ALIBABA CLOUD

Alibaba Cloud

DataWorks

数据集成

文档版本：20220713

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.数据集成概述

数据集成是稳定高效、弹性伸缩的数据同步平台，致力于提供复杂网络环境下、丰富的异构数据源之间高速稳定的数据移动及同步能力。

费用说明

运行数据集成任务产生的费用由两部分组成：

- DataWorks相关收费
  - 数据集成资源组（独享数据集成资源组，如果使用公共数据集成（调试）资源组，则还包括公共数据集成资源组费用）。
  - 任务调度资源组费用（独享调度资源组，公共调度资源组）。
  - 公网流量费用（如果任务走公网进行的数据传输）。
  - DataWorks的版本使用费（如使用相关收费版本）。

> **说明** 此类费用体现在DataWorks产品相关账单中。

- 非DataWorks的收费

  除上述DataWorks相关收费外，也可能会产生由数据同步任务相关配置引起的账单和费用，如同步上下游数据库、计算引擎系统计算和存储费用，所需网络服务费用（如高速通道、共享带宽、EIP）等等，此类收费不属于DataWorks相关收费范畴。账单也不会体现在DataWorks产品下。请在任务配置后，确认您使用DataWorks以外相关资源所产生的任务及相关费用问题。

> **说明** DataWorks计费项说明详情可参考文档：计费逻辑说明。

使用限制

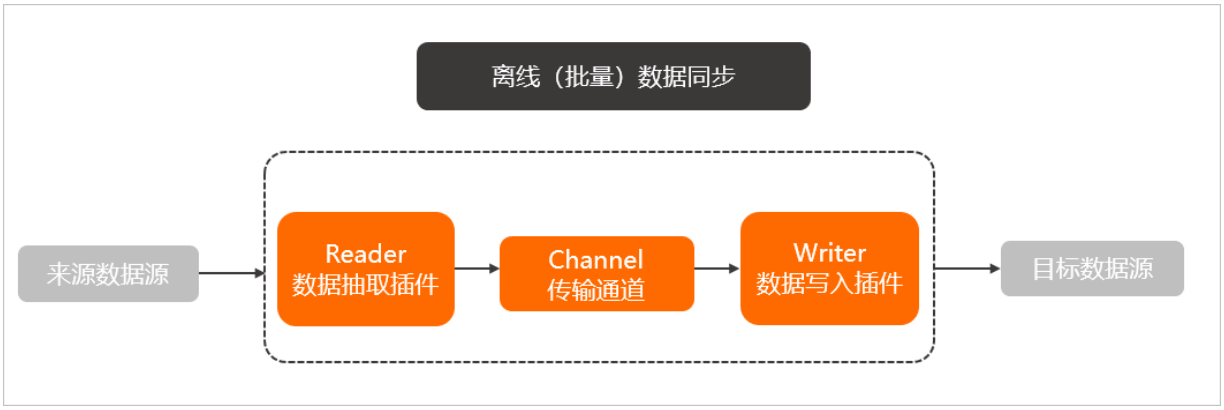
- 支持且仅支持结构化（例如RDS、DRDS等）、半结构化、无结构化（OSS、TXT等，要求具体同步数据必须抽象为结构化数据）的数据的同步。即数据集成仅支持传输能够抽象为逻辑二维表的数据同步，不支持同步OSS中存放完全非结构化的数据（例如一段MP3）至MaxCompute。
- 支持单地域内及部分跨地域的数据存储相互同步、交换的数据同步需求。

  部分地域之间可以通过经典网络传输，但不能保证其连通性。如果测试经典网络不通，建议您使用公网方式进行连接。
- 数据集成仅完成数据同步（传输），本身不提供数据流的消费方式。
- 数据集成同步仅支持at least once，不支持exact once，即不能保证数据重复，只能依赖主键+目的端能力来保证。

离线（批量）同步简介

> **说明** DataWorks的离线同步暂不支持跨时区同步数据。如果同步任务中同步的数据源与使用的DataWorks资源组不在同一个时区，则会导致同步的数据有误。

数据集成主要用于离线（批量）数据同步。离线（批量）的数据通道通过定义数据来源和去向的数据源和数据集，提供一套抽象化的数据抽取插件（Reader）、数据写入插件（Writer），并基于此框架设计一套简化版的中间数据传输格式，从而实现任意结构化、半结构化数据源之间数据传输。



数据同步的开发模式

您可以通过以下两种模式进行数据同步开发：

- **向导模式**：提供向导式的开发引导，通过可视化的填写和下一步的引导，助您快速完成数据同步任务的配置工作。向导模式的学习成本低，但无法支持部分高级功能。详情请参见通过向导模式配置离线同步任务。
- **脚本模式**：您可以通过直接编写数据同步的JSON脚本来完成数据同步开发，适合高级用户，学习成本较高。脚本模式可以提供更丰富灵活的能力，实现精细化的配置管理。详情请参见通过脚本模式配置离线同步任务。

> 说明
> - 向导模式生成的代码可以转换为脚本模式，该转换为单向操作，转换完成后无法恢复至向导模式。
> - 代码编写前，您需要配置数据源和创建目标表。

网络连通说明

数据集成通过数据集成资源组将您的数据从源端同步到目标端，但在数据同步前，您需要先保障数据集成资源组（执行数据同步的机器）与数据库的网络连通性。

数据集成支持复杂网络环境下的数据库进行异构数据源间的同步，您可以根据数据库所在网络环境，选择对应的网络解决方案，来保障同步使用的资源组与您数据库的网络连通，详情可参考文档：配置资源组与网络连通。

目前的数据库的网络类型大致分为：经典网络、专有网络（VPC）和本地IDC网络：

- 经典网络：统一部署在阿里云的公共基础网络内，网络的规划和管理由阿里云负责。
- 专有网络：基于阿里云构建出一个隔离的网络环境。您可以完全掌控自己的虚拟网络，包括选择自有的IP地址范围，划分网段以及配置路由表、网关。

  专有网络构建了一个隔离的网络环境，您可以自定义IP地址范围、网段、网关等参数。随着专有网络应用的推广，数据集成提供了RDS（MySQL、PostgreSQL和SQLServer）、PolarDB、DRDS、HybridDB for MySQL、AnalyticDB for PostgreSQL和AnalyticDB for MySQL3.0数据源之间的反向代理自动检测功能。在专有网络下，您无需购买一台和VPC同网络的ECS，即可通过反向代理自动检测连通网络。

  PPAS、OceanBase、Redis、MongoDB、Memcache、TableStore和HBase等阿里云其它非RDS的数据库，在专有网络下配置数据同步任务时，需要购买同网络的ECS，才可以通过ECS连通网络。
- 本地IDC网络：您自身构建机房的网络环境，与阿里云网络隔离。

经典网络和专有网络相关问题请参见经典网络和VPC常见问题。

> 说明 网络连接可以支持公网连接，请注意公网带宽的速度和相关网络费用消耗。无特殊情况不建议使用，关于公网计费详情可参考文档：公网流量计费说明。

基本概念

- 并发数

  并发数是数据同步任务中，可以从源并行读取或并行写入数据存储端的最大线程数。
- 限速

  限速是数据集成同步任务可以达到的传输速度限制。
- 脏数据

  脏数据是对于业务没有意义，格式非法或者同步过程中出现问题的数据。单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。例如，源端是VARCHAR类型的数据写到INT类型的目标列中，导致因为转换不合理而无法写入的数据。您可以在同步任务配置时，控制同步过程中是否允许脏数据产生，并且支持控制脏数据条数，即当脏数据超过指定条数时，任务失败退出。
- 数据源

  DataWorks所处理的数据的来源，可能是一个数据库或数据仓库。DataWorks支持各种类型的数据源，并且支持数据源之间的转换。

  在数据集成同步任务配置前，您可以在DataWorks**数据源管理**页面，配置好您需要同步的源端和目标端数据库或数据仓库的相关信息，并在同步过程中，通过选择数据源名称来控制同步读取和写入的数据库或数据仓库。

参考文档

- 数据同步任务配置详情请参见创建数据同步任务。
- 如果需要处理OSS等非结构化数据，请参见MaxCompute访问OSS数据。
- DataWorks可以通过免费传输能力（默认任务资源组）进行海量数据上云，但默认资源组无法实现传输速度存在较高要求或复杂环境中的数据源同步上云的需求。您可以新增独享数据集成资源或者数据集成自定义资源组运行数据同步任务，解决DataWorks默认资源组与您的数据源不通的问题，或实现更高速度的传输能力。详情请参见新增和使用独享数据集成资源组和新增和使用自定义数据集成资源组。

2.同步前准备
2.1. 数据源配置与管理
2.1.1. 数据源身份认证
2.1.1.1. 配置第三方身份认证

DataWorks的数据同步功能支持第三方身份认证机制，您需要提前在DataWorks的认证文件管理页面上传认证文件，并在配置数据源同时开启第三方认证功能，使得只有可信的应用和服务才能访问数据资源。本文为您介绍如何上传和引用认证文件。

背景信息

第三方认证机制用于用户和服务的强身份验证，通过该机制，可以有效的避免不受信任的程序或服务来获取数据访问权限，提高数据同步过程中访问数据资源的安全性。DataWorks的认证文件管理功能，可以为您提供统一入口来管理认证文件，您可以通过认证文件管理页面上传和查看认证文件引用关系等。

使用限制

目前仅支持Kerberos认证机制（后续会逐步支持其他认证机制，敬请期待），详情请参见附录：配置Kerberos认证。

注意事项

一般情况下证书有自己的有效期，请留意您上传的证书有效期，如果证书过期将会导致对应的数据同步任务因无法获得授权而失败，请及时更换为最新有效期的证书。

上传认证文件

使用认证功能前，需要提前准备好认证文件并上传至认证文件管理页面。

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
4. 在左侧导航栏，单击**数据源 > 认证文件管理**。
5. 单击**认证文件管理**页面右上方的**上传kerberos文件**。
6. 在**上传kerberos文件**对话框，单击**上传文件**选择对应文件并填写相应**文件描述**，单击**确定**。

引用认证文件

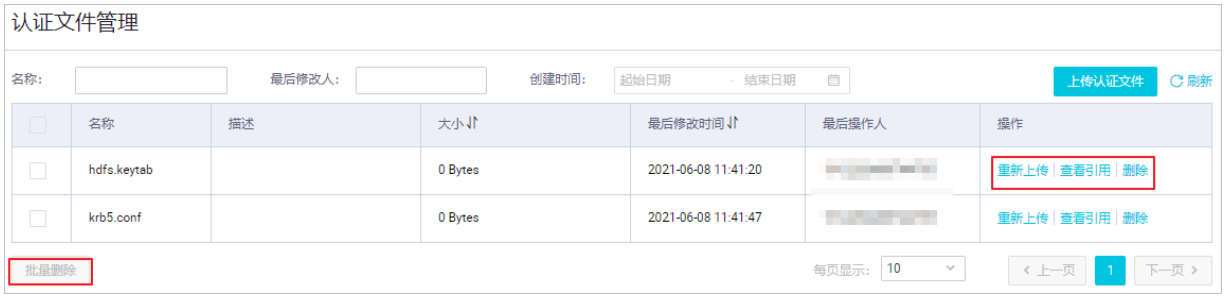
如果需要使用第三方身份认证功能，请在配置数据源页面选择开启特殊身份认证方式，配置相关参数并引用相关认证文件。目前DataWorks仅支持Kerberos认证，具体介绍请参见附录：配置Kerberos认证。

以下以HDFS数据源配置Kerberos认证为例，列举Kerberos认证的关键配置，关于数据源详细配置，请参见配置数据源。

配置项	说明
特殊认证方式	选择**特殊认证方式**为**Kerberos认证**。
keytab文件	在下拉列表选择已上传的keytab文件。如果需要重新上传认证文件，请单击**新增认证文件**。
conf文件	在下拉列表选择已上传的conf文件。如果需要重新上传认证文件，请单击**新增认证文件**。
principal	Kerberos认证的主体，包括主名称、实例和领域三部分，格式为：主名称/实例名@领域名。例如****/hadoopclient@**.*** 。

其他操作

您还可以在**认证文件管理**页面进行认证文件的批量删除、重新上传和查看引用等操作。



2.1.1.2. 附录：配置Kerberos认证

DataWorks的数据同步功能目前仅支持Kerberos认证，配置kerberos认证后，可以仅对受信任的应用和服务提供认证，使得只有经过认证的应用和服务才能访问数据资源。本文为您介绍Kerberos的认证机制。

背景信息

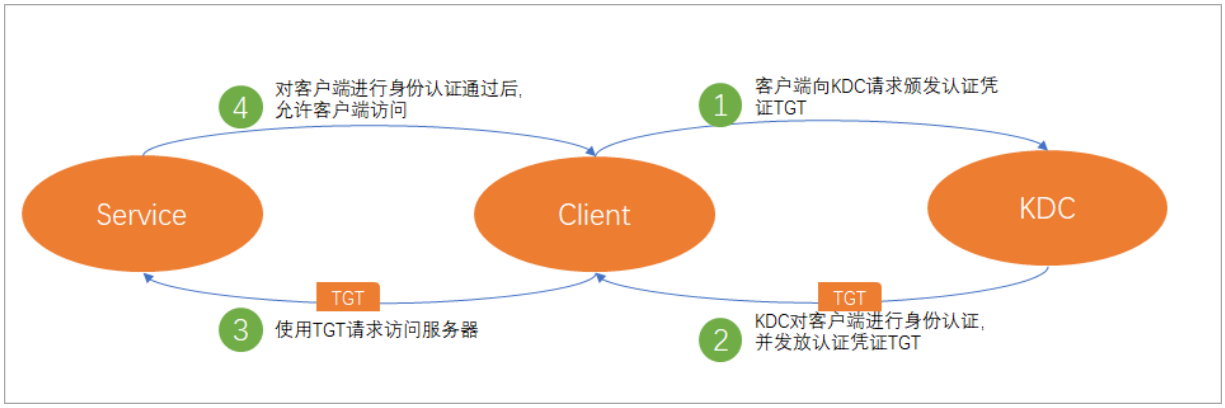
Kerberos协议主要用于计算机网络的身份鉴别（Authentication），其特点是用户只需输入一次身份验证信息就可以凭借此验证获得的票据（Ticket-Granting Ticket，网络授权凭证）访问多个服务，即SSO（Single Sign On，单点登录）。使用Kerberos协议时，会在每个Client和Service之间建立共享密钥，服务之间使用密钥进行通信，避免不受信任的服务或应用访问数据资源，因此该协议具有较高的安全性。

使用限制

- Kerberos认证功能仅支持CDH集群6.X版本，其他版本或者自建集群未经过Kerberos认证测试，可能会导致认证失败。
- Kerberos认证功能仅支持HBase、HDFS和Hive数据源（后续会逐步支持其他数据源类型，敬请期待）。
- Kerberos认证功能仅支持在独享数据集成资源组上使用。

Kerberos认证原理

Kerberos是一种基于对称密钥的第三方认证协议，客户端和服务器均是依赖KDC（Kerberos的服务端程序，即密钥分发中心）来进行身份认证。有关Kerberos的详细介绍请参见概述。



如上图所示，在DataWorks上进行Kerberos认证分为如下四个阶段：

1. 客户端请求TGT：当客户端用户（Principle）访问已开启Kerberos认证的数据源时，会先向KDC请求颁发一个认证凭证TGT，作为客户端向KDC请求特定服务的身份证明。
2. KDC发放TGT：KDC收到请求后，先对客户端进行身份认证，认证通过后，KDC会已加密形式为客户端发放一个有使用期限的认证凭证TGT。
3. 客户端请求访问服务器：客户端获取TGT后，会根据需要访问的服务名称向服务器请求访问特定服务资源。
4. 服务器认证客户端：服务器收到请求后，先对客户端进行身份认证，认证通过后，才会允许客户端正常访问服务资源。

Kerberos认证过程中需要使用keytab认证文件和krb5.conf配置文件完成认证行为，其中krb5.conf文件主要用于存储KDC服务器的相关配置，keytab文件用于存储资源主体的身份验证凭据，包含pricipals和加密principle key。使用Kerberos认证前，需要先将这两个文件上传到认证文件管理页面，并在数据源配置页面完成认证文件引用和配置，即可使用Kerberos认证。上传认证文件和不同数据源类型的Kerberos配置参考请参见配置第三方身份认证和支持Kerberos认证的数据源。

支持Kerberos认证的数据源

Kerberos支持的数据源类型及配置指引如下所示：

数据源类型	配置指引
HBase	配置HBase数据源
HDFS	配置HDFS数据源
Hive	配置Hive数据源

2.1.2. 访问模式配置说明：RAM角色授权模式

2.1.2.1. DataWorks数据集成服务关联角色

数据集成支持RAM角色授权模式。本文为您介绍如何获取DataWorks数据集成相关的RAM角色列表、删除服务关联角色，以及子账号如何创建服务关联角色所需要的权限。

应用场景

当您通过RAM角色授权模式创建DataWorks数据源时，请选择相关的自定义RAM角色来访问数据源，例如OSS。

您需要授权DataWorks服务为AliyunServiceRoleForDataWorksDI服务的关联角色，以获取与DataWorks数据集成相关的RAM角色列表，供您选择。

AliyunServiceRoleForDataWorksDI介绍

- 角色名称： AliyunServiceRoleForDataWorksDI
- 角色权限策略： AliyunServiceRolePolicyForDataWorksDI
- 权限说明：允许DataWorks访问与DataWorks数据集成相关的RAM角色列表。
- 使用该权限的作用：罗列与DataWorks数据集成相关的RAM角色列表。

```
{
    "Version": "1",
    "Statement": [
        {
            "Action": [
                "ram:ListRoles",
                "ram:GetRole"
            ],
            "Resource": "*",
            "Effect": "Allow"
        }
    ]
}
```

删除服务关联角色

您可以随时删除AliyunServiceRoleForDataWorksDI角色。如果您删除了该角色，则相关任务在DataWorks创建数据源时，无法罗列并选择DataWorks数据集成相关的RAM角色。详情请参见删除服务关联角色。

子账号创建服务关联角色所需要的权限

子账号被授权DataWorksFullAccess策略或如下策略，即可创建服务关联角色AliyunServiceRoleForDataWorksDI。

```
{
    "Version": "1",
    "Statement": [
        {
            "Action": "dataworks:*",
            "Resource": "*",
            "Effect": "Allow"
        },
        {
            "Action": "ram:CreateServiceLinkedRole",
            "Resource": "*",
            "Effect": "Allow",
            "Condition": {
                "StringEquals": {
                    "ram:ServiceName": "di.dataworks.aliyuncs.com"
                }
            }
        }
    ]
}
```

2.1.2.2. 通过RAM角色授权模式配置数据源

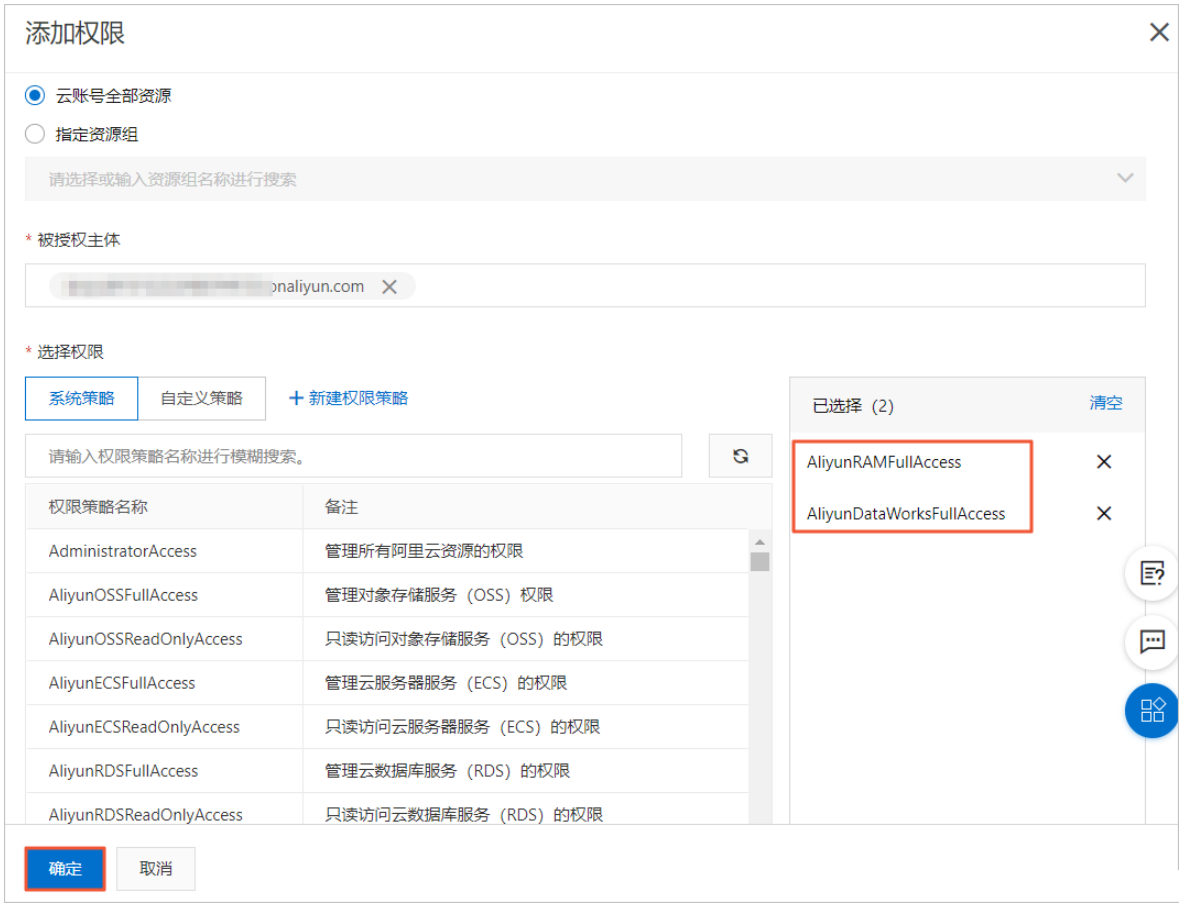
本文以OSS数据源为例，为您介绍如何通过RAM角色授权模式配置数据源，提高云上数据的安全性。

前提条件

如果您登录的是RAM用户并需要完成本文的所有操作流程，请首先为RAM用户授予DataWorksFullAccess和AliyunRamFullAccess权限。如果您使用的是阿里云主账号，请忽略该前提条件。

1. RAM用户登录RAM访问控制。
2. 在左侧导航栏，单击**用户**。
3. 单击相应RAM用户后的**添加权限**。

4. 在**添加权限**对话框中，选择**授权范围**为**云账号全部资源**，在**选择权限**区域分别单击DataWorksFullAccess和AliyunRamFullAccess。



5. 单击**确定**。

背景信息

数据源是数据同步任务的基础，同时关系着企业云上数据的安全性。DataWorks支持您使用更高安全系数的**RAM角色授权模式**配置并访问OSS、AnalyticDB for MySQL 2.0、LogHub、OTS和Hologres等部分数据源，以提升云上数据的安全性，避免数据源被滥用、密钥泄露等情况。

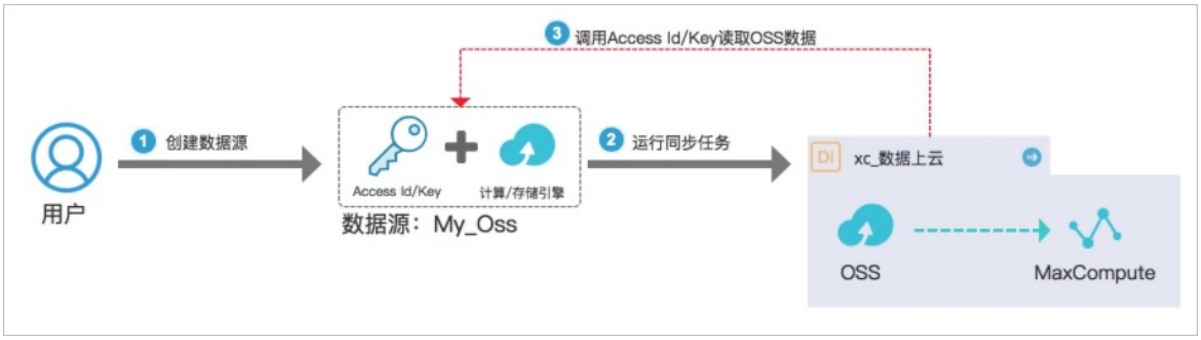
数据源的访问模式包括RAM角色授权模式和Access Key模式。本文为您介绍的是通过RAM角色授权模式配置数据源，您可以根据业务需求进行选择。Access Key模式和RAM角色授权模式的实现原理如下：

- **Access Key模式**

  在安全性较低的AK（AccessKeyID和AccessKeySecret）模式下，您只需要在页面输入阿里云主账号或子账号的AK，即可完成配置。

以OSS数据源为例，您在**配置数据源**页面输入具有访问OSS某个Bucket权限的账号AK，即可完成配置。

在执行、调度运行同步任务时，您可以通过该AK来访问OSS并读写数据。

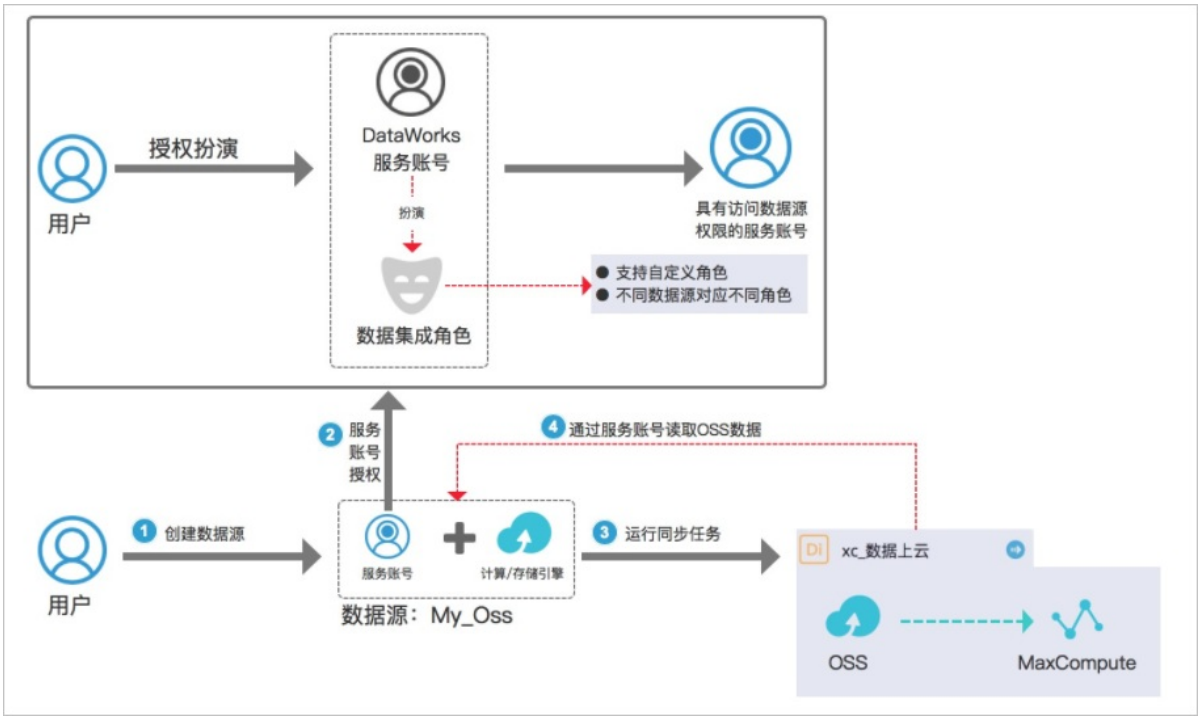


> 说明 Access Key模式下，一旦该云账号的AK被泄露，会导致OSS数据同时被泄露。

- **RAM角色授权模式**

  RAM角色授权模式致力于提供安全性更高的数据源访问方式，并无需生成AK，能够有效地规避AK泄露的风险。

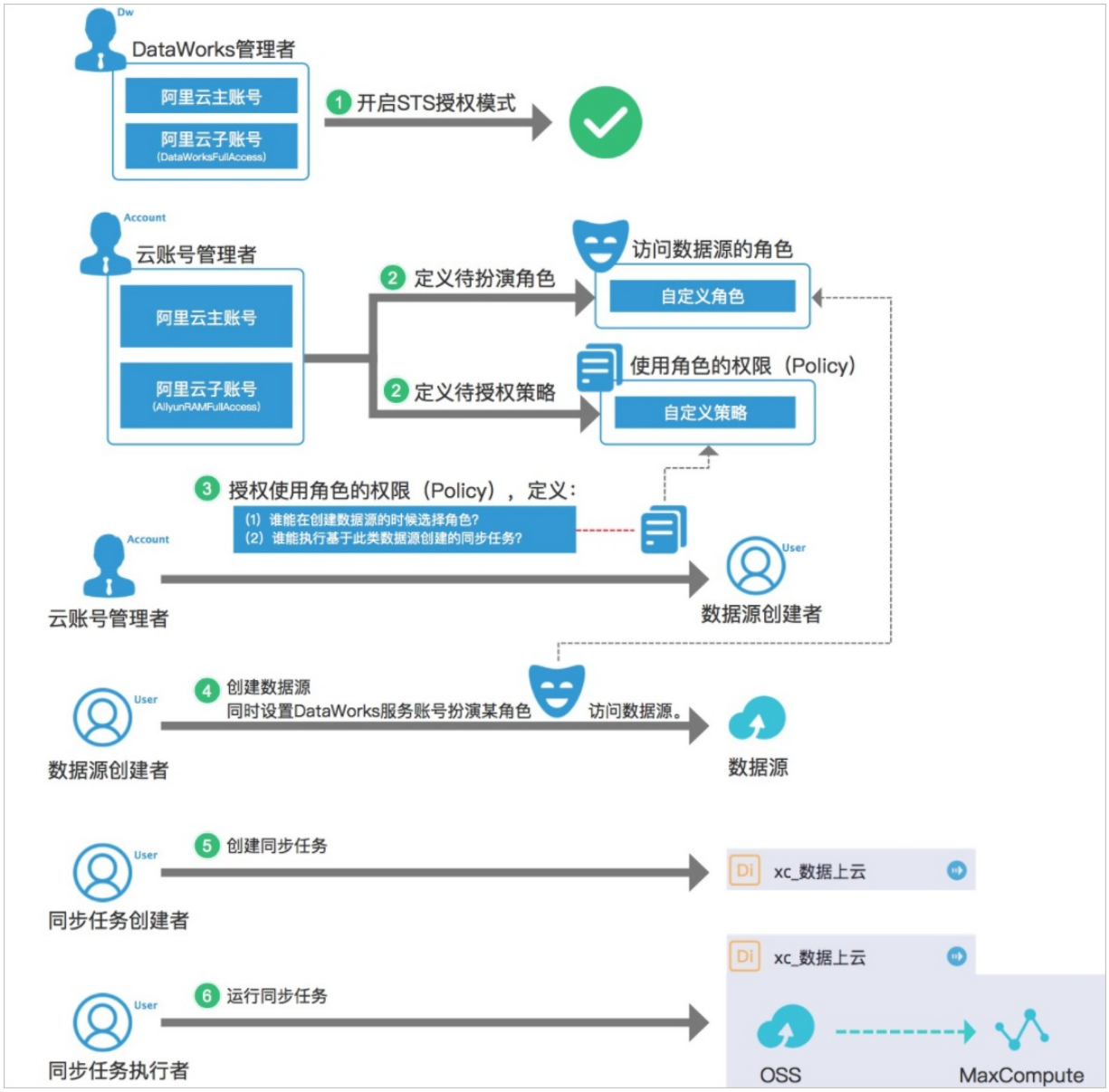
在**RAM角色授权模式**下，您只需要授权DataWorks服务账号为具有访问OSS权限的角色，即可实现无AK访问OSS数据源。



同时，为了兼顾企业级用户的诉求，允许您对不同数据源设置具有能够权限范围的角色，实现更专业的权限管控。

流程介绍

为方便子账号可以替代主账号完成全链路操作，本说明增加阿里云子账号在每一个步骤的操作条件。RAM角色授权模式的操作流程如下。



1. 阿里云主账号或RAM用户（被授予DataWorksFullAccess权限）登录DataWorks数据集成页面，开启**RAM角色授权模式**。
2. 阿里云主账号或RAM用户（被授予 AliyunRamFullAccess角色）登录RAM访问控制，分别定义待扮演角色和待授权策略。
   - 待扮演角色：您需要创建自定义角色供DataWorks服务账号扮演。扮演后，DataWorks服务账号即可在角色具备的权限范围内，访问OSS数据源。
   - 待授权策略：您需要创建包含 PassRole相关权限点的策略，用于授权某个使用者使用某个角色创建数据源或运行同步任务的权限。
3. 阿里云主账号或RAM用户（被授予 AliyunRamFullAccess权限）登录RAM访问控制，授权第4步和第6步的RAM用户使用角色的权限。

   > **说明** 如果未被授权的RAM用户通过**RAM角色授权模式**创建数据源，则通过RAM角色授权模式数据源配置的同步任务都将运行失败。

4. 数据源的创建者登录DataWorks数据集成页面，以**RAM角色授权模式**创建数据源，实现运行同步任务时，以DataWorks服务账号扮演某个角色访问OSS数据源的最终效果。

> **说明** 本步骤的创建者必须装在步骤3中被授权后，才能进行操作。

5. 数据同步任务的创建者登录DataStudio（数据开发）页面，基于配置的数据源创建同步任务。
6. 执行者在DataStudio或运维中心页面，执行数据同步任务。

> **说明** 本步骤的创建者必须装在步骤3中被授权后，才能进行操作。

操作步骤

1. 开启RAM角色授权模式。

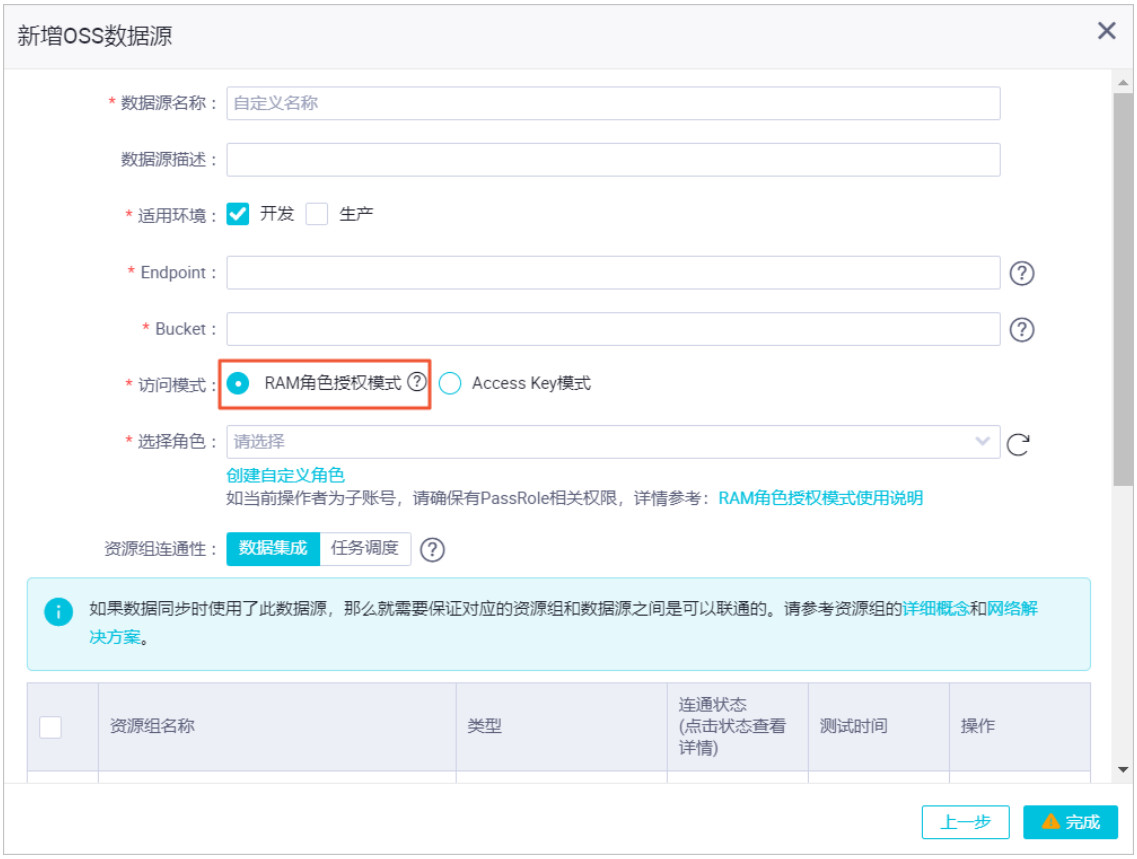
   在初次选择**RAM角色授权模式**时，需要执行一次性开启RAM角色授权，以便允许DataWorks服务账号以角色扮演的方式访问数据源。一次性开启RAM角色授权的操作如下：

   i. 阿里云主账号或RAM用户（被授予 DataWorksFullAccess权限）登录数据源管理页面。

   ii. 单击页面右上方的**新增数据源**。

   iii. 在**新增数据源**对话框中，选择数据源类型为**OSS**。

   iv. 在**新增OSS数据源**对话框中，选择**访问模式**为**RAM角色授权模式**。



v. 在**警告**对话框中，单击**开启授权**，进行授权操作即可。

2. 创建待扮演角色。

   根据实际的安全场景，您需要自定义不同的角色，以匹配不同的数据源。

   > **说明** 仅阿里云主账号和被授予 AliyunRAMFullAccess权限的RAM用户可以操作本步骤。

   本文以下述场景为例，为您介绍如何创建待扮演角色。

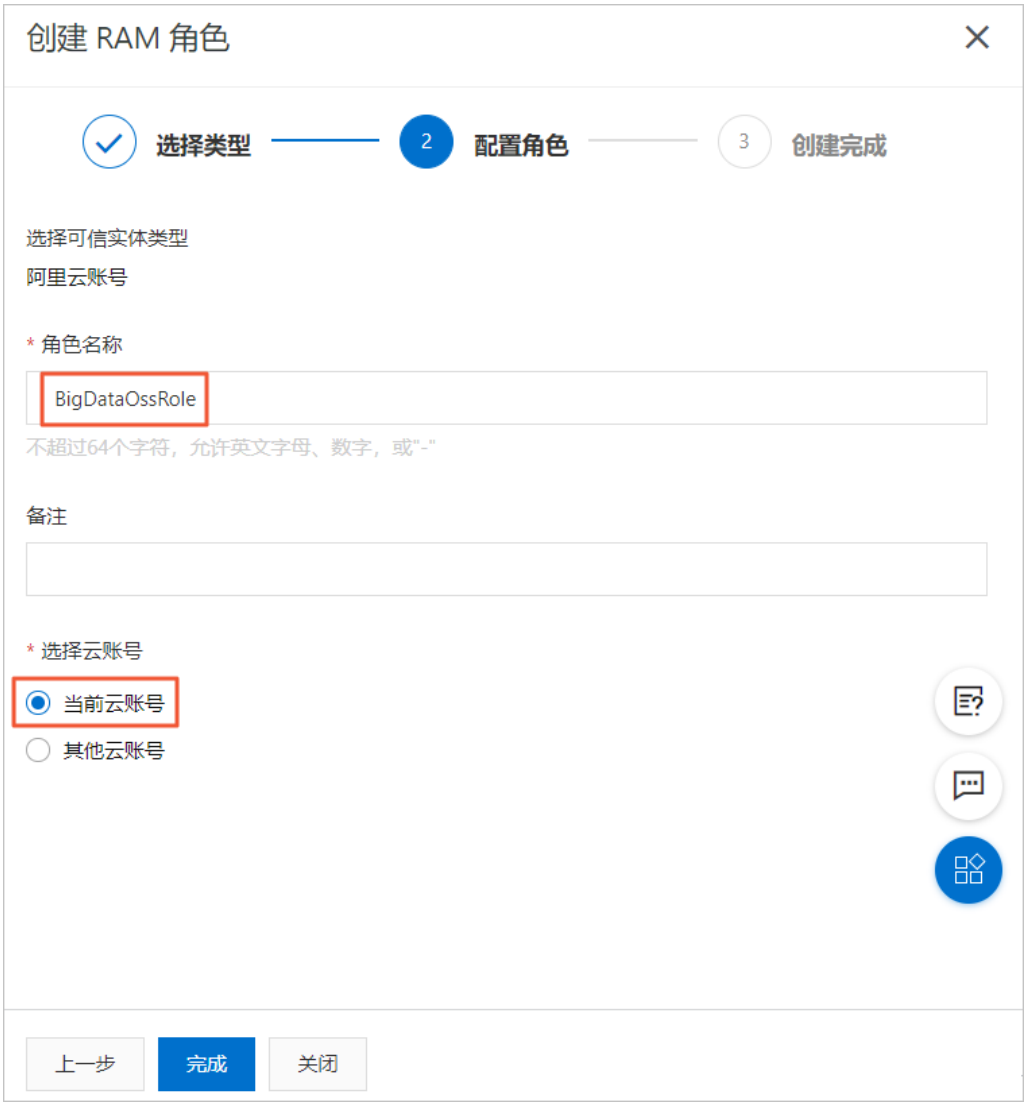
   某企业有100个Bucket，存放该企业的所有数据，但大数据团队只需要使用其中2个Bucket的数据。如果使用预设的AliyunDataWorksAccessingOSSRole角色，可能导致其它98个Bucket被大数据团队访问，存在管理风险。

   因此，云账号负责人可以为大数据团队创建自定义角色 BigDataOssRole，并限制可以使用角色的人员为大数据团队的相关人鱼，实现团队间的权限管控。

   i. 登录RAM访问控制。
   ii. 在左侧导航栏，单击**RAM角色管理**。
   iii. 单击**创建RAM角色**。

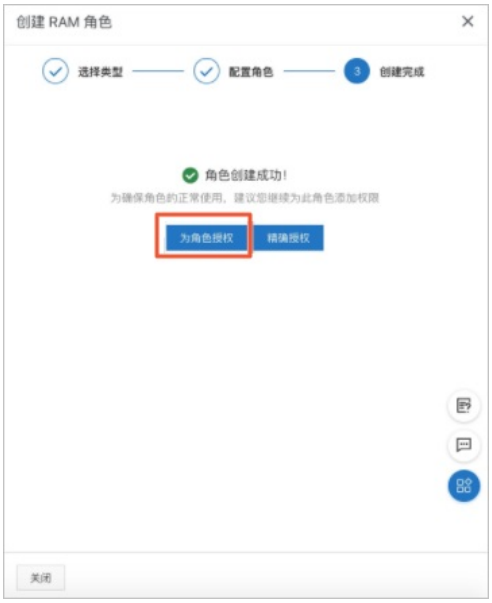
iv. 在**创建RAM角色**对话框的**选择类型**区域，选择**当前可信实体类型**为**阿里云账号**，单击**下一步**。

v. 在**配置角色**区域，输入**角色名称**为BigDataOssRole，并**选择云账号**为**当前云账号**。



vi. 单击**完成**。

vii. 在**创建完成**区域，单击**为角色授权**。



viii. 在**添加权限**对话框中，单击**新建权限策略**，创建一个策略。详情请参见创建自定义权限策略。

以策略（Policy）内容为完全管理某2个Bucket内的数据，即可读写为例，策略内容如下。

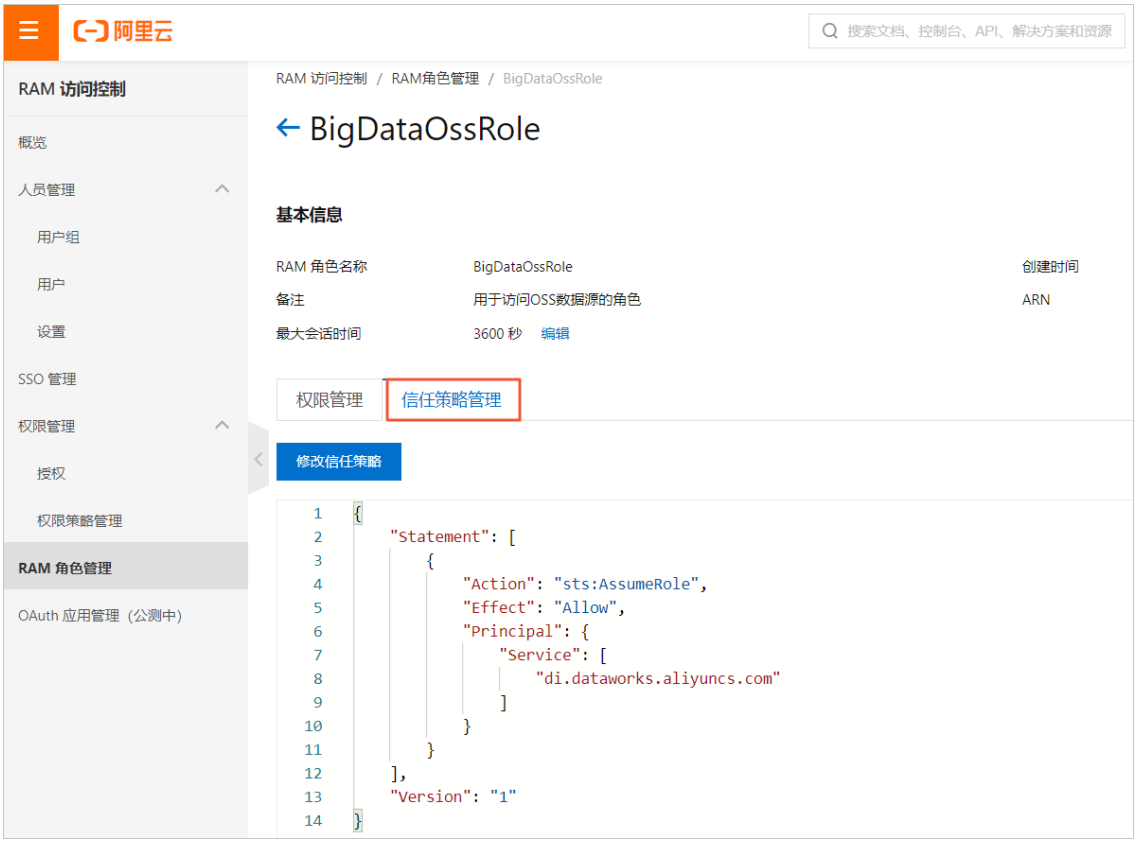
```
{
    "Version": "1",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "oss:GetObject",
                "oss:ListObjects",
                "oss:GetObjectMetadata",
                "oss:GetObjectMeta",
                "oss:GetBucketAcl",
                "oss:GetBucketInfo",
                "oss:PutObject",
                "oss:DeleteObject",
                "oss:PutBucket"
            ],
            "Resource": [
                "acs:oss:*:*:bucket_name_1",
                "acs:oss:*:*:bucket_name_2"
            ]
        }
    ]
}
```

ix. 在RAM角色管理页面，单击BigDataOSSRole。

在BigDataOSSRole的**基本信息**页面，单击**信任策略管理** > **修改信任策略**，修改BigDataOSSRole角色的授信策略为如下内容，单击**确定**，以便允许DataWorks数据集成服务账号扮演该角色。

> **注意** 本步骤必须设置，否则将无法使用角色。

```
{
    "Statement": [
        {
            "Action": "sts:AssumeRole",
            "Effect": "Allow",
            "Principal": {
                "Service": [
                    "di.dataworks.aliyuncs.com"
                ]
            }
        }
    ],
    "Version": "1"
}
```



3. 授权可以使用角色的人员。

确认所有待扮演角色后，您需要将包含PassRole权限点的Policy授权给相关人员，相关人员才可以通过该角色创建数据源和执行同步任务。同时，您可以根据实际场景要求配置人员和角色的对应关系，即配置哪些人员可以使用哪些角色。

- 策略模板一：您可以参照如下模板创建策略，该模板允许被授权者使用所有与DataWorks数据集成相关的角色，请谨慎授权。

```
{
            "Action": "ram:PassRole",
            "Resource": "*",
            "Effect": "Allow",
            "Condition": {
                "StringEquals": {
                    "acs:Service": "di.dataworks.aliyuncs.com"
                }
            }
}
```

- 策略模板二：您可以根据实际的安全场景需求，自定义包含PassRole权限点的Policy，配置人员和角色的对应关系。

  > **说明** 仅阿里云主账号和被授予AliyunRAMFullAccess权限的RAM用户可以操作本步骤。

  场景示例：如上文待扮演角色的场景示例所述，当云账号管理者为大数据团队定义了BigDataOssRole角色后，需要指定仅相关人员才能使用该角色。您可以自定义Policy为BigDataOssRoleAllowUse，并授权给相关人员。

  创建策略的操作如下：

  a. 登录页面，单击**创建权限策略**。

b. 在**新建自定义权限策略**页面，输入**策略名称**为**BigDataOssRoleAllowUse**，并选择**配置模式**为**脚本配置**，输入如下策略。

```
{
    "Version": "1",
    "Statement": [
        {
            "Action": "ram:PassRole",
            "Resource": "acs:ram::19122324****:role/BigDataOssRole",
            "Effect": "Allow",
            "Condition": {
                "StringEquals": {
                    "acs:Service": "oss.aliyuncs.com",
                    "acs:Service": "di.dataworks.aliyuncs.com"
                }
            }
        }
    ]
}
```

> **说明** 请替换上述策略中的UID（19122324****）为当前登录的阿里云主账号的UID。

c. 管理员授权BigDataOssRoleAllowUse策略给允许使用BigDataOssRole角色的RAM用户。

被授权BigDataOssRoleAllowUse策略的RAM用户可以通过BigDataOssRole角色来创建数据源（将BigDataOssRole作为数据源的访问身份）、运行同步任务。

4. 创建数据源。

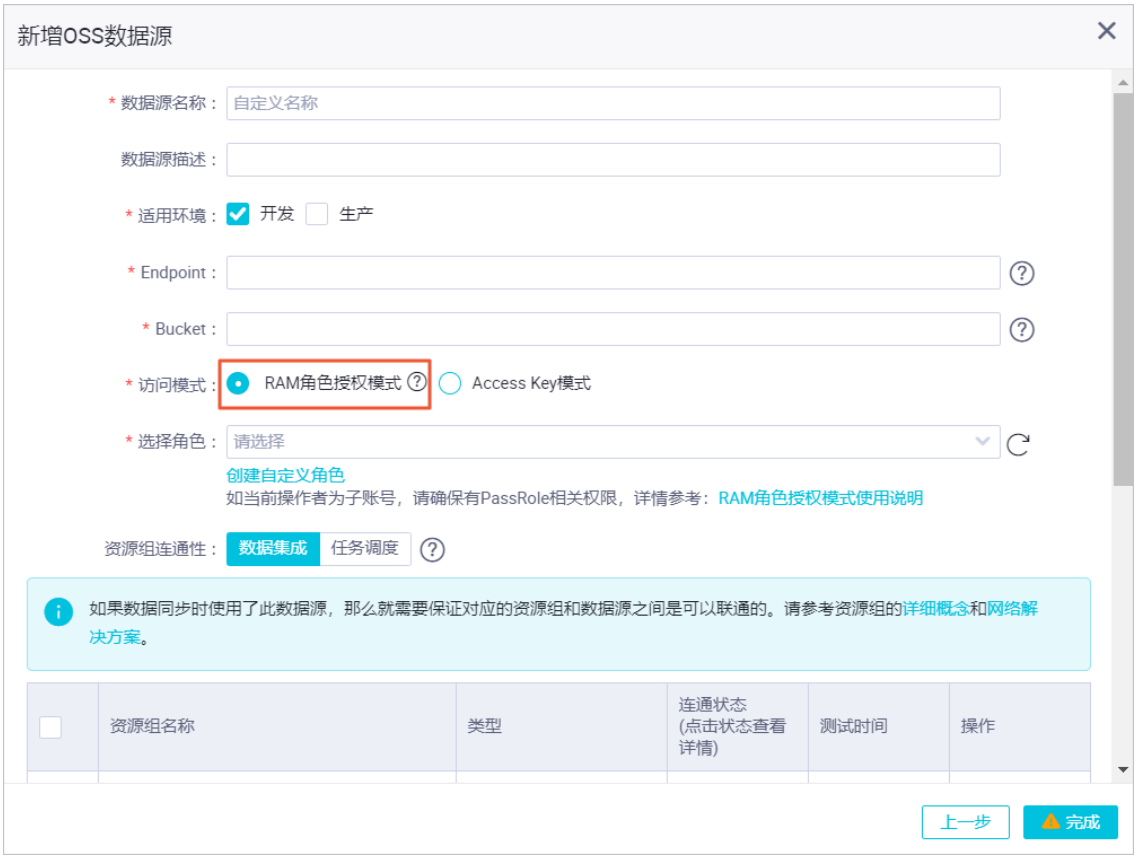
当云账号管理者完成对数据源创建者的授权后，即可创建数据源。

i. 阿里云主账号和RAM用户（被授予 DataWorksFullAccess权限）登录数据源管理页面。

ii. 单击页面右上方的**新增数据源**。

iii. 在**新增数据源**对话框中，选择数据源类型为**OSS**。

iv. 在**新增OSS数据源**对话框中，选择**访问模式**为**RAM角色授权模式**，配置各项参数。



参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Endpoint	OSS Endpoint信息，格式为 `http://oss.aliyuncs.com` ，OSS服务的Endpoint和地域有关。访问不同的地域时，需要填写不同的域名。 **说明** Endpoint的正确的填写格式为 `http://oss.aliyuncs.com` ，但 `http://oss.aliyuncs.com` 在OSS前加上Bucket值，以点号的形式连接。例如 `http://xxx.oss.aliyuncs.com` ，测试连通性可以通过，但同步会报错。
Bucket	相应的OSS Bucket信息，指存储空间，是用于存储对象的容器。 您可以创建一个或多个存储空间，每个存储空间可添加一个或多个文件。 您可以在数据同步任务中查找此处输入的存储空间中相应的文件，没有添加的存储空间，则不能查找其中的文件。
访问模式	此处选择**RAM角色授权模式**，通过STS授权的方式允许云产品服务账号扮演相关角色来访问数据源，具备更高安全性。

参数	描述
选择角色	从**选择角色**下拉列表中选择RAM角色。

v. 在**数据集成**页签下，单击相应资源组后的**测试连通性**。

数据同步时，一个任务只能使用一种资源组。您需要测试每种资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

vi. 测试连通性通过后，单击**完成**。

创建数据源后，开发人员可以进入DataStudio（**数据开发**）页面，基于该数据源创建同步任务，详情请参见配置同步任务。

在DataStudio（**数据开发**）页面执行任务或调度运行任务时，请确保任务执行者在步骤3（授权可以使用角色的人员）中，已被授权，避免任务运行失败。

2.1.3. 管理数据源权限

您可以在数据源管理页面，分享数据源权限给相应的工作空间，并进入被分享的工作空间查看该数据源。本文为您介绍如何管理数据源权限及查看分享的数据源。

背景信息

通常数据源会承载数据的具体地址、账户和密码等敏感信息，但普通开发人员仅需要引用数据源进行数据访问。如果数据源中透露过多敏感信息或允许所有人修改数据源，会造成安全风险。例如，多人修改数据源信息，可能导致数据源报错，以致于引用该数据源的任务运行失败。

因此数据集成提供更加严格的权限管控功能，仅支持数据源的创建者对数据源进行权限管理，指定某个工作空间或者某个人对数据源的权限。

进入数据源管理页面

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源**，进入**数据源管理**页面。
2. 在**数据源管理**页面，单击相应数据源后的**权限管理**。
3. 在**数据源权限管理**对话框中，配置各项参数。

序号	参数	描述

序号	参数	描述
①	**工作空间**	为您展示当前操作人员参与的所有工作空间，并且展示空间内的所有人员。您可以针对整个工作空间，或者工作空间下的某些人员进行数据源分享： ○ 如果数据源未设置过权限，则继承以前数据源的权限控制。 ○ 当针对某个工作空间进行权限设置时，将会影响该工作空间内的所有人员。即使后续添加新成员，也会遵循同样的权限。但是，给工作空间授权后，您可以针对某个具体的人员单独设置不同的权限。例如，将一个数据源分享给整个工作空间为**不可编辑**权限，但设置其中的某个成员为**可编辑**权限。 ○ 支持控制本工作空间成员的权限分享。 ○ 仅数据源的创建者可以编辑、分享数据源，其他人员（包括项目管理员）均无法编辑。 ○ 项目管理员同样需要设置权限后，才可以使用数据源。
②	**工作空间类型**	为您展示各工作空间的类型，包括**简单**和**标准**两种类型。
③	**权限**	数据源的权限控制包括以下三种类型： ○ **无权限**：相应的工作空间或个人无法查看该数据源。 ○ **不可编辑**：可以使用该数据源，但不允许进行编辑，且无法查看数据源内的详细配置信息。 ○ **可编辑**：可以使用和编辑该数据源。 **注意** 由于数据源分享只是进行引用投射，此处是针对原始数据源进行编辑，需要注意开放编辑权限的适用人员。
④	批量操作	批量选中工作空间或成员后，您可以设置**批量不可编辑**、**批量可编辑**和**批量无权限**。

4. 单击**确定**。

   数据源支持跨工作空间分享，分享规则如下：

   - 从简单模式工作空间分享至简单模式工作空间：
     - 源端的简单模式工作空间升级为标准模式工作空间时：源工作空间会分享生产环境的数据源。
     - 目标简单模式工作空间升级为标准模式工作空间时：数据源变为两个，分别对应开发环境和生产环境，且内容一致。
   - 从简单模式工作空间分享至标准模式工作空间：数据源在标准模式工作空间下变为两个，分别对应开发环境和生产环境，且内容一致。
   - 从标准模式工作空间分享至标准模式工作空间：不同环境的数据源，分享后在目标工作空间内，仍然对应各自的环境。
   - 从标准模式工作空间分享至简单模式工作空间：
     - 源工作空间的生产或开发环境下的数据源，均支持分享。分享至目标数据源后，仅保留一个数据源，且最新分享的数据源会覆盖之前的数据源。
     - 如果目标工作空间升级为标准模式工作空间，则该数据源会变为两个，分别对应开发和生产环境，且内容一致，均与源数据源保持分享关系。

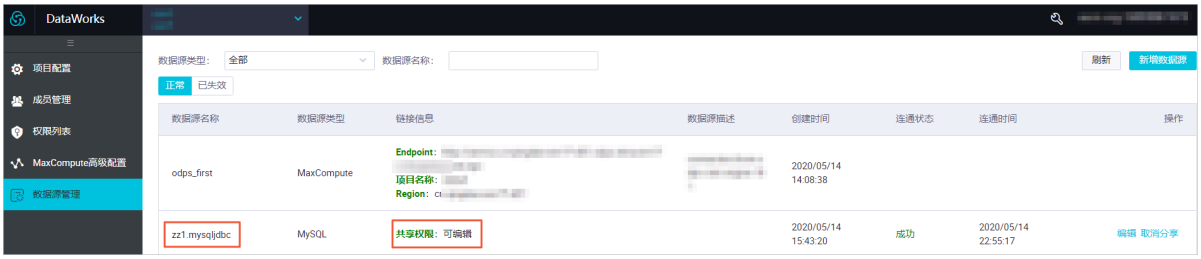
查看分享的数据源

单击顶部的工作空间名称区域，切换至上文已添加权限的工作空间。进入该工作空间的**数据源**页面，即可查看**正常**和**已失效**的被分享的数据源：

- **正常**

  您可以在该模块查看被分享的数据源的**数据源名称**、**数据源类型**、**链接信息**、**数据源描述**、**创建时间**、**连通状态**和**连通时间**等信息。

被分享的数据源的**链接信息**下，会显示共享权限。被分享的数据源命名格式为源分享工作空间名称.数据源名称。



如果被分享的数据源在当前工作空间下有编辑权限，则会在该数据源的操作栏下显示**编辑**。

- **已失效**

  单击**已失效**，即可查看以前分享过，但目前已经失效的数据源。

  您可以查看**取消方**和**取消共享时间**，以便查找数据源不能使用的原因。

2.1.4. 数据源开发和生产环境隔离

数据源隔离模式可以满足标准模式下，开发环境和生产环境的数据隔离需求。

背景信息

同一个名称的数据源存在开发环境和生产环境两套配置，您可以通过数据源隔离使其在不同环境隔离使用。

> **说明** 目前只有标准模式的工作空间支持数据源隔离。

配置数据同步任务时会使用开发环境的数据源，提交生产运行时会使用生产环境的数据源。如果您要将任务提交到生产环境调度，同一个数据源名需要同时添加生产环境和开发环境的数据源配置。离线数据源的详情请参见支持的数据源与读写插件。

新增数据源隔离模式后，对工作空间有以下影响：

- 简单模式：数据源功能和界面与之前保持一致。
- 标准模式：数据源界面按照数据源隔离模式进行相应调整，增加了适用环境的参数。
- 简单模式升级成标准模式：进行模式升级时，会提示对数据源进行升级，将数据源拆分成生产环境和开发环境隔离的模式。

操作步骤

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
4. 在左侧导航栏，单击**数据源**，进入**工作空间管理 > 数据源管理**页面。

页面功能	描述
批量新增数据源	目前仅支持MySQL、SQLServer和Oracle数据源。 您可以下载相应数据源的模板，根据模板中的格式填写内容，选择上传文件进行新建操作，文本框中会显示添加详情。模板内容：显示数据源类型、数据源名称、数据源描述、环境类别（0开发、1生产）、链接地址。
新增数据源	○ 开发环境可用的数据源：可以在新建数据同步节点时选择并在开发环境运行，但无法提交到生产环境或在生产环境运行。 ○ 生产环境可用的数据源：只允许在生产环境运行时使用，不可以在新建数据同步节点时选择。 **说明** 同一个开发环境和生产环境的数据源名称必须一致。
适用环境	简单模式下的工作空间不显示该配置。
操作	○ **新建**：如果不存在适用环境下的数据源，显示**新建** ○ **编辑**和**删除**：如果存在适用环境下的数据源，则显示**编辑**和**删除**按钮。 ■ 删除开发环境和生产环境的数据源：需确认是否存在生产环境关联的同步任务，操作不可逆，删除后，在开发环境配置同步任务时此数据源不可见。 如果生产环境在使用此数据源配置的同步任务，删除后，生产环境任务不可正常运行。请删除同步任务后再删除此数据源。 ■ 删除开发环境的数据源：需确认是否存在生产环境关联的同步任务，操作不可逆，删除后，在开发环境配置同步任务时此数据源不可见。 如果生产环境在使用此数据源配置的同步任务，删除后，任务编辑时将不能获取到元数据信息，但生产环境任务可以正常运行。 ■ 删除生产环境的数据源：需确认是否存在生产环境关联的同步任务，删除后，在开发环境使用此数据源配置的同步任务将不能提交生产发布。 如果生产环境在使用此数据源配置的同步任务，删除后，生产环境任务不可正常运行。
选择	勾选后，可以进行批量测试连通性和批量删除操作。

2.2. 资源组与网络连通

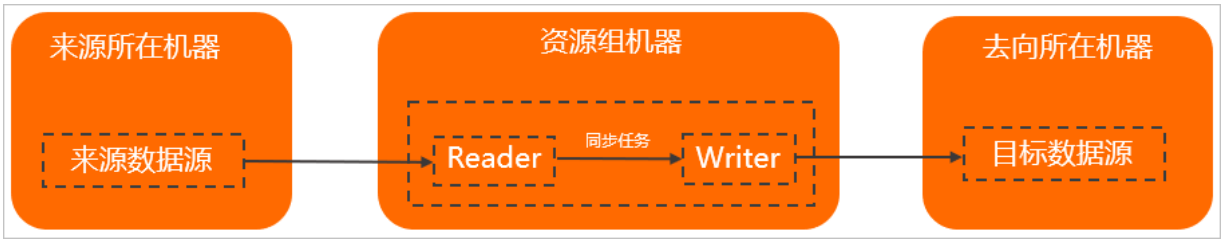
2.2.1. 资源规划与配置

使用DataWorks的数据集成功能进行数据同步时，数据同步任务会运行占用一定的资源组，本文为您介绍资源组的基本概念和分类，以及连通性和性能问题，通过对比各类资源组，助力您根据自身需求选择更合适的资源组类型。

基本概念

资源组是指数据集成的批数据同步任务运行所在的计算资源。通常资源组以机器的形式展现，即CPU、内存和带宽的大小。

执行数据同步任务的流程为：先从数据来源所在的机器抽取数据至资源组所在的机器，再推送至目标数据源所在的机器。



资源组分类

资源组大体可分为独享数据集成资源组、自定义数据集成资源组两类。

- 独享数据集成资源组：

  购买后可独占使用的资源组。在任务高并发执行且无法错峰运行，需要独享的资源组来保障数据快速、稳定地传输时，您可以选择独享资源组。

  更多独享资源组的介绍可参见独享数据集成资源组概述，独享资源组的使用可参见新增和使用独享数据集成资源组。

- 自定义数据集成资源组：

  如果您有富余的服务器资源，也可将此部分资源作为DataWorks中任务运行的资源组使用，DataWorks支持自定义资源组。

  更多自定义资源组的介绍可参见自定义资源组概述，自定义资源组的使用可参见新增和使用自定义数据集成资源组。

资源规划关键：连通性和性能

您在使用资源组时，需要关注资源组的连通性和性能两方面：

- 连通性

  由于执行数据同步任务流程的要求，需要保证资源组可以访问数据源（来源数据源和目标数据源）所在的网络，且不会因白名单限制等原因不能访问数据源。您需要在保证网络连通的前提下，再使用数据集成，如果网络不通，会导致数据同步任务无法执行。

  连通性问题是资源组最重要的问题，您在选用资源组时，需结合数据源网络环境和各资源组的网络连通解决方案，选择合适的资源组。各资源组支持的网络连通解决方案可参见配置资源组与网络连通。

- 性能

  执行同步任务时，会占用资源组所在机器的CPU、内存和网络等资源。如果资源不足，会导致任务无法启动、启动后长时间等待资源，或启动后传输速率较慢，不能及时产出结果等问题。

  您需要给予同步任务充足的资源，以保障任务的顺利运行。建议您使用独享资源组，将任务运行在独立的环境中，无需受到公共资源池的影响。独享资源组的性能指标可参见独享数据集成资源组计费说明：包年包月。

资源组对比与建议

两种类型的资源组适用于不同的场景，以下通过在资源组归属、网络连通、收费方式等维度为您对比展示各资源组的现状，您可以在执行任务时，根据自身需求选择合适的资源组。

类别	独享数据集成资源组	自定义数据集成资源组
机器资源归属	由DataWorks维护，是自己的租户独享使用的计算资源。	由您自己维护，是属于您的IDC机器。
网络	支持VPC、公网和任意网络下的阿里云产品。	支持VPC、公网和任意网络下的阿里云产品。
收费方式	根据机器的规格，包年包月计费。	DataWorks版本按月使用收费。
支持的数据源	全部数据源	全部数据源
安全性	高	根据您自身机器所处的环境决定
任务执行的效率 指任务是否能够分到足够的计算资源，能否以最高性能运行。	高	根据您自身机器所处的环境决定
可靠性 指任务是否能够按时启动。执行任务时，网络资源是否被其它租户占用，导致任务不能按时产出结果。	高	根据您自身机器所处的环境决定
适用场景	大量、重要的生产级别的任务。	使用自定义资源组的场景如下： • 如果您自身已有计算资源，可以对接阿里云重复使用，无需重新购买。 • 需要同步的数据源全部在IDC内。

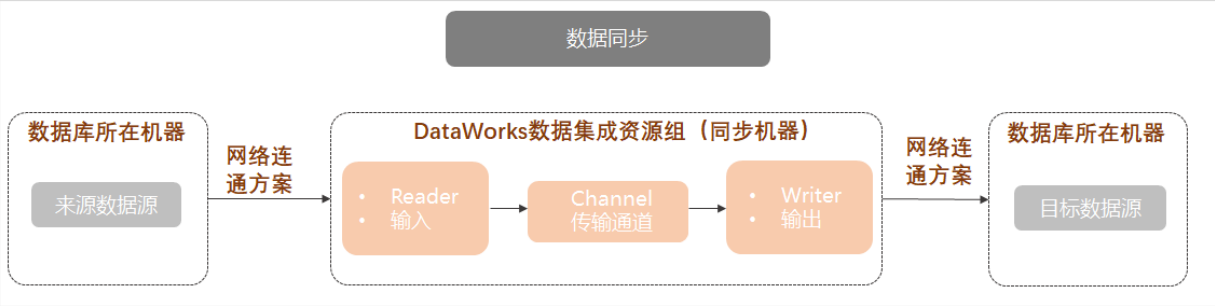
类别	独享数据集成资源组	自定义数据集成资源组
推荐指数	★★★★★	★

根据上表两种类型的对比，推荐您使用独享数据集成资源组来执行同步任务。

2.2.2. 配置资源组与网络连通

在数据同步任务配置前，您需要确保用于执行同步任务的独享数据集成资源组与您将要同步的数据来源端与目的端数据库的网络连通性，您可以根据数据库所在网络环境，选择合适的网络解决方案来实现网络连通。本文为您介绍数据库在不同网络环境中时，可选用的网络连通方案。

背景信息



如上图所示，进行数据同步前，需通过合适的网络连通方案将数据库和资源组间的网络打通。本文重点关注通过独享数据集成资源组访问数据库的场景。

> 说明 如果数据库存在白名单限制，您还需要将独享数据集成资源组的EIP添加至数据库白名单。获取EIP详情请参见：添加白名单。

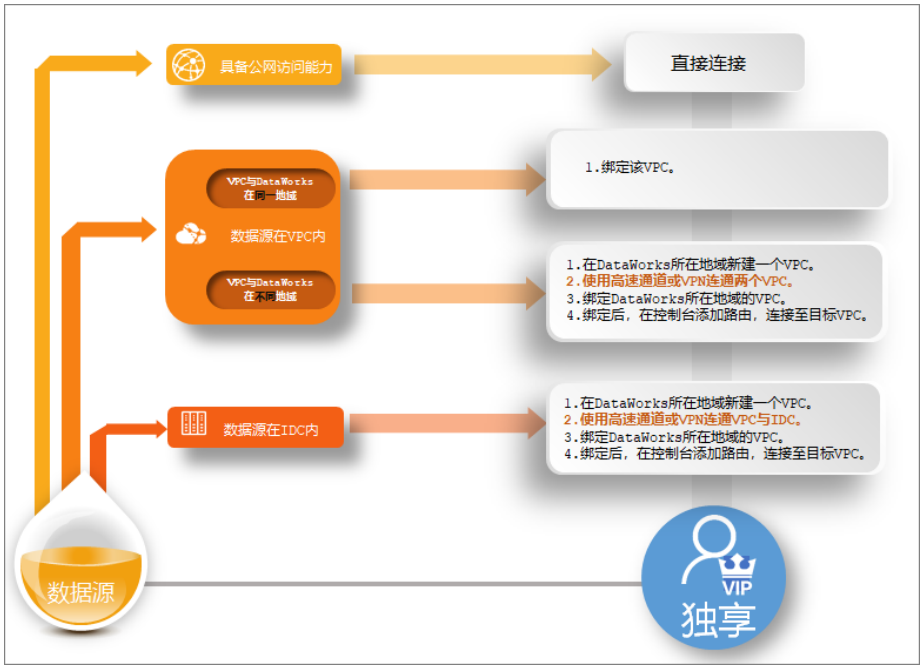
购买合适规格的资源组

购买并选择合适规格的独享数据集成资源组，购买详情请参见：新增和使用独享数据集成资源组。

> 说明
> - 不同规格的资源组支持的实时同步任务数存在上限，您需要根据业务需要选择合适规格的资源组。
> - 离线和实时同步任务推荐使用不同的资源组，以便任务分开执行。如果选择同一个资源组，任务混跑会带来资源抢占、运行态互相影响等问题。例如，CPU、内存、网络等互相影响，可能会导致离线任务变慢或实时任务延迟等问题，甚至在资源不足的极端情况下，可能会出现任务被OOM KILLER杀掉等问题。

配置网络连通

网络连通方案选择取决于数据库与DataWorks工作空间（独享资源组，即同步机器）间的环境关系，包括以下场景：



同步网络选择	数据源所在环境	数据源与DataWorks工作空间关系	网络连通通用逻辑	配置示例
通过**VPC（内网）**同步	数据库在阿里云 • ECS自建数据库 • 阿里云数据库云产品	同阿里云主账号、同地域 	独享数据集成资源组绑定数据源所在VPC即可。	场景一：数据库与DataWorks工作空间同阿里云主账号、同地域
		不在同一个同阿里云主账号下或不在同一个地域下 	1. 先通过网络连通工具（云企业网、高速通道、VPN网关）连通数据库所在地域与DataWorks工作空间所在地域的网络环境。 2. 为独享数据集成资源组绑定当前阿里云主账号下已经与数据库网络连通的专有网络。 3. 为独享数据集成资源组添加一条自定义路由并指向目标数据库IP地址。详情可参见添加路由。	• 场景二：数据库与DataWorks工作空间同阿里云主账号、不同的地域 • 场景三：数据库与DataWorks工作空间使用不同阿里云主账号
	数据库不在阿里云 • IDC数据库 • 非阿里云的云数据库	不涉及		场景四：数据库在IDC
通过**公网**访问数据库	-		独享数据集成资源组有公网访问能力，可直接连通。	-

同步网络选择	数据源所在环境	数据源与DataWorks工作空间关系	网络连通通用逻辑	配置示例

> **说明**
>
> - 无论上述哪种场景，如果数据库存在白名单访问控制，您都需要将资源组绑定的交换机网段添加至数据库白名单中。获取数据库需要添加的白名单详情请参见：添加白名单。
> - 独享数据集成资源组暂不支持经典网络环境下的数据库同步，建议您将经典网络环境下的数据库迁移至VPC环境。
> - 无法保障公网环境下的数据同步速度，建议通过内网同步。

各场景网络连通配置示例

下文以使用阿里云RDS数据库为例，网络连通配置如下。获取RDS的专有网络信息，详情请参见切换专有网络VPC和虚拟交换机。

- **通过VPC（内网）访问数据库**

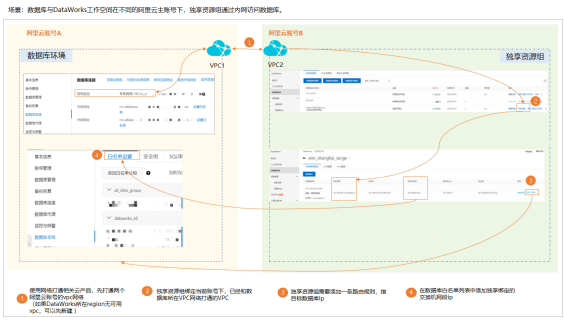
  场景一：数据库与DataWorks工作空间同阿里云主账号、同地域

网络连通配置操作	配置操作图示
i. 网络绑定：独享数据集成资源组可绑定数据源所在VPC。 ii. 数据库白名单配置：将独享资源组绑定的交换机网段IP添加到数据库白名单中。	

  场景二：数据库与DataWorks工作空间同阿里云主账号、不同的地域

网络连通配置操作	配置操作图示
i. 配置两个地域间的网络连通。 使用高速通道、VPN网关连通两个地域的VPC。 ii. 配置数据源和资源组的网络连通。 a. 为独享数据集成资源组绑定当前账号下，已与数据库网络连通的VPC。 b. 在控制台添加路由，连接至目标VPC，详情可参见添加路由。	

  场景三：数据库与DataWorks工作空间使用不同阿里云主账号

网络连通配置操作	配置操作图示
i. 配置两个阿里云主账号间的网络连通。 使用高速通道、VPN网关或云企业网连通两个阿里云主账号的VPC。 ii. 配置数据源和资源组的网络连通。 a. 为独享数据集成资源组绑定当前账号下已与数据库网络连通的VPC。 b. 在控制台添加路由，连接至目标VPC，详情可参见添加路由。	

- 数据库不在阿里云

  场景四：数据库在IDC

  i. 配置两个环境的网络连通。

  使用高速通道、VPN网关连通IDC与阿里云专有网络。

  ii. 配置数据源和资源组的网络连通。

  a. 为独享数据集成资源组绑定当前账号下已与数据库网络连通的VPC。

  b. 在控制台添加路由，连接至目标VPC，详情可参见添加路由。

后续步骤

1. 网络连通配置。
   i. 完成网络连通配置后，您可根据数据库是否开启白名单设置，如果开启了白名单，则您需要将资源组相应的IP地址段添加至数据库的白名单中，避免资源组无法正常读写数据库的数据，详情可参见添加白名单。
   ii. 如果您的数据库是ECS自建数据库，您还需安全组配置，详情可参见场景示例：ECS自建数据库的安全组配置。
2. 数据同步任务配置，操作详情请参见以下文档。
   - 数据同步解决方案。
   - 离线数据同步。
   - 实时数据同步。

2.2.3. 添加白名单

保障资源组与数据源之间网络连通后，您还需保障资源组与数据源之间不会因为白名单的限制而无法进行数据访问，例如，部分数据源设置白名单后会不允许白名单外的IP访问，您需要将资源组的IP添加至数据源的白名单中。本文为您介绍白名单相关问题。

前提条件

您需保障数据源与数据集成资源组间的网络时连通状态：

- 如果您目前需要连通的数据库网络较复杂（不是同步同阿里云账号下同Region的数据库），您需要根据数据集成资源组网络能力来选择合适的方案访问您特殊场景下的数据库，详情可参见配置资源组与网络连通。
- 如果您用独享数据集成资源组同步同阿里云账号，同Region下的VPC环境数据库，您购买添加独享数据集成资源组后，还需完成网络配置和工作空间绑定，操作详情可参见新增和使用独享数据集成资源组。

如果您已完成网络连通配置，但资源组仍然无法访问您的数据库，则数据库可能开启了白名单配置，限制了数据集成资源组的访问。此时您需要将资源组的相应IP地址段添加至数据库白名单中。

背景信息

如果数据集成资源组和您需要访问的数据源网络已连通，如配置资源组与网络连通中所述，但资源组仍然无法访问您的数据库，则可能是因为您的数据库有白名单限制，您需要获取资源组的IP地址与网段，添加至数据库的白名单中。

为保证数据源的数据库的安全稳定，大部分数据源开启了白名单的限制（例如，RDS、MongoDB和Redis等常见的数据源），此种场景下，您需要将DataWorks资源组的IP地址或网段添加至数据源的白名单中，对数据集成资源组的访问IP放行。添加白名单时：

- 使用独享数据集成资源组：

  您需要添加资源组的EIP地址及网段、绑定的交换机网段至数据库的白名单内。您需获取相应的IP地址及网段后添加至数据库的白名单中，详情可参见使用独享数据集成资源组执行任务需要在数据库添加的IP白名单和云产品白名单配置注意事项。

- 使用公共数据集成（调试）资源组：

  您需要数据库给底层运行机器授予访问权限，在数据库添加DataWorks工作空间所在区域的白名单，详情可参见使用公共数据集成（调试）资源组执行任务时需要在数据库添加的IP白名单和云产品白名单配置注意事项。

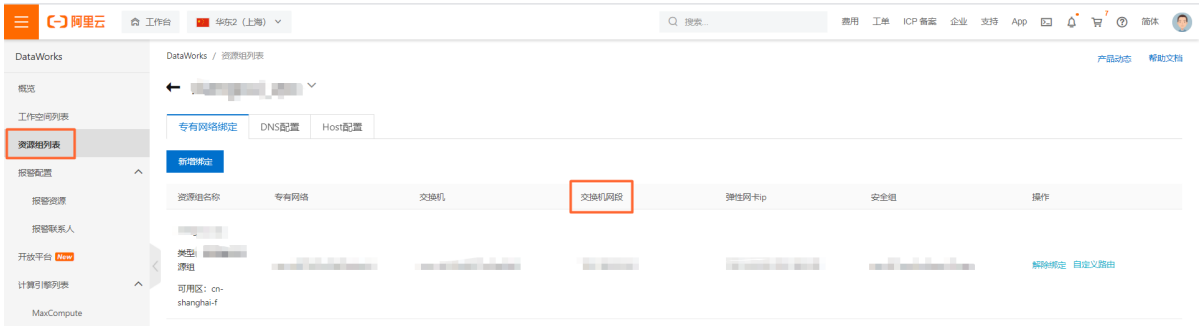
- 使用自定义数据集成资源组：

  您需要数据库给自定资源组机器授权，添加自定义机器内网IP或外网IP至数据源的白名单列表，详情可参见使用自定义数据集成资源组执行任务时需要在数据库添加的IP白名单和云产品白名单配置注意事项。

使用独享数据集成资源组执行任务需要在数据库添加的IP白名单

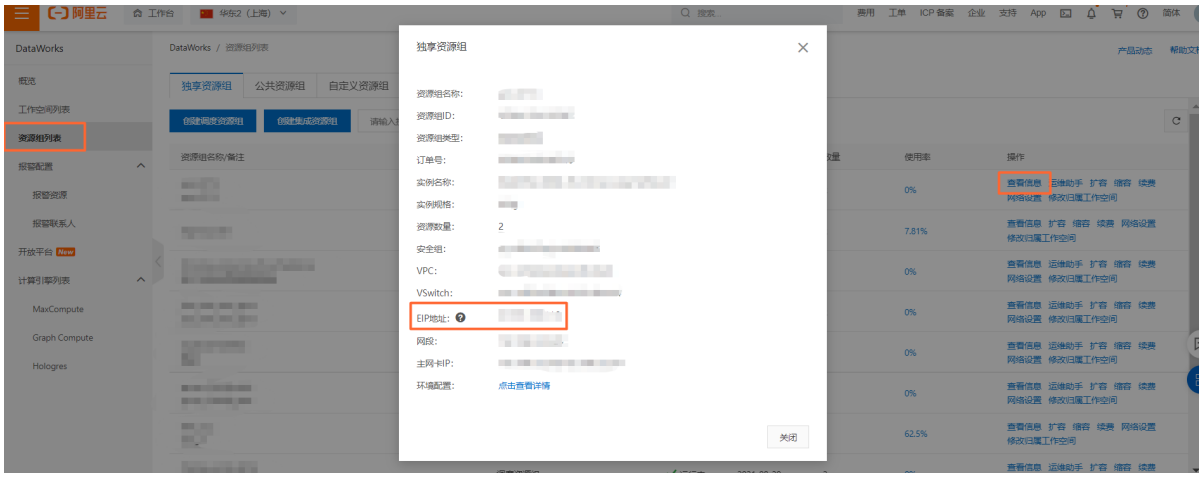
- 如果您使用独享数据集成资源组走VPC内网同步数据，请在数据库白名单列表中添加独享数据集成资源组绑定的交换机网段。获取独享绑定的交换机网段信息如下：

  在DataWorks控制台**独享资源组**页签下，单击目标独享数据集成资源组后的**网络设置**，查看**交换机网段**并将其添加至数据库白名单列表中。



- 如果您使用独享数据集成资源组走公网同步数据，请在数据库白名单列表中添加独享数据集成资源组本身的EIP地址。获取独享数据集成资源组EIP地址如下：

  在DataWorks控制台**独享资源组**页签下，单击数据集成资源组后的**查看信息**，复制对话框中的**EIP地址**至数据库白名单列表中。



> **说明** 如果您之后对独享数据集成资源组进行了扩容操作，请检查此处待添加的EIP是否有变动，为了避免任务执行出错，请在扩容后第一时间更新数据库添加的白名单。

使用公共数据集成（调试）资源组执行任务时需要在数据库添加的IP白名单

使用公共数据集成（调试）资源组时，您需要根据DataWorks所在的地域，添加不同的IP地址段至数据库的白名单中。查看不同地域的IP地址段信息的操作步骤如下：

1. 以开发者身份登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 单击左上角的区域，切换相应的工作空间区域。
4. 根据工作空间所在的区域查看相应的资源组IP地址信息，后续需将这些IP地址段添加至数据库的白名单中。

地域	白名单
华东1（杭州）	11.193.215.0/24,11.194.110.0/24,11.194.73.0/24,11.196.23.0/24,11.197.247.0/24,11.193.102.0/24,100.104.0.0/16,118.31.157.0/24,47.97.53.0/24,47.99.12.0/24,47.99.13.0/24,114.55.197.0/24,11.197.246.0/24
华东2（上海）	10.152.69.0/24,10.153.136.0/24,11.115.106.0/24,11.192.97.0/24,11.192.98.0/24,11.193.102.0/24,11.193.109.0/24,11.193.252.0/24,11.218.89.0/24,11.218.96.0/24,11.219.217.0/24,11.219.218.0/24,11.219.219.0/24,11.219.233.0/24,11.219.234.0/24,10.117.28.203,10.117.39.238,10.27.63.15,10.27.63.38,10.27.63.41,10.27.63.60,10.46.64.81,10.46.67.156,100.104.0.0/16,118.178.142.154,118.178.56.228,118.178.59.233,118.178.84.74,120.27.160.26,120.27.160.81,121.43.110.160,121.43.112.137,47.100.129.0/24,47.101.107.0/24,47.102.181.128/26,47.102.181.192/26,47.102.234.0/26,47.102.234.64/26,106.15.14.0/24,10.143.32.0/22
华南1（深圳）	100.106.46.0/24,100.106.49.0/24,10.152.27.0/24,10.152.28.0/24,11.192.96.0/24,11.193.103.0/24,11.193.104.0/24,11.196.76.0/24,11.192.91.0/24,100.104.0.0/16,120.76.104.0/24,120.76.91.0/24,120.78.45.0/24,47.106.63.0/26,47.106.63.128/26,47.106.63.192/26,47.106.63.64/26,120.77.195.128/26,120.77.195.192/26,120.77.195.64/26,47.112.86.0/26
西南1（成都）	11.195.52.0/24,11.195.55.0/24,47.108.46.0/26,47.108.46.128/26,47.108.46.192/26,47.108.46.64/26,47.108.22.0/24,100.104.0.0/16
华北3（张家口）	11.193.235.0/24,100.104.0.0/16,47.92.185.0/26,47.92.185.64/26,47.92.185.128/26,47.92.185.192/26,47.92.22.0/24
中国（香港）	10.152.162.0/24,11.192.196.0/24,11.193.11.0/24,11.193.118.0/24,100.104.0.0/16,47.75.228.0/24,47.89.61.0/24,47.244.92.128/26,47.244.92.192/26,47.56.45.0/26,47.56.45.64/26,47.91.171.0/25,47.101.109.0/26,47.56.45.128/26,47.56.45.192/26,47.90.24.0/26,47.90.24.64/26
亚太东南1（新加坡）	11.193.162.0/24,11.193.163.0/24,11.193.8.0/24,11.197.188.0/24,11.193.158.0/24,11.193.220.0/24,11.192.152.0/23,11.192.40.0/26,10.151.234.0/26,10.151.238.0/26,10.152.248.0/26,100.106.10.0/26,100.106.35.0/26,100.104.0.0/16,47.74.161.0/24,47.74.162.0/24,47.88.235.0/25,47.88.147.0/24,47.74.203.0/24,161.117.146.128/26,161.117.146.192/26,161.117.164.0/26,161.117.164.64/26,47.74.206.0/26,47.74.206.128/26,47.74.206.192/26,47.74.206.64/26
亚太东南2（悉尼）	11.192.100.0/24,11.192.134.0/24,11.192.135.0/24,11.192.184.0/24,11.192.99.0/24,11.193.165.0/24,100.104.0.0/16,47.91.60.0/24,47.91.50.0/25,47.91.49.128/25,47.91.49.0/25
华北2（北京）	11.193.75.0/24,100.106.48.0/24,11.193.82.0/24,11.193.99.0/24,11.197.231.0/24,10.152.167.0/24,10.152.168.0/24,11.193.50.0/24,11.195.172.0/22,100.104.0.0/16,47.93.110.0/24,47.94.185.0/24,47.95.63.0/24,47.94.49.0/24,182.92.144.0/24,182.92.32.128/26,39.107.7.0/26
美国西部1（硅谷）	10.152.160.0/24,11.193.216.0/24,100.104.0.0/16,47.89.224.0/24,47.88.108.0/24,47.89.124.0/26,47.89.124.128/26,47.89.124.192/26,47.89.124.64/26
美国东部1（弗吉尼亚）	47.88.98.0/26,47.88.98.64/26,47.88.98.128/26,47.88.98.192/26,47.252.91.0/26,47.252.91.128/26,47.252.91.192/26,47.252.91.64/26,47.252.71.128/26,47.252.71.192/26,47.252.90.0/26,47.252.90.64/26,10.128.134.0/24,11.193.203.0/24,11.194.68.0/24,11.194.69.0/24,100.104.0.0/16
亚太东南3（吉隆坡）	11.193.188.0/24,11.193.189.0/24,11.214.81.0/24,11.221.206.0/24,11.221.205.0/24,11.221.207.0/24,100.104.0.0/16,47.254.212.0/24,47.250.29.0/26,47.250.29.128/26,47.250.29.192/26,47.250.29.64/26

地域	白名单
欧洲中部1（法兰克福）	11.192.116.0/24,11.192.170.0/24,11.193.167.0/24,11.192.169.0/24,11.193.106.0/24,11.192.168.0/24,100.104.0.0/16,47.91.82.0/24,47.91.83.0/24,47.91.84.0/24,47.254.138.0/24,47.254.180.0/26,47.254.180.128/26,47.254.180.192/26,47.254.180.64/26
亚太东北1（日本）	100.105.55.0/24,11.192.147.0/24,11.192.149.0/24,11.199.250.0/24,11.59.59.0/24,11.192.148.0/24,100.104.0.0/16,47.91.0.128/26,47.91.0.192/26,47.91.27.128/26,47.91.12.0/24,47.91.13.0/24,47.91.9.0/24,47.91.27.0/26,47.245.18.128/26,47.245.18.192/26,47.245.51.0/26,47.245.51.64/26,47.245.51.128/26,47.245.51.192/26
中东东部1（迪拜）	11.192.107.0/24,11.192.127.0/24,11.192.88.0/24,11.193.246.0/24,100.104.0.0/16,47.91.116.0/24
亚太南部1（孟买）	11.194.10.0/24,11.246.70.0/24,11.246.71.0/24,11.246.73.0/24,11.246.74.0/24,11.59.62.0/24,11.194.11.0/24,100.104.0.0/16,149.129.164.0/24,149.129.165.192/26,147.139.23.0/26,147.139.23.128/26,147.139.23.64/26,147.139.21.0/26,147.139.21.128/26,147.139.21.192/26,147.139.21.64/26
英国（伦敦）	11.199.93.0/24,100.104.0.0/16,8.208.17.0/24,8.208.72.0/26,8.208.72.128/26,8.208.72.192/26,8.208.72.64/26
亚太东南5（雅加达）	11.194.49.0/24,11.194.50.0/24,11.200.93.0/24,11.200.97.0/24,11.59.135.0/24,11.200.95.0/26,10.143.32.0/22,100.104.0.0/16,149.129.228.0/24,47.89.94.128/27,47.89.94.160/27,47.89.94.192/27,47.89.94.224/27,47.89.95.128/26,149.129.229.0/26,149.129.229.128/26,149.129.229.192/26,149.129.229.64/26,147.139.156.0/26,147.139.156.128/26,147.139.156.64/26,149.129.230.192/26
华北2（政务云）	11.194.116.0/24,100.104.0.0/16,39.107.188.0/24 如果IP地址段添加不成功，请添加下述IP地址： 11.194.116.160,11.194.116.161,11.194.116.162,11.194.116.163,11.194.116.164,11.194.116.165,11.194.116.167,11.194.116.169,11.194.116.170,11.194.116.171,11.194.116.172,11.194.116.173,11.194.116.174,11.194.116.175,39.107.188.0/24,100.104.0.0/16
华东2（上海）金融云	140.205.46.128/25,140.205.48.0/25,140.205.48.128/25,140.205.49.0/25,140.205.49.128/25,11.192.156.0/25,11.192.157.0/25,11.192.164.0/25,11.192.165.0/25,11.192.166.0/25,11.192.167.0/25,106.11.245.0/26,106.11.245.128/26,106.11.245.192/26,106.11.245.64/26,140.205.39.0/24,106.11.225.0/24,106.11.226.0/24,106.11.227.0/24,106.11.242.0/24,100.104.0.0/16

使用自定义数据集成资源组执行任务时需要在数据库添加的IP白名单

使用自定义数据集成资源组执行数据同步任务时，请添加自定义数据集成资源组的机器IP至数据库的白名单内。

> **说明** 自定义数据集成资源组扩容后，为避免任务由于白名单问题导致报错，请第一时间将扩容机器IP添加到数据库白名单列表中。

云产品白名单配置注意事项

以阿里云云数据库RDS为例，您需要添加数据集成资源组相应的IP地址段到数据库白名单列表中，在配置白名单前您可以先了解以下问题。

目前云产品支持通用模式IP白名单和高安全模式IP白名单配置，您添加白名单时配置的白名单分组可能会影响数据集成资源组与数据库的网络连通：

- 如果您目前数据库设置的为通用模式IP白名单：
  - 通用模式IP白名单不区分经典网络和专有网络白名单分组。
  - 不同类型的资源组需要添加的白名单，建议放在不同的白名单分组中，以免混淆。

    > **说明** 在通用白名单模式下，设置的IP地址，既可通过经典网络，也可通过专有网络访问RDS实例。

- 如果您目前数据库设置的为高安全模式IP白名单模式：

- 高安全模式区分经典网络和专有网络白名单分组。

  > **说明** 在高安全白名单模式下，白名单分组需指定网络隔离模式，例如设置在经典网络的白名单IP地址，不可从专有网络访问RDS实例，反之亦然。

- 使用独享资源组VPC内网直接连接数据库，使用专有网络白名单分组。
- 使用公共数据集成（调试）访问 VPC网络数据源（例如，实例模式配置的专有网络 RDS MySQL），使用专有网络白名单分组。
- 使用公网连接地址直接访问数据库，走的是经典网络白名单分组。

- 如果您在数据库将白名单模式从通用模式IP白名单模式切换为高安全模式IP白名单模式：

  RDS会将通用模式IP白名单复制分为2份，分别放到经典网络和专有网络白名单分组类型里面。

其他白名单配置注意事项：

- 设置白名单不会影响RDS实例的正常运行。
- 默认的IP白名单分组（default ）不能删除，只能清空。
- 请勿修改或删除系统自动生成的分组，避免影响相关产品的使用。例如**ali_dms_group**（DMS产品IP地址白名单分组）、**hdm_security_ips**（DAS产品IP地址白名单分组）。

  > **说明** 建议您在数据库配置白名单时，单独为DataWorks白名单新建一个白名单分组。

- 默认的IP白名单只包含127.0.0.1，表示任何IP均无法访问该RDS实例。

RDS白名单配置详情可参见通过客户端、命令行连接RDS MySQL实例。其他类型的数据源类似，可参考各数据源数据库的白名单配置步骤，分别添加对应的白名单。

后续步骤

如果您的数据库是ECS自建数据库，那么完成白名单配置后，您还需要配置自建数据库的安全组，才能保障资源组能正常读写数据库的数据，详情可参见场景示例：ECS自建数据库的安全组配置。

2.2.4. 场景示例：ECS自建数据库的安全组配置

如果您使用的是ECS自建数据库，必须添加安全组才能保证资源组可正常读写数据源的数据，本文为您介绍选择不同区域的DataWorks时，如何配置ECS自建数据库的安全组。

前提条件

1. 已完成资源组与数据源之间的网络连通配置，详情可参见配置资源组与网络连通。
2. 数据库开启白名单时，已将资源组对应的IP地址段添加至数据库白名单中，详情可参见添加白名单。

添加安全组

进行ECS自建数据源的数据同步时，如果同步任务运行在：

- 独享数据集成资源组上

  需要将独享数据集成资源组的EIP地址段或独享数据集成资源组绑定的交换机网段添加至ECS安全组上。

  - 通过外网同步数据时，需获取并添加独享资源组本身的EIP网段至ECS安全组上。
  - 通过VPC内网同步数据时，需获取并添加独享数据集成资源组绑定的交换机网段至ECS安全组上。

  独享数据集成资源组的相关信息获取可参见使用独享数据集成资源组执行任务需要在数据库添加的IP白名单，添加安全组规则的操作可参见添加安全组规则。

3.离线数据同步

3.1. 支持的数据源与读写插件

数据集成是稳定高效、弹性伸缩的数据同步平台，为阿里云大数据计算引擎（MaxCompute、AnalyticDB for PostgreSQL和Hologres等）提供离线、批量数据的进出通道。

注意

- 在网络可达的情况下，支持跨账号同步数据。即A账号下的MySQL数据库的数据可以同步至B账号的MongoDB等数据库中。
- 部分数据源仅支持脚本模式，详情请参见具体的Reader和Writer文档。
- 目前不支持导入Excel格式的数据，您可以修改Excel的格式为CSV格式，再进行导入。
- 数据源需要在连通网络的前提下进行数据同步，详情请参见配置资源组与网络连通。

数据源类型	抽取（Reader）	导入（Writer）
AWS S3	AWS S3 Reader	不支持
AnalyticDB for MySQL 2.0	AnalyticDB for MySQL 2.0 Reader	AnalyticDB for MySQL 2.0 Writer
AnalyticDB for MySQL 3.0	AnalyticDB for MySQL 3.0 Reader	AnalyticDB for MySQL 3.0 Writer
AnalyticDB for PostgreSQL	AnalyticDB for PostgreSQL Reader	AnalyticDB for PostgreSQL Writer
ApsaraDB For Oceanbase 说明 目前该数据源仅支持使用独享数据集成资源组。	ApsaraDB For Oceanbase Reader	ApsaraDB For Oceanbase Writer
ClickHouse 说明 目前该数据源仅支持使用独享数据集成资源组。	ClickHouse Reader	ClickHouse Writer
DataHub	DataHub Reader	DataHub Writer
DB2 说明 目前该数据源仅支持使用独享数据集成资源组。	DB2 Reader	DB2 Writer
DM（达梦） 说明 目前该数据源仅支持使用独享数据集成资源组。	DM Reader	DM Writer
DRDS	DRDS Reader	DRDS Writer
Elasticsearch	Elasticsearch Reader	Elasticsearch Writer
FTP	FTP Reader	FTP Writer

数据源类型	抽取（Reader）	导入（Writer）
GBase8a 说明 目前该数据源仅支持使用独享数据集成资源组。	支持	支持
HBase 说明 目前该数据源仅支持使用独享数据集成资源组。	• HBase Reader • HBase20xsql Reader	• HBase Writer • HBase 11xsql Writer
HDFS 说明 目前该数据源仅支持使用独享数据集成资源组。	HDFS Reader	HDFS Writer
Hive 说明 目前该数据源仅支持使用独享数据集成资源组。	Hive Reader	Hive Writer
Hologres 说明 目前该数据源仅支持使用独享数据集成资源组。	支持	支持
HybridDB for MySQL	HybridDB for MySQL Reader	HybridDB for MySQL Writer
Kafka 说明 目前该数据源仅支持使用独享数据集成资源组。	Kafka Reader	Kafka Writer
KingbaseES（人大金仓） 说明 目前该数据源仅支持使用独享数据集成资源组。	KingbaseES Reader	KingbaseES Writer
Lindorm 说明 目前该数据源仅支持使用独享数据集成资源组。	Lindorm Reader	Lindorm Writer
LogHub（SLS）	LogHub（SLS） Reader	LogHub（SLS） Writer
MaxCompute	MaxCompute Reader	MaxCompute Writer
MaxGraph	不支持	Maxgraph Writer

数据源类型	抽取（Reader）	导入（Writer）
Memcache	不支持	Memcache Writer
MetaQ 说明 目前该数据源仅支持使用独享数据集成资源组。	MetaQ Reader	不支持
MongoDB 说明 目前该数据源仅支持使用独享数据集成资源组。	MongoDB Reader	MongoDB Writer
MySQL 说明 目前MySQL 8.0仅支持新增和使用独享数据集成资源组。	MySQL Reader	MySQL Writer
OpenSearch 说明 目前该数据源仅支持使用独享数据集成资源组。	不支持	OpenSearch Writer
Oracle	Oracle Reader	Oracle Writer
OSS	OSS Reader	OSS Writer
OTSStream	OTSStream Reader	不支持
PolarDB	PolarDB Reader	PolarDB Writer
PostgreSQL	PostgreSQL Reader	PostgreSQL Writer
Prometheus 说明 目前该数据源仅支持使用独享数据集成资源组。	Prometheus Reader	不支持
Redis	不支持	Redis Writer
RestAPI（HTTP形式） 说明 目前该数据源仅支持使用独享数据集成资源组。	RestAPI Reader	RestAPI Writer
SAP HANA 说明 目前该数据源仅支持使用独享数据集成资源组。	SAP HANA Reader	SAP HANA WRITER

数据源类型	抽取（Reader）	导入（Writer）
StarRocks 说明 目前该数据源仅支持使用独享数据集成资源组。	StarRocks Reader	StarRocks Writer
Stream	Stream Reader	Stream Writer
SQL Server	SQL Server Reader	SQL Server Writer
Table Store（OTS）	Table Store（OTS）Reader	Table Store（OTS）Writer
TSDB 说明 目前该数据源仅支持使用独享数据集成资源组。	不支持	TSDB Writer
Vertica 说明 目前该数据源仅支持使用独享数据集成资源组。	Vertica Reader	Vertica Writer

3.2. 离线同步任务配置

3.2.1. 通过向导模式配置离线同步任务

本文为您介绍如何通过数据集成向导模式配置任务。

前提条件

1. 已完成数据源配置。您需要在数据集成同步任务配置前，配置好您需要同步的源端和目标端数据库，以便在同步任务配置过程中，可通过选择数据源名称来控制同步任务的读取和写入数据库。离线同步支持的数据源及其配置详情请参见支持的数据源与读写插件。

   说明 数据源相关能力介绍详情请参见：数据源概述。

2. 已购买合适规格的独享数据集成资源组。详情请参见：新增和使用独享数据集成资源组。
3. 独享数据集成资源组与数据源网络已打通。详情请参见：配置资源组与网络连通。

开发流程

1. 新建离线同步节点
2. 配置离线同步任务
   i. 配置同步网络链路
   ii. 选择要同步的具体对象
   iii. 配置字段的映射关系
   iv. 配置作业速率上限、脏数据检查规则等信息
   v. 配置调度属性
3. 提交并发布任务
4. 任务运行与管理

新建离线同步节点

1. 创建业务流程。详情请参见：管理业务流程。
2. 创建离线同步节点。

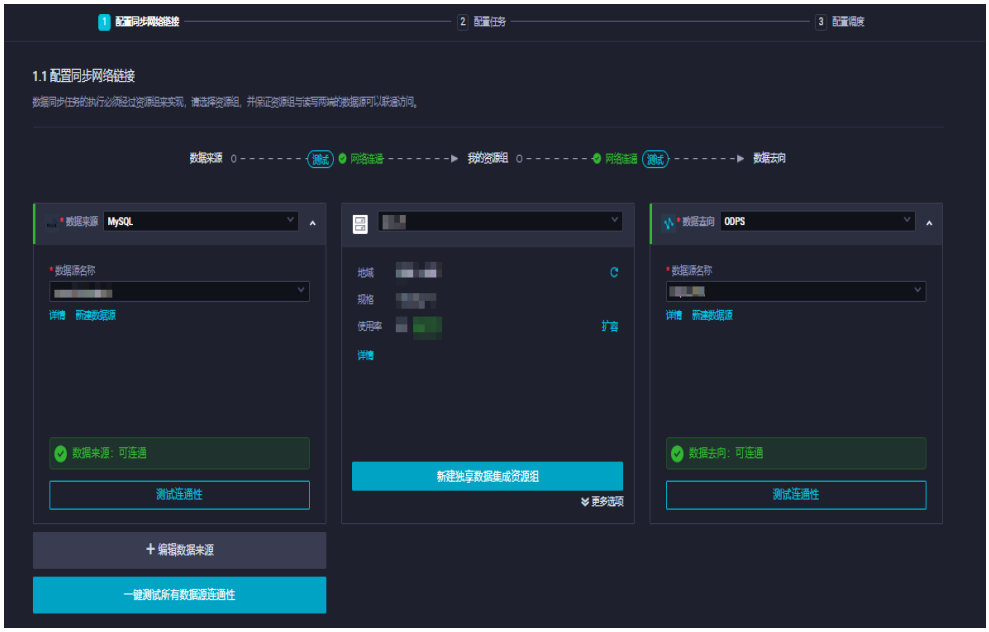
   你可以通过以下两种方式创建离线同步节点：

   - 方式一：展开业务流程，右键单击**数据集成 > 新建节点 > 离线同步**。
   - 方式二：双击业务流程名称，将**数据集成**目录下的**离线同步**节点直接拖拽至右侧业务流程编辑面板。

3. 根据界面提示创建离线同步节点。

配置离线同步任务

1. 配置同步网络链接。



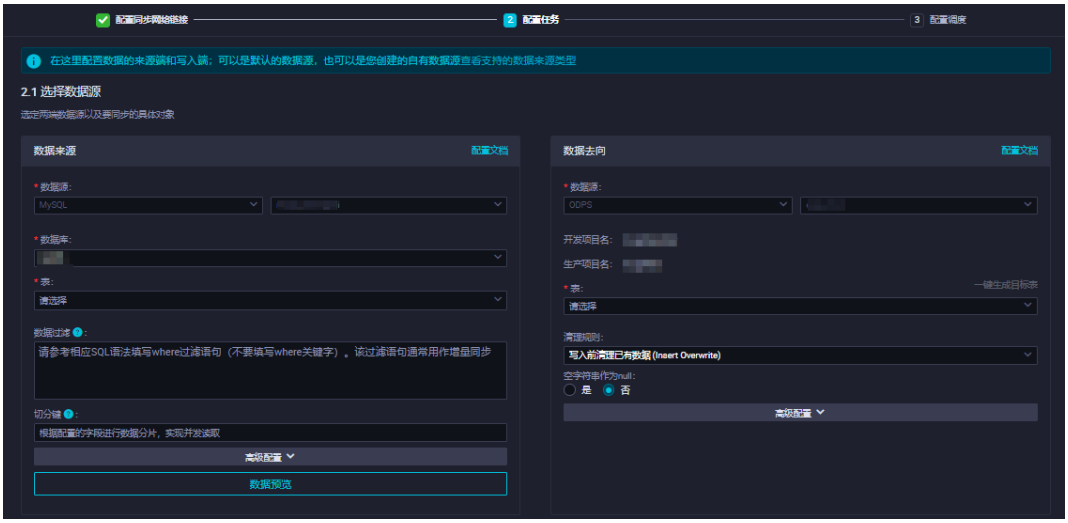
   选择离线同步任务的数据来源和数据去向，以及用于执行同步任务的资源组，并测试连通性。

   > **说明**
   >
   > - 还支持同步源端分库分表数据至目标单表，详情请参见：分库分表同步。
   > - 若数据源与数据库网络不通，请参考界面提示或文档进行网络连通配置。详情请参见：配置资源组与网络连通。

2. 单击**下一步**，配置同步任务。

i. 选择要同步的具体对象。

在选择数据源区域，配置任务读取与写入的表，及同步的数据范围。



配置同步范围：

- 当您在**数据过滤**文本框配置过滤条件时，执行同步任务时将只同步满足过滤条件的数据。同时，过滤条件可以结合调度参数使用，实现过滤条件随任务调度时间的动态变化，进而实现增量数据的同步。不同插件增量同步配置方式不同，关于增量数据同步配置详情请参见：配置数据增量同步。

  > **说明**
  >
  > 在数据过滤和目标表相关配置中定义的变量，当单击下一步，配置调度属性时，您可以为此处定义的变量赋值。实现增量或全量数据写入目标表对应时间分区等功能，关于调度参数的使用详情请参见：调度参数概述。

- 当不配置数据过滤条件时，默认同步全量数据。

ii. 配置字段映射关系。

支持同名映射、同行映射。在使用过程中，您还可以：

- 为目标字段赋值：通过**添加一行**的方式，为目标表添加常量、变量。例如，'123'，'${变量名}'。

  > **说明** 在此处定义的变量，当单击下一步，配置调度时，可以为变量赋值。关于调度参数的使用详情请参见：调度参数概述。

- 编辑源端字段：单击类型右侧的图标，即可使用源端数据库支持的函数，对字段进行函数处理，例如，通过Max(id)控制仅同步最大值。

  > **说明** MaxCompute Reader暂不支持使用函数。

> **注意** 配置字段映射关系后，任务将根据字段映射关系，将源端字段写入目标端对应类型的字段中。同步过程中可能存在源端与目标端字段类型不匹配，产生脏数据，导致数据无法正常写入目标端，同步过程中关于脏数据的容忍条数，请参考下一步**通道控制**进行配置。

iii. 通道控制。

您可通过通道配置，控制数据同步过程相关属性。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入目标端的最大线程数。 **说明** 由于资源规格等原因，实际执行时并发数可能小于等于此处配置的并发数，收费将按照实际执行的并发数收费。详情请参见：性能指标。

参数	描述
同步速率	■ 限流：您可以设置同步速率，以保护读取端数据库，避免抽取速度过大，给源库造成太大的压力。限速最小配置为1MB/S，最高上限为30MB/s。 ■ 不限流：在不限流的情况下，任务将在所配置的并发数的限制基础上，提供现有硬件环境下最大的传输性能。 **说明** 流量度量值是数据集成本身的度量值，不代表实际网卡流量。通常，网卡流量是通道流量膨胀的1至2倍，实际流量膨胀取决于具体的数据存储系统传输序列化情况。
错误记录数控制（脏数据控制）	脏数据的最大容忍条数。 ■ 配置为0，表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ■ 允许脏数据并设置其阈值时： ■ 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 ■ 若产生的脏数据超出阈值范围，同步任务将失败退出。 **说明** 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据。单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。您可以在同步任务配置时，控制同步过程中是否允许脏数据产生，并且支持控制脏数据条数，即当脏数据超过指定条数时，任务失败退出。
分布式处理能力	■ 开启：分布式执行模式可以将您的任务切片分散到多台执行节点上并发执行，进而做到同步速度随执行集群规模做水平扩展，突破单机执行瓶颈。 ■ 未开启：配置的并发数据仅仅是单机上的进程并发，无法利用多机联合计算。 如果您对于同步性能有比较高的诉求可以使用分布式模式。 另外分布式模式也可以使用机器的碎片资源，对资源利用率友好。 **注意** ■ 并发数大于等于8个才能开启分布式处理能力。 ■ 部分数据源支持分布式模式执行任务，详情请参见各插件配置文档：支持的数据源与读写插件。

3. 单击**下一步**，配置调度属性。
   - 配置调度参数：您在上述配置中定义的变量均可以在此处进行赋值，支持赋值常量与变量。
   - 配置时间属性：用于定义任务在生产环境的周期调度方式。您可以在调度配置的时间属性区域，配置任务生成周期实例的方式、调度类型、调度周期等属性。
   - 配置资源属性：任务的运行依赖于调度资源组，您可以在调度配置的资源属性区域，选择任务调度运行时需要使用的资源组。
4. 单击**完成配置**。

提交并发布任务

若任务需要进行周期性调度运行，您需要将任务发布至生产环境。关于任务发布，详情请参见：发布任务。

任务运行与管理

任务发布至生产环境后，调度系统会按照调度配置自动运行任务，生成周期实例。同时，支持您手动触发补数据或测试，生产补数据实例或测试实例。详情请参见：查看并管理周期任务。

> 说明
> 如果您需要同步历史数据至目标表对应分区中，您可以使用补数据功能，调度参数将根据补数据配置的业务日期自动替换为具体的值。

后续步骤

您还可以进入数据质量规则页面，对数据同步写入的目标表配置数据质量校验规则。详情请参见数据质量概述。

3.2.2. 通过脚本模式配置离线同步任务

本文为您介绍如何通过数据集成的脚本模式配置任务。

前提条件

1. 已完成数据源配置。您需要在数据集成同步任务配置前，配置好您需要同步的源端和目标端数据库，以便在同步任务配置过程中，可通过选择数据源名称来控制同步任务的读取和写入数据库。离线同步支持的数据源及其配置详情请参见支持的数据源与读写插件。

   > 说明 数据源相关能力介绍详情请参见：数据源概述。

2. 已购买合适规格的独享数据集成资源组。详情请参见：新增和使用独享数据集成资源组。
3. 独享数据集成资源组与数据源网络已打通。详情请参见：配置资源组与网络连通。

开发流程

1. 新建离线同步节点
2. 配置离线同步任务
   i. 配置同步网络链路
   ii. 转脚本并导入模板
   iii. 编辑脚本，配置数据来源与去向、作业速率上限、脏数据检查规则等信息
   iv. 配置调度属性
3. 提交并发布任务
4. 任务运行与管理

新建离线同步节点

1. 创建业务流程。详情请参见：管理业务流程。
2. 创建离线同步节点。

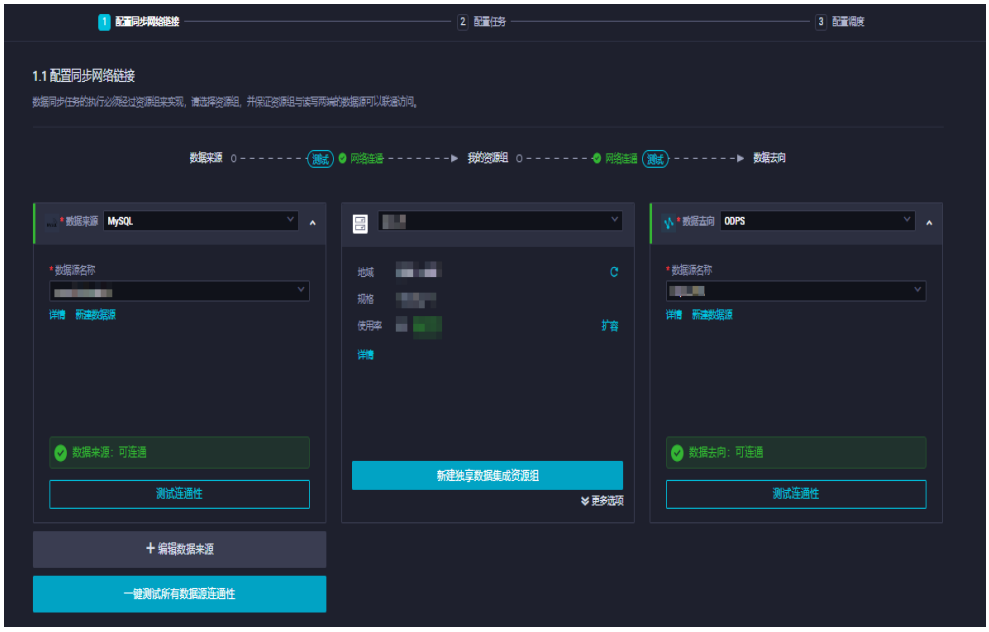
   你可以通过以下两种方式创建离线同步节点：

   - 方式一：展开业务流程，右键单击**数据集成 > 新建节点 > 离线同步**。
   - 方式二：双击业务流程名称，将**数据集成**目录下的**离线同步**节点直接拖拽至右侧业务流程编辑面板。

3. 根据界面提示创建离线同步节点。

配置离线同步任务

1. 配置同步网络链接。



选择离线同步任务的数据来源和数据去向，以及用于执行同步任务的资源组，并测试连通性。

> 说明
> - 支持同步源端分库分表数据至目标单表，详情请参见：分库分表同步。
> - 若数据源与数据库网络不通，请参考界面提示或文档进行网络连通配置。详情请参见：配置资源组与网络连通。

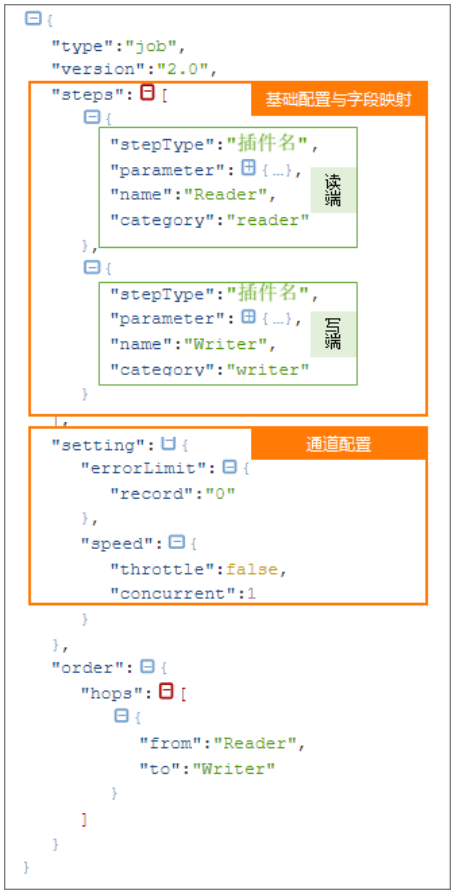
2. 转为脚本模式。

单击工具栏中的转换脚本图标。

如果脚本还未配置，您可以通过单击工具栏中的图标，根据界面提示快速导入脚本模板。

3. 编辑脚本，配置同步任务。

脚本模式通用配置如下：



i. 配置读端与写端的基础信息与字段映射关系。

各读写插件参数配置不同，详情请参见各个插件文档：支持的数据源与读写插件。

通过配置参数您可以：

- 为目标字段赋值：支持在源表待同步字段（column）上，为目标表增加常量与变量。例如，‘123’、'${变量名}'。在此处定义的变量，当单击下一步，配置调度时，可以为变量赋值。关于调度参数的使用详情请参见：调度参数概述。
- 编辑源端表字段：您可使用源端数据库支持的函数，对字段进行函数处理，例如，通过Max(id)控制仅同步最大值。

  > **说明** MaxCompute Reader暂不支持使用函数。

- 配置同步范围：部分插件可利用过滤参数实现增量同步，例如，通过MySQL Reader插件同步MySQL数据时，使用MySQL Reader插件的where参数结合DataWorks调度参数实现增量同步。关于增量数据同步配置详情请参见：配置数据增量同步。

  > **说明**
  > - 具体插件是否支持增量同步，及对应增量同步的具体实现，各个插件存在差异，请以实际插件文档为准。
  > - 支持增量同步相关参数的插件在配置同步任务时，若不配置数据过滤条件，则默认同步全量数据。
  > - 在数据过滤和目标表相关配置中定义的变量，当单击下一步，配置调度属性时，您可以为此处定义的变量赋值。实现增量或全量数据写入目标表对应时间分区等功能，关于调度参数的使用详情请参见：调度参数概述。

ii. 通道控制。

您可以在setting域中进行效率配置，主要包括并发数设置、同步速率设置、同步脏数据设置和同步资源组设置等信息。

参数	描述
executeMode（分布式处理能力）	■ distribute：开启分布式处理能力。分布式执行模式可以将您的任务切片分散到多台执行节点上并发执行，进而做到同步速度随执行集群规模做水平扩展，突破单机执行瓶颈。 ■ null：不开启分布式处理能力。配置的并发数据仅仅是单机上的进程并发，无法利用多机联合计算。 注意 ■ 并发数大于等于8个才能开启分布式处理能力。 ■ 部分数据与支持分布式模式执行任务，详情请参见具体插件配置文档。
concurrent（任务期望最大并发数）	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。 说明 由于资源规格等原因，实际执行时并发数可能小于等于此处配置的并发数，收费将按照实际执行的并发数收费。详情请参见：性能指标。
throttle（同步速率）	■ True：限流。以保护读取端数据库，避免抽取速度过大，给源库造成太大的压力。限速最小配置为1MB/S，最高上限为30MB/s。 说明 throttle设置为true时，您还需要设置mbps（同步速率）参数。 ■ False：不限流。在不限流的情况下，任务将在所配置的并发数的限制基础上，提供现有硬件环境下最大的传输性能。 说明 流量度量值是数据集成本身的度量值，不代表实际网卡流量。通常，网卡流量是通道流量膨胀的1至2倍，实际流量膨胀取决于具体的数据存储系统传输序列化情况。
errorLimit（错误记录数控制）	脏数据的最大容忍条数。 ■ 配置为0，表示不允许脏数据存在。如果同步过程中产生脏数据，任务将失败退出。 ■ 允许脏数据并设置其阈值时： ■ 若产生的脏数据在阈值范围内，同步任务将忽略脏数据（即不会写入目标端），并正常执行。 ■ 若产生的脏数据超出阈值范围，同步任务将失败退出。 说明 脏数据认定标准：脏数据是对业务没有意义，格式非法或者同步过程中出现问题的数据。单条数据写入目标数据源过程中发生了异常，则此条数据为脏数据。 因此只要是写入失败的数据均被归类于脏数据。 例如，源端是VARCHAR类型的数据写到INT类型的目标列中，则会因为转换不合理导致脏数据不会成功写入目的端。您可以在同步任务配置时，控制同步过程中是否允许脏数据产生，并且支持控制脏数据条数，即当脏数据超过指定条数时，任务失败退出。

4. 单击**下一步**，配置调度属性。
   - 配置调度参数：您在上述配置中定义的变量均可以在此处进行赋值，支持赋值常量与变量。
   - 配置时间属性：用于定义任务在生产环境的周期调度方式。您可以在调度配置的时间属性区域，配置任务生成周期实例的方式、调度类型、调度周期等属性。
   - 配置资源属性：任务的运行依赖于调度资源组，您可以在调度配置的资源属性区域，选择任务调度运行时需要使用的资源组。

提交并发布任务

若任务需要进行周期性调度运行，您需要将任务发布至生产环境。关于任务发布，详情请参见：发布任务。

任务运行与管理

任务发布至生产环境后，调度系统会按照调度配置自动运行任务，生成周期实例。同时，支持您手动触发补数据或测试，生产补数据实例或测试实例。详情请参见：查看并管理周期任务。

> **说明**
>
> 如果您需要同步历史数据至目标表对应分区中，您可以使用补数据功能，调度参数将根据补数据配置的业务日期自动替换为具体的值。

后续步骤

您还可以进入数据质量规则页面，对数据同步写入的目标表配置数据质量校验规则。详情请参见数据质量概述。

3.2.3. 通过OpenAPI创建离线同步任务

本文为您介绍如何使用OpenAPI创建数据集成同步任务，同步来源端数据至去向端。

前提条件

- 已创建Maven项目，详情请参见创建Maven项目。
- 已创建业务流程，详情请参见创建业务流程。
- 已创建同步任务需要的数据源，详情请参见配置数据源。

使用限制

- DataWorks当前仅支持使用OpenAPI创建数据集成离线同步任务。
- 调用CreateDISyncTask创建数据集成同步任务，仅支持使用脚本模式配置同步任务内容，详情请参见通过脚本模式配置离线同步任务。
- DataWorks暂不支持使用OpenAPI创建业务流程，您需要使用现有的业务流程创建数据同步任务。

配置环境依赖及账号认证

- 配置Maven依赖。
  i. 打开Maven项目下的*pom.xml*文件，添加 `aliyun-java-sdk-core` 。

```
<dependency>
  <groupId>com.aliyun</groupId>
  <artifactId>aliyun-java-sdk-core</artifactId>
  <version>4.5.20</version>
</dependency>
```

  ii. 打开Maven项目下的*pom.xml*文件，添加 `aliyun-java-sdk-dataworks-public` 。

```
<dependency>
  <groupId>com.aliyun</groupId>
  <artifactId>aliyun-java-sdk-dataworks-public</artifactId>
  <version>3.3.18</version>
</dependency>
```

- 客户端认证。

  使用OpenAPI创建数据同步任务前，需要调用如下语句对登录阿里云的账号相关信息进行认证。如果阿里云的账号信息认证通过，则继续执行后续任务，如果认证不通过，则该调用会报错，您需要根据实际报错处理相关问题。

```
DefaultProfile profile = DefaultProfile.getProfile(
                         "regionId",           //DataWorks工作空间所在的地域，例如cn-hangzhou。
                         "<yourAccessKeyId>",  //登录DataWorks工作空间的阿里云账号的AccessKey ID。
                         "<yourAccessSecret>"); //登录DataWorks工作空间的阿里云账号的AccessKey Secret。
IAcsClient client = new DefaultAcsClient(profile);
```

您可以登录DataWorks控制台鼠标悬停至右上角的用户头像，单击**AccessKey管理**，进入**AccessKey管理**页面获取AccessKey ID和AccessKey Secret。

配置流程

完成上述配置环境依赖及账号认证后，您可以通过OpenAPI调用相关接口，创建数据同步任务，同步来源端数据至去向端。配置流程如下：

1. 创建数据集成任务。
2. 配置任务的调度依赖。
3. 提交数据集成任务。
4. 发布同步任务至生产环境。

配置步骤

1. 创建数据集成任务。

   调用*CreateDISyncTask*接口，创建数据集成任务。如下代码仅示例部分参数的配置，更多参数详情请参见CreateDISyncTask。

```
public void createFile() throws ClientException{
        CreateDISyncTaskRequest request = new CreateDISyncTaskRequest();
        request.setProjectId(181565L);
        request.setTaskType("DI_OFFLINE");
        request.setTaskContent("{\"type\":\"job\",\"version\":\"2.0\",\"steps\":[{\"stepType\":\"mysql\",
\"parameter\":{\"envType\":1,\"datasource\":\"dh_mysql\",\"column\":[\"id\",\"name\"],\"tableComment\":\"
same表comment\",\"connection\":[{\"datasource\":\"dh_mysql\",\"table\":[\"same\"]}],\"where\":\"\",\"split
Pk\":\"id\",\"encoding\":\"UTF-8\"},\"name\":\"Reader\",\"category\":\"reader\"},{\"stepType\":\"odps\",\
"parameter\":{\"partition\":\"pt=${bizdate}\",\"truncate\":true,\"datasource\":\"odps_first\",\"envType\"
:1,\"column\":[\"id\",\"name\"],\"emptyAsNull\":false,\"tableComment\":\"same表comment\",\"table\":\"same\
"},\"name\":\"Writer\",\"category\":\"writer\"}],\"setting\":{\"errorLimit\":{\"record\":\"\"},\"speed\":
{\"throttle\":false,\"concurrent\":2}},\"order\":{\"hops\":[{\"from\":\"Reader\",\"to\":\"Writer\"}]}}");
        request.setTaskParam("{\"FileFolderPath\":\"业务流程/new_biz/数据集成\",\"ResourceGroup\":\"S_res_g
roup_280749521950784_1602767279794\"}");
        request.setTaskName("new_di_task_0607_1416");
        String akId = "XXX";
        String akSecret = "XXXX";
        String regionId = "cn-hangzhou";
        IClientProfile profile = DefaultProfile.getProfile(regionId, akId, akSecret);
        DefaultProfile.addEndpoint("cn-hangzhou","dataworks-public","dataworks.cn-hangzhou.aliyuncs.com")
;
        IAcsClient client;
        client = new DefaultAcsClient(profile);
        CreateDISyncTaskResponse response1 = client.getAcsResponse(request);
        Gson gson1 = new Gson();
        System.out.println(gson1.toJson(response1));
}
```

2. 设置任务的调度依赖。

   调用*UpdateFile*接口，设置数据集成任务的调度依赖，参数详情请参见UpdateFile。

3. 提交数据集成任务。

   调用*SubmitFile*接口，提交数据集成任务至调度系统的开发环境。任务提交后，Response会返回deploymentId，您可以调用GetDeployment接口，通过deploymentId获取本次发布包的详细信息。

```
public void submitFile() throws ClientException{
        SubmitFileRequest request = new SubmitFileRequest();
        request.setProjectId(78837L);
        request.setProjectIdentifier("zxy_8221431");
      // 此节点ID为创建节点时返回的ID，对应数据库File表的file_id。
        request.setFileId(501576542L);
        request.setComment("备注");
        SubmitFileResponse acsResponse = client.getAcsResponse(request);
      //调用GetDeployment接口，获取本次发布的具体情况。
      Long deploymentId = acsResponse.getData();
        log.info(acsResponse.toString());
}
```

上述代码仅示例部分参数的配置，更多参数详情请参见Submit FileGet Deployment。

4. 发布同步任务到生产环境。

   调用DeployFile接口，发布数据集成同步任务至生产环境。

   > **说明** 仅标准模式的工作空间涉及执行该发布操作。

```
public void deploy() throws ClientException{
        DeployFileRequest request = new DeployFileRequest();
        request.setProjectIdentifier("zxy_8221431");
        request.setFileId(501576542L);
        request.setComment("备注");
        //NodeId和file_id二选一。NodeId的值为调度配置中基础属性的节点ID。
        request.setNodeId(700004537241L);
        DeployFileResponse acsResponse = client.getAcsResponse(request);
         //调用GetDeployment接口，获取本次发布的具体情况。
        Long deploymentId = acsResponse.getData();
        log.info(acsResponse.getData().toString());
    }
```

上述代码仅示例部分参数的配置，更多参数详情请参见DeployFile。

5. 获取发布包详情。

   任务发布后，Response会返回deploymentId，您可以调用GetDeployment接口，通过deploymentId获取本次发布包的详细信息。当GetDeployment接口的返回参数Status取值为 7时，则表示此次发布成功。

```
public void getDeployment() throws ClientException{
        GetDeploymentRequest request = new GetDeploymentRequest();
        request.setProjectId(78837L);
        request.setProjectIdentifier("zxy_8221431");
      //DeploymentId为提交或发布的返回值。
        request.setDeploymentId(2776067L);
        GetDeploymentResponse acsResponse = client.getAcsResponse(request);
        log.info(acsResponse.getData().toString());
    }
```

上述代码仅示例部分参数的配置，更多参数详情请参见Get Deployment。

修改同步任务的相关配置

成功创建数据集成同步任务后，您可以调用UpdateDISyncTask接口更新任务的Content，或通过TaskParam来更新使用的独享资源组。更新后，您需要重新提交、发布同步任务，详情请参见配置流程。

3.2.4. 同步场景示例

3.2.4.1. 数据增量同步

本文以同步业务RDS数据库的数据至MaxCompute为例，为您介绍如何对不同场景的数据进行增量同步。

背景信息

根据需要同步的数据在写入后是否发生变化，分为恒定的存量数据（通常是日志数据）和持续更新的数据（例如人员表中，人员的状态会发生变化）。

根据幂等性原则（一个任务多次运行的结果一致，则该任务支持重跑调度。如果该任务出现错误，脏数据较容易清理），每次导入数据都是导入至一张单独的表或分区中，或者覆盖历史记录。

本文定义任务测试时间是2016年11月14日，在14日进行增量同步，同步历史数据至分区ds=20161113中。增量同步的场景配置了自动调度，把增量数据在15日凌晨同步至分区ds=20161114中。数据中的时间字段optime用来表示该数据的修改时间，从而判断这条数据是否为增量数据。

使用说明

- 部分数据源暂无增量同步方案，例如HBase、OTSStream数据源等。具体数据源是否支持增量同步可以看具体的Reader插件文档。
- 每个插件实现增量同步的所配置的参数可能不同，具体参数配置可以参考对应的Reader插件文档，详情可参考：支持的数据源与读写插件。例如：

数据库类型	增量同步需配置的参数	支持的语法
MySQL Reader	where **说明** 向导模式配置时，需要配置的界面参数名为：数据过滤。	数据库语法 **说明** 可与调度参数结合实现每日读取指定时间区间的数据。
MongoDB Reader	query **说明** 向导模式配置时，需要配置的界面参数名为：检索查询条件。	基本与数据库一致 **说明** 可与调度参数结合实现每日读取指定时间区间的数据。
OSS Reader	Object	指定路径 **说明** 与调度参数结合实现每日读取指定文件数据。
...	...	...

- 调度参数将根据任务运行的业务时间实现参数值的动态替换，实现每日数据增量同步。关于调度参数的使用详情可参考文档：调度参数概述。

  如下图，将每日MySQL增量数据写入到MaxCompute表对应分区中。

新建业务流程

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。

4. 在**数据开发**面板，右键单击**业务流程**，选择**新建业务流程**。
5. 在**新建业务流程**对话框中，输入**业务名称**和**描述**。

   > **说明** 节点名称的长度不能超过128个字符。

6. 单击**新建**，即可完成业务流程的创建。

对恒定的存量数据进行增量同步

由于数据生成后不会发生变化，因此可以很方便地根据数据的生成规律进行分区。较常见的是根据日期进行分区，例如每天1个分区。

1. 在RDS数据库中，执行下述语句准备数据。

```
drop table if exists oplog;
create table if not exists oplog(
optime DATETIME,
uname varchar(50),
action varchar(50),
status varchar(10)
 );
Insert into oplog values(str_to_date('2016-11-11','%Y-%m-%d'),'LiLei','SELECT','SUCCESS');
Insert into oplog values(str_to_date('2016-11-12','%Y-%m-%d'),'HanMM','DESC','SUCCESS');
```

   上述的两条数据作为历史数据，需要先进行一次全量数据同步，将历史数据同步至昨天的分区。

2. 在**数据开发**页面，右键单击业务流程下的**表**，选择**新建表**。
3. 在**新建表**对话框中，输入**表名**（ods_oplog），单击**提交**。
4. 双击ods_oplog表，在右侧的编辑页面单击**DDL模式**，输入下述建表语句。

```
--创建好MaxCompute表，按天进行分区。
create table if not exists ods_oplog(
 optime datetime,
 uname string,
 action string,
 status string
) partitioned by (ds string);
```

5. 配置同步历史数据的任务，详情请参见创建同步任务。

   测试同步任务成功后，单击节点编辑页面右侧的**调度配置**，勾选**暂停调度**并重新提交或发布，避免任务自动调度执行。

6. 执行下述语句，向RDS源头表中插入数据作为增量数据。

```
insert into oplog values(CURRENT_DATE,'Jim','Update','SUCCESS');
insert into oplog values(CURRENT_DATE,'Kate','Delete','Failed');
insert into oplog values(CURRENT_DATE,'Lily','Drop','Failed');
```

7. 配置同步增量数据的任务。

   在**数据来源**中设置数据过滤为 `date_format(optime,'%Y%m%d')=${bdp.system.bizdate}` ，在**数据去向**中设置分区信息为 `${bdp.system.bizdate}` 。

   > **说明** 通过配置数据过滤，在15日凌晨进行同步时，您可以查询14日源头表全天新增的数据，并同步至目标表的增量分区中。

8. 查看同步结果。

   单击节点编辑页面右侧的**调度配置**，设置任务的调度周期为天调度。提交或发布任务后，第2天任务将自动调度执行。执行成功后，即可查看MaxCompute目标表的数据。

对持续更新的数据进行增量同步

根据数据仓库反映历史变化的特点，建议每天对人员表、订单表等会发生变化的数据进行全量同步，即每天保存的都是全量数据，方便您获取历史数据和当前数据。

真实场景中因为某些特殊情况，需要每天只进行增量同步。但MaxCompute不支持Update语句修改数据，只能通过其它方式实现。下文将为您介绍两种同步策略（全量同步、增量同步）的具体操作。

1. 执行下述语句准备数据。

```
drop table if exists user ;
create table if not exists user(
    uid int,
    uname varchar(50),
    deptno int,
    gender VARCHAR(1),
    optime DATETIME
    );
--历史数据
insert into user values (1,'LiLei',100,'M',str_to_date('2016-11-13','%Y-%m-%d'));
insert into user values (2,'HanMM',null,'F',str_to_date('2016-11-13','%Y-%m-%d'));
insert into user values (3,'Jim',102,'M',str_to_date('2016-11-12','%Y-%m-%d'));
insert into user values (4,'Kate',103,'F',str_to_date('2016-11-12','%Y-%m-%d'));
insert into user values (5,'Lily',104,'F',str_to_date('2016-11-11','%Y-%m-%d'));
--增量数据
update user set deptno=101,optime=CURRENT_TIME  where uid = 2; --null改成非null
update user set deptno=104,optime=CURRENT_TIME  where uid = 3; --非null改成非null
update user set deptno=null,optime=CURRENT_TIME  where uid = 4; --非null改成null
delete from user where uid = 5;
insert into user(uid,uname,deptno,gender,optime) values (6,'Lucy',105,'F',CURRENT_TIME);
```

2. 进行数据同步。
   - 每天全量同步
     a. 执行下述语句创建MaxCompute表，新建表的详情请参见新建表。

```
--全量同步
create table ods_user_full(
    uid bigint,
    uname string,
    deptno bigint,
    gender string,
    optime DATETIME
) partitioned by (ds string);
```

     b. 配置全量同步任务。

> **说明** 需要每天都全量同步，因此任务的调度周期需要配置为天调度。

     c. 运行任务，并查看同步后MaxCompute目标表的结果。

     因为每天都是全量同步，没有全量和增量的区别，所以第2天任务自动调度执行成功后，即可看到数据结果。

   - 每天增量同步

     不推荐使用该方式，只有在不支持Delete语句、无法通过SQL语句查看被删除的数据等场景才会考虑。虽然实际上大多使用逻辑删除数据，将Delete转化为Update进行处理。但仍会限制一些特殊的业务场景不能实现，导致数据不一致。并且同步后需要合并新增数据和历史数据。

     准备数据

     需要创建两张表，一张写当前的最新数据，一张写增量数据。

```
--结果表
create table dw_user_inc(
    uid bigint,
    uname string,
    deptno bigint,
    gender string,
    optime DATETIME
);
```

```
--增量记录表
create table ods_user_inc(
    uid bigint,
    uname string,
    deptno bigint,
    gender string,
    optime DATETIME
)
```

a. 配置同步任务，将全量数据直接写入结果表。

> **说明** 只需要执行一次，执行成功后需要单击页面右侧的**调度配置**，勾选**暂停调度**。

b. 配置同步任务，将增量数据写入到增量表。设置数据过滤，即**where**参数配置为 `date_format(optime,'%Y%m%d')=${bdp.system.bizdate}` 。

c. 执行下述语句合并数据。

```
insert overwrite table dw_user_inc
select
--所有select操作，如果ODS表有数据，说明发生了变动，以ODS表为准。
case when b.uid is not null then b.uid else a.uid end as uid,
case when b.uid is not null then b.uname else a.uname end as uname,
case when b.uid is not null then b.deptno else a.deptno end as deptno,
case when b.uid is not null then b.gender else a.gender end as gender,
case when b.uid is not null then b.optime else a.optime end as optime
from
dw_user_inc a
full outer join ods_user_inc b
on a.uid  = b.uid ;
```

查看执行结果会发现Delete的记录没有同步成功。

每天增量同步的优点是同步的增量数据量较小，但可能出现数据不一致的情况，并且需要通过额外的计算进行数据合并。

如果不是必要情况，对持续更新的数据进行每天全量同步即可。如果希望历史数据仅保留一定的时间，自动删除超出保留时间的数据，您可以设置Lifecycle。

3.2.4.2. 分库分表同步

数据集成支持分库分表。您可以在一个任务中配置多个数据源多张表后，同步至一个目标表中。

背景信息

配置分库分表同步时，请确保所有表的Schema与同步配置中第一个数据源的第一张表保持一致。

分库分表支持MySQL（支持向导模式）、SQL Server、Oracle、PostgreSQL、PolarDB和AnalyticDB等类型的数据源。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 创建离线同步节点。
   i. 鼠标悬停至+新建图标，单击**数据集成 > 离线同步**。

      您也可以找到相应的业务流程，右键单击**数据集成**，选择**新建 > 离线同步**。
   ii. 在**新建节点**对话框中，输入**节点名称**，并选择**目标文件夹**。

> **说明** 节点名称必须是大小写字母、中文、数字、下划线（_）和小数点（.），且不能超过128个字符。

iii. 单击**提交**。

3. 配置离线同步节点。

   您可以通过向导模式和脚本模式配置分库分表同步：

   - 如果您通过向导模式配置分库分表同步，请单击**数据来源**区域的**添加分库分表+**，选择对应的**数据源**和表。详情请参见通过向导模式配置离线同步任务。

     > **说明** 仅MySQL支持通过向导模式添加分库分表，其他数据库请切换至脚本模式配置分库分表同步，详情请参见通过脚本模式配置离线同步任务。

     您还可以单击**新建数据源**进行新建，详情请参见配置MySQL数据源。

   - 如果您通过脚本模式配置分库分表同步，示例如下。详情请参见通过脚本模式配置离线同步任务。

     > **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {
            "plugin": "mysql",
            "parameter": {
                "connection": [
                    {
                        "table": [ //分库分表的Table列表1
                            "tbl1",
                            "tbl2",
                            "tbl3"
                        ],
                        "datasource": "datasourceName1" //分库分表的数据源1
                    },
                    {
                        "table": [ //分库分表的Table列表2
                            "tbl4",
                            "tbl5",
                            "tbl6"
                        ],
                        "datasource": "datasourceName2" //分库分表的数据源2
                    }
                ],
                "singleOrMulti": "multi",
                "splitPk": "db_id",
                "column": [
                    "id", "name", "age"
                ],
                "where": "1 < id and id < 100"
            }
        },
        "writer": {
        }
    }
}
```

4. 提交节点。

   > **注意** 您需要设置节点的**重跑属性**和**依赖的上游节点**，才可以提交节点。

   i. 单击工具栏中的图标。

   ii. 在**提交新版本**对话框中，输入**备注**。

   iii. 单击**确认**。

   如果您使用的是标准模式的工作空间，提交成功后，请单击右上方的**发布**。具体操作请参见发布任务。

5. 测试节点，详情请参见查看并管理周期任务。

3.3. 整库迁移与批量上云

整库迁移是帮助提升用户效率、降低用户使用成本的一种快捷工具，它可以快速把来源数据源内所有表一并上传至目标数据源，可节省大量初始化数据上云的批量任务创建时间。

支持的数据源

当前DataWorks支持各类数据源的数据整库迁移至MaxCompute，支持的数据源类型和操作指导链接如下。

去向数据源	来源数据源	操作指导链接

去向数据源	来源数据源	操作指导链接
MaxCompute	• MySQL • PostgreSQL • SQL Server • Oracle • PolarDB • AnalyticDB for MySQL2.0 • AnalyticDB for MySQL 3.0 • AnalyticDB for PostgreSQL • HybridDB for MySQL • DRDS（PolarDB-X） • DM • Hive • DB2	• 整库离线同步（周期性全量） • 整库离线同步（周期性增量） • 整库离线同步（一次性全量） • 整库离线同步（一次性增量） • 整库离线同步（一次性全量周期性增量）

3.4. 离线数据同步任务调优

3.4.1. 性能调优配置

本文为您介绍影响数据同步速度的因素、如何通过调整同步任务的并发配置来实现同步速度最大化、作业的限速选项，以及数据同步过慢的场景。

DataWorks数据集成支持任意位置和网络环境下的数据源之间的实时与离线数据互通，是一站式数据同步的全栈平台，让您能在各种云和本地数据存储中每天同步数据。

DataWorks具有极强的数据传输性能，支持400多对异构数据源之间的数据互通，确保您可以专注于构建大数据解决方案的核心问题。

数据同步速度的影响因素

影响数据同步速度的因素如下：

- 来源端数据源
  - 数据库的性能：CPU、内存、SSD硬盘、网络和硬盘等。
  - 并发数：数据源并发数越高，数据库负载越高。
  - 网络：网络的带宽（吞吐量）、网速。通常，数据库的性能越好，它可以承载的并发数越高，您可以为数据同步作业配置越多的并发数据抽取。
- 数据集成的同步任务配置
  - 传输速度：是否设置任务同步速度上限值。
  - 并发：从源并行读取或并行写入数据存储端的最大线程数。
  - WAIT资源。
  - Bytes的设置：单个线程的Bytes=1048576，在网速比较敏感时，会出现超时现象，此时建议将Bytes设置的较小。
  - 查询语句是否建索引。
- 目的端数据源
  - 性能：CPU、内存、SSD 硬盘、网络和硬盘。
  - 负载：目的数据库负载过高会影响同步任务数据写入效率。
  - 网络：网络的带宽（吞吐量）、网速。

数据源端和目的端数据库的性能、负载和网络情况主要由您自己关注和调优，下文为您介绍在数据集成产品中同步任务的优化配置。

并发

向导模式下，通过界面化配置并发数，指定任务所使用的并行度。通过脚本模式配置并发数的示例如下。

```
"setting": {
      "speed": {
        "concurrent": 10
      }
    }   }
```

限速

数据集成同步任务默认不限速，任务将在所配置的并发数的限制上以最高能达到的速度进行同步。另一方面，考虑到速度过高可能对数据库造成过大的压力从而影响生产，数据集成同时提供了限速选项，您可以按照实际情况调优配置（建议选择限速之后，最高速度上限不应超过30 MB/s）。脚本模式通过如下示例代码配置限速，代表1 MB/s的传输带宽。

```
"setting": {
      "speed": {
         "throttle": true // 是否限流。
        "mbps": 1,  // 具体速率值。
      }
    }
```

- throttle包括*true*和*false*:
  - 当throttle设置为*true*时，表示限速，您必须设置mbps具体的数据值。如果没有设置mbps，程序运行将会出错或者速率异常。
  - 当throttle设置为*false*时，表示不限速，则mbps的配置无意义。
- 流量度量值是数据集成本身的度量值，不代表实际网卡流量。通常，网卡流量往往是通道流量膨胀的1至2倍，实际流量膨胀取决于具体的数据存储系统传输序列化情况。
- 半结构化的单个文件没有切分键的概念，多个文件可以设置作业速率上限来提高同步的速度，但作业速率上限和文件的个数有关。

  例如，有n个文件，作业速率上限最多设置为n MB/s:
  - 如果设置n+1 MB/s，还是以n MB/s速度同步。
  - 如果设置为n-1 MB/s，则以n-1 MB/s速度同步。
- 关系型数据库设置作业速率上限和切分键后，才能根据作业速率上限将表进行切分。关系型数据库通常只支持数值型作为切分键，但Oracle数据库支持以数值型和字符串类型作为切分键。

数据同步过慢的场景

- 场景一：同步任务使用公共调度（WAIT）资源时，一直在等待状态。
  - 场景示例

    在DataWorks中对任务进行测试时，出现任务一直等待的状态，同时提示系统内部错误。

    例如使用默认资源组，完成从RDS同步数据至MaxCompute的任务，共等待了约800秒，但是日志显示任务只运行了18秒。现在运行其它同步任务进行测试，也一直处于等待中 。

    显示的等待日志如下所示。

    ```
    2017-01-03 07:16:54 : State: 2(WAIT) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
    ```
  - 解决方法

    因为您使用的是公共调度资源，公共资源能力是受限的。可能不只是单个用户的2~3个任务在使用，很多项目也在使用。任务实际运行10秒，但是延长到800秒，是因为您的任务下发执行时，发现资源不足，需等待获取资源。

    如果您对同步速度和等待时间比较敏感，建议在低峰期配置同步任务。通常，晚上零点到3点同步任务较多，您可以避开零点到3点的时间段，便可相对减少等待资源的情况。
- 场景二：提高多个任务导入数据到同一张表的同步速度。
  - 场景示例

    想要将多个数据源的表同步至一张表中，所以将同步任务设置成串行任务，但是最后发现同步时间很长。

- 解决方法

  可以同时启动多个任务，同时往一个数据库写入数据，请注意以下问题：

  - 确保目标数据库负载能力是能够承受的，避免不能正常工作。
  - 在配置工作流任务时，可以选择单个任务节点配置分库分表任务，或在一个工作流中设置多个节点同时执行。
  - 如果任务执行时，出现等待资源（WAIT）情况，可以低峰期配置同步任务，保证任务有较高的执行优先级。

- 场景三：数据同步任务where条件没有索引，导致全表扫描同步变慢。
  - 场景示例

    执行的SQL如下所示。

    ```
    select bid,inviter,uid,createTime from `relatives` where createTime>='2016-10-2300:00:00'and reateTime<'2016-10-24 00:00:00';
    ```

    从2016-10-25 11:01:24.875开始执行，到2016-10-25 11:11:05.489开始返回结果。同步程序在等待数据库返回SQL查询结果，MaxCompute需等待很久才能执行。

  - 分析原因

    where条件查询时，createTime列没有索引，导致查询全表扫描。

  - 解决方法

    建议where条件使用有索引相关的列，提高性能，索引也可以补充添加。

3.4.2. 数据同步任务调优

数据同步任务调度运行时，您可能会遇到实例的执行时间超过预期的情况。本文为您介绍如何在数据同步任务实例执行慢、时间差异大等不满足预期的情况下进行任务调优。

前提条件

正式开始数据同步任务调优前，请首先获取任务的运行日志和属性信息。

针对数据同步任务，DataWorks的调度资源分为一级调度资源和二级运行资源。

- 一级调度资源： 您可以进入**运维中心 > 周期任务运维 > 周期实例**页面，右键单击相应节点，选择**查看更多详情**，即可查看该节点的**属性**。
- 二级运行资源：您可以进入**数据集成 > 同步资源管理 > 资源组**页面，新增和查看二级任务运行资源。

背景信息

通常数据同步任务执行慢的场景分为以下三种：

- 任务开始运行的时间和调度时间差异比较大。
- 任务长时间处于WAIT状态。
- 任务同步的速率慢。

场景一：任务开始运行时间和调度时间差异较大

在任务开始运行时间和调度时间差异较大的情况下，您首先需要获取任务的运行日志和**属性**信息。经过对比发现， **运行日志**中开始运行的时间和节点**属性**中的调度时间有差异，时间主要耗费在等待调度上。

问题示例

1. 在**运维中心**中的**周期任务运维 > 周期任务**页面，右键单击相应节点，选择**查看更多详情**，查看节点的**属性**，发现调度时间在00:00。但是开始运行的时间在00:29，推测时间主要消耗在等待调度上。

2. 在**周期任务运维 > 周期实例**页面，右键单击相应实例，选择**查看运行日志**。查看任务从00:29开始运行，在00:30运行结束，整个任务执行仅花费了1分钟。说明本次任务本身的执行无问题。

解决方法

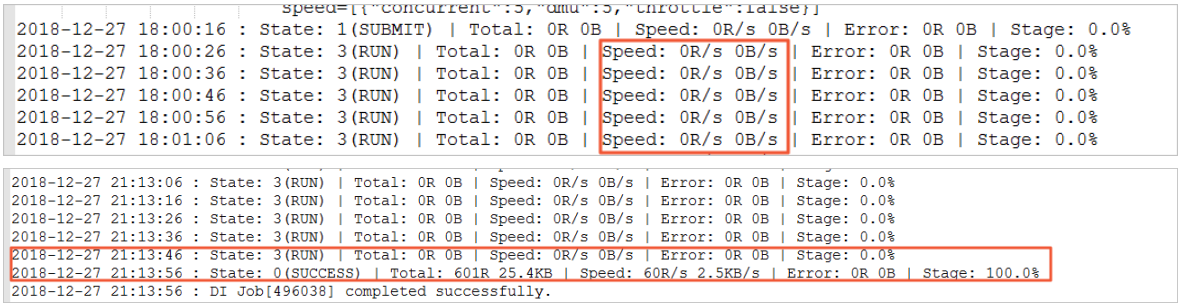
1. 首先建议您确认工作空间下是否有较多的任务同时进行调度。默认资源组下的一级调度资源有限，如果有较多的任务同时进行调度，会导致其它任务排队等待。
2. 通常每天0点~2点是业务调度的高峰期，建议您设置的业务运行时间尽量避开高峰期 。

场景二：同步任务一直运行，但速率为0

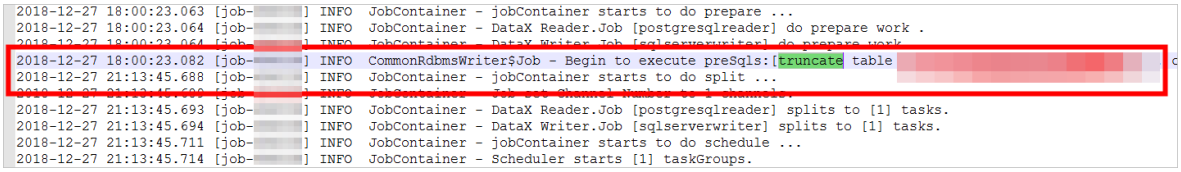
查看运行日志时，发现任务长时间处于运行状态，但速率为0。通常是由于拉取的SQL执行比较慢（源数据库CPU负载高或网络流量占用高），或在拉取SQL前进行truncate等操作，导致处理时间较长。

问题示例

1. 查看任务运行日志，任务长时间执行，但速率为0，从18:00开始到21:13结束。



2. 查看运行日志中存在truncate操作记录，从18:00开始到21:13结束。



解决方法

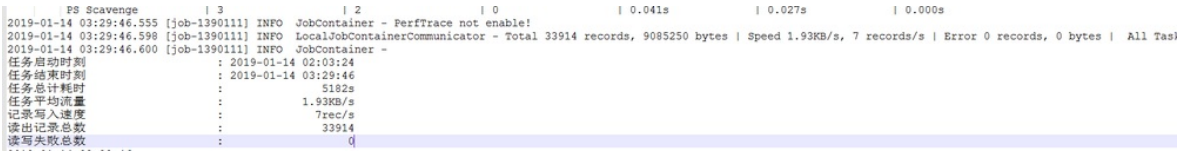
如问题示例所示，可能是truncate操作导致的同步任务慢，您需要检查源数据库truncate慢的原因。

场景三：同步任务的速率较低

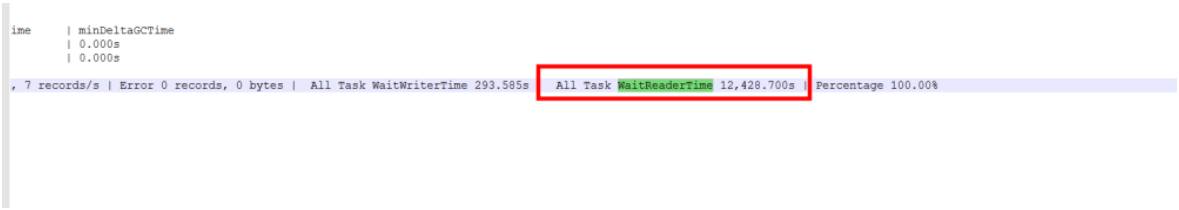
查看运行日志时，发现任务同步速率不为0，但是速率较低。

问题示例

1. 获取运行日志后，查看日志中的同步速率较低，约为1.93kb/s。



2. 查看运行日志中的同步时间消耗字段WaitWriterTime、WaitReaderTime等信息，发现WaitReaderTime时间较长，主要在等待读数据。



解决方法

针对速率比较慢的情况，您可以查看主要在等Writer还是Reader，如果是读写过程较慢，请查看对应的源数据库或目标数据库的负载情况。

3.5. 离线任务资源的使用说明

本文为您介绍数据集成离线任务并发度的配置方法、并发度和资源的占用关系，以及同步速度。

并发度的配置方法

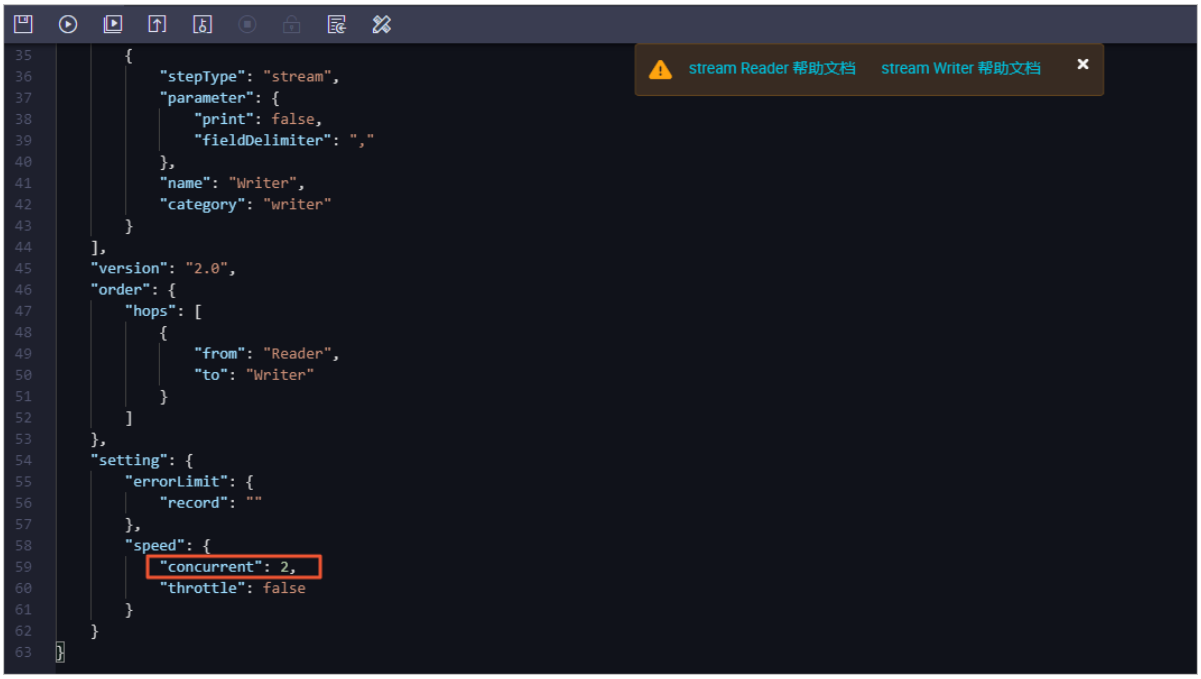
数据集成的离线同步任务主要通过设置并发度，来控制任务的占用和同步速度。离线同步任务包括向导模式和脚本模式：

- 通过向导模式配置离线同步任务，详情请参见通过向导模式配置离线同步任务。

  在向导模式编辑页面的**通道控制**区域，您可以通过配置**任务期望最大并发数**来控制离线任务的并发度。

- 通过脚本模式配置离线同步任务，详情请参见通过脚本模式配置离线同步任务。

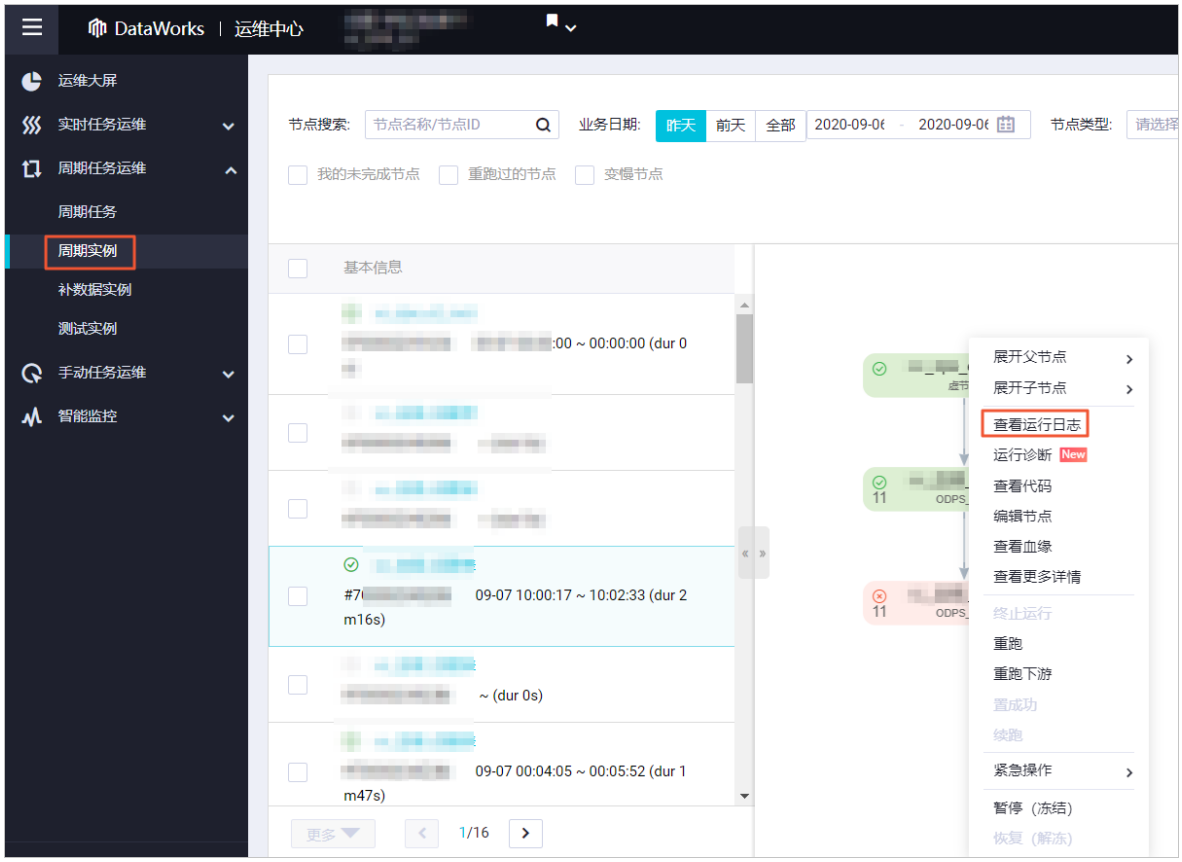
  在脚本模式的编辑页面，您可以在JSON结构的配置文本中，通过JSON路径*$.setting.speed.concurrent*设置离线任务的并发度。



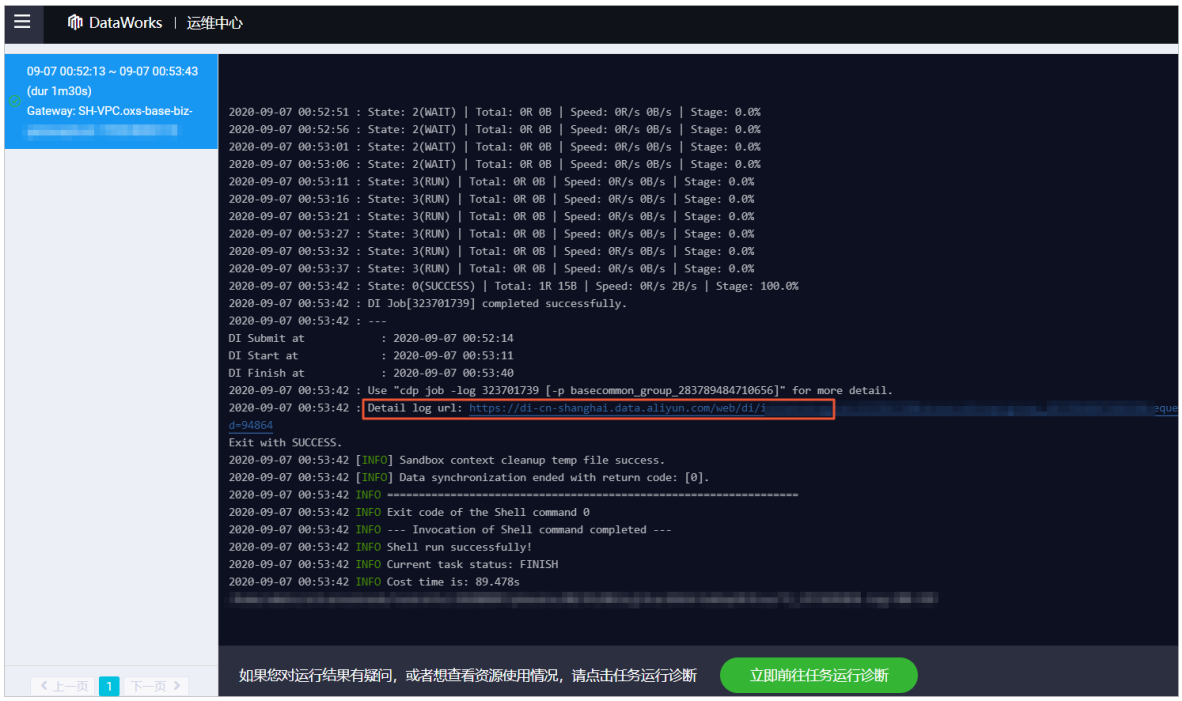
出于性能的考虑和具体数据源读取的限制，同步任务实际运行时的并发度可能小于配置的**任务最大期望并发数**和任务实际运行时的并发度不一致。

查看任务实际运行并发度的操作如下：

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入运维中心**。
4. 在左侧导航栏，单击**周期任务运维 > 周期实例**。
5. 单击相应的数据同步节点，在右侧打开DAG图。右键单击该节点，选择**查看运行日志**。



6. 在节点的运行日志页面，单击Detail log url链接。



7. 在数据同步任务的详情日志页面，查找形式为 `JobContainer - Job set Channel-Number to 2 channels.` 的日志，此处的channels即为任务实际运行的并发度。

```
[ResponseError]:


  AccessDenied
  The bucket you access does not belong to you.
  5F550DA521F3CE34313E361D
  new-dataworks-workshop.oss-cn-shanghai-internal.aliyuncs.com

2020-09-07 00:26:15.462 [job-323683754] INFO  OdpsWriter$Job - blockSizeInMB=64.
2020-09-07 00:26:15.462 [job-323683754] INFO  JobContainer - jobContainer starts to do prepare ...
2020-09-07 00:26:15.463 [job-323683754] INFO  JobContainer - DataX Reader.Job [ossreader] do prepare work .
2020-09-07 00:26:15.491 [job-323683754] INFO  OssReader$Job - add object [user_log.txt] as a candidate to be read.
2020-09-07 00:26:15.493 [job-323683754] INFO  JobContainer - DataX Writer.Job [odpswriter] do prepare work .
2020-09-07 00:26:15.494 [job-323683754] INFO  IdAndKeyUtil - Try to get accessId/accessKey from your config.
2020-09-07 00:26:15.494 [job-323683754] INFO  OdpsWriter$Job - accessId:[[illegible]] .
2020-09-07 00:26:18.045 [job-323683754] INFO  OdpsUtil - Try to truncate [illegible]
2020-09-07 00:26:19.435 [job-323683754] INFO  OdpsUtil - Try to start sql[illegible]
alter table xc_ods_raw_log_d add IF NOT EXISTS partition(dt='20200906');
] .
2020-09-07 00:26:21.860 [job-323683754] INFO  JobContainer - jobContainer starts to do split ...
2020-09-07 00:26:21.860 [job-323683754] INFO  JobContainer - Job set Channel-Number to 2 channels.
2020-09-07 00:26:22.120 [job-323683754] INFO  UnstructuredSplitUtil - File to be read:[{"end":[illegible]36,"filePath":"user_log
2020-09-07 00:26:22.121 [job-323683754] INFO  UnstructuredSplitUtil - File to be read:[{"end":[illegible]
2020-09-07 00:26:22.121 [job-323683754] INFO  JobContainer - DataX Reader.Job [ossreader] splits to [2] tasks.
```

并发度和资源的占用关系

在独享资源组中，占用关系包括并发度和CPU、内存的占用关系：

- 并发度和CPU的占用关系

  在独享资源组中，并发度和CPU的占用关系为1:0.5，即拥有一台4 vCPU 8 GiB规格的ECS机器，其独享资源组的并发额度为8。最多能够同时运行8个并发度为1的离线同步任务，或4个并发度为2的离线同步任务。

  当新提交至独享资源组的任务所需要的并发度大于独享资源组剩余的并发度额度时，新提交的任务将等待独享资源组中正在运行的任务结束，直至剩余的并发度额度满足新提交任务的并发度需求。

  > **说明** 如果新提交任务设置的并发度超过独享资源组的最大并发额度，例如，向一台拥有4 vCPU 8 GiB规格的ECS机器的独享资源组提交一个并发度设置为10的任务，该任务将永远处于等待资源的状态。由于资源组根据任务被提交的先后顺序分配资源，后续提交的任务也将无法运行。

- 并发度和内存的占用关系

  在独享资源组中，单个任务的并发度和内存的占用关系为Min{768+（并发数-1）*256，8029} MB。但是，您可以在任务中通过设置，覆盖其对应关系。如果是脚本模式，请在JSON结构的配置文本中，通过JSON路径*$.setting.jvmOption*进行设置。

```
"setting": {
    "errorLimit": {
        "record": "0"
    },
    "speed": {
        "throttle": false,
        "concurrent": 1
    },
    "jvmOption": "-Xms1024m -Xmx1024m"
}
```

  您需要确保所有正在运行的任务使用内存的总和，比独享资源组中所有机器的内存总量小1 GB以上，任务才能平稳运行。如果未满足该条件，会因为Linux系统的OOM Killer机制强制停止任务的运行。

  > **说明** 如果您未使用脚本模式加大任务的内存，则只需要考虑独享资源组并发度的额度对任务提交的限制。

同步速度

不同数据源的并发读写速度会有很大的差异。下文为您介绍典型数据源在独享资源组中，单并发的同步速度：

- 不同数据源的Writer插件对应的单并发平均速度

Writer	单并发平均速度（KB/s）
AnalyticDB for PostgreSQL	147.8

Writer	单并发平均速度（KB/s）
AnalyticDB for MySQL	181.3
ClickHouse	5259.3
DataHub	45.8
DRDS	93.1
Elasticsearch	74.0
FTP	565.6
GDB	17.1
HBase	2395.0
hbase20xsql	37.8
HDFS	1301.3
Hive	1960.4
HybridDB for MySQL	323.0
HybridDB for PostgreSQL	116.0
Kafka	0.9
LogHub	788.5
MongoDB	51.6
MySQL	54.9
ODPS	660.6
Oracle	66.7
OSS	3718.4
OTS	138.5
PolarDB	45.6
PostgreSQL	168.4
Redis	7846.7
SQLServer	8.3
Stream	116.1
TSDB	2.3
Vertica	272.0

- 不同数据源的Reader插件对应的单并发平均速度

Reader	单并发平均速度（KB/s）
AnalyticDB for PostgreSQL	220.3

Reader	单并发平均速度（KB/s）
AnalyticDB for MySQL	248.6
DRDS	146.4
Elasticsearch	215.8
FTP	279.4
HBase	1605.6
hbase20xsql	465.3
HDFS	2202.9
Hologres	741.0
HybridDB for MySQL	111.3
HybridDB for PostgreSQL	496.9
Kafka	3117.2
LogHub	1014.1
MongoDB	361.3
MySQL	459.5
ODPS	207.2
Oracle	133.5
OSS	665.3
OTS	229.3
OTSStream	661.7
PolarDB	238.2
PostgreSQL	165.6
RDBMS	845.6
SQLServer	143.7
Stream	85.0
Vertica	454.3

4.实时数据同步

4.1. 实时同步能力说明

DataWorks为您提供的实时数据同步功能，方便您使用单表或整库同步方式，将源端数据库中部分或全部表的数据变化实时同步至目标数据库中，实现目标库实时保持和源库的数据对应。

使用限制

- 目前Groovy函数、多路输出处于研发阶段，上线日期待定。
- 实时同步仅支持使用独享数据集成资源组。

功能概述

实时同步支持的能力如下图所示：

功能	描述
10+数据源间的数据同步	实时同步支持10+种数据源星型链路组合，您可以将多种输入及输出数据源搭配组成同步链路进行数据同步。详情请参见实时同步支持的数据源。 **说明** 实时同步任务不支持同步视图。

功能	描述
复杂网络环境下的数据同步	支持云数据库，本地IDC、ECS自建数据库或非阿里云数据库等环境下的数据同步。您可以根据数据库所在网络环境，选择合适的网络解决方案来实现数据源与资源组的网络连通。在配置同步任务前，您需要确保数据集成资源组与您将同步的数据来源端与目标端网络环境已经连通，对应数据库环境与网络连通配置详情请参见：配置资源组与网络连通。 **说明** 离线和实时同步的资源组推荐使用不同资源组，以便任务分开执行。避免混跑带来的资源抢占、运行态互相影响等问题。例如，CPU、内存、网络等互相影响，可能会导致离线任务变慢、实时任务延迟，在资源不足的极端情况下可能还会出现任务被OOM KILLER杀掉等问题。
数据同步场景	离线同步支持单表实时增量数据同步至目标端单表、分库分表实时增量数据同步至目标端单表、整库（多表）实时增量数据同步至目标多表，同步方案介绍请参见：选择同步方案概述。
实时同步任务配置	实时同步任务配置时支持的能力如下，您无需编写代码，直接通过托拽的方式即可进行任务的开发。业务新手也能够轻松上手。 • 数据转换 单表增量同步场景下，您可以对输入的数据源进行**数据过滤**、**字符串替换**和**数据脱敏**处理，再将处理后的数据输出至目标数据库。 **说明** 目前Groovy函数、多路输出处于研发阶段，上线日期待定。 • 自定义目标schema名或表名 实时同步默认将增量数据写入到目标端的同名schema或同名表中，如果schema或表不存在，将默认新建，并支持您自定义目标schema或表名。 • 为目标字段赋值 实时同步默认同名映射，将源端字段写入目标端同名字段中，未映射成功的字段将无法同步。同时，支持您为目标表新增字段并为字段赋值常量或变量。 **说明** 数据集成实时同步在同步MySQL、Oracle、Loghub和PolarDB类型的数据至DataHub或Kafka时，会在同步的目标端添加5个附加列，进行元数据管理、排序、去重等操作。详情请参见实时同步字段格式。 • 定义DDL/DML消息处理策略 来源数据源会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的处理策略。 **说明** 关于实时同步各目标端支持的DML与DDL操作详情请参见：支持的DML及DDL操作。

功能	描述
实时同步任务运维	• 支持对同步任务设置监控报警 支持对业务延迟、Failover、DDL策略、心跳检查设置监控报警。并通过邮件、短信、电话和钉钉等方式将报警信息发送给报警接收人，方便您及时发现并处理任务异常。详情请参见：实时同步任务运行与管理。 支报警疲劳度控制。为了避免短时间内产生大量报警，DataWorks支持您设置当前规则在指定时间间隔内只发送一次报警信息。 • 支持断点续传 支持断点续传或从指定同步起始位置开始同步。即当您重启实时同步任务时，您无需指定位点，任务会自动从失败位点开始读取数据。 说明 如果您需要指定源端开始同步数据的位置，实时同步也支持您指定实时同步任务同步数据的起始位置。 • 支持定义脏数据及对任务的影响 ○ 当不允许脏数据产生时，则同步任务执行过程中如果产生脏数据，任务将失败退出。 ○ 当允许脏数据并设置其阈值时：同步任务将忽略脏数据（即不会写入目标端），并正常执行。 说明 脏数据相关介绍详情请参见：基本概念。

4.2. 实时同步支持的数据源

实时同步支持输入、输出和转换三种类型的插件。

使用限制

- 实时同步不支持在界面直接运行，您需要保存、提交实时同步节点后，在生产环境运行该节点。
- 实时同步仅支持运行在独享数据集成资源组上，详情请参见独享数据集成资源组。
- 实时同步任务不支持同步视图。

单表实时同步支持的数据源

单表实时同步支持输入、输出和转换三种类型的插件，单表实时同步任务配置详情请参见单表增量数据实时同步。

说明 单表实时同步支持如下转换插件，您可以根据业务需求，对输入数据源进行数据过滤、字符串替换和数据脱敏处理后，再将数据输出至目标数据库。

- 配置数据过滤转换
- 配置字符串替换
- 配置数据脱敏

数据源	单表实时读	单表实时写
AnalyticDB for MySQL 3.0	-	配置AnalyticDB MySQL输出
DataHub	DataHub输入	配置DataHub输出
Elasticsearch	-	配置Elasticsearch输出
Hologres	-	Hologres输出
Kafka	Kafka输入	配置Kafka输出

数据源	单表实时读	单表实时写
LogHub（SLS）	LogHub（SLS）输入	-
MaxCompute	-	MaxCompute输出
MySQL	MySQL Binlog输入	-
PolarDB	PolarDB MySQL输入	-

整库实时同步支持的数据源

> **说明** 为避免同步过程由于数据库环境（权限）导致同步失败退出，建议您在配置任务前，先参考数据库环境准备概述，配置好数据库相关环境。

数据源	整库实时读	整库实时写
AnalyticDB for MySQL 3.0	-	整库同步至AnalyticDB MySQL 3.0 **说明** 仅支持同步PolarDB、MySQL、OceanBase数据源至AnalyticDB MySQL。
配置ApsaraDB for OceanBase数据源	OceanBase环境准备	-
DataHub	-	整库实时至DataHub **说明** 仅支持同步PolarDB、OceanBase、MySQL及Oracle数据源至DataHub。
DRDS（polardb-x）	-	-
Hologres	-	整库同步至Hologres **说明** 仅支持同步PolarDB、Oracle、MySQL、SQL Server数据源至Hologres。
Kafka	-	整库同步至Kafka **说明** 仅支持同步MySQL、Oracle和PolarDB数据源至Kafka。
MaxCompute	-	整库同步至MaxCompute **说明** 仅支持同步MySQL、Oracle、OceanBase和PolarDB数据源至MaxCompute。
MongoDB	支持	-
MySQL	MySQL环境准备	-

数据源	整库实时读	整库实时写
Oracle	-	-
PolarDB	PolarDB环境准备	-
PostgreSQL	PostgreSQL环境准备	-

4.3. 同步单表数据

4.3.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。

参数	描述
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

> **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次进入管控后弹出的界面弹框中进行授权。
3. 绑定专有网络VPC。
   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

> **说明** 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。

4. （可选）配置Host。

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

   ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

   说明
   - 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
   - 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

   如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

   i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

   ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将输入数据源与输出数据源的网络、账号权限等准备工作完成。如果您使用的数据源不涉及配置操作，则可以直接进行下一步的添加数据源，准备好需要使用的数据源，以便创建执行后续的实时数据同步任务。

- 配置数据源请参见配置数据源（输入为PolarDB）及配置数据源（输入为MySQL）。
- 添加数据源请参见添加数据源。

4.3.2. 配置数据源（输入为PolarDB）

实时同步单表数据时，当输入数据源为PolarDB时，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

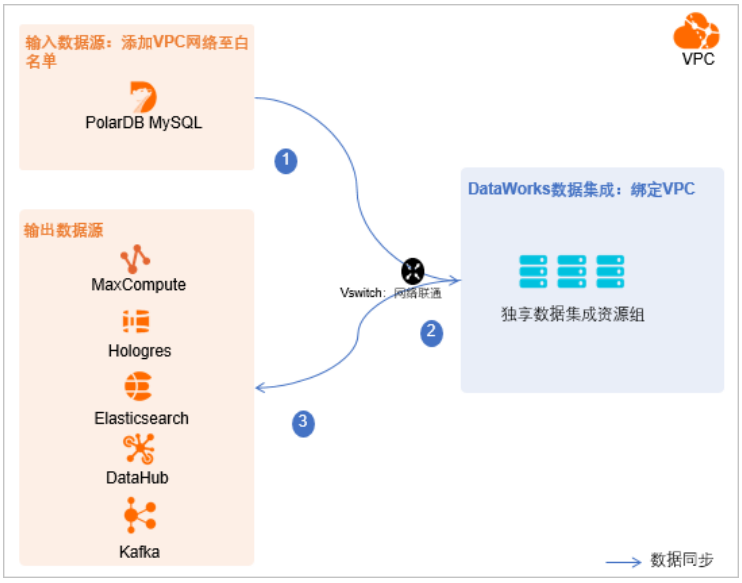
- 数据源准备：已购买输入数据源PolarDB MySQL、输出数据源。输出数据源可以为MaxCompute、Hologres、Elasticsearch、DataHub及Kafka。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将输入数据源的数据同步至输出数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  输入数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

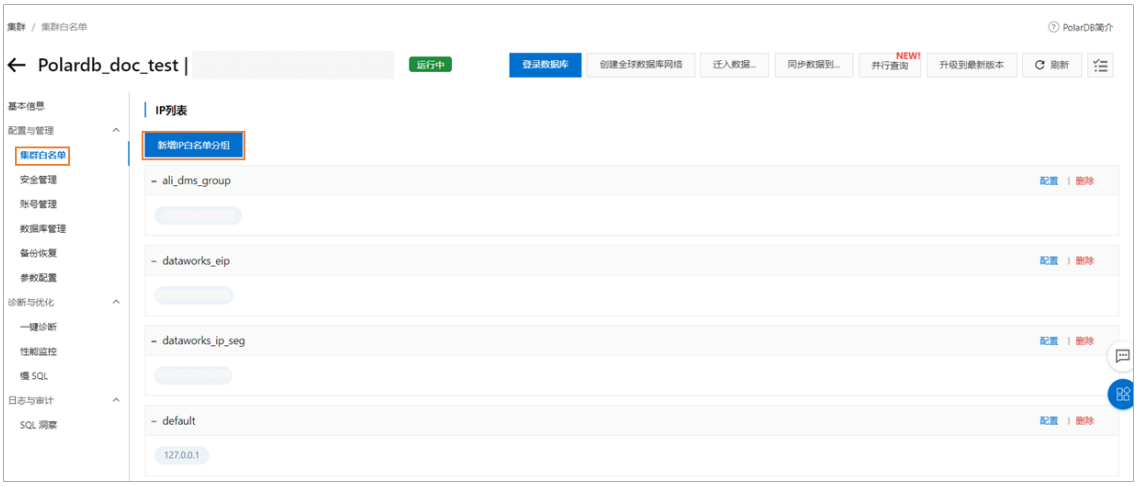
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，输入数据源、资源实例、输出数据源彼此间已可网络联通，且不存在访问限制。您可将输入数据源和输出数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联输入和输出数据源。添加数据源操作可参见添加数据源。

4.3.3. 配置数据源（输入为MySQL）

实时同步单表数据时，当输入数据源为MySQL时，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买输入数据源MySQL、输出数据源。输出数据源可以为MaxCompute、Hologres、Elasticsearch、DataHub及Kafka。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

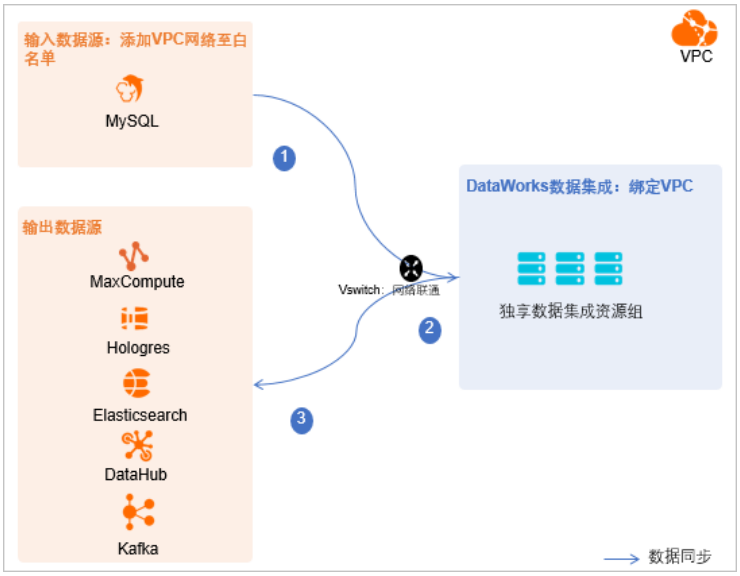
```
select version();
```

> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步输入数据源的数据至输出数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，输入数据源、资源实例、输出数据源彼此间已可网络联通，且不存在访问限制。您可将输入数据源和输出数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联输入和输出数据源。添加数据源操作可参见添加数据源。

4.3.4. 添加数据源

将输入数据源表的数据同步至输出数据源表的过程中，配置数据同步任务前，您需将输入数据源和输出数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行输入和输出的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通输入数据源和输出数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加输入数据源

添加输入数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。实时同步单表数据支持的输入数据源及配置如下：

- 配置MySQL数据源
- 配置DataHub数据源
- 配置LogHub（SLS）数据源
- 配置PolarDB数据源
- 配置SQLServer数据源

添加输出数据源

实时同步单表数据支持的输出数据源及配置如下：

- 配置MaxCompute数据源
- 配置Hologres数据源
- 配置Elasticsearch数据源
- 配置DataHub数据源
- 配置Elasticsearch数据源

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将输入数据源的数据同步至输出数据源中。操作详情可参见通用配置流程。

4.3.5. 通用配置流程

完成网络、资源、输入及输出数据源的准备配置后，您可创建实时同步节点，同步目标输入表数据至输出表。本文为您介绍如何创建单表增量数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

1. 已完成数据源配置。您需要在数据集成同步任务配置前，配置好您需要同步的源端和目标端数据库，以便在同步任务配置过程中，可通过选择数据源名称来控制同步任务的读取和写入数据库。离线同步支持的数据源及其配置详情请参见支持的数据源与读写插件。

   > **说明** 数据源相关能力介绍详情请参见：数据源概述。

2. 已购买合适规格的独享数据集成资源组。详情请参见：新增和使用独享数据集成资源组。
3. 独享数据集成资源组与数据源网络已打通。详情请参见：配置资源组与网络连通。
4. 数据库环境已准备完成：您可以基于您需要进行的同步配置，在同步任务执行前，授予数据源配置的账号在数据库进行相应操作的权限。详情请参见：数据库环境准备概述。

注意事项

单表实时同步仅支持同步单个或多个表数据至目标单表，如果您需要同步数据至多个表，您可以采用以下方案：

- 如果您需要对同步过程中的数据进行过滤、字符串替换或数据脱敏，您可以创建多个单表数据实时同步任务。
- 如果您需要同步多个表至多个表，除了可以创建多个单表实时同步任务，部分数据源还可以选择配置整库实时同步任务。详情请参见：配置整库增量数据实时同步。
- 如果您需要先同步全量数据再将增量数据实时同步至目标端，您可选择同步解决方案。详情请参见：同步解决方案能力说明。

操作流程

1. 创建单表实时同步节点
2. 配置资源组
3. 配置单表实时同步任务
4. 提交并发布实时同步任务
5. 运行并管理实时同步任务

创建单表实时同步节点

1. 创建业务流程。详情请参见：管理业务流程。
2. 创建实时同步节点。
   i. 你可以通过以下两种方式创建实时同步节点。
   - 方式一：展开业务流程，右键单击**数据集成 > 新建节点 > 实时同步**。
   - 方式二：双击业务流程名称，将**数据集成**目录下的**实时同步**节点直接拖拽至右侧业务流程编辑面板。

ii. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，并配置节点存放路径及名称等信息。

配置资源组

实时同步任务仅支持使用独享数据集成资源组，您可以在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。在**资源组**下拉框选择已经与数据库网络连通的独享数据集成资源组。

> **说明** 建议实时同步任务与离线同步任务运行在不同的资源组上，避免混跑带来资源互相抢占、运行态互相影响等问题，例如，CPU、内存、网络等互相影响，导致离线同步任务突然变慢、实时同步任务延迟、以及在资源不足的极端情况下可能会出现任务被OOM Killer杀掉等问题。

配置单表实时同步任务

1. 配置输入数据源。
   i. 在实时同步任务编辑页面左侧的**输入**区域，拖拽目标输入数据源组件至右侧面板。
   ii. 单击输入组件，在右侧的**节点配置**对话框配置相关信息。
   目前，同步单表数据支持的输入数据源类型及其配置如下：
   - 配置MySQL输入
   - 配置DataHub输入
   - 配置LogHub（SLS）输入
   - 配置Kafka输入
   - 配置PolarDB输入
2. （可选）配置数据转换方式。
   在实时同步数据的过程中，如果您希望将输入的数据转换为需要使用的格式进行输出，则可以配置数据转换方式。
   i. 在实时同步任务编辑页面左侧的**转换**区域，拖拽需要使用的数据转换方式组件至右侧面板。
   ii. 单击转换组件，在右侧的**节点配置**对话框配置相关信息。
   目前，同步单表数据支持的转换方式及其配置如下：
   - 配置数据过滤转换
   - 配置字符串替换
   - 配置数据脱敏
3. 配置输出数据源。
   i. 在实时同步任务编辑页面左侧的**输出**区域，拖拽目标输出数据源组件至右侧面板。

ii. 单击输出组件，在右侧的**节点配置**对话框配置相关信息。

目前，同步单表数据支持的输出数据源类型及其配置如下：

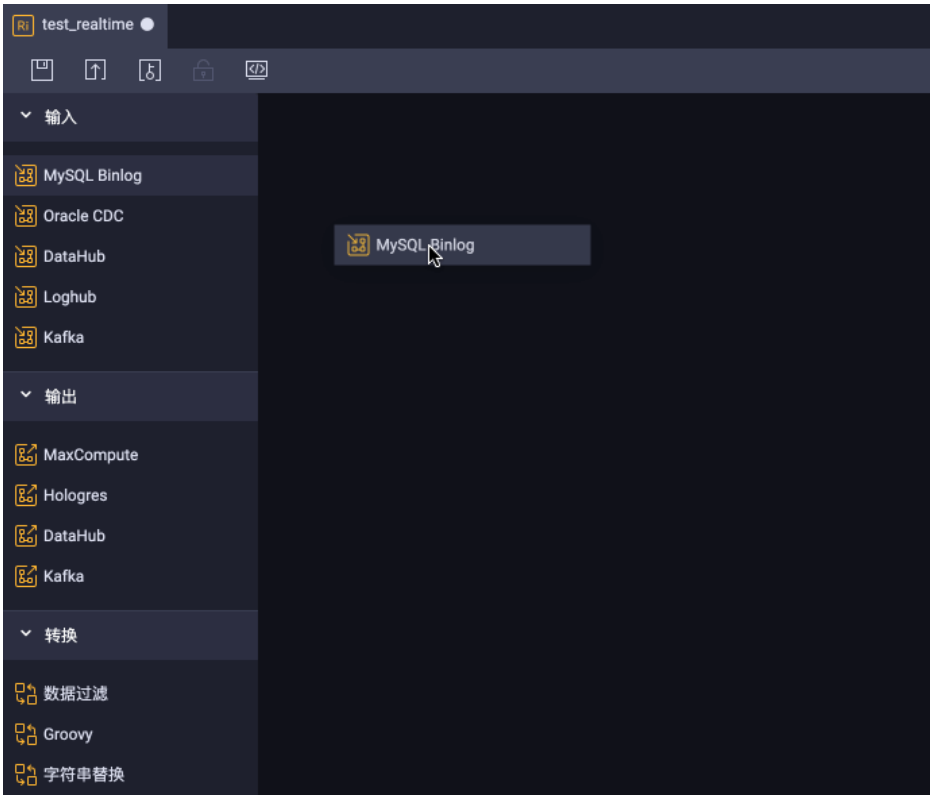
- 配置MaxCompute输出
- 配置Hologres输出
- 配置DataHub输出
- 配置Kafka输出
- 配置Elasticsearch输出

4. 配置输入、输出组件的上下游关系。

添加完输入、输出组件后，您可以根据业务需求，通过连线完成相应的组件关系连接，数据会根据连线从上游组件同步至下游组件。

- 场景一：只进行实时同步，不进行数据转换。

若您不需要进行转换操作，您可以参考下图进行配置。

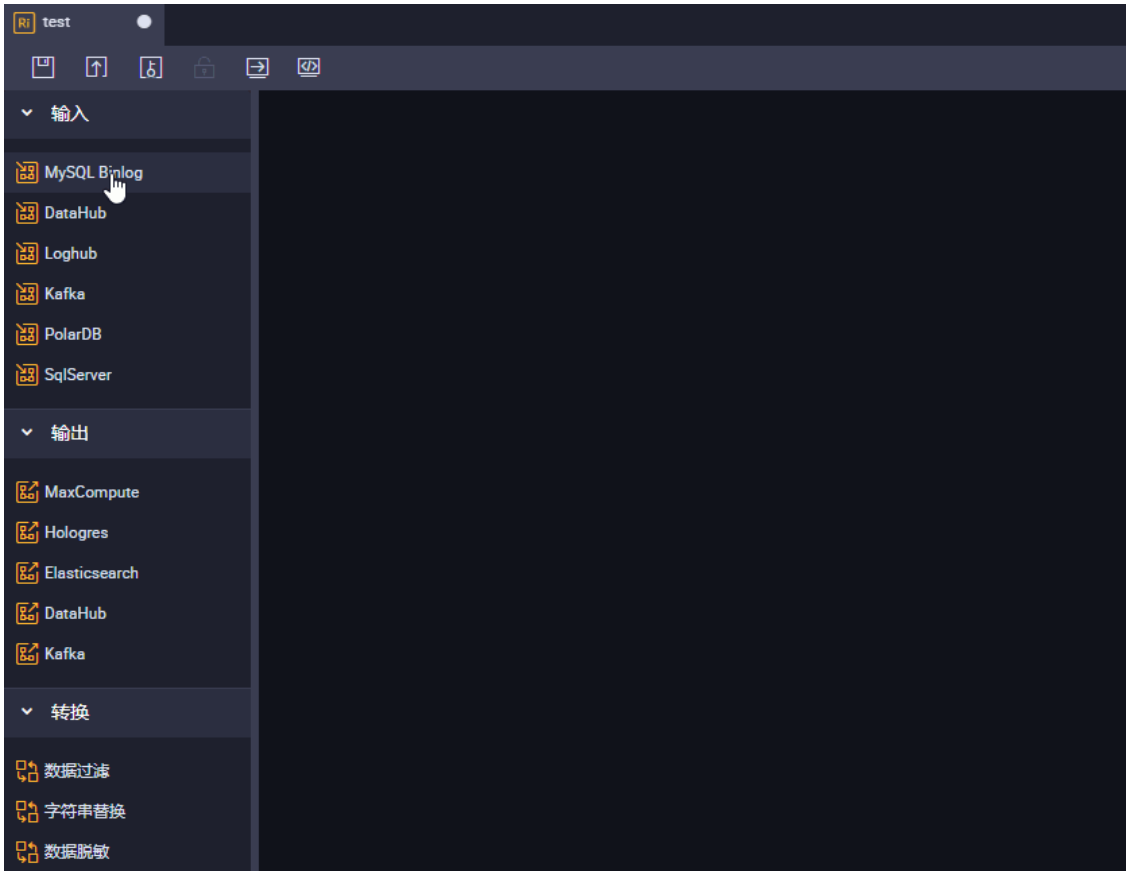


> **说明** 数据同步方向：将上游组件连线至下游组件，数据由输入数据源MySQL Binlog同步至输出数据源MaxCompute。

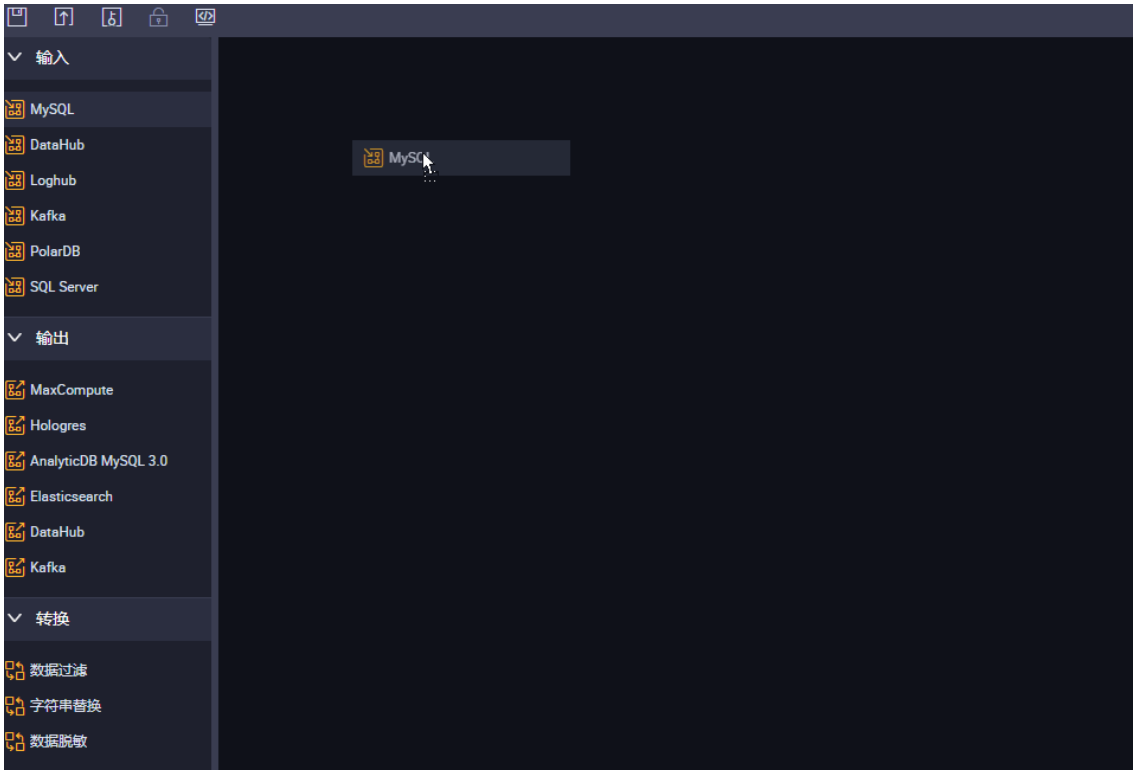
- 场景二：进行数据同步，并且需要在同步过程中进行数据转换。

您可以参考下图，在输入数据源与输出数据源之间拖拽相应的转换节点，并为节点之间设置依赖关系。

- 示例1：对MySQL Binlog数据源的数据进行数据过滤处理后，再同步至下游的输出数据源MaxCompute。



- 示例三：上游的输入数据源MySQL通过数据脱敏后，再同步至下游的输出数据源MaxCompute。



提交并发布实时同步任务

1. 提交并发布节点任务。
   i. 单击工具栏中的图标，保存节点。
   ii. 单击工具栏中的图标，提交节点任务。
   iii. 在**提交新版本**对话框中，输入**变更描述**。
   iv. 单击**确定**。

   如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

运行并管理实时同步任务

任务配置完成后，您可以在**运维中心 > 实时任务运维 > 实时同步任务**面板启动并管理对应任务，详情请参见实时同步任务运行与管理。

4.3.6. 输入

4.3.6.1. 实时同步字段格式

本文为您介绍数据集成实时同步字段的格式及附加列。

数据集成实时同步MySQL或Oracle数据的记录格式如下。

_sequence_id_	_operation_type_	_execute_time_	_before_image_	_after_image_	字段1	字段2	字段3
数据变更位点	操作类型（I/D/U）	数据对应的时间戳	是否是变更前（Y/N）	是否是变更后（Y/N）	真实数据字段1	真实数据字段2	真实数据字段3

数据集成实时同步在同步MySQL、Oracle、Loghub和PolarDB类型的数据至DataHub或Kafka时，会在同步的目标端添加5个附加列，以进行元数据管理、排序去重等操作。

参数	类型	描述
_sequence_id_	STRING	同步记录的位点，由binlog name和offset组成。
_operation_type_	STRING	操作类型，取值分别如下： • I：INSERT操作 • D：DELETE操作 • U：UPDATE操作
_execute_time_	LONG	数据产生时间戳，即binlog时间戳。
_before_image_	STRING	是否更新前的记录，取值为Y或N。
_after_image_	STRING	是否更新后的记录，取值为Y或N。

对于INSERT、UPDATE和DELETE等不同的操作类型，增量数据记录中的_before_image_和_after_image_定义如下：

- 当操作类型为INSERT时，生成的记录为更新后的记录，_before_image_取值为N，_after_image_取值为Y。
- 当操作类型为UPDATE时，数据集成会将其拆分为两条记录。一条是更新前记录，一条是更新后记录。这两条增量数据的_sequence_id_、_operation_type_及_execute_time_对应的值一致。

  第一条增量数据是更新前的值，所以_before_image_取值为Y，_after_image_取值为N。第二条增量数据是更新后的值，所以_before_image_取值为N，_after_image_取值为Y。
- 当操作类型为DELETE时，增量数据中为已经删除的数据，所以_before_image_取值为Y，_after_image_取值为N。

4.3.6.2. 配置MySQL输入

MySQL输入基于Binlog实时订阅的方式，实时读取您配置的MySQL数据库表数据。本文为您介绍，如何配置MySQL输入，以及配置输入之前需要准备的网络环境及账号权限。

前提条件

配置MySQL输入之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：新增数据源之前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，并进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

  详情请参见配置白名单。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

  ```
  select version();
  ```

  > **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。
- 准备账号并授权：

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

  详情请参见创建账号并配置账号权限。
- 开启MySQL：仅实时同步数据时需要开启MySQL。实时同步数据详细介绍请参见实时同步概述。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：
  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

  详情请参见开启MySQL的Binlog。

使用限制

- 数据集成不支持同步MySQL只读库实例的数据。
- DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。
- 仅MySQL Binlog输入数据源支持同步多个表数据至目标单表，并且选择同步的多个输入源表的类型及Schema必须相同。例如，同步的多个表均为MySQL Binlog表。

配置MySQL输入

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至  图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输入 > MySQL**并拖拽至编辑面板。
6. 单击**MySQL**节点，在**节点配置**对话框中，配置各项参数。

参数	描述
数据源	选择已经配置好的MySQL数据源，此处仅支持MySQL数据源。 如果未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面进行新建。
表	选择当前数据源下需要同步的表名称。您可以单击右侧的**数据预览**进行确认。 您可以实现分库分表的场景，配置的库和表会在该任务中同时进行实时同步。 **注意** 分库分表中的数据表的Schema请保持一致，以避免执行报错。
输出字段	选择需要同步的字段，包括**管理字段**和**数据字段**： ○ **管理字段**：为方便进行数据管理、排序和去重等操作，会在同步的目标端自动添加附加字段。 ○ **数据字段**：显示要同步的原始表中对应的字段。 详情请参见实时同步字段格式。

MySQL节点支持分库分表，您可以单击**添加分库分表数据源**，从下拉列表中选择相应的**数据源**和**表**，添加多个数据源，同时进行同步。

> 注意 所选表的Schema需要保持一致，否则执行会报错。

7. 单击工具栏中的图标。

4.3.6.3. 配置DataHub输入

DatahubStream Reader插件通过DataHub SDK实时读取DataHub数据的功能。

背景信息

插件启动后一直运行，等待源端DataHub有数据后进行读取。DatahubStream Reader插件有以下两个功能：

- 实时读取。
- 根据DataHub Shard个数并发读取。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

> 注意 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**输入 > DataHub**并拖拽至编辑面板。
6. 单击**DataHub**节点，在**节点配置**对话框中，配置各项参数。

参数	描述
数据源	选择已经配置好的DataHub数据源，此处仅支持DataHub数据源。 如果未配置数据源，请单击右侧的**新建数据源**，跳转至**工作空间管理 > 数据源管理**页面进行新建。详情请参见配置DataHub数据源。
Topic	选择当前数据源下需要同步的Topic。您可以单击右侧的**数据预览**进行确认。
输出字段	选择需要同步的字段。

7. 单击工具栏中的图标。

4.3.6.4. 配置LogHub（SLS）输入

LogHubStream Reader插件基于LogHub（SLS） SDK实时抽取您配置的LogHub（SLS） topic数据，并支持合并和分裂Shard。合并或分裂Shard后，数据可能会增加，但保证不会丢失数据。

背景信息

LogHub（SLS）实时同步读取插件，会自带以下元数据字段：

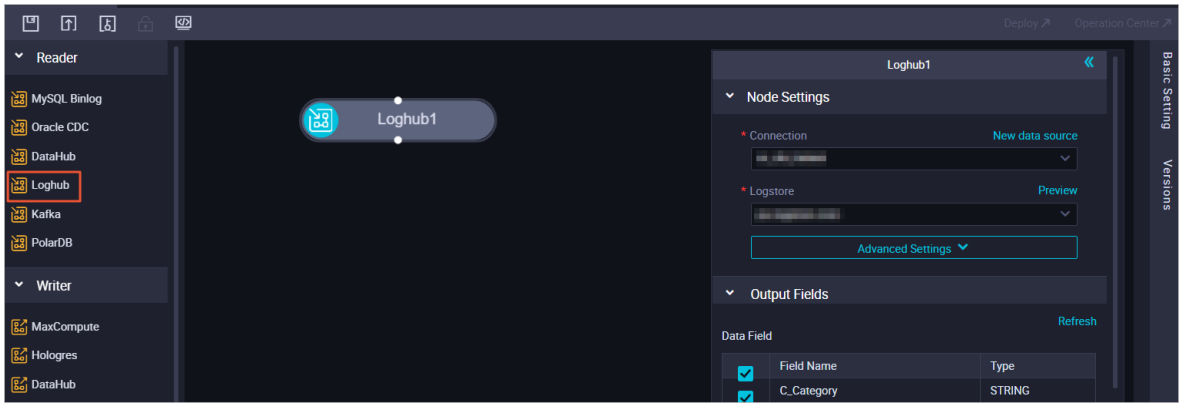
LogHub（SLS）实时同步字段	LogHub（SLS）原始字段
C_Category	category
C_Topic	topic
C_Source	source
C_MachineUUID	machineUUID
C_HostName	__hostname__
C_Path	__path__
C_LogTime	logTime

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**输入 > LogHub**并拖拽至编辑面板。
6. 单击**LogHub**节点，在**节点配置**对话框中，配置各项参数。



参数	描述

参数	描述
数据源	选择已经配置好的LogHub（SLS）数据源，此处仅支持LogHub（SLS）数据源。 如果未配置数据源，请单击右侧的**新建数据源**，跳转至**工作空间管理 > 数据源管理**页面进行新建。详情请参见配置LogHub（SLS）数据源。
Logstore	选择当前数据源下需要同步的Logstore。您可以单击右侧的**数据预览**进行确认。
高级配置	您可以在此设置是否拆分Logstore中的数据。如果选择**拆分**，则需要设置**拆分规则**。
输出字段	选择需要同步的字段。

7. 单击工具栏中的图标。

4.3.6.5. 配置Kafka输入

Kafka插件基于Kafka SDK实时读取Kafka数据。

背景信息

> 说明
> - 支持阿里云Kafka，以及>=0.10.2且<=2.2.x的自建Kafka版本。
> - 对于<0.10.2版本Kafka，由于Kafka不支持检索分区数据offset，且Kafka数据结构可能不支持时间戳，因此会引发同步任务延时统计错乱，造成无法正确重置同步位点。

如果您需要使用SASL鉴权模式，请联系技术支持并提供SSL根证书以及SASL鉴权文件，需要将其配置到同步任务运行环境中，同时您需要在同步任务引用的kafka数据源设置中，加入如下扩展参数：

```
{ "java.security.auth.login.config": "/home/admin/kafka_client_jaas.conf",
 "ssl.truststore.location": "/home/admin/kafka.client.truststore.jks",
 "ssl.truststore.password": "KafkaOnsClient", "security.protocol": "SASL_SSL",
 "sasl.mechanism": "PLAIN", "ssl.endpoint.identification.algorithm": "" }
```

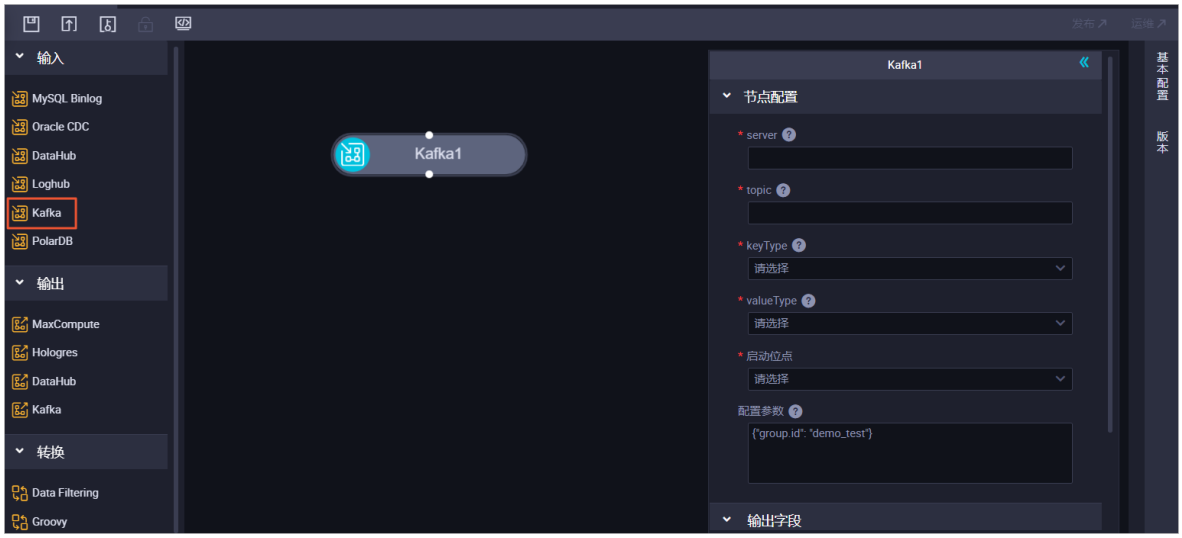
kafka数据源配置详情请参考：配置Kafka数据源。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > 注意 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**输入 > Kafka**并拖拽至编辑面板。
6. 单击**Kafka**节点，在**节点配置**对话框中，配置各项参数。



参数	描述
server	Kafka的Broker Server地址，格式为 `ip:port` 。
topic	Kafka的Topic名称，是Kafka处理资源的消息源的不同分类。 每条发布至Kafka集群的消息都有一个类别，该类别被称为Topic，一个Topic是对一组消息的归纳。 说明 一个Kafka输入仅支持一个Topic。
keyType	Kafka的Key的类型。
valueType	Kafka的Value的类型。
启动位点	控制启动实时同步任务时开始同步数据的起始位点。 ○ EARLIEST：从每个分区的起始位点开始同步。 ○ LATEST：从每个分区任务启动时刻的最新位点开始同步。 ○ TIMESTAMP：根据运维中心启动任务时指定的起始位点开始同步。 ○ GROUP_OFFSETS：根据配置参数中指定的group.id对应的kafka群组位点开始同步。 说明 如果是重启任务，依靠任务缓存位点或者在运维页指定重启位点决定起始位点，该参数无效。
输出模式	定义解析kafka记录的方式 ○ 单行输出：以无结构字符串或者JSON对象解析kafka记录，一个kafka记录解析出一个输出记录。 ○ 多行输出：以JSON数组解析kafka记录，一个JSON数组元素解析出一个输出记录，因而一个kafka记录可能解析出多个输出记录。 说明 目前只在部分地域支持该配置项，如发现无该配置项请耐心等待功能在对应地域发布。
数组所在位置路径	当输出模式设置为多行输出时，指定JSON数组在kafka记录value中的路径，路径支持以 `a.a1` 的格式引用特定JSON对象中的字段或者以 `a[0].a1` 的格式引用特定JSON数组中的字段，如果该配置项为空，则将整个kafka记录value作为一个JSON数组解析。

参数	描述
配置参数	创建Kafka数据生产客户端KafkaProducer可以指定扩展参数，例如bootstrap.servers、auto.commit.interval.ms、session.timeout.ms等，各版本Kafka集群支持的KafkaProducer参数可以参考Kafka官方文档，您可以基于kafkaConfig控制KafkaProducer写入数据的行为。实时同步Kafka输出节点，KafkaProducer的默认acks参数为all，如果对性能有更高要求可以再配置参数中指定acks覆盖默认值。acks取值如下： ○ 0：不进行写入成功确认。 ○ 1：确认主副本写入成功。 ○ all：确认所有副本写入成功。
输出字段	您可以自定义Kafka数据对外输出的字段名： ○ 单击**添加更多字段**，输入**字段名**，并选择**类型**，即可新增自定义字段。 **取值方式**支持从kafka记录中取得字段值的方式，单击右侧⇌按钮可以在两类取值方式间切换。 ■ 预置取值方式：提供6种可选预置从kafka记录中取值的方式： ■ value：消息体 ■ key：消息键 ■ partition：分区号 ■ offset：偏移量 ■ timestamp：消息的毫秒时间戳 ■ headers：消息头

参数	描述
	■ JSON解析取值：可以通过.（获取子字段）和[]（获取数组元素）两种语法，获取复杂JSON格式的内容，同时为了兼容历史逻辑，支持在选择JSON解析取值时使用例如__value__这样以两个下划线开头的字符串获取kafka记录的特定内容作为字段值。 Kafka的数据示例如下。 `{ "a": { "a1": "hello" }, "b": "world", "c":[ "xxxxxxx", "yyyyyyy" ], "d":[ { "AA":"this", "BB":"is_data" }, { "AA":"that", "BB":"is_also_data" } ] }` ■ 不同情况下，输出字段的取值为： ■ 如果同步kafka记录value，取值方式填写__value__。 ■ 如果同步kafka记录key，取值方式填写__key__。 ■ 如果同步kafka记录partition，取值方式填写__partition__。 ■ 如果同步kafka记录offset，取值方式填写__offset__。 ■ 如果同步kafka记录timestamp，取值方式填写__timestamp__。 ■ 如果同步kafka记录headers，取值方式填写__headers__。 ■ 如果同步a1的数据"hello"，取值方式填写a.a1。 ■ 如果同步b的数据"world，取值方式填写b。 ■ 如果同步c的数据"yyyyyyy"，取值方式填写c[1]。 ■ 如果同步AA的数据"this"，取值方式填写d[0].AA。 ○ 鼠标悬停至相应字段，单击显示的图标，即可删除该字段。

7. 单击工具栏中的图标。

> **说明** 一个Kafka输入仅支持一个Topic。

4.3.6.6. 配置PolarDB输入

PolarDB输入插件支持PolarDB MySQL数据库，暂不支持PolarDB PostgreSQL数据库。

前提条件

目前实时同步处于灰度阶段，如果您需要使用相关功能，请提交工单进行开通。

背景信息

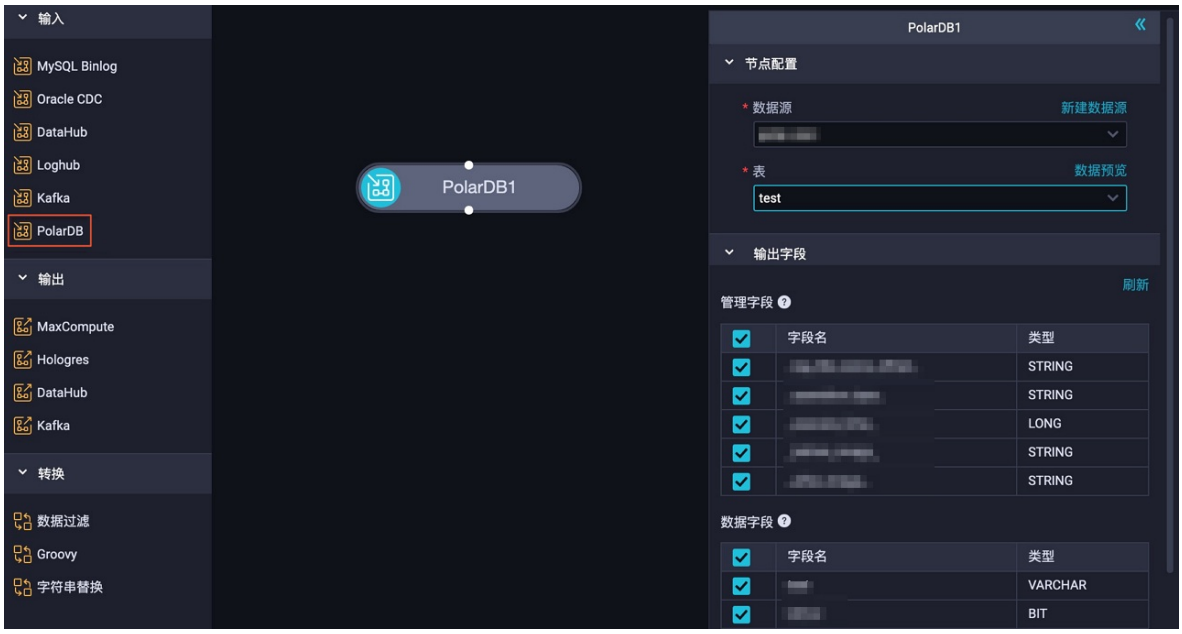
实时同步任务不支持同步视图。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > 注意 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输入 > PolarDB**并拖拽至编辑面板。
6. 单击**PolarDB**节点，在**节点配置**对话框中，配置各项参数。



参数	描述
数据源	选择已经配置好的PolarDB数据源，此处仅支持PolarDB数据源。 如果未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面进行新建。
表	选择当前数据源下需要同步的表名称。您可以单击右侧的**数据预览**进行确认。
输出字段	选择需要同步的字段。

7. 单击工具栏中的图标。

4.3.7. 输出

4.3.7.1. 配置MaxCompute输出

大数据计算服务MaxCompute（原名ODPS）为您提供完善的数据导入方案，能够快速解决海量数据的计算问题。

前提条件

配置MaxCompute输出节点前，您需要先配置好相应的输入或转换数据源，详情请参见实时同步能力说明。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输出 > MaxCompute**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**MaxCompute**节点，在**节点配置**对话框中，配置各项参数。

参数	描述
数据源	选择已经配置好的MaxCompute数据源，此处仅支持MaxCompute数据源。 如果您未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面新建，详情请参见配置MaxCompute数据源。
表	选择当前数据源下需要同步的表名称。 您可以单击右侧的**一键建表**创建新表，也可以单击**数据预览**进行确认。 **注意** 新建目标数据表前，请先连线输入节点，并确认有输出字段。
分区方式	包括**时间自动分区**及**根据字段内容动态分区**。其中**时间自动分区**是根据_*execute_time*_字段进行分区的，详情请参见实时同步字段格式。**根据字段内容动态分区**通过指定源端表某字段与目标MaxCompute表分区字段对应关系，实现源端对应字段所在数据行写入到MaxCompute表对应的分区中。
分区讯息	为您展示MaxCompute分区表的信息。
字段映射	单击**字段映射**，设置源端和目标端字段的映射。同步任务会根据字段的映射关系同步数据。

如果您需要新建表，请单击**一键建表**后，在**新建数据表**对话框中，配置各项参数。

参数	描述
表名称	实时同步写入的MaxCompute表的名称。
生命周期	实时同步写入的MaxCompute表的生命时间长度，详情请参见生命周期。
数据字段结构	实时同步写入的MaxCompute表的字段结构。如果您需要新增字段，请单击**添加**。

参数	描述
分区设置	实时同步写入的MaxCompute表的分区信息。 实时同步写入MCompute表支持**时间自动分区**与**根据字段内容动态分区**两种分区方式 ◦ **时间自动分区**：根据_*execute_time*_字段将数据写入到对应时间分区中，详情请参见实时同步字段格式，  注意 ■ 您最少需要设置二级分区（月和年），最多支持设置五级分区（分钟、小时、天、月和年）。 ■ 关于MaxCompute表的介绍可参考文档：分区 ◦ **根据字段内容动态分区**：通过指定源端表某字段与目标MaxCompute表分区字段对应关系，实现源端对应字段所在数据行写入到MaxCompute表对应的分区中。  例如：配置MaxCompute表分区字段值来源为源端字段A，当A字段值为aa时，实时同步会将数据写入到MaxCompute表对应的aa分区中，当A字段值为bb时，实时同步会将数据写入到MaxCompute表对应的bb分区中。

7. 单击工具栏中的图标。

4.3.7.2. 配置Hologres输出

您可以通过交互式分析Hologres的实时写入能力，构建实时数仓。

前提条件

配置Hologres输出节点前，您需要先配置好相应的输入或转换数据源，详情请参见实时同步支持的数据源。

操作步骤

1. 进入**数据开发**页面。

i. 登录DataWorks控制台。
ii. 在左侧导航栏，单击**工作空间列表**。
iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。

2. 鼠标悬停至**+新建**图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输出 > Hologres**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**Hologres**节点，在**节点配置**对话框中，配置各项参数。

参数	描述
数据源	选择已经配置好的Hologres数据源，此处仅支持Hologres数据源。 如果您未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面进行新建，详情请参见配置Hologres数据源。
表	选择当前数据源下需要同步的数据表名称。 您可以单击右侧的**一键建表**新建数据表，也可以单击**数据预览**进行确认。
动态时间分区	当Hologres表为分区表时，您需要设置动态时间分区。 该动态时间分区会解析来源映射字段的数据值，根据yyyymmdddhhmmss的格式解析。解析后，您可以在目标表中使用变量拼凑为字符串格式的动态分区。来源值不同，最终的目标分区也不同。 例如，来源数据为20200816，目标分区格式为{yyyy}-{mm}-{dd}，则最终写入的目标分区为2020-08-16。
作业类型	包括**重放**和**插入**两种类型： • **重放**表示镜像功能。即源端 `INSERT` 一条记录，Hologres中同样 `INSERT` 一条数据。源端进行 `UPDATE` 或 `DELETE` 操作，Hologres会执行相应的 `UPDATE` 或 `DELETE` 操作。 • **插入**表示将Hologres作为流存储，通过 `INSERT` 保存源端同步过来的数据。
写入冲突策略	包括**覆盖**和**忽略**两种类型： • **覆盖**：使用源端同步过来的新数据覆盖已有的数据。 • **忽略**：忽略源端同步过来的新数据，保留已有的数据。
字段映射	单击**字段映射**，设置源端和目标端字段的映射。同步任务会根据字段的映射关系同步数据。

7. 单击工具栏中的图标。

4.3.7.3. 配置AnalyticDB MySQL输出

您可以通过AnalyticDB MySQL的实时写入能力，构建实时数仓。

前提条件

配置AnalyticDB MySQL输出节点前，您需要先配置好相应的输入或转换数据源，详情请参见实时同步支持的数据源。

操作步骤

1. 进入**数据开发**页面。

i. 登录DataWorks控制台。
ii. 在左侧导航栏，单击**工作空间列表**。
iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。

2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > 注意 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输出 > AnalyticDB MySQL**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**AnalyticDB MySQL**节点，在**节点配置**对话框中，配置各项参数。

参数	描述
数据源	选择已经配置好的AnalyticDB MySQL数据源，此处仅支持AnalyticDB MySQL数据源。 如果您未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面进行新建，详情请参见配置AnalyticDB for MySQL 3.0数据源。
表	选择当前数据源下需要同步的数据表名称。 您可以单击右侧的**一键建表**新建数据表，也可以单击**数据预览**进行确认。
写入模式	当前仅支持配置为**重放（replay）**，即以行为单位更新数据。
字段映射	单击**字段映射**，设置源端和目标端字段的映射。同步任务会根据字段的映射关系同步数据。

7. 单击工具栏中的图标，保存配置。

4.3.7.4. 配置DataHub输出

DataHub是流式数据（Streaming Data）的处理平台，为您提供发布、订阅和分发流式数据的功能，让您可以轻松构建基于流式数据的分析和应用。

前提条件

配置DataHub输出节点前，您需要先配置好相应的输入或转换数据源，实时同步支持的数据源。

背景信息

DataHub Writer通过DataHub服务的Java SDK向DataHub写入数据，使用的日志服务Java SDK版本如下。

```
<dependency>
    <groupId>com.aliyun.datahub</groupId>
    <artifactId>aliyun-sdk-datahub</artifactId>
    <version>2.5.1</version>
</dependency>
```

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**输出 > DataHub**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**DataHub**节点，填写**节点配置**对话框中的参数。

参数	描述
数据源	选择已经配置好的DataHub数据源，此处仅支持DataHub数据源。 如果未配置数据源，请单击右侧的**新建数据源**，跳转至**工作空间管理 > 数据源管理**页面进行新建。详情请参见配置DataHub数据源。
Topic	选择当前数据源下需要同步的Topic。您可以单击右侧的**数据预览**进行确认。
批量条数	支持批量同步的数量。
字段映射	映射源端和目标端的字段，进行同步任务时，会根据字段的映射关系同步数据。

7.

4.3.7.5. 配置Kafka输出

Kafka输出节点只需要选择表，进行字段映射即可完成配置。

前提条件

配置Kafka输出节点前，您需要先配置好相应的输入或转换数据源，实时同步支持的数据源。

操作步骤

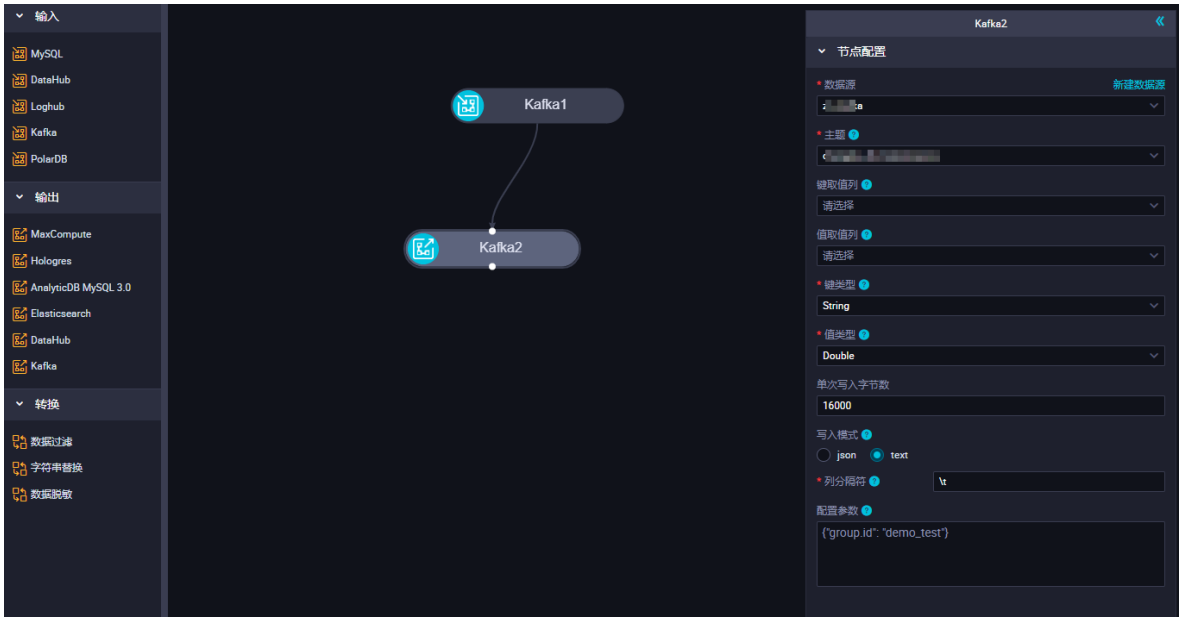
1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至**+新建**图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。

5. 在实时同步节点的编辑页面，鼠标单击**输出 > Kafka**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**Kafka**节点，在**节点配置**对话框中，配置各项参数。



参数	描述
数据源	选择已经配置好的Kafka数据源，此处仅支持Kafka数据源。如果未配置数据源，请单击右侧的**新建数据源**，跳转至**工作空间管理 > 数据源管理** 页面进行新建。详情请参见：配置Kafka数据源。
主题	Kafka的Topic名称，是Kafka处理资源的消息源（feeds of messages）的不同分类。 每条发布至Kafka集群的消息都有一个类别，该类别被称为Topic，一个Topic是对一组消息的归纳。 **说明** 一个Kafka输入仅支持一个Topic。
键取值列	指定哪些源端列的值拼接后作为Kafka记录Key，如果选择多列，使用逗号作为分隔符拼接列值。如果不选，则Key为空。
值取值列	指定哪些源端列的值拼接后作为Kafka记录Value。如果不填写，默认将所有列拼起来作为Value。拼接方式取决于选择的写入模式，详情请参见：Kafka Writer的参数说明。
键类型	Kafka的Key的类型，决定了初始化KafkaProducer时的key.serializer配置，可选值包括STRING、BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。
值类型	Kafka的Value的类型，决定了初始化KafkaProducer时的value.serializer配置，可选值包括STRING、BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。
单次写入字节数	一次写入请求包含的字节数，建议设置大于16000。
写入模式	该配置项决定将源端列拼接作为写入Kafka记录Value的格式，可选值为text和json。 ○ 配置为text，将所有列按照列分隔符进行拼接。 ○ 配置为json，将所有列拼接为JSON字符串。 例如，列配置为col1、col2和col3，源端某记录这三列的值为a、b和c，写入模式配置为text、列分隔符配置为 `#` 时，对应写入Kafka的记录Value为字符串 `a#b#c` ；写入模式配置为json时，写入Kafka的记录Value为字符串 `{"col1":"a","col2":"b","col3":"c"}` 。

参数	描述
列分隔符	当写入模式配置为text，将源端列按照该配置项指定列分隔符拼接作为写入Kafka记录的Value，支持配置单个或者多个字符作为分隔符，支持以 `\u0001` 格式配置unicode字符，支持 `\t` 、 `\n` 等转义字符。默认值为 `\t` 。
配置参数	创建Kafka数据消费客户端KafkaConsumer可以指定扩展参数，例如bootstrap.servers、acks、linger.ms、session.timeout.ms等，您可以基于kafkaConfig控制KafkaProducer消费数据的行为。实时同步Kafka输出节点，KafkaProducer的默认acks参数为all，如果对性能有更高要求可以在配置参数中指定acks覆盖默认值。acks取值如下： ○ 0：不进行写入成功确认。 ○ 1：确认主副本写入成功。 ○ all：确认所有副本写入成功。

7.

4.3.7.6. 配置Elasticsearch输出

您可以通过Elasticsearch的实时写入能力，构建实时数仓。

前提条件

配置Elasticsearch输出节点前，您需要先配置好相应的输入或转换数据源，详情请参见实时同步支持的数据源。

使用限制

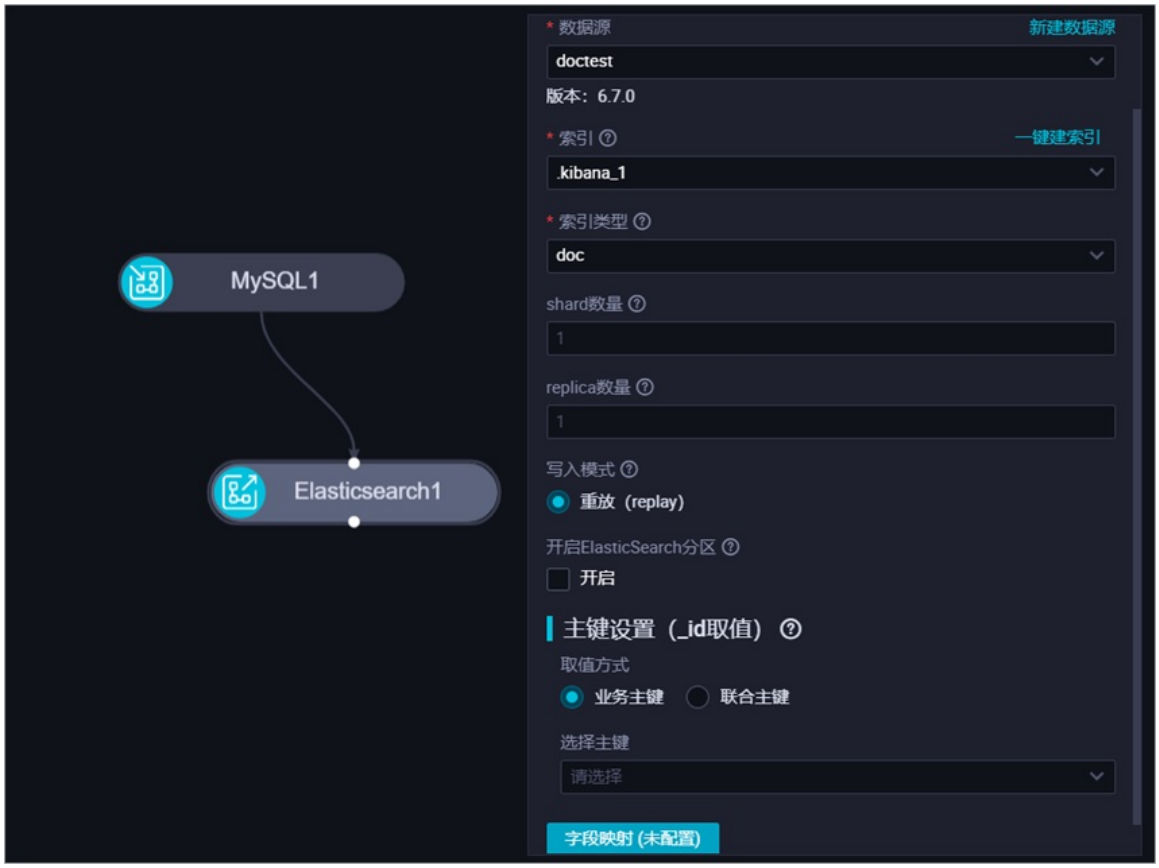
DataWorks平台目前仅支持配置阿里云Elasticsearch5.x、6.x、7.x版本数据源，不支持配置自建Elasticsearch数据源。

操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

   > **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
4. 单击**提交**。
5. 在实时同步节点的编辑页面，单击**输出 > Elasticsearch**并拖拽至编辑面板，连线已配置好的输入或转换节点。
6. 单击**Elasticsearch**节点，在**节点配置**对话框中，配置各项参数。



参数	描述
数据源	选择已经配置好的Elasticsearch数据源，此处仅支持Elasticsearch数据源。 如果您未配置数据源，请单击右侧的**新建数据源**，进入**工作空间管理 > 数据源管理**页面新建，详情请参见配置Elasticsearch数据源。

参数	描述
索引	选择当前数据源下需要同步的索引名称。 您可以单击右侧的一键建索引新建索引，可以直接使用默认生成的索引信息进行新建索引，也可以修改索引名、索引类型、动态参数映射、分片数、副本数以及建索引的语句，然后进行建立索引。 ○ **索引类型**：在Elasticsearch7.x以上版本是没有的，5.x，6.x已经其他更低版本是支持索引类型的，可以自定义配置。 ○ **动态映射状态**：配置的是Elasticsearch索引根目录的参数dynamic的值，dynamic含义是Elasticsearch动态加字段，Elasticsearch做自动类型推导。 ■ 版本低于7.10时，取值包括：true、false、strict。 ■ 7.10以上版本时，取值包括：true、false、strict、runtime。 其中。 ■ true表示可以被存储且被索引到。 ■ false表示可以被存储但是不能被索引到 ■ strict表示新的字段出现，抛异常，不存储 ■ runtime表示新字段加入到运行时字段中，字段不会被索引。 详情可参见dynamic。 ○ **shard数量**：shard数量代表索引分片，Elasticsearch可以把一个完整的索引分成多个分片，这样可以把一个大的索引拆分成多个，分布到不同的节点上，构成分布式搜索。分片的数量只能在索引创建前指定，并且索引创建后不能更改。详情可参见分片（shards）。 ○ **replica数量**：replica数量是shard的数据副本，负责容错，以及承担读请求负载。容量不够、数据不需要备份多份、写入性能不够的时候，replica默认就设成1。 ○ **建索引语句**：properties里面是字段的配置，可以自定义修改字段的类型。
开启Elasticsearch分区	Elasticsearch的routing分区机制是否开启，routing是一个可变值，默认是文档的_id ，也可以设置成一个自定义的值。routing通过hash函数生成一个数字，根据主分片的数量取余最终获得文档所在分片的位置。
主键设置	设置同步时Elasticsearch index上_id的取值方式。 ○ 业务主键：将源头表中一个列作为主键。 ○ 联合主键：将源头表中多个列合并起来作为主键。
字段映射	设置源端和目标端字段的映射。同步任务会根据字段的映射关系同步数据

7. 单击工具栏中的图标。

4.3.8. 转换

4.3.8.1. 配置数据过滤转换

数据过滤插件可以对数据进行规则过滤，例如过滤字段的大小等，符合规则的数据才会被保留。

前提条件

配置数据过滤节点前，您需要先配置好相应的输入节点，详情请参见实时同步支持的数据源。

操作步骤

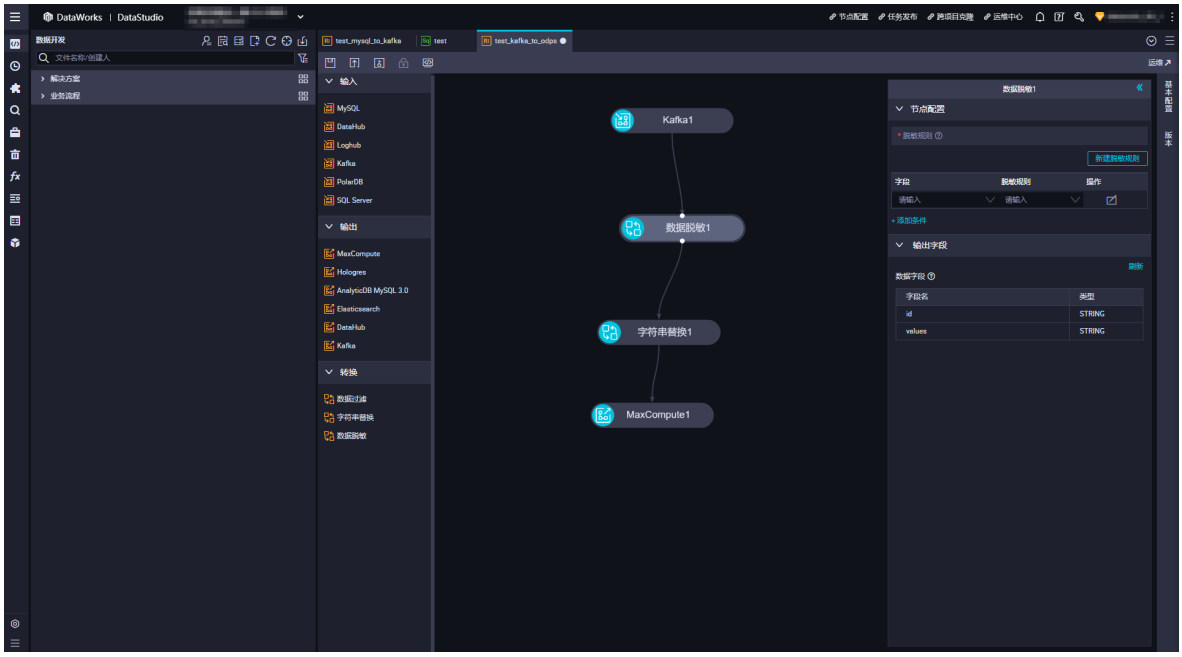
1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

> **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**转换 > 数据过滤**并拖拽至编辑面板，连线已配置好的输入节点。
6. 单击**数据过滤**节点，在**节点配置**对话框中，配置各项参数。



- **节点配置**

  规则：对数据源中的数据进行规则过滤，满足规则的数据才会被保留。

- **输出字段**

  展示最终的输出字段和类型。

7.

4.3.8.2. 配置字符串替换

您可以通过字符串替换节点替换字符串类型的字段。

前提条件

配置字符串替换节点前，您需要先配置好相应的输入节点，详情请参见实时同步支持的数据源。

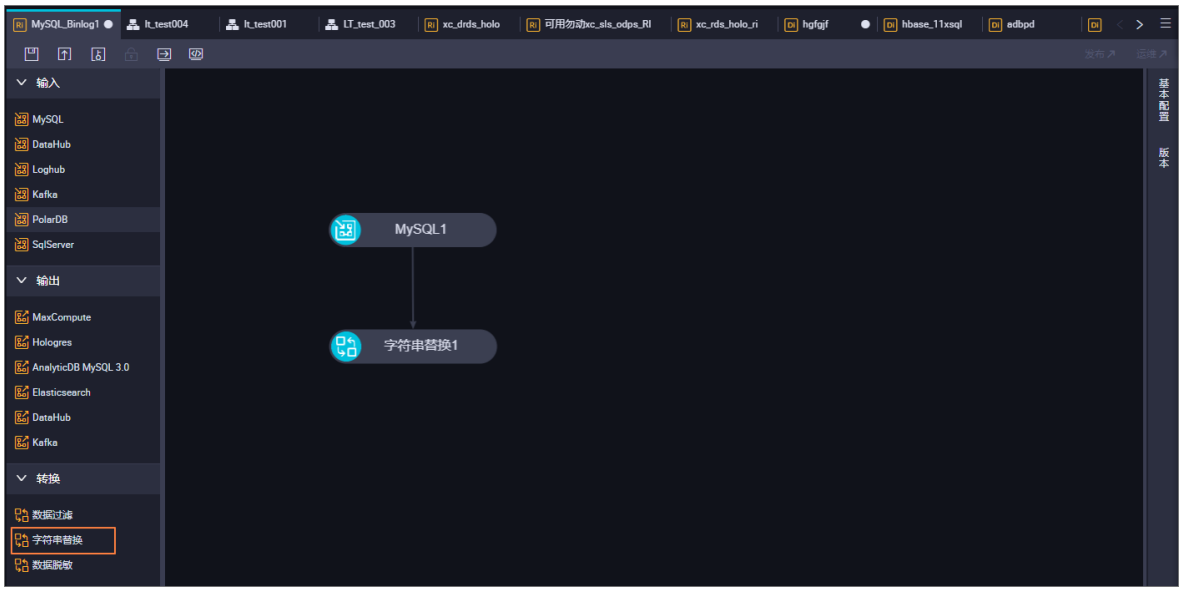
操作步骤

1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至**+新建**图标，单击**数据集成 > 实时同步**。

   您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。
3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

> **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**转换 > 字符串替换**并拖拽至编辑面板，连线已配置好的输入节点。
6. 单击**字符串替换**节点，在**节点配置**对话框中，配置各项参数。



参数	描述
规则	规则包括以下参数： ○ **字段**：选择前一个步骤的输入字段。 ○ **正则匹配**：是否用于正则表达式匹配。 ○ **原字符串**：查找的原字符串。 ○ **新字符串**：替换后的新字符串。 ○ **大小写敏感**：原字符串查找是否大小写敏感。
添加条件	您可以添加多个字符串替换规则。
输出字段	替换后的最终输出字段。

7. 单击工具栏中的图标。

4.3.8.3. 配置数据脱敏

数据脱敏可以对实时同步的单表数据进行脱敏，然后存储到指定的数据库位置。

前提条件

配置数据脱敏节点前，您需要先配置好相应的输入节点，详情请参见实时同步支持的数据源。

操作步骤

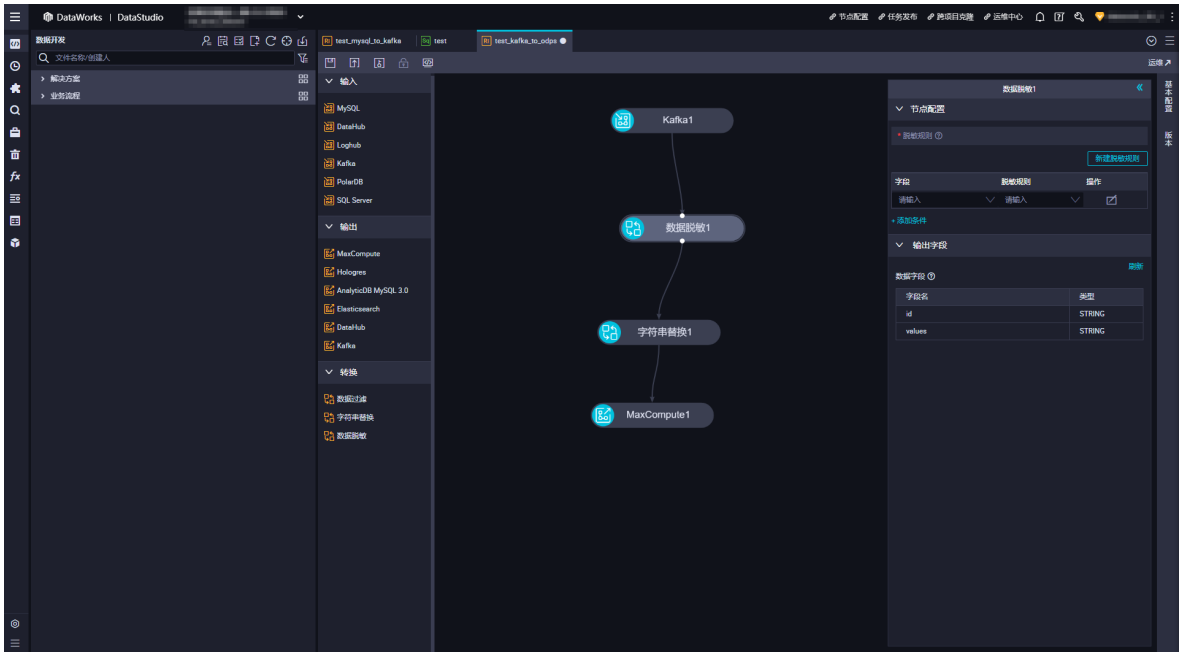
1. 进入**数据开发**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
2. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

您也可以展开目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

3. 在**新建节点**对话框中，选择同步方式为**单表（Topic）到单表（Topic）ETL**，输入**节点名称**，并选择**目标文件夹**。

> **注意** 节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。

4. 单击**提交**。
5. 在实时同步节点的编辑页面，鼠标单击**转换 > 数据脱敏**并拖拽至编辑面板，连线已配置好的输入节点。
6. 单击**数据脱敏**节点，在**数据脱敏**对话框中，配置各项参数。



i. 新建**脱敏规则**：单击**新建脱敏规则**，在弹出来的**新建脱敏规则**对话框中选择需要设置的**敏感数据类型**、**脱敏规则名称**、**脱敏方式**、**安全域**和**替换字符集**。

a. 新建脱敏规则

a. 配置基础信息

参数	描述
敏感数据类型	■ 默认为**选择已有**，右侧下拉框选择已创建的敏感数据类型（包括内置和所有用户创建的敏感数据类型）。 ■ 可切换**新增类型**，右侧输入框可输入敏感数据类型名称（输入字符限制1~30字，包括：中文、英文、数字）。 用户输入**新增敏感数据类型**，系统会判断文字与已有敏感数据类型名称是否相同（包括：内置和该租户下所有用户配置的敏感数据类型），如果名称相同则提示**敏感字段类型重复**。 **说明** 内置敏感数据类型：手机号、身份证号、银行卡号、邮箱_内置、IP、车牌号、 邮政编码、座机号、MAC地址、地址、姓名、公司名、民族、星座、性别、国籍。
脱敏规则名称	该文本框会自动代入用户填写的**敏感数据类型**（输入字符限制1~30字，包括：中文、英文、数字），您也可以修改名称，当与该租户下所有用户创建的脱敏规则出现重名时，提示**规则名称重复**。

b. 配置脱敏方式与规则：DataWorks支持的**脱敏方式**包括**假名**、**HASH**和**掩盖**三种方式。

■ **假名**

假名脱敏会将一个值替换成一个具有相同特征的脱敏信息。脱敏后数据和脱敏前数据的格式保持一致。

■ 当选择的**敏感数据类型**为内置敏感数据类型（手机号、身份证号、银行卡号、邮箱_内置、IP、车牌号、邮政编码、座机号、MAC地址、地址、姓名、公司名）时，用户需要配置**安全域**。

安全域：可选范围0~9，不同安全域的脱敏策略规则不一致，即相同的待脱敏数据在不同的安全域脱敏出来的结果信息不一致。例如，原始数据为a123，安全域设置为0时，脱敏成b124，安全域设置为1时，脱敏成c234。原始数据相同时，如果安全域相同则脱敏后的数据也是相同的。

■ 当选择的**敏感数据类型**为非内置时，用户需要配置**替换字符集**。

替换字符集：遇到字符集中的字符，即会被替换为其他相同类型的字符，不支持中文，若需要脱敏的数据不符合字符集范围则不脱敏（可输入大写字母、小写字母、和数字，多个字符请用英文逗号隔开），例如，敏感数据脱敏前是0~3的数字和a~d的字母组成，那么脱敏后也会脱敏成在这个范围内的数字和字母。

■ **哈希**

可将原始数据加密成固定长度的数据。HASH脱敏方式需要选择**安全域**。

安全域：可选范围0~9，不同安全域的脱敏策略规则不一致，即相同的待脱敏数据在不同的安全域脱敏出来的结果信息不一致。例如，原始数据为a123，安全域设置为0时，脱敏成b124，安全域设置为1时，脱敏成c234。原始数据相同时，如果安全域相同则脱敏后的数据也是相同的。

■ **掩盖**

掩盖脱敏是对部分信息进行掩盖，将对应位置上的字符用“*”替换，达到脱敏的效果。

■ **推荐方式**：下拉框可选择**只展示前1位和最后1位**（默认选中）、**只展示前3位和最后2位**、**只展示前3位和最后4位**。

■ **自定义**：自定义设置提供了更加灵活的设置方式，可以在前中后三段设置是否脱敏，以及需要脱敏（或者不脱敏）的字符长度。最多可添加10个分段，至少要有1个分段是**剩余位数**。

图标	描述
①	可选择**位数**、**剩余位数**。
②	输入范围为【1，100】。
③	可选择**脱敏**、**不脱敏**。

例如，脱敏前3位，剩余位数不脱敏。

例如，脱敏后3位，剩余位数不脱敏。

c. 验证脱敏配置结果：您可以在**样本数据**文本框中输入脱敏前样本数据（输入字符限制0~100字符）单击**脱敏验证**，在**脱敏效果**中会返回脱敏后的数据。

b. 单击**确定**，可以在脱敏规则下拉框中选择该脱敏规则，同时新建的脱敏规则会同步到数据保护伞脱敏规则页面。

ii. 单击**添加条件**可新增一行配置数据字段的脱敏规则。

- 在**字段**列下拉框中选择数据脱敏节点的上个节点的输出字段。
- 在**脱敏规则**列下拉框中为字段选择在**数据保护伞>数据脱敏配置**列表中所有已生效的脱敏规则。
- 在**操作**列单击**编辑**。
  - 如果是当前用户创建的脱敏规则，在实时同步任务未提交前，可以单击**编辑**在弹出的**编辑脱敏规则**窗口修改脱敏规则，并支持输入**样本数据**进行**脱敏验证**。
  - 如果是非当前用户创建的脱敏规则，单击**编辑**可以查看脱敏规则配置详情，并支持输入**样本数据**进行**脱敏验证**。
- 在**操作**列单击**删除**可以删除一行字段。

iii. **输出字段**：展示要同步的原始表中对应字段和类型。

4.4. 同步整库数据至MaxCompute

4.4.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表** > **独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

   > **说明** 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的实时数据同步任务。数据源的配置可参见配置数据源（来源为PolarDB）、配置数据源（来源为Oracle）及配置数据源（来源为MySQL）。

4.4.2. 配置数据源（来源为PolarDB）

实时同步PolarDB的数据至MaxCompute时，来源数据源为PolarDB，去向数据源为MaxCompute，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

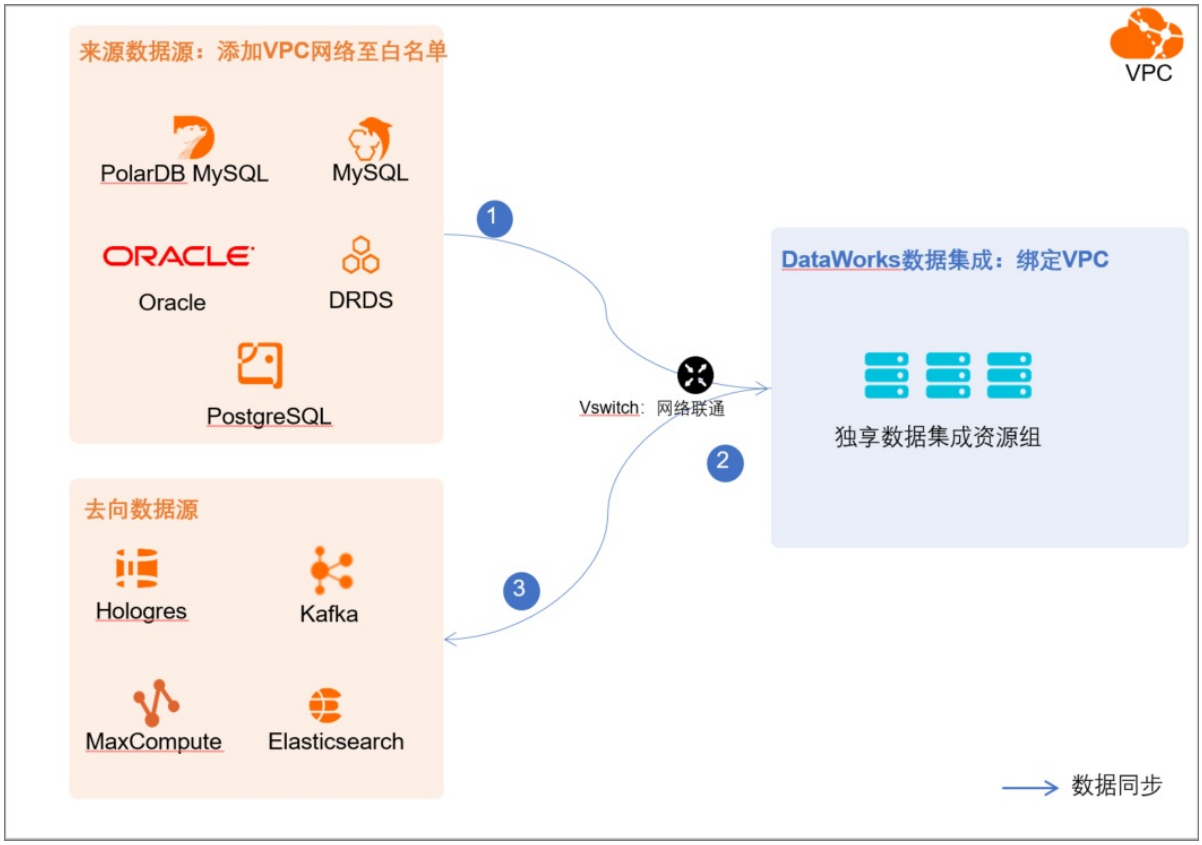
- 数据源准备：已购买来源数据源PolarDB MySQL和去向数据源MaxCompute。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

配置来源数据源：PolarDB

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

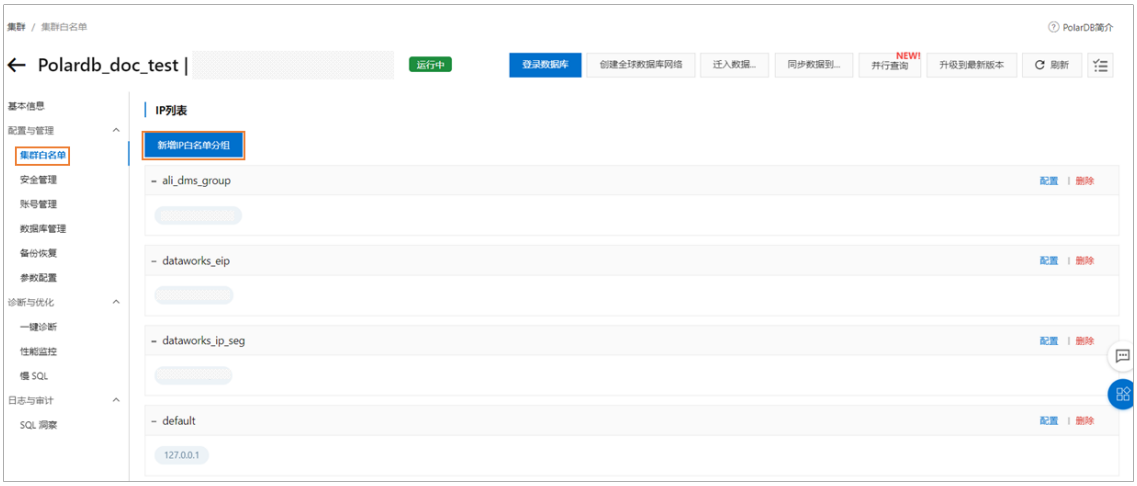
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner 账号登录客户端。

操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

使用Project Owner账号在客户端执行以下命令。

```
setproject odps.sql.acid.table.enable=true;
```

3. （可选）开启使用数据2.0。

如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

```
setproject odps.sql.type.system.odps2=true;
```

4. 创建账号。

此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.4.3. 配置数据源（来源为Oracle）

实时同步Oracle的数据至MaxCompute时，来源数据源为Oracle，去向数据源为MaxCompute，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

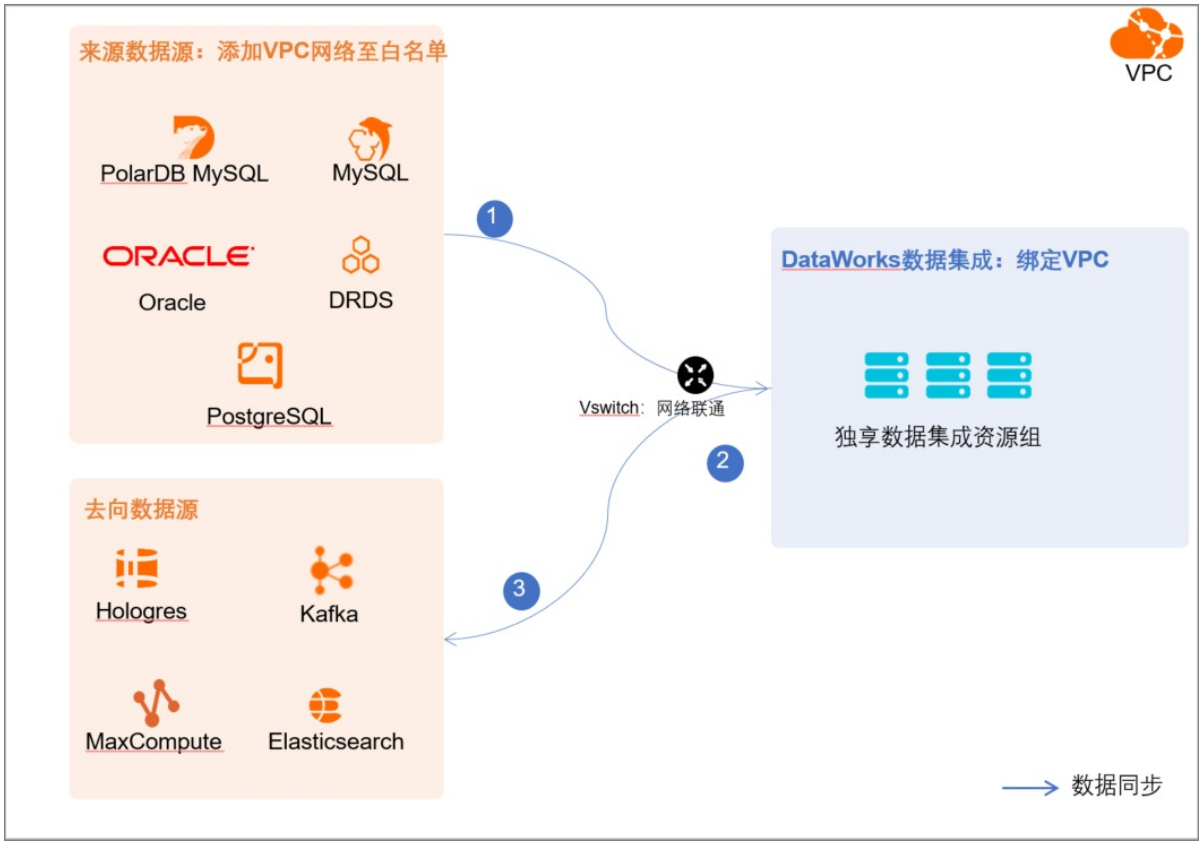
- 准备数据源：已购买来源数据源Oracle、去向数据源MaxCompute。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。

- Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

配置来源数据源：Oracle

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换`'同步账号'`为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner账号登录客户端。

操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

使用Project Owner账号在客户端执行以下命令。

```
setproject odps.sql.acid.table.enable=true;
```

3. （可选）开启使用数据2.0。

如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

```
setproject odps.sql.type.system.odps2=true;
```

4. 创建账号。

此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.4.4. 配置数据源（来源为MySQL）

实时同步MySQL的数据至MaxCompute时，来源数据源为MySQL，去向数据源为MaxCompute，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源MaxCompute。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

  ```
  select version();
  ```

  > **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

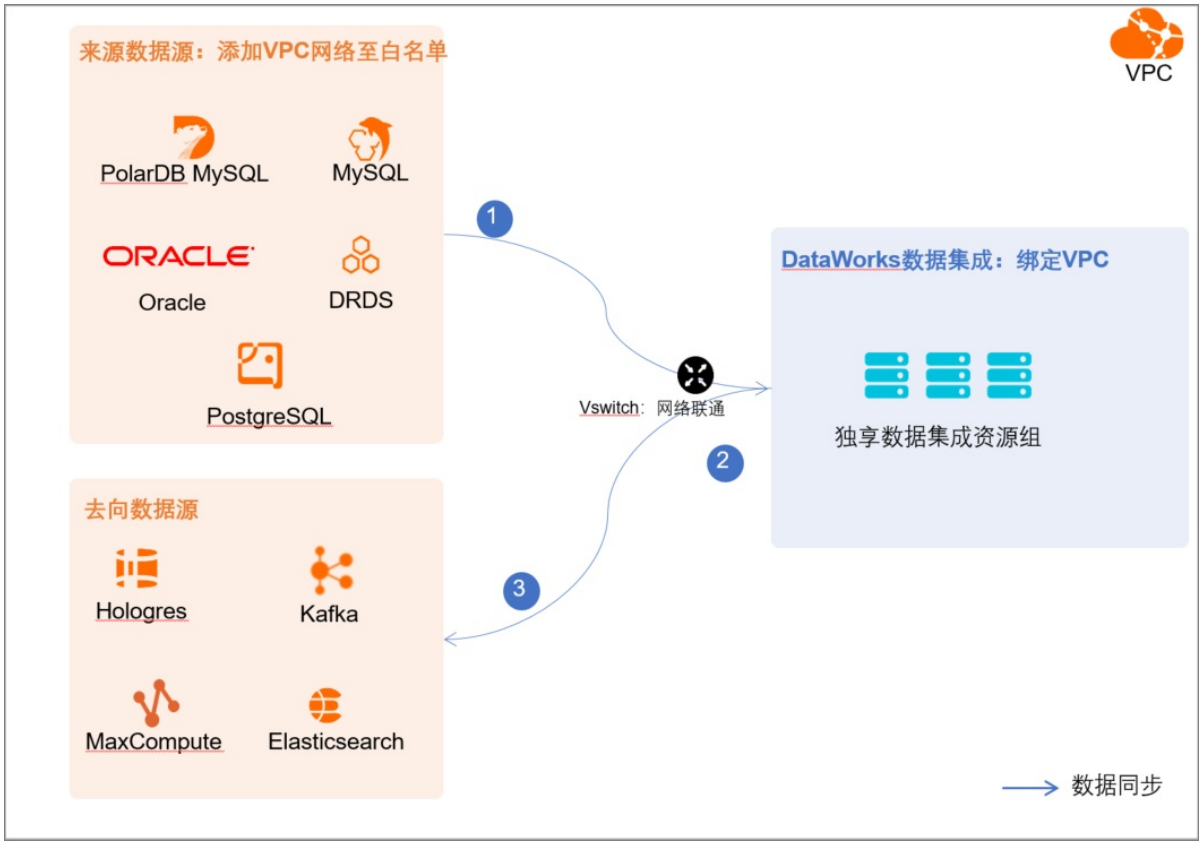
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

配置来源数据源：MySQL

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner 账号登录客户端。

   操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

   使用Project Owner账号在客户端执行以下命令。

   ```
   setproject odps.sql.acid.table.enable=true;
   ```

3. （可选）开启使用数据2.0。

   如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

   ```
   setproject odps.sql.type.system.odps2=true;
   ```

4. 创建账号。

   此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

   创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.4.5. 添加数据源

将来源数据源的数据同步至MaxCompute数据源的过程中，配置实时同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：Oracle

添加Oracle数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置Oracle数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加去向数据源：MaxCompute

操作详情可参见配置MaxCompute数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置并管理实时同步任务。

4.4.6. 配置并管理实时同步任务

完成数据源、网络、资源的准备配置后，您可创建实时同步节点，同步数据至MaxCompute。本文为您介绍如何创建数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

创建实时数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

使用限制

- 实时数据同步任务仅支持使用独享数据集成资源组。
- 实时同步节点目前仅支持同步PolarDB、Oracle、MySQL数据源至MaxCompute。
- 实时数据同步任务暂不支持同步没有主键的表。

创建实时同步任务

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 创建**业务流程**。

   如果您已有**业务流程**，则可以忽略该步骤。

   i. 鼠标悬停至+新建图标，选择**新建业务流程**。

   ii. 在**新建业务流程**对话框，输入**业务名称**。

   iii. 单击**新建**。
5. 创建实时同步节点。

   i. 鼠标悬停至+新建图标，选择**数据集成 > 实时同步**。

   您也可以找到目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

ii. 在**新建节点**对话框中，配置各项参数。

参数	描述
节点类型	默认为**实时同步**。
同步方式	选择**数据库迁至MaxCompute**，用于迁移目标数据库下的部分或所有表至MaxCompute中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放实时同步节点的目录。

iii. 单击**提交**，进入实时同步任务编辑页面。

6. 选择资源组。

i. 在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。

ii. 在**资源组**下拉框，选择需要使用的资源组。

> **说明**
>
> 实时数据同步任务仅支持使用独享数据集成资源组。
>
> 如果您没有可用的独享数据集成资源组，请单击**新建独享资源组**创建。详情请参见独享数据集成资源组概述。

7. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、Oracle和PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

iv. 单击**下一步**。

8. 选择目标数据源并配置目标表格式。

i. 在**设置目标表**页面，选择**目标MaxCompute（ODPS）数据源**和**写入模式**。

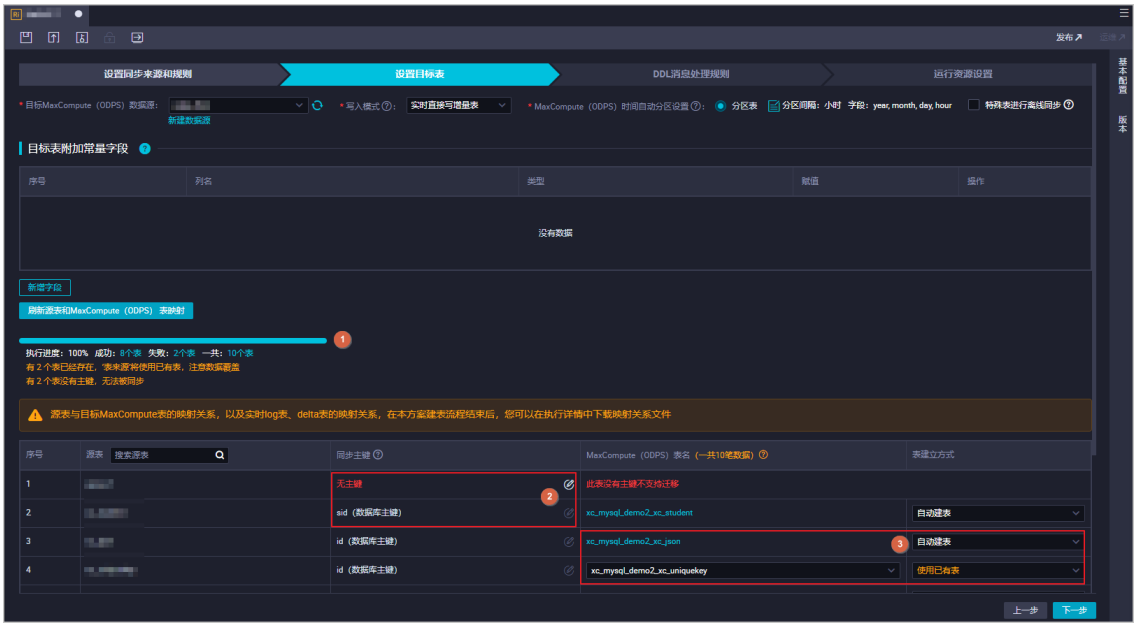
ii. 单击**MaxCompute（ODPS）时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置（支持天级别分区）。此处可以选择写入MaxCompute分区表或者是非分区表。

iii. （可选）目标表新增字段。

如果您希望为目标表中的所有同步表新增统一的字段，则可以在**目标表附加常量字段**区域，单击**新增字段**添加。

iv. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

v. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。 **说明** 实时数据同步任务暂不支持同步没有主键的表。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。 **说明** 如果源表为无主键表，您可以无主键字样后的编辑入口，为源表手动指定主键，以便后续进行增全量同步。

vi. 单击**下一步**

如果您前一步中目标数据源使用的**表建立方式**为**自动建表**，则需要在弹出的**自动建表**对话框，单击**开始建表**，批量创建目标MaxCompute表。

9. DDL消息处理规则设置。

   来源数据源会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的规则。

   > **说明** 该规则为初次执行实时同步任务时的DDL消息处理规则，后续如果您需要修改规则，则可以进入实时任务的运维配置页面修改，详情请参见管理实时同步任务。

   i. 在**DDL消息处理规则**页签，配置**实时同步DDL消息处理策略**。

   不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

   ii. 单击**下一步**。

10. 运行资源设置。

    i. 在**运行资源设置**页面，配置各项参数。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为15。
目标端写入并发数	数据同步任务内，可以从来源表并行读取或写入数据至目标端的最大线程数。最大并发数为32。请根据您的资源组大小和目标端实际规模合理设置。

    ii. 单击**完成配置**。

提交并发布实时同步任务

1. 提交并发布节点任务。

   i. 单击工具栏中的图标，保存节点。

   ii. 单击工具栏中的图标，提交节点任务。

iii. 在**提交新版本**对话框中，输入**变更描述**。

iv. 单击**确定**。

如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

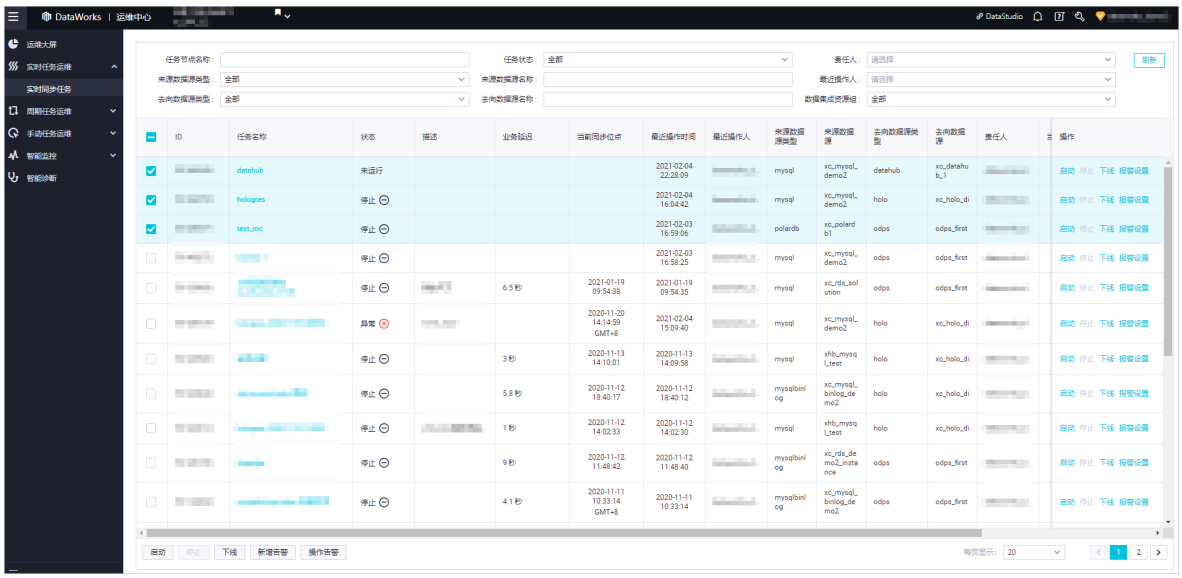
执行实时同步任务

1. 进入运维页面。

   提交或发布节点成功后，单击节点编辑页面右上方的**运维中心**，进入**实时任务运维 > 实时同步任务**页面。

2. 查看实时同步任务详情。

   在**实时同步任务**页面，单击相应**任务名称**，查看运维任务的详细信息。



3. 执行实时同步任务。

   i. 单击目标实时同步任务**操作**列的**启动**。

   ii. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点任务的日期和时间。
时区	从**时区**下拉列表中选择执行任务的时区。
Failover	您可以设置在固定时间内，任务的Failover超过指定次数时，自动结束任务。 **说明** 如果您不配置Failover的次数，将根据5分钟Failover超过100次来自动结束任务，避免频繁启动任务占用系统资源。
脏数据策略	■ **零容忍，不允许**：只要同步任务中包含脏数据，则任务自动结束。 ■ **不限制**：无论同步任务中是否包含脏数据，任务均可正常执行。 ■ **有限控制**：指定可包含固定数值的脏数据，超出该数值时任务自动结束。
实时同步DDL消息处理策略	您可以根据需求修改已配置的DDL消息处理策略。详情请参见配置DDL消息处理策略。

管理实时同步任务

- 停止运行中的任务。

  单击相应任务后的**停止**。在停止对话框中，单击**停止**。
- 下线非运行状态的任务。

  单击相应任务后的**下线**。在下线对话框中，单击**下线**。
- 查看任务的报警信息。

  单击相应任务后的**报警设置**，在**报警设置**页面查看**报警事件**及**报警规则**。

- 为任务新增告警。
  i. 选中需要新增告警的任务，单击**实时同步任务**页面下方的**新增告警**。
  ii. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称。
描述	新建规则的描述信息。
指标	产生报警的指标项： ■ **任务状态** ■ **业务延迟** ■ **Failover** ■ **脏数据** ■ **DDL错误**
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。
报警间隔	设置报警的时间间隔，默认为5分钟发一次报警。
WARNING	产生相应报警时，可以选择通过**邮件**、**短信**和**钉钉**发送报警通知。 **说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
CRITICAL	
接收人（非钉钉）	选择报警通知的接收人。

  iii. 单击**确定**。
- 批量修改目标任务中指定类型的所有告警。
  i. 选中需要操作告警的任务，单击**实时同步任务**页面下方的**操作告警**。
  ii. 在**操作告警**对话框中，选则需要修改的**操作类型**和**告警指标**。
  iii. 单击**确定**。

4.5. 同步整库数据至Hologres

4.5.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。
- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。

2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次次进入管控后弹出的界面弹框中进行授权。

3. 绑定专有网络VPC。

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

说明 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。

4. （可选）配置Host。

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

   如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   > **说明** 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

   i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

   ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的实时数据同步任务。目前同步数据至Hologres的来源数据源仅支持PolarDB、Oracle、MySQL，您可以根据实际需求选择合适的数据源。数据源的配置可参见配置数据源（来源为PolarDB）、配置数据源（来源为Oracle）、配置数据源（来源为MySQL）。

4.5.2. 配置数据源（来源为PolarDB）

实时同步PolarDB的数据至Hologres时，来源数据源为PolarDB，去向数据源为Hologres，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

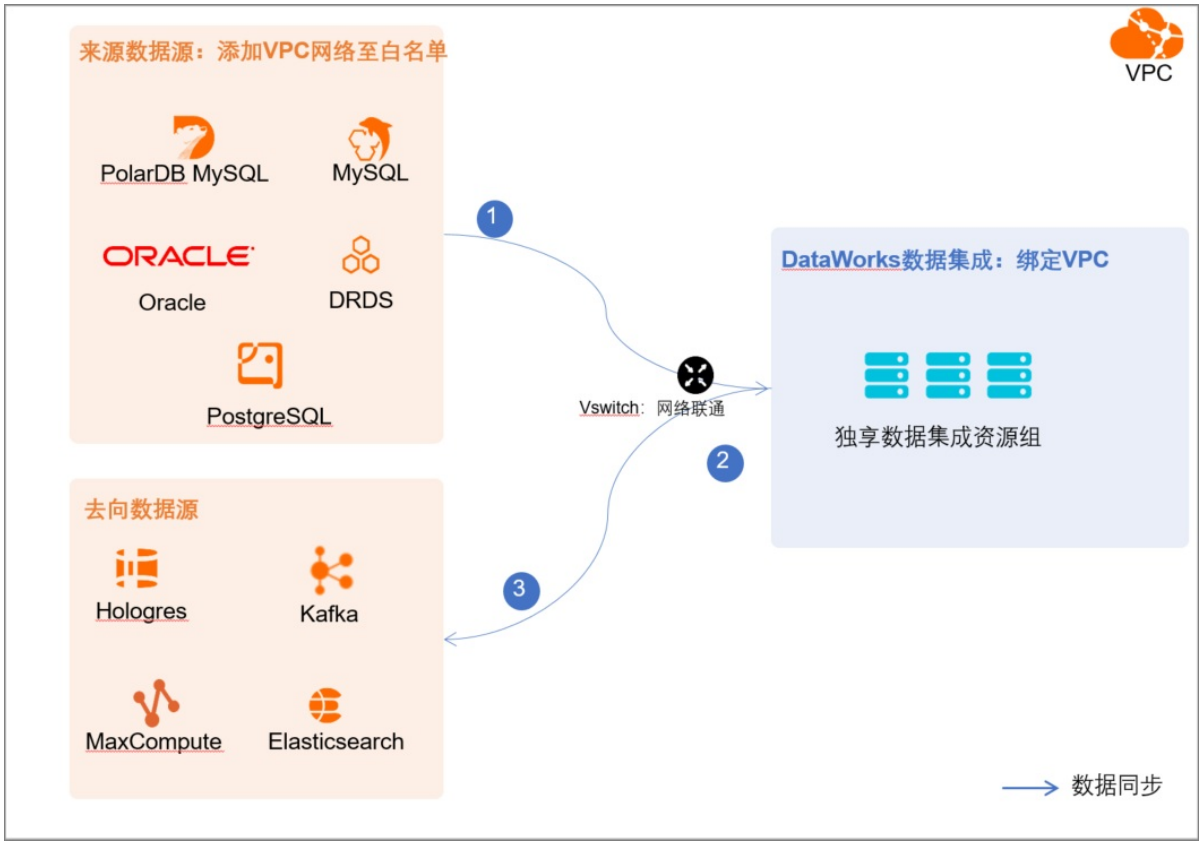
- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源Hologres。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

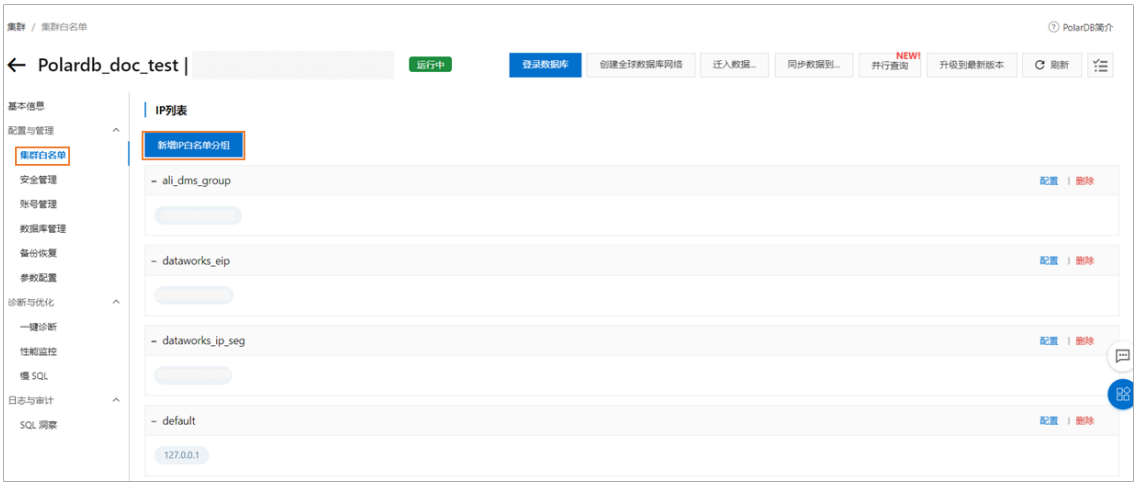
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.5.3. 配置数据源（来源为Oracle）

实时同步Oracle的数据至Hologres时，来源数据源为Oracle，去向数据源为Hologres，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

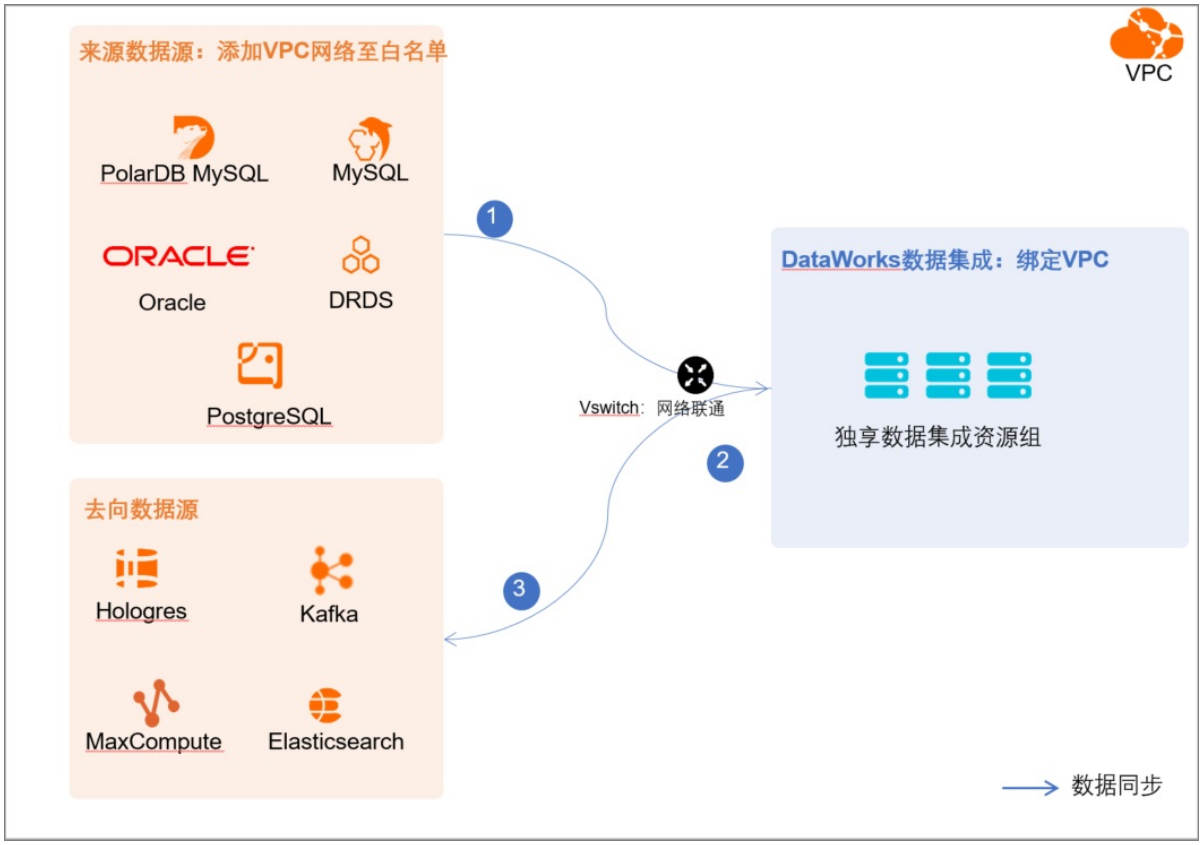
- 准备数据源：已购买来源数据源Oracle、去向数据源Hologres。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。

- Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

操作步骤

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换`'同步账号'`为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.5.4. 配置数据源（来源为MySQL）

实时同步MySQL的数据至Hologres时，来源数据源为MySQL，去向数据源为Hologres，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源Hologres。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。

- 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

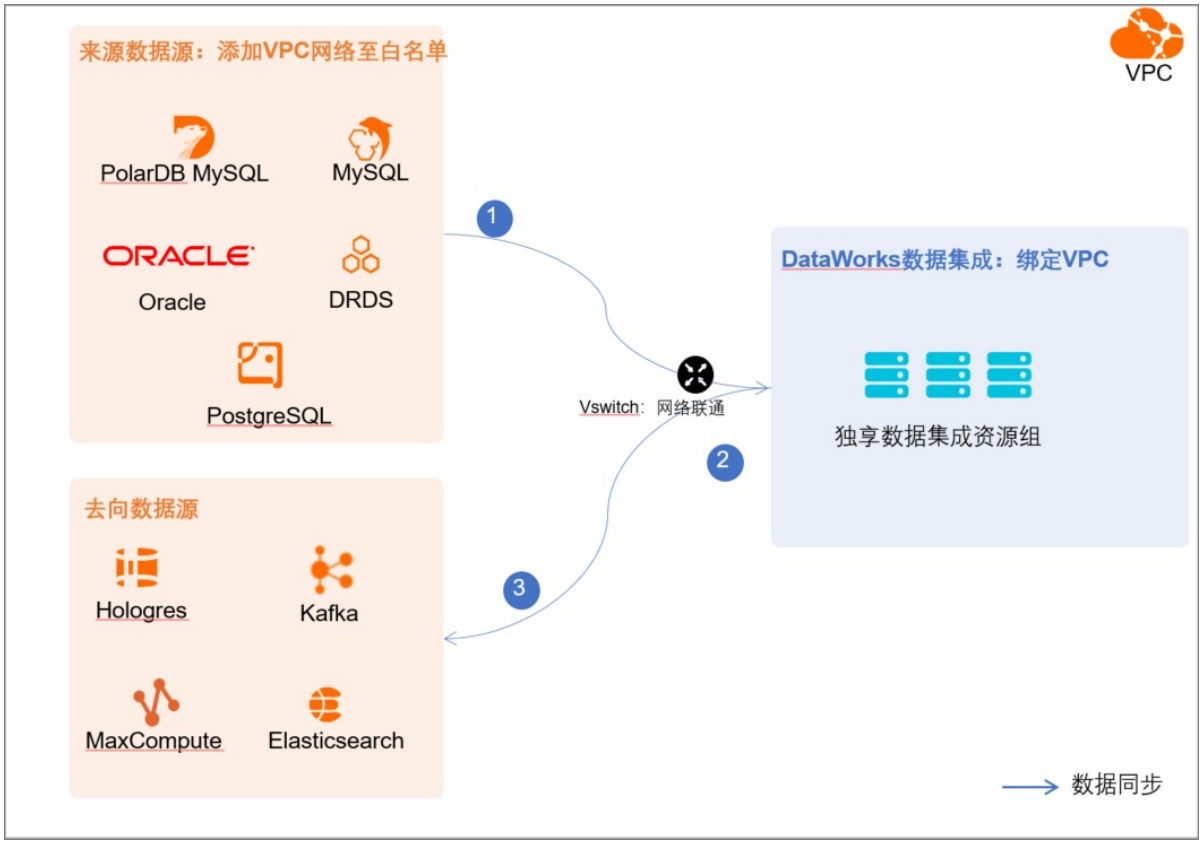
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。

- Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的`5.x`或`8.x`版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

   a. 登录DataWorks控制台。

   b. 在左侧导航栏，单击**资源组列表**。

   c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

   d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

   e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

   f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的`SELECT, REPLICATION SLAVE, REPLICATION CLIENT`权限。

   i. 创建账号。

   操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.5.5. 添加数据源

将来源数据源的数据同步至Hologres数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：Oracle

添加Oracle数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置Oracle数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：SQL Server

添加SQL Server数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置数据源（来源为SQL Server）。

添加去向数据源：Hologres

操作详情可参见配置Hologres数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置并管理实时同步任务。

4.5.6. 配置并管理实时同步任务

完成数据源、网络、资源的准备配置后，您可创建实时同步节点，同步数据至Hologres。本文为您介绍如何创建数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

创建实时数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

使用限制

- 实时数据同步任务仅支持使用独享数据集成资源组。
- 实时同步节点目前仅支持同步PolarDB、Oracle、MySQL数据源至Hologres。
- 实时数据同步任务暂不支持同步没有主键的表。

创建实时同步任务

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 创建**业务流程**。

   如果您已有**业务流程**，则可以忽略该步骤。

   i. 鼠标悬停至 +新建 图标，选择**新建业务流程**。

   ii. 在**新建业务流程**对话框，输入**业务名称**。

iii. 单击**新建**。

5. 创建实时同步节点。

i. 鼠标悬停至+新建图标，选择**数据集成 > 实时同步**。

您也可以找到目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

ii. 在**新建节点**对话框中，配置各项参数。

参数	描述
节点类型	默认为**实时同步**。
同步方式	选择**数据库迁至Hologres**，用于迁移目标数据库下的部分或所有表至Hologres中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放实时同步节点的目录。

iii. 单击**提交**，进入实时同步任务编辑页面。

6. 选择资源组。

i. 在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。

ii. 在**资源组**下拉框，选择需要使用的资源组。

> **说明**
> 实时数据同步任务仅支持使用独享数据集成资源组。
> 如果您没有可用的独享数据集成资源组，请单击**新建独享资源组**创建。详情请参见独享数据集成资源组概述。

7. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、Oracle和PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

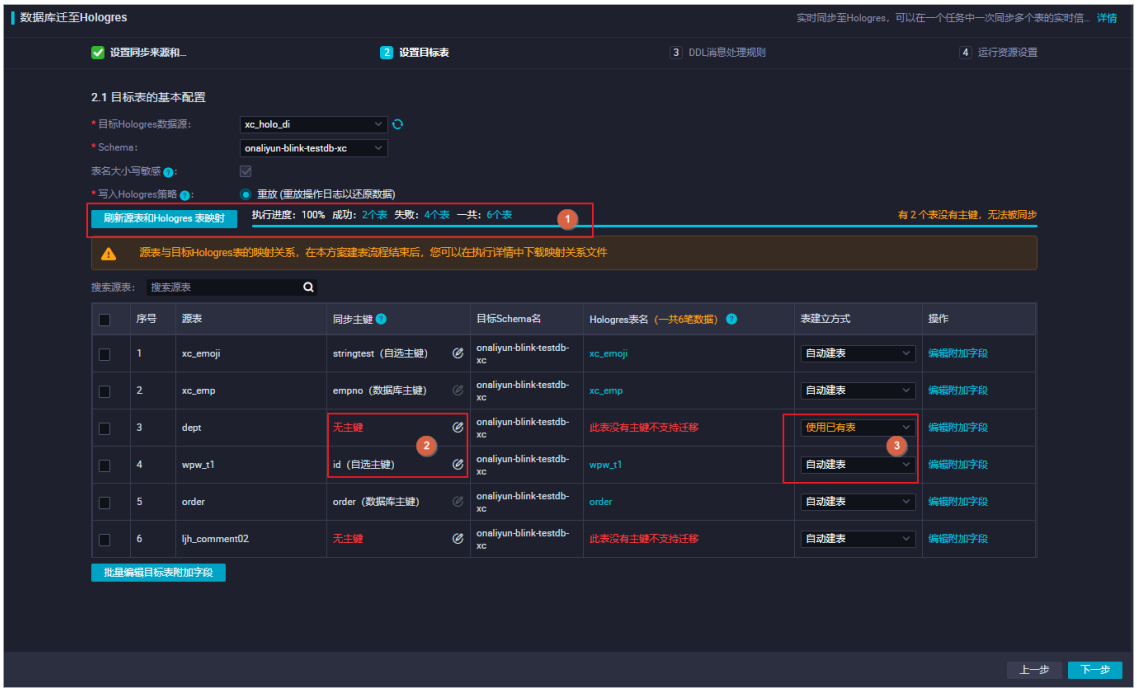
iv. 单击**下一步**。

8. 选择目标数据源并配置目标表格式。

i. 在**设置目标表**页面，选择**目标Hologres数据源**和**写入Hologres策略**。

ii. 单击**刷新源表和Hologres表映射**，创建需要同步的源表和目标Hologres表的映射关系。

iii. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	包括**自动建表**和**使用已有表**。

iv. 单击**下一步**

如果您前一步中目标数据源使用的**表建立方式**为**自动建表**，则需要在弹出的**自动建表**对话框，单击**开始建表**，批量创建目标Hologres表。

9. DDL消息处理规则设置。

来源数据源会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的规则。

说明 该规则为初次执行实时同步任务时的DDL消息处理规则，后续如果您需要修改规则，则可以进入实时任务的运维配置页面修改，详情请参见本文档的*执行实时同步任务*小节。

i. 在**DDL消息处理规则**页签，配置**实时同步DDL消息处理策略**。

不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

10. 运行资源设置。

i. 在**运行资源设置**页面，配置各项参数。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为15。
目标端写入并发数	数据同步任务内，可以从来源表并行读取或写入数据至目标端的最大线程数。最大并发数为32。请根据您的资源组大小和目标端实际规模合理设置。

ii. 单击**完成配置**。

提交并发布实时同步任务

1. 提交并发布节点任务。

i. 单击工具栏中的图标，保存节点。

ii. 单击工具栏中的图标，提交节点任务。

iii. 在**提交新版本**对话框中，输入**变更描述**。

iv. 单击**确定**。

如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

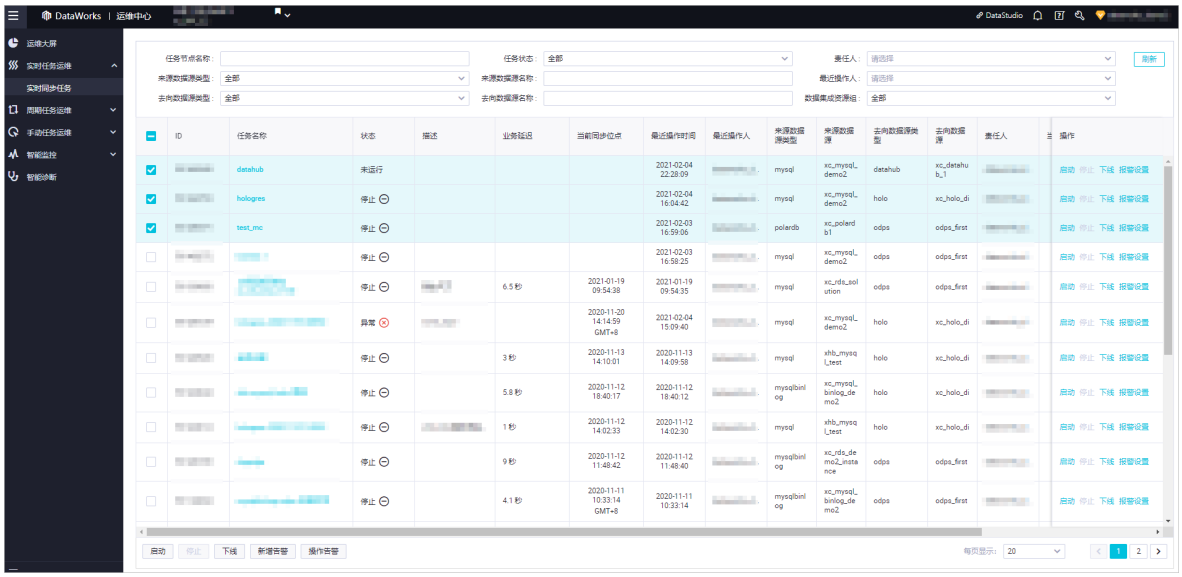
执行实时同步任务

1. 进入运维页面。

   提交或发布节点成功后，单击节点编辑页面右上方的**运维中心**，进入**实时任务运维 > 实时同步任务**页面。

2. 查看实时同步任务详情。

   在**实时同步任务**页面，单击相应**任务名称**，查看运维任务的详细信息。



3. 执行实时同步任务。

   i. 单击目标实时同步任务**操作**列的**启动**。

ii. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点任务的日期和时间。
时区	从**时区**下拉列表中选择执行任务的时区。
Failover	您可以设置在固定时间内，任务的Failover超过指定次数时，自动结束任务。 **说明** 如果您不配置Failover的次数，将根据5分钟Failover超过100次来自动结束任务，避免频繁启动任务占用系统资源。
脏数据策略	■ **零容忍，不允许**：只要同步任务中包含脏数据，则任务自动结束。 ■ **不限制**：无论同步任务中是否包含脏数据，任务均可正常执行。 ■ **有限控制**：指定可包含固定数值的脏数据，超出该数值时任务自动结束。
实时同步DDL消息处理策略	您可以根据需求修改已配置的DDL消息处理策略。详情请参见配置DDL消息处理策略。

管理实时同步任务

- 停止运行中的任务。

  单击相应任务后的**停止**。在停止对话框中，单击**停止**。
- 下线非运行状态的任务。

  单击相应任务后的**下线**。在下线对话框中，单击**下线**。
- 查看任务的报警信息。

单击相应任务后的**报警设置**，在**报警设置**页面查看**报警事件**及**报警规则**。

- 为任务新增告警。
  i. 选中需要新增告警的任务，单击**实时同步任务**页面下方的**新增告警**。
  ii. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称。
描述	新建规则的描述信息。
指标	产生报警的指标项： ■ **任务状态** ■ **业务延迟** ■ **Failover** ■ **脏数据** ■ **DDL错误**
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。
报警间隔	设置报警的时间间隔，默认为5分钟发一次报警。
WARNING	产生相应报警时，可以选择通过**邮件**、**短信**和**钉钉**发送报警通知。 **说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
CRITICAL	
接收人（非钉钉）	选择报警通知的接收人。

  iii. 单击**确定**。
- 批量修改目标任务中指定类型的所有告警。
  i. 选中需要操作告警的任务，单击**实时同步任务**页面下方的**操作告警**。
  ii. 在**操作告警**对话框中，选则需要修改的**操作类型**和**告警指标**。
  iii. 单击**确定**。

4.5.7. 常见问题

以下为您介绍实时同步数据至Hologres操作失败的常见问题和解决方案。

- PolarDB数据源网络联通性测试失败怎么办？
- Oracle数据源网络联通性测试失败怎么办？
- MySQL数据源网络联通性测试失败怎么办？
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

PolarDB数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源PolarDB时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

Oracle数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源Oracle时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

MySQL数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源MySQL时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX` 。
- 可能原因：来源数据源PolarDB没有开启binlog。
- 如何处理：PolarDB开启binlog，详细操作可参见配置数据源（来源为PolarDB）。并进行至少一条数据的变更，同时切换数据集成实时同步开始点位到当前时间。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation` 。
- 可能原因：来源数据源PolarDB没有给进行数据同步的账号开启所需权限，或对接的PolarDB数据库不是主节点。
- 如何处理：参见配置数据源（来源为PolarDB）的操作授予权限，或者检查PolarDB是否是主节点（读写库），目前实时任务不支持从PolarDB备节点抓取数据。

实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first` 。
- 可能原因：来源数据源PolarDB未打开loose_polar_log_bin参数。
- 如何处理：需要打开loose_polar_log_bin参数，详细操作可参见配置数据源（来源为PolarDB）。

4.6. 同步整库数据至AnalyticDB MySQL

4.6.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

   > **说明** 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的实时数据同步任务。目前同步数据至AnalyticDB MySQL的来源数据源仅支持PolarDB及MySQL，您可以根据实际需求选择合适的数据源。数据源的配置可参见配置数据源（来源为PolarDB）或配置数据源（来源为MySQL）。

4.6.2. 配置数据源（来源为PolarDB）

实时同步PolarDB的数据至AnalyticDB MySQL时，来源数据源为PolarDB，去向数据源为AnalyticDB MySQL，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

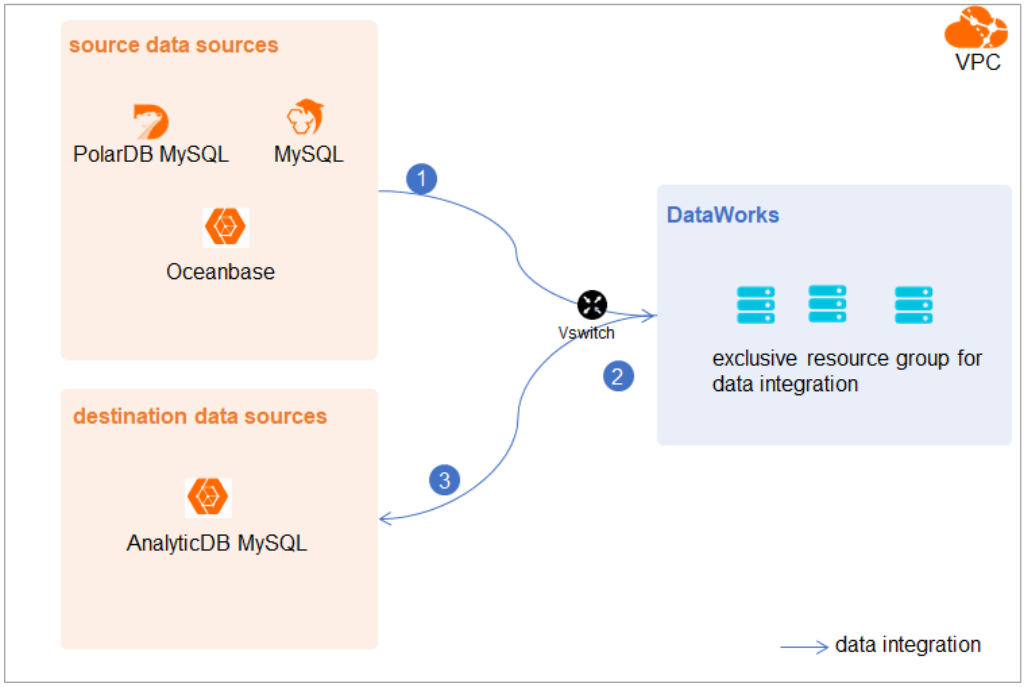
- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源AnalyticDB MySQL。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

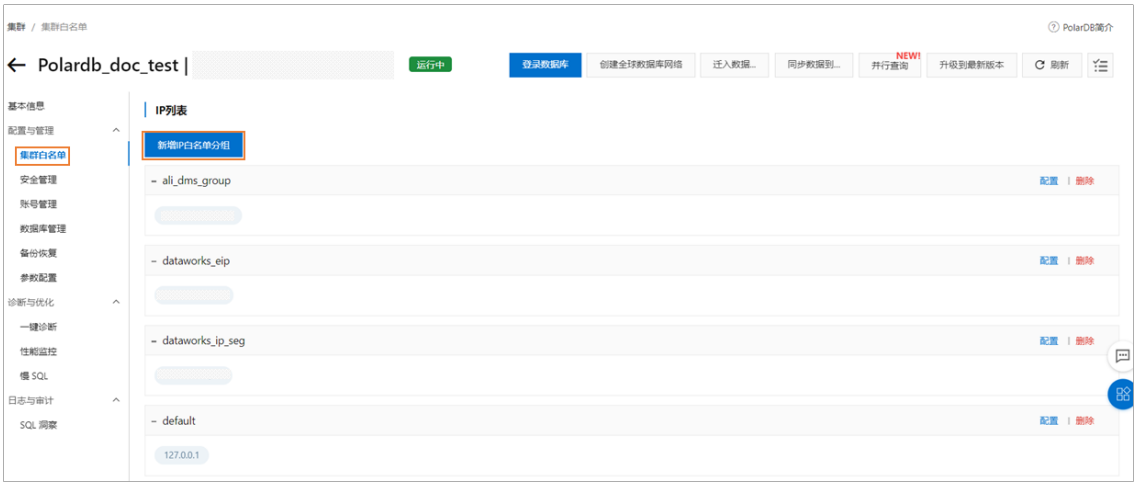
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.6.3. 配置数据源（来源为MySQL）

实时同步MySQL的数据至AnalyticDB MySQL时，来源数据源为MySQL，去向数据源为AnalyticDB MySQL，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源AnalyticDB MySQL。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

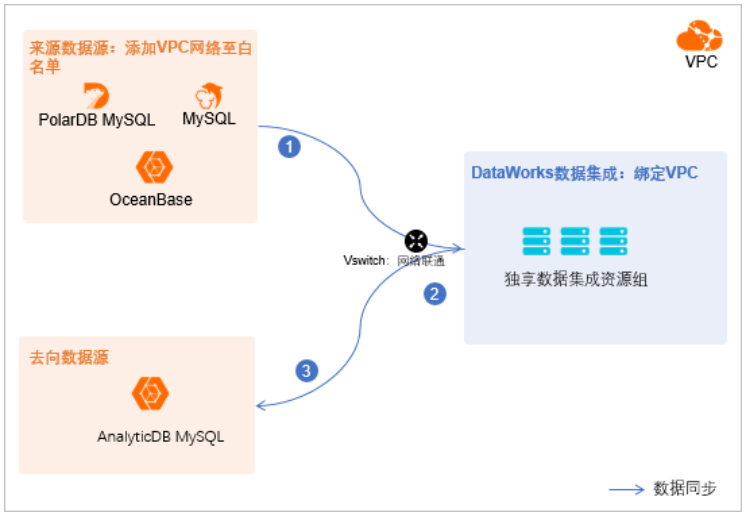
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.6.4. 配置数据源（来源为OceanBase）

同步OceanBase的数据至AnalyticDB MySQL 3.0时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

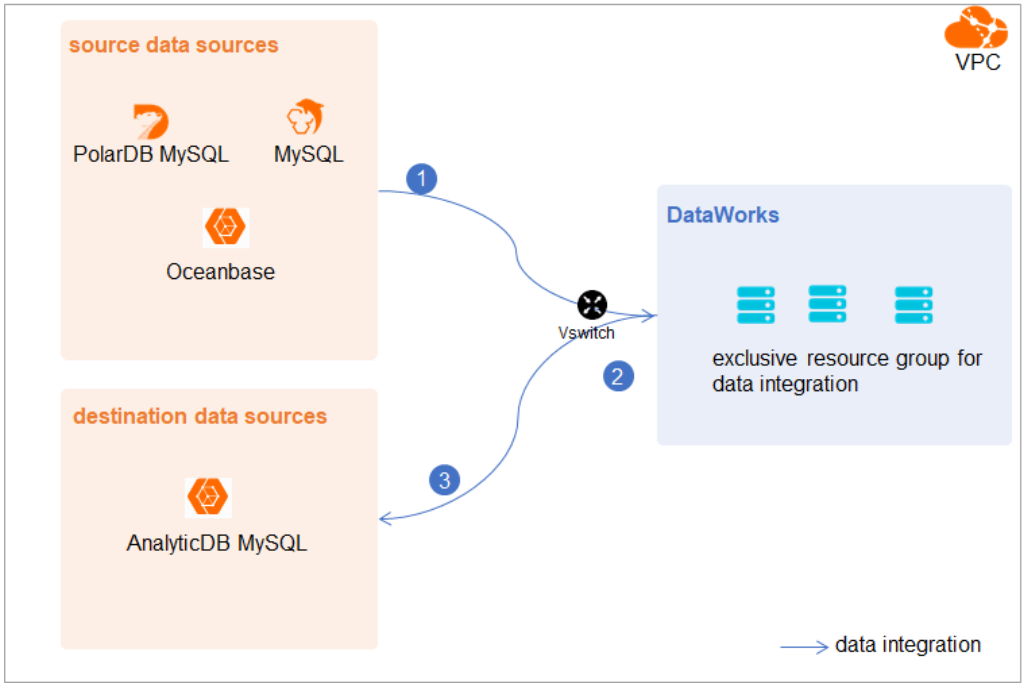
- 准备数据源：已购买来源数据源OceanBase、去向数据源AnalyticDB MySQL 3.0。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

使用限制

OceanBase为分布式关系型数据库，可以使物理分布不同的多个数据库上的数据，被整合为一个完整的逻辑数据库。但实时同步OceanBase的数据至AnalyticDB MySQL，目前仅支持同步单个物理库的数据，不支持同步逻辑库数据。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至OceanBase的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至OceanBase集群的白名单中，详情请参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有OceanBase的相关操作权限，详情请参见新建账号。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.6.5. 添加数据源

将来源数据源的数据同步至AnalyticDB MySQL数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加去向数据源：AnalyticDB MySQL

添加AnalyticDB MySQL数据源，详情请参见配置AnalyticDB for MySQL 3.0数据源

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置并管理实时同步任务。

4.6.6. 配置并管理实时同步任务

完成数据源、网络、资源的准备配置后，您可创建实时同步节点，同步数据至AnalyticDB MySQL。本文为您介绍如何创建数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

创建实时数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为MySQL）
- 配置数据源（来源为OceanBase）
- 添加数据源

使用限制

- 实时数据同步任务仅支持使用独享数据集成资源组。
- 实时同步节点目前仅支持同步PolarDB、MySQL、OceanBase数据源至AnalyticDB MySQL。
- 实时数据同步任务暂不支持同步没有主键的表。

创建实时同步任务

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 创建**业务流程**。

   如果您已有**业务流程**，则可以忽略该步骤。

   i. 鼠标悬停至 +新建 图标，选择**新建业务流程**。

   ii. 在**新建业务流程**对话框，输入**业务名称**。

   iii. 单击**新建**。
5. 创建实时同步节点。

   i. 鼠标悬停至 +新建 图标，选择**数据集成 > 实时同步**。

   您也可以找到目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

ii. 在**新建节点**对话框中，配置各项参数。

参数	描述
节点类型	默认为**实时同步**。
同步方式	选择**数据库迁至AnalyticDB MySQL 3.0**，用于迁移目标数据库下的部分或所有表至AnalyticDB MySQL中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放实时同步节点的目录。

iii. 单击**提交**，进入实时同步任务编辑页面。

6. 选择资源组。

i. 在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。

ii. 在**资源组**下拉框，选择需要使用的资源组。

> **说明**
>
> 实时数据同步任务仅支持使用独享数据集成资源组。
>
> 如果您没有可用的独享数据集成资源组，请单击**新建独享资源组**创建。详情请参见独享数据集成资源组概述。

7. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、PolarDB和OceanBase类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

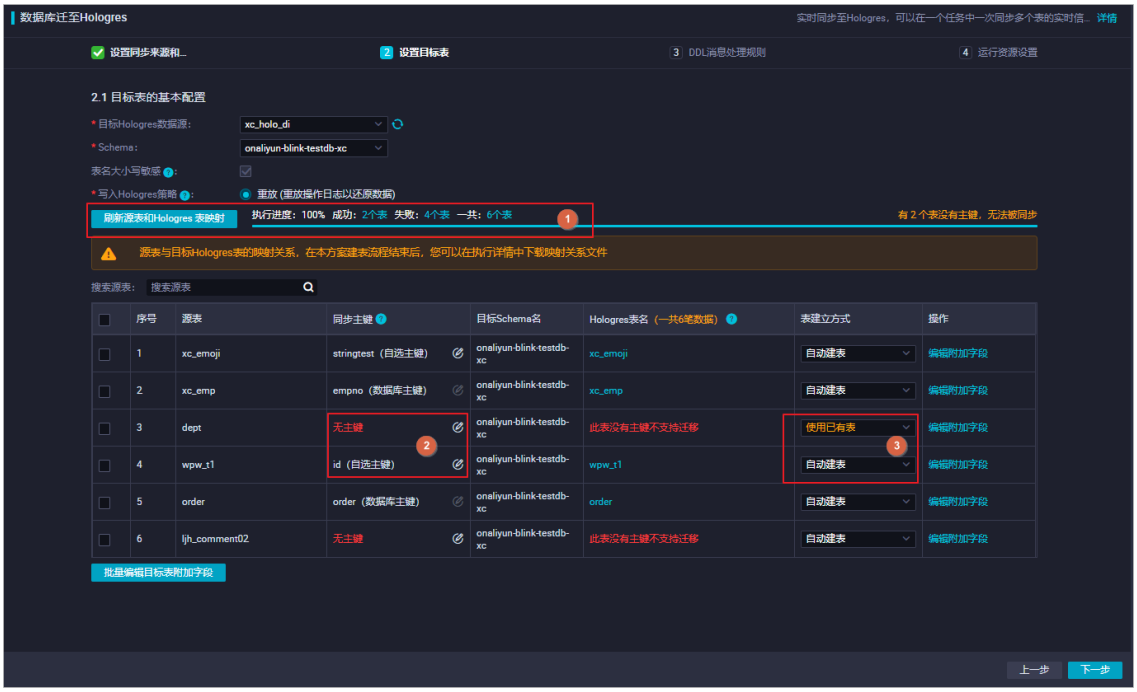
iv. 单击**下一步**。

8. 选择目标数据源并配置目标表格式。

i. 在**设置目标表**页面，选择**目标AnalyticDB MySQL 3.0数据源**。

ii. 单击**刷新源表和AnalyticDB MySQL 3.0 表映射**，创建需要同步的源表和目标AnalyticDB MySQL表的映射关系。

iii. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。 **说明** 实时数据同步任务暂不支持同步没有主键的表。
③	选择的**表建立方式**，取值如下： ■ 当**表建立方式**选择**使用已有表**时，**AnalyticDB MySQL 3.0表名**列显示自动创建的AnalyticDB MySQL表名称。您也可以在下拉列表中选择需要使用的表名称。 ■ 当**表建立方式**选择**自动建表**时，显示自动创建的AnalyticDB MySQL表名称。您可以单击表名称，查看和修改建表语句。

iv. 单击**下一步**

如果您前一步中目标数据源使用的**表建立方式**为**自动建表**，则需要在弹出的**自动建表**对话框，单击**开始建表**，批量创建目标AnalyticDB MySQL表。

9. DDL消息处理规则设置。

来源数据源会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的规则。

> **说明** 该规则为初次执行实时同步任务时的DDL消息处理规则，后续如果您需要修改规则，则可以进入实时任务的运维配置页面修改，详情请参见本文档的*执行实时同步任务*小节。

i. 在**DDL消息处理规则**页签，配置**实时同步DDL消息处理策略**。

不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

10. 运行资源设置。

i. 在**运行资源设置**页面，配置各项参数。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为15。
目标端写入并发数	数据同步任务内，可以从来源表并行读取或写入数据至目标端的最大线程数。最大并发数为32。请根据您的资源组大小和目标端实际规模合理设置。

ii. 单击**完成配置**。

提交并发布实时同步任务

1. 提交并发布节点任务。

i. 单击工具栏中的图标，保存节点。

ii. 单击工具栏中的图标，提交节点任务。

iii. 在**提交新版本**对话框中，输入**变更描述**。

iv. 单击**确定**。

如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

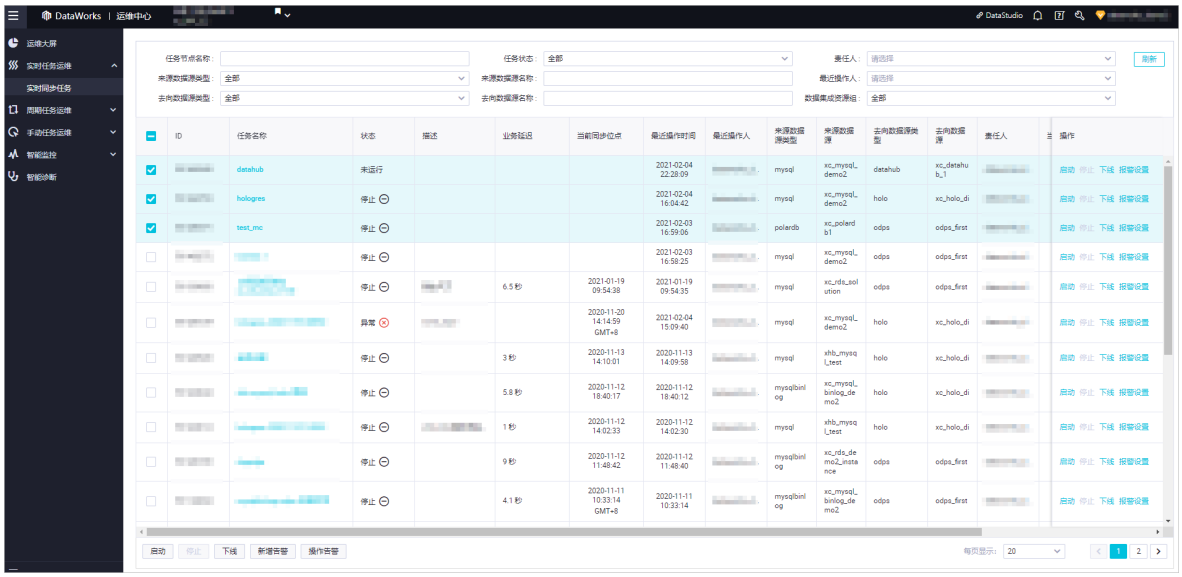
执行实时同步任务

1. 进入运维页面。

   提交或发布节点成功后，单击节点编辑页面右上方的**运维中心**，进入**实时任务运维 > 实时同步任务**页面。

2. 查看实时同步任务详情。

   在**实时同步任务**页面，单击相应**任务名称**，查看运维任务的详细信息。



3. 执行实时同步任务。

   i. 单击目标实时同步任务**操作**列的**启动**。

ii. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点任务的日期和时间。
时区	从**时区**下拉列表中选择执行任务的时区。
Failover	您可以设置在固定时间内，任务的Failover超过指定次数时，自动结束任务。 **说明** 如果您不配置Failover的次数，将根据5分钟Failover超过100次来自动结束任务，避免频繁启动任务占用系统资源。
脏数据策略	■ **零容忍，不允许**：只要同步任务中包含脏数据，则任务自动结束。 ■ **不限制**：无论同步任务中是否包含脏数据，任务均可正常执行。 ■ **有限控制**：指定可包含固定数值的脏数据，超出该数值时任务自动结束。
实时同步DDL消息处理策略	您可以根据需求修改已配置的DDL消息处理策略。详情请参见配置DDL消息处理策略。

管理实时同步任务

- 停止运行中的任务。

  单击相应任务后的**停止**。在停止对话框中，单击**停止**。
- 下线非运行状态的任务。

  单击相应任务后的**下线**。在下线对话框中，单击**下线**。
- 查看任务的报警信息。

单击相应任务后的**报警设置**，在**报警设置**页面查看**报警事件**及**报警规则**。

- 为任务新增告警。
  i. 选中需要新增告警的任务，单击**实时同步任务**页面下方的**新增告警**。
  ii. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称。
描述	新建规则的描述信息。
指标	产生报警的指标项： ■ **任务状态** ■ **业务延迟** ■ **Failover** ■ **脏数据** ■ **DDL错误**
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。
报警间隔	设置报警的时间间隔，默认为5分钟发一次报警。
WARNING	产生相应报警时，可以选择通过**邮件**、**短信**和**钉钉**发送报警通知。 **说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
CRITICAL	
接收人（非钉钉）	选择报警通知的接收人。

  iii. 单击**确定**。
- 批量修改目标任务中指定类型的所有告警。
  i. 选中需要操作告警的任务，单击**实时同步任务**页面下方的**操作告警**。
  ii. 在**操作告警**对话框中，选则需要修改的**操作类型**和**告警指标**。
  iii. 单击**确定**。

4.6.7. 常见问题

以下为您介绍实时同步数据至AnalyticDB MySQL 3.0操作失败的常见问题和解决方案。

- PolarDB数据源网络联通性测试失败怎么办？
- MySQL数据源网络联通性测试失败怎么办？
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

PolarDB数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源PolarDB时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

MySQL数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源MySQL时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX` 。
- 可能原因：来源数据源PolarDB没有开启binlog。
- 如何处理：PolarDB开启binlog，详细操作可参见配置数据源（来源为PolarDB）。并进行至少一条数据的变更，同时切换数据集成实时同步开始点位到当前时间。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation` 。
- 可能原因：来源数据源PolarDB没有给进行数据同步的账号开启所需权限，或对接的PolarDB数据库不是主节点。
- 如何处理：参见配置数据源（来源为PolarDB）的操作授予权限，或者检查PolarDB是否是主节点（读写库），目前实时任务不支持从PolarDB备节点抓取数据。

实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first` 。
- 可能原因：来源数据源PolarDB未打开loose_polar_log_bin参数。
- 如何处理：需要打开loose_polar_log_bin参数，详细操作可参见配置数据源（来源为PolarDB）。

4.7. 同步整库数据至DataHub

4.7.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。

2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

   > **说明** 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的实时数据同步任务。目前同步数据至DataHub的来源数据源仅支持PolarDB、OceanBase、MySQL及Oracle，您可以根据实际需求选择合适的数据源。数据源的配置可参见配置数据源（来源为PolarDB）、配置数据源（来源为OceanBase）、配置数据源（来源为MySQL）或配置数据源（来源为Oracle）。

4.7.2. 配置数据源（来源为PolarDB）

实时同步PolarDB的数据至DataHub时，来源数据源为PolarDB，去向数据源为DataHub，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

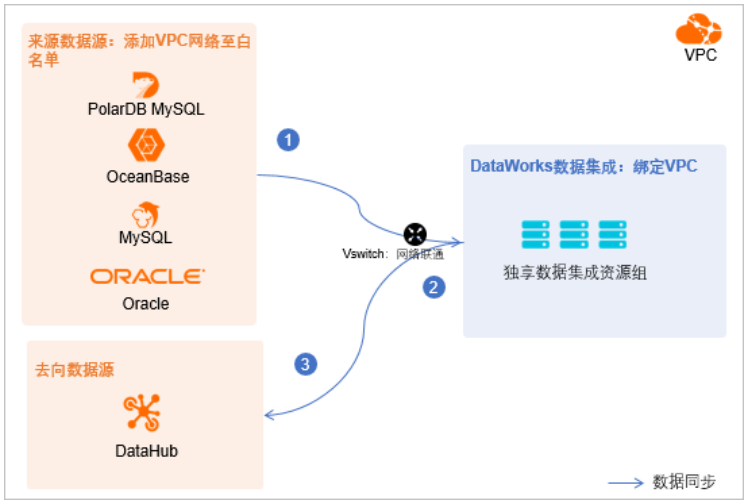
- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源DataHub。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

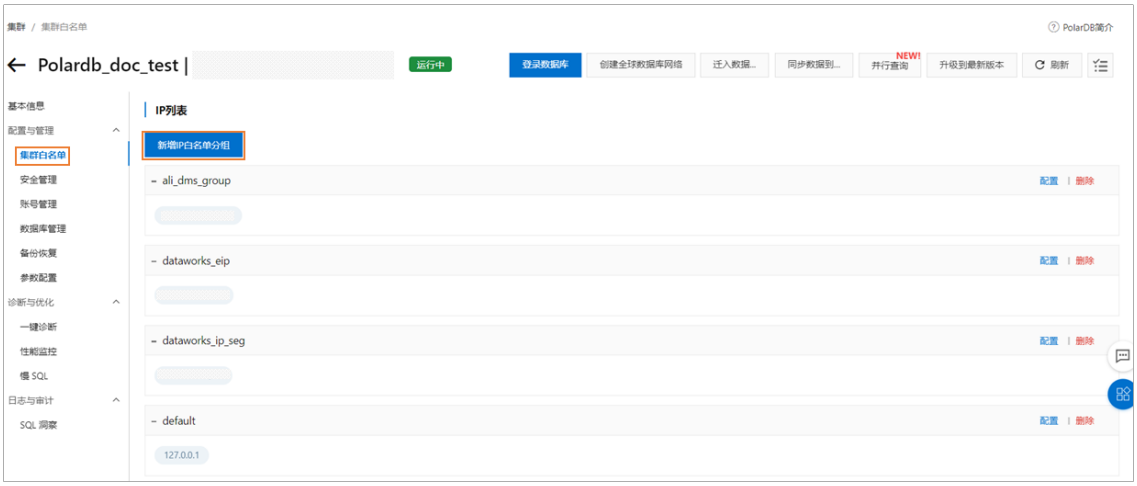
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.7.3. 配置数据源（来源为MySQL）

实时同步MySQL的数据至DataHub时，来源数据源为MySQL，去向数据源为DataHub，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源DataHub。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

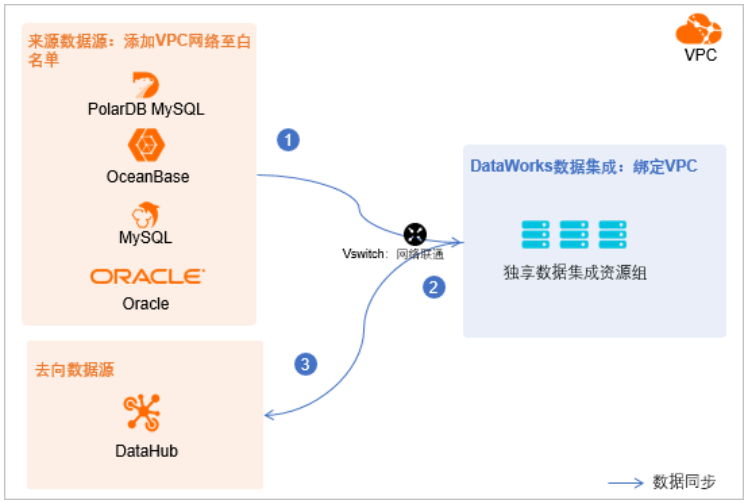
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.7.4. 配置数据源（来源为OceanBase）

同步OceanBase的数据至DataHub时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

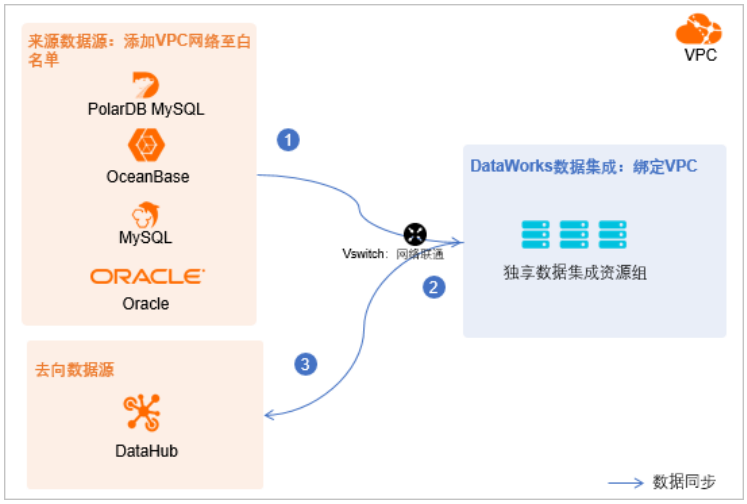
- 准备数据源：已购买来源数据源OceanBase、去向数据源DataHub。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

使用限制

OceanBase为分布式关系型数据库，可以使物理分布不同的多个数据库上的数据，被整合为一个完整的逻辑数据库。但实时同步OceanBase的数据至DataHub，目前仅支持同步单个物理库的数据，不支持同步逻辑库数据。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至OceanBase的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至OceanBase集群的白名单中，详情请参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有OceanBase的相关操作权限，详情请参见新建账号。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.7.5. 配置数据源（来源为Oracle）

实时同步Oracle的数据至DataHub时，来源数据源为Oracle，去向数据源为DataHub，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

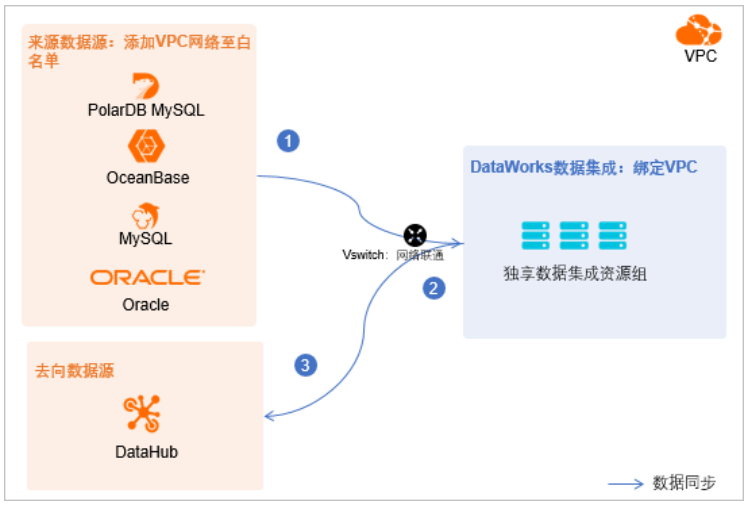
- 准备数据源：已购买来源数据源Oracle、去向数据源DataHub。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。
  - Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。

- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- 目前仅华南1（深圳）地域开放实时整库同步Oracle数据源至DataHub功能。如果其他地域需要使用该功能，请提交工单申请。
- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

操作步骤

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换`'同步账号'`为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.7.6. 添加数据源

将来源数据源的数据同步至DataHub数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：OceanBase

添加OceanBase数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置ApsaraDB for OceanBase数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加去向数据源：DataHub

操作详情可参见配置DataHub数据源。

添加来源数据源：Oracle

添加Oracle数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置Oracle数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置并管理实时同步任务。

4.7.7. 配置并管理实时同步任务

完成数据源、网络、资源的准备配置后，您可创建实时同步节点，同步数据至DataHub。本文为您介绍如何创建数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

创建实时数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为OceanBase）
- 配置数据源（来源为MySQL）
- 配置数据源（来源为Oracle）
- 添加数据源

使用限制

- 实时数据同步任务仅支持使用独享数据集成资源组。
- 实时同步节点目前仅支持同步PolarDB、OceanBase、MySQL及Oracle数据源至DataHub。
- 目前仅华南1（深圳）地域开放实时整库同步Oracle数据源至DataHub功能。如果其他地域需要使用该功能，请提交工单申请。

使用说明

DataHub不同数据类型对应操作的支持情况，不同数据类型的分片策略、数据格式及相关消息示例。详情请参见：附录：DataHub消息格式。

创建实时同步任务

1. 登录DataWorks控制台。

2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 创建**业务流程**。

   如果您已有**业务流程**，则可以忽略该步骤。

   i. 鼠标悬停至**+新建**图标，选择**新建业务流程**。

   ii. 在**新建业务流程**对话框，输入**业务名称**。

   iii. 单击**新建**。
5. 创建实时同步节点。

   i. 鼠标悬停至**+新建**图标，选择**数据集成 > 实时同步**。

   您也可以找到目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

   ii. 在**新建节点**对话框中，配置各项参数。

参数	描述
节点类型	默认为**实时同步**。
同步方式	选择**数据库迁至DataHub**，用于迁移目标数据库下的部分或所有Topic至DataHub中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放实时同步节点的目录。

   iii. 单击**提交**，进入实时同步任务编辑页面。
6. 选择资源组。

   i. 在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。

   ii. 在**资源组**下拉框，选择需要使用的资源组。

> **说明**
>
> 实时数据同步任务仅支持使用独享数据集成资源组。
>
> 如果您没有可用的独享数据集成资源组，请单击**新建独享资源组**创建。详情请参见独享数据集成资源组概述。

7. 选择来源数据源并配置同步规则。

   i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、OceanBase和PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

iii. 在**设置表名到Topic的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

映射规则包括**源表名和目标Topic转换规则**和**目标Topic规则**：

- **源表名和目标Topic转换规则**：转换表名为目标Topic，进行字符串替换。
- **目标Topic规则**：支持对转换后的Topic添加前缀和后缀。

iv. 单击**下一步**。

8. 选择目标数据源并配置目标Topic格式。

i. 在**设置目标Topic**页面，选择**目标DataHub数据源**、**DataHub写入模式**及**分片策略**。

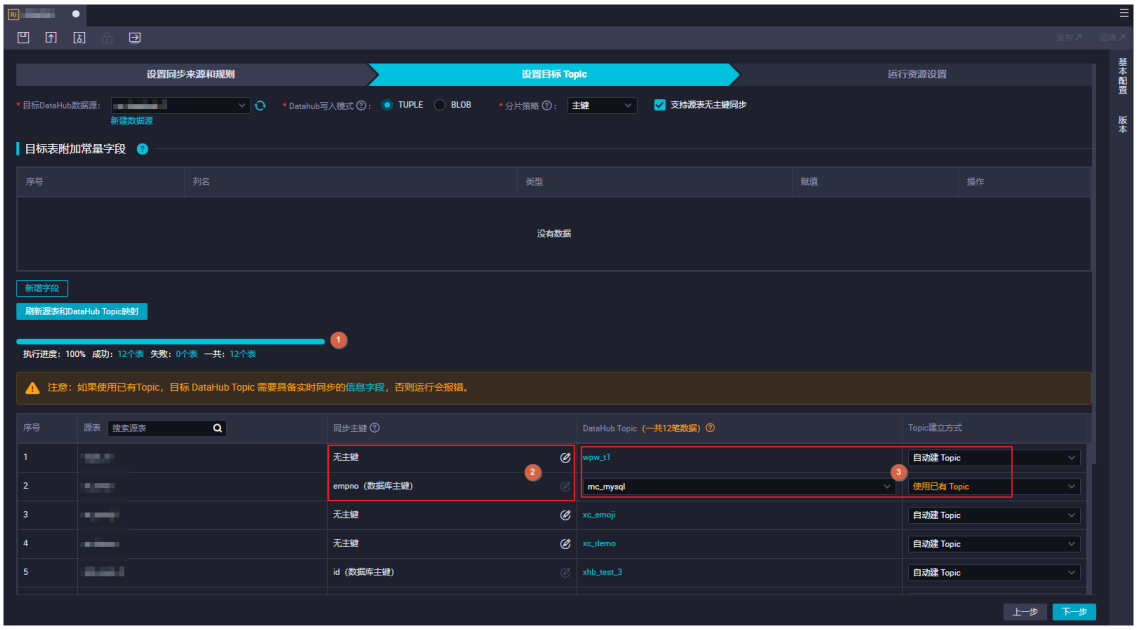
如果您需要同步无主键的来源表，则可以勾选**支持源表无主键同步**。

ii. （可选）目标表新增字段。

如果您希望为目标表中的所有同步表新增统一的字段，则可以在**目标表附加常量字段**区域，单击**新增字段**添加。

iii. 单击**刷新源表和DataHhub Topic映射**，创建需要同步的源表和目标DataHhub Topic的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键： ■ 当在**设置目标Topic**页面勾选了**支持源表无主键同步**，则无主键的表可以正常同步。您可以选择单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。 ■ 当在**设置目标Topic**页面未勾选**支持源表无主键同步**，则无主键的表同步时会出现异常，您需要在同步任务中删除无主键的表，才能继续执行同步任务。
③	选择的**Topic建立方式**不同，此处显示的DataHub Topic也不同： ■ 当**Topic建立方式**选择**自动建Topic**时，显示自动创建的Topic名称。您可以单击Topic名称，编辑Topic信息。 ■ 当**Topic建立方式**选择**使用已有Topic**时，请在下拉列表中选择需要的Topic。

v. 单击**下一步**。

如果您前一步中目标数据源使用的**Topic建立方式**为**自动建Topic**，则需要在弹出的**自动建表**对话框，单击**开始建表**，批量创建目标DataHub Topic。

9. 运行资源设置。

i. 在**运行资源设置**页面，配置各项参数。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为**15**。
目标端写入并发数	数据同步任务内，可以从来源表并行读取或写入数据至目标端的最大线程数。最大并发数为32。请根据您的资源组大小和目标端实际规模合理设置。

ii. 单击**完成配置**。

提交并发布实时同步任务

1. 提交并发布节点任务。
   i. 单击工具栏中的图标，保存节点。
   ii. 单击工具栏中的图标，提交节点任务。
   iii. 在**提交新版本**对话框中，输入**变更描述**。
   iv. 单击**确定**。

   如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

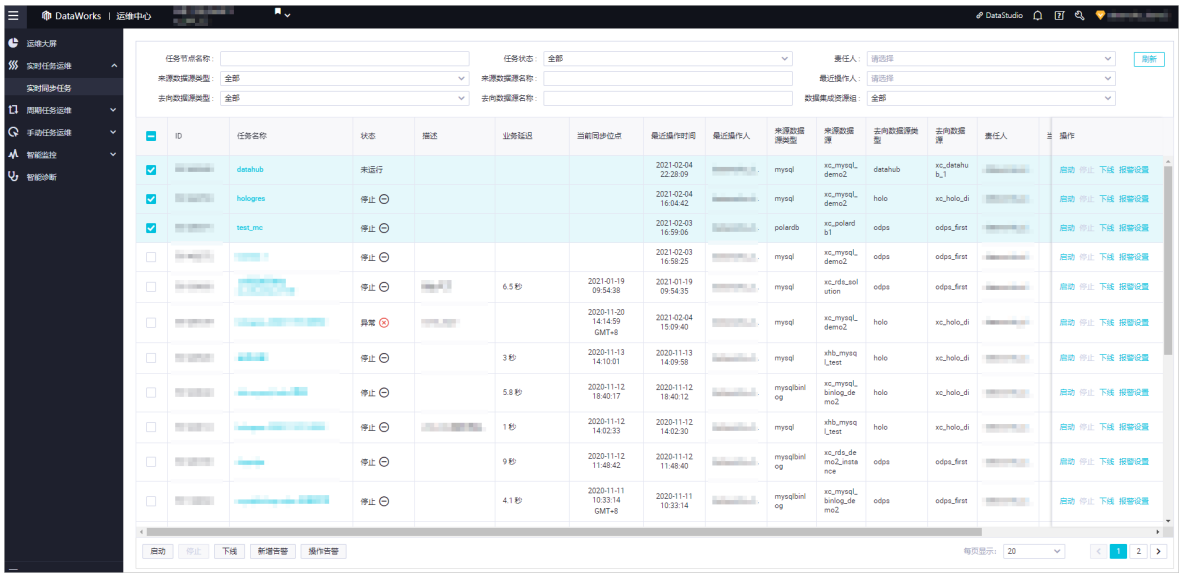
执行实时同步任务

1. 进入运维页面。

   提交或发布节点成功后，单击节点编辑页面右上方的**运维中心**，进入**实时任务运维 > 实时同步任务**页面。

2. 查看实时同步任务详情。

   在**实时同步任务**页面，单击相应**任务名称**，查看运维任务的详细信息。



3. 执行实时同步任务。
   i. 单击目标实时同步任务**操作**列的**启动**。

ii. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点任务的日期和时间。
时区	从**时区**下拉列表中选择执行任务的时区。
Failover	您可以设置在固定时间内，任务的Failover超过指定次数时，自动结束任务。 **说明** 如果您不配置Failover的次数，将根据5分钟Failover超过100次来自动结束任务，避免频繁启动任务占用系统资源。
脏数据策略	■ **零容忍，不允许**：只要同步任务中包含脏数据，则任务自动结束。 ■ **不限制**：无论同步任务中是否包含脏数据，任务均可正常执行。 ■ **有限控制**：指定可包含固定数值的脏数据，超出该数值时任务自动结束。

iii. 单击**确定**。

管理实时同步任务

- 停止运行中的任务。

  单击相应任务后的**停止**。在停止对话框中，单击**停止**。
- 下线非运行状态的任务。

  单击相应任务后的**下线**。在下线对话框中，单击**下线**。
- 查看任务的报警信息。

  单击相应任务后的**报警设置**，在**报警设置**页面查看**报警事件**及**报警规则**。
- 为任务新增告警。

  i. 选中需要新增告警的任务，单击**实时同步任务**页面下方的**新增告警**。

  ii. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称。
描述	新建规则的描述信息。

参数	描述
指标	产生报警的指标项： ■ **任务状态** ■ **业务延迟** ■ **Failover** ■ **脏数据** ■ **DDL错误**
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。
报警间隔	设置报警的时间间隔，默认为5分钟发一次报警。
WARNING	产生相应报警时，可以选择通过**邮件**、**短信**和**钉钉**发送报警通知。
CRITICAL	**说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
接收人（非钉钉）	选择报警通知的接收人。

iii. 单击**确定**。

- 批量修改目标任务中指定类型的所有告警。
  i. 选中需要操作告警的任务，单击**实时同步任务**页面下方的**操作告警**。
  ii. 在**操作告警**对话框中，选则需要修改的**操作类型**和**告警指标**。
  iii. 单击**确定**。

4.7.8. 附录：DataHub消息格式

本文为您介绍DataHub不同数据类型对应操作的支持情况，不同数据类型的分片策略、数据格式及相关消息示例。

不同数据类型对应操作的支持情况

Topic是DataHub订阅和发布的最小单位，用户可以用Topic来表示一类或者一种流数据。目前支持Tuple与Blob两种类型：

DataHub类型	写入DML消息	写入上游心跳消息	写入DDL消息	源表和目标topic映射方式	数据类型
Tuple	支持	不支持	不支持	单表对单topic	DataHub支持的类型
Blob	支持	支持	支持	单库（多表）对单topic	Blob二进制数据

- Tuple类型由于schema各字段在topic创建后无法更改，所以适用于schema固定，且源表无add column、drop column等改变schema的DDL操作场景。Tuple类型不支持保留上游传递的DDL消息以及心跳消息，即Tuple不能将此类消息透穿给消费DataHub的下游。而且源表和topic的映射方式为单表对单topic，如果源表数量过多，则需要创建大量的topic，将不便于下游消费。
- Blob类型由于不存在schema，topic内只存放Blob二进制数据，因此有较大的自由度。支持存放源表的DDL消息和源端的心跳消息，可传递给下游消费，且采用单库（多表）对单topic的映射方式，不论源表数量多少仅需创建一个topic，方便下游消费。更适用于DataHub作为中间消息队列进行整库迁移的场景。

不同数据类型分片策略

Shard表示对DataHub的一个topic进行数据传输的并发通道，单个Shard写入速率有上限，多个Shard可以提高写入性能，但DataHub仅能保证单个Shard在消费时的有序性，不保证多个Shard之间消息的顺序。因此，为了既能通过增加Shard数量提高写入性能，又能够保证多个Shard之间消息的有序性，同时避免数据倾斜，现针对Blob和Tuple类型提供以下分片策略。

场景	Tuple	Blob
有主键（包含自定义主键）	按主键进行分片	按主键进行分片
顺序保证	同一主键消息保证有序	同一主键消息保证有序
无主键	随机分片	按表名进行分片
顺序保证	不保证有序	同一表消息保证有序

同步数据格式

- Tuple

  数据类型为DataHub Tuple topic自身支持的类型。当使用数据集成创建的topic时，会增加部分元数据列。

  其中 `_sequence_id_` 、`_excute_time_` 、`_source_table_` 、`_before_image_` 、`_after_image_` 为元信息列。

参数	描述
_sequence_id_	string类型，由数字组成，每条消息唯一（update before和update after共用一个sequence id）。
_excute_time_	数据产生时间。
_source_table_	数据源表名。
_before_image_	前镜像（update before和delete为Y，update after和insert为N）。
_after_image_	后镜像（update before和delete为N，update after和insert为Y）。

  示例：下表为一条Insert、Update、Delete语句同步到DataHub的结果。

_sequence_id_	_operation_type_	_excute_time_	_before_image_	_after_image_
164999161068800000 0	I	1649991726000	N	Y
164999161068800000 1	U	1649991756000	Y	N
164999161068800000 1	U	1649991756000	N	Y
164999161068800000 2	D	1649991774000	Y	N

- Blob

  Blob类型的消息格式为JSON字符串转化的二进制数据，其对应的JSON格式如下：

```
{
```

```
{
    "schema": { //变更的元数据信息，仅指定列名与列类型信息
        "dataColumn": [//变更的数据列信息，更新目标表记录内容
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "binData",
                "type": "BYTES"
            },
            {
                "name": "ts",
                "type": "DATE"
            }
        ],
        "primaryKey": [
            "pkName1",
            "pkName2"
        ],
        "source": {
            "dbType": "mysql",
            "dbVersion": "1.0.0",
            "dbName": "myDatabase",
            "schemaName": "mySchema",
            "tableName": "tableName"
        }
    },
    "payload": {
        "before": {
            "dataColumn":{
                "id": 111,
                "name":"scooter",
                "binData": "[base64 string]",
                "ts": 1590315269000
            }
        },
        "after": {
            "dataColumn":{
                "id": 222,
                "name":"donald",
                "binData": "[base64 string]",
                "ts": 1590315269000
            }
        },
          "sequenceId":XXX//字符串类型，用于增全量数据合并的数据排序，
        "op": "INSERT/UPDATE/DELETE/TRANSACTION_BEGIN/TRANSACTION_END/CREATE/ALTER/ERASE/QUERY/TRUNCATE/RENAME/CINDEX/DINDEX/GTID/XACOMMIT/XAROLLBACK/MHEARTBEAT..."//大小写敏感，
        "timestamp": {
            "eventTime": 1,//必选，记录的变更时间，13为时间戳，ms精度
            "systemTime": 2,//可选，oracle CDC等部分数据源存在
            "checkpointTime": 3//可选，部分数据库如oceanbase等数据源包含
        },
        "ddl": {
            "text": "ADD COLUMN ...",
            "ddlMeta": "[SQLStatement serialized binary, expressed in base64 string]"
        }
    },
    "version":"1.0.0"
}
```

- Blob字段说明

> **注意** 消息中的所有字段类型范围为StreamX定义的BOOLEAN、DOUBLE、DATE、BYTES、LONG，STRING六种类型。

```
BOOLEAN：取值为true，false
DATE：取值为13为整形，时间精确到ms级
BYTES：存储bytes类型，格式为base64编码后的字符串
BASE64编解码使用java.util.Base64中的接口实现：
String text = "测试text123";
//编码
Base64.getEncoder().encodeToString(text.getBytes("UTF-8"))
//编码
Base64.getDecoder().decode(encodedText)//解码
```

一级元素	二级元素	说明
schema	dataColumn	JSONArray类型，数据列的类型信息。dataColumn记录上游数据变更记录的所有列和对应的列类型信息。变更操作包括数据库对数据的更改（新增、删除及修改）和数据库表结构等变更。 ■ name：列名 ■ type：列类型
	primaryKey	List类型，主键信息。 pk：主键名。
	source	Object 类型，源端数据库或表信息。 ■ dbType：String类型，数据库类型 ■ dbVersion：String类型，数据库版本 ■ dbName：String类型，数据库名 ■ schemaName：String类型，Schema名（针对Postgres和SQL Server等） ■ tableName：String 类型，数据表名
	before	JSONObject类型，修改前的数据。例如：数据源端为mysql，做了一次记录的update操作，before字段存储记录被update之前的数据内容。 ■ 在从源端读取到更新、删除操作消息时，在写入记录中填充该字段。 ■ dataColumn：JSONObject类型，表示数据信息。格式为列名：列值， 列名为字符串，列值取决于本身类型，BYTES类型使用Base64 String进行表示，DATE类型采用long表示的13位时间戳，其余类型的值均为本身类型。
	after	修改后的数据。格式同before相同。 **说明** 在更新、插入操作时必填。

一级元素	二级元素	说明
payload	op	操作类型。取值如下： ■ INSERT：数据插入 ■ UPDATE_BEFOR：数据更新前 ■ UPDATE_AFTER：数据更新后 ■ DELETE：数据删除 ■ TRANSACTION_BEGIN：数据库事务开始 ■ TRANSACTION_END：数据库事务结束 ■ CREATE：数据库建表 ■ ALTER：数据库表变更 ■ QUERY：数据库变更的原始SQL ■ TRUNCATE：数据库表清空 ■ RENAME：数据库表重命名 ■ CINDEX：创建索引 ■ DINDEX：删除索引 ■ MHEARTBEAT：用于在源端无新增数据时标识同步仍正常进行的心跳消息
	timestamp	JSONObject 类型，本条数据的相关时间戳。 ■ eventTime：Long类型，记录源端库发生变更的时间，毫秒精度的13位时间戳。 ■ systemTime：Long类型，同步任务处理该条变更消息的时间，毫秒精度的13位时间戳。 ■ checkpointTime：Long类型，重置同步位点时的设置时间，毫秒精度的13位时间戳，一般与eventTime值一致。
	ddl	该字段只在更改数据库的表结构时才会填充数据，更改数据（包括新增、删除和修改）时对应的ddl直接填充为null。 ■ text：String类型，数据库DDL语句文本。 ■ ddlMeta：String类型，使用FastSQL对DDL进行解析后生成的SQLStatement Object进行序列化的二进制表示，并使用Base64编码为String存储。 开启ddl支持时，需要传递的SQLStatement序列化对象，下游链路反序列化解析对象后，还原成目标数据源的ddl语句做变更。
version	无	格式的版本号。

- Blob序列化说明

  本文定义的JSON格式，一条消息对应一个JSONObject，JSONObject内部按照消息格式，逐级映射为相应的格式（JSONObject，JSONArray，相应类型的value等）。

  整个JSONObject中每个字段的存放类型均按照上述字段说明。序列化将JSONObject转换为String（如fastJSON的toJSONString方法）然后再采用String的getBytes(Charsets.UTF_8)方法，指定UTF_8字符集转化为byte[]。

相关消息的JSON样例

- Insert：

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "comment",
                "type": "STRING"
            }
        ],
        "source": {
            "dbName": "yunshi_db",
            "dbType": "MySQL",
            "tableName": "t_shiyu_pk"
        },
        "primaryKey": [
            "id",
            "name"
        ]
    },
    "payload": {
        "op": "INSERT",
        "after": {
            "dataColumn": {
                "name": "joe",
                "comment": "comment",
                "id": 1
            }
        },
        "sequenceId": "1605339516000000004",
        "timestamp": {
            "eventTime": 1605339932000,
            "systemTime": 1605339932736,
            "checkpointTime": 1605339932000
        }
    },
    "version": "0.0.1"
}
```

- update before:

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "comment",
                "type": "STRING"
            }
        ],
        "source": {
            "dbName": "yunshi_db",
            "dbType": "MySQL",
            "tableName": "t_shiyu_pk"
        },
        "primaryKey": [
            "id",
            "name"
        ]
    },
    "payload": {
        "op": "UPDATE_BEFOR",
        "before": {
            "dataColumn": {
                "name": "joe",
                "comment": "comment",
                "id": 1
            }
        },
        "sequenceId": "1605339516000000005",
        "timestamp": {
            "eventTime": 1605339934000,
            "systemTime": 1605339934951,
            "checkpointTime": 1605339934000
        }
    },
    "version": "0.0.1"
}
```

- update after:

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "comment",
                "type": "STRING"
            }
        ],
        "source": {
            "dbName": "yunshi_db",
            "dbType": "MySQL",
            "tableName": "t_shiyu_pk"
        },
        "primaryKey": [
            "id",
            "name"
        ]
    },
    "payload": {
        "op": "UPDATE_AFTER",
        "after": {
            "dataColumn": {
                "name": "joe",
                "comment": "com1",
                "id": 1
            }
        },
        "sequenceId": "1605339516000000005",
        "timestamp": {
            "eventTime": 1605339934000,
            "systemTime": 1605339934951,
            "checkpointTime": 1605339934000
        }
    },
    "version": "0.0.1"
}
```

- delete:

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "comment",
                "type": "STRING"
            }
        ],
        "source": {
            "dbName": "yunshi_db",
            "dbType": "MySQL",
            "tableName": "t_shiyu_pk"
        },
        "primaryKey": [
            "id",
            "name"
        ]
    },
    "payload": {
        "op": "DELETE",
        "before": {
            "dataColumn": {
                "name": "joe",
                "comment": "com1",
                "id": 1
            }
        },
        "sequenceId": "1605339516000000006",
        "timestamp": {
            "eventTime": 1605339937000,
            "systemTime": 1605339937671,
            "checkpointTime": 1605339937000
        }
    },
    "version": "0.0.1"
}
```

- Heartbeat：

```
{
    "schema": {},
    "payload": {
        "op": "MHEARTBEAT",
        "timestamp": {
            "eventTime": 1605339953629,
            "checkpointTime": 1605339953629
        }
    },
    "version": "0.0.1"
}
```

- DDL：

```
{
    "schema": {
        "source": {
            "dbName": "yunshi_db",
```

```
            "dbType": "MySQL",
            "tableName": "t_shiyu_nopk"
        }
    },
    "payload": {
        "op": "ALTER",
        "sequenceId": "1605339516000000035",
        "ddl": {
            "text": "alter table t_shiyu_nopk add column holo text",
            "ddlMeta": "rO0ABXNyACljb20uYWxpYmFiYS5kaS5wbHVnaW4uY2VudGVyLm1ldGEuRERMTWV0YQLb5Cx/YWXtAgACTAA
HZGRsVGV4dHQAEkxqYXZhL2xhbmcvU3RyaW5nO0wACXN0YXRlbWVudHQAKkxjb20vYWxpYmFiYS9mYXN0c3FsL3NxbC9hc3QvU1FMU3RhdG
VtZW50O3hwdAAtYWx0ZXIgdGFibGUgdF9zaGl5dV9ub3BrIGFkZCBjb2x1bW4gaG9sbyB0ZXh0c3IAPGNvbS5hbGliYWJhLmZhc3RzcWwuc
3FsLmFzdC5zdGF0ZW1lbnQuU1FMQWx0ZXJUYWJsZVN0YXRlbWVudBQPP3vMUl2cAgAPSQAHYnVja2V0c1oABmlnbm9yZVoAF2ludmFsaWRh
dGVHbG9iYWxJbmRleGVzWgAPbWVyZ2VTbWFsbEZpbGVzWgAHb2ZmbGluZVoADnJlbW92ZVBhdGl0aW5nWgATdXBkYXRlR2x
vYmFsSW5kZXhlc1oAD3VwZ3JhZGVQYXRpdGluZ0wAC2NsdXN0ZXJlZEJ5dAAQTGphdmEvdXRpbC9MaXN0OO0wABWl0ZW1zcQB+AAZMAAlwYX
J0aXRpb250ACxMY29tL2FsaWJhYmEvZmFzdHNxbC9zcWwvYXN0L1NRTFBhcnRpdGlvbkJ5O0wACHNvcnRlZEJ5cQB+AAZMAAx0YWJsZU9wd
GlvbnNxAH4ABkwAC3RhYmxlU291cmNldAA6TGNvbS9hbGliYWJhL2Zhc3RzcWwvc3FsL2FzdC9zdGF0ZW1lbnQvU1FMRXhwclRhYmxlU291
cmNlO3hyACxjb20uYWxpYmFiYS5mYXN0c3FsLnNxbC5hc3QuU1FMU3RhdGVtZW50SW1wbEOxUUDVCJMGAgADWgAJYWZ0ZXJTZW1pTAAGZGJ
UeXBldAAcTGNvbS9hbGliYWJhL2Zhc3RzcWwvRGJUeXBlO0wACWhlYWRIaW50c3EAfgAGeHIAKWNvbS5hbGliYWJhLmZhc3RzcWwuc3FsLm
FzdC5TUUxPYmplY3RJbXBs5LvqLFggFVECAAVJAAxzb3VyY2VDb2x1bW5JAApzb3VyY2VMaW5lTAAKYXR0cmlidXRlc3QAD0xqYXZhL3V0a
WwvTWFwO0wABGhpbnR0ACxMY29tL2FsaWJhYmEvZmFzdHNxbC9zcWwvYXN0L1NRTENvbW1lbnRIaW50O0wABnBhcmVudHQAJ0xjb20vYWxp
YmFiYS9mYXN0c3FsL3NxbC9hc3QvU1FMT2JqZWN0O3hwAAAAAAAAAABwcHAAfnIAGmNvbS5hbGliYWJhLmZhc3RzcWwuRGJUeXBlAAAAAAAA
AAAASAAB4cgAOamF2YS5sYW5nLkVudW0AAAAAAAAAABIAAHhwdAAFbXlzcWxwAAAAAAAAAAAAAAAAAAc3IAE2phdmEudXRpbC5BcnJheUxpc3
R4gdIdmcdhnQMAAUkABHNpemV4cAAAAAB3BAAAAAB4c3EAfgAUAAAAAXcEAAAAAXNyADxjb20uYWxpYmFiYS5mYXN0c3FsLnNxbC5hc3Quc
3RhdGVtZW50LlNRTEFsdGVyVGFibGVBZGRDb2x1bW41lT6CFe//BAIABloAB2Nhc2NhZGVaAAVmaXJzdEwAC2FmdGVyQ29sdW1udAAlTGNv
bS9hbGliYWJhL2Zhc3RzcWwvc3FsL2FzdC9TUUxOYW1lO0wAB2NvbHVtbnNxAH4ABkwAC2ZpcnN0Q29sdW1ucQB+ABhMAAhyZXN0cmljdHQ
AE0xqYXZhL2xhbmcvQm9vbGVhbjt4cQB+AAsAAAAAAAAAAHBwcQB+AAhMAAhyZXN0cmljdHQAE0xqYXZhL2xhbmcvQm9vbGVhbjt4cQB+AB5MAAdjb21tZW50cQB+AB5MAAtjb2
1wcmVzc2lvbnQALkxjb20vYWxpYmFiYS9mYXN0c3FsL3NxbC9hc3QvZXhwci9TUUxDaGFyRXhwcjtMAAtjb25zdHJhaW50c3EAfgAGTAAIZ
GF0YVR5cGV0ACllMY29tL2FsaWJhYmEvZmFzdHNxbC9zcWwvYXN0L1NRTERhdGFUeXBlO0wABmRiVHlwZXEAfgAKTAALZGVmYXVsdEV4cHJx
AH4AHkwACWRlbGltaXRlcnEAfgAeTAASZGVsaW1pdGVyVG9rZW5pemVycQB+AB5MAAZlbmFibGVxAH4AGUwABmVuY29kZXEAfgAfTAAGZm9
ybWF0cQB+AB5MABBnZW5lcmF0ZWRBbGF3c0FzcQB+AB5MAAhpZGVudGl0eXQARExjb20vYWxpYmFiYS9mYXN0c3FsL3NxbC9hc3Qvc3RhdG
VtZW50L1NRTENvbHVtbkRlZmluaXRpb24kSWRlbnRpdHk7TAASanNvbkluZGV4QXR0cnNFeHBycQB+AB5MAAhtYXBwZWRCeXEAfgAGTAAEb
mFtZXEAfgAYTAAMbmxwVG9rZW5pemVycQB+AB5MAAhvblVwZGF0ZXEAfgAeTAAEcmVseXEAfgAZTAAMc2VxdWVuY2VUeXBldAAvTGNvbS9h
bGliYWJhL2Zhc3RzcWwvc3FsL2FzdC9BdXRvSW5jcmVtZW50VHlwZTtMAARzdGVwcQB+AB5MAAdzdG9yYWdlcQB+AB5MAAl1bml0Q291bnR
xAH4AHkwACXVuaXRJbmRleHEAfgAeTAAIdmFsaWRhdGVxAH4AGUwACXZhbHVlVHlwZXEAfgAeeHEAfgALAAAAAAAAAABwcHEAfgAaAAAAAA
AAAAAAAHBwcHBwcHBzcQB+ABQAAAAAdwQAAAAAeHNyADpjb20uYWxpYmFiYS5mYXN0c3FsLnNxbC5hc3Quc3RhdGVtZW50LlNRTENoYXJhY
3RlckRhdGFUeXBlqtJac/d+04cCAAVaAAloYXNCaW5hcnlMAAtjaGFyU2V0TmFtZXEAfgABTAAIY2hhclR5cGVxAH4AAUwABWhpbnRzcQB+AAZ4cgArY29tLmFsaWJhYmEuZmFzdHNxbC5zcWwuYXN0LlNRTERhdGFUeXBlSW1wbEWL29pc1gZFAgAJSgAObmF
tZUhhc2hDb2RlNjRaAAh1bnNpZ25lZFoAEXdpdGhMb2NhbFRpbWVab25lWgAIemVyb2ZpbGxMAAlhcmd1bWVudHNxAH4ABkwABmRiVHlwZX
EAfgAKTAAHaW5kZXhCeXEAfgAeTAAEbmFtZXEAfgABTAAMd2l0aFRpbWVab25lcQB+ABl4cQB+AAsAAAAAAAAAAHBwcQB+ACP6BPTvGZVAf
gAAAHNxAH4AFAAAAAB3BAAAAAB4cHB0AAR0ZXh0cABwcHBwcQB+ABJwcHBwcHBwc3IAMmNvbS5hbGliYWJhLmZhc3RzcWwuc3FsLmFz
dC5leHByLlNRTElkZW50aWZpZXJFeHBy3DXH1zvWbgkCAARKAApoYXNoQ29kZTY0TAAEbmFtZXEAfgABTAAOcmVzb2x2ZWRDb2x1bW5xAH4
ADkwAE3Jlc29sdmVkT3duZXJPYmplY3RxAH4ADnhyACdjb20uYWxpYmFiYS5mYXN0c3FsLnNxbC5hc3QuU1FMRXhwckltcGxs2ypmFJxWrQ
IAAHhxAH4ACwAAAAAAAAAcHBwQCnxzH5tIDl0AARob2xvcHBwcHBwcHBweHBweHBzcQB+ABQAAAAAdwQAAAAeHNxAH4AFAAAAAB3B
AAAAAB4c3IAOGNvbS5hbGliYWJhLmZhc3RzcWwuc3FsLmFzdC5zdGF0ZW1lbnQuU1FMRXhwclRhYmxlU291cmNlRHD7eYJ4eswCAAVMAAdj
b2x1bW5zcQB+AAZMAAR1eHBycQB+AB5MAApwYXJ0aXRpb25zcQB+AAZMAAhzYW1wbGluZ3QAOExjb20vYWxpYmFiYS9mYXN0c3FsL3NxbC9
hc3Qvc3RhdGVtZW50L1NRTFRhYmxlU2FtcGxpbmc7TAAMc2NoZW1hT2JqZWN0dAAxTGNvbS9hbGliYWJhL2Zhc3RzcWwvc3FsL3JlcG9zaX
RvcnkvU2NoZW1hT2JqZWN0O3hyADhjb20uYWxpYmFiYS5mYXN0c3FsLnNxbC5hc3Quc3RhdGVtZW50LlNRTFRhYmxlU291cmNlSW1wbAqEM
enTm5zUAgAESgAPYWxpYXNIYXNoQ29kZTY0TAAFYWxpYXNxAH4AAUwACWZsYXNoYmFja3EAfgAeTAAFaGludHNxAH4ABnhxAH4ACwAAAAAA
AAAcHBwAAAAAAAAABwcHBwc3EAfgAqAAAAAAAAAABwcHEAfgA0NH7o4UvP9Dt0AAx0X3NoaXllX25vcGtwcHBwcA==="
        },
        "timestamp": {
            "eventTime": 1605342109000,
            "systemTime": 1605342109259,
            "checkpointTime": 1605342109000
        }
    },
    "version": "0.0.1"
}
```

4.8. 同步整库数据至Kafka

4.8.1. 资源规划与配置

当前使用DataWorks的实时数据同步任务同步数据时，仅支持使用独享数据集成资源组。本文为您介绍使用实时数据同步任务同步数据时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用实时数据同步任务同步数据时，当前仅支持使用独享数据集成资源组。因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续同步任务使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表** > **独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

   > **说明** 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC

   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

   ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

   > **说明**
   > - 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
   > - 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

   如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

> **说明** 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的实时数据同步任务。目前同步数据至Kafka的来源数据源仅支持MySQL，数据源的配置可参见配置数据源（来源为MySQL）。

4.8.2. 配置数据源（来源为MySQL）

实时同步MySQL的数据至Kafka时，来源数据源为MySQL，去向数据源为Kafka，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源Kafka。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

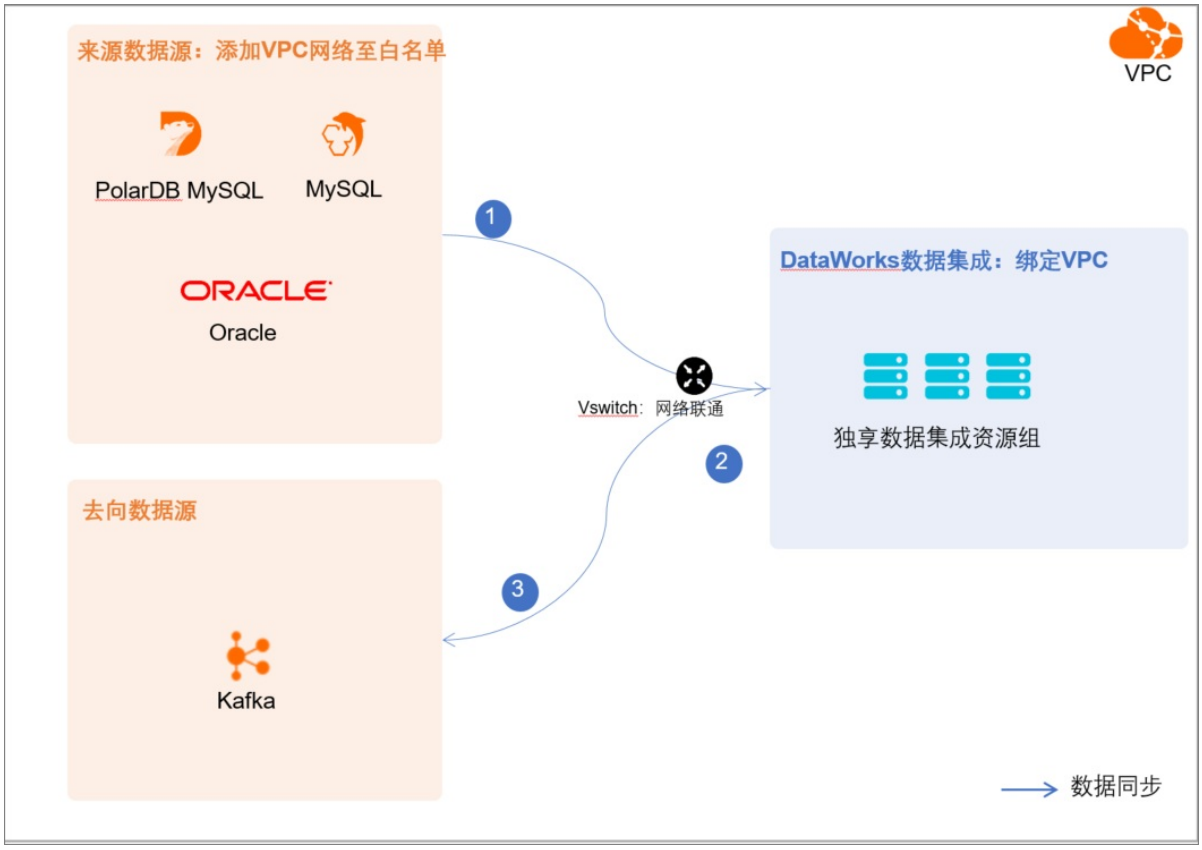
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

4.8.3. 添加数据源

将来源数据源的数据同步至去向数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加去向数据源：Kafka

操作详情可参见配置Kafka数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置并管理实时同步任务。

4.8.4. 配置并管理实时同步任务

完成数据源、网络、资源的准备配置后，您可创建实时同步节点，同步数据至Kafka。本文为您介绍如何创建数据实时同步任务，并在创建完成后查看任务运行情况。

前提条件

创建实时数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为MySQL）
- 添加数据源

使用限制

- 实时数据同步任务仅支持使用独享数据集成资源组。
- 实时同步节点目前仅支持同步MySQL、Oracle和PolarDB数据源至Kafka。

注意事项

- 对于源端同步表有主键的场景，同步时会使用主键值作为kafka记录的key，确保同主键的变更有序写入kafka的同一分区。
- 对于源端同步表无主键的场景，如果选择了支持无主键表同步选项，则同步时kafka记录的key为空。如果要确保表的变更有序写入kafka，则选择写入的kafka topic必须是单分区。如果选择了自定义同步主键，则同步时使用其他非主键的一个或几个字段的联合，代替主键作为kafka记录的key。
- 如果在kafka集群发生响应异常的情况下，仍要确保有主键表同主键的变更有序写入kafka的同一分区，则需要在配置kafka数据源时，在扩展参数表单中加入如下配置。

  `{"max.in.flight.requests.per.connection":1,"buffer.memory": 100554432}` 。

  > **注意** 添加配置后同步性能会大幅下降，需要在性能和严格保序可靠性之间做好权衡。

- 实时同步写入kafka的消息总体格式、同步任务心跳消息格式及源端更改数据对应的消息格式，详情请参见：附录：消息格式。

创建实时同步任务

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 创建**业务流程**。

   如果您已有**业务流程**，则可以忽略该步骤。

   i. 鼠标悬停至**+新建**图标，选择**新建业务流程**。

   ii. 在**新建业务流程**对话框，输入**业务名称**。

   iii. 单击**新建**。
5. 创建实时同步节点。

   i. 鼠标悬停至**+新建**图标，选择**数据集成 > 实时同步**。

   您也可以找到目标业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。

ii. 在**新建节点**对话框中，配置各项参数。

参数	描述
节点类型	默认为**实时同步**。
同步方式	选择**整库实时同步至Kafka**，用于迁移目标数据库下的部分或所有表至Kafka中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放实时同步节点的目录。

iii. 单击**提交**，进入实时同步任务编辑页面。

6. 选择资源组。

i. 在实时同步任务编辑页面的右侧导航栏，单击**基本配置**。

ii. 在**资源组**下拉框，选择需要使用的资源组。

> **说明**
>
> 实时数据同步任务仅支持使用独享数据集成资源组。
>
> 如果您没有可用的独享数据集成资源组，请单击**新建独享资源组**创建。详情请参见独享数据集成资源组概述。

7. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**、**数据源**及**编码**格式。

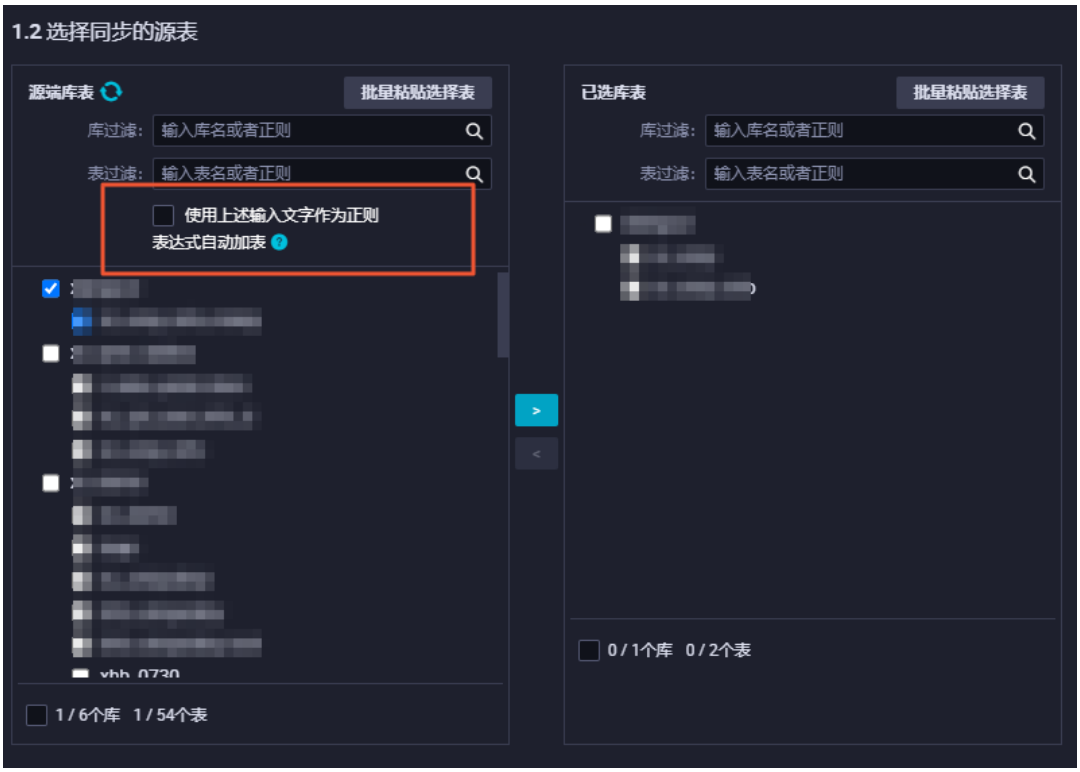
> **说明** 仅支持选择MySQL、Oracle和PolarDB数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

iii. 在**择同步的源表**区域搜索框下方，勾选**使用上述输入文字作为正则表达式自动加表**复选框，仅MySQL支持通过正则表达自动加表。

勾选后可以在源端库表的**库过滤**搜索框和**表过滤**搜索框中填写正则表达式，实时同步运行过程中会自动识别源端数据库binlog中数据库名和表名符合正则表达式的变更，识别到这类变更后按照下面设置的**表名到Topic的映射规则**计算源端变更应该写入的**kafka topic**名称，如果kafka topic尚未建立，则会自动创建对应topic后将变更写入。



库名及表名支持的正则规则如下所示。

正则表达式	描述
.	匹配除换行符以外的所有字符
x?	匹配0次或一次x字符串
x*	匹配0次或多次x字符串,但匹配可能的最少次数
x+	匹配1次或多次x字符串,但匹配可能的最少次数
.*	匹配0次或多次的任何字符
.+	匹配1次或多次的任何字符
{m}	匹配刚好是m个的指定字符串
{m,n}	匹配在m个以上n个以下的指定字符串
{m,}	匹配m个以上的指定字符串
[]	匹配符合 [] 内的字符
[^]	匹配不符合 [] 内的字符
\d	匹配一个数字的字符,和 [0-9] 语法一样
\d+	匹配多个数字字符串,和 [0-9]+ 语法一样
\D	非数字,其他同 \d
\D+	非数字,其他同 \d+
\w	英文字母或数字的字符串,和 [a-zA-Z0-9_] 语法一样
\w+	和 [a-zA-Z0-9_]+ 语法一样
\W	非英文字母或数字的字符串,和 [^a-zA-Z0-9_] 语法一样
\W+	和 [^a-zA-Z0-9_]+ 语法一样
\s	空格,和 [\n\t\r\f] 语法一样
\s+	和 [\n\t\r\f]+ 一样
\S	非空格,和 [^\n\t\r\f] 语法一样
\S+	和 [^\n\t\r\f]+ 语法一样
\b	匹配以英文字母,数字为边界的字符串
\B	匹配不以英文字母,数值为边界的字符串
(a|b|c)	匹配符合a字符或是b字符或是c字符的字符串，注意或表达式必须用括号包围，否则将导致运行时错误
\	对"$^[].?+|*{}"具有特殊含义的正则规则字符转义

iv. 在**设置表名到Topic的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

映射规则包括**源表名和目标Topic转换规则**和**目标Topic规则**：

- **源表名和目标Topic转换规则**：转换表名为目标Topic，进行字符串替换。
- **目标Topic规则**：支持对转换后的Topic添加前缀和后缀。

v. 单击**下一步**。

8. 选择目标数据源并配置目标Topic格式。

i. 在**设置目标Topic**页面，选择**目标Kafka数据源**。

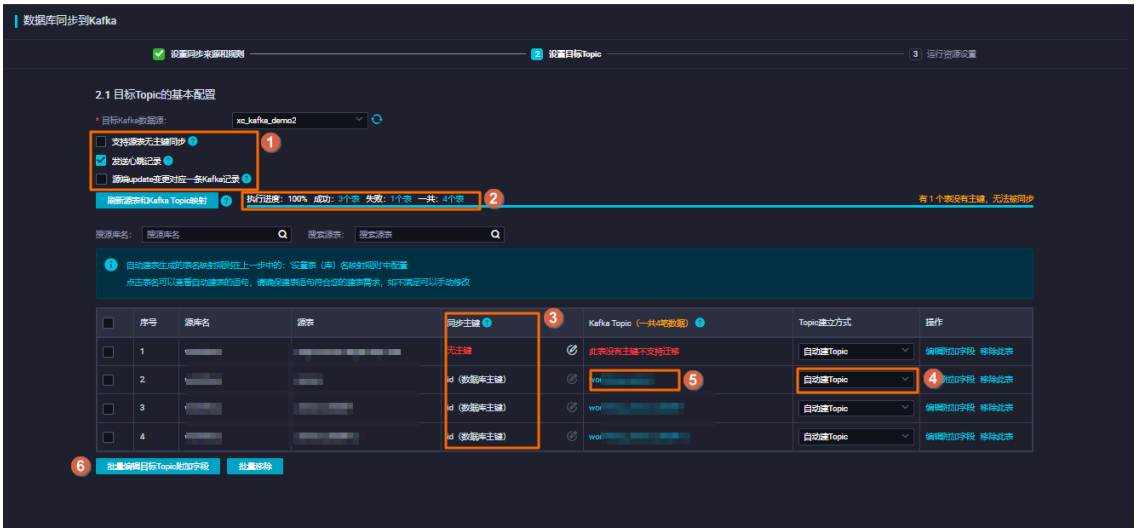
- 如果您需要同步无主键的来源表，则可以勾选**支持源表无主键同步**。
- 如果您希望消费Kafka中数据时能够准确判断同步任务进度，则可以勾选**发送心跳记录**，同步任务时将定时产生op字段作为**MHEARTBEAT**的同步任务心跳记录写入Kafka的topic中，心跳记录具体格式请参考附录：消息格式。

ii. （可选）目标表新增字段。

如果您希望为目标表中的所有同步表新增统一的字段，则可以在**目标表附加常量字段**区域，单击**新增字段**添加。

iii. 单击**刷新源表和Kafka Topic映射**，创建需要同步的源表和目标Kafka Topic的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	■ 勾选**支持源表无主键同步**后，源表没有主键，也可以向下游同步，但是同步数据时kafka记录的key将使用空值，只有当写入的kafka topic是单分区，才能确保变更有序写入。 ■ 勾选**发送心跳记录**后，实时同步任务将每隔5秒往Kafka中写入一条带有当前时间信息的记录。这样即使源端没有读取到新数据，Kafka中最新数据的时间信息也会持续更新，您可以根据Kafka中读取到的最新数据的时间判断实时同步的进度。 ■ 勾选**源端update变更对应一条Kafka记录**后，源端关系型数据库一条记录的一次update变更，变更前和变更后的数据将保存在一条Kafka记录中；如果未勾选，源端关系型数据库一条记录的一次update变更，将保存在两条Kafka记录中，分别保存变更前和变更后的数据。写入Kafka消息的消息结构及各字段含义详情请参见：附录：消息格式。
②	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。

序号	描述
③	■ 如果来源库有主键，则同步数据时会使用该主键值作为kafka记录的key，确保同主键的变更有序写入kafka的同一分区。 ■ 如果来源库没有主键： ■ 当在**设置目标Topic**页面勾选了**支持源表无主键同步**，则无主键的表可以正常同步。此时写入kafka记录的key将使用空值，只有当写入的kafka topic是单分区，才能确保变更有序写入，此外，您还可以选择单击图标自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键作为kafka记录的key。 ■ 当在**设置目标Topic**页面未勾选**支持源表无主键同步**，则无主键的表同步时会出现异常，您需要在同步任务中删除无主键的表或者选择单击图标自定义主键才能继续执行同步任务。
④	包括**使用已有Topic**和**自动建Topic**。
⑤	选择的**Topic建立方式**，取值如下： ■ 当**Topic建立方式**选择**使用已有Topic**时，您可以在Kafka Topic列的下拉列表中选择需要使用的Topic名称。 ■ 当**Topic建立方式**选择**自动建Topic**时，显示自动创建的Kafka Topic名称。您可以单击Topic名称，查看和修改建Topic名称和注释。
⑥	在**批量编辑目标Topic附加字段**表单中给目标Kafka Topic增加字段。也可以单击**操作**列的**编辑附加字段**进行单表附加字段的设置。 **说明** 批量修改仅针对**Topic建立方式**选择**自动建Topic**的Topic生效。

v. 单击**下一步**。

如果您前一步中目标数据源使用的**Topic建立方式**为**自动建Topic**，则需要在弹出的**自动建表**对话框，单击**开始建表**，批量创建目标Kafka Topic。

9. 运行资源设置。

i. 在**运行资源设置**页面，配置各项参数。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为**15**。
目标端写入并发数	数据同步任务内，可以从来源表并行读取或写入数据至目标端的最大线程数。最大并发数为32。请根据您的资源组大小和目标端实际规模合理设置。

ii. 单击**完成配置**。

提交并发布实时同步任务

1. 提交并发布节点任务。

i. 单击工具栏中的图标，保存节点。

ii. 单击工具栏中的图标，提交节点任务。

iii. 在**提交新版本**对话框中，输入**变更描述**。

iv. 单击**确定**。

如果您使用的是标准模式的工作空间，任务提交成功后，需要将任务发布至生产环境进行发布。请单击顶部菜单栏左侧的**任务发布**。具体操作请参见发布任务。

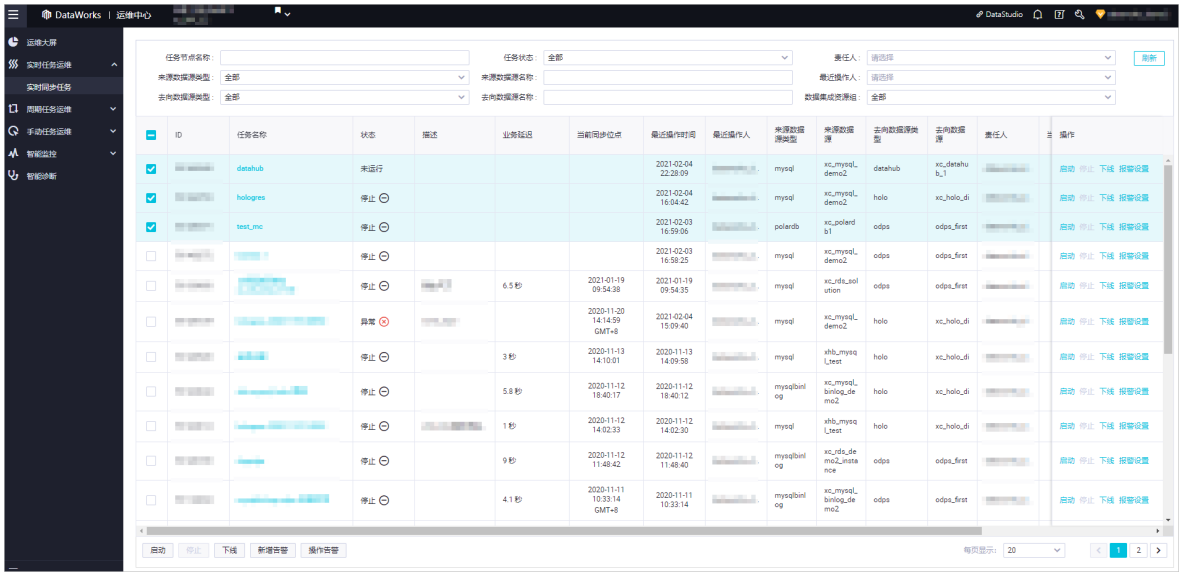
执行实时同步任务

1. 进入运维页面。

   提交或发布节点成功后，单击节点编辑页面右上方的**运维中心**，进入**实时任务运维 > 实时同步任务**页面。

2. 查看实时同步任务详情。

   在**实时同步任务**页面，单击相应**任务名称**，查看运维任务的详细信息。



3. 执行实时同步任务。

   i. 单击目标实时同步任务**操作**列的**启动**。

ii. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点任务的日期和时间。
时区	从**时区**下拉列表中选择执行任务的时区。
Failover	您可以设置在固定时间内，任务的Failover超过指定次数时，自动结束任务。 **说明** 如果您不配置Failover的次数，将根据5分钟Failover超过100次来自动结束任务，避免频繁启动任务占用系统资源。
脏数据策略	■ **零容忍，不允许**：只要同步任务中包含脏数据，则任务自动结束。 ■ **不限制**：无论同步任务中是否包含脏数据，任务均可正常执行。 ■ **有限控制**：指定可包含固定数值的脏数据，超出该数值时任务自动结束。

iii. 单击**确定**。

管理实时同步任务

- 停止运行中的任务。

  单击相应任务后的**停止**。在停止对话框中，单击**停止**。
- 下线非运行状态的任务。

  单击相应任务后的**下线**。在下线对话框中，单击**下线**。
- 查看任务的报警信息。

  单击相应任务后的**报警设置**，在**报警设置**页面查看**报警事件**及**报警规则**。
- 为任务新增告警。

  i. 选中需要新增告警的任务，单击**实时同步任务**页面下方的**新增告警**。

  ii. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称。
描述	新建规则的描述信息。

参数	描述
指标	产生报警的指标项： ■ **任务状态** ■ **业务延迟** ■ **Failover** ■ **脏数据** ■ **DDL错误**
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。
报警间隔	设置报警的时间间隔，默认为5分钟发一次报警。
WARNING	产生相应报警时，可以选择通过**邮件**、**短信**和**钉钉**发送报警通知。 **说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
CRITICAL	
接收人（非钉钉）	选择报警通知的接收人。

iii. 单击**确定**。

- 批量修改目标任务中指定类型的所有告警。
  i. 选中需要操作告警的任务，单击**实时同步任务**页面下方的**操作告警**。
  ii. 在**操作告警**对话框中，选则需要修改的**操作类型**和**告警指标**。
  iii. 单击**确定**。

后续步骤

完成配置实时同步任务的操作后，执行同步任务会将源端数据库读取的数据，以JSON格式写入到Kafka topic中，您可以通过附录：消息格式获取写入Kafka的消息的状态及变更等信息。

4.8.5. 附录：消息格式

本文介绍写入Kafka消息的消息结构及各字段含义。

背景信息

同步整库数据至kafka任务，是将从上游数据源读取的数据，按照下面描述的JSON格式写入到Kafka的topic。消息总体格式包括变更记录的列信息、以及数据变更前后的状态信息等。为确保消费Kafka中数据时能够准确判断同步任务进度，同步任务还将定时产生op字段作为MHEARTBEAT的同步任务心跳记录写入Kafka的topic中。以下为您介绍写入Kafka的消息总体格式、同步任务心跳消息格式及源端更改数据对应的消息格式，关于字段类型及参数说明等信息，详情请参见字段类型和参数说明。

消息总体格式

写入Kafka消息的总体格式如下所示：

```
{
    "schema": { //变更的元数据信息，仅指定列名与列类型信息
        "dataColumn": [//变更的数据列信息，更新目标表记录内容
            {
                "name": "id",
                "type": "LONG"
            },
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "binData",
```

```
                "type": "BYTES"
            },
            {
                "name": "ts",
                "type": "DATE"
            },
            {
              "name":"rowid",// 数据源为Oracle时，rowid会放在数据列中
              "type":"STRING"
            }
        ],
        "primaryKey": [
            "pkName1",
            "pkName2"
        ],
        "source": {
            "dbType": "mysql",
            "dbVersion": "1.0.0",
            "dbName": "myDatabase",
            "schemaName": "mySchema",
            "tableName": "tableName"
        }
    },
    "payload": {
        "before": {
            "dataColumn":{
                "id": 111,
                "name":"scooter",
                "binData": "[base64 string]",
                "ts": 1590315269000,
                "rowid": "AAIUMPAAFAACxExAAE"//字符串类型,Oracle的rowid信息
            }
        },
        "after": {
            "dataColumn":{
                "id": 222,
                "name":"donald",
                "binData": "[base64 string]",
                "ts": 1590315269000,
                "rowid": "AAIUMPAAFAACxExAAE"//字符串类型,Oracle的rowid信息
            }
        },
        "sequenceId":"XXX",//字符串类型，用于增全量数据合并的数据排序,
        "scn":"xxxx",//字符串类型,Oracle的scn信息
        "op": "INSERT/UPDATE_BEFOR/UPDATE_AFTER/DELETE/TRANSACTION_BEGIN/TRANSACTION_END/CREATE/ALTER/ERASE/Q
UERY/TRUNCATE/RENAME/CINDEX/DINDEX/GTID/XACOMMIT/XAROLLBACK/MHEARTBEAT...",//大小写敏感,
        "timestamp": {
            "eventTime": 1,//必选，记录源端库发生变更的时间，毫秒精度的13位时间戳
            "systemTime": 2,//可选，同步任务处理该条变更消息的时间，毫秒精度的13位时间戳
            "checkpointTime": 3//可选，重置同步位点时的设置时间，毫秒精度的13位时间戳，一般等于eventTime
        },
        "ddl": {
            "text": "ADD COLUMN ...",
            "ddlMeta": "[SQLStatement serialized binary, expressed in base64 string]"
        }
    },
    "version":"1.0.0"
}
```

同步任务心跳消息格式

```
{
    "schema": {
        "dataColumn": null,
        "primaryKey": null,
        "source": null
    },
    "payload": {
        "before": null,
        "after": null,
        "sequenceId": null,
        "timestamp": {
            "eventTime": 1620457659000,
            "checkpointTime": 1620457659000
        },
        "op": "MHEARTBEAT",
        "ddl": null
    },
    "version": "0.0.1"
}
```

源端更改数据对应的消息格式

- 源端插入数据对应的Kafka消息格式：

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "job",
                "type": "STRING"
            },
            {
                "name": "sex",
                "type": "STRING"
            },
            {
                "name": "#alibaba_rds_row_id#",
                "type": "LONG"
            }
        ],
        "primaryKey": null,
        "source": {
            "dbType": "MySQL",
            "dbName": "pkset_test",
            "tableName": "pkset_test_no_pk"
        }
    },
    "payload": {
        "before": null,
        "after": {
            "dataColumn": {
                "name": "name11",
                "job": "job11",
                "sex": "man",
                "#alibaba_rds_row_id#": 15
            }
        },
        "sequenceId": "1620457642589000000",
        "timestamp": {
            "eventTime": 1620457896000,
            "systemTime": 1620457896977,
            "checkpointTime": 1620457896000
        },
        "op": "INSERT",
        "ddl": null
    },
    "version": "0.0.1"
}
```

- 源端更新数据对应的Kafka消息格式：

  当未勾选**源端update变更对应一条Kafka记录**时，源端更新数据对应的Kafka消息格式包含两条Kafka消息，分别描述更新前的数据状态和更新后的数据状态。具体消息格式如下：

- 更新前的数据状态消息格式：

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "job",
                "type": "STRING"
            },
            {
                "name": "sex",
                "type": "STRING"
            },
            {
                "name": "#alibaba_rds_row_id#",
                "type": "LONG"
            }
        ],
        "primaryKey": null,
        "source": {
            "dbType": "MySQL",
            "dbName": "pkset_test",
            "tableName": "pkset_test_no_pk"
        }
    },
    "payload": {
        "before": {
            "dataColumn": {
                "name": "name11",
                "job": "job11",
                "sex": "man",
                "#alibaba_rds_row_id#": 15
            }
        },
        "after": null,
        "sequenceId": "16204576425890000001",
        "timestamp": {
            "eventTime": 1620458077000,
            "systemTime": 1620458077779,
            "checkpointTime": 1620458077000
        },
        "op": "UPDATE_BEFOR",
        "ddl": null
    },
    "version": "0.0.1"
}
```

- 更新后的数据状态消息格式：

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "job",
                "type": "STRING"
            },
            {
                "name": "sex",
                "type": "STRING"
            },
            {
                "name": "#alibaba_rds_row_id#",
                "type": "LONG"
            }
        ],
        "primaryKey": null,
        "source": {
            "dbType": "MySQL",
            "dbName": "pkset_test",
            "tableName": "pkset_test_no_pk"
        }
    },
    "payload": {
        "before": null,
        "after": {
            "dataColumn": {
                "name": "name11",
                "job": "job11",
                "sex": "woman",
                "#alibaba_rds_row_id#": 15
            }
        },
        "sequenceId": "16204576425890000001",
        "timestamp": {
            "eventTime": 1620458077000,
            "systemTime": 1620458077779,
            "checkpointTime": 1620458077000
        },
        "op": "UPDATE_AFTER",
        "ddl": null
    },
    "version": "0.0.1"
}
```

- 源端删除数据对应的Kafka消息格式：

```
{
    "schema": {
        "dataColumn": [
            {
                "name": "name",
                "type": "STRING"
            },
            {
                "name": "job",
                "type": "STRING"
            },
            {
                "name": "sex",
                "type": "STRING"
            },
            {
                "name": "#alibaba_rds_row_id#",
                "type": "LONG"
            }
        ],
        "primaryKey": null,
        "source": {
            "dbType": "MySQL",
            "dbName": "pkset_test",
            "tableName": "pkset_test_no_pk"
        }
    },
    "payload": {
        "before": {
            "dataColumn": {
                "name": "name11",
                "job": "job11",
                "sex": "woman",
                "#alibaba_rds_row_id#": 15
            }
        },
        "after": null,
        "sequenceId": "1620457642589000002",
        "timestamp": {
            "eventTime": 1620458266000,
            "systemTime": 1620458266101,
            "checkpointTime": 1620458266000
        },
        "op": "DELETE",
        "ddl": null
    },
    "version": "0.0.1"
}
```

字段类型

写入Kafka topic中的消息将从源端读取数据映射为BOOLEAN、DOUBLE、DATE、BYTES、LONG，STRING六种类型，再以不同的JSON格式写入kafka topic中。

类型	说明
BOOLEAN	对应JSON中的布尔类型，取值为true，false
DATE	对应JSON中的数值类型，取值为13位数字时间戳，精确到毫秒（ms）级。
BYTES	对应JSON中的字符串类型，写入Kafka前会先对字节数组进行base64编码转换为字符串，消费时需要进行base64解码（编码Base64.getEncoder().encodeToString(text.getBytes("UTF-8"))；解码Base64.getDecoder().decode(encodedText)）。

类型	说明
STRING	对应JSON中的字符串类型
LONG	对应JSON中的数值类型
DOUBLE	对应JSON中的数值类型

参数说明

以下为您介绍写入Kafka的消息中的各个字段的含义及说明。

一级元素	二级元素	说明
schema	dataColumn	JSONArray类型，数据列的类型信息。dataColumn记录上游数据变更记录的所有列和对应的列类型信息。变更操作包括数据库对数据的更改（新增、删除及修改）和数据库表结构等变更。 • name：列名 • type：列类型
	primaryKey	List类型，主键信息。 pk：主键名。
	source	Object 类型，源端数据库或表信息。 • dbType：String类型，数据库类型 • dbVersion：String类型，数据库版本 • dbName：String类型，数据库名 • schemaName：String类型，Schema名（针对Postgres和SQL Server等） • tableName：String 类型，数据表名
	before	JSONObject类型，修改前的数据。例如：数据源端为mysql，做了一次记录的update操作，before字段存储记录被update之前的数据内容。 • 在从源端读取到更新、删除操作消息时，在写入记录中填充该字段。 • dataColumn：JSONObject类型，表示数据信息。格式为列名：列值， 列名为字符串，列值BOOLEAN、DOUBLE、DATE、BYTES、LONG，STRING。
	after	修改后的数据。格式同before相同。
	sequenceId	字符串类型，Streamx产生，用于增量数据和全量数据合并的数据排序，每个streamx record都是唯一的。 **说明** 对于从源端读取的更新操作消息，会生成两条写入记录，一条update before记录和一条update after记录，这两条记录的sequenceId相同。
	scn	当源端为Oracle数据库时有效，对应Oracle的scn信息。

一级元素	二级元素	说明
payload	op	对应源端读取到的消息类型，取值如下： • INSERT：数据插入 • UPDATE_BEFOR：数据更新前 • UPDATE_AFTER：数据更新后 • DELETE：数据删除 • TRANSACTION_BEGIN：数据库事务开始 • TRANSACTION_END：数据库事务结束 • CREATE：数据库建表 • ALTER：数据库表变更 • QUERY：数据库变更的原始SQL • TRUNCATE：数据库表清空 • RENAME：数据库表重命名 • CINDEX：创建索引 • DINDEX：删除索引 • MHEARTBEAT：用于在源端无新增数据时标识同步仍正常进行的心跳消息
	timestamp	JSONObject 类型，本条数据的相关时间戳。 • eventTime：Long类型，记录源端库发生变更的时间，毫秒精度的13位时间戳。 • systemTime：Long类型，同步任务处理该条变更消息的时间，毫秒精度的13位时间戳。 • checkpointTime：Long类型，重置同步位点时的设置时间，毫秒精度的13位时间戳，一般与eventTime值一致。
	ddl	该字段只在更改数据库的表结构时才会填充数据，更改数据（包括新增、删除和修改）时对应的ddl直接填充为null。 • text：String类型，数据库DDL语句文本。 • ddlMeta：String类型，将数据库ddl类型变更记录到一个Java对象，使用对象序列化后再进行base64编码得到的字符串。
version	无	格式的版本号。

4.9. 创建、编辑、提交和运维实时同步节点

DataWorks支持实时同步数据，本文为您介绍如何创建、编辑、提交和运维实时同步节点。

使用限制

目前除以下地域外，其余地域均已开通实时同步能力。

- 马来西亚（吉隆坡）：建设中。
- 迪拜：未开通。
- 英国（伦敦）：未开通。

创建实时同步节点

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据开发**。
4. 鼠标悬停至+新建图标，单击**数据集成 > 实时同步**。

   您也可以找到相应的业务流程，右键单击**数据集成**，选择**新建 > 实时同步**。实时同步支持的数据源请参见实时同步支持的数据源。
5. 在**新建节点**对话框中，配置各项参数。

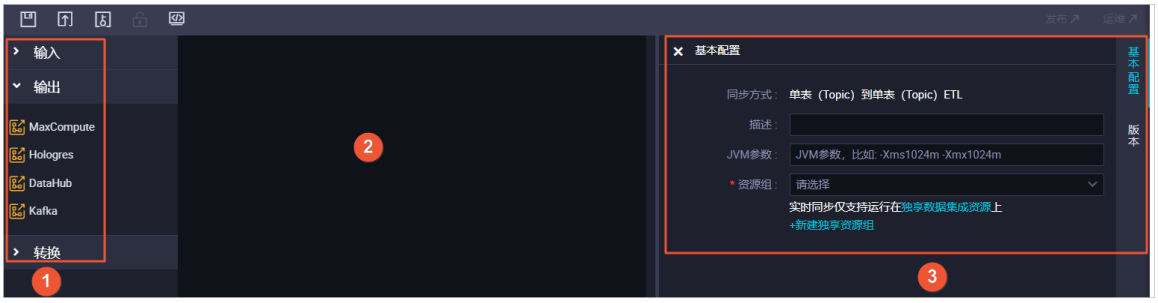
参数	描述
节点类型	默认为**实时同步**。
同步方式	包括**单表（Topic）到单表（Topic）ETL**、**数据库迁至Hologres**、**数据库迁至MaxCompute**和**数据库迁至DataHub**： ○ **单表（Topic）到单表（Topic）ETL**：实时同步单个表至一个或多个表中，支持同步过程中变换数据。 ○ **数据库迁至Hologres**：迁移一个整库下的所有或部分表至Hologres中，支持Hologres下自动创建目标表。 ○ **数据库迁至MaxCompute**：迁移一个整库下的所有或部分表至MaxCompute中。 ○ **数据库迁至DataHub**：迁移一个整库下的所有或部分Topic至DataHub中。
节点名称	节点名称必须是大小写字母、中文、数字、下划线（_）以及英文句号（.），且不能超过128个字符。
目标文件夹	存放节点的目录。

6. 单击**提交**。

编辑实时同步节点

选择不同的**同步方式**，实时同步节点的编辑页面也不同：

- 当选择**同步方式**为**单表（Topic）到单表（Topic）ETL**时，操作如下：
  i. 双击打开实时同步节点的编辑页面，单击右侧的**基本配置**，从**资源组**下拉列表中选择需要使用的资源组。

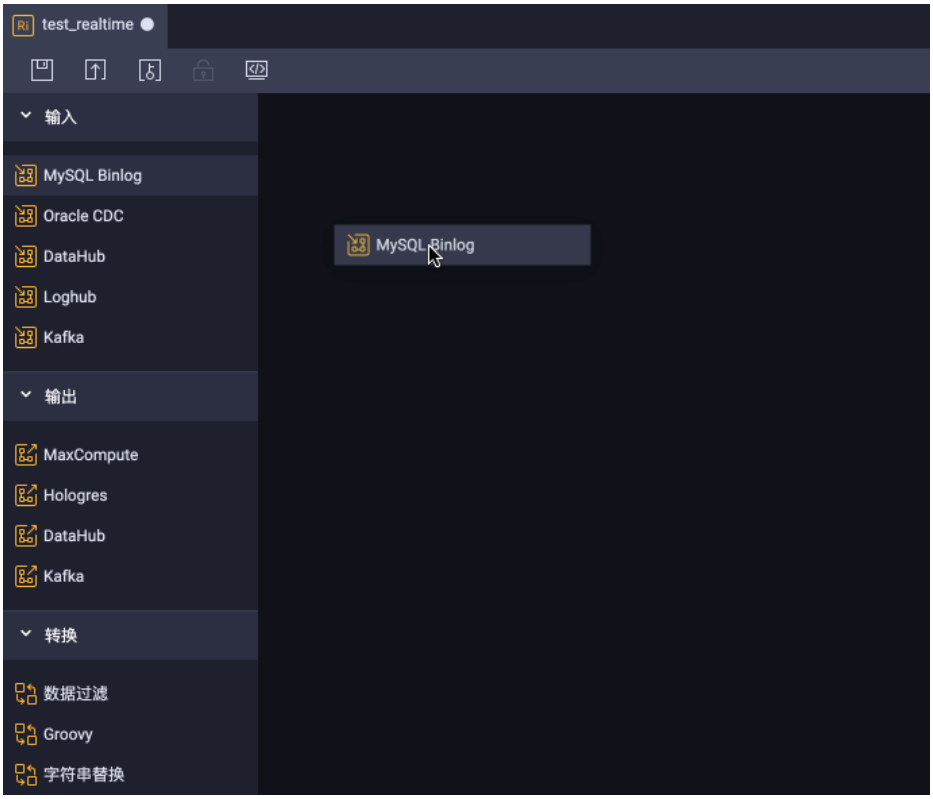


序号	描述
①	组件区域，包括**输入**、**输出**和**转换**三大模块。
②	节点的图形化编辑区域，您可以拖拽组件至该区域进行编辑。

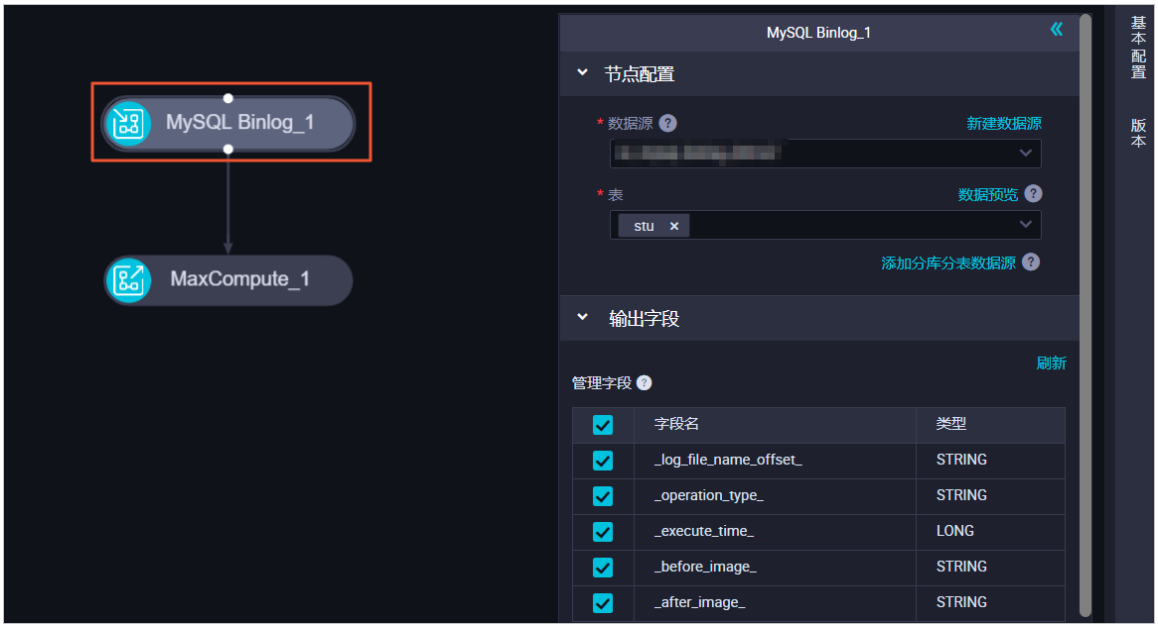
序号	描述
③	属性配置区域，单击组件或右侧的**基本配置**时，会显示相应的属性配置面板。 **注意** 请务必选择**资源组**，否则提交节点时会报错。实时同步仅支持运行在独享数据集成资源组上，详情请参见新增和使用独享数据集成资源组。

ii. 根据自身需求，从组件区域拖拽相应的组件至节点的编辑区域，并通过连线完成相应的节点关系连接，数据会根据连线从上游同步至下游。

下图为您展示新建MySQL数据实时同步至MaxCompute的过程。



iii. 单击相应的节点，在**节点配置**对话框中，配置各项参数。详情请参见实时同步。



iv. 单击工具栏中的图标。

- 当选择**同步方式**为**数据库迁至Hologres**时，操作如下：

i. 双击打开实时同步节点的编辑页面，单击右侧的**基本配置**，从**资源组**下拉列表中选择需要使用的资源组。

> **注意** 请务必选择**资源组**，否则提交节点时会报错。实时同步仅支持运行在独享数据集成资源组上，详情请参见新增和使用独享数据集成资源组。

ii. 在**数据来源**区域，选择**类型**和**数据源**。

iii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表和部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iv. （可选）在**设置同步规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**表名转换规则**和**目标表名规则**：

- **表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

vi. 在**设置目标表**页面，选择**目标Hologres数据源**和该数据源下的**Schema**。

vii. 单击**刷新源表和Hologres表映射**，创建需要同步的源表和目标Hologres表的映射关系。

viii. 查看任务的执行进度和表来源，单击**下一步**。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	包括**自动建表**和**使用已有表**。 **说明** 暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。
③	选择的**表建立方式**不同，此处显示的Hologres表名也不同： ■ 当**表建立方式**选择**自动建表**时，单击**下一步**，会显示**自动建表**对话框。请单击**开始建表**，创建成功后，单击**完成**。您可以单击表名称，查看和修改建表语句。 ■ 当**表建立方式**选择**使用已有表**时，请在下拉列表中选择需要的表。

ix. 在**运行资源设置**页面，配置**来源端读取支持最大连接数**和**目标端写入并发数**，并单击工具栏中的图标。

- 当选择**同步方式**为**数据库迁至MaxCompute**时，操作如下：

i. 双击打开实时同步节点的编辑页面，单击右侧的**基本配置**，从**资源组**下拉列表中选择需要使用的资源组。

ii. 在**数据来源**区域，选择**类型**和**数据源**。

iii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表和部分表进行同步。

注意 如果选中的表没有主键，将无法进行实时同步。

iv. 在**设置同步规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**表名转换规则**和**目标表名规则**：

- **表名转换规则**：转换表名为目标表名，进行字符串替换。

- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

vi. 在**设置目标表**页面，选择**目标MaxCompute（ODPS）数据源**，单击**MaxCompute（ODPS）时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置（支持天和小时级别的分区）。

vii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

viii. 查看任务的执行进度和表来源，单击**下一步**。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	包括**自动建表**和**使用已有表**。 **说明** 暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当**表建立方式**选择**自动建表**时，单击**下一步**，会显示**自动建表**对话框。请单击**开始建表**，创建成功后，单击**完成**。您可以单击表名称，查看和修改建表语句。 ■ 当**表建立方式**选择**使用已有表**时，请在下拉列表中选择需要的表。

ix. 在**运行资源设置**页面，配置**来源端读取支持最大连接数**和**目标端写入并发数**，并单击工具栏中的图标。

- 当选择**同步方式**为**数据库迁至DataHub**时，操作如下：

i. 双击打开实时同步节点的编辑页面，单击右侧的**基本配置**，从**资源组**下拉列表中选择需要使用的资源组。

ii. 在**数据来源**区域，选择**类型**和**数据源**。

iii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表和部分表进行同步。

注意 如果选中的表没有主键，将无法进行实时同步。

iv. 在**设置同步规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和Topic转换规则**和**目标Topic规则**。

v. 单击**下一步**。

vi. 在**设置目标表**页面，选择**目标DataHub数据源**，单击**刷新源表和DataHub Topic映射**，创建需要同步的源表和目标Topic的映射关系。

vii. 查看任务的执行进度和Topic来源，单击**下一步**。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的Topic数量较多，会导致执行进度较慢，请耐心等待。
②	包括**自动建表**和**使用已有Topic**。
③	选择的**Topic建立方式**不同，此处显示的DataHub Topic也不同： ■ 当**Topic建立方式**选择**自动建表**时，单击**下一步**，会显示**自动建表**对话框。请单击**开始建表**，创建成功后，单击**完成**。 ■ 当**Topic建立方式**选择**使用已有Topic**时，请在下拉列表中选择需要的Topic。

viii. 在**运行资源设置**页面，配置**来源端读取支持最大连接数**和**目标端写入并发数**，并单击工具栏中的图标。

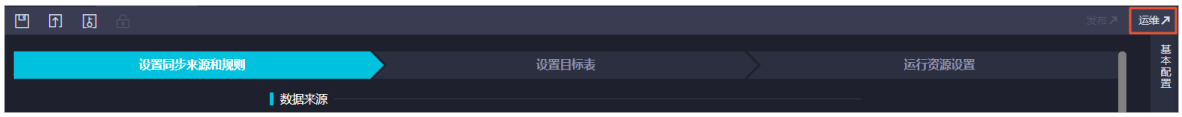
提交实时同步节点

1. 在实时同步节点的编辑页面，单击工具栏中的图标。
2. 在**提交新版本**对话框中，输入**变更描述**。
3. 单击**确认**。

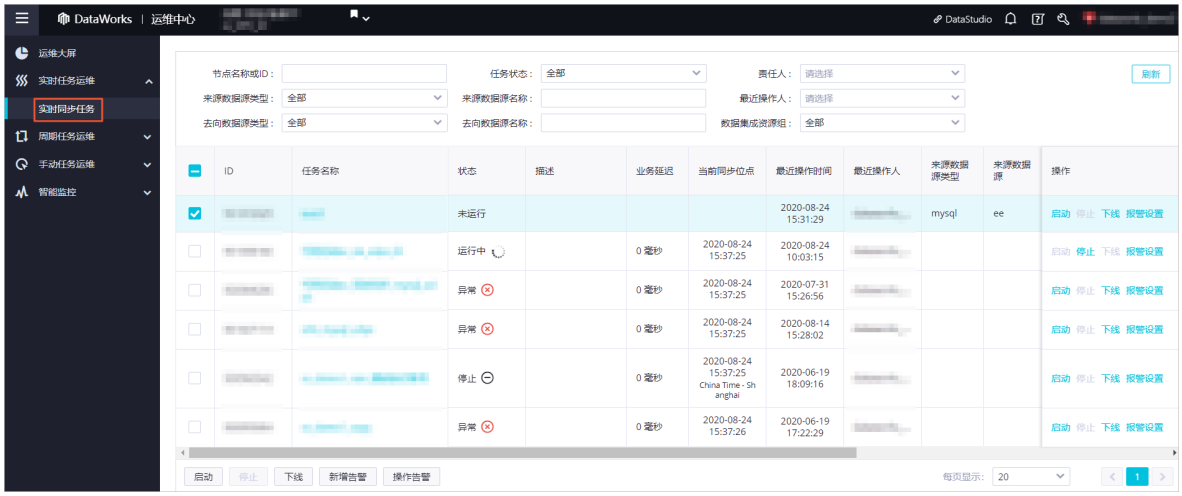
   如果您使用的是标准模式的工作空间，提交成功后，请单击右上方的**发布**。详情请参见发布管理。

运维实时同步节点

1. 提交或发布节点成功后，单击页面右上方的**运维**，进入**实时任务运维 > 实时同步任务**页面。



2. 您可以在**实时同步任务**页面，单击相应的**任务名称**，查看详细的运维信息。



您可以在该页面对实时同步节点进行**启动**、**停止**、**下线**和**报警设置**等操作：

- 启动非运行状态的任务：
  a. 单击相应任务后的**启动**。
  b. 在**启动**对话框中，配置各项参数。

参数	描述
是否重置位点	如果选中该参数，请设置下次启动的时间位点。即**启动时间点位**和**时区**为必选项。
启动时间点位	选择启动节点的日期和时间。
时区	从**时区**下拉列表中选择时区。
任务自动结束	■ 配置脏数据的最大容忍条数。如果您配置为0，表示严格不允许脏数据存在。如果不配置，则代表容忍脏数据。 ■ 如果您不配置Failover次数，将根据5分钟Failover 100次来自动结束任务，避免频繁启动占用系统资源。

  c. 单击**确认**。

- 停止运行中的任务：
  a. 单击相应任务后的**停止**。
  b. 在确认对话框中，单击**停止**。

- 下线非运行状态的任务：
  a. 单击相应任务后的**下线**。
  b. 在确认对话框中，单击**下线**。
- 单击相应任务后的**报警设置**，您可以在该页面查看**报警时间**和**报警规则**。
- 新增告警：
  a. 选中需要新增告警的任务，单击页面下方的**新增告警**。
  b. 在**新建规则**对话框中，配置各项参数。

参数	描述
名称	新建规则的名称，必填项。
描述	对新建规则进行简单描述。
指标	包括**任务状态**、**业务延迟**、**Failover**、**脏数据**和**DDL错误**。 **说明** ■ Failover：任务运行出错，各种原因都有可能。 ■ 脏数据：正常读取的数据，但无法正常写入。关于脏数据的完整定义，您可以参考：基本概念。 ■ DDL不支持：由于不支持来源DDL操作，导致的错误。
阈值	设置**WARNING**和**CRITICAL**的阈值，默认值为5分钟。 **说明** 无心跳：指管控与执行层之间的信号中断了，有可能是管控信号网络断了，或者任务异常挂了。
报警间隔	设置报警的时间间隔，默认值为5分钟内只发一次报警。
WARNING	包括**邮件**、**短信**和**钉钉**。 **说明** 可使用短信告警的地域为：新加坡、马来西亚（吉隆坡）、德国（法兰克福）。其他地域如果希望通过短信方式报警，可提交工单联系阿里云DataWorks技术人员咨询办理。
CRITICAL	
接收人（非钉钉）	从**接收人（非钉钉）**下拉列表中选择接收人。

  c. 单击**确认**。
- 操作告警：
  a. 选中需要操作告警的任务，单击页面下方的**操作告警**。
  b. 在**操作告警**对话框中，选中**操作类型**和**告警指标**。
  选中要操作的告警类型后，其对应的所有规则会被批量修改。
  c. 单击**确认**。

5.同步解决方案

5.1. 概述

DataWorks为您提供多种数据源之间进行不同数据同步场景的同步解决方案，包括实时数据同步、离线全量同步、离线增量同步等同步场景，助力企业数据更高效、更便捷的一键上云。

背景信息

实际业务场景下，数据同步通常不能通过一个或多个简单离线同步或者实时同步任务完成，而是由多个离线同步、实时同步和数据处理等任务组合完成，这就会导致数据同步场景下的配置复杂度非常高。为了解决上述问题，DataWorks提出了面向业务场景的同步任务配置化方案，支持不同数据源的一键同步功能，例如“一键实时同步至Elasticsearch”、“一键实时同步至Hologres”和“一键实时同步至MaxCompute”功能等，通过此类功能，您只需要进行简单的配置，就可以完成一个复杂业务场景。

以业务数据库数据同步到MaxCompute数据仓库为例，当有大量的数据存储在数据库系统里，需要将数据库中的全量及增量数据同步到MaxCompute进行数仓分析时，数据集成传统方式是通过全量同步或者依赖数据库表中的modify_time等字段进行增量同步。但实际生产场景下，数据库表中不一定存在modify_time等字段，因此不能使用传统的基于JDBC抽取的方式进行增量同步。而“一键实时同步至MaxCompute”功能实现了数据库全增量实时同步到MaxCompute以及在MaxCompute上进行自动的全增量合并，可以使整个同步场景化繁为简。

同步任务配置化方案具有如下优势：

- 全量数据初始化。
- 增量数据实时写入。
- 增量数据和全量数据定时自动合并写入新的全量表分区。

使用限制

DataWorks的同步解决方案暂不支持跨时区同步数据。如果同步任务中同步的数据源与使用的DataWorks资源组不在同一个时区，则会导致同步的数据有误。

支持的数据源

DataWorks的同步解决方案支持多种数据源间的多种数据同步，详情可见下表。

去向数据源	来源数据源	数据源配置指导	同步任务配置指导
Elasticsearch	• MySQL • PolarDB MySQL 说明 当前仅支持PolarDB MySQL。	配置数据源（来源为MySQL）	配置查看整库离线同步任务
Hologres	• PolarDB MySQL • Oracle • MySQL • PolarDB-X	• 配置数据源（来源为PolarDB） • 配置数据源（来源为Oracle） • 配置数据源（来源为MySQL） • 配置数据源（来源为DRDS）	配置查看数据同步任务
MaxCompute	• PolarDB MySQL • Oracle • MySQL • PolarDB-X	• 配置数据源（来源为PolarDB） • 配置数据源（来源为Oracle） • 配置数据源（来源为MySQL）	一键实时同步至MaxCompute

资源使用与费用

使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

独享数据集成资源组的性能指标如下表。

规格	离线同步最大并发线程数	单表实时同步最大任务数	整库多表实时同步最大任务数	分库分表实时同步最大任务数
4c8g	8	3	3	不支持
8c16g	16	6	6	1
12c24g	24	9	9	1
16c32g	32	12	12	2
24c48g	48	18	18	3

不同地域的各个规格的独享数据集成资源组的定价可参考计费标准。实际支付价格以订单页面为准。

您可结合待同步的数据量估算并购买独享数据集成资源组，独享数据集成资源组的详细介绍可参见独享数据集成资源。任务调度可运行在公共调度资源组上，或运行在您已购买的独享调度资源组上。

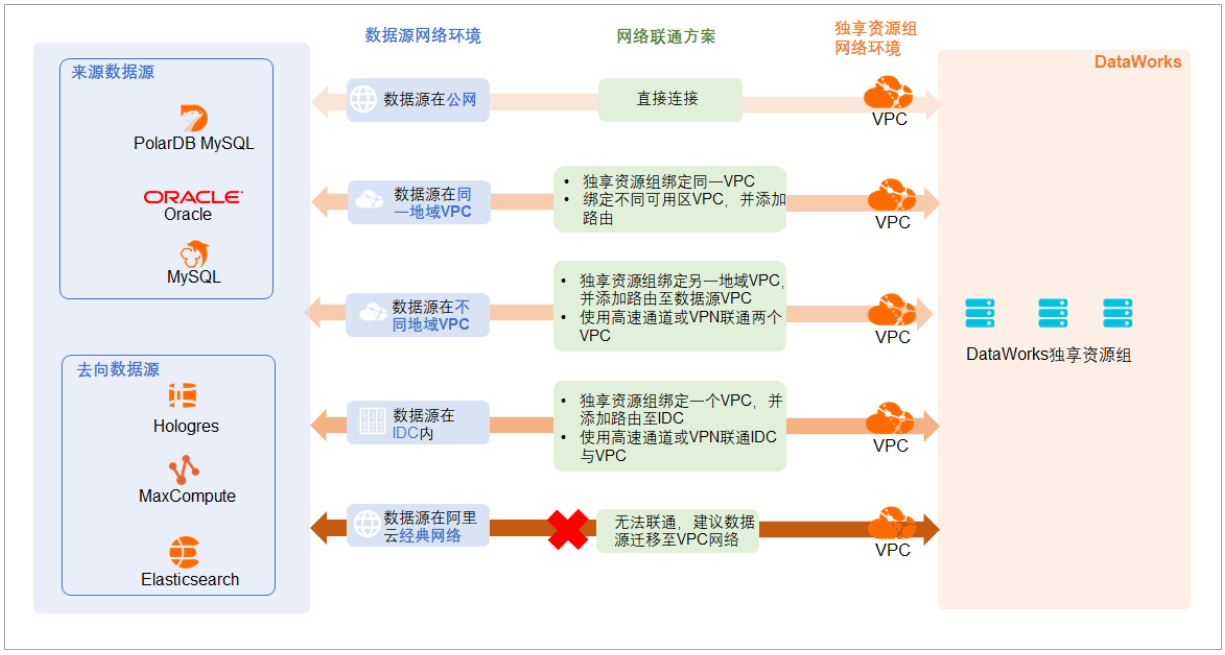
> 说明
>
> - 同步解决方案本身功能不收费，其是一堆具体子任务的组合，每个子任务按照各自归属的具体类型进行收费。（比如，内部生成的离线和实时同步子任务，使用的是数据集成独享资源组和调度资源组，费用按照对应的资源组来计算）。
> - 产生的其他任务，比如一键实时同步至MaxCompute解决方案需要定期做全增量数据周期合并，因此会消耗MaxCompute计算资源。这部分费用由MaxCompute直接收取，费用与用户同步全量数据大小、合并周期正相关。具体费用可以参考MaxCompute计费项与计费方式概述。

网络联通方案

DataWorks的网络连通性解决方案详细可参见配置网络连通。以下为您概要介绍数据源与独享资源组之间的网络联通方案。

独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

根据您实际的来源数据源、去向数据源所在网络环境不同，有不同的网络联通方案与独享资源组进行网络联通。



- 当数据源处于公网环境中时：

  数据源所在的公网环境可与独享资源组绑定的VPC直接连接。

- 当数据源与独享资源组处于同一地域的VPC网络环境中时：
  - 独享资源组与数据源在同一可用区时，可绑定数据源所在的VPC网络。
  - 独享资源组与数据源不在同一可用区时，可绑定一个VPC后并配置路由，将独享资源组路由至数据源所在的VPC网络。
- 当数据源与独享资源组处于不同地域的VPC网络环境中时：
  - 独享资源组可绑定一个VPC后并配置路由，将独享资源组路由至数据源所在的VPC网络。
  - 使用高速通道或VPN网关，联通独享资源组绑定的VPC与数据源所在VPC。
- 当数据源处于IDC网络环境中时：
  - 独享资源组可绑定一个VPC后并配置路由，将独享资源组路由至数据源所在的IDC网络。
  - 使用高速通道或VPN网关，联通独享资源组绑定的VPC与数据源所在IDC网络。
- 当数据源处于阿里云经典网络环境中时：

  经典网络与独享资源组使用的VPC网络无法联通，建议您将数据源的网络环境迁移至VPC中。

操作流程

使用DataWorks的同步解决方案的操作流程一般包含以下几个流程：

1. 资源规划与配置

   根据待同步的数据量和网络情况，评估规划并购买您需要使用的独享数据集成资源组和调度资源组，根据规划配置好资源，保障网络连通性。

2. 配置数据源

   网络通畅后，您还需根据待同步的数据源情况，配置好数据源的可访问性，保障后续数据同步不会因为白名单等限制导致同步失败。

3. 添加数据源

   分别将来源数据源和去向数据源添加至DataWorks，便于后续创建同步方案时直接关联使用。

4. 配置查看数据同步任务

   创建同步方案，并根据同步场景设置同步细节参数。

> **说明**
> - 支持对已创建的同步解决方案增加和删除表，如果实时同步任务正在运行需要先终止任务，配置增加和删除表后**提交执行**解决方案，系统会自动新建离线任务和更新实时任务。请参见：增加或删除已运行任务的同步表。
> - 同步解决方案的操作流程中，在**设置目标表**时，当**表建立方式**为**自动建表**时，您可以单击表名为您弹出建表语句或者配置框，并且允许您手工修改。请您仔细检查是否满足需求。

不同数据源的详细同步流程和操作步骤可参见对应指导文档：

- 同步数据至Elasticsearch
- 同步数据至Hologres
- 同步数据至MaxCompute

5.2. 选择同步解决方案

数据集成支持数据同步解决方案功能，您可以通过配置同步规则，一次性实时同步或离线同步数据至相应的数据源中。同步解决方案支持整库内批量同步多张表，也支持全量、增量数据一体化同步，先同步全量数据，再实时同步增量数据。

进入同步解决方案页面

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
4. 在左侧导航栏，单击**同步任务**，进入**解决方案任务列表**页面。

   您可以在该页面新建同步解决方案，或查看已创建的方案及当前执行的状态。

新建同步解决方案时，您可以根据需要选择对应的数据同步解决方案，详情可参见下文的选择同步方案。查看同步解决方案状态时，包括以下状态：

- **未运行**：任务未开始执行。您可以单击相应方案后的**开始执行**，执行当前方案。

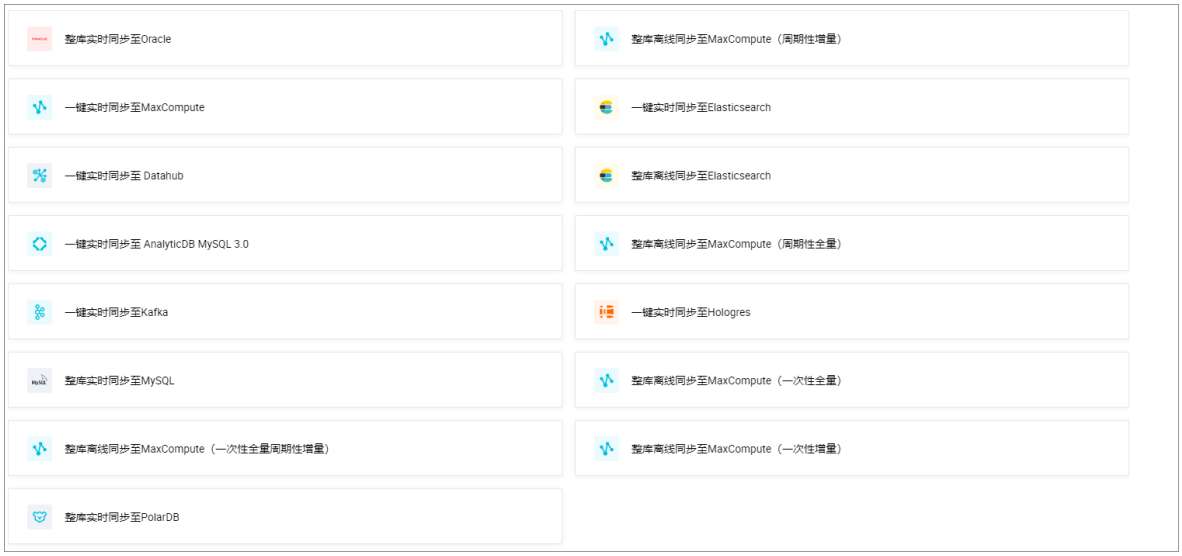
  > **说明** 仅单击**未运行**状态后的**任务配置**，您可以编辑任务。其它状态下的任务配置页面，仅支持查看。

- **运行中**：任务在运行中，无法终止。您需要等待任务执行结束。
- **异常**：任务出错，无法正常运行。此时您需要单击相应任务后的**执行详情**进行排查。
- **成功**：任务成功运行。您可以单击相应任务后的**执行详情**，查看运行结果。

选择同步方案

1. 在同步解决方案页面的右上角单击**新建任务**。
2. 在新建任务页面选择数据同步的来源与去向后，选择可用的同步方案。

   支持的来源数据源与去向数据源，可参见支持的数据源，当前支持的同步方案如下所示。



根据数据同步的目的数据源类型进行分类，当前DataWorks数据同步支持以下几种数据同步方案：

- 同步数据至DataHub：当前支持的数据同步解决方案为**一键实时同步至DataHub**。
- 同步数据至Elasticsearch：当前支持的数据同步解决方案为**一键实时同步至Elasticsearch**。
- 同步数据至Hologres：当前支持的数据同步解决方案为**一键实时同步至Hologres**。
- 同步数据至AnalyticDB MySQL3.0：当前支持的数据同步解决方案为**一键实时同步至AnalyticDB MySQL3.0**。
- 同步数据至Kafka：当前支持的数据同步解决方案为**一键实时同步至Kafka**。
- 同步数据至MaxCompute：当前支持的数据同步解决方案为
  - **一键实时同步至MaxCompute**
  - **整库离线同步至MaxCompute（周期性全量）**
  - **整库离线同步至MaxCompute（周期性增量）**
  - **整库离线同步至MaxCompute（一次性全量）**
  - **整库离线同步至MaxCompute（一次性增量）**
  - **整库离线同步至MaxCompute（一次性全量周期增量）**

5.3. 同步数据至DataHub

5.3.1. 资源规划与配置

当前使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表** > **独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次进入管控后弹出的界面弹框中进行授权。
3. 绑定专有网络VPC。
   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

> **说明** 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。

4. （可选）配置Host。

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明
- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。数据源的配置可参见配置数据源（来源为MySQL）、配置数据源（来源为PolarDB）和配置数据源（来源为Oracle）。

5.3.2. 配置数据源（来源为MySQL）

同步MySQL的数据至DataHub时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源DataHub。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。

- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

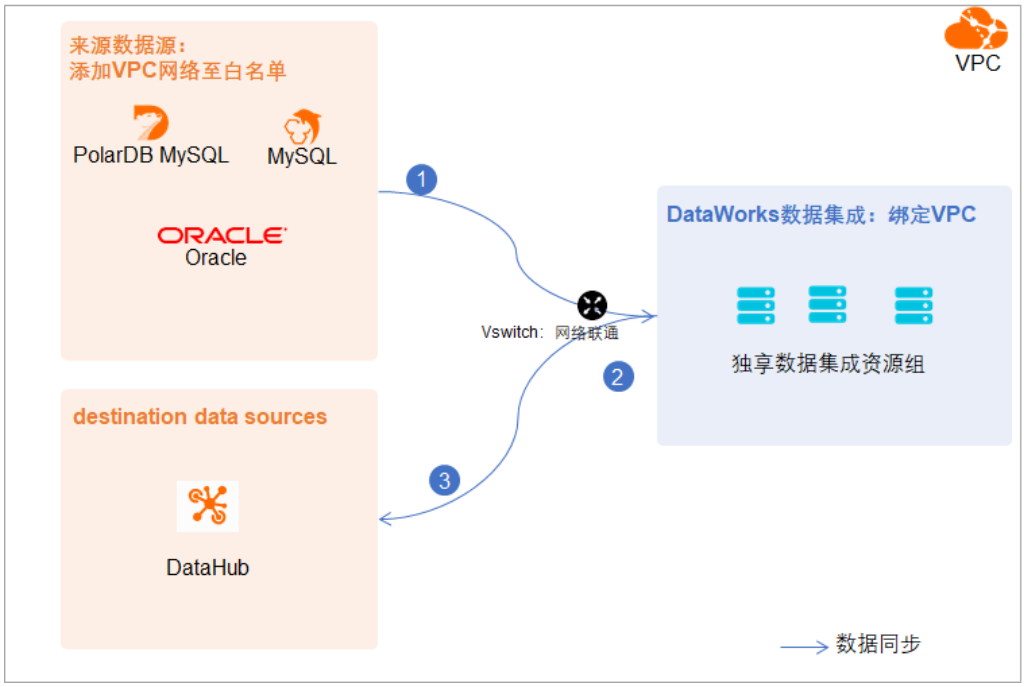
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将独享数据集成资源组在网络配置时所绑定的交换机网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。

- Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

   a. 登录DataWorks控制台。

   b. 在左侧导航栏，单击**资源组列表**。

   c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

   d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

   e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

   f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

   i. 创建账号。

   操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.3.3. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至DataHub时，您需要参考本文在数据源中配置好网络、白名单、权限等配置，为后续的数据同步方案执行做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源DataHub。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。

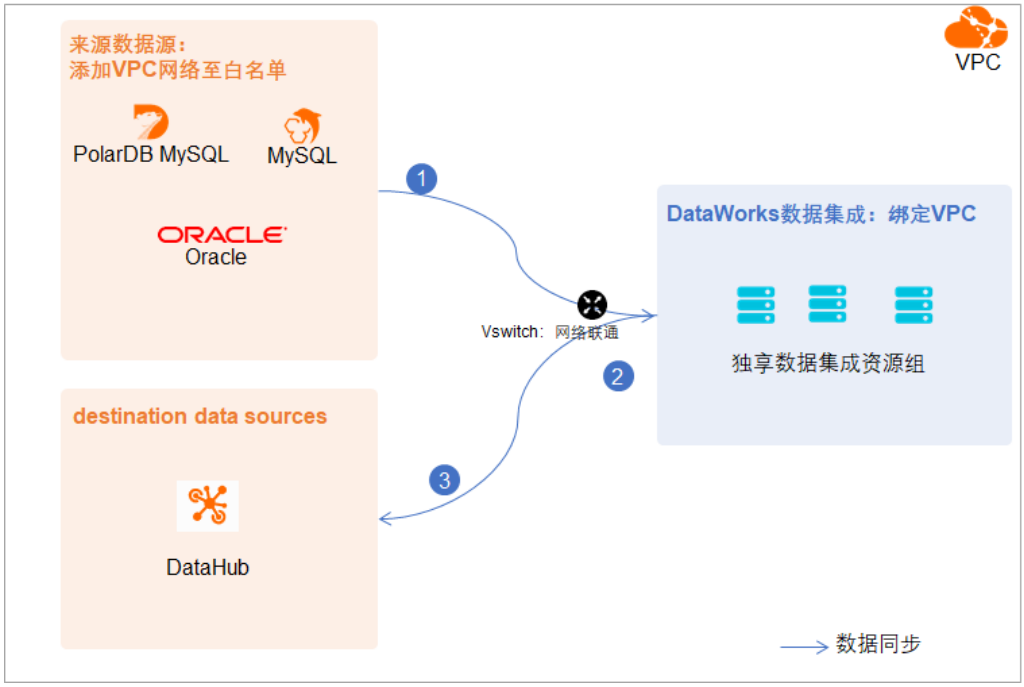
- 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

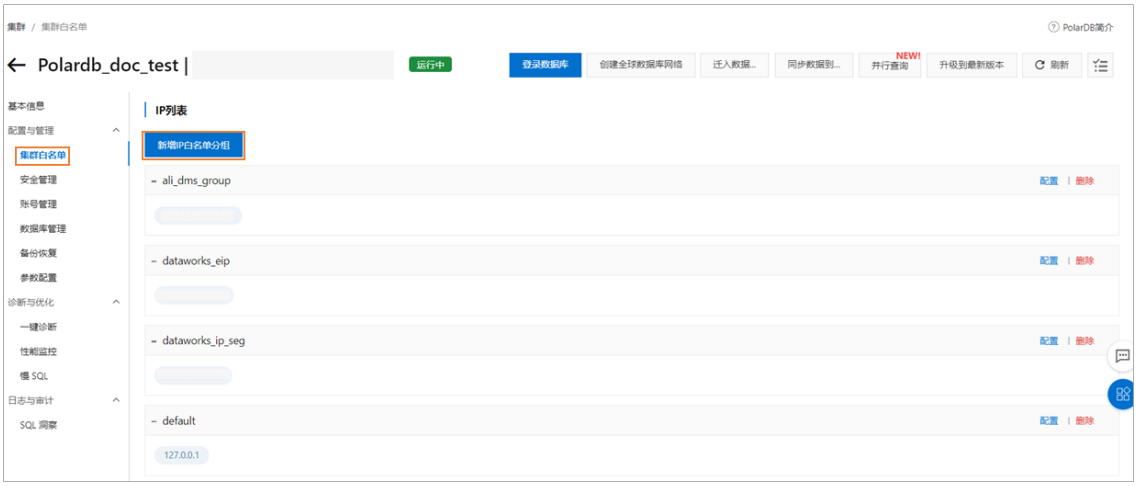
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.3.4. 配置数据源（来源为Oracle）

同步Oracle的数据至DataHub时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

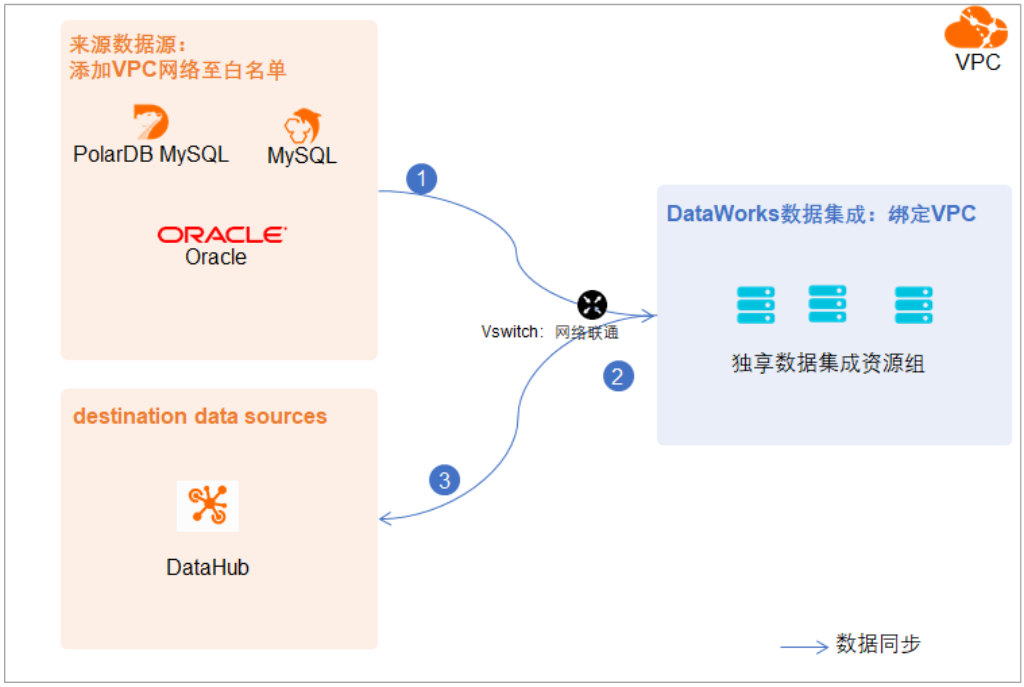
- 准备数据源：已购买来源数据源Oracle、去向数据源DataHub。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。
  - Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。

- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换`'同步账号'`为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.3.5. 添加数据源

将来源数据源的数据同步至DataHub数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：Oracle

添加Oracle数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置数据源（来源为Oracle）。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置查看整库实时同步任务。

5.3.6. 配置查看整库实时同步任务

完成数据源、网络、资源的准备配置后，您可以创建并执行整库实时同步任务，开始进行数据同步。本文为您介绍如何创建整库实时同步任务，先将指定数据库中的部分或全部表的数据离线同步至DataHub中，再将后续新增的数据实时同步至DataHub中，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为MySQL）
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 添加数据源

背景信息

DataWorks为您提供了整库实时同步数据至DataHub的解决方案，轻松助力企业实现同步整库数据至DataHub后，再将持续新增的数据实时同步至DataHub中。同时，您可以实时查看创建的同步任务详情，监控任务的运行状况及业务库数据的更新情况，用于后续做数据检索、数据分析或数据开发。

整库实时同步数据的优势如下：

- 整库级别同步。

  无需逐个创建表级别的数据同步任务，支持直接创建库级别的同步任务，选择目标库的部分或全部表数据进行同步。

- 同步规则配置灵活。
  - 您可以根据业务需求灵活配置不同DDL消息的处理规则。例如，针对来源端删除表的DDL消息，如果您将同步数据的处理方式配置为**忽略**，则进行实时同步时，DataWorks收到相关删除表的DDL消息时，会忽略该类消息，目标端的对应表则不会被删除。
  - 您可以编辑已经配置运行的同步任务，为该任务快速添加表或删除已有同步表。
  - 您可以根据业务需求，配置目标索引的同步规则，选择是否添加同步源表的实时新增字段至目标索引。添加字段至目标索引后，该字段后续可被搜索。
- 配置操作简单。

  您无需进行创建同步任务、创建数据库、创建表、创建相互依赖以及执行参数对齐等复杂操作，通过简单的产品配置向导，即可完成对应功能的配置。

- 实现海量数据的实时更新，自动化运维管理效率较高。

适用场景

适用于需要实时监测业务库数据的更新情况，便于上层应用对实时数据进行检索分析或数据开发的场景。

使用限制

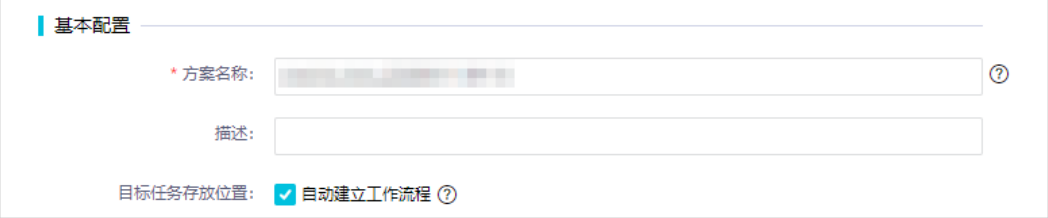
- 目前仅支持整库实时同步MySQL、PolarDB、Oracle类型的数据库至DataHub。
- 整库实时同步解决方案仅支持使用独享资源组。

创建整库实时同步任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 在**新建同步解决方案**对话框中，单击**一键实时同步至DataHub**。
4. 完成方案名称等基本信息配置。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

5. 选择来源数据源并配置同步规则。
   i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明**
>
> 目前仅支持整库实时同步MySQL、PolarDB、Oracle类型的数据库至DataHub。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择同步目标数据源的部分或全部表。

iii. 在**设置表名到Topic的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标Topic转换规则**和**目标索引名规则**：

- **源表名和目标Topic转换规则**：转换表名为目标Topic，进行字符串替换。
- **目标Topic规则**：支持对转换后的Topic添加前缀和后缀。

iv. 单击**下一步**。

6. 选择目标数据源并配置目标Topic。

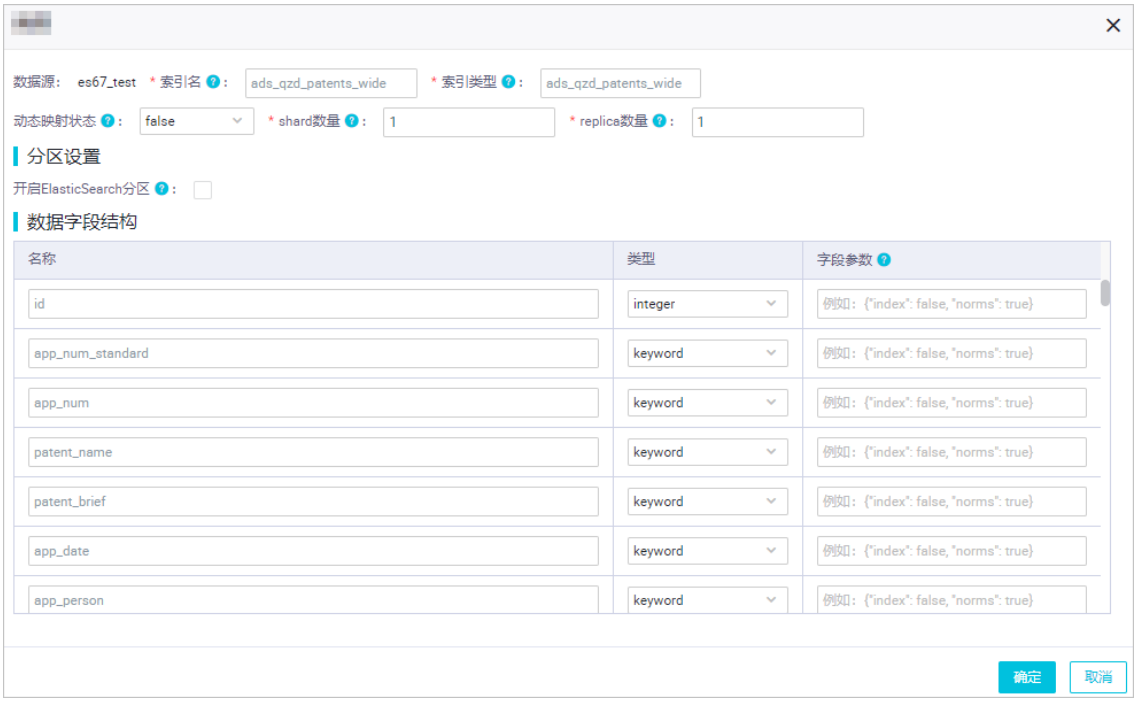
i. 在**设置目标Topic**页签，选择**目标DataHub数据源**。

ii. 单击**刷新源表和DataHub Topic映射**，创建需要同步的源表和目标DataHub Topic的映射关系。

iii. 查看任务的执行进度和表来源。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	选择的**Topic建立方式**，取值如下： ■ 当**Topic建立方式**选择**自动建Topic**时，**DataHub Topic**列显示自动创建的DataHub Topic名称。您可以单击Topic名称，修改建立Topic的相关配置。 ■ 当**Topic建立方式**选择**使用已有Topic**时，请在**DataHub Topic**列对应的下拉列表中，选择需要使用的Topic。

当**Topic建立方式**选择**自动建Topic**时，您可以单击创建的DataHub Topic名称，根据业务需求修改目标Topic的相关参数。



- **同时创建生产环境Topic**：用于同时创建生产环境Topic。标准模式下显示该选项，且默认为选中状态。
- **生命周期**：用于设置Topic的生命周期。默认为7天。
- **数据字段结构**：用于设置映射的目标Topic中字段的类型及扩展属性。

说明 当创建了目标Topic后，如果不修改相关参数，则系统会按照默认值的相应规则进行数据同步。

iv. 单击**下一步**。

7. 运行资源设置。

在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	选择全量离线任务使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为15。

8. 单击**完成配置**，完成整库实时同步任务的创建。

> **说明**
> - 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
> - 实时同步任务仅支持运行在独享数据集成资源组上，详情请参见新增和使用独享数据集成资源组。

执行整库实时同步任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的整库实时同步任务。

查看任务运行状态及结果

- 在**解决方案任务列表**页面，选择已运行任务后的**更多 > 执行详情**，查看当前解决方案整库实时同步任务过程中，各子任务节点的运行详情。

步骤	说明	起始时间	结束时间	状态
1	批量创建Datahub Topic	2021-05-20 17:49:09	2021-05-20 17:49:12	成功 执行详情
2	创建DataWorks工作流程	2021-05-20 17:49:12	2021-05-20 17:49:12	成功
3	创建DataWorks虚拟节点	2021-05-20 17:49:12	2021-05-20 17:49:12	成功 执行详情
4	创建全量同步任务节点	2021-05-20 17:49:12	2021-05-20 17:49:15	成功 执行详情
5	提交发布DataWorks虚拟节点	2021-05-20 17:49:15	2021-05-20 17:49:18	成功 执行详情
6	提交发布全量同步任务节点	2021-05-20 17:49:18	2021-05-20 17:49:34	成功 执行详情
7	全量同步任务节点批量冒烟执行	2021-05-20 17:49:34	2021-05-20 17:50:42	成功 执行详情
8	创建DataWorks实时同步节点	2021-05-20 17:50:42	2021-05-20 17:50:43	成功 执行详情
9	提交发布DataWorks实时同步节点	2021-05-20 17:50:43	2021-05-20 17:50:44	成功
10	启动DataWorks实时同步节点	2021-05-20 17:50:44	2021-05-20 17:50:50	成功 执行详情

- 单击子任务节点后的**执行详情**，可以单击对话框中的任务链接，进入子节点的数据开发页面。

管理整库实时同步任务

- 查看任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 查看配置**，可查看任务的配置信息。

- 编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 修改配置**，可编辑任务的配置信息。

  对于已运行成功的实时同步任务，您可以单击相应任务后的**更多 > 修改配置**，快速添加或删除同步表。方法如下：

  在**设置同步来源和规则**页签的**选择同步的源表**区域，选择添加或删除目标同步表。修改配置后，保存并运行同步任务，则可以快速添加新表或删除已有同步表。

- 修改任务优先级。

单击相应任务后的**更多 > 修改优先级**。在**修改优先级**对话框中，输入需要配置的优先级数值，单击**确定**。优先级取值范围为1~8，数值越大优先级越高。

> **说明** 优先级相同的任务，按照提交时间的先后顺序执行。

- 删除任务。

  单击相应任务后的**更多 > 删除**。在删除对话框中，单击**确认**。

> **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

5.4. 同步数据至ElasticSearch

5.4.1. 资源规划与配置

当前使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。

参数	描述
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

说明 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

说明 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC
   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。数据源的配置可参见配置数据源（来源为MySQL）及配置数据源（来源为PolarDB）。

5.4.2. 配置数据源（来源为MySQL）

同步MySQL的数据至ElasticSearch时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源ElasticSearch。

- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

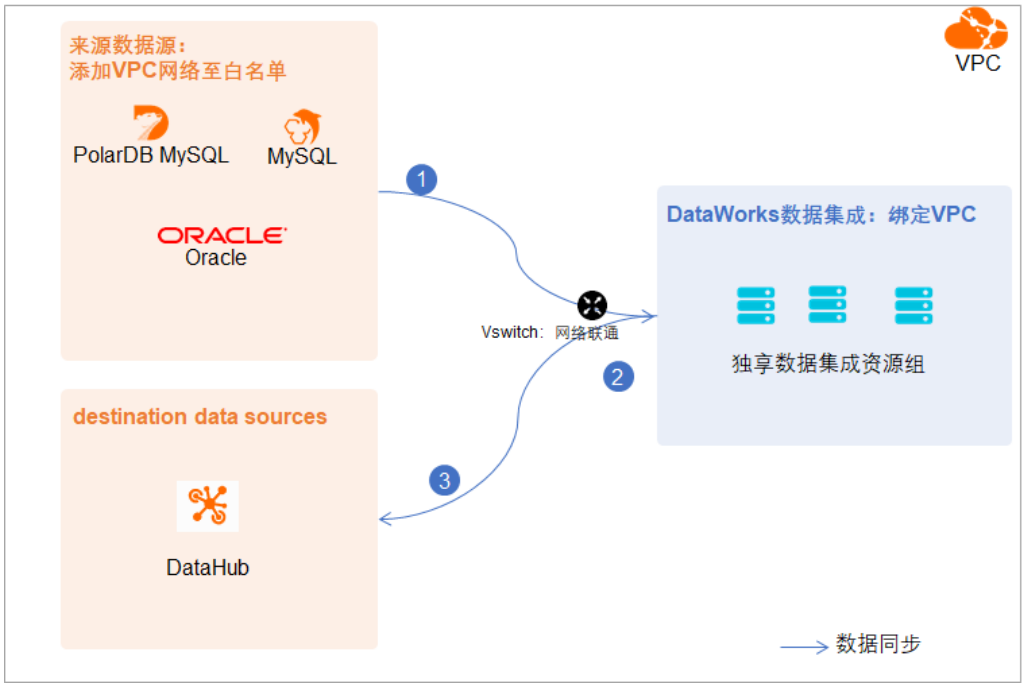
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将独享数据集成资源组在网络配置时所绑定的交换机网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。

- Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

   a. 登录DataWorks控制台。

   b. 在左侧导航栏，单击**资源组列表**。

   c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

   d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

   e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

   f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

   i. 创建账号。

   操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.4.3. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至Elasticsearch时，您需要参考本文在数据源中配置好网络、白名单、权限等配置，为后续的数据同步方案执行做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源Elasticsearch。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。

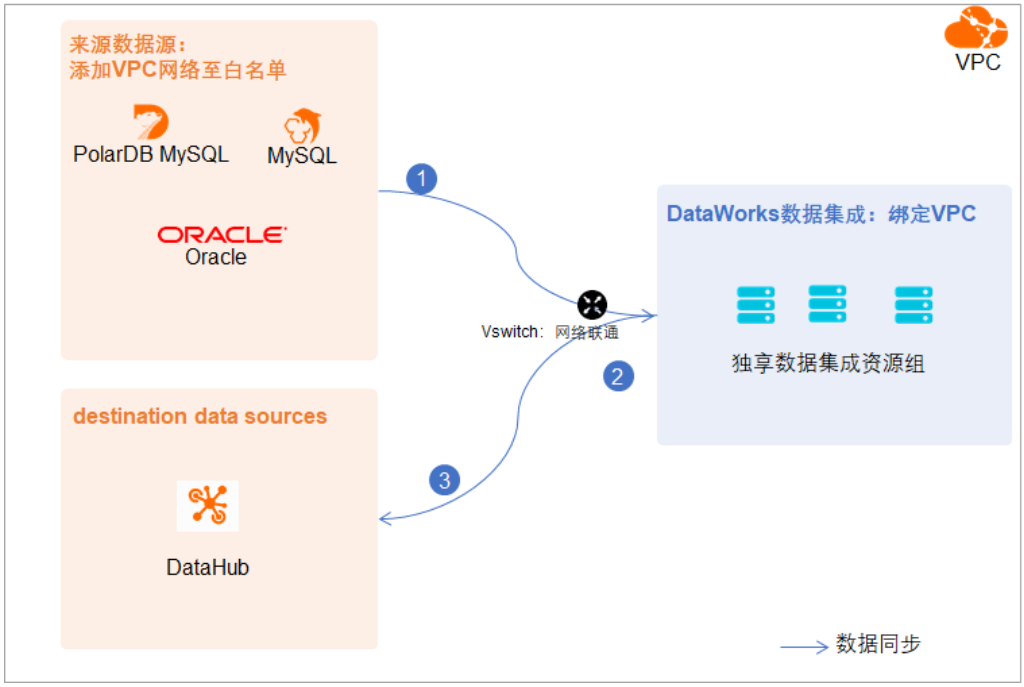
- 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将独享数据集成资源组在网络配置时所绑定的交换机网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

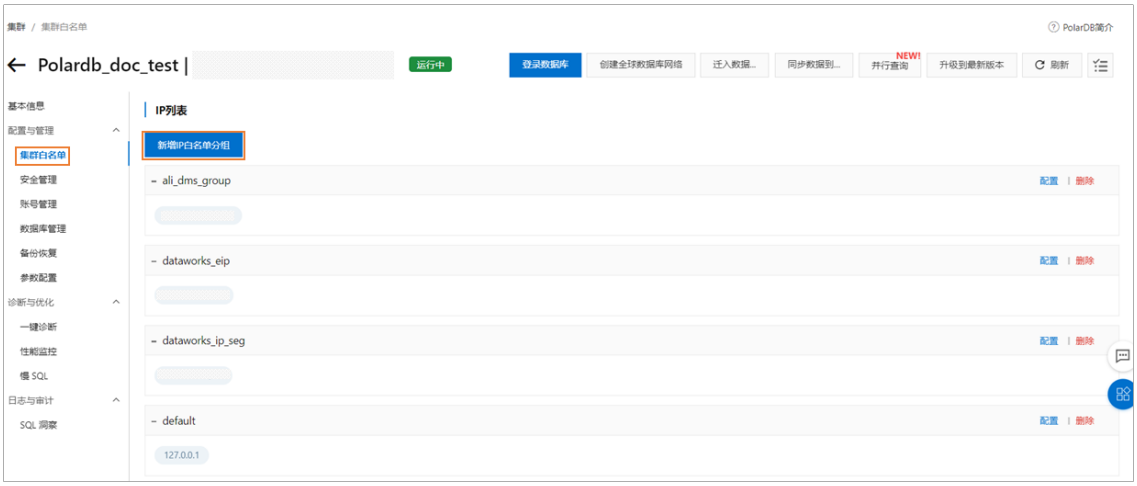
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.4.4. 添加数据源

将来源数据源的数据同步至Elasticsearch数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加去向数据源：Elasticsearch

操作详情可参见配置Elasticsearch数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置查看整库离线同步任务或配置查看整库实时同步任务。

5.4.5. 配置查看整库离线同步任务

完成数据源、网络、资源的准备配置后，您可以创建并执行整库离线同步任务，开始进行数据同步。本文为您介绍如何创建整库离线同步任务，将指定数据库中的部分或全部表的数据同步至Elasticsearch中，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为MySQL）

- 配置数据源（来源为PolarDB）
- 添加数据源

背景信息

企业的实时数据一般存储在大数据引擎中，实时数据通常会产生许多非结构化的日志数据，这类日志数据以及企业的离线数据可以使用通过Elasticsearch全托管方式提供的冷热存储方案进行存储。基于该功能，DataWorks为您提供了整库离线同步数据至Elasticsearch的解决方案，轻松助力企业同步整库数据至Elasticsearch中。同时，您可以查看创建的同步任务详情，监控任务的运行状况，提高自动化运维管理效率。

整库离线同步任务可以将业务库数据全量或增量同步至ElasticSearch中，用于做数据检索、数据分析或后续进行数据开发。其优势如下：

- 整库级别同步。

  无需逐个创建表级别的数据同步任务，支持直接创建库级别的同步任务，选择目标库的部分或全部表数据进行同步。
- 同步方式多样。

  支持全量、增量以及全量和增量相结合的方式同步数据。同时，支持对同步任务进行周期性调度配置。
- 配置操作简单。

  您无需进行创建同步任务、创建数据库、创建表、创建相互依赖以及执行参数对齐等复杂操作，通过简单的产品配置向导，即可完成对应功能的配置。
- 成本较低，自动化运维管理效率较高。

使用限制

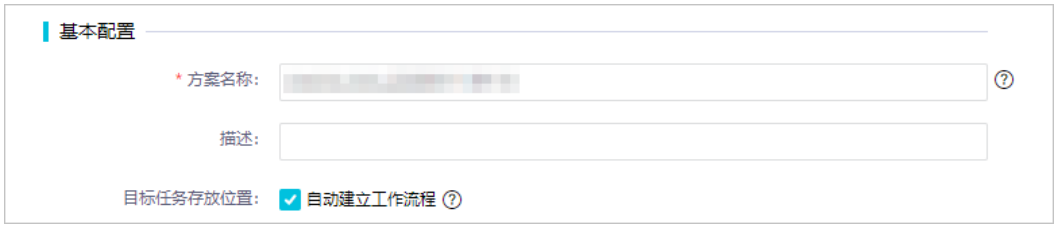
- 目前仅支持整库离线同步MySQL、SQLServer、PolarDB类型的数据库至Elasticsearch。
- 整库离线同步解决方案仅支持使用独享数据集成资源组。

创建整库离线同步任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 在**新建同步解决方案**对话框中，单击**整库离线同步至Elasticsearch**。
4. 完成方案名称等基本信息配置。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

5. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、SQLServer、PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择同步目标数据源的部分或全部表。

> **注意** 如果选中的表没有主键，则在创建源表和目标Elasticsearch索引的映射关系时，需要为该表自定义主键（例如，使用非主键的一个或几个字段的联合代替主键）进行同步数据时去重，详情请参见选择目标数据源并配置目标索引。

iii. 在**设置表名到索引名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标索引名转换规则**和**目标索引名规则**：

- **源表名和目标索引名转换规则**：转换源表名为目标索引名，进行字符串替换。
- **目标索引名规则**：支持对转换后的索引名添加前缀和后缀。

iv. 单击**下一步**。

6. 选择目标数据源并配置目标索引。

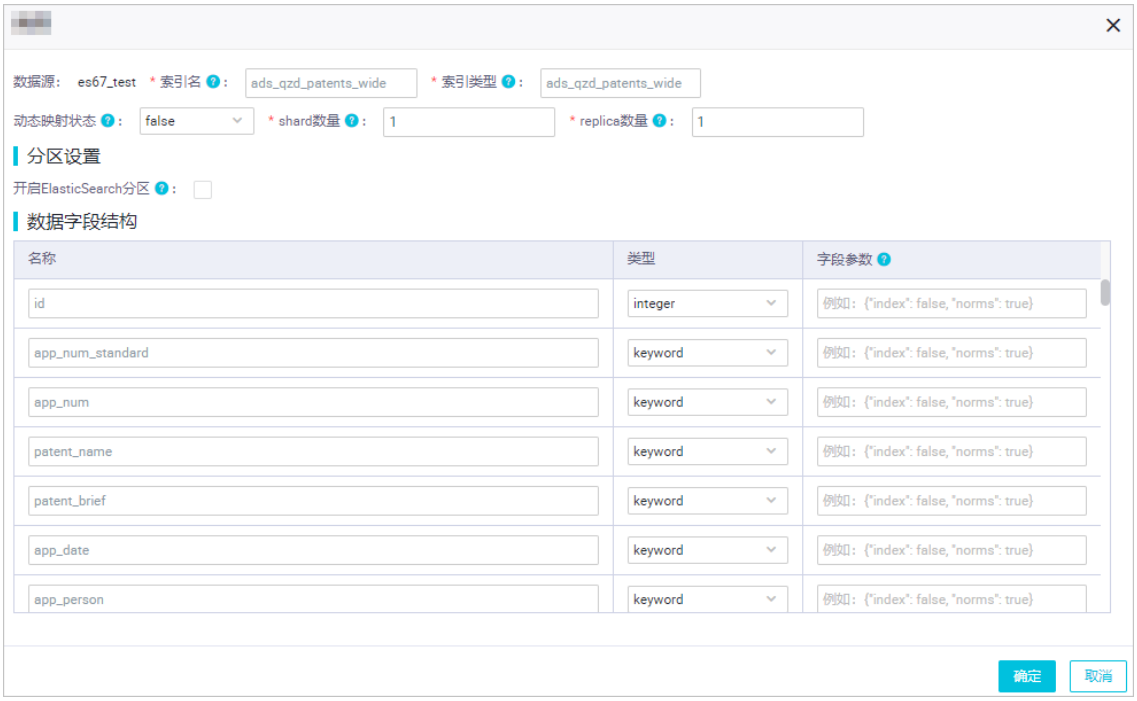
i. 在**设置目标索引**页签，选择**目标Elasticsearch数据源**。

ii. 单击**刷新源表和Elasticsearch索引映射**，创建需要同步的源表和目标Elasticsearch索引的映射关系。

iii. 查看任务的执行进度和表来源。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。 **说明** 使用如下方式进行数据同步的表，必须设置主键： ■ 使用**增量同步**方式同步数据。 ■ 使用**全量同步**方式同步数据，并且**写入Elasticsearch策略**配置为**更新**。 同步方式详情请参见选择同步方案。
③	选择的**索引建立方式**，取值如下： ■ 当**索引建立方式**选择**自动建索引**时，**Elasticsearch索引名**列显示自动创建的Elasticsearch索引名称。您可以单击索引名称，修改建立索引的相关配置。 ■ 当**索引建立方式**选择**使用已有索引**时，请在**Elasticsearch索引名**列对应的下拉列表中，选择需要使用的索引。同时您可以单击**设置同步规则**，查看源表字段与目标索引的映射情况。

当**索引建立方式**选择**自动建索引**时，您可以单击创建的Elasticsearch索引名称，根据业务需求修改目标索引的相关参数。



- **动态映射状态**：用于在同步数据时，是否将数据源表的新增字段添加至目标索引中。取值如下：
  - **true**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，会将该字段同步添加至目标索引中，后续该字段可以被搜索。该取值为默认值。
  - **false**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，会将该字段同步添加至目标索引中，但后续该字段不可以被搜索。
  - **strict**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，将拒绝同步该字段至目标索引并产生异常报错，您可以在日志信息中查看报错详情。

  更多动态映射的内容，详情请参见动态映射。
- **shard数量及replica数量**：索引的主分片及副本分片，用于将一个完整的索引分成多个分片，分布至不同的Elasticsearch节点上，构成分布式搜索，提升Elasticsearch的查询效率，详情请参见基本概念。

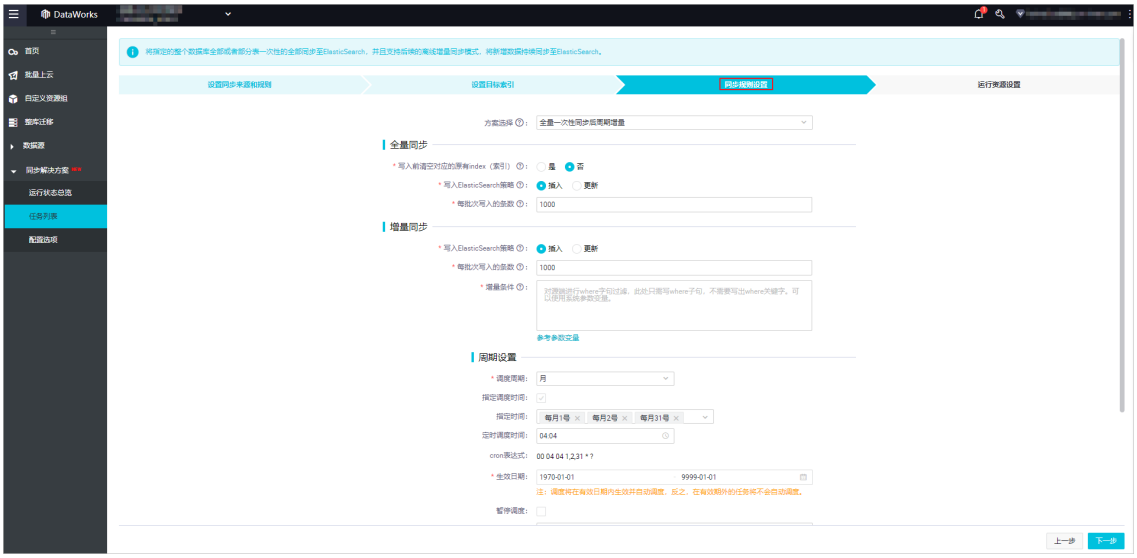
  > **说明** shard数量及replica数量参数，配置运行后不可更改，默认取值为 *1*。
- **分区设置**：您可以选择同步的数据源表的某一列做为分区列，该参数需要和**shard数量**、**replica数量**配合使用。默认不开启分区功能。
- **数据字段结构**：用于设置映射的目标索引中字段的类型及扩展属性，详情请参见Elasticsearch的字段类型。

> **说明** 当创建了目标Elasticsearch索引后，如果不修改相关参数，则系统会按照默认值的相应规则进行数据同步。

iv. 单击**下一步**。

7. 同步规则设置。

i. 在**同步规则设置**页签，进行同步**方案选择**。



同步方案描述如下表所示。

同步方案	描述
只全量一次性同步	只执行一次同步操作，将来源数据源的所有数据，全量同步至Elasticsearch中。
只增量一次性同步	只执行一次同步操作，按照指定的过滤条件，将来源数据源的增量数据同步至Elasticsearch中。
周期性全量同步	按照配置的周期任务，每次执行任务时都将来源数据源的所有数据，全量同步至Elasticsearch中。
周期性增量同步	按照指定的过滤条件和周期任务，每次执行任务时仅将增量数据同步至Elasticsearch中。
全量一次性同步后周期增量	先将来源数据源的所有数据全量同步至Elasticsearch，再按照指定的过滤条件和周期任务，后续每次执行任务时仅将增量数据同步至Elasticsearch中。

ii. 配置同步方案参数。

选择同步方案后，不同同步方案需配置不同的同步参数，包括**全量同步**参数、**增量同步**参数及**周期设置**参数。具体如下：

- **全量同步**

  仅当**方案选择**配置为**只全量一次性同步**、**周期性全量同步**或**全量一次性同步后周期增量**时，需要配置该参数。

参数	描述
写入前清空对应的原有index（索引）	取值如下： ■ **是**：写入数据前会清空索引中原有的数据。 ■ **否**：写入数据前不会清空索引中原有的数据。 **注意** 配置该参数为是时，则会在写入数据前删除目标索引中所有的数据，请谨慎选择。
写入Elasticsearch策略	取值如下： ■ **插入**：为默认值，同步数据时直接向目标索引中插入数据。 ■ **更新**：同步数据时，如果有相同的主键，则更新主键数据；如果没有相同的主键，则直接插入数据。 更新数据时，是先将原有的一行数据全部删除后再重新插入。
每批次写入的条数	每次批量写入Elasticsearch的数据条数，即攒够一定条数的数据后，一次性写入Elasticsearch。默认为**1000**。您可以根据实际网络情况及数据量大小进行合理配置，减少不必要的网络开销。

- **增量同步**

  仅当**方案选择**配置为**只增量一次性同步**、**周期性增量同步**或**全量一次性同步后周期增量**时，需要配置该参数。

参数	描述
写入ElasticSearch策略	取值如下： ■ **插入**：为默认值，同步数据时直接向目标索引中插入数据。 ■ **更新**：同步数据时，如果有相同的主键，则更新主键数据；如果没有相同的主键，则直接插入数据。 更新数据时，是先将原有的一行数据全部删除后再重新插入。
每批次写入的条数	每次批量写入Elasticsearch的数据条数，即攒够一定条数的数据后，一次性写入Elasticsearch。默认为**1000**。您可以根据实际网络情况及数据量大小进行合理配置，减少不必要的网络开销。
增量条件	对来源数据源进行增量内容同步的过滤条件。您可以参考调度参数概述进行配置。

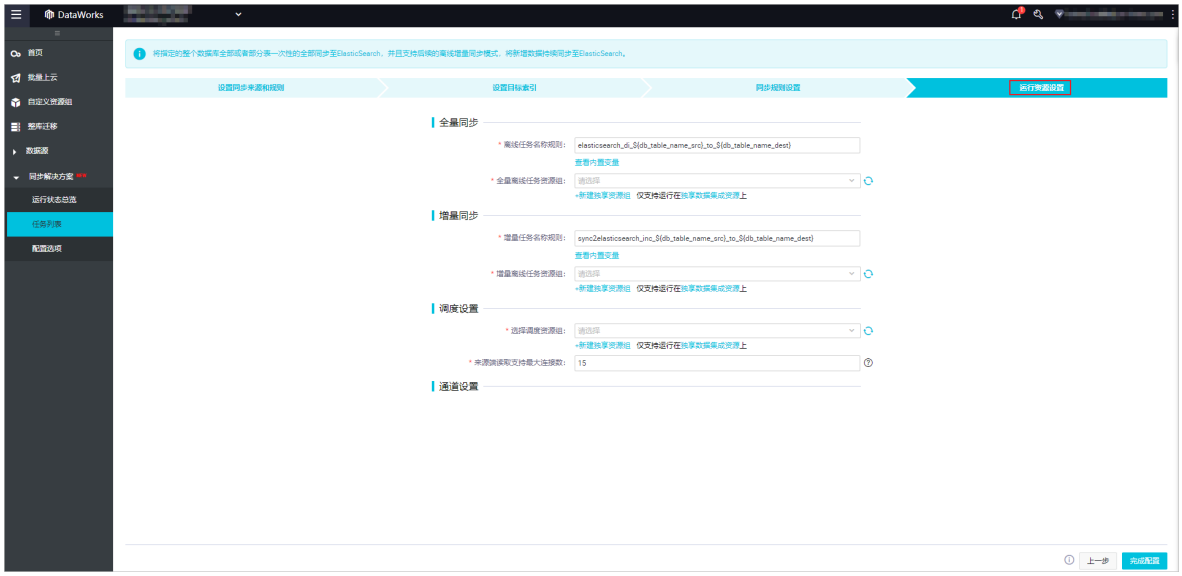
- **周期设置**

参数	描述
调度周期	包括**分钟**、**小时**、**日**、**周**及**月**，详细的配置说明请参考时间属性配置说明。
生效日期	调度任务将在配置的有效日期内生效并自动调度，在有效期外任务将不会自动调度。
暂停调度	暂停调度后，目标任务在配置的周期内不会执行。通常用于某个任务暂时不用执行，但后面还会继续使用的场景。
重跑属性	取值如下： ■ **运行成功或失败后皆可重跑**如果节点任务多次重跑不会影响结果，则可以选择该选项。 ■ **运行成功或失败后皆不可重跑**如果节点任务无论运行成功或失败，重跑都会影响结果，则可以选择该选项。 当选择该选项时，如果系统出现故障，则在故障恢复后系统也不会自动重跑节点任务。

iii. 单击**下一步**。

8. 运行资源设置。

在**运行资源设置**页签，配置各项参数。



- **全量同步**

仅当**同步规则设置**页签的**方案选择**配置为**只全量一次性同步**、**周期性全量同步**或**全量一次性同步后周期增量**时，需要配置该参数。

参数	描述
离线任务同步规则	全量同步时的离线任务名称。创建解决方案后，会生成一个离线任务用于读取全量数据。
全量离线任务资源组	目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **增量同步**

仅当**同步规则设置**页签的**方案选择**配置为**只增量一次性同步**、**周期性增量同步**或**全量一次性同步后周期增量**时，需要配置该参数。

参数	描述
增量任务名称规则	增量同步时的离线任务名称。创建解决方案后，会生成一个离线任务用于读取增量数据。
增量离线任务资源组	目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **调度设置**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为15。

9. 单击**完成配置**，完成整库离线同步任务的创建。

执行整库离线同步任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的整库离线同步任务。

查看任务运行状态及结果

- 在**解决方案任务列表**页面，选择已运行任务后的**更多 > 执行详情**，查看当前解决方案整库离线同步任务过程中，各子任务节点的运行详情。

步骤	说明	起始时间	结束时间	状态
1	批量创建Datahub Topic	2021-05-20 17:49:09	2021-05-20 17:49:12	成功 执行详情
2	创建DataWorks工作流程	2021-05-20 17:49:12	2021-05-20 17:49:12	成功
3	创建DataWorks虚拟节点	2021-05-20 17:49:12	2021-05-20 17:49:12	成功 执行详情
4	创建全量同步任务节点	2021-05-20 17:49:12	2021-05-20 17:49:15	成功 执行详情
5	提交发布DataWorks虚拟节点	2021-05-20 17:49:15	2021-05-20 17:49:18	成功 执行详情
6	提交发布全量同步任务节点	2021-05-20 17:49:18	2021-05-20 17:49:34	成功 执行详情
7	全量同步任务节点批量冒烟执行	2021-05-20 17:49:34	2021-05-20 17:50:42	成功 执行详情
8	创建DataWorks实时同步节点	2021-05-20 17:50:42	2021-05-20 17:50:43	成功 执行详情
9	提交发布DataWorks实时同步节点	2021-05-20 17:50:43	2021-05-20 17:50:44	成功
10	启动DataWorks实时同步节点	2021-05-20 17:50:44	2021-05-20 17:50:50	成功 执行详情

- 单击子任务节点后的**执行详情**，可以单击对话框中的任务链接，进入子节点的数据开发页面。

管理整库离线同步任务

- 查看或编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 查看配置**或**更多 > 修改配置**，可查看或编辑当前任务的配置信息。

  > **说明** 仅单击**未运行**状态后的**更多 > 修改配置**，您可以编辑任务。其它状态下的任务配置页面，仅支持查看。

- 修改任务优先级。

  单击相应任务后的**更多 > 修改优先级**。在**修改优先级**对话框中，输入需要配置的优先级数值，单击**确定**。优先级取值范围为1~8，数值越大优先级越高。

  > **说明** 优先级相同的任务，按照提交时间的先后顺序执行。

- 删除任务。

  单击相应任务后的**更多 > 删除**。在删除对话框中，单击**确认**。

  > **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

5.4.6. 配置查看整库实时同步任务

完成数据源、网络、资源的准备配置后，您可以创建并执行整库实时同步任务，开始进行数据同步。本文为您介绍如何创建整库实时同步任务，先将指定数据库中的部分或全部表的数据离线同步至Elasticsearch中，再将后续新增的数据实时同步至Elasticsearch中，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为MySQL）
- 配置数据源（来源为PolarDB）
- 添加数据源

背景信息

您可以使用通过Elasticsearch全托管方式提供的冷热存储方案存储企业的实时数据。基于该功能，DataWorks为您提供了整库实时同步数据至Elasticsearch的解决方案，轻松助力企业实现同步整库数据至Elasticsearch后，再将持续新增的数据实时同步至Elasticsearch中。同时，您可以实时查看创建的同步任务详情，监控任务的运行状况及业务库数据的更新情况，用于后续做数据检索、数据分析或数据开发。

整库实时同步数据的优势如下：

- 整库级别同步。

  无需逐个创建表级别的数据同步任务，支持直接创建库级别的同步任务，选择目标库的部分或全部表数据进行同步。

- 同步规则配置灵活。
  - 您可以根据业务需求灵活配置不同DDL消息的处理规则。例如，针对来源端删除表的DDL消息，如果您将同步数据的处理方式配置为**忽略**，则进行实时同步时，DataWorks收到相关删除表的DDL消息时，会忽略该类消息，目标端的对应表则不会被删除。
  - 您可以编辑已经配置运行的同步任务，为该任务快速添加表或删除已有同步表。
  - 您可以根据业务需求，配置目标索引的同步规则，选择是否添加同步源表的实时新增字段至目标索引。添加字段至目标索引后，该字段后续可被搜索。

- 配置操作简单。

  您无需进行创建同步任务、创建数据库、创建表、创建相互依赖以及执行参数对齐等复杂操作，通过简单的产品配置向导，即可完成对应功能的配置。

- 实现海量数据的实时更新，自动化运维管理效率较高。

适用场景

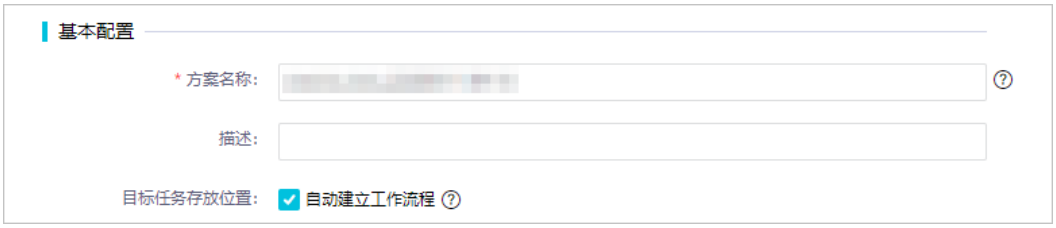
适用于需要实时监测业务库数据的更新情况，便于上层应用对实时数据进行检索分析或数据开发的场景。

使用限制

- 目前仅支持整库实时同步MySQL、PolarDB类型的数据库至Elasticsearch。
- 整库实时同步解决方案仅支持使用独享资源组。

创建整库实时同步任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。
   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 在**新建同步解决方案**对话框中，单击**一键实时同步至Elasticsearch**。
4. 完成方案名称等基本信息配置。
   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

5. 选择来源数据源并配置同步规则。
   i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明**
> 目前仅支持整库实时同步MySQL、PolarDB类型的数据库至Elasticsearch。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择同步目标数据源的部分或全部表。

> **注意** 如果选中的表没有主键，则在创建源表和目标Elasticsearch索引的映射关系时，需要为该表自定义主键（例如，使用非主键的一个或几个字段的联合代替主键）进行同步数据时去重，详情请参见选择目标数据源并配置目标索引。

iii. 在**设置表名到索引名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标索引名转换规则**和**目标索引名规则**：

- **源表名和目标索引名转换规则**：转换表名为目标索引名，进行字符串替换。
- **目标索引名规则**：支持对转换后的索引名添加前缀和后缀。

iv. 单击**下一步**。

6. 选择目标数据源并配置目标索引。

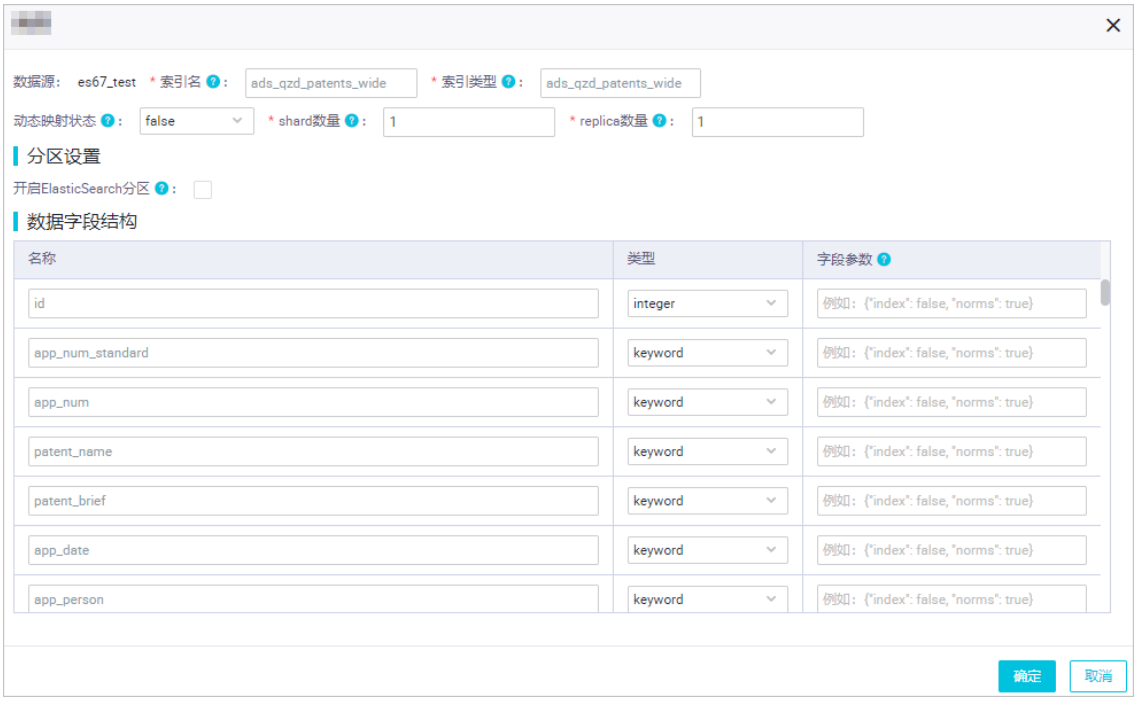
i. 在**设置目标索引**页签，选择**目标Elasticsearch数据源**。

ii. 单击**刷新源表和Elasticsearch索引映射**，创建需要同步的源表和目标Elasticsearch索引的映射关系。

iii. 查看任务的执行进度和表来源。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	选择的**索引建立方式**，取值如下： ■ 当**索引建立方式**选择**自动建索引**时，**Elasticsearch索引名**列显示自动创建的Elasticsearch索引名称。您可以单击索引名称，修改建立索引的相关配置。 ■ 当**索引建立方式**选择**使用已有索引**时，请在**Elasticsearch索引名**列对应的下拉列表中，选择需要使用的索引。同时您可以单击**设置同步规则**，查看源表字段与目标索引的映射情况。

当**索引建立方式**选择**自动建索引**时，您可以单击创建的Elasticsearch索引名称，根据业务需求修改目标索引的相关参数。



- **动态映射状态**：用于在同步数据时，是否将数据源表的新增字段添加至目标索引中。取值如下：
  - **true**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，会将该字段同步添加至目标索引中，后续该字段可以被搜索。该取值为默认值。
  - **false**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，会将该字段同步添加至目标索引中，但后续该字段不可以被搜索。
  - **strict**：当映射的目标索引检测到同步的数据源表中存在新添加的字段时，将拒绝同步该字段至目标索引并产生异常报错，您可以在日志信息中查看报错详情。

  更多动态映射的内容，详情请参见动态映射。
- **shard数量及replica数量**：索引的主分片及副本分片，用于将一个完整的索引分成多个分片，分布至不同的Elasticsearch节点上，构成分布式搜索，提升Elasticsearch的查询效率，详情请参见基本概念。

  > **说明** shard数量及replica数量参数，配置运行后不可更改，默认取值为*1*。
- **分区设置**：您可以选择同步的数据源表的某一列做为分区列，该参数需要和**shard数量**、**replica数量**配合使用。默认不开启分区功能。
- **数据字段结构**：用于设置映射的目标索引中字段的类型及扩展属性，详情请参见Elasticsearch的字段类型。

> **说明** 当创建了目标Elasticsearch索引后，如果不修改相关参数，则系统会按照默认值的相应规则进行数据同步。

iv. 单击**下一步**。

7. DDL消息处理规则设置。

   来源数据源（MySQL、PolarDB、SQLServer）会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的规则。

> **说明** 该规则为初次执行实时同步任务时的DDL消息处理规则，后续如果您需要修改规则，可以手动停止任务，进入实时任务的配置页面修改，详情请参见管理整库实时同步任务。

i. 在**DDL消息处理规则**页签，配置**实时同步DDL消息处理策略**。

不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

8. 运行资源设置。

在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

9. 单击**完成配置**，完成整库实时同步任务的创建。

执行整库实时同步任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的整库实时同步任务。

查看任务运行状态及结果

- 在**解决方案任务列表**页面，选择已运行任务后的**更多 > 执行详情**，查看当前解决方案整库实时同步任务过程中，各子任务节点的运行详情。

步骤	说明	起始时间	结束时间	状态
1	批量创建Datahub Topic	2021-05-20 17:49:09	2021-05-20 17:49:12	成功 执行详情
2	创建DataWorks工作流程	2021-05-20 17:49:12	2021-05-20 17:49:12	成功
3	创建DataWorks虚拟节点	2021-05-20 17:49:12	2021-05-20 17:49:12	成功 执行详情
4	创建全量同步任务节点	2021-05-20 17:49:12	2021-05-20 17:49:15	成功 执行详情
5	提交发布DataWorks虚拟节点	2021-05-20 17:49:15	2021-05-20 17:49:18	成功 执行详情
6	提交发布全量同步任务节点	2021-05-20 17:49:18	2021-05-20 17:49:34	成功 执行详情
7	全量同步任务节点批量冒烟执行	2021-05-20 17:49:34	2021-05-20 17:50:42	成功 执行详情
8	创建DataWorks实时同步节点	2021-05-20 17:50:42	2021-05-20 17:50:43	成功 执行详情
9	提交发布DataWorks实时同步节点	2021-05-20 17:50:43	2021-05-20 17:50:44	成功
10	启动DataWorks实时同步节点	2021-05-20 17:50:44	2021-05-20 17:50:50	成功 执行详情

- 单击子任务节点后的**执行详情**，可以单击对话框中的任务链接，进入子节点的数据开发页面。

管理整库实时同步任务

- 查看任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 查看配置**，可查看任务的配置信息。

- 编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 修改配置**，可编辑任务的配置信息。

  对于已运行成功的实时同步任务，您可以单击相应任务后的**更多 > 修改配置**，快速添加或删除同步表，以及动态添加列：

  - 快速添加或删除同步表。

    在**设置同步来源和规则**页签的**选择同步的源表**区域，选择添加或删除目标同步表。修改配置后，保存并运行同步任务，则可以快速添加新表或删除已有同步表。

  - 动态添加列。
    - 配置**DDL消息处理规则**的**新增列**为**正常处理**，实时同步任务将会自动监听同步源表列的变化，如果识别到有新增列，则自动将该列的数据同步至目标Elasticsearch索引。详情请参见配置DDL消息处理规则。
    - 配置目标Elasticsearch索引的**动态映射状态**为**true**时，则目标Elasticsearch索引会将同步源表中的新增列实时添加至索引中，后续该列可被搜索。详情请参见目标索引配置规则。

- 修改任务优先级。

  单击相应任务后的**更多 > 修改优先级**。在**修改优先级**对话框中，输入需要配置的优先级数值，单击**确定**。优先级取值范围为1~8，数值越大优先级越高。

  > **说明** 优先级相同的任务，按照提交时间的先后顺序执行。

- 删除任务。

  单击相应任务后的**更多 > 删除**。在删除对话框中，单击**确认**。

  > **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

5.5. 同步数据至Hologres

5.5.1. 资源规划与配置

当前使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次进入管控后弹出的界面弹框中进行授权。
3. 绑定专有网络VPC。
   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

   > **说明** 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。
4. （可选）配置Host。

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明
- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。目前同步数据至MaxCompute的来源数据源仅支持PolarDB、Oracle及MySQL，您可以根据实际需求选择合适的数据源。详细的来源数据源配置请参见配置数据源（来源为PolarDB）、配置数据源（来源为Oracle）或配置数据源（来源为MySQL）。

5.5.2. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至Hologres时，您需要参考本文在数据源中配置好网络、白名单、权限等配置，为后续的数据同步方案执行做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源Hologres。本文以阿里云PolarDB MySQL作为来源数据源进行示

例。

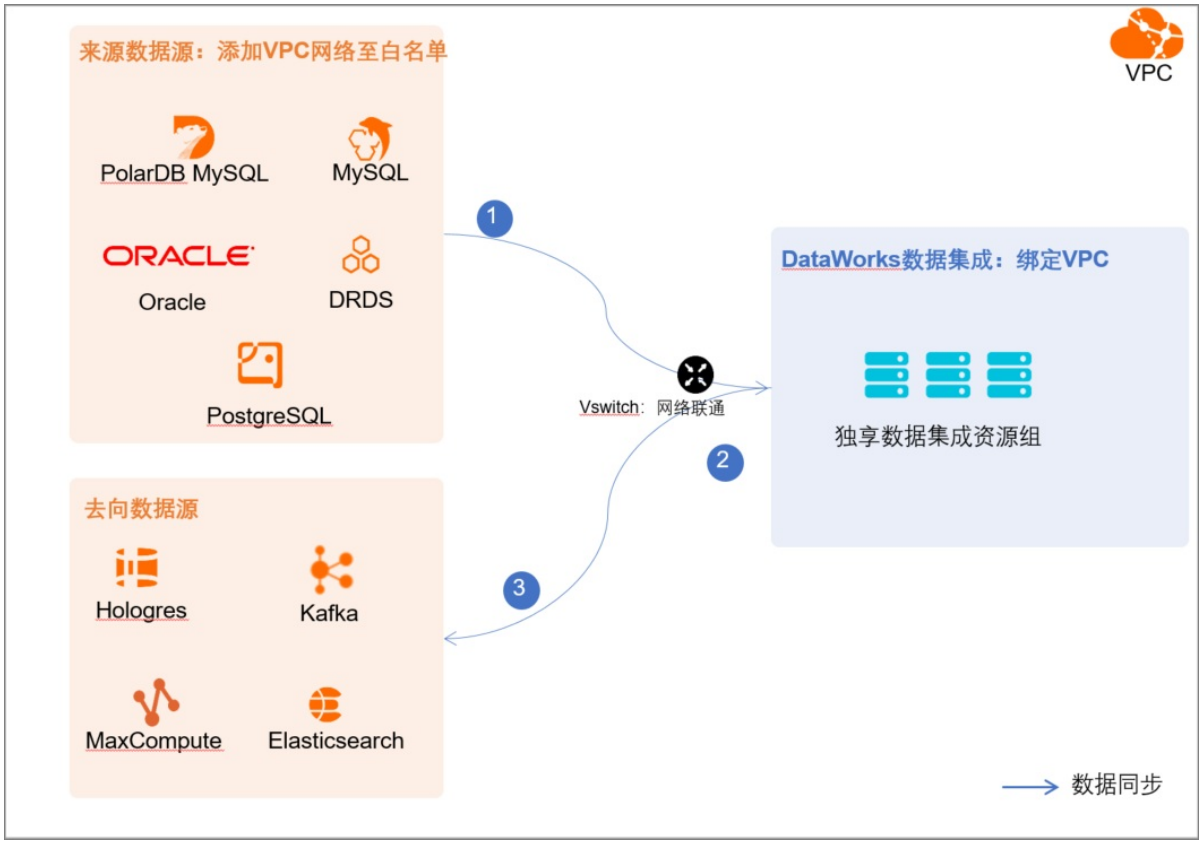
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

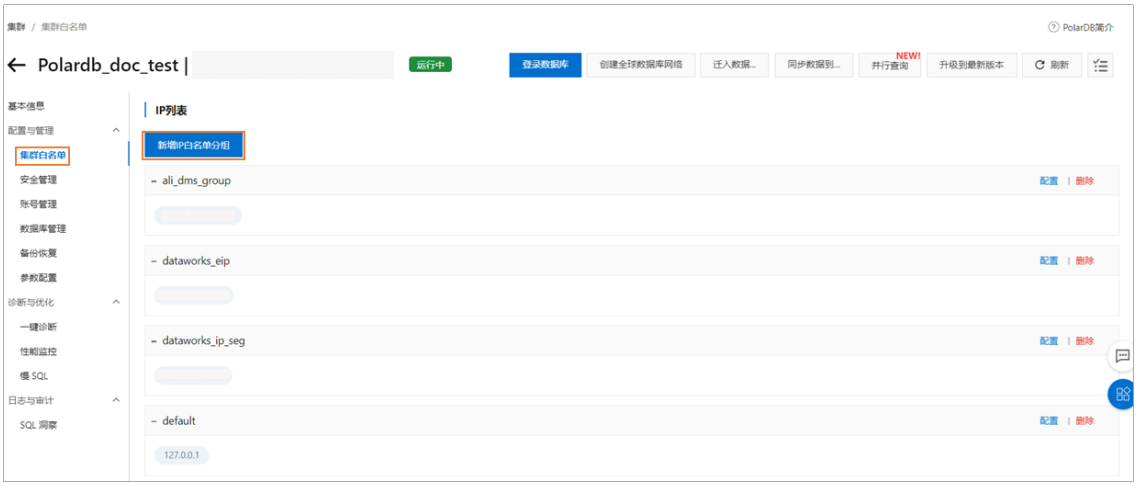
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.5.3. 配置数据源（来源为Oracle）

同步Oracle的数据至Hologres时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

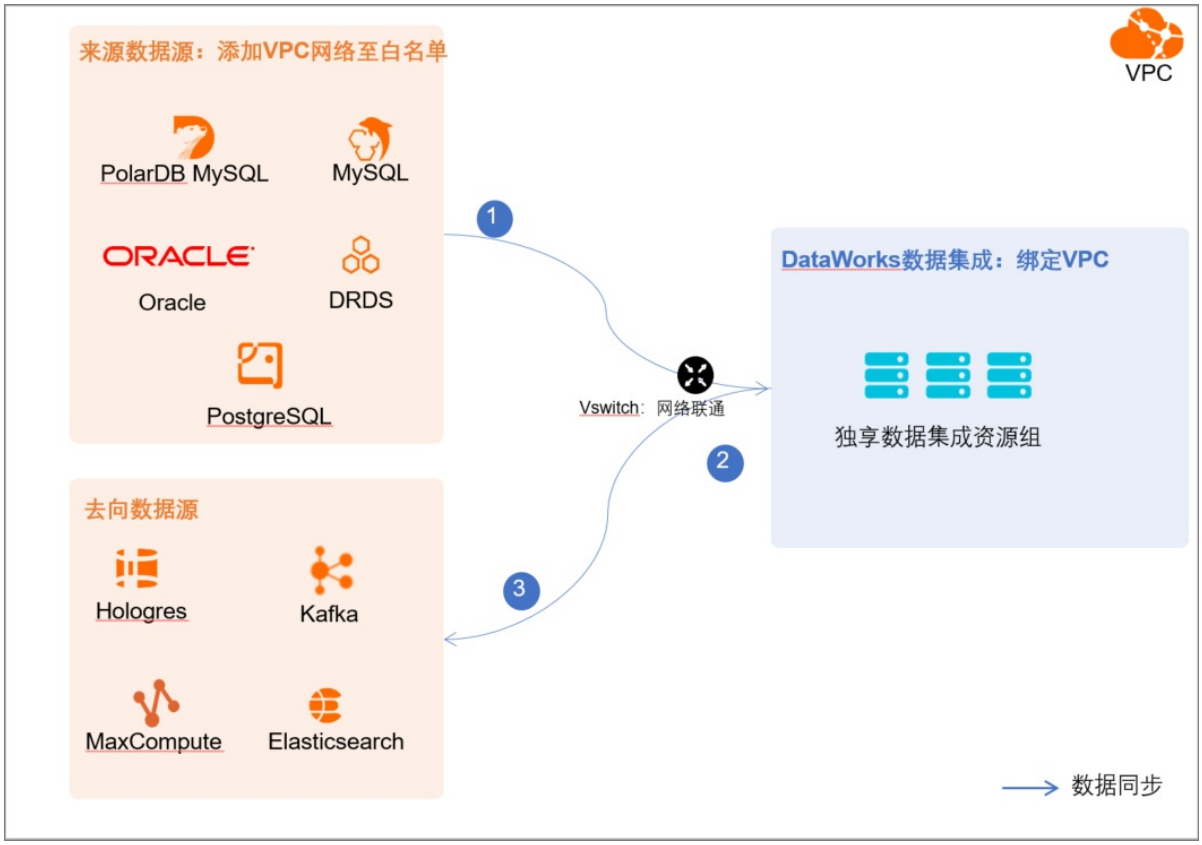
- 准备数据源：已购买来源数据源Oracle、去向数据源Hologres。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。

- Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

操作步骤

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.5.4. 配置数据源（来源为MySQL）

同步MySQL的数据至Hologres时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源Hologres。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。

- 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

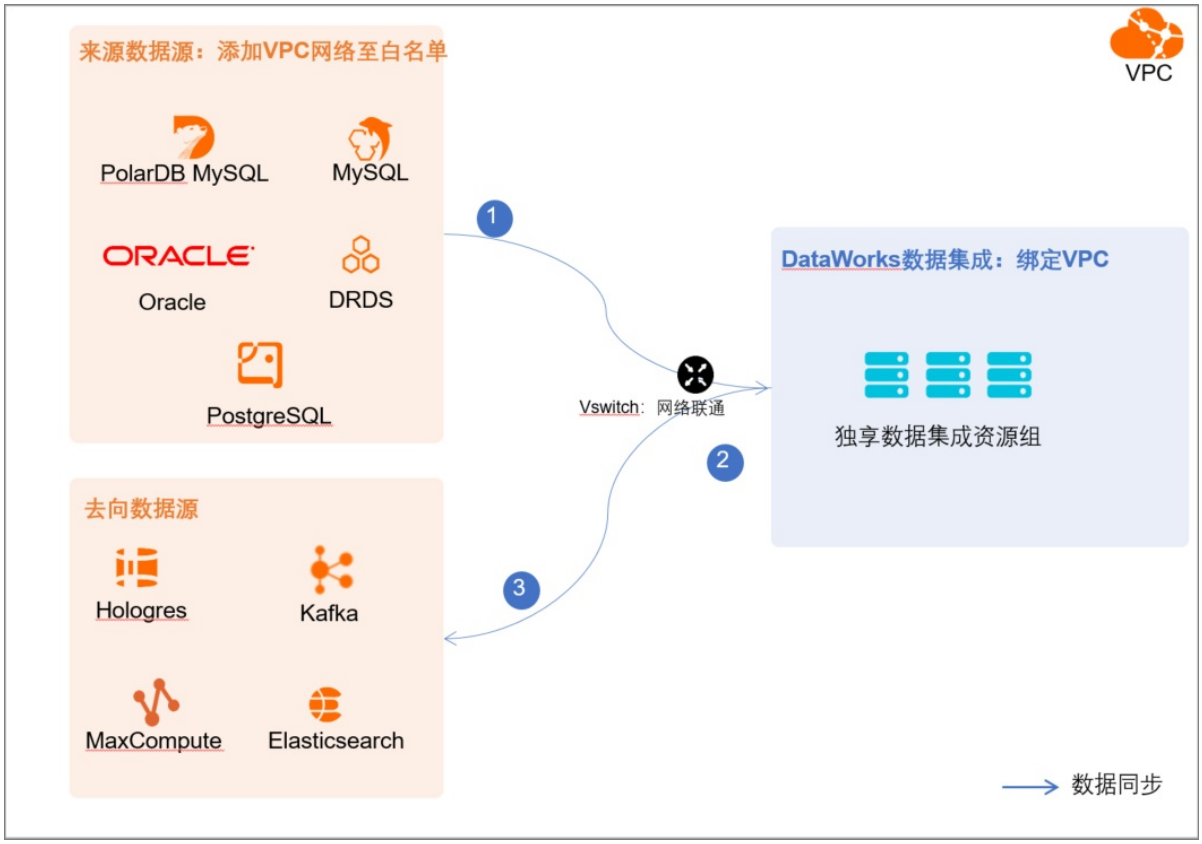
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。

- Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

   a. 登录DataWorks控制台。

   b. 在左侧导航栏，单击**资源组列表**。

   c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

   d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

   e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

   f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

   i. 创建账号。

   操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.5.5. 配置数据源（来源为DRDS）

同步DRDS的数据至Hologres时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源DRDS、去向数据源Hologres。

> **说明** 当前对来源数据源DRDS有如下限制要求：
> - 实例类型：仅支持DRDS1.0非只读实例。
> - 目前仅支持使用实例模式配置的数据源，如果您使用JDBC连接串配置的数据源，任务运行将会出错。
> - 实例的存储类型：仅支持PolarDB（即租户侧PolarDB MySQL）和存量的用户RDS（新购已不支持），不支持RDS MySQL（即私有定制RDS MySQL）。
>
> 创建DRDS 1.0实例的操作可参见云原生分布式数据库DRDS文档的*购买DRDS实例*创建实例。

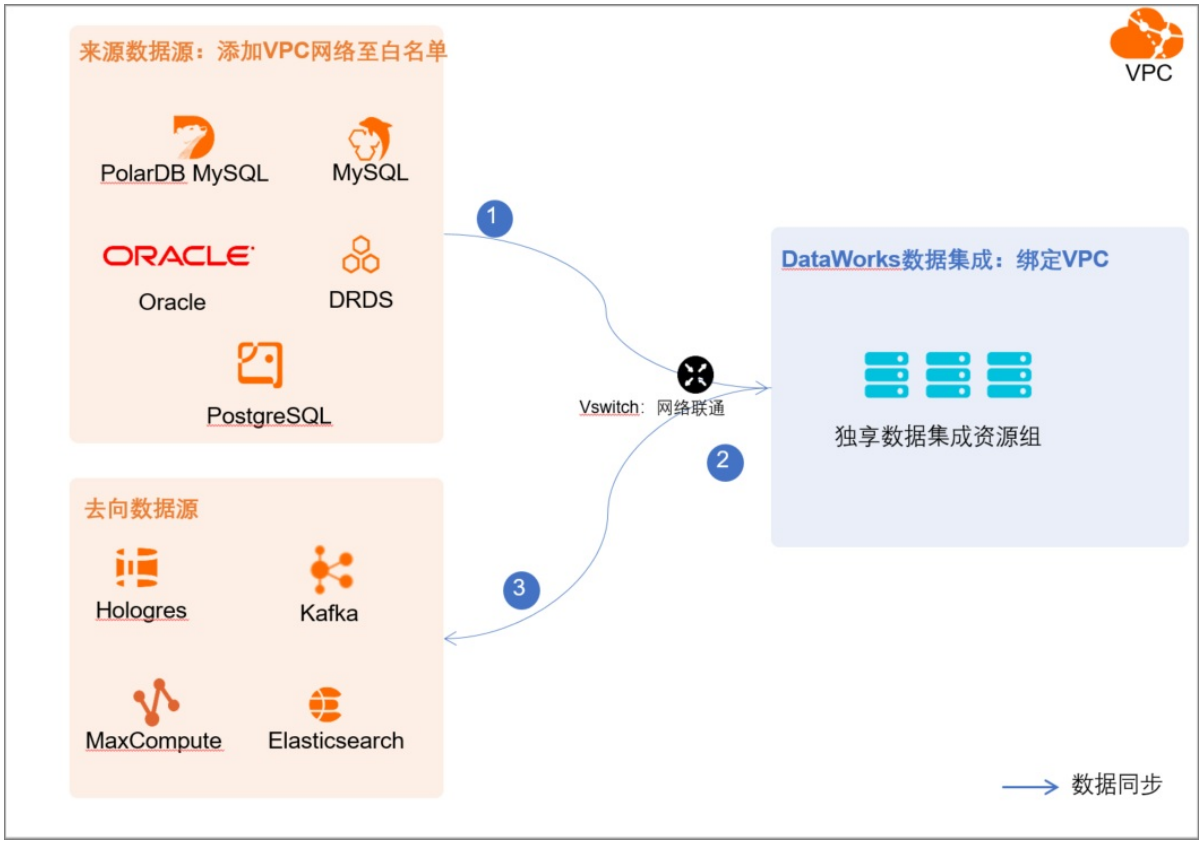
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至DRDS的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至DRDS的白名单中。

操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，创建账号的操作详情可参见账号管理。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.5.6. 配置数据源（来源为PostgreSQL）

同步PostgreSQL的数据至Hologres时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

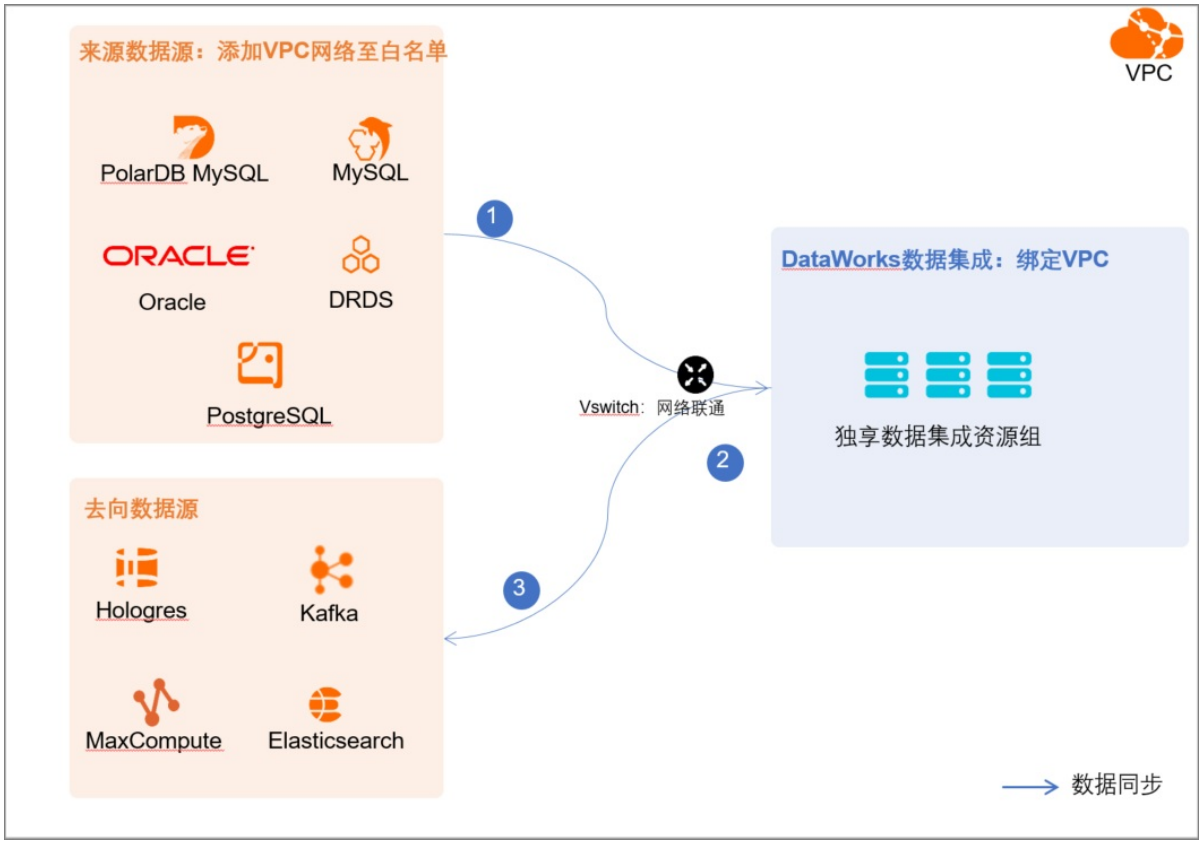
- 准备数据源：已购买来源数据源PostgreSQL、去向数据源Hologres。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  目前仅支持配置PostgreSQL数据源为PostgreSQL `10`、`11`、`12`、`14.1`版本。您可以通过如下语句查看PostgreSQL数据库的版本。

  ```
  show server_version
  ```

使用限制

数据集成实时同步任务存在如下约束与限制：

- 数据集成对`ADD COLUMN`进行了特别支持：
  - 约束：`ADD COLUMN`时不能有`ADD COLUMN`和`DROP COLUMN`或者其他DDL的组合。

    > 注意 `ADD COLUMN`时其他`DROP COLUMN`、`RNAME COLUMN`等`ALTER COLUMN`的行为将使数据同步任务不能正常工作。

  - 限制：除了`ADD COLUMN`外，无法识别用户的其他DDL操作。
- 不支持`ALTER TABLE/CREATE TABLE`。
- 不支持TEMPRORARY表和UNLOGGED表复制，PostgreSQL数据库没有提供机制对这两种类型的表进行log解析订阅。
- 不支持Sequences复制（`serial/bigserial/identity`）。

- 不支持TRUNCATE操作。
- 不支持大对象复制（Bytea）。
- 不支持视图、物化视图、外部表复制。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PostgreSQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

      a. 登录DataWorks控制台。

      b. 在左侧导航栏，单击**资源组列表**。

      c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

      d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

      e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

      f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至PostgreSQL实例的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `REPLICATION`、`LOGIN` 权限。

   > **说明**
   >
   > 实时同步只支持逻辑复制机制，逻辑复制使用发布和订阅模型，其中一个或多个订阅者订阅发布者节点上的一个或多个发布。订阅者从他们订阅的发布中提取数据。
   >
   > 表的逻辑复制通常从对发布者数据库上的数据进行快照并将其复制到订阅者开始。完成后，发布者上的更改会实时发送给订阅者。

   i. 创建账号。

      操作详情请参见创建数据库和账号。

ii. 配置权限。

检查账号是否有 `replication` 权限。

```
select userepl from pg_user where usename='xxx'
```

预期返回结果为True，返回False则表示无权限，您可以通过如下语句进行授权。

```
ALTER USER <user> REPLICATION;
```

3. 检查是否支持备库。

```
SELECT pg_is_in_recovery()
```

目前仅支持主库，预期返回结果为False，返回True时表示是备库，实时同步不支持备库，需修改数据源配置信息为主库的信息，请参见配置PostgreSQL数据源。

4. 检查 `wal_level` 是否为 `logical` 。

```
show wal_level
```

`wal_level` 指定了 `wal_log` 的级别，预期返回结果为logical，否则不支持逻辑复制机制。

5. 检查是否可以启动 `wal_sender` 进程。

```
-- 查询 max_wal_senders
show max_wal_senders;
-- 查询 pg_stat_replication 数量
select count(*) from pg_stat_replication
```

当 `max_wal_senders` 不为空，且 `max_wal_senders` 值大于 `pg_stat_replication` 数量时，则表示有空闲可用的 `wal_sender` 进程。PostgreSQL数据库会为同步数据程序启动 `wal_sender` 进程来给订阅者发送日志。

5.5.7. 添加数据源

将来源数据源的数据同步至Hologres数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：Oracle

添加Oracle数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置Oracle数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：PolarDB-X（原DRDS）

添加PolarDB-X（原DRDS）数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置DRDS数据源。

添加去向数据源：Hologres

操作详情可参见配置Hologres数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置查看数据同步任务。

5.5.8. 配置查看数据同步任务

完成数据源、网络、资源的准备配置后，您可创建并执行数据同步任务，开始进行数据同步。本文为您介绍如何创建数据同步任务，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

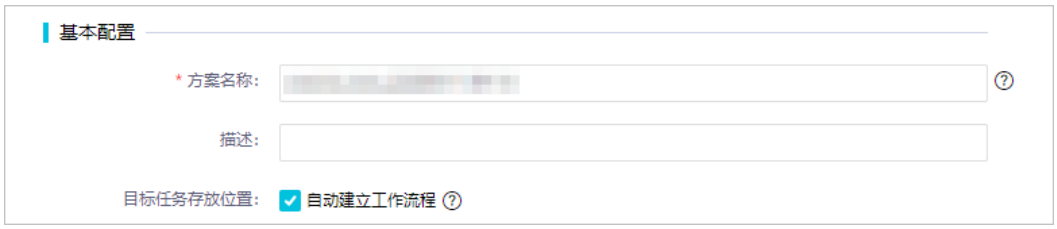
- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 配置数据源（来源为DRDS）
- 添加数据源

创建同步解决方案任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 在**新建同步解决方案**对话框中，单击**一键实时同步至Hologres**。
4. 完成方案名称等基本信息配置。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

5. 选择来源数据源并配置同步规则。

i. 在**数据来源**区域，选择**类型**和**数据源**。

> **说明** 仅支持选择MySQL、Oracle和PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 > 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

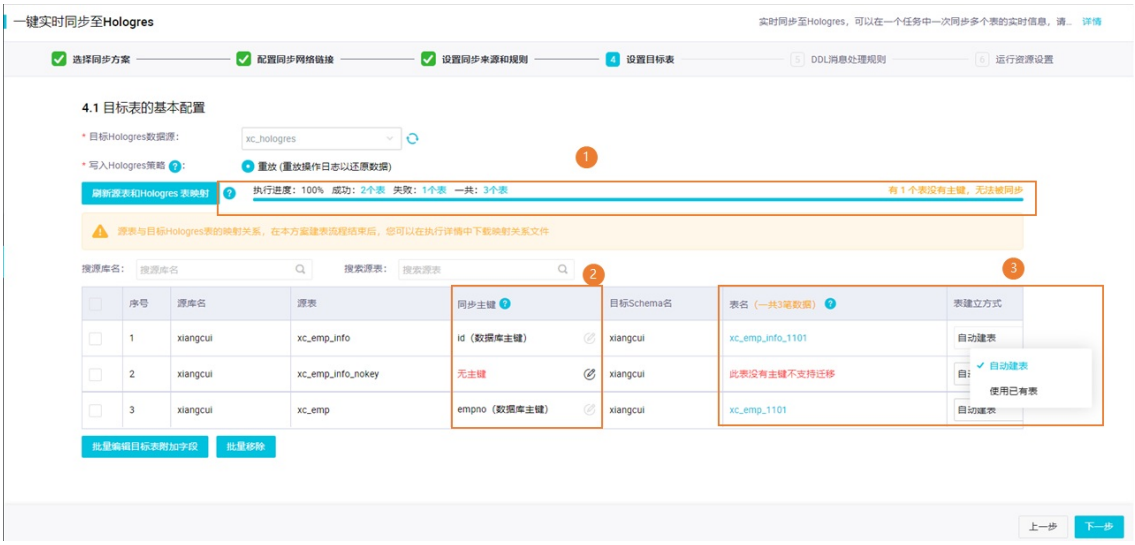
iv. 单击**下一步**。

6. 选择目标数据源并配置目标表格式。

i. 在**设置目标表**页面，选择**目标Hologres数据源**和**写入Hologres策略**。

ii. 单击**刷新源表和Hologres表映射**，创建需要同步的源表和目标Hologres表的映射关系。

iii. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	同步Hologres时，不支持同步无主键表，如果源端表无主键，您可以在此处通过自选该表主键来满足同步要求。
③	支持**自动建表**和**使用已有表**。 选择的**表建立方式**，取值如下： ■ 当**表建立方式**选择**使用已有表**时，**Hologres表名**列显示自动创建的Hologres表名称。您也可以在下拉列表中选择需要使用的表名称。 ■ 当**表建立方式**选择**自动建表**时，显示自动创建的Hologres表名称。您可以单击表名称，查看和修改建表语句。

iv. 单击**下一步**。

7. 运行资源设置。

在**运行资源设置**页面，配置各项参数。

参数	描述
选择实时任务独享资源组	分别选择实时任务和全量离线任务需要使用的独享资源组。目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情可参见资源规划与配置。
选择全量离线任务独享资源组	
选择调度资源组	选择运行任务时使用的调度资源组。

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于读取全量数据，再生成实时任务持续读取实时增量数据。

8. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**开始执行**，运行创建的数据同步解决方案任务。

如果任务执行失败，您可查看任务运行失败的错误提示，参考以下常见问题进行排查处理。

- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.
- 离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

管理数据同步解决方案任务

- 查看或编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**任务配置**，查看或编辑任务。

  > **说明** 仅单击**未运行**状态后的**任务配置**，您可以编辑任务。其它状态下的任务配置页面，仅支持查看。

- 删除任务。

  单击相应任务后的**删除**。在**删除**对话框中，单击**确定**。

  > **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

5.5.9. 增加或删除已运行任务的同步表

同步数据至Hologres解决方案为您提供了一键增加及删除同步表功能，方便您为已成功配置运行的同步任务快速添加新表或删除已有同步表。本文为您介绍如何增加或删除已运行任务的同步表。

前提条件

已创建并运行同步数据至Hologres解决方案的任务，详情请参见配置查看数据同步任务。

同步任务新增表

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。

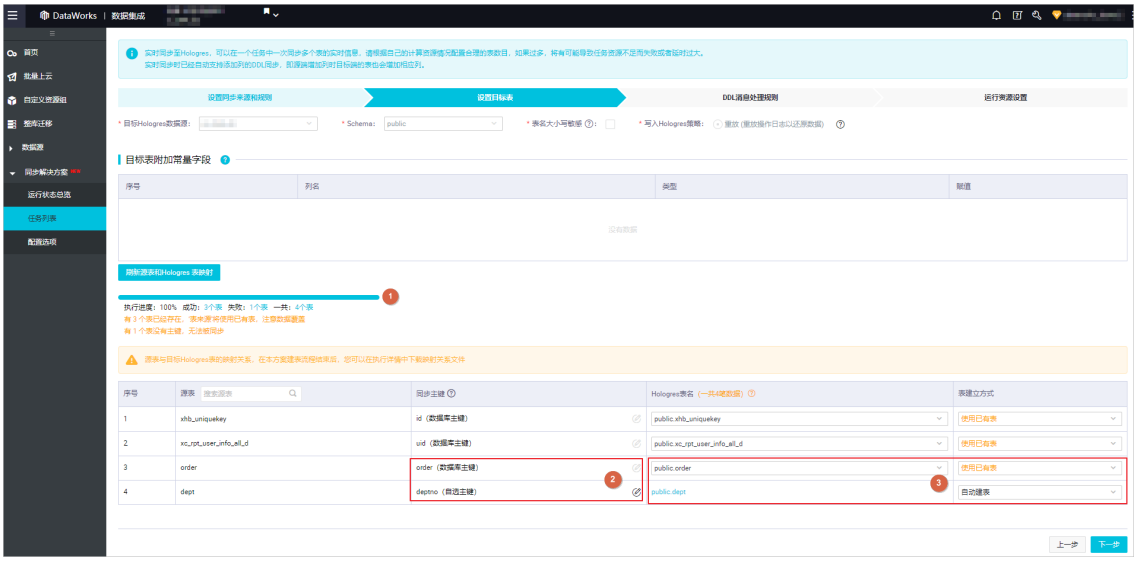
2. 在**解决方案任务列表**页面，选择目标同步任务后的**更多 > 修改配置**，进入任务配置页面。
3. 新增同步源表并更新源表至目标表的映射关系。

i. 在**设置同步来源和规则**页面的**选择同步的源表**区域，选中需要新增的同步**源表**，单击 > 图标，将其移动至**已选源表**。

ii. 单击**下一步**。

iii. 在**设置目标表**页面，单击**刷新源表和Hologres表映射**，更新需要同步的源表和目标Hologres表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	选择的**表建立方式**，取值如下： ■ 当**表建立方式**选择**使用已有表**时，**Hologres表名**列显示自动创建的Hologres表名称。您也可以在下拉列表中选择需要使用的表名称。 ■ 当**表建立方式**选择**自动建表**时，显示自动创建的Hologres表名称。您可以单击表名称，查看和修改建表语句。

4. 单击**下一步**。
5. 配置**DDL消息处理规则**。

   来源数据源（例如，MySQL）会包含许多DDL操作，进行实时同步时，您可以在**DDL消息处理规则**页面，根据业务需求，修改不同类别DDL消息同步至目标端的规则。

i. 配置**实时同步DDL消息处理策略**。

不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

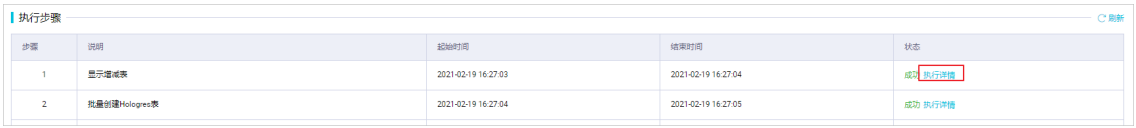
参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

7. 单击**完成配置**，返回**解决方案任务列表**页面。
8. 单击上述修改任务**操作**列的**更多 > 提交执行**在**提交执行**对话框，单击**确定**，运行当前任务。

   提交执行任务时，会和上一次运行成功的任务对应的表做对比，当发现新增表时则会执行新增表的添加流程。

> **说明** 重置实时同步任务位点并启动运行时，会存在一个新增表追加变更数据的过程，即任务位点时间重置到新增表全量数据初始化时的时间。例如，您的同步任务8点开始运行，到9点时运行未结束。9点时新增了一张表，则全量数据初始化在9点开始执行，此过程耗时1小时，即全量数据初始化在10点完成。此时，已经正在运行的实时同步任务会先停止，然后重置任务位点到9点，进行追加增量数据，9点到10点之间所有变更表的增量数据会被重新同步至Hologres目标表，一键新增表只保证数据的最终一致性。

9. 查看同步任务新增表详情。
   i. 进入**任务列表**页面，单击目标任务**操作**列的**执行详情**，进入任务的执行详情页。
   ii. 在**执行步骤**区域，单击**显示增减表**后的**执行详情**。



   显示增减表的状态为**成功**时，表明新增的表已成功添加至同步任务。
   iii. 查看同步任务新增的同步表。

同步任务删除表

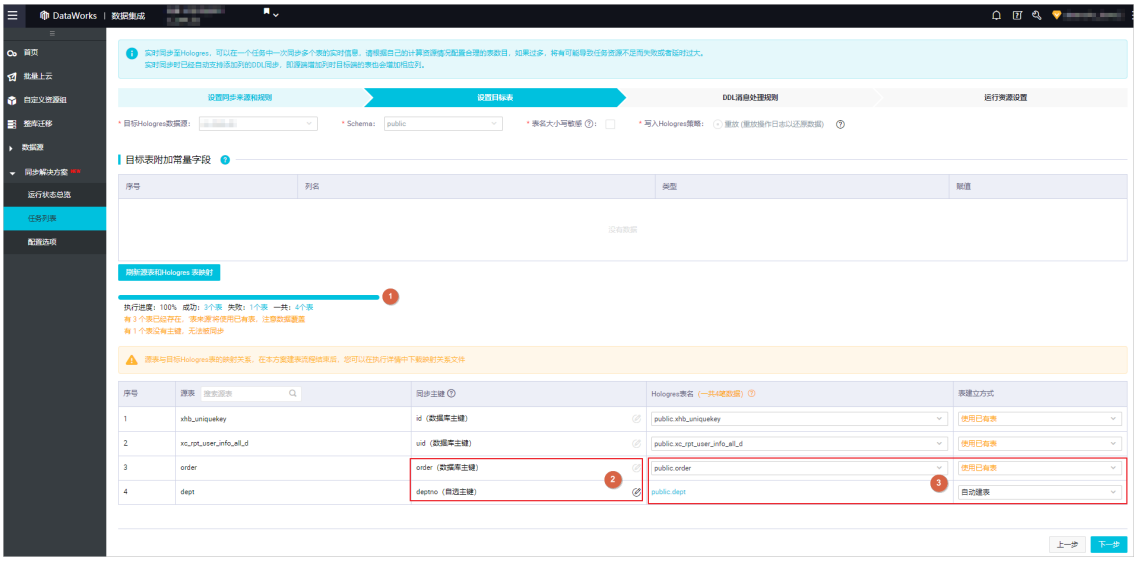
1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。
   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，选择目标同步任务后的**更多 > 修改配置**，进入任务配置页面。
3. 删除同步源表并更新源表至目标表的映射关系。

i. 在**设置同步来源和规则**页面的**选择同步的源表**区域，选中需要删除的**已选源表**，单击<图标，将其移回至**源表**。

ii. 单击**下一步**。

iii. 在**设置目标表**页面，单击**刷新源表和Hologres表映射**，更新需要同步的源表和目标Hologres表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	选择的**表建立方式**，取值如下： ■ 当**表建立方式**选择**使用已有表**时，**Hologres表名**列显示自动创建的Hologres表名称。您也可以在下拉列表中选择需要使用的表名称。 ■ 当**表建立方式**选择**自动建表**时，显示自动创建的Hologres表名称。您可以单击表名称，查看和修改建表语句。

4. 单击**下一步**。
5. 配置**DDL消息处理规则**。

   来源数据源（例如，MySQL）会包含许多DDL操作，进行实时同步时，您可以在**DDL消息处理规则**页面，根据业务需求，修改不同类别DDL消息同步至目标端的规则。

i. 配置**实时同步DDL消息处理策略**。

不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

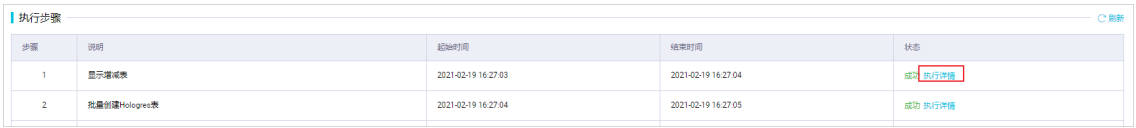
参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

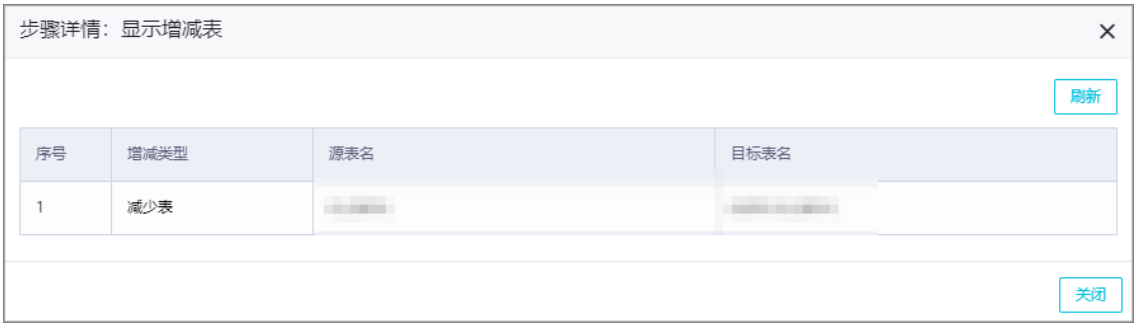
7. 单击**完成配置**，返回**解决方案任务列表**页面。
8. 单击上述修改任务**操作**列的**更多 > 提交执行**在**提交执行**对话框，单击**确定**，运行当前任务。

   删除已运行同步任务中的源表时，会将目标源表从实时同步任务中移除。重新提交执行同步任务时，会直接在重启同步任务的时间点继续同步数据。
9. 查看同步任务删除表详情。
   i.
   ii. 在**执行步骤**区域，单击**显示增减表**后的**执行详情**。



显示增减表的状态为**成功**时，表明目标源表已成功从同步任务中删除。

iii. 查看此次执行任务删除的同步表。



5.5.10. 常见问题

以下为您介绍同步数据至Hologres解决方案操作失败的常见问题和解决方案。

- PolarDB数据源网络联通性测试失败怎么办？
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.
- 离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.
- 离线任务，运行报错：errorCode:NoSuchTopic, errorMessage:The specified topic name does not exist.

PolarDB数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源PolarDB时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

Oracle数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源Oracle时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

MySQL数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源MySQL时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX` 。
- 可能原因：来源数据源PolarDB没有开启binlog。
- 如何处理：PolarDB开启binlog，详细操作可参见配置数据源（来源为PolarDB）。并进行至少一条数据的变更，同时切换数据集成实时同步开始点位到当前时间。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation` 。
- 可能原因：来源数据源PolarDB没有给进行数据同步的账号开启所需权限，或对接的PolarDB数据库不是主节点。
- 如何处理：参见配置数据源（来源为PolarDB）的操作授予权限，或者检查PolarDB是否是主节点（读写库），目前实时任务不支持从PolarDB备节点抓取数据。

实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first` 。
- 可能原因：来源数据源PolarDB未打开loose_polar_log_bin参数。
- 如何处理：需要打开loose_polar_log_bin参数，详细操作可参见配置数据源（来源为PolarDB）。

离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.

- 报错内容：数据同步任务执行时，离线任务运行失败，错误提示为 `com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns` 。
- 可能原因：DataWorks的离线引擎插件未升级到最新版本。
- 如何处理：请提交工单联系技术支持，帮您将离线引擎插件升级到最新版本。

离线任务，运行报错：errorCode:NoSuchTopic, errorMessage:The specified topic name does not exist.

- 报错内容：执行数据同步任务时，离线任务运行失败，错误提示为 `errorCode:NoSuchTopic, errorMessage:The specified topic name does not exist.` 。
- 可能原因：
  - 数据同步任务中使用的Hologres目标表不存在。
  - 使用数据同步任务同步了数据源表至Hologres的外部表。目前，Hologres Writer不支持写入数据至Hologres外部表。
- 如何处理：您需要使用Hologres的内部表作为同步任务的目标表。如果Hologres目标表不存在，请在配置数据同步任务时使用**自动建表**，创建可用的Hologres目标表，详情请参见配置Hologres目标表。

5.6. 同步数据至AnalyticDB MySQL 3.0

5.6.1. 资源规划与配置

当前使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表** > **独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

   > **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次进入管控后弹出的界面弹框中进行授权。
3. 绑定专有网络VPC。

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

说明 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。

4. （可选）配置Host。

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

   如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   > **说明** 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

   i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

   ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。目前同步数据至AnalyticDB MySQL 3.0的来源数据源仅支持PolarDB及MySQL，您可以根据实际需求选择合适的数据源。详细的来源数据源配置请参见配置数据源（来源为PolarDB）、配置数据源（来源为OceanBase）或配置数据源（来源为MySQL）。

5.6.2. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至AnalyticDB MySQL 3.0时，您需要参考本文在数据源中配置好网络、白名单、权限等配置，为后续的数据同步方案执行做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

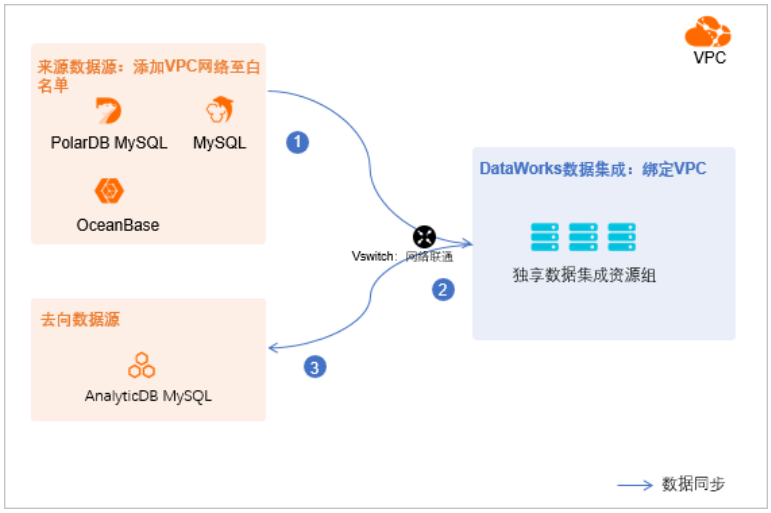
- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源AnalyticDB MySQL 3.0。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将独享数据集成资源组在网络配置时所绑定的交换机网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

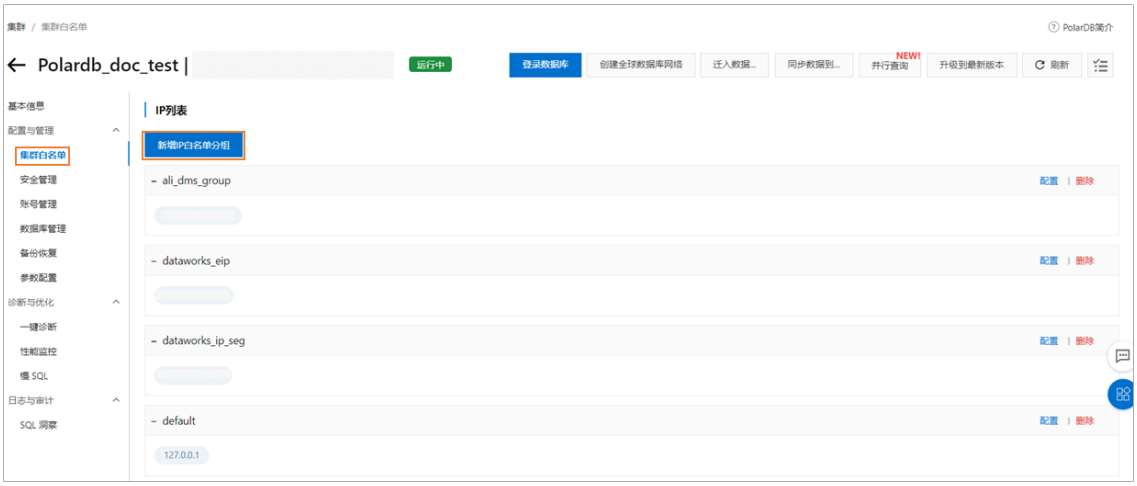
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.6.3. 配置数据源（来源为MySQL）

同步MySQL的数据至AnalyticDB MySQL 3.0时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源AnalyticDB MySQL 3.0。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

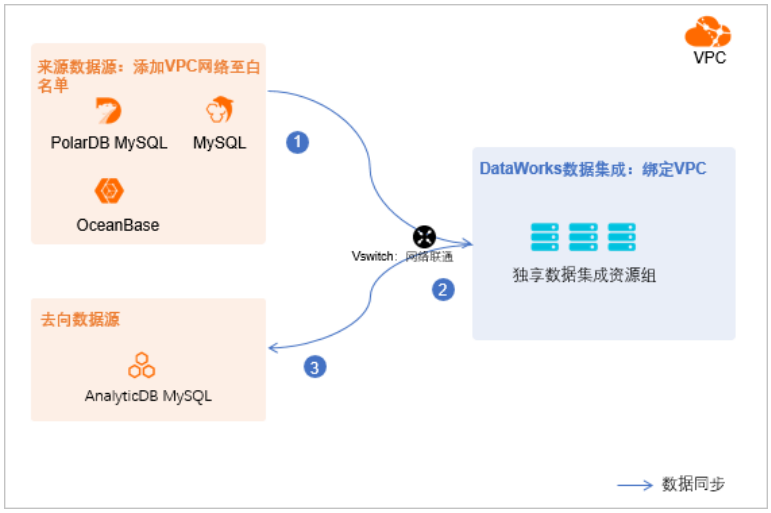
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.6.4. 配置数据源（来源为OceanBase）

同步OceanBase的数据至AnalyticDB MySQL 3.0时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

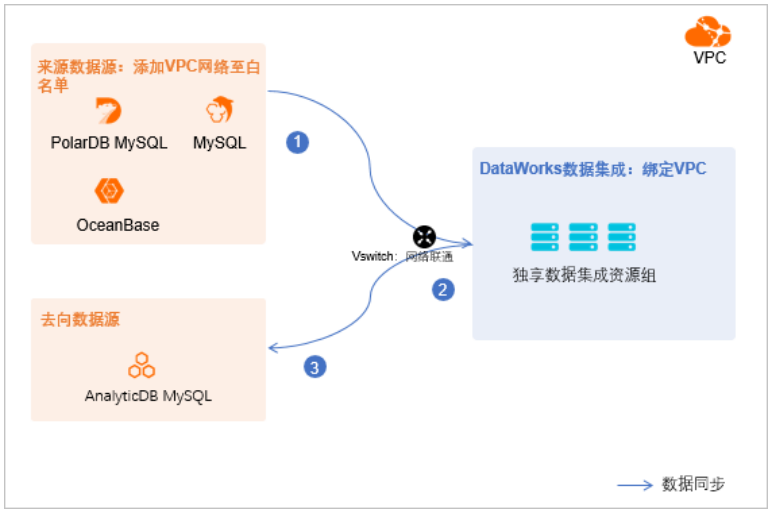
- 准备数据源：已购买来源数据源OceanBase、去向数据源AnalyticDB MySQL 3.0。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

使用限制

OceanBase为分布式关系型数据库，可以使物理分布不同的多个数据库上的数据，被整合为一个完整的逻辑数据库。但实时同步OceanBase的数据至AnalyticDB MySQL 3.0，目前仅支持同步单个物理库的数据，不支持同步逻辑库数据。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至OceanBase的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至OceanBase集群的白名单中，详情请参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有OceanBase的相关操作权限，详情请参见新建账号。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.6.5. 添加数据源

将来源数据源的数据同步至AnalyticDB MySQL 3.0数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：OceanBase

添加OceanBase数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置ApsaraDB for OceanBase数据源。

添加去向数据源：AnalyticDB MySQL 3.0

添加AnalyticDB MySQL 3.0数据源，详情请参见配置AnalyticDB for MySQL 3.0数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置查看数据同步任务。

5.6.6. 配置查看数据同步任务

完成数据源、网络、资源的准备配置后，您可创建并执行数据同步任务，开始进行数据同步。本文为您介绍如何创建数据同步任务，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

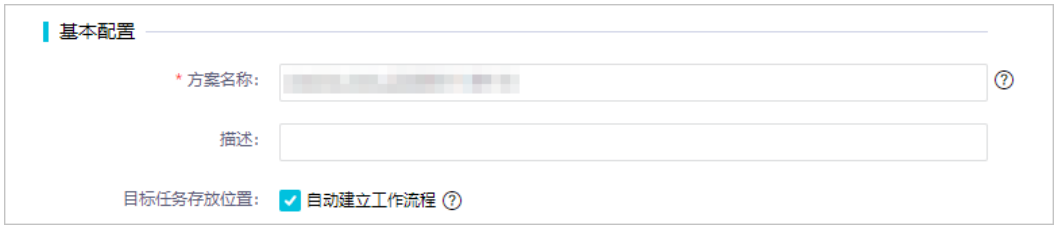
- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为MySQL）
- 配置数据源（来源为OceanBase）
- 添加数据源

创建同步解决方案任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 在**新建同步解决方案**对话框中，单击**一键实时同步至 AnalyticDB MySQL 3.0**。
4. 完成方案名称等基本信息配置。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。

参数	描述
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

5. 选择来源数据源并配置同步规则。
   i. 在**数据来源**区域，选择**类型**和**数据源**。

   > **说明** 仅支持选择MySQL、OceanBase和PolarDB类型的数据源。

   ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 图标，将其移动至**已选源表**。

   该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

   > **注意** 如果选中的表没有主键，将无法进行实时同步。

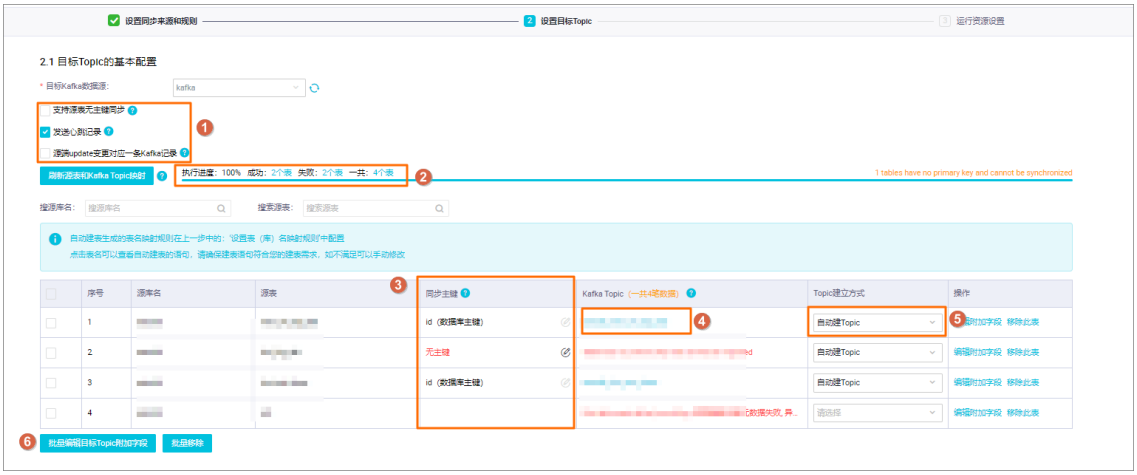
   iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

   同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

   - **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
   - **目标表名规则**：支持对转换后的表名添加前缀和后缀。

   iv. 单击**下一步**。
6. 选择目标数据源并配置目标表格式。
   i. 在**设置目标表**页面，选择**目标AnalyticDB for MySQL3.0数据源**。
   ii. 单击**刷新源表和AnalyticDB for MySQL3.0表映射**，创建需要同步的源表和目标AnalyticDB MySQL 3.0表的映射关系。

iii. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 如果来源库有主键，则同步数据时会直接使用该主键进行去重。 ■ 如果来源库没有主键，则您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
③	选择的**表建立方式**，取值如下： ■ 当**表建立方式**选择**使用已有表**时，**AnalyticDB for MySQL3.0表名**列显示自动创建的AnalyticDB MySQL 3.0表名称。您也可以在下拉列表中选择需要使用的表名称。选择**使用已有表**时，**表类型**及**分布字段列**不可更改。 ■ 当**表建立方式**选择**自动建表**时，显示自动创建的AnalyticDB MySQL 3.0表名称。您可以单击表名称，查看和修改建表语句。同时，您还可以为创建的表自定义**表类型**及**分布字段列**。

iv. 单击**下一步**。

7. DDL消息处理规则设置。

来源数据源（MySQL、PolarDB、OceanBase）会包含许多DDL操作，进行实时同步时，您可以根据业务需求，对不同的DDL消息设置同步至目标端的规则。

说明 该规则为初次执行实时同步任务时的DDL消息处理规则，后续如果您需要修改规则，可以手动停止任务，进入实时任务的配置页面修改，详情请参见管理数据同步解决方案任务。

i. 在**DDL消息处理规则**页签，配置**实时同步DDL消息处理策略**。

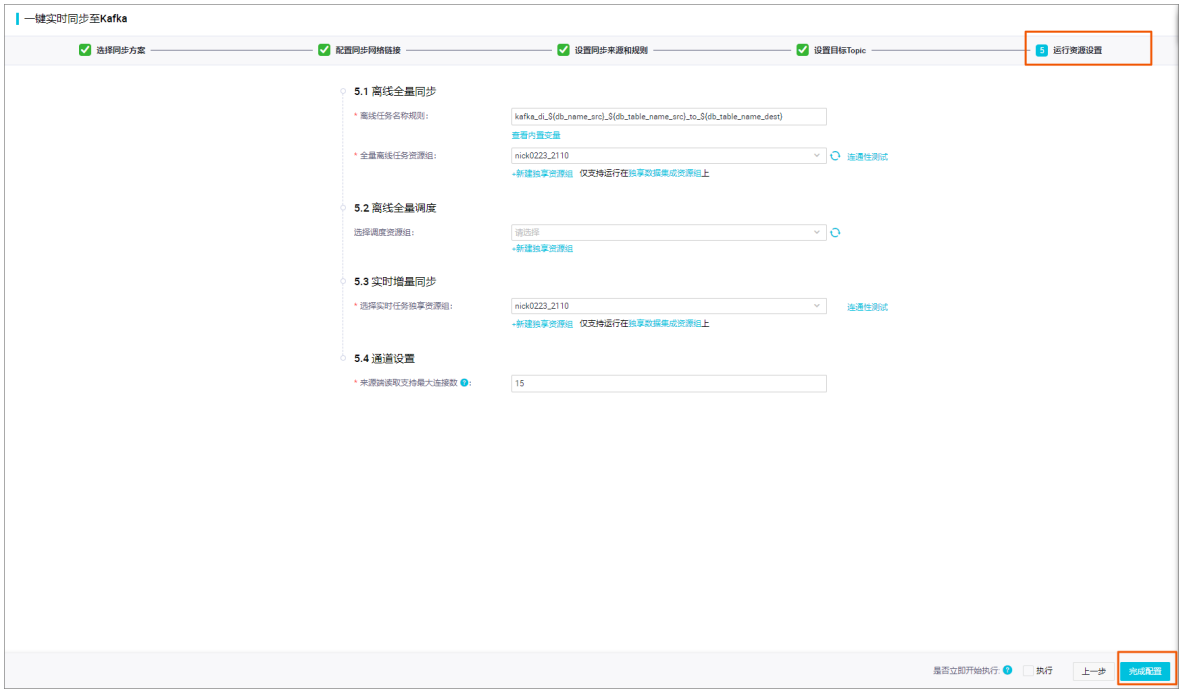
不同DDL消息处理策略如下表所示。

DDL消息类型	处理策略
新建表	DataWorks收到对应类型的DDL消息时，处理策略如下： ■ **正常处理**：将相应消息继续下发给目标数据源，由目标数据源来处理。因为不同目标数据源对DDL消息处理策略可能会不同，因此DataWorks只执行转发操作。 ■ **忽略**：直接丢弃该消息，不再向目标数据源发送。 ■ **告警**：直接丢弃该消息，同时会在实时同步日志中记录告警信息，指明该消息因执行出错被丢弃。 ■ **出错**：实时同步任务直接显示出错状态并终止运行。
删除表	
新增列	
删除列	
重命名表	
重命名列	
修改列类型	
清空表	

ii. 单击**下一步**。

8. 运行资源设置。

   在**运行资源设置**页面，配置各项参数。



- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为**20**。
目的端写入支持最大连接数	目的端写入支持最大连接不能小于来源端读取支持最大连接数。请根据数据库资源的实际情况合理配置。默认为**45**。

9. 单击**完成配置**，完成整库实时同步任务的创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

如果任务执行失败，您可查看任务运行失败的错误提示，参考以下常见问题进行排查处理。

- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.
- 离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。

- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

管理数据同步解决方案任务

- 查看或编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 查看配置**或**更多 > 修改配置**，可查看或编辑任务的配置信息。

  说明 仅单击**未运行**状态后的**任务配置**，您可以编辑任务。其它状态下的任务配置页面，仅支持查看。

- 删除任务。

单击相应任务后的**更多 > 删除**。在**删除**对话框中，单击**确定**。

> **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

- 修改任务优先级。

  单击相应任务后的**更多 > 修改优先级**。在**修改优先级**对话框中，输入需要配置的优先级数值，单击**确定**。优先级取值范围为1~8，数值越大优先级越高。

  > **说明** 优先级相同的任务，按照提交时间的先后顺序执行。

5.6.7. 常见问题

以下为您介绍同步数据至AnalyticDB MySQL 3.0解决方案操作失败的常见问题和解决方案。

PolarDB数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源PolarDB时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

OceanBase数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源OceanBase时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

MySQL数据源网络联通性测试失败怎么办？

- 错误现象：添加数据源MySQL时，网络连通性测试失败。
- 如何处理：切换到jdbc连接串，同时检查白名单配置，以及独享资源组的VPC配置。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX` 。
- 可能原因：来源数据源PolarDB没有开启binlog。
- 如何处理：PolarDB开启binlog，详细操作可参见配置数据源（来源为PolarDB）。并进行至少一条数据的变更，同时切换数据集成实时同步开始点位到当前时间。

实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation` 。
- 可能原因：来源数据源PolarDB没有给进行数据同步的账号开启所需权限，或对接的PolarDB数据库不是主节点。
- 如何处理：参见配置数据源（来源为PolarDB）的操作授予权限，或者检查PolarDB是否是主节点（读写库），目前实时任务不支持从PolarDB备节点抓取数据。

实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.

- 报错内容：数据同步任务执行时，实时任务运行失败，错误提示为 `com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first` 。
- 可能原因：来源数据源PolarDB未打开loose_polar_log_bin参数。
- 如何处理：需要打开loose_polar_log_bin参数，详细操作可参见配置数据源（来源为PolarDB）。

离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.

- 报错内容：数据同步任务执行时，离线任务运行失败，错误提示为 `com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns` 。
- 可能原因：DataWorks的离线引擎插件未升级到最新版本。
- 如何处理：请提交工单联系技术支持，帮您将离线引擎插件升级到最新版本。

5.7. 同步数据至MaxCompute

5.7.1. 准备工作

5.7.1.1. 资源规划与配置

当前使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。

4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

> **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
> - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
> - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

> **说明** 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

绑定专有网络

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。绑定专有网络的操作如下。

> **注意** 4c8g类型的独享数据集成资源组最多支持绑定2个专有网络，其他规格的独享数据集成资源组最多支持绑定3个专有网络。

1. 登录DataWorks控制台。
2. 在**资源组列表**的**独享资源组**页签下，单击相应资源组后的**网络设置**，进入专有网络绑定页面。

   绑定前，请首先使用阿里云主账号进行RAM授权（仅主账号有权限），让DataWorks拥有访问您的云资源的权限。您可以通过云资源访问授权页面进行授权。也可以在主账号首次次进入管控后弹出的界面弹框中进行授权。
3. 绑定专有网络VPC。

i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

参数说明如下：

参数	配置说明（数据源与独享资源同账号同地域）	配置说明（数据源与独享资源在不同账号或不同地域）
专有网络	如果您的数据源与独享资源组在同一个阿里云账号下，建议配置为数据源所在的VPC。 如果不在同一个阿里云账号下，则与不在同一地域场景一致。	如果您的数据源与独享资源不在同一地域，例如，数据源不在阿里云VPC网络环境中，您可单击**创建专有网络**，为独享资源组创建一个VPC。创建完成后这里配置为新建的VPC或选择已经与目标数据库网络打通的VPC。 **说明** 在创建专有网络的场景下，您还需通过VPN或高速通道等方式，将独享资源组绑定的VPC与数据源所在VPC网络打通，并手动添加路由指向目标数据库IP，保障两个网络间可达。
可用区	选择数据库所在可用区。	选择已经与目标数据库网络联通的可用区。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。 **说明** 绑定数据源所在VPC后，绑定VPC下任意一个交换机，会自动添加路由至整个VPC网段，实现独享数据集成资源组在该VPC下网络可达。	选择已经与目标数据库网络联通的交换机，若没有可用交换机，可单击**创建交换机**为独享资源组创建交换机。创建完成后这里配置为创建的交换机。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

ii. 单击**确定**，完成绑定VPC操作。

说明 如果数据源和独享资源组不在同一个地域，或不在同一个阿里云账号下，则您需要绑定专有网络后，再添加路由规则指向目标数据库IP地址。

4. （可选）配置Host。

如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

5. （可选）配置DNS。

   如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

   > **说明** 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

   i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

   ii. 如果您需要修改之前配置的DNS，您可单击左下角的**修改**。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。目前同步数据至MaxCompute的来源数据源仅支持PolarDB、Oracle及MySQL，您可以根据实际需求选择合适的数据源。详细的来源数据源配置请参见配置数据源（来源为PolarDB）、配置数据源（来源为Oracle）或配置数据源（来源为MySQL）。

5.7.1.2. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至MaxCompute时，来源数据源为PolarDB，去向数据源为MaxCompute，执行数据同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

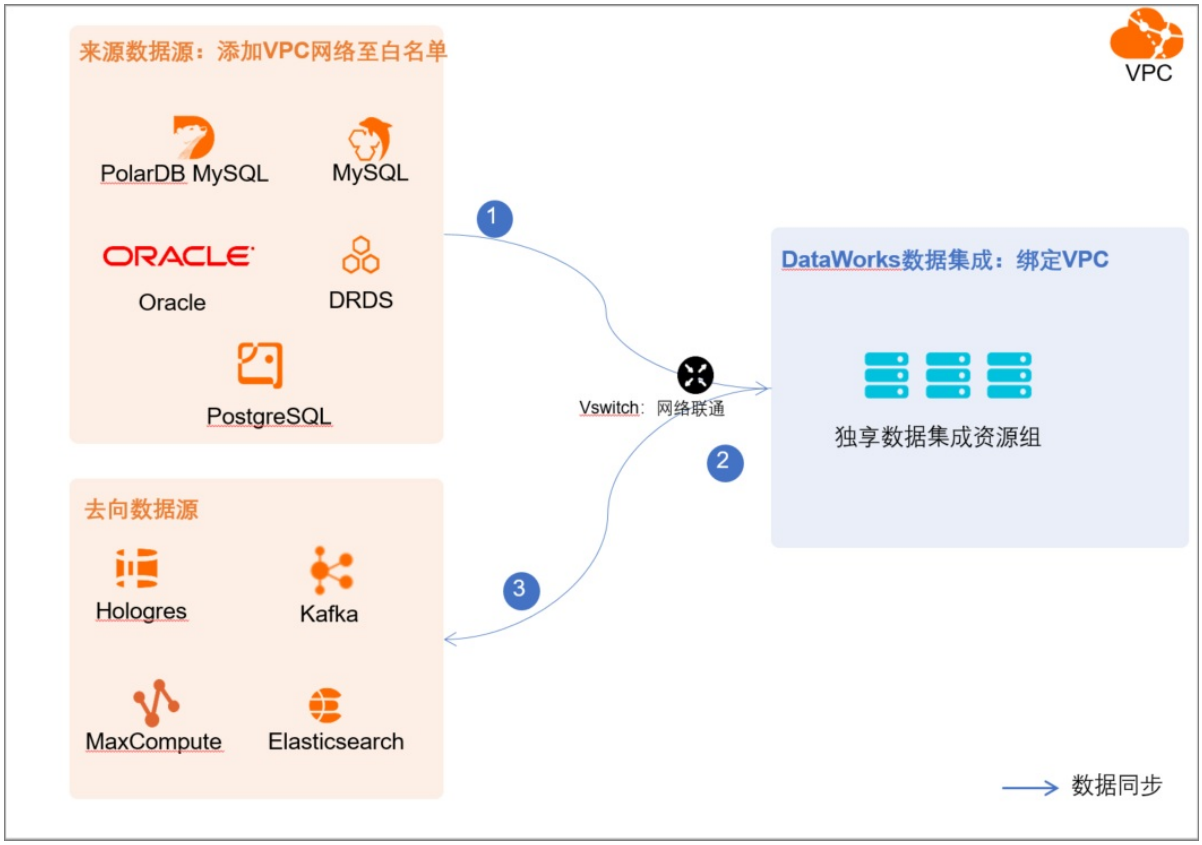
- 数据源准备：已购买来源数据源PolarDB MySQL和去向数据源MaxCompute。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

配置来源数据源：PolarDB

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

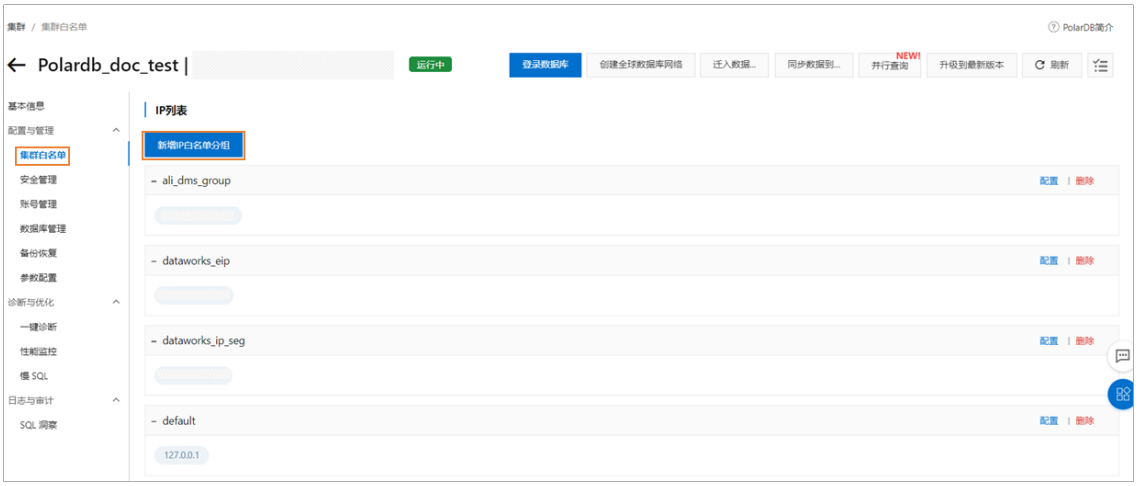
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner 账号登录客户端。

操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

使用Project Owner账号在客户端执行以下命令。

```
setproject odps.sql.acid.table.enable=true;
```

3. （可选）开启使用数据2.0。

如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

```
setproject odps.sql.type.system.odps2=true;
```

4. 创建账号。

此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.7.1.3. 配置数据源（来源为Oracle）

同步Oracle的数据至MaxCompute时，来源数据源为Oracle，去向数据源为MaxCompute，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

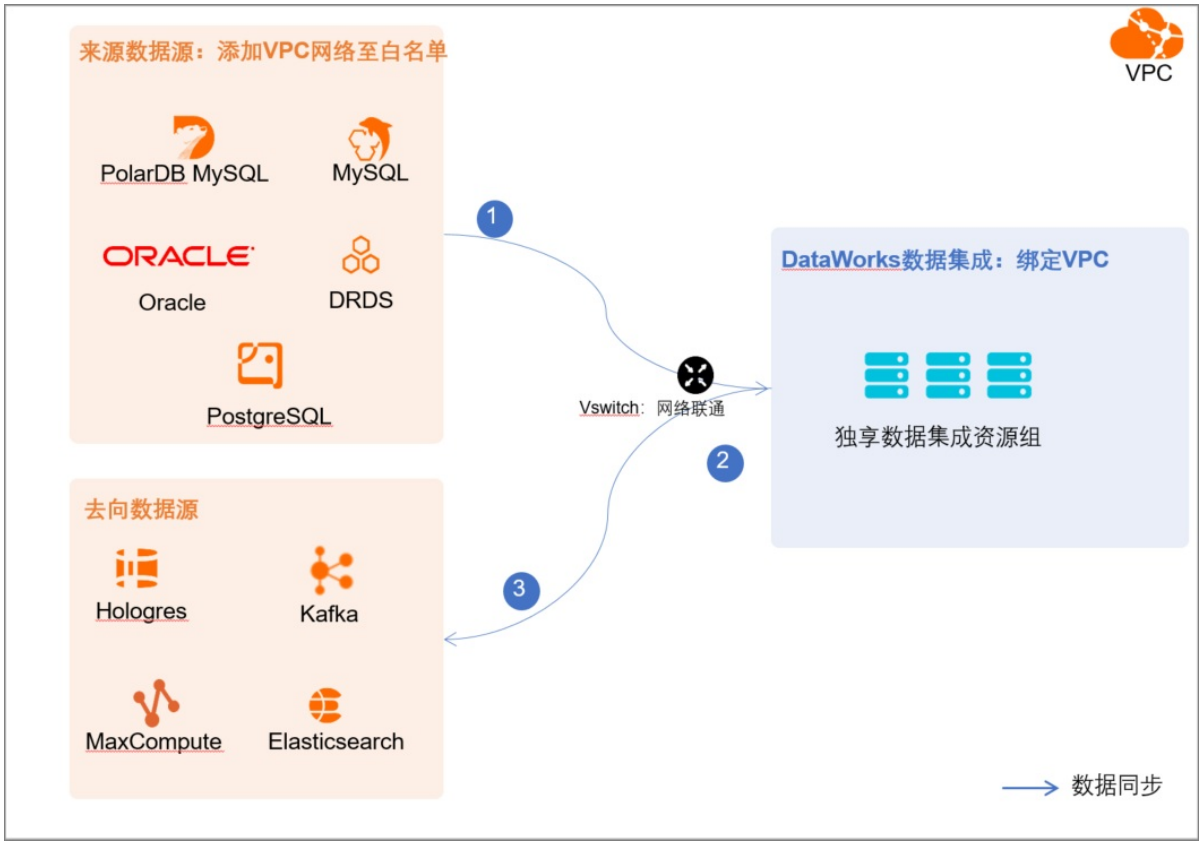
- 准备数据源：已购买来源数据源Oracle、去向数据源MaxCompute。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

  > **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。

- Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
- 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

  例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

  使用数据集成时推荐开启主键列或唯一索引列补充日志。

  - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
  - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

```
select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
```

- 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
- 当 `supplemental_log_data_pk` 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。
- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

配置来源数据源：Oracle

1. 配置白名单。

将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

i. 创建账号。

操作详情请参见创建Oracle账号。

ii. 配置权限。

您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换`'同步账号'`为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner账号登录客户端。

操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

使用Project Owner账号在客户端执行以下命令。

```
setproject odps.sql.acid.table.enable=true;
```

3. （可选）开启使用数据2.0。

如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

```
setproject odps.sql.type.system.odps2=true;
```

4. 创建账号。

此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.7.1.4. 配置数据源（来源为MySQL）

同步MySQL的数据至MaxCompute时，来源数据源为MySQL，去向数据源为MaxCompute，执行数据同步任务前，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源MaxCompute。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

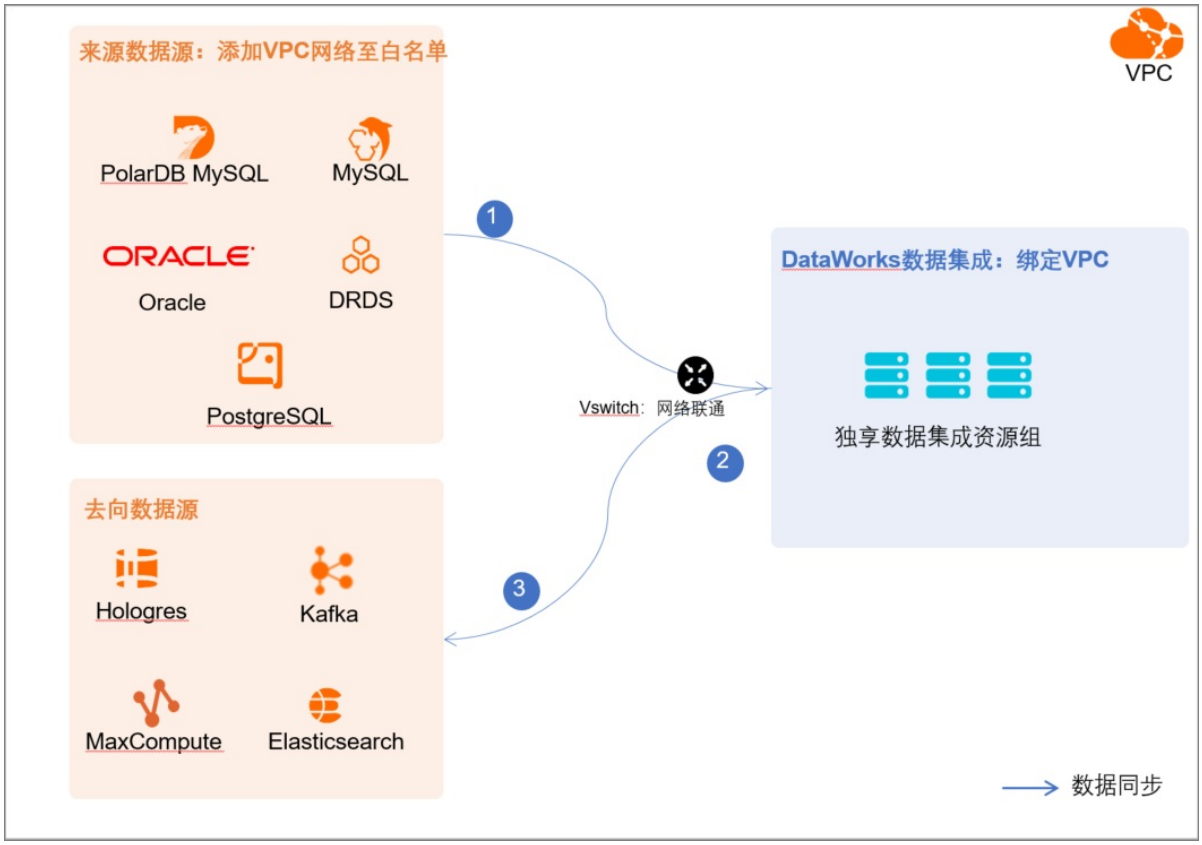
- 工具准备：进行MaxCompute数据源的项目属性配置时，需使用MaxCompute客户端，您需要提前下载客户端并熟悉客户端操作。操作详情可参见使用客户端（odpscmd）连接。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：

  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

- DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL Binlog实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。
- DDL的增加列不支持在表中间加列，会引起列错乱的数据质量问题，仅支持在表末尾加列的DDL操作。

配置来源数据源：MySQL

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

配置去向数据源：MaxCompute

1. 使用MaxCompute的Project Owner账号登录客户端。

   操作详情可参见使用客户端（odpscmd）连接。

2. 打开项目的acid属性。

   使用Project Owner账号在客户端执行以下命令。

   ```
   setproject odps.sql.acid.table.enable=true;
   ```

3. （可选）开启使用数据2.0。

   如果需要使用MaxCompute数据2.0类型中的timestamp类型，您需要使用Project Owner账号在客户端执行以下命令开启数据2.0。

   ```
   setproject odps.sql.type.system.odps2=true;
   ```

4. 创建账号。

   此账号在后续添加数据源时需配置使用，用于对接MaxCompute进行数据同步操作。操作详情可参见准备阿里云账号。

   创建完成后，您可记录下此账号的Accesskey ID和Accesskey Secret，便于后续配置使用。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.7.1.5. 添加数据源

将来源数据源的数据同步至MaxCompute数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加来源数据源：Oracle

添加Oracle数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置Oracle数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加去向数据源：MaxCompute

操作详情可参见配置MaxCompute数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见一键实时同步至MaxCompute。

5.7.2. 一键实时同步至MaxCompute

完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建一键实时数据同步任务，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

使用限制

仅支持与当前工作空间同地域的自建MaxCompute数据源，跨地域的MaxCompute项目在测试数据源服务连通性时可以正常连通，但同步任务执行时，在MaxCompute建表阶段会报引擎不存在的错误。

> **说明** 使用自建MaxCompute数据源时，DataWorks项目仍然需要绑定MaxCompute引擎，否则将无法创建MaxCompute SQL节点，导致全量同步标done节点创建失败。

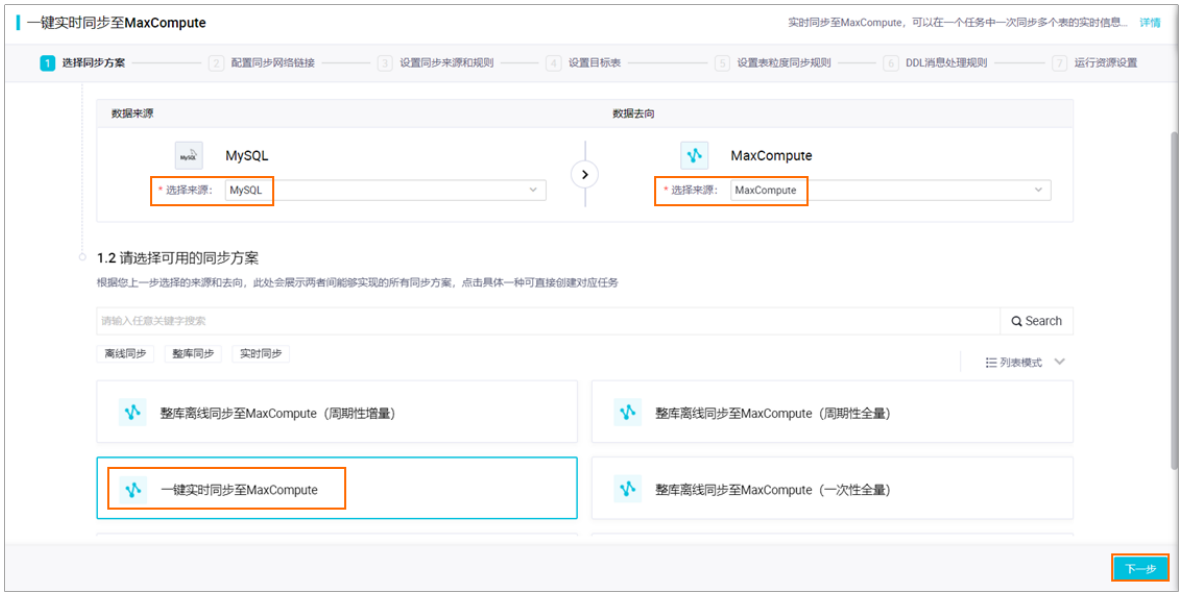
计费说明

一键实时同步至MaxCompute解决方案需要定期做全增量数据周期合并，因此会消耗MaxCompute计算资源。这部分费用由MaxCompute直接收取，费用与用户同步全量数据大小、合并周期正相关。具体费用请参考计费项与计费方式概述。

创建同步解决方案任务

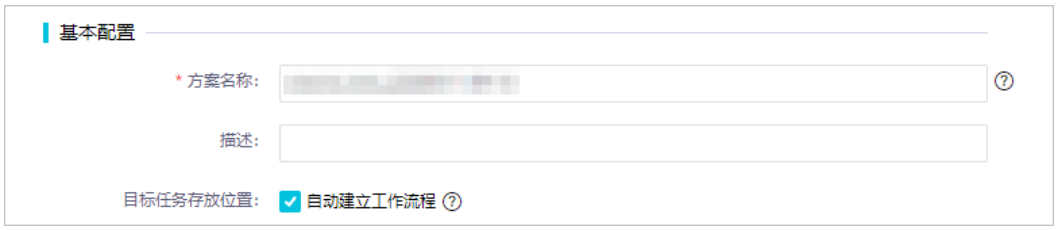
1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**一键实时同步至MaxCompute方案**）。

   操作详情请参见选择同步方案。



2. 完成方案名称等基本信息配置。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

3. 选择来源数据源并配置同步规则。

   i. 在**数据来源**区域，选择**类型**和**数据源**。

   > **说明** 仅支持选择MySQL、Oracle和PolarDB类型的数据源。

ii. 在**选择同步的源表**区域，选中需要同步的**源表**，单击 图标，将其移动至**已选源表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

> **注意** 如果选中的表没有主键，将无法进行实时同步。

iii. 在**设置表名的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

iv. 单击**下一步**。

4. 选择目标数据源并配置目标表格式。

i. 在**设置目标表**页面，选择**目标MaxCompute（ODPS）数据源**和**写入模式**。

ii. 单击**MaxCompute（ODPS）时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置（支持天级别分区）。此处可以选择写入MaxCompute分区表或者是非分区表。

iii. （可选）**特殊表进行离线同步**：是否将无主键表设置为全量离线数据同步任务。

iv. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

v. 查看任务的执行进度和表来源。

序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。 **说明** 如果源表为无主键表，您可以无主键字样后的编辑入口，为源表手动指定主键，以便后续进行增全量同步。
④	支持您通过**编辑附加字段**为目标MaxCompute表在源表字段基础上增加写入的目标MaxCompute表字段。 **说明** 如果您初次使用**自动建表**方式时选择了附加字段，系统在自动建表时会加上对应列。如果您在已有表上增加附加字段，需要保证已有表中已经存在对应的列名，系统才会往对应列里写值，否则不会修改已有表的表结构去追加列。

vi. 单击**下一步**。

5. 运行资源设置。

   在**运行资源设置**页面，配置各项参数。

参数	描述
同步引擎	支持**默认内嵌引擎**。
选择实时任务独享资源组	从下拉列表中选择实时任务的独享资源组。 **说明** 实时同步任务仅支持运行在独享数据集成资源组上，详情请参见新增和使用独享数据集成资源组。
实时同步任务名称	实时同步任务的名称。
选择调度资源组	分别选择任务调度和全量离线任务需要使用的独享资源组。目前解决方案仅支持使用独享数据集成资源组，详情请参见新增和使用独享数据集成资源组。
选择全量离线任务独享资源组	
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于读取全量数据，再生成实时任务持续读取实时增量数据。

6. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

管理数据同步解决方案任务

- 查看或编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**任务配置**，查看或编辑任务。

  说明 仅单击**未运行**状态后的**任务配置**，您可以编辑任务。其它状态下的任务配置页面，仅支持查看。

- 删除任务。

  单击相应任务后的**删除**。在**删除**对话框中，单击**确定**。

  说明 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

5.7.3. 整库离线同步（周期性全量）

完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建周期全量数据同步任务，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

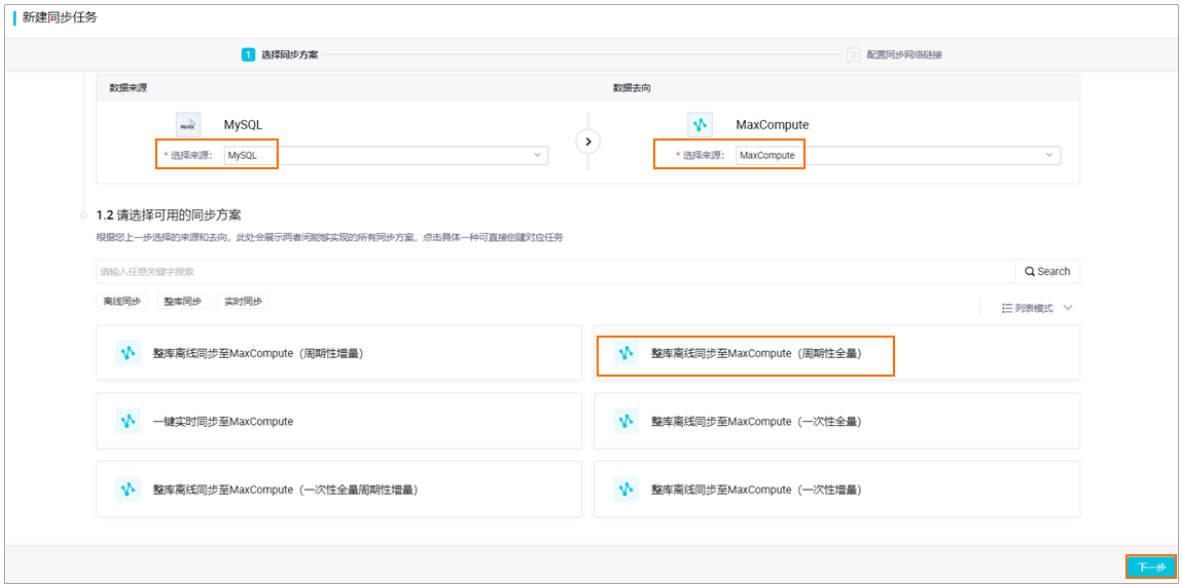
应用场景

整库离线同步（周期性全量）适用于需要将某些表的全量数据周期性的同步到MaxCompute的场景，适用于大量表的周期性同步任务，本数据同步解决方案中，您可以选用分批上传的功能来降低数据同步负载，并支持丰富的调度时间配置，便于数据同步任务周期运行。

创建同步解决方案任务

1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**整库离线同步至MaxCompute（周期性全量）**）。

   操作详情请参见选择同步方案。



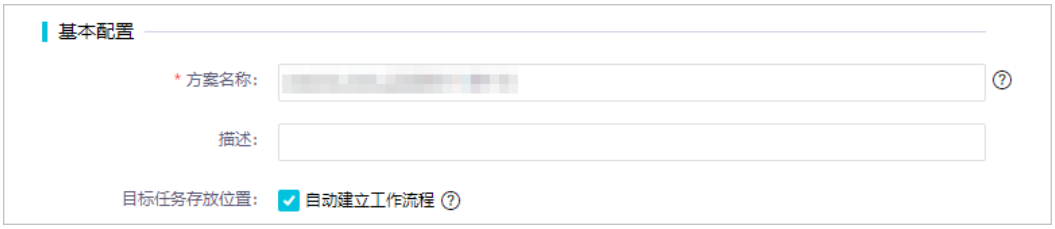
2. 配置同步网络链接。

   根据界面提示选择数据来源、独享数据集成资源组、数据去向，并测试网络连通性，完成后单击**下一步**。请务必参考资源规划与配置提前规划好所用的资源组与网络连通方案，并将数据源添加至DataWorks中，完成白名单等网络连通所需配置，避免网络连通测试失败。

3. 设置同步来源和规则。

i. 配置基本信息。

在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

ii. 确认数据来源信息。

页面展示上述步骤选择的数据来源结果并为您默认选择了数据来源的编码信息，您需确认数据来源是否正确、是否需要修改编码类型。

iii. 选择同步的源表。

根据界面提示，选择待同步的来源数据源中的源表。后续选中的表即会通过同步方案的配置从来源数据源同步至MaxCompute。

> 注意 如果选中的表没有主键，将无法进行实时同步。

iv. 设置表名的映射规则。

单击**添加规则**，选择相应的规则进行添加。同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

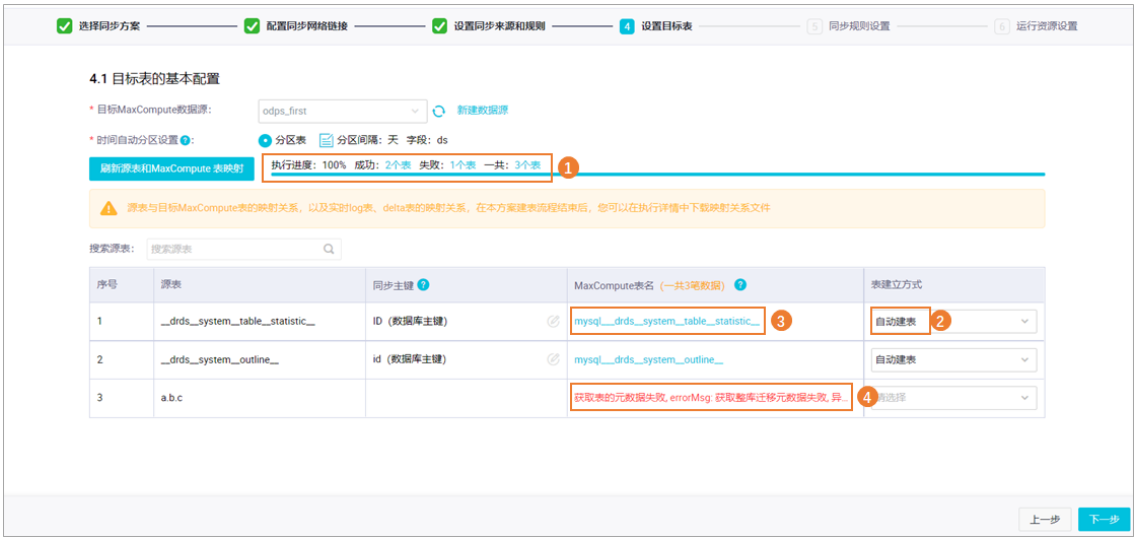
4. 设置目标表。

i. 页面展示上述步骤选择的去向数据源，确认去向数据源正确。

ii. 单击**时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置。

iii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。
④	暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。

v. 单击**下一步**。

5. 设置同步规则。

i. 配置全量同步的数据规则。

参数	描述
写入前清空对应的原有表	根据实际情况选择是否要打开此开关。打开后，后续进行数据同步时，每次向MaxCompute写入数据前，均会删除MaxCompute表内原有的数据，建议您谨慎打开此开关。
同步并发配置	您可以选择后续进行数据同步时，同步并发任务采取分批并发同步还是整批并发同步。如果待同步的数据表过多，建议选择**分批上传**，避免数据库负载过高，影响同步任务运行。
设置分批周期 （通过单位时间同步的表数量控制分批周期）	当上述**同步并发配置**设置为**分批上传**时，您需要根据分批同步的规划，设置分批周期，时间间隔从15分钟到若干小时不等。 例如，您规划本同步任务后续定时每天凌晨5：00开始运行，待同步的数据表有300个。从开始运行到当天结束共有19h可以运行任务，为了避免负载过高，可以将这300个表同步任务分为6批运行，从凌晨6点开始，每3h运行一批，每批同步50个表。 **说明** 进行分批上传时，您需要规划好整个任务的调度周期与这里分批周期，分批周期加和要小于任务可运行的时间。以上面的例子为例，所有的分批周期之和（每批3h，共6批，加和共18h）小于任务可运行时间19h。

ii. 配置数据同步的调度周期。

根据任务运行需求设置数据同步任务的调度周期，包括**调度周期**、**定时调度时间**、**生效日期**等。调度参数的配置与一般节点任务的调度参数类似，参数详情可参见时间属性配置说明。

i. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页面，确认同步任务的任务名、所需的同步资源和调度资源，并配置**来源端读取支持最大连接数**。

说明 **来源端读取支持最大连接数**需结合来源端的数据库能力，设置源库允许支持的最大JDBC连接数。如果设置的连接数过大，有可能在读取源端数据库的数据时，造成源端数据库因连接数过大而无法正常连接读数。

7. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

5.7.4. 整库离线同步（周期性增量）

完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建周期增量数据同步任务，并在创建完成后查看任务运行情况。

前提条件

创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

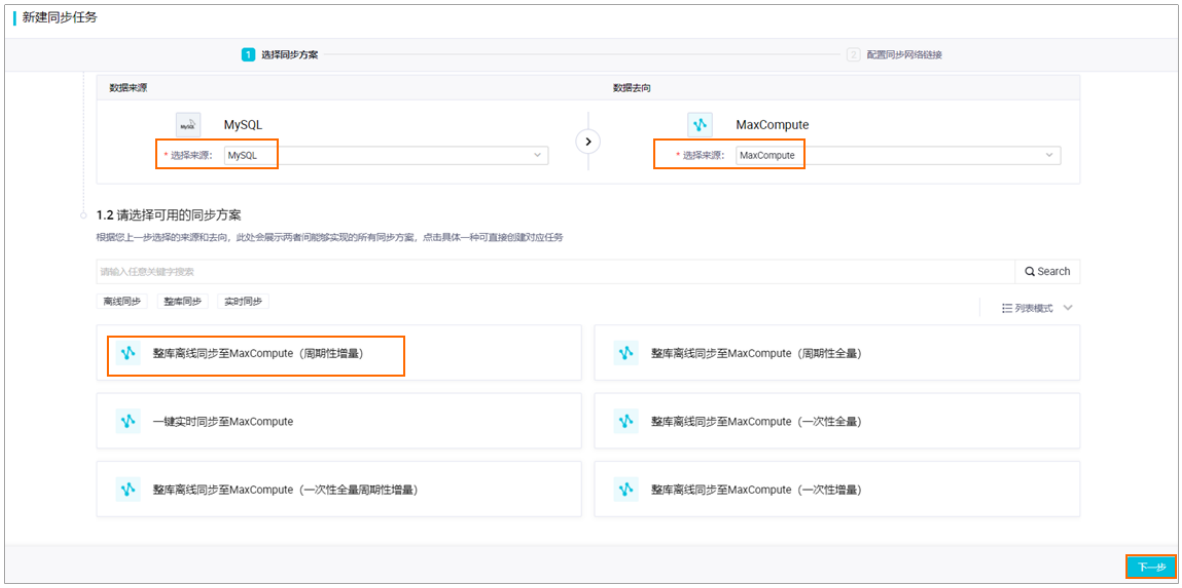
应用场景

整库离线同步（周期性全量）适用于需要将某些表的增量数据周期性的同步到MaxCompute的场景。本数据同步解决方案中，您可以通过where条件来抽取增量数据同步至MaxCompute，并支持丰富的调度时间配置，便于数据同步任务周期运行。

创建同步解决方案任务

1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**整库离线同步至MaxCompute（周期性增量）**）。

   操作详情请参见选择同步方案。



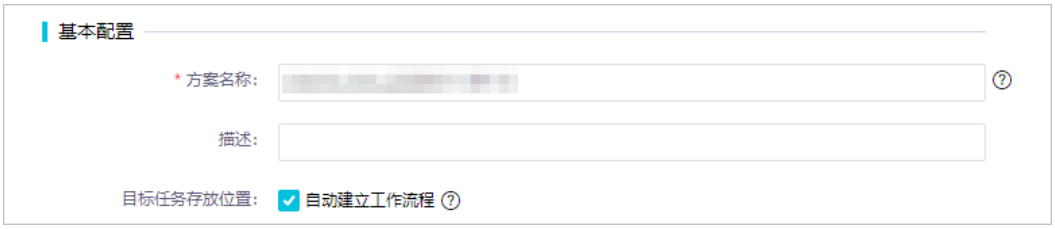
2. 配置同步网络链接。

   根据界面提示选择数据来源、独享数据集成资源组、数据去向，并测试网络连通性，完成后单击**下一步**。请务必参考资源规划与配置提前规划好所用的资源组与网络连通方案，并将数据源添加至DataWorks中，完成白名单等网络连通所需配置，避免网络连通测试失败。

3. 设置同步来源和规则。

i. 配置基本信息。

在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

ii. 确认数据来源信息。

页面展示上述步骤选择的数据来源结果并为您默认选择了数据来源的编码信息，您需确认数据来源是否正确、是否需要修改编码类型。

iii. 选择同步的源表。

根据界面提示，选择待同步的来源数据源中的源表。后续选中的表即会通过同步方案的配置从来源数据源同步至MaxCompute。

> 注意 如果选中的表没有主键，将无法进行实时同步。

iv. 设置表名的映射规则。

单击**添加规则**，选择相应的规则进行添加。同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

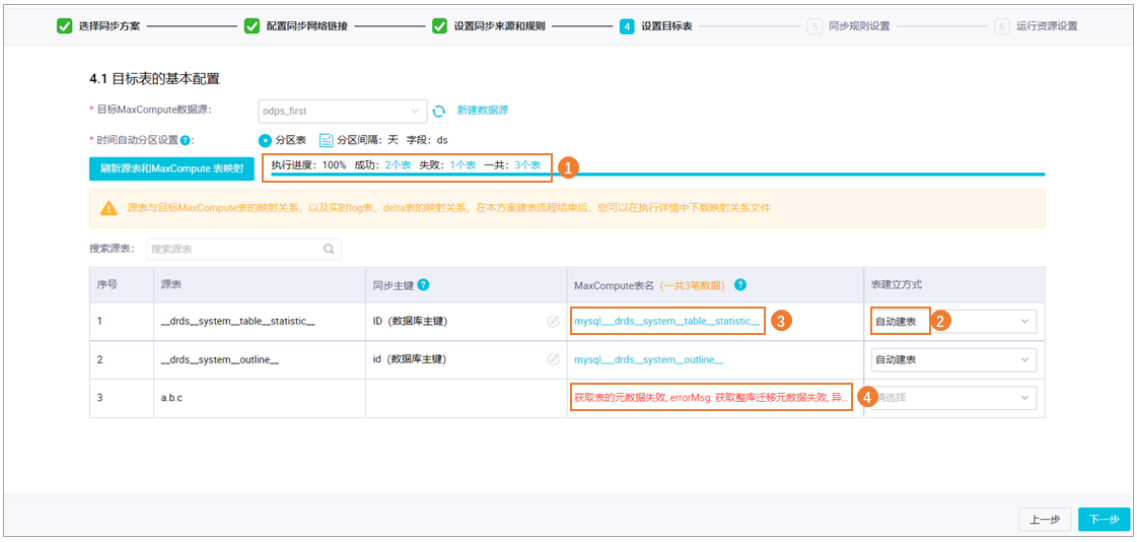
4. 设置目标表。

i. 页面展示上述步骤选择的去向数据源，确认去向数据源正确。

ii. 单击**时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置。

iii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。
④	暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。

v. 单击**下一步**。

5. 设置同步规则。

i. 配置增量同步的数据规则。

您可通过where语句对待同步的数据表进行过滤，且只需在增量条件框中填写where子句，无需写where关键字。同时，在写where子句时，您可以使用系统内置变量，例如使用 `${bizdate}` 指代业务日期、使用 `${cyctime}` 指代定时时间等，系统内置变量的类型与使用方法请参见调度参数概述。

ii. 配置数据同步的调度周期。

根据任务运行需求设置数据同步任务的调度周期，包括**调度周期**、**定时调度时间**、**生效日期**等。调度参数的配置与一般节点任务的调度参数类似，参数详情可参见时间属性配置说明。

iii. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页面，确认同步任务的任务名、所需的同步资源和调度资源，并配置**来源端读取支持最大连接数**。

> **说明** **来源端读取支持最大连接数**需结合来源端的数据库能力，设置源库允许支持的最大JDBC连接数。如果设置的连接数过大，有可能在读取源端数据库的数据时，造成源端数据库因连接数过大而无法正常连接读数。

7. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

5.7.5. 整库离线同步（一次性全量）

整库离线同步（一次性全量）适用于需要将某些表的全量数据一次性的同步到MaxCompute的场景。完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建一次性全量数据同步任务，并在创建完成后查看任务运行情况。

前提条件

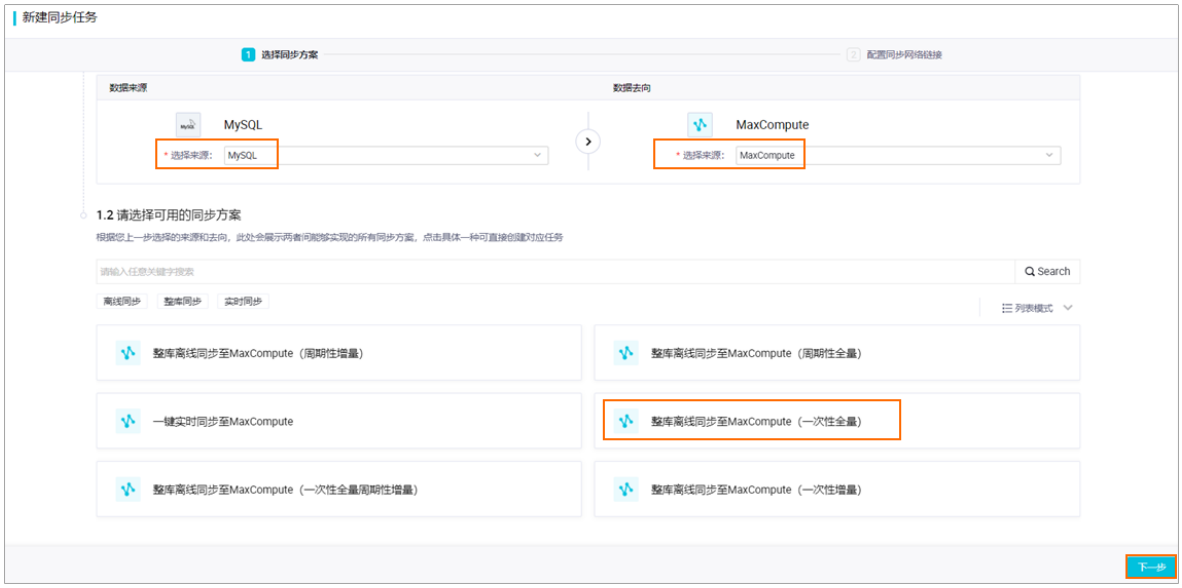
创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

创建同步解决方案任务

1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**整库离线同步至MaxCompute（一次性全量）**）。

操作详情请参见选择同步方案。



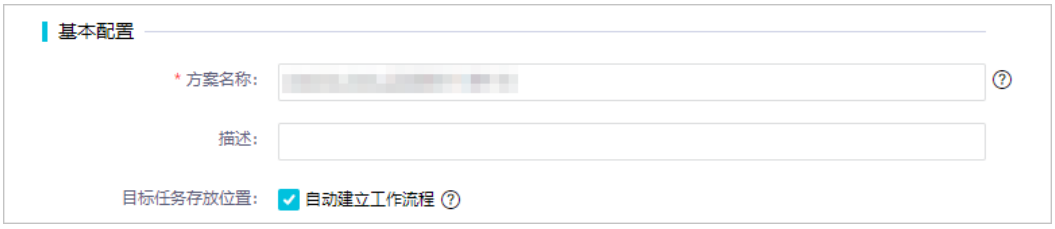
2. 配置同步网络链接。

   根据界面提示选择数据来源、独享数据集成资源组、数据去向，并测试网络连通性，完成后单击**下一步**。请务必参考资源规划与配置提前规划好所用的资源组与网络连通方案，并将数据源添加至DataWorks中，完成白名单等网络连通所需配置，避免网络连通测试失败。

3. 设置同步来源和规则。

   i. 配置基本信息。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

   ii. 确认数据来源信息。

   页面展示上述步骤选择的数据来源结果并为您默认选择了数据来源的编码信息，您需确认数据来源是否正确、是否需要修改编码类型。

   iii. 选择同步的源表。

   根据界面提示，选择待同步的来源数据源中的源表。后续选中的表即会通过同步方案的配置从来源数据源同步至MaxCompute。

   > **注意** 如果选中的表没有主键，将无法进行实时同步。

iv. 设置表名的映射规则。

单击**添加规则**，选择相应的规则进行添加。同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

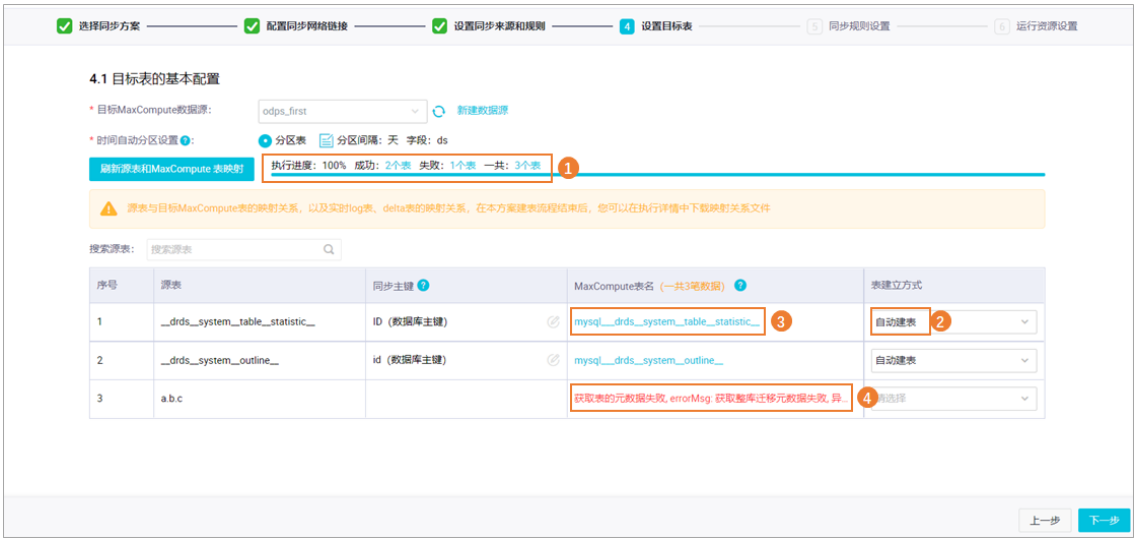
4. 设置目标表。

i. 页面展示上述步骤选择的去向数据源，确认去向数据源正确。

ii. 单击**时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置。

iii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。
④	暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。

v. 单击**下一步**。

5. 设置同步规则。

i. 配置增量同步的数据规则。

参数	描述
写入前清空对应的原有表	根据实际情况选择是否要打开此开关。打开后，后续进行数据同步时，每次向MaxCompute写入数据前，均会删除MaxCompute表内原有的数据，建议您谨慎打开此开关。

ii. 单击**下一步**。

6. 运行资源设置。

   在**运行资源设置**页面，确认同步任务的任务名、所需的同步资源和调度资源，并配置**来源端读取支持最大连接数**。

   > **说明** **来源端读取支持最大连接数**需结合来源端的数据库能力，设置源库允许支持的最大JDBC连接数。如果设置的连接数过大，有可能在读取源端数据库的数据时，造成源端数据库因连接数过大而无法正常连接读数。

7. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

5.7.6. 整库离线同步（一次性增量）

整库离线同步（一次性增量）适用于需要将某些表的增量数据一次性同步到MaxCompute的场景，且同步过程中可以对数据按where条件抽取。完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建一次性增量数据同步任务，并在创建完成后查看任务运行情况。

前提条件

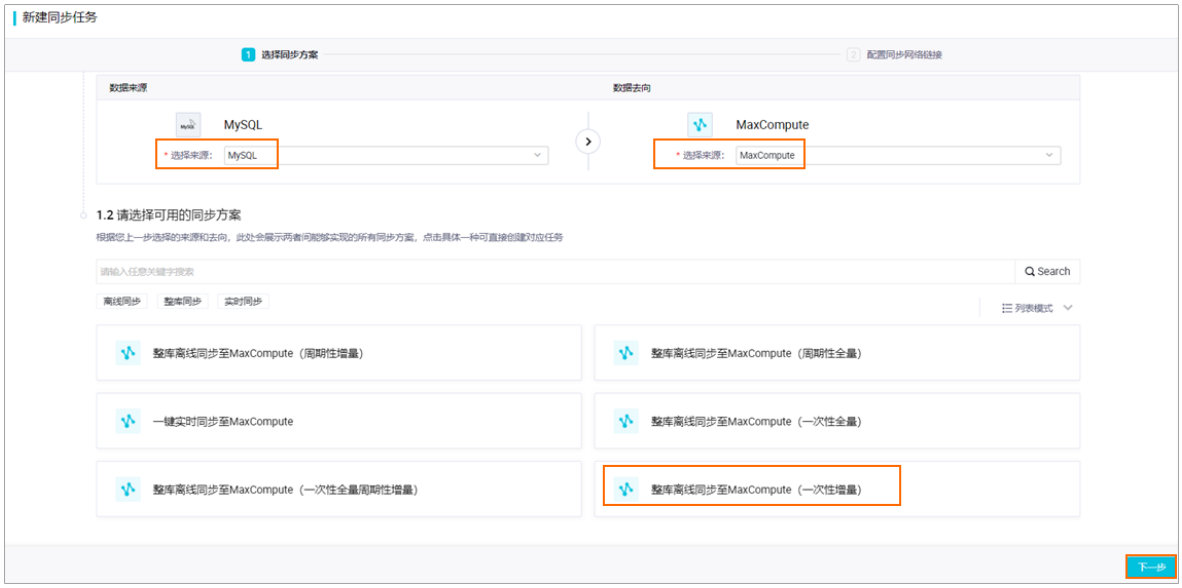
创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

创建同步解决方案任务

1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**整库离线同步至MaxCompute（一次性增量）**）。

   操作详情请参见选择同步方案。



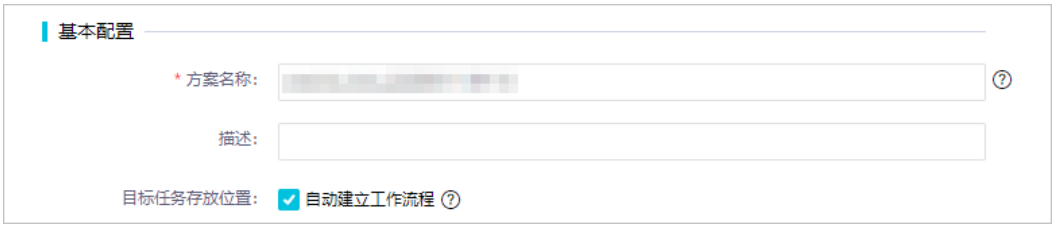
2. 配置同步网络链接。

   根据界面提示选择数据来源、独享数据集成资源组、数据去向，并测试网络连通性，完成后单击**下一步**。请务必参考资源规划与配置提前规划好所用的资源组与网络连通方案，并将数据源添加至DataWorks中，完成白名单等网络连通所需配置，避免网络连通测试失败。

3. 设置同步来源和规则。

   i. 配置基本信息。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

   ii. 确认数据来源信息。

   页面展示上述步骤选择的数据来源结果并为您默认选择了数据来源的编码信息，您需确认数据来源是否正确、是否需要修改编码类型。

   iii. 选择同步的源表。

   根据界面提示，选择待同步的来源数据源中的源表。后续选中的表即会通过同步方案的配置从来源数据源同步至MaxCompute。

   > 注意 如果选中的表没有主键，将无法进行实时同步。

iv. 设置表名的映射规则。

单击**添加规则**，选择相应的规则进行添加。同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

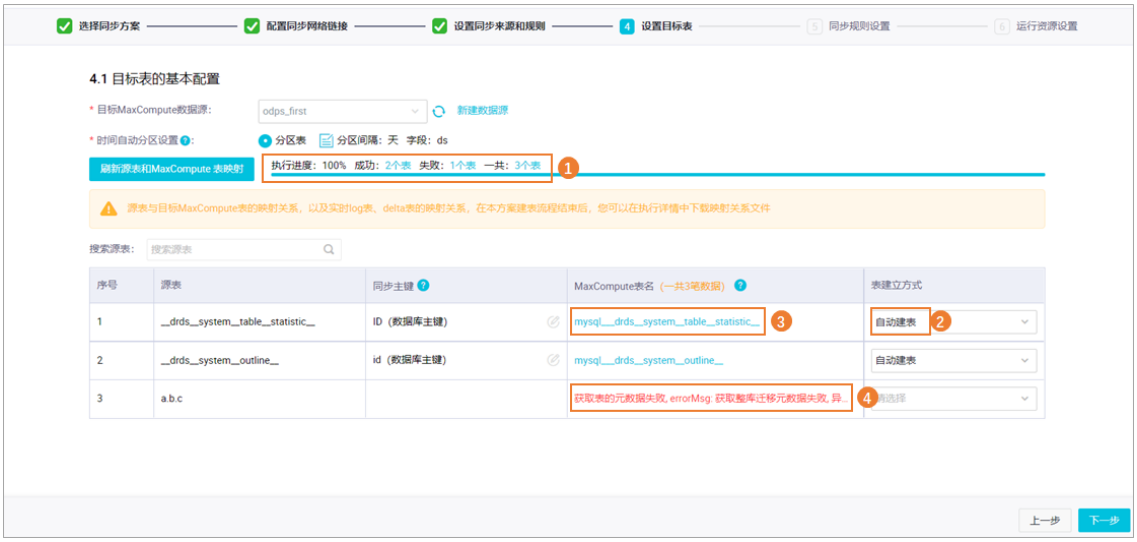
4. 设置目标表。

i. 页面展示上述步骤选择的去向数据源，确认去向数据源正确。

ii. 单击**时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置。

iii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。
④	暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。

v. 单击**下一步**。

5. 设置同步规则。

i. 配置增量同步的数据规则。

您可通过where语句对待同步的数据表进行过滤，且只需在增量条件框中填写where子句，无需写where关键字。同时，在写where子句时，您可以使用系统内置变量，例如使用 `${bizdate}` 指代业务日期、使用 `${cyctime}` 指代定时时间等，系统内置变量的类型与使用方法请参见调度参数概述。

ii. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页面，确认同步任务的任务名、所需的同步资源和调度资源，并配置**来源端读取支持最大连接数**。

> **说明** **来源端读取支持最大连接数**需结合来源端的数据库能力，设置源库允许支持的最大JDBC连接数。如果设置的连接数过大，有可能在读取源端数据库的数据时，造成源端数据库因连接数过大而无法正常连接读数。

7. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。
- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

5.7.7. 整库离线同步（一次性全量周期性增量）

整库离线同步（一次性全量周期性增量）适用于需要将某些表数据进行全量同步后周期性增量同步到MaxCompute的场景。完成数据源、网络、资源的准备工作配置后，您可以创建并执行同步任务。本文为您介绍如何创建一次性全量周期性增量数据同步任务，并在创建完成后查看任务运行情况。

前提条件

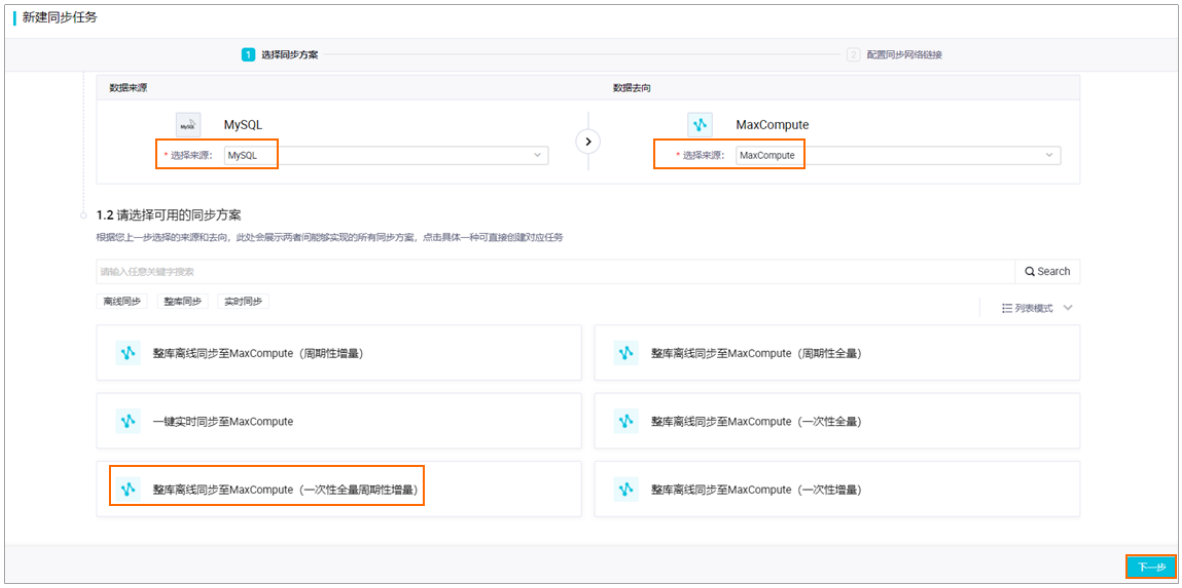
创建数据同步任务前，需检查已完成以下准备操作。

- 资源规划与配置
- 配置数据源（来源为PolarDB）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为MySQL）
- 添加数据源

创建同步解决方案任务

1. 进入同步解决方案页面后，根据待同步的来源数据源和去向数据源（本场景去向数据源为MaxCompute），选择同步解决方案（本场景为**整库离线同步至MaxCompute（一次性全量周期性增量）**）。

操作详情请参见选择同步方案。



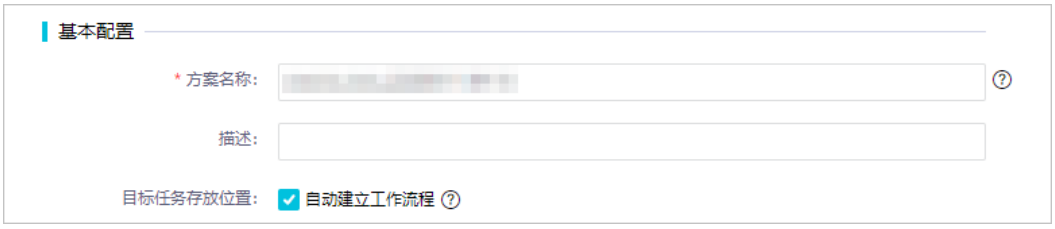
2. 配置同步网络链接。

   根据界面提示选择数据来源、独享数据集成资源组、数据去向，并测试网络连通性，完成后单击**下一步**。请务必参考资源规划与配置提前规划好所用的资源组与网络连通方案，并将数据源添加至DataWorks中，完成白名单等网络连通所需配置，避免网络连通测试失败。

3. 设置同步来源和规则。

   i. 配置基本信息。

   在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

   ii. 确认数据来源信息。

   页面展示上述步骤选择的数据来源结果并为您默认选择了数据来源的编码信息，您需确认数据来源是否正确、是否需要修改编码类型。

   iii. 选择同步的源表。

   根据界面提示，选择待同步的来源数据源中的源表。后续选中的表即会通过同步方案的配置从来源数据源同步至MaxCompute。

   > **注意** 如果选中的表没有主键，将无法进行实时同步。

iv. 设置表名的映射规则。

单击**添加规则**，选择相应的规则进行添加。同步规则包括**源表名和目标表名转换规则**和**目标表名规则**：

- **源表名和目标表名转换规则**：转换表名为目标表名，进行字符串替换。
- **目标表名规则**：支持对转换后的表名添加前缀和后缀。

v. 单击**下一步**。

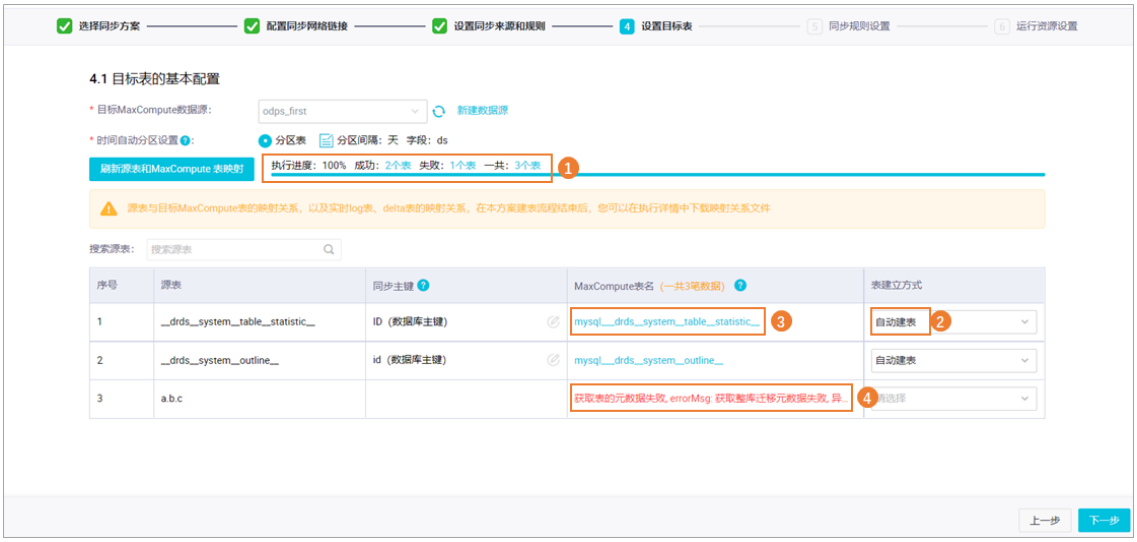
4. 设置目标表。

i. 页面展示上述步骤选择的去向数据源，确认去向数据源正确。

ii. 单击**时间自动分区设置**后的图标，在**编辑**对话框中，修改目标MaxCompute分区的设置。

iii. 单击**刷新源表和MaxCompute（ODPS）表映射**，创建需要同步的源表和目标MaxCompute表的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	**表建立方式**包括**自动建表**和**使用已有表**。
③	选择的**表建立方式**不同，此处显示的MaxCompute表名也不同： ■ 当选择**表建立方式**为**自动建表**时，显示自动创建的MaxCompute表名称。您可以单击表名称，查看和修改建表语句。 ■ 当选择**表建立方式**为**使用已有表**时，请在下拉列表中选择需要的表。
④	暂不支持同步没有主键的表。但只要选择的表中包括有主键的表，会正常执行流程，没有主键的表会被忽略。

v. 单击**下一步**。

5. 设置同步规则。

i. 配置增量同步的数据规则。

参数	描述
写入前清空对应的原有表	根据实际情况选择是否要打开此开关。打开后，后续进行数据同步时，每次向MaxCompute写入数据前，均会删除MaxCompute表内原有的数据，建议您谨慎打开此开关。

ii. 配置增量同步的数据规则。

您可通过where语句对待同步的数据表进行过滤，且只需在增量条件框中填写where子句，无需写where关键字。同时，在写where子句时，您可以使用系统内置变量，例如使用 `${bizdate}` 指代业务日期、使用 `${cyctime}` 指代定时时间等，系统内置变量的类型与使用方法请参见调度参数概述。

iii. 配置增量同步的数据规则。

您可通过where语句对待同步的数据表进行过滤，且只需在增量条件框中填写where子句，无需写where关键字。同时，在写where子句时，您可以使用系统内置变量，例如使用 `${bizdate}` 指代业务日期、使用 `${cyctime}` 指代定时时间等，系统内置变量的类型与使用方法请参见调度参数概述。

iv. 配置数据同步的调度周期。

根据任务运行需求设置数据同步任务的调度周期，包括**调度周期**、**定时调度时间**、**生效日期**等。调度参数的配置与一般节点任务的调度参数类似，参数详情可参见时间属性配置说明。

v. 单击**下一步**。

6. 运行资源设置。

在**运行资源设置**页面，确认同步任务的任务名、所需的同步资源和调度资源，并配置**来源端读取支持最大连接数**。

> **说明** **来源端读取支持最大连接数**需结合来源端的数据库能力，设置源库允许支持的最大JDBC连接数。如果设置的连接数过大，有可能在读取源端数据库的数据时，造成源端数据库因连接数过大而无法正常连接读数。

7. 单击**完成配置**，完成数据同步解决方案任务创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。

- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

5.8. 同步数据至Kafka

5.8.1. 资源规划与配置

当使用DataWorks的同步解决方案时，数据集成任务仅支持使用独享数据集成资源组，调度资源可根据业务需求选用公共资源或独享调度资源组。本文为您介绍使用同步解决方案时，需要使用的资源及相关配置。

背景信息

- 资源准备与规划：

  使用同步解决方案进行数据同步时，数据集成操作运行在数据集成资源组实例和调度资源组实例上。其中数据集成资源组当前仅能使用独享数据集成资源组，因此，在进行数据同步前，您需要购买独享数据集成资源组，并将资源组添加至DataWorks中，便于后续数据集成任务关联使用。

  独享数据集成资源组的详细介绍可参见独享数据集成资源。

- 网络联通：

  独享数据集成资源组本质上为一组资源实例，购买添加完成后的初始状态下，与其他产品的网络并不联通，因此您需要为独享数据集成资源组绑定网络环境，为后续与数据源进行数据同步做好网络联通的准备。

购买独享数据集成资源组

1. 登录DataWorks控制台。
2. 选择相应地域后，在左侧导航栏，单击**资源组列表**。
3. 在**独享资源组**页面，单击**创建独享资源组**。
4. 在**创建独享资源组**对话框中，单击**订单号**后的**购买**，跳转至购买页面。
5. 进入购买页面后，请根据实际需要，选择相应的**地域**、**独享资源类型**、**资源数量**和**计费周期**，单击**立即购买**。

   > **说明** 此处的**独享资源类型**选择**独享数据集成资源**：
   > - 独享资源组不支持跨地域使用，即华东2（上海）地域的独享资源，只能给华东2（上海）地域的工作空间使用。
   > - 独享资源组的规格和性能请参见独享数据集成资源组计费说明：包年包月。

6. 确认订单信息无误后，勾选**《DataWorks独享资源（包年包月）服务协议》**，单击**去支付**。

新增独享数据集成资源组

1. 在**资源组列表 > 独享资源组**页面，单击**创建独享资源组**。
2. 在**创建独享资源组**对话框中，配置各项参数。

参数	描述
资源组类型	资源的使用类型。独享资源包括**独享调度资源**和**独享数据集成资源**两种类型，分别适用于通用任务调度和数据同步任务专用。
资源组名称	资源的名称，租户内唯一，请避免重复。 **说明** 租户即主账号，一个租户（主账号）下可以有多个用户（子账号）。
资源组备注	对资源进行简单描述。
订单号	此处选择购买的独享资源订单。如果没有购买，请单击**购买**，跳转至售卖页进行购买。

3. 配置完成后，单击**确定**。

说明 独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为**运行中**。

网络配置

独享资源部署在DataWorks托管的专有网络（VPC）中，与其他网络环境不连通。使用独享资源时，您需进行网络配置，为独享资源绑定一个可与数据源连通的VPC，通过此VPC实现与数据源网络连通。

1. 单击相应资源后的**网络设置**。

说明 绑定VPC前，您需要进行RAM授权，让DataWorks拥有访问云资源的权限。

2. 绑定专有网络VPC
   i. 单击**专有网络绑定**页面左上方的**新增绑定**，在**新增专有网络绑定**对话框中，配置各项参数，不同网络环境下各参数的配置说明如下。

   参数说明如下：

参数	配置说明（数据源与独享资源在同一VPC）	配置说明（数据源与独享资源不在同一VPC）
专有网络	如果您的数据源在阿里云VPC的网络环境中，建议配置为数据源所在的VPC。	如果您的数据源与独享资源不在同一VPC，例如，数据源不在阿里云VPC网络环境中，或需要将数据源与独享数据集成资源分别部署在不同VPC网络中时，您可单击**创建专有网络**，为独享数据资源创建一个VPC。创建完成后这里配置为新建的VPC。
交换机	专有网络配置为数据源所在VPC时，建议选择与数据源绑定的交换机。	专有网络配置为其他VPC，或没有可用交换机时，可单击**创建交换机**，为独享资源组单独创建一个交换机。创建完成后这里配置为创建的交换机。 **说明** 此种场景下，后续还需配置交换机路由，保障独享数据集成资源与数据源之间网络连通。
安全组	安全组指定了独享资源实例需要允许或禁止哪些公网或内网的访问。您可根据业务需求选择已有的安全组，或单击**创建安全组**为独享资源实例创建安全组。创建安全组的详细参数配置可参见添加安全组规则。	

   ii. 单击**确定**，完成绑定VPC操作。

3. （可选）配置Host

   如果您的数据源无法通过IP直接访问，例如，数据源通过Host的域名托管，通过Host域名直接被外部访问时，您需要参考以下步骤配置Host，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

i. 单击**Host配置**，在Host配置页面左上方单击**新增**，在**新增域名Host配置**对话框中配置各项参数。参数说明如下。

参数	配置说明
IP地址	配置为数据源的实际IP地址。
Host域名	配置为数据源对外提供访问服务的Host域名。如果有多个Host域名时，可换行依次添加。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。

ii. 如果有多个IP地址需要添加，可继续单击**新增**继续添加。

说明

- 新增的Host配置中，IP、域名不能与之前配置的Host中的IP或域名重复。
- 一个Host配置中，IP与域名为1对多的关系，即IP可以对应多个Host域名，但是同一个Host域名只能指向一个IP。

4. （可选）配置DNS

如果您的数据源无法通过IP直接访问，例如，数据源通过负载均衡域名直接被外部访问，通过内部域名解析服务器将域名解析至实际数据源IP地址时，您需要参考以下步骤配置DNS，否则在添加数据源时，使用数据源的Host域名进行连通性测试时会失败。

说明 如果同一个域名同时配置了Host和DNS，系统会优先根据Host的访问规则访问数据源。

i. 单击**DNS配置**，在DNS配置页面左下角单击**添加**，配置完成DNS各项参数后单击**保存**。参数说明如下。

参数	配置说明
Domain	非必配项。如果数据源有统一的一级域名，可在此处配置为数据源对外提供访问的域名的一级域名。 例如，数据源1对外的访问域名为domain1.example.com，数据源2对外的访问域名为domain2.example.com，此处建议配置为example.com。 **说明** 此处的域名需包含数字、字母、连字符（-）、点（.），且必须以字母开头，以字母或者数字结尾。
NameServer	配置为数据源提供域名解析的解析服务的IP地址。如果有多个域名解析服务器时，可换行依次添加。

ii. 如果您需要修改之前配置的DNS，您可单击右下角的**修改**。

完成独享数据集成资源的网络配置后，您还需添加独享资源组的EIP地址、专有网络的弹性网卡IP至数据库的白名单。

后续步骤

资源规划配置完成后，您可继续配置数据源，将来源数据源与去向数据源的网络、账号权限等准备工作完成，以便创建执行后续的数据同步任务。数据源的配置可参见配置数据源（来源为MySQL）、配置数据源（来源为Oracle）、配置数据源（来源为PolarDB）。

5.8.2. 配置数据源（来源为MySQL）

实时同步MySQL的数据至Kafka时，来源数据源为MySQL，去向数据源为Kafka，执行同步任务前，您需要参考本文在数据源中配置好网络、白名单等配置，为后续的数据同步做好网络环境和账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL、去向数据源Kafka。

- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

```
select version();
```

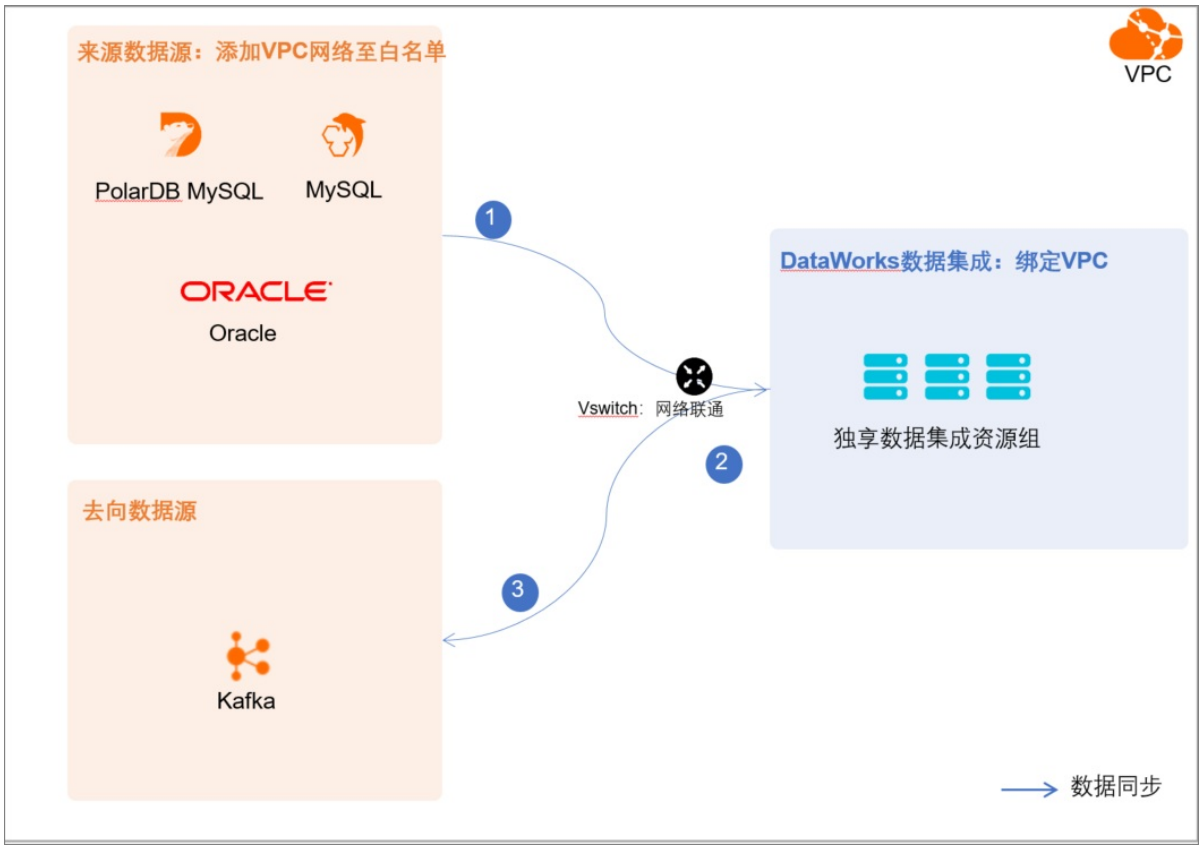
> **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

Binlog日志的格式如下：

- Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
- Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
- Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

使用限制

DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至MySQL的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

   a. 登录DataWorks控制台。

   b. 在左侧导航栏，单击**资源组列表**。

   c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

   d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

   e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

   f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至MySQL集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账号用于后续执行操作，此账号需要拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

   i. 创建账号。

   操作详情请参见创建MySQL账号。

ii. 配置权限。

您可以参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '密码'; //创建同步账号并设置密码，使其可以通过任意主机登录数据库。%表示任意主机。
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%'; //授权同步账号数据库的 SELECT, REPLICATION SLAVE, REPLICATION CLIENT权限。
```

`*.*` 表示授权同步账号对所有数据库的所有表拥有上述权限。您也可以指定授权同步账号对目标数据库的指定表拥有上述权限。例如，授权同步账号对*test*数据库的*user*表拥有上述权限，则可以使用 `GRANT SELECT, REPLICATION CLIENT ON test.user TO '同步账号'@'%';` 语句。

> **说明** `REPLICATION SLAVE` 语句为全局权限，不能指定授权同步账号对目标数据库的指定表拥有相关权限。

3. 开启MySQL的Binlog。

检查Binlog是否开启并查询Binlog格式，操作如下：

- 使用如下语句检查Binlog是否开启。

```
show variables like "log_bin";
```

返回结果为*ON*时，表明已开启Binlog。

- 如果您使用备用库同步数据，则还可以通过如下语句检查Binlog是否开启。

```
show variables like "log_slave_updates";
```

返回结果为*ON*时，表明备用库已开启Binlog。

如果返回的结果与上述结果不符，请参考*MySQL官方文档*开启Binlog。

使用如下语句查询Binlog的使用格式。

```
show variables like "binlog_format";
```

返回结果说明：

- 返回*ROW*，表明开启的Binlog格式为*ROW*。
- 返回*STATEMENT*，表明开启的Binlog格式为*STATEMENT*。
- 返回*MIXED*，表明开启的Binlog格式为*MIXED*。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.8.3. 配置数据源（来源为Oracle）

同步Oracle的数据至Kafka时，您需要参考本文，在数据源中完成网络、白名单及权限等相关配置，为后续执行数据同步方案做好网络环境及账号权限的准备。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

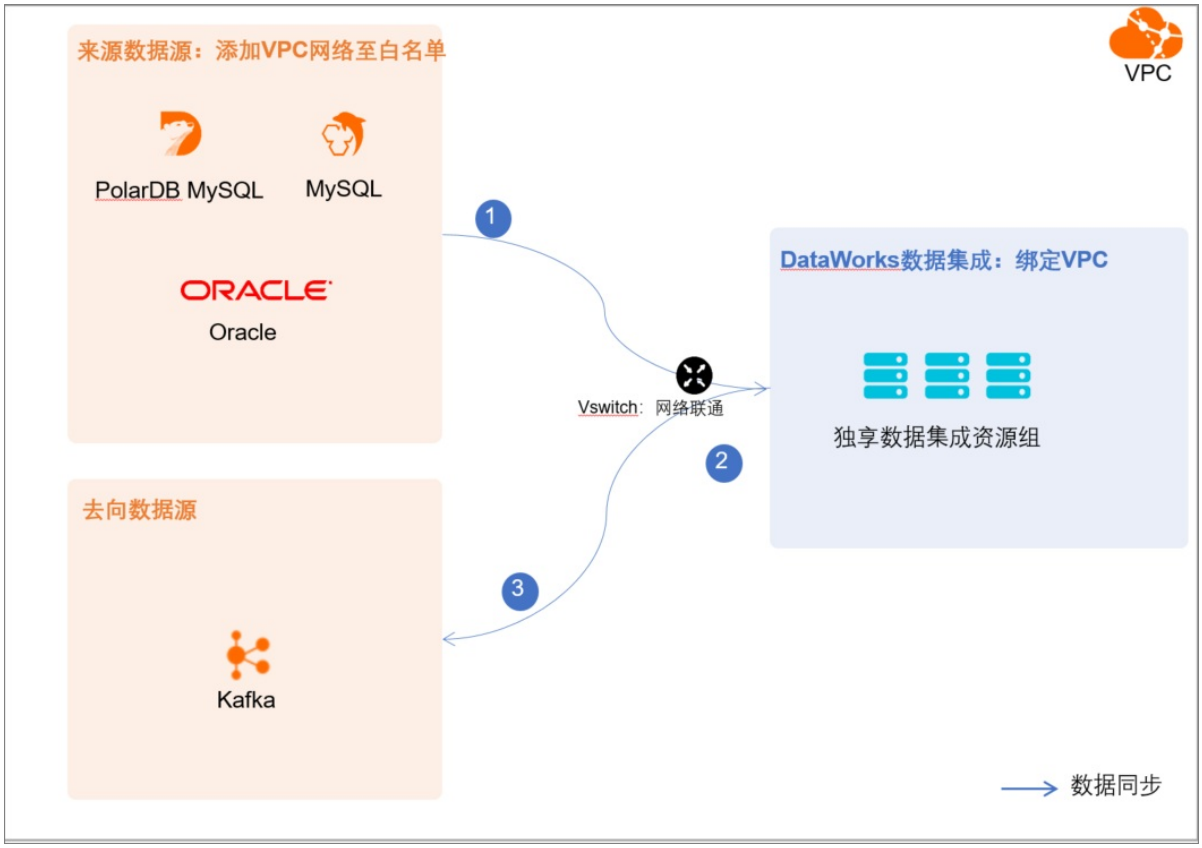
- 准备数据源：已购买来源数据源Oracle、去向数据源Kafka。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

同步来源数据源的数据至去向数据源时，您需要保障数据源与DataWorks的独享数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。同时，需要确保Oracle数据源中不存在数据集成不支持的数据库版本、字符编码及数据类型。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 查看当前使用的数据库版本是否为DataWorks数据集成实时同步任务所支持的版本。

  DataWorks的数据集成实时同步Oracle数据是基于Oracle Logminer日志分析工具实现的。实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

  i. 您可以通过如下任意语句查看Oracle数据库的版本。

  - 语句一：

    ```
    select * from v$version;
    ```

  - 语句二：

    ```
    select version from v$instance;
    ```

  ii. 如果查看到的Oracle数据库版本为 `12c` 、 `18c` 或 `19c` ，则您需要使用如下语句进一步确认该数据库是否为 `cdb` 类型的数据库。DataWorks数据集成实时同步任务暂不支持使用 `cdb` 类型的Oracle数据库。

  ```
  select name,cdb,open_mode,con_id from v$database;
  ```

> **说明** 如果当前使用的数据库版本不是DataWorks数据集成实时同步任务支持的Oracle数据库版本，请尽快更换为数据集成实时同步任务支持的Oracle数据库版本，否则会导致数据集成任务无法执行。

- 日志权限

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。

  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。
  - Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
  - 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

    例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

    使用数据集成时推荐开启主键列或唯一索引列补充日志。

    - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
    - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

  DataWorks数据集成实时同步Oracle数据前，您需要确保已为数据库开启归档日志及补充日志。查看当前使用的数据库是否开启数据库级别的归档日志及补充日志的SQL语句如下。

  ```
  select log_mode, supplemental_log_data_pk, supplemental_log_data_ui from v$database;
  ```

  - 当 `log_mode` 的返回结果为*ARCHIVELOG*，则表示数据库的归档日志已开启，当返回结果不为*ARCHIVELOG*，则表示数据库的归档日志未开启，您需要参考本文操作步骤的*开启归档日志*，开启归档日志。
  - 及 `supplemental_log_data_ui` 的返回结果为*YES*，则表示数据库的补充日志已开启，当返回结果为*FALSE*，则表示数据库的补充日志未开启，您需要参考本文操作步骤的*开启补充日志*，开启补充日志。

- 检查数据库的字符编码格式

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

- 检查是否包含不支持的数据类型

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

使用限制

- Oracle仅支持在主库中为主库或备库开启补充日志。
- 当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。
- 当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。
- 实时同步Oracle数据目前仅支持配置Oracle数据源为Oracle的 `10g` 、 `11g` 、 `12c non cdb` 、 `18c non cdb` 或 `19c non cdb` 版本数据库，不支持配置为Oracle的 `12c cdb` 、 `18c cdb` 及 `19c cdb` 版本数据库。数据库容器CDB（Container Database）是Oracle 12c及之后版本的数据库新特性，用于承载多个可插拔数据库PDB（Pluggable Database）。

注意事项

- DataWorks数据集成实时同步任务，目前对于Oracle主库支持订阅联机重做日志（Online Redo），对于Oracle备库仅支持订阅归档日志。因此，对于时效性要求比较高的实时同步任务，建议订阅主库的实时增量变更。订阅Oracle备库时，Oracle日志的产生到可以被获取的最短延迟时间取决于Oracle的自动切换归档日志的时间，不能保证时效性。
- Oracle数据库的归档日志建议保留3天。当写入大批量数据至Oracle数据库时，实时同步数据的速度可能会慢于日志生成的速度，方便在同步任务出现问题时，为追溯数据预留足够的时间。您可以通过分析归档日志排查问题并恢复数据。
- DataWorks数据集成实时同步任务，不支持对Oracle数据库中无主键的表进行 `truncate` 操作。对于无主键表进行日志分析（即 `logminer` 操作）是根据 `Rowid` 进行回查，当遇到 `truncate` 操作时会修改原表的 `Rowid` ，该操作会导致同步任务运行报错。

- 在规格为 `24 vCPU 192 GiB` 的DataWorks上运行实时同步任务时，如果非 `update` 等操作日志较多，并且速度达到约每秒记录3~5W条数据的极限速度，则Oracle服务器的单核CPU使用率最高可以达到25%～35%；如果处理 `update` 等操作日志，则处理实时同步消息的DataWorks机器可能会存在性能瓶颈，Oracle服务器的单核CPU使用率仅可以达到1%～5%。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至Oracle的白名单中，操作如下：

   i. 查看并记录独享数据资源组所在的VPC网络。

      a. 登录DataWorks控制台。

      b. 在左侧导航栏，单击**资源组列表**。

      c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

      d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

      e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

      f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

   ii. 将上述步骤中记录的独享数据集成资源组的EIP地址和网段添加至Oracle集群的白名单中。

2. 创建账号并配置账号权限。

   您需要规划一个数据库的登录账户用于后续执行操作，此账号需要拥有Oracle的相关操作权限。

   i. 创建账号。

      操作详情请参见创建Oracle账号。

   ii. 配置权限。

      您可以参考以下命令为账号添加相关权限。如下执行语句在实际使用时，请替换 `'同步账号'` 为上述创建的账号。

```
grant create session to '同步账号';  //授权同步账号登录数据库。
grant connect to '同步账号';  //授权同步账号连接数据库。
grant select on nls_database_parameters to '同步账号';  //授权同步账号查询数据库的nls_database_parameters系统配置。
grant select on all_users to '同步账号';  //授权同步账号查询数据库中的所有用户。
grant select on all_objects to '同步账号';  //授权同步账号查询数据库中的所有对象。
grant select on DBA_MVIEWS to '同步账号';  //授权同步账号查看数据库的物化视图。
grant select on DBA_MVIEW_LOGS to '同步账号';  //授权同步账号查看数据库的物化视图日志。
grant select on DBA_CONSTRAINTS to '同步账号';  //授权同步账号查看数据库所有表的约束信息。
grant select on DBA_CONS_COLUMNS to '同步账号';  //授权同步账号查看数据库中所有表指定约束中所有列的相关信息。
grant select on all_tab_cols to '同步账号';  //授权同步账号查看数据库中表、视图和集群中列的相关信息。
grant select on sys.obj$ to '同步账号';  //授权同步账号查看数据库中的对象。sys.obj$表是Oracle字典表中的对象基础表，存放Oracle的所有对象。
grant select on SYS.COL$ to '同步账号';  //授权同步账号查看数据库表中列的定义信息。SYS.COL$用于保存表中列的定义信息。
grant select on sys.USER$ to '同步账号';  //授权同步账号查看数据库的系统表。sys.USER$是用户会话的默认服务。
grant select on sys.cdef$ to '同步账号';  //授权同步账号查看数据库的系统表。
grant select on sys.con$ to '同步账号';  //授权同步账号查看数据库的约束信息。sys.con$记录了Oracle的相关约束信息。
grant select on all_indexes to '同步账号';  //授权同步账号查看数据库的所有索引。
grant select on v_$database to '同步账号';  //授权同步账号查看数据库的v_$database视图。
grant select on V_$ARCHIVE_DEST to '同步账号';  //授权同步账号查看数据库的V_$ARCHIVE_DEST视图。
grant select on v_$log to '同步账号';  //授权同步账号查看数据库的v_$log视图。v_$log用于显示控制文件中的日志文件信息。
grant select on v_$logfile to '同步账号';  //授权同步账号查看数据库的v_$logfile视图。v_$logfile包含有关Redo日志文件的信息。
grant select on v_$archived_log to '同步账号';  //授权同步账号查看数据库的v$archived_log视图。v$archived_log包含有关归档日志的相关信息。
grant select on V_$LOGMNR_CONTENTS to '同步账号';  //授权同步账号查看数据库的V_$LOGMNR_CONTENTS视图。
grant select on DUAL to '同步账号';  //授权同步账号查看数据库的DUAL表。DUAL是用来构成select语法规则的虚拟表，Oracle的中DUAL中仅保留一条记录。
grant select on v_$parameter to '同步账号';  //授权同步账号查看数据库的v_$parameter视图。v$parameter是Oracle的动态字典表，保存了数据库参数的设置值。
grant select any transaction to '同步账号';  //授权同步账号查看数据库的任意事务。
grant execute on SYS.DBMS_LOGMNR to '同步账号';  //授权同步账号使用数据库的Logmnr工具。Logmnr工具可以帮助您分析事务，并找回丢失的数据。
grant alter session to '同步账号';  //授权同步账号修改数据库的连接。
grant select on dba_objects to '同步账号';  //授权同步账号查看数据库的所有对象。
grant select on v_$standby_log to '同步账号';  //授权同步账号查看数据库的v_$standby_log视图。v_$standby_log包含备用库的归档日志。
grant select on v_$ARCHIVE_GAP to '同步账号';  //授权同步账号查询缺失的归档日志。
```

如果您涉及使用离线全量同步数据，还需要执行如下命令，授权同步账号所有表的查询权限。

```
grant select any table to '同步账号';
```

Oracle 12c及之后的版本需要执行如下命令，授权同步账号可以进行日志挖掘。Oracle 12c之前的版本，内置日志挖掘功能，无需执行该命令。

```
grant LOGMINING TO '同步账号';
```

3. 开启归档日志、补充日志并切换Redo日志文件。

   您需要进入主库执行如下操作：

   i. 开启归档日志，SQL语句如下。

```
shutdown immediate;
startup mount;
alter database archivelog;
alter database open;
```

ii. 开启补充日志。

您可以根据需要选择开启合适的补充日志，SQL语句如下。

```
alter database add supplemental log data(primary key) columns; //为数据库的主键列开启补充日志。
alter database add supplemental log data(unique) columns; //为数据库的唯一索引列开启补充日志。
```

iii. 切换Redo日志文件。

开启补充日志后，您需要多次（一般建议执行5次）执行如下命令，切换Redo日志文件。

```
alter system switch logfile;
```

> **说明** 多次执行上述命令切换Redo日志文件，是保证当前日志文件被写满后可以切换至下一个日志文件。使执行过的操作记录不会丢失，便于后续恢复数据。

4. 检查数据库的字符编码。

您需要在当前使用的数据库中，执行如下命令检查数据库的字符编码。

```
select * from v$nls_parameters where PARAMETER IN ('NLS_CHARACTERSET', 'NLS_NCHAR_CHARACTERSET');
```

- v$nls_parameters用于存放数据库参数的设置值。
- NLS_CHARACTERSET及NLS_NCHAR_CHARACTERSET为数据库字符集和国家字符集，表明Oracle中两大类字符型数据的存储类型。

当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。如果数据库中包含不支持的字符编码，请进行修改后再执行数据同步。

5. 检查数据库表的数据类型。

您可以使用查看表的SQL相关语句（SELECT）查询数据库表的数据类型。示例查看*'tablename'*表数据类型的语句如下。

```
select COLUMN_NAME,DATA_TYPE from all_tab_columns where TABLE_NAME='tablename';
```

- COLUMN_NAME：表的列名称。
- DATA_TYPE：对应列的数据类型。
- all_tab_columns：存放数据库表所有列相关信息的视图。
- TABLE_NAME：需要查询的目标表的名称。执行上述语句时，请替换*'tablename'*为实际需要查看的表名称。

您也可以执行 `select * from 'tablename';` ，查询目标表的所有信息，获取数据类型。

当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。如果表里包含这些字段类型，请将该表从实时同步任务列表中移除，或修改表字段类型后再执行数据同步。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.8.4. 配置数据源（来源为PolarDB）

将PolarDB的数据同步至Kafka时，您需要参考本文在数据源中配置好网络、白名单、权限等配置，为后续的数据同步方案执行做好网络环境和账号权限的准备。

前提条件

在进行数据源配置前，请确保已完成以下规划与准备工作。

- 数据源准备：已购买来源数据源PolarDB MySQL、去向数据源Kafka。本文以阿里云PolarDB MySQL作为来源数据源进行示例。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：进行数据集成前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，网络联通后参考本文进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。

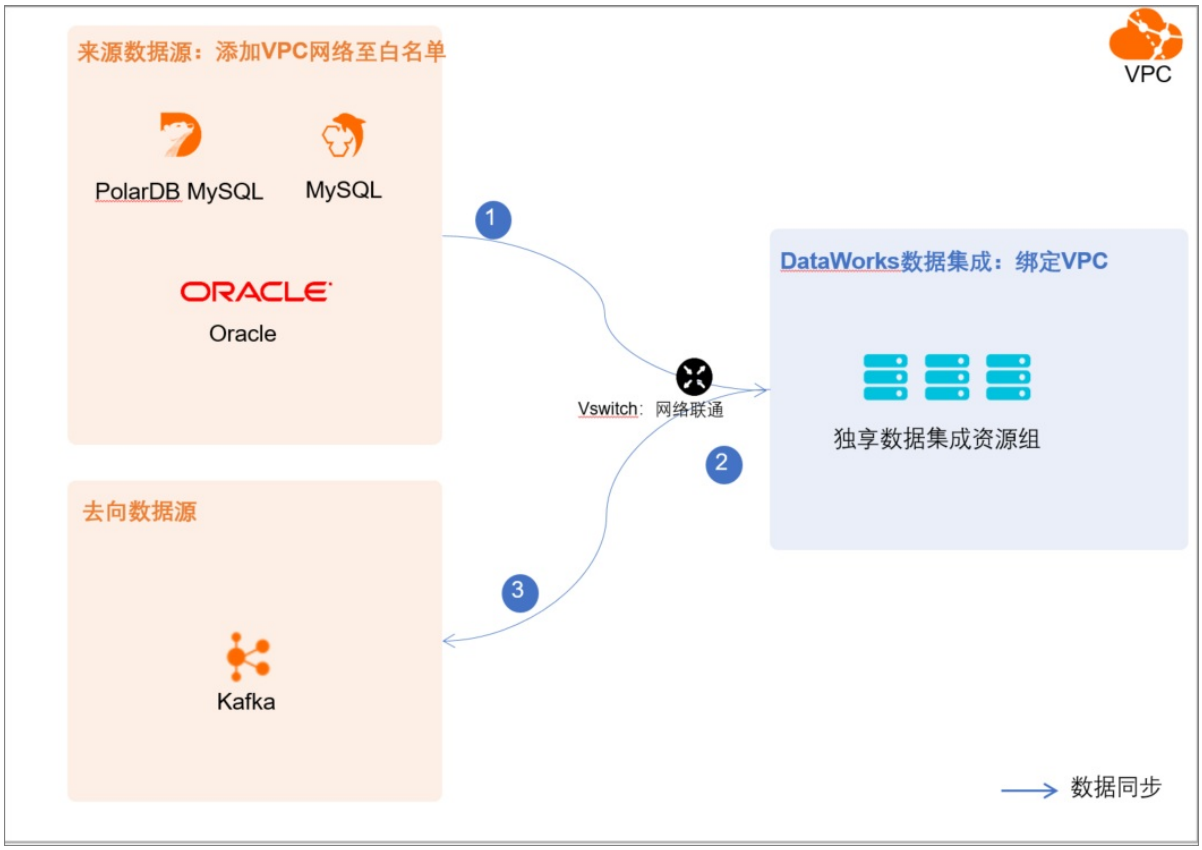
- 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

背景信息

将来源数据源的数据同步至去向数据源时，您需要保障数据源与DataWorks的数据集成资源组在网络上是联通的，且不存在账号权限的访问限制。

- 网络白名单

  以下以使用同一VPC网络环境为例，您需要将数据集成资源组所在的VPC网段添加至白名单中，保障数据集成资源组可访问数据源。



- 账号权限

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

- 其他访问限制。

  来源数据源为阿里云PolarDB MySQL时，您需要开启Binlog。阿里云PolarDB MySQL是一款完全兼容MySQL的云原生数据库，默认使用了更高级别的物理日志代替Binlog，但为了更好地与MySQL生态融合，PolarDB支持开启Binlog的功能。

使用限制

- 目前仅支持使用同步方案同步PolarDB MySQL类型的数据源，不支持同步其他类型的PolarDB数据源。文中均使用PolarDB代指PolarDB MySQL类型的数据源。
- PolarDB目前只能用主节点（读写库）进行实时同步。

操作步骤

1. 配置白名单。

   将独享数据资源组所在的VPC网段添加至PolarDB集群白名单中，操作如下：

i. 查看并记录独享数据资源组所在的VPC网络。

a. 登录DataWorks控制台。

b. 在左侧导航栏，单击**资源组列表**。

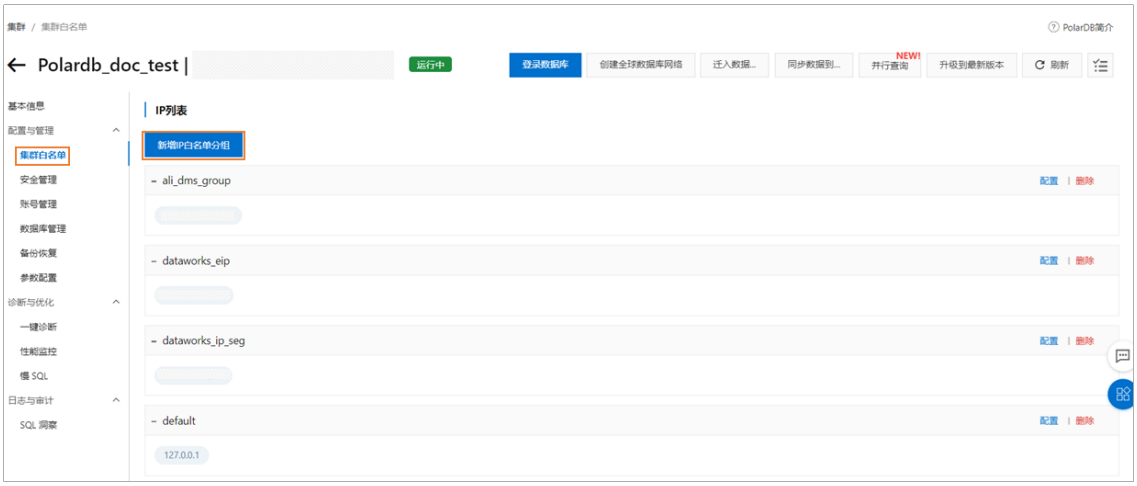
c. 在**独享资源组**页签下，单击目标数据集成资源组后的**查看信息**。

d. 复制对话框中的**EIP地址**和**网段**至数据库白名单。

e. 在**独享资源组**页签下，单击目标数据集成资源组后的**网络设置**。

f. 在**专有网络绑定**页签，查看**交换机网段**并将其添加至数据库的白名单中。

ii. 将上述步骤中记录的独享数据资源组的EIP和网段添加至PolarDB的白名单中。



操作详情可参见设置白名单。

2. 创建账号并配置账号权限。

您需要规划一个数据库的登录账户用于后续执行操作，此账户需拥有数据库的 `SELECT, REPLICATION SLAVE, REPLICATION CLIENT` 权限。

i. 创建账号。

操作详情可参见创建数据库账号。

ii. 配置权限。

您可参考以下命令为账号添加此权限，或直接给账号赋予 `SUPER` 权限。

```
-- CREATE USER '同步账号'@'%' IDENTIFIED BY '同步账号';
GRANT SELECT, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO '同步账号'@'%';
```

3. 开启PolarDB的开启Binlog。

操作详情可参见开启Binlog。

后续步骤

配置完成数据源后，来源数据源、资源实例、去向数据源彼此间已可网络联通，且不存在访问限制。您可将来源数据源和去向数据源添加至DataWorks的数据源列表中，便于后续创建数据同步方案时关联来源和去向数据源。

添加数据源操作可参见添加数据源。

5.8.5. 添加数据源

将来源数据源的数据同步至Kafka数据源的过程中，配置数据同步任务前，您需将来源数据源和去向数据源分别添加至DataWorks中，便于后续创建数据同步任务时进行来源和去向的配置。

前提条件

添加数据源前，您需检查已完成以下准备工作。

- 数据源开通：已购买开通来源数据源和去向数据源。
- 对接账号创建与授权：已在数据源数据库中创建好可对接访问的账号。

注意事项

DataWorks支持简单模式和标准模式两种环境模式，其中简单模式是不区分开发和生产环境，标准模式支持开发环境和生产环境隔离。

如果您使用的DataWorks是标准模式，您需要参考以下步骤，为开发环境和生产环境分别添加数据源。

添加来源数据源：MySQL

添加MySQL数据源时，需要根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情请参见配置MySQL数据源。

添加来源数据源：Oracle

添加Oracle数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置数据源（来源为Oracle）。

添加来源数据源：PolarDB MySQL

添加PolarDB MySQL数据源时，需根据您的规划，指定数据源与DataWorks的网络对接类型、对接账号密码等信息。您可根据规划进行配置添加，操作详情可参见配置PolarDB数据源。

如果添加PolarDB数据源时，联通性测试失败，可参考PolarDB数据源网络联通性测试失败怎么办？排查处理。

添加去向数据源：Kafka

操作详情可参见配置Kafka数据源。

后续步骤

添加完成数据源后，您可以创建并执行数据同步任务，将来源数据源的数据同步至去向数据源中。

操作详情可参见配置查看数据同步任务。

5.8.6. 配置查看数据同步任务

完成数据源、网络、资源的准备配置后，您可创建并执行数据同步任务，开始进行数据同步。本文为您介绍如何创建数据同步任务，并在创建完成后查看任务运行情况。

前提条件

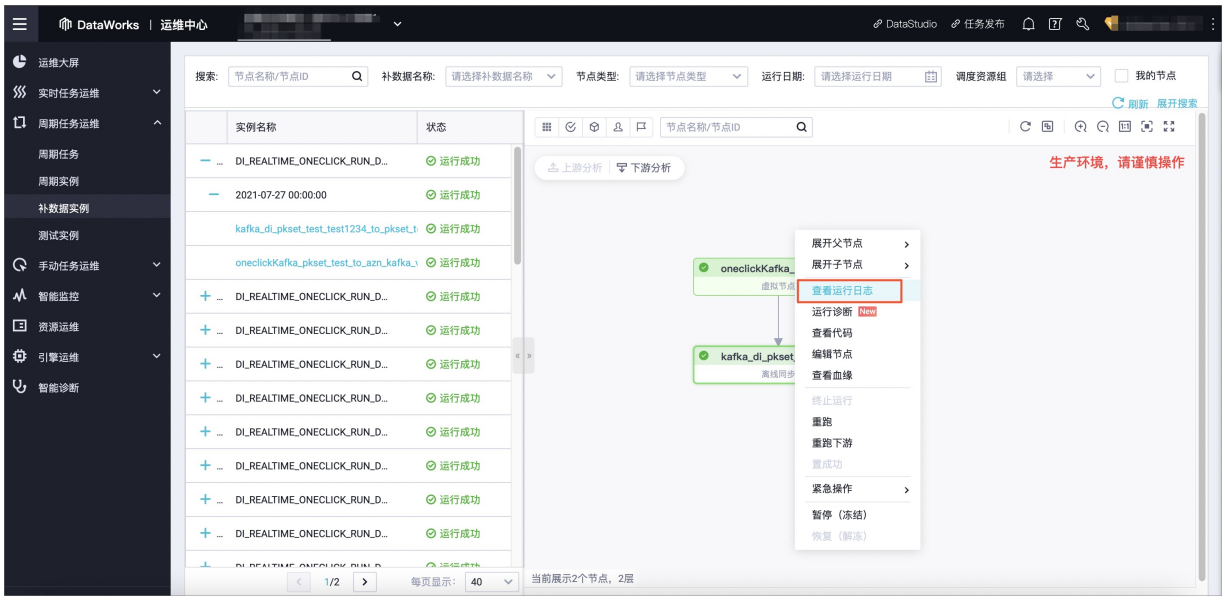
创建数据同步任务前，需检查已完成以下准备操作。

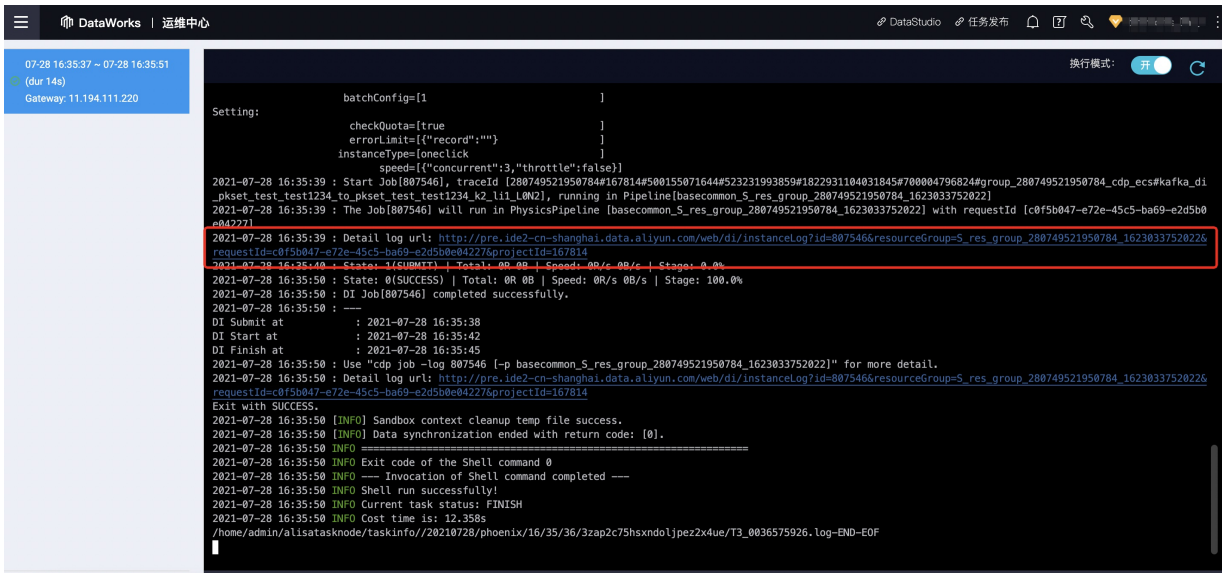
- 资源规划与配置
- 配置数据源（来源为MySQL）
- 配置数据源（来源为Oracle）
- 配置数据源（来源为PolarDB）
- 添加数据源

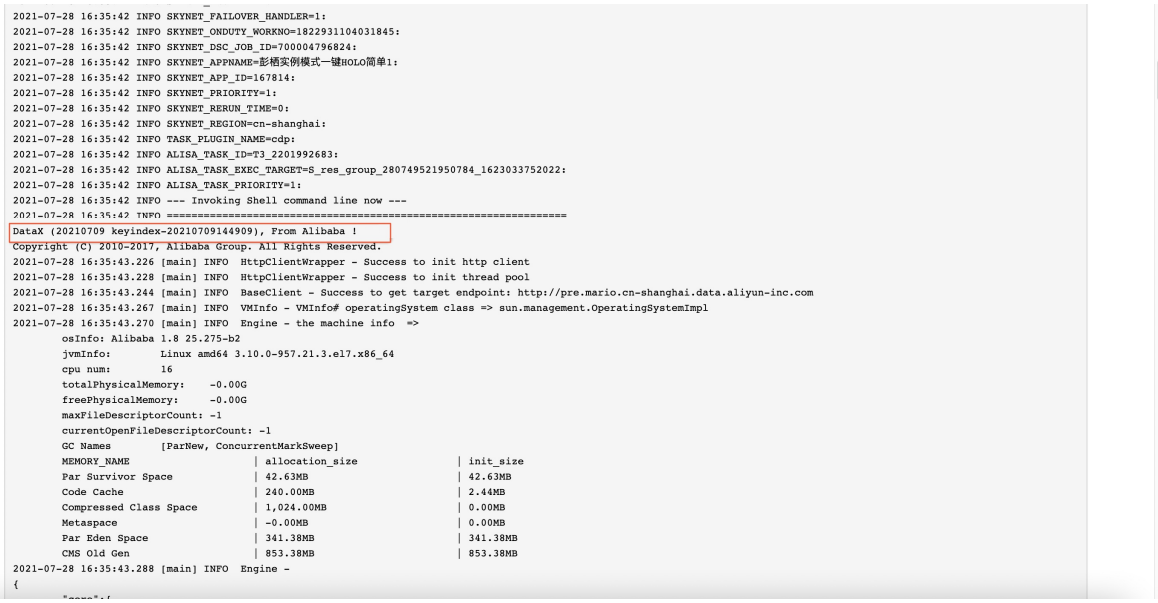
背景信息

独享集成资源组离线同步插件**DataX**版本必须大于等于20210726203000，实时同步插件**StreamX**版本必须大于等于202107121400，否则运行增量和全量同步Kafka数据时可能失败或者存在数据格式错误。

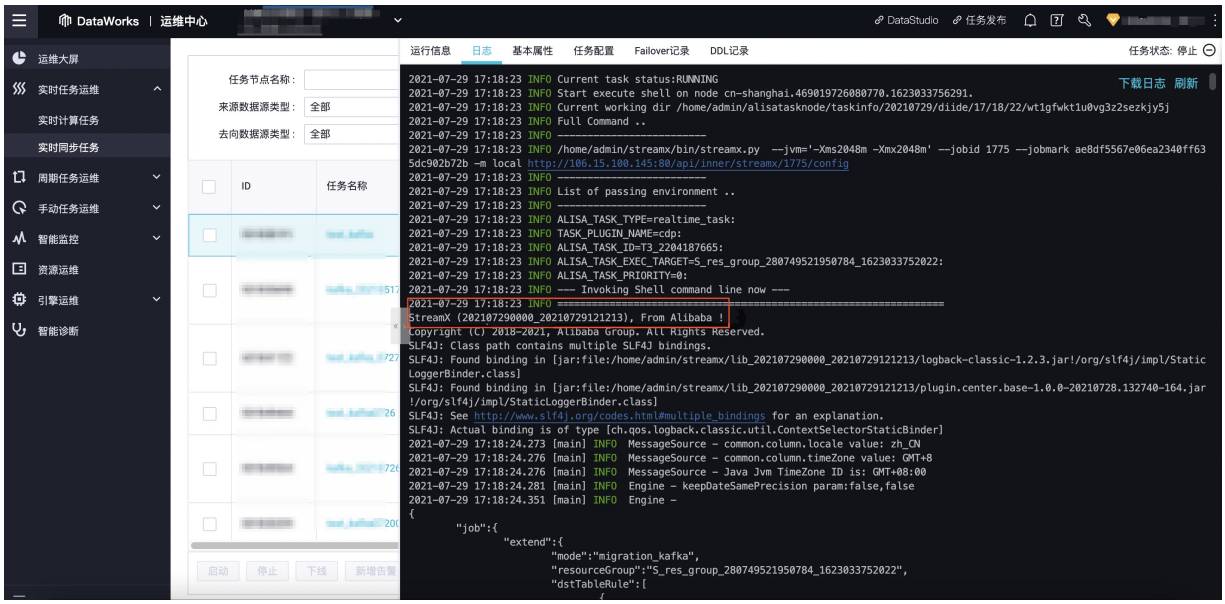
离线同步插件版本：在**运维中心>离线同步任务**日志中搜索 `Detail log url` 跳转到**离线同步详情**日志页，然后搜索 `DataX( ..... ),From Alibaba!` 格式文本，例如，`DataX (20210709_keyindex-20210709144909), From Alibaba !”`，截图中框起来的内容即离线同步插件**DataX**的版本。







实时同步插件版本：在**运维中心>实时同步任务**日志中搜索 `StreamX( ..... ),From Alibaba!` 格式文本，例如，`StreamX (202107290000_20210729121213), From Alibaba !` ，截图中内容即实时同步插件**StreamX**的版本。



注意事项

- 对于源端同步表有主键的场景，同步时会使用主键值作为kafka记录的key，确保同主键的变更有序写入kafka的同一分区。
- 对于源端同步表无主键的场景，如果选择了支持无主键表同步选项，则同步时kafka记录的key为空。如果要确保表的变更有序写入kafka，则选择写入的kafka topic必须是单分区。如果选择了自定义同步主键，则同步时使用其他非主键的一个或几个字段的联合，代替主键作为kafka记录的key。
- 如果在kafka集群发生响应异常的情况下，仍要确保有主键表同主键的变更有序写入kafka的同一分区，则需要在配置kafka数据源时，在扩展参数表单中加入如下配置。

  `{"max.in.flight.requests.per.connection":1,"buffer.memory": 100554432}` 。

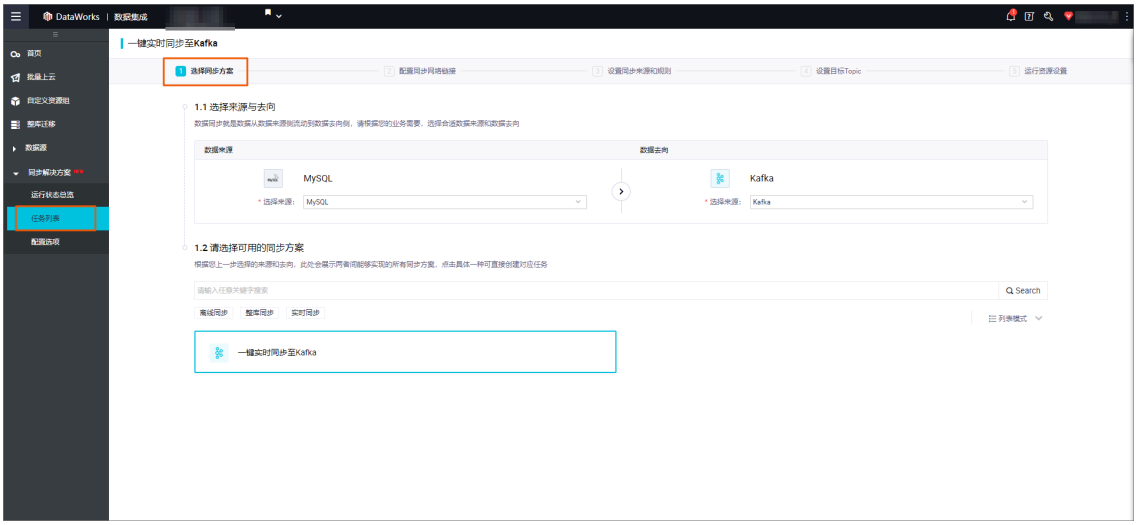
  > 注意 添加配置后同步性能会大幅下降，需要在性能和严格保序可靠性之间做好权衡。

- 实时同步写入kafka的消息总体格式、同步任务心跳消息格式及源端更改数据对应的消息格式，详情请参见：附录：消息格式。

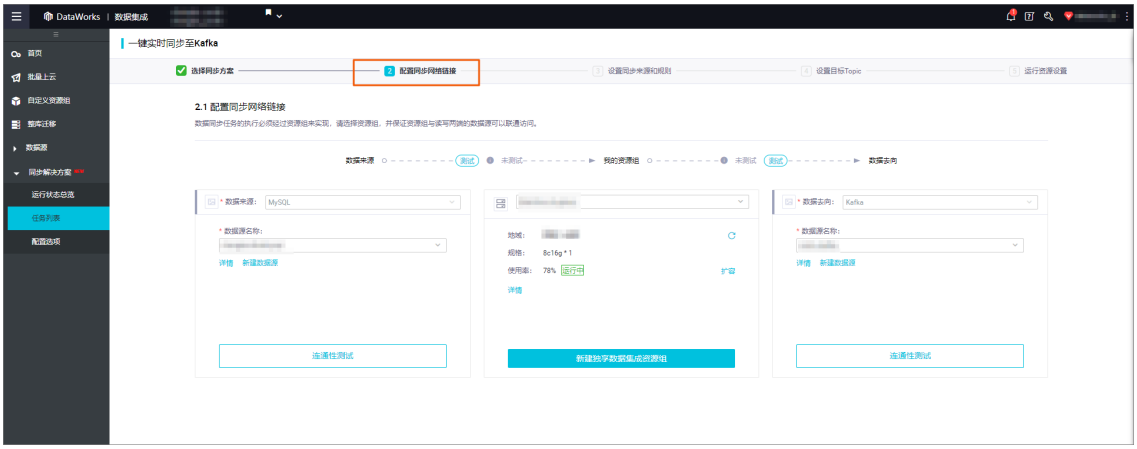
创建同步解决方案任务

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。
   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，单击右上方的**新建任务**。
3. 选择同步方案。
   i. 选择数据来源与去向。
   在**数据来源**和**数据去向**区域，选择**数据源**类型。

   > **说明** 一键实时同步至Kafka方案仅支持选择MySQL、Oracle和PolarDB类型的数据源。

   ii. 在**选择可用的同步方案**区域，单击**一键实时同步至Kafka**。
   iii. 单击**下一步**。
4. 配置同步网络链接。
   i. 选择**数据来源**和**数据去向**的**数据源名称**。如果下拉列表中没有数据源，单击**新建数据源**进行创建，详情请参见：配置MySQL数据源、配置Oracle数据源、配置PolarDB数据源。

   ii. 在**我的资源组**区域选择**独享数据集成资源组**。如果列表中没有资源组，单击**新建独享数据集成资源组**，选择相应**规格**、**资源数量**、**计费周期**，单击**确认购买**跳转到支付页面完成支付，请参见：资源规划与配置。

   > **说明**
   > 地域默认是该工作空间所在区域。
   > 购买后独享集成资源组默认与该工作空间绑定。

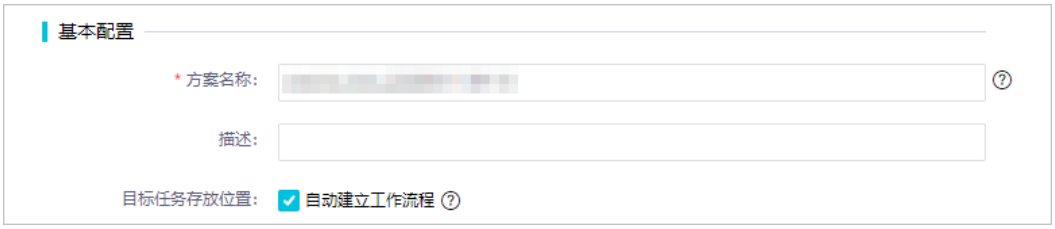
iii. 单击**连通性测试**验证独享数据集成资源组和数据源的网络连通性。详情请参见：配置资源组与网络连通。如果网络不通，请根据弹出的**网络连通性诊断工具**逐步诊断。

iv. 单击**下一步**。

5. 选择来源数据源并配置同步规则。

i. 完成方案名称等基本信息配置。

在**基本配置**区域，配置各项参数。



参数	描述
方案名称	同步解决方案的名称，最多支持50个字符。
描述	对当前方案进行简单描述，最多支持50个字符。
目标任务存放位置	默认创建一个新的业务流程，所有任务均以clone_database_源端数据源名称+to+目标数据源名称的命名方式存放至**数据集成**目录下。 您也可以取消**自动建立工作流程**，在**选择位置**下拉列表中指定存放目标任务的路径。

ii. 在**数据来源**区域，选择数据来源的**编码**格式。

iii. 在**选择同步的源表**区域，选中需要同步的**源端库表**，单击 > 图标，将其移动至**已选库表**。

该区域会为您展示所选数据源下所有的表，您可以选择整库全表或部分表进行同步。

iv. 在**设置表名到Topic的映射规则**区域，单击**添加规则**，选择相应的规则进行添加。

同步规则包括**源表名和目标Topic转换规则**和**目标Topic规则**：

- **源表名和目标Topic转换规则**：转换表名为目标表名，进行字符串替换。
- **目标Topic规则**：支持对转换后的表名添加前缀和后缀。

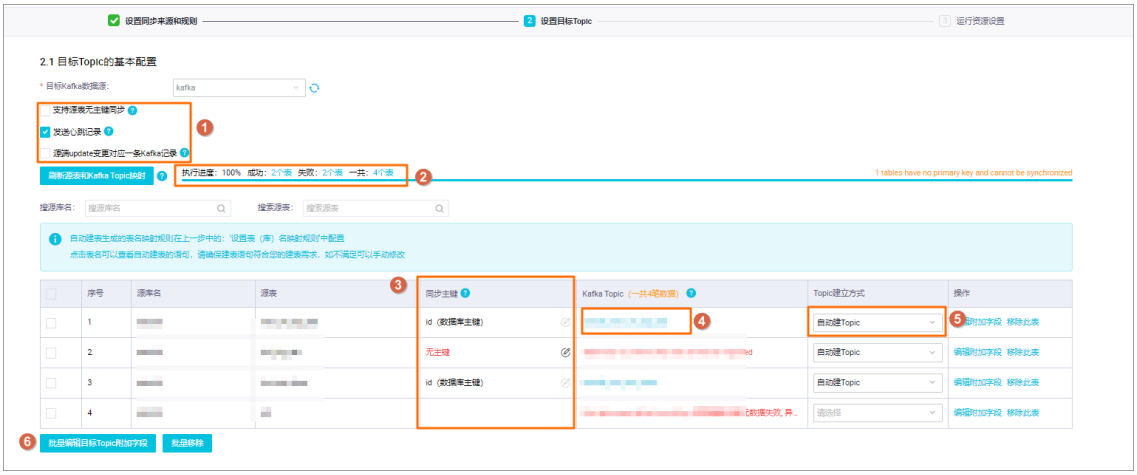
v. 单击**下一步**。

6. 配置目标Topic格式。

i. **目标Kafka数据源**默认为已配置好的数据源。

ii. 单击**刷新源表和Kafka Topic映射**，创建需要同步的源表和目标Kafka Topic的映射关系。

iii. 查看任务的执行进度和表来源。



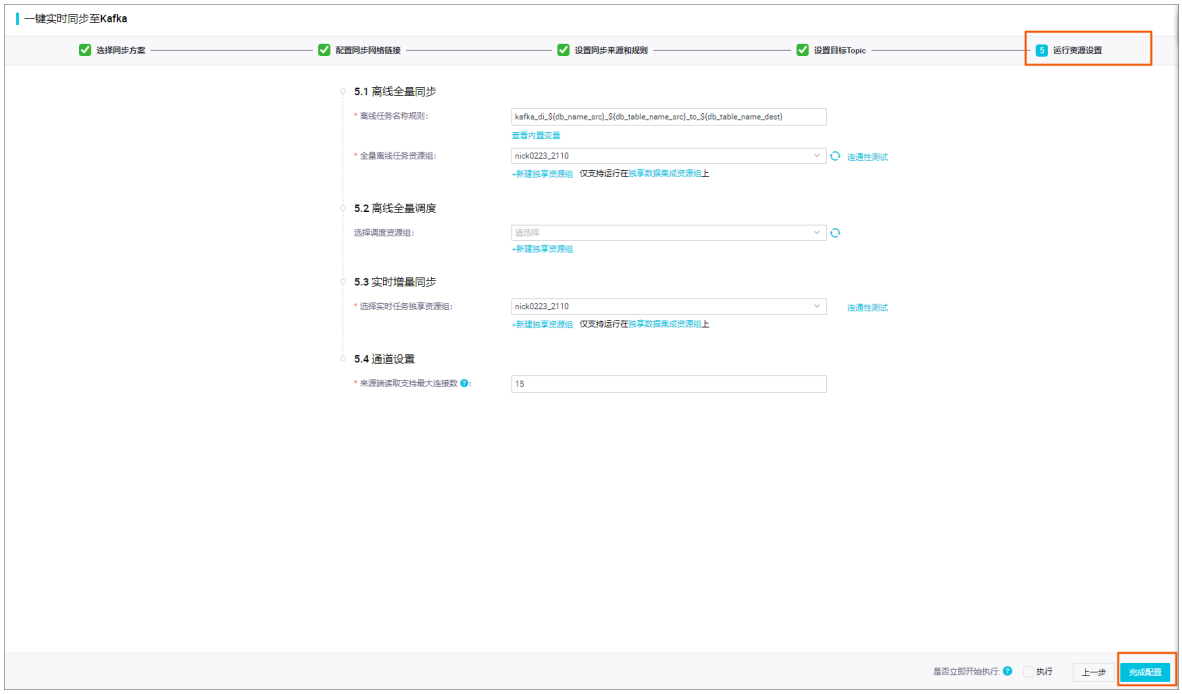
序号	描述
①	■ 勾选**支持源表无主键同步**后，源表没有主键，也可以向下游同步，但是同步数据时kafka记录的key将使用空值，只有当写入的kafka topic是单分区，才能确保变更有序写入。 ■ 勾选**发送心跳记录**后，实时同步任务将每隔5秒往Kafka中写入一条带有当前时间信息的记录。这样即使源端没有读取到新数据，Kafka中最新数据的时间信息也会持续更新，您可以根据Kafka中读取到的最新数据的时间判断实时同步的进度。 ■ 勾选**源端update变更对应一条Kafka记录**后，源端关系型数据库一条记录的一次update变更，变更前和变更后的数据将保存在一条Kafka记录中；如果未勾选，源端关系型数据库一条记录的一次update变更，将保存在两条Kafka记录中，分别保存变更前和变更后的数据。写入Kafka消息的消息结构及各字段含义详情请参见：附录：消息格式。
②	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
③	■ 如果来源库有主键，则同步数据时会使用该主键值作为kafka记录的key，确保同主键的变更有序写入kafka的同一分区。 ■ 如果来源库没有主键： ■ 当勾选了**支持源表无主键同步**，则无主键的表可以正常同步。此时写入kafka记录的key将使用空值，只有当写入的kafka topic是单分区，才能确保变更有序写入，此外，您还可以选择单击图标自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键作为kafka记录的key。 ■ 当在**设置目标Topic**页面未勾选**支持源表无主键同步**，则无主键的表同步时会出现异常，您需要在同步任务中删除无主键的表或者选择单击图标自定义主键才能继续执行同步任务。

序号	描述
④	包括**使用已有Topic**和**自动建Topic**。
⑤	选择的**Topic建立方式**，取值如下： ■ 当**Topic建立方式**选择**使用已有Topic**时，您可以在Kafka Topic列的下拉列表中选择需要使用的Topic名称。 ■ 当**Topic建立方式**选择**自动建Topic**时，显示自动创建的Kafka Topic名称。您可以单击Topic名称，查看和修改建Topic名称和注释。
⑥	在**批量编辑目标Topic附加字段**表单中给目标Kafka Topic增加字段。也可以单击**操作**列的**编辑附加字段**进行单表附加字段的设置。 **说明** 批量修改仅针对**Topic建立方式**选择**自动建Topic**的Topic生效。

iv. 单击**下一步**。

7. 运行资源设置。

   在**运行资源设置**页面，配置各项参数。



- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

8. 单击**完成配置**，完成整库实时同步任务的创建。

执行数据同步解决方案任务

在**解决方案任务列表**页面，单击相应任务后的**提交执行**，运行创建的数据同步解决方案任务。

如果任务执行失败，您可查看任务运行失败的错误提示，参考以下常见问题进行排查处理。

- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.PositionNotFoundException: can't find start position for XXX
- 实时任务，运行报错：com.alibaba.otter.canal.parse.exception.CanalParseException: command : 'show master status' has an error! pls check. you need (at least one of) the SUPER,REPLICATION CLIENT privilege(s) for this operation
- 实时任务，运行报错：com.alibaba.datax.plugin.reader.mysqlbinlogreader.MysqlBinlogReaderException: The mysql server does not enable the binlog write function. Please enable the mysql binlog write function first.
- 离线任务，运行报错：com.alibaba.datax.common.exception.DataXException: Code:[HoloWriter-02], Description:[Invalid config parameter in your configuration.]. - Field _log_file_name_offset_ not allow null but not present in user configured columns.

查看运行状态及结果

- 在**解决方案任务列表**页面，单击已运行任务后的**执行详情**，查看当前解决方案数据同步过程中各子任务节点的运行详情。

- 单击子任务节点后的**执行详情**，可在弹窗中单击任务链接进入子节点的数据开发页面。

管理数据同步解决方案任务

- 查看或编辑任务。

  在**解决方案任务列表**页面，单击相应任务后的**更多 > 查看配置**或**更多 > 修改配置**，可查看或编辑任务的配置信息。

- 删除任务。

  单击相应任务后的**更多 > 删除**。在**删除**对话框中，单击**确定**。

  > **说明** 仅删除当前任务的配置记录，已经生成的表和任务不受影响。

- 修改任务优先级。

  单击相应任务后的**更多 > 修改优先级**。在**修改优先级**对话框中，输入需要配置的优先级数值，单击**确定**。优先级取值范围为1~8，数值越大优先级越高。

  > **说明** 优先级相同的任务，按照提交时间的先后顺序执行。

写入Kafka消息格式定义

完成配置实时同步任务的操作后，执行同步任务会将源端数据库读取的数据，以JSON格式写入到Kafka topic中。除了会将设置的源端表中已有数据全部写入Kafka对应Topic中，还会启动实时同步将增量数据持续写入Kafka对应Topic中，同时源端表增量DDL变更信息也会以JSON格式写入Kafka对应Topic中。您可以通过附录：消息格式获取写入Kafka的消息的状态及变更等信息。

> **说明** 通过离线同步任务写入Kafka的数据JSON结构中的payload.sequenceId、payload.timestamp.eventTIme和payload.timestamp.checkpointTime字段均设置为-1。

5.8.7. 增加或删除已运行任务的同步表

同步数据至Kafka解决方案为您提供了一键增加及删除同步表功能，方便您为已成功配置运行的同步任务快速添加新表或删除已有同步表。本文为您介绍如何增加或删除已运行任务的同步表。

前提条件

已创建并运行同步数据至Kafka解决方案的任务，详情请参见配置查看数据同步任务。

同步任务新增表

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。

   操作详情可参见选择同步解决方案。

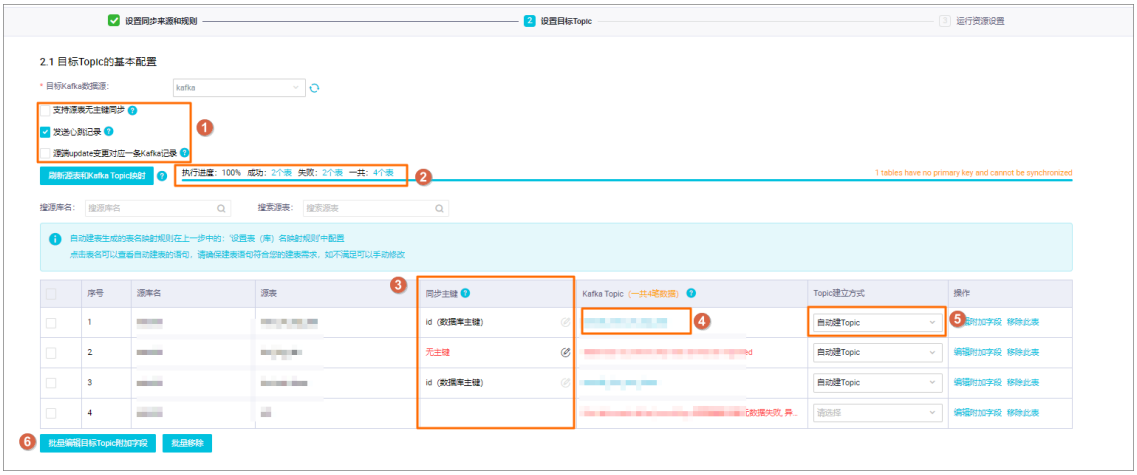
2. 在**解决方案任务列表**页面，选择目标同步任务后的**更多 > 修改配置**，进入任务配置页面。
3. 新增同步源表并更新源表至目标表的映射关系。

i. 在**设置同步来源和规则**页面的**选择同步的源表**区域，选中需要新增的同步**源端库表**，单击 > 图标，将其移动至**已选库表**。

ii. 单击**下一步**。

iii. 在**设置目标表**页面，单击**刷新源表和Kafka Topic映射**，更新需要同步的源表和目标Kafka Topic的映射关系。

iv. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 勾选**支持源表无主键同步**后，源表没有主键，也可以向下游同步，但是同步数据时不会进行去重。 ■ 勾选**发送心跳记录**后，实时同步任务将每隔5秒往Kafka中写入一条带有当前时间信息的记录。这样即使源端没有读取到新数据，Kafka中最新数据的时间信息也会持续更新，您可以根据Kafka中读取到的最新数据的时间判断实时同步的进度。
③	■ 如果来源库有主键，同步数据时会直接使用该主键进行去重。 ■ 如果勾选**支持源表无主键同步**，那么源表没有主键，您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
④	包括**使用已有Topic**和**自动建Topic**。
⑤	选择的**Topic建立方式**，取值如下： ■ 当**Topic建立方式**选择**使用已有Topic**时，**Kafka Topic**列显示自动创建的Kafka Topic名称。您也可以在下拉列表中选择需要使用的Topic名称。 ■ 当**Topic建立方式**选择**自动建Topic**时，显示自动创建的Kafka Topic名称。您可以单击Topic名称，查看和修改建Topic名称和注释。
⑥	在**批量编辑目标Topic附加字段**表单中给目标Kafka Topic增加字段。也可以单击**操作**列的**编辑附加字段**进行单表附加字段的设置。 **说明** 批量修改仅针对**Topic建立方式**选择**自动建Topic**的Topic生效。

4. 单击**下一步**。
5. 运行资源设置。

   在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

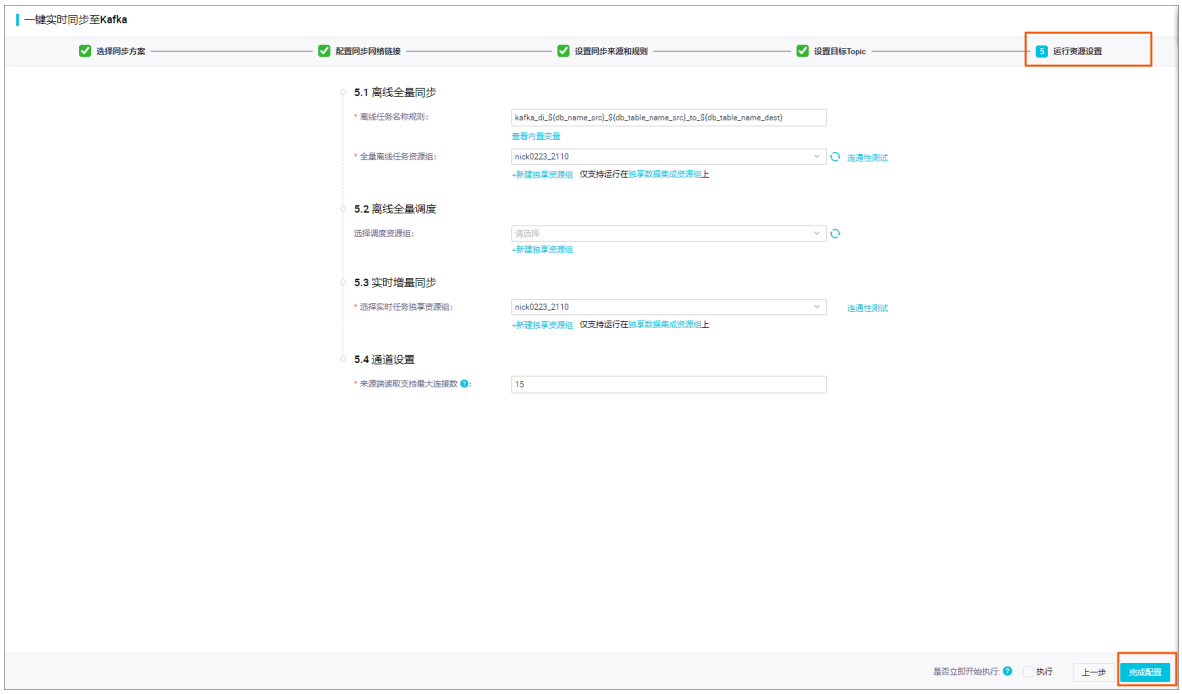
- **通道设置**

参数	描述

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

6. 运行资源设置。

   在**运行资源设置**页面，配置各项参数。



- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为**20**。

7. 单击**完成配置**，返回**解决方案任务列表**页面。
8. 单击上述修改任务**操作**列的**更多 > 提交执行**在**提交执行**对话框，单击**确定**，运行当前任务。

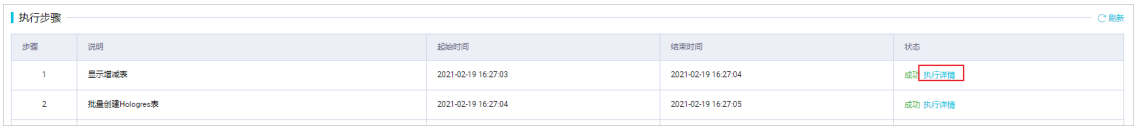
   提交执行任务时，会和上一次运行成功的任务对应的表做对比，当发现新增表时则会执行新增表的添加流程。


执行步骤 刷新

步骤	说明	起始时间	结束时间	状态
1	显示增减表	2021-04-15 19:47:37	2021-04-15 19:47:37	成功 执行详情
2	批量创建Kafka Topic	2021-04-15 19:47:37	2021-04-15 19:47:37	成功 执行详情
3	创建DataWorks工作流程	2021-04-15 19:47:37	2021-04-15 19:47:38	成功
4	创建DataWorks虚拟节点	2021-04-15 19:47:38	2021-04-15 19:47:39	成功 执行详情
5	创建全量同步任务节点	2021-04-15 19:47:39	2021-04-15 19:47:40	成功 执行详情
6	提交发布DataWorks虚拟节点	2021-04-15 19:47:40	2021-04-15 19:47:43	成功 执行详情
7	提交发布全量同步任务节点	2021-04-15 19:47:43	2021-04-15 19:47:47	成功 执行详情
8	全量同步任务节点批量冒烟执行	2021-04-15 19:47:47	2021-04-15 19:50:09	成功 执行详情
9	创建DataWorks实时同步节点	2021-04-15 19:50:10	2021-04-15 19:50:10	成功 执行详情
10	停止DataWorks实时同步节点	2021-04-15 19:50:10	2021-04-15 19:53:29	成功 执行详情
11	提交发布DataWorks实时同步节点	2021-04-15 19:53:29	2021-04-15 19:53:31	成功
12	启动DataWorks实时同步节点	2021-04-15 19:53:31	2021-04-15 19:54:16	成功 执行详情

说明 重置实时同步任务位点并启动运行时，会存在一个新增表追加变更数据的过程，即任务位点时间重置到新增表全量数据初始化时的时间。例如，您的同步任务8点开始运行，到9点时运行未结束。9点时新增了一张表，则全量数据初始化在9点开始执行，此过程耗时1小时，即全量数据初始化在10点完成。此时，已经正在运行的实时同步任务会先停止，然后重置任务位点到9点，进行追加增量数据，9点到10点之间所有变更表的增量数据会被重新同步至Kafka目标表，一键新增表只保证数据的最终一致性。

9. 查看同步任务新增表详情。
   i. 进入**任务列表**页面，单击目标任务**操作**列的**执行详情**，进入任务的执行详情页。
   ii. 在**执行步骤**区域，单击**显示增减表**后的**执行详情**。


执行步骤 刷新

步骤	说明	起始时间	结束时间	状态
1	显示增减表	2021-02-19 16:27:03	2021-02-19 16:27:04	成功 执行详情
2	批量创建Hologres表	2021-02-19 16:27:04	2021-02-19 16:27:05	成功 执行详情

显示增减表的状态为**成功**时，表明新增的表已成功添加至同步任务。

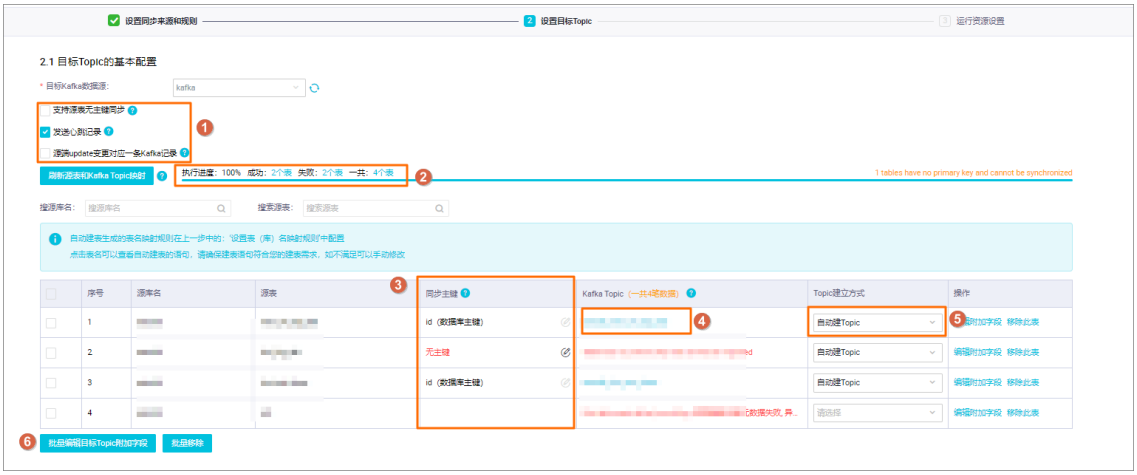
iii. 查看同步任务新增的同步表。

同步任务删除表

1. 登录并进入数据集成页面，单击**同步解决方案 > 任务列表**，进入同步解决方案页面。
   操作详情可参见选择同步解决方案。
2. 在**解决方案任务列表**页面，选择目标同步任务后的**更多 > 修改配置**，进入任务配置页面。
3. 删除同步源表并更新源表至目标表的映射关系。
   i. 在**设置同步来源和规则**页面的**选择同步的源表**区域，选中需要删除的**已选库表**，单击<图标，将其移回至**源端库表**。

   ii. 单击**下一步**。
   iii. 在**设置目标表**页面，单击**刷新源表和Kafka Topic映射**，更新需要同步的源表和目标Kafka Topic的映射关系。
   iv.

v. 查看任务的执行进度和表来源。



序号	描述
①	显示映射关系的创建进度。 **说明** 如果同步的表数量较多，会导致执行进度较慢，请耐心等待。
②	■ 勾选**支持源表无主键同步**后，源表没有主键，也可以向下游同步，但是同步数据时不会进行去重。 ■ 勾选**发送心跳记录**后，实时同步任务将每隔5秒往Kafka中写入一条带有当前时间信息的记录。这样即使源端没有读取到新数据，Kafka中最新数据的时间信息也会持续更新，您可以根据Kafka中读取到的最新数据的时间判断实时同步的进度。
③	■ 如果来源库有主键，同步数据时会直接使用该主键进行去重。 ■ 如果勾选**支持源表无主键同步**，那么源表没有主键，您需要单击图标，自定义主键，即使用其他非主键的一个或几个字段的联合，代替主键进行同步数据时进行去重判断。
④	包括**使用已有Topic**和**自动建Topic**。
⑤	选择的**Topic建立方式**，取值如下： ■ 当**Topic建立方式**选择**使用已有Topic**时，**Kafka Topic**列显示自动创建的Kafka Topic名称。您也可以在下拉列表中选择需要使用的Topic名称。 ■ 当**Topic建立方式**选择**自动建Topic**时，显示自动创建的Kafka Topic名称。您可以单击Topic名称，查看和修改建Topic名称和注释。
⑥	在**批量编辑目标Topic附加字段**表单中给目标Kafka Topic增加字段。也可以单击**操作**列的**编辑附加字段**进行单表附加字段的设置。 **说明** 批量修改仅针对**Topic建立方式**选择**自动建Topic**的Topic生效。

4. 单击**下一步**。
5. 运行资源设置。

   在**运行资源设置**页签，配置各项参数。

- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

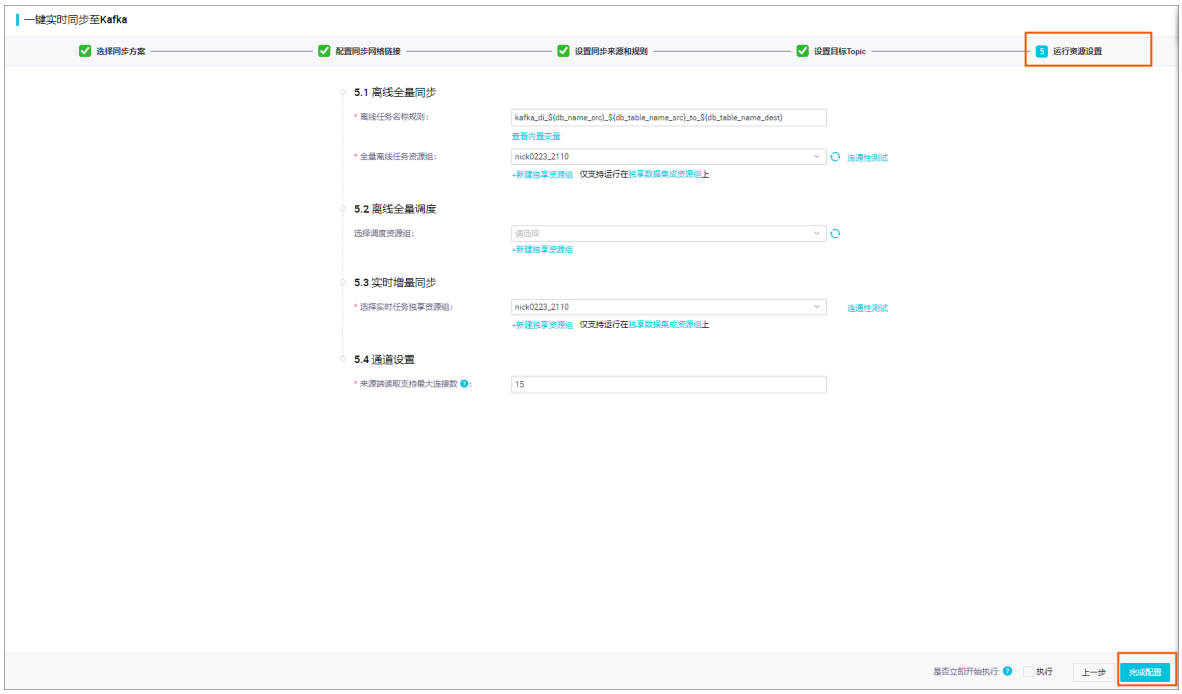
- **通道设置**

参数	描述

参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为20。

6. 运行资源设置。

   在**运行资源设置**页面，配置各项参数。



- **离线全量同步**

参数	描述
离线任务名称规则	全量同步时的离线任务名称。创建解决方案后，会先生成一个离线任务用于同步全量数据，再生成实时任务实时同步增量数据。
全量离线任务资源组	运行全量离线任务需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **离线全量调度**

参数	描述
选择调度资源组	选择运行任务时使用的调度资源组。 目前解决方案仅支持使用独享调度资源组，此处可配置为准备操作中已购买并配置的独享调度资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **实时增量同步**

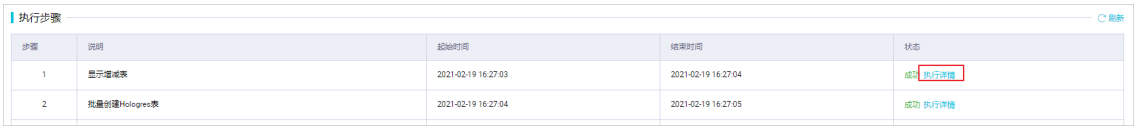
参数	描述
选择实时任务独享资源组	选择运行实时任务时需要使用的独享数据集成资源组。 目前解决方案仅支持使用独享数据集成资源组，此处可配置为准备操作中已购买并配置的独享数据集成资源组，详情请参见资源规划与配置。 **说明** 如果您没有购买独享资源组，可单击**+新建独享资源组**，创建新的独享资源组。

- **通道设置**

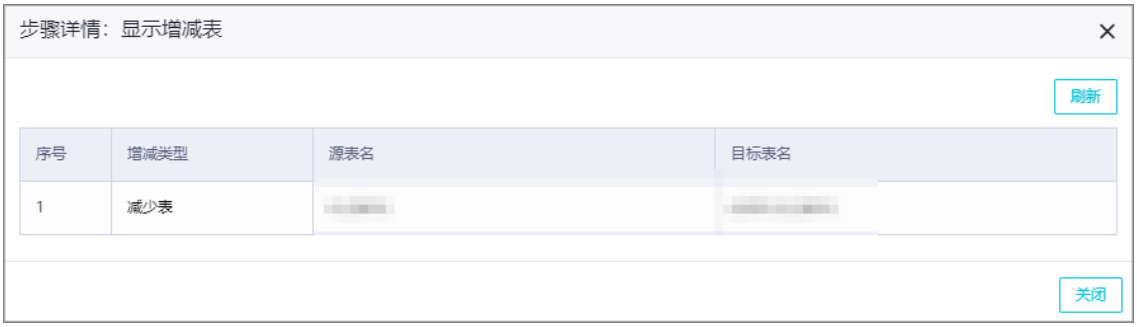
参数	描述
来源端读取支持最大连接数	读取端的最大连接数，即来源端数据库的JDBC连接数。请根据数据库资源的实际情况合理配置。默认为**20**。

7. 单击**完成配置**，返回**解决方案任务列表**页面。
8. 单击上述修改任务**操作**列的**更多 > 提交执行**在**提交执行**对话框，单击**确定**，运行当前任务。

   删除已运行同步任务中的源表时，会将目标源表从实时同步任务中移除。重新提交执行同步任务时，会直接在重启同步任务的时间点继续同步数据。
9. 查看同步任务删除表详情。
   i.
   ii. 在**执行步骤**区域，单击**显示增减表**后的**执行详情**。



   显示增减表的状态为**成功**时，表明目标源表已成功从同步任务中删除。
   iii. 查看此次执行任务删除的同步表。



5.9. 查看同步任务运行状态

运行状态总览为您展示所选时间周期中，目标同步解决方案任务的整体运行状态分布、资源使用率，以及离线和实时子任务的运行状态分布、同步速率、同步数据及任务延迟情况，帮助您快速查看已运行任务的分布及运行情况，及时发现并处理异常任务，提升任务的运维效率。

进入运行状态总览

1. 登录DataWorks控制台。
2. 在左侧导航栏，单击**工作空间列表**。
3. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
4. 在左侧导航栏，选择**同步解决方案 > 运行状态总览**。

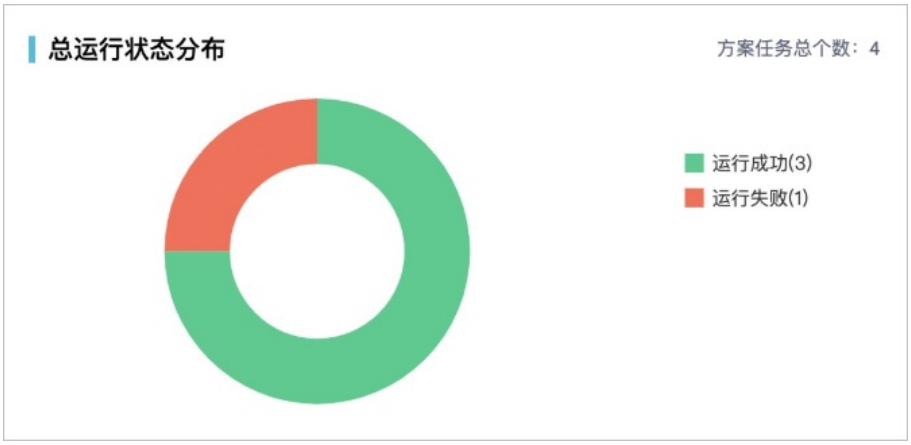
查看运行状态概况

您可以根据业务需求，选择查看所选时间周期中目标解决方案同步任务的运行状态概况。

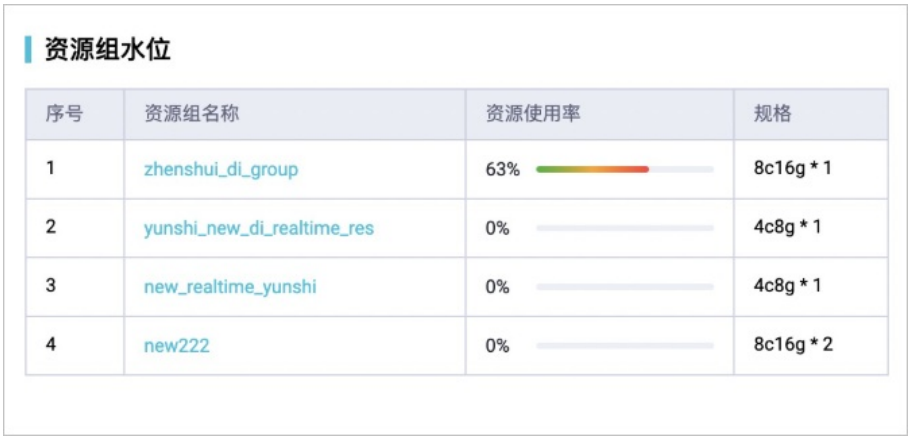
- 时间周期包括**最近一周**、**最近48小时**及**最近24小时**，默认选择**最近24小时**。
- 解决方案同步任务可以选择**整库离线同步至ElasticSearch**、**一键实时同步至Hologres**、**一键实时同步至MaxCompute**及**一键实时同步至ElasticSearch**的部分或全部类型任务，默认选择所有类型的同步任务。

运行状态总览的主要内容如下：

- **总运行状态分布**区域展示当前所选时间周期，目标任务的个数及运行结果的状态分布，统计的是任务运行成功及失败状况。您可以单击饼图中的某个色块，进入相应状态的任务列表，在任务列表页面查看相应任务的执行详情。任务的执行详情内容介绍请参见查看任务执行详情。



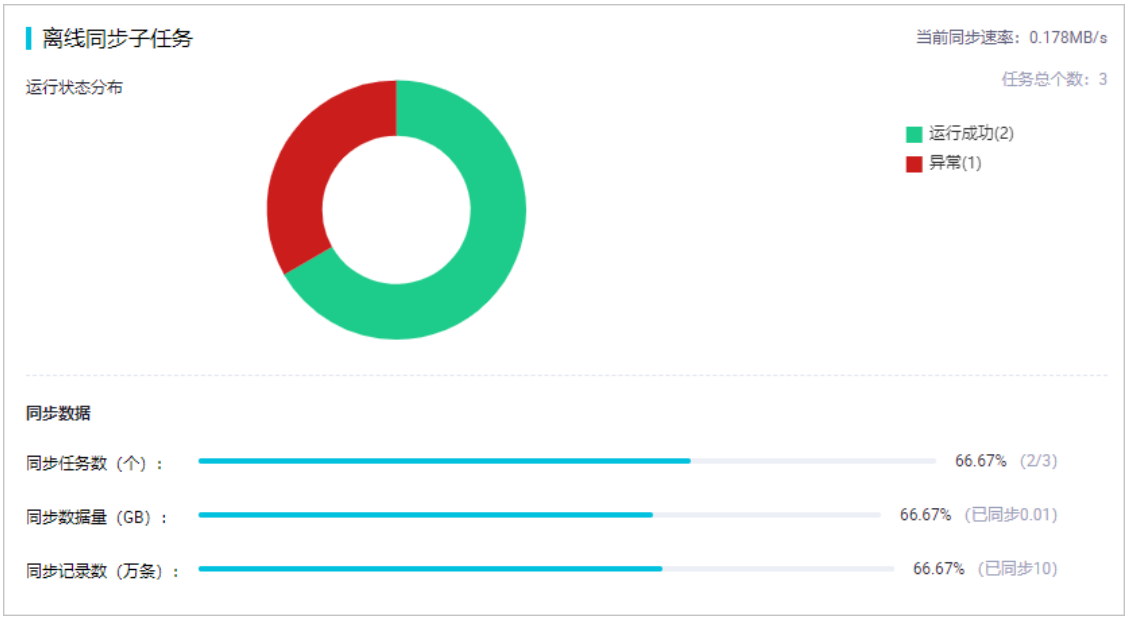
- **资源组水位**区域展示当前登录阿里云的账号所使用的资源组规格及使用情况。您可以单击资源组名称，进入目标资源组详情页面，查看资源组的基本信息及使用详情。资源组的详细内容介绍请参见查看独享资源组的使用详情。



- **离线同步子任务**区域展示当前所选时间周期，目标解决方案任务包含的离线同步任务个数、同步速率、运行结果的状态分布，以及同步数据情况。

> 说明　离线同步子任务的统计情况每小时更新一次。

  - 运行结果统计的是任务运行成功及异常状况。
  - 同步数据说明如下：
    - 同步任务数：显示当前同步任务中运行成功的任务个数。
    - 同步数据量：显示当前同步任务中的已同步数据量，包含已运行成功任务同步的数据量及运行中任务已同步的数据量。
    - 同步记录数：显示当前同步任务中的已同步数据的条数记录。



- **实时同步子任务**区域展示当前所选周期，目标解决方案任务包含的实时同步任务个数、同步速率、运行结果的状态分布，以及运行中任务延迟排行TOP10的延迟情况。您可以单击任务名称，进入实时任务运维页面查看任务详情。

查看任务执行详情

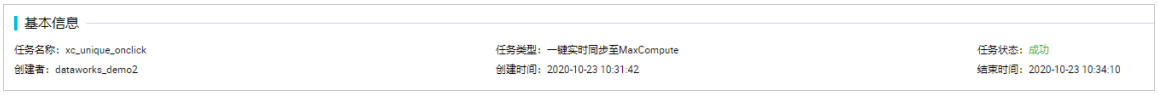
1. 在数据集成界面左侧导航栏，选择**同步解决方案 > 任务列表**，进入任务列表页面。

   任务列表页面，为您展示了同步解决方案任务的**任务ID**、**任务类型**、**任务名称**、**描述**、**创建时间**、运行**状态**、**创建者**、及可执行的相关**操作**等信息。您可以根据不同条件进行筛选，在该页面显示符合相应条件的同步任务。

2. 在目标同步任务**操作**列，选择**更多 > 执行详情**，查看任务执行的详细信息。

   执行详情页面的主要内容如下：

   - **基本信息**区域展示当前查看任务的执行状态、运行时间等信息。

   - **执行数据**区域展示当前查看任务的**同步前置环境准备**、**全量离线同步**、**实时同步**等子任务的执行状态。您可以根据执行状态，判断各个子任务是否正常运行，便于快速定位同步解决方案任务执行的阻塞点。状态如下：

     - 显示图标，则表示任务执行成功。
     - 显示图标，则表示任务执行异常。
     - 显示图标，则表示任务等待运行。

     示例一：下图为您展示了该解决方案同步任务运行成功。

     示例二：下图为您展示了该解决方案同步任务在**同步前置环境准备**子任务处运行异常，阻碍了**全量离线同步**和**实时同步**子任务的运行，使**全量离线同步**和**实时同步**长期处于待运行状态，导致整个解决方案同步任务运行失败。您可以根据执行数据快速判断解决方案同步任务正常执行的阻塞点，并进行异常清理。

   - **全量离线同步**区域展示当前查看的解决方案同步任务中，**全量离线同步**子任务同步的来源数据源、当前同步速率、同步数据、及所使用资源组的详细信息。

- 同步数据说明如下：
  - 同步任务数：显示当前同步任务中运行成功的任务个数。
  - 同步数据量：显示当前同步任务中的已同步数据量，包含已运行成功任务同步的数据量及运行中任务已同步的数据量。
  - 同步记录数：显示当前同步任务中的已同步数据的条数记录。
- 资源组的详细内容介绍请参见查看独享资源组的使用详情。

在该区域查看资源组的详细信息，步骤如下：

a. 单击资源组名称。

b. 在**全量离线同步详情**对话框，单击**详情**，即可跳转至该资源组的详情页，查看资源组的基本信息、使用率、使用趋势等详细内容。

- **实时同步**区域展示当前查看的解决方案同步任务中，**实时同步**子任务的任务名称、当前同步速率、同步数据的延迟情况、及所使用资源组的详细信息。资源组的详细内容介绍请参见查看独享资源组的使用详情。

在该区域查看资源组的详细信息，步骤如下：

a. 单击资源组名称。

b. 在**实时同步详情**对话框，单击**详情**，即可跳转至该资源组的详情页，查看资源组的基本信息、使用率、使用趋势等详细内容。

- **执行步骤**区域展示当前查看的解决方案同步任务，从创建任务到启动执行离线任务、实时任务的所有流程步骤。您可以在该区域查看所有步骤的执行时间及执行状态。

执行步骤

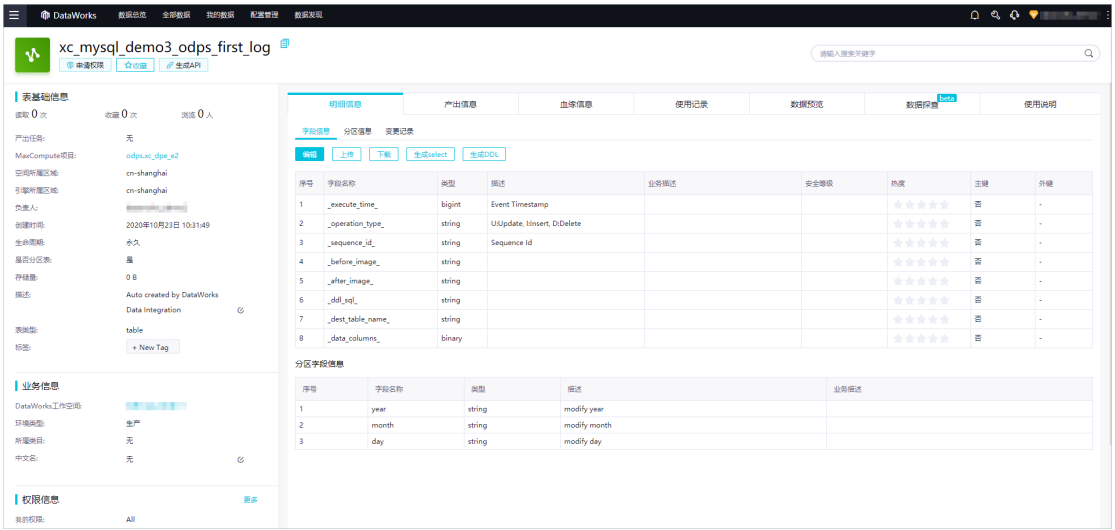
步骤	说明	起始时间	结束时间	状态
1	创建MaxCompute Log表	2020-10-23 10:31:48	2020-10-23 10:31:50	成功 执行详情
2	创建MaxCompute Delta表	2020-10-23 10:31:50	2020-10-23 10:31:52	成功 执行详情
3	创建MaxCompute Base表	2020-10-23 10:31:52	2020-10-23 10:31:54	成功 执行详情
4	创建DataWorks工作流程	2020-10-23 10:31:54	2020-10-23 10:31:54	成功
5	创建DataWorks虚拟节点	2020-10-23 10:31:55	2020-10-23 10:31:55	成功 执行详情
6	创建全量同步任务节点	2020-10-23 10:31:55	2020-10-23 10:31:56	成功 执行详情

查询目标步骤的任务详情，步骤如下：

a. 单击相应步骤**状态**列的**执行详情**。

b. 在显示的对话框单击任务名称，即可查看该步骤的详细内容。

示例查看**创建MaxCompute Log表**的详细内容。



6.附录
6.1. 配置数据源
6.1.1. 配置AnalyticDB for MySQL 2.0数据源

AnalyticDB for MySQL 2.0为您提供其它数据源向AnalyticDB for MySQL 2.0写入数据的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入数据集成页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 创建数据源。

   您可以通过以下两种方式创建数据源。

   > **说明**
   > - 数据集成模块仅支持创建生产环境的数据源。创建后，会在工作空间管理模块的**数据源管理**页面同步创建数据源。并且其中一处生产数据源变动，另一处会同步更新。
   > - 工作空间管理模块的**数据源管理**页面支持创建生产或开发环境数据源。
   > - 数据集成同步解决方案仅支持使用生产数据源。

   - 方式一：单击**数据集成 > 数据源**进入数据源列表页面，单击**新增数据源**。
   - 方式二：
     a. 单击右上角的**工作空间管理**图标进入进入**工作空间配置**页面。
     b. 在左侧导航栏，单击**数据源管理**，在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**AnalyticDB for MySQL 2.0**。
4. 在**新增AnalyticDB for MySQL 2.0数据源**对话框，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
连接Url	AnalyticDB for MySQL 2.0连接信息，格式为 `Address:Port` 。
数据库	AnalyticDB for MySQL 2.0的数据库名称。
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AceessKey Secret	访问密钥中的AccessKey Secret，相当于登录密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。
7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置AnalyticDB for MySQL 2.0数据源，您可以继续下一个教程。在该教程中，您将学习如何配置AnalyticDB for MySQL 2.0插件。

6.1.2. 配置SQLServer数据源

SQLServer数据源为您提供读取和写入SQLServer双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

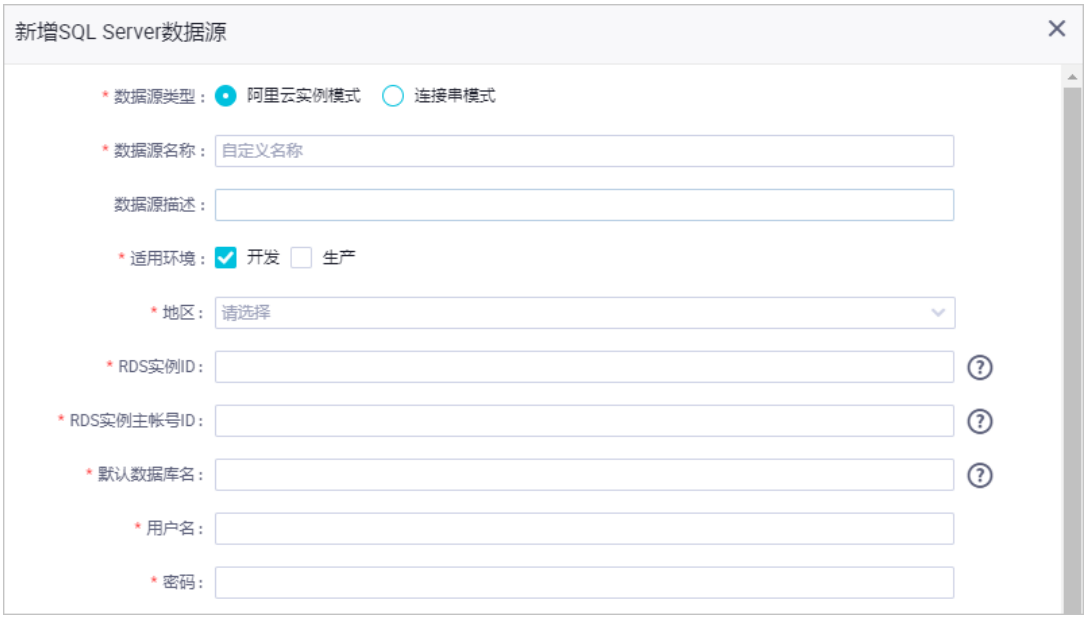
标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**SQLServer**。
4. 在**新增SQLServer数据源**对话框中，配置各项参数。
   i. 配置数据源的基本信息。

      SQLServer数据源包括**阿里云实例模式**和**连接串模式**两种类型。

- 以新增**阿里云实例模式**类型的数据源为例，配置数据源的基本信息。



参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	选择购买RDS的地区。
RDS实例ID	您可以进入RDS控制台，查看RDS实例ID。
RDS实例主账号ID	实例购买者登录DataWorks控制台，鼠标悬停至右上角的用户头像，查看账号ID。
默认数据库名	此处配置的是该数据源对应的默认数据库名称。后续配置同步任务的说明如下： ■ 配置整库同步（包含实时和离线）或同步解决方案任务时，您可以选择相应RDS实例下所有具有权限的数据库。 ■ 配置离线同步任务，当您选择使用多个数据库时，则每个数据库均需要配置一个数据源。
用户名	登录数据库的用户名称。
密码	登录数据库的密码。密码中避免使用@符号。

说明 您需要先添加RDS白名单才能连接成功，详情请参见添加白名单。

- 以新增**连接串模式**类型的数据源为例，配置数据源的基本信息。

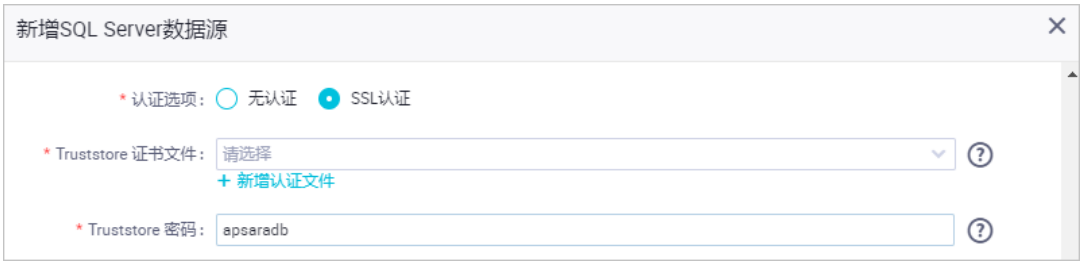
参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:sqlserver://ServerIP:Port;DatabaseName=Database` 。此连接串中的Database为本数据源的默认数据库，但在配置同步任务时，您可以使用相应RDS实例下所有的数据库。
用户名	登录数据库的用户名。
密码	登录数据库的密码。

ii. 配置数据源的认证信息。

第三方认证机制用于用户和服务的强身份验证，通过该机制，可以有效的避免不受信任的程序或服务来获取数据访问权限，提高数据同步过程中访问数据资源的安全性。DataWorks在配置数据源时为您提供了开启第三方认证功能（即**认证选项**配置为**SSL认证**），开启该功能后，只有可信的应用和服务才能访问数据资源。

> **说明** 使用**SSL认证**时，您需要提前在DataWorks的认证文件管理页面上传认证文件。上传和引用认证文件，详情请参见配置第三方身份认证。如果访问数据源时您无需对其他应用或服务进行认证，则可将**认证选项**配置为**无认证**。

配置认证信息如下图所示。



参数	描述
Truststore 证书文件	Truststore用于保存一些可信任的证书。当访问SSL服务器时对该证书进行认证，以确保访问本数据源的应用或服务是可信任的。 您可以单击**新增认证文件**上传认证文件。
Truststore 密码	■ 如果您配置的数据源使用的是阿里云RDS MySQL、RDS SQLserver、RDS PostgreSQL数据库，则密码固定为*apsaradb*。 ■ 如果您配置的数据源使用的是自建的RDS数据库，则**Truststore 密码**为您所配置的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> **说明**
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置SQLServer数据源，您可以继续下一个教程。在该教程中，您将学习如何配置SQLServer插件。详情请参见SQLServer Reader和SQLServer Writer。

6.1.3. 配置MongoDB数据源

MongoDB是目前仅次于Oracle、MySQL的文档型数据库，为您提供读取和写入MongoDB双向通道的功能，您可以通过脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选中数据源类型为**MongoDB**。
4. 在**新增MongoDB数据源**对话框中，配置各项参数。

   MongoDB数据源包括**阿里云实例模式**和**连接串模式**两种类型：

   > **说明** 如果DataWorks和MongoDB数据源不在同一个阿里云主账号名下，需要使用连接串模式。

   - **阿里云实例模式**：通常使用经典网络类型，同区域的经典网络可以连通，跨区域的经典网络不保证可以连通。

参数	描述

参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。 **说明** 如果您尚未授权数据集成系统默认角色，需要主账号前往RAM进行角色授权后，再刷新该页面。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	购买MongoDB时选择的区域。
实例ID	您可以在MongoDB控制台查看MongoDB实例ID。
数据库名	您可以在MongoDB控制台新建数据库，设置相应的数据名、用户名和密码。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- **连接串模式**：通常使用公网类型，可能产生一定的费用。

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
访问地址	格式为 `host:port` 。如果此处您需要同时添加多个地址，请单击**添加访问地址**进行添加。 **说明** 添加的访问地址必须全部为公网地址或全部为私网地址，不可以公网、私网地址混合。
数据库名	该数据源对应的数据库名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

如果您使用的是连接串模式的数据库，请通过下述操作添加MongoDB数据源：

a. 选择数据源类型为**连接串模式**。

b. 在**新增MongoDB数据源**对话框中，配置各项参数，其中访问地址填写您的内网地址。

c. 添加完成后，无需进行连通性测试，单击**完成**。

d. 添加自定义资源组，将任务运行在自定义资源组上，详情请参见新增和使用自定义数据集成资源组。

> 注意
>
> - MongoDB云数据库仅支持经典网络连通。
> - VPC环境的MongoDB云数据库，添加连接串模式数据源类型并保存。
> - VPC环境不支持测试连通性。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > 说明
   >
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置MongoDB数据源，您可以继续下一个教程。在该教程中，您将学习如何配置MongoDB插件，详情请参见MongoDB Reader和MongoDB Writer。

6.1.4. 配置DataHub数据源

DataHub数据源作为数据中枢，为您提供完善的数据导入方案，能够快速解决海量数据的计算问题。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

DataHub同步数据时，会根据DataHub Field的数据类型同步到对应的数据类型中，DataHub仅支持**BIGINT**、**STRING**、**BOOLEAN**、**DOUBLE**、**TIMESTAMP**、**DECIMAL**数据类型。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**DataHub**。
4. 在**新增DataHub数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源的简单描述，不超过80个字。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
DataHub Endpoint	默认只读，从系统配置中自动读取。
DataHub Project	对应的DataHub Project标识。
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AceessKey Secret	访问密钥AccessKey Secret，相当于登录密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置DataHub数据源，您可以继续下一个教程。在该教程中，您将学习如何配置DataHub Writer插件。详情请参见DataHub Writer。

6.1.5. 配置达梦（DM）数据源

达梦（DM）数据源为您提供读取和写入DM双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**DM**。
4. 在**新增DM数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 说明 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:dm://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

6.1.6. 配置DRDS数据源

DRDS数据源为您提供读取和写入DRDS双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**DRDS**。
4. 在**新增DRDS数据源**对话框中，配置各项参数。

   DRDS数据源包括**阿里云数据库（DRDS）**和**连接串模式**两种类型：

   - 以新增**阿里云数据库（DRDS）**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**阿里云数据库（DRDS）**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
实例ID	您可以登录DRDS控制台查看相应的实例ID。
主账号ID	实例购买者登录控制台，鼠标悬停至右上角的用户头像，单击**安全设置**，即可查看实例的主账号ID。
数据库名	数据库对应的名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**连接串模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置DRDS数据源，您可以继续下一个教程。在该教程中，您将学习如何配置DRDS插件。详情请参见DRDS Reader和DRDS Writer。

6.1.7. 配置FTP数据源

FTP数据源为您提供读取和写入FTP双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**FTP**。
4. 在**新增FTP数据源**对话框，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Portocol	目前仅支持FTP和SFTP协议。
Host	FTP的主机Host。
Port	如果选择FTP协议，端口默认为21。如果选择SFTP协议，端口默认为22。
用户名	访问该FTP服务的用户名。
密码	访问该FTP服务的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试**

连通性。详情请参见配置资源组与网络连通。

> **说明**
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置FTP数据源，您可以继续下一个教程。在该教程中，您将学习如何配置FTP插件。详情请参见FTP Reader和FTP Writer。

6.1.8. 配置HDFS数据源

HDFS是一个分布式文件系统，为您提供读取和写入HDFS双向通道的功能，本文为您介绍如何配置HDFS数据源。

背景信息

使用限制：目前不支持阿里云文件存储HDFS版。

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

当底层存储为OSS时，请注意以下问题：

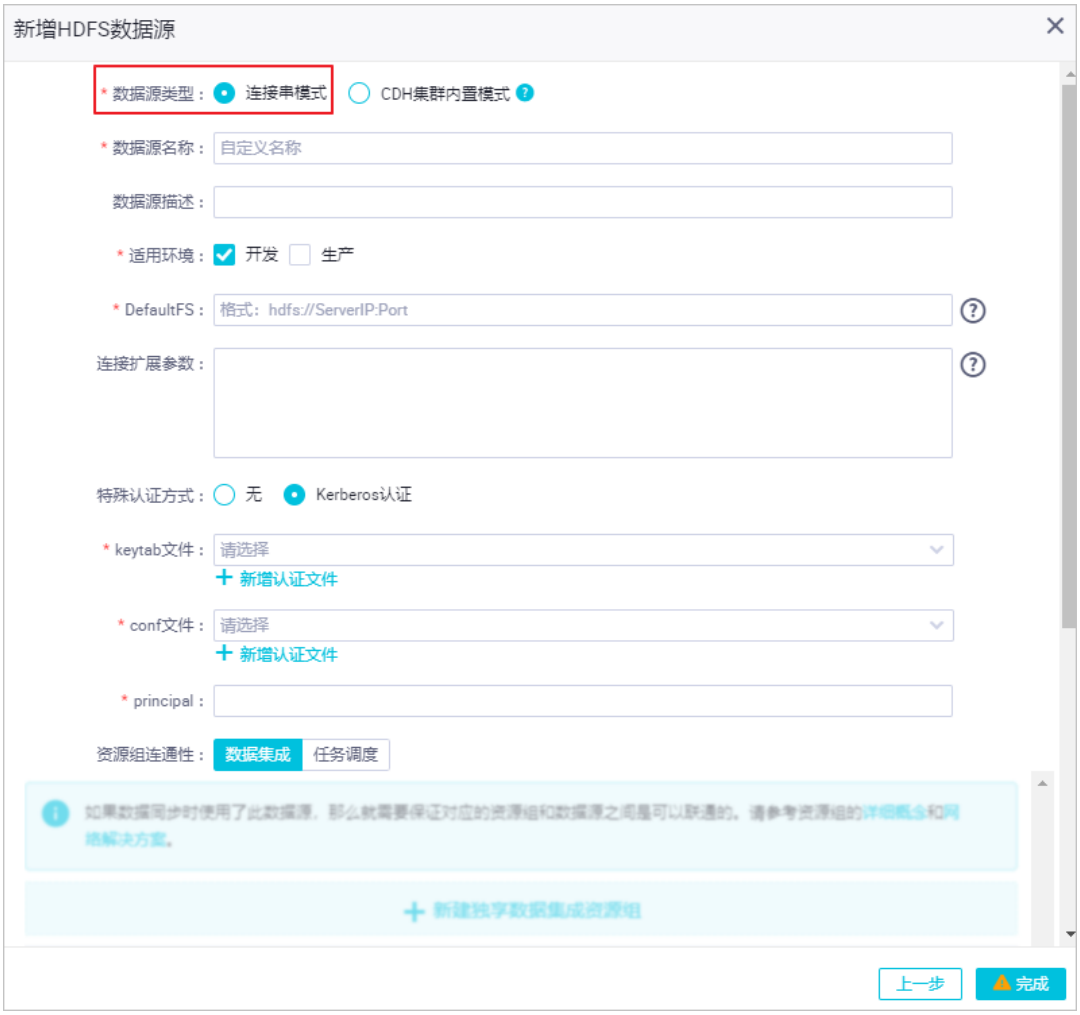
- defaultFS的配置请以oss://为前缀。例如，`oss://IP:PORT`或`oss://nameservice`。
- 您需要在高级参数中配置连接OSS服务时需要的参数，示例如下。

```
{
        "hadoopConfig":{
            "fs.oss.accessKeyId":"<yourAccessKeyId>",
                "fs.oss.accessKeySecret":"<yourAccessKeySecret>",
                "fs.oss.endpoint":"oss-cn-<yourRegion>-internal.aliyuncs.com"
        }
    }
```

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**HDFS**。
4. 在**新增HDFS数据源**对话框中，配置各项参数。

   HDFS数据源包括**连接串模式**和**CDH集群内置模式**两种类型：
   - 以新增**HDFS > 连接串模式**类型的数据源为例。



参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
DefaultFS	Hadoop HDFS文件系统中nameNode节点地址，格式为 `hdfs://ServerIP:Port`。
连接扩展参数	HDFS插件的hadoopConfig配置参数，您可以配置与Hadoop相关的高级参数，例如Hadoop HA的配置。
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。

参数	描述
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

- 以新增**HDFS > CDH集群内置模式**类型的数据源为例。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
选择CDH集群	选择已创建的CDH集群。

参数	描述
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明**
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置HDFS数据源，您可以继续下一个教程。在该教程中，您将学习如何配置HDFS插件。详情请参见HDFS Reader和HDFS Writer。

6.1.9. 配置LogHub（SLS）数据源

LogHub（SLS）数据源作为数据中枢，为您提供读取和写入LogHub（SLS）双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**LogHub**。
4. 在**新增LogHub数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
网络连接类型	取值如下： ○ **阿里云VPC** ○ **本地IDC** ○ **公网** ○ **经典网络**
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
LogHub Endpoint	通常格式为 `http://cn-shanghai.log.aliyun.com` 。详情请参见服务入口。 **说明** Endpoint地址仅支持填写相应的地址，不支持空格、斜线（/）等多余的符号。

参数	描述
Project	该数据源对应的Project。
AccessKey ID	当前阿里云登录账号的AccessKey ID，您可以进入用户信息管理页面获取。
AceessKey Secret	当前阿里云登录账号的AceessKey Secret，相当于登录密码。您可以进入用户信息管理页面获取。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > 说明
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置LogHub（SLS）数据源，您可以继续下一个教程。在该教程中，您将学习如何配置LogHub（SLS）插件。详情请参见LogHub（SLS） Reader和LogHub（SLS） Writer。

6.1.10. 配置Memcache（OCS）数据源

Memcache（原名OCS）数据源为您提供其它数据源向Memcache写入数据的功能，目前仅支持脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Memcache（OCS）**。
4. 在**新增Memcache（OCS）数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Proxy Host	进入Memcache控制台的基本信息页面，查看机器的IP或Host相应的**Proxy Host**。
Port	Memcache的连接端口，默认为11211。
用户名	该数据源对应的用户名。
密码	该数据源对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> **说明**
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Memcache数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Memcache Writer插件。详情请参见Memcache（OCS） Writer。

6.1.11. 配置MySQL数据源

MySQL数据源为您提供读取和写入MySQL双向通道的功能，方便您后续可以通过向导模式和脚本模式配置数据同步任务。本文为您介绍，配置数据源之前需要准备的网络环境及账号权限，以及在DataWorks中如何新增MySQL数据源。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源MySQL。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：新增数据源之前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，并进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

  详情请参见配置白名单。
- 查看当前使用的数据库版本是否为MySQL `5.x` 或 `8.x` 版本。您可以通过如下语句查看。

  ```
  select version();
  ```

  > **说明** DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。如果当前使用的数据库版本不是RDS的 `5.x` 或 `8.x` 版本的MySQL，请更换为使用RDS的 `5.x` 或 `8.x` 版本的MySQL，否则会导致数据集成任务无法执行。
- 准备账号并授权：

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

  详情请参见创建账号并配置账号权限。
- 开启MySQL：仅实时同步数据时需要开启MySQL。实时同步数据详细介绍请参见实时同步概述。

  来源数据源为MySQL时，您需要开启Binlog。Binlog是记录所有数据库表结构变更（例如执行CREATE、ALTER操作）以及表数据修改（例如执行INSERT、UPDATE、DELETE等）的日志。方便您可以通过Binlog日志中的内容，查看数据库的变更历史、进行数据库增量备份和恢复以及主从数据库的复制。

  Binlog日志的格式如下：
  - Statement：基于SQL语句的复制。Binlog中会保存每条修改数据的SQL语句。
  - Row：基于行的复制。Binlog中不保存SQL语句上下文的相关信息，仅保存被修改的记录。
  - Mixed：混合模式复制。Statement与Row的结合，一般的语句修改使用Statement格式（例如函数），Statement无法完成复制的操作，则采用Row格式保存Binlog。MySQL会根据执行的每条SQL语句自主识别使用哪种格式。

  详情请参见开启MySQL的Binlog。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

使用限制

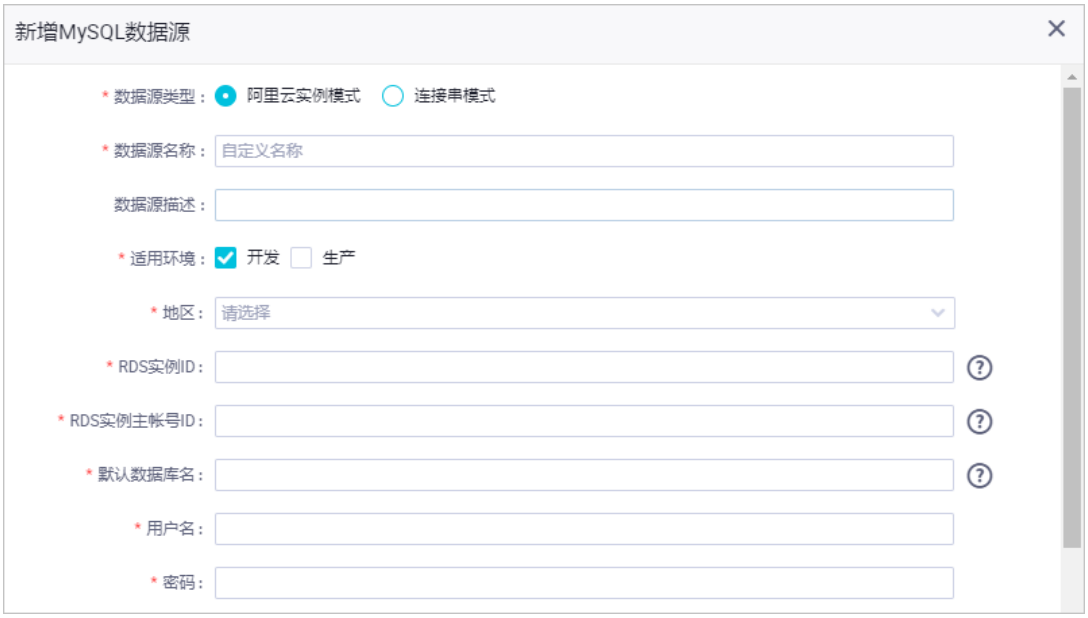
DataWorks的数据集成实时同步MySQL数据是基于实时订阅MySQL实现的，实时同步MySQL数据目前仅支持配置MySQL数据源为RDS的 `5.x` 或 `8.x` 版本的MySQL，不支持配置为DRDS的MySQL。

新增MySQL数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**MySQL**。
4. 在**新增MySQL数据源**对话框中，配置各项参数。
   i. 配置数据源的基本信息。

      MySQL数据源包括**阿里云实例模式**和**连接串模式**两种类型。

- 以新增**阿里云实例模式**类型的数据源为例，配置数据源的基本信息。



参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	选择相应的地域。
RDS实例ID	您可以进入RDS控制台，查看RDS实例ID。
RDS实例主账号ID	实例购买者登录DataWorks控制台，鼠标悬停至右上角的用户头像，查看账号ID。
默认数据库名	此处配置的是该数据源对应的默认数据库名称。后续配置同步任务的说明如下： ■ 配置整库同步（包含实时和离线）或同步解决方案任务时，您可以选择相应RDS实例下所有具有权限的数据库。 ■ 配置离线同步任务，当您选择使用多个数据库时，则每个数据库均需要配置一个数据源。
用户名	登录数据库的用户名称。
密码	登录数据库的密码。密码中避免使用@符号。

说明 您需要先添加RDS白名单才可以与RDS数据库连接成功，详情请参见添加白名单。

- 以新增**连接串模式**类型的数据源为例，配置数据源的基本信息。

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。此连接串中的Database为本数据源的默认数据库，后续在配置同步任务时，您可以使用相应RDS实例下，所有当前用户有登录权限的数据库。 **说明** 离线同步任务使用该数据源时默认访问本数据源的指定数据库。同步解决方案使用该数据源时可以读取对应实例下的所有数据库。
用户名	登录数据库的用户名。
密码	登录数据库的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置MySQL数据源，您可以继续下一个教程。在该教程中，您将学习如何配置MySQL插件。详情请参见MySQL Reader和MySQL Writer。

6.1.12. 配置Oracle数据源

Oracle数据源为您提供读取和写入Oracle双向通道的功能，方便您后续可以通过向导模式和脚本模式配置数据同步任务。本文为您介绍，配置数据源之前需要准备的网络环境及账号权限，以及在DataWorks中如何新增Oracle数据源。

前提条件

配置数据源之前，请确保已完成以下规划与准备工作。

- 准备数据源：已购买来源数据源Oracle。
- 资源规划与准备：已购买独享数据集成资源组，并完成资源配置。详情可参见资源规划与配置。
- 网络环境评估与规划：新增数据源之前，您需根据业务情况，打通数据源、独享数据集成资源组之间的网络，并进行交换机、白名单等网络环境下的访问配置。
  - 如果数据源和独享数据集成资源组均处于同地域的同一VPC网络中，数据源与资源组间的网络天然联通。
  - 如果数据源和独享数据集成资源组均处于不同的网络环境中，您需要通过VPN网关等方式，将数据源与资源组间的网络打通。

  详情请参见配置白名单。
- 准备账号并授权：

  您需要规划一个可访问数据源的账号，用于后续数据集成过程中访问数据源并进行数据提取、写入的同步操作。

  详情请参见创建账号并配置账号权限。
- 开启补充日志：

  来源数据源为Oracle时，您需要开启数据库级别的归档日志、Redo日志及补充日志。
  - 归档日志：Oracle通过归档日志保存所有的重做历史记录，用于在数据库出现故障时完全恢复数据库。
  - Redo日志：Oracle通过Redo日志来保证数据库的事务可以被重新执行，从而使得在故障（例如断电）之后，数据可以被恢复，因此您需要为数据库开启并切换Redo日志。
  - 补充日志：补充日志是对Redo日志中信息的补充。在Oracle中，Redo日志用于记录被修改的字段的值，而补充日志是对Redo日志中变更记录的补充信息，可以确保Oracle的Redo日志包含描述所有数据更改的完整信息，以便在进行数据恢复、数据同步等操作时，可以追溯到完整的语句及相关变更。Oracle数据库的某些功能要求启用补充日志才能正常或更好的工作，因此您需要为数据库开启补充日志。

    例如，如果未启用补充日志，执行UPDATE命令后，Redo日志中只会记录通过UPDATE命令更改后的字段值，启用补充日志后，则Redo日志中会记录被修改字段，修改前的值、修改后的值以及修改目标字段的条件值。当数据库发生故障（例如断电）时，您可以基于此修改信息恢复数据。

    使用数据集成时推荐开启主键列或唯一索引列补充日志。
    - 开启主键列的补充日志后，如果数据库有任何更新，则组成主键的所有列都会被记录在日志中。
    - 开启唯一索引列的补充日志后，如果组成唯一键或位图索引的任何列被修改，则组成该唯一键或位图索引的列都会被记录在日志中。

  Oracle仅支持在主库中为主库或备库开启补充日志。详情请参见开启补充日志并切换Redo日志文件。
- 检查数据库的字符编码格式：

  您需要确保Oracle中不能包含数据集成不支持的字符编码格式，防止同步数据失败。当前数据集成同步数据时，仅支持UTF8、AL32UTF8、AL16UTF16及ZHS16GBK编码格式。

  详情请参见检查数据库的字符编码。
- 检查数据库表的数据类型：

  您需要确保Oracle中不能包含数据集成不支持的数据类型，防止同步数据失败。当前数据集成进行实时同步时，不支持LONG、BFILE、LONG RAW及NCLOB数据类型 。

  详情请参见检查数据库表的数据类型。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

新增Oracle数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Oracle**。
4. 在**新增Oracle数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
JDBC URL	JDBC连接信息，格式为 `jdbc:oracle:thin:@ServerIP:Port:Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。

6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > 说明
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Oracle数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Oracle插件，详情请参见Oracle Reader和Oracle Writer。

6.1.13. 配置OSS数据源

对象存储Object Storage Service（简称OSS），是阿里云对外提供的海量、安全和高可靠的云存储服务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

如果您想对OSS产品有更深了解，请参见OSS产品概述。

OSS Java SDK请参见阿里云OSS Java SDK。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**OSS**。
4. 在**新增OSS数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Endpoint	OSS Endpoint信息，格式为 `http://oss.aliyuncs.com` ，OSS服务的Endpoint和区域有关。访问不同的区域时，需要填写不同的域名。 **说明** Endpoint的正确的填写格式为 `http://oss.aliyuncs.com` ，但 `http://oss.aliyuncs.com` 在OSS前加上Bucket值，以点号的形式连接。例如 `http://xxx.oss.aliyuncs.com` ，测试连通性可以通过，但同步会报错。
Bucket	相应的OSS Bucket信息，指存储空间，是用于存储对象的容器。 您可以创建一个或多个存储空间，每个存储空间可添加一个或多个文件。 您可以在数据同步任务中查找此处输入的存储空间中相应的文件，没有添加的存储空间，则不能查找其中的文件。

参数	描述
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AceessKey Secret	访问密钥中的AccessKey Secret，相当于登录密码。

> **注意** 准备OSS数据时，如果数据为CSV文件，则必须为标准格式的CSV文件。例如，如果列内容在半角引号（"）内，需要替换成两个半角引号（""），否则会造成文件被错误分割。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明**
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置OSS数据源，您可以继续下一个教程。在该教程中，您将学习如何配置OSS插件，详情请参见OSS Reader和OSS Writer。

6.1.14. 配置OTS数据源

表格存储Table Store（简称OTS）是构建在阿里云飞天分布式系统之上的NoSQL数据存储服务，为您提供海量结构化数据的存储和实时访问。

背景信息

表格存储的更多详情请参见什么是表格存储。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**OTS**。
4. 在**新增OTS数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Endpoint	Table Store服务对应的Endpoint。
Table Store实例名称	Table Store服务对应的实例名称。
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AccessKey Secret	访问密钥中的AccessKey Secret，相当于登录密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置OTS数据源，您可以继续下一个教程。在该教程中，您将学习如何配置OTS Reader插件。详情请参见Table Store（OTS） Reader。

6.1.15. 配置PostgreSQL数据源

PostgreSQL数据源为您提供读取和写入PostgreSQL双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

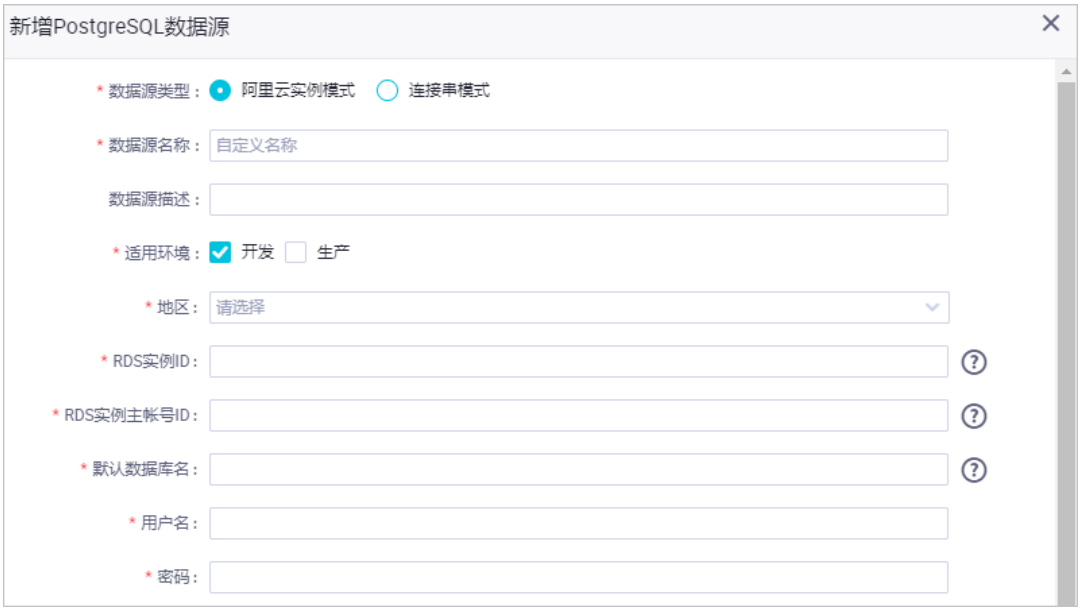
标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**PostgreSQL**。
4. 在**新增PostgreSQL数据源**对话框中，配置各项参数。
   i. 配置数据源的基本信息。

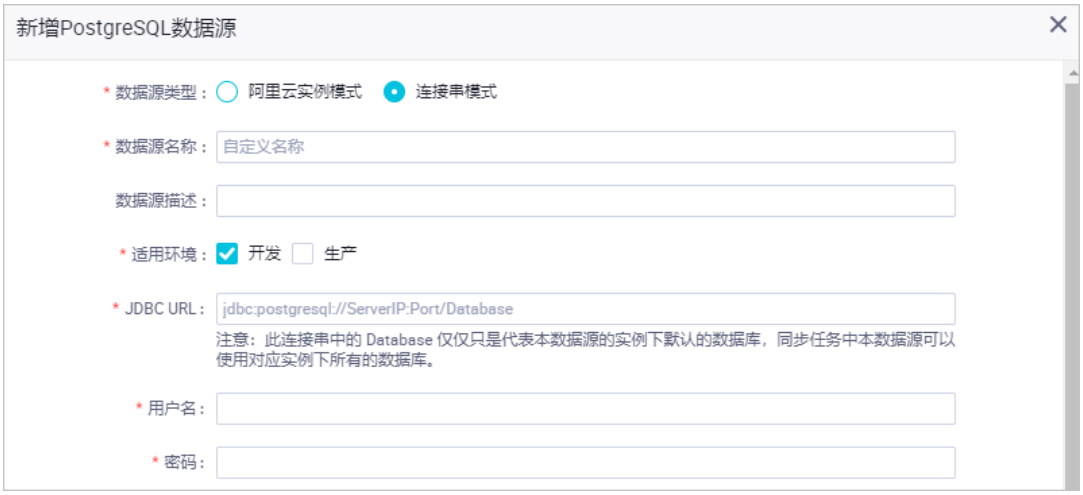
      PostgreSQL数据源包括**阿里云实例模式**和**连接串模式**两种类型。

- 以新增**阿里云实例模式**类型的数据源为例，配置数据源的基本信息。



参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	选择购买实例的地域。
RDS实例ID	您可以进入RDS控制台，查看RDS实例ID。
RDS实例主账号ID	实例购买者登录DataWorks控制台，鼠标悬停至右上角的用户头像，单击**安全设置**，查看账号ID。
默认数据库名	此处配置的是该数据源对应的默认数据库名称。后续配置同步任务的说明如下： ■ 配置整库同步（包含实时和离线）或同步解决方案任务时，您可以选择相应RDS实例下所有具有权限的数据库。 ■ 配置离线同步任务，当您选择使用多个数据库时，则每个数据库均需要配置一个数据源。
用户名	登录数据库的用户名称。
密码	登录数据库的密码。密码中避免使用@符号。

- 以新增**连接串模式**类型的数据源为例，配置数据源的基本信息。



参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:postgresql://ServerIP:Port/Database`。此连接串中的Database为本数据源的默认数据库，但在配置同步任务时，您可以使用相应RDS实例下所有的数据库。
用户名	登录数据库的用户名。
密码	登录数据库的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置PostgreSQL数据源，您可以继续下一个教程。在该教程中，您将学习如何配置PostgreSQL插件。详情请参见PostgreSQL Reader和PostgreSQL Writer。

6.1.16. 配置Redis数据源

Redis数据源为您提供读取和写入Redis双向通道的功能，您可以通过脚本模式配置同步任务。

背景信息

Redis是文档型的NoSQL数据库，为您提供持久化的内存数据库服务。Redis基于高可靠双机热备架构和可以无缝扩展的集群架构，满足高读写性能场景，以及容量需要弹性变化的业务需求。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Redis**。
4. 在**新增Redis数据源**对话框中，配置各项参数。

   Redis数据源类型包括**阿里云实例模式**和**连接串模式**两种类型：

   - 以新增**Redis > 阿里云实例模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	购买Redis时选择的区域。
Redis实例ID	您可以进入Redis控制台，查看Redis实例ID。
Redis访问密码	Redis Server的访问密码，如果没有则不填。

- 以新增**Redis > JDBC连接串模式**类型的数据源为例。

参数	描述

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。 选择该类型的数据源需要使用自定义调度资源才能进行同步，您可以单击**帮助手册**查看详情。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
服务器地址	格式为 `host:port` 。
添加服务器地址	单击**添加服务器地址**，即可添加格式为 `host:port` 的服务器地址。
Redis访问密码	Redis的服务访问密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Redis数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Redis Writer插件，详情请参见Redis Writer。

6.1.17. 配置HybridDB for MySQL数据源

HybridDB for MySQL数据源为您提供读取和写入HybridDB for MySQL的双向功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。

3. 在**新增数据源**对话框中，选择数据源类型为**HybridDB for MySQL**。
4. 在**新增HybridDB for MySQL数据源**对话框中，配置各项参数。

   HybridDB for MySQL数据源包括**阿里云数据库（AnalyticDB）**和**连接串模式**两种类型：

   - 以新增**HybridDB for MySQL > 阿里云数据库（AnalyticDB）**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**HybridDB for MySQL > 阿里云数据库（AnalyticDB）**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对新建的数据源进行简单描述。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
实例ID	您可以进入HybridDB for MySQL控制台，查看实例ID。
主账号ID	实例购买者登录控制台，进入**安全设置**页面查看实例账号ID。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**HybridDB for MySQL > 连接串模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**HybridDB for MySQL > 连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。
用户名	输入数据库对应的用户名。
密码	输入数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置HybridDB for MySQL数据源，您可以继续下一个教程。在该教程中，您将学习如何配置HybridDB for MySQL插件，详情请参见HybridDB for MySQL Reader和HybridDB for MySQL Writer。

6.1.18. 配置AnalyticDB for PostgreSQL数据源

AnalyticDB for PostgreSQL数据源提供读取和写入AnalyticDB for PostgreSQL的双向功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

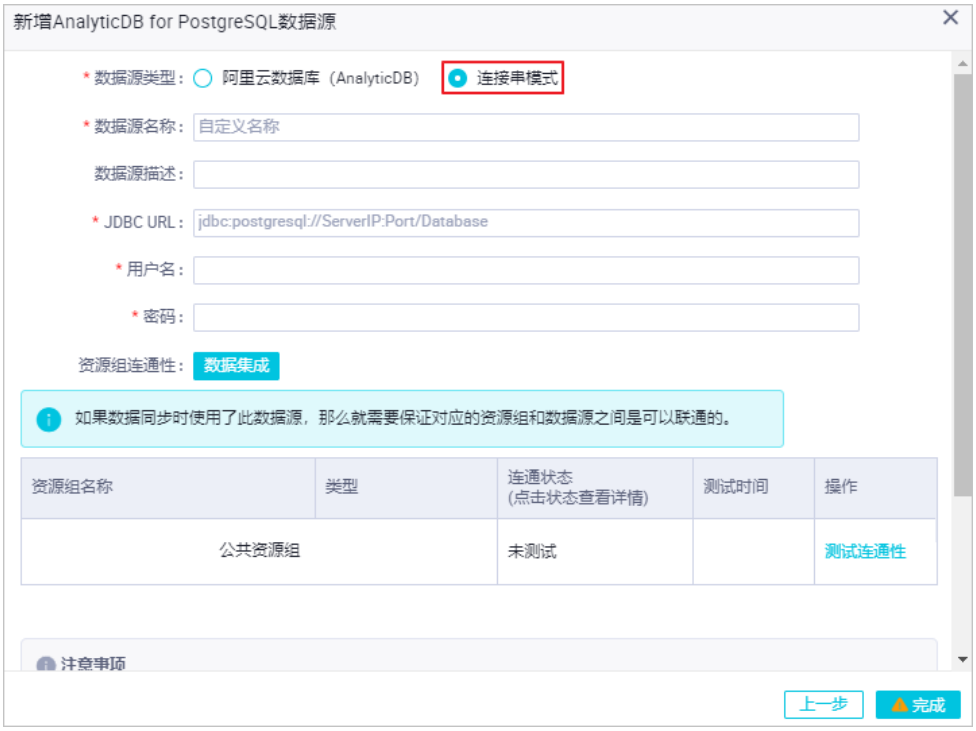
操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**AnalyticDB for PostgreSQL**。
4. 在**新增AnalyticDB for PostgreSQL数据源**对话框中，配置各项参数。

   AnalyticDB for PostgreSQL数据源包括**阿里云数据库（AnalyticDB）**和**连接串模式**两种类型：
   - 以新增**AnalyticDB for PostgreSQL > 阿里云数据库（AnalyticDB）**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**AnalyticDB for PostgreSQL > 阿里云数据库（AnalyticDB）**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
实例ID	您可以进入AnalyticDB for PostgreSQL的控制台，查看相应的实例ID。
主账号ID	您可以进入AnalyticDB for PostgreSQL控制台的**安全设置**页面，查看相应的信息。
数据库名	该数据源对应的数据库名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**AnalyticDB for PostgreSQL > 连接串模式**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**AnalyticDB for PostgreSQL > 连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:postgresql://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置AnalyticDB for PostgreSQL数据源，您可以继续下一个教程。在该教程中，您将学习如何配置AnalyticDB for PostgreSQL插件，详情请参见AnalyticDB for PostgreSQL Reader和AnalyticDB for PostgreSQL Writer。

6.1.19. 配置PolarDB数据源

PolarDB数据源为您提供读取和写入PolarDB双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

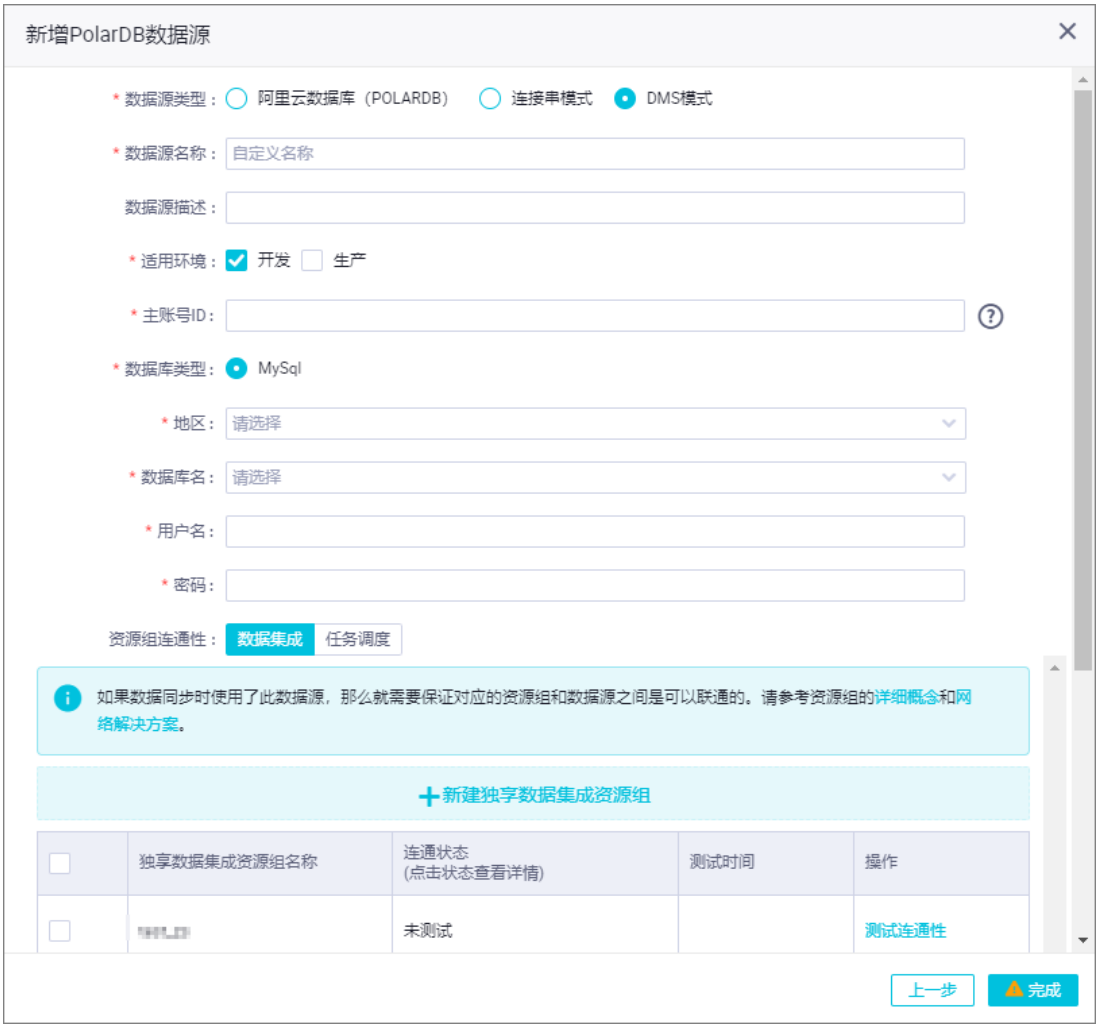
1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**PolarDB**。
4. 在**新增PolarDB数据源**对话框中，配置各项参数。

   PolarDB数据源包括**阿里云数据库（PolarDB）**、**连接串模式**、**DMS模式**三种类型：

   - 以新增**PolarDB > 阿里云数据库（PolarDB）**类型的数据源为例。

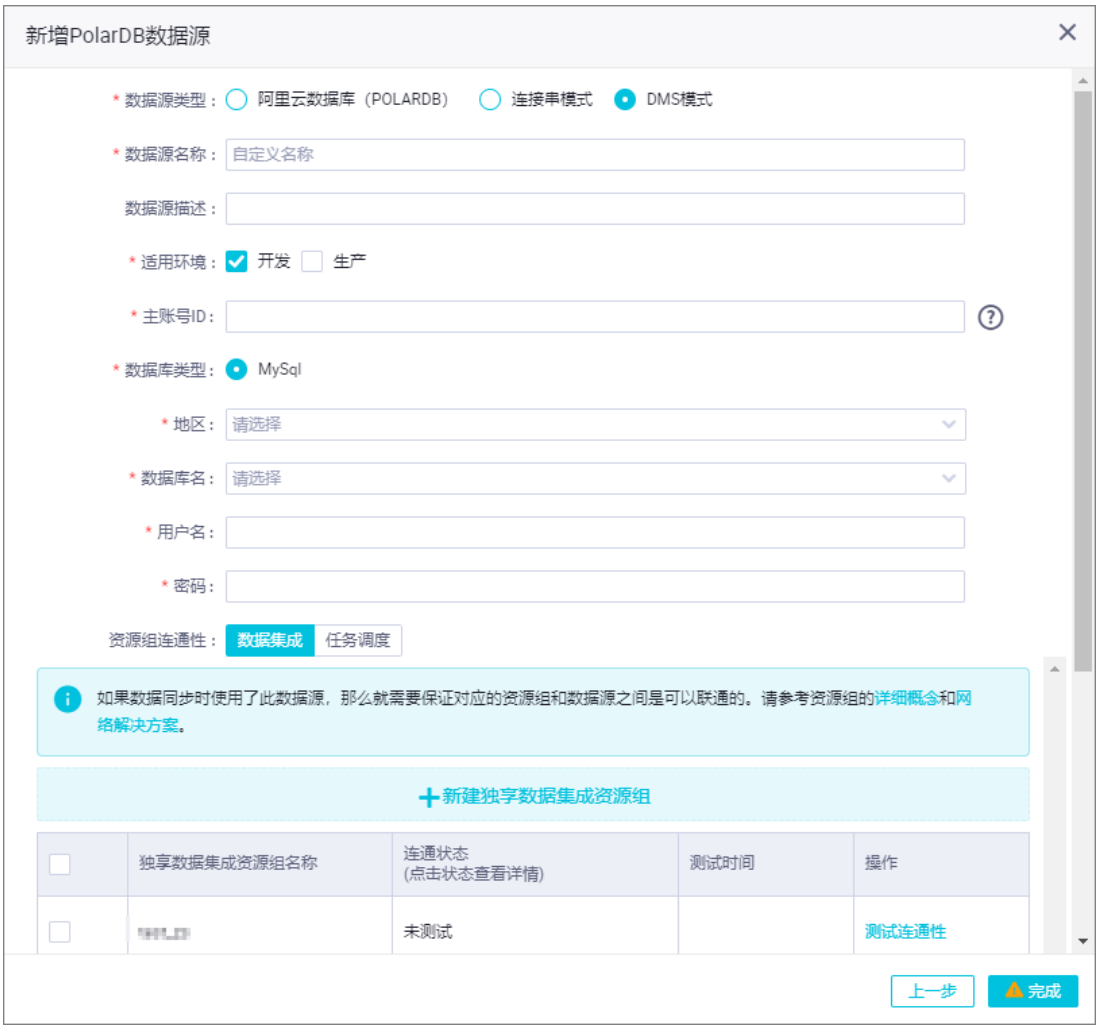
参数	描述
数据源类型	当前选择的数据源类型为**阿里云数据库（PolarDB）**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
集群ID	您可以进入PolarDB控制台，查看集群ID。
PolarDB实例主账号	实例购买者登录控制台，进入**安全设置**页面，即可查看实例账号ID。
数据库名	该数据源对应的数据库名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**PolarDB > 连接串模式**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**JDBC连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
数据库类型	包括**MySQL**和**PostgreSQL**。
JDBC URL	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**PolarDB > DMS模式**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**DMS模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
主账号ID	实例购买者登录控制台，进入**安全设置**页面查看实例账号ID。
数据库类型	默认为**MySQL**。
地区	在下拉列表选择数据源所对应的地域。
数据库名	在下拉列表选择对应的数据库。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置PolarDB数据源，您可以继续下一个教程。在该教程中，您将学习如何配置PolarDB插件。详情请参见PolarDB Reader和 PolarDB Writer。

6.1.20. 配置AnalyticDB for MySQL 3.0数据源

AnalyticDB MySQL 3.0数据源为您提供读取和写入AnalyticDB MySQL 3.0双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。

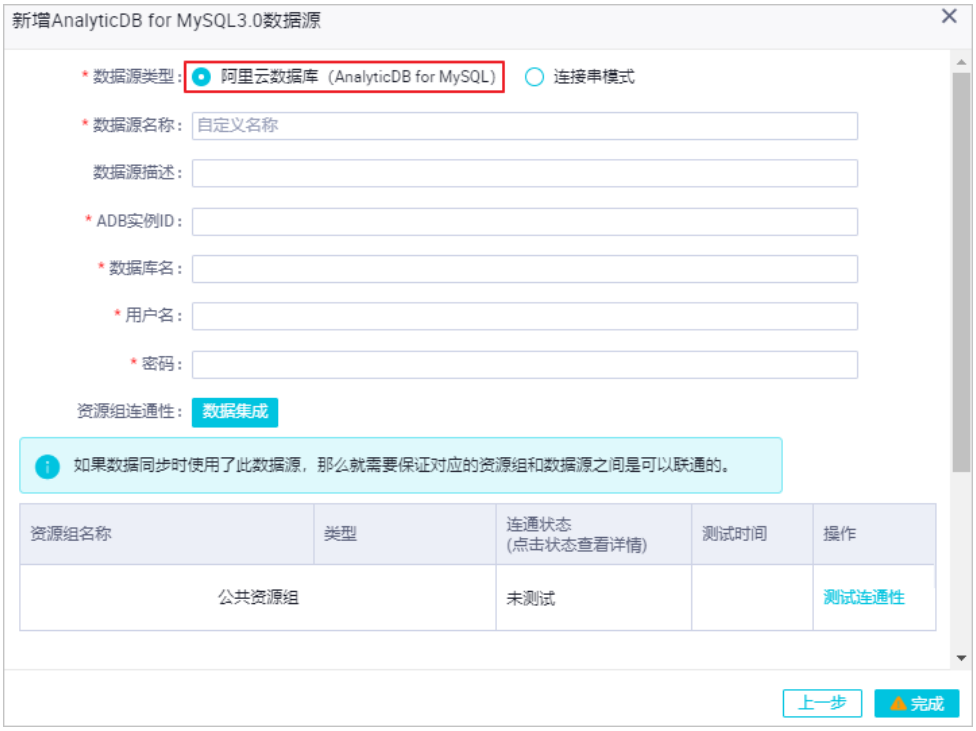
iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。

iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。

2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**AnalyticDB MySQL 3.0**。
4. 在**新增AnalyticDB MySQL 3.0数据源**对话框中，配置各项参数。

   数据源类型包括**阿里云数据库（AnalyticDB for MySQL）**和**连接串模式**两种类型，您可以根据自身需求进行选择。

   - 以新增**AnalyticDB MySQL 3.0 > 阿里云数据库（AnalyticDB for MySQL）**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**AnalyticDB MySQL 3.0 > 阿里云数据库（AnalyticDB for MySQL）**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地域	选择购买实例的地域。
ADB实例ID	您可以进入AnalyticDB MySQL 3.0控制台，查看实例ID。
数据库名	该数据源对应的数据库名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

   - 以新增**AnalyticDB MySQL 3.0 > 连接串模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**AnalyticDB MySQL 3.0 > 连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。
7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置AnalyticDB MySQL 3.0数据源，您可以继续下一个教程。在该教程中，您将学习如何配置AnalyticDB MySQL 3.0插件。详情请参见AnalyticDB MySQL 3.0 Reader和AnalyticDB MySQL 3.0 Writer。

6.1.21. 配置ClickHouse数据源

ClickHouse数据源为您提供读取和写入ClickHouse双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

> **说明** 仅支持阿里云ClickHouse。

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**ClickHouse**。
4. 在**新增ClickHouse数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。

参数	描述
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:clickhouse://ServerIP:Port/Database` 。 **说明** Port只支持HTTP协议的8123端口。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。
7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置ClickHouse数据源，您可以继续下一个教程。在该教程中，您将学习如何配置ClickHouse插件。详情请参见ClickHouse Reader和ClickHouse Writer。

6.1.22. 配置Data Lake Analytics（DLA）数据源

本文为您介绍如何配置Data Lake Analytics（DLA）数据源。

背景信息

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Data Lake Analytics（DLA）**。
4. 在**新增Data Lake Analytics（DLA）数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
连接Url	格式为 `Address:Port` 。
数据库	该数据源对应的数据库名称。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

6.1.23. 配置MaxCompute数据源

MaxCompute数据源作为数据中枢，为您提供读取和写入MaxCompute双向通道的功能。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

大数据计算服务MaxCompute（原名ODPS）提供完善的数据导入方案，能够更快速地解决海量数据计算问题。绑定MaxCompute实例的工作空间，系统会自动生成一个默认的数据源（odps_first）。同时，新绑定一个计算引擎实例，便会产生一个默认的计算引擎数据源，命名格式为0_regionId_引擎名称。

默认的数据源和计算引擎数据源对应的MaxCompute项目名称，均为当前工作空间对应的计算引擎MaxCompute项目名称。您可以单击右上方的用户信息，在修改AccessKey信息页面切换默认数据源的访问密钥，但需要注意以下问题：

- 仅支持阿里云账号间切换访问密钥。
- 切换时必须没有任务在运行中（数据集成或数据开发等一切和DataWorks相关的任务），您自行添加的MaxCompute数据源可以使用子账号访问密钥。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**MaxCompute（ODPS）**。
4. 在**新增MaxCompute（ODPS）数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
ODPS Endpoint	默认只读，从系统配置中自动读取。
Tunnel Endpoint	MaxCompute Tunnel服务的连接地址，详情请参见Endpoint。
ODPS项目名称	MaxCompute（ODPS）项目名称。
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AccessKey Secret	访问密钥中的AccessKey Secret，相当于登录密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试**

连通性。详情请参见配置资源组与网络连通。

> **说明**
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置MaxCompute数据源，您可以继续下一个教程。在该教程中，您将学习如何配置MaxCompute插件。详情请参见MaxCompute Reader和MaxCompute Writer。

6.1.24. 配置Hive数据源

Hive数据源为您提供读取和写入Hive双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

当底层存储为OSS时，请注意以下问题：

- defaultFS的配置请以oss://为前缀。例如，`oss://IP:PORT`或`oss://nameservice`。
- 您需要在高级参数中配置连接OSS服务时需要的参数，示例如下。

```
{
        "hiveConfig":{
            "fs.oss.accessKeyId":"<yourAccessKeyId>",
                "fs.oss.accessKeySecret":"<yourAccessKeySecret>",
                "fs.oss.endpoint":"oss-cn-<yourRegion>-internal.aliyuncs.com"
        }
    }
```

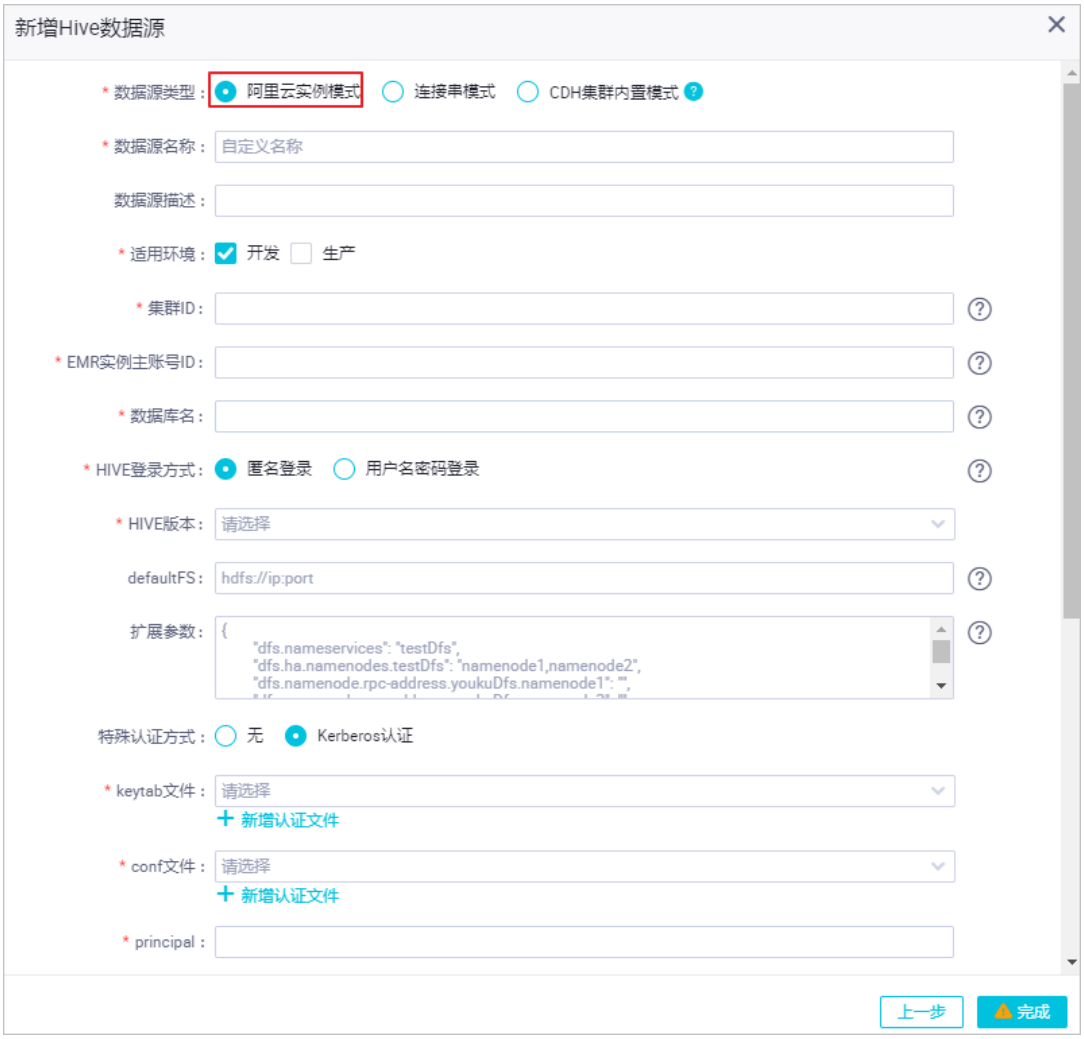
使用限制

- Hive数据源仅支持使用独享数据集成资源组。配置独享数据集成资源组详情请参见新增和使用独享数据集成资源组。
- 访问Hive数据源时目前仅支持使用Kerberos身份认证方式，如果访问数据源时不需要进行身份认证，则新增数据源时**特殊认证方式**选择**无**。

新建Hive数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Hive**。
4. 在**新增Hive数据源**对话框中，配置各项参数。

   Hive数据源包括**阿里云实例模式**，**连接串模式**，**CDH集群内置模式**三种类型：
   - 以新增**Hive > 阿里云实例模式**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	选择相应的地域。
集群ID	您可以登录EMR控制台，查看集群ID。
EMR实例主账号ID	实例购买者登录控制台，进入**安全设置**页面查看实例主账号ID。
数据库名	数据库的名称。
Hive登录方式	包括**用户名密码登录**和**匿名登录**。 如果您选择**用户名密码登录**，需要输入**HIVE用户名**和**HIVE密码**。
HIVE版本	选择需要使用的Hive版本。

参数	描述
defaultFS	Hadoop HDFS文件系统处于action状态的namenode节点地址。格式为 `hdfs://ip:port` 。
扩展参数	Hive高级参数配置，例如HA的相关配置，示例如下。 `"hadoopConfig":{` `"dfs.nameservices": "testDfs",` `"dfs.ha.namenodes.testDfs": "namenode1,namenode2",` `"dfs.namenode.rpc-address.youkuDfs.namenode1": "",` `"dfs.namenode.rpc-address.youkuDfs.namenode2": "",` `"dfs.client.failover.proxy.provider.testDfs` `"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider"` `}`
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

- 以新增**Hive > 连接串模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
HIVE JDBC URL	Hive元数据库的JDBC URL。 如果您选择kerberos认证方式，您需要在**HIVE JDBC URL**配置中拼接**principal**，例如：`jdbc:hive2://***.**.*.***:10000/default;principal=hive/**@**.***`。
数据库名	访问的Hive数据库的名称。您可以在Hive客户端执行命令`show databases`，查看已经创建的数据库。
Hive登录方式	包括**用户名密码登录**和**匿名登录**。 如果您选择**用户名密码登录**，需要输入**HIVE用户名**和**HIVE密码**。
HIVE版本	选择需要使用的Hive版本。

参数	描述
metastoreUris	格式为 `thrift://ip1:port1,thrift://ip2:por2` 。
defaultFS	Hadoop HDFS文件系统处于action状态的namenode节点地址。格式为 `hdfs://ip:port` 。
扩展参数	Hive高级参数配置，例如HA的相关配置，示例如下。 `"hadoopConfig":{` `"dfs.nameservices": "testDfs",` `"dfs.ha.namenodes.testDfs": "namenode1,namenode2",` `"dfs.namenode.rpc-address.youkuDfs.namenode1": "",` `"dfs.namenode.rpc-address.youkuDfs.namenode2": "",` `"dfs.client.failover.proxy.provider.testDfs` `"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider"` `}`
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。 如果您选择kerberos认证方式，您需要在**HIVE JDBC URL**配置中拼接**principal**，例如： `jdbc:hive2://***.**.*.***:10000/default;principal=hive/**@**.***` 。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

- 以新增**Hive > CDH集群内置模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**CDH集群内置模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
选择CDH集群	选择需要使用的CDH集群。
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。

参数	描述
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

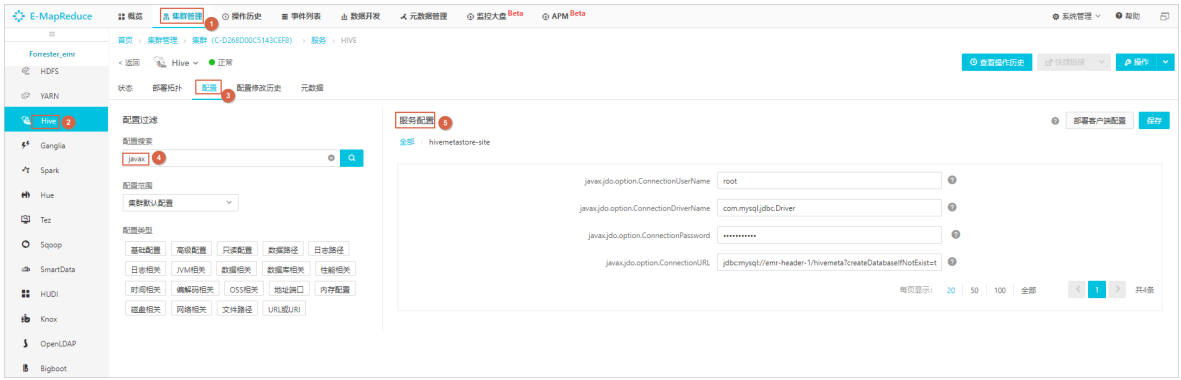
   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明**
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

在EMR控制台获取Hive配置

1. 登录E-MapReduce控制台。
2. 在顶部菜单栏，单击**集群管理**。
3. 单击相应集群后的**详情**，进入**集群基础信息**页面。
4. 在左侧导航栏，单击**集群服务 > Hive**。
5. 单击**配置**页签。
6. 在**配置搜索**下输入javax，单击图标，查看**服务配置**。



6.1.25. 配置GBase8a数据源

GBase8a数据源为您提供读取和写入GBase8a双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源，并进行隔离，以保护您的数据安全。

> **注意** GBase8a数据源仅支持使用新增和使用独享数据集成资源组和自定义资源组，不支持使用公共资源组。

操作步骤

1. 进入**数据源管理**页面。

i. 登录DataWorks控制台。
ii. 在左侧导航栏，单击**工作空间列表**。
iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。

2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**GBase8a**。
4. 在**新增GBase8a数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
连接Url	JDBC连接信息，格式为 `jdbc:mysql://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明**
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置GBase8a数据源，您可以继续下一个教程。在该教程中，您将学习如何配置GBase8a插件，详情请参见Gbase8a Reader和Gbase8a Writer。

6.1.26. 配置Hologres数据源

Hologres数据源为您提供读取和写入Hologres双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源开发和生产环境隔离功能，您可以分别添加开发环境和生产环境的数据源并进行隔离，以保护您的数据安全。

> **注意** 目前Hologres数据源仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Hologres**。
4. 在**新增Hologres数据源**对话框中，配置各项参数。

参数	描述
数据源类型	目前仅支持**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
实例ID	输入要同步的Hologres实例ID，您可以进入Hologres引擎管理页面进行查看。
数据库名	输入Hologres的数据库名称。
AccessKey ID	访问密钥中的AccessKey ID，您可以进入用户信息管理页面进行复制。
AccessKey Secret	访问密钥AccessKey Secret，相当于登录密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试**

连通性。详情请参见配置资源组与网络连通。

> **说明**
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Hologres数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Hologres插件。详情请参见Hologres Reader和Hologres Writer。

6.1.27. 配置HBase数据源

HBase数据源为您提供读取和写入HBase双向通道的功能，您可以通过脚本模式配置同步任务。

使用限制

- HBase数据源目前仅支持使用Kerberos身份认证方式（后续会逐步支持其他认证方式，敬请期待）。
- 不同网络连通条件下，HBase各版本的数据同步功能支持情况如下：

版本	独享资源组与数据源通过公网连通	独享资源组与数据源通过VPC内网连通
单机版（0.94.x）	支持	支持
标准版（1.1和2.0）	不支持	支持
增强版	支持	支持

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**HBase**。
4. 在**新增HBase数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。

参数	描述
配置信息	HBase集群提供给客户端连接的配置信息。 您可以转换hbase-site.xml为JSON格式，并补充scan的cache、batch等更多HBase客户端配置，以优化集群与客户端的交互。 根据使用的HBase版本，您需要配置不同的配置信息： **说明** 当前支持的HBase版本包含单机版（0.94.x）、标准版（1.1和2.0）和增强版。各版本的详细介绍可参见HBase版本选择。 ○ 如果您使用的是HBase单机版或标准版时，则使用默认的配置信息，您只需要输入对应的ZK信息。 `{ "hbase.rootdir": "hdfs://localhost:9000/hbase", "hbase.zookeeper.quorum": "localhost" }` ○ 如果您使用的是HBase增强版时，则使用增强版特有的endpoint形式，不再使用zookeeper.quorum参数连接。 请手动修改HBase增强版（lindorm）数据源的配置信息，示例如下。 **配置信息栏填写：** `{ "hbase.client.connection.impl" : "com.alibaba.hbase.client.AliHBaseUEConnection", "hbase.client.endpoint" : "host:30020", "hbase.client.username" : "root", "hbase.client.password" : "root" }`
特殊认证方式	选择数据源是否需要进行身份认证。默认选择**无**。目前认证方式仅支持选择**Kerberos认证**。关于Kerberos认证详细介绍请参见附录：配置Kerberos认证。
keytab文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的keytab文件。 如果没有可用的keytab文件，请单击**新增认证文件**进行添加。
conf文件	如果**特殊认证方式**选择为**Kerberos认证**，请选择需要使用的conf文件。 如果没有可用的conf文件，请单击**新增认证文件**进行添加。
principal	填写Kerberos认证的主体，即Kerberos账户，格式为：主名称/实例名称@领域名。例如****/hadoopclient@**.*** 。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在单击资源组列表下，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要在每种资源组上单独测试连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。详情请参见配置资源组与网络连通。

> 说明
> - 仅支持**独享数据集成资源组**测试连通性，详情请参见新增和使用独享数据集成资源组。
> - 如果您使用的是HBase增强版（lindorm），请忽略测试连通性时出现的找不到AliHBase类的报错。
> - 如果您使用的是HBase单机版（0.94.x），请忽略连通性失败问题，可以先执行数据同步任务。

7. 测试连通性通过后，单击**完成**。

> **说明** Hbase测试连通性会出现失败问题，目前后端服务已支持了多版本兼容。您需要手动在Hbase数据源的配置中，增加HbaseVersion的属性。示例如下：

```
{
"hbase.zookeeper.quorum": "my-zk:2181",
"hbaseVersion": "2.0.14"
}
```

后续步骤

现在，您已经学习了如何配置HBase数据源，您可以继续下一个教程。在该教程中，您将学习如何配置HBase插件。详情请参见HBase Reader和HBase Writer。

6.1.28. 配置Elasticsearch数据源

Elasticsearch数据源为您提供读取和写入Elasticsearch双向通道的功能，您可以通过脚本模式配置同步任务。

背景信息

DataWorks平台目前仅支持配置阿里云Elasticsearch5.x、6.x、7.x版本数据源，不支持配置自建Elasticsearch数据源。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Elasticsearch**。
4. 在**新增Elasticsearch数据源**对话框中，配置各项参数。

   Elasticsearch数据源包括**阿里云实例模式**和**连接串模式**两种类型：

- 以新增**阿里云实例模式**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地域	购买Elasticsearch时选择的地域。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

- 以新增**连接串模式**类型的数据源为例。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Endpoint	格式为 `http://esxxxx.elasticsearch.aliyuncs.com:9200` 。
安全验证	此处可以选择**用户密码登录**或**匿名登录**。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> **说明**
>
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Elasticsearch数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Elasticsearch插件。详情请参见Elasticsearch Reader和Elasticsearch Writer。

6.1.29. 配置Vertica数据源

Vertica数据源为您提供读取和写入Vertica双向通道的功能，您可以通过脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Vertica**。
4. 在**新增Vertica数据源**对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:vertica://ServerIp:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置Vertica数据源，您可以继续下一个教程。在该教程中，您将学习如何配置Vertica插件。详情请参见Vertica Reader和Vertica Writer。

6.1.30. 配置RestAPI数据源

当数据源支持通过API方式对接调用时，您可使用阿里云DataWorks的RestAPI（HTTP形式）的方式对接数据源，进行数据集成。本文为您介绍通过此种方式进行数据集成前，如何添加RestAPI（HTTP形式）类的数据源。

使用限制

目前该数据源仅支持新增和使用独享数据集成资源组。

操作步骤

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

如果您使用的DataWorks为标准模式，您需要参考以下操作步骤分别添加开发环境的数据源及生产环境的数据源。

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**RestAPI**。
4. 在新增RestAPI数据源对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
网络连接类型	根据对接数据源的网络环境选择合适的网络连接类型。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Url	填写RESTful请求地址。
默认请求头	每次请求传给该地址的header内容。

参数	描述
验证方法	使用RestAPI方式对接数据源时，支持以下三种验证方式：Basic Auth、Token Auth和Aliyun API Signature。您可以根据数据源API实际支持的验证方式选择对应的验证方式并配置验证参数。 ○ Basic Auth：基础验证。 如果数据源API支持用户名和密码的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的用户名和密码，后续数据集成过程中对接数据源时，通过Basic Auth协议传递给RESTful地址，完成验证。 ○ Token Auth：Token验证。 如果数据源API支持Token的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的固定Token值，后续数据集成过程中对接数据源时，通过传入header中进行验证，例如：{“Authorization”:” Bearer TokenXXXXXX”}。 ○ Aliyun API Signature：阿里云API签名验证。 如果数据源为阿里云产品，且此阿里云产品的API支持通过AccessKey和AccessSecret的方式进行验证，您可选择此种种验证方式，并在选择完成后配置用于验证的AccessKey和AccessSecret。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每种资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明** 仅支持新增和使用独享数据集成资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

完成添加数据源后，您可进行配置数据集成任务节点，当前支持通过向导模式和脚本模式进行配置。

6.1.31. 配置SAP HANA数据源

SAP HANA是一款支持企业预置型部署和云部署模式的内存计算平台，为您提供高性能的数据查询功能。SAPHANA数据源为您提供读取和写入SAP HANA双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

使用限制

目前该数据源仅支持新增和使用独享数据集成资源组。

操作步骤

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

如果您使用的DataWorks为标准模式，您需要参考以下操作步骤分别添加开发环境的数据源及生产环境的数据源。

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**SAP HANA**。
4. 在**新增SAP HANA数据**源对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
网络连接类型	根据对接数据源的网络环境选择合适的网络连接类型。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:sap://host:port?currentschema=SCHEMA` 。您可以在配置SAP HANA数据库时获取到相关信息。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每种资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明 仅支持新增和使用独享数据集成资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

完成添加数据源后，您可进行配置数据集成任务节点，当前支持通过向导模式和脚本模式进行配置。

6.1.32. 配置KingbaseES数据源

KingbaseES数据源为您提供读取和写入KingbaseES双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。本文为您介绍如何将KingbaseES数据源添加至DataWorks，为后续进行数据同步做好对接准备。

背景信息

人大金仓数据库管理系统KingbaseES（简称金仓数据库或KingbaseES）是北京人大金仓信息技术股份有限公司在国家“863”计划数据库重大专项和北京市科技计划重大项目支持下研发成功的具有自主知识产权的国产大型通用数据库管理系统（DBMS）。系统具有完整的大型通用数据库管理系统特征，提供完备的数据库管理功能，支持1000个以上并发用户、TB级数据量、GB级大对象。系统可运行于Windows、Linux、麒麟以及UNIX等多种操作系统平台，具有标准通用、稳定高效、安全可靠、兼容易用等特点。

使用限制

目前该数据源仅支持新增和使用独享数据集成资源组。

操作步骤

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

如果您使用的DataWorks为标准模式，您需要参考以下操作步骤分别添加开发环境的数据源及生产环境的数据源。

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**KingbaseES**。
4. 在**新增KingbaseES数据**源对话框中，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
网络连接类型	根据对接数据源的网络环境选择合适的网络连接类型。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:sap://host:port?currentschema=SCHEMA` 。您可以在配置SAP HANA数据库时获取到相关信息。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每种资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> 说明 仅支持新增和使用独享数据集成资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

完成添加数据源后，您可进行配置数据集成任务节点，当前支持通过向导模式和脚本模式进行配置。

6.1.33. 配置ApsaraDB for OceanBase数据源

ApsaraDB for OceanBase数据源提供读取和写入ApsaraDB for OceanBase数据的双向功能，您可以使用ApsaraDB for OceanBase数据源配置同步任务同步数据。本文为您介绍如何创建ApsaraDB for OceanBase数据源。

背景信息

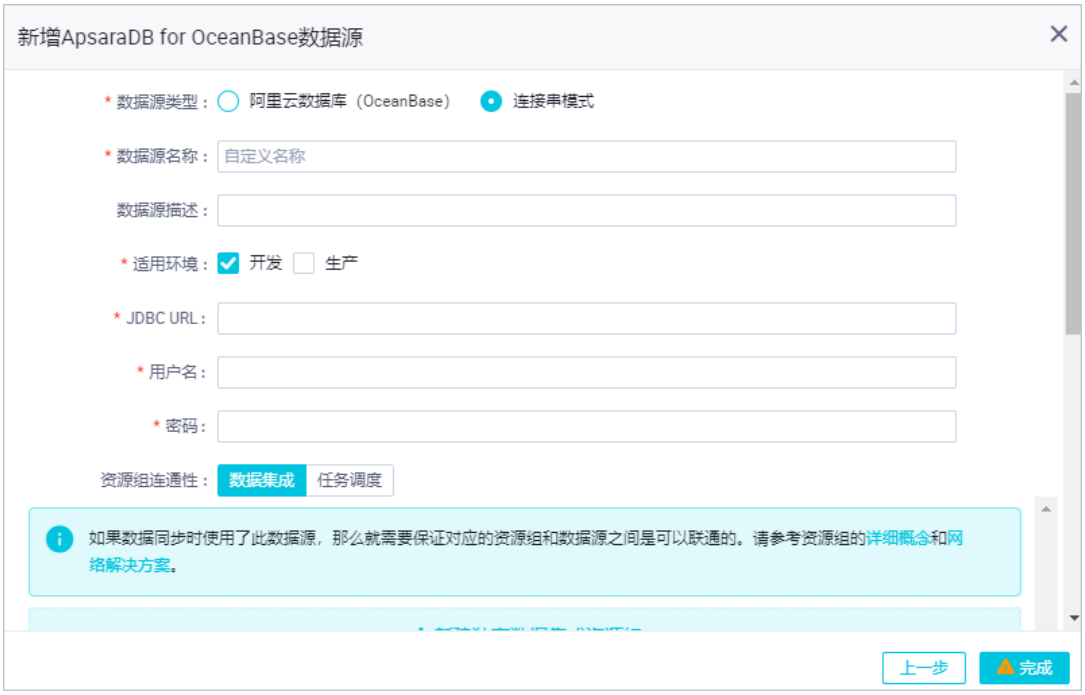
标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

新增数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**ApsaraDB for OceanBase**。
4. 在**新增ApsaraDB for OceanBase数据源**对话框中，配置各项参数。
   ApsaraDB for OceanBase数据源包括**阿里云数据库（OceanBase）**和**连接串模式**两种类型：
   - 以新增**ApsaraDB for OceanBase > 阿里云数据库（OceanBase）**类型的数据源为例。

参数	描述
数据源类型	当前选择的数据源类型为**ApsaraDB for OceanBase > 阿里云数据库（OceanBase）**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	新增数据源所在的地域。
集群	您可以进入OceanBase控制台的基本信息页面，查看相应的集群ID。
租户	您可以进入OceanBase控制台的基本信息页面，查看相应的租户ID。
主账号ID	您可以使用购买实例的账号登录OceanBase控制台，进入**安全设置**页面查看账号ID。
数据库名	该数据源对应的数据库名称。
用户名	登录数据库的用户名。
密码	登录数据库的密码。

- 以新增**ApsaraDB for OceanBase > 连接串模式**类型的数据源为例。



参数	描述
数据源类型	当前选择的数据源类型为**ApsaraDB for OceanBase > 阿里云数据库（OceanBase）**。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:oceanbase://ip:port/database` 。
用户名	登录数据库的用户名。
密码	登录数据库的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6.
7. 测试连通性通过后，单击**完成**。

后续步骤

您可以使用创建的数据源，执行数据集成同步任务，详情请参见数据集成概述。

6.1.34. 配置Kafka数据源

Kafka作为分布式消息队列，广泛应用于日志收集、监控数据聚合、流式数据处理、在线和离线分析等大数据领域，您可以使用Kafka数据源配置同步任务同步数据。本文为您介绍如何创建Kafka数据源。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

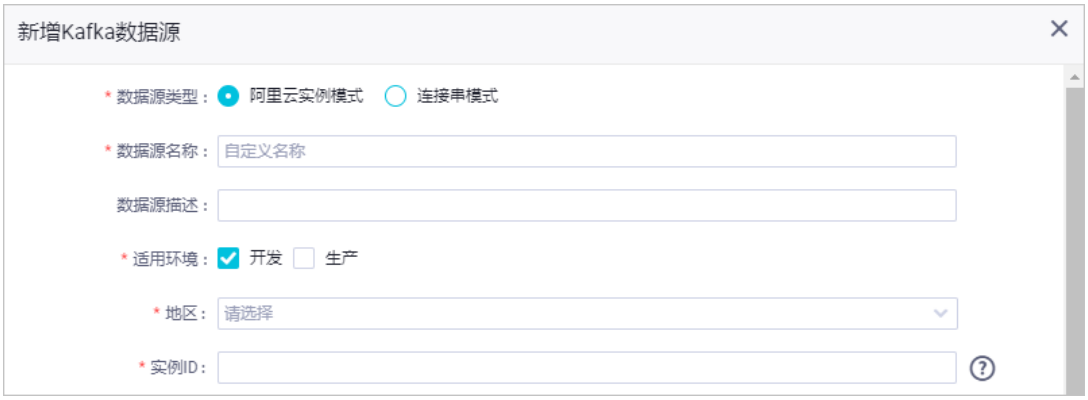
注意事项

支持阿里云Kafka，以及>=0.10.2且<=2.2.x的自建Kafka版本。

> **说明** 对于<0.10.2版本Kafka，由于Kafka不支持检索分区数据offset，并且Kafka数据结构可能不支持时间戳，进而无法支持数据同步。

新增数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**Kafka**。
4. 在**新增Kafka数据源**对话框中，配置各项参数。
   i. 配置数据源的基本信息。
   Kafka数据源包括**阿里云实例模式**和**连接串模式**两种类型。
   - 以新增**阿里云实例模式**类型的数据源为例，配置数据源的基本信息。



参数	描述
数据源类型	当前选择的数据源类型为**阿里云实例模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
地区	选择需要同步的Kafka实例所在的地域。
实例ID	输入需要同步的Kafka实例ID。您可以登录Kafka管控台，进入**实例列表**页面，获取实例ID。

- 以新增**连接串模式**类型的数据源为例，配置数据源的基本信息。

参数	描述
数据源类型	当前选择的数据源类型为**连接串模式**。
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
Kafka集群地址	Kafka集群Broker的IP地址和端口，即Kafka实例的**接入点**信息。您可以登录Kafka管控台，进入**实例列表**页面，单击实例名称，在实例详情页获取**接入点**信息。 多个Broker地址之间使用逗号（,）分隔，示例为 `10.0.0.1:9092,10.0.0.2:9092` 。

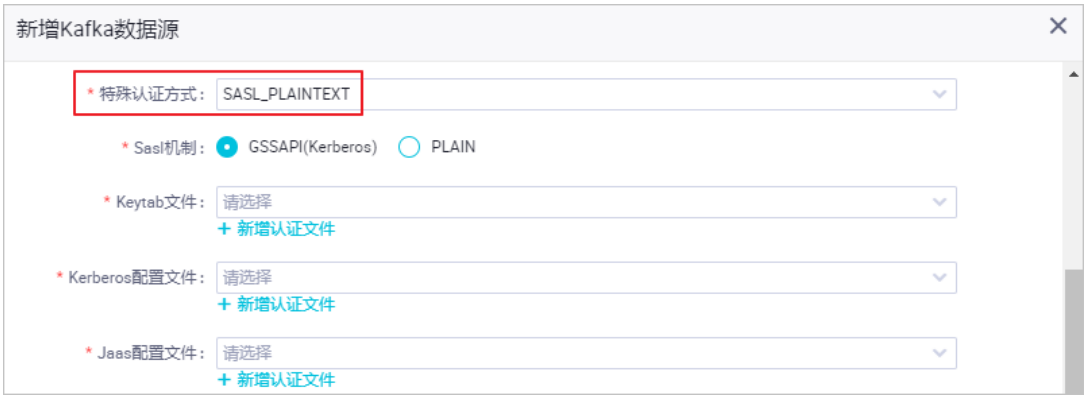
ii. 配置数据源的认证信息。

第三方认证机制用于用户和服务的强身份验证，通过该机制，可以有效的避免不受信任的程序或服务来获取数据访问权限，提高数据同步过程中访问数据资源的安全性。DataWorks在配置Kafka数据源时为您提供了三种第三方认证方式（即**特殊认证方式**配置为**SASL_PLAINTEXT**、**SASL_SSL**或**SSL**），开启第三方认证功能后，只有可信的应用和服务才能访问数据资源。

> **说明**
> - 使用第三方认证功能时，您需要提前在DataWorks的认证文件管理页面上传认证文件。上传和引用认证文件，详情请参见配置第三方身份认证。
> - 认证文件只对离线同步任务读取Kafka数据源时有效，实时同步如有需求，请提交工单。
> - 如果访问数据源时您无需对其他应用或服务进行认证，则可将**特殊认证方式**配置为None。
> - **阿里云实例模式**和**连接串模式**类型的数据源，认证方式的配置相同。

不同认证方式的配置如下：

- **SASL_PLAINTEXT**是一种简单的用户名和密码认证机制。使用**SASL_PLAINTEXT**方式进行认证的配置如下。

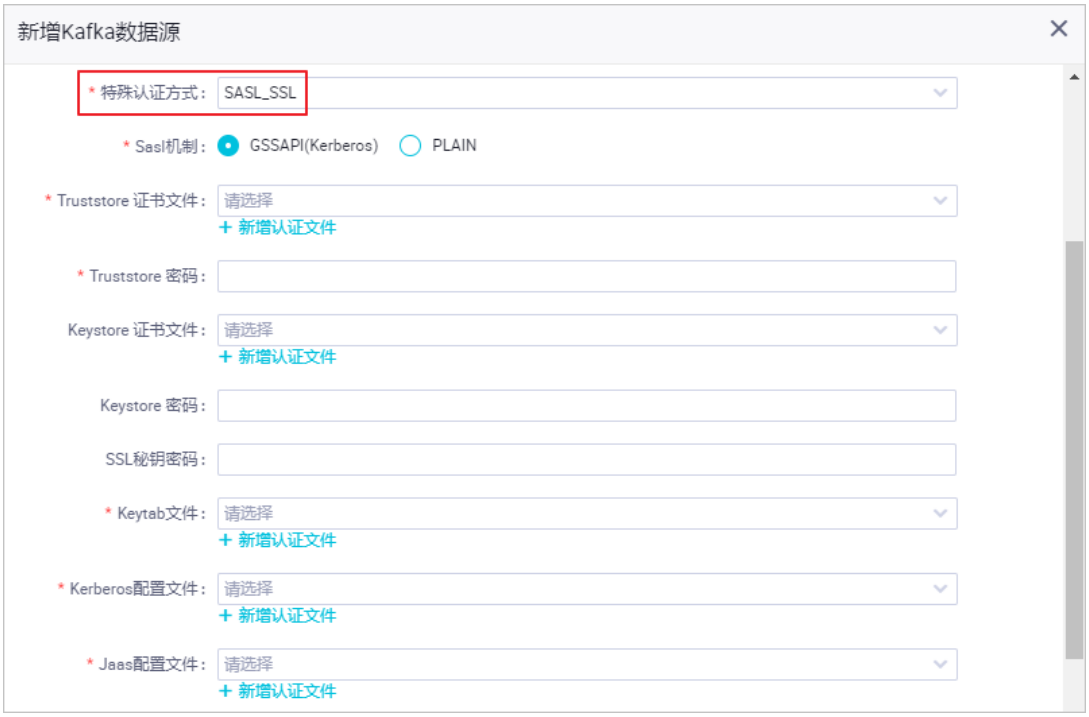


参数	描述
Sasl机制	目前支持使用**GSSAPI(Kerberos)**和**PLAIN**（指通过用户名和密码认证）两种简单验证和安全层Sasl（Simple Authentication and Security Layer）认证机制。
Keytab文件	密钥表文件，用于存储其他应用或服务的密钥信息，后续在**Jaas配置文件**中会引用该文件对其他应用或服务进行身份验证。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** 仅当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，需要配置该参数。

参数	描述
Kerberos配置文件	该文件包含密钥分发中心KDC（Key Distribute Center）的地址信息，用于设置Java的安全认证系统参数 `java.security.krb5.conf` 。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** ■ 仅当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，需要配置该参数。 ■ 典型的Kerberos配置文件格式如下： `[libdefaults]` `kdc_realm = EMR.232619.COM` `default_realm = EMR.232619.COM` `udp_preference_limit = 1` `kdc_tcp_port = 88` `kdc_udp_port = 88` `dns_lookup_kdc=false` `[realms]` `EMR.232619.COM = {` `kdc = 106.15.184.100:88` `}`

参数	描述
Jaas配置文件	该文件包含认证和授权信息，用于设置Java的安全鉴权系统参数 `java.security.auth.login.config` 。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** ■ 当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，数据集成功能在使用**Jaas配置文件**时会引用**Keytab文件**的密钥信息进行认证。 ■ 典型的JAAS配置文件格式如下： `KafkaClient { org.apache.kafka.common.security.plain.PlainLoginModule required username="Lisa" password="secret"; };` 注意`JaasContextName`务必指定为`KafkaClient`。

■ **SASL_SSL**是一种主要用于客户端到服务器端进行简单认证的认证方式。使用**SASL_SSL**方式进行认证的配置如下。

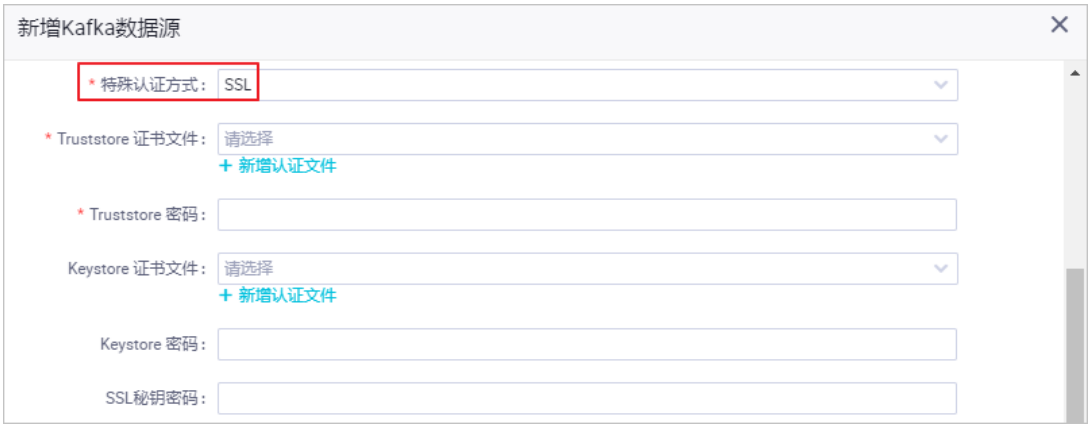


参数	描述
Sasl机制	目前支持使用**GSSAPI(Kerberos)**和**PLAIN**（指通过用户名和密码认证）两种简单验证和安全层Sasl（Simple Authentication and Security Layer）认证机制。

参数	描述
Truststore证书文件	存放Kafka集群CA证书的文件，用于访问SSL服务器时对该证书进行认证，以确保访问本数据源的应用或服务是可信任的。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** ■ CA证书是由电子商务认证中心CA（Certificate Authority）颁发的数字证书，用于验证访问源可信任性。 ■ 如果您使用的是阿里kafka实例，请参考SSL证书算法升级说明获取证书。
Truststore密码	读取Kafka集群CA证书内容时使用的密码。 **说明** 如果您使用的是阿里kafka实例，密码固定为KafkaOnsClient。
Keystore 证书文件	存放Kafka集群可信的CA证书和密钥的密钥库文件。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。
Keystore 密码	读取**Keystore 证书文件**内容的密码。
SSL密钥密码	读取**Keystore 证书文件**中指定密钥对的密码。
Keytab文件	密钥表文件，用于存储其他应用或服务的密钥信息，后续在**Jaas配置文件**中会引用该文件对其他应用或服务进行身份验证。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** 仅当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，需要配置该参数。

参数	描述
Kerberos配置文件	该文件包含密钥分发中心KDC（Key Distribute Center）的地址信息，用于设置Java的安全认证系统参数 `java.security.krb5.conf` 。您可以选择DataWorks中已上传的文件，也可以单击**新增认证文件**，上传新的认证文件。 **说明** ■ 仅当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，需要配置该参数。 ■ 典型的Kerberos配置文件格式如下： `[libdefaults]` `kdc_realm = EMR.232619.COM` `default_realm = EMR.232619.COM` `udp_preference_limit = 1` `kdc_tcp_port = 88` `kdc_udp_port = 88` `dns_lookup_kdc=false` `[realms]` `EMR.232619.COM = {` `kdc = 106.15.184.100:88` `}`
Jaas配置文件	该文件包含认证和授权信息，用于设置Java的安全鉴权系统参数 `java.security.auth.login.config` 。 **说明** ■ 当**Sasl机制**参数配置为**GSSAPI(Kerberos)**时，数据集成功能在使用**Jaas配置文件**时会引用**Keytab文件**的密钥信息进行认证。 ■ 典型的JAAS配置文件格式如下： `KafkaClient {` `org.apache.kafka.common.security.plain.PlainLoginModule required` `username="Lisa"` `password="secret";` `};` **注意**`JaasContextName`**务必指定为** `KafkaClient`。

- SSL主要用于客户端到服务器端的认证。使用SSL方式进行认证的配置如下。



参数	描述
Truststore证书文件	存放Kafka集群CA证书的文件，用于访问SSL服务器时对该证书进行认证，以确保访问本数据源的应用或服务是可信任的。 **说明** ■ CA证书是由电子商务认证中心CA（Certificate Authority）颁发的数字证书，用于验证访问源可信任性。 ■ 如果您使用的是阿里kafka实例，请参考SSL证书算法升级说明获取证书。
Truststore密码	读取Kafka集群CA证书内容时使用的密码。 **说明** 如果您使用的是阿里kafka实例，密码固定为KafkaOnsClient。
Keystore 证书文件	存放Kafka集群可信的CA证书和密钥的密钥库文件。
Keystore 密码	读取**Keystore 证书文件**内容的密码。
SSL密钥密码	读取**Keystore 证书文件**中指定密钥对的密码。

5. （可选）配置数据源的扩展参数。

   您可以选择为当前数据源配置扩展参数，即配置Kafka消费者和生产者的相关参数，格式为JSON格式。

   示例如下：

   - 配置发往每个分区（Partition）的消息缓存量（消息内容的字节数总和）为16342。

- 配置每条消息在缓存中的最长时间为10毫秒。

```
{
"batch.size":"16342",
"linger.ms":"10"
}
```

> **说明** 如果使用脚本模式配置的离线同步任务，或使用单表同步配置的实时同步任务中，配置的消费者或生产者参数，与**扩展参数**中配置的参数相同但取值不同，则**扩展参数**配置的参数优先级低于同步任务中配置的参数。

6. 测试数据源与资源组的连通性。
   i. 选择**资源组连通性**类型为**数据集成**。
   ii. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   > **说明**
   > - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   > - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

7. 测试连通性通过后，单击**完成**。

后续步骤

您可以使用创建的数据源，执行数据集成同步任务，详情请参见数据集成概述。

6.1.35. 配置DB2数据源

阿里云DataWorks的DB2数据源作为数据中枢，为您提供读取和写入DB2数据库的双向通道，在对接DB2数据库进行数据读写前，您需要先在DataWorks上配置好DB2数据源，配置完成后您可以通过向导模式或脚本模式配置同步任务，进行数据读写。本文为您介绍如何配置DB2数据源。

前提条件

已准备好用于对接DB2数据的数据集成资源组。

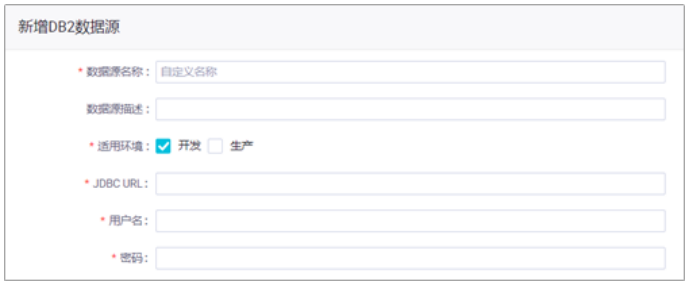
读写DB2数据库时，您需要使用数据集成资源组，并在配置数据源过程中调测数据集成资源组与数据源之间的网络连通性，保障所用的数据集成资源组与数据源间网络是连通的。建议您使用独享数据集成资源组，详情可见独享数据集成资源组概述。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**DB2**。
4. 在**新增DB2数据源**对话框，配置各项参数。



参数	描述
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	JDBC连接信息，格式为 `jdbc:db2://ServerIP:Port/Database` 。
用户名	数据库对应的用户名。
密码	数据库对应的密码。

5. 测试网络连通性。

i. 选择**数据集成**页签。

ii. 选择待测试的资源组。

- 建议您使用独享数据集成资源组，如果您还未购买独享数据集成资源组，可单击上方的**新建独享数据集成资源组**，根据界面提示快速购买。
- 如果您想使用自定义资源组，可单击**更多选项**，根据界面提示操作，然后相应的数据集成资源组。

iii. 单击相应资源组操作列的**测试连通性**。

数据同步时，一个任务只能使用一种资源组。如果您要测试每种资源组的连通性，以保证同步任务使用的数据集成资源组都能够与数据源连通，需要测试多个数据集成资源组与数据源间的网络连通性时，您可以多选后单击**批量测试连通性**。

6. 测试连通性通过后，单击**完成**。

后续步骤

完成数据源配置后，您可以开始配置数据集成任务，进行DB2数据源读写任务的配置。配置时会使用DataWorks提供的Reader和Writer插件，您需先学习如何配置DB2的Reader和Writer插件，详情请参见DB2 Reader和DB2 Writer。

6.1.36. 配置AWS S3数据源

Simple Storage Service（简称S3）是一种专为从任意位置存储和检索任意数量的数据而构建的对象存储，您可以使用S3数据源配置同步任务同步数据。本文为您介绍如何创建S3数据源。

前提条件

已准备好用于对接S3数据的数据集成资源组。

读取S3数据库时，您需要使用数据集成资源组，并在配置数据源过程中调测数据集成资源组与数据源之间的网络连通性，保障所用的数据集成资源组与数据源间网络是连通的。仅支持您使用独享数据集成资源组，详情可见独享数据集成资源组概述。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

使用限制

暂不支持大陆及中国香港地区的S3数据源。

新增数据源

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**进入数据集成**。
   iv. 在左侧导航栏，单击**数据源 > 数据源列表**，进入**工作空间管理 > 数据源管理**页面。
2. 在**数据源管理**页面，单击右上角的**新增数据源**。
3. 在**新增数据源**对话框中，选择数据源类型为**S3**。
4. 在**新增S3数据源**对话框中，配置各项参数。

i. 配置数据源的基本信息。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
Endpoint	S3的地址信息，格式为 `http://s3.ap-northeast-1.amazonaws.com` 。您可以在S3管控台查询。
Bucket	相应的S3 Bucket信息，指存储空间，是用于存储对象的容器。 您可以创建一个或多个存储空间，每个存储空间可添加一个或多个文件。 您可以在数据同步任务中查找此处输入的存储空间中相应的文件，没有添加的存储空间，则不能查找其中的文件。
AccessKey ID	访问密钥中的AccessKey ID。
AceessKey Secret	访问密钥中的AccessKey Secret，相当于登录密码。

5. 测试数据源与资源组的连通性。

i. 选择**资源组连通性**类型为**数据集成**。

ii. 在资源组列表，单击相应资源组后的**测试连通性**。

数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

> **说明**
>
> - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

6. 测试连通性通过后，单击**完成**。

后续步骤

现在，您已经学习了如何配置S3数据源，您可以继续下一个教程。在该教程中，您将学习如何配置S3插件。详情请参见AWS S3 Reader。

6.1.37. 配置StarRocks数据源

StarRocks数据源为您提供读取和写入StarRocks双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

背景信息

标准模式的工作空间支持数据源隔离功能，您可以分别添加并隔离开发环境和生产环境的数据源，以保护您的数据安全。详情请参见数据源开发和生产环境隔离。

操作步骤

1. 进入**数据源管理**页面。
   i. 登录DataWorks控制台。
   ii. 在左侧导航栏，单击**工作空间列表**。
   iii. 选择工作空间所在地域后，单击相应工作空间后的**数据集成**。
   iv. 在左侧导航栏，单击**数据源**，进入**数据源管理**页面。
2. 在**数据源管理**页面，单击**新增数据源**。
3. 在**新增数据源**对话框，选择数据源类型为**StarRocks**。
4. 在**新增StarRocks数据源**对话框，配置各项参数。

参数	描述
数据源名称	数据源名称必须以字母、数字、下划线（_）组合，且不能以数字和下划线（_）开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。
适用环境	可以选择**开发**或**生产**环境。 **说明** 仅标准模式工作空间会显示该配置。
JDBC URL	StarRocks连接信息，格式为 `jdbc:mysql://<ip>:<port>/<dbname>` 。连接串中填写FE IP、Mysql Port（一般默认是9030）。
Username	登录数据库的用户名称。
Password	登录数据库的密码。

5. 选择**资源组连通性**类型为**数据集成**。
6. 在资源组列表，单击相应资源组后的**测试连通性**。

   数据同步时，一个任务只能使用一种资源组。您需要测试每个资源组的连通性，以保证同步任务使用的数据集成资源组能够与数据源连通，否则将无法正常执行数据同步任务。如果您需要同时测试多种资源组，请选中相应资源组后，单击**批量测试连通性**。详情请参见配置资源组与网络连通。

   说明
   - （推荐）资源组列表默认仅显示独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
   - 如果您需要测试公共资源组或自定义资源组的连通性，请在资源组列表右下方，单击**更多选项**，在**警告**对话框单击**确定**，资源组列表会显示可供选择的公共资源组和自定义资源组。

后续步骤

现在，您已经学习了如何配置StarRocks数据源，您可以继续下一个教程。在该教程中，您将学习如何配置StarRocks插件，详情请参见StarRocks Reader和StarRocks Writer。

6.2. 配置Reader插件

6.2.1. DRDS Reader

DRDS Reader插件实现了从DRDS（分布式RDS）读取数据。本文为您介绍DRDS Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

目前DRDS的插件仅适配MySQL引擎的场景。DRDS是一套分布式MySQL数据库，并且大部分通信协议遵守MySQL使用场景。

> **注意** DRDS下的MySQL8.0版本仅支持使用独享数据集成资源组。

DRDS Reader通过JDBC连接器连接至远程的DRDS数据库，根据您配置的信息生成查询SQL语句，发送至远程DRDS数据库，执行该SQL语句并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

对于您配置的table、column、where等信息，DRDS Reader将其拼接为SQL语句发送至DRDS数据库。不同于普通的MySQL数据库，DRDS作为分布式数据库系统，无法适配所有MySQL的协议，包括复杂的Join等语句。

类型转换列表

DRDS Reader支持大部分DRDS类型，但也存在个别类型没有支持的情况，请注意检查您的数据类型 。

DRDS Reader针对DRDS类型的转换列表，如下所示。

类型分类	DRDS数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔类	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	所选取的需要同步的表。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息，默认使用所有列配置，例如[*]： • 支持列裁剪，即可以导出部分列。 • 支持列换序，即可以不根据表Schema信息的顺序导出列。 • 支持常量配置，您需要按照MySQL的语法格式。例如，`["id", "`table`", "1", "'bazhen.csy'", "null", "to_char(a + 1)", "2.3", "true"]`。示例中的参数说明如下： ○ id为普通列名。 ○ table包含保留的列名。 ○ 1为整型数字常量。 ○ bazhen.csy为字符串常量。 ○ null为空指针。 ○ to_char(a + 1)为计算字符串长度函数表达式。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示您指定同步的列集合，不允许为空 。	是	无
where	筛选条件，DRDS Reader根据指定的column、table、where条件拼接SQL，并根据该SQL进行数据抽取： • where条件可以有效地进行业务增量同步。 • where条件不配置或者为空时，视作全表同步数据 。 例如，在测试时使用where条件指定实际业务场景。通常会选择当天的数据进行同步，您可以指定where条件为`STRTODATE('${bdp.system.bizdate}','%Y%m%d') <= taday AND taday < DATEADD(STRTODATE('${bdp.system.bizdate}', '%Y%m%d'), interval 1 day)`。	否	无

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

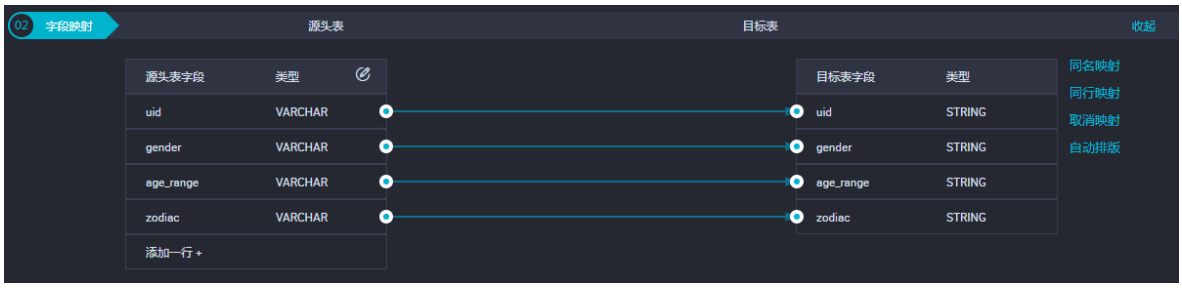
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段。鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其它空行会被忽略。
添加一行	添加一行的功能如下所示： ○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为**未识别**。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

参数	描述
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从DRDS数据库同步抽取数据作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        {
            "stepType":"drds",//插件名
            "parameter":{
                "datasource":"",//数据源名
                "column":[//列名
                    "id",
                    "name"
                ],
                "where":"",//过滤条件
                "table":"",//表名
                "splitPk": ""//切分键
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",//插件名
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1,//并发数
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}:"Writer"
            }
        ]
    }
}
```

使用说明

- 一致性视图问题

  DRDS本身属于分布式数据库，对外无法提供一致性的多库多表视图。不同于MySQL等单库单表同步，DRDS Reader无法抽取同一个时间切片的分库分表快照信息，即DRDS Reader抽取底层不同的分表将获取不同的分表快照，无法保证强一致性。

- 数据库编码问题

  DRDS本身的编码设置非常灵活，包括指定编码到库、表、字段级别，甚至可以设置不同编码。优先级从高到低为字段、表、库、实例。建议您在库级别将编码统一设置为UTF-8。

  DRDS Reader底层使用JDBC进行数据抽取，JDBC天然适配各类编码，并在底层进行了编码转换。因此DRDS Reader不需要您指定编码，可以自动获取编码并转码。

  对于DRDS底层写入编码和其设定的编码不一致的混乱情况，DRDS Reader对此无法识别，该类情况的导出结果有可能为乱码。

- 增量数据同步的方式

  DRDS Reader使用JDBC SELECT语句完成数据抽取工作，因此您可以使用`SELECT...WHERE...`进行增量数据抽取，方式如下：

  - 数据库在线应用写入数据库时，填充modify字段为更改时间戳，包括新增、更新、删除（逻辑删除）。对于这类应用，DRDS Reader只需要where条件后跟上一同步阶段时间戳即可。
  - 对于新增流水型数据，DRDS Reader在where条件后跟上一阶段最大自增ID即可。

  对于业务上无字段区分新增、修改数据的情况，DRDS Reader无法进行增量数据同步，只能同步全量数据。

- 不支持在where语句中配置物理表相关的筛选条件。

6.2.2. HBase Reader

HBase Reader插件实现了从HBase中读取数据，本文为您介绍HBase Reader支持的数据类型、字段映射和数据源等参数及配置示例。

在底层实现上，HBase Reader通过HBase的Java客户端连接远程HBase服务，并通过Scan方式读取您指定的rowkey范围内的数据，将读取的数据使用数据集成自定义的数据类型拼装为抽象的数据集，并传递给下游Writer处理。

使用限制

- HBase Reader不支持读取phoenix写入的数据，phoenix有特殊处理。
- HBase Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

支持的功能

- 支持HBase0.94.x、HBase1.1.x和HBase2.x版本
  - 如果您的HBase版本为HBase0.94.x，Reader端的插件请选择094x。

    ```
    "reader": {
            "plugin": "094x"
        }
    ```

  - 如果您的HBase版本为HBase1.1.x或HBase2.x，Reader端的插件请选择11x。

    ```
    "reader": {
            "plugin": "11x"
        }
    ```

    说明 HBase1.1.x插件当前可以兼容HBase 2.0，如果您在使用上遇到问题请提交工单。

- 支持normal和multiVersionFixedColumn模式

- normal模式：把HBase中的表当成普通二维表（横表）进行读取，获取最新版本数据。

```
hbase(main):017:0> scan 'users'
ROW                                   COLUMN+CELL
lisi                                  column=address:city, timestamp=1457101972764, value=beijing
lisi                                  column=address:contry, timestamp=1457102773908, value=china
lisi                                  column=address:province, timestamp=1457101972736, value=beijing
lisi                                  column=info:age, timestamp=1457101972548, value=27
lisi                                  column=info:birthday, timestamp=1457101972604, value=1987-06-17
lisi                                  column=info:company, timestamp=1457101972653, value=baidu
xiaoming                              column=address:city, timestamp=1457082196082, value=hangzhou
xiaoming                              column=address:contry, timestamp=1457082195729, value=china
xiaoming                              column=address:province, timestamp=1457082195773, value=zhejiang
xiaoming                              column=info:age, timestamp=1457082218735, value=29
xiaoming                              column=info:birthday, timestamp=1457082186830, value=1987-06-17
xiaoming                              column=info:company, timestamp=1457082189826, value=alibaba
2 row(s) in 0.0580 seconds }
```

读取后的数据如下所示。

rowKey	address:city	address:contry	address:province	info:age	info:birthday	info:company
lisi	beijing	china	beijing	27	1987-06-17	baidu
xiaoming	hangzhou	china	zhejiang	29	1987-06-17	alibaba

- multiVersionFixedColumn模式：把HBase中的表当成竖表进行读取。读出的每条记录是四列形式，依次为rowKey、family:qualifier、timestamp和value。读取时需要明确指定要读取的列，把每一个cell中的值，作为一条记录（record），若有多个版本则存在多条记录。

```
hbase(main):018:0> scan 'users',{VERSIONS=>5}
ROW                                   COLUMN+CELL
lisi                                  column=address:city, timestamp=1457101972764, value=beijing
lisi                                  column=address:contry, timestamp=1457102773908, value=china
lisi                                  column=address:province, timestamp=1457101972736, value=beijing
lisi                                  column=info:age, timestamp=1457101972548, value=27
lisi                                  column=info:birthday, timestamp=1457101972604, value=1987-06-17
lisi                                  column=info:company, timestamp=1457101972653, value=baidu
xiaoming                              column=address:city, timestamp=1457082196082, value=hangzhou
xiaoming                              column=address:contry, timestamp=1457082195729, value=china
xiaoming                              column=address:province, timestamp=1457082195773, value=zhejiang
xiaoming                              column=info:age, timestamp=1457082218735, value=29
xiaoming                              column=info:age, timestamp=1457082178630, value=24
xiaoming                              column=info:birthday, timestamp=1457082186830, value=1987-06-17
xiaoming                              column=info:company, timestamp=1457082189826, value=alibaba
2 row(s) in 0.0260 seconds }
```

读取后的数据（4列） 如下所示。

rowKey	column:qualifier	timestamp	value
lisi	address:city	1457101972764	beijing
lisi	address:contry	1457102773908	china
lisi	address:province	1457101972736	beijing
lisi	info:age	1457101972548	27
lisi	info:birthday	1457101972604	1987-06-17
lisi	info:company	1457101972653	beijing
xiaoming	address:city	1457082196082	hangzhou
xiaoming	address:contry	1457082195729	china
xiaoming	address:province	1457082195773	zhejiang
xiaoming	info:age	1457082218735	29
xiaoming	info:age	1457082178630	24
xiaoming	info:birthday	1457082186830	1987-06-17
xiaoming	info:company	1457082189826	alibaba

支持的数据类型

支持读取HBase数据类型及HBase Reader针对HBase类型的转换列表如下表所示。

类型分类	数据集成column配置类型	数据库数据类型
整数类	long	short、int和long
浮点类	double	float和double
字符串类	string	binary_string和string
日期时间类	date	date

类型分类	数据集成column配置类型	数据库数据类型
字节类	bytes	bytes
布尔类	boolean	boolean

参数说明

参数	描述	是否必选	默认值
haveKerberos	haveKerberos值为true时，表示HBase集群需要kerberos认证。 **说明** • 如果该值配置为true，必须要配置以下kerberos认证相关参数： ○ kerberosKeytabFilePath ○ kerberosPrincipal ○ hbaseMasterKerberosPrincipal ○ hbaseRegionserverKerberosPrincipal ○ hbaseRpcProtection • 如果HBase集群没有kerberos认证，则不需要配置以上参数。	否	*false*
hbaseConfig	连接HBase集群需要的配置信息，JSON格式。必填的配置为hbase.zookeeper.quorum，表示HBase的ZK链接地址。同时可以补充更多HBase client的配置，例如设置scan的cache、batch来优化与服务器的交互。 **说明** 如果是云HBase的数据库，需要使用内网地址连接访问。	是	无
mode	读取HBase的模式，支持normal模式和multiVersionFixedColumn模式。	是	无
table	读取的HBase表名（大小写敏感）。	是	无
encoding	编码方式，UTF-8或GBK，用于对二进制存储的HBase byte[]转为String时的编码。	否	*utf-8*

参数	描述	是否必选	默认值
column	要读取的HBase字段，normal模式与multiVersionFixedColumn模式下必填。 • normal模式下 name指定读取的HBase列，除rowkey外，必须为列族:列名的格式。type指定源数据的类型，format指定日期类型的格式。value指定当前类型为常量，不从HBase读取数据，而是根据value值自动生成对应的列。配置格式如下所示： `"column": [ { "name": "rowkey", "type": "string" }, { "value": "test", "type": "string" } ]` normal模式下，对于您指定的Column信息，type必须填写，name和value必须选择其一。 • multiVersionFixedColumn模式 name指定读取的HBase列，除rowkey外，必须为列族:列名的格式，type指定源数据的类型，format指定日期类型的格式。multiVersionFixedColumn模式下不支持常量列。配置格式如下所示： `"column": [ { "name": "rowkey", "type": "string" }, { "name": "info:age", "type": "string" } ]`	是	无
maxVersion	指定在多版本模式下的HBase Reader读取的版本数，取值只能为-1或大于1的数字，-1表示读取所有版本。	multiVersionFixedColumn模式下必填项	无

参数	描述	是否必选	默认值
range	指定HBase Reader读取的rowkey范围。 • startRowkey：指定开始rowkey。 • endRowkey：指定结束rowkey。 • isBinaryRowkey：指定配置的startRowkey和endRowkey转换为 byte[]时的方式，默认值为*false*。如果为true，则调用 `Bytes.toBytesBinary(rowkey)` 方法进行转换。如果为false，则调用 `Bytes.toBytes(rowkey)` 。配置格式如下所示： `"range": { "startRowkey": "aaa", "endRowkey": "ccc", "isBinaryRowkey":false }`	否	无
scanCacheSize	HBase Reader每次从HBase读取的行数和列数。	否	*256*
scanBatchSize	HBase Reader每次从HBase读取的行数和列数。	否	*100*

向导开发介绍

暂不支持向导开发模式开发。

脚本开发介绍

配置一个从HBase抽取数据到本地的作业（normal模式），使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"hbase",//插件名。
            "parameter":{
                "mode":"normal",//读取HBase的模式，支持normal模式、multiVersionFixedColumn模式。
                "scanCacheSize":"256",//HBase client每次RPC从服务器端读取的行数。
                "scanBatchSize":"100",//HBase client每次RPC从服务器端读取的列数。
                "hbaseVersion":"094x/11x",//HBase版本。
                "column":[//字段。
                    {
                        "name":"rowkey",//字段名。
                        "type":"string"//数据类型。
                    },
                    {
                        "name":"columnFamilyName1:columnName1",
                        "type":"string"
                    },
                    {
                        "name":"columnFamilyName2:columnName2",
                        "format":"yyyy-MM-dd",
                        "type":"date"
                    },
                    {
                        "name":"columnFamilyName3:columnName3",
                        "type":"long"
                    }
                ],
                "range":{//指定HBase Reader读取的rowkey范围。
                    "endRowkey":"",//指定结束rowkey。
                    "isBinaryRowkey":true,//指定配置的startRowkey和endRowkey转换为byte[]时的方式，默认值为false。
                    "startRowkey":""//指定开始rowkey。
                },
```

```
                "maxVersion":"",//指定在多版本模式下的HBase Reader读取的版本数。
                "encoding":"UTF-8",//编码格式。
                "table":"",//表名。
                "hbaseConfig":{//连接HBase集群需要的配置信息，JSON格式。
                    "hbase.zookeeper.quorum":"hostname",
                    "hbase.rootdir":"hdfs://ip:port/database",
                    "hbase.cluster.distributed":"true"
                }
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1,//作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.3. HBase20xsql Reader

HBase20xsql Reader插件实现了从Phoenix（Hbase中的SQL表）中读取数据，本文为您介绍HBase20xsql Reader支持的数据类型、参数说明及配置示例。

前提条件

配置HBase20xsql Reader插件前，请确认数据源已配置完成，详情请参见：配置HBase数据源。

实现原理

HBase20xsql Reader插件通过Phoenix轻客户端连接Phoenix的查询服务器（QueryServer），并根据用户配置信息生成SQL查询语句，然后发送至QueryServer读取HBase数据源中的数据，完成后将返回结果按照DataX自定义的数据类型拼装为抽象的数据集，最终传递给下游Writer处理。

使用限制

- 仅支持HBase2.x和Phoenix5.x版本。
- 切分表时仅支持对单个列进行切分，且该列必须是表主键。
- 按照作业的并发个数平均切分表时，仅支持以整型和字符串类型作为切分列。
- 表名、schema名及列名大小写敏感，需要与Phoenix表实际大小写保持一致。
- 仅支持通过Phoenix QueryServer读取数据，因此您的Phoenix必须启动QueryServer服务才能使用HBase20xsql Reader插件。

类型转换

HBase20xsql Reader支持大部分Phoenix类型，但也存在个别类型没有支持的情况，请注意检查你的类型。

HBase20xsql Reader针对Phoenix类型的转换列表，如下所示。

DataX内部类型	Phoenix数据类型
long	INTEGER、TINYINT、SMALLINT、BIGINT
double	FLOAT、DECIMAL、DOUBLE
string	CHAR、VARCHAR
date	DATE、TIME、TIMESTAMP
bytes	BINARY、VARBINARY
boolean	BOOLEAN

参数说明

参数	描述	是否必选	默认值
queryServerAddress	HBase20xsql Reader插件需要通过Phoenix轻客户端去连接Phoenix QueryServer，因此，您需要在此处填写对应的QueryServer地址。如果HBase增强版（Lindorm）用户需要透传user、password参数，可以在queryServerAddress后增加对应的可选属性。格式为：`http://127.0.0.1:8765;user=root;password=root`。	是	无
serialization	QueryServer使用的序列化协议。	否	PROTOBUF
table	所要读取的表名（大小写敏感）。	是	无
schema	表所在的schema。	否	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息，空值表示读取所有列。默认值为空值。	否	全部列
splitKey	读取表时对表进行切分，如果指定splitKey，表示您希望使用splitKey代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高了数据同步的效能。您可以选择两种不同的切分方式，如果splitPoint为空，默认根据方法一自动切分： • 方法一：根据splitKey找到最大值和最小值，然后按照指定的concurrent数平均切分。 **说明** 仅支持以整型和字符串类型作为切分列。 • 方式二：根据设置的splitPoint进行切分。然后按照指定的concurrent同步数据。	是	无
splitPoints	根据切分列的最大值和最小值切分时不能保证避免数据热点，因此，建议切分点根据Region的startkey和endkey进行设置，保证每个查询对应单个Region。	否	无
where	筛选条件，支持对表查询增加过滤条件。HBase20xsql Reader根据指定的column、table、where条件拼接SQL，并根据该SQL进行数据抽取。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。配置该项后，除queryserverAddress参数必须设置外，HBase20xsql Reader会直接忽略column、table、where和splitKey条件的配置，使用该项配置的内容对数据进行筛选。	否	无

向导开发介绍

暂不支持向导开发模式开发。

脚本开发介绍

脚本配置示例如下，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"hbase20xsql",//插件名。
            "parameter":{
                "queryServerAddress": "http://127.0.0.1:8765",  //填写连接Phoenix QueryServer地址。
                "serialization": "PROTOBUF",  //QueryServer序列化格式。
                "table": "TEST",    //读取表名。
                "column": ["ID", "NAME"],   //所要读取列名。
                "splitKey": "ID"    //切分列，必须是表主键。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1,//作业并发数。
            "mbps":"12"//限流。
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.4. HDFS Reader

HDFS Reader提供了读取分布式文件系统数据存储的能力。在底层实现上，HDFS Reader获取分布式文件系统上文件的数据，并转换为数据集成传输协议传递给Writer。

> 注意 HDFS Reader仅支持使用新增和使用独享数据集成资源组。

使用限制

目前不支持阿里云文件存储HDFS版。

背景信息

HDFS Reader实现了从Hadoop分布式文件系统HDFS中，读取文件数据并转为数据集成协议的功能 。

例如，TextFile是Hive建表时默认使用的存储格式，数据不进行压缩。本质上TextFile是以文本的形式将数据存放在HDFS中，对于数据集成而言，HDFS Reader在实现上与OSS Reader有很多相似之处。

ORCFile的全名是Optimized Row Columnar File，是对RCFile的优化，该文件格式可以提供一种高效的方法来存储Hive数据。HDFS Reader利用Hive提供的OrcSerde类，读取解析ORCFile文件的数据。

使用HDFS Reader时，请注意以下事项：

- 由于连通默认资源组到HDFS的网络链路比较复杂，建议您使用独享数据集成资源组完成数据同步任务。您需要确保您的独享数据集成资源组具备HDFS的namenode和datanode的网络访问能力。
- HDFS默认情况下，使用网络白名单进行数据安全。基于此种情况，建议您使用独享数据集成资源组完成针对HDFS的数据同步任务。
- 您通过脚本模式配置HDFS同步作业，并不依赖HDFS数据源网络连通性测试通过，针对此类错误您可以临时忽略。
- 数据集成同步进程以admin账号启动，您需要确保操作系统的admin账号具备访问相应HDFS文件的读写权限。

支持的功能

HDFS Reader支持以下功能：

- 支持TextFile、ORCFile、rcfile、sequence file、csv和parquet格式的文件，且要求文件内容存放的是一张逻辑意义上的二维表。
- 支持多种类型数据读取（使用String表示），支持列裁剪，支持列常量。
- 支持递归读取、支持正则表达式 `*` 和 `?` 。
- 支持ORCFile数据压缩，目前支持SNAPPY和ZLIB两种压缩方式。
- 支持SequenceFile数据压缩，目前支持lZO压缩方式。
- 多个File可以支持并发读取。
- CSV类型支持压缩格式有gzip、bz2、zip、lzo、lzo_deflate和snappy。
- 目前插件中Hive版本为1.1.1，Hadoop版本为2.7.1（Apache适配JDK1.6），在Hadoop 2.5.0、Hadoop 2.6.0和Hive 1.2.0测试环境中写入正常。

> **注意** HDFS Reader暂不支持单个File多线程并发读取，此处涉及到单个File内部切分算法。

支持的数据类型

由于这些文件表的元数据信息由Hive维护，并存放在Hive自己维护的元数据库（如MySQL）中。目前HDFS Reader不支持对Hive元数据的数据库进行访问查询，因此您在进行类型转换时，必须指定数据类型。

RCFile、ParquetFile、ORCFile、TextFile和SequenceFile中的类型，会默认转为数据集成支持的内部类型，如下表所示。

类型分类	数据集成column配置类型	Hive数据类型
整数类	long	tinyint、smallint、int和bigint
浮点类	double	float和double
字符串类	string	string、char、varchar、struct、map、array、union和binary
日期时间类	date	date和timestamp
布尔类	boolean	boolean

> **说明**
>
> - long：HDFS文件中的整型类型数据，例如*123456789*。
> - double：HDFS文件中的浮点类型数据，例如*3.1415*。
> - bool：HDFS文件中的布尔类型数据，例如*true*、*false*，不区分大小写。
> - date：HDFS文件中的时间类型数据，例如*2014-12-31 00:00:00*。

Hive支持的数据类型TIMESTAMP可以精确到纳秒级别，所以TextFile、ORCFile中TIMESTAMP存放的数据类似于 `2015-08-21 22:40:47.397898389` 。如果转换的类型配置为数据集成的DATE，转换之后会导致纳秒部分丢失。所以如果需要保留纳秒部分的数据，请配置转换类型为数据集成的字符串类型。

参数说明

参数	描述	是否必选	默认值
path	要读取的文件路径，如果要读取多个文件，可以使用简单正则表达式匹配，例如 `/hadoop/data_201704*` ，如果文件以时间命名且较为规律，则可以结合调度参数使用，调度参数将根据业务时间动态替换，详情请参见调度参数概述： • 当指定单个HDFS文件时，HDFS Reader暂时只能使用单线程进行数据抽取。 • 当指定多个HDFS文件时，HDFS Reader支持使用多线程进行数据抽取，线程并发数通过作业并发数concurrent指定。 **说明** 实际启动的并发数是您的HDFS待读取文件数量和您配置作业并发数两者中的小者。 • 当指定通配符，HDFS Reader尝试遍历出多个文件信息。例如指定/代表读取/目录下所有的文件，指定/bazhen/代表读取bazhen目录下游所有的文件。HDFS Reader目前只支持 `*` 和 `?` 作为文件通配符，语法类似于通常的Linux命令行文件通配符。 请注意以下事项： • 数据集成会将一个同步作业所有待读取文件视作同一张数据表。您必须自己保证所有的File能够适配同一套schema信息，并且提供给数据集成权限可读。 • 注意分区读取：Hive在建表时，可以指定分区。例如创建分区 `partition(day="20150820", hour="09")` ，对应的HDFS文件系统中，相应的表的目录下则会多出/20150820和/09两个目录且/20150820是/09的父目录。 分区会列成相应的目录结构，在按照某个分区读取某个表所有数据时，则只需配置好JSON中path的值即可。例如需要读取表名叫mytable01下分区day为20150820这一天的所有数据，则配置如下。 `"path": "/user/hive/warehouse/mytable01/20150820/*"`	是	无
defaultFS	Hadoop HDFS文件系统namenode节点地址。公共资源组不支持Hadoop高级参数HA的配置。	是	无

参数	描述	是否必选	默认值
fileType	文件的类型，目前仅支持您配置为*TEXT*、*ORC*、*RC*、*SEQ*、*CSV*和*parquet*。HDFS Reader能够自动识别文件的类型，并使用对应文件类型的读取策略。HDFS Reader在做数据同步前，会检查您配置的路径下所有需要同步的文件格式是否和fileType一致，如果不一致任务会失败。 fileType可以配置的参数值列表如下所示： • *TEXT*：表示TextFile文件格式。 • *ORC*：表示ORCFile文件格式。 • *RC*：表示RCFile文件格式。 • *SEQ*：表示SequenceFile文件格式。 • *CSV*：表示普通HDFS文件格式（逻辑二维表）。 • *PARQUET*：表示普通Parquet文件格式。 由于TextFile和ORCFile是两种不同的文件格式，所以HDFS Reader对这两种文件的解析方式也存在差异，这种差异导致Hive支持的复杂复合类型（例如map、array、struct和union）在转换为数据集成支持的String类型时，转换的结果格式略有差异，以map类型为例： • ORCFile map类型经HDFS Reader解析，转换成数据集成支持的STRING类型后，结果为 `{job=80, team=60, person=70}` 。 • TextFile map类型经HDFS Reader解析，转换成数据集成支持的STRING类型后，结果为 `{job:80, team:60, person:70}` 。 如上述转换结果所示，数据本身没有变化，但是表示的格式略有差异。所以如果您配置的文件路径中要同步的字段在Hive中是复合类型的话，建议配置统一的文件格式 。 最佳实践建议： • 如果需要统一复合类型解析出来的格式，建议您在Hive客户端将TextFile格式的表导成ORCFile格式的表。 • 如果是Parquet文件格式，后面的parquetSchema则必填，此属性用来说明要读取的Parquet格式文件的格式。 对于您指定的column信息，type必须填写，index和value必须选择其一。	是	无
column	读取字段列表，type指定源数据的类型，index指定当前列来自于文本第几列（以0开始），value指定当前类型为常量。不从源头文件读取数据，而是根据value值自动生成对应的列。默认情况下，您可以全部按照STRING类型读取数据，配置为 `"column": ["*"]` 。 您也可以指定column字段信息（文件数据列和常量列配置二选一），配置如下。 `{` `"type": "long",` `"index": 0` `//从本地文件文本第一列（下标索引从0开始计数）获取INT字段，index表示从数据文件中获取列数据。` `},` `{` `"type": "string",` `"value": "alibaba"` `//HDFS Reader内部生成alibaba的字符串字段作为当前字段，value表示常量列。` `}` **说明** • index从0开始（下标索引从0开始计数），表示从本地文本文件第一列开始读取数据。 • 建议您指定待读取的每一列数据的下标和类型，避免配置 `column *` 通配符。	是	无

参数	描述	是否必选	默认值
fieldDelimiter	读取的字段分隔符，HDFS Reader在读取TextFile数据时，需要指定字段分割符，如果不指定默认为逗号（,）。HDFS Reader在读取ORCFile时，您无需指定字段分割符，Hive本身的默认分隔符为\u0001。 **说明** • 如果您想将每一行作为目的端的一列，分隔符请使用行内容不存在的字符。例如，不可见字符\u0001。 • 分隔符不能使用\n。	否	,
encoding	读取文件的编码配置。	否	utf-8
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据集成提供nullFormat定义哪些字符串可以表示为null。 例如您配置 `nullFormat:"null"` ，如果源头数据是null，数据集成会将其视作null字段。 **说明** 字符串的null（n、u、l、l四个字符）和实际的null不同。	否	无
compress	当fileType（文件类型）为csv下的文件压缩方式，目前仅支持gzip、bz2、zip、lzo、lzo_deflate、hadoop-snappy和framing-snappy压缩。 **说明** • LZO存在lzo和lzo_deflate两种压缩格式。您在配置时，请注意不要配置错误。 • 由于snappy目前没有统一的stream format，数据集成目前仅支持最主流的hadoop-snappy（hadoop上的snappy stream format）和framing-snappy（google建议的snappy stream format）。 • ORC文件类型下无需填写。	否	无
parquetSchema	如果您的文件格式类型为Parquet，在配置column配置项的基础上，您还需配置parquetSchema，具体表示parquet存储的类型说明。您需要确保填写parquetSchema后，整体配置符合JSON语法。 `message MessageType名 {` `是否必填，数据类型，列名;` `.....;` `}` parquetSchema的配置格式说明如下： • MessageType名：填写名称。 • 是否必填：required表示非空，optional表示可为空。推荐全填optional。 • 数据类型：Parquet文件支持BOOLEAN、Int32、Int64、Int96、FLOAT、DOUBLE、BINARY（如果是字符串类型，请填BINARY）和fixed_len_byte_array类型。 • 每行列设置必须以分号结尾，最后一行也要写上分号。 配置示例如下所示。 `"parquetSchema": "message m { optional int32 minute_id; optional int32 dsp_id; optional int32 adx_pid; optional int64 req; optional int64 res; optional int64 suc; optional int64 imp; optional double revenue; }"`	否	无

参数	描述	是否必选	默认值
csvReaderConfig	读取CSV类型文件参数配置，Map类型。读取CSV类型文件使用的CsvReader进行读取，会有很多配置，不配置则使用默认值。 常见配置如下所示。 `"csvReaderConfig":{ "safetySwitch": false, "skipEmptyRecords": false, "useTextQualifier": false }` 所有配置项及默认值，配置时csvReaderConfig的map中请严格按照以下字段名字进行配置。 `boolean caseSensitive = true; char textQualifier = 34; boolean trimWhitespace = true; boolean useTextQualifier = true;//是否使用csv转义字符。 char delimiter = 44;//分隔符 char recordDelimiter = 0; char comment = 35; boolean useComments = false; int escapeMode = 1; boolean safetySwitch = true;//单列长度是否限制100,000字符。 boolean skipEmptyRecords = true;//是否跳过空行。 boolean captureRawRecord = true;`	否	无

<table>
<tr><th>参数</th><th>描述</th><th>是否必选</th><th>默认值</th></tr>
<tr><td>hadoopConfig</td><td>hadoopConfig中可以配置与Hadoop相关的一些高级参数，例如HA的配置。公共资源组不支持Hadoop高级参数HA的配置。
<pre>"hadoopConfig":{
"dfs.nameservices": "testDfs",
"dfs.ha.namenodes.testDfs": "namenode1,namenode2",
"dfs.namenode.rpc-address.youkuDfs.namenode1": "",
"dfs.namenode.rpc-address.youkuDfs.namenode2": "",
"dfs.client.failover.proxy.provider.testDfs":
"org.apache.hadoop.hdfs.server.namenode.ha.ConfiguredFailoverProxyProvider",
"dfs.data.transfer.protection": "integrity",
"dfs.datanode.use.datanode.hostname" :"true",
"dfs.client.use.datanode.hostname":"true"
}</pre>
说明
<pre>"hadoopConfig":{ "dfs.data.transfer.protection":
"integrity", "dfs.datanode.use.datanode.hostname"
:"true", "dfs.client.use.datanode.hostname":"true" }</pre>
上述参数为hdfs reader插件中配置的kerberos认证。如果您在HDFS数据源已经配置了kerberos认证，则在hdfs reader插件中无需重复配置。配置HDFS数据源，详情请参见配置HDFS数据源</td><td>无</td><td></td></tr>
<tr><td>haveKerberos</td><td>是否有Kerberos认证，默认为false。例如用户配置为true，则配置项kerberosKeytabFilePath和kerberosPrincipal为必填。</td><td>否</td><td>false</td></tr>
<tr><td>kerberosKeytabFilePath</td><td>Kerberos认证keytab文件的绝对路径。如果haveKerberos为true，则必选。</td><td>否</td><td>无</td></tr>
<tr><td>kerberosPrincipal</td><td>Kerberos认证Principal名，如****/hadoopclient@**.*** 。如果haveKerberos为true，则必选。
说明 由于Kerberos需要配置keytab认证文件的绝对路径，您需要在独享数据集成资源组上使用此功能。配置示例如下。
<pre>"haveKerberos":true,
"kerberosKeytabFilePath":"/opt/datax/**.keytab",
"kerberosPrincipal":"**/hadoopclient@**.**"</pre></td><td>否</td><td>无</td></tr>
</table>

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

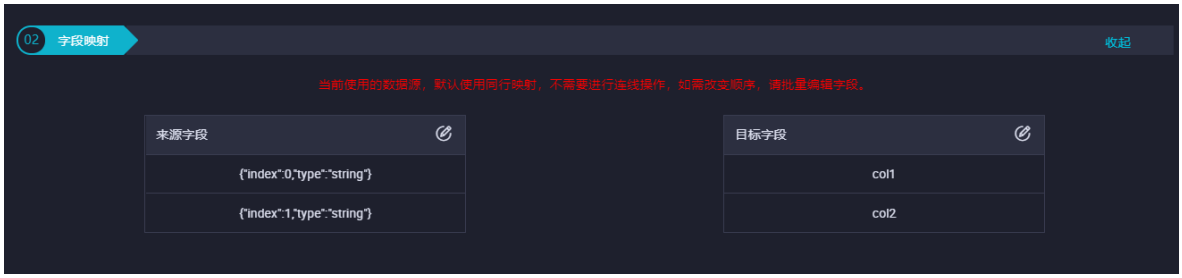
您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
文件路径	即上述参数说明中的path。
文本类型	即上述参数说明中的fileType。读取的文件类型，目前支持您配置为*TEXT*、*ORC*、*RC*、*SEQ*、*CSV*和*parquet*。
字段分隔符	即上述参数说明中的fieldDelimiter，默认值为（,）。
文件编码	即上述脚本模式参数说明中的encoding，默认值为utf-8。
Kerberos认证	是否有Kerberos认证，默认为否，当配置为是的时候，配置项Keytab文件路径和Principal名为必填。详情请参见附录：配置Kerberos认证。
是否忽略(文件不存在时)	用户配置的文件，或者文件夹不存在时，是否忽略（即日志中不报异常信息），是表示文件不存在则忽略，否表示文件不存在则任务出错。默认值为否。
NullFormat	即上述参数说明中的NullFormat。定义哪些字符串可以表示为null。
HadoopConfig	hadoopConfig中可以配置与Hadoop相关的一些高级参数，例如HA的配置。公共资源组不支持Hadoop高级参数HA的配置。

2. 字段映射，即上述参数说明中的column。默认使用**同行映射**。您可以单击图标手动编辑源表字段，一行表示一个字段，首尾空行会被采用，其它空行会被忽略。



   > **说明** index从0开始（下标索引从0开始计数），表示从本地文本文件第一列开始读取数据。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从HDFS抽取数据到本地的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "hdfs",//插件名
            "parameter": {
                "path": "",//要读取的文件路径
                "datasource": "",//数据源
                "hadoopConfig":{
                "dfs.data.transfer.protection": "integrity",
               "dfs.datanode.use.datanode.hostname" :"true",
                "dfs.client.use.datanode.hostname":"true"
                 },
                "column": [
                    {
                        "index": 0,//序列号,index从0开始（下标索引从0开始计数），表示从本地文本文件第一列开始读取数据。
                        "type": "string"//字段类型
                    },
                    {
                        "index": 1,
                        "type": "long"
                    },
                    {
                        "index": 2,
                        "type": "double"
                    },
                    {
                        "index": 3,
                        "type": "boolean"
                    },
                    {
                        //日期格式
```

```
                        "format": "yyyy-MM-dd HH:mm:ss", //日期格式
                        "index": 4,
                        "type": "date"
                    }
                ],
                "fieldDelimiter": ","//列分隔符
                "encoding": "UTF-8",//编码格式
                "fileType": ""//文本类型
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": ""//错误记录数
        },
        "speed": {
            "concurrent": 3,//作业并发数
            "throttle": true //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

parquetSchema的HDFS Reader配置示例如下。

> **说明**
>
> - fileType配置项必须设置为parquet。
> - 如果您要读取parquet文件中的部分列，需在parquetSchema配置项中，指定完整schema结构信息，并在column中根据下标，筛选需要的同步列进行列映射。

```
"reader":  {
    "name": "hdfsreader",
    "parameter": {
        "path": "/user/hive/warehouse/addata.db/dw_ads_rtb_monitor_minute/thedate=20170103/hour_id=22/*",
        "defaultFS": "h10s010.07100.149:8020",
        "column": [
            {
                "index": 0,
                "type": "string"
            },
            {
                "index": 1,
                "type": "long"
            },
            {
                "index": 2,
                "type": "double"
            }
        ],
        "fileType": "parquet",
        "encoding": "UTF-8",
        "parquetSchema": "message m { optional int32 minute_id; optional int32 dsp_id; optional int32 adx_pid
; optional int64 req; optional int64 res; optional int64 suc; optional int64 imp; optional double revenue; }"
    }
}
```

6.2.5. MongoDB Reader

本文为您介绍MongoDB Reader支持的数据类型、字段映射和数据源等参数及配置示例。

MongoDB Reader插件通过MongoDB的Java客户端MongoClient，进行MongoDB的读操作。最新版本的Mongo已经将DB锁的粒度，从DB级别降低至document级别，配合MongoDB强大的索引功能，即可达到高性能读取MongoDB的需求。

> **说明**
>
> - 如果您使用的是云数据库MongoDB版，MongoDB默认会有root账号。出于安全策略的考虑，数据集成仅支持使用MongoDB数据库对应账号进行连接。您添加使用MongoDB数据源时，也请避免使用root作为访问账号。
> - query不支持JS语法。

MongoDB Reader通过数据集成框架从MongoDB并行地读取数据，通过主控的Job程序，按照指定规则对MongoDB中的数据进行分片并行读取，然后将MongoDB支持的类型通过逐一判断转换为数据集成支持的类型。

类型转换列表

MongoDB Reader支持大部分MongoDB类型，但也存在部分没有支持的情况，请注意检查您的数据类型。

MongoDB Reader针对MongoDB类型的转换列表，如下表所示。

类型分类	MongoDB数据类型
LONG	INT、LONG、document.INT和document.LONG
DOUBLE	DOUBLE和document.DOUBLE
STRING	STRING、ARRAY、document.STRING、document.ARRAY和COMBINE
DATE	DATE和document.DATE
BOOLEAN	BOOL和document.BOOL
BYTES	BYTES和document.BYTES

> **说明** document类型为嵌入文档类型，即OBJECT类型。

COMBINE类型的使用如下：

使用MongoDB Reader插件读出数据时，支持将MongoDB document中的多个字段合并成一个JSON串。

例如，导入MongoDB中的字段至MaxCompute，有字段如下（下文均省略了value使用key来代替整个字段）的三个document，其中a、b是所有document均有的公共字段，x_n是不固定字段。

```
doc1: a b x_1 x_2
doc2: a b x_2 x_3 x_4
doc3: a b x_5
```

配置文件中要明确指出需要一一对应的字段，需要合并的字段则需另取名称（不可以与document中已存在字段同名），并指定类型为COMBINE，如下所示。

```
"column": [
{
"name": "a",
"type": "string",
},
{
"name": "b",
"type": "string",
},
{
"name": "doc",
"type": "combine",
}
]
```

最终导出的MaxCompute结果如下所示。

odps_column1	odps_column2	odps_column3
a	b	{x_1,x_2}
a	b	{x_2,x_3,x_4}
a	b	{x_5}

> **说明**
>
> 使用COMBINE类型合并MongoDB Document中的多个字段后，输出结果映射至MaxCompute时会自动删除公共字段，仅保留Document的特有字段。
>
> 例如，a、b为所有Document均有的公共字段，Document文件 `doc1: a b x_1 x_2` 使用COMBINE类型合并字段后，输出结果本应该为*{a,b,x_1,x_2}*，该结果映射至MaxCompute后，会删除公共字段a和b，最终输出的结果为*{x_1,x_2}*。

使用限制

- 数据同步任务时，从源并行读取或并行写入数据存储端的最大线程数为1。
- 切分key必须是整数，否则可能会导致切片不连续，出现丢数据的风险。
- MongoDB版本限制：仅支持4.x版本。

参数说明

参数	描述
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。

参数	描述
collectionName	MonogoDB的集合名。
column	MongoDB的文档列名，配置为数组形式表示MongoDB的多个列。 • *name*：column的名字。 • *type*支持的类型包括： ○ *string*：表示字符串。 ○ *long*：表示整型数。 ○ *double*表示浮点数。 ○ *date*表示日期。 ○ *bool*表示布尔值。 ○ *bytes*：表示二进制序列。 ○ *arrays*：以JSON字符串格式读出，例如["a","b","c"]。 ○ *array*：以分隔符splitter分隔的方式读出，例如 `a,b,c` ，推荐使用*arrays*格式。 ○ *combine*使用MongoDB Reader插件读出数据时，支持合并MongoDB document中的多个字段为一个JSON串。 • *splitter*：因为MongoDB支持数组类型，但数据集成框架本身不支持数组类型，所以MongoDB读出来的数组类型，需要通过该分隔符合并成字符串。
batchSize	批量获取的记录数，该参数为选填参数。默认值为 `1000` 条。
cursorTimeoutInMs	游标超时时间，该参数为选填参数。默认值为 `1000 * 60 * 10 = 600000` 。如果*cursorTimeoutInMs*配置为负值，则表示游标永不超时。 **说明** • 不推荐您设置游标永不超时。如果客户端程序意外退出，永不超时的游标将一直存在于MongoDB服务器中，直到服务重启。 • 如果出现游标超时，您可以执行如下操作： ○ 减小批量获取的记录数*batchSize*。 ○ 增加游标超时时间*cursorTimeoutInMs*。
query	您可以通过该配置型来限制返回MongoDB数据范围，仅支持以下时间格式，不支持直接使用时间戳类型的格式。例如您可以配置 `"query":"{'operationTime':{'$gte':ISODate('${last_day}T00:00:00.424+0800')}}"` ，限制返回operationTime大于等于${last_day}零点的数据。此处${last_day}为DataWorks调度参数，其中*last_day*格式为 `yyyy-mm-dd` 。您可以根据需要具体使用其它MongoDB支持的条件操作符号（$gt、$lt、$gte和$lte等），逻辑操作符（and和or等），函数（max、min、sum、avg和ISODate等）。该参数为选填参数。

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
集合名称	即上述参数说明中的collectionName。
批量条数	从MongoDB批量获取的数据条数，默认值为1000。
游标超时时间	游标超时时间，默认值是3600000毫秒，如果配置为负数，则永不超时。
检索查询条件	即上述参数说明中的query。您可以通过该配置项来限制返回MongoDB数据范围。

2. 字段映射，即上述参数说明中的column。默认使用**同行映射**。您可以单击图标手动编辑源表字段。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。 **说明** 当前任务期望最大并发数仅支持配置为1。

参数	描述
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

配置一个从MongoDB抽取数据到本地的作业，详情请参见上述参数说明。

> **注意**
> - 实际运行时，请删除下述代码中的注释。
> - 暂时不支持取出array中的指定元素。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "datasource": "datasourceName", //数据源名称。
                "collectionName": "tag_data", //集合名称。
                "query": "", // 数据查询过滤。
                "column": [
                    {
                        "name": "unique_id", //字段名称。
                        "type": "string" //字段类型。
                    },
                    {
                        "name": "sid",
                        "type": "string"
                    },
                    {
                        "name": "user_id",
                        "type": "string"
                    },
                    {
                        "name": "auction_id",
                        "type": "string"
                    },
                    {
                        "name": "content_type",
                        "type": "string"
                    },
                    {
                        "name": "pool_type",
                        "type": "string"
                    },
                    {
                        "name": "frontcat_id",
                        "type": "array",
                        "splitter": ""
```

```
                    },
                    {
                        "name": "categoryid",
                        "type": "array",
                        "splitter": ""
                    },
                    {
                        "name": "gmt_create",
                        "type": "string"
                    },
                    {
                        "name": "taglist",
                        "type": "array",
                        "splitter": " "
                    },
                    {
                        "name": "property",
                        "type": "string"
                    },
                    {
                        "name": "scorea",
                        "type": "int"
                    },
                    {
                        "name": "scoreb",
                        "type": "int"
                    },
                    {
                        "name": "scorec",
                        "type": "int"
                    },
                    {
                        "name": "a.b",
                        "type": "document.int"
                    },
                    {
                        "name": "a.b.c",
                        "type": "document.array",
                        "splitter": " "
                    }
                ]
            },
            "stepType": "mongodb"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
```

```
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.6. DB2 Reader

本文为您介绍DB2 Reader支持的数据类型、字段映射和数据源等参数及配置示例。

> 注意 DB2 Reader仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

背景信息

DB2 Reader插件实现了从DB2读取数据。在底层实现上，DB2 Reader通过JDBC连接远程DB2数据库，并执行相应的SQL语句，从DB2库选取数据。

DB2 Reader通过JDBC连接器连接至远程的DB2数据库，根据您配置的信息生成查询SQL语句，发送至远程DB2数据库，执行该SQL语句并返回结果。然后使用**数据集成**自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理：

- 对于您配置的table、column、where等信息，DB2 Reader将其拼接为SQL语句发送至DB2数据库。
- 对于您配置的querySql信息，DB2 Reader直接将其发送至DB2数据库。

DB2 Reader使用的DB2驱动版本为IBM Data Server Driver for JDBC and SQLJ 4.11.77。DB2驱动和数据库服务之间的版本映射请参见官网文档。

类型转换列表

DB2 Reader支持大部分DB2类型，但也存在个别类型没有支持的情况，请注意检查您的数据类型。

DB2 Reader针对DB2类型的转换列表，如下所示。

类型分类	DB2数据类型
整数类	SMALLINT
浮点类	DECIMAL、REAL和DOUBLE
字符串类	CHAR、CHARACTER、VARCHAR、GRAPHIC、VARGRAPHIC、LONG VARCHAR、CLOB、LONG VARGRAPHIC和DBCLOB
日期时间类	DATE、TIME和TIMESTAMP
布尔类	无
二进制类	BLOB

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
jdbcUrl	描述的是到DB2数据库的JDBC连接信息，jdbcUrl按照DB2官方规范，DB2格式为 `jdbc:db2://ip:port/database` ，并可以填写连接附件控制信息。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无
table	所选取的需要同步的表，一个作业只能支持一个表同步。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息，默认使用所有列配置。例如[*]： • 支持列裁剪，即可以导出部分列。 • 支持列换序，即可以不根据表Schema信息顺序导出列。 • 支持常量配置，您需要遵循DB2的SQL语法格式。例如，`["id", "1", "'const name'", "null", "upper('abc_lower')", "2.3" , "true"]`： ○ id为普通列名。 ○ 1为整型数字常量。 ○ 'const name'为字符串常量（需要加上一对单引号）。 ○ null为空指针。 ○ upper('abc_lower')为函数表达式。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示您指定同步的列集合，不允许为空 。	是	无
splitPk	DB2 Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片。数据同步系统会启动并发任务进行数据同步，以提高数据同步的效能： • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整形数据切分，不支持浮点、字符串和日期等其它类型。如果您指定其它非支持类型，DB2 Reader将报错。	否	*""*
where	筛选条件，DB2 Reader根据指定的column、table、where条件拼接SQL，并根据该SQL抽取数据。 在实际业务场景中，通常会选择当天的数据进行同步，可以指定where条件为`gmt_create>$bizdate`。where条件可以有效地进行业务增量同步。如果该值为空，代表同步全表所有的信息。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当您配置该项后，数据同步系统会忽略table、column等配置，直接使用该配置项的内容筛选数据。 例如，您需要在多表Join后同步数据，使用`select a,b from table_a join table_b on table_a.id = table_b.id`。当您配置querySql时，DB2 Reader直接忽略table、column、where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据同步系统和服务器端的网络交互次数，能够较大的提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1024*

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

配置一个从DB2数据库同步抽取数据作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        {
            "stepType":"db2",//插件名
            "parameter":{
                "password":"",//密码
                "jdbcUrl":"",//DB2数据库的JDBC连接信息
                "column":[
                    "id"
                ],
                "where":"",//筛选条件。
                "splitPk":"",//数据分片
                "table":"",//表名
                "username":""//用户名
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数
        },
        "speed":{
             "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
             "concurrent":1 //作业并发数。
             "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

补充说明

- 主备同步数据恢复问题

  主备同步问题指DB2使用主从灾备，备库从主库不间断通过binlog恢复数据。由于主备数据同步存在一定的时间差，特别在网络延迟等情况下，会导致备库同步恢复的数据不是一份当前时间的完整镜像，与主库有较大差别。

- 一致性约束

  DB2在数据存储划分中属于RDBMS系统，对外可以提供强一致性数据查询接口。例如，一次同步任务启动运行过程中，有其它数据源写入数据至该库。由于数据库本身的快照特性，DB2 Reader完全不会获取到写入的更新数据。

  上述是在DB2 Reader单线程模型下数据同步一致性的特性，DB2 Reader可以根据您配置信息并发抽取数据，因此不能严格保证数据的一致性。

  DB2 Reader根据splitPk切分数据后，会先后启动多个并发任务完成数据同步。多个并发任务不属于同一个读事务且存在时间间隔。对于原始数据而言，多并发读取的快照数据可能存在不完整、不一致的情况。

针对多线程的一致性快照需求，目前在技术上无法实现，只能从工程角度解决。工程化的方式存在取舍，在此提供以下解决思路，您可根据自身情况进行选择：

- 使用单线程同步，即不再进行数据切片。缺点是速度比较慢，但是能够很好保证一致性。
- 关闭其他数据写入方，保证当前数据为静态数据，例如锁表、关闭备库同步等。缺点是可能影响在线业务。

- 数据库编码问题

  DB2 Reader底层使用JDBC进行数据抽取，JDBC天然适配各类编码，并在底层进行了编码转换。因此DB2 Reader不需您指定编码，可以自动识别编码并转码。

- 增量数据同步的方式

  DB2 Reader使用JDBC SELECT语句完成数据抽取工作，因此可以使用`SELECT...WHERE...`进行增量数据抽取，方式如下：

  - 数据库在线应用写入数据库时，填充modify字段为更改时间戳，包括新增、更新、删除（逻辑删除）。对于该类应用，DB2 Reader只需要where条件后跟上一同步阶段时间戳即可。
  - 对于新增流水型数据，DB2 Reader在where条件后跟上一阶段最大自增ID即可。

  对于业务上无字段区分新增、修改数据的情况，DB2 Reader无法进行增量数据同步，只能同步全量数据。

- SQL安全性

  DB2 Reader提供querySql语句交给您自己实现SELECT抽取语句，DB2 Reader本身对querySql不进行任何安全性校验。

6.2.7. MySQL Reader

本文为您介绍MySQL Reader支持的数据类型、字段映射和数据源等参数及配置示例。

前提条件

开始配置MySQL Reader插件前，请首先配置好数据源，详情请参见配置MySQL数据源。

背景信息

MySQL Reader插件通过JDBC连接器连接至远程的MySQL数据库，根据您配置的信息生成查询SQL语句，发送至远程MySQL数据库，执行该SQL语句并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

在底层实现上，MySQL Reader插件通过JDBC连接远程MySQL数据库，并执行相应的SQL语句，从MySQL库中抽取数据。

MySQL Reader插件支持读取表和视图。表字段可以依序指定全部列、指定部分列、调整列顺序、指定常量字段和配置MySQL的函数，例如now()等。

类型转换列表

MySQL Reader针对MySQL类型的转换列表，如下所示。

类型分类	MySQL数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔型	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

> 注意
>
> - 除上述罗列字段类型外，其它类型均不支持。
> - MySQL Reader插件将tinyint（1）视作整型。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称，一个数据集成任务只能同步数据到一张目标表。 table用于配置范围的高级用法示例如下： • 您可以通过配置区间读取分库分表，例如 `'table_[0-99]'` 表示读取 `'table_0'` 、 `'table_1'` 、 `'table_2'` 直到 `'table_99'` 。 • 如果您的表数字后缀的长度一致，例如 `'table_000'` 、 `'table_001'` 、 `'table_002'` 直到 `'table_999'` ，您可以配置为 `'"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'` 。 **说明** 任务会读取匹配到的所有表，具体读取这些表中column配置项指定的列。如果表不存在，或者读取的列不存在，会导致任务失败。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪：列可以挑选部分列进行导出。 • 支持列换序：列可以不按照表schema信息顺序进行导出。 • 支持常量配置：您需要按照MySQL SQL语法格式，例如 `["id","table","1","'mingya.wmy'","'null'","to_char(a+1)","2.3","true"]` 。 ◦ id为普通列名。 ◦ table为包含保留字的列名。 ◦ 1为整型数字常量。 ◦ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ◦ 关于null： ▪ `" "` 表示空。 ▪ `null` 表示null。 ▪ `'null'` 表示null这个字符串。 ◦ to_char(a+1)为计算字符串长度函数。 ◦ 2.3为浮点数。 ◦ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	MySQL Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点和日期等其他类型 。如果您指定其他非支持类型，忽略splitPk功能，使用单通道进行同步。 • 如果不填写splitPk，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据 。	否	无

参数	描述	是否必选	默认值
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。如果不填写where语句，包括不提供where的key或value，数据同步均视作同步全量数据。 • 不可以将where条件指定为limit 10，这不符合MySQL SQL WHERE子句约束。	否	无
querySql（高级模式，向导模式不提供）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。配置该项后，数据同步系统会忽略tables、columns和splitPk配置项，直接使用该项配置的内容对数据进行筛选。例如，需要进行多表join后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，MySQL Reader直接忽略table、column、where和splitPk条件的配置，querySql优先级大于table、column、where和splitPk选项。datasource通过它解析出用户名和密码等信息。 **说明** querySql需要区分大小写，例如，写为querysql会不生效。	否	无
singleOrMulti（仅适用于分库分表）	表示分库分表，向导模式转换成脚本模式主动生成此配置 `"singleOrMulti":"multi"` ，但配置脚本任务模板不会直接生成此配置必须手动添加，否则只会识别第一个数据源。	是	*multi*

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

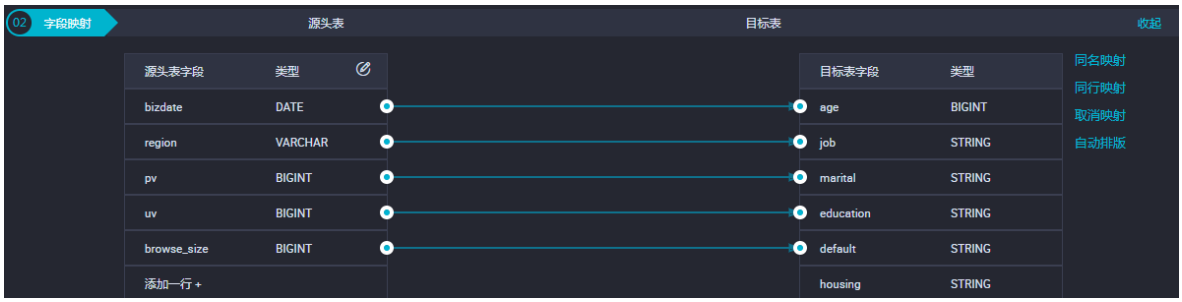
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。

参数	描述
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

本文为您提供单库单表和分库分表的配置示例：

- 配置单库单表

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"mysql",//插件名。
            "parameter":{
                "column":[//列名。
                    "id"
                ],
                "connection":[
                    {   "querySql":["select a,b from join1 c join join2 d on c.id = d.id;"], //使用字符串的形式，将querySql写在connection中。
                        "datasource":"",//数据源。
                        "table":[//表名，即使只有一张表，也必须以[]的数组形式书写。
                            "xxx"
                        ]
                    }
                ],
                "where":"",//过滤条件。
                "splitPk":"",//切分键。
                "encoding":"UTF-8"//编码格式。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1,//作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

- 配置分库分表

> **说明** 分库分表是指在MySQL Reader端可以选择多个MySQL数据表，且表结构保持一致。

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {
            "plugin": "mysql",
            "parameter": {
                "connection": [
                    {
                        "table": [
                            "tbl1",
                            "tbl2",
                            "tbl3"
                        ],
                        "datasource": "datasourceName1"
                    },
                    {
                        "table": [
                            "tbl4",
                            "tbl5",
                            "tbl6"
                        ],
                        "datasource": "datasourceName2"
                    }
                ],
                "singleOrMulti": "multi",
                "splitPk": "db_id",
                "column": [
                    "id", "name", "age"
                ],
                "where": "1 < id and id < 100"
            }
        },
        "writer": {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
        "setting":{
            "errorLimit":{
               "record":"0"//错误记录数。
        },
            "speed":{
               "throttle":false,//false代表不限流，下面的限流的速度不生效；true代表限流。
               "concurrent":1,//作业并发数。
        }
    },
     "order":{
        "hops":[
        {
            "from":"Reader",
            "to":"Writer"
        }
        ]
    }
    }
}
```

6.2.8. Oracle Reader

本文为您介绍Oracle Reader支持的数据类型、字段映射和数据源等参数及配置示例。

Oracle Reader插件实现了从Oracle读取数据。在底层实现上，Oracle Reader通过JDBC连接远程Oracle数据库，并执行相应的SQL语句，从Oracle数据库中选取数据。

> 说明
> - RDS和DRDS不提供Oracle存储引擎。
> - Oracle Reader插件使用ojdbc7-12.1.0.2.jar驱动，支持的Oracle版本请参见Oracle官网。

Oracle Reader通过JDBC连接器连接至远程的Oracle数据库，根据您配置的信息生成查询语句，并发送至远程Oracle数据库。然后使用数据集成自定义的数据类型，将该SQL执行返回结果拼装为抽象的数据集，并传递给下游Writer处理。

- 对于您配置的table、column和where信息，Oracle Reader将其拼接为SQL语句，发送至Oracle数据库。
- 对于您配置的querySql信息，Oracle直接将其发送至Oracle数据库。

类型转换列表

Oracle Reader支持大部分Oracle类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型 。

Oracle Reader针对Oracle类型的转换列表，如下所示。

类型分类	Oracle数据类型
整数类	NUMBER、RAWID、INTEGER、INT和SMALLINT
浮点类	NUMERIC、DECIMAL、FLOAT、DOUBLE PRECISIOON和REAL
字符串类	LONG、CHAR、NCHAR、VARCHAR、VARCHAR2、NVARCHAR2、CLOB、NCLOB、CHARACTER、CHARACTER VARYING、CHAR VARYING、NATIONAL CHARACTER、NATIONAL CHAR、NATIONAL CHARACTER VARYING、NATIONAL CHAR VARYING和NCHAR VARYING
日期时间类	TIMESTAMP和DATE
布尔型	BIT和BOOL
二进制类	BLOB、BFILE、RAW和LONG RAW

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项输入的内容必须和添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息。默认使用所有列配置，例如["*"]。 • 支持列裁剪，即可以导出部分列。 • 支持列换序，即可以不根据表Schema信息的顺序导出列。 • 支持常量配置，您需要按照JSON格式进行配置。 `["id", "1", "'mingya.wmy'", "null", "to_char(a + 1)", "2.3" , "true"]` ○ id为普通列名。 ○ 1为整型数字常量。 ○ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ○ null为空指针。 ○ to_char(a + 1)为表达式。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示填写，不允许为空。	是	无
splitFactor	切分因子，可以配置同步数据的切分份数，如果配置了多并发，会按照并发数 * splitFactor份来切分。例如，并发数=5，splitFactor=5，则会按照5*5=25份来切分，在5个并发线程上执行。 **说明** 建议取值范围：1～100，过大会导致内存溢出。	否	5
splitMode	切分模式，包括： • **averageInterval**：平均采样，根据**splitPK**找到最大值和最小值，然后按照切分数目平均切分。 • **randomSampling**：随机采样，在所有数据中随机找到一定数目作为切分点。 **说明** • splitPk为字符串类型时，**splitMode**需要配为**randomSampling**。 • splitMode切分模式为**averageInterval**时，仅支持splitPk配置为数值类型。	否	**randomSampling**
splitPk	Oracle Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，可以提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • splitPk支持数值类型、字符串类型，浮点和日期等其它类型。 • 如果不填写splitPk，将视作您不对单表进行切分，Oracle Reader使用单通道同步全量数据。 **说明** splitPK字段在视图的情况下不能使用ROWID。	否	无
where	筛选条件，Oracle Reader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。例如，在测试时指定where条件为row_number()。 • where条件可以有效地进行业务增量同步 。 • where条件不配置或为空时，将视作全表同步数据 。	否	无

参数	描述	是否必选	默认值
querySql（高级模式，向导模式不支持）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置来自定义筛选SQL。当您配置该项后，数据同步系统就会忽略table和column等配置，直接使用该配置项的内容对数据进行筛选。例如，需要进行多表Join后同步数据，则使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，Oracle Reader直接忽略table、column和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据同步系统和服务器端的网络交互次数，能够较大的提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	1,024

向导开发介绍

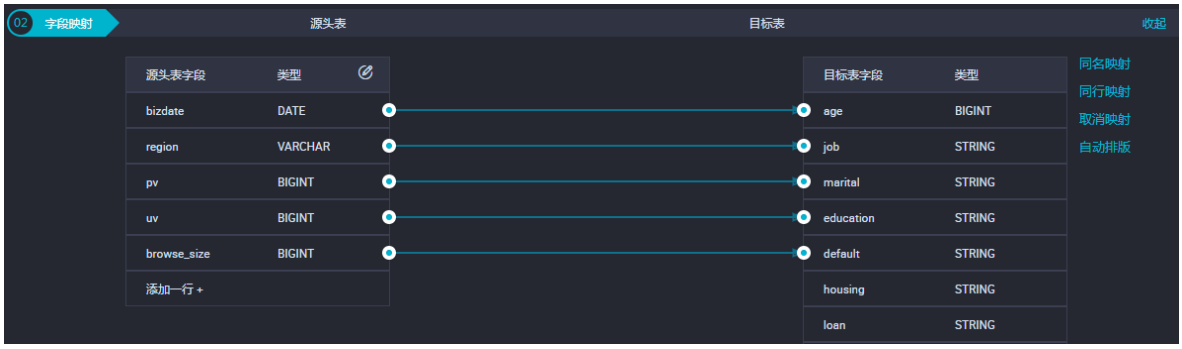
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，向导模式仅支持类型为整型的字段。 如果需要配置为字符串，浮点和日期等其它类型的字段，请切换成脚本模式。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为**未识别**。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从Oracle数据库同步抽取数据的作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"oracle",
            "parameter":{
                "fetchSize":1024,//该配置项定义了插件和数据库服务器端每次批量数据获取条数。
                "datasource":"",//填写添加的数据源名。
                "column":[//列名。
                    "id",
                    "name"
                ],
                "where":"",//筛选条件。
                "splitPk":"",//切分键。
                "table":""//表名。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
} "to":"Writer"
            }
        ]
    }
}
```

使用说明

- 主备同步数据恢复问题

  主备同步问题指Oracle使用主从灾备，当主库报错切换至备库后，备库从主库不断地通过binlog恢复数据。由于主备数据同步存在一定的时间差，在网络延迟等特定情况下，会导致备库同步恢复的数据与主库有较大差别，从备库同步的数据不是一份当前时间的完整镜像。

- 一致性约束

  Oracle在数据存储划分中属于RDBMS系统，对外可以提供强一致性数据查询接口。例如，在一次同步任务启动运行的过程中，当该库存在其它数据写入方写入数据时，由于数据库本身的快照特性，Oracle Reader不会获取到写入的新数据。

上述是在Oracle Reader单线程模型下实现数据同步的一致性。Oracle Reader根据您配置的信息并发抽取数据，则不能严格保证数据一致性。

当Oracle Reader根据splitPk进行数据切分后，会先后启动多个并发任务完成数据同步。多个并发任务相互之间不属于同一个读事务，同时多个并发任务存在时间间隔。因此该数据并不是完整的、一致的数据快照信息。

针对多线程的一致性快照需求，目前在技术上无法实现，只能从工程角度解决。工程化的方式存在取舍，在此提供以下解决思路，您可以根据自身情况进行选择。

- 使用单线程同步，即不再进行数据切片。缺点是速度比较慢，但是能够很好保证一致性。
- 关闭其它数据写入方，保证当前数据为静态数据。例如，锁表、关闭备库同步等。缺点是可能影响在线业务。

- 数据库编码问题

  Oracle Reader底层使用JDBC进行数据抽取，JDBC天然适配各类编码，并在底层进行了编码转换。因此Oracle Reader无需您指定编码，可以自动获取编码并转码。

- 增量数据同步的方式

  Oracle Reader使用JDBC SELECT语句完成数据抽取工作，因此您可以使用`SELECT...WHERE...`进行增量数据抽取，方式如下：

  - 数据库在线应用写入数据库时，填充modify字段为更改时间戳，包括新增、更新、删除（逻辑删除）。对于该类应用，Oracle Reader只需要where条件后跟上一同步阶段时间戳即可。
  - 对于新增流水型数据，Oracle Reader在where条件后跟上一阶段最大自增ID即可。

  对于业务上无字段区分新增、修改数据的情况，Oracle Reader无法进行增量数据同步，只能同步全量数据。

- SQL安全性

  Oracle Reader为您提供querySql功能，您可以自行实现SELECT抽取语句。Oracle Reader本身对querySql不进行任何安全性校验。

6.2.9. OSS Reader

本文为您介绍OSS Reader支持的数据类型、字段映射和数据源等参数及配置示例。

OSS Reader插件提供了读取OSS数据存储的能力。在底层实现上，OSS Reader使用OSS官方Java SDK获取OSS数据，并转换为数据同步传输协议传递给Writer。OSS Reader支持OSS中的BIGINT、DOUBLE、STRING、DATETIME和BOOLEAN数据类型。

OSS Reader实现了从OSS读取数据并转为数据集成协议的功能，OSS本身是无结构化数据存储。对于数据集成而言，OSS Reader支持的功能如下：

- 支持且仅支持读取TXT格式的文件，且要求TXT中schema为一张二维表。
- 支持类CSV格式文件，自定义分隔符。
- 支持多种类型数据读取（使用String表示），支持列裁剪、列常量。
- 支持递归读取、支持文件名过滤。
- 支持文本压缩，现有压缩格式为gzip、bzip2和zip。

  > **说明** 一个压缩包不允许多文件打包压缩。

- 多个Object可以支持并发读取。

OSS Reader暂时不能实现以下功能：

- 单个Object（File）支持多线程并发读取。
- 单个Object在压缩情况下，从技术上无法支持多线程并发读取。
- 单个Object（File）超过100 GB。

参考文档如下：

- 如果您想对OSS产品有更深了解，请参见OSS产品概述。
- OSS Java SDK的详细介绍，请参见阿里云OSS Java SDK。
- 处理OSS等非结构化数据的详细介绍，请参见处理非结构化数据。

支持的数据类型

类型分类	数据集成column配置类型	数据库数据类型
整数类	LONG	LONG
字符串类	STRING	STRING
浮点类	DOUBLE	DOUBLE
布尔类	BOOLEAN	BOOL
日期时间类	DATE	DATE

参数说明

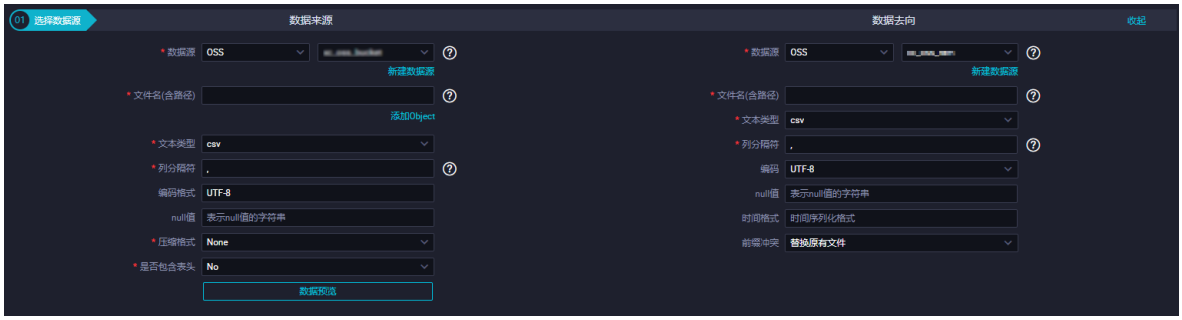
参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
Object	OSS的Object信息，此处可以支持填写多个Object。例如xxx的bucket中有yunshi文件夹，文件夹中有ll.txt文件，则Object直接填yunshi/ll.txt。 • 当指定单个OSS Object时，OSS Reader暂时只能使用单线程进行数据抽取。后期将考虑在非压缩文件情况下针对单个Object可以进行多线程并发读取。 • 当指定多个OSS Object时，OSS Reader支持使用多线程进行数据抽取。可以根据具体要求配置线程并发数。 • 当指定通配符时，OSS Reader尝试遍历出多个Object信息。例如配置为 `abc*[0-9]` 时，可以匹配到 `abc0` 、 `abc1` 、 `abc2` 、 `abc3` 等；配置为 `abc?.txt` 时，可以匹配到以 `abc` 开头、 `.txt` 结尾、中间有1个任意字符的文件。 配置通配符会导致内存溢出，通常不建议您进行配置。详情请参见OSS产品概述。 **说明** • 数据同步系统会将一个作业下同步的所有Object视作同一张数据表。您必须保证所有的Object能够适配同一套Schema信息。 • 请注意控制单个目录下的文件个数，否则可能会触发系统OutOfMemoryError报错。如果遇到此情况，请将文件拆分到不同目录后再尝试进行同步。	是	无

参数	描述	是否必选	默认值
column	读取字段列表，type指定源数据的类型，index指定当前列来自于文本第几列（以0开始），value指定当前类型为常量，不是从源头文件读取数据，而是根据value值自动生成对应的列。 默认情况下，您可以全部按照String类型读取数据，配置如下。 `json` `"column": ["*"]` 您可以指定column字段信息，配置如下。 `json` `"column":` `{` `"type": "long",` `"index": 0 //从OSS文本第一列获取int字段。` `},` `{` `"type": "string",` `"value": "alibaba" //从OSSReader内部生成alibaba的字符串字段作为当前字段。` `}` **说明** 对于您指定的column信息，type必须填写，index/value必须选择其一。	是	全部按照STRING类型读取。
fieldDelimiter	读取的字段分隔符。 **说明** OSS Reader在读取数据时，需要指定字段分割符，如果不指定默认为（,），界面配置中也会默认填写为（,）。 如果分隔符不可见，请填写Unicode编码。例如，\u001b、\u007c。	是	,
compress	文本压缩类型，默认不填写（即不压缩）。支持压缩类型为gzip、bzip2和zip。	否	*不压缩*
encoding	读取文件的编码配置。	否	*utf-8*
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据同步系统提供nullFormat定义哪些字符串可以表示为null。例如，您配置 `nullFormat="null"` ，那么如果源头数据是 `"null"` ，数据同步系统会视作null字段。针对空字符串，需要加一层转义： `\N=\\N` 。	否	无
skipHeader	类CSV格式文件可能存在表头为标题情况，需要跳过。默认不跳过，压缩文件模式下不支持skipHeader。	否	*false*
csvReaderConfig	读取CSV类型文件参数配置，Map类型。读取CSV类型文件使用的CsvReader进行读取，会有很多配置，不配置则使用默认值。	否	无

向导开发介绍

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
文件名（含路径）	即上述参数说明中的Object。 **说明** 假如您的OSS文件名有根据每天的时间命名的部分，例如aaa/20171024abc.txt，关于Object系统参数可以设置为aaa/${bdp.system.bizdate}abc.txt。
列分隔符	即上述参数说明中的fieldDelimiter，默认值为（,）。
编码格式	即上述参数说明中的encoding，默认值为utf-8。
null值	即上述参数说明中的nullFormat，将要表示为空的字段填入文本框，如果源端存在则将对应的部分转换为空。
压缩格式	即上述参数说明中的compress，默认值为不压缩。
是否包含表头	即上述参数说明中的skipHeader，默认值为No。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下所示，具体参数填写请参见参数说明。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"oss",//插件名。
            "parameter":{
                "nullFormat":"",//定义可以表示为null的字符串。
                "compress":"",//文本压缩类型。
                "datasource":"",//数据源。
                "column":[//字段。
                    {
                        "index":0,//列序号。
                        "type":"string"//数据类型。
                    },
                    {
                        "index":1,
                        "type":"long"
                    },
                    {
                        "index":2,
                        "type":"double"
                    },
                    {
                        "index":3,
                        "type":"boolean"
                    },
                    {
                        "format":"yyyy-MM-dd HH:mm:ss", //时间格式。
                        "index":4,
                        "type":"date"
                    }
                ],
                "skipHeader":"",//类CSV格式文件可能存在表头为标题情况，需要跳过。
                "encoding":"",//编码格式。
                "fieldDelimiter":",",//字段分隔符。
                "fileFormat": "",//文本类型。
                "object":[]//object前缀。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
```

```
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":""//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

ORC或Parquet文件读取OSS

目前通过复用HDFS Reader的方式完成OSS读取ORC或Parquet格式的文件，在OSS Reader已有参数的基础上，增加了Path、FileFormat等扩展配置参数。

- 以ORC文件格式读取OSS，示例如下。

```
{
      "stepType": "oss",
      "parameter": {
        "datasource": "",
        "fileFormat": "orc",
        "path": "/tests/case61/orc__691b6815_9260_4037_9899_****",
        "column": [
          {
            "index": 0,
            "type": "long"
          },
          {
            "index": "1",
            "type": "string"
          },
          {
            "index": "2",
            "type": "string"
          }
        ]
      }
    }
```

- 以Parquet文件格式读取OSS，示例如下。

```
{
"type": "job",
"version": "2.0",
"steps": [
{
"stepType": "oss",
"parameter": {
```

```
"nullFormat": "",
"compress": "",
"fileFormat": "parquet",
"path": "/*",
"parquetSchema": "message m { optional BINARY registration_dttm (UTF8); optional Int64 id; optional BINARY
first_name (UTF8); optional BINARY last_name (UTF8); optional BINARY email (UTF8); optional BINARY gender (
UTF8); optional BINARY ip_address (UTF8); optional BINARY cc (UTF8); optional BINARY country (UTF8); option
al BINARY birthdate (UTF8); optional DOUBLE salary; optional BINARY title (UTF8); optional BINARY comments
(UTF8); }",
"column": [
{
"index": "0",
"type": "string"
},
{
"index": "1",
"type": "long"
},
{
"index": "2",
"type": "string"
},
{
"index": "3",
"type": "string"
},
{
"index": "4",
"type": "string"
},
{
"index": "5",
"type": "string"
},
{
"index": "6",
"type": "string"
},
{
"index": "7",
"type": "string"
},
{
"index": "8",
"type": "string"
},
{
"index": "9",
"type": "string"
},
{
"index": "10",
"type": "double"
},
{
"index": "11",
"type": "string"
},
{
"index": "12",
"type": "string"
}
],
"skipHeader": "false",
```

```
"encoding": "UTF-8",
"fieldDelimiter": ",",
"fieldDelimiterOrigin": ",",
"datasource": "wpw_demotest_oss",
"envType": 0,
"object": [
"wpw_demo/userdata1.parquet"
]
},
"name": "Reader",
"category": "reader"
},
{
"stepType": "odps",
"parameter": {
"partition": "dt=${bizdate}",
"truncate": true,
"datasource": "0_odps_wpw_demotest",
"envType": 0,
"column": [
"id"
],
"emptyAsNull": false,
"table": "wpw_0827"
},
"name": "Writer",
"category": "writer"
}
],
"setting": {
"errorLimit": {
"record": ""
},
"locale": "zh_CN",
"speed": {
"throttle": false,
"concurrent": 2
}
},
"order": {
"hops": [
{
"from": "Reader",
"to": "Writer"
}
]
}
}
```


6.2.10. FTP Reader

本文为您介绍FTP Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

FTP Reader为您提供读取远程FTP文件系统数据存储的功能。在底层实现上，FTP Reader获取远程FTP文件数据，并转换为数据同步传输协议传递给Writer。

本地文件内容存放的是一张逻辑意义上的二维表，例如CSV格式的文本信息。

FTP Reader实现了从远程FTP文件读取数据并转为数据同步协议的功能，远程FTP文件本身是无结构化数据存储。对于数据同步而言，目前FTP Reader支持的功能如下所示：

- 支持且仅支持读取TXT的文件，并要求TXT中的schema为一张二维表。
- 支持类CSV格式文件，自定义分隔符。

- 支持多种类型数据读取（使用STRING表示）、支持列裁剪和列常量。
- 支持递归读取、支持文件名过滤。
- 支持文本压缩，现有压缩格式为gzip、bzip2、zip、lzo和lzo_deflate。
- 多个File可以支持并发读取。

暂时不支持以下功能：

- 单个File支持多线程并发读取，此处涉及到单个File内部切分算法。
- 单个File在压缩情况下，从技术上无法支持多线程并发读取。

类型转换列表

远程FTP文件本身不提供数据类型，该类型是DataX FtpReader定义。

DataX内部类型	远程FTP文件数据类型
LONG	LONG
DOUBLE	DOUBLE
STRING	STRING
BOOLEAN	BOOLEAN
DATE	DATE

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
path	远程FTP文件系统的路径和文件名信息，需要填写包含路径和文件后缀的完整文件路径和文件名。这里可以支持填写多个路径。 • 当指定单个远程FTP文件，FTP Reader暂时只能使用单线程进行数据抽取。后期会在非压缩文件情况下针对单个File进行多线程并发读取。 • 当指定多个远程FTP文件，FTP Reader支持使用多线程进行数据抽取。线程并发数通过通道数指定。 • 当指定通配符，FTP Reader尝试遍历出多个文件信息。例如，指定/代表读取/目录下所有的文件，指定/bazhen/代表读取bazhen目录下游所有的文件。FTP Reader目前仅支持星号（*）作为文件通配符，但支持使用自定义参数配合调度，灵活生成任务名。 **说明** • 通常不建议您使用星号（*），易导致任务运行报JVM内存溢出的错误。 • 数据同步会将一个作业下同步的所有Text File视作同一张数据表。您必须自己保证所有的File能够适配同一套Schema信息。 • 您必须保证读取文件为类CSV格式，并且提供给数据同步系统权限可读。 • 如果Path指定的路径下没有符合匹配的文件抽取，同步任务将报错。	是	无

参数	描述	是否必选	默认值
column	读取字段列表，type指定源数据的类型，index指定当前列来自于文本第几列（以0开始），value指定当前类型为常量，不从源头文件读取数据，而是根据value值自动生成对应的列。 默认情况下，您可以全部按照STRING类型读取数据，配置为 `"column":["*"]` 。您可以指定column字段信息，配置如下。 `{ "type": "long", "index": 0 //从远程FTP文件文本第一列获取INT字段。 }, { "type": "string", "value": "alibaba" //从FTP Reader内部生成alibaba的字符串字段作为当前字段。 }` 对于您指定的column信息，type必须填写，index和value必须选择其一。	是	无
fieldDelimiter	读取的字段分隔符。 **说明** FTP Reader在读取数据时，需要指定字段分割符，如果不指定会默认为（,），界面配置也会默认填写（,）。	是	,
skipHeader	类CSV格式文件可能存在表头为标题情况，需要跳过。默认不跳过，压缩文件模式下不支持skipHeader。	否	false
encoding	读取文件的编码配置。	否	*utf-8*
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据同步提供nullFormat定义哪些字符串可以表示为null。 例如，您配置 `nullFormat:"null"` ，如果源头数据是null，则数据同步视作null字段。	否	无
markDoneFileName	标档文件名，数据同步前检查标档文件。如果标档文件不存在，等待一段时间重新检查标档文件，如果检查到标档文件开始执行同步任务。	否	无
maxRetryTime	表示检查标档文件重试次数，默认重试60次，每一次重试间隔为1分钟，共60分钟。	否	*60*
csvReaderConfig	读取CSV类型文件参数配置，Map类型。读取CSV类型文件使用的CsvReader进行读取，会有很多配置，不配置则使用默认值。	否	无
fileFormat	读取的文件类型，默认情况下文件作为csv格式文件进行读取，内容被解析为逻辑上的二维表结构处理。如果您配置为binary，则表示按照纯粹二进制格式进行复制传输。 通常在FTP、OSS等存储之间进行目录结构对等复制时使用，通常无需配置该项。	否	无

向导开发介绍

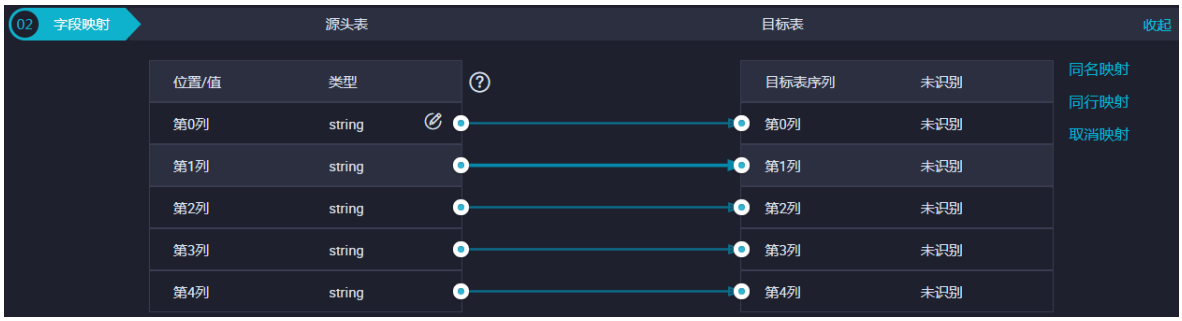
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
文件路径	即上述参数说明中的path，需要填写包含路径和文件后缀的完整文件路径和文件名。
文本类型	读取的文件类型，默认情况下文件作为csv格式文件进行读取。
列分隔符	即上述参数说明中的fieldDelimiter，默认值为（,）。
编码格式	即上述参数说明中的encoding，默认值为*utf-8*。
null值	即上述参数说明中的nullFormat，定义表示null值的字符串。
压缩格式	文本压缩类型，默认值为不压缩。
是否包含表头	即上述参数说明中的skipHeader，默认值为*No*。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，将鼠标放至需要删除的字段上，即可单击**删除**按钮进行删除。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从FTP数据库同步抽取数据作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"ftp",//插件名。
            "parameter":{
                "path":[],//文件路径。
                "nullFormat":"",//null值。
                "compress":"",//压缩格式。
                "datasource":"",//数据源。
                "column":[//字段。
                    {
                        "index":0,//序列号。
                        "type":""//字段类型。
                    }
                ],
                "skipHeader":"",//是否包含表头。
                "fieldDelimiter":",",//列分隔符。
                "encoding":"UTF-8",//编码格式。
                "fileFormat":"csv"//文本类型。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
        "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.11. Table Store（OTS） Reader

本文为您介绍OTS Reader支持的数据类型、读取方式、字段映射和数据源等参数及配置示例。

OTS Reader插件实现了从Table Store（OTS）读取数据，通过您指定的抽取数据范围，可以方便地实现数据增量抽取的需求。
目前支持以下三种抽取方式：

- 全表抽取
- 范围抽取
- 指定分片抽取

Table Store是构建在阿里云飞天分布式系统之上的NoSQL数据库服务，提供海量结构化数据的存储和实时访问。Table Store以实例和表的形式组织数据，通过数据分片和负载均衡技术，实现规模上的无缝扩展。

OTS Reader通过Table Store官方Java SDK连接到Table Store服务端，获取并按照数据同步官方协议标准转为数据同步字段信息传递给下游Writer端。

OTS Reader会根据Table Store的表范围，按照数据同步并发的数目N，将范围等分为N份Task。每个Task都会有一个OTS Reader线程来执行。

目前OTS Reader支持所有Table Store类型，OTS Reader针对Table Store的类型转换表，如下所示。

类型分类	Table Store数据类型
整数类	INTEGER
浮点类	DOUBLE
字符串类	STRING
布尔型	BOOLEAN
二进制类	BINARY

> **说明** Table Store本身不支持日期型类型。应用层通常使用Long报错时间的Unix TimeStamp。

参数说明

参数	描述	是否必选	默认值
endpoint	OTS Server的EndPoint（服务地址），详情请参见服务地址。	是	无
accessId	Table Store的AccessKey ID。	是	无
accessKey	Table Store的AccessKey Secret。	是	无
instanceName	Table Store的实例名称，实例是您使用和管理Table Store服务的实体。 您在开通Table Store服务后，需要通过管理控制台来创建实例，然后在实例内进行表的创建和管理。 实例是Table Store资源管理的基础单元，Table Store对应用程序的访问控制和资源计量都在实例级别完成。	是	无
table	所选取的需要抽取的表名称，这里有且只能填写一张表。在Table Store不存在多表同步的需求。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息。由于Table Store本身是NoSQL系统，在OTS Reader抽取数据过程中，必须指定相应的字段名称。 • 支持普通的列读取，例如{"name":"col1"} • 支持部分列读取，如果您不配置该列，则OTS Reader不予读取。 • 支持常量列读取，例如{"type":"STRING","value":"DataX"}。使用type描述常量类型，目前支持String、Int、Double、Bool、Binary（使用Base64编码填写）、INF_MIN（Table Store的系统限定最小值，如果使用该值，您不能填写value属性，否则报错）、INF_MAX（Table Store的系统限定最大值，如果使用该值，您不能填写value属性，否则报错）。 • 不支持函数或者自定义表达式，由于Table Store本身不提供类似SQL的函数或者表达式功能，OTS Reader也不能提供函数或表达式列功能。	是	无

<table>
<tr><th>参数</th><th>描述</th><th>是否必选</th><th>默认值</th></tr>
<tr><td>begin和end</td><td>
begin和end配置项必须配对使用，用于表示抽取Table Store表数据的范围。

begin和end描述的是Table Store PrimaryKey的区间分布状态，并且必须保证该配置项覆盖到目标表所有类型的PrimaryKey。对于无限大小的区间，您可以使用 <code>{"type":"INF_MIN"}</code> 和 <code>{"type":"INF_MAX"}</code> 分别指代begin和end，其中type表示抽取数据的类型。

说明
• 请确保主键个数和begin、end配置项对应。例如，Table Store中有n（n大于等于1）个主键，则begin和end分别有n个对应的取值。
• 如果对包含多个主键的Table Store表抽取数据，当抽取数据时扫描到第一个主键的区间范围为 <code>（INF_MIN，INF_MAX）</code> 时，会对Table Store表进行全表数据抽取，并且不会再匹配其他主键进行数据抽取。

例如，对一张主键为 <code>[DeviceID，SellerID]</code> 的双主键Table Store表进行数据抽取，begin和end的配置如下所示。
• 示例一：
抽取DeviceID的范围为 <code>（INF_MIN，INF_MAX）</code> 、SellerID的范围为 <code>（0，9999）</code> 并且数据类型为INT的数据，配置如下。
<pre>"range": {
      "begin": [
        {"type":"INF_MIN"}, //指定抽取DeviceID的最小值。
        {"type":"INT", "value":"0"} //指定抽取SellerID的最小值。
      ],
      "end": [
        {"type":"INF_MAX"}, //指定抽取DeviceID的最大值。
        {"type":"INT", "value":"9999"} //指定抽取SellerID的最大值。
      ]
    }</pre>
• 示例二：
抽取Table Store表的全表数据，配置如下。
<pre>"range": {
      "begin": [
        {"type":"INF_MIN"}, //指定抽取deviceID的最小值。
        {"type":"INF_MIN"} //指定抽取SellerID的最小值。
      ],
      "end": [
        {"type":"INF_MAX"}, //指定抽取deviceID的最大值。
          {"type":"INF_MAX"} //指定抽取SellerID的最大值。
      ]
    }</pre>
</td><td>是</td><td>空</td></tr>
</table>

参数	描述	是否必选	默认值
split	该配置项属于高级配置项，是您自己定义切分配置信息，普通情况下不建议使用。 适用场景：通常在Table Store数据存储发生热点，使用OTS Reader自动切分的策略不能生效的情况下，使用您自定义的切分规则。 split指定在Begin、End区间内的切分点，且只能是partitionKey的切分点信息，即在split仅配置partitionKey，而不需要指定全部的PrimaryKey。 如果对一张主键为 `[DeviceID, SellerID]` 的Table Store进行抽取任务，配置如下。 `"range": { "begin": { {"type":"INF_MIN"}, //指定DeviceID最小值。 {"type":"INF_MIN"} //指定SellerID最小值。 }, "end": { {"type":"INF_MAX"}, //指定DeviceID抽取最大值。 {"type":"INF_MAX"} //指定SellerID抽取最大值。 }, // 您指定的切分点。如果指定了切分点，Job将按照begin、end和split进行Task的切分。切分的列只能是Partition Key（ParimaryKey的第一列）。 //支持INF_MIN、INF_MAX、STRING和INT。 "split":[{"type":"STRING", "value":"1"}, {"type":"STRING", "value":"2"}, {"type":"STRING", "value":"3"}, {"type":"STRING", "value":"4"}, {"type":"STRING", "value":"5"}] }`	否	无

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

配置一个从Table Store同步抽取数据到本地的作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"ots",//插件名。
            "parameter":{
                "datasource":"",//数据源。
                "column":[//字段。
                    {
                        "name":"column1"//字段名。
                    },
                    {
                        "name":"column2"
                    },
                    {
                        "name":"column3"
                    },
                    {
                        "name":"column4"
                    },
                    {
```

```
                    {
                        "name":"column5"
                    }
                ],
                "range":{
                    "split":[
                        {
                            "type":"INF_MIN"
                        },
                        {
                            "type":"STRING",
                            "value":"splitPoint1"
                        },
                        {
                            "type":"STRING",
                            "value":"splitPoint2"
                        },
                        {
                            "type":"STRING",
                            "value":"splitPoint3"
                        },
                        {
                            "type":"INF_MAX"
                        }
                    ],
                    "end":[
                        {
                            "type":"INF_MAX"
                        },
                        {
                            "type":"INF_MAX"
                        },
                        {
                            "type":"STRING",
                            "value":"end1"
                        },
                        {
                            "type":"INT",
                            "value":"100"
                        }
                    ],
                    "begin":[
                        {
                            "type":"INF_MIN"
                        },
                        {
                            "type":"INF_MIN"
                        },
                        {
                            "type":"STRING",
                            "value":"begin1"
                        },
                        {
                            "type":"INT",
                            "value":"0"
                        }
                    ]
                },
                "table":""//表名。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
```

```
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//false代表不限流，下面的限流的速度不生效，true代表限流。
            "concurrent":1 //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.12. AnalyticDB for MySQL 3.0 Reader

本文为您介绍AnalyticDB for MySQL 3.0 Reader支持的数据类型、字段映射和数据源等参数及配置示例。

AnalyticDB for MySQL 3.0 Reader插件实现了从AnalyticDB for MySQL 3.0读取数据。在底层实现上，AnalyticDB for MySQL 3.0 Reader通过JDBC连接远程AnalyticDB for MySQL 3.0数据库，并执行相应的SQL语句，从AnalyticDB for MySQL 3.0库中读取数据。

数据类型转换

AnalyticDB for MySQL 3.0 Reader针对AnalyticDB for MySQL 3.0类型的转换列表，如下表所示。

类型分类	AnalyticDB for MySQL 3.0类型
整数类	INT、INTEGER、TINYINT、SMALLINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔类	BOOLEAN

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	所选取的需要同步的表。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息，默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表组织结构信息的顺序进行导出。 • 支持常量配置，您需要按照MySQL的语法格式，例如 `["id", "`table`", "1", "'bazhen.csy'", "null", "to_char(a + 1)", "2.3" , "true"]` 。 ○ id为普通列名。 ○ table包含保留的列名。 ○ 1为整型数字常量。 ○ bazhen.csy为字符串常量。 ○ null为空指针。 ○ to_char(a + 1)为计算字符串长度函数表达式。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示您指定同步的列集合，不允许为空 。	是	无
splitPk	AnalyticDB for MySQL 3.0 Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点和日期等其他类型 。如果您指定其他非支持类型，忽略splitPk功能，使用单通道进行同步。 • 如果不填写splitPk，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据 。	否	无
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。如果不填写where语句，包括不提供where的key或value，数据同步均视作同步全量数据。 • 不可以将where条件指定为limit 10，这不符合MySQL SQL WHERE子句约束。	否	无

向导开发介绍

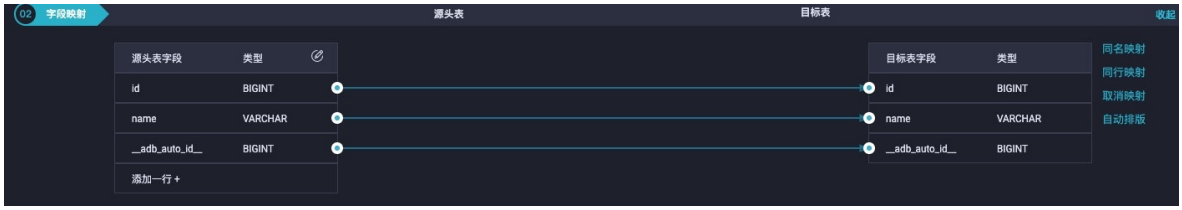
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	○ 可以输入常量，输入的值需要使用英文单引号，例如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。

参数	描述
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置示例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "analyticdb_for_mysql", //插件名。
            "parameter": {
                "column": [ //列名。
                    "id",
                    "value",
                    "table"
                ],
                "connection": [
                    {
                        "datasource": "xxx", //数据源。
                        "table": [ //表名。
                            "xxx"
                        ]
                    }
                ],
                "where": "", //过滤条件。
                "splitPk": "", //切分键。
                "encoding": "UTF-8" //编码格式。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0" //同步过程中的错误记录限制数。
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
                      "concurrent":1 //作业并发数。
                    "mbps":"12"//限流
        }
    }
}
```

6.2.13. ClickHouse Reader

本文为您介绍ClickHouse Reader支持的数据类型、字段映射和数据源等参数及配置示例。

> **说明** 仅支持阿里云ClickHouse。

ClickHouse Reader插件实现了从ClickHouse读取数据。在底层实现上，ClickHouse Reader通过JDBC连接远程ClickHouse数据库，并执行相应的SQL语句，从ClickHouse库中读取数据。

使用限制

- ClickHouse Reader仅支持使用独享数据集成资源组概述，不支持使用公共资源组和自定义资源组概述。
- ClickHouse Reader使用JDBC连接ClickHouse，且仅支持使用JDBC Statement读取数据。
- ClickHouse Reader支持筛选部分列、列换序等功能，您可以自行填写列。
- 您需要确认驱动和您的ClickHouse服务之间的兼容能力，数据库驱动使用如下版本。

```
<dependency>
    <groupId>ru.yandex.clickhouse</groupId>
    <artifactId>clickhouse-jdbc</artifactId>
    <version>0.2.4.ali2-SNAPSHOT</version>
</dependency>
```

背景信息

ClickHouse Reader面向ETL开发工程师，通过ClickHouse Reader从ClickHouse读取数据。在底层实现上，ClickHouse Reader通过JDBC连接远程ClickHouse数据库，ClickHouse Reader会根据您的配置生成相应的ClickHouse SQL查询语句，匹配数据集成框架各Writer的写入协议，并利用各引擎暴露的写入接口写入其他引擎。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	所选取的需要同步的表。使用JSON数据进行描述。 **说明** table必须包含在connection配置单元中。	是	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据同步系统和服务器端的网络交互次数，能够提升数据抽取性能。 **说明** fetchSize值过大会造成数据同步进程OOM，需要根据ClickHouse负载情况递增。	否	1,024
column	需要读取的ClickHouse数据，字段之间用英文逗号分隔。例如"column": ["id", "name", "age"]。 **说明** column配置项必须指定，不能为空。	是	无
jdbcUrl	到源端数据库的JDBC连接信息，jdbcUrl包含在connection配置单元中。 • 在一个数据库上只能配置一个值。 • jdbcUrl的格式和ClickHouse官方一致，并可以连接附加参数信息。例如：jdbc:clickhouse://localhost:3306/test?user=root&password=&useUnicode=true&characterEncoding=gbk&autoReconnect=true&failOverReadOnly=false。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无

参数	描述	是否必选	默认值
splitPk	ClickHouse进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 **说明** 当配置了splitPk时，fetchSize参数为必填项。	否	无
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 where条件可以有效地进行业务增量同步。如果不填写where语句，包括不提供where的key或value，数据同步均视作同步全量数据。	否	无

向导开发介绍

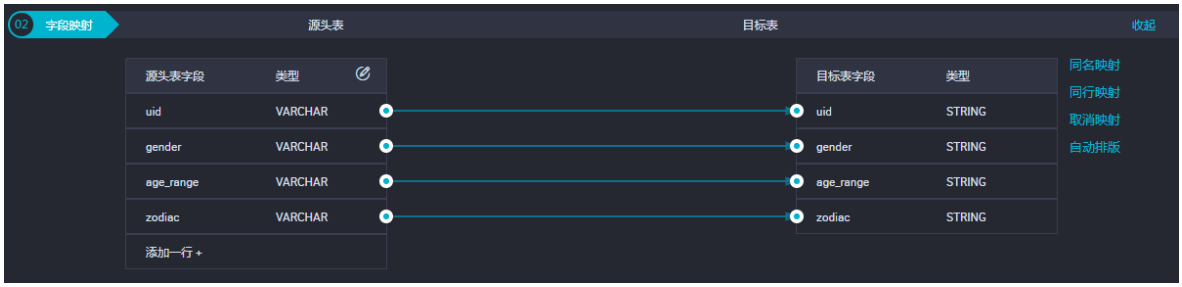
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	即上述参数说明中的where。您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	即上述参数说明中的splitPk。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段。鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其它空行会被忽略。
添加一行	添加一行的功能如下所示： ○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为**未识别**。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置示例如下，详情请参见上述参数说明。

> **说明** 实际执行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "clickhouse", //插件名。
            "parameter": {
                "fetchSize":1024,//该配置项定义了插件和数据库服务器端每次批量数据获取条数。
                "datasource": "example",
                "column": [   //列名。
                    "id",
                    "name"
                ],
                "where": "",    //过滤条件。
                "splitPk": "",  //切分键。
                "table": ""    //表名。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "clickhouse",
            "parameter": {
                "postSql": [
                    "update @table set db_modify_time = now() where db_id = 1"
                ],
                "datasource": "example",    //数据源。
                "batchByteSize": "67108864",
                "column": [
                    "id",
                    "name"
                ],
                "writeMode": "insert",
                "encoding": "UTF-8",
                "batchSize": 1024,
                "table": "ClickHouse_table",
                "preSql": [
                    "delete from @table where db_id = -1"
                ]
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "executeMode": null,
        "errorLimit": {
            "record": "0"  //同步过程中的错误记录限流数。
        },
        "speed": {
         "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.2.14. SQL Server Reader

本文为您介绍SQL Server Reader支持的数据类型、字段映射和数据源等参数及配置示例。

SQL Server Reader插件从SQL Server读取数据。在底层实现上，SQL Server Reader通过JDBC连接远程SQL Server数据库，并执行相应的SQL语句，从SQL Server库中读取数据。

SQL Server Reader通过JDBC连接器连接至远程的SQL Server数据库，根据您配置的信息生成查询SQL语句，发送至远程SQL Server数据库，执行该SQL并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

SQL Server Reader支持大部分SQL Server类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型。

SQL Server驱动版本支持列表

SQL Server Reader使用驱动版本是com.microsoft.sqlserver sqljdbc4 4.0，驱动能力具体请参见官网文档。该驱动支持的SQL Server版本如下所示：

版本	支持性（是/否）
SQL Server 2016	是
SQL Server 2014	是
SQL Server 2012	是
PDW 2008R2 AU34	是
SQL Server 2008 R2	是
SQL Server 2008	是
SQL Server 2019	否
SQL Server 2018	否

类型转换列表

SQL Server Reader针对SQL Server的类型转换列表，如下所示。

类型分类	SQL Server数据类型
整数类	BIGINT、INT、SMALLINT和TINYINT
浮点类	FLOAT、DECIMAL、REAL和NUMERIC
字符串类	CHAR、NCHAR、NTEXT、NVARCHAR、TEXT、VARCHAR、NVARCHAR（MAX）和VARCHAR（MAX）
日期时间类	DATE、DATETIME和TIME
布尔型	BIT
二进制类	BINARY、VARBINARY、VARBINARY（MAX）和TIMESTAMP

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称，一个作业只能支持一个表同步。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表schema信息顺序进行导出。 • 支持常量配置，您需要按照MySQL SQL语法格式，例如 `["id", "table","1", "'mingya.wmy'", "'null'", "to_char(a+1)", "2.3" , "true"]` 。 ○ id为普通列名。 ○ table为包含保留字的列名。 ○ 1为整型数字常量。 ○ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ○ 'null'为字符串。 ○ to_char(a + 1)为函数表达式。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	SQL Server Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片。数据同步系统因此会启动并发任务进行数据同步，这样可以提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其他类型 。如果您指定其他非支持类型，SQL Server Reader将报错。	否	无
where	筛选条件，SQL Server Reader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。例如在测试时，可以将where条件指定为limit 10。在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create > $bizdate` 。 • where条件可以有效地进行业务增量同步。 • where条件为空，视作同步全表所有的信息。	否	无
querySql	使用格式： `"querysql" : "查询statement",` 在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当配置此项后，数据同步系统就会忽略tables、columns配置项，直接使用这项配置的内容对数据进行筛选，例如需要进行多表join后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，SQL Server Reader直接忽略column、table和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1024*

- 对于您配置的table、column和where等信息，SQL Server Reader将其拼接为SQL语句发送至SQL Server数据库。
- 对于您配置的querySql信息，SQL Server直接将其发送至SQL Server数据库。

向导开发介绍

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table，选择需要同步的表。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从SQL Server数据库同步抽取数据的作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"sqlserver",//插件名。
            "parameter":{
                "datasource":"",//数据源。
                "column":[//字段。
                    "id",
                    "name"
                ],
                "where":"",//筛选条件。
                "splitPk":"",//如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片。
                "table":""//数据表。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

如果您想使用querySql查询，Reader部分脚本代码示例如下（SQL Server数据源是sql_server_source，待查询的表是dbo.test_table，待查询的列是name）。

```
{
    "stepType": "sqlserver",
    "parameter": {
        "querySql": "select name from dbo.test_table",
        "datasource": "sql_server_source",
        "column": [
            "name"
        ],
        "where": "",
        "splitPk": "id"
    },
    "name": "Reader",
    "category": "reader"
},
```

补充说明

- 主备同步数据恢复问题

  主备同步问题指SQL Server使用主从灾备，备库从主库不间断通过binlog恢复数据。由于主备数据同步存在一定的时间差，特别在于某些特定情况，例如网络延迟等问题，导致备库同步恢复的数据与主库有较大差别，从备库同步的数据不是一份当前时间的完整镜像。

- 一致性约束

  SQL Server在数据存储划分中属于RDBMS系统，对外可以提供强一致性数据查询接口。例如一次同步任务启动运行过程中，当该库存在其他数据写入方写入数据时，由于数据库本身的快照特性，SQL Server Reader完全不会获取到写入的更新数据。

  上述是在SQL Server Reader单线程模型下数据同步一致性的特性，SQL Server Reader可以根据您配置的信息使用并发数据抽取，因此不能严格保证数据一致性。

  当SQL Server Reader根据splitPk进行数据切分后，会先后启动多个并发任务完成数据同步。多个并发任务相互之间不属于同一个读事务，同时多个并发任务存在时间间隔，因此这份数据并不是完整的、一致的数据快照信息。

  针对多线程的一致性快照需求，目前在技术上无法实现，只能从工程角度解决。工程化的方式存在取舍，在此提供以下解决思路，您可以根据自身情况进行选择。

  - 使用单线程同步，即不再进行数据切片。缺点是速度比较慢，但是能够很好保证一致性。
  - 关闭其他数据写入方，保证当前数据为静态数据，例如锁表、关闭备库同步等。缺点是可能影响在线业务。

- 数据库编码问题

  SQL Server Reader底层使用JDBC进行数据抽取，JDBC天然适配各类编码，并在底层进行了编码转换。因此SQL Server Reader不需您指定编码，可以自动获取编码并转码。

- 增量数据同步的方式

  SQL Server Reader使用JDBC SELECT语句完成数据抽取工作，因此您可以使用`SELECT...WHERE...`进行增量数据抽取，方式如下：

  - 数据库在线应用写入数据库时，填充modify字段为更改时间戳，包括新增、更新、删除（逻辑删除）。对于该类应用，SQL Server Reader只需要where条件后跟上一同步阶段时间戳即可。
  - 对于新增流水型数据，SQL Server Reader在where条件后跟上一阶段最大自增ID即可。

  对于业务上无字段区分新增、修改数据的情况，SQL Server Reader无法进行增量数据同步，只能同步全量数据。

- SQL安全性

  SQL Server Reader提供querySql语句交给您自己实现SELECT抽取语句，SQL Server Reader本身对querySql不进行任何安全性校验。

6.2.15. Lindorm Reader

本文为您介绍Lindorm Reader支持的实现原理、参数定义及配置示例。

背景信息

LindormReader 插件提供了从Lindorm读取数据。在底层实现上，LindormReader通过Lindorm的Java客户端连接远程的Lindorm服务，并且支持通过对应的API读取Table或者WideColumn类型的数据，并将读取的数据使用DataX自定义的数据类型封装为抽象的数据集，传递给下游Writer处理。

> **说明**
>
> - LindormReader的必填配置项configuration，可以通过Lindorm集群控制台查看连接Lindorm的相关配置项进行获取，并以JSON格式填写相关信息。
> - Lindorm为多模数据库，LindormReader支持读取table和widecolumn类型的数据，关于table和widecolumn类型的详细介绍请参见Lindorm使用文档，您也可以通过钉钉咨询Lindorm值班人员。

使用限制

Lindorm Reader仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

类型转换

Lindorm Reader支持大部分Lindorm类型，但也存在个别没有支持的情况，请注意检查您的数据类型。

Lindorm Reader针对Lindorm类型的转换列表，如下所示。

类型分类	数据类型
整数类	INT、LONG、SHORT
浮点类	DOUBLE、FLOAT、DOUBLE
字符串类	STRING
日期时间类	DATE
布尔类	BOOLEAN
二进制类	BINARYSTRING

参数说明

参数	描述	是否必选	默认值
configuration	表示每个lindorm集群提供给DataX客户端连接的配置信息，可以通过lindorm集群控制台查询，获取到配置信息后可以联系lindorm数据库管理员将其转换为如下JSON格式：*{"key1":"value1","key2":"value2"}*。 例如：*{"lindorm.zookeeper.quorum":"????","lindorm.zookeeper.property.clientPort":"????"}* **说明** 如果是手工编写的JSON代码，则需要将JSON格式中*value*值的双引号转义为\"。	是	无
mode	表示数据读取模式，包括固定列模式FixedColumn和动态列模式DynamicColumn。默认选择FixedColumn。	是	FixedColumn
tablemode	包括普通表模式table和宽表模式wideColumn。默认为table，如果选择table模式，可不填写。	否	默认不填写
table	表示所要读取的lindorm表名。lindorm表名对大小写敏感。	是	无
namespace	表示所要读取的lindorm表的命名空间。lindorm表的命名空间对大小写敏感。	是	无
encoding	编码方式，取值为UTF-8或GBK。一般用于将二进制存储的lindorm byte[]类型转换为String类型。	否	UTF-8
selects	当前读取的Table类型数据不支持自动切割分片，默认单并发运行，因此需要手动配置selects参数进行数据切片，例如： `selects": [ "where(compare(\"id\", LESS, 5))", "where(and(compare(\"id\", GREATER_OR_EQUAL, 5), compare(\"id\", LESS, 10)))", "where(compare(\"id\", GREATER_OR_EQUAL, 10))" ],`	否	无

参数	描述	是否必选	默认值
columns	读取字段列表。读取字段列表支持列裁剪和列换序，列裁剪指可以选择部分列进行导出，列换序指可以不按照表schema信息顺序进行导出。 • table类型的表，只需要填写列名即可，会自动从表的meta获取schema信息。示例如下： table**类型**： ["id", "name", "age", "birthday", "gender"] • widecolumn类型的表。示例如下： Widecolumn**类型**： ["STRING|rowkey", "INT|f:a", "DOUBLE|f:b"]	是	无

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

- 配置一个Lindorm Table（对应SDK中的TableService模型）抽取数据到本地的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "lindorm",
            "parameter": {
                "mode": "FixedColumn",
            "caching": 128,
                "configuration": {    //lindorm控制台中与连接相关的配置项，以JSON格式填写
                    "lindorm.client.username": "",
                    "lindorm.client.seedserver": "seddserver.et2sqa.tbsite.net:30020",
                    "lindorm.client.namespace": "namespace",
                    "lindorm.client.password": ""
                },
                "columns": [
                    "id",
                    "name",
                    "age",
                    "birthday",
                    "gender"
                ],
                "envType": 1,
                "datasource": "_LINDORM",
                "namespace": "namespace",
                "table": "lindorm_table"
```

```
            },
            "name": "lindormreader",
            "category": "reader"
        },
        {
            "stepType": "mysql",
            "parameter": {
                "postSql": [],
                "datasource": "_IDB.TAOBAO",
                "session": [],
                "envType": 1,
                "columns": "columns": [
                    "id",
                    "name",
                    "age",
                    "birthday",
                    "gender"
                ],
         "selects": [
                    "where(compare(\"id\", LESS, 5))",
                    "where(and(compare(\"id\", GREATER_OR_EQUAL, 5), compare(\"id\", LESS, 10)))",
                    "where(compare(\"id\", GREATER_OR_EQUAL, 10))"
                ],
                "socketTimeout": 3600000,
                "guid": "",
                "writeMode": "insert",
                "batchSize": 1024,
                "encoding": "UTF-8",
                "table": "",
                "preSql": []
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "jvmOption": "",
        "executeMode": null,
        "errorLimit": {
            "record": "0"
        },
        "speed": {
        //设置传输速度，单位为byte/s，DataX运行会尽可能达到该速度但是不超过它.
        "byte": 1048576
      }
      //出错限制
      "errorLimit": {
        //出错的record条数上限，当大于该值即报错。
        "record": 0,
        //出错的record百分比上限 1.0表示100%，0.02表示2%
        "percentage": 0.02
      }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

- 配置一个Lindorm wideColumn（对应SDK中的WideColumnService模型）抽取数据到本地的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "lindorm",
            "parameter": {
                "mode": "FixedColumn",
                "configuration": {  //lindorm控制台中与连接相关的配置项，以JSON格式填写
                    "lindorm.client.username": "",
                    "lindorm.client.seedserver": "seddserver.et2sqa.tbsite.net:30020",
                    "lindorm.client.namespace": "namespace",
                    "lindorm.client.password": ""
                },
                "columns": "columns": [
                   "STRING|rowkey",
                      "INT|f:a",
                      "DOUBLE|f:b"
                ],
                "envType": 1,
                "datasource": "_LINDORM",
                "namespace": "namespace",
                "table": "wideColumn"
            },
            "name": "lindormreader",
            "category": "reader"
        },
        {
            "stepType": "mysql",
            "parameter": {
                "postSql": [],
                "datasource": "_IDB.TAOBAO",
                "session": [],
                "envType": 1,
                "column": [
                    "id",
                    "value"
                ],
                "socketTimeout": 3600000,
                "guid": "",
                "writeMode": "insert",
                "batchSize": 1024,
                "encoding": "UTF-8",
                "table": "",
                "preSql": []
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "jvmOption": "",
        "executeMode": null,
        "errorLimit": {
            "record": "0"
        },
        "speed": {
        //设置传输速度，单位为byte/s，DataX运行会尽可能达到该速度但是不超过它。
        "byte": 1048576
```

```
            "byte": 1048576
        }
          //出错限制
          "errorLimit": {
          //出错的record条数上限，当大于该值即报错。
          "record": 0,
          //出错的record百分比上限 1.0表示100%，0.02表示2%。
          "percentage": 0.02
        }
      },
      "order": {
          "hops": [
              {
                  "from": "Reader",
                  "to": "Writer"
              }
          ]
      }
}
```

6.2.16. LogHub（SLS） Reader

本文为您介绍LogHub（SLS） Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

日志服务（Log Service）是针对实时数据的一站式服务，为您提供日志类数据采集、消费、投递及查询分析功能，全面提升海量日志的处理、分析能力。LogHub（SLS） Reader是使用日志服务的Java SDK消费LogHub（SLS）中的实时日志数据，并转换日志数据为数据集成传输协议传递给Writer。

实现原理

LogHub（SLS） Reader通过日志服务Java SDK消费LogHub（SLS）中的实时日志数据，具体使用的日志服务Java SDK版本，如下所示。

```
<dependency>
    <groupId>com.aliyun.openservices</groupId>
    <artifactId>aliyun-log</artifactId>
    <version>0.6.7</version>
</dependency>
```

日志库（Logstore）是日志服务中日志数据的采集、存储和查询单元，Logstore读写日志必定保存在某一个分区（Shard）上。每个日志库分为若干个分区，每个分区由MD5左闭右开区间组成，每个区间范围不会相互覆盖，并且所有的区间的范围是MD5整个取值范围，每个分区可以提供一定的服务能力：

- 写入：5 MB/s，2000 次/s。
- 读取：10 MB/s，100 次/s。

LogHub（SLS） Reader消费Shard中的日志，具体消费过程（GetCursor、BatchGetLog相关API）如下所示：

- 根据时间区间范围获得游标。
- 通过游标、步长参数读取日志，同时返回下一个位置游标。
- 不断移动游标进行日志消费。
- 根据Shard进行任务的切分并发执行。

类型转换列表

LogHub（SLS） Reader针对LogHub（SLS）类型的转换列表，如下所示。

数据集成内部类型	LogHub（SLS）数据类型
STRING	STRING

参数说明

参数	描述	是否必选	默认值
endPoint	日志服务入口endPoint是访问一个项目（Project）及其内部日志数据的URL。它和Project所在的阿里云地域（Region）及Project名称相关。各地域的服务入口请参见服务入口。	是	无
accessId	访问日志服务的访问密钥，用于标识用户。	是	无
accessKey	访问日志服务的访问密钥，用来验证用户的密钥。	是	无
project	目标日志服务的项目名称，是日志服务中的资源管理单元，用于隔离和控制资源。	是	无
logstore	目标日志库的名称，logstore是日志服务中日志数据的采集、存储和查询单元。	是	无
batchSize	一次从日志服务查询的数据条数。	否	*128*
column	每条数据中的列名，此处可以配置日志服务中的元数据作为同步列。日志服务支持日志主题、采集机器唯一标识、主机名、路径和日志时间等元数据。 **说明** 列名区分大小写。元数据的写法请参见日志服务机器组。	是	无
beginDateTime	数据消费的开始时间位点，即日志数据到达LogHub（SLS）的时间。该参数为时间范围（左闭右开）的左边界，yyyyMMddHHmmss格式的时间字符串（例如20180111013000），可以和DataWorks的调度时间参数配合使用。 例如，您在节点编辑页面右侧的**调度参数**中配置 `beginDateTime=${yyyymmdd-1}` ，则在**日志开始时间**处配置为${beginDateTime}000000，表示获取的日志开始时间为业务日期的0点0分0秒。 **说明** beginDateTime和endDateTime需要互相组合配套使用。	是	无
endDateTime	数据消费的结束时间位点，为时间范围（左闭右开）的右边界，yyyyMMddHHmmss格式的时间字符串（例如20180111013010），可以和DataWorks的调度时间参数配合使用。 例如，您在节点编辑页面右侧的**调度参数**中配置endDateTime=${yyyymmdd}，则在**日志结束时间**处配置为${endDateTime}000000，表示获取的日志结束时间为业务日期后一天的0点0分0秒。 **说明** 上一周期的endDateTime需要和下一周期的beginDateTime保持一致，或晚于下一周期的beginDateTime。否则，可能无法拉取部分区域的数据。	是	无

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
Logstore	目标日志库的名称。
日志开始时间	数据消费的开始时间位点，即日志数据到达LogHub（SLS）的时间。时间范围（左闭右开）的左边界，yyyyMMddHHmmss格式的时间字符串（例如20180111013000），可以和DataWorks的调度时间参数配合使用。
日志结束时间	数据消费的结束时间位点，时间范围（左闭右开）的右边界，yyyyMMddHHmmss格式的时间字符串（例如20180111013010），可以和DataWorks的调度时间参数配合使用。
批量条数	一次从日志服务查询的数据条数。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。

参数	描述
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下所示，具体参数填写请参见参数说明。

```
{
 "type":"job",
 "version":"2.0",//版本号。
 "steps":[
     {
         "stepType":"LogHub",//插件名。
         "parameter":{
             "datasource":"",//数据源。
             "column":[//字段。
                 "col0",
                 "col1",
                 "col2",
                 "col3",
                 "col4",
                 "C_Category",
                 "C_Source",
                 "C_Topic",
                 "C_MachineUUID", //日志主题。
                 "C_HostName", //主机名。
                 "C_Path", //路径。
                 "C_LogTime" //事件时间。
             ],
             "beginDateTime":"",//数据消费的开始时间位点。
             "batchSize":"",//一次从日志服务查询的数据条数。
             "endDateTime":"",//数据消费的结束时间位点。
             "fieldDelimiter":",",//列分隔符。
             "logstore":""//: 目标日志库的名字。
         },
         "name":"Reader",
         "category":"reader"
     },
     {
         "stepType":"stream",
         "parameter":{},
         "name":"Writer",
         "category":"writer"
     }
 ],
 "setting":{
     "errorLimit":{
         "record":"0"//错误记录数。
     },
     "speed":{
         "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
     }
 },
 "order":{
     "hops":[
         {
             "from":"Reader",
             "to":"Writer"
         }
     ]
 }
}
```

说明 如果元数据配置JSON中有tag前缀，需要删除tag前缀。例如，`__tag__:__client_ip__` 需要修改为 `__client_ip__` 。

6.2.17. OTSReader-Internal

本文为您介绍OTSReader-Internal支持的数据类型、字段映射和数据源等参数及配置示例。

表格存储Table Store（简称OTS）是构建在阿里云飞天分布式系统之上的NoSQL数据库服务，提供海量结构化数据的存储和实时访问。Table Store以实例和表的形式组成数据，通过数据分片和负载均衡技术，实现规模上的无缝扩展。

OTSReader-Internal主要用于OTS Internal模型的表数据导出，而另外一个插件OTS Reader则用于OTS Public模型的数据导出。

OTS Internal模型支持多版本，所以该插件提供两种模式数据的导出：

- 多版本模式：因为Table Store本身支持多版本，特此提供一个多版本模式，将多版本的数据导出。

  导出方案：Reader插件将Table Store的一个Cell展开为一个一维表的4元组，分别是主键（PrimaryKey，包含1-4列）、ColumnName、Timestamp和Value（原理和HBase Reader的多版本模式类似），将这个4元组作为Datax record中的4个Column传输给消费端（Writer）。
- 普通模式：和HBase Reader普通模式一致，只需导出每行数据中每列的最新版本的值，详情请参见HBase Reader中HBase Reader支持的normal模式内容。

OTS Reader通过Table Store官方Java SDK连接至OTS服务端，并通过SDK读取数据。OTS Reader本身对读取过程做了很多优化，包括读取超时重试、异常读取重试等。

目前OTS Reader支持所有Table Store类型，OTSReader-Internal针对Table Store类型的转换列表，如下所示。

数据集成内部类型	Table Store数据类型
LONG	INTEGER
DOUBLE	DOUBLE
STRING	STRING
BOOLEAN	BOOLEAN
BYTES	BINARY

参数说明

参数	描述	是否必选	默认值
mode	插件的运行方式，支持*normal*和*multiVersion*，分别表示普通模式和多版本模式。	是	无
endpoint	OTS Server的endpoint（服务地址）。	是	无
accessId	Table Store的accessId。	是	无
accessKey	Table Store的accessKey。	是	无
instanceName	Table Store的实例名称，实例是您使用和管理Table Store服务的实体。 您在开通Table Store服务后，需要通过管理控制台来创建实例，然后在实例内进行表的创建和管理。实例是Table Store资源管理的基础单元，Table Store对应用程序的访问控制和资源计量都在实例级别完成。	是	无
table	选取的需要抽取的表名称，这里有且只能填写一张表。在Table Store中不存在多表同步的需求。	是	无
range	导出的范围，读取的范围是[begin,end)，左闭右开的区间。 • begin小于end，表示正序读取数据。 • begin大于end，表示反序读取数据。 • begin和end不能相等。 • type支持的类型有STRING、INT和BINARY，BINARY输入的方式采用二进制的Base64字符串形式传入，INF_MIN表示无限小，INF_MAX表示无限大。	否	从表的开始位置读取到表的结束位置

参数	描述	是否必选	默认值
range:{"begin"}	导出的起始范围，这个值的输入可以填写空数组或PK前缀，也可以填写完整的PK。正序读取数据时，默认填充PK后缀为INF_MIN，反序为INF_MAX。 该配置是OTS主键的值范围，用于进行数据过滤。 如果没有配置开始的值，则默认最小值。 binary类型的PrimaryKey列比较特殊，因为JSON不支持直接输入二进制数，所以系统定义：如果您要传入二进制，必须使用（Java）Base64.encodeBase64String方法，将二进制转换为一个可视化的字符串，然后将这个字符串填入value中，Java示例如下： • `byte[] bytes = "hello".getBytes();`：构造一个二进制数据，这里使用字符串hello的byte值。 • `String inputValue = Base64.encodeBase64String(bytes)`：调用Base64方法，将二进制转换为可视化的字符串。 上面的代码执行之后，可以获得inputValue为 `"aGVsbG8="`。 最终写入配置 `{"type":"binary","value" : "aGVsbG8="}`。	否	从表的开始位置读取数据
range:{ "end" }	导出的结束范围，这个值的输入可以填写空数组或PK前缀，也可以填写完整的 PK。正序读取数据时，默认填充PK后缀为INF_MAX，反序为INF_MIN。 binary类型的PrimaryKey列比较特殊，因为JSON不支持直接输入二进制数，所以系统定义：如果您要传入二进制，必须使用（Java）Base64.encodeBase64String方法，将二进制转换为一个可视化的字符串，然后将这个字符串填入value中，Java示例如下。 • `byte[] bytes = "hello".getBytes();`：构造一个二进制数据，这里使用字符串hello的byte值。 • `String inputValue = Base64.encodeBase64String(bytes)`：调用Base64方法，将二进制转换为可视化的字符串。 上面的代码执行之后，可以获得inputValue为 `"aGVsbG8="`。 最终写入配置 `{"type":"binary", "value":"aGVsbG8="}`。	否	读取到表的结束位置
range:{"split"}	当前用户数据较多时，需要开启并发导出，Split可以将当前范围的数据按照切分点切分为多个并发任务。 **说明** • split中的输入值只能PK的第一列（分片建），且值的类型必须和PartitionKey一致。 • 值的范围必须在begin和end之间。 • split内部的值必须根据begin和end的正反序关系而递增或者递减。	否	空切分点
column	指定要导出的列，支持普通列和常量列。 格式（支持多版本模式） 普通列格式： `{"name":"{your column name}"}`	是	无
timeRange（仅支持多版本模式）	请求数据的Time Range，读取的范围为[begin,end)，左闭右开的区间。 **说明** begin必须小于end。	否	默认读取全部版本
timeRange:{"begin"}（仅支持多版本模式）	请求数据的Time Range开始时间，取值范围是0~LONG_MAX。	否	默认为0
timeRange:{"end"}（仅支持多版本模式）	请求数据的Time Range结束时间，取值范围是0~LONG_MAX。	否	默认为 *Long Max(9223372036854775806L)*

参数	描述	是否必选	默认值
maxVersion（仅支持多版本模式）	请求的指定Version，取值范围是1~INT32_MAX。	否	默认读取所有版本

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务：

- 多版本模式

```
{
  "type": "job",
  "version": "2.0",
  "steps": [
       {
       "stepType": "ots-internal",
            "parameter": {
                "mode": "multiVersion",
                "endpoint": "",
                "accessId": "",
                "accessKey": "",
                "instanceName": "",
                "table": "",
                "range": {
                    "begin": [
                        {
                            "type": "string",
                            "value": "a"
                        },
                        {
                            "type": "INF_MIN"
                        }
                    ],
                    "end": [
                        {
                            "type": "string",
                            "value": "g"
                        },
                        {
                            "type": "INF_MAX"
                        }
                    ],
                    "split": [
                        {
                            "type": "string",
                            "value": "b"
                        },
                        {
                            "type": "string",
                            "value": "c"
                        }
                    ]
                },
                "column": [
                    {
                        "name": "attr1"
                    }
                ],
                "timeRange": {
                    "begin": 1400000000,
                    "end": 1600000000
                },
                "maxVersion": 10
            }
        }
    },
    "writer": {}
]
}
```

- 普通模式

```
{
  "type": "job",
  "version": "2.0"
```

```
 "version": "2.0",
 "steps":
[
      {
          "stepType": "otsr-internal",
           "parameter": {
              "mode": "normal",
              "endpoint": "",
              "accessId": "",
              "accessKey": "",
              "instanceName": "",
              "table": "",
              "range": {
                  "begin": [
                      {
                          "type": "string",
                          "value": "a"
                      },
                      {
                          "type": "INF_MIN"
                      }
                  ],
                  "end": [
                      {
                          "type": "string",
                          "value": "g"
                      },
                      {
                          "type": "INF_MAX"
                      }
                  ],
                  "split": [
                      {
                          "type": "string",
                          "value": "b"
                      },
                      {
                          "type": "string",
                          "value": "c"
                      }
                  ]
              },
              "column": [
                  {
                      "name": "pk1"
                  },
                  {
                      "name": "pk2"
                  },
                  {
                      "name": "attr1"
                  },
                  {
                      "type": "string",
                      "value": ""
                  },
                  {
                      "type": "int",
                      "value": ""
                  },
                  {
                      "type": "double",
                      "value": ""
                  },
                  {
```

```
                        "type": "binary",
                        "value": "aGVsbG8="
                    }
                ]
            }
        }
    },
    "writer": {}
    ]
}
```

6.2.18. Stream Reader

本文为您介绍Stream Reader支持的数据类型、字段映射和数据源等参数及配置示例。

Stream Reader插件实现了从内存中自动产生数据的功能，主要用于数据同步的性能测试和基本的功能测试。

Stream Reader支持的数据类型，如下所示。

数据类型	类型描述
string	字符型
long	长整型
date	日期类型
bool	布尔型
bytes	字节型

参数说明

参数	描述	是否必选	默认值
column	产生的源数据的列数据和类型，可以配置多列。可以配置产生随机字符串，并制定范围，示例如下。 `"column" : [ { "random": "8,15" }, { "random": "10,10" } ]` 配置项说明如下： • "random": "8, 15"：表示随机产生8~15位长度的字符串。 • "random": "10, 10"：表示随机产生10位长度的字符串。	是	无
sliceRecordCount	表示循环产生column的份数。	是	无

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

配置一个从内存中读数据的同步作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",//插件名。
            "parameter":{
                "column":[//字段。
                    {
                        "type":"string",//数据类型。
                        "value":"field"//值。
                    },
                    {
                        "type":"long",
                        "value":100
                    },
                    {
                        "dateFormat":"yyyy-MM-dd HH:mm:ss",//时间格式。
                        "type":"date",
                        "value":"2014-12-12 12:12:12"
                    },
                    {
                        "type":"bool",
                        "value":true
                    },
                    {
                        "type":"bytes",
                        "value":"byte string"
                    }
                ],
                "sliceRecordCount":"100000"//表示循环产生column的份数。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//false代表不限流，下面的限流的速度不生效；true代表限流。
            "concurrent":1,//作业并发数。
           "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.19. HybridDB for MySQL Reader

本文为您介绍HybridDB for MySQL Reader支持的数据类型、字段映射和数据源等参数及配置示例。

HybridDB for MySQL Reader插件支持读取表和视图。表字段可以依序指定全部列、部分列、调整列顺序、指定常量字段和配置HybridDB for MySQL的函数，如now()等。

HybridDB for MySQL Reader插件从HybridDB for MySQL读取数据。在底层实现上，HybridDB for MySQL Reader通过JDBC连接远程HybridDB for MySQL数据库，并执行相应的SQL语句，从HybridDB for MySQL库中选取数据。

HybridDB for MySQL Reader插件通过JDBC连接器连接至远程的HybridDB for MySQL数据库，根据您配置的信息生成查询SQL语句，发送至远程HybridDB for MySQL数据库，执行该SQL语句并返回结果。然后使用数据集成支持的数据类型将其拼装为抽象的数据集，传递给下游Writer处理。

类型转换列表

HybridDB for MySQL Reader针对HybridDB for MySQL类型的转换列表，如下所示。

类型分类	HybridDB for MySQL数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔型	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

> **说明**
> - 除上述罗列字段类型外，其他类型均不支持。
> - HybridDB for MySQL Reader插件将tinyint（1）视作整型。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称，一个数据集成任务只能同步数据到一个目的表。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表Schema信息顺序进行导出。 • 支持常量配置，您需要按照SQL语法格式，例如 `["id", "table","1","'mingya.wmy'","'null'", "to_char(a+1)","2.3","true"]` 。 ○ id为普通列名。 ○ table为包含保留字的列名。 ○ 1为整型数字常量。 ○ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ○ 'null'为字符串常量。 ○ to_char(a+1)为计算字符串长度函数。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	HybridDB for MySQL Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，从而提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其他类型 。如果您指定其他非支持类型，忽略splitPk功能，使用单通道进行同步。 • 如果splitPk不填写，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据 。	否	无
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。如果不填写where语句，包括不提供where的key或value，数据同步均视作同步全量数据。 • 不可以将where条件指定为limit 10，不符合SQL WHERE子句约束。	否	无
querySql（高级模式，向导模式不提供）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL 。当配置此项后，数据同步系统就会忽略column、table和where配置项，直接使用该项配置的内容对数据进行筛选。例如需要进行多表join后同步数据，使用 `["id","table","1","'mingya.wmy'","'null'","to_char(a+1)","2.3","true"]` 。当您配置querySql时，HybridDB for MySQL Reader直接忽略column、table和where和splitPk条件的配置，querySql优先级大于table、column、where、splitPk选项。datasource会使用它解析出用户名和密码等信息。	否	无
singleOrMulti（只适合分库分表）	表示分库分表，向导模式转换成脚本模式主动生成此配置 `"singleOrMulti":"multi"` ，但配置脚本任务模板不会直接生成此配置，您需要手动添加，否则只会识别第一个数据源。singleOrMulti只是前端在用，后端没有用这个进行分库分表判断。	是	*multi*

向导模式

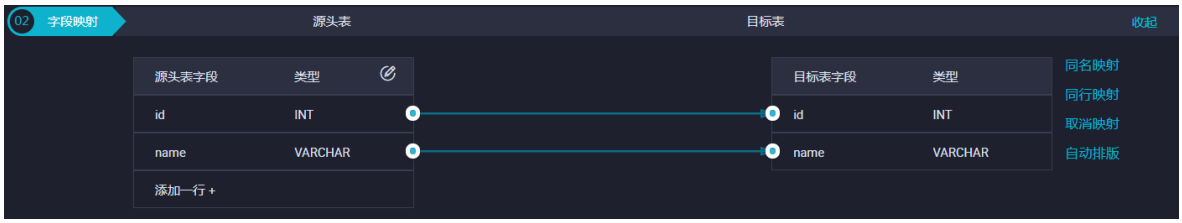
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键和数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，将鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。



参数	描述
同名映射	单击**同名映射**，即可根据名称建立相应的同行映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，即可在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，即可取消建立的映射关系。
自动排版	单击**自动排版**，即可根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。

参数	描述
添加一行	添加一行的功能如下所示： ○ 可以输入常量，输入的值需要使用英文单引号，例如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

单库单表的脚本样例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {
                "datasource": "px_aliyun_hymysql",//数据源名。
                "column": [//源端列名。
                    "id",
                    "name",
                    "sex",
                    "salary",
                    "age",
                    "pt"
                ],
                "where": "id=10001",//过滤条件。
                "splitPk": "id",//切分键。
                "table": "person"//源端表名。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "parameter": {}
    ],
    "version": "2.0",//版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "concurrent": 7,//并发数。
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps": 1 //限流值。
        }
    }
}
```

6.2.20. AnalyticDB for PostgreSQL Reader

本文为您介绍AnalyticDB for PostgreSQL Reader支持的数据类型、读取方式、字段映射和数据源等参数及配置示例。

AnalyticDB for PostgreSQL Reader插件从AnalyticDB for PostgreSQL读取数据。在底层实现上，AnalyticDB for PostgreSQL Reader通过JDBC连接远程AnalyticDB for PostgreSQL数据库，并执行相应的SQL语句，从AnalyticDB for PostgreSQL库中选取数据。RDS为您提供AnalyticDB for PostgreSQL存储引擎。

AnalyticDB for PostgreSQL Reader通过JDBC连接器连接至远程的AnalyticDB for PostgreSQL数据库，根据您配置的信息生成查询SQL语句，发送至远程AnalyticDB for PostgreSQL数据库，执行该SQL并返回结果。然后使用数据同步自定义的数据类型将其拼装为抽象的数据集，传递给下游Writer处理。

- 对于您配置的table、column和where等信息，AnalyticDB for PostgreSQL Reader将其拼接为SQL语句，发送至AnalyticDB for PostgreSQL数据库。
- 对于您配置的querySql信息，AnalyticDB for PostgreSQL直接将其发送至AnalyticDB for PostgreSQL数据库。

类型转换列表

AnalyticDB for PostgreSQL Reader支持大部分AnalyticDB for PostgreSQL类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型。

AnalyticDB for PostgreSQL Reader针对AnalyticDB for PostgreSQL的类型转换列表，如下所示。

类型分类	AnalyticDB for PostgreSQL数据类型
整数类	BIGINT、BIGSERIAL、INTEGER、SMALLINT和SERIAL
浮点类	DOUBLE、PRECISION、MONEY、NUMERIC和REAL
字符串类	VARCHAR、CHAR、TEXT、BIT和INET
日期时间类	DATE、TIME和TIMESTAMP
布尔型	BOOL
二进制类	BYTEA

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项输入的内容必须和添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表Schema信息顺序进行导出。 • 支持常量配置，您需要按照SQL语法格式。例如，`["id", "table","1","'mingya.wmy'","'null'", "to_char(a+1)","2.3","true"]`。 ○ id为普通列名。 ○ table为包含保留字的列名。 ○ 1为整型数字常量。 ○ ‘mingya.wmy’为字符串常量（注意需要加上一对单引号）。 ○ 'null'为字符串常量。 ○ to_char(a+1)为计算字符串长度函数。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	AnalyticDB for PostgreSQL Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片。数据同步会启动并发任务进行数据同步，以提高数据同步的效能。 • 通常表主键较为均匀，切分出的分片不易出现数据热点，所以推荐splitPk用户使用表主键。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其他类型。如果您指定其他非支持类型，忽略splitPk功能，使用单通道进行同步。 • 如果不填写splitPk，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据。	否	无

参数	描述	是否必选	默认值
where	筛选条件，AnalyticDB for PostgreSQLReader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。例如测试时，可以将where条件指定实际业务场景，往往会选择当天的数据进行同步，将where条件指定为 `id>2 and sex=1` 。 • where条件可以有效地进行业务增量同步。 • where条件不配置或者为空，视作全表同步数据。	否	无
querySql（高级模式，向导模式不提供）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当配置此项后，数据同步系统就会忽略column、table等配置项，直接使用该项配置的内容对数据进行筛选。例如需要进行多表join后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。 当您配置querySql时，AnalyticDB for PostgreSQL Reader直接忽略column、table和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*512*

向导开发介绍

1. 选择数据源。

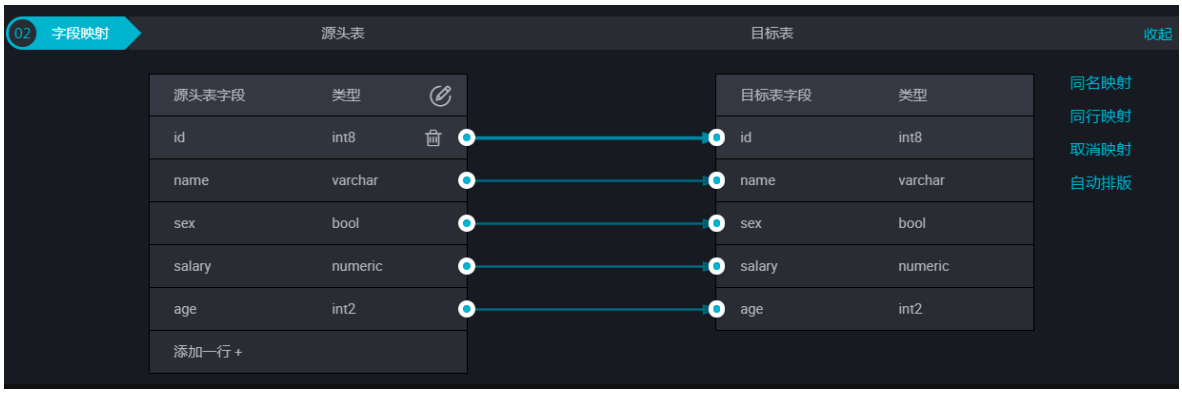
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table，选择需要同步的表。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤，SQL语法与选择的数据源一致。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键的设置跟数据同步里的选择来源有关，在配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，将鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. **通道控制**

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。

参数	描述
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {
                "datasource": "test_004",//数据源名称。
                "column": [//源端表的列名。
                    "id",
                    "name",
                    "sex",
                    "salary",
                    "age"
                ],
                "where": "id=1001",//过滤条件。
                "splitPk": "id",//切分键。
                "table": "public.person"//源端表名。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "parameter": {},
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",//版本号
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "concurrent": 6,//并发数。
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
           "mbps":"12"//限流
        }
    }
}
```

6.2.21. PolarDB Reader

本文为您介绍PolarDB Reader支持的数据类型、字段映射和数据源等参数及配置示例。

PolarDB Reader插件通过JDBC连接器连接至远程的PolarDB数据库，根据您配置的信息生成查询SQL语句，发送至远程PolarDB数据库，执行该SQL语句并返回结果。然后使用数据同步自定义的数据类型将其拼装为抽象的数据集，传递给下游Writer处理。

在底层实现上，PolarDB Reader插件通过JDBC连接远程PolarDB数据库，并执行相应的SQL语句，从PolarDB库中读取数据。

PolarDB Reader插件支持读取表和视图。表字段可以依序指定全部列、指定部分列、调整列顺序、指定常量字段和配置PolarDB的函数，例如now()等。

类型转换列表

PolarDB Reader针对PolarDB类型的转换列表，如下所示。

类型分类	PolarDB数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔型	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

> **说明**
> - 除上述罗列字段类型外，其它类型均不支持。
> - PolarDB Reader插件将tinyint（1）视作整型。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表Schema信息顺序进行导出。 • 支持常量配置，您需要按照SQL语法格式，例如 `["id", "table","1","'mingya.wmy'","'null'", "to_char(a+1)","2.3","true"]` 。 ○ id为普通列名。 ○ table为包含保留字的列名。 ○ 1为整型数字常量。 ○ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ○ 'null'为字符串常量。 ○ to_char(a+1)为计算字符串长度函数。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	PolarDB Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，从而提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其他类型 。如果您指定其他非支持类型，忽略plitPk功能，使用单通道进行同步。 • 如果splitPk不填写，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据 。	否	无
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。如果不填写where语句，包括不提供where的key或value，数据同步均视作同步全量数据。 • 将where条件指定为limit 10不符合WHERE子句约束，不建议使用。	否	无
querySql（高级模式，向导模式不提供）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当配置该项后，数据同步系统就会忽略column、table和where配置项，直接使用该项配置的内容对数据进行筛选。例如需要进行多表 join 后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，PolarDB Reader直接忽略column、table和where条件的配置，querySql优先级大于table、column、where、splitPk选项。datasource会使用它解析出用户名和密码等信息。	否	无

参数	描述	是否必选	默认值
singleOrMulti（只适合分库分表）	表示分库分表，向导模式转换成脚本模式会主动生成 `"singleOrMulti":"multi"` 配置，但脚本模式不会自动生成，您需要手动添加。如果不添加该配置，则仅识别第1个数据源。 **说明** 仅前端使用singleOrMulti，后端没有使用该参数判断分库分表。	是	*multi*

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键和数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，将鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

单库单表的脚本示例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {
                "datasource": "test_005",//数据源名。
                "column": [//源端列名。
                    "id",
                    "name",
                    "age",
                    "sex",
                    "salary",
                    "interest"
                ],
                "where": "id=1001",//过滤条件。
                "splitPk": "id",//切分键。
                "table": "PolarDB_person"//源端表名。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "parameter": {}
    ],
    "version": "2.0",//版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "concurrent": 6,//并发数。
            "throttle": true//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
      "mbps":"12",//限流
        }
    }
}
```

6.2.22. Elasticsearch Reader

本文为您介绍Elasticsearch Reader的工作原理、功能和参数。

使用限制

- DataWorks平台目前仅支持配置阿里云Elasticsearch5.x、6.x、7.x版本数据源，不支持配置自建Elasticsearch数据源。
- 不支持同步scaled_float类型的字段。

工作原理

Elasticsearch在公共资源组上支持Elasticsearch5.x版本，在独享数据集成资源组上支持Elasticsearch5.x、6.x和7.x版本。独享数据集成资源组的详情请参见新增和使用独享数据集成资源组。

Elasticsearch Reader的工作原理如下：

- 通过Elasticsearch的_searchscrollslice（即游标分片）方式实现，slice结合数据集成任务的task多线程分片机制使用。
- 根据Elasticsearch中的Mapping配置，转换数据类型。

更多详情请参见Elasticsearch官方文档。

基本配置

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"0" //错误记录数。
        },
        "jvmOption":"",
        "speed":{
            "concurrent":3,//并发数
            "throttle":true,//
                     "mbps":"12",//限流
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "column":[ //读取列。
                    "id",
                    "name"
                ],
                "endpoint":"", //服务地址。
                "index":"",  //索引。
                "password":"",  //密码。
                "scroll":"",  //scroll标志。
                "search":"",  //查询query参数，与Elasticsearch的query内容相同，使用_search api，重命名为search。
                "type":"default",
                "username":""  //用户名。
            },
            "stepType":"elasticsearch"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter":{ },
            "stepType":"stream"
        }
    ],
    "type":"job",
    "version":"2.0" //版本号。
}
```

高级功能

- 支持全量拉取

  支持将Elasticsearch中一个文档的所有内容拉取为一个字段。

- 支持提取半结构化到结构化数据

分类	描述
产生背景	Elasticsearch中的数据特征为字段不固定，且有中文名、数据使用深层嵌套的形式。为更好地方便下游业务对数据的计算和存储需求，特推出从半结构化到结构化的转换解决方案。
实现原理	将Elasticsearch获取到的JSON数据，利用JSON工具的路径获取特性，将嵌套数据扁平化为一维结构的数据。然后将数据映射至结构化数据表中，拆分Elasticsearch复合结构数据至多个结构化数据表。
解决方案	◦ JSON有嵌套的情况，通过path路径来解决： ▪ 属性 ▪ 属性.子属性 ▪ 属性[0].子属性 ◦ 附属信息有一对多的情况，需要进行拆表拆行处理，进行遍历。 属性[*].子属性 ◦ 数组归并，一个字符串数组内容，归并为一个属性，并进行去重。 属性[] 去重 ◦ 多属性合一，将多个属性合并为一个属性。 属性1,属性2 ◦ 多属性选择处理 属性1|属性2

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
index	Elasticsearch中的index名。	是	无
type	Elasticsearch中index的type名。	否	*index名*
search	Elasticsearch的query参数。	是	无
pageSize	每次读取数据的条数。	否	*100*
scroll	Elasticsearch的分页参数，设置游标存放时间。 • 设置的过小时，如果获取两页数据间隔时间超出scroll，会导致游标过期，进而丢失数据。 • 设置的过大时，如果同一时刻发起的查询过多，超出服务端 `max_open_scroll_context` 配置时，会导致数据查询报错。	是	无
sort	返回结果的排序字段。	否	无
retryCount	失败后重试的次数。	否	*300*
connTimeOut	客户端连接超时时间。	否	*600,000*
readTimeOut	客户端读取超时时间。	否	*600,000*
multiThread	http请求，是否有多线程。	否	*true*

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

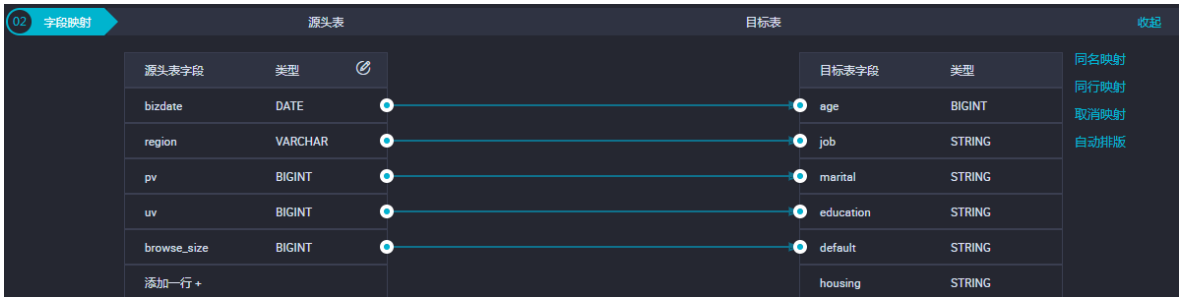
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	通常填写您配置的数据源名称。
索引	Elasticsearch中的index名。
检索查询条件	Elasticsearch的query参数。
分页大小	每次读取数据的条数，默认为100。
游标时间	分页参数，设置游标存放时间。
高级配置	**高级配置**包括以下内容： ○ **排序方式**：返回结果的排序字段。 ○ **全文作为一列**：是否将Elasticsearch的数据拉取为一个字段。 例如，Elasticsearch Reader需要读取Elasticsearch的所有数据作为一列同步至MaxCompute，则需要设置**全文作为一列**。设置Elasticsearch Reader中的column为content，content为hits[]中的一行信息_source全内容。 **说明** _id属性为Elasticsearch数据的固有属性，目前无法通过数据集成同步任务单独抽取并写入目的端。您可以将Elasticsearch Reader中**全文作为一列**参数配置为**是**（即JSON脚本中**full**参数配置为**true**），将Elasticsearch中的每个数据都作为一个字段同步到目的端，然后在目的端使用 `get_json_object` 函数或其他JSON处理函数，将 `_id` 值单独取出来做后续处理。 ○ **拆多行数组列名**：是否将数组进行列拆多行的处理，需要辅助设置子属性。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。

   说明 来源或目标端有Lindom、HBase、Tair、Elasticsearch数据源，字段无需连线，直接编辑即可保存。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从Elasticsearch读取数据的JSON示例，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"0" //错误记录数。
        },
        "jvmOption":"",
        "speed":{
            "concurrent":3,
            "throttle":false
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "column":[ //读取列。
                    "id",
                    "name"
                ],
                "endpoint":"http://es-cn-xxx.elasticsearch.aliyuncs.com:9200", //服务地址。
                "index":"aliyun_es_xx",  //索引。
                "password":"*******",  //密码。
                "multiThread":true,
                "scroll":"5m",  //scroll标志。
                "pageSize":5000,
                "connTimeOut":600000,
                "readTimeOut":600000,
                "retryCount":30,
                "retrySleepTime":"10000",
                "search":{
                            "range":{
                                "gmt_modified":{
                                    "gte":0
                                }
                            }
                        },  //查询query参数，与Elasticsearch的query内容相同，使用_search api，重命名为search。
                "type":"doc",
                "username":"aliyun_di"  //用户名。
            },
            "stepType":"elasticsearch"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter":{ },
            "stepType":"stream"
        }
    ],
    "type":"job",
    "version":"2.0" //版本号。
}
```

配置数据集成资源组

1. 单击数据同步任务编辑页面右侧的**数据集成资源组配置**。
2. 根据提示选择对应的**独享数据集成资源组**。

> **说明**
>
> - （推荐）**数据集成资源组配置**页面默认支持选择独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。
> - 如果您需要选择公共资源组，请在页面右下方单击**更多选项**，在弹出的**警告**对话框单击**确认**，在**数据集成资源组配置**子页面进行选择。关于自定义数据集成资源组和公共资源组，详情请参见公共资源组。

6.2.23. AnalyticDB for MySQL 2.0 Reader

本文为您介绍AnalyticDB for MySQL 2.0 Reader支持的数据类型、字段映射和数据源等参数及配置示例。

AnalyticDB for MySQL 2.0 Reader插件实现了从AnalyticDB for MySQL 2.0读取数据。在底层实现上，AnalyticDB for MySQL 2.0 Reader通过JDBC连接远程AnalyticDB for MySQL 2.0数据库，并根据AnalyticDB for MySQL 2.0的推荐分页大小，执行相应的SQL语句，从AnalyticDB for MySQL 2.0库中分批读取数据。

数据类型转换

AnalyticDB for MySQL 2.0类型	DataX类型	MaxCompute类型
BIGINT	LONG	BIGINT
TINYINT	LONG	INT
TIMESTAMP	DATE	DATETIME
VARCHAR	STRING	STRING
SMALLINT	LONG	INT
INT	LONG	INT
FLOAT	STRING	DOUBLE
DOUBLE	STRING	DOUBLE

AnalyticDB for MySQL 2.0类型	DataX类型	MaxCompute类型
DATE	DATE	DATETIME
TIME	DATE	DATETIME

> **说明** 不支持multivalue，会直接异常退出。

使用限制

当前版本，在大批量数据导出并且配置较低的机器上，会出现超时的情况。

- 当前mode=Select时，上限为30万行。
- 当前mode=ODPS时，上限为1亿行。
- 50列以上为AnalyticDB for MySQL 2.0本身的限制，需要联系AnalyticDB for MySQL 2.0的管理员进行手动调整。
- Java版本需要1.8及以上，编译转码 `native2ascii LocalStrings.properties > LocalStrings_zh_CN.properties` 。

参数说明

参数	描述	是否必选	默认值
table	需要导出的表的名称。	是	无
column	列名，如果没有，则为全部。	否	*
limit	限制导出的记录数。	否	无
where	where条件，方便添加筛选条件，此处的String会被直接作为SQL条件添加到查询语句中，例如 `where id < 100` 。	否	无
mode	目前支持Select和ODPS2种导入类型。 • Select：使用limit分页。 • ODPS：使用ODPS DUMP来导出数据，需要有ODPS的访问权限。	否	*Select*
odps.accessKey	当mode=ODPS时必填，AnalyticDB for MySQL 2.0访问ODPS使用的云账号AccessKey，需要有Describe、Create、Select、Alter、Update和Drop权限。	否	无
odps.accessId	当mode=ODPS时必填，AnalyticDB for MySQL 2.0访问ODPS使用的云账号AccessID，需要有Describe、Create、Select、Alter、Update和Drop权限。	否	无
odps.odpsServer	当mode=ODPS时必填，ODPS API地址。	否	无
odps.tunnelServer	当mode=ODPS时必填，ODPS Tunnel地址。	否	无
odps.project	当mode=ODPS时必填，ODPS Project名称。	否	无

参数	描述	是否必选	默认值
odps.accountType	当mode=ODPS时生效，ODPS访问账号类型。	否	*aliyun*

配置文件示例

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "ads",
            "parameter": {
                "datasource": "ads_demo",
                "table": "th_test",
                "column": [
                    "id",
                    "testtinyint",
                    "testbigint",
                    "testdate",
                    "testtime",
                    "testtimestamp",
                    "testvarchar",
                    "testdouble",
                    "testfloat"
                ],
                "odps": {
                    "accessId": "<yourAccessKeyId>",
                    "accessKey": "<yourAccessKeySecret>",
                    "account": "*********@aliyun.com",
                    "odpsServer": " http://service.cn.maxcompute.aliyun-inc.com/api",
                    "tunnelServer": "http://dt.cn-shanghai.maxcompute.aliyun-inc.com",
                    "accountType": "aliyun",
                    "project": "odps_test"
                },
                "mode": "ODPS"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "concurrent": 2,
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
                  "mbps":"12"//限流
        }
    }
}
```

6.2.24. Kafka Reader

Kafka Reader通过Kafka服务的Java SDK从Kafka读取数据。

背景信息

Apache Kafka是一个快速、可扩展、高吞吐和可容错的分布式发布订阅消息系统。Kafka具有高吞吐量、内置分区、支持数据副本和容错的特性，适合在大规模消息处理的场景中使用。

> 注意
> - 支持阿里云Kafka，以及>=0.10.2且<=2.2.x的自建Kafka版本。
> - 对于<0.10.2版本Kafka，由于Kafka不支持检索分区数据offset，且Kafka数据结构可能不支持时间戳，所以无法进行数据同步。
> - Kafka Reader仅支持使用新增和使用独享数据集成资源组，不支持使用公共资源组和自定义资源组。

实现原理

Kafka Reader通过Kafka Java SDK读取Kafka中的数据，使用的日志服务Java SDK版本如下所示。

```
<dependency>
   <groupId>org.apache.kafka</groupId>
   <artifactId>kafka-clients</artifactId>
   <version>2.0.0</version>
</dependency>
```

主要涉及的Kafka SDK调用方法如下，详情请参见Kafka官方文档：

- 使用KafkaConsumer作为消息消费的客户端。

  ```
  org.apache.kafka.clients.consumer.KafkaConsumer<K,V>
  ```

- 根据unix时间戳查询Kafka点位offSet。

  ```
  Map<TopicPartition,OffsetAndTimestamp> offsetsForTimes(Map<TopicPartition,Long> timestampsToSearch)
  ```

- 定位到开始点位offSet。

  ```
  public void seekToBeginning(Collection<TopicPartition> partitions)
  ```

- 定位到结束点位offSet。

  ```
  public void seekToEnd(Collection<TopicPartition> partitions)
  ```

- 定位到指定点位offSet。

  ```
  public void seek(TopicPartition partition,long offset)
  ```

- 客户端从服务端拉取poll数据。

  ```
  public ConsumerRecords<K,V> poll(final Duration timeout)
  ```

> 说明 Kafka Reader消费数据使用了自动点位提交机制。

参数说明

参数	描述	是否必选
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是
server	Kafka的broker server地址，格式为ip:port。 您可以只配置一个server，但请务必保证Kafka集群中所有broker的IP地址都可以连通DataWorks。	是

参数	描述	是否必选
topic	Kafka的Topic，是Kafka处理资源的消息源（feeds of messages）的聚合。	是
column	需要读取的Kafka数据，支持常量列、数据列和属性列： • 常量列：使用单引号包裹的列为常量列，例如 `["'abc'", "'123'"]` 。 • 数据列 ○ 如果您的数据是一个JSON，支持获取JSON的属性，例如 `["event_id"]` 。 ○ 如果您的数据是一个JSON，支持获取JSON的嵌套子属性，例如 `["tag.desc"]` 。 • 属性列 ○ __key__表示消息的key。 ○ __value__表示消息的完整内容 。 ○ __partition__表示当前消息所在分区。 ○ __headers__表示当前消息headers信息。 ○ __offset__表示当前消息的偏移量。 ○ __timestamp__表示当前消息的时间戳。 完整示例如下。 `"column": [ "__key__", "__value__", "__partition__", "__offset__", "__timestamp__", "'123'", "event_id", "tag.desc" ]`	是
keyType	Kafka的Key的类型，包括BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。	是
valueType	Kafka的Value的类型，包括BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。	是
beginDateTime	数据消费的开始时间位点，为时间范围（左闭右开）的左边界。yyyymmddhhmmss格式的时间字符串，可以配合**时间属性**使用。详情请参见调度参数概述。 **说明** Kafka 0.10.2及以上的版本支持该功能。	需要和beginOffset二选一。 **说明** beginDateTime和endDateTime配合使用。
endDateTime	数据消费的结束时间位点，为时间范围（左闭右开）的右边界。yyyymmddhhmmss格式的时间字符串，可以配合**时间属性**使用。详情请参见调度参数概述。 **说明** Kafka 0.10.2及以上的版本支持该功能。	需要和endOffset二选一。 **说明** endDateTime和beginDateTime配合使用。

参数	描述	是否必选
beginOffset	数据消费的开始时间位点，您可以配置以下形式： • 数字形式（例如15553274），表示开始消费的点位。 • seekToBeginning：表示从开始点位消费数据。 • seekToLast：表示从上次的偏移位置读取数据。 • seekToEnd：表示从最后点位消费数据，会读取到空数据。	需要和beginDateTime二选一。
endOffset	数据消费的结束位点，用于控制结束数据消费任务退出的时间。	需要和endDateTime二选一。
skipExceedRecord	Kafka使用 `public ConsumerRecords<K, V> poll(final Duration timeout)` 消费数据，一次poll调用获取的数据可能在endOffset或者endDateTime之外。skipExceedRecord用于控制是否写出多余的数据至目的端。由于消费数据使用了自动点位提交，建议您： • Kafka 0.10.2之前版本：建议配置skipExceedRecord为false。 • Kafka 0.10.2及以上版本：建议配置skipExceedRecord为true。	否，默认值为*false*。
partition	Kafka的一个Topic有多个分区（partition），正常情况下数据同步任务是读取Topic（多个分区）一个点位区间的数据。您也可以指定partition，仅读取一个分区点位区间的数据。	否，无默认值。
kafkaConfig	创建Kafka数据消费客户端KafkaConsumer可以指定扩展参数，例如bootstrap.servers、auto.commit.interval.ms、session.timeout.ms等，您可以基于kafkaConfig控制KafkaConsumer消费数据的行为。	否
encoding	当keyType或valueType配置为STRING时，将使用该配置项指定的编码解析字符串。	否，默认为UTF-8。
waitTIme	消费者对象从Kafka拉取一次数据的最大等待时间，单位为秒。	否，默认为60。
stopWhenPollEmpty	该配置项可选值为true/false。当配置为true时，如果消费者从Kafka拉取数据返回为空（一般是已经读完主题中的全部数据，也可能是网络或者Kafka集群可用性问题），则立即停止任务，否则持续重试直到再次读到数据。	否，默认为true。

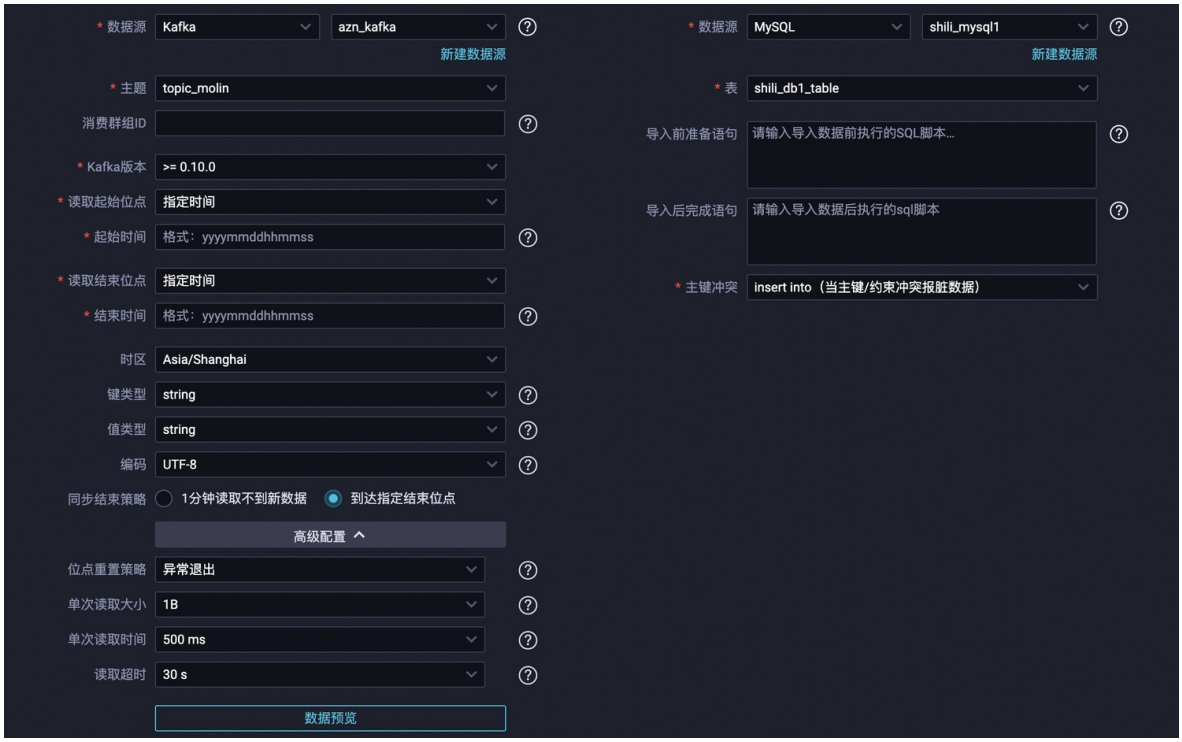
kafkaConfig参数说明如下。

参数	描述
fetch.min.bytes	指定消费者从broker获取消息的最小字节数，即有足够的数据时，才将其返回给消费者。
fetch.max.wait.ms	等待broker返回数据的最大时间，默认500毫秒。fetch.min.bytes和fetch.max.wait.ms先满足哪个条件，便按照该方式返回数据。
max.partition.fetch.bytes	指定broker从每个partition中返回给消费者的最大字节数，默认为1 MB。
session.timeout.ms	指定消费者不再接收服务之前，可以与服务器断开连接的时间，默认是30秒。
auto.offset.reset	消费者在读取没有偏移量或者偏移量无效的情况下（因为消费者长时间失效，包含偏移量的记录已经过时并被删除）的处理方式。默认为latest（消费者从最新的记录开始读取数据），您可以更改为earliest（消费者从起始位置读取partition的记录）。
max.poll.records	单次调用poll方法能够返回的消息数量。
key.deserializer	消息Key的反序列化方法，例如org.apache.kafka.common.serialization.StringDeserializer。
value.deserializer	数据Value的反序列化方法，例如org.apache.kafka.common.serialization.StringDeserializer。
ssl.truststore.location	SSL根证书的路径。
ssl.truststore.password	根证书Store的密码。如果是Aliyun Kafka，则配置为KafkaOnsClient。

参数	描述
security.protocol	接入协议，目前支持使用SASL_SSL协议接入。
sasl.mechanism	SASL鉴权方式，如果是Aliyun Kafka，使用PLAIN。
java.security.auth.login.config	SASL鉴权文件路径。

向导开发介绍

1. 选择数据源。配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
主题	即上述参数说明中的topic，可以在下拉列表中选择要读取的Kafka主题名称。
消费群组ID	即上述参数说明中的kafkaConfig的JSON结构中的group.id字段，是初始化Kafka Consumer时的group.id配置。 为了确保同步时消费位点的正确性，请避免该参数与其他消费进程重复。如果不指定该参数，则每次执行同步自动生成datax_开头的随机字符串作为group.id。
Kafka版本	Kafka版本可以选择>=0.10.2且<=2.2.x。该参数主要影响读取起始位点和读取结束位点的可选项。

参数	描述
读取起始位点	即上述参数说明中的beginOffset。 ○ 指定时间：数据写入Kafka的时候自动生成一个unixtime时间戳作为该数据的时间记录。同步任务通过获取用户配置的yyyymmddhhmmss数值，将该值转成unixtimestamp后从kafka中读取相应数据。例如，"beginDateTime": "20210125000000"。 ○ 分区起始位点：从kafka topic每个分区没有删除的位点最小的数据开始抽取数据。 ○ 群组当前位点：从任务配置上面指定的消费群组ID保存的位点开始读取数据，一般是使用这个群组ID读数据的进程上次停止的位点（最好确保使用这个群组ID的进程只有配置的这个数据集成任务，避免共用群组ID造成数据丢失），如果使用群组当前位点，一定要配置消费群组ID，否则数据集成任务会随机生成一个群组ID，而新的群组ID因为没有保存过位点，所以会从分区起始位点开始读取。 **说明** 如果选择除了指定时间之外的选项时对应beginOffset。如果选择指定时间，生成任务时将不会填充beginOffset参数，而是采用起始时间表单填充beginDateTime参数来决定读取的起始位点。
起始时间	即上述参数说明中的beginDateTime，当读取起始位点选择指定时间时，支持以yyyyMMddHHmmss格式的时间字符串指定具体时间，为时间范围（左闭右开）的左边界，例如20210513000000，可以配合时间属性使用。详情请参见调度参数。
读取结束位点	即上述参数说明中的endOffset。 **说明** 如果选择除了指定时间之外的选项时对应endOffset。数据读取结束位点如果选择指定时间，生成任务配置时将不会填充endOffset参数，而是采用结束时间表单填充endDateTime参数来决定读取的结束位点。
结束时间	即上述参数说明中的endDateTime，当读取结束位点选择指定时间时，支持以yyyyMMddHHmmss格式的时间字符串指定具体时间，为时间范围（左闭右开）的右边界，例如20210514000000，可以配合时间属性使用。详情请参见调度参数。
时区	当读取起始位点或读取结束位点配置指定时间时，决定时间字符串对应时区。
键类型	即上述参数说明中的keyType，Kafka的Key的类型，决定了初始化KafkaConsumer时的key.deserializer配置，可选值包括STRING、BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。
值类型	即上述参数说明中的valueType，Kafka的Value的类型，决定了初始化KafkaConsumer时的value.deserializer配置，可选值包括STRING、BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。
编码	即上述参数说明中的encoding，当键类型或者值类型配置为STRING时，决定读取时的编码。
同步结束策略	○ 当配置为1分钟读取不到新数据时，如果消费者1分钟从Kafka拉取数据返回为空（一般是已经读完主题中的全部数据，也可能是网络或者Kafka集群可用性原因），则立即停止任务，否则持续重试直到再次读到数据。 ○ 当配置为**到达指定结束位点**时，如果数据集成任务读取到的Kafka记录业务时间或者位点满足上面读取结束位点配置时，则任务结束，否则无限重试读取Kafka记录。
位点重置策略	即上述参数说明中的kafkaConfig的JSON结构中的auto.offset.reset字段，通过控制kafka消费者的auto.offset.reset参数控制找不到位点时的重置策略，建议配置为“异常退出”，避免读取丢失数据。
单次读取大小	即上述参数说明中的kafkaConfig的JSON结构中的fetch.min.bytes字段，通过控制kafka消费者的fetch.min.bytes参数控制单次读取批大小。
单次读取时间	即上述参数说明中的kafkaConfig的JSON结构中的fetch.max.wait.ms字段，通过控制kafka消费者的fetch.max.wait.ms参数控制单次读取时间。

参数	描述
读取超时	即上述参数说明中的kafkaConfig的JSON结构中的session.timeout.ms字段，通过控制kafka消费者的session.timeout.ms参数控制读取超时时间。

2. 字段映射，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标悬停至需要删除的字段上，即可单击**删除**图标进行删除。

   当源头表字段配置为如下两个下划线（__）开头的6个字符串时，对应Kafka记录的特定内容。具体如下所示：

源头表字段	描述
__key__	Kafka记录的key值。
__value__	Kafka记录的value值。
__partition__	Kafka记录所在分区编号，是一个大于等于0的整数。
__headers__	Kafka记录的headers序列化得到的JSON字符串。
__offset__	Kafka记录所在分区的偏移量，是一个大于等于0的整数。
__timestamp__	Kafka记录的毫秒时间戳。

   源头表字段也可配置为上述6个字符串之外的字符串，此时将Kafka记录作为JSON字符串进行解析，将源头表字段配置的字符串作为JSON路径读取对应内容作为字段值写入对应的目标表字段，例如以下示例为Kafka记录的value值，当源头表字段配置为data.name时，将会读取"bob"作为这个字段的值并写入对应目标表。

```
{
  "data": {
    "name": "bob",
    "age": 35
  }
}
```

   源头表和目标表的映射关系如下所示：

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。

参数	描述
添加一行	○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发示例

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

从Kafka读取数据的JSON配置，如下所示。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "kafka",
            "parameter": {
                "server": "host:9093",
                "column": [
                    "__key__",
                    "__value__",
                    "__partition__",
                    "__offset__",
                    "__timestamp__",
                    "'123'",
                    "event_id",
                    "tag.desc"
                ],
                "kafkaConfig": {
                    "group.id": "demo_test"
                },
                "topic": "topicName",
                "keyType": "ByteArray",
                "valueType": "ByteArray",
                "beginDateTime": "20190416000000",
                "endDateTime": "20190416000006",
                "skipExceedRecord": "false"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {
                "print": false,
                "fieldDelimiter": ","
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent": 1,//并发数
            "mbps":"12"//限流
        }
    }
}
```

使用SASL鉴权

如果需要使用SASL鉴权或SSL鉴权，请在定义kafka数据源时进行相关配置，详情请参考：配置Kafka数据源。

6.2.25. MaxCompute Reader

本文为您介绍MaxCompute Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

MaxCompute Reader插件实现了从MaxCompute读取数据的功能，有关MaxCompute的详细介绍请参见MaxCompute简介。

根据您配置的源头项目、表、分区和表字段等信息，可以通过Tunnel从MaxCompute系统中读取数据。常用的Tunnel命令请参见Tunnel命令操作。

MaxCompute Reader支持读取分区表、非分区表，不支持读取虚拟视图。DataWorks不支持对MaxCompute分区表进行字段映射，您需要对分区字段进行单独配置。例如，读取t0表，其分区为pt=1，ds=hangzhou，则您需要在配置中配置该值。表字段既可以依序指定全部列、部分列，也可以调整列顺序、指定常量字段和指定分区列（分区列不是表字段）。

> 说明
> - MaxCompute Reader不支持数据过滤功能。如果您在数据同步过程中，需要过滤符合条件的数据，请创建新表并写入过滤数据后，同步新表中的数据。
> - MaxCompute Reader不支持同步外部表。

支持的数据类型

MaxCompute Reader针对MaxCompute的类型转换列表，如下所示。

类型分类	数据集成配置类型	数据库数据类型
整数类	LONG	BIGINT、INT、TINYINT和SMALLINT
布尔类	BOOLEAN	BOOLEAN
日期时间类	DATE	DATETIME、TIMESTAMP和DATE
浮点类	DOUBLE	FLOAT、DOUBLE和DECIMAL
二进制类	BYTES	BINARY
复杂类	STRING	ARRAY、MAP和STRUCT

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称。脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	读取数据表的表名称（大小写不敏感）。	是	无

参数	描述	是否必选	默认值
partition	读取的数据所在的分区信息。 • ODPS的分区配置支持linux shell通配符，`*` 表示0个或多个字符，`?` 表示任意一个字符。 • 默认情况下，读取的分区必须存在，如果分区不存在则运行的任务会报错。如果您希望当分区不存在时任务仍然执行成功，则可以切换至脚本模式执行任务，并在ODPS的Parameter中添加 `"successOnNoPartition": true` 配置。 例如，分区表*test*包含*pt=1,ds=hangzhou*、*pt=1,ds=shanghai*、*pt=2,ds=hangzhou*、*pt=2,ds=beijing*四个分区，则读取不同分区数据的配置如下： • 如果您需要读取*pt=1,ds=hangzhou*分区的数据，则分区信息的配置为 `"partition":"pt=1,ds=hangzhou"` 。 • 如果您需要读取*pt=1*中所有分区的数据，则分区信息的配置为 `"partition":"pt=1,ds=*"` 。 • 如果您需要读取整个*test*表所有分区的数据，则分区信息的配置为 `"partition":"pt=*,ds=*"` 。 此外，您还可以根据实际需求设置分区数据的获取条件： • 如果您需要指定最大分区，则可以添加 `/*query*/ ds=(select MAX(ds) from DataXODPSReaderPPR)` 配置信息。 • 如果需要按条件过滤，则可以添加相关条件 `/*query*/ pt+表达式` 配置。例如 `/*query*/ pt>=20170101 and pt<20170110` 表示获取pt分区中，20170101日期之后（包含20170101日期），至20170110日期之前（不包含20170110日期）的所有数据。 **说明** `/*query*/` 表示将其后填写的内容识别为一个where条件。	如果表为分区表，则必填。如果表为非分区表，则不能填写。	无
column	读取MaxCompute源头表的列信息。例如表test的字段为id、name和age： • 如果您需要依次读取id、name和age，则应该配置为 `"column":["id","name","age"]` 或者配置为 `"column":["*"]` 。 **说明** 不推荐您配置抽取字段为（*），因为它表示依次读取表的每个字段。如果您的表字段顺序调整、类型变更或者个数增减，您的任务会存在源头表列和目的表列不能对齐的风险，则直接导致您的任务运行结果不正确甚至运行失败。 • 如果您想依次读取name和id，则应该配置为 `"coulumn":["name","id"]` 。 • 如果您想在源头抽取的字段中添加常量字段（以适配目标表的字段顺序）。例如，您想抽取的每一行数据值为age列对应的值，name列对应的值，常量日期值1988-08-08 08:08:08，id列对应的值，则您应该配置为 `"column":["age","name","'1988-08-08 08:08:08'","id"]` ，即常量列首尾用符号 `'` 包住即可。 内部实现上识别常量是通过检查您配置的每一个字段，如果发现有字段首尾都有 `'` ，则认为其是常量字段，其实际值为去除 `'` 之后的值。 **说明** ◦ MaxCompute Reader抽取数据表不是通过MaxCompute的Select SQL语句，所以不能在字段上指定函数。 ◦ column必须显示指定同步的列集合，不允许为空。	是	无

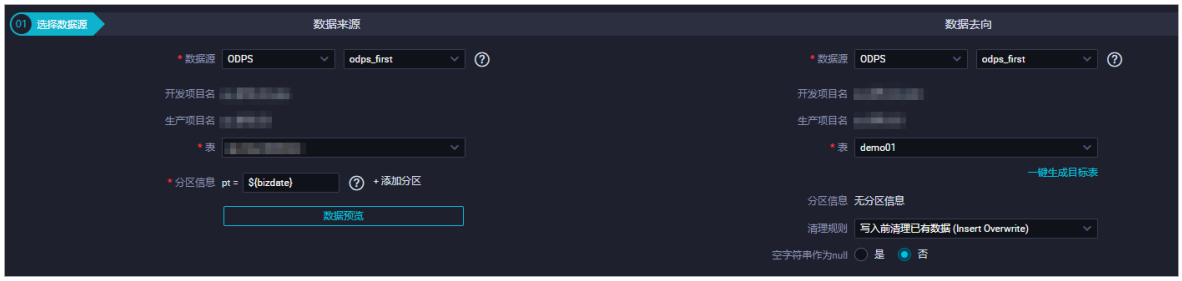
向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
开发项目名	默认不可以修改。 **说明** 仅标准模式的工作空间会显示该配置。
生产项目名	默认不可以修改。
表	即上述参数说明中的table，选择需要同步的表。
分区信息	如果您每日增量数据限定在对应日期的分区中，可以使用分区进行每日增量。例如，配置分区pt的值为${bizdate}。 **说明** DataWorks不支持对MaxCompute分区表进行字段映射，您需要单独配置分区字段。

说明 如果是指定所有的列，可以在column配置，例如 `"column": [""]` 。partition支持配置多个分区和通配符的配置方法：

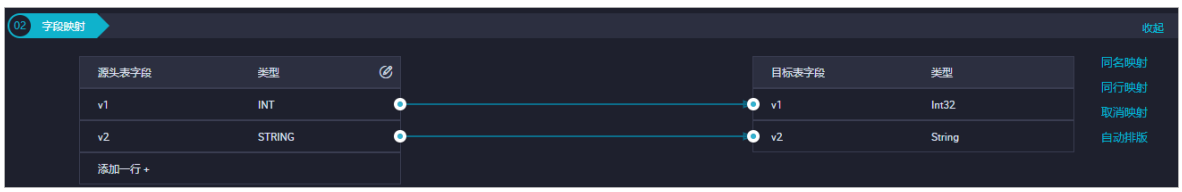
- `"partition":"pt=20140501/ds=*"` 代表ds中的所有的分区。
- `"partition":"pt=top?"` 中的问号（?）代表前面的字符是否存在，指pt=top和pt=to两个分区。

您可以输入需要同步的分区列，例如MaxCompute的分区为pt=${bdp.system.bizdate}，您可以直接添加分区名称pt至源表字段中（可能会有未识别的标志，直接忽视进行下一步）：

- 如果需要同步所有的分区，配置分区值为pt=*。
- 如果需要同步某个分区，可以直接选择您要同步的时间值。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。

参数	描述
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其它空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从MaxCompute抽取数据到本地的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "type":"job",
    "version":"2.0",
    "steps":[
        {
            "stepType":"odps",//插件名。
            "parameter":{
                "partition":[],//读取数据所在的分区。
                "isCompress":false,//是否压缩。
                "datasource":"",//数据源。
                "column":[//源头表的列信息。
                    "id"
                ],
                "emptyAsNull":true,
                "table":""//表名。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

如果您需要指定MaxCompute的Tunnel Endpoint，可以通过脚本模式手动配置数据源。将上述示例中的 `"datasource":""` 替换为数据源的具体参数，示例如下。

```
"accessId":"********************",
"accessKey":"********************",
"endpoint":"http://service.eu-central-1.maxcompute.aliyun-inc.com/api",
"odpsServer":"http://service.eu-central-1.maxcompute.aliyun-inc.com/api",
"tunnelServer":"http://dt.eu-central-1.maxcompute.aliyun.com",
"project":"*****",
```

6.2.26. Prometheus Reader

Prometheus是时间序列数据库，由SoundCloud开发并维护，是Google BorgMon监控系统的开源版本。Prometheus Reader插件实现了从Prometheus读取数据。

> 注意
> - Prometheus Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。
> - Prometheus Reader仅支持脚本模式配置方式。

实现原理

在底层实现上，Prometheus Reader通过HTTP请求连接到Prometheus实例，用 `/api/v1/query_range` 接口获取原始数据点。整个同步的过程通过Metric和时间段进行切分，组合为一个迁移Task。

使用限制

- 指定起止时间会被自动转为整点时刻，例如2019-4-18的 `[3:35, 4:55)` ，会被转为 `[3:00, 4:00)` 。
- 目前仅支持兼容Prometheus 2.9.x版本。
- 时间上切分的粒度，默认只有10s。

  `/api/v1/query_range` 接口对查询的数据点数量有所限制。如果查询的时间范围过大，会报 `exceeded maximum resolution of 11,000 points per timeseries` 的异常。因此插件中默认选择10s作为查询的切分粒度。即使原始数据点的存储粒度为毫秒级，也只会查询出10,000个数据点，可满足 `/api/v1/query_range` 接口的限制。

支持的数据类型

类型分类	数据集成数据类型	TSDB数据类型
字符串	STRING	TSDB数据点序列化字符串，包括TIMESTAMP、METRIC、TAGS和VALUE。

参数说明

参数	描述	是否必选	默认值
endpoint	Prometheus的HTTP连接地址。	是，格式为 `http://IP:Port` 。	无
column	数据迁移任务需要迁移的Metric列表。	是	无
beginDateTime	配合endDateTime使用，用于指定哪个时间段内的数据点需要被迁移。	是，格式为 `yyyyMMddHHmmss` 。	无 **说明** 指定起止时间会自动忽略分钟和秒，转为整点时刻。例如2019-4-18的 `[3:35, 4:55)` 会被转为 `[3:00, 4:00)` 。
endDateTime	配合beginDateTime使用，用于指定哪个时间段内的数据点需要被迁移。	是，格式为 `yyyyMMddHHmmss` 。	无 **说明** 指定起止时间会自动忽略分钟和秒，转为整点时刻。例如2019-4-18的 `[3:35, 4:55)` 会被转为 `[3:00, 4:00)` 。

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个从Prometheus数据库同步的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,//并发数
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "endpoint": "http://localhost:9090",
                "column": [
                    "up"
                ],
                "beginDateTime": "20190520150000",
                "endDateTime": "20190520160000"
            },
            "stepType": "prometheus"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {},
            "stepType": ""
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

性能测试报告

通道数	数据集成速度（Rec/s）	数据集成流量（MB/s）
1	45,000	5.36
2	55,384	6.60
3	60,000	7.15

6.2.27. PostgreSQL Reader

本文为您介绍PostgreSQL Reader支持的数据类型、读取方式、字段映射和数据源等参数及配置示例。

背景信息

PostgreSQL Reader插件从PostgreSQL读取数据。在底层实现上，PostgreSQL Reader通过JDBC连接远程PostgreSQL数据库，并执行相应的SQL语句，从PostgreSQL库中选取数据。RDS提供PostgreSQL存储引擎。

PostgreSQL Reader通过JDBC连接器连接至远程的PostgreSQL数据库，根据您配置的信息生成查询SQL语句，发送至远程PostgreSQL数据库，执行该SQL并返回结果。再使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理：

- 对于您配置的table、column和where等信息，PostgreSQL Reader将其拼接为SQL语句发送至PostgreSQL数据库。
- 对于您配置的querySql信息，PostgreSQL直接将其发送至PostgreSQL数据库。

注意事项

当PostgreSQL中表名称、字段名称是以数字开头，或者名称中包含大小写英文字母、中划线（-）时需要使用双引号（""）进行转义，不进行转义会导致PostgreSQL Reader插件读取PostgreSQL数据失败。但是在PostgreSQL Reader插件中，双引号（""）为JSON关键字，因此，您需要使用反斜线（\）再次对双引号（""）进行转义。例如，表名称为 `123Test` ，则转义后表名称为 `\"123Test\"` 。

> **说明**
> - 双引号（""）中，前引号（"）和后引号（"）均需使用反斜线（\）进行转义。
> - 向导模式不支持转义，您需要转换为脚本模式进行转义。

使用脚本模式进行转义的代码示例如下。

```
"parameter": {
    "datasource": "abc",
    "column": [
        "id",
        "\"123Test\"", //添加转义符
],
"where": "",
"splitPk": "id",
"table": "public.wpw_test"
},
```

类型转换列表

PostgreSQL Reader支持大部分PostgreSQL类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型。

PostgreSQL Reader针对PostgreSQL的类型转换列表，如下所示。

类型分类	PostgreSQL数据类型
整数类	BIGINT、BIGSERIAL、INTEGER、SMALLINT和SERIAL
浮点类	DOUBLE、PRECISION、MONEY、NUMERIC和REAL
字符串类	VARCHAR、CHAR、TEXT、BIT和INET
日期时间类	DATE、TIME和TIMESTAMP
布尔型	BOOL
二进制类	BYTEA

> **说明**
> - 除上述罗列字段类型外，其它类型均不支持。
> - MONEY、INET和BIT需要您使用 `a_inet::varchar` 类似的语法进行转换。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表schema信息顺序进行导出。 • 支持常量配置，您需要按照MySQL SQL语法格式，例如 `["id", "table","1", "'mingya.wmy'", "'null'", "to_char(a+1)", "2.3" , "true"]` 。 ○ id为普通列名。 ○ table为包含保留字的列名。 ○ 1为整形数字常量。 ○ 'mingya.wmy'为字符串常量（注意需要加上一对单引号）。 ○ 'null'为字符串。 ○ to_char(a+1)为计算字符串长度函数。 ○ 2.3为浮点数。 ○ true为布尔值。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	PostgreSQL Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步会启动并发任务，以提高数据同步的效能： • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其它类型 。如果您指定其它非支持类型，忽略splitPk功能，使用单通道进行同步。 • 如果splitPk不填写，包括不提供splitPk或者splitPk值为空，数据同步视作使用单通道同步该表数据 。	否	无
where	筛选条件，PostgreSQL Reader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。例如在测试时，您可以使用where条件指定实际业务场景，通常会选择当天的数据进行同步，指定where条件为 `id>2 and sex=1` ： • where条件可以有效地进行业务增量同步。 • where条件不配置或者为空，视作全表同步数据。	否	无
querySql（高级模式，向导模式不提供）	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当配置该项后，数据同步系统会忽略tables、columns和splitPk配置项，直接使用该配置的内容筛选数据。例如需要进行多表JOIN后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，PostgreSQL Reader直接忽略table、column和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够较大的提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*512*

向导开发介绍

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
表	即上述参数说明中的table，选择需要同步的表。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的**column**。

左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。

参数	描述
添加一行	○ 可以输入常量，输入的值需要使用英文单引号，例如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从PostgreSQL数据库同步抽取数据作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"postgresql",//插件名。
            "parameter":{
                "datasource":"",//数据源。
                "column":[//字段。
                    "col1",
                    "col2"
                ],
                "where":"",//筛选条件。
                "splitPk":"",//用splitPk代表的字段进行数据分片，数据同步会启动并发任务进行数据同步。
                "table":""//表名。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true, //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

补充说明

- 主备同步数据恢复问题

  主备同步问题指PostgreSQL使用主从灾备，备库从主库不间断通过binlog恢复数据。由于主备数据同步存在一定的时间差，特别在于某些特定情况，例如网络延迟等问题，导致备库同步恢复的数据与主库有较大差别，从备库同步的数据不是一份当前时间的完整镜像。

- 一致性约束

  PostgreSQL在数据存储划分中属于RDBMS系统，对外可以提供强一致性数据查询接口。例如一次同步任务启动运行过程中，当该库存在其他数据写入方写入数据时，由于数据库本身的快照特性，PostgreSQL Reader完全不会获取到写入的更新数据。

  上述是在PostgreSQL Reader单线程模型下数据同步一致性的特性，PostgreSQL Reader可以根据您配置的信息使用并发数据抽取，因此不能严格保证数据一致性。

  当PostgreSQL Reader根据splitPk进行数据切分后，会先后启动多个并发任务完成数据同步。多个并发任务相互之间不属于同一个读事务，同时多个并发任务存在时间间隔，因此这份数据并不是完整的、一致的数据快照信息。

针对多线程的一致性快照需求，目前在技术上无法实现，只能从工程角度解决。工程化的方式存在取舍，在此提供以下解决思路，您可以根据自身情况进行选择。

- 使用单线程同步，即不再进行数据切片。缺点是速度比较慢，但是能够很好保证一致性。
- 关闭其它数据写入方，保证当前数据为静态数据，例如锁表、关闭备库同步等。缺点是可能影响在线业务。

- 数据库编码问题

  PostgreSQL在服务器端仅支持EUC_CN和UTF-8两种简体中文编码，PostgreSQL Reader底层使用JDBC进行数据抽取，JDBC天然适配各类编码，并在底层进行了编码转换。因此PostgreSQL Reader不需您指定编码，可以自动获取编码并转码。

  对于PostgreSQL底层写入编码和其设定的编码不一致的混乱情况，PostgreSQL Reader对此无法识别，也无法提供解决方案，导出结果有可能为乱码。

- 增量数据同步的方式

  PostgreSQL Reader使用JDBC SELECT语句完成数据抽取工作，因此可以使用 `SELECT…WHERE…` 进行增量数据抽取，方式如下：

  - 数据库在线应用写入数据库时，填充modify字段为更改时间戳，包括新增、更新、删除（逻辑删除）。对于该类应用，PostgreSQL Reader只需要where条件后跟上一同步阶段时间戳即可。
  - 对于新增流水型数据，PostgreSQL Reader在where条件后跟上一阶段最大自增ID即可。

  对于业务上无字段区分新增、修改数据的情况，PostgreSQL Reader无法进行增量数据同步，只能同步全量数据。

- SQL安全性

  PostgreSQL Reader提供querySql语句交给您自己实现SELECT抽取语句，PostgreSQL Reader本身对querySql不进行任何安全性校验。

6.2.28. OTSStream Reader

本文为您介绍OTSStream Reader支持的数据类型、读取方式、字段映射和数据源等参数及配置示例。

背景信息

OTSStream Reader插件主要用于导出Table Store的增量数据。您可以将增量数据看作操作日志，除数据本身外还附有操作信息。

与全量导出插件不同，增量导出插件只有多版本模式，且不支持指定列。使用插件前，您必须确保表上已经开启Stream功能。您可以在建表时指定开启，也可以使用SDK的UpdateTable接口开启。

开启Stream的方法，如下所示。

```
SyncClient client = new SyncClient("", "", "", "");
建表的时候开启：
CreateTableRequest createTableRequest = new CreateTableRequest(tableMeta);
createTableRequest.setStreamSpecification(new StreamSpecification(true, 24)); // 24代表增量数据保留24小时。
client.createTable(createTableRequest);
如果建表时未开启，您可以通过UpdateTable开启：
UpdateTableRequest updateTableRequest = new UpdateTableRequest("tableName");
updateTableRequest.setStreamSpecification(new StreamSpecification(true, 24));
client.updateTable(updateTableRequest);
```

您使用SDK的UpdateTable功能，指定开启Stream并设置过期时间，即开启了Table Store增量数据导出功能。开启后，Table Store服务端就会将您的操作日志额外保存起来，每个分区有一个有序的操作日志队列，每条操作日志会在一定时间后被垃圾回收，该时间即为您指定的过期时间。

Table Store的SDK提供了几个Stream相关的API用于读取这部分的操作日志，增量插件也是通过Table Store SDK的接口获取到增量数据，默认情况下会将增量数据转化为多个6元组的形式（pk、colName、version、colValue、opType和sequenceInfo）导入至MaxCompute中。

列模式

在Table Store多版本模型下，表中的数据组织为行>列>版本三级的模式， 一行可以有任意列，列名并不是固定的，每一列可以含有多个版本，每个版本都有一个特定的时间戳（版本号）。

您可以通过Table Store的API进行一系列读写操作，Table Store通过记录您最近对表的一系列写操作（或数据更改操作）来实现记录增量数据的目的，所以您也可以把增量数据看作一批操作记录。

Table Store支持PutRow、UpdateRow和DeleteRow操作：

- PutRow：写入一行，如果该行已存在即覆盖该行。
- UpdateRow：更新一行，不更改原行的其它数据。更新包括新增或覆盖（如果对应列的对应版本已存在）一些列值、删除某一列的全部版本、删除某一列的某个版本。
- DeleteRow：删除一行。

Table Store会根据每种操作生成对应的增量数据记录，Reader插件会读出这些记录，并导出为数据集成的数据格式。

同时，由于Table Store具有动态列、多版本的特性，所以Reader插件导出的一行不对应Table Store中的一行，而是对应Table Store中的一列的一个版本。即Table Store中的一行可能会导出很多行，每行包含主键值、该列的列名、该列下该版本的时间戳（版本号）、该版本的值、操作类型。如果设置isExportSequenceInfo为true，还会包括时序信息。

转换为数据集成的数据格式后，定义了以下四种操作类型：

- U（UPDATE）：写入一列的一个版本。
- DO（DELETE_ONE_VERSION）：删除某一列的某个版本。
- DA（DELETE_ALL_VERSION）：删除某一列的全部版本，此时需要根据主键和列名，删除对应列的全部版本。
- DR（DELETE_ROW）：删除某一行，此时需要根据主键，删除该行数据。

假设该表有两个主键列，主键列名分别为pkName1， pkName2，示例如下。

pkName1	pkName2	columnName	timestamp	columnValue	opType
pk1_V1	pk2_V1	col_a	1441803688001	col_val1	U
pk1_V1	pk2_V1	col_a	1441803688002	col_val2	U
pk1_V1	pk2_V1	col_b	1441803688003	col_val3	U
pk1_V2	pk2_V2	col_a	1441803688000	—	DO
pk1_V2	pk2_V2	col_b	—	—	DA
pk1_V3	pk2_V3	—	—	—	DR
pk1_V3	pk2_V3	col_a	1441803688005	col_val1	U

假设导出的数据如上，共7行，对应Table Store表内的3行，主键分别是（pk1_V1，pk2_V1），（pk1_V2， pk2_V2），（pk1_V3， pk2_V3）：

- 对于主键为（pk1_V1，pk2_V1）的一行，包括写入col_a列的两个版本和col_b列的一个版本等操作。
- 对于主键为（pk1_V2，pk2_V2）的一行，包括删除col_a列的一个版本和删除col_b列的全部版本等操作。
- 对于主键为（pk1_V3，pk2_V3）的一行，包括删除整行和写入col_a列的一个版本等操作。

行模式

您可以通过行模式导出数据，该模式将用户每次更新的记录，抽取成行的形式导出，需要设置mode属性并配置列名。

```
"parameter": {
  #parameter中配置下面三项配置（例如datasource、table等其它配置项照常配置）。
  "mode": "single_version_and_update_only", # 配置导出模式。
  "column":[  #按照需求添加需要导出TableStore中的列，您可以自定义设置配置个数。
          {
             "name": "uid"  #列名示例，可以是主键或属性列。
          },
          {
             "name": "name"  #列名示例，可以是主键或属性列。
          },
  ],
  "isExportSequenceInfo": false, #single_version_and_update_only模式下只能是false。
}
```

行模式导出的数据更接近于原始的行，易于后续处理，但需要注意以下问题：

- 每次导出的行是从用户每次更新的记录中抽取，每一行数据与用户的写入或更新操作一一对应。如果用户存在单独更新某些列的行为，则会出现有一些记录只有被更新的部分列，其它列为空的情况。
- 行模式不会导出数据的版本号（即每列的时间戳），也无法进行删除操作。

数据类型转换列表

目前OTSStream Reader支持所有的Table Store类型，其针对Table Store类型的转换列表，如下所示。

类型分类	OTSStream数据类型
整数类	INTEGER
浮点类	DOUBLE
字符串类	STRING
布尔类	BOOLEAN
二进制类	BINARY

参数说明

参数	描述	是否必选	默认值
dataSource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
dataTable	导出增量数据的表的名称。该表需要开启Stream，可以在建表时开启，或者使用UpdateTable接口开启。	是	无
statusTable	Reader插件用于记录状态的表的名称，这些状态可用于减少对非目标范围内的数据的扫描，从而加快导出速度。statusTable是Reader用于保存状态的表，如果该表不存在，Reader会自动创建该表。一次离线导出任务完成后，您无需删除该表，该表中记录的状态可用于下次导出任务中： • 您无需创建该表，只需要给出一个表名。Reader插件会尝试在您的instance下创建该表，如果该表不存在即创建新表。如果该表已存在，会判断该表的Meta是否与期望一致，如果不一致会抛出异常。 • 在一次导出完成之后，您无需删除该表，该表的状态可以用于下次的导出任务。 • 该表会开启TTL，数据自动过期，会认为其数据量很小。 • 针对同一个instance下的多个不同的dataTable的Reader配置，可以使用同一个statusTable，记录的状态信息互不影响。 您配置一个类似TableStoreStreamReaderStatusTable的名称即可，请注意不要与业务相关的表重名。	是	无
startTimestampMillis	增量数据的时间范围（左闭右开）的左边界，单位为毫秒： • Reader插件会从statusTable中找对应startTimestampMillis的位点，从该点开始读取开始导出数据。 • 如果statusTable中找不到对应的位点，则从系统保留的增量数据的第一条开始读取，并跳过写入时间小于startTimestampMillis的数据。	否	无
endTimestampMillis	增量数据的时间范围（左闭右开）的右边界，单位为毫秒： • Reader插件从startTimestampMillis位置开始导出数据后，当遇到第一条时间戳大于等于endTimestampMillis的数据时，结束导出数据，导出完成。 • 当读取完当前全部的增量数据时，即使未达到endTimestampMillis，也会结束读取。	否	无
date	日期格式为yyyyMMdd，例如20151111，表示导出该日的数据。如果没有指定date，则必须指定startTimestampMillis和endTimestampMillis，反之也成立。例如，采云间调度仅支持天级别，所以提供该配置，作用与startTimestampMillis和endTimestampMillis类似。	否	无

参数	描述	是否必选	默认值
isExportSequenceInfo	是否导出时序信息，时序信息包含了数据的写入时间等。默认该值为*false*，即不导出。	否	*false*
maxRetries	从TableStore中读增量数据时，每次请求的最大重试次数，默认为30次。重试之间有间隔，重试30次的总时间约为5分钟，通常无需更改。	否	*30*
startTimeString	任务的开始时间，即增量数据的时间范围（左闭右开）的左边界，格式为 `yyyymmddhh24miss` ，单位为秒。	否	无
endTimeString	任务的结束时间，即增量数据的时间范围（左闭右开）的右边界，格式为 `yyyymmddhh24miss` ，单位为秒。	否	无
mode	导出模式，设置为single_version_and_update_only时为行模式，默认不设置为列模式。	否	无

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的dataSource，通常选择您配置的数据源名称。
表	即上述参数说明中的dataTable。
开始时间	任务的开始时间，即增量数据的时间范围（左闭右开）的左边界。格式为 `yyyymmddhh24miss` ，单位为秒。 **说明** 开始时间需要设置为最近七天（包括当天）。
结束时间	任务的结束时间，即增量数据的时间范围（左闭右开）的右边界。格式为 `yyyymmddhh24miss` ，单位为秒。
状态表	用于记录状态的表的名称。
最大重试次数	即上述参数说明中的maxRetries，默认值为30。
导出时序信息	即上述参数说明中的isExportSequenceInfo，默认值为*false*。

2. 字段映射。

左侧的源头表字段和右侧的目标表字段为一一对应的关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除。

参数	说明
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置样例如下所示，具体参数填写请参见参数说明。使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"otsstream",//插件名。
            "parameter":{
                "statusTable":"TableStoreStreamReaderStatusTable",//用于记录状态的表的名称。
                "maxRetries":30,//从 TableStore 中读增量数据时，每次请求的最大重试次数，默认为30。
                "isExportSequenceInfo":false,//是否导出时序信息。
                "datasource":"$srcDatasource",//数据源。
                "startTimeString":"${startTime}${hh}",//增量数据的时间范围（左闭右开）的左边界。参数配置中配置startTime=${yyyymmdd} hh=$[hh24miss],表示ots读取开始时间为业务日期的定时时间
                "table":"",//表名。
                "endTimeString":"${endTime}${hh}"//增量数据的时间范围（左闭右开）的右边界。参数配置中配置endTime=${yyyymmdd} hh=$[hh24miss]，表示ots读取结束时间为业务日期的定时时间
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.2.29. MetaQ Reader

本文为您介绍MetaQ Reader支持的数据类型、字段映射和数据源等参数及配置示例。

> 注意 MetaQ Reader仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

背景信息

消息队列Message Queue（简称MQ）是阿里巴巴集团自主研发的专业消息中间件。消息队列基于高可用分布式集群技术，为您提供消息发布订阅、消息轨迹查询、定时（延时）消息、资源统计和监控报警等消息云服务。消息队列为分布式应用系统提供异步解耦的功能，同时具备海量消息堆积、高吞吐等互联网应用所需要的特性，是阿里巴巴集团双11使用的核心产品。

MetaQ Reader使用消息队列的Java SDK消费消息队列中的实时数据，将数据转换为数据集成传输协议传递给Writer。

实现原理

MetaQ Reader通过消息队列服务的Java SDK订阅MetaQ中的实时消息数据，使用的Java SDK版本如下所示。

```
<dependency>
            <groupId>com.taobao.metaq.final</groupId>
            <artifactId>metaq-client</artifactId>
            <version>4.0.1</version>
        </dependency>
        <dependency>
            <groupId>com.aliyun.openservices</groupId>
            <artifactId>ons-sdk</artifactId>
            <version>1.3.1</version>
        </dependency>
```

类型转换列表

MetaQ Reader针对MetaQ类型的转换列表，如下所示。

数据集成数据类型	消息队列数据类型
STRING	STRING

参数说明

参数	描述	是否必选
accessId	访问消息队列的访问密钥，用于标识用户。	是
accessKey	访问消息队列的访问密钥，用来验证用户的密钥。	是
consumerId	Consumer是消息的消费者，也称为消息订阅者，负责接收并消费消息。 consumerId是一类Consumer的标识，该类Consumer通常接收并消费一类消息，且消费逻辑一致。	是
topicName	消息主题，一级消息类型，通过topic对消息进行分类。	是
subExpression	消息子主题。	是
onsChannel	用于进行消息队列鉴权。	是
unitName	接收消息的目标单元。常用单元如下： • sh：中心 • unsz：深圳单元 • us：美国 • en-us：欧洲 • rg-ru：俄罗斯 • zbyk：张北优酷 • unzbyun：张北云 • unshyun：上海云 • lazada-sg：新加坡lazada • lazada-my：马来西亚lazada • lazada-vn：越南lazada • lazada-ph：菲律宾lazada • lazada-th：泰国lazada • lazada-id：印尼lazada	否

参数	描述	是否必选
instanceName	Consumer的实例名称。	否
domainName	消息队列的接入点。	是
contentType	消息的类型，支持*singlestringcolumn*（消息为STRING类型）、*text*（消息为文本类型）和*json*（消息为JSON类型）。	是
beginOffset	任务开始读取的Offset，支持begin（从最开始），lastRead（上次读取的Offset）。	是
nullCurrentOffset	上次Offset为空时，开始读取的地方。支持begin（从最开始），current（当前offset）。	是
fieldDelimiter	分隔符模式下消息字符串的列分隔符，例如逗号等。支持控制字符，例如 *u0001*。	是
column	读取的字段列表。	是
beginDateTime	数据消费的开始时间位点，为时间范围（左闭右开）的左边界。 beginDateTime是yyyyMMddHHmmss格式的时间字符串，可以和DataWorks的调度时间参数配合使用。	否 **说明** beginDateTime和endDateTime配合使用。
endDateTime	数据消费的结束时间位点，为时间范围（左闭右开）的右边界。 endDateTime是yyyyMMddHHmmss格式的时间字符串，可以和DataWorks的调度时间参数配合使用。	

功能说明

配置一个从消息队列读取数据的示例，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "job": {
        "content": [
            {
                "reader": {
                    "name": "metaqreader",
                    "parameter": {
                        "accessId": "<yourAccessKeyId>",
                        "accessKey": "<yourAccessKeySecret>",
                        "consumerId": "Test01",
                        "topicName": "test",
                        "subExpression": "*",
                        "onsChannel": "ALIYUN",
                        "domainName": "***.aliyun.com",
                        "contentType": "singlestringcolumn",
                        "beginOffset": "lastRead",
                        "nullCurrentOffset": "begin",
                        "fieldDelimiter": ",",
                        "column": [
                            "col0"
                        ],
                        "fieldDelimiter": ","
                    }
                },
                "writer": {
                    "name": "streamwriter",
                    "parameter": {
                        "print": false
                    }
                }
            }
        ]
    }
}
```

6.2.30. Hive Reader

Hive Reader插件实现了从Hive读取数据的功能，本文为您介绍Hive Reader的工作原理、参数和示例。

背景信息

Hive是基于Hadoop的数据仓库工具，用于解决海量结构化日志的数据统计。Hive可以将结构化的数据文件映射为一张表，并提供SQL查询功能。Hive的本质是一个SQL解析引擎，其底层通过MapReduce实现数据分析，使用HDFS存储处理的数据，将HQL或SQL语句转化为MapReduce程序并在Yarn上运行。

Hive Reader插件通过访问Hive元数据库，解析出您配置的数据表的HDFS文件存储路径、文件格式、分隔符等信息后，再通过读取HDFS文件的方式读取Hive中的表数据。

Hive Reader插件通过访问HiveMetastore服务，获取您配置的数据表的元数据信息。您可以基于HDFS文件和Hive JDBC两种方式读取数据：

- 基于HDFS文件读取数据

  Hive Reader插件通过访问HiveMetastore服务，解析出您配置的数据表的HDFS文件存储路径、文件格式、分隔符等信息后，再通过读取HDFS文件的方式读取Hive中的表数据。

  Hive Reader的底层逻辑和HDFS Reader插件一致，读取数据后，再通过Hive JDBC将文件中的数据导入到目标表中。您可以在Hive Reader插件参数中配置HDFS Reader相关参数，配置的参数会透传给HDFS Reader插件。

- 基于Hive JDBC读取数据

  Hive Reader插件通过Hive JDBC客户端连接HiveServier2服务读取数据。Hive Reader支持通过where条件过滤数据，并支持直接通过SQL读取数据。

使用限制

- Hive Reader仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。
- Hive Reader支持的版本请参见下文的版本支持汇总。
- 目前仅支持读取TextFile、ORCFile和ParquetFile三种格式的文件。

支持的数据类型

类型分类	S3数据类型
字符串类	CHAR、VARCHAR、STRING
整数类	TINYINT、SMALLINT、INT、INTEGER、BIGINT
浮点类	FLOAT、DOUBLE、DECIMAL
日期时间类	TIMESTAMP、DATE
布尔型	BOOLEAN

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，必须与添加的数据源名称保持一致。	是	无
table	表名，用于指定需要同步的表。 说明 请注意大小写。	是	无
readMode	读取方式： • 基于HDFS文件方式读取数据，配置为 `"readMode":"hdfs"` 。 • 基于Hive JDBC方式读取数据，配置为 `"readMode":"jdbc"` 。 说明 • 基于Hive JDBC方式读取数据时，支持使用WHERE条件做数据过滤。基于HDFS文件方式读取数据时，不支持使用WHERE条件做数据过滤。 • 基于Hive JDBC方式读取数据时，不支持用户自定义并发数。基于HDFS文件方式读取数据时，支持用户自定义并发数，。	否	无
partition	Hive表的分区信息： • 如果您基于Hive JDBC读取数据，无需配置该参数。 • 如果您读取的Hive表是分区表，您需要配置partition信息。同步任务会读取partition对应的分区数据。 Hive Reader支持使用星号（*）作为通配符，格式为 `pt1=a,pt2=b,...` 。 • 如果您的Hive表是非分区表，则无需配置partition。	否	无

参数	描述	是否必选	默认值
column	需要读取的字段列，例如 `"column": ["id", "name"]` 。 • 支持列裁剪：即可以导出部分列。 • 支持列换序，即可以不根据表Schema信息顺序导出列。 • 支持配置分区列。 • 支持配置常量。 • column必须显示指定同步的列集合，不允许为空。	是	无
querySql	当您基于Hive JDBC方式读取数据时，可以直接配置querySql读取数据。	否	无
where	当您基于Hive JDBC方式读取数据时，可以通过设置where条件过滤数据。	否	无

向导开发介绍

在**数据开发**页面，双击打开新建的数据同步节点，即可在右侧的编辑页面配置任务。详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

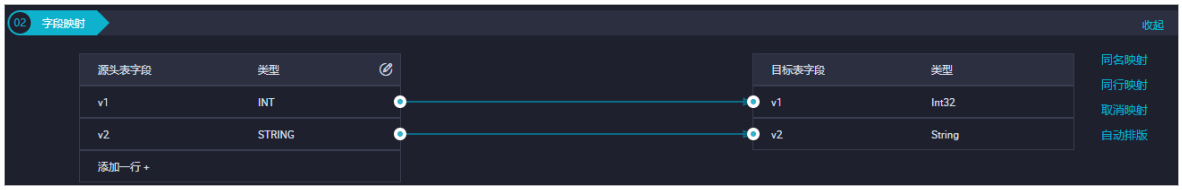
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常选择您配置的数据源名称。
表	即上述参数说明中的table。
读取Hive方法	即上述参数说明中的readMode，包括**基于HDFS文件读取数据**和**基于Hive JDBC读取数据**。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

参数	描述
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为自定义。 ○ 可以输入分区列名，以同步分区列。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。 **说明** 基于Hive JDBC方式读取数据时，不支持用户自定义并发数。基于HDFS文件方式读取数据时，支持用户自定义并发数。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

您可以基于HDFS文件和Hive JDBC读取数据：

- 基于HDFS文件读取数据

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "hive",
            "parameter": {
                "partition": "pt1=a,pt2=b,pt3=c", //分区信息
                "datasource": "hive_not_ha_****", //数据源名称
                "column": [ //需要读取的字段列
                    "id",
                    "pt2",
                    "pt1"
                ],
                "readMode": "hdfs", //读取方式
                "table": "part_table_1"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "hive",
            "parameter": {
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "" //错误记录数
        },
        "speed": {
            "concurrent": 2, //作业并发数
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    }
}
```

- 基于Hive JDBC读取数据

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "hive",
            "parameter": {
                "querySql": "select id,name,age from part_table_1 where pt2='B'",
                "datasource": "hive_not_ha_****",  //数据源名称
                "column": [ //需要读取的字段列
                    "id",
                    "name",
                    "age"
                ],
                "where": "",
                "table": "part_table_1",
                "readMode": "jdbc" //读取方式
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "hive",
            "parameter": {
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "concurrent": 2,  //作业并发数
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    }
}
```

配置数据集成资源组

1. 单击数据同步任务编辑页面右侧的**数据集成资源组配置**。
2. 根据提示选择对应的**独享数据集成资源组**。

> **说明** **数据集成资源组配置**页面默认支持选择独享数据集成资源组，为确保数据同步的稳定性和性能要求，推荐使用独享数据集成资源组。Hive数据源不支持使用公共资源组和自定义资源组。

版本支持汇总

Hive Reader支持的版本如下。

```
0.8.0
0.8.1
0.9.0
0.10.0
0.11.0
0.12.0
0.13.0
0.13.1
0.14.0
1.0.0
1.0.1
1.1.0
1.1.1
1.2.0
1.2.1
1.2.2
2.0.0
2.0.1
2.1.0
2.1.1
2.2.0
2.3.0
2.3.1
2.3.2
2.3.3
2.3.4
2.3.5
2.3.6
2.3.7
3.0.0
3.1.0
```

```
3.1.1
3.1.2
0.8.1-cdh4.0.0
0.8.1-cdh4.0.1
0.9.0-cdh4.1.0
0.9.0-cdh4.1.1
0.9.0-cdh4.1.2
0.9.0-cdh4.1.3
0.9.0-cdh4.1.4
0.9.0-cdh4.1.5
0.10.0-cdh4.2.0
0.10.0-cdh4.2.1
0.10.0-cdh4.2.2
0.10.0-cdh4.3.0
0.10.0-cdh4.3.1
0.10.0-cdh4.3.2
0.10.0-cdh4.4.0
0.10.0-cdh4.5.0
0.10.0-cdh4.5.0.1
0.10.0-cdh4.5.0.2
0.10.0-cdh4.6.0
0.10.0-cdh4.7.0
0.10.0-cdh4.7.1
0.12.0-cdh5.0.0
0.12.0-cdh5.0.1
0.12.0-cdh5.0.2
0.12.0-cdh5.0.3
0.12.0-cdh5.0.4
0.12.0-cdh5.0.5
0.12.0-cdh5.0.6
0.12.0-cdh5.1.0
0.12.0-cdh5.1.2
0.12.0-cdh5.1.3
0.12.0-cdh5.1.4
0.12.0-cdh5.1.5
0.13.1-cdh5.2.0
0.13.1-cdh5.2.1
0.13.1-cdh5.2.2
0.13.1-cdh5.2.3
0.13.1-cdh5.2.4
0.13.1-cdh5.2.5
0.13.1-cdh5.2.6
0.13.1-cdh5.3.0
0.13.1-cdh5.3.1
0.13.1-cdh5.3.2
0.13.1-cdh5.3.3
0.13.1-cdh5.3.4
0.13.1-cdh5.3.5
0.13.1-cdh5.3.6
0.13.1-cdh5.3.8
0.13.1-cdh5.3.9
0.13.1-cdh5.3.10
1.1.0-cdh5.3.6
1.1.0-cdh5.4.0
1.1.0-cdh5.4.1
1.1.0-cdh5.4.2
1.1.0-cdh5.4.3
1.1.0-cdh5.4.4
1.1.0-cdh5.4.5
1.1.0-cdh5.4.7
1.1.0-cdh5.4.8
1.1.0-cdh5.4.9
1.1.0-cdh5.4.10
1.1.0-cdh5.4.11
1.1.0-cdh5.5.0
```

```
1.1.0-cdh5.5.1
1.1.0-cdh5.5.2
1.1.0-cdh5.5.4
1.1.0-cdh5.5.5
1.1.0-cdh5.5.6
1.1.0-cdh5.6.0
1.1.0-cdh5.6.1
1.1.0-cdh5.7.0
1.1.0-cdh5.7.1
1.1.0-cdh5.7.2
1.1.0-cdh5.7.3
1.1.0-cdh5.7.4
1.1.0-cdh5.7.5
1.1.0-cdh5.7.6
1.1.0-cdh5.8.0
1.1.0-cdh5.8.2
1.1.0-cdh5.8.3
1.1.0-cdh5.8.4
1.1.0-cdh5.8.5
1.1.0-cdh5.9.0
1.1.0-cdh5.9.1
1.1.0-cdh5.9.2
1.1.0-cdh5.9.3
1.1.0-cdh5.10.0
1.1.0-cdh5.10.1
1.1.0-cdh5.10.2
1.1.0-cdh5.11.0
1.1.0-cdh5.11.1
1.1.0-cdh5.11.2
1.1.0-cdh5.12.0
1.1.0-cdh5.12.1
1.1.0-cdh5.12.2
1.1.0-cdh5.13.0
1.1.0-cdh5.13.1
1.1.0-cdh5.13.2
1.1.0-cdh5.13.3
1.1.0-cdh5.14.0
1.1.0-cdh5.14.2
1.1.0-cdh5.14.4
1.1.0-cdh5.15.0
1.1.0-cdh5.16.0
1.1.0-cdh5.16.2
1.1.0-cdh5.16.99
2.1.1-cdh6.1.1
2.1.1-cdh6.2.0
2.1.1-cdh6.2.1
2.1.1-cdh6.3.0
2.1.1-cdh6.3.1
2.1.1-cdh6.3.2
2.1.1-cdh6.3.3
```

6.2.31. Vertica Reader

Vertica是一款基于列存储的MPP架构的数据库，Vertica Reader插件实现了从Vertica读取数据的功能。本文为您介绍Vertica Reader的实现原理、参数和示例。

> **注意** Vertica Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

实现原理

在底层实现上，Vertica Reader通过JDBC连接远程Vertica数据库，并执行相应的SQL语句，从Vertica数据库中读取数据。

Vertica Reader通过JDBC连接器连接至远程的Vertica数据库，根据您配置的信息生成查询SQL语句，发送至远程Vertica数据库，执行该SQL并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

- 对于您配置的table、column和where等信息，Vertica Reader将其拼接为SQL语句发送至Vertica数据库。
- 对于您配置的querySql信息，Vertica直接将其发送至Vertica数据库。

Vertica Reader通过Vertica数据库驱动访问Vertica，您需要确认Vertica驱动和您的Vertica服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>com.vertica</groupId>
    <artifactId>vertica-jdbc</artifactId>
    <version>7.1.2</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
jdbcUrl	描述的是到对端数据库的JDBC连接信息，使用JSON的数组进行描述，并支持一个库填写多个连接地址。 如果配置多个连接地址，Vertica Reader可以依次验证IP的可连接性，直到选择一个合法的IP。如果全部连接失败，则Vertica Reader报错。 **说明** jdbcUrl必须包含在connection配置单元中。 jdbcUrl的格式和Vertica官方一致，并可以连接附件控制信息。例如，`jdbc:vertica://1**.0.0.1:3306/database`。	否	无
username	数据源的用户名。	否	无
table	选取的需要同步的表名称。	是	无
password	数据源指定用户名的密码。	否	无
table	选取的需要同步的表。使用JSON的数组进行描述，支持同时读取多张表。 当配置为多张表时，您需要保证多张表的schema结构一致，Vertica Reader不检查表的逻辑是否统一。 **说明** table必须包含在connection配置单元中。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表schema信息顺序进行导出。 • 支持常量配置。 • column必须显示指定同步的列集合，不允许为空。	是	无

参数	描述	是否必选	默认值
splitPk	Vertica Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点、日期等其它类型 。如果您指定其它非支持类型，Vertica Reader将报错。 • 如果设置splitPk为空，底层将视作您不允许对单表进行切分，因此使用单通道进行抽取。	否	无
where	筛选条件，Vertica Reader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。 例如在测试时，可以指定where条件。在实际业务场景中，通常会选择当天的数据进行同步，可以将where条件指定为 `gmt_create > $bizdate` 。 • where条件可以有效地进行业务增量同步。 • where条件不配置或者为空，视作全表同步数据。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。配置该项后，数据同步系统会忽略tables、columns和splitPk配置项，直接使用该项配置的内容对数据进行筛选。 当您配置querySql时，Vertica Reader直接忽略table、column和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够较大地提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个从Vertica读取数据的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "vertica", //插件名。
            "parameter": {
                "datasource": "", //数据源名。
                "username": "",
                "password": "",
                "where": "",
                "column": [ //字段。
                    "id",
                    "name"
                ],
                "splitPk": "id",
                "connection": [
                    {
                        "table": [ //表名。
                            "table"
                        ],
                        "jdbcUrl": [
                            "jdbc:vertica://host:port/database"
                        ]
                    }
                ]
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {
                "print": false,
                "fieldDelimiter": ","
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0" //错误记录数。
        },
        "speed": {
            "throttle": true, //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent": 1, //作业并发数。
                        "mbps":"12"//限流
        }
    }
}
```

6.2.32. Gbase8a Reader

本文为您介绍Gbase8a Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

Gbase8a是一款基于列存储的新型分析型数据库，Gbase8a Reader插件实现了从Gbase8a读取数据的功能。

> **注意** 目前Gbase8a Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

Gbase8a Reader通过JDBC连接器连接至远程的Gbase8a数据库，根据您配置的信息生成查询SQL语句，发送至远程Gbase8a数据库，执行该SQL并返回结果。然后使用数据同步自定义的数据类型拼装返回的结果为抽象的数据集，传递给下游Writer处理。

- 对于您配置的table、column和where等信息，Gbase8a Reader将其拼接为SQL语句发送至Gbase8a数据库。
- 对于您配置的querySql信息，Gbase8a直接将其发送至Gbase8a数据库。

Gbase8a Reader通过MySQL数据库驱动访问Gbase8a（复用MySQL协议），您需要确认驱动和您的Gbase8a服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>mysql</groupId>
    <artifactId>mysql-connector-java</artifactId>
    <version>5.1.22</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
datasource	如果您使用的DataWorks版本支持添加Gbase8a数据源，即可在此处根据数据源名称引用您添加的Gbase8a数据源。 包括jdbcUrl和username两种配置方式。	否	无
jdbcUrl	指的是到对端数据库的JDBC连接信息。使用JSON的数组描述，并支持一个库填写多个连接地址。 如果配置了多个，Gbase8a Reader可以依次探测IP的可连接性，直到选择一个合法的IP。 如果全部连接失败，则Gbase8a Reader报错。 **说明** jdbcUrl必须包含在connection配置单元中。 jdbcUrl根据Gbase8a官方规范，可以填写连接附件控制信息。例如，`jdbc:mysql://127.0.0.1:3306/database`，需要和username配置方式二选一。	否	无
username	数据源的用户名。	否	无
password	数据源指定用户名的密码。	否	无
table	选取的需要同步的表。使用JSON的数组进行描述，支持同时读取多张表。 当配置为多张表时，您需要保证多张表的schema结构一致，Gbase8a Reader不检查表的逻辑是否统一。 **说明** table必须包含在connection配置单元中。	是	无

参数	描述	是否必选	默认值
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪：列可以挑选部分列进行导出。 • 支持列换序：列可以不按照表schema信息顺序进行导出。 • 支持常量配置：例如，`'123'` 。 • 支持函数列：例如，`date('now')` 。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	Gbase8a Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点和日期等其它类型 。如果您指定其它非支持类型，则忽略splitPk功能，使用单通道进行同步。 • 如果设置splitPk值为空，底层将视作您不允许对单表进行切分，因此使用单通道进行抽取。	否	空
where	筛选条件，Gbase8a Reader根据指定的column、table、where条件拼接SQL，并根据该SQL进行数据抽取。 例如，在进行测试时，可以指定where条件为limit 10。在实际业务场景中，通常会选择当天的数据进行同步，指定where条件为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。 • where条件不配置或为空，则视作全表同步数据。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。配置该项后，数据同步系统会忽略tables、columns和splitPk配置项，直接使用该项配置的内容对数据进行筛选。 当您配置querySql时，Gbase8a Reader直接忽略table、column、where和splitPk条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够较大地提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个从Gbase8a读取数据的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "gbase8a", //插件名。
            "parameter": {
                "datasource": "", //数据源名。
                "username": "",
                "password": "",
                "where": "",
                "column": [ //字段。
                    "id",
                    "name"
                ],
                "splitPk": "id",
                "connection": [
                    {
                        "table": [ //表名。
                            "table"
                        ],
                        "jdbcUrl": [
                            "jdbc:mysql://host:port/database"
                        ]
                    }
                ]
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {
                "print": false,
                "fieldDelimiter": ","
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0" //错误记录数。
        },
        "speed": {
            "throttle": true, //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent": 1, //作业并发数。
            "mbps":"12"//限流
        }
    }
}
```

6.2.33. DataHub Reader

阿里云流数据处理平台DataHub是流式数据（Streaming Data）的处理平台，为您提供发布（Publish）、订阅（Subscribe）和分发流式数据的功能，让您可以轻松构建基于流式数据的分析和应用。

DataHub Reader通过DataHub的Java SDK读取DataHub中的数据，具体使用的Java SDK版本，如下所示。

```
<dependency>
    <groupId>com.aliyun.DataHub</groupId>
    <artifactId>aliyun-sdk-DataHub</artifactId>
    <version>2.9.1</version>
</dependency>
```

参数说明

参数	描述	是否必选
endpoint	DataHub的endpoint。	是
accessId	访问DataHub的用户accessId。	是
accessKey	访问DataHub的用户accessKey。	是
project	目标DataHub的项目名称。project是DataHub中的资源管理单元，用于资源隔离和控制。	是
topic	目标DataHub的topic名称。	是
batchSize	一次读取的数据量，默认为*1,024*条。	否
beginDateTime	数据消费的开始时间位点。该参数是时间范围（左闭右开）的左边界，yyyyMMddHHmmss格式的时间字符串，可以和DataWorks的调度时间参数配合使用。 **说明** beginDateTime和endDateTime需要互相组合配套使用。	是
endDateTime	数据消费的结束时间位点。该参数是时间范围（左闭右开）的右边界，yyyyMMddHHmmss格式的时间字符串，可以和DataWorks的调度时间参数配合使用。 **说明** beginDateTime和endDateTime需要互相组合配套使用。	是

向导开发介绍

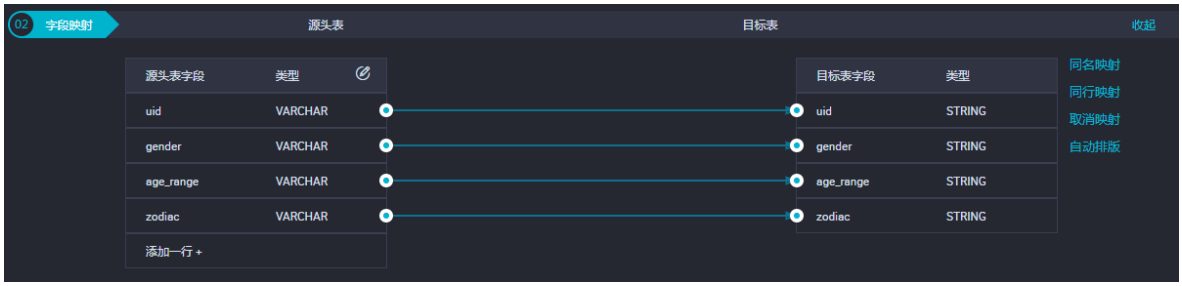
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	在下拉列表中选择您配置的数据源名称。
主题	即上述参数说明中的topic。
消费开始时间	即上述参数说明中的beginDateTime。
消费结束时间	即上述参数说明中的endDateTime。
批量条数	即上述参数说明中的batchSize。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段。鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其它空行会被忽略。
添加一行	添加一行的功能如下所示： ○ 可以输入常量，输入的值需要使用英文单引号。例如，'abc'、'123'等。 ○ 可以配合调度参数使用。例如，${bizdate}等。 ○ 可以输入关系数据库支持的函数。例如，now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为**未识别**。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从DataHub读取数据的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "job": {
        "content": [
            {
                "reader": {
                    "name": "DataHubreader",
                    "parameter": {
                        "endpoint": "xxx" //DataHub的endpoint。
                        "accessId": "xxx", //访问DataHub的用户accessId。
                        "accessKey": "xxx", //访问DataHub的用户accessKey。
                        "project": "xxx", //目标DataHub的项目名称。
                        "topic": "xxx" //目标DataHub的topic名称。
                        "batchSize": 1000, //一次读取的数据量。
                        "beginDateTime": "20180910111214", //数据消费的开始时间位点。
                        "endDateTime": "20180910111614", //数据消费的结束时间位点。
                        "column": [
                            "col0",
                            "col1",
                            "col2",
                            "col3",
                            "col4"
                        ]
                    }
                },
                "writer": {
                    "name": "streamwriter",
                    "parameter": {
                        "print": false
                    }
                }
            }
        ]
    }
}
```

6.2.34. ApsaraDB For OceanBase Reader

ApsaraDB For OceanBase是阿里云和蚂蚁金服自主研发的金融级分布式关系数据库，本文为您介绍ApsaraDB For OceanBase Reader的实现原理、参数说明及配置示例。

> **注意** 目前ApsaraDB For OceanBase Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

背景信息

ApsaraDB For OceanBase在金融行业创造了三地五中心的城市级故障自动无损容灾的新标准，在普通硬件上实现了金融高可用。同时具备在线水平扩展能力，是在功能、稳定性、可扩展性、性能方面都经历严格检验的国产数据库。

ApsaraDB For OceanBase Reader实现了从ApsaraDB For OceanBase读取数据，支持Oracle和MySQL两种模式的同步功能。

在底层实现上，ApsaraDB For OceanBase Reader通过JDBC连接远程ApsaraDB For OceanBase数据库，并执行相应的SQL语句，从ApsaraDB For OceanBase库中选取数据。

ApsaraDB For OceanBase Reader通过Java客户端连接器连接至远程的OceanBase数据库，根据您配置的信息生成查询SQL语句，发送至远程OceanBase数据库，执行该SQL并返回结果。然后使用数据集成自定义的数据类型拼装返回的结果为抽象的数据集，传递给下游Writer处理。

- 对于您配置的table、column和where等信息，ApsaraDB For OceanBase Reader将其拼接为SQL语句发送至ApsaraDB For OceanBase数据库。
- 对于您配置的querySql信息，ApsaraDB For OceanBase直接将其发送至ApsaraDB For OceanBase数据库。

> **说明** ApsaraDB For OceanBase包括Oracle和MySQL两种租户模式，您在配置where数据过滤条件、column中的函数列时，需要符合对应租户模式的SQL语法约束，否则SQL语句可能执行失败。

ApsaraDB For OceanBase Reader通过OceanBase数据库驱动访问ApsaraDB For OceanBase，您需要确认驱动和您的ApsaraDB For OceanBase服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>com.alipay.OceanBase</groupId>
    <artifactId>OceanBase-connector-java</artifactId>
    <version>3.1.0</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
datasource	如果您使用的DataWorks版本支持添加ApsaraDB For OceanBase数据源，即可在此处根据数据源名称引用您添加的ApsaraDB For OceanBase数据源。 包括jdbcUrl和username两种配置方式。	是	无
jdbcUrl	到对端数据库的JDBC连接信息。使用JSON的数组描述，并支持一个库填写多个连接地址。 如果配置了多个，ApsaraDB For OceanBase Reader可以依次探测IP的可连接性，直到选择一个合法的IP。 如果全部连接失败，则ApsaraDB For OceanBase Reader报错。 **说明** jdbcUrl必须包含在connection配置单元中。 jdbcUrl根据ApsaraDB For OceanBase官方规范，可以填写连接附件控制信息。例如 `jdbc:OceanBase://127.0.0.1:3306/database`，需要和username配置方式二选一。	否	无

参数	描述	是否必选	默认值
username	数据源的用户名。	否	无
password	数据源指定用户名的密码。	否	无
table	选取的需要同步的表。使用JSON的数组进行描述，支持同时读取多张表。 当配置为多张表时，您需要保证多张表的Schema结构一致，ApsaraDB For OceanBase Reader不检查表的逻辑是否统一。 **说明** table必须包含在connection配置单元中。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息 。默认使用所有列配置，例如[*]。 • 支持列裁剪：可以导出部分列。 • 支持列换序：可以不按照表Schema信息顺序进行导出。 • 支持常量配置：例如 `'123'` 。 • 支持函数列：例如 `date('now')` 。 • column必须显示指定同步的列集合，不允许为空。	是	无
splitPk	ApsaraDB For OceanBase Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。 • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持字符串、浮点和日期等其它类型 。如果您指定其它非支持类型，ApsaraDB For OceanBase Reader将报错。 • 如果设置splitPk值为空，底层将视作您不允许对单表进行切分，因此使用单通道进行抽取。	否	空
where	ApsaraDB For OceanBase Reader根据指定的column、table、where条件拼接SQL，并根据该SQL进行数据抽取。 例如，在进行测试时，可以指定where条件为limit 10。在实际业务场景中，通常会选择当天的数据进行同步，指定where条件为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。 • where条件不配置或为空，则视作全表同步数据。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。配置该项后，数据同步系统会忽略tables、columns和splitPk配置项，直接使用该项配置的内容对数据进行筛选。 当您配置querySql时，ApsaraDB For OceanBase Reader直接忽略table、column、where和splitPk条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据集成和服务器端的网络交互次数，能够较大地提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个从ApsaraDB For OceanBase读取数据的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "apsaradb_for_OceanBase", //插件名
            "parameter": {
                "datasource": "", //数据源名
                "where": "",
                "column": [ //字段
                    "id",
                    "name"
                ],
                "splitPk": ""
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "stream",
            "parameter": {
                "print": false,
                "fieldDelimiter": ","
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0" //错误记录数
        },
        "speed": {
            "throttle": true, //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent": 1, //作业并发数
            "mbps":"12"//限流
        }
    }
}
```

6.2.35. Hologres Reader

Hologres Reader实现了从交互式分析（Hologres）数仓导出数据的功能，您可以根据数据集成标准协议从Hologres表中导出数据至其它数据源。

背景信息

> 注意 Hologres Reader仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

Hologres Reader通过PSQL读取Hologres表中的数据，根据表的Shard Count发起多个并发，每个Shard对应一个Select并发任务：

- Hologres在创建表时，在同一个 `CREATE TABLE` 事务中，通过 `CALL set_table_property('table_name', 'shard_count', 'xx')` 配置表的Shard Count。

  默认情况下，使用数据库默认的Shard Count，具体数值取决于Hologres实例的配置。
- Select语句通过表的内置列hg_shard_id的Shard筛选数据。

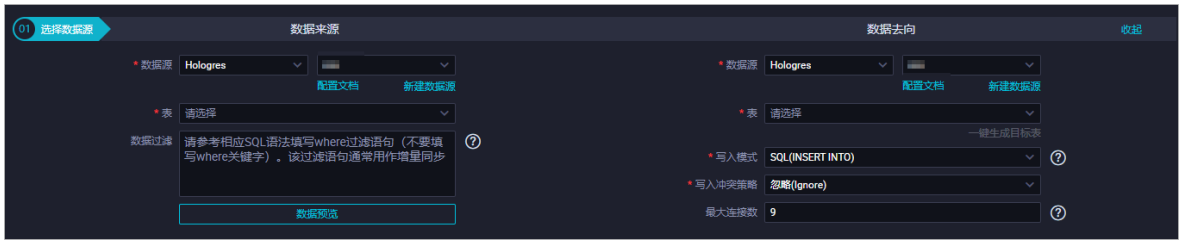
参数说明

参数	描述	是否必选	默认值
endpoint	目标交互式分析（Hologres）实例对应的endpoint，格式为 `instance-id-region-endpoint.hologres.aliyuncs.com:port` 。您可以从交互式分析实例的管理页面获取。 endpoint包括经典网络、公网和VPC三种网络类型，请根据数据集成资源组和Hologres实例所在的网络环境选择正确的endpoint类型，否则会出现网络不通或者性能受限的情况： • 经典网络示例： `instance-id-region-endpoint-internal.hologres.aliyuncs.com:port` • 公网示例： `instance-id-region-endpoint.hologres.aliyuncs.com:port` • VPC示例： `instance-id-region-endpoint-vpc.hologres.aliyuncs.com:port` 通常建议数据集成资源组和Hologres实例配在同一个地域的同一个可用区，以保证网络端口连通，实现最大性能。	是	无
accessId	访问Hologres的accessId。	是	无
accessKey	访问Hologres的accessKey，请确保该密钥对目标表有写入权限。	是	无
database	Hologres实例内部数据库的名称。	是	无
table	Hologres的表名称，如果是分区表，请指定父表的名称。	是	无
column	定义导入目标表的数据列，必须包含目标表的主键集合。例如 `["*"]` 表示全部列。	是	无
partition	针对分区表，表示分区Column以及对应的Value，格式为 `column=value` 。 **注意** • 目前Hologres仅支持LIST分区，分区Column仅支持单个Column分区，且仅支持INT4或TEXT类型。 • 请确认该参数和表DDL的分区配置匹配。 • 请确认对应的子表已经创建，且已经导入数据。	否	*空，表示非分区表。*
fetchSize	指定使用Select语句一次性读取数据的条数。	否	*1,000*

向导开发介绍

1. 选择数据源。

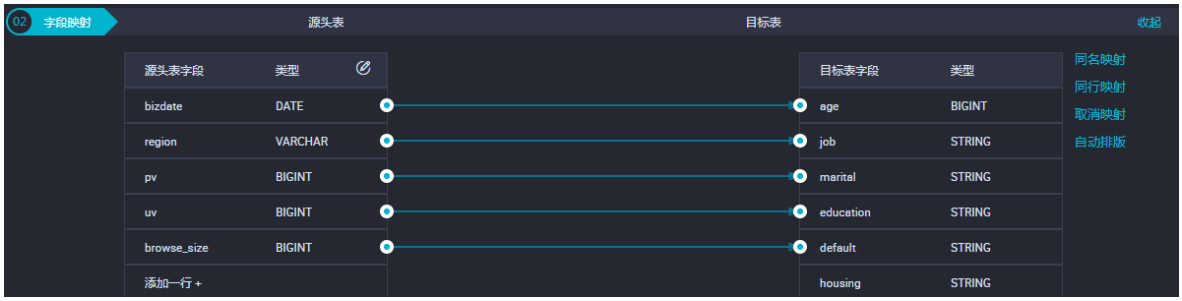
配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	通常输入您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，SQL语法与选择的数据源一致，请勿填写where关键字。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	◦ 可以输入常量，输入的值需要使用英文单引号，例如'abc'、'123'等。 ◦ 可以配合调度参数使用，例如${bizdate}等。 ◦ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ◦ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

- 配置非分区表

- 配置从Hologres非分区表读取数据至内存，如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"holo",//插件名。
            "parameter":{
                "endpoint": "instance-id-region-endpoint.hologres.aliyuncs.com:port",
                "accessId": "***************", //访问Hologres的accessId。
                "accessKey": "******************", //访问Hologres的accessKey。
                "database": "postgres",
                "table": "holo_reader_****",
                "column" : [ //字段。
                    "tag",
                    "id",
                    "title"
                ]
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1,//作业并发数。
                  "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

- Hologres表的DDL语句，如下所示。

```
begin;
drop table if exists holo_reader_basic_src;
create table holo_reader_basic_src(
  tag text not null,
  id int not null,
  title text not null,
  body text,
  primary key (tag, id));
  call set_table_property('holo_reader_basic_src', 'orientation', 'column');
  call set_table_property('holo_reader_basic_src', 'shard_count', '3');
commit;
```

- 配置分区表
  - 配置从内存产生的数据同步至Hologres分区表的子表，示例为通过SDK模式导入的配置。

> **说明** 请注意partition的配置。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"holo",//插件名。
            "parameter":{
                "endpoint": "instance-id-region-endpoint.hologres.aliyuncs.com:port",
                "accessId": "***************", //访问Hologres的accessId。
                "accessKey": "******************", //访问Hologres的accessKey。
                "database": "postgres",
                "table": "holo_reader_basic_****",
                "partition": "tag=foo",
                "column" : [
                    "*"
                ],
                "fetchSize": "100"
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1,//作业并发数。
                  "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

- Hologres表的DDL语句，如下所示。

```
begin;
drop table if exists holo_reader_basic_part_src;
create table holo_reader_basic_part_src(
  tag text not null,
  id int not null,
  title text not null,
  body text,
  primary key (tag, id))
  partition by list( tag );
  call set_table_property('holo_reader_basic_part_src', 'orientation', 'column');
  call set_table_property('holo_reader_basic_part_src', 'shard_count', '3');
commit;
create table holo_reader_basic_part_src_1583161774228 partition of holo_reader_basic_part_src for values
in ('foo');
# 确保分区表子表已经创建且导入数据。
postgres=# \d+ holo_reader_basic_part_src
                        Table "public.holo_reader_basic_part_src"
 Column |  Type   | Collation | Nullable | Default | Storage  | Stats target | Description
--------+---------+-----------+----------+---------+----------+--------------+-------------
 tag    | text    |           | not null |         | extended |              |
 id     | integer |           | not null |         | plain    |              |
 title  | text    |           | not null |         | extended |              |
 body   | text    |           |          |         | extended |              |
Partition key: LIST (tag)
Indexes:
    "holo_reader_basic_part_src_pkey" PRIMARY KEY, btree (tag, id)
Partitions: holo_reader_basic_part_src_1583161774228 FOR VALUES IN ('foo')
```

6.2.36. GDB Reader

本文为您介绍GDB（Graph Database） Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

图数据库GDB（Graph Database）支持Property Graph图模型，是一种用于处理高度连接数据查询与存储的实时、可靠的在线数据库服务。GDB支持Apache TinkerPop Gremlin查询语言，可以帮您快速构建基于高度连接的数据集的应用程序。

> 说明
> - 开始配置GDB Reader插件前，请首先配置好数据源，详情请参见。
> - 由于点和边的数据集成任务的配置不同，请您分别进行配置。

使用限制

- 导出GDB需要配置点任务和边任务，分别导出点数据和边数据。
- 导出任务根据点或边的类型名称遍历数据，您需要确保待导出数据的类型名称。
- GDB中点和边的主键ID字段是字符串类型，请配置导出为STRING类型。如果配置LONG等数值类型，GDB Reader会转换为配置的目标类型，但转换失败时会丢失该记录。
- 您需要配置GDB导出的属性值为存储类型。如果存储类型和配置类型不一致，GDB Reader会转换为目标类型，但可能会转换失败导致该记录丢失。
- 导出点SET属性的一个值时，不能保证多次导出都是同一个值。
- 导出所有属性使用JSON格式输出时，仅包含一个属性值的SET属性会被作为普通属性输出。
- 如果未对示例的字段名或枚举值进行特别说明，请注意大小写敏感。
- GDB服务端仅支持UTF-8编码格式，导出的数据均为UTF-8编码格式。
- GDB需要升级至1.0.20或更高版本才支持SET属性。使用SET属性时，请确认实例的版本。

参数说明

参数	描述	是否必选	默认值
host	GDB实例的连接域名。您可以在图数据库GDB控制台，单击相应实例后的**管理**，查看**内网地址**（即host）。	是	无
port	GDB实例的连接端口。	是	*8182*
username	GDB实例的账号名。	是	无
password	GDB实例账号的密码。	是	无
labels	类型名，即点或边的名称。支持读取多个名称的数据，使用数组表示，例如["label1", "label2"]。	是	无
labelType	数据的Label类型： • 枚举值VERTEX表示点。 • 枚举值EDGE表示边。	是	无
column	点或边的字段映射关系配置。	是	无
column -> name	点或边的映射关系的字段名。读取属性时必选，请提供属性名。	是	无
column -> type	点或边的映射关系的字段值类型： • 主键ID、类型名Label在GDB中均为STRING类型。如果您配置为STRING类型，会转换失败。 • 普通属性支持INT、LONG、FLOAT、DOUBLE、BOOLEAN和STRING等类型。 • GDB Reader会尽量转换读取的数据为配置要求的类型，但转换失败会导致该条记录错误。	是	无

参数	描述	是否必选	默认值
column -> columnType	GDB点或边数据对应到的点或边的映射关系字段，包括以下枚举值： • 公共枚举值： ○ primaryKey：表示该字段为主键ID。 ○ primaryLabel：表示该字段为名称Label。 • 点枚举值： ○ ：vertexProperty：当labelType为点时，表示该字段为点的属性。 ○ vertexJsonProperty：当labelType为点时，表示该字段为点的属性的集合，使用JSON格式封装。当配置该类型时，所有属性会打包至该列，column中不能再包含其它属性的类型。 vertexJsonProperty格式如下。 `{ "properties":[ {"k":"name","t":"string","v":"tom","c":"set"}, {"k":"name","t":"string","v":"jack","c":"set"}, {"k":"sex","t":"string","v":"male","c":"single"} ] }` 上述导出的属性包含多值属性name，有两个属性值和一个单值属性。如果GDB中的多值属性仅包含一个属性值，导出会被当作单值属性。 • 当labelType为边时的边枚举值： ○ srcPrimaryKey：表示该字段为起点主键ID。 ○ dstPrimaryKey：表示该字段为终点主键ID。 ○ srcPrimaryLabel：表示该字段为起点名称Label。 ○ dstPrimaryLabel：表示该字段为终点名称Label。 ○ edgeProperty：表示该字段为边的属性。 ○ edgeJsonProperty：表示该字段为边的属性集合，使用JSON格式封装。配置该类型时，所有属性均会打包至该列，column中不能再包含其它属性的类型。 edgeJsonProperty格式如下。 `{ "properties":[ {"k":"name","t":"string","v":"tom"}, {"k":"sex","t":"string","v":"male"} ] }` 边不支持多值属性，无c字段。	是	无

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

配置写入GDB的数据同步作业时，请分别配置点和边，详情请参见参数说明：

- 点配置示例

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"100"  //错误记录数,表示脏数据的最大容忍条数。
        },
        "jvmOption":"",
        "speed":{
            "concurrent":3,
            "throttle":true,///当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "host": "gdb-xxxxxx.aliyuncs.com", //GDB实例的连接地址。
                "port": 8182, //GDB实例的连接端口。
                "username": "gdb", //GDB实例的用户名。
                "password": "gdb", //GDB实例用户名对应的密码。
                "labelType": "VERTEX", // Label类型，使用VERTEX表示点。
                "labels": ["label1", "label2"],  // Label名的列表，为空表示导出所有的点。
                "column": [
                    {
                        "name": "id",               // 字段名。
                        "type": "string",           // 字段类型。
                        "columnType": "primaryKey"  // 字段分类，表示点的主键ID，GDB中是STRING类型。
                    },
                    {
                        "name": "label",              // 字段名。
                        "type": "string",             // 字段类型。
                        "columnType": "primaryLabel"  // 字段分类，表示点的Label名，GDB中是STRING类型。
                    },
                    {
                        "name": "age",                   // 属性字段名。
                        "type": "int",                   // 属性字段类型。
                        "columnType": "vertexProperty"   // 字段分类，表示点的属性，GDB中基础类型属性。
                    }
                ]
            },
            "stepType":"gdb"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter":{
                "print": true
            },
            "stepType":"stream"
        }
    ]
}
```

- 边配置示例

```
{
```

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"100"  //错误记录数,表示脏数据的最大容忍条数。
        },
        "jvmOption":"",
        "speed":{
            "concurrent":3,
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps":"12"//限流
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "host": "gdb-xxxxxx.aliyuncs.com", //GDB实例的连接地址。
                "port": 8182, //GDB实例的连接端口。
                "username": "gdb", //GDB实例用户名。
                "password": "gdb", //GDB实例用户名对应的密码。
                "labelType": "EDGE", // Label类型，VERTEX表示点。
                "labels": ["label1", "label2"],  // Label名列表，为空表示导出所有的边。
                "column": [
                    {
                        "name": "id",               // 字段名。
                        "type": "string",           // 字段类型。
                        "columnType": "primaryKey"  // 字段分类，表示边的主键ID，GDB中是STRING类型。
                    },
                    {
                        "name": "label",              // 字段名。
                        "type": "string",             // 字段类型。
                        "columnType": "primaryLabel"  // 字段分类，表示边的Label名，GDB中是STRING类型。
                    },
                    {
                        "name": "srcId",               // 字段名。
                        "type": "string",              // 字段类型。
                        "columnType": "srcPrimaryKey"  // 字段分类，表示边关联点中起点的ID，GDB中是STRING类型。
                    },
                    {
                        "name": "srcLabel",               // 字段名。
                        "type": "string",                 // 字段类型。
                        "columnType": "srcPrimaryLabel"   // 字段分类，表示边关联点中起点的Label名，GDB中是STRIN
G类型。
                    },
                    {
                        "name": "dstId",                    // 字段名。
                        "type": "string",                   // 字段类型。
                        "columnType": "dstPrimaryKey"       // 字段分类，表示边关联点中终点的ID，GDB中是STRING类
型。
                    },
                    {
                        "name": "dstLabel",                 // 字段名。
                        "type": "string",                   // 字段类型。
                        "columnType": "dstPrimaryLabel"     // 字段分类，表示边关联点中终点的Label名，GDB中是STR
ING类型。
                    },
```

```
                    {
                        "name": "weight",               // 属性字段名。
                        "type": "double",               // 属性字段类型。
                        "columnType": "edgeProperty"    // 字段分类，表示边的属性。
                    }
                ]
            },
            "stepType":"gdb"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter":{
                "print": true
            },
            "stepType":"stream"
        }
    ]
}
```

6.2.37. RestAPI Reader

本文为您介绍RestAPI Reader支持的数据类型、字段映射和数据源等参数及配置示例，在创建数据集成任务节点前，您可参考本文了解数据集成过程中对数据源进行数据抽取所需的参数及支持的数据类型。

背景信息

RestAPI Reader插件提供了读取RESTful接口数据的能力。RestAPI Reader从RESTful地址中获取数据，转换为数据集成支持的数据类型，然后传递给下游的Writer。RestAPI Reader支持JSON格式的返回结果，并可以从中读取INT、BOOLEAN、DATE、DOUBLE、FLOAT、LONG、STRING数据类型。

数据类型转换列表

类型分类	数据集成column配置类型
整数类	LONG，INT
字符串类	STRING
浮点类	DOUBLE，FLOAT
布尔类	BOOLEAN
日期时间类	DATE

参数说明

进行数据集成时，您需要添加数据源后再配置数据源的来源与去向，并在配置过程中设置好集成的数据及数据类型等信息，整个数据集成包含数据提取（使用reader插件提取数据来源的数据）和数据写入（使用writer插件将集成的数据写入数据去向的数据源中）。

以下为您介绍使用reader插件提取RestAPI类型数据源的数据时，需要配置的参数。

> **说明** 以下的参数包含在添加数据源和配置数据集成任务节点的过程中。
>
> 当前插件暂不支持使用调度参数。

参数	描述	是否必选	默认值
url	RESTful接口地址。	是	无

参数	描述	是否必选	默认值
dataMode	RESTful请求返回的结果JSON数据的格式。 • oneData：从返回的JSON中取其1条数据。 • multiData：从返回的JSON中取一个JSON数组，传递多条数据给writer。	是	无
responseType	返回结果的数据格式，目前仅支持JSON格式。	是	JSON
column	读取字段列表，type指定源数据的类型，name指定当前column数据获取的JSON路径。您可以指定column字段信息，配置如下。 "column":[{"type":"long","name":"a.b" //从a.b路径中查找数据},{"type":"string","name":"a.c"//从a.c路径中查找数据}] 对于您指定的column信息，type和name必须填写。	是	无
dataPath	从返回结果中查询单个JSON对象或者JSON数组的路径。	否	无
method	请求方法，支持get或post两种方式。	是	无
customHeader	传递给RESTful接口的header信息。	否	无
parameters	传递给RESTful接口的参数信息。 • get方法填入 `abc=1&def=1` 。 • post方法填入JSON类型参数。	否	无
dirtyData	当从指定的column json路径中找不到数据时的处理方式。 • dirty：当一条数据解析时遇到column找不时这条数据置为脏数据。 • null：当一条数据解析时遇到column找不到时，这个column设置为null。	是	dirty
requestTimes	从RESTful地址中请求数据的次数。 • single：只进行一次请求。 • multiple：进行多次请求。	是	single
requestParam	若requestTimes设为multiple时，需要指定循环的参数，例如pageNumber，插件会根据设置的startIndex、endIndex、step三个参数循环传递pageNumber参数给RESTful接口，进行多次请求。	否	无
startIndex	循环请求的起点，起点包含在循环请求之内。	否	无
endIndex	循环请求的终点，终点包含在循环请求之内。	否	无
step	循环请求的步长。	否	无
authType	验证方法。包括： • Basic Auth：基础验证。 如果数据源API支持用户名和密码的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的用户名和密码，后续数据集成过程中对接数据源时，通过Basic Auth协议传递给RESTful地址，完成验证。 • Token Auth：Token验证。 如果数据源API支持Token的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的固定Token值，后续数据集成过程中对接数据源时，通过传入header中进行验证，例如：{“Authorization”:” Bearer TokenXXXXXX”}。 • Aliyun API Signature：阿里云API签名验证。 如果数据源为阿里云产品，且此阿里云产品的API支持通过AccessKey和AccessSecret的方式进行验证，您可选择此种种验证方式，并在选择完成后配置用于验证的AccessKey和AccessSecret。	否	无

参数	描述	是否必选	默认值
authUsername/authPassword	Basic Auth验证的用户名和密码。	否	无
authToken	Token Auth验证的token。	否	无
accessKey/accessSecret	Aliyun API签名验证的账户信息。	否	无

配置示例：向导模式

1. 选择数据源。

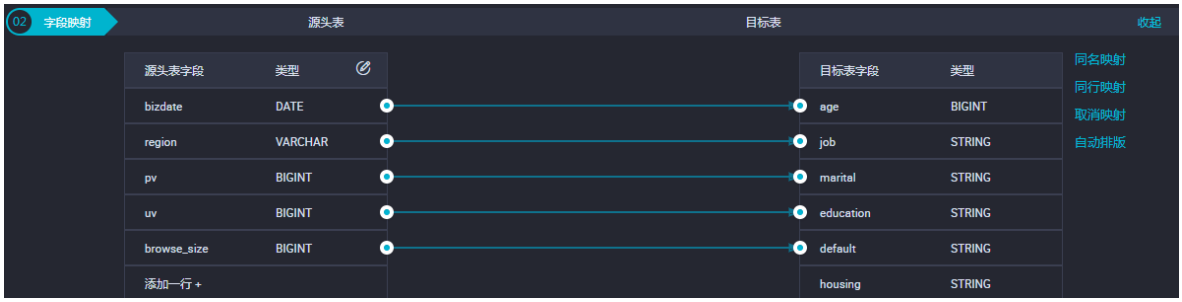
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	数据源类型选择**RestAPI**后，选择需要读取的数据表。
请求Method	即上述参数说明中的method。
返回数据结构	即上述参数说明中的dataMode。
数据存储JSON路径	即上述参数说明中的dataPath。
数据格式	即上述参数说明中的responseType，目前仅支持JSON格式。
脏数据处理	即上述参数说明中的dirtyData，默认值为dirty。
header	即上述参数中的customHeader。

参数	描述
请求参数	即上述参数中的parameters。
请求次数	即上述参数中的requestTimes。
StartIndex/Step/EndIndex	与上述参数中同名参数相同。

2. 字段映射。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

脚本配置示例如下所示。

```
{
    "type":"job",
    "version":"2.0",
    "steps":[
        {
            "stepType":"restapi",
            "parameter":{
                "url":"http://127.0.0.1:5000/get_array5",
                "dataMode":"oneData",
                "responseType":"json",
                "column":[
                    {
                        "type":"long",
                        "name":"a.b"  //从a.b路径中查找数据
                    },
                    {
                        "type":"string",  //从a.c路径中查找数据
                        "name":"a.c"
                    }
                ],
                "dirtyData":"null",
                "method":"get",
                "defaultHeader":{
                    "X-Custom-Header":"test header"
                },
                "customHeader":{
                    "X-Custom-Header2":"test header2"
                },
                "parameters":"abc=1&amp;def=1"
            },
            "name":"restapireader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":""
        },
        "speed":{
            "throttle":true,  //当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1,  //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

脚本模式配置说明如下：

Restapi插件发出http(s)请求后，会获得请求响应body（body是一个json），dataPath用来配置从body中提取数据的json path路径。
举2个示例如下：
以接口返回数据body如下举例，其中业务数据在DATA内，且接口一次返回了多行数据（DATA是一个数组）：

```
{
    "HEADER": {
        "BUSID": "bid1",
        "RECID": "uuid",
        "SENDER": "dc",
        "RECEIVER": "pre",
        "DTSEND": "202201250000"
    },
    "DATA": [
        {
            "SERNR": "sernr1"
        },
        {
            "SERNR": "sernr2"
        }
    ]
}
```

如果需要将DATA中的多行数据抽取为多条同步记录，则需要将 column 配置为 "column": ["SERNR"]，dataMode 配置为 "dataMode": "multiData"，dataPath 配置为 "dataPath": "DATA"
以接口返回数据body如下举例，其中业务数据在content.DATA内，且接口一次返回了1行数据（DATA是一个对象）：

```
{
    "HEADER": {
        "BUSID": "bid1",
        "RECID": "uuid",
        "SENDER": "dc",
        "RECEIVER": "pre",
        "DTSEND": "202201250000"
    },
    "content": {
        "DATA": {
            "SERNR": "sernr2"
        }
    }
}
```

如果需要将content.DATA中的一行数据抽取为一条同步记录，则需要将 column 配置为 "column": ["SERNR"]，dataMode 配置为 "dataMode": "oneData"，dataPath 配置为 "dataPath": "content.DATA"

6.2.38. SAP HANA Reader

本文为您介绍SAP HANA Reader支持的数据类型、字段映射和数据源等参数及配置示例，在创建数据集成任务节点前，您可参考本文了解数据集成过程中对数据源进行数据抽取所需的参数及支持的数据类型。

背景信息

SAP HANA Reader插件通过JDBC连接器连接至远程的SAP HANA数据库，根据您配置的信息生成查询SQL语句，发送至远程SAP HANA数据库，执行该SQL语句并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

类型转换列表

SAP HANA Reader针对SAP HANA类型的转换列表，如下所示。

类型分类	数据源的数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL

类型分类	数据源的数据类型
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔型	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

注意
- 除上述罗列字段类型外，其它类型均不支持。
- SAP HANA Reader插件将tinyint（1）视作整型。

参数说明

参数	描述
username	用户名。
password	密码。
column	需要同步的字段名称。如果需要同步所有列，请使用星号（*）。
table	需要同步的表名。
jdbcUrl	连接HANA的JDBC URL。例如，jdbc:sap://127.0.0.1:30215?currentschema=TEST。
splitPk	HANA表中的某个字段作为同步的切分字段，切分字段有助于多并发同步HANA表。 切分字段需要是数值整型的字段，如果没有该类型，则可以不填。

配置示例：向导模式

1. 选择数据源。

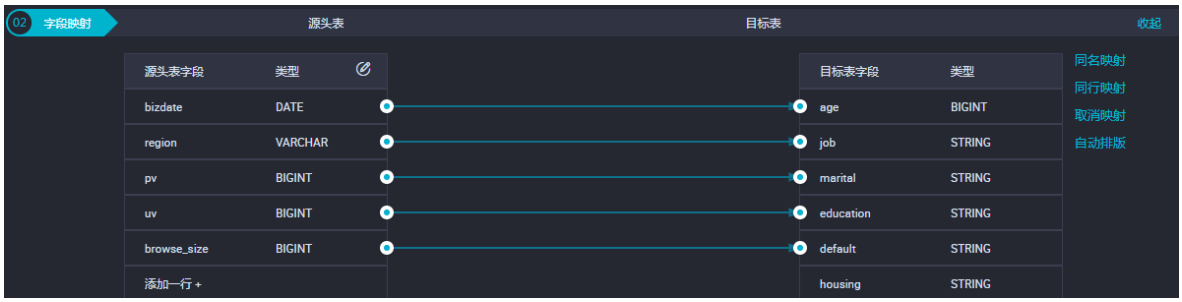
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。

参数	描述
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

本文为您提供单库单表和分库分表的配置示例：

- 配置单库单表

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"saphana",//插件名。
            "parameter":{
                "column":[//列名。
                    "id"
                ],
                "connection":[
                    {   "querySql":["select a,b from join1 c join join2 d on c.id = d.id;"], //使用字符串的形式，将querySql写在connection中。
                        "datasource":"",//数据源。
                        "table":[//表名，即使只有一张表，也必须以[]的数组形式书写。
                            "xxx"
                        ]
                    }
                ],
                "where":"",//过滤条件。
                "splitPk":"",//切分键。
                "encoding":"UTF-8"//编码格式。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为flase时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

- 配置分库分表

  > **说明** 分库分表是指在SAP HANA Reader端可以选择多个SAP HANA数据表，且表结构保持一致。

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {
            "plugin": "saphana",
            "parameter": {
                "connection": [
                    {
                        "table": [
                            "tbl1",
                            "tbl2",
                            "tbl3"
                        ],
                        "datasource": "datasourceName1"
                    },
                    {
                        "table": [
                            "tbl4",
                            "tbl5",
                            "tbl6"
                        ],
                        "datasource": "datasourceName2"
                    }
                ],
                "singleOrMulti": "multi",
                "splitPk": "db_id",
                "column": [
                    "id", "name", "age"
                ],
                "where": "1 < id and id < 100"
            }
        },
        "writer": {
        }
    }
}
```

6.2.39. KingbaseES Reader

本文为您介绍KingbaseES Reader支持的数据类型、字段映射和数据源等参数及配置示例，在创建数据集成任务节点前，您可参考本文了解数据集成过程中对数据源进行数据抽取所需的参数及支持的数据类型。

背景信息

KingbaseES Reader插件通过JDBC连接器连接至远程的KingbaseES数据库，根据您配置的信息生成查询SQL语句，发送至远程KingbaseES数据库，执行该SQL语句并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理。

类型转换列表

KingbaseES Reader针对KingbaseES类型的转换列表，如下所示。

类型分类	数据源的数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL

类型分类	数据源的数据类型
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP、TIME和YEAR
布尔型	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

注意
- 除上述罗列字段类型外，其它类型均不支持。
- KingbaseES Reader插件将tinyint（1）视作整型。

参数说明

参数	描述
username	用户名。
password	密码。
column	需要同步的字段名称。如果需要同步所有列，请使用星号（*）。
table	需要同步的表名。
jdbcUrl	连接KingbaseES的JDBC URL。例如，jdbc:sap://127.0.0.1:30215?currentschema=TEST。
splitPk	KingbaseES表中的某个字段作为同步的切分字段，切分字段有助于多并发同步KingbaseES表。 切分字段需要是数值整型的字段，如果没有该类型，则可以不填。

配置示例：向导模式

1. 选择数据源。

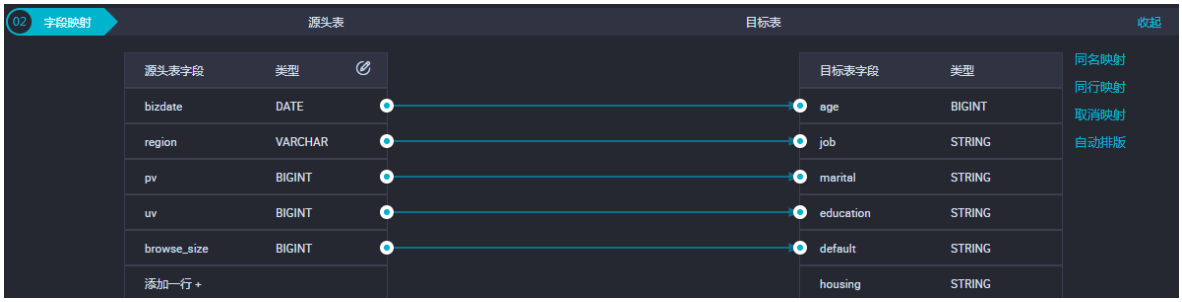
   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
数据过滤	您将要同步数据的筛选条件，暂时不支持limit关键字过滤。SQL语法与选择的数据源一致。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

参数	描述
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

本文为您提供单库单表和分库分表的配置示例：

- 配置单库单表

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"kingbasees",//插件名。
            "parameter":{
                "column":[//列名。
                    "id"
                ],
                "connection":[
                    {   "querySql":["select a,b from join1 c join join2 d on c.id = d.id;"], //使用字符串的形式，将querySql写在connection中。
                        "datasource":"",//数据源。
                        "table":[//表名，即使只有一张表，也必须以[]的数组形式书写。
                            "xxx"
                        ]
                    }
                ],
                "where":"",//过滤条件。
                "splitPk":"",//切分键。
                "encoding":"UTF-8"//编码格式。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

- 配置分库分表

> **说明** 分库分表是指在KingbaseES Reader端可以选择多个KingbaseES数据表，且表结构保持一致。

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {
            "plugin": "kingbasees",
            "parameter": {
                "connection": [
                    {
                        "table": [
                            "tbl1",
                            "tbl2",
                            "tbl3"
                        ],
                        "datasource": "datasourceName1"
                    },
                    {
                        "table": [
                            "tbl4",
                            "tbl5",
                            "tbl6"
                        ],
                        "datasource": "datasourceName2"
                    }
                ],
                "singleOrMulti": "multi",
                "splitPk": "db_id",
                "column": [
                    "id", "name", "age"
                ],
                "where": "1 < id and id < 100"
            }
        },
        "writer": {
        }
    }
}
```

6.2.40. DM Reader

本文为您介绍DM Reader支持的数据类型、字段映射和数据源等参数及配置示例。

> **注意** DM Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

背景信息

DM Reader插件从DM读取数据。在底层实现上，DM Reader通过JDBC连接远程DM数据库，并执行相应的SQL语句，从DM库中读取数据。目前DM Reader支持读取DM、DB2、PPAS和Sybase等数据库的数据。DM Reader是一个通用的关系数据库读插件，您可以通过注册数据库驱动等方式，增加任意多样的关系数据库读支持。

DM Reader通过JDBC连接器连接至远程的DM数据库，并根据您配置的信息生成查询SQL语句，发送至远程DM数据库，执行该SQL并返回结果。然后使用数据同步自定义的数据类型拼装为抽象的数据集，传递给下游Writer处理：

- 对于您配置的table、column和where等信息，DM Reader将其拼接为SQL语句发送至DM数据库。
- 对于您配置的querySql信息，DM直接将其发送至DM数据库。

DM Reader支持大部分通用的关系数据库数据类型，例如数字、字符等。但也存在部分类型没有支持的情况，请注意检查您的数据类型，根据具体的数据库进行选择。

参数说明

参数	描述	是否必选	默认值
jdbcUrl	描述的是到对端数据库的JDBC连接信息，jdbcUrl按照DM官方规范，并可以填写连接附件控制信息。请注意不同的数据库JDBC的格式不同，数据集成会根据具体JDBC的格式选择合适的数据库驱动读取数据： • DM格式：`jdbc:dm://ip:port/database` • DB2格式：`jdbc:db2://ip:port/database` • PPAS格式：`jdbc:edb://ip:port/database` • 进入DM Reader对应目录，${DATAX_HOME}为数据集成主目录。 • 在DM Reader插件目录下有*plugin.json*配置文件，在此文件中注册您具体的数据库驱动，放在drivers数组中。DM Reader插件在任务执行时会动态选择合适的数据库驱动连接数据库。 ``` { "name": "rdbmsreader", "class": "com.alibaba.datax.plugin.reader.rdbmsreader.RdbmsReader", "description": "useScene: prod. mechanism: Jdbc connection using the database, execute select sql, retrieve data from the ResultSet. warn: The more you know about the database, the less problems you encounter.", "developer": "alibaba", "drivers": ["dm.jdbc.driver.DmDriver", "com.ibm.db2.jcc.DB2Driver", "com.sybase.jdbc3.jdbc.SybDriver", "com.edb.Driver"] } ``` - 在DM Reader插件目录下有libs子目录，您需要将您具体的数据库驱动放到libs目录下。 ``` $tree . |-- libs | |-- Dm7JdbcDriver16.jar | |-- commons-collections-3.0.jar | |-- commons-io-2.4.jar | |-- commons-lang3-3.3.2.jar | |-- commons-math3-3.1.1.jar | |-- datax-common-0.0.1-SNAPSHOT.jar | |-- datax-service-face-1.0.23-20160120.024328-1.jar | |-- db2jcc4.jar | |-- druid-1.0.15.jar | |-- edb-jdbc16.jar | |-- fastjson-1.1.46.sec01.jar | |-- guava-r05.jar | |-- hamcrest-core-1.3.jar | |-- jconn3-1.0.0-SNAPSHOT.jar | |-- logback-classic-1.0.13.jar | |-- logback-core-1.0.13.jar | |-- plugin-rdbms-util-0.0.1-SNAPSHOT.jar | `-- slf4j-api-1.7.10.jar |-- plugin.json |-- plugin_job_template.json `-- rdbmsreader-0.0.1-SNAPSHOT.jar	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无

参数	描述	是否必选	默认值
table	所选取的需要同步的表。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息，默认使用所有列配置，例如[*]: • 支持列裁剪，即列可以挑选部分列进行导出。 • 支持列换序，即列可以不按照表schema信息顺序进行导出。 • 支持常量配置，您需要按照JSON格式 `["id","1", "'bazhen.csy'", "null", "to_char(a + 1)", "2.3" , "true"]` 。 ◦ id为普通列名。 ◦ 1为整型数字常量。 ◦ 'bazhen.csy'为字符串常量。 ◦ null为空指针。 ◦ to_char(a + 1)为函数表达式。 ◦ 2.3为浮点数。 ◦ true为布尔值。 • column必须显示您指定同步的列集合，不允许为空 。	是	无
splitPk	DM Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片。数据同步系统会启动并发任务进行数据同步，以提高数据同步的效能: • 推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，切分出来的分片也不容易出现数据热点。 • 目前splitPk仅支持整型数据切分，不支持浮点、字符串和日期等其他类型。如果您指定其他非支持类型，DM Reader将报错。 • 如果不填写splitPk，将视作您不对单表进行切分，DM Reader使用单通道同步全量数据。	否	空
where	筛选条件，DM Reader根据指定的column、table和where条件拼接SQL，并根据该SQL进行数据抽取。例如在做测试时，可以将where条件指定为limit 10。 在实际业务场景中，通常会选择当天的数据进行同步，可以将where条件指定为 `gmt_create>$bizdate` : • where条件可以有效地进行业务增量同步。 • where条件不配置或为空时，则视作全表同步数据。	否	无
querySql	在部分业务场景中，where配置项不足以描述所筛选的条件，您可以通过该配置型来自定义筛选SQL。当您配置该项后，数据同步系统会忽略column、table等配置，直接使用该配置项的内容对数据进行筛选。 例如，需要进行多表join后同步数据，使用 `select a,b from table_a join table_b on table_a.id = table_b.id` 。当您配置querySql时，DM Reader直接忽略column、table和where条件的配置。	否	无
fetchSize	该配置项定义了插件和数据库服务器端每次批量数据获取条数，该值决定了数据同步系统和服务器端的网络交互次数，能够提升数据抽取性能。 **说明** fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1,024*

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

配置一个从DM数据库同步抽取数据作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
                  "concurrent":1 //作业并发数。
                  "mbps":"12",//限流
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "connection": [
                    {
                        "jdbcUrl": [
                            "jdbc:dm://ip:port/database"
                        ],
                        "table": [
                            "table"
                        ]
                    }
                ],
                "username": "username",
                "password": "password",
                "table": "table",
                "column": [
                    "*"
                ],
                "preSql": [
                    "delete from XXX;"
                ]
            },
            "stepType": "rdbms"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {},
            "stepType": "stream"
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

6.2.41. AWS S3 Reader

AWS S3 Reader插件用于从S3数据库读取数据。本文为您介绍AWS S3 Reader支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

AWS S3 Reader插件用于从S3读取数据。在底层实现上，AWS S3 Reader使用Amazon官网提供的Java SDK获取S3数据，并转换为数据同步传输协议传递给Writer。

AWS S3是非结构化数据存储。对于数据集成而言，AWS S3 Reader支持的功能如下：

- 支持且仅支持读取TXT格式的文件，且要求TXT中schema为一张二维表。
- 支持类CSV格式文件，自定义分隔符。
- 支持多种类型数据读取，支持列裁剪和列常量。
- 支持递归读取、文件名过滤。
- 支持文本压缩，现有压缩格式为gzip、bzip2和zip。
- 支持多个Object并发读取。

使用限制

- 暂不支持大陆及中国香港地区的S3数据源。
- AWS S3 Reader暂时不支持以下功能：
  - 单个Object（File）不支持多线程并发读取。
  - 单个Object在压缩状态下，不支持多线程并发读取。
  - 单个Object（File）不能超过100 GB。

支持的数据类型

类型分类	数据集成column配置类型	S3数据类型
整数类	LONG	LONG
浮点类	DOUBLE	DOUBLE
字符串类	STRING	STRING
日期时间类	DATE	DATE
布尔型	BOOL	BOOL

脚本开发介绍

- 参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项输入的内容必须和添加的数据源名称保持一致。	是	无

参数	描述	是否必选	默认值
Object	S3的Object信息，支持填写多个Object。例如，bucket中有*test*文件夹，文件夹中有*ll.txt*文件，则Object填写*test/ll.txt*。 ○ 当指定单个S3 Object时，AWS S3 Reader目前只支持单线程进行数据抽取。 ○ 当指定多个S3 Object时，AWS S3 Reader支持使用多线程进行数据抽取。线程并发数通过通道数指定。 ○ 当指定通配符时，AWS S3 Reader尝试遍历出多个Object信息。例如，配置abc[0-9]表示abc0、abc1、abc2、abc3等。配置通配符会导致内存溢出，通常不建议您配置通配符。 **说明** ○ 数据同步系统会将一个作业下同步的所有Object视作同一张数据表。您必须保证所有的Object能够适配同一套Schema信息。 ○ 请注意控制单个目录下的文件个数，否则可能会触发系统OutOfMemoryError报错。如果遇到此情况，请将文件拆分到不同目录后再尝试进行同步。	是	无
column	读取字段列表，type指定源数据的类型，index指定当前列来自于文本第几列（以0开始），value指定当前类型为常量，不是从源头文件读取数据，而是根据value值自动生成对应的列。 默认情况下，您可以全部按照String类型读取数据，配置如下。 `column": ["*"]` 您可以指定column字段信息，配置如下。 `"column": { "type": "long", "index": 0 //从S3文本第一列获取int字段。 }, { "type": "string", "value": "alibaba" //从S3 Reader内部生成alibaba的字符串字段作为当前字段。 }` **说明** 对于您指定的column信息，type必须填写，index和value必须选择其一。	是	全部按照STRING类型读取。
fieldDelimiter	读取的字段分隔符。 **说明** AWS S3 Reader在读取数据时，需要指定字段分割符，如果不指定默认为（,），界面配置中也会默认填写为（,）。 如果分隔符不可见，请填写Unicode编码。例如，\u001b、\u007c。	是	默认值：（，）
compress	文本压缩类型，默认不填写（即不压缩）。支持压缩类型为gzip、bzip2和zip。	否	不压缩
encoding	读取文件的编码配置。	否	utf-8
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据同步系统提供nullFormat定义哪些字符串可以表示为null。例如，您配置`nullFormat="null"`，那么如果源头数据是`null"`，数据同步系统会视作null字段。针对空字符串，需要加一层转义：`\N=\\N`。	否	无

参数	描述	是否必选	默认值
skipHeader	CSV格式文件通过**skipHeader**配置是否读取表头内容。 ○ True：同步数据源的时候读取表头内容。 ○ False：同步数据源的时候不读取表头内容。 **说明** 压缩文件模式下不支持**skipHeader**。	否	false
csvReaderConfig	读取CSV类型文件参数配置，Map类型。读取CSV类型文件使用的CsvReader进行读取，会有很多配置，不配置则使用默认值。	否	无

- 配置一个从AWS S3数据库同步抽取数据的作业。使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。示例脚本如下。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"s3",//插件名。
            "parameter":{
                "nullFormat":"",//定义可以表示为null的字符串。
                "compress":"",//文本压缩类型。
                "datasource":"",//数据源。
                "column":[//字段。
                    {
                        "index":0,//列序号。
                        "type":"string"//数据类型。
                    },
                    {
                        "index":1,
                        "type":"long"
                    },
                    {
                        "index":2,
                        "type":"double"
                    },
                    {
                        "index":3,
                        "type":"boolean"
                    },
                    {
                        "format":"yyyy-MM-dd HH:mm:ss", //时间格式。
                        "index":4,
                        "type":"date"
                    }
                ],
                "skipHeader":"",//类CSV格式文件可能存在表头为标题情况，需要跳过。
                "encoding":"",//编码格式。
                "fieldDelimiter":",",//字段分隔符。
                "fileFormat": "",//文本类型。
                "object":[]//object前缀。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
```

```
    ],
    "setting":{
        "errorLimit":{
            "record":""//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1 //作业并发数。
            "mbps":"12",//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述脚本模式参数说明中的datasource，通常输入您配置的数据源名称。
文件名（含路径）	即上述脚本模式参数说明中的object。 **说明** 假如您的S3文件名中有根据每天的时间命名的部分，例如，*aaa/20171024abc.txt*，关于Object系统参数可以设置为*aaa/${bdp.system.bizdate}abc.txt*。
列分隔符	即上述脚本模式参数说明中的fieldDelimiter，默认值为（,）。
编码格式	即上述脚本模式参数说明中的encoding，默认值为utf-8。
null值	即上述脚本模式参数说明中的nullFormat，将要表示为空的字段填入文本框，如果源端存在则将对应的部分转换为空。
压缩格式	即上述脚本模式参数说明中的compress，默认值为不压缩。
是否包含表头	即上述脚本模式参数说明中的skipHeader，默认值为否。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

6.2.42. StarRocks Reader

StarRocks Reader插件用于从StarRocks数据库读取数据。本文为您介绍StarRocks Reader支持的数据类型、字段映射和数据源等参数及配置示例。

使用限制

支持EMR-StarRocks 2.1版本。

实现原理

StarRocks Reader通过JDBC连接器连接至远程的StarRocks数据库，根据您配置的信息生成查询语句，并发送至远程StarRocks数据库。然后使用数据集成自定义的数据类型，将该SQL执行返回结果拼装为抽象的数据集，并传递给下游Writer处理。

支持的数据类型

StarRocks Reader支持大部分StarRocks类型，包括数值类型、字符串类型、日期类型。

参数说明

参数	描述	是否必选	默认值
datasource	StarRocks数据源名称。	是	无
selectedDatabase	StarRocks数据库名称。	否	StarRocks数据源内配置的数据库名称。
column	所配置的表中需要同步的列名集合。	是	无
where	筛选条件，在实际业务场景中，往往会选择当天的数据进行同步，将where条件指定为 `gmt_create>$bizdate` 。 • where条件可以有效地进行业务增量同步。 • where语句，包括不提供where的key或value，数据同步均视作同步全量数据。	否	无
table	选取的需要同步的表名称。	是	无
splitPk	StarRocks Reader进行数据抽取时，如果指定splitPk，表示您希望使用splitPk代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。推荐splitPk用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。	否	无

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述脚本模式参数说明中的datasource，通常输入您配置的数据源名称。
数据库	即上述脚本模式参数说明中的selectedDatabase，选择要读取的数据库名称。
表	即上述参数说明中的table。
数据过滤	填写where过滤语句对读取数据进行过滤，不需要填写where关键词。

参数	描述
切分键	您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 **说明** 切分键与数据同步中的选择来源有关，配置数据来源时才显示切分键配置项。

2. 字段映射，即上述参数说明中的column。

   左侧的源头表字段和右侧的目标表字段为一一对应关系。单击**添加一行**可以增加单个字段，鼠标放至需要删除的字段上，即可单击**删除**图标进行删除 。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

参数	描述
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下所示，具体参数填写请参见参数说明。

```
{
    "stepType": "starrocks",
    "parameter": {
        "selectedDatabase": "didb1",
        "datasource": "starrocks_datasource",
        "column": [
            "id",
            "name"
        ],
        "where": "id>100",
        "table": "table1",
        "splitPk": "id"
    },
    "name": "Reader",
    "category": "reader"
}
```

6.3. 配置Writer插件

6.3.1. AnalyticDB for MySQL 2.0 Writer

本文为您介绍AnalyticDB for MySQL 2.0 Writer支持的数据类型、字段映射和数据源等参数及配置示例。

前提条件

数据集成通过实时导入的方式将数据导入AnalyticDB for MySQL 2.0中，要求您必须提前在AnalyticDB for MySQL 2.0中创建好实时表（普通表）。实时导入方式效率高，且流程简单。

开始配置AnalyticDB for MySQL 2.0 Writer插件前，请首先配置好数据源，详情请参见配置AnalyticDB for MySQL 2.0数据源。

类型转换列表

AnalyticDB for MySQL 2.0 Writer针对AnalyticDB for MySQL 2.0类型的转换列表，如下所示。

类型	AnalyticDB for MySQL 2.0数据类型
整数类	INT、TINYINT、SMALLINT、BIGINT
浮点类	FLOAT和DOUBLE
字符串类	VARCHAR
日期时间类	DATE和TIMESTAMP
布尔类	BOOLEAN

参数说明

参数	描述	必选	默认值
连接url	AnalyticDB for MySQL 2.0连接信息，格式为Address:Port。	是	无
数据库	AnalyticDB for MySQL 2.0的数据库名称。	是	无
Access Id	AnalyticDB for MySQL 2.0对应的AccessKey Id。	是	无
Access Key	AnalyticDB for MySQL 2.0对应的AccessKey Secret。	是	无
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	目标表的表名称。	是	无
partition	目标表的分区名称，当目标表为普通表，需要指定该字段。	否	无
writeMode	Insert模式，在主键冲突情况下新的记录会覆盖旧的记录。	是	无
column	目的表字段列表，可以为["*"]，或者具体的字段列表，例如["a", "b", "c"]。	是	无
suffix	AnalyticDB for MySQL 2.0 url配置项的格式为 `ip:port` ，此部分为您定制的连接串，是可选参数。实际在AnalyticDB for MySQL 2.0数据库访问时，会变成JDBC数据库连接串。例如配置suffix为 `autoReconnect=true&failOverReadOnly=false&maxReconnects=10` 。	否	无
batchSize	AnalyticDB for MySQL 2.0提交数据写的批量条数，当writeMode为insert时，该值才会生效。	writeMode为insert时，为必选。	无
bufferSize	DataX数据收集缓冲区大小，缓冲区的目的是积累一个较大的Buffer，源头的数据首先进入到此Buffer中进行排序，排序完成后再提交至AnalyticDB for MySQL 2.0。排序是根据AnalyticDB for MySQL 2.0的分区列模式进行的，排序的目的是数据顺序对AnalyticDB for MySQL 2.0服务端更友好（出于性能考虑）。 BufferSize缓冲区中的数据会经过batchSize批量提交至AnalyticDB for MySQL 2.0，通常需要设置bufferSize为batchSize数量的多倍。当writeMode为insert时，该值才会生效。	writeMode为insert时，为必选。	默认不配置不开启此功能。

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	选择AnalyticDB for MySQL 2.0，系统将自动关联配置AnalyticDB for MySQL 2.0数据源时设置的数据源名称。
表	选择AnalyticDB for MySQL 2.0中的一张表，将Reader数据库中的数据同步至该表中。

参数	描述
导入模式	根据AnalyticDB for MySQL 2.0中表的更新方式设置导入模式，包括**批量导入**和**实时导入**。 **说明** 批量导入不支持从非MaxCompute数据源批量导入数据至AnalyticDB for MySQL 2.0。请配置两个同步任务，先将数据导入MaxCompute，再批量导入AnalyticDB for MySQL 2.0。
导入规则	**写入前清理已有数据**：导数据之前，清空表或者分区的所有数据，相当于 `insert overwrite` 。
一级分区	默认，不可以修改。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",
    "steps":[
        {
            "stepType":"stream",
            "parameter":{
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"ads",//插件名。
            "parameter":{
                "partition":"",//目标表的分区名称。
                "datasource":"",//数据源。
                "column":[//字段。
                     "id"
                ],
                "writeMode":"insert",//写入模式。
                "batchSize":"256",//一次性批量提交的记录数大小。
                "table":"",//表名。
                "overWrite":"true"//AnalyticDB for MySQL 2.0写入是否覆盖当前写入的表，true为覆盖写入，false为不覆盖
。（追加）写入。当 writeMode 为 Load 时，该值才会生效。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.2. DataHub Writer

本文为您介绍DataHub Writer支持的数据类型、字段映射和数据源等参数及配置示例。

DataHub是实时数据分发平台、流式数据（Streaming Data）的处理平台，提供对流式数据的发布（Publish）、订阅（Subscribe）和分发功能，让您可以轻松构建基于流式数据的分析和应用。

DataHub服务基于阿里云自研的飞天平台，具有高可用、低延迟、高可扩展和高吞吐的特点。它与阿里云流计算引擎StreamCompute无缝连接，您可以轻松使用SQL进行流数据分析。DataHub同时提供分发流式数据至MaxCompute（原ODPS）、OSS等云产品的功能。

> 注意 STRING字符串仅支持UTF-8编码，单个STRING列最长允许1 MB。

参数配置

通过Channel将Source与Sink连接，在Writer端的Channel要对应Reader端的Channel类型。通常Channel包括Memory-Channel和File-channel两种类型，如下配置即File通道。

```
"agent.sinks.dataXSinkWrapper.channel": "file"
```

参数说明

参数	描述	是否必选	默认值
accessId	DataHub的accessId。	是	无
accessKey	DataHub的accessKey。	是	无
endPoint	对DataHub资源的访问请求，需要根据资源所属服务，选择正确的域名。	是	无
maxRetryCount	任务失败的最多重试次数。	否	无
mode	Value是STRING类型时，写入的模式。	是	无
parseContent	解析内容。	是	无
project	项目（Project）是DataHub数据的基本组织单元，一个Project下包含多个Topic。 **说明** DataHub的项目空间与MaxCompute的项目相互独立，您在MaxCompute中创建的项目不能复用于DataHub，需要单独创建。	是	无
topic	Topic是DataHub订阅和发布的最小单位，您可以用Topic来表示一类或者一种流数据。	是	无
maxCommitSize	为提高写出效率，DataX会积累Buffer数据，待积累的数据大小达到maxCommitSize 大小（单位MB）时，批量提交到目的端。默认是1,048,576，即1 MB数据。另外datahub侧对于一次request请求写入的数据条数限制是10000条，超出10000条数据会超出限制导致任务出错，请根据您单条数据平均数据量*10000条的数据总量来从侧方面进行写入datahub的数据条数控制。	否	*1MB*

向导开发介绍

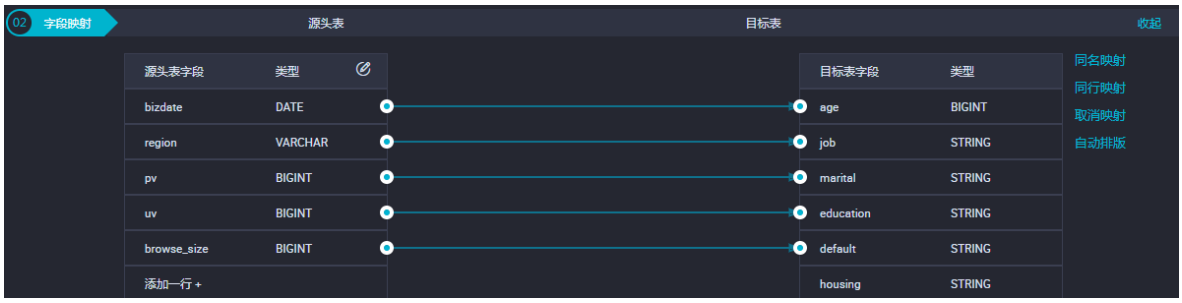
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	在下拉列表中选择您配置的数据源名称。

参数	描述
主题	即上述参数说明中的topic。
一次提交的数据量	一次向DataHub提交的数据量，单位byte。
最大重试次数	即上述参数说明中的maxRetryCount。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个从内存中读数据的同步作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "version": "2.0",//版本号。
    "steps": [
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "datahub",//插件名。
            "parameter": {
                "datasource": "",//数据源。
                "topic": "",//Topic是DataHub订阅和发布的最小单位，您可以用Topic来表示一类或者一种流数据。
                "maxRetryCount": 500,//任务失败的重试的最多次数。
                "maxCommitSize": 1048576//待积累的数据Buffer大小达到maxCommitSize大小（单位MB）时，批量提交至目的端
。
                 //datahub侧对于一次request请求写入的数据条数限制是10000条，超出10000条数据会超出限制导致任务出错，请根
据您单条数据平均数据量*10000条数据的数据总量来从侧方面进行单次写入datahub的数据条数控制。比如每条数据10 k，那么此参数的设置
值要低于10*10000 k。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": ""//错误记录数。
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":20, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.3.3. DB2 Writer

本文为您介绍DB2 Writer支持的数据类型、字段映射和数据源等参数及配置示例。

> 注意 DB2 Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

背景信息

DB2 Writer插件为您提供写入数据至DB2数据库的目标表的功能。在底层实现上，DB2 Writer通过JDBC连接远程DB2数据库，执行相应的 `insert into` 语句，写入数据至DB2，内部会分批次提交入库。

DB2 Writer面向ETL开发工程师，使用DB2 Writer从数仓导入数据至DB2。同时DB2 Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

DB2 Writer通过数据同步框架获取Reader生成的协议数据，通过 `insert into` （当主键/唯一性索引冲突时，冲突的行会写不进去）语句，写入数据至DB2。另外出于性能考虑采用了 `PreparedStatement + Batch` ，并且设置了 `rewriteBatchedStatements=true` ，将数据缓冲到线程上下文Buffer中，当Buffer累计到预定阈值时，才发起写入请求。

> **说明** 整个任务至少需要具备 `insert into` 的权限，是否需要其它权限，取决于您配置任务时在preSql和postSql中指定的语句。

DB2 Writer使用的DB2驱动版本为IBM Data Server Driver for JDBC and SQLJ 4.11.77。DB2驱动和数据库服务之间的版本映射请参见官网文档。

类型转换列表

DB2 Writer支持大部分DB2类型，但也存在个别没有支持的情况，请注意检查您的数据类型。

DB2 Writer针对DB2类型的转换列表，如下所示。

类型分类	DB2数据类型
整数类	SMALLINT
浮点类	DECIMAL、REAL和DOUBLE
字符串类	CHAR、CHARACTER、VARCHAR、GRAPHIC、VARGRAPHIC、LONG VARCHAR、CLOB、LONG VARGRAPHIC和DBCLOB
日期时间类	DATE、TIME和TIMESTAMP
布尔类	无
二进制类	BLOB

参数说明

参数	描述	是否必选	默认值
jdbcUrl	描述的是到DB2数据库的JDBC连接信息，JDBC URL遵循DB2官方规范。DB2格式为jdbc:db2://ip:port/database，您还可以填写连接附件控制信息。	是	无
username	数据源的用户名 。	是	无
password	数据源指定用户名的密码 。	是	无
table	所选取的需要同步的表 。	是	无
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如："column": ["id", "name", "age"]。如果要依次写入全部列，使用（*）表示。例如"column": ["*"] 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句，目前仅允许执行一条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与Db2的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1024*

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

配置一个写入DB2的数据同步作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"db2",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之后率先执行的SQL语句。
                "password":"",//密码。
                "jdbcUrl":"jdbc:db2://ip:port/database",//DB2数据库的JDBC连接信息。
                "column":[
                    "id"
                ],
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "username":"",//用户名。
                "preSql":[]//执行数据同步任务之前执行的SQL语句。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.4. DRDS Writer

本文为您介绍DRDS Writer支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

DRDS Writer插件为您提供将数据写入DRDS表的功能。在底层实现上，DRDS Writer通过JDBC连接远程DRDS数据库的Proxy，执行相应的 `replace into` 语句，写入数据至DRDS。

> **说明**
> - 执行的SQL语句是 `replace into` ，为避免数据重复写入，需要您的表具备主键（Primary Key）或唯一性索引（Unique index）。
> - 开始配置DRDS Writer插件前，请首先配置好数据源，详情请参见配置DRDS数据源。
> - DataWorks不支持DRDS下的MySQL8.0版本。

DRDS Writer面向ETL开发工程师，从数据仓库导入数据至DRDS。同时DRDS Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

DRDS Writer通过数据同步框架获取Reader生成的协议数据，通过 `replace into` （没有遇到主键/唯一性索引冲突时，与 `insert into` 行为一致，冲突时会用新行替换原有行所有字段）语句写入数据至DRDS。DRDS Writer累积一定数据，提交给DRDS的Proxy，该Proxy内部决定数据是写入一张还是多张表，以及多张表写入时如何路由数据。

> **说明** 整个任务至少需要具备 `replace into` 的权限。是否需要其它权限，取决于您配置任务时在preSql和postSql中指定的语句。

类型转换列表

DRDS Writer支持大部分DRDS类型，但也存在个别类型没有支持的情况，请注意检查您的数据类型。

DRDS Writer针对DRDS类型的转换列表，如下所示。

类型分类	DRDS数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT、BIGINT和YEAR
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔类	BIT和BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

参数说明

参数	描述	必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	所选取的需要同步的表。	是	无
writeMode	选择导入模式，包括*insert ignore*和*replace into*： • *insert ignore*：当主键或约束冲突时，忽略来源数据。 • *replace into*：当主键或约束冲突时，使用来源数据替换目标数据。	否	*insert ignore*
column	目标表需要写入数据的字段，字段之间用英文逗号（,）分隔，例如"column": ["id", "name", "age"]。如果要依次写入全部列，使用（*）表示，例如"column": ["*"]。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。 例如 `delete * from table xxx;` ，表示同步写入数据前清理xxx表的数据。请根据自身需求进行配置。	否	无

参数	描述	必选	默认值
postSql	执行数据同步任务之后执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。 例如 `delete * from table xxx where xx=xx;` ，表示同步数据后，删除符合某条件的数据。请根据自身需求进行配置。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与DRDS的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

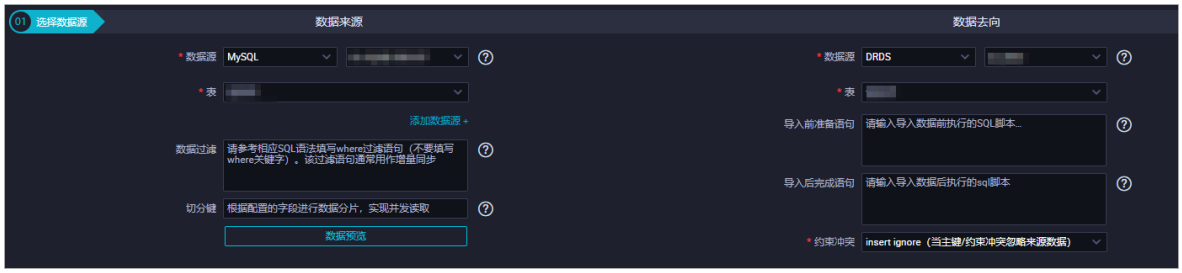
向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

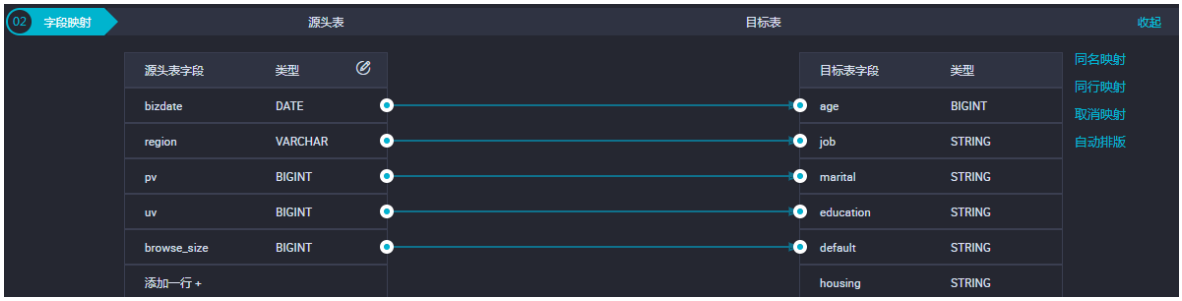
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个写入DRDS的数据同步作业，脚本开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
                },
        {
            "stepType":"drds",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之后执行的SQL语句。
                "datasource":"",//数据源。
                "column":[//列名。
                "id"
                ],
                "writeMode":"insert ignore",
                "batchSize":"1024",//一次性批量提交的记录数大小。
                "table":"test",//表名。
                "preSql":[]//执行数据同步任务之前执行的SQL语句。
                },
            "name":"Writer",
            "category":"writer"
                }
                ],
    "setting":{
        "errorLimit":{
        "record":"0"//错误记录数。
            },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
                }
            },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
                }
            ]
        }
    }
```

6.3.5. FTP Writer

本文为您介绍FTP Writer支持的数据类型、字段映射和数据源等参数及配置示例。

FTP Writer实现了向远程FTP文件写入CSV格式的一个或多个文件。在底层实现上，FTP Writer将数据集成传输协议下的数据转换为CSV格式，并使用FTP相关的网络协议写出至远程FTP服务器。

> **说明** 开始配置FTP Writer插件前，请首先配置好数据源，详情请参见配置FTP数据源。

写入FTP文件内容存放的是一张逻辑意义上的二维表，例如CSV格式的文本信息。

FTP Writer实现了从数据集成协议转为FTP文件功能，FTP文件本身是无结构化数据存储。目前FTP Writer支持的功能如下：

- 支持且仅支持写入文本类型（不支持BLOB，如视频数据）的文件，且要求文本中schema为一张二维表。
- 支持类CSV和TEXT格式的文件，自定义分隔符。

- 写出时不支持文本压缩。
- 支持多线程写入，每个线程写入不同子文件。

暂时不支持以下两种功能：

- 单个文件不能支持并发写入。
- FTP本身不提供数据类型，FTP Writer均将数据以STRING类型写入FTP文件。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
timeout	连接FTP服务器连接超时时间，单位毫秒。	否	*60,000*（1分钟）
path	FTP文件系统的路径信息，FTP Writer会写入Path目录下多个文件。	是	无
fileName	FTP Writer写入的文件名，该文件名会添加随机的后缀作为每个线程写入实际文件名。	是	无
singleFileOutput	FtpWriter写入的文件名受fileName控制，默认行会添加随机的后缀作为每个线程写入实际文件名。如果您不需要默认添加的随机后缀，您可以将singleFileOutput配置为true，写出文件名将会是您指定的完整文件名。	否	*false*
writeMode	FTP Writer写入前数据清理处理模式： • *truncate*：写入前清理目录下，fileName前缀的所有文件。 • *append*：写入前不做任何处理，数据集成FTP Writer直接使用filename写入，并保证文件名不冲突。 • *nonConflict*：如果目录下有fileName前缀的文件，直接报错。	是	无
fieldDelimiter	写入的字段分隔符。	是，单字符	无
skipHeader	类CSV格式文件可能存在表头为标题情况，需要跳过。默认不跳过，压缩文件模式下不支持skipHeader。	否	*false*
compress	支持*gzip*和*bzip2*两种压缩形式。	否	无压缩
encoding	读取文件的编码配置。	否	*utf-8*
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据集成提供nullFormat定义哪些字符串可以表示为null。 例如您配置 `nullFormat="null"` ，如果源头数据是null，数据集成视作null字段。	否	无
dateFormat	日期类型的数据序列化到文件中时的格式，例如"dateFormat":"yyyy-MM-dd"。	否	无
fileFormat	文件写出的格式，包括CSV和TEXT两种，CSV是严格的CSV格式，如果待写数据包括列分隔符，则会按照CSV的转义语法转义，转义符号为双引号。TEXT格式是用列分隔符简单分割待写数据，对于待写数据包括列分隔符情况下不做转义。	否	*TEXT*
header	header:txt文本（包括csv、text等）写出时的表头，脚本模式支持配置表头信息，例如"header":["id","name","age"]，表示将id、name、age作为表头写入ftp文件的第一行。	否	无
markDoneFileName	标档文件名，同步任务结束后生成标档文件，根据此标档文件可以判断同步任务是否成功。此处应配置为绝对路径。	否	无

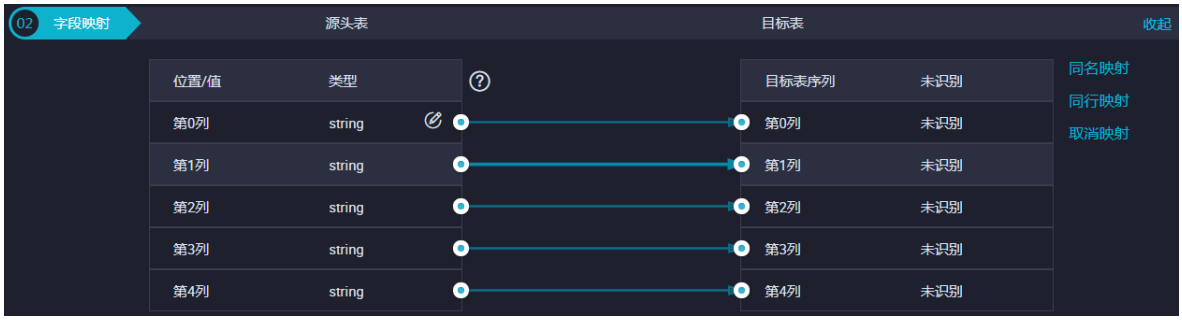
向导开发介绍

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
文件路径	即上述参数说明中的path。
文本类型	读取的文件类型，默认情况下文件作为csv格式文件进行读取。
列分隔符	即上述参数说明中的fieldDelimiter，默认值为（,）。
编码格式	即上述参数说明中的encoding，默认值为*utf-8*。
null值	即上述参数说明中的nullFormat，定义表示null值的字符串。
时间格式	即上述参数说明中的dateFormat。
前缀冲突	即上述参数说明中的writeMode，定义表示null值的字符串。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



配置	说明
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个写入FTP数据库的同步作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"ftp",//插件名。
            "parameter":{
                "path":"",//文件路径。
                "fileName":"",//文件名称。
                "nullFormat":"null",//null值。
                "dateFormat":"yyyy-MM-dd HH:mm:ss",//时间格式。
                "datasource":"",//数据源。
                "writeMode":"",//写入模式。
                "fieldDelimiter":",",//列分隔符。
                "encoding":"",//编码格式。
                "fileFormat":""//文本类型。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.6. HBase Writer

本文为您介绍HBase Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

HBase Writer插件实现了向HBase中写入数据。在底层实现上，HBase Writer通过HBase的Java客户端连接远程HBase服务，并通过put方式写入HBase。

使用限制

- HBase Reader不支持读取phoenix写入的数据，phoenix有特殊处理。
- HBase Reader仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

支持的功能

- 支持HBase0.94.x、HBase1.1.x和HBase2.x版本

- 如果您的HBase版本为HBase0.94.x，Writer端的插件请选择094x。

```
"writer": {
        "hbaseVersion": "094x"
    }
```

- 如果您的HBase版本为HBase1.1.x或HBase2.x，Writer端的插件请选择hbase11x。

```
"writer": {
        "hbaseVersion": "11x"
    }
```

> **说明** HBase1.1.x插件当前可以兼容HBase 2.0，如果您在使用上遇到问题请提交工单。

- 支持源端多个字段拼接作为rowkey

  目前HBase Writer支持源端多个字段拼接作为HBase表的rowkey。

- 写入HBase的版本支持

  写入HBase的时间戳（版本）支持：

  - 当前时间作为版本。
  - 指定源端列作为版本。
  - 指定一个时间作为版本。

支持的数据类型

支持读取HBase数据类型，HBase Writer针对HBase类型的转换列表，如下表所示。

> **说明**
> - column的配置需要和HBase表对应的列类型保持一致。
> - 除下表中罗列的字段类型外，其它类型均不支持。

类型分类	数据库数据类型
整数类	INT、LONG和SHORT
浮点类	FLOAT和DOUBLE
布尔类	BOOLEAN
字符串类	STRING

参数说明

参数	描述	是否必选	默认值
haveKerberos	haveKerberos值为true时，表示HBase集群需要kerberos认证。 **说明** - 如果该值配置为true，必须要配置以下kerberos认证相关参数： - kerberosKeytabFilePath - kerberosPrincipal - hbaseMasterKerberosPrincipal - hbaseRegionserverKerberosPrincipal - hbaseRpcProtection - 如果HBase集群没有kerberos认证，则不需要配置以上参数。	否	*false*

参数	描述	是否必选	默认值
hbaseConfig	连接HBase集群需要的配置信息，JSON格式。必填的配置为hbase.zookeeper.quorum，表示HBase的ZK链接地址。同时可以补充更多HBase client的配置，例如设置scan的cache、batch来优化与服务器的交互。 **说明** 如果是云HBase的数据库，需要使用内网地址连接访问。	是	无
mode	写入HBase的模式，目前仅支持*normal*模式，后续考虑动态列模式。	是	无
table	要写入的HBase表名（大小写敏感）。	是	无
encoding	编码方式，UTF-8或GBK，用于STRING转HBase byte[]时的编码。	否	*utf-8*
column	要写入的HBase字段： • index：指定该列对应Reader端column的索引，从0开始。 • name：指定HBase表中的列，格式必须为列族：列名。 • type：指定写入的数据类型，用于转换HBase byte[]。	是	无
rowkeyColumn	要写入的HBase的rowkey列： • index：指定该列对应Reader端column的索引，从0开始。如果是常量，index为-1。 • type：指定写入的数据类型，用于转换HBase byte[]。 • value：配置常量，常作为多个字段的拼接符。HBase Writer会将rowkeyColumn中所有列按照配置顺序进行拼接作为写入HBase的rowkey，不能全为常量。 配置格式如下所示。 `"rowkeyColumn": [ { "index":0, "type":"string" }, { "index":-1, "type":"string", "value":"_" } ]`	是	无
versionColumn	指定写入HBase的时间戳。支持当前时间、指定时间列或指定时间（三者选一），如果不配置则表示用当前时间。 • index：指定对应Reader端column的索引，从0开始，需保证能转换为LONG。 • type：如果是Date类型，会尝试用yyyy-MM-dd HH:mm:ss和yyyy-MM-dd HH:mm:ss SSS解析。如果是指定时间，则index为-1。 • value：指定时间的值，LONG类型。 配置格式如下所示。 `"versionColumn":{ "index":1 }` `"versionColumn":{ "index":-1, "value":123456789 }`	否	无

参数	描述	是否必选	默认值
nullMode	读取的数据为null值时，您可以通过以下两种方式解决： • *skip*：表示不向HBase写该列。 • *empty*：写入HConstants.EMPTY_BYTE_ARRAY，即new byte [0]。	否	*skip*
walFlag	HBase Client向集群中的RegionServer提交数据时（Put/Delete操作），首先会先写WAL（Write Ahead Log）日志（即HLog，一个RegionServer上的所有Region共享一个HLog），只有当WAL日志写成功后，才会接着写MemStore，最后客户端被通知提交数据成功。 如果写WAL日志失败，客户端则被通知提交失败。关闭（false）放弃写WAL日志，从而提高数据写入的性能。	否	*false*
writeBufferSize	设置HBase Client的写Buffer大小，单位字节，配合autoflush使用。 autoflush（默认处于关闭状态）： • 开启（true）：表示HBase Client在写的时候有一条put就执行一次更新。 • 关闭（false）：表示HBase Client在写的时候只有当put填满客户端写缓存时，才实际向HBase服务端发起写请求。	否	*8M*

向导开发介绍

暂不支持向导开发模式开发。

脚本开发介绍

配置一个从本地写入hbase1.1.x的作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"hbase",//插件名。
            "parameter":{
                "mode":"normal",//写入HBase的模式。
                "walFlag":"false",//关闭（false）放弃写WAL日志。
                "hbaseVersion":"094x",//Hbase版本。
                "rowkeyColumn":[//要写入的HBase的rowkey列。
                    {
                        "index":"0",//序列号。
                        "type":"string"//数据类型。
                    },
                    {
                        "index":"-1",
                        "type":"string",
                        "value":"_"
                    }
                ],
                "nullMode":"skip",//读取的为null值时，如何处理。
                "column":[//要写入的HBase字段。
                    {
                        "name":"columnFamilyName1:columnName1",//字段名。
                        "index":"0",//索引号。
                        "type":"string"//数据类型。
                    },
                    {
```

```
                        {
                            "name":"columnFamilyName2:columnName2",
                            "index":"1",
                            "type":"string"
                        },
                        {
                            "name":"columnFamilyName3:columnName3",
                            "index":"2",
                            "type":"string"
                        }
                    ],
                    "encoding":"utf-8",//编码格式。
                    "table":"",//表名。
                    "hbaseConfig":{//连接HBase集群需要的配置信息，JSON格式。
                        "hbase.zookeeper.quorum":"hostname",
                        "hbase.rootdir":"hdfs: //ip:port/database",
                        "hbase.cluster.distributed":"true"
                    }
                },
                "name":"Writer",
                "category":"writer"
            }
        ],
        "setting":{
            "errorLimit":{
                "record":"0"//错误记录数。
            },
            "speed":{
                "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
                "concurrent":1, //作业并发数。
                "mbps":"12"//限流
            }
        },
        "order":{
            "hops":[
                {
                    "from":"Reader",
                    "to":"Writer"
                }
            ]
        }
}
```

6.3.7. HBase11xsql Writer

本文为您介绍HBase11xsql Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

HBase11xsql Writer插件实现了向Hbase中的SQL表（phoenix）批量导入数据。因为Phoenix对rowkey进行了数据编码，如果您直接使用HBaseAPI写入，需要手动转换数据，麻烦且易错。HBase11xsql Writer插件为您提供了单间的SQL表的数据导入方式。

在底层实现上，通过Phoenix的JDBC驱动，执行UPSERT语句向HBase写入数据。

> 注意 HBase11xsql Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

使用限制

Writer中的列的定义顺序必须与Reader的列顺序匹配，Reader中的列顺序定义了输出的每一行中，列的组织顺序。而Writer的列顺序，定义的是在收到的数据中，Writer期待的列的顺序。示例如下：

Reader的列顺序为c1，c2，c3，c4。

Writer的列顺序为x1，x2，x3，x4。

则Reader输出的列c1就会赋值给Writer的列x1。如果Writer的列顺序是x1，x2，x4，x3，则c3会赋值给x4，c4会赋值给x3。

支持的功能

支持带索引的表的数据导入，可以同步更新所有的索引表。

限制

HBase11xsql Writer插件的限制如下所示：

- 仅支持1.x系列的HBase。
- 仅支持通过phoenix创建的表，不支持原生HBase表。
- 不支持带时间戳的数据导入。

实现原理

通过Phoenix的JDBC驱动，执行UPSERT语句向表中批量写入数据。因为使用上层接口，所以可以同步更新索引表。

参数说明

参数	描述	是否必选	默认值
plugin	插件名字，必须是hbase11xsql。	是	无
table	要导入的表名，大小写敏感，通常phoenix表都是大写表名。	是	无
column	列名，大小写敏感。通常phoenix的列名都是大写。 **说明** • 列的顺序必须与Reader输出的列的顺序一一对应。 • 不需要填写数据类型，会自动从phoenix获取列的元数据。	是	无
hbaseConfig	hbase集群地址，zk为必填项，格式为ip1, ip2, ip3。 **说明** • 多个IP之间使用英文的逗号分隔。 • znode是可选的，默认值是/hbase。	是	无
batchSize	批量写入的最大行数。	否	*256*
nullMode	读取到的列值为null时，您可以通过以下两种方式进行处理： • *skip*：跳过该列，即不插入该列（如果之前已经存在，则会被删除）。 • *empty*：插入空值，值类型的空值是0，varchar的空值是空字符串。	否	*skip*

脚本开发介绍

脚本配置示例如下，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
  "type": "job",
  "version": "1.0",
  "configuration": {
    "setting": {
      "errorLimit": {
        "record": "0"
      },
      "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"1"//限流
      }
    },
    "reader": {
      "plugin": "odps",
      "parameter": {
        "datasource": "",
        "table": "",
        "column": [],
        "partition": ""
      }
    },
    "plugin": "hbase11xsql",
    "parameter": {
      "table": "目标hbase表名，大小写有关",
      "hbaseConfig": {
        "hbase.zookeeper.quorum": "目标hbase集群的ZK服务器地址，向PE咨询",
        "zookeeper.znode.parent": "目标hbase集群的znode，向PE咨询"
      },
      "column": [
        "columnName"
      ],
      "batchSize": 256,
      "nullMode": "skip"
    }
  }
}
```

常见问题

Q：并发设置多少比较合适？速度慢时增加并发有用吗？

A：数据导入进程默认JVM的堆大小是2GB，并发（channel数）是通过多线程实现的，开过多的线程有时并不能提高导入速度，反而可能因为过于频繁的GC导致性能下降。一般建议并发数（channel）为5-10。

Q：batchSize设置多少比较合适？

A：默认是256，但应根据每行的大小来计算最合适的batchSize。通常一次操作的数据量在2MB~4MB左右，用该值除以行大小，即可得到batchSize。

6.3.8. HDFS Writer

本文为您介绍HDFS Writer支持的数据类型、字段映射和数据源等参数及配置示例。

HDFS Writer提供向HDFS文件系统指定路径中写入TextFile文件、 ORCFile文件以及ParquetFile格式文件，文件内容可以与Hive中的表关联。开始配置HDFS Writer插件前，请首先配置好数据源，详情请参见配置Hive数据源。

> **说明** HDFS Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组。

使用限制

- 目前HDFS Writer仅支持TextFile、ORCFile和ParquetFile三种格式的文件，且文件内容存放的必须是一张逻辑意义上的二维表。

- 由于HDFS是文件系统，不存在schema的概念，因此不支持对部分列写入。
- 目前不支持DECIMAL、BINARY、ARRAYS、MAPS、STRUCTS和UNION等Hive数据类型。
- 对于Hive分区表目前仅支持一次写入单个分区。
- 对于TextFile，需要保证写入HDFS文件的分隔符与在Hive上创建表时的分隔符一致，从而实现写入HDFS数据与Hive表字段关联。
- 目前插件中的Hive版本为1.1.1，Hadoop版本为2.7.1（Apache为适配JDK1.7）。在Hadoop2.5.0、Hadoop2.6.0和Hive1.2.0测试环境中写入正常。

实现过程

HDFS Writer的实现过程如下所示：

1. 根据您指定的path，创建一个HDFS文件系统上不存在的临时目录。

   创建规则：path_随机。
2. 将读取的文件写入这个临时目录。
3. 全部写入后，将临时目录下的文件移动到您指定的目录（在创建文件时保证文件名不重复）。
4. 删除临时目录。如果在此过程中，发生网络中断等情况造成无法与HDFS建立连接，需要您手动删除已经写入的文件和临时目录。

> **说明** 数据同步需要使用Admin账号，并且有访问相应文件的读写权限。

数据类型转换

目前HDFS Writer支持大部分Hive类型，请注意检查您的数据类型。

HDFS Writer针对Hive数据类型的转换列表，如下所示。

> **说明** column的配置需要和Hive表对应的列类型保持一致。

类型分类	数据库数据类型
整数类	TINYINT、SMALLINT、INT和BIGINT
浮点类	FLOAT和DOUBLE
字符串类	CHAR、VARCHAR和STRING
布尔类	BOOLEAN
日期时间类	DATE和TIMESTAMP

参数说明

参数	描述	是否必选	默认值
defaultFS	Hadoop HDFS文件系统namenode节点地址，例如 `hdfs://127.0.0.1:9000` 。公共资源组不支持Hadoop高级参数HA的配置，请新增和使用自定义数据集成资源组。	是	无
fileType	文件的类型，目前仅支持您配置为*text*、*orc*和*parquet*： • *text*：表示Hive中的存储表，TextFile文件格式。 • *orc*：表示Hive中的压缩表，ORCFile文件格式。 • *parquet*：表示普通Parquet File文件格式。	是	无

参数	描述	是否必选	默认值
path	存储到Hadoop HDFS文件系统的路径信息，HDFS Writer会根据并发配置在path目录下写入多个文件。 为了与Hive表关联，请填写Hive表在HDFS上的存储路径。例如Hive上设置的数据仓库的存储路径为 `/user/hive/warehouse/` ，已建立数据库test表hello，则对应的存储路径为 `/user/hive/warehouse/test.db/hello` 。	是	无
fileName	HDFS Writer写入时的文件名，实际执行时会在该文件名后添加随机的后缀作为每个线程写入实际文件名。	是	无
column	写入数据的字段，不支持对部分列写入。 为了与Hive中的表关联，需要指定表中所有字段名和字段类型，其中name指定字段名，type指定字段类型。 您可以指定column字段信息，配置如下。 `"column": [ { "name": "userName", "type": "string" }, { "name": "age", "type": "long" } ]`	是（如果filetype为parquet，此项无需填写）	无
writeMode	HDFS Writer写入前数据清理处理模式： • *append*：写入前不做任何处理，数据集成HDFS Writer直接使用filename写入，并保证文件名不冲突。 • *nonConflict*：如果目录下有fileName前缀的文件，直接报错。 • `truncate` ：写入前清理fileName名称前缀匹配的所有文件。例如， `"fileName": "abc"` ，将清理对应目录所有abc开头的文件。 **说明** Parquet格式文件不支持Append，所以只能是noConflict。	是	无
fieldDelimiter	HDFS Writer写入时的字段分隔符，需要您保证与创建的Hive表的字段分隔符一致，否则无法在Hive表中查到数据。	是（如果filetype为parquet，此项无需填写）	无
compress	HDFS文件压缩类型，默认不填写，则表示没有压缩。 其中text类型文件支持gzip和bzip2压缩类型。	否	无
encoding	写文件的编码配置。	否	无压缩

参数	描述	是否必选	默认值
parquetSchema	写Parquet格式文件时的必填项，用来描述目标文件的结构，所以此项当且仅当fileType为parquet时生效，格式如下。 `message MessageType名 {` `是否必填，数据类型，列名;` `.....;` `}` 配置项说明如下： • MessageType名：填写名称。 • 是否必填：required表示非空，optional表示可为空。推荐全填optional。 • 数据类型：Parquet文件支持BOOLEAN、INT32、INT64、INT96、FLOAT、DOUBLE、BINARY（如果是字符串类型，请填BINARY）和FIXED_LEN_BYTE_ARRAY等类型。 **说明** 每行列设置必须以分号结尾，最后一行也要写上分号。 示例如下。 `message m {` `optional int64 id;` `optional int64 date_id;` `optional binary datetimestring;` `optional int32 dspId;` `optional int32 advertiserId;` `optional int32 status;` `optional int64 bidding_req_num;` `optional int64 imp;` `optional int64 click_num;` `}`	否	无
hadoopConfig	hadoopConfig中可以配置与Hadoop相关的一些高级参数，例如HA的配置。公共资源组不支持Hadoop高级参数HA的配置，请新增和使用自定义数据集成资源组。 `"hadoopConfig":{` `"dfs.nameservices": "testDfs",` `"dfs.ha.namenodes.testDfs":` `"namenode1,namenode2",` `"dfs.namenode.rpc-address.youkuDfs.namenode1":` `"",` `"dfs.namenode.rpc-address.youkuDfs.namenode2":` `"",` `"dfs.client.failover.proxy.provider.testDfs":` `"org.apache.hadoop.hdfs.server.namenode.ha.Conf` `iguredFailoverProxyProvider"` `}`	否	无
	针对Parquet文件进行同步的模式。使用fields支持array、map和struct等复杂类型。可选值包括fields和columns。 配置dataxParquetMode为fields时，支持hdfs over oss，即HDFS的存储为OSS，OSS的数据存储格式为parquet。此时您可以在hadoopConfig中增加OSS相关的参数，详情如下： • fs.oss.accessKeyId：访问OSS的AccessKeyID。 • fs.oss.accessKeySecret：访问OSS的AccessKeySecret。		

参数	描述	是否必选	默认值
dataxParquetMode	• fs.oss.endpoint：访问OSS的endpoint。 访问OSS的示例如下所示。 ```json "writer": { "name": "hdfswriter", "parameter": { "defaultFS": "oss://test-bucket", "fileType": "parquet", "path": "/datasets/oss_demo/kpt", "fileName": "test", "writeMode": "truncate", "encoding": "UTF-8", "hadoopConfig": { "fs.oss.accessKeyId": "the-access-id", "fs.oss.accessKeySecret": "the-access-key", "fs.oss.endpoint": "oss-cn-hangzhou.aliyuncs.com" }, "parquetSchema": "message test {\n required int64 id;\n optional binary name (UTF8);\n optional int64 gmt_create;\n required group map_col (MAP) {\n repeated group key_value {\n required binary key (UTF8);\n required binary value (UTF8);\n }\n }\n required group array_col (LIST) {\n repeated group list {\n required binary element (UTF8);\n }\n }\n required group struct_col {\n required int64 id;\n required binary name (UTF8);\n }\n}", "dataxParquetMode": "fields" } } ```	否	*columns*

参数	描述	是否必选	默认值
haveKerberos	是否有Kerberos认证，默认为*false*。如果您配置为*true*，则配置项kerberosKeytabFilePath和kerberosPrincipal为必填。	否	*false*
kerberosKeytabFilePath	Kerberos认证keytab文件的绝对路径。	如果haveKerberos为*true*，则必选。	无
kerberosPrincipal	Kerberos认证Principal名，如****/hadoopclient@**.*** 。如果haveKerberos为*true*，则必选。 由于Kerberos需要配置keytab认证文件的绝对路径，您需要在自定义资源组上使用此功能。配置示例如下。 `"haveKerberos":true,` `"kerberosKeytabFilePath":"/opt/datax/**.keytab",` `"kerberosPrincipal":"**/hadoopclient@**.**"`	否	无

向导开发介绍

暂不支持向导开发模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "version": "2.0",//版本号。
    "steps": [
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "hdfs",//插件名。
            "parameter": {
                "path": "",//存储到Hadoop HDFS文件系统的路径信息。
                "fileName": "",//HDFS Writer写入时的文件名。
                "compress": "",//HDFS文件压缩类型。
                "datasource": "",//数据源。
                "column": [
                    {
                        "name": "col1",//字段名。
                        "type": "string"//字段类型。
                    },
                    {
                        "name": "col2",
                        "type": "int"
                    },
                    {
                        "name": "col3",
                        "type": "double"
                    },
                    {
                        "name": "col4",
                        "type": "boolean"
                    },
                    {
```

```
                        "name": "col5",
                        "type": "date"
                    }
                ],
                "writeMode": "",//写入模式。
                "fieldDelimiter": ",",//列分隔符。
                "encoding": "",//编码格式。
                "fileType": "text"//文本类型。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": ""//错误记录数。
        },
        "speed": {
            "concurrent": 3,//作业并发数。
            "throttle": false //false代表不限流，下面的限流的速度不生效；true代表限流。
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.3.9. Memcache（OCS） Writer

本文为您介绍Memcache（OCS） Writer支持的数据类型、字段映射和数据源等参数及配置示例。

云数据库Memcache版（ApsaraDB for Memcache，原简称OCS）是一种高性能、高可靠、可平滑扩容的分布式内存数据库服务。基于飞天分布式系统及高性能存储，并提供了双机热备、故障恢复、业务监控和数据迁移等方面的全套数据库解决方案。

云数据库Memcache版支持即开即用的方式快速部署，对于动态Web、APP应用，可以通过缓存服务减轻对数据库的压力，从而提高网站整体的响应速度。

云数据库Memcache版与本地MemCache的异同点如下：

- 相同点：云数据库Memcache版兼容Memcached协议，与您的环境兼容，可以直接用于云数据库Memcache版服务。
- 不同点：云数据库Memcache版的硬件和数据部署在云端，有完善的基础设施、网络安全保障和系统维护等服务。所有服务只需要按量付费即可。

Memcache Writer基于Memcached协议的数据写入Memcache通道。

Memcache Writer目前支持一种格式的写入方式，不同写入方式的类型转换方式不一致。

- text：Memcache Writer将来源数据序列化为STRING类型格式，并使用您的fieldDelimiter作为间隔符。
- binary：目前暂不支持。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无

参数	描述	是否必选	默认值
writeMode	Memcache Writer写入方式，具体如下： • *set*：存储这个数据。 • *add*：存储这个数据，当且仅当这个key不存在时（目前不支持）。 • *replace*：存储这个数据，当且仅当这个key存在时（目前不支持）。 • *append*：将数据存放在已存在的key对应的内容后面，忽略exptime（目前不支持）。 • *prepend*：将数据存放在已存在的key对应的内容的前面，忽略 exptime（目前不支持）。	是	无
writeFormat	Memcache Writer写出数据的格式，目前仅支持TEXT数据写入方式。 TEXT：将源端数据序列化为文本格式，其中第一个字段作为Memcache写入的key，后续所有字段序列化为String类型，使用您指定的fieldDelimiter作为间隔符，将文本拼接为完整的字符串再写入Memcache。 例如源头数据如下所示。 `| ID | NAME | COUNT|` `| ---- |:------|:-----|` `| 23 | "CDP" | 100 |` 如果您指定fieldDelimiter为\^，则写入Memcache的格式如下。 `| KEY (OCS) | VALUE(OCS) |` `| --------- |:---------- |` `| 23 | CDP\^100 |`	否	无
expireTime	Memcache值缓存失效时间，目前MemCache支持两类过期时间。 • Unix时间（自1970.1.1开始到现在的秒数），该时间指定了到未来某个时刻的数据失效。 • 相对当前时间的秒数，该时间指定了从现在开始多长时间后数据失效。 **说明** 如果过期时间的秒数大于60*60*24*30（即30天），则服务端认为是Unix时间。	否	*0*，0永久有效
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与MySQL的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个写入Memcache的数据同步作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"ocs",//插件名
            "parameter":{
                "writeFormat":"text",//Memcache Writer写出数据格式。
                "expireTime":1000,//Memcache值缓存失效时间。
                "indexes":0,
                "datasource":"",//数据源。
                "writeMode":"set",//写入模式。
                "batchSize":"256"//一次性批量提交的记录数大小。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.10. MongoDB Writer

本文为您介绍MongoDB Writer支持的数据类型、写入方式、字段映射和数据源等参数和配置示例。

背景信息

MongoDB Writer插件利用MongoDB的Java客户端MongoClient进行MongoDB的写操作。最新版本的Mongo已经将DB锁的粒度从DB级别降低到Document级别，配合MongoDB强大的索引功能，基本可以满足数据源向MongoDB写入数据的需求。针对数据更新的需求，也可以通过配置业务主键的方式实现。

> **说明**
>
> - 在开始配置MongoDB Writer插件前，请首先配置好数据源，详情请参见配置MongoDB数据源。
> - 如果您使用的是云数据库MongoDB版，MongoDB默认会有用root账号。
> - 出于安全策略的考虑，数据集成仅支持使用 MongoDB数据库对应账号进行连接。您在添加使用MongoDB数据源时，请避免使用root作为访问账号。

MongoDB Writer通过数据集成框架获取Reader生成的协议数据，然后将支持的类型通过逐一判断转换为MongoDB支持的类型。数据集成本身不支持数组类型，但MongoDB支持数组类型，并且数组类型具有强大的索引功能。

您可以通过参数的特殊配置，将字符串转换为MongoDB中的数组。转换类型后，即可并行写入MongoDB。

类型转换列表

MongoDB Writer支持大部分MongoDB类型，但也存在部分没有支持的情况，请注意检查您的数据类型。

MongoDB Writer针对MongoDB类型的转换列表，如下所示。

类型分类	MongoDB数据类型
整数类	INT和LONG
浮点类	DOUBLE
字符串类	STRING和ARRAY
日期时间类	DATE
布尔型	BOOL
二进制类	BYTES

> **说明** 此处DATE类型，写入至MongoDB后即为DATETIME类型。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
collectionName	MongoDB的集合名。	是	无
column	MongoDB的文档列名，配置为数组形式表示MongoDB的多个列。 • name：Column的名字。 • type：Column的类型。 • splitter：特殊分隔符，当且仅当要处理的字符串要用分隔符分隔为字符数组Array时，才使用此参数。通过此参数指定的分隔符，将字符串分隔存储到MongoDB的数组中。	是	无
writeMode	指定了传输数据时是否覆盖的信息，包括*isReplace*和*replaceKey*： • *isReplace*：当设置为true时，表示针对相同的replaceKey做覆盖操作。当设置为false时，表示不覆盖。 • *replaceKey*：replaceKey指定了每行记录的业务主键，用来做覆盖时使用（不支持replaceKey为多个键，通常指Mongo中的主键）。	否	无
preSql	表示数据同步写出MongoDB前的前置操作，例如清理历史数据等。如果preSql为空，表示没有配置前置操作。配置preSql时，需要确保preSql符合JSON语法要求。	否	无

执行数据集成作业时，会首先执行您已配置的preSql。完成preSql的执行后，才可以进入实际的数据写出阶段。preSql本身不会影响写出的数据内容。数据集成通过preSql参数，可以具备幂等执行特性。例如，您的preSql在每次任务执行前都会清理历史数据（根据您的业务规则进行清理）。此时，如果任务失败，您只需要重新执行数据集成作业即可。

preSql的格式要求如下：

- 需要配置type字段，表示前置操作类别，支持drop和remove，例如 `"preSql":{"type":"remove"}` ：
  - *drop*：表示删除集合和集合内的数据，collectionName参数配置的集合即是待删除的集合。
  - *remove*：表示根据条件删除数据。

- *json*：您可以通过JSON控制待删除的数据条件，例如 `"preSql":{"type":"remove", "json":"{'operationTime':{'$gte':ISODate('${last_day}T00:00:00.424+0800')}}"}` 。此处的 `${last_day}` 为DataWorks调度参数，格式为 `$[yyyy-mm-dd]` 。您可以根据需要具体使用其它MongoDB支持的条件操作符号（$gt、$lt、$gte和$lte等）、逻辑操作符（and和or等）或函数（max、min、sum、avg和ISODate等）。

  数据集成通过如下MongoDB标准API执行您的数据，删除query。

  ```
  query=(BasicDBObject) com.mongodb.util.JSON.parse(json);
  col.deleteMany(query);
  ```

  > **说明** 如果您需要条件删除数据，建议优先使用JSON配置形式。

- *item*：您可以在item中配置数据过滤的列名（name）、条件（condition）和列值（value）。例如 `"preSql":{"type":"remove","item":[{"name":"pv","value":"100","condition":"$gt"},{"name":"pid","value":"10"}]}` 。

  数据集成会基于您配置的item条件项，构造查询query条件，进而通过MongoDB标准API执行删除。例如 `col.deleteMany(query);` 。

- 不识别的preSql，无需进行任何前置删除操作。

向导开发介绍

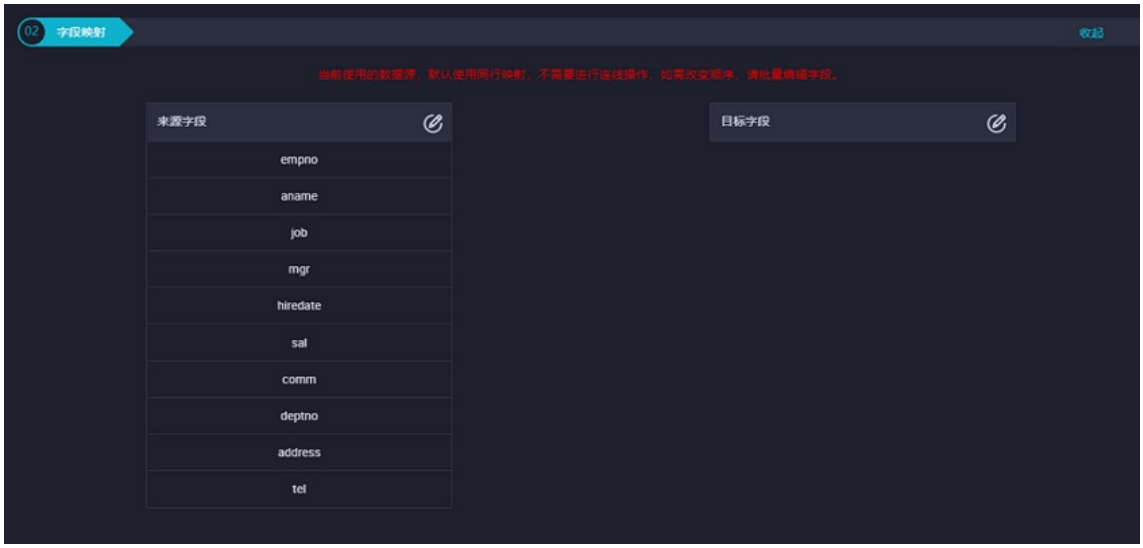
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常输入您配置的数据源名称。
集合名称	即上述参数说明中的collectionName。
写入模式（是否覆盖）	即上述参数说明中的writeMode。
前置条件	即上述参数说明中的preSql。表示数据同步写出MongoDB前的前置操作，例如清理历史数据等。如果preSql为空，表示没有配置前置操作。配置preSql时，需要确保preSql符合JSON语法要求。

2. 字段映射，即上述参数说明中的column。默认使用**同行映射**。您可以单击图标手动编辑目标表字段。



3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

配置写入MongoDB的数据同步作业，详情请参见上述参数说明。

```
{
    "type": "job",
    "version": "2.0",//版本号。
    "steps": [
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "mongodb",//插件名。
            "parameter": {
                "datasource": "",//数据源名。
```

```
                "column": [
                    {
                        "name": "_id",//列名。
                        "type": "ObjectId"//数据类型。如果replacekey为_id，则此处的type必须配置为ObjectID。如果配置为string，会无法进行替换。
                    },
                    {
                        "name": "age",
                        "type": "int"
                    },
                    {
                        "name": "id",
                        "type": "long"
                    },
                    {
                        "name": "wealth",
                        "type": "double"
                    },
                    {
                        "name": "hobby",
                        "type": "array",
                        "splitter": " "
                    },
                    {
                        "name": "valid",
                        "type": "boolean"
                    },
                    {
                        "name": "date_of_join",
                        "format": "yyyy-MM-dd HH:mm:ss",
                        "type": "date"
                    }
                ],
                "writeMode": {//写入模式。
                    "isReplace": "true",
                    "replaceKey": "_id"
                },
                "collectionName": "datax_test"//连接名称。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {//错误记录数。
            "record": "0"
        },
        "speed": {
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent": 1,//作业并发数。
            "mbps": "1"//限流的速度。
        },
       "jvmOption": "-Xms1024m -Xmx1024m"
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.3.11. MySQL Writer

本文为您介绍MySQL Writer支持的数据类型、字段映射和数据源等参数及配置示例。

前提条件

开始配置MySQL Writer插件前，请首先配置好数据源，详情请参见配置MySQL数据源。

背景信息

MySQL Writer插件实现了写入数据至MySQL数据库目标表的功能。在底层实现上，MySQL Writer通过JDBC连接远程MySQL数据库，并执行相应的 `insert into` 或 `replace into` 语句，写入数据至MySQL。数据库本身采用InnoDB引擎，以分批次提交数据入库。

MySQL Writer作为数据迁移工具，为数据库管理员等用户提供服务。根据您配置的writeMode，通过数据同步框架获取Reader生成的协议数据。

> **说明** 整个任务必须具备 `insert/replace into` 的权限。您可以根据配置任务时，在preSql和postSql中指定的语句，判断是否需要其它权限。

类型转换列表

目前MySQL Writer支持大部分MySQL类型，但也存在个别类型没有支持的情况，请注意检查您的数据类型。

MySQL Writer针对MySQL类型的转换列表，如下所示。

类型分类	MySQL数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT、BIGINT和YEAR
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔型	BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无

参数	描述	是否必选	默认值
writeMode	选择导入模式，可以支持*insert into*、*on duplicate key update*和*replace into*三种方式： • *insert into*：当主键/唯一性索引冲突时会写不进去冲突的行，以脏数据的形式体现。 如果您通过脚本模式配置任务，请设置writeMode为*insert*。 • *on duplicate key update*：没有遇到主键/唯一性索引冲突时，与 `insert into` 行为一致。冲突时会用新行替换已经指定的字段的语句，写入数据至MySQL。 如果您通过脚本模式配置任务，请设置writeMode为*update*。 • *replace into*：没有遇到主键/唯一性索引冲突时，与 `insert into` 行为一致。冲突时会先删除原有行，再插入新行。即新行会替换原有行的所有字段。 如果您通过脚本模式配置任务，请设置writeMode为*replace*。	否	*insert into*
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔，例如 `"column": ["id", "name", "age"]` 。如果要依次写入全部列，使用星号（*）表示， 例如 `"column": ["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，执行前清空表中的旧数据（truncate table tablename）。 **说明** 当有多条SQL语句时，不支持事务。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，加上某一个时间戳 `alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP` 。 **说明** 当有多条SQL语句时，不支持事务。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与MySQL的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

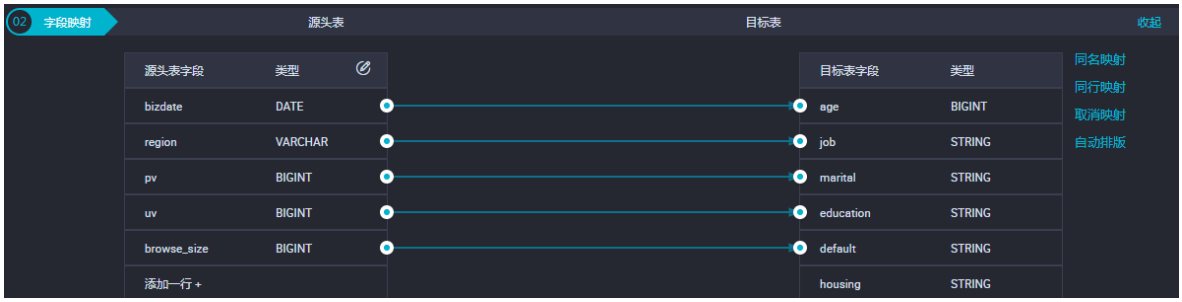
您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置示例如下，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"mysql",//插件名。
            "parameter":{
                "postSql":[],//导入后的准备语句。
                "datasource":"",//数据源。
                "column":[//列名。
                    "id",
                    "value"
                ],
                "writeMode":"insert",//写入模式，您可以设置为insert、replace或update。
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "preSql":[
                     "delete from XXX;" //导入前的准备语句。
                   ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{//错误记录数。
            "record":"0"
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流，控制同步的最高速率，防止对上游/下游数据库读取/写入压力过大。
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.12. Oracle Writer

本文为您介绍Oracle Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

Oracle Writer插件实现了写入数据到Oracle主库的目标表的功能。在底层实现上，Oracle Writer通过JDBC连接远程Oracle数据库，并执行相应的 `insert into...` SQL语句，将数据写入Oracle。

> **说明**
>
> - 开始配置Oracle Writer插件前，请首先配置好数据源，详情请参见配置Oracle数据源。
> - Oracle Writer插件使用ojdbc6-12.1.1.jar驱动，支持的Oracle版本请参见Oracle官网。

Oracle Writer面向ETL开发工程师，使用Oracle Writer从数仓导入数据至Oracle。同时Oracle Writer也可以作为数据迁移工具，为数据库管理员等用户提供服务。

Oracle Writer通过数据同步框架获取Reader生成的协议数据，然后通过JDBC连接远程Oracle数据库，并执行相应的SQL语句，将数据写入Oracle。

类型转换列表

Oracle Writer支持大部分Oracle类型，但也存在个别类型没有支持的情况，请注意检查您的数据类型。

Oracle Writer针对Oracle类型的转换列表，如下所示。

类型分类	Oracle数据类型
整数类	NUMBER、RAWID、INTEGER、INT和SMALLINT
浮点类	NUMERIC、DECIMAL、FLOAT、DOUBLE PRECISION和REAL
字符串类	LONG、CHAR、NCHAR、VARCHAR、VARCHAR2、NVARCHAR2、CLOB、NCLOB、CHARACTER、CHARACTER VARYING、CHAR VARYING、NATIONAL CHARACTER、NATIONAL CHAR、NATIONAL CHARACTER VARYING、NATIONAL CHAR VARYING和NCHAR VARYING
日期时间类	TIMESTAMP和DATE
布尔型	BIT和BOOL
二进制类	BLOB、BFILE、RAW和LONG RAW

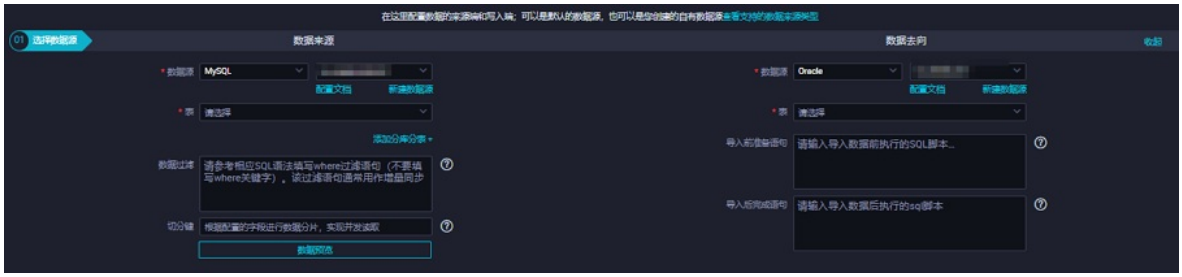
参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	目标表名称，如果表的schema信息和上述配置username不一致，请使用schema.table的格式填写table信息。	是	无
writeMode	选择导入模式，仅支持*insert into*。当主键或唯一性索引冲突时，会写不进去冲突的行，以脏数据的形式体现。	否	*insert into*
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如 `"column": ["id",”name”,”age”]` 。如果要依次写入全部列，使用*表示。例如 `"column": ["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与MySQL的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

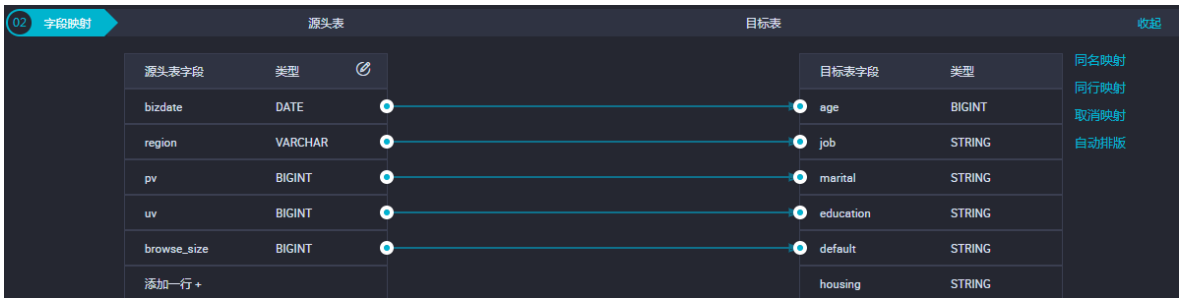
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。

2. 字段映射，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

配置一个写入Oracle的作业，使用脚本进行开发的详情请参见通过脚本模式配置离线同步任务。

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"oracle",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之后执行的SQL语句。
                "datasource":"",
                "session":[],//数据库连接会话参数。
                "column":[//字段。
                    "id",
                    "name"
                ],
                "encoding":"UTF-8",//编码格式。
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "preSql":[]//执行数据同步任务之前执行的SQL语句。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.13. OSS Writer

本文为您介绍OSS Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

背景信息

OSS Writer插件为您提供向OSS写入类CSV格式的一个或多个表文件的功能，写入的文件个数和您的任务并发及同步的文件数有关。

> **说明** 开始配置OSS Writer插件前，请首先配置好数据源，详情请参见配置OSS数据源。

写入OSS中的是一张逻辑意义上的二维表，例如CSV格式的文本信息。如果您想对OSS产品有更深入的了解，请参见OSS产品概述。

OSS Writer实现了从数据同步协议转为OSS中的文本文件功能，OSS本身是无结构化数据存储，目前OSS Writer支持的功能如下：

- 支持且仅支持写入文本文件，并要求文本文件中的Schema为一张二维表。
- 支持类CSV格式文件，自定义分隔符。
- 支持多线程写入，每个线程写入不同的子文件。
- 文件支持滚动，当文件大于某个size值时，支持文件切换。

OSS Writer暂时不能实现以下功能：

- 单个文件不能支持并发写入。
- OSS本身不提供数据类型，OSS Writer均以STRING类型写入OSS对象。

参数说明

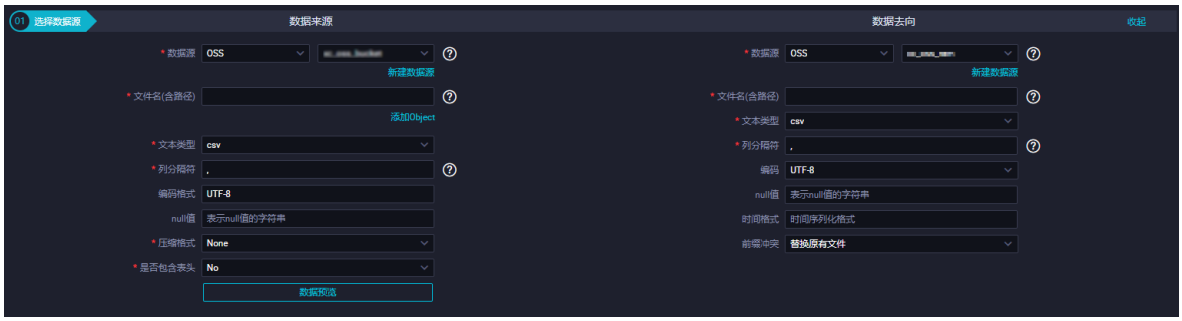
参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
object	OSS Writer写入的文件名，OSS使用文件名模拟目录的实现。OSS对于Object的名称有以下限制： • 使用 `"object": "datax"` ，写入的Object以datax开头，后缀添加随机字符串。 • 使用 `"object": "cdo/datax"` ，写入的Object以 `/cdo/datax` 开头，后缀随机添加字符串，OSS模拟目录的分隔符为（/）。 如果您不需要后缀随机UUID，建议您配置 `"writeSingleObject" : "true"` ，详情请参见writeSingleObject说明。	是	无
writeMode	OSS Writer写入前，数据的处理： • *truncate*：写入前清理Object名称前缀匹配的所有Object。例如 `"object":"abc"` ，将清理所有abc开头的Object。 • *append*：写入前不进行任何处理，数据集成OSS Writer直接使用Object名称写入，并使用随机UUID的后缀名来保证文件名不冲突。例如您指定的Object名为数据集成，实际写入为DI_****_****_****。 • *nonConflict*：如果指定路径出现前缀匹配的Object，直接报错。例如 `"object":"abc"` ，如果存在abc123的Object，将直接报错。	是	无
writeSingleObject	OSS写数据时，是否写单个文件： • *true*：表示写单个文件。 • *false*：表示写多个文件。	否	*false*
fileFormat	文件写出的格式，包括*csv*和*text*： • *csv*仅支持严格的*csv*格式。如果待写数据包括列分隔符，则会根据*csv*的转义语法转义，转义符号为双引号（"）。 • *text*格式指使用列分隔符简单分割待写数据，对于待写数据包括列分隔符情况下不进行转义。 **说明** 支持写入parquet文件类型，若使用此文件类型，必须增加parquetschema参数定义数据类型。	否	*text*
compress	写入OSS的数据文件的压缩格式（需使用脚本模式任务配置）。 **说明** csv、text文本类型不支持压缩，parquet/orc文件支持zip、snappy等压缩。	否	无

参数	描述	是否必选	默认值
fieldDelimiter	写入的字段分隔符。	否	,
encoding	写出文件的编码配置。	否	*utf-8*
nullFormat	文本文件中无法使用标准字符串定义null（空指针），数据同步系统提供nullFormat定义可以表示为null的字符串。例如，您配置 `nullFormat="null"` ，如果源头数据是 `null` ，数据同步系统会视作null字段。	否	无
header	OSS写出时的表头，例如， `["id", "name", "age"]` 。	否	无
maxFileSize（高级配置，向导模式不支持）	OSS写出时单个Object文件的最大值，默认为10,000*10MB，类似于在打印log4j日志时，控制日志文件的大小。OSS分块上传时，每个分块大小为10MB（也是日志轮转文件最小粒度，即小于10MB的maxFileSize会被作为10MB），每个OSS InitiateMultipartUploadRequest支持的分块最大数量为10,000。 轮转发生时，Object名字规则是在原有Object前缀加UUID随机数的基础上，拼接_1,_2,_3等后缀。	否	100,000 **说明** 默认单位为MB。 配置示例："maxFileSize":300，表示设置单个文件大小为300M。
suffix（高级配置，向导模式不支持）	数据同步写出时，生成的文件名后缀。例如，配置suffix为*.csv*，则最终写出的文件名为fileName****.csv。	否	无

向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
文件名（含路径）	即上述参数说明中的Object，填写OSS文件夹的路径，其中不要填写bucket的名称。
文本类型	包括*csv*、*text*和parquet。
列分隔符	即上述参数说明中的fieldDelimiter，默认值为（,）。
编码格式	即上述参数说明中的encoding，默认值为*utf-8*。

参数	描述
null值	即上述参数说明中的nullFormat，将要表示为空的字段填入文本框，如果源端存在则将对应的部分转换为空。
时间格式	日期类型的数据序列化到Object时的格式，例如 `"dateFormat": "yyyy-MM-dd"` 。
前缀冲突	有同样的文件时，可以选择替换、保留或报错。

2. 字段映射。左侧的源头表字段和右侧的目标表字段为一一对应的关系。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置示例如下所示，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"oss",//插件名。
            "parameter":{
                "nullFormat":"",//数据同步系统提供nullFormat，定义哪些字符串可以表示为null。
                "dateFormat":"",//日期格式。
                "datasource":"",//数据源。
                "writeMode":"",//写入模式。
                "writeSingleObject":"false", //表示是否将同步数据写入单个oss文件。
                "encoding":"",//编码格式。
                "fieldDelimiter":",",//字段分隔符。
                "fileFormat":"",//文本类型。
                "object":""//Object前缀。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

ORC或Parquet文件写入OSS

目前通过复用HDFS Writer的方式完成OSS写ORC或Parquet格式的文件。在OSS Writer已有参数的基础上，增加了Path、FileFormat等扩展配置参数，参数含义请参见HDFS Writer。

ORC或Parquet文件写入OSS的示例如下：

> **注意** 以下仅为示例，请根据您自己具体的列名称和类型修改对应的参数，请勿直接复制使用。

- 以ORC文件格式写入OSS。

```
{
      "stepType": "oss",
      "parameter": {
        "datasource": "",
        "fileFormat": "orc",
        "path": "/tests/case61",
        "fileName": "orc",
        "writeMode": "append",
        "column": [
          {
            "name": "col1",
            "type": "BIGINT"
          },
          {
            "name": "col2",
            "type": "DOUBLE"
          },
          {
            "name": "col3",
            "type": "STRING"
          }
        ],
        "writeMode": "append",
        "fieldDelimiter": "\t",
        "compress": "NONE",
        "encoding": "UTF-8"
      }
    }
```

- 以Parquet文件格式写入OSS，示例如下。

```
{
      "stepType": "oss",
      "parameter": {
        "datasource": "",
        "fileFormat": "parquet",
        "path": "/tests/case61",
        "fileName": "test",
        "writeMode": "append",
        "fieldDelimiter": "\t",
        "compress": "SNAPPY",
        "encoding": "UTF-8",
        "parquetSchema": "message test { required int64 int64_col;\n required binary str_col (UTF8);\nrequi
red group params (MAP) {\nrepeated group key_value {\nrequired binary key (UTF8);\nrequired binary value (U
TF8);\n}\n}\nrequired group params_arr (LIST) {\n  repeated group list {\n    required binary element (UTF8
);\n  }\n}\nrequired group params_struct {\n  required int64 id;\n required binary name (UTF8);\n }\nrequir
ed group params_arr_complex (LIST) {\n  repeated group list {\n    required group element {\n required int6
4 id;\n required binary name (UTF8);\n}\n  }\n}\nrequired group params_complex (MAP) {\nrepeated group key_
value {\nrequired binary key (UTF8);\nrequired group value {\n  required int64 id;\n required binary name (
UTF8);\n  }\n}\n}\nrequired group params_struct_complex {\n  required int64 id;\n required group detail {\n
  required int64 id;\n required binary name (UTF8);\n  }\n  }\n}",
        "dataxParquetMode": "fields"
      }
    }
```

6.3.14. PostgreSQL Writer

本文为您介绍PostgreSQL Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

PostgreSQL Writer插件实现了向PostgreSQL写入数据。在底层实现上，PostgreSQL Writer通过JDBC连接远程PostgreSQL数据库，并执行相应的SQL语句，将数据写入PostgreSQL。

> **说明** 开始配置PostgreSql Writer插件前，请首先配置好数据源，详情请参见配置PostgreSQL数据源。

- 对于您配置的table、column和where等信息，PostgreSQL Writer将其拼接为SQL语句发送至PostgreSQL数据库。
- 对于您配置的querySql信息，PostgreSQL直接将其发送至PostgreSQL数据库。

注意事项

当PostgreSQL中表名称、字段名称是以数字开头，或者名称中包含大小写英文字母、中划线（-）时需要使用双引号（""）进行转义，不进行转义会导致PostgreSQL Writer插件写入数据至PostgreSQL失败。但是在PostgreSQL Writer插件中，双引号（""）为JSON关键字，因此，您需要使用反斜线（\）再次对双引号（""）进行转义。例如，表名称为 `123Test` ，则转义后表名称为 `\"123Test\"` 。

> **说明**
> - 双引号（""）中，前引号（"）和后引号（"）均需使用反斜线（\）进行转义。
> - 向导模式不支持转义，您需要转换为脚本模式进行转义。

使用脚本模式进行转义的代码示例如下。

```
"parameter": {
    "datasource": "abc",
    "column": [
        "id",
        "\"123Test\"", //添加转义符
],
"where": "",
"splitPk": "id",
"table": "public.wpw_test"
},
```

类型转换列表

PostgreSQL Writer支持大部分PostgreSQL类型，请注意检查您的数据类型。

PostgreSQL Writer针对PostgreSQL的类型转换列表，如下所示。

数据集成内部类型	PostgreSQL数据类型
LONG	BIGINT、BIGSERIAL、INTEGER、SMALLINT和SERIAL
DOUBLE	DOUBLE、PRECISION、MONEY、NUMERIC和REAL
STRING	VARCHAR、CHAR、TEXT、BIT和INET
DATE	DATE、TIME和TIMESTAMP
BOOLEAN	BOOL
BYTES	BYTEA

> **说明**
> - 除上述罗列字段类型外，其它类型均不支持。
> - MONEY、INET和BIT需要您使用 `a_inet::varchar` 类似的语法进行转换。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无

参数	描述	是否必选	默认值
writeMode	选择导入模式，目前支持insert和copy两种方式： • *insert*：执行PostgreSQL的 `insert into...values...` 语句，将数据写入PostgreSQL中。当数据出现主键/唯一性索引冲突时，待同步的数据行写入PostgreSQL失败，当前记录行成为脏数据。建议您优先选择insert模式。 • *copy*：PostgreSQL提供copy命令，用于表与文件（标准输出，标准输入）之间的相互复制。数据集成支持使用 `copy from` 将数据加载到表中。建议您在遇到性能问题时再尝试使用该模式。	否	*insert*
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如 `"column":["id","name","age"]` 。如果要依次写入全部列，使用（*）表示，例如 `"column":["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据集成与PostgreSQL的网络交互次数，并提升整体吞吐量。但是该值设置过大可能会造成数据集成运行进程OOM情况。	否	*1,024*
pgType	PostgreSQL特有类型的转化配置，支持bigint[]、double[]、text[]、Jsonb和JSON类型。配置示例如下。 `{ "job": { "content": [{ "reader": {...}, "writer": { "parameter": { "column": [// 目标表字段列表 "bigint_arr", "double_arr", "text_arr", "jsonb_obj", "json_obj"], "pgType": { // 特殊的类型设置，key为目标表的字段名，value为字段类型。 "bigint_arr": "bigint[]", "double_arr": "double[]", "text_arr": "text[]", "jsonb_obj": "jsonb", "json_obj": "json" } } } }] } }`	否	无

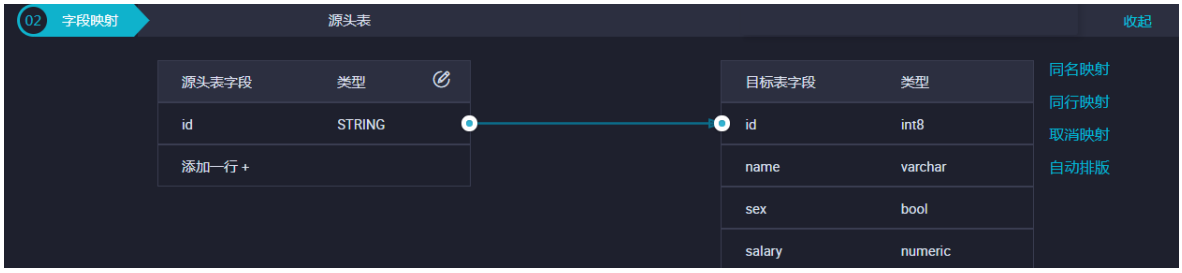
向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
导入模式	即上述参数说明中的writeMode，包括*insert*和*copy*两种模式。

2. 字段映射，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，详情请参见上述参数说明。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"postgresql",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之后率先执行的SQL语句。
                "datasource":"//数据源。
                    "col1",
                    "col2"
                ],
                "table":"",//表名。
                "preSql":[]//执行数据同步任务之前率先执行的SQL语句。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.15. Redis Writer

Redis Writer是基于数据集成框架实现的Redis写入插件，可以通过Redis Writer从数仓或者其它数据源导入数据至Redis。

Redis（REmote DIctionary Server）是一个可以基于内存也可以持久化的日志型、高性能、支持网络的key-value存储系统，可以用作数据库、高速缓存和消息队列代理。Redis支持较丰富的存储value类型，包括String（字符串）、List（链表）、Set（集合）、ZSet（sorted set有序集合）和Hash（哈希类型）。Redis详情请参见redis.io。

Redis Writer与Redis Server之间的交互基于Jedis实现，Jedis是Redis官方首选的Java客户端开发包。

> **说明**
>
> - 当前仅支持向集群化（Cluster）部署的Redis中导入数据，且数据导入过程仅支持使用**独享数据集成资源组**。
> - 开始配置Redis Writer插件前，请首先配置好数据源，详情请参见配置Redis数据源。
> - 使用Redis Writer向Redis写入数据时，如果Value类型是List，重跑同步任务的同步结果不是幂等的。因此，如果Value类型是List，重跑同步任务时，需要您手动清空Redis上相应的数据。

参数说明

参数	描述	是否必选	默认值
expireTime	Redis value值缓存失效时间，单位为秒。如果不填该配置项，则该配置项取值为默认值 `0`，表示永久有效。 expireTime的配置方式有以下两种： • **seconds**方式：指定了从现在开始多长时间后数据失效，取值是失效时间相对当前时间的秒数。 • **unixtime**方式：指定了从1970.1.1开始多长时间后数据失效。取值是失效时间相对1970.1.1时间的秒数。 **说明** 如果expireTime的取值大于60*60*24*30（即失效时间超过30天），则服务端均将expireTime以**unixtime**方式进行配置。	否	*0*（0表示永久有效）
keyFieldDelimiter	写入Redis的Key分隔符。例如key=key1\u0001id，如果有多个Key需要拼接时，该值为必填项。如果只有一个Key，则可以忽略该配置项。	否	*\u0001*
dateFormat	写入Redis时，Date的时间格式为yyyy-MM-dd HH:mm:ss。	否	无
datasource	数据源名称。该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
writeMode	Redis Writer写入Redis的value类型包含以下5种： • 字符串（string） • 字符串列表（list） • 字符串集合（set） • 有序字符串集合（zset） • 哈希（hash） 不同的value类型，writeMode配置会略有差异，详细说明可参见writeMode参数说明。 **说明** 配置Redis Writer时，您需要配置writeMode为支持的5种写入数据类型中的1种类型，且只能配置1种。如果您没有配置，则writeMode取值为默认值 `string`。	否	*string*
keyIndexes	指定作为key的源端列的列序号。列序号从0开始（即第1列的序号是0，第2列的序号是1，依次类推）。 • 源端的某一列作为Redis的key时：配置为对应列的序号即可，例如，第1列作为key，则配置为 `0`。 • 源端的连续多列组合作为Redis的key时：配置为对应多列的序号数组，例如，第2列至第4列组合作为key，则配置为 `[1,3]`。 **说明** 配置keyIndexes后，Redis Writer会将其余的列作为Value。如果您只想同步源表的某几列作为Key，某几列作为Value，则无需同步所有字段，在Reader插件端指定好column进行列筛选即可。	是	无
batchSize	一次性批量提交的记录数大小。该值可以极大减少数据同步系统与Redis的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,000*
timeout	写入Redis的超时时间，单位为毫秒。	否	*30,000*
redisMode	Redis的模式。当前仅支持集群部署模式的Redis，此配置项需配置为ClusterMode。 **说明** 仅支持使用独享数据集成资源组。	否	无

参数	描述	是否必选	默认值
column	写入Redis的column配置。对于Redis对应类型为string，set操作时： • 如果此column没有配置，那么value的格式是分隔符连接的字符串。(csv格式，假设id的值为1，name的值为"小王"，age的值为18，sex的值为男，Redis的value结果示例："18：：男")； • 如果配置了column，且按照如下格式配置，比如"column": [{"index":"0", "name":"id"}, {"index":"1", "name":"content"}]，这样Redis的value数据写出到Redis后，以 {"id":"对应源头列的值","name":"对应源头列的值"} 的JSON形式存储，假设id的值为1，name的值为"小王"，Redis的value结果示例{"id":1,"name":"小王"}"	否	无

writeMode参数说明

配置Redis Writer时，您需要配置writeMode为支持的5种写入数据类型中的1种类型，且只能配置1种。如果您没有配置，则writeMode取值为默认值 `string` 。

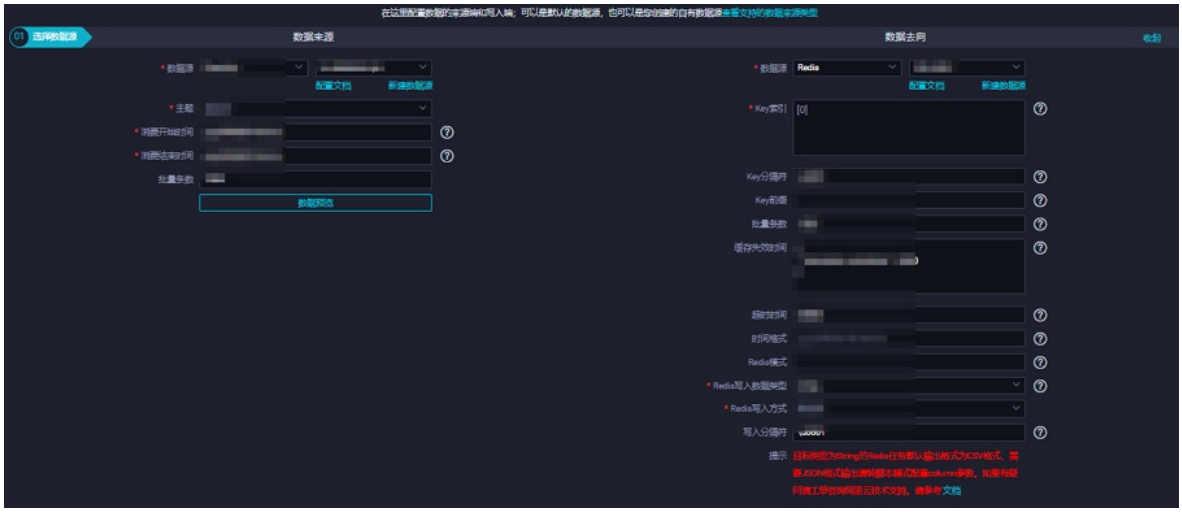
value类型	type参数（必选）	mode参数（必选）	valueFieldDelimiter参数（非必选）	writeMode配置样例
字符串（string）	type需配置为 `string` 。	mode为写入模式参数，value为字符串（string）时： • mode需配置为 `set` 。 • 如果需存储的数据已经存在，则覆盖原有的数据。	valueFieldDelimiter为value之间的分隔符，默认值为 `\u0001` 。 • 该配置项主要用于源数据每行超过两列的情况，例如有三列时，各列通过分隔符分割样例为value1\u0001value2\u0001value3。 • 如果源数据只有两列（即key和value）时，则无需配置。	`"writeMode":{ "type": "string", "mode": "set", "valueFieldDelimiter": "\u0001" }`
字符串列表（list）	type需配置为 `list` 。	mode为写入模式参数，value为字符串列表（list）时，可配置为： • `lpush` ，表示在list最左边存储数据。 • `rpush` ，表示在list最右边存储数据。		`"writeMode":{ "type": "list", "mode": "lpush|rpush", "valueFieldDelimiter": "\u0001" }`
字符串集合（set）	type需配置为 `set` 。	mode为写入模式参数，value为字符串集合（set）时： • mode需配置为 `sadd` ，表示向set集合中存储数据。 • 如果需存储的数据已经存在，则覆盖原有的数据。		`"writeMode":{ "type": "set", "mode": "sadd", "valueFieldDelimiter": "\u0001" }`

value类型	type参数（必选）	mode参数（必选）	valueFieldDelimiter参数（非必选）	writeMode配置样例
有序字符串集合（zset）	type需配置为 `zset` 。	mode为写入模式参数，value为有序字符串集合（zset）时： • mode需配置为 `zadd` ，表示向zset有序集合中存储数据。 • 如果需存储的数据已经存在，则覆盖原有的数据。	无需配置此参数。	`"writeMode":{ "type": "zset", "mode": "zadd" }` **说明** 当value类型为zset时，数据源的每行记录均需遵循相应的规范。即每行记录除key外，只能有1对score和value，并且score必须在value前面，Redis Writer方可解析出column对应的是score或value。
哈希（hash）	type需配置为 `hash` 。	mode为写入模式参数，value为哈希（hash）时： • mode需配置为 `hset` ，表示向hash有序集合中存储数据。 • 如果需存储的数据已经存在，则覆盖原有的数据。	无需配置此参数。	`"writeMode":{ "type": "hash", "mode": "hset" }` **说明** 当value类型为hash时，数据源的每行记录都需遵循相应的规范。即每行记录除key外，只能有1对attribute和value，并且attribute必须在value前面，Redis Writer方可解析出column对应的是attribute或value。

向导开发介绍

1. 选择数据源。

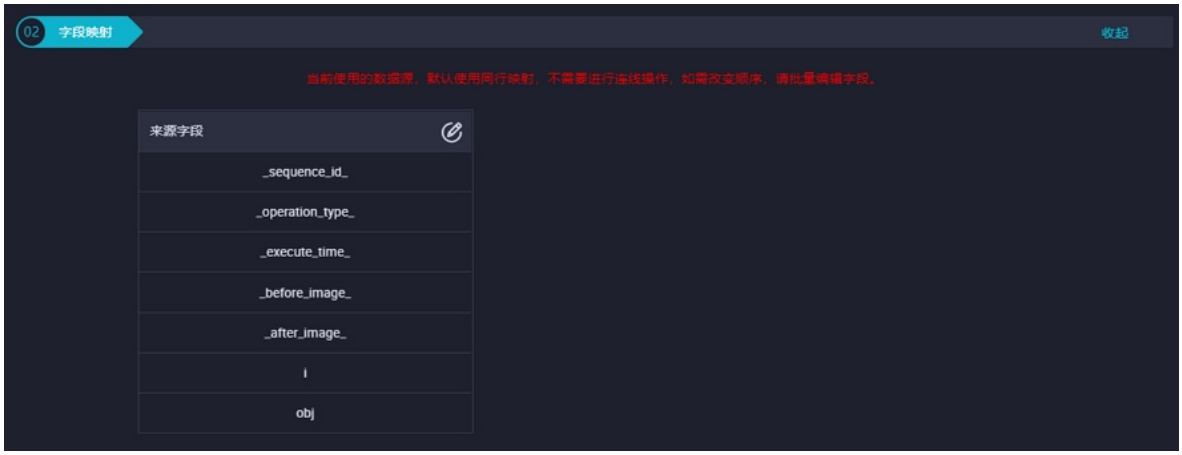
   配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
Key索引	即上述参数说明中的keyIndexes。指定作为key的源端列的列序号，列序号从0开始（即第1列的序号是0，第2列的序号是1，依次类推）。

参数	描述
Key分隔符	即上述参数说明中的keyFieldDelimiter。写入Redis的Key分隔符。
Key前缀	Key的前缀，例如，前缀为 `prefix::` ，key为 `1::2` ，最后生成的key为 `prefix::1::2` 。
批量条数	即上述参数说明中的batchSize。
缓存失效时间	即上述参数说明中的expireTime。
超时时间	即上述参数说明中的timeout。
时间格式	即上述参数说明中的dateFormat。
Redis模式	即上述参数说明中的redisMode。
Redis写入数据类型	即上述参数说明中的writeMode。
Redis写入方式	Redis Writer写入Redis的方式包含：set、lpush、rpush、sadd、zadd、hset。详情请参见：writeMode参数说明。
写入分隔符	即上述参数说明中的**keyFieldDelimiter**。

2. 字段映射，即上述参数说明中的column。默认使用同行映射。您可以单击图标手动编辑目标表字段。



3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

参数	描述
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

以下以读取MySQL数据并写入Redis为例，为您示例读取端MySQL Reader和写入端Redis Writer的脚本代码样例。写入Redis的数据同步作业，具体参数填写请参见参数说明。

> **说明** 通过脚本模式开发的通用流程可参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",  //版本号
    "steps":[
        { //以下为读取端代码样例，读取端的参数详情可查看对应数据源的Reader插件文档。
            "stepType":"mysql",
            "parameter": {
                "envType": 0,
                "datasource": "xc_mysql_demo2",
                "column": [
                    "id",
                    "value",
                    "table"
                ],
                "connection": [
                    {
                        "datasource": "xc_mysql_demo2",
                        "table": []
                    }
                ],
                "where": "",
                "splitPk": "",
                "encoding": "UTF-8"
            },,
            "name":"Reader",
            "category":"reader"
        },
        {//以下为写入端代码样例。
            "stepType":"redis",                     //Redis Writer的插件名，配置为redis。
            "parameter":{                           //以下为Redis Writer的主要参数，各参数的配置详情可参见参数说明。
                "expireTime":{                      //Redis value值缓存失效时间，可配置为seconds类型或unixtime类型。"seconds":"1000"
                            },
                "keyFieldDelimiter":"u0001",        //写入Redis的key的分隔符。
                "dateFormat":"yyyy-MM-dd HH:mm:ss",//写入Redis时，Date的时间格式。
                "datasource":"xc_mysql_demo2",      //数据源名称，需与添加的数据源名称保持一致。
                "envType": 0,                       //环境类型，开发环境：1，生产环境：0。
                "writeMode":{                       //写入模式。
                    "type":"string"                 //value类型。
                    "mode":"set",                   //value是某类型时，写入的模式。
                    "valueFieldDelimiter":"u0001", //value之间的分隔符。
                            },
                "keyIndexes":[0,1],                 //用于源端到Redis的映射，指定源端需要作为key的列（第1列从0开始），如果源端第1列、第2列组合作为Redis的key，这里配置为[0,1]。
                "batchSize":"1000"                  //一次性批量提交的记录数大小。
```

```
        "column": [                          // 对于redis类型为string，set操作，如果此column没有配置那么value的格式是分隔符连接的字符串(csv格式，假设ID的值为1，name的值为"小王"，age的值为18，sex的值为男，redis的value结果示例："18::男")；如果配置了column，且按照如下格式配置，则redis 的value将把原列的列名和值写入成JSON格式，假设id的值为1，name的值为"小王"，age的值为18，sex的值为男，redis的value结果示例{"id":1,"name":"小王","age":18,"sex":"男"}
                {
                "name": "id",
                "index": "0"
                },
                {
                "name": "name",
                "index": "1"
                },
                {
                "name": "age",
                "index": "2"
                },
                {
                "name": "sex",
                "index": "3"
                }
            ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"                            //错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.16. SQL Server Writer

本文为您介绍SQL Server Writer支持的数据类型、字段映射和数据源等参数及配置示例。

SQL Server Writer插件实现了写入数据至SQL Server主库的目标表的功能。在底层实现上， SQL Server Writer通过JDBC连接远程SQL Server数据库，并执行相应的 `insert into` 语句，将数据写入SQL Server，数据库本身会分批次提交数据入库。

> **说明** 开始配置SQL Server Writer插件前，请首先配置好数据源，详情请参见配置SQL Server数据源。

SQL Server Writer面向ETL开发工程师，通过SQL Server Writer从数仓导入数据至SQL Server。同时SQL Server Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

SQL Server Writer通过数据同步框架获取Reader生成的协议数据，通过 `insert into` （当主键/唯一性索引冲突时，冲突的行会写不进去）语句，写入数据至SQL Server。另外出于性能考虑采用了 `PreparedStatement + Batch` ，并且设置了 `rewriteBatchedStatements=true` ，将数据缓冲到线程上下文Buffer中。当Buffer累计到预定阈值时，才发起写入请求。

> 说明
> - 目标表所在数据库必须是主库才能写入数据。
> - 整个任务至少需要具备 `insert into` 的权限，是否需要其它权限，取决于您配置任务时在preSql和postSql中指定的语句。

SQL Server驱动版本支持列表

SQL Server Writer使用驱动版本是com.microsoft.sqlserver sqljdbc4 4.0，驱动能力具体请参见官网文档。该驱动支持的SQL Server版本如下所示：

版本	支持性（是/否）
SQL Server 2016	是
SQL Server 2014	是
SQL Server 2012	是
PDW 2008R2 AU34	是
SQL Server 2008 R2	是
SQL Server 2008	是
SQL Server 2019	否
SQL Server 2018	否

类型转换列表

SQL Server Writer支持大部分SQL Server类型，但也存在个别没有支持的情况，请注意检查您的数据类型。

SQL Server Writer针对SQL Server的类型转换列表，如下所示。

类型分类	SQL Server数据类型
整数类	BIGINT、INT、SMALLINT和TINYINT
浮点类	FLOAT、DECIMAL、REAL和NUMERIC
字符串类	CHAR、NCHAR、NTEXT、NVARCHAR、TEXT、VARCHAR、NVARCHAR（MAX）和VARCHAR（MAX）
日期时间类	DATE、TIME和DATETIME
布尔类	BIT
二进制类	BINARY、VARBINARY、VARBINARY（MAX）和TIMESTAMP

参数说明

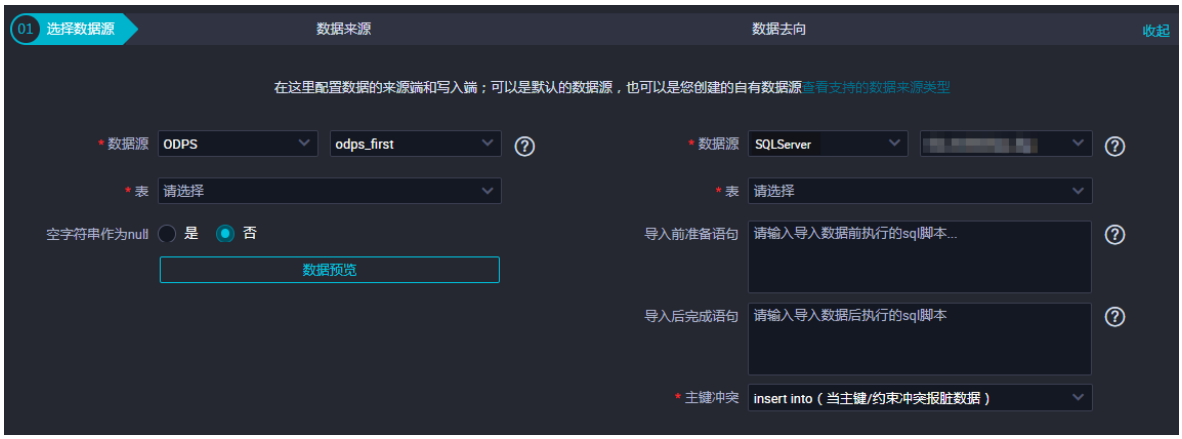
参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如 `"column": ["id","name","age"]` 。如果要依次写入全部列，使用*表示，例如 `"column": ["*"]` 。	是	无

参数	描述	是否必选	默认值
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
writeMode	选择导入模式，可以支持*insert*方式。 当主键/唯一性索引冲突时，数据集成视为脏数据但保留原有的数据。	否	*insert*
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与SQL Server的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

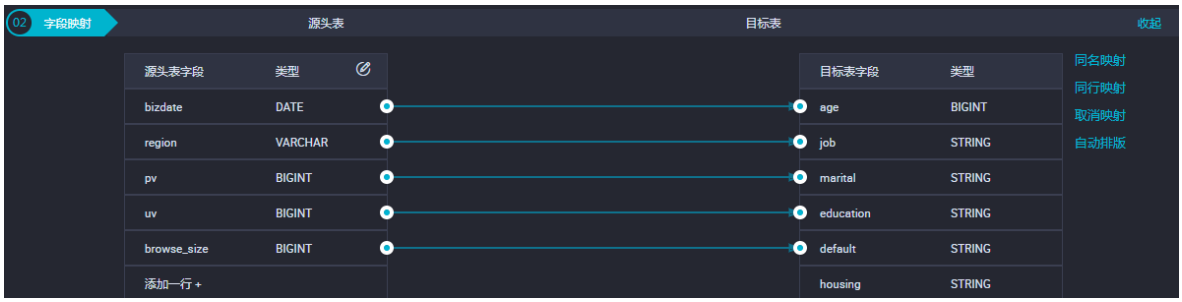
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。

参数	描述
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

配置写入SQL Server的作业，具体参数填写请参见参数说明。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"sqlserver",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之后率先执行的SQL语句。
                "datasource":"",//数据源。
                "column":[//字段。
                    "id",
                    "name"
                ],
                "table":"",//表名。
                "preSql":[]//执行数据同步任务之前率先执行的SQL语句。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.17. Lindorm Writer

本文为您介绍Lindorm Writer的实现原理、参数定义及配置示例。

背景信息

Lindorm Writer插件实现了将数据写入Lindorm的功能。在底层实现上，Lindorm Writer通过Lindorm的Java客户端远程连接Lindorm服务，并支持通过对应的API将数据写入Lindorm的table类型或wideColumn类型的表中。

> 说明
> - LindormWriter的必填配置项configuration，可以通过Lindorm集群控制台查看连接Lindorm相关的配置项进行获取，并以JSON格式填写相关信息。
> - Lindorm为多模数据库，Lindorm Writer支持数据写入table和wideColumn类型的模型表，关于table和wideColumn类型的详细介绍请参见Lindorm使用文档，您也可以通过钉钉咨询Lindorm值班人员。

使用限制

Lindorm Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

类型转换

Lindorm Writer支持大部分Lindorm类型，但也存在个别没有支持的情况（例如，），请注意检查您的数据类型。

Lindorm Writer针对Lindorm类型的转换列表，如下所示。

类型分类	数据类型
整数类	INT、LONG、SHORT
浮点类	DOUBLE、FLOAT、DOUBLE
字符串类	STRING
日期时间类	DATE
布尔类	BOOLEAN
二进制类	BINARYSTRING

参数说明

参数	描述	是否必选	默认值
configuration	每个lindorm集群提供给DataX客户端连接的配置信息，可以通过lindorm集群控制台查询，获取到配置信息后可以联系lindorm数据库管理员将其转换为如下JSON格式：*{"key1":"value1","key2":"value2"}*。 例如：*{"lindorm.zookeeper.quorum":"????","lindorm.zookeeper.property.clientPort":"????"}* > 说明 如果是手写的JSON代码，则需要将双引号转义为\"。	是	无
dynamicColumn	表示dynamic动态列模式，该模式配置较为复杂，一般不使用该模式。可选值包括true和false，默认选择false。	是	false
table	表示所要读取的lindorm表名。lindorm表名对大小写敏感。	是	无
namespace	表示所要读取的lindorm表的命名空间。lindorm表的命名空间对大小写敏感。	是	无
encoding	编码方式，取值为UTF-8或GBK。一般用于将二进制存储的lindorm byte[]类型转换为String类型。	否	UTF-8

参数	描述	是否必选	默认值
columns	读取字段列表。读取字段列表支持列裁剪和列换序，列裁剪指可以选择部分列进行导出，列换序指可以不按照表schema信息顺序进行导出。 • table类型的表，只需要填写列名即可，会自动从表的meta获取schema信息。 • widecolumn类型或table类型的表。	是	无

向导开发介绍

暂不支持向导开发模式。

脚本开发介绍

- 配置一个数据源为MySQL，需要写入数据到Lindorm Table（对应SDK中的TableService模型）的作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
  "type": "job",
  "version": "2.0",
  "steps": [
    {
      "stepType": "mysql",
      "parameter": {
        "checkSlave": true,
        "datasource": " ",
        "envType": 1,
        "column": [
          "id",
          "value",
          "table"
        ],
        "socketTimeout": 3600000,
        "masterSlave": "slave",
        "connection": [
          {
            "datasource": " ",
            "table": []
          }
        ],
        "where": "",
        "splitPk": "",
        "encoding": "UTF-8",
        "print": true
      },
      "name": "mysqlreader",
      "category": "reader"
    },
    {
      "stepType": "lindorm",
      "parameter": {
        "configuration":  {
          "lindorm.client.seedserver": "xxxxxxx:30020",
          "lindorm.client.username": "xxxxxx",
          "lindorm.client.namespace": "default",
          "lindorm.client.password": "xxxxxx"
        },
        "nullMode": "skip",
        "datasource": "",
        "writeMode": "api",
```

```
        "envType": 1,
        "columns": [
          "id",
          "name",
          "age",
          "birthday",
          "gender"
        ],
        "dynamicColumn": "false",
        "table": "lindorm_table",
        "encoding": "utf8",
      },
      "name": "Writer",
      "category": "writer"
    }
  ],
  "setting": {
    "jvmOption": "",
    "executeMode": null,
    "errorLimit": {
      "record": "0"
    },
    "speed": {
      //设置传输速度，单位为byte/s，DataX运行会尽可能达到该速度但是不超过它。
      "byte": 1048576
    },
    //出错限制
    "errorLimit": {
      //出错的record条数上限，当大于该值即报错。
      "record": 0,
      //出错的record百分比上限 1.0表示100%，0.02表示2%。
      "percentage": 0.02
    }
  },
  "order": {
    "hops": [
      {
        "from": "Reader",
        "to": "Writer"
      }
    ]
  }
}
```

- 配置一个数据源为MySQL，需要写入数据到Lindorm wideColumn（对应SDK中的WideColumnService模型）作业，使用脚本开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "mysql",
            "parameter": {
                "envType": 0,
                "datasource": " ",
                "column": [
                     "id",
                    "name",
                    "age",
                    "birthday",
```

```
                    "gender"
                ],
                "connection": [
                    {
                        "datasource": " ",
                        "table": []
                    }
                ],
                "where": "",
                "splitPk": "",
                "encoding": "UTF-8"
            },
            "name": "Reader",
            "category": "reader"
        },
      {
          "stepType": "lindorm",
            "parameter": {
                 "configuration":  {
                  "lindorm.client.seedserver": "xxxxxxx:30020",
                  "lindorm.client.username": "xxxxxx",
                  "lindorm.client.namespace": "default",
                  "lindorm.client.password": "xxxxxx"
                },
            "writeMode": "api",
            "namespace": "default",
            "table": "xxxxxx",
            "encoding": "utf8",
            "nullMode": "skip",
            "dynamicColumn": "false",
            "caching": 128,
            "columns": [
                  "ROW|STRING",
                  "cf:id|STRING",
                  "cf:age|INT",
                  "cf:birthday|STRING"
                ]
              }
        }
    ],
    "setting": {
        "jvmOption": "",
                "errorLimit": {
                        "record": "0"
                },
                "speed": {
                    "concurrent": 3,
                    "throttle": false
                }
    }
    "order": {
            "hops": [
               {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.3.18. Elasticsearch Writer

本文为您介绍Elasticsearch Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

使用限制

DataWorks平台目前仅支持配置阿里云Elasticsearch5.x、6.x、7.x版本数据源，不支持配置自建Elasticsearch数据源。

背景信息

Elasticsearch在公共资源组上支持Elasticsearch5.x版本，在独享数据集成资源组上支持Elasticsearch5.x、6.x和7.x版本。独享数据集成资源组的详情请参见新增和使用独享数据集成资源组。

Elasticsearch是遵从Apache开源条款的一款开源产品，是当前主流的企业级搜索引擎。Elasticsearch是一个基于Lucene的搜索和数据分析工具，它提供分布式服务。Elasticsearch核心概念同数据库核心概念的对应关系如下所示。

```
Relational DB（实例） -> Databases（数据库） -> Tables（表） -> Rows（一行数据） -> Columns（一行数据的一列）
Elasticsearch        -> Index              -> Types       -> Documents       -> Fields
```

Elasticsearch中可以有多个索引或数据库，每个索引可以包括多个类型或表，每个类型可以包括多个文档或行，每个文档可以包括多个字段或列。Elasticsearch Writer插件使用Elasticsearch的Rest API接口，批量把从Reader读入的数据写入Elasticsearch中。

参数说明

参数	描述	是否必选	默认值
endpoint	Elasticsearch的连接地址，通常格式为 `http://example.com:9999` 。	否	无
accessId	Elasticsearch的username，用于与Elasticsearch建立连接时的鉴权。 **说明** AccessID和AccessKey为必填项，如果不填写会产生报错。如果您使用的是自建Elasticsearch，不设置basic验证，则无需账号密码，此处AccessId和AccessKey填写随机值即可。	否	无
accessKey	Elasticsearch的password。	否	无
index	Elasticsearch中的index名。	否	无
indexType	Elasticsearch中index的type名。	否	*Elasticsearch*
cleanup	是否删除所配索引中已有数据，清理数据的方法为删除并重建对应索引，默认值为*false*，表示保留已有索引中的数据。	否	*false*
batchSize	每次批量数据的条数。	否	*1,000*
trySize	失败后重试的次数。	否	*30*
timeout	客户端超时时间。	否	*600,000*
discovery	启用节点发现将轮询并定期更新客户机中的服务器列表。	否	*false*
compression	HTTP请求，开启压缩。	否	*true*
multiThread	HTTP请求，是否有多线程。	否	*true*
ignoreWriteError	忽略写入错误，不重试，继续写入。	否	*false*
ignoreParseError	忽略解析数据格式错误，继续写入。	否	*true*
alias	Elasticsearch的别名类似于数据库的视图机制，为索引my_index创建一个别名my_index_alias，对my_index_alias的操作与my_index的操作一致。 配置alias表示在数据导入完成后，为指定的索引创建别名。	否	无

参数	描述	是否必选	默认值
aliasMode	数据导入完成后增加别名的模式，包括*append*（增加模式）和*exclusive*（只留这一个）： • aliasMode为*append*时，表示追加当前索引至别名alias映射中（一个别名对应多个索引）。 • aliasMode为*exclusive*时，表示首先删除别名alias，再添加当前索引至别名alias映射中（一个别名对应一个索引）。 后续会转换别名为实际的索引名称，别名可以用来进行索引迁移和多个索引的查询统一，并可以用来实现视图的功能。	否	*append*
splitter	如果待插入目标端数据列类型是array数组类型，则使用指定分隔符（-,-），将源头数据进行拆分写出。 例如，源头列是数组 `["a", "b", "c", "d"]` ，使用分隔符（-,-）拆分后是字符串类型数据 `a-,-b-,-c-,-d` ，最终写出至Elasticsearch对应Filed列中。	否	-,-
settings	创建index时的settings，与Elasticsearch官方一致。	否	无
column	column用来配置文档的多个字段Filed信息，具体每个字段项可以配置name（名称）、type（类型）等基础配置，以及Analyzer、Format和Array等扩展配置。 Elasticsearch所支持的字段类型如下所示。 `- id  //type id对应Elasticsearch中的_id，可以理解为唯一主键。写入时，相同id的数据会被覆盖，且不会被索引。` `- string` `- text` `- keyword` `- long` `- integer` `- short` `- byte` `- double` `- float` `- date` `- boolean` `- binary` `- integer_range` `- float_range` `- long_range` `- double_range` `- date_range` `- geo_point` `- geo_shape` `- ip` `- token_count` `- array` `- object` `- nested` 列类型的说明如下： • 列类型为text类型时，可以配置analyzer（分词器）、norms和index_options等参数，示例如下。 `{` `"name": "col_text",` `"type": "text",` `"analyzer": "ik_max_word"` `}`	是	无

参数	描述	是否必选	默认值
	• 列类型为日期Date类型时，可以配置Format或Timezone和origin其中一个参数，分别表示日期序列化格式和时区。 ○ 如果配置origin，插件会按照此配置更新index的mappings，按照原格式写入Elasticsearch，建议都加上origin。 ○ 如果想让数据集成帮助您进行时区转换，需要删除origin，添加Timezone参数。 示例如下。 `{ "name": "col_date", "type": "date", "format": "yyyy-MM-dd HH:mm:ss", "origin": true }` • 列类型为地理形状geo_shape时，可以配置tree（geohash或quadtree）、precision（精度）属性，示例如下。 `{ "name": "col_geo_shape", "type": "geo_shape", "tree": "quadtree", "precision": "10m" }` 如果您在列Filed中配置了array属性，且值为*true*时，则表示数组列。Elasticsearch Writer会使用splitter配置的分隔符（一个任务仅支持配置一种切分分隔符），将对应源端数据进行拆分，转换为字符串数组形式最终写出至目的端，示例如下。 `{ "name": "col_integer_array", "type": "integer", "array": true }`		
dynamic	如果为*true*，则使用Elasticsearch的自动mappings，而非使用数据集成的mappings。 Elasticsearch 7.x版本的默认type为_doc。使用Elasticsearch的自动mappings时，请配置_doc和esVersion为7。 您需要转换为脚本模式，添加一个版本参数：`"esVersion": "7"`。	否	*false*

参数	描述	是否必选	默认值
actionType	表示Elasticsearch在数据写出时的action类型，目前数据集成支持*index*和*update*两种actionType，默认值为*index*： • *index*：底层使用了Elasticsearch SDK的Index.Builder构造批量请求。Elasticsearch *index*插入时，需要首先判断插入的文档数据中是否指定ID： ◦ 如果没有指定ID，Elasticsearch会默认生成一个唯一ID。该情况下会直接添加文档至Elasticsearch中。 ◦ 如果已指定ID，会进行更新（替换整个文档），且不支持针对特定Field进行修改。 **说明** 此处的更新并非Elasticsearch中的更新（替换部分指定列替换）。 • *update*：底层使用了Elasticsearch SDK的Update.Builder构造批量请求。Elasticsearch *update*更新的逻辑为：每次update都会调用InternalEngine中的get方法，来获取整个文档信息，从而实现针对特定字段进行修改。该逻辑导致每次更新都需要获取一遍原始文档，对性能有较大影响，但可以更新用户指定的列。如果匹配的文档不存在，则执行文档插入操作。	否	*index*
other_params	此参数属于高级用法，可以支持一些其他参数，例如，normalizer。	否	无

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，具体参数请参见上文的参数说明。

```
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
            },
            "stepType": "stream"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {
                "endpoint": "http://example.com:9999",
                "accessId": "xxxx",
```

```
"accessKey": "yyyy",
"index": "test-1",
"type": "default",
"cleanup": true,
"settings": {
        "number_of_shards": 1,
        "number_of_replicas": 0
},
"discovery": false,
"batchSize": 1000,
"splitter": ",",
"column": [
    {
        "name": "pk",
        "type": "id"
    },
    {
        "name": "col_ip",
        "type": "ip"
    },
    {
        "name": "col_double",
        "type": "double"
    },
    {
        "name": "col_long",
        "type": "long"
    },
    {
        "name": "col_integer",
        "type": "integer"
    },
    {
        "name": "col_keyword",
        "type": "keyword"
    },
    {
        "name": "col_text",
        "type": "text",
        "analyzer": "ik_max_word"
    },
    {
        "name": "col_geo_point",
        "type": "geo_point"
    },
    {
        "name": "col_date",
        "type": "date",
        "format": "yyyy-MM-dd HH:mm:ss"
    },
    {
        "name": "col_nested1",
        "type": "nested"
    },
    {
        "name": "col_nested2",
        "type": "nested"
    },
    {
        "name": "col_object1",
        "type": "object"
    },
    {
        "name": "col_object2",
        "type": "object"
```

```
                    },
                    {
                        "name": "col_integer_array",
                        "type": "integer",
                        "array": true
                    },
                    {
                        "name": "col_geo_shape",
                        "type": "geo_shape",
                        "tree": "quadtree",
                        "precision": "10m"
                    }
                ]
            },
            "stepType": "elasticsearch"
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

> **说明** VPC环境的Elasticsearch运行在默认资源组会存在网络不通的情况。您需要使用独享数据集成资源或自定义资源，才能连通VPC进行数据同步。添加两种资源的详情请参见独享数据集成资源和新增自定义资源组。

6.3.19. LogHub（SLS） Writer

本文为您介绍LogHub（SLS） Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

LogHub（SLS） Writer使用LogService的Java SDK，可以将数据集成Reader中的数据推送到指定的LogService LogHub（SLS）上，供其它程序消费。

> **说明** 由于LogHub（SLS）无法实现幂等，FailOver重跑任务时会引起数据重复。

LogHub（SLS） Writer通过数据集成框架获取Reader生成的数据，然后将数据集成支持的类型通过逐一判断转换成STRING类型。当达到您指定的batchSize时，会使用LogService Java SDK一次性推送至LogHub（SLS）。默认情况下，一次推送*1,024*条数据，batchSize的最大值为*4,096*。

类型转换列表

LogHub（SLS） Writer针对LogHub（SLS）类型的转换列表，如下所示。

数据集成内部类型	LogHub（SLS）数据类型
LONG	STRING
DOUBLE	STRING
STRING	STRING
DATE	STRING
BOOLEAN	STRING
BYTES	STRING

参数说明

参数	描述	是否必选	默认值

参数	描述	是否必选	默认值
endpoint	日志服务入口endPoint是访问一个项目（Project）及其内部日志数据的URL。它和Project所在的阿里云地域（Region）及Project名称相关。各地域的服务入口请参见：服务入口。	是	无
accessKeyId	访问日志服务的**AccessKeyId**。	是	无
accessKeySecret	访问日志服务的AccessKeySecret。	是	无
project	目标日志服务的项目名称。	是	无
logstore	目标日志库的名称，logstore是日志服务中日志数据的采集、存储和查询单元。	是	无
topic	目标日志服务的topic名称。	否	空字符串
batchSize	LogHub（SLS）一次同步的数据条数，默认1,024条。 **说明** 一次性同步至LogHub（SLS）的数据大小不要超过5M，请根据您的单条数据量大小调整一次性推送的条数。	否	*1,024* 即一次推送1,024条，您可以修改该配置值。
column	每条数据中的column名称。	是	无

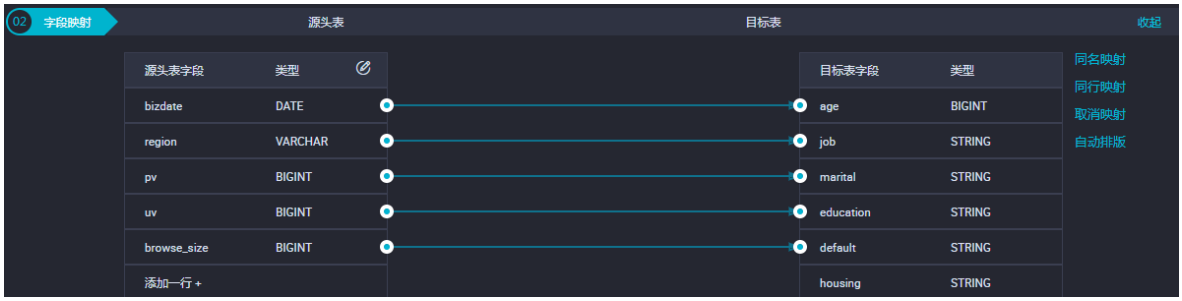
向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	在下拉列表中选择您配置的数据源名称。
Logstore	即上述参数说明中的logstore。
Topic	即上述参数说明中的topic。
批量条数	即上述参数说明中的batchSize。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，具体参数的填写请参见上述的参数说明。

```
{
    "type": "job",
    "version": "2.0",//版本号。
    "steps": [
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "LogHub",//插件名。
            "parameter": {
                "datasource": "",//数据源。
                "column": [//字段。
                    "col0",
                    "col1",
                    "col2",
                    "col3",
                    "col4",
                    "col5"
                ],
                "topic": "",//选取topic。
                "batchSize": "1024",//一次性批量提交的记录数大小。
                "logstore": ""//目标LogService LogStore的名称。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": ""//错误记录数。
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":3, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

6.3.20. OpenSearch Writer

本文为您介绍OpenSearch Writer支持的数据类型、字段映射和数据源等参数及配置示例。

> 注意 OpenSearch Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

实现原理

OpenSearch Writer插件用于向OpenSearch中插入或更新数据。OpenSearch Writer将处理好的数据导入OpenSearch，以搜索的方式输出。

在底层实现上，OpenSearch Writer通过OpenSearch对外提供的开放搜索接口。

> 说明
> - V3版本使用二方包，依赖pom为：com.aliyun.opensearch aliyun-sdk-opensearch 2.1.3。
> - 如果您需要使用OpenSearchWriter插件，请务必使用JDK 1.6-32及以上版本，使用 `java -version` 查看Java版本号。
> - 目前公共资源组不支持连接VPC环境，如果是VPC环境可能会存在网络问题。

插件特点

OpenSearch的列是无序的，因此OpenSearch Writer写入时，需严格按照指定的列的顺序写入。如果指定的列比OpenSearch的列少，则其余列使用默认值或null。

例如，您需要导入的字段列表有b、c两个字段，但OpenSearch表中的字段有a、b、c三列，在列配置中可以写为"column": ["c","b"]，表示会把Reader的第一列和第二列导入OpenSearch的c字段和b字段，而OpenSearch表中新插入的a字段会被置为默认值或null。

补充说明如下：

- 列配置错误的处理

  为保证写入数据的可靠性，避免多余列数据丢失造成数据质量故障。对于写入多余的列，OpenSearch Writer将报错。例如OpenSearch表字段为a、b、c，如果OpenSearch Writer写入的字段多于3列，OpenSearch Writer将报错。

- 表配置注意事项

  OpenSearch Writer一次只能写入一个表。

- 任务重跑和Failover

  重跑后会自动根据ID覆盖。所以插入OpenSearch的列中，必须有一个ID，该ID是OpenSearch的一行记录的唯一标识。唯一标识一样的数据，会被覆盖掉。

类型转换列表

OpenSearch Writer支持大部分OpenSearch类型，请注意检查您的数据类型。

OpenSearch Writer针对OpenSearch类型的转换列表，如下所示。

类型分类	OpenSearch数据类型
整数类	INT
浮点类	DOUBLE和FLOAT
字符串类	TEXT、LITERAL和SHORT_TEXT
日期时间类	INT
布尔类	LITERAL

参数说明

参数	描述	是否必选	默认值
accessId	访问密钥中的AccessKey ID。	是	无
accessKey	访问密钥中的AccessKey Secret，相当于登录密码。	是	无
host	OpenSearch连接的服务地址，您可以在应用详情页面进行查看。	是	无
indexName	OpenSearch项目的名称。	是	无
table	写入数据的表名，不能填写多张表，因为DataX不支持同时导入多张表。	是	无

参数	描述	是否必选	默认值
column	需要导入的字段列表。当导入全部字段时，可以配置为 `"column":["*"]` 。当需要插入部分OpenSearch列时，填写需要插入的列，例如： `"column":["id","name"]` 。 OpenSearch支持列筛选、列换序，例如：表有a、b和c三个字段，只需同步c，b两个字段，则可以配置为 `["c","b"]` 。导入过程中，字段a自动补空，设置为null。	是	无
batchSize	单次写入的数据条数。OpenSearch写入为批量写入，通常OpenSearch的优势在于查询，写入的每秒处理事务数（TPS）不高，请根据账号申请的资源进行设置。 通常OpenSearch的单条数据小于1 MB，单次写入小于2 MB。	如果是分区表，该选项必填。如果是非分区表，该选项不可填写。	*300*
writeMode	OpenSearch Writer通过配置"writeMode":"add/update"，保证写入的幂等性： • "add"：当出现写入失败再次运行时，OpenSearch Writer将清理该条数据，并导入新数据（原子操作）。 • "update"：表示该条插入数据以修改的方式插入（原子操作）。 **说明** OpenSearch的批量插入并非原子操作，有可能会部分成功，部分失败。writeMode参数的选择较为重要，目前V3版本暂不支持update操作。	是	无
ignoreWriteError	忽略写错误。 配置示例： `"ignoreWriteError":true` 。OpenSearch为批量写入，是否忽略当前批次的写失败。如果忽略，则继续执行其它的写操作。如果不忽略，则直接结束当前任务，并返回错误。建议使用默认值。	否	*false*
version	OpenSearch的版本信息，例如 `"version":"v3"` 。由于V2版本对于push操作的限制较多，建议使用V3版本。	否	*v2*

脚本开发介绍

配置写入OpenSearch的数据同步作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {},
        "writer": {
            "plugin": "opensearch",
            "parameter": {
                "accessId": "*********",
                "accessKey": "********",
                "host": "http://yyyy.aliyuncs.com",
                "indexName": "datax_xxx",
                "table": "datax_yyy",
                "column": [
                "appkey",
                "id",
                "title",
                "gmt_create",
                "pic_default"
                ],
                "batchSize": 500,
                "writeMode": add,
                "version":"v2",
                "ignoreWriteError": false
            }
        }
    }
}
```

6.3.21. Table Store（OTS） Writer

本文为您介绍Table Store（OTS） Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

表格存储（Table Store）是构建在阿里云飞天分布式系统之上的NoSQL数据库服务，提供海量结构化数据的存储和实时访问。Table Store以实例和表的形式组织数据，通过数据分片和负载均衡技术，实现规模上的无缝扩展。

Table Store Writer通过Table Store官方Java SDK连接到Table Store服务端，并通过SDK写入Table Store服务端 。Table Store Writer本身对于写入过程进行诸多优化，包括写入超时重试、异常写入重试、批量提交等功能。

目前Table Store Writer支持所有Table Store类型，其针对Table Store类型的转换列表，如下所示。

类型分类	Table Store数据类型
整数类	INTEGER
浮点类	DOUBLE
字符串类	STRING
布尔类	BOOLEAN
二进制类	BINARY

> **说明** 您需要将INTEGER类型的数据，在脚本模式中配置为INT类型，DataWorks会将其转换为INTEGER类型。如果您直接配置为INTEGER类型，日志将会报错，导致任务无法顺利完成。

注意事项

OTS列由主键列primaryKey+普通列column组成，源端列顺序需要和OTS目的端主键列+普通列保持一致，否则会产生列映射错误。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
endPoint	Table Store Server的服务地址。	是	无
accessId	Table Store的AccessKey ID。	是	无
accessKey	Table Store的AccessKey Secret。	是	无
instanceName	Table Store的实例名称。 实例是您使用和管理Table Store服务的实体。开通Table Store服务后，需要通过管理控制台创建实例后，在实例内进行表的创建和管理。实例是Table Store资源管理的基础单元，Table Store对应用程序的访问控制和资源计量都在实例级别完成。	是	无
table	所选取的需要抽取的表名称，此处能且只能填写一张表。在Table Store中不存在多表同步的需求。	是	无
primaryKey	Table Store的主键信息，使用JSON的数组描述字段信息。Table Store本身是NoSQL系统，在Table Store Writer导入数据过程中，必须指定相应的字段名称。 数据同步系统本身支持类型转换的，因此对于源头数据非STRING/INT，Table Store Writer会进行数据类型转换。配置示例如下。 `"primaryKey" : [ {"name":"pk1", "type":"string"}, {"name":"pk2", "type":"int"} ],` **说明** • Table Store的PrimaryKey仅支持STRING和INT类型，不支持BINARY类型。因此Table Store Writer本身也限定填写STRING和INT两种类型。 • primaryKey参数与column参数配置的列名称不能重复。	是	无
column	所配置的表中需要同步的列名集合，使用JSON的数组描述字段信息。 使用格式为： `"column" : [ {"name":"col2", "type":"INT"}, {"name":"col3", "type":"STRING"} ],` 其中的name为写入的Table Store列名称，type为写入的类型。Table Store类型支持STRING、INT、DOUBLE、BOOL和BINARY类型。 **说明** primaryKey参数与column参数配置的列名称不能重复。	是	无

参数	描述	是否必选	默认值
writeMode	writeMode表示数据写入表格存储的格式，目前支持以下两种模式： • *PutRow*：对应于Table Store PutRow API，插入数据到指定的行。如果该行不存在，则新增一行。如果该行存在，则覆盖原有行。 • *UpdateRow*：对应于Table Store UpdateRow API，更新指定行的数据。如果该行不存在，则新增一行。如果该行存在，则根据请求的内容在这一行中新增、修改或者删除指定列的值。	是	无
requestTotalSizeLimitation	该配置限制写入Table Store时单行数据的大小，配置类型为数字。	否	*1MB*
attributeColumnSizeLimitation	该配置限制写入Table Store时单个属性列的大小，配置类型为数字。	否	*2MB*
primaryKeyColumnSizeLimitation	该配置限制写入Table Store时单个主键列的大小，配置类型为数字。	否	*1KB*
attributeColumnMaxCount	该配置限制写入Table Store时属性列的个数，配置类型为数字。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个写入Table Store作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"ots",//插件名。
            "parameter":{
                "datasource":"",//数据源。
                "column":[//字段。
                    {
                        "name":"columnName1",//字段名。
                        "type":"INT"//数据类型。
                    },
                    {
                        "name":"columnName2",
                        "type":"STRING"
                    },
                    {
                        "name":"columnName3",
                        "type":"DOUBLE"
                    },
                    {
                        "name":"columnName4",
                        "type":"BOOL"
                    },
                    {
                        "name":"columnName5",
```

```
                        "type":"BINARY"
                    }
                ],
                "writeMode":"",//写入模式。
                "table":"",//表名。
                "primaryKey":[//Table Store的主键信息。
                    {
                        "name":"pk1",
                        "type":"STRING"
                    },
                    {
                        "name":"pk2",
                        "type":"INT"
                    }
                ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.22. Stream Writer

本文为您介绍Stream Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

Stream Writer插件实现了从Reader端读取数据，并向屏幕上打印数据或直接丢弃数据的功能。该插件主要用于数据同步的性能测试和基本的功能测试。

参数说明

print

- 描述：是否向屏幕打印输出。
- 必选：否。
- 默认值：*true*。

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

配置一个从Reader端读取数据并向屏幕打印的作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"stream",//插件名。
            "parameter":{
                "print":false,//是否向屏幕打印输出。
                "fieldDelimiter":","//列分隔符。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.23. HybridDB for MySQL Writer

本文为您介绍HybridDB for MySQL Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

HybridDB for MySQL Writer插件实现了写入数据至MySQL数据库目标表的功能。在底层实现上，HybridDB for MySQL Writer通过JDBC连接远程HybridDB for MySQL数据库，并执行相应的 `insert into` 或 `replace into` 语句，将数据写入HybridDB for MySQL。数据库本身采用InnoDB引擎，分批次提交数据入库。

> **说明** 开始配置HybridDB for MySQL Writer插件前，请首先配置好数据源，详情请参见配置HybridDB for MySQL数据源。

HybridDB for MySQL Writer面向数据开发工程师，通过HybridDB for MySQL Writer从数仓导入数据至HybridDB for MySQL。同时，HybridDB for MySQL Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。HybridDB for MySQL Writer通过数据同步框架获取Reader生成的协议数据。

> **说明** 整个任务至少需要具备 `insert into` 的权限，是否需要其它权限，取决于您配置任务时在preSql和postSql中指定的语句。

类型转换列表

目前HybridDB for MySQL Writer支持大部分HybridDB for MySQL类型，请注意检查您的数据类型。

HybridDB for MySQL Writer针对HybridDB for MySQL类型的转换列表，如下所示。

类型分类	HybridDB for MySQL数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT、BIGINT和YEAR
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔类	BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

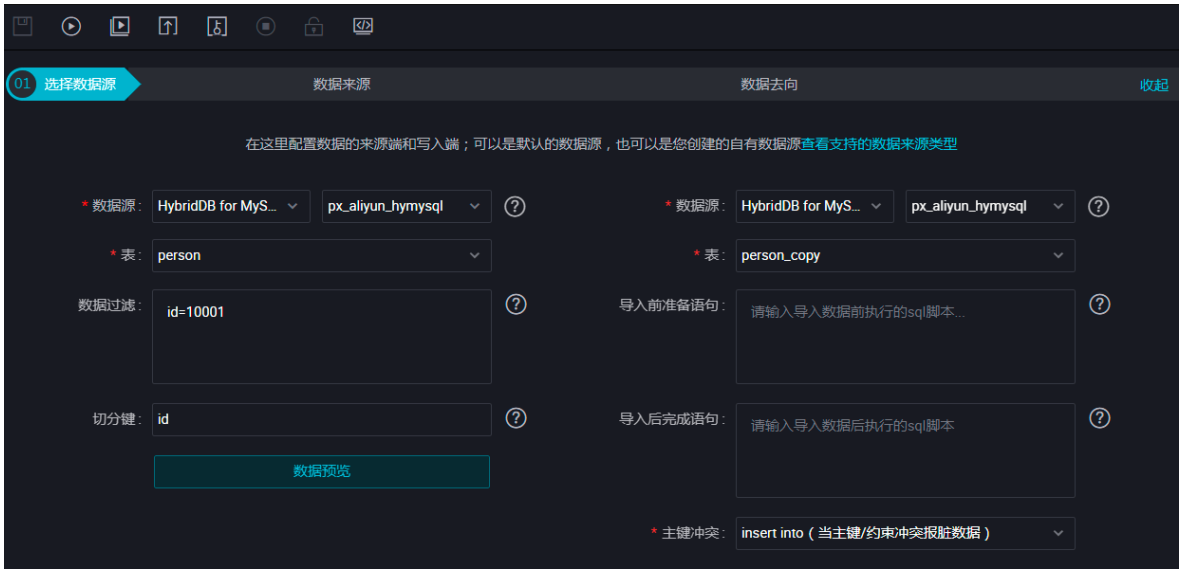
参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
writeMode	选择导入模式，目前支持insert和replace两种方式。 • *replace*：没有遇到主键或唯一性索引冲突时，与*insert*行为一致，冲突时会用新行替换原有行所有字段。 • *insert*：当没有遇到主键或唯一性索引冲突时，数据正常写入。当主键或唯一性索引冲突时会写不进去冲突的行，以脏数据的形式体现。	否	*insert*
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如 `"column":["id","name","age"]` 。如果要依次写入全部列，使用*表示，例如 `"column":["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与HybridDB for MySQL的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

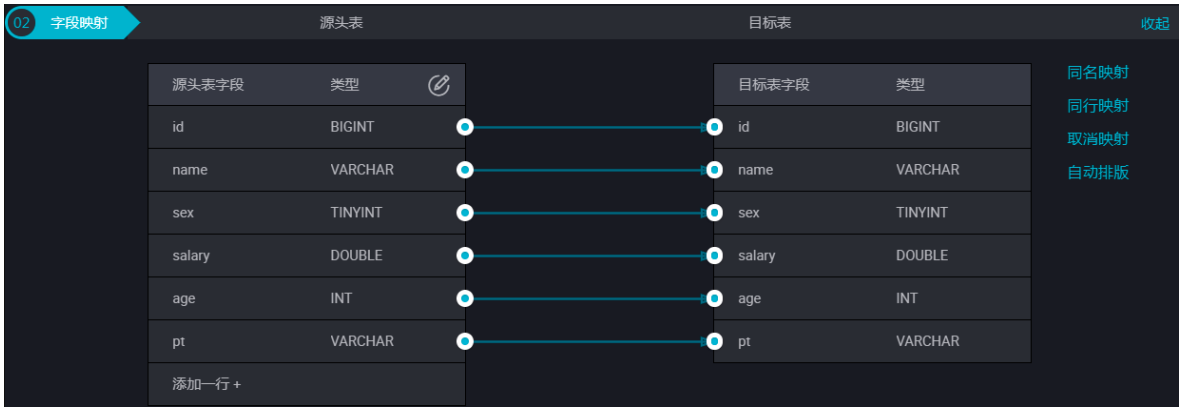
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常选择您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {},
        {
            "parameter": {
                "postSql": [],//导入后的完整语句。
                "datasource": "px_aliyun_hy***",//数据源名。
                "column": [//目标端列名。
                    "id",
                    "name",
                    "sex",
                    "salary",
                    "age",
                    "pt"
                ],
                "writeMode": "insert",//写入模式。
                "batchSize": 256,//一次性批量提交的记录数大小。
                "encoding": "UTF-8",//编码格式。
                "table": "person_copy",//目标表名。
                "preSql": []//导入前的准备语句。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",//版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "concurrent": 7,//并发数。
            "throttle": true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "mbps": 1,//限流值。
        }
    }
}
```

6.3.24. AnalyticDB for PostgreSQL Writer

本文为您介绍AnalyticDB for PostgreSQL Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

AnalyticDB for PostgreSQL Writer插件实现了向AnalyticDB for PostgreSQL写入数据。在底层实现上，AnalyticDB for PostgreSQL Writer通过JDBC连接远程AnalyticDB for PostgreSQL数据库，并执行相应的SQL语句，向AnalyticDB for PostgreSQL库中写入数据。RDS在公共云提供AnalyticDB for PostgreSQL存储引擎。

> **说明** 开始配置AnalyticDB for PostgreSQL Writer插件前，请首先配置好数据源，详情请参见配置AnalyticDB for PostgreSQL数据源。

类型转换列表

AnalyticDB for PostgreSQL Writer支持大部分AnalyticDB for PostgreSQL类型，但也存在部分类型没有支持的情况，请注意检查您的类型。

AnalyticDB for PostgreSQL Writer针对AnalyticDB for PostgreSQL的类型转换列表，如下所示。

类型分类	AnalyticDB for PostgreSQL数据类型
LONG	BIGINT、BIGSERIAL、INTEGER、SMALLINT和SERIAL
DOUBLE	DOUBLE、PRECISION、MONEY、NUMERIC和REAL
STRING	VARCHAR、CHAR、TEXT、BIT和INET
DATE	DATE、TIME和TIMESTAMP
BOOLEAN	BOOL
BYTES	BYTEA

> **说明**
> - 除上述罗列字段类型外，其它类型均不支持。
> - MONEY、INET和BIT需要您使用 `a_inet::varchar` 类似的语法进行转换。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
writeMode	选择导入模式，可以支持insert和copy方式。 • insert：执行PostgreSQL的 `insert into...values...` 语句，将数据写出到PostgreSQL中。当数据出现主键/唯一性索引冲突时，待同步的数据行写入PostgreSQL失败，当前记录行成为脏数据。建议您优先选择insert模式。 • copy：PostgreSQL提供copy命令，用于表与文件（标准输出，标准输入）之间的相互复制。数据集成支持使用 `copy from` ，将数据加载到表中。建议您在遇到性能问题时再尝试使用该模式。	否	*insert*
column	目标表需要写入数据的字段，字段之间用英文逗号分隔。例如 `"column":["id","name","age"]` 。如果要依次写入全部列，使用*表示，例如"column":["*"]。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据集成与AnalyticDB for PostgreSQL的网络交互次数，并提升整体吞吐量。但是该值设置过大可能会造成数据集成运行进程OOM情况。	否	*1,024*

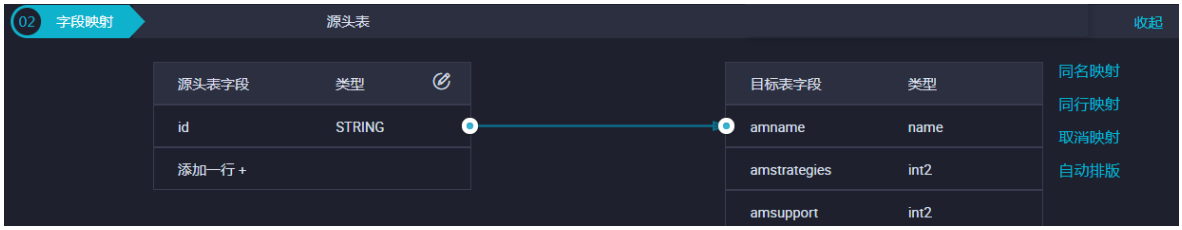
向导开发介绍

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常选择您配置的数据源名称。
表	即上述参数说明中的table，选择需要同步的表。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
导入模式	即上述参数说明中的writeMode，包括**insert**和**copy**两种模式。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "parameter": {
                "postSql": [],//导入后的完成语句。
                "datasource": "test_004",//数据源名。
                "column": [//目标表的列名。
                    "id",
                    "name",
                    "sex",
                    "salary",
                    "age"
                ],
                "table": "public.person",//目标表的表名。
                "preSql": []//导入前的准备语句。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",//版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":6, //作业并发数。
            "mbps":"12"//限流
        }
    }
}
```

6.3.25. PolarDB Writer

本文为您介绍PolarDB Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

PolarDB Writer插件实现了写入数据到PolarDB数据库目标表的功能。在底层实现上，PolarDB Writer通过JDBC连接远程PolarDB数据库，并执行相应的 `insert into…` 或 `replace into…` 的SQL语句将数据写入PolarDB。内部会分批次提交入库，需要数据库本身采用innodb引擎。

> **说明** 在开始配置PolarDB Writer插件前，请首先配置好数据源，详情请参见配置PolarDB数据源。

PolarDB Writer面向ETL开发工程师，使用PolarDB Writer从数仓导入数据到PolarDB。同时PolarDB Writer也可以作为数据迁移工具为DBA等用户提供服务。PolarDB Writer通过数据同步框架获取Reader生成的协议数据，根据您配置的writeMode生成。

> **说明** 整个任务至少需要具备 `insert/replace into…` 的权限，是否需要其它权限，取决于您配置任务时在preSql和postSql中指定的语句。

类型转换列表

类似于PolarDB Reader，目前PolarDB Writer支持大部分PolarDB类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型。

PolarDB Writer针对PolarDB类型的转换列表，如下所示。

类型分类	PolarDB数据类型
整数类	INT、TINYINT、SMALLINT、MEDIUMINT、BIGINT和YEAR
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT和LONGTEXT
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔型	BOOL
二进制类	TINYBLOB、MEDIUMBLOB、BLOB、LONGBLOB和VARBINARY

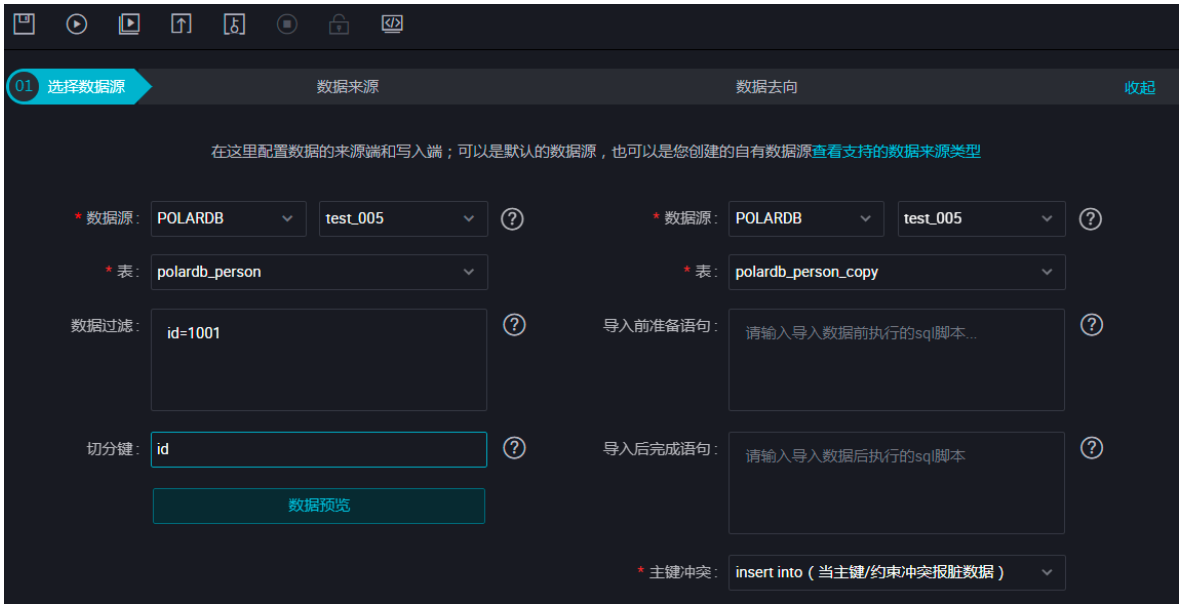
参数说明

参数	描述	必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
writeMode	选择导入模式，可以支持insert，replace和update方式。 • replace：没有遇到主键/唯一性索引冲突时，与insert行为一致，冲突时会用新行替换原有行所有字段。 • insert：当主键或唯一性索引冲突时会写不进去冲突的行，以脏数据的形式体现。没有遇到主键/唯一性索引冲突时，直接插入数据。 • update：没有遇到主键或唯一性索引冲突时，与insert行为一致，冲突时会用新行替换已经指定的字段的语句写入数据到PolarDB。	否	*insert*
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔。例如 `"column": ["id", "name", "age"]` 。如果要依次写入全部列，使用（*）表示。例如 `"column": [" *"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与PolarDB的网络交互次数，并提升整体吞吐量。但是该值设置过大可能会造成数据同步运行进程OOM情况。	否	*1,024*

向导开发介绍

1. 选择数据源

配置同步任务的**数据来源**和**数据去向**。



参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table，选择需要同步的表。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。例如，`update table set gmt_modify=now();`。
主键冲突	即上述参数说明中的writeMode，选择需要的导入模式。

2. **字段映射**，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应的关系。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. **通道控制**

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

4. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下，详情请参见上述参数说明。

```
{
    "type": "job",
    "steps": [
        {
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "parameter": {
                "postSql": [],//导入后完成语句。
                "datasource": "test_005",//数据源名称。
                "column": [//目标列名。
                    "id",
                    "name",
                    "age",
                    "sex",
                    "salary",
                    "interest"
                ],
                "writeMode": "insert",//写入模式。
                "batchSize": 256,//一次性批量提交的记录数大小。
                "encoding": "UTF-8",//编码格式。
                "table": "PolarDB_person_copy",//目标表名。
                "preSql": []//导入前准备语句。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",//版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {//错误记录数。
            "record": ""
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":6, //作业并发数。
            "mbps":"12"//限流
        }
    }
}
```

6.3.26. TSDB Writer

TSDB Writer插件实现了将数据点写入到阿里巴巴云原生多模数据库Lindorm TSDB数据库中，本文为您介绍TSDB Writer支持的数据类型、字段映射和数据源等参数及配置示例。

> 注意 TSDB Writer仅支持使用独享数据集成资源组，不支持使用公共数据集成（调试）资源组和自定义数据集成资源组。

背景信息

时间序列数据库TSDB（Time Series Database）是一种高性能、低成本、稳定可靠的在线时序数据库服务。提供高效读写、高压缩比存储、时序数据插值及聚合计算，广泛应用于物联网（IoT）设备监控系统 、企业能源管理系统（EMS）、生产安全监控系统和电力检测系统等行业场景。

TSDB 提供千万级时序数据秒级写入，高压缩比低成本存储、预降采样、插值、多维聚合计算，查询结果可视化功能；解决由于设备采集点数量巨大，数据采集频率高，造成的存储成本高，写入和查询分析效率低的问题。

> **说明** HiTSDB已更名为云原生多模数据库Lindorm TSDB。Lindorm TSDB兼容大部分HiTSDB的HTTP API并提供原生SQL能力，TSDBWriter插件使用HTTP API方式写入，要使用原生SQL能力需要提前在Lindorm TSDB进行建表。

实现原理

TSDB Writer插件通过TSDB客户端 hitsdb-client 连接TSDB实例，并将数据点通过HTTP API方式写入。关于写入接口，详情请参见TSDB的SDK文档：SDK参考。

使用限制

- TSDB Writer目前仅支持Lindorm TSDB全部版本以及HiTSDB 2.4.x 及以上版本，其他版本暂不保证兼容。
- TSDB Writer仅支持使用独享数据集成资源组，不支持使用公共数据集成（调试）资源组和自定义数据集成资源组。
- TSDB Writer仅支持脚本模式配置任务。

支持的数据类型

当sourceDbType为TSDB，即源端数据源为TSDB Reader或者OpenTSDB Reader时，插件会将源端数据按照JSON字符串格式直接写入；当sourceDbType为RDB，即源端数据源为关系型数据库，插件会按照关系型数据库的Record进行解析，以下内容为您介绍，当sourceDbType为RDB时，columnType的配置及与其对应位置的column可写入的数据类型。

数据模型	columnType配置类型	数据类型
数据标签	tag	字符串类型。Tag描述数据源的特征，通常不随时间变化
数据产生时间	timestamp	时间戳类型。Timestamp代表数据产生的时间点，可以写入时指定，也可由系统自动生成
数据内容	field_string	该Field的value是字符串类型。Field描述数据源的量测指标，通常随着时间不断变化，
	field_double	该Field的value是数值类型。Field描述数据源的量测指标，通常随着时间不断变化，
	field_boolean	该Field的value是布尔类型。Field描述数据源的量测指标，通常随着时间不断变化，

参数说明

数据源	参数	描述	是否必选	默认值
公共参数	sourceDbType	数据源的类型。	否	*TSDB* > **说明** 目前支持TSDB和RDB两个取值。其中，TSDB包括OpenTSDB、Prometheus和TimeScale。RDB包括MySQL、Oracle、PostgreSQL、DRDS等。
	endpoint	TSDB 的 HTTP 连接访问地址，请登录产品管理控制台获取。	是，格式为 http://IP:Port。	无

数据源	参数	描述	是否必选	默认值
	database	导入的TSDB数据库名。	否	*default* **说明** TSDB需要提前创建数据库。
	username	TSDB数据库用户名， TSDB配置了鉴权时需要填写。	否	无
	batchSize	每次批量写入数据的条数。设置过大的batchSize需要更多的任务内存。	否，数据类型为INT，需要确保大于0。	100
数据源为TSDB	maxRetryTime	失败后重试的次数。	否，数据类型为INT，需要确保大于1。	*3*
	ignoreWriteError	如果设置为*true*，则忽略写入错误，继续写入。如果多次重试后仍写入失败，则终止写入任务。	否，数据类型为BOOL。	*false*
数据源为RDB	table	要导入TSDB的表名（metric）。如果multiField为false则不需要填写，对应的metric需要写到column字段	当multiField为true时必选。	无
	multiField	使用HTTP API多值（多个field）方式写入TSDB。 **说明** 如果使用Lindorm TSDB原生SQL能力访问HTTP API方式写入的数据，需要在TSDB进行预建表，否则只能使用HiTSDB HTTP API方式查询数据，详情请参见多值数据查询。	必选。	*false* **说明** 目前TSDB版本使用多值写入时，此值需要指定为true。
	column	关系型数据库中表的字段名。	是	无 **说明** 此处的字段顺序，需要和Reader插件中配置的column字段的顺序保持一致。
	columnType	关系型数据库中表字段，映射到TSDB中的类型。支持的类型如下所示： • timestamp：该字段为时间戳。 • tag：该字段为tag。 • field_string：该Field的value是字符串类型 • field_double：该Field的value是数值类型。 • field_boolean：该Field的value是布尔类型。	是	无 **说明** 此处的字段顺序，需要和Reader插件中配置的column字段的顺序保持一致。

数据源	参数	描述	是否必选	默认值
	batchSize	每次批量写入数据的条数。	否，数据类型为INT，需要确保大于0。	*100*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

- RDB到TSDB的默认配置（推荐）

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "stream",//您可将stream插件替换为具体的RDB类型插件json，RDB类型数据库包括MySQL、Oracle、PostgreSQL、DRDS等。
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "tsdb",
            "parameter": {
                "endpoint": "http://localhost:8242",
                "username": "xxx",
                "password": "xxx",
                "sourceDbType": "RDB",
                "batchSize": 256,
                "columnType": [
                    "tag",
                    "tag",
                    "field_string",
                    "field_double",
                    "timestamp",
                    "field_bool"
                ],
                "column": [
                    "tag1",
                    "tag2",
                    "field1",
                    "field2",
                    "timestamp",
                    "field3"
                ],
                "multiField": "true",
                "table": "testmetric",
                "ignoreWriteError": "false",
                "database": "default"
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

- 从支持OpenTSDB 协议的数据库同步抽取数据到TSDB:

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "opentsdb",
            "parameter": {
                "endpoint": "http://localhost:4242",
                "column": [
                    "m1",
                    "m2",
                    "m3",
                    "m4",
                    "m5",
                    "m6"
                ],
                "startTime": "2019-01-01 00:00:00",
                "endTime": "2019-01-01 03:00:00"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "tsdb",
            "parameter": {
                "endpoint": "http://localhost:8242"
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

- 使用 OpenTSDB（单值）协议写入TSDB（不推荐）：

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "stream",//您可将stream插件替换为具体的RDB类型插件json，RDB类型数据库包括MySQL、Oracle、PostgreSQL、DRDS等。
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "tsdb",
            "parameter": {
                "endpoint": "http://localhost:8242",
                "username": "xxx",
                "password": "xxx",
                "sourceDbType": "RDB",
                "batchSize": 256,
                "columnType": [
                    "tag",
                    "tag",
                    "field_string",
                    "field_double",
                    "timestamp",
                    "field_boolean"
                ],
                "column": [
                    "tag1",
                    "tag2",
                    "field_metric_1",
                    "field_metric_2",
                    "timestamp",
                    "field_metric_3"
                ],
                "ignoreWriteError": "false"
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

> **说明** 转换到的TSDB表名（metric）由column中field对应的列名决定，对于上诉配置一行关系型数据将会写入三个metric（field_metric_1，field_metric_2，field_metric_3）。

性能报告

- 性能数据特征
  - Metric：指定一个Metric为m。
  - tagkv：前4个tagkv全排列，形成 `10*20*100*100=2,000,000` 条时间线，最后IP对应2,000,000条时间线，从1开始自增。

tag_k	tag_v
zone	z1~z10
cluster	c1~c20
group	g1~100
app	a1~a100
ip	ip1~ip2,000,000

  - value：度量值为[1, 100]区间内的随机值。
  - interval：采集周期为10秒，持续摄入3小时，总数据量为 `3*60*60/10*2,000,000=2,160,000,000` 个数据点。
- 性能测试结果

通道数	数据集成速度（Rec/s）	数据集成流量（MB/s）
1	129,753	15.45
2	284,953	33.70
3	385,868	45.71

6.3.27. AnalyticDB for MySQL 3.0 Writer

本文为您介绍AnalyticDB for MySQL 3.0 Writer支持的数据类型、字段映射和数据源等参数及配置示例。

开始配置AnalyticDB for MySQL 3.0 Writer插件前，请先配置好数据源，详情请参见配置AnalyticDB for MySQL 3.0数据源。

类型转换列表

AnalyticDB for MySQL 3.0 Writer针对AnalyticDB for MySQL 3.0类型的转换列表，如下所示。

类型	AnalyticDB for MySQL 3.0数据类型
整数类	INT、INTEGER、TINYINT、SMALLINT和BIGINT
浮点类	FLOAT、DOUBLE和DECIMAL
字符串类	VARCHAR
日期时间类	DATE、DATETIME、TIMESTAMP和TIME
布尔类	BOOLEAN

参数说明

参数	描述	是否必选	默认值

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
writeMode	选择导入模式，可以支持*insert* 和*replace* 两种方式。 • *insert*：当没有遇到主键/唯一性索引冲突时，数据正常写入。当主键/唯一性索引冲突时会写不进去冲突的行，以脏数据的形式体现。 • *replace*：没有遇到主键/唯一性索引冲突时，与 `insert` 行为一致。冲突时会先删除原有行，再插入新行。即新行会替换原有行的所有字段。	否	*insert*
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔，例如 `"column": ["id", "name", "age"]` 。如果要依次写入全部列，使用*表示， 例如 `"column": ["*"]` 。 **说明** 如果字段名中包含select，请在字段名前后加上反引号。例如，item_select_no需要写为`item_select_no`。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如清除旧数据。 **说明** 当有多条SQL语句时，不支持事务。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句，例如加上某一个时间戳。 **说明** 当有多条SQL语句时，不支持事务。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与MySQL的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

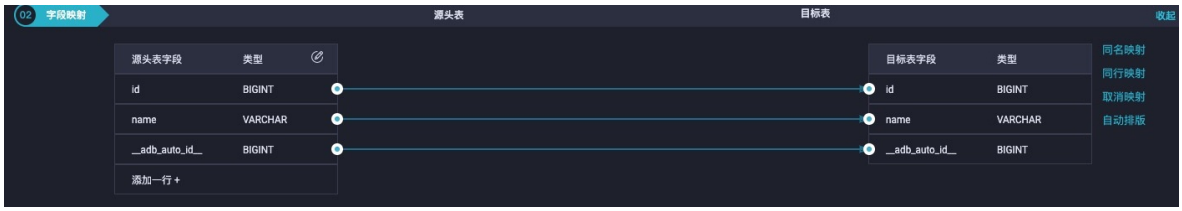
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。

参数	描述
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
主键冲突	即上述参数说明中的writeMode，可以选择需要的导入模式。
批量插入条数	即上述参数说明中的batchSize，提交数据写的批量条数，当wirteMode为insert时，该值才会生效。

2. 字段映射，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应的关系。



3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置示例如下，详情请参见上述参数说明。

> **注意** 实际运行时，请删除下述代码中的注释。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "stream",
            "parameter": {},
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "analyticdb_for_mysql", //插件名。
            "parameter": {
                "postSql": [], //导入后的准备语句。
                "tableType": null, //保留字段，默认空。
                "datasource": "hangzhou_ads", //数据源名称。
                "column": [ //同步字段。
                    "id",
                    "value"
                ],
                "guid": null,
                "writeMode": "insert", //写入模式，请参见writeMode参数说明。
                "batchSize": 2048, //批量写入的大小，请参见batchSize参数说明。
                "encoding": "UTF-8", //编码格式。
                "table": "t5", //写入的表名。
                "preSql": [] //导入前的准备语句。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0", //配置文件格式的版本号。
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0" //错误记录数。
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
                        "concurrent":2, //作业并发数。
                        "mbps":"12"//限流
        }
    }
}
```

6.3.28. GDB Writer

本文为您介绍GDB Writer支持的数据类型、字段映射和数据源等参数及配置示例。

图数据库（Graph Database，简称GDB）是一种支持属性图模型，用于处理高度连接数据查询与存储的实时可靠的在线数据库，支持TinkerPop Gremlin查询语言，可以帮您快速构建基于高度连接的数据集的应用程序。

> 注意
> - GDB Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。
> - 由于点和边的数据集成任务的配置不同，请您分别配置点和边的数据集成任务。

使用限制

- 必须先运行点的同步任务，运行成功后，才可以运行边的同步任务。
- 点有以下约束规则：
  - 点必须具备类型名（即点名称，对应label）。
  - 点的主键ID为必选，必须保证在点范围内唯一，且类型必须是STRING（如果不是STRING类型，GDB Writer插件会强制转换）。
  - 请谨慎选择点的主键映射规则idTransRule。如果选择*none*，需要保证点的ID在全局点的范围内唯一。
- 边有以下约束规则：
  - 边必须具备类型名（即边名称，对应label）。
  - 边的主键ID为可选。
    - 如果填写，则需要保证在全局边范围内唯一。
    - 如果不填写，则GDB服务端默认生成一个UUID，类型必须是STRING（如果不是STRING类型，GDB Writer插件会强制转换）。
  - 请谨慎选择边的主键映射规则idTransRule。如果选择*none*，需要保证边的ID在全局点边的范围内唯一。
  - 边必须选择srcIdTransRule和dstIdTransRule，且必须和导入点时选择的idTransRule一致。
- 示例的字段名或枚举值，如果没有特殊说明，均为大小写敏感。
- 目前GDB服务端仅支持UTF-8编码格式，要求来源数据均为UTF-8编码格式。
- 由于网络限制，运行数据集成任务时，只能使用独享资源组模式，请您提前购买并绑定GDB实例所在的专有网络（VPC）。调度任务可以使用公共资源组。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
label	类型名，即点/边名称。 label支持从源列中读取，例如${0}，表示取第1列字段作为label名，源列索引从0开始。	是	无
labelType	label的类型： • 枚举值VERTEX表示点。 • 枚举值EDGE表示边。	是	无
srcLabel	• 当label为边时，表示起点的点名称。 label为边、srcIdTransRule为*none*时，可以不填写，否则为必填项。 • label为点时，则不填。	否	无
dstLabel	• 当label为边时，表示终点的点名称。 label为边、dstIdTransRule为*none*时，可以不填写，否则为必填项。 • label为点时，则不填。	否	无

参数	描述	是否必选	默认值
writeMode	导入ID重复时的处理模式。 • 枚举值*INSERT*表示会报错，错误记录数加1。 • 枚举值*MERGE*表示用新值覆盖旧值。	是	*INSERT*
idTransRule	主键ID的转换规则。 • 枚举值*labelPrefix*表示将映射的值转换为`{label名}-{源字段}`。 • 枚举值*none*表示映射的值不进行转换。	是	*none*
srcIdTransRule	当label为边时，表示起点的主键ID转换规则。 • 枚举值*labelPrefix*表示将映射的值转换为`{label名}-{源字段}`。 • 枚举值*none*表示映射的值不进行转换，此时可以不填写srcLabel。	label为边时必选	*none*
dstIdTransRule	当label为边时，表示终点的主键ID转换规则。 • 枚举值*labelPrefix*表示将映射的值转换为`{label名}-{源字段}`。 • 枚举值*none*表示映射的值不进行转换，此时可以不填写dstLabel。	label为边时必选	*none*

参数	描述	是否必选	默认值
column	点/边字段映射关系配置。 • name：点/边的字段名。 • value：点/边字段映射的值，仅脚本模式支持字符串自定义拼接。 ○ *${N}*表示直接映射源端值，N为源端column索引，从0开始。 ○ *${0}*表示映射源端column第1个字段。 ○ *test-${0}*表示源端值进行拼接转换，*${0}*值前/后可以添加固定字符串。 ○ *${0}-${1}*表示进行多字段拼接，也可以在任意位置添加固定字符串，例如 `test-${0}-test1-${1}-test2` 。 • *type*：点/边字段映射值的类型。 主键ID仅支持STRING类型，GDB Writer会进行强制转换，源ID必须保证可以转换为STRING类型。 普通属性支持类型：INT、LONG、FLOAT、DOUBLE、BOOLEAN和STRING。 • *columnType*：点/边映射字段的分类，支持的枚举值如下所示。 ○ 公共枚举值 *primaryKey*：label为点/边时，表示该字段是主键ID。 ○ 点枚举值 ■ *vertexProperty*：label为点时，表示该字段是点的普通属性。 ■ *vertexJsonProperty*：label为点时，表示是点JSON属性，value结构请参见properties示例。 ○ 边枚举值 ■ *srcPrimaryKey*：label为边时，表示该字段是起点主键ID。 ■ *dstPrimaryKey*：label为边时，表示该字段是终点主键ID。 ■ *edgeProperty*：label为边时，表示该字段是边的普通属性。 ■ *edgeJsonProperty*：label为边时，表示是边JSON属性，value结构请参见properties示例。 properties示例 `{"properties":[ {"k":"name","t":"string","v":"tom"}, {"k":"age","t":"int","v":"20"}, {"k":"sex","t":"string","v":"male"} ]}`	是	*无*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个写入GDB的数据同步作业，通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务：

- 点配置示例

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
```

```
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"100"  //错误记录数，表示脏数据最大容忍条数。
        },
        "jvmOption":"",
        "speed":{
             "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":3, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "column":[
                    "*"
                ],
                "datasource":"_ODPS",
                "emptyAsNull":true,
                "guid":"",
                "isCompress":false,
                "partition":[],
                "table":""
            },
            "stepType":"odps"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter": {
                "datasource": "testGDB", // 数据源名称。
                "label": "person", //label名，即点名称。
                "srcLabel": "", // 点类型时此字段无需关注。
                "dstLabel": "", // 点类型时此字段无需关注。
                "labelType": "VERTEX", //label类型，"VERTEX"表示点。
                "writeMode": "INSERT", //导入ID重复时处理方式。
                "idTransRule": "labelPrefix", //点的主键转换规则。
                "srcIdTransRule": "none", // 点类型时此字段无需关注。
                "dstIdTransRule": "none", // 点类型时此字段无需关注。
                "column": [
                    {
                        "name": "id", //字段名。
                        "value": "${0}", //${0}表示取源端第1个字段值，支持拼接，0是源端column索引号。
                        "type": "string", //字段类型。
                        "columnType": "primaryKey" //字段分类，primaryKey表示是主键。
                    }, //点的主键，字段名必须是ID且类型是STRING，该记录必须存在。
                    {
                        "name": "person_age",
                        "value": "${1}", //${1}表示取源端第2个字段值，同上支持拼接。
                        "type": "int",
                        "columnType": "vertexProperty" //字段分类，vertexProperty表示是点的属性。
                    }, //点的属性，支持INT、LONG、FLOAT、DOUBLE、BOOLEAN和STRING类型。
                    {
                        "name": "person_credit",
                        "value": "${2}", //${2}表示取源端第3个字段值，同上支持拼接。
                        "type": "string",
                        "columnType": "vertexProperty"
                    }, //点的属性。
                ]
            }
```

```
            "stepType":"gdb"
        }
    ],
    "type":"job",
    "version":"2.0"
}
```

- 边配置示例

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"100" //错误记录数，表示脏数据的最大容忍条数。
        },
        "jvmOption":"",
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":3, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "steps":[
        {
            "category":"reader",
            "name":"Reader",
            "parameter":{
                "column":[
                    "*"
                ],
                "datasource":"_ODPS",
                "emptyAsNull":true,
                "guid":"",
                "isCompress":false,
                "partition":[],
                "table":""
            },
            "stepType":"odps"
        },
        {
            "category":"writer",
            "name":"Writer",
            "parameter": {
                "datasource": "testGDB", // 数据源名称。
                "label": "use", //label名，即边名称。
                "labelType": "EDGE", //label类型，EDGE表示边。
                "srcLabel": "person", //起点的点名称。
                "dstLabel": "software", //终点的点名称。
                "writeMode": "INSERT", //导入ID重复时的处理方式。
                "idTransRule": "labelPrefix", //边的主键转换规则。
                "srcIdTransRule": "labelPrefix", //起点的主键转换规则。
                "dstIdTransRule": "labelPrefix", //终点的主键转换规则。
                "column": [
                    {
                        "name": "id", //字段名。
                        "value": "${0}", //${0}表示取源端第1个字段值，支持拼接。
                        "type": "string", //字段类型。
                        "columnType": "primaryKey" //字段分类，primaryKey表示该字段是主键。
```

```
                    }, //边的主键，字段名必须是ID且类型是STRING，该记录选填。
                    {
                        "name": "id",
                        "value": "${1}", //支持拼接，注意映射规则要与录入点时一致。
                        "type": "string",
                        "columnType": "srcPrimaryKey" //字段分类，srcPrimaryKey表示是起点主键。
                    }, //起点的主键，字段名必须是ID且类型是STRING，该记录必须存在。
                    {
                        "name": "id",
                        "value": "${2}", //支持拼接，注意映射规则要与录入点时一致。
                        "type": "string",
                        "columnType": "dstPrimaryKey" //字段分类，dstPrimaryKey表示是终点主键。
                    }, //终点的主键，字段名必须是ID且类型是STRING，该记录必须存在。
                    {
                        "name": "person_use_software_time",
                        "value": "${3}", //支持拼接。
                        "type": "long",
                        "columnType": "edgeProperty" //字段分类，edgeProperty表示边的属性。
                    }, //边的属性，支持INT、LONG、FLOAT、DOUBLE、BOOLEAN和STRING类型。
                    {
                        "name": "person_regist_software_name",
                        "value": "${4}", //支持拼接。
                        "type": "string",
                        "columnType": "edgeProperty"
                    }, //边属性
                    {
                        "name": "id",
                        "value": "${5}", //支持拼接。
                        "type": "long",
                        "columnType": "edgeProperty"
                    }, //边的属性，字段名是ID。与主键ID不同，该字段为普通属性，可选。
                ]
            }
            "stepType":"gdb"
        }
    ],
    "type":"job",
    "version":"2.0"
}
```

6.3.29. MaxCompute Writer

本文为您介绍MaxCompute Writer支持的数据类型、字段映射和数据源等参数及配置示例。

前提条件

开始配置MaxCompute Writer插件前，请首先配置好数据源，详情请参见配置MaxCompute数据源。

背景信息

MaxCompute Writer插件用于实现向MaxCompute中插入或更新数据，主要适用于开发者将业务数据导入MaxCompute，适合于TB、GB等数量级的数据传输。MaxCompute的详情请参见什么是MaxCompute。

在底层实现上，您可以根据配置的源头项目、表、分区、字段等信息，通过Tunnel写入数据至MaxCompute。常用的Tunnel命令请参见Tunnel命令。

对于MySQL、MaxCompute等强Schema类型的存储，数据集成会逐步读取源数据至内存中，并根据目的端数据源的类型，将源头数据转换为目的端对应的格式，写入目的端存储。

如果数据转换失败，或数据写出至目的端数据源失败，则将数据作为脏数据，您可以配合脏数据限制阈值使用。

> **说明** 当数据有null值时，MaxCompute Writer不支持VARCHAR类型。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，该配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	写入的数据表的表名称（大小写不敏感），不支持填写多张表。	是	无
partition	需要写入数据表的分区信息，必须指定到最后一级分区。例如把数据写入一个三级分区表，必须配置到最后一级分区，例如 `pt=20150101, type=1, biz=2`： • 对于非分区表，该值务必不要填写，表示直接导入至目标表。 • MaxCompute Writer不支持数据路由写入，对于分区表请务必保证写入数据到最后一级分区。	如果表为分区表，则必填。如果表为非分区表，则不能填写。	无
column	需要导入的字段列表。当导入全部字段时，可以配置为 `"column": ["*"]` 。当需要插入部分MaxCompute列，则填写部分列，例如 `"column": ["id","name"]`： • MaxCompute Writer支持列筛选、列换序。例如一张表中有a、b和c三个字段，您只同步c和b两个字段，则可以配置为 `"column": ["c","b"]` ，在导入过程中，字段a自动补空，设置为null。 • column必须显示指定同步的列集合，不允许为空。	是	无
truncate	通过配置 `"truncate": "true"` 保证写入的幂等性。即当出现写入失败再次运行时，MaxCompute Writer将清理前述数据，并导入新数据，可以保证每次重跑之后的数据都保持一致 。 因为利用MaxCompute SQL进行数据清理工作，SQL无法保证原子性，所以truncate选项不是原子操作。当多个任务同时向一个Table或Partition清理分区时，可能出现并发时序问题，请务必注意。 针对该类问题，建议您尽量不要多个作业DDL同时操作同一个分区，或者在多个并发作业启动前，提前创建分区。	是	无

向导开发介绍

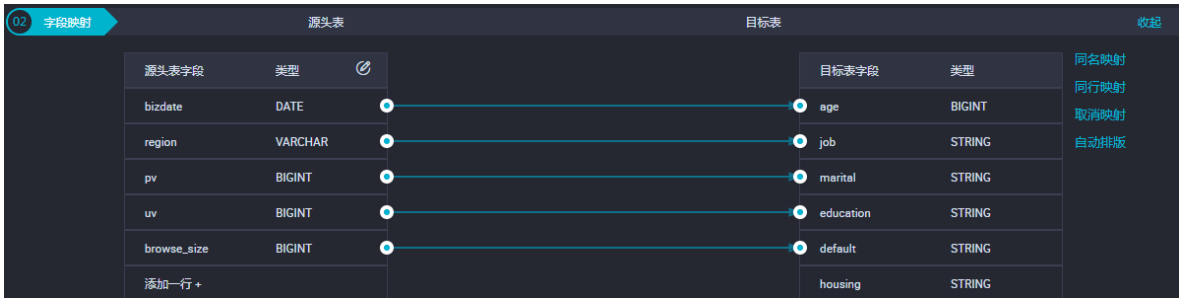
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。如果为分区表，需要配置写入的分区，分区配置可结合调度参数使用，调度参数使用功能详情可参考文档：调度参数概述。
清理规则	清理规则包括： ○ **写入前清理已有数据**：导数据之前，清空表或者分区的所有数据，相当于 `insert overwrite` 。 ○ **写入前保留已有数据**：导数据之前，不清理任何数据，每次运行数据均为追加的数据，相当于 `insert into` 。 **说明** ○ MaxCompute通过Tunnel服务读取数据，同步任务本身不支持数据过滤，需要读取某一个表或分区内的数据。 ○ MaxCompute通过Tunnel服务写出数据，未使用MaxCompute的Insert SQL语句写出数据。数据同步任务执行成功后，完整数据才可以对表可见。请注意配置任务的依赖关系。
空字符串作为null	设置是否允许空字符串作为null。
同步完成才可见	从数据写入开始到数据写入完成是一完整过程。您可以选择是否开启同步完成可见。 **说明** 开启了**分布式处理能力**的任务，该功能的配置将不生效。 ○ 当选择**是**：仅当同步任务执行完成，新同步到MaxCompute的数据才能被查询到。 **说明** 同步完成可见可支持的最大数据量为1TB，当数据存在严重倾斜时无法保证达到此同步上限。 ○ 当选择**否**：同步任务运行过程中，即有部分新同步到MaxCompute的数据可以被查询到。 **说明** 分区自动创建、数据清理动作为数据同步的前置操作，不包含在此写入过程中，因此无论是否开启同步完成可见，您都可能在同步任务运行过程中看到自动创建的分区，或已经完成数据清理的表或分区。我们不建议您使用分区是否存在、数据条数是否为0作为数据同步任务完成的依据。如果希望判断同步任务是否结束，建议使用调度依赖为同步任务添加下游依赖的MaxCompute SQL节点，节点中创建特定的MaxCompute表或分区，您可以根据特定的表或分区是否存在来判断同步任务是否完成。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。

参数	描述
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下，详情请参见上述参数说明。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"odps",//插件名。
            "parameter":{
                "partition":"",//分区信息。
                "truncate":true,//清理规则。
                "compress":false,//是否压缩。
                "datasource":"odps_first",//数据源名。
            "column": [//源端列名。
                "id",
                "name",
                "age",
                "sex",
                "salary",
                "interest"
                ],
                "emptyAsNull":false,//空字符串是否作为null。
                "table":""//表名。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数，表示脏数据的最大容忍条数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

如果您需要指定MaxCompute的Tunnel Endpoint，可以通过脚本模式手动配置数据源：将上述示例中的 `"datasource":"",` 替换为数据源的具体参数，示例如下。

```
"accessId":"<yourAccessKeyId>",
 "accessKey":"<yourAccessKeySecret>",
 "endpoint":"http://service.eu-central-1.maxcompute.aliyun-inc.com/api",
 "odpsServer":"http://service.eu-central-1.maxcompute.aliyun-inc.com/api",
"tunnelServer":"http://dt.eu-central-1.maxcompute.aliyun.com",
"project":"**********",
```

补充说明

- 关于列筛选的问题

  通过配置MaxCompute Writer，可以实现MaxCompute本身不支持的列筛选、重排序和补空等操作。例如需要导入的字段列表，当导入全部字段时，可以配置为 `"column": ["*"]` 。

  MaxCompute表有a、b和c三个字段，您只同步c和b两个字段，可以将列配置为 `"column": ["c","b"]` ，表示会把Reader的第一列和第二列导入MaxCompute的c字段和b字段，而MaxCompute表中新插入的a字段会被置为null。

- 列配置错误的处理

  为保证写入数据的可靠性，避免多余列数据丢失造成数据质量故障。对于写入多余的列，MaxCompute Writer将报错。例如MaxCompute表字段为a、b和c，如果MaxCompute Writer写入的字段多于三列，MaxCompute Writer将报错。

- 分区配置注意事项

  MaxCompute Writer仅提供写入到最后一级分区的功能，不支持写入按照某个字段进行分区路由等功能。假设表一共有三级分区，在分区配置中必须指明写入至某个三级分区。例如，写入数据至一个表的第三级分区，可以配置为 `pt=20150101, type=1, biz=2` ，但不能配置为 `pt=20150101, type=1` 或者 `pt=20150101` 。

- 任务重跑和failover

  MaxCompute Writer通过配置 `"truncate": true` ，保证写入的幂等性。即当出现写入失败再次运行时，MaxCompute Writer将清理前述数据，并导入新数据，以保证每次重跑之后的数据都保持一致。如果在运行过程中，因为其它的异常导致任务中断，便不能保证数据的原子性，数据不会回滚也不会自动重跑，需要您利用幂等性的特点进行重跑，以确保数据的完整性。

  > **说明** truncate为true的情况下，会将指定分区或表的数据全部清理，请谨慎使用。

6.3.30. Hive Writer

Hive Writer插件实现了从Hive写出数据至HDFS的功能，本文为您介绍Hive Writer的工作原理、参数和示例。

背景信息

Hive是基于Hadoop的数据仓库工具，用于解决海量结构化日志的数据统计。Hive可以将结构化的数据文件映射为一张表，并提供SQL查询功能。

> **注意** Hive Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。Hive Writer支持的版本请参见下文的版本支持汇总。

Hive的本质是转化HQL或SQL语句为MapReduce程序：

- Hive处理的数据存储在HDFS中。
- Hive分析数据底层的实现是MapReduce。
- Hive的执行程序运行在Yarn上。

实现原理

Hive Writer插件通过访问Hive Metastore服务，解析出您配置的数据表的HDFS文件存储路径、文件格式和分隔符等信息。通过读取HDFS文件的方式，从Hive写出数据至HDFS。再通过Hive JDBC客户端执行SQL语句，加载HDFS文件中的数据至Hive表。

Hive Writer底层的逻辑和HDFS Writer插件一致，您可以在Hive Writer插件参数中配置HDFS Writer相关的参数，配置的参数会透传给HDFS Writer插件。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，必须与添加的数据源名称保持一致。	是	无

参数	描述	是否必选	默认值
column	需要写出的字段列，例如 `"column": ["id", "name"]` 。 • 支持列裁剪：即可以导出部分列。 • column必须显示指定同步的列集合，不允许为空。 • 不支持列换序。	是	无
table	需要写出的Hive表名。 **说明** 请注意大小写。	是	无
partition	Hive表的分区信息： • 如果您写出的Hive表是分区表，您需要配置partition信息。同步任务会写出partition对应的分区数据。 • 如果您的Hive表是非分区表，则无需配置partition。	否	无
writeMode	Hive表数据的写出模式。数据写出至HDFS文件后，Hive Writer插件会执行 `LOAD DATA INPATH (overwrite) INTO TABLE` ，加载数据至Hive表中。 writeMode用于表示数据的加载行为： • 如果writeMode为*truncate*，表示先清空数据再加载。 • 如果writeMode为*append*，表示保留原有数据。 • 如果writeMode为其它，则表示数据写出至HDFS文件，无需再加载数据至Hive表。 **说明** writeMode是高危参数，请您注意数据的写出目录和writeMode行为，避免误删数据。 加载数据行为需要配合hiveConfig使用，请注意您的配置。	是	无
hiveConfig	您可以在hiveConfig中配置进一步的Hive扩展参数，包括hiveCommand、jdbcUrl、username和password： • hiveCommand：表示Hive客户端工具的全路径，执行 `hive -e` 后，会执行writeMode相关联的 `LOAD DATA IN PATH` 数据加载操作。 Hive相关的访问信息由hiveCommand对应的客户端保证。 • jdbcUrl、username和password表示Hive的JDBC访问信息。HiveWriter会通过Hive JDBC驱动访问Hive后，执行writeMode相关联的 `LOAD DATA INPATH` 数据加载操作。 `"hiveConfig": { "hiveCommand": "", "jdbcUrl": "", "username": "", "password": "" }` • Hive Writer插件底层通过HDFS客户端，写入数据至HDFS文件。您也可以通过hiveConfig配置HDFS客户端的高级参数。	是	无

向导开发介绍

在**数据开发**页面，双击打开新建的数据同步节点，即可在右侧的编辑页面配置任务。详情请参见通过向导模式配置离线同步任务。

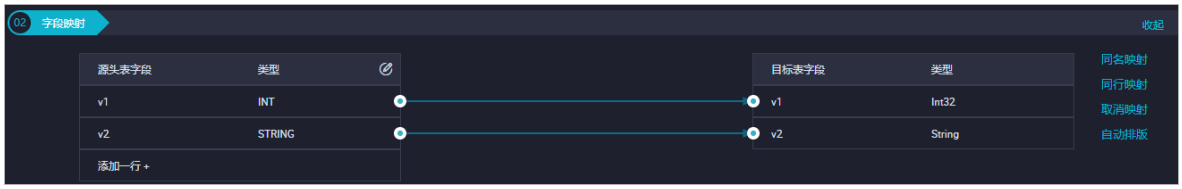
您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常选择您配置的数据源名称。
表	即上述参数说明中的table。
分区信息	您需要指定到最后一级分区，且仅支持写入单个分区。
写入模式	即上述参数说明中的writeMode。
parquet shchema	*parquet*类型存储文件的模式。示例配置如下： `message tmp{` `REQUIRED BINARY  id;` `REQUIRED BINARY name;` `REQUIRED BINARY cyctime1;` `}` 仅当Hive表底层存储文件为*parquet*类型时，需要配置该参数。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。

脚本开发介绍

配置一个从Hive写出数据的JSON示例，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "hive",
            "parameter": {
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "hive",
            "parameter": {
                "partition": "year=a,month=b,day=c", // 分区配置
                "datasource": "hive_ha_shanghai", // 数据源
                "table": "partitiontable2", // 目标表
                "column": [ // 列配置
                    "id",
                    "name",
                    "age"
                ],
                "writeMode": "append" // 写入模式
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":2, //作业并发数。
            "mbps":"12"//限流
        }
    }
}
```

版本支持汇总

Hive Reader支持的版本如下。

```
0.8.0
0.8.1
0.9.0
0.10.0
0.11.0
0.12.0
0.13.0
0.13.1
0.14.0
1.0.0
1.0.1
1.1.0
```

```
1.1.1
1.2.0
1.2.1
1.2.2
2.0.0
2.0.1
2.1.0
2.1.1
2.2.0
2.3.0
2.3.1
2.3.2
2.3.3
2.3.4
2.3.5
2.3.6
2.3.7
3.0.0
3.1.0
3.1.1
3.1.2
0.8.1-cdh4.0.0
0.8.1-cdh4.0.1
0.9.0-cdh4.1.0
0.9.0-cdh4.1.1
0.9.0-cdh4.1.2
0.9.0-cdh4.1.3
0.9.0-cdh4.1.4
0.9.0-cdh4.1.5
0.10.0-cdh4.2.0
0.10.0-cdh4.2.1
0.10.0-cdh4.2.2
0.10.0-cdh4.3.0
0.10.0-cdh4.3.1
0.10.0-cdh4.3.2
0.10.0-cdh4.4.0
0.10.0-cdh4.5.0
0.10.0-cdh4.5.0.1
0.10.0-cdh4.5.0.2
0.10.0-cdh4.6.0
0.10.0-cdh4.7.0
0.10.0-cdh4.7.1
0.12.0-cdh5.0.0
0.12.0-cdh5.0.1
0.12.0-cdh5.0.2
0.12.0-cdh5.0.3
0.12.0-cdh5.0.4
0.12.0-cdh5.0.5
0.12.0-cdh5.0.6
0.12.0-cdh5.1.0
0.12.0-cdh5.1.2
0.12.0-cdh5.1.3
0.12.0-cdh5.1.4
0.12.0-cdh5.1.5
0.13.1-cdh5.2.0
0.13.1-cdh5.2.1
0.13.1-cdh5.2.2
0.13.1-cdh5.2.3
0.13.1-cdh5.2.4
0.13.1-cdh5.2.5
0.13.1-cdh5.2.6
0.13.1-cdh5.3.0
0.13.1-cdh5.3.1
0.13.1-cdh5.3.2
```

```
0.13.1-cdh5.3.3
0.13.1-cdh5.3.4
0.13.1-cdh5.3.5
0.13.1-cdh5.3.6
0.13.1-cdh5.3.8
0.13.1-cdh5.3.9
0.13.1-cdh5.3.10
1.1.0-cdh5.3.6
1.1.0-cdh5.4.0
1.1.0-cdh5.4.1
1.1.0-cdh5.4.2
1.1.0-cdh5.4.3
1.1.0-cdh5.4.4
1.1.0-cdh5.4.5
1.1.0-cdh5.4.7
1.1.0-cdh5.4.8
1.1.0-cdh5.4.9
1.1.0-cdh5.4.10
1.1.0-cdh5.4.11
1.1.0-cdh5.5.0
1.1.0-cdh5.5.1
1.1.0-cdh5.5.2
1.1.0-cdh5.5.4
1.1.0-cdh5.5.5
1.1.0-cdh5.5.6
1.1.0-cdh5.6.0
1.1.0-cdh5.6.1
1.1.0-cdh5.7.0
1.1.0-cdh5.7.1
1.1.0-cdh5.7.2
1.1.0-cdh5.7.3
1.1.0-cdh5.7.4
1.1.0-cdh5.7.5
1.1.0-cdh5.7.6
1.1.0-cdh5.8.0
1.1.0-cdh5.8.2
1.1.0-cdh5.8.3
1.1.0-cdh5.8.4
1.1.0-cdh5.8.5
1.1.0-cdh5.9.0
1.1.0-cdh5.9.1
1.1.0-cdh5.9.2
1.1.0-cdh5.9.3
1.1.0-cdh5.10.0
1.1.0-cdh5.10.1
1.1.0-cdh5.10.2
1.1.0-cdh5.11.0
1.1.0-cdh5.11.1
1.1.0-cdh5.11.2
1.1.0-cdh5.12.0
1.1.0-cdh5.12.1
1.1.0-cdh5.12.2
1.1.0-cdh5.13.0
1.1.0-cdh5.13.1
1.1.0-cdh5.13.2
1.1.0-cdh5.13.3
1.1.0-cdh5.14.0
1.1.0-cdh5.14.2
1.1.0-cdh5.14.4
1.1.0-cdh5.15.0
1.1.0-cdh5.16.0
1.1.0-cdh5.16.2
1.1.0-cdh5.16.99
2.1.1-cdh6.1.1
```

```
2.1.1-cdh6.2.0
2.1.1-cdh6.2.1
2.1.1-cdh6.3.0
2.1.1-cdh6.3.1
2.1.1-cdh6.3.2
2.1.1-cdh6.3.3
```

6.3.31. Maxgraph Writer

本文为您介绍Maxgraph Writer支持的数据类型、字段映射和数据源等参数及配置示例。

> **说明** 如果您需要导入自己的MaxCompute表至Maxgraph，请先在源端MaxCompute项目中授予Maxgraph build账号读取源端MaxCompute表的权限。请联系Maxgraph管理员提供关于授权的Maxgraph build账号。

您可以通过创建数据和上线数据等操作，导入MaxCompute表至Maxgraph。

1. Maxgraph读取您的MaxCompute表，根据MaxCompute表中的列到Maxgraph中点或边的映射关系，执行一个MapReduce作业。该MapReduce作业会将数据处理为Maxgraph内部存储需要的数据格式。
2. Maxgraph将上一步处理好的数据载入Maxgraph的存储。

参数说明

参数	描述	是否必选	默认值
endpoint	Maxgraph的URL。	是	无
graphName	图实例的名称。	是	无
accessId	用户名。	是	无
accessKey	用户密码。	是	无
label	标签名，即点或边的名称。	是	无
labelType	标签类型，只能选择vertex或edge其中一种类型。	是	无
srcLabel	边的起点标签，仅在导入边时使用。	是	无
dstLabel	边的终点标签，仅在导入边时使用。	是	无
splitSize	创建数据过程中MapReduce作业的分片大小。	否	*256MB*
onlineMode	数据上线模式，包括partition和type。两者区别如下： • 以partition模式上线，在上线过程中可能会查询到旧数据与新数据混合的结果，但是保证最终一致性，上线速度较快。 • 以type模式上线，只有数据上线成功后才能查询到新上线的数据，在数据上线过程中只能查询到旧的数据（假如存在旧的数据），上线速度较慢。	否	*type*
column	点的属性名，仅导入点时使用。	是	无
name	属性的名称。	仅导入边时必填	无

参数	描述	是否必选	默认值
propertyType	属性的类型，包括srcPrimaryKey、dstPrimaryKey和edgeProperty。	仅导入边时必填	无
srcPrimaryKey	起点主键，仅导入边时使用。	仅导入边时必填	无
dstPrimaryKey	终点主键，仅导入边时使用。	仅导入边时必填	无
edgeProperty	边的属性，如果边没有属性，可以不填。	否	无

功能说明

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

Maxgraph中点和边的导入配置不一致，所以Maxgraph Writer需要区分点和边。

- 点配置示例

```
{
  "job": {
    "setting": {
      "speed": {
        "channel": 1 //配置channel为1即可。
      },
      "errorLimit": {
        "record": 1000
      }
    },
    "content": [
      {
        "reader": {
          "name": "odpsreader",
          "parameter": {
            "accessId": "*****",
            "accessKey": "*****",
            "project": "maxgraph_dev",
            "table": "maxgraph_demo_person",
            "column": [ //对应MaxCompute表的column名称，与Writer配置项中的column一一对应。
              "id",
              "name",
              "age"
            ],
            "packageAuthorizedProject": "biggraph_dev",
            "splitMode": "record",
            "odpsServer": "******"
          }
        },
        "writer": {
          "name": "maxgraphwriter",
          "parameter": {
            "endpoint": "http://graph.alibaba.net",
            "graphName": "xxx",
            "accessId": "xxx",
            "accessKey": "xxx",
            "label": "person",
            "labelType": "vertex",
            "onlineMode": "partition",
            "splitSize": "256",
            "column": [ //对应Maxgraph vertex的属性名，与reader配置项中的column一一对应。
              "id",
              "name",
              "age"
            ]
          }
        }
      }
    ]
  }
}
```

- 边配置示例

```
{
  "job": {
    "setting": {
      "speed": {
        "channel": 1 //配置channel为1即可。
      },
      "errorLimit": {
        "record": 1000
      }
```

```
    },
    "content": [
      {
        "reader": {
          "name": "odpsreader",
          "parameter": {
            "accessId": "*****",
            "accessKey": "*****",
            "project": "maxgraph_dev",
            "table": "maxgraph_demo_knows",
            "column": [
              "person_id",
              "person_id2",
              "weight",
              "id"
            ],
            "packageAuthorizedProject": "biggraph_dev",
            "splitMode": "record",
            "odpsServer": "*****"
          }
        },
        "writer": {
          "name": "maxgraphwriter",
          "parameter": {
            "endpoint": "http://graph.alibaba.net",
            "graphName": "xxx",
            "accessId": "xxx",
            "accessKey": "xxx",
            "label": "knows",
            "labelType": "edge",
            "srcLabel": "person",
            "dstLabel": "person",
            "onlineMode": "partition",
            "splitSize": "256",
            "column": [
              {
                "name": "id", //对应的在graph中的属性名。
                "propertyType": "srcPrimaryKey" //对应该属性的类型，主要用来标识是起点、终点还是边的属性。
              },
              {
                "name": "id",
                "propertyType": "dstPrimaryKey"
              },
              {
                "name": "weight",
                "propertyType": "edgeProperty"
              },
              {
                "name": "id",
                "propertyType": "edgeProperty"
              }
            ]
          }
        }
      }
    ]
  }
}
```

6.3.32. Kafka Writer

Kafka Writer通过Kafka服务的Java SDK向Kafka写入数据，本文为您介绍Kafka Writer的实现原理、参数和示例。

> 注意 目前Kafka Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

Apache Kafka是一个快速、可扩展、高吞吐和可容错的分布式发布订阅消息系统。Kafka具有高吞吐量、内置分区、支持数据副本和容错的特性，适合在大规模消息处理的场景中使用。

实现原理

Kafka Writer通过Java SDK向Kafka中写入数据，使用的Kafka服务Java SDK版本如下。

```
<dependency>
   <groupId>org.apache.kafka</groupId>
   <artifactId>kafka-clients</artifactId>
   <version>2.0.0</version>
</dependency>
```

参数说明

参数	描述	是否必选
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须要与添加的数据源名称保持一致。	是
server	Kafka的server地址，格式为ip:port。	是
topic	Kafka的topic，是Kafka处理资源的消息源（feeds of messages）的不同分类。 每条发布至Kafka集群的消息都有一个类别，该类别被称为topic，一个topic是对一组消息的归纳。	是
valueIndex	Kafka Writer中作为Value的那一列。如果不填写，默认将所有列拼起来作为Value，分隔符为fieldDelimiter。	否
writeMode	当未配置valueIndex时，该配置项决定将源端读取记录的所有列拼接作为写入kafka记录Value的格式，可选值为text和JSON，默认值为text。 • 配置为text，将所有列按照fieldDelimiter指定分隔符拼接。 • 配置为JSON，将所有列按照column参数指定字段名称拼接为JSON字符串。 例如源端记录有三列，值为a、b和c，writeMode配置为text、fieldDelimiter配置为#时，写入kafka的记录Value为字符串a#b#c；writeMode配置为JSON、column配置为[{"name":"col1"},{"name":"col2"},{"name":"col3"}]时，写入kafka的记录Value为字符串{"col1":"a","col2":"b","col3":"c"}。 如果配置了valueIndex，该配置项无效。	否

参数	描述	是否必选
column	目标表需要写入数据的字段，字段间用英文逗号分隔。例如："column": ["id", "name", "age"]。 当未配置valueIndex，并且writeMode选择JSON时，该配置项定义源端读取记录的列值在JSON结构中的字段名称。例如，"column": [{"name":id"}, {"name":"name"}, {"name":"age"}]。 • 当源端读取记录列的个数多于column配置的字段名个数时，写入时进行截断。例如： 源端记录有三列，值为a、b和c，column配置为[{"name":"col1"}, {"name":"col2"}]时，写入kafka的记录Value为字符串{"col1":"a","col2":"b"}。 • 当源端读取记录列的个数少于column配置的字段名个数时，多余column配置字段名填充null或者nullValueFormat指定的字符串。例如： 源端记录有两列，值为a和b，column配置为[{"name":"col1"}, {"name":"col2"},{"name":"col3"}]时，写入kafka的记录Value为字符串{"col1":"a","col2":"b","col3":null}。如果配置了valueIndex，或者writeMode配置为text，该配置项无效。 如果配置了valueIndex，或者writeMode配置为text，该配置项无效。	当未配置valueIndex，并且writeMode配置为JSON时必选
partition	指定写入Kafka topic指定分区的编号，是一个大于等于0的整数。	否
keyIndex	Kafka Writer中作为Key的那一列。 keyIndex参数取值范围是大于等于0的整数，否则任务会出错。	否
keyIndexes	源端读取记录中作为写入kafka记录Key的列的序号数组。 列序号从0开始，例如[0,1,2]，会将配置的所有列序号的值用逗号连接作为写入kafka记录的Key。如果不填写，写入kafka记录Key为null，数据轮流写入topic的各个分区中，与keyIndex参数只能二选一。	否
fieldDelimiter	当writeMode配置为text，并且未配置valueIndex时，将源端读取记录的所有列按照该配置项指定列分隔符拼接作为写入kafka记录的Value，支持配置单个或者多个字符作为分隔符，支持以 \u0001格式配置unicode字符，支持 \t、\n等转义字符。默认值为 \t。 如果writeMode未配置为text或者配置了valueIndex，该配置项无效。	否
keyType	Kafka的Key的类型，包括BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。	是
valueType	Kafka的Value的类型，包括BYTEARRAY、DOUBLE、FLOAT、INTEGER、LONG和SHORT。	是
nullKeyFormat	keyIndex或者keyIndexes指定的源端列值为null时，替换为该配置项指定的字符串，如果不配置不做替换。	否
nullValueFormat	当源端列值为null时，组装写入kafka记录Value时替换为该配置项指定的字符串，如果不配置不做替换。	否
acks	初始化Kafka Producer时的acks配置，决定写入成功的确认方式。默认acks参数为all。acks取值如下： • 0：不进行写入成功确认。 • 1：确认主副本写入成功。 • all：确认所有副本写入成功。	否

向导模式开发

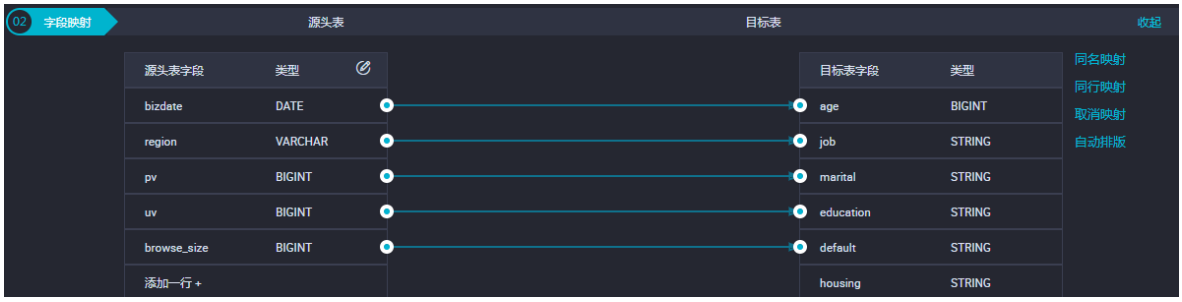
1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
主题	即上述参数说明中的topic。
键取值列	即上述参数说明中的keyIndexes，指定写入Kafka记录的Key的取值方式。
写入模式	即上述参数说明中的writeMode，指定写入Kafka记录的格式。
分隔符	即上述参数说明中的fieldDelimiter，当写入模式为text时决定写入Kafka记录的Value的连接字符。
null键替代字符串	即上述参数说明中的nullKeyFormat，指定null的替代方案。
null值替代字符串	即上述参数说明中的nullValueFormat，指定null的替代方案。
写入成功确认方式	即上述参数说明中的acks，指定写入成功确认方式。
单次写入大小	决定初始化Kafka Producer时的batch.size和linger.ms参数，控制单次写入数据量，默认值为batch.size=16384、linger.ms=10。
写入超时时间	决定初始化Kafka Producer时的timeout.ms、request.timeout.ms和metadata.fetch.timeout.ms参数，控制单次写入超时时间，默认值为timeout.ms=30000，request.timeout.ms=30000，metadata.fetch.timeout.ms=60000。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。
   - 当写入模式为text时，映射到目标表字段的源头表字段的值使用分隔符连接作为Kafka记录的value。
   - 当写入模式为JSON时，目标表字段名将决定源端读取记录的列值在JSON结构中的字段名，例如源头表字段有两列，值为a和b，目标表字段配置为col1和col2，那么写入kafka的记录Value为字符串{"col1":"a","col2":"b"}。

> **说明** 目标表字段名称必须使用字母、数字或者下划线，否则任务运行时会失败。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本模式开发

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

向Kafka写入数据的JSON配置，如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"Kafka",//插件名。
            "parameter":{
                    "server": "ip:9092", //Kafka的server地址。
                    "keyIndex": 0, //作为Key的列。需遵循驼峰命名规则，k小写
                    "valueIndex": 1, //作为Value的某列。目前只支持取来源端数据的一列或者该参数不填（不填表示取来源所有数
据）
        //例如想取odps的第2、3、4列数据作为kafkaValue，请新建odps表将原odps表数据做清洗整合写新odps表后使用新表同步。
                    "keyType": "Integer", //Kafka的Key的类型。
                    "valueType": "Short", //Kafka的Value的类型。
                    "topic": "t08", //Kafka的topic。
                    "batchSize": 1024 //向kafka一次性写入的数据量。
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
                     "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示
限流。
                     "concurrent":1, //作业并发数。
                     "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

使用SASL鉴权

如果需要使用SASL鉴权或SSL鉴权，请在定义kafka数据源时进行相关配置，详情请参考：配置Kafka数据源。

6.3.33. Vertica Writer

Vertica是一款基于列存储的MPP架构的数据库，Vertica Writer插件实现了向Vertica写入数据的功能。本文将为您介绍Vertica Writer的实现原理、参数和示例。

> 注意 Vertica Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

实现原理

在底层实现上，Vertica Writer通过JDBC连接远程Vertica数据库，并执行相应的 `insert into...` 语句，写入数据至Vertica，内部会分批次提交入库。

Vertica Writer面向ETL开发工程师，通过Vertica Writer从数仓导入数据至Vertica。同时，Vertica Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

Vertica Writer通过数据同步框架获取Reader生成的协议数据，根据您的配置生成相应的SQL插入语句：

- `insert into...`：当主键或唯一性索引冲突时，会写不进去冲突的行。
- 目标表所在数据库必须是主库才能写入数据。

> **说明** 整个任务需要至少具备 `insert into...` 的权限。是否需要其它权限，取决于您配置任务时，在preSql和postSql中指定的语句。

- Vertica Writer不支持配置writeMode参数。
- Vertica Writer通过Vertica数据库驱动访问Vertica，您需要确认Vertica驱动和您的Vertica服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>com.vertica</groupId>
    <artifactId>vertica-jdbc</artifactId>
    <version>7.1.2</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
jdbcUrl	描述的是到对端数据库的JDBC连接信息，jdbcUrl包含在connection配置单元中： • 在一个数据库上只能配置一个值，不支持同一个数据库存在多个主库的情况（双主导入数据情况）。 • jdbcUrl的格式和Vertica官方一致，并可以连接附加参数信息。例如，`jdbc:vertica://127.0.0.1:3306/database`。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无
table	选取的需要同步的表名称，使用JSON的数组进行描述。 **说明** table必须包含在connection配置单元中。	是	无
column	目标表需要写入数据的字段，字段之间用英文逗号分隔，例如 `"column": ["id", "name", "age"]`。	是	无
preSql	写入数据至目标表前，会先执行此处的标准语句。如果SQL中有需要操作的表名称，请使用 `@table` 表示，以便在实际执行SQL语句时，对变量按照实际表名称进行替换。	否	无
postSql	写入数据至目标表后，会执行此处的标准语句。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与Vertica的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个Vertica写入数据的作业，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"vertica",//插件名。
            "parameter":{
                "datasource": "数据源名",
                "username": "",
                "password": "",
                "column": [//字段。
                    "id",
                    "name"
                ],
                "connection": [
                    {
                        "table": [//表名。
                            "vertica_table"
                        ],
                        "jdbcUrl": "jdbc:vertica://ip:port/database"
                    }
                ],
                "preSql": [ //执行数据同步任务之前率先执行的SQL语句。
                    "delete from @table where db_id = -1"
                ],
                "postSql": [//执行数据同步任务之后率先执行的SQL语句。
                    "update @table set db_modify_time = now() where db_id = 1"
                ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
                    "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
                    "concurrent":1, //作业并发数。
                    "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.34. Gbase8a Writer

本文为您介绍Gbase8a Writer支持的数据类型、字段映射和数据源等参数及配置示例。

Gbase8a是一款基于列存储的新型分析型数据库，Gbase8a Writer插件实现了写入数据至Gbase8a数据库的功能。

> 注意 Gbase8a Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

在底层实现上，Gbase8a Writer通过JDBC连接远程Gbase8a数据库，并执行相应的 `insert into` 语句，将数据写入Gbase8a，数据会分批次提交入库。

Gbase8a Writer面向ETL开发工程师，通过Gbase8a Writer从数仓导入数据至Gbase8a。同时Gbase8a Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

Gbase8a Writer通过数据集成框架获取Reader生成的协议数据，根据您的配置生成相应的SQL插入语句。

使用限制

- `insert into...`：当主键或唯一性索引冲突时，无法写入冲突的行。
- 目的表所在数据库必须是主库才能写入数据。

  > 说明 整个任务需要至少具备 `insert into...` 的权限。是否需要其它权限，取决于您配置任务时，在preSql和postSql中指定的语句。

- Gbase8a Writer不支持配置writeMode参数。
- Gbase8a Writer通过MySQL数据库驱动访问Gbase8a，您需要确认驱动和您的Gbase8a服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>mysql</groupId>
    <artifactId>mysql-connector-java</artifactId>
    <version>5.1.22</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
jdbcUrl	描述的是到对端数据库的JDBC连接信息，jdbcUrl包含在connection配置单元中。 • 在一个数据库上只能配置一个值，不支持同一个数据库存在多个主库的情况（双主导入数据情况）。 • jdbcUrl的格式和Gbase8a官方一致，并可以连接附加参数信息。例如，`jdbc:mysql://127.0.0.1:3306/database`。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无
table	需要同步写出的表名称，使用JSON的数组进行描述。 说明 table必须包含在connection配置单元中。	是	无
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔。例如，`"column": ["id", "name", "age"]`。 说明 column配置项必须指定，不能为空。	是	无

参数	描述	是否必选	默认值
preSql	写入数据至目标表前，会先执行此处的标准语句。如果SQL中有需要操作的表名称，请使用 `@table` 表示，以便在实际执行SQL语句时，对变量按照实际表名称进行替换。	否	无
postSql	写入数据至目标表后，会执行此处的标准语句。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与Gbase8a的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

向Gbase8a写入数据的JSON配置，如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"gbase8a",//插件名。
            "parameter":{
                "datasource": "数据源名",
                "username": "",
                "password": "",
                "column": [//字段。
                    "id",
                    "name"
                ],
                "connection": [
                    {
                        "table": [//表名。
                            "Gbase8a_table"
                        ],
                        "jdbcUrl": "jdbc:mysql://ip:port/database"
                    }
                ],
                "preSql": [ //执行数据同步任务之前率先执行的SQL语句。
                    "delete from @table where db_id = -1"
                ],
                "postSql": [//执行数据同步任务之后率先执行的SQL语句。
                    "update @table set db_modify_time = now() where db_id = 1"
                ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.35. ClickHouse Writer

ClickHouse是面向联机分析处理（OLAP）和列式存储的开源的数据库管理系统（DBMS），本文为您介绍ClickHouse Writer的实现原理、参数说明及配置示例。

使用限制

- 仅支持阿里云ClickHouse。
- ClickHouse Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。
- ClickHouse Writer使用JDBC连接ClickHouse，且仅支持使用JDBC Statement插入数据。
- ClickHouse Writer支持筛选部分列、列换序等功能，您可以自行填写列。
- 考虑到ClickHouse负载问题，ClickHouse Writer使用INSERT模式时，建议您限流系统吞吐量（TPS）最高为1,000。
- ClickHouse Writer在所有Task写入任务后，Job Post单进程执行Flush工作，保证数据在ClickHouse整体更新。
- 您需要确认驱动和您的ClickHouse服务之间的兼容能力，数据库驱动使用如下版本。

```
<dependency>
    <groupId>ru.yandex.clickhouse</groupId>
    <artifactId>clickhouse-jdbc</artifactId>
    <version>0.2.4.ali2-SNAPSHOT</version>
</dependency>
```

背景信息

ClickHouse Writer实现了向ClickHouse写出数据的功能。在底层实现上，ClickHouse Writer通过JDBC连接远程ClickHouse数据库，并执行相应的 `insert into` 语句，写入数据至ClickHouse。

ClickHouse Writer面向ETL开发工程师，通过ClickHouse Writer从数仓导入数据至ClickHouse。同时ClickHouse Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

ClickHouse Writer通过数据集成框架获取Reader生成的协议数据，并利用ClickHouse暴露的INSERT接口写入ClickHouse，根据您的配置生成相应的SQL插入语句。

参数说明

参数	描述	是否必选	默认值
jdbcUrl	到对端数据库的JDBC连接信息，jdbcUrl包含在connection配置单元中。 • 在一个数据库上只能配置一个值。 • jdbcUrl的格式和ClickHouse官方一致，并可以连接附加参数信息。例如，`jdbc:clickhouse://127.0.0.1:3306/database`。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无
table	需要同步写出的表名称，使用JSON的数组进行描述。 **说明** table必须包含在connection配置单元中。	是	无
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔。例如 `"column": ["id", "name", "age"]`。 **说明** column配置项必须指定，不能为空。	是	无
preSql	写入数据至目标表前，会先执行此处的标准语句。	否	无
postSql	写入数据至目标表后，会执行此处的标准语句。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与ClickHouse的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

> **说明** 实际执行时，请删除下述代码中的注释。

向ClickHouse写出数据的JSON配置，如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"clickhouse",//插件名。
            "parameter":{
                "username": "",
                "password": "",
                "column": [//字段。
                    "id",
                    "name"
                ],
                "connection": [
                    {
                        "table": [//表名。
                            "ClickHouse_table"
                        ],
                        "jdbcUrl": "jdbc:clickhouse://ip:port/database"
                    }
                ],
                "preSql": [ //执行数据同步任务之前率先执行的SQL语句。
                    "TRUNCATE TABLE IF EXISTS tablename"
                ],
                "postSql": [//执行数据同步任务之后率先执行的SQL语句。
                    "ALTER TABLE tablename UPDATE col1=1 WHERE col2=2"
                ],
                "batchSize": "1024",
                "batchByteSize": "67108864",
                "writeMode": "insert"
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.36. ApsaraDB For OceanBase Writer

ApsaraDB For OceanBase是阿里云和蚂蚁金服自主研发的金融级分布式关系数据库，本文将为您介绍ApsaraDB For OceanBase Writer的实现原理、参数说明及配置示例。

> **注意** ApsaraDB For OceanBase Writer仅支持使用新增和使用独享数据集成资源组，不支持使用默认资源组和自定义资源组。

背景信息

ApsaraDB For OceanBase在金融行业创造了三地五中心的城市级故障自动无损容灾的新标准，在普通硬件上实现了金融高可用。同时具备在线水平扩展能力，是在功能、稳定性、可扩展性、性能方面都经历严格检验的国产数据库。

ApsaraDB For OceanBase Writer面向ETL开发工程师，通过ApsaraDB For OceanBase Writer从数仓导入数据至ApsaraDB For OceanBase。同时ApsaraDB For OceanBase Writer可以作为数据迁移工具，为数据库管理员等用户提供服务。

ApsaraDB For OceanBase Writer通过数据集成框架获取Reader生成的协议数据，根据您的配置生成相应的SQL插入语句。

使用限制

- `insert into...`：当主键或唯一性索引冲突时，无法写入冲突的行。Oracle租户模式下仅支持`insert into`。
- `insert into...ON DUPLICATE KEY UPDATE...`：当主键或唯一性索引未冲突时，与`insert into`的行为一致。当主键或唯一性索引冲突时，新行会替换指定的字段。MySQL租户模式下支持`insert into...`和`insert into...ON DUPLICATE KEY UPDATE...`。
- 目的表所在数据库必须是主库才能写入数据。

  > **说明** 整个任务需要至少具备`insert into...`的权限。是否需要其它权限，取决于您配置任务时，在preSql和postSql中指定的语句。

- 建议您使用batch的方式批量写入数据，当行数累计到预定阈值时，再发起写入请求。
- ApsaraDB For OceanBase包括Oracle和MySQL两种租户模式，您在配置preSql、postSql时，需要符合对应租户模式的SQL语法约束，否则SQL语句可能执行失败。
- ApsaraDB For OceanBase Writer通过OceanBase数据库驱动访问ApsaraDB For OceanBase，您需要确认驱动和您的ApsaraDB For OceanBase服务之间的兼容能力。数据库驱动使用如下版本。

```
<dependency>
    <groupId>com.alipay.OceanBase</groupId>
    <artifactId>OceanBase-connector-java</artifactId>
    <version>3.1.0</version>
</dependency>
```

参数说明

参数	描述	是否必选	默认值
datasource	如果您使用的DataWorks版本支持添加ApsaraDB For OceanBase数据源，即可在此处根据数据源名称引用您添加的ApsaraDB For OceanBase数据源。 包括jdbcUrl和username两种配置方式。	否	无
jdbcUrl	到对端数据库的JDBC连接信息，jdbcUrl包含在connection配置单元中。 • 在一个数据库上只能配置一个值，不支持同一个数据库存在多个主库的情况（双主导入数据情况）。 • jdbcUrl的格式和ApsaraDB For OceanBase官方一致，并可以连接附加参数信息。例如，`jdbc:mysql://127.0.0.1:3306/database`。	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无

参数	描述	是否必选	默认值
table	需要同步写出的表名称，使用JSON的数组进行描述。 说明 table必须包含在connection配置单元中。	是	无
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔。例如，`"column": ["id", "name", "age"]`。 说明 column配置项必须指定，不能为空。	是	无
writeMode	控制写入数据至目标表使用的模式，包括`insert into`和`ON DUPLICATE KEY UPDATE`。	是	无
preSql	写入数据至目标表前，会先执行此处的标准语句。如果SQL中有需要操作的表名称，请使用`@table`表示，以便在实际执行SQL语句时，对变量按照实际表名称进行替换。	否	无
postSql	写入数据至目标表后，会执行此处的标准语句。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与服务器端的网络交互次数，并提升整体吞吐量。 说明 fetchSize值过大（>2048）可能造成数据同步进程OOM。	否	*1,024*

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

向ApsaraDB For OceanBase写出数据的JSON配置，如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"apsaradb_for_OceanBase",//插件名。
            "parameter":{
                "datasource": "数据源名",
                "column": [//字段。
                    "id",
                    "name"
                ],
                "table": "apsaradb_for_OceanBase_table",//表名。
                "preSql": [ //执行数据同步任务之前率先执行的SQL语句。
                    "delete from @table where db_id = -1"
                ],
                "postSql": [//执行数据同步任务之后率先执行的SQL语句。
                    "update @table set db_modify_time = now() where db_id = 1"
                ],
                "writeMode": "insert",
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.37. Hologres Writer

Hologres Writer实现了导入数据至实时数仓Hologres的功能，您可以把多种数据源的数据导入Hologres进行实时分析。

使用限制

- Hologres Writer不支持写入数据至Hologres的外部表。
- Hologres Writer仅支持使用新增和使用独享数据集成资源组，不支持使用使用公共资源组和自定义资源组。

实现原理

Hologres Writer通过数据同步框架获取Reader生成的协议数据，根据writeMode的配置选择不同的写入方式，并根据conflictMode的配置决定写入数据时的冲突解决策略：

- （推荐使用）writeMode为SQL模式时，通过PSQL的 `INSERT INTO` 命令（JDBC方式）写入数据，能提供更好的写入性能，建议您采用此写入模式同步数据。
- writeMode为SDK模式时，通过Hologres的写入接口来进行同步数据，后续该模式将逐步不再使用，建议使用SQL模式写入数据。

您可以通过配置conflictMode，决定新导入的数据和已有数据的主键发生冲突时，如何处理新导入的数据：

- conflictMode为Replace模式时，新数据覆盖旧数据。
- conflictMode为Ignore模式时，忽略新数据。

在不同的writeMode下，conflictMode的实现方式也不同。writeMode为SDK 模式时，通过设置Hologres Table属性来改变此次导入的冲突解决模式。

> 注意 conflictMode仅适用于有主键的表。

参数说明

参数	描述	是否必选	默认值
endpoint	目标交互式分析（Hologres）实例对应的endpoint，格式为 `instance-id-region-endpoint.hologres.aliyuncs.com:port` 。您可以从交互式分析实例的管理页面获取。 endpoint包括公网、经典网络和VPC三种网络类型，请根据数据集成资源组和Hologres实例所在的网络环境选择正确的endpoint类型，否则会出现网络不通或者性能受限的情况： • 公网示例： `instance-id-region-endpoint.hologres.aliyuncs.com:port` • 经典网络示例： `instance-id-region-endpoint-internal.hologres.aliyuncs.com:port` • VPC示例： `instance-id-region-endpoint-vpc.hologres.aliyuncs.com:port` 通常建议数据集成资源组和Hologres实例在同一个地域的同一个可用区，以确保网络连通，实现最大性能。	是	无
accessId	访问Hologres的accessId。	是	无
accessKey	访问Hologres的accessKey，请确保该密钥对目标表有写入权限。	是	无
database	Hologres实例内部数据库的名称。	是	无
table	Hologres的表名称，目前支持表名称中包含Schema，例如 `schema_name.table_name` 。	是	无
writeMode	writeMode包括SDK和SQL（INSERT INTO），详情请参见实现原理。 在脚本模式中，SDK的可选配置如下： • maxCommitSize：定义HoloWriter聚合的最大值。该参数可选，默认值为*1,048,576* Byte。 • maxRetryCount：定义HoloWriter通过SDK模式写入的最大重试次数。该参数可选，默认值为*500*。 • retryInterval：定义HoloWriter通过SDK模式写入出错时的重试间隔，单位为ms。该参数可选，默认值为1,000。	是	无
conflictMode	conflictMode包括Replace和Ignore，详情请参见实现原理。	是	无

参数	描述	是否必选	默认值
column	定义导入目标表的数据列，必须包含目标表的主键集合。例如 `["*"]` 表示全部列。	是	无
partition	针对分区表，表示分区Column以及对应的Value，格式为 `column=value` 。 **说明** • 目前Hologres仅支持LIST分区，分区Column仅支持单个Column分区，且仅支持INT4或TEXT类型。 • 请确认该参数和表DDL的分区配置匹配。	否	空，表示非分区表
truncate	写入Holo表之前是否需要清空目标表。 **说明** 仅脚本模式支持配置该参数。配置方式请参见脚本开发介绍。 • true：清空目标表。 **说明** 目前仅支持清空非分区表和静态分区表。不支持清空动态分区表，如果您是动态分区表，并且设置了参数值为true，同步任务将会异常退出。 • false：不清空目标表。	否	false

向导开发介绍

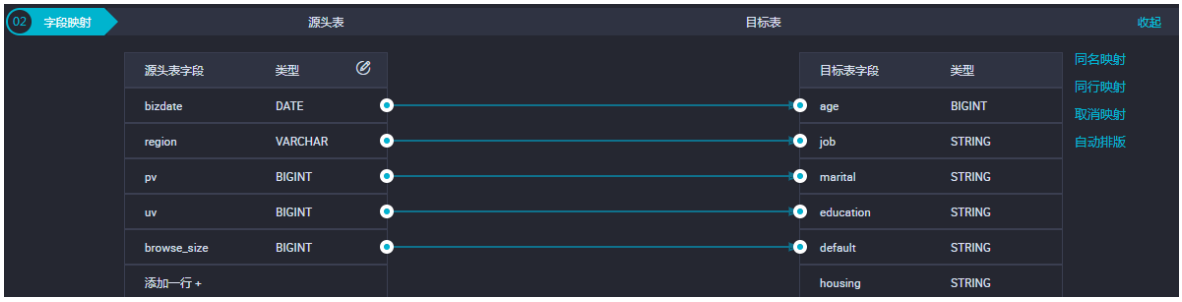
1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	通常输入您配置的数据源名称。第一次数据同步需要先创建Hologres数据源，详情请参见配置Hologres数据源。

参数	描述
表	需要同步的Hologres目标表，可以通过一键生成目标表自动创建，也可以提前在Hologres中创建。即上述参数说明中的table。 说明 ○ 一键生成目标表创建的仅是基础表，请根据实际业务修改建表语句并添加相应的索引，以获得更好的性能，详情请参见CREATE TABLE。 ○ 支持自动路由分区，若是通过一键生成目标表创建表，需要将表创建为分区表，以实现分区自动路由的功能，若是无对应分区子表则会自动创建，详情请参见CREATE PARTITION TABLE。
分区信息	Hologres的分区键。
写入模式	即上述参数说明中的writeMode。 ○ （推荐）**SQL(INSERT INTO)**，默认使用SQL模式，性能好。 ○ **SDK**，后续逐步不再使用，不推荐使用此模式。
写入冲突策略	即上述参数说明中的conflictMode。
最大连接数	JDBC使用的默认连接数，仅在SQL模式下使用，在开启任务时请确保实例有充足的空闲连接。默认9个连接，即一个任务使用9个连接。

2. 字段映射，即上述参数说明中的column，左侧的源头表字段和右侧的目标表字段为一一对应关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。配置非分区表和分区表的示例如下：

- 配置非分区表
  - 配置从内存产生的数据导入至Hologres普通表，示例为通过JDBC模式导入的配置。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "mysql",
            "parameter": {
                "envType": 0,
                    "datasource": "<mysql_source_name>",
                "column": [
                    "<column1>",
                    "<column2>",
                    ......,
                    "<columnN>"
                ],
                "connection": [
                    {
                        "datasource": "<mysql_source_name>",//mysql数据源名
                        "table": [
                            "<mysql_table_name>"
                        ]
                    }
                ],
                "where": "",
                "splitPk": "",
                "encoding": "UTF-8"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "holo",
            "parameter": {
                "maxConnectionCount": 9,
                "datasource": "<holo_sink_name>",//Hologres数据源名称
                "truncate":true,//清理规则。
                "conflictMode": "ignore",
                "envType": 0,
                "column": [
                    "<column1>",
                    "<column2>",
```

```
                    ......,
                    "<columnN>"
                ],
                "writeMode": "insert",
                "table": "<holo_table_name>"
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "executeMode": null,
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "concurrent": 2,//作业并发数
            "throttle": false//限流
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

- Hologres表的DDL语句，如下所示。

```
begin;
drop table if exists mysql_to_holo_test;
create table mysql_to_holo_test(
  tag text not null,
  id int not null,
  body text not null,
  brrth date,
  primary key (tag, id));
  call set_table_property('mysql_to_holo_test', 'orientation', 'column');
  call set_table_property('mysql_to_holo_test', 'distribution_key', 'id');
  call set_table_property('mysql_to_holo_test', 'clustering_key', 'birth');
commit;
```

- 配置分区表

> **说明**
> - 目前Hologres仅支持LIST分区，分区Column仅支持单个Column分区，且仅支持INT4或TEXT类型。
> - 请确认该参数和表DDL的分区配置匹配。

- 配置从内存产生的数据同步至Hologres分区表的子表，示例为通过SQL模式导入的配置。

```
{
  "type": "job",
  "version": "2.0",
  "steps": [
    {
      "stepType": "mysql",
      "parameter": {
        "envType": 0,
        "datasource": "<mysql_source_name>",
        "column": [
          "<column1>",
```

```
          "<column2>",
            ......,
          "<columnN>"
        ],
        "connection": [
          {
            "datasource": "<mysql_source_name>",
            "table": [
              "<mysql_table_name>"
            ]
          }
        ],
        "where": "",
        "splitPk": "<mysql_pk>",//mysql的pk字段
        "encoding": "UTF-8"
      },
      "name": "Reader",
      "category": "reader"
    },
    {
      "stepType": "holo",
      "parameter": {
        "maxConnectionCount": 9,
        "partition": "<partition_key>",//Hologres分区键
        "datasource": "<holo_sink_name>",//Hologres数据源名
        "conflictMode": "ignore",
        "envType": 0,
        "column": [
          "<column1>",
          "<column2>",
            ......,
          "<columnN>"
        ],
        "writeMode": "insert",
        "table": "<holo_table_name>"
      },
      "name": "Writer",
      "category": "writer"
    }
  ],
  "setting": {
    "executeMode": null,
    "errorLimit": {
      "record": ""
    },
    "speed": {
      "concurrent": 2,//作业并发数
      "throttle": false//限流
    }
  },
  "order": {
    "hops": [
      {
        "from": "Reader",
        "to": "Writer"
      }
    ]
  }
}
```

- Hologres表的DDL语句，如下所示。

```
BEGIN;
CREATE TABLE public.hologres_parent_table(
  a text ,
  b int,
  c timestamp,
  d text,
  ds text,
  primary key(ds,b)
  )
  PARTITION BY LIST(ds);
CALL set_table_property('public.hologres_parent_table', 'orientation', 'column');
CREATE TABLE public.holo_child_1 PARTITION OF public.hologres_parent_table FOR VALUES IN('20201215');
CREATE TABLE public.holo_child_2 PARTITION OF public.hologres_parent_table FOR VALUES IN('20201216');
CREATE TABLE public.holo_child_3 PARTITION OF public.hologres_parent_table FOR VALUES IN('20201217');
COMMIT;
```

6.3.38. RestAPI Writer

本文为您介绍RestAPI Writer支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

RestAPI Writer插件提供了向RESTful接口进行请求并且写入数据的功能。Writer插件获取到Reader端传递过来的数据后，根据column配置生成相应的JSON数据，对RESTful接口发起请求并传递数据。

参数说明

进行数据集成时，您需要添加数据源后再配置数据源的来源与去向，并在配置过程中设置好集成的数据及数据类型等信息，整个数据集成包含数据提取（使用reader插件提取数据来源的数据）和数据写入（使用writer插件将集成的数据写入数据去向的数据源中）。

以下为您介绍使用Writer插件写入RestAPI类型数据源的数据时，需要配置的参数。

> **说明** 以下的参数包含在添加数据源和配置数据集成任务节点的过程中。
>
> 当前插件暂不支持使用调度参数。

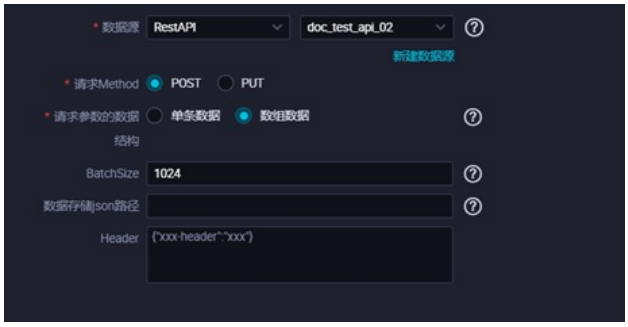
参数	描述	是否必须	默认值
url	RESTful接口地址。	是	无
dataMode	RESTful请求传递的JSON数据的格式。 • oneData：一次请求只传递一条数据，有几条数据就进行几次请求。 • multiData：一次请求传递一批数据，根据reader端切分的task数确定请求次数。	是	无
column	生成JSON数据对应的字段路径列表，type指定源数据的类型，name指定当前column数据放置的JSON路径。您可以指定column字段信息，配置如下。 "column":[{"type":"long","name":"a.b" //放置column数据到路径a.b},{"type":"string","name":"a.c"//放置column数据到路径a.c}] **说明** 对于您指定的column信息，type和name必须填写。	是	无
dataPath	数据结果放置的JSON对象的路径。	否	无
method	请求方法，支持post和put。	是	无
customHeader	传递给RESTful接口的header信息。	否	无

参数	描述	是否必须	默认值
authType	验证方法。 • Basic Auth：基础验证。 如果数据源API支持用户名和密码的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的用户名和密码，后续数据集成过程中对接数据源时，通过Basic Auth协议传递给RESTful地址，完成验证。 • Token Auth：Token验证。 如果数据源API支持Token的方式进行验证，您可选择此种验证方式，并在选择完成后配置用于验证的固定Token值，后续数据集成过程中对接数据源时，通过传入header中进行验证，例如：{“Authorization”:” Bearer TokenXXXXXX”}。 • Aliyun API Signature：阿里云API签名验证。 如果数据源为阿里云产品，且此阿里云产品的API支持通过AccessKey和AccessSecret的方式进行验证，您可选择此种种验证方式，并在选择完成后配置用于验证的AccessKey和AccessSecret。	否	无
authUsername/authPassword	Basic Auth验证的用户名密码。	否	无
authToken	Token Auth验证的token。	否	无
accessKey/accessSecret	Aliyun API签名验证的账户信息。	否	无
batchSize	dataMode为multiData时，一次请求最大的数据条数。	是	512

配置示例：向导模式

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	数据源类型选择**RestAPI**后，选择数据需要写入的数据表。
请求Method	即上述参数说明中的method。
返回数据结构	即上述参数说明中的dataMode。
batchSize	即上述参数说明中的batchSize。
数据存储JSON路径	即上述参数说明中的dataPath。
Header	即上述参数说明中的customHeader。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

脚本配置示例如下所示。

```
{
    "type":"job",
    "version":"2.0",
    "steps":[
        {
            "stepType":"stream",
            "parameter":{
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"restapi",
            "parameter":{
                "url":"http://127.0.0.1:5000/writer1",
                "dataMode":"oneData",
                "responseType":"json",
                "column":[
                    {
                        "type":"long", //放置column数据到路径a.b
                        "name":"a.b"
                    },
                    {
                        "type":"string", //放置column数据到路径a.c
                        "name":"a.c"
                    }
                ],
                "method":"post",
                "defaultHeader":{
                    "X-Custom-Header":"test header"
                },
                "customHeader":{
                    "X-Custom-Header2":"test header2"
                },
                "parameters":"abc=1&amp;def=1",
                "batchSize":256
            },
            "name":"restapiwriter",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":"0" //错误记录数。
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.39. SAP HANA WRITER

本文为您介绍SAP HANA Writer支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

SAP HANA Writer插件实现了写入数据至SAP HANA数据库目标表的功能。在底层实现上，SAP HANA Writer通过JDBC连接远程SAP HANA数据库，并执行相应的 `insert into` 或 `replace into` 语句，写入数据至SAP HANA。数据库本身采用InnoDB引擎，以分批次提交数据入库。

SAP HANA Writer作为数据迁移工具，为数据库管理员等用户提供服务。根据您配置的writeMode，通过数据同步框架获取Reader生成的协议数据。

> **说明** 整个任务必须具备 `insert/replace into` 的权限。您可以根据配置任务时，在preSql和postSql中指定的语句，判断是否需要其它权限。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔，例如 `"column": ["id", "name", "age"]` 。 如果要依次写入全部列，使用星号（*）表示，例如 `"column": ["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，执行前清空表中的旧数据： `truncate table tablename` 。 > **说明** 当有多条SQL语句时，不支持SQL事务原子性。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，加上某一个时间戳： `alter table tablenameadd colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP` 。 > **说明**	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与数据源的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1024*

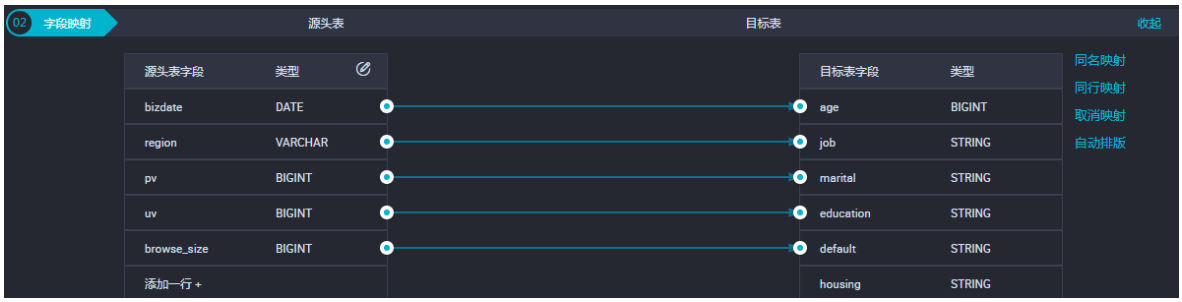
配置示例：向导模式

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。

参数	描述
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

脚本配置示例如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"saphana",//插件名。
            "parameter":{
                "postSql":[],//导入后的准备语句。
                "datasource":"",//数据源。
                "column":[//列名。
                    "id",
                    "value"
                ],
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "preSql":[
                     "delete from XXX;" //导入前的准备语句。
                   ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{//错误记录数。
            "record":"0"
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时，表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.40. KingbaseES Writer

本文为您介绍KingbaseES Writer支持的数据类型、字段映射和数据源等参数及配置示例。

背景信息

KingbaseES Writer插件实现了写入数据至KingbaseES数据库目标表的功能。在底层实现上，KingbaseES Writer通过JDBC连接远程KingbaseES数据库，并执行相应的 `insert into` 或 `replace into` 语句，写入数据至KingbaseES。数据库本身采用InnoDB引擎，以分批次提交数据入库。

KingbaseES Writer作为数据迁移工具，为数据库管理员等用户提供服务。根据您配置的writeMode，通过数据同步框架获取Reader生成的协议数据。

> **说明** 整个任务必须具备 `insert/replace into` 的权限。您可以根据配置任务时，在preSql和postSql中指定的语句，判断是否需要其它权限。

参数说明

参数	描述	是否必选	默认值
datasource	数据源名称，脚本模式支持添加数据源，此配置项填写的内容必须与添加的数据源名称保持一致。	是	无
table	选取的需要同步的表名称。	是	无
column	目标表需要写入数据的字段，字段之间用英文所逗号分隔，例如 `"column": ["id", "name", "age"]` 。 如果要依次写入全部列，使用星号（*）表示，例如 `"column": ["*"]` 。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，执行前清空表中的旧数据： `truncate table tablename` 。 > **说明** 当有多条SQL语句时，不支持SQL事务原子性。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前向导模式仅允许执行一条SQL语句，脚本模式可以支持多条SQL语句。例如，加上某一个时间戳： `alter table tablenameadd colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP` 。 > **说明**	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据同步系统与数据源的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程OOM异常。	否	*1024*

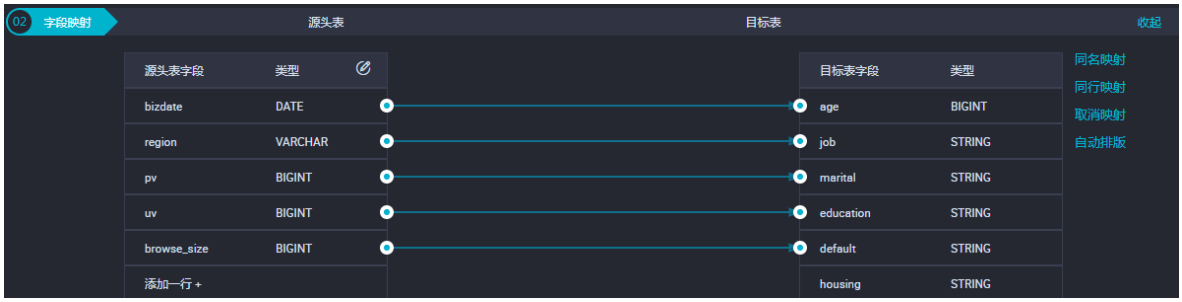
配置示例：向导模式

1. 选择数据源。

配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。



参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。

参数	描述
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	暂不支持，敬请期待。

配置示例：脚本模式

脚本配置示例如下所示。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"kingbasees",//插件名。
            "parameter":{
                "postSql":[],//导入后的准备语句。
                "datasource":"",//数据源。
                "column":[//列名。
                    "id",
                    "value"
                ],
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "preSql":[
                     "delete from XXX;" //导入前的准备语句。
                   ]
            },
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{//错误记录数。
            "record":"0"
        },
        "speed":{
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":1, //作业并发数。
            "mbps":"12"//限流
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

6.3.41. DM Writer

本文为您介绍DM Writer支持的数据类型、写入方式、字段映射和数据源等参数及配置示例。

> 注意 DM Writer仅支持使用新增和使用独享数据集成资源组，不支持使用公共资源组和自定义资源组。

背景信息

DM Writer插件实现了写入数据至DM主库的目的表的功能。在底层实现上，DM Writer通过数据集成框架获取Reader生成的协议数据，通过JDBC连接远程DM数据库，并执行相应的 `insert into...` 的SQL语句，将数据写入DM。DM Writer是一个通用的关系数据库写插件，您可以通过注册数据库驱动等方式，增加任意多样的关系数据库写支持。

DM Writer面向ETL开发工程师，通过DM Writer从数仓导入数据至DM。同时DM Writer也可以作为数据迁移工具，为数据库管理员等用户提供服务。

目前DM Writer支持数字、字符等大部分通用的关系数据库类型，但也存在部分类型没有支持的情况，请注意检查您的数据类型。

参数说明

参数	描述	是否必选	默认值
jdbcUrl	描述的是到对端数据库的JDBC连接信息，JDBCUrl按照DM官方规范，并可填写连接附件控制信息。请注意不同的数据库JDBC的格式不同，数据集成会根据具体JDBC的格式选择合适的数据库驱动写入数据： • DM格式：`jdbc:dm://ip:port/database` • DB2格式：`jdbc:db2://ip:port/database` • PPAS格式：`jdbc:edb://ip:port/database`	是	无
username	数据源的用户名。	是	无
password	数据源指定用户名的密码。	是	无
table	目标表名称，如果表的Schema信息和上述配置username不一致，请使用 `schema.table` 的格式填写table信息。	是	无
column	所配置的表中需要同步的列名集合。以英文逗号（,）进行分隔。 说明 建议您不要使用默认列情况。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句，目前只允许执行一条SQL语句，例如清除旧数据。 说明 当有多条SQL语句时，不支持事务。	否	无
postSql	执行数据同步任务之后执行的SQL语句，目前只允许执行一条SQL语句，例如加上某一个时间戳。 说明 当有多条SQL语句时，不支持事务。	否	无
batchSize	一次性批量提交的记录数大小，该值可以极大减少数据集成与DM（达梦）的网络交互次数，并提升整体吞吐量。但是该值设置过大可能会造成数据集成运行进程OOM情况。	否	*1024*

功能说明

配置一个写入DM的作业，通过脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "oracle",
            "parameter": {
                "datasource": "aaa",
                "column": [
                    "PROD_ID",
                  "name"
                ],
                "where": "",
                "splitPk": "",
                "encoding": "UTF-8",
                "table": "PENGXI.SALES"
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "rdbms",
            "parameter": {
                "connection": [
                    {
                        "jdbcUrl": "jdbc:dm://ip:port/database",
                        "table": [
                            "table"
                        ]
                    }
                ],
                "username": "username",
                "password": "password",
                "table": "table",
                "column": [
                   "id",
                  "name"
                ],
                "preSql": [
                    "delete from XXX;"
                ]
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "version": "2.0",
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": ""
        },
        "speed": {
            "throttle":true,//当throttle值为false时，mbps参数不生效，表示不限流；当throttle值为true时,表示限流。
            "concurrent":2, //作业并发数。
            "mbps":"12"//限流
        }
    }
}
```

```
```

DM Writer增加新的数据库支持的操作如下：

1. 进入DM Writer对应目录，此处*${DATAX_HOME}*为数据集成主目录，即*${DATAX_HOME}/plugin/writer/RDBMS Writer*。
2. 在DM Writer插件目录下有*plugin.json*配置文件，在此文件中注册您具体的数据库驱动，具体放在drivers数组中。DM Writer插件在任务执行时，会动态选择合适的数据库驱动连接数据库。

```
{
    "name": "RDBMS Writer",
    "class": "com.alibaba.datax.plugin.reader.RDBMS Writer.RDBMS Writer",
    "description": "useScene: prod. mechanism: Jdbc connection using the database, execute select sql, re
trieve data from the ResultSet. warn: The more you know about the database, the less problems you encount
er.",
    "developer": "alibaba",
    "drivers": [
        "dm.jdbc.driver.DmDriver",
        "com.ibm.db2.jcc.DB2Driver",
        "com.sybase.jdbc3.jdbc.SybDriver",
        "com.edb.Driver"
    ]
}
```

3. 在DM Writer插件目录下有libs子目录，您需要将您具体的数据库驱动放至*libs*目录下。

```
$tree
.
|-- libs
|   |-- Dm7JdbcDriver16.jar
|   |-- commons-collections-3.0.jar
|   |-- commons-io-2.4.jar
|   |-- commons-lang3-3.3.2.jar
|   |-- commons-math3-3.1.1.jar
|   |-- datax-common-0.0.1-SNAPSHOT.jar
|   |-- datax-service-face-1.0.23-20160120.024328-1.jar
|   |-- db2jcc4.jar
|   |-- druid-1.0.15.jar
|   |-- edb-jdbc16.jar
|   |-- fastjson-1.1.46.sec01.jar
|   |-- guava-r05.jar
|   |-- hamcrest-core-1.3.jar
|   |-- jconn3-1.0.0-SNAPSHOT.jar
|   |-- logback-classic-1.0.13.jar
|   |-- logback-core-1.0.13.jar
|   |-- plugin-rdbms-util-0.0.1-SNAPSHOT.jar
|   `-- slf4j-api-1.7.10.jar
|-- plugin.json
|-- plugin_job_template.json
`-- RDBMS Writer-0.0.1-SNAPSHOT.jar
```

6.3.42. StarRocks Writer

本文为您介绍StarRocks Writer支持的数据类型、字段映射和数据源等参数及配置示例。

使用限制

支持EMR-StarRocks 2.1版本。

实现原理

StarRocks Writer插件实现了写入数据至StarRocks数据库目标表的功能。采用StreamLoad CSV方式进行批量导入。

支持的数据类型

StarRocks Writer支持大部分StarRocks类型，包括数值类型、字符串类型、日期类型。

参数说明

参数	描述	是否必选	默认值
datasource	StarRocks数据源名称。	是	无
selectedDatabase	StarRocks数据库名称。	否	StarRocks数据源内配置的数据库名称。
loadProps	StarRocks StreamLoad请求参数。使用StreamLoad CSV导入，此处可选择配置导入参数。如果无特殊配置则使用{}。可配置参数包括： • column_separator：CSV导入列分隔符，默认\t。 • row_delimiter：CSV导行分隔符，默认\n。 • 如果您的数据中本身包含\t、\n，则需自定义使用其他字符作为分隔符，使用特殊字符示例如下： `{ "column_separator": "\\x01", "row_delimiter": "\\x02"}`	是	无
column	所配置的表中需要同步的列名集合。	是	无
loadUrl	填写StarRocks FrontEnd IP、Http Port（一般默认是8030），如果有多个FrontEnd节点，可全部配置上，并使用逗号（，）分隔。	是	无
table	选取的需要同步的表名称。	是	无
preSql	执行数据同步任务之前率先执行的SQL语句。例如，执行前清空表中的旧数据（truncate table tablename）。	否	无
postSql	执行数据同步任务之后执行的SQL语句。	否	无

向导开发介绍

打开新建的数据同步节点，即可进行同步任务的配置，详情请参见通过向导模式配置离线同步任务。

您需要在数据同步任务的编辑页面进行以下配置：

1. 选择数据源。

   配置同步任务的**数据来源**和**数据去向**。

参数	描述
数据源	即上述参数说明中的datasource，通常填写您配置的数据源名称。
数据库	即上述参数说明中的selectedDatabase，选择要读取的数据库名，为数据源JDBC中配置的数据库。
表	即上述参数说明中的table。
导入前准备语句	即上述参数说明中的preSql，输入执行数据同步任务之前率先执行的SQL语句。
导入后完成语句	即上述参数说明中的postSql，输入执行数据同步任务之后执行的SQL语句。
LoadUrls	即上述参数说明中的loadUrl，填写FE IP，Http Port（一般默认是8030），如果有多个FE节点，可全部配置上，使用逗号（，）分隔。
StreamLoad请求参数	即上述参数说明中的loadProps，使用StreamLoad CSV导入，此处可选择配置导入参数。如果无特殊配置则使用{}。可配置参数包括： ○ column_separator：CSV导入列分隔符，默认\t。 ○ row_delimiter：CSV导行分隔符，默认\n。 ○ 如果您的数据中本身包含\t、\n，则需自定义使用其他字符作为分隔符，使用特殊字符示例如下： `{ "column_separator": "\\x01", "row_delimiter": "\\x02"}`

2. 字段映射，即上述参数说明中的column。左侧的源头表字段和右侧的目标表字段为一一对应的关系。

参数	描述
同名映射	单击**同名映射**，可以根据名称建立相应的映射关系，请注意匹配数据类型。
同行映射	单击**同行映射**，可以在同行建立相应的映射关系，请注意匹配数据类型。
取消映射	单击**取消映射**，可以取消建立的映射关系。
自动排版	可以根据相应的规律自动排版。
手动编辑源表字段	请手动编辑字段，一行表示一个字段，首尾空行会被采用，其他空行会被忽略。
添加一行	单击**添加一行**，您可以输入以下类型的字段： ○ 可以输入常量，输入的值需要使用英文单引号，如'abc'、'123'等。 ○ 可以配合调度参数使用，例如${bizdate}等。 ○ 可以输入关系数据库支持的函数，例如now()、count(1)等。 ○ 如果您输入的值无法解析，则类型显示为未识别。

3. 通道控制。

参数	描述
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
分布式处理能力	数据同步时，可以将任务切片分散到多台执行节点上并发执行，提高同步速率。该模式下，配置较大任务并发数会增加数据存储访问压力，如需使用该功能，请提前评估数据存储的访问负载。该功能仅支持在独享数据集成资源组配置，详情请参见独享数据集成资源组概述和新增和使用独享数据集成资源组。

脚本开发介绍

脚本配置示例如下，使用脚本模式开发的详情请参见通过脚本模式配置离线同步任务。

脚本配置样例如下所示，具体参数填写请参见参数说明。

```
{
    "stepType": "starrocks",
    "parameter": {
        "selectedDatabase": "didb1",
        "loadProps": {
            "row_delimiter": "\\x02",
            "column_separator": "\\x01"
        },
        "datasource": "starrocks_public",
        "column": [
            "id",
            "name"
        ],
        "loadUrl": [
            "1.1.1.1:8030"
        ],
        "table": "table1",
        "preSql": [
            "truncate table table1"
        ],
        "postSql": [
        ]
    },
    "name": "Writer",
    "category": "writer"
}
```