阿里云
DataWorks

数据汇聚

文档版本：20191209

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. **阿里云文档中所有内容，包括但不限于图片、架构设计、页面布局、文字描述，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
	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。	 **禁止:** 重置操作将丢失用户配置数据。
	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。
	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。
	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。	执行`cd /d C:/window`命令，进入Windows系统文件夹。
*## *	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

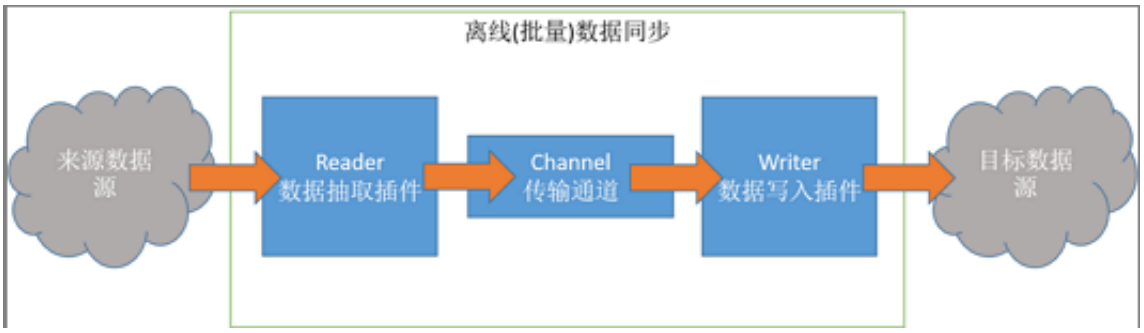
1 数据集成

1.1 数据集成概述

数据集成是阿里集团对外提供的稳定高效、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数据高速稳定的数据移动及同步能力。

离线（批量）数据同步简介

离线（批量）的数据通道主要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（称之为Reader）、数据写入插件（称之为Writer），并基于此框架设计一套简化版的中间数据传输格式，从而达到任意结构化、半结构化数据源之间数据传输的目的。



支持的数据源类型

数据集成提供丰富的数据源支持，如下所示。

- **文本存储（FTP/SFTP/OSS/多媒体文件等）。**
- **数据库（RDS/DRDS/MySQL/PostgreSQL等）。**
- **NoSQL（Memcache/Redis/MongoDB/HBase等）。**
- **大数据（MaxCompute/AnalyticDB/HDFS等）。**
- **MPP数据库（HybridDB for MySQL等）。**

更多详情请参见 支持的数据源。

说明:
由于每个数据源的配置信息差距较大，需要根据使用情况详细查询参数配置信息。所以在数据源配置、作业配置页面提供了详细描述，请您根据自身情况进行查询使用。

同步开发说明

同步开发提供向导模式和脚本模式两种开发模式。

- 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，帮助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，但无法享受到一些高级功能。
- 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，做精细化的配置管理。

说明：

- 向导模式生成的代码可以转换为脚本模式，此转换为单向操作，转换完成后无法恢复到向导模式，因为脚本模式能力是向导模式的超集。
- 代码编写前需要完成数据源的配置和目标表的创建。

网络类型说明

网络类型分为经典网络、专有网络（VPC）和本地IDC网络（规划中）。

- 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，更适合对网络易用性要求比较高的客户。
- 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，以及配置路由表和网关。
- 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。

补充说明：

- 网络连接可以支持公网连接，网络类型选择经典网络即可。需要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。
- 规划中的网络连接，进行数据同步，可以使用本地新增运行资源+脚本模式的方案进行数据同步传输。您也可以使用Shell+DataX方案。
- 专有网络VPC是构建一个隔离的网络环境，可以自定义IP地址范围、网段、网关等。随着专有网络安全性提高，专有网络运用越来越广，所以数据集成提供了RDS-MySQL、RDS-SQL Server、RDS-PostgreSQL，在专有网络下不需要购买一台和VPC同网络的ECS，系统通过反向代理会自动检测从而网络能够互通。对于阿里云其他的数据库PPAS、OceanBase、Redis、MongoDB、Memcache、Tabl eStore、HBase等，后续也会提供支持。所以非RDS的数据源在专有网络下配置数据集成的同步任务需要购买同网络的ECS，这样可以通过ECS连通网络。

约束与限制

- 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）的数据同步。也就是说，数据集成支持传输能够抽象为逻辑

二维表的数据同步，其他完全非结构化数据，例如OSS中存放的一段MP3，数据集成暂不支持将其同步到MaxCompute，这个功能会在后期实现。

- 支持单个和部分跨Region地域内数据存储相互同步、交换的数据同步需求。

  部分地域通过经典网络是可以传输的，但不能保证。如果必须使用且测试经典网络不通，可以考虑使用公网方式连接。
- 仅完成数据同步（传输），本身不提供数据流的消费方式。

参考文档

- 数据同步任务配置的详细介绍请参见**创建数据同步任务**。
- 如果处理像OSS等非结构化数据的详细介绍请参见*MaxCompute***访问***OSS***数据**。

1.2 基本概念

本文将为您介绍并发数、限速、脏数据和数据源等基本概念。

并发数

并发数是数据同步任务内，可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最高能达到的传输速度。

脏数据

脏数据对于业务没有意义或者格式非法的数据。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。

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。

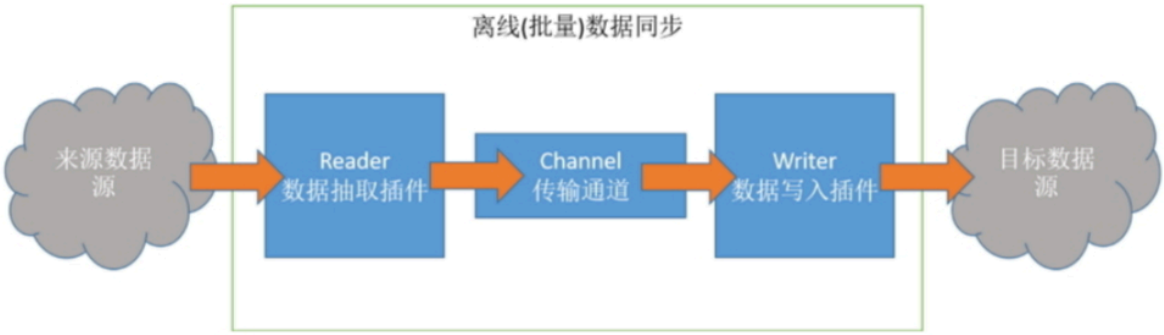
1.3 创建数据集成任务

本文为您介绍创建数据集成任务的流程和操作步骤。

数据集成是阿里巴巴集团对外提供的可跨异构数据存储系统、可靠、安全、低成本、可弹性扩展的数据同步平台，为数据源提供不同网络环境下的全量/增量数据进出通道。

Reader插件通过远程连接数据库，并执行相应的SQL语句，将数据从数据库中Select出来，从底层实现了从数据库读取数据。

Writer插件通过远程连接数据库，并执行相应的SQL语句，将数据写入数据库，从底层实现了向数据库写入数据。



数据集成任务准备工作

创建阿里云账号

1. 开通阿里云主账号，并创建账号的访问密钥，即AccessKeys。
2. 开通MaxCompute，自动产生一个默认的MaxCompute数据源，并使用主账号登录DataWorks。
3. 创建工作空间。您可在工作空间中协作完成工作流，共同维护数据和任务等，因此使用DataWorks前需要先创建工作空间。

> **说明:**
> 如果您想通过子账号创建数据集成任务，可以赋予其相应的权限。详情请参见准备RAM子账号。

创建源端和目标端数据库和表

1. 您可以使用建表语句或直接通过客户端建表，不同的数据源库创建数据库和表，请参见相应数据库的官方文档进行创建。
2. 给相关数据库和表赋予读写的权限。

> **说明:**
> 通常至少需要赋予Reader端读的权限，赋予Writer端增、删、改的权限，建议提前赋予数据库中的表足够的权限。

数据集成任务操作步骤

创建数据源

1. 从数据库获取相关的数据源信息。
2. 在界面配置相关的数据源。

> **说明:**

- 界面配置数据源只支持一部分，如果在界面找不到相关的配置数据源界面可以直接脚本模式配置，将相关的数据源信息填写在JSON脚本中。
- 支持数据源的情况，请参见支持的数据源类型。
- 如何配置数据源和注意细节请参见数据源配置。

创建自定义资源组（可选）

1. 创建自定义资源组，详情请参见资源组。
2. 添加服务器。
3. 安装Agent。
4. 检查连通性。

说明:

- 网络环境不通或者DataWorks提供的资源满足不了任务运行条件的情况下，您可以选择添加自定义资源组。
- 建议无论是专有网络还是经典网络，都选择专有网络的添加形式。
- 配置自定义资源组的方式，请参见新增调度资源。
- 最佳实践
  - （仅一端不通）数据源网络不通的情况下的数据同步。
  - （两端都不通）数据源网络不通的情况下的数据同步。

配置数据集成任务

1. 配置同步任务的读取端，每个Reader插件的配置细节请参见配置*Reader*插件。
2. 配置同步任务的写入端，每个Writer插件的配置细节请参见配置*Writer*插件。
3. 配置同步任务读写端的映射关系。
4. 配置同通道控制，您可以在该步骤切换相关的资源组。

说明:

- 您可以通过向导模式和脚本模式配置同步任务，详情请参见向导模式配置和脚本模式配置 。
- 配置任务时，您可以对您的任务进行速度调优，详情请参见优化配置。
- 向导模式可以转换成脚本模式，脚本模式不能转换成向导模式，我们已为您提供全部插件的模板。

运行数据集成任务

1. 您可以直接在界面运行数据集成任务，日志不会保存。

2. 提交之前需要进行调度配置，提交后一般第二天产生实例。详情请参见调度配置模块的文档。

> 说明:
>
> - 您配置任务时可以设置相关调度参数。
> - 测试同步任务时，不能直接调用调度配置中的参数，您需要提交后，才可以自动调用调度配置中配置的参数。

查看运行日志

您可以进入运维中心页面，查看运行结果。

> 说明:
>
> - 您可以进入运维中心找到DAG图，右键查看运行日志。
> - 在同步任务是幂等可自动重跑的前提下，如果您的任务运行失败，可以配置调度重跑，这样失败的任务可以自动重跑，增加系统稳定性。

1.4 支持的数据源

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 MySQL 2.0和OSS等）提供离线、批量数据的进出通道。

数据同步支持的数据源类型如下表所示。

数据源分类	数据源类型	抽取（Reader）	导入（Writer）	支持方式	支持类型
关系型数据库	*MySQL*	支持，详情请参见配置*MySQL Reader*	支持，详情请参见配置*MySQL Writer*	向导/脚本	阿里云/自建
	SQL Server	支持，详情请参见配置*SQL Server Reader*	支持，详情请参见配置*SQL Server Writer*	向导/脚本	阿里云/自建
	PostgreSQL	支持，详情请参见配置*PostgreSQL Reader*	支持，详情请参见配置*PostgreSQL Writer*	向导/脚本	阿里云/自建
	Oracle	支持，详情请参见配置*Oracle Reader*	支持，详情请参见配置*Oracle Writer*	向导/脚本	自建
	DM	支持	支持	脚本	自建

数据源分类	数据源类型	抽取（Reader）	导入（Writer）	支持方式	支持类型
	DRDS	支持，详情请参见*配置DRDS Reader*	支持，详情请参见*配置DRDS Writer*	向导/脚本	阿里云
	POLARDB	支持，详情请参见*配置POLARDB Reader*	支持，详情请参见*配置POLARDB Writer*	向导/脚本	阿里云
	HybridDB for MySQL	支持，详情请参见*配置HybridDB for MySQL Reader*	支持，详情请参见*配置HybridDB for MySQL Writer*	向导/脚本	阿里云
	AnalyticDB for PostgreSQL	支持，详情请参见*配置AnalyticDB for PostgreSQL Reader*	支持，详情请参见*配置AnalyticDB for PostgreSQL Writer*	向导/脚本	阿里云
	DB2	支持，详情请参见*配置DB2 Reader*	支持，详情请参见*配置DB2 Writer*	脚本	自建
	RDS for PPAS	支持	支持	脚本	阿里云
大数据存储	*MaxCompute*	支持，详情请参见*配置MaxCompute Reader*	支持，详情请参见*配置MaxCompute Writer*	向导/脚本	阿里云
	DataHub	不支持	支持，详情请参见*配置DataHub Writer*	脚本	阿里云
	AnalyticDB for MySQL 2.0	支持，详情请参见*配置AnalyticDB for MySQL 2.0 Reader*	支持，详情请参见*配置AnalyticDB for MySQL 2.0 Writer*	向导/脚本	阿里云
	Elasticsearch	支持，详情请参见*配置Elasticsearch Reader*	支持，详情请参见*配置Elasticsearch Writer*	脚本	阿里云
非结构化存储	*OSS*	支持，详情请参见*配置OSS Reader*	支持，详情请参见*配置OSS Writer*	向导/脚本	阿里云
	HDFS	支持，详情请参见*配置HDFS数据源*	支持，详情请参见*配置HDFS Writer*	脚本	自建

数据源分类	数据源类型	抽取（Reader）	导入（Writer）	支持方式	支持类型
	FTP	支持，详情请参见配置FTP Reader	支持，详情请参见配置HDFS Writer	向导/脚本	自建
NoSQL	MongoDB	支持，详情请参见配置MongoDB数据源	支持，详情请参见配置MongoDB Writer	脚本	阿里云/自建
	Memcache（OCS）	不支持	支持，详情请参见配置Memcache（OCS） Writer	脚本	阿里云/自建 Memcached
	Redis	不支持	支持，详情请参见配置Redis Writer	脚本	阿里云/自建
	Table Store（OTS）	支持，详情请参见配置Table Store（OTS） Reader	支持，详情请参见配置Table Store（OTS） Writer	脚本	阿里云
	OpenSearch	不支持	支持，详情请参见配置OpenSearch Writer	脚本	阿里云
	HBase	支持，详情请参见配置HBase Reader	支持，详情请参见配置HBase Writer	脚本	阿里云/自建
消息队列	LogHub	支持，详情请参见配置LogHub Reader	支持，详情请参见配置LogHub Writer	向导/脚本	阿里云
性能测试	Stream	支持，详情请参见配置Stream Reader	支持，详情请参见配置Stream Writer	脚本	-
其它数据源	SAP HANA	支持，详情请参见配置SAP HANA Reader	不支持	脚本	第三方

1.5 数据源测试连通性

本文将为您介绍支持连通性测试的数据源类型，以及数据源连通性测试常见问题示例。

数据源	数据源类型	网络类型	是否支持测试连通性	是否添加自定义资源组
MySQL	云数据库	经典网络	支持	-
		专有网络	支持	-
	连接串模式（数据集成网络可直接连通）		支持	-
	连接串模式（数据集成网络不可直接连通）		不支持	是
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
SQL Server	云数据库	经典网络	支持	-
		专有网络	支持	-
	连接串模式（数据集成网络可直接连通）		支持	-
	连接串模式（数据集成网络不可直接连通）		不支持	是
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
PostgreSQL	云数据库	经典网络	支持	-
		专有网络	支持	-
	连接串模式（数据集成网络可直接连通）		支持	-
	连接串模式（数据集成网络不可直接连通）		不支持	是
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
Oracle	连接串模式（数据集成网络可直接连通）		支持	-
	连接串模式（数据集成网络不可直接连通）		不支持	是
	ECS自建	经典网络	支持	-

数据源	数据源类型	网络类型	是否支持测试连通性	是否添加自定义资源组
		专有网络	不支持	是
DRDS	云数据库	经典网络	支持	-
		专有网络	排期中	是
HybridDB for MySQL	云数据库	经典网络	支持	-
		专有网络	排期中	是
HybridDB for PostgreSQL	云数据库	经典网络	支持	-
		专有网络	排期中	是
MaxCompute（对应odps数据源）	云数据库	经典网络	支持	-
AnalyticDB AnalyticDB for MySQL 2.0	云数据库	经典网络	支持	-
		专有网络	排期中	是
OSS	云数据库	经典网络	支持	-
		专有网络	支持	-
Hdfs	连接串模式（数据集成网络可直接连通）		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
FTP	连接串模式（数据集成网络可直接连通）		支持	-
	连接串模式（数据集成网络不可直接连通）		不支持	是
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
MongoDB	云数据库	经典网络	支持	-
		专有网络	排期中	是
	连接串模式（数据集成网络可直接连通）		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	是

数据源	数据源类型	网络类型	是否支持测试连通性	是否添加自定义资源组
Memcache	云数据库	经典网络	支持	-
		专有网络	排期中	是
Redis	云数据库	经典网络	支持	-
		专有网络	排期中	是
	连接串模式（数据集成网络可直接连通）		支持	-
	ECS自建	经典网络	支持	-
		专有网络	不支持	是
Table Store（对应OTS数据源）	云数据库	经典网络	支持	-
		专有网络	排期中	是
DataHub	云数据库	经典网络	支持	-
		专有网络	不支持	-

说明:

- 添加自定义资源组的详情，请参见新增任务资源。
- 使用独享资源组的详情，请参见*#unique_85*。

上述表格中的-表示没有该说法。不支持并不代表不能配置同步任务，只是单击测试连通性无效，需要添加自定义资源组。

以关系数据库JDBCUrl为例，约束如下：

- 本地IDC
  - 连接串模式（数据集成网络可直接连通）：支持测试连通性和通过JDBCUrl模式添加数据源。
  - 连接串模式（数据集成网络不可直接连通）：不支持测试连通性。支持通过JDBCUrl模式添加数据源，需要使用数据集成自定义资源组同步任务。

    您也可以使用高速通道打通本地IDC网络和已有的专有网络，并提交工单。

- ECS自建数据源
  - 连接串模式（数据集成网络可直接连通）：支持测试连通性和通过JDBCUrl模式添加数据源。
  - 经典网络：
    - 如果和DataWorks在相同的区域，支持测试连通性和通过JDBCUrl模式添加数据源。
    - 如果和DataWorks在不同的区域，则不支持测试连通性。支持通过JDBCUrl模式添加数据源，需要使用数据集成自定义资源组同步任务。
    - 经典网络ECS上自建的数据源，不保证默认资源组网络可通，建议使用数据集成自定义资源组同步任务。
  - 专有网络VPC内部地址：不支持测试连通性。支持通过JDBCUrl模式添加数据源，需要使用数据集成自定义资源组或独享数据集成资源同步任务。
- 阿里云产品：
  - 实例模式添加数据源：
    - RDS（MySQL、PostgreSQL和SQLServer）、POLARDB、DRDS、HybridDB for MySQL、AnalyticDB for PostgreSQLr和AnalyticDB for MySQL3.0支持反向VPC，支持测试连通性和使用默认资源组同步任务。
    - Redis、MongoDB和AnalyticDB for MySQL2.0支持实例模式添加数据源，不支持反向VPC和测试连通性，需要使用数据集成自定义资源组或独享数据集成资源同步任务。
  - 连接串模式（数据集成网络可直接连通）：支持测试连通性和通过JDBCUrl模式添加数据源。
  - 经典网络：
    - 如果和DataWorks在相同的区域，支持测试连通性和通过JDBCUrl模式添加数据源。
    - 如果和DataWorks在不同的区域，则不支持测试连通性。支持通过JDBCUrl模式添加数据源，需要使用数据集成自定义资源组同步任务。
  - 专有网络VPC内部地址：不支持测试连通性。支持通过JDBCUrl模式添加数据源，需要使用数据集成自定义资源组或独享数据集成资源同步任务。

  例如MaxCompute、OSS和LogHub等其它区域中心化服务的产品，包括3种类型的endpoint，您按照自身需求进行选择即可。

说明:

- HDFS、Redis和MongoDB等其它数据源对应连接地址的约束和关系型数据库一致。

- 选择数据源连接地址时，需要和任务配置模式（包括向导模式和脚本模式）、任务实际执行资源组（包括默认资源组、自定义资源组和独享数据集成资源组）配合，让运行任务的资源组可以访问数据源。
- 由于数据存储的特性，建议HBase数据源和HDFS数据源使用自定义资源组或独享数据集成资源组。
- 金融云的数据源支持网络连通和通过实例模式添加数据源。如果网络不通，请使用自定义资源组同步任务。

调度集群

- 目前调度集群在华东2、华南1、中国（香港）、新加坡均有部署，以调度集群在华东2为准和用户数据源进行对比。假设您的MongoDB数据源在华北经典网络，以调度集群在华东2经典网络为准，跨区域连接不通。
- OXS集群和ECS集群内网不通。

  RDS的调度集群是OXS，OXS集群和内网中国大陆所有区域的RDS互通。其它数据源由另外一套ECS经典网络的调度集群调度。

  例如RDS同步至自建数据库测试时，RDS和自建数据库数据源测试连通性均可以成功。但实际调度时，RDS会下发至OXS调度集群，自建数据库会下发至ECS集群，RDS和ECS集群不通，所以测试失败。通常建议您将RDS改为MySQL>JDBC方式，以保证都可以调度ECS集群，网络连接成功。

如何查看任务下发执行集群

- **出现RDS作为数据源时，任务会到OXS集群同步。**

您可以通过以下方式确认任务运行的资源组：

- **任务运行在默认的资源组上，日志中会出现如下信息。**

```
running in Pipeline[basecommon_ group_xxxxxxxxx]
```

- **任务运行在数据集成自定义资源组上，日志中会出现如下信息。**

```
running in Pipeline[basecommon_xxxxxxxxx]
```

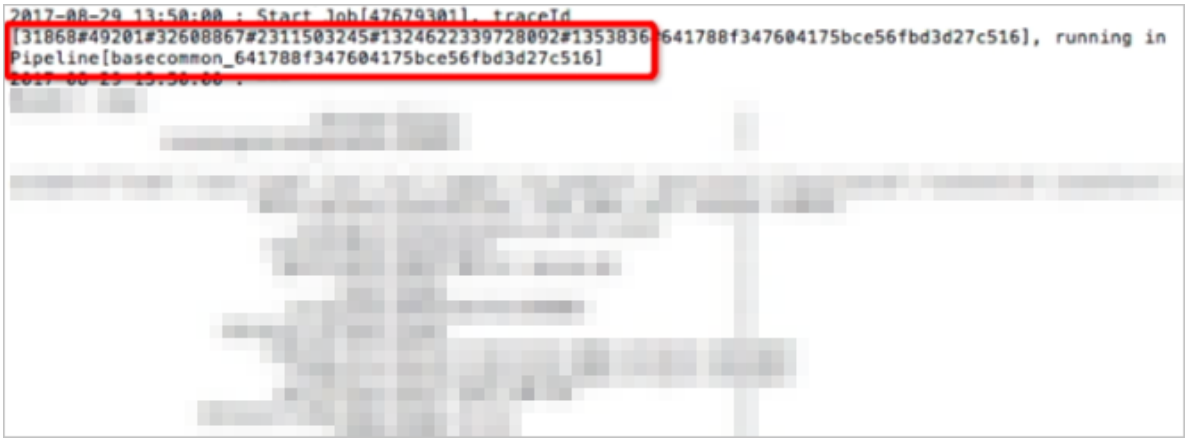
- **任务运行在独享数据集成资源上，日志中会出现如下信息。**

```
running in Pipeline[basecommon_S_res_group_xxx]
```

如果需要切换资源组，请参见*#unique_85/unique_85_Connect_42_section_442_xzp_r7r*。

- **当数据源为其它数据源在ECS调度集群时。**

- 当调度集群为自定义调度资源时，日志如下图所示（非常重要，用于判断是否为自定义资源组）。



- 进入数据集成测试页面，直接单击运行，统一由ECS调度集群调度。因为RDS作为数据源跨区域时，需要在OXS调度集群执行。所以可能出现RDS相关任务手动运行成功，但调度失败的情况，此时需要您选择调度运维 > 测试运行。

测试连通性失败的常见场景

当测试连通性失败时，需要确认数据源区域、网络类型、RDS白名单是否添加完整实例ID、数据库名称和用户名是否正确。如果您的测试连通性失败，可以首先参见及如何排查数据集成问题进行排查。常见错误示例如下所示：

- 数据库密码错误，如下所示。

- 网络不通错误，如下所示。

  ```
  "com.mysql.jdbc.exceptions.jdbc4.CommunicationsException:
  Communications link failure
  ```

- 同步过程中出现网络断开等情况。

  首先要查看完整日志，确定是哪个调度资源，是否是自定义资源。

  如果是自定义资源，请确认自定义资源组的IP是否添加至相应数据源的白名单，例如RDS白名单。

  > 说明:
  > MongoDB有白名单限制，需要进行添加。

  确认两端是否通过数据源连通性、RDS和MongoDB的白名单是否添加完整。

  > 说明:
  > 如果白名单不完整，则不能保证任务成功。如果任务下发至已经添加的调度服务器上会成功，没添加的会失败。

- 任务显示成功，但是日志出现8000断开报错。

  出现上述报错，是因为您使用的自定义调度资源组，没有对10.116.134.123的访问8000端口在安全组内网入方向放行，添加后重新运行即可。

测试连通性失败的示例

示例一

- 问题现象

  测试连接失败，测试数据源连通性失败。连接数据库失败，数据库连接串：`jdbc:mysql://xx.xx.xx.x:xxxx/t_uoer_bradef`，用户名：xxxx_test，异常消息：`Access denied for user 'xxxx_test'@'%' to database 'yyyy_demo'`。

- 排查思路

  1. 确认其添加的信息是否有问题。
  2. 密码、白名单或者用户的账号是否具有对应数据库的权限，RDS管控台可以添加授权。

示例二

- 问题现象

  测试连接失败，测试数据源连通性失败，报错如下。

  ```
  error message: Timed out after 5000 ms while waiting for a server
  that matches ReadPreferenceServerSelector{readPreference=primary}.
  Client view of cluster state is {type=UNKNOWN, servers=[(xxxxxxxxxx
  ), type=UNKNOWN, state=CONNECTING, exception={com.mongodb.MongoSocke
  tReadException: Prematurely reached end of stream}}]
  ```

- 排查思路

  非VPC的MongoDB，添加MongoDB数据源测试连通性要添加相应的白名单，详情请参见添加白名单。

1.6 数据源配置

1.6.1 数据源隔离

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。

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

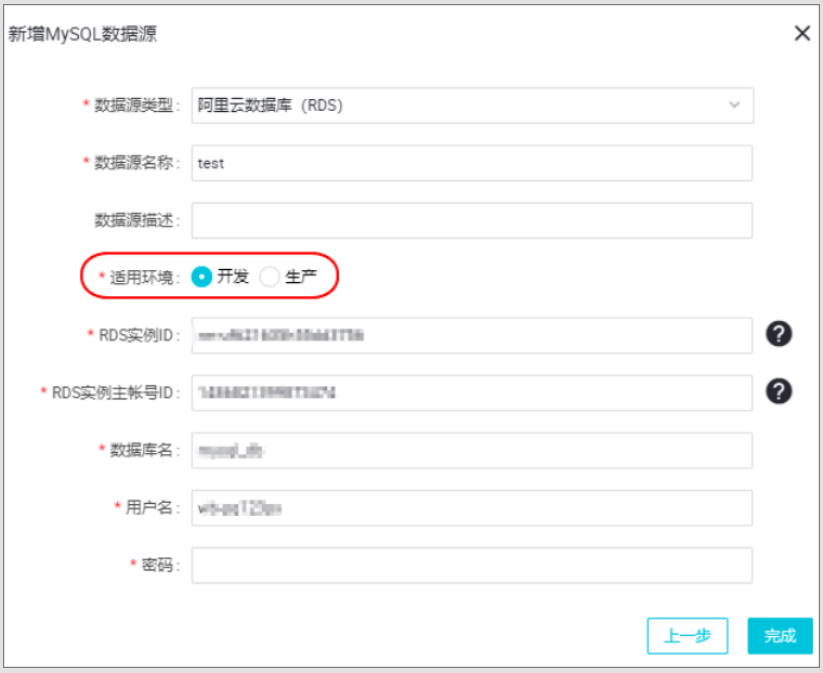
> 说明:
> 目前只有标准模式的工作空间支持数据源隔离。

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。

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，详情请参见数据源配置。
- 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。

页面功能	说明
多库多表搬迁	单击多库多表搬迁，可直接跳转至批量上云页面。 **说明：** 必须确保生产环境和开发环境都存在数据源，且数据源测试连通性成功，方可在批量上云页面选择相应的数据源。
批量新增数据源	目前仅支持MySQL、SQLServer和Oracle数据源。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，您可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。

页面功能	说明
新增数据源	· 开发环境可用的数据源：可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相同。 
适用环境	简单模式下的工作空间不显示此列，标准模式下的工作空间，用以展现数据源适用的环境。

页面功能	说明
操作	· 整库迁移批量配置：该按钮仅对开发环境的数据源显示。 · 新建：若不存在适用环境下的数据源，显示新建按钮。 · 编辑/删除：若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。 若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。 若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。

1.6.2 配置AnalyticDB for MySQL 2.0数据源

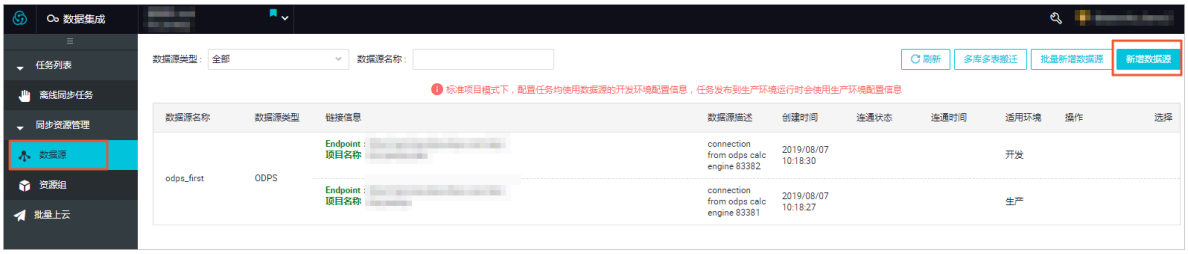
AnalyticDB for MySQL 2.0为您提供其他数据源向AnalyticDB for MySQL 2.0写入的功能，但不能读取数据，支持数据集成中的向导模式和脚本模式。

> **说明:**
> 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. 以项目管理员身份进入*DataWorks*管理控制台，单击相应工作空间操作栏中的进入数据集成。

2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新建数据源弹出框中，选择数据源类型为AnalyticDB for MySQL 2.0。
4. 填写AnalyticDB for MySQL 2.0数据源的各配置项。

配置	说明
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此配置。

配置	说明
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/ AceessKey Secret	访问密钥（AccessKey ID和AccessKey Secret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性功能，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置AnalyticDB for MySQL 2.0数据源，您可以继续学习下一个教程。在该教程中您将学习如何配置AnalyticDB for MySQL 2.0插件，详情请参见配置*AnalyticDB for MySQL 2.0 Reader*和配置*AnalyticDB for MySQL 2.0 Writer*。

1.6.3 配置SQL Server数据源

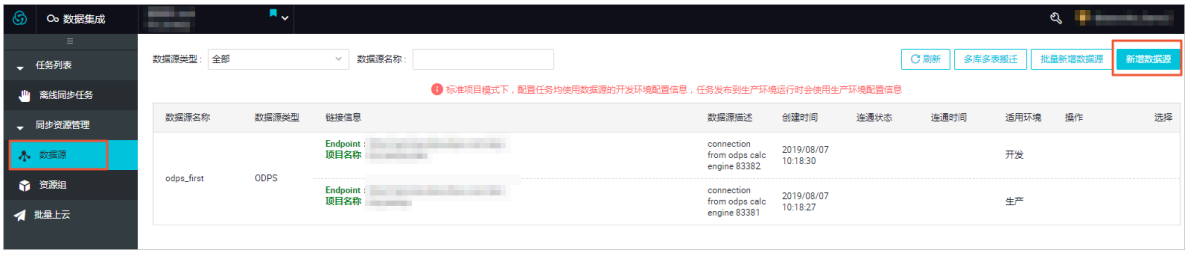
SQL Server数据源为您提供读取和写入SQL Server双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

> 说明:
>
> 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。
>
> 目前仅支持SQL Server 2005及以上版本。如果是VPC环境下的SQL Server，需要注意以下问题：
>
> - 自建的SQL Server数据源。
>   - 不支持测试连通性，但仍支持配置同步任务，创建数据源时，直接单击完成即可。
>   - 必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。
> - 通过RDS创建的SQL Server数据源。
>
>   您无需选择网络环境，系统会根据您填写的RDS实例信息，自动进行判断。

操作步骤

1. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。

2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为SQL Server。

4. 填写SQL Server数据源的各配置项。

SQL Server数据源类型包括阿里云数据库（RDS）、连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通），您可以根据自身需求进行选择。

- 以新增SQL Server > 阿里云数据库（RDS）类型的数据源为例。

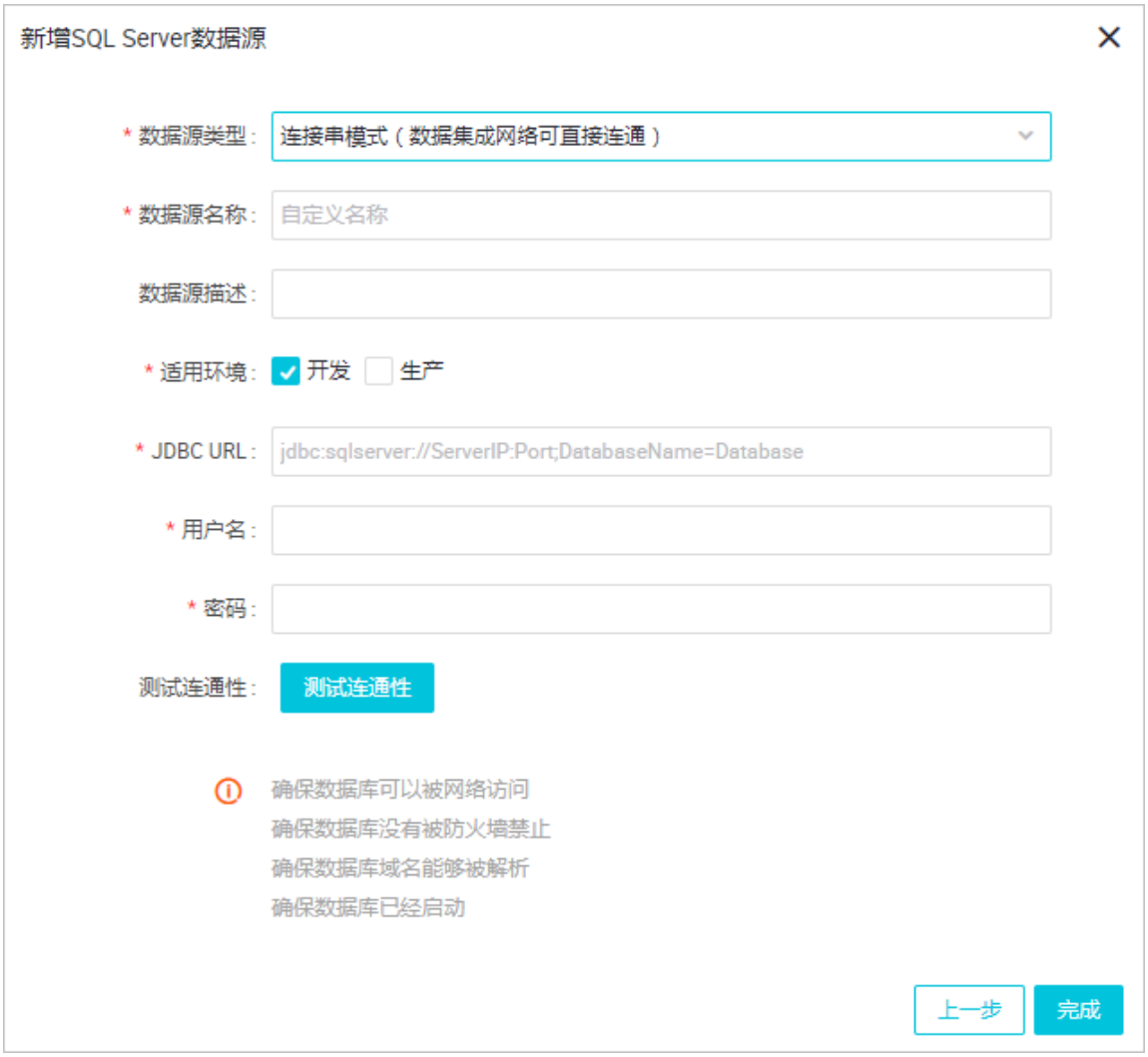
配置	说明
数据源类型	当前选择的数据源类型为SQL Server > 阿里云数据库（RDS）。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。
数据源描述	对数据源进行简单描述，不得超过80个字符。

配置	说明
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此配置。
地区	选择相应的Region。
RDS实例ID	您可以进入RDS的控制台，查看RDS的实例ID。
RDS实例主账号ID	输入购买RDS实例的主账号ID。
数据库名	填写对应的数据库名称。
用户名/密码	数据库对应的用户名和密码。

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SQL Server > 连接串模式（数据集成网络可直接连通）类型的数据源为例。**



配置	说明
数据源类型	当前选择的数据源类型为SQL Server > 连接串模式（数据集成网络可直接连通）。
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此配置。
JDBC URL	JDBC连接信息，格式为`jdbc:sqlserver://ServerIP:Port;DatabaseName=Database`。

配置	说明
用户名/密码	数据库对应的用户名和密码。

- 以新增SQL Server > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为SQL Server > 连接串模式（数据集成网络不可直接连通）。
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。

配置	说明
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此配置。
资源组	可以用于执行同步任务，通常添加资源组时可以绑定多台机器。详情请参见新增任务资源。
JDBC URL	JDBC连接信息，格式为jdbc:sqlserver://ServerIP:Port;DatabaseName=Database。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络/专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置SQL Server数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置SQL Server插件，详情请参见 配置*SQL Server Writer*和配置*SQL Server Reader*。

1.6.4 配置MongoDB数据源

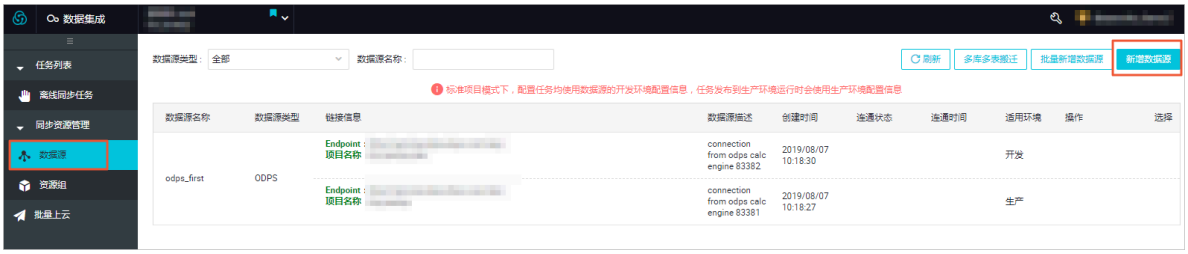
MongoDB是目前仅次于Oracle、MySQL的文档型数据库，为您提供读取和写入MongoDB双向通道的功能，您可以通过脚本模式配置同步任务。

说明:
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. 以项目管理员身份进入*DataWorks*管理控制台，单击相应工作空间操作栏中的进入数据集成。

2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为MongoDB。

4. 填写MongoDB数据源的各配置项。

MongoDB数据源类型包括实例模式（阿里云数据源）和连接串模式（数据集成网络可直接连通）。

- 实例模式（阿里云数据源）：通常使用经典网络类型，同地区的经典网络可以连通，跨地区的经典网络不保证可以连通。
- 连接串模式（数据集成网络可直接连通）：通常使用公网类型，可能产生一定的费用。

以新增MongDB > 实例模式（阿里云数据源）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为MongDB > 实例模式（阿里云数据源）。 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，然后刷新此页面。
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此配置。
地区	是指在购买MongoDB时所选择的区域。
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。

配置	说明
用户名/密码	数据库对应的用户名和密码。

以新增MongDB > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为MongDB > 连接串模式（数据集成网络可直接连通）。
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。

配置	说明
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此配置。
访问地址	格式为host:port。如果此处您需要同时添加多个地址，可以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、私网地址混杂。
数据库名	该数据源对应的数据库名称。
用户名/密码	数据库对应的用户名和密码。

说明:

连接串模式（数据集成网络不可直接连通）的数据库，可以通过下述操作添加MongoDB数据源。

a. 选择数据源类型为连接串模式（数据集成网络可直接连通）。

b. 填写新增MongoDB数据源对话框中的配置项，其中访问地址填写您的内网地址。

c. 添加完成后，不需要进行连通性测试，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，详情请参见新增任务资源。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

说明:

- VPC环境的MongoDB云数据库，添加连接串模式（数据集成网络可直接连通）数据源类型并保存。
- VPC环境不支持测试连通性。

后续步骤

现在，您已经学习了如何配置MongoDB数据源，您可以继续学习下一个教程。在该教程中您将学习如何配置MongoDB插件。详情请参见配置*MongoDB Reader*和配置*MongoDB Writer*。

1.6.5 配置DataHub数据源

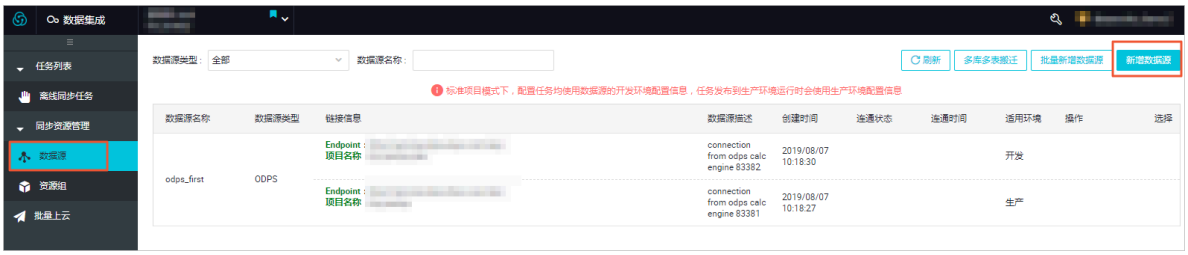
DataHub数据源作为数据中枢，为您提供从其它数据源写入数据至DataHub的功能，支持DataHub Writer插件。

DataHub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。

> **说明:**
> **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。**

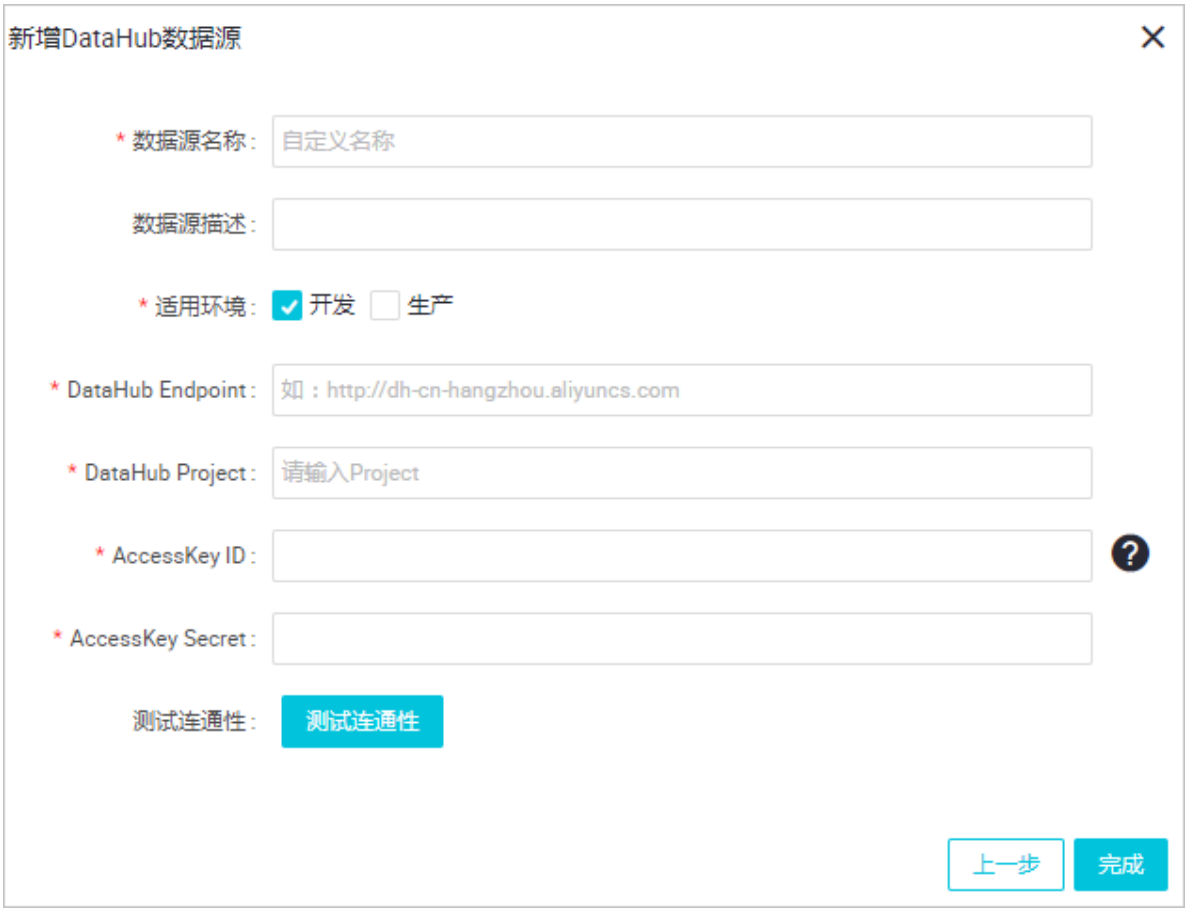
操作步骤

1. **以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。**
2. **选择同步资源管理 > 数据源，单击新增数据源。**



3. **在新增数据源弹出框中，选择数据源类型为DataHub。**

4. 填写DataHub数据源的各配置项。



配置	说明
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此配置。
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/ AceessKey Secret	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性功能，可以判断输入的信息是否正确 。

后续步骤

现在，您已经学习了如何配置DataHub数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置DataHub Writer插件，详情请参见配置*DataHub Writer*。

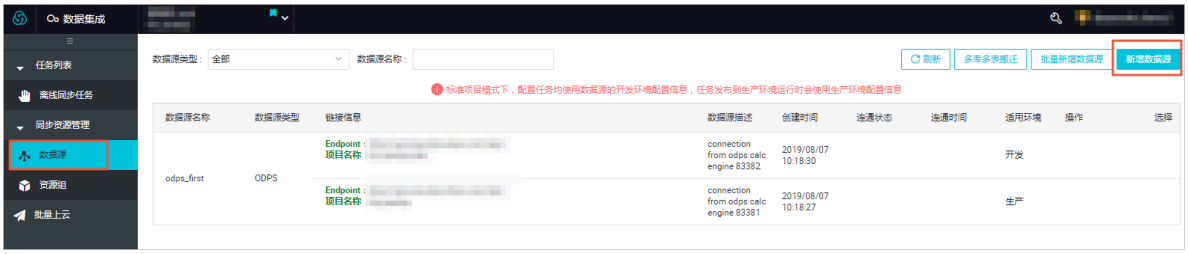
1.6.6 配置DM数据源

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。

> **说明:**
> 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. 以项目管理员身份进入*DataWorks*管理控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。

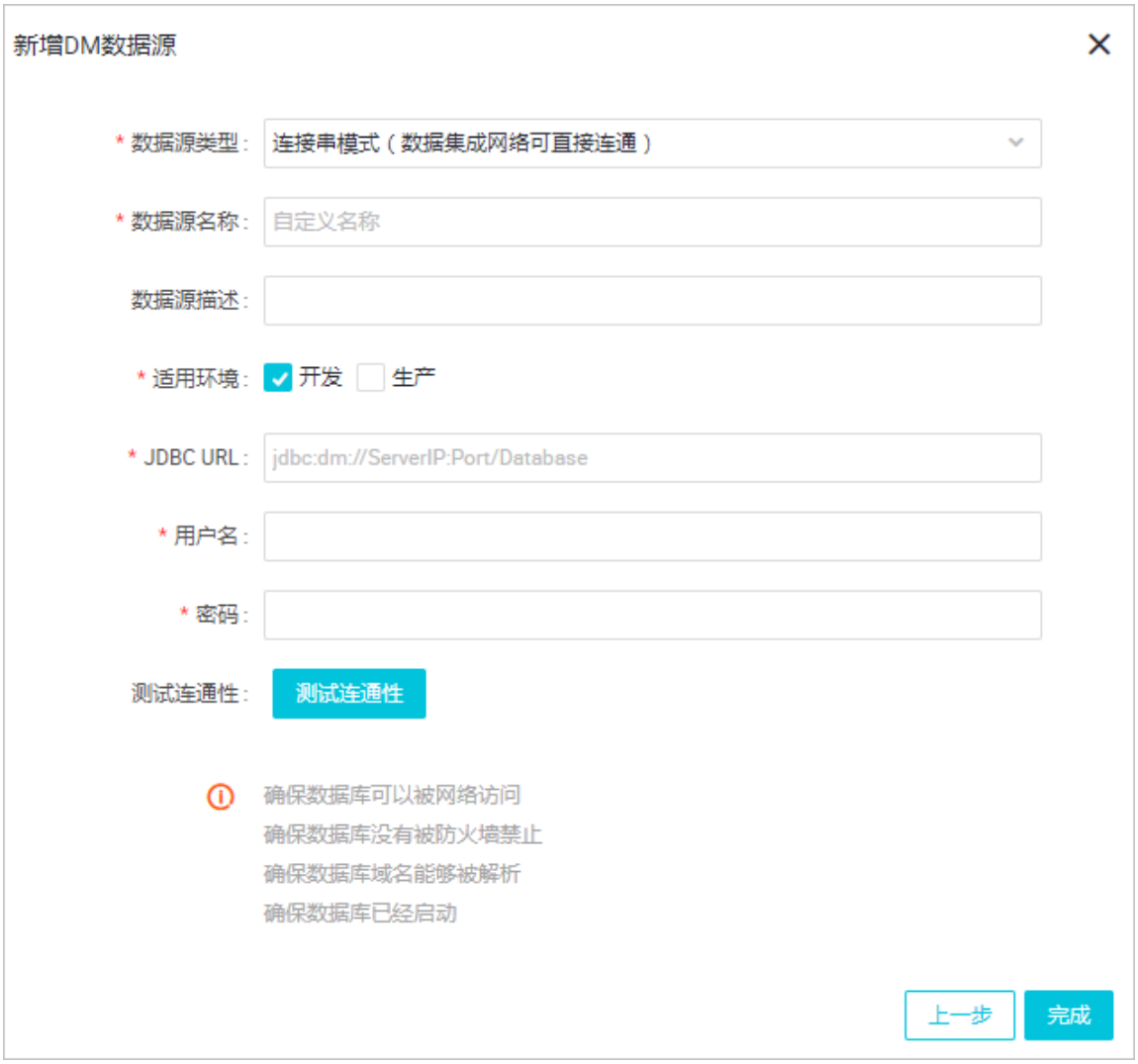


3. 在新增数据源弹出框中，选择数据源类型为DM。

4. 填写DM数据源的各配置项。

DM数据源类型包括连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通），您可以根据自身需求进行选择。

- 以新增DM > 连接串模式（数据集成网络可直接连通）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为DM > 连接串模式（数据集成网络可直接连通）。
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此配置。

配置	说明
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

- 以新增DM > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为DM > 连接串模式（数据集成网络不可直接连通）。选择此类型的数据源，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。

配置	说明
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此配置。
资源组	可以用于执行同步任务，通常添加资源组时可以绑定多台机器。详情请参见新增任务资源。
JDBC URL	JDBC连接信息，格式为jdbc:mysql://ServerIP:Port/Database。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性能力，可以判断输入的信息是否正确 。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络和连接串模式（数据集成网络不可直接连通）下，目前不支持数据源连通性测试，直接单击完成。

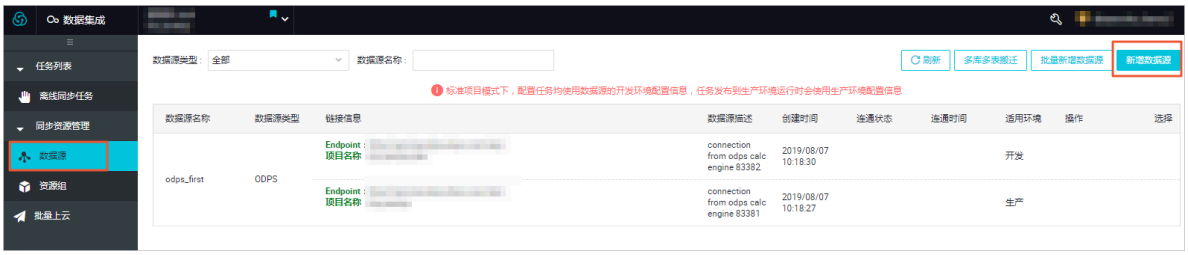
1.6.7 配置DRDS数据源

DRDS（分布式RDS）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。

说明:
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. 以项目管理员身份进入DataWorks控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为DRDS。
4. 填写DRDS数据源的各配置项。

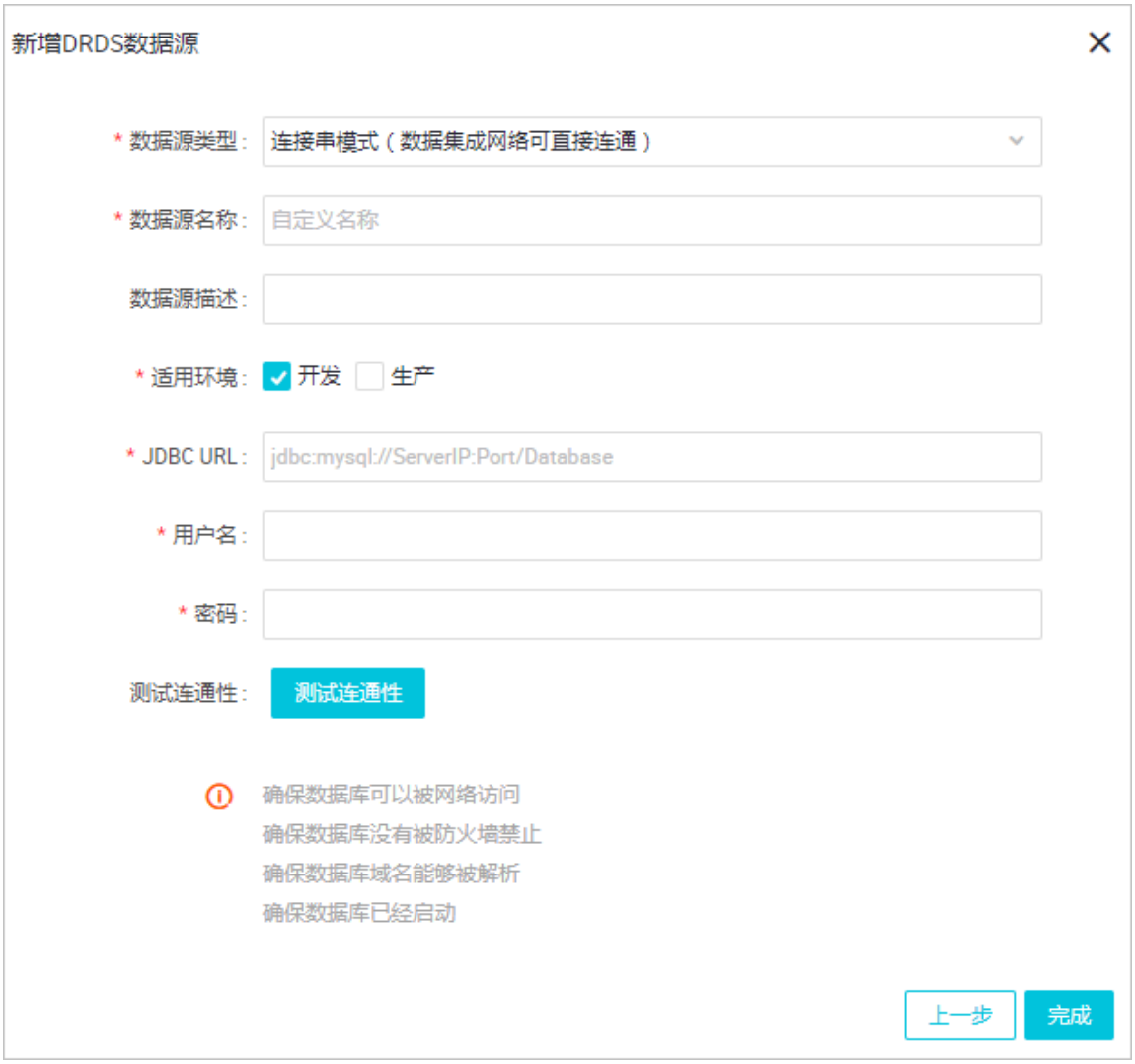
   DRDS数据源类型包括阿里云数据库（DRDS）和连接串模式（数据集成网络可直接连通），您可以根据自身需求进行选择。

   - 以新增DRDS > 阿里云数据库（DRDS）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为DRDS > 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。

配置	说明
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此配置。
实例ID	您可以进入DRDS控制台查看相关实例ID。
主账号ID	您可以进入DRDS控制台的安全设置页面，查看相应的信息。
数据库名	填写数据库对应的名称。
用户名/密码	数据库对应的用户名和密码。

- 以新增DRDS > 连接串模式（数据集成网络可直接连通）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为DRDS > 连接串模式（数据集成网络可直接连通）。

配置	说明
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此配置。
JDBC URL	JDBC连接信息，格式为`jdbc:sqlserver://ServerIP:Port;DatabaseName=Database`。
用户名/密码	数据库对应的用户名和密码。

说明:
对于连接串模式（数据集成网络可直接连通）的数据源，您需要添加白名单才能连接成功。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络或专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置DRDS数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置DRDS插件，详情请参见配置*DRDS Writer*和配置*DRDS Reader*。

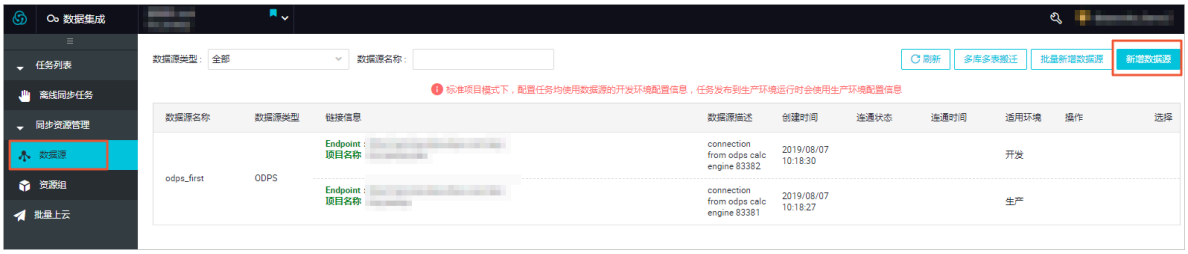
1.6.8 配置FTP数据源

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。

说明:
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. 以项目管理员身份进入*DataWorks*管理控制台，单击对应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。

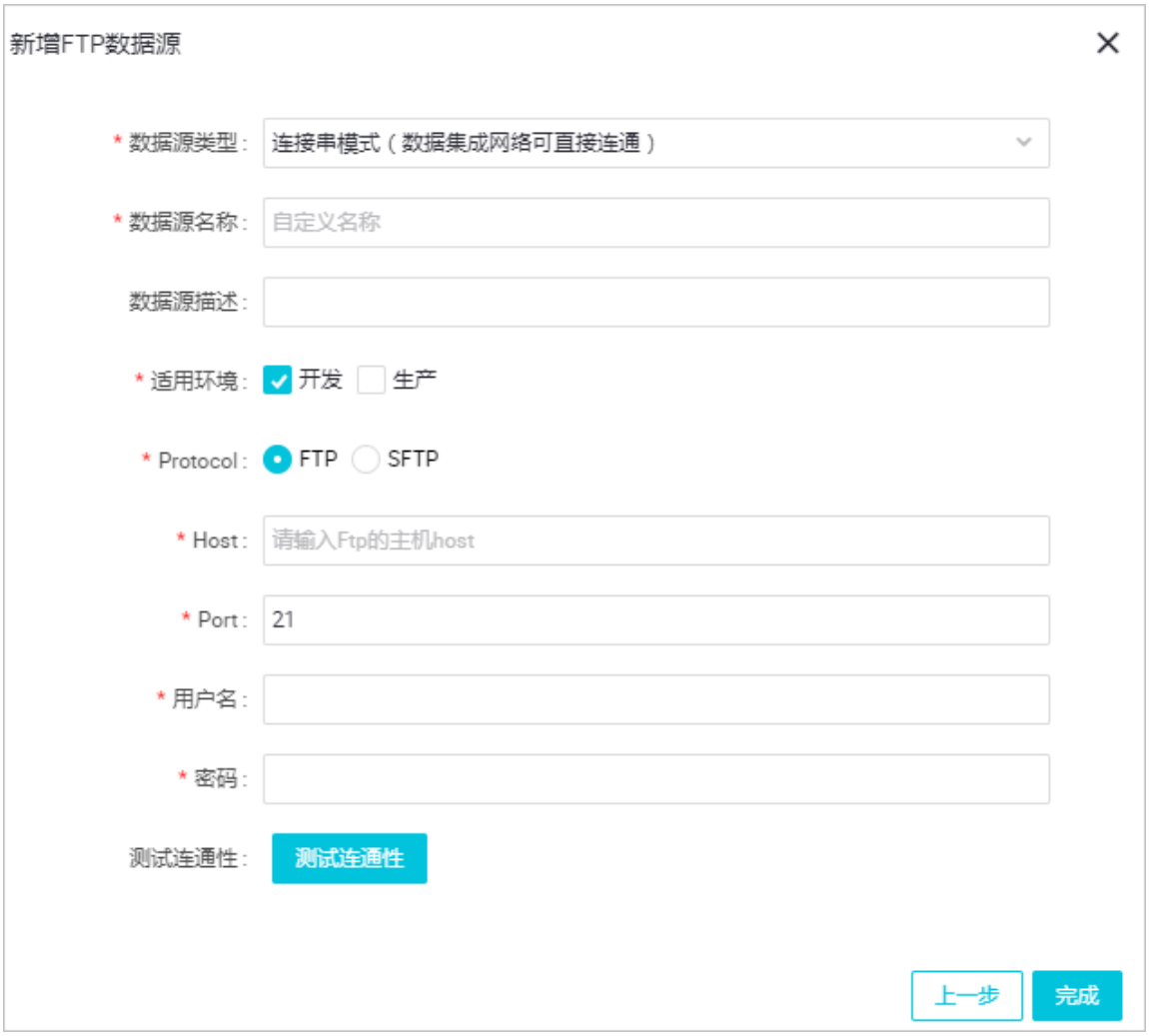


3. 在新增数据源弹出框中，选择数据源类型为FTP。

4. 填写FTP数据源的各配置项。

FTP数据源类型包括连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通），您可以根据自身需求进行选择。

- 以新增FTP > 连接串模式（数据集成网络可直接连通）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为FTP > 连接串模式（数据集成网络可直接连通）。
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此配置。

配置	说明
Portocol	目前仅支持FTP和SFTP协议。
Host	对应FTP主机的IP地址。
Port	如果选择的是FTP协议，则端口默认为21。如果选择的是SFTP协议，则端口默认为22。

配置	说明
用户名/密码	访问该FTP服务的账号密码。

- 以新增FTP > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为FTP > 连接串模式（数据集成网络不可直接连通）。 选择此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。

配置	说明
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此配置。
资源组	可以用于执行同步任务，通常添加资源组时可以绑定多台机器。详情请参见新增任务资源。
Portocol	目前仅支持FTP和SFTP协议。
Host	对应FTP主机的IP地址。
Port	如果选择的是FTP协议，则端口默认为21。如果选择的是SFTP协议，则端口默认为22。
用户名/密码	访问该FTP服务的账号密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性能力，可以判断输入的信息是否正确 。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络目前不支持数据源连通性测试，直接单击完成。

后续步骤

现在，您已经学习了如何配置FTP数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置FTP插件，详情请参见配置*FTP Reader* 和配置*FTP Writer*。

1.6.9 配置HDFS数据源

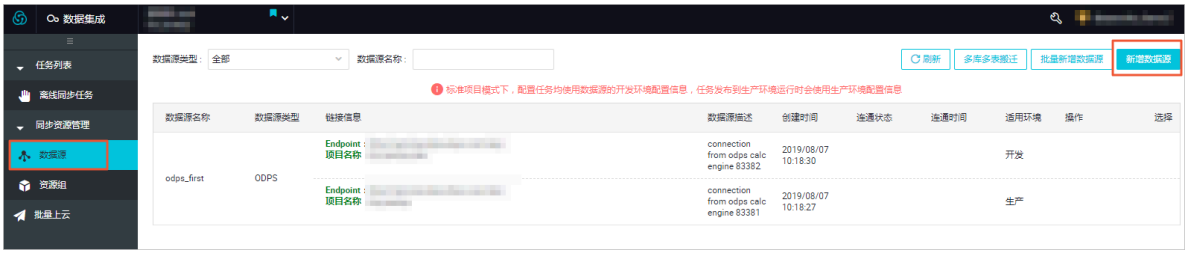
HDFS是一个分布式文件系统，它为您提供读取和写入HDFS双向通道的功能，您可以通过脚本模式配置同步任务。

说明:
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. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。

2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为HDFS。

4. 填写HDFS数据源的各配置项。

配置	说明
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此配置。
DefaultFS	nameNode节点地址，格式为`hdfs://ServerIP:Port`。

5. 单击测试连通性。

6. 测试连通性通过后，单击完成。

   提供测试连通性能力，可以判断输入的信息是否正确 。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络目前不支持数据源连通性测试，直接单击完成。

后续步骤

现在，您已经学习了如何配置HDFS数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置HDFS插件。详情请参见配置*HDFS Reader*和配置*HDFS Writer*。

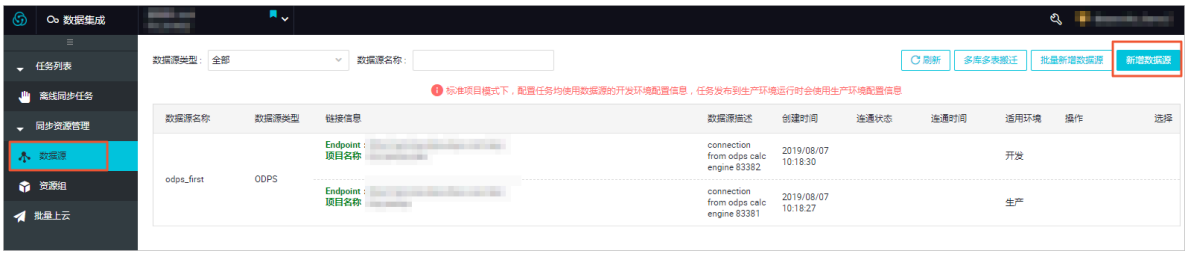
1.6.10 配置LogHub数据源

LogHub数据源作为数据中枢，为您提供读取和写入LogHub双向通道的功能，支持Reader和Writer插件。

> **说明:**
> 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。

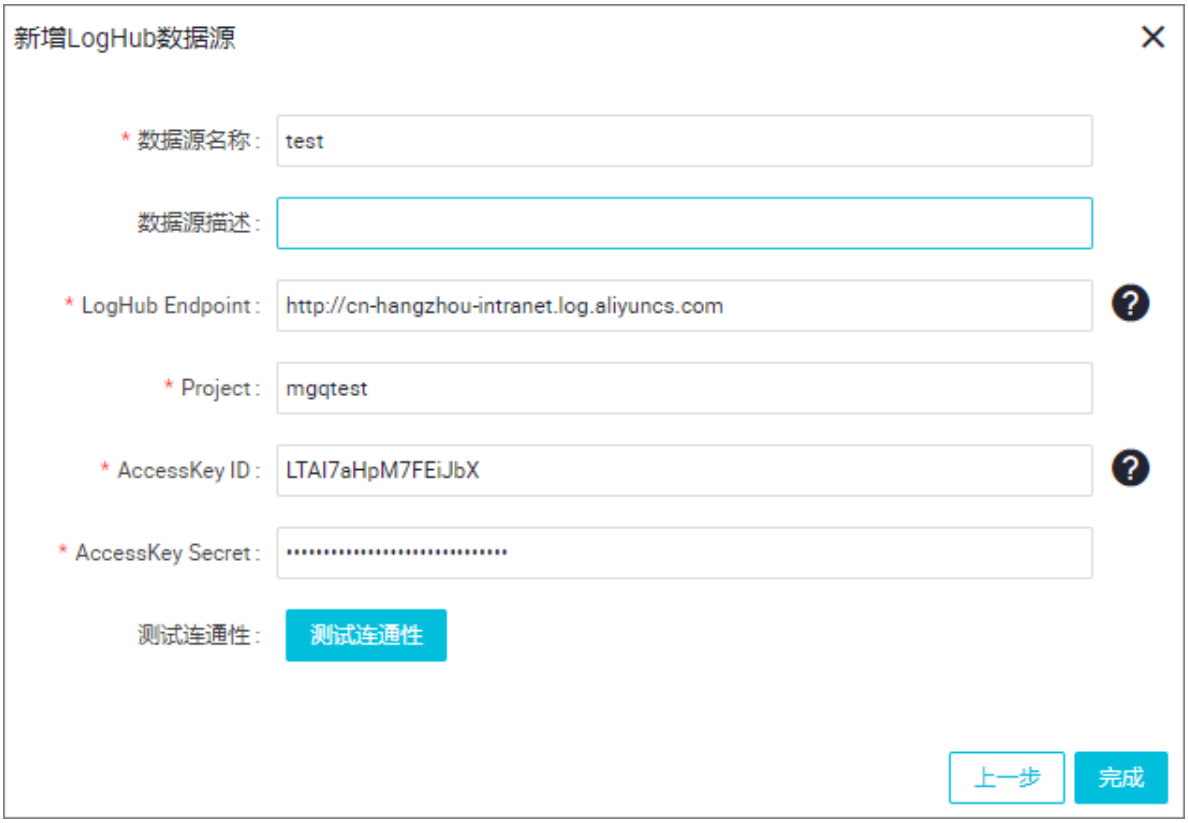
操作步骤

1. 以项目管理员身份登录*DataWorks*控制台，单击对应工作空间操作栏中的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为LogHub。

4. 填写LogHub数据源的各配置项。



配置	说明
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此配置。
LogHub Endpoint	通常格式为`http://cn-shanghai.log.aliyun.com`。详情请参见服务入口。
Project	输入对应的Project。
AccessID/AceessKey	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击确定。

   提供测试连通性功能，可以判断输入的信息是否正确 。

后续步骤

现在，您已经学习了如何配置LogHub数据源，您可以继续学习下一个教程。在该教程中您将学习如何配置LogHub插件，详情请参见配置*LogHub Reader*和配置*LogHub Writer*。

1.6.11 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，支持Reader和Writer插件。

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。

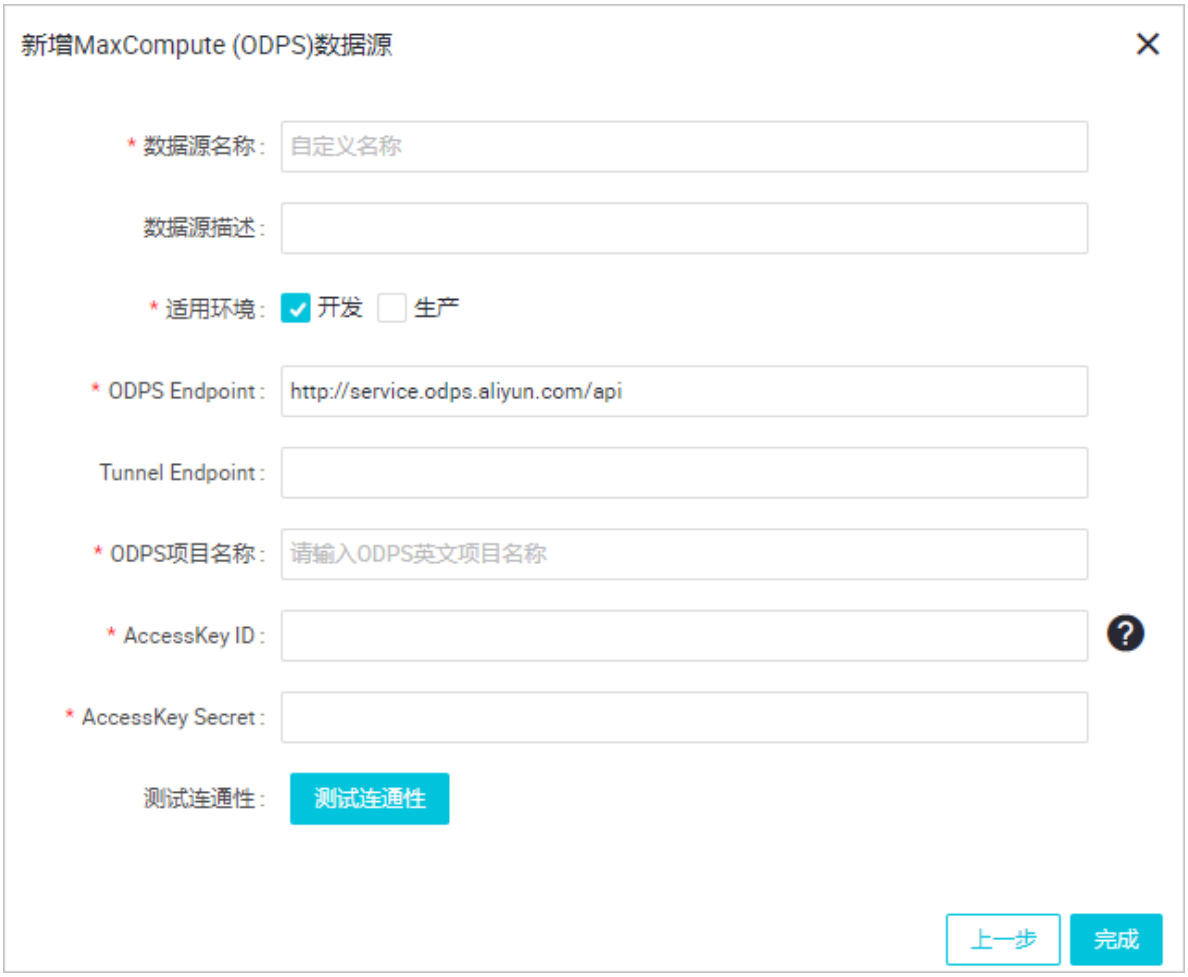
> **说明:**
>
> - 标准模式的工作空间支持数据源隔离功能，您可以分别添加开发环境和生产环境的数据源并进行隔离，以保护您的数据安全。
> - 每个项目空间系统都将生成一个默认的数据源（odps_first），对应的MaxCompute项目名称为当前项目空间对应的计算引擎MaxCompute项目名称。您可以单击右上方的用户信息，在修改AccessKey信息页面切换默认数据源的AK，但需要注意以下问题：
>   - 只能从主账号AK切换到主账号AK。
>   - 切换时当前必须没有任务在运行中（数据集成或数据开发等一切和DataWorks相关的任务），您自行添加的MaxCompute数据源可以使用子账号AK。

操作步骤

1. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。

3. 在新增数据源弹出框中，选择数据源类型为MaxCompute（ODPS）。

4. 填写MaxCompute数据源的各配置项。



配置	说明
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此配置。
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#unique_93。
ODPS项目名称	MaxCompute（ODPS）项目名称。
AccessID/AceessKey	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。

6. 测试连通性通过后，单击完成。

   提供测试连通性功能，可以判断输入的信息是否正确 。

后续步骤

现在，您已经学习了如何配置MaxCompute数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置MaxCompute插件。详情请参见配置*MaxCompute Reader*和配置*MaxCompute Writer*。

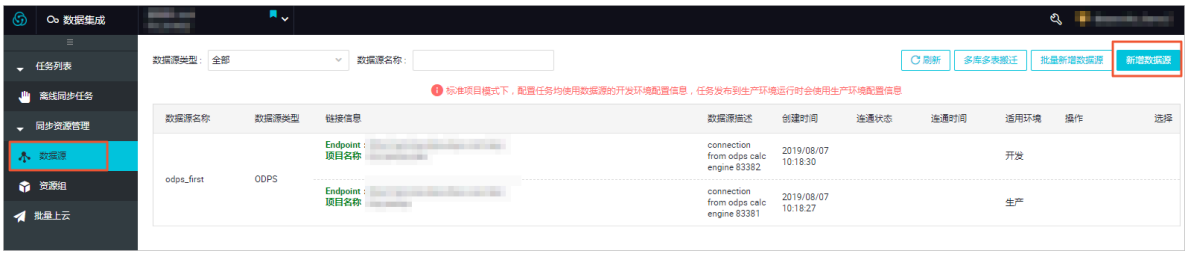
1.6.12 配置Memcached数据源

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。

> **说明:**
> 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. 以项目管理员身份登录*DataWorks*控制台，单击对应工作空间操作栏中的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为Memcached。

4. 填写Memcached数据源的各配置项。

配置	说明
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此配置。
Proxy Host	相应的Memcache Proxy。
Port	相应的Memcache端口，默认为11211。
用户名/密码	对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性功能，可以判断输入的各项信息是否正确 。

后续步骤

现在，您已经学习了如何配置Memcache数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置Memcache Writer插件，详情请参见配置Memcache（OCS） *Writer*。

1.6.13 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，您可以通过向导模式和脚本模式配置同步任务。

说明:

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。

如果是在VPC环境下的MySQL，需要注意以下问题：

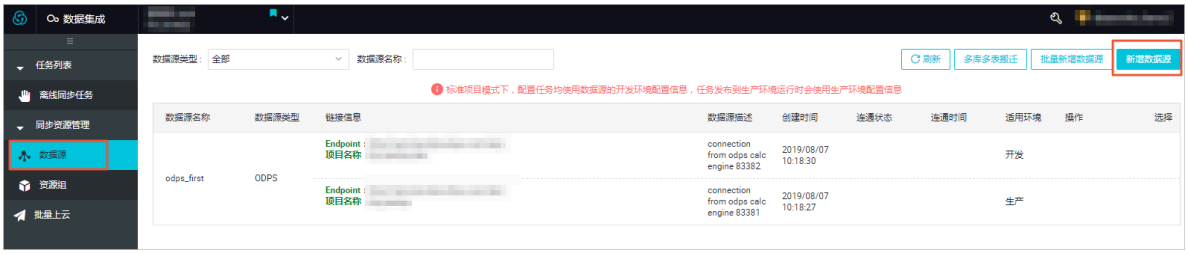
- 自建的MySQL数据源
  - 不支持测试连通性，但仍支持配置同步任务，创建数据源时单击完成即可。
  - 必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。
- 通过RDS创建的MySQL数据源

  您无需选择网络环境，系统会自动根据您填写的RDS实例信息进行判断。

目前DataWorks数据集成驱动无法直接支持MySQL 8.0版本。如果您使用MySQL 8.0，请新增任务资源，详情请参见*RDBMS*添加关系型数据库驱动最佳实践，配合配置*RDBMS Reader*及配置*RDBMS Writer*，完成与MySQL数据库的连接和读写。

操作步骤

1. 以项目管理员身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为MySQL。

4. 填写MySQL数据源的各配置项。

MySQL数据源类型包括阿里云数据库（RDS）、连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通）。

以新增MySQL > 阿里云数据库（RDS）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 阿里云数据库（RDS）。
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。

配置	说明
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此配置。
地区	选择相应的区域。
RDS实例ID	您可以进入RDS管控台，查看RDS的实例ID。
RDS实例主账号ID	输入购买RDS实例的主账号ID。
数据库名	填写对应的数据库名称。
用户名/密码	数据库对应的用户名和密码。

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MySQL > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 连接串模式（数据集成网络可直接连通）。
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此配置。
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。

配置	说明
用户名/密码	数据库对应的用户名和密码。

以新增MySQL > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

说明:

连接串模式（数据集成网络不可直接连通）的数据源不支持测试连通性。

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 连接串模式（数据集成网络不可直接连通）。

配置	说明
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此配置。
资源组	可以用于执行同步任务，通常添加资源组时可以绑定多台机器。详情请参见新增任务资源。
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络/专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置MySQL数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置MySQL插件，详情请参见配置*MySQL Reader*和配置*MySQL Writer*。

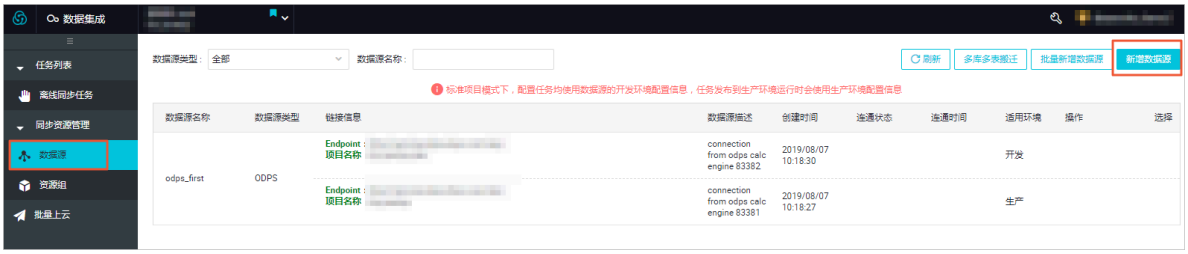
1.6.14 配置Oracle数据源

Oracle数据源为您提供读取和写入Oracle双向通道的功能，您可以通过向导模式和脚本模式配置同步任务。

说明:
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. 以项目管理员身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。

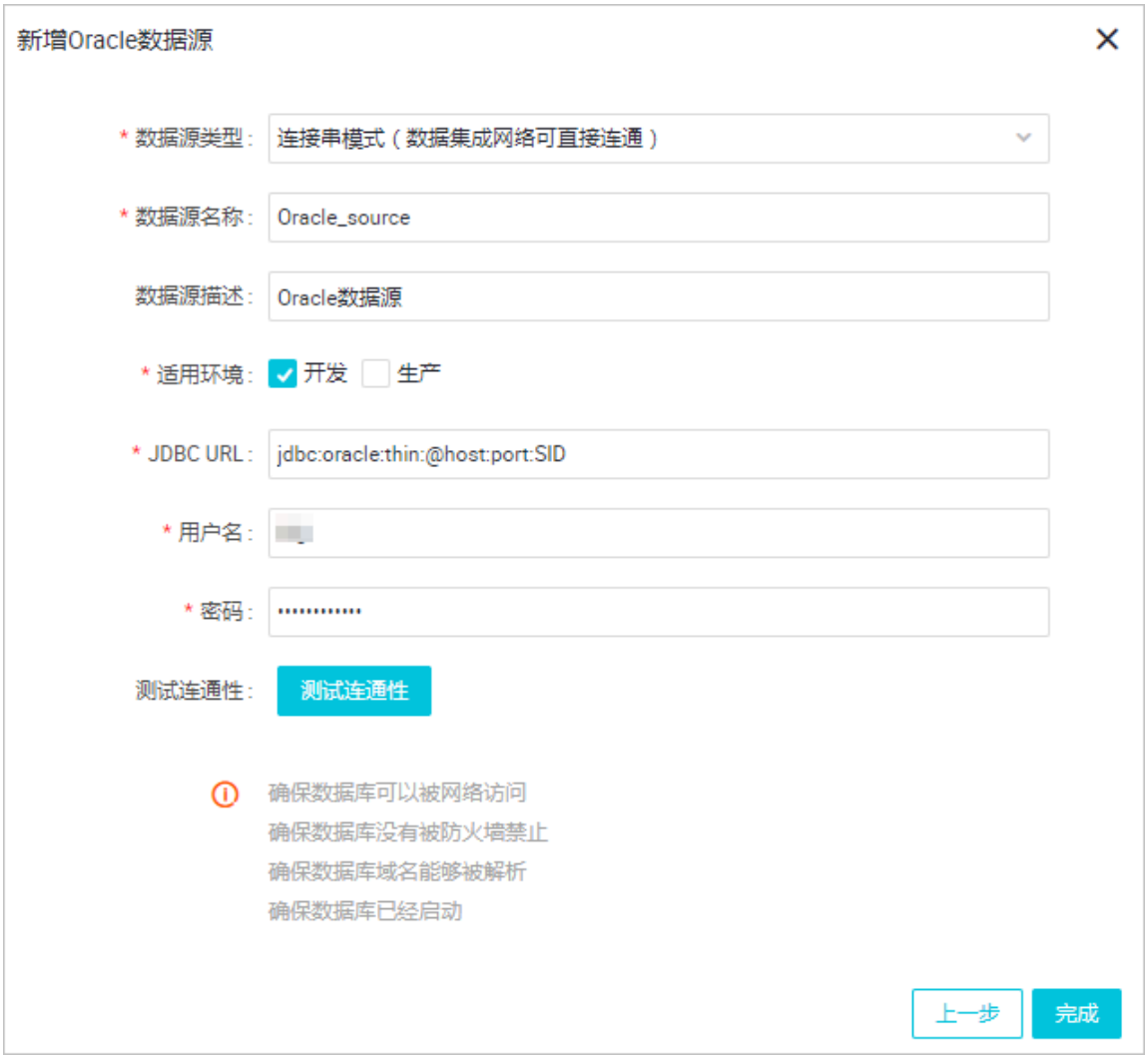


3. 在新增数据源弹出框中，选择数据源类型为Oracle。

4. 填写Oracle数据源的各配置项。

   Oracle数据源类型分为连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通），您可以根据自身情况进行选择。

   - 以新增Oracle > 连接串模式（数据集成网络可直接连通）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为Oracle > 连接串模式（数据集成网络可直接连通）。
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此配置。

配置	说明
JDBC URL	**JDBC连接信息，格式为**`jdbc:oracle:thin:@host:port:SID`**或**`jdbc:oracle:thin:@//host:port/service_name`。

配置	说明
用户名/密码	数据库对应的用户名和密码。

- 以新增Oracle > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为Oracle > 连接串模式（数据集成网络不可直接连通）。 **说明:** 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进行查看。
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。

配置	说明
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此配置。
资源组	选择相应的资源组，您也可以新增自定义资源组。
JDBC URL	JDBC连接信息，格式为`jdbc:oracle:thin:@host:port:SID`或`jdbc:oracle:thin:@//host:port/service_name`。
用户名/密码	数据库对应的用户名和密码。

5. 因该数据源不支持测试连通性，直接单击完成即可。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络目前不支持数据源连通性测试，直接单击完成。

后续步骤

现在，您已经学习了如何配置Oracle数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置Oracle插件，详情请参见配置*Oracle Reader*和配置*Oracle Writer*。

1.6.15 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。

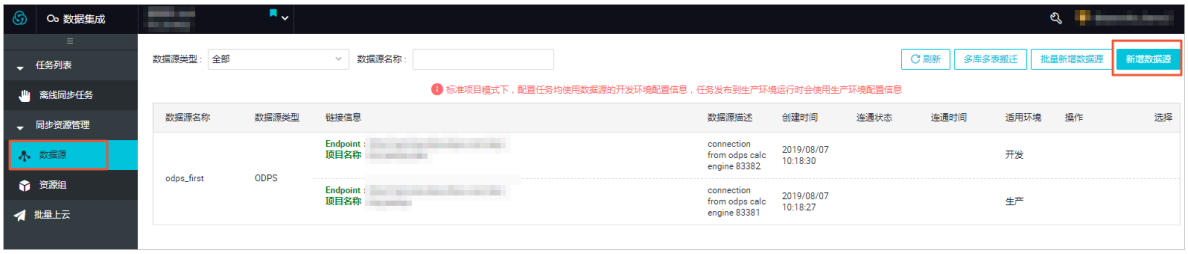
说明:

- 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。
- 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. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。

2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为OSS。

4. 填写OSS数据源的各配置项。

配置	说明
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此配置。

配置	说明
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。 您可在数据同步任务中查找此处填写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/ AceessKey Secret	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

 说明:
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。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络目前不支持数据源连通性测试，直接单击完成。

后续步骤

现在，您已经学习了如何配置OSS数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置OSS插件，详情请参见配置*OSS Reader*和配置*OSS Writer*。

1.6.16 配置Table Store（OTS）数据源

表格存储（Table Store）是构建在阿里云飞天分布式系统之上的NoSQL数据存储服务，提供海量结构化数据的存储和实时访问。

> **说明:**
>
> - **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。**
> - **如果您想对表格存储有更深入的了解，请参见*#unique_98*。**

操作步骤

1. **以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。**
2. **选择同步资源管理 > 数据源，单击新增数据源。**

3. **在新增数据源弹出框中，选择数据源类型为Table Store（OTS）。**

4. 填写Table Store（OTS）数据源的各配置项。

配置	说明
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此配置。
Endpoint	Table Store服务对应的Endpoint。
Table Store实例ID	Table Store服务对应的实例ID。
AccessID/AceessKey	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。

- 专有网络目前不支持数据源连通性测试，直接单击完成。

后续步骤

现在，您已经学习了如何配置OTS数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置Table Store（OTS） Reader插件，详情请参见配置*Table* Store（OTS） *Reader*。

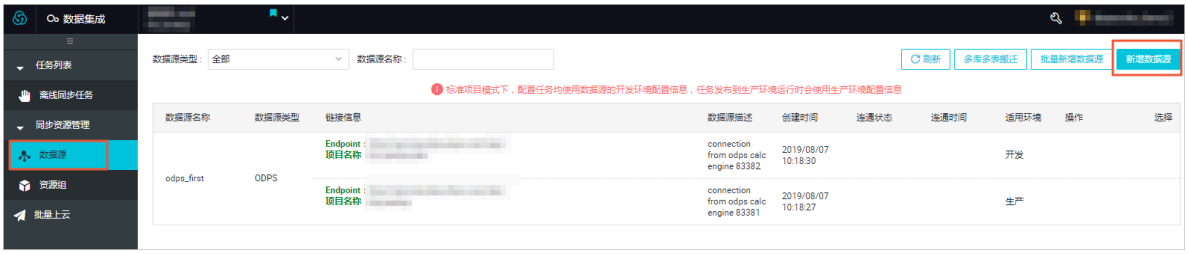
1.6.17 配置PostgreSQL数据源

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。

> **说明:**
> 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。
>
> 如果是在VPC环境下的PostgreSQL，需要注意以下问题：
>
> - 自建的PostgreSQL数据源：
> - 不支持测试连通性，但仍支持配置同步任务，创建数据源时单击完成即可。
> - 必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。
> - 通过RDS创建的PostgreSQL数据源。
>
> 您无需选择网络环境，系统自动根据您填写的RDS实例信息进行判断。

操作步骤

1. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为PostgreSQL。

4. 填写PostgreSQL数据源的各配置项。

   PostgreSQL数据源类型分为阿里云数据库（RDS）、连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通），您可以根据自身情况进行选择。

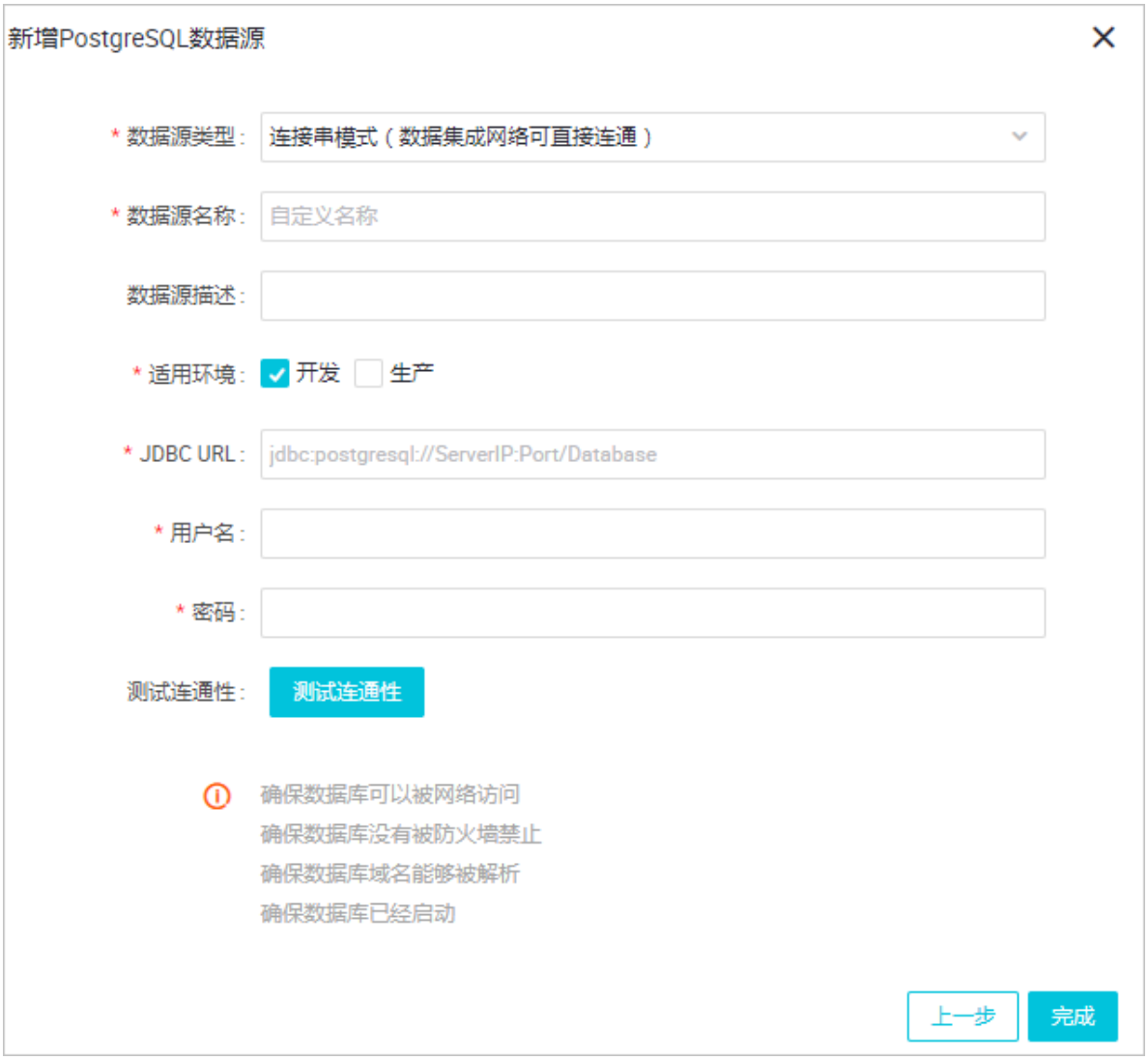
   以新增PostgreSQL > 阿里云数据库（RDS）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为PostgreSQL > 阿里云数据库（RDS）。
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。

配置	说明
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此配置。
地区	选择相应的Region。
RDS实例ID	您可以进入RDS管控台，查看RDS的实例ID。
RDS实例主账号ID	输入购买RDS实例的主账号的ID。
数据库名	填写对应的数据库名称。

配置	说明
用户名/密码	数据库对应的用户名和密码。

以新增PostgreSQL > 连接串模式（数据集成网络可直接连通）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为PostgreSQL > 连接串模式（数据集成网络可直接连通）。
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此配置。

配置	说明
JDBC URL	JDBC连接信息，格式为`jdbc:postgresql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

以新增PostgreSQL > 连接串模式（数据集成网络不可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为PostgreSQL > 连接串模式（数据集成网络不可直接连通）。 选择此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。

配置	说明
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此配置。
资源组	可以用于执行同步任务，通常添加资源组时可以绑定多台机器。详情请参见新增任务资源。
JDBC URL	JDBC连接信息，格式为`jdbc:postgresql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络/专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置PostgreSQL数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置PostgreSQL插件，详情请参见配置*PostgreSQL Reader*和配置*PostgreSQL Writer*。

1.6.18 配置Redis数据源

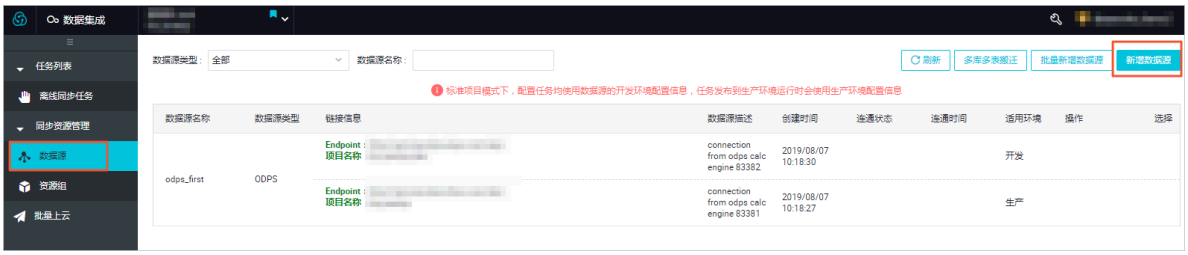
Redis数据源为您提供读取和写入Redis双向通道的功能，您可以通过脚本模式配置同步任务。

Redis是文档型的NoSQL数据库，提供持久化的内存数据库服务，基于高可靠双机热备架构及可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。

说明:
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. 以项目管理员身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 单击左侧导航栏中的同步资源管理 > 数据源，进入数据源页面。
3. 单击右上角的新增数据源。



4. 在新增数据源对话框中，选择数据源类型为Redis。
5. 填写Redis数据源的各配置项。

   Redis数据源类型包括实例模式（阿里云数据源）和连接串模式（数据集成网络可直接连通），您可以根据自身需求进行选择。

   - 以新增Redis > 实例模式（阿里云数据源）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为Redis > 实例模式（阿里云数据源）。

配置	说明
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此配置。
地区	填写购买Redis时所选择的区域。
Redis实例ID	您可以进入Redis管控台，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 > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为Redis > 连接串模式（数据集成网络不可直接连通）。 选择此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。

配置	说明
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此配置。
服务器地址	格式为host:port。
添加访问地址	添加访问地址，格式为host:port。
Redis访问密码	Redis的服务访问密码。

6. 单击测试连通性。
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，您可以继续学习下一个教程。在该教程中您将学习如何配置Redis Writer插件，详情请参见配置*Redis Writer*。

1.6.19 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的双向能力，本文将为您介绍如何配置HybridDB for MySQL数据源。

说明:
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。

您可以通过向向导模式配置和脚本模式配置配置同步任务。

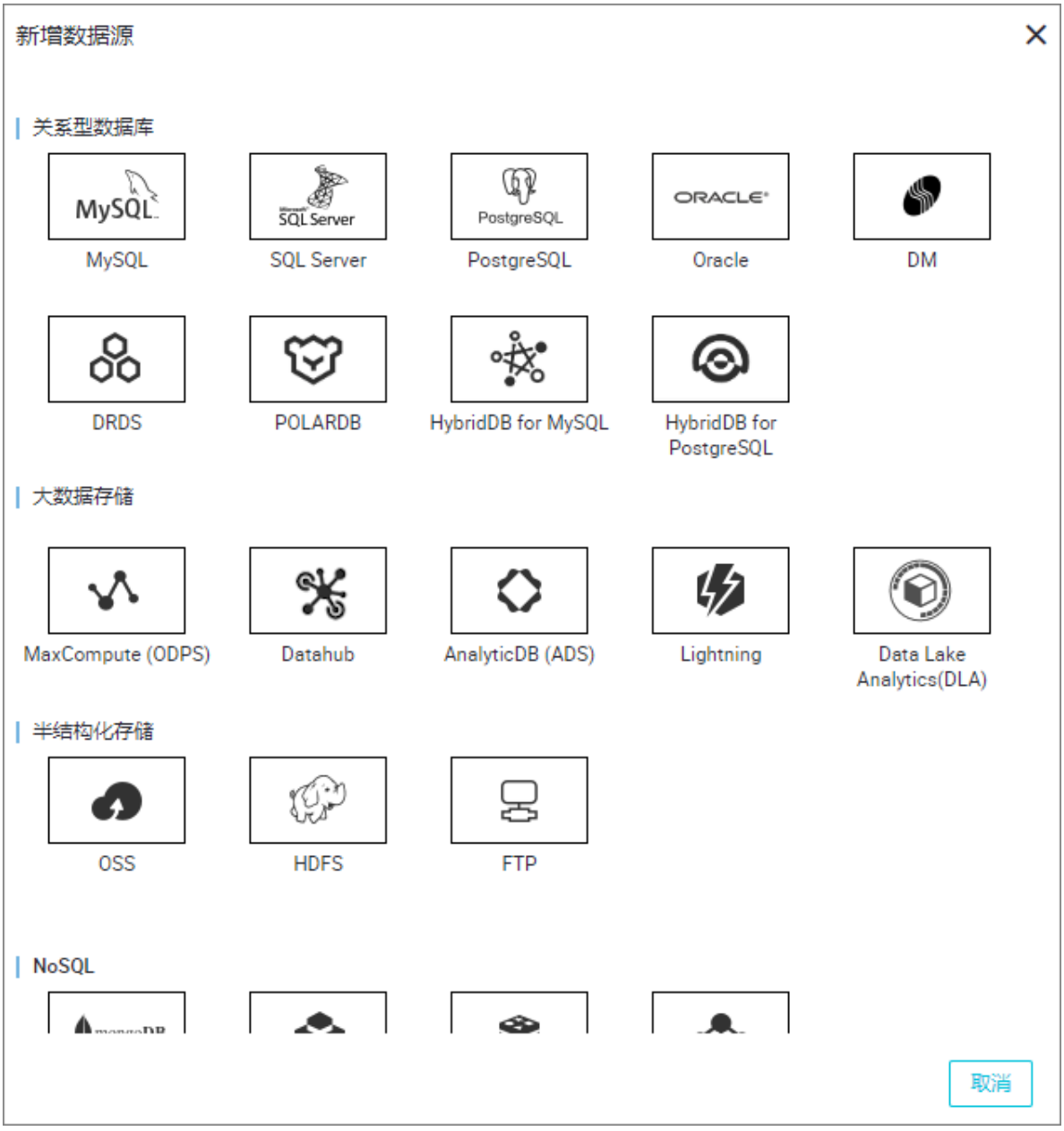
说明:
如果是在VPC环境下的HybridDB for MySQL，需要注意以下问题。

- 自建的MySQL数据源
  - 不支持测试连通性，但仍支持配置同步任务，创建数据源时单击确认即可。
  - 必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。

- 对于通过实例ID创建的HybridDB for MySQL数据源，您无需选择网络环境，系统自动根据您填写的HybridDB for MySQL实例信息进行判断。

操作步骤

1. 以项目管理员身份进入*DataWorks*管理控制台，单击对应项目操作栏中的进入数据集成。
2. 单击数据源 > 新增数据源，弹出支持的数据源。



3. 在新增数据源弹出框中，选择数据源类型为HybridDB for MySQL。

4. 填写HybridDB for MySQL数据源的各配置项。

配置	说明
数据源类型	当前选择的数据源类型为阿里云数据源（HybridDB）。
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	分为开发环境和生产环境。
实例ID	您可进入HybridDB for MySQL管控台，查看相关的实例ID。

配置	说明
主账号ID	您可在HybridDB for MySQL管控台安全设置中查看相应的信息。 
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击确定。

> **说明:**
> 您需要先添加白名单才能连接成功，详情请参看添加白名单文档。

测试连通性说明

- 经典网络下，能够提供测试连通性能力，可以判断输入的用户名/密码、实例ID/JDBC URL是否正确。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的实例ID、主账号ID、用户名/密码是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络/专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的JDBC URL、用户名/密码是否正确。

后续步骤

现在，您已经学习了如何配置HybridDB for MySQL数据源，您可以继续学习下一个教程。在该教程中您将学习如何通过配置HybridDB for MySQL插件，详情请参见配置*HybridDB for MySQL Reader*和配置*HybridDB for MySQL Writer*。

1.6.20 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。

说明:

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。

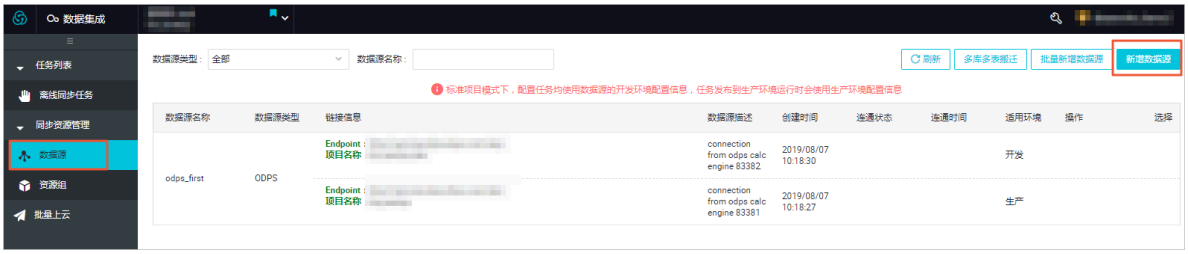
如果是在VPC环境下的AnalyticDB for PostgreSQL，需要注意以下问题：

- **自建的PostgreSQL数据源**
  - **不支持测试连通性，但仍支持配置同步任务，创建数据源时单击完成即可。**
  - **必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。**
- **通过实例ID创建的AnalyticDB for PostgreSQL数据源**

  您无需选择网络环境，系统自动根据您填写的RDS实例信息进行判断。

操作步骤

1. **以项目管理员身份登录*DataWorks*控制台，单击对应工作空间后的进入数据集成。**
2. **选择同步资源管理 > 数据源，单击新增数据源。**

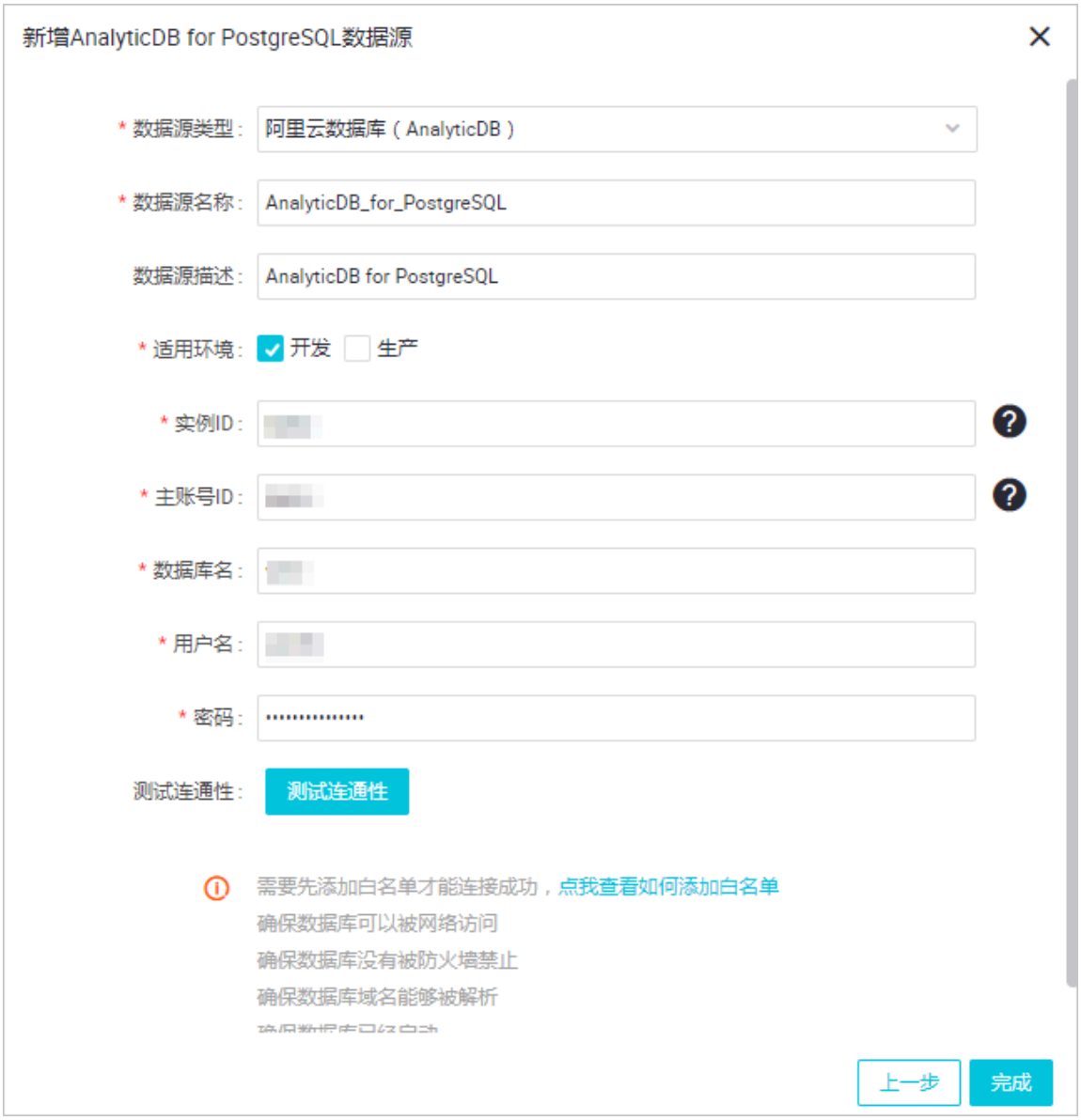


3. **在新增数据源弹出框中，选择数据源类型为AnalyticDB for PostgreSQL。**

4. 填写AnalyticDB for PostgreSQL数据源的各配置项。

   AnalyticDB for PostgreSQL数据源类型包括阿里云数据库（AnalyticDB）和连接串模式（数据集成网络可直接连通）。

   - 以新增AnalyticDB for PostgreSQL > 阿里云数据库（AnalyticDB）类型的数据源为例。



配置	说明
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。

配置	说明
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此配置。
RDS实例ID	您可以进入AnalyticDB for PostgreSQL的控制台，查看相应的实例ID。

配置	说明
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。 

- 以新增AnalyticDB for PostgreSQL > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为AnalyticDB for PostgreSQL > 连接串模式（数据集成网络可直接连通）。
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此配置。
JDBC URL	JDBC连接信息，格式为`jdbc:postgresql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

说明:
您需要先添加白名单才能连接成功，详情请参见添加白名单。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。
- 专有网络下，如果您将VPC内部地址作为JDBC URL添加数据源，测试连通性会报告失败。
- 经典网络/专有网络下，如果您将数据源的公网地址作为JDBC URL添加数据源，可以判断输入的信息是否正确。

后续步骤

现在，您已经学习了如何配置AnalyticDB for PostgreSQL数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置AnalyticDB for PostgreSQL插件，详情请参见配置*AnalyticDB for PostgreSQL Reader*和配置*AnalyticDB for PostgreSQL Writer*。

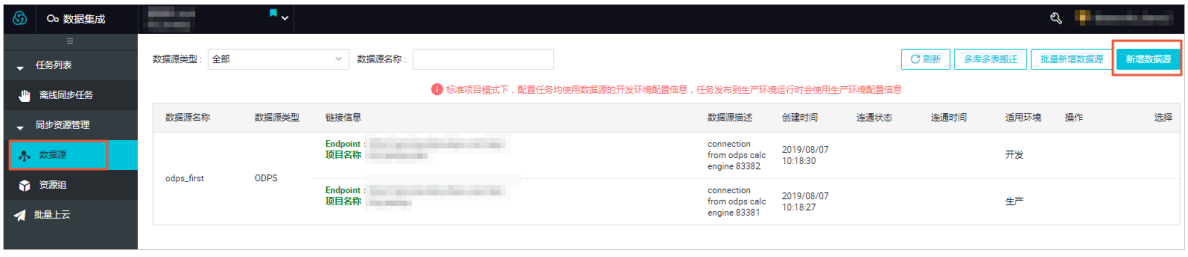
1.6.21 配置POLARDB数据源

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。

> **说明:**
>
> **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。**
>
> **当前POLARDB数据源不支持自定义资源组，请使用默认资源组。如果您需要使用自定义资源组，请选择添加连接串模式（数据集成网络不可直接连通）类型的*MySQL*数据源。如果您的数据源是在VPC环境下的POLARDB，需要注意以下问题：**
>
> - **自建的POLARDB数据源**
>   - **不支持测试连通性，但仍支持配置同步任务，创建数据源时单击完成即可。**
>   - **必须使用自定义调度资源组运行对应的同步任务，请确保自定义资源组可以连通您的自建数据库，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。**
> - **通过实例ID创建的POLARDB数据源**
>
>   **您无需选择网络环境，系统自动根据您填写的POLARDB实例信息进行判断。**

操作步骤

1. **以项目管理员身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。**
2. **选择同步资源管理 > 数据源，单击新增数据源。**



3. **在新增数据源弹出框中，选择数据源类型为POLARDB。**

4. 填写POLARDB数据源的各配置项。

配置	说明
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此配置。

配置	说明
集群ID	您可以进入POLARDB控制台，查看集群ID。
POLARDB实例主账号ID	输入购买POLARDB实例的主账号ID。
数据库名	POLARDB中创建的数据库名。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

> **说明:**
> 您需要先添加白名单才能连接成功，详情请参见添加白名单。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。
- 专有网络以添加实例ID形式能够添加成功，提供相关反向代理功能。

后续步骤

现在，您已经学习了如何配置POLARDB数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置POLARDB插件，详情请参见配置*POLARDB Reader*和配置*POLARDB Writer*。

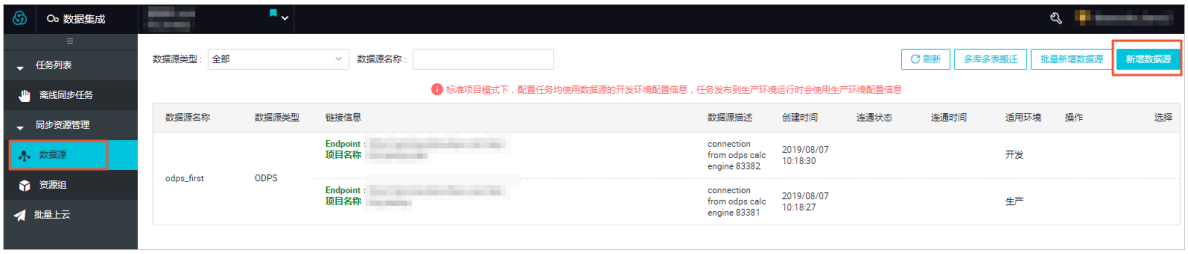
1.6.22 配置AnalyticDB for MySQL数据源

本文将为您介绍如何配置AnalyticDB for MySQL数据源。

> **说明:**
> 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

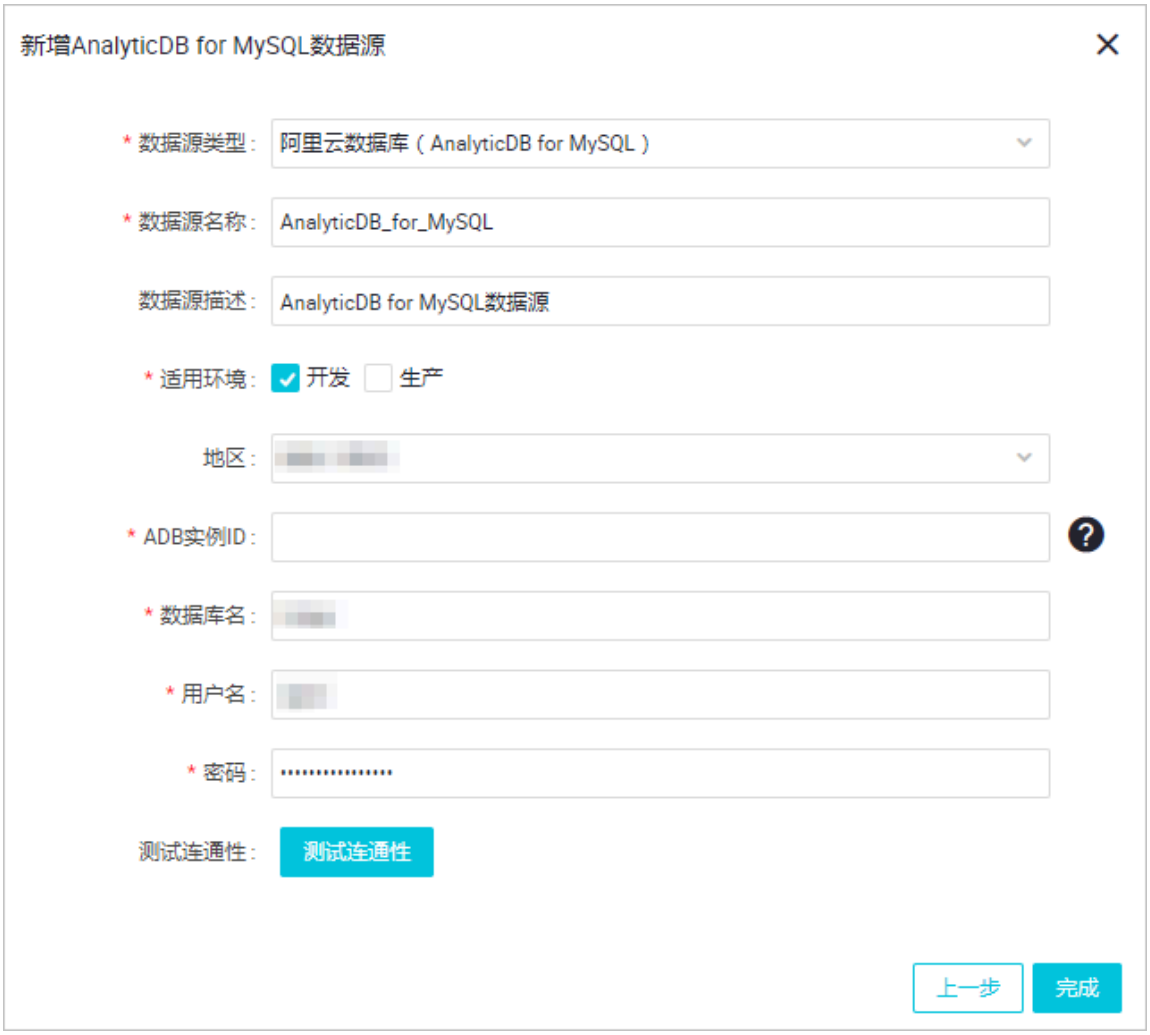
1. 以项目管理员身份进入*DataWorks*控制台，单击对应工作空间操作栏中的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为AnalyticDB for MySQL。

4. 填写AnalyticDB for MySQL数据源的各配置项。

   - 以新增AnalyticDB for MySQL > 阿里云数据库（AnalyticDB for MySQL）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为AnalyticDB for MySQL > 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此配置。
地区	选择数据源所在地区。
ADS实例ID	您可以进入RDS控制台，查看RDS实例ID。

配置	说明
数据库名	您可以新建数据库，设置相应的数据名、用户名和密码。
用户名/密码	数据库对应的用户名和密码。

- 以新增AnalyticDB for MySQL > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为AnalyticDB for MySQL > 连接串模式（数据集成网络可直接连通）。
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此配置。
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。

配置	说明
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络下，能够提供测试连通性能力，可以判断输入的信息是否正确。
- 专有网络下，如果您使用实例模式配置数据源，可以判断输入的信息是否正确。

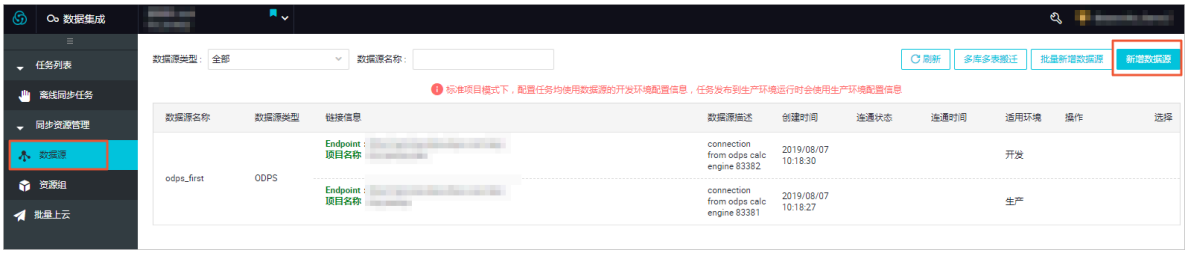
1.6.23 配置Data Lake Analytics（DLA）数据源

本文将为您介绍如何配置Data Lake Analytics（DLA）数据源。

> **说明:**
> 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。

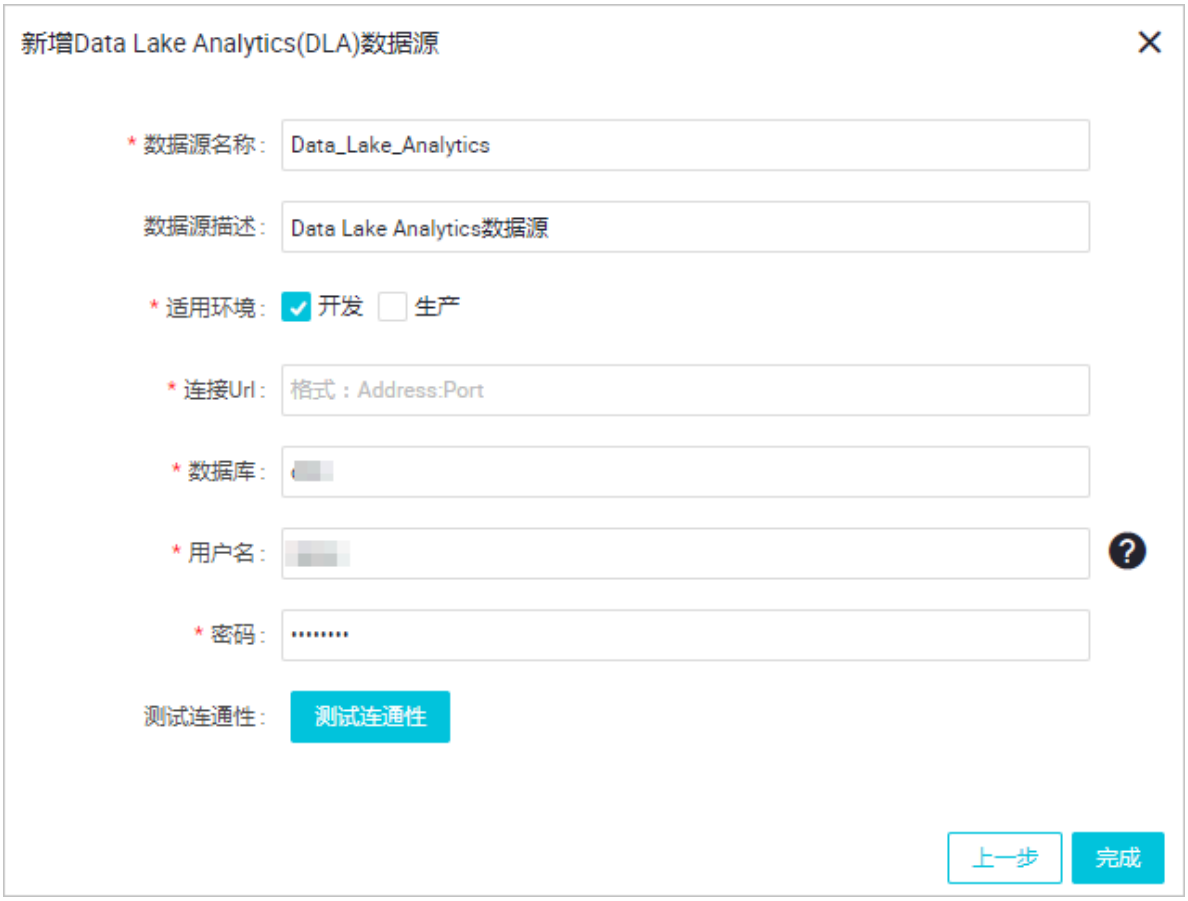
操作步骤

1. 以项目管理员身份进入*DataWorks*控制台，单击对应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为Data Lake Analytics（DLA）。

4. 填写Data Lake Analytics（DLA）数据源的各配置项。



配置	说明
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此配置。
连接Url	格式为`Address:Port`。
数据库	填写对应的数据库名称。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

测试连通性说明

- 经典网络ECS上自建的数据源，建议使用数据集成自定义资源组，默认资源组不保证网络可通。

- 专有网络VPC目前不支持数据源连通性测试，直接单击完成。

  如果是专有网络VPC，需要使用独享资源，详情请参见*#unique_85*和*#unique_101*。

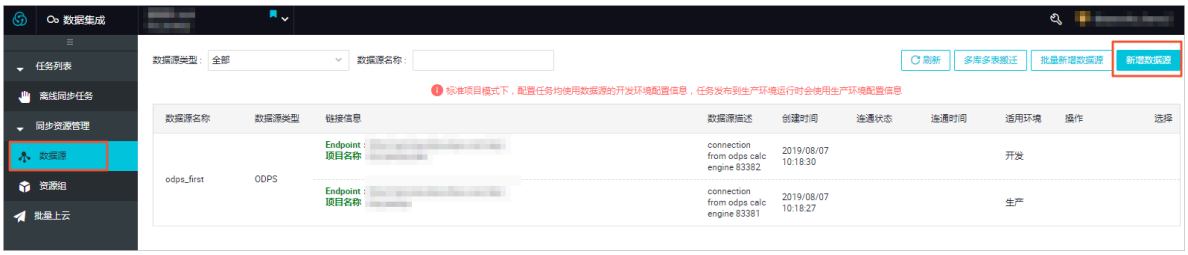
1.6.24 配置AnalyticDB for MySQL 3.0数据源

AnalyticDB for MySQL 3.0数据源为您提供读取和写入AnalyticDB for MySQL 3.0双向通道的功能，可以通过向导模式和脚本模式配置同步任务。

> **说明:**
> 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. 以项目管理员身份登录*DataWorks*控制台，单击对应工作空间操作栏中的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。

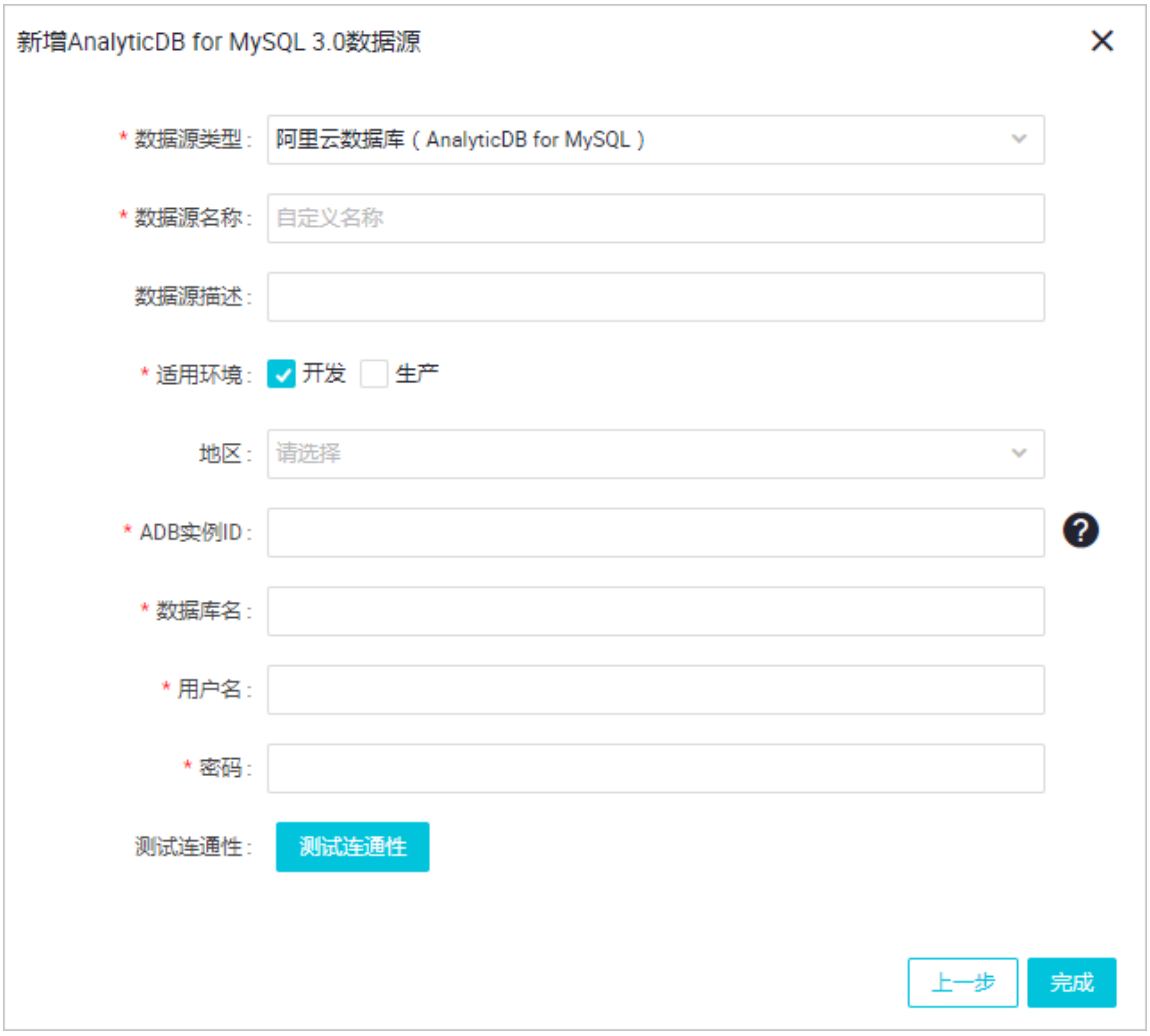


3. 在新建数据源弹出框中，选择数据源类型为AnalyticDB for MySQL 3.0。

4. 填写AnalyticDB for MySQL 3.0数据源的各配置项。

   AnalyticDB for MySQL 3.0数据源类型包括阿里云数据库（AnalyticDB for MySQL）和连接串模式（数据集成网络可直接连通）。

   - 以新增MySQL > 阿里云数据库（AnalyticDB for MySQL）类型的数据源为例。



配置	说明
数据源类型	当前选择的数据源类型为MySQL > 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此配置。

配置	说明
地区	选择相应的区域。
ADS实例ID	您可以进入ADS管控台，查看ADS的实例ID。
数据库名	填写对应的数据库名称。
用户名/密码	数据库对应的用户名和密码。

- 以新增MySQL > 连接串模式（数据集成网络可直接连通）类型的数据源为例。

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 连接串模式（数据集成网络可直接连通）。
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此配置。

配置	说明
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

   提供测试连通性功能，可以判断输入的信息是否正确 。

后续步骤

现在，您已经学习了如何配置AnalyticDB for MySQL 3.0数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置AnalyticDB for MySQL 3.0插件，详情请参见配置*AnalyticDB for MySQL 3.0 Reader*和配置*AnalyticDB for MySQL 3.0 Writer*。

1.6.25 配置GDB数据源

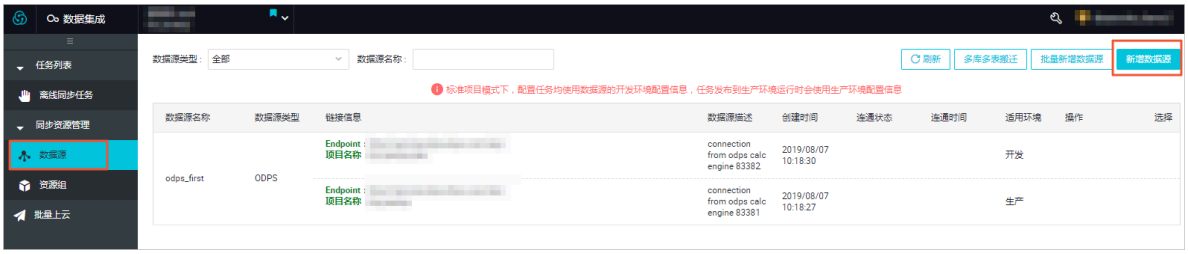
GDB数据源为您提供写入GDB单向通道的功能，您可以通过脚本模式配置同步任务。

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。

> **说明:**
> 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. 以项目管理员身份进入*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为GDB。

4. 填写GDB数据源的各配置项。

配置	说明
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此配置。
图实例域名	图实例域名（host），对应图数据库GDB > 实例列表 > 实例管理 > 基本信息 > 内网地址。
图实例端口	图实例端口（port），对应图数据库GDB > 实例列表 > 实例管理 > 基本信息 > 内网端口。
图实例账号	图实例账号（username），对应图数据库GDB > 实例列表 > 实例管理 > 账号管理 > 用户账号。
图实例密码	图实例密码（password），对应图数据库GDB > 实例列表 > 实例管理 > 账号管理 > 用户账号密码。

5. 单击测试连通性。

> 说明:
> 测试连通性使用的是默认资源组，此时网络不通，测试结果为失败。提交任务时，需要使用数据集成独享资源，方可正常运行。

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GDB数据源，您可以继续学习下一个教程。在该教程中，您将学习如何配置GDB Writer插件，详情请参见配置*GDB Writer*。

1.7 作业配置

1.7.1 配置Reader插件

1.7.1.1 脚本模式配置

本文将为您介绍如何通过数据集成的脚本模式进行任务配置。

任务配置的操作步骤如下所示：

1. 新建数据源。
2. 新建数据同步节点。
3. 导入模板。
4. 配置同步任务的读取端。
5. 配置同步任务的写入端。
6. 配置字段的映射关系。
7. 配置作业速率上限、脏数据检查规则等信息。
8. 配置调度属性。

> 说明:
> 下文将为您介绍操作步骤的具体实现，以下每个步骤都会跳转到对应的指导文档中，请在完成当前步骤后，单击链接回到本文，继续下一步操作。

新建数据源

同步任务支持多种同构、异构数据源间的数据传输。首先，将需要同步的数据源在数据集成中完成注册。注册完成后，在数据集成配置同步任务时，可以直接选择数据源。数据集成支持同步的数据源类型请参见 支持的数据源。

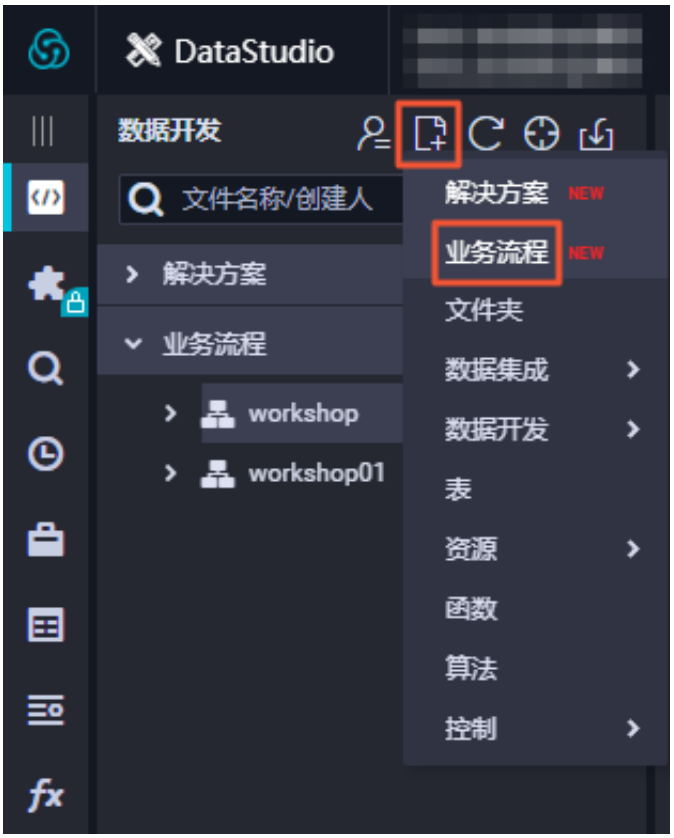
确认需要同步的数据源已经被数据集成支持后，可以开始在数据集成中注册数据源。详细的数据源注册步骤请参见配置数据源信息。

> 说明:
>
> - 有部分数据源数据集成不支持测试连通性，数据源测试连通性的支持详情请参见数据源测试连通性。
> - 很多时候，数据源都是创建在本地，没有公网IP或网络无法直达。在这种情况下，配置数据源的时候测试连通性会直接失败，数据集成支持新增任务资源来解决这种网络不可达的情况。但在新建同步任务的时候只能选择脚本模式（因为网络不可直达，在向导模式中就无法获取表结构等信息）。

新建数据同步节点

> 说明:
>
> 本文主要为您介绍向导模式下的同步任务配置，在数据集成中新建同步任务时请选择脚本模式。

1. 以开发者身份进入*DataWorks*控制台，单击对应工作空间操作栏中的进入数据开发。
2. 进入DataStudio（数据开发）页面，选择新建 > 业务流程。



3. 在新建业务流程对话框中，填写业务流程名称和描述，单击新建。

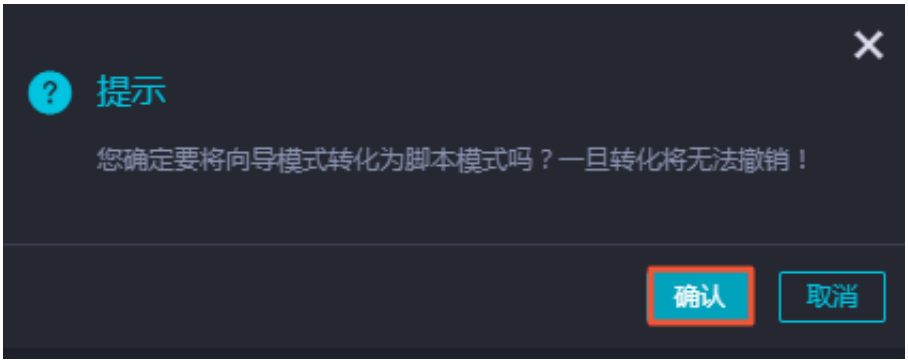
4. 展开业务流程，右键单击数据集成，选择新建数据集成节点 > 数据同步，输入节点名称。

5. 单击提交。

导入模板

1. 成功创建数据同步节点后，单击工具栏中的转换脚本。

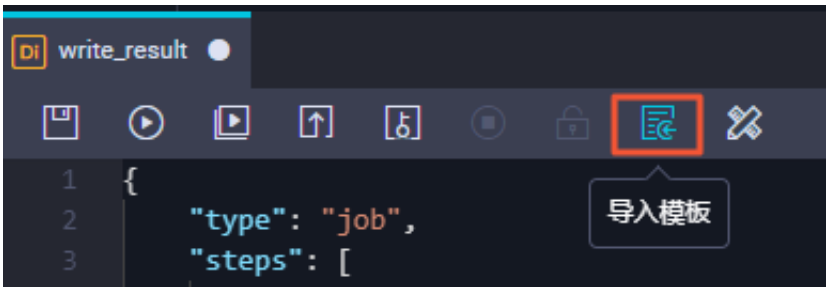
2. 单击提示对话框中的确认，即可进入脚本模式进行开发。



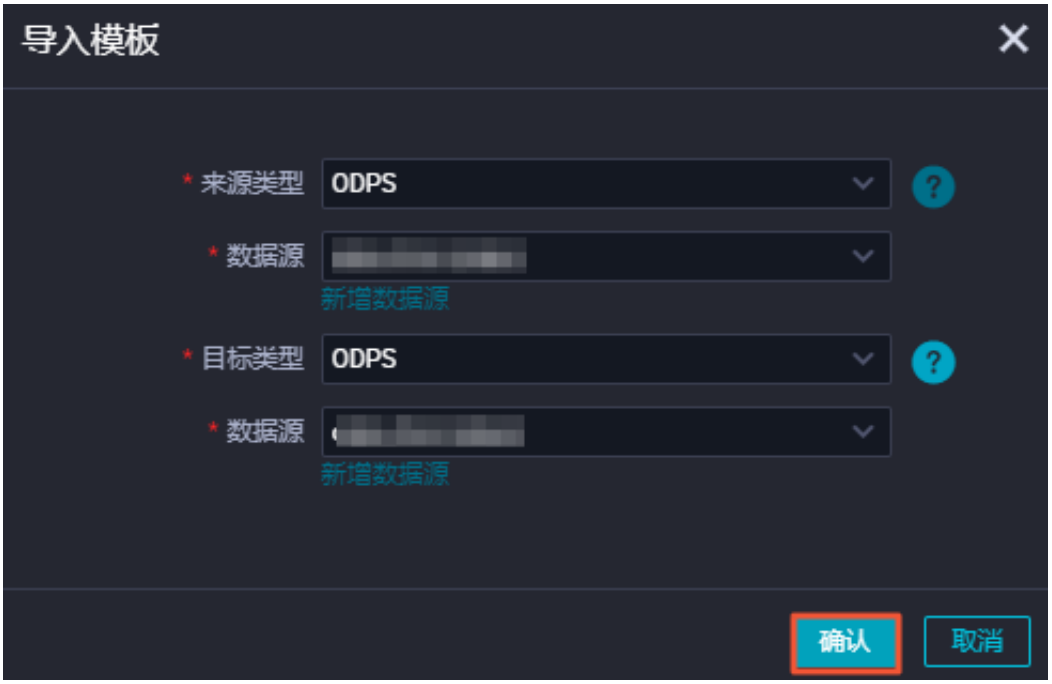
说明:

脚本模式支持更多功能，例如网络不可达情况下的同步任务编辑。

3. 单击工具栏中的导入模板。



4. 在导入模板对话框中，选择来源类型、数据源、目标类型及数据源。



5. 单击确认。

配置同步任务的读取端

新建同步任务完成后，通过导入模板已生成了基本的读取端配置。此时您可以继续手动配置数据同步任务的读取端数据源，以及需要同步的表信息等。

```
{"type": "job",
    "version": "2.0",
    "steps": [   //上述配置为整个同步任务头端代码，可以不进行修改。
        {
            "stepType": "mysql",
            "parameter": {
                "datasource": "MySQL",
                "column": [
                    "id",
                    "value",
                    "table"
                ],
                "socketTimeout": 3600000,
                "connection": [
                    {
                        "datasource": "MySQL",
                        "table": [
                            "`case`"
                        ]
                    }
                ],
                "where": "",
                "splitPk": "",
                "encoding": "UTF-8"
            },
            "name": "Reader",
            "category": "reader"     //说明分类为reader读取端。
        },    //以上配置为读取端配置。
```

配置项说明如下：

- `type`：**指定本次提交的同步任务，仅支持Job参数，所以您只能填写为Job。**
- `version`：**目前所有Job支持的版本号为1.0或2.0。**

> **说明：**
>
> - **选择读取端的数据源时，请参见配置*Reader*中的脚本开发介绍。**
> - **很多任务在配置读取端数据源时，需要进行数据增量同步。此时可以结合DataWorks提供的*#unique_20*来获取相对日期，以完成获取增量数据的需求。**

配置同步任务的写入端

配置完成读取端数据源信息后，可以继续手动配置数据同步任务的写入端数据源，以及需要同步的表信息等。

```
{
  "stepType": "odps",
  "parameter": {
      "partition": "",
```

```
        "truncate": true,
        "compress": false,
        "datasource": "odps_first",
        "column": [
            "*"
         ],
         "emptyAsNull": false,
         "table": ""
       },
       "name": "Writer",
       "category": "writer"    //说明分类为writer写入端。
    }
 ],    //以上配置为读取端配置。
```

说明:

- **选择写入端的数据源时，请参见配置*Writer*。**
- **很多任务在写入时，需要选择写入模式。例如覆盖写入还是追加写入，针对不同的数据源，有不同的写入模式。**

配置字段的映射关系

脚本模式仅支持同行映射，可以在同行建立相应的映射关系，请注意匹配数据类型。

说明:

请注意列与列之间映射的字段类型是否数据兼容。

配置通道控制

当上述步骤都配置完成后，则需进行效率配置。setting域描述的是Job配置参数中除源端、目的端外，有关Job全局信息的配置参数。您可以在setting域中进行效率配置，主要包括同步并发数设置、同步速率设置、同步脏数据设置和同步资源组设置等信息。

```
"setting": {
        "errorLimit": {
            "record": "1024"    //脏数据条目设置。
        },
        "speed": {
            "throttle": false,    //是否进行限速。
            "concurrent": 1,  //同步并发数设置。
        }
    },
```

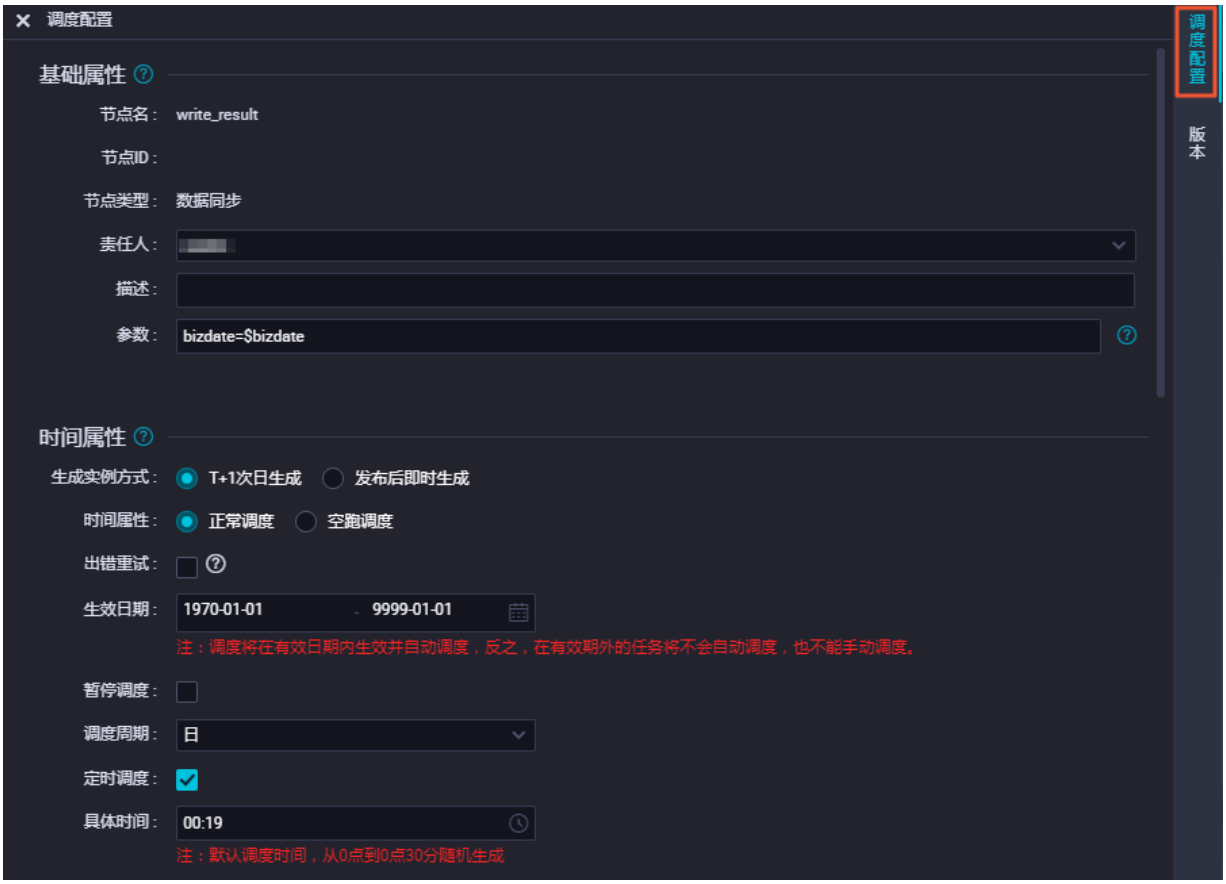
配置	说明
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。

配置	说明
错误记录数	错误记录数，表示脏数据的最大容忍条数。
任务资源组	单击当前页面右上角的配置任务资源组，即可指定资源组配置。 任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见#unique_109和新增任务资源。

配置调度属性

数据同步节点中，经常需要使用调度参数进行数据过滤，下文将为您介绍如何在同步任务中配置调度参数。

进入数据同步节点编辑界面，单击右侧的调度配置。



您可以设置数据同步节点的运行周期、运行时间和调度依赖等属性。由于数据同步节点是ETL工作的开始，所以没有上游节点，此时建议使用工作空间根节点作为上游。

完成数据同步节点的配置后，请保存并提交节点。

1.7.1.2 向导模式配置

本文将为您介绍如何通过数据集成向导模式进行任务配置。

任务配置的操作步骤如下所示：

1. **新建数据源。**
2. **新建数据同步节点。**
3. **选择数据来源。**
4. **选择数据去向。**
5. **配置字段的映射关系。**
6. **配置作业速率上限、脏数据检查规则等信息。**
7. **配置调度属性。**

> **说明：**
> **下文将为您介绍操作步骤的具体实现，以下每个步骤都会跳转到对应的指导文档中。请在完成当前步骤后，单击链接回到本文，继续下一步操作。**

新建数据源

同步任务支持多种同构、异构数据源间的数据传输。首先，将需要同步的数据源在数据集成中完成注册。注册完成后，在数据集成配置同步任务时，可以直接选择数据源。数据集成支持同步的数据源类型请参见 支持的数据源。

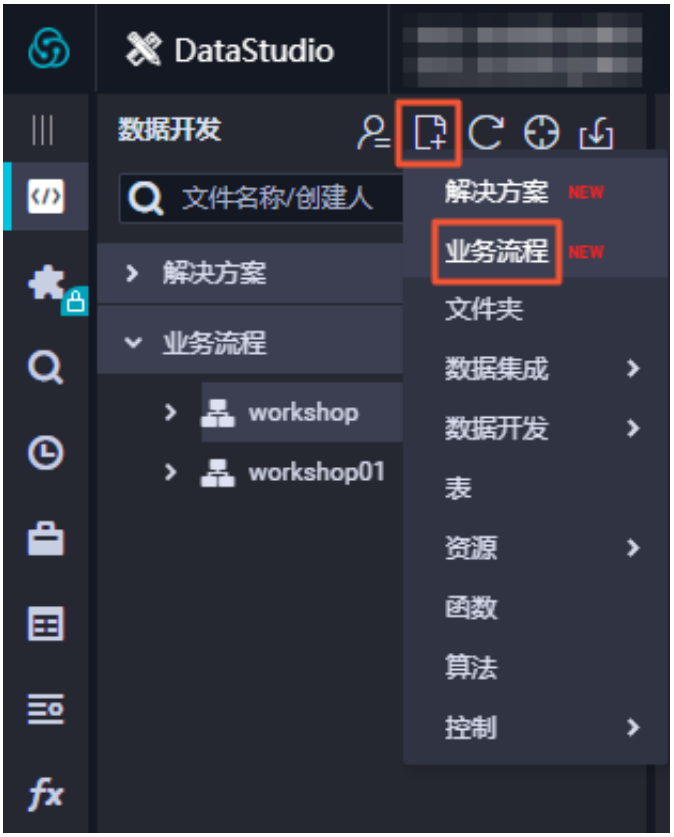
确认需要同步的数据源已经被数据集成支持后，可以开始在数据集成中注册数据源。详细的数据源注册步骤请参见配置数据源信息。

> **说明：**
> - **有部分数据源数据集成不支持测试连通性，数据源测试连通性的支持详情请参见数据源测试连通性。**
> - **很多时候，数据源都是创建在本地，没有公网IP或网络无法直达。在这种情况下，配置数据源的时候测试连通性会直接失败，数据集成支持新增任务资源来解决这种网络不可达的情况。但在新建同步任务的时候只能选择脚本模式（因为网络不可直达，在向导模式中就无法获取表结构等信息）。**

新建数据同步节点

> **说明：**
> **本文主要为您介绍向导模式下的同步任务配置，在数据集成中新建同步任务时请选择向导模式。**

1. 以开发者身份进入*DataWorks*管理控制台，单击对应工作空间操作栏中的进入数据开发。
2. 进入DataStudio（数据开发）页面，选择新建 > 业务流程。



3. 在新建业务流程对话框中，填写业务流程名称和描述，单击新建。

4. 展开业务流程，右键单击数据集成，选择新建数据集成节点 > 数据同步，输入节点名称，单击提交。

选择数据来源

新建数据同步节点后，首先需要配置数据同步节点的读取端数据源，以及需要同步的表等信息。

说明:

- **选择读取端的数据源时，请参见配置*Reader*。**
- **很多任务在配置读取端数据源时，需要进行数据增量同步。此时可以结合DataWorks提供的*#unique_20*来获取相对日期，以完成获取增量数据的需求。**

选择数据去向

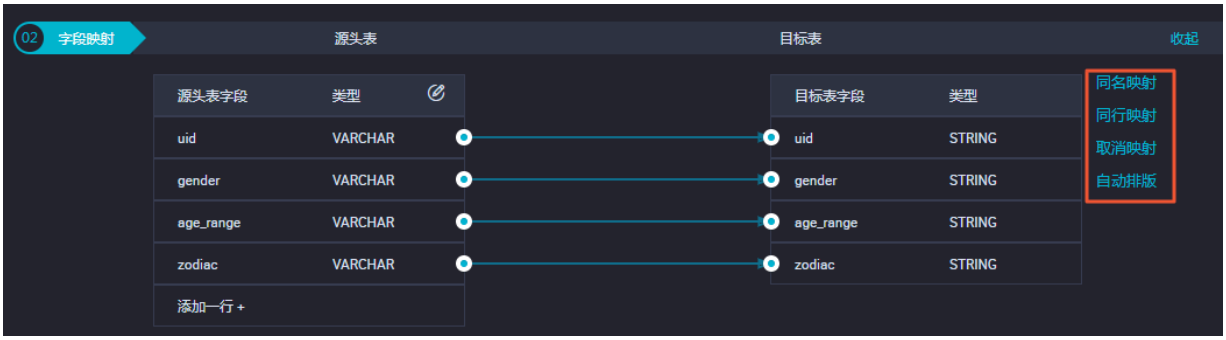
配置完成读取端数据源信息后，可以配置右侧的写入端数据源，以及需要写入的表信息等。

说明:

- **选择写入端的数据源时，请参见配置*Writer*。**
- **很多任务在写入时，需要选择写入模式。比如覆盖写入还是追加写入，针对不同的数据源，有不同的写入模式。**

配置字段的映射关系

选择好数据来源和数据去向后，需要指定读取端和写入端列的映射关系，可以选择同名映射、同行映射、取消映射或自动排版。



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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

说明:

请注意列与列之间映射的字段类型是否数据兼容。

配置通道控制

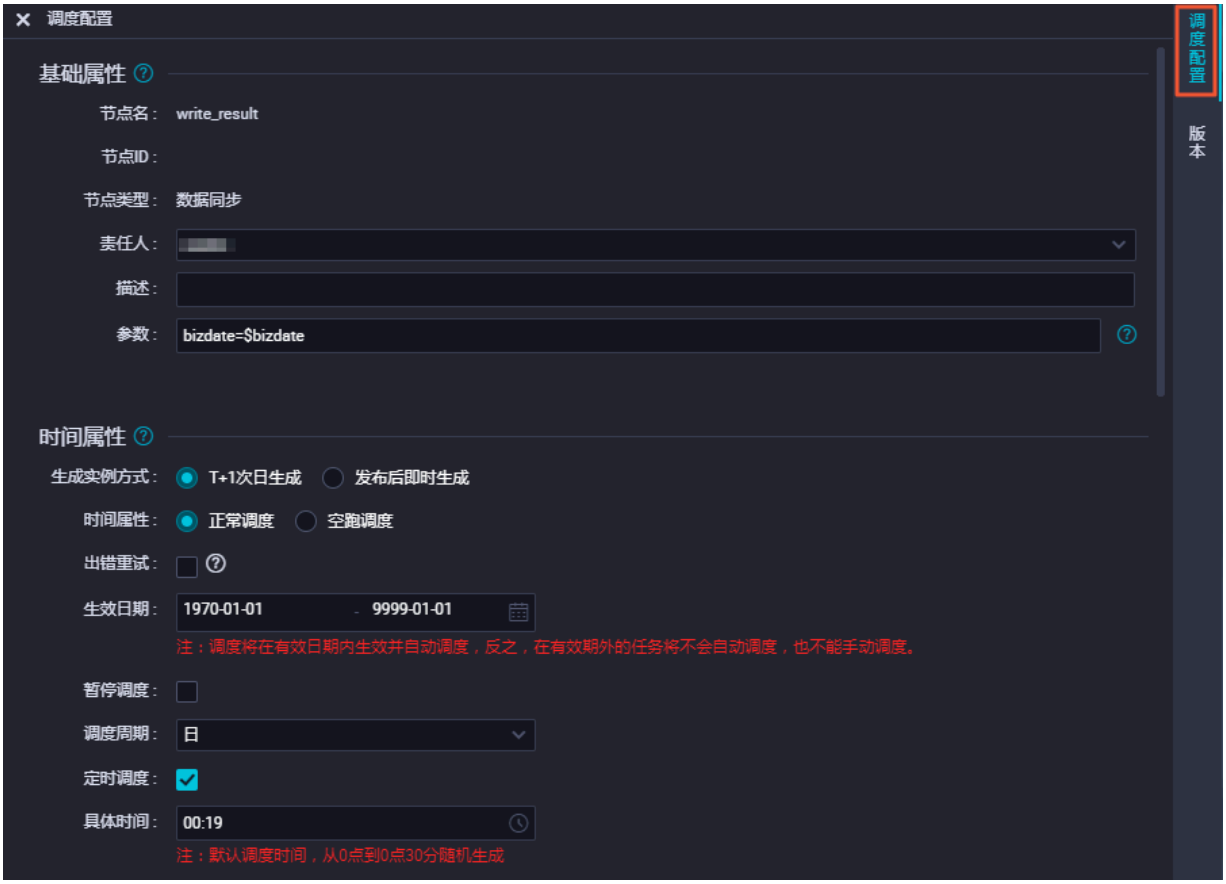
配置完成上述操作后，需要进行通道控制。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和**新增任务资源**。

配置调度属性

数据同步节点中，经常需要使用调度参数进行数据过滤，下文将为您介绍如何在同步任务中配置调度参数。

进入数据同步节点编辑界面，单击右侧的调度配置。



您可以通过${变量名}的方式声明调度参数变量。当变量声明完成后，在调度的参数属性中写上变量的初始化值，此处变量初始化的值以$[]为标识，其中的内容可以填时间表达式或者一个常量。

例如在代码中写了${today}，在调度参数中赋值today=$[yyyymmdd]，则可获取到当天的日期。如果需要对日期进行加减操作，请参见#unique_20。

您可以设置数据同步节点的运行周期、运行时间和调度依赖等属性。由于数据同步节点是ETL工作的开始，所以没有上游节点，此时建议使用工作空间根节点作为上游。

在同步任务中使用自定义调度参数

在同步任务中只需要在代码中声明如下参数即可。

- bizdate：获取到业务日期，运行日期-1。
- cyctime：获取到当前运行时间，格式为yyyymmddhhmiss。
- Dataworks提供了两个系统默认调度参数bizdate和cyctime 。

完成数据同步节点的配置后，请保存并提交节点。

1.7.1.3 配置DRDS Reader

DRDS Reader插件实现了从DRDS（分布式RDS）读取数据。在底层实现上，DRDS Reader通过JDBC连接远程DRDS数据库，并执行相应的SQL语句，从DRDS库中选取数据。

DRDS的插件目前仅适配了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。

简而言之，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，DRDS暂时无法支持。

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

类型分类	DRDS数据类型
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
datasourc	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
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
where	筛选条件，DRDS Reader根据指定的column、table、where条件拼接SQL，并根据这个SQL进行数据抽取。例如在测试时，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，可以将where条件指定为STRTODATE('${bdp.system.bizdate}', '%Y%m%d') <= taday AND taday < DATEADD(STRTODATE('${bdp.system.bizdate}', '%Y%m%d'), interval 1 day)。 · 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 。	否	无

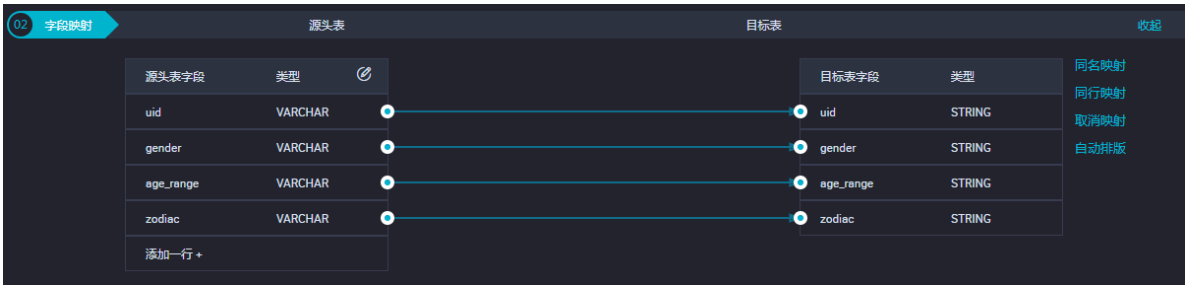
向导开发介绍

1. **配置同步任务的数据来源和数据去向。**

配置	说明
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 。



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。
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	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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个从DRDS数据库同步抽取数据作业。

```
{
    "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"
        },
        {//此处以stream为例，如果您需要其他插件，可以查找相应的文档。
            "stepType":"stream",//插件名
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
```

```
    ],
    "setting":{
        "errorLimit":{
            "record":"0"//错误记录数。
        },
        "speed":{
            "throttle":false,//是否限流。
            "concurrent":1,//并发数。
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}:"Writer"
            }
        ]
    }
}
```

补充说明

- **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，也无法提供解决方案，这类情况的导出结果有可能为乱码。

- 增量数据同步

  DRDS 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，有以下几种方式：

  - 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。

- SQL安全性

  DRDS Reader提供querySql语句交给您自己实现SELECT抽取语句，DRDS Reader本身对querySql不进行任何安全性校验。

1.7.1.4 配置HBase Reader

HBase Reader插件实现了从HBase中读取数据。本文将为您介绍HBase Reader支持的功能、数据类型和参数，以及配置示例。

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。

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

类型分类	数据集成column配置类型	数据库数据类型
字符串类	string	binary_string和string
日期时间类	date	date
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/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 client每次RPC从服务器端读取的行数。	否	256
scanBatchSize	HBase client每次RPC从服务器端读取的列数。	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）。

```
{
    "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"
            },
            {//下面是关于Reader的模板，您可以查看相应的读插件文档。
                "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"
                }
            ]
        }
```

```
}
```

1.7.1.5 配置HDFS Reader

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 。

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。

示例如下：

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，这种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。

说明：

- **由于打通默认资源组到HDFS的网络链路比较复杂，建议您使用自定义资源组完成数据同步任务。您需要确保您的自定义资源组具备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。基于该情况，建议您使用自定义资源组完成针对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

目前HDFS Reader支持的功能如下所示：

- **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sequence file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。

说明:

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

说明如下：

- 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。

说明:

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	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*。 · 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速度指定。 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速度两者中的小者。 · 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目前仅支持（*）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。 说明: · 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（partition）。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，并且/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。默认资源组不支持Hadoop高级参数HA的配置，请新增自定义资源，详情请参见新增任务资源。	是	无

参数	描述	必选	默认值
fileType	文件的类型，目前只支持您配置为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sequence file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 file文件格式。  说明: 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。	是	无

参数	描述	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:** 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
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（,）。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。 · 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

参数	描述	必选	默认值
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包括lzo和lzo_deflate两种压缩格式，您在配置时需要注意。 · 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）。 · rc表示rcfile文件格式。 · orc文件类型下无需填写。	否	无

参数	描述	必选	默认值
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parquetSchem后，整体配置符合JSON语法。parquetScema的配置格式说明如下： `message MessageType名 {` `是否必填，数据类型，列名;` `.....;` `}` · 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。 · 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
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;`	否	无

参数	描述	必选	默认值
kerberosKeytabFilePath	Kerberos认证keytab文件的绝对路径。如果haveKerberos为true，则必选。	否	无
kerberosPrincipal	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

配置一个从HDFS抽取数据到本地的作业，详情请参见上述参数说明。

```
{
    "type": "job",
    "version": "2.0",
    "steps": [
        {
            "stepType": "hdfs",//插件名。
            "parameter": {
                "path": "",//需要读取的文件路径。
                "datasource": "",//数据源。
                "column": [
                    {
                        "index": 0,//序列号。
                        "type": "string"//字段类型。
                    },
                    {
                        "index": 1,
                        "type": "long"
                    },
                    {
                        "index": 2,
                        "type": "double"
                    },
                    {
                        "index": 3,
                        "type": "boolean"
                    },
                    {
                        "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"
        },
        { //下面是关于Writer的模板，您可以查找相应的写插件文档。
            "stepType": "stream",
            "parameter": {},
            "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"
            }
        ]
    }
}
```

parquetSchema的HDFS Reader配置样例如下。

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_m
inute/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; }"
    }
}
```

1.7.1.6 配置MaxCompute Reader

本文将为您介绍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。

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/表字段等信息，在底层实现上可通过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。当读取分区表时，需要指定出具体的分区配置，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。当读取非分区表时，您不能提供分区配置。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）。

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column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和timestamp
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	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	读取数据表的表名称（大小写不敏感）。	是	无
partition	读取数据所在的分区信息，支持linux shell通配符，包括表示0个或多个字符，?代表一个字符是否存在。例如现在有分区表test，其存在pt=1/ds=hangzhou、pt=1/ds=shanghai、pt=2/ds=hangzhou和pt=2/ds=beijing四个分区。 · 如果您想读取pt=1/ds=shanghai分区的数据，则应该配置为"partition":"pt=1/ds=shanghai"。 · 如果您想读取pt=1下的所有分区，则应该配置为"partition":"pt=1/ds=*"。 · 如果您想读取整个test表的所有分区的数据，则应该配置为"partition":"pt=*/ds=*"。	如果表为分区表，则必填。如果表为非分区表，则不能填写。	无

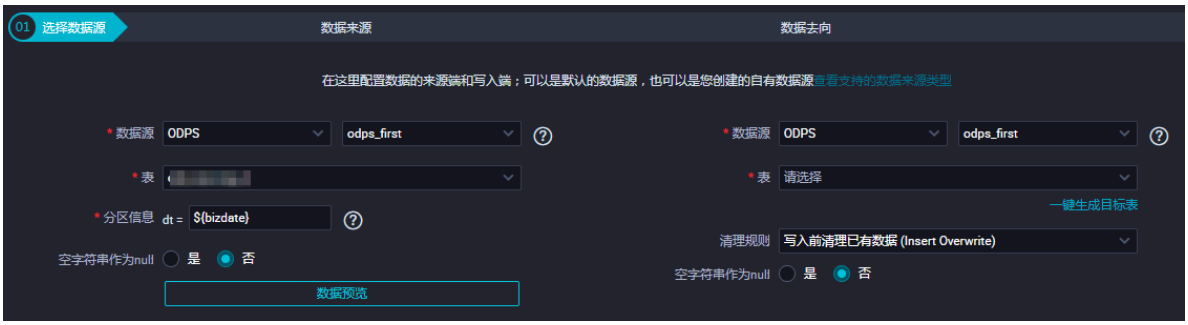
参数	描述	必选	默认值
column	读取MaxCompute源头表的列信息。例如现在有表test，其字段为id、name和age。 · 如果您想依次读取id、name和age，则应该配置为"column":["id","name","age"]或者配置为"column":["*"]。 说明: 不推荐您配置抽取字段为（*），因为它表示依次读取表的每个字段。如果您的表字段顺序调整、类型变更或者个数增减，您的任务会存在源头表列和目的表列不能对齐的风险，则直接导致您的任务运行结果不正确甚至运行失败。 · 如果您想依次读取name和id，则应该配置为"coulumn":["name","id"]。 · 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，那么您应该配置为"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#unique_112。

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。
分区信息	填写相应的分区信息。
压缩	可以选择压缩或不压缩。
空字符串是否作为null	可以选择空字符串是否作为null处理。

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。

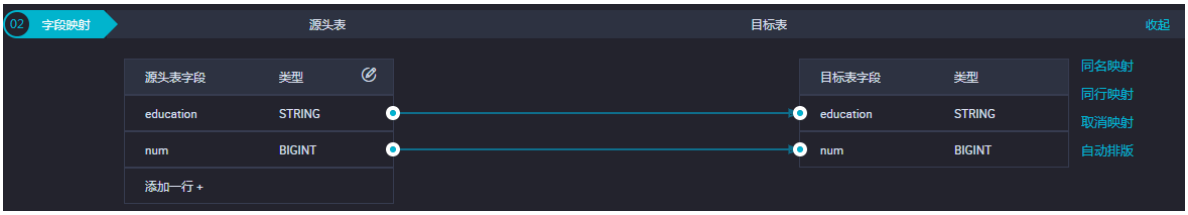
- "partition":"pt=20140501/ds=*"代表ds中的所有的分区。
- "partition":"pt=top?"中的?代表前面的字符是否存在，指pt=top和pt=to两个分区。

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。
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	· 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。

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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个从MaxCompute抽取数据到本地的作业，详情请参见上述参数说明。

```
{
    "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"
        },
        { //下面是关于Writer的模板，您可以查看相应的写插件文档。
            "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"
            }
        ]
```

```
    }
}
```

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":"*****",
```

1.7.1.7 配置MongoDB Reader

本文将为您介绍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级别降低到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，然后逐一判断MongoDB支持的类型，将其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

类型分类	MongoDB数据类型
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。

  例如将MongoDB中的字段导入至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
a	b
a	b

odps_column1	odps_column2
a	b

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
collectionName	MonogoDB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：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。	是	无
query	您可以通过该配置型来限制返回MongoDB数据范围，仅支持时间类型。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${last_day}零点的数据，此处的${last_day}为DataWorks调度参数。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），详情请参见MongoDB查询语法。	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MongoDB抽取数据到本地的作业，详情请参见上述参数说明。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
            "reader": {
            "plugin": "mongodb", //插件名称。
```

```
            "parameter": {
                "datasource": "datasourceName", //数据源名称。
                "collectionName": "tag_data", //集合名称。
                "query":"",
                "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": " "
                        }
                ]
            }
        },
        { //下面是关于Writer的模板，您可以查找相应的写插件文档。
            "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"
            }
        ]
    }
}
```

说明:

暂时不支持取出array中的指定元素。

1.7.1.8 配置DB2 Reader

本文将为您介绍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。

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，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直接将其发送到DB2数据库。

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	—
二进制类	BLOB

参数说明

参数	描述	必选	默认值
datasourc	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	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DB2 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，数据同步系统因此会启动并发任务进行数据同步，这样可以大大提供数据同步的效能。 · 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、字符串和日期等其他类型。如果您指定其他非支持类型，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，并根据这个SQL进行数据抽取。在实际业务场景中，往往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。如果该值为空，代表同步全表所有的信息。	否	无

参数	描述	必选	默认值
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。当您配置了这项后，数据同步系统就会忽略table、column等配置，直接使用这个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，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。

```
{
    "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":"",//splitPk代表的字段进行数据分片。
                "table":"",//表名。
                "username":""//用户名。
            },
            "name":"Reader",
            "category":"reader"
        },
        { //下面是关于Writer的模板，可以查找相应的写插件文档。
            "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"
            }
        ]
    }
}
```

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**

  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，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，DB2 Reader完全不会获取到写入更新数据。

  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使用并发数据抽取，因此不能严格保证数据一致性。

  当DB2 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

  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，有以下几种方式：

  - 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。

1.7.1.9 配置MySQL Reader

本文将为您介绍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。

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库中选取数据。

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

类型分类	MySQL数据类型
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。
- MySQL Reader插件将tinyint(1)视作整型。
- 目前MySQL Reader暂不支持MySQL 8.0及以上版本。

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Job只能同步一张表。	是	无

参数	描述	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：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

参数	描述	必选	默认值
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
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 。当配置此项后，数据同步系统就会忽略tables、columns和splitPk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，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。	否	无

参数	描述	必选	默认值
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。singleOrMulti仅前端使用，后端没有用此进行分库分表判断。	是	multi

向导开发介绍

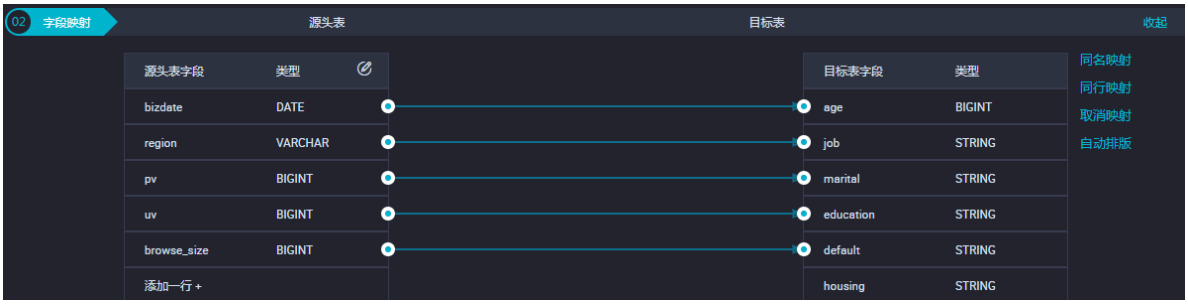
1. 选择数据源。

   配置同步任务的数据来源和数据去向。

配置	说明
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 。



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。
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，如’abc’、’123’等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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",
    "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。
                            "xxx"
                        ]
                    }
                ],
                "where":"",//过滤条件。
                "splitPk":"",//切分键。
                "encoding":"UTF-8"//编码格式。
            },
            "name":"Reader",
            "category":"reader"
        },
        {//下面是关于writer的模板，您可以查找相应的写插件文档。
            "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"
            }
        ]
    }
}
```

分库分表的脚本样例如下，详情请参见上述参数说明。

说明:

分库分表是指在MySQL Reader端可以选择多个MySQL数据表，且表结构要一致。

```
{
    "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": {
        }
    }
}
```

1.7.1.10 配置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举例。

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。

公共云上RDS/DRDS不提供Oracle存储引擎，Oracle Reader目前更多用于专有云数据迁移、数据集成项目。

简单来说，Oracle Reader通过JDBC连接器连接到远程的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CDP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，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
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，这样可以大大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支持数字类型、字符串类型，浮点和日期等其他类型。 · 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。	否	无
where	筛选条件，Oracle Reader根据指定的column、table和where条件拼接SQL，并根据这个SQL进行数据抽取。例如在做测试时，可以将where条件指定为row_number()。在实际业务场景中，往往会选择当天的数据进行同步，可以将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。当您配置这项后，数据同步系统就会忽略table和column等配置，直接使用这个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，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	1024

向导开发介绍

1. 选择数据源。

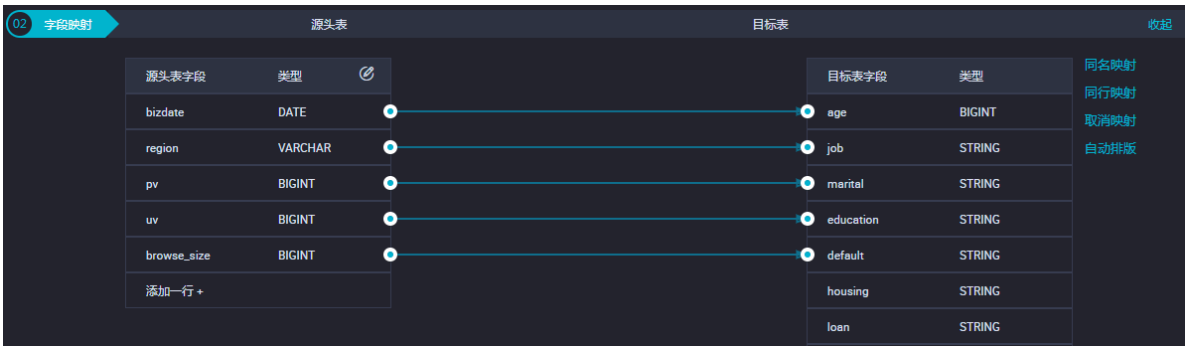
   配置同步任务的数据来源和数据去向。

配置	说明
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。

配置	说明
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 。



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。
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，如’abc’、’123’等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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"
        },
        {//此处以stream为例，如果您需要使用其他插件，可以查找对应的插件填写相应的内容。
            "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"
            }
        ]
    }
} "to":"Writer"
            }
        ]
    }
}
```

补充说明

- **主备同步数据恢复问题**

  主备同步问题指Oracle使用主从灾备，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

  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，Oracle Reader完全不会获取到写入更新数据。

  上述是在Oracle Reader单线程模型下数据同步一致性的特性，Oracle 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

  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不需您指定编码，可以自动获取编码并转码。

- 增量数据同步

  Oracle 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，有以下几种方式：

  - 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提供querySql语句交给您自己实现SELECT抽取语句，Oracle Reader本身对querySql不进行任何安全性校验。

1.7.1.11 配置OSS Reader

本文将为您介绍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。

- 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。

OSS Reader实现了从OSS读取数据并转为数据集成/DataX协议的功能，OSS本身是无结构化数据存储。对于数据集成/DataX而言，目前OSS Reader支持的功能如下所示。

- 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。

OSS Reader支持OSS中的BIGINT、DOUBLE、STRING、DATATIME和BOOLEAN数据类型。

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	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。线程并发数通过通道数指定。 · 当指定通配符时，OSS 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。若遇到此情况，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	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（,）。	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

参数	描述	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。	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。

配置	说明
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。
Object前缀	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系统参数就可以设置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。

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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和**新增任务资源**。

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。

```
{
```

```
    "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"
        },
        {//下面是关于Writer的模板，可以查看相应的写插件文档。
            "stepType":"stream",
            "parameter":{},
            "name":"Writer",
            "category":"writer"
        }
    ],
    "setting":{
        "errorLimit":{
            "record":""//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"
```

```
            }
        ]
    }
}
```

ORC/Parquet文件读取OSS

目前通过复用HDFS Reader的方式完成OSS读取ORC/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，参数含义请参见**配置*HDFS Reader*。

- **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_a
a8e61dc7e4b",
        "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：**

```
{
      "stepType": "oss",
      "parameter": {
        "datasource": "",
        "fileFormat": "parquet",
        "path": "/tests/case61/parquet",
        "parquetSchema": "message test { required int64 int64_col;
\n required binary str_col (UTF8);\nrequired group params (MAP) {\
nrepeated group key_value {\nrequired binary key (UTF8);\nrequired
 binary value (UTF8);\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 }\nrequired group params_arr_complex (LIST) {\
n  repeated group list {\n    required group element {\n required
int64 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_comple
```

```
x {\n  required int64 id;\n required group detail {\n  required
int64 id;\n required binary name (UTF8);\n  }\n  }\n}",
        "column": [
          {
            "index": 0,
            "type": "long"
          },
          {
            "index": "1",
            "type": "string"
          },
          {
            "index": "2",
            "type": "string"
          },
          {
            "index": "3",
            "type": "string"
          },
          {
            "index": "4",
            "type": "string"
          },
          {
            "index": "5",
            "type": "string"
          },
          {
            "index": "6",
            "type": "string"
          },
          {
            "index": "7",
            "type": "string"
          }
        ]
      }
    }
```

1.7.1.12 配置FTP Reader

本文将为您介绍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。

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。

暂时不支持以下两种功能。

- 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。

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	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	必选	默认值
path	远程FTP文件系统的路径信息，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目前只支持*作为文件通配符。  说明: · 通常不建议您使用*，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	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/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： 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

参数	描述	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。	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，通常不需配置此项。	否	无

向导开发介绍

1. 选择数据源。

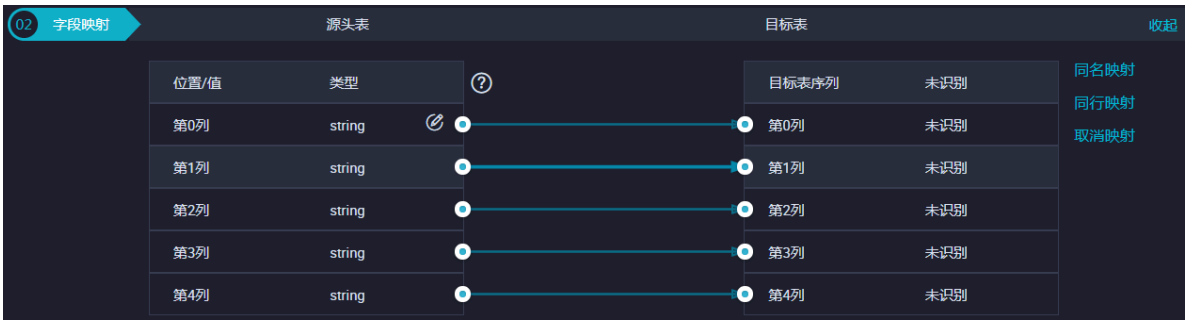
   配置同步任务的数据来源和数据去向。

配置	说明
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。

配置	说明
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，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。



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。

配置	说明
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。

配置	说明
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见#unique_109和新增任务资源。

脚本开发介绍

配置一个从FTP数据库同步抽取数据作业。

```
{
    "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"
        },
        {//Reader的模板如下所示，您可以查看相应数据源的读插件文档。
            "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"
            }
        ]
    }
```

```
}
```

1.7.1.13 配置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举例。

OTS Reader插件实现了从Table Store（OTS）读取数据，通过您指定抽取数据范围可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。

简而言之，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	MySQL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。应用层一般使用Long报错时间的Unix TimeStamp。

参数说明

参数	描述	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

参数	描述	必选	默认值
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	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

参数	描述	必选	默认值
begin/end	该配置项必须配对使用，用于支持Table Store表范围抽取。begin/end中描述的是OTS PrimaryKey的区间分布状态，而且必须保证区间覆盖到所有的 PrimaryKey，需要指定该表下所有的PrimaryKey范围，对于无限大小的区间，可以使用{"type":"INF_MIN"}, {"type":"INF_MAX"}指代。例如对一张主键为[DeviceID, SellerID]的Table Store进行抽取任务，begin/end的配置如下所示。 `"range": { "begin": [ {"type":"INF_MIN"}, //指定 deviceID 最小值 {"type":"INT", "value":"0"} //指定 SellerID 最小值 ], "end": [ {"type":"INF_MAX"}, //指定 deviceID 抽取最大值 {"type":"INT", "value":"9999"} //指定 SellerID 抽取最大值 ] }` 如果要对上述表抽取全表，可以使用如下配置。 `"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 ] }`	是	空
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": {`	否	无

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。

```
{
    "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"
        },
        { //下面是关于Writer的模板，可以找相应的写插件文档
            "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
            "dmu":1//DMU值
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
```

```
}
```

1.7.1.14 配置PostgreSQL Reader

本文将为您介绍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。

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。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，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。**

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，其他类型均不支持。**
- **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，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

参数	描述	是否必选	默认值
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，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tables、columns和splitPk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，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

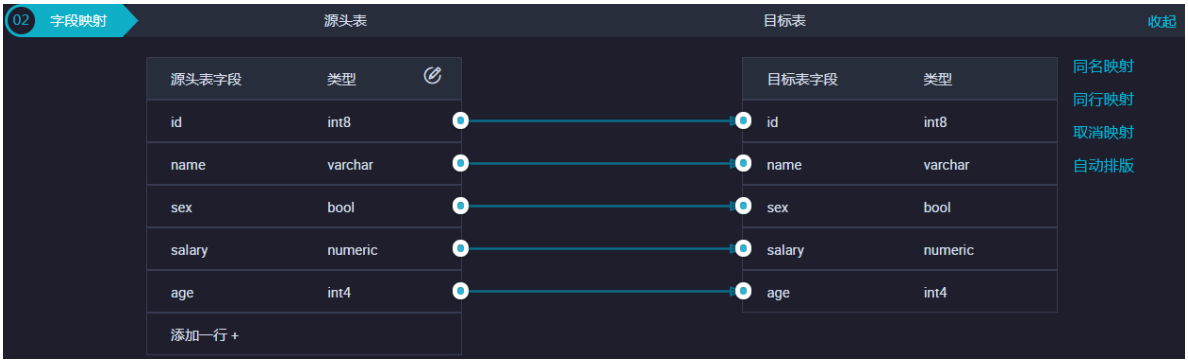
1. **选择数据源。**

   配置同步任务的数据来源和数据去向。

配置	说明
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，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。



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。
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，如’abc’、’123’等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个从PostgreSQL数据库同步抽取数据作业。

```
{
    "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"
        },
        { //下面是关于writer的模板，您可以查找相应的写插件文档。
            "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"
                }
            ]
        }
}
```

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

  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

  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，有以下几种方式：

  - 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。

1.7.1.15 配置SQL Server Reader

本文将为您介绍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。

- 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。

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 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	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，一个作业只能支持一个表同步。	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

参数	描述	必选	默认值
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

向导开发介绍

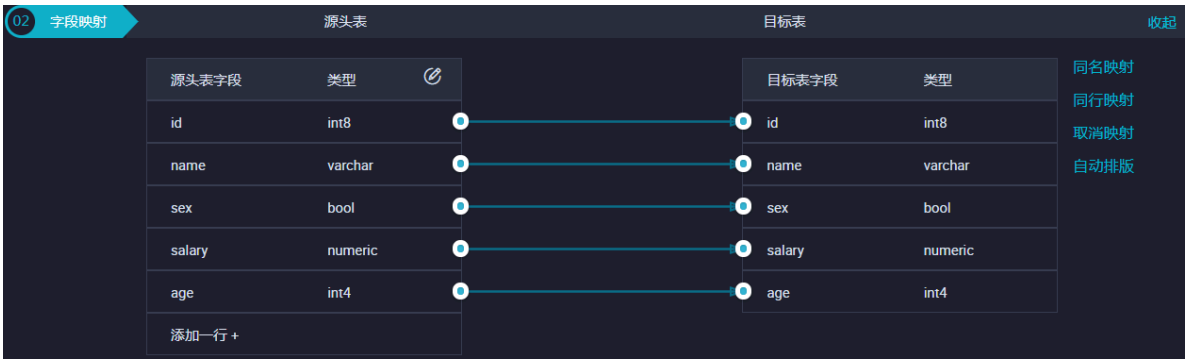
1. **选择数据源。**

   配置同步任务的数据来源和数据去向。

配置	说明
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。



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。

配置	说明
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，如’abc’、’123’等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和**新增任务资源**。

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"
        },
        {//下面是关于Writer的模板，您可以查找相应的写插件文档。
            "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"
            }
        ]
    }
}
```

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

  SQL Server 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，有以下几种方式：

  - 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。

1.7.1.16 配置LogHub Reader

本文将为您介绍LogHub Reader支持的数据类型、字段映射和数据源等参数及配置示例。

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，全面提升海量日志处理/分析能力。LogHub Reader是使用日志服务的Java SDK消费LogHub中的实时日志数据，并将日志数据转换为数据集成传输协议传递给Writer。

实现原理

LogHub Reader通过日志服务Java SDK消费LogHub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，每个分区可提供一定的服务能力。

- 写入：5MB/s，2000次/s。
- 读取：10MB/s，100次/s。

LogHub 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 Reader针对LogHub类型的转换列表，如下所示。

DataX内部类型	LogHub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所在的阿里云区域（Region）及Project名称相关。各Region的服务入口请参见#unique_113。	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字，是日志服务中的资源管理单元，用于资源隔离和控制。	是	无
logstore	目标日志库的名字，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column的名字，这里可以配置日志服务中的元数据作为同步列，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

参数	描述	是否必选	默认值
beginDateTime	数据消费的开始时间位点，即日志数据到达Loghub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。 说明： beginDateTime和endDateTime需要互相组合配套使用。	和beginTimestampMillis选择一种	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。 说明： 请尽量保证周期之间重合：即上周期的endDateTime时间和下周期的beginDateTime时间一致，或比下周期的beginDateTime时间晚。否则，可能造成部分区域数据无法拉取。	和endTimestampMillis选择一种	无
beginTimestampMillis	数据消费的开始时间位点。该参数为时间范围（左闭右开）的左边界，单位毫秒。 说明： beginTimestampMillis和endTimestampMillis组合配套使用。 -1表示日志服务游标的最开始CursorMode.BEGIN。推荐使用beginDateTime模式。	和beginDateTime选择一种	无
endTimestampMillis	数据消费的结束时间位点，为时间范围（左闭右开）的右边界，单位毫秒。 说明： endTimestampMillis和beginTimestampMillis组合配套使用。 -1表示日志服务游标的最后位置CursorMode.END。推荐使用endDateTime模式。	和endDateTime选择一种	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。

配置	说明
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。
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。

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。
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，如'abc'、'123'等。 • 可以配合调度参数使用，如${bizdate}等。 • 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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。

```
{
 "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",
                 "=Topic",//日志主题。
                 "HostName",//主机名。
                 "Path",//路径。
                 "LogTime"//日志时间。
             ],
             "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。
             "encoding":"UTF-8",//编码格式。
             "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":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"
            }
        ]
    }
}
```

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__`。

1.7.1.17 配置OTSReader-Internal

本文将为您介绍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，详情请参见配置*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	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

参数	描述	必选	默认值
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
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

参数	描述	必选	默认值
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

参数	描述	必选	默认值
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)
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

- **多版本模式**

```
{
    "type": "job",
    "version": "1.0",
    "configuration": {
        "reader": {
            "plugin": "otsreader-internalreader",
            "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": "1.0",
    "configuration": {
        "reader": {
            "plugin": "otsreader-internalreader",
            "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": {}
}
```

1.7.1.18 配置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。

OTSStream Reader插件主要用于Table Store增量数据的导出，增量数据可以看作操作日志，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。您可以在建表的时候指定开启，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，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有PutRow、UpdateRow和DeleteRow三类数据更改操作。

- 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Datax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Datax的数据格式后，定义了4种操作类型，如下所示：

- 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，示例如下。

```
"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	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

参数	描述	必选	默认值
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	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。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	增量数据的时间范围（左闭右开）的左边界，格式为yyyymmddhh24miss，单位为毫秒。	否	无
endTimeString	增量数据的时间范围（左闭右开）的右边界，格式为yyyymmddhh24miss，单位为毫秒。	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	增量数据的时间范围（左闭右开）的左边界。
结束时间	增量数据的时间范围（左闭右开）的右边界。
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。

参数	描述
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，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	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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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。

```
{
    "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}",//增量数据的时间范围（左闭右开）的左边界。
                "table":"",//表名。
                "endTimeString":"${endTime}"//增量数据的时间范围（左闭右开）的右边界。
            },
            "name":"Reader",
            "category":"reader"
        },
        {
            "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"
            }
        ]
    }
}
```

1.7.1.19 配置RDBMS Reader

本文为您介绍RDBMS 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。

RDBMS Reader插件从RDBMS读取数据。在底层实现上，RDBMS Reader通过JDBC连接远程RDBMS数据库，并执行相应的SQL语句，从RDBMS库中读取数据。目前RDBMS Reader支持读取DM、DB2、PPAS和Sybase等数据库的数据。RDBMS 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。

RDBMS Reader通过JDBC连接器连接至远程的RDBMS数据库，并根据您配置的信息生成查询SQL语句，发送至远程RDBMS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，RDBMS Reader将其拼接为SQL语句发送至RDBMS数据库。
- 对于您配置的`querySql`信息，RDBMS直接将其发送至RDBMS数据库。

RDBMS 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	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RDBMS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的格式是不同的，DataX会根据具体JDBC的格式选择合适的数据库驱动完成数据读取。 · DM格式：jdbc:dm://ip:port/database · DB2格式：jdbc:db2://ip:port/database · PPAS格式：jdbc:edb://ip:port/database RDBMS Writer可以通过以下方式增加新的数据库支持。 · 进入RDBMS Reader对应目录，${DATAX_HOME}为DataX主目录，即*${DATAX_HOME}/plugin/reader/rdbmswriter*。 · 在RDBMS 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。RDBMS 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rdbmsreader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	是	无

参数	描述	必选	默认值
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	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	RDBMS 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，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，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，RDBMS Reader将报错。 · 如果不填写splitPk，将视作您不对单表进行切分，RDBMS Reader使用单通道同步全量数据。	否	空
where	筛选条件，RDBMS 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，往往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

参数	描述	必选	默认值
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，RDBMS 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RDBMS数据库同步抽取数据作业。

```
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,
            "throttle": false
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "column": [
                    {
                        "type": "string",
```

```
                        "value": "field"
                    },
                    {
                        "type": "long",
                        "value": 100
                    },
                    {
                        "dateFormat": "yyyy-MM-dd HH:mm:ss",
                        "type": "date",
                        "value": "2014-12-12 12:12:12"
                    },
                    {
                        "type": "bool",
                        "value": true
                    },
                    {
                        "type": "bytes",
                        "value": "byte string"
                    }
                ],
                "sliceRecordCount": "10"
            },
            "stepType": "stream"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {
                "connection": [
                    {
                        "jdbcUrl": "jdbc:dm://ip:port/database",
                        "table": [
                            "table"
                        ]
                    }
                ],
                "username": "username",
                "password": "password",
                "table": "table",
                "column": [
                    "*"
                ],
                "preSql": [
                    "delete from XXX;"
                ]
            },
            "stepType": "rdbms"
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

1.7.1.20 配置Stream Reader

本文将为您介绍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	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"
        },
        { //下面是关于Writer的模板，您可以查找相应的写插件文档。
            "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"
            }
        ]
    }
```

```
}
```

1.7.1.21 配置HybridDB for MySQL Reader

本文将为您介绍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。

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	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Job只能同步一张表。	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

参数	描述	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，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

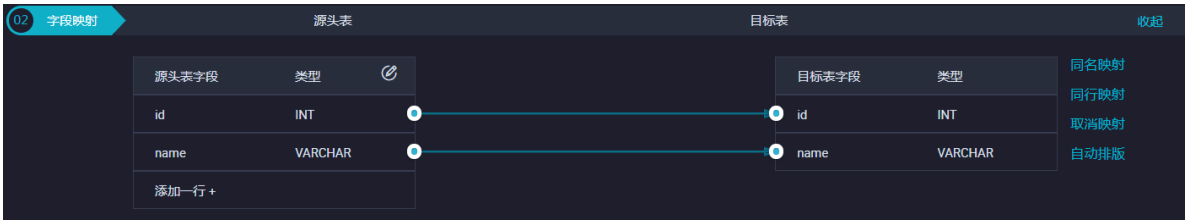
1. **选择数据源。**

   配置同步任务的数据来源和数据去向。

配置	说明
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，即可单击删除按钮进行删除。



配置	说明
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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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,//同步速度限流。
            "mbps": 1,//限流值。
        }
    }
```

```
}
```

1.7.1.22 配置AnalyticDB for PostgreSQL Reader

本文将为您介绍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提供AnalyticDB for PostgreSQL存储引擎。

AnalyticDB for PostgreSQL Reader通过JDBC连接器连接到远程的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	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	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

参数	描述	必选	默认值
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。 · 因为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
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

参数	描述	必选	默认值
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

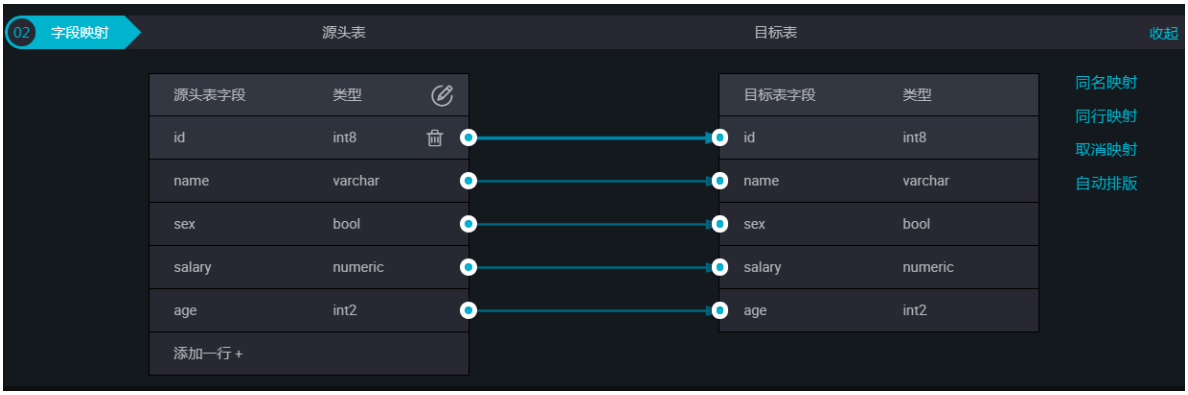
1. 选择数据源。

   配置同步任务的数据来源和数据去向。

配置	说明
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，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，即可单击删除按钮进行删除。



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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和**新增任务资源**。

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": false,//同步速率是否限流。
        }
    }
}
```

1.7.1.23 配置POLARDB Reader

本文将为您介绍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	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

参数	描述	必选	默认值
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	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。

配置	说明
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。

配置	说明
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，即可单击删除按钮进行删除。

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。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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": false,//同步速率限流。
        }
    }
}
```

1.7.1.24 配置Elasticsearch Reader

本文将为您介绍Elasticsearch Reader的工作原理、功能和参数。

工作原理

- **通过Elasticsearch的_search+scroll+slice（即游标+分片）方式实现，slice结合DataX job的task多线程分片机制使用。**
- **根据Elasticsearch中的mapping配置，进行数据类型转换。**

更多详情请参见*Elasticsearch*官方文档。

基本配置

```
{
    "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":"", //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。

- **支持半结构化到结构化数据的提取**

分类	说明
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。

分类	说明
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
endpoint	Elasticsearch的连接地址。	是	无
username	http auth中的username。	否	空
password	http auth中的password。	否	空
index	Elasticsearch中的index名。	是	无
type	Elasticsearch中index的type名。	否	index名
pageSize	每次读取数据的条数。	否	100
search	Elasticsearch的query参数。	是	无
scroll	Elasticsearch的分页参数，设置游标存放时间。	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

参数	描述	是否必选	默认值
connTimeOut	客户端连接超时时间。	否	600000
readTimeOut	客户端读取超时时间。	否	600000
multiThread	http请求，是否有多线程。	否	true
column	Elasticsearch所支持的字段类型，样例中包含了全部。	是	无
full	是否支持将Elasticsearch的数据拉取为一个字段。	否	false
multi	是否支持将数组进行列拆多行的处理，需要辅助设置子属性。	否	false

补充配置：

```
"full":false,
        "multi": {
          "multi": true,
          "key":"crn_list[*]"
        }
```

1.7.1.25 配置AnalyticDB for MySQL 2.0 Reader

本文将为您介绍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

AnalyticDB for MySQL 2.0类型	DataX类型	MaxCompute类型
SMALLINT	LONG	INT
INT	LONG	INT
FLOAT	STRING	DOUBLE
DOUBLE	STRING	DOUBLE
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

参数	描述	是否必选	默认值
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
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": "*********",
                    "accessKey": "*********",
                    "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": false,
        }
    }
```

```
}
```

1.7.1.26 配置Kafka Reader

Kafka Reader通过Kafka服务的Java SDK从Kafka读取数据。

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。

消费消息的详情参见订阅者最佳实践。

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，您可以参见Kafka官方了解接口的功能和限制。

- **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	说明	是否必填
server	Kafka的broker 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

参数	说明	是否必填
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，可以和#unique_118配合使用。Kafka 0.10.2以上的版本支持此功能。	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，可以和#unique_118配合使用。Kafka 0.10.2以上的版本支持此功能。	需要和endOffset二选一。 说明: endDateTime和beginDateTime配合使用。
beginOffset	数据消费的开始时间位点，您可以配置以下形式。 · 例如15553274的数字形式，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，用来控制什么时候应该结束数据消费任务退出。	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用来控制这些多余的数据是否写出到目的端。由于消费数据使用了自动点位提交，建议： · Kafka 0.10.2之前版本：建议skipExceedRecord配置为false。 · Kafka 0.10.2及以上版本：建议skipExceedRecord配置为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。

参数	说明	是否必填
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

kafkaConfig参数说明如下：

- fetch.min.bytes：指定消费者从broker获取消息的最小字节数，即等到有足够的数据时才把它返回给消费者。
- fetch.max.wait.ms：等待broker返回数据的最大时间，默认500ms。fetch.min.bytes和fetch.max.wait.ms哪个条件先得到满足，便按照哪种方式返回数据。
- 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1MB。
- 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30s。
- 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），可更改为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。
- 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。

配置示例如下：

```
{
    "group.id": "demo_test",
    "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": ""
}
```

脚本开发示例

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": false,
            "concurrent": 1,
            "dmu": 1
        }
    }
}
```

1.7.1.27 配置InfluxDB Reader

InfluxDB是由InfluxData开发的开源时序型数据库，它由Go写成，致力于高性能地查询与存储时序型数据。InfluxDB Reader插件实现了从InfluxDB读取数据。

目前InfluxDB Reader仅支持脚本模式配置，更多详情请参见*InfluxDB*。

实现原理

在底层实现上，InfluxDB Reader通过Java Client，将SQL查询请求发送到InfluxDB实例，扫描出指定的数据点。整个同步的过程通过Database、Metric和时间段进行切分，组合为一个迁移Task。

约束限制

- **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InfluxDB 0.9及以上版本。**

支持的数据类型

类型分类	数据集成column配置类型	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	InfluxDB的HTTP连接地址。	是，格式为`http://IP:Port`。	无
database	指定InfluxDB的数据库。	是	无
username	用于连接InfluxDB的账号。	是	无
password	用于连接InfluxDB的密码。	是	无

参数	描述	是否必选	默认值
column	数据迁移任务需要迁移的Metric列表。	是	无
beginDateTime	和endDateTime配合使用，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	和beginDateTime配合使用，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InfluxDB数据库同步的作业。

```json
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,
            "throttle": true
        }
    },
```

```
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "endpoint": "http://host:8086",
                "database": "",
                "username": "",
                "password": "",
                "column": [
                    "xc"
                ],
                "endDateTime": "20190515180000",
                "beginDateTime": "20190515170000"
            },
            "stepType": "influxdb"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {},
            "stepType": ""
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

1.7.1.28 配置OpenTSDB Reader

OpenTSDB是主要由Yahoo维护、可扩展、分布式的时序数据库，OpenTSDB Reader插件实现了从OpenTSDB读取数据。

OpenTSDB与阿里巴巴自研TSDB的关系与区别，请参见相比*OpenTSDB*优势。

目前OpenTSDB Reader仅支持脚本模式配置方式。

实现原理

在底层实现上，OpenTSDB Reader通过HTTP请求连接到OpenTSDB实例，用`/api/config`接口获取其底层存储HBase的连接信息。然后通过AsyncHBase框架连接HBase，以Scan的方式将数据点扫描出来。整个同步的过程通过Database、Metric和时间段进行切分，即某个Metric在某一个小时内的数据迁移，组合成一个迁移Task。

约束限制

- **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OpenTSDB 2.3.x版本。**

- 不可直接使用/api/query查询获取数据点，需要连接OpenTSDB的底层存储。

  因为通过OpenTSDB的HTTP接口（/api/query）读取数据，在数据量较大的情况下，会导致OpenTSDB的异步框架报CallBack过多的异常。所以通过连接底层HBase存储，以Scan的方式扫描数据点，可避免此问题。且通过指定Metric和时间范围，可顺序扫描HBase表，提高查询效率。

支持的数据类型

类型分类	数据集成column配置类型	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	OpenTSDB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	和endDateTime配合使用，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	和beginDateTime配合使用，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OpenTSDB数据库同步抽取数据到本地的作业。

```json
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,
            "throttle": true
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                "endpoint": "http://host:4242",
                "column": [
                    "xc"
                ],
                "beginDateTime": "20190101000000",
                "endDateTime": "20190101030000"
            },
            "stepType": "opentsdb"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {},
            "stepType": ""
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

```
```

性能报告

- **性能数据特征**

  从Metric、时间线、Value和采集周期四个方面进行描述。

  - **Metric：指定一个Metric为m。**
  - **tagkv：前四个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	215,428	25.65
2	424,994	50.60
3	603,132	71.81

1.7.1.29 配置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。

目前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。

约束限制

- 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column配置类型	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	和endDateTime配合使用，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)。

参数	描述	是否必选	默认值
endDateTime	和beginDateTime配合使用，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。

```
```json
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,
            "throttle": true
        }
    },
    "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

1.7.1.30 配置AnalyticDB for MySQL 3.0 Reader

本文将为您介绍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

参数	描述	是否必选	默认值
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

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

向导开发介绍

1. 选择数据源。

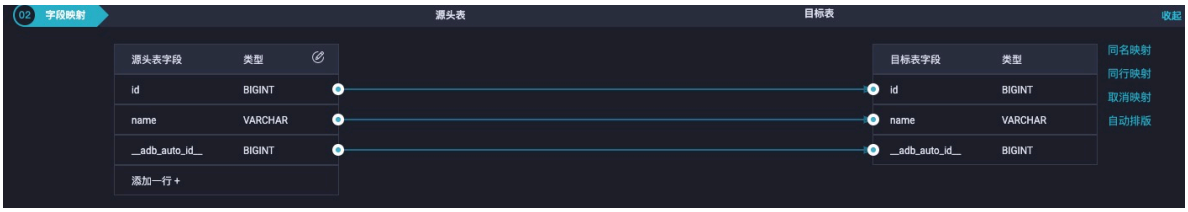
   配置同步任务的数据来源和数据去向。

配置	说明
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。

配置	说明
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 。



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。
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，如’abc’、’123’等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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"
```

```
        },
        { //下面是关于writer的模板，您可以查找相应的写插件文档。
            "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": {
            "concurrent": 2, //任务记录数。
            "throttle": false //false代表不限流，下面的限流的速度不生效，
true代表限流。
        }
    }
}
```

1.7.1.31 配置MetaQ Reader

本文将为您介绍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。

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	是否必填
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
domainNam	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（从一开始）和lastRead（上次读取的offset）	是

参数	描述	是否必填
nullCurrentOffset	上次Offset为空时，开始读取的位置，支持begin（从一开始）和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

功能说明

配置一个从消息队列读取数据的示例，详情请参见上述参数说明。

```
{
    "job": {
        "content": [
            {
                "reader": {
                    "name": "metaqreader",
                    "parameter": {
                        "accessId": "xxxxxxxxxx",
                        "accessKey": "xxxxxxxxxxxxxxx",
                        "consumerId": "Test01",
                        "topicName": "test",
                        "subExpression": "*",
                        "onsChannel": "ALIYUN",
                        "domainName": "xxx.aliyun.com",
                        "contentType": "singlestringcolumn",
                        "beginOffset": "lastRead",
                        "nullCurrentOffset": "begin",
                        "fieldDelimiter": ",",
                        "column": [
                            "col0"
                        ],
                        "fieldDelimiter": ","
                    }
                },
                "writer": {
                    "name": "streamwriter",
                    "parameter": {
                        "print": false
                    }
                }
            }
        ]
    }
```

```
}
```

1.7.1.32 配置Hive Reader

Hive Reader插件实现了从Hive读取数据的功能，本文将为您介绍Hive Reader的工作原理、参数和示例。

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，并提供类SQL查询功能。

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。**
- **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 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底层的逻辑和HDFS Reader插件一致，完成数据的读取后，您可以在Hive Reader插件参数中配置HDFS Reader相关的参数，配置的参数会透传给HDFS Reader插件。

参数说明

参数	描述	必选	默认值
jdbcUrl	Hive元数据库的地址。目前Hive Reader仅支持访问MySQL类型的Hive元数据库。 您需要确保任务执行节点具备Hive元数据库的网络和访问权限。	是	无
username	Hive元数据库的用户名。	是	无
password	Hive元数据库的密码。	是	无

参数	描述	必选	默认值
column	需要读取的字段列，例如"column": ["id", "name"]。 · 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。 · column必须显示指定同步的列集合，不允许为空。	是	无
table	需要读取的Hive表名。 说明: 请注意大小写。	是	无
partition	· 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。 · 如果您的Hive表是非分区表，则无需配置partition。	否	无

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

配置一个从Hive读取数据的JSON示例。

```
{
    "order": {
    "hops": [
    {
```

```
    "from": "Reader",
    "to": "Writer"
    }
    ]
    },
    "setting": {
    "errorLimit": {
    "record": "0"
    },
    "speed": {
    "concurrent": 1,   //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。
    }
    },
    "steps": [
    {
    "category": "reader",
    "name": "Reader",
    "parameter": {
    "username": "",  //Hive元数据库的用户名。
    "password": "", //Hive元数据库的密码。
    "jdbcUrl": "jdbc:mysql://host:port/database", //Hive元数据库的地址。

    "table": "",  //需要读取的Hive表名。
    "partition": "",  //如果是分区表，需要配置分区信息。
    "column": [
    "id",
    "name"
    ]
    },
    "stepType": "hive"  //插件名。
    },
    {
    "category": "writer",
    "name": "Writer",
    "parameter": {},
    "stepType": "stream"
    }
    ],
    "type": "job",
    "version": "2.0"  //版本号。
    }
```

1.7.1.33 配置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。本文将为您介绍Vertica Reader的实现原理、参数和示例。

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。

实现原理

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

参数	描述	是否必选	默认值
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
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

参数	描述	是否必选	默认值
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

从Vertica读取数据的JSON配置，如下所示。

```
{
    "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": 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。
        }
    }
}
```

1.7.1.34 配置SAP HANA Reader

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。您可以直接对大量实时业务数据进行查询和分析，无需对业务数据进行建模、聚合等操作。本文将为您介绍SAP HANA Reader支持的参数及配置示例。

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（*）。**
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。**

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HANA同步任务至MaxCompute的示例。

```
{
    "type": "job",
    "steps": [
        {
            "stepType": "saphana",
            "parameter": {
                "username": "用户名",
                "password": "密码",
                "column": [
                    "字段名1",
                    "字段名2",
                    "字段名3"
                ],
                "connection": [
                    {
                        "table": [
                            "需要同步的表名"
                        ],
                        "jdbcUrl": [
                            "jdbc:sap://127.0.0.1:30215?currentschema=
TEST"
                        ]
                    }
                ],
                "splitPk": "字段1" //splitPk代表的字段进行数据分片。
            },
            "name": "Reader",
            "category": "reader"
        },
        {
            "stepType": "odps",
            "parameter": {
                "partition": "",
                "truncate": true,
                "datasource": "example", //数据源。
                "column": [
                    "*"
                ],
                "table": ""
            },
            "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": false //false代表不限流，下面的限流的速度不生效，
true代表限流。
        }
    }
}
```

1.7.2 配置Writer插件

1.7.2.1 配置AnalyticDB for MySQL 2.0 Writer

本文将为您介绍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。

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。

如果数据源来源是RDS for SQLServer，详细的导入操作，请参见使用数据集成迁移。

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。

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

参数	描述	必选	默认值
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（请参见MySQL支持的JDBC控制参数）。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，排序完成后再提交到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批量提交到ADB中，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。

配置项	说明
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。
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`。 · 写入前保留已有数据：导数据之前，不清理任何数据，每次运行数据都是追加进去的，相当于`insert into`。
一级分区	默认，不可修改。

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。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。

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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

```
{
    "type":"job",
    "version":"2.0",
    "steps":[ //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
        {
            "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":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"
            }
        ]
```

```
    }
}
```

1.7.2.2 配置DataHub Writer

本文将为您介绍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。

说明:

STRING字符串仅支持UTF-8编码，单个STRING列最长允许1MB。

参数配置

通过Channel将Source与Sink连接起来，所以在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。** **详情请参见*DataHub*访问域名。**	**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**

参数	描述	是否必选	默认值
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的工作空间相互独立，您在MaxCompute中创建的项目不能复用于DataHub，需要独立创建。	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。 详情请参见*Project*及*Topic*的数量限制。	是	无
maxCommitSize	为提高写出效率，DataX-On-Flume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1MB数据。	否	1MB
batchSize	为提高写出效率，DataX-On-Flume会积累Buffer数据，待积累的数据条数达到batchSize大小（单位条数）时，批量提交到目的端。默认1,024，即1,024条数据。	否	1,024
maxCommitInterval	为提高写出效率，DataX-On-Flume会积累Buffer数据，待积累的数据条数达到maxCommitSize、batchSize大小限制时，批量提交到目的端。 如果数据采集源头长时间没有产出数据，为了保证数据的及时投递，增加了maxCommitInterval参数（单位毫秒），即Buffer数据的最长时间，超过此时间会强制投递。默认30,000，即30秒。	否	30,000
parseMode	日志解析模式，目前有不解析default模式和csv模式。不解析即采集到的一行日志，直接作为DataX的Record单列Column写出。csv模式支持配置一个列分隔符，一行日志通过分隔符分隔成DataX的Record的多列。	否	default

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": [
        { //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "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，批量提交到目的端。
            },
            "name": "Writer",
            "category": "writer"
        }
    ],
    "setting": {
        "errorLimit": {
            "record": ""//错误记录数。
        },
        "speed": {
            "concurrent": 20,//并发线程数。
            "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"
            }
        ]
    }
}
```

1.7.2.3 配置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。

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，将数据写入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，是否需要其他权限，取决于您配置任务时在preSql和postSql中指定的语句。

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	—
二进制类	BLOB

参数说明

参数	描述	必选	默认值
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 。	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

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

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。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        { //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"db2",//插件名。
            "parameter":{
                "postSql":[],//执行数据同步任务之前率先执行的SQL语句。
                "password":"",//密
                "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后执行的SQL语句。
            },
            "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"
            }
        ]
    }
}
```

1.7.2.4 配置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。

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。**

DRDS Writer面向ETL开发工程师，使用DRDS Writer从数仓导入数据至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）的语句写入数据至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，是否需要其他权限，取决于您配置任务时在preSql和postSql中指定的语句。

类似于MySQL Writer，目前DRDS 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。

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

类型分类	DRDS数据类型
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，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
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。 • 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。 • 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。	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

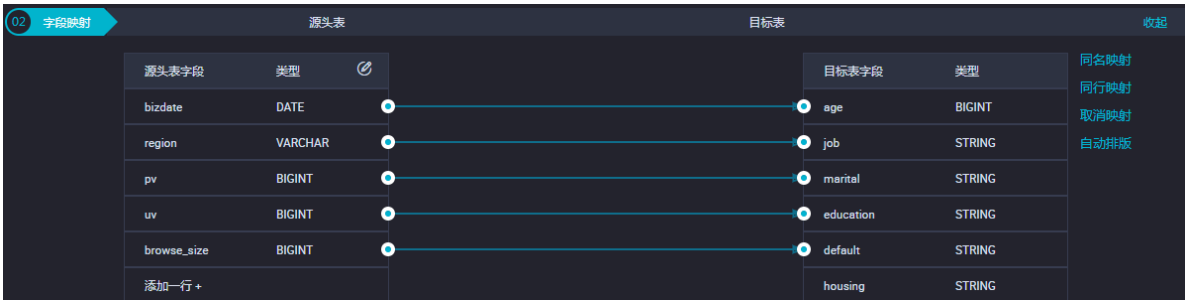
1. **选择数据源。**

   配置同步任务的数据来源和数据去向。

配置	说明
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。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。



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。

配置	说明
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个写入DRDS的数据同步作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
```

```
    "steps":[
        {//下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "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":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,//并发数。
                }
            },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
                }
            ]
        }
    }
```

1.7.2.5 配置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	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
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

参数	描述	必选	默认值
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	txt写出时的表头，例如['id', 'name', 'age']。	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。

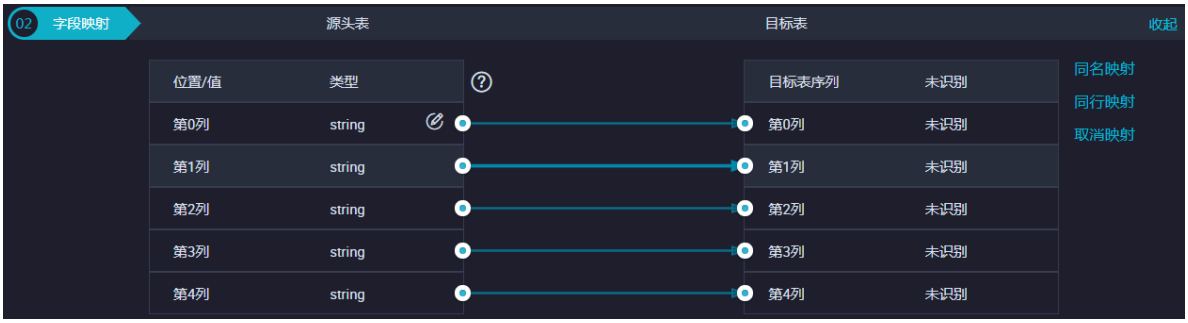
   配置同步任务的数据来源和数据去向。

配置	说明
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。

配置	说明
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。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。



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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个写入FTP数据库的同步作业。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        { //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "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":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"
                }
            ]
        }
}
```

1.7.2.6 配置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。

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，其他类型均不支持。

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	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	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
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	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)。** `"range": { "startRowkey": "aaa", "endRowkey": "ccc", "isBinaryRowkey":false }` `"column": [ { "index":1, "name": "cf1:q1", "type": "string" }, { "index":2, "name": "cf1:q2", "type": "string" } ]`	否	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**	否	无

参数	描述	必选	默认值
walFlag	HBae 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HBae Client的写Buffer大小，单位字节，配合autoflush使用。 autoflush： · 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。

```
{
    "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"
                    }
                ],
                "writeMode":"api",//写入模式。
                "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":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"
            }
        ]
    }
```

```
}
```

1.7.2.7 配置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举例。

HBase11xsql Writer实现了向Hbase中的SQL表（phoenix）批量导入数据的功能。Phoenix因为对rowkey做了数据编码，所以直接使用HBaseAPI进行写入会面临手工数据转换的问题，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。

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

参数	描述	是否必选	默认值
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：跳过这一列，即不插入这一列（如果该行的这一列之前已经存在，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。

```
{
  "type": "job",
  "version": "1.0",
  "configuration": {
    "setting": {
      "errorLimit": {
        "record": "0"
      },
      "speed": {
        "mbps": "1",
        "concurrent": "1"
      }
    },
    "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"
```

```
    }
  }
}
```

约束限制

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。

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，用这个值除以行大小，即可得到batchSize。

1.7.2.8 配置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*HDFS*数据源。

说明：

HBase1.1.x插件目前可以兼容HBase 2.0，如果您在使用上遇到问题请提交工单。

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。

功能限制

- 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。

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	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。默认资源组不支持Hadoop高级参数HA的配置，请新增任务资源。	是	无
fileType	文件的类型，目前仅支持您配置为text、orc和parquet。 • text：表示TextFile文件格式。 • orc：表示ORCFile文件格式。 • parquet：表示普通parquet file文件格式。	是	无
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

参数	描述	必选	默认值
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。 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，orc类型文件支持SNAPPY压缩类型（需要您安装SnappyCodec）。	否	无
encoding	写文件的编码配置。	否	无压缩

参数	描述	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。默认资源组不支持Hadoop高级参数HA的配置，请新增任务资源。 `"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`	否	无

参数	描述	必选	默认值
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。 说明: 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

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": [
        { //以下为Writer模板，您可以查找相应数据源的写插件文档。
            "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"
            }
        ]
    }
}
```

1.7.2.9 配置MaxCompute Writer

本文将为您介绍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。

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。

说明:

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。MaxCompute的详情请参见*#unique_110*。

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:

当数据中有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，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	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

参数	描述	是否必选	默认值
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无法做到原子性，所以truncate选项不是原子操作。因此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。 针对这类问题，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

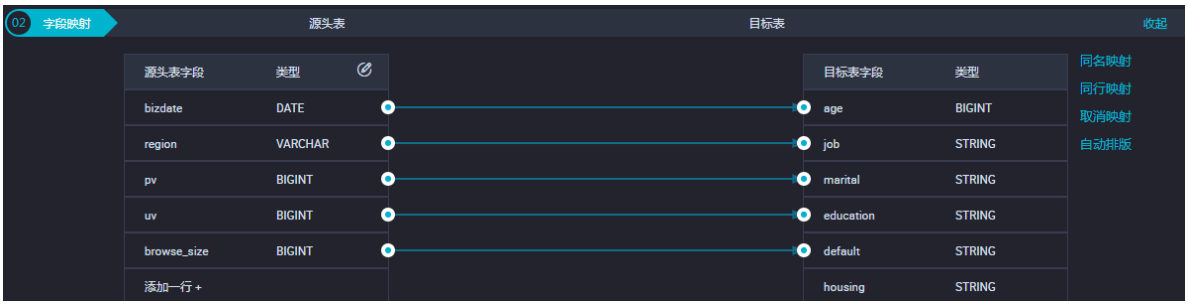
1. 选择数据源。

   配置同步任务的数据来源和数据去向。

参数	描述
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。

参数	描述
分区信息	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中所有的分区。 · "partition":"pt=top?"中的?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。
清理规则	· 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。 · 写入前保留已有数据：导数据之前，不清理任何数据，每次运行数据都是追加进去的，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，没有使用MaxCompute的insert SQL语句进行数据写出。数据同步任务执行成功后，方可对表可见完整数据。请注意建立好任务依赖关系。
空字符串是否作null	默认值为否。

2. 字段映射，即上述参数说明中的column。



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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

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":[
        {//下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "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":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"
            }
        ]
    }
}
```

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":"**********",
```

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3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3级分区，那么在分区配置中就必须指明写入到某个三级分区，例如把数据写入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，这样可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，因为其他的异常导致了任务中断，便不能保证数据的原子性，数据不会回滚也不会自动重跑，需要您利用幂等性这一特点重跑，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。

1.7.2.10 配置Memcache（OCS） Writer

本文将为您介绍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**
`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**

参数	描述	是否必选	默认值
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。

```
{
    "type":"job",
    "version":"2.0",//版本号。
    "steps":[
        { //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
            "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":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"
            }
        ]
    }
}
```

1.7.2.11 配置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。

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，写入到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，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
collectionName	MonogoDB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。 · 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，一般是指Monogo中的主键）。	否	无

参数	描述	是否必选	默认值
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。 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读取或写出阶段。preSql本身不会影响读取和写出的数据内容。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），详情请参见MongoDB查询语法。 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);`	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": {
            "jvmOption": "-Xms1024m -Xmx1024m",
            "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": "1"//限流的速度。
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
}
```

1.7.2.12 配置MySQL Writer

本文将为您介绍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。

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，将数据写入MySQL。数据库本身采用InnoDB引擎，以将数据分批次提交入库。

说明:

- **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。**
- **目前MySQL Writer暂不支持MySQL 8.0及以上版本。**

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，判断是否需要其他权限。

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	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 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。 • 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。 • 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。	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。 说明: 当有多条SQL语句时，不支持事务。	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

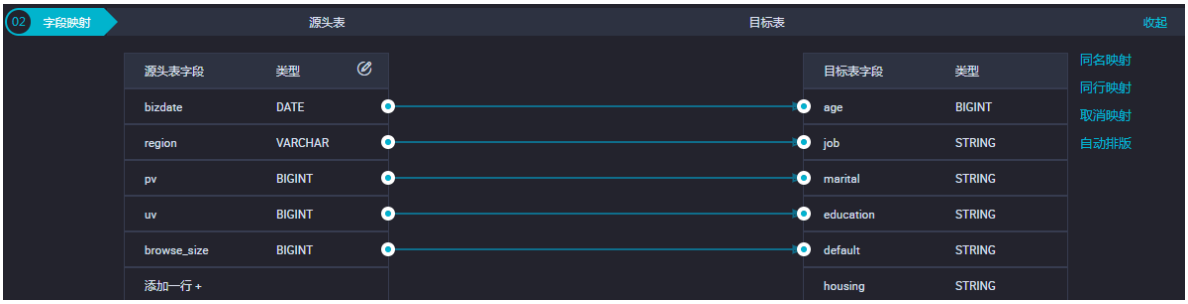
1. 选择数据源。

   配置同步任务的数据来源和数据去向。

配置	说明
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。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。



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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

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":[ //下面是关于Writer的模板，您可以查找相应数据源的写插件文档。
        {
            "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。
                "batchSize":1024,//一次性批量提交的记录数大小。
                "table":"",//表名。
                "preSql":[]//导入前的准备语句。
            },
            "name":"Writer",
```

```
                "category":"writer"
            }
        ],
        "setting":{
            "errorLimit":{//错误记录数。
                "record":"0"
            },
            "speed":{
                "throttle":false,//是否限流。
                "concurrent":1,//并发数。
            }
        },
        "order":{
            "hops":[
                {
                    "from":"Reader",
                    "to":"Writer"
                }
            ]
        }
}
```

1.7.2.13 配置Oracle Writer

本文将为您介绍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

类型分类	Oracle数据类型
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

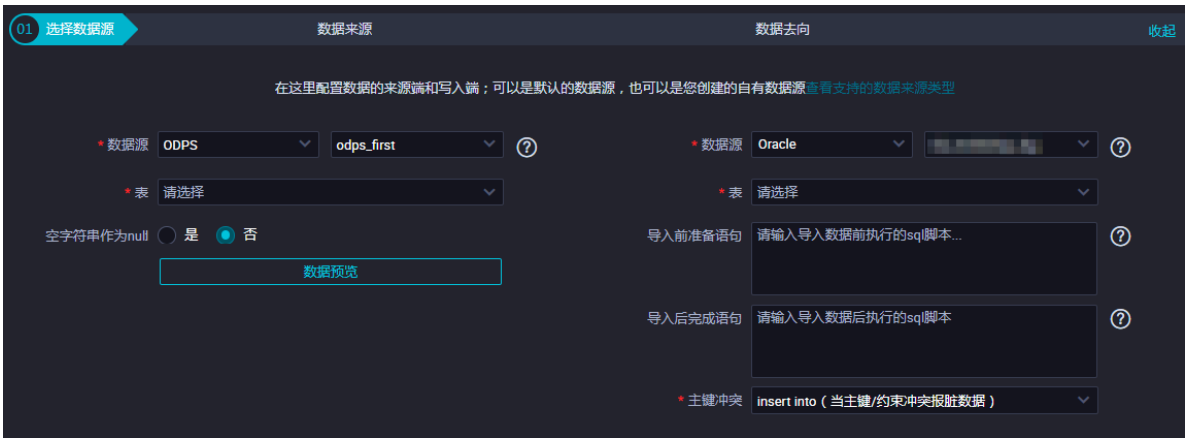
参数说明

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，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，目前仅支持insert into方式。 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。	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

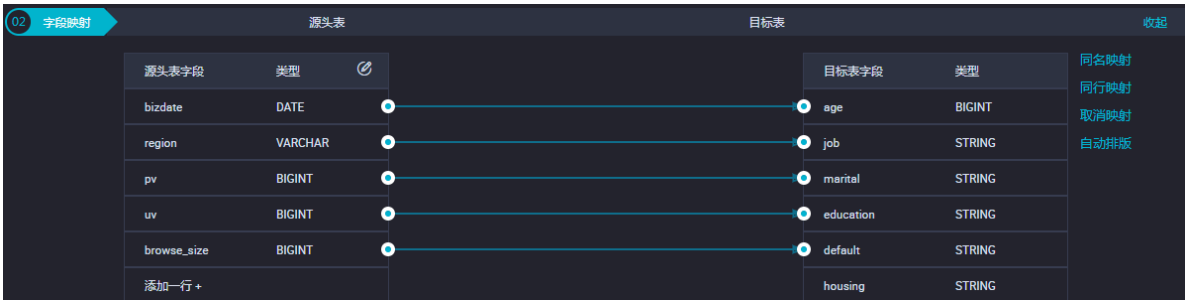
1. **选择数据源。**

   配置同步任务的数据来源和数据去向。



配置	说明
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。**



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。

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置一个写入Oracle的作业。

```
{
    "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":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,//并发数。
            }
        },
        "order":{
            "hops":[
                {
                    "from":"Reader",
                    "to":"Writer"
                }
            ]
        }
}
```

1.7.2.14 配置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。

OSS Writer插件提供了向OSS写入类CSV格式的一个或者多个表文件的功能，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。如果您想对OSS产品有更深入的了解，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*。

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支持的功能如下所示：

- **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。当文件大于某个行数值时，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，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 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写入前数据清理处理。 - 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

参数	描述	是否必选	默认值
fileFormat	文件写出的格式，包括csv和text。 · 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进行转义。	否	text
fieldDelimiter	读取的字段分隔符。	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"，数据同步系统会视作null字段。	否	无
header（高级配置，向导模式不支持）	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log4j日志打印时根据日志文件大小轮转。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	100,000MB
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。

配置	说明
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。**
Object前缀	**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`**和**`text`**。**
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。**
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"`**。**

配置	说明	
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，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。

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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

脚本配置示例如下所示，具体参数填写请参见参数说明。

```
{
    "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。
                "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":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"
            }
        ]
    }
}
```

ORC/Parquet文件写入OSS

目前通过复用HDFS Writer的方式完成OSS写ORC/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，参数含义请参见配置*HDFS Writer*。

- **以ORC文件格式写入OSS，示例如下：**

```
{
      "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);\nrequired group params (MAP) {\
nrepeated group key_value {\nrequired binary key (UTF8);\nrequired
 binary value (UTF8);\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 }\nrequired group params_arr_complex (LIST) {\
n  repeated group list {\n    required group element {\n required
int64 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_comple
x {\n  required int64 id;\n required group detail {\n  required
int64 id;\n required binary name (UTF8);\n  }\n  }\n}",
        "dataxParquetMode": "fields"
      }
```

```
    }
```

1.7.2.15 配置PostgreSQL Writer

本文将为您介绍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。

PostgreSQL Writer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。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，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。**

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，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，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

参数	描述	是否必选	默认值
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

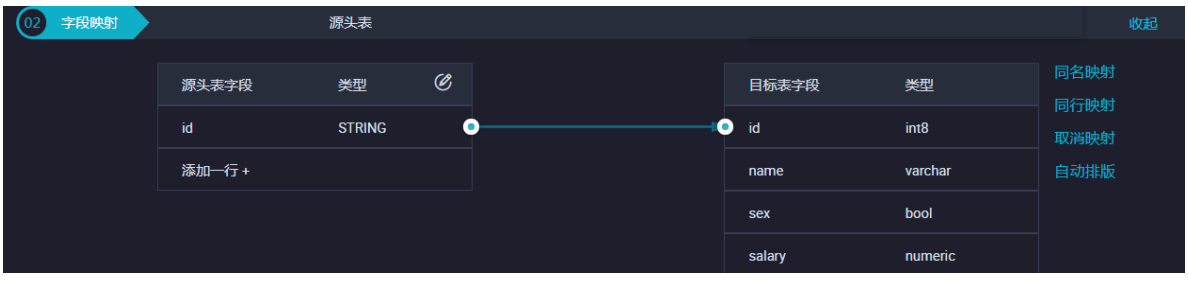
1. 选择数据源。

   配置同步任务的数据来源和数据去向。

配置	说明
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。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，即可选择删除按钮进行删除。



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。

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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",
    "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":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"
            }
        ]
    }
}
```

1.7.2.16 配置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:

- **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
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
keyIndexes	keyIndexes表示源端需要作为key（第1列是从0开始）的列。如果是第1列和第2列需要组合作为key，则keyIndexes的值为[0,1]。  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，则不需要同步所有字段，在Reader插件端指定好column进行列筛选即可。	是	无
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1个key，则可以忽略该配置项。	否	\u0001
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
expireTime	Redis value值缓存失效时间（如果需要永久有效，则可以不填该配置项），单位为秒。 • seconds：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。 • unixtime：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表示永久有效）
timeout	写入Redis的超时时间，单位为毫秒。	否	30,000
dateFormat	写入Redis时，Date的时间格式为yyyy-MM-dd HH:mm:ss。	否	无

参数	描述	是否必选	默认值
writeMode	Redis支持丰富的value类型，包括字符串（string）、字符串列表（list）、字符串集合（set）、有序字符串集合（zset）和哈希（hash）。Redis Writer支持上述5种类型的写入，根据不同的value类型，writeMode配置会略有差异，writeMode的配置说明如下表所示。 说明: 您在配置Redis Writer时，只能配置以下5种类型中的1种。您需要配置其中一种写入数据类型，如果您不填，默认数据类型是string。	否	string

您在配置Redis Writer时，只能配置以下5种类型中的1种。

类型	配置	描述	是否必选
字符串（string） `"writeMode":{ "type": "string", "mode": "set", " valueField Delimiter": "\ u0001" }`	type	value为string类型。	是
	mode	value为string类型时，写入的模式。	是，可选值为set（如果需存储的数据已经存在，则覆盖原有的数据）。
	valueField Delimiter	该配置项主要考虑的是源数据每行超过两列的情况。如果您的源数据只有两列，即key和value时，可以忽略该配置项，无需填写。 value类型为string时，value之间的分隔符。例如value1\u0001value2\u0001value3。	否，默认值为\u0001。 则
字符串列表（list） `"writeMode":{ "type": " list", "mode": " lpush|rpush", "valueField Delimiter": "\ u0001" }`	type	value为list类型。	是
	mode	value为list类型时，写入的模式。	是，可选值为lpush（在list最左边存储数据）和rpush（在list最右边存储数据）。
	valueField Delimiter	value为string类型时，value之间的分隔符。例如value1\u0001value2\u0001value3。	否，默认值为\u0001。
字符串集合（set） `"writeMode":{ "type": " set", "mode": " sadd", " valueField Delimiter": "\ u0001" }`	type	value为set类型。	是
	mode	value为set类型时，写入的模式。	是，可选值：sadd（向set集合中存储数据，如果已经存在则覆盖）。
	valueField Delimiter	value类型是string时，value之间的分隔符，例如value1\u0001value2\u0001value3。	否，默认值为\u0001。
有序字符串集合（zset） `"writeMode":{`	type	value为zset类型。 说明:	是

脚本开发介绍

配置写入Redis的数据同步作业，具体参数填写请参见参数说明。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        {
            "stepType":"stream",
            "parameter":{},
            "name":"Reader",
            "category":"reader"
        },
        {
            "stepType":"redis",//插件名。
            "parameter":{
                "expireTime":{//Redis value值缓存失效时间。
                    "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":"",//数据源。
                "writeMode":{//写入模式。
                    "mode":"",//value是某类型时，写入的模式。
                    "valueFieldDelimiter":"",//value之间的分隔符。
                    "type":""//value类型。
                },
                "keyIndexes":[//主键索引。
                    0,
                    1
                ],
                "batchSize":"1000"//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":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"
            }
        ]
    }
}Writer"
            }
        ]
    }
```

```
}
```

1.7.2.17 配置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。**

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

类型分类	SQL Server数据类型
布尔类	BIT
二进制类	BINARY、VARBINARY、VARBINARY（MAX）和TIMESTAMP

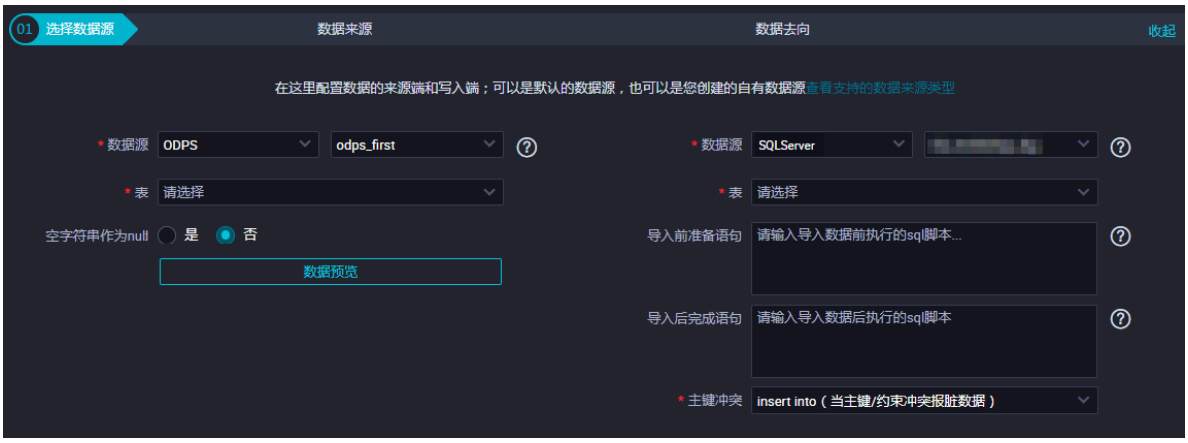
参数说明

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
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。**

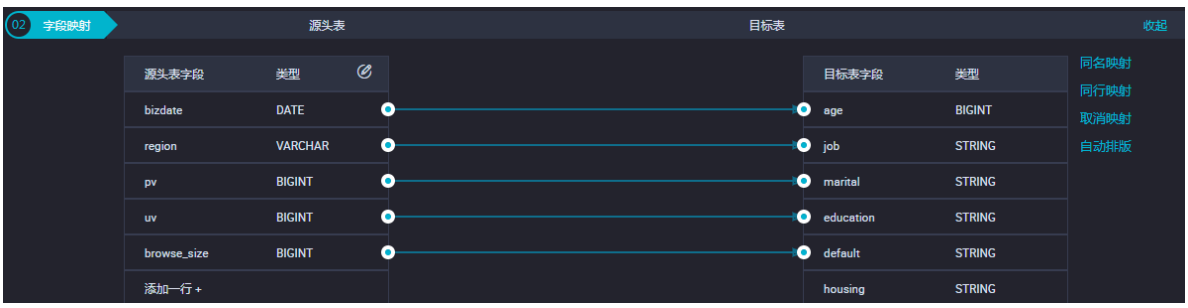
   配置同步任务的数据来源和数据去向。



配置	说明
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。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。



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。

配置	说明
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

配置写入SQL Server的作业，具体参数填写请参见参数说明。

```
{
    "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":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"
            }
        ]
    }
}
```

1.7.2.18 配置Elasticsearch Writer

本文将为您介绍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。

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/数据库，每个索引可以包括多个类型/表，每个类型可以包括多个文档/行，每个文档可以包括多个字段/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://xxxx.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，不需要账号密码，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

参数	描述	是否必选	默认值
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
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）。	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。示例如下： 源头列是字符串类型数据a-,-b-,-c-,-d，使用分隔符（-,-）拆分后是数组["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

参数	描述	是否必选	默认值
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" } · 列类型为日期Date类型时，可以配置Format、Timezone或origin参数，分别表示日期序列化格式和时区，示例如下。 { "name": "col_date", "type": "date", "format": "yyyy-MM-dd HH:mm:ss", "origin": true } 说明:	是	无

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插入的逻辑为：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

脚本开发介绍

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": {
            "concurrent": 1,
```

```
            "throttle": false
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {
                 //下面是关于Reader的模板，您可以查找相应的读插件文档。
            },
            "stepType": "stream"
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {
                "endpoint": "http://xxxx.com:9999",
                "accessId": "xxxx",
                "accessKey": "yyyy",
                "index": "test-1",
                "type": "default",
                "cleanup": true,
                "settings": {
                    "index": {
                        "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:

目前VPC环境的Elasticsearch仅能使用自定义调度资源，运行在默认资源组会存在网络不通的情况。添加自定义资源组的具体操作请参见新增任务资源。

1.7.2.19 配置LogHub Writer

本文将为您介绍LogHub 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 Writer使用SLS的Java SDK，可以将DataX Reader中的数据推送到指定的SLS LogHub上，供其他程序消费。

说明:

由于LogHub无法实现幂等，FailOver重跑任务时会引起数据重复。

LogHub Writer通过Datax框架获取Reader生成的数据，然后将Datax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SLS Java SDK一次性推送至LogHub。默认情况下，一次推送1,024条数据，batchSize值最大为4,096。

类型转换列表

LogHub Writer针对LogHub类型的转换，如下表所示。

DataX内部类型	LogHub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
endpoint	SLS地址。	是	无
accessKeyId	访问SLS的AccessKeyId。	是	无
accessKeySecret	访问SLS的AccessKeySecret。	是	无
project	目标SLS的项目名称。	是	无
logstore	目标SLS LogStore的名称。	是	无
topic	选取topic。	否	空字符串
batchSize	每次批量数据的条数。	否	1,024
column	每条数据中的column名称。	是	无

向导开发介绍

暂不支持向导模式开发。

脚本开发介绍

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": [
        { //下面是关于Reader的模板，您可以查找相应的读插件文档。
            "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SLS LogStore的名字
            },
            "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。
            "dmu": 1//DMU值
        }
    },
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    }
```

```
}
```

1.7.2.20 配置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。

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。数据传输的速率取决于OpenSearch表对应账号的每秒查询率（QPS）。

实现原理

在底层实现上，OpenSearch Writer通过OpenSearch对外提供的开放搜索接口，关于接口的更多详情请参见开放搜索。

说明:

- **V2版本请参见请求结构。**
- **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。**
- **目前默认资源组不支持连接VPC环境，如果是VPC环境可能会存在网络问题。**

插件特点

关于列顺序的问题

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。

- **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。

- 任务重跑和failover

  重跑后会自动根据ID覆盖。

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	阿里云系统登录ID。	是	无
accessKey	阿里云系统登录Key。	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。通常生产的服务地址为：`http://opensearch-cn-internal.aliyuncs.com/`，测试的服务地址为：`http://opensearch-cn-corp.aliyuncs.com/`。	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1MB，单次写入小于2MB。	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。

```
{
    "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
            }
        }
    }
}
```

1.7.2.21 配置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。

简而言之，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，如下表所示。

类型分类	Table Store数据类型
整数类	INTEGER
浮点类	DOUBLE

类型分类	Table Store数据类型
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。

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
endPoint	Table Store 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。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，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

参数	描述	是否必选	默认值
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。 说明: Table Store的PrimaryKey仅支持STRING和INT两种类型，因此Table Store Writer本身也限定填写上述两种类型。 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"} ],`	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: `{"name":"col2", "type":"INT"},` 其中的name指定写入的Table Store列名，type指定写入的类型。Table Store类型支持STRING、INT、DOUBLE、BOOL和BINARY类型。	是	无
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

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。

```
{
    "type":"job",
    "version":"2.0",//版本号
    "steps":[
        { //下面是关于Reader的模板，您可以查看相应的读插件文档。
            "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":"BOOLEAN"
                    },
                    {
                        "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":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
            "dmu":1//DMU值
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

1.7.2.22 配置RDBMS Writer

本文为您介绍RDBMS 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。

RDBMS Writer插件实现了写入数据至RDBMS主库的目的表的功能。在底层实现上，RDBMS Writer通过DataX框架获取Reader生成的协议数据，通过JDBC连接远程RDBMS数据库，并执行相应的`insert into...`的SQL语句，将数据写入RDBMS。RDBMS 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。

RDBMS Writer面向ETL开发工程师，通过RDBMS Writer从数仓导入数据至RDBMS。同时RDBMS 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。

类型转换

目前RDBMS 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RDBMS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的格式是不同的，DataX会根据具体jdbc的格式选择合适的数据库驱动完成数据读取。 · 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

参数	描述	是否必选	默认值
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	1024

功能说明

配置一个写入RDBMS的作业。

```
{
    "job": {
        "setting": {
            "speed": {
                "channel": 1
            }
        },
        "content": [
            {
                "reader": {
                    "name": "streamreader",
                    "parameter": {
                        "column": [
                            {
                                "value": "DataX",
                                "type": "string"
                            },
                            {
                                "value": 19880808,
                                "type": "long"
                            },
                            {
                                "value": "1988-08-08 08:08:08",
                                "type": "date"
                            },
                            {
                                "value": true,
                                "type": "bool"
                            },
                            {
                                "value": "test",
                                "type": "bytes"
                            }
                        ],
                        "sliceRecordCount": 1000
                    }
                },
                "writer": {
                    "name": "RDBMS Writer",
                    "parameter": {
                        "connection": [
                            {
                                "jdbcUrl": "jdbc:dm://ip:port/database
",
                                "table": [
                                    "table"
                                ]
                            }
                        ],
```

```
                        "username": "username",
                        "password": "password",
                        "table": "table",
                        "column": [
                            "*"
                        ],
                        "preSql": [
                            "delete from XXX;"
                        ]
                    }
                }
            }
        ]
    }
}
```

RDBMS Writer增加新的数据库支持的操作如下。

1. **进入RDBMS Writer对应目录，这里${DATAX_HOME}为DataX主目录，即***${DATAX_HOME}/plugin/writer/RDBMS Writer*。
2. **在RDBMS 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。RDBMS 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, 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"
    ]
}
```

3. **在RDBMS 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
`-- RDBMS Writer-0.0.1-SNAPSHOT.jar
```

1.7.2.23 配置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举例。

Stream Writer 插件实现了从 Reader 端读取数据并向屏幕上打印数据或者直接丢弃数据，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

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":[
        { //下面是关于Reader的模板，可以找相应的读插件文档
            "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":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
            "dmu":1//DMU值
        }
    },
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    }
}
```

1.7.2.24 配置HybridDB for MySQL Writer

本文将为您介绍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	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 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。 • 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。	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

参数	描述	必选	默认值
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。

配置	说明
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。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。

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。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

配置	说明
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。

配置	说明
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

96811

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": [
        {
            "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,//false代表不限流，下面的限流的速度不生效，true代表限流。
            "mbps": 1,//限流值。
        }
    }
```

```
}
```

1.7.2.25 配置AnalyticDB for PostgreSQL Writer

本文将为您介绍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，从AnalyticDB for PostgreSQL库中选取数据。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。

简而言之，AnalyticDB for PostgreSQL Writer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SELECT SQL语句，发送至远程AnalyticDB for PostgreSQL数据库。然后使用CDP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等信息，AnalyticDB for PostgreSQL Writer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 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。

PAnalyticDB 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

类型分类	AnalyticDB for PostgreSQL数据类型
BYTES	BYTEA

说明:

- 除上述罗列字段类型外，其他类型均不支持。
- 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

参数	描述	是否必选	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	1024

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。

配置	说明
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。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，即可选择删除按钮进行删除。

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。
自动排版	可以根据相应的规律自动排版。

3. 通道控制。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和**新增任务资源**。

脚本开发介绍

```
{
    "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": {
            "concurrent": 6,//并发数。
            "throttle": false,//同步速率是否限流。
        }
    }
}
```

1.7.2.26 配置POLARDB Writer

本文将为您介绍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举例。

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，他们使用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，是否需要其他权限，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、T ext、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

参数	描述	必选	默认值
writeMode	选择导入模式，可以支持insert/replace方式。 · 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。 · 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。 · INSERT INTO table (a,b,c) VALUES (1,2,3) 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	1024

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。

配置	说明
数据源	即上述参数说明中的datasource，一般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。
主键冲突	即上述参数说明中的writeMode，可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，单击添加一行可增加单个字段，鼠标放至需要删除的字段上，即可单击删除图标进行删除。

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。
自动排版	可以根据相应的规律自动排版。

3. 通道控制

配置	说明
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。

配置	说明
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

脚本开发介绍

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
                "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": {
            "concurrent": 6,//并发数
            "throttle": false,//同步速率限流
        }
    }
```

```
}
```

1.7.2.27 配置TSDB Writer

TSDB Writer插件实现了将数据点写入阿里巴巴自主研究的TSDB数据库。

时间序列数据库（Time Series Database , 简称TSDB）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提供百万级时序数据秒级写入、高压缩比低成本存储、预降采样、插值和多维聚合计算、查询结果可视化功能。TSDB可以解决由于设备采集点数量大、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等问题。

目前仅支持脚本模式配置方式，更多详情请参见时序时空数据库文档。

实现原理

TSDB Writer通过HTTP连接TSDB实例，并通过`/api/put`接口将数据点写入。

约束限制

目前仅支持兼容TSDB 2.4.x及以上版本。

支持的数据类型

类型分类	数据集成column配置类型	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

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、InfluxDB、Prom。RDB包括MySQL、Oracle、PostgreSQL**

数据源	参数	描述	是否必选	默认值
数据源为TSDB	endpoint	TSDB的HTTP连接地址。	是，格式为http://IP:Port。	无
	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
	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	endpoint	TSDB的HTTP连接地址。	是，格式为http://IP:Port。	无
	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。 • metric_num：该Metric的value是数值类型。 • metric_string：该Metric的value是字符串类型。	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

配置一个同步数据至TSDB的作业。

```json
{
    "order": {
        "hops": [
            {
                "from": "Reader",
                "to": "Writer"
            }
        ]
    },
    "setting": {
        "errorLimit": {
            "record": "0"
        },
        "speed": {
            "concurrent": 1,
            "throttle": true
        }
    },
    "steps": [
        {
            "category": "reader",
            "name": "Reader",
            "parameter": {},
            "stepType": ""
        },
        {
            "category": "writer",
            "name": "Writer",
            "parameter": {
                "endpoint": "http://localhost:8242",
                "sourceDbType": "RDB",
                "batchSize": 256,
                "column": [
                    "name",
                    "type",
                    "create_time",
                    "price"
                ],
                "columnType": [
                    "tag",
                    "tag",
                    "timestamp",
                    "metric_num"
                ]
            },
```

```
            "stepType": "tsdb"
        }
    ],
    "type": "job",
    "version": "2.0"
}
```

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

1.7.2.28 配置AnalyticDB for MySQL 3.0 Writer

本文将为您介绍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	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 into和replace into两种方式。 • 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。 • 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。	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。 说明: 当有多条SQL语句时，不支持事务。	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。

配置	说明
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。
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。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。

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。
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，如'abc'、'123'等。 · 可以配合调度参数使用，如${bizdate}等。 · 可以输入关系数据库支持的函数，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。

配置	说明
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。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

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": "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": {
            "concurrent": 2, //任务并发数。
            "throttle": false //false表示不限流，所设置的限流速度不生效，
true表示限流。
        }
    }
}
```

1.7.2.29 配置Hive Writer

Hive Writer插件实现了从Hive写出数据至HDFS的功能，本文将为您介绍Hive Writer的工作原理、参数和示例。

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，并提供类SQL查询功能。

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。**
- **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文件的方式从Hive写出数据至HDFS。

Hive Writer底层的逻辑和HDFS Writer插件一致，完成数据的写出后，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	必选	默认值
jdbcUrl	Hive元数据库的地址。目前Hive Reader仅支持访问MySQL类型的Hive元数据库。 您需要确保任务执行节点具备Hive元数据库的网络和访问权限。	是	无
username	Hive元数据库的用户名。	是	无

参数	描述	必选	默认值
password	Hive元数据库的密码。	是	无
column	需要写出的字段列，例如"column": ["id", "name"]。 · 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。 · column必须显示指定同步的列集合，不允许为空。	是	无
table	需要写出的Hive表名。 说明: 请注意大小写。	是	无
partition	· 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

脚本开发介绍

配置一个从Hive写出数据的JSON示例。

```
{
        "order": {
            "hops": [
                {
                    "from": "Reader",
                    "to": "Writer"
```

```
                }
            ]
        },
        "setting": {
            "errorLimit": {
                "record": "0"
            },
            "speed": {
                "concurrent": 1,
                "throttle": false
            }
        },
        "steps": [
            {
                "category": "reader",
                "name": "Reader",
                "parameter": {},
                "stepType": "stream"
            },
            {
                "category": "writer",
                "name": "Writer",
                "parameter": {
                    "username": "",
                    "password": "",
                    "jdbcUrl": "jdbc:mysql://host:port/database",
                    "table": "",
                    "partition": "",
                    "column": [
                        "id",
                        "name"
                    ],
                    "writeMode": "append",
                    "hiveConfig": {
                        "hiveCommand": "",
                        "jdbcUrl": "",
                        "username": "",
                        "password": ""
                    }
                },
                "stepType": "hive"
            }
        ],
        "type": "job",
        "version": "2.0"
    }
```

1.7.2.30 配置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。

说明:

- **开始配置GDB Writer插件前，请首先配置好数据源，详情请参见配置*GDB*数据源。**

· 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。

约束限制

· 必须先运行点的同步任务，运行成功后，方可运行边的同步任务。
· 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*#unique_85*，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

参数	描述	是否必选	默认值
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
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。	是	无

脚本开发介绍

配置写入GDB的数据同步作业，详情请参见上述参数说明。

- **点配置示例**

```
{
    "order":{
        "hops":[
            {
                "from":"Reader",
                "to":"Writer"
            }
        ]
    },
    "setting":{
        "errorLimit":{
            "record":"100"   //错误记录数，表示脏数据最大容忍条数。
        },
        "jvmOption":"",
        "speed":{
            "concurrent":3,
            "throttle":false
        }
    },
    "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":{
            "concurrent":3,
            "throttle":false
        }
    },
    "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"
}
```

1.7.2.31 配置Kafka Writer

Kafka Writer通过Kafka服务的Java SDK向Kafka写入数据，本文将为您介绍Kafka Writer的实现原理、参数和示例。

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>
```

参数说明

参数	说明	是否必填
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
KeyIndex	Kafka Writer中作为Key的那一列。	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
fieldDelimiter	源头的多列数据通过列分隔符拼接后，写出至Kafka。默认值为\t	否

参数	说明	是否必填
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
batchSize	向kafka一次性写入的数据量，默认为1,024条。	否

向导模式开发

暂不支持向导模式开发。

脚本模式开发

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。
                    "valueIndex": 1, //作为Value的某列。
                    "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":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"
            }
        ]
    }
}
```

1.7.2.32 配置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。

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。

实现原理

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。

- **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	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，字段之间用英文所逗号分隔，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

向Vertica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":"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后率先执行的SQL语句。
                    "delete from @table where db_id = -1"
                ],
                "postSql": [//执行数据同步任务之前率先执行的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":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"
            }
        ]
    }
```

```
}
```

1.7.3 优化配置

本文将为您介绍数据同步速度的影响因素，如何通过调整同步作业的并发配置来达到最大化同步速度，作业限速和不限速的区别，以及自定义资源组的注意事项。

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、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复制10TB级的数据。

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所示：

- **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，可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，建议设置得较小。**
  - **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，以下将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30MB/s）。脚本模式通过如下示例代码配置限速，代表1MB/s的传输带宽。

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

说明:

- 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nMB/s，如果设置n+1MB/s还是以nMB/s速度同步，如果设置为n-1MB/s，则以n-1MB/s速度同步。
- 关系型数据库设置作业速率上限和切分键才能根据作业速率上限将表进行切分，关系型数据库只支持数值型作为切分键，但Oracle数据库支持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。

    例如一个RDS到MaxCompute的数据同步任务执行完成，共等待了约800s，但是日志显示任务只运行了18s。使用的是默认资源组。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。很多项目都在使用，不只是单个用户的2-3个任务。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到一张表里，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，同时往一个数据库进行写入，需注意以下问题：

    - **确保目标数据库负载能力是能够承受，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，这样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>='2016-10-23 00:00:00'and reateTime<'2016-10-24 00:00:
    00';
    ```

    假设从2016-10-25 11:01:24开始执行上述语句，到2016-10-25 11:11:05才开始返回结果，同步任务执行时间较长。

  - 分析原因

    检查where语句，发现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。

1.8 常见配置

1.8.1 添加安全组

本文将为您介绍选择不同区域的DataWorks时，如何添加需要的安全组。

通常情况下，如果您使用的是ECS自建数据库，则必须添加安全组才能保证数据源连通性正常。

为保证数据库的安全稳定，在开始使用某些数据库的实例前，您需要将访问数据库的IP地址或IP段加到目标实例的白名单或安全组中。添加白名单的详情请参见添加白名单。

添加安全组

- 如果您的ECS上的自建数据源同步任务运行在自定资源组上，需要给自定资源组机器授权，将自定义机器内/外网IP和端口添加到ECS安全组上。
- 如果您的ECS上的自建数据源运行在默认的资源组上，需要给默认的机器授权。根据您的ECS的机器区域来选择添加您的安全组内容，例如您的ECS是华北2，安全组便添加华北2（北京）：sg-2ze3236e8pcbxw61o9y0和1156529087455811内容，并且只能在华北2添加数据源，如下表所示。

区域	授权对象	账号ID
华东1（杭州）	sg-bp13y8iuj33uqpqvgqw2	1156529087455811
华东2（上海）	sg-uf6ir5g3rlu7thymywza	1156529087455811

区域	授权对象	账号ID
华南1（深圳）	sg-wz9ar9o9jgok5tajj7ll	1156529087455811
亚太东南1（新加坡）	sg-t4n222njci99ik5y6dag	1156529087455811
中国（香港）	sg-j6c28uqpqb27yc3tjmb6	1156529087455811
美国西部 1（硅谷）	sg-rj9bowpmdvhyl53lza2j	1156529087455811
美国东部1	sg-0xienf2ak8gs0puz68i9	1156529087455811
华北2（北京）	sg-2ze3236e8pcbxw61o9y0	1156529087455811

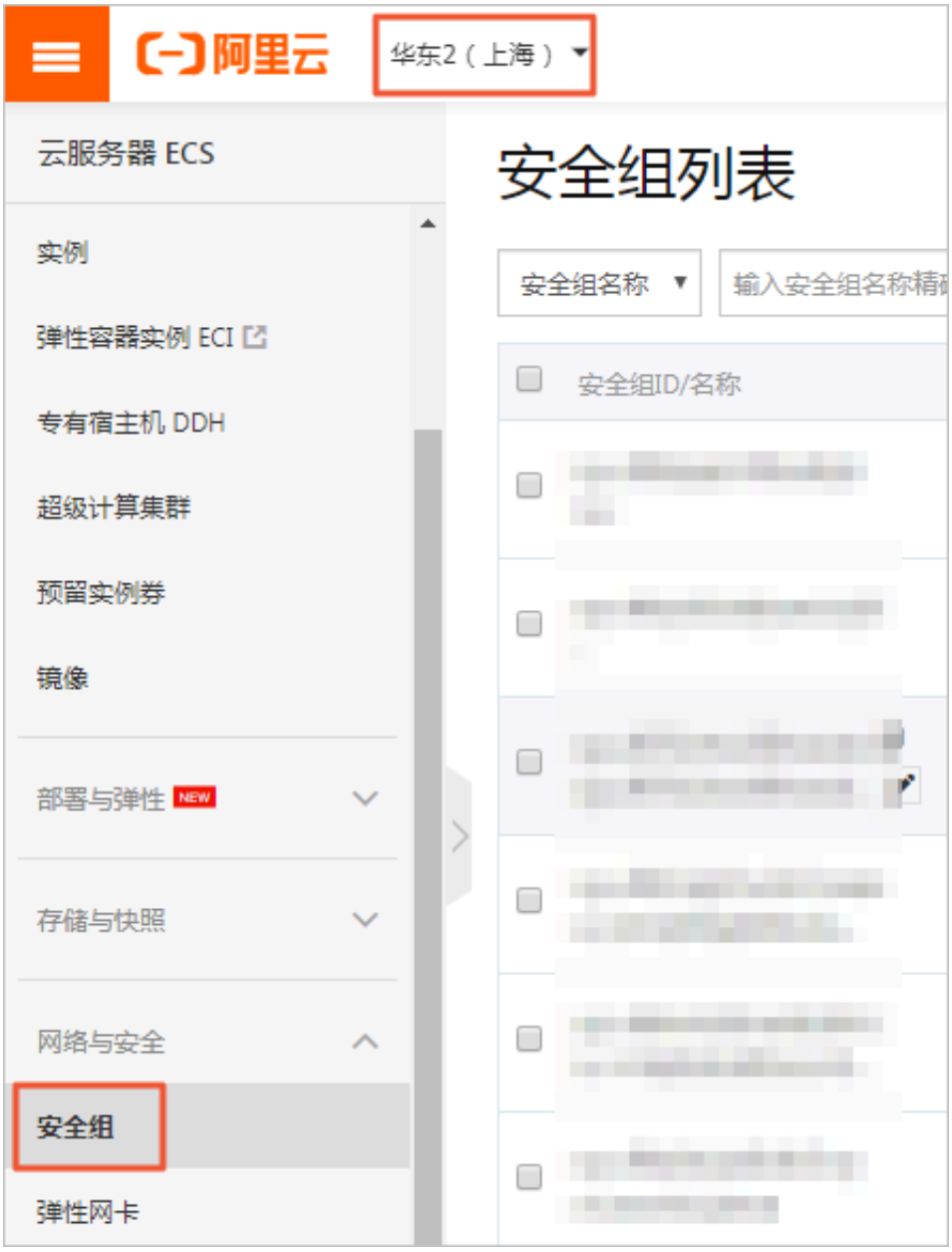
说明:

VPC环境的ECS不支持添加上面的安全组，因为上面的都是经典网络类型的IP，会存在网络类型不同的问题。

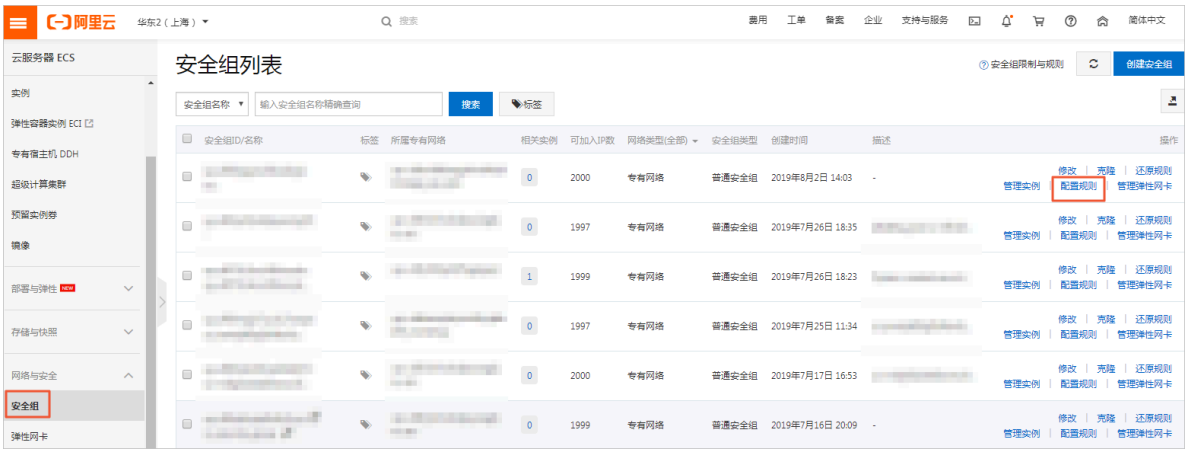
ECS添加安全组

1. **登录云服务器ECS的管理控制台。**

2. 进入网络和安全 > 安全组页面，选择目标区域。



3. 找到要配置授权规则的安全组，单击操作列下的配置规则。



4. 进入安全组规则 > 入方向页面，单击右上角的添加安全组规则。

5. 填写添加安全组规则对话框中的配置。

6. 单击确定。

1.8.2 添加白名单

本文将为您介绍选择不同区域的DataWorks时，如何添加需要的不同白名单的内容。

通常情况下，如果您使用的是RDS数据源，则必须添加白名单才能保证数据连通性正常。

为保证数据库的安全稳定，在开始使用某些数据库实例前，您需要将访问数据库的IP地址或者IP段加到目标实例的白名单中。

> **说明:**
> **添加白名单功能仅对数据集成生效，其它类型任务不支持白名单功能。**

添加白名单

1. **以开发者身份登录*DataWorks*控制台，单击工作空间列表，进入工作空间列表页面。**
2. **选择工作空间区域。**

   目前DataWorks支持多个区域，此处要选择的工作空间区域和您购买不同区域的MaxCompute有关，例如华东2（上海）、华南1（深圳）、中国（香港）等，都代表您开通这些区域的MaxCompute，您可以手动切换不同的区域入口。

3. 根据工作空间所在的区域选择相应的白名单。

   目前一部分数据源有白名单的限制，需要对数据集成的访问IP进行放行，例如RDS、MongoDB和Redis等常见的数据源，需要在相应的控制台对这边的IP进行开放。通常添加白名单有以下两种情况：

   - 同步任务运行在自定资源组上，需要给自定资源组机器授权，将自定义机器内/外网IP添加数据源的白名单列表。
   - 同步任务运行在默认资源组上，需要给底层运行机器授予访问权限，根据您选择DataWorks的区域来填写您需要添加的白名单，内容如下表所示。

区域	白名单
华东1（杭州）	100.64.0.0/10,11.193.102.0/24,11.193.215.0/24,11.194.110.0/24,11.194.73.0/24,118.31.157.0/24,47.97.53.0/24,11.196.23.0/24,47.99.12.0/24,47.99.13.0/24,114.55.197.0/24,11.197.246.0/24,11.197.247.0/24
华东2（上海）	11.193.109.0/24,11.193.252.0/24,47.101.107.0/24,47.100.129.0/24,106.15.14.0/24,10.117.28.203,10.143.32.0/24,10.152.69.0/24,10.153.136.0/24,10.27.63.15,10.27.63.38,10.27.63.41,10.27.63.60,10.46.64.81,10.46.67.156,11.192.97.0/24,11.192.98.0/24,11.193.102.0/24,11.218.89.0/24,11.218.96.0/24,11.219.217.0/24,11.219.218.0/24,11.219.219.0/24,11.219.233.0/24,11.219.234.0/24,118.178.142.154,118.178.56.228,118.178.59.233,118.178.84.74,120.27.160.26,120.27.160.81,121.43.110.160,121.43.112.137,100.64.0.0/10
华南1（深圳）	100.106.46.0/24,100.106.49.0/24,10.152.27.0/24,10.152.28.0/24,11.192.91.0/24,11.192.96.0/24,11.193.103.0/24,100.64.0.0/10,120.76.104.0/24,120.76.91.0/24,120.78.45.0/24
西南1（成都）	11.195.52.0/24,11.195.55.0/24,47.108.22.0/24,100.64.0.0/10
华北3（张家口）	11.193.235.0/24,47.92.22.0/24,100.64.0.0/10
中国（香港）	10.152.162.0/24,11.192.196.0/24,11.193.11.0/24,100.64.0.0/10,11.192.196.0/24,47.89.61.0/24,47.91.171.0/24,11.193.118.0/24,47.75.228.0/24

区域	白名单
亚太东南1（新加坡）	100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.153.0/24,11.192.40.0/24,11.193.8.0/24,100.64.0.0/10,100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.40.0/24,47.88.147.0/24,47.88.235.0/24,11.193.162.0/24,11.193.163.0/24,11.193.220.0/24,11.193.158.0/24,47.74.162.0/24,47.74.203.0/24,47.74.161.0/24,11.197.188.0/24
亚太东南2（澳洲、悉尼）	11.192.100.0/24,11.192.134.0/24,11.192.135.0/24,11.192.184.0/24,11.192.99.0/24,100.64.0.0/10,47.91.49.0/24,47.91.50.0/24,11.193.165.0/24,47.91.60.0/24
华北2（北京）	100.106.48.0/24,10.152.167.0/24,10.152.168.0/24,11.193.50.0/24,11.193.75.0/24,11.193.82.0/24,11.193.99.0/24,100.64.0.0/10,47.93.110.0/24,47.94.185.0/24,47.95.63.0/24,11.197.231.0/24,11.195.172.0/24,47.94.49.0/24,182.92.144.0/24
美国西部1	10.152.160.0/24,100.64.0.0/10,47.89.224.0/24,11.193.216.0/24,47.88.108.0/24
美国东部1	11.193.203.0/24,11.194.68.0/24,11.194.69.0/24,100.64.0.0/10,47.252.55.0/24,47.252.88.0/24
亚太东南3（马来西亚、吉隆坡）	11.193.188.0/24,11.221.205.0/24,11.221.206.0/24,11.221.207.0/24,100.64.0.0/10,11.214.81.0/24,47.254.212.0/24,11.193.189.0/24
欧洲中部1（德国、法兰克福）	11.192.116.0/24,11.192.168.0/24,11.192.169.0/24,11.192.170.0/24,11.193.106.0/24,100.64.0.0/10,11.192.116.14,11.192.116.142,11.192.116.160,11.192.116.75,11.192.170.27,47.91.82.22,47.91.83.74,47.91.83.93,47.91.84.11,47.91.84.110,47.91.84.82,11.193.167.0/24,47.254.138.0/24
亚太东北1（日本）	100.105.55.0/24,11.192.147.0/24,11.192.148.0/24,11.192.149.0/24,100.64.0.0/10,47.91.12.0/24,47.91.13.0/24,47.91.9.0/24,11.199.250.0/24,47.91.27.0/24
中东东部1（阿联酋、迪拜）	11.192.107.0/24,11.192.127.0/24,11.192.88.0/24,11.193.246.0/24,47.91.116.0/24,100.64.0.0/10
亚太东南1（印度、孟买）	11.194.10.0/24,11.246.70.0/24,11.246.71.0/24,11.246.73.0/24,11.246.74.0/24,100.64.0.0/10,149.129.164.0/24,11.194.11.0/24
英国	11.199.93.0/24,100.64.0.0/10

区域	白名单
亚太东南5（印度尼西亚、雅加达）	11.194.49.0/24,11.200.93.0/24,11.200.95.0/24,11.200.97.0/24,100.64.0.0/10,149.129.228.0/24,10.143.32.0/24,11.194.50.0/24
华北2（政务云）	11.194.116.0/24,100.64.0.0/10 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

RDS添加白名单

RDS数据源可以通过以下两种方式进行配置：

- RDS实例形式

  通过RDS实例创建数据源，目前支持测试连通性（其中包括VPC环境的RDS）。如果RDS实例形式测试连通性失败，可以尝试用JDBCUrl形式添加数据源。

- JDBCUrl形式

  JDBCUrl中的IP请优先填写内网地址，如果没有内网地址请填写外网地址。其中，内网地址是走阿里云机房，内网地址同步时同步速度会更快，外网地址同步时的同步速度受限于您开通的外网带宽。

配置RDS白名单

数据集成连接RDS同步数据需要使数据库标准协议连接数据库。RDS默认允许所有IP连接，但如果您在RDS配置指定了IP白名单，您需要添加数据集成执行节点IP白名单。如果您没有指定RDS白名单，不需要给数据集成提供白名单。

如果您设置了RDS的IP白名单，请进入RDS控制台，并导航至安全控制，根据上面的白名单列表进行设置，详情请参见白名单设置。

说明：

如果使用自定义资源组调度RDS的数据同步任务，必须把自定义资源组的机器IP添加至RDS的白名单中。

1.8.3 新增任务资源

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。您可以新增自定义的任务资源运行数据同步任务，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。

项目管理员可以在数据集成 > 同步资源管理 > 资源组页面新增或修改任务资源。

当默认任务资源无法与您的复杂的网络环境连通时，可以通过数据集成自定义资源的部署，打通任意网络环境之间的数据传输同步，详情请参见（一端不通）数据源网络不通的情况下的数据同步和（两端不通）数据源网络不通的情况下的数据同步。

说明:

- 您在数据集成 > 同步资源管理 > 资源组页面增加的任务资源，只能给当前工作空间作为数据同步资源组使用，不会显示在资源组列表。目前该页面添加的任务资源不支持手动业务流程数据同步节点。
- 添加自定义资源时一台机器只能添加一个自定义资源组，每个自定义资源组只能选择一种网络类型。
- 注册服务器时，只有华东2可以选择经典网络的方式注册（输入主机名），建议您优先使用专有网络VPC。其他Region只能选择专有网络方式注册（输入UUID）。
- 自定义资源组上运行的部分文件需要Admin权限。例如，在您自己写的Shell脚本任务中调用自定义ECS上的Shell文件、SQL文件等。
- 因为调度资源组主要用于调度任务，资源有限，并不适合用来完成计算任务，所以不推荐在调度资源组上安装数据处理模块。MaxCompute具有海量数据处理能力，推荐您通过MaxCompute进行大数据计算。

购买云服务器ECS

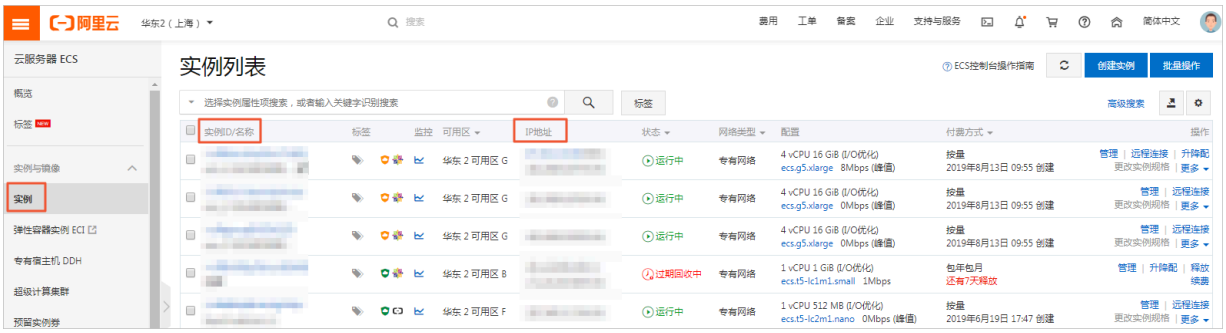
购买ECS云服务器的具体操作请参见购买ECS云服务器。

说明:

- 使用CentOS 6、CentOS 7或Aliyun OS。
- 如果您添加的ECS需要执行MaxCompute任务或同步任务，需要检查当前ECS的Python版本是否是Python2.6或2.7（CentOS 5的版本为Python 2.4，其它OS自带Python 2.6以上版本）。
- 请确保ECS有访问公网能力，您可将是否ping通`www.aliyun.com`作为衡量标准。
- 建议ECS的配置为8核16G。

查看ECS主机名和内网IP地址

您可以进入云服务器ECS > 实例页面，查看购买的ECS主机名和IP。



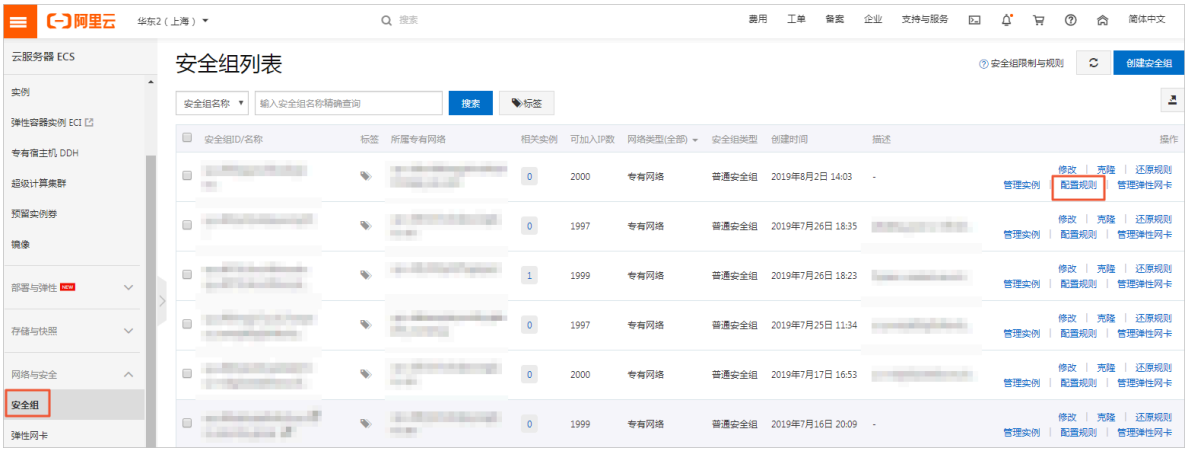
开通8000端口，以便读取日志

说明:

如果您的ECS是VPC专有网络类型，则无需开通8000端口。下述步骤仅适用于经典网络。

1. 添加安全组规则

进入云服务器ECS > 网络和安全 > 安全组页面，单击配置规则，进入配置规则页面 。



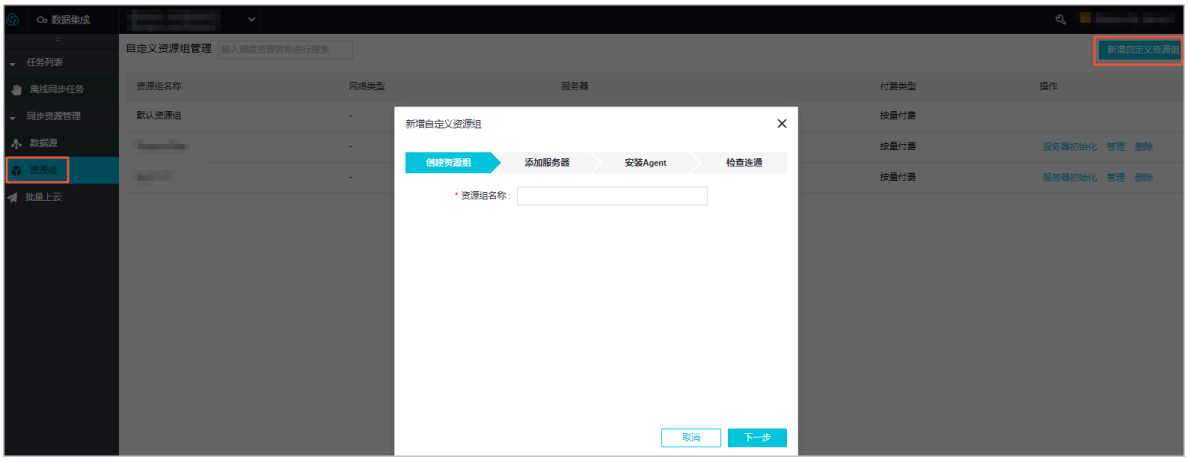
2. 进入安全组规则 > 入方向页面，单击右上角的添加安全组规则。

3. 填写添加安全组规则对话框中的配置信息，配置IP为数据集成的固定IP，访问端口为8000。

新增任务资源

1. 以开发者身份进入DataWorks管理控制台，单击对应工作空间后的进入数据集成。

2. 进入同步资源管理 > 资源组页面，单击右上角的新增自定义资源。



3. 在新增自定义资源组对话框中，填写资源组名称，单击下一步。

4. 在添加服务器对话框中，填写购买的ECS云服务器的主机IP等信息。

配置	说明
网络类型	目前除上海区域支持经典网络外，其他区域仅支持专有网络。

配置	说明
服务器名称/ECS UUID	· 选择经典网络时，需要填写服务器名称。登录ECS，执行`hostname`取返回值。 · 选择专有网络时，需要填写ECS UUID。登录ECS，执行`dmidecode | grep UUID`，取返回值。
机器IP	请输入内网机器IP。
机器CPU（核）	推荐的自定义资源组机器CPU配置至少为4核。
机器内存（GB）	推荐的自定义资源组机器内存配置至少为8GB RAM和80GB磁盘。

说明:

填写专有网络下的ECS作为服务器时，需要填写ECS的UUID作为服务器名称。登录到ECS机器执行`dmidecode | grep UUID`即可获取。

例如执行`dmidecode | grep UUID`，返回结果是UUID: 713F4718-8446-4433-A8EC-6B5B62D7****，则对应的UUID为713F4718-8446-4433-A8EC-6B5B62D7****。

5. 安装Agent并初始化。

如果是新添加的服务器，请按照如下步骤进行操作。

a. SSH登录ECS服务器，保持在root用户下。

b. 执行下述命令：

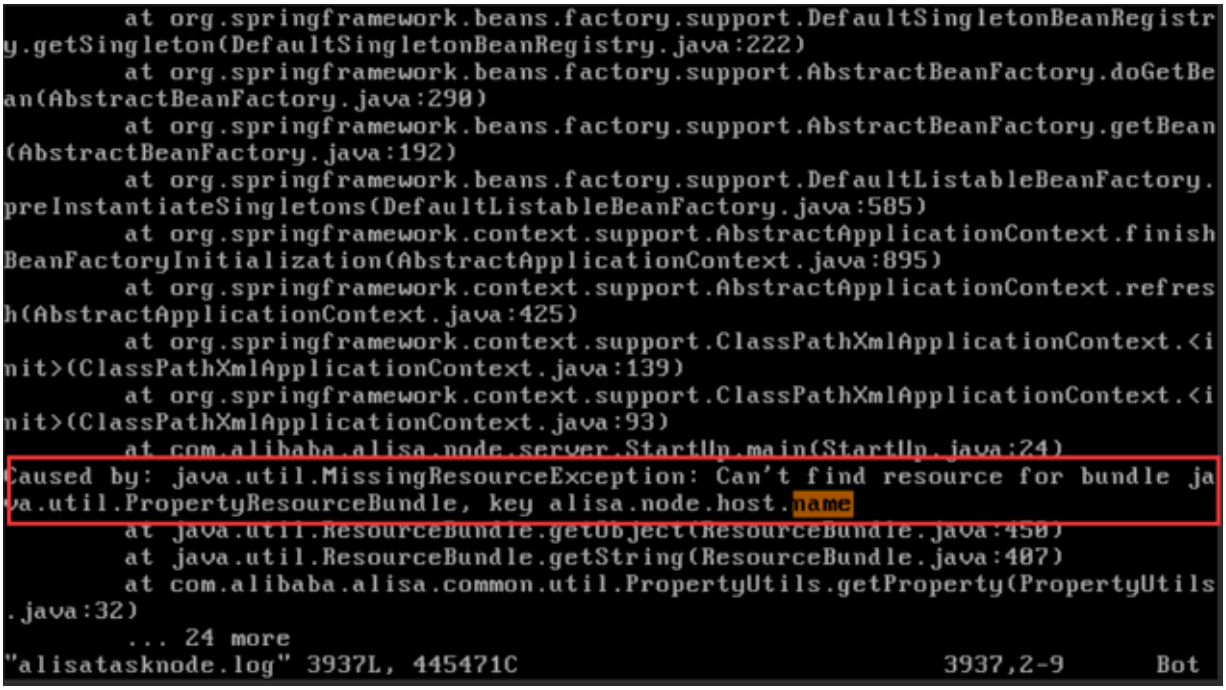
```
chown admin:admin /opt/taobao  //用于给admin用户授予/opt/taobao目录权限。
wget https://alisaproxy.shuju.aliyun.com/install.sh --no-check-certificate
sh install.sh --user_name=*****19d --password=****h1bm --enable_uuid=false
```

c. 稍后在添加服务器页面，单击刷新，查看服务状态是否转为可用。

d. 开通服务器的8000端口。

> **说明：**
> 如果执行install.sh过程中出错或需要重新执行，请在install.sh的同一个目录下执行`rm -rf install.sh`，删除已经生成的文件。然后执行`install.sh`。上面的初始化界面对于每个用户的命令都不一样，请根据自己的初始化界面执行相关命令。

执行完上述操作后，如果服务状态一直是停止，您可能碰到以下问题。



上图的错误原因是没有绑定host，请参见以下步骤进行修改。

1. **切换到admin账号。**
2. **执行`hostname -i`，查看host的绑定情况。**
3. **执行`vim/etc/hosts`，添加IP地址和主机名。**
4. **刷新页面服务状态，查看ECS服务器注册是否成功。**

> **说明:**
>
> - **如果刷新后还是停止状态，您可以重启alisa。**
>
> **切换到admin账号，执行下述命令。**
>
> ```
> /home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart
> ```
>
> - **命令中涉及到您的AK信息，请不要轻易泄露。**

数据同步选择任务资源组

在数据同步任务中的通道控制选择任务资源组。

使用限制

- 自定义任务资源所在的ECS服务器的时间与当前互联网时间差必须在2分钟之内，否则会导致部署的自定义任务资源服务请求接口超时服务异常，无法执行任务。
- 如果您发现alisatasknode日志中有超时报错信息`response code is not 200`，通常是因为某个时段访问服务接口不稳定的异常导致。只要不是持续10分钟异常，自定义资源组服务器就依然可以正常服务。您可以查阅日志`/home/admin/alisatasknode/logs/heartbeat.log`进行确认。

1.9 整库迁移

1.9.1 整库迁移概述

本文将为您介绍整库迁移的任务生成规则和约束限制。

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，它可以快速把一个MySQL数据库内所有表一并上传到MaxCompute的工作，节省大量初始化数据上云的批量任务创建时间。

假设数据库内有100张表，您原本可能需要配置100次数据同步任务，但有了整库迁移便可以一次性完成。同时，由于数据库的表设计规范性的问题，此工具并无法保证一定可以一次性完成所有表按照业务需求进行同步的工作，即它有一定的约束性。

任务生成规则

完成配置后，根据选择的需要同步的表，依次创建MaxCompute表，生成数据同步任务。

MaxCompute表的表名、字段名和字段类型根据高级配置生成，如果没有填写高级配置，则与MySQL表的结构完全相同。表的分区为pt，格式为yyyymmdd。

生成的数据同步任务是按天调度的周期任务，会在第二天凌晨自动运行，传输速率为1M/s，它在细节上会因为同步的方式、并发配置等有所不同，您可以在同步任务目录树的clone_database > 数据源名称 > mysql2odps_表名中找到生成的任务，然后对其进行更加个性化的编辑操作。

说明:

建议您当天对数据同步任务进行冒烟测试，相关任务节点可以在运维中心 > 任务管理中的project_etl_start > 整库迁移 > 数据源名称 下找到所有此数据源生成的同步任务，然后右键单击，测试相应的节点即可。

约束限制

由于数据库的表设计规范性的问题，整库迁移具有一定的约束性。

- 目前仅提供MySQL和Oracle数据源的整库迁移至MaxCompute，后续Hadoop、Hive数据源功能会逐渐开放。
- 仅提供每日增量、每日全量的上传方式。

  如果您需要一次性同步历史数据，则此功能无法满足您的需求，建议如下：

  - 建议您配置为每日任务，而非一次性同步历史数据。您可以通过调度提供的补数据，来对历史数据进行追溯，这样可避免全量同步历史数据后，还需要做临时的SQL任务来拆分数据。
  - 如果您需要一次性同步历史数据，可以在任务开发页面进行任务的配置，然后单击运行。完成后通过SQL语句进行数据的转换，因为这两个操作均为一次性行为。

  如果您每日增量上传有特殊业务逻辑，而非一个单纯的日期字段可以标识增量，则此功能无法满足您的需求，建议如下：

  - 数据库数据的增量上传有两种方式：通过binlog（DTS产品可提供）和数据库提供数据变更的日期字段来实现。

    目前数据集成支持的为后者，所以要求您的数据库有数据变更的日期字段，通过日期字段，系统会识别您的数据是否为业务日期当天变更，即可同步所有的变更数据。
  - 为了更方便地增量上传，建议您在创建所有数据库表的时候都有gmt_create和gmt_modify字段，同时为了效率更高，建议增加id为主键。
- 整库迁移提供分批和整批迁移的方式

  分批上传为时间间隔，目前不提供数据源的连接池保护功能，此功能正在规划中。

  - 为了保障对数据库的压力负载，整库迁移提供了分批迁移的方式，您可以按照时间间隔把表拆分为几批运行，避免对数据库的负载过大，影响正常的业务能力。建议如下：
    - 如果您有主、备库，建议同步任务全部同步备库数据。
    - 批量任务中每张表都会有1个数据库连接，上限速度为1M/s。如果您同时运行100张表的同步任务，就会有100个数据库进行连接，建议您根据自己的业务情况谨慎选择并发数。
  - 如果您对任务传输效率有自己特定的要求，此功能无法实现您的需求。所有生成任务的上限速度均为1M/s。

- 仅提供整体的表名、字段名和字段类型映射

  整库迁移会自动创建MaxCompute表，分区字段为pt，类型为字符串String，格式为yyyymmdd。

  > **说明:**
  > 选择表时必须同步所有字段，它不能对字段进行编辑。

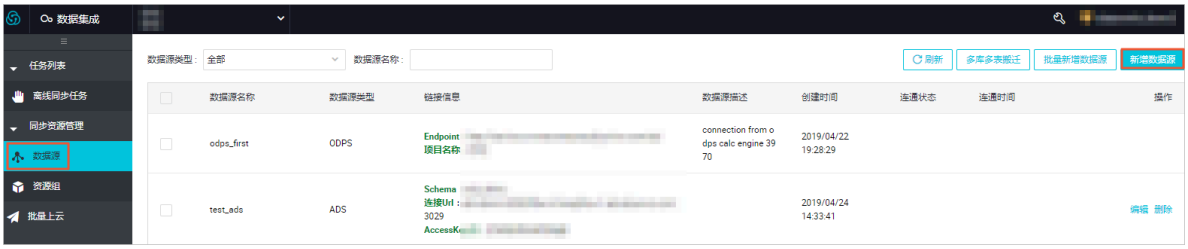
1.9.2 配置MySQL整库迁移

本文将为您介绍如何通过整库迁移功能，将MySQL数据整库迁移至MaxCompute。

整库迁移是为了提升用户效率、降低用户使用成本的一种快捷工具，它可以快速把MySQL数据库内所有表一并上传至MaxCompute。整库迁移的详细介绍请参见整库迁移概述。

操作步骤

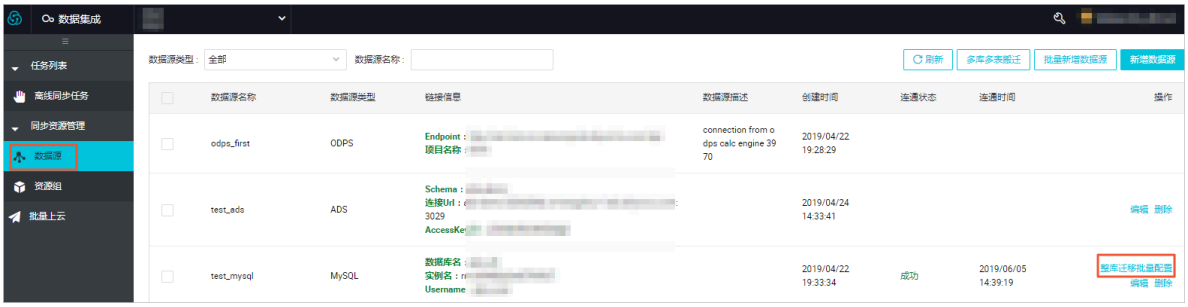
1. 登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择左侧导航栏中的同步资源管理 > 数据源，进入数据源管理页面。



3. 单击右上角的新增数据源，添加一个面向整库迁移的MySQL数据源（clone_database）。

4. 单击测试连通性，验证数据源访问正确无误后，确认并保存该数据源。

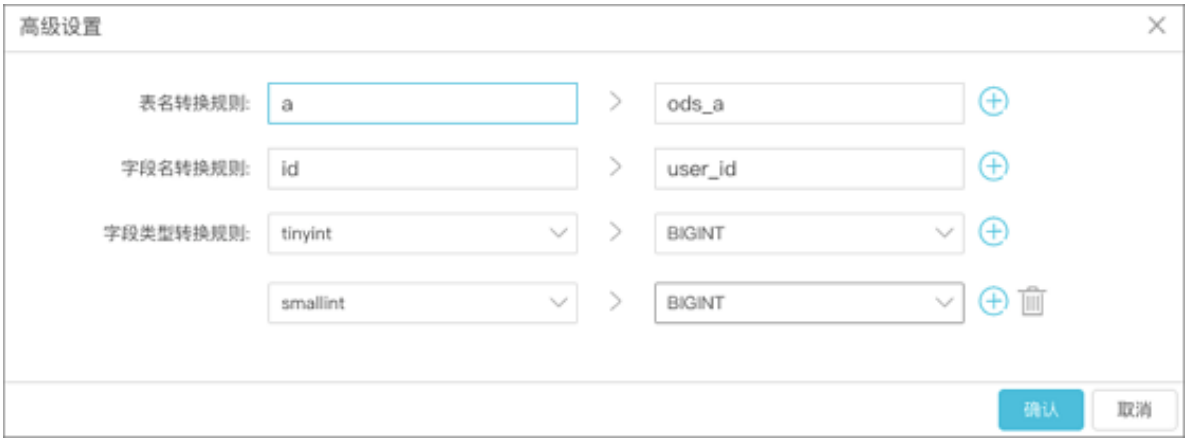
5. 新增数据源成功后，即可在数据源列表中看到新增的MySQL数据源（clone_database）。单击对应MySQL数据源后的整库迁移批量配置，即可进入对应数据源的整库迁移功能页面。



整库迁移页面主要分3块功能区域。

序号	功能区域	说明
1	待迁移表筛选区	此处将MySQL数据源（clone_database）下的所有数据库表以表格的形式展现出来，您可以根据实际需要批量选择待迁移的数据库表。
2	高级设置	此处提供了MySQL数据表和MaxCompute数据表的表名称、列名称、列类型的映射转换规则。
3	迁移模式、并发控制区	此处可以控制整库迁移的模式（全量、增量）、并发度配置（分批上传、整批上传）、提交迁移任务进度状态信息等。

6. 单击高级设置，您可以根据具体的需求选择转换规则。例如MaxCompute端建表时统一增加了ods_这一前缀。



7. 在迁移模式、并发控制区中，选择同步方式为每日增量，并配置增量字段为gmt_modified，数据集成默认会根据您选择的增量字段生成具体每个任务的增量抽

取where条件，并配如${bdp.system.bizdate}的DataWorks调度参数，形成针对每天的数据抽取条件。

数据集成抽取MySQL库表的数据是通过JDBC连接远程MySQL数据库，并执行相应的SQL语句，将数据从MySQL库中Select出来。由于是标准的SQL抽取语句，可以配置Where子句控制数据范围。此处您可以查看到增量抽取的Where条件如下所示：

```
STR_TO_DATE('${bdp.system.bizdate}', '%Y%m%d') <= gmt_modified AND
gmt_modified < DATE_ADD(STR_TO_DATE('${bdp.system.bizdate}', '%Y%m%d
'), interval 1 day)
```

为了对源头MySQL数据源进行保护，避免同一时间点启动大量数据同步作业带来数据库压力过大，此处选择分批上传模式，并配置从每日0点开始，每1小时启动3个数据库表同步。

最后单击提交任务，查看迁移进度信息，以及每一个表的迁移任务状态。

8. 单击a1表对应的迁移任务，跳转至数据开发页面，查看迁移结果。

此时便完成了将一个MySQL数据源（clone_database）整库迁移到MaxCompute的工作。这些任务会根据配置的调度周期（默认天调度）被调度执行，您也可以使用DataWorks调度补数据功能完成历史数据的传输。通过数据集成 > 整库迁移功能可以极大减少您初始化上云的配置、迁移成本。

查看整库迁移a1表任务执行成功的日志。

```
PHASE                 |    AVERAGE RECORDS |      AVERAGE BYTES |        MAX RECORDS | MAX RECORD`S BYTES |        MAX TASK ID |
MAX TASK INFO
READ_TASK_DATA        |              56345 |           128.12K |              56345 |           128.12K |              0-0-0 |
a1,jdbcUrl:[jdbc:mysql://dataxtest.mysql.rds.aliyuncs.com:3306/base_cdp]
2017-05-11 20:43:47.907 [job-31340023] INFO  LocalJobContainerCommunicator - Total 56345 records, 128121 bytes | Speed 62.56KB/
s, 28172 records/s | Error 0 records, 0 bytes |  All Task WaitWriterTime 0.486s |  All Task WaitReaderTime 0.082s | Percentage
 100.00%
2017-05-11 20:43:47.908 [job-31340023] INFO  LogReportUtil - report datax log is turn off
2017-05-11 20:43:47.908 [job-31340023] INFO  JobContainer -
任务启动时刻                    : 2017-05-11 20:43:42
任务结束时刻                    : 2017-05-11 20:43:47
任务总计耗时                    :                  5s
任务平均流量                    :           62.56KB/s
记录写入速度                    :          28172rec/s
读出记录总数                    :              56345
读写失败总数                    :                  0
2017-05-11 20:43:47 INFO ==================================================================
2017-05-11 20:43:47 INFO Exit code of the Shell command 0
2017-05-11 20:43:47 INFO --- Invocation of Shell command completed ---
2017-05-11 20:43:47 INFO Shell run successfully!
```

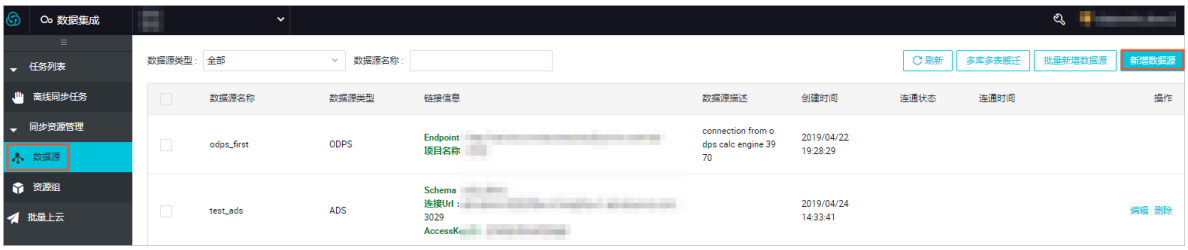
1.9.3 配置Oracle整库迁移

本文将为您介绍如何通过整库迁移功能，将Oracle数据整库迁移至MaxCompute。

整库迁移是为了提升用户效率、降低用户使用成本的一种快捷工具，它可以快速把Oracle数据库内所有表一并上传至MaxCompute。整库迁移的详细介绍请参见整库迁移概述。

操作步骤

1. **登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。**
2. **选择左侧菜单栏中的同步资源管理 > 数据源，进入数据源管理页面。**



3. 单击右上角的新增数据源，添加一个面向整库迁移的Oracle数据源（clone_database）。

4. 单击测试连通性，验证数据源访问正确无误后，单击完成。

5. 新增数据源成功后，即可在数据源列表中查看新增的Oracle数据源（clone_database）。单击相应Oracle数据源后的整库迁移批量配置，即可进入对应数据源的整库迁移功能页面。

   整库迁移页面主要分3块功能区域。

序号	功能区域	说明
1	待迁移表筛选区	此处将Oracle数据源（clone_database）下的所有数据库表以表格的形式展现出来，您可根据实际需要批量选择待迁移的数据库表。
2	高级设置	此处提供了Oracle数据表和MaxCompute数据表的表名称、列名称、列类型的映射转换规则。
3	迁移模式、并发控制区	并发控制区：此处可以控制整库迁移的模式（全量、增量）、并发度配置（分批上传、整批上传）、提交迁移任务进度状态信息等。

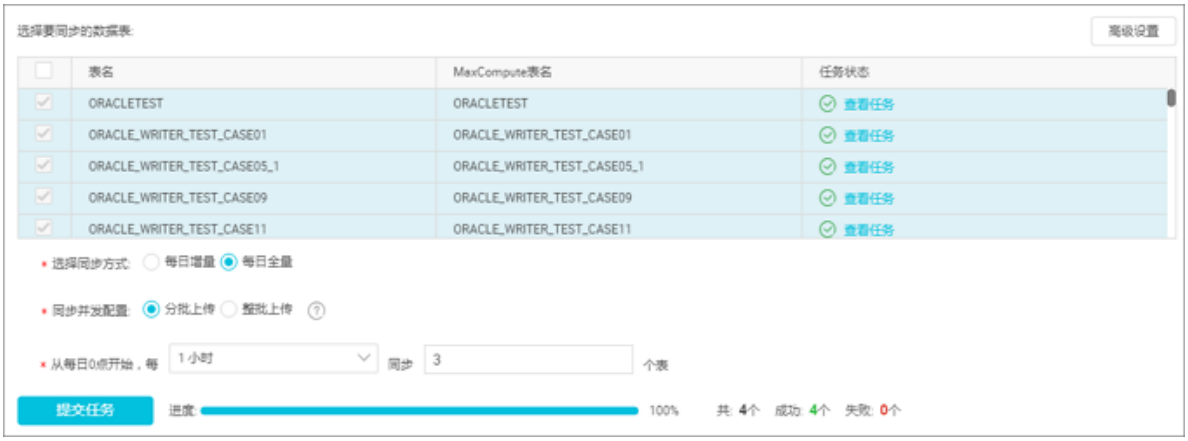
6. 单击高级设置，您可以根据具体的需求选择转换规则。
7. 在迁移模式、并发控制区中，选择同步方式为每日全量。

> **说明:**

如果您的表中有日期字段，可以选择同步方式为每日增量，并配置增量字段为日期字段，数据集成默认会根据您选择的增量字段生成具体每个任务的增量抽取where条件，并配合DataWorks调度参数，例如${bdp.system.bizdate}形成针对每天的数据抽取条件。

为了对源头Oracle数据源进行保护，避免同一时间点启动大量数据同步作业导致数据库压力过大，此处选择分批上传模式，并配置从每日0点开始，每1小时启动3个数据库表同步。

最后单击提交任务，这里可以看到迁移进度信息，以及每一个表的迁移任务状态。



8. 单击表对应的查看任务，跳转至数据集成的任务开发页面，您可查看任务的运行详情。

此时便完成了将一个Oracle数据源（clone_database）整库迁移到MaxCompute的工作。这些任务会根据配置的调度周期（默认天调度）被调度执行，您也可以使用DataWorks调度补数据功能完成历史数据的传输。通过数据集成 > 整库迁移功能可以极大减少您初始化上云的配置、迁移成本。

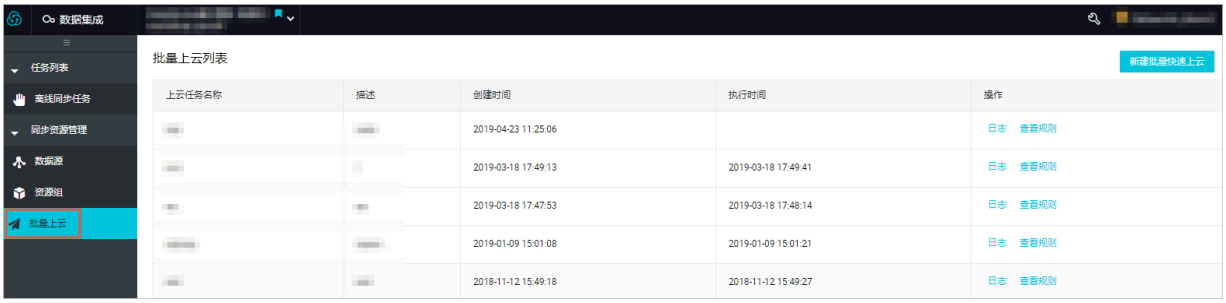
1.10 批量上云

1.10.1 批量上云

批量上云是帮您提升效率、降低使用成本的一种快捷工具，它可以快速把MySQL、Oracle、SQL Server数据库内的所有表一并上传至MaxCompute中，节省大量初始化数据上云的批量任务创建时间。

您可以灵活地配置表名转换、字段名转换、字段类型转换、目标表新增字段、目标表字段赋值、数据过滤、目标表名前缀等规则，来满足您的业务需求。

您可以进入数据集成 > 批量上云页面，查看您配置的批量上云任务。



说明:

- **批量上云列表中，操作栏下的日志和规则只能查看不能修改。**
- **如果您提交规则后，没有提交任务，则没有运行时间，并且此配置规则无效。**

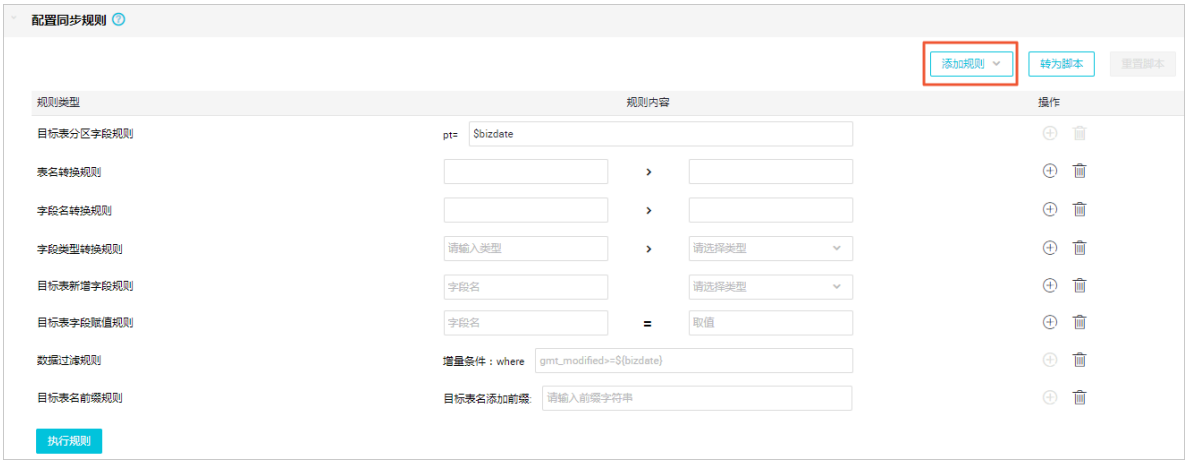
操作步骤

1. **单击批量上云页面右上角的新建批量快速上云。**
2. **选择同步的数据源。**

   选择添加成功的同步数据源，此处可以选择多个数据源并且数据源类型相同，例如均是MySQL、Oracle或SQL Server，详情请参见批量新增数据源。

3. 配置同步规则。

您可以根据自身需求选择相应的规则配置，然后执行规则，并检查DDL和同步脚本确认规则效果。



说明：

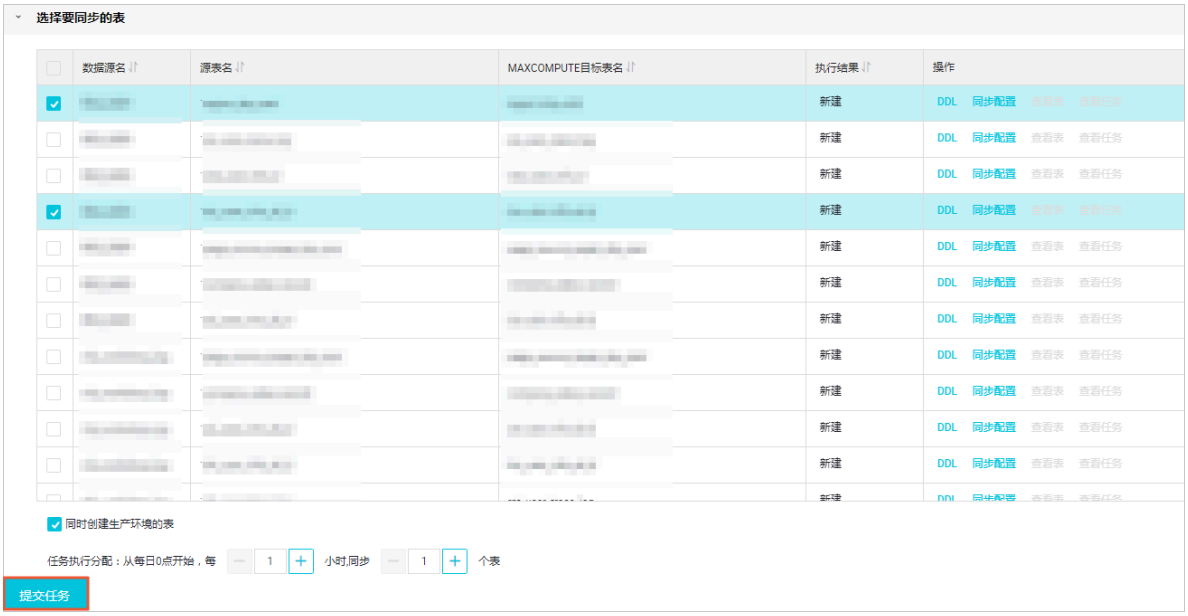
- 如果界面中的规则无法满足您的需求，可以尝试脚本模式。
- 配置完规则后，您必须执行规则并提交任务，否则您配置的规则在刷新或关闭浏览器后没有相关的记录。
- 如果您需要在批量上云时对表前缀进行设置，请参见批量上云时给目标表名加上前缀。

操作	配置	说明
添加规则	目标表分区字段规则	展现分区的内容，符合调度参数配置，详情请参见#unique_20。
	表名转换规则	选择您的数据库表名的任何词，转换成您需要的内容。
	字段名转换规则	选择您的表中字段名的任何词，转换成您需要的内容。
	字段类型转换规则	选择您的数据源表中具有的数据类型，转换成您需要的数据类型。
	目标表新增字段规则	可以在MaxCompute表中增加一列，根据您的需求设置名称。
	目标表字段赋值规则	给您增加的字段赋值。
	数据过滤规则	针对您选择的数据源，对表中的数据进行过滤。
	目标表名前缀规则	给表名添加一个前缀。

操作	配置	说明
转为脚本	配置规则时可以转为脚本模式配置，与UI模式相比，单个规则可以指定作用范围。但UI模式转为脚本模式后，无法反向转换回UI配置模式。	
重置脚本	转换脚本后才能重置脚本，单击后提供统一的脚本模板。	
执行规则	单击执行规则，可以看到规则对DDL脚本和同步脚本的影响，此按钮不会创建任务，仅提供DDL和同步脚本的预览。 您可以选择一部分表检查对应的DDL和同步脚本，确认是否符合规则。	

4. 选择要同步的表并提交。

   您可以选择多个表进行批量提交，MaxCompute表会根据上面配置规则生成。如果执行失败，将鼠标放到执行结果上，会提示相关的原因。



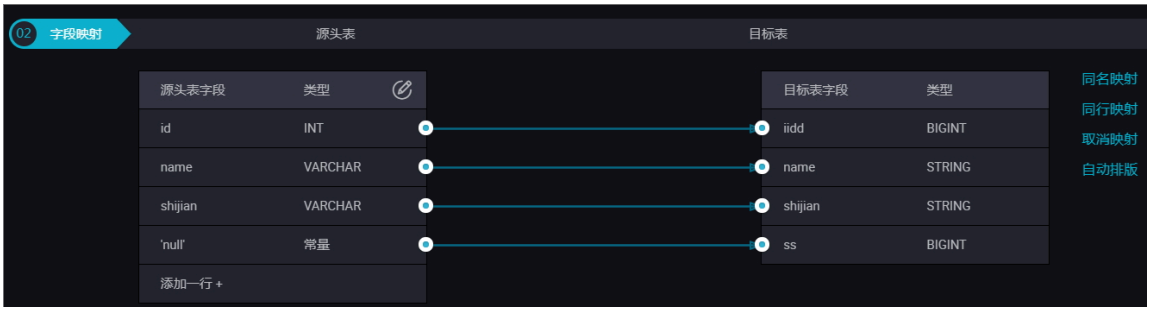
配置	说明
DDL	单击后可以查看相关建表语句，只能查看不能修改。
同步配置	单击后可以查看您配置的任务，以脚本模式展现。
查看表	单击后可以查看MaxCompute建表的具体情况。
查看任务	提交成功后，您可以进入数据开发 > 业务流程页面，查看您的批量上云任务。

5. 查看任务。

您选择几个数据源，便会产生几个业务流程，通常命名规则是clone_database_数据源名。每张表会产生一个同步任务，命名规则是数据源名2odps_表名。

a. 选择数据源：根据批量上云生成的MySQL同步至MaxCompute（ODPS）的同步任务，数据过滤条件在配置数据过滤规则后产生。

b. 字段映射：目标端是根据您配置相关字段规则而产生，可以根据您配置的规则进行查看。



c. 通道配置。

配置	说明
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。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

说明:

任务的具体配置请参见配置*Reader*插件和配置*Writer*插件。

6. 运行任务。

   直接单击运行，同步任务会立刻运行。您也可以单击提交，将同步任务提交到调度系统中，调度系统会按照配置属性在从第二天开始自动定时执行，详情请参见调度配置。

说明:

- 简单模式：提交之后直接到生产环境。
- 标准模式：提交后到开发环境，然后发布到生产环境。

1.10.2 批量新增数据源

本文将为您介绍如何批量新增数据源。

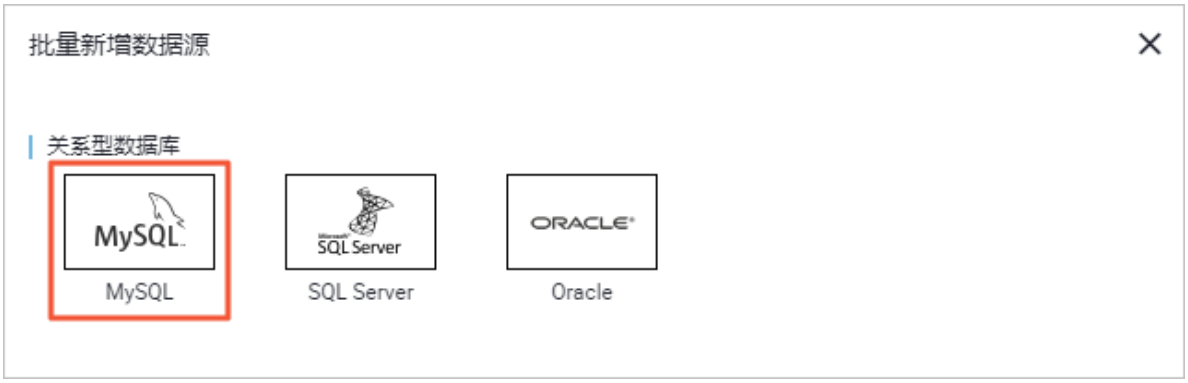
说明:

- 快速上云目前仅支持MySQL、Oracle和SQL Server三种类型的数据源。
- 批量新增数据源目前仅支持连接串模式（数据集成网络可直接连通）。

- 添加MySQL、Oracle和SQL Server数据源后，需要测试连通性。当连通状态为成功时，批量上云选择同步数据源列表才能选择该数据源。

1. 以项目管理员身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 在数据集成 > 同步资源管理 > 数据源页面，单击批量新增数据源。

3. 在批量新增数据源对话框中，选择相应的数据源类型，选择文件进行上传，本文以新增MySQL数据源为例。



配置	说明
数据源类型	仅支持连接串模式（数据集成网络可直接连通）。
脚本上传	首先单击模板下载，在模板中添加您的数据源名称、数据源描述、链接地址、用户名和密码。 说明: 通常会有一个默认的数据源mysql_001_di_test，您可以直接删除添加您自己的数据源。
选择文件	单击选择文件，选择修改好的模板。
开始创建	文件上传成功后，单击开始创建，您上传的结果会在文本框中展示，例如成功个数、失败的个数、原因等。

4. 上传成功后，单击关闭。
5. 在数据源列表页面，勾选相应的数据源，单击批量测试连通性。

> 说明:
> 必须保证数据源的连通状态为成功，方可进行批量上云操作，详情请参见批量上云。

1.11 最佳实践

1.11.1 （一端不通）数据源网络不通的情况下的数据同步

数据集成通过部署Agent，可以打通任意网络环境之间的数据传输同步。本文将为您介绍如何在仅一端数据源无法连通的情况下，进行数据同步。

两端数据源均无法连通的情况请参见（两端不通）数据源网络不通的情况下的数据同步。

场景说明

复杂网络环境主要包含以下两种情况：

- 数据的来源端和目标端有一端为私网环境。
  - VPC环境（除RDS）<->公网环境
  - 金融云环境<->公网环境
  - 本地自建无公网环境<->公网环境
- 数据的来源端和目标端均为私网环境。
  - VPC环境（除RDS）<->VPC环境（除RDS）
  - 金融云环境<->金融云环境
  - 本地自建无公网环境<->本地自建无公网环境
  - 本地自建无公网环境<->VPC环境（除RDS）
  - 本地自建无公网环境<->金融云环境

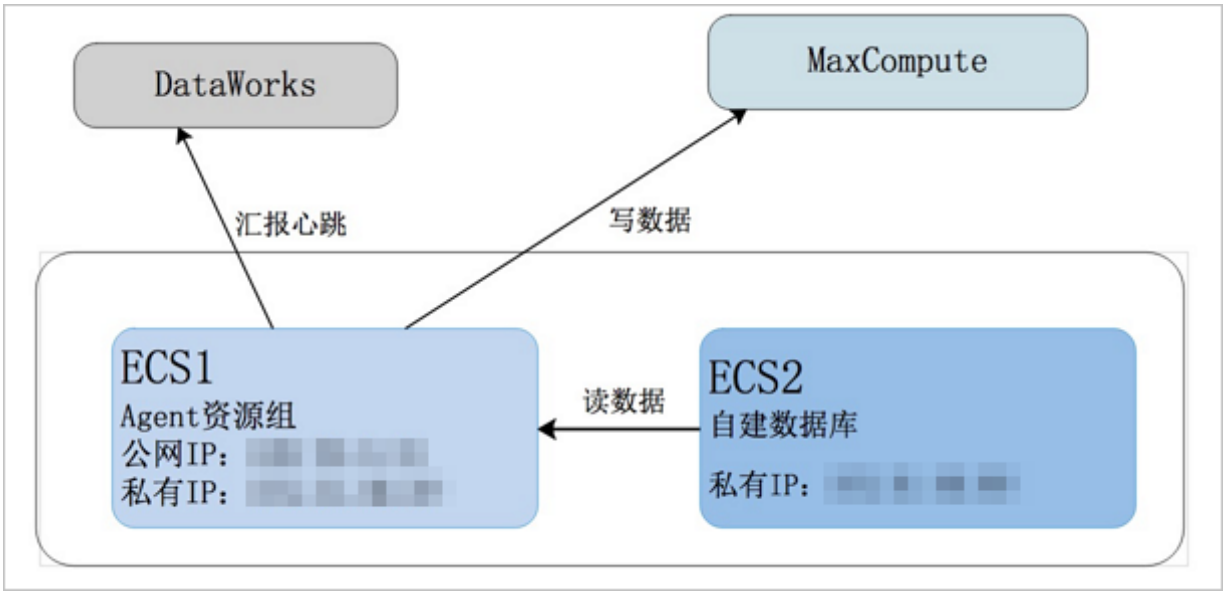
实现逻辑

针对第一种复杂网络环境，可以在私网环境的一端相同网络环境下的机器上部署数据集成Agent，通过Agent与外部公网连通。私网环境通常有以下两种情况：

- 购买云服务ECS上搭建的数据库，没有分配公网IP或弹性公网IP。
- 本地IDC机房无公网IP。

云服务ECS

此场景下的数据同步方式，如下图所示。



- 由于ECS2服务器无法访问公网，所以需要准备1台和ECS2在同一网段，并且可以访问公网的机器ECS1部署Agent。
- 将ECS1作为资源组，并且同步任务运行在该机器上。

说明:

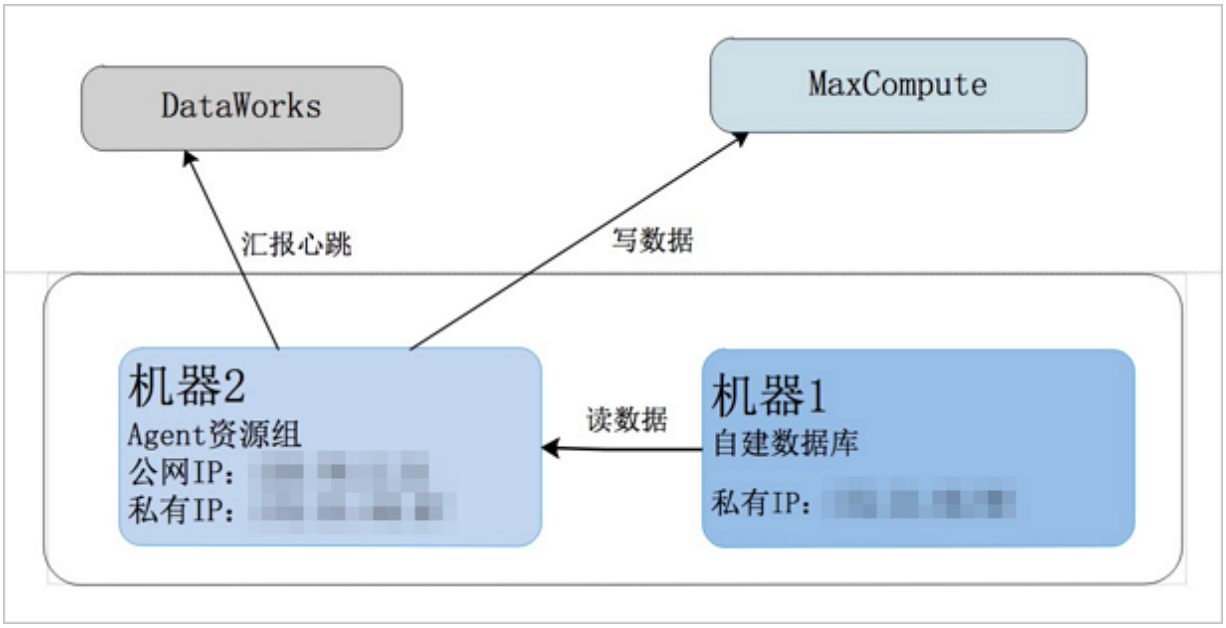
您需要给数据库赋权限，让ECS2服务器能访问到相应的数据库，方可读取该数据库的数据至ECS1中。授权命令如下所示：

```
grant all privileges on *.* to 'demo_test'@'%' identified by '密码';
 --> %号代表给所有 IP 授权<br>
```

ECS2上的自建数据源同步任务运行在自定资源组上，需要给自定资源组机器授权，添加ECS2机器内/外网IP和端口至ECS1的安全组，详情请参见添加安全组。

无公网IP本地IDC机房

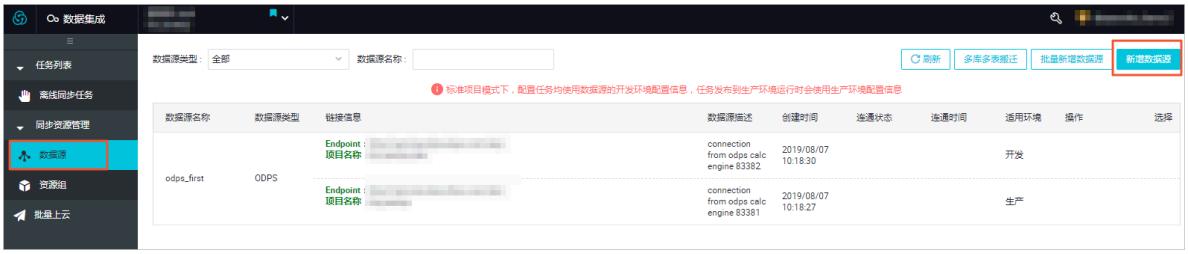
此场景下的数据同步方式，如下图所示。



- 由于机器1无法访问公网，所以需要准备1台和机器1在同一网段，并且可以访问公网的机器2部署Agent。
- 将机器2作为任务资源组，并且同步任务运行在该机器上。

配置数据源

1. 以开发者身份登录*DataWorks*控制台，单击相应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



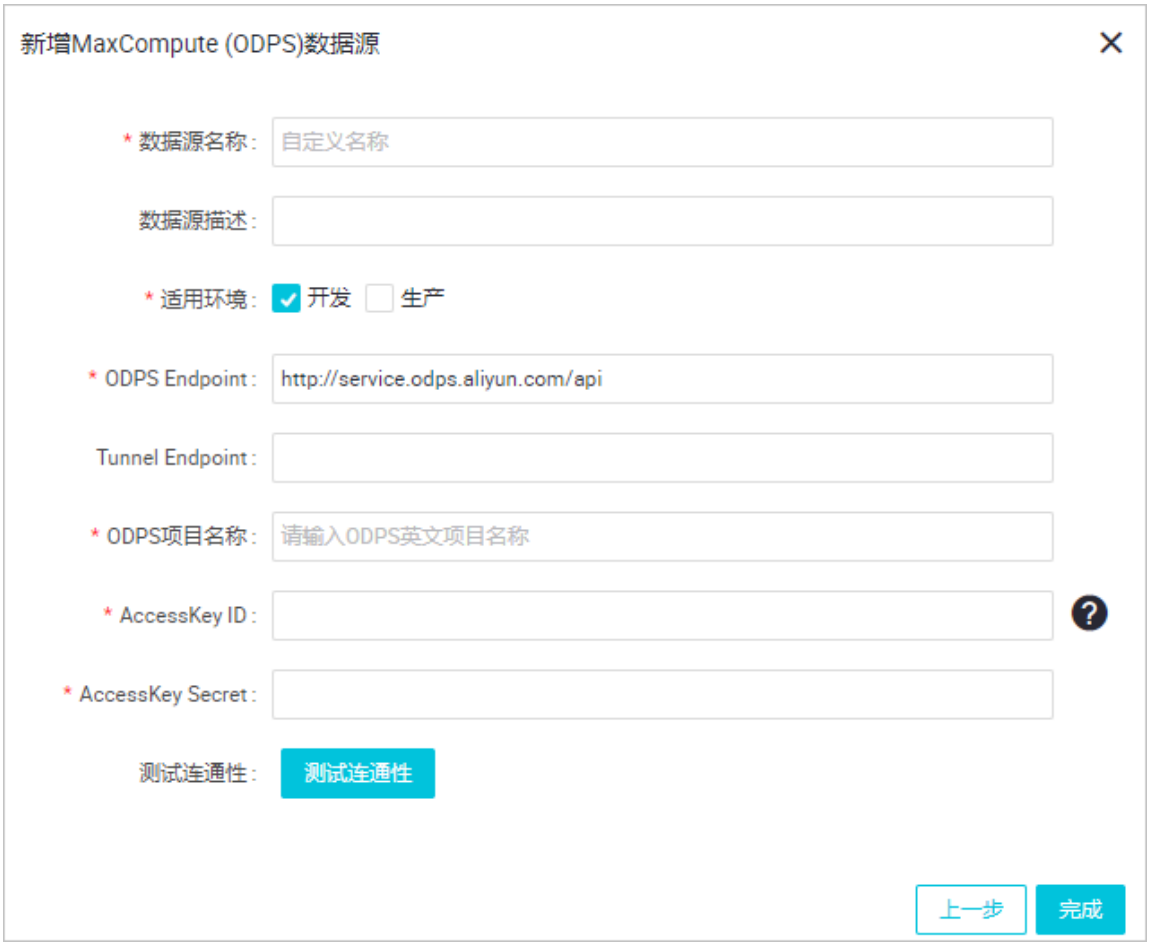
3. 此处需要新增连接串模式（数据集成网络不可直接连通）类型的MySQL数据源作为来源端数据源，新增MaxCompute数据源作为目标端数据源。

   - 来源端数据源

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 连接串模式（数据集成网络不可直接连通）。
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。

配置	说明
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此配置。
资源组	选择部署Agent的机器，通过Agent与外部公网连通，特殊网络环境的数据源可以将同步任务运行在资源组上。详情请参见新增任务资源。
JDBC URL	JDBC连接信息，格式为`jdbc:mysql://ServerIP:Port/Database`。
用户名/密码	数据库对应的用户名和密码。

- 目标端数据源



配置	说明
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。

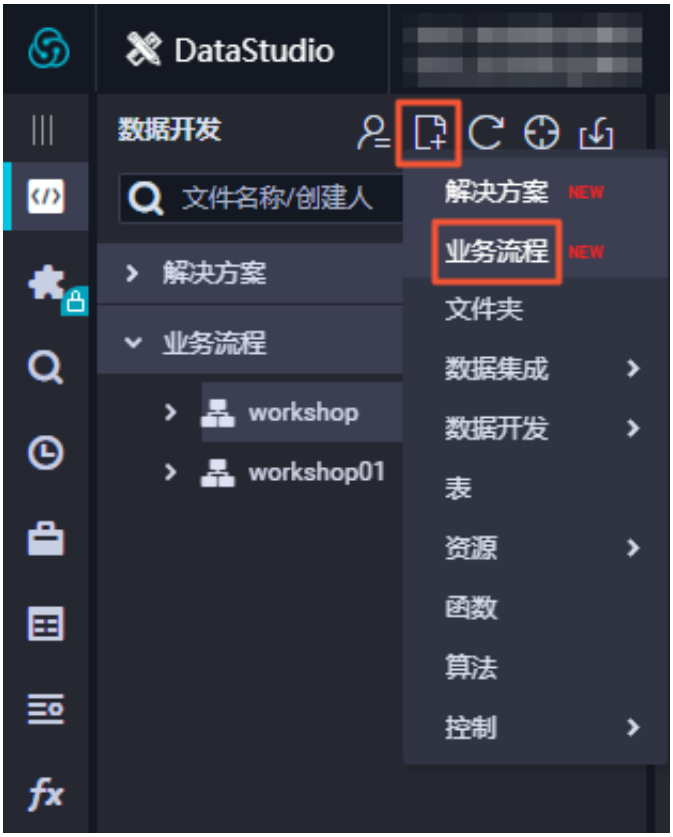
配置	说明
适用环境	可以选择开发或生产环境。 说明: 仅标准模式工作空间会显示此配置。
ODPS Endpoint	默认只读，从系统配置中自动读取。
Tunnel Endpoint	MaxCompute Tunnel服务的连接地址，详情请参见#unique_93。
ODPS项目名称	MaxCompute（ODPS）项目的名称。
AccessID/AceessKey	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

4. 源端数据源配置完成后，直接单击完成。

   目标数据源配置完成后，单击测试连通性。连通成功后，单击完成。

新建数据同步节点

1. 单击左上角的DataWorks图标，选择全部产品 > DataStudio（数据开发）。
2. 进入DataStudio（数据开发）页面，选择新建 > 业务流程。



3. 在新建业务流程对话框中，填写业务流程名称和描述，单击新建。

4. 展开业务流程，右键单击数据集成，选择新建数据集成节点 > 数据同步，输入节点名称。

5. 单击提交。

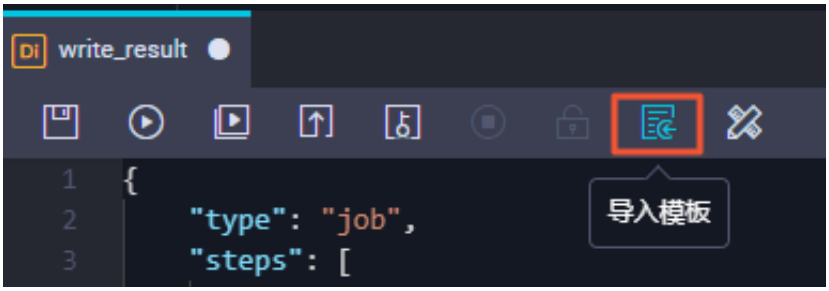
导入模板

1. 成功创建数据同步节点后，单击工具栏中的转换脚本。

2. 单击提示对话框中的确认，即可进入脚本模式进行开发。

> **说明:**
> 脚本模式支持更多功能，例如网络不可达情况下的同步任务编辑。

3. 单击工具栏中的导入模板。



4. 选择从来源端的MySQL数据源同步至目标端的ODPS数据源的导入模板，单击确认。

配置	说明
来源类型	选择MySQL。
数据源	选择新建的来源端的数据源。
目标类型	选择ODPS。
数据源	选择新建的目标端的数据源。

5. 导入模板后，根据自身需求进行代码的编辑。

```
{
"type": "job",
"configuration": {
 "setting": {
   "speed": {
     "concurrent": "1",//作业并发数。
     "mbps": "1"//作业速率上限。
   },
   "errorLimit": {
     "record": "0"//错误记录数。
   }
```

```
 },
 "reader": {
   "parameter": {
     "splitPk": "id",//切分键。
     "column": [//目标端的列名。
       "name",
       "tag",
       "age",
       "balance",
       "gender",
       "birthday"
     ],
     "table": "source",//源端的表名。
     "where": "ds = '20171218'",//过滤条件。
     "datasource": "private_source"//数据源名称，需要和添加的数据源名保持
一致。
   },
   "plugin": "mysql"
 },
 "writer": {
   "parameter": {
     "partition": "ds='${bdp.system.bizdate}'",//分区信息。
     "truncate": true,
     "column": [//目标端的列。
       "name",
       "tag",
       "age",
       "balance",
       "gender",
       "birthday"
     ],
     "table": "random_generated_data",//目标端的表名。
     "datasource": "odps_mrtest2222"//数据源名称，需要和添加的数据源名保持
一致。
   },
   "plugin": "odps"
 }
},
"version": "1.0"
}
```

运行同步任务

您可以通过以下两种方式运行任务：

- **在数据同步节点的编辑页面，直接单击运行。**
- **调度运行，提交调度的步骤请参见调度配置。**

1.11.2 （两端不通）数据源网络不通的情况下的数据同步

数据集成通过部署Agent，可以打通任意网络环境之间的数据传输同步。本文将为您介绍如何在两端数据源均无法连通的情况下，进行数据同步。

仅一端数据源无法连通的情况请参见（一端不通）数据源网络不通的情况下的数据同步。

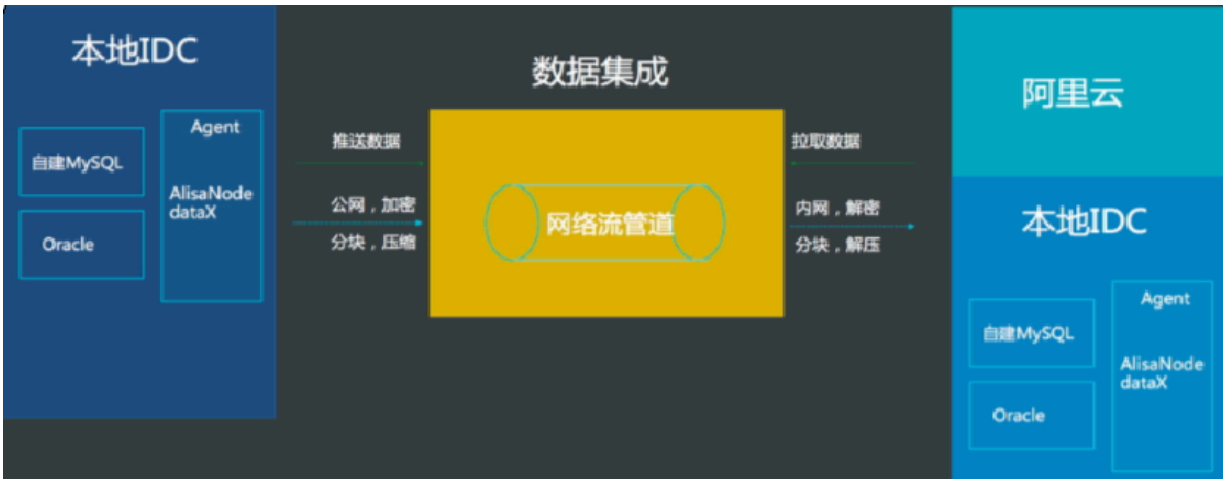
场景说明

复杂网络环境主要包含以下两种情况：

- 数据的来源端和目标端有一端为私网环境。
  - VPC环境（除RDS）<->公网环境
  - 金融云环境<->公网环境
  - 本地自建无公网环境<->公网环境
- 数据的来源端和目标端均为私网环境。
  - VPC环境（除RDS）<->VPC环境（除RDS）
  - 金融云环境<->金融云环境
  - 本地自建无公网环境<->本地自建无公网环境
  - 本地自建无公网环境<->VPC环境（除RDS）
  - 本地自建无公网环境<->金融云环境

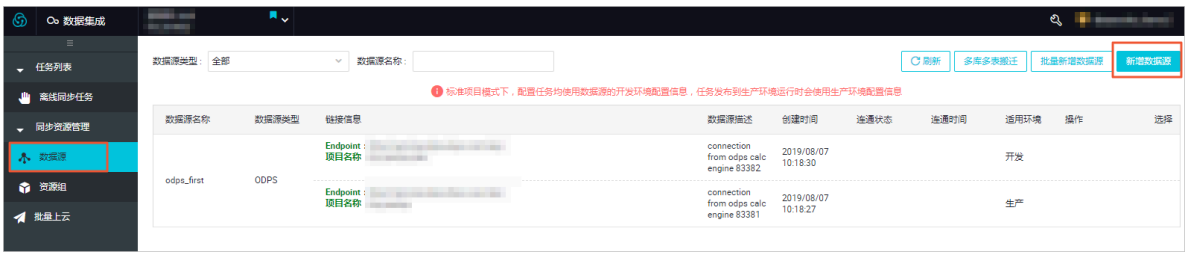
实现逻辑

针对第二种复杂网络环境，可以在两端数据源的相同网络环境下，均部署数据集成Agent。来源端Agent负责推送数据至数据集成服务端，目标端Agent负责拉取数据至本地，且数据在传输过程中进行数据的分块、压缩和加密，以保障数据传输的及时性和安全性。



配置数据源

1. 以开发者身份登录*DataWorks*控制台，单击对应工作空间后的进入数据集成。
2. 选择同步资源管理 > 数据源，单击新增数据源。



3. 在新增数据源弹出框中，选择数据源类型为FTP。

4. 填写FTP数据源的各配置项。

此处选择连接串模式（数据集成网络不可直接连通）类型的数据源。

添加源端和目标端的数据源

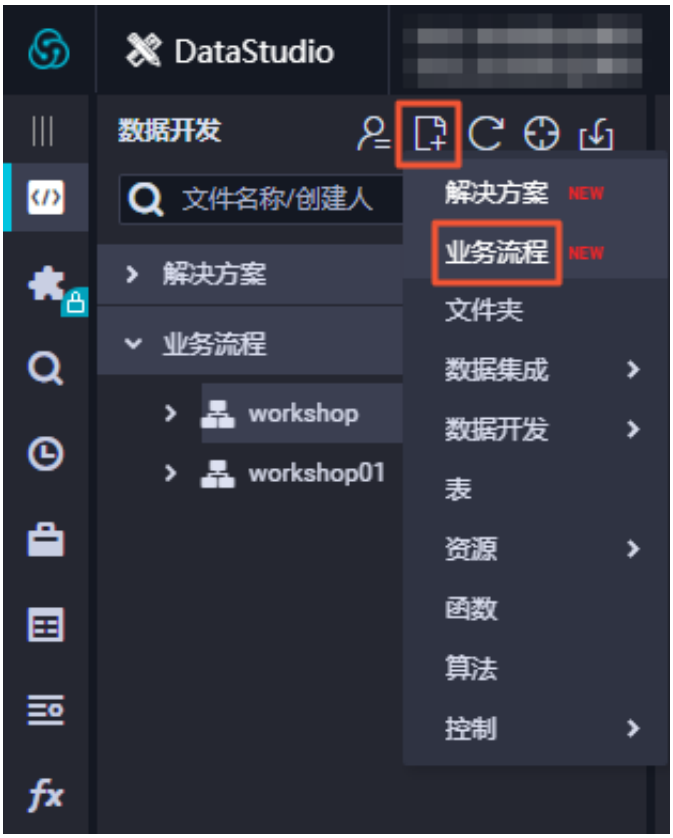
配置	说明
数据源类型	当前选择的数据源类型为FTP > 连接串模式（数据集成网络不可直接连通）。 选择此类型的数据源需要使用自定义调度资源才能进行同步，您可以单击帮助手册查看详情。
数据源名称	数据源名称必须以字母、数字、下划线组合，且不能以数字和下划线开头。

配置	说明
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此配置。
资源组	是选择部署Agent的机器，来源端Agent负责推送数据至数据集成服务端，目标端Agent负责拉取数据至本地。详情请参见新增任务资源。
Portocol	目前仅支持FTP和SFTP协议。
Host	对应FTP主机的IP地址。
Port	如果选择的是FTP协议，则端口默认为21。如果选择的是SFTP协议，则端口默认为22。
用户名/密码	访问该FTP服务的账号密码。

5. 单击完成。

新建数据同步节点

1. 单击左上角的DataWorks图标，选择全部产品 > DataStudio（数据开发）。
2. 进入DataStudio（数据开发）页面，选择新建 > 业务流程。



3. 在新建业务流程对话框中，填写业务流程名称和描述，单击新建。
4. 展开业务流程，右键单击数据集成，选择新建数据集成节点 > 数据同步，输入节点名称。

5. 单击提交。

导入模板

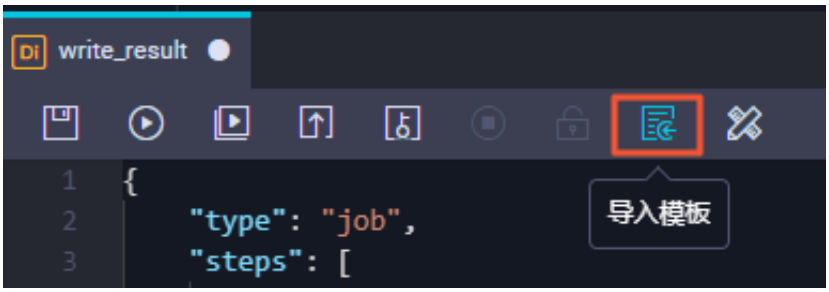
1. 成功创建数据同步节点后，单击工具栏中的转换脚本。

2. 单击提示对话框中的确认，即可进入脚本模式进行开发。

> **说明:**

脚本模式支持更多功能，例如网络不可达情况下的同步任务编辑。

3. 单击工具栏中的导入模板。



4. 选择从来源端的FTP数据源同步至目标端的FTP数据源的导入模板，单击确认。

配置	说明
来源类型	选择FTP。
数据源	选择新建的来源端的数据源。
目标类型	选择FTP。
数据源	选择新建的目标端的数据源。

5. 导入模板后，根据自身需求进行代码的编辑。

运行同步任务

您可以通过以下两种方式运行任务：

- 在数据同步节点的编辑页面，直接单击运行。
- 调度运行，提交调度的步骤请参见调度配置。

1.11.3 数据增量同步

需要同步的数据根据数据写入后是否会发生变化，分为不会发生变化的数据（通常是日志数据）和会变化的数据（人员表，例如人员的状态会发生变化）。

示例说明

针对以上两种数据场景，需要设计不同的同步策略，本文以把业务RDS数据库的数据同步至MaxCompute为例。

根据等幂性原则（1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到一张单独的表/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。

不变的数据进行增量同步

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. 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，需先做一次全量数据同步，将历史数据同步到昨天的分区。

2. 执行下述语句创建MaxCompute表。

```
--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);
```

3. 配置同步历史数据的任务，详情请参见#unique_144。

   测试同步任务成功后，单击右侧的调度配置，勾选暂停调度并重新提交/发布，避免任务自动调度执行。

4. 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');
```

5. 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，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。

6. 查看同步结果。

   单击右侧的调度配置，设置任务的调度周期为天调度。提交/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。

会变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，需要每天只做增量同步。但MaxCompute不支持Update语句修改数据，只能用其它方法实现。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。

数据准备

```
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);
```

- **每天全量同步**

  1. **执行下述语句创建MaxCompute表，新建表的详情请参见***#unique_145/unique_145_Connect_42_section_lgp_ld4_q2b*。

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

  2. **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。

3. 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

不推荐使用此方式，只有如不支持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。虽然实际上很少直接删除数据，都是使用逻辑删除，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。

数据准备

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
)
```

1. 配置同步任务，将全量数据直接写入结果表。

   说明:

   只需执行一次，执行成功后需单击页面右侧的调度配置，勾选暂停调度。

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

3. 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，并且需要用额外的计算进行数据合并。

如果不是必要情况，会变化的数据进行每天全量同步即可。如果对历史数据希望只保留一定的时间，超出时间的进行自动删除，可以设置Lifecycle。

1.11.4 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调优的方法。

场景分类

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。

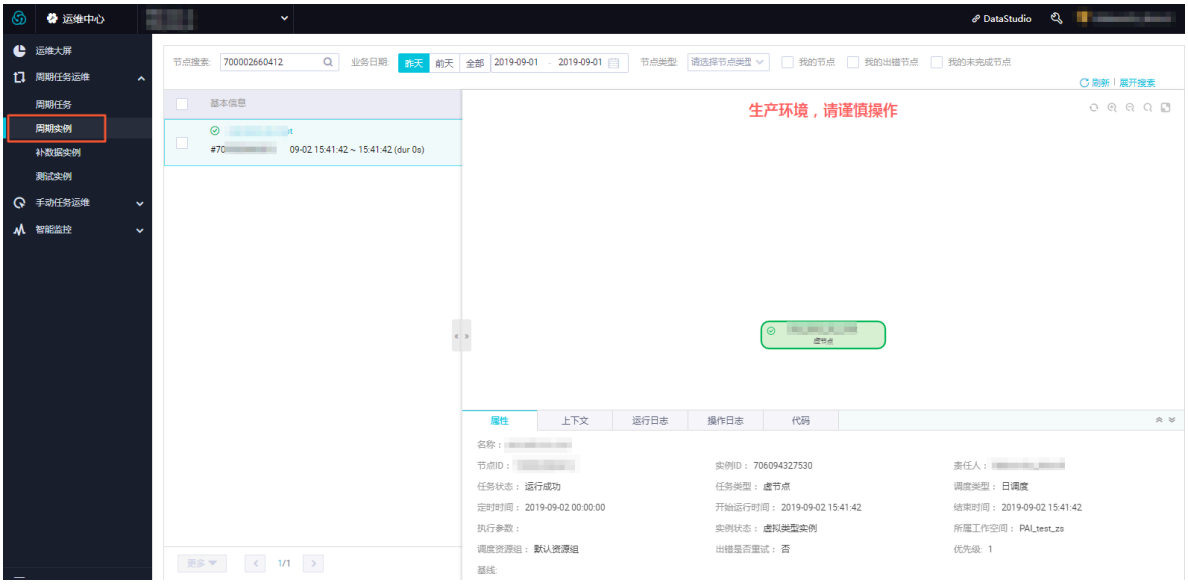
前提条件

正式开始数据同步任务调优前，请首先收集下列信息：

- 任务运行日志（从日志开始打印到结束）
- 任务的属性标签页信息

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。



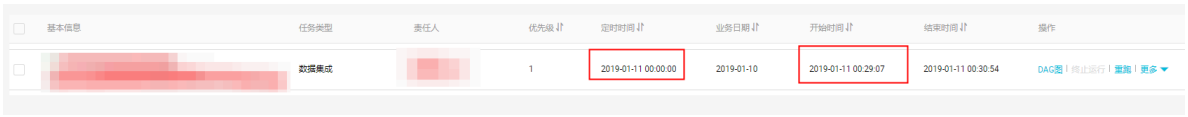
- 二级运行资源：您可以进入数据集成 > 同步资源管理 > 资源组页面，新增和查看二级任务运行资源。

场景一：任务开始运行时间和调度时间差异较大

在该场景下，您首先需要任务运行日志和任务属性标签页信息 。对比分析发现， 运行日志中开始running的时间和属性节点的调度时间是有差异的，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页面查看用户任务的属性标签页查，发现调度时间在00:00， 但是开始运行时间在00:29，怀疑时间主要消耗在等待调度上。



2. 在实例页面右键查看用户任务运行日志，任务从00:29分开始运行，00:30执行结束，整个任务执行仅仅花费了1分钟。说明本次任务本身执行无问题。

问题解法

1. 首先建议您观察您的项目下是否有较多的任务同时调度。默认资源组下的一级调度资源有限，同时调度的任务较多会有其他任务排队等待。
2. 通常每天0点-2点是业务调度的高峰期， 建议您的业务运行时间尽量避开高峰期 。

场景二：任务同步速率慢

在该场景下，通过分析运行日志，通常有以下两种情况：

- 任务一直在运行，但速率是0。
- 任务速率较低。

任务速率为0

查看运行日志，看到任务长时间处于run的状态，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，任务长时间在run，速率为0，从18:00开始到21:13结束。

```
2018-12-27 18:00:16 : State: 1(SUBMIT) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 18:00:26 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 18:00:36 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 18:00:46 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 18:00:56 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 18:01:06 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
```

```
2018-12-27 21:13:06 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 21:13:16 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 21:13:26 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 21:13:36 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 21:13:46 : State: 3(RUN) | Total: 0R 0B | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0%
2018-12-27 21:13:56 : State: 0(SUCCESS) | Total: 601R 25.4KB | Speed: 60R/s 2.5KB/s | Error: 0R 0B | Stage: 100.0%
2018-12-27 21:13:56 : DI Job[496038] completed successfully.
```

2. 查看运行日志信息有truncate操作记录，从18:00开始到21:13 truncate操作结束。

```
2018-12-27 18:00:23.063 [job-xxxxxx] INFO  JobContainer - jobContainer starts to do prepare ...
2018-12-27 18:00:23.064 [job-xxxxxx] INFO  JobContainer - DataX Reader.Job [postgresqlreader] do prepare work .
2018-12-27 18:00:23.082 [job-xxxxxx] INFO  CommonRdbmsWriter$Job - Begin to execute preSqls:[truncate table
2018-12-27 21:13:45.688 [job-xxxxxx] INFO  JobContainer - jobContainer starts to do split ...
2018-12-27 21:13:45.693 [job-xxxxxx] INFO  JobContainer - DataX Reader.Job [postgresqlreader] splits to [1] tasks.
2018-12-27 21:13:45.694 [job-xxxxxx] INFO  JobContainer - DataX Writer.Job [sqlserverwriter] splits to [1] tasks.
2018-12-27 21:13:45.711 [job-xxxxxx] INFO  JobContainer - jobContainer starts to do schedule ...
2018-12-27 21:13:45.714 [job-xxxxxx] INFO  JobContainer - Scheduler starts [1] taskGroups.
```

问题解法

如上所示，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。

任务速率慢

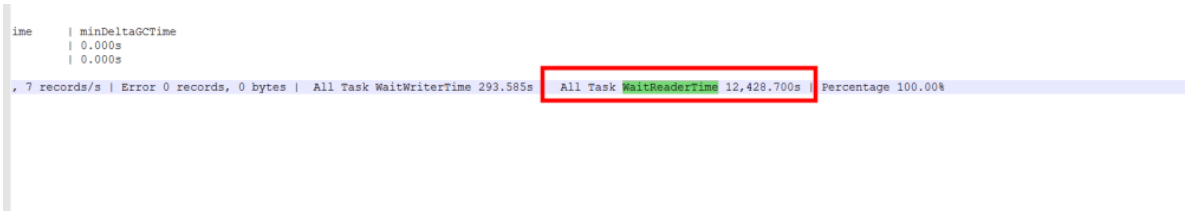
查看运行日志，看到任务同步速率不为0，但是速率慢。

问题示例

1. 获取运行日志后，查看日志中信息同步速率确实比较慢， 约为1.93kb/s。

```
PS Scavenge          | 3              | 2              | 0              | 0.041s         | 0.027s         | 0.000s
2019-01-14 03:29:46.555 [job-1390111] INFO  JobContainer - PerfTrace not enable!
2019-01-14 03:29:46.598 [job-1390111] INFO  LocalJobContainerCommunicator - Total 33914 records, 9085250 bytes | Speed 1.93KB/s, 7 records/s | Error 0 records, 0 bytes |  All Tas
2019-01-14 03:29:46.600 [job-1390111] INFO  JobContainer -
任务启动时刻                    : 2019-01-14 02:03:24
任务结束时刻                    : 2019-01-14 03:29:46
任务总计耗时                    :               5182s
任务平均流量                    :            1.93KB/s
记录写入速度                    :              7rec/s
读出记录总数                    :              33914
读写失败总数                    :                  0
```

2. 查看运行日志中同步时间消耗字段WaitWriterTime、WaitReaderTime的信息， 发现WaitReaderTime时长较长，主要在等待读数据。



问题解法

针对速率比较慢的情况， 可以看下主要在等Writer还是Reader，如果是读或写慢，需要查看对应的源数据库或目的数据库的负载情况。

1.11.5 通过数据集成导入数据到Elasticsearch

本文将为您介绍如何通过数据集成对离线Elasticsearch进行数据导入的操作。

准备工作

1. 准备阿里云账号，并创建账号的访问密钥，即AccessID和AccessKey。详情请参见#unique_11。
2. 开通MaxCompute，自动产生一个默认的MaxCompute数据源，并使用主账号登录DataWorks。
3. 创建工作空间，您可以在工作空间中协作完成业务流程，共同维护数据和任务等，因此使用DataWorks之前需要先创建一个工作空间。详情请参见#unique_149。

   > 说明:
   > 如果您想通过子账号创建数据集成任务，可以赋予其相应的权限。详情请参见#unique_150和#unique_151。

4. 配置好相关的数据源，详情请参见数据源配置。

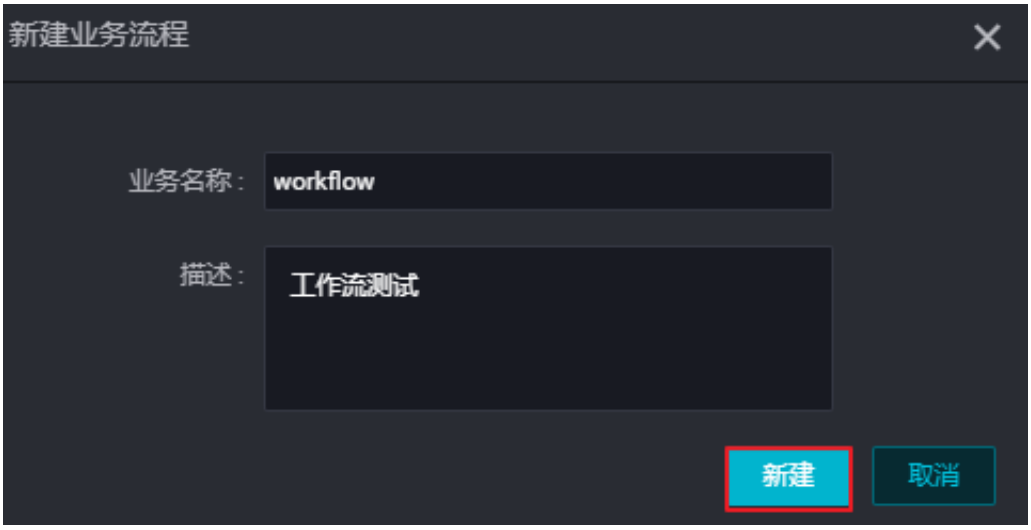
新建数据同步节点

1. 以开发者身份登录DataWorks控制台，单击相应工作空间后的进入数据开发。

2. 新建业务流程。

a. 右键单击业务流程，选择新建业务流程。

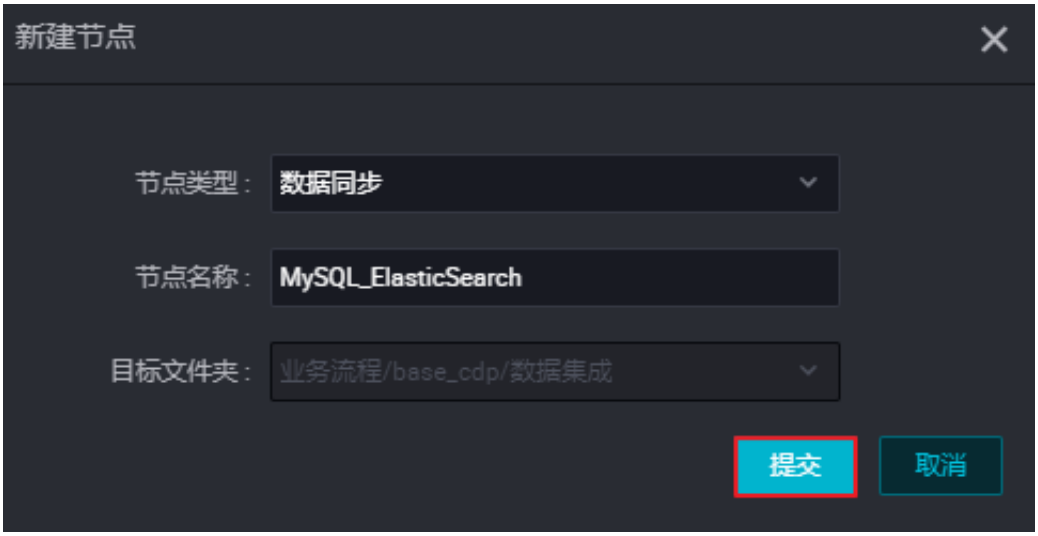
b. 在新建业务流程对话框中，输入业务名称和描述。



c. 单击新建，即可完成业务流程的创建。

3. 展开业务流程，右键单击数据集成，选择新建数据集成节点 > 数据同步。

4. 填写新建节点对话框中的配置，单击提交。

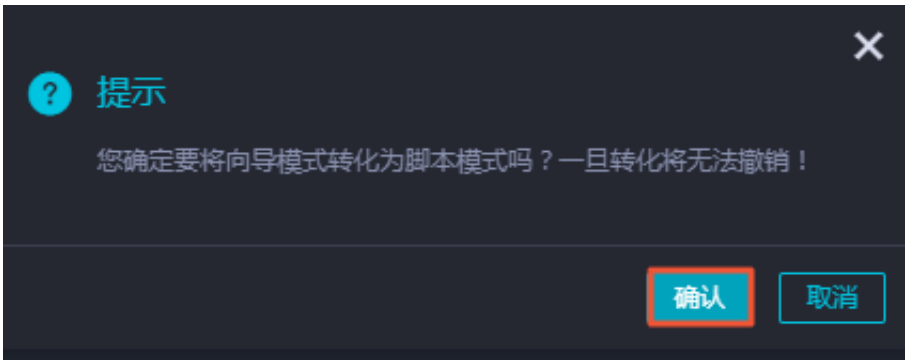


配置	说明
节点类型	默认数据同步类型。
节点名称	填写该节点名称。
目标文件夹	默认放在相应的业务流程下。

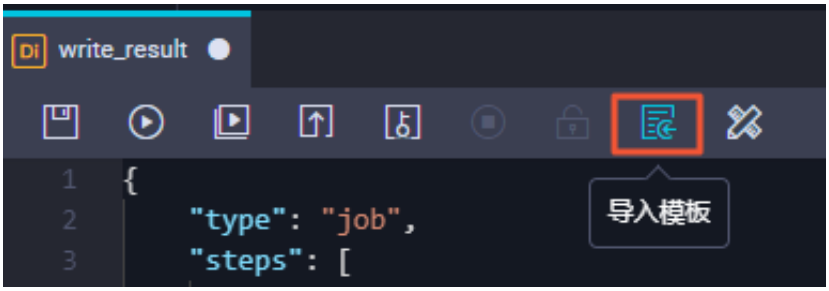
配置数据同步节点

1. 成功创建数据同步节点后，单击工具栏中的转换脚本。

2. 单击提示对话框中的确认，即可进入脚本模式进行开发。



3. 单击工具栏中的导入模板。



4. 填写导入模板对话框中的配置。

配置	说明
来源类型	此处选择MySQL类型。
数据源	选择配置好的MySQL数据源。
目标类型	此处选择Elasticsearch类型。
数据源	选择配置好的Elasticsearch数据源。

5. 单击确认，根据自身情况进行配置。

```
{
"configuration": {
"setting": {
  "speed": {
    "concurrent": "1", //作业并发数。
    "mbps": "1" //作业速率上限。
  }
},
"reader": {
  "parameter": {
    "connection": [
```

```
        {
          "table": [
            "`es_table`" //源端表名。
          ],
          "datasource": "px_mysql_OK" //数据源名，建议和添加的数据源名保持一
致。
        }
      ],
      "column": [ //源端表的列名。
        "col_ip",
        "col_double",
        "col_long",
        "col_integer",
        "col_keyword",
        "col_text",
        "col_geo_point",
        "col_date"
      ],
      "where": "", //过滤条件。
    },
    "plugin": "mysql"
},
"writer": {
    "parameter": {
      "cleanup": true, //是否在每次导入数据到Elasticsearch时清空原有数据，全
量导入/重建索引的时候需要设置为true，同步增量的时候必须为false。此处为同步增
量，需要设置为false。
      "accessKey": "nimda", //如果使用了X-PACK插件，需要填写password；如
果未使用，则填空字符串即可。阿里云Elasticsearch使用了X-PACK插件，需要填写
password。
      "index": "datax_test", // Elasticsearch的索引名称，如果之前没有，插件
会自动创建。
      "alias": "test-1-alias", //数据导入完成后写入别名。
      "settings": {
        "index": {
          "number_of_replicas": 0,
          "number_of_shards": 1
        }
      },
      "batchSize": 1000, //每次批量数据的条数。
      "accessId": "default", //如果使用了X-PACK插件，需要填写username；如
果未使用，则填空字符串即可。阿里云Elasticsearch使用了X-PACK插件，需要填写
username。
      "endpoint": "http://xxx.xxxx.xxx:xxxx", //Elasticsearch的连接地
址，可以在控制台查看。
      "splitter": ",", //如果插入数据是array，则使用指定分隔符。
      "indexType": "default", //Elasticsearch中相应索引下的类型名称。
      "aliasMode": "append", //数据导入完成后增加别名的模式，append（增加模
式），exclusive（只留这一个）。
      "column": [ //Elasticsearch中的列名，顺序和Reader中的Column顺序一致。
        {
          "name": "col_ip",//对应于TableStore中的属性列：name。
          "type": "ip"//文本类型，采用默认分词。
        },
        {
          "name": "col_double",
          "type": "string"
        },
        {
          "name": "col_long",
          "type": "long"
        },
        {
          "name": "col_integer",
```

```
          "type": "integer"
        },
        {
          "name": "col_keyword",
          "type": "keyword"
        },
        {
          "name": "col_text",
          "type": "text"
        },
        {
          "name": "col_geo_point",
          "type": "geo_point"
        },
        {
          "name": "col_date",
          "type": "date"
        }
      ],
      "discovery": false//是否自动发现，设置为true。
    },
    "plugin": "elasticsearch"//Writer插件的名称：ElasticsearchWriter，不需要修改。
}
},
"type": "job",
"version": "1.0"
}
```

6. 配置完成后，单击保存并运行。

说明：

- Elasticsearch仅支持以脚本模式导入数据。
- 如果想选择新模板，可以单击工具栏中的导入模板。一旦导入新模板，原有内容将会被全部覆盖。
- 同步任务保存后，直接单击运行，任务会立刻运行。您也可以单击提交，提交同步任务至调度系统中，调度系统会按照配置属性在从第二天开始自动定时执行。

参考文档

其它类型的数据源配置同步任务的详情，请参见下述文档：

- 配置*Reader*插件。
- 配置*Writer*插件。

1.11.6 日志服务（Loghub）通过数据集成投递数据

本文将以LogHub数据同步至MaxCompute为例，为您介绍如何通过数据集成功能同步LogHub数据至数据集成已支持的目的端数据源（如MaxCompute、OSS、OTS、RDBMS和DataHub等）。

> **说明:**
> **此功能已在华北2、华东2、华南1、中国（香港）、美西1、亚太东南1、欧洲中部1、亚太东南2、亚太东南3、亚太东北1、亚太南部1等多个地域发布。**

支持场景

- **支持跨地域的LogHub与MaxCompute等数据源的数据同步。**
- **支持不同阿里云账号下的LogHub与MaxCompute等数据源间的数据同步。**
- **支持同一阿里云账号下的LogHub与MaxCompute等数据源间的数据同步。**
- **支持公共云与金融云账号下的LogHub与MaxCompute等数据源间的数据同步。**

跨阿里云账号的特别说明

以B账号进入数据集成配置同步任务，将A账号的LogHub数据同步至B账号的MaxCompute为例。

1. **用A账号的AccessId和Accesskey创建LogHub数据源。**

   此时B账号可以拖A账号下所有SLS Project的数据。

2. **用A账号下子账号A1的AccessId和Accesskey创建LogHub数据源。**

   - **A给A1赋权日志服务的通用权限，即`AliyunLogFullAccess`和`AliyunLogReadOnlyAccess`，详情请参见授权*RAM*子用户访问日志服务资源。**
   - **A给A1赋权日志服务的自定义权限。**

     主账号A进入RAM控制台 > 策略管理页面，选择自定义授权策略 > 新建授权 > 空白模板。

     相关的授权请参见访问控制*RAM*和*RAM*子用户访问。

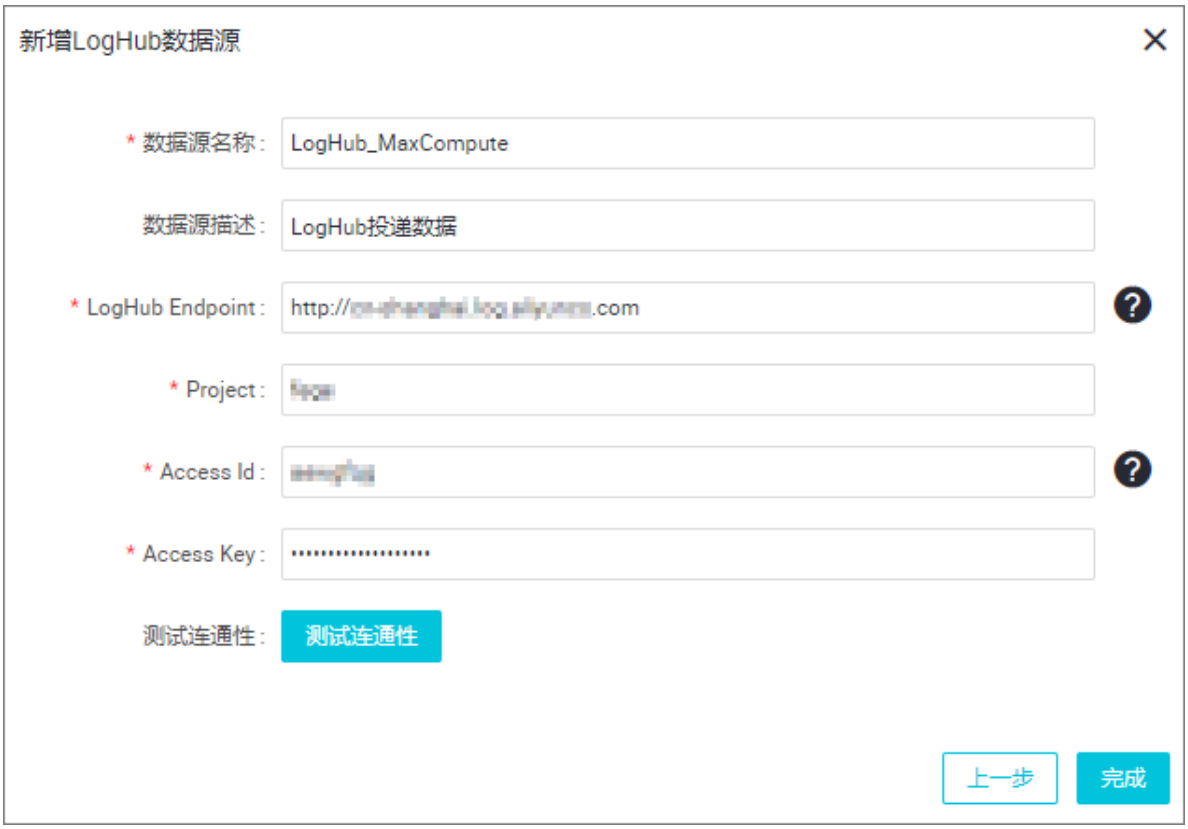
     根据下述策略进行授权后，B账号通过子账号A1只能同步日志服务project_name1以及project_name2的数据。

```
{
"Version": "1",
"Statement": [
{
"Action": [
"log:Get*",
"log:List*",
"log:CreateConsumerGroup",
"log:UpdateConsumerGroup",
```

```
"log:DeleteConsumerGroup",
"log:ListConsumerGroup",
"log:ConsumerGroupUpdateCheckPoint",
"log:ConsumerGroupHeartBeat",
"log:GetConsumerGroupCheckPoint"
],
"Resource": [
"acs:log:*:*:project/project_name1",
"acs:log:*:*:project/project_name1/*",
"acs:log:*:*:project/project_name2",
"acs:log:*:*:project/project_name2/*"
],
"Effect": "Allow"
}
]
}
```

新增数据源

1. **B账号或B的子账号以开发者身份登录*DataWorks*控制台，单击对应项目下的进入数据集成。**
2. **进入同步资源管理 > 数据源页面，单击右上角的新增数据源。**
3. **选择数据源类型为LogHub，填写新增LogHub数据源对话框中的配置。**



配置	说明
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。
LogHub Endpoint	LogHub的Endpoint，格式为`http://yyy.com`。

配置	说明
Project	详情请参见服务入口。
Access Id/Access Key	即访问密钥，相当于登录密码。您可以填写主账号或子账号的Access Id和Access Key。

4. 单击测试连通性。
5. 测试连通性通过后，单击确定。

通过向导模式配置同步任务

1. 进入数据开发 > 业务流程页面，单击左上角的新建数据同步节点。
2. 填写新建数据同步节点对话框中的配置，单击提交，进入数据同步任务配置页面。
3. 选择数据来源。

配置	说明
数据源	填写LogHub数据源的名称。
Logstore	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。
日志开始时间	数据消费的开始时间位点，为时间范围（左闭右开）的左边界，为yyyyMMddHHmmss格式的时间字符串（比如20180111013000），可以和DataWorks的调度时间参数配合使用。

配置	说明
日志结束时间	数据消费的结束时间位点，为时间范围（左闭右开）的右边界，为yyyyMMddHHmmss格式的时间字符串（比如20180111013010），可以和DataWorks的调度时间参数配合使用。
批量条数	一次读取的数据条数，默认为256。

数据预览默认收起，您可以单击进行预览。

> **说明:**
> 数据预览是选择LogHub中的几条数据展现在预览框，可能您同步的数据会跟您的预览的结果不一样，因为您同步的数据会指定开始时间和结束时间。

4. 选择数据去向。

选择MaxCompute数据源及目标表。

配置	说明
数据源	填写配置的数据源名称。
表	选择需要同步的表。
分区信息	此处需同步的表是非分区表，所以无分区信息。

配置	说明
清理规则	· 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。 · 写入前保留已有数据：导数据之前不清理任何数据，每次运行数据都是追加进去的，相当于insert into。
压缩	默认选择不压缩。
空字符串作为null	默认选择否。

5. 字段映射。

   选择字段的映射关系。需对字段映射关系进行配置，左侧源头表字段和右侧目标表字段为一一对应的关系。

6. 通道控制。

   配置作业速率上限和脏数据检查规则。

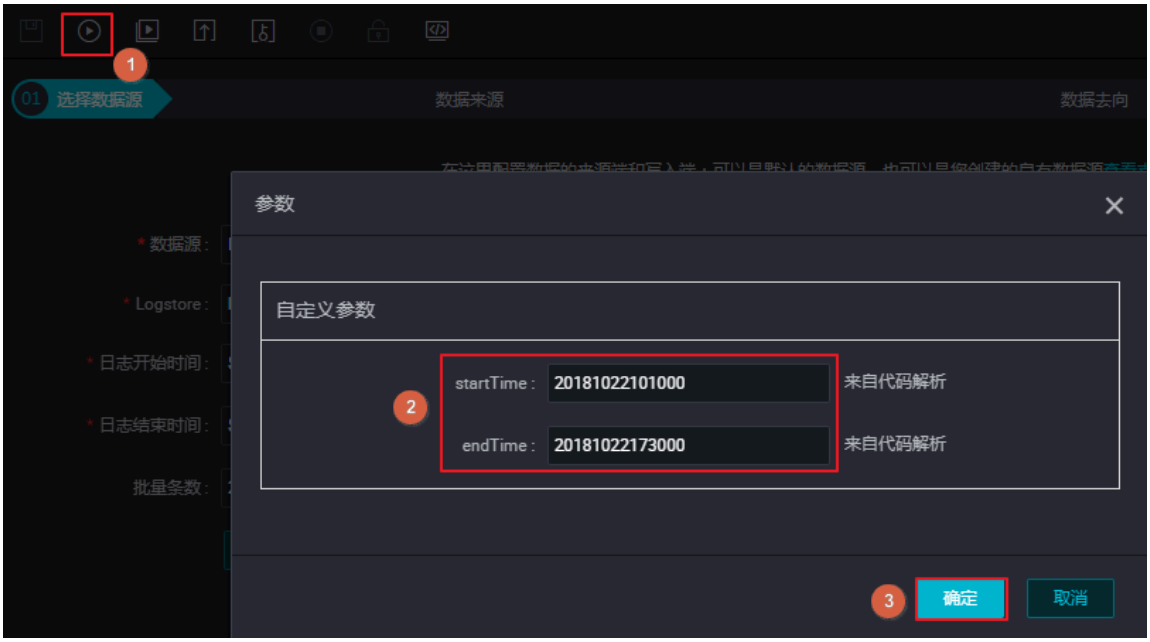
配置	说明
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

7. 运行任务。

您可以通过以下两种方式运行任务。

- 直接运行（一次性运行）

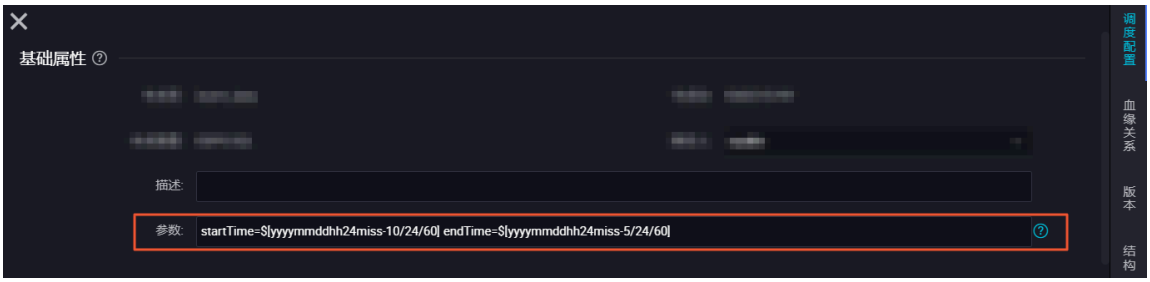
  单击任务上方的运行按钮，将直接在数据集成页面运行任务，运行之前需要配置自定义参数的具体数值。



如上图所示，代表同步10:10到17:30这段时间的LogHub记录到MaxCompute。

- 调度运行

  单击提交按钮，将同步任务提交到调度系统中，调度系统会按照配置属性在从第二天开始自动定时执行。

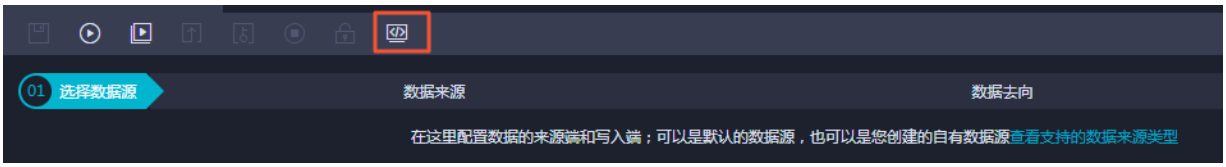


如上图所示，设置开始时间和结束时间：startTime=$[yyyymmddhh24miss-10/24/60]系统前10分钟到 endTime=$[yyyymmddhh24miss-5/24/60]系统前5分钟时间。

如上图所示，设置按分钟调度，从00:00~23:59每5分钟调度一次。

通过脚本模式配置同步任务

如果您需要通过脚本模式配置此任务，单击工具栏中的转换脚本，选择确认即可进入脚本模式。



您可以根据自身进行配置，示例脚本如下。

```
{
"type": "job",
"version": "1.0",
"configuration": {
"reader": {
"plugin": "loghub",
"parameter": {
"datasource": "loghub_lzz",//数据源名，需要和您添加的数据源名一致。
"logstore": "logstore-ut2",//目标日志库的名字，LogStore是日志服务中日志数据的采集、存储和查询单元。
"beginDateTime": "${startTime}",//数据消费的开始时间位点，为时间范围（左闭右开）的左边界。
"endDateTime": "${endTime}",//数据消费的开始时间位点，为时间范围（左闭右开）的右边界。
"batchSize": 256,//一次读取的数据条数，默认为256。
"splitPk": "",
"column": [
"key1",
```

```
"key2",
"key3"
]
}
},
"writer": {
"plugin": "odps",
"parameter": {
"datasource": "odps_first",//数据源名，需要和您添加的数据源名一致。
"table": "ok",//目标表名。
"truncate": true,
"partition": "",//分区信息。
"column": [//目标列名。
"key1",
"key2",
"key3"
]
}
},
"setting": {
"speed": {
"mbps": 8,//作业速率上限。
"concurrent": 7//并发数。
}
}
}
```

1.11.7 DataHub通过数据集成批量导入数据

本文将为您介绍如何通过数据集成对离线DataHub进行数据的导入操作。

数据集成是阿里巴巴集团提供的数据同步平台。该平台具备可跨异构数据存储系统、可靠、安全、低成本、可弹性扩展等特点，可以为20多种数据源提供不同网络环境下的离线（全量/增量）数据进出通道。数据源类型的详情请参见 支持的数据源。

准备工作

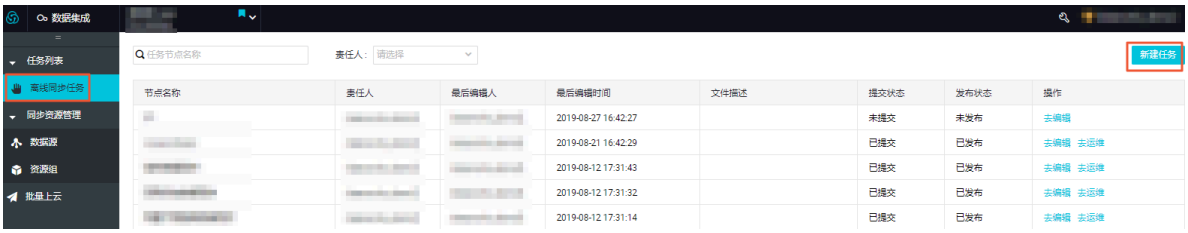
1. **准备阿里云账号，并创建账号的访问密钥，即AccessID和AccessKey。详情请参见**#unique_11。
2. **开通MaxCompute，自动产生一个默认的MaxCompute数据源，并使用主账号登录DataWorks。**
3. **创建工作空间，您可以在工作空间中协作完成业务流程，共同维护数据和任务等，因此使用DataWorks之前需要先创建一个工作空间。详情请参见**#unique_149。

> **说明:**
> **如果您想通过子账号创建数据集成任务，可以赋予其相应的权限。详情请参见**#unique_150**和**#unique_151。

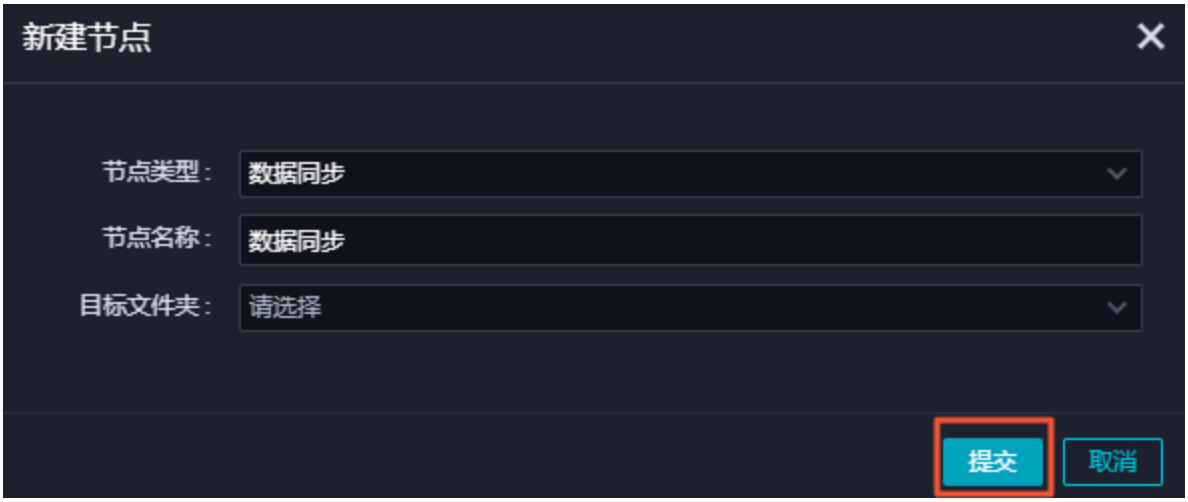
操作步骤

以Stream同步数据至DataHub的脚本模式为例，操作如下。

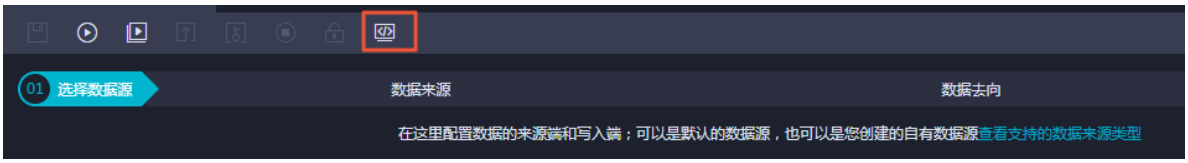
1. 以开发者身份登录*DataWorks*控制台，单击对应工作空间后的进入数据集成。
2. 进入任务列表 > 离线同步任务页面，单击右上角的新建任务。



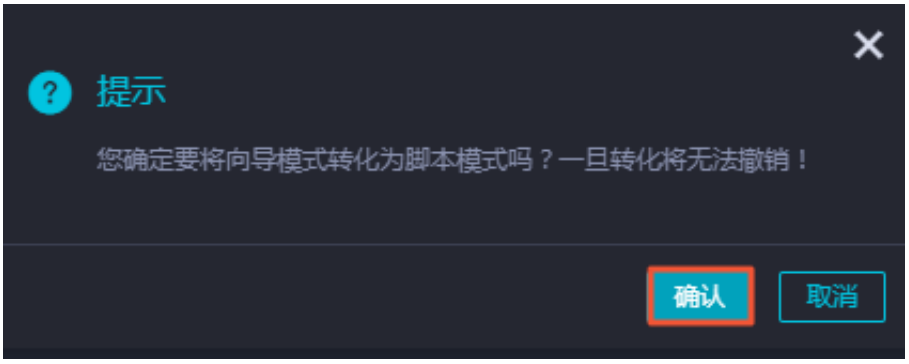
3. 在新建节点对话框中，填写节点名称并选择目标文件夹，单击提交。



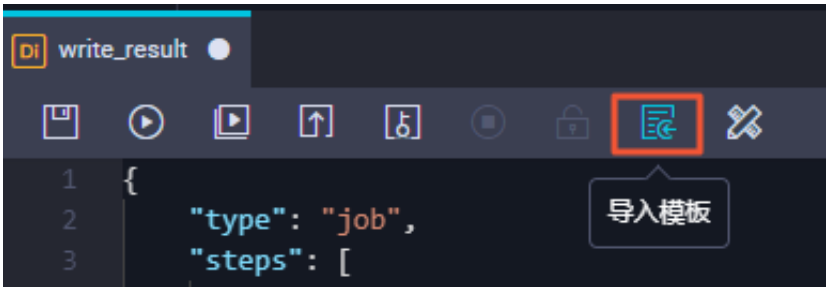
4. 成功创建数据同步节点后，单击工具栏中的转换脚本。



5. 单击提示对话框中的确认，即可进入脚本模式进行开发。



6. 单击工具栏中的导入模板。



7. 在导入模板对话框中，选择来源类型、数据源、目标类型及数据源。

配置	说明
来源类型	此处选择Stream类型。
目标类型	此处选择DataHub类型。
数据源	选择配置好的数据源。 **说明:** 如果没有提前配置数据源，可以单击新增数据源进行新增操作。

8. 单击确认生成初始脚本，您可以根据自身需求进行配置。

```
{
"type": "job",
"version": "1.0",
"configuration": {
 "setting": {
   "errorLimit": {
     "record": "0"
   },
   "speed": {
     "mbps": "1",
```

```
      "concurrent": 1,//作业并发数。
      "throttle": false
    }
  },
  "reader": {
    "plugin": "stream",
    "parameter": {
      "column": [//源端列名。
        {
          "value": "field",//列属性。
          "type": "string"
        },
        {
          "value": true,
          "type": "bool"
        },
        {
          "value": "byte string",
          "type": "bytes"
        }
      ],
      "sliceRecordCount": "100000"
    }
  },
  "writer": {
    "plugin": "datahub",
    "parameter": {
      "datasource": "datahub",//数据源名。
      "topic": "xxxx",//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。
      "mode": "random",//随机写入。
      "shardId": "0",//Shard 表示对一个Topic进行数据传输的并发通道，每个Shard会有对应的ID。
      "maxCommitSize": 524288,//为了提高写出效率，待攒数据大小达到maxCommitSize大小（单位MB）时，批量提交到目的端。默认是1,048,576，即1MB数据。
      "maxRetryCount": 500
    }
  }
}
}
```

9. 单击保存并运行。

> **说明:**
>
> - DataHub仅支持以脚本模式导入数据。
> - 如果需要选择新模板，可以单击工具栏中的导入模板，导入新模板后，会覆盖原有模板的所有内容。
> - 保存同步任务后，直接单击运行，任务会立刻运行。
>
>   您也可以单击提交，将同步任务提交到调度系统中。调度系统会按照配置属性，从第2天开始自动定时执行。

参考文档

其它数据源的配置同步任务详情，请参见下述文档：

- 配置*Reader*插件。
- 配置*Writer*插件。

1.11.8 OTSStream配置同步任务

OTSStream插件主要用于导出Table Store增量数据。增量数据可以看作操作日志，除数据本身外还附有操作信息。

OTSStream插件与全量导出插件不同，增量导出插件仅支持多版本模式，且不支持指定列，详情请参见配置*OTSStream Reader*。

说明:

OTSStream配置同步任务时，请注意以下问题：

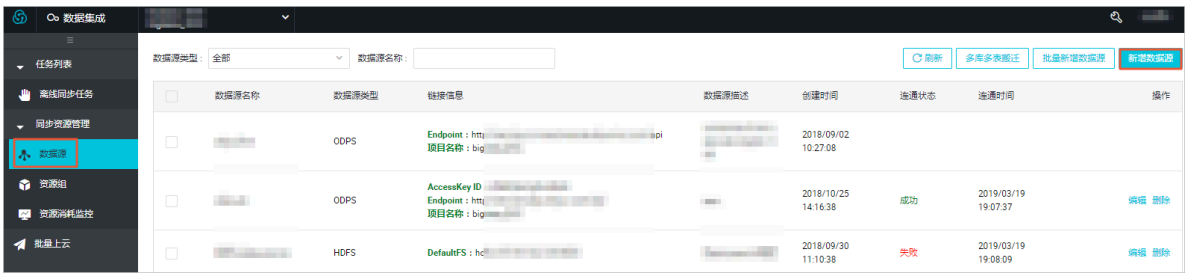
- **当前时间的前5分钟之前和24小时之内是可读数据。**
- **设置的结束时间不能超过系统显示的时间，即您设置的结束时间要比运行时间早5分钟。**
- **配置日调度会出现数据丢失的情况。**
- **不可以配置周期调度和月调度。**

示例如下：

开始时间和结束时间要包含操作Table Store表的时间，例如20171019162000您向Table Store插入2条数据，则开始时间设置为`20171019161000`，结束时间设置为`20171019162600`。

新增数据源

1. **以项目管理员身份登录*DataWorks*控制台，单击对应工作空间后的进入数据集成。**
2. **进入同步资源管理 > 数据源页面，单击右上角的新增数据源。**



3. **在新增数据源弹出框中，选择数据源类型为Table Store（OTS）。**

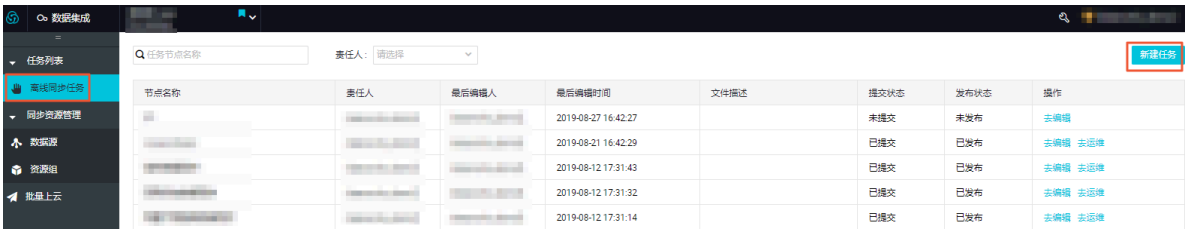
4. 填写Table Store（OTS）数据源的各配置项。

配置	说明
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此配置。
Endpoint	Table Store服务对应的Endpoint。
Table Store实例ID	Table Store服务对应的实例ID。
AccessID/AceessKey	访问密钥（AccessKeyID和AccessKeySecret），相当于登录密码。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

通过向导模式配置同步任务

1. **进入任务列表 > 离线同步任务页面，单击右上角的新建任务。**



2. **在新建节点对话框中，填写节点名称并选择目标文件夹，单击提交。**
3. **进入数据同步节点配置页面，选择数据来源。**

配置	说明
数据源	填写数据源的名称。
表	导出增量数据的表的名称。该表需要开启Stream，您可以在建表时开启，或使用UpdateTable接口开启。
开始时间	增量数据的时间范围（左闭右开）的左边界，格式为yyyymmddhh24miss，单位毫秒。
结束时间	增量数据的时间范围（左闭右开）的右边界，格式为yyyymmddhh24miss，单位毫秒。
状态表	用于记录状态的表的名称。

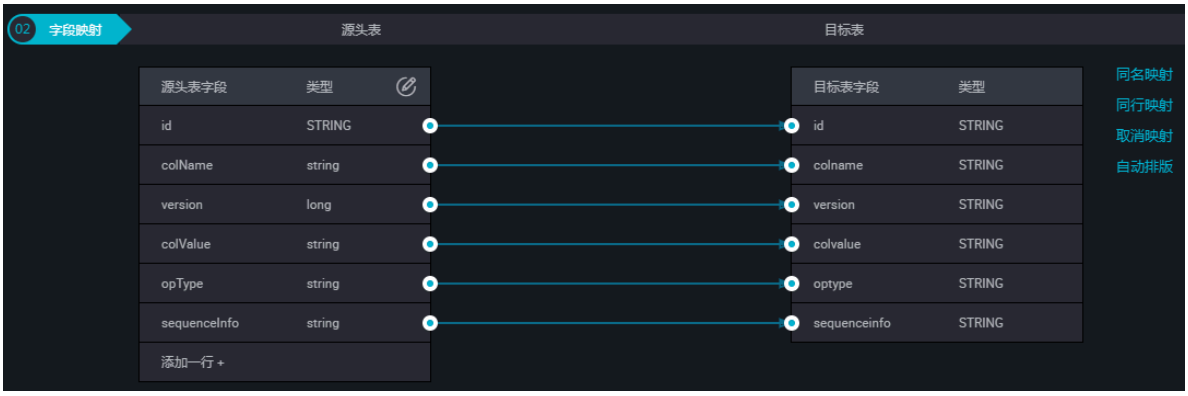
配置	说明
最大重试次数	从TableStore中读增量数据时，每次请求的最大重试次数，默认是30。
导出时序信息	是否导出时序信息，时序信息包含了数据的写入时间等。

4. 选择数据去向。

配置	说明
数据源	填写配置的数据源名称。
表	选择需要同步的表。
分区信息	此处需同步的表是非分区表，所以无分区信息。
清理规则	· 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`。 · 写入前保留已有数据：导数据之前，不清理任何数据，每次运行数据都是追加进去的，相当于`insert into`。
空字符串作为null	默认值为否。

5. 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。



6. 通道控制。

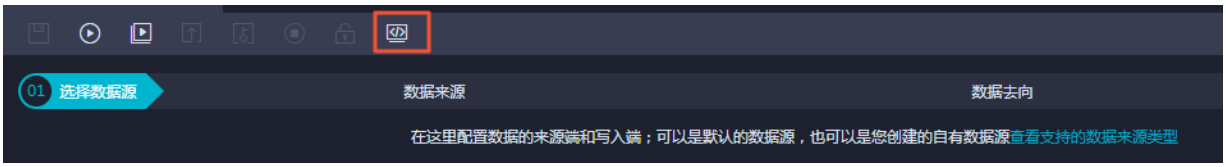
配置	说明
任务期望最大并发数	数据同步任务内，可以从源并行读取或并行写入数据存储端的最大线程数。向导模式通过界面化配置并发数，指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库，以避免抽取速度过大，给源库造成太大的压力。同步速率建议限流，结合源库的配置，请合理配置抽取速率。
错误记录数	错误记录数，表示脏数据的最大容忍条数。
任务资源组	任务运行的机器，如果任务数比较多，使用默认资源组出现等待资源的情况，建议购买独享数据集成资源或添加自定义资源组，详情请参见*#unique_109*和新增任务资源。

7. 保存并运行任务。

单击任务上方的运行按钮，将直接在数据集成页面运行任务，运行之前需要配置自定义参数的具体数值。

通过脚本模式配置同步任务

如果您需要通过脚本模式配置此任务，单击工具栏中的转换脚本，选择确认即可进入脚本模式。



您可以根据自身进行配置，示例脚本如下。

```
{
  "type": "job",
  "version": "1.0",
  "configuration": {
    "reader": {
      "plugin": "otsstream",
      "parameter": {
        "datasource": "otsstream",//数据源名，需要与您添加的数据源名称保持一致。
        "dataTable": "person",//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。
        "startTimeString": "${startTime}",//增量数据的时间范围（左闭右开）的左边界，格式为yyyymmddhh24miss，单位毫秒。
        "endTimeString": "${endTime}",//运行时间。
        "statusTable": "TableStoreStreamReaderStatusTable",//用于记录状态的表的名称。
```

```
        "maxRetries": 30,//请求的最大重试次数。
        "isExportSequenceInfo": false,
      }
    },
    "writer": {
      "plugin": "odps",
      "parameter": {
        "datasource": "odps_first",//数据源名。
        "table": "person",//目标表名。
        "truncate": true,
        "partition": "pt=${bdp.system.bizdate}",//分区信息。
        "column": [//目标列名。
          "id",
          "colname",
          "version",
          "colvalue",
          "optype",
          "sequenceinfo"
        ]
      }
    },
    "setting": {
      "speed": {
        "mbps": 7,//作业速率上限。
        "concurrent": 7//并发数。
      }
    }
  }
}
```

说明:

- 关于运行时间参数和结束时间参数，有两种表现形式（配置任务选择其中一种）。
  - `"startTimeString": "${startTime}"`

    增量数据的时间范围（左闭右开）的左边界，格式为yyyymmddhh24miss，单位毫秒。

    `"endTimeString": "${endTime}"`

    增量数据的时间范围（左闭右开）的右边界，格式为yyyymmddhh24miss，单位毫秒。
  - `"startTimestampMillis":""`

    增量数据的时间范围（左闭右开）的左边界，单位毫秒。

    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1条开始读取，并跳过写入时间小于startTimestampMillis的数据。
  - `"endTimestampMillis":" "`

    增量数据的时间范围（左闭右开）的右边界，单位毫秒。

    Reade插件startTimestampMilli位置开始导出数据后，当遇到第1条时间戳大于等于endTimestampMilli的数据时，结束导出数据，导出完成。

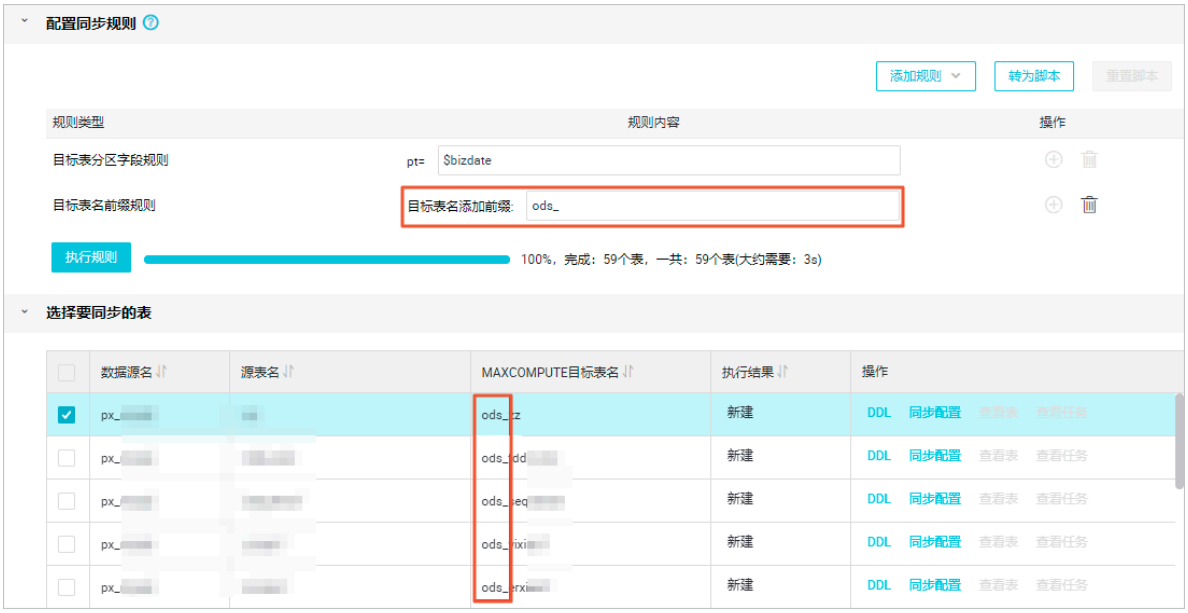
    当读取完当前全部的增量数据时，结束读取，即使未达endTimestampMillis。
- 如果配置isExportSequenceInfo项为true，如“isExportSequenceInfo”: true，则会导出时序信息，目标会多出1行，目标字段列则多1列。时序信息包含了数据的写入时间等，默认该值为false，即不导出。

1.11.9 批量上云时给目标表名加上前缀

本文将为您介绍如何在批量上云时，给目标表名加上前缀。

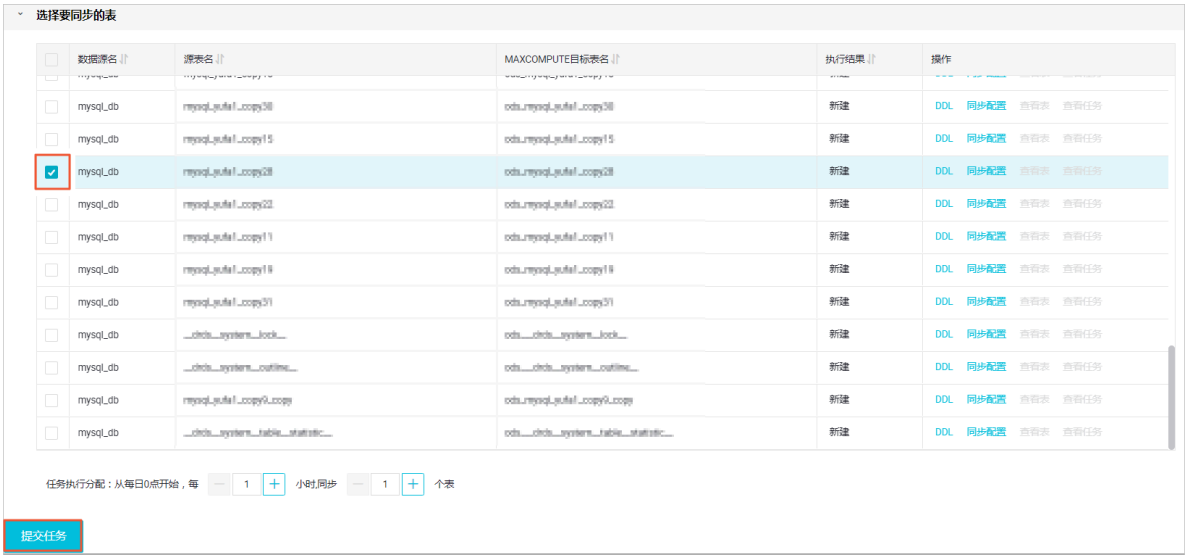
1. 请参见批量上云添加数据源。
2. 新建批量快速上云任务，并选择您创建的数据源。

3. 单击添加规则，选择表名转换规则，输入您的表名转换正则表达式。本示例中使用`(.+)`匹配所有表头，使用`ods_$1`表示给表头加上前缀`ods_`。



4. 完成设置后，单击执行规则，您即可下方选择要同步的表处看到，表名已经进行了转换。

5. 勾选您需要同步的表，单击执行任务，即可在批量上云时完成对表前缀的设置。



1.11.10 RDBMS添加关系型数据库驱动最佳实践

RDBMS Reader插件通过JDBC连接远程RDBMS数据库，并执行相应的SQL语句将数据从RDBMS库中SELECT出来。目前支持达梦、DB2、PPAS、Sybase数据库的读取。RDBMS Reader是一个通用的关系数据库读插件，您可以通过添加、注册数据库驱动等方式增加各种关系型数据库的读支持。

背景信息

*RDBMS Reader*通过JDBC连接器连接到远程的RDBMS数据库，根据您配置的信息生成查询SQL语句并发送到远程RDBMS数据库，将该SQL执行返回的结果使用DataX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。

对于您配置的Table、Column、Where等信息，RDBMS Reader将其拼接为SQL语句发送到RDBMS数据库。对于您配置的querySql信息，RDBMS直接将其发送到RDBMS数据库。

目前RDBMS Reader支持大部分通用的关系数据库类型如数字、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库类型。

准备工作

在添加关系型数据库驱动前，您需要已购买ECS服务器作为您的自定义资源组资源，建议购买规格如下：

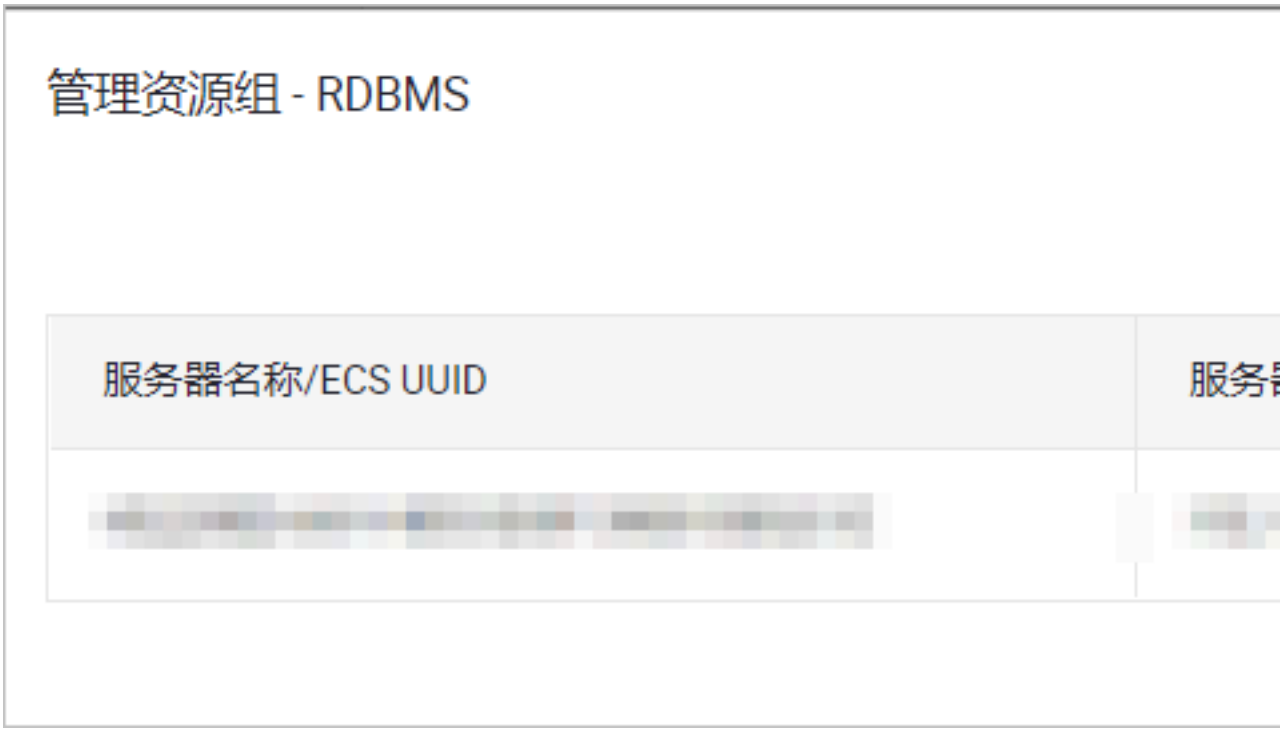
- 使用CentOS 6、CentOS 7或AliyunOS。
- 如果您添加的ECS需要执行MaxCompute任务或同步任务，需要检查当前ECS的python版本是否是Python2.6或2.7的版本（CentOS 5的Python版本为2.4，其它OS自带2.6以上版本）。
- 请确保ECS有访问公网能力，可以是否能ping通 www.aliyun.com 作为衡量标准。
- 建议ECS的配置为8核16G。

添加自定义资源组

首先您可以参考新增任务资源添加自定义资源组：

1. 创建项目后，单击对应项目的进入数据集成。
2. 选择数据集成页面里的资源组 > 新增资源组。

3. 遵照步骤提示完成Agent安装与初始化，待服务器为可用状态时，则说明自定义资源组完成。



说明:

如果刷新后还是停止状态，您可以重启alisa命令。切换到admin账号，执行下述命令:

```
/home/admin/alisatasknode/target/alisatasknode/bin/serverct1
restart
```

添加MySQL驱动

我们以添加MySQL驱动为例，说明添加关系型数据库驱动操作步骤。

1. **进入RDBMS Reader对应目录，${DATAX_HOME}为DataX主目录，即/home/admin/datax3/plugin/reader/rdbmsreader目录。**

```
[root@izbp1czjkv9fpzmsbv0qcdz rdbmsreader]# pwd
/home/admin/datax3/plugin/reader/rdbmsreader
[root@izbp1czjkv9fpzmsbv0qcdz rdbmsreader]# ls
libs plugin.json rdbmsreader-0.0.1-SNAPSHOT.jar
```

2. **在RDBMS Reader插件目录下找到plugin.json配置文件，在此文件中注册您具体的数据库驱动，如下面的"com.mysql.jdbc.Driver"，放在drivers数组中，如下所示。RDBMS 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。**

```
[root@izbp1czjkv9fpzmsbv0qcdz rdbmsreader]# vim plugin.json
{
    "name": "rdbmsreader",
    "class": "com.alibaba.datax.plugin.reader.rdbmsreader.RdbmsReade
r",
```

```
    "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.sybase.jdbc3.jdbc.
SybDriver", "com.edb.Driver","com.mysql.jdbc.Driver"]
}
```

3. **在rdbmsreader插件目录下找到libs子目录，将您下载的mysql的jar包上传上去，如下图的***mysql-connector-java-5.1.47.jar*

   ```
   [root@                    ~]# cd /home/admin/datax
   [root@                    libs]# ll
   total 19964
   -rwxr-xr-x 1 admin admin 2783513 Dec 18  2017 byte-bud
   -rwxr-xr-x 1 admin admin   31084 Dec 18  2017 byte-bud
   -rwxr-xr-x 1 admin admin  518641 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin  185140 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin  284220 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin  412739 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin   62050 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin 1599627 Dec 18  2017 commons-
   -rwxr-xr-x 1 admin admin   95324 Dec 18  2017 datax-co
   -rwxr-xr-x 1 admin admin 3528544 Dec 18  2017 db2jcc4.
   -rwxr-xr-x 1 admin admin  818729 Dec 18  2017 Dm7JdbcD
   -rwxr-xr-x 1 admin admin 1952759 Dec 18  2017 druid-1.
   -rwxr-xr-x 1 admin admin  667170 Dec 18  2017 edb-jdbc
   -rwxr-xr-x 1 admin admin  372746 Dec 18  2017 fastjson
   -rwxr-xr-x 1 admin admin  934783 Dec 18  2017 guava-r0
   -rwxr-xr-x 1 admin admin   45024 Dec 18  2017 hamcrest
   -rwxr-xr-x 1 admin admin 1467326 Dec 18  2017 hsqldb-2
   -rwxr-xr-x 1 admin admin  855824 Dec 18  2017 jackcess
   -rwxr-xr-x 1 admin admin 1006392 Dec 18  2017 jconn3-1
   -rwxr-xr-x 1 admin admin  264600 Dec 18  2017 logback-
   -rwxr-xr-x 1 admin admin  418870 Dec 18  2017 logback-
   -rwxr-xr-x 1 admin admin  533647 Dec 18  2017 mockito-
   -rw-r--r-- 1 root  root  1007502 Aug  7 14:59 mysql-co
   -rwxr-xr-x 1 admin admin   54393 Dec 18  2017 objenesi
   -rwxr-xr-x 1 admin admin   96744 Dec 18  2017 plugin-r
   -rwxr-xr-x 1 admin admin   32119 Dec 18  2017 slf4j-ap
   ```

配置RDBMS数据同步任务

目前通过*RDBMS Reader*插件只能在脚本模式中配置同步任务，配置示例如下所示：

```
{
"job": {
        "setting": {
            "speed": {
                "byte": 1048576
            },
            "errorLimit": {
                "record": 0,
                "percentage": 0.02
            }
        },
        "content": [
            {
                "reader": {
                    "name": "rdbmsreader",
                    "parameter": {
                        "username": "xxxxx",
                        "password": "yyyyyy",
                        "column": [
                            "*",
                        ],
                        "splitPk": "id",
                        "connection": [
                            {
                                "table": [
                                    "a2"
                                ],
                                "jdbcUrl": [
                                    "jdbc:mysql://xxx.mysql.yy.
aliyuncs.com:3306/xxx"  //直接
配置您的SQL地址                                    ]
                            }
                        ],

                        "where": ""
                    }
                },
                "writer": {  //writer部分根据您的
需要配置即可                    "name": "streamwriter",
                    "parameter": {
                        "print": true
                    }
                }
            }
        ]
    }
}
```

1.11.11 独享数据集成资源组最佳实践

在数据集成任务高并发执行且无法错峰运行的情况下，企业需要独享的计算资源来保障数据快速、稳定地传输，此时您可以选择独享数据集成资源。

说明:

购买的独享数据集成资源和新增的数据源必须在同地域同购买购可用区，暂不支持跨可用区：

- **目前购买的独享数据集成资源需要和您的VPC可用区进行绑定，即独享数据集成资源需要和数据源在同一个可用区。**
- **独享数据集成资源和独享调度资源绑定VPC时，选择需要访问的数据源所绑定的交换机。**
- **独享数据集成资源绑定VPC后，独享数据集成资源能够访问您的VPC对应可用区的数据源，暂时不能直接访问您的VPC其他可用区内的数据源。**
- **建议您在购买创建独享数据集成资源时，确认好可用区。**
- **独享数据集成资源无法访问阿里云经典网络。如果您的数据源是经典网络，建议使用默认资源组进行同步任务运行。**
- **目前正在开发独享数据集成资源支持一个VPC多个可用区的网络打通功能。**

购买独享数据集成资源

1. **登录DataWorks控制台，进入资源列表 > 独享资源页面。**
2. **如果您在该地域未购买过独享资源，单击右上角的新增独享资源。**

3. 单击新增独享资源对话框中订单号后的购买，即可跳转至购买页面。

4. 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、独享调度资源、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。

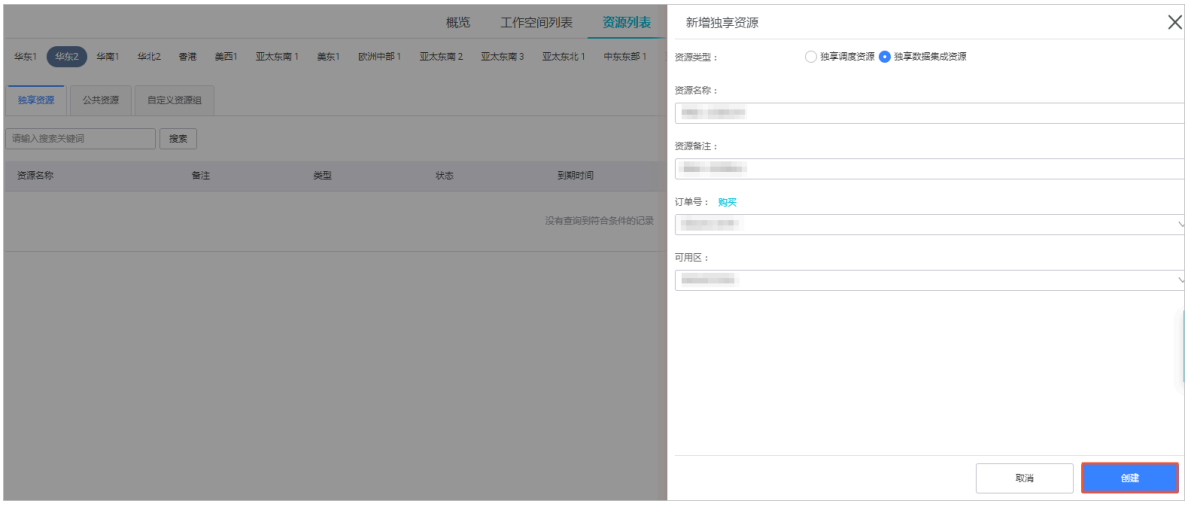
5. 确认订单信息无误后，勾选《DataWorks独享资源（包年包月）服务协议》，单击去支付。

> **说明：**
> 独享资源不支持跨地域使用，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。

新增独享数据集成资源

1. 进入资源列表 > 独享资源页面，单击右上角的新增独享资源。

2. 填写新增独享资源对话框中的配置。



配置	说明
资源类型	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，可以单击购买，跳转至售卖页进行购买。
可用区	单个地域提供了不同机器的可用区，请根据自身情况进行选择。

3. 配置完成后，单击创建，即可新增独享资源。

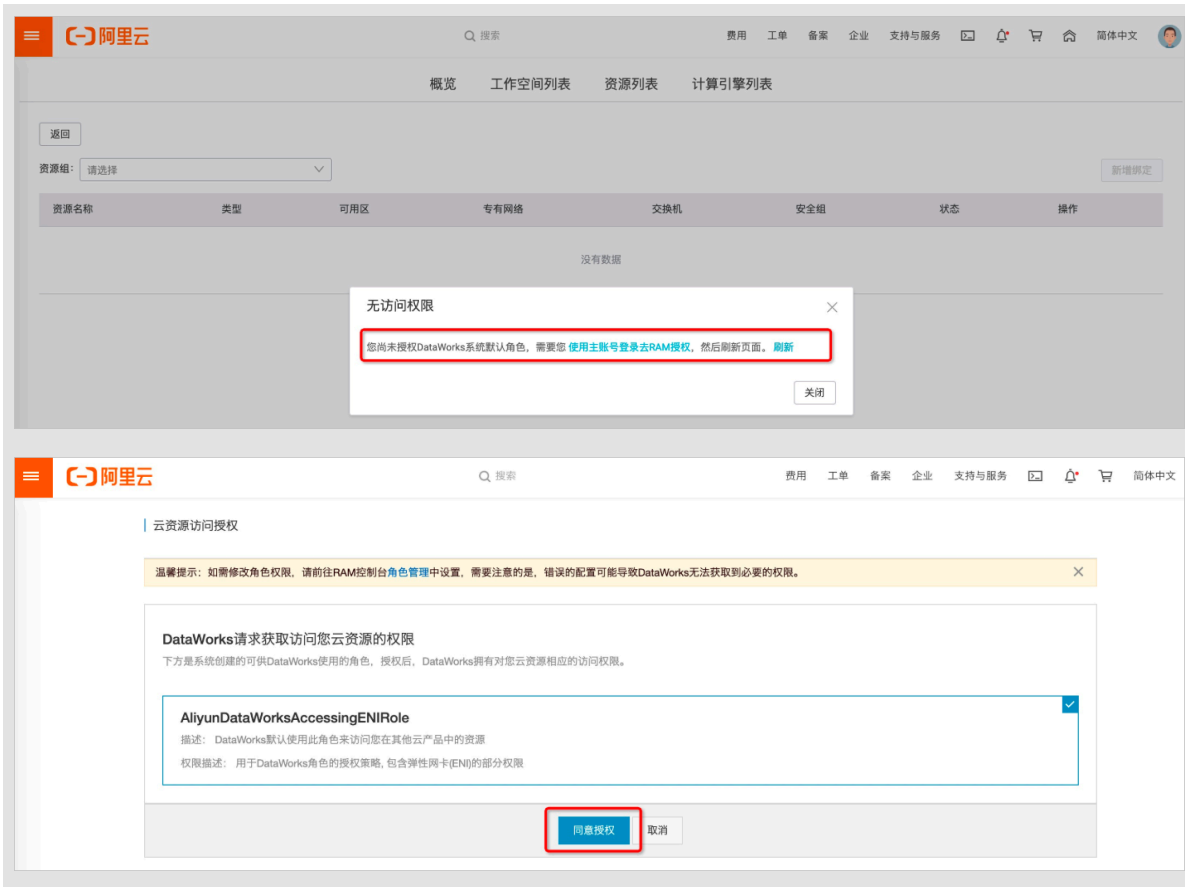
说明:
独享资源在20分钟内完成环境初始化，请耐心等待其状态更新为运行中。

专有网络绑定

独享资源部署在DataWorks托管的专有网络（VPC）中，如果需要与您自己的专有网络连通，需要进行专有网络绑定操作。

1. 单击相应资源后的专有网络绑定。

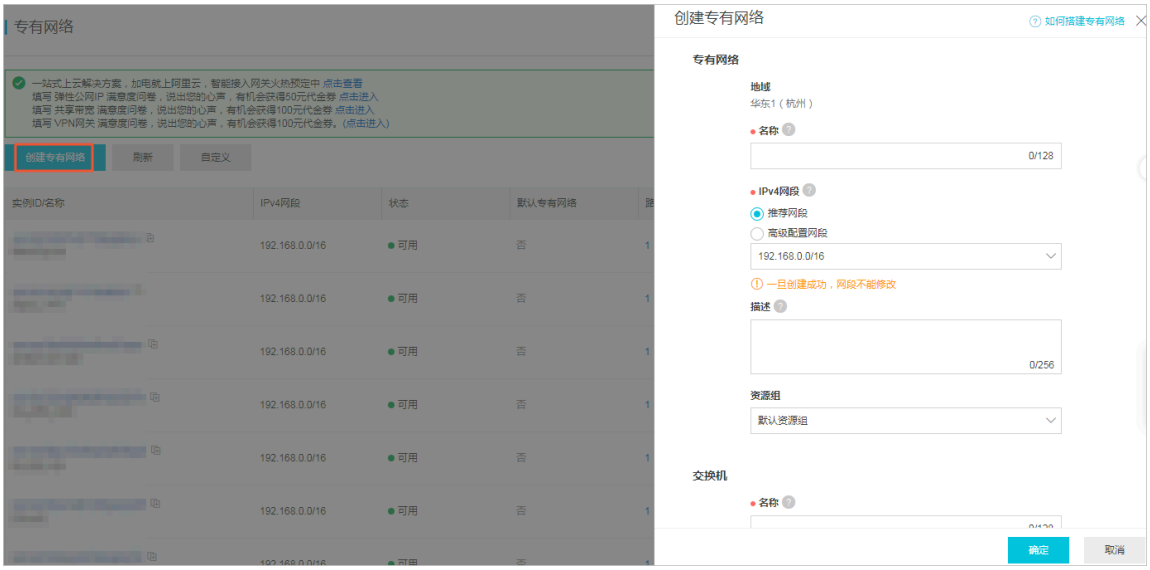
说明:
绑定前，需要进行RAM授权，让DataWorks拥有访问您的云资源的权限。


阿里云
概览
工作空间列表
资源列表
计算引擎列表
返回
资源组：
请选择
新增绑定
资源名称
类型
可用区
专有网络
交换机
安全组
状态
操作
没有数据
无访问权限
您尚未授权DataWorks系统默认角色，需要您 使用主账号登录去RAM授权，然后刷新页面。 刷新
关闭
阿里云
云资源访问授权
温馨提示：如需修改角色权限，请前往RAM控制台角色管理中设置，需要注意的是，错误的配置可能导致DataWorks无法获取到必要的权限。
DataWorks请求获取访问您云资源的权限
下方是系统创建的可供DataWorks使用的角色，授权后，DataWorks拥有对您云资源相应的访问权限。
AliyunDataWorksAccessingENIRole
描述： DataWorks默认使用此角色来访问您在其他云产品中的资源
权限描述： 用于DataWorks角色的授权策略，包含弹性网卡(ENI)的部分权限
同意授权
取消

2. 授权完成后，单击右上角的新增绑定，填写新增专有网络绑定对话框中的配置。

- 如果没有可用的专有网络，您可以单击创建专有网络，跳转至专有网络控制台中的专有网络页面进行新建。

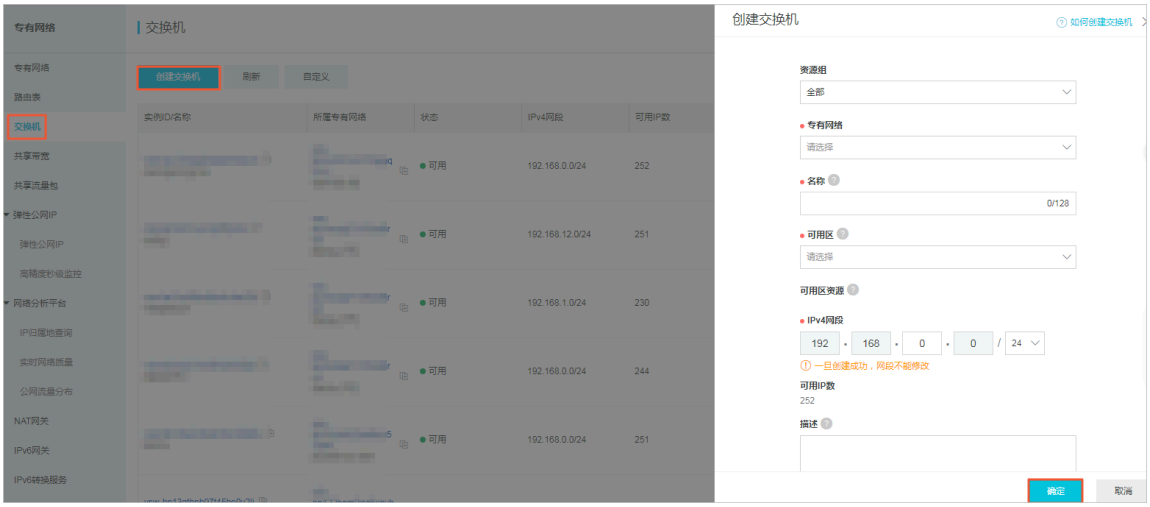
  单击创建专有网络，填写创建专有网络对话框中的配置，单击确定。



创建完成后，即可跳转至专有网络列表页面进行查看。

- **如果没有可用的交换机，您可以单击创建交换机，跳转至专有网络控制台中的交换机页面进行新建。**

  单击创建交换机，填写创建交换机对话框中的配置，单击确定。

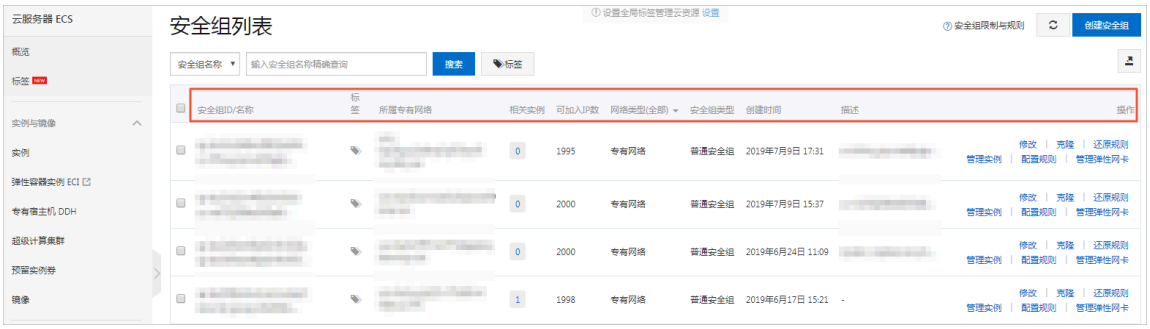


创建完成后，即可跳转至交换机列表页面进行查看。

- **如果没有可用的安全组，您可以单击创建安全组，跳转至ECS控制台中的安全组列表页面进行新建。**

  单击创建安全组，填写创建安全组对话框中的配置，单击确定。

创建完成后，即可跳转至安全组列表页面进行查看。



3. **配置完成后，单击创建。**

购买RDS实例

1. **鼠标悬浮至左上角的图标，单击云数据库RDS版，进入实例列表页面。**

2. **单击右上角的创建实例。**

3. 选择购买页面的各配置项。

说明：

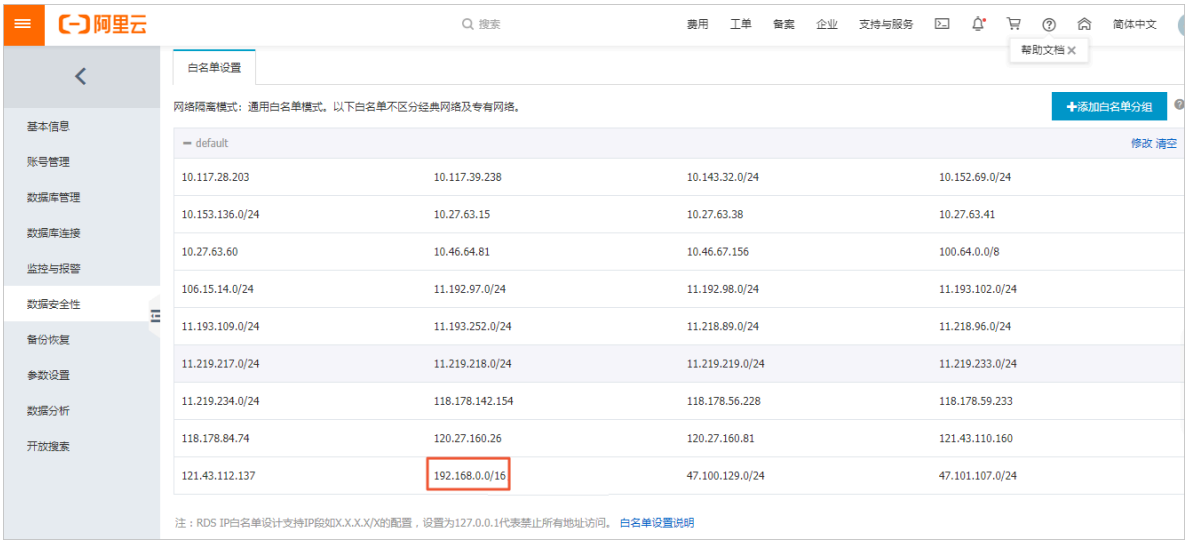
配置过程中，需特别注意版本、可用区和网络类型的选择，必须与上文的配置保持一致。

4. 配置完成后，单击立即购买。
5. 确认订单无误后，勾选《关系型数据库RDS服务条款》，单击去支付。
6. 购买完成后，即可返回实例列表页面进行查看。

设置白名单

1. 在实例列表页面，单击新建的RDS实例ID。
2. 单击左侧导航栏中的数据安全性。
3. 在白名单设置页签中，单击default白名单分组中的修改。

4. 在修改白名单分组对话框中，输入相应地域的白名单，并添加上文创建的专有网络的IP。



5. 进入账号管理和数据库管理页面，分别创建账号和数据库。

新增数据源

1. 以项目管理员身份进入*DataWorks*控制台，单击对应工作空间操作栏中的进入数据集成。
2. 单击数据源 > 新增数据源，弹出支持的数据源。
3. 在新增数据源弹出框中，选择数据源类型为MySQL。

4. 填写MySQL数据源的各配置项。

MySQL数据源类型分为阿里云数据库（RDS）、连接串模式（数据集成网络可直接连通）和连接串模式（数据集成网络不可直接连通）。

本文选择MySQL > 阿里云数据库（RDS）类型的数据源。

配置	说明
数据源类型	当前选择的数据源类型为MySQL > 阿里云数据库（RDS）。
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。

配置	说明
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此配置。
地区	选择相应的地区。
RDS实例ID	即上文创建的RDS的实例ID，您可以进入RDS控制台进行查看。
RDS实例主账号ID	您可以进入在RDS控制台安全设置页面进行查看。
用户名/密码	数据库对应的用户名和密码。

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。

5. 单击测试连通性。
6. 测试连通性通过后，单击确定。

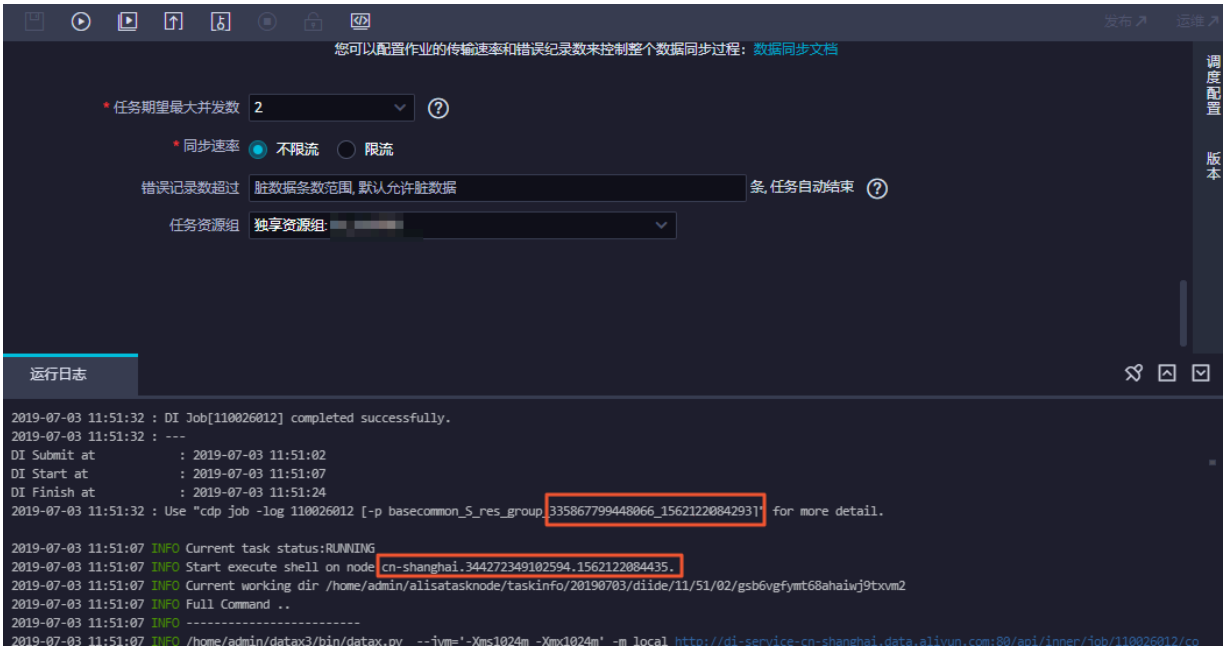
修改归属工作空间

独享资源需绑定归属的工作空间，方可被任务真正使用。一个独享资源可分配给多个工作空间使用。

1. 进入DataWorks控制台的资源列表页面。
2. 单击相应资源后的修改归属工作空间。

3. 在修改归属对话框中勾选需要的工作空间，单击确定。

绑定工作空间后，即可在数据同步任务中使用独享数据集成资源。



1.12 常见问题

1.12.1 如何排查数据集成问题

当通过数据集成实现某操作出现问题时，首先要定位问题的相关信息。例如查看运行资源、数据源信息，确认配置任务的区域等。

查看运行资源

- **任务运行在默认的资源组上，日志中会出现如下信息：**

```
running in Pipeline[basecommon_ group_xxxxxxxxx]
```

- **任务运行在数据集成自定义资源组上，日志中会出现如下信息：**

```
running in Pipeline[basecommon_xxxxxxxxx]
```

- **任务运行在独享数据集成资源上，日志中会出现如下信息：**

```
running in Pipeline[basecommon_S_res_group_xxx]
```

查看数据源信息

当出现数据集成问题时，您可以参见添加数据源典型问题场景进行排查。

需要查看的数据源的相关信息如下：

1. 确认是什么数据源之间的同步。

2. 确认是什么环境的数据源。

   阿里云数据库、连接串模式（数据集成网络可直接连通）、连接串模式（数据集成网络不可直接连通）、VPC网络环境的数据源（RDS或其它数据源）、金融云环境（VPC环境、经典网络）。

3. 确认数据源测试连通性是否成功。

   请参见配置数据源文档，确认数据源的相关信息是否正确。通常填错的情况如下：

   - 多个数据源库填错。
   - 填写的信息中加了空格或特殊字符。
   - 不支持测试连通性的问题，例如连接串模式（数据集成网络不可直接连通）的数据源、除RDS的VPC环境的数据源。

确认配置任务的区域

进入DataWorks控制台，可以查看相关的区域，例如华东2、华南1、中国（香港）、亚太东南1、欧洲中部1、亚太东南2等，通常默认是华东2。

> **说明:**
> 购买MaxCompute后，方可查看相应的区域。

界面模式报错，复制排查码

如果报错，请复制排查码，并提供给处理人员。

1.12.2 如何排查数据同步报错问题

编辑数据同步节点时报错

- 问题描述：编辑数据同步节点时，数据源访问失败，获取表结构失败，异常消息为`table XXX does not exists error with code: CDP_DATASOURCE_ERROR`。但实际上，表存在且周期运维任务正常运行。
- 解决方法：标准模式下的工作空间，包括生产环境和开发环境的数据源。请您进入数据源页面，添加开发环境的数据源，提交MaxCompute表至开发环境即可。

数据同步任务异常

- 问题描述：00:00~23:59的定时任务依赖的上游节点是08:00~23:59的定时任务。 08:00过后，同时启动了9个实例。出现异常的是9个实例的其中之一，其余的8个是正常的。

- **解决方法：同样的任务不同的实例在运行时冲突，容易导致Duplicate entry。如果该调度频繁，建议设置自依赖，以避免此类情况。**

无法连接数据源

- **问题描述：数据集成中配置Oracle数据源，测试连通性提示连接成功。但配置数据集成任务并执行时，提示无法连接数据源，错误信息如下。**

```
ErrorMessage:Code:[DBUtilErrorCode-10], Description:[连接数据库失败. 请检查您的账号、密码、数据库名称、IP、Port或者向DBA寻求帮助（注意网络环境）.].
```

  - **数据库连接失败。根据您配置的连接信息无法从`jdbc:oracle:thin:@X.X.X.X:XXXXj`中找到可以连接的JDBCUrl，请检查您的配置并进行修改。**
  - **java.lang.Exception：DataX无法连接对应的数据库。可能原因如下所示：**

    ■ 配置的`ip`、`port`、`database`或`JDBC`错误，导致无法连接。

    ■ 配置的`username`或`password`错误，导致鉴权失败。

    请和数据库管理员确认该数据库的连接信息是否正确。

```
at com.alibaba.datax.plugin.rdbms.util.DBUtil$2.call(DBUtil.java:76)
at com.alibaba.datax.plugin.rdbms.util.DBUtil$2.call(DBUtil.java:52)
at com.alibaba.datax.common.util.RetryUtil$Retry.call(RetryUtil.java
:163)
at com.alibaba.datax.common.util.RetryUtil$Retry.doRetry(RetryUtil.
java:111)
at com.alibaba.datax.common.util.RetryUtil.executeWithRetry(
RetryUtil.java:31)
at com.alibaba.datax.plugin.rdbms.util.DBUtil.chooseJdbcUrl(DBUtil.
java:52)
at com.alibaba.datax.plugin.rdbms.reader.util.OriginalConfPretreat
mentUtil.dealJdbcAndTable(OriginalConfPretreatmentUtil.java:105)
at com.alibaba.datax.plugin.rdbms.reader.util.OriginalConfPretreat
mentUtil.simplifyConf(OriginalConfPretreatmentUtil.java:69)
at com.alibaba.datax.plugin.rdbms.reader.util.OriginalConfPretreat
mentUtil.doPretreatment(OriginalConfPretreatmentUtil.java:43)
at com.alibaba.datax.plugin.rdbms.reader.CommonRdbmsReader$Job.init(
CommonRdbmsReader.java:70)
at com.alibaba.datax.plugin.reader.oraclereader.OracleReader$Job.
init(OracleReader.java:40)
at com.alibaba.datax.core.job.JobContainer.initJobReader(JobContain
er.java:1005)
at com.alibaba.datax.core.job.JobContainer.init(JobContainer.java:
434)
at com.alibaba.datax.core.job.JobContainer.start(JobContainer.java:
210)
at com.alibaba.datax.core.Engine.start(Engine.java:96)
at com.alibaba.datax.core.Engine.entry(Engine.java:246)
```

```
at com.alibaba.datax.core.Engine.main(Engine.java:279)
```

- **解决方法：从Oracle同步数据至MySQL，确认MySQL数据源是否连通公网。如果有，修改为JDBC的连接方式进行测试，底层会根据您配置的连接串的方式选择数据同步的机器。**

  如果MySQL选择实例的方式进行配置，会被分配至不支持公网的机器上，会导致Oracle的连接出现问题。

列包含关键字

- **问题描述：日志中报SQL语句执行失败（列包含关键字）。**

```
2017-05-31 14:15:20.282 [33881049-0-0-reader] ERROR ReaderRunner -
Reader runner Received Exceptions:com.alibaba.datax.common.exception
.DataXException: Code:[DBUtilErrorCode-07]
```

- **错误解读：读取数据库数据失败，请检查您配置的column、table、where、querySql，或者向数据库管理员寻求帮助。**

  执行的SQL如下所示。

```
select **index**,plaid,plarm,fget,fot,havm,coer,ines,oumes from xxx
```

  错误信息如下所示。

```
You have an error in your SQL syntax; check the manual that
corresponds to your MySQL server version for the right syntax to use
 near **index**,plaid,plarm,fget,fot,havm,coer,ines,oumes from xxx
```

- **排查思路：**

  1. **本地运行SQL语句**`select **index**,plaid,plarm,fget,fot,havm,coer,ines, oumes from xxx`**，查看其结果，通常也会有相应的报错。**
  2. **字段中有关键字index，可以通过添加单引号或修改字段解决该问题。**

表名带有双引号包单引号

- **问题描述：日志中报SQL语句执行失败（表名带有双引号包单引号）。**

```
com.alibaba.datax.common.exception.DataXException: Code:[DBUtilErro
rCode-07]
```

- **错误解读：**

  读取数据库数据失败，请检查您配置的column、table、where、querySql，或者向数据库管理员寻求帮助。

  执行的SQL如下所示。

```
select /_+read_consistency(weak) query_timeout(100000000)_/ _ from**
 'ql_ddddd_[0-31]' **where 1=2
```

  错误信息如下所示。

```
You have an error in your SQL syntax; check the manual that
corresponds to your MySQL server version for the right syntax to
use near ‘‘ql_live_speaks[0-31]’ where 1=2’ at line 1 - com.mysql.
jdbc.exceptions.jdbc4.MySQLSyntaxErrorException: You have an error
in your SQL syntax; check the manual that corresponds to your MySQL
 server version for the right syntax to use near **’‘ql_ddddd_[0-31
]’ where 1=2’ **
```

- **排查思路：配置表名时，需要双引号包单引号。通常配置常量是双引号包单引号，例如**`“table”:[“‘qlddddd[0-31]’”]`**，直接去掉其中的单引号。**

测试数据源连通性失败（Access denied for）

- **问题描述：连接数据库失败。**

  数据库连接串为`jdbc:mysql://xx.xx.xx.x:3306/t_demo`**，用户名为fn_test，异常消息为**`Access denied for user ‘fn_test’@’%’ to database ‘t_demo’`**。**

- **排查思路：**
  - **通常出现Access denied for异常，是因为填写的信息有问题，请确认您填写的信息。**
  - **确认白名单或用户的账号是否具有对应数据库的权限，RDS管控台可以添加相应的白名单和授权。**

路由策略问题

- **问题描述：路由策略有问题，运行的池子oxs和ECS集群。**

```
2017-08-08 15:58:55 : Start Job[xxxxxxx], traceId **running
in Pipeline[basecommon_group_xxx_cdp_oxs]**ErrorMessage:Code:[
DBUtilErrorCode-10]
```

- **错误解读：**

  连接数据库失败，请检查您的账号、密码、数据库名称、IP、Port或者向数据库管理员寻求帮助（注意网络环境）。

- **排查思路：数据库连接失败，因为根据您配置的连接信息，无法从jdbc:oracle:thin:@xxx.xxxxx.x.xx:xxxx:prod中找到可连接的JDBCUrl，请检查您的配置并进行修改。**

DataX无法连接对应的数据库

- **问题描述：ava.lang.Exception：DataX无法连接对应的数据库。**
- **错误解读：**

  出现该错误可能的原因如下：

  - **配置的**`ip`、`port`、`database`**或**`JDBC`**错误，无法连接。**
  - **配置的**`username`**或**`password`**错误，鉴权失败。请和数据库管理员确认该数据库的连接信息是否正确。**

- **排查思路：**

  情况一：

  - **Oracle同步的RDS-PostgreSQL直接单击运行，不能在调度中运行，因为运行的集群不同。**
  - **添加RDS的数据源时，改成添加普通JDBC形式的数据源，则Oracle同步的RDS-PostgreSQL可以在调度中运行。**

  情况二：

  - **VPC环境的RDS-PostgreSQL不能运行在自定义资源组上，因为VPC环境的RDS有反向代理功能，这样与用户自定义资源组存在网络问题。因此，通常VPC环境的RDS直接运行在DataWorks默认的资源即可。如果默认资源不能满足您的需要，要运行在自己的资源上，可以将VPC环境的RDS作为VPC环境JDBC形式的数据源，购买一个同网段的ECS。**

    详情请参见*VPC*环境数据同步配置。
  - **通常VPC环境的RDS映射出的URL为**`jdbc:mysql://100.100.70.1:4309/xxx`**，**`100`**开头的IP是后台映射出来的，如果是一个域名的表现形式则为非VPC环境。**

HBase Writer不支持DATE类型

- **问题描述：HBase Writer不支持DATE类型。**

```
HBase同步到hbase:2017-08-15 11:19:29 : State: 4(FAIL) | Total: 0R 0B
 | Speed: 0R/s 0B/s | Error: 0R 0B | Stage: 0.0% ErrorMessage:Code:[
Hbasewriter-01]
```

- **错误解读：您填写的参数值不合法。**

  Hbasewriter不支持DATE类型，目前支持的类型包括STRING、BOOLEAN、SHORT、INT、LONG、FLOAT和DOUBLE。

- **排查思路：**

  - **HBase的Writer不支持DATE类型，所以在Writer中不能配置DATE类型。**
  - **直接配置STRING类型，因为HBase没有数据类型的概念，底层通常是BYTE数组。**

JSON格式配置错误

- **错误描述：column配置错误。**

  经DataX智能分析，该任务最可能的错误原因如下。

```
com.alibaba.datax.common.exception.DataXException: Code:[Framework-
02]
```

- **错误解读：DataX引擎运行过程出错，详情请参见DataX运行结束时的错误诊断信息。**

```
java.lang.ClassCastException: com.alibaba.fastjson。JSONObject cannot
 be cast to java.lang.String
```

- **排查思路：发现其JSON配置有问题。**

```
writer端:
“column”:[
{
“name”:“busino”,
“type”:“string”
}
]
```

  正确的写法：

```
“column”:[
{
“busino”
}
```

```
]
```

JSON List编写缺少[]

- **问题描述：JSON List编写缺少[]。**

  经DataX智能分析，该任务最可能的错误原因如下所示。

  ```
  com.alibaba.datax.common.exception.DataXException: Code:[Framework-
  02]
  ```

- **错误解读：DataX引擎运行过程出错，详情请参见DataX运行结束时的错误诊断信息。**

  ```
  java.lang.String cannot be cast to java.util.List - java.lang.String
   cannot be cast to java.util.List
  at com.alibaba.datax.common.exception.DataXException.asDataXExc
  eption(DataXException.java:41)
  ```

- **排查思路：少了[]，list型变成其它的形式，找到对应的地方填上[]即可解决问题。**

权限问题

- **缺少delete权限。**
  - **问题描述：MaxCompute同步至RDS-MySQL，报错如下。**

    ```
    ErrorMessage:Code:[DBUtilErrorCode-07]
    ```

  - **错误解读：**

    读取数据库数据失败，请检查您配置的column、table、where、querySql，或者向数据库管理员寻求帮助。

    执行的SQL如下所示。

    ```
    delete from fact_xxx_d where sy_date=20170903
    ```

    具体错误信息如下所示。

    ```
    **DELETE command denied** to user ‘xxx_odps’@‘[xx.xxx.xxx.xxx](
    http://xx.xxx.xxx.xxx)’ for table ‘fact_xxx_d’ - com.mysql.jdbc.
    exceptions.jdbc4.MySQLSyntaxErrorException: DELETE command denied
     to user ‘xxx_odps’@‘[xx.xxx.xxx.xxx](http://xx.xxx.xxx.xxx)’ for
    table 'fact_xxx_d’
    ```

  - **排查思路：DELETE command denied to没有删除此表的权限，到相应的数据库设置相关表的删除权限。**

- 缺少drop权限。
  - 问题描述：读取数据库数据失败。

    ```
    Code:[DBUtilErrorCode-07]
    ```

  - 错误解读：请检查您配置的column、table、where、querySql或者向数据库管理员寻求帮助。

    执行的SQL为`truncate table be_xx_ch`

    具体错误信息如下所示。

    ```
    **DROP command denied to user** ‘xxx’@‘[xxx.xx.xxx.xxx](http://
    xxx.xx.xxx.xxx)’ for table ‘be_xx_ch’ - com.mysql.jdbc.exceptions
    .jdbc4.MySQLSyntaxErrorException: DROP command denied to user ‘xxx
    ’@‘[xxx.xx.xxx.xxx](http://xxx.xx.xxx.xxx)’ for table 'be_xx_ch’
    ```

  - 排查思路：

    MySQL Writer配置执行前准备语句truncate删除表中的数据报上面错误，是因为没有drop的权限。

- AnalyticDB for MySQL权限问题。

  ```
  2016-11-04 19:49:11.504 [job-12485292] INFO  OriginalConfPretreat
  mentUtil - Available jdbcUrl:jdbc:mysql://100.98.249.103:3306/
  AnalyticDB for MySQL_rdb?yearIsDateType=false&zeroDateTimeBehavior=
  convertToNull&tinyInt1isBit=false&rewriteBatchedStatements=true.
  2016-11-04 19:49:11.505 [job-12485292] WARN  OriginalConfPretreat
  mentUtil
  ```

  您的配置文件中的列配置存在一定的风险，因为您未配置读取数据库表的列，当您的表字段个数、类型有变动时，可能影响任务正确性甚至会运行出错。请检查您的配置并进行修改。

  ```
  2016-11-04 19:49:11.528 [job-12485292] INFO Writer$Job
  ```

  如果是MaxCompute>AnalyticDB for MySQL的数据同步，您需要完成以下两方面的授权：

  - AnalyticDB for MySQL官方账号至少需要有需要同步的表的describe和select权限，因为AnalyticDB for MySQL系统需要获取MaxCompute需要同步表的结构和数据信息。
  - 您配置的AnalyticDB for MySQL数据源访问账号密钥，需要拥有向指定的AnalyticDB for MySQL数据库发起load data的权限，您可以在AnalyticDB for MySQL系统中添加授权。

  ```
  2016-11-04 19:49:11.528 [job-12485292] INFO Writer$Job
  ```

  如果是RDS（或其它非MaxCompute数据源）>AnalyticDB for MySQL的数据同步，实现逻辑为先将数据装载至MaxCompute临时表，再从MaxCompute临时表同步

至AnalyticDB for MySQL，中转MaxCompute项目为example_project，中转项目账号为someone@example.com。您需要完成以下两方面的授权：

- AnalyticDB for MySQL官方账号需要至少具备同步的表（即MaxCompute临时表）的describe和select权限，因为AnalyticDB for MySQL系统需要获取MaxCompute需要同步的表的结构和数据信息，此部分部署时已经完成授权。
- 中转MaxCompute对应的账号someone@example.com，需要具备向指定的AnalyticDB for MySQL数据库发起load data的权限，您可以在AnalyticDB for MySQL系统中添加授权。

排查思路：

出现此问题是因为没有设置load data权限。

中转项目账号为someone@example.com，权限方面：AnalyticDB for MySQL官方账号至少需要拥有需要同步的表（即MaxCompute临时表）的describe和select权限，因为AnalyticDB for MySQL系统需要获取MaxCompute需要同步的表的结构和数据信息，此部分部署时已经完成授权，登录AnalyticDB for MySQL管控台给AnalyticDB for MySQL授予load data的权限。

白名单问题

- 没有添加白名单导致测试连通性失败。

```
error message: **Timed out after 5000** ms while waiting for a
server that matches ReadPreferenceServerSelector{readPreference=
primary}. Client view of cluster state is {type=UNKNOWN, servers=[{[
address:3717=dds-bp1afbf47fc7e8e41.mongodb.rds.aliyuncs.com](http
://address:3717=dds-bp1afbf47fc7e8e41.mongodb.rds.aliyuncs.com),
type=UNKNOWN, state=CONNECTING, exception={com.mongodb.MongoSocke
tReadException: Prematurely reached end of stream}}, {[address:3717=
dds-bp1afbf47fc7e8e42.mongodb.rds.aliyuncs.com](http://address:3717=
dds-bp1afbf47fc7e8e42.mongodb.rds.aliyuncs.com), type=UNKNOWN, state
=CONNECTING,** exception={com.mongodb.MongoSocketReadException:
Prematurely reached end of stream**}}]
```

排查思路：非VPC环境的MongoDB，添加数据源时报Timed out after 5000，白名单添加有问题。

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，请避免使用root作为访问账号。

- **白名单不全。**

```
for Code:[DBUtilErrorCode-10]
```

错误解读：连接数据库失败，请检查您的账号、密码、数据库名称、IP、Port或者向数据库管理员寻求帮助（注意网络环境）。

错误信息如下所示。

```
java.sql.SQLException: Invalid authorization specification,  message
 from server: "#**28000ip not in whitelist, client ip is xx.xx.xx.xx
".**
2017-10-17 11:03:00.673 [job-xxxx] ERROR RetryUtil - Exception when
calling callable
```

排查思路：未添加用户自己的资源至白名单内。

数据源信息填写错误

- **脚本模式配置缺少相应数据源信息（could not be blank）。**

```
2017-09-06 12:47:05 [INFO] Success to fetch meta data for table with
 **projectId [43501]** **项目ID **and instanceId **[mongodb]数据源
名.**
2017-09-06 12:47:05 [INFO] Data transport tunnel is CDP.
2017-09-06 12:47:05 [INFO] Begin to fetch alisa account info for
3DES encrypt with parameter account: [zz_683cdbcefba143b7b709067b362
d4385].
2017-09-06 12:47:05 [INFO] Begin to fetch alisa account info for
3DES encrypt with parameter account: [zz_683cdbcefba143b7b709067b362
d4385].
[Error] Exception when running task, message:** Configuration
property [accessId]通常是odps数据源要填写的信息 could not be blank!**
```

排查思路：报错显示没有相应的accessId信息，通常出现这种现象是脚本模式，查看用户配置的JSON代码，是否忘记写相应的数据源名。

- **数据源配置错误或未配置数据源。**

```
2017-10-10 10:30:08 INFO
 ==================================================
File “/home/admin/synccenter/src/Validate.py”, line 16, in notNone
raise Exception(“Configuration property [%s] could not be blank!” %
 (context))
**Exception: Configuration property [username] could not be blank!**
```

排查思路：

- **根据正常打出的日志进行对比。**

```
[56810] and instanceId(instanceName) [spfee_test_mysql]…
```

```
2017-10-09 21:09:44 [INFO] Success to fetch meta data for table
with projectId [56810] and instanceId [spfee_test_mysql].
```

- 由rds-mysql显示的信息可见，调数据源失败且报用户为空，说明数据源的位置配置错误或未配置数据源。

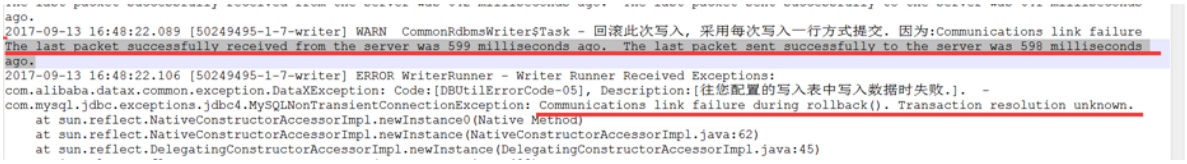
- DRDS连接数据超时。

  MaxCompute同步数据到DRDS，经常出现下述错误。

  ```
  [2017-09-11 16:17:01.729 [49892464-0-0-writer] WARN CommonRdbm
  sWriter$Task
  ```

  回滚此次写入，采用每次写入一行方式提交，原因如下。

  ```
  com.mysql.jdbc.exceptions.jdbc4.CommunicationsException: **
  Communications link failure  **
  The last packet successfully received from the server was 529
  milliseconds ago.  The last packet sent successfully to the server
  was** 528 milliseconds ago**.
  ```

  排查思路：

  DataX客户端的超时。您可以在添加DRDS数据源时加上`?useUnicode=true&characterEncoding=utf-8&socketTimeout=3600000`超时参数。

  示例如下。

  ```
  jdbc:mysql://10.183.80.46:3307/ae_coupon?useUnicode=true&characterE
  ncoding=utf-8&socketTimeout=3600000
  ```

- 系统内部错误。

  排查思路：通常是在开发环境改错了JSON格式并直接保存，导致报错位系统内部问题。界面显示为空白，碰到该问题，直接提供您的工作空间名称和节点名称进行咨询。

脏数据

- 脏数据（String[“”]不能转为Long）。

  ```
  2017-09-21 16:25:46.125 [51659198-0-26-writer] ERROR WriterRunner -
  Writer Runner Received Exceptions:
  ```

```
com.alibaba.datax.common.exception.DataXException: Code:[Common-01]
```

错误解读：同步数据出现业务脏数据情况，数据类型转换错误。String[“”]不能转为Long。

排查思路：String[“”]不能转为Long：两张表格中的建表语句一致，报上述错误是因为字段类型中的空字段不能转换成Long类型，直接配置为String类型。

- **脏数据（Out of range value）。**

```
2017-11-07 13:58:33.897 [503-0-0-writer] ERROR StdoutPluginCollector

脏数据:
{“exception”:“Data truncation: Out of range value for column ‘id’ at
 row 1”,“record”:[{“byteSize”:2,“index”:0,“rawData”:-3,“type”:“LONG
”},{“byteSize”:2,“index”:1,“rawData”:-2,“type”:“LONG”},{“byteSize”:2
,“index”:2,“rawData”:“其他”,“type”:“STRING”},{“byteSize”:2,“index”:
3,“rawData”:“其他”,“type”:“STRING”}],“type”:“writer”}
```

排查思路：mysql2mysql，源端设置的是smallint(5)，目标端是int(11) unsigned，因为smallint(5)范围有负数，unsigned不允许有负数，所以产生脏数据。

- 脏数据（存储emoji）。

  数据表配置成了可以存储emoji的，同步时报脏数据。

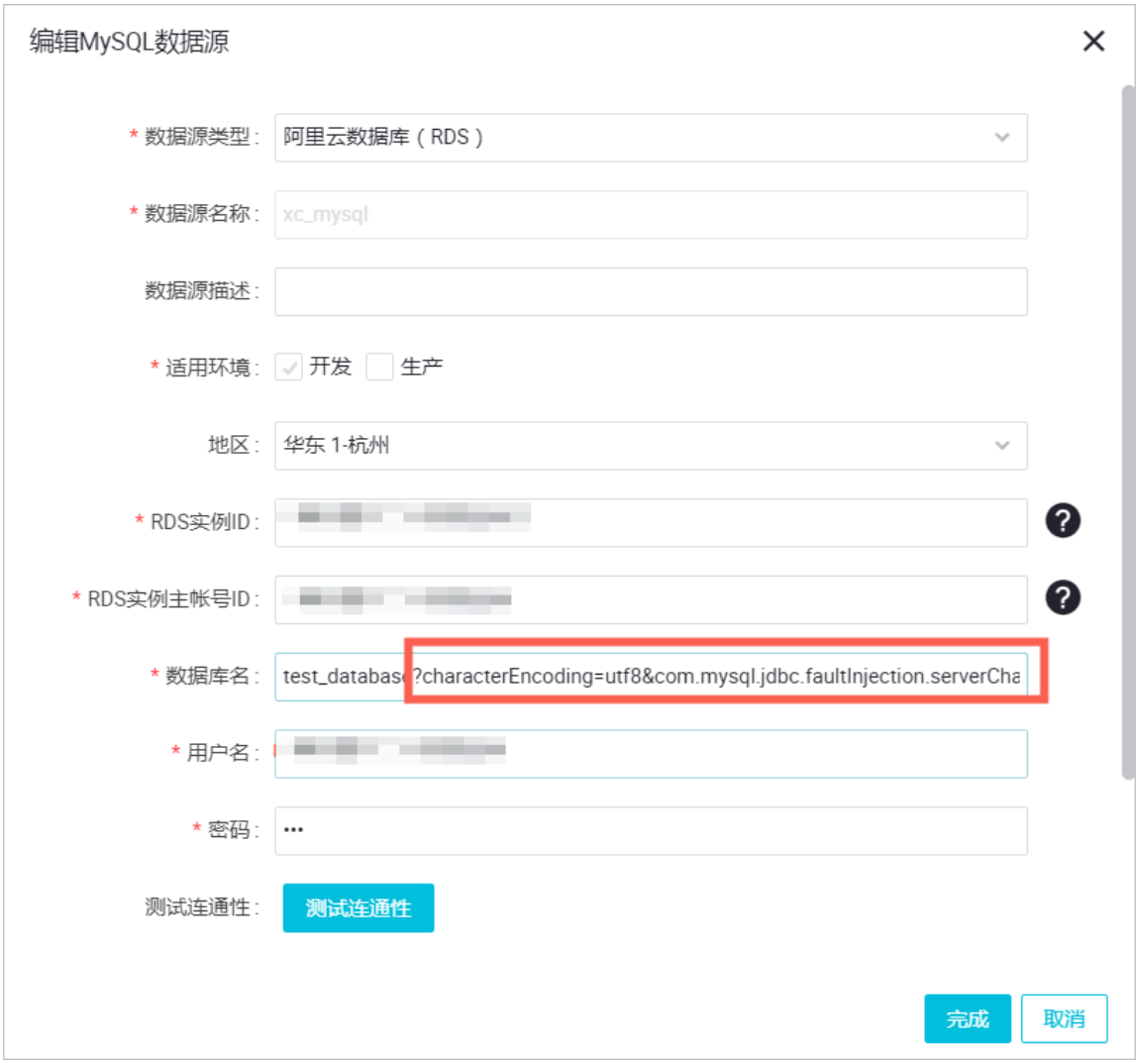
  排查思路：同步emoji时报错脏数据，需要修改编码格式：

  - JDBC形式添加数据源

    ```
    jdbc:mysql://xxx.x.x.x:3306/database?characterEncoding=utf8&com.
    mysql.jdbc.faultInjection.serverCharsetIndex=45
    ```

  - 实例ID形式添加数据源

    在数据库名后拼接`?characterEncoding=utf8&com.mysql.jdbc.faultInjection.serverCharsetIndex=45`。



- 空字段引起的脏数据。

```
{“exception”:“Column ‘xxx_id’ cannot be null”,“record”:[{“byteSize
”:0,“index”:0,“type”:“LONG”},{“byteSize”:8,“index”:1,“rawData”:-1,“
type”:“LONG”},{“byteSize”:8,“index”:2,“rawData”:641,“type”:“LONG”}
```

经DataX智能分析，该任务最可能的错误原因如下所示。

```
com.alibaba.datax.common.exception.DataXException: Code:[Framework-
14]
```

错误解读：DataX传输脏数据超过用户预期，该错误通常是由于源端数据存在较多业务脏数据导致。请仔细检查DataX汇报的脏数据日志信息，或者您可以适当调大脏数据阈值。

脏数据条数检查不通过，限制是1条，但实际上捕获了7条。

排查思路：设置Column ‘xxx_id’ cannot be null字段不能为空，但数据中用空数据导致脏数据，修改其数据或对字段进行修改。

- Data too long for column ‘flash’字段设置太小引起的脏数据。

```
2017-01-02 17:01:19.308 [16963484-0-0-writer] ERROR StdoutPlug
inCollector
脏数据:
{“exception”:“Data truncation: Data too long for column ‘flash’ at
row 1”,“record”:[{“byteSize”:8,“index”:0,“rawData”:1,“type”:“LONG”},
{“byteSize”:8,“index”:3,“rawData”:2,“type”:“LONG”},{“byteSize”:8,“
index”:4,“rawData”:1,“type”:“LONG”},{“byteSize”:8,“index”:5,“rawData
”:1,“type”:“LONG”},{“byteSize”:8,“index”:6,“rawData”:1,“type”:“LONG
”}
```

排查思路：设置`Data too long for column ‘flash’`字段设置太小，但数据中数据太大导致脏数据，修改其数据或对字段进行修改。

- read-only数据库权限设置问题，设置只读权限。

```
2016-11-02 17:27:38.288 [12354052-0-8-writer] ERROR StdoutPlug
inCollector
脏数据:
{“exception”:“The MySQL server is running with the --read-only
option so it cannot execute this statement”,“record”:[{“byteSize”:
3,“index”:0,“rawData”:201,“type”:“LONG”},{“byteSize”:8,“index”:1
,“rawData”:1474603200000,“type”:“DATE”},{“byteSize”:8,“index”:2,“
rawData”:“9月23号12点” ,“type”:“STRING”},{“byteSize”:5,“index”:3,“
rawData”:“12:00”,“type”:“STRING”}
```

排查思路：设置`read-only`模式，同步数据全为脏数据，修改其数据库模式，运行可以写入。

- 分区错误。

  参数配置为$[yyyymm] 报错，日志如下所示。

  ```
  [2016-09-13 17:00:43]2016-09-13 16:21:35.689 [job-10055875] ERROR
  Engine
  ```

  经DataX智能分析，该任务最可能的错误原因如下。

  ```
  com.alibaba.datax.common.exception.DataXException: Code:[OdpsWriter-
  13]
  ```

  错误解读：执行MaxCompute SQL时抛出异常，可重试。MaxCompute目的表在运行MaxCompute SQL时抛出异常，请联系MaxCompute管理员处理。SQL内容如下所示。

  ```
  alter table db_rich_gift_record add IF NOT EXISTS
        partition(pt='${thismonth}');
  ```

  排查思路：由于加了单引号，调度参数替换无效。

  解决方法：’${thismonth}’去掉引号调度参数。

- column没有配成数组形式。

  ```
  Run  command failed.
  com.alibaba.cdp.sdk.exception.CDPException: com.alibaba.fastjson.
  JSONException: syntax error, **expect {,** actual error, pos 0
  at com.alibaba.cdp.sdk.exception.CDPException.asCDPException(
  CDPException.java:23)
  ```

  排查思路：JSON有问题，如下所示。

  ```
  "plugin": “mysql”,**
  “parameter”: {
  “datasource”: “xxxxx”,
  ** “column”: “uid”,**
  “where”: “”,
  “splitPk”: “”,
  “table”: “xxx”
  }
  **“column”: “uid”,-----没有配成数组形式**
  ```

- JDBC的格式填写错误。

  排查思路：JDBC格式填错，正确的格式为`jdbc:mysql://ServerIP:Port/Database`。

- 测试连通性失败。

  排查思路：

  - 防火墙对IP和端口账号是否有相关的限制。
  - 安全组的端口开发情况。

- **日志中报uid[xxxxxxxx]问题。**

```
Run  command failed.
com.alibaba.cdp.sdk.exception.CDPException: RequestId[F9FD049B-xxxx-
xxxx-xxx-xxxx] Error: CDP server encounter problems, please contact
 us, reason: 获取实例的网络信息发生异常，请检查RDS购买者id和RDS实例名,uid
[xxxxxxxx],instance[rm-bp1cwz5886rmzio92]ServiceUnavailable : The
request has failed due to a temporary failure of the server.
RequestId : F9FD049B-xxxx-xxxx-xxx-xxxx
```

排查思路：通常RDS同步至MaxCompute时，如果报上述错误，您可以直接将RequestId：F9FD049B-xxxx-xxxx-xxx-xxxx复制给RDS人员。

- **MongoDB中的query参数错误。**

MongoDB同步到MySQL报下面的问题，排查出JSON没有写好，是JSON中的query参数没有配置好。

```
Exception in thread “taskGroup-0” com.alibaba.datax.common.exception
.DataXException: Code:[Framework-13]
```

错误解读：DataX插件运行时出错，具体原因请参见DataX运行结束时的错误诊断信息。

```
org.bson.json.JsonParseException: Invalid JSON input. Position: 34.
Character: ‘.’.
```

排查思路：

- **错误示例：**`“query”:”{‘update_date’:{’$gte’:new Date().valueOf()/1000}}”`，**不支持如**`new Date()`**的参数。**
- **正确示例：**`“query”:”{‘operationTime’{’$gte’:ISODate(’${last_day}T00:00:00.424+0800’)}}”`。

- **Cannot allocate memory**

```
2017-10-11 20:45:46.544 [taskGroup-0] INFO  TaskGroupContainer -
taskGroup[0] taskId[358] attemptCount[1] is started
Java HotSpot™ 64-Bit Server VM warning: INFO: os::commit_memory
(0x00007f15ceaeb000, 12288, 0) failed; error=‘**Cannot allocate
memory’** (errno=12)
```

排查思路：内存不够。如果运行在自己的资源上，需要自行添加内存。如果运行在阿里巴巴的资源上，请提交工单进行咨询。

- **max_allowed_packet参数错误。**

错误信息如下所示。

```
Packet for query is too large (70 > -1). You can change this value
on the server by setting the max_allowed_packet’ variable. - **com
.mysql.jdbc.PacketTooBigException: Packet for query is too large (
```

```
70 > -1). You can change this value on the server by setting the
max_allowed_packet' variable.**
```

排查思路：

- **max_allowed_packet参数用来控制其通信缓冲区的最大长度。MySQL根据配置文件会限制server接受的数据包大小。有时候大的插入和更新会被max_allowed_packet参数限制掉，导致失败。**
- **max_allowed_packet参数的设置不宜设置过大，通常10m=10_1024_1024。**

- **HTTP Status 500读取日志失败。**

```
Unexpected Error:
Response is com.alibaba.cdp.sdk.util.http.Response@382db087[proxy
=HTTP/1.1 500 Internal Server Error [Server: Tengine, Date: Fri,
27 Oct 2017 16:43:34 GMT, Content-Type: text/html;charset=utf-8,
Transfer-Encoding: chunked, Connection: close,
**HTTP Status 500** - Read timed out**type** Exception report**
message****Read timed out****description****The server encountered
 an internal error that prevented it from fulfilling this request.+
+**exception**
java.net.SocketTimeoutException: Read timed out
```

排查思路：

调度运行报500的问题，如果是运行在默认资源上，读取日志失败，请直接联系技术支持帮您解决。如果运行在您自己的资源上，请重启Alisa即可。

> **说明:**
> **如果刷新后还是停止状态，您可以重启Alisa命令：切换到admin账号执行**`/home/admin/alisatatasknode/target/alisatatasknode/bin/serverct1 restart`。

- **hbasewriter参数：hbase.zookeeper.quorum配置错误。**

```
2017-11-08 09:29:28.173 [61401062-0-0-writer] INFO  ZooKeeper -
Initiating client connection, connectString=xxx-2:2181,xxx-4:2181
,xxx-5:2181,xxxx-3:2181,xxx-6:2181 sessionTimeout=90000 watcher=
hconnection-0x528825f50x0, quorum=node-2:2181,node-4:2181,node-5:
2181,node-3:2181,node-6:2181, baseZNode=/hbase
Nov 08, 2017 9:29:28 AM org.apache.hadoop.hbase.zookeeper.Recoverabl
eZooKeeper checkZk
WARNING: **Unable to create ZooKeeper Connection**
```

排查思路：

- **错误示例：“hbase.zookeeper.quorum”：“xxx-2,xxx-4,xxx-5,xxxx-3,xxx-6”**
- **正确示例：“hbase.zookeeper.quorum”：“您的zookeeperIP地址”**

- 没有找到相应的文件。

  经DataX智能分析，该任务最可能的错误原因如下所示。

  ```
  com.alibaba.datax.common.exception.DataXException: Code:[HdfsReader-
  08]
  ```

  错误解读：您尝试读取的文件目录为空。未能找到需要读取的文件，请确认您的配置项。

  ```
  path:  /user/hive/warehouse/tmp_test_map/*
  at com.alibaba.datax.common.exception.DataXException.asDataXExc
  eption(DataXException.java:26)
  ```

  排查思路：按照path找到相应的地方，检查对应的文件。如果没有找到文件，则对文件进行处理。

- 表不存在。

  经DataX智能分析，该任务最可能的错误原因如下所示。

  ```
  com.alibaba.datax.common.exception.DataXException: Code:[MYSQLErrCo
  de-04]
  ```

  错误解读：表不存在，请检查表名或者联系数据库管理员确认该表是否存在。

  表名为：xxxx，执行的SQL为`select * from xxxx where 1=2;`

  错误信息如下所示。

  ```
  Table ‘darkseer-test.xxxx’ doesn’t exist - com.mysql.jdbc.exceptions
  .jdbc4.MySQLSyntaxErrorException: Table ‘darkseer-test.xxxx’ doesn’t
   exist
  ```

  排查思路：`select * from xxxx where 1=2`判断表xxxx是否存在问题，如果有问题则需要对表进行处理。

1.12.3 添加数据源典型问题场景

DataWorks添加数据源的典型问题可以分为连通性问题、参数问题和权限问题。

连通性问题

连通性问题主要体现为测试连通性失败。

- 如果您使用的是RDS数据源，建议您首先为*RDS*添加白名单。
- 如果您使用的是ECS上自建数据库，建议您首先为*ECS*添加安全组。
- 问题现象

  添加MySQL数据源时，网络类型选择为经典网络，单击测试连通性时失败报错：测试连接失败，测试数据源联通性失败，连接数据库失败，数据库连接串…异常消息：Communications

`link failure. The last packet sent successfully to the server was 0 milliseconds ago.The dirver has not received any packets from the server`。

解决方案

出现上述报错通常都是网络连通性问题导致。建议检查您的网络是否可达、防火墙是否对该IP/端口有相关限制，以及安全组是否已配置对IP/端口放通。

- **问题现象：**

  添加阿里云MongoDB数据源，测试数据源连通性失败，报错如下：

```
error message: Timed out after 5000 ms while waiting for a server
that matches ReadPreferenceServerSelector{readPreference=primary}.
Client view of cluster state is {type=UNKNOWN, servers=[..] error
with code: PROJECT_DATASOURCE_CONN_ERROR
```

  问题解法

  处理此类问题时，首先需要确定您的DataWorks工作空间所处地域。使用阿里云MongoDB，需要确定网络类型是否为VPC。VPC环境下MongoDB不支持数据连通性测试（使用方案一可以避免该问题）。

  VPC环境下阿里云MongoDB数据同步有两种方案：

  - **方案一：通过公网进行数据同步**

    1. **数据源配置时，数据源类型选择连接串模式（数据集成网络可直接连通）。**
    2. **VPC环境下，您的MongoDB需要开通公网访问。**
    3. **在MongoDB上放行相关白名单IP，详情请参见添加白名单。**
    4. **进行数据连通性测试。**

  - **方案二：配置自定义资源组，从内网进行数据同步**

    1. **准备一台和MongoDB同区域、同网络的ECS作为调度资源，详情请参见新增任务资源。**
    2. **将这台ECS的IP加入MongoDB的白名单或者安全组。**
    3. **数据源测试连通时直接确定保存（不支持测试连通性）。**
    4. **修改资源组为自定义调度资源，测试运行。**

  说明：

  请务必添加相应的白名单。

- 问题现象

  添加自建MongoDB数据源，测试数据源连通性失败。

  问题解法

  1. 数据源配置时，数据源类型选择连接串模式（数据集成网络可直接连通）。
  2. 如果是VPC环境下ECS上自建的MongoDB，需要开通公网访问。
  3. 确认网络和端口之间是否能够连通，检查ECS的防火墙和安全组设置。
  4. 确认自建的数据库涉及的安全访问限制、权限的限制和远程登录的情况。
  5. 确认访问地址`host:port`、数据库名和用户名是否填写正确。

     > 说明:
     >
     > 添加MongoDB数据源时，使用的用户名必须是用户需要同步的这张表所在的数据库创建的用户名，不能用root。
     >
     > 例如需要导入name表，name表在test库，则此处数据库名称填写为test。
     >
     > 用户名为指定数据库中创建的用户名，不要使用root。例如之前指定的是test库，则用户名需使用test数据库中创建的账户。

- 问题现象

  VPC环境下添加Redis数据源，测试数据源连通性失败，报错如下。

  问题解法

  Redis添加数据源时如果没有公网IP，需要保证数据源和DataWorks工作空间地域一致，通过新增调度资源完成数据源的打通。

- 问题现象

  添加MongoDB数据源，已经配置白名单，测试数据源连通性仍然失败，报错如下：

  ```
  error message: Timed out after 5000 ms while waiting for a server
  that matches ReadPreferenceServerSelector{readPreference=primary}
  ```

  问题解法

  VPC网络的MongoDB数据源和Dataworks的默认资源组在内网上是不通的，所以无法直接进行同步任务，需要通过公网或者自定义资源组的方式进行连通。

- 问题现象

  Docker中安装的MySQL如何添加到数据源?

  问题解法

  Docker中安装的MySQL直接用服务器的公网IP组成的JDBC地址是无法连接的，连通性测试无法通过。您需要将MySQl的端口映射到宿主机上，使用映射出的端口链接。

- 问题现象

  配置Redis数据源失败，测试数据源连通性失败报错如下：

  ```
  error message: java.net.SocketTimeoutException: connect timed out
  ```

  问题解法

  目前DataWorks不支持Redis通过内网添加数据源。建议您为Redis数据源开通公网访问能力。数据源配置时，选择连接串模式（数据集成网络可直接连通），通过公网连接。

- 问题现象

  新增阿里云RDS数据源时，测试连通性不通。

  问题解法

  1. 当RDS数据源测试连通性不通时，需要到自己的RDS上添加数据同步机器IP白名单，详情请参见添加白名单。

     
     说明:
     如果使用自定义资源组调度RDS的数据同步任务，必须把自定义资源组的机器IP也加到RDS的白名单中 。

  2. 确保添加的信息正确：RDS实例ID和RDS实例主帐号ID、用户名、密码数据库名必须确保正确。

- 问题现象

  新增自建ECS中的MySQL数据源时，数据源测试连通性不通。

  问题解法

  1. 确认网络和端口之间是否能连通，检查ECS的防火墙以及安全组设置。
  2. 确认自建的数据库涉及的安全访问限制、权限的限制和远程登录的情况。
  3. 确认添加的用户名、密码、JDBC URL中的IP地址和端口的信息是否正确。
  4. 在VPC的环境下购买的ECS，只能用脚本模式运行任务，在添加数据源时测试连通性不能成功。购买ECS后，您可以添加自定义资源，将同步任务下发到相应的资源组运行。

参数问题

- **问题现象：**

  添加MySQL类型数据源时，单击测试连通性，报错如下：

  ```
  测试连接失败，测试数据源连通性失败，连接数据库失败…异常消息：No suitable direver found for...
  ```

  问题解法

  出现上述情况可能是JDBC URL格式填写错误导致，JDBC URL在填写时，请不要在URL中添加空格或任何特殊字符。正确格式为：`jdbc:mysql://ServerIP:Port/Database`。

- **问题现象**

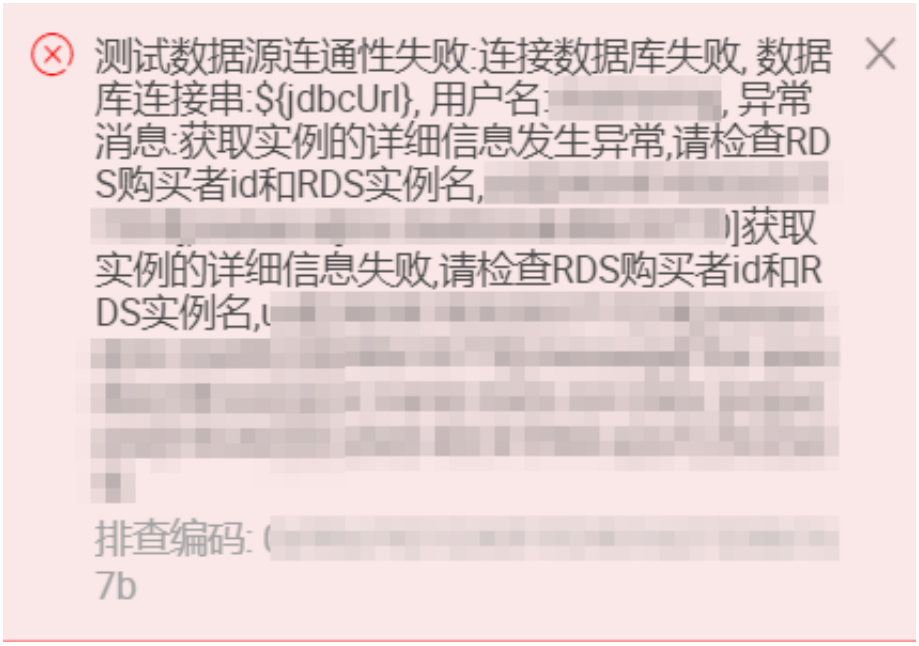
  使用用户名root添加MongoDB数据源时报错。

  问题解法

  添加MongoDB数据源时，使用的用户名必须是用户需要同步的这张表所在的数据库创建的用户名，不能用root。例如需要导入name表，name表在test库，则此处数据库名称填写为test。用户名为指定数据库中创建的用户名，不要使用root。例如之前指定的是test库，则用户名需使用test数据库中创建的账户。

- **问题现象**

  添加RDS数据源失败，数据库连接不上，报错如下。

  问题解法

  需要检查填写的UID是否为子账号的UID，此处要填写RDS所属主账号的UID方可成功添加数据源。

- **问题现象**

  加ODPS默认数据源时报测试连通性失败。

  问题解法

  ODPS默认数据源无需添加，默认为odps_fisrt。

- **问题现象**

  DataWorks的数据源支持HybridDB for PostgreSQL吗?

  问题解法

  支持，添加时选择关系型数据库PostgreSQL即可。

- **问题现象**

  没有外网地址的DRDS实例，配置数据源时，是否支持将实例的内网地址映射为自定义的域名?

  问题解法

  需要严格按照格式来，目前不支持域名映射的方式。

- **问题现象**

  添加RDS数据源时，为什么白名单已添加，依然报错提示`user not exist ip white list reference`**。**

  问题解法

  出现这种情况通常是由于用户名输入错误。您可以参见创建账号和数据库检查自己输入的用户名是否正确。

权限问题

- **问题现象**

  添加ADS数据源时，测试数据连通性报错如下：

  ```
  连接数据库失败，数据库连接串:${jdbcUrl}，用户名:XXXXXX，异常消息:You don
  't have privilege for connecting database 'dw', userId=RAM$XXX,
  schemaId=XX
  ```

  问题解法

  首先，您需要检查在数据源中填写的子账号是否有ADS的访问权限。分析型数据库用户基于阿里云帐号进行认证，用户建立的数据库属于该用户，用户也可以授权给其他用户访问其数据库下的表，所以连接的用户是需要在ADS上进行授权的，具体的说明参见用户账号类型与用户管理。

- **问题现象**

  子帐户无权限查看数据源，无法创建数据源，提示您没有权限进行此操作。

  问题解法

  只有项目管理员权限的RAM子账户才可以增删改数据源。

1.12.4 同步任务等待槽位

问题描述

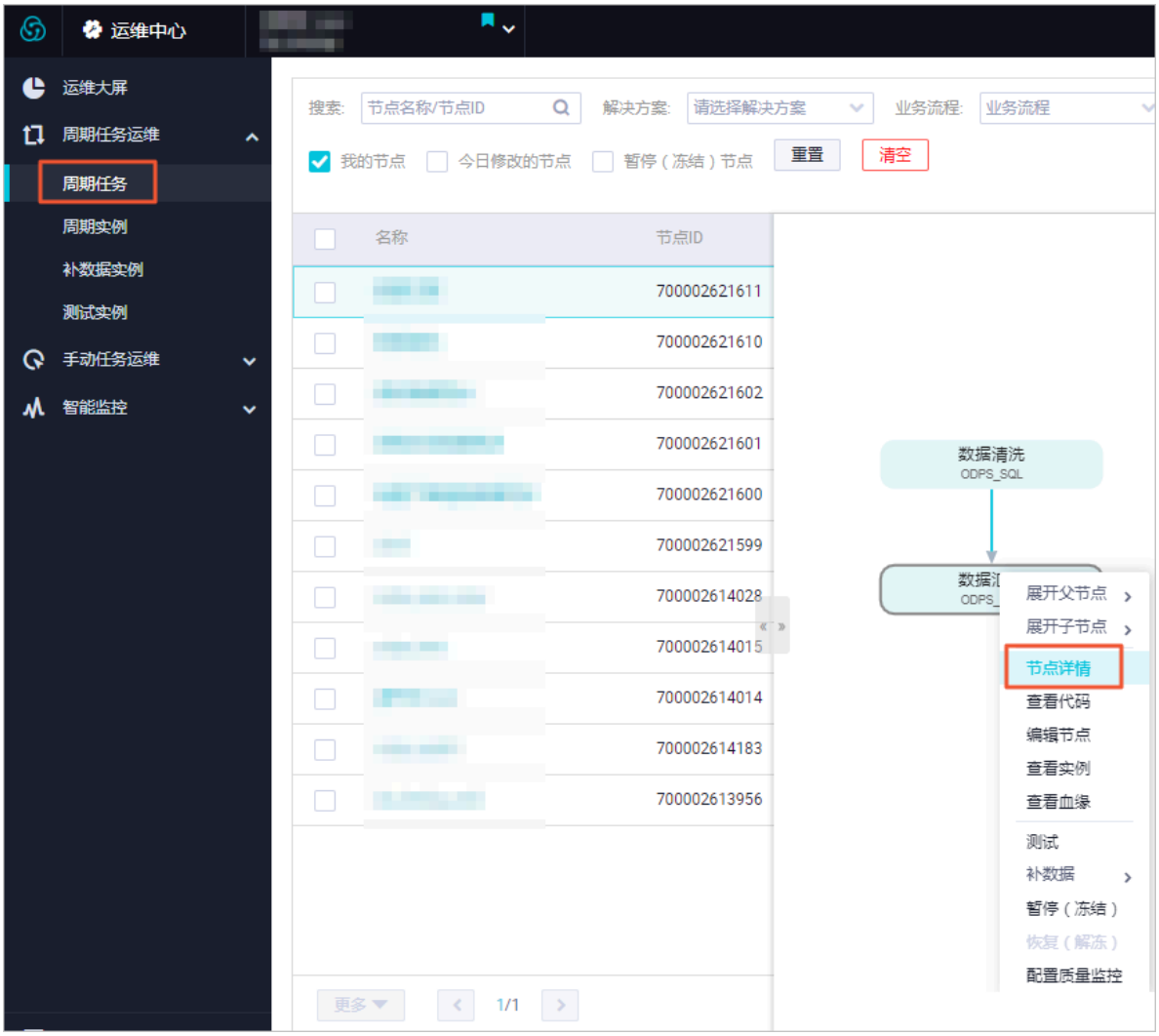
任务未正常运行，日志提示目前实例还没有产生日志信息，在等待槽位。

问题原因

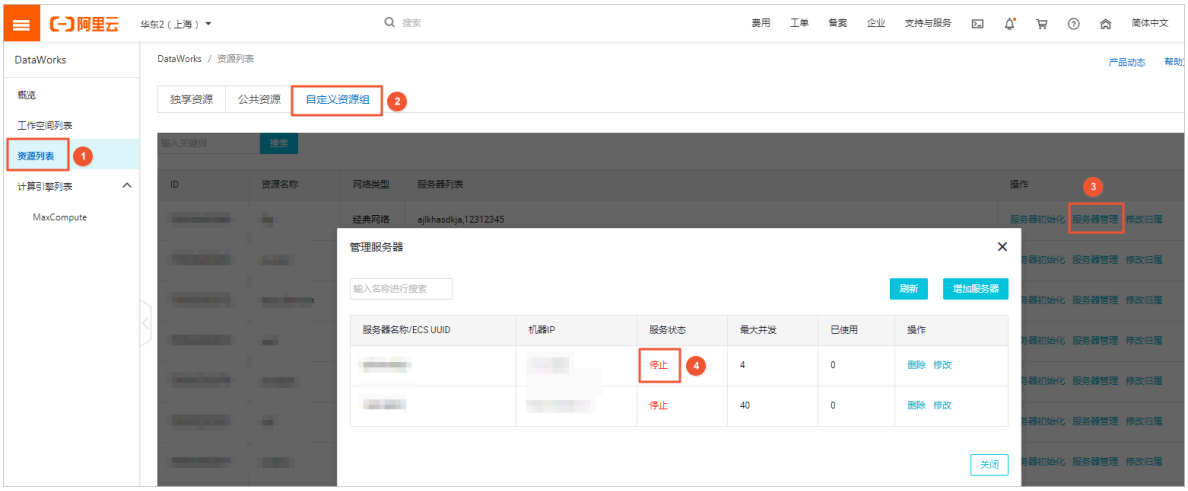
出现上述提示的原因是任务的配置调度使用的是自定义资源，但目前没有可用的自定义资源。

解决方法

1. **您可以进入DataWorks > 运维中心 > 周期任务运维 > 周期任务页面，右键单击DAG图中没有按照预期进行调度的任务，选择节点详情，查看任务使用的资源组。**



2. **进入资源列表 > 自定义资源组页面，找到任务使用的调度资源，单击服务器管理，查看服务器的状态是否停止，或是否被其他任务占用。**



3. 如果以上排查无法解决问题，可以执行下述命令重启服务。

```
su - admin
/home/admin/alisatasknode/target/alisatasknode/bin/serverctl restart
```

1.12.5 编码格式设置问题

数据集成的同步任务设置编码格式后，如果数据含有表情符，在运行任务时可能出现同步失败且产生脏数据或同步成功但数据乱码的问题。

同步失败且产生脏数据

问题描述

数据集成任务失败，且因编码问题产生脏数据，报错日志如下所示。

```
016-11-18 14:50:50.766 [13350975-0-0-writer] ERROR StdoutPlug
inCollector - 脏数据:<br>
{"exception":"Incorrect string value: '\\xF0\\x9F\\x98\\x82\\xE8\\
xA2...' for column 'introduction' at row 1","record":[{"byteSize":8,"
index":0,"rawData":9642,"type":"LONG"},
{"byteSize":33,"index":1,"rawData":"A公司出来的女汉子，扛得了箱子，招待好顾
客![1](http://docs-aliyun.cn-hangzhou.oss.aliyun-inc.com/assets/pic/
56134/cn_zh/149872864****/%E5%9B%BE%E7%89%877.png)
被自己感动cry","type":"STRING"},
{"byteSize":8,"index":4,"rawData":0,"type":"LONG"}],"type":"writer"}
2016-11-18 14:50:51.265 [13350975-0-0-writer] WARN  CommonRdbmsWriter$
Task - 回滚此次写入，采用每次写入一行方式提交:java.sql.BatchUpdateException
: Incorrect string value: '\xF0\x9F\x88\xB6\xEF\xB8...' for column '
introduction' at row 1
```

问题原因

在对数据库做相应的编码格式设置或添加数据源时，未将编码设置为utf8mb4**。只有**utf8mb4**编码支持同步表情符。**

解决方法

- **添加JDBC格式的数据源时，需要修改**utf8mb4**的设置，例如**jdbc:mysql://xxx.x.x.x:3306/database?com.mysql.jdbc.faultInjection.serverCharsetIndex=45**。这样，在数据源设置表情符可以同步成功。**
- **将数据源编码格式改成**utf8mb4**。例如在RDS控制台修改RDS的数据库编码格式。**

> **说明:**
> **如果需要设置RDS数据源编码格式**set names utf8mb4**，在添加数据源时必须使用无公网IP+连接串方式。**

同步成功但数据乱码

问题描述

数据同步任务虽然成功，但数据乱码。

问题原因

发生乱码的原因有以下三种：

- **源端的数据本身就是乱码。**
- **数据库和客户端的编码不一样。**
- **浏览器编码不一样，导致预览失败或乱码。**

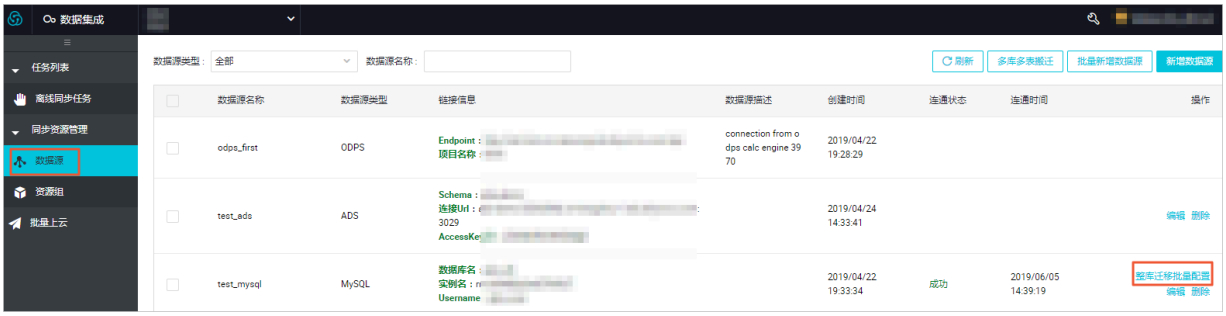
解决方法

您可以针对产生乱码的不同原因，选择相应的解决方法：

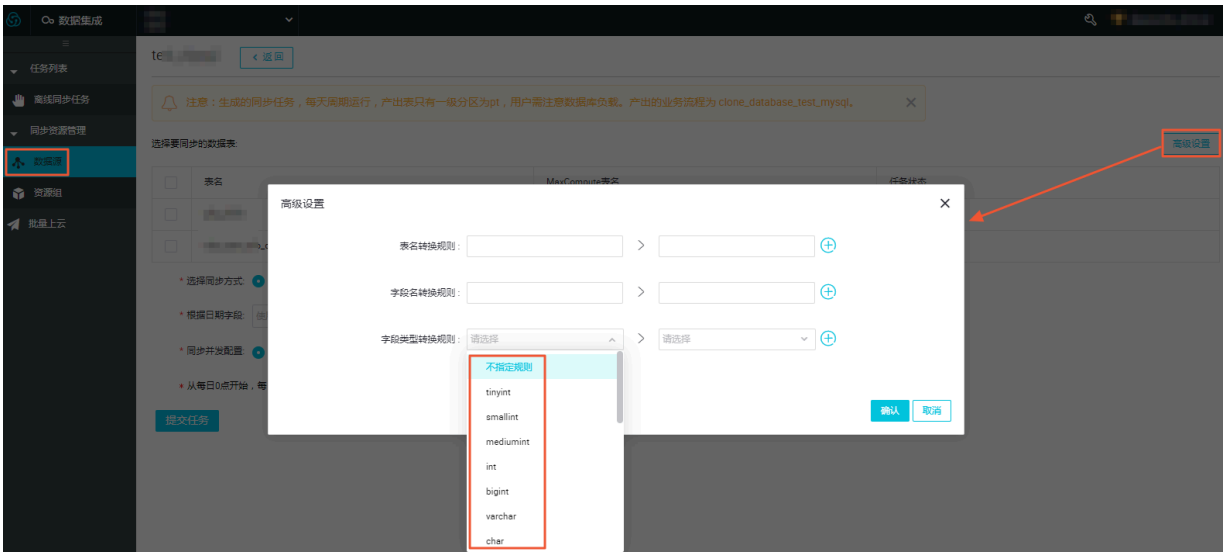
- **如果您的原始数据乱码，需首先处理好原始数据，再进行同步任务。**
- **数据库和客户端编码格式不一致，需先修改编码格式。**
- **浏览器编码和数据库或客户端编码格式不一致，需先统一编码格式，然后进行数据预览。**

1.12.6 整库迁移数据类型

整库迁移目前仅支持MySQL（包括RDS中的MySQL）、Oracle数据源同步至MaxCompute，可以从已经添加好的MySQL/Oracle数据源中进入整库迁移页面。



下面仅针对整库迁移中高级设置的数据类型进行介绍。



整库迁移源端MySQL支持的数据源类型包括TINYINT、SMALLINT、MEDIUMINT、INT、BIGINT、VARCHAR、CHAR、TINYTEXT、TEXT、MEDIUMTEXT、LONGTEXT、YEAR、FLOAT、DOUBLE、DECIMAL、DATE、DATETIME、TIMESTAMP、TIME和BOOL。

目标端MaxCompute支持的数据源类型包括BIGINT、STRING、DOUBLE、DATETIME和BOOLEAN。

上述MySQL支持的数据类型均支持与MaxCompute数据源类型之间的转换。

> 说明:
> MySQL中的BIT，如果是bit(2)以上，则目前不支持与BIGINT、STRING、DOUBLE、DATETIME和BOOLEAN等类型转换。如果是bit(1)，则会被转换成BOOLEAN。

1.12.7 RDS同步失败转换成JDBC格式

问题描述

从RDS（MySQL/SQL Server/PostgreSQL）同步到自建MySQL/SQL Server/PostgreSQL时，报错为：DataX无法连接对应的数据库。

解决方法

以 RDS（MySQL）的数据同步到自建SQL Server为例，操作如下：

1. 新建一个数据源，将数据源配置为MySQL>JDBC 格式。
2. 使用新数据源配置同步任务，重新执行即可。

> 说明:
> 如果是在RDS（MySQL）>RDS（SQL Server）等云产品之间同步时，建议选择RDS（MySQL）>RDS（SQL Server）数据源来配置同步任务。

1.12.8 同步表列名是关键字任务失败

问题描述

用户做同步任务时，同步的表的列名是关键字，导致任务失败。

解决方法

以MySQL数据源为例。

1. 新建一张表aliyun，建表语句如下：

```
create table aliyun (`table` int ,msg varchar(10));
```

2. 创建视图，给table列取别名。

```
create view v_aliyun as  select `table` as col1,msg as col2 from
aliyun;
```

> 说明：
> - table是MySQL的关键字，在数据同步时，拼接出来的代码会报错。因此通过创建视图，给table列起别名，以绕过此限制。
> - 不建议使用关键字作为表的列名。

3. 上述语句给有关键字的列取了别名，那么在配置数据同步任务时，可以选择v_aliyun视图来代替aliyun这张表 。

> 说明：
> - MySQL的转义符是`关键字`。
> - Oracle和PostgreSQl的转义符是"关键字"。
> - SQlServer的转义符是[关键字]。

1.12.9 数据同步任务如何自定义表名

数据背景

表是按天分的（如orders_20170310、orders_20170311和orders_20170312），每天一个表， 表结构一致。

实现需求

创建数据同步任务，将表数据导入至MaxCompute中。希望只需要创建一个同步任务，自定义表名，实现每天凌晨自动从源数据库读取昨天的表数据（例如今天是2017年3月15日，自动从源数据库中读取orders_20170314的表的数据导入，以此类推）。

实现方式

1. 登录*DataWorks*控制台，单击对应工作空间操作栏中的进入数据开发。
2. 通过向导模式创建数据同步任务，配置时数据来源表先选一个表名如orders_20170310。详情请参见*#unique_144*。

3. 单击转换脚本按钮，将向导模式转换为脚本模式。

4. 在脚本模式中，改用来源表的表名为变量，如orders_${tablename}。

   在任务的参数配置中，给变量tablename赋值。由数据背景得知，表名是按天区分，而需求是每天读取昨天的表，所以赋值为$[yyyymmdd-1]。

   > **说明:**
   > 您也可以改用来源表的表名为变量时，直接写为orders_${bdp.system.bizdate}。

完成上述配置后，保存并提交，然后再进行后续操作。

1.12.10 使用用户名root添加MongoDB数据源报错

问题描述

使用用户名root添加MongoDB数据源时报错。

问题原因

添加MongoDB数据源时，使用的用户名必须是用户需要同步的这张表所在的数据库创建的用户名，不能用root。

解决方法

例如需要导入name表，name表在test库，则此处数据库名称填写为test。

用户名为指定数据库中创建的用户名，不要使用root。例如之前指定的是test库，则用户名需使用test数据库中创建的账户。

1.12.11 自定义资源组常见问题

本文将为您介绍自定义资源组在使用、配置文件、命令等方面的常见问题和解决方案。

应用场景

- 保证运行资源：由于集群共享默认资源组，会存在水位变高导致任务长时间等待的情况。如果您对任务有较高的资源使用需求，可以使用自定义资源组来自建任务运行集群。
- 连通网络：由于默认资源组无法连通VPC环境下的数据库，您可以使用自定义资源组进行网络连通。
- 用于调度资源组：调度槽位资源紧张的情况下，您可以使用自定义资源组。

- 提升并发能力：默认资源组的运行槽位有限，您可以通过自定义资源组扩大槽位资源，允许更多的并发任务同时调度运行。

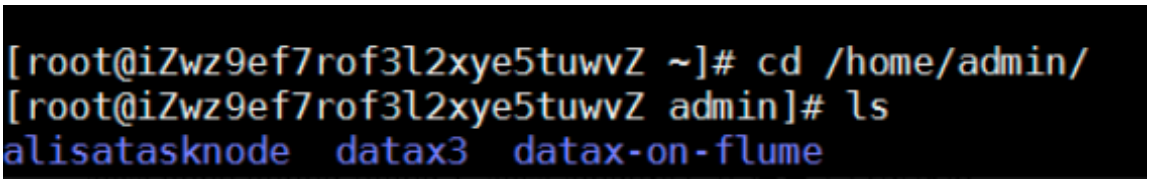
使用限制

- 1台ECS只能注册到1个自定义资源组下，1个自定义资源组可以添加多个ECS。
- 经典网络和专有网络注册的区别为：经典网络是主机名称，专有网络是UUID。
- 1个自定义资源中只允许存在1种网络类型。
- 不支持运行手动任务实例。
- ECS需要具备公网访问能力，ECS可以配置公网IP、EIP、NAT网关SNAT。

配置文件

通过DataWorks界面引导，完成自定义资源组的安装后，您可以登录ECS查看agent插件的下述信息。

- 默认安装路径：*/home/admin/*， 默认路径下通常会有以下目录信息。
  - alisatasknode：agent有关配置和命令所在目录。
  - datax和datax-on-flume：数据同步插件库和配置所在目录。



- agent有关命令

  当前支持对agent进程进行stop/start/restart等命令操作，具体操作命令如下：

```
/home/admin/alisatasknode/target/alisatasknode/bin/serverctl start/
stop/restart
```

- 运行日志

  agent运行日志有以下2个存放路径：

  - */home/admin/alisatasknode/taskinfo/*：存放Shell脚本运行的日志信息，和DataWorks节点运行日志页面中查看的结果一致。
  - */home/admin/alisatasknode/logs*：alisatasknode.log日志文件中存放的是agent插件的运行信息，如接收到的任务运行/kill操作、agent心跳状态等。
  - */home/admin/datax3/log*：存放DataX任务的详细运行日志，遇到任务执行失败，可以查看该部分日志查找原因。

监控手段

您可以通过下述方法监控agent进程的运行状态，在监控到agent进程退出后，可以及时进行恢复。

1. **root用户登录到ECS机器。**
2. **执行命令**`wget https://alisaproxy.shuju.aliyun.com/install_monitor.sh --no-check-certificate`。
3. **执行命令**`sh install_monitor.sh`，**监控日志默认存放在***/home/admin/alisatasknode/monitor/monitor.log***中。**

DataWorks调度分类

自定义资源组在DataWorks调度体系中使用， 当前DataWorks调度体系分为一级调度资源和二级运行资源。

- **一级调度资源：进入运维中心 > 周期实例 > 属性页面查看一级调度资源，用来调度实例。**
- **二级运行资源：进入数据集成 > 同步资源管理 > 资源组页面查看二级任务运行资源。**

自定义资源组的使用

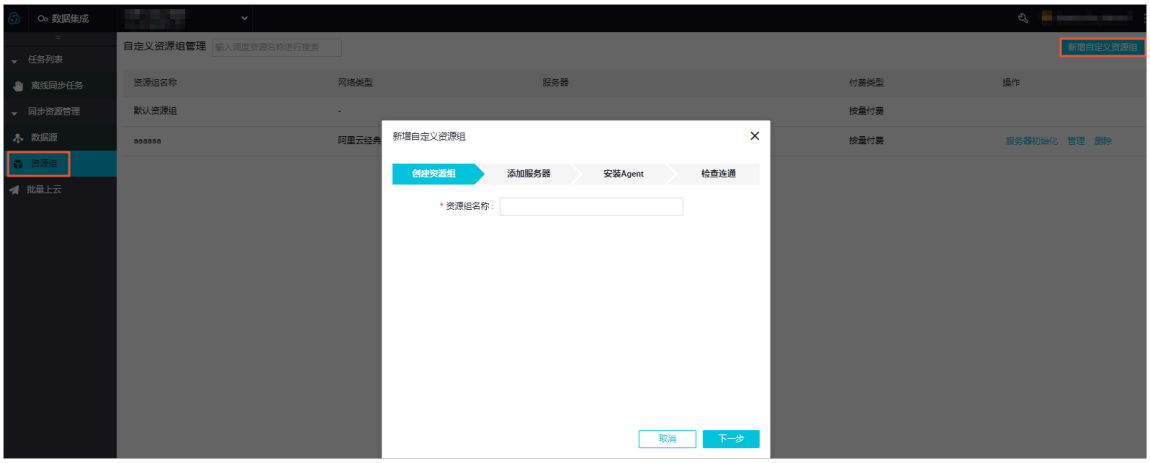
- **配置一级调度资源**

  登录DataWorks控制台，进入资源列表页面，创建自定义资源组。

  > **说明:**
  > **在该页面创建的调度资源适用于Shell任务，配置的是一级调度资源。**

- **配置二级运行资源**

  1. **登录DataWorks控制台，进入数据集成 > 同步资源管理 > 资源组页面新增自定义资源组。**



> **说明:**

此页面添加的资源组仅用于数据同步任务，配置的是二级运行资源。

2. 新建完成后，能且只能在数据开发页面配置任务时，在通道控制 > 任务资源组配置中选择。

自定义资源组的常见问题

- 常见问题一：添加ECS资源组时，报错gateway already exists。

1. 报错提示对应ECS已经在gateway中注册过，因为1个ECS只能添加到1个自定义资源组下，所以需要您在资源列表和数据集成 > 同步资源管理 > 资源组页面，查看是否存在同名或同UUID的自定义资源组。
2. 如果工作空间中没有发现对应的自定义资源组，则收集request ID信息进行咨询。

- 常见问题二：添加自定义资源组后状态不可用。
  - 查看alisatasknode.log日志，确认是否有心跳上报302的情况。如果上报心跳302，则可以排查下述问题。
    - 查看UUID是否一致。对比自定义资源组页面的UUID信息和ECS上执行命令`dmidecode |grep UUID`的结果是否一致。

      说明：
      此处的UUID区分大小写。
    - 如果UUID不一致，需要填写正确的UUID并重新安装agent。

      说明：

dmidecode命令在3.0.5版本及之前版本会是大写的方式显示UUID，如果升级到3.1.2或以上版本，则会小写显示UUID，此时会导致心跳异常，需要重新安装agent进行恢复。

- 确认config.properties中配置的用户名及密码是否和自定义资源组添加界面一致，如果不一致，请参见agent插件安装界面给出的命令重新安装agent。
- 如果UUID和密码信息正确，请查看config.properties中的字段node.uuid.enable，针对VPC类型的ECS，该字段的值需要为true。如果VPC类型下node.uuid.enable的值为false，则修改为true，重启agent进程即可。

- 查看alisatasknode.log日志，确认是否存在connection timeout的信息。如果存在，则可以进行如下处理：
  1. 查看ECS是否有公网能力，如公网IP、EIP、NAT网关SNAT IP，可以执行`ping www.taobao.com`确认是否可以连通。
  2. 如果ECS有公网能力，请查看ECS的安全组配置中内网出方向或公网出方向是否进行访问限制。如果有访问限制，需要对gateway的IP和端口放行。
- 场景三：ECS状态正常，但Shell任务执行失败。
  - 任务执行异常，运行日志显示如下：

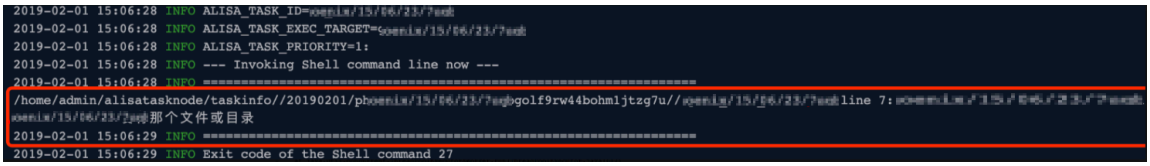
```
2019-01-17 19:32:36 INFO ALISA_TASK_EXEC_TARGET ALISA_TASK_EXEC_TARGET
2019-01-17 19:32:36 INFO ALISA_TASK_PRIORITY=1:
2019-01-17 19:32:36 INFO --- Invoking Shell command line now ---
2019-01-17 19:32:36 INFO
================================================================
The task process was abnormal exit, system set task rerun!!!
/home/admin/alisatasknode/taskinfo//20190117/phoenix/19/32/35/usa5vzt2fso5fcyu3q95fpn9/T3_0699121848.log-END-EOF
```

    您可以进行如下操作：

    - 查看alisatasknode.log日志中具体的报错信息，可以根据T3_0699121848关键字进行搜索。
    - 登录ECS，切换到admin用户下，执行`python -V`命令查看Python版本是否为2.7或2.6版本。

      说明：
      agent当前支持的是Python2.7或2.6版本，通常Python版本不对会导致replace user hive conf error的错误。

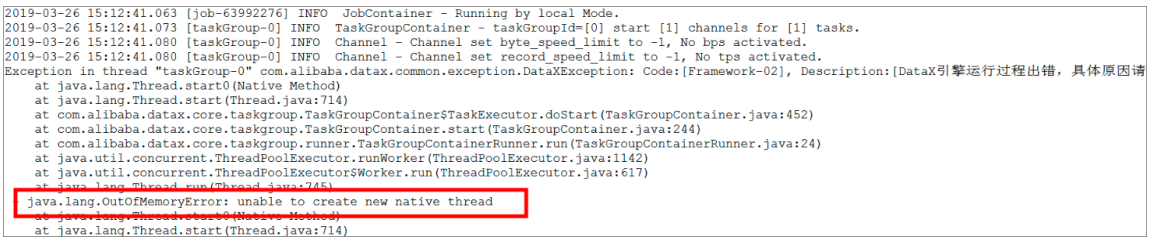
- 查看DataWorks运行日志，找不到对应的文件。



您可以登录ECS并切换到admin用户，执行命令sh -x脚本名，确认是否可以正常执行，根据报错信息进行调试。

- 常见错误四：自定义资源组下任务执行OOM。
  - 报错问题

    获取用户运行报错日志如下图所示，提示无法分配内存给作业线程。



- 排查思路

  自定义资源组创建时设置的内存数决定了资源组可提供的槽位能力。资源组内的系统进程和agent进程也会占用一部分内存，不能将ECS实例所有内存都用于槽位资源，会导致高并发下某些作业OOM。

- 解决方法

  建议您调小自定义资源组下的内存数设置，可以考虑预留2G的空间给系统和agent进程使用。如果还有其它进程，则需预留更多的内存。