阿里云 DataWorks

快速开始

文档版本: 20190816

为了无法计算的价值 | [] 阿里云

<u>法律声明</u>

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。

- 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。
- 未经阿里云事先书面许可,任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分 或全部,不得以任何方式或途径进行传播和宣传。
- 3. 由于产品版本升级、调整或其他原因,本文档内容有可能变更。阿里云保留在没有任何通知或者 提示下对本文档的内容进行修改的权利,并在阿里云授权通道中不时发布更新后的用户文档。您 应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
- 4. 本文档仅作为用户使用阿里云产品及服务的参考性指引,阿里云以产品及服务的"现状"、"有缺陷"和"当前功能"的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引,但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的,阿里云不承担任何法律责任。在任何情况下,阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害,包括用户使用或信赖本文档而遭受的利润损失,承担责任(即使阿里云已被告知该等损失的可能性)。
- 5. 阿里云网站上所有内容,包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计,均由阿里云和/或其关联公司依法拥有其知识产权,包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意,任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外,未经阿里云事先书面同意,任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称(包括但不限于单独为或以组合形式包含"阿里云"、Aliyun"、"万网"等阿里云和/或其关联公司品牌,上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司)。
- 6. 如若发现本文档存在任何错误,请与阿里云取得直接联系。

通用约定

格式	说明	样例
•	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。
A	该类警示信息可能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,恢复业务所需 时间约10分钟。
Ê	用于补充说明、最佳实践、窍门等,不 是用户必须了解的内容。	道 说明: 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	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。	单击 确定。
courier 字体	命令。	执行 cd /d C:/windows 命令,进 入Windows系统文件夹。
##	表示参数、变量。	bae log listinstanceid Instance_ID
[]或者[a b]	表示可选项,至多选择一个。	ipconfig[-all -t]
	表示必选项,至多选择一个。	<pre>swich {stand slave}</pre>

目录

法律声明	I
通用约定	I
1入门概述	1
2 建表并上传数据	3
3 创建业务流程	10
4 创建同步任务	16
5 设置周期和依赖	24
6 运行及排错	27
7 使用临时查询快速查询SQL(可选)	

1入门概述

本模块将指引您快速完成一个完整的数据开发和运维操作。

▋ 说明:

- 如果您是第一次使用DataWorks,请确认已经根据准备工作模块的操作,准备好账号和工作空间角色等内容。然后进入DataWorks控制台,单击对应工作空间后的进入数据开发,即可开始数据开发操作。
- ·本模块的操作在标准模式的工作空间下进行。如果您是简单模式的工作空间,操作步骤同标准 模式。但在提交任务时,不会区分开发环境和生产环境。

通常情况下,通过DataWorks的工作空间实现数据开发和运维,包含以下操作:

- 1. 建表并上传数据
- 2. 创建业务流程
- 3. 创建同步任务
- 4. 设置周期和依赖
- 5. 运行及排错

您也可以选择直接使用DataWorks临时查询功能,快速编写SQL语句操作MaxCompute。详情请 参见#unique_10。 数据开发和运维的基本流程,如图 1-1: 流程图所示。

图 1-1: 流程图



在正式开始操作DataWorks前,您可以参见DataWorks V2.0系列详解视频对DataWorks V2.0各 功能模块特性进行深入学习。

- · DataWorks V2.0版本概述与最佳实践
- · DataWorks V2.0前生后世
- ・ DataWorks V2.0常见问题与难点分析
- · DataWorks V2.0数据开发功能与用法解析
- · DataWorks V2.0数据集成简介与最佳实践
- · DataWorks V2.0智能监控简介与最佳实践
- · DataWorks V2.0数据服务功能及用法解析
- · DataWorks V2.0数据质量简介及最佳实践
- · DataWorks V2.0数据安全简介与最佳实践
- · Function Studio简介与使用指导

2 建表并上传数据

本文将以创建表bank_data和result_table为例,为您介绍如何通过DataWorks V2.0创建表并上 传数据。



其中表bank_data用于存储业务数据,表result_table用于存储数据分析后产生的结果。

创建表bank_data

- 1. #unique_12后, 单击对应工作空间操作栏下的进入数据开发。
- 2. 进入DataStudio(数据开发)页面,选择新建 > 表。



3. 在新建表对话框中,填写表名为bank_data。

新建表		×
数据库类型:	MaxCompute	
表名:	bank_data	

- 4. 单击提交。
- 5. 进入新建表页面,选择DDL模式。
- 6. 在DDL模式对话框中输入建表语句, 单击生成表结构, 并确认操作。

	DDL模式	×
	<pre>1 CREATE TABLE `bank_data` (2</pre>	
	 9 、 当前操作会覆盖掉页面所有操作,确认覆盖? 10 、n 11 、c 12 、c 13 、c 14 、pdays、double COMMENT、与上一次联系的时间间隔,, 15 、previous、double COMMENT、之前与客户联系的次数,, 16 、poutcome、string COMMENT、之前市场活动的结果、 	
満加字段 上移 下移 字段英文名	2 生成表结构	取消

创建表的更多SQL语法请参见#unique_13。

本示例的建表语句如下所示:

CREATE TAB (LE IF NOT EXISTS bank_data	
age job	BIGINT COMMENT '年龄', STRING COMMENT '工作类型',	

STRING	COMMENT	'婚否',
STRING	COMMENT	'教育程度',
STRING	COMMENT	'是否有信用卡',
STRING	COMMENT	'房贷',
STRING	COMMENT	'贷款',
STRING	COMMENT	'联系途径',
STRING	COMMENT	'月份',
STRING	COMMENT	'星期几',
STRING	COMMENT	'持续时间'、
BIGINT	COMMENT	'本次活动联系的次数',
DOUBLE	COMMENT	'与上一次联系的时间间隔',
DOUBLE	COMMENT	'之前与客户联系的次数',
STRING	COMMENT	'之前市场活动的结果',
DOUBLE	COMMENT	'就业变化速率',
DOUBLE	COMMENT	'消费者物价指数',
DOUBLE	COMMENT	'消费者信心指数',
DOUBLE	COMMENT	'欧元存款利率',
DOUBLE	COMMENT	'职工人数',
BIGINT	COMMENT	'是否有定期存款'
	STRING STRING STRING STRING STRING STRING STRING STRING BIGINT DOUBLE DOUBLE DOUBLE DOUBLE DOUBLE DOUBLE DOUBLE DOUBLE BIGINT	STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT STRING COMMENT BIGINT COMMENT DOUBLE COMMENT

7. 表结构生成后,输入表的中文名,并分别提交到开发环境和提交到生产环境。

bank_data x	
DDL模式 从开发环境加载 提交到开发环境 从生产环境加载 提交到生产环境	
表名 bank_data	
中文名用户信息表	
	C
描述	

8. 创建成功后,您可以在左侧导航栏的表管理中,输入表名进行搜索。搜索成功后,双击表名,即 可查看表信息。



创建表result_table

1. 进入DataStudio(数据开发)页面,选择新建>表。



- 2. 在新建表对话框中,填写表名为result_table。
- 3. 进入新建表页面,选择DDL模式。
- 4. 在DDL模式对话框中输入建表语句,单击生成表结构,并确认操作。

本示例的建表语句如下所示:

```
CREATE TABLE IF NOT EXISTS result_table
(
education STRING COMMENT '教育程度',
num BIGINT COMMENT '人数'
);
```

- 5. 表结构生成后,输入表的中文名,并分别提交到开发环境和提交到生产环境。
- 创建成功后,您可以在左侧导航栏的表管理中,输入表名进行搜索。搜索成功后,双击表名,即 可查看表信息。

本地数据上传至bank_data

DataWorks支持以下操作:

- · 将保存在本地的文本文件中的数据,上传至工作空间的表中。
- ·通过数据集成模块,将业务数据从多个不同的数据源导入至工作空间。

▋ 说明:

本文将使用本地文件作为数据来源。本地文本文件上传有以下限制:

- · 文件类型: 仅支持.txt、.csv和.log文件类型。
- ・文件大小:不超过10M。
- ・操作对象:支持分区表导入和非分区表导入,但不支持分区值为中文。

以导入本地文件banking.txt至DataWorks为例,操作如下:

1. 单击导入。

6	💸 DataStudio	~
		துகாக திடு பெ
s	数据开发	Q 文件名称/创建人
*	组件管理	> 解决方案
Q	—————————————————————————————————————	▼ 业务流程 問
©	运行历史	> 🛃 workshop > 🛃 works
â	手动业务流程 💷	▶ <mark> </mark> 数据集成
⊞	公共表	▶ 🗤 数据开发
⊒⁰	表管理	● vij start 我锁定 07-04 16:07 ● <mark>Soj</mark> insert_data 我锁定 07-11 [·]
fx	函数列表	> 圖 表
	MaxCompute资源	> <mark>∅</mark> 资源 > <mark>∱</mark> 函数
Σ	MaxCompute函数	····································
亩	回收站	▶ @ 控制

2. 在数据导入向导对话框中,选择要导入数据的表,单击下一步。

数据导入向导			×
选择要导入数据的表: bank_data			
名称	类型	描述	
age	bigint	年龄	
job	string	工作类型	
marital	string	婚否	
education	string	教育程度	
default	string	是否有信用卡	
		—————————————————————————————————————	网消

3. 单击浏览...,选择本地数据文件,配置导入信息。确认无误后,单击下一步。

数据导入向	导									×
选择数据导入方式: 💽 上传本地文件 🔘 来自数据服务 🔵 来自数据分析的电子表格										
	选择文件:					ž	技只 … 武城			
选	择分隔符:	0 逗号								
原	始字符集:	GBK								
Ę	入起始行:	1								
首	行为标题:	V								
数据预览 由	于数据量太大									
44	blue-coll ar	married	basic.4y	unknow n	yes	no	cellular	aug	thu	210
53	techni cian	marri ed	unkno wn	no	no	no	cellul ar	nov	fri	138
28	mana geme	single	univer sity.d	no	yes	no	cellul	jun	thu	339
								上一步	下一步	取消

4. 选择目标表字段与源字段的匹配方式(本示例选择按位置匹配),单击按位置匹配。

数据导入向导		×
选择目标表字段与源字段的匹配方式: 💿 按位置匹配 💿 掛	这称匹配	
目标字段	源字段	
age		
job		
marital		
education		
default		
	上一步 导入数据	取消

文件导入后,系统将返回数据导入成功的条数或失败的异常。

其他数据导入方式

・ 创建数据同步任务

此方式适用于保存在RDS、MySQL、SQL Server、PostgreSQL、MaxCompute、OCS、 DRDS、OSS、Oracle、FTP、DM、HDFS和MongoDB等多种数据源中的各种数据。

通过DataWorks创建数据同步的具体操作,请参见创建同步任务。

・本地文件上传

此方式适用于文件大小不超过10M、文件类型为.txt和.csv的数据,目标支持分区表和非分区 表,但不支持中文作为分区。

通过DataWorks进行本地文件上传,具体操作请参见本地数据上传至bank_data。

· 使用Tunnel命令上传文件

此方式适用于任意大小的的本地文件和其他资源文件等。

通过MaxCompute客户端提供的Tunnel命令,来进行数据的上传及下载。当本地数据文件 需要上传至分区表时,可以通过客户端Tunnel命令方式进行上传。详情请参见Tunnel命令操 作。

后续步骤

现在,您已经学习了如何创建表并上传数据,您可以继续学习下一个教程。在该教程中,您将学习 如何创建业务流程,对工作空间的数据进行计算与分析。详情请参见<mark>创建业务流程</mark>。

3 创建业务流程

本文将以创建业务流程为例,为您介绍如何在业务流程中创建节点并配置依赖关系,以方便的设计 来展现数据分析的步骤和顺序。并简要说明如何利用数据开发功能,对工作空间的数据进行深入分 析和计算。

DataWorks的数据开发功能支持在业务流程中,通过可视化拖拽来完成节点间的依赖设置。以操 作业务流程的方式,实现对数据的处理和相互依赖。目前支持ODPS SQL、ODPS Script、ODPS Spark、PyODPS、虚拟节点、ODPS MR和Shell等多种节点类型,详情请参见#unique_18。

前提条件

开始本操作前,请确保您已根据建表并上传数据的操作,在工作空间中准备好业务数据 表bank_data和其中的数据,以及结果表result_table。

创建业务流程

- 1. #unique_12后,单击对应工作空间操作栏下的进入数据开发。
- 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。

新建业务流程		×
业务名称:	works	
描述:	快速开始	
	新建	取消

4. 单击新建。

新建节点并配置依赖关系

本节将在业务流程中创建一个虚拟节点(start)和odps_sql节点(insert_data),并将依赖关系 配置为insert_data依赖于start。



使用虚拟节点时,需要注意以下几点:

- · 虚拟节点属于控制类型节点,在业务流程运行过程中,不会对数据产生任何影响,仅用于实现 对下游节点的运维控制。
- · 虚拟节点在被其他节点依赖的情况下,如果被运维人员手动设置为运行失败,则下游未运行的 节点将因此无法被触发运行。在运维过程中,可以防止上游错误数据进一步蔓延。详情请参 见#unique_19。
- · 业务流程中,虚拟节点的上游节点,通常会设置为工作空间根节点。工作空间根节点的格式 为工作空间名_root。

综上所述,通常建议设计业务流程时,默认创建一个虚拟节点作为业务流程的根节点来控制整个工 作流。 1. 进入业务流程开发面板,并向面板中拖入一个虚拟节点,填写节点名称为start,单击节点名称。

新建节点			×
节点类型:	虚拟节点	Ý	
节点名称:	start		
目标文件夹:	业务流程/works	~	
		提交	取消

- 2. 向面板中拖入一个ODPS_SQL节点,填写节点名称为insert_data,单击提交。
- 3. 拖拽连线,将start节点设置为insert_data节点的上游节点。

6	X DataStudio	∼
	数据开发 2 日 0 0	🗸 works 🗙
(/)	Q 文件名称/创建人	E ⊙
≮ a	> 解決方案	◇ 节点组 C
Q	▶ 业务流程 器	◇ 数据集成
G	> ♣ workshop > ♣ works	□ 数据同步
۵		◇ 数据开发
≡	▼ 🕢 数据开发	Seq ODPS SQL
_	● 🔰 start 我锁定 07-04 16:07	Se ODPS Script
≞≏	● Sog insert_data 我锁定 07-11 11	Sep ODPS Spark
fx	▶ 🔳 表	Py PyODPS
_	> 🧭 资源	▼ 虚拟节点
	> 🔂 函数	
Σ	▶ 🧮 算法	Sh Shell
亩	▶ 👩 控制	P→ AnalyticDB for PostgreSQL
		AnalyticDB for MySQL
		O Data Lake Analytics

配置虚拟节点的上游依赖

在业务流程中,虚拟节点通常作为整个业务流程的控制器,是整个业务流程中所有节点的上游。通 常会设置业务流程中的虚拟节点依赖整个工作空间的根节点。

- 1. 双击虚拟节点,单击右侧的调度配置。
- 2. 单击使用工作空间根节点,设置虚拟节点的上游节点为工作空间根节点。

X 调度配置							调
调度周期:	B						配署
定时调度:							
具体时间:	00:20						版本
cron表达式:	00 20 00 * * ?						
依赖上—周期:							
调度依赖 ⑦ 一							
自动解析:	● 是 ● 否 解析输入输出						
広範的 上游节点・	法检入公共占给出2次式给出表2		(体田工作 空间相节占)				
1909013-1.03 137/// -			BO BETTELIENK DAM				
父节点輸出名称	父节点输出表名	节点名	父节点ID	责任人	来源	操作	

3. 配置完成后,单击左上角的 🛄 进行保存。

在ODPS_SQL节点中编辑代码

本节将在ODPS_SQL节点(insert_data)中,通过SQL代码,查询不同学历的单身人士贷款买房的数量,并将保存加过,以便后续节点继续分析或展现。

SQL语句如下所示,具体语法说明请参见#unique_20。

```
INSERT OVERWRITE TABLE result_table --数据插入到result_table中。
SELECT education
, COUNT(marital) AS num
FROM bank_data
WHERE housing = 'yes'
    AND marital = 'single'
GROUP BY education
```

运行并调试ODPS_SQL节点

1. 在insert_data节点中编辑好SQL语句后,单击保存,防止代码丢失。

2. 单击运行,查看运行日志和结果。

Sq insert_data •
<pre>1odps sql 2***********************************</pre>
15 16 运行日志
<pre>metrics_inner_time_ms:</pre>

提交业务流程

1. 运行并调试好ODPS_SQL节点insert_data后,返回业务流程页面,单击提交。

2. 在提交对话框中, 勾选需要提交的节点, 填写备注, 并勾选备注。

📕 works	×			
6		\mathbb{N}		
~ 节点组	C			
∨ 数据集				
回数据	提交			×
~ 数据				
Sq ODF	请选择节点		节点名称	
Sc ODF			start	
Sp ODF			insert_data	
Ру РуС	备注	works		
vī 虚兆				
Mr ODF		🔽 忽略輔	1入输出不一致的告 答	
sh She				
o–o Ana o–o Pos				
				提交取消

3. 单击提交。

后续步骤

现在,您已经学习了如何创建和提交业务流程,您可以继续学习下一个教程。在该教程中,您将学 习如何通过创建同步任务,将数据回流至不同类型的数据源中。详情请参见创建数据同步任务。

4 创建同步任务

本文将为您介绍如何创建同步任务,从而将MaxCompute中的数据导出至MySQL数据源中。

背景信息

在DataWorks中,通常通过数据集成功能,将系统中产生的业务数据定期导入至工作区。SQL任务 进行计算后,再将计算结果定期导出至您指定的数据源中,以便进一步展示或运行使用。



目前数据集成功能支持从RDS、MySQL、SQL

Server、PostgreSQL、MaxCompute、OCS、DRDS、OSS、Oracle、FTP、DM、HDFS和MongoDB等数据源中,将数据导入工作空间或将数据从工作空间导出。详细的数据源类型列表请参见#unique_22。

前提条件

- ·如果您使用的是ECS上自建的数据库,需要在自己的ECS添加安全组。详情请参见添加安全组。
- ·如果您使用的是RDS/MongoDB等数据源,需要在RDS/MongoDB等控制台添加白名单。详情 请参见添加白名单。

📕 说明:

如果是通过自定义资源组调度RDS的数据同步任务,必须把自定义资源组的机器IP也加入RDS的白名单中。

新增数据源



仅项目管理员角色可以新建数据源,其他角色的成员仅可查看数据源。

- 1. 以项目管理员身份登录DataWorks控制台,进入工作空间列表页面。
- 2. 单击相应工作空间操作下的进入数据集成。

- 3. 选择同步资源管理 > 数据源,单击新增数据源。
- 4. 在新增数据源弹出框中,选择数据源类型为MySQL。
- 填写新增MySQL数据源对话框中的配置,此处以创建连接串模式(数据集成网络可直接连通)类型为例。

新增MySQL数据源		×
* 数据源类型:	连接串模式 (数据集成网络可直接连通) ~	
* 数据源名称:	clone_database	
数据源描述:		
*适用环境:	✔ 开发 生产	
* JDBC URL :	jdbc:mysql://ServerIP:Port/Database	
* 用户名:	- 165	
* 密码 :		
测试连通性:	测试连通性	
0	确保数据库可以被网络访问	
	确保数据库没有被防火墙禁止	
	确保数据库域名能够被解析	
	确保数据库已经启动	
	上一步	完成

配置	说明
数据源类型	连接串模式(数据集成网络可直接连通)。
数据源名称	字母、数字、下划线组合,且不能以数字和下划线开头。例如 abc_123。
数据源描述	不超过80个字符。
适用环境	可以选择开发或生产环境。 道 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。				
	前 前 認 時 記 前 明 記 的 MySQL数据库对应的信息。不同数据源类型对 应 的 配 置 说明: 一 。 的 明 : 一 。 一 》 的 明 : 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》 》				

6. 单击测试连通性。

7. 如果测试连通性成功,单击完成。

确认作为目标的MySQL数据库中有表

在MySQL数据库中创建表odps_result, 建表语句如下所示:

```
CREATE TABLE `ODPS_RESULT` (
  `education` varchar(255) NULL ,
  `num` int(10) NULL
);
```

建表完成后,可以执行desc odps_result;语句, 查看表详情。

新建并配置同步节点

本节将新建一个同步节点write_result并进行配置,目的是用来把表result_table中的数据写入至

自己的MySQL数据库中。

1. 切换至数据开发面板,新建一个同步节点write_result。

A works X	
f) 🕑 🔍 🖈	
◇ 节点组	
~ 数据集成	
回 数据同步	
✓ 数据 ^开	
新建节点 Sa ODF	×
Sc ODF	
节点类型: Sp ODF	数据同步 🗸 🗸
节点名称: Py PyC	write_result
📊 📩 目标文件夹:	业务流程/works V
Mr ODF	
Sh She	

2. 设置write_result节点的上游节点为insert_data节点。

🛃 works 🗙	
Image:	
◇ 节点组	C
~ 数据集成	
▶ 数据同步	
◇ 数据开发	
Sq ODPS SQL	
Sc ODPS Script	
Sp ODPS Spark	
Py PyODPS	Vi start
Ⅵ 虚拟节点	
Mr ODPS MR	
Sh Shell	+
⊶ AnalyticDB for Or PostgreSQL	Sq insert_data
AnalyticDB for MySQL	
💿 Data Lake Analytic	Di write_result

3. 选择数据来源。

选择MaxCompute数据源及源头表result_table,单击下一步。

Di wr	ite_re:	sult	Sq insert	_deta	×	works												
면	۲		ſ	٤]			\$											ŝ
01	选择	款据源				款据并	ŧ源						救援去向					调度
				在	田配酒	数据的	来源油和写	入键:可	기운되다	的数据源。	地可以 長	1/2001	白右数探疫音	i ta ini ta	680 M			
						owonin o												版
		• 数据源	ODPS			odp	s_first	~	0			数据源:	数据源英型			0		
		•表	result_t	able				~										
	5	分区信息	: 无分区信	18														
		压缩	: 💿 不压	\$\$C)	围缩													
倅	宇符串	作为null	: 📀 是() 香														
					8	螺预览												

4. 选择数据去向。

选择MySQL数据源及目标表odps_result,单击下一步。

🛛 write_result 🌑	🔄 insert_data 🗙 🏯 works 🛛 🗙			≡
	1 I I I I I I			
01 选择数据源	救錫来源		数据去向	收起 調
	在这里配置数据的来源端和写入端:可	以导默认的数据源,也可以导感创建的	自有数据源音看支持的数据来源受司	「「「「」」の「「」」の「「」」の「「」」の「「」」の「「」」の「「」」の「
		-		版本
• 数据源:	ODPS v odps_first v	⑦ *数据源:	MySQL 🗸 works 🗸	0
*表:	result_table ~	* 表:	odps_result ~	
分区偏息:	无分区偏息	导入前准备语句:	请输入导入数据前执行的sql脚本	0
压缩:	● 不压缩 ○ 压缩			
空字符串作为null:	● 是 () 香	导入后完成语句:	请输入导入数据后执行的sol脚本	0
	数据预度			
		* 主變冲突:	insert into (当主键/约束冲突报脏数据) ~	

5. 字段映射。

选择字段的映射关系。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。

Di wr	ite_result	۰	Sq inse	rt_data	A. 1	works	,									
Ľ	۲	Þ	ſ↑]	٦			\$									
																洞度
02	字段映	4	>			源头#					目标表					配置
														NAR STRANGT		
			193	头表字段		炭	빞	Ø			目标表字段	英型		CHEIREIT JIRRENS		版本
			ed	ucation			RING	•		•	education	VARCHAR				
			nui			BIC	SINT	•		•	num	INT				
			澎	ம—ர∙												

6. 通道控制。

单击下一步,配置作业速率上限和脏数据检查规则。

03 通道控制 窓可以配置作业的传输速率和错误纪录数来控制整个数据同步过程:数据同步文档 ・任务期望最大并发数 3 ② ・日好速率 不現流 ● 限流 10 MB/s 错误记录数超过 脏数据条数范围,默认允许脏数据 条,任务自动结束 ② 任务资源組 新认资源组

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_26和#unique_27。

7. 预览保存。

完成上述配置后,上下滚动鼠标即可查看任务配置。确认无误,单击保存。

🔲 write_result 🌘	Sq insert_data	× 👗	works	×									≡
•••	f) [j]		- 9										
01 选择数据源			救据 来 源					数据去向					调度
	存	文田配留 約	如果的史源論	in写λ装:可L	计导致计的	教授酒、也可以寻知	BANISIZACI	白右数据源有异	专 编的 题 题	2. 大学 2. 大			配置
			A. BRIEF 3-7-828-3894	H-37 GHI (-34		Market 1 (21-3 Multiple)	-038243	H T3 X AHBA C C					版本
* 数据源:	ODPS		odps_first		0	* 30	अद्राप्त :	MySQL		works	0		
· *表:	result_table						*表:	odps_result					

提交数据同步任务

同步任务保存后,返回业务流程。单击顶部菜单栏中的提交,将同步任务提交至调度系统中。调度 系统会根据配置的属性,从第二天开始自动定时执行。

🚑 works 🗙 🗾 oss	数据同步
厨 ⊙ 🔍 🖈	<u>ا</u>
◇ 节点组 C	
✓ 数据集成	
□ 数据同步	
✓ 数据开发	
Sq ODPS SQL	Vi start
Sc ODPS Script	
Sp ODPS Spark	
Py PyODPS	
☑ 虚拟节点	Sq insert_data
Mr ODPS MR	
Sh Shell	
AnalyticDB for PostgreSQL	write_result
AnalyticDB for MySQL	

后续步骤

现在,您已经学习了如何创建同步任务,将数据导出至不同类型的数据源中,您可以继续学习下一 个教程。在该教程中,您将学习如何设置同步任务的调度属性和依赖关系。详情请参见设置周期和 依赖。

5 设置周期和依赖

本文将为您介绍DataWorks的调度配置和依赖。

📋 说明:

下文的操作以创建数据同步节点中创建的write_result为例,将其调度周期配置为周调度。

DataWorks具有强大的调度能力,支持根据时间、依赖关系的节点触发机制。DataWorks保障每 日千万级别的任务,根据DAG关系,准确、准时运行。且支持分钟、小时、天、周和月多种调度周 期配置,详情请参见调度配置介绍。

配置同步节点的调度属性

- 1. 创建数据同步节点后,双击需要配置的数据同步节点(write_result)。
- 2. 单击右侧的调度配置,为节点配置调度属性。

× 调度都	12 12		调度
生	;成实例方式:	● T+1次日生成 ○ 发布后即时生成	配置
	时间属性:	● 正常调度 ○ 空跑调度	
	出错重试:		版本
	生效日期:	1970-01-01 📅	
	暂停调度:		
	调度周期:	周 ~	
	定时调度:		
	指定时间:	星期二 × V	
	具体时间:	02.00	
	cron表达式 :	00 00 02 ?* 2	
依	赖上—周期:		

配置	说明
生成实例方式	可以根据自身需求,选择T+1次日生成或发布后即时生成。
时间属性	可以根据自身需求,选择正常调度或空跑调度。
出错重试	勾选后即开启。
生效日期	节点的有效日期,根据自身需求进行设置。
暂停调度	勾选后即为暂停状态。
调度周期	节点的运行周期(月/周/天/小时/分钟),例如以周为调度周期进行调 度。
定时调度	默认勾选定时调度。

配置	说明
指定时间/具体时间	指定节点运行的具体时间,例如将节点配置为在每周二的凌晨2点开始 运行。
cron表达式	默认为00 00 02 1 * ?,不可以更改。
依赖上一周期	根据自身需求,选择是否依赖上一周期。

配置数据同步节点的依赖属性

完成数据同步节点的调度属性的配置后,继续配置数据同步节点的依赖属性。

DI write_result ×	Sq insert_data 🗙	🏯 works	×								Ξ
• • •	1 1 🖲	6									
01 选择数据源	•	数据来源	资源组: default_								调度配
	在这里配置	收益的未透識	调度依赖 ⑦								Ē K
* 数据源:	ODPS ~	odps_firs	依赖的上游节点	请输入父节点题	出名称或输出表名		+	使用项目根节点			
*表:	result_table		父节点输出名称		父节点输出表 名	节点名	父节点) D	责任人	来源	操作	
分区信息:	无分区信息		_out	11322729		insert_da ta			手动添 加	删除	
压缩:	💽 不压缩 🔵 压缩										
空字符串作为 null	 ● 是 ○ 否 		本节点的输出	寄输入节点输出 谷		4					
		\$24版于REVEL	输出名称	输出表名	下游节。 称	点名 下諸	笻点ID	责任人	未渡	操作	
			bz.11322831_o t	- 0					系统 默 认源 加		

依赖属性中可以配置节点的上游依赖,表示即使当前节点的实例已经到定时时间,也必须等待上游 节点的实例运行完毕,才会触发运行。

如上图所示的配置表明:当前节点的实例将在上游insert_data节点的实例运行完毕后,才会触发执行。

在调度系统中,每一个项目中默认会创建一个projectname_root节点作为根节点。如果本节点没有上游节点,可以直接依赖根节点。

提交数据同步节点

保存数据同步节点write_result,单击提交,将其提交到调度系统中。

	⊙	Þ	ſ	ե	P	<u>ଜ</u>	
01	先择数据	源				数据来源	数据去向
						在这里配置数据的来源端和写入端;可以是默认的数据源,也可以是您创建	的自有数据源查看支持的数据来源类型

节点只有提交至调度系统中,才会从第二天开始,自动根据调度属性配置的周期,在各时间点生成 实例,然后定时运行。

如果是23:30以后提交的节点,则调度系统从第三天开始,才会自动周期生成实例并定时运行。

后续步骤

现在,您已经学习了如何设置数据同步节点的调度属性和依赖关系,您可以继续学习下一个教程。 在该教程中,您将学习如何对提交的节点进行周期运维,并查看日志排错。详情请参见运行及排 错。

6运行及排错

本文将为您介绍如何实现节点的运行、运维,并查看日志进行排错。

在设置周期和依赖的操作中,您配置了每周二凌晨2点执行同步节点。提交节点后,需要到第二天 才能看到调度系统自动执行的结果。DataWorks为您提供测试运行、补数据和周期运行3种触发方 式,帮助您确认实例运行的定时时间、相互依赖关系、数据结果产出是否符合预期。

- ·测试运行:手动触发方式。如果您仅需确认单个节点的定时情况和运行,建议您使用测试运行。 详情请参见测试运行。
- · 补数据运行: 手动触发方式。如果您需要确认多个节点的定时情况和相互依赖关系,或者需要从 某个根节点开始重新执行数据分析计算,建议您使用补数据运行。详情请参见补数据运行。
- 周期运行:系统自动触发方式。提交成功的节点,调度系统在第二天0点起会自动触发当天不同时间点的运行实例,并在定时时间达到时检查各实例的上游实例是否运行成功,如果定时时间已到并且上游实例全部运行成功,则当前实例会自动触发运行,无需人工干预。详情请参见周期运行。

▋ 说明:

手动触发和自动调度的调度系统与周期生成实例的规则一致。

- ・无论周期选择天/小时/分钟/月/周,节点在每一个日期都会有对应实例生成。
- · 仅在指定日期的对应实例,会定时运行并生成运行日志。
- ·非指定日期的对应实例不会实际运行,而是在满足运行条件时,将状态直接转换为成功。因此 不会有运行日志生成。

关于任务运维的更多操作和功能说明,请参见任务运维。

测试运行

1. 单击左上角的图标,选择全部产品 > 运维中心(工作流),进入运维中心页面。



2. 单击左侧导航栏中的周期任务,找到需要运行的节点。单击相应节点后的周期任务。

⑤ 🏶 运维中心								& DataStudio 🔍 😽	2
=									
③ 运维大屏	搜索: 节点名称/节点D Q, 解决	方案: 请选择	▶ 业务流程:	请选择 >	节点类型: 请选择	▼ 责任人:	请选择责任人 > 基线:	请选择 🖌	
ᢏ 任务列表		▶(冻结)节点	重置清空						
(字) 周期任务								○ 刷新 收起	搜索
③ 手动任务	名称	节点JD	修改日期↓	任务类型	责任人	调度类型	资源组 🏹 🧃	蹭 操作	
任务运维	10.000.000	700002549371	2019-07-04 16:34:36	ODPS_SQL	1000	日调度	默认资源组	DAG图 测试 补数据 🔻	更多 🔻
173周期实例		700002549370	2019-07-04 16:34:36	ODPS_SQL	100	日调度	默认资源组	DAG图 测试 补数据 🔻	更多 🔻

3. 在冒烟测试对话框中,填写冒烟测试名称,并选择业务日期,单击选择业务日期。

冒烟测试	×
如果业务日期选择昨天之前,则立即执行任务。	
如果业务日期选择昨天,则需等到定时时间才能执行任务。	
* 冒烟测试名称: P_rpt_u	
*选择业务日期: 2019-07-10	
	确定取消

4. 自动跳转至测试实例页面,查看节点的运行状态。

\$	🤗 运维中心	₽ DataStudio	থ্
•	运维大屏	200002549371 Q 予応課題 第近評 × またた 第近時また × 近行日離 2019-07-11 回 业务日席 第近時日期 回 近行状态 第述時	~
•	任务列表	基地 潮迅择 ✓ □ 我的节点 □ 我今天期试的节点 □ 智停(添给)节点 量置 清空	
6	周期任务	c	刷新 收
ß	手动任务	基本/原目 生产环境,请谨慎操作 C	⊕ ⊝
•	任务运维	2700 ~ (dur 0s)	
3	周期实例		
ß	手动实例		
R	测试实例		
5	补数据实例		

搜索: 基线:	70 Q 节点类型: 请选择 请选择 >	✓ 责任人: 请选择责任人 ✓ 试的节点 目智停(冻结)节点 重量	运行日期: 2019-07-11 清空	並务日期: 请选择日	期 箇 运行状态:	(○) 前所 收起換案
	基本值息 ②t #70t 07-11153454 ~ 153454 (dur 0s)			生产环境,清谨慎操作		୯ ହେ ରୁ ଜ
	ã.▼ < 1/1 >			重都近行日志 重都近行日志 重着代码 嶋城市県 重重加添 現し進行 重成の 暫停(所法) 防災(所法)	节点00 节点名称 调度类型 更任人 运行状态 所配工作空间 开始时间 结束时间	2 700002547993 ● workhopstart 日間度 运行成功 E est_workahop001 2 2019-07-11 15:34-54 金目更多详情

5. 选择测试实例页面中相应的实例并单击,即可看到实例DAG图。

・右键单击实例,可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。
 ・双击实例,即可弹出节点的属性、运行日志、操作日志、代码等信息。

📕 说明:

- ·测试运行是手动触发节点,只要到定时的时间,立即运行,自动忽略实例的上游依赖关系。
- ·根据前文所述的实例生成规则,配置为每周二凌晨2点运行的节点write_result,测试运行
 时选择的业务日期是周一(业务日期=运行日期-1),实例会在2点真正运行。如果不是周一,则实例在2点转换为成功状态,且没有日志生成。

补数据运行

如果需要确认多个节点的定时情况和相互依赖关系,或者需要从某个根节点开始重新执行数据分析 计算,可以进行补数据运行。

- 1. 进入运维中心 > 任务列表 > 周期任务页面。
- 2. 单击相应节点后的补数据 > 当前节点。

⑤ ※ 运维中心	•							& DataStudio 🖏 🔻	
Ξ									
③ 运维大屏	搜索: 节点名称/节点ID Q,	解决方案: 请选择	∨ 业务流程:	请选择	▼ 节点类型: 请选择	▼ 责任人:	请选择责任人 >	基线 请选择 >	
ᇦ 任务列表	1 我的节点 1 今日修改的节点	暂停(冻结)节点 重	置清空						
③ 周期任务								○ 刷新 收起搜	续
③ 手动任务	名称	节点ID	化胰日始剂	任务类型	责任人	调度类型	资源组 🍸	报警 操作	
✔ 任务运维		700002549371	2019-07-04 16:34:36	ODPS_SQL	-	日调度	默认资源组	DAG图 测试 补数据 ▼ 更	18 🔻
13、周期实例		700002549370	2019-07-04 16:34:36	ODPS_SQL	-	日调度	默认资源组	DAG图 测试 补数据 ▼ 更	iø 🔻
		700002549369	2019-07-04 16:34:36	ODPS_SQL	1000	日调度	默认资源组	当前节点	18 🔻
③ 手动实例		700002549359	2019-07-04 16:34:35	数据集成	-	日调度	同步资源组 默认资源组	当前节点及下游节点	18 -
		700002549360	2019-07-04 16:34:35	数据集成	100	日调度	同步资源组: 默认资源组	海重口 只模式	18 🔻
		700002547993	2019-07-04 16:34:34	虚节点	1.00	日调度	默认资源组	DAG图 测试 补数据 ▼ 更	搜索 更多 ▼ 更多 ▼ 更多 ▼ 更多 ▼ 更多 ▼ 更多 ▼ 更多 ▼

3. 填写补数据对话框中的配置, 单击确定。

补数据		×
* 补数据名称:	P_write_result_20180723_221754	
* 选择业务日期:	2018-07-15 - 2018-07-22 🗇	
* 当前任务:	write_result	
* 是否并行:	不并行 🗸	
		确定取消

配置	说明
补数据名称	填写补数据名称。
选择业务日期	选择补数据的业务日期为2018-07-15到2018- 07-22。
当前任务	默认为当前节点,不可以更改。
是否并行	可以选择不并行或指定允许几组任务同时运 行。

- 4. 自动跳转至补数据实例页面,单击相应的实例,即可看到实例DAG图。
 - ・右键单击实例,可以查看该实例的依赖关系和详细信息,并进行终止运行、重跑等具体操作。
 - ・双击实例,即可弹出节点的属性、运行日志、操作日志、代码等。

- 说明:

- ·补数据任务的实例依赖前一天,例如补2017-09-15到2017-09-18时间段内的任务,如果15 号的实例运行失败了,则16号的实例也不会运行。
- ·根据前文所述的实例生成规则,配置为每周二凌晨2点运行的节点write_result,补数据运行时选择的业务日期是周一(业务日期=运行日期-1),实例会在2点真正运行。如果不是周一,则实例在2点转换为成功状态,且没有日志生成。

周期自动运行

周期自动运行,由系统根据所有节点的调度配置自动触发,所以页面没有操作入口。您可以通过以 下两种方式查看实例信息和运行日志:

- · 进入运维中心 > 任务运维 > 周期实例页面,选择业务日期或运行日期等参数,搜 索write_result节点对应的实例,然后右键查看实例信息和运行日志。
- ·选择周期实例页面中相应的节点实例并单击,即可看到实例DAG图。
 - 右键单击实例,可以查看该实例的依赖关系和详细信息并进行终止运行、重跑等具体操作。
 - 双击实例,即可弹出节点的属性、运行日志、操作日志、代码等。

🗾 说明:

- 如果上游节点未运行,下游节点也不会运行。
- 如果节点的实例初始状态为未运行,当定时时间到达时,调度系统会检查该实例的全部上游 实例是否运行成功。
- 只有上游实例全部运行成功,且定时时间到达的实例,才会被触发运行。
- 处于未运行状态的实例,请确认上游实例已经全部成功且已到定时时间。

7 使用临时查询快速查询SQL(可选)

如果您已经创建了MaxCompute项目(DataWorks工作空间),可以直接使用DataWorks临时 查询功能,快速书写SQL语句操作MaxCompute。

关于临时查询功能的具体信息,请参见#unique_35。

进入临时查询

点击DataWorks控制台工作空间列表,选择您需要进入的项目,点击进入数据开发。

= (-)阿里云	华东1(杭州)▼	Q 搜索		義用	工单 音	宴 企业	支持与服务	>_	۵.	Ħ	0 8	简体中文	0
		概览 工作空间	列表 资源列表 计	算引擎列表									
请输入工作空间/显示名	搜索										创建工	作空间 剧赛	例表
工作空间名称/显示名	模式	创建时间	管理员		状态	Я	通服务		操作				
-	标准模式 (开发跟生产隔离)	2019-07-26 17:10:46	101010-0010-0		正常	0	• 🔨		工作空间 进入数据	配置 进 集成 进	主入数据开发 主入数据服务	修改服务 更多 👻	
	简单模式(单环境)	2019-05-30 11:40:00	Sector Sector		I.S	0	o €.		工作空间 进入数据	配置 进 集成 进	主入数据开发 主入数据服务	修改服务 更多 ▼	

直接点击临时查询,右键临时查询,点击新建节点 > ODPS SQL。

Data	DataStudio
•	文件名称/创建人
*	✓ 临时查☆ 新建 节点 > ODPS SQL
Q	Solution 新建文件夹 Shell
ତ	
×	
▦	
I	
fx	<
Σ	
Ξ	

在弹框中输入节点名称,点击提交,创建您的临时查询节点。

节点类型: ODPS SQL ~ /	
节点名称: test1	
目标文件夹: 临时查询 ~	
した。 して、「たい」の「たい」の「たい」の「たい」の「たい」の「たい」の「たい」の「たい」の	

运行SQL

现在您可以在刚刚创建的临时查询节点中运行MaxCompute支持的SQL语句了,我们以运行一个DDL语句创建表为例。

输入建表语句,点击运行即可。

```
create table if not exists sale_detail
(
shop_name string,
customer_id string,
total_price double
)
partitioned by (sale_date string,region string);
-- 创建一张分区表sale_detail
```



在弹框中您可以看到本次运行的费用预估,继续点击运行。

成本估计	×
▲ 按量付费用户每次运行都会产生相应费用,请谨慎进行。小于1分钱按1分钱估算,实际以账单为准	
sql语句	预估费用
create table if not exists sale_detail (shop_name string, customer_id string, total_price double) partitioned b	¥ 0 RMB
	运行取消

您可以在下方的日志窗口,看到运行情况和最终结果:本次运行成功,结果为OK。

 odps sql odps sql wither: create time: 2019-03-20 11:02:49 minimum estring, create table if not exists sale_detail	Sq test1													
<pre>1odps sql 2</pre>			∢	►		C	20	\$						
<pre>4create time:2019-03-20 11:02:49 5************************************</pre>	1 2 3	odj *** aut	ps sq1 ***** thor:] *****										
<pre>6 create table if not exists sale_detail 7 (8 shop_name string, 9 customer_id string, 10 total_price double 11) 12 partitioned by (sale_date string,region string); 12 partitioned by (sale_date string,region string); 12 partitioned by (sale_date string,region string); 13 partitioned by (sale_date string,region string); 14 partitioned by (sale_date string,region string); 15 partitioned by (sale_date string,region string); 16 cm cm</pre>	4 5	CP(eate 1 *****	time:2 *****	019-03 *****	3-20 1 *****	1:02: ****							
8 shop_name string, 9 customer_id string, 10 total_price double 11) 12 partitioned by (sale_date string, region string); 运行日志 2019-03-20 11:09:38 start to get jobId: 2019-03-20 11:09:38 get jobid:2019032003093 ID = 2019032003093 OK 2019-03-20 11:09:38 INFO	6 7	creat	te tal	ole if	not e	exists	sale	_detail						
12 partitioned by (sale_date string, region string); 运行日志 2019-03-20 11:09:38 start to get jobId: 2019-03-20 11:09:38 get jobid:2019032003093 ID = 2019032003093 OK 2019-03-20 11:09:38 INFO ====================================	8 9 10 11	snop custo tota	_name omer_i l_prio	ids ced	tring tring ouble	,								
运行日志 2019-03-20 11:09:38 start to get jobId: 2019-03-20 11:09:38 get jobid:2019032003093 ID = 2019032003093 OK 2019-03-20 11:09:38 INFO	12	part:	itione	ed by	(sale	_date	strin	g,regio	n strin	g);				
2019-03-20 11:09:38 start to get jobId: 2019-03-20 11:09:38 get jobid:2019032003093 ID = 2019032003093 OK 2019-03-20 11:09:38 INFO ====================================	运行	志												
2019-03-20 11:09:38 INFO ====================================	2019-03 2019-03 ID = 20 OK	-20 11 -20 11 190320	:09:38 :09:38 03093	start get jo	to get bid:20	jobId: 1903200	3093							
2017-03-20 11.07.30 1000 0050 0100 15. 2.3435	2019-03 2019-03 2019-03 2019-03 2019-03 2019-03	-20 11 -20 11 -20 11 -20 11 -20 11 -20 11	:09:38 :09:38 :09:38 :09:38 :09:38 :09:38	INFO = INFO E INFO - INFO S INFO C INFO C	Exit co Inv Shell r Current	eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee	entering the She of Sh tessful tatus: 2.345s	======= ell comma ly! FINISH	and O	.eted	-	 		

使用同样的方法,您也可以执行查询语句。