ALIBABA CLOUD

阿里云

数据湖分析

DataWorks

文档版本：20211117

阿里云

法律声明

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

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

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

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

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

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

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

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

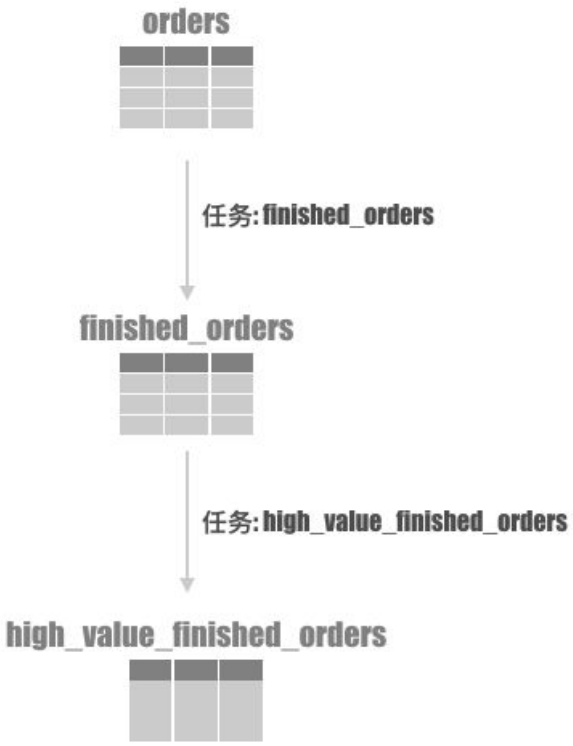
目录

1.调度DLA Presto任务

DataWorks是基于MaxCompute作为计算和存储引擎的用于工作流可视化开发和托管调度运维的海量数据离线加工分析平台，支持按照时间和依赖关系的任务全面托管调度。本文主要介绍如何通过DataWorks调度DLA任务。

目的

任务调度中一个重要的功能是任务之间的依赖，为演示这个功能，本文会在DataWorks中创建两个DLA任务。表和任务之间的关系如下图所示：



- **任务一**

  从orders表查询出已经完成的订单：`o_orderstatus = 'F'`，并将其写入`finished_orders`表。

- **任务二**

  从finished orders中查询出总价大于10000的订单：`o_totalprice > 10000`，并将其写入`high_value_finished_orders`表。

本文中源数据orders表存储在OSS dlaossfile1 Bucket的dla Object中，即`oss://dlaossfile1/dla/`，单击orders.txt下载源数据。空文件finished_orders.txt和high_value_finished_orders.txt存储位置为`oss://dlaossfile1/dla/finished_orders/`。

前提条件

1. 您已经开通DLA、DataWorks以及OSS服务，且DLA、DataWorks、OSS所属Region相同。本文中三个服务所属Region均为华东1（杭州）。
2. 创建项目空间，详情请参见创建工作空间，本教程项目空间为liujing_dataworks_test。
3. 在DLA中创建服务访问点，详情请参见设置服务访问点。

操作步骤

1. 在DLA中创建OSS Schema

```
CREATE SCHEMA dataworks_demo with DBPROPERTIES(
 CATALOG = 'oss',
LOCATION = 'oss://dlaossfile1/dla/');
```

2. 在DLA中创建指向OSS文件的外表
   - orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS orders (
  O_ORDERKEY INT,
  O_CUSTKEY INT,
  O_ORDERSTATUS STRING,
  O_TOTALPRICE DOUBLE,
  O_ORDERDATE DATE,
  O_ORDERPRIORITY STRING,
  O_CLERK STRING,
  O_SHIPPRIORITY INT,
  O_COMMENT STRING
)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
 STORED AS TEXTFILE LOCATION 'oss://dlaossfile1/dla/';
```

   - finished_orders表

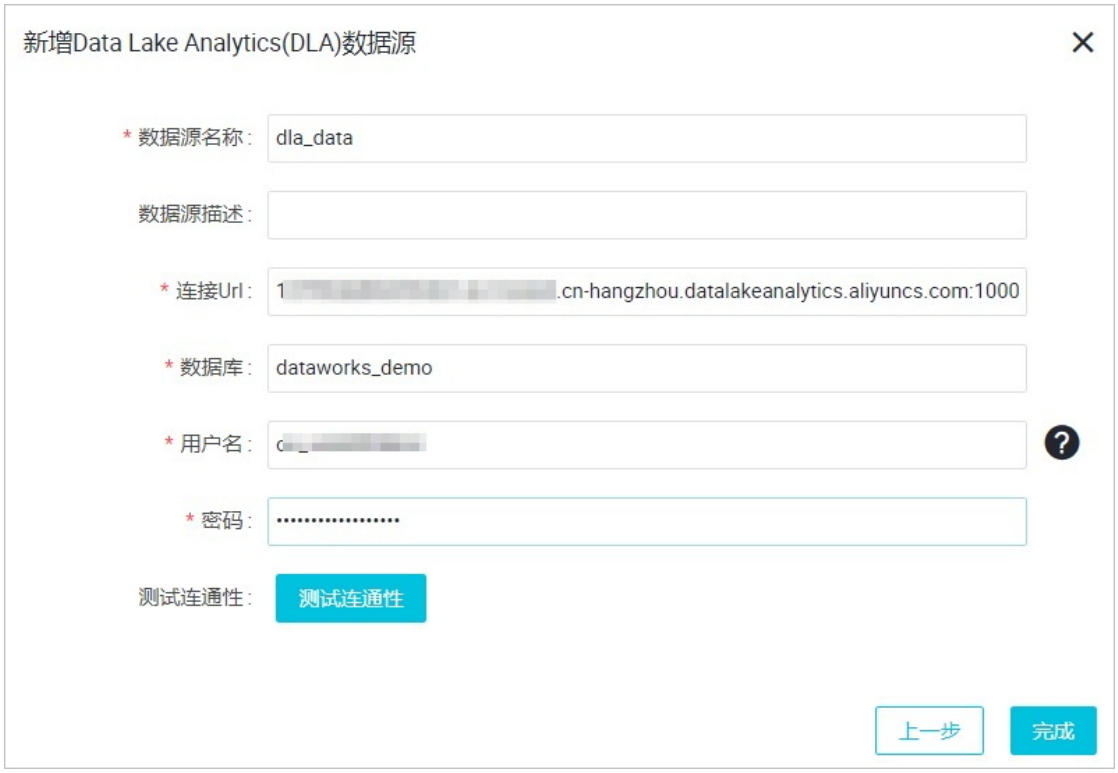
```
CREATE EXTERNAL TABLE IF NOT EXISTS finished_orders (
   O_ORDERKEY INT,
    O_TOTALPRICE DOUBLE
)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
 STORED AS TEXTFILE
 LOCATION 'oss://dlaossfile1/dla/finished_orders/';
```

   - high_value_finished_orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS high_value_finished_orders (
  O_ORDERKEY INT,
   O_TOTALPRICE DOUBLE
)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
 LOCATION 'oss://dlaossfile1/dla/finished_orders/';
```

3. 在DataWorks中添加DLA数据源
   i. 登录DataWorks控制台，单击对应项目栏中的**进入数据集成**。
   ii. 单击**新增数据源**，数据源选择**Data Lake Analytics(DLA)**。

iii. 在**新增Data Lake Analytics(DLA)数据源**页面，进行参数配置。



参数配置说明如下表所示：

数据源名称	为数据源指定一个名字，便与后续管理。
数据源描述	添加数据源描述，该项为可选填项。
连接Url	DLA的服务访问点地址信息，由 `Address:Port` 组成。可通过设置服务访问点获取 `Address:Port` 。
数据库名	DLA中创建的OSS连接，本教程为dataworks_demo。
用户名	登录DLA使用的用户名。
密码	登录DLA使用的用户名对应的密码。

iv. 添加DataWorks中的沙箱白名单

前往DataWorks的配置页面： https://setting-cn-hangzhou.data.aliyun.com/#/project在安全设置区域，有一个“沙箱白名单“配置，需要把我们刚刚添加的DLA的域名信息添加进去

v. 修改DLA白名单。

由于DataWorks中对DLA数据源有白名单限制，您需要根据DLA所属Region，将下表中对应的IP地址或者IP地址段加入DLA的白名单。

华东1（杭州）	100.64.0.0/8,11.193.102.0/24,11.193.215.0/24,11.194.110.0/24,11.194.73.0/24,118.31.157.0/24,47.97.53.0/24,11.196.23.0/24,47.99.12.0/24,47.99.13.0/24,114.55.197.0/24,11.197.246.0/24,11.197.247.0/24

华东2（上海）	11.193.109.0/24,11.193.252.0/24,47.101.107.0/24,47.100.129.0/24,106.15.14.0/24,10.117.28.203,10.117.39.238,10.143.32.0/24,10.152.69.0/24,10.153.136.0/24,10.27.63.15,10.27.63.38,10.27.63.41,10.27.63.60,10.46.64.81,10.46.67.156,11.192.97.0/24,11.192.98.0/24,11.193.102.0/24,11.218.89.0/24,11.218.96.0/24,11.219.217.0/24,11.219.218.0/24,11.219.219.0/24,11.219.233.0/24,11.219.234.0/24,118.178.142.154,118.178.56.228,118.178.59.233,118.178.84.74,120.27.160.26,120.27.160.81,121.43.110.160,121.43.112.137,100.64.0.0/8
华南1（深圳）	100.106.46.0/24,100.106.49.0/24,10.152.27.0/24,10.152.28.0/24,11.192.91.0/24,11.192.96.0/24,11.193.103.0/24,100.64.0.0/8,120.76.104.0/24,120.76.91.0/24,120.78.45.0/24
中国香港	10.152.162.0/24,11.192.196.0/24,11.193.11.0/24,100.64.0.0/8,11.192.196.0/24,47.89.61.0/24,47.91.171.0/24,11.193.118.0/24,47.75.228.0/24
亚太东南1（新加坡）	100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.153.0/24,11.192.40.0/24,11.193.8.0/24,100.64.0.0/8,100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.40.0/24,47.88.147.0/24,47.88.235.0/24,11.193.162.0/24,11.193.163.0/24,11.193.220.0/24,11.193.158.0/24,47.74.162.0/24,47.74.203.0/24,47.74.161.0/24,11.197.188.0/24
亚太东南2（澳洲，悉尼）	11.192.100.0/24,11.192.134.0/24,11.192.135.0/24,11.192.184.0/24,11.192.99.0/24,100.64.0.0/8,47.91.49.0/24,47.91.50.0/24,11.193.165.0/24,47.91.60.0/24
华北2（北京）	100.106.48.0/24,10.152.167.0/24,10.152.168.0/24,11.193.50.0/24,11.193.75.0/24,11.193.82.0/24,11.193.99.0/24,100.64.0.0/8,47.93.110.0/24,47.94.185.0/24,47.95.63.0/24,11.197.231.0/24,11.195.172.0/24,47.94.49.0/24,182.92.144.0/24
美国西部1	10.152.160.0/24,100.64.0.0/8,47.89.224.0/24,11.193.216.0/24,47.88.108.0/24
美国东部1	11.193.203.0/24,11.194.68.0/24,11.194.69.0/24,100.64.0.0/8,47.252.55.0/24,47.252.88.0/24
亚太东南3（马来西亚，吉隆坡）	11.193.188.0/24,11.221.205.0/24,11.221.206.0/24,11.221.207.0/24,100.64.0.0/8,11.214.81.0/24,47.254.212.0/24,11.193.189.0/24

欧洲中部1（德国，法兰克福）	11.192.116.0/24,11.192.168.0/24,11.192.169.0/24,11.192.170.0/24,11.193.106.0/24,100.64.0.0/8,11.192.116.14,11.192.116.142,11.192.116.160,11.192.116.75,11.192.170.27,47.91.82.22,47.91.83.74,47.91.83.93,47.91.84.11,47.91.84.110,47.91.84.82,11.193.167.0/24,47.254.138.0/24
亚太东北1（日本）	100.105.55.0/24,11.192.147.0/24,11.192.148.0/24,11.192.149.0/24,100.64.0.0/8,47.91.12.0/24,47.91.13.0/24,47.91.9.0/24,11.199.250.0/24,47.91.27.0/24
中东东部1（阿联酋，迪拜）	11.192.107.0/24,11.192.127.0/24,11.192.88.0/24,11.193.246.0/24,47.91.116.0/24,100.64.0.0/8
亚太东南1（印度，孟买）	11.194.10.0/24,11.246.70.0/24,11.246.71.0/24,11.246.73.0/24,11.246.74.0/24,100.64.0.0/8,149.129.164.0/24,11.194.11.0/24
英国	11.199.93.0/24,100.64.0.0/8
亚太东南5（印度尼西亚，雅加达）	11.194.49.0/24,11.200.93.0/24,11.200.95.0/24,11.200.97.0/24,100.64.0.0/8,149.129.228.0/24,10.143.32.0/24,11.194.50.0/24
华北2（政务云）	11.194.116.0/24,100.64.0.0/8 如果IP地址段添加不成功，请添加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

vi. 完成上述参数配置后，单击**测试连通性**进行连通性测试，测试通过后单击**完成**。

4. **在DataWorks中创建DLA调度任务**

   i. 登录DataWorks控制台，单击对应项目栏中的**进入数据开发**。

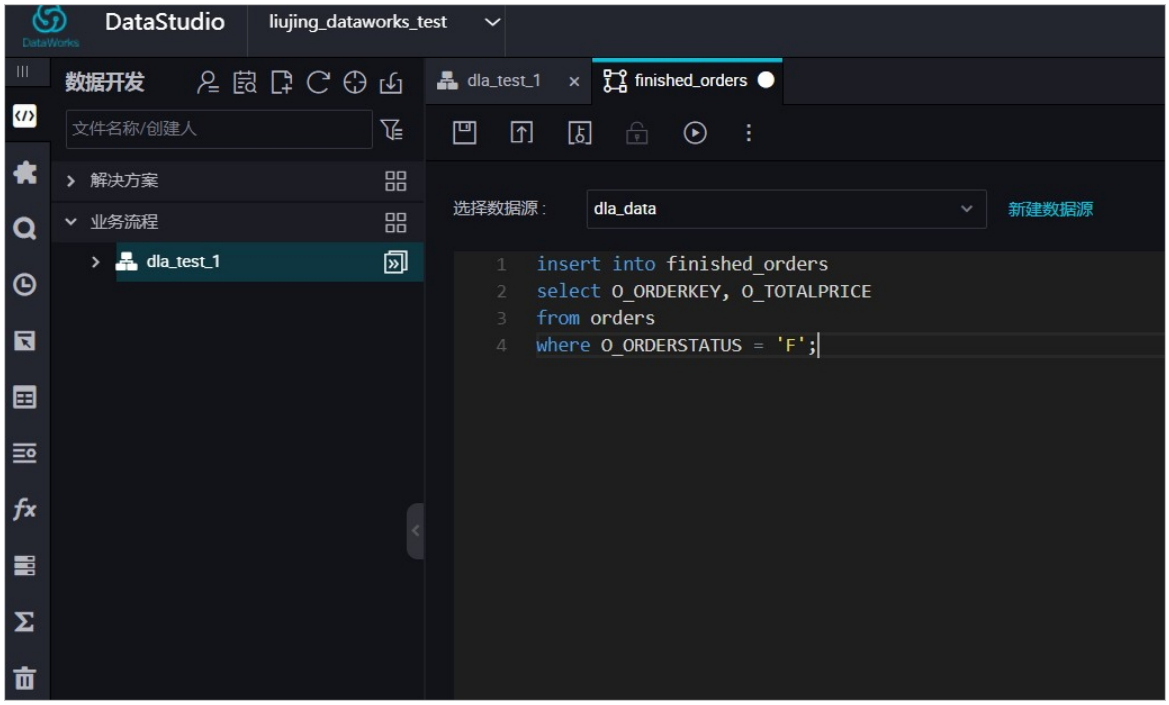
   ii. 右键单击**业务流程**新建一个流程，本教程新建业务流程dla_test_1。

   iii. 新建一个DLA任务，本教程为finished_orders。

   iv. 单击DLA任务finished_orders，**选择数据源**设置为DLA数据源。

   > **说明** 可参考上述步骤，创建多个任务。本文创建两个任务：finished_orders、high_value_finished_orders。

5. **在DataWorks中执行DLA调度任务**



- 任务一：从orders表查询出已经完成的订单：`o_orderstatus = 'F'`，并将其写入finished_orders表。

```
insert into finished_orders
select O_ORDERKEY, O_TOTALPRICE
from orders
where O_ORDERSTATUS = 'F';
```

- 任务二：从finished_orders中查询出总价大于10000的订单：`o_totalprice > 10000`，并将其写入high_value_finished_orders表。

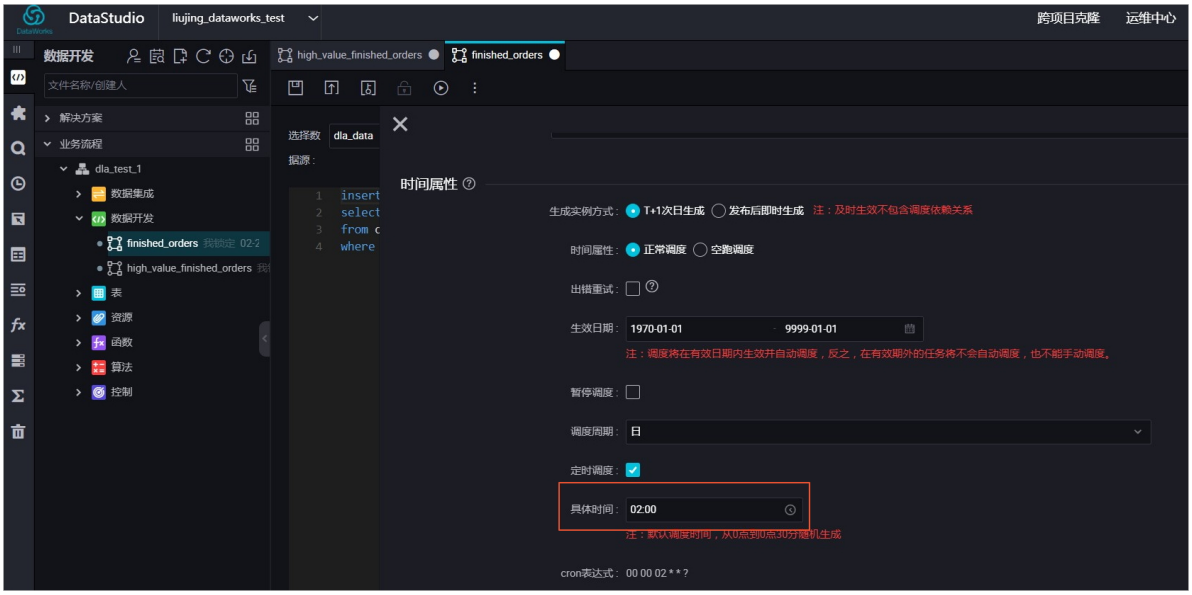
```
insert into high_value_finished_orders
select * from finished_orders
where O_TOTALPRICE > 10000;
```

后续操作

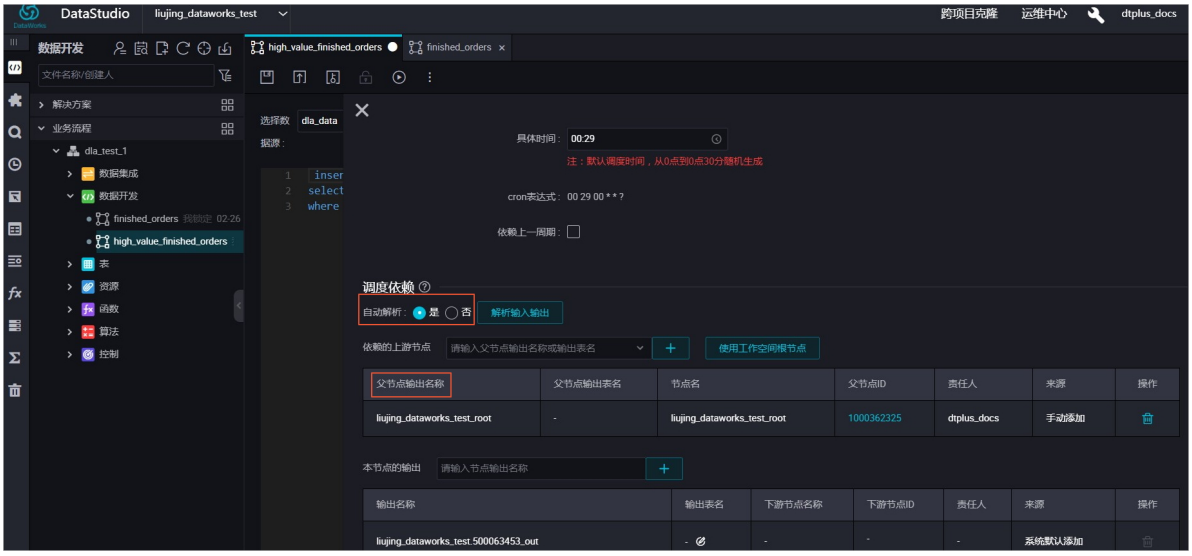
- **任务配置**

  DataWorks支持按照时间、依赖关系的任务触发机制，支持多个任务按照指定的依赖关系在指定的时间运行。

  例如，每天凌晨2点运行finished_orders任务：



finished_orders任务成功运行之后再运行high_value_finished_orders任务：



- **任务发布**

  任务配置好之后，就可以进行任务的发布和运维，详情请参见发布任务。

2.循环调度DLA Presto任务

2.1. 背景信息以及准备工作

本文档主要介绍了循环调度DLA Presto任务的背景信息和准备工作。

背景信息

DLA作为无服务化的大数据分析服务，通过标准的SQL语句直接对存储在阿里云对象存储服务（Object Storage Service，简称 OSS）、表格存储（Table Store）中的数据进行清洗。例如，使用DLA对OSS中的历史数据按天进行清洗。DataWorks是基于MaxCompute作为计算和存储引擎的用于工作流可视化开发和托管调度运维的海量数据离线加工分析平台，支持按照时间和依赖关系的任务全面托管调度。DLA用户可以通过DataWorks强大的任务托管调度功能，调度执行DLA任务，使用DataWorks、任务依赖关系管理、任务运维等全方位强大的功能。

在使用DLA对OSS中的历史数据按天进行清洗时，由于数据清洗的SQL是固定的，只是每次执行的时候需要传入不同的日期，因此我们可以通过DataWorks来循环调度DLA数据清洗任务。针对上述场景，我们需要在DataWorks中完成以下工作：

- 部署一个赋值节点，该节点负责输出日期值，并作为下游循环节点的输入。
- 部署一个循环节点，该节点包含用来做数据清洗的一个或者一组SQL。其中，日期取值是一个变量，每次循环的输入值由赋值节点提供。

任务调度中一个重要的功能是任务之间的依赖，为演示这个功能，本教程以在DataWorks中循环调度从orders表查询出已经完成的订单`o_orderstatus = 'F'`，并将其写入finished_orders表为例，为您介绍详细的实施步骤。

准备工作

使用DataWorks循环调度DLA任务之前，您需要先通过以下操作在OSS中准备测试数据、在DLA中创建OSS Schema和表以及创建DataWorks项目空间。

步骤一：在OSS中准备测试数据

1. 开通OSS服务
2. 创建存储空间
3. 上传测试数据

   单击orders下载测试数据。

步骤二：创建OSS Schema

```
CREATE SCHEMA dataworks_demo with DBPROPERTIES(
 CATALOG ='oss',
 LOCATION ='oss://bucket-name/dla/'
);
```

location：文件所在的OSS Bucket的目录，需以`/`结尾。

步骤三：创建OSS表

- orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS orders (
  O_ORDERKEY INT,
  O_CUSTKEY INT,
  O_ORDERSTATUS STRING,
  O_TOTALPRICE DOUBLE,
  O_ORDERDATE DATE,
  O_ORDERPRIORITY STRING,
  O_CLERK STRING,
  O_SHIPPRIORITY INT,
  O_COMMENT STRING
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
STORED AS TEXTFILE
LOCATION 'oss://bucket-name/dla/';
```

- finished_orders表

```
CREATE EXTERNAL TABLE IF NOT EXISTS finished_orders (
   O_ORDERKEY INT,
   O_TOTALPRICE DOUBLE
)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY '|'
 STORED AS TEXTFILE
 LOCATION 'oss://bucket-name/dla/finished_orders/';
```

步骤四：创建DataWorks项目空间

参考DataWorks准备工作，创建项目空间。本教程项目空间为dla_project。

2.2. 实施步骤

本文档主要介绍了循环调度DLA Presto任务的操作步骤。

步骤一：新增Data Lake Analytics（DLA）数据源

1. 登录DataWorks控制台，单击对应项目栏中的**进入数据集成**。
2. 单击**新增数据源**，数据源选择**Data Lake Analytics（DLA）**。
3. 在**新增Data Lake Analytics（DLA）数据源**页面，进行以下参数配置。详细的参数配置如下表所示：

参数名称	参数说明
数据源名称	为数据源指定一个名字，便与后续管理。
数据源描述	添加数据源描述，该项为可选填项。
连接Url	DLA的服务访问点地址信息，由 Address:Port 组成。可通过设置服务访问点获取 Address:Port 。
数据库名	DLA中创建的OSS连接，本教程为dataworks_demo。
用户名	登录DLA使用的用户名。
密码	登录DLA使用的用户名对应的密码。

4. 修改DLA白名单。

   由于DataWorks中对DLA数据源有白名单限制，您需要根据DLA所属Region，将下表中对应的IP地址或者IP地址段加入DLA的白名单。

华东1（杭州）	100.64.0.0/8,11.193.102.0/24,11.193.215.0/24,11.194.110.0/24,11.194.73.0/24,118.31.157.0/24,47.97.53.0/24,11.196.23.0/24,47.99.12.0/24,47.99.13.0/24,114.55.197.0/24,11.197.246.0/24,11.197.247.0/24

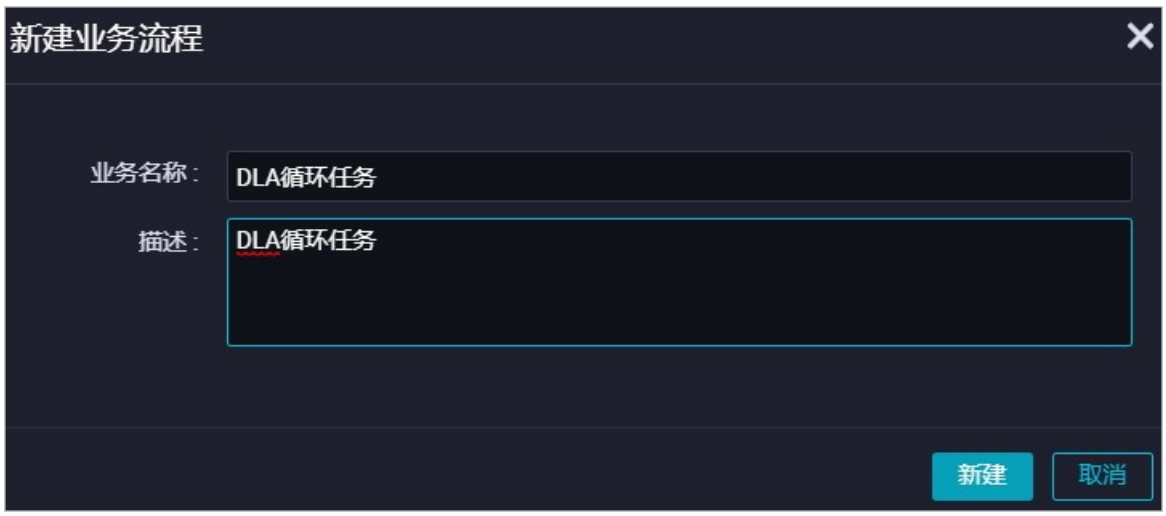
华东2（上海）	11.193.109.0/24,11.193.252.0/24,47.101.107.0/24,47.100.129.0/24,106.15.14.0/24,10.117.28.203,10.117.39.238,10.143.32.0/24,10.152.69.0/24,10.153.136.0/24,10.27.63.15,10.27.63.38,10.27.63.41,10.27.63.60,10.46.64.81,10.46.67.156,11.192.97.0/24,11.192.98.0/24,11.193.102.0/24,11.218.89.0/24,11.218.96.0/24,11.219.217.0/24,11.219.218.0/24,11.219.219.0/24,11.219.233.0/24,11.219.234.0/24,118.178.142.154,118.178.56.228,118.178.59.233,118.178.84.74,120.27.160.26,120.27.160.81,121.43.110.160,121.43.112.137,100.64.0.0/8
华南1（深圳）	100.106.46.0/24,100.106.49.0/24,10.152.27.0/24,10.152.28.0/24,11.192.91.0/24,11.192.96.0/24,11.193.103.0/24,100.64.0.0/8,120.76.104.0/24,120.76.91.0/24,120.78.45.0/24
中国香港	10.152.162.0/24,11.192.196.0/24,11.193.11.0/24,100.64.0.0/8,11.192.196.0/24,47.89.61.0/24,47.91.171.0/24,11.193.118.0/24,47.75.228.0/24
亚太东南1（新加坡）	100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.153.0/24,11.192.40.0/24,11.193.8.0/24,100.64.0.0/8,100.106.10.0/24,100.106.35.0/24,10.151.234.0/24,10.151.238.0/24,10.152.248.0/24,11.192.40.0/24,47.88.147.0/24,47.88.235.0/24,11.193.162.0/24,11.193.163.0/24,11.193.220.0/24,11.193.158.0/24,47.74.162.0/24,47.74.203.0/24,47.74.161.0/24,11.197.188.0/24
亚太东南2（澳洲，悉尼）	11.192.100.0/24,11.192.134.0/24,11.192.135.0/24,11.192.184.0/24,11.192.99.0/24,100.64.0.0/8,47.91.49.0/24,47.91.50.0/24,11.193.165.0/24,47.91.60.0/24
华北2（北京）	100.106.48.0/24,10.152.167.0/24,10.152.168.0/24,11.193.50.0/24,11.193.75.0/24,11.193.82.0/24,11.193.99.0/24,100.64.0.0/8,47.93.110.0/24,47.94.185.0/24,47.95.63.0/24,11.197.231.0/24,11.195.172.0/24,47.94.49.0/24,182.92.144.0/24
美国西部1	10.152.160.0/24,100.64.0.0/8,47.89.224.0/24,11.193.216.0/24,47.88.108.0/24
美国东部1	11.193.203.0/24,11.194.68.0/24,11.194.69.0/24,100.64.0.0/8,47.252.55.0/24,47.252.88.0/24
亚太东南3（马来西亚，吉隆坡）	11.193.188.0/24,11.221.205.0/24,11.221.206.0/24,11.221.207.0/24,100.64.0.0/8,11.214.81.0/24,47.254.212.0/24,11.193.189.0/24

欧洲中部1（德国，法兰克福）	11.192.116.0/24,11.192.168.0/24,11.192.169.0/24,11.192.170.0/24,11.193.106.0/24,100.64.0.0/8,11.192.116.14,11.192.116.142,11.192.116.160,11.192.116.75,11.192.170.27,47.91.82.22,47.91.83.74,47.91.83.93,47.91.84.11,47.91.84.110,47.91.84.82,11.193.167.0/24,47.254.138.0/24
亚太东北1（日本）	100.105.55.0/24,11.192.147.0/24,11.192.148.0/24,11.192.149.0/24,100.64.0.0/8,47.91.12.0/24,47.91.13.0/24,47.91.9.0/24,11.199.250.0/24,47.91.27.0/24
中东东部1（阿联酋，迪拜）	11.192.107.0/24,11.192.127.0/24,11.192.88.0/24,11.193.246.0/24,47.91.116.0/24,100.64.0.0/8
亚太东南1（印度，孟买）	11.194.10.0/24,11.246.70.0/24,11.246.71.0/24,11.246.73.0/24,11.246.74.0/24,100.64.0.0/8,149.129.164.0/24,11.194.11.0/24
英国	11.199.93.0/24,100.64.0.0/8
亚太东南5（印度尼西亚，雅加达）	11.194.49.0/24,11.200.93.0/24,11.200.95.0/24,11.200.97.0/24,100.64.0.0/8,149.129.228.0/24,10.143.32.0/24,11.194.50.0/24
华北2（政务云）	11.194.116.0/24,100.64.0.0/8 如果IP地址段添加不成功，请添加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

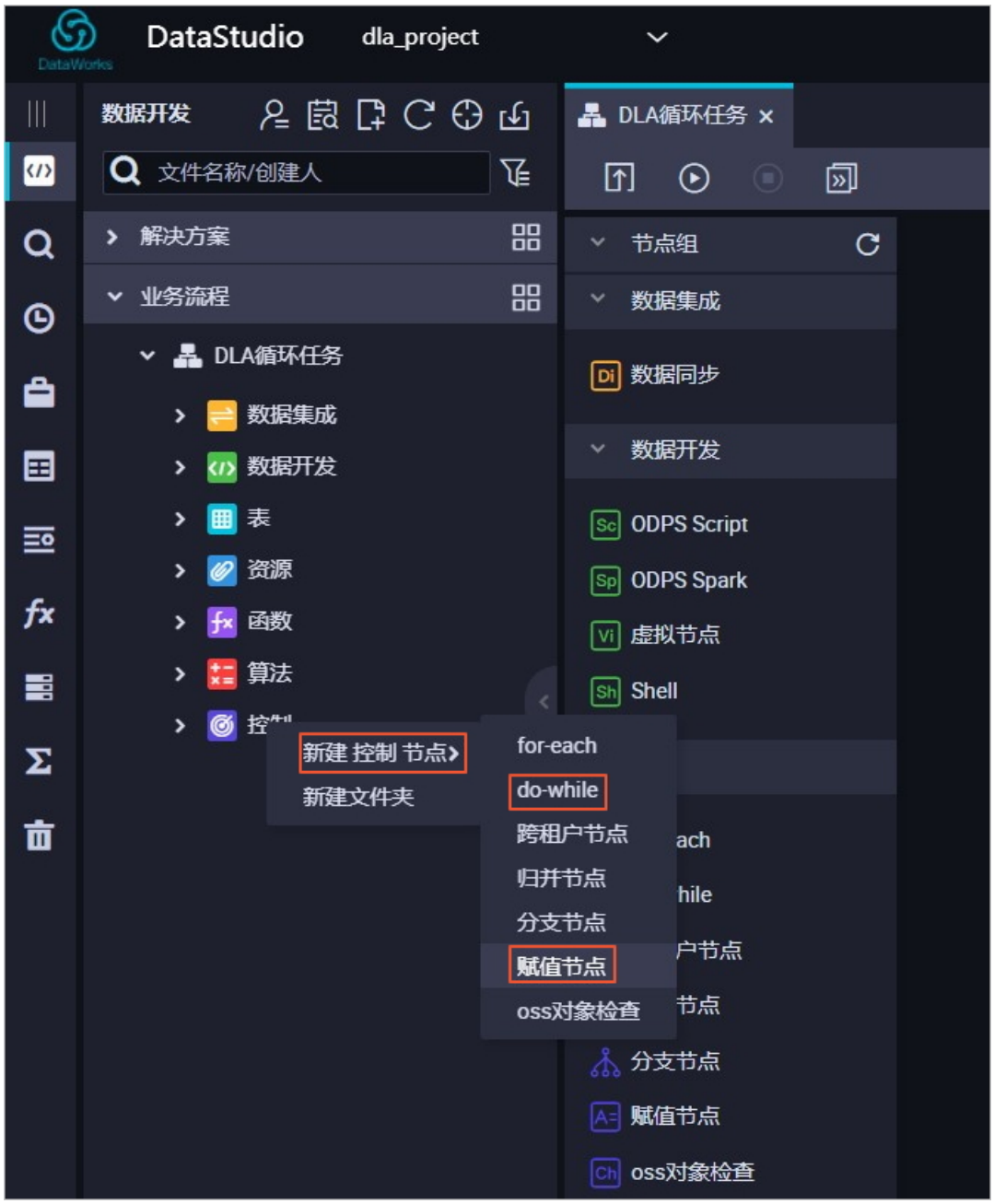
5. 完成上述参数配置后，单击**测试连通性**进行连通性测试，测试通过后单击**完成**。

步骤二：新建业务流程和节点

1. 登录DataWorks控制台，单击对应项目栏中的**进入数据开发**。
2. 右键单击**业务流程**新建一个流程，本教程新建业务流程**DLA循环任务**。



3. 在新建的业务流程下，创建一个**赋值节点**和一个循环节点**do-while**。

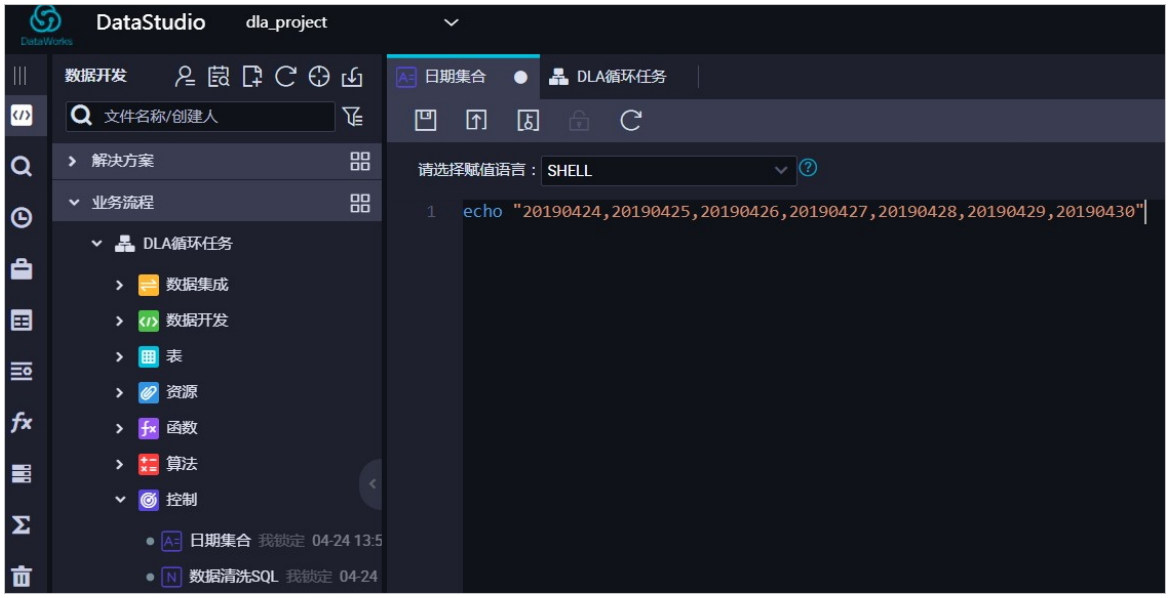


步骤三：配置赋值节点

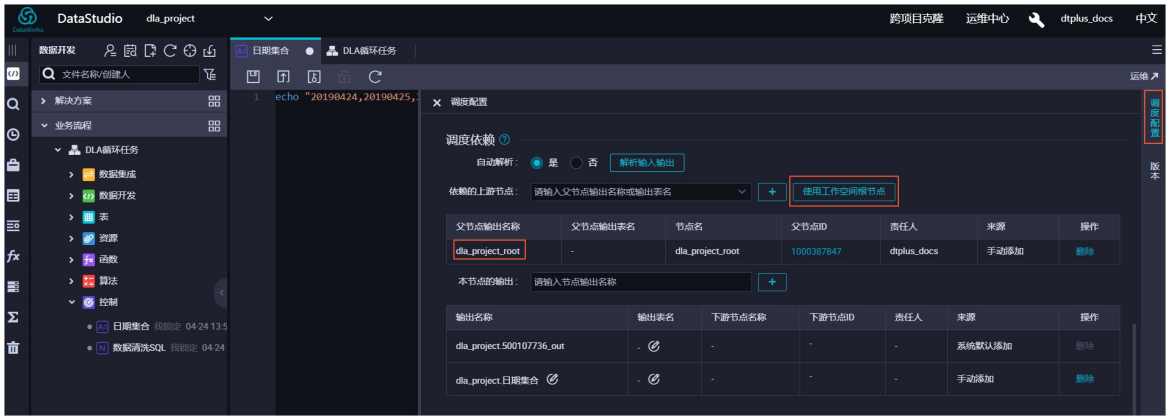
1. 双击**日期集合**节点进入编辑页面，选择**SHELL**语言，将要执行的日期值写在一个数组里并保存设置。

   日期值之间以英文逗号分隔，且不含空格：

```
echo "20190424,20190425,20190426,20190427,20190428,20190429,20190430"
```



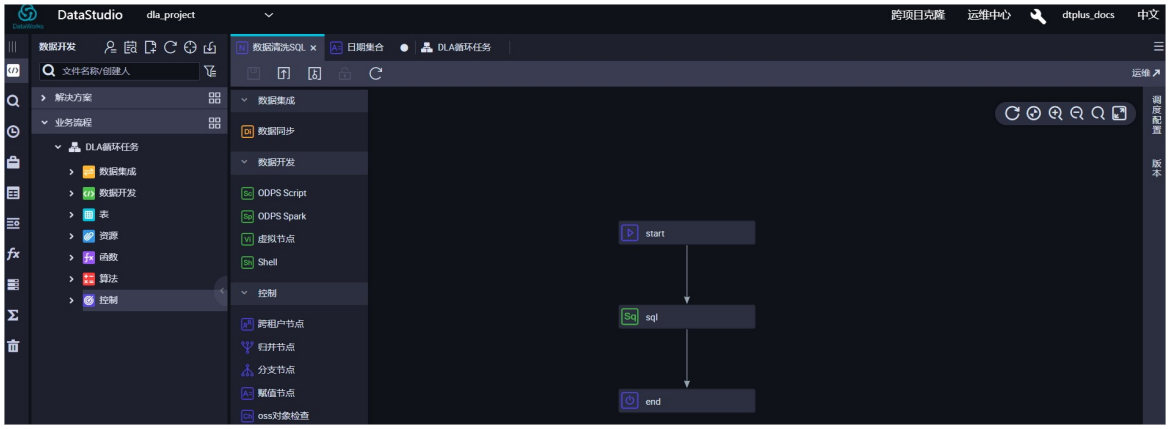
2. 单击**调度配置**为**赋值节点**设置一个上游节点。这里可以将**当前工作空间的根节点**设置为上游节点，比如，本教程工作空间名为dla_project，则上游节点为dla_project_root。



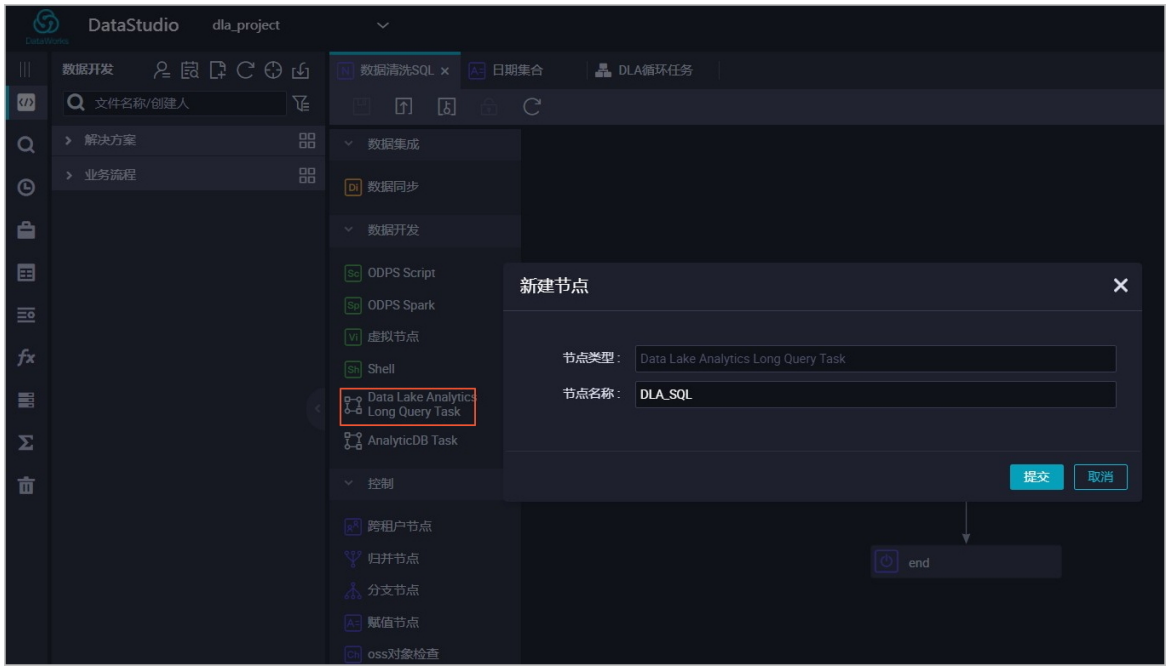
3. 单击保存，保存配置。

步骤四：配置do-while循环节点

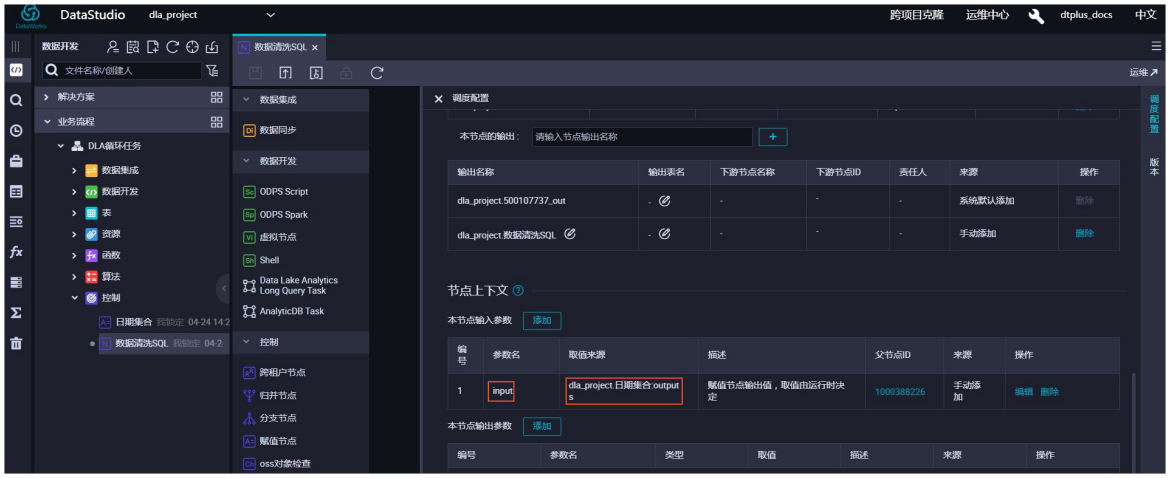
1. 双击do-while循环节点进入编辑页面。



2. 新建一个DLA任务。



3. 单击**调度配置**，在**调度配置**页面设置依赖关系和节点上下文。上游节点设置为赋值节点**日期集合**，本节点的输入为赋值节点的输出。



步骤五：配置DLA_SQL节点

```
INSERT INTO finished_orders
SELECT O_ORDERKEY
    ,O_TOTALPRICE
FROM   orders
WHERE  pure_date = ${dag.input[${dag.offset}]}
AND    O_ORDERSTATUS = 'F';
```

- `pure_date` 的值是从赋值节点读入，每次读取赋值节点的输出结果数组中的一个值。
- `dag.offset` 是DataWorks的系统保留变量，每一次循环次数相对于第一次的偏移量，即第一次循环中offset为0、第二次为1、第三次为2...第n次为n-1。
- `dag.input` 变量是用户配置的循环节点的上下文依赖。循环节点内部节点如果需要引用上下文依赖的值，可以直接通过 `dag.$ctxKey` 的方式来引用。本教程中，上文配置的上下文依赖key为input，因此可以使

用 {dag.input} 来引用这个值。

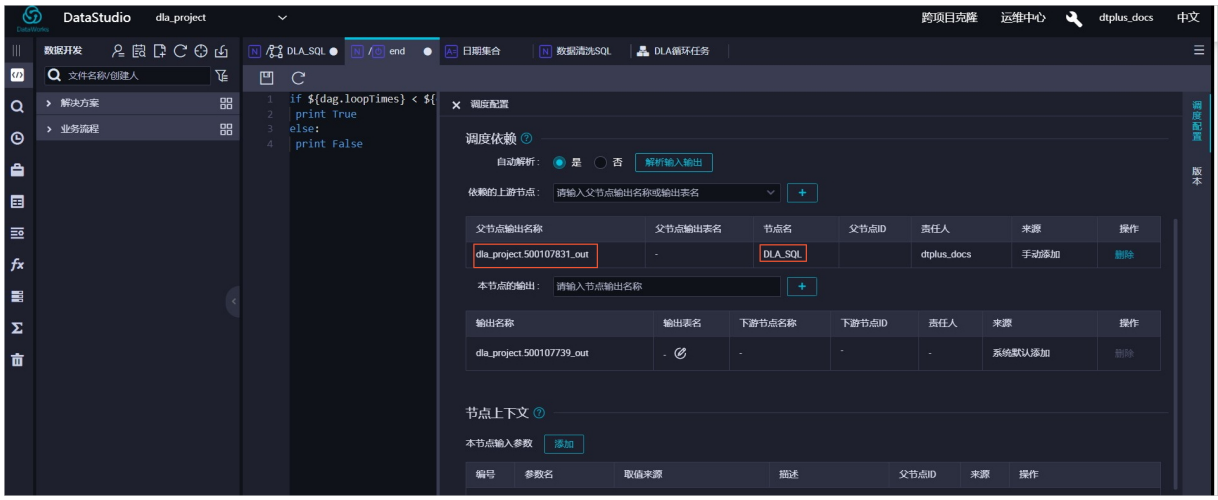
- dag.input[$dag.offset] 节点数据集初始化的输出是一个表格，可以用偏移量的方式来获取表格数据的某一行。由于每次循环中的值是递增的，所以最后输出的数据应该是 {dag.input[0]} 、 ${dag.input[1]} 以此类推达到遍历数据集的效果。

步骤六：设置end节点

end节点用于控制循环的结束，将 dag.loopTimes 和 dag.input.length 进行比较， dag.loopTimes 小于 dag.input.length 则输出 True 继续循环； dag.loopTimes 不小于 dag.input.length 则输出 False 退出循环。 dag.input.length 变量，标识上下文参数 input 数组的行数，是系统自动根据节点配置的上下文下发的变量。

```
if ${dag.loopTimes} < ${dag.input.length}:
 print True
else:
 print False
```

在**调度配置**页面，设置end节点的**上游节点**为DLA_SQL节点。

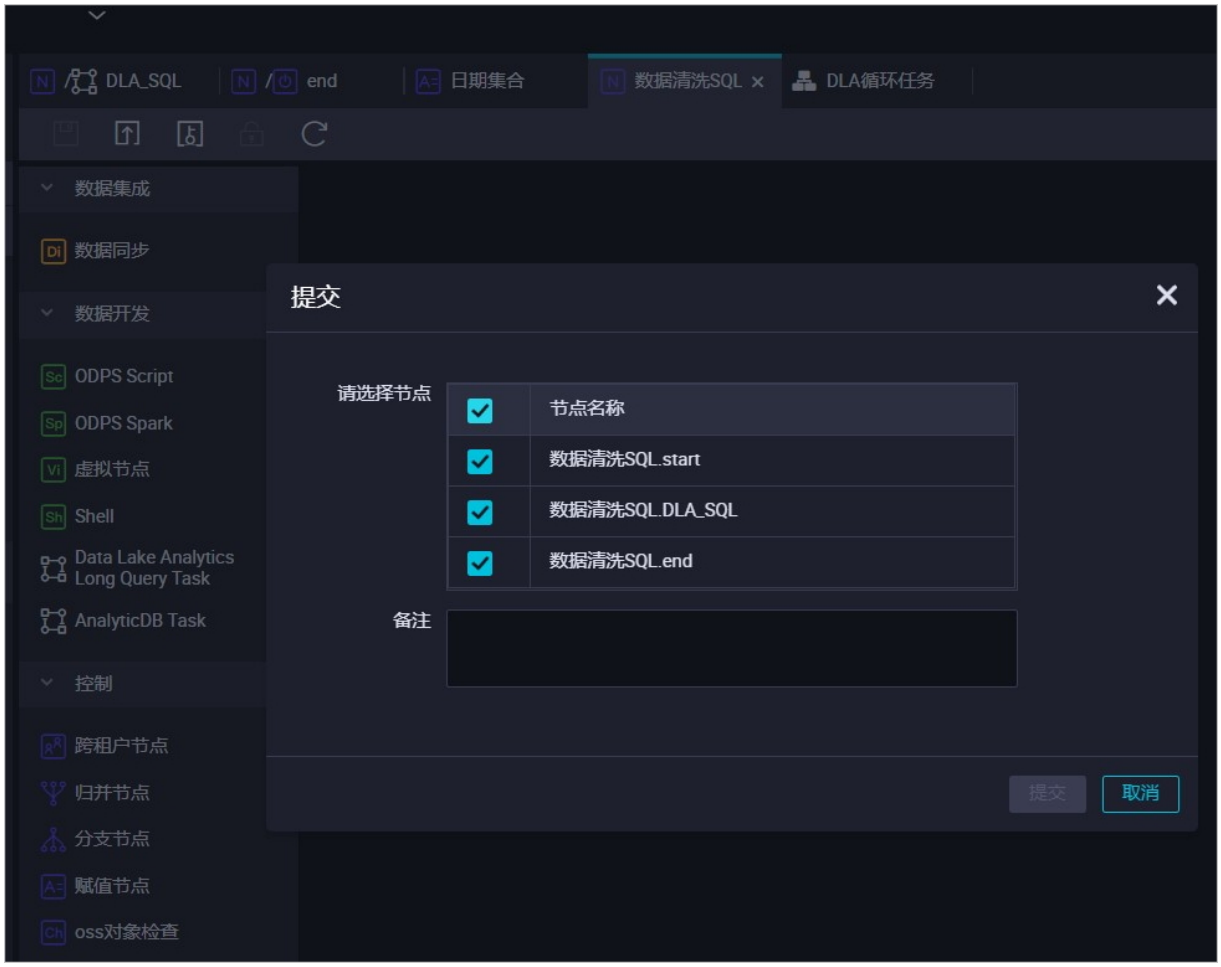


设置完成并保存后，可以看到循环节点变更为以下形式。

步骤七：发布任务

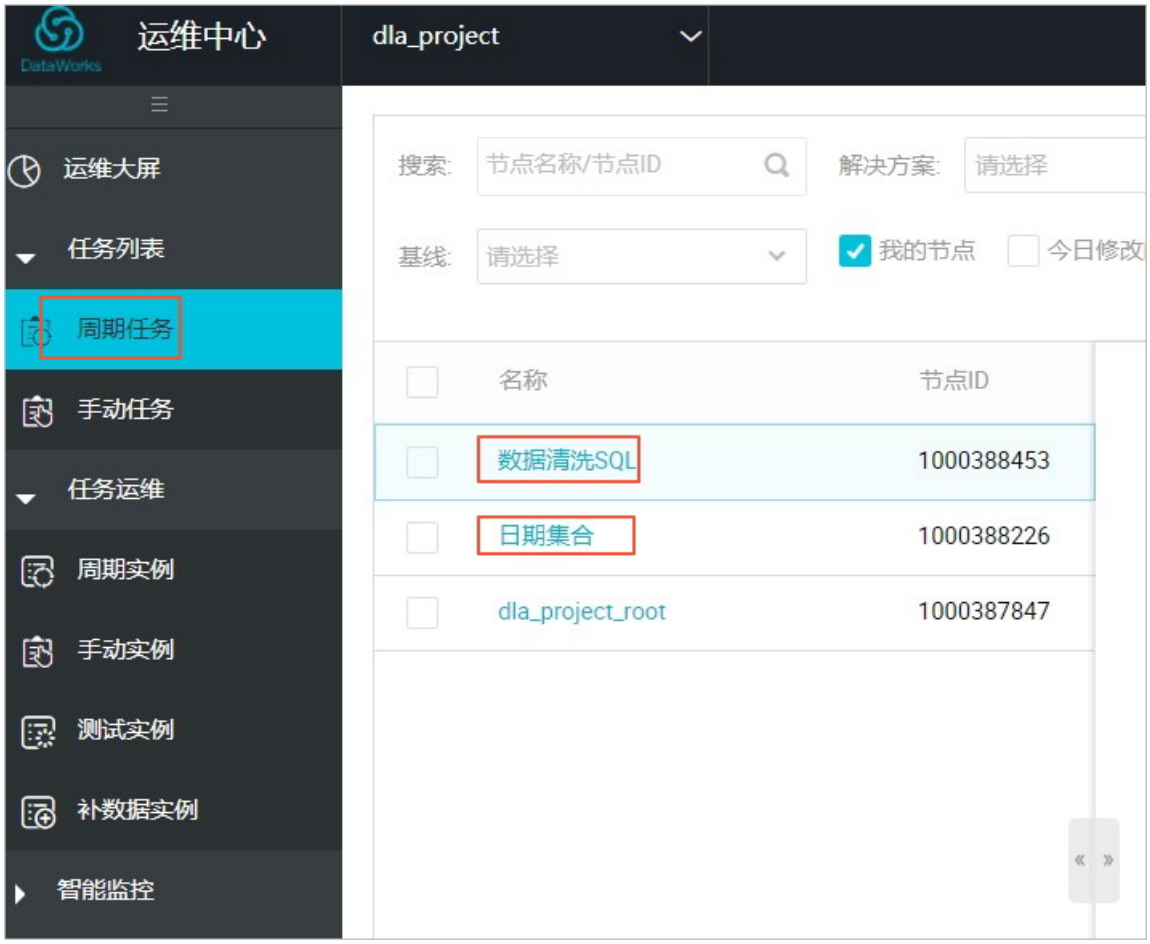
目前DataWorks的开发界面暂不支持运行循环节点，需要提交循环节点后在**运维中心**测试运行。

分别单击**日期集合**和**数据清洗SQL**页面上的**提交按钮**进行提交。提交**数据清洗SQL**时，注意勾选所有节点。



步骤八：运行任务

1. 进入**运维中心**页面，在**周期任务**列表中可以看到刚刚提交的两个作业。



2. 右键单击**日期集合>补数据>当前节点及下游节点**手动执行该组任务。

   任务执行后可以看到每个节点的运行状态。