阿里云
DataWorks

使用教程

文档版本：20190906

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
	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。
	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。	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 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。	`swich` *`{stand | slave}`*

目录

法律声明.....I
通用约定.....I
1 Workshop.....1
1.1 Workshop教程介绍.....1
1.2 环境准备.....2
1.3 数据采集.....7
1.4 数据加工.....13
1.5 数据质量监控.....19
1.6 数据可视化展现.....30
1.7 通过Function Studio开发UDF.....39
2 搭建互联网在线运营分析平台.....44
2.1 业务场景与开发流程.....44
2.2 环境准备.....46
2.3 数据准备.....53
2.4 数据建模与开发.....58
2.4.1 新建数据表.....58
2.4.2 设计工作流.....65
2.4.3 节点配置.....67
2.4.4 任务提交与测试.....73
2.5 数据可视化展现.....80
3 数据质量保障教程.....94
3.1 数据质量教程概述.....94
3.2 数据质量管理流程.....96
3.3 数据资产定级.....97
3.4 离线数据加工卡点.....98
3.5 数据质量风险监控.....101
3.6 数据及时性监控.....116
4 实现窃电用户自动识别教程.....121
4.1 窃电用户自动识别概述.....121
4.2 环境准备.....122
4.3 数据准备.....128
4.4 数据加工.....142
4.5 数据建模.....155

1 Workshop

1.1 Workshop教程介绍

本模块将为您介绍DataWorks的设计思路和核心功能，以帮您深入了解阿里云DataWorks。

教程概述

教程时长：2小时，采用在线学习的方式。

教程对象：面向Java工程师、产品运营等Dataworks所有的新老用户。只需熟悉标准SQL，即可快速掌握DataWorks的基本技能，无需对数据仓库和 MaxCompute的原理过多了解。建议您进一步学习Dataworks教程，深入了解Dataworks的基本概念及功能，详情请参见#unique_5。

教程目标：以常见的真实的海量日志数据分析任务为教程背景，争取在完成教程后，您对DataWorks的主要功能有所了解，能够按照教程演示内容，独立完成数据采集 、数据开发和任务运维等数据岗位常见的任务。

开发流程

Workshop教程涉及的具体开发流程如下：

1. 环境准备：准备操作过程中需要的MaxCompute、DataWorks等环境。详情请参见#unique_6。
2. 数据采集：学习如何从不同的数据源同步数据至MaxCompute中、如何快速触发任务运行、如何查看任务日志等。详情请参见数据采集。
3. 数据加工：学习如何运行数据流程图、如何新建数据表、如何新建数据流程任务节点、如何配置任务的周期调度属性。详情请参见数据加工。
4. 数据质量监控：学习如何给任务配置数据质量的监控规则，以保证任务运行的质量问题。详情请参见#unique_9。
5. 数据可视化展现：学习如何通过Quick BI创建网站用户分析画像的仪表板，实现所需数据的可视化展现。详情请参见数据可视化展现。
6. 通过Function Studio开发UDF：学习如何通过Function Studio开发UDF，并将其提交至DataStudio的开发环境。详情请参见#unique_11。

DataWorks简介

DataWorks是数加平台&DataWorks团队倾力9年打造的一款一站式大数据研发平台，以MaxCompute为主要计算引擎，上层有机融合数据集成、数据建模、数据开发、运维监控、数据

管理、数据安全和数据质量等产品功能，同时与算法平台PAI打通，完善了从大数据开发到数据挖掘、机器学习的完整链路。

学习答疑

如果您在学习过程中遇到问题，可以加入钉钉群：11718465进行咨询。

1.2 环境准备

为保证您可以顺利完成本次实验，请您首先确保自己云账号已开通大数据计算服务MaxCompute和数据工场DataWorks。

前提条件

- 阿里云账号注册，详情请参见#unique_13。
- 实名认证，详情请参见#unique_14或#unique_15。

背景信息

本次实验涉及的阿里云产品如下：

- 大数据计算服务MaxCompute
- 数据工场DataWorks

开通大数据计算服务MaxCompute

> 说明:
> 如果您已经开通MaxCompute，请跳过此步骤，直接创建DataWorks工作空间。

1. 登录阿里云官网，单击右上角的登录，填写您的阿里云账号和密码。
2. 选择产品分类 > 大数据 > 大数据计算 > MaxCompute，进入MaxCompute产品详情页。
3. 单击立即购买。
4. 进入按量付费页面，选择区域和规格类型，单击立即购买。

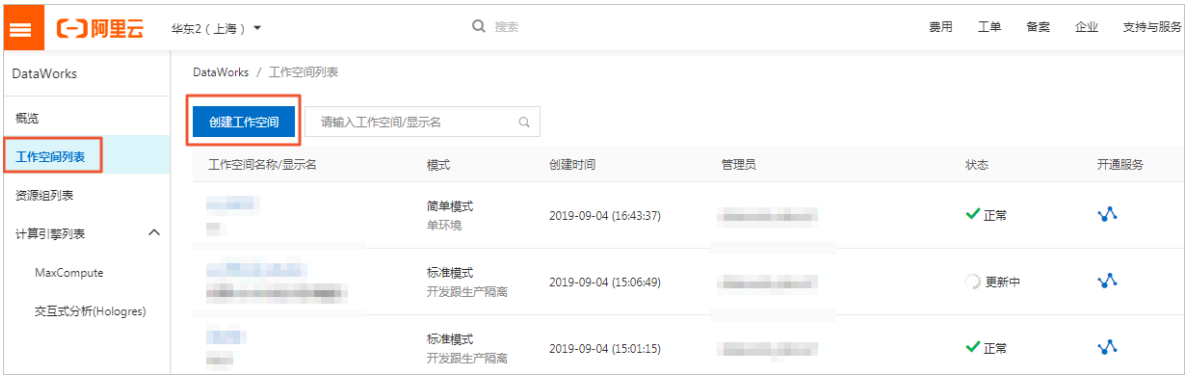
创建工作空间

> 说明:
> 因本实验提供的数据资源都在华东2，建议您将工作空间创建在华东2，以避免工作空间创建在其他Region，添加数据源时出现网络不可达的情况。

1. 使用主账号登录DataWorks控制台。

2. 单击控制台概览 > 常用功能下的创建工作空间。

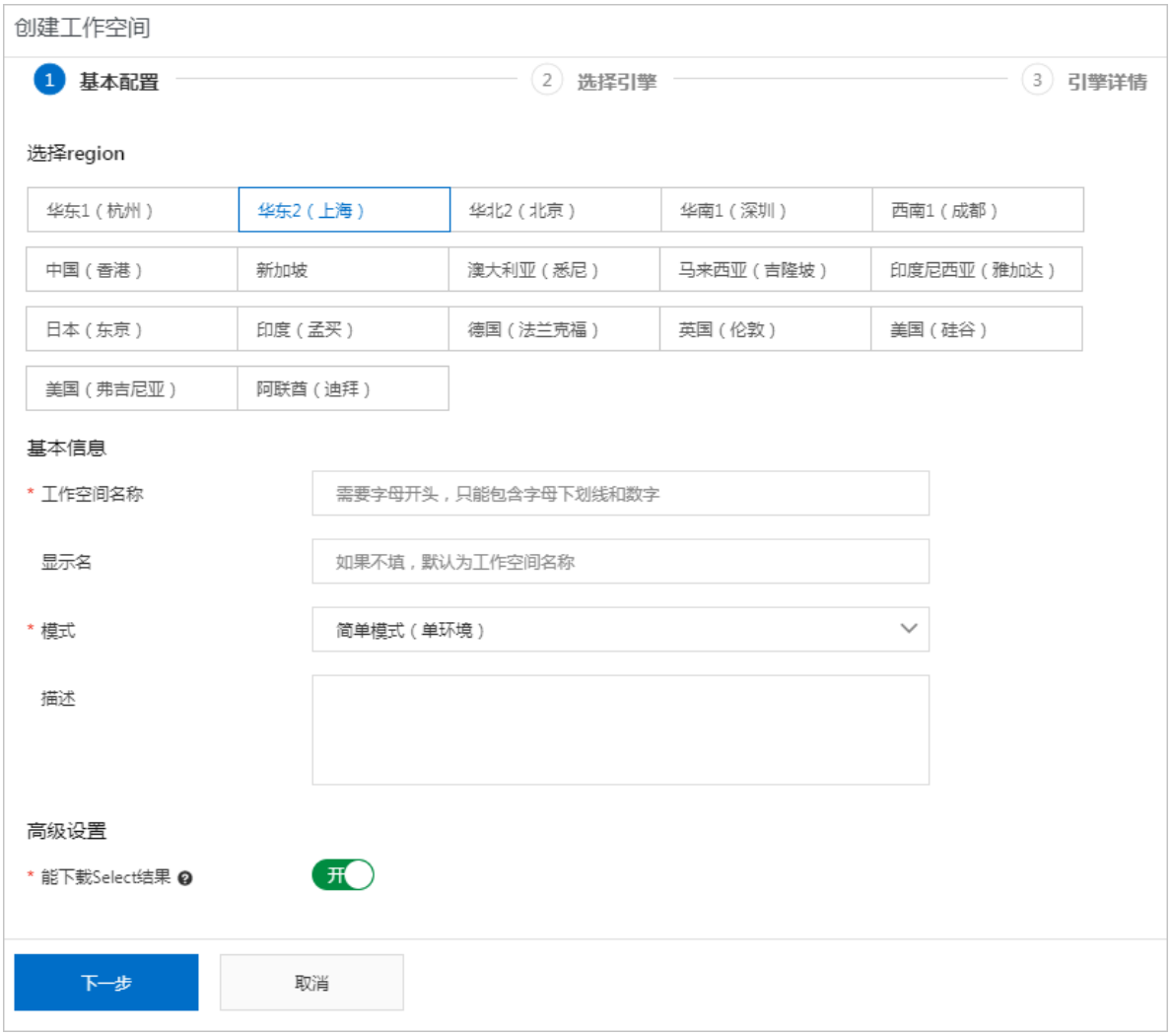
您也可以进入工作空间列表页面，单击创建工作空间。



说明：

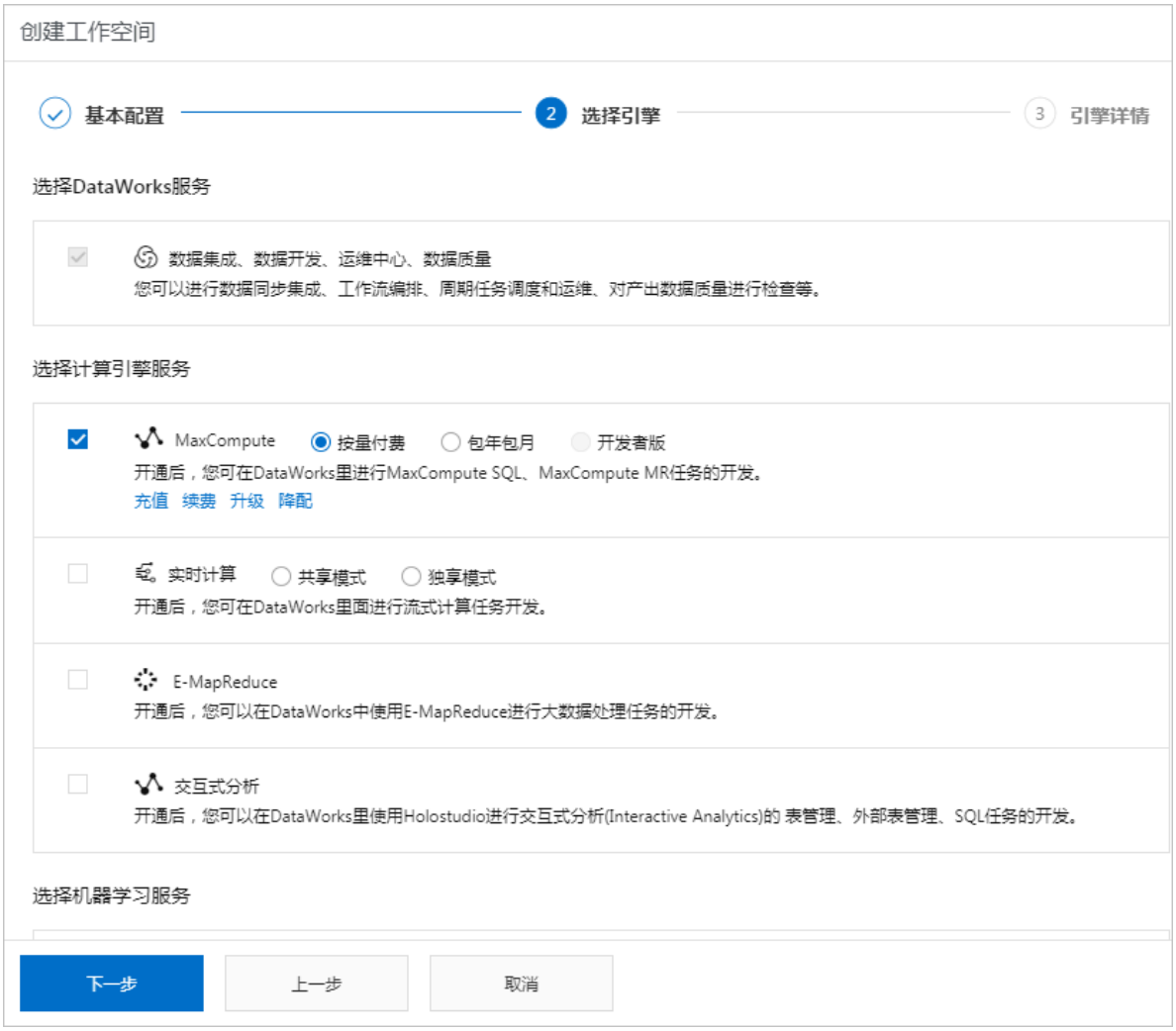
从工作空间列表页面创建工作空间时，需提前选择区域，在创建工作空间对话框中不会显示选择region。

3. 填写创建工作空间对话框中的基本配置，单击下一步。

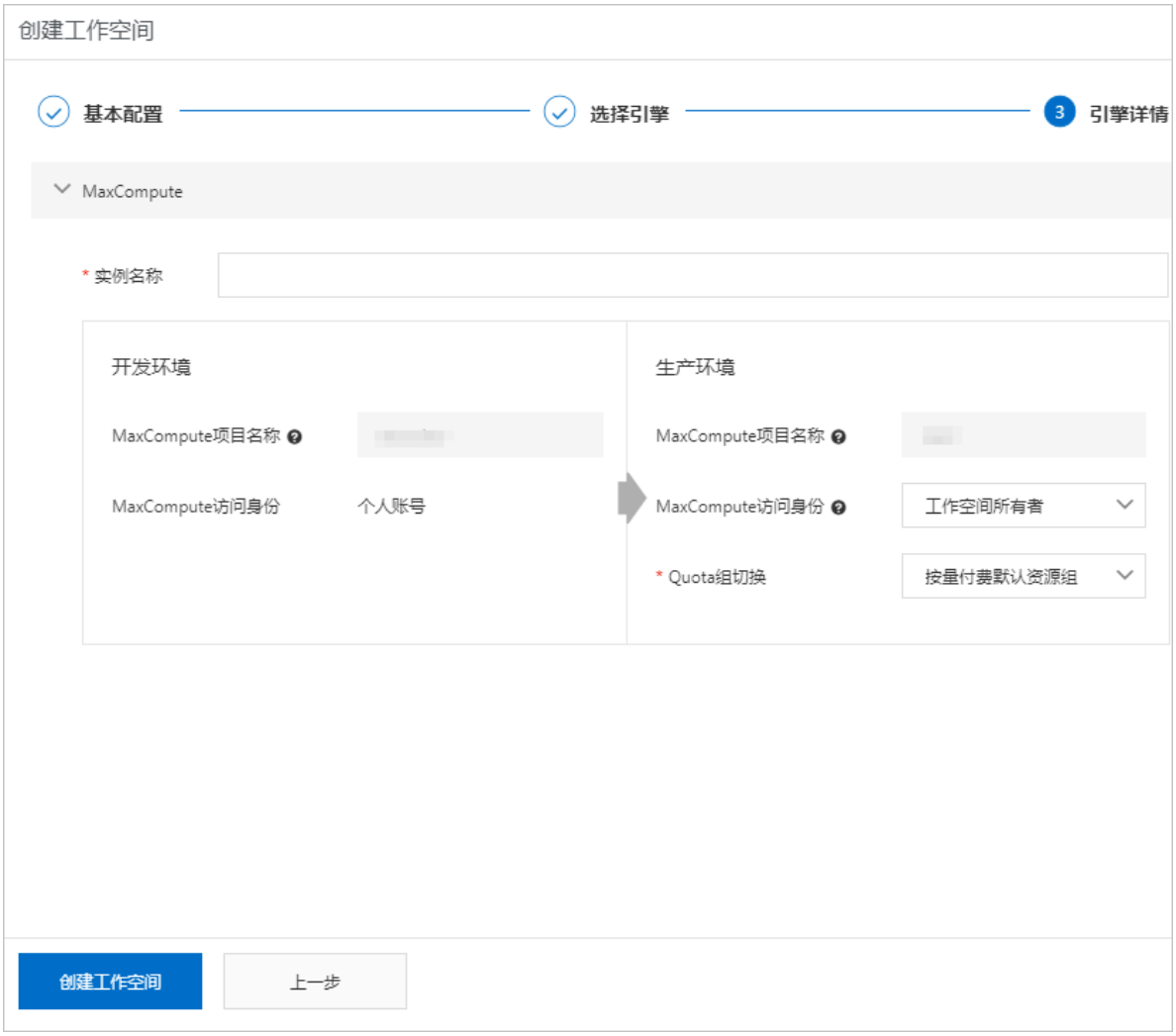


4. 进入选择引擎界面，选择相应引擎后，单击下一步。

   DataWorks已正式商用，如果该区域没有开通，需要首先开通正式商用的服务。默认选中数据集成、数据开发、运维中心和数据质量。



5. 进入引擎详情页面，填写选购引擎的配置。



分类	配置	说明
MaxCompute	实例名称	实例名称不能超过27个字符，仅支持字母、中文开头，仅包含中文、字母、下划线和数字。
	MaxCompute项目名称	默认与DataWorks工作空间的名称一致。
	MaxCompute访问身份	包括个人账号和工作空间所有者，开发环境默认为个人账号，生产环境推荐使用工作空间所有者。
	Quota组切换	Quota用来实现计算资源和磁盘配额。

6. 配置完成后，单击创建工作空间。

工作空间创建成功后，即可在工作空间列表页面查看相应内容。

1.3 数据采集

本文将为您介绍如何通过DataWorks采集日志数据至MaxCompute。

新建数据源

> 说明:
> 根据本次实验模拟的场景，您需要分别创建OSS数据源和RDS数据源。

- 新建OSS数据源。

1. 单击相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，单击新增数据源。

3. 选择数据源类型为OSS，填写新增OSS数据源对话框中的配置。

配置	说明
数据源名称	填写oss_workshop_log。
数据源描述	对数据源进行简单描述。
适用环境	勾选开发。 说明: 开发环境的数据源创建完成后，需要勾选生产，以同样方式创建生产环境的数据源，否则任务生产执行会报错。
Endpoint	填写`http://oss-cn-shanghai-internal.aliyuncs.com`。
Bucket	填写dataworks-workshop。
AccessKey ID	填写LTAINEhd4MZ8pX64。
AccessKey Secret	填写lXnzUngTSebt3SfLYxZxoSjGAK6IaF。

4. 单击测试连通性。
5. 连通性测试通过后，单击完成。

> 说明:
> - 如果测试连通性失败，请检查您的用户名/密码和工作空间所在区域。
> - 建议将工作空间创建在华东2，其他区域不保证网络可达。

- 如果您无法使用内网Endpoint连接数据源，可以改用公网Endpoint。

· 新建RDS数据源。

1. 单击相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，单击新增数据源。
3. 选择数据源类型MySQL，填写新增MySQL数据源对话框中的配置。

配置	说明
数据源类型	选择阿里云数据库（RDS）。
数据源名称	填写rds_workshop_log。
数据源描述	填写rds日志数据同步。
适用环境	勾选开发。 说明: 开发环境的数据源创建完成后，需要勾选生产，以同样方式创建生产环境的数据源，否则任务生产执行会报错。
RDS实例ID	填写rm-bp1z69dodhh85z9qa。
RDS实例主账号ID	填写1156529087455811。
数据库名	填写workshop。
用户名	填写workshop。
密码	填写workshop#2017。

4. 单击测试连通性。
5. 连通性测试通过后，单击完成。

新建业务流程

1. 单击左上角的图标，选择全部产品 > DataStudio（数据开发）。
2. 右键单击业务流程，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。
4. 单击新建，即可完成业务流程的创建。

5. 进入业务流程开发面板，并向面板中拖入1个虚拟节点（workshopstart）和2个数据同步节点（oss_数据同步和rds_数据同步）。填写相应配置后，单击提交。

6. 拖拽连线将workshop_start节点设置为2个数据同步节点的上游节点。

配置workshop_start节点

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

   由于新版本给每个节点都设置了输入输出节点，所以需要给workshop_start节点设置一个输入。此处设置其上游节点为工作空间根节点，通常命名为工作空间名_root。

3. 配置完成后，单击左上角的进行保存。

新建表

1. 右键单击业务流程workshop下的表，选择新建表。

2. 在新建表对话框中填写表名，单击提交。

   此处需要创建2张表（ods_raw_log_d和ods_user_info_d），分别存储同步过来的OSS日志数据和RDS日志数据。

   输入表名称。

3. 打开创建的表，单击DDL模式，分别填写以下相应的建表语句。

```
--创建OSS日志对应目标表
CREATE TABLE IF NOT EXISTS  ods_raw_log_d (
    col STRING
)
PARTITIONED BY (
    dt STRING
);
```

```
--创建RDS对应目标表
CREATE TABLE IF NOT  EXISTS ods_user_info_d (
    uid STRING COMMENT '用户ID',
    gender STRING COMMENT '性别',
    age_range STRING COMMENT '年龄段',
    zodiac STRING COMMENT '星座'
)
PARTITIONED BY (
    dt STRING
```

```
);
```

4. 建表语句输入完成后，单击生成表结构并确认覆盖当前操作。

5. 返回建表页面后，在基本属性中输入表的中文名。

6. 完成设置后，分别单击提交到开发环境和提交到生产环境，提交后即可在表中查看信息。

配置数据同步任务

说明:

标准项目模式下，不建议数据集任务在开发环境下运行（开发面板直接运行），建议将其发布至生产环境后再操作测试运行，以获取完整的运行日志。同时，数据产出至生产环境后，您可以申请表权限，以读取写入开发环境中的表数据。

- 配置oss_数据同步节点。

1. 双击oss_数据同步节点，进入节点配置页面。
2. 选择数据来源。

配置	说明
数据源	选择OSS > oss_workshop_log数据源。
Object前缀	选择user_log.txt。
文本类型	选择text类型。
列分隔符	选择列分隔符为（|）。
编码格式	默认为UTF-8格式。
null值	表示null值的字符串。
压缩格式	包括None、Gzip、Bzip2和Zip四种类型，此处选择None。
是否包含表头	默认为No。

3. 选择数据去向。

配置	说明
数据源	选择ODPS > odps_first数据源。
表	选择数据源中的ods_raw_log_d表。

配置	说明
分区信息	默认配置为${bizdate}。
清理规则	默认为写入前清理已有数据。
空字符串作为null	此处勾选否。

说明:

odps_first数据源写入到当前工作空间下的MaxCompute项目中。

4. 配置字段映射，连接需要同步的字段。

5. 配置通道控制，作业速率上限建议配置为10MB/s。

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_18和#unique_19。

6. 确认当前任务的配置情况，可以进行修改。确认无误后，单击左上角的保存。

7. 关闭当前任务，返回业务流程配置面板。

- 配置rds_数据同步节点。

1. 双击rds_数据同步节点，进入节点配置页面。

2. 选择数据来源。

配置	说明
数据源	选择MySQL > rds_workshop_log数据源。
表	选择数据源中的ods_user_info_d表。

配置	说明
数据过滤	该数据过滤语句通常用作增量同步，此处可不填。
切分键	默认为uid。

3. 选择数据去向。

配置	说明
数据源	选择ODPS > odps_first数据源。
表	选择数据源中的ods_user_info_d表。
分区信息	默认配置为${bizdate}。
清理规则	默认为写入前清理已有数据。
空字符串作为null	此处勾选否。

4. 配置字段映射，默认同名映射。

5. 配置通道控制，作业速率上限建议配置为10MB/s。

6. 确认当前任务的配置情况，可以进行修改。确认无误后，单击左上角的保存。

7. 关闭当前任务，返回业务流程配置面板。

提交工作流任务

1. 单击左上角的提交按钮，提交当前业务流程。

2. 选择提交对话框中需要提交的节点，填写备注，勾选忽略输入输出不一致的告警。单击提交，待显示提交成功即可。

运行工作流任务

1. 单击运行。

2. 右键选中rds_数据同步任务，选择查看日志。

   当日志中出现如下字样，表示同步任务运行成功，并成功同步数据。

3. 右键选中oss_数据同步任务，选择查看日志，确认方法与rds_数据同步任务一致。

确认数据是否成功导入MaxCompute

1. 单击左侧导航栏中的临时查询。
2. 选择新建 > ODPS SQL。

3. 编写并执行SQL语句，查看导入ods_raw_log_d和和ods_user_info_d的记录数。

说明:

SQL语句如下所示，其中分区列需要更新为业务日期。例如，任务运行的日期为20180717，则业务日期为20180716。

```
--查看是否成功写入MaxCompute
select count(*) from ods_raw_log_d where dt=业务日期;
select count(*) from ods_user_info_d where dt=业务日期;
```

后续步骤

现在，您已经学习了如何进行日志数据同步，完成数据的采集，您可以继续学习下一个教程。在该教程中您将学习如何对采集的数据进行计算与分析。详情请参见#unique_20。

1.4 数据加工

本文将为您介绍如何通过DataWorks，将已经采集至MaxCompute的日志数据进行加工并进行用户画像。

前提条件

开始本实验前，请首先完成#unique_22中的操作。

新建数据表

1. 右键单击业务流程下的表，选择新建表。
2. 在新建表对话框中填写表名，单击提交。

   此处需要创建3张表，分别为ODS层表（ods_log_info_d）、DW层表（dw_user_info_all_d）和RPT层表（rpt_user_info_d）。
3. 打开创建的表，单击DDL模式，分别填写以下相应的建表语句。

```
--创建ODS层表
CREATE TABLE IF NOT EXISTS ods_log_info_d (
  ip STRING COMMENT 'ip地址',
  uid STRING COMMENT '用户ID',
  time STRING COMMENT '时间yyyymmddhh:mi:ss',
```

```
  status STRING COMMENT '服务器返回状态码',
  bytes STRING COMMENT '返回给客户端的字节数',
  region STRING COMMENT '地域，根据ip得到',
  method STRING COMMENT 'http请求类型',
  url STRING COMMENT 'url',
  protocol STRING COMMENT 'http协议版本号',
  referer STRING COMMENT '来源url',
  device STRING COMMENT '终端类型 ',
  identity STRING COMMENT '访问类型 crawler feed user unknown'
)
PARTITIONED BY (
  dt STRING
);
```

```
--创建DW层表
CREATE TABLE IF NOT EXISTS dw_user_info_all_d (
  uid STRING COMMENT '用户ID',
  gender STRING COMMENT '性别',
  age_range STRING COMMENT '年龄段',
  zodiac STRING COMMENT '星座',
  region STRING COMMENT '地域，根据ip得到',
  device STRING COMMENT '终端类型 ',
  identity STRING COMMENT '访问类型 crawler feed user unknown',
  method STRING COMMENT 'http请求类型',
  url STRING COMMENT 'url',
  referer STRING COMMENT '来源url',
  time STRING COMMENT '时间yyyymmddhh:mi:ss'
)
PARTITIONED BY (
  dt STRING
);
```

```
--创建RPT层表
CREATE TABLE IF NOT EXISTS rpt_user_info_d (
  uid STRING COMMENT '用户ID',
  region STRING COMMENT '地域，根据ip得到',
  device STRING COMMENT '终端类型 ',
  pv BIGINT COMMENT 'pv',
  gender STRING COMMENT '性别',
  age_range STRING COMMENT '年龄段',
  zodiac STRING COMMENT '星座'
)
PARTITIONED BY (
  dt STRING
);
```

4. 建表语句输入完成后，单击生成表结构并确认覆盖当前操作。
5. 返回建表页面后，在基本属性中输入表的中文名。
6. 完成设置后，分别单击提交到开发环境和提交到生产环境。

设计业务流程

业务流程workshop及依赖关系的配置可以参见#unique_23。向画布中拖入三个ODPS SQL节点，依次命名为ods_log_info_d、dw_user_info_all_d和rpt_user_info_d，并配置如下图所示的依赖关系。

创建用户自定义函数

1. 下载ip2region.jar。
2. 右键单击资源，选择新建资源 > JAR。
3. 单击选择文件，选择已经下载到本地的ip2region.jar，单击提交。
4. 资源上传至DataWorks后，单击提交。
5. 右键单击函数，选择新建函数。
6. 输入函数名称getregion，选择所属的业务流程，单击提交。
7. 填写注册函数对话框的配置，单击保存并提交。

配置	说明
函数类型	选择函数类型。
函数名	填写getregion。
责任人	选择责任人。
类名	填写org.alidata.odps.udf.Ip2Region。
资源列表	填写ip2region.jar。
描述：	填写IP地址转换地域。
命令格式	填写getregion（‘ip’）。
参数说明	填写IP地址。

配置ODPS SQL节点

- 配置ods_log_info_d节点。
  1. 双击ods_log_info_d节点，进入节点配置页面。
  2. 编写处理逻辑。

     SQL逻辑如下所示：

```
INSERT OVERWRITE TABLE ods_log_info_d PARTITION (dt=${bdp.system.
bizdate})
SELECT ip
  , uid
```

```
  , time
  , status
  , bytes --使用自定义UDF通过ip得到地域
  , getregion(ip) AS region --通过正则把request差分为三个字段
  , regexp_substr(request, '(^[^ ]+ )') AS method
  , regexp_extract(request, '^[^ ]+ (.*) [^ ]+$') AS url
  , regexp_substr(request, '([^ ]+$)') AS protocol --通过正则清晰
refer，得到更精准的url
  , regexp_extract(referer, '^[^/]+://([^/]+){1}') AS referer --通
过agent得到终端信息和访问形式
  , CASE
    WHEN TOLOWER(agent) RLIKE 'android' THEN 'android'
    WHEN TOLOWER(agent) RLIKE 'iphone' THEN 'iphone'
    WHEN TOLOWER(agent) RLIKE 'ipad' THEN 'ipad'
    WHEN TOLOWER(agent) RLIKE 'macintosh' THEN 'macintosh'
    WHEN TOLOWER(agent) RLIKE 'windows phone' THEN 'windows_phone'
    WHEN TOLOWER(agent) RLIKE 'windows' THEN 'windows_pc'
    ELSE 'unknown'
  END AS device
  , CASE
    WHEN TOLOWER(agent) RLIKE '(bot|spider|crawler|slurp)' THEN '
crawler'
    WHEN TOLOWER(agent) RLIKE 'feed'
    OR regexp_extract(request, '^[^ ]+ (.*) [^ ]+$') RLIKE 'feed'
THEN 'feed'
    WHEN TOLOWER(agent) NOT RLIKE '(bot|spider|crawler|feed|slurp
)'
    AND agent RLIKE '^[Mozilla|Opera]'
    AND regexp_extract(request, '^[^ ]+ (.*) [^ ]+$') NOT RLIKE '
feed' THEN 'user'
    ELSE 'unknown'
  END AS identity
  FROM (
    SELECT SPLIT(col, '##@@')[0] AS ip
    , SPLIT(col, '##@@')[1] AS uid
    , SPLIT(col, '##@@')[2] AS time
    , SPLIT(col, '##@@')[3] AS request
    , SPLIT(col, '##@@')[4] AS status
    , SPLIT(col, '##@@')[5] AS bytes
    , SPLIT(col, '##@@')[6] AS referer
    , SPLIT(col, '##@@')[7] AS agent
  FROM ods_raw_log_d
  WHERE dt = ${bdp.system.bizdate}
) a;
```

3. 单击左上角的保存按钮。

- 配置dw_user_info_all_d节点。

  1. 双击dw_user_info_all_d节点，进入节点配置页面。
  2. 编写处理逻辑，SQL逻辑如下所示：

```
INSERT OVERWRITE TABLE dw_user_info_all_d PARTITION (dt='${bdp.
system.bizdate}')
SELECT COALESCE(a.uid, b.uid) AS uid
  , b.gender
  , b.age_range
  , b.zodiac
  , a.region
  , a.device
  , a.identity
  , a.method
```

```
  , a.url
  , a.referer
  , a.time
FROM (
  SELECT *
  FROM ods_log_info_d
  WHERE dt = ${bdp.system.bizdate}
) a
LEFT OUTER JOIN (
  SELECT *
  FROM ods_user_info_d
  WHERE dt = ${bdp.system.bizdate}
) b
ON a.uid = b.uid;
```

3. 单击左上角的保存按钮。

- 配置rpt_user_info_d节点。

1. 双击rpt_user_info_d节点，进入节点配置页面。
2. 编写处理逻辑，SQL逻辑如下所示：

```
INSERT OVERWRITE TABLE rpt_user_info_d PARTITION (dt='${bdp.system
.bizdate}')
SELECT uid
  , MAX(region)
  , MAX(device)
  , COUNT(0) AS pv
  , MAX(gender)
  , MAX(age_range)
  , MAX(zodiac)
FROM dw_user_info_all_d
WHERE dt = ${bdp.system.bizdate}
GROUP BY uid;
```

3. 单击左上角的保存按钮。

提交业务流程

1. 单击提交按钮，提交业务流程中已配置完成的节点。

2. 选择提交对话框中需要提交的节点，勾选忽略输入输出不一致的告警。

3. 单击提交。

运行业务流程

1. 单击运行，验证代码逻辑。

2. 待所有任务运行完成显示绿色箭头后，单击左侧导航栏中的临时查询，进入临时查询面板。
3. 右键单击临时查询，选择新建节点 > 临时查询。

4. 编写并执行SQL语句，查询任务运行结果，确认数据产出。

   查询语句如下所示：

```
---查看rpt_user_info_d数据情况。
select * from rpt_user_info_d where dt=业务日期 limit 10;
```

发布业务流程

提交业务流程后，表示任务已进入开发环境。由于开发环境的任务不会自动调度，您需要将配置完成的任务发布至生产环境。

> 说明：
>
> · 将任务发布至生产环境前，您需要对代码进行测试，确保其正确性。
>
> · 如果您使用的是简单模式的工作空间，则没有发布按钮，您可以直接前往运维中心。

1. 单击发布，进入发布页面。
2. 选择待发布任务，单击添加到待发布。
3. 进入待发布列表，单击全部打包发布。
4. 在发布包列表页面查看已发布内容。

在生产环境运行任务

1. 任务发布成功后，单击运维中心。
2. 选择任务列表中的workshop业务流程。
3. 右键单击DAG图中的workshop_start节点，选择补数据 > 当前节点及下游节点。
4. 勾选需要补数据的任务，输入业务日期，单击确定。

   单击确定后，自动跳转至补数据任务实例页面。
5. 单击刷新，直至SQL任务都运行成功即可。

后续步骤

现在，您已经学习了如何创建SQL任务、如何处理原始日志数据。您可以继续学习下一个教程，学习如何对开发完成的任务设置数据质量监控，保证任务运行的质量。详情请参见#unique_24。

1.5 数据质量监控

本文主要阐述在使用数据工场的过程中如何监控数据质量，设置表的质量监控规则，监控提醒等。

前置条件

在进行本实验前，请先完成实验#unique_22和#unique_20。

数据质量

数据质量（DQC），是支持多种异构数据源的质量校验、通知、管理服务的一站式平台。数据质量以数据集（DataSet）为监控对象，目前支持MaxCompute数据表和DataHub实时数据流的监控，当离线MaxCompute数据发生变化时，数据质量会对数据进行校验，并阻塞生产链路，以避免问题数据污染扩散。同时，数据质量提供了历史校验结果的管理，以便您对数据质量分析和定级。在流式数据场景下，数据质量能够基于Datahub数据通道进行断流监控，第一时间告警给订阅用户，并且支持橙色、红色告警等级，以及告警频次设置，最大限度减少冗余报警。

数据质量的使用流程：针对已有的表进行监控规则配置，配置完规则后可以进行试跑，验证此规则是否试用。当试跑成功后，可将此规则和调度任务进行关联。关联成功后，每次调度任务代码运行完毕，都会触发数据质量的校验规则，以提升任务准确性。在关联调度后，可根据业务情况，对重要的表进行订阅。订阅成功后，此表的数据质量一旦出问题，都会有邮件或者报警进行通知。

说明:

数据质量会产生额外的计算费用，在使用时请注意。关于数据质量的详细介绍请参见#unique_26。

新增表规则配置

如果已完成《日志数据上传》、《用户画像》实验，请确认您是否已拥有数据表：ods_raw_log_d、ods_user_info_d、ods_log_info_d、dw_user_info_all_d、rpt_user_info_d。完成确认后，您可以开始配置表规则。

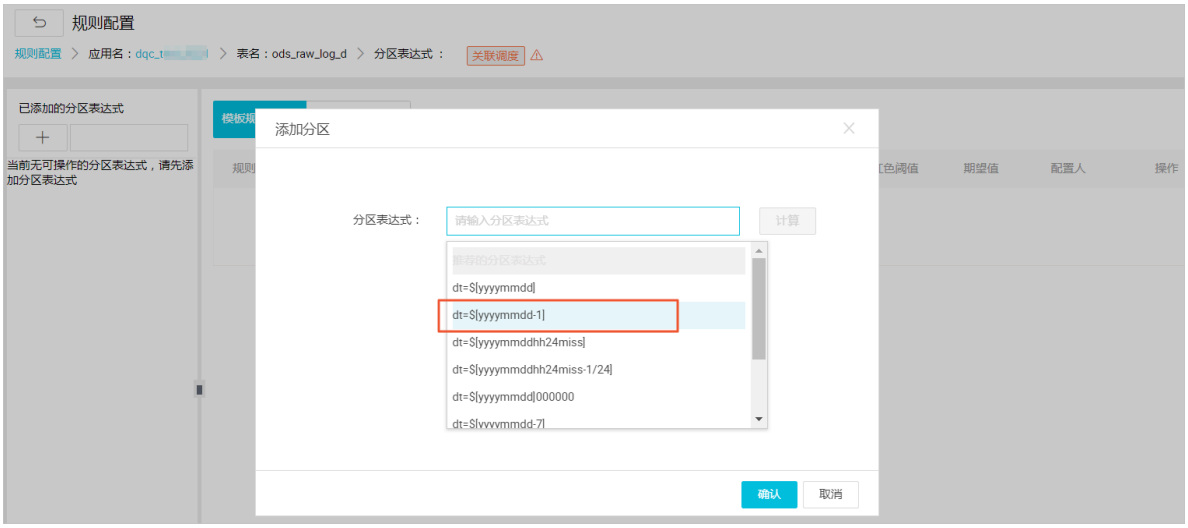
- ods_raw_log_d

在数据质量中可以看到该项目下的所有表信息，为ods_raw_log_d表进行数据质量的监控规则配置。

选择ods_raw_log_d表，单击配置监控规则，将会进入如下页面。

ods_raw_log_d表的数据来源为oss_workshop_log，数据是从OSS中获取到的日志数据，其分区格式为`${bdp.system.bizdate}`（获取到前一天的日期）。

对于此类每天产出的日志数据，您可以配置表的分区表达式。分区表达式有如下几种，您可以选择 `dt=$[yyyymmdd-1]` 。表达式的详细解读参见#unique_28。



说明:

如果表中无分区列，可以配置无分区，请根据真实的分区值配置对应的分区表达式。

单击确认后，选择创建规则。

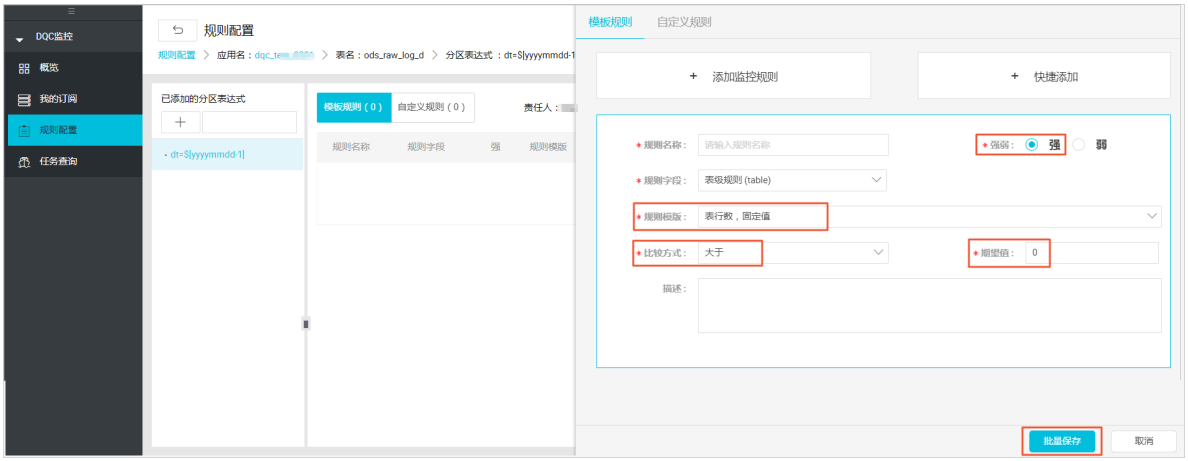
单击添加监控规则，会出现一个提示窗，用于配置规则。

这张表里的数据来源于OSS上传的日志文件，作为源头表，您需要尽早判断此表分区中是否有数据。如果这张表中没有数据，则需要阻止后续任务运行。如果来源表没有数据，后续任务运行无意义。

说明:

只有强规则下红色报警会导致任务阻塞，阻塞会将任务的实例状态置为失败。

在配置规则时，选择模板类型为表行数，将规则的强度设置为强，比较方式设置为期望值大于0，设置完毕后单击批量保存按钮即可。



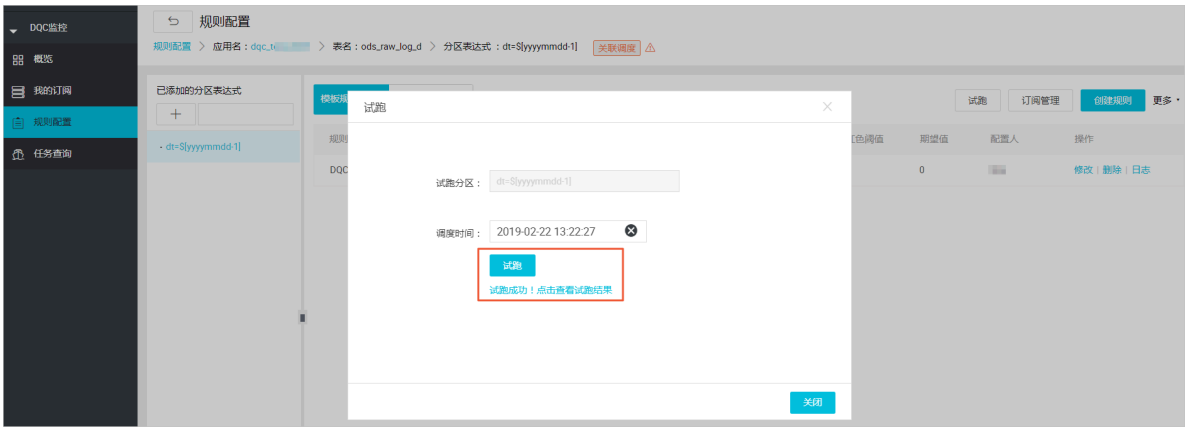
说明:

此配置主要是为了避免分区中没有数据，导致下游任务的数据来源为空的问题。

规则试跑

右上角有一个节点试跑的按钮，可以在规则配置完毕后，进行规则校验，试跑按钮可立即触发数据质量的校验规则。

单击试跑按钮后，会提示一个弹窗，确认试跑日期。单击试跑后，下方会有一个提示信息，单击提示信息，可跳转至试跑结果。



说明：

可根据试跑结果，来确认此次任务产出的数据是否符合预期。建议每个表规则配置完毕后，都进行一次试跑操作，以验证表规则的适用性。

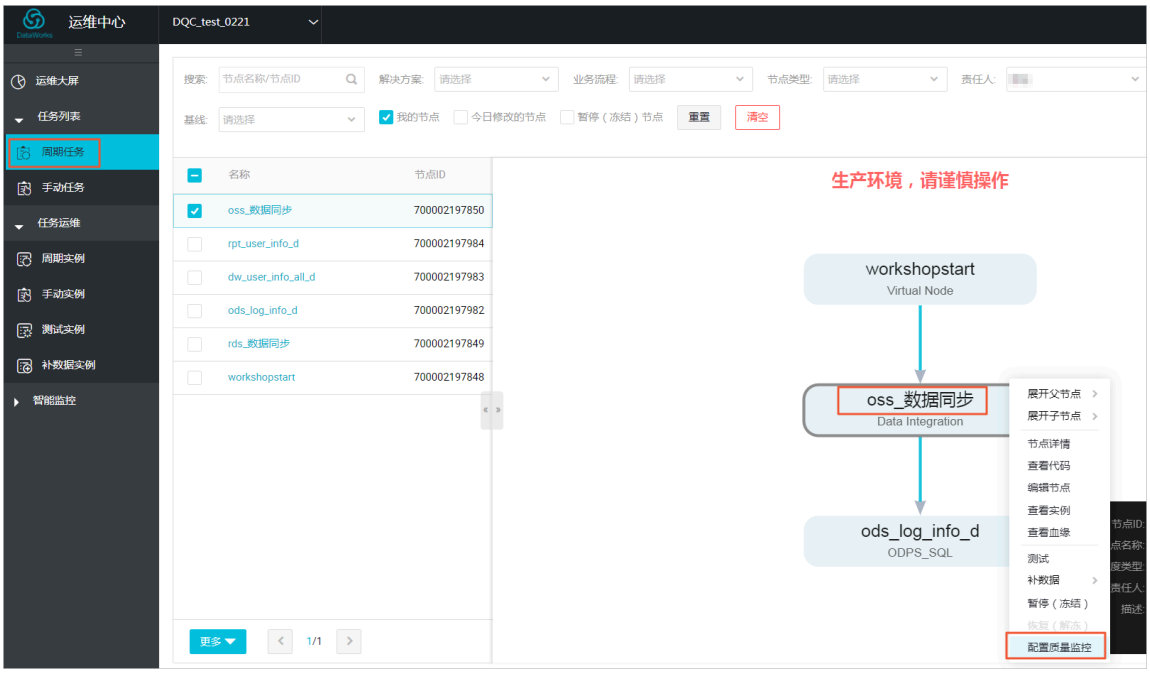
在规则配置完毕，且试跑又都成功的情况下。您需要将表和其产出任务进行关联，这样每次表的产出任务运行完毕后，都会触发数据质量规则的校验，以保证数据的准确性。

关联调度

数据质量支持和调度任务关联。在表规则和调度任务绑定后，任务实例运行完毕，都会触发数据质量的检查。表规则和任务关联调度有两种方式：

- 在运维中心的任务中进行表规则关联。
- 在数据质量的规则配置界面进行关联。
- 在运维中心页面关联表规则

  在运维中心 > 周期任务中，找到oss_数据同步任务，右键单击选择配置质量监控。



在弹窗中输入监控的表名，以及分区表达式。此处输入的表名为ods_raw_log_d。分区表达式参照上文规则配置的内容，即为`dt=$[yyyymmdd-1]`。

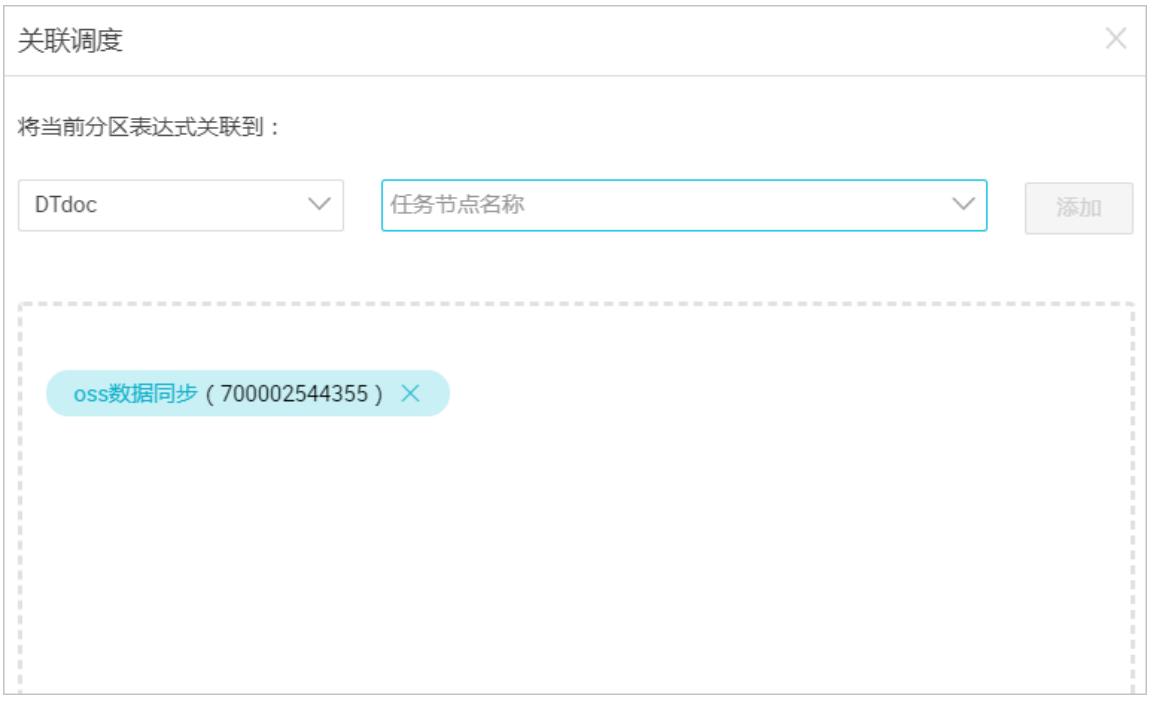
单击配置，即可跳转至规则配置界面。

- **在数据质量页面关联表规则**

  在表规则配置界面，单击关联调度，配置规则与任务的绑定关系。

单击关联调度，可以与已提交到调度的节点任务进行绑定，系统会根据血缘关系给出推荐绑定的任务，也支持自定义绑定。

选中搜索结果后，单击添加，添加完毕后即可完成与调度节点任务的绑定。

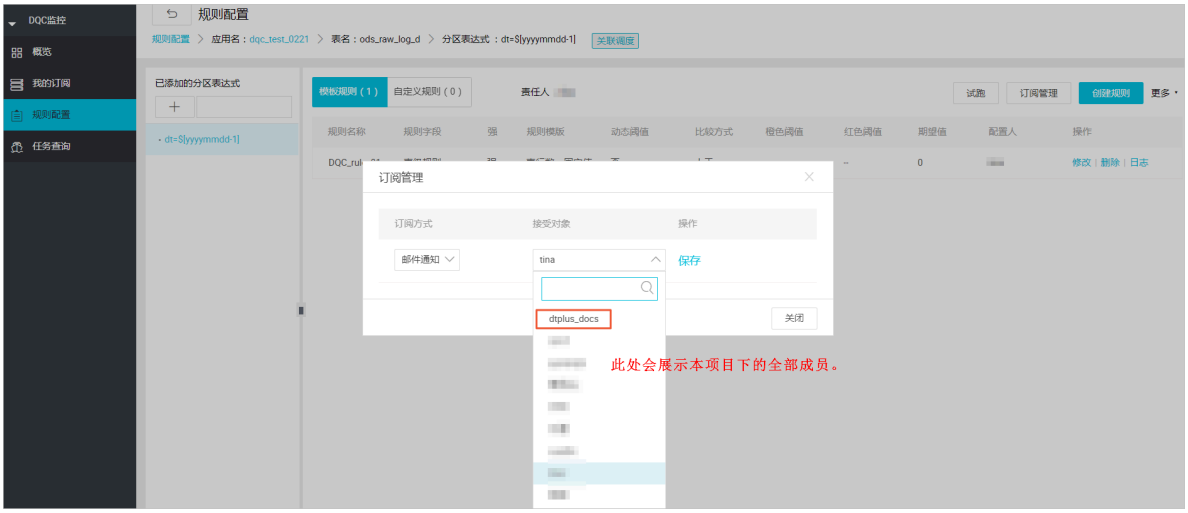


关联调度后，表名后面的小图标会变成蓝色。

配置任务订阅

关联调度后，每次调度任务运行完毕，都会触发数据质量的校验。数据质量支持设置规则订阅，可以针对重要的表及其规则设置订阅，设置订阅后会根据数据质量的校验结果进行告警，从而实现对校验结果的跟踪。若数据质量校验结果异常，则会根据配置的告警策略进行通知。

单击订阅管理，设置接收人以及订阅方式，目前支持邮件通知及邮件和短信通知。



订阅管理设置完毕后，可以在我的订阅中进行查看及修改。

建议将全部规则订阅，避免校验结果无法及时通知。

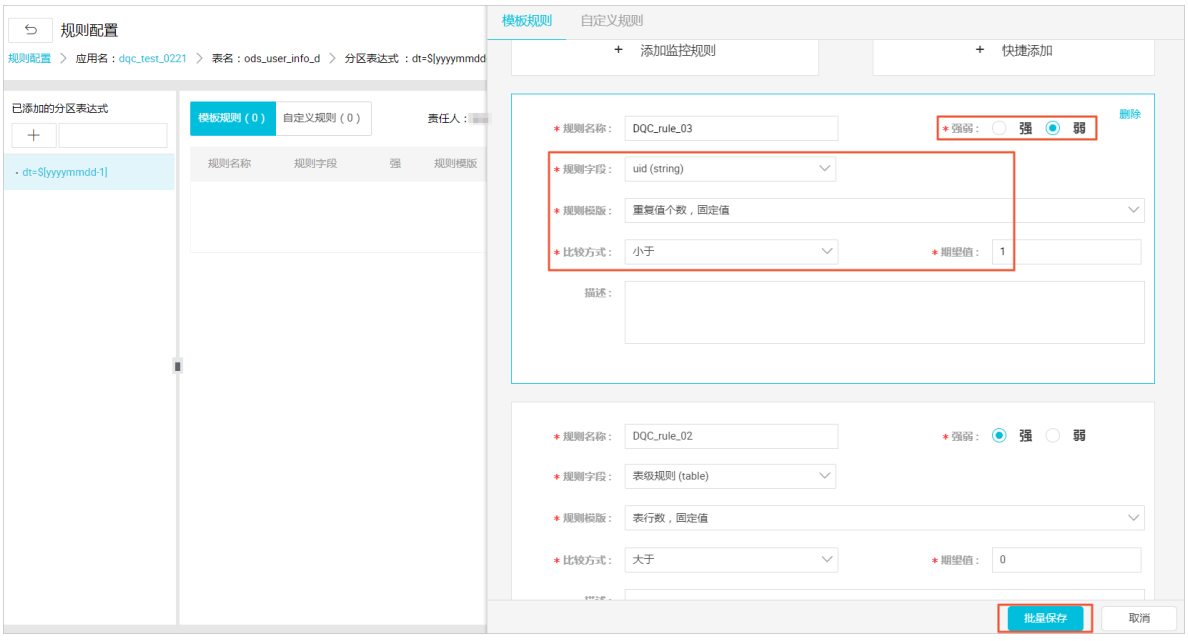
- ods_user_info_d

  ods_user_info_d是用户信息表。您在配置规则的时候，需要配置表的行数校验和主键唯一性校验，避免数据重复。

  同上，您需要先配置一个分区字段的监控规则，监控的时间表达式为：`dt=$[yyyymmdd-1]`。配置成功后，在已添加的分区表达式中可以看到成功的分区配置记录。

分区表达式配置完毕后，单击右侧的创建规则，进行数据质量的校验规则配置。添加表行数的监控规则，规则强度设置为强，比较方式设置为期望值大于0。

添加列级规则，设置主键列（uid）为监控列。模板类型为：字段重复值个数校验，规则设置为弱，比较方式设置为字段重复值个数小于1。设置完毕后，单击批量保存按钮即可。



说明:

此配置主要是为了避免数据重复，导致下游数据被污染的情况。

请不要忘记先试跑，再关联调度，最后规则订阅。

ods_log_info_d

ods_log_info_d数据主要来源于解析ods_raw_log_d表里的数据。鉴于日志中的数据无法配置过多监控，只需配置表数据不为空的校验规则即可。首先，配置表的分区表达式为`dt=$[yyyymmdd-1]`**。**

配置表数据不为空的校验规则，规则强度设置为强，比较方式设置为期望值不等于0。设置完毕后，单击批量保存。

- dw_user_info_all_d

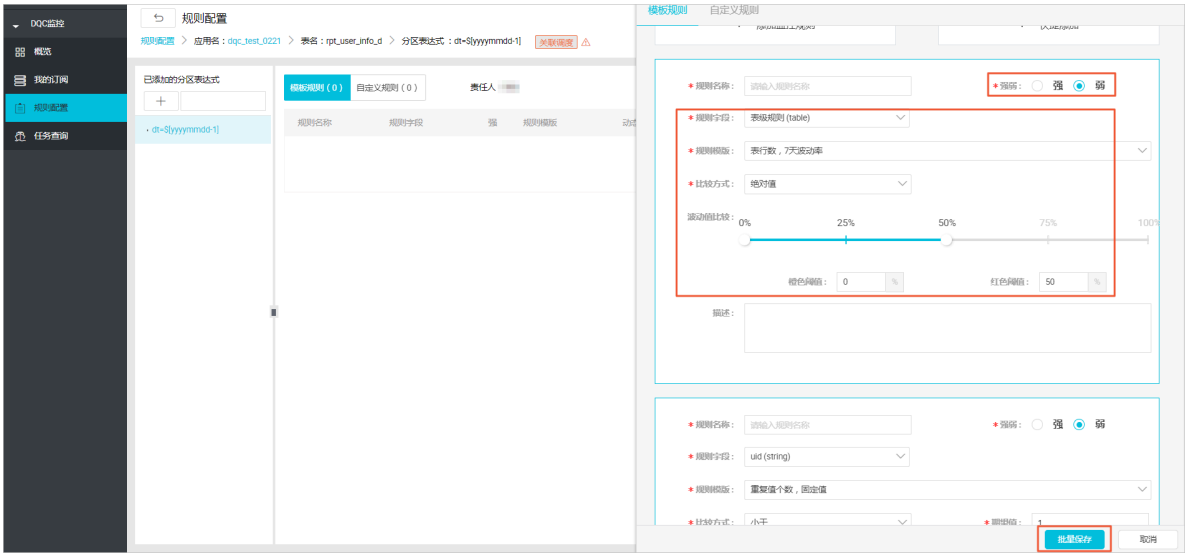
  dw_user_info_all_d表是针对ods_user_info_d和ods_log_info_d表的数据汇总，由于流程较为简单，ods层已配置了表行数不为空的规则，所以此表不进行数据质量监控规则的配置，以节省计算资源。

- rpt_user_info_d

  rpt_user_info_d表是数据汇总后的结果表。根据此表的数据，您可以进行表行数波动监测，以及针对主键进行唯一值校验。首先，配置表的分区表达式dt=$[yyyymmdd-1]。

然后配置监控规则，单击右侧创建规则 > 添加监控规则。添加列级规则，设置主键列（uid）为监控列，模板类型为：字段重复值个数校验，规则设置为弱，比较方式设置为字段重复值个数小于1。

继续添加监控规则和表级规则，模板类型为：SQL任务表行数，7天波动检测；规则强度设置为弱，橙色阈值设置成0%，红色阈值设置成50%（此处阈值范围根据业务逻辑进行设置），配置完毕后，单击批量保存即可。



说明:

此处您监控表行数是为了查看每日UV的波动，以便及时了解应用动态。

在设置表规则强度时，数据仓库中越底层的表，设置强规则的次数越多。这是因为ODS层的数据作为数仓中的原始数据，一定要保证其数据的准确性，避免因ODS层的数据质量太差而影响其他层的数据，及时止损。

数据质量还提供任务查询功能，以便查看已配置规则的校验结果。

1.6 数据可视化展现

通过补数据完成数据表rpt_user_info_d加工后，您可以通过Quick BI创建网站用户分析画像的仪表板，实现该数据表的可视化。

前提条件

在开始试验前，请确认您已经完成了#unique_29。单击进入Quick BI控制台。

背景信息

rpt_user_info_d表包含了region、device、gender、age、zodiac等字段信息。您可以通过仪表板展示用户的核心指标、周期变化、用户地区分布、年龄与星座分布和记录。为查看数据在日期上的变化，建议您在补数据时至少选择一周的时间。

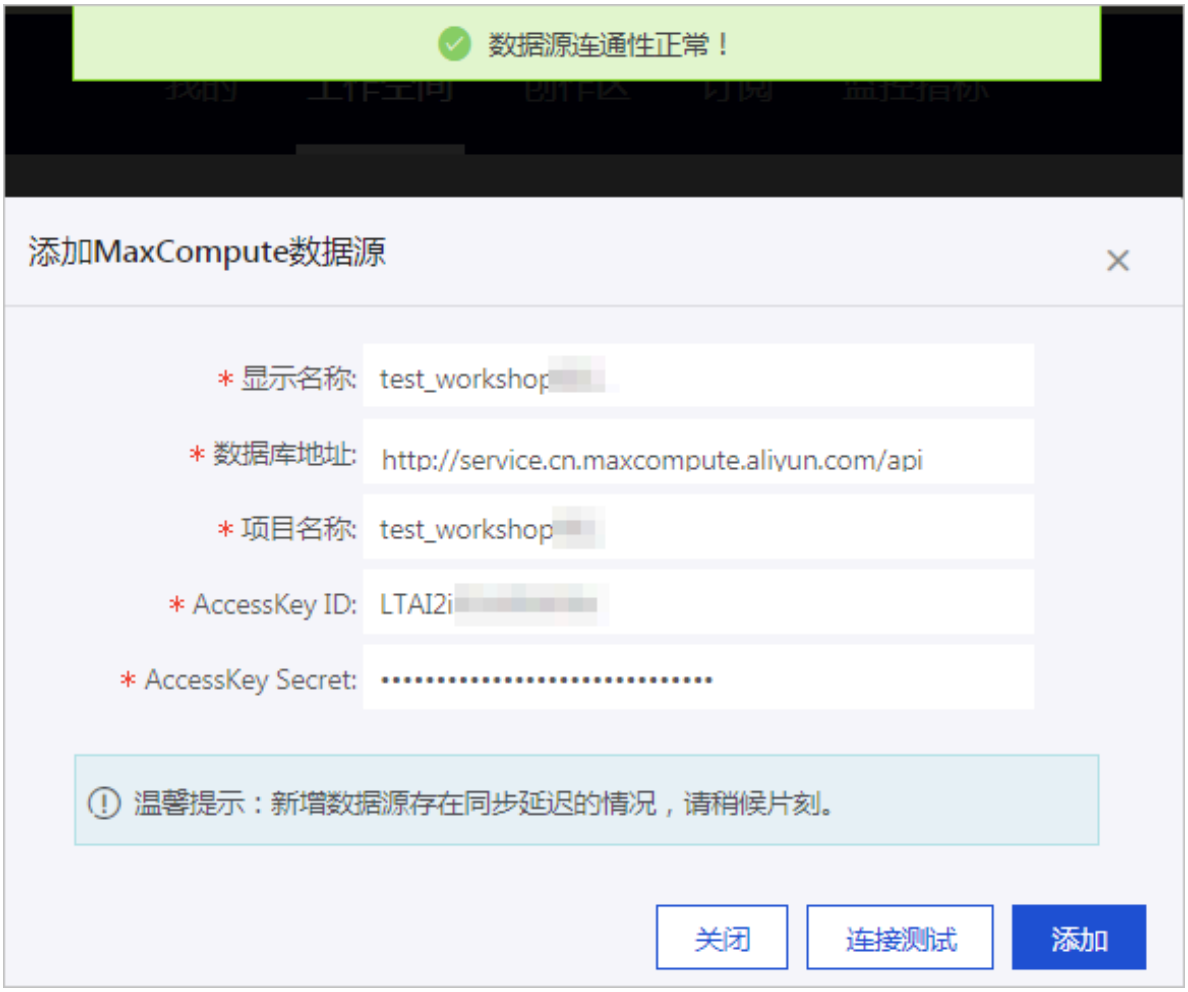
操作步骤

1. 单击进入默认空间，您也可以使用自己的个人空间。

2. 选择数据源 > 新建数据源 > 云数据库 > MaxCompute。

3. 输入您的MaxCompute项目名称以及您的AcessKey信息，数据库地址使用默认地址即可，关于数据库地址详情请参见#unique_30。

   完成填写后，单击连接测试，待显示数据源连通性正常后单击添加即可。



4. 找到您刚添加的数据源的rpt_user_info_d表，单击创建数据集。

选择您想放置的数据集位置，单击确定。

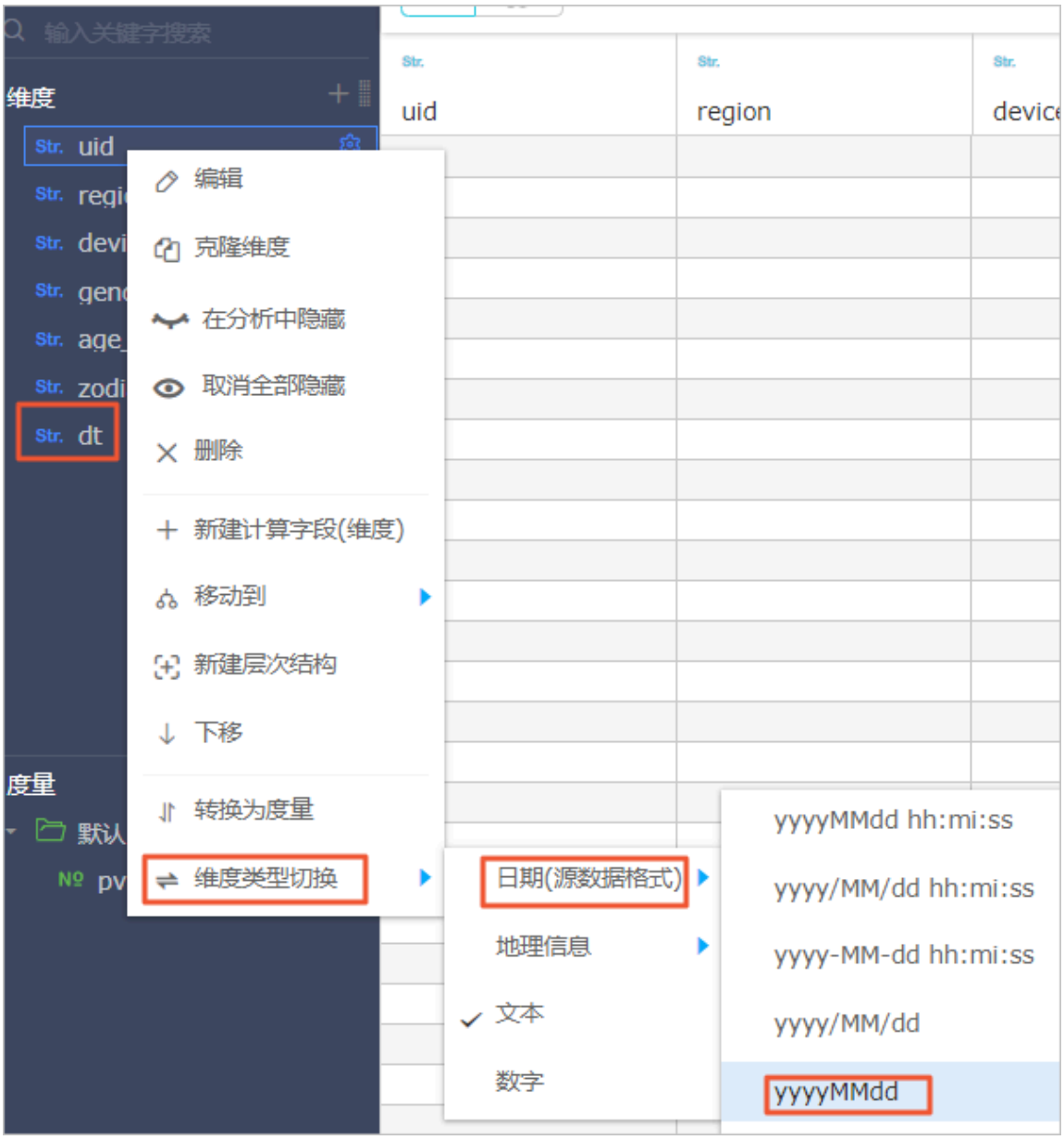
5. 进入数据集列表页，单击您刚刚创建的数据集，对数据集进行编辑。

常见的数据集加工包括：维度、度量的切换、修改维度的类型、增加计算字段、创建层次结构、修改字段的数据类型、更改度量聚合方式、制作关联模型。

6. 转换字段的维度类型。完成转换后，您可以根据字段中具体的数值进行过滤筛选。

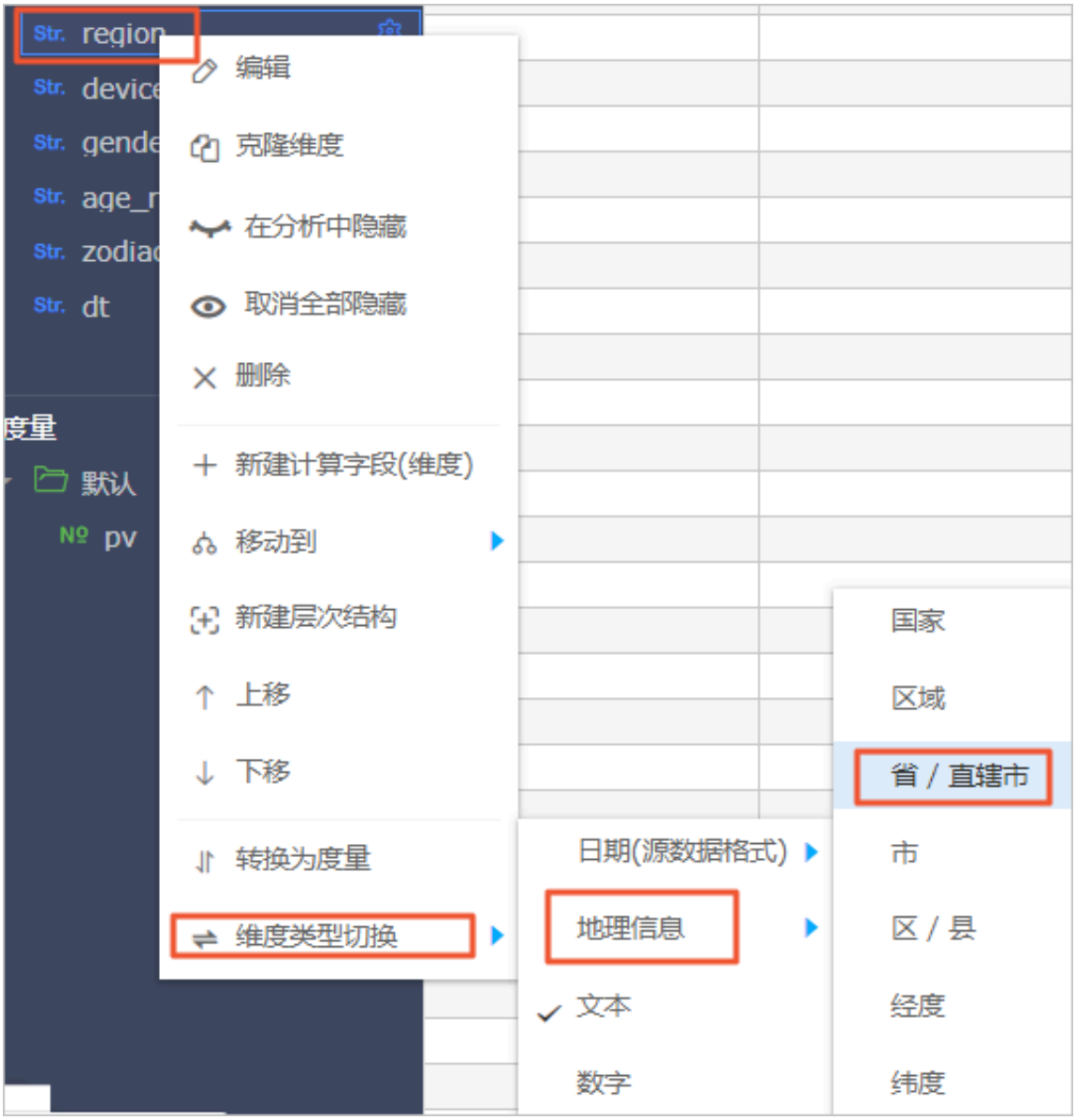
a) 转换日期字段的维度类型。

右键单击dt字段，选择维度类型切换 > 日期（源数据格式） > yyyyMMdd。



b) 转换地理信息字段的维度类型。

右键单击region字段，选择维度类型切换 > 地理信息 > 省/直辖市。转换成功后，在左侧维度栏中会看到字段前多一个地理位置图标。


Str. region
Str. device
Str. gende
Str. age_r
Str. zodiac
Str. dt
度量
默认
№ pv
编辑
克隆维度
在分析中隐藏
取消全部隐藏
删除
新建计算字段(维度)
移动到
新建层次结构
上移
下移
转换为度量
维度类型切换
日期(源数据格式)
地理信息
文本
数字
国家
区域
省 / 直辖市
市
区 / 县
经度
纬度

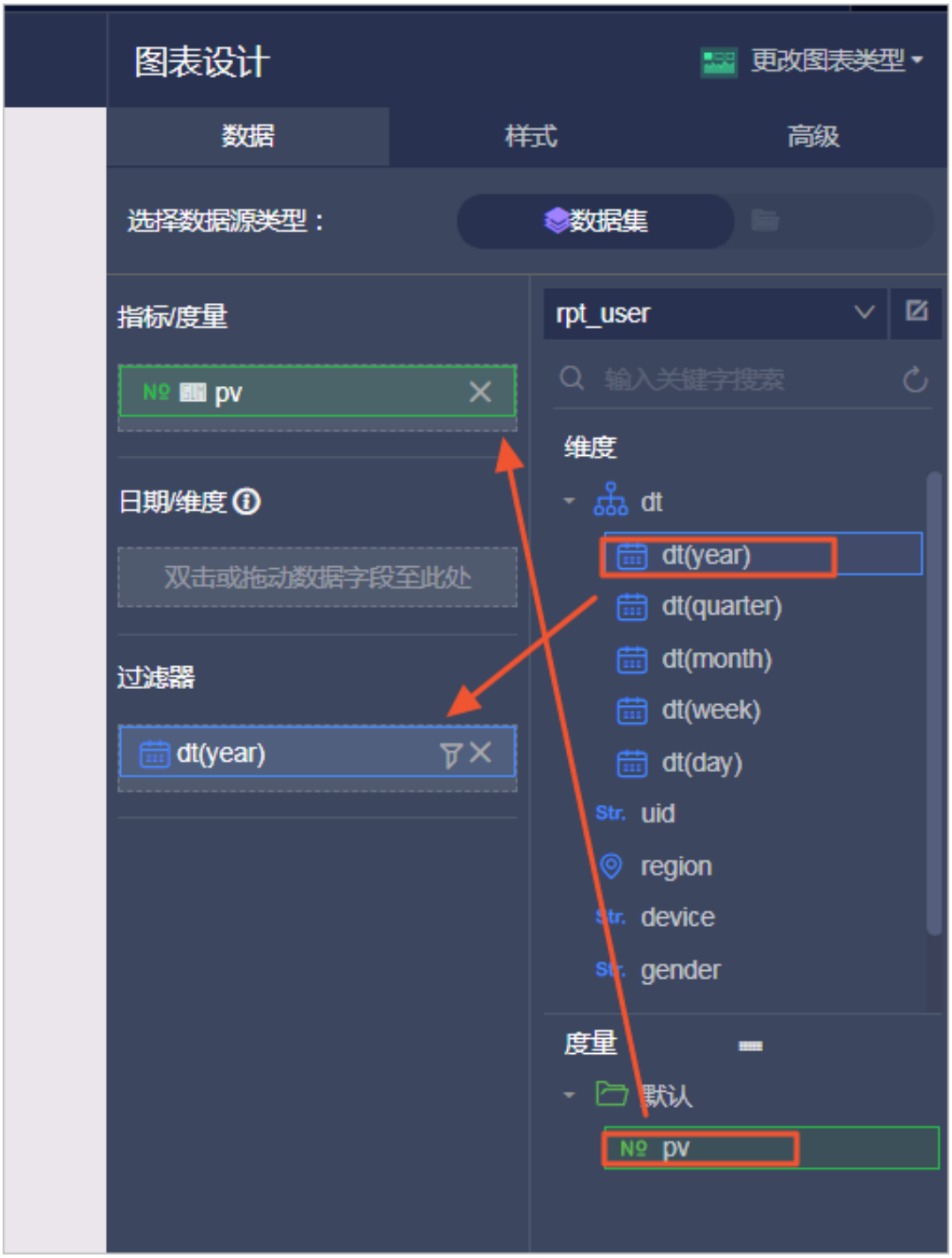
7. 制作仪表板。

随着数据的更新，让报表可视化地展现最新数据，这个过程叫制作仪表板。仪表板的制作流程为： 确定内容、布局和样式，制作图表，完成动态联动查询。

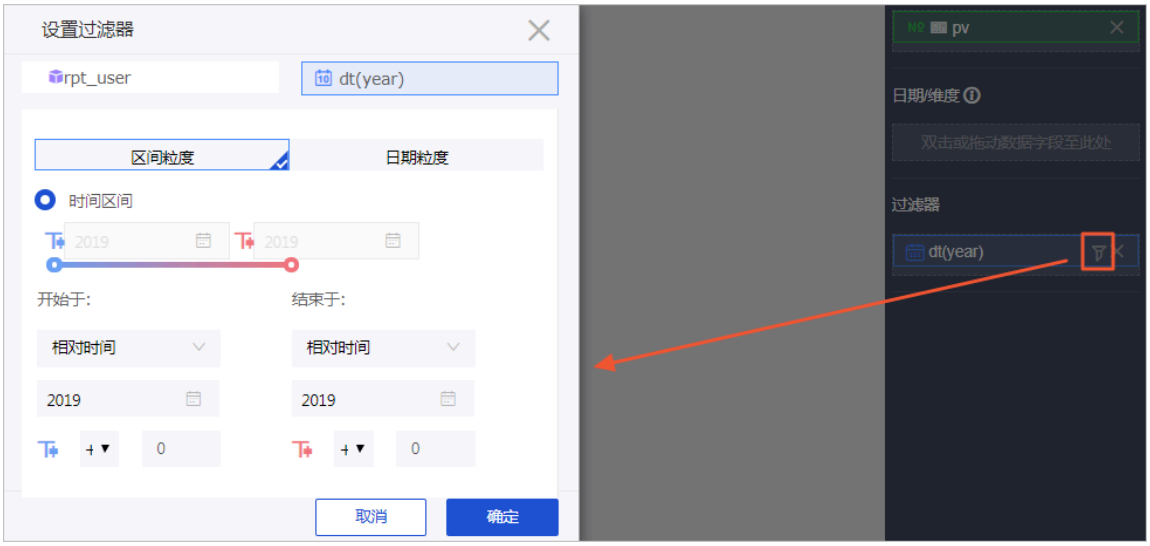
a) 单击rpt_user数据集后的新建仪表板，选择常规模式，进入仪表板编辑页。

b) 从仪表板空间中向空白区拖入1个指标看板。

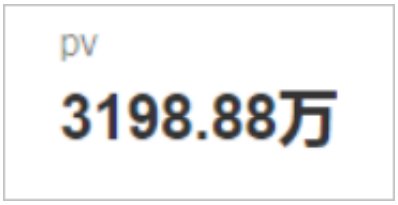
选择数据来源为数据集rpt_user，选择度量为pv。



由于数据表rpt_user_info_d为分区表，因此必须在过滤器处选择筛选的日期，本例中筛选为2019~2019年，完成设置后单击更新。



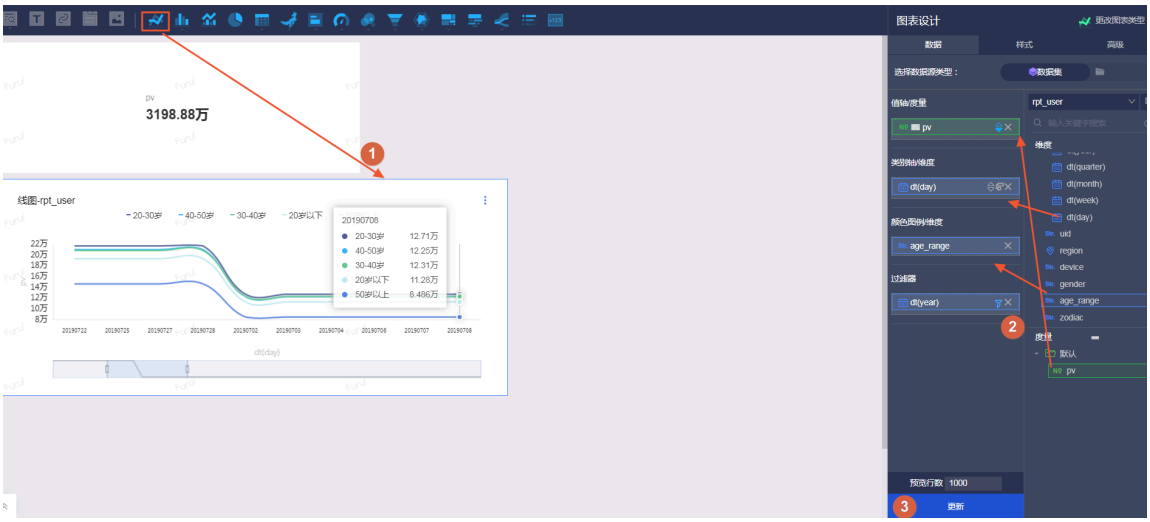
完成后可以看到当下数据。



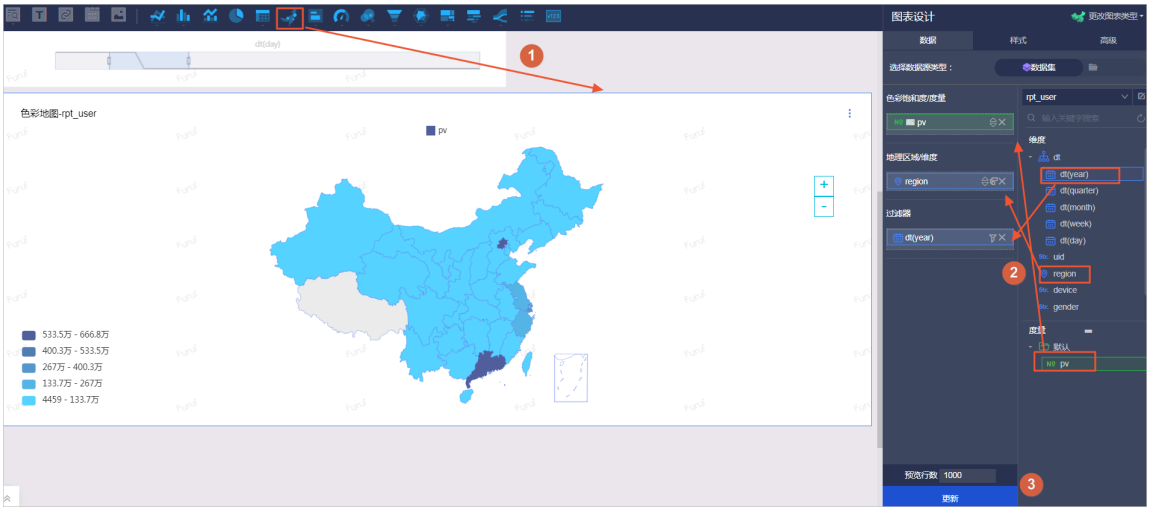
c) 制作趋势图：将图表区域内的线图拖拽到左侧画布。

参数配置如下，完成之后单击更新：

- 值轴/度量：pv
- 类别轴/维度：dt（day）
- 颜色图例/维度：age_range
- 过滤器：dt（year）



d) 制作色彩地图：单击图表区域内的色彩地图，并选择数据源来源为数据集rpt_user，选择地理区域/维度为region、色彩饱和度/度量为pv，选择完成后单击更新，结果如下。



e) 完成配置后，单击保存及预览，即可看到展示效果。

1.7 通过Function Studio开发UDF

本文将为您介绍如何通过Function Studio开发UDF，并将其提交至DataStudio的开发环境。

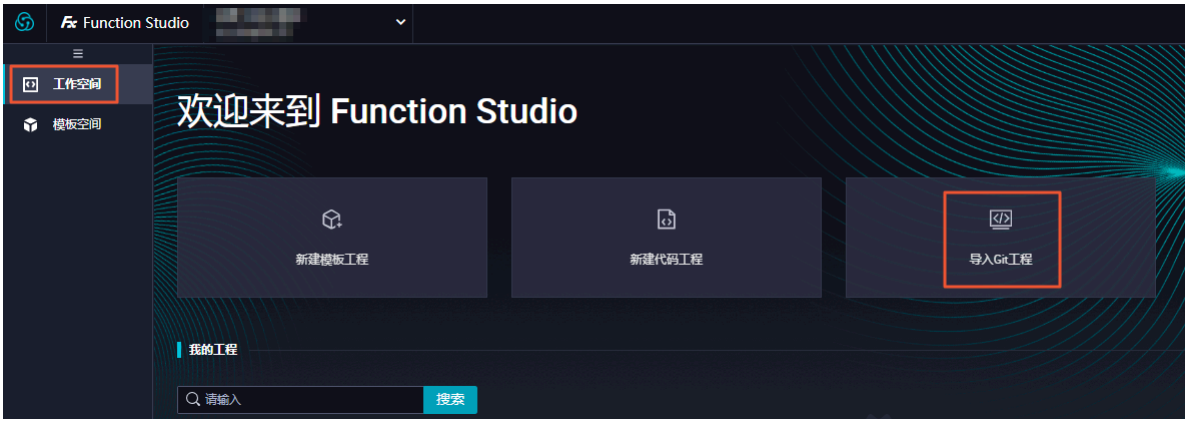
新建工程

如果您已经有Git代码，可以直接导入Git代码创建工程。此处仅支持Code中的代码导入。

1. 登录DataWorks控制台，进入DataStudio（数据开发）页面。

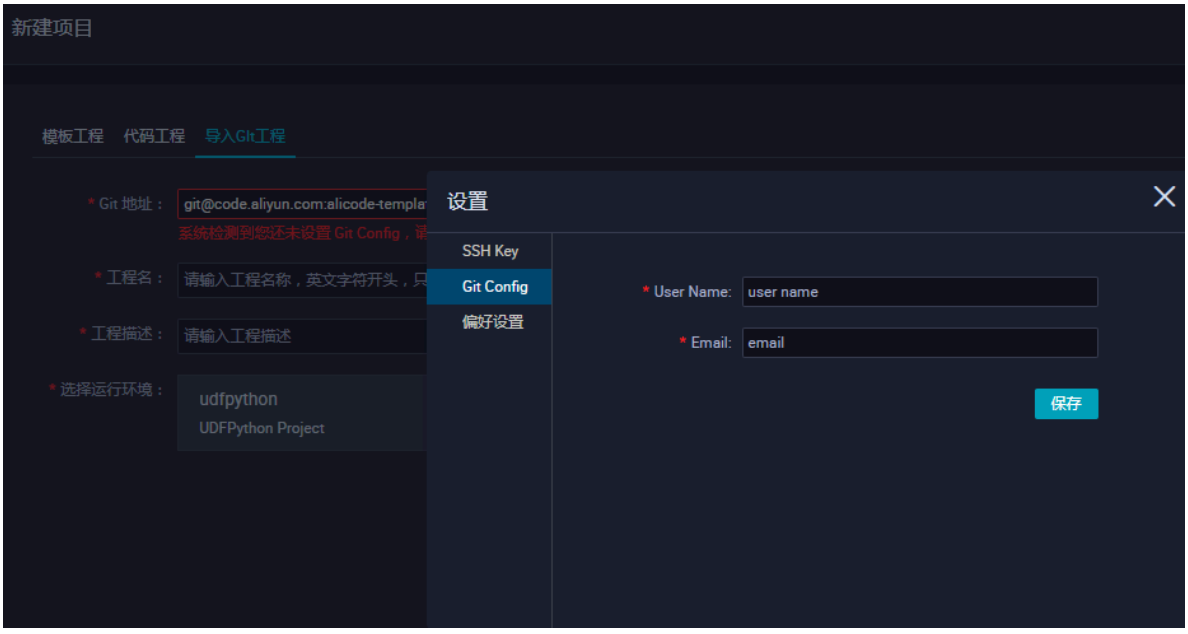
2. 单击左上角的图标，鼠标悬停至全部产品，选择数据开发 > Function Studio。

3. 单击工作空间页面的导入Git工程。

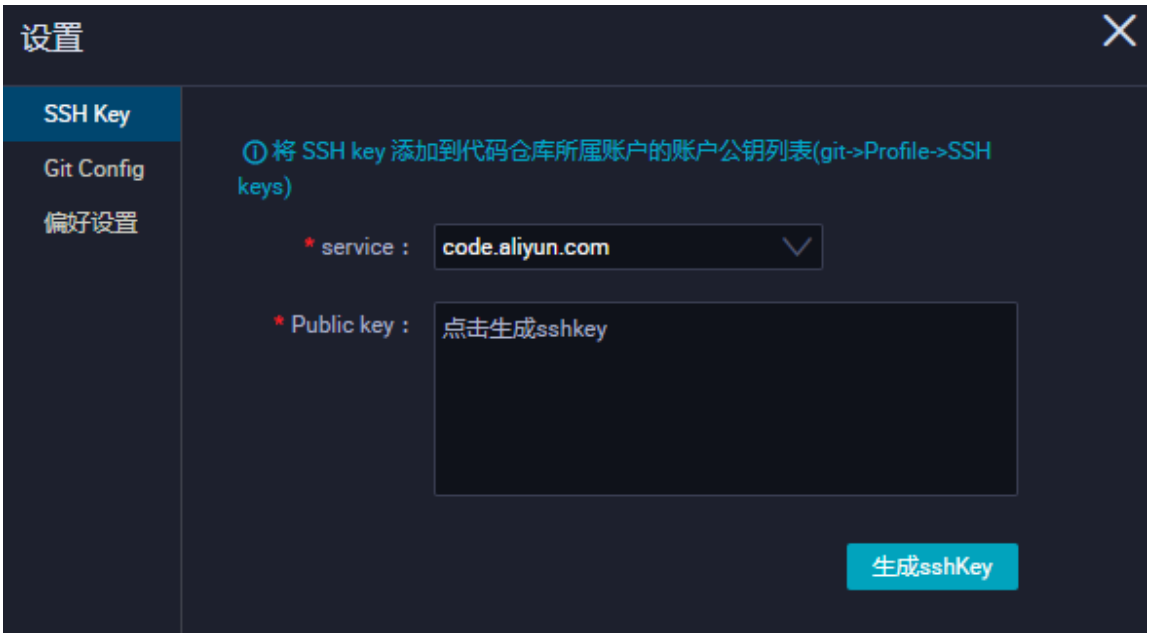


4. 填写新建项目对话框中的Git地址、工程名和工程描述，并选择运行环境。

新创建的工程默认未关联Git服务，会弹出设置对话框，请首先进行SSH KEY、Git Config和偏好设置的配置，单击保存。



- 选择SSH Key中的service为`code.aliyun.com`，单击生成sshKey，即可生成Public key，单击保存。



- 填写Git Config中的User Name和Email，单击保存。

- 根据自身需求选择偏好设置中的编辑器字号，单击保存。

> 说明:
> 如果工程创建完成后，需要修改相关信息，可以鼠标悬停至顶部菜单栏中的设置进行修改。

5. 单击提交。
   工程创建完成后，Function Studio会自动拉取该工程。

新建SSH密钥

设置好SSH KEY、Git Config和偏好设置后，可以新增SSH密钥。

1. 访问Code页面，单击左侧导航栏中的设置。
2. 进入设置页面，选择SSH公钥 > 增加SSH密钥。

3. 在增加SSH密钥对话框中填写前文生成的Public key，单击增加密钥。

测试需要运行的类

1. 打开需要运行的类，单击右上角的运行按钮进行测试。

2. 在Run/Debug Configurations对话框中，手动添加测试类的信息。

3. 添加完成后，单击Run，即可看到输出的测试信息。

说明:

- 第一次启动时速度较慢，之后的启动速度会逐渐接近本地编辑器的体验。
- 如果需要运行的类已经存在，直接在右上角进行选择，单击运行按钮即可。

提交函数和资源至DataStudio开发环境

确认代码无误后，可以提交函数和资源至DataStudio开发环境。

- 提交资源至DataStudio开发环境。
  1. 鼠标悬停至提交按钮，单击提交资源至DataStudio开发环境。
  2. 选择提交资源至DataStudio开发环境对话框中的目标业务空间和目标业务流程，并填写资源。
  3. 单击确认。
- 提交函数至DataStudio开发环境。
  1. 鼠标悬停至提交按钮，单击提交函数至DataStudio开发环境。
  2. 选择提交函数至DataStudio开发环境对话框中的目标业务空间、目标业务流程和类名，并填写资源和函数名。
  3. 单击确认。

当资源和函数都提交至DataStudio开发环境后，即可直接在SQL节点中使用。

2 搭建互联网在线运营分析平台

2.1 业务场景与开发流程

本教程基于大数据时代在线运营分析的平台的基础需求，为开发者提供从数据高并发写入存储、便捷高效的数据加工处理到数据分析与展示的全链路解决方案。本教程助您了解并实操阿里云的大数据产品，完成在线运营分析平台的搭建。

业务场景

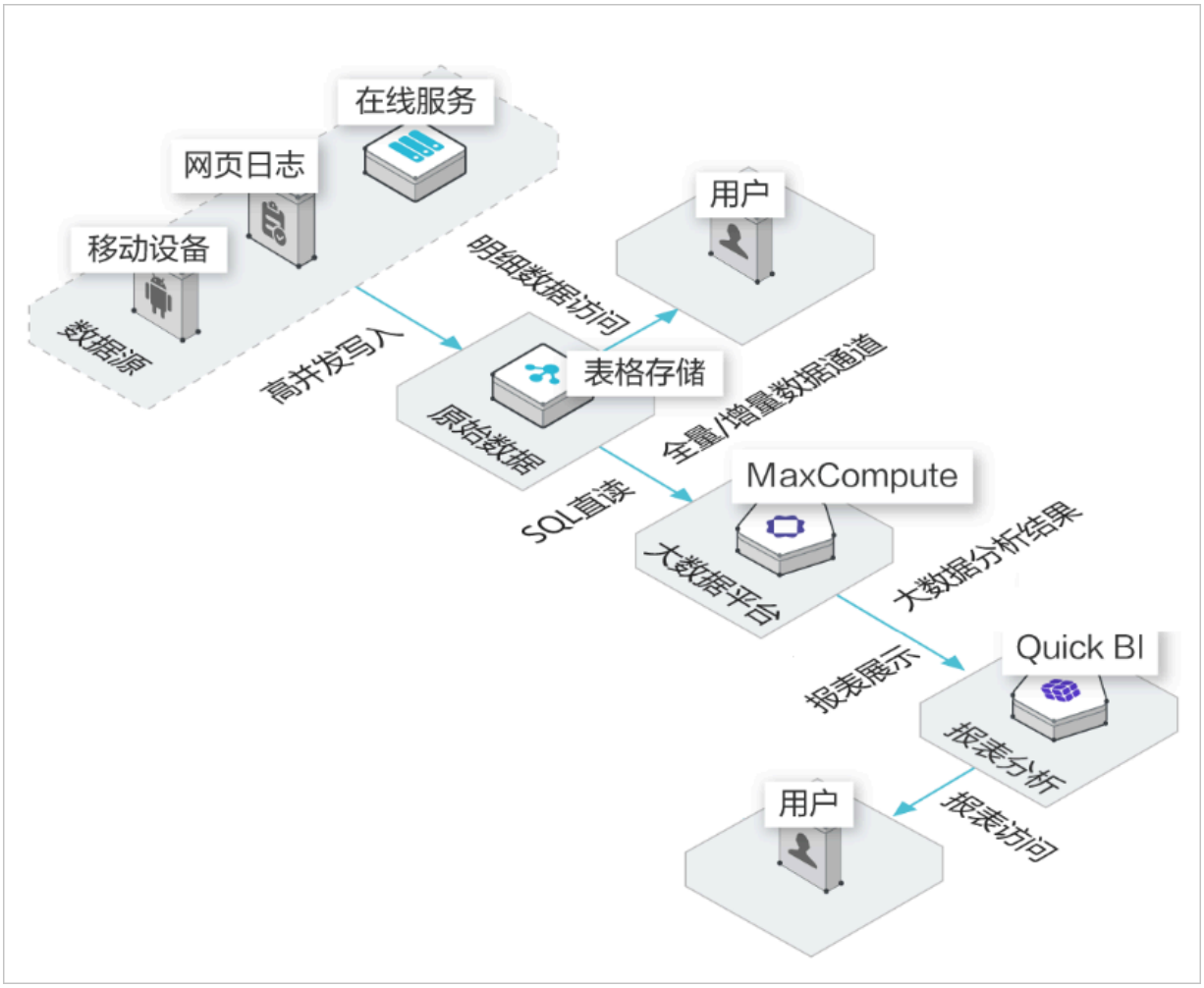
本节的示例说明基于一份真实的网站日志数据集，数据来源于某网站上的HTTP访问日志数据。基于这份网站日志，您可以实现如下分析需求：

- 统计并展现网站的PV和UV，并能够按照用户的终端类型（如Android、iPad、iPhone、PC等）分别统计。

  说明:
  浏览次数（PV）和独立访客（UV）是衡量网站流量的两项最基本指标。用户每打开一个网站页面，记录一个PV，多次打开同一页面PV累计多次。独立访客是指一天内访问网站的不重复用户数，一天内同一访客多次访问网站只计算一次。

- 统计并展现网站的流量来源地域。

开发流程



本教程涉及的具体开发流程如下：

- 步骤一：#unique_34。
- 步骤二：#unique_35。
- 步骤三：#unique_36。
- 步骤四：#unique_37。
- 步骤五：#unique_38。
- 步骤六：任务提交与测试。
- 步骤七：#unique_40。

整体数仓研发的规划建议请参见#unique_41。

2.2 环境准备

为保证您可以顺利完成本教程，请您首先确保自己云账号已开通表格存储TableStore、大数据计算服务MaxCompute、数据工场DataWorks和智能分析套件Quick BI。

前提条件

- 阿里云账号注册，详情请参见#unique_13。
- 实名认证，详情请参见#unique_14或#unique_15。

背景信息

本教程涉及的阿里云产品如下：

- 表格存储 TableStore
- 大数据计算服务 MaxCompute
- 数据工场 DataWorks
- 智能分析套件Quick BI

说明：

在本教程中，表格存储服务选择华北2区域。

操作步骤

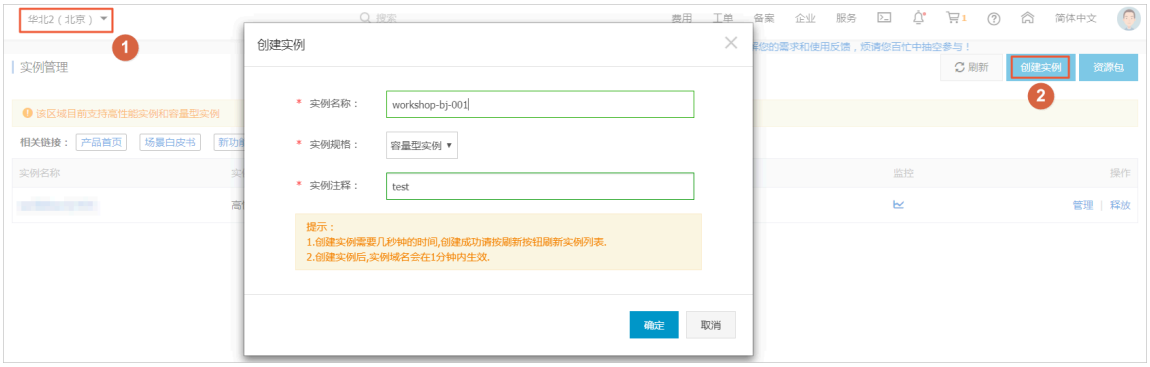
1. 创建表格存储实例

a) 进入表格存储TableStore产品详情页，单击立即开通。

b) 进入开通页面后，单击立即开通。

c) 单击管理控制台。

d) 选择区域为华北2（北京），单击创建实例。填写实例名称，实例规格请选择容量型实例，单击确定完成创建。



说明：

实例名称在表格存储同一个区域内必须全局唯一，建议您选用自己可辨识且符合规则的名称。实例名称在MaxCompute数据处理中也会被实用，本例中为`workshop-bj-001`，关于实例的详细解释请参见#unique_43。

e) **完成创建后，您可以在实例列表 > 实例管理中看到您刚刚创建的实例，状态为运行中。**

2. **开通大数据计算服务MaxCompute**

a) **进入MaxCompute产品详情页，单击立即购买。**

b) **选择按量付费，选择区域为华东2（上海），规格类型为默认的标准版，单击立即购买。**

说明：

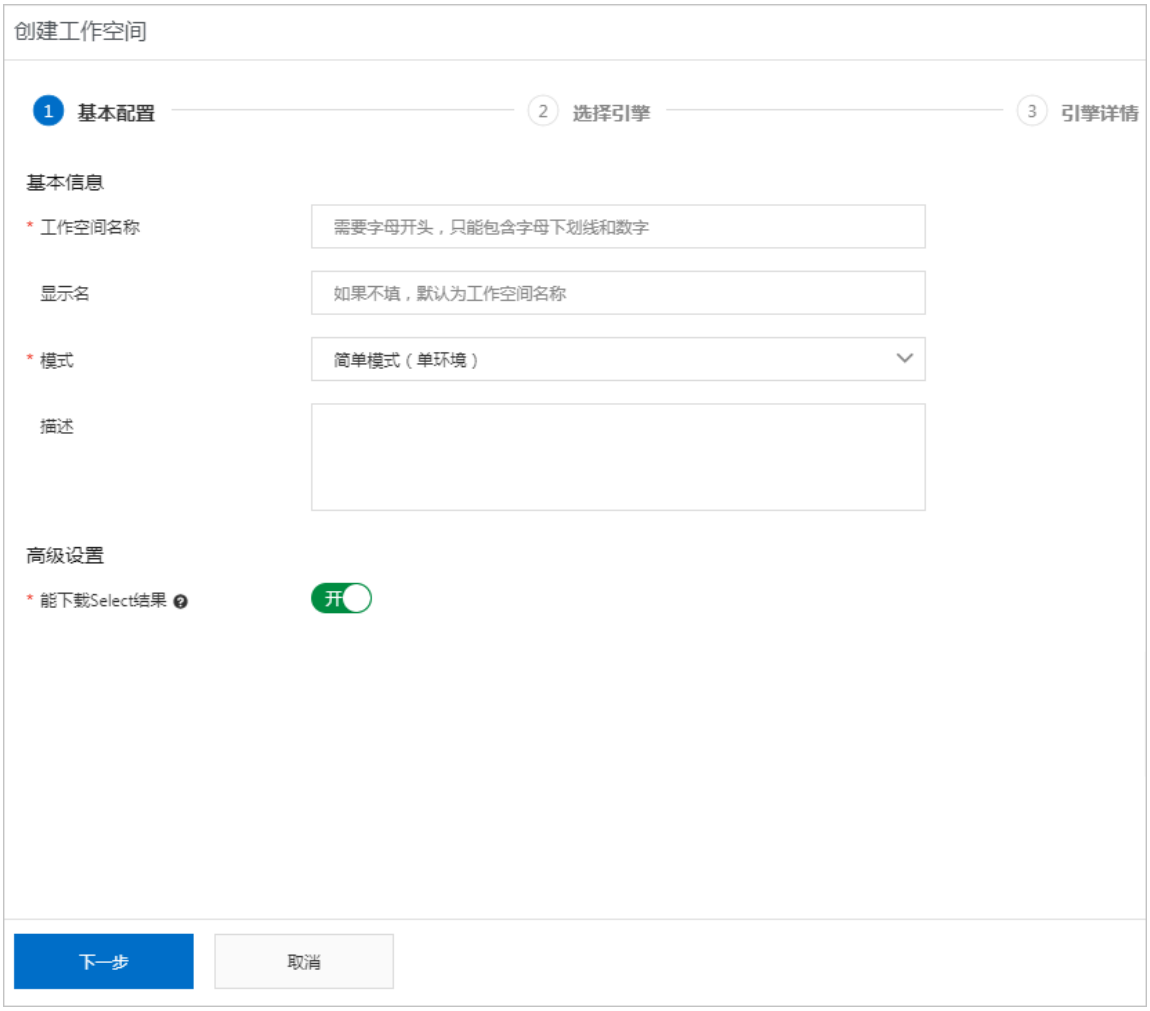
选择MaxCompute区域与表格存储相同可以节省您的流量费用，因此您可以选择区域为华北2（北京）。本例中MaxCompute区域选择为华东2（上海），以便为您展示跨地域的外部表使用过程。

3. 创建DataWorks工作空间

a) 进入DataWorks工作空间列表，选择区域为华东1，单击创建工作空间。

b) 填写创建工作空间对话框中的基本配置，单击下一步。

为方便使用，本教程中DataWorks工作空间模式为简单模式（单环境）。在简单模式下，DataWorks工作空间与MaxCompute项目一一对应，详情请参见#unique_44。

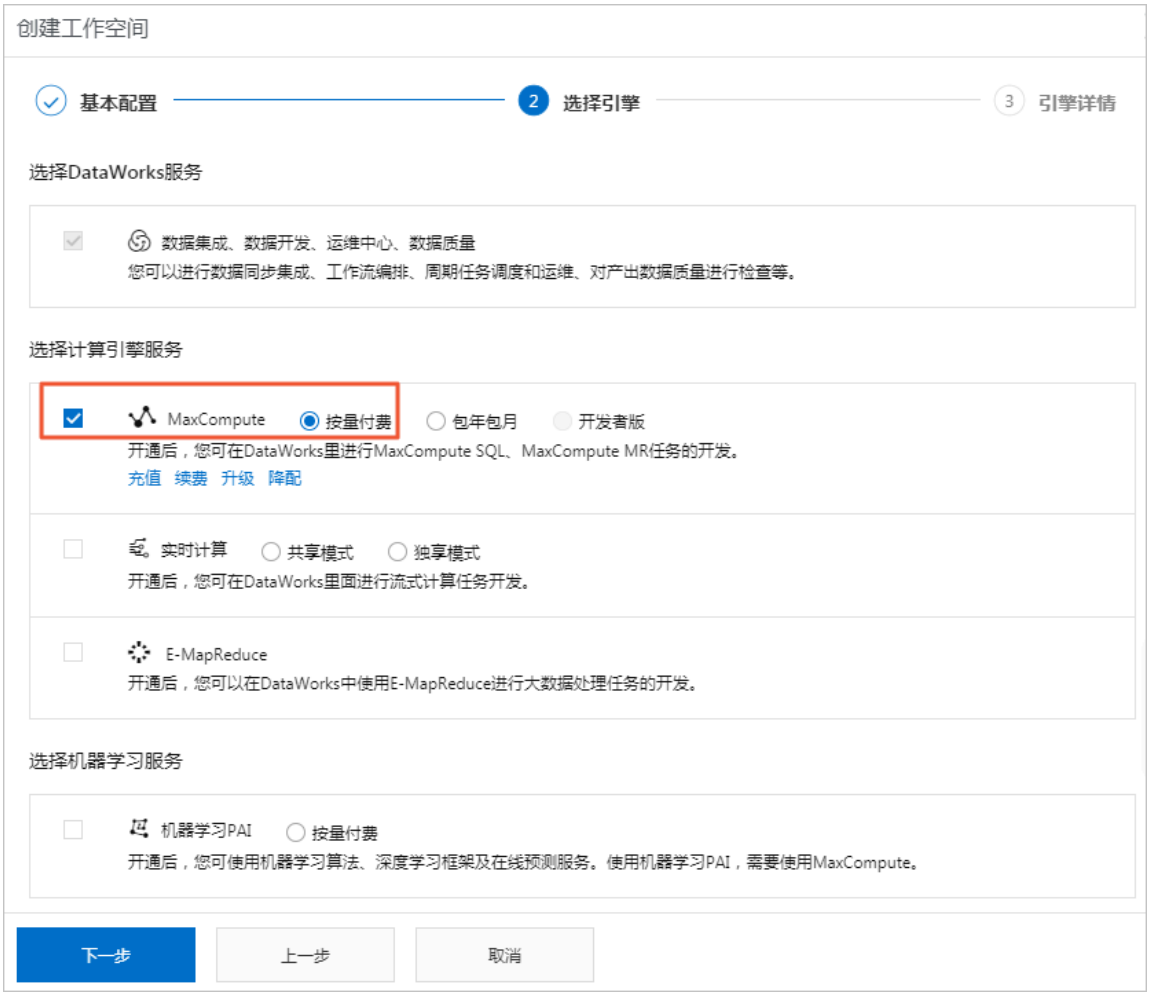


说明:

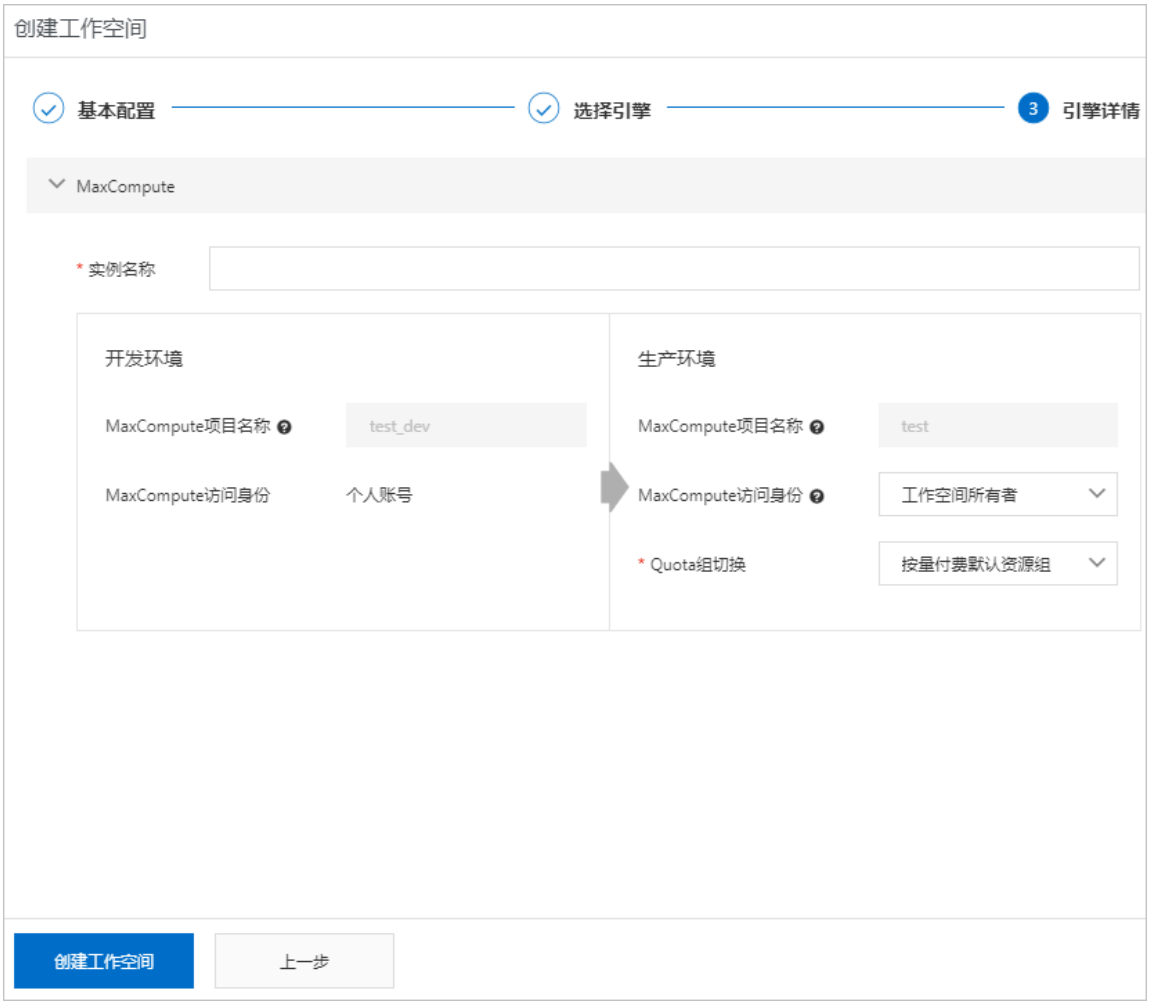
工作空间名称全局唯一，建议您使用易于区分的名称。

c) 进入选择引擎界面，选择相应引擎后，单击下一步。

选择计算引擎服务为MaxCompute、按量付费。



d) 进入引擎详情页面，填写选购引擎的配置。



分类	配置	说明
MaxCompute	实例名称	实例名称不能超过27个字符，仅支持字母、中文开头，仅包含中文、字母、下划线和数字。
	MaxCompute项目名称	默认与DataWorks工作空间的名称一致。
	MaxCompute访问身份	包括个人账号和工作空间所有者，开发环境默认为个人账号，生产环境推荐使用工作空间所有者。
	Quota组切换	Quota用来实现计算资源和磁盘配额。
PAI	使用GPU	默认不使用，如果需要使用，请前往工作空间配置页面开启GPU使用。

e) 配置完成后，单击创建工作空间。

工作空间创建成功后，即可在工作空间列表页面查看相应内容。

4. 开通Quick BI

a) 进入Quick BI产品详情页，单击管理控制台。

b) 进入控制台后，单击高级版30天试用申请或专业版30天试用申请。勾选同意Quick BI服务协议，单击开通试用。成功开通Quick BI专业版试用后的界面如下图所示。

说明:

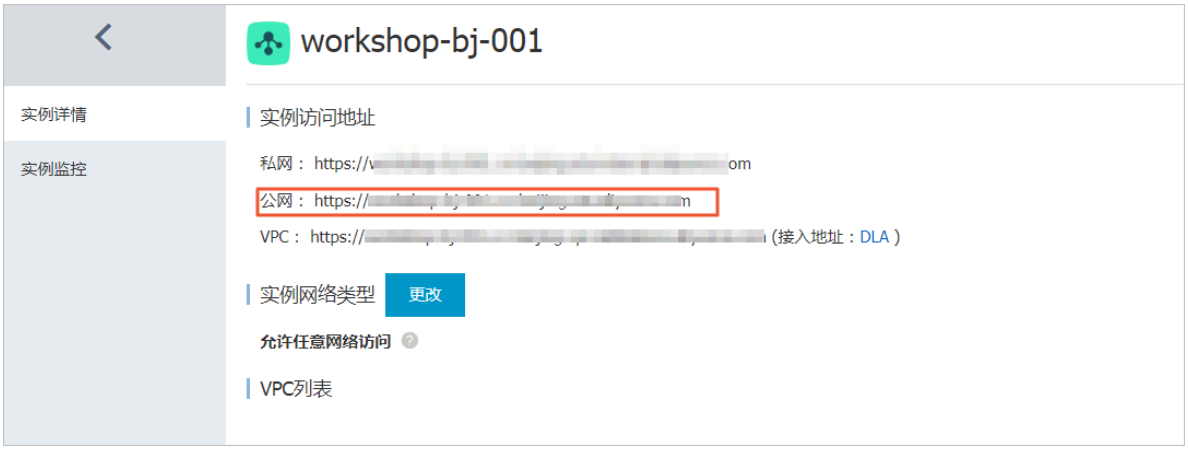
您可以选择使用个人空间或默认空间，推荐您使用默认空间。

2.3 数据准备

在数据准备阶段，您需要通过数据Demo包生成出模拟真实环境的数据，以便后续数据开发使用。

前提条件

1. 创建华北2区域的表格存储示例，同时记录实例名称和实例访问地址。详情请参见环境准备。
2. 单击表格存储控制台中的实例名称后，您可以获得实例访问地址。对于跨区域的访问，建议您使用公网地址。



3. 使用主账号登录安全信息管理控制台，获取并记录您的AccessKey ID和AccessKey Secret信息。

> **说明:**
> AccessKey ID和AccessKey Secret是您访问阿里云API的密钥，具有该账户完全的权限，请您妥善保管。

操作步骤

1. 下载数据Demo包

   数据Demo包下载地址如下，本例中使用环境为Windows7 64位：

   - Mac下载地址
   - Linux 下载地址
   - Windows7 64位 下载地址

2. 配置Demo环境

完成下载后，您需要解压下载包，编辑conf文件夹内的app.conf文件。

app.conf文件内容示例如下，其中endpoint信息即为您的实例访问地址。

```
endpoint = "https://workshop-bj-001.cn-beijing.ots.aliyuncs.com"
instanceName = "workshop-bj-001"
accessKeyId = "LTAIF24u7g******"
accessKeySecret = "CcwFeF3sWTPy0wsKULMw34Px******"
usercount = "200"
daysCount = "7"
```

3. 启动Demo准备测试数据

启动Windows CMD命令行工具，进入您解压缩Demo包的路径，您可以使用`workshop_demo.exe -h`命令查看Demo包命令用法。

```
workshop_demo.exe -h 会列出该demo的相关命令:
* prepare:  准备测试数据，创建数据表，根据conf中的用户数量，为用户生成一周的行为日志数据。
* raw ${userid} ${date} ${Top条数}:  查询指定用户的日志明细。
```

* new/day_active/month_active/day_pv/month_pv：在结果表中查询上述几种类型的报表数据（新增：new，日活：day_active，月活：month_active，日PV：day_pv，月PV：month_pv）。

执行workshop_demo.exe prepare命令生成准备数据。

```
C:\Users [redacted]\workshop_demo>workshop_demo.exe prepare
OTSObjectAlreadyExist Requested table already exists.
OTSObjectAlreadyExist Requested table already exists.
Prepare the metric data
Prepare the metric data
Prepare the metric data
Prepare the metric data
Prepare the metric data
Prepare User data
finished one round
finished one round
finished one round
finished one round
finished one round
total insert data count is: 41757
```

在此过程中，Demo包会自动帮助您在表格存储中创建表，结构如下：

· 原始日志数据表：user_trace_log

列名	类型	说明
md5	STRING	用户uid的md5值 undefined前8位，表格存储主键。
uid	STRING	用户uid，表格存储主键。
ts	BIGINT	用户操作时间戳，表格存储主键。
ip	STRING	IP地址。
status	BIGINT	服务器返回状态码。
bytes	BIGINT	返回给客户端的字节数。
device	STRING	终端型号。
system	STRING	系统版本：ios xxx/android xxx。
customize_event	STRING	自定义事件：登录/退出/购买/注册/点击/后台/切换用户/浏览。
use_time	BIGINT	APP单次使用时长，当事件为退出、后台、切换用户时有该项。

列名	类型	说明
customize_event_content	STRING	用户关注的内容信息。

- 分析结果表： analysis_result

列名	类型	说明
metric	STRING	报表的类型：'new'、'day_active'、'month_active'、'day_pv'、'month_pv'，表格存储主键。
ds	STRING	时间yyyy-mm-dd或yyyy-mm，表格存储主键。
num	BIGINT	对应的数据值。

4. 数据验证

· 用户明细查询

表格数据对应的日期对应于您创建表格的时间，例如您创建数据时间为2019年6月15日，则可以使用`workshop_demo.exe raw 00010 "2019-06-15" 20`查看20条用户明细数据。

```
C [redacted]nloads\workshop_demo>workshop_demo.exe raw 00010 "2019-06-1
5" 20
       uid                          Date    bytes      customize_event
      device                    ip   status          system
     00010         2019-06-14 11:56:47 PM      759               regist
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 11:26:34 PM      252     backstage    369
   iPad min2       157.249.67.241      200          ios11
     00010         2019-06-14 11:21:30 PM      427        browse travel
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 11:16:03 PM      764        switch    185
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 11:06:03 PM      436                click
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 10:36:54 PM      131                click
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 10:22:26 PM      778        switch     73
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 10:06:29 PM      535     backstage    179
   iPad min2       157.249.67.241      200          ios11
     00010         2019-06-14 09:56:11 PM      668                click
   iPad min2       157.249.67.241      200          ios11
     00010         2019-06-14 09:20:45 PM      354               regist
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 09:15:37 PM      989                click
   iPad min2       157.249.67.241      200          ios11
     00010         2019-06-14 08:51:17 PM      460        logout    462
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 08:26:06 PM      887       comment  funny
   iPad min2       157.249.67.241      200          ios11
     00010         2019-06-14 08:10:34 PM      278       browse finance
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 07:56:00 PM      480                click
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 07:30:11 PM       68                click
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 07:15:09 PM      398        browse   news
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 07:11:21 PM       21                click
    iPhone6s      222.133.108.234      200          ios10
     00010         2019-06-14 06:35:07 PM      207        browse  photo
iPhone7 Plus        61.103.79.217      200          ios11
     00010         2019-06-14 06:24:43 PM      261               regist
iPhone7 Plus        61.103.79.217      200          ios11
```

说明:

由于表格存储是SchemaFree结构，表的属性列不需要预先定义。customize_event 中不同的事件对应了不同的内容，因此Demo中将事件-内容进行对齐显示。

- 报表结果查询

  您可以使用`workshop_demo.exe day_active`命令查看日活数据。

```
C:\ workshop_demo>workshop_demo.exe day_active
          metric                  ds                 num
      day_active          2019-05-19             1416104
      day_active          2019-05-20             1416540
      day_active          2019-05-21             1422314
      day_active          2019-05-22             1422411
      day_active          2019-05-23             1428480
      day_active          2019-05-24             1431989
      day_active          2019-05-25             1436218
      day_active          2019-05-26             1437886
      day_active          2019-05-27             1440633
      day_active          2019-05-28             1444736
      day_active          2019-05-29             1450520
      day_active          2019-05-30             1451543
      day_active          2019-05-31             1457510
      day_active          2019-06-01             1458998
      day_active          2019-06-02             1466801
      day_active          2019-06-03             1468898
      day_active          2019-06-04             1473173
      day_active          2019-06-05             1479770
      day_active          2019-06-06             1483101
      day_active          2019-06-07             1484922
      day_active          2019-06-08             1485347
      day_active          2019-06-09             1492034
      day_active          2019-06-10             1499914
      day_active          2019-06-11             1495458
      day_active          2019-06-12             1500697
      day_active          2019-06-13             1508061
      day_active          2019-06-14             1509108
      day_active          2019-06-15             1510583
      day_active          2019-06-16             1518355
      day_active          2019-06-17             1520938
```

2.4 数据建模与开发

2.4.1 新建数据表

为方便在MaxCompute上对数据进行加工处理，首先您需要在MaxCompute上建立数据表，用于承载原始数据及加工后的数据。

前提条件

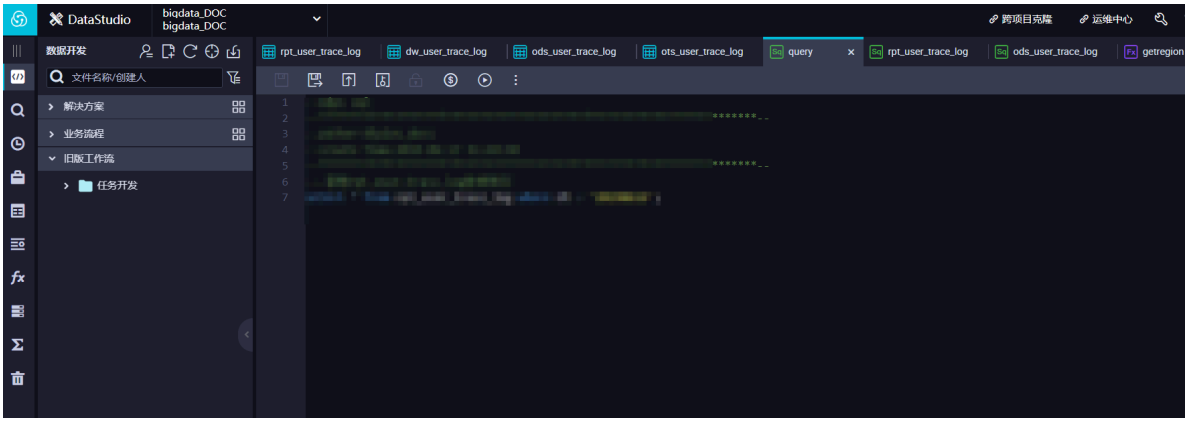
请您参见环境准备章节，完成数据计算服务MaxCompute的开通和DataWorks工作空间的创建。

操作步骤

1. 进入DataWorks工作空间

进入DataWorks工作空间列表，选择区域为华东1，双击您创建好的工作空间（项目）。

双击之后，即可进入工作空间的数据开发界面。



2. 新建数据表

本示例通过DataWorks#unique_48功能新建数据表。为了与表格存储联动，创建的OTS表类型为MaxCompute外部表，作为原始数据提供层。为满足外部表的授权条

件，当MaxCompute和TableStore的Owner是同一个账号时，您可以单击此处一键授权，详情请参见#unique_49。

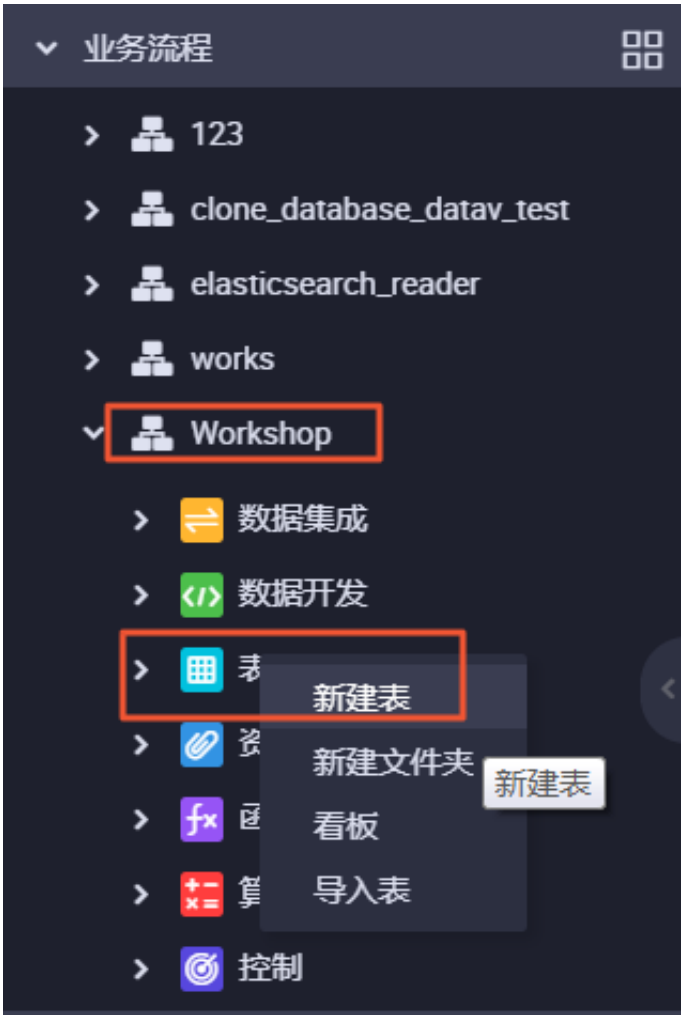
a) 创建业务流程

A. 右键单击业务流程，选择新建业务流程。

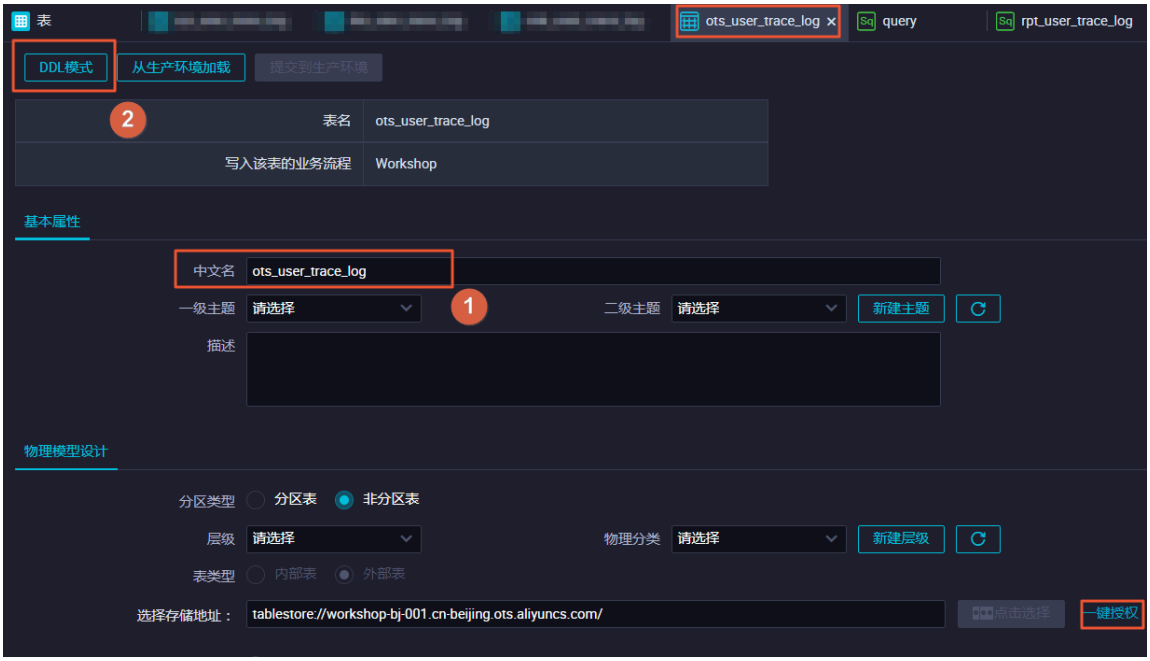
B. 填写业务名称和描述，单击新建。本教程中，业务流程名为Workshop。

b) 创建外部表ots_user_trace_log

双击您新建的业务流程，右键单击表，选择新建表输入您的表名ots_user_trace_log。



填写您的表中文名称，如果您之前未进行一键授权，此时也可以继续完成授权。然后单击DDL模式，开始编辑建表语句。



本例使用的建表语句如下，请您参考环境准备章节，根据自己的表格存储实例访问地址参数填写LOCATION地址。完成填写后单击提交到生产环境。

```
CREATE EXTERNAL TABLE `ots_user_trace_log` (
    `md5` string COMMENT '用户uid的md5值前8位',
    `uid` string COMMENT '用户uid',
    `ts` bigint COMMENT '用户操作时间戳',
    `ip` string COMMENT 'ip地址',
    `status` bigint COMMENT '服务器返回状态码',
    `bytes` bigint COMMENT '返回给客户端的字节数',
    `device` string COMMENT '终端型号',
    `system` string COMMENT '系统版本ios xxx/android xxx',
    `customize_event` string COMMENT '自定义事件：登录/退出/购买/注册/点击/后台/切换用户/浏览',
    `use_time` bigint COMMENT 'APP单次使用时长，当事件为退出、后台、切换用户时有该项',
    `customize_event_content` string COMMENT '用户关注内容信息，在customize_event为浏览和评论时 包含该列'
)
STORED BY 'com.aliyun.odps.TableStoreStorageHandler'
WITH SERDEPROPERTIES (
    'tablestore.columns.mapping'=':md5,:uid,:ts, ip,status,bytes,device,system,customize_event,use_time,customize_event_content',
    'tablestore.table.name'='user_trace_log'
)
LOCATION 'tablestore://workshop-bj-001.cn-beijing.ots.aliyuncs.com/';
```

说明：

如果您使用LOCATION 'tablestore://workshop-bj-001.cn-beijing.ots.aliyuncs.com/'报错，显示网络不同，可尝试更换为LOCATION 'tablestore://workshop-bj-001.cn-beijing.ots-interna.aliyuncs.com/'。

c) 创建ods_user_trace_log表

建表方法同上，建表语句如下，完成填写后单击提交到生产环境。ods_user_trace_log为ODS层表，相关数仓模型定义请参见#unique_50。

```
CREATE TABLE IF NOT EXISTS ods_user_trace_log (
    md5 STRING COMMENT '用户uid的md5值前8位',
    uid STRING COMMENT '用户uid',
    ts BIGINT COMMENT '用户操作时间戳',
    ip STRING COMMENT 'ip地址',
    status BIGINT COMMENT '服务器返回状态码',
    bytes BIGINT COMMENT '返回给客户端的字节数',
    device STRING COMMENT '终端型号',
    system STRING COMMENT '系统版本ios xxx/android xxx',
    customize_event STRING COMMENT '自定义事件：登录/退出/购买/注册/点击/后台/切换用户/浏览',
    use_time BIGINT COMMENT 'APP单次使用时长，当事件为退出、后台、切换用户时有该项',
    customize_event_content STRING COMMENT '用户关注内容信息，在customize_event为浏览和评论时 包含该列'
)
PARTITIONED BY (
    dt STRING
```

```
);
```

d) 创建dw_user_trace_log表

建表方法同上，建表语句如下，完成填写后单击提交到生产环境。dw_user_trace_log为DW层表，相关数仓模型定义请参见#unique_51。

```
CREATE TABLE IF NOT EXISTS dw_user_trace_log (
    uid STRING COMMENT '用户uid',
    region STRING COMMENT '地域，根据ip得到',
    device_brand string comment '设备品牌',
    device STRING COMMENT '终端型号',
    system_type STRING COMMENT '系统类型，Android、IOS、ipad、
Windows_phone',
    customize_event STRING COMMENT '自定义事件：登录/退出/购买/注册/点
击/后台/切换用户/浏览',
    use_time BIGINT COMMENT 'APP单次使用时长，当事件为退出、后台、切换用
户时有该项',
    customize_event_content STRING COMMENT '用户关注内容信息，在
customize_event为浏览和评论时 包含该列'
)
PARTITIONED BY (
    dt STRING
);
```

e) 创建rpt_user_trace_log表

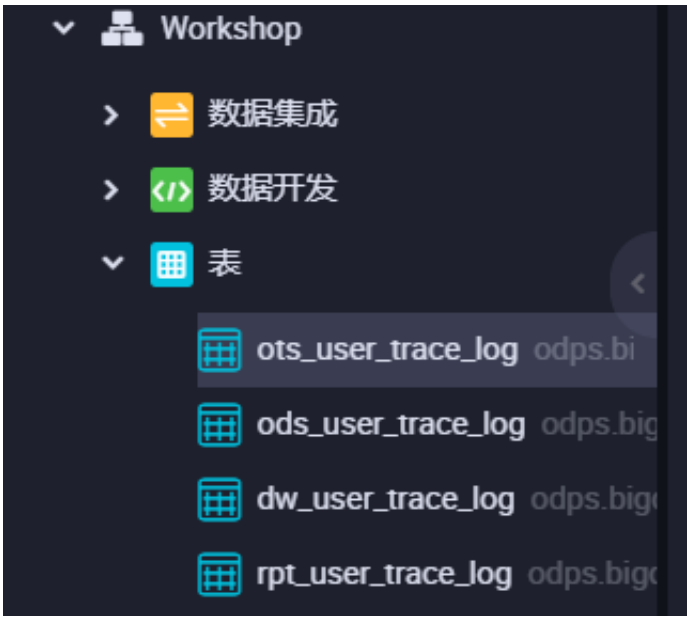
建表方法同上，建表语句如下，完成填写后单击提交到生产环境。rpt_user_trace_log为ADS层表，相关数仓模型定义请参见#unique_52。

```
CREATE TABLE IF NOT EXISTS rpt_user_trace_log (
    country STRING COMMENT '国家',
    province STRING COMMENT '省份',
    city STRING COMMENT '城市',
    device_brand string comment '设备品牌',
    device STRING COMMENT '终端型号',
    system_type STRING COMMENT '系统类型，Android、IOS、ipad、
Windows_phone',
    customize_event STRING COMMENT '自定义事件：登录/退出/购买/注册/点
击/后台/切换用户/浏览',
    use_time BIGINT COMMENT 'APP单次使用时长，当事件为退出、后台、切换用
户时有该项',
    customize_event_content STRING COMMENT '用户关注内容信息，在
customize_event为浏览和评论时 包含该列',
    pv bigint comment '浏览量',
    uv bigint comment '独立访客'
)
PARTITIONED BY (
    dt STRING
```

```
);
```

3. 验证建表结果

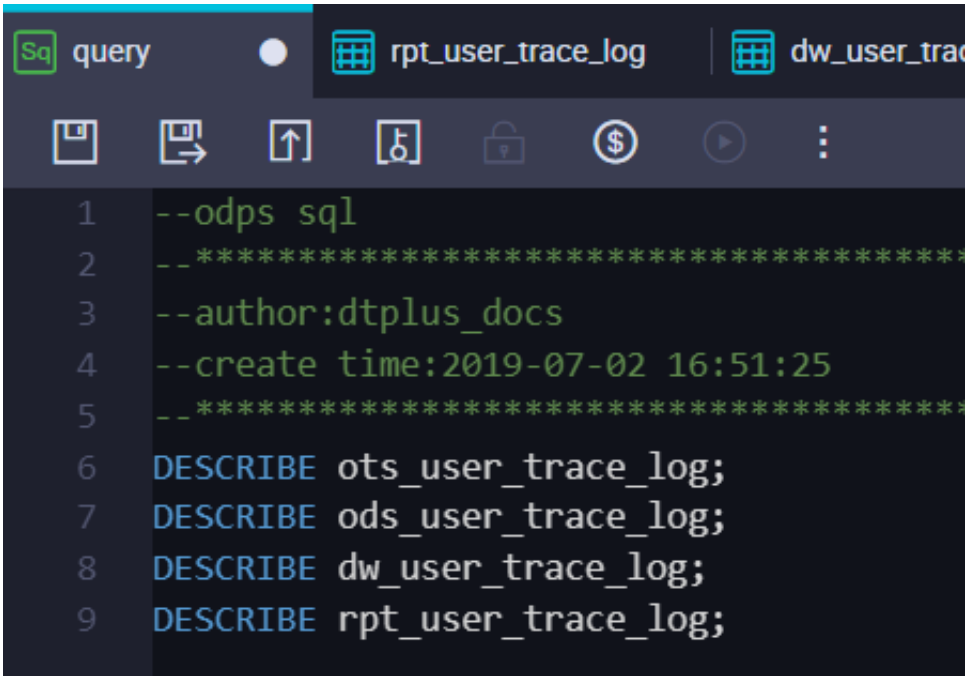
完成建表后，您可以在自己的工作量下看到新建的4张表。



使用数据开发 > 新建数据开发节点 > ODPS SQL，在新建的ODPS SQL节点中写入下列表查询SQL语句。

```
DESCRIBE ots_user_trace_log;
DESCRIBE ods_user_trace_log;
DESCRIBE dw_user_trace_log;
DESCRIBE rpt_user_trace_log;
```

单击运行，查询建表结果。

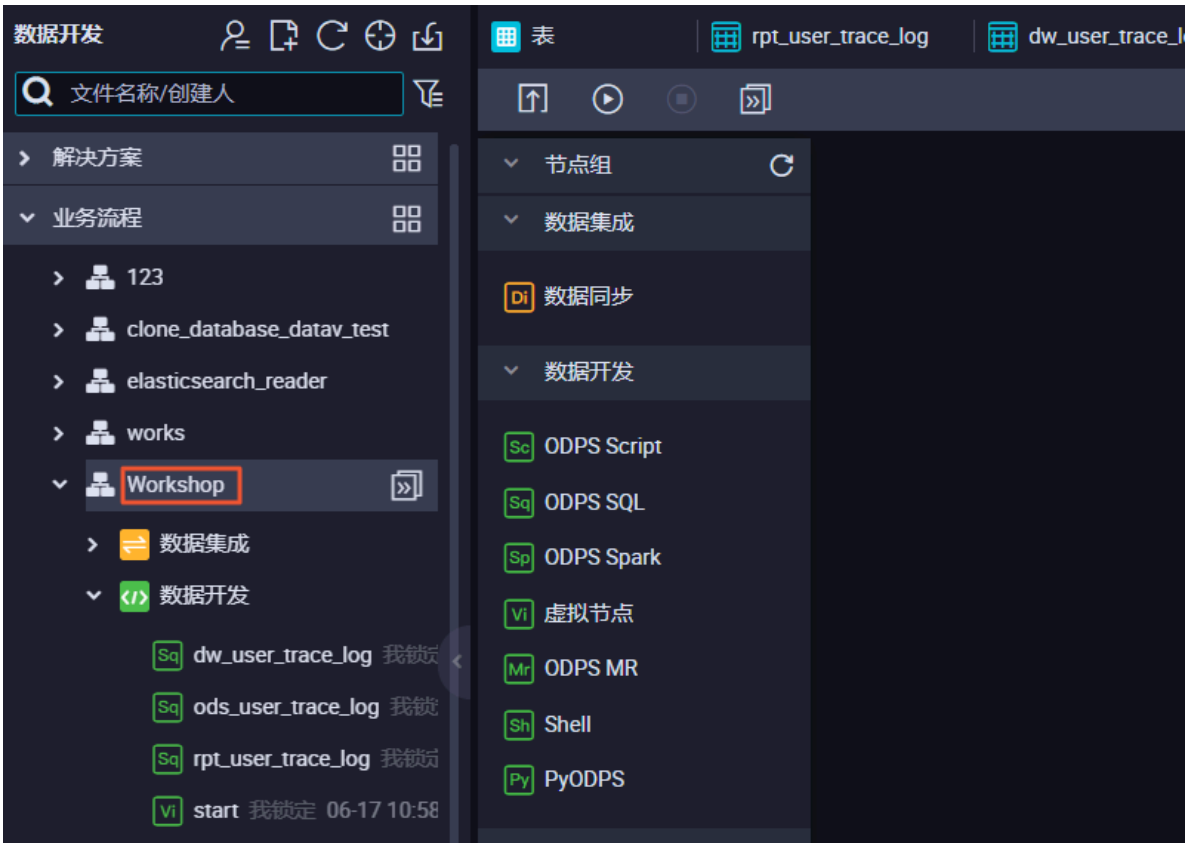


2.4.2 设计工作流

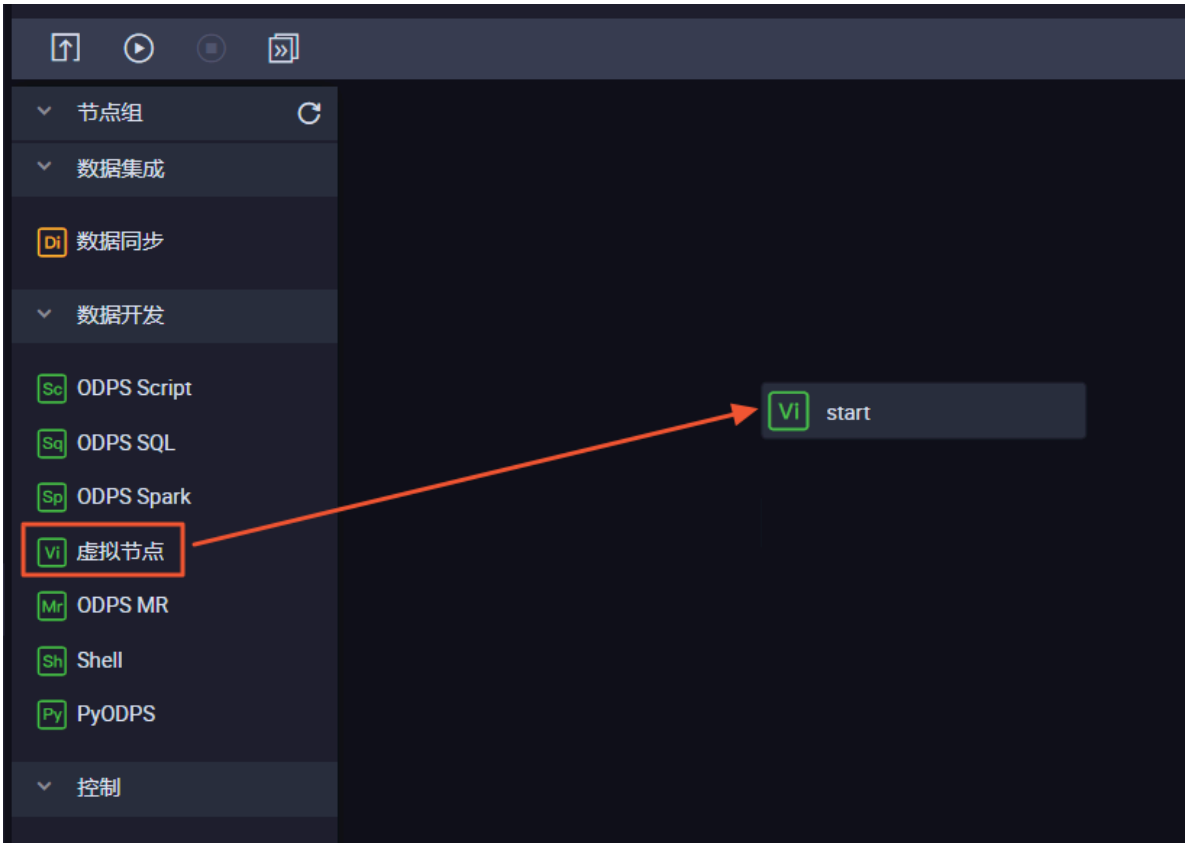
通过设计工作流，您可以明确在整体数据开发过程中各任务节点的排布。对于本教程中这种较为简单的单数据流场景，您可以选择每个数据表（数仓层次）对应一个工作流。

操作步骤

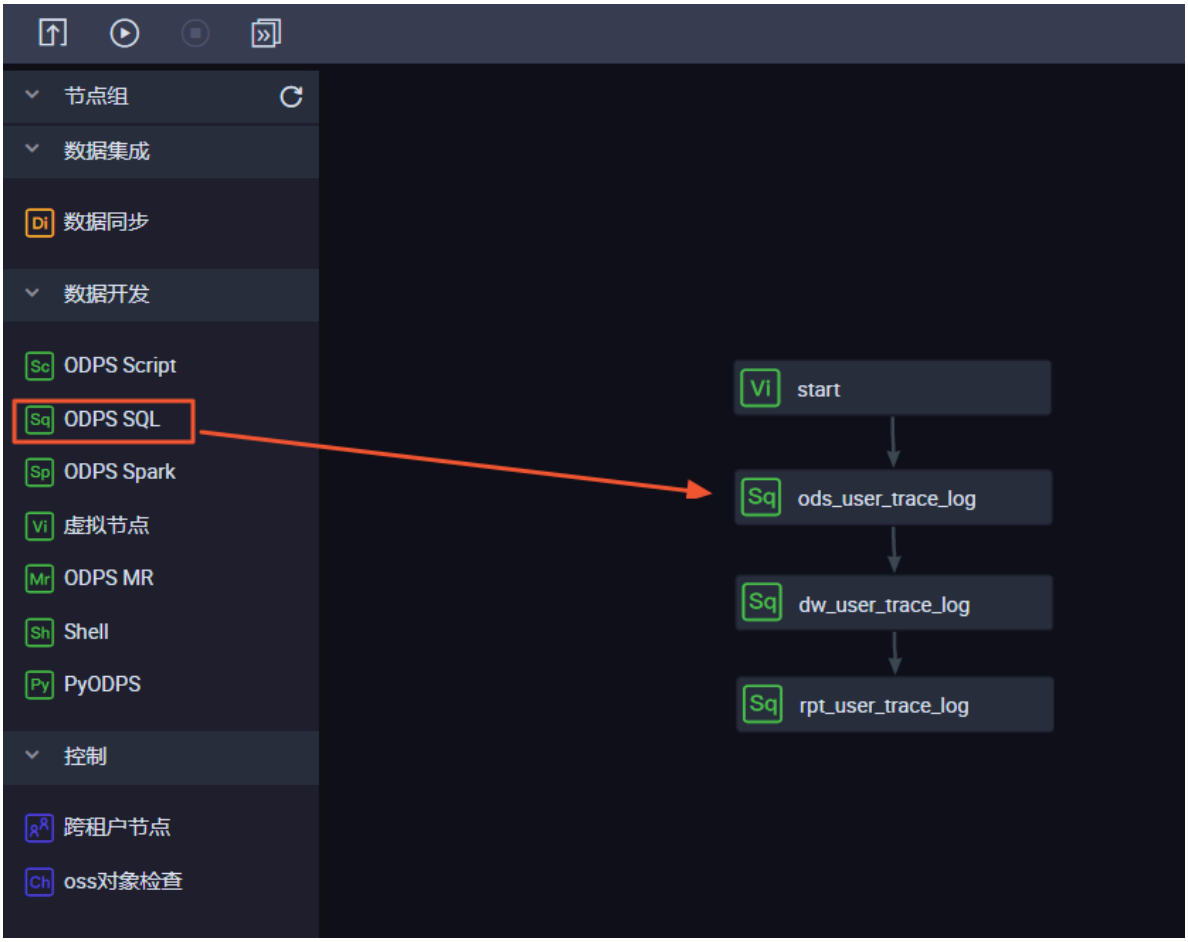
1. 双击您的业务流程，打开画布面板。



2. 向画布中拖入1个虚拟节点。



3. 向画布中拖入3个ODPS SQL节点，依次命名为ods_user_trace_log、dw_user_trace_log、rpt_user_trace_log。通过连接不同节点，配置依赖关系如下。



说明：
ods_user_trace_log、dw_user_trace_log、rpt_user_trace_log分别代表数据仓库的ODS、CDM和ADS层，详情请参见#unique_52。

2.4.3 节点配置

完成工作流设计后，您需要对每个数据开发节点进行配置，填写SQL处理语句。

前提条件

由于本次数据开发过程中需要使用UDF自定义函数，您首先需要完成自定义函数的注册。

操作步骤

1. 添加资源并创建自定义函数

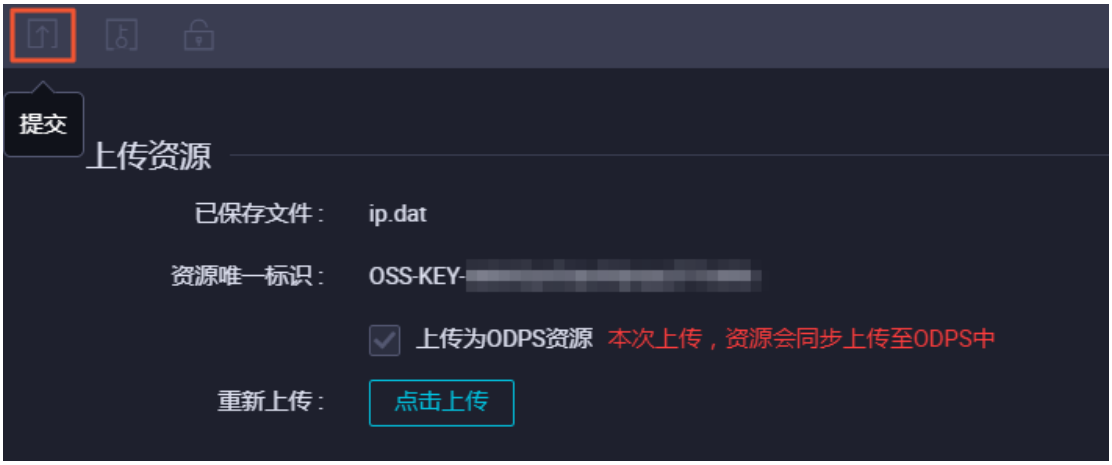
a) 单击此处，下载用于IP地转换的自定义函数Java包getaddr.jar以及地址库ip.dat。

本教程不关注IP地址转换的自定义函数内容。如果您有兴趣了解，请参见MaxCompute中实现IP地址归属地转换。

b) 右键单击您的业务流程下的资源，单击新建资源。

- File类型对应地址库ip.dat。您需要勾选大文件（内容超过500KB）及上传为ODPS资源，然后点击上传。

上传完成后，请务必记得单击提交。

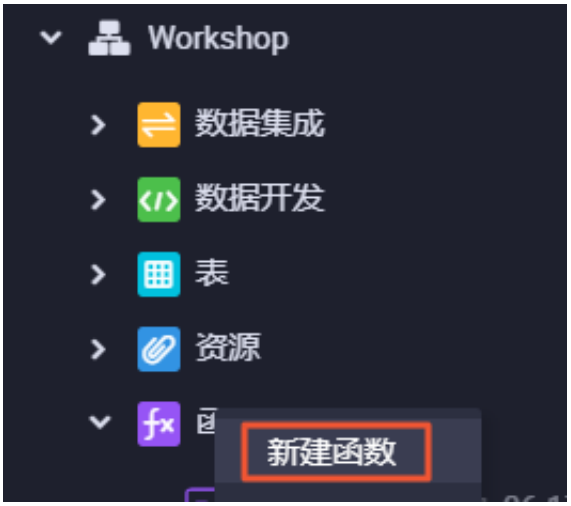


- JAR类型对应Java包getaddr.jar。您需要勾选上传为ODPS资源，然后点击上传。

上传完成后，请务必记得单击提交。

c) 注册函数。

在您的业务流程下右键单击函数，选择新建函数。



请依次填写函数名为getregion，类名为odps.test.GetAddr，资源列表为getaddr.jar,ip.dat，命令格式为getregion(ip string)。填写完成后，单击提交。

2. 配置ODPS SQL节点

a) 双击ods_user_trace_log节点，进入节点配置界面，编写处理逻辑。

SQL代码如下。

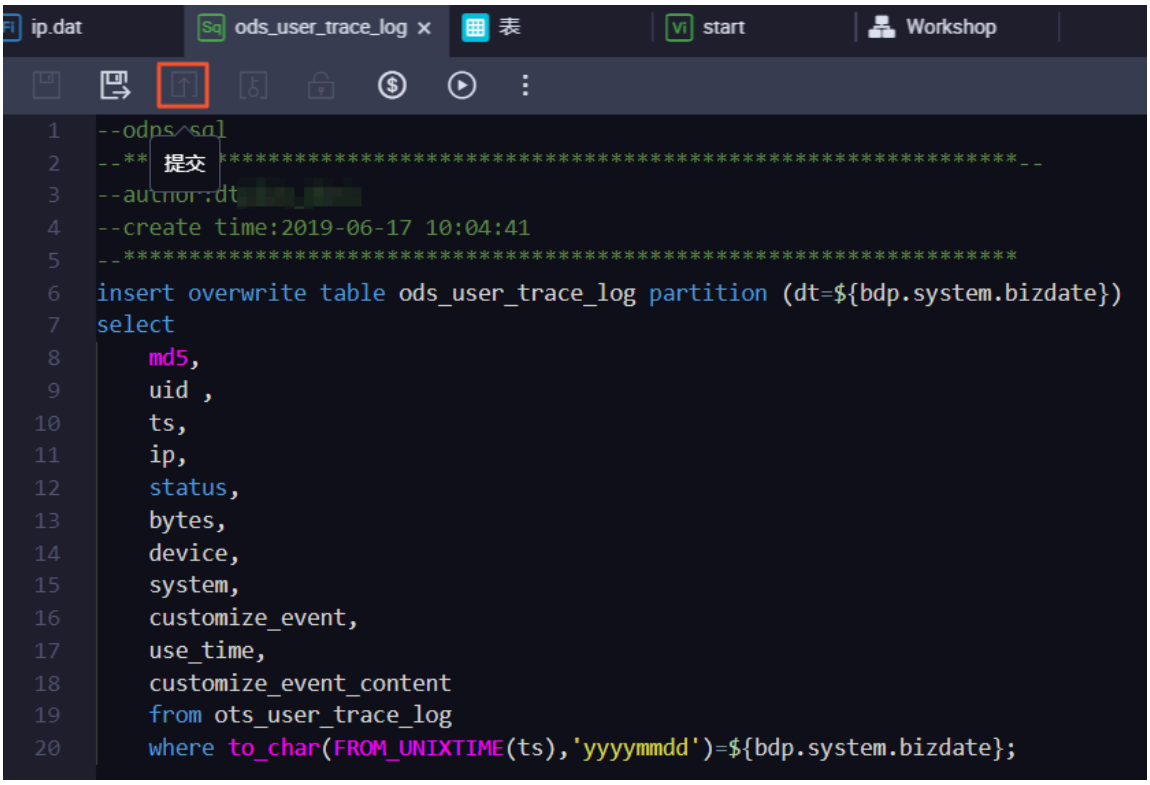
```
insert overwrite table ods_user_trace_log partition (dt=${bdp.
system.bizdate})
select
    md5,
    uid ,
    ts,
    ip,
    status,
    bytes,
    device,
    system,
    customize_event,
```

```
    use_time,
    customize_event_content
    from ots_user_trace_log
    where to_char(FROM_UNIXTIME(ts),'yyyymmdd')=${bdp.system.
bizdate};
```

说明:

关于${bdp.system.bizdate}释义请参见#unique_55。

b) 完成代码编写后，单击提交。



3. 配置dw_user_trace_log节点

您可以使用与ods_user_trace_log节点一样的方法配置dw_user_trace_log节点，SQL代码如下。

```
INSERT OVERWRITE TABLE dw_user_trace_log PARTITION (dt=${bdp.system.
bizdate})
SELECT uid, getregion(ip) AS region
    , CASE
        WHEN TOLOWER(device) RLIKE 'xiaomi' THEN 'xiaomi'
        WHEN TOLOWER(device) RLIKE 'meizu' THEN 'meizu'
        WHEN TOLOWER(device) RLIKE 'huawei' THEN 'huawei'
        WHEN TOLOWER(device) RLIKE 'iphone' THEN 'iphone'
        WHEN TOLOWER(device) RLIKE 'vivo' THEN 'vivo'
        WHEN TOLOWER(device) RLIKE 'honor' THEN 'honor'
        WHEN TOLOWER(device) RLIKE 'samsung' THEN 'samsung'
        WHEN TOLOWER(device) RLIKE 'leeco' THEN 'leeco'
        WHEN TOLOWER(device) RLIKE 'ipad' THEN 'ipad'
        ELSE 'unknown'
    END AS device_brand, device
    , CASE
        WHEN TOLOWER(system) RLIKE 'android' THEN 'android'
```

```
        WHEN TOLOWER(system) RLIKE 'ios' THEN 'ios'
        ELSE 'unknown'
    END AS system_type, customize_event, use_time, customize_
event_content
FROM ods_user_trace_log
WHERE dt = ${bdp.system.bizdate};
```

4. **配置rpt_user_trace_log节点**

   您可以使用与ods_user_trace_log节点一样的方法配置rpt_user_trace_log节点，SQL代码如下。

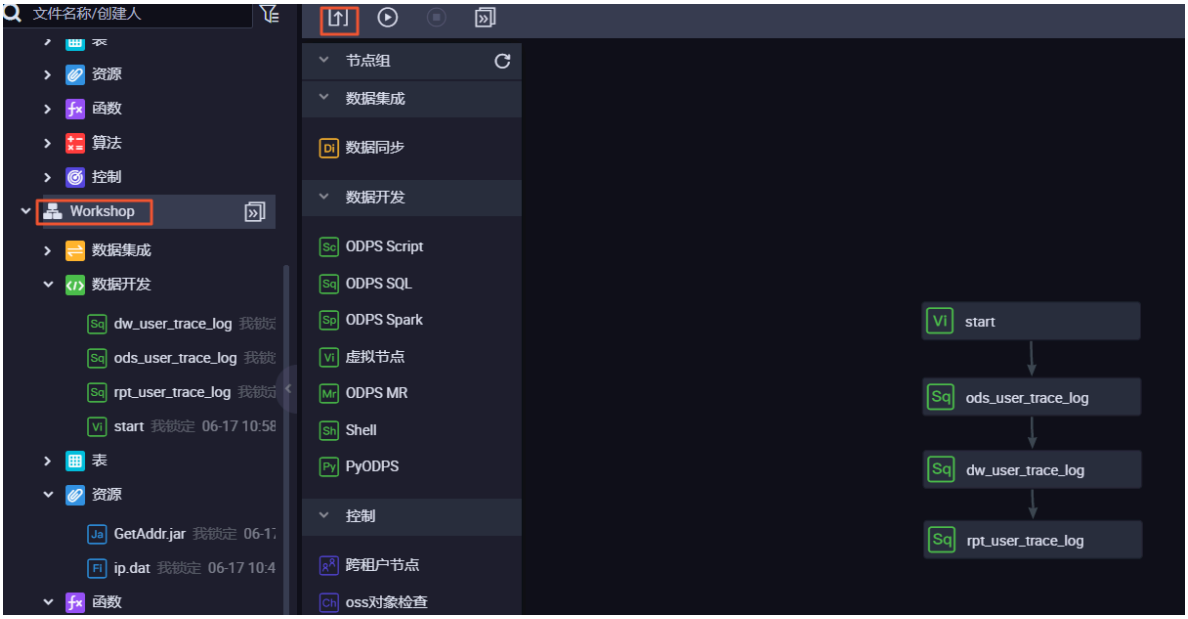
```
INSERT OVERWRITE TABLE rpt_user_trace_log PARTITION (dt=${bdp.system
.bizdate})
SELECT split_part(split_part(region, ',', 1),'[',2) AS country
    , trim(split_part(region, ',', 2)) AS province
    , trim(split_part(region, ',', 3)) AS city
    , MAX(device_brand), MAX(device)
    , MAX(system_type), MAX(customize_event)
    , FLOOR(AVG(use_time / 60))
    , MAX(customize_event_content), COUNT(uid) AS pv
    , COUNT(DISTINCT uid) AS uv
FROM dw_user_trace_log
WHERE dt = ${bdp.system.bizdate}
GROUP BY uid,
    region;
```

2.4.4 任务提交与测试

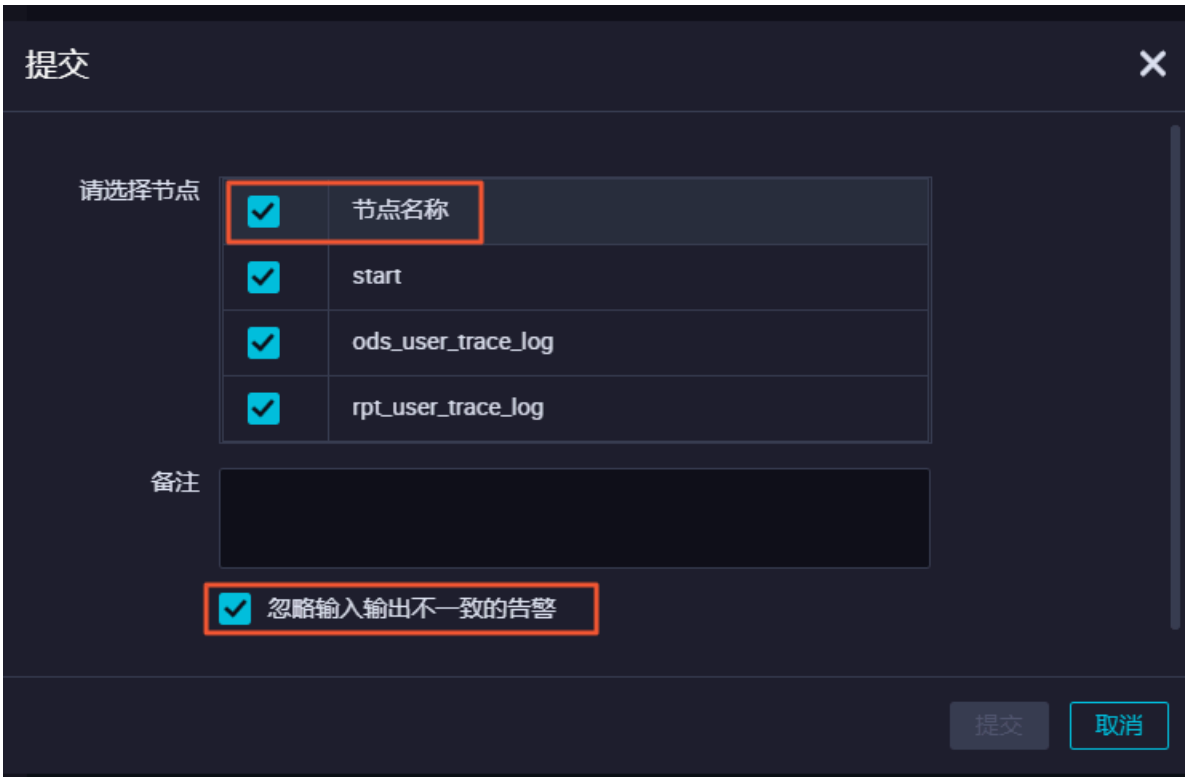
在您完成节点配置后，还需要提交任务到运维中心，才能对任务进行测试。

操作步骤

1. 双击您的业务流程名称，单击提交。

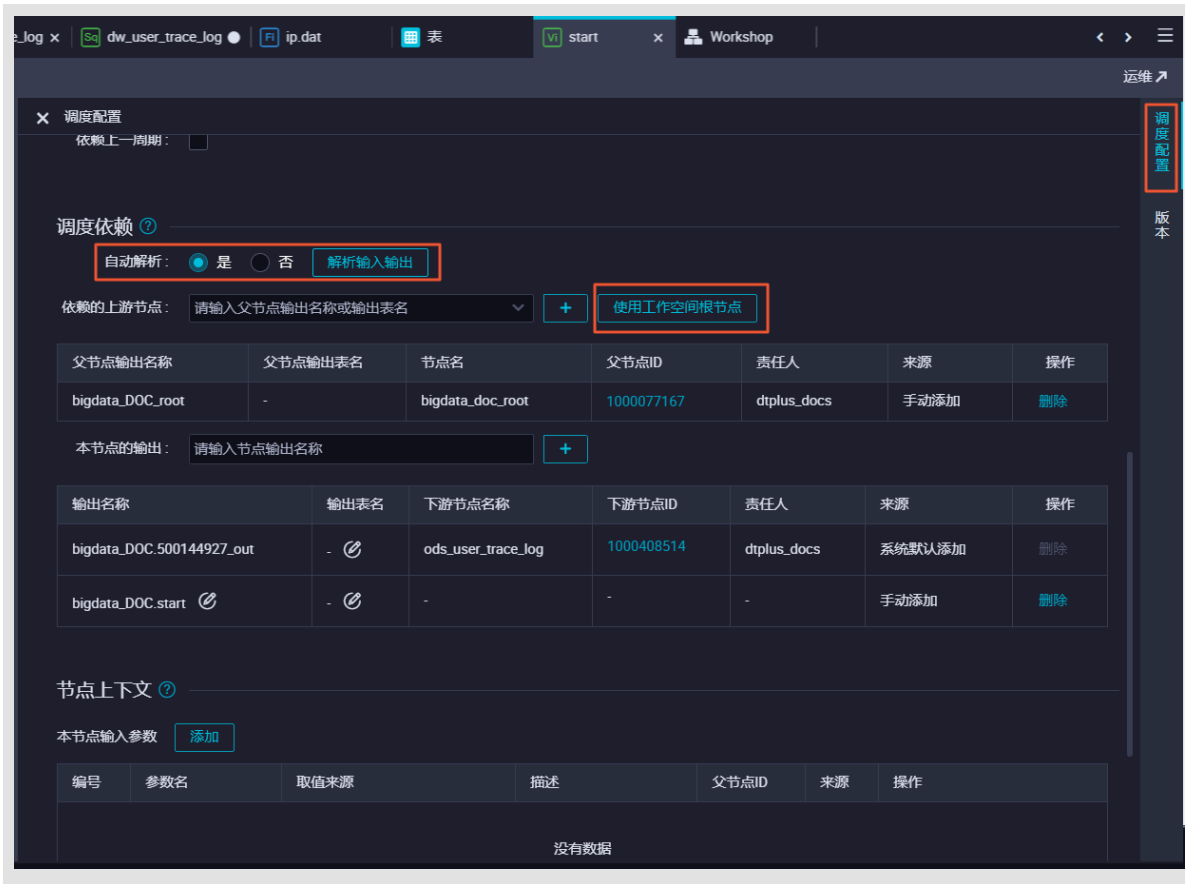


勾选所有可提交节点及忽略输入输出不一致的告警。如果您的节点在配置完成后已经提交完毕且无更新，此处您会发现没有可以提交的节点，直接跳过本步骤。



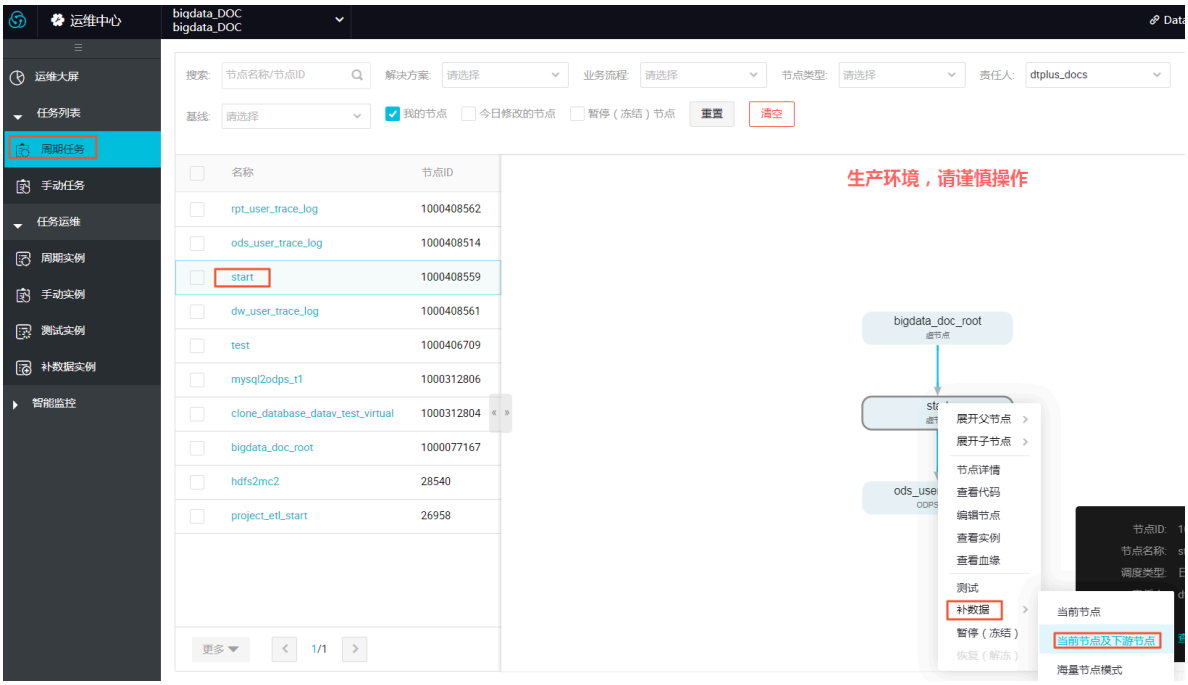
说明:

如果您在提交过程中报错，很可能是节点调度配置对输入输出的自动解析有误。您可以重新编辑节点的调度配置页面，选择自动解析为否后，手动删除错误的输入。对于VI虚拟节点，建议您勾选使用工作空间根节点。

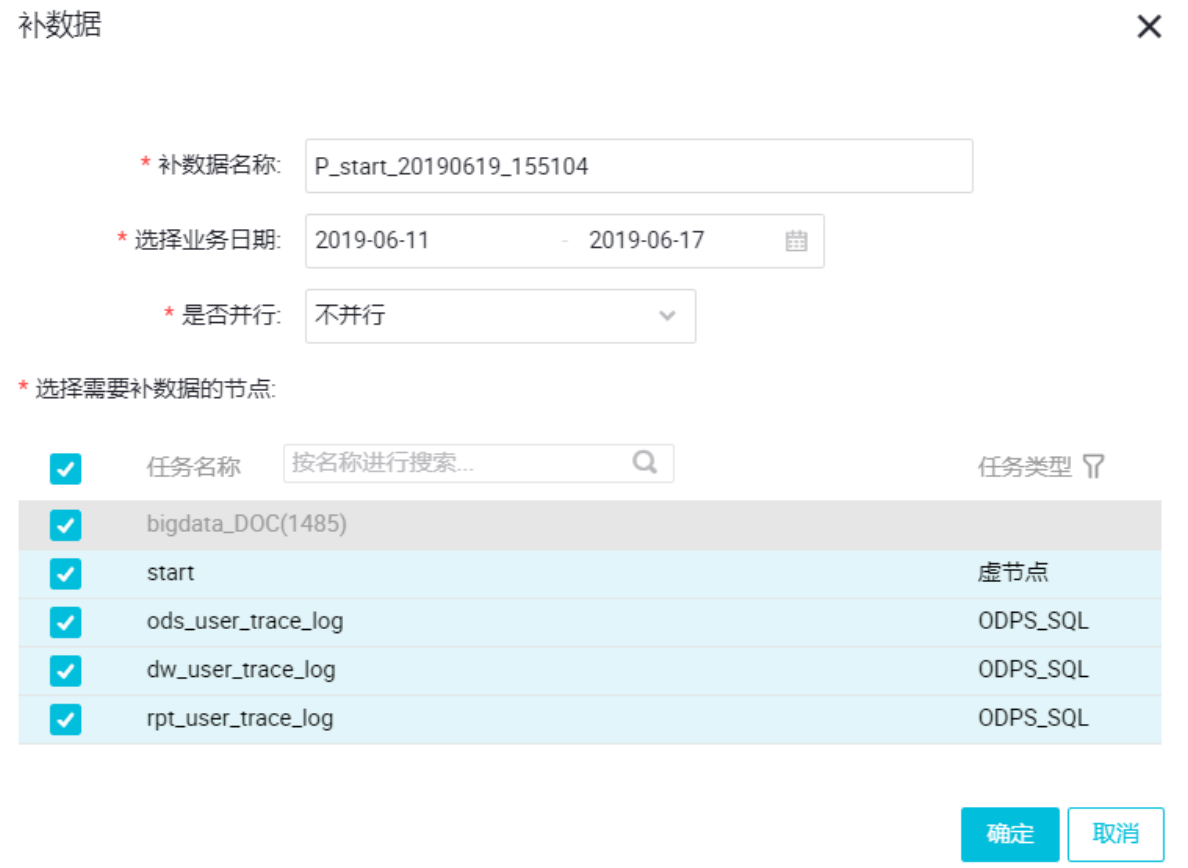


2. 单击右上角的运维中心进行界面切换。

3. 双击任务列表 > 周期任务中您的虚拟节点后，在右侧界面右键单击虚拟节点（本例中名为start）。在弹框中单击补数据 > 当前节点及下游节点。



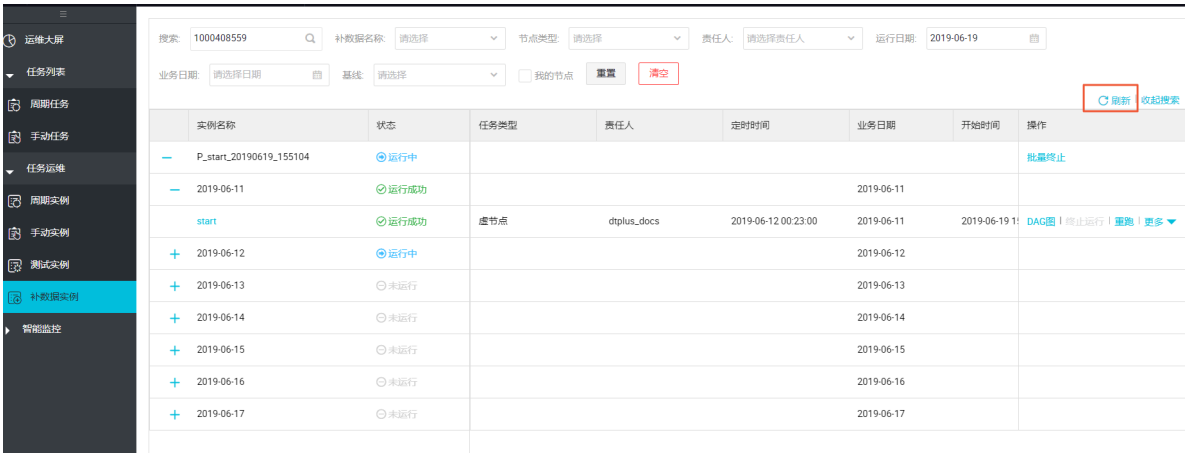
在弹框中勾选所有节点，选择业务日期为最近一周，单击确定。



说明:

关于补数据实例的详情请参见#unique_56。

4. 在补数据实例中，您可以查看补数据实例的运行情况，并通过单击刷新查看实时状态。

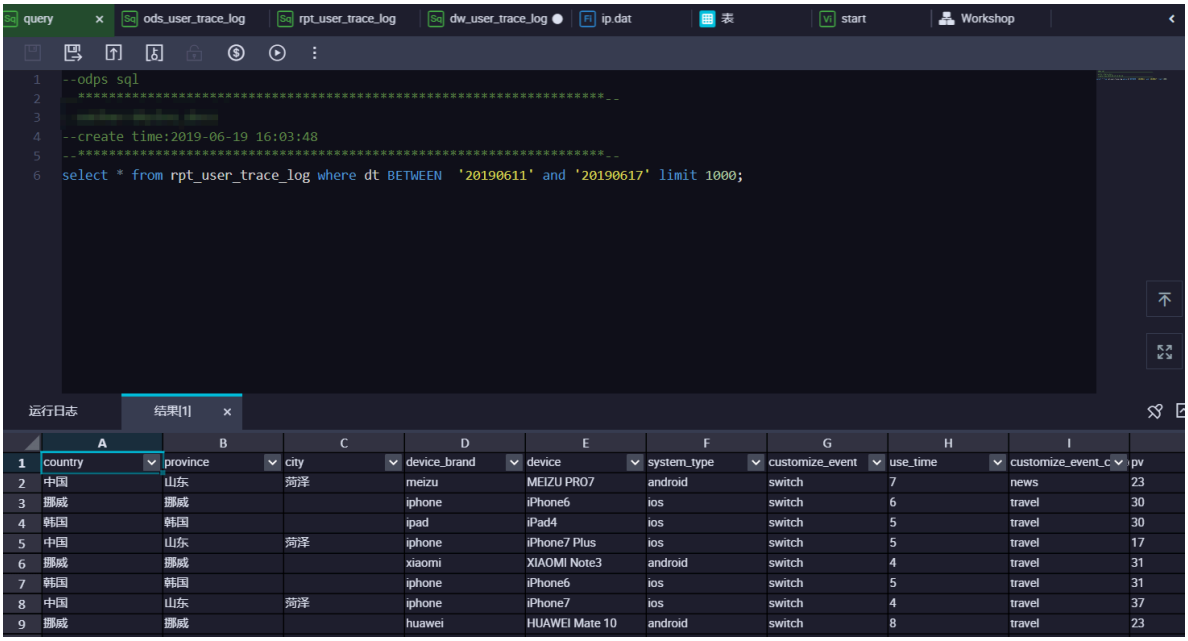


如果运行状态异常，您可以右键单击出错节点，单击查看运行日志进行排查。


生产环境，请谨慎操作
展开父节点
展开子节点
查看运行日志
查看代码
编辑节点
查看血缘
终止运行
重跑
重跑下游
置成功
暂停（冻结）

5. 待补数据实例运行完成后，您可以使用数据开发 > 新建数据开发节点 > ODPS SQL，在新建的ODPS SQL节点中写入下列SQL语句来确认数据是否成功写入rpt_user_trace_log表。

   SQL语句：`select * from rpt_user_trace_log where dt BETWEEN '20190611' and '20190617' limit 1000;`



	A	B	C	D	E	F	G	H	I	
1	country	province	city	device_brand	device	system_type	customize_event	use_time	customize_event_c	pv
2	中国	山东	菏泽	meizu	MEIZU PRO7	android	switch	7	news	23
3	挪威	挪威		iphone	iPhone6	ios	switch	6	travel	30
4	韩国	韩国		ipad	iPad4	ios	switch	5	travel	30
5	中国	山东	菏泽	iphone	iPhone7 Plus	ios	switch	5	travel	17
6	挪威	挪威		xiaomi	XIAOMI Note3	android	switch	4	travel	31
7	韩国	韩国		iphone	iPhone6	ios	switch	5	travel	31
8	中国	山东	菏泽	iphone	iPhone7	ios	switch	4	travel	37
9	挪威	挪威		huawei	HUAWEI Mate 10	android	switch	8	travel	23

2.5 数据可视化展现

数据表rpt_user_trace_log加工完成后，您可以通过Quick BI创建网站用户分析画像的仪表板，实现该数据表的可视化。

前提条件

在开始实验前，请确认您已经完成了环境准备和数据建模与开发的全部步骤。单击进入Quick BI控制台。

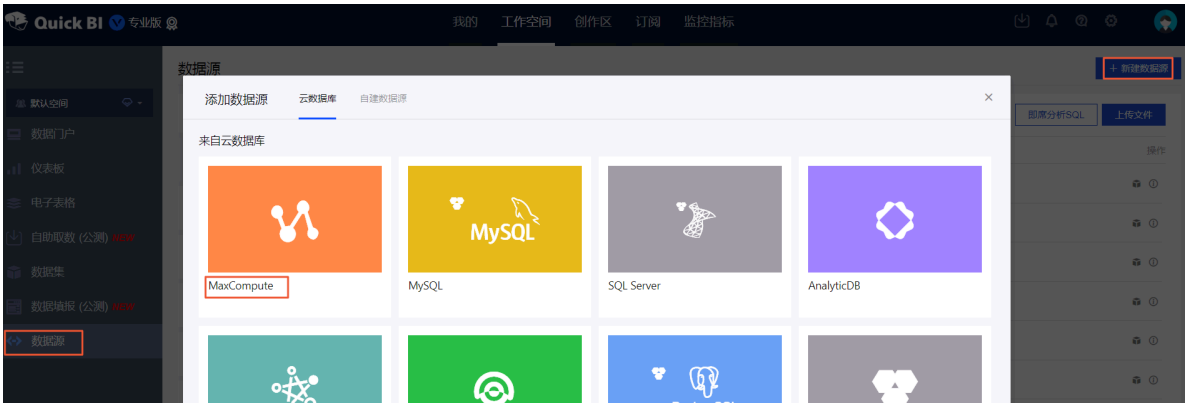
背景信息

rpt_user_trace_log表包含了country、province、city、device_brand、use_time、pv等字段信息。您可以通过仪表板展示用户的核心指标、周期变化、用户地区分布和记录。

操作步骤

1. 单击进入默认空间，您也可以使用自己的个人空间。

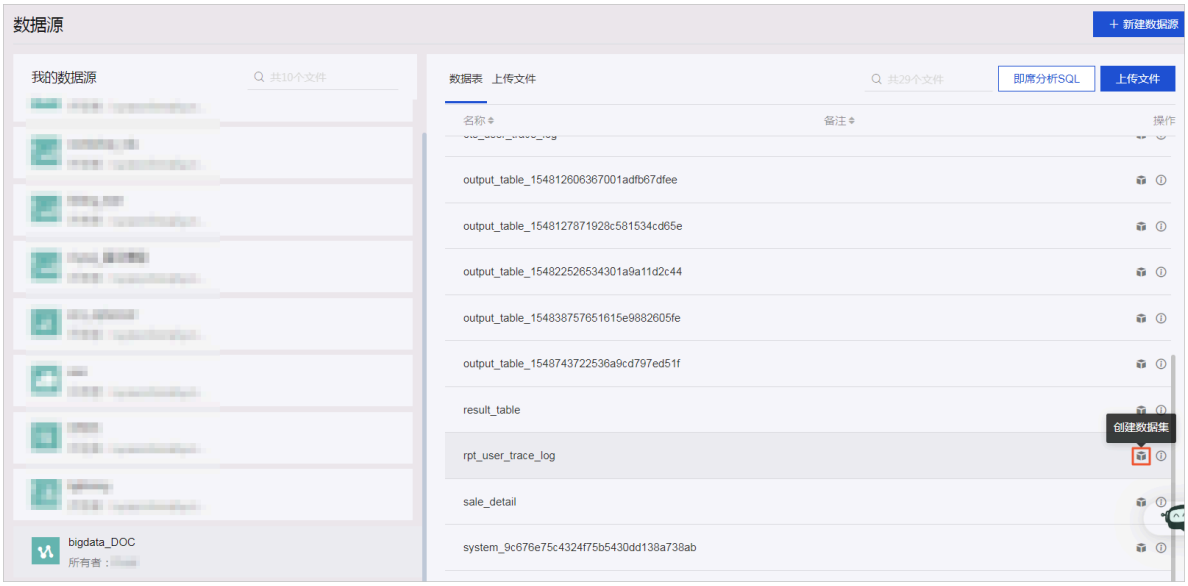
2. 选择数据源 > 新建数据源 > 云数据库 > MaxCompute。



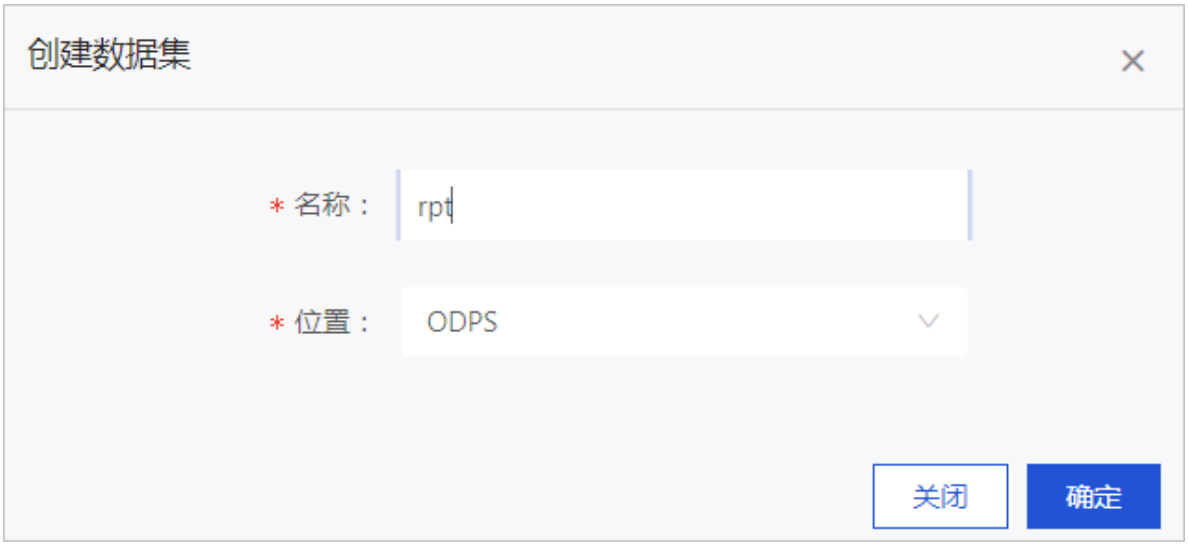
3. 输入您的MaxCompute项目名称以及您的AcessKey信息，数据库地址使用默认地址即可，关于数据库地址详情请参见#unique_58。

   完成填写后，单击连接测试，待显示数据源连通性正常后单击添加即可。

4. 找到您刚添加的数据源的rpt_user_trace_log表，单击创建数据集。



选择您想放置的数据集位置，单击确定。



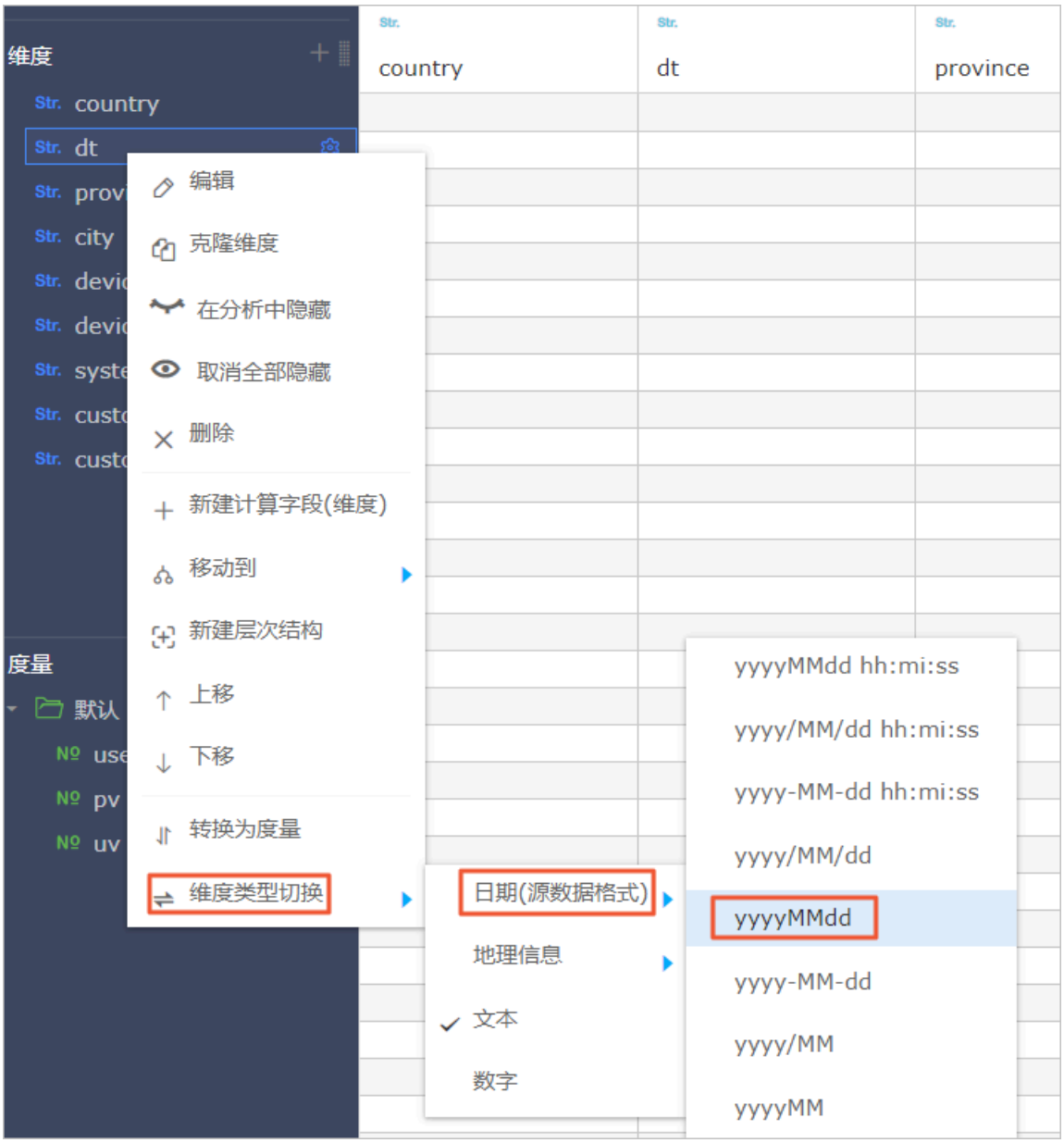
5. 进入数据集列表页，单击您刚刚创建的数据集，对数据集进行编辑。

常见的数据集加工包括：维度、度量的切换、修改维度的类型、增加计算字段、创建层次结构、修改字段的数据类型、更改度量聚合方式、制作关联模型。

6. 转换字段的维度类型。

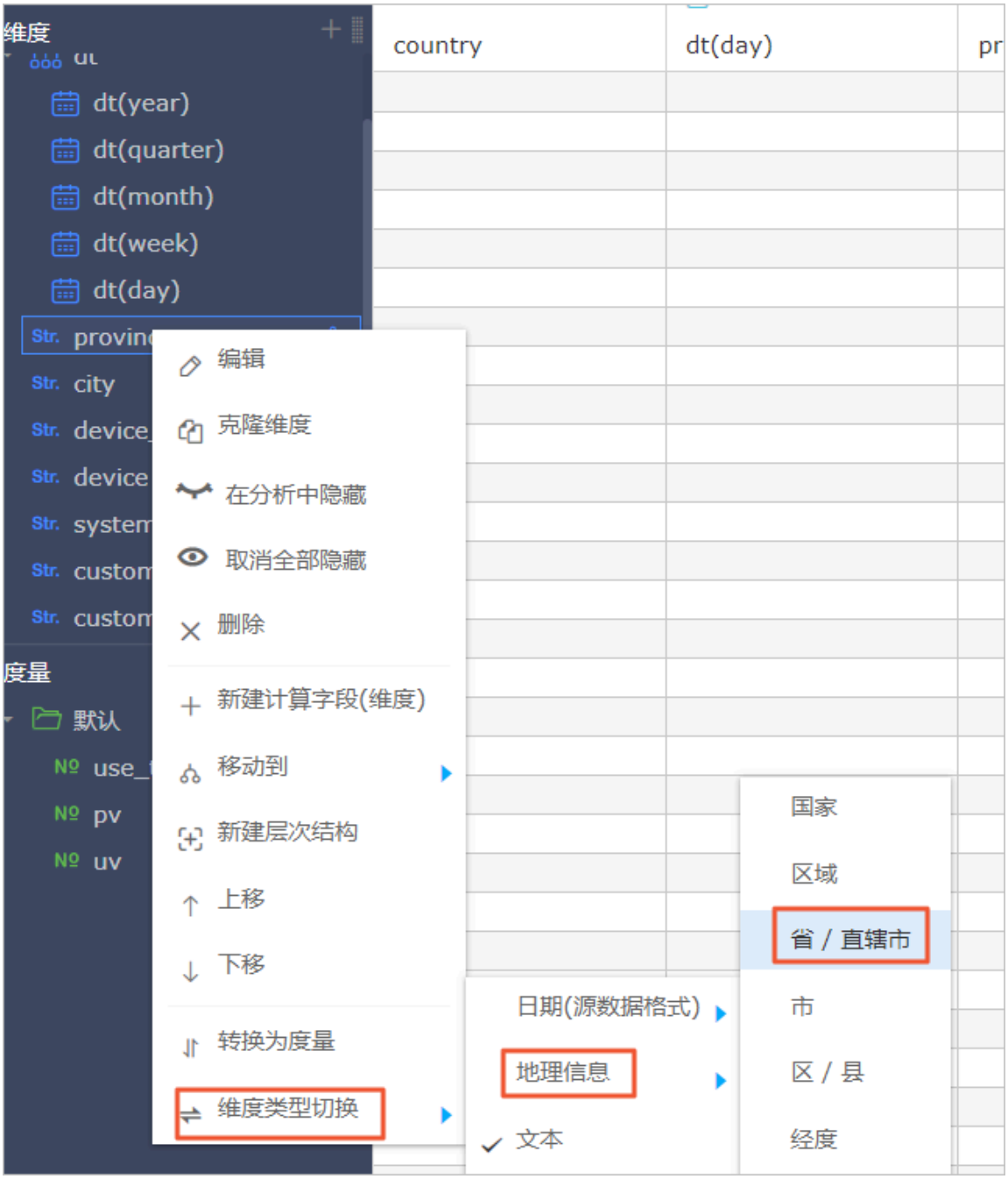
a) 转换日期字段的维度类型。

右键单击dt字段，选择维度类型切换 > 日期（源数据格式） > yyyyMMdd。

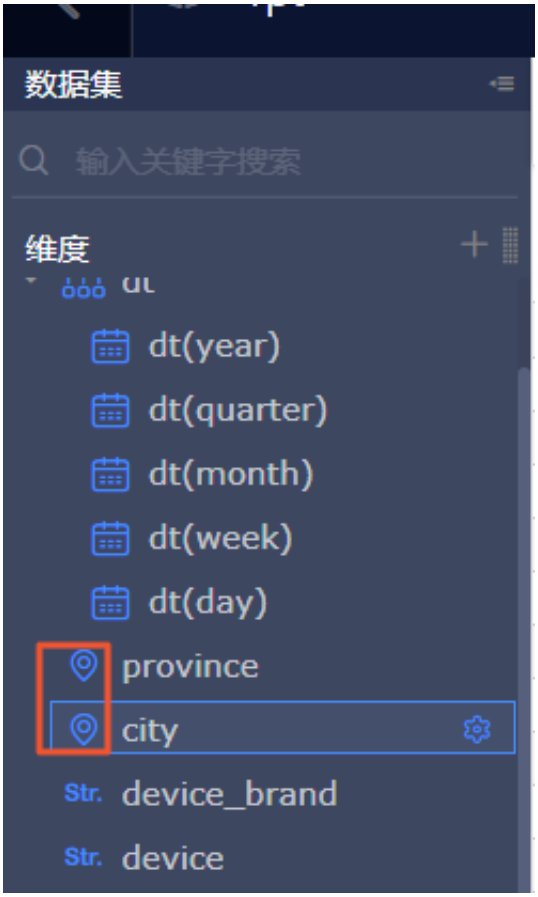


b) 转换地理信息字段的维度类型。

右键单击province字段，选择维度类型切换 > 地理信息 > 省/直辖市。

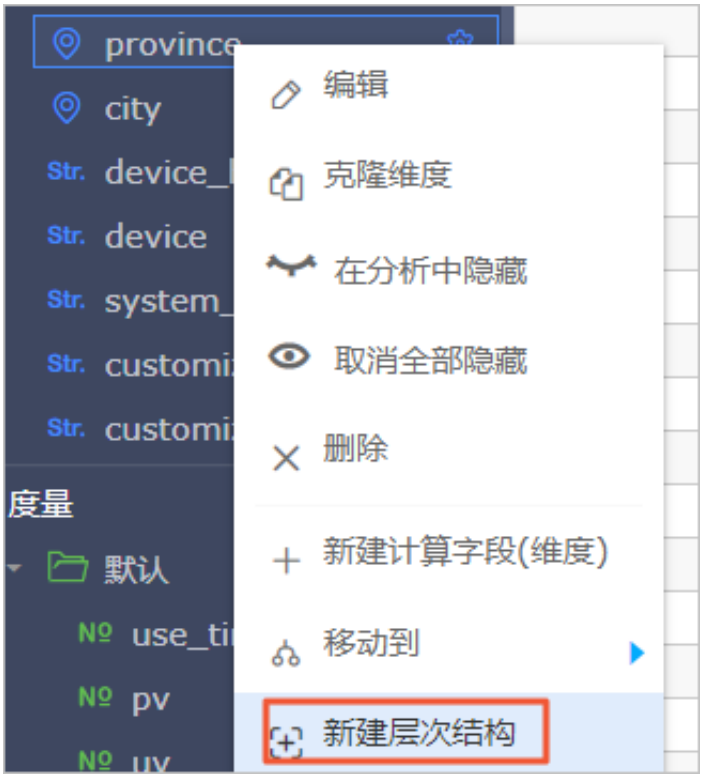


右键单击city字段，选择维度类型切换 > 地理信息 > 市。转换成功后，在左侧维度栏中会看到字段前多一个地理位置图标。



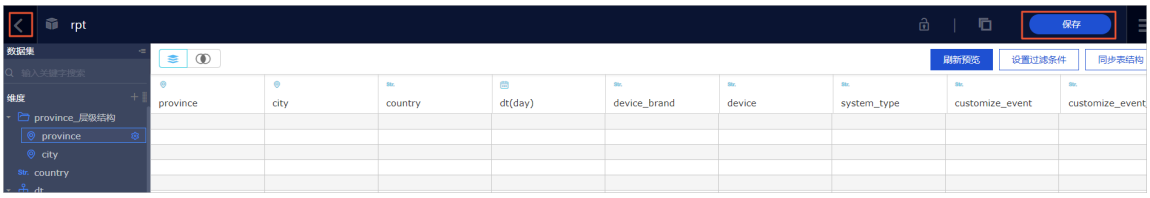
c) **新建层次结构。**

右键province，单击新建层次结构，在弹框中单击确定。



然后，把city字段移到province层次结构的树下。

完成上述操作后，单击保存，返回数据集列表。



7. 制作仪表板。

随着数据的更新，让报表可视化地展现最新数据，这个过程叫制作仪表板。仪表板的制作流程如下：

a. 确定内容。

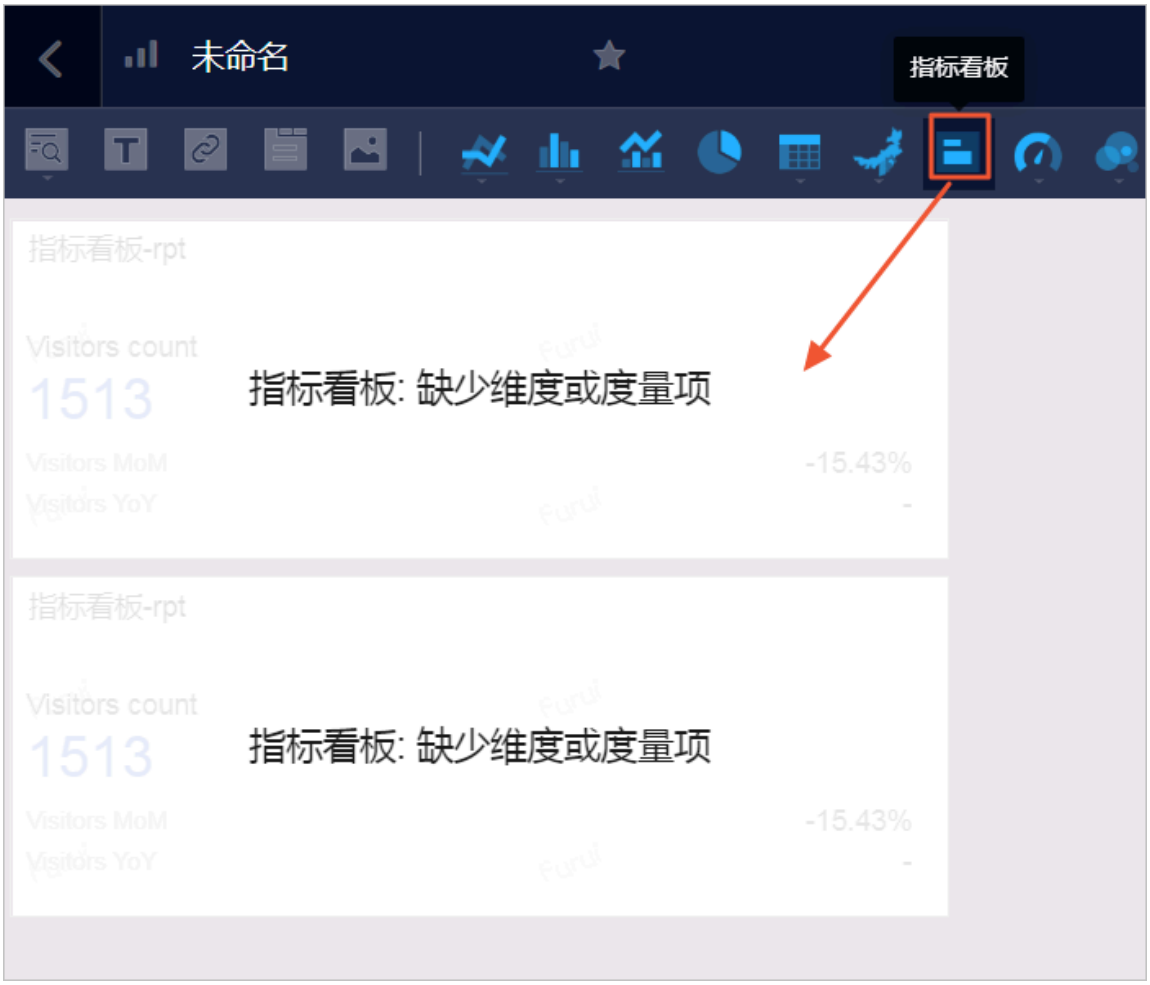
b. 确定布局和样式。

c. 制作图表。

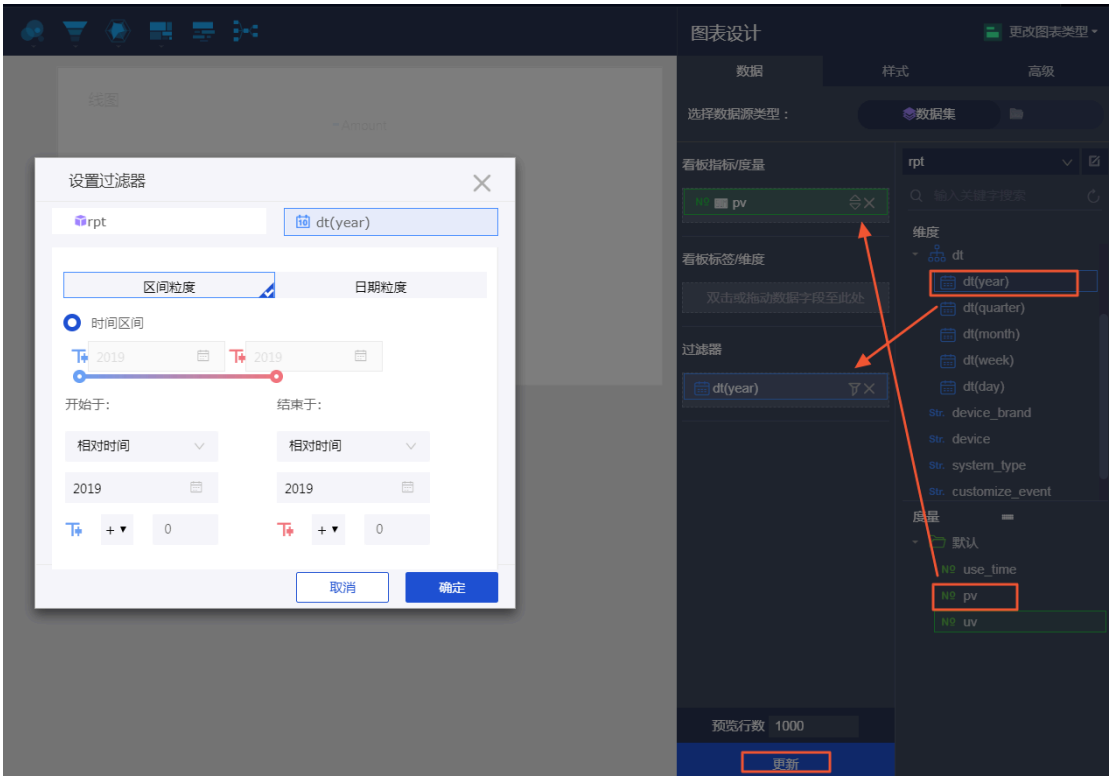
d. 实现动态联动查询。

a) 单击rpt数据集后的新建仪表板，选择常规模式，进入仪表板编辑页。

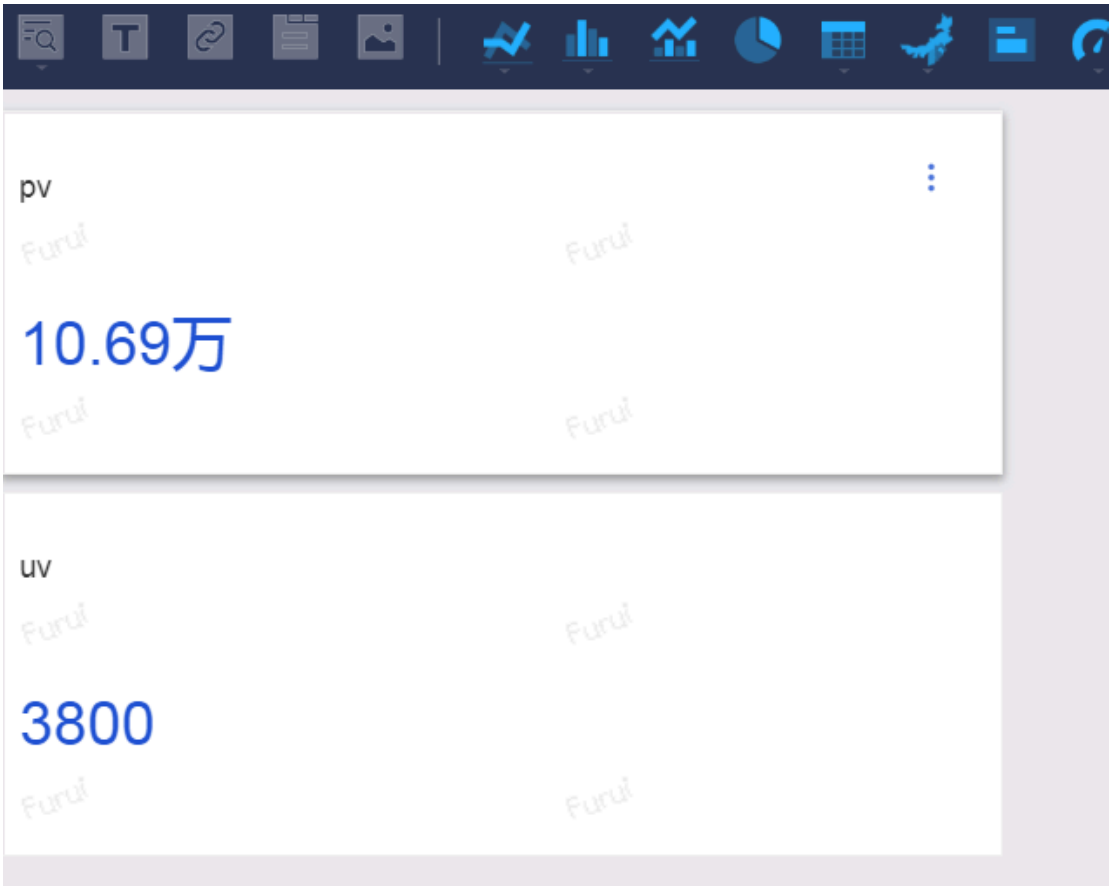
b) 从仪表板空间中向空白区拖入2个指标看板，调整布局成一排。



- 指标看板一：选择数据来源为数据集rpt，选择度量为pv。由于数据表rpt_user_trace_log为分区表，因此必须在过滤器处选择筛选的日期，本例中筛选为2019~2019年，完成设置后单击更新。



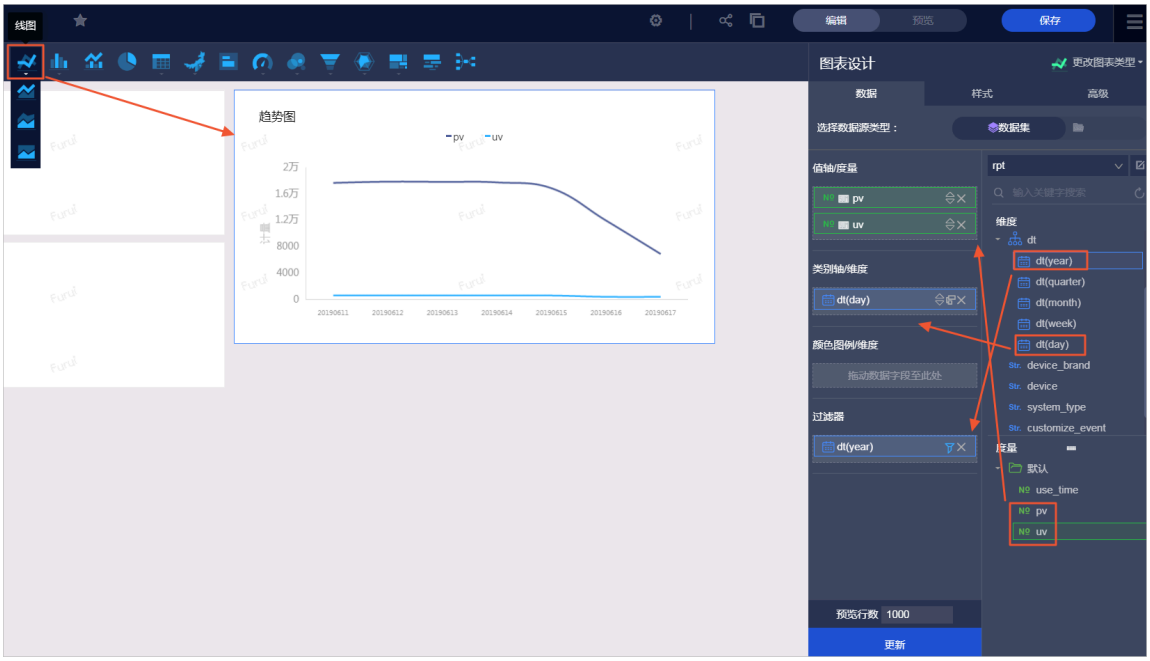
- 指标看板二：选择数据来源为来自数据集rpt，选择度量为uv，其他操作同上。完成设置后单击更新样式处设置指标看板显示的名称，显示效果如下。



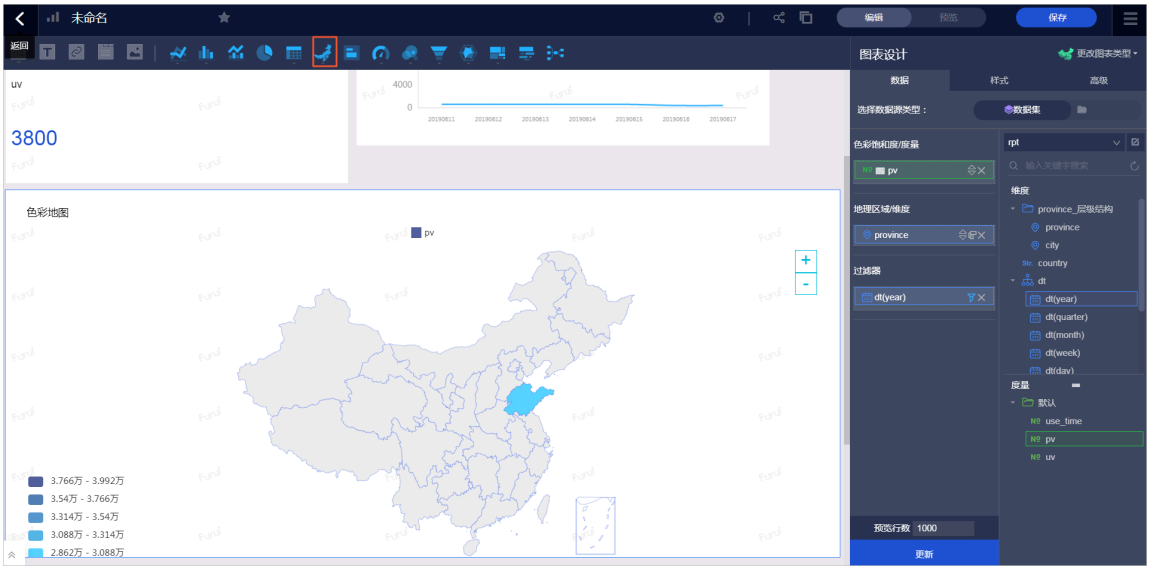
c) 制作趋势图：将图表区域内的线图拖拽到左侧画布。

参数配置如下，完成之后单击更新：

- 值轴/度量：pv、uv
- 类别轴/维度：dt（day）
- 过滤器：dt（year）



d) 制作色彩地图：单击图表区域内的色彩地图，并选择数据源来源为数据集rpt_user_trace_log，选择地理区域/维度为province（地区）、色彩饱和度/度量为pv，选择完成后单击更新，结果如下。



e) 完成配置后，单击保存及预览，即可看到展示效果。


RPT
预览
保存
pv
10.69万
uv
3800
趋势图
pv
uv
20190611
20190612
20190613
20190614
20190615
20190616
20190617
色彩地图
pv

3 数据质量保障教程

背景信息

本教程基于一份真实的网站日志数据集，数据来源于某网站上的HTTP访问日志数据。基于这份网站日志，您可以实现如下分析需求：统计并展现网站的浏览次数（PV）和独立访客（UV），并能够按照用户的终端类型（如Android、iPad、iPhone、PC等）和地域分别统计。

在整体数据链路的处理过程中，为保证最终产出数据的质量，您需要对数据仓库的ODS、CDM和ADS层的数据分别进行监控。数据仓库分层的定义请参见#unique_52。本教程基于教程《搭建互联网在线运营分析平台》，ods_user_trace_log、dw_user_trace_log、rpt_user_trace_log分别代表数据仓库的ODS、CDM和ADS层，详情请参见设计工作流。

如何衡量数据质量

3.1 数据质量教程概述

数据质量是数据分析结论有效性和准确性的基础。本文为您介绍数据质量保障教程的业务场景以及如何衡量数据质量的高低。

前提条件

在开始本教程前，请您首先完成教程《搭建互联网在线运营分析平台》，详情请参见#unique_61。

说明:

由于数据质量当前仅在华东2区域开放，请您在华东2区域创建DataWorks工作空间，完成教程。

业务场景

要保证业务数据质量，首先您需要明确数据的消费场景和加工链路。

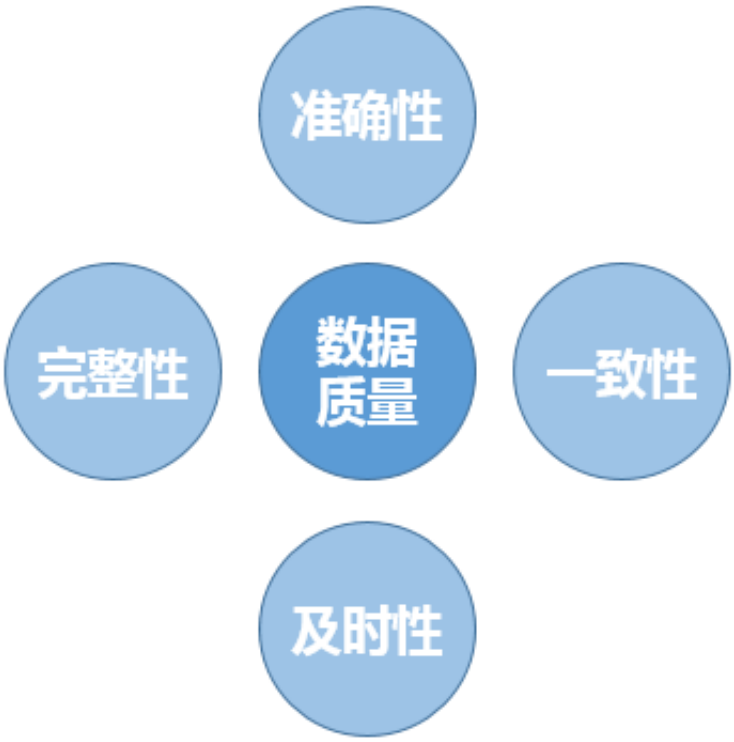
本教程基于一份真实的网站日志数据集，数据来源于某网站上的HTTP访问日志数据。基于这份网站日志，您可以实现如下分析需求：统计并展现网站的浏览次数（PV）和独立访客（UV），并能够按照用户的终端类型（如Android、iPad、iPhone、PC等）和地域分别统计。

在整体数据链路的处理过程中，为保证最终产出数据的质量，您需要对数据仓库的ODS、CDM和ADS层的数据分别进行监控。数据仓库分层的定义请参见#unique_52。本教程基于教程《搭建互联网在线运营分析平

台》，ods_user_trace_log、dw_user_trace_log、rpt_user_trace_log分别代表数据仓库的ODS、CDM和ADS层，详情请参见设计工作流。

如何衡量数据质量

数据质量可以从完整性、准确性、一致性和及时性共四个角度进行评估。详情请参见#unique_62。



在本教程中，您将学会通过数据质量风险监控，保证数据的完整性、准确性。通过数据及时性监控，保证数据的及时性。

- 完整性

  完整性是指数据的记录和信息是否完整、不缺失。数据的缺失包括数据记录的缺失（表行数异常）和记录中某字段信息的缺失（字段出现空值）。在本教程中，您需要重点关注数据的生产环节（MaxCompute外部表引用的表格存储数据）和加工环节（数仓CDM及ADS层）中表行数是否大于0、表行数波动是否正常以及字段是否出现空值或重复的情况。

- 准确性

  准确性是指数据记录中信息和数据是否准确、不存在错误或异常。例如，在本教程中，如果UV、PV数值小于0，则明显是错误数据。

- 一致性

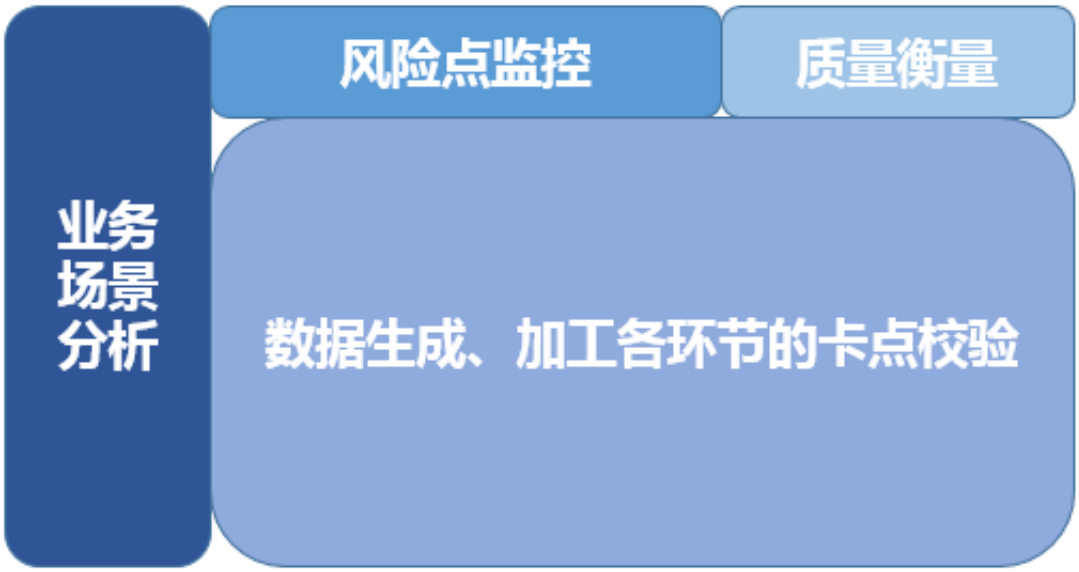
  对于不同的业务流程和节点，同一份数据必须保持一致性。例如表字段的province字段中如果有浙江、ZJ两种表述，在您`group by province`时会出现两条记录。

- 及时性

  及时性主要体现在最终ADS层的数据可以及时产出。为保证及时性，您需要确保整条数据加工链路上的每个环节都可以及时产出数据。本教程将利用DataWorks智能监控功能保证数据加工每个环节的及时性。

3.2 数据质量管理流程

数据质量的管理流程包括业务数据资产定级、加工卡点、风险点监控、及时性监控，您可以构建属于自己的数据质量保障体系。



数据质量的管理流程如下：

1. #unique_64。
2. 离线数据加工卡点。
3. #unique_66。
4. #unique_67。

3.3 数据资产定级

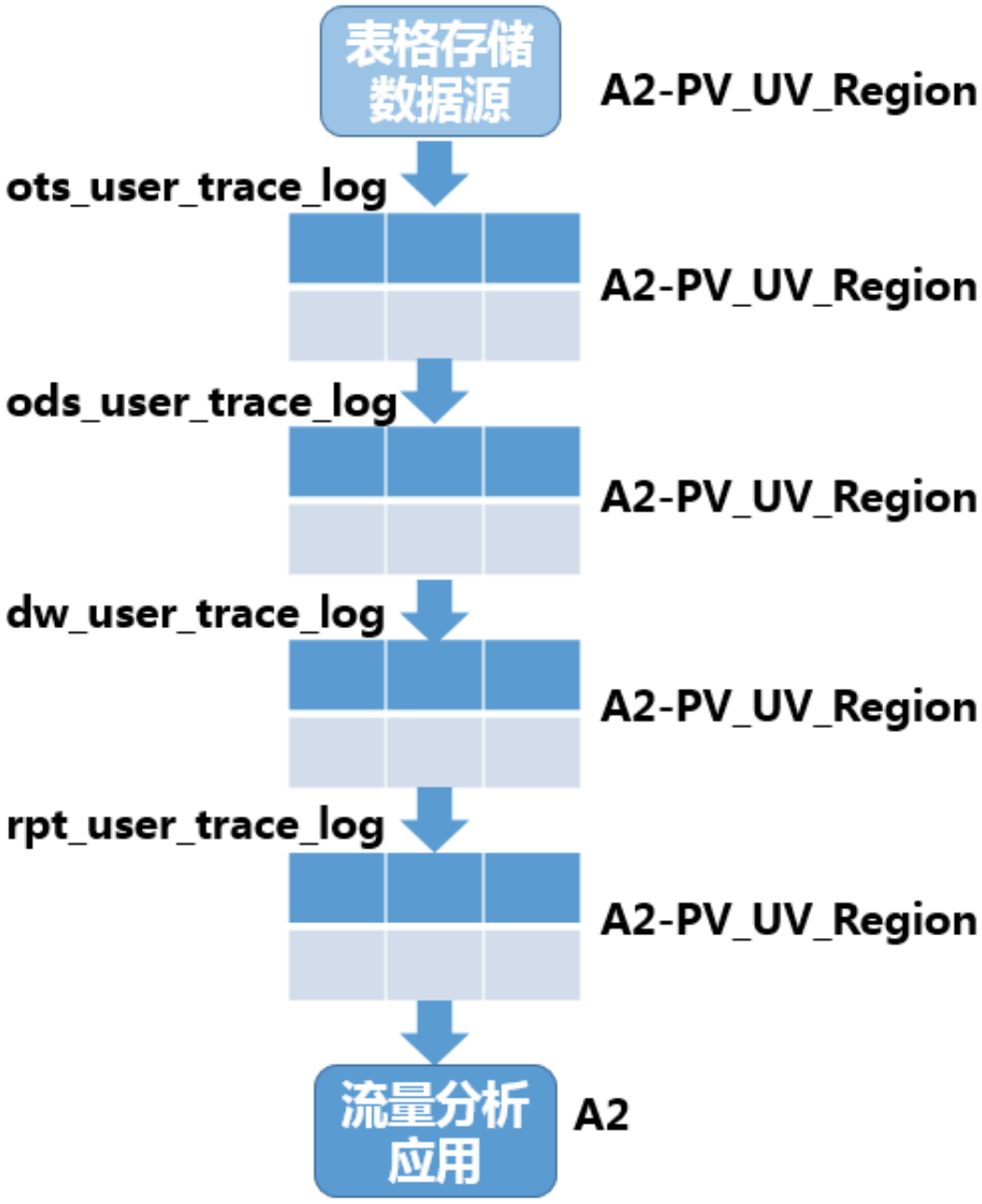
数据的资产等级，可以根据数据质量不满足完整性、准确性、一致性、及时性后对业务的影响程度进行划分。

数据等级定义如下：

- 毁灭性质：数据一旦出错，将会引起重大资产损失，面临重大收益损失等。
- 全局性质：数据直接或间接用于企业级业务和效果评估、重要决策等。
- 局部性质：数据直接或间接用于某些业务线的运营、报告等，若出现问题会给业务线造成一定的影响或造成工作效率降低。
- 一般性质：数据主要用于日常数据分析，出现问题带来的影响极小。
- 未知性质：无法明确数据的应用场景。

资产等级可以用Asset进行标记：毁灭性质为A1，全局性质为A2，局部性质为A3，一般性质为A4，未知性质为Ax。重要程度为：A1>A2>A3>A4>Ax。

在数据流转链路上，您需要整理各个表是被哪些应用业务消费。通过给这些应用业务划分数据资产等级，结合数据的上下游依赖关系，将整个链路打上某一类资产等级的标签。在本教程中，互联网在线运营分析平台只存在一个应用：统计并展现网站的PV和UV，并能够按照用户的终端类型和地域进行统计，命名为PV_UV_Region。假设该应用会直接影响整个企业的重要业务决策，您可以定级应用为A2，从而整个数据链路上的表的数据等级，都可以标记为A2-PV_UV_Region。



说明:

当前MaxCompute暂无配套资产等级打标工具，您可以使用其他工具完成打标。

3.4 离线数据加工卡点

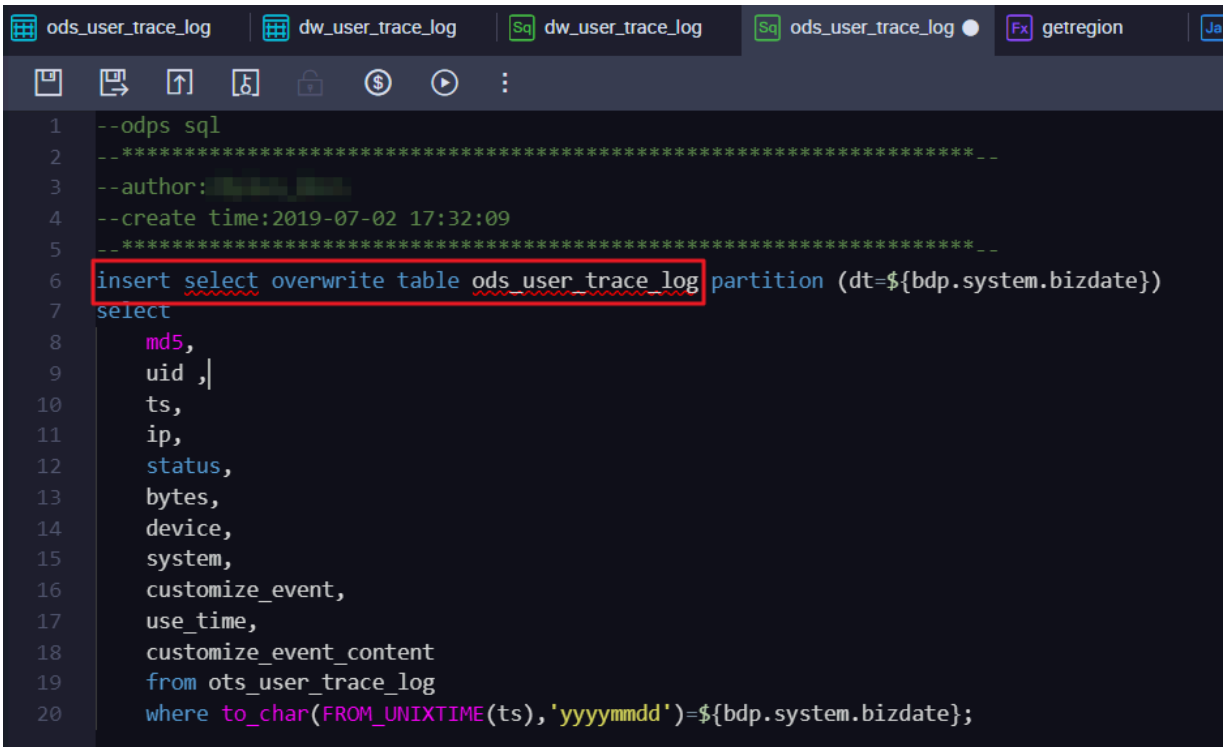
离线数据加工卡点，主要指在业务系统的数据生成过程中进行的卡点校验。

代码提交的卡点校验

代码提交卡点主要包括您在提交代码时，手动或自动进行SQL扫描，检查您的SQL逻辑。校验规则分类如下：

- 代码规范类规则，如表命名规范、生命周期设置、表注释等。
- 代码质量类规则，如分母为0提醒、NULL值参与计算影响结果提醒、插入字段顺序错误等。
- 代码性能类规则，如分区裁剪失效、扫描大表提醒、重复计算检测等。

您在使用DataWorks数据开发功能时，如果代码中有语法错误，会出现如下红色波浪线提示。



关于SQL代码、表命名、生命周期、注释的其他规范，请参见#unique_69及#unique_70。

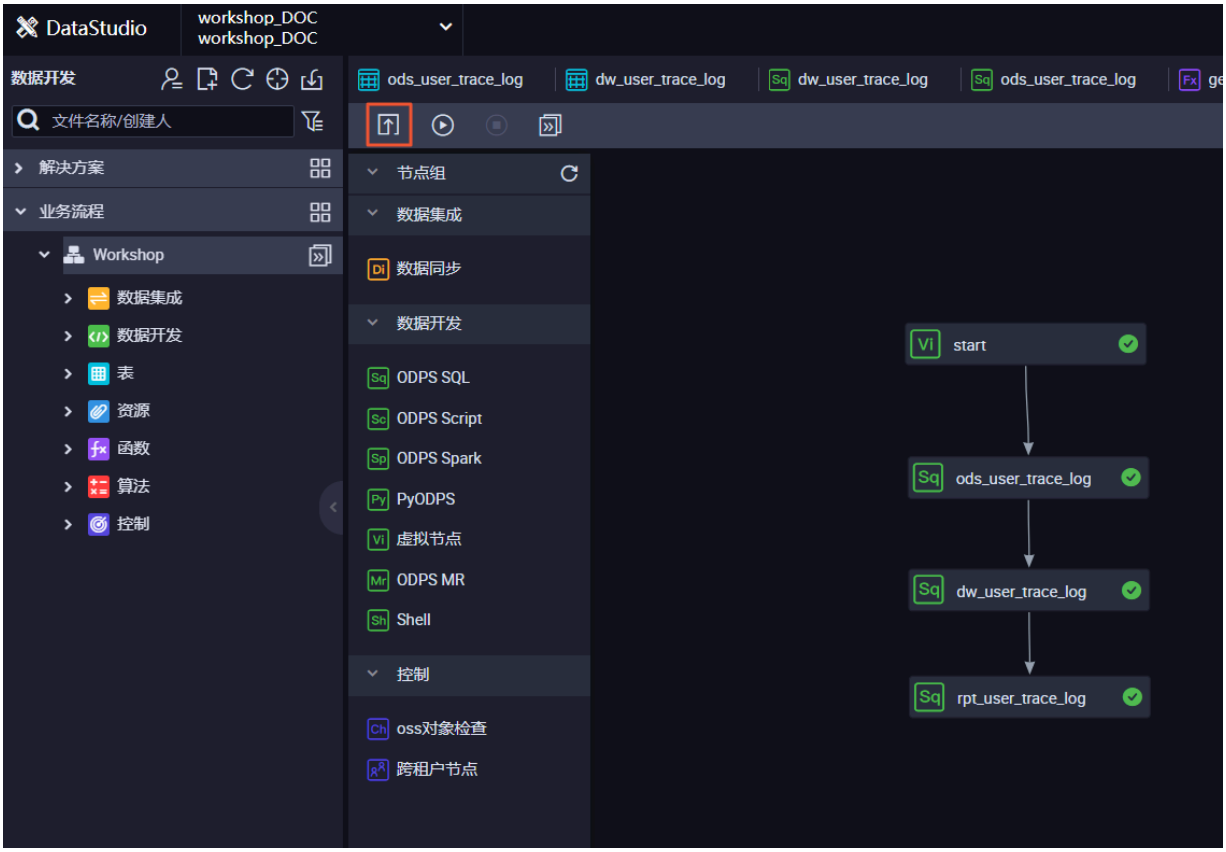
任务发布卡点

为保证线上数据的准确性，每次变更都需要经过测试再发布到线上生产环境，且生产环境测试通过后才算发布成功。发布上线前的测试包括代码审查和回归测试。对于资产等级较高的应用，必须在完成回归测试之后，才允许任务发布，本教程中应用为A2等级，属于高资产级别应用。

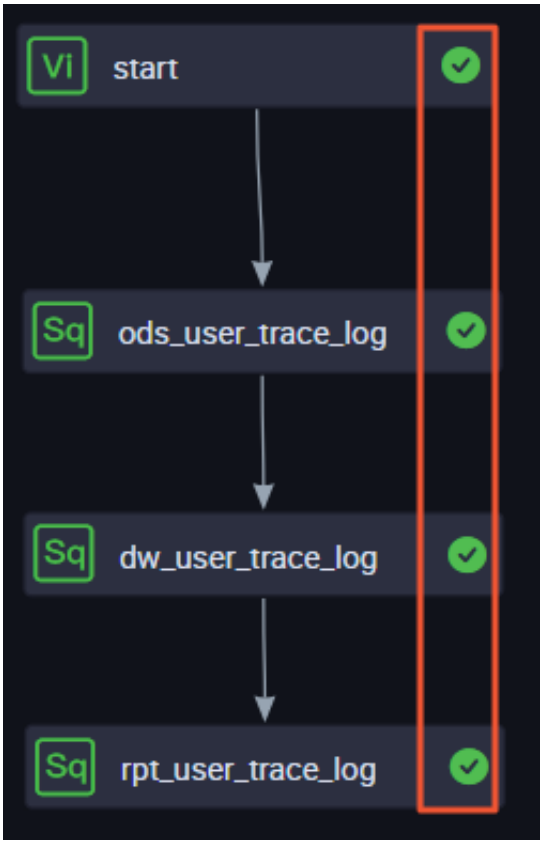
回归测试需保证您能充分模拟真实环境进行测试：

- 对于标准模式项目，您可使用SQL语句将数据从生产环境复制开发环境，运行业务流程。
- 对于简单模式的项目，您可以直接运行业务流程，观察是否有报错，详情请参见#unique_71。

在本教程中，由于使用简单模式，您只需提交任务。



完成运行后，如果所有节点都显示绿色图标，则表示业务流程测试通过。



相关人员通告

在进行更新操作前，需要通知下游变更原因、变更逻辑、变更时间等信息。下游对此次变更没有异议后，再按照约定时间执行发布变更，将变更对下游的影响降到最低。例如，在本教程中，如果表格存储数据源的表结构发生了变更，您需要通知ots_user_trace_log、ods_user_trace_log、dw_user_trace_log、rpt_user_trace_log表的责任人，及时更新表结构。

3.5 数据质量风险监控

数据质量风险监控主要针对数据的准确性、一致性和完整性。本教程使用DataWorks数据质量（DQC）功能，完成数仓各层次的数据质量监控。

前提条件

您需要首先完成教程《搭建互联网在线运营分析平台》，并保证您的DataWorks工作空间创建区域为华东2上海，详情参见#unique_61。您需要完成数据资产定级，本教程中定义为A2，详情请参见#unique_64。

说明:

数据质量风险监控理论规范，请参见#unique_73。

背景信息

数据质量监控和数据资产等级对应，您可以根据以下因素细化您的监控配置，数据质量使用详情请参见#unique_26。

- 监控分类：数据量、主键、离散值、汇总值、业务规则和逻辑规则。
- 监控粒度：字段级别、表级别。
- 监控层次：ODS、CDM、ADS三层数据，其中ODS和DWD层主要偏重数据的完整和一致性。DWS和ADS层数据量较小、逻辑复杂，偏重数据的准确性。

以下为不同数据资产等级和数仓层次数据的数据质量监控建议，仅供参考。

数据质量DQC监控规范									
监控分类					数据量	主键	离散值	汇总值	业务逻辑、规则
适合场景					所有非临时表都建议配置该项监控。	对于存在业务主键、逻辑主键的表需配置该监控。	维表、事实表中的维度值、状态值、可枚举的值需配置该监控。	汇总统计表中的汇总值需配置该监控。	1、重要指标的异常值监控。例如，正常UID长度是否为32位。 2、字段间的平衡值监控。例如，字段a与字段b满足一一对应关系等。 3、多表关联监控。例如两张表左关联，关联不上的记录数应等于0。
监控粒度					表级数据量监控	字段级	字段级	字段级	字段级/表级
常用监控规则					表行数波动/自助规则表行数>固定值	模板规则的字段空值、重复值/自定义规则监控联合主键空值、重复值情况	离散值分组个数/离散值分组个数波动/离散值状态值波动	模板规则的单字段大于0/自定义规则判断字段等于0所占的比例等	自定义规则
层次	表类型				规则配置				
ODS/DWD	离线表	A2	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	不涉及	需监控
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	不涉及	需监控
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	不涉及	需监控
		A3	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	不涉及	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
		A4	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	不涉及	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
		Ax	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	不涉及	不涉及	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	不涉及	不涉及	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	不涉及	不涉及	不涉及
DWS/ADS	离线表	A2	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	需监控	需监控
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	需监控	需监控
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	需监控	需监控
		A3	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	需监控	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	需监控	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	需监控	不涉及
		A4	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	需监控	不涉及	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	需监控	不涉及	不涉及
		Ax	增量表	有周期规律	模板表行数波动率	空值、重复值唯一性	不涉及	不涉及	不涉及
				无周期规律	自助表行数>固定值	空值、重复值唯一性	不涉及	不涉及	不涉及
			全量表		模板表行数波动率	空值、重复值唯一性	不涉及	不涉及	不涉及

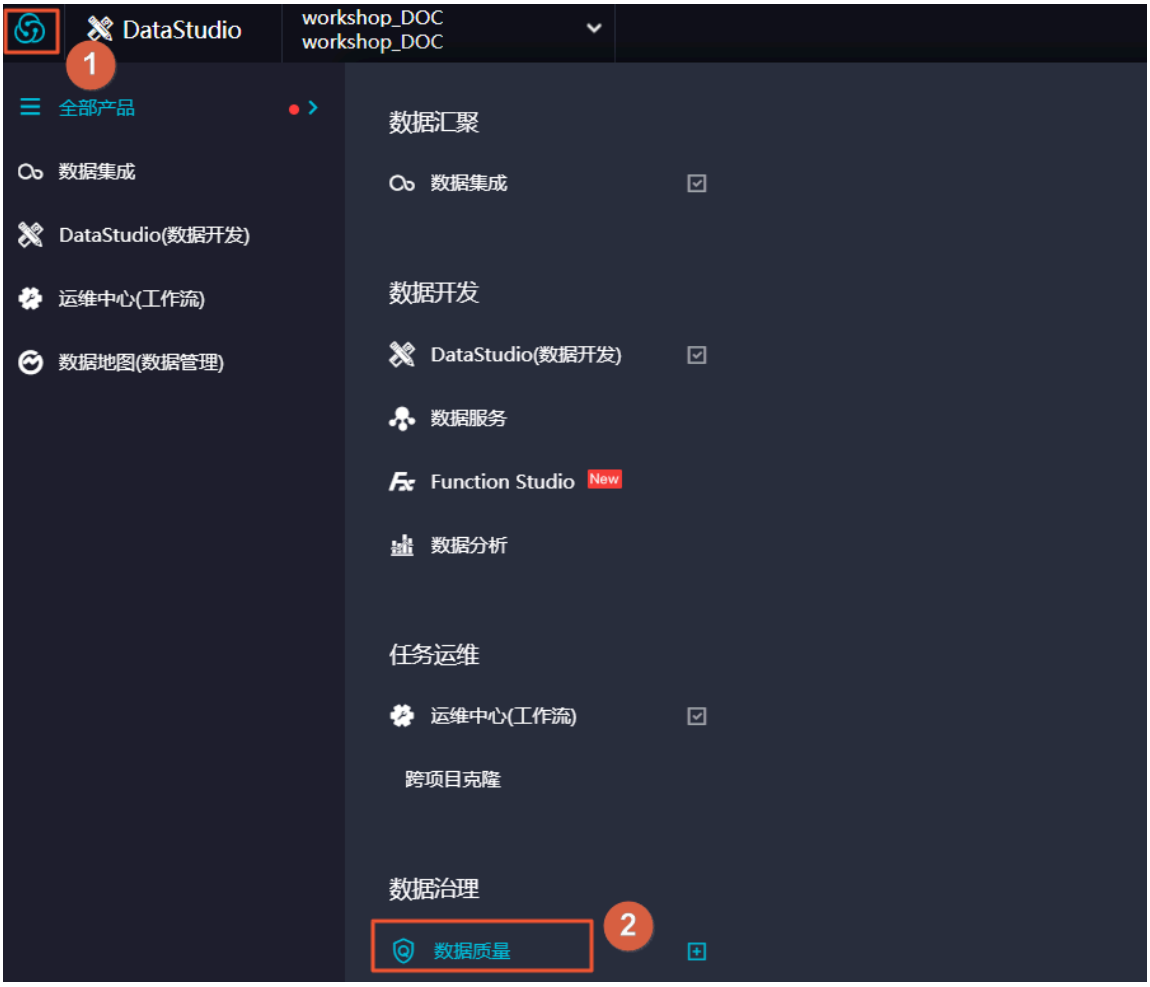
操作步骤

1. ODS层数据质量监控

ODS层表里的数据来源于OSS上的日志文件，作为源头表，您需要尽早判断此表分区中是否有数据。如果这张表中没有数据，后续任务运行无意义，则需要阻止后续任务运行。

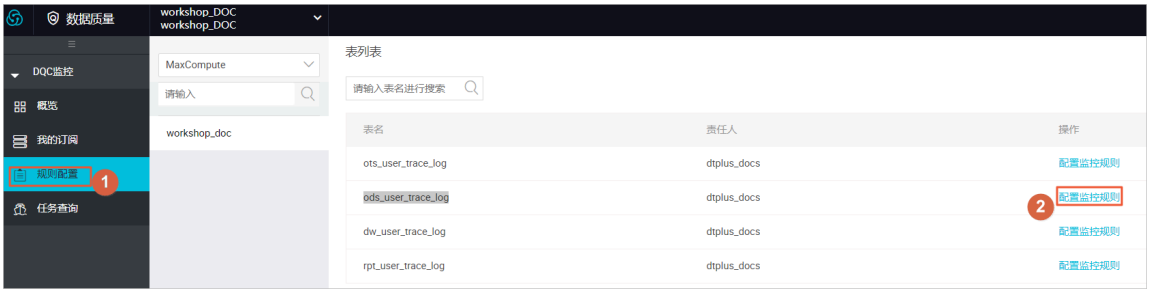
a) 进入数据质量。

在您的数据开发页面，单击左上角图标，选择数据质量。



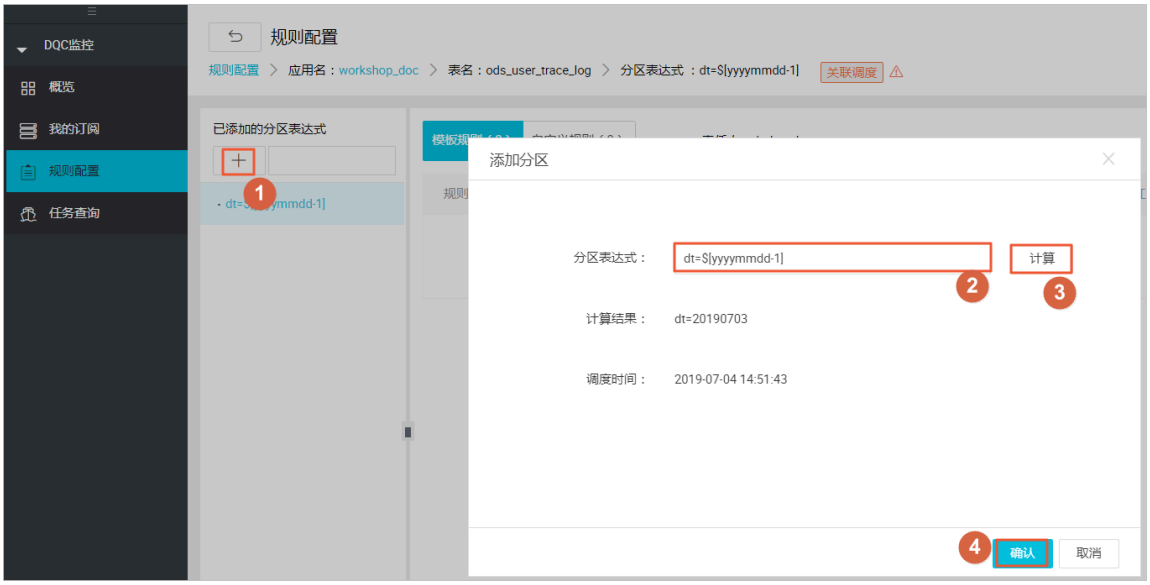
b) 进入ods_user_trace_log规则配置页面。

在规则配置页面找到代表外部数据源的表ods_user_trace_log，单击配置监控规则。



c) 添加分区。

单击+，在分区表达式一栏输入 dt=$[yyyymmdd-1]，对应表ods_user_trace_log的分区格式${bdp.system.bizdate}（获取到前一天的日期）。分区表达式的详细信息请参见#unique_55，若表中无分区列，可以配置无分区。



d) 单击创建规则。

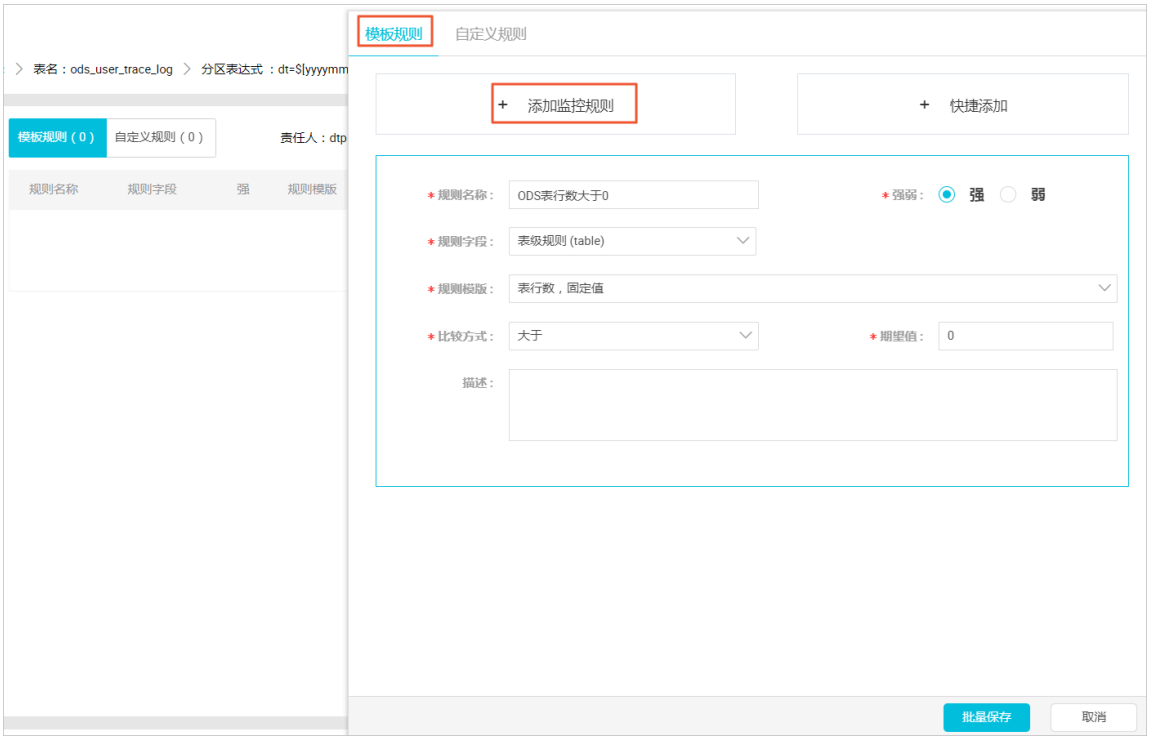
e) 监控表行数大于0。

单击模板规则 > 添加监控规则。

在配置规则时，选择规则模板为表行数、固定值，将规则的强度设置为强，比较方式设置为期望值大于0。目的为保证ODS层分区内存在表数据。

说明:

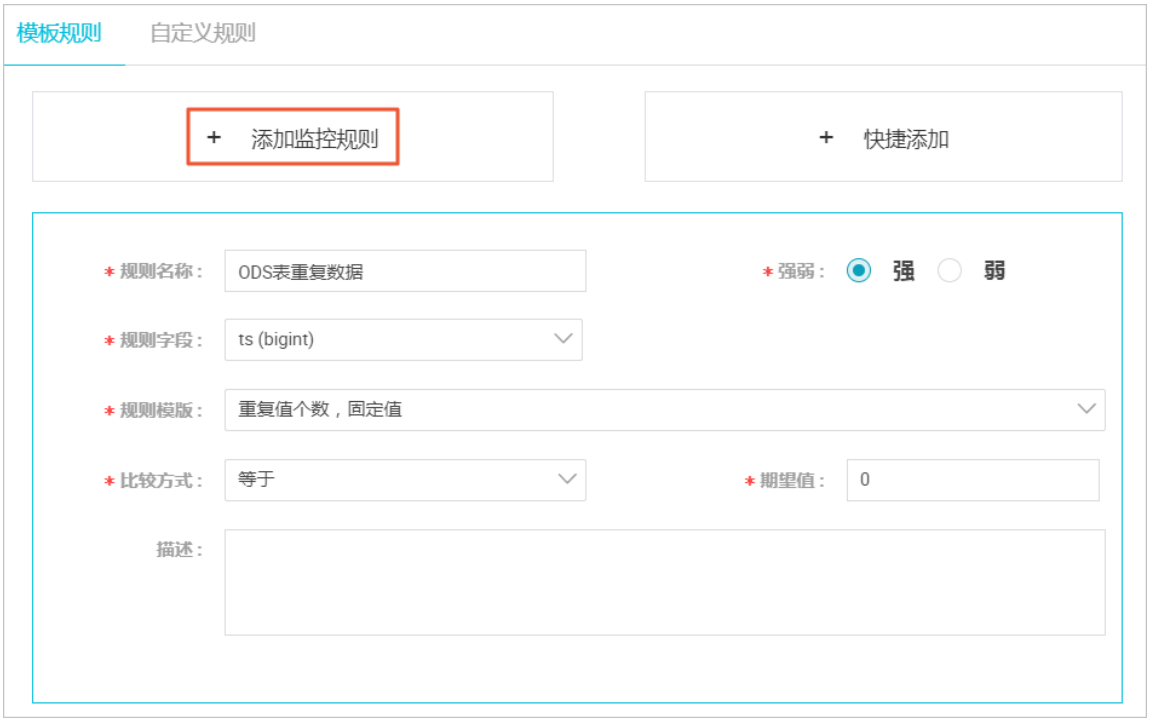
只有强规则下红色报警会导致任务阻塞，阻塞会将任务的实例状态置为失败。



f) 监控重复数据。

单击添加监控规则。

配置规则时，选择规则字段为ts(bigint)，规则模板为重复值个数、固定值，将规则的强度设置为强，比较方式设置为期望值等于0。ts(bigint)值为用户时间戳，目的是避免ODS层出现重复的数据。



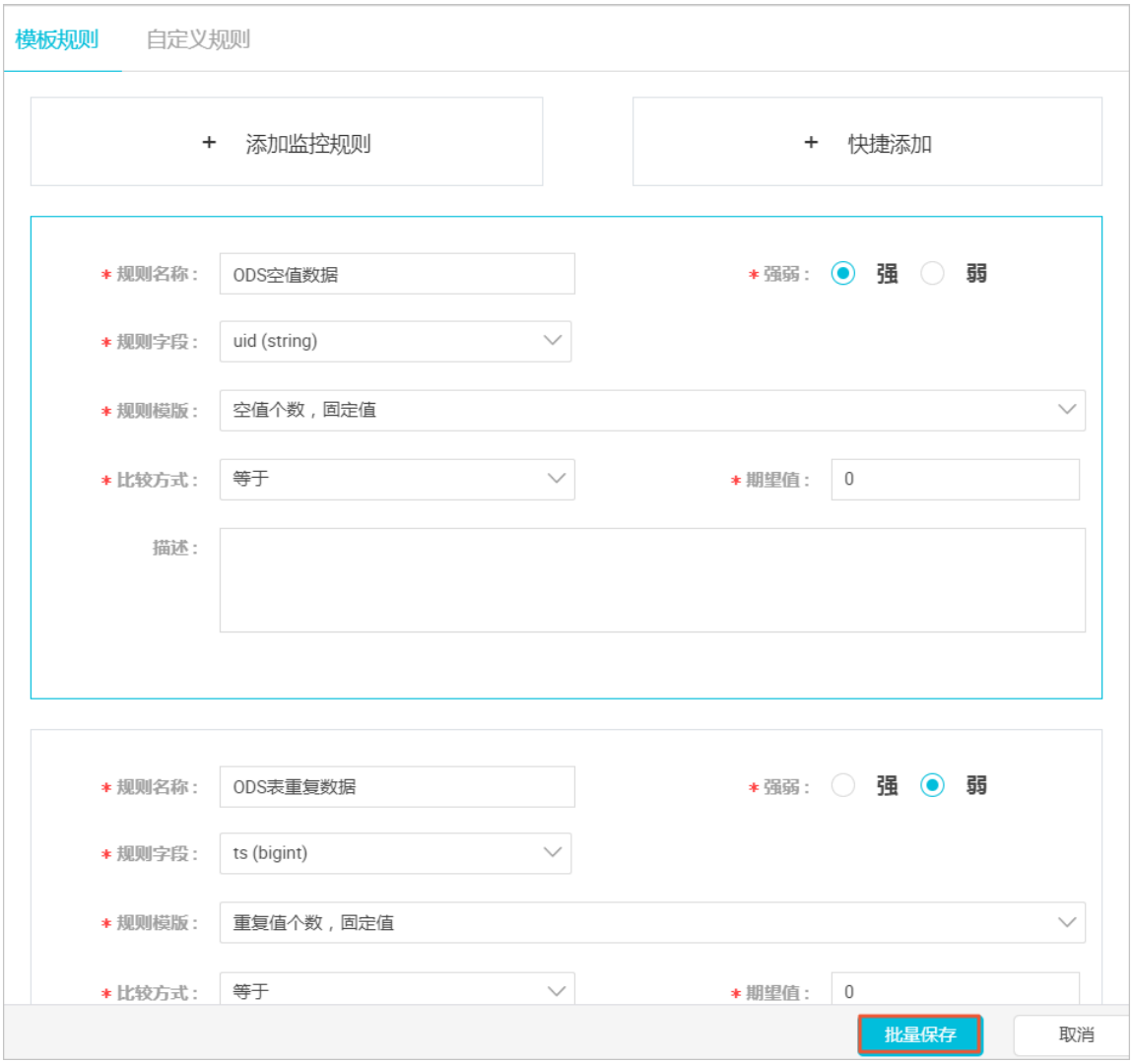
g) 监控空值数据。

单击添加监控规则。

配置规则时，选择规则字段为uid(string)，规则模板为空值个数、固定值，将规则的强度设置为强，比较方式设置为期望值等于0。uid(string)值为用户ID，目的是避免出现用户ID为空值的脏数据。

h) 批量保存规则。

完成上述操作后，单击批量保存。

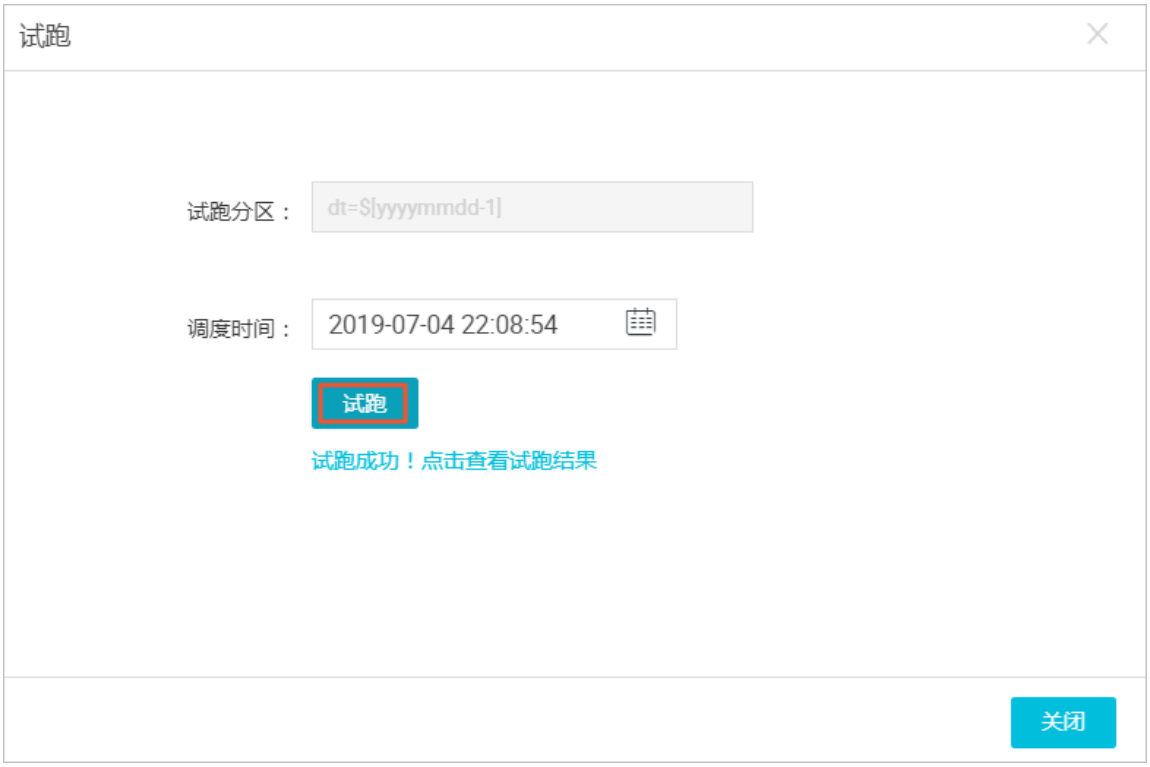


i) 规则试跑。

右上角有一个节点试跑的按钮，可以在规则配置完毕后，进行规则校验。单击试跑按钮，可立即触发数据质量的校验规则。

j) 查看试跑结果。

单击试跑后，您可以单击试跑成功！点击查看试跑结果。



在弹出的页面中，您可以查看表数据是否已符合您的规则。

说明：

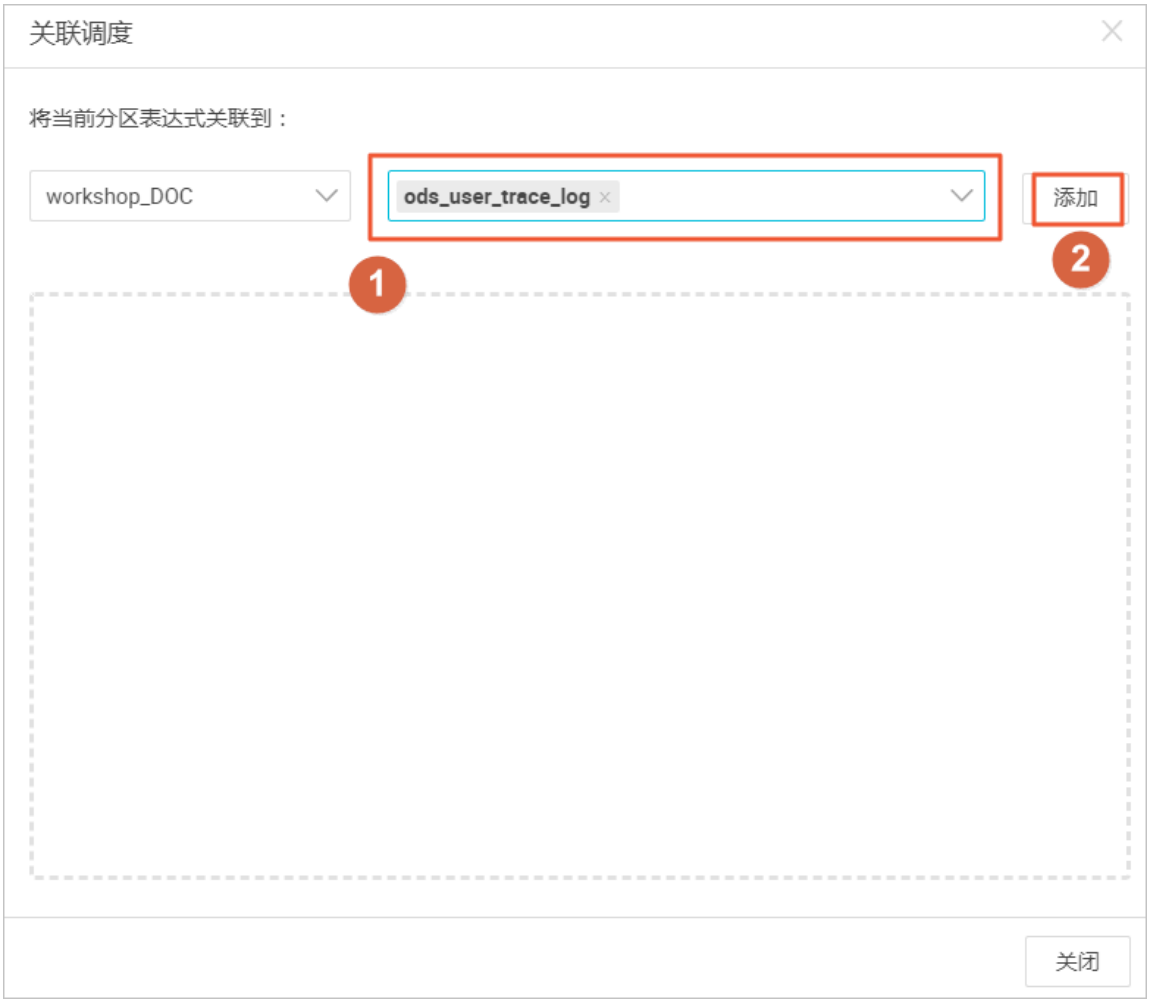
可根据试跑结果，来确认此次任务产出的数据是否符合预期。建议每个表规则配置完毕后，都进行一次试跑操作，以验证表规则的适用性。

k) 关联调度。

在规则配置完毕，且试跑又都成功的情况下。您需要将表和其产出任务进行关联，这样每次表的产出任务运行完毕后，都会触发数据质量规则的校验，以保证数据的准确性。在表规则和调度任务绑定后，任务实例运行完毕，都会触发数据质量的检查。

在表规则配置界面，单击关联调度，配置规则与任务的绑定关系。

在弹框中输入您需要关联的任务节点名称，单击添加。



关联调度后，表名后的小图标会变成蓝色。

l) 配置任务订阅。

关联调度后，每次调度任务运行完毕，都会触发数据质量的校验。数据质量支持设置规则订阅，可以针对重要的表及其规则设置订阅，设置订阅后会根据数据质量的校验结果进行告警，从而实现对校验结果的跟踪。若数据质量校验结果异常，则会根据配置的告警策略进行通知。

单击订阅管理，设置接收人以及订阅方式，目前支持邮件通知、邮件和短信通知、钉钉群机器人和钉钉群机器人@ALL四种方式。

订阅管理设置完毕后，可以在我的订阅中进行查看及修改，建议您订阅所有规则。

2. CDM层数据质量监控

CDM层数据质量监控配置方法与ODS层相同，区别在于监控规则的配置。

a) 添加分区表达式。

进入dw_user_trace_log表的规则配置页面，同样配置分区为dt=$[yyyymmdd-1]，完成添加后您可以在界面中看到已添加的分区表达式。

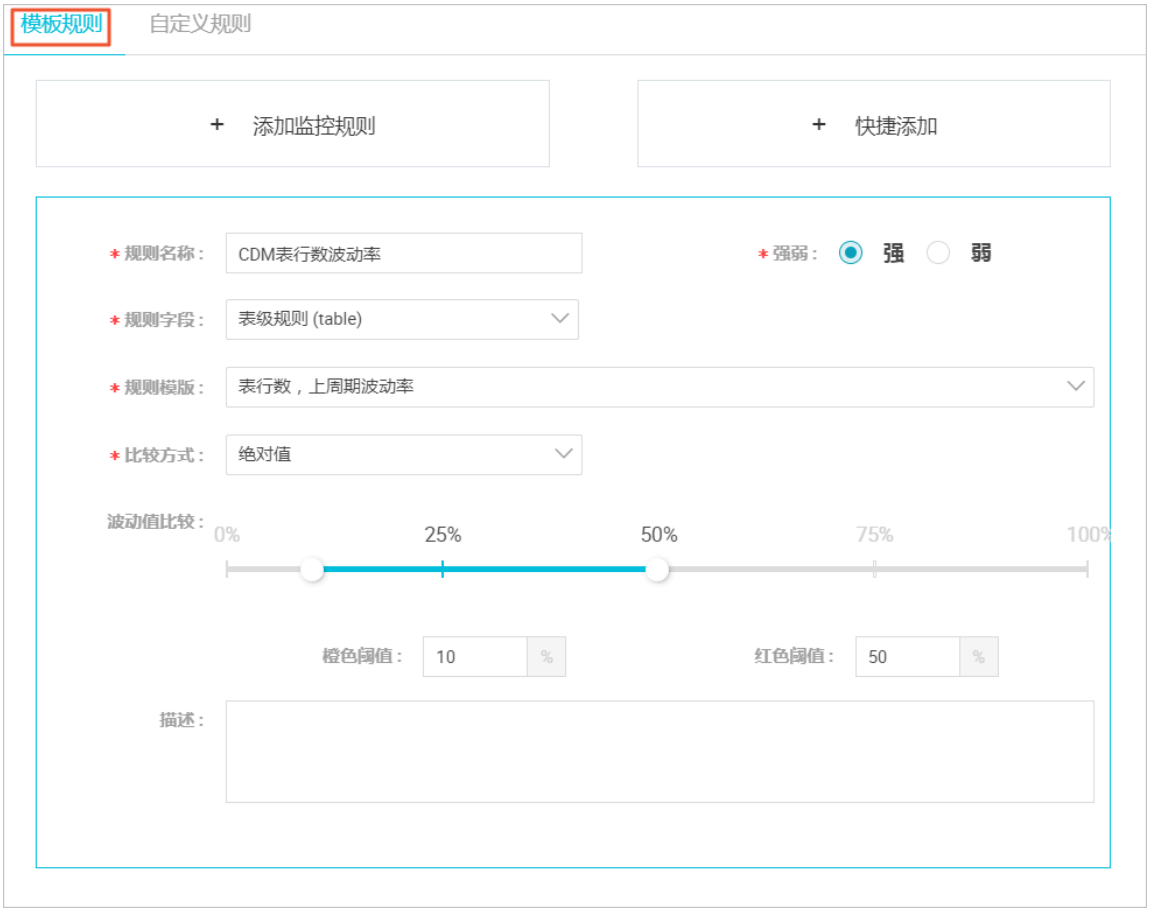
b) 监控表行数及空值数据。

表行数和空值数据的监控规则配置与ODS层相同，完成配置后如下图所示。

c) 监控表行数波动率。

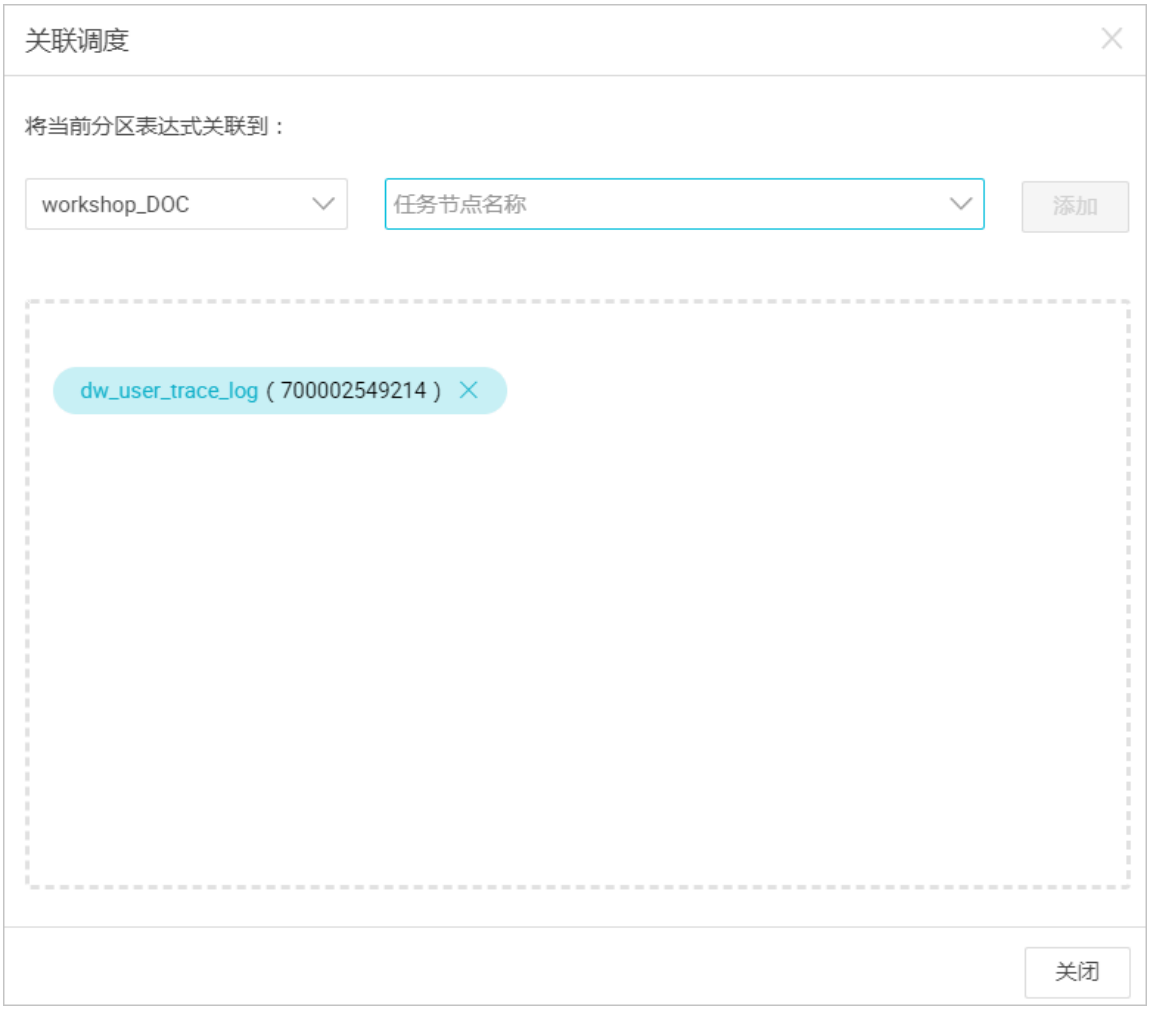
监控表行数波动率主要是为了避免出现突发的大量脏数据的污染。配置规则时，选择规则字段为表级规则(table)，规则模板为表行数、上周期波动率，将规则的强度设置为强，比较方

式设置为绝对值。橙色阈值为10，红色阈值为50，代表当表行数波动率到达50%时，会产生红色报警。



d) 规则试跑并关联调度。

方法同ODS层。


关联调度
将当前分区表达式关联到：
workshop_DOC
任务节点名称
添加
dw_user_trace_log (700002549214)
关闭

3. ADS层数据质量监控

ADS层数据质量监控配置方法与ODS层相同，区别在于监控规则的配置。

a) 添加分区表达式。

进入rpt_user_trace_log表的规则配置页面，同样配置分区为dt=$[yyyymmdd-1]，完成添加后您可以在界面中看到已添加的分区表达式。

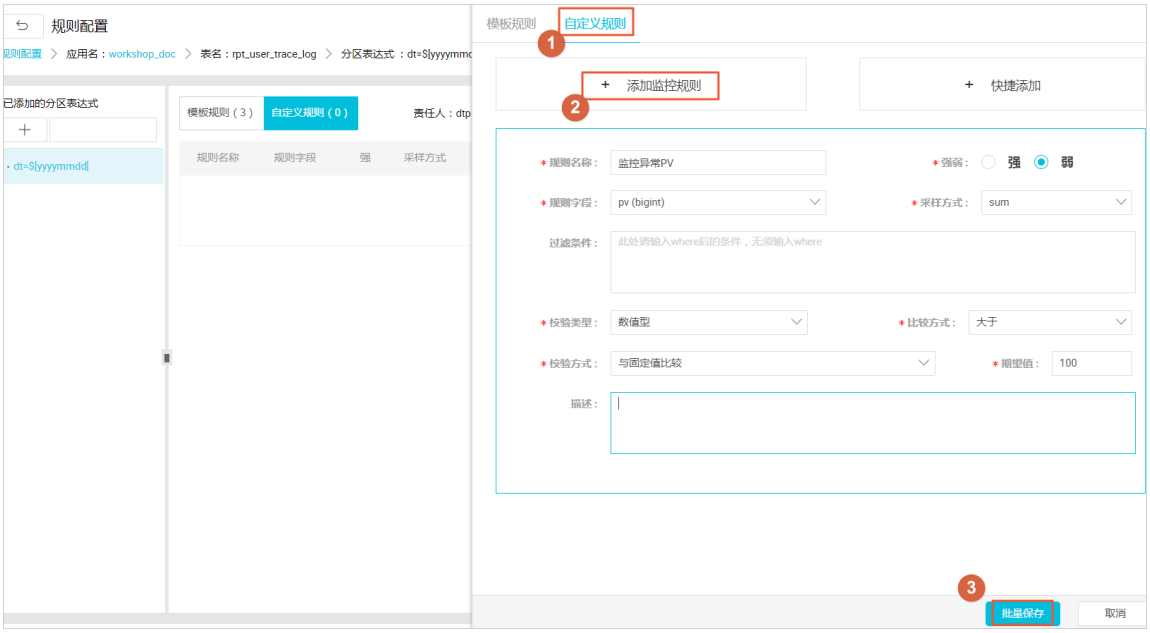
b) 监控表行数、波动率及空值数据。

监控表行数、波动率和空值数据的监控规则配置与CDM层相同。由于在数仓分层中，越靠近应用层数据越少、约束性越低，强弱选择为弱，完成配置后如下图所示。

c) 监控表异常PV。

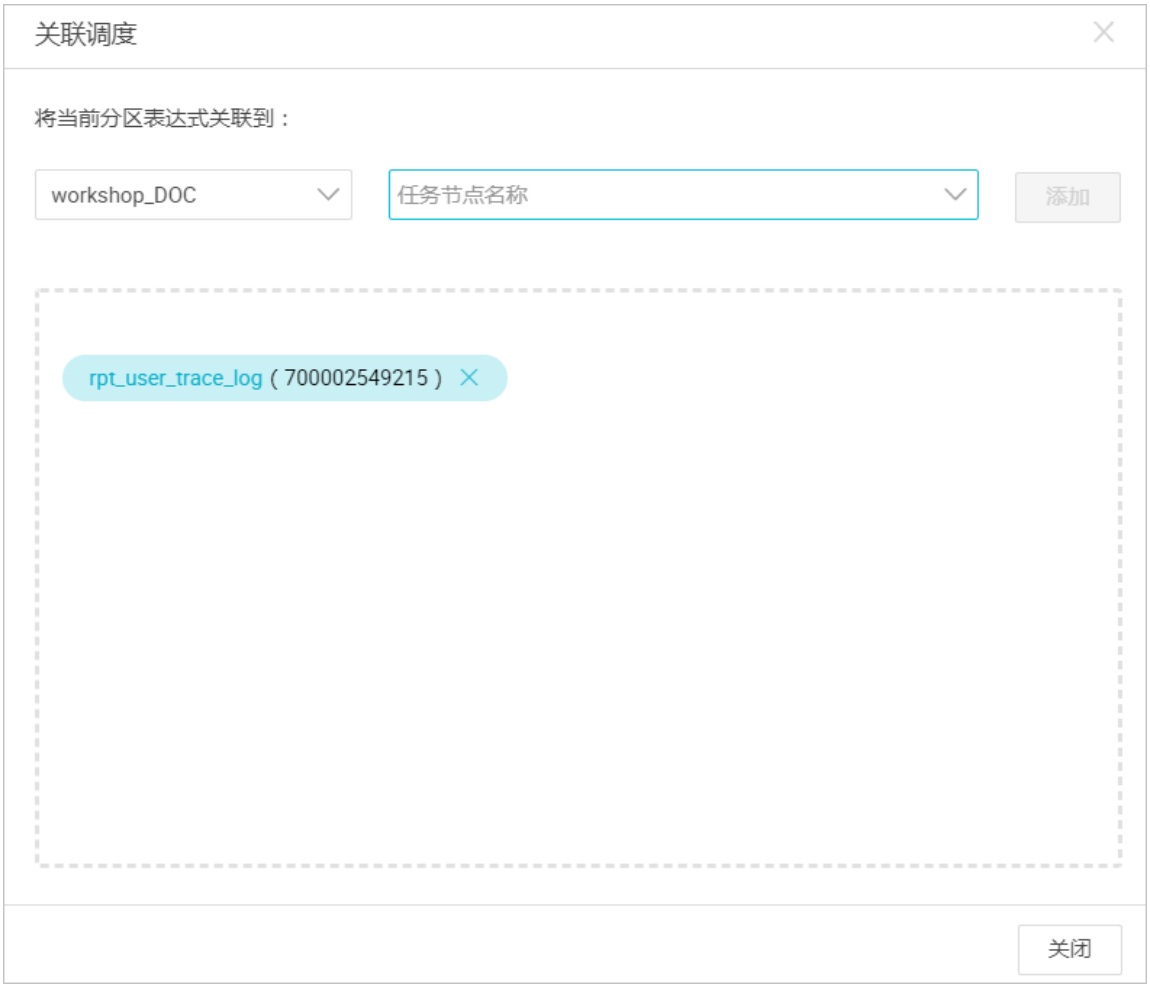
您可以利用自定义规则功能监控ADS层的应用数据。配置规则时，选择规则字段为pv(bigint)，采样方式为sum，将规则的强度设置为弱，比较方式设置为大于期望

值100。这样，当PV和异常锐减到100时，您可以及时收到告警。完成配置后，单击批量保存。



d) 规则试跑并关联调度。

方法同ODS层。



3.6 数据及时性监控

基于MaxCompute的离线任务会对数据产出有时间要求，在确保数据准确性的前提下，您需要进一步让数据能够及时提供服务。本文为您介绍如何使用DataWorks智能监控功能完成数据及时性的监控。

前提条件

本文为您演示基础版DataWorks的基本智能监控功能：规则管理。如果您想使用完整的智能监控功能，需至少购买标准版DataWorks，详情请参见#unique_75。关于DataWorks智能监控功能详情请参见#unique_76。

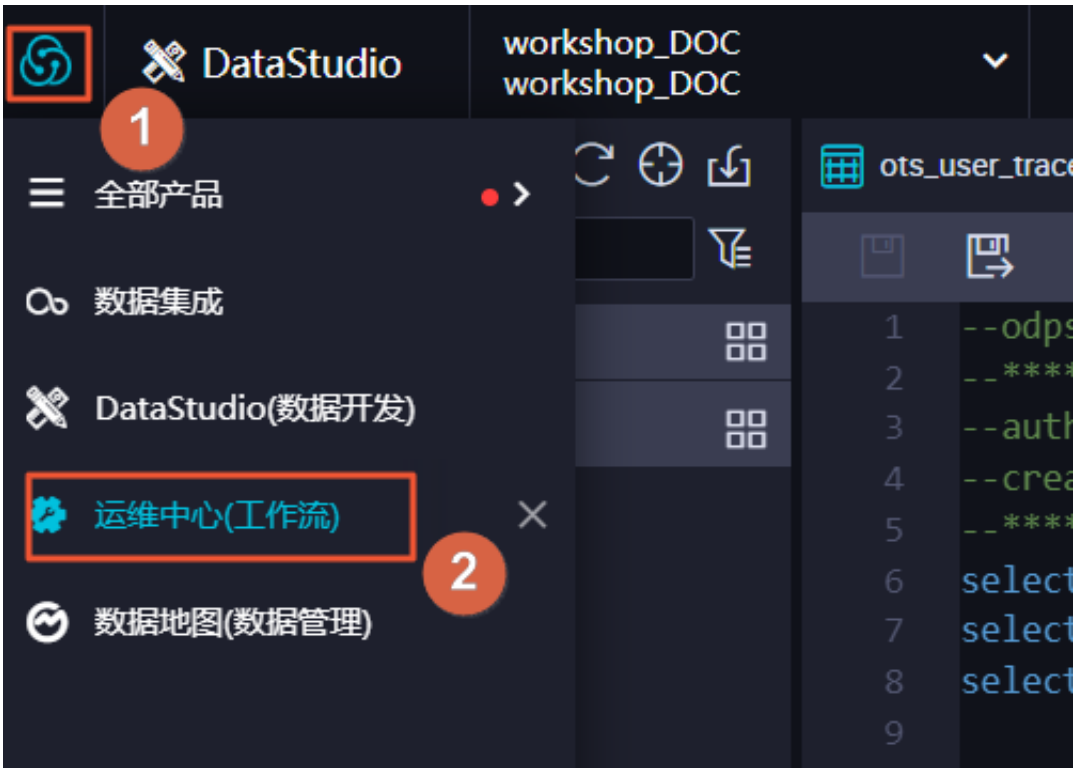
背景信息

在对数据产出及时性监控前，您需要首先确定调度任务的优先级。数据资产等级越高的任务节点，优先级越高，您可以给予更加严格的数据及时性监控和告警规则。

操作步骤

1. 进入规则管理页面。

   在DataStudio页面单击运维中心（工作流）。

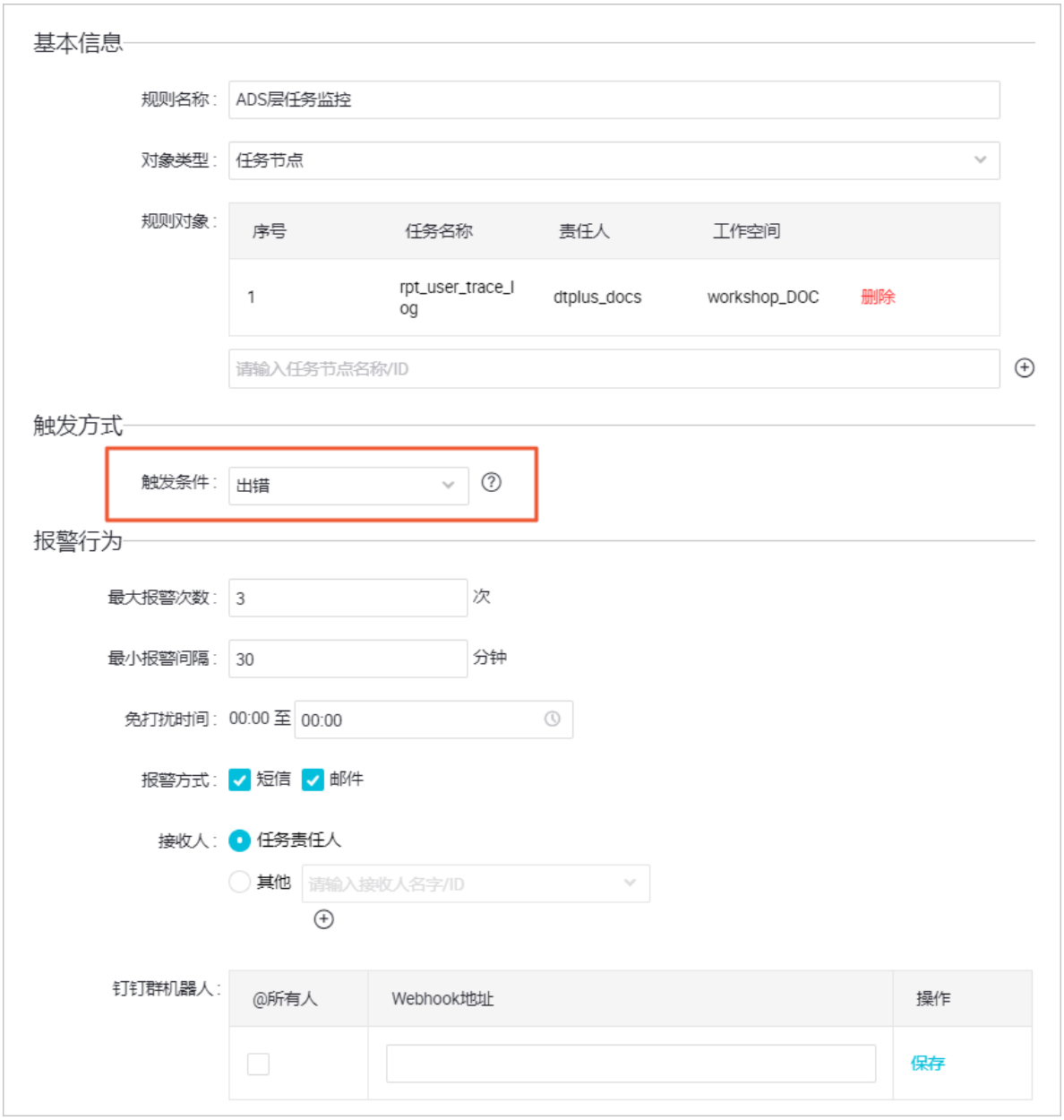


   单击智能监控 > 规则管理。关于规则管理的详情请参见#unique_77。

2. 新建自定义规则。

单击新建自定义规则，输入参数后单击确定即可。

在本例中，监控整个业务流程每次运行时间不可超过30分钟。如果运行时间超过30分钟，则每30分钟上报一次告警。连续上报3次告警，系统自动以邮件及短信的方式来上报。对于重要的任务节点，您还可以单独设置任务节点规则，并定义其他触发条件。



3. **数据及时性优化。**

   通常，影响数据按时产出的主要原因和优化方式如下表所示。

问题原因	问题优化
计算资源不足 **· 资源总量不足。例如，资源上限为500，但您提交了需要1000资源的任务。** **· 资源分配不合理，重要任务未优先分配资源。**	**扩容计算资源，或让您的核心计算任务独占资源。**

问题原因	问题优化
代码执行效率低 · 代码冗余，例如扫描所有分区。 · 节点任务配置不合理，例如出现长尾问题。	分级错峰：高峰时段让低优先级任务延迟启动。
缺少问题紧急预案，运维人员无法应对。	在任务正式运行前，进行充分的测试。

4 实现窃电用户自动识别教程

4.1 窃电用户自动识别概述

本教程将为您介绍如何通过DataWorks配合机器学习的方式，实现窃电用户的自动识别，保障用户的安全用电。

传统的识别窃电或计量装置故障的方法包括定期巡检、定期校验电表、用户举报窃电等，对人的依赖性较强，且查找窃电漏电的目标不明确。

目前，很多供电局的营销稽查人员、用电检查人员和计量工作人员主要利用计量异常报警功能和电能量数据查询功能，开展用户用电情况的在线监控工作。通过采集电量异常、负荷异常、终端报警、主站报警和线损异常等信息，建立数据分析模型，来实时监测窃漏电情况和发现计量装置的故障。根据报警事件发生前后，客户计量点有关的电流、电压和负荷数据情况等数据情况，构建基于指标的用电异常分析模型，实现检查客户是否存在窃电、违章用电，及计量装置故障等情况。

上述防窃电漏电的查询方法，虽然可以获得用电异常的某些信息，但由于终端误报或漏报过多，无法真正快速精确地定位窃电漏电用户。并且采用上述方法建模时，需要专家根据其知识和经验，来判断模型各输入指标权重，主观性较强。

现有的电力计量自动化系统能够采集到各项电流、电压、功率因数等用电负荷数据，以及用电异常等终端报警信息。异常告警信息和用电负荷数据能够反映用户的用电情况，同时稽查工作人员也会通过在线稽查系统和现场稽查，查找窃电漏电用户，并录入系统。

通过上述数据信息，提取出窃电漏电用户的关键特征，构建窃漏电用户的识别模型，便可自动检查、判断用户是否存在窃电漏电行为，降低稽查工作人员的工作量，并保障用户的正常、安全用电。

窃电用户自动识别教程涉及的具体开发流程如下：

1. #unique_80
2. 数据准备
3. 数据加工
4. 数据建模

4.2 环境准备

为保证您可以顺利完成本次实验，请您首先确保自己云账号已开通大数据计算服务MaxCompute、数据工场DataWorks和机器学习PAI。

前提条件

- 阿里云账号注册，详情请参见#unique_13。
- 实名认证，详情请参见#unique_14或#unique_15。

背景信息

本次实验涉及的阿里云产品如下：

- 大数据计算服务MaxCompute
- 数据工场DataWorks
- 机器学习PAI

开通大数据计算服务MaxCompute

> 说明:
> 如果您已经开通MaxCompute，请跳过此步骤，直接创建DataWorks工作空间。

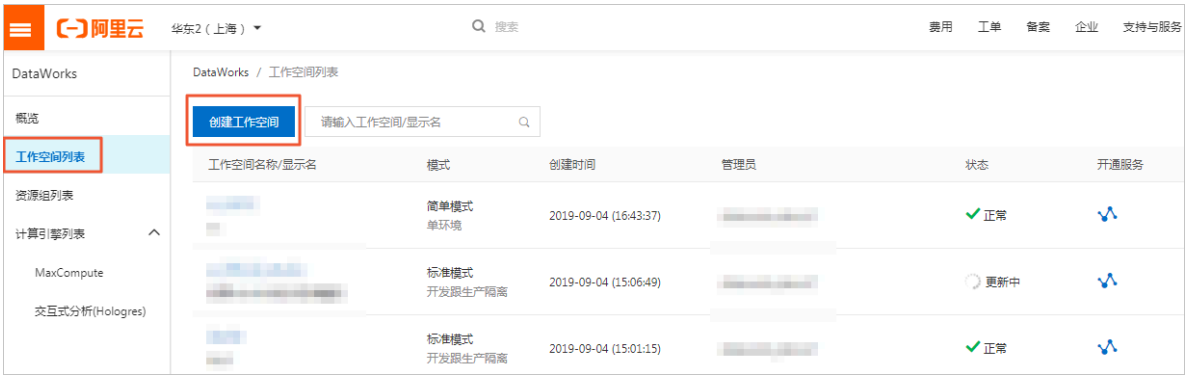
1. 登录阿里云官网，单击右上角的登录，填写您的阿里云账号和密码。
2. 选择产品分类 > 大数据 > 大数据计算 > MaxCompute，进入MaxCompute产品详情页。
3. 单击立即购买。
4. 进入按量付费页面，选择区域和规格类型，单击立即购买。

创建工作空间

1. 使用主账号登录DataWorks控制台。

2. 单击控制台概览 > 常用功能下的创建工作空间。

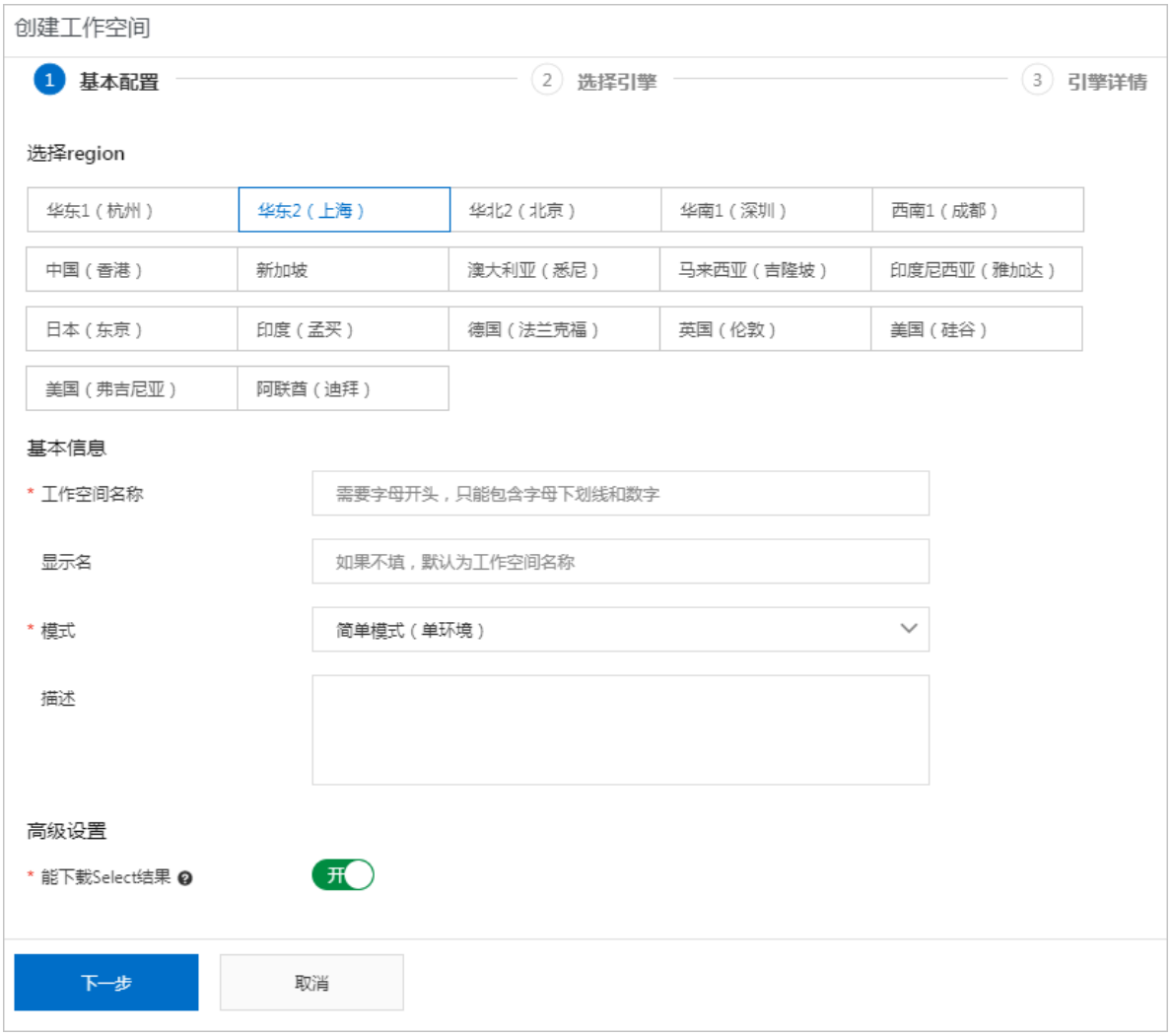
您也可以进入工作空间列表页面，单击创建工作空间。



说明：

从工作空间列表页面创建工作空间时，需提前选择区域，在创建工作空间对话框中不会显示选择region。

3. 填写创建工作空间对话框中的基本配置，单击下一步。



分类	配置	说明
选择region	所有支持DataWorks的区域。	您可以选择与MaxCompute服务一致的区域。
基本信息	工作空间名称	工作空间名称的长度需要在3到27个字符，以字母开头，且只能包含字母下划线和数字。
	显示名	显示名不能超过27个字符，只能字母、中文开头，仅包含中文、字母、下划线和数字。

分类	配置	说明
	模式	工作空间模式是DataWorks新版推出的新功能，分为简单模式和标准模式，双项目开发模式的区别请参见#unique_85。 · 简单模式：指一个Dataworks工作空间对应一个MaxCompute项目，无法设置开发和生产环境，只能进行简单的数据开发，无法对数据开发流程以及表权限进行强控制。 · 标准模式：指一个Dataworks工作空间对应两个MaxCompute项目，可以设置开发和生产双环境，提升代码开发规范，并能够对表权限进行严格控制，禁止随意操作生产环境的表，保证生产表的数据安全。
	描述	对创建的工作空间进行简单描述。
高级设置	能下载select结果	控制数据开发中查询的数据结果是否能够下载，如果关闭无法下载select的数据查询结果。

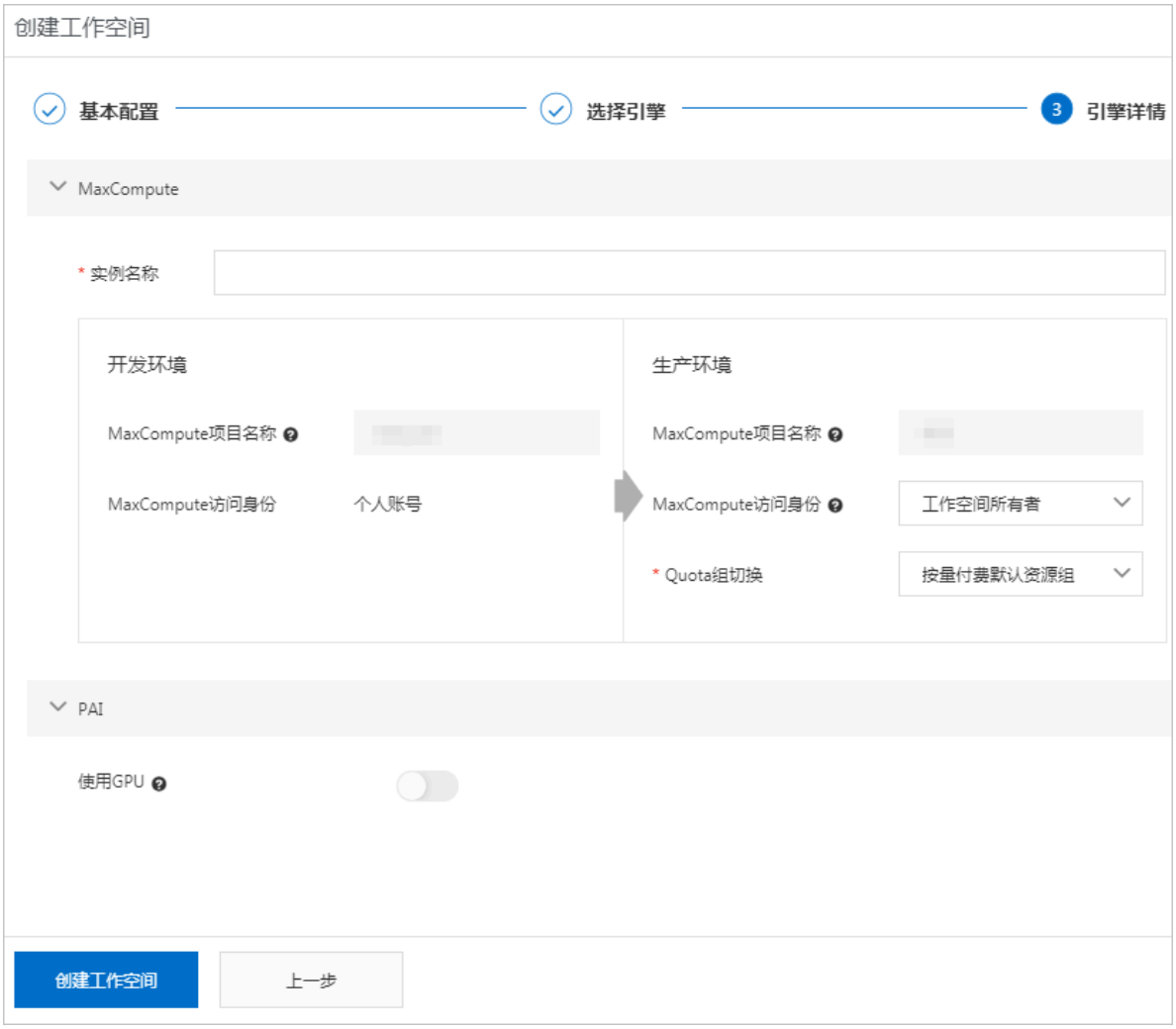
4. 进入选择引擎界面，选择相应引擎后，单击下一步。

   DataWorks已正式商用，如果该区域没有开通，需要首先开通正式商用的服务。默认选中数据集成、数据开发、运维中心和数据质量。

说明:

此处需要勾选机器学习PAI，以开通机器学习。

5. 进入引擎详情页面，填写选购引擎的配置。



分类	配置	说明
MaxCompute	实例名称	实例名称不能超过27个字符，仅支持字母、中文开头，仅包含中文、字母、下划线和数字。
	MaxCompute项目名称	默认与DataWorks工作空间的名称一致。
	MaxCompute访问身份	包括个人账号和工作空间所有者，开发环境默认为个人账号，生产环境推荐使用工作空间所有者。
	Quota组切换	Quota用来实现计算资源和磁盘配额。
PAI	使用GPU	默认不使用，如果需要使用，请前往工作空间配置页面开启GPU使用。

6. 配置完成后，单击创建工作空间。

工作空间创建成功后，即可在工作空间列表页面查看相应内容。

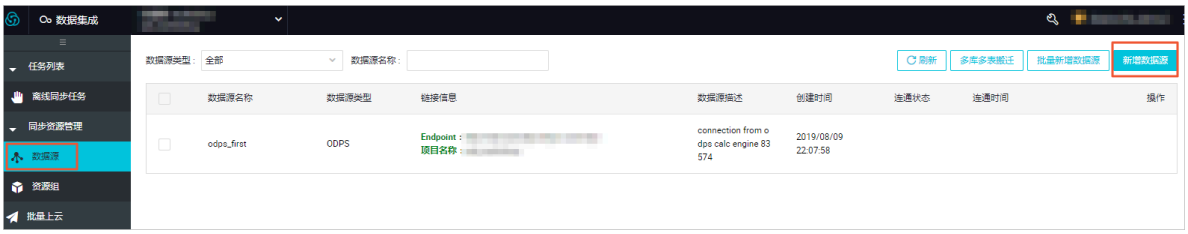
4.3 数据准备

在数据准备阶段，您需要同步原始数据至MaxCompute。

新建数据源

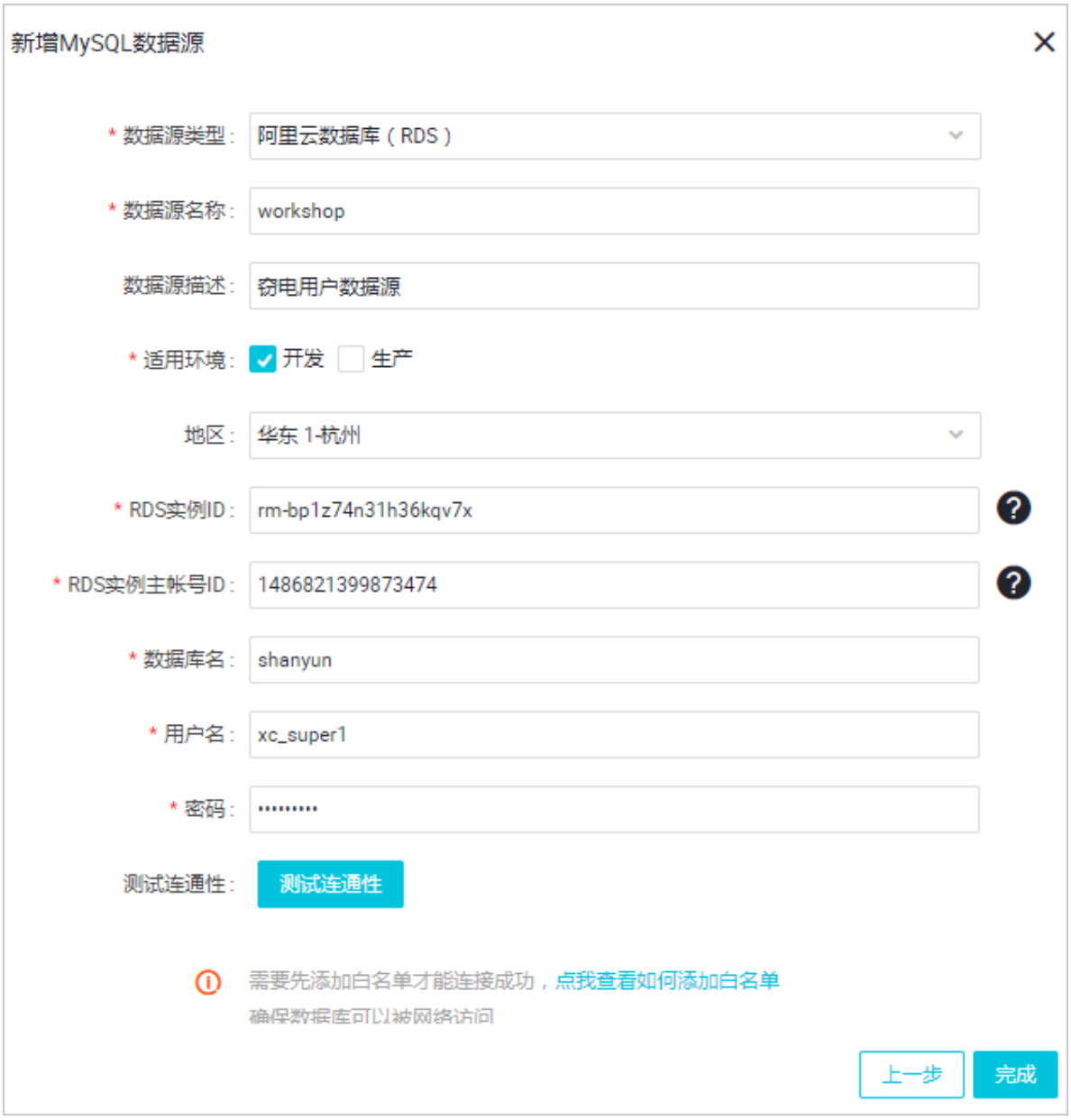
> **说明:**
> 本次实验需要创建MySQL数据源。

1. 单击相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 > 阿里云数据库（RDS）对话框中的配置。



您可以使用自己的数据源，也可以根据下表中的配置进行填写。

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。此处名称填写为workshop。
数据源描述	对数据源进行简单描述，不得超过80个字符。

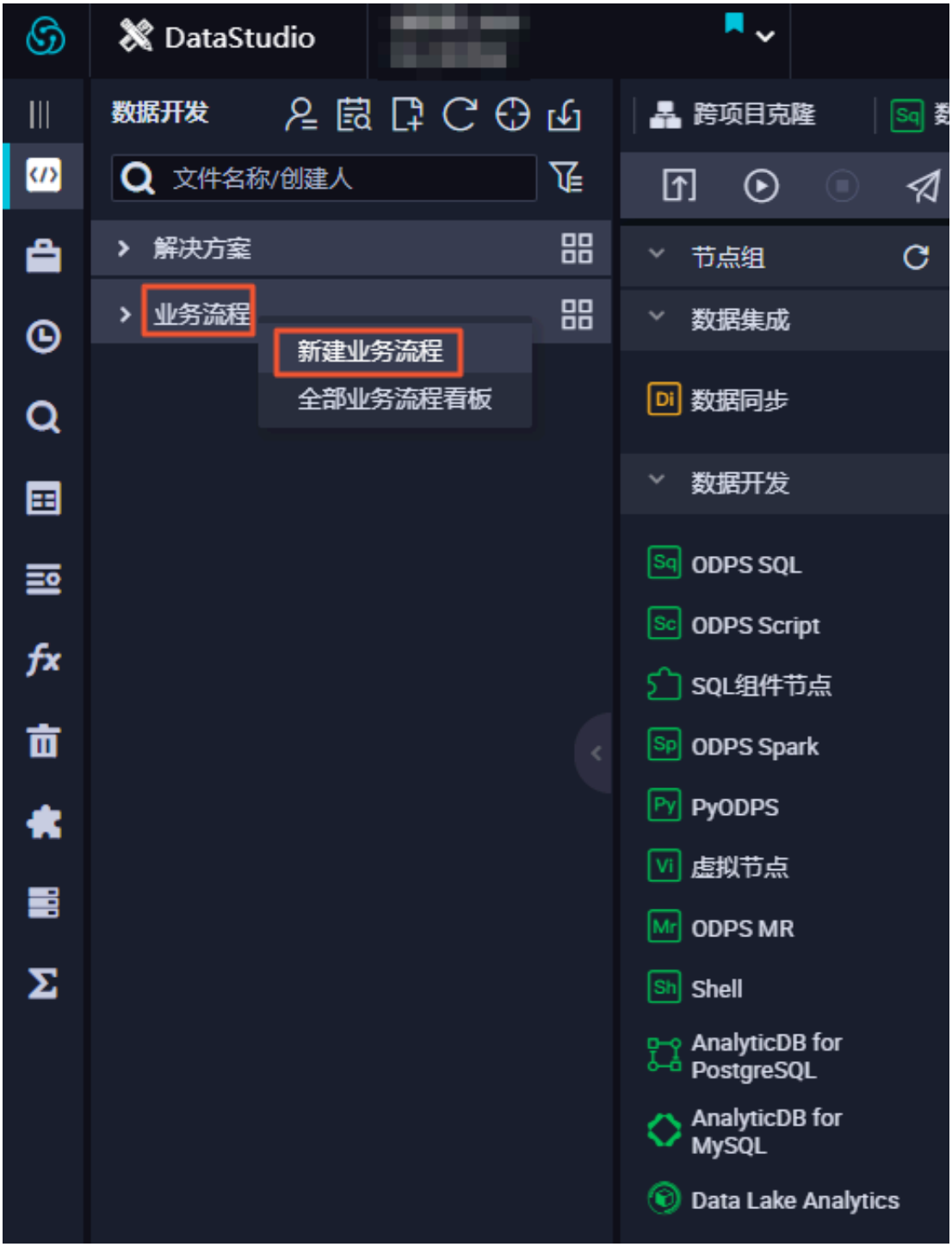
配置	说明
适用环境	可以选择开发或生产环境。 说明: 仅标准模式工作空间会显示此配置。
地区	选择华东1-杭州。
RDS实例ID	填写rm-bp1z74n31h36kqv7x。
RDS实例主账号ID	填写1486821399873474。
数据库名	填写shanyun。
用户名	填写xc_super1。
用户名/密码	填写xc_super1。

5. 单击测试连通性。
6. 测试连通性通过后，单击完成。

新建业务流程

1. 单击左上角的图标，选择全部产品 > DataStudio（数据开发）。

2. 右键单击业务流程，选择新建业务流程。

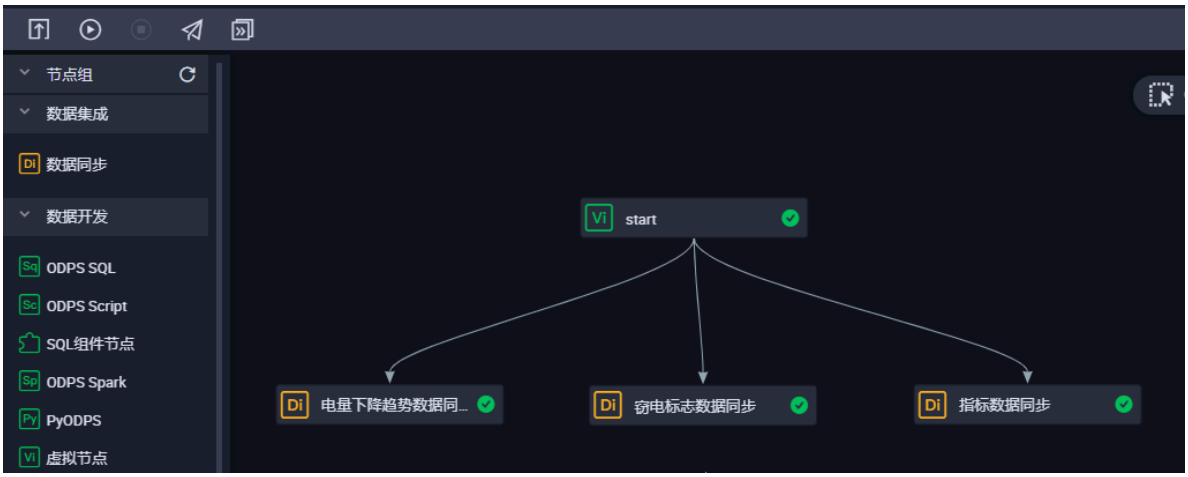


3. 在新建业务流程对话框中，填写业务流程名称和描述。

4. 单击新建，即可完成业务流程的创建。

5. 进入业务流程开发面板，并向面板中拖入一个虚拟节点（start）和3个数据同步节点（电量下降趋势数据同步、窃电标志数据同步和指标数据同步）分别填写相应的配置后，单击提交。

6. 拖拽连线将start节点设置为3个数据同步节点的上游节点。

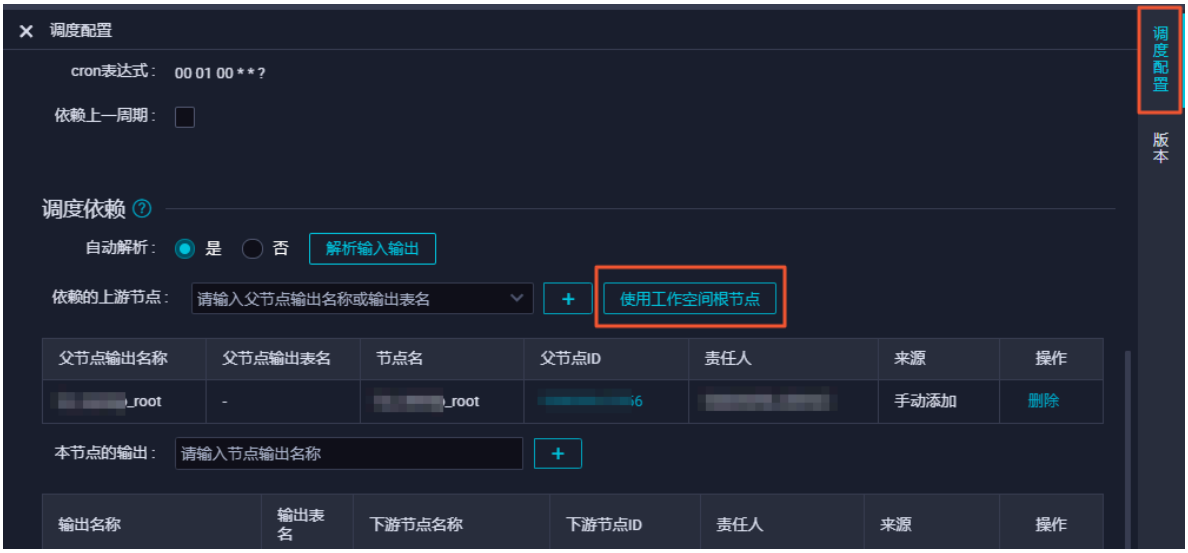


配置start节点

1. 双击虚拟节点，单击右侧的调度配置。

2. 设置start节点的上游节点为工作空间根节点。

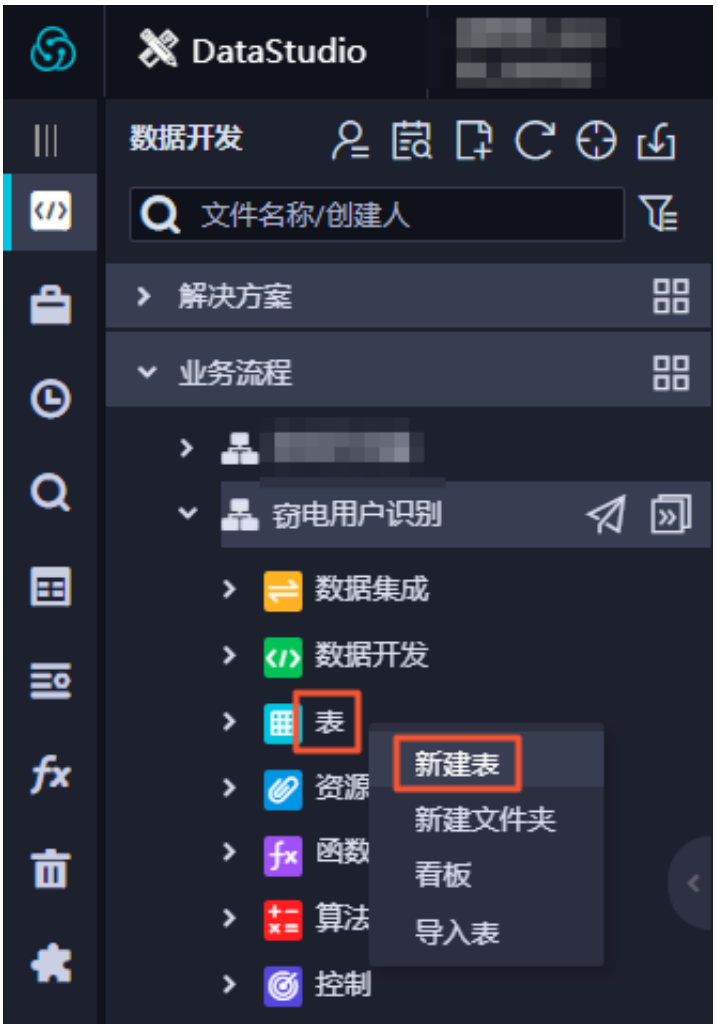
   由于新版本给每个节点都设置了输入输出节点，所以需要给start节点设置一个输入。此处设置其上游节点为工作空间根节点，通常命名为工作空间名_root。



3. 配置完成后，单击左上角的  进行保存。

新建表

1. 右键单击业务流程下的表，选择新建表。



2. 在新建表对话框中填写表名，单击提交。

   此处需要创建3张表，分别存储同步过来的电量下降趋势数据、指标数据和窃电标志数据（trend_data、indicators_data和steal_flag_data）。

3. 打开创建的表，单击DDL模式，分别填写以下相应的建表语句。

```
--电量下降趋势表
CREATE TABLE `trend_data` (
    `uid` bigint,
    `trend` bigint
)
PARTITIONED BY (dt string);
```

```
--指标数据
CREATE TABLE `indicators_data` (
    `uid` bigint,
    `xiansun` bigint,
    `warnindicator` bigint
)
COMMENT '*'
```

```
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

```
--窃电标志数据
CREATE TABLE `steal_flag_data` (
    `uid` bigint,
    `flag` bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

4. 建表语句输入完成后，单击生成表结构并确认覆盖当前操作。
5. 返回建表页面后，在基本属性中输入表的中文名。
6. 完成设置后，分别单击提交到开发环境和提交到生产环境。

配置数据同步节点

配置电量下降趋势数据同步节点。

a) 双击电量下降趋势数据同步节点，进入节点配置页面。

b) 选择数据来源。

配置	说明
数据源	选择MySQL > workshop。
表	选择MySQL数据源中的表trending。
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，此处可以不填。

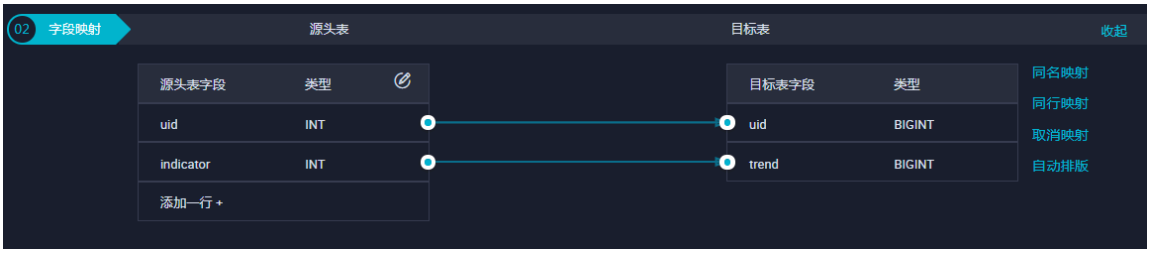
配置	说明
切分键	读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。此处可以不填。

c) 选择数据去向。

配置	说明
数据源	选择ODPS > odps_first。
表	选择ODPS数据源中的表trend_data。
分区信息	输入要同步的分区列，此处默认为`dt=${bdp.system.bizdate}`。
清理规则	选择写入前清理已有数据。

配置	说明
空字符串是否作null	选择null。

d) 配置字段映射。



e) 配置通道控制。

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_86和#unique_87。

f) 确认当前节点的配置情况，可以进行修改。确认无误后，单击左上角的保存。

提交业务流程

1. 打开业务流程配置面板，单击左上角的提交。

2. 选择提交对话框中需要提交的节点，填写备注，勾选忽略输入输出不一致的告警。

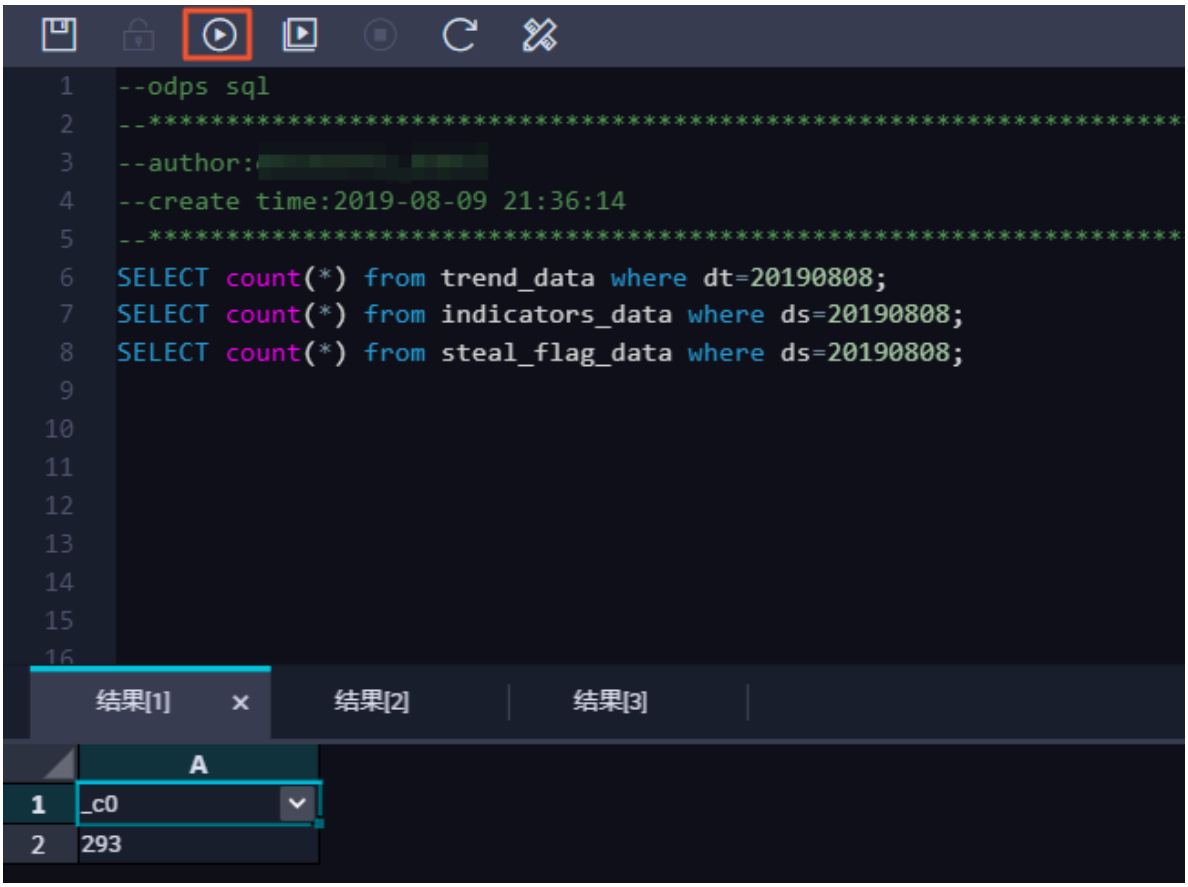
3. 单击提交，待显示提交成功即可。

确认数据是否成功导入MaxCompute

1. 单击左侧导航栏中的临时查询，进入临时查询面板。

2. 右键单击临时查询，选择新建节点 > ODPS SQL。

3. 编写并执行SQL语句，查看导入表trend_data、indicators_data和steal_flag_data的记录数。



说明:

SQL语句如下所示，其中分区列需要更新为业务日期。例如，任务运行的日期为20190809，则业务日期为201900808。

```
--查看是否成功写入MaxCompute
SELECT count(*) from trend_data where dt=业务日期;
SELECT count(*) from indicators_data where ds=业务日期;
SELECT count(*) from steal_flag_data where ds=业务日期;
```

后续步骤

现在，您已经学习了如何进行数据同步，完成数据的采集，您可以继续学习下一个教程。在该教程中您将学习如何对采集的数据进行计算与分析。详情请参见。

4.4 数据加工

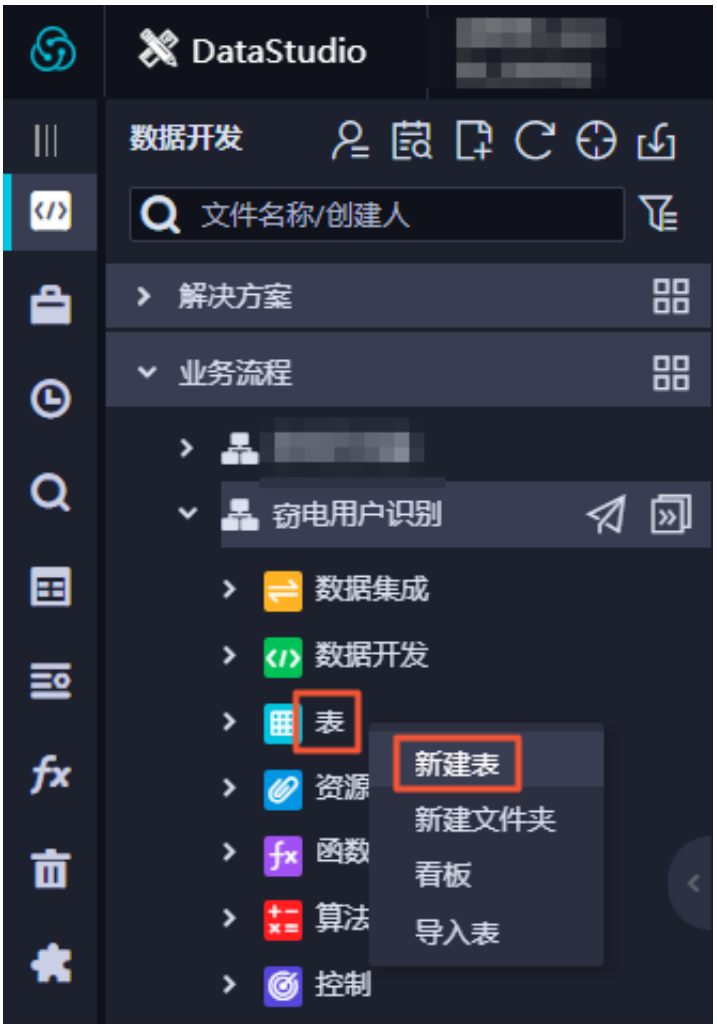
本文将为您介绍如何通过DataWorks，将已经采集至MaxCompute的数据进行加工，获取清洗后的数据。

前提条件

开始本文的操作前，请首先完成数据准备中的操作。

新建表

1. 右键单击业务流程下的表，选择新建表。



2. 在新建表对话框中填写表名，单击提交。

   此处需要创建的数据表，如下所示：

   - 创建3张表，分别存储同步过来的电量下降趋势数据、指标数据和窃电标志数据清洗之后的数据（clean_trend_data、clean_indicators_data和clean_steal_flag_data）。
   - 创建1张表，存储汇聚后的数据（data4ml）。

3. 打开创建的表，单击DDL模式，分别填写以下相应的建表语句。

```
--清洗后的电量下降趋势数据
CREATE TABLE `clean_trend_data` (
    `uid` bigint,
    `trend` bigint
)
PARTITIONED BY (dt string)
```

```
LIFECYCLE 7;
```

```
--清洗后的指标数据
CREATE TABLE `clean_indicators_data` (
    `uid` bigint,
    `xiansun` bigint,
    `warnindicator` bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

```
--清洗后的窃电标志数据
CREATE TABLE `clean_steal_flag_data` (
    `uid` bigint,
    `flag` bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

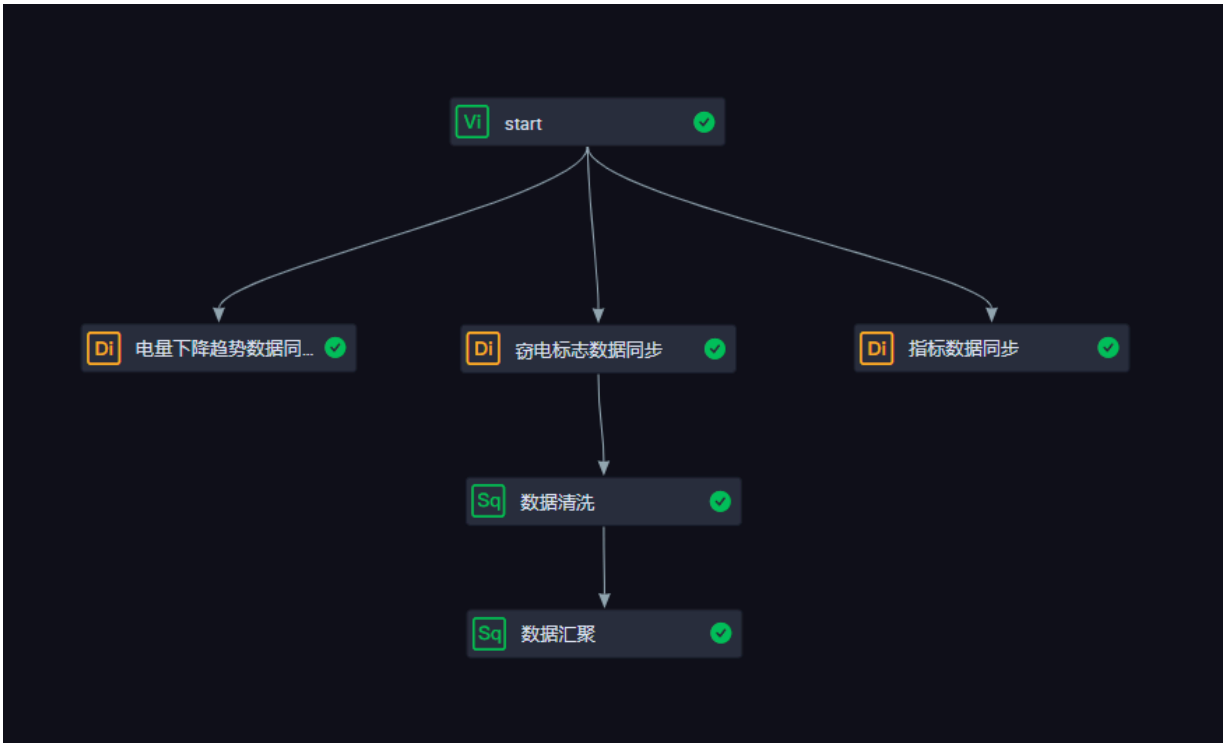
```
--汇聚后的数据
CREATE TABLE `data4ml` (
    `uid` bigint,
    `trend` bigint,
    `xiansun` bigint,
    `warnindicator` bigint,
    `flag` bigint
)
COMMENT '*'
PARTITIONED BY (ds string)
LIFECYCLE 36000;
```

4. 建表语句输入完成后，单击生成表结构并确认覆盖当前操作。
5. 返回建表页面后，在基本属性中输入表的中文名。
6. 完成设置后，分别单击提交到开发环境和提交到生产环境。

设计业务流程

业务流程的新建及依赖关系的配置请参见新建业务流程。

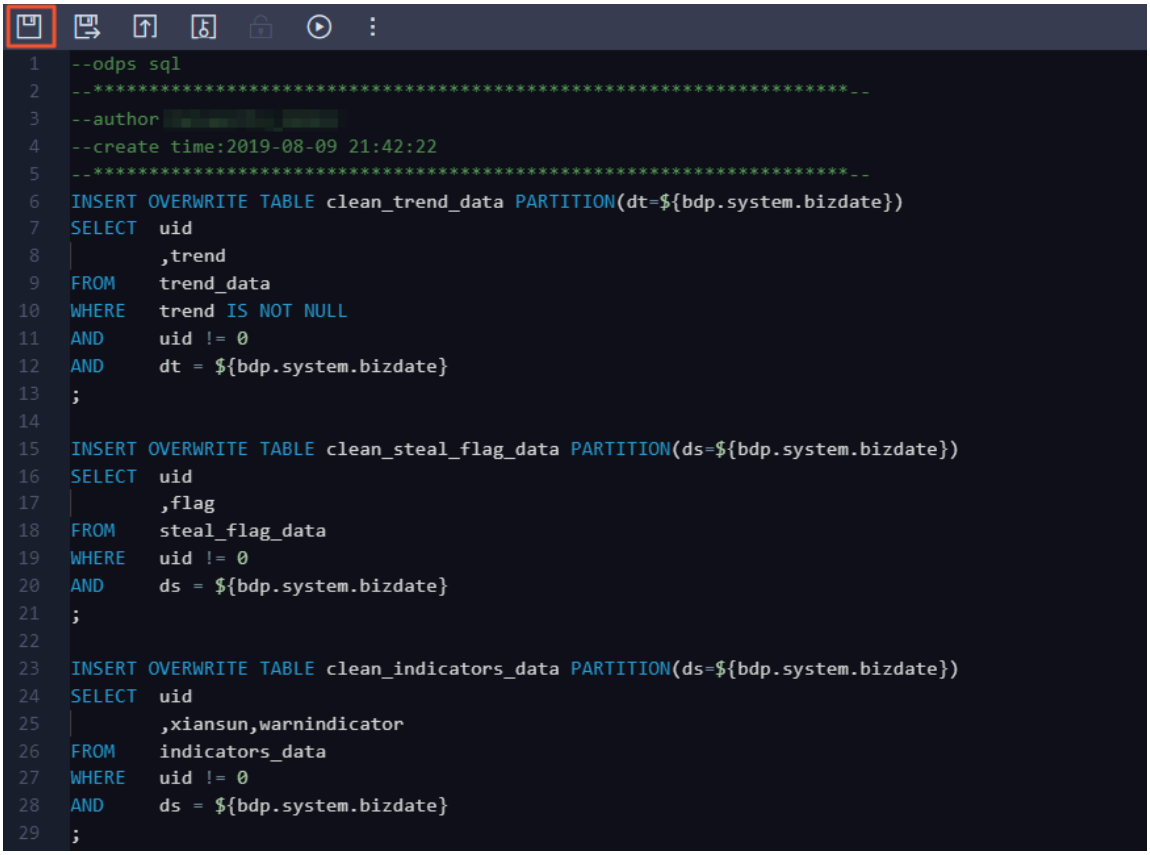
进入业务流程开发面板，并向面板中拖入2个ODPS SQL节点，依次命名为数据清洗和数据汇聚，并配置如下图所示的依赖关系。



配置ODPS SQL节点

- 配置数据清洗节点。

  1. 双击数据清洗节点，进入节点配置页面。
  2. 编写处理逻辑。



```
--odps sql
--********************************************************************--
--author
--create time:2019-08-09 21:42:22
--********************************************************************--
INSERT OVERWRITE TABLE clean_trend_data PARTITION(dt=${bdp.system.bizdate})
SELECT  uid
        ,trend
FROM    trend_data
WHERE   trend IS NOT NULL
AND     uid != 0
AND     dt = ${bdp.system.bizdate}
;

INSERT OVERWRITE TABLE clean_steal_flag_data PARTITION(ds=${bdp.system.bizdate})
SELECT  uid
        ,flag
FROM    steal_flag_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;

INSERT OVERWRITE TABLE clean_indicators_data PARTITION(ds=${bdp.system.bizdate})
SELECT  uid
        ,xiansun,warnindicator
FROM    indicators_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;
```

SQL逻辑如下所示：

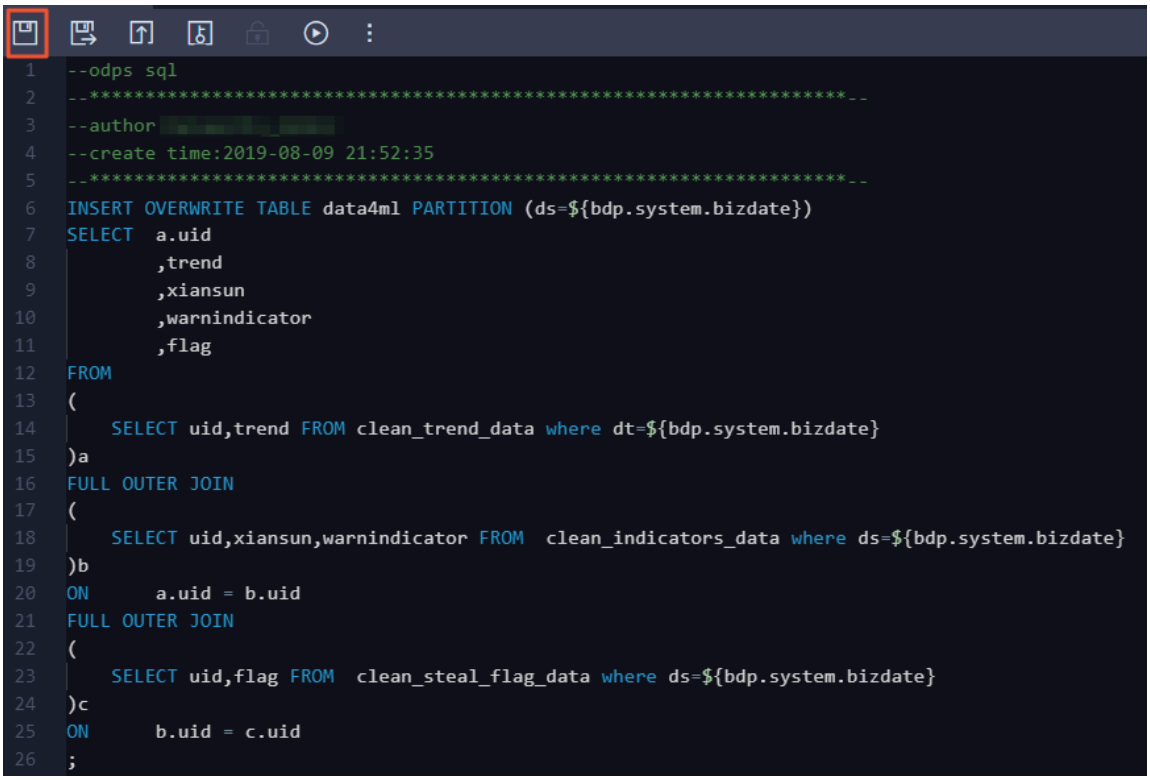
```
INSERT OVERWRITE TABLE clean_trend_data PARTITION(dt=${bdp.system.
bizdate})
SELECT  uid
        ,trend
FROM    trend_data
WHERE   trend IS NOT NULL
AND     uid != 0
AND     dt = ${bdp.system.bizdate}
;

INSERT OVERWRITE TABLE clean_steal_flag_data PARTITION(ds=${bdp.
system.bizdate})
SELECT  uid
        ,flag
FROM    steal_flag_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;

INSERT OVERWRITE TABLE clean_indicators_data PARTITION(ds=${bdp.
system.bizdate})
SELECT  uid
```

```
        ,xiansun,warnindicator
FROM    indicators_data
WHERE   uid != 0
AND     ds = ${bdp.system.bizdate}
;
```

3. 单击左上角的保存按钮。

- 配置数据汇聚节点。

1. 双击数据汇聚节点，进入节点配置页面。
2. 编写处理逻辑。



```
--odps sql
--********************************************************************--
--author
--create time:2019-08-09 21:52:35
--********************************************************************--
INSERT OVERWRITE TABLE data4ml PARTITION (ds=${bdp.system.bizdate})
SELECT  a.uid
        ,trend
        ,xiansun
        ,warnindicator
        ,flag
FROM
(
    SELECT uid,trend FROM clean_trend_data where dt=${bdp.system.bizdate}
)a
FULL OUTER JOIN
(
    SELECT uid,xiansun,warnindicator FROM  clean_indicators_data where ds=${bdp.system.bizdate}
)b
ON      a.uid = b.uid
FULL OUTER JOIN
(
    SELECT uid,flag FROM  clean_steal_flag_data where ds=${bdp.system.bizdate}
)c
ON      b.uid = c.uid
;
```

SQL逻辑如下所示：

```
INSERT OVERWRITE TABLE data4ml PARTITION (ds=${bdp.system.bizdate
})
SELECT  a.uid
        ,trend
        ,xiansun
        ,warnindicator
        ,flag
FROM
(
    SELECT uid,trend FROM clean_trend_data where dt=${bdp.system.
bizdate}
)a
FULL OUTER JOIN
(
    SELECT uid,xiansun,warnindicator FROM  clean_indicators_data
where ds=${bdp.system.bizdate}
)b
ON      a.uid = b.uid
FULL OUTER JOIN
(
```

```
    SELECT uid,flag FROM  clean_steal_flag_data where ds=${bdp.
system.bizdate}
)c
ON      b.uid = c.uid
;
```

3. 单击左上角的保存按钮。

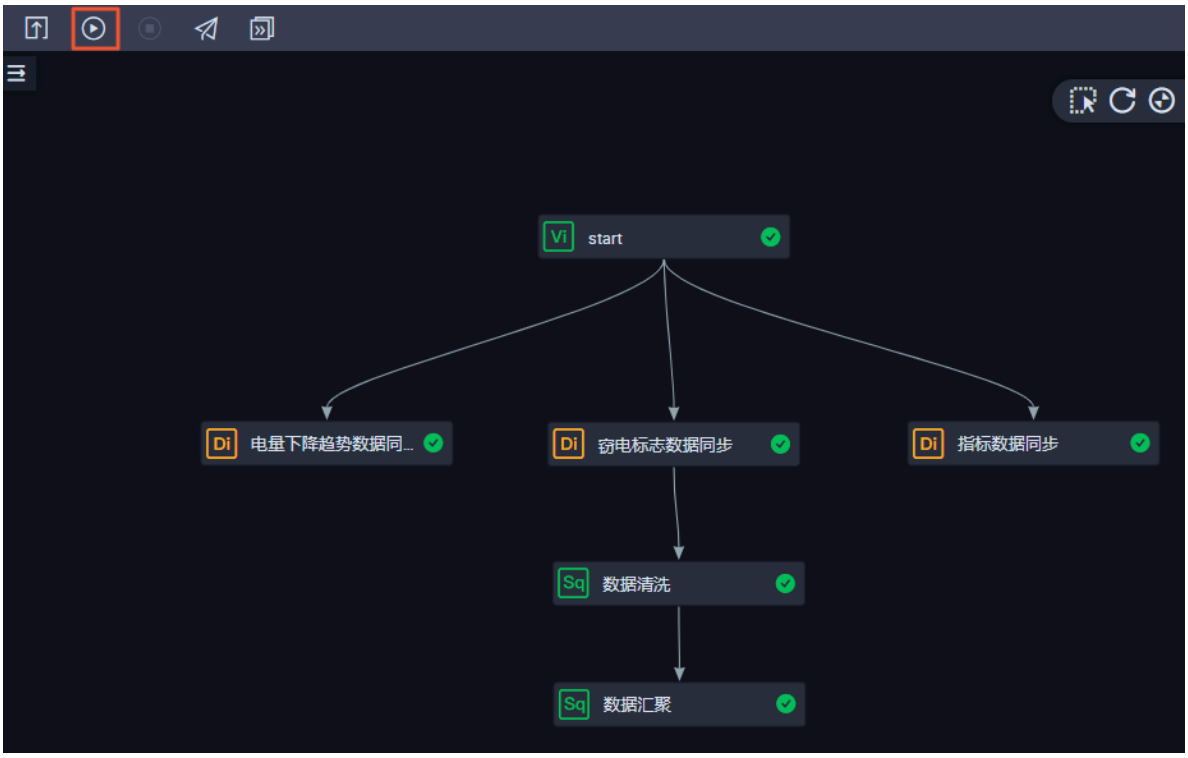
提交业务流程

1. 打开业务流程配置面板，单击左上角的提交。
2. 选择提交对话框中需要提交的节点，填写备注，勾选忽略输入输出不一致的告警。

3. 单击提交，待显示提交成功即可。

运行业务流程

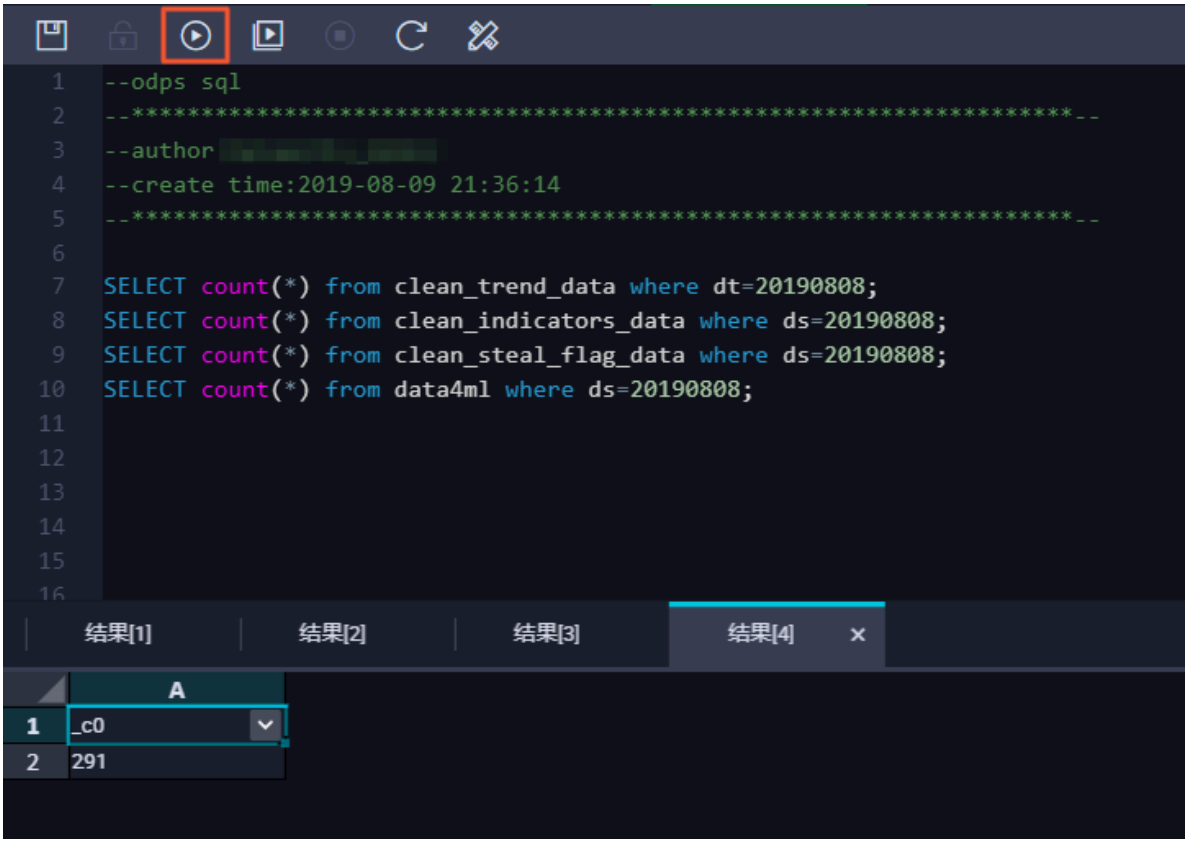
1. 打开业务流程配置面板，单击左上角的运行。



2. 单击左侧导航栏中的临时查询，进入临时查询面板。

3. 右键单击临时查询，选择新建节点 > ODPS SQL。

4. 编写并执行SQL语句，查看导入表clean_trend_data、clean_indicators_data、clean_steal_flag_data和data4ml的记录数。



说明:

SQL语句如下所示，其中分区列需要更新为业务日期。例如，任务运行的日期为20190809，则业务日期为201900808。

```
--查看是否成功写入MaxCompute
SELECT count(*) from clean_trend_data where dt=业务日期;
SELECT count(*) from clean_indicators_data where ds=业务日期;
SELECT count(*) from clean_steal_flag_data where ds=业务日期;
SELECT count(*) from data4ml where ds=业务日期;
```

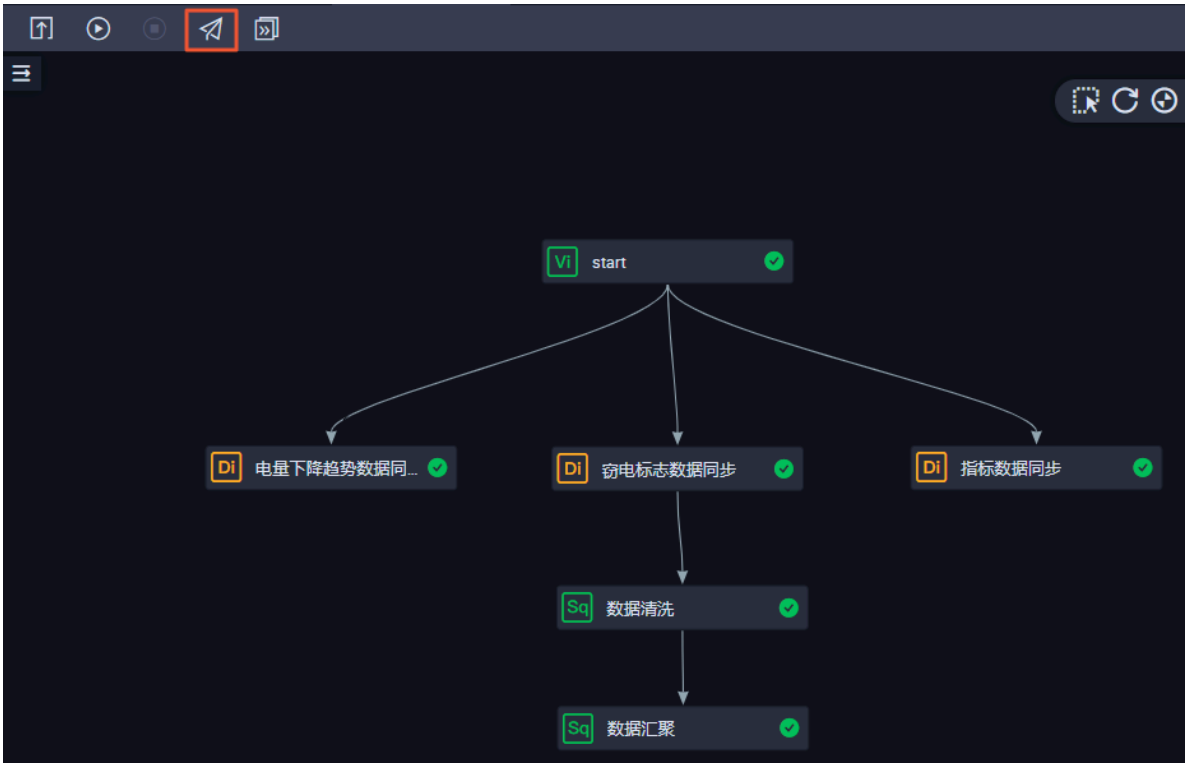
发布业务流程

提交业务流程后，表示任务已进入开发环境。由于开发环境的任务不会自动调度，您需要将配置完成的任务发布至生产环境。

说明:

将任务发布至生产环境前，您需要对代码进行测试，确保其正确性。

1. 打开业务流程配置面板，单击左上角的发布，进入发布页面。



2. 选择待发布任务，单击添加到待发布。

3. 进入右上角的待发布列表，单击全部打包发布。

4. 在发布包列表页面查看已发布内容。

在生产环境运行任务

1. 任务发布成功后，单击右上角的运维中心。

2. 选择周期任务运维 > 周期任务中的相应节点。

3. 右键单击DAG图中的start节点，选择补数据 > 当前节点及下游节点。

4. 勾选需要补数据的任务，输入业务日期，单击确定。

单击确定后，自动跳转至补数据实例页面。

5. 单击刷新，直至SQL任务都运行成功即可。

后续步骤

现在，您已经学习了如何创建SQL任务、如何处理原始数据。您可以继续学习下一个教程，学习如何通过机器学习，载入处理好的数据并构建窃漏电用户的识别模型。详情请参见数据建模。

4.5 数据建模

本文将为您介绍如何载入DataWorks中处理好的数据，通过机器学习构建窃漏电用户的识别模型。

前提条件

开始本文的操作前，请首先完成数据加工中的操作。

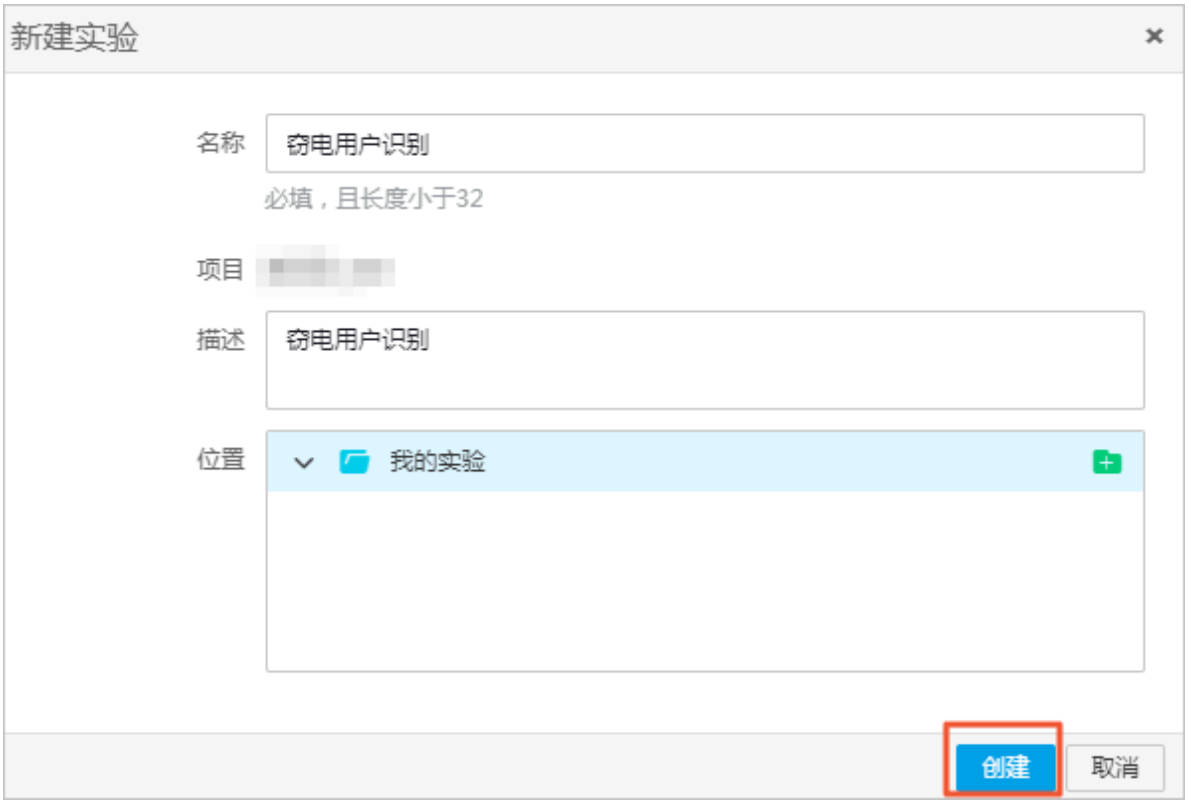
新建实验

1. 进入机器学习控制台，单击左侧导航栏中的Studio-可视化建模。

2. 单击相应工作空间后的进入机器学习。

3. 单击左侧菜单栏中的实验，右键单击我的实验，选择新建空白实验。

4. 填写新建实验对话框中的名称和描述。

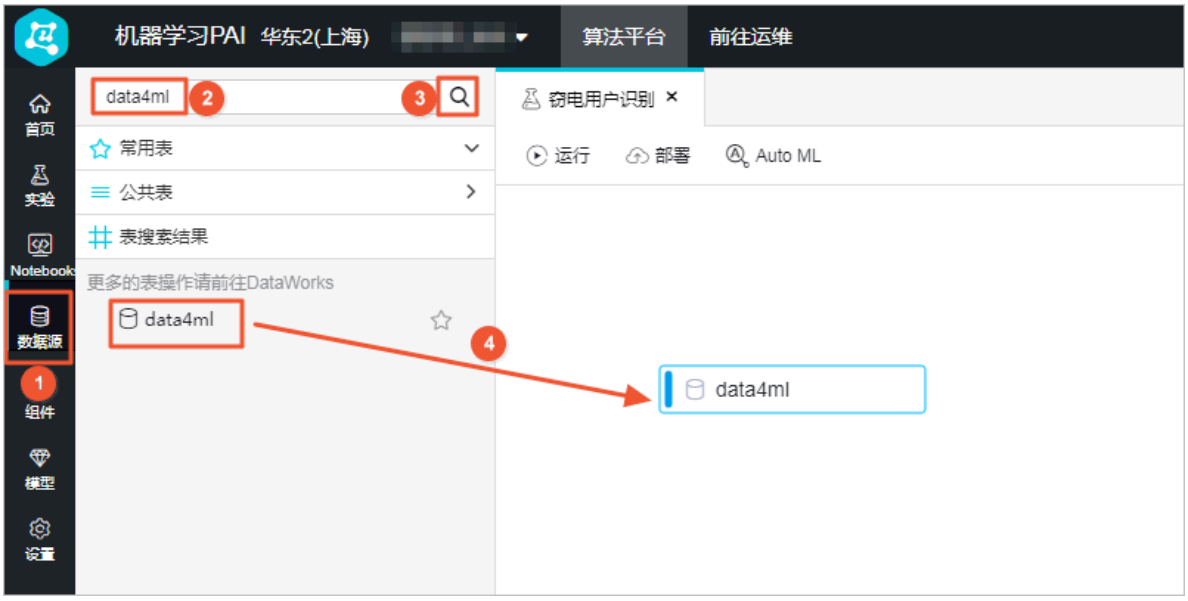


5. 单击创建。

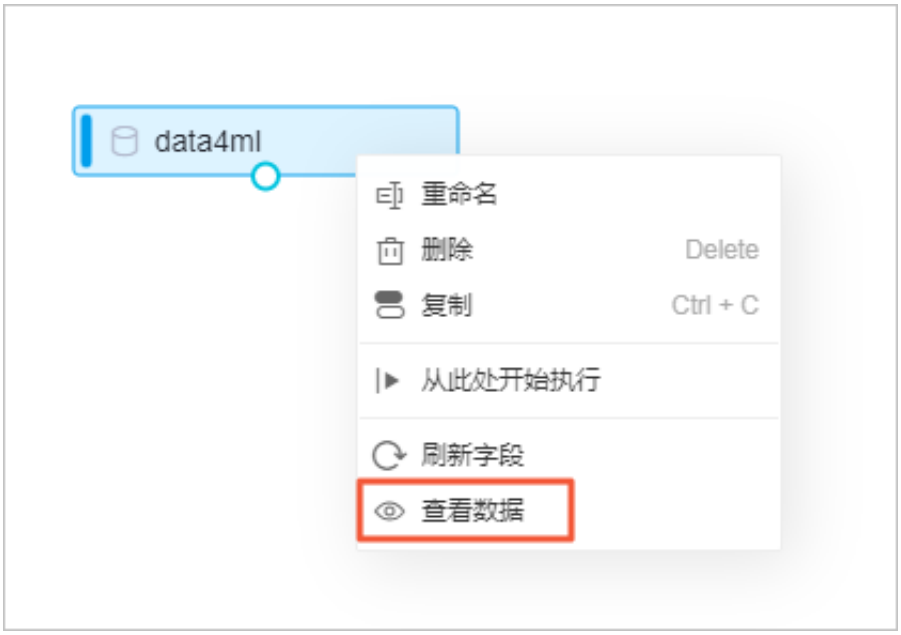
载入数据集

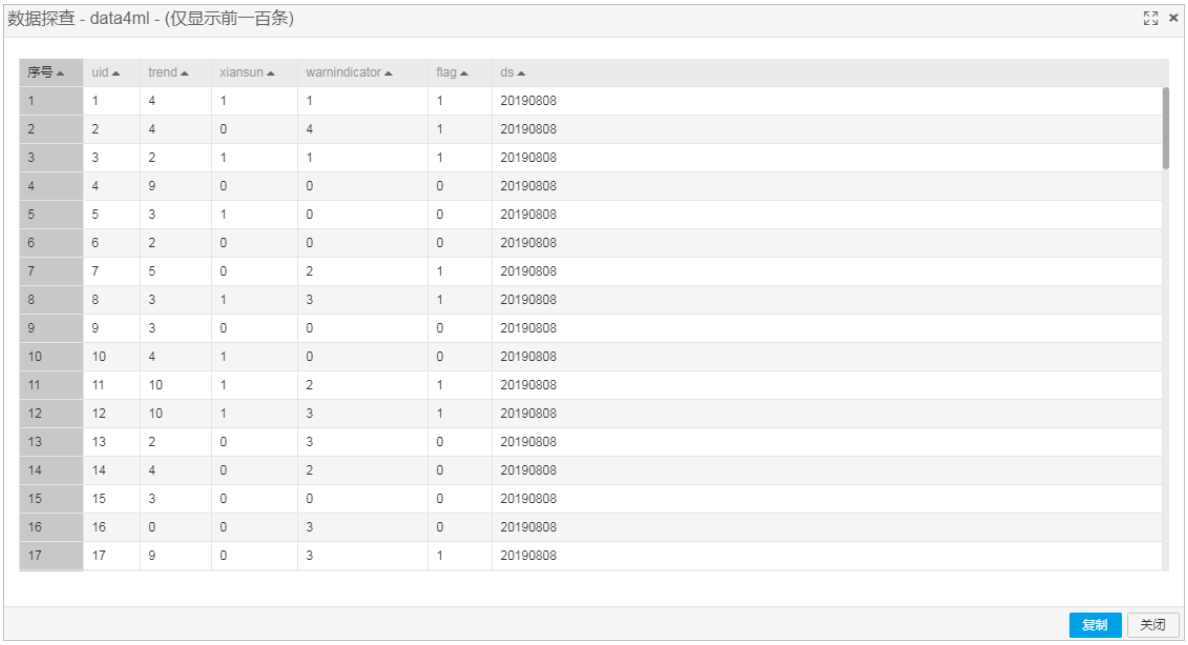
1. 单击左侧导航栏中的数据源。
2. 在搜索框输入数据加工中最终输出的data4ml表，单击搜索图标。

3. 拖拽表搜索结果下的data4ml表至右侧画布。



右键单击读数据表，选择查看数据，即可查看载入的结果数据。数据包括1个用户的电量趋势下降指标、线损指标和告警类指标数量等3个窃电漏电指标，以及用户是否真实窃电漏电的数据。




数据探查 - data4ml - (仅显示前一百条)

序号	uid	trend	xiansun	warnindicator	flag	ds
1	1	4	1	1	1	20190808
2	2	4	0	4	1	20190808
3	3	2	1	1	1	20190808
4	4	9	0	0	0	20190808
5	5	3	1	0	0	20190808
6	6	2	0	0	0	20190808
7	7	5	0	2	1	20190808
8	8	3	1	3	1	20190808
9	9	3	0	0	0	20190808
10	10	4	1	0	0	20190808
11	11	10	1	2	1	20190808
12	12	10	1	3	1	20190808
13	13	2	0	3	0	20190808
14	14	4	0	2	0	20190808
15	15	3	0	0	0	20190808
16	16	0	0	3	0	20190808
17	17	9	0	3	1	20190808

复制 关闭

进行数据探索

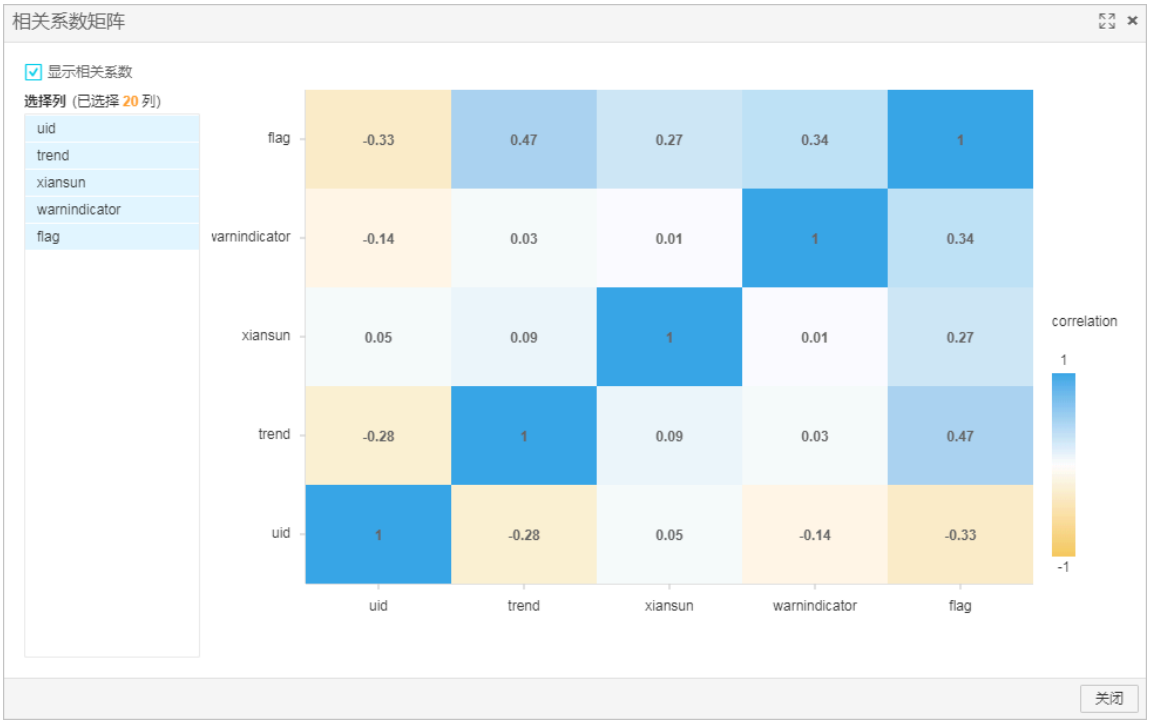
1. 相关性分析

a) 单击左侧导航栏中的组件，拖拽统计分析 > 相关系数矩阵至右侧画布。

b) 连线读数据表中ODPS源的输出和相关系数矩阵的输入。

c) 右键单击相关系数矩阵，选择从此处开始执行。

d) 待运行完成后，右键单击相关系数矩阵，选择查看分析报告。



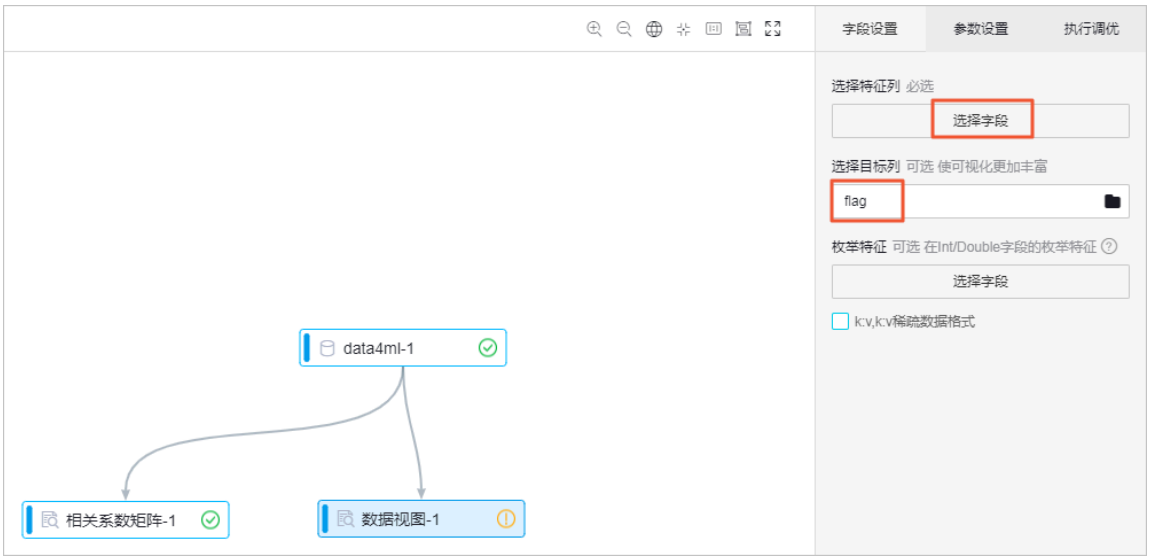
如相关系数矩阵图所示，3个窃电漏电指标对于最终是否为窃电用户的关系都不是特别明显，即决定用户是否为窃电用户的特征并不具有单一性。

2. 特征分析

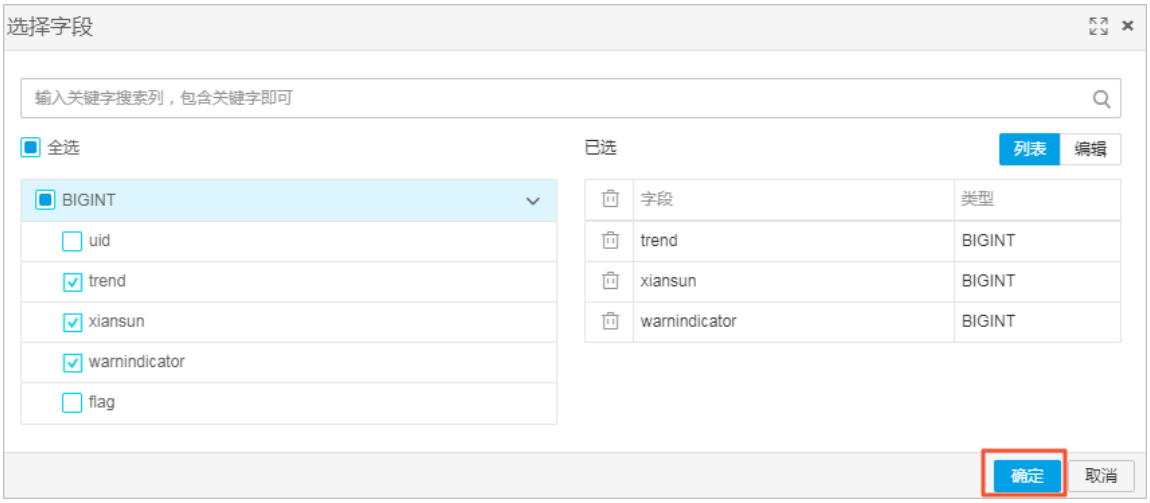
a) 单击左侧导航栏中的组件，拖拽统计分析 > 数据视图至右侧画布。

b) 连线读数据表中ODPS源的输出和数据视图的输入。

c) 双击数据视图，选择右侧的字段设置 > 选择特征列，单击选择字段，并选择目标列为flag。

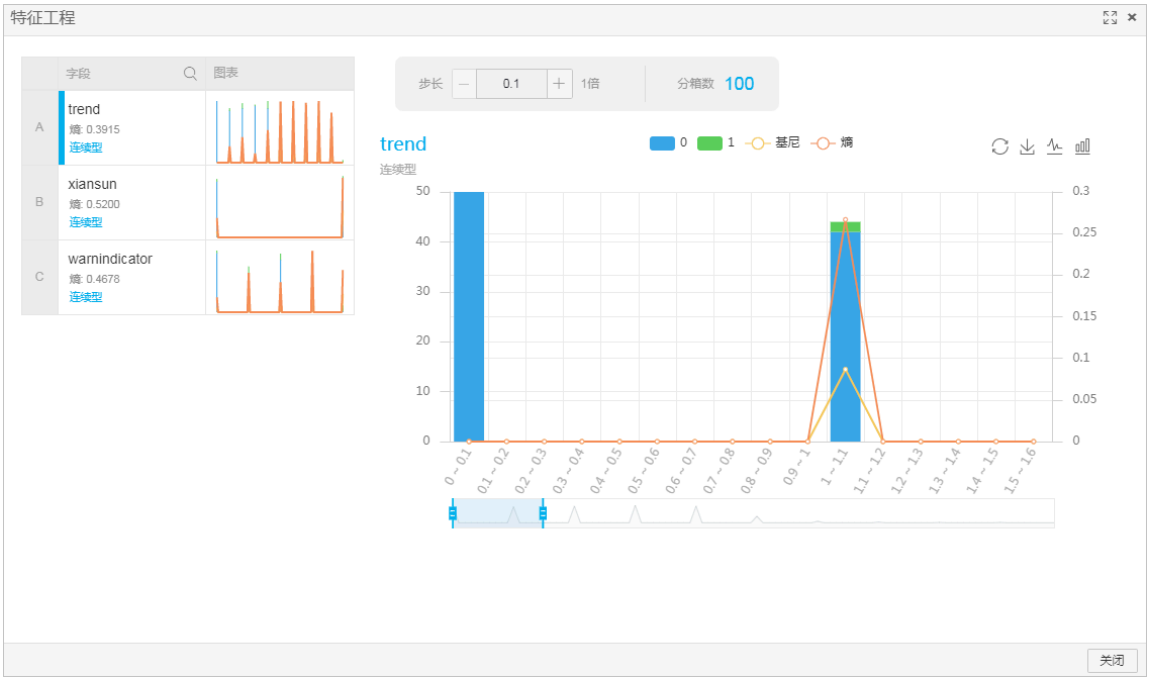


d) 在选择字段对话框中，选择trend、xiansun和warnindicator3个字段，单击确定。



e) 右键单击 数据视图，选择从此处开始执行。

f) 执行完成后，选择查看分析报告，即可查看各个特征和标签列在数据分布上的关系。

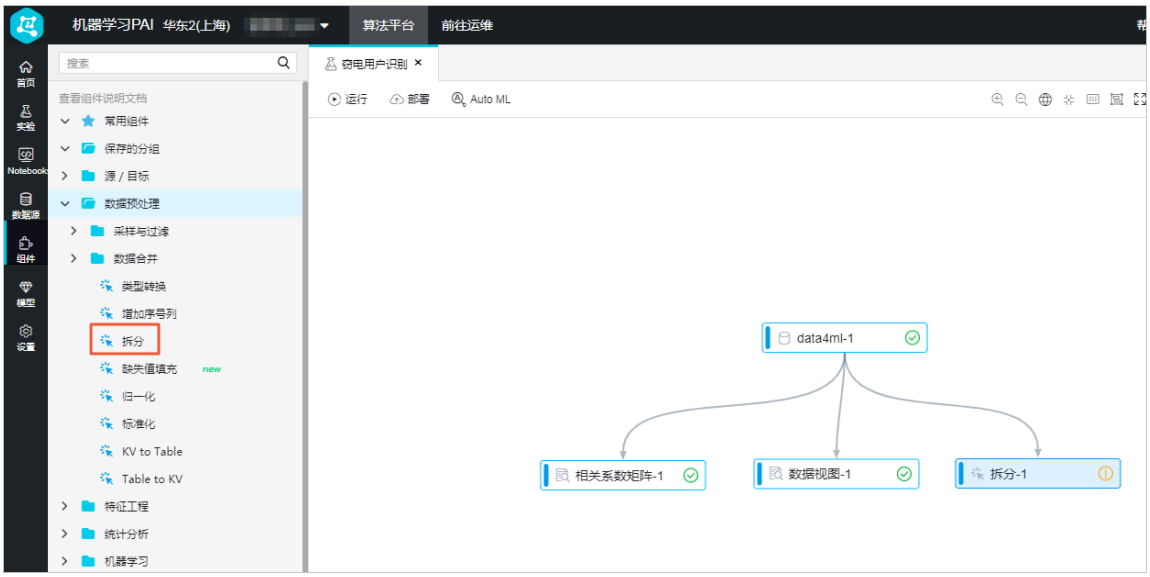


进行数据建模

完成简单的探索性分析之后，即可开始选择合适的算法模型进行数据建模。

1. 通过拆分组件，将数据分为训练集和测试集。

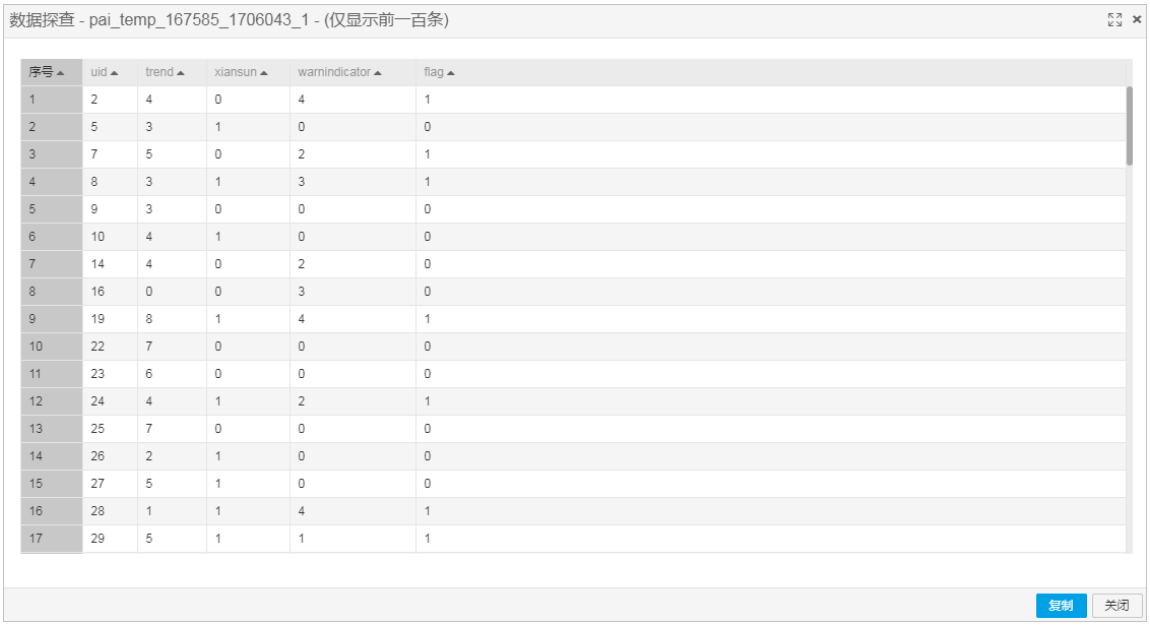
   a) 单击左侧导航栏中的组件，拖拽数据预处理 > 拆分至右侧画布。



   b) 连线读数据表中ODPS源的输出和拆分的输入。

   c) 右键单击拆分，选择从此处开始执行。

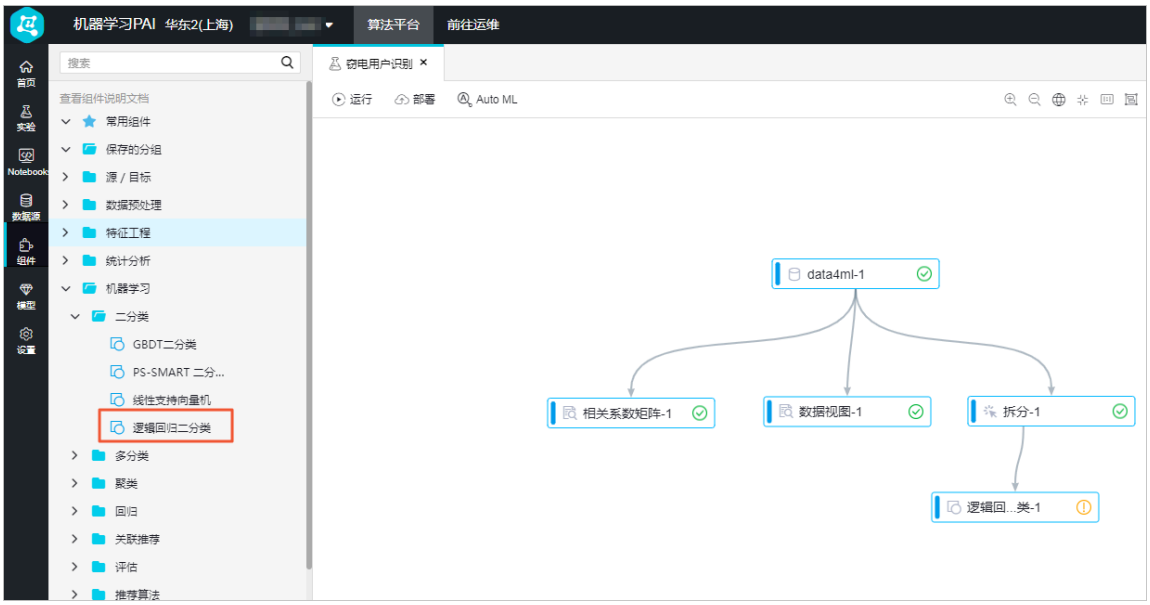
   d) 待运行完成后，右键单击拆分，选择查看数据 > 查看输出桩。



序号	uid	trend	xiansun	warnindicator	flag
1	2	4	0	4	1
2	5	3	1	0	0
3	7	5	0	2	1
4	8	3	1	3	1
5	9	3	0	0	0
6	10	4	1	0	0
7	14	4	0	2	0
8	16	0	0	3	0
9	19	8	1	4	1
10	22	7	0	0	0
11	23	6	0	0	0
12	24	4	1	2	1
13	25	7	0	0	0
14	26	2	1	0	0
15	27	5	1	0	0
16	28	1	1	4	1
17	29	5	1	1	1

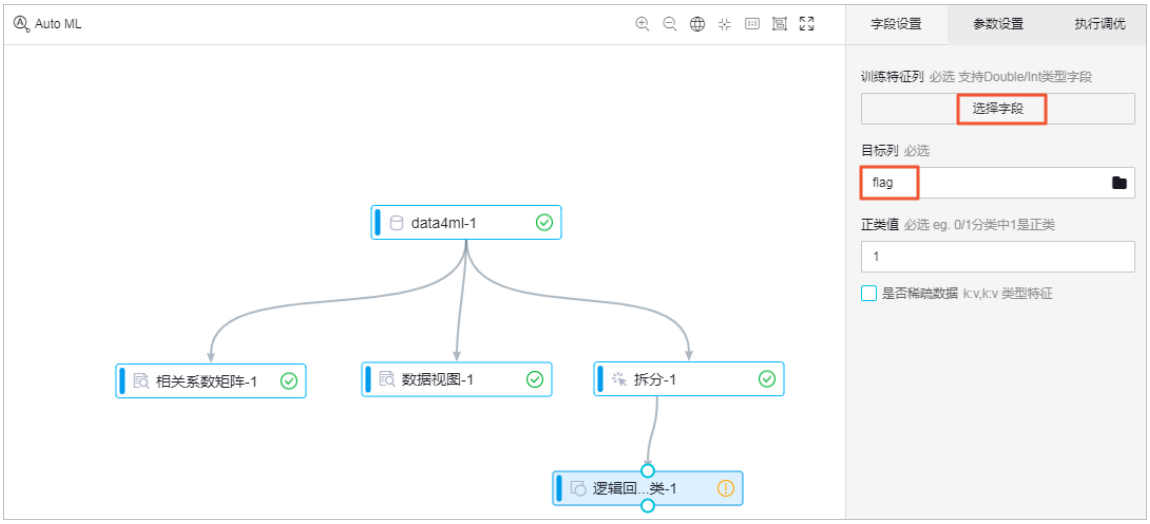
2. 通过逻辑回归二分类组件，对数据进行回归建模。

a) 单击左侧导航栏中的组件，拖拽机器学习 > 二分类 > 逻辑回归二分类至右侧画布。

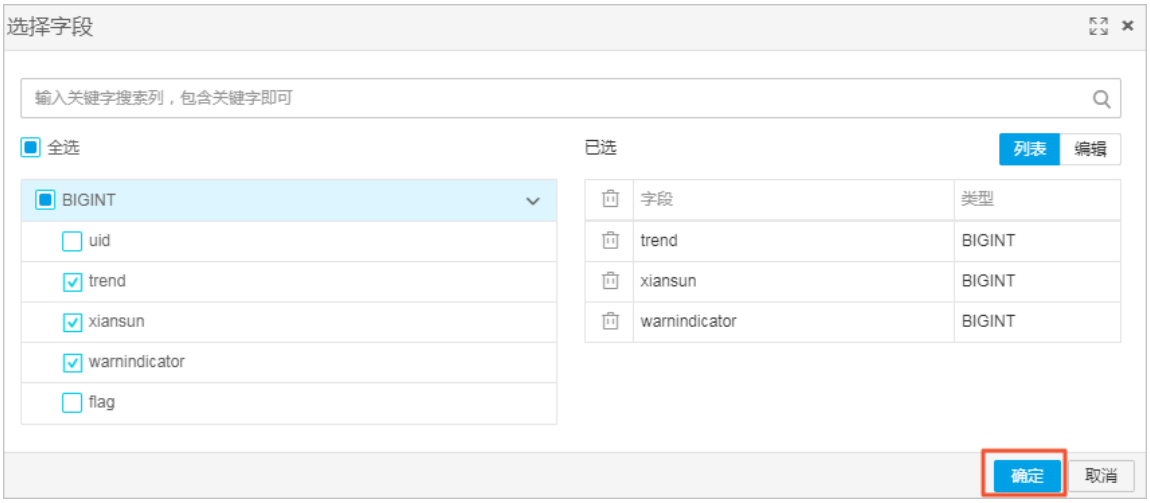


b) 连线拆分中的输出表1和逻辑回归二分类的训练表。

c) 双击逻辑回归二分类，选择右侧的字段设置 > 选择特征列，单击选择字段，并选择目标列为flag。

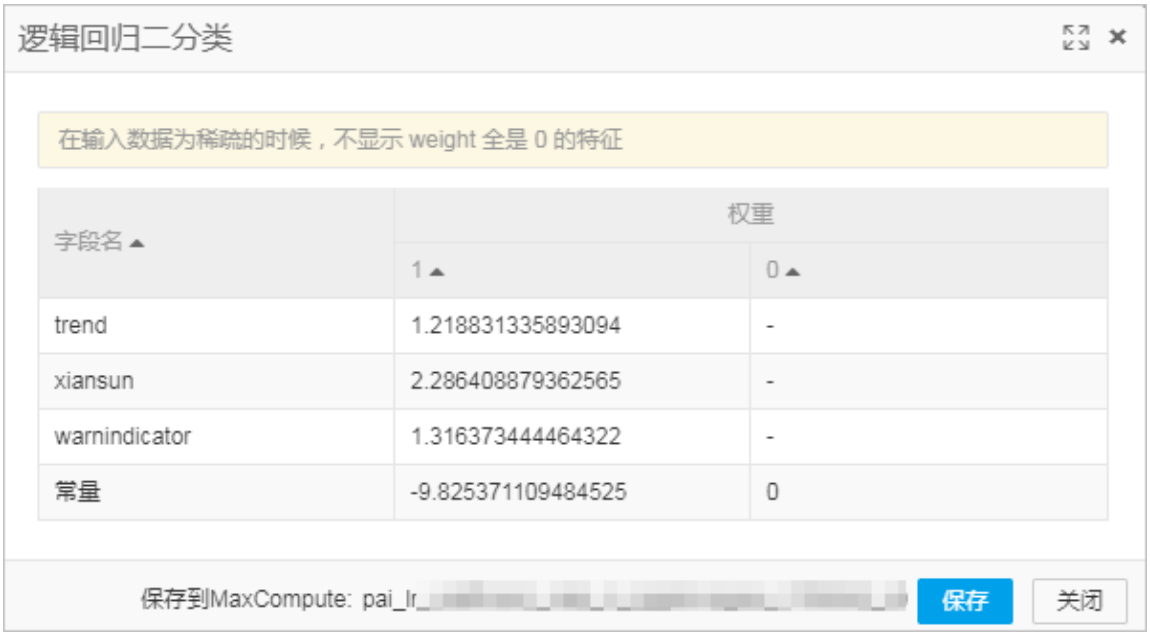


d) 在选择字段对话框中，选择trend、xiansun和warnindicator3个字段，单击确定。



e) 右键单击 逻辑回归二分类，选择从此处开始执行。

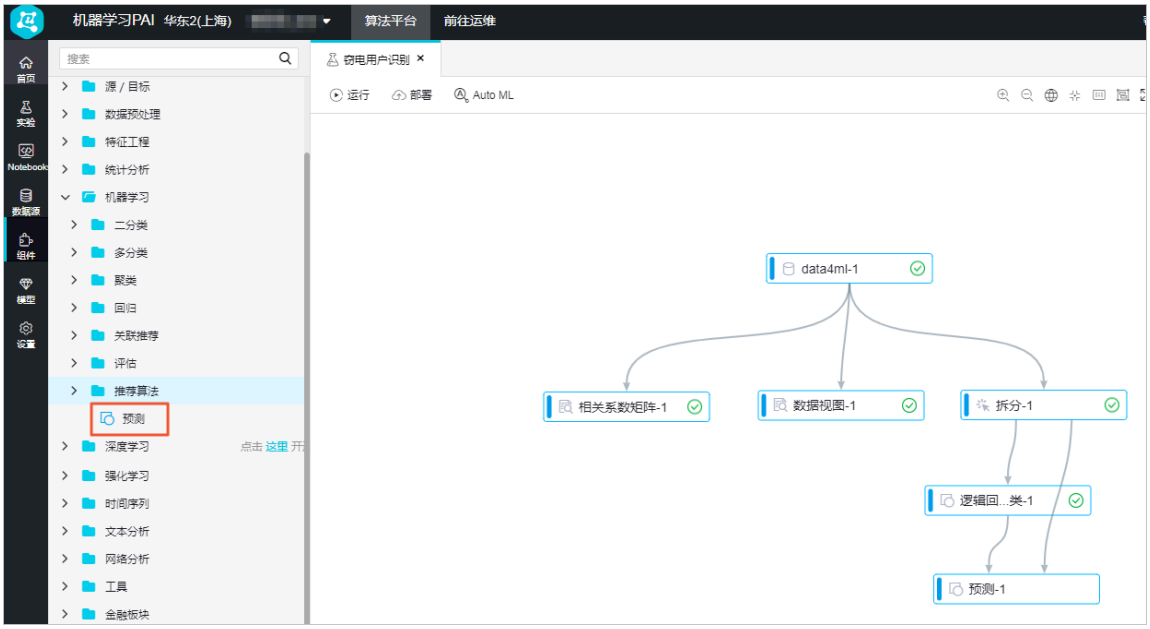
f) 执行完成后，选择模型选项 > 查看模型，即可查看数据模型。



预测和评估回归模型

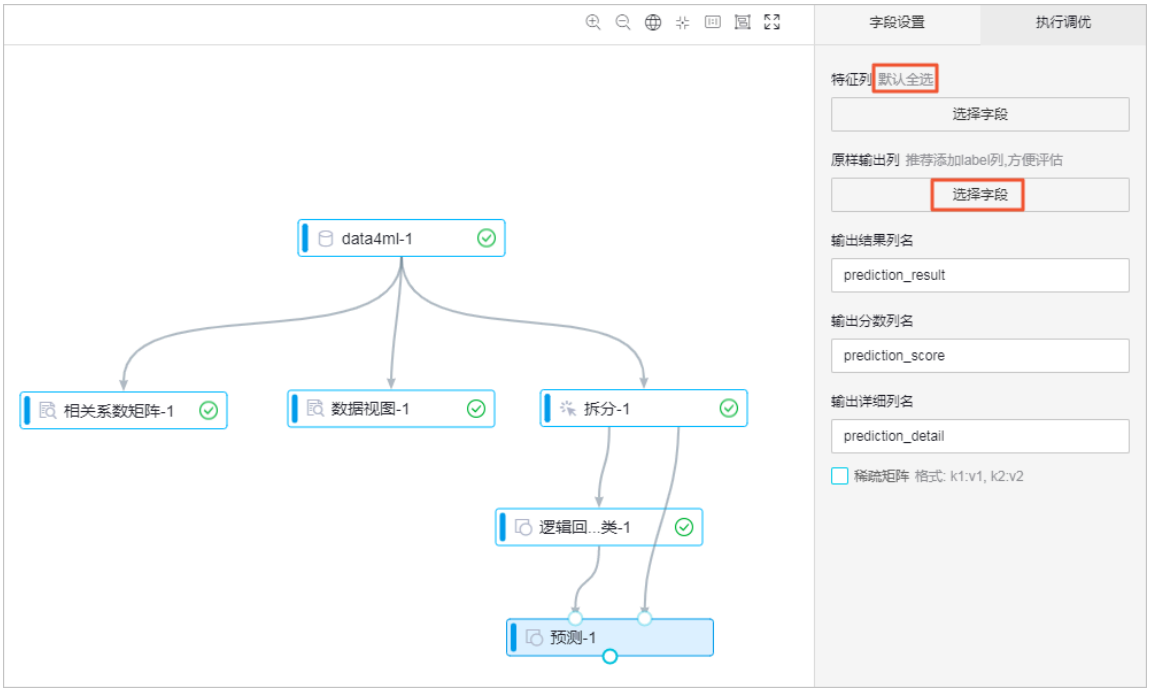
1. 通过预测组件，预测该模型在测试数据集上的效果。

   a) 单击左侧导航栏中的组件，拖拽机器学习 > 预测至右侧画布。



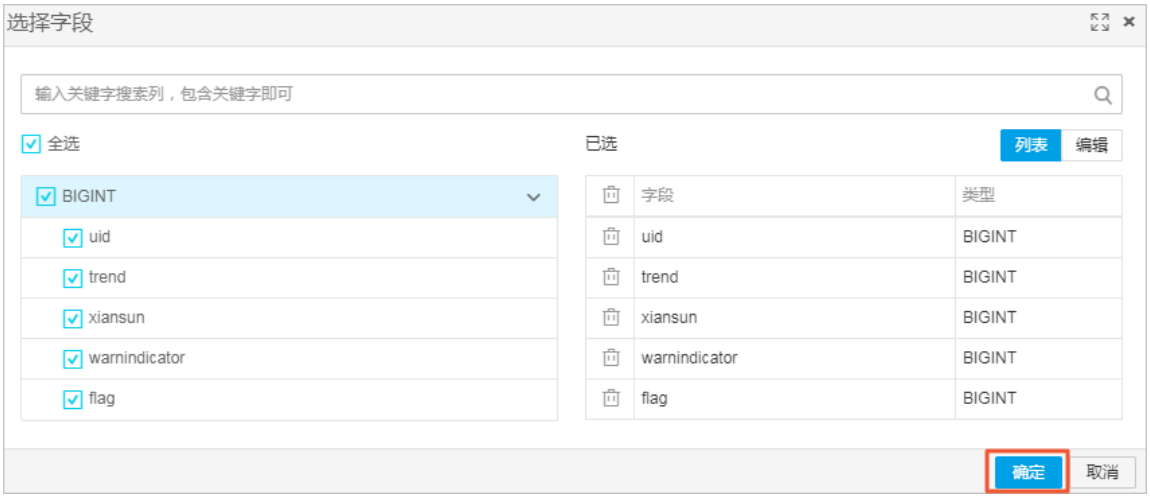
   b) 连线逻辑回归二分类中的逻辑回归模型和预测中的模型结果输入。连线拆分中的输出表2和预测的预测数据输入。

   c) 双击预测，进行右侧的字段设置。



   特征列默认全选，单击原样输出列下的选择字段。

   d) 在选择字段对话框中，全选5个字段，单击确定。

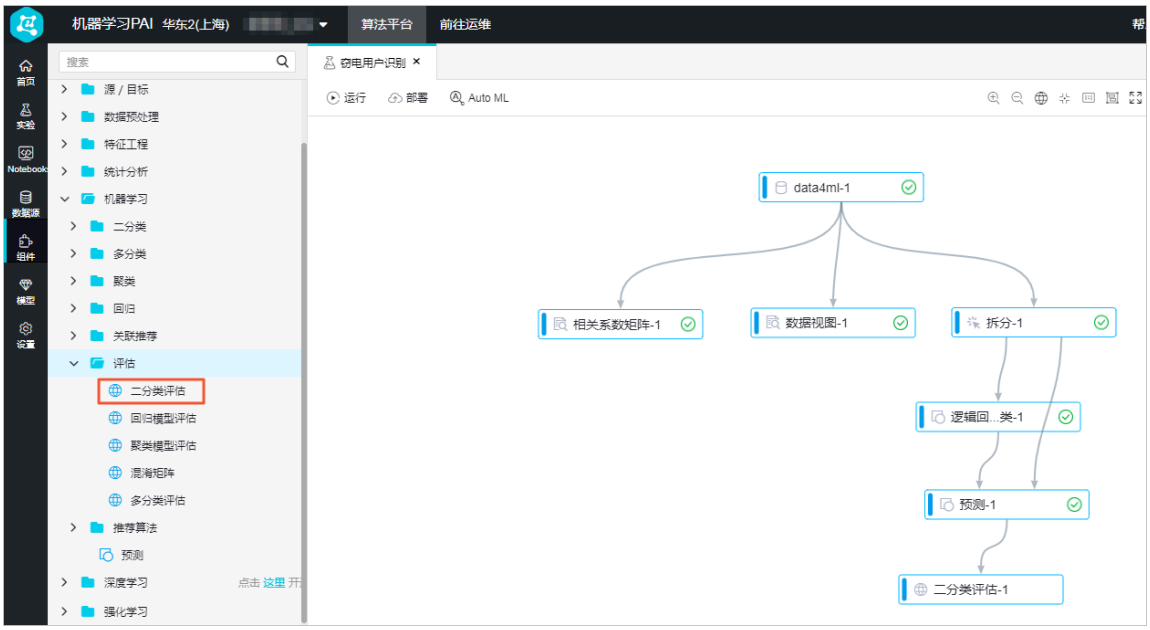


e) 右键单击预测，选择从此处开始执行。

f) 执行完成后，选择查看数据。

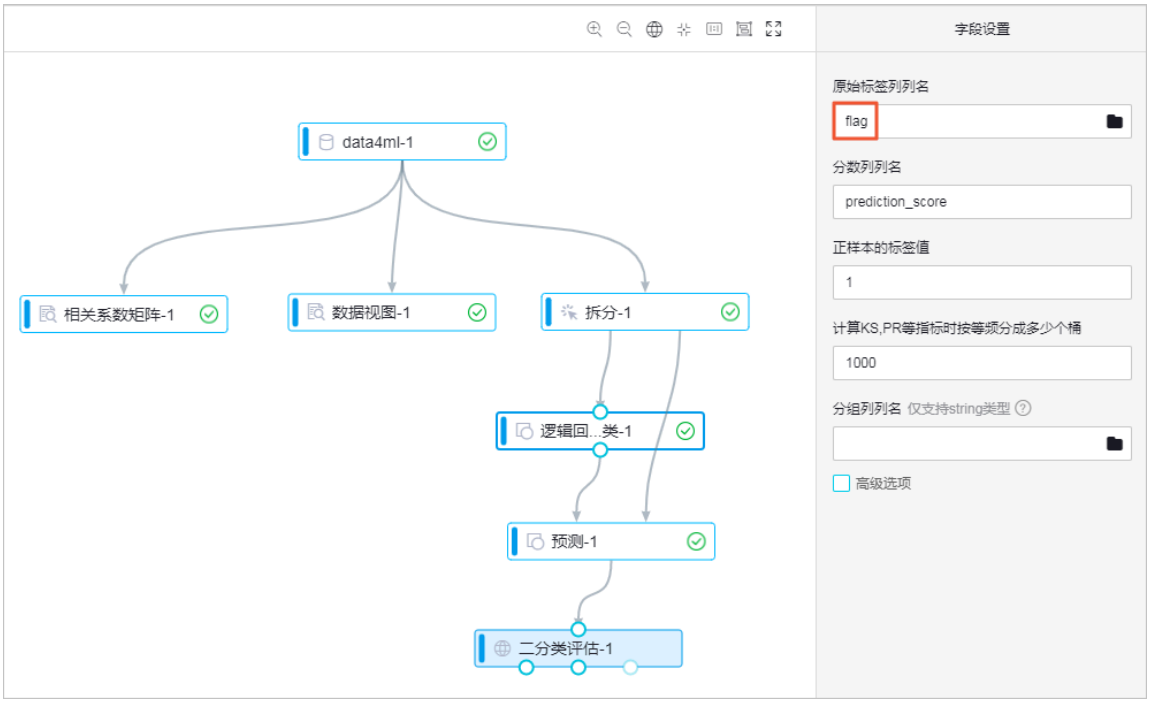
2. 通过二分类评估组件，获取模型效果。

a) 单击左侧导航栏中的组件，拖拽机器学习 > 评估 > 二分类评估至右侧画布。



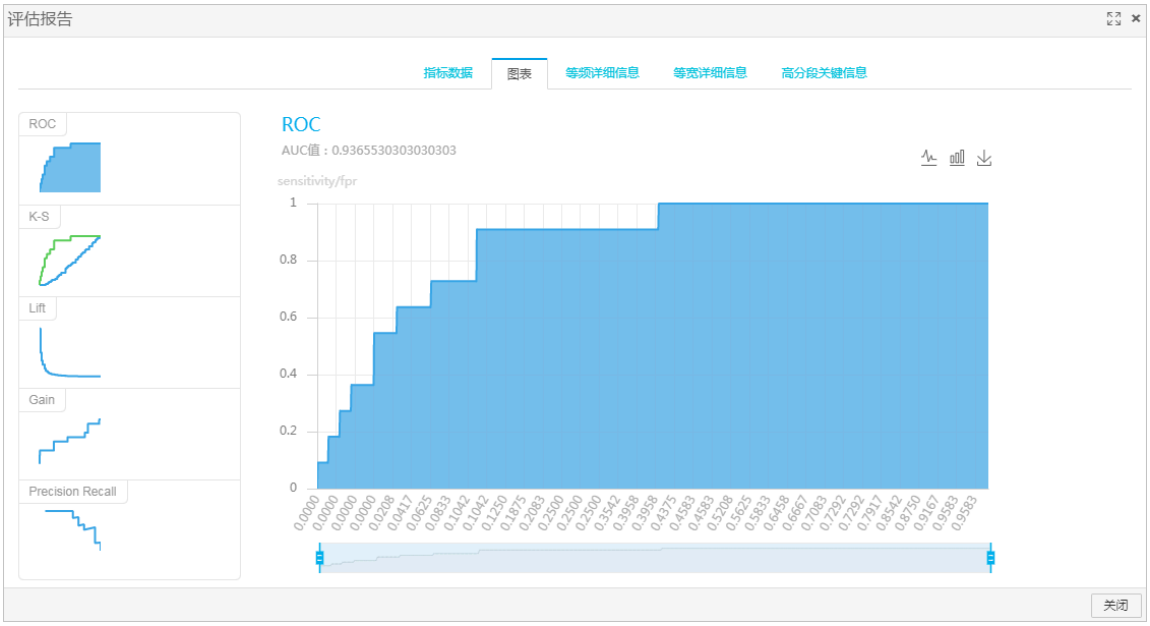
b) 连线预测中的预测结果输出和二分类评估中的输入。

c) 双击二分类评估，选择右侧的字段设置 > 原始标签列列名为flag。



d) 右键单击二分类评估，选择从此处开始执行。

e) 执行完成后，选择查看评估报告，即可查看模型效果。



后续步骤

至此，您已通过机器学习PAI完成了用户窃电行为的识别。您还可以通过EAS在线部署，将该服务部署为可在线调用的服务，为电网提供用户窃电行为的在线识别服务。