阿里云 DataWorks

使用教程

文档版本: 20190906

为了无法计算的价值 | []阿里云

<u>法律声明</u>

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

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

通用约定

格式	说明	样例
•	该类警示信息将导致系统重大变更甚至 故障,或者导致人身伤害等结果。	禁止: 重置操作将丢失用户配置数据。
A	该类警示信息可能导致系统重大变更甚 至故障,或者导致人身伤害等结果。	▲ 警告: 重启操作将导致业务中断,恢复业务所需 时间约10分钟。
Ê	用于补充说明、最佳实践、窍门等,不 是用户必须了解的内容。	道 说明: 您也可以通过按Ctrl + A选中全部文件。
>	多级菜单递进。	设置 > 网络 > 设置网络类型
粗体	表示按键、菜单、页面名称等UI元素。	单击 确定。
courier 字体	命令。	执行 cd /d C:/windows 命令,进 入Windows系统文件夹。
##	表示参数、变量。	bae log listinstanceid Instance_ID
[]或者[a b]	表示可选项,至多选择一个。	ipconfig [-all -t]
	表示必选项,至多选择一个。	<pre>swich {stand slave}</pre>

目录

法律声明	I
通用约定	I
1 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	
1.7 通过Function Studio开发UDF	
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	
2.2 环境准备	
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	
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	
3.3 数据资产定级	
3.4 离线数据加工卡点	
3.5 数据质量风险监控	
3.6 数据及时性监控	116
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	
4.2 环境准备	
4.3 数据准备	
4.4 数据加工	
4.5 数据建模	

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。
- 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。
- 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。
- 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

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

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。

■ (-)阿里云	③ 全球		Q 搜索			费用	工单	备案	企业	支持
DataWorks	常用工作空间							:	全部工作	空间
概览	-			华东1					4	峰东2
工作空间列表	创建时间:2019-08-3 计算引擎:MaxComp	21 11:13:05 oute	PPセンテ始由い		创建时间:2019-09-04 16:43:37 计算引擎: MaxCompute		Èthai			
计算引擎列表 个	T作空		进入数据开发		加芬使快: Appstudio adis 开友 数据自注 如 工作空间配置	1836 763636	进入费	如据开发		
MaxCompute	进入数	据服务	进入数据集成		进入数据服务		进入数	如据集成		
交互式分析(Hologres)	100.000			华北2	10.0				4	峰南1
	创建时间:2019-08- 计算引擎:MaxComp 服务模块:AppStudi	30 20:33:29 bute PAI计算引擎 5 数据开发 数据管理 数据	服务 运维中心		创建时间:2019-08-13 16:11:39 计算引擎:MaxCompute 服务模块:AppStudio 数振开发 数据管理 数	据服务 运线	主中心			
	工作空	间配置	进入数据开发		工作空间配置		进入数	如据开发		
	进入数	据服务	进入数据集成		进入数据服务		进入费	如据集成		
	常用功能 〇 ² 创建工作空间	× 一键CDN								

您也可以进入工作空间列表页面,单击创建工作空间。

■ (-)阿里云	华东2(上海) ▼	Q 搜索			费用	工单	备室	企业	支持与服务
DataWorks	DataWorks / 工作空间列表								
概览	创建工作空间 请输入工	作空间/显示名 Q							
工作空间列表	工作空间名称/显示名	模式	创建时间	管理员		状态		开	画服务
	-	简单模式 单环境	2019-09-04 (16:43:37)			✔ 正常		V	<u>.</u>
MaxCompute 조百式分析(Hologres)	1.000 (0.00) (0.000) (0.000)	标 准模式 开发跟生产隔离	2019-09-04 (15:06:49)	-) 更新中	7	V	<u>.</u>
	1010	标 曲模式 开发跟生产隔离	2019-09-04 (15:01:15)			✔ 正常		V	ς

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。

创建工作空间							
1 基本配置	2 选择引擎 3 引擎详情						
选择region							
华东1(杭州)	华东2((上海)	华北2(北京)	华南1 (深圳)	西南1(成	邹)	
中国(香港)	新加坡		澳大利亚 (悉尼)	马来西亚(吉隆坡)	印度尼西亚	(雅加达)]
日本(东京)	印度(;	孟买)	德国 (法兰克福)	英国 (伦敦)	美国(硅谷	;)	
美国 (弗吉尼亚)	阿联酋	(迪拜)					
基本信息							
* 工作空间名称		需要字母开头	, 只能包含字母下划线和数字	2			
显示名		如果不埴,默	从为工作空间名称				
* 模式		简单模式(单环	不境)		\sim		
描述							
立れい平							
回级设直 * 部下裁Salart结里 ❹		Ŧ					
BD Frankrickender A							
下一步	取	消					

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。

创建工作空间
 ✓ 基本配置 ② 选择引擎 ③ 引擎详情
选择DataWorks服务
数据集成、数据开发、运维中心、数据质量 您可以进行数据同步集成、工作流编排、周期任务调度和运维、对产出数据质量进行检查等。
选择计算引擎服务
✓ MaxCompute ● 按量付费 ○ 包年包月 ● 开发者版 开通后,您可在DataWorks里进行MaxCompute SQL、MaxCompute MR任务的开发。 充值 续费 升级 降配
髦。实时计算 ○ 共享模式 ○ 独享模式 开通后,您可在DataWorks里面进行流式计算任务开发。
↓ E-MapReduce 开通后,您可以在DataWorks中使用E-MapReduce进行大数据处理任务的开发。
☆ 交互式分析 开通后,您可以在DataWorks里使用Holostudio进行交互式分析(Interactive Analytics)的表管理、外部表管理、SQL任务的开发。
选择机器学习服务
下一步 上一步 取消

5. 进入引擎详情页面,填写选购引擎的配置。

创建日	C作空间							
\bigcirc	基本配置			— 🕢 选	择引擎		3	引擎详情
~	MaxCompute							
	* 实例名称							
	开发环境				生产环境			
	MaxCompute项	目名称 🛛			MaxCompute项目	目名称 🛛		
	MaxCompute访	问身份	个人账号		MaxCompute访问	回身份 ❷	工作空间所有者	\sim
					* Quota组切换		按量付费默认资源组	~
êIJ	建工作空间	上一步						

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



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。 道 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lXnzUngTSebt3SfLYxZx oSjGAK6IaF。

- 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 '性別',
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

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。

送明:

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。
- 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。

送明:
 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 '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(adiac)
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。

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。

=		表列表		
➡ DQC监控	MaxCompute V			
品 概览	済输入 Q			
吕 我的订阅	dqc	表名	责任人	操作
創業		ods_raw_log_d	April 2 March 199	配置监控规则
① 任务查询				〈上一页 】 下一页 〉

选择ods_raw_log_d表,单击配置监控规则,将会进入如下页面。

三 ✔ DQC监控	5 规则配置											
品 概 览	规则配置 > 应用名:dqc_test_022	1 〉 表名:ods_ra	iw_log_d 〉 分区表i	去式 :	关联调度 ⚠							
吕 我的订阅	已添加的分区表达式	模板规则(0)	自定义规则(0)									
一 规则配置								100 av 200 av	100 M 100			
① 任务查询	加分区表达式	规则名称	规则字段	强	规则模版	动态阈值	比较万式	橙色阈值	红色阈值	期望值	配置人	操作
							没有	 牧据				

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。

	〉表名	:ods_raw_log_d 〉 分区表达式	:: 关联调度 企						
已添加的分区表达式 + 当前无可操作的分区表达式,请先添 加分区表达式	<mark>模板類</mark> 规则	添加分区			×	词值	期望值	配置人	操作
		分区表达式:	評論入分区表达式 dt=S[yyyymmdd] dt=S[yyyymmdd] dt=S[yyyymmdd]hb24miss] dt=S[yyyymmddhb24miss]/24] dt=S[yyyymmdd]000000 dt=S[yyyymmdd]000000 dt=S[yyyymmdd]7]	计算					
~				御认見	び肖				

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。

+ 快捷添加

单击添加监控规则,会出现一个提示窗,用于配置规则。

三 ▼ DQC监控	∽ 规则配置				模板规则自定义规	见则		
品 概览	规则配置 〉 应用名:dqc_	🐚 〉 表名:ods_raw_log_d	〉 分区表达式	: dt=S[yyyymmdd-1	+	- 添加监控规则	+ 快捷添加	
a anita	已添加的分区表达式	模板规则(0) 自定义	规则(0)	责任人:				
	- dt=S[yyyymmdd-1]	规则名称 规则	1字段 强	规则模版	* 规则名称:	请输入规则名称	*强弱: 💿 强 🔵 弱	
U THERE					*规则字段:	表级规则 (table) V		
					▶规则模版:	表行数,固定值		^
					*比较方式:	相比上一周期表大小变化相比上一周期表行数变化		1
					描述:	相比上一周期表行数不变		
						表大小,上周期差值 表行数,上周期差值		
						✓ 表行数,固定值		
						表大小, 回走值 表行数, 1天差值		L
						表大小,上周期差值		v

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。

三 ↓ DQC监控	5 规则配置			模板规则 自定义规		
品 概览	规则配置 > 应用名:dqc_tem_^_	≧ 〉 表名:ods_raw_log_d 〉 分区	表达式 :dt=Slyyyymmdd	1 +	添加监控规则	+ 快捷添加
吕 我的订阅	已添加的分区表达式	楼板规则(0) 自定义规则(0)	责任人:			
」規則配置	+ dt=Slvvvvmmdd-11	规则名称 规则字段	强 规则模版	* 规则名称 :	请输入规则名称	* 366 : 9 35 55
① 任务查询				* 规则字段:	表级规则 (table) V	
				+ 规则模版 :	表行数 , 固定值	~
				* 比较方式 :	大于 🗸	* 烱盤值: 0
				描述:		
		1				
						批量保存取消
- 						

文档版本: 20190906

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。

- DQC监控	5 规则配置) 77 1										
品 <i>概</i> 览	规则配置 > 应用者: 0qc1	〉 病名:0ds_rav	v_log_d > 551≧%b	232C : 0T=3	slakkare (关联调度						
8 36010	已添加的分区表达式	模板规则(1)	自定义规则(0)		责任人:					1	试跑 订阅管理	创建規則 更多・
一規则配置	+	1000 400	40.014.00	20	47777114414			atte de Atte da	fan de Alto de			
① 任务查询	 dt=S[yyyymmdd-1] 	规则省称	规则子段	9 <u>9</u>	规则模版	动态同归	KARDE	橙色腐值	红色同胞	期望值	配置人	操作
		DQC_rule_01	表级规则	强	表行数,固定值	否	大于	-	-	0	1000	修改 删除 日志

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。

	∽ 规则配置														
品 概览	规则配置 〉 应用名:dqc_t	> 表名	: ods_raw_log_d	〉 分区表达:	式 : dt=\$[yyyymi	ndd-1] 关联i									
我的订阅	已添加的分区表达式	模板规	试跑						>	K		试跑 订	间管理	创建规则	更多・
① 任务查询	- dt=S[yyyymmdd-1]	规则								[色阈值	期望值	配置人	3	桑作	
		DQC		试跑分区:							0		1	^像 改 删除 日志	
				·调除时间: 【	2019-02-22 试验成功!点时	13:22:27			(B)						
	实例详情														
➡ DQC监控	应用 :dqc表名 :ods,	_raw_log_d > dt=	S[yyyymmdd-1]	苗 201	9-02-22 13:24:49	更多								周	制制
	规则名称 规则字段	强/弱	采样方式	过滤条件	校验类型	校验方式	比较方式	橙色阈值	红色阈值	期望值	历史结果	采样结果	状态	操作	
	DQC_rule_01 ·	强	table_count		数值型		大于		**	0		570386	正常	查看历史 结果	
① 任务查询															



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。

Software Software	DQC_test_0221 V
Ξ ③ 运维大屏	
✔ 任务列表	基総:
🕞 周期任务	
(1) 手动任务	E 2000 E 200 E 200
→ 任务运维	rpLuser_info_d 700002197984
	dw_user_info_all_d 700002197983
	ods_log_info_d 700002197982
All All All All All All All All All	rds. 数据同步 700002197849
▶ 智能监控	workshopstart 700002197848 ■ OSS 数据同步 展开父节点 >
	C Data Integration 展开子节点 > 共子学点 >
	고 전체에 가지 않는 것이 같이 같이 있다. 이 가지 않는 것이 같이 같이 있는 것이 같이 없는 것이 같이 않는 것이 없는 것이 없 않는 것이 없는 것이 않는 것이 없는 것이 않는 것이 없는 것이 없는 것이 없는 것이 없는 것이 않는 것이 없는 것이 않는 것이 없는 것이 없는 것이 않이 않는 것이 않이 않이 않는 것이 않는 것이 않이
	編載で点
	ods_log_info_d 音響血線 高四 高四 高四 高四 高四 高四 高四 高四 高四 高四
	OUFS_SQL 測试 度失型 补数据 > cr.i
	間子(海話) 「描述」
	更多▼ 〈 1/1 〉

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]。

✔ DQC监控	切り 切り 切り 切り 切り 切り 切り の こ ロット ロー) 宇夕 Lodo roy	u log d \ ART	+-+	Changemented 1			
器 概 览	规则配置 > 应用名:oqc_test_0221	/ 表台:00S_ra	w_log_a / 刀込衣i		-Slaaker 1	医联调度 / 小		
吕 我的订阅	已添加的分区表达式	模板规则(1)	自定义规则(0)		责任人:			
(≜) 规则配置	+	1001-0250		20	400 00 d.148 U.C.	-1	11 observe - D	
① 任务查询	- dt=S[yyyymmdd-1]	规则名称	规则子段	强	规则模成	动公鸣但	比较方式	位巴國但
		DQC_rule_01	表级规则	强	表行数 , 固定值	否	大于	



- ▼ DQC监控 品 概%	 5 規则配置 規则配置 > 应用名:dqc_test_023 	21 〉 表名:ods_u:	ser_info_d 〉 分区表	転式 : (dt=S[yyyymmdd-1]	关联调度							
S Anita	已添加的分区表达式	494637899 (0)	自定义规则(0)		责任人:						试胞 订阅管理	创建规则	更多,
	- dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色褐值	期望值	配置人	操作	
							没有	牧 掘					

- 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。

➡ DQC监控	与 规则配置 规则配置 > 应用名:dqc_test_02	21 〉 表名:ods_ra	w_log_d 〉 分区表	达式 : dt	=S[yyyymmdd·1]	关联调度							
器 我的订阅	已添加的分区表达式	模板规则(1)	自定义规则(0)		责任人:					1	北海 订阅管理	创建规则	s.
	+ dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色閾值	红色阈值	期望值	配置人	操作	
		DQC_rule_01	表级规则	强	表行数,固定值	否	大于			0	100	修改 删除 日志	

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		\times
将当前分区表达式关联到:		
DTdoc 🗸	任务节点名称	添加
oss数据同步(7000025443	55) ×	

关联调度后,表名后面的小图标会变成蓝色。

		21 〉 表名:ods_rav	v_log_d 〉 分区表达	去式 : dt	=S[yyyymmdd-1]	关联调度						
	已添加的分区表达式	根板规则(1)	自定义规则(0)		责任人:					ja Ja	(跑)订阅管理	创建规则 更多,
無則配置	+ dt=S[yyyymmdd-1]	规则名称	规则字段	强	規則模版	动态阈值	比较方式	橙色閾值	红色阈值	期望值	配置人	操作
① 125000		DQC_rule_01	表级规则	强	表行数 , 固定值	否	大于	-	-	0	-	修改 删除 日志

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。

■ DQC监控		〉 表名:ods_ra	w_log_d 〉 分区表达	s ≣ C : dt	\$[yyyymmdd-1]	(联调度						
	已添加的分区表达式	模板规则(1)	自定义规则 (0)		责任人 :					ž	式跑 订阅管理	创建规则 更多,
<u>角</u> 任务查询	- dt=S[yyyymmdd-1]	规则名称 DQC_rule_01	规则字段表级规则	强强	规则模版 表行数 , 固定值	动态阈值	比较方式	橙色阈值 	红色阈值 	期望值 0	配置人	操作 修改 翻除 日志
✔ DQC监控 品 概览	5	〉 表名:ods_ra	w_log_d 〉 分区表达	th : JEZ	=S[yyyymmdd-1]	关联调度						
我的订阅	已添加的分区表达式	校板规则(1)	自定义规则(0)	30	责任人		11.45 miles	a art	177 (7. 33) (7.	1048.41	试跑 订阅管理	创建规则 更多・
① 任务查询	• dt=\$]yyyymmdd-1]	DQC_rul	1001+12 101管理	30		14/1224091E	T	X		0		修改 删除 日志
			订阅方式 邮件通知 >		接受对象 tina	^	操作 保存					
					dtplus_docs	Q 		关闭				
					11111	此处会展	示本项目下	《的全部成员	•			
✓ DQC监控 BR 概策	5 規则配置 規則配置 > 应用名:dqc_test_0221	〉 表名:ods_ra	w_log_d 〉 分区表i	达式 : dt	:=S[yyyymmdd-1]	关联调度						
3 我的订阅	已添加的分区表达式	模板规则(1)	自定义规则(0)		责任人:						试胞 订阅管理	創建規则更多・
负 任务查询	- dt=S[yyyymmdd-1]	规则名称 DQC_rul **	规则字段 1)阅管理	强	規則模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值 0	配置人	操作 修改 删除 日志
			订阅方式		接受对象		操作					
			邮件通知 ∧ ✓ 邮件通知 邮件和短信通知 钉钉哔机器人 钉钉哔机器人。) ALL		~	保存	关闭				

订阅管理设置完毕后,可以在我的订阅中进行查看及修改。

- DQC监控	我的订阅		分区表达式	责任人	操作
品 概览	MaxCompute > 请输入表名 C	清空	dt=\$[yyyymmdd-1]		上次结果 通知方式▼ 取消订阅
🗧 我的订阅	MaxCompute项目 🏹	表名			
自规则配置	dqc_test_0221	ods_raw_log_d			
① 任务查询	DQC_test_0221	ods_user_info_d			

建议将全部规则订阅,避免校验结果无法及时通知。

 \cdot 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。

→ DQC监控	∽ 规则配置											
品 概览	規則配置 > 应用名:dqc_test_02	21 〉 表名:ods_user_info_	_d 〉 分区表达式	t : dt=S[yyyymmdd-1]	关联调度 △							
吕 我的订阅	已添加的分区表达式 十	模板规则(0) 自定义	2规则(0)	责任人:						试跑 订阅管理	创建规则	更多,
日 規則配置 ① 任务查询	• dt=S[yyyymmdd-1]	规则名称 规则	川字段 引	虽 规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作	
						没有要	刘居					

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。

5 规则配置				模板规则	自定义共	见则				
规则配置 〉 应用名:dqc_test_02	21 〉 表名:ods_u	ser_info_d 〉 分区表	長达式 : dt=\$[yyyymm	ld	+	- 添加监控规则		+ 快	捷添加	
已添加的分区表达式	模板规则(0)	自定义规则(0)	责任人							_
- dt=S[yyyymmdd-1]	规则名称	规则字段	强规则模拟		★规则名称:	DQC_rule_02		*强弱: ●	强 🔾 弱	
					*规则字段:	表级规则 (table)	\sim			
				_	* 規则模版:	表行数 , 固定值				\sim
					* 比较方式:	大于	\sim	* 期墾值: 0		
					描述:					
	•									

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。

5 規则配置 規则配置 > 应用名:dqc_test_0	221 〉 表名:ods_u	iser_info_d 〉 分区	表达式 : dt	-s[yyyymmdd	機械規则 自定义规则 + 添加监控规则 + 快捷添加	
已添加的分区表达式 + - dt-S[yyyymmdd-1]	楼板规规则(0) 规则名称	自定义规则(0) 规则字段	强	责任人: 规则惯版	 ・規則名称: DQC_rule_03 ・短節: 强 ● 弱 ・規則守僚: uid (string) ・規則昭誕: 重复值个数,回定值 ・比较方式: 小子 ・出版5: 	
	*				 ・規則名称: DQC_rule_02 ・強弱: ④ 强) 弱 ・規則字段: 表現規則(table) ・規則相談: 表行数,固定值 ・比较方式: 大子 	
					waa. 批量保存	取消

🗾 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]。

- DQC监控	ら 規则配置												
品 概范	规则配置 > 应用名:dqc_test_022	1 〉表名:ods_lo	g_info_d 〉 分区表	达式 : dt	=S[yyyymmdd-1]	关联调度							
8 8010	已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:					ì	式跑 订阅管理	创建规则	更多・
(言) 规则配置	+	10.01	400000-00	20	40.0114910-	-1-1-30 /**	11 statester 10	an a 18 m	100.00 200.000	110.48.41	3073 000 /	49.14	
凸 任务查询	 dt=S[yyyymmdd-1] 	規则省称	规则子拔	58k	\$6291965R	和小小肉组	75692/3375	位巴姆坦	KLEO/6918	A812101	RGBEA	1981 F	
							没有影	数据					

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。

_ DQC监控	5 规则配置					模板规则 自定义的				
铝 概览	规则配置 〉 应用名:dqc_test_02	21 〉 表名:ods_k	og_info_d 〉 分区表	送式 : dt=!	Slyyyymmdd-1		• 添加监控规则		+ 快捷添加	
要我的订阅	已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:					
① 任务查询	- dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	*规则名称:	请输入规则名称		*强弱: • 强	55
						*规则字段: *规则模版:	表极规则 (table) 表行数,固定值	~		~
						*比较方式:	不等于	~	◆ 期望值: 0	
						描述:				
	· · · ·									
									批量保存	取消

· 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]。

= ▼ DQC监控 品 概览	5 規则配置 規則配置 > 应用名:dqc_test_022	21 〉 表名:rpt_us	er_info_d 〉 分区表	li达式:dt	t=S[yyyymmdd-1]	▲ 美歌调度							
	已添加的分区表达式	模板规则 (0)	自定义规则(0)		责任人:						试跑 订阅管理	创建规则	更多,
① 任务查询	- dt=S[yyyymmdd-1]	规则名称	规则字段	<u>3</u> ;	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作	
							没有数	斑唇					

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。

_ ↓ DQC监控	5 规则配置					模板规则	自定义规	观则			
品 概览	规则配置 〉 应用名:dqc_test_0221 〉 表名:rpt_user_info_d 〉 分区表达式:dt=Slyyyymmdd-						+	添加监控规则		+ 快捷添加	
n the the test of test	已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:						
規則配置	+	规则名称	规则字段	强	规则模版	*	規则名称:	请输入规则名称		*强弱: 〇 强 💿 弱	删除
① 任务查询						*	规则字段:	uid (string)	~		
						*	规则模版:	重复值个数 , 固定值			\sim
						*	比较方式:	小于	\sim	*期經值: 1	
							描述:				

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。

	← DQC监控	5 規则配置			. de Olassamandal 8	1	模板	<mark>规则</mark> 自定义的	現100 1960-1000 11-96949			. Kneikinn	
	88 46233	HUNINGER / MARKET OCCUSE	21 / 表名:rpt_user_	nto_d / 为这表达式。	: at-siyyyymmaa-i								
		已添加的分区表达式	ARREARING (O) E	主 义规则(0)	责任人			*规则名称:	请输入规则名称		*	1995: 🕜 强 🧿	弱
	日 7600 maga	• dt=S(yyyymmdd-1)	规则名称	规则字段	强	规则模版	स्त्रा	* 规则字段:	表级规则 (table)	~			
								* 3030160355 :	表行数,7天波动率				~
								*比较方式:	绝对值	\sim			
								波动值比较: 0	1% 25	5%	50%	75%	100%
								C)		0		
									橙色间值: (%	红	色阈值: 50	8
・ 現地 名称: 第10公 300 15 10 ・ 第66 : 一 預 ④ 第5 ・ 現地 5 20 15 10 ・ 1 ・ 第66 : 一 預 ④ 第5 ・ 現地 5 20 15 10 ・ 1 ・ 1 ・ 1 ・ 1 ・ 1 ・ 1 ・ 1 ・ 1 ・ 1								描述:					
・ 服務 ・													
・NEWESHE: INDALAGNESHE ・SIGE: 預加 ・ SIGE: 預加 ・ SIGE: 預加 ・ SIGE: 預加 ・ SIGE: 可加 ・ SIGE: SIGE: 可加 ・ SIGE: SIGE: 可加 ・ SIGE:													
・ 第二の公司の計 ・ 第二の公司の計 ・ 第二の ・ ・ ・ ・ ・ ・ ・ ・ ・ ・ ・								+ 1000a Cr 2/r .				2027. ZE (22
								• 202201-0-000 :			•	2009: 🕛 194 🕒	999
・利利時代語: 屋空道: ・ ・ 和田田代語: 屋空道: ・ ・ ・ 田田田代語: ・ 田田田代語: ・ ・ 田田田代語: ・ ・ 田田田代語: ・ ・ 田田田代語: ・ 田田田代: ・ 田田代: ・ 田田代: ・ 田田田代: ・ 田田代: ・ ・ 田田田代: ・ 田田代: ・ 田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ 田田代: ・ 田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ 田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ 田田田代: ・ ・ 田田田代: ・ ・ ・ 田田田代: ・ ・ ・ ・ ・ ・ ・ ・								* KENE-9-EX :	uid (string)	~			
* Etstofrat: /h-F								* 102301452352 :	重复值个数,固定值				~
								*出始方式:	小干	~	* 10	照应: 1 批型保存	RoH



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。

儿外关了但,于山廷该洲松,内亚小妖相孙廷也且正用泪干山孙相称了	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
---------------------------------	--------	---------	--------------------

	✓ 数据源车通性正常!	
171 EUXE	F工问 UNFA DIQ 通过的小	
添加MaxCompute数据	原	×
* 显示名称:	test_workshop	
* 数据库地址:	http://service.cn.maxcompute.aliyun.com/api	
*项目名称:	test_workshop	
* AccessKey ID:	LTAI2i	
* AccessKey Secret:		
① 温馨提示:新增数	居源存在同步延迟的情况,请稍候片刻。	
	关闭 连接测试	添加

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

数据表 上传文件		Q 共7个文件	即席分析SQL 上传文件
名称 ♦	备注♦		操作
bank_data			î ()
dw_user_info_all_d			î ()
ods_log_info_d			î ()
ods_raw_log_d			ŵ ()
ods_user_info_d			ŵ ()
result_table			1) (1) (1) (1) (1) (1) (1) (1) (1) (1) (
rpt_user_info_d			1

选择您想放置的数据集位置,单击确定。

创建数据集					×
	*名称:	rpt_user			
		0.000			
	*1 <u>\/</u> 吉:	ODPS		~	
				关闭	确定
- := 数据集 全部 我的 🔍 默认空间 \bigcirc 根目录 > ODPS 数据门户 名称 💲 ■■ 仪表板 rpt_user NEW \mathbf{T} rpt_user_info_d 电子表格 自定义sql5 自助取数 (公测) NEW 自定义sql5 🏹 数据集 rpt rpt_user_trace_log
- 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。

Q 输入关键	字搜索					
		Str.		Str.		Str.
维度		uid		re	gion	device
Str. uid	愈	_				
str. regio	⊘ 编辑					
str. devi	C) 克隆维度					
str. geno	▲ ▲ 左公析市陶蔵					
str. age_						
Str. zodi	⊙ 取消全部隐藏					
Str. dt	★ 删除					
	+ 新建计算字段(维展	度)				
	よ 移动到	•				
	🕄 新建层次结构					
	↓ 下移					
度量 - 🗁 默认	↓ 转换为度量				yyyyMMdd hh:m	i:ss
[№] pv	₽ 维度类型切换	•	日期(源数据格式)) 🕨	yyyy/MM/dd hh:	mi:ss
			地理信息	•	yyyy-MM-dd hh:	mi:ss
			✔ 文本		yyyy/MM/dd	
			数字		ууууMMdd	

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. regior			
Str. device	⊘ 编辑		
str. gende	12 克隆维度		
str. age_r	_		
Str. zodiac	✔ 在分析中隐藏		
str. dt	◎ 取消全部隐藏		
	★ 墨琴		
 在 量	× 10010		
	+ 新建计算字段(维度)		
N5 b∧	ぬ 移动到 🔹 🕨		
	3)新建层次结构		国家
	↑ 上移		区域
	↓ 下移		省 / 直辖市
	↓ 转换为度量	日期(源数据格式) 🕨	市
	⇒ 维度类型切换	地理信息	区/县
		✔ 文本	经度
		数字	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。

数据集	全部 现的		名称	▶ ∨ Q 共18个文件	+ 新建数据集
根目录 > ODPS					
名称 🛊		创建者 🗘	修改人/修改时间	数相同原	新建仪表板
rpt_user	f New Linfo.d	144	144	No. of Street,	⊠ .ıl ≢ :

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。

设置过滤器		\times		№ 🖬 pv 🛛 🗙
<pre>pt_user</pre>	🔟 dt(year)			日期/维度 🛈
区间粒度	✓ 日期粒度			双击或拖动数据字段至此处
1 时间区间				过滤器
Tr 2019	019 🛱			🛗 dt(year)
开始于:	结束于:			
相对时间 🗸 🗸	相对时间 🗸 🗸		-	
2019	2019			
↓ + ▼ 0	T ∳ + ▼ 0			
	取消 确	淀		

完成后可以看到当下数据。

pv	
3198.88万	

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)

圆 🛛 🖉 🗮 🖸 💌 🛉 🕸 🕲 🗳 🗒 🔞 单 🚔 🚔 📾 🖌 💷 🔤	图表设计		*	更改图表类型	1
	数据 选择数据源关型:	JEHH JEHH	数宏集	高级))	
3198.88万 3198.75	值种度量		rpt_user Oee入teentate		1
erd red d	and the second s		继度	, ter)	
14122-171_User = 10.500 = 20.400 = 20.400		≑ex ▼	🛗 dt(mont 🛗 dt(wee)	() ()	
201577 - 2015776 - 2015776 201577 - 40-509 12.2057	統色超例增度 Sw. age_range		■ dt(day) ≈ uid ⊗ region		
1875 1975 1975 1275 1275	istialiaa dt(vear)		 device gender age_rang 		
10.7 877 10.877 20.8773 20.8773 20.8773 20.8773 20.8778 20.8774 20.8778 20.877		2	se zodiac	_	
روديم ل المراجع			- CD BACK		
	预进行数 1000 3 更新				

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。

ē t 2 🛱 🖬	🕺 🏨 🖀 🕚	🗏 😼 🗐 🤌 🛎 🖷	₩ 🗶 🖽 🔤		图表设计		🐋 更改图表类型 •
		dt(day)	•		数据	样式	商級
Fund	Fund	Furul			选择数据源类型:	(Constant)	
Anticipation of upper					色彩地和皮/皮量	rpt_user	~ 12
巴来2008-rpt_user		purd 🔳 pv		euri	NO 📰 pv	⇔x <u>Q</u> №λ.≭	
					地理区域维度	- 🚠 at	
		. 📥	2	+	region	⊖ @× \ 🛱 d	t(year)
			the second	-	过滤器		t(month)
		1			🛗 dt(year)		t(week) t(day)
		A Sand	1 (2) () () () () () () () () (2 Str. uid	ion
		- ashi	and and a			Str. der	rice
			Jone San Star			str. ger	nder
■ 533.5万 - 666.8万 ゃび ■ 400.3万 - 533.5万		Ford	har and a			- LD 180	
 267万 - 400.3万 133.7万 - 267万 							<u> </u>
4459 - 133.775			erri				
					预览行数 1000	3	
*					更新		

e) 完成配置后, 单击保存及预览, 即可看到展示效果。

		Ford	्य					
		Park	3198.88万					
线图-rpt_user			- 20-30岁 · 40-	50岁 - 30-40岁	- 20岁以下 - 50岁			
2275 2075 1875 1675 1475	eard	pund	60	d.	Ford	ędid	Ford	ŧσ
1275 1075 875		enerthe and a strange	and all a start and a start	i Anon Anon Anon Ano	international and the second se	internet and a set of the set of the	and a second and a second a s	Name and Alabama
und.				dt(day)				
色彩地图-rpt_user							D V	Ford
						800	Red .	Z
						Ford	Ares E	Pard
						eard	CARE A	eard
533.575 - 666.875 400.375 - 533.575 26775 - 400.375								
4459 - 133.775							eard 🥈	

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。

🜀 🗲 Function S	Studio 🖌		
=			
① 工作空间			
◆ 模板空间	欢迎来到 Function St		
	¢:	۵	Ø
	新建模板工程	新建代码工程	导入Git工程
	我的工程		
	Q 请输入 搜索	지 아이지 아이지 아이지 않는 환성가 환성가 환성가 환성하는	a for fair fair fair an ann an ann an ann an ann ann ann an

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。

工作空间 > 新建项目				
新建项目				
模板工程 代码工程	星 导入GIt工程			
* Git 地址:	请输入 Git 地址			
* 工程名:	请输入工程名称,英文字符开头,只能包含数字	、英文字符、_、-		
* 工程描述:	请输入工程描述			
*选择运行环境:	udfpython UDFPython Project	udfjava VDFJava Proje	ect 🗸	
提交				

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。

新建项目					
模板工程 代码工					
* Git 地址:	git@code.aliyun.com:alicode-templa	设置			×
* 工程名:	案號检測到認知来及量Git Comg, 请输入工程名称,英文字符开头,只	SSH Key Git Config	* User Name:	user name	
* 工程描述:	请输入工程描述	偏好设置	* Email:	email	
* 选择运行环境:				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。

设置		×
SSH Key		
Git Config 偏好设置	 ● 将 SSH key 添加到代码仓库所属账户的账户公钥列表(git->Profile->SSH keys) * service : code.aliyun.com * Public key : 点击生成sshkey 	
	生成sshKey	

·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。
 - 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。
 - 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

沿田日 .
见明:

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

☰ (-) 阿里云		CDN Q	中国站 🗸 🗋 购物车 控制台
最新活动 产品 解决方案 定价 ET大脑 数据智能 安全 云	市场 支持与服务 合作伙伴		
表格存储。 电磁频分			
表格存储(TableStore)是阿里云自研的面向海量结构化数据存储	的Serverless NoSQL多模型数据库,被		
广泛用于社交、物联网、人工智能、元数据和大数据等业务场景。 及开创性的消息模型Timeline,可提供PB级存储、千万TPS以及毫	提供兼容HBase的WideColumn模型以 秒级延迟的服务能力。		
立即开通 产品文档	当 <u>产品价格</u> 权威指南	表格存储全新升级发布>>	
Serverless	冬数据模型		分布式架构
	支持WideColumn和Timelin	ne模型,	能无缝扩展到支持PB级存储和千万
	便于不同类型数据抽象,未 展更多NoSQL数据模型。		TPS服务能力,同时能提供高开发低延迟访问。
灵活的数据分析			
	Serverless	مړمړه	灵活的数据分析
	大需目主管理水位和容量, 性伸缩,底层提供存储和试	完全弹	对接多种大数据计算引擎,与 MaxCompute、StreamCompute和
	短期展现及日初贝料均衡。		的。 發打通。

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。

表格存储开通	
产品名称:	表格存储
计费方式:	按实际使用量后付费
计费项:	外网下行流量,预留读写吞吐量,数据存储 查看计费标准
	☑ 我已阅读并同意 表格存储开通协议
	立即开通

c) 单击管理控制台。

▲从订单 开通完成 ★麦 开通成功 !	一支付				
→ 共喜 - 开通成功!		确认订单		开通完成	
您订购的【表格存储】正在努力开通中,一般需要1-5分钟,请您耐心等待 管理控制台		秋喜,开通f 新喜,开通f 新客, 新書 新客 和名 新客 和名 新客 和名 和 和 和名 和 和 和 和 和 和 和 和 和 和 和 和	戊功! E在努力开通中,一般需要1-5分钟,请	変形心等待	

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。

华北2(北京) 🔻	Qį	· · · · · · · · · · · · · · · · · · ·	费用	工单	备案	企业	服务	>_	Ū, È	1 (?		简体中文	. ()
1 家例管理	创建实例			\times	解您的需	求和使用	反馈,划	「清您百忙	中抽空参	5! C 刷新	創研	实例	资源包
	* 实例名称:	workshop-bi-001									-	2	
 ● 该区域目前支持高性能实例和容量型实例 相关链接: 产品首页 场暑白皮书 新功 	* 立例抑格·	· 如 早 刑 穴 (A) ▼											
实例名称		甘重主火 内,						监护					操作
R R R R R R R R R R R R R R R R R R R	* 头彻注释:	test						⊵				管理	释放
	提示: 1.创建实例需要 2 创建实例后 实	】沙钟的时间,创建成功清按刷新按钮刷新实例列表。 刚域名会在1分钟内生效											
			确定	取消									

📕 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。

表指存储Table Store	实例管理					C	③调查问卷:表悟存储(TableStore) 到新 创建实例 资源包
实例列表	该区域目前支持高性能实例和简单	3量型实例					
全部实例	相关链接:产品首页 场景白	皮书 新功能:多元索引 🎯 🗌	新功能:通道服务 🐻				
审计日本	实例名称	实例规格	实例注释	状态	创建时间	监控	操作
活动页面	workshop-bj-001	高性能实例	workshop1	运行中	2019-06-14 17:27:02	R	管理 释放

2. 开通大数据计算服务MaxCompute

说明:

a) 进入MaxCompute产品详情页, 单击立即购买。

(-)	阿里云									
刮	产品	解决方案	定价	ET大脑	数据智能	安全	云市场	支持与服务	合作伙伴	
		大数排	計	算服約	务・M	axC	omp	ute 💿	观频简介	
	MaxCompute (原ODPS)是一项大数据计算服务,它能提供快速、完全托管的PB级数据仓库解决方案, 使您可以经济并高效的分析处理海量数据。欢迎加入钉钉交流群11782920。 激您参加 MaxCompute 问卷调研 赠送100元太产品无门槛代全卷 >>									
		立即购到	2	管理控	****	线	 F实验			

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。

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。

■ (-)阿里云	华东1(杭州)▼	Q 搜索			费	用 工单	留寫	企业	支持与服务	Þ.,	Ū,	Ä	0	ଛ	简体中文	0
DataWorks	DataWorks / 工作空间列表													产品	动态 帮	助文档
概范	创建工作空间 谐编入	工作空间/显示名 0.														С
工作空间列表	工作空间名称/显示名	模式	创建时间	曾理员		状态		Ŧ	画服务		操作					
资源组列表 计算引擎列表 ^	No. of Concession, Name	简单模式 <mark>Lite</mark> 单环境	2019-08-21 (11:13:05)	100410-0012		✔ 正常		V	V		工作的进入器	空间配置 数据集成	2 修改] 2 进入3	服务 进入 数据服务	数据开发 更多	
MaxCompute	Annali pel la	标准模式 开发跟生产隔离	2019-08-21 (10:40:57)	Real Property lies		✔ 正常		V	「「「		工作5 进入3	空间配置 数据集成	E 修改目 乾 进入調	服务 进入 数据服务	数据开发 更多	

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。

创建工作空间		
1 基本配置	2 选择引擎	3 引擎详情
基本信息		
* 工作空间名称	需要字母开头,只能包含字母下划线和数字	
显示名	如果不埴,默认为工作空间名称	
* 模式	简单模式(单环境)	\checkmark
描述		
高级设置		
* 能下载Select结果 🛛	π	
下一步	取消	
~		
道 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。
创建工作空间
基本配置 3 引擎详情
选择DataWorks服务
数据集成、数据开发、运维中心、数据质量 您可以进行数据同步集成、工作流编排、周期任务调度和运维、对产出数据质量进行检查等。
选择计算引擎服务
✓ MaxCompute ● 按量付费 ○ 包年包月 开发者版 开通后,您可在DataWorks里进行MaxCompute SQL、MaxCompute MR任务的开发。 充值 续费 升级 降配
 ☑。奕时计算 ○ 共享模式 ○ 独享模式 开通后,您可在DataWorks里面进行流式计算任务开发。
E-MapReduce 开通后,您可以在DataWorks中使用E-MapReduce进行大数据处理任务的开发。
选择机器学习服务
□ 机器学习PAI ○ 按量付费 开通后,您可使用机器学习算法、深度学习框架及在线预测服务。使用机器学习PAI,需要使用MaxCompute。
下一步

d) 进入引擎详情页面,填写选购引擎的配置。

创建工	工作空间					
\bigcirc	基本配置		⊘ 选择]擎	3	引擎详情
\sim	MaxCompute					
	* 实例名称					
	开发环境			生产环境		
	MaxCompute项目名称 🚱	test_dev		MaxCompute项目名称 🥝	test	
	MaxCompute访问身份	个人账号		MaxCompute访问身份 🔮	工作空间所有者	\sim
				* Quota组切换	按量付费默认资源组	\sim
ê li	建工作空间 上一步					

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。



🏷 Quick Bl 💟 专业版 😡		我的	工作空间 自	创作区 订阅	监控指标
Quick Bi していたいのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでのでの	最近相关 与我共享	我的收藏	暂时没有数据哦	А е, (tyt): tyt):	
 ● 默认空间 ● 创建:2017/12/13 10:02:12 □ 项目负责人: ■ 33人 ◆管理员 	创作流程 1. 获取数据	> 2. 数据建模		 > 3. 数据分析 (V表板 丰富的可视化组 	件支持
申请加入更多项目空间吧! 加入项目	数据源 连接云上及本地多 种数调库以及文件	数据集 简易的自助建模实 现复杂的数据模型		目助数据分析 电子表格 类Excel式的体验 杂行业报表	制作复

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。

<	🐟 workshop-bj-001
实例详情	实例访问地址
实例监控	私网: https://www.om 公网: https://www.initialized for a final fina

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。

道 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。

名称	修改日期	类型	大小
L conf	2019/6/17 10:07	文件夹	10.007 //0
worksnop_demo.exe	2017/12/18 16:58	应用程序	12,307 KB

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_d

emo.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。

workshop_demo;workshop_demo.exe prepare
ted table already exists.
ted table already exists.
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_acti ve'、'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	nloads\w	orkshop de	emo>v	vorkshop	demo.exe	raw	00010 '	2019-06-
5" 20		1 =						
uid		Da	ate	bytes	cust	tomiz	e_event	:
device		ip sta	atus	-	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	backs	stage	369)
iPad min2	157.249.67	. 241	200		ios11			
00010	2019-06-14	11:21:30	PM	427	bı	rowse	trave]	L
iPhone6s	222.133.108	. 234	200		ios10			
00010	2019-06-14	11:16:03	PM	764	SI	witch	185	5
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	SI	witch	73	3
iPhone6s	222.133.108	. 234	200		ios10			
00010	2019-06-14	10:06:29	PM	535	backs	stage	179	3
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	10	ogout	462	2
iPhone6s	222.133.108	. 234	200		ios10	-		
00010	2019-06-14	08:26:06	PM	887	coi	nment	funny	4
iPad min2	157.249.67	. 241	200		ios11		-	
00010	2019-06-14	08:10:34	PM	278	bro	owse	finance	è
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	bı	rowse	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	bi	rowse	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			



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>worksh	<pre>nop_demo.exe day_active</pre>
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)。

	ł	戰览 工作空间列表 资源	原列表 计算引擎列	俵		
华东2 华南1 华北2 香港 美西1	亚太东南1 美东1 欧洲中部1 亚	2太东南 2 亚太东南 3 亚太东北 1	中东东部1 亚太南部1	亚太东南 5 英国		
搜索						
工作空间名称/显示名	模式	创建的时间	管理员	状态	开通服务	操作
Repair Control of Cont	简单模式(单环境)	2019-04-23 16:47:31	10100	正常	Co	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼
	简单模式(单环境)	2019-02-26 14:15:17	10 million	正常	∞ 🔥	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼
Press (M)	标准模式(开发跟生产隔离)	2019-01-30 10:18:52	1011	正常	0。 🔨 본	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼
10.00	简单模式(单环境)	2019-01-10 13:46:08	1000	正常	0。 🔨 본	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼
La para a	简单模式(单环境)	2018-12-28 15:03:49	101-011	正常	∞ 🔥	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 マ
NUMBER OF STREET	简单模式(单环境)	2018-12-10 20:22:30		正常	Co	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼
bigdata_DOC bigdata_DOC	简单模式(单环境)	2018-09-02 10:26:59	10000	正常	∞ 🔨	工作空间配置 进入数据开发 修改服务 进入数据集成 进入数据服务 更多 ▼

双击之后,即可进入工作空间的数据开发界面。

\odot	X DataStudio biqda	ta_DOC ta_DOC	~						❷ 跨项目克隆	∂ 运维中心 🔍
Ш	数据开发 2 口(С 🕀 Ю	# rpt_user_trace_log	dw_user_trace_log	ds_user_trace_log	ts_user_trace_log	Sq query X	sq rpt_user_trace_log	Sq ods_user_tra	ce_log
(I)	Q 文件名称/创建人	™	🖱 🗳 🗗	li 🗇 🕲 🖸						
Q	> 解决方案	80								
G	▶ 业务流程	88								
	∨ 旧版工作流									
	› 🛅 任务开发			start for any longitudes						
≡										
∎										
fx										
Σ										
亩										

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。

6	💸 DataStudio	biqdata_DO bigdata_DO	C C		~
	数据开发	[‡ C O	শ্	🖽 rpt_u	iser_trace
(/)	Q 文件名称/创建人		T		Ē}
Q	> 解决方案				odps
6	▼ 业务流程			3	auth
\sim	› 🛃 123		新	建业务流移	
Ê	> 🖧 clone_databa	ase_datav_test	£	命业务流档	E有权
⊞	> 📇 elasticsearch	1_reader			select

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。

▦ 表				. .		tt ots_user_t	race_log ×	Sq query	Sq rpt_user_t	race_log
DDL模式	从生产环境加载									
	2	表名	ots_user_trac	e_log						
	写,	入该表的业务流程	Workshop							
基本属性										
	中文名	ots_user_trace_log								
	一级主题	请选择		1	二级主题	请选择		新建主题	C	
	描述									
物理模型设计 										
	分区类型	○ 分区表 🌘	非分区表							
	层级	请选择			物理分类	请选择		新建层级	C	
	表类型									
	选择存储地址:	tablestore://works	hop-bj-001.cn-b	eijing.ots.aliyuncs.co	om/					一键授权

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'
   行台/切秧用尸/砌宽<sup>+</sup>,
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。

数据开发 名日の	ك 🗈	表	🛛 🧮 rpt_us	er_trace_log	dw_user_trace_l
Q 文件名称/创建人	™	1			
▶ 解决方案 {		/ 节点组	C		
▶ 业务流程	品 ~	数据集成			
> 🛃 123 > 🛃 clone_database_datav_test	C	Di 数据同步			
> 📇 elasticsearch_reader	~	数据开发			
> 🚣 works		CODPS Scrip	t		
Y 🛃 Workshop [»)	odps sql			
> 😑 数据集成	6	DDPS Spark	c		
▼ (1) 数据开发	0	Ⅵ 虚拟节点			
Sa dw_user_trace_log ∄		Mr ODPS MR			
Sq ods_user_trace_log		sh Shell			
Sq rpt_user_trace_log ∄		Py PyODPS			
Vi Start 我锁定 06-171	0:58				

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。

1		€		»
	节点	a		C
	数据	集成		
Di	数据	司步		
~	数据	肝发		
Sc	ODPS	S Scrip	t	
Sq	ODPS	S SQL]_	
Sp	ODPS	S Spar	k	
Vi	虚拟	节点		
Mr	ODPS	S MR		
Sh Gu	Shell	ne		
Ру	PyOL	P5		
~	控制			
<mark>۶</mark> ۹	跨租	中节点		
Ch	ossX	橡检	查	

▋ 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。

▼ 业务流程		
> 🚣 123		
> 📇 clone_database_	datav_test	
> 📇 elasticsearch_rea	ader	
> 🚑 works		
🗸 🛃 Workshop	»	
> 럳 数据集成		
> ₩ 数据开发		
▶ 🔠 表		
> 💋 资源	新建资源 🔰	File
> 🗾 函数	新建文件夹	Archive
> 🧮 算法	看板	JAR
> 🞯 控制		Python

· 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。
新建资源		×
资源名称:		
目标文件夹:	业务流程/Workshop/资源	
资源类型:	File	
	✓ 大文件(内容超过500KB)	
	✓ 上传为ODPS资源本次上传,资源会同步上传至ODPS中	
上传文件:	点击上传	
	·····································	取消

上传完成后,请务必记得单击提交。

提交 上 <i>佳</i> 盗酒	
上に反応	
已保存文件:	ip.dat
资源唯一标识:	OSS-KEY-
	✓ 上传为ODPS资源本次上传,资源会同步上传至ODPS中
重新上传:	点击上传

・ 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。

新建资源		×
资源名称:	资源类型为JAR时文件名需要加后缀名.jar	
目标文件夹:	业务流程/Workshop/资源	
资源类型:	JAR	
	✓ 上传为ODPS资源本次上传,资源会同步上传至ODPS中	
上传文件:	点击上传	
	确定	取消

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。

🗸 🛃 Workshop	
> 😑 数据集成	
> 🕢 数据开发	
▶ 🔠 表	
> 💋 资源	
▶ 1 2 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。

	ע ר
提交。	
函数类型:	其他函数
函数名:	getregion
责任人:	dtplus_docs ~
类名:	odps.test.GetAddr
资源列表:	getaddr.jar,ip.dat
描述:	
命令格式:	getregion(ip string)
参数说明:	
返回值:	
实例 :	

- 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};
```

使用教程 / 2 搭建互联网在线运营分析平台

📕 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。

Q 文件名称/创建人	🖸 💿 🔍
/ Ⅲ 衣	· 世紀 C
> 💋 资源	
> 🔂 函数	◇ 数据集成
> 🧮 算法	
> ◎ 控制	
✓ 🛃 Workshop 🔊	◇ 数据开发
▶ 🔁 数据集成	Se ODPS Script
▶ 🗤 数据开发	Sej ODPS SQL
Sag dw_user_trace_log 我統	Sp ODPS Spark VI start
Sq ods_user_trace_log 我锁	「」「虚拟节点
Sag rpt_user_trace_log 我锁行 <	Mr ODPS MR Sq ods_user_trace_log
vi start 我锁定 06-17 10:58	Sh Shell
▶ 🗰 表	Py PyODPS Sq dw_user_trace_log
✔ 💋 资源	v #2##
Ja GetAddr.jar 我锁定 06-1.	Sq rpt_user_trace_log
Fi ip.dat 我锁定 06-17 10:4	
✔ 💤 函数	Ch oss对象检查

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。

提交						×
请选择节点	V V	节点名称 start ods_user_trace_b	log			
备注	 ✓ ✓ 28路線 	rpt_user_trace_lo 动物出不一致的告	og 蓉			
					 提交	取消

🗾 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。

e_log ×	: So dw_user_trac	æ_log ●	🕞 ip.da	at 🛛	■ 表	Vi star	t x	🛃 Wo	rkshop				<	≡
													运	推↗
×	调度配置 依赖上一周期:													调度配置
	调度依赖 🕐 —													版木
	自动解析:) 是	● 否	解析输入输出	H I									
	依赖的上游节点:	请输入父	节点输出	名称或输出表名			使用工作的	2间根节;	<u>ال</u>					
	父节点输出名称		热守父	俞出表名	节点名		父节点ID		责任人		来源	操作		
	bigdata_DOC_root				bigdata_doc_roo	ıt			dtplus_c	locs	手动添加			
	本节点的输出:	请输入节	点输出名	你										
	输出名称			输出表名	下游节点名称		下游节点ID		责任人		来源	操作		
	bigdata_DOC.5001	44927_ou	t	- Ø	ods_user_trace_	log			dtplus_do	cs	系统默认添加			
	bigdata_DOC.start	C		- C							手动添加			
:	节点上下文 😨	添加												
	编号 参数名		取	值来源		描述		兌	古点ID	来源	操作			
	没有数据													

2. 单击右上角的运维中心进行界面切换。

💸 DataStudio	bigdata_DOC bigdata_DOC						₽ 跨项目克隆	@运维中心
数据开发	≗₿С⊕⊎	Sq ods_user_trace_log	sq rpt_user_trace_log	🔄 dw_user_trace_log 🌑 🥅 ip.dat	表	Vi start x	A Workshop	
Q 文件名称/创建/	V V		🕆 C :					
থ <u>আ</u> ব⊼ ১ <u>আর</u> এভগল			×	调度配置				
 ✓ 山政 → <u>採</u>函数 				依赖上一周期:				
> 🔚 算法								
				油中片本 🥥				

 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。

⑤ 参 运维中心	data_DOC v data_DOC	🖉 Data
≡ ⑦ 运维大屏	検索: 市点名称/市点印 Q 解決方案: 演造择 ~ 业务添程 清选择 ~ 市点择型: 清选择 ~ 市点保型: 清选择 ~ 青在人: dtplus_docs	~
✔ 任务列表	基此 時选择 ✓ ✓ ● <td< th=""><th></th></td<>	
同期任务	□ 名称 ^{节点D} ^{节点D} 生产环境,请谨慎操作	
→ 任务运维	rpt_user_trace_Jog 1000408562	
同期实例	ods.user_trace.log 1000408514	
	dw_user_trace_log 1000408561 bigdata_doc_root	
○ 补数据实例	test 1000406709	
▶ 智能监控	mysql2odps_t1 1000312806 clone_database_datav_test_virtual 1000312804 《 >	
	bigdata_doc_root 1000077167 展开学节点 >	
	hdfs2mc2 28540 ods_use 重着代码	
	projection 4時間 市内 市内 第二日 査書次例 市点 市点	訊D: 1) \$称: st
	三百四四 《词度》	型:E
	Proggia → 当前节点 当前节点 更多マ く 1/1	<u>ع</u>
	「「「「」」「「」」「「」」」「「」」」「「」」」「「」」」「「」」」「「」」」「」」」「「」」」「」」」「」」」「」」」「」」」」	

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。

补数据		×
* 补数据名称:	P start 20190619 155104	
*选择业务日期:	2019-06-11 - 2019-06-17	
* 是否并行:	不并行	
*选择需要补数据的节点:		
✓ 任务名称	安名称进行搜索 Q	任务类型 7
Join State Digdata_DOC(1	485)	
🗸 start		虚节点
ods_user_trace	_log	ODPS_SQL
dw_user_trace_	log	ODPS_SQL
rpt_user_trace_	log	ODPS_SQL

取消

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。

=								
	搜索: 1000408559 Q. 补数3	諸名称: 请选择	▼ 节点类型: 请送	峰 > 责任	E人: 请选择责任人	~ 运行日期: 2019	-06-19	
▼ 任务列表	业务日期: 请选择日期 苗 基約	请选择	∨ 3的节点	重置清空				
局期任务								C 刷新 收起搜索
(1) 手动任务	实例名称	状态	任务类型	责任人	定时时间	业务日期	开始时间	操作
任各运维	 P_start_20190619_155104 	④运行中						批量终止
	- 2019-06-11	⊘运行成功				2019-06-11		
■ 「 割 手动实例	start	⊘运行成功	虚节点	dtplus_docs	2019-06-12 00:23:00	2019-06-11	2019-06-19 1	DAG图 终止运行 重跑 更多 🔻
副演奏例	+ 2019-06-12	●运行中				2019-06-12		
补数据实例	+ 2019-06-13	⊖未运行				2019-06-13		
▶ 智能监控	+ 2019-06-14	◎未运行				2019-06-14		
	+ 2019-06-15	◎未运行				2019-06-15		
	+ 2019-06-16	⊖未运行				2019-06-16		
	+ 2019-06-17	◎未运行				2019-06-17		

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。



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;

Sq query		×	Sq od:	s_user_tra	ace_log	5] rpt_user_	trace_log	Sq dw_user_tra	ace_	_log 🌒 📄 ip.dat	■ 表	Vi start		📲 Worksho	p		<
	Ē)	ſ	لة]	\$	lacksquare												
1 2 3 4 5 6	od ** cr ** sele	ps s eate **** ct *	ql ***** time ***** fron	2:2019- ******** n rpt_u	06-19 1 *******	**** 16:0 ****	******** 3:48 ******** log wher	******** *****************************		:**: 5 11	********** * and `20190617	' limit 1000;					i in an	
																		₹
运行	日志		ŝ	吉果[1]	×													\$ E
		А			в		С		D		E		G					
1 0	ountry		~	province		~	city	~	device_brand	~ (device 🗸	system_type 🗸 🗸	customize_event 🗸	use_tin	ne 🗸	customiz	e_event_c 🗸	pv
2 中	国			山东			菏泽		meizu		MEIZU PRO7	android	switch			news		23
3 捌	隇			挪威					iphone	i	iPhone6	ios	switch	6		travel		30
4 朝	润			韩国					ipad	i	iPad4	ios	switch	5		travel		30
5 👎	·Ξ			山东			菏泽		iphone	i	iPhone7 Plus	ios	switch	5		travel		17
6 捌	戚			挪威					xiaomi		XIAOMI Note3	android	switch	4		travel		31
7 朝	国			韩国					iphone	i	iPhone6	ios	switch	5		travel		31
8 4	国			山东			菏泽		iphone	i	iPhone7	ios	switch	4		travel		37
9 捌	隇			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控制合。

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。

😍 Quick Bl V 专业版 🧕		我的	工作空间	创作区	订阅	监控指标
Upper State 加入更多 + 工作空间 加入更多 + ・ ヘ人空间 ・ ・	最近相关 与我共享	我的收藏	暂时没有数据	щ , ķţţ开ź	使用吧 !	
 2017/12/13 10:02:12 项目负责人: 	创作流程 1. 获取数据	> 2. 数据建模		> 3. 数据	分析	
33人 《管理员	12	ŵ	4	(火表) 丰富的 自助费	扳 9可视化组件支 数据分析	诗
11 申请加入更多项目空间吧! 加入项目	数据源 连接云上及本地多 种数据库以及文件	数据集 简易的自助建模实 现复杂的数据模型		电子: 类Exc	表格 el式的体验制f V报表	複

2. 选择数据源 > 新建数据源 > 云数据库 > MaxCompute。

😍 Quick Bl 📎 रुभक्ष (<u>ģ</u>	我的工作空间(训作区 订阅 监控指标		L C C C
:=	数据源				+ 新建数据源
🛝 默认空间 🔍 -	添加数据源 云数据库 自	建数据源			× 即席分析SQL 上传文件
💷 数据门户	来自云数据库				操作
11 仪表板					G (1)
惑 电子表格	5.8	*	*	<u></u>	* 0
		MySQL	<i>66</i> 9	\sim	
☆ 数据集	MaxCompute	MySQL	SQL Server	AnalyticDB	•• U
数据填报 (公测) 4844					0
↔ 数据源					a ()
	م ج ک	0	* (P		÷ 0

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。

	✓ 数据源连通性正常!						
添加MaxCompute数据源							
* 显示名称:	bigdata_DOC						
* 数据库地址:	* 数据库地址: http://service.odps.aliyun.com/api						
* 项目名称:	bigdata_DOC						
* AccessKey ID:	LTAIF24u7g						
* AccessKey Secret:							
① 温馨提示:新增数排	居源存在同步延迟的情况,请稍候片刻。						
	关闭 连接测试	添加					

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

数据源				+ 新建数据源
我的数据源	Q 共10个文件	数据表 上传文件	Q 共29个文件	即席分析SQL 上传文件
The state of		名称	备注 ≑	操作
C		output_table_154812606367001adfb67dfee		î ()
and the company.		output_table_1548127871928c581534cd65e		ũ ()
2 101,8280 1887 - Landard March		output_table_154822526534301a9a11d2c44		ŵ ()
and the second second		output_table_154838757651615e9882605fe		a ()
D -		output_table_1548743722536a9cd797ed51f		ū ()
and the second second		result_table		♀_① 创建数据集
Carlo Constanting		rpt_user_trace_log		D
The second second		sale_detail		11 O
bigdata_DOC 所有者:		system_9c676e75c4324f75b5430dd138a738ab		ū ()

选择您想放置的数据集位置,单击确定。

创建数据集				×
	* 名称:	rpt		
	* 位置:	ODPS		~
			÷	关闭 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。

😍 Quick Bl 🔮 专业版	😍 Quick BI 😢 专业版 🧟					
:=	数据集 全部 我的					
🔊 默认空间 🛛 🗢 🗸	根目录 > ODPS					
💷 数据门户	名称 👙					
■■ 仪表板	rpt NEW					
📚 电子表格	rpt_user_trace_log					
★ 数据集	*					

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。

			Str. S	Str.	Str.
维度		+ 🎚	country c	dt	province
Str.	count	ry			
Str.	dt	命			
Str.	provi				
Str.	city	C 克隆维度			
Str.	devid	❤ 左公析山陷碛			
Str.	devid				
Str.	syste	◎ 取消全部隐藏			
Str.	custo	★ 删除			
Str.	custo	~			
		+ 新建计算字段(维度	度)		
		A 移动到	•		
ст. П		£ 新建层次结构		-	
度重 - 🗁	默认	↑ 上移		yyyymmaa nn:m	1:55
ŀ	l⁰ use	↓ 下移		yyyy/MM/dd hh:	mi:ss
N	lo b∧			yyyy-MM-dd hh:	mi:ss
N	l⁰ uv	↓ 转换为度量		yyyy/MM/dd	
		➡ 维度类型切换	▶ 日期(源数据格式)		
				ууууммаа	
				yyyy-MM-dd	
			→ 文本	voov/MM	
			数字	y y y y / I'II'I	
				ууууММ	

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。

 iii dt(year) iii dt(quarter) iii dt(month) iii dt(day) iiii dt(day) iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii	维度 * 品品 al	+ 🏼 cou	untry	dt(day)	pr
 iii dt(quarter) iii dt(month) iii dt(av) Str. city Str. city Q: 売隆维度 A: 在分析中隐藏 A: 在分析中隐藏 A: 在分析中隐藏 A: 在分析中隐藏 A: 在分析中隐藏 A: 在分析中隐藏 A: 和強计算字段(维度) A: 移动到 A: 移动到 A: 移动到 A: 移动到 A: 移动到 A: 移动到 A: 新建法算字段(维度) A: 移动到 A: 下移 D: 許換为度量 - · · · · · · · · · · · · · · · · · · ·	🛗 dt(ye	ar)			
 	🛗 dt(qu	arter)			
 iii dt(week) iii dt(day) iii d	🛗 dt(ma	onth)			
 fit dt(day) \$tr. city st. city \$tr. device st. device \$tr. device st. device \$tr. device st. system \$tr. custor \$tr. custor \$tr. custor \$tr. custor \$tr. custor \$tr. custor \$tr. prov \$tr. custor \$tr. custor \$tr. bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright: \$fr: bright:	🛗 dt(we	ek)			
Str. provint	🛗 dt(da	у)			
Str. city	Str. provine				
Str. device ① 克隆维度 Str. device → 在分析中隐藏 Str. custor ● 取消全部隐藏 Str. custor × 删除 # 新建计算字段(维度) ● Nº use ● 移动到 Nº use ● 新建层次结构 ● 上移 □ ↓ 下移 □ □ 宇梭为度量 □ ↓ 生態 □ ↓ 生態 □ ● 维度类型切换 ● ● 文本 经度	Str. city	⊘ 编辑			
Str. device Str. syster	str. device	2 克隆维度			
Str. system Str. custom	str. device	▶ 大公城市陶蔵			
Str. custom ● 取消全部隐藏	^{Str.} system	* 往刀仰中脑腿			
Str. custon × 删除	Str. custorr	● 取消全部隐藏			
g量 + 新建计算字段(维度) Nº use_ & 移动到 Nº pv ※ 新建层次结构 Nº uv ① * 上移 区域 ↓ 下移 □ 小 转换为度量 □ ↓ * 转换为度量 ○ ↓ * 支车 经度	Str. custor	★ 删除			
 + 新建计算字段(维度) N² use_1 & 移动到 N² pv S 新建层次结构 ① 上移 ① 下移 ① 「 卡移 ① 「 市 ① 「 均方度量 ① 「 4度类型切换 ○ 文本 	度量				
Nº use_ ▲ 移动到 」 Nº pv チ 新建层次结构 Nº uv チ 上移 ◆ 下移 「 」 转换为度量 日期(源数据格式) ★ 文本 经度	- 🗁 默认	+ 新建计算字段(维度)			
№ pv ① 新建层次结构 国家 □ ↑ 上移 区域 □ ↓ 下移 □ □ □ ↓ 下移 □ □ □ ↓ 转换为度量 □ □ □ ↓ 转换为度量 □ □ □ ↓ 軟度类型切换 √ 文本 经度 □	[№] use_t	よ 移动到	•		_
 № uv ↑ 上移 ↓ 下移 ↓ 「转换为度量 ↓ 维度类型切换 ↓ 並理信息 ↓ 公本 经度 	[№] pv	。 新建民次结构		国家	
 ↑ 上移 ↓ 下移 ↓ 「转换为度量 ↓ \$\$\$#\$	Nº uv	出机建运入印码		区域	
 ↓ 下移 ↓ 转换为度量 ↓ 维度类型切换 ↓ 维度类型切换 ↓ 文本 ④ (/ 直辖市) ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○		↑ 上移			
 ↓ 转换为度量 ↓ 堆度类型切换 ↓ 堆度类型切换 ↓ 文本 经度 		1 下移		省 / 直辖市	
 ↓ 转换为度量 地理信息 人 文本 经度 		V · · · ·	日期(源数据格] 市	
→ 维度类型切换 → → ▲		↓ 转换为度量			
		、维度类刑切场	地埋信息	▶ 区/县	
			- 🗸 文本	经度	

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。

🤇 🛱 rpt							Û		保存
数据集 = 0 いんンギリロンIPで	۵ ا								件 同步表结构
维度 +	e province	eity.	ar.	dt(day)	ax device brand	ar. device	sk.	sk.	s.
- 🗁 province_层级结构	province	city	country	((ddy)	device_brand	Gevice	ayatem_type	customize_event	customize_event
province 3									
👳 city									
Str. country									
<u> </u>									

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。

:=	数据集 全部 网的		名称	✓ Q 共17个文件	+ 新建数据集
🛝 默认空间 🛛 🗢 👻	根目录 > ODPS				
■ 数据门户	名称 👙	创建吉 💠	修改人/修改时间	数据源	9570/27145
(以表板	rpt wew rpt_user_trace_log	Furui	-	bigdata_DOC MaxCompute	
◎ 电子表格		四百濟	-	ECS_QBI_DB	
● 自助取数 (公測) мен/	No. Spectropol, and prove	~~ <u>01000 580</u>		SQL Server	
育 数据集	F	冯国章		workshop MySQL	⊠ al ≋ :

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。

<	.11	未命谷	3			*			指	标看机	⊽	
ļά		ି		1	<u>i</u>	*	٩		#	Ē	Ô	¢,
		ot										
		int #	标着板	デ 缶丸ノレダ	往度式	唐雷	而	×				
		15		с. щ > =	·æ/义卒/	》又里						
									-			
	旨板- 内	ot										
		int 指	标看板	ī: 缺少约	维度或	渡量	项					
									-			

 ・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_t race_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。

			图表设计		
			数据		
			选择数据源类型:	◎数据集	
г	设置过滤器	×	看板指标/度量		
	₽rpt	🔟 dt(year)	Nº 📰 pv		
			看板标签/维度	⇒ ∰ d	t
	区间粒度	🖌 日期粒度	双击或拖动数据字段	至此处	dt(year) dt(quarter)
L	 ● 時间区间 ■ 1 2019 ■ 1 	019 📾	过滤器		
L	● 开始于:	-•• 结束于:	dt(year)	ਲ਼× Str. d	
	相对时间 🗸	相对时间 🗸			
	2019	2019			
L	T+ ▼ 0	T ♠ + ▼ 0		● 展量	
L		取消 确定			use_time pv
			7711/5-4-1000		
			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。

T 2	-	🛃	ψ,	*		Å	6
pv Furul						÷	
10.69万 ^{euroi}			FURU				
uv Furui							
3800 eurol							

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)

(戦略) (大) (利用) (利用) (利用) (利用) (利用) (利用) (利用) (利用			© « โ	编辑 预览	保存
🏒 փ. 🖀 💌 🏛	A 🗉 🙆 🧶 🐺 🕭) 🗮 🚍 🦮		图表设计	💉 更改图表类型 -
				数据	样式 高级
Fund	趋势图	-pv Furth-uv	Furth	选择数据源类型:	◆数据集 ■
	275			值轴/度量	rpt v E
	1.6万 ^{尺山山} 1.2万 太 8000	Furd	Fund	110 ■ pv ⇔×	Q 输入关键字按索 () 维度 - 品 dt
	ورير <i>رياً</i> 4000 0	FUT ^{UÅ} 20190612 20190613 20190614 20190615 ;	F.J.T.Å 0130616 20130617	关别地/维度 ☐	iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
				颜色图例维度 拖动数据字段至此处	Str. device_brand Str. device Str. system type
				过始器	Str. customize_event IE最 - う 武认 N2 use_time N2 py
				预选行数 1000	
				更新	

 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。

< al RPT 🔺						1911 1923	保存
pv eved		植物画 での ^の		_{est} o≓pv –uv			
10.69万		555 50 ⁻⁰¹ 1675 1275	Eurid	Eurol	Rain		
uv esta		9000 우네네 4000			Fund		
3800 ***		_{१उ} र्ज 0 201	0611 (20190612 :	20190613 (13 ¹³⁾ 20190614	20190615 / ^{.d} 20190616	20190617	
色彩地图 194 ⁻⁰		est ∎ PV				Ford	
						++ 	
			535222				
		Las B	And the second				
Portal 3.76675 - 3.99275 3.5475 - 3.76675		mart	when the way				
3.31477 - 3.5477 3.08877 - 3.31477 2.86277 - 3.08877							

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。

Vi start	0
Ļ	
Sq ods_user_trace_log	0
Sq dw_user_trace_log	0
Sq rpt_user_trace_log	0

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。

🔠 ods_	user_trace_log	dw_u	user_trace_lo	og Sq	dw_user_trace_log	Sq ods	_user_trace_log 🌒	Fx getregion	Ja
		ه 🔓	\$) :					
	odps sql		******		*****		****		
	author:								
	create t	ime:2019-	07-02 17	32:09					
	********	*******	*******	*******	************	********	*************	-	
	insert <u>sel</u>	ect overw	rite tabl	le ods_us	<u>er trace log</u>	partition	(dt=\${bdp.sys	stem.bizdate})	
	select								
	md5,								
	uid ,								
	ts,								
	ip,								
	status	,							
	bytes,								
	device	,							
	system	,							
	custom	ize_event	,						
	use ti	me,							
	custom	ize event	content						
	from o	ts user t	- race log						
	where	to_char(F	ROM_UNIXT	IME(ts),	'yyyymmdd')=٩	{bdp.syst	em.bizdate};		

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。



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监控制	「范		
		监	拉分类		数据量	主键	离散值	汇总值	业务逻辑、规则
					所有非临时表都建议配	对于存在业务主键、逻辑主	维表、事实表中的维	汇总统计表中的汇总	1、重要指标的异常值监控。例如,正常UID长度
					置该项监控。	键的表需配置该监控。	度值、状态值、可枚	值需配置该监控。	是否为32位。
							举的值需配置该监控		2、字段间的平衡值监控。例如,字段a与字段b满
		這	[台场景				•		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
					表行数波动/自助规则表	模板规则的字段空值、重复	离散值分组个数/离	模板规则的单字段大	自定义规则
					行数>固定值	值/自定义规则监控联合主	散值分组个数波动/	于0/自定义规则判断	
		常用	监控规则			键空值、重复值情况	离散值状态值波动	字段等于0所占的比例	
								4	
层次			表类型				规则	記堂	
			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
		A2		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
			3	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
			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	不涉及
		A3	10 2000	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
ODS/DWD	赢 线 表			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
000, 010	Ind souther		恼暑夷	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
		A4	11 2 11	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
			1	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
			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
		Ax		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
			4	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
			松县地	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
		A2	相里衣	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
			4	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
			松島市	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
		A3	相里衣	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
DWC (ADC	± 100 40 ≠		4	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
DWS/ADS	商线衣		茶田市	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
		A4	相里衣	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
			做具本	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
		Ax	相重农	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
			4	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。

🜀 💥 DataStudio	workshop_DOC workshop_DOC	
☰ 全部产品	●〉 数据汇聚	
○ 数据集成	Co数据集成	
X DataStudio(数据开发)		
送 运维中心(工作流)	数据开发	
数据地图(数据管理)	💸 DataStudio(数据开发)	
	→ 数据服务	
	🗲 Function Studio New	
	畫数据分析	
	任务运维	
	🤔 运维中心(工作流)	
	跨项目克隆	
	数据治理	
	◎ 数据质量	Ð

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。

	⑤ Ø 数据质量	workshop_DOC workshop_DOC			
			表列表		
	→ DQC监控	MaxCompute ~			
l	品 概览	请输入 🔍	请输入表名进行搜索 🔍		
	吕 我的订阅	workshop_doc	表名	责任人	操作
	自 规则配置		ots_user_trace_log	dtplus_docs	配置监控规则
	① 任务查询		ods_user_trace_log	dtplus_docs	2 配置监控规则
1			dw_user_trace_log	dtplus_docs	配置监控规则
			rpt_user_trace_log	dtplus_docs	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。

=	
▼ DQC监控	
品 概 览	規則配置 > 应用名:workshop_doc > 表名:ods_user_trace_log > 分区表达式:dt=Styyyymmdd-11 关联调度 ▲
吕 我的订阅	
● 规则配置	
① 任务查询	- dt=
	分区表达式: dt=S[yyyymmdd-1] 计算
	计算结果: dt=20190703
	调度时间: 2019-07-04 14:51:43
	*

d) 单击创建规则。

⑤ ② 数据质量	workshop_DOC vorkshop_DOC v											ළු 👳 dtp	lus_docs
≡	● 规则配置 规则配置 > 应用名:workshop_dk	oc 〉 表名:ods_use	er_trace_log 〉 分[区表达式	: dt=\$[yyyymmdd-1	美联调度 △							
品 概念						[
王 邦的订阅	已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:dtplus_d	ocs					试跑 订阅管理	创建规则	更多・
創 規則配置	+	470 This America	1001-0-00	30	4777741444147	an and a loss of the		atte da Atte da	tan da 270 kita	40.45	7000 (10.10	-
① 任务查询	- dt=S[yyyymmdd-1]	规则名称	规则学段	599	规则视版	动态阈值	比級力式	橙色圆值	红色网灿	周月端於加加	BCNR人	操作	
							没有	数据					

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。


〉 表名:ods_u	ser_trace_log 〉 分	区表达式	: dt=\$[yyyymm	模板规则 自定义	現则				
模板规则(0)	自定义规则(0)		责任人:dtp	ŀ	+ 添加监控规则		+	快捷添加	
规则名称	规则字段	强	规则模版	* 规则名称 :	0DS表行数大于0		*强弱:	• 强 .	5
				*规则字段:	表级规则 (table)	\sim			
				* 规则模版 :	表行数,固定值				\sim
				*比较方式:	大于	\sim	*期望值:	0	
				描述:					
								批量保存	取消

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。

+	· 添加监控规则	+ 快捷添加
*规则名称:	ODS表重复数据	*强弱: 🖲 强 🔵 弱
*规则字段:	ts (bigint)	
★规则模版:	重复值个数,固定值	
*比较方式:	等于 🗸	★ 期望值: 0
描述:		

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。

+	添加监控规则	+ 快捷添加
* 规则名称 :	ODS空值数据	*强弱: • 强 35
*规则字段:	uid (string)	
* 规则模版:	空值个数,固定值	\checkmark
* 比较方式:	等于 ~	*期望值: 0
描述:		

h) 批量保存规则。

完成上述操作后,单击批量保存。

	+ 添加监控规则	+ 快捷添加
*规则名称:	ODS空值数据	*强弱: 💿 强 🔵 弱
*规则字段:	uid (string)	
*规则模版:	空值个数,固定值	
*比较方式:	等于 🗸 🗸	*期望值: 0
描述:		
描述:		
描述:		
描述 : * 规则名称 :	ODS表重复数据	*强弱: 〇 强 • 弱
描述: * 规则名称: * 规则字段:	ODS表重复数据 ts (bigint)	*强弱: 〇 强 💿 弱

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。

〉 表名:ods_user	_trace_log 〉 分区	表达式:	dt=\$[yyyymmdd-1]	关联调度 ⚠						
模板规则(3)	自定义规则(0)		责任人:dtplus_docs	5				试题	订阅管理	创建规则更多
规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作
ODS表行数大于 0	表级规则	强	表行数 , 固定值	否	大于	-	-	0	101.011	修改 删除 日志
0DS表重复数据	ts	弱	重复值个数 , 固定 值	否	等于	-		0	Are 111	修改 删除 日志
0DS空值数据	uid	强	空値个数,固定値	否	等于			0	April 1941	修改 删除 日志

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。

试跑	>	<
试跑分区:	dt=S[yyyymmdd-1]	
调度时间:	2019-07-04 22:08:54	
	试跑	
	一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一一	
	反关	

在弹出的页面中, 您可以查看表数据是否已符合您的规则。

Ξ	灾伤心关键														
▼ DQC监控													E126		
品 概览	应用 :workshop.	.doc 現名:ods_	user_trace_log >	dt=S[yyyymmdd-1	1] 💷 20	019-07-04 22:09:	01 更多								MGGYI
吕 我的订阅	规则名称	規則字段	强/弱	采样方式	过滤条件	校验类型	校验方式	比较方式	橙色阈值	红色阈值	期望值	历史结果	采样结果	状态	操作
自規則配置	ODS表重复 数据	ts	80	table_count- count_distinct	-	数值型	-	等于	-	-	0	-	0	正常	查看历史 结果
① 任务查询	0DS空值数 据	uid	强	null_value	-	数值型		等于		-	0	-	0	正常	查看历史 结果
	0DS表行数 大于0		强	table_count		数值型		大于			0		0	红色异常	<u> 書</u> 看历史 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。

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。

≡ ✔ DQC监控 品 概览	知则配置 規则配置 > 应用名:workshop_do	oc 〉 表名:ods_use	r_trace_log 〉 分	区表达式	: dt=\$[yyyymmdd-1]	【关联调度】 ⚠			
我的订阅	已添加的分区表达式	模板规则(3)	自定义规则(0)		责任人:dtplus_doo	s			
■ 元代前正	- dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色
		ODS表行数大于 0	表级规则	强	表行数 , 固定值	否	大于	-	-
		ODS表重复数据	ts	弱	重复值个数 , 固定 值	否	等于		
		ODS空值数据	uid	强	空值个数,固定值	否	等于	-	

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。

С	〉 表名:ods_use	r_trace_log 〉 分[☑表达式 :	dt=\$[yyyymmdd-1]	关联调度						
[模板规则(3)	自定义规则(0)		责任人:dtplus_doc	s				试	泡 订阅管理	创建规则 更多,
	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作
	ODS表行数大于 0	表级规则	强	表行数 , 固定值	否	大于	-	-	0	dtplus_docs	修改 删除 日志
	ODS表重复数据	ts	弱	重复值个数 , 固定 值	否	等于			0	dtplus_docs	修改 删除 日志
	ODS空值数据	uid	强	空値个数 , 固定値	否	等于		-	0	dtplus_docs	修改 删除 日志
	ODS表重复数据 ODS空值数据	ts uid	弱强	重复值个数,回定 值 空值个数,固定值	否	等于			0	dtplus_docs dtplus_docs	修改 删除 日志 修改 删除 日志

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。

ち 規則配置 規則配置 加則配置 シ 应用名:workshop_doc > 表名:ods_user_trace_log > 分区表达式:dtsSlyyyymmdd-1] 実際調度											
已添加的分区表达式	模板规则(3	3) 自定义规则(0)		责任人:dtplus_d	ocs				i	式跑 订阅管理	創建規则 更多・
- dt=\$[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作
	ODS表行 0	订阅管理					×	-	0	10000	修改 删除 日志
	ODS表重	订阅方式		接受对象		操作			0	-	修改 删除 日志
	ODS空催	邮件通知		Approxime 1		修改删除		-	0		修改 删除 日志
		邮件通知 ∨		请选择	\sim	保存					
							关闭				

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。

6	◎ 数据质量					ಲ್ಗೆ
•	≡ DQC监控	我的订阅		分区表达式	责任人	操作
88	概览	MaxCompute 〜 清船入表名 〇	清空	dt=\$[yyyymmdd-1]	10010-000	上次结果 通知方式▼
8	我的订阅	MaxCompute项目 分	表名			
â	規則配置	•	ods_user_trace_log			
₫	任务查询					

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。

E DQC监控 BB 概览	 ・ 規则配置 ・ ・ ・	oc 〉 表名:dw_us	er_trace_log 〉 分区	【表达式	: dt=S[yyyymmdd-1	美联调度 企						
吕 我的订阅	已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:dtplus_	docs					试跑 订阅管理	创建规则
■ 規则配置 ① 任务查询	- dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色调值	红色阈值	期望值	配置人	操作
							没有	敗据				

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。

5 規则配置 規则配置 > 应用名:workshop_do	oc 〉 表名:dw_use	r_trace_log 〉 分区	【表达式 :	dt=\$[yyyymmdd-1]	关联调度						
已添加的分区表达式	模板规则(2)	自定义规则(0)		责任人:dtplus_doo	:5				试	跑订阅管理	创建规则更多、
- dt=S[yyyymmdd-1]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作
	CDM表行数大 于0	表级规则	强	表行数 , 固定值	否	大于	-		0	10/10/100	修改 删除 日志
	CDM表空值数 据	uid	36	空值个数,固定值	否	等于			0	10110-001	修改 删除 日志

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

-	+ 添加监控规则			+ 快捷演	忝加	
* 规则名称 :	CDM表行数波动率			*强弱: 💿 强	5 5	
★规则字段:	表级规则 (table)	\sim				
*规则模版:	表行数,上周期波动率					\sim
*比较方式:	绝对值	\sim				
波动值比较:	0%	25%	50%	75%		100

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。

描述:

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	
dw_user_trace_log (700002549	214) ×	
		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。

5 规则配置								
规则配置 〉 应用名:workshop_do	oc 〉 表名:rpt_use	er_trace_log 〉 分区	表达式	: dt=\$[yyyymmdd]	关联调度 🛆			
已添加的分区表达式	模板规则(0)	自定义规则(0)		责任人:dtplus_	.docs			
- dt=\$[yyyymmdd]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值
						没有到	牧据	

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。

5 規则配置 規则配置 > 应用名:workshop_do	c 〉 表名:rpt_user	_trace_log 〉 分区	【表达式 :(ft=S[yyyymmdd]	关联调度						
已添加的分区表达式	模板规则(3)	自定义规则(0)		责任人:dtplus_doc	s					询问管理	创建规则更多・
- dt=S[yyyymmdd]	规则名称	规则字段	强	规则模版	动态阈值	比较方式	橙色阈值	红色阈值	期望值	配置人	操作
	ADS表行数波动 率	表级规则	弱	表行数 , 上周期波 动率	否	绝对值	10%	50%	-	dtplus_docs	修改 删除 日志
	ADS表行数大于 0	表级规则	弱	表行数 , 固定值	否	大于			0	dtplus_docs	修改 删除 日志
	ADS表空值为0	pv	88	空値个数,固定値	否	等于			0	dtplus_docs	修改 删除 日志

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

1 1 *∘									
5 规则配置					模板规则 自定义共	RQU			
见则配置 > 应用名:workshop	_doc 〉 表名:rpt_us	ser_trace_log > 🖇)区表达元	t:dt=S[yyyymmc					
已添加的分区表达式	樽板抑則(3)	自完义抑则(0)		書任人:dto	2	· 添加监控规则		+ 快捷添加	
+	Beachers (5)			MITV: ub					
 dt=S[yyyymmdd] 	规则名称	规则字段	强	采样方式	*规则名称:	监控异常PV	* 强弱	: 🗌 强 🧕) 55
					*规则字段:	pv (bigint)	* 采样方式	: sum	\sim
					过滤条件:	此处请输入where后的条件,无须输入where			
					* 校验类型:	数值型	*比较方式:	大于	\sim
	*				* 校验方式:	与固定值比较	\sim	* 期望值:	100
					描述:	1			
								3	
								批量保存	取消

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

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 V 任务节点名称	∨ 添加
<code>rpt_user_trace_log</code> (700002549215) \times	
	关闭

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)。

6	X DataStudio	wor wor	kshop_D(kshop_D(DC DC		~
≡	全部产品	• >	€ €	€1	🖽 ots_u	iser_trace
				V.		⊡
Co	数据集成					odps
×	DataStudio(数据开发)			88		****
	运维中心(工作流)	×			4	crea
		2			6	select
େ	数据地图(数据管理)					select
					8	select
					9	

单击智能监控 > 规则管理。关于规则管理的详情请参见#unique_77。

⑤ * 运维中心										Ø	DataStudio
≡ ① 运维大屏	规则管理										
▶ 任务列表	规则名称:	请输入规则名称	规则对	İ象 : 规则	对象名称/ID	ť	创建人:	dtplus_docs		接收人:	请选择接收
▶ 任务运维	触发条件: 🔽 🕫	宅成 🔽 未完成 🔽 出错	<mark>- 7</mark> 周期未完成 <mark>-</mark> 超日	せ 捜索							
- 智能监控 1	规则	IID 规则名称	创建人	类型	对象类型	规则对象	角生	发条件	报警方式	接收人	
日 規則管理 2	26314	4 节点孤立报警		全局规则	任务节点	所有节点	-		短信,邮件	任务责任人	
△ 报警信息	2631	5 节点成环报警		全局规则	任务节点	所有节点			短信,邮件	任务责任人	

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。

规	则名称:	Workshop业务流移	呈监控			
对	象类型:	业务流程				~
规	则对象:	序号	业务流程	责任人	工作空间	
		1	Workshop	-	workshop_DOC 🖁	除
		请输入业务流程名	称/ID			•
触发方式——						
触	发条件:	超时	~ 💿	受监控的任务在打	指定时间后仍未正常结束 ×	
开始	运行起:	30	分钟			
段警行为──						
最大报	警次数:	3	次			
最小报	警问隔:	30	分钟			
免打	「扰时间:(00:00 至 00:00		0		
报	警方式:	✓ 短信 ✓ 邮件				
	接收人: (• 任务责任人				
		○ 其他 请输入指	制收人名字/ID	~		
		(+)				
钉钉郡	¥机器人:	@所有人	Webhook地址			操作
						保存

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。

基本信息	L						
	规则名称:	ADS层任务监控					
	对象类型:	任务节点				~	
	规则对象:	序号	任务名称	责任人	工作空间		
		1	rpt_user_trace_l og	dtplus_docs	workshop_DOC	删除	
		请输入任务节点名称	:/ID				Ð
触发方式							
	触发条件:	出错	~ ?				
报警行为]			1			
最	支报警次数:	3	次				
륲	妙报警间隔:	30	分钟				
	免打扰时间:	00:00 至 00:00		0			
	报警方式:	🗸 短信 🔽 邮件					
	接收人:	• 任务责任人					
		○ 其他 请输入接到	女人名字/ID	~			
		G					
Ŧ]钉群机器人:	@所有人	Webhook地址			操作	
						保存	

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。

■ (-)阿里云	③ 全球		Q 搜索			费用	工单	备案	企业	支持
DataWorks	常用工作空间							:	全部工作	空间
概览	stage and day			华东1	14 C				华	练2
工作空间列表 资源组列表	创建时间:2019-0 计算引擎:MaxCo 服务模块:AppStu)8-21 11:13:05 mpute udio 数据开发 数据管理 数	腐服务 运维中心		创建时间:2019-09-04 16:43:37 计算引擎:MaxCompute 服务模块:AppStudio 数据开发 数据管理 数	据服务 运线	主中心			
计算引擎列表 个	. If	F空间配置	进入数据开发		工作空间配置		进入数	数据开发		
MaxCompute	进入	、数据服务	进入数据集成		进入数据服务		进入数	改据集成		
交互式分析(Hologres)	1000,000			华北2	-				4	痈1
	创建时间:2019-0 计算引擎:MaxCo 服务模块:AppStu	8-30 20:33:29 mpute PAI计算引擎 udio 数据开发 数据管理 数	据服务 运维中心		创建时间:2019-08-13 16:11:39 计算引擎:MaxCompute 服务模块:AppStudio 数据开发 数据管理 数	据服务 运线	建中心			
	Iť	F空间配置	进入数据开发		工作空间配置		进入数	数据开发		
	进入	、数据服务	进入数据集成		进入数据服务		进入数	数据集成		
	常用功能	间 × 一键CDN								

您也可以进入工作空间列表页面,单击创建工作空间。

☰ (-)阿里云	华东2(上海) 🔻	Q 搜索			费用	工単	备案	企业	支持与服务
DataWorks	DataWorks / 工作空间列表								
概览	创建工作空间 请输入]	[作空间/显示名 Q							
工作空间列表	工作空间名称/显示名	模式	创建时间	管理员		状态		开	画服务
	-	简单模式 单环境	2019-09-04 (16:43:37)	-		✔ 正常		V	<u>.</u>
MaxCompute 太百寸分析(Hologres)		标 准模式 开发跟生产隔离	2019-09-04 (15:06:49)	-) 更新中	Þ	V	V
	101	标 准模式 开发跟生产隔离	2019-09-04 (15:01:15)	-		✔ 正常		V	ς

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。

创建工作空间						
1 基本配置			2 选择引擎		3 引擎详	情
选择region						
华东1(杭州)	华东2(上海)	华北2(北京)	华南1 (深圳)	西南1(成都)	
中国(香港)	新加坡		澳大利亚 (悉尼)	马来西亚(吉隆坡)	印度尼西亚(雅加达)	
日本(东京)	印度(注	孟买)	德国 (法兰克福)	英国 (伦敦)	美国(硅谷)	
美国 (弗吉尼亚)	阿联酋	(迪拜)				
基本信息						
* 工作空间名称		需要字母开头	, 只能包含字母下划线和数字	2		
显示名		如果不埴,默;	人为工作空间名称			
* 模式		简单模式(单现	⊼境)		\sim	
描述						
古 加 : 元 平						
回奴设直 * 能下裁Select结里 Ø		Ŧ				
20 I 200 circology 🖉						
下一步	取	消				

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。

创建工作空间
选择DataWorks服务
数据集成、数据开发、运维中心、数据质量 您可以进行数据同步集成、工作流编排、周期任务调度和运维、对产出数据质量进行检查等。
选择计算引擎服务
✓ MaxCompute ● 按量付费 ○ 包年包月 ● 开发者版 开通后,您可在DataWorks里进行MaxCompute SQL、MaxCompute MR任务的开发。 充值 续费 升级 降配
 ☑。实时计算 ○ 共享模式 ○ 独享模式 开通后,您可在DataWorks里面进行流式计算任务开发。
■ E-MapReduce 开通后,您可以在DataWorks中使用E-MapReduce进行大数据处理任务的开发。
○ 交互式分析 开通后,您可以在DataWorks里使用Holostudio进行交互式分析(Interactive Analytics)的表管理、外部表管理、SQL任务的开发。
选择机器学习服务
✓
下一步 上一步 取消
~

■ 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。

创建	工作空间					
\bigcirc	基本配置		✓ 选择	引擎 ————————————————————————————————————	3	引擎详情
~	MaxCompute					
	* 实例名称					
	开发环境			生产环境		
	MaxCompute项目名称 🛛			MaxCompute项目名称 🛿		
	MaxCompute访问身份	个人账号		MaxCompute访问身份 🥑	工作空间所有者	\sim
				* Quota组切换	按量付费默认资源组	\sim
~	PAI					
	使用GPU 🛛					
Û	健工作空间 上一步					

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



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。

⑤ C₀数据集成	-	~							ನಿ 📕	
= → 任务列表	数据源类型:	全部	> 数据源名称:				C周新	多库多表搬迁	批量新增数据源	新增数据源
- 高线同步任务		数据源名称	数据源类型	链接信息	数据源描述	创建时间	连通状态	连通时间		操作
		odos first	ODPS	Endpoint :	 connection from o	2019/08/09				
办 数据源				项目名称:	574	22:07:58				
🛃 批量上云										

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。

新增MySQL数据源		×
* 数据源类型:	阿里云数据库 (RDS) ~	
* 数据源名称:	workshop	
数据源描述:	窃电用户数据源	
* 适用环境:	✔ 开发 生产	
地区:	华东 1-杭州 ~	
* RDS实例ID :	rm-bp1z74n31h36kqv7x	?
* RDS实例主帐号ID :	1486821399873474	?
* 数据库名 :	shanyun	
* 用户名:	xc_super1	
* 密码 :		
测试连通性:	测试连通性	
0	需要先添加白名单才能连接成功, <mark>点我查看如何添加白名单</mark> 确保数据库可以被网络访问	
	上一步	完成

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。

~ 数据开发			
Sq ODPS SQL			
Sc ODPS Script			
🗂 SQL组件节点	新建节点		×
Sp ODPS Spark			
Py PyODPS	节点类型:	虚拟节点	
Ⅵ 虚拟节点	节点名称:	start	
Mr ODPS MR	月标文件来·	业务流程/窗由用户识别	
Sh Shell			
P→ AnalyticDB for PostgreSQL			
AnalyticDB for MySQL			提交取消
	4 5		
	A N		
◇ 节点组	C		
◇ 数据集成			
Di 数据同步			
◇ 数据开发	新建节石		×
			~
C cou 细件若古	节点类型:	数据同步	
	节点名称:	电量下降趋势数据同步	
	目标文件夹:	业务流程/窃电用户识别	
M ODPS MR			現消

	A N		
~ 节点组	C		
~ 数据集成			
DI 数据同步			
> 数据开发	新建节点		×
Sq ODPS SQL			
Sc ODPS Script	节点类型:	数据同步	
分 SQL组件节点	节点名称:	窃电标志数据同步	
SP ODPS Spark	目标文件夹:	业务流程/窃电用户识别	
「Y PyODPS			
M DDPS MR			提交取消

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。

f 💿 🔍 🖈	
* 节点组 C	
~ 数据集成	
回 数据同步	
◇ 数据开发	Vi start 🥥
Sq ODPS SQL	
Sc ODPS Script	
♪ SQL组件节点	
Sp ODPS Spark	
Py PyODPS	□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
☑ 虚拟节点	

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。

X 调度配置							调度
cron表达式:	00 01 00 * * ?						配置
依赖上—周期:							版本
调度依赖 🕐 🗉							
自动解析:	● 是 ● 否 🕅	f输入输出					
依赖的上游节点:	请输入父节点输出名称	或输出表名	✓ + 使用工作	空间根节点			
父节点输出名称	父节点输出表名	节点名	父节点ID	责任人	来源	操作	
_root		_root			手动添加		
本节点的输出:	请输入节点输出名称						
输出名称	输出表名	下游节点名称	下游节点ID	责任人	来源	操作	

3. 配置完成后,单击左上角的 🛄 进行保存。

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。

DDL模式 从开发环境加载	我 提交到开发环境 从生产环境加载	提交到生产环境
	表名 trend_data	
中文名	电量下降趋势表	
一级主题	请选择 >	二級主题 清选择 ダン 新建主题 C
描述		
物理模型设计		
分区类型	● 分区表 ○ 非分区表	生命周期
层级	请选择	物理分类 请选择 新建层级 С

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。

01 选择数据源	数据来源		
	在这里配置数据的来源端和写力	入端 ; 可	以是默认的
* 数据源	MySQL 🛛 workshop	~	?
*表	[°] trending [°] ×	~	
	添加對	数据源 +	
数据过滤	请参考相应SQL语法填写where过滤语句(不要填 where关键字)。该过滤语句通常用作增量同步	写	?
切分键	根据配置的字段进行数据分片,实现并发读取		?
	数据预览		

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。

	数据去向			收起
也可以是您创建的自	有数据源查看支持的数据来》	原类型		
* 数据源	odps 🗸	odps_first	~	0
*表	trend_data		~	
			一键生成目标表	
* 分区信息	dt = \${bizdate}	0		
清理规则	写入前清理已有数据 (Insert	t Overwrite)	~	
空字符串作为null	● 是 💿 否			

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。

源失表字段 美型 グ 目标表字段 美型 同名映射 uid INT ● uid BICINT 可行映射 indicator INT ● 0 trend BICINT 取消映射	02 字段映射		源头表			目标表		收起
uid INT ● ● III indicator INT ● ● uid BIGINT 取消映射		源头表字段	类型	Ø		目标表字段	类型	同名映射 同行映射 取消映射
indicator INT O trend BIGINT 自动排版		uid	INT	•		🕨 uid	BIGINT	
		indicator	INT	•		trend	BIGINT	自动排版
添加—行 +		添加—行 +						

e) 配置通道控制。

03 通道控制	
您可以配置作业的传输速率和错误纪录数来控制整个数据同步过程:数据同步文档	
* 任务期望最大并发数 2 ⑦	
*同步速率 💿 不限流 🔘 限流	
错误记录数超过 脏数据条数范围,默认允许脏数据 条,任务自动结束 🧿	
任务资源组 默认资源组 🗸	

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

F • • 4	
◇ 节点组 С	
~ 数据集成	L.R
回 数据同步	
◇ 数据开发	Vī start 🥥
Sq ODPS SQL	
Sc ODPS Script	
♪ SQL组件节点	
Sp ODPS Spark	$\downarrow \qquad \downarrow \qquad \downarrow$
Py PyODPS	DI 电量下降趋势数据同… 🧇 DI 窃电标志数据同步 🤡 DI 指标数据同步

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。

提交			×
请选择节点		节点名称	
		电量下降趋势数据同步	
		窃电标志数据同步	
		指标数据同步	
备注			
	✔ 忽略轴	认输出不一致的告 警	
			提交取消

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。

	6 🖸		C 🗱		
	odps so	ql			
	******	*****	*****	****	******
	author:				
	create	time:2019-0	8-09 21:36:	14	
	******	*****	*****	*****	******
	SELECT CO	ount(*) from	trend data	where dt=2	0190808;
	SELECT co	ount(*) from	indicators	data where	ds=20190808;
	SELECT C	ount(*) from	steal flag	_ data where	ds=20190808;
			_ 0.	-	
16		<u> </u>			
	结果[1] x	: 结果[2]	结	里[3]	
	2424[1] X				
	A				
1	_c0	~			
2	293				



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。

DDL模式 从开发环境加载	え 提交到开发环境 が	人生产环境加载 提交到生产现	積	
	表名 trend_data			
基本属性				
中文名	电量下降趋势表			
一级主题	请选择 >	二级主题	请选择	新建主题
描述				
物理模型设计				
分区类型	● 分区表 ● 非分区表	生命周期		
层级	请选择 ~	物理分类	请选择	新建层级 C
表类型	● 内部表 ○ 外部表			

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。



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。



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。

提交				×
请选择节点		节点名称		
		数据清洗		
		数据汇聚		
备注	数据加工			
	✔ 忽略辅	ì入输出不一致的告警		
				交取消

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。





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。



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。

🜀 🖪 任务发布	100		~						<i>€</i> Da	taStudio 🕜 這	維中心 🔍	timetaleni :
												0
6月 创建发布包	创建发布	包										☆ 待发布列表
日二 发布包列表	解决方案:	请选择		业务流程:	窃电用户识别 🗸 🗸	提交人:	全部		节点ID: 请输入节	点ID		
	节点类型:	请选择		变更类型:	清选择 🗸 🗸	提交时间小	▶于等于: үүүү-№	IM-DD		家		
			名称		提交人	节点类型	变更类型	节点状态	提交时间	开发环境测试	操作	
		700002621611	数据汇聚			ODPS SQL			2019-08-13 10:38:24	未測试	查看 发布 漆	
		700002621610	数据清洗			ODPS SQL	-		2019-08-13 10:38:21	未測试	查看 发布 滚	
		700002621602	指标数据同	步	-	数据同步			2019-08-13 10:05:53	未测试	查看 发布 汤	
		700002621601	窃电标志器	文据同步		数据同步	100		2019-08-13 10:05:49	未測试	查看 发布 汤	
		700002621600	电量下降键	当势数据同步		数据同步			2019-08-13 10:05:46	未测试	查看 发布 涩	
	添加到待	持发布 打开待发	布发布	选中项					< <u>+</u> -	-页 👖 下—]	页 > 每页 10	显示:

3. 进入右上角的待发布列表,单击全部打包发布。

待发布 5 项	全部打包发布]	×
待发布		操作	
ID: 700002621611 提交人 : o2 节点状态: 检查通过	名称: 数据汇聚 □ 节点类型: ODPS SQL 变更类型:	查看	移除
ID: 700002621610 提交人 : o2 节点状态: 检查通过	名称: 数据清洗 节点类型: ODPS SQL 变更类型:	查看	移除
ID: 700002621602 提交人 : o2 节点状态: 检查通过	名称: 指标数据同步 节点类型: 数据同步 变更类型:	查看	移除
ID: 700002621601 提交人 : o2 节点状态: 检查通过	名称: 窃电标志数据同步 节点类型:数据同步 变更类型:	查看	移除
ID: 700002621600 提交人 : o2 节点状态: 检查通过	名称: 电量下降趋势数据 同步 □ 节点类型: 数据同步 变更类型:	查看	移除

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。

6	🤔 运维中心		•					& DataStudio 🔍 📕	-
e	运维大屏	搜索:	节点名称/节点ID Q	解决方案: 请选择解决7	5案 🗸 业务流程: 业务流程	▼ 节点类型: 请选择节点	类型 → 责任人:	✓ 基线: 请选择基线	~
a	周期任务运维 、	•	我的节点 🗌 今日惨改的节点	暂停(冻结)节点	重置清空				
	周期实例	-	名称	节点ID		生立	环语 法诺情竭作	C	刷新 收起搜索 C
	补数据实例 测试实例		数据汇聚	tension 1		±r	「小現,用哇俱採TF		
ର	手动任务运维		数据清洗	Rest of the local division of the local divi	start				
Ŵ	智能监控 🗸 🗸		指标数据同步	Total Second					
			电量下降趋势数据同步	-	● 电量下降趋势…	● 窃电标志数据同步 **#■#	↓ 指标数据同步		
			start		XVB2004	***Bally	and and a		

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。

方案 🗸 业务流程:	业务流程	✓ 节点类型:	请选择节点类型	✔ 责任人:
重置 清空				
			生产环境	竟,请谨慎操作
start 应节点 电量下降趋势 数据集成	展开父节点 > 展开子节点 > 节点详情 查看代码 编辑节点 查看血缘 测试 补数据 > 暂停(冻结) 恢复(解冻)	一部では のででので、 のでで、 のでで、 ので、 ので、 ので、 ので、	步	指标数据同步 数据集成

4. 勾选需要补数据的任务, 输入业务日期, 单击确定。

补数据				×
* 补数据名称:	P			
* 选择业务日期:	2019-08-12	- 2019-08-12	Ē	
* 是否并行:	不并行	~		
* 选择需要补数据的节点:				
✔ 任务名称	按名称进行搜索	Q		任务类型 🍸
start				虚节点
✓ 电量下降趋势	数据同步			数据集成
✓ 窃电标志数据	同步			数据集成
✓ 指标数据同步				数据集成
✓ 数据清洗				ODPS_SQL
✓ 数据汇聚				ODPS_SQL
				确定 取消

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。

■ (-)阿里云	华东2(上海) ▼		Q 搜索			费用 工单	备案 企业 支持与服务	e ý. A	⑦	简体中文
机器学习PAI	机器学习PAI / Studio-可	视化建模								帮助
概览	PAI 可视化	建模								
Studio-可视化建模	封装常用机器学习算法	级丰富的可视化组件,用户表	无需代码基础,通过拖挂	如規即可训练模型。						
算法发布	创建项目 项目名称	• • 请输入	Q							(
DSW-Notebook建模 EAS-模型在线服务 NKW	项目名称	显示名	付携模式	所屬区域	项目管理员	MaxCompute资源	创建时间	开启GPU 🚺	操作	
AI市场	1000	100	I/O后付费	华东 2 (上海)	datawor*******	100.000	2019-08-09 22:07:56		进入机器	学习
	10,000	4000,000	CU预付费	华东 2 (上海)	datawor*******	10.000	2019-08-07 10:18:24		进入机器	学习
	and the	4000	I/O后付费	华东 2 (上海)	datawor*******	10,000	2019-08-07 10:17:22		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	机器学习F	PAI 华东2(上海)		算法平台	前往运维
ŝ	搜索		Q		
百页	✓ / / 我的实	验			
る 実验		④ 从模板新建实验			
		♪ 导入实验			
Notebook:		① 新建空白实验			
8		+ 新建文件夹			
数据源		■ 重命名			
ற் 		直 删除			
₩		评估对比			
模型					
(0) # 1					

4. 填写新建实验对话框中的名称和描述。

新建实验		×
名称	窃电用户识别	
	必填, 且长度小于32	
项目		_
描述	窃电用户识别	
位置	✓ 「 我的实验	
	创建取消	É

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。

🖯 data4ml		
- 0-	□ 重命名	
	卣 删除	Delete
	8 复制	Ctrl + C
	▶ 从此处开始执行	Ŧ
	○ 刷新字段	
	 • • ē	

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

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。

2	机器学习PAI 华东2(上海)	100 ,00	-	算法平台	前往运维
分前	搜索	Q	ā i	1电用户识别 ×	
A	> 数据预处理		٤ð	云行 ⑦ 部署	🗞 Auto ML
安全	> 📄 特征工程				
Notebook	✓ / 统计分析				
● 数据源	➡ 数据视图				
<u> </u>	EQI加力差				
ピッ 组件	EQ 经验概率密度图				data4mi
₩	🔂 全表统计				
模型	🖸 卡方拟合性检验			(
¢	🖸 卡方独立性检验		_	¥	
ig 🖬	▶ 箱线图			ð 相关系数矩阵	≨-1 ①
	10 散点图				
	10 相关系数矩阵				
	🖸 双样本T检验				
	议 单样本T检验				

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
 - 机器学习PAI 华东2(上海) æ 算法平台 前往运维 Q ☆ 颜 搜索 🛆 窃电用户识别 🗙 1 源/目标 ●运行 ④部署 @ Auto ML 込 実验 📄 数据预处理 特征工程 <u>6</u> 统计分析 🖸 数据视图 日 数据源 🖯 data4ml \odot 🖸 协方差 ற் ஆடி 🖸 经验概率密度图 🖸 全表统计 ❤ 概型 🖸 卡方拟合性检验 © F 🖸 卡方独立性检验 🖸 数据视图-1 🖸 相关系数矩阵-1 \odot 🖸 箱线图 10 散点图 🖸 相关系数矩阵 🔁 双样本T检验 🖸 单样本T检验 🖸 正态检验
 - 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。

	⊕ ⊖ ⊕	 国 23	字段设置	参数设置	执行调优
			选择特征列 必道	选择字段	
			选择目标列 可读	た使可视化更加非	
			tx≄ntni ⊓uz	选择字段	NK HATE ()
data4ml-1			kv,kv稀疏器	饮据格式	

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

选择字段				кл КИ	ĸ
输入关键字搜索列,包含关键字即可				Q	
■ 全选		已选		列表编辑	
BIGINT	~	Î	字段	类型	
uid		ů	trend	BIGINT	
✓ trend		Û	xiansun	BIGINT	
✓ xiansun		Û	warnindicator	BIGINT	
✓ warnindicator					
flag					
				确定取消	

- 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。

4	机器学习PAI 华东2(上海)	▼ 算法平行	前往运维			幕
<u>ନ</u>	搜索	Q 🖉 窃电用户识别	×			
em A	查看组件说明文档	 逐运行 ④ 	邮晷 🔍 Auto ML			€ € ⊕ ⊹ ⊡ 🛙 🖸
樂脸	 保存的分组 					
Notebook	> 📄 源/目标					
日 数据源	∨ 🧧 数据预处理					
A.	> 📄 采样与过滤					
组件	> 📄 数据合并					
\\$	🔆 美型转换					
1222	🔆 増加序号列					
(2) 101	柒 拆分				🖯 data4ml-1 🛛 📀	
	🔆 缺失值填充 new				\wedge	
	¾ 归─化					
	⅔ 标准化					
	🔆 KV to Table				+	
	犠 Table to KV			🗟 相关系数矩阵-1 🛛 📀	🖸 数据视图-1 🛛 🕑	柒拆分-1 🕕
	> 📄 特征工程					
	> 📄 统计分析					
	> 📄 机器学习					

- 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。

RB Ud Tend Xansul wanindicator flag 1 2 4 0 4 1 2 5 3 1 0 0 3 5 0 1 0 0 4 8 3 1 0 0 5 9 3 1 3 1 6 10 4 0 0 0 7 14 4 0 2 0 8 0 0 0 0 7 14 4 0 2 0 8 16 0 0 0 0 10 2 7 0 0 0 11 23 6 0 0 0 12 24 1 0 0 0 14 26 1 0 0 0 16	数据探查	- pai_te	mp_16758	85_1706043	3_1 - (仅显示前一	百条)	ку X
Pic uida tenda kansura wamindicatora flaga 1 2 4 0 4 1 2 5 3 1 0 0 3 7 5 3 1 0 0 4 6 3 1 3 1 1 5 9 3 0 0 1 1 6 9 3 0 0 0 1 6 10 4 1 0 0 1 7 14 4 0 0 0 1 10 1 1 0 0 1 1 11 2 7 0 0 0 1 11 2 1 0 0 1 1 12 4 1 0 0 1 1 13 5 1 0							
1 2 4 0 4 1 2 5 3 1 0 0 3 7 5 0 2 1 4 8 3 1 3 1 4 8 3 1 3 1 5 9 3 1 3 1 6 0 3 1 3 1 6 1 3 1 3 1 7 4 3 1 3 1 7 14 4 1 3 1 7 14 3 1 3 1 8 1 3 3 1 3 10 2 7 0 0 1 11 3 6 1 1 1 14 2 1 0 0 1 15 7 5 1 1 1	序号▲	uid 🔺	trend 🔺	xiansun 🔺	warnindicator 🔺	flag 🔺	
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 3 0 0 9 9 8 1 4 1 10 22 7 0 0 0 11 23 6 0 0 0 12 24 4 12 1 1 13 25 7 0 0 0 14 28 1 0 0 0 15 27 5 1 0 0 16 28 1 1 1 1	1	2	4	0	4	1	
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 3 0 0 9 19 8 1 4 1 10 22 7 0 0 0 11 23 6 0 0 0 12 24 4 1 0 0 13 25 7 0 0 0 14 20 1 0 0 0 14 23 1 0 0 0 15 7 5 1 1 1 16 28 1 1 1 1	2	5	3	1	0	0	
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 0 0 9 19 8 1 4 1 10 22 7 0 0 0 11 23 6 0 0 0 12 24 4 1 0 0 13 25 7 0 0 0 14 26 2 1 0 0 15 27 5 1 0 0 16 28 1 4 1 1 17 29 5 1 1 1	3	7	5	0	2	1	
§ 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 0 0 13 25 7 0 0 0 14 26 2 1 0 0 15 7 5 1 0 0 16 23 1 1 1 1 17 29 5 1 1 1	4	8	3	1	3	1	
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 4 1 1 17 29 5 1 1 1	5	9	3	0	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 7 5 1 0 0 16 23 1 1 1 1 17 29 5 1 1 1	6	10	4	1	0	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 16 28 1 0 0 0 16 29 5 1 0 0 17 29 5 1 1 1	7	14	4	0	2	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 4 1 1 17 29 5 1 1 1	8	16	0	0	3	0	
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 4 1 1 17 29 5 1 1 1	9	19	8	1	4	1	
11 23 6 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 4 1 1 17 29 5 1 1 1	10	22	7	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 4 1 17 29 5 1 1 1	11	23	6	0	0	0	
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	12	24	4	1	2	1	
14 26 2 1 0 0 15 27 5 1 0 0 16 28 1 1 4 1 17 29 5 1 1 1	13	25	7	0	0	0	
15 27 5 1 0 0 16 28 1 1 4 1 17 29 5 1 1 1	14	26	2	1	0	0	
16 28 1 1 4 1 17 29 5 1 1 1	15	27	5	1	0	0	
17 29 5 1 1 17 29 5 1 1	16	28	1	1	4	1	
复刻 关闭	17	29	5	1	1	1	
起刻 关闭							
						复刻	关闭

- 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。

选择字段				кл X
輸入关键字搜索列,包含关键字即可全选		已选		Q 列表 编辑
BIGINT	~	Û	字段	类型
uid		ů	trend	BIGINT
✓ trend		Ē	xiansun	BIGINT
✓ xiansun		Ē	warnindicator	BIGINT
varnindicator				
flag				
				确定 取消

- 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。

逻辑回归二分类			К Л К Л	×							
在輸入数据为稀疏的时候,不显	在输入数据为稀疏的时候,不显示 weight 全是 0 的特征										
今纪夕 .		权重									
	1.	0 🔺									
trend	1.218831335893094	-									
xiansun	2.286408879362565	-									
warnindicator	1.316373444464322	-									
常量	-9.825371109484525	-9.825371109484525 0									
保存到MaxCompute: p	ai_lr	保存	关闭								

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。

Ø	ł	几器学习PAI 华东2(上海)	***.*	•	算法平台	前往运维	
୍ଥିତ	搜	**************************************	Q	Å Ø	电用户识别 ×		
首页	>	📄 源 / 目标		● iz	行 ① 部署	Q Auto ML	€ ⊖ ⊕ ⊹ □ ፬ 2
込 実验	>	📄 数据预处理					
60	>	📄 特征工程					
Notebook	>	📄 统计分析					
 ₩/探/新	\sim	📶 机器学习					
A.	>	📄 二分类					
ep 组件	>	▶ 多分类					
₩	>	📄 聚类					Udata4111-1
機理	>	二 回归					
(Q) (Q)	>	▶ 关联推荐					
	>	🖿 评估					
	>	▶ 推荐算法					
		🖸 预测					
	>	▶ 深度学习	点击 <mark>这里</mark> 开				
	>	📄 强化学习					
	>	📄 时间序列					□ 逻辑回兴-1 / ⊘
	>	📄 文本分析					
	>	📄 网络分析					
	>	I I					17 预测-1
	>	📄 金融板块					

- 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。

选择字段			K2 K3
输入关键字搜索列,包含关键字即可			Q
✔ 全选	已选		列表 编辑
V BIGINT V	Ē	字段	类型
∨ uid	Ô	uid	BIGINT
✓ trend	Ô	trend	BIGINT
✓ xiansun	Ô	xiansun	BIGINT
varnindicator	Ô	warnindicator	BIGINT
✓ flag	Ô	flag	BIGINT
			确定取消

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。

			-	-				
茅号▲	uid 🔺	trend 🔺	xiansun 🔺	warnindicator 🔺	flag 🔺	prediction_result	prediction_score -	prediction_detail
	1	4	1	1	1	0	0.7936818743516113	{ "0": 0.79368187
2	3	2	1	1	1	0	0.9777937746755442	{ "0": 0.97779377
ļ	4	9	0	0	0	1	0.758433607940023	{ "0": 0.24156639
Ļ	6	2	0	0	0	0	0.9993815703148501	{ "0": 0.99938157
;	11	10	1	2	1	1	0.9993127315069457	{ "0": 0.00068726
i	12	10	1	3	1	1	0.9998156465709368	{ "0": 0.00018435
	13	2	0	3	0	0	0.9688889850628143	{ "0": 0.96888898
	15	3	0	0	0	0	0.9979107883787329	{ "0": 0.99791078
)	17	9	0	3	1	1	0.9938992931601628	{ "0": 0.00610070
0	18	0	0	2	0	0	0.9992484524299784	{ "0": 0.99924845
1	20	2	0	4	0	0	0.8930436506958285	{ "0": 0.89304365
2	21	3	0	1	0	0	0.9922517024361496	{ "0": 0.99225170
3	35	2	1	4	1	1	0.5409565812943579	{ "0": 0.45904341
4	38	6	0	1	0	0	0.7678141618637798	{ "0": 0.76781416
5	39	1	0	3	0	0	0.9905983080109843	{ "0": 0.99059830
6	45	4	1	0	0	0	0.9348465344845021	{ "0": 0.93484653
7	51	1	1	3	0	0	0.9145898338779723	{ "0": 0.91458983

- 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。

2	;	机器学习PAI 华东2(上海	9)	•	算法平台	前往运维						帮
ଜ	19	蒙	Q	🛆 窃电	用户识别 ×							
首页	>	📄 源 / 目标		 运行 	ī ① 部署	🔍 Auto ML			Ð	∈ ⊕	: :	E 23
会 実验	>	📄 数据预处理										
6	>	▶ 特征工程										
Notebook 設据源	>	统计分析						data4ml-1	\odot			
	×.	┌── 机器学习										
	,											
组件	,	● 珍元突 ■ 認光										
₩22	,							4		Ì		
¢	>	 					🗟 相关系数矩阵-1 📀	🖻 数据视图-1	⊘ % 拆	ற் -1	_	\odot
182	`	🗸 🔽 评估								/		
		🌐 二分类评估										
		① 回归模型评估							🛛 🖸 逻辑回	类-1 /	\odot	
		聚类模型评估)			
		⑦ 混淆矩阵								Ļ		
		🌐 多分类评估							🛛 🔂 预测-1		\odot	
	>	■ 推荐算法)		
		い							ý l		_	
	>	■ 深度字対	点击 这里 升						① 二分类评估·	1		
	>	■ 强化学习										

- 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。