Alibaba Cloud

DataWorks 教程

文档版本: 20220104



法律声明

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

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

通用约定

格式	说明	样例
⚠ 危险	该类警示信息将导致系统重大变更甚至故 障,或者导致人身伤害等结果。	
▲ 警告	该类警示信息可能会导致系统重大变更甚 至故障,或者导致人身伤害等结果。	警告 重启操作将导致业务中断,恢复业务 时间约十分钟。
〔) 注意	用于警示信息、补充说明等,是用户必须 了解的内容。	大意 权重设置为0,该服务器不会再接受新 请求。
? 说明	用于补充说明、最佳实践、窍门等 <i>,</i> 不是 用户必须了解的内容。	⑦ 说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击设置> 网络> 设置网络类型。
粗体	表示按键、菜单、页面名称等UI元素。	在 结果确认 页面,单击 确定 。
Courier字体	命令或代码。	执行 cd /d C:/window 命令,进入 Windows系统文件夹。
斜体	表示参数、变量。	bae log listinstanceid
[] 或者 [alb]	表示可选项,至多选择一个。	ipconfig [-all -t]
{} 或者 {a b}	表示必选项,至多选择一个。	switch {act ive st and}

目录

1.简单用户画像分析(MaxCompute版)	06
1.1. Workshop介绍	06
1.2. 准备环境	06
1.3. 采集数据	08
1.4. 加工数据	23
1.5. 配置数据质量监控	34
1.6. 数据可视化展现	40
1.7. 通过Function Studio开发UDF	48
2.简单用户画像分析(EMR版)	53
2.1. 准备环境	53
2.2. 采集数据	57
2.3. 加工数据	68
2.4. 收集和查看元数据	75
2.5. 配置数据质量监控	75
3.搭建互联网在线运营分析平台	79
3.1. 业务场景与开发流程	79
3.2. 环境准备	80
3.3. 数据准备	86
3.4. 数据建模与开发	90
3.4.1. 新建数据表	90
3.4.2. 设计工作流	95
3.4.3. 节点配置	97
3.4.4. 任务提交与测试	04
3.5. 数据可视化展现	08
4.实现窃电用户自动识别教程 1	20
4.1. 窃电用户自动识别概述	20

	4.2.	准备环境	 120
	4.3.	准备数据	 122
	4.4.	加工数据	 130
	4.5.	数据建模	 139
5	.对接	使用CDH	 152

1.简单用户画像分析(MaxCompute 版)

1.1. Workshop介绍

本模块为您介绍DataWorks的设计思路和核心功能,帮助您深入了解阿里云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。只需要熟悉标准SQL,无需对数据仓 库和MaxCompute的原理过多了解,即可快速掌握DataWorks的基本技能。建议您进一步学习DataWorks教程,深入了解DataWorks的基本概念及功能,详情请参见什么是DataWorks。

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。按照教程演示内容,独立通过MaxCompute计算引擎完成数据采集、数据开发和任务运维等 数据岗位常见的任务。

DataWorks简介

DataWorks是一站式大数据研发平台,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,请申请加入钉钉群进行咨询。

1.2. 准备环境

为保证您可以顺利完成本次实验,请您首先确保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。
- 实名认证。

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

⑦ 说明 如果您已经开通MaxCompute,请跳过该步骤,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,输入您的阿里云账号和密码。

2. 鼠标悬停至顶部菜单栏中的产品,单击大数据 > 大数据计算与分析 > MaxCompute,进入

MaxCompute产品详情页。

- 3. 单击立即开通。
- 4. 在购买页面,选择地域,并选中服务协议,单击确认订单并支付。

? 说明

- 。购买页面默认提供的规格类型为MaxCompute按量计费标准版和DataWorks基础版。
- MaxCompute的项目管理和查询编辑集成DataWorks的功能,因此需要同时开通DataWorks 服务。DataWorks基础版为0元开通,如果您不使用数据集成、不执行调度任务,则不会产 生费用。
- 选择地域时,您需要考虑的最主要因素是MaxCompute与其它阿里云产品之间的关系。例如,ECS所在地域、数据所在地域等。

创建工作空间

⑦ 说明 因本实验提供的数据资源都在华东2(上海),建议您将工作空间创建在华东2(上海),以 避免工作空间创建在其它区域,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. 在概览页面,单击右侧的创建工作空间。

您也可以单击左侧导航栏中的工作空间列表,切换至相应的区域后,单击创建工作空间。

3. 配置创建工作空间对话框中的基本配置,单击下一步。

⑦ 说明 本教程以标准模式的工作空间为例进行操作。

4. 进入选择引擎界面,勾选MaxCompute引擎后,单击下一步。

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

5. 进入引擎详情页面, 配置选购引擎的参数。

分类	参数	描述	
	实例显示名称	实例显示名称不能超过27个字符,仅支持字母开 头,仅包含字母、数字和下划线(_)。	
	Quota组切换	Quota用于实现计算资源和磁盘配额。	
MaxCompute	MaxCompute数据类型	该选项设置后,将在5分钟内生效。详情请参见 <mark>数</mark> 据类型版本说明	
	MaxCompute项目名称	默认与DataWorks工作空间的名称一致。	
	MaxCompute访问身份	包括阿里云主账号和任务负责人。	

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. 采集数据

本文为您介绍如何通过DataWorks采集日志数据至MaxCompute。

背景信息

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,并准备好相应的数据表。

? 说明

- 您可以直接使用本实验提供的数据源,也可以使用自己的数据源。
- 因本实验提供的数据资源在华东2(上海),建议您使用华东2(上海)的工作空间。以避免工作空间创建在其它区域,添加数据源时出现网络不可达的情况。

新建OSS数据源

- 1. 进入数据源管理页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 单击相应工作空间后的进入数据集成。

如果您已在DataWorks的某个功能模块,请单击左上方的**一**图标,选择**全部产品 > 数据汇聚 > 数据集成**,进入**数据集成**页面。

- iv. 在左侧导航栏, 单击数据源, 进入工作空间管理 > 数据源管理页面。
- 2. 在数据源管理页面,单击右上方的新增数据源。
- 3. 在新增数据源对话框中,选择数据源类型为OSS。
- 4. 在新增OSS数据源对话框, 配置各项参数。

新增OSS数据源					<	
* 数据源名称: oss_workshop.	R : oss_workshop_log				^	
数据源描述:	数据源描述:					
* 适用环境: 🗹 开发 📃 🕚	生产					
* Endpoint : http://oss-cn-si	hanghai-interna	l.aliyuncs.com			?	
* Bucket : new-dataworks	s-workshop				?	
* AccessKey ID : LTAI4FvGT3iU4	4xjKotpUMAjS				?	
* AccessKey Secret :						
资源组连通性:数据集成日	£务调度 ?					
 如果数据同步时使用了此数据源, 决方案。 	那么就需要保讨	E对应的资源组和数据源之间	是可以联通的。请	参考资源组的详细机	既念和网络解	
资源组名称		类型	连通状态 (点击状态查看 详情)	测试时间	操作	
公共	资源组		未测试		测试连通性	•
				_	上一步	
参数	描述					
数据源名称	输入oss_workshop_log。					
数据源描述	对数据源进行简单描述。					
	勾选 开发	o				
适用环境	⑦ 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	输入 http://oss-cn-shanghai-internal.aliyuncs.com 。					
Bucket	输入new-dataworks-workshop。					
AccessKey ID	输入LT Al4FvGT 3iU4xjKotpUMAjS。					
AccessKev Secret	输λ9RSI	JoRmNxpRC9EhC4m9	PiuG7Izy7px。			

5. 在资源组列表,单击相应资源组后的测试连通性。

数据同步时,一个任务只能使用一种资源组。您需要测试每个资源组的连通性,以保证同步任务使用的 数据集成资源组能够与数据源连通,否则将无法正常执行数据同步任务。如果您需要同时测试多种资源 组,请选中相应资源组后,单击**批量测试连通性**。详情请参见选择网络连通方案。

- ? 说明
 - (推荐)资源组列表默认仅显示独享数据集成资源组,为确保数据同步的稳定性和性能要求,推荐使用独享数据集成资源组。
 - 如果您需要测试公共资源组或自定义资源组的连通性,请在资源组列表右下方,单击更多选项,在警告对话框单击确定,资源组列表会显示可供选择的公共资源组和自定义资源组。
- 6. 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AccessKey ID、AccessKey Secret和工作空间所在区域。
 - 如果您无法使用内网Endpoint连接数据源,请改用公网Endpoint。

新建RDS数据源

- 1. 单击当前页面左上角的 国 图标,选择全部产品 > 数据汇聚 > 数据集成。
- 2. 在左侧导航栏, 单击数据源 > 数据源列表, 进入工作空间管理 > 数据源管理页面。
- 3. 在数据源管理页面,单击右上方的新增数据源。
- 4. 在新增数据源对话框中,选择数据源类型为MySQL。
- 5. 在新增MySQL数据源对话框中, 配置各项参数。

新增MySQL数据源		×
* 数据源名称:	rds_workshop_log	*
数据源描述:	rds日志数据同步	- 1
* 适用环境:	✔ 开发 生产	
* 地域:	name (an anterpa) v	
* RDS实例ID:	rm-bp1z69dodhh85z9qa	?
* RDS实例主账号ID:	1156529087455811	?
* 数据库名:	workshop	
* 用户名:	workshop	
* 密码:		
资源组连通性:	数据集成 数据服务 任务调度 ?	
; 如果数据同步时像 决方案。	使用了此数据源,那么就需要保证对应的资源组和数据源之间是可以联通的。请参考资源组的 <mark>详细概念和网</mark>	络解
		•
		步 完成

参数	描述	
数据源类型	选择 阿里云实例模式 。	
数据源名称	输入rds_workshop_log。	
数据源描述	输入RDS日志数据同步。	
适用环境	勾选开发。 ⑦ 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。	
地区	选择RDS实例所在的区域。	
RDS实例ID	输入rm-bp1z69dodhh85z9qa。	
RDS实例主账号ID	输入1156529087455811。	
数据库名	输入workshop。	
用户名	输入workshop。	
密码	输入workshop#2017。	

6. 在资源组列表,单击相应资源组后的测试连通性。

数据同步时,一个任务只能使用一种资源组。您需要测试每个资源组的连通性,以保证同步任务使用的 数据集成资源组能够与数据源连通,否则将无法正常执行数据同步任务。如果您需要同时测试多种资源 组,请选中相应资源组后,单击**批量测试连通性**。详情请参见选择网络连通方案。

- ? 说明
 - (推荐)资源组列表默认仅显示独享数据集成资源组,为确保数据同步的稳定性和性能要求,推荐使用独享数据集成资源组。
 - 如果您需要测试公共资源组或自定义资源组的连通性,请在资源组列表右下方,单击更多选项,在警告对话框单击确定,资源组列表会显示可供选择的公共资源组和自定义资源组。
- 7. 测试连通性通过后,单击完成。

创建业务流程

- 1. 单击当前页面左上方的 图标,选择全部产品 > 数据开发 > DataStudio (数据开发)。
- 2. 在数据开发面板,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, 输入业务名称和描述。

↓ 注意 业务名称不能超过128个字符,且必须是大小写字母、中文、数字、下划线(_)以及小数点(.)。

- 4. 单击**新建**。
- 5. 进入业务流程开发面板, 鼠标单击虚拟节点并拖拽至右侧的编辑页面。
- 6. 在新建节点对话框中,输入节点名称为workshop_start,单击提交。

f • • 1	»	
◇ 节点组		
◇ 数据集成		
回 离线同步		
✓ MaxCompute	新建节点	×
Sq ODPS SQL		
Sp ODPS Spark	节点类型:	虚拟节点
Py PyODPS	节点名称:	
Sc ODPS Script	目标文件夹:	业务流程
Mr ODPS MR	1	
◇ 通用		
Ch oss对象检查		
Sh Shell		
▼ 虚拟节点		
🔊 跨租户节点		

以同样的方式新建两个离线同步节点,节点名称分别为oss_数据同步和rds_数据同步。

7. 通过拖拽连线,将workshop_start节点设置为两个离线同步节点的上游节点。

•	Vi workshop_start	
● Di rds_数据同步	•	Di oss_数据同步

配置workshop_start节点

- 1. 在数据开发页面,双击相应业务流程下的虚拟节点。打开该节点的编辑页面,单击右侧的调度配置。
- 2. 在**调度依赖**区域,单击**使用工作空间根节点**,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称_root。

★ 调度配置							週
定时调度:							E 配 置
具体时间:	00:04						
							版本
cron表达式:	00 04 00 * * ?						
依赖上—周期:							
谷源属性 ⑦ —							
文[[5] 南江 🔍							
调度资源组:	公共调度资源组						
油库休格 ②							
,同反100款 🔮							
自动解析 💿 是	· · · · · · · · · · · · · · · · · · ·						
依赖的上游节点	请输入 ◇ 节点输出名称或输出	表名 🗸					
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	
p_roo_	t -	-	10000	1000003-00003	手动添加	删除	

3. 配置完成后,单击工具栏中的凹图标。

新建表

1. 在数据开发页面打开新建的业务流程,右键单击MaxCompute,选择新建>表。



 在新建表对话框中,输入表名,单击提交。
 此处需要创建两张表(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。

↓ 注意 表名必须以字母开头,不能包含中文或特殊字符,且不能超过64个字符。

- 3. 通过DDL模式新建表。
 - 新建ods_raw_log_d表。

在表的编辑页面单击DDL模式,输入下述建表语句。



```
CREATE TABLE IF NOT EXISTS ods_raw_log_d (
   col STRING
)
PARTITIONED BY (
 dt STRING
);
```

○ 新建ods_user_info_d表。

在表的编辑页面单击DDL模式,输入下述建表语句。

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

4. 单击生成表结构,并确认覆盖当前操作。

5. 返回建表页面,在基本属性中输入表的中文名。

6. 完成设置后,分别单击提交到开发环境和提交到生产环境。

配置离线同步节点

⑦ 说明 标准模式的工作空间下,不建议离线同步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,您可以申请数据权限,以读取写入开发环境中的表数据。

1. 配置oss_数据同步节点。

i. 在数据开发页面,双击oss_数据同步节点,进入节点配置页面。

ii. 选择数据来源。

01 选择数据源		数据	居 来 源		
		在	这里配置数据的来源	端和写入端;	可以是
* 数据源	OSS		oss_workshop_log		?
* Object前缀	user_log.txt				
				添加Object	:
* 文本类型	text				
* 列分隔符	I				
编码格式	UTF-8				
null值	表示null值的字符串				
*压缩格式	None				
* 是否包含表头	No				
		数据	滪览		

参数	描述
数据源	选择OSS > oss_workshop_log数据源。
Object前缀	输入OSS文件夹的路径,请勿填写Bucket的名称。示例为user_log.txt。
文本类型	选择text类型。
列分隔符	输入列分隔符为 。
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。

iii. 选择数据去向。

01 选择数据源	数据来源		数据去向			
• 数据源	OSS v oss_workshop_log v	0	* 数据源	ODPS 🗸	odps_first ~	0
* Object前缀	user_log.txt		生产项目名			
		t.	*表	ods_raw_log_d		
* 文本类型	text ~					
• 列分隔符			• 分区信息	dt = \${bizdate}	0	
编码格式	UTF-8		清理规则	写入前清理已有数据 (Ins	sert Overwrite) 🗸 🗸	
nuli值	表示null值的字符串		空字符串作为null	● 是 🧿 香		
* 压缩格式	None ~					
* 是否包含表头	No ~					
	数据预览					

参数	描述
数据源	选择ODPS > odps_first数据源。
表	选择数据源中的ods_raw_log_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。

? 说明

- odps_first数据源是工作空间绑定MaxCompute实例时,系统自动生成的默认数据源。
- odps_first数据源写入至当前工作空间下的MaxCompute项目中。

iv. 配置字段映射。



v. 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 🛛 🗸 🖉
	*同步速率 💿 不限流 🕜 限流 🕐
	错误记录数超过 註数据条数范围,默认允许脏数据 条,任务自动结束 ⑦
参数	描述
任务期望最大并发数	数据同步任务内,可以从源并行读取或并行写入数据存储端的最大线程数。向导 模式通过界面化配置并发数,指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库,以避免抽取速度过大,给源库造成太大的 压力。同步速率建议限流,结合源库的配置,请合理配置抽取速率。
错误记录数	错误记录数,表示脏数据的最大容忍条数。

vi. 单击页面右侧的**调度配置**, 在**调度依赖 > 本节点的输出**区域, 输入本节点的输出名称为工作空间 名称.ods_raw_log_d, 单击 ➡图标。

↓ 注意 不建议您使用中文作为本节点输出名称,会减少自动推荐功能的准确性。

× 调度配置										调度
										記置
调度依赖 ⑦										版本
依赖的上游节点 请输入父节点输出名称	或輸出表名		+ 使用工作空	间根节点	自动推荐					
父节点输出名称	父节点输出表名	节点名		父节点		责任人		来源	操作	
_root			pot					手动添加		
本节点的输出	raw_log_d	+	3							
輸出名称		输出表名	下游节点名称		下游节点ID	责任人	来源		操作	
.500855460_out		- Ø					系统默认	添加		
.oss_数据同步 化		- @					手动添加			

vii. 确认当前节点的配置无误后,单击工具栏中的图图标。

viii. 关闭当前任务,返回业务流程配置面板。

2. 配置rds_数据同步节点。

i. 在数据开发页面,双击rds_数据同步节点,进入节点配置页面。

ii. 选择数据来源。

			6		
01 选择数据源			数 据来 源		
			在这里配置数	居的来源端和写入端	;可以是默认
* 数	居源 MySQL		✓ rds_worksl	hop_log	0
	*表 [·] ods_us	er_info_d' ×]
				添加数据源	÷
数据	过濾 请参考相 字)。该	应SQL语法填译 过滤语句通常/	where过滤语句(不 用作增量同步	要填写where关键	0
រារ	分键 uid				?
			数据预览]

参数	描述
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。
数据过滤	该数据过滤语句通常用作增量同步,此处可以不填。
切分键	默认为uid。

iii. 选择数据去向。

01 选择数据源	数据来源					数据去向	
* 数据源	MySQL v rds_	workshop_log	0	* 数据源	ODPS	✓ odps_first	0
* _表	'ods_user_info_d' ×			生产项目名			
					ods_user_info_d		
数据过渡	请参考相应SQL语法填写where这 where关键字)。该过诸语句通知	地谚语句 (不要填写 常用作增量同步	0	* 分区信息	dt = \${bizdate}		
+11-2428	nid		A	清理规则	写入前清理已有数据	(Insert Overwrite)	
2017.00	数据预选			空字符串作为null	● 是 🧿 否		
		描述					
		JEL KL					
数据源		选择ODI	PS > odps_first数据源。				
表		选择数据	弱源中的ods_user_info_d表。				
分区信息		默认配置	乱为\${bizdate}。				
清理规则		默认为写	6入前清理已有数据 。				
空字符串作为nul	l	此处勾进	西。				

- iv. 配置字段映射。
- v. 配置通道控制。
- vi. 单击页面右侧的**调度配置**, 在**调度依赖 > 本节点的输出**区域, 输入本节点的输出名称为工作空间 名称.ods_user_info_d, 单击 ■图标。

添加成功后,您可以删除不规范的输出名称。

↓ 注意 不建议您使用中文作为本节点输出名称,会减少自动推荐功能的准确性。

X 测度化器								1 調度配置
调度休赖 ③ 依赖的上脚节点 请输入交节点输出条约或输出条名		+ 使用工作空	间根节点	<u>ķ</u>				版本
父节点输出各称 父节点输出表名 root -	节点名	mot	父节点ID	责任人	_	来源	授作	
本节点的输出 ods_user_info_d 2	[+	3						
輸出名称	输出表名	下游节点名称	下游节点ID) 责任人	来源		操作	
out	- Ø				系统默认清	ābo		
.rds_数据同步 C	- @				手动添加			

vii. 确认当前节点的配置无误后,单击工具栏中的 图标。

viii. 关闭当前任务,返回业务流程配置面板。

提交业务流程

1. 在数据开发页面,双击相应的业务流程打开编辑页面,单击工具栏中的回图标。

厨 ⊙ ● র	
◇ 节点组 С	
◇ 数据集成	
▶ 离线同步	Vi start
✓ MaxCompute	
Sq ODPS SQL	
௺ SQL组件节点	
Sp ODPS Spark	
Py PyODPS	Di oss 数据同步
Sc ODPS Script	
Mr ODPS MR	
◇ 通用	

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

运行业务流程

1. 在数据开发页面,双击相应的业务流程打开编辑页面,单击工具栏中的 <>>> 图图标。



2. 右键单击rds_数据同步节点,选择查看日志。

当日志中出现如下字样,表示同步节点运行成功,并成功同步数据。



3. 右键单击oss_数据同步节点,选择查看日志,确认方法与rds_数据同步节点一致。

确认数据是否成功导入MaxCompute

- 1. 在数据开发页面的左侧导航栏,单击临时查询,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ods_raw_log_d和ods_user_info_d的记录数。

⑦ 说明 SQL语句如下所示,其中分区列需要更新为业务日期。例如,任务运行的日期为 20180717,则业务日期为20180716,即任务运行日期的前一天。

--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;

		sql						

	autho							84 ⁶⁶⁹ 6
	create time:2018-07-18 22:44:38							

	select	count(*) fro	m ods	raw log d	where o	t=2018071	5:	
	select	count(*) from	m ods	usan info	d where	dt-20180	716.	
			003			2010100		
		(4) B(1)		44-99 (p)				
XEA	丁口志	后来[1]	~	结末(4)	×			
	A							
1	_c0	~						
2	570386							

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。详情请参见数据加工。

1.4. 加工数据

本文为您介绍如何通过DataWorks计算和分析已采集的数据。

前提条件

开始本实验前,请首先完成采集数据中的操作。

新建数据表

- 1. 进入数据开发页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 选择工作空间所在地域后,单击相应工作空间后的进入数据开发。
- 2. 在数据开发页面,打开新建的业务流程。右键单击MaxCompute,选择新建>表。
- 3. 在新建表对话框中, 输入表名, 单击提交。

此处需要创建三张表,分别为数据运营层表(ods_log_info_d)、数据仓库层表 (dw_user_info_all_d)和数据产品层表(rpt_user_info_d)。

4. 通过DDL模式新建表。

○ 新建ods_log_info_d表。

双击ods_log_info_d表,在右侧的编辑页面单击DDL模式,输入下述建表语句。

```
--创建数据运营层(ODS)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
);
```

○ 新建dw_user_info_all_d表。

双击dw_user_info_all_d表,在右侧的编辑页面单击DDL模式,输入下述建表语句。

```
--创建数据仓库层(DW)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
);
```

○ 新建rpt_user_info_d表。

双击rpt_user_info_d表,在右侧的编辑页面单击DDL模式,输入下述建表语句。

```
---创建数据产品层 (RPT) 表

CREATE TABLE IF NOT EXISTS rpt_user_info_d (

uid STRING COMMENT '用户ID',

region STRING COMMENT '地域, 根据ip得到',

device STRING COMMENT '终端类型 ',

pv BIGINT COMMENT 'Pv',

gender STRING COMMENT '性別',

age_range STRING COMMENT '生粉段',

zodiac STRING COMMENT '星座'

)

PARTITIONED BY (

dt STRING

);
```

5. 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。

- 6. 返回建表页面后, 在基本属性中输入表的中文名。
- 7. 完成设置后,分别单击提交到开发环境和提交到生产环境。

⑦ 说明 如果您使用的是简单模式的工作空间,页面仅显示提交到生产环境。

设计业务流程

业务流程节点间依赖关系的配置请参见采集数据。

双击新建的业务流程打开编辑页面, 鼠标单击ODPS SQL并拖拽至右侧的编辑页面。在新建节点对话框中, 输入**节点名称**, 单击提交。

此处需要新建三个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. 新建资源。
 - i. 下载ip2region.jar。

ii. 在数据开发页面打开业务流程,右键单击MaxCompute,选择新建 > 资源 > JAR。



iii. 在新建资源对话框中, 输入资源名称, 并选择目标文件夹。

新建资源		×
资源名称:	ip2region.jar	
目标文件夹:	业务流程/workshop/MaxCompute	~
资源类型:	JAR	~
上传文件:	✓ 上传为ODPS资源本次上传,资源会同步上传至ODPS中 点击上传	
		以消

? 说明

- 请选中上传为ODPS资源。
- 资源名称无需与上传的文件名保持一致。
- 资源名称命名规范: 1~128个字符,字母、数字、下划线、小数点,大小写不敏感, JAR资源的后缀为.jar, Python资源的后缀为.py。
- iv. 单击点击上传,选择已经下载至本地的ip2region.jar,单击打开。
- v. 单击确定。
- vi. 单击工具栏中的 图标。
- 2. 注册函数。
 - i. 在数据开发页面打开业务流程,右键单击MaxCompute,选择新建 > 函数。
 - ii. 在新建函数对话框中, 输入函数名称(示例为get region), 单击提交。
 - iii. 在注册函数对话框中, 配置各项参数。

🖱 🗗 🕼 🗗 🗋	
注册函数	
函数类型:	其他函数
MaxCompute引擎实例:	2
函数名:	getregion
责任人:	• •
类名:	org.alidata.odps.udf.lp2Region
资源列表:	ip2region.jar
描述:	IP地址转换地域
命令格式:	getregion("ip')
参数说明 :	P地址

参数	描述
函数类型	选择函数类型。
MaxCompute引擎实例	默认不可以修改。
函数名	新建函数时输入的函数名称。
责任人	选择责任人。
类名	输入 org.alidata.odps.udf.Ip2Region 。
资源列表	输入 ip2region.jar 。
描述	输入IP地址转换地域。
命令格式	输入 getregion('ip') 。
参数说明	输入IP地址。

iv. 分别单击工具栏中的凹和可图标。

配置ODPS SQL节点

- 1. 配置ods_log_info_d节点。
 - i. 双击ods_log_info_d节点,进入节点配置页面。

ii. 在节点编辑页面,编写如下SQL语句。

```
INSERT OVERWRITE TABLE ods log info d PARTITION (dt=${bdp.system.bizdate})
SELECT ip
  , uid
  , time
  , status
  , bytes
  , getregion(ip) AS region --使用自定义UDF通过IP得到地域。
  , regexp substr(request, '(^[^ ]+ )') AS method --通过正则把request差分为3个字段。
  , regexp extract(request, '^[^ ]+ (.*) [^ ]+$') AS url
  , regexp substr(request, '([^ ]+$)') AS protocol
  , regexp_extract(referer, '^[^/]+://([^/]+){1}') AS referer <mark>--通过正则清晰</mark>refer,得
到更精准的URL。
  , CASE
   WHEN TOLOWER (agent) RLIKE 'android' THEN 'android' --通过agent得到终端信息和访问形
式。
   WHEN TOLOWER (agent) RLIKE 'iphone' THEN 'iphone'
   WHEN TOLOWER (agent) RLIKE 'ipad' THEN 'ipad'
   WHEN TOLOWER (agent) RLIKE 'macintosh' THEN 'macintosh'
   WHEN TOLOWER(agent) RLIKE 'windows phone' THEN 'windows phone'
   WHEN TOLOWER (agent) RLIKE 'windows' THEN 'windows pc'
   ELSE 'unknown'
  END AS device
  , CASE
   WHEN TOLOWER(agent) RLIKE '(bot|spider|crawler|slurp)' THEN 'crawler'
   WHEN TOLOWER (agent) RLIKE 'feed'
   OR regexp extract (request, '^[^]+ (.*) [^]+$') 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, '##00')[0] AS ip
   , SPLIT(col, '##00')[1] AS uid
    , SPLIT(col, '##00')[2] AS time
    , SPLIT(col, '##00')[3] AS request
    , SPLIT(col, '##00')[4] AS status
   , SPLIT(col, '##00')[5] AS bytes
    , SPLIT(col, '##00')[6] AS referer
   , SPLIT(col, '##00')[7] AS agent
  FROM ods raw log d
  WHERE dt = ${bdp.system.bizdate}
) a;
```

ⅲ. 单击工具栏中的凹图标。

2. 配置dw_user_info_all_d节点。

i. 双击dw_user_info_all_d节点,进入节点配置页面。

ii. 在节点编辑页面,编写如下SQL语句。

```
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;
```

iii. 单击工具栏中的凹图标。

3. 配置rpt_user_info_d节点。

i. 双击rpt_user_info_d节点,进入节点配置页面。

ii. 在节点编辑页面,编写如下SQL语句。

```
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(gender)
, MAX(age_range)
, MAX(zodiac)
FROM dw_user_info_all_d
WHERE dt = ${bdp.system.bizdate}
GROUP BY uid;
```

ⅲ. 单击工具栏中的Ⅲ图标。

提交业务流程

- 1. 在业务流程的编辑页面,单击回图标,提交业务流程中已配置完成的节点。
- 2. 在提交对话框中,选择需要提交的节点,选中忽略输入输出不一致的告警。
- 3. 单击提交。

运行业务流程

- 1. 在业务流程的编辑页面,单击工具栏中的 图标,验证代码逻辑。
- 2. 待所有任务运行完成显示绿色箭头后, 在左侧导航栏, 单击临时查询。
- 3. 在临时查询面板,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,查询任务运行结果,确认数据产出。

1odps sql 2************************************											
3	author:: create tim *********			42:29 ***********							
6	6 select * from rpt_user info_d where dt=20190630 limit 10;										
7											
			_								
运	行日志	结果[1]	×								
	А		в	С		D	E		F		Н
1	uid 🗸 🗸	region	~	device	✓ pv	~	gender	∨ ag	e_range 🗸 🗸	zodiac 🗸 🗸	dt
2	0016359810821	湖北省		windows_pc	1		女	30	-40岁	巨蟹座	20190630
3	0016359814159	未知		windows_pc	5		女	30	-40岁	巨蜡座	20190630
4	001d9e7863049	浙江省		iphone	21		女	40	-50岁	双鱼座	20190630
5	001d9e7866387	河南省		windows_pc	1		女	40	-50岁	双鱼座	20190630
6	001d9e7869725	未知		windows_pc	1		女	40	-50岁	双鱼座	20190630
7	001dce2983544	湖北省		unknown	2		女	20	-30岁	水瓶座	20190630
8	001dce2986882	广东省		windows_pc	3		女	20	-30岁	水瓶座	20190630
9	0026c84ad1206	台湾省		windows_pc	1		女	20	岁以下	天秤座	20190630
10	0026c84ad4544	福建省		windows_pc	126		女	20	岁以下	天秤座	20190630
	0007 04 17000	to 7th de-			2		4	20		Telet	20100620

查询语句如下所示,通常默认业务日期为运行日期的前一天。

```
---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,您需要发布配置完成的任 务至生产环境。

? 说明

- 发布任务至生产环境前,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
 图标。您在提交任务后,单击
 图标,进入运维
 中心页面。
- 1. 在业务流程的编辑页面,单击工具栏中的图图标,进入发布页面。
- 2. 选择待发布任务,单击添加到待发布。

6	✓ 任务发布		••							& DataStudio	@ 运维中心 🔍 🛑	-
¢‡	三 创建发布包	创建发布包										0 ⑦ 待发布列表
83	发布包列表	解决方案: 节点类型:	请选择 请选择	✓ 业务流程: ✓ 变更类型:	workshop ~ 请选择 ~	提交人: 全部 提交时间小于等	ŧ∓: YYYY-MM-DD	* #	点D: 请输入节点D			
				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	
				rds_数据同步		离线同步	新增		2019-11-11 09:53:46	未測试		
				oss_数据同步		离线同步	新増		2019-11-11 09:53:41	未測试		
				start01		虚拟节点	新増		2019-11-11 09:53:37	未測试	查看 发布 添加到待发布	
				ip2region.jar	insertion in the lateral in	JAR	新増		2019-11-06 13:20:49	未測试	查看 发布 添加到待发布	
		添加到待	援布 打开待发布	发布选中项							1 下-页)	每页显示: 10 ~

- 3. 单击右上角的待发布列表,进入列表后,单击全部打包发布。
- 4. 在确认发布对话框中,单击发布。
- 5. 在左侧导航栏,单击发布包列表,查看发布状态。

在生产环境运行任务

- 任务发布成功后,单击右上角的运维中心。
 您也可以进入业务流程的编辑页面,单击工具栏中的前往运维,进入运维中心页面。
- 2. 在左侧导航栏,单击周期任务运维>周期任务,进入周期任务页面,单击workshop业务流程。
- 3. 双击DAG图中的虚节点展开业务流程,右键单击workshop_start节点,选择**补数据 > 当前节点及下游** 节点。



4. 选中需要补数据的任务,输入业务日期,单击确定,自动跳转至补数据实例页面。

5. 单击刷新,直至SQL任务全部运行成功即可。

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。详情请参见配置数据质量监控。

1.5. 配置数据质量监控

本文为您介绍如何监控数据质量、设置表的质量监控规则和监控提醒等。

前提条件

在进行本实验前,请确保已采集并加工数据。详情请参见采集数据和加工数据。

背景信息

数据质量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。

数据质量开发流程

- 针对已有的表进行监控规则配置,配置完成后进行试跑,验证该规则是否适用。
 您可以根据试跑结果,确认此次任务产出的数据是否符合预期。建议每个表的监控规则配置完成后,都 进行一次试跑操作,以验证表规则的适用性。
- 2. 试跑成功后,将该规则和调度任务进行关联。

在监控规则配置完成且试跑成功的情况下,您需要将表和其产出任务进行关联,以便每次表的产出任务 运行完成后,都会触发数据质量规则的校验,以保证数据的准确性。

3. 关联调度后,每次调度任务代码运行完成,都会触发数据质量的校验规则,以提升任务准确性。

数据质量支持设置规则订阅,您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。如果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。

⑦ 说明

- 每张表在完成规则的配置后,都需要进行试跑、关联调度和规则订阅等操作。
- 数据质量会产生额外的计算费用,更多详情请参见数据质量概述。

配置数据表的监控规则

如果已经完成数据采集和数据加工实验,请确认您已拥有数据

表: ods_raw_log_d、ods_user_info_d、ods_log_info_d、dw_user_info_all_d和rpt_user_info_d。确认 后,进行以下操作:

- 1. 进入表ods_raw_log_d的监控规则页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 选择工作空间所在地域后,单击相应工作空间后的进入数据开发。

 - v. 在左侧导航栏,单击监控规则,从数据源下拉列表中选择MaxCompute。

vi. 在**引擎/数据库实例**下拉框中选择待配置监控规则表所在的引擎实例,在过滤后的表列表中找到待 配置监控规则的表,例如本教程的ods_raw_log_d表。

vii. 单击ods_raw_log_d表后的配置监控规则。

- 2. 配置表ods_raw_log_d的监控规则。
 - i. 在已添加的分区表达式模块,单击+,添加分区表达式。

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。在**添加分区**对话框中,选择 dt=\$[yyyymmdd-1],单击**确认**。分区表达式的详情请参见<mark>配置调度参数</mark>。

⑦ 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。

分区表达式 请输入	+ Q											
		规则名称	规则字段	强/弱	规则模板	模板路径	动态阈值	比较方式	橙色			
								没有数据				
					添加分区				×			
					分区表达式:	dt=\$[yyyymmdd-1]		计算]			
								确认	取消			

- ii. 单击创建规则,默认在模板规则对话框。
- iii. 单击添加监控规则,选择规则模板为表行数,固定值,设置规则的强度为强、比较方式为期望值
 大于0。

表ods_raw_log_d的数据来源于OSS上传的日志文件,作为源头表,您需要尽早判断该表的分区中 是否存在数据。如果该表没有数据,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。

⑦ 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。

配置完成后,单击批量保存。

⑦ 说明 该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。

iv. 单击**试跑**,在**试跑**对话框中,选择**调度时间**,单击**试跑**。

试跑可以立即触发数据质量的校验规则,对配置完成的规则进行校检。试跑完成后,单击**试跑成 功!点击查看试跑结果**,即可跳转至试跑结果页面。

v. 进行关联调度。

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,任务实例运行完成都会触发数据质量 的检查。您可以通过以下两种方式进行表规则和任务的关联调度:

■ 在运维中心页面关联表规则

单击左上方的**≣**图标,选择**全部产品 > 运维中心**。

在左侧导航栏,单击**周期任务运维 > 周期任务**。在DAG图中,右键单击oss_数据同步任务, 选择配置质量监控。

- 名	称	节点ID	生产环境,请谨慎操作
os 🔽	ss_数据同步	700002197850	
rpt	t_user_info_d	700002197984	
dw	w_user_info_all_d	700002197983	Workshopstart Virtual Node
od	ds_log_info_d	700002197982	
rds	ls_数据同步	700002197849	
wc	orkshopstart	700002197848	oss_数据同步 Data Integration ods_log_info_d
更多 ▼	< 1/1 >		

在配置质量监控对话框中,选择表名(ods_raw_log_d)和分区表达式dt=\$[yyyymmdd-1]),单击添加。

■ 在数据质量页面关联表规则

在表的监控规则页面,单击关联调度,配置规则与任务的绑定关系。

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

在**关联调度**对话框中,输入节点ID或节点名称,单击**添加**。添加完成后,即可完成与调度节点任务的绑定。

关联调度		×
将当前分区表达式关联到:		
DTdoc 🗸	任务节点名称	
oss数据同步(7000025443	55) ×	
vi. 订阅任务。

在表的监控规则页面,单击**订阅管理**,设置接收人以及订阅方式。数据质量支持**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。

订阅管理设置完成后,在左侧导航栏,单击我的订阅,查看和修改已订阅的任务。

⑦ 说明 建议订阅全部规则,避免校验结果无法及时通知。

3. 配置ods_user_info_d表规则。

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

- i. 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查看成功的分区配置记录。
- ii. 分区表达式配置完成后,单击创建规则,配置数据质量的校验规则。

分别添加表级规则和列规则:

■ 选择规则字段为表级规则。

创建规则					
模板规则 自定义规则					
添加监控规则	快捷添加				
* 规则名称 :	请输入规则名称		删除		
* 强弱 :	• 强)弱				
* 动态阈值:	○ 是 (● 否				
* 规则来源 :	内置模板 🗸 🗸				
* 规则字段:	表级规则(table)	~			
* 规则模板 :	表行数,固定值	~			
* 比较方式:	大于	~			
* 期望值:	0				
描述:					
WEXTO DO	·····································				
	/F				

选择规则模板为表行数,固定值、强弱为强、比较方式为大于以及期望值为0。

■ 选择规则字段为uid。

添加监控规则 快捷添加 * 规则名称: 请输入规则名称 删除 * 强弱: 9 强 ④ 弱 * * 动态阈值: 0 是 ④ 否 * * 规则来源: 内置模板 ~ * * 规则字段: uid(string) ~ * 规则模板: 重复值个数,固定值 ● * 比较方式: 小于 ~ * 期望值: 1 描述:	模板规则自	定义规则		
 * 规则名称: 请输入规则名称 ● 强 ● 弱 * 动态阈值: ● 是 ● 否 * 规则来源: 内置模板 ▼ * 规则字段: uid(string) ▼ * 规则模板: 重复值个数,固定值 ▼ * 比较方式: 小于 ▼ * 期望值: 1 描述: □ 	添加监控规则	快捷添加		
 * 强弱: ・ 强 ● 弱 * 动态阈值: ● 是 ● 否 * 规则来源: * 内置模板 × * 规则字段: uid(string) × * 规则模板: 重复值个数,固定值 × * 比较方式: 小子 × 期望值: 1 描述: 	* 规则名称 :	请输入规则名称		删除
 * 动态阈值: ○ 是 ● 否 * 规则来源: 内置模板 × * 规则字段: uid(string) × * 规则模板: 重复值个数,固定值 × * 比较方式: 小于 × * 期望值: 1 描述: □ 	* 强弱:	◯ 强 (● 弱		
 * 规则来源: 内置模板 × * 规则字段: uid(string) × * 规则模板: 重复值个数,固定值 × * 比较方式: 小于 × * 期望值: 1 描述: 	★ 动态阈值:	○是 ⑧ 否		
 * 规则字段: uid(string) * 规则模板: 重复值个数,固定值 * 比较方式: 小子 * 期望值: 1 描述: 	* 规则来源:	内置模板 🗸		
* 规则模板: 重复值个数,固定值 ~ * 比较方式: 小于 ~ * 期望值: 1 描述:	* 规则字段:	uid(string)	~	
* 比较方式: 小于 v * 期望值: 1 描述:	* 规则模板 :	重复值个数,固定值	~	
* 期望值: 1 描述:	* 比较方式:	小于	~	
描述:	* 期望值:	1		
	描述:			

添加列级规则,设置主键列(uid)为监控列。选择**模板类型**为重复值个数,固定值、强弱为弱、比较方式为小于以及期望值为1。

iii. 配置完成后,单击**批量保存**。

⑦ 说明 该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。

4. 配置ods_log_info_d表规则。

ods_log_info_d数据主要来源于解析ods_raw_log_d表中的数据。鉴于日志中的数据无法配置过多监控,只需要配置表数据不为空的校验规则即可。

i. 配置表的分区表达式为dt=\$[yyyymmdd-1]。

监控规则 → ods_log_info_d ods_log_info_d			
分区表达式 + 请蝓入 Q dt=\$[yyyymmdd-1]	 创建规则 ▲ 关联调度 试 责任人: 模板规则 (0) 自定义规则 (0) 	跑 订阅管理 >)	上一次校验结果
	规则名称 规则字段	發 强/弱 规则模板	模板路径

ii. 单击创建规则, 在对话框中单击添加监控规则。

创建规则			
模板规则自我	定义规则		
添加监控规则	快捷添加		
* 规则名称 :	请输入规则名称		删除
* 强弱号:	● 强 ○ 弱		
* 动态阈值:	○ 是		
* 规则来源 :	内置模板 🗸		
* 规则字段 :	表级规则(table)	~	
* 规则模板 :	表行数, 固定值	~	
* 比较方式:	不等于	~	
* 期望值:	0		
描述:			

配置表数据不为空的校验规则,选择规则强度为强、规则模板为表级规则、比较方式为不等于、期望值为0。

- iii. 配置完成后,单击**批量保存**。
- 5. 配置dw_user_info_all_d表规则。

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,所以该表无需进行数据质量监控规则的配置,以节省计算资 源。

6. 配置rpt_user_info_d表规则。

rpt_user_info_d表是数据汇总后的结果表。根据该表的数据,您可以进行表行数波动监测和针对主键进 行唯一值校验。

- i. 单击已添加的分区表达式模块的+, 配置表的分区表达式为dt=\$[yyyymmdd-1]。
- ii. 单击创建规则, 在添加监控规则对话框中添加列级规则。设置主键列(uid)为监控列,选择规则 模板为重复值个数, 固定值、强弱为弱、比较方式为小于以及期望值为1。

iii. 继续添加监控规则和表级规则,选择规则模板为表行数,7天波动率、强弱为弱,设置橙色阈
 值为1%、红色阈值为50%(此处阈值范围根据业务逻辑进行设置)。

模板规则自	定义规则	
添加监控规则	快捷添加	
* 规则名称 :	请输入规则名称	删除
* 强弱:	○ 强 ● 弱	
* 动态阈值:	○ 是 (● 否	
* 规则来源:	内置模板 🗸	
* 规则字段 :	表级规则(table) ~	
* 规则模板 :	表行数,7天波动率 🗸 🗸	
* 比较方式:	绝对值 ~	
* 波动值比较:	0% 25% 50% 75% 10	00%
	橙色阈值: 1 96 红色阈值: 50 9	6
描述:		

? 说明

- 橙色阈值和红色阈值必须大于0%。
- 此处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。

iv. 配置完成后, 单击**批量保存**。

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它层的数据,及时止损。数据质量还为您提供**任务查询**功能,以便查看已配置规则的校验结果,详情请参见查看监控任务。

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 Bl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加工数据。单击进入Quick Bl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。

🎨 Quick BI 🔮 专业版 இ		我的	工作空间	创作区	订阅	监控指标
Quick BI していたいです。 していたいです。 T作空间 加入更タ +	最近相关 与我共享	我的收藏		t		
▲ 个人空间 创建:2018/6/417:19:34 项目负责人: 1人 ◆管理员			暂时没有数据	哦 , 快快开始	使用吧!	
 2 默认空间 ◎ 创建: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		4	仪表 丰富的 自助费	扳 9可视化组件支 奴据分析	持
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: 送Exc 杂行时	表格 el式的体验制作 V报表	復

2. 选择数据源 > 新建数据源 > 云数据库 > MaxCompute。

😍 Quick Bl 💿 रु±फ्र		我的 工作空间 仓			
≔	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			X 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					a 0
念 电子表格	•		*		a 0
□ 自助取数 (公测) MEW		MySQL			
♠ 数据集	MaxCompute	MySQL	SQL Server	AnalyticDB	
■ 数据填报 (公測) ₩EW					
<→ 数据源					û û
	• *	\odot	C C C C C C C C C C C C C C C C C C C		@ ()

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Endpoint。

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。

1×03 T.1.		
添加MaxCompute数据》	泉	×
* 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 (j
ods_log_info_d				î (j
ods_raw_log_d				î (j
ods_user_info_d				î (j
result_table				↑ ①
rpt_user_info_d				i

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		
	* 位置:	ODPS	~	
			¥	和一种宁

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。

:=	数据集 全部 我的
🔊 🔊 🗢 🖉 🖉 🖉	根目录 > ODPS
■ 数据门户	名称 🜲
■■ 仪表板	rpt_user NEW
📚 电子表格	ipi_aser_inite_a
🕑 自助取数 (公测) MEW	自定义sql5 自定义sql5
★ 数据集 ★ ● ★	rpt rpt_user_trace_log

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。

i. 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.
3	助足		uid		regi	ion	devic
	Str. uid	☆ 编辑					
	str. regi						
	str. devi	2 克隆维度					
	str. geno str. age_	✔ 在分析中隐藏					
	^{Str.} zodi	⊙ 取消全部隐藏					
	Str. dt	★ 删除					
		+ 新建计算字段(维)	度)				
		よ 移动到	•				
		🕃 新建层次结构					
		↓ 下移					
E -	望 〇 默认	↓ 转换为度量			_	yyyyMMdd hh:mi	:ss
	[№] pv	≓ 维度类型切换	•	日期(源数据格式)		yyyy/MM/dd hh:r	ni:ss
				地理信息	•	уууу-MM-dd hh:r	ni:ss
				✓ 文本		yyyy/MM/dd	
				数字		ууууMMdd	

ii. 转换地理信息字段的维度类型。

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

str. region	ক্র		
str. device	⊘ 编辑		
str. gende	2) 克隆维度		
^{str.} age_r			
str. zodiac	✔ 在分析中隐藏		
str. dt	⊙ 取消全部隐藏		
	★ 删除		
度量			
- 🗁 默认	+ 新建计算字段(维度)		
[№] pv	よ 移动到 🍡 🕨		
	于 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。

i. 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 我的		61	▶ ∨ Q 共18个文件	十 新建数据组
相目录 > COPS				
	创建者 🗘	修改人/修改时间	数据源	新建仪表板
TPLUSET NEW TPLUSET.Info.d	August 1	144	No. of Belleville Textures	⊠ <mark>,, </mark> ≋ ;

? 说明

- 如果您使用的是标准版或高级版的Quick BI,则单击新建仪表板后默认进入常规模式仪表板。
- 如果您使用的是专业版Quick BI,则单击新建仪表板后可以选择进入常规模式或全屏模 式仪表板。当前文档以常规模式仪表板示例。

ii. 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		×	N9 📾 pv 🛛 🗙
<pre>prpt_user</pre>	ti dt(year)		日期/维度①
区间粒度	日期粒度		双击或拖动数据字段至此处
● 时间区间			过滤器
T∎ 2019 🛱 T₽	2019		tt(year)
开始于:	结束于:		
相对时间 🗸	相对时间 >		
2019	2019		
T• + ▼ 0	T ∳ + ▼ 0		
	取消	角定 ····································	

完成后可以看到当下数据。

^{⊳∨} 3198.88万

- iii. 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)

御 国 國 麗 國 [史]字 秋 @ 晋 礼 御 ⑦ 希 神 帝 神 帝 作 前 图	图表设计		💉 DODSE
	数据	Ĵz¥ł	商级
	选择数据源类型:	\$ 305	194 III III
》 3198 88万	值物度量	rpt_u	ser ~
	не 📰 ру	⇔ × <u></u>	
	Mana/Mar	NHER	🔲,,
线图-pt_user :	🛗 dt(day)	⊜€×	🛗 dt(month) 🛗 dt(week)
-20-30# -40-50# -30-40# 20#0XF 20190708	颜色图例维度		
2075 • 0.45 509 12 27/05	💷 age_range		region
10.75 10	过滤器		device
1477 1277 1077	💼 dt(year)		<pre>x gender x age_range</pre>
107 27 310012 201072 201073 201073 201073 201073 201073 201073 201073 201073		2	
dt(day)		-	: −
¢ ¢			
s			

iv. 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。



v. 完成配置后, 单击保存及预览, 即可看到展示效果。

		_{€3} c3					
		3198.88万 197 ³					
线图-rpt_user		eard - 20-30)岁 - 40-50岁 - 30-4	0岁 - 20岁以下 - 50岁以上			
22万 20万 18万 16万 16万	end.						
14/7 1275 1075 875		and a set of the set of			and an an an		
19 19 19 19	100	they and the set of the set of the	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	n tay) n(day) N ^{CI}	en ber ber ber ber ber		
色彩地图-rpt_user						pv	
					Ford		-
					110	Barrow Con	Ford
					evel	C ALLE	
533.575 - 666.875 400.375 - 533.575 26775 - 400.375							
133.7万 - 267万 4459 - 133.7万						7. 7 Mill 😥	

1.7. 通过Function Studio开发UDF

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

使用限制

目前仅华北2(北京)、华东2(上海)、华南1(深圳)和华东1(杭州)地域支持Function Studio。

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. 进入数据开发页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 选择工作空间所在地域后,单击相应工作空间后的进入数据开发。
- 2. 单击左上方的 图标,选择 2部产品 > 数据开发 > Function Studio。
- 3. 在工作空间页面, 单击导入Git工程。
- 在新建项目对话框中,输入Git地址、工程名和工程描述,并选择运行环境。
 其中Git地址仅支持配置为阿里云的Git地址,您可以在云计算服务平台中创建。
 新创建的工程默认未关联Git服务,会弹出设置对话框,请首先进行SSH KEY、Git Config和偏好设置的配置,单击保存。
 - 。选择SSH Key中的service为 code.aliyun.com , 单击生成sshKey, 即可生成Public key, 单击保存。
 - 填写Git Config中的User Name和Email, 单击保存。
 - 根据自身需求选择偏好设置中的编辑器字号,单击保存。

⑦ 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。

Θ	管理控制台	产品与	产品与服务 ▼				
	ш						
ñ	首页		在增加 SSH 密钥之前需要先	生成感钥。		2 +增加 SSH 密钥	
۵	个人设置						
	邮箱		标题	指纹	增加时间		
	通知		Function Studio		增加时间大约1小时之前	删除	
a,	SSH 公钥 🚺						
	密码						

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。

增加 SSH 密钥 在这里粘贴 SSH 公钥。如何生成	密钥请点击 SSH 帮助页面了解更多。
公钥	ssh-rsa AAAAB3N == Function Studio
标题	Function Studio
增加密钥	取消

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。

	Fx Function Studio		S					-		
工程	文件 编辑 版本 查看	调试设	18 2	彩布 模板 帮助 反馈		Unna	med 🗸			
പ	MaxCompute工程	⊿ ≔	🛓 Ip2R	legionjava ×						
יים	test (j)			package org.alidata.odps.udf;			1772-		i i	2
	> .alicode			import com.aliyun.odps.udf.UDF; import com.sithub.isrod.gauxy (PZrope:						ţ.
۲Ľ	settings			import com.github.jarod.qqwry.QQWry;						
	✓ src ✓ main			import java.io.IOException;						
	 ✓ java > com.github.jarod.qqwry ✓ org.alidata.odps.udf 			public final class IpDRegion extends UDF { private QDMry qeery;						
	▲ Ip2Region.java			public Ip2Region() throws IOException {						
	> resources			qqwry = new QQWry();						
	> test									
	> target			public String evaluate(String ip) {						
	🛓 .classpath			<pre>ir (ip = null) { return null; } iP/One zone = qawry.findIP(ip); </pre>						
	♦ .gitignore			return zone.getProvince();						
	≡ .project									
	pom.xml			<pre>public static vold main(String[] args) throws IOException (// Set (V)/Text of more simples String [p = " ;; System.out.println(new Ip2Region().evaluate(ip)); } }</pre>						

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

Run/Debug Configurations		×
添加 删除	Name: Unnamed	
✓ ➡ Application	t Main alaas	ann all data a data u déla 20 anian
Unnamed		organoata.odps.uorap.z.vegion
	VM options:	
	Program arguments:	
	Environment Variables:	
	JRE:	1.8 - SDK
	PORT:	7001
	机器:	4vCPU , 8G内存 ~ /
	Pre-Launch Option: 👔	清选择 ◆
	开启HOTCODE:	● 是 ○ 否
		Cancel Apply Run

3. 添加完成后,单击Run,即可看到输出的测试信息。

6	🗲 Function Studio	♥ ▼
工務	呈文件 编辑 版本 查看 调试 设	置 发布 模板 帮助 反馈
	MaxCompute工程	🛓 Ip2Region.java 🗴
Ŷ	test () > .alicode > .settings > src > main > java > com.github.jarod.qqwry > org.alidata.odps.udf () Ip2Region.java > resources > test > target () .classpath > .gitignore E .project	<pre>package org.alidata.odps.udf; import com.aliyun.odps.udf.UDF; import com.github.jarod.qqwry.IPZone; import com.github.jarod.qqwry.QQWry; import java.io.IOException; public final class Ip2Region extends UDF { private QQWry qqwry; public Ip2Region() throws IOException { qqwry = new QQWry(); } public String evaluate(String ip) { if (ip == null) { return null; } IPZone zone = qqwry.findIP(ip); return zone.getProvince(); } } </pre>
	輸出	
*	Starting Language Server 100% Sta 新江名 Disconnected from the target VM, address	rting Java Language Server - Importing Maven project(s) s: pcsoxsproductd58dlb15ptz4js83pgrae-80.r-app-data.aliyun.com, transport: socket
	🗐 OUT 🗼 RUN 🗮 PROBLEM 🗖 Termin	nal 🕴 Version Control

? 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。
 - i. 鼠标悬停至提交按钮,单击提交资源至DataStudio开发环境。



- ii. 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。
- iii. 单击确认。
- 提交函数至DataStudio开发环境。
 - i. 鼠标悬停至提交按钮,单击提交函数至DataStudio开发环境。
 - ii. 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。
 - iii. 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.简单用户画像分析(EMR版) 2.1. 准备环境

为保证您可以顺利完成本次实验,请您首先确保云账号已开通E-MapReduce(简称为EMR)、数据工场 Dat aWorks和数据存储OSS。

前提条件

- 注册阿里云账号。
- 进行个人实名认证或企业实名认证。
- 您在工作空间配置页面添加E-MapReduce计算引擎实例后,当前页面才会显示EMR目录。详情请参见配置 工作空间。
- 您已创建阿里云EMR集群,且集群所在的安全组中入方向的安全策略包含以下策略。
 - 授权策略: 允许
 - 协议类型: 自定义 TCP
 - 。端口范围: 8898/8898
 - 。 授权对象: 100.104.0.0/16
- 如果EMR启用了Ranger,则使用DataWorks进行EMR的作业开发前,您需要在EMR中修改配置,添加白名 单配置并重启Hive,否则作业运行时会报错Cannot modify spark.yarn.queue at runtime或Cannot modify SKYNET_BIZDATE at runtime。
 - i. 白名单的配置通过EMR的自定义参数,添加Key和Value进行配置,以Hive组件的配置为例,配置值如下。

```
hive.security.authorization.sqlstd.confwhitelist.append=tez.*|spark.*|mapred.*|mapred
uce.*|ALISA.*|SKYNET.*
```

⑦ 说明 其中 ALISA.* 和 SKYNET.* 为DataWorks专有的配置。

ii. 白名单配置完成后需要重启服务,重启后配置才会生效。重启服务的操作详情请参见重启服务。

 已开通独享调度资源组,并且独享调度资源组需要绑定EMR所在的VPC专有网络,详情请参见新增和使用 独享调度资源组。

⑦ 说明 仅支持使用独享调度资源组运行该类型任务。

背景信息

本次实验涉及的阿里云产品如下:

- E-MapReduce
- Dat aWorks
- **OSS**

操作步骤

- 1. 创建EMR集群。
 - i. 登录E-MapReduce控制台。

ii. 选择华东2(上海)区域,单击创建集群。

? 说明

- 由于源数据存储在华东2(上海),建议EMR集群创建在相同的区域。
- 您可以通过一键购买和自定义购买两种方式创建EMR集群,本文以自定义购买为例。
- iii. 在自定义购买 > 软件配置对话框中,选择集群类型为Hadoop,其它配置项默认无需修改。单 击下一步:硬件配置。
- iv. 在硬件配置对话框中,选择付费类型为按量付费,并进行网络配置和实例配置,单击下一步: 基础配置。
- v. 在基础配置对话框中, 输入集群名称, 并选择密钥对, 单击下一步: 确定。

EMR默认选项不开启挂载公网,创建集群后只能通过内网访问EMR集群。本次实验的Workshop操作 中不涉及挂载公网,直接单击**挂载公网说明**对话框中的**继续下一步**即可。如果您需要公网访问, 请进入ECS控制台挂载EIP。

- vi. 在确认对话框中,确认订单无误后,勾选《E-MapReduce服务条款》,单击创建。
- 2. 初始化集群。

购买成功后,即可进入集群管理页面进行查看,集群初始化需要几分钟的时间。

- i. 集群初始化成功后, 单击顶部菜单栏中的数据开发。
- ii. 在数据开发页面, 单击新建项目。
- iii. 在新建项目对话框中, 输入项目名称和项目描述。

② 说明 请使用主账号创建项目,该项目用于关联DataWorks工作空间。

ⅳ. 单击创建。

3. 创建DataWorks工作空间。

⑦ 说明 因本实验提供的数据资源都在华东2(上海),建议您将工作空间创建在华东2(上海),以避免工作空间创建在其它区域,添加数据源时出现网络不可达的情况。

i. 鼠标悬停至EMR控制台左上角的═图标,单击产品与服务 > 大数据(数加) > DataWorks。

- ii. 在左侧导航栏, 单击工作空间列表。
- iii. 在**工作空间列表**页面, 鼠标悬停至左上角的地域, 单击需要创建工作空间的地域。

iv. 单击创建工作空间,进行基本配置,单击下一步。

分类	参数	描述
	工作空间名称	工作空间名称的长度需要在3~27个字符,以字 母开头,且只能包含字母下划线和数字。
基本信息	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。
	模式	包括 简单模式 和 标准模式 ,本文以创建简单模 式的工作空间为例。
	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	设置是否允许下载数据开发中查询的数据结果。

v. 在选择引擎对话框中,选中E-MapReduce引擎,单击下一步。

DataWorks已正式商用,如果该地域没有开通,您需要首先开通正式商用服务。

vi. 在引擎详情对话框中, 配置各项参数。

创建工作空间		
✓ 基本配置		3 引擎详情
✓ E-MapReduce		
*实例显示名称		
* Access ID	请输入Access Id	
* Access Key	请输入Access Key	
* EmrUserID		
* 集群ID:	清选择	
* 项目ID:	请选择	
* YARN资源队列:	default	
* Endpoint:		

参数	描述
实例显示名称	自定义实例名称。
Access ID	已经授权可以访问EMR集群的账号的AccessKey ID。
Access Key	已经授权可以访问EMR集群的账号的AccessKey Secret。
EmrClusterID	集群ID,从EMR端获取。
集群ID	当前集群创建者的用户ID。
项目ID	当前集群下的项目ID。
YARN资源队列	当前集群下的队列名称。如果无特殊需求,请输入 <i>default</i> 。
Endpoint	EMR的Endpoint,从EMR端获取。

vii. 配置完成后,单击创建工作空间。

- 4. 购买OSS并创建Bucket。
 - i. 购买OSS,详情请参见<mark>开通OSS服务</mark>。
 - ii. 登录OSS控制台。
 - iii. 在左侧导航栏,单击Bucket列表。
 - iv. 在Bucket列表页面,单击创建Bucket。
 - v. 在创建Bucket对话框中,配置各项参数,单击确定。

⑦ 说明 此处需要选择区域为华东2(上海),更多参数说明请参见创建存储空间。

- vi. 单击相应的Bucket名称,进入Bucket的文件管理页面。
- vii. 在新建目录对话框中, 输入目录名, 单击确定。

② 说明 此处需要新建三个目录,分别存放同步过来的外部OSS数据源、RDS数据源和JAR资源。

2.2. 采集数据

本文为您介绍如何通过DataWorks采集日志数据至EMR引擎。

前提条件

开始本文的操作前,请准备好需要使用的环境。详情请参见准备环境。

背景信息

根据本次实验模拟的场景,您需要分别新建OSS数据源、RDS数据源,用于存储数据。同时需要新建私有 OSS数据源,用于存储同步后的数据。

新建OSS数据源

- 1. 进入数据源管理页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 单击相应工作空间后的进入数据集成。

如果您已在DataWorks的某个功能模块,请单击左上方的<mark></mark>图标,选择**全部产品 > 数据汇聚 > 数** 据集成,进入数据集成页面。

iv. 在左侧导航栏, 单击数据源, 进入工作空间管理 > 数据源管理页面。

- 2. 在数据源管理页面,单击右上方的新增数据源。
- 3. 在新增数据源对话框中,选择数据源类型为OSS。
- 4. 在新增OSS数据源对话框中,配置各项参数。此处您可以直接按照示例参数进行填写。

参数	描述
数据源名称	输入数据源名称,示例为oss_workshop_log。
数据源描述	对数据源进行简单描述。
Endpoint	输入Endpoint,示例为 http://oss-cn-shanghai-internal.aliyuncs .com 。
Bucket	输入Bucket名称,示例为new-dataworks-workshop。
AccessKey ID	输入访问密钥中的AccessKey ID,示例为LTAl4FvGT3iU4xjKotpUMAjS。
AccessKey Secret	输入访问密钥中的AccessKey Secret , 示例 为9RSUoRmNxpRC9EhC4m9PjuG7Jzy7px。

5.

6. 连通性测试通过后,单击完成。

新建RDS数据源

- 1. 在数据源管理页面,单击右上方的新增数据源。
- 2. 在新增数据源对话框中,选择数据源类型为MySQL。
- 3. 在新增MySQL数据源对话框中,配置各项参数。此处您可以直接按照示例参数进行填写。

参数	描述
数据源类型	选择 阿里云实例模式 。
数据源名称	输入数据源名称,示例为rds_workshop_log。
数据源描述	对数据源进行简单描述。
地区	选择RDS实例所在的区域,示例为华东2-上海。
RDS实例ID	输入rm-2ev0681lc7042g16u。
RDS实例主账号ID	输入5600815724958382。
数据库名	输入数据库名称,示例为workshop。
用户名	输入用户名,示例为workshop。
密码	输入密码,示例为workshop#2017。

4.

5. 连通性测试通过后,单击完成。

新建私有OSS数据源

本次实验将EMR引擎的数据存储在OSS数据源中。

- 1. 在数据源管理页面,单击右上方的新增数据源。
- 2. 在新增数据源对话框中,选择数据源类型为OSS。
- 3. 在新增OSS数据源对话框中,配置各项参数。

参数	描述
数据源名称	输入数据源的名称。
数据源描述	对数据源进行简单描述。
Endpoint	输入 http://oss-cn-shanghai-internal.aliyuncs.com 。
Bucket	您在环境准备中创建的OSS Bucket的名称,示例为dw-emr-demo。
AccessKey ID	当前登录账号的AccessKey ID,您可以进入 安全信息管理 页面复制 AccessKey ID。

参数	描述
AccessKey Secret	输入当前登录账号的AccessKey Secret。

4.

5. 连通性测试通过后,单击完成。

新建业务流程

- 1. 单击左上方的 图标,选择全部产品 > 数据开发 > DataStudio (数据开发)。
- 2. 在数据开发面板,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, 输入业务名称和描述。

↓ 注意 业务名称不能超过128个字符,且必须是大小写字母、中文、数字、下划线(_)以及英 文句号(.)。

- 4. 单击**新建**。
- 5. 进入业务流程开发面板,鼠标单击虚**拟节点**并拖拽至右侧的编辑页面。在**新建节点**对话框中,输入**节 点名称**为workstart,单击提交。

F 💿 🔊 🖈			
◇ 节点组 (9		
∨ 数据集成			
Di] 离线同步			
MaxCompute	新建节点		×
Sq ODPS SQL			
Sp ODPS Spark	节点类型:	虚拟节点	
Py PyODPS	节点名称:		
Sc ODPS Script	目标文件夹:	业务流程	
	1		
◇ 通用		した。 たいで、 たいで たいで たいで たいで たいで たいで たいで たいで	取消
Ch oss对象检查			
Sh Shell			
▼ 「虚拟节点			
解 跨租户节点			

以同样的方式新建两个离线同步节点,**节点名称**分别为Log2oss和User2oss。

6. 通过拖拽连线,将workstart节点设置为两个离线同步节点的上游节点。

Vi workstar	t 🕑
	
Di User2oss 🛛 😒	Di Log2oss 🥑

配置workstart节点

- 1. 在数据开发页面,双击相应业务流程下的虚拟节点。打开该节点的编辑页面,单击右侧的调度配置。
- 2. 在调度依赖区域,单击使用工作空间根节点,设置workstart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workstart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。

★ 调度配置						週
定时调度:						度配罟
具体时间:	00:04					
						版
cron表达式:	00 04 00 **?					4
依赖上—周期:						
N						
资源属性 ⑦ -						
调度资源组:	公共调度资源组					
油度広格 ②						
师授113款 ♥						
自动解析 💿 是	否 所析输入输出					
依赖的上游节点	请输入父节点输出名称或输出表名					
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	
			-	手动添加		

3. 配置完成后,单击左上方的图图标。

配置离线同步节点

- 1. 同步RDS数据源的用户信息至自建的OSS。
 - i. 在**数据开发**页面,双击User2oss节点,进入节点配置页面。

ii. 选择数据来源。

Di User2oss 🔹 🚣 EMR_v	vorkshop	
	3 💿 🔒 🖾	
01 选择数据源	数提来源	
	在这里配置数据的来源端和写入端:	可以是默认的数
* 数据源	MySQL V rds_workshop_log V (0
*表	`ods_user_info_d` ×	
	添加跌搬源 +	
数据过滤	请参考相应SQL语法填写where过滤语句(不要填写where关键字)。该过 (滤语句通常用作增量同步	0
切分键	uid	0
	数据预览	

参数	描述
数据源	选择 MySQL > rds_workshop_log 数据源。
表	选择数据源中的ods_user_info_d。
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。此处可以不填 写。
切分键	建议使用主键或有索引的列作为切分键,仅支持类型为整型的字段。此处 设置切分键为uid。

iii. 选择数据去向。



参数	描述
数据源	选择前文创建的OSS数据源,此处示例为 OSS > dw_emr_demo 数据 源。
Object前缀	根据您自建OSS的目录进行输入,示例 为ods_user_info_d/user_\${bizdate}/user_\${bizdate}.txt。
文本类型	选择text类型。
列分隔符	输入列分隔符为 。
编码格式	默认为UTF-8格式。
null值	表示null值的字符串,此处可以不填写。
时间格式	时间序列化格式,此处可以不填写。
前缀冲突	此处选择 替换原有文件 。

iv. 配置字段映射。

02 字段映射								
	源头表字段	美型				目标表序列	类型	同名映
	uid	VARCHAR	•	•	•	第0列	自定义	同行映開
	gender	VARCHAR	•)ı	•	第1列	自定义	
	age_range	VARCHAR	•)ı	•	第2列	自定义	
	zodiac	VARCHAR	•		•	第3列	自定义	
	添加—行+							

v. 配置通道控制,单击工具栏中的 图标。

	訂以配置作业的传输速率和错误纪录数末控制整个数据同	步过程:数据同步文档
*任务期望最大并发数	2 🗸	0
• 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 ??

vi. 单击工具栏中的**圆**图标,在已有的脚本中手动添加参数"writeSingleObject": "true"和"suffix": ".txt"。



? 说明

- writeSingleObject和suffix参数仅支持脚本模式进行添加。
- 存储的路径object需要与自建OSS中的目录一致。

vii. 配置完成后,单击工具栏中的凹图标。

- 2. 同步OSS数据源的日志信息至自建的OSS。
 - i. 在数据开发页面,双击Log2oss节点,进入节点配置页面。

ii. 选择数据来源。

参数	描述			
数据源	选择OSS > oss_workshop_log数据源。			
Object前缀	输入user_log.txt。			
文本类型	选择text类型。			
列分隔符	输入列分隔符为 。			
编码格式	默认为UTF-8格式。			
null值	表示null值的字符串,此处可以不填写。			
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。			

iii. 选择数据去向。

参数	描述
数据源	选择前文创建的OSS数据源,此处示例为OSS > dw_emr_demo数据源。
Object前缀	根据您自建OSS的目录进行输入,示例 为ods_raw_log_d/user_log_\${bizdate}/user_log _\${bizdate}.txt。
文本类型	选择text类型。
列分隔符	输入列分隔符为 。
编码格式	默认为UTF-8格式。
null值	表示null值的字符串,此处可以不填写。
时间格式	时间序列化格式,此处可以不填写。
前缀冲突	此处选择 替换原有文件 。

iv. 配置字段映射。

↓ 注意 源数据表中只有一列数据,此处需要删除其它映射过来的空列。

v. 配置通道控制,单击工具栏中的回图标。

- vi. 单击工具栏中的operation图标,在已有的脚本中手动添加参数"writeSingleObject": "true"和"suffix": ".txt"。
 - ? 说明
 - writeSingleObject和suffix参数仅支持脚本模式进行添加。
 - 存储的路径object需要与自建OSS中的目录一致。

vii. 配置完成后,单击工具栏中的凹图标。

新建表

- 1. 在数据开发页面打开新建的业务流程,右键单击EMR,选择新建 > EMR Hive。
- 2. 在新建节点对话框中, 输入节点名称, 单击提交。

此处需要新建两个EMR Hive节点(ods_user_info_d和ods_raw_log_d),分别新建存储同步过来的OSS 日志数据和RDS日志数据的两张表。



- 3. 分别在EMR Hive节点中,选择EMR引擎并输入建表语句,单击保存并运行各建表语句。
 - 新建ods_user_info_d表。

双击ods_user_info_d节点,在右侧的编辑页面输入下述建表语句。



⑦ 说明 上述代码中的location为示例路径,需要输入您建立的文件夹的路径名称。

○ 新建ods_raw_log_d表。

双击ods_raw_log_d节点,在右侧的编辑页面输入下述建表语句。

```
--创建oss日志对应目标表
CREATE EXTERNAL TABLE IF NOT EXISTS ods_raw_log_d
(
    `col` STRING
) PARTITIONED BY (
    dt STRING
);
ALTER TABLE ods_raw_log_d ADD IF NOT EXISTS PARTITION (dt=${bizdate})
LOCATION 'oss://dw-emr-demo/ods_raw_log_d/user_log_${bizdate}/';
```

⑦ 说明 上述代码中的location为示例路径,需要输入您建立的文件夹的路径名称。

4. 查看数据同步结果。

建表语句运行成功后,分别在两个EMR Hive节点中输入查询语句。

⑦ 说明 查询语句中的分区列需要更新为业务日期。例如,任务运行的日期为20191107,则业务日期为20191106,即任务运行日期的前一天。

○ 查询ods_user_info_d表的数据。

SELECT * from ods_user_info_d where dt=业务日期; --业务日期为任务运行日期的前一天。

选择EMR引擎 请选择EMR引擎,或者点此新建引擎										
1	6				a circy		0_0/000	+ L = = = = =		
1	<pre>18 SELECT * from ods_user_info_d where dt=20191106;</pre>									
运	行日志	组	吉果[1]	×						
/	A			в		С	D		E	
1	ods_user_info_	d.uii 🗸	ods_user	_info_	d.g€ ✔ :od:	s_user_info_d.ac 🗸	ods_user_info_	d.zo 🗸 🖉	ods_user_info_d.dt 🗸	
2	001635981082	21	女		30-	-40岁	巨蟹座		20191106	
3	001635981415	59	女		30-	-40岁	巨蟹座		20191106	
4	001635981749	97	女		30-	-40岁	巨蟹座		20191106	
5	001d9e786304	19	女		40-	-50岁	双鱼座		20191106	
6	001d9e786638	37	女		40-	-50岁	双鱼座		20191106	
7	001d9e786972	25	女		40-	-50岁	双鱼座		20191106	
8	001dce298354	4	女		20-	-30岁	水瓶座		20191106	
9	001dce298688	2	女		20-	-30岁	水瓶座		20191106	
10	0026c84ad120	6	X +		20	タ以下	大杆崖		20191106	
11	0026684ad454	4	× ×		20	タ以下 1911万	大杆崖		20191106	
12	0026c84ad/88	2	x •		20		大杆崖		20191106	
13	002726101046	10	55 10		50		双丁座		20191106	
14	0027261d1379	10	95 10		50		从丁座 羽之広		20191106	
15	002/20101/13	0	5		50	9WL	从于座		20191100	

○ 查询ods_raw_log_d表的数据。

SELECT * from ods_raw_log_d where dt=**业务日期;** --**业务日期为任务运行日期的前一天。**



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。详情请参见加工数据。

2.3. 加工数据

本文将为您介绍如何通过DataWorks中的EMR Hive节点加工采集的日志数据。

前提条件

开始本实验前,请首先完成采集数据中的操作。

在OSS上传资源

- 1. 下载ip2region-emr.jar存放至本地。
- 2. 登录OSS控制台。
- 3. 在左侧存储空间列表中,单击目标存储空间(示例为dw-emr-demo)。
- 4. 单击**文件管理**,打开在 环境准备章节新建的用于存储JAR资源的目录,示例的目录名为ip2region。
- 5. 单击上传文件,在上传文件对话框中,设置上传文件的参数。

	Q. 1938/XX	i、控制台、API、解决方案和资	11.50.55 奥用 工単 备業 企业 支持 首网 🖾 🎝 🏹 🧿 滴体 🌘
对象存储	对象存储 / dw-emr-demo / 文件管理		上传文件 ④
RIS PHESIS Pucket SR: Q House Ho	dw-emr-demo 概定 文件管理 2 話始近章 域名管理 图片处理 事件通知 @ 1 10文件 解練目 卵片管理 健収 単原原件 単 解新 文件名 (Object Name) く 1 ip2region/ ip2region-emrjar	融計算 云伊緒阿关 管 文件大小 4.616MB	上作別 当前目录 指定目录 555://dw-emr-demo/p2region/
< 1/3 >			注意: Bucket 下若存在同各文件,將被新上传的文件覆盖。
参数		描述	
上传到		选择 当 前 <i>mo/ip2</i>	前目录 , 示例的目录路径为 <i>oss://dw-emr-de</i> h2region/。
文件ACL		默认为 组 的读写机	继承Bucket ,即单个文件的读写权限以Bucket 权限为准。
上传文件		单击直捂 件。	接上传,选择已下载的ip2region-emr.jar文

设计业务流程

业务流程节点间依赖关系的配置请参见采集数据。

双击新建的业务流程打开编辑页面, 鼠标单击EMR Hive并拖拽至右侧的编辑页面。在**新建节点**对话框中, 输入**节点名称**, 单击提交。

此处需要新建3个EMR Hive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。



配置EMR Hive节点

- 1. 配置ods_log_info_d节点。
 - i. 双击ods_log_info_d节点,进入节点配置页面。
 - ii. 在节点编辑页面,编写如下语句。

```
    ⑦ 说明 如果您的工作空间绑定多个EMR引擎,需要选择EMR引擎。如果仅绑定一个EMR引擎,则无需选择。
    --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 'B容器返回状态码',
        bytes STRING COMMENT '地域,根据ip得到',
        region STRING COMMENT 'http请求类型',
        url STRING COMMENT 'url',
        protocol STRING COMMENT 'bttp协议版本号'.
```

```
Prococor origino commun
                          THEORING MINIT J
 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
):
create function getregion as 'org.alidata.emr.udf.Ip2Region'
using jar 'oss://dw-emr-demo/ip2region/ip2region-emr.jar';
ALTER TABLE ods_log_info_d ADD IF NOT EXISTS PARTITION (dt=${bizdate});
set hive.vectorized.execution.enabled = false;
INSERT OVERWRITE TABLE ods log info d PARTITION (dt=${bizdate})
SELECT ip
  , uid
  . tm
  , status
  , bytes
  , getregion(ip) AS region -- 使用自定义UDF通过ip得到地域。
  , regexp_extract(request, '(^[^ ]+) .*') AS method --通过正则把request差分为三个字段
  , regexp_extract(request, '^[^ ]+ (.*) [^ ]+$') AS url
  , regexp_extract(request, '.* ([^ ]+$)') AS protocol
  , regexp extract(referer, '^[^/]+://([^/]+){1}') AS referer --通过正则清洗refer,
得到更精准的url。
  , CASE
   WHEN lower(agent) RLIKE 'android' THEN 'android' --通过agent得到终端信息和访问形式
0
   WHEN lower(agent) RLIKE 'iphone' THEN 'iphone'
   WHEN lower(agent) RLIKE 'ipad' THEN 'ipad'
   WHEN lower (agent) RLIKE 'macintosh' THEN 'macintosh'
   WHEN lower (agent) RLIKE 'windows phone' THEN 'windows phone'
   WHEN lower (agent) RLIKE 'windows' THEN 'windows pc'
   ELSE 'unknown'
 END AS device
  , CASE
   WHEN lower(agent) RLIKE '(bot|spider|crawler|slurp)' THEN 'crawler'
   WHEN lower (agent) RLIKE 'feed'
   OR regexp extract (request, '^[^]+ (.*) [^]+$') RLIKE 'feed' THEN 'feed'
   WHEN lower(agent) NOT RLIKE '(bot|spider|crawler|feed|slurp)'
   AND agent RLIKE '^[Mozilla|Opera]'
   AND regexp_extract(request, '^[^ ]+ (.*) [^ ]+$') NOT RLIKE 'feed' THEN 'user'
   ELSE 'unknown'
  END AS identity
  FROM (
   SELECT SPLIT(col, '##00')[0] AS ip
   , SPLIT(col, '##00')[1] AS uid
    , SPLIT(col, '##00')[2] AS tm
    , SPLIT(col, '##00')[3] AS request
    , SPLIT(col, '##00')[4] AS status
    , SPLIT(col, '##00')[5] AS bytes
    , SPLIT(col, '##00')[6] AS referer
    , SPLIT(col, '##00')[7] AS agent
   FROM ods raw log d
 WHERE dt = {bizdate}
```

) a;

- ⅲ. 单击工具栏中的Ⅲ。
- 2. 配置dw_user_info_all_d节点。
 - i. 双击dw_user_info_all_d节点,进入节点配置页面。
ii. 在节点编辑页面,编写如下语句。

② **说明** 如果您的工作空间绑定多个EMR引擎,需要**选择EMR引擎**。如果仅绑定一个EMR引擎,则无需选择。

```
--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
);
ALTER TABLE dw_user_info_all_d ADD IF NOT EXISTS PARTITION (dt = ${bizdate});
INSERT OVERWRITE TABLE dw user info all d PARTITION (dt=${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 = ${bizdate}
) a
LEFT OUTER JOIN (
 SELECT *
 FROM ods_user_info_d
 WHERE dt = {bizdate}
) b
ON a.uid = b.uid;
```

- ⅲ. 单击工具栏中的Ⅲ。
- 3. 配置rpt_user_info_d节点。
 - i. 双击rpt_user_info_d节点,进入节点配置页面。

ii. 在节点编辑页面,编写如下语句。

⑦ 说明 如果您的工作空间绑定多个EMR引擎,需要选择EMR引擎。如果仅绑定一个EMR引擎,则无需选择。

```
--创建RPT层表
CREATE TABLE IF NOT EXISTS rpt user info d (
 uid STRING COMMENT '用户ID',
 region STRING COMMENT '地域,根据ip得到',
 device STRING COMMENT '终端类型 ',
 pv BIGINT COMMENT 'pv',
 gender STRING COMMENT '性别',
 age_range STRING COMMENT '年龄段',
 zodiac STRING COMMENT '星座'
)
PARTITIONED BY (
 dt STRING
);
ALTER TABLE rpt user info d ADD IF NOT EXISTS PARTITION (dt=${bizdate});
INSERT OVERWRITE TABLE rpt user info d PARTITION (dt=${bizdate})
SELECT uid
 , MAX(region)
  , MAX(device)
  , COUNT(0) AS pv
  , MAX(gender)
  , MAX(age range)
  , MAX(zodiac)
FROM dw user info all d
WHERE dt = ${bizdate}
GROUP BY uid;
```

ⅲ. 单击工具栏中的Ⅲ。

提交业务流程

- 1. 在业务流程的编辑页面,单击),运行业务流程。
- 2. 待业务流程中的所有节点后出现 💽 , 单击 🖬 , 提交运行成功的业务流程。
- 3. 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。
- 4. 单击提交。

在生产环境运行任务

- 任务发布成功后,单击右上角的运维中心。
 您也可以进入业务流程的编辑页面,单击工具栏中的前往运维,进入运维中心页面。
- 2. 单击左侧导航栏中的周期任务运维 > 周期任务,进入周期任务页面,单击workstart虚节点。
- 3. 在右侧的DAG图中,右键单击workstart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, 自动跳转至补数据实例页面。
- 5. 单击刷新,直至SQL任务全部运行成功即可。

后续步骤

现在,您已经学习了如何创建EMR Hive节点、如何处理原始日志数据。您可以继续下一个教程,学习如何在数据地图模块开启元数据收集功能,并查看数据表信息。详情请参见收集和查看元数据。

2.4. 收集和查看元数据

本文为您介绍如何在数据地图模块开启元数据收集功能,并查看数据表信息。

前提条件

开始本实验前,请首先完成加工数据中的操作。

背景信息

元数据是数据的描述数据,可以为数据说明其属性(名称、大小、数据类型等),或结构(字段、类型、长度等),或其相关数据(位于何处、拥有者、产出任务、访问权限等)。DataWorks中元数据主要指库、表相关的信息,元数据管理对应的主要应用是数据地图。

开启元数据收集

- 1.
- 2.
- _.
- 3.
- 4.
- 5.
- 6.
- 7.
- 8. 在E-MapReduce元数据采集页面,单击新建的采集器后的运行全量获取。

单击页面右上角的刷新,待EMR采集实例的运行状态显示为收集成功即可。

查看数据表信息

- 1. 在当前页面的顶部菜单栏,单击全部数据。
- 2. 在全部数据页面,单击E-MapReduce。
- 3. 在E-MapReduce页签下,单击表名(rpt_user_info_d),查看该表的详情。

您也可以在顶部搜索框中输入关键字进行搜索,查看E-MapReduce表详情。

4. 单击血缘信息,查看该表的上下游血缘详情。

后续步骤

现在,您已经学习了如何在数据地图模块开启元数据收集功能,并查看数据表信息。您可以继续下一个教程,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。详情请参见配置数据质量监控。

2.5. 配置数据质量监控

⑦ 说明 全量采集E-MapReduce元数据后,系统会开启自动增量采集,自动同步表中新增的元数据。

本文为您介绍如何配置表ods_log_info_d的数据质量监控规则。

前提条件

在进行本实验前,请首先采集元数据,详情请参见收集和查看元数据。

操作步骤

- 1. 进入数据开发页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏,单击工作空间列表。
 - iii. 选择工作空间所在地域后,单击相应工作空间后的进入数据开发。
- 2. 进入ods_log_info_d表的监控规则页面。
 - i. 单击左上角的≣图标,选择全部产品>数据质量。
 - ii. 在左侧导航栏,单击**监控规则**,从数据源下拉列表中选择EMR。
 - iii. 单击ods_log_info_d表后的配置监控规则。
- 3. 添加分区表达式。
 - i. 在已添加的分区表达式模块,单击+。
 - ii. 添加分区对话框中,选择分区表达式为dt=\$[yyyymmdd-1],并选择相应的数据质量插件。
 - iii. 单击计算,即可查看调度结果。
 - ⅳ. 确认无误后,单击**确认**。
- 4. 创建规则。
 - i. 选中分区后,单击右上角的**创建规则**。
 - ii. 在模板规则对话框中,单击添加监控规则。
 - iii. 配置监控规则。

创建规则			
模板规则 自定!	之规则		
添加监控规则	央捷添加		
* 规则名称:	请输入规则名称		删除
* 强弱 :	• 强) 弱		
* 动态阈值:	○ 是 ● 否		
* 规则来源:	内置模板 🗸 🗸		
* 规则字段 :	表级规则(table)	~	
* 规则模板 :	表行数,固定值	~	
* 比较方式:	大于	~	
* 期望值:	0		
描述:			
批量添加取消	 		
参数		描述	
规则名称		新建规则的名称。	
强弱		设置规则的强度为强。	
		根据自身需求,选择是否开启动态	阈值。
动态阈值		⑦ 说明 您需要购买DataWo	rks企业版及
			77 170 0
		包括内置模板和规则模板库。	
规则来源		⑦ 说明 您需要购买DataWc 以上版本,才可以选择规则模板	ırks企业版及 库 。
规则字段		请选择 表级规则(table)。	

参数	描述
规则模板	请选择 表行数,固定值 。
比较方式	请选择大于。
期望值	设置为0,即比较方式为期望值大于0。

ⅳ. 配置完成后, 单击批量添加。

5. 进行试跑。

i. 单击页面右上角的**试跑**。

ii. 在试跑对话框中,选择调度时间和资源组,单击试跑。

iii. 试跑完成后,单击**试跑成功! 点击查看试跑结果**,即可跳转至试跑结果页面。

- 6. 进行关联调度。
 - i. 在ods_log_info_d表的监控规则页面,单击关联调度。

ii. 在关联调度对话框中,输入节点ID或节点名称,单击添加。

iii. 添加完成后,即可完成与调度节点任务的绑定,则任务实例运行完成都会触发数据质量的检查。

- 7. 配置任务订阅。
 - i. 在ods_log_info_d表的监控规则页面,单击订阅管理。
 - ii. 在订阅管理对话框中,设置订阅方式和接受对象。
 目前支持的订阅方式包括邮件通知、邮件和短信通知、钉钉群机器人和钉钉群机器人@ALL。
 - iii. 设置完成后,单击保存,您可以进入我的订阅页面进行查看和修改。

3.搭建互联网在线运营分析平台 3.1. 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。本教程帮助您了解并操作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

本文的示例基于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(例如, Android、iPad、iPhone和PC等)分别统 计。

⑦ 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(UV)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:环境准备。
- 步骤二: 数据准备。
- 步骤三: 新建数据表。
- 步骤四:设计工作流。
- 步骤五:节点配置。
- 步骤六: 任务提交与测试。
- 步骤七: 数据可视化展现。

3.2. 环境准备

本文为您介绍开始本教程前的环境准备工作,需要开通表格存储(TableStore)、大数据计算服务(MaxCompute)、数据工厂(DataWorks)和智能分析套件(Quick BI)。

前提条件

- 已注册阿里云账号。如果您还没有注册阿里云账号,请进入阿里云官网,单击免费注册,即可进入阿里云 账号注册页面创建新的阿里云账号。
- 已实名认证。如果您还没有实名认证,请进入实名认证页面对账号进行实名认证。

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

⑦ 说明 在本教程中,表格存储服务选择华北2(北京)。

操作步骤

- 1. 创建表格存储实例。
 - i. 进入表格存储TableStore产品详情页, 单击免费开通。
 - ii. 在云产品开通页页面, 勾选我已阅读并同意表格存储(按量付费)服务协议并单击立即开通。

云j	产品开通页	当前环境是预发环境
志	₹格存储 (按	量(7费)
基本	固定模块	要指存储
配置	开通说明	表格存储产品免费开通,开通后即可使用
	☑ 我!	已阅读并同意 表格存储 (按量付费) 服务协议
		立即开通

iii. 单击管理控制台。

_	-	_
_		T
	×.	

确认订单	开通完成
 	

iv. 单击创建实例。在创建实例页面,选择地区为华北2(北京)。填写实例名称,实例规格请选 择容量型实例,单击确定。

创建实例		×
地区:	华北2 (北京)	
实例名称: *	Instance Name	
实例规格:	容量型实例 > 容量型实例:适合对于读性能不敏感,期望高性价比海量 数据存储。	
实例注释:	实例注释最多256个字。	
✓ 创建实例免费, 计费按使用量	收取。可以开启更多丰富数据访问管理能力。	
	确定取消	

⑦ 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数据处理中也会被使用,本例中为workshop-bj-001,关于实例的详细解释请参见实例。

v. 完成创建后,单击左侧导航栏全部实例可以看到您刚刚创建的实例,状态为运行中。

2. 开通大数据计算服务MaxCompute。

i. 进入MaxCompute产品详情页, 单击立即购买。

ii. 选择按量计费,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。

⑦ 说明 MaxCompute区域与表格存储区域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)**。本教程中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. 开通DataWorks。

- i. 进入DataWorks产品详情页,单击立即购买。
- ii. 选择区域为**华东2(上海)**,单击**立即购买**。

⑦ 说明 MaxCompute区域与表格存储区域相同可以节省您的流量费用,因此您也可以选择 区域为华北2(北京)。本教程中MaxCompute区域选择为华东2(上海),以便为您展示跨 地域访问数据的使用过程。

4. 创建DataWorks工作空间。

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

= (-)阿里云	华	东1(杭州) 💙 2		Q 搜索		费用	工单	备案	企业
DataWorks		DataWorks / 工作空间列表							
概览		当前使用的是企业版,版本到期	旧为 2020年5月10日 。					版	本升级
资源组列表		创建工作空间 请输入工作	F空间/显示名	٩					
计算引擎列表	^	工作空间名称/显示名	模式	创建时间	管理员	状态		开通服务	p.
MaxCompute Graph Compute			简单模式 单环境	2019-11-20 (15:32:43)	$(a_{1},a_{2},a_{3},a_{$	✓ 正常		w.	
交互式分析		10.000	标 准模式 开发跟生产隔离	2019-11-12 (18:07:43)	10000	✔ 正常		w.	

ii. 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<mark>简单模式和标准模式的区别</mark>。

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) ~	
描述		
高级设置	_	
* 能下载Select结果 ❷	π	
下一步 取	- Fire	
⑦ 说明 工作空间名称	家全局唯一,建议您使用易于区分的名称。	

iii. 进入选择引擎界面,选择相应引擎后,单击下一步。

✓ 基:	本配置 2 选择引擎 3 引	挙详 情
选择Data	aWorks服务	
>	数据集成、数据开发、运维中心、数据质量 忽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 ○ 包年包月 ● 按量付费 开发者版 去购买 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。 	
	韩 章 交互式分析 包年包月 去购买 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。	
	2	

iv. 进入**引擎详情**页面,填写选购引擎的配置。

选择计算引擎服务为MaxCompute、按量付费。

创建工作空间		
✓ 基本配置		引擎
✓ MaxCompute		
* 实例显示名称	注絵)の周日二々分	
* Quota组切换	按量付费默认资源组	~
* MaxCompute数据类型 🥝	1.0数据类型(面向已有使用1.	0数据类型用户) 🗸
* MaxCompute项目名称 🥥		
* MaxCompute访问身份 🥥	阿里云主账号	~
创建工作空间	モー步取消	Э Н Ш Н Ш
分类	配置	
	实例显示名称	实例名称不能超过27个字符,仅支持字母开头, 仅包含字母、下划线和数字。
	Quota组切换	Quota用来实现计算资源和磁盘配额。
MaxCompute	MaxCompute数据类型	MaxCompute项目的数据类型版本。
	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。

v. 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。

- 5. 开通Quick BI。
 - i 讲λOuick RI产品详情而 单击管理控制台

ii. 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,单击免费试用。

⑦ 说明 您可以选择使用个人空间或默认空间,推荐您使用默认空间。

3.3. 数据准备

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

前提条件

- 创建华北2(北京)区域的表格存储实例,同时记录实例名称和实例访问地址。单击表格存储控制台中的 实例名称,即可获得实例访问地址。对于跨区域的访问,建议您使用公网地址。详细操作请参见环境准 备。
- 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。

```
② 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:

- o Mac下载地址
- o Linux下载地址
- Windows7 64位下载地址
- 2. 配置Demo环境。

完成下载后, 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
👢 conf	2019/6/17 10:07	文件夹	
workshop_demo.exe	2017/12/18 16:58	应用程序	12,367 KB

app.conf文件内容示例如下。

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

其中,需要配置的参数如下:

○ endpoint:表格存储实例的访问网络地址,建议您使用公网地址。

○ instanceName: 表格存储实例的名称。

- accessKeyld和accessKeySecret:访问阿里云的密钥。
- 3. 启动Demo准备测试数据。
 - i. 启动Windows CMD命令行工具,进入您解压缩Demo包的路径,执行如下语句查看Demo包命令用法。

workshop_demo.exe -h

该命令会列出该demo的相关命令,如下。

workshop demo.exe -h

- * 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)。
- ii. 执行如下命令生成准备数据。

workshop_demo.exe prepare

结果如下。

C:\Users	<pre>\workshop_demo; workshop_demo.exe prepare</pre>
OTSObjectAlreadyExist	Requested table already exists.
OTSObjectAlreadyExist	Requested table already exists.
Prepare the metric dat	a internet and the second s
Prepare the metric dat	ta di
Prepare the metric dat	a a a a a a a a a a a a a a a a a a a
Prepare the metric dat	a a a a a a a a a a a a a a a a a a a
Prepare the metric dat	ta di
Prepare User data	
finished one round	
total insert data coun	nt is: 41757

在此过程中, Demo包会自动帮助您在表格存储中创建表, 结构如下:

○ 原始日志数据表: user_trace_log

列名	类型	说明
md5	STRING	用户uid的md5值undefined前8 位,表格存储主键。
uid	STRING	用户uid,表格存储主键。
ts	BIGINT	用户操作时间戳,表格存储主 键。
ip	STRING	IP地址。
status	BIGINT	服务器返回状态码。
bytes	BIGINT	返回给客户端的字节数。

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

○ 分析结果表: analysis_result

列名	类型	说明
metric	STRING	报表的类型:'new'、 'day_active'、'month_active'、 'day_pv'、'month_pv' <i>,</i> 表格存 储主键。
ds	STRING	时间yyyy-mm-dd或yyyy-mm <i>,</i> 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

通过如下语句查询指定用户在某一日期指定条数的明细数据。表格数据的日期对应于您创建表格的时间。

raw \${userid} \${date} \${Top**条数**}

其中, \${userid}为用户ID, \${date}为指定日期, \${Top条数}为指定查询条数。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\workshop_c	demo>wo	prkshop_de	emo.exe raw	00010 '	2019-06-
5" 20						
uid	Γ)ate	bytes	customiz	e_event	
device	ip st	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	I PM	252	backstage	365)
iPad min2	157.249.67.241	200		ios11		
00010	2019-06-14 11:21:30) PM	427	browse	trave]	
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 11:16:03	3 PM	764	switch	185	5
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 11:06:03	3 PM	436		click	ζ
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 10:36:54	I PM	131		click	ζ
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 10:22:20	S PM	778	switch	73	}
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 10:06:29	PM (535	backstage	179)
iPad min2	157.249.67.241	200		ios11		
00010	2019-06-14 09:56:11	PM	668		click	ζ
iPad min2	157.249.67.241	200		ios11		
00010	2019-06-14 09:20:45	5 PM	354		regist	
iPhone6s	222.133.108.234	200		ios10	-	
00010	2019-06-14 09:15:37	' PM	989		click	ζ
iPad min2	157.249.67.241	200		ios11		
00010	2019-06-14 08:51:17	' PM	460	logout	462	2
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 08:26:00	S PM	887	comment	funny	
iPad min2	157.249.67.241	200		ios11		
00010	2019-06-14 08:10:34	I PM	278	browse	finance)
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 07:56:00) PM	480		click	< compared with the second sec
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 07:30:11	PM	68		click	(
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 07:15:09	PM (398	browse	news	•
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 07:11:21	PM	21		click	(
iPhone6s	222.133.108.234	200		ios10		
00010	2019-06-14 06:35:07	PM	207	browse	photo)
iPhone7 Plus	61.103.79.217	200		ios11		
00010	2019-06-14 06:24:43	B PM	261		regist	
iPhone7 Plus	61.103.79.217	200		ios11		

⑦ 说明 由于表格存储是SchemaFree结构,表的属性列不需要预先定义。Customize_Event 中不同的事件对应了不同的内容,因此Demo中将事件、内容进行对齐显示。

○ 报表结果查询

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

您可以使用	workshop	demo.exe	day_	active	命令查看 日	ヨ活数据。
-------	----------	----------	------	--------	--------	-------

3.4. 数据建模与开发

3.4.1. 新建数据表

本文为您介绍如何在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

- 已开通MaxCompute服务并创建DataWorks工作空间(本教程使用为简单模式工作空间),详情请参见<mark>环 境准备</mark>。
- 已具备访问Tablestore数据的权限。当MaxCompute和Tablestore的所有者是同一个账号时,您可以单击 此处一键授权。如果不是,您可以自定义授权,详情请参见OTS外部表。

操作步骤

- 1. 进入DataWorks数据开发界面。
 - i. 进入DataWorks工作空间列表,选择区域为华东2(上海)。
 - ii. 单击已创建好的工作空间后的进入数据开发,进入工作空间的数据开发界面。
- 2. 新建业务流程。

i. 右键单击**业务流程**,选择新建业务流程。



- ii. 填写**业务名称和描述**,单击新建。本教程中,业务流程名为Workshop。
- 3. 新建数据表。
 - i. 创建外部表ots_user_trace_log。
 - a. 单击新建的业务流程Workshop,右键单击**MaxCompute**,选择**新建 > 表**,输入表 名ots_user_trace_log,单击**提交**。

▼ 业务流程		
🗸 🚠 Workshop		
> 럳 数据集成		
> 🚺 MaxCompute		0000 001
2 通用	新建	ODPS SQL
、 💷 白完议	新建文件夹	SQL组件节点
		ODPS Spark
		PyODPS 2
		ODPS Script
		ODPS MR
		PyODPS 3
		表
		资源 >
		函数

b. 填写创建表的中文名,然后单击DDL模式。

· 表					ts_user_t	race_log ×	Sq query	Sq rpt_user	_trace_log
DDL模式 从生产环境加载									
2	表名	ots_user_trace_lo	g						
10. 10.	******	-							
基本属性									
中文名	ots_user_trace_loo								
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-beijin	g.ots.aliyuncs.cor	m/					一键授权

c. 在DDL模式页面, 输入建表语句, 单击生成表结构。

外部表ots_user_trace_log的建表语句如下。

```
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/评论',
   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/';
```

 STORED BY:必选参数,值为 com.aliyun.odps.TableStoreStorageHandler ,是 MaxCompute内置处理Tablestore数据的StorageHandler,定义了MaxCompute和 Tablestore的交互。

- SERDEPROPERITES:必选参数,是提供参数选项的接口,在使用 TableStoreStorageHandler时,以下选项必须指定:
 - tablestore.columns.mapping:用于描述MaxCompute将访问的Tablestore表的列,包括 主键和属性列。

? 说明

- 以冒号(:)开头的参数值为Tablestore主键,例如示例中的 :md5 和 :ui
 d,其它参数值均为属性列。
- 在指定映射时,您必须提供指定Tablestore表的所有主键,只需提供需要通过 MaxCompute访问的属性列。提供的属性列必须是Tablestore表的列,否则即 使外表可以创建成功,查询时也会报错。
- tablestore.table.name: 需要访问的Tablestore表名。如果指定的Tablestore表名错误 (不存在),则会报错,MaxCompute不会主动创建Tablestore表。
- LOCATION:用来指定Tablestore的访问地址。请您根据环境准备,将自己的表格存储实例 访问地址参数填写在此。

⑦ 说明 如果您使用公网地址LOCATION 'tablestore://workshop-bj-001.cnbeijing.ots.aliyuncs.com/'报错,显示网络不同,可尝试更换为经典网地址LOCATION 'tablestore://workshop-bj-001.cn-beijing.ots-internal.aliyuncs.com/'。

d. 单击提交到生产环境,完成表的创建。

⑦ 说明 如果您使用的是标准模式工作空间,请先单击提交到开发环境,然后单击提 交到生产环境。

ii. 创建ods_user_trace_log表。

建表方法同上,建表语句如下。ods_user_trace_log为ODS层表。

```
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/评论',
   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
);
```

iii. 创建dw_user_trace_log表。

建表方法同上,建表语句如下。dw_user_trace_log为DW层表。

```
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/评论',

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

);
```

iv. 创建rpt_user_trace_log表。

建表方法同上,建表语句如下。rpt_user_trace_log为ADS层表。

```
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/评论',
   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
);
```

4. 验证建表结果。

- i. 完成建表后,您可以在Workshop业务流程MaxCompute > 表下看到新建的4张表。
- ii. 右键单击业务流程中MaxCompute下的数据开发,选择新建 > ODPS SQL。
- iii. 在新建节点页面,输入节点名称,单击提交新建ODPS SQL节点。

iv. 在新建的ODPS SQL节点中输入如下SQL语句,单击 回图标。

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

返回表的结构信息如下:

+ Owner: TableComment:		Pr	oject: 	
<pre> CreateTime: 2 LastDDLTime: LastModifiedTime: </pre>		020-06-16 1 020-06-16 1 020-06-16 1	8:56:46 8:56:46 8:56:46	
• InternalTable: YE	is įs	ize: 0	1	
• Native Columns:			<u> </u>	
Field	Туре	Label	Comment	
<pre>country country province city device_brand device system_type customize_event use_time customize_event_c pv uv +</pre>	string string string string string string bigint content s bigint	 	国家 省份 城市 设备品牌 终端型号 系统类型, Android、IOS、ipad、Windows_phone 自定义事件:登录/退出/购买/注册/点击/后台/切换用户/浏览 APP単次使用时长,当事件为退出、后台、切换用户时有该项 川用户关注内容信息,在customize_event为浏览和评论时包含该列 浏览量 独立访客	
Partition Columns				
dt	string		I	
OK 2020-06-16 19:56:10 2020-06-16 19:56:10 2020-06-16 19:56:10) INFO ====) INFO Exit) INFO	code of th Invocation	e Shell command 0 of Shell command completed	

3.4.2. 设计工作流

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

操作步骤

- 1. 双击您的业务流程, 打开画布面板。
- 2. 向画布中拖入1个虚拟节点,命名为start。

Dat aWorks

→ 节点组	C
~ 数据集成	
□ 离线同步	
 MaxCompute 	
Sq ODPS SQL	
SQL组件节点	
SP ODPS Spark	
Sc ODPS Script	
Mr ODPS MR	
◇ 数据服务	
▶ 数据服务	
~ 通用	
Ch oss对象检查	
5 for-each	
N do-while	
♀ 归并节点 ↓ ☆支节点	
Sh Shell	
☑ 盧拟节点	
💦 跨租户节点	

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。



② **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,详情请参见数仓分层。

3.4.3. 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语句。

前提条件

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,详细请参见注册自定义 函数。

注册自定义函数

- 1. 添加资源
 - i. 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。

关于IP地址转换的自定义函数,详情请参见MaxCompute中实现IP地址归属地转换。

ii. 右键单击WorkShop业务流程下的MaxCompute,选择新建>资源。需要分别新建File和JAR类型的资源。



- File类型上传地址库ip.dat。
 - a. 输入资源名称,选中大文件(内容超过500KB)及上传为ODPS资源,然后单击点击上 传。

新建资源			×
* 资源名称:			
目标文件夹:	业务流程/test/MaxCompute/testworkshop777/资源		
资源类型:	File		
上传文件:			
		确定 取	肖

b. 单击**提交**。

ſ	لم			
_				
	上传资	獂		
			已保存文件:	ip.dat
			资源唯一标识:	OSS-KEY-yruhgfj9qtmk81ax2fhoc4ua
				☑ 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。
 - a. 您需要勾选上传为ODPS资源,然后单击点击上传。

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/test/MaxCompute/testworkshop777/资源		
资源类型:	JAR		
上传文件:	✓ 上传为ODPS资源本次上传,资源会同步上传至ODPS中 点击上传		
		确定	取消

b. 上传完成后, 单击提交。

? 说明 提交时,请忽略血缘不一致信息。

- 2. 注册函数
 - i. 在业务流程下右键单击MaxCompute,选择新建 > 函数,将函数命名为getregion。
 - ii. 在注册函数页面,依次填写类名为odps.test.GetAddr,资源列表为getaddr.jar,ip.dat,命令格式为getregion(ip string),保存后单击 提交函数注册。

	х ,
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:	

配置节点

- 1. 配置虚拟节点start。
 - i. 双击start节点,进入节点配置页面。

ii. 单击右侧的调度配置,在调度依赖区域下单击使用工作空间根节点完成配置。

X 调度配置							
依赖上一周期:							
调度依赖 ⑦							
自动解析: 🧿 是 🔵	否 解析输入输出						
依赖的上游节点: 请输入父节点	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
my_project_simple_root		my_project_sim	ple_root		dtplus_docs	手动添加	
本节点的输出: 请输入节点转	創出名称						
輸出名称	輸出表	名 下游节点	名称	下游节点ID	责任人	来源	操作
my_project_simple.500642587_ou	ut - 🖉	ods_user	_trace_log		tina	系统默认添加	
my_project_simple.start	- 0					手动添加	

- iii. 在时间属性区域选择重跑属性为运行成功或失败后皆可重跑。
- iv. 单击 **了** 按钮,完成节点提交。
- 2. 配置ODPS SQL节点ods_user_trace_log
 - i. 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};
```

```
⑦ 说明 关于${bdp.system.bizdate}释义请参见配置调度参数。
```

ii. 完成代码编写后,单击右侧的调度配置,选择自动解析为否。

×	调度配置										1
	暫停调度:										度配量
	调度周期:	H									
	定时调度:										鳥
	員体时间:	00:00									关系
	cron表达式:	00 00 00 **?									版本
	依赖上—周期:										
											结构
	调度依赖①										
	目初解析:										
	依赖的上游节点:	清榆入父节点输出名称或输出	出表名		× + (9	明工作空间相	時点				
	自动推荐										
	父节点输出名称		父节点输出表	洺	节点名	父节点ID		责任人	来源	操作	
	my_project_simple.start				start			tina	手动添加		
	大共占约给山.	(344) # 544000755									
		店细人口///细口与44									
	輸出名称			輸出表名	下游节点名称		下游节点ID	责任人	来源	操作	
	my_project_simple.5006	43288_out		- @					系统默认添加		
	my_project_simple.ods_	user_trace_log 🕜		- Ø	dw_user_trace_log			tina	手动添加		
Γ_	节点上下文 🕜 ——										
	大节占给入会教 沃加										
	THE REPORT OF THE PARTY OF THE										

iii. 手动删除错误的依赖关系。

调度依赖 🕜 ———							
自动解析:	○ 是 💽 否 解析输入输出						
依赖的上游节点:	请输入父节点输出名称或输出表名		+ 使用工作会	间根节点			
自动推荐							
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
my_project_simple.50064	13286_out		start		tina	手动添加	删除

iv. 按照业务流程顺序搜索正确的上游节点,例如此处为start,并单击添加。

调度依赖 ⑦ —						
休毅的上游节点: • · · · · · · · · · · · · · · · · · ·	編出表名	2) ✓ + (t)	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
my_project_simple.start		start		tina	手动添加	
本节点的输出: 请输入节点输出名称		+				

v. 在时间属性区域选择重跑属性为运行成功或失败后皆可重跑。

vi. 完成后, 单击提交。

ip.dat	🔄 ods_user_trace_log 🗙 📃 表 🛛 🗤 start 🛛 🚑 Workshop
	🖳 🗔 🗟 🕲 E
	odbe~eaf
	** 提交 ********************************
	autror:dt
	create time:2019-06-17 10:04:41
	<pre>insert overwrite table ods_user_trace_log partition (dt=\${bdp.system.bizdate})</pre>
	select
	md5,
	uid ,
	ts,
	ip,
	status,
	bytes,
	device,
	system,
	customize_event,
	use_time,
	customize_event_content
	from ots_user_trace_log
	<pre>where to_char(FROM_UNIXTIME(ts),'yyyymmdd')=\${bdp.system.bizdate};</pre>

3. 配置ODPS SQL节点dw_user_trace_log

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ODPS SQL节点rpt_user_trace_log

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;
```

5. 验证配置结果。

双击业务流Workshop,打开画布面板。单击 🕟 按钮。运行成功如下图所示。



```
如果运行状态异常,请右键单击出错节点,单击查看运行日志进行排查。
```

Vi start	0
Ļ	
Sq ods_user_trace	_log 🥑
Sq dw_user_trace_	打开节点
	查看节点血缘关系
Sq rpt_user_trace_	运行节点
	运行节点及下游
	运行到该节点
	查看日志

3.4.4. 任务提交与测试

您完成节点配置后,需要将任务提交到运维中心进行测试。

操作步骤

- (可选)提交业务流程。如果您的节点在配置完成后已经提交完毕且无更新,请跳过本步骤。
 i. 双击业务流程名称Workshop,单击回图标。
 - ii. 勾选所有可提交节点及**忽略输入输出不一致的告警**, 单击提交。

提交			×
请选择	节点	节点名称	
		start	
		ods_user_trace_log	
		rpt_user_trace_log	
f	备注		
	✓ 忽略	输入输出不一致的告答	
		tetta a	小出
		tex 4	
? 说明	标准空间模式	式下,提交通过后,需要单击 发布 将任务发布至生产环境。	

2. 单击右上角的运维中心。

						❷ 跨项目克隆	@运维中心
log	Sq dw_user_trace_log 🌑	🗐 ip.dat	│ 🔜 表	Vi start	× 🛃 Works	hop	
×	调度配置						

- 3. 在左侧导航栏,单击周期任务运维 > 周期任务,双击节点列表中的虚拟节点start。
- 4. 在右侧流程图上,右键单击虚拟节点start,选择**补数据 > 当前节点及下游节点**。

展开父节点 展开子节点 ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_user_trace_log ods_scu 编辑节点 查看实例 查看血缘 测试 补数据 暂停 (冻结) 新辑节点 暂停 (冻结) 新辑节点	start				
展开子节点 → 市点详情 立者代码 金者代码 编辑节点 查看实例 查看血缘 测试 补数据 新管 (冻结) 新田节点又下游节点 海晶节点 本数据 新田节点 当前节点及下游节点 海晶节点 海晶节点 本数据 新田节点 当前节点及下游节点 海晶节点 海晶节点 二 四PS_SQL	点件型	展开父节点	>		
节点洋情 立者代码 应DPS_SQL 章者代码 编辑节点 查看实例 查看面缘 测试 补数据 补数据 暂停(冻结) 皆前节点及下游节点 海晶节点 海晶节点 近前节点及下游节点 海晶节点 海晶节点 海晶节点 黄音点段 黄音点 黄音点段 黄音点 黄音 黄音 黄音 黄音 黄音 黄音 黄音 黄音		展开子节点	>		
ods_user_trace_log oDPS_SQL	+	节点详情		+	+
编辑节点 查看实例 查看血缘 测试 补数据	ods_user_tra ODPS_SQL	查看代码		(_user_trace_log ODPS_SQL	rpt_user_trace_log ODPS_SQL
查看实例 查看血缘 迎试 当前节点 Y数据) 当前节点及下游节点 描停(冻结) 海雷节点建式		编辑节点			
查看血缘 测试 补数据 当前节点 暂停(冻结) 当前节点及下游节点		查看实例			
 測试 补数据 当前节点 当前节点及下游节点 海鼻节点模式 		查看血缘			
补数据 当前节点 暂停(冻结) 当前节点及下游节点		测试			
暂停(冻结) 当前节点及下游节点 海星节点模式		补数据	>	当前节点	
海星节点模式		暂停 (冻结)		当前节点及下游节点	
物語(解析)		恢复 (解冻)		海量节点模式	

5. 在补数据页面,选中所有需要补数据的节点,选择业务日期为过去一周,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	1-1-1 1-1-1	
* 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
bigdata_DOC(1	485)			
🔽 start				虚节点
ods_user_trace	e_log			ODPS_SQL
dw_user_trace	_log			ODPS_SQL
rpt_user_trace	_log			ODPS_SQL
				确定取消

⑦ 说明 关于补数据实例的详情请参见执行补数据并管理补数据实例。

6. 在左侧导航栏,单击**补数据实例**,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	た 700003169435 Q 补数据条称: 南西縣补数編名称 ◇ 节点開墾 南西縣市由純型 ◇ 銀行日時 2020-03-17 □ 開始)市成										
	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	结束时间	REGION	操作	
	P_start_20200317_135530	◎ 运行中								批量终止	
-	2020-03-16 00:00:00	@运行中				2020-03-16 00:00:					
	start	◎ 运行成功	虚节点	tina	2020-03-17 00:11:00	2020-03-16 00:00:	2020-03-17 13:58:07	2020-03-17 13:58:07		DAG题 终止运行 重第 更多 🔻	

生产环境,请谨慎操作										
	展开父节点									
\odot	展开子节点 >									
	查看运行日志									
	查看代码 编辑节点									
	查看血缘									
	重跑									
	重砲ト游									
	暂停 (冻结)									

如果运行状态异常,右键单击出错节点,选择查看运行日志进行排查。

- 7. 补数据实例运行完成后,验证结果。
 - i. 在左侧导航栏,单击业务流程Workshop > MaxCompute,右键单击数据开发,选择新建 > ODPS SQL,新建名为query的SQL节点。

ii. 输入如下SQL语句,查询2019年6月11日到2019年6月17日之间表rpt_user_trace_log中的数据,确 认数据是否成功写入rpt_user_trace_log表。

select * from rpt_user_trace_log where dt BETWEEN '20190611' and '20190617' limit 1
000;

ⅲ. 单击⊙图标。

查询结果如下。

Sq que	ny x	Sq od	s_user_tra	ace_log	Sq rp	t_user_trace_log	Sq dw_user_tra	ice_lo	g 🌒 F ip.dat	📃 表		Vi start		🛛 📇 Worksho	p		
	(<u>ት</u> [{		\$	\odot												
	<pre>1odps sql 2 3 4create time:2019-06-19 16:03:48 5 5 6 select * from rpt_user_trace_log where dt BETWEEN '20190611' and '20190617' limit 1000;</pre>																
																	5.7 2 3
运	行日志	ŧ	結果[1]	×													\$\$ E
	A																
1	country	~	province		✓ city	~	device_brand	✓ dev	vice 🗸	system_type	>	customize_event	✓ use_t	ime 🗸	customize_event	_c 🗸 p\	
	中国		山东		菏泽		meizu	ME	EIZU PRO7	android		switch	7		news	23	3
	挪威		挪威				iphone	iPh	hone6	ios		switch	6		travel	30	
	韩国		韩国				ipad	iPa	ad4	ios		switch	5		travel	30	
	中国		山东		菏泽		iphone	iPh	hone7 Plus	ios		switch	5		travel	17	7
6	挪威		挪威				xiaomi	XI/	AOMI Note3	android		switch	4		travel	31	1
7	韩国		韩国				iphone	iPh	hone6	ios		switch	5		travel	31	1
8	中国		山东		菏泽		iphone	iPh	hone7	ios		switch	4		travel	37	7
	挪威		挪威				huawei	HU	JAWEI Mate 10	android		switch	8		travel	23	3

3.5. 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 Bl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。进入Quick Bl控制台。

背景信息

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

操作步骤

1. 单击进入默认空间,您也可以使用自己的个人空间。
| 😍 Quick Bl 😵 专业版 🔵 | | 我的 | 工作空间 | 创作区 | 订阅 | 监控指标 |
|---|------------------------------------|-----------------------------|--------|-------------------------|---|----------------|
| Quick Bi していたちに 工作空间 加入更多 + ・ | 最近相关 与我共享 | 我的收藏 | 暂时没有数据 | 建 度,快快开始 | 山 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 一 | |
| ■ 默认空间 ◎ 創建: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 | 仪表
丰富的
自助教 | 板
的可视化组件支
数据分析 | i j |
| []
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 | | | 表格
cel式的体验制作
业报表 | 復 |

- 2. 新建MaxCompute数据源。
 - i. 单击左侧导航栏上的数据源,进入数据源页面。
 - ii. 单击右上角的新建数据源。选择云数据库 > MaxCompute。
 - iii. 在**添加MaxCompute数据源**页面,配置数据源连接参数。

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。

	✓ 数据源连通性正常!						
添加MaxCompute数据源							
* 显示名称:	bigdata_DOC						
* 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:新增数据源存在同步延迟的情况,请稍候片刻。							
	关闭 连接测试	添加					

- 显示名称:数据源配置列表的显示名称。
- 数据库地址:此处有默认地址,通常无需修改。

⑦ 说明 数据库地址根据Region不同而变化,详细对应信息请参见Endpoint。

- 项目名称: MaxCompute项目名称。
- AccessKey ID: 您账号的AccessKey ID。
- AccessKey Secret: 您账号的AccessKey Secret。

iv. 单击连接测试,进行数据源连通性测试。

⑦ 说明 如果连通正常,系统会给出连通成功提示。

v. 单击**添加**,完成数据源添加。

成功添加完成后,页面自动跳转到**数据源**管理页面,并在页面右侧展示出数据源所包含的所有数据 表。

3. 在数据源管理页面找到rpt_user_trace_log表,单击创建数据集。

教程·搭建互联网在线运营分析平台

Dat aWorks

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
and state to an other state.		名称 \$ 5.6_50550559	备注◆	操作
C and the second second		output_table_154812606367001adfb67dfee		ŵ ()
and the second second		output_table_1548127871928c581534cd65e		i
20,000 (000)		output_table_154822526534301a9a11d2c44		ŵ ()
an and an		output_table_154838757651615e9882605fe		â ()
F 1		output_table_1548743722536a9cd797ed51f		i
		result_table		⑦ ① 创建数据集
a the second second		rpt_user_trace_log		1
A REAL PROPERTY OF		sale_detail		· ()
bigdata_DOC 所有者:		system_9c676e75c4324f75b5430dd138a738ab		a ()

输入数据集的**名称和位置**,单击确定。

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

4. 单击左侧导航栏上的**数据集**,进入**数据集**页面。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。详情请参见概述。

5. 转换字段的维度类型。

i. 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。



ii. 转换地理信息字段的维度类型。

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

维度 * 446 at	+ 🏼 count	ry	dt(day)	pr
🛗 dt(ye	ar)				
🛗 dt(au	(arter)				
☐(1-	onth)				
dt(we	eek)				
dt(da	y)				
Str. provine	。 伯母	1			
str. city	ØHH44				
str. device	⑦ 克隆维度				
str. device	✔ 左分析山隐藏				
^{Str.} system					
Str. custor	◎ 取消全部隐藏				
Str. custom	★ 删除				
度量					
- 🗁 默认	+ 新建计具子段(维度)				
[№] use_i	。移动到				
Nº pv	。 新建尼次结构			国家	
Nº uv	出 新建运入印码			区域	
	↑ 上移				
	」 下移			省 / 直辖市	
	¥ ' '	日期(源数据格	式) ⊾	市	
	↓ 转换为度量			_ / _	
	,始度举刑扣场	地理信息	•	区/县	
	₩ 使大至 则 探	✓ 文本		经度	
		-			-

b. 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。



- iii. 新建层次结构。
 - a. 右键province, 单击新建层次结构, 在弹框中单击确定。



b. 将city字段移到province层次结构的树下。

维度	+ 🎚
- 🗁 province_层级结构	
province	ŝ
🦁 city	

c. 完成上述操作后,单击保存,返回数据集列表。

<	🗊 rpt									保存
	王 -	۵ ا								件 同步素结构
体度		e province	eitu	a.	dt(daw)	an device brand	an douico	au sustam tuna	austomizo event	sustemite event
- 🖻	province_层级结构	province	city	country	ut(uay)	device_brand	Gevice	system_type	customize_event	custornize_event
ser.										

6. 制作仪表板。

即随着数据的更新,让报表可视化地展现最新数据。

i. 单击rpt数据集后的新建仪表板图标,选择常规模式,进入仪表板编辑页。

数据集 全部 我的		名和	▶ ∨ Q 共17个文件	+ 新建数据集
惯目录 ≻ ODPS				
名称 🗣	创建者 🔷	修改人/修改时间	数据源	新建仪表板
rpt.user_trace_log	Furui		bigdata_DOC MaxCompute	⊠ ,,] ≉ :
Terrar Statement (1997)	冯国童	1000	ECS_QBI_DB SQL Server	⊠ ,ıl ⊜ :
7	冯国章		workshop MySQL	⊠ ,ıl ≋ :

ii. 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	未命名		*		指标着	板	
ا <mark>ث</mark> +	T	2	R 🐳	di 26	(A 📮	Ô	¢,
		nt 指标	看板: 缺少维	<u></u> 度或度量	项			
		指标	看板: 缺少维	 度或度量	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_trace_log为分区 表,因此必须在过滤器处选择筛选的日期,本例中筛选为2019~2019年,完成设置后单击更 新。



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

ন্ T 🖉	-	*	ψ,	**	Ť	-	C
pv Fu ^{ruli}						:	
10.69万 ^{Furdi}			Furui				
uv Fu ^{rul}							
3800 Purul							

iii. 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)



iv. 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。



v. 完成配置后,单击**保存**及**预览**,即可看到展示效果。

< al RPT 🔺					ê ◎ Da ≪ 🖸	1993 (1 993)	fit fit
pv Ford		線時間 でが		-uv			
10.69万		1.05 1.05	rad	FUN	1000		
uv Forti		要 8000 900 ³ 4000			Fund		
3800 M		e.e.d 0 20190	611 C 20190612	20190613 Col ^{CJ} 20190614	20190615 / ⁽⁾ 20190616	20190617	
色彩地图 199		ed P V				eard	
						euro -	
		Less-					
3.76675 - 3.99275		mart	TA -				
3.34/7 - 3.766/7 3.314/7 - 3.54/7 3.088/7 - 3.314/7 2.862/7 - 3.088/7							

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、用电检查和计量工作人员,利用计量异常报警和电能量数据查询功能来在线 监控用电情况。通过采集电量异常、负荷异常、线损异常、终端报警、主站报警信息,建立数据分析模型, 工作人员可以实时监测窃漏电情况并发现计量装置故障。根据报警事件发生前后,客户计量点有关的电流、 电压和负荷等数据情况,构建基于指标的用电异常分析模型,检查是否存在窃电、违章用电及计量装置故障 等情况。

虽然上述防窃电漏电的查询方法可以获得用电异常信息,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。同时,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、功率等用电负荷数据及用电异常等终端报警信息。此外,稽查工作人员还可以通过在线稽查系统和现场稽查,查找窃电漏电用户数据并录入系统。

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,即可自动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。

4.2. 准备环境

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

- 注册阿里云账号。
- 进行个人实名认证或企业实名认证。

背景信息

本次实验涉及的阿里云产品如下:

- MaxCompute
- Dat aWorks
- PAI

开通大数据计算服务MaxCompute

⑦ 说明 如果您已经开通MaxCompute,请跳过该步骤,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,输入您的阿里云账号和密码。
- 2. 鼠标悬停至顶部菜单栏中的**产品**,单击**大数据 > 大数据计算与分析 > MaxCompute**,进入 MaxCompute产品详情页。
- 3. 单击立即开通。

- 4. 在购买页面,选择地域,并选中服务协议,单击确认订单并支付。
 - ? 说明
 - 购买页面默认提供的规格类型为MaxCompute按量计费标准版和DataWorks基础版。
 - MaxCompute的项目管理和查询编辑集成DataWorks的功能,因此需要同时开通DataWorks 服务。DataWorks基础版为0元开通,如果您不使用数据集成、不执行调度任务,则不会产 生费用。
 - 选择地域时, 您需要考虑的最主要因素是MaxCompute与其它阿里云产品之间的关系。例 如, ECS所在地域、数据所在地域等。

创建工作空间

- 1. 使用主账号登录DataWorks控制台。
- 2. 在概览页面,单击右侧的创建工作空间。

您也可以单击左侧导航栏中的工作空间列表,切换至相应的区域后,单击创建工作空间。

3. 在创建工作空间对话框, 配置各项参数, 单击下一步。

分类	参数	描述
	工作空间名称	工作空间名称的长度需要在3~23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23个字符,只能字母、中文开头, 仅包含中文、字母、下划线(_)和数字。
		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 :
基本信息		 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。
	模式	 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两种环境,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。
		详情请参见简单模式和标准模式的区别。
	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。此参数在 工作空间创建完成后可以在工作空间配置页面进行 修改,详情可参考文档:安全设置。

4. 在选择引擎界面,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。 ⑦ 说明 此处需要同时勾选机器学习PAI和MaxCompute。

5. 进入引擎详情页面, 配置选购引擎的参数。

分类	参数	描述
	实例显示名称	实例显示名称需要以字母开头,只能包含字母、数 字和下划线(_)。
	Quota组切换	Quota用于实现计算资源和磁盘配额。
	MaxCompute数据类型	该选项设置后将在5分钟内生效,数据类型模式的 详情请参见 <mark>数据类型版本说明</mark> 。如果您不清楚模式 的选择,建议与工作空间管理员确认后再进行选 择。
MaxCompute	是否加密	您可以设置不加密和加密。
	MaxCompute项目名称	开发环境的默认名称为DataWorks工作空间的名称 dev,生产环境的默认名称与DataWorks工作空 间名称一致。
	MaxCompute访问身份	开发环境的MaxCompute访问身份默认为 任务负 责人,不可以修改。 生产环境的MaxCompute访问身份包括 阿里云主 账号和 阿里云子账 号。

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. 准备数据

在数据准备阶段,您需要同步原始数据至MaxCompute。

准备数据源

- 1. 通过RDS创建MySQL实例,获取RDS实例ID。详情请参见创建RDS MySQL实例。
- 2. 在RDS控制台添加白名单,详情请参见添加白名单。

⑦ 说明 如果是通过自定义资源组调度RDS的数据同步任务,必须把自定义资源组的机器IP也加入RDS的白名单中。

- 3. 下载本教程使用的原始数据indicators_data、steal_flag_data和trend_data。
- 4. 上传原始数据至RDS数据源,详情请参见将Excel的数据导入数据库。

新增数据源

⑦ 说明 本次实验需要创建MySQL数据源。

1. 进入数据源管理页面。

> 文档版本: 20220104

i. 登录DataWorks控制台。

ii. 在左侧导航栏, 单击工作空间列表。

iii. 选择工作空间所在地域后, 单击相应工作空间后的进入数据集成。

iv. 在左侧导航栏,单击数据源,进入工作空间管理>数据源管理页面。

2. 在数据源管理页面,单击右上角的新增数据源。

3. 在新增数据源对话框中,选择数据源类型为MySQL。

4. 在新增MySQL数据源对话框中, 配置各项参数。

参数	描述		
数据源类型	当前选择的数据源类型为 MySQL> 阿里云实例模式 。		
数据源名称	数据源名称必须以字母、数字、下划线组合,且不能以数字和下划线开 头。		
数据源描述	对数据源进行简单描述,不得超过80个字符。		
	可以选择 开发 或 生产 环境。		
适用环境	⑦ 说明 仅标准模式工作空间会显示该配置。		
地区	选择相应的区域。		
RDS实例ID	您可以进入RDS控制台,查看RDS实例ID。		
RDS实例主账号ID	实例购买者登录控制台,进入 安全设置 页面,即可查看实例账号ID。		
数据库名			
	数据库的名称。		
用户名	数据库的名称。 数据库对应的用户名。		

5. 单击测试连通性。

6.

7. 测试连通性通过后,单击完成。

新建业务流程

- 1. 单击当前页面左上角的 图标,选择全部产品 > 数据开发 > 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, 输入业务名称和描述。

数据开发

⑦ 说明 业务名称的长度不能超过128个字符,且必须是大小写字母、中文、数字、下划线(_)以及小数点(.)。

- 4. 单击**新建**。
- 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)和三个离线同步节点(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**。

∨ 数据集成			
回 离线同步			
✓ MaxCompute			
Sq ODPS SQL			
௺ SQL组件节点	新建节点		×
Sp ODPS Spark			
Py PyODPS	节点类型:	虚拟节点	
Sc ODPS Script	节点名称:	start	
Mr ODPS MR	目标文件夹:	业务流程/	
∨ 通用			
Ch oss对象检查			設置
🛐 for-each			
N do-while			
❣ 归并节点			
📩 分支节点			
▲ 赋值节点			
Sh Shell			
☑ 虚拟节点			
▶ 跨租户节点			

6. 拖拽连线将start节点设置为三个离线同步节点的上游节点。

চি 💿 🔍 🖈			
◇ 节点组 C			
✓ 数据集成			
□] 离线同步			
✓ MaxCompute		Vi start	
Sq ODPS SQL			
௺ SQL组件节点			
Sp ODPS Spark			
Py PyODPS			
Sc ODPS Script	Di 电量下降趋势数据同步	Di 窃电标志数据同步	Di 指标数据同步
Mr ODPS MR			
◇ 通用			
ch oss对象检查			
テ for-each			
N do-while			
❣ 归并节点			

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。

× 调度	配置									调度
C	ron表达式:	00 01 00 * * ?								配置
依赖	负上—周期:									
										本
调度	を依赖 ⑦ -									
	自动解析:	● 是 ●	否 解析	输入输出						
依赖	納上游节点:	请输入父	节点输出名称。	成输出表名		+ 使用工作的	空间根节点			
父节	节点输出名称	父节点	輸出表名	节点名	5	2节点ID	责任人	来源	操作	
	_root			_root				手动添加		
本节	5点的输出:	请输入节点	渝出名称							
输出	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,单击左上角的图图标。

新建表

- 1. 打开新建的业务流程,单击MaxCompute左侧的展开图标,打开MaxCompute。
- 2. 右键单击MaxCompute下的表,单击新建表。
- 3. 在新建表对话框中, 输入表名, 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)。

⑦ 说明 表名不能超过64个字符,且必须以字母开头,不能包含中文或特殊字符。

4. 打开创建的表,单击DDL模式,分别输入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;
```

- 5. 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。
- 6. 返回建表页面后,在基本属性中输入表的中文名。
- 7. 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 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		
基本属性				
	中文名电量下降趋势表	Ē		
-	-级主题 请选择		二级主题 请选择	→ 新建主题 C
	描述			
物理模型设计				
£	这类型 💿 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
	表类型 💿 内部表 🤇			

配置离线同步节点

- 1. 配置电量下降趋势数据同步节点。
 - i. 双击电量下降趋势数据同步节点,进入节点配置页面。

ii. 选择数据来源。



参数	描述
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。
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。

iii. 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	持的数据来源	赎 型			
* 数据源	ODPS		odps_first		0	
*表	trend_data					
				键生成目标表		
* 分区信息	dt = \${bizdate}		?			
清理规则	写入前清理已有	与数据 (Insert	Overwrite)			
空字符串作为nuli(● 是 🧿 व	Б Т				
参数		描述				
数据源		选择ODPS	> odps_fir	st 。		

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。					
空字符串作为null	选择否。					

iv. 配置字段映射。

02 字段映射		源头表			目标表	麦			收起
	源头表字段	类型	Ø		B	标表字段	类型	同名映射	
	uid	INT	•)	• uic	d	BIGINT	同行映射取消映射	
	indicator	INT	•	•	• tre	end	BIGINT	自动排版	
	添加——行 +								

v. 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 💙 🕐
	・同步速率 💿 不暇流 🔘 限流 🕐
	错误记录数超过 胜数据会数范围,默认允许驻数据 条,任务自动结束 ②
参数	描述
任务期望最大并发数	数据同步任务内,可以从源并行读取或并行写入数据存储端的最大线程数。向导 模式通过界面化配置并发数,指定任务所使用的并行度。
同步速率	设置同步速率可以保护读取端数据库,以避免抽取速度过大,给源库造成太大的 压力。同步速率建议限流,结合源库的配置,请合理配置抽取速率。
错误记录数	错误记录数,表示脏数据的最大容忍条数。

vi. 确认当前节点的配置无误后,单击左上角的凹图标。

提交业务流程

1. 打开业务流程配置面板,单击左上角的回进行提交。

2. 选择提交对话框中需要提交的节点, 输入备注, 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	窃电用户证	只别		
	✔ 忽略输	i入输出不一致的告警		
			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. 在数据开发页面的左侧导航栏,单击临时查询,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。

	Ē	\$	\odot			С	8		
		odps	sql						
					*****		****		******
		auth	ior:						
		crea	ite t	ime:2	019-08	3-09	21:36	5:14	
					*****		****		******
		SELECT	cou	nt(*)	from	tren	d_dat	t <mark>a where dt</mark>	=20190808;
		SELECT	cou	nt(*)	from	indi	cator	rs_data whe	re ds=20190808;
		SELECT	cou	nt(*)	from	stea	1_f1a	ag_data whe	re ds=20190808;
16	_								
	结	課[1]	×	4	吉果[2]			结果[3]	
		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,您已经学习了如何通过数据同步采集数据,您可以继续下一个教程。在该教程中,您将学习如何对采 集的数据进行计算与分析。

4.4. 加工数据

本文为您介绍如何通过DataWorks加工采集至MaxCompute的数据,并获取清洗后的数据。

前提条件

开始本文的操作前,请首先完成准备数据中的操作。

新建表

- 1. 进入数据开发页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击工作空间列表。
 - iii. 选择工作空间所在地域后,单击相应工作空间后的进入数据开发。
- 2. 在**数据开发**页面,单击相应业务流程左侧的**、**图标,展开该业务流程。

- 3. 右键单击MaxCompute,选择新建>表。
- 4. 在新建表对话框中, 输入表名, 单击提交。

↓ 注意 表名必须以字母开头,不能包含中文或特殊字符,且不能超过64个字符。

此处需要创建的数据表,如下所示:

- 创建三张表,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)。
- 创建表data4ml,存储汇聚后的数据。
- 5. 打开创建的表,单击DDL模式,分别输入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

--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;
```

6. 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。

- 7. 返回建表页面后,在基本属性中输入表的中文名。
- 8. 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 从开发环境加速	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			
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 グ 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 イン 新建居	级 С
表类型				

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,并向面板中拖入两个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。

	Vi start		
Di 电量下降趋势数据同步	Di 窃电标志数据同步	Di 指标数据同步	
	Sq 数据清洗		
	Sq 数据汇聚		

配置ODPS SQL节点

- 配置数据清洗节点。
 - i. 双击数据清洗节点,进入节点配置页面。
 - ii. 编写处理逻辑。

	🗳 🖸									
1	odps	sql								
2										
З	autho									
4	create time:2019-08-09 21:42:22									
5										
6	INSERT	OVERWRITE TABLE clean_trend_data PARTITION(dt=\${bdp.system.bizdate})								
7	SELECT	uid								
8		, trend								
9	FROM	trend_data								
10	WHERE	trend IS NOT NULL								
11	AND	uid != 0								
12	AND	dt = >{bap.system.bizdate}								
17	j									
14	TNCEDT	O(EPURTTE TARLE clear ctool flog data PARTITION(dc-C(hdp custom birdata))								
16	SELECT	uid								
17		.flag								
18	FROM	steal flag data								
19	WHERE	uid != 0								
20	AND	ds = \${bdp.system.bizdate}								
21	;									
22										
23	INSERT	OVERWRITE TABLE clean_indicators_data PARTITION(ds=\${bdp.system.bizdate})								
24	SELECT	uid								
25		,xiansun,warnindicator								
26	FROM	indicators_data								
27	WHERE	uid != 0								
28	AND	ds = \${bdp.system.bizdate}								
29	;									

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}
;
```

ⅲ. 单击工具栏中的Ⅲ图标。

• 配置数据汇聚节点。

- i. 双击数据汇聚节点,进入节点配置页面。
- ii. 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=${bdp.syste
m.bizdate}
)b
ON
      a.uid = b.uid
FULL OUTER JOIN
(
   SELECT uid, flag FROM clean steal flag data where ds=${bdp.system.bizdate}
) C
       b.uid = c.uid
ON
;
```

ⅲ. 单击工具栏中的Ⅲ图标。

提交业务流程

- 1. 打开业务流程配置面板,单击工具栏中的面图标。
- 2. 选择提交对话框中需要提交的节点, 输入备注, 并选中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	入输出不一致的告警	
			提交取消

3. 单击提交,待显示提交成功即可。

运行业务流程

- 1. 打开业务流程配置面板,单击工具栏中的 图标。
- 2. 在左侧导航栏,单击临时查询。
- 3. 在临时查询页面,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。

	£ () (D 🗈		С	83					
	od	ps s	ql								
		****	*****	*****	*****	******	****	*****	*****	*****	
	au	thor									
	cr	eate	time:	2019-08	3-09 2	21:36:14					
				*****	*****	*****	****			*************	
	SELE	CT c	ount(*) from	clea	n_trend_data	a whe	re dt=201	90808;		
	SELE	CT c	ount(*) from	clea	n_indicator:	s_dat	a where d	s=2019	0808 ;	
	SELE	CT o	ount(*) from	clea	n_steal_fla	g_dat	a where d	s=2019	0808;	
	SELE	CT c	ount(*) from	data	4ml where d	 s=201	90808;			
				r							
	结果[1]			结果[2]		结果[3]		结果[4]	×		
		Α									
1	_c0		~								
2	291										

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20190808。

```
--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。

⑦ 说明 将任务发布至生产环境前,您需要对代码进行测试,确保其正确性。

- 1. 打开业务流程配置面板,单击工具栏中的 ☑图标。
- 2. 在创建发布包页面,选中待发布的任务,单击添加到待发布。

教程·实现窃电用户自动识别教程

😚 🖪 任务发布	100.00		~					<i>₽</i> Dat	aStudio 🕜 运	雄中心 🕄	time to be a little of the lit
											0
① 创建发布包	创建发布	副社会中心 									
□□ 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	: 请输入节点ID		
	节点类型	请选择	◇ 交更类型:	请选择 ~	提交时间小	、于等于: үүүү-№	IM-DD	證	<u>ŝ</u>		
			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 渴	加到待发布
		700002621610	数据清洗		ODPS SQL			2019-08-13 10:38:21	未測试		§加到待发布
		700002621602	指标数据同步		数据同步			2019-08-13 10:05:53	未測試		ā加到待发布
		700002621601	窃电标志数据同步		数据同步			2019-08-13 10:05:49	未測試		动到待发布
		700002621600	电量下降趋势数据同步		数据同步			2019-08-13 10:05:46	未測試		动到待发布
	添加到作	春发布 打开待发 [。]	布发布选中项						・ 上一页 🚺 下一		還示:

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的内容。

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	会 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	央方案 🗸 业务流程: 业务流利	E マ 节点类型: 请选择节点	类型 → 责任人:	 基线 请选择基线 > 	
\$\$\$	实时任务运维	~	✓ 我	的节点 🗌 今日修改的节点 🗌 暫	停(冻结)节点	重置 清空				
а	周期任务运维	^							○ 刷新 收起搜索	
	周期任务			名称	节点ID		生产	环境,请谨慎操作	ତେଭ୍ର୍ର	2
	周期实例			数据汇聚	Internation of the					
	补数据实例			数据清洗	-					
	測試实例			指标数据同步	The second second	start 盧市庄				
ଡ	手动任务运维	~	-							
*	智能监控	~		创电标应器组织同步		-				
				电量下降趋势数据同步		电量下降趋势 _{数据集成}	窃电标志数据同步 _{数编集成}	指标数据同步 ^{数据集成}		
				start	in the second second					

- 节点类型: 请选择节点类型 责任人: 方案 业务流程: 业务流程 \sim \sim × 重置 湷空 生产环境,请谨慎操作 start 虚节点 展开父节点 > 展开子节点 > 节点详情 电量下降趋势 窃电标志数据同步 指标数据同步 数据集成 数据集成 数据集成 查看代码 编辑节点 查看实例 查看血缘 3) 测试 当前节点 补数据 > 当前节点及下游节点 暂停(冻结) 海量节点模式 恢复(解冻)
- 3. 右键单击DAG图中的start节点,选择**补数据 > 当前节点及下游节点**。

4. 选中需要补数据的任务,并选择业务日期。

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 耳	硝

5. 单击确定

6. 在**补数据实例**页面,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。

4.5. 数据建模

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

前提条件

开始本文的操作前,请首先完成加工数据中的操作。

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。

	华东2(上海) ▼		Q 搜索			费用 工单	医白素 企业	支持与服务 🛛	Ĵ Ĝ	₩ 0	۵ (简体中文
机器学习PAI	机器学习PAI / Studio-可核	選挙习NI / Studio可吸化建模										
概范	PAI 可视化	AI可视化建模										
Studio-可视化建模 ~	封装常用机器学习算法》	及丰富的可视化组件,用户列	需代码基础,通过拖拉	按即可训练模型。								
算法发布	创建项目 项目名称	➤ 请输入	Q									
EAS-模型在线服务 WW	项目名称	显示名	付费模式	所屬区域	项目管理员	MaxCompute资源	创建时间		开启GPU	0	操作	
AI市场		10.000	I/O后付费	华东 2 (上海)	datawor*******	10.000	2019-08-09 22:	07:56			进入机器	醫学习
	10,000	400 m	CU预付费	华东 2 (上海)	datawor*******	10000	2019-08-07 10:	18:24			进入机器	醫学习
	in the local data	-	I/O后付费	华东 2 (上海)	datawor*******	1.000	2019-08-07 10:	17:22			进入机器	後学习

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。

2	机器学习F	PAI 华东2(上海)	***.***	算法平台	前往运维
	搜索		Q		
ΗW	🗸 🔽 我的实	验			
る 実验		④ 从模板新建实验	È		
		🔊 导入实验			
Notebook		① 新建空白实验			
		➡ 新建文件夹			
数据源		■ 重命名			
合 组件		直 删除			
₩		👳 评估对比			
模型					
ڻ هت					

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	\$10.m	
描述	窃电用户识别	
位置	✓ / 我的实验	
	创建取消	≝

5. 单击创建。

载入数据集

- 1. 单击左侧导航栏中的数据源。
- 2. 在搜索框输入加工数据中最终输出的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。

2	机器学习PAI 华东2(上海)	-	▼ 算法平台	前往运维
ភ្ល	data4ml 2	<u>3</u> Q	🛆 窃电用户识别 🗙	
自见	☆ 常用表	~	● 运行 ④ 部署	& Auto ML
≦à 実验	☰ 公共表	>		
<u>6</u>	井 表搜索结果			
Notebook	更多的表操作请前往DataWorks			
日 数据源	🖯 data4ml	☆ 4		
1 组件] data4ml
❤ 様型				
(ý) 1111				

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。数据包括电量趋势下降指标、线损指标 和告警类指标数量等窃电漏电指标,以及用户是否真实窃电漏电的数据。

🖯 data4ml		
• •	□□ 重命名	
	向 删除	Delete
	■ 复制	Ctrl + C
	▶ 从此处开始执行	
	○ 刷新字段	
	 查看数据 	

数据探查	据探查 - data4ml - (仅显示前一百条)						кл Х
序号▲	uid 🔺	trend 🔺	xiansun 🔺	warnindicator 🔺	flag 🔺	ds 🔺	
1	1	4	1	1	1	20190808	_
2	2	4	0	4	1	20190808	
3	3	2	1	1	1	20190808	
4	4	9	0	0	0	20190808	
5	5	3	1	0	0	20190808	
6	6	2	0	0	0	20190808	
7	7	5	0	2	1	20190808	
8	8	3	1	3	1	20190808	
9	9	3	0	0	0	20190808	
10	10	4	1	0	0	20190808	
11	11	10	1	2	1	20190808	
12	12	10	1	3	1	20190808	
13	13	2	0	3	0	20190808	
14	14	4	0	2	0	20190808	
15	15	3	0	0	0	20190808	
16	16	0	0	3	0	20190808	
17	17	9	0	3	1	20190808	
						复制	关闭

进行数据探索

- 1. 相关性分析
 - i. 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)	1 10	-	算法	平台	前往运维
分前	搜索		Q	ā 5	阳用户ù	₋ 別×	
A	>	数据预处理		۰ì	运行	① 部署	
実验	>	特征工程					
Notebook	~ [🧧 统计分析					
		资数据视图					
		🖸 协方差					
じゅ 组件		🖸 经验概率密度图				6	data4ml
₩		🖸 全表统计					
模型		🔁 卡方拟合性检验				6	
¢		🔯 卡方独立性检验				ł	
ig i i		10 箱线图			ð 相关	系数矩阵	F-1 ()
		100 散点图					
		🖸 相关系数矩阵					
		🖸 双样本T检验					
		🖸 单样本T检验					

ii. 连线读数据表中ODPS源的输出和相关系数矩阵的输入。

iii. 右键单击相关系数矩阵,选择从此处开始执行。



iv. 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个窃电漏电指标本身和最终是否为窃电用户的关系都不太明显,即用于 判断用户是否为窃电用户的特征并不具有单一性。

2. 特征分析

i. 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。



ii. 连线读数据表中ODPS源的输出和数据视图的输入。

- ⊕ ⊖ ⊕ ⊹ ⊡ 🗉 🖸 字段设置 参数设置 执行调优 选择特征列 必选 选择字段 选择目标列 可选 使可视化更加丰富 flag 枚挙特征 可选 在Int/Double字段的枚举特征 ⑦ 选择字段 k:v,k:v稀疏数据格式 🖯 data4ml-1 \odot 🖸 相关系数矩阵-1 📀 🖸 🖸 数据视图-1
- iii. 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。

iv. 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				× دی
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	~	Ē	字段	类型
uid		Û	trend	BIGINT
✓ trend		Û	xiansun	BIGINT
✓ xiansun		Ē	warnindicator	BIGINT
varnindicator				
🔲 flag				
				确定 取消

v. 右键单击数据视图,选择从此处开始执行。
	字段	Q	間表	步长	- 0.1 + 1倍 分積数 100	
A	trend 熵: 0.3915 连续型			trend	■ 0 ■ 1 華尼 満 ○ 上	<u>/r </u>
В	xiansun 嫜: 0.5200 连续型			连续型		0.3
С	warnindicator 熵: 0.4678 连续型			30		0.2
				20		0.15
				10		0.05
					2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 - 2 -	

vi. 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。
 - i. 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。



ii. 连线读数据表中ODPS源的输出和拆分的输入。

iii. 右键单击拆分,选择从此处开始执行。

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

iv. 待运行完成后,右键单击**拆分**,选择查看数据 > 查看输出桩。

- 2. 通过逻辑回归二分类组件,对数据进行回归建模。
 - i. 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。

4	机器学习PAI 华东2(上海)	▼ 算法平台	前往运维
	搜索 Q	🖉 窃电用户识别 ×	
2	查看组件说明文档 ▼ ◆ 営用組件	④运行 ④ 部署	Q, Auto ML
988 202	✓ ☑ 保存的分组		
Notebook	 > > 源 / 目标 > 数据预处理 		
数据源	> 🖿 特征工程		
组件 (19)件	> 统计分析		data4mi-1 📀
耀	✓ ⊆ 二分类		
÷	G GBDT二分类		
	 送性支持向量机 逻辑回归二分类 		民 相关系数矩阵-1 ○ 民 数据视图-1 ○ 承 拆分-1 ○
	> 🖿 多分类		
	> 📄 聚美		
	 大式性体 评估 		
	> _ 推荐算法		

ii. 连线拆分中的输出表1和逻辑回归二分类的训练表。

 iii. 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。

®₄ Auto ML	$\oplus \oplus \oplus *$		字段设置	参数设置	执行调优
 ☐ data4mi-1 ② ☑ data4mi-1 ② ☑ data4mi-1 ② ☑ 数据视图-1 ② ☑ 数据视图-1 ② ☑ 逐 逐 图 视图-1 	⅔ 拆分-1 編回类-1 〔	0	训练特征列 必选 目标列 必选 flag 正类值 必选 eg. 1 量否解疏数时	5 支持Double/Int类 选择字段 0/1分类中1是正类 属 k:v,k:v 类型特征	型字段

iv. 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				X الاع الاع
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	~	Ē	字段	类型
uid		Ē	trend	BIGINT
✓ trend		Ē	xiansun	BIGINT
✓ xiansun		Ē	warnindicator	BIGINT
varnindicator				
flag				
				确定 取消

- v. 右键单击 逻辑回归二分类,选择从此处开始执行。
- vi. 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			кл х КИ х							
	在输入数据为稀硫的时候,不显示 weight 全是 0 的特征										
		权重									
		1 🔺	0 🔺								
	trend	1.218831335893094	-								
	xiansun	2.286408879362565	-								
	warnindicator	1.31637344464322	-								
	常量	-9.825371109484525	0								
	保存到MaxCompute: pai	_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。
 - i. 单击左侧导航栏中的组件,拖拽机器学习 > 预测至右侧画布。

2	ŧ	机器学习PAI	华东2(上海)	***.**		算法平台	前往运维								
ଜ	搜	索		Q	△ 窃电	用户识别 ×									
首页	>	📄 源/目标			 运行 	5 ① 部署	@ Auto ML					Đ	ର୍ ⊕	* 0	
会	>	📄 数据预处理	1												
9	>	📄 特征工程													
Notebook	>	📄 统计分析													
日本	\sim	📁 机器学习													
et.	>	📄 二分类													
组件	>	📄 多分类								data (m) 1					
₩	>	📄 聚类													
8442	>	■ 回归								\wedge					
© #=	>	📄 关联推荐	l												
	>	📄 评估						(
	>	📄 推荐算法									0	15-45-45-44	*		0
		🖸 预测						EQ 相大系数6	80年-1 ♥			- (c 4/6 3/	- 1	-	
	>	■ 深度学习	-	点击 <mark>这里</mark> 开								/			
	>	📄 强化学习									_	ļ			
	>	📄 时间序列										〕逻辑回…类	-1 /	\odot	
	>	📄 文本分析													
	>	📄 网络分析										¢	Ļ		
	>	🖿 I具									ſ	🖸 预测-1	,]
	>	📄 金融板块													, ,

- ii. 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。
- iii. 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。

iv. 在选择字段对话框中,全选5个字段,单击确定。

block (加) Add X 健学即可 (小) Add X 健学即可 (小) BIGINT (小) Vid (小)	选择字段			K7 X
父 鈕DD小小<	输入关键字搜索列,包含关键字即可			Q
♥ BIGINT♥学段类型♥ uididididBIGINT♥ trendItrendBIGINT♥ xiansunIIxiansunBIGINT♥ warnindicatorIIwarnindicatorBIGINT♥ flagIIflagBIGINT	✓ 全选	已选		列表编辑
V uidIIIuidBIGINTV trendIIItrendBIGINTV xiansunIIIxiansunBIGINTV warnindicatorIIIwarnindicatorBIGINTIIIIflagBIGINTBIGINT	V BIGINT V	11	字段	类型
Image: rendImage: rendBIGINTImage: rendImage: rendBIGINTImage: rendImage: rend	✓ uid	Ĥ	uid	BIGINT
Y xiansunIIxiansunBIGINTV warnindicatorIIwarnindicatorBIGINTI flagIIflagBIGINT	✓ trend		trend	BIGINT
V warnindicatorIIIBIGINTIIIflagIIIBIGINT	✓ xiansun	Ē	xiansun	BIGINT
✓ flag ⊡ flag BIGINT	varnindicator	Ē	warnindicator	BIGINT
	✓ flag	ů	flag	BIGINT

v. 右键单击预测,选择从此处开始执行。

vi. 执行完成后,选择查看数据。

序号 ▲	uid 🔺	trend 🔺	xiansun 🔺	warnindicator 🔺	flag 🔺	prediction_result	prediction_score 🔺	prediction_detail
1	1	4	1	1	1	0	0.7936818743516113	{ "0": 0.793681874
2	3	2	1	1	1	0	0.9777937746755442	{ "0": 0.977793774
3	4	9	0	0	0	1	0.758433607940023	{ "0": 0.24156639
4	6	2	0	0	0	0	0.9993815703148501	{ "0": 0.99938157
5	11	10	1	2	1	1	0.9993127315069457	{ "0": 0.00068726
6	12	10	1	3	1	1	0.9998156465709368	{ "0": 0.00018435
7	13	2	0	3	0	0	0.9688889850628143	{ "D": 0.96888898
В	15	3	0	0	0	0	0.9979107883787329	{ "0": 0.99791078
9	17	9	0	3	1	1	0.9938992931601628	{ "0": 0.00610070
10	18	0	0	2	0	0	0.9992484524299784	{ "D": 0.99924845
11	20	2	0	4	0	0	0.8930436506958285	{ "D": 0.89304365
12	21	3	0	1	0	0	0.9922517024361496	{ "0": 0.99225170
13	35	2	1	4	1	1	0.5409565812943579	{ "0": 0.45904341
14	38	6	0	1	0	0	0.7678141618637798	{ "0": 0.76781416
15	39	1	0	3	0	0	0.9905983080109843	{ "0": 0.99059830
16	45	4	1	0	0	0	0.9348465344845021	{ "D": 0.93484653
17	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。

i. 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	ł	几器学习PAI 华东2(上海	ī)	r.	算法平台	前往运维						帮
ଜ	搜	素	Q	🛆 窃电	用户识别 ×							
首页	>	🧰) 目标) 运行 	- - ⑦ 部署	Q. Auto ML		Đ	€ @	*		57
<u>ぶ</u> 実验	>	📄 数据预处理				v						
	>	📄 特征工程										
Notebook	>	📄 统计分析					🛛 🖂 data (m) d					
	\sim	🔽 机器学习					Uata4III-1					
A.	>	📄 二分类										
组件	>	🧧 多分类							_			
Ŧ	>	■ 聚类										
1012 0	>						♥ ● ● ● ●	○	¥ ⇔.1		0	
(Q) (Q)	>	▶ 关联推荐					EQ BOADCEPT	· · · · · · · · · · · · · · · · · · ·]			
	~	🧧 评估										
		🌐 二分类评估										
		④ 回归模型评估						🛛 🛛 逻辑回)	\odot		
		① 聚类模型评估)				
		① 混淆矩阵						Ç	Ļ			
		③ 多分类评估						6 预测-1		\odot]	
	>	▶ 推荐算法]			
		13 预测						ſ				
	>	📄 深度学习	点击 <mark>这里</mark> 开					⊕ 二分类评估	1			
	>	📄 强化学习										

- ii. 连线预测中的预测结果输出和二分类评估中的输入。
- iii. 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。

€ € ⊕ ⊹ Ⅲ ፬ 23	字段设置
 ☐ data4ml-1 ④ data4ml-1 ④ data4ml-1 ④ data4ml-1 ④ 振分-1 ● 振分-1 ● 振行-1 ● 正 近期回, ** 折分-1 ● 正 近期回, ** 折分-1 ● 正 近期回, ** 近 	原始标签列列名 flag 分数列列名 prediction_score 正样本的标签値 1 计算KS,PR等指标时技等级分成多少个桶 1000 分組列列名(仅支持string类型) 。 高級选项

iv. 右键单击二分类评估,选择从此处开始执行。

v. 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,提供用户窃电行为的在线识别服务。

5.对接使用CDH

Dat aWorks 提供了与CDH(Cloudera's Distribution Including Apache Hadoop,以下简称CDH)集群对接的能力,在保留继续使用CDH集群作为存储和计算引擎的前提下,您可以使用Dat aWorks的任务开发、调度、数据地图(元数据管理)和数据质量等一系列的数据开发和治理功能。本文为您介绍如何对接使用CDH。

前提条件

● 已部署CDH。

支持非阿里云ECS环境部署的CDH,但需要确保部署CDH集群的ECS和阿里云网络可达。通常您可以使用高速通道、VPN等网络连通方案,来保障网络可达。

• 已开通DataWorks服务并创建好对接使用CDH的工作空间。

⑦ 说明 对接使用CDH的工作空间无需绑定计算引擎,在创建工作空间时可跳过选择引擎步骤,其 他步骤的操作详情可参见创建工作空间。

- 拥有一个有工作空间的管理员权限的账号,在DataWorks中新增CDH引擎配置的操作仅空间管理员可操作。为账号授权空间管理员权限的操作可参见。成员及角色管理
- 已购买并创建DataWorks的独享调度资源组。详情可参见独享资源组模式。

在DataWorks中对接使用CDH引擎时,主要配置流程为:

- 1. Step1: 获取CDH集群配置信息
- 2. Step2: 配置网络联通
- 3. Step3: 在DataWorks中新增CDH集群配置

对接配置完成后,您可在DataWorks上开发CDH引擎的数据开发任务并运行,并在运行后通过DataWorks的运维中心查看任务运行情况。详情可参见使用DataWorks进行数据开发和运维监控配置。

同时您可使用DataWorks的数据质量、数据地图功能,进行数据和任务管理。详情可参见数据质量规则配置和数据地图配置。

使用限制

- 在DataWorks中使用CDH相关功能,必须使用DataWorks的独享调度资源组。
- 您需要先保障CDH集群和独享调度资源组的网络可达后再进行后续的相关操作。
- 目前DataWorks支持的CDH版本有: cdh6.1.1、cdh5.16.2、cdh6.2.1和cdh6.3.2。

Step1: 获取CDH集群配置信息

获取CDH版本信息,用于后续DataWorks中新增CDH引擎配置。
 登录Cloudera Manager,在主界面集群名称旁可查看当前部署的CDH集群版本,如下图所示。

Cloudera Manager 群集 - 主机 - 诊断	▼ 审核 图表▼ 备份▼ 管理▼	搜索
主页 状态 所有运行状况问题 01 配置 🗲 11	▼ 所有最新命令	
您正在非生产模式下运行 Cloudera Manager,该模式使用嵌入式	PostgreSQL 数据库。请在移入生产环境之前切换为使用支	持的外部数据库。更多详细信息 🖸
Cluster 1 (CDH 6.1.1, Parcel)	图表	30分钟 1小时 2小时
◎ ■3主机 チ2	已完成的 Impala 队列	
♥ H HBase	puo	
❶ ➡ HDFS ● 1	280	
🗢 😵 Hive 💌	neries	
🗢 🕂 Hue 🔑 🔎 👻	o 06:30 06:45	
🗢 🦞 Impala 🕞	- Impala, 各 Impala Daemons 中的总 查询 0	
Key-Value Stor	HDFS IO	群集网络 IO
📀 👩 Oozie 👻	20b/s	
🗢 💏 Solr 🔍		6 97.7K/s
🗢 🛟 Spark		48.8K/s
S III YARN (MR2 In	کتر اور کی میں مرتبط کر کی کر	£ 06:30 06:45
🗢 🗼 ZooKeeper 🔑 1	=各 DataNodes 1.9b/s =各 DataNodes 1b/s	■各网络接口 18.8K/s ■各网络接口 134K/s

- 2. 获取Host地址与组件地址信息,用于后续DataWorks中新增CDH引擎配置。
 - 方式一: 使用DataWorks JAR包工具获取。
 - a. 登录Cloudera Manager, 下载工具JAR包。

wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar

b. 运行工具JAR包。

```
export PATH=$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
java -jar dw-tools.jar <user> <password>
```

其中 <user> 和 <password> 分别是Cloudera Manager的用户名和密码。

c. 在运行结果中查看并记录CDH的Host地址和组件地址信息。

[root@cdh-header-1-cn-shanghai ~]# wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
2021-01-08 18:52:55 https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar
Resolving dataworks-public-tools.oss-cn-shanghai.aliyuncs.com (dataworks-public-tools.oss-cn-shanghai.aliyuncs.com) 106.14.228.176
Connecting to dataworks-public-tools.oss-cn-shanghai.aliyuncs.com (dataworks-public-tools.oss-cn-shanghai.aliyuncs.com) 106.14.228.176 :443
connected.
HTTP request sent, awaiting response 200 OK
Length: 6743456 (6.4M) [application/java-archive]
Saving to: 'dw-tools.jar.1'
100%[===================================
2021-01-08 18:52:55 (36.0 MB/s) - 'dw-tools.jar.1' saved [6743456/6743456]
[root@cdh-header-1-cn-shanghai ~]# export PATH=\$PATH:/usr/java/jdk1.8.0_181-cloudera/bin
[root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin
Hosts:
192.168.22.217 cdh-header-1-cn-shanghai
192.168.22.219 cdh-worker-2-cn-shanghai
192.168.22.218 cdh-worker-1-cn-shanghai
Urls:
HiveServer2: jdbc:hive2://cdh-header-1-cn-shanghai:10000
Hive Metastore: thrift://cdh-header-1-cn-shanghai:9083
YARN ResourceManager: http://cdh-header-1-cn-shanghai:8032
Impala Daemon: jdbc:impala://cdh-worker-1-cn-shanghai:21050

○ 方式二:在Cloudera Manager页面手动查看。

登录Cloudera Manager,在主机(Hosts)下拉菜单中选择角色(Roles),根据关键字和图标识别出 需要配置的服务,然后看左侧对应的主机(Host),按照格式补全要填写的地址。默认端口号可以参 考方法一的输出结果样例。

Cloudera Manager ## -	主机 ▼ 诊断 ▼	审核 图表	备份▼	管理 🗸						搜索	.	山 。 支持	▪ admin ◄
角色	所有主机角色												
主机 cdh-header-1-cn-shanghai	主机模板 磁盘概述 Parcel	B SNN	SNN	😵 G	😵 HMS 😵 HS 🔅 SS 🖧 G	H LB	HS III JHS	¥ ICS	¥ ISS	🛠 LHBI	CAP	C ES	СНМ
cdh-worker-[1-2]-cn-shanghai	2 H RS	🖻 DN 🛛 😵 G	¥ ID	¢≇G	III NM								
此表按分配了相同角色的主机分组。													

其中

- HS2: HiveServer2
- HMS: Hive Metastore
- ID: Impala Daemon
- RM: YARN ResourceManager
- 3. 获取配置文件,用于后续上传至DataWorks。
 - i. 登录Cloudera Manager。
 - ii. 在状态页面,单击集群的下拉菜单中的查看客户端配置 URL。



iii. 在对话框中下载YARN的配置包。

Cloudera Manager 群集 - 主机 -	诊断 → 审核 图表 → 备份	} - 管理 -		援索 島 広 支持 ▼ admin ▼
主页 状态 所有运行状况问题 01 i				添加群集
您正在非生产模式下运行 Cloudera Manager,该模式使	客户端配置 URL		×	
Cluster 1 (CDH 6.1.1, Parcel)	▲名称	类型	URL	30分钟 1小时 2小时 6小时 12小时 1天 7d 30d 🖋 -
◎ ■3主机 ≯2	# YARN (MR2 Included)	YARN (MR2 Included)	/cmf/services/22/client-config	
H HBase	H HBase	HBase	/cmf/services/23/client-config	
0 🖸 HDFS 01 🗲 2	si Solr	Solr	/cmf/services/18/client-config	
🗢 🐐 Hive	- HDES	HDES	★/cmf/services/21/client-config	
● (i) Hue 📕 1		horo	- your serves ziy cherr coming	
♥ Impala	👻 🖌 😵 Hive	Hive	/cmf/services/24/client-config	
Key-Value Stor				
🗢 🧿 Oozie	-		14.17	
🗢 🎲 Solr	-		关闭	
🔿 🛟 Spark	· ·		tes	
III YARN (MR2 In	▲ 06:30			
🗢 🟦 ZooKeeper 🖉 🥕 1	▲ =各 DataNodes 1.9b/s	=各 DataNodes 1b/s	=各网络接□ 57.7K/s =各网	月络接口 19.7K/s

- 4. 获取CDH集群的网络信息,用于后续与DataWorks的独享调度资源组网络联通配置。
 - i. 登录部署CDH集群的ECS控制台。
 - ii. 在实例列表中找到部署CDH集群的ECS实例,在**实例详情**中查看并记录**安全组、专有网络、虚拟交** 换机信息。

云服务器 ECS	实例详惯	监控	安全组	云盘	快照	弹性网卡	远程命令	》/文件 排	製作记录	健康诊断		
概览	Â											
事件	基本信	Ū.							10	新健康状态 🔤	启动 重启	停止 配置安全组规则
标签	bw-te	est-node	ejs 🖄 📀	运行中								
自助问题排查时可	实例ID					远程	连接	地域	华东	1 (杭州)		
发送命令/文件 (云助手) 107	公网IP	_				转换为弹性的	公网IP	所在可用区	杭州	可用区		
	安全组				1	加入多	全组	主机名	test			修改实例主机名
実例与镜像 ヘ	标签	-				相当	關标签	创建时间	202	1年1月8日 20:34	:00	
实例	描述	-				修改实例	利描述	自动释放时间	8] -			释放设置
镜像	CPU&P	存 2 相	₹ 4GiB					云盘	1			重新初始化云盘
20040-00-00-00-00-00-00-00-00-00-00-00-00	操作系统	6 Cer	ntOS 7.2 64位	1		更换操作	作系统	快照	0			
钾性脊髓头例 ECI LS	实例规格	ecs	.t6-c1m2.larg	ge(性能约束到	定例()	更改实行	列规格	镜像ID	cent	tos_7_02_64_20	G_alibase_201	创建自定义镜像
专有宿主机 DDH	实例规格	號 ecs	.t6					当前使用带数	₹ 20N	lbps (峰值)		按量付赛实例更改带宽
超级计算集群												
预留实例券	网络信	息									绑定辅助	弹性网卡 更换专有网络
节省计划 🚥	网络类型	专行	有网络					RDMA IP				
	弹性网+	eni	-bp15a451q2	2qutp3ji4oc				弹性IP实例II	D -			
网络与安全 へ	专有网络				Ľ			虚拟交换机			Ľ	
安全组	±私网IF	192	2.168.0.58					辅助私网IP	-			

Step2: 配置网络联通

DataWorks的独享调度资源组购买创建完成后,默认与其他云产品网络不可达,在对接使用CDH时,您需获 取部署CDH集群的网络信息,将独享调度资源组绑定至CDH集群所在的VPC网络中,保障CDH集群与独享调度 资源组的网络联通。

- 1. 进入独享资源组网络配置页面。
 - i. 登录DataWorks控制台。
 - ii. 在左侧导航栏, 单击资源组列表, 默认进入独享资源组页签。
 - iii. 单击已购买的独享调度资源组后的网络设置。
- 2. 绑定VPC。

在专有网络绑定页签,单击新增绑定,在配置页面选择上述步骤记录的CDH集群所在VPC、交换机、安

全组。

3. 配置Host。

在Host 配置页签,单击批量修改,在对话框中配置为上述步骤中记录的Host 地址信息。

DataWorks / 资源组列表 ← sm_duxiang_for_adb ∽	<pre>[root@cdh-header-1-cn-shanghai ~]# export PATH=\$PATH:/usr/java/jdk1.8.0_181-cloudera/bin [root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin Hosts: 192. 217 cdh-header-1-cn-shanghai 192. 218 cdh-worker-2-cn-shanghai 192. 218 cdh-worker-1-cn-shanghai Urls: HiveServer2: jpbc:hive2://cdh-header-1-cn-shanghai:10000 Hive Metastore: fhrift://cdh-header-1-cn-shanghai:10003 YARN ResourceMargaer: http://cdh-header-1-cn-shanghai:8832 Impala Daemon: jdbc:impala://cdh-worker-1-cn-shanghai:21050</pre>
专有网络绑定 DNS配置 Host配置 ① 如果数据序是域名配置的,将会首先从下列Host中做域名 新留 2 批量修改词	exet [注] Host配F,优先极高于DNS配置 名Host配置 X
IPtété	ddh-header-1-cn-shanghai cdh-header-1-cn-shanghai cdh-header-1-cn-shanghai

Step3:在DataWorks中新增CDH集群配置

只有工作空间管理员才能进行新增CDH集群配置操作,操作时请使用拥有空间管理员权限的账号。

- 1. 进入项目空间管理页面。
 - i.
 - ii. 在左侧导航栏,单击工作空间列表。
 - iii. 在对应工作空间的操作列单击工作空间配置。
 - iv. 在右侧配置页面中单击**更多设置**。
- 2. 在项目空间管理页面,单击CDH集群配置。



3. 在CDH集群配置页面单击立即新增,在新增CDH集群配置对话框中,填写上述步骤Step2:配置网络 联通中记录的组件地址信息。

新增CDH集群配置			
集群信息			
* 集群名称:	doctest		
* CDH集群版本:	5.16.2 ~		
Hive			
版本选择:	1.1.0 🗸		
* HiveServer2: 0	jdbc:hive2:// <host>:<port>/<database></database></port></host>		
* Metastore : 💿	thrift:// <host_1>:<port_1>,thrift://<host_2>:<port_2></port_2></host_2></port_1></host_1>		[root@cdh-header-1-cn-shanghai ~]# export PATH=\$PATH:/usr/java/jdk1.8.0_181-cloudera/bin [root@cdh-header-1-cn-shanghai ~]# java -jar dw-tools.jar admin admin
Impela		\setminus	Hosts: 192.168.22.217 cdh-header-1-cn-shanghai
版本选择:	2.12.0 🗸		192.168.22.219 cdh-worker-2-cn-shanghai 192.168.22.218 cdh-worker-1-cn-shanghai
* JDBC地址: 🜖	jdbc:impala:// <host>:<port>/<schema></schema></port></host>		
Spark			HiveServer2: jdbc:hive2://cdh-header-1-cn-shanghai:10000
版本选择:	1.6.0 ~		Hive Metastore: thrift://cdn-header-1-ch-shanghal:9083 YARN ResourceManager: http://cdh-header-1-cn-shanghal:8032
Yern			Impala Daemon: jdbc:impala://cdh-worker-1-cn-shanghai:21050
版本选择:	26.0 ~	\land	
* yarn.resourcemanager	address: 0 http:// <host>:<port></port></host>		
jobhistory.webapp.addre	ess: () http:// <host><<pre>http://<host></host></pre></host>		
MapReduce	yarn.resourcemar	nag	mager.address地址的端口修改为8088即为jobhistory.webapp.address
版本选择:	26.0 🗸		
Presto	Presto非CDH默认组件,	需要	需要根据实际部署情况填写访问地址
版本选择:	0.244.1		
* JDBC地址: 🧿	jdbc:presto:// <host>:<port>/<catalog>/<schema></schema></catalog></port></host>		

其中:

- 集群名称: 可自定义集群名称。
- 版本信息:根据实际情况选择对应的CDH和组件版本。
- 地址信息: 根据上述步骤中记录的地址信息填写。其中:
 - Yarn的jobhistory.webapp.address信息: yarn.resourcemanager.address地址的端口修改 为8088即为jobhistory.webapp.address。
 - Presto的JDBC地址: Presto非CDH默认组件,需要根据实际部署情况填写访问地址。
- 4. 上传配置文件并授权给其他工作空间。

1462			Coopera manager and a				
配置文件: •	t@core aite.aml		王贞 US ANDRUSHE 00	ALC: NO.	客户编記言 URL		
配置文件: •	上的hdfs-site.xml						
配置文件: •	上桥mapred-site ami		Chatter Tatters Convent		8.6	RU	(IRL
配置文件: •	L thysers eite anni		6 2. La	-	and the second	white fully excluded	A romthanvises/22/otent contig
In the second se	h Manufacture and a second second		a Areas		H Hitess	HEase	A rand/services/23/client-config
alloci+1	Tracound properties		0 9 mm		() Sol	Sale	A rowt/services/10/client-config
200文/#: 0	上的presto jka		0 minus 191		C HOPS	HDFS	A rowthervices/21/chere-config
-			© V impairs		Se then	Hote	A /cmt/services/34/client-carely
TUZHBBB			G of any local line .				
海北工作空间: 0	161642 ×	~	O Down				AM .
	请输入工作空间名称进行活动 (不能删除当给工作空间						

5. 配置访问身份的映射关系。

如果您希望在运行任务时,对不同云账号在CDH集群内可访问的数据进行数据权限隔离,则可开 启Kerberos账号(principal)认证,并配置云账号与Kerberos账号的权限映射关系。

⑦ 说明 Kerberos账号为CDH集群的访问账号。CDH集群通过Sentry或Ranger组件为Kerberos 账号进行不同权限的配置,实现数据权限隔离。与Kerberos账号存在映射关系的云账号拥有相同的CDH集群数据访问权限。请填写格式为 实例名@领域名 的Kerberos账号(principal),例 如,cdn test@HADOOP.COM。

映射配置	
* 认证类型:	○ 无认证方式
	● Kerberos账号 (principal) 认证
* 配置文件:	上传krb5.conf
* 配置文件:	上传hive site.xml
✓ 配置引擎权限映射	
* 云账号: 请选择	✓ *kerberos账号: 请填写账号
● * 上传文件: 上传ke	eytab文件
添加	

6. 单击确定,完成新增CDH集群配置。

完成新增CDH集群配置后,已授权的工作空间中可新增此CDH引擎,用于后续编辑并运行数据开发等任务。

Step4:在DataWorks中新增CDH引擎

- 1. 在项目空间管理页面, 单击工作空间配置。
- 2. 在计算引擎信息区域的CDH页签单击增加实例,在弹窗中配置实例信息。

新增引擎实例时,可选择使用**快捷模式**或安全模式访问模式,安全模式可以实现不同云账号运行任务时的数据权限隔离。不同访问模式的配置界面如下:

• 快捷模式的实例信息配置。

增加CDH引擎	实例		×
* 实例显示名称:	- ⁷	輸入实例显示名称 1	Â
* 访问模式:	甘	2 ~	- 1
集群信息	请	选择未开启Kerberos或LDAP认证的集群!	- 1
* 选择集群:	С	DH_CLUSTER 3	- 1
集群版本:	6.1	.1	- 1
Hive	2.1.1	HiveServer2: jdbc:hive2://cdh-header-1-cn-shanghai:10000 Metastore: thrift://cdh-header-1-cn-shanghai:9083	
Presto	0.244.1	JDBC地址: jdbc:presto://cdh-header-1-cn-shanghai:8080	
Impala	3.1.0	JDBC地址: jdbc:impala://cdh-worker-1-cn-shanghai:21050	
Spark	2.4	配置文件: 🗸 已上传	
Yarn	3.0.0	yarn.resourcemanager.address: http://cdh-header-1-cn- shanghai:8032 jobhistory.webapp.address: http://cdh-header-1-cn-shanghai:803	88
MapReduce	3.0.0	配置文件: 🗸 已上传	
访问身份		4	
* 认证类型:	Q	无认证方式	- 1
* 账号:	加	根据认证类型填写账号	- 1
网络连通性			- 1
请添加	中享调度资源	组以实现DataWorks与CDH集群的连通!	
* 独享调度资源	组: du 请 享	oc_test 5 刷新 思参考此文档对独享资源组进行网络配置。如当前地域未购买独 调度资源,请购买后再进行配置。	
测试网络连通性	:	测试连通性 6	•
		确定	取消

○ 安全模式的实例信息配置。

增加CDH引導	歐例		×
* 实例显示名称	: i	输入实例显示1	
* 访问模式:	安	全模式 2 ~	
集群信息	请	选择已开启Kerberos或LDAP认证的集群!	
* 选择集群:	C	DH_CLUSTER 3	
集群版本:	6.1	.1	
Hive	2.1.1	HiveServer2; jdbc:hive2://cdh-header-1-cn-shanghai:10000 Metastore: thrift://cdh-header-1-cn-shanghai:9083	
Presto	0.244.1	JDBC地址: jdbc:presto://cdh-header-1-cn-shanghai:8080	
Impala	3.1.0	JDBC地址; jdbc:impala://cdh-worker-1-cn-shanghai:21050	
Spark	2.4	配置文件: 🗸 已上传	
Yarn	3.0.0	yarn.resourcemanager.address: http://cdh-header-1-cn- shanghai:8032 jobhistory.webapp.address: http://cdh-header-1-cn-shanghai:808	8
MapReduce	3.0.0	配置文件: ✓ 已上传	
 访问身份 * 调度访问身份 网络连通性 	: C	任务责任人 🔵 阿里云主账号 🔵 阿里云子账号	
请添加	虫享调度资源 :	且以实现DataWorks与CDH集群的连通!	
* 独享调度资源	组: do 请 调	pc_test 5 刷新 思参考此文档对独享资源组进行网络配置。如当前地域未购买独享 度资源,请购买后再进行配置。	
测试网络连通性	:	则武连通性 6	
		确定	取消

- i. 填写实例显示名称。
- ii. 选择访问模式
 - 快捷模式

该访问模式使用便捷,多个云账号对应一个集群账号,多个账号均可访问同一个集群账号内的数据,无法实现不同云账号运行任务时的数据权限隔离。

■ 安全模式

该访问模式允许您配置云账号与CDH集群账号的身份映射关系,实现不同云账号运行任务时的数据权限隔离。

iii. 选择上述新增的CDH集群配置。

如果上一步**访问模式**选择**快捷模式**,则此处选择未开启Kerberos认证的CDH集群。如果**访问模** 式选择**安全模式**,则此处需要选择已开启Kerberos认证的CDH集群。您可以进入工作空间配置查看 CDH集群是否开启Kerberos认证。

- iv. 设置访问集群的认证信息。
 - 快捷模式

当前仅支持指定特定账号,建议使用admin或hadoop账号。该账号仅用于下发任务。

■ 安全模式

您可以根据需求选择**调度访问身份**。该身份用于在任务提交调度后自动调度运行任务,并且需 要配置云账号与CDH集群账号的身份映射,详情请参见<mark>配置访问身份映射</mark>。

⑦ 说明 在DataStudio页面,运行任务所使用的身份均为当前已登录云账号映射的集群访问身份。因此,除了需要为调度访问身份配置身份映射外,建议为项目空间开发成员也配置身份映射,避免页面运行任务失败。

- v. 选择已经购买好的独享调度资源组。
- vi. 单击测试连通性。

如果网络连通测试失败,可能是因为独享调度资源组没有绑定CDH集群所在的专有网络,或者独享 调度资源组没有设置Host,请参见Step2:配置网络联通检查独享调度资源组的网络配置。

3. 单击确定, 创建计算引擎实例。

此步骤会触发独享调度资源组的初始化(安装访问CDH集群的客户端以及上传配置文件),您需要等 待**独享资源组初始化状态** 从**准备中**变成完成,CDH引擎实例才创建完成。

4. 在创建的CDH引擎实例页面单击**测试服务连通性**, DataWorks会运行测试任务测试客户端和配置文件是 否正确安装。

如果测试结果显示失败,您可以查看日志并提交工单联系DataWorks技术支持。

使用DataWorks进行数据开发

完成新增CDH引擎后,您就可以在DataStudio(数据开发)中创建Hive、Spark、MapReduce、Impala或者 Presto任务节点,直接运行任务或者设置周期调度运行任务。以下以创建并运行一个Hive任务为例,为您介 绍在DataWorks中如何进行CDH引擎的数据开发和运行。

1. 进入DataStudio页面。

i.

- ii. 在左侧导航栏, 单击**工作空间列表**。
- iii. 在对应工作空间的操作列单击进入数据开发。
- 2. 创建业务流程,根据界面提示填写业务流程信息。
- 3. 单击创建好的业务流程,在CDH引擎文件夹上右键选择新建 > CDH Hive。



4. 在右侧代码编辑框中编写Hive SQL,完成代码编辑后单击顶部 💽 运行图标,选择调度资源组并确认,

运行完毕后可以查看Hive SQL的运行结果。

- 5. 如果想要设置任务周期调度,单击右侧的调度配置,在弹窗中设置时间属性、资源属性和调度依赖,完成后单击提交任务,提交成功后任务就可以按照配置周期调度运行,调度配置详情可参见配置基础属性。
- 6. 在运维中心中可以查看提交的周期任务,在周期实例中查看任务周期调度的运行情况。详细可参见查看 周期任务。

运维监控配置

CDH引擎的任务支持使用DataWorks运维中心的智能监控功能,通过自定义报警规则、配置任务告警,根据 设置的报警规则自动触发任务运行异常报警。自定义报警规则操作可参见自定义规则,配置任务告警操作可 参见基线管理。

数据质量规则配置

在DataWorks上使用CDH引擎时,可使用DataWorks的数据质量服务进行数据查、对比、质量监控、SQL扫描和智能报警等功能,数据质量服务的详细操作可参见数据质量概述。

数据地图配置

在DataWorks上使用CDH引擎时,可使用DataWorks的数据地图服务采集CDH集群中Hive数据库、表、字段、 分区元数据,便于实现全局数据检索、元数据详情查看、数据预览、数据血缘和数据类目管理等功能。

⑦ 说明 当前仅支持Hive数据库。

DataWorks上数据地图功能的详细介绍与配置指导可参见概述。

如果要您希望可以实时感知CDH集群中Hive元数据的变更,或者要在数据地图中查看血缘和元数据变更记录,需要将DataWorks的Hive Hook嵌入到目标集群,并通过阿里云日志服务采集Hive Hook产生的日志。

配置Hive Hook后,元数据变更消息会被记录到HS2和HMS服务器的日志文

件/*tmp/hive/hook.event.*.log*中,使用阿里云日志服务采集后供DataWorks读取,下载DataWorks小工 具 dw-tools.jar ,在同一目录下创建 config.json 文件并补全配置项的值,最后执行工具一键创建日 志采集。

配置Hive Hook和采集Hive Hook日志的操作步骤如下。

- 1. 配置Hive Hook。
 - i. 登录HS2和HMS服务器,并进入/var/lib/hive目录,下载DataWorks Hive Hook。

```
# CDH 6.x 版本下载 dataworks-hive-hook-2.1.1.jar
wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dataworks-hive-hoo
k-2.1.1.jar
# CDH 5.x 版本下载 dataworks-hive-hook-1.1.0-cdh5.16.2.jar
wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dataworks-hive-hoo
k-1.1.0-cdh5.16.2.jar
```

- ii. 登录Cloudera Manager首页后,进入Hive > 配置,把Hive辅助JAR目录配置项设置为 /var/lib/h ive 。
- iii. 在Hive > 配置 中,将 hive-site.xml 的Hive服务高级配置代码段(安全阀)配置项添加以下内容。

```
<property>
<name>hive.exec.post.hooks</name>
<value>com.cloudera.navigator.audit.hive.HiveExecHookContext,org.apache.hadoop.hi
ve.ql.hooks.LineageLogger,com.aliyun.dataworks.meta.hive.hook.LineageLoggerHook</va
lue>
</property>
```

iv. 在**Hive > 配置**中,将 hive-site.xml 的Hive Metastore Server高级配置代码段(安全阀)配置 项添加以下内容。

```
<property>
  <name>hive.metastore.event.listeners</name>
  <value>com.aliyun.dataworks.meta.hive.listener.MetaStoreListener</value>
  </property>
  <name>hive.metastore.pre.event.listeners</name>
  <value>com.aliyun.dataworks.meta.hive.listener.MetaStorePreAuditListener</value>
  </property>
```

v. 配置完成后,根据Cloudera Manager的提示部署客户端配置,然后重启Hive服务。

(?) 说明 如果重启失败,保留日志用于排查问题,为防止影响正常作业可以先去掉上面两个步骤添加的配置再重启恢复Hive服务。如果添加配置后重启成功,查看服务器/tmp/hive/下是 否产生名称以 hook.event 开头的日志文件,例如 hook.event.1608728145871.log 。

2. 采集Hive Hook日志。

i. 登录Cloudera Manager, 下载工具JAR包。

wget https://dataworks-public-tools.oss-cn-shanghai.aliyuncs.com/dw-tools.jar

ii. 在小工具所在目录创建 config.json , 根据以下文件内容要求修改并保存文件。

```
// config.json
{
    "accessId": "<accessId>",
    "accessKey": "<accessKey>",
    "endpoint": "cn-shanghai-intranet.log.aliyuncs.com",
    "project": "onefall-test-pre",
    "clusterId": "1234",
    "ipList": "192.168.0.1,192.168.0.2,192.168.0.3"
```

其中:

- accessId: 阿里云账号的AccessKey ID。
- accessKey: 阿里云账号的AccessKey Secret。
- endpoint:填写为日志服务project的访问域名中的私网域名,详细可参见服务入口。
- project:填写为使用的阿里云日志服务的project名称,您可参见管理Project获取日志服务的 project名称。
- clusterId: 填写为DataWorks生成的CDH集群ID, 可以提交工单获取此ID。
- ipList:填写为HS2和HMS的所有服务器的IP列表(即部署了DataWorks Hive Hook的所有服务器 IP),多个IP使用英文逗号(,)分隔。
- iii. 运行配置文件。

java -cp dw-tools.jar com.aliyun.dataworks.tools.CreateLogConfig config.json

iv. 安装客户端。

wget http://logtail-release-cn-shanghai.oss-cn-shanghai.aliyuncs.com/linux64/logtai
l.sh -0 logtail.sh; chmod 755 logtail.sh; ./logtail.sh install cn-shanghai

其中cn-shanghai改为日志服务对应的Region。

 完成上述步骤后,在阿里云日志服务的指定project下会生成名为hive-event日志库、名为hive-eventconfig的logtail配置以及名为hive-servers的机器组。您可以查看并记录阿里云账号ID、日志服务的 endPoint和Project信息,将这些信息通过提交工单提供给DataWorks技术人员,由技术人员进行后续的 配置。