

数据开发治理平台 WeData

实践教程



腾讯云

【 版权声明 】

©2013–2024 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】

及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。

您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或95716。

文档目录

实践教程

对接 WeData 平台执行周期性调度任务

数据质量自定义模板使用

实践教程

对接 WeData 平台执行周期性调度任务

最近更新时间：2024-01-19 17:30:42

背景

本文为您介绍如何搭配腾讯云数据开发治理平台 Wedata 和 TI-kit CLI 工具，进行任务周期性调度，该功能适用于同时有数据治理需求和机器学习需求的业务场景，可以在 Wedata 页面进行数据开发任务和机器学习任务的统一调度。

操作流程

您可以按照如下最佳实践流程实现 TI-ONE 和 Wedata 平台（数据开发治理平台 WeData 是位于云端的一站式数据开发治理平台，详细介绍请查看 [文档](#)）的对接。TI-ONE 机器学习任务调度能力当前仅支持 Wedata 广州地域的企业版。

步骤一 准备工作

1. 创建用户及项目

在 Wedata 产品内需要首先创建用户及项目，详情操作指引请查看 [创建用户及项目](#)。

2. 配置自定义调度资源组

启用 TI-ONE 对接功能需要首先配置企业版自定义调度资源组，详情操作指引请查看 [自定义调度资源组列表](#)。

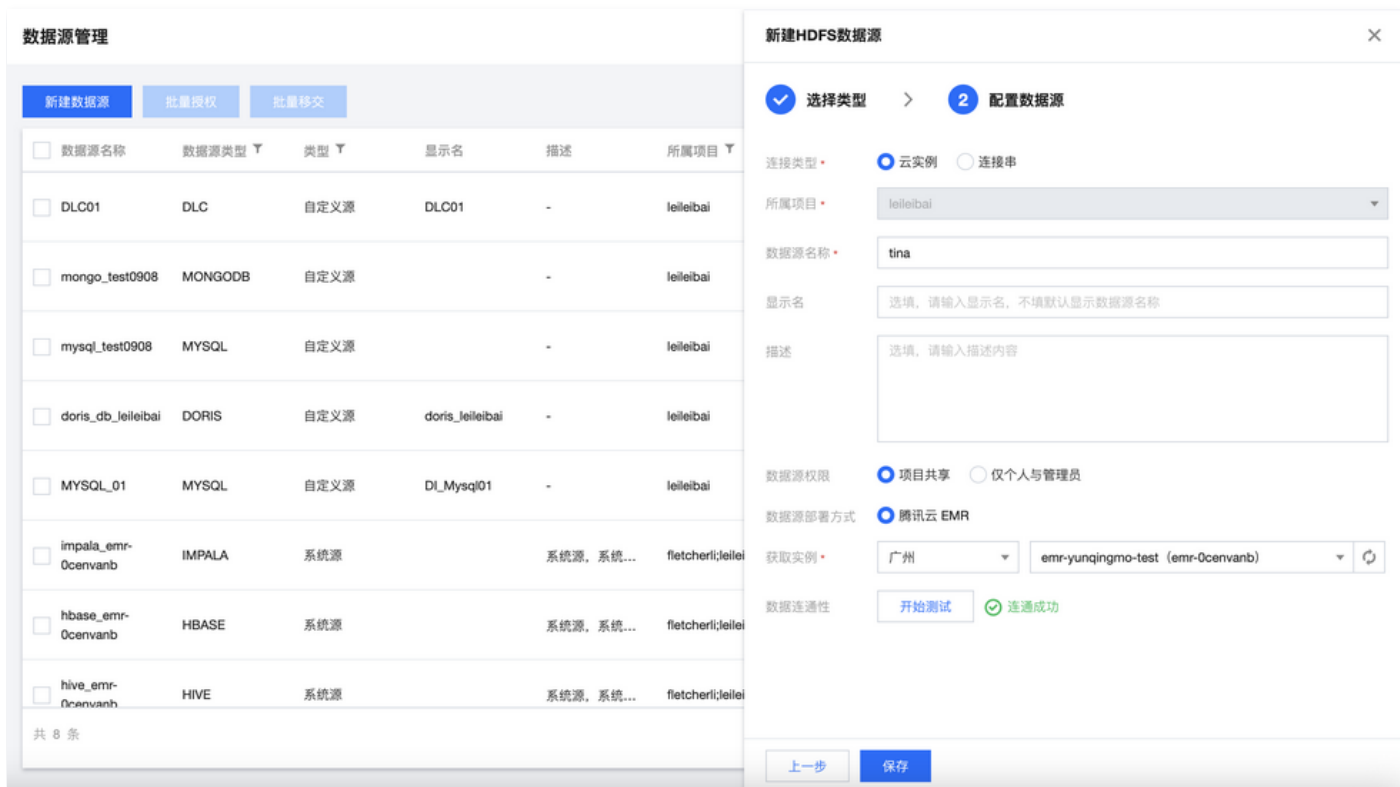
步骤二 初始化环境

在 Wedata 项目空间中添加执行资源组，添加服务器后需要同时安装 Wedata Agent 和 TI-ONE CLI 机器学习环境。登录机器安装完毕后，可以看到资源组节点的状态为正常。

← tina				
添加服务器				
所属网络	内网IP	状态	添加时间	操作
vpc-k4hwgy3		正常	2022-09-29 16:30:19	初始化 监控 退役 删除
共 1 条				
				10 条 / 页
				1 / 1 页

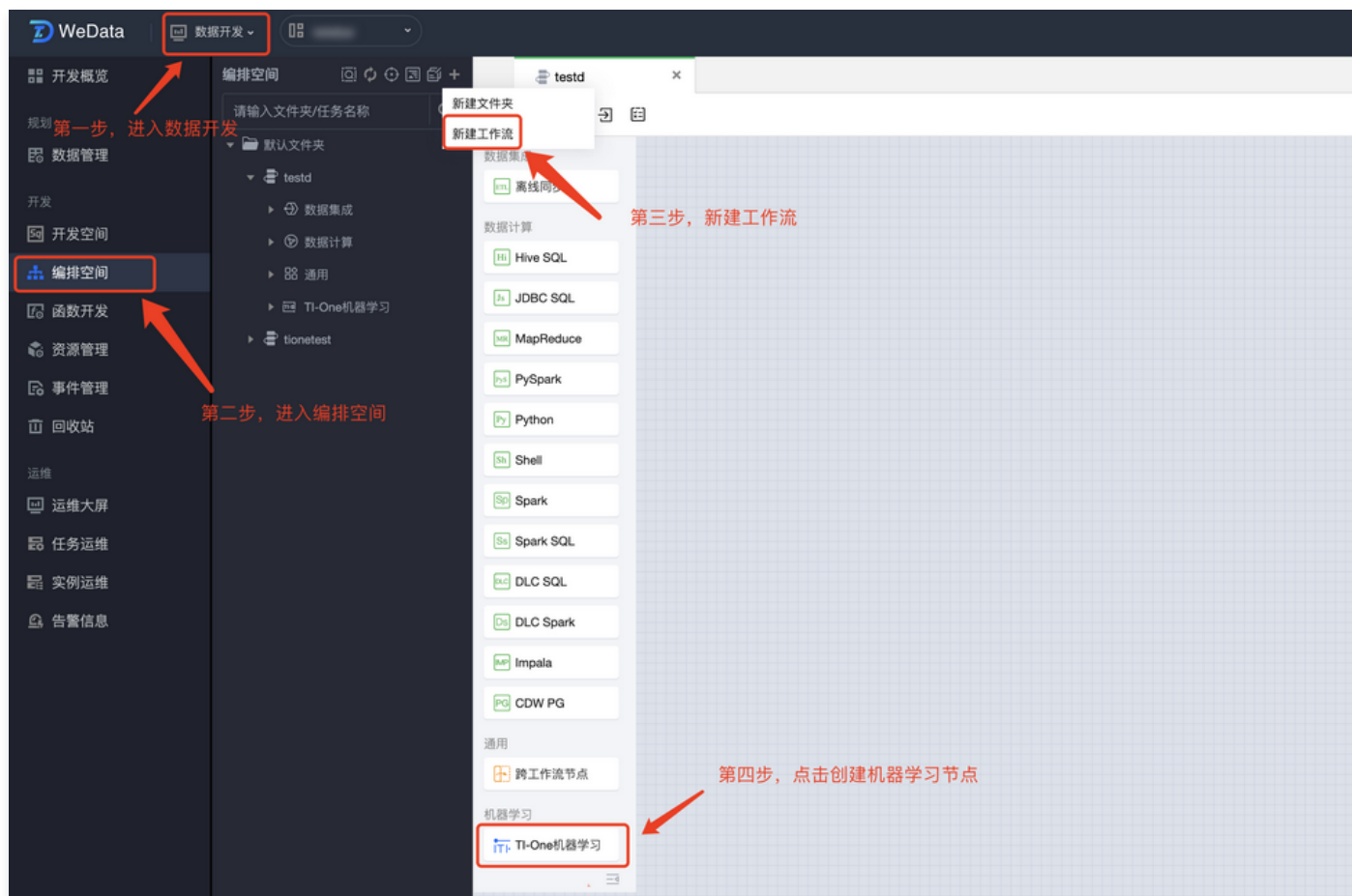
步骤三 添加数据源

在 Wedata 项目空间中添加数据源，添加一个 HDFS 或者 HIVE 的数据源，测试连通性，需要注意，创建完成后，要授权给需要使用的项目。数据源创建详细操作指引请查看 [文档](#)。



步骤四 机器学习节点配置

1. 进入数据开发 > 编排空间，创建工作流，在工作流编排面板中，单击 TI-ONE 机器学习节点创建。



新建任务

所属工作流testd

任务名称test

任务类型TI-ONE机器学习

确认取消

2. Wedata 中的机器学习节点本质上是一个安装了机器学习任务执行环境 Tikit 的 Shell 节点，用户需要在这个节点中编写 Tikit 命令，用于调度 TIONE 算力提交训练任务。
3. 进入节点配置页面后，单击**机器学习属性**，可进行数据源配置和算法开发配置。其中数据源配置可下拉选择当前训练任务所关联的数据源（若机器学习节点上游连接了其他节点，可在下方展示上游父任务数据源），下拉后会展示数据源 ID，该 ID 可用于脚本开发和训练任务提交。
4. 在提交训练任务前，我们需要准备训练代码，TIONE 提供轻量便捷的交互式开发环境 Notebook，可点击右侧进入 TIONE Notebook 进行代码编写（跳转至 TIONE Notebook 实例创建页面后，会默认带上选中数据源的网络信息，若数据源为 HDFS，也会默认在数据目录中选中该数据源）。若当前机器学习任务关联了某个 Notebook 实例，可直接下拉选中，页面会显示快速跳转链接和实例运行状态。

testtina 模型训练 test testd test1 特征处理

```
1 #!/bin/bash
2 #*****
3 ##author: wedata_sjzl
4 ##create time: 2022-11-01 15:19:49
5 ##用于在Linux环境下，以命令行的方式 发起创建任务、查询任务
6 ##使用前请先执行tikit init --secretid=xxx --secretkey=xxx
7 ##secretId和secretKey是腾讯云的访问密钥，可以在腾讯云控制台获取
8 ##使用tikit -h 获取更多帮助信息
9 #*****
10 tikit -h
```

机器学习属性

数据源配置

数据源名称①hive_eml

数据源ID: 4079

上游父任务数据源

任务名称	类型	数据源名称	数据源ID
特征处理1	HIVE	hive_emr-0c...	4079
特征处理2	HIVE	hive_emr-0c...	4079

共 2 条

算法开发

Notebook实例①ddd

查看实例 ddd

实例状态已停止

运行日志

任务属性 调度设置 版本 机器学习属性

步骤五 使用 TICLI 编写训练任务提交命令

1. 进入机器学习节点后，使用前请先执行 `tikit init --secretid=xxx --secretkey=xxx`，进行初始化。`secretId` 和 `secretKey` 是腾讯云的访问密钥，获取方式：进入控制台，单击右上角头像，进入访问管理 > API 密钥管理获取。
2. 使用前输入 `tikit -h`，获取 `tikit` CLI 工具各命令的运行方式。

-
3. 根据当前所需的任务类型提交任务，可在当前 shell 节点运行命令测试，任务提交后可在运行日志中打印对应的 TI-ONE 任务 URL，可前 [TI-ONE 控制台](#) 查看训练任务详情。

步骤六 提交 workflow 进行周期调度

当完成 workflow 开发后，可以配置 workflow 周期调度参数，并且将 workflow 整体提交。提交完成后，可在 [任务运维](#) 模块查看 workflow 和任务，当生成周期性实例后，可在 [实例运维](#) 页面查看实例详情。调度相关详细操作指引请查看 [workflow 列表](#)。

testtina

模型训练

test

testd

test1

特征

数据集成

离线同步

数据计算

Hive SQL

JDBC SQL

MapReduce

PySpark

Python

Shell

Spark

Spark SQL

DLC SQL

DLC Spark

Impala

CDW PG

通用

跨工作流节点

特征处理2

模型训练

统一调度

对工作流下所有任务设置统一的调度配置，支持常规和crontab方式，常规方式可对工作流下的任务调度配置进行单独修改，crontab方式不支持对任务调度配置进行单独修改，设置后原来的配置将会被覆盖，请谨慎操作！

调度策略

配置方式①

常规

crontab

调度周期

周期

一次性

天

周期说明

生效时间

2022-11-01 ~ 2099-12-31

执行时间

00:00

调度计划

每天00:00执行一次

自依赖

并行

无序串行

有序串行

工作流自依赖①

是

否

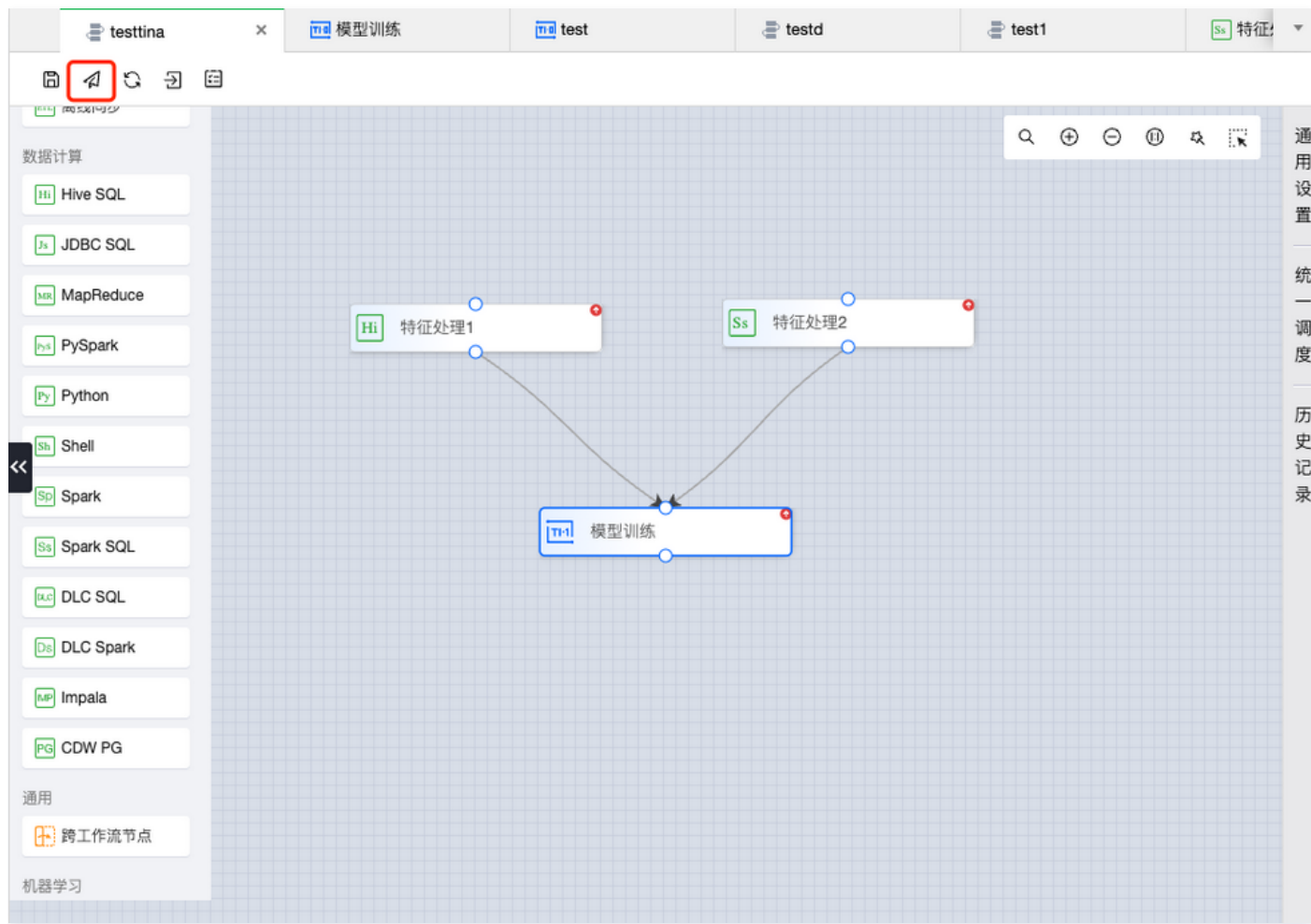
重置

保存

通用设置

统一调度

历史记录



数据质量自定义模板使用

最近更新时间：2024-05-22 10:53:02

背景

腾讯云数据开发治理平台 WeData 数据质量支持自定义模板创建和批量管理，帮助您根据业务场景定制化表质量检测逻辑。本文将为您介绍如何通过自定义模板页面新建规则模板、并根据自定义规则模板在数据监控页面对表创建检测规则。

操作流程



步骤一 准备工作

1. 创建用户及项目
在 WeData 产品内需要首先创建用户及项目，详情操作指引请查看 [创建用户及项目](#)。
2. 创建调度资源组
运行质量检测任务需要创建调度资源组，详情操作指引请查看 [调度资源组](#)。

步骤二 创建自定义模板

1. 进入数据质量 > 规则模板，单击自定义模板，新增模板并保存。

SQL 表达式：

```
select count(table1.${table_1.column_2}) AS count
from ${table_1} table1
join ${table_2} table2
on table1.${table_1.column_1} = table2.${table_2.column_1}
where table1.${table_1.column_3} >= ${param_1} and table1.${table_1.column_3} <= ${param_2}
and table2.${table_2.column_2} >= ${param_3} and table2.${table_2.column_2} <= ${param_4};
```

解释说明：

- 上文中一共出现了两张表：\${table_1} 和 \${table_2}，
 - \${table_1} 表示监控规则扫描的主表；
 - \${table_2} 表示同数据源同数据库下的其他表（实际使用时也可以选择主表自己）；
- 使用了表1四个字段，分别为：
 - \${table_1.column_1}：用于与表2关联；
 - \${table_1.column_2}：用于结果计数；
 - \${table_1.column_3}：用于过滤条件，大于等于参数1，小于等于参数2；
 - \${table_1.column_4}：表示表1的分区字段，可极大的节省计算资源，避免扫描全量数据；
- 使用了表2两个字段，分别为：
 - \${table_1.column_1}：用于与表1关联；
 - \${table_1.column_2}：用于过滤条件，大于等于参数3，小于等于参数4；
- 使用了4个 where 参数，分别为：
 - \${param_1}：SQL 中表1字段3的最小值；
 - \${param_2}：SQL 中表1字段3的最大值；
 - \${param_3}：SQL 中表2字段2的最小值；
 - \${param_4}：SQL 中表2字段2的最大值。
- 最终计算结果：为符合条件的表1字段2的个数，为一个数值。

截图示例：

质量概览

规则模板

数据监控

运维管理

质量报告

规则模板

系统模板

自定义模板

新增

批量删除

全部(7)

模板名称	模板描述	模板类型	维度
比较固定值大小_分区	-	表级	准确性
多表关联_不用别名	-	表级	准确性
多表关联模板	-	表级	准确性
比较固定值大小	-	表级	准确性
表行数	-	表级	准确性
中文	-	字段级	有效性
中文检测	-	字段级	有效性

共 7 条

创建模板

模板名称

多表关联

模板类型

表级

模板维度

准确性

适用引擎

请选择

模板描述 (可选)

请输入

SQL表达式

☒ 需要关联其他库表数据 (目前仅支持关联一个)
 ☒ 添加where过滤参数

勾选

```

1 select count(table1.${table_1.column_2}) AS count
2 from ${table_1} table1
3 join ${table_2} table2
4 on table1.${table_1.column_1} = table2.${table_2.column_1}
5 where table1.${table_1.column_3} >= ${param_1} and table1.${table_1.column_3}
6 and table2.${table_2.column_2} >= ${param_3} and table2.${table_2.column_2} <=
                    
```

填入 SQL

保存

取消

步骤三 创建质量规则

1. 进入数据监控，找到需要监控的表，单击配置监控任务。

质量概览

规则模板

数据监控

运维管理

质量报告

单表新增规则

选择监控对象

数据源

hive_emr-b2owxpm

数据库

default

监控表

aa

为监控对象配置监控规则

新增监控规则

批量设置执行策略

批量设置订阅信息

批量删除

执行策略配置数量 0/0

订阅信息配置数量 0/0

规则名称	规则描述	监控对象	模板类型	规则模板	检测范围	触发条件	触发等级	执行策略 (选...)	订阅信息 (选...)	操作
------	------	------	------	------	------	------	------	-------------	-------------	----

2. 单击新增规则，规则类型选择自定义模板，选中刚创建的模板，根据模板变量选择库表参数及 where 参数，配置好触发条件及等级，单击保存。

规则类型

自定义模板

选择模版

多表关联

select count(table1.\${table_1.column_2}) AS count from
\${table_1} table1 join \${table_2} table2 on
table1.\${table_1.column_1} = table2.\${table_2.column_1}
where table1.\${table_1.column_3} >= \${param_1} and
table1.\${table_1.column_3} <= \${param_2} and
table2.\${table_2.column_2} >= \${param_3} and
table2.\${table_2.column_2} <= \${param_4} and
table1.\${table_1.column_4} = '\${yyyy-MM-dd-1d}';

适用引擎

HIVE

库表参数 ①

table_1(当前表)

已选中的所有表

column_1

id(bigint)

column_2

id(bigint)

column_3

id(bigint)

column_4

id(bigint)

表 1 以及表 1 的四个参数

表 2 以及表 2 的两个参数

table_2(关联表)

default.ods_gf_chain_user_info

column_1(关联字段)

id(int)

column_2

user_id(bigint)

where 参数 ①

param_1

1

param_2

100

param_3

1

param_4

100

where 参数

触发条件

等于

1

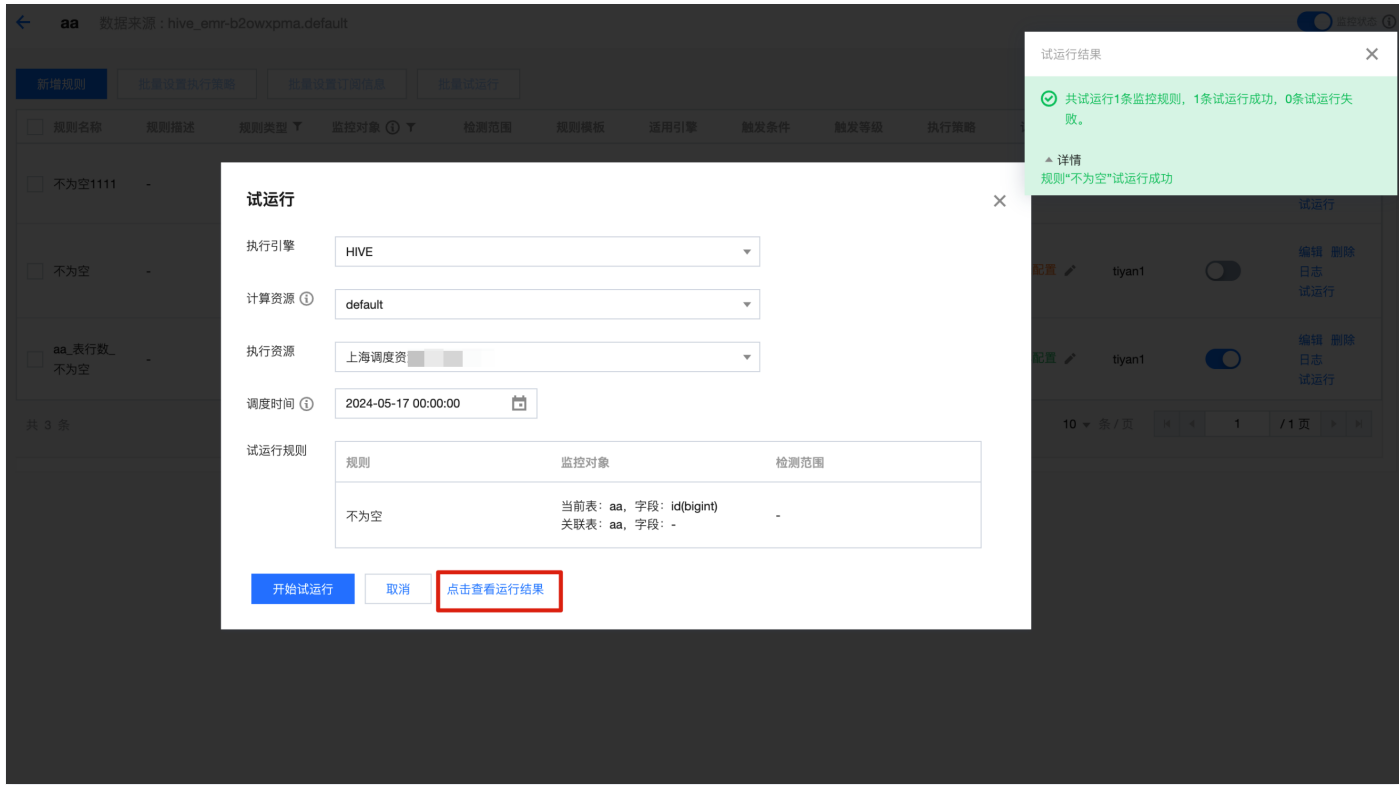
添加

注意:

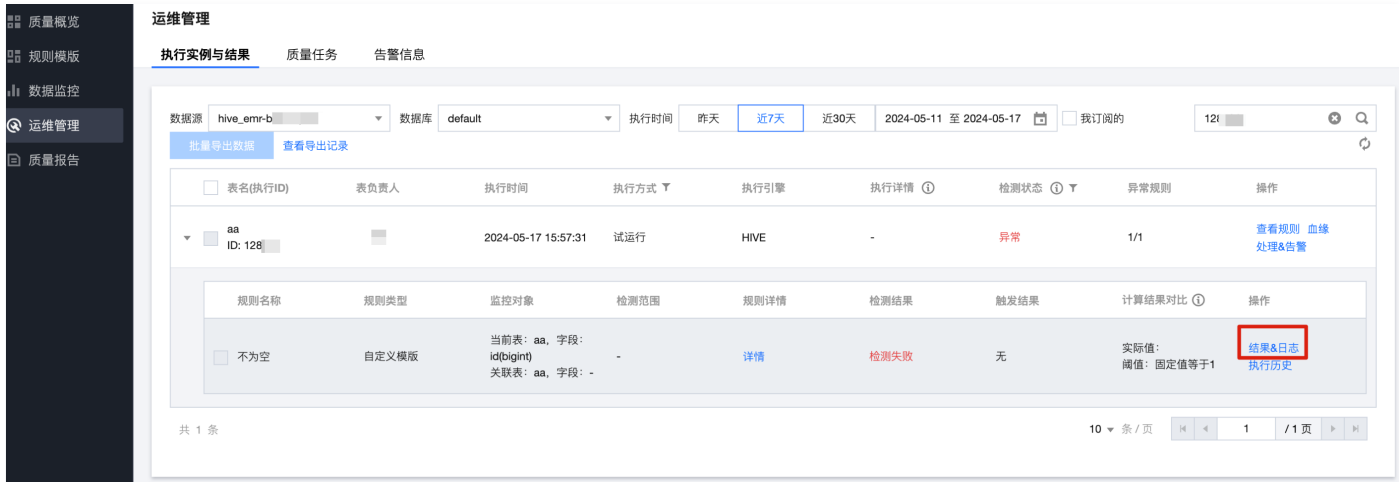
使用自定义模板前请先分析每个字段是什么含义，再进行映射。

步骤四 测试运行

1. 单击**试运行**，选择执行引擎、计算资源、执行资源，在验证规则中选择刚创建的规则。



2. 单击查看运行结果，跳转到运维管理页面查看运行结果。



3. 单击结果&日志，查看运行日志。

其中 EXECUTING SQL : xxxxxx，打印的是提交给 hive/spark/dlc 引擎进行质量检测 SQL。

质量概览

规则模板

数据监控

运维管理

质量报告

运维管理

执行实例与结果

质量任务

告警信息

数据源hive_emr-b2o数据库default执行时间

批量导出数据查看导出记录

表名(执行ID)

aa

ID: 128

表负责人

执行时间

2024-05-17 15:57:31

执行方式

试运行

规则名称

不为空

规则类型

自定义模版

监控对象

当前表: aa, 字段: id(bigint)

检测范围

-

关联表: aa, 字段: -

aa

ID: 128

表负责人

执行时间

2024-05-17 15:49:03

执行方式

试运行

规则名称

不为空

规则类型

自定义模版

监控对象

当前表: aa, 字段: id(bigint),id(bigint),id(bigint)

检测范围

-

关联表: aa, 字段: id(bigint),id(bigint)

aa

ID: 128

表负责人

执行时间

2024-05-17 15:48:56

执行方式

试运行

结果 & 日志

结果

日志

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

[2024-05-17 15:57:34]-[INFO] the rule-sql list : [{"ruleExecId":35994(

[2024-05-17 15:57:34]-[INFO] ===== START<2024-05-17 15:

[2024-05-17 15:57:34]-[INFO] tion.

[2024-05-17 15:57:34]-[INFO] om security sc

[2024-05-17 15:57:35]-[INFO] ccess, user: h

[2024-05-17 15:57:35]-[INFO] onType: ldap

[2024-05-17 15:57:35]-[INFO] jdbcUrl: jdbc:

[2024-05-17 15:57:35]-[INFO] ss.

[2024-05-17 15:57:35]-[INFO] 3591, curRunDe

[2024-05-17 15:57:35]-[INFO]

[2024-05-17 15:57:35]-[INFO] -service. rule

[2024-05-17 15:57:35]-[INFO] p://172.16.0.2

[2024-05-17 15:57:35]-[INFO] :{"statusCode"

[2024-05-17 15:57:35]-[INFO] succeed.

[2024-05-17 15:57:41]-[INFO]

[2024-05-17 15:57:49]-[INFO] ct * from (sel

[2024-05-17 15:57:49]-[INFO] *****