

数据开发治理平台 WeData

最佳实践



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的内容。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

最佳实践

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

最佳实践

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](#)。

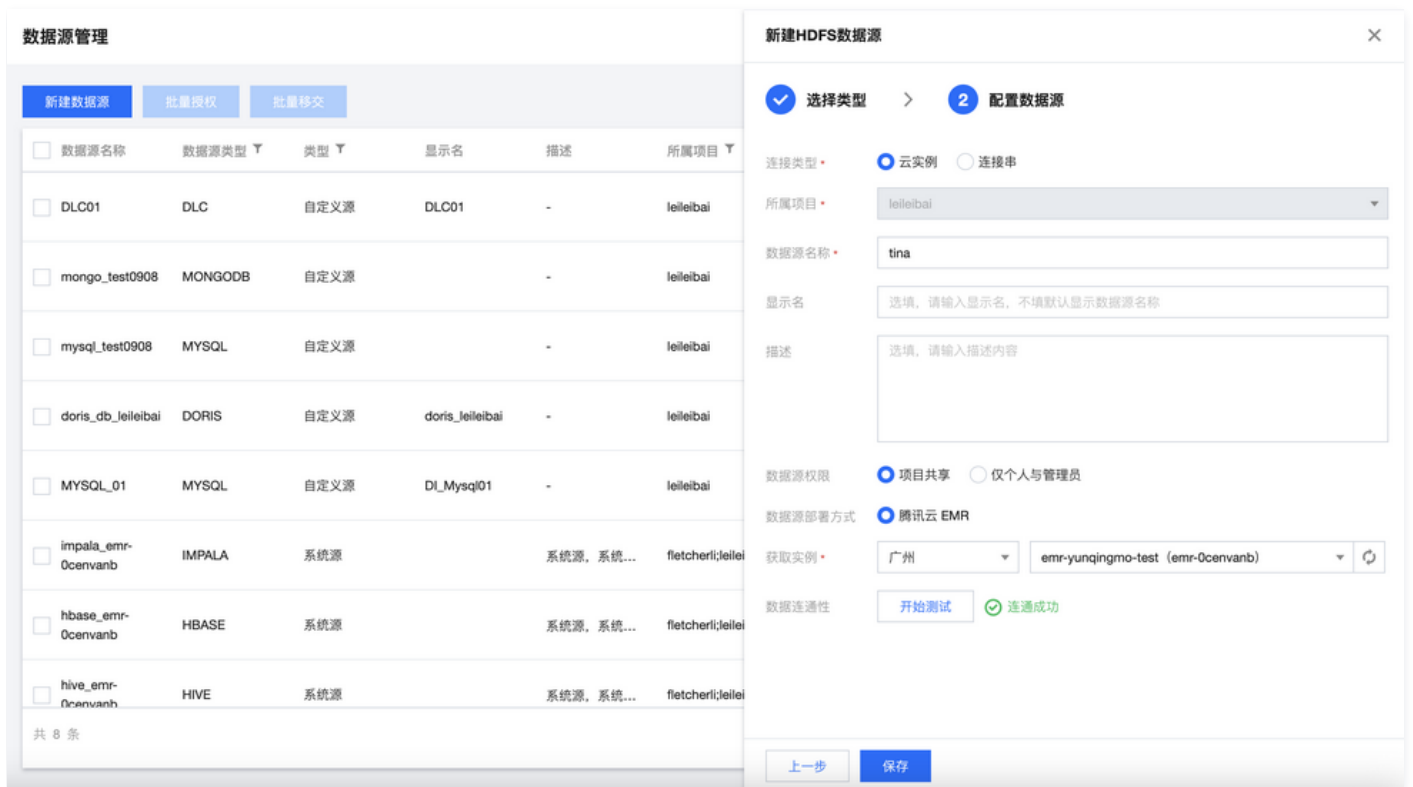
步骤二 初始化环境

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。

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

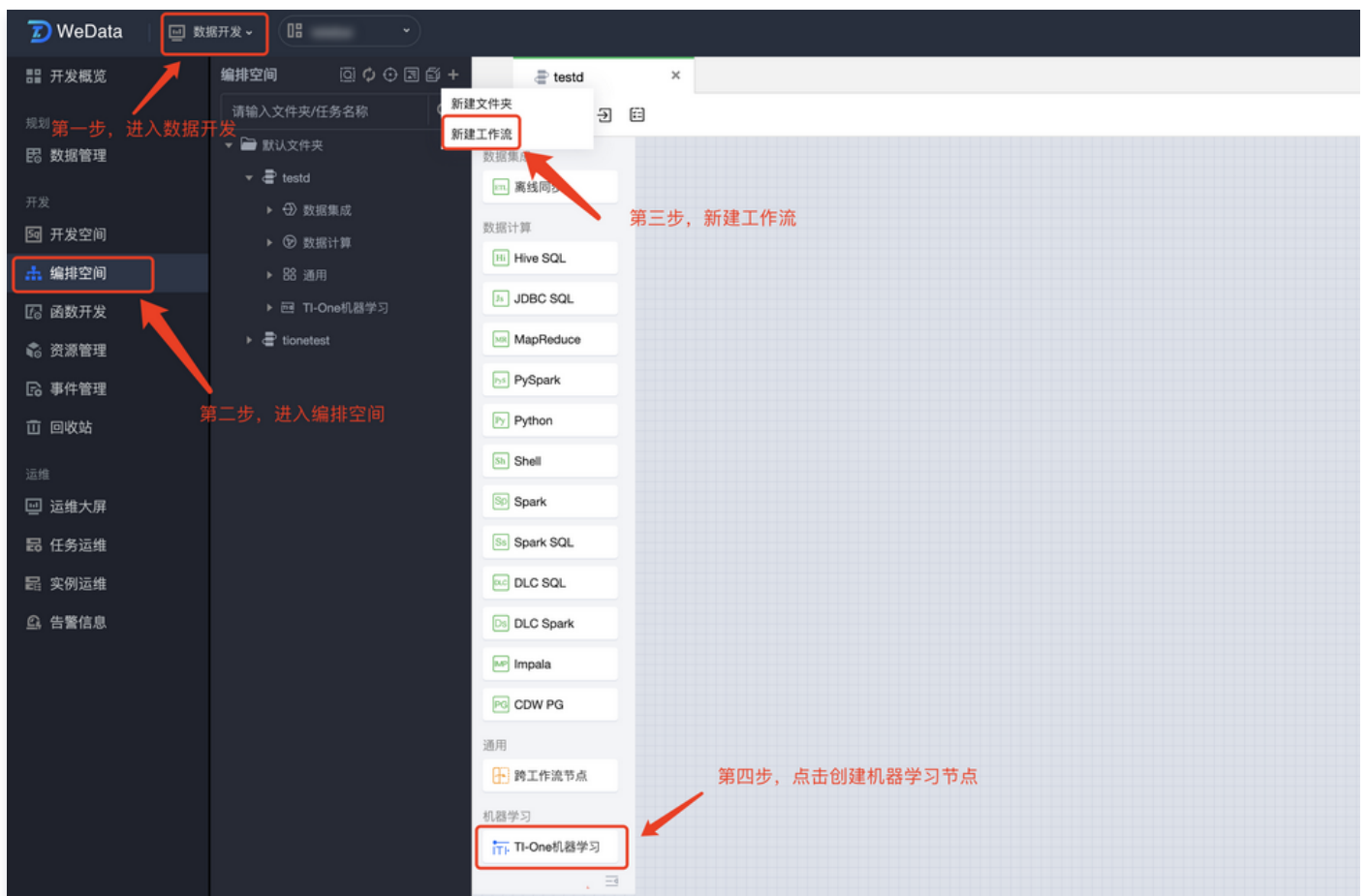
步骤三 添加数据源

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](#)。



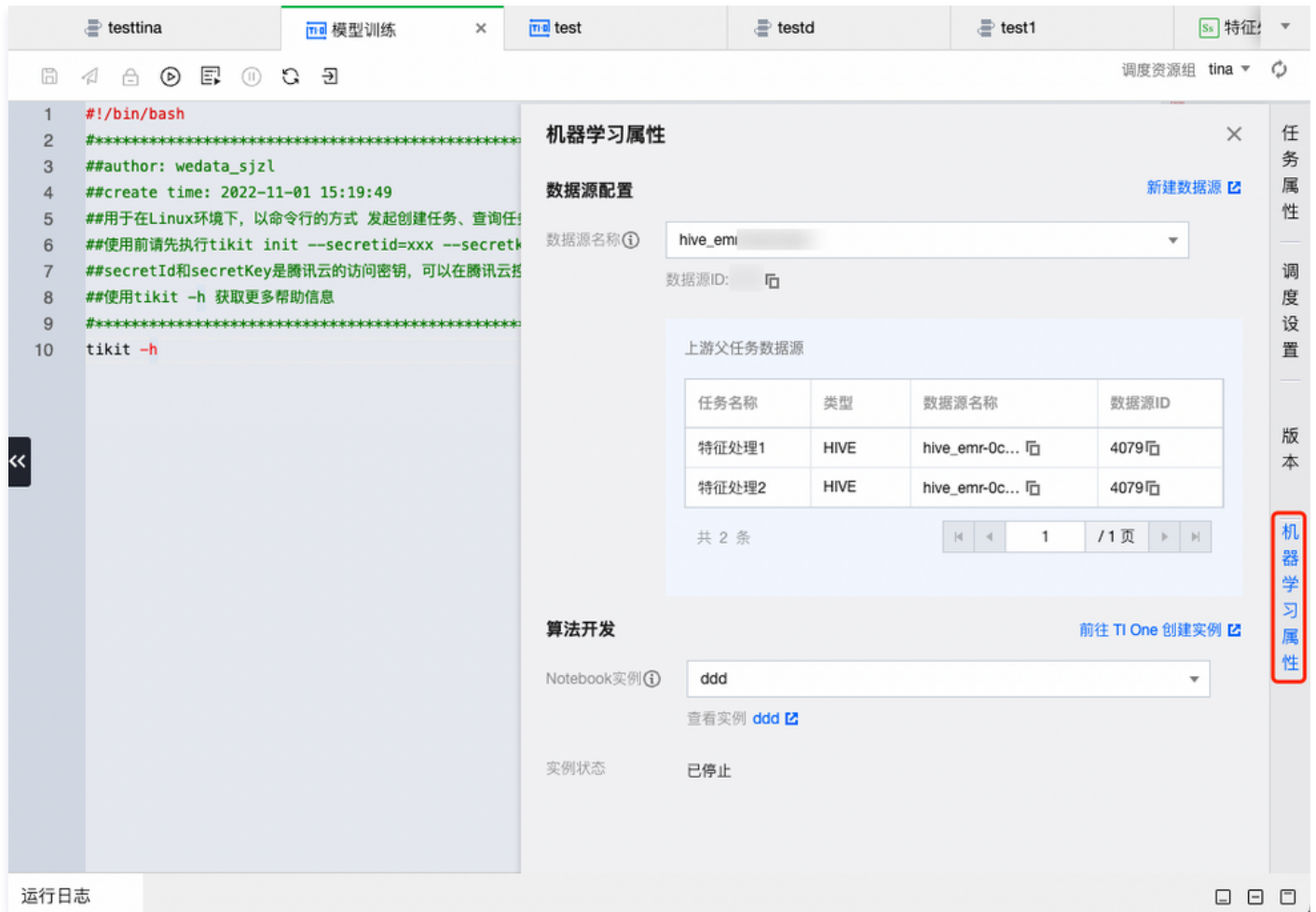
步骤四 机器学习节点配置

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。





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。



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](#)。

The screenshot displays the WeData interface with a workflow editor on the left and a '统一调度' (Unified Scheduling) configuration panel on the right. The workflow editor shows a sequence of tasks: '特征处理2' (Feature Processing 2) followed by '模型训练' (Model Training). The '统一调度' panel is currently open, showing configuration options for the selected task.

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

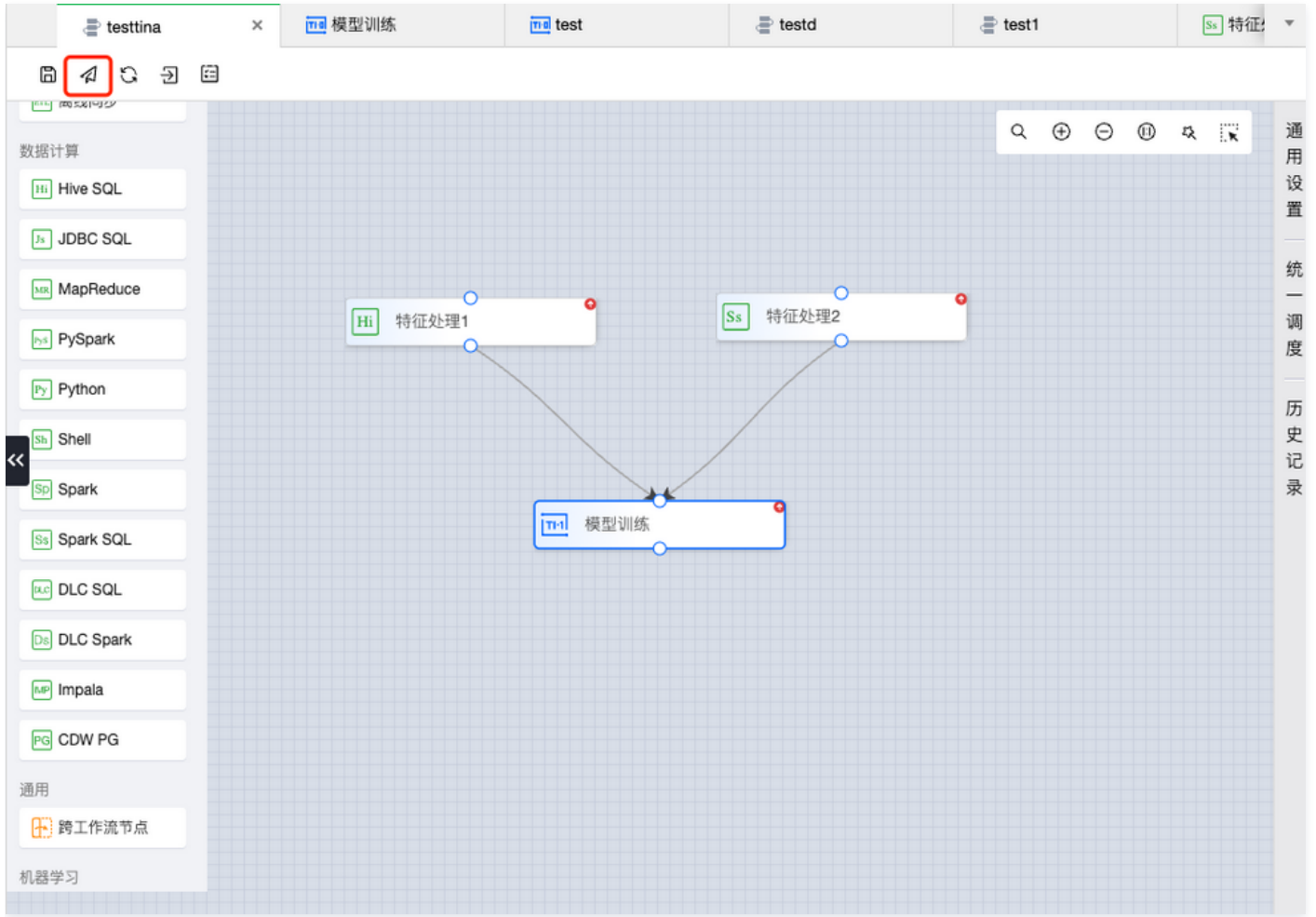
执行时间 00:00

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

通用设置 **统一调度** 历史记录



数据质量自定义模板使用

最近更新時間：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。

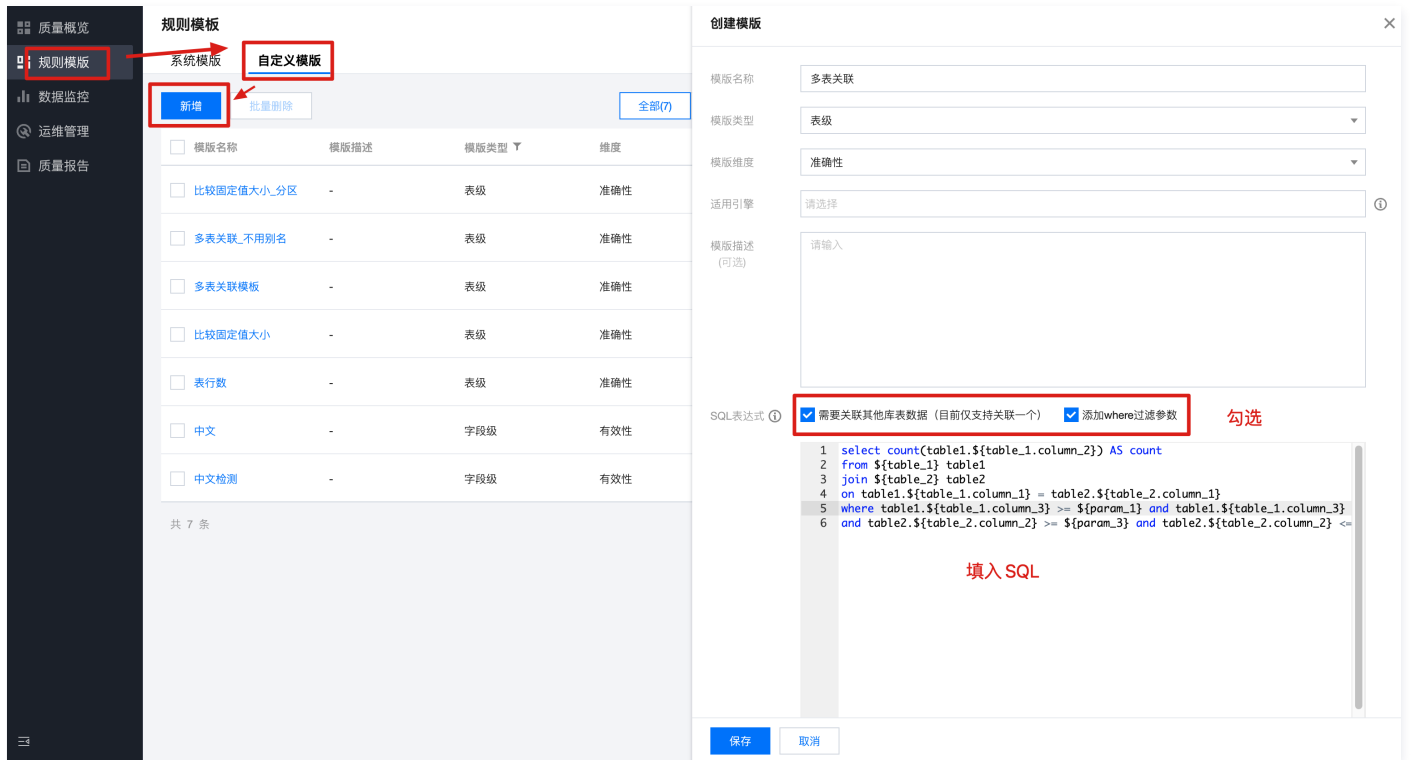
SQL 表达式：

```
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：

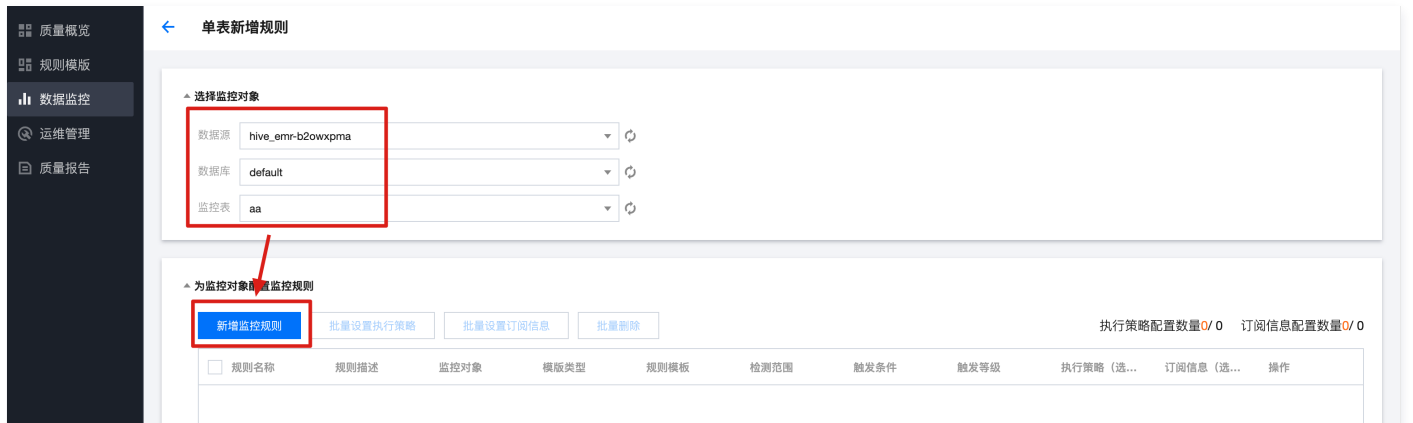
- 上文中一共出现了两张表： $\${table_1}$ 和 $\${table_2}$ ，
 - $\${table_1}$ 表示监控规则扫描的主表；
 - $\${table_2}$ 表示同数据源同数据库下的其他表（实际使用时也可以选择主表自己）；
- 使用了表1四个字段，分别为：
 - $\${table_1.column_1}$ ：用于与表2关联；
 - $\${table_1.column_2}$ ：用于结果计数；
 - $\${table_1.column_3}$ ：用于过滤条件，大于等于参数1，小于等于参数2；
 - $\${table_1.column_4}$ ：表示表1的分区字段，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:



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。



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 自定义模版

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

表1以及表1的四个参数

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

param_1	1
param_2	100
param_3	1
param_4	100

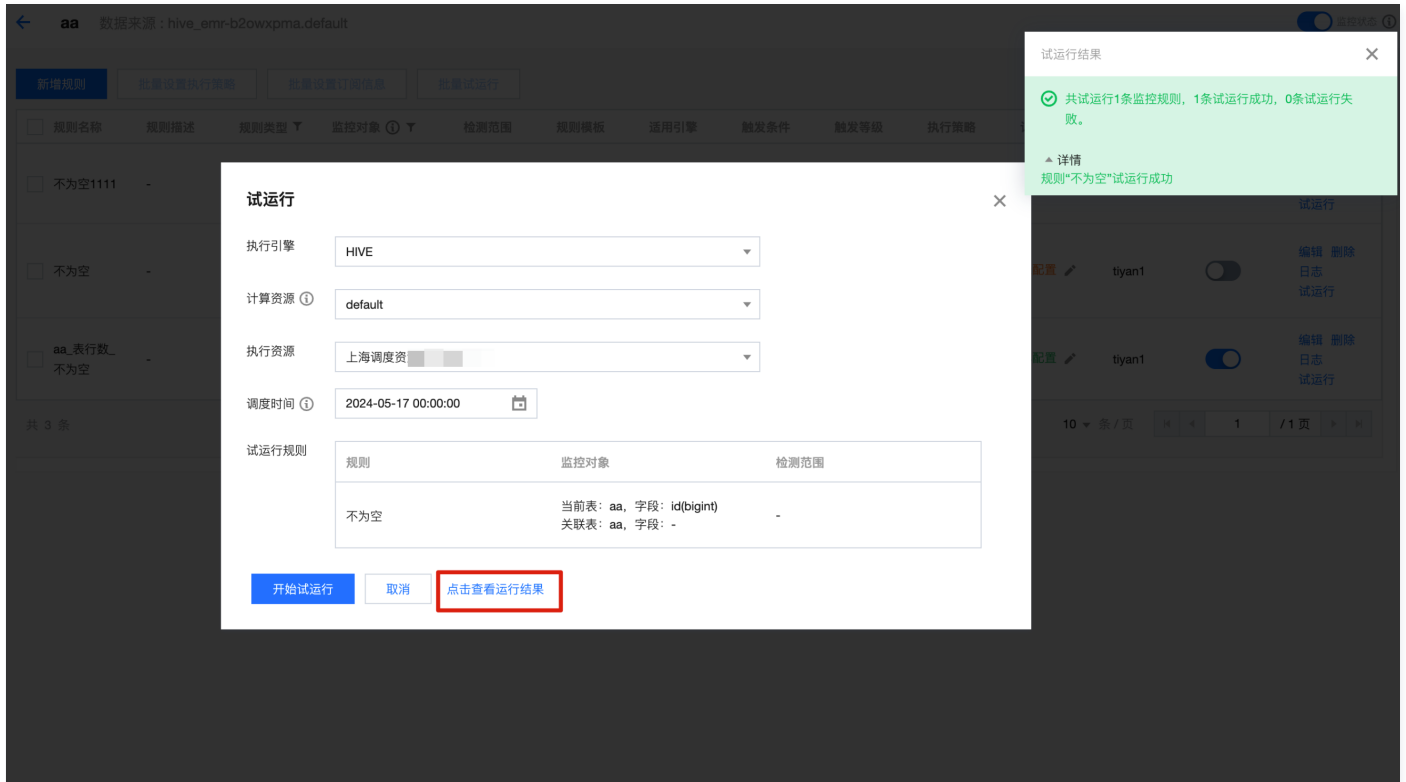
触发条件 等于 1 添加

注意:

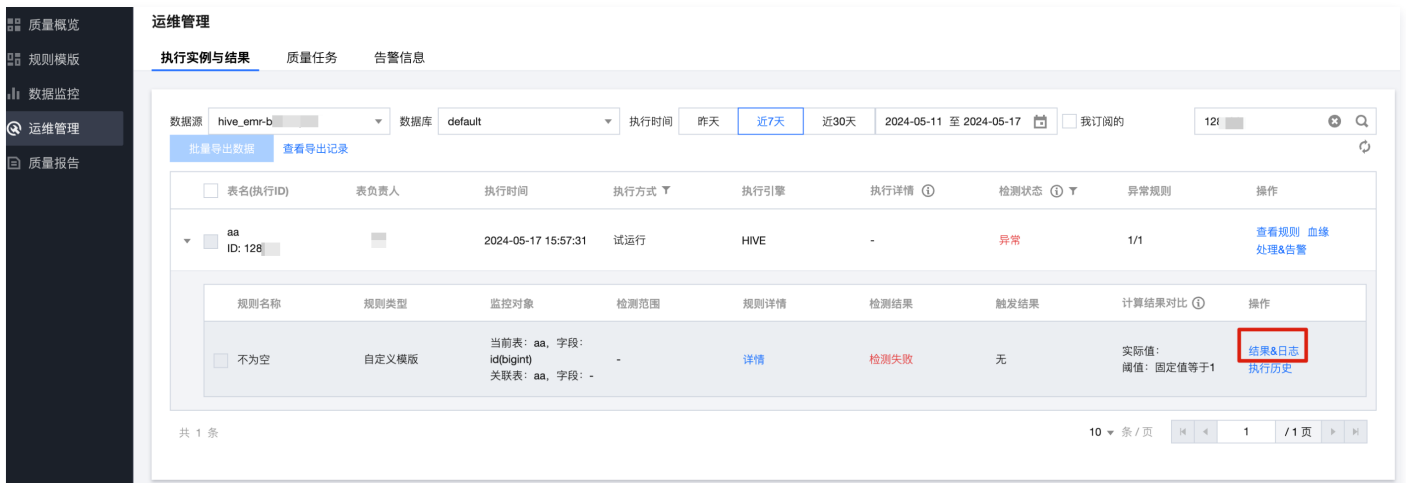
使用自定义模板前请先分析每个字段是什么含义，再进行映射。

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: [按钮]

表名(执行ID)	表负责人	执行时间	执行方式
aa ID: 128		2024-05-17 15:57:31	试运行

规则名称	规则类型	监控对象	检测范围
不为空	自定义模版	当前表: aa, 字段: id(bigint) 关联表: aa, 字段: -	-

规则名称	规则类型	监控对象	检测范围
不为空	自定义模版	当前表: aa, 字段: id(bigint),id(bigint),id(bigint) 关联表: aa, 字段: id(bigint),id(bigint)	-

结果&日志

结果 **日志**

```
1 [2024-05-17 15:57:34]-[INFO] the rule-sql list : [{"ruleExecId":35994(
2
3 [2024-05-17 15:57:34]-[INFO] ===== START<2024-05-17 15:
4
5 [2024-05-17 15:57:34]-[INFO] .....
6
7 [2024-05-17 15:57:34]-[INFO] .....om security sc
8
9 [2024-05-17 15:57:35]-[INFO] .....ccess. user: h
10
11 [2024-05-17 15:57:35]-[INFO] .....onType: ldap
12
13 [2024-05-17 15:57:35]-[INFO] .....dbcUrl: jdbc:h
14
15 [2024-05-17 15:57:35]-[INFO] .....ss.
16
17 [2024-05-17 15:57:35]-[INFO] .....3591, curRunDe
18
19 [2024-05-17 15:57:35]-[INFO] .....
20
21 [2024-05-17 15:57:35]-[INFO] .....-service. rule
22
23 [2024-05-17 15:57:35]-[INFO] .....p://172.16.0.2
24
25 [2024-05-17 15:57:35]-[INFO] .....:{"statusCode'
26
27 [2024-05-17 15:57:35]-[INFO] .....succeed.
28
29 [2024-05-17 15:57:41]-[INFO] .....ct * from (sel
30
31 [2024-05-17 15:57:49]-[INFO] .....
32
33 [2024-05-17 15:57:49]-[INFO] .....*****
34
```