WeData Data Development Platform Practice Tutorial

Copyright Notice

Trademark Notice

Service Notice

This document provides an overview of the as–is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre–sales consultation and technical after–sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Practice Tutorial

Integrating WeData Platform to Execute Periodic Scheduling Tasks

Last updated：2025-04-09 14:29:15

Background

This document introduces how to collocate Tencent Cloud Data Development and Governance Platform (Wedata) and TI-kit CLI tool to perform periodic scheduling of tasks. This feature is applicable to business scenarios that have both data governance needs and machine learning needs. You can perform unified scheduling of data development tasks and machine learning tasks on the Wedata page.

Operation Process

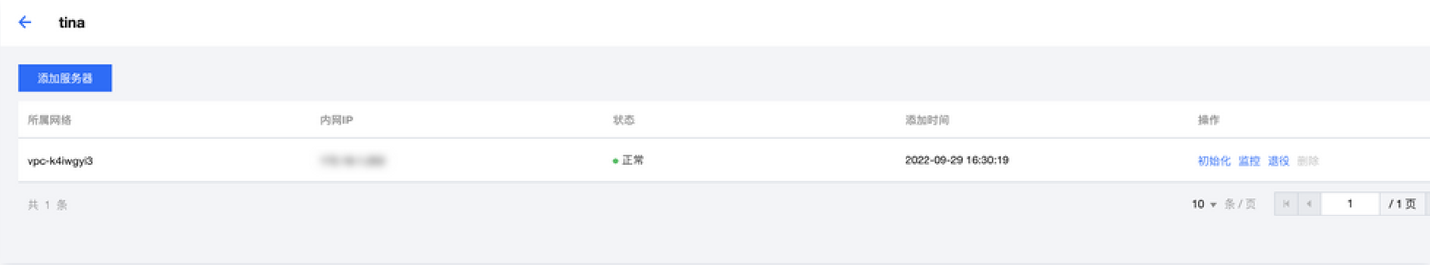
You can follow the following practice process to integrate TI-ONE and Wedata platform (WeData, a one-stop data development and governance platform located in the cloud. For details, see Overview). Currently, TI-ONE's machine learning task scheduling capability only supports the enterprise edition of Wedata in Guangzhou region.

Step 1 Preparations

1. Create a user and a project
   Within the Wedata product, you need to first create a user and a project. For details, see Project Creation and Member and Role Management.
2. Configure a custom scheduling resource group
   To enable the TI-ONE integration feature, you need to configure the enterprise edition custom scheduling resource group first. For details, see List of Custom Scheduling Resource Groups.

Step Two Initialize Environment

Add an execution resource group in the Wedata project space. After adding a server, you need to install Wedata Agent and TI-ONE CLI machine learning environment simultaneously. After logging in to the machine and completing the installation, you can see that the status of the resource group node is normal.



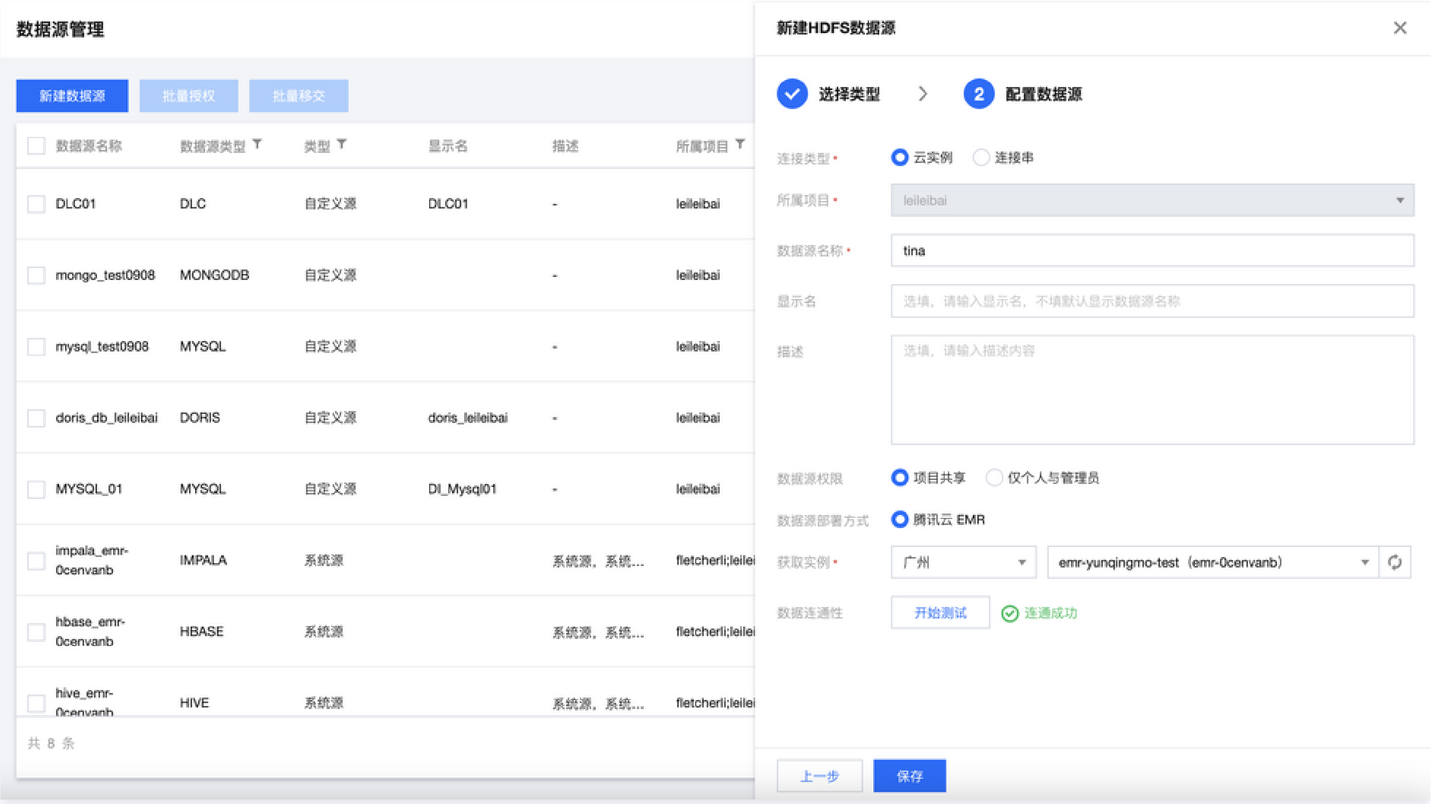
tina

所属网络	内网IP	状态	添加时间	操作
vpc-k4iwgyi3	[redacted]	● 正常	2022-09-29 16:30:19	初始化 监控 退役 删除

共 1 条

Step Three Add Data Source

Add a data source in the Wedata project space. Add an HDFS or HIVE data source. Test connectivity. Pay attention that after creation, you need to authorize it to the projects that need to use it. Please refer to Data Source Management for data source creation.



数据源管理

□	数据源名称	数据源类型	类型	显示名	描述	所属项目
□	DLC01	DLC	自定义源	DLC01	-	leileibai
□	mongo_test0908	MONGODB	自定义源		-	leileibai
□	mysql_test0908	MYSQL	自定义源		-	leileibai
□	doris_db_leileibai	DORIS	自定义源	doris_leileibai	-	leileibai
□	MYSQL_01	MYSQL	自定义源	DI_Mysql01	-	leileibai
□	impala_emr-0cenvanb	IMPALA	系统源		系统源，系统...	fletcherli;leilei
□	hbase_emr-0cenvanb	HBASE	系统源		系统源，系统...	fletcherli;leilei
□	hive_emr-0cenvanb	HIVE	系统源		系统源，系统...	fletcherli;leilei

共 8 条

新建HDFS数据源

✓ 选择类型 > 2 配置数据源

连接类型 * ● 云实例 ○ 连接串

所属项目 * leileibai

数据源名称 * tina

显示名 选填，请输入显示名，不填默认显示数据源名称

描述 选填，请输入描述内容

数据源权限 ● 项目共享 ○ 仅个人与管理员

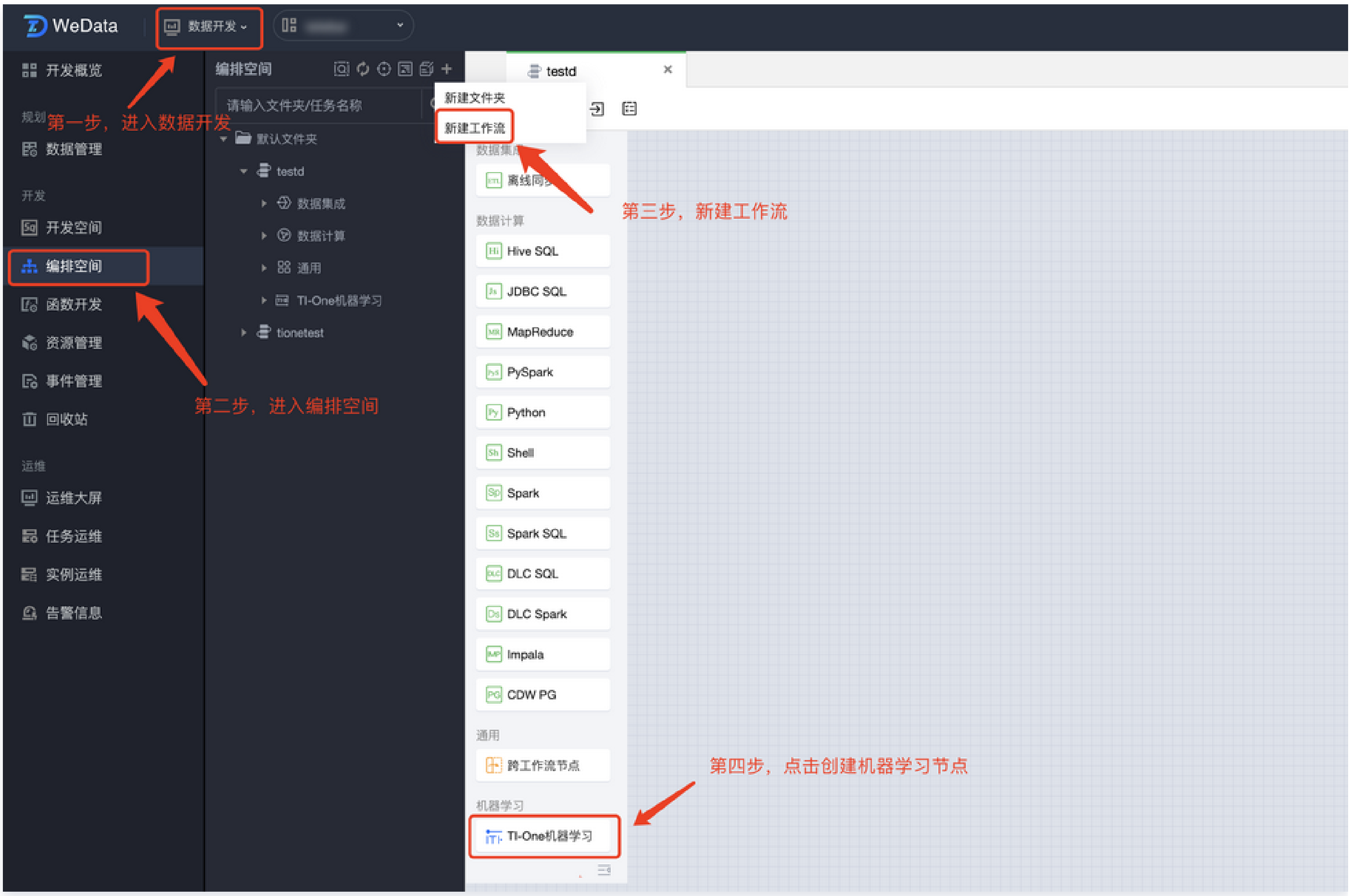
数据源部署方式 ● 腾讯云 EMR

获取实例 * 广州 emr-yunqingmo-test（emr-0cenvanb）

数据连通性 ✓ 连通成功

Step Four Machine Learning Node Configuration

1. Enter **Data Development** > **Orchestration Space**, create a workflow. In the workflow orchestration panel, click to create a **TI-ONE Machine** learning node.

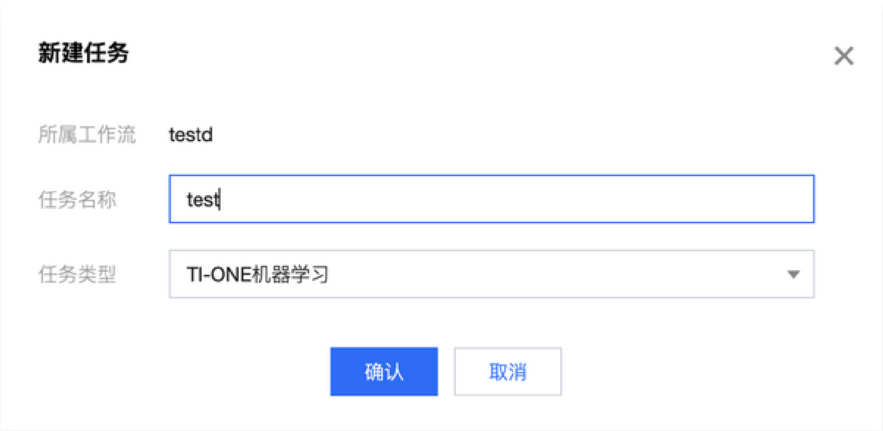


第一步，进入数据开发

第二步，进入编排空间

第三步，新建工作流

第四步，点击创建机器学习节点



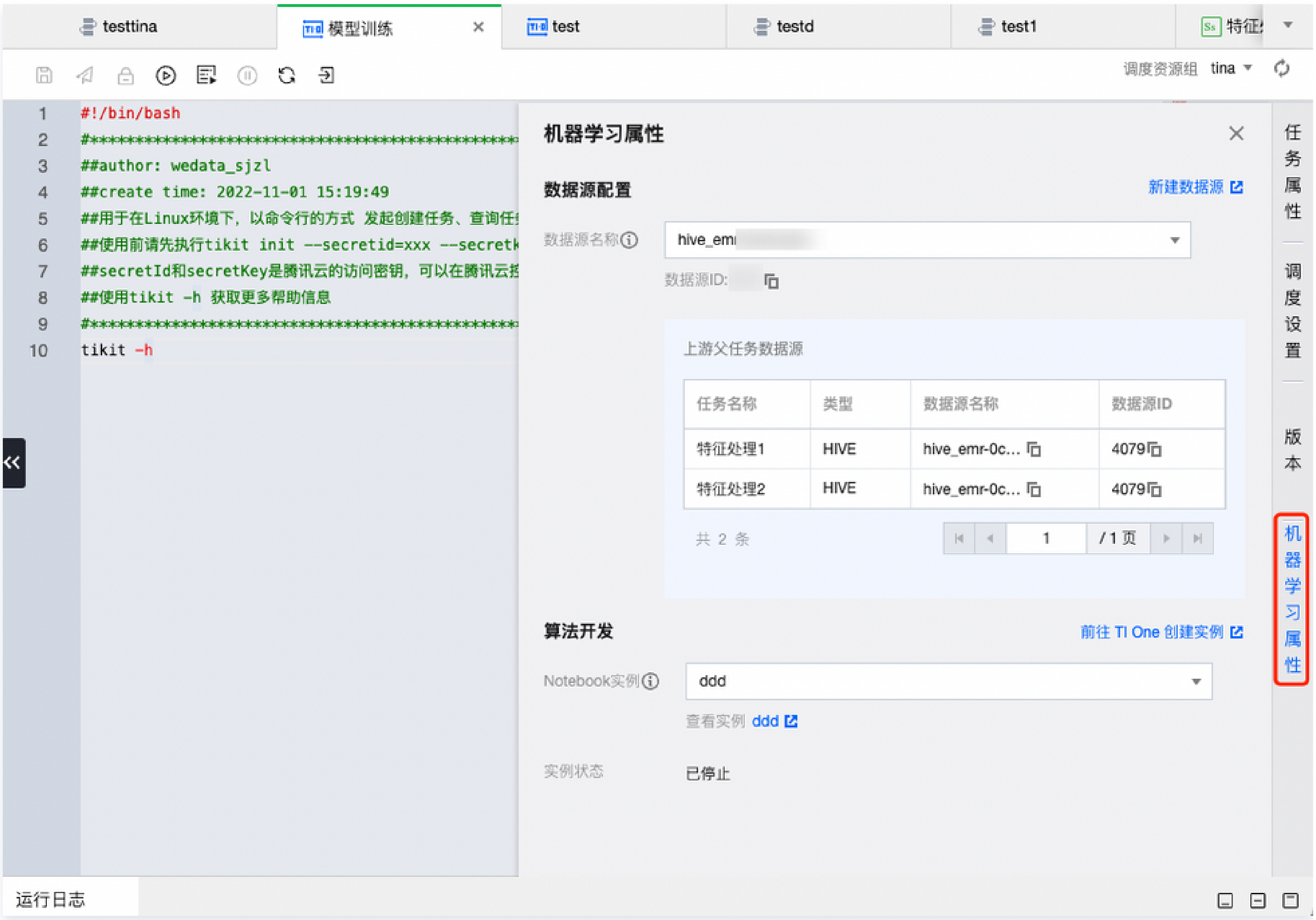
新建任务

所属工作流 testd

任务名称 test

任务类型 TI-ONE机器学习

2. The Machine Learning Node in Wedata is essentially a Shell node with the Tikit execution environment for machine learning tasks installed. Users need to write Tikit commands in this node to schedule TIONE computing power for submitting training tasks.
3. After entering the node configuration page, click **Machine Learning Attribute** to configure the data source and algorithm development. Among them, for data source configuration, you can drop down to select the data source associated with the current training task (if the Machine Learning Node is upstream connected to other nodes, the upstream parent task data source can be displayed below). After dropping down, the data source ID will be displayed, which can be used for script development and training task submission.
4. Before submitting a training task, we need to prepare the training code. TIONE provides a lightweight and convenient interactive development environment, Notebook. You can click on the right to enter TIONE Notebook for code writing. (After navigating to the TIONE Notebook instance creation page, the network information of the selected data source will be carried by default. If the data source is HDFS, it will also be selected by default in the data catalog.) If the current machine learning task is associated with a certain Notebook instance, it can be directly dropped down to select. The page will display a quick jump link and the instance running status.



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

机器学习属性

数据源配置　新建数据源

数据源名称：hive_emr[redacted]

数据源ID: [redacted]

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　1　/ 1 页

算法开发　前往 TI One 创建实例

Notebook实例：ddd

查看实例 ddd

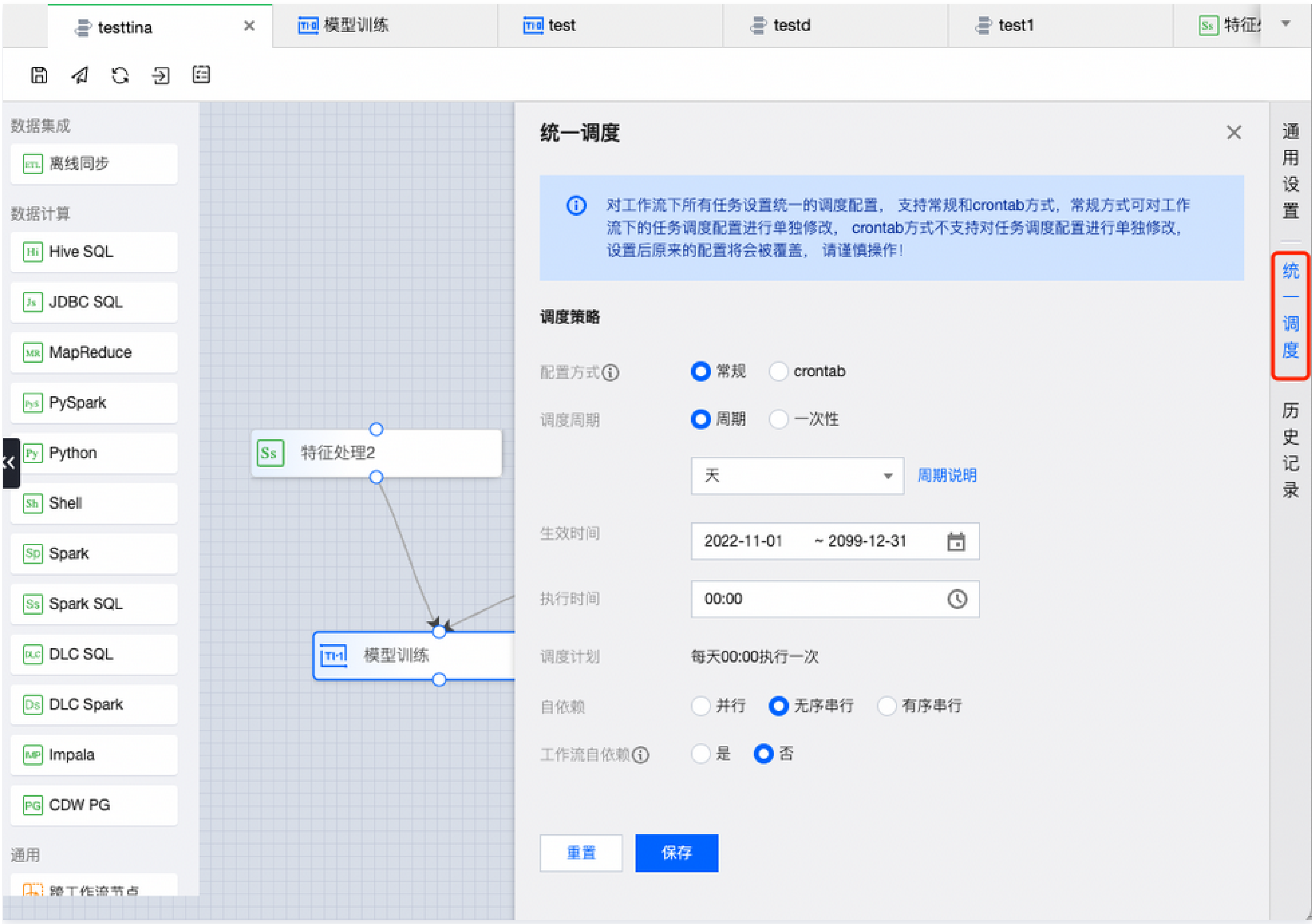
实例状态：已停止

Step 5 Write a Training Task Submission Command Using TICLI

1. After entering the Machine Learning Node, execute `tikit init --secretid=xxx --secretkey=xxx` for initialization before use. secretId and secretKey are the access keys of Tencent Cloud. Method for obtaining: Enter the console, click on the avatar in the upper right corner, and enter **CAM** > **API Key Management** to obtain.
2. Before using, input tikit –h to get the running modes of each command of the tikit CLI tool.
3. Submit tasks according to the currently required task type. Command testing can be run on the current shell node. After task submission, the corresponding TI-ONE task URL can be printed in the running log. You can go to the TI-ONE Console to view the training task details.

Step 6 Submit Workflow for Periodic Scheduling

After the workflow development is completed, you can configure the workflow periodic scheduling parameters and submit the overall workflow. After submission, you can view the workflow and tasks in the **Task Ops** module. Once the periodic instance is generated, you can view the instance details on the **Instance Ops** page. For detailed operation guide related to scheduling, please refer to Task Ops.



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：● 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

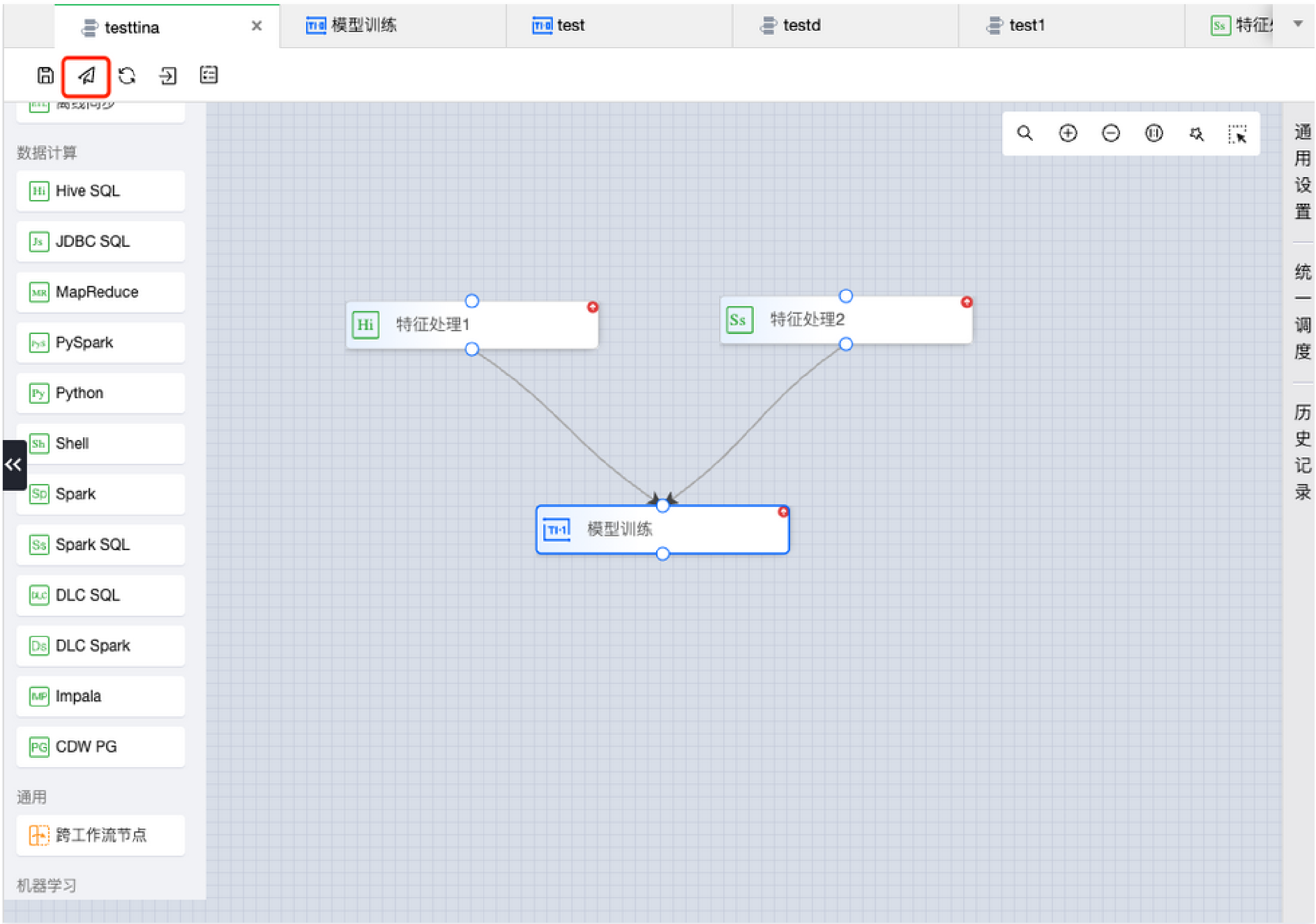
执行时间：00:00

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：○ 是 ● 否

重置 保存


testtina
模型训练
test
testd
test1
特征
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
机器学习
特征处理1
特征处理2
模型训练
通用设置
统一调度
历史记录

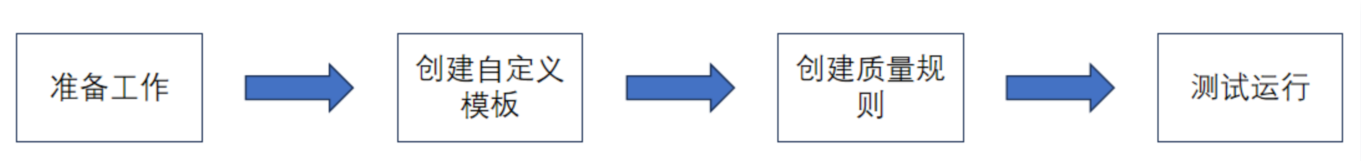
Using Custom Templates for Data Quality

Last updated：2025-04-09 14:30:09

Background

Tencent Cloud Data Development and Governance Platform Wedata's data quality supports creating custom templates and batch management, helping you customize table quality inspection logic based on business scenarios. This document introduces how to create rule templates through the custom template page and create detection rules for tables on the data monitoring page according to the custom rule templates.

Operation Process



Step 1 Preparations

1. Create a user and a project
   Within the Wedata product, you need to first create a user and a project. For detailed operation guide, check Preparations.
2. Create a scheduling resource group
3. Running quality inspection tasks requires creating a scheduling resource group. For detailed operation guide, check Scheduling Resource Group.

Step Two Create Custom Template

1. Enter **Data Quality** > **Rule Template**, click **Custom Template**, add a template and **save.**

SQL expression:

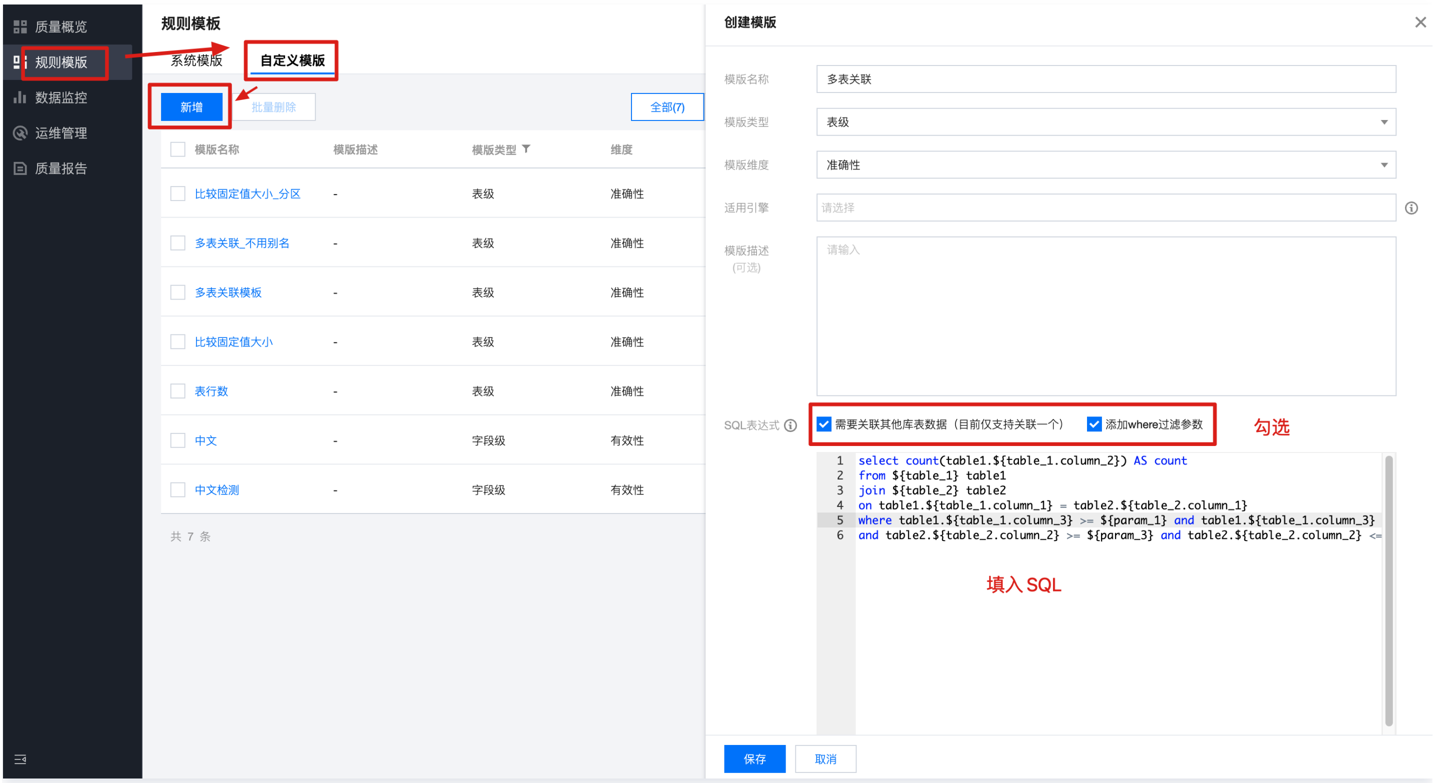
```
select count(table1.${table_1.column_2}) AS count
from ${table_1} table1
join ${table_2} table2
on table1.${table_1.column_1} = table2.${table_2.column_1}
where table1.${table_1.column_3} >= ${param_1} and table1.${table_1.column_3} <= ${param_2}
and table2.${table_2.column_2} >= ${param_3} and table2.${table_2.column_2}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

Explanation:

- Two tables appear in the previous context: ${table_1} and ${table_2}.
  - ${table_1} indicates the primary table scanned by the monitoring rule;
  - ${table_2} refers to other tables in the same data source and database (you can also choose the primary table itself in actual use);
- Four fields of Table 1 are used, respectively:
  - ${table_1.column_1}: used for association with Table 2;
  - ${table_1.column_2}: used for result counting;
  - ${table_1.column_3}: used for filtering conditions, greater than or equal to Parameter 1, less than or equal to Parameter 2;
  - ${table_1.column_4}: represents the partition field of Table 1, which can save computing resources significantly and avoid scanning full data;
- Two fields of Table 2 are used, respectively:
  - ${table_1.column_1}: used for association with Table 1;
  - ${table_1.column_2}: used for filtering conditions, greater than or equal to Parameter 3, less than or equal to Parameter 4;
- Used 4 where parameters, which are:

  - ○ ${param_1}: minimum value of Field 3 in Table 1 in SQL;
  - ○ ${param_2}: maximum value of Field 3 in Table 1 in SQL;
  - ○ ${param_3}: minimum value of Field 2 in Table 2 in SQL;
  - ○ ${param_4}: maximum value of Field 2 in Table 2 in SQL.
- Final calculation result: the count of eligible Field 2 in Table 1, a number.

Screenshot example:



□ 模版名称	模版描述	模版类型	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

模版名称: 多表关联

模版类型: 表级

模版维度: 准确性

适用引擎: 请选择

模版描述（可选）: 请输入

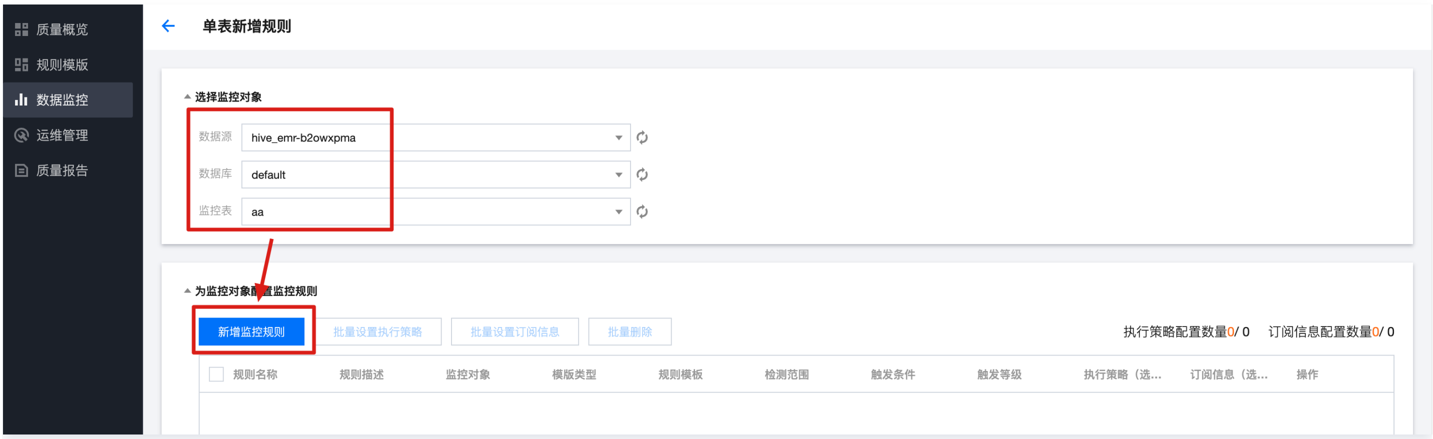
SQL表达式: ☑ 需要关联其他库表数据（目前仅支持关联一个） ☑ 添加where过滤参数 勾选

```
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}
and table2.${table_2.column_2} >= ${param_3} and table2.${table_2.column_2} <=
```

填入 SQL

Step 3 Create a Quality Rule

1. 1. Enter Data Monitoring, find the table to be monitored, and click Configure Monitoring Task.



选择监控对象

数据源: hive_emr-b2owxpma

数据库: default

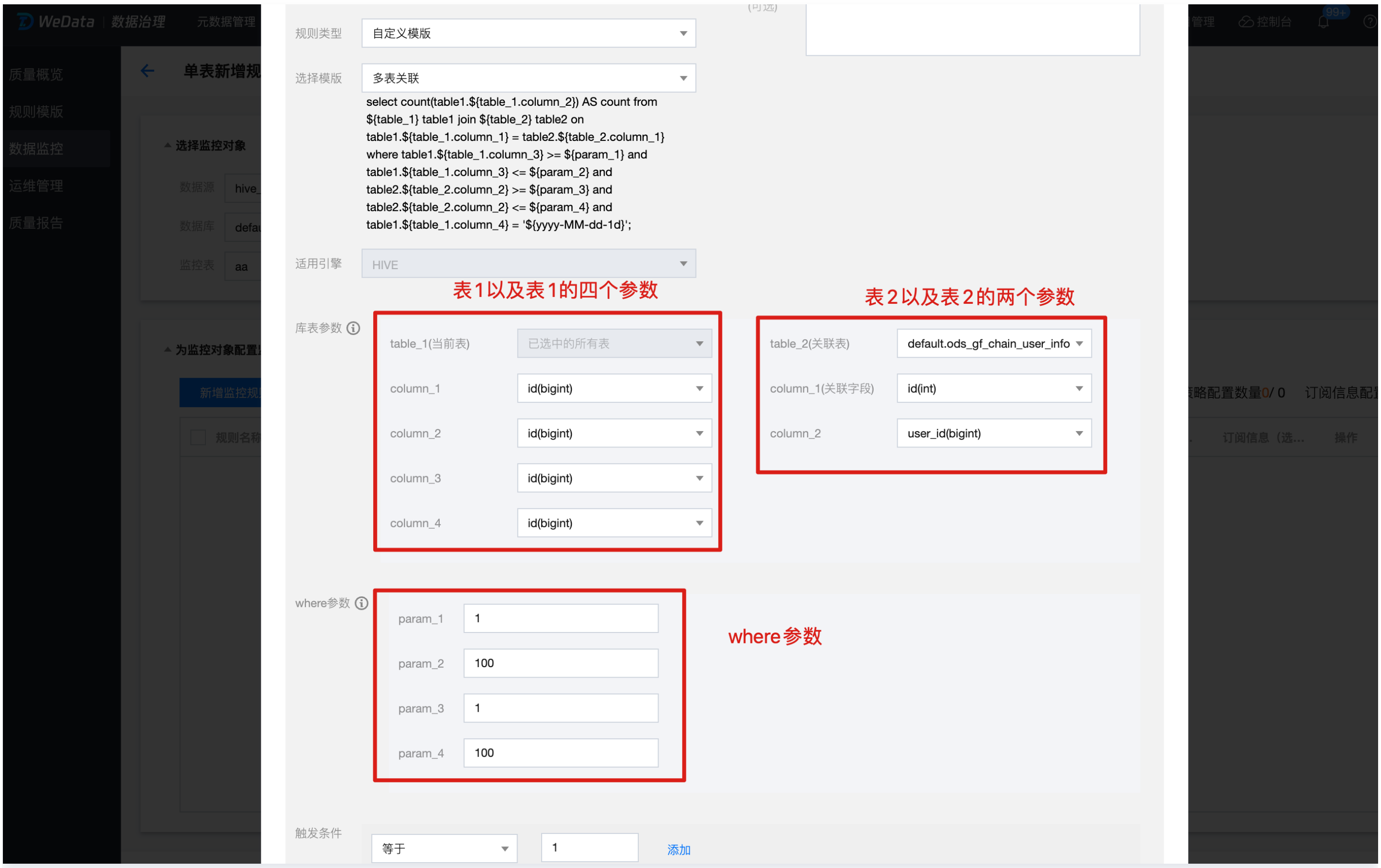
监控表: aa

为监控对象配置监控规则

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	模版类型	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. Click **Add Rule**, select Custom Template for the rule type, select the newly created template, choose database and table parameters and where parameter based on the template variables, configure the trigger conditions and level, and click **Save**.



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

库表参数

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

where参数

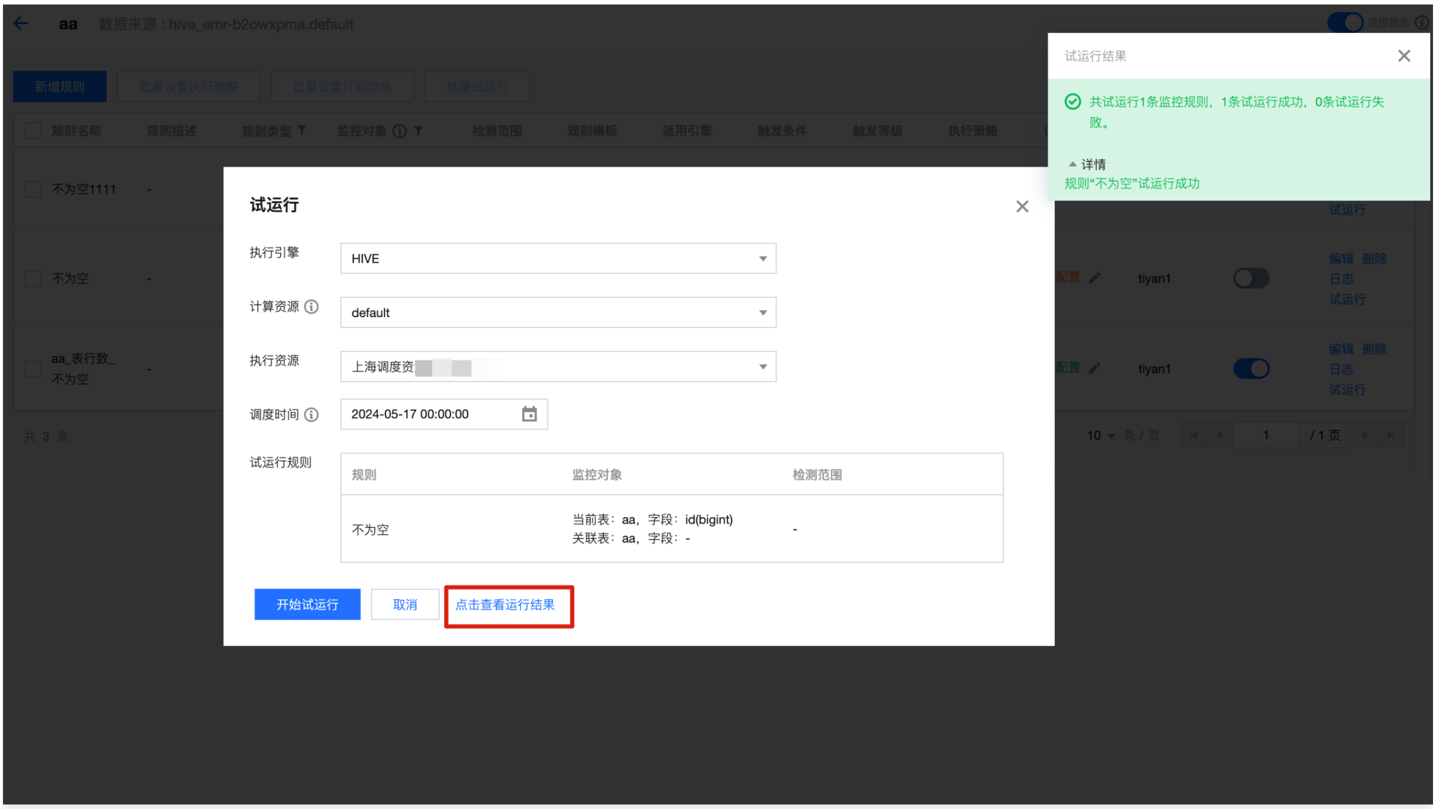
触发条件: 等于 1 添加

⚠ **Notes:**

Please first analyze what each field means before using a custom template and then map them.

Step 4 Test-Run

1. 1. Click **trial run**, select an execution engine, computational resource, and execution resource, and select the rule just created in the validation rules.



试运行结果

共试运行1条监控规则，1条试运行成功，0条试运行失败。

详情
规则"不为空"试运行成功

试运行

执行引擎: HIVE

计算资源: default

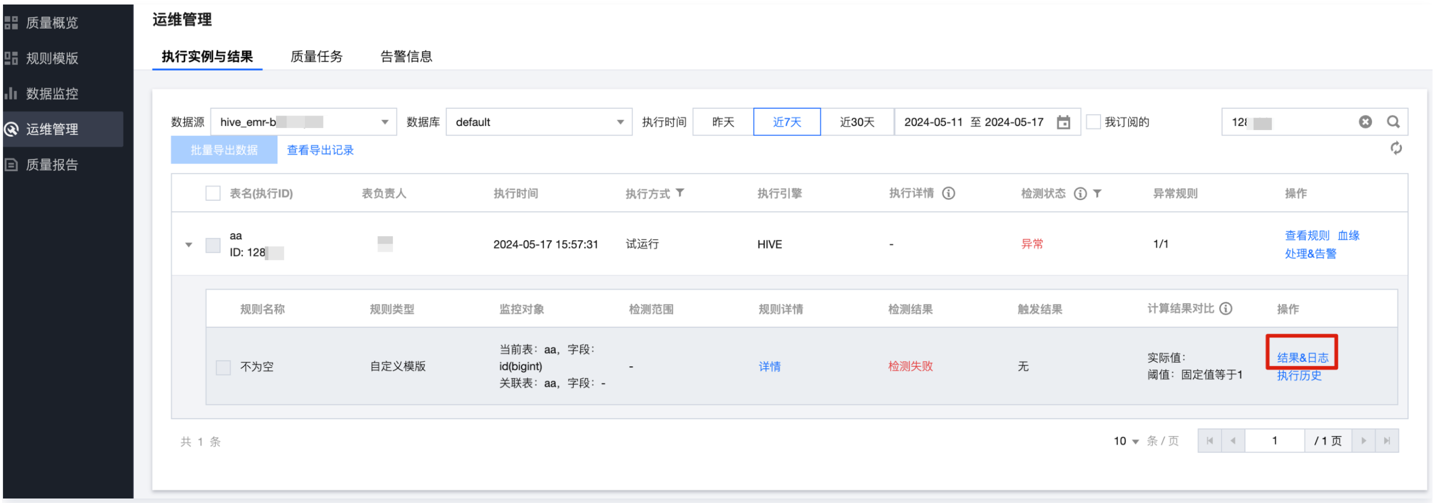
执行资源: 上海调度资[redacted]

调度时间: 2024-05-17 00:00:00

试运行规则

规则	监控对象	检测范围
不为空	当前表：aa，字段：id(bigint) 关联表：aa，字段：-	-

2. 1. Click **view execution results**, navigate to **Ops management** page to view execution results.



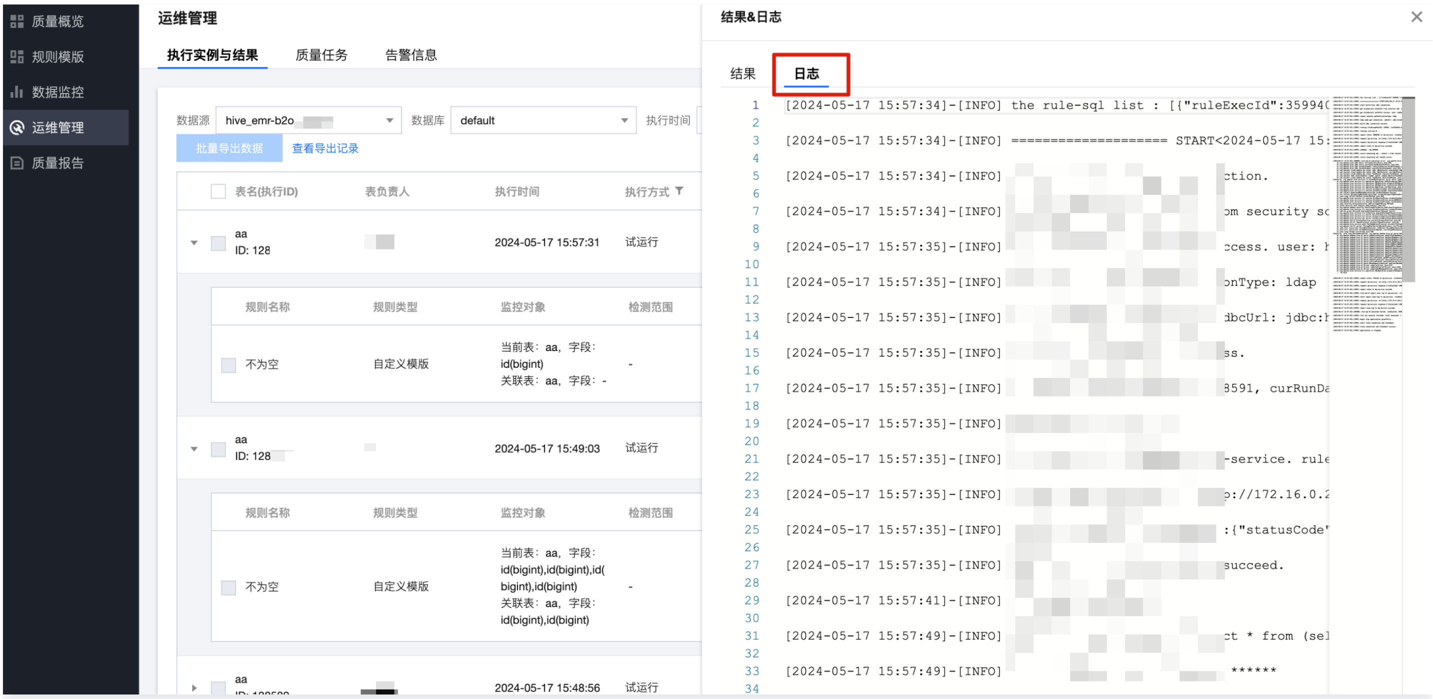
数据源: hive_emr-b[redacted] 数据库: default 执行时间: 近7天 2024-05-11 至 2024-05-17 □ 我订阅的 12[redacted]

表名(执行ID)	表负责人	执行时间	执行方式	执行引擎	执行详情	检测状态	异常规则	操作
aa ID: 128[redacted]	[redacted]	2024-05-17 15:57:31	试运行	HIVE	-	异常	1/1	查看规则 血缘 处理&告警

规则名称	规则类型	监控对象	检测范围	规则详情	检测结果	触发结果	计算结果对比	操作
不为空	自定义模版	当前表：aa，字段：id(bigint) 关联表：aa，字段：-	-	详情	检测失败	无	实际值： 阈值：固定值等于1	结果&日志 执行历史

3. 1. Click **Results & Logs** to view running logs.

Among them, EXECUTING SQL: xxxxxx prints the SQL submitted to the hive/spark/dlc engine for quality inspection.



运维管理

数据源 hive_emr-b2o[redacted] 数据库 default 执行时间

表名(执行ID)	表负责人	执行时间	执行方式
aa ID: 128[redacted]	[redacted]	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：-	-

表名(执行ID)	表负责人	执行时间	执行方式
aa ID: 128[redacted]	[redacted]	2024-05-17 15:49:03	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),id(bigint),id(bigint),id(bigint) 关联表：aa，字段：id(bigint),id(bigint)	-

表名(执行ID)	表负责人	执行时间	执行方式
aa ID: 1	[redacted]	2024-05-17 15:48:56	试运行

结果&日志

```
[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] [redacted] ction.

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

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

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

[2024-05-17 15:57:35]-[INFO] [redacted] dbcUrl: jdbc:h

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

[2024-05-17 15:57:35]-[INFO] [redacted] 8591, curRunDa

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

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

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

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

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

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

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

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

```