数据集成

实践教程

腾讯云

文档目录

实践教程
DLC 数据实时导入与小文件合并

最近更新时间：2025-06-04 10:24:12

业务场景

通过 DataInLong 数据集成将业务数据源实时导入至 DLC iceberg 表的过程中，伴随着实时同步过程的推进，目标系统端会不断生成小文件。对于目标系统内已生成的小文件，基于周期合并的方式可避免由于小文件的累积造成目标系统 DLC 引擎查询效率恶化。

操作场景

本文以 MySQL 实时同步至 DLC iceberg 表为例，介绍实时任务配置及小文件合并操作实践。

操作步骤

创建目标表

进入 DLC 控制台，根据以下语句创建 DLC 原生表（内表），DLC 内表默认为 iceberg 表。详细创建 DLC 原生表属性及数据优化，详情请参见 数据优化。

```
CREATE TABLE IF NOT EXISTS
 `db_name`.`new_table_name`(
`column_name1` column_type1,
`column_name2` column_type2
);
```

配置项目空间

> **说明：**
> 若您使用的是 WeData 产品，配置项目空间操作请参见 项目列表。

1. 进入 DataInLong 控制台，单击**项目列表 > 新建**，新建项目空间。

2. 您可以参考下表配置项目空间信息。

参数	说明
项目名称/标识	项目命名与唯一标识，其中唯一标识创建后不可修改。

高级设置 – 项目成员	为此创建的项目中添加其他项目成员，创建者默认加入项目空间。
成员角色	批量为项目成员配置角色（此处默认为前面添加的成员添加统一的角色，后续可项目管理模块修改）。

配置集成资源组

1. 进入 DataInLong 控制台 选择**集成资源**并单击**创建**，进入集成资源组购买页。

> **说明**
> 若您使用的是 WeData 产品，请点击进入 WeData 控制台 。

2. 购买集成资源组。

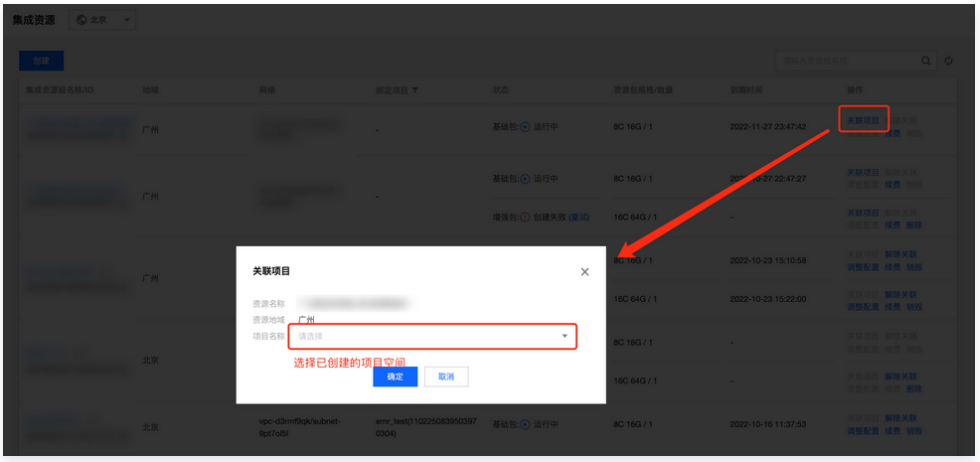
说明

- 离线资源包与实时资源包可根据实际数据情况配置规格、以及数量。
- 资源组网络建议选择 MySQL 和 DLC 所在网络；若 MySQL 和 DLC 不在一个 VPC 环境，可为 VPC 配置开通公网，详细操作参见 资源组配置公网 。

3. 购买完成后，返回控制台并关联资源组与项目空间。

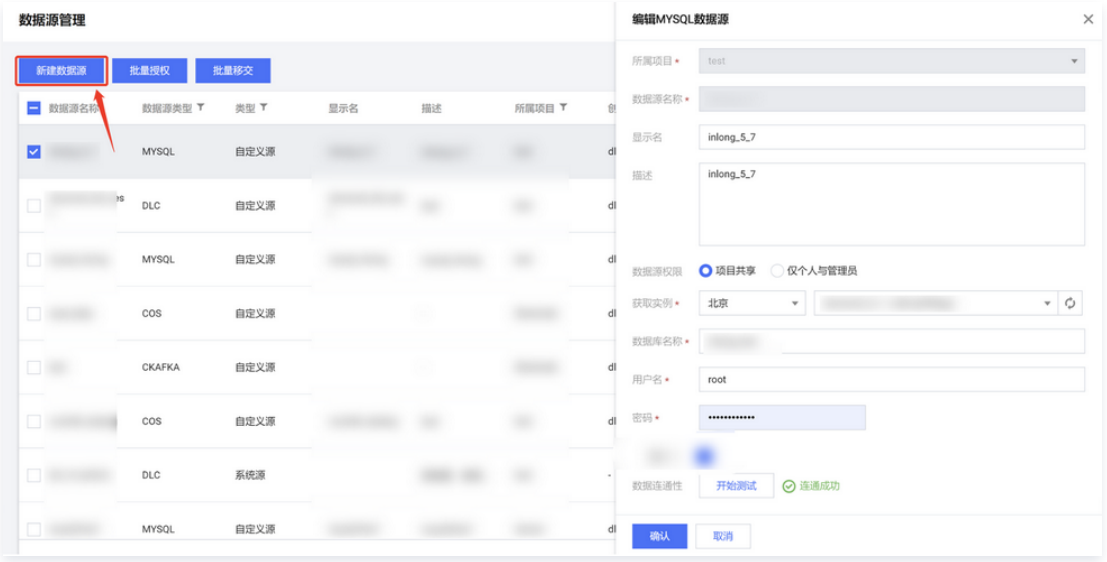
说明

若在购买页面内已经关联资源组与项目空间，可忽略此步骤。



配置数据源

1. 配置 MySQL 数据源。
   进入**项目管理**模块，选择**数据源管理 > 新建数据源 > 选择 MySQL**。以 MySQL 数据源为例，数据连通性测试成功后，单击**保存**。

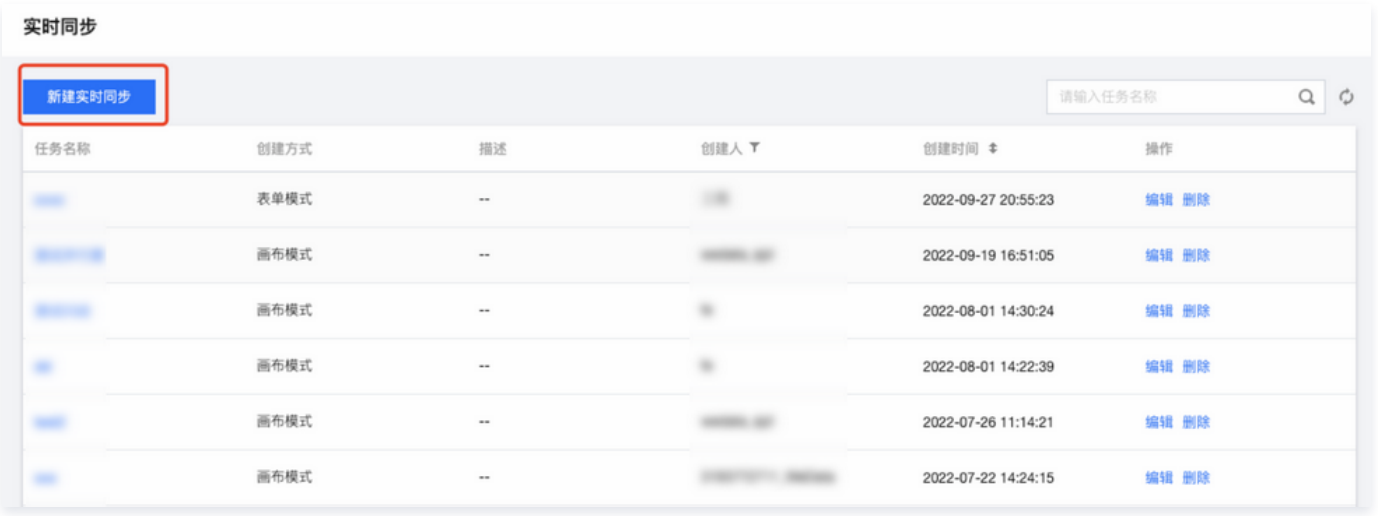


2. 配置 DLC 数据源。
   进入 **项目管理模块**，选择**数据源管理 > 新建数据源 > 选择 DLC**，配置数据源参数，并在连通性测试成功后，即可单击**保存**。

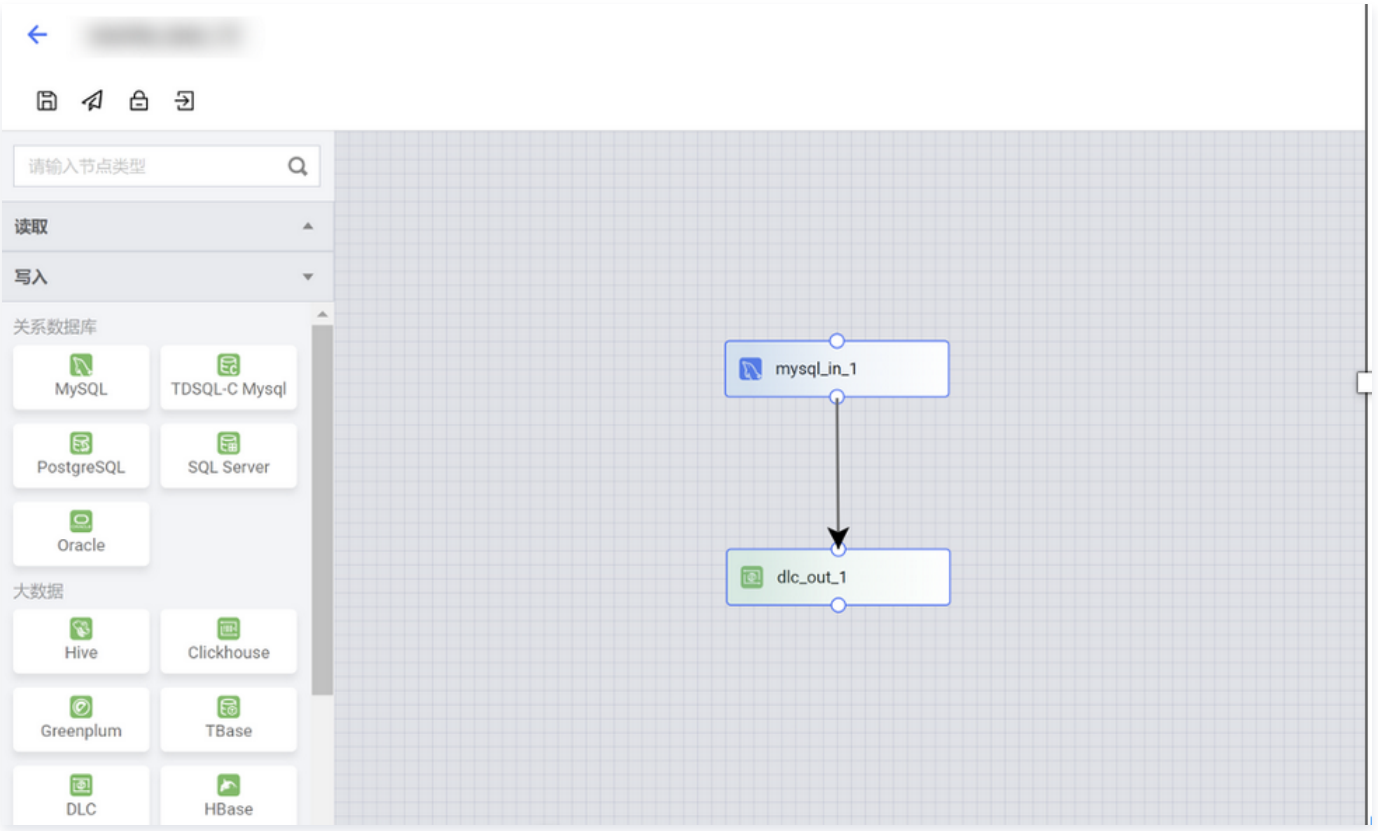
参数	说明
JDBC URL	格式参考： `jdbc:dlc:dlc.internal.tencentcloudapi.com?task_type=SparkSQLTask&database_name=&datasource_connection_name=DataLakeCatalog®ion=ap-beijing&data_engine_name=test_engine` 若需要使用小文件合并，数据源的访问域名必须使用 `dlc.internal.tencentcloudapi.com`，`task_type` 必须使用 `SparkSQLTask`，`data_engine_name` 指定的引擎会用于实时同步后的小文件合并。 **注意：** 小文件合并会使用此处配置的 DLC 数据引擎并在合并的时候占用部分资源，请合理配置资源，如果不启动小文件合并，该 DLC 引擎不会被使用。
secretId/secretKey	填写拥有引擎和 SQL 执行权限的账号或者子账号的密钥。可在 API 密钥管理 中查看。

配置实时同步任务

1. 创建任务。
   进入数据集成模块，创建**实时同步任务**，在弹出的提示框中输入任务名称和备注，选择 **画布模式**或**表单模式**，并单击**完成**。本介绍以画布模式为例。



2. 编辑任务。
   单击新建的实时同步任务名称，进入任务编辑界面，通过拖拽分别新建读取数据源和写入数据源，并通过连线指定数据流向。



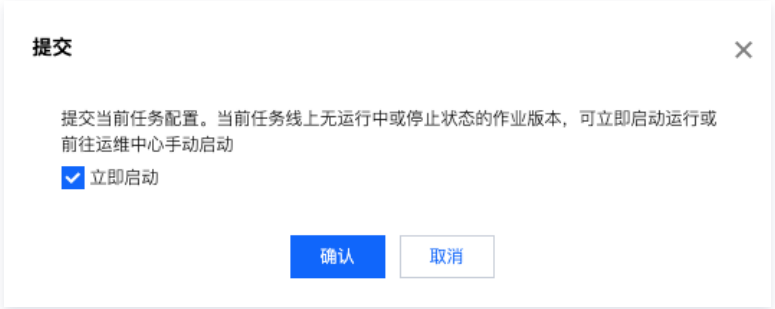
3. 配置 MySQL 节点。
   双击画布中的 MySQL 节点，对读取数据源进行配置。如下图选择需要同步的数据库表，读取模式选择**全量模式**，完成后单击**保存**。

4. 配置 DLC 节点。
双击画布中的 DLC 节点，对 DLC 写入数据源进行配置。如下图选中需要写入的库表，根据业务需求选择写入模式，并指定唯一键。例子中指定唯一键为 ID 和 MySQL 的主键保持一致。

下拉至底部，配置 MySQL 与 DLC 表字段映射，完成后单击**保存**。

5. 任务保存与提交。
   - 配置完节点后，单击**任务数据**配置集成资源组。此资源组为 配置集成资源组 步骤3中已关联至本空间的资源组。

   - 完成后，单击**提交**按钮，并在弹窗口中勾选**立即启动**。



6. 查看并运维实时任务。

- 提交任务后，可进入**实时运维**页面查看并监控任务状态。

- 单击**运行监控**，可查看当前任务数据指标统计、以及配置监控告警等。

存量任务处理

1. 如果存量实时同步任务需要添加小文件合并功能，首先需按照步骤一修改表属性。

> **说明**
> - ALTER TABLE `db_name`.`new_table_name` SET TBLPROPERTIES ('write.compact.enable' = 'true', 'write.compact.snapshot.interval' = '20');其中合并周期参数‘write.compact.snapshot.interval’需要根据业务需求进行调整。
> - 如果存量表已经存在大量的小文件，推荐手动将小文件合并到一定数量之下后，再启动定时合并功能。

2. 将实时同步任务**停止**再**运行**即可。

MySQL 分库分表同步至 Hive

最近更新时间：2024-07-09 22:01:41

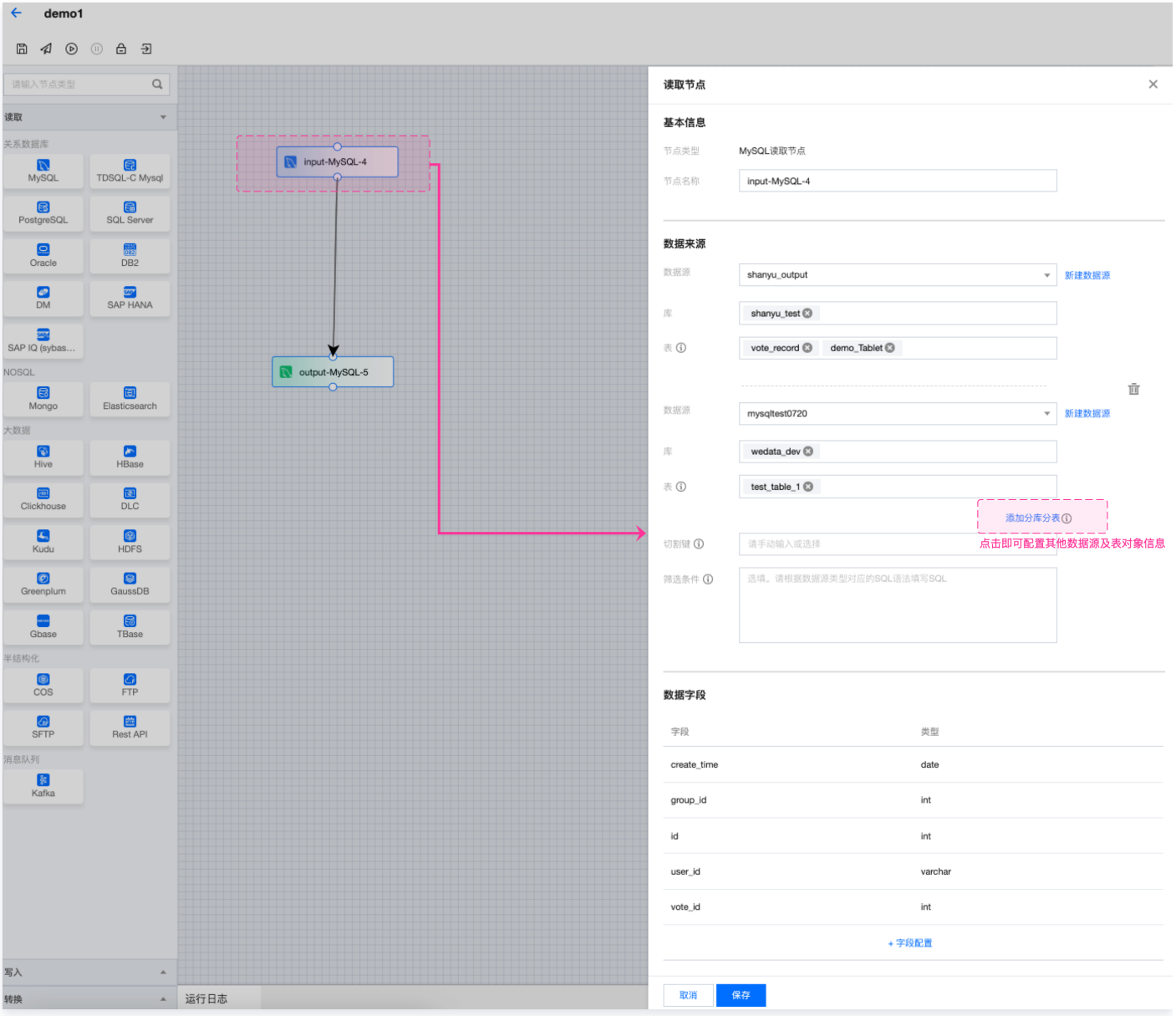
目前离线及实时数据同步任务支持 MySQL 分库分表同步至目标数据源中。本实践介绍 MySQL 分库分表离线同步至 Hive 数据源。

适用场景

业务层对基础同结构的业务数据使用分库分表的方式存储在不同的 MySQL 数据库实例或同一实例的不同表内。应用层需要对分布在不同数据库内表同时同步到数仓 ODS 层中统一存储。

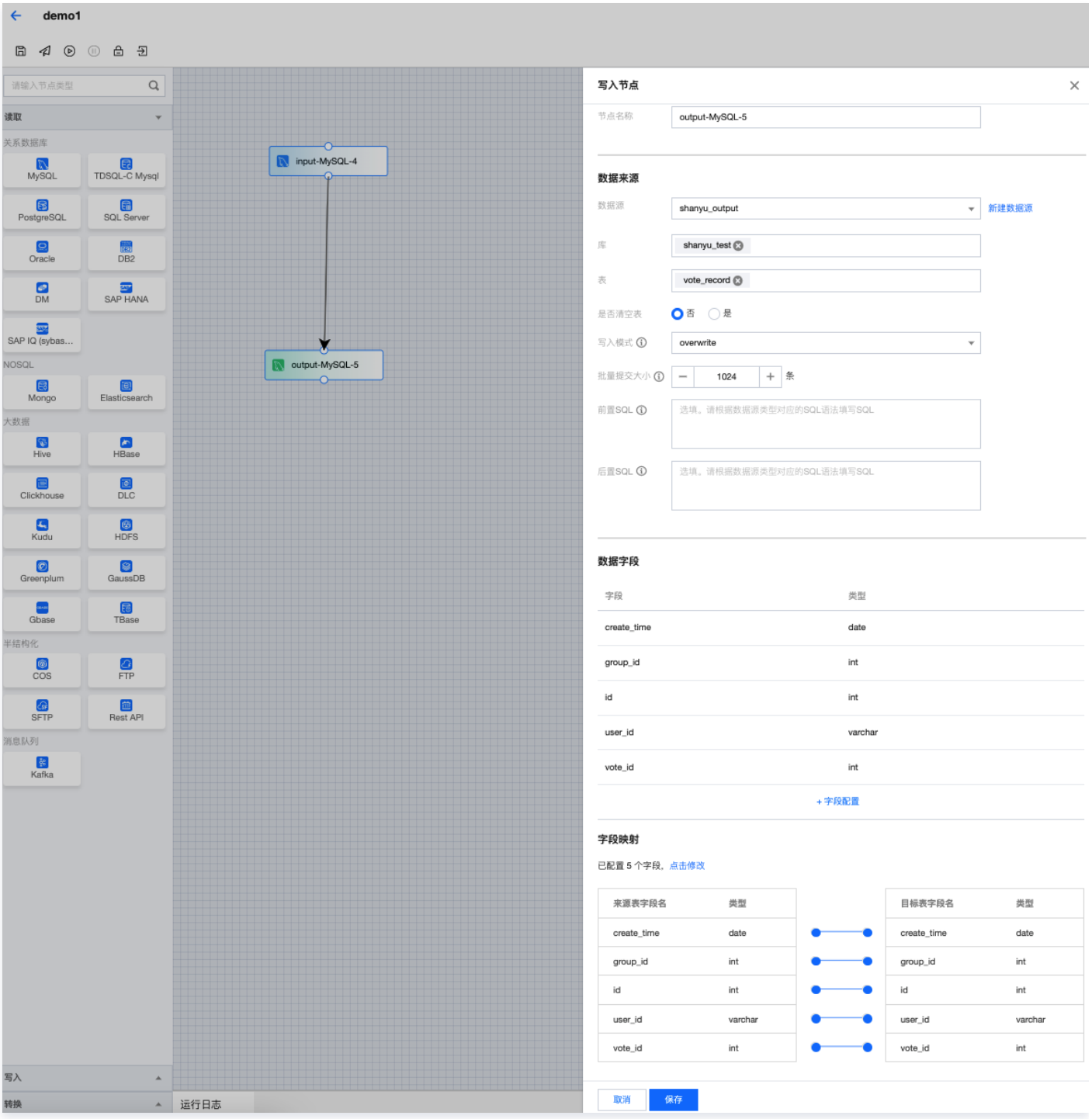
操作步骤

1. 创建离线同步任务，并从离线同步列表中点击对应任务名称进入画布配置界面。
2. 拖拽 MySQL 读取节点，默认 MySQL 读取节点支持一个数据源，单个数据源（库）内可选择多张 MySQL 表。若存在分库情况，可单击**添加分库分表**，即可创建多个数据源并添加对应的表对象。



> **说明：**
> 分库分表情况下选择的多个表对象需保证 Schema 信息一致（包括字段名称、字段类型）。**数据字段**模块内系统默认展示第一个数据源的第一张表的元数据字段信息，若多表间字段不一致可能会导致运行失败。

3. 拖拽 Hive 写入节点，并配置读写字段映射。



4. 保存任务信息，进行测试运行或提交至运维中心。

SQL Server 数据离线同步至 Hive 分区

最近更新时间：2024-07-09 22:01:41

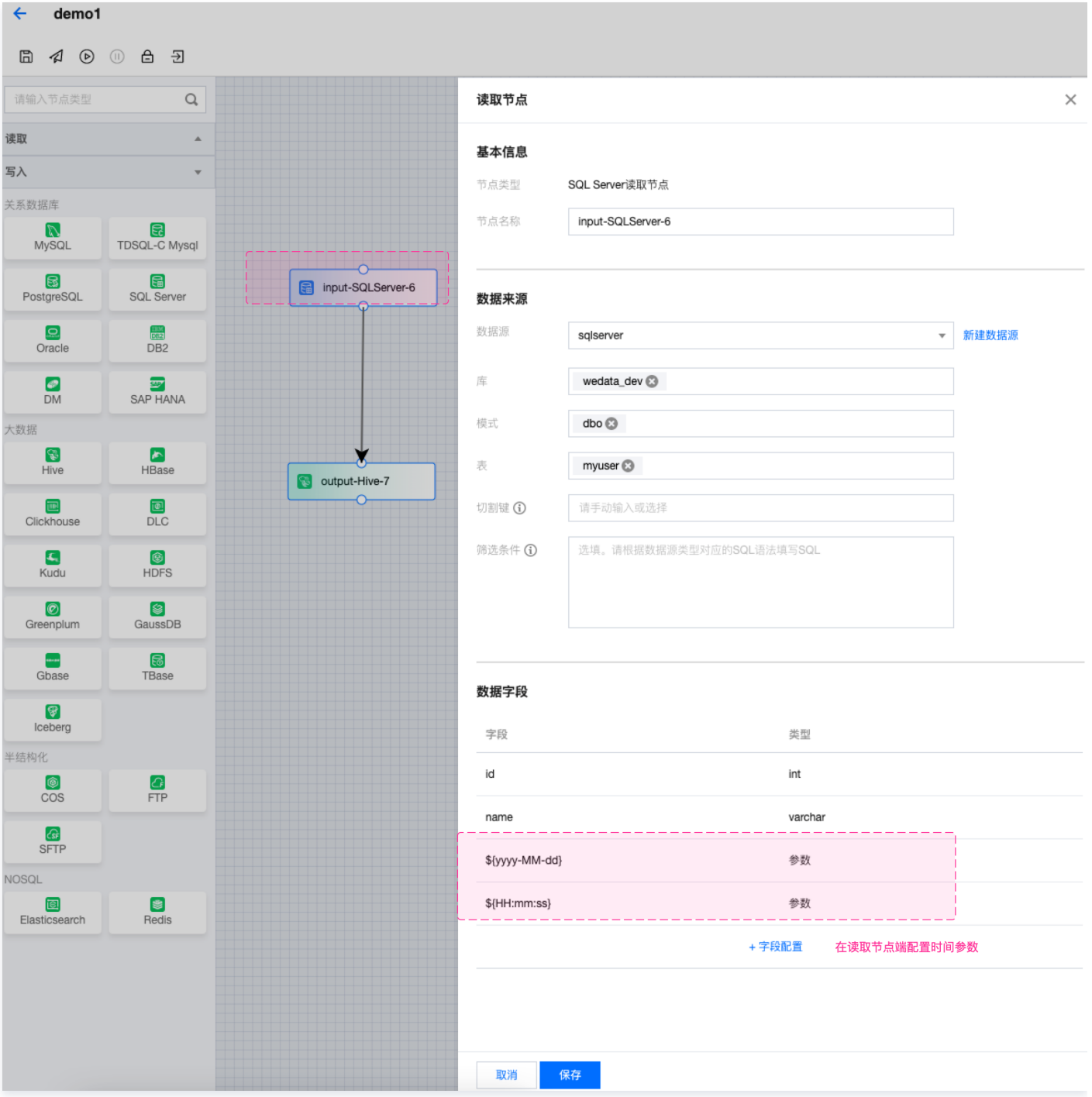
离线周期同步任务支持周期性更新表分区数据，本实践介绍 SQL Server 表数据定时更新至 Hive 分区。

适用场景

周期性创建或更新 Hive 分区表数据，将数据表中内容按照周期任务实例计划调度时间写入到对应一级或多级分区。

操作步骤

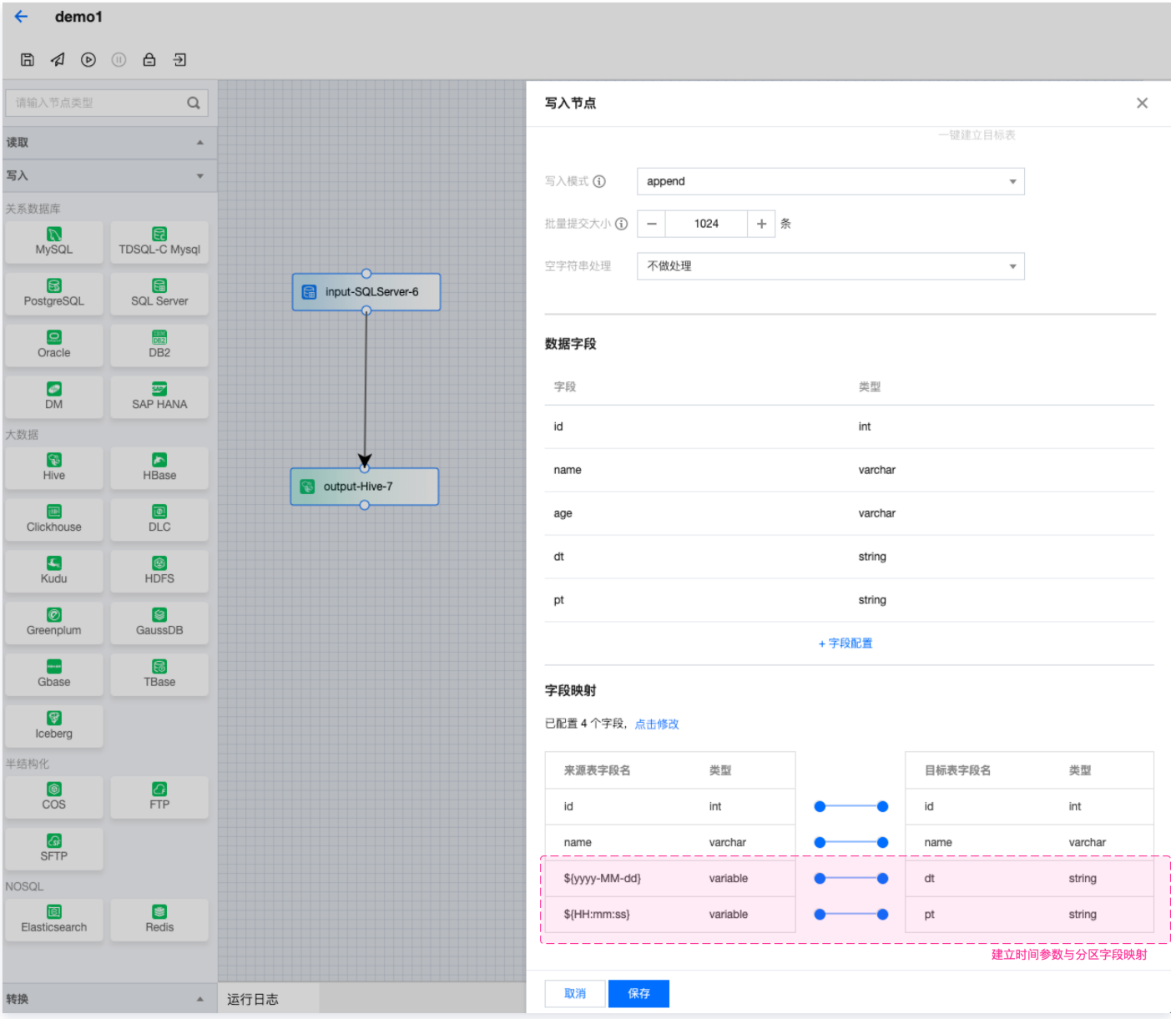
1. 创建离线同步任务，并从离线同步列表中单击对应任务名称进入画布配置界面。
2. 拖拽 SQL Server 读取节点，配置数据来源信息并在**数据字段**模块内配置时间参数。



说明：

- 时间参数定义方式请参见 时间参数说明 。
- 时间参数类型需配置为“参数”。

3. 拖拽 Hive 写入节点，配置数据源信息并在**字段映射**模块内建立时间参数与分区字段映射。



4. 保存**任务信息**，进行测试运行或提交至运维中心。

MySQL + gh-ost 实时同步至 Kafka

最近更新时间：2023-12-28 08:54:41

业务场景

- gh-ost 的业务场景是在 MySQL 中进行在线表结构变更，即 Online DDL，而不影响业务的正常运行。它可以解决传统的 alter table 或 create index 等命令导致的表锁、性能下降、同步延迟等问题。它适用于需要对表进行修改的场景，例如增加新列、添加索引、修改字段类型等。
- 本实践介绍使用 gh-ost 变更 MySQL 的表后，实时同步它的 DDL 变更记录到 Kafka。

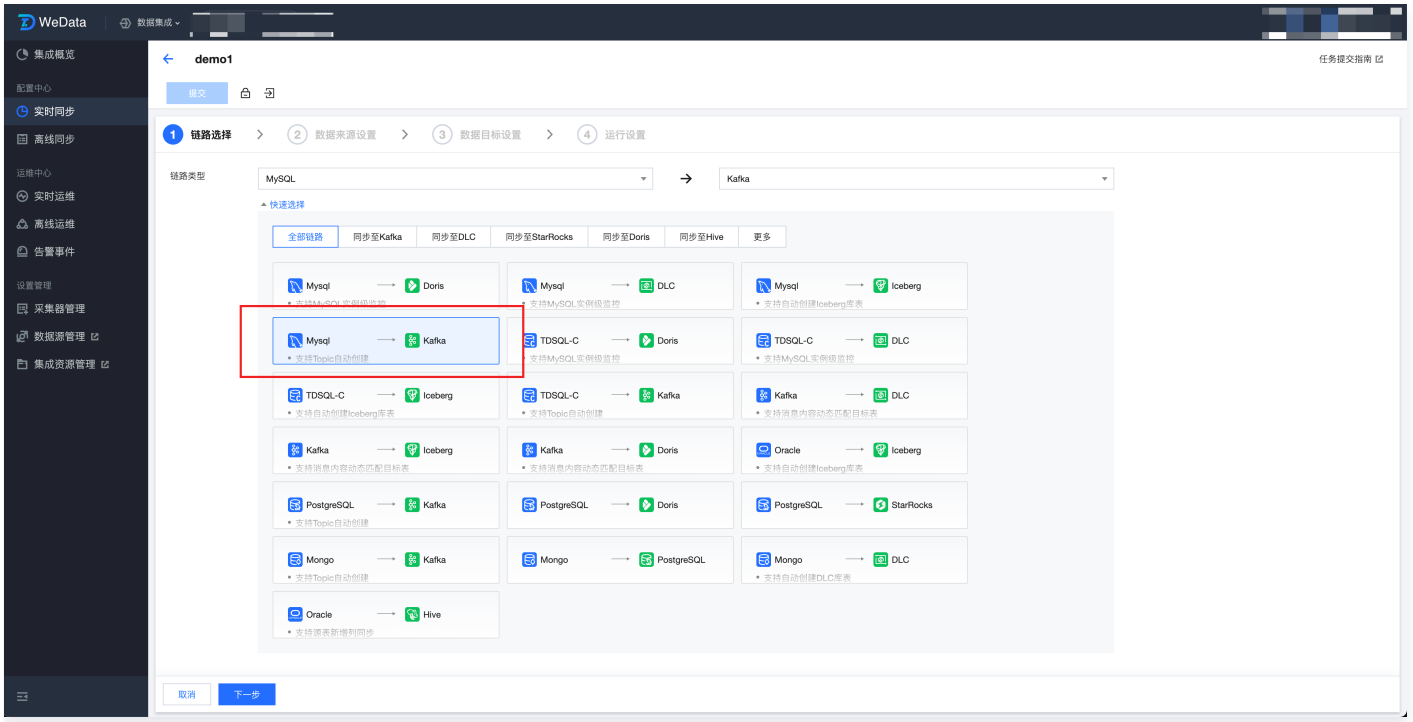
gh-ost 使用前提

MySQL 环境准备：数据集成 MySQL 环境准备与数据库配置 。

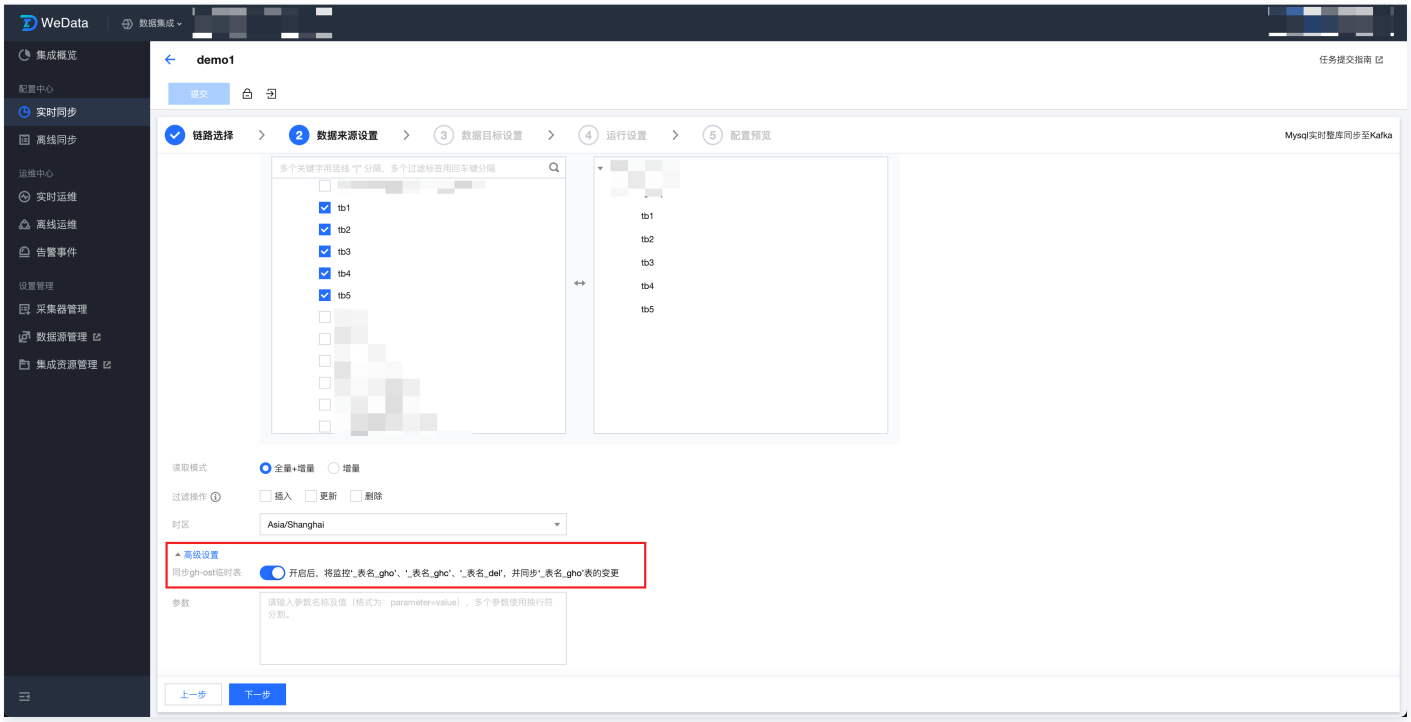
1. gh-ost 必须能访问 MySQL。
2. 如果 MySQL 是腾讯云的 CDB，在 gh-ost 执行命令的参数里面需要添加 `--aliyun-rds` 。
3. gh-ost 工具执行过程中，生成的临时表规则为 `^_(.*)_(gho|ghc|del)$` ，其中 (.*) 是变更表的名称，不支持自定义临时表的名称。
4. 其他 gh-ost 限制，可以参考：gh-ost/requirements-and-limitations.md at master · github/gh-ost · GitHub。

操作步骤

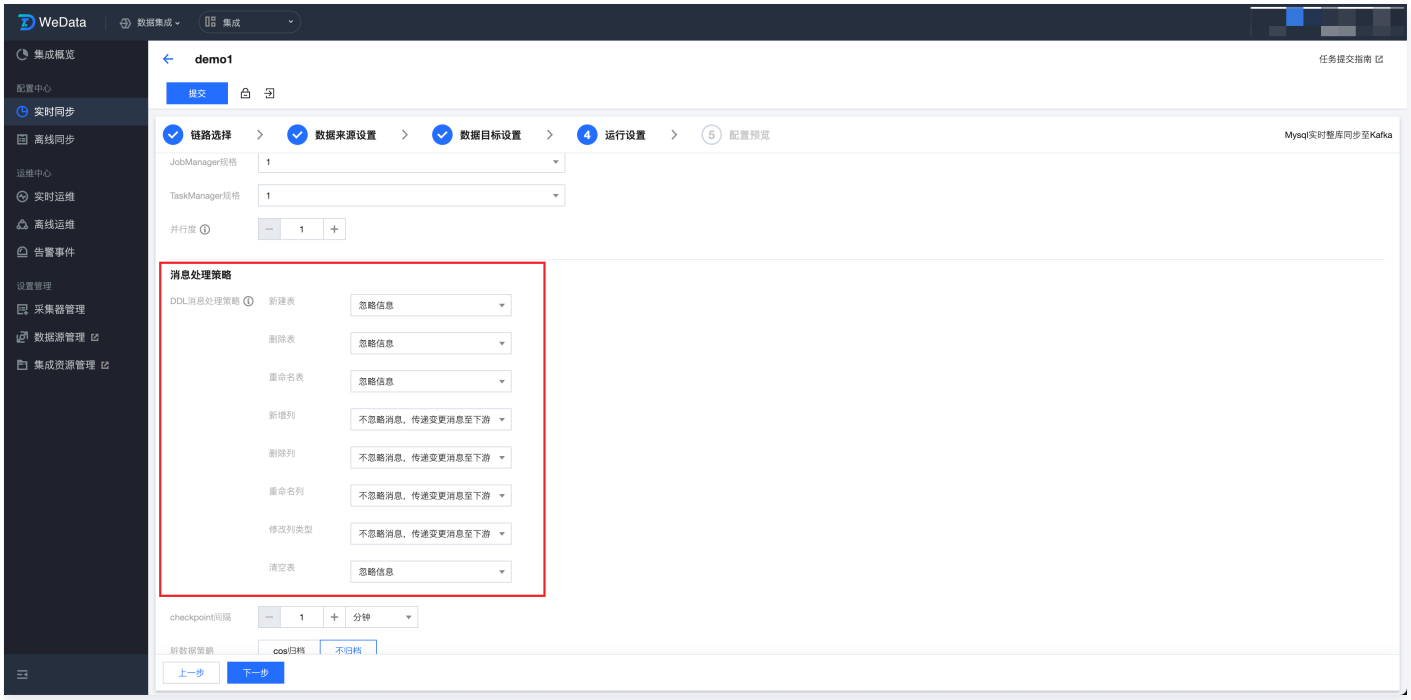
1. 创建 MySQL 实时同步到 Kafka 的整库同步任务。



2. 在任务配置中的**数据来源设置**，**高级设置**开启**同步 gh-ost 临时表**功能。



3. 在任务配置中的**运行设置**，**消息处理策略**选择需要的 DDL 消息处理策略，这里选择下发新增列，删除列，重命名列和修改列类型。配置完其余选项后，提交任务。



4. 使用 gh-ost 变更原始表。

4.1 查看表结构。

```
mysql> desc tb1;
+-------+--------------+------+-----+---------+-------+
| Field | Type         | Null | Key | Default | Extra |
+-------+--------------+------+-----+---------+-------+
| id    | int          | NO   | PRI | NULL    |       |
| name  | varchar(255) | YES  |     | NULL    |       |
+-------+--------------+------+-----+---------+-------+
```

```
2 rows in set (0.00 sec)
```

4.2 使用 gh-ost 工具在 源表 上增加一个字段。

gh-ost 工具下载：Releases · github/gh-ost · GitHub
下载后解压，执行下面命令，为 `tb1` 表添加 `c` 字段。

```
./gh-ost \
--max-load=Threads_running=25 \
--critical-load=Threads_running=1000 \
--chunk-size=1000 \
--throttle-control-replicas="" \
--max-lag-millis=1500 \
--user="root" \ // 用户名
--password="test" \ // 密码
--host=127.0.0.1 \ // mysql 的ip
--allow-on-master \
--database="databaseName" \ // 需要变更表所在的数据库名称
--table="tb1" \ // 需要变更的表名称
--verbose \
--alter="engine=innodb" \
--switch-to-rbr \
--allow-master-master \
--exact-rowcount \
--concurrent-rowcount \
--default-retries=120 \
--panic-flag-file=/tmp/ghost.panic.flag \
--default-retries=120 \
--alter="add column c varchar(255);" \ // alter 变更语句
--approve-renamed-columns \
--initially-drop-ghost-table \
--initially-drop-old-table \
--ok-to-drop-table \
--aliyun-rds \ // 使用腾讯云CDB，需要添加这个参数
--execute
```

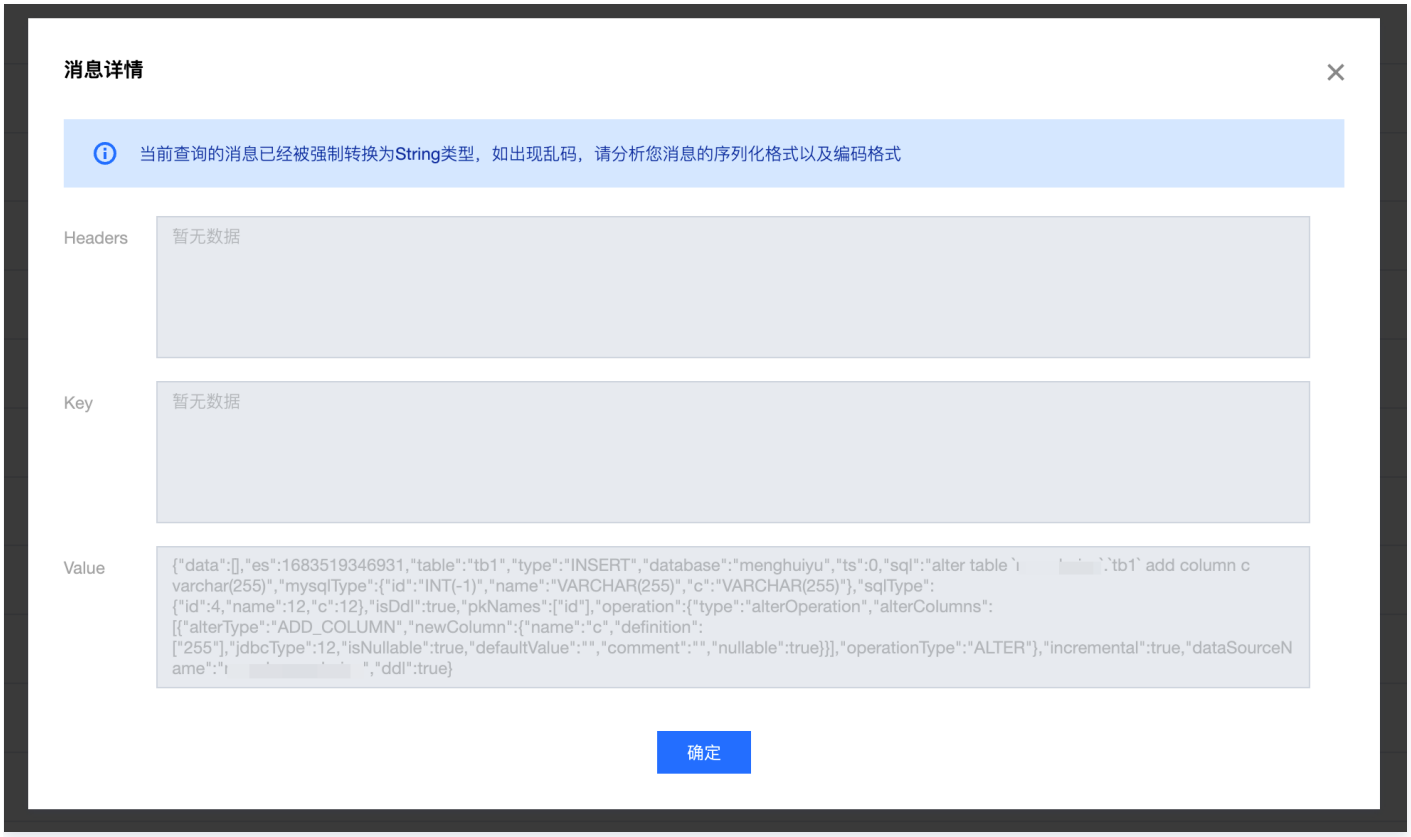
4.3 查看变更后的表结构。

```
mysql> desc tb1;
+-------+--------------+------+-----+---------+-------+
| Field | Type         | Null | Key | Default | Extra |
+-------+--------------+------+-----+---------+-------+
| id    | int          | NO   | PRI | NULL    |       |
| name  | varchar(255) | YES  |     | NULL    |       |
| c     | varchar(255) | YES  |     | NULL    |       |
+-------+--------------+------+-----+---------+-------+
3 rows in set (0.00 sec)
```

5. 在 Kafka 上查看结果。

```
{"data":[],"es":1683519346931,"table":"tb1","type":"INSERT","database":"databaseName","ts":0,"sql":"alter
table `databaseName`.`tb1` add column c varchar(255)","mysqlType":
{"id":"INT(-1)","name":"VARCHAR(255)","c":"VARCHAR(255)"},"sqlType":
{"id":4,"name":12,"c":12},"isDdl":true,"pkNames":["id"],"operation":
{"type":"alterOperation","alterColumns":[{"alterType":"ADD_COLUMN","newColumn":{"name":"c","definition":
["255"],"jdbcType":12,"isNullable":true,"defaultValue":"","comment":"","nullable":true}}],"operationType":
"ALTER"},"incremental":true,"dataSourceName":"databaseName","ddl":true}
```

示例如下：


消息详情
当前查询的消息已经被强制转换为String类型，如出现乱码，请分析您消息的序列化格式以及编码格式
Headers
暂无数据
Key
暂无数据
Value
{"data":[],"es":1683519346931,"table":"tb1","type":"INSERT","database":"menghuiyu","ts":0,"sql":"alter table `
`.`tb1` add column c
varchar(255)","mysqlType":{"id":"INT(-1)","name":"VARCHAR(255)","c":"VARCHAR(255)"},"sqlType":
{"id":4,"name":12,"c":12},"isDdl":true,"pkNames":["id"],"operation":{"type":"alterOperation","alterColumns":
[{"alterType":"ADD_COLUMN","newColumn":{"name":"c","definition":
["255"],"jdbcType":12,"isNullable":true,"defaultValue":"","comment":"","nullable":true}}],"operationType":"ALTER"},"incremental":true,"dataSourceN
ame":"
","ddl":true}
确定

将 MySQL 数据同步至 Doris
概述

最近更新时间：2024-06-18 14:47:41

本文将为您介绍如何通过 DataInLong 数据集成将 MySQL 中的数据实时导入 Doris 中，此处以腾讯云数据库 MySQL 为例，介绍实时同步任务的配置及操作实践。

> **注意：**
> 建议提前合理分配及配置网络环境，确保 MySQL、Doris、数据集成资源之间网络互通。
> - 若 MySQL、Doris、DataInLong 集成资源处于同一个 VPC 内：此时网络联通，可直接使用（**推荐方式**）。
> - 若与集成资源位于不同 VPC：需购买对等连接打通集成与数据源所在 VPC。
> - 若数据源位于 IDC 或其他经典网络环境下：可购买 VPN 连接或专线网关打通集成资源与 MySQL/Doris 集群所在 VPC。
> - 若数据源可开通公网：可购买 NAT 网关，允许集成资源通过网关连通数据源所在 VPC。NAT 网关配置流程请参见 集成资源组配置公网 。

步骤1：登录注册

登录腾讯云官网。如果没有账号，请参考 账号注册教程 。

步骤2：提前准备好 MySQL 数据库实例

数据集成目前支持 MySQL 数据库版本为：5.6，5.7，8.0.x。具体操作请参见 源端 MySQL 准备 。

步骤3：提前开通好云数据仓库 TCHouse-D 集群

建议选择内核版本：1.1、1.2或以上。具体操作请参见 目标端 Doris 准备 。

步骤4：购买 DataInlong，并配置项目空间和集成资源

建议选择内核版本：1.1、1.2或以上。具体操作请参见 配置 DataInlong 项目空间及集成资源 。

步骤5：在 WeData 或 DataInlong 配置数据实时同步任务

- 需要进行单表同步时，可配置单表实时同步任务，具体操作请参见 配置单表实时同步任务 。
- 需要进行整库迁移时，可配置整库实时迁移任务，具体操作请参见 配置整库实时迁移任务 。

步骤6：进入运维中心进行任务运维

具体操作请参见 任务运维 。

源端 MySQL 准备

最近更新时间：2024-08-16 16:27:21

新建 MySQL 数据源实例

> **说明**
> 目前数据集成支持 MySQL 数据库版本为：5.6，5.7，8.0.x。

使用腾讯云 MySQL 时

1. 您可登录 云数据库 TencentDB 控制台，进入 MySQL 实例列表。
2. 单击**新建**购买指定数据库版本的云实例。

> **注意：**
> 购买 MySQL 云实例时，建议配置 MySQL、Doris 处于同一个 VPC。

使用非腾讯云 MySQL 时

您可以在 MySQL 数据库中通过如下语句查看当前 MySQL 数据库版本，检查当前待同步的 MySQL 是否符合版本要求。

```
select version();
```

创建账号并赋权

> **注意：**
> 为保证实时数据同步顺利进行，您必须定义一个对 Debezium MySQL 连接器监控的所有数据库具有适当权限的 MySQL 用户。该 MySQL 账号必须拥有数据库的 SELECT、REPLICATION SLAVE 和 REPLICATION CLIENT 权限。

使用腾讯云 MySQL 时

1. 您可登录 云数据库 TencentDB 控制台，单击 **实例 ID/名称** 进入实例详情页。
2. 进入**数据库管理 > 账号管理** 页面，单击**创建账号**来新增账号，**修改权限**来配置账号权限。

使用非腾讯云 MySQL 时

您需要通过 SQL 语句授予并刷新账号权限。

```
mysql> GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT ON *.* TO 'user' IDENTIFIED BY
'password';
mysql>FLUSH PRIVILEGES;
```

说明：

1. 启用 `scan.incremental.snapshot.enabled` 时不再需要 RELOAD 权限（默认启用）。
2. 查看更多关于权限说明。

创建数据库

使用腾讯云 MySQL 时

1. 您可登录 云数据库 TencentDB 控制台，单击 **实例 ID/名称** 进入实例详情页。
2. 进入**数据库管理 > 数据库列表** 页面，单击 **创建数据库** 来创建 MySQL 数据库。

3. 在数据库登录跳转页面输入**账号管理**中已创建好的账号名和密码。

使用非腾讯云 MySQL 时

您可以通过 MySQL Client 等客户端进行建库操作。

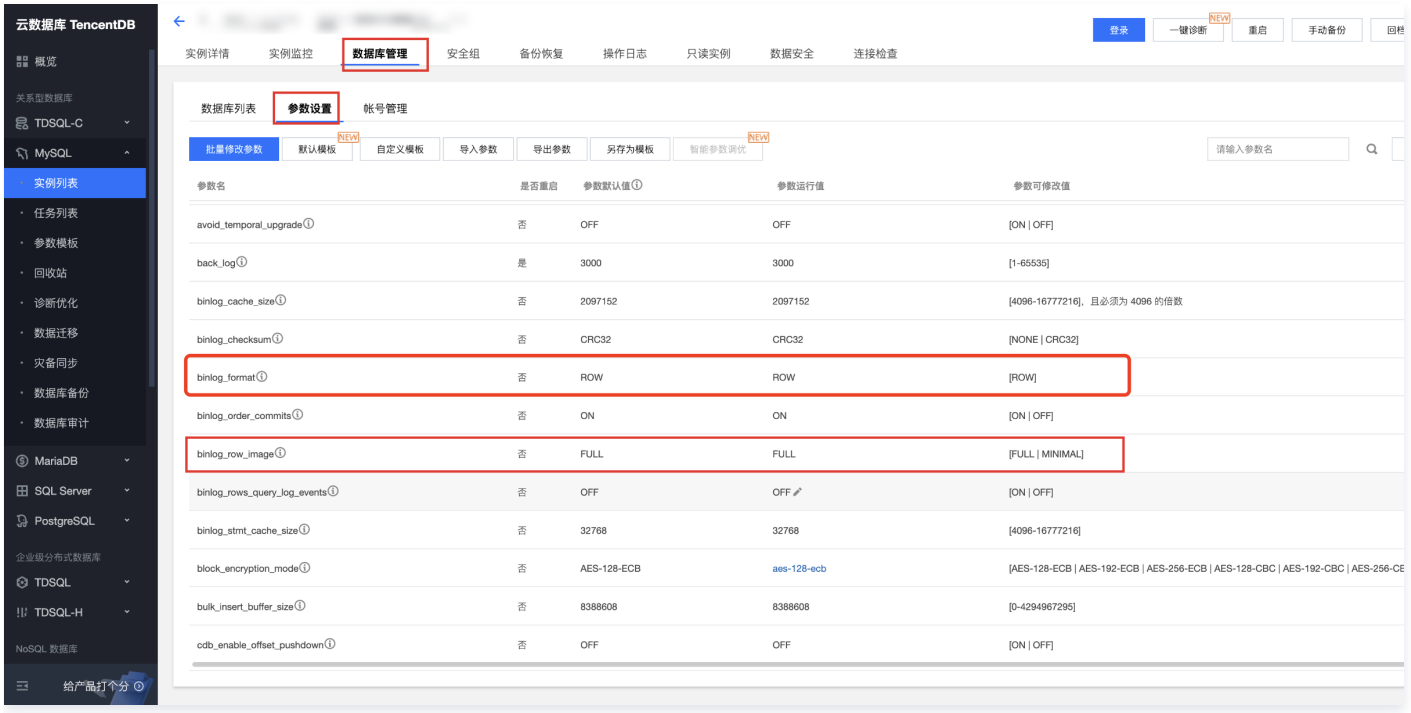
开启 Binlog 并确认 Binlog 格式

数据集成基于 MySQL binlog 进行数据同步时，要求 MySQL 服务器开启 binlog，并将 binlog 格式符配置为 ROW、将 binlog_row_image 配置格式为 FULL。

使用腾讯云 MySQL 时

腾讯云 MySQL 默认已开启 Binlog，可进入对应 MySQL 数据库管理–参数设置 页面设置并管理对应 binlog 参数：

> **说明：**
> MySQL 8.X 默认 binlog_format 为 ROW，无需额外配置。



使用非腾讯云 MySQL 时

可通过以下命令查看并配置 binlog 格式。

```
show variables like "binlog_format";
show variables like "binlog_row_image";
```

创建数据表

使用腾讯云 MySQL 时

1. 您可进入 数据库管理，登录腾讯云 MySQL 。

2. 在数据库管理页面新建数据表。

使用非腾讯云 MySQL 时

您可以通过 MySQL Client 等客户端进行建表操作。

目标端 Doris 准备

最近更新时间：2024-06-14 16:36:51

购买 Doris 集群

- 购买方式一：进入 腾讯云官网，登录后在官网首页单击**立即选购**。
- 购买方式二：登录 腾讯云数据仓库 TCHouse-D 控制台，在集群列表页新建集群。

- 进入集群购买页面，按需选购 FE、BE 的集群资源规格，详细操作步骤可参见 新建集群 。

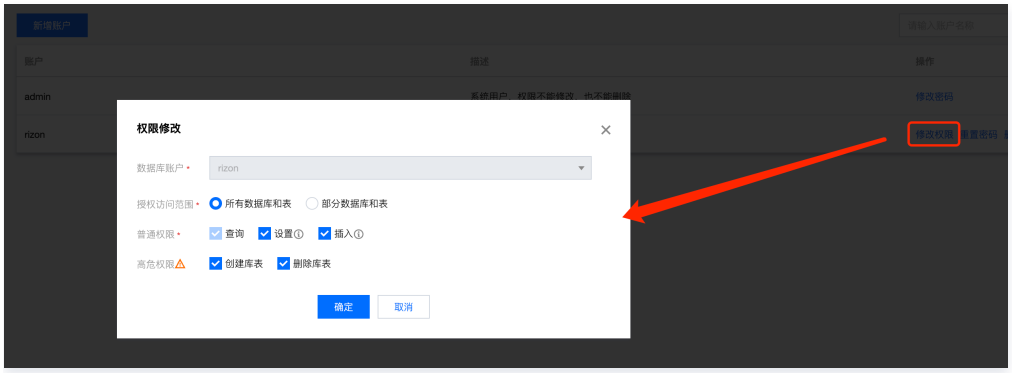
> **注意：**
> - 建议生产集群配置如下：
> - FE：开启高可用，3节点，每个节点16核64G。
> - BE：3节点，每个节点16核64G。
> - 购买 Doris 集群时，建议配置 MySQL、Doris 处于同一个 VPC 内。

创建账户并赋权

1. 集群购买完毕后，可进入 腾讯云数据仓库 TCHouse-D 控制台 集群列表，单击**集群 ID/名称**进行集群管理。

2. 可在“集群管理”模块查看集群具体信息、查看 BE/FE 监控指标、配置备份恢复策略等，详细操作说明可参见 集群常规操作 。
3. 为更方便的进行 doris 库表操作，需进入集群详情页，单击**账户管理**，创建库表并进行用户授权。单击**新增账户**，输入账户、名称后即可完成账号创建。

4. 账号创建完毕后，在账户列表单击**修改权限**按钮，即可为账号赋予相应的数据权限，支持为账户授予全部库表权限或指定表的权限，详细操作请参见 账户管理和权限管理 。



> **注意：**
> 授权所有数据库和表后，也将获取外部数据源的权限，但外部数据源仅支持查询，不支持插入、增删库表等。

进入 CDW Studio

- 进入 CDW Studio 时，需指定集群和登录账号。
  - Admin 用户登录后，默认有所有库表的权限。
  - 其他账号登录后，只能操作自己有权限的库表（可参见上文账号管理）。

> **说明：**
> **若忘记密码：**
> - 普通账户可在**集群详情页 > 账户管理**中直接重置密。
> - Admin 账户可通过 提交工单 联系我们重置密码。

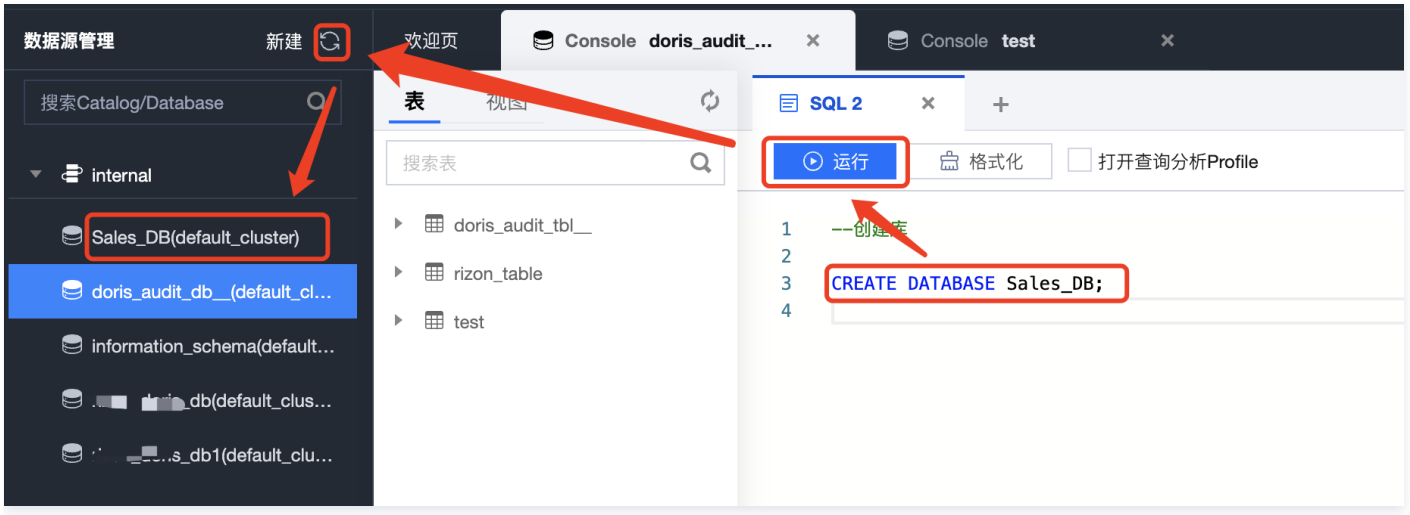
创建 Doris 数据库

> **注意：**
> 集群创建成功后，会初始化内置catalog（Internal）和2个系统数据库，请勿对系统数据库进行增删改等操作：
> 1. 系统库 doris_audit_db__：审计日志数据库，用于记录 Doris 系统的操作日志和安全事件。通过审计日志，可以追踪系统的操作历史和安全事件，保证系统的安全性和可靠性。
> 2. 系统库 information_schema：即 Doris 的元数据数据库，可通过查询 information_schema 来获取系统的元数据信息，了解系统的结构和属性，方便进行数据管理和查询操作。

在任意库下的 SQL 编译框中输入建库的 SQL 语句，并单击**运行**后，即可完成建库操作：

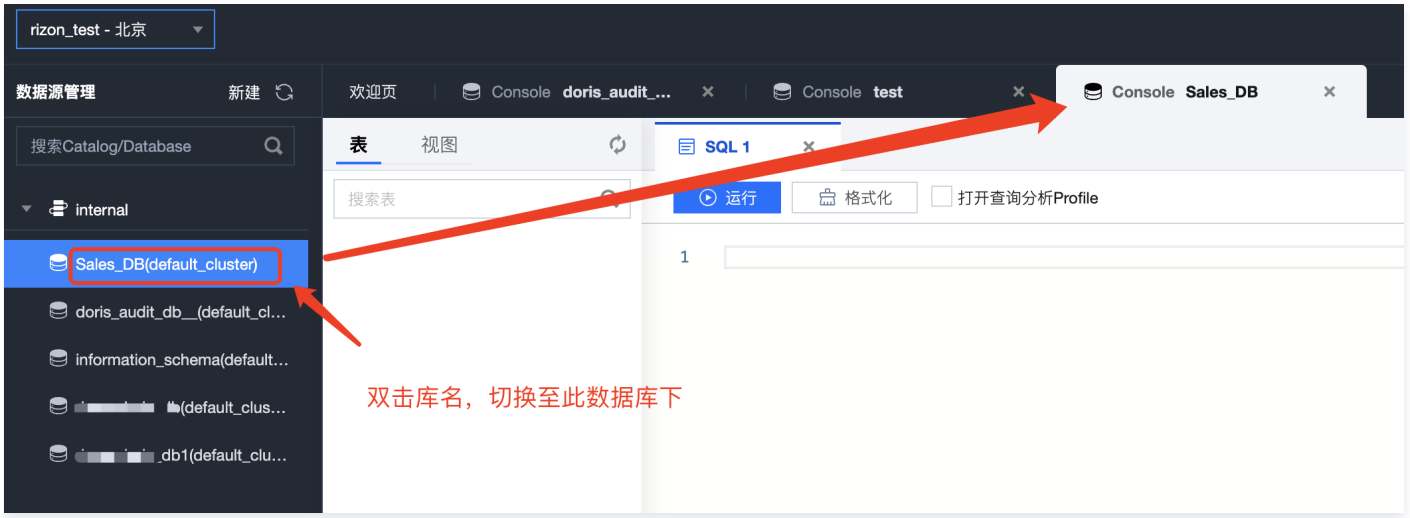
```
CREATE DATABASE Sales_DB;
```

建库完成后，单击数据源管理中的**刷新**，即可在目录树中回显新建的数据库。



创建 Doris 数据表

双击左侧库名，即可切换至此数据库下，然后可在弹出的 SQL 编译框中进行建表操作。



Doris 表模型选择

Doris 支持 Aggregate、Unique 和 Duplicate三种表模型，数据模型在建表时就已经确定，且无法修改。所以，**选择一个合适的数据模型非常重要**。

> **说明：**
> 具体表模型说明可参考文档：Doris 数据表和数据模型 。MySQL 表同步 Doris 时的选型建议如下：
> - 源端有主键的：建议创建 Unique 表模型接入数据。
> - 源端没有主键的：建议创建 Duplicate 表模型接入数据。

> **注意：**
> - key 列在表中的顺序需要和在 key 定义的顺序保持一致。
> - String/text 类型的列不适合作为 key 列，可以转为 varchar 类型。

1. **Aggregate 模型**：可以通过预聚合，极大地降低聚合查询时所需扫描的数据量和查询的计算量，非常适合有固定模式的报表类查询场景，但该模型对 count(*) 查询很不友好。
   建表 Demo 如下：

```
CREATE TABLE site_visit
(
    siteid      INT,
    city        SMALLINT,
    username    VARCHAR(32),
    pv BIGINT   SUM DEFAULT '0'
)
AGGREGATE KEY(siteid, city, username)
DISTRIBUTED BY HASH(siteid) BUCKETS 10;
```

2. **Unique 模型**：针对需要唯一主键约束的场景，Unique key 相同时，新记录覆盖旧记录，可以保证主键唯一性约束，适用于有更新需求的分析业务。
   建表 Demo 如下：

```
CREATE TABLE sales_order
(
    orderid     BIGINT,
    status      TINYINT,
    username    VARCHAR(32),
    amount      BIGINT DEFAULT '0'
)
UNIQUE KEY(orderid)
```

```
DISTRIBUTED BY HASH(orderid) BUCKETS 10;
```

3. **Duplicate 模型**：相同的行不会合并，适合任意维度的 Ad-hoc 查询。虽然无法利用预聚合的特性，但是不受聚合模型的约束，可以发挥列存模型的优势（列裁剪、向量执行等）。

   建表 Demo 如下：

```
CREATE TABLE session_data
(
    visitorid   SMALLINT,
    sessionid   BIGINT,
    visittime   DATETIME,
    city        CHAR(20),
    province    CHAR(20),
    ip          varchar(32),
    brower      CHAR(20),
    url         VARCHAR(1024)
)
DUPLICATE KEY(visitorid, sessionid)
DISTRIBUTED BY HASH(sessionid, visitorid) BUCKETS 10;
```

Doris 表分区分桶建议

Doris 支持两层的数据划分，第一层是 Partition（分区），支持 Range 和 List 的划分方式。第二层是 Bucket（分桶），仅支持 Hash 的划分方式。一个表可以不分区，但必须分桶。

> **说明：**
>
> 具体分区、分桶说明可参考文档：数据分区和分桶，简而言之：
>
> - 数据量较小的场景：推荐使用一级分片表（即指不分区的分桶表）。
> - 数据量较大或数据有明确日期归属的场景（如事实表）：推荐使用两级分片表，即既分区又分桶的表。

1. 创建一级分片表的示例：

```
Create Table `user_login_new`
(
    `loginId` bigint NOT NULL COMMENT '登录Id',
    `userAcount` varchar(255)NOT NULL COMMENT '用户账号',
    `onlineTime` decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长，单位：小时',
    `androidVersion` varchar
)
UNIQUE KEY (`loginid`)
COMMENT '用户归因登录表'
DISTRIBUTED BY HASH(`loginid`) buckets 8;
```

> **注意：**
>
> 创建分桶的语法：DISTRIBUTED BY HASH(`loginid`) buckets 8;
>
> 其中两个关键点：
>
> 1. 分桶列：分桶列有一定要求：必须是 key 列，类型不能是 string,text。
> 2. 分桶数量：根据数据大小确定，最好保证分桶后每个桶的大小在1-10G。

2. 创建两级分片表的示例

```
Create Table `user_login_beifen`
(
    `loginId` bigint NOT NULL COMMENT '登录Id',
    `loginTime` date NOT NULL COMMENT '登录时间',
    `loginIp` varchar(255)NOT NULL COMMENT '登录ip',
```

```
    `loginProvince` varchar(255)COMMENT '登录地区',
    `onlineTime` decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长，单位：小时'
)
UNIQUE KEY (`loginId`,`loginTime`)
COMMENT '用户归因登录表'
PARTITION BY RANGE(`loginTime`)
(
    PARTITION `p201701` VALUES LESS THAN ("2017-02-01"),
    PARTITION `p201702` VALUES LESS THAN ("2017-03-01"),
    PARTITION `p201703` VALUES LESS THAN ("2017-04-01")
)
DISTRIBUTED BY HASH(`loginId`) buckets 8;
```

注意：

上为创建 Range 类型分区的语法示例，有几个关键点：

1. 分区列可以是一列或多列，但都需要是key列。
2. 创建分区时不可添加范围重叠的分区。
3. 数据写入前数据归属的分区要提前创建好。

以 Doris 也支持预先创建分区、自动创建分区，即动态分区特性：例子如下：

```
Create Table user_login_beifen2
(
    loginId bigint NOT NULL COMMENT '登录Id',
    loginTime date NOT NULL COMMENT '登录时间',
    loginIp varchar(255)NOT NULL COMMENT '登录ip',
    loginProvince varchar(255)COMMENT '登录地区',
    onlineTime decimal(10, 2) NOT NULL DEFAULT '0.00' COMMENT '玩家在线时长，单位：小时'
)
UNIQUE KEY (loginId,loginTime)
COMMENT '用户归因登录表'
PARTITION BY RANGE(loginTime)()
DISTRIBUTED BY HASH(loginId) buckets 8
PROPERTIES
(
    "dynamic_partition.time_unit" = "DAY",
    "dynamic_partition.start" = "-10",
    "dynamic_partition.end" = "3",
    "dynamic_partition.buckets" = "8",
    "dynamic_partition.create_history_partition" = "true",
    "dynamic_partition.prefix" = "p"
);

--上述建表demo预先创建好今天之前10天至后3天的分区表，按loginTime列分区，分区名以'p'开头，每个分区内分成8个桶。
```

Doris 索引使用建议

如果经常对某列进行精确匹配过滤并且列的基数比较高，建议在此列上创建 bloom filter 索引。

建表时尽量避免的操作

特别注意：截止当前最近 Doris 版本（1.2.4.2），以下表相关功能还不完善，**不建议生产使用。**

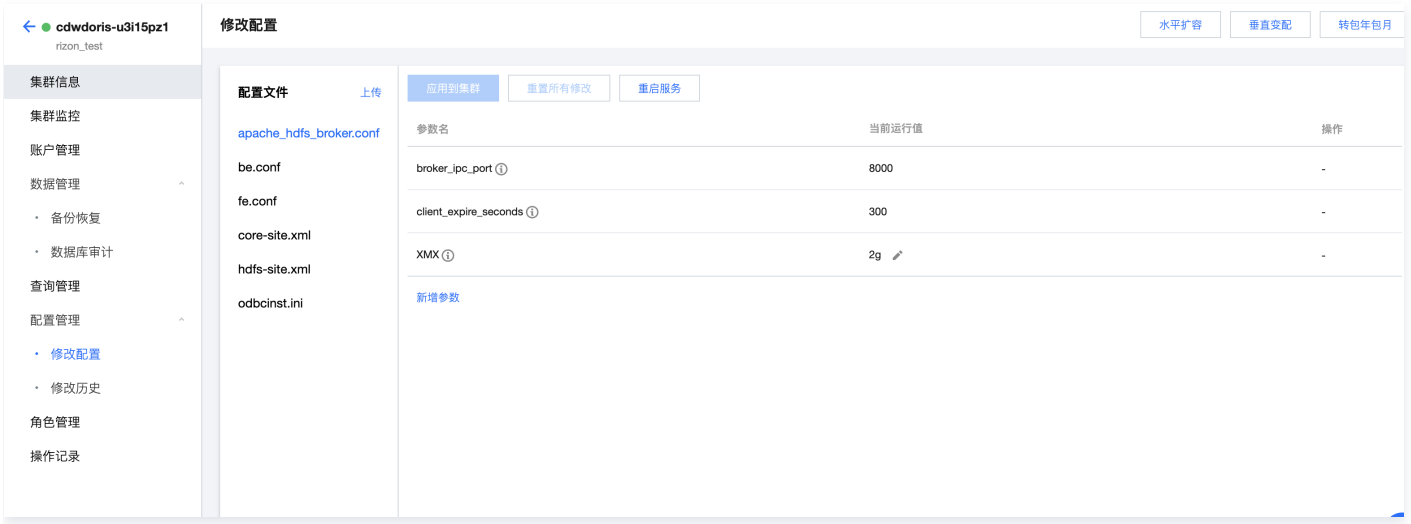
注意：

- 不建议使用 “Merge-on-Write” 功能。
- 不建议使用 “auto bucket” 功能。

- 不建议使用 “动态Schema表” 功能。

集群配置建议

创建集群时，会初始化以下5个配置文件，说明如下：



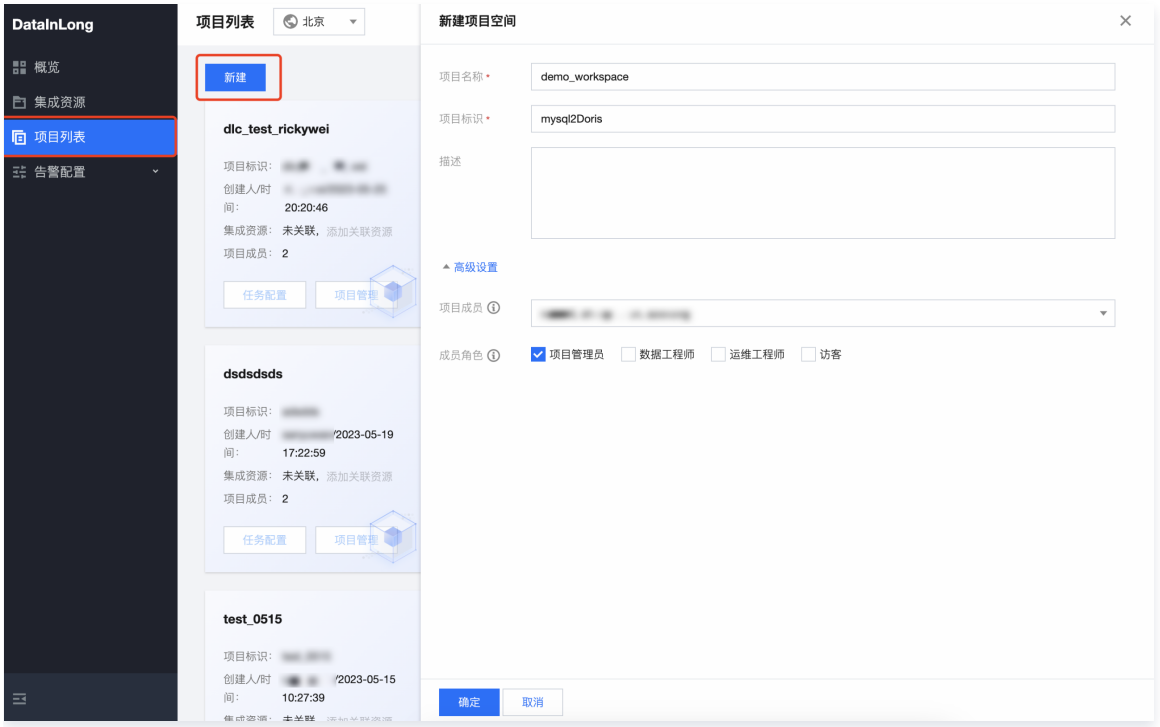
配置文件	配置建议
apache_hdfs_broker.conf	建议保持默认配置不变
be.conf	大多数配置保持默认配置不变，需特殊注意的参数如下： • compaction_task_num_per_disk：每个磁盘可并发执行的 compaction 任务数量，默认值是2，如果想要提高导入速度可适当调大。具体可参考社区文档 BE 配置项 。 • disable_auto_compaction=true：建议不要调整。
fe.conf	大多数配置保持默认配置不变，需特殊注意的参数如下： • max_running_txn_num_per_db: 控制同一个 DB 的并发导入个数，默认值100，当集群中有过多的导入任务正在运行时，可适当调大。
core-site.xml	建议保持默认配置不变。
hdfs-site.xml	建议保持默认配置不变。
odbcinst.ini	建议保持默认配置不变。

配置 DataInlong 项目空间及集成资源

最近更新时间：2024-08-15 21:18:52

创建项目空间

进入 DataInlong 控制台，单击**项目列表 > 新建**，创建并配置项目及所包含成员。



购买集成资源组并关联项目

1. 进入 DataInlong 控制台，选择集成资源并单击**创建**。

2. 购买并配置集成资源。

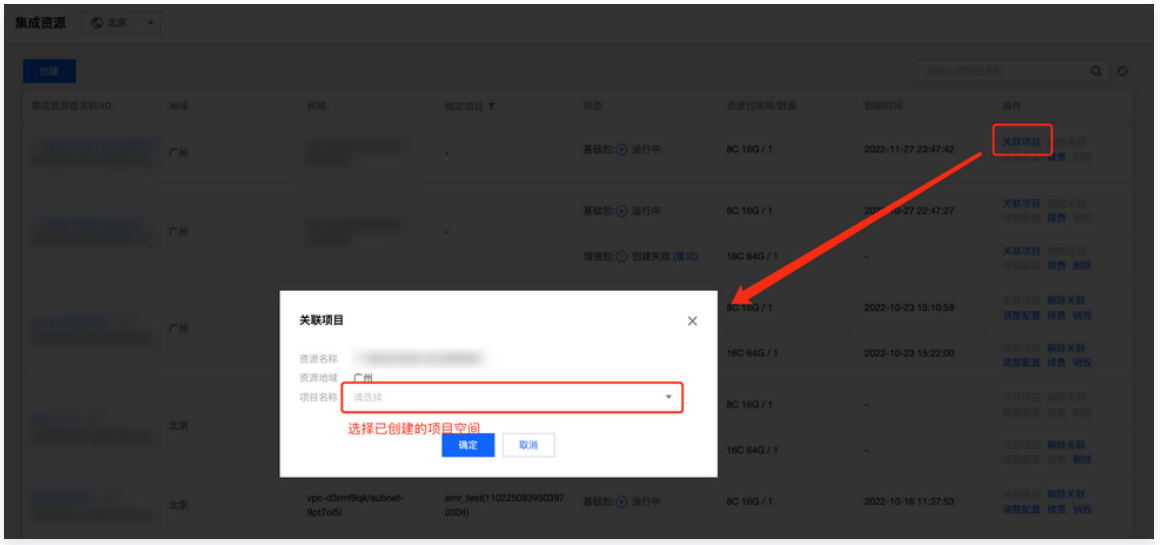
说明：

- 离线资源包与实时资源包可根据实际数据情况配置规格以及数量。
- 资源组网络建议和 MySQL 及 Doris 在同一个 VPC 下，若不在一个 VPC 下，可为 VPC 配置开通公网，可参见 资源组配置公网 。
- 购买完成后，返回控制台并关联资源组与项目空间。

3. 关联资源组和项目空间

说明：

若在购买页面内已经关联资源组与项目空间，可忽略此步骤。

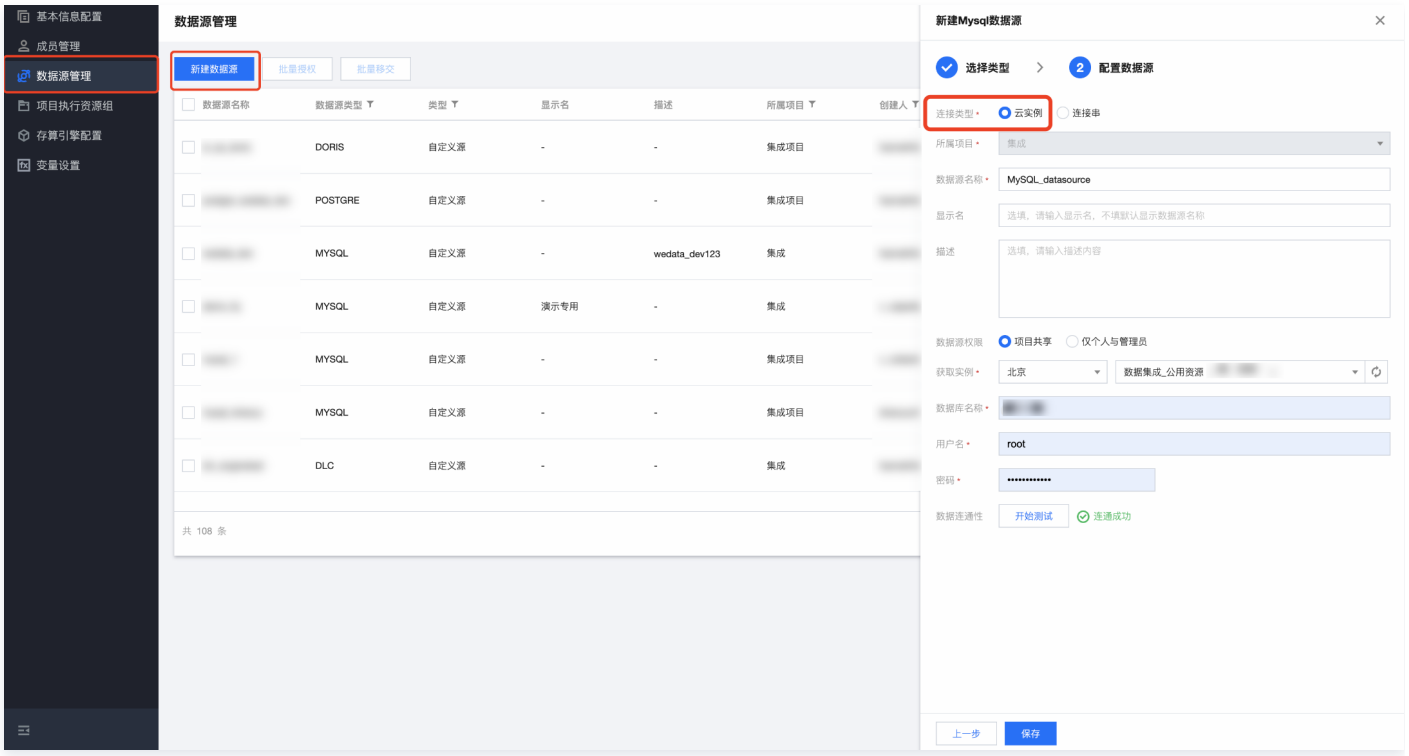


进入项目空间并注册数据源

配置 MySQL 数据源

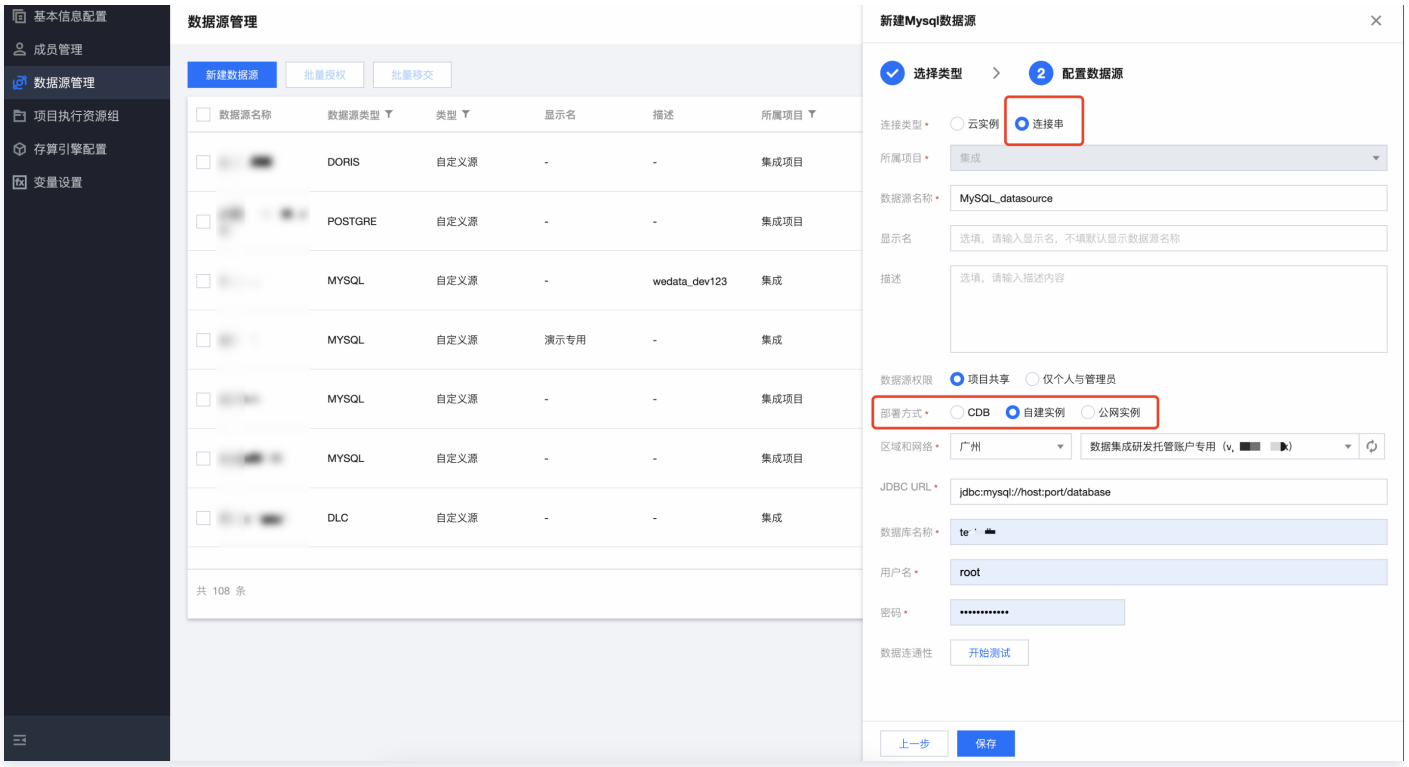
支持注册腾讯云 MySQL 或本地自建 MySQL 数据源。进入 项目管理 模块，选择**数据源管理 > 新建数据源 > 选择 MySQL**。

- 使用腾讯云 MySQL 时：可通过配置“云实例”直接关联已购买地域数据库云实例构建数据源。



参数	参数说明
连接类型	腾讯云 MySQL 数据库通过云实例方式添加，此方式下可直接获取当前账号下的 MySQL 数据。
数据源名称	新建的数据源的名称，由用户自定义且不可为空。命名以字母开头，可包含字母、数字、下划线。长度在20字符以内。
显示名	数据源在产品中使用时的显示名称，不填默认显示数据源名称。
描述	选填，对本数据源的描述。
数据源权限	项目共享表示当前数据源项目所有成员均可使用 ，仅个人和管理员表示该数据源仅创建人和项目管理员可用。
获取实例	选择账户下云数据库实例所在的地域、实例名称及 ID 信息。
数据库名	填入云实例下数据库名称；此数据库后续将作为后续数据源的默认数据库。
用户	连接数据库的用户名称。
密码	连接数据库的密码。
连通性	测试是否能够连通所配置的数据库。 **注意：** • 若连通性测试不通过，可以继续创建数据源，但后续数据读写时会报错。 • 如果连通性测试不通过，可能是因为 WeData 被数据库所在网络防火墙禁止，请参见 添加腾讯云 MySQL 数据库安全组 。

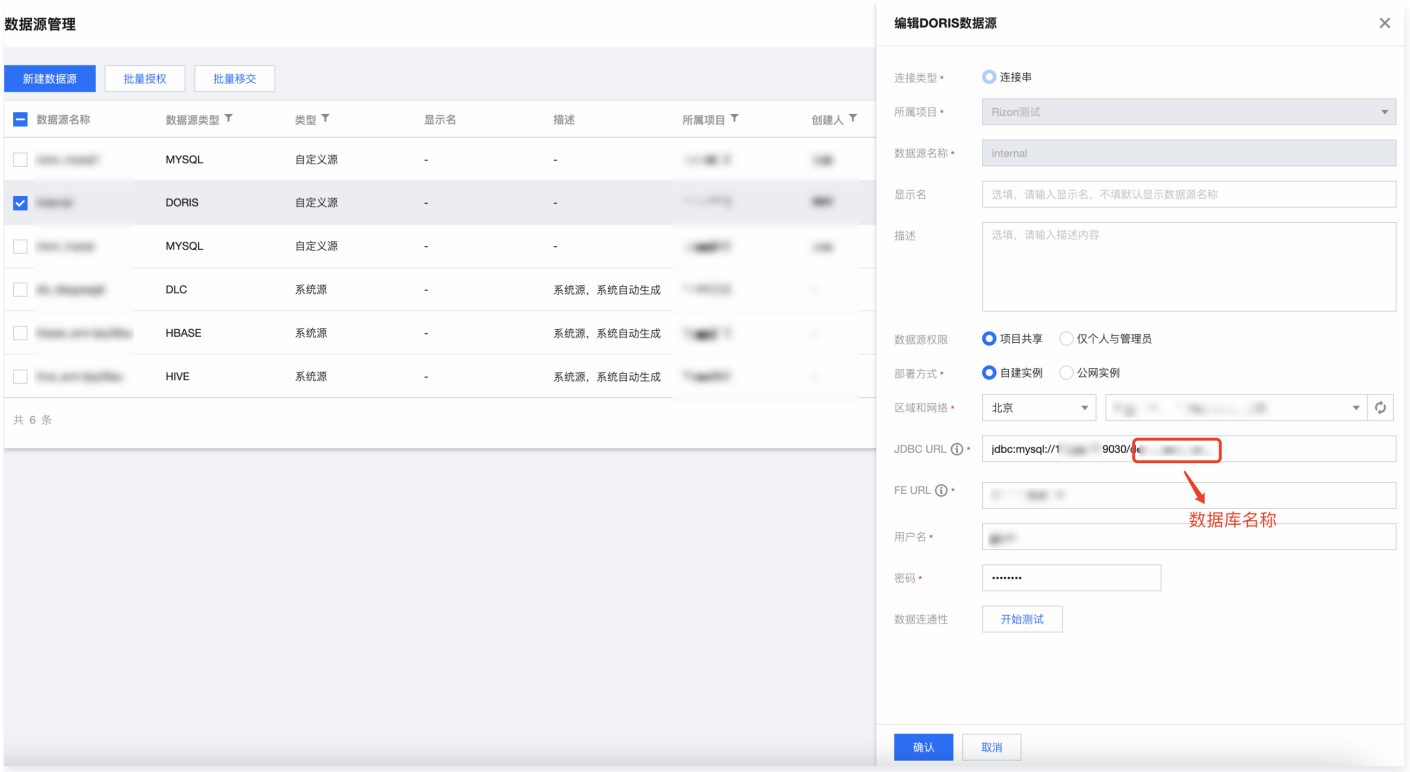
- 使用非腾讯云 MySQL 时：通过“连接串” JDBC 方式添加自建数据库作为数据源。



参数	说明
连接类型	非腾讯云数据库实例可通过连接串方式连接。
数据源名称	新建的数据源的名称，由用户自定义且不可为空。命名以字母开头，可包含字母、数字、下划线。长度在20字符以内。
显示名	数据源在产品中使用时的显示名称，不填默认显示数据源名称。
描述	选填，对本数据源的描述。
数据源权限	项目共享表示当前数据源项目所有成员均可使用 ，仅个人和管理员表示该数据源仅创建人和项目管理员可用。
部署方式	• CDB：仅适用于使用腾讯云数据库。 • 自建实例：适用于自建且开通 VPC 环境内的 MySQL 集群。 • 公网实例：适用于开通了公网的 MySQL 集群。
数据库名	填入数据库名称；此数据库后续将作为后续数据源的默认数据库。
用户	连接数据库的用户名称。
密码	连接数据库的密码。
连通性	测试是否能够连通所配置的数据库。 **注意：** **若连通性测试不通过，可以继续创建数据源，但后续数据读写时会报错。**

配置 Doris 数据源

进入 项目管理 模块，选择**数据源管理 > 新建数据源 > 选择 Doris**，配置数据源参数并在连通性测试成功后即可**保存**。

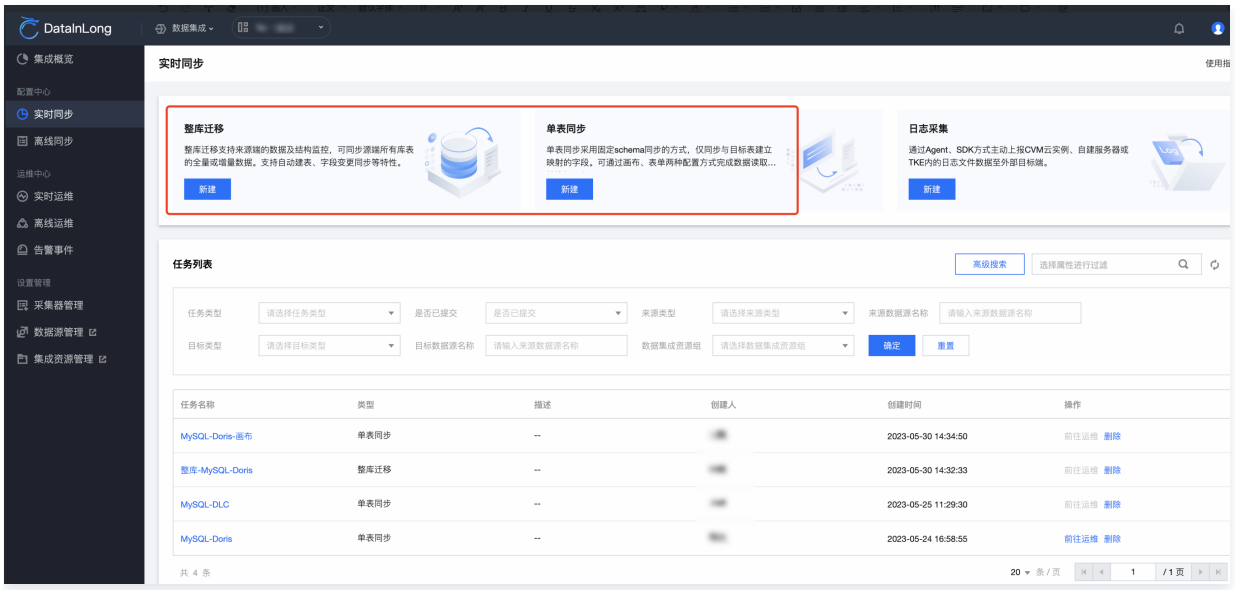


参数	说明
数据源名称	新建的数据源的名称，由用户自定义且不可为空。命名以字母开头，可包含字母、数字、下划线。长度在20字符以内。
描述	选填，对本数据源的描述。
数据源权限	项目共享表示当前数据源项目所有成员均可使用 ，仅个人和管理员表示该数据源仅创建人和项目管理员可用。
部署方式	• 自建实例：位于 VPC 环境内 Doris 实例。 • 公网实例：可使用公网访问的实例。
区域和网络	选择账户下云数据库实例所在的地域、实例名称及 ID 信息。
JDBC URL	用于连接 Doris 数据源的连接串信息： • 端口若为自建实例请填入内网IP地址和端口，多个地址间逗号（,）分隔，例如：jdbc:mysql://内网IP:port/参数。 • 若为公网实例请填入公网ip地址和端口，例如：jdbc:mysql://公网IP:port/参数。 注：上述“参数”填写数据源下的任一“数据库名称”即可，用于校验连通性。
FE URL	输入 fe http 地址，格式为：IP 地址:http 端口（无需 https:// 或 http://前缀），多个地址之间使用逗号（,）分隔，例如：172.17.16.3:8030,172.17.16.4:8030。
用户名	连接数据源的用户名称。
密码	连接数据源的密码。
数据连通性	测试是否能够连通所配置的数据库。

配置单表实时同步任务

最近更新时间：2024-08-15 21:18:52

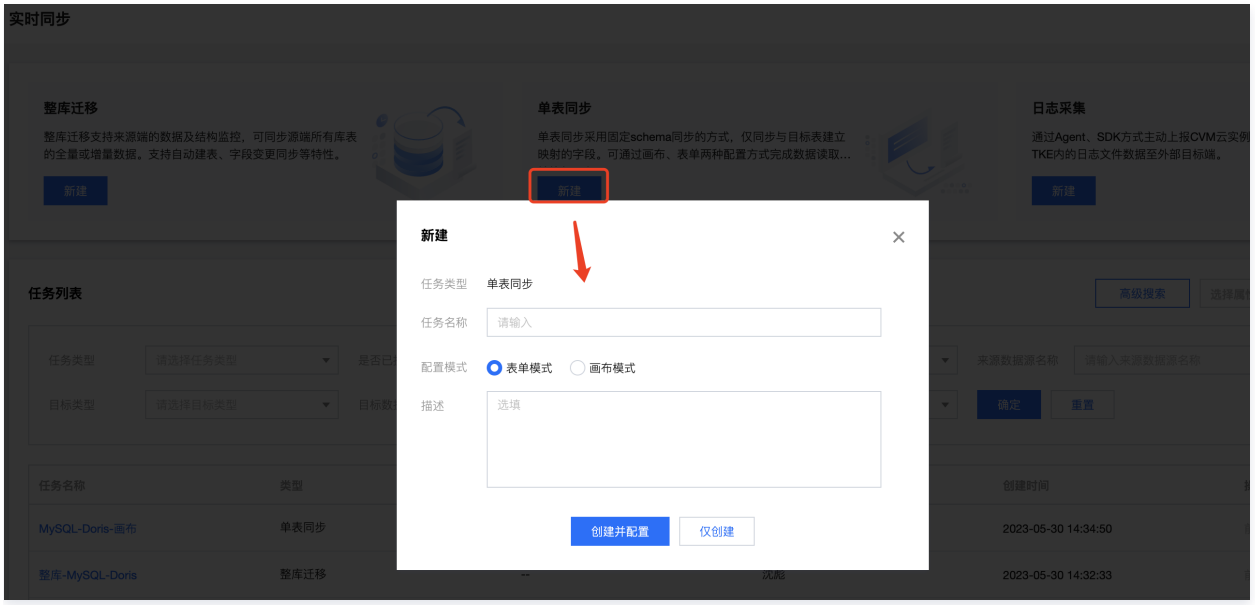
数据集成支持单表同步、整库迁移两种数据同步方式，下文将介绍单表实时同步的操作方法。



创建单表同步任务

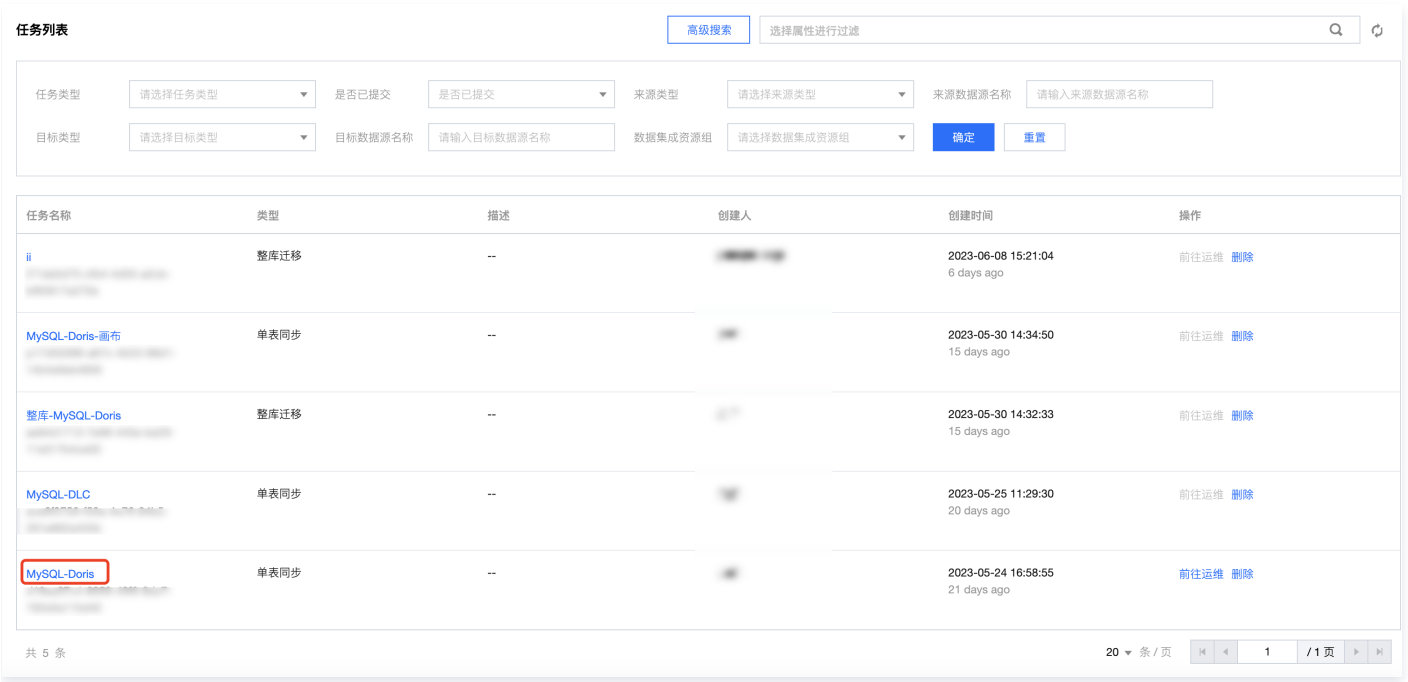
1. 进入**数据集成 > 实时同步**页面，单击**新建**创建单表同步任务。

2. 在弹出的提示框中输入任务名称和备注，选择表单模式或画布模式后（此处 demo 选择“单表模式”），单击“创建并配置”或“仅创建”完成任务创建。

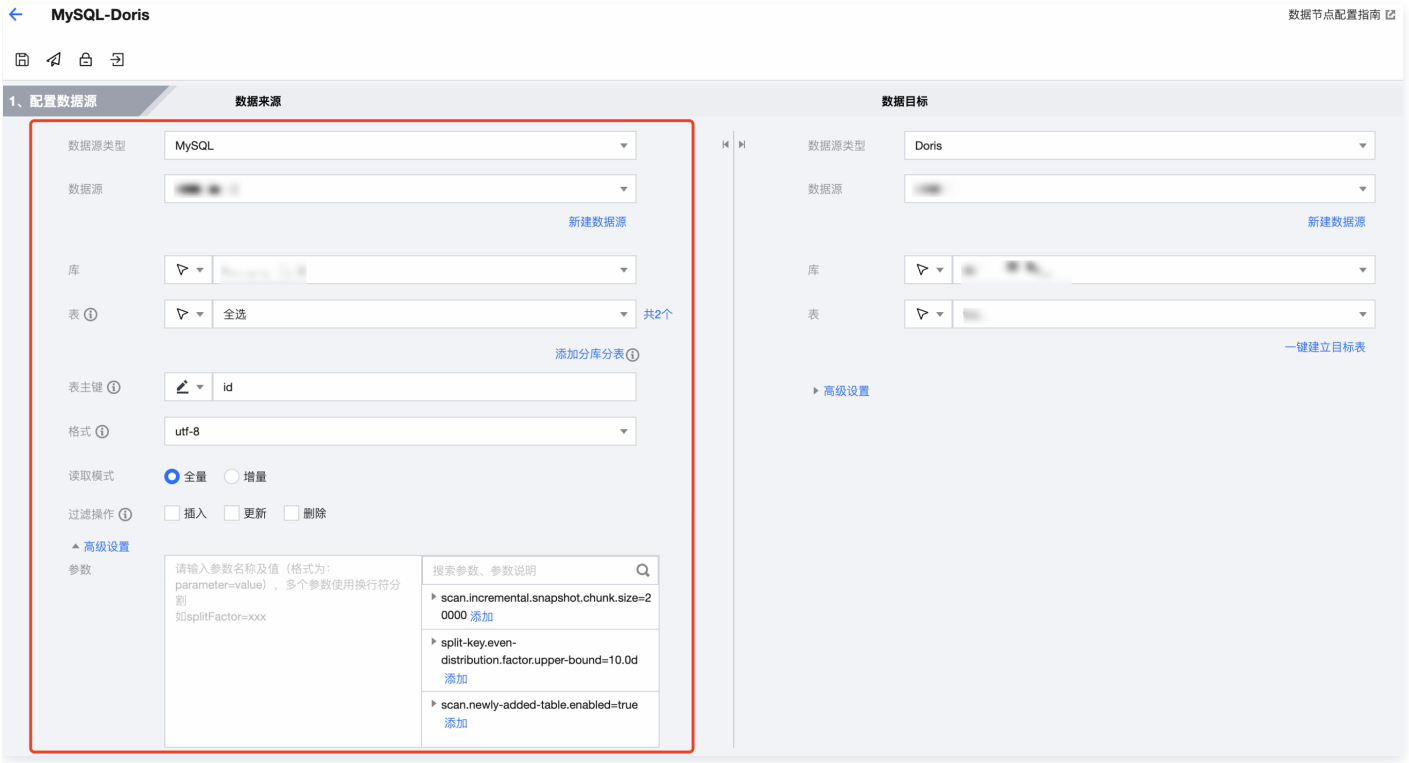


编辑任务

创建完成任务后，可在任务列表页面单击新建的实时同步任务名称进入任务编辑界面，本文以单表模式为例，进行作业配置。



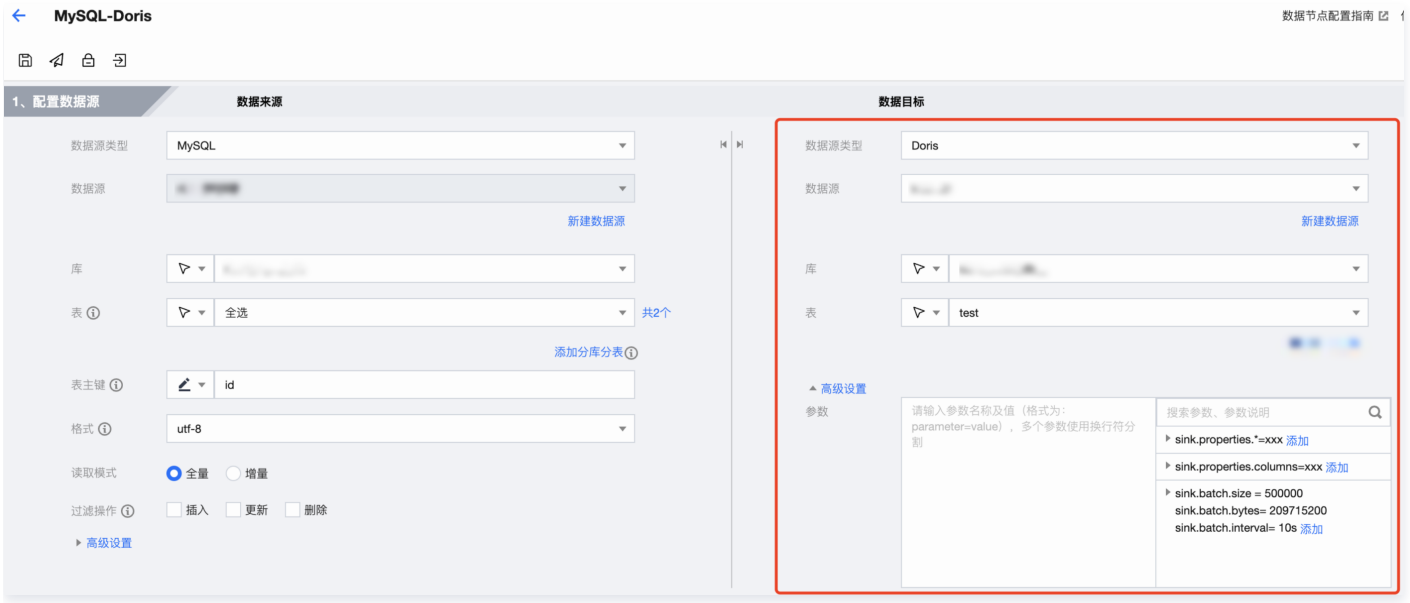
配置 MySQL 信息



参数	说明
数据源类型	可支持多种，此处选择 MySQL。
数据源	可选择上文“数据源管理”中已注册好的 MySQL 数据源。
库	单选，支持选择值、输入值/表达式两种方式： • 选择值时：可通过下拉列表选择在“数据源管理”中注册的数据库。 • 输入值/表达式时：可手动输入选定数据源下的已有库。

表	可多选，支持选择值、输入值/表达式两种方式： • 选择值时：可通过下拉列表选择。 • 输入值/表达式时：可手动输入。 ⚠ **注意：** 选择多个表时需保证多表 schema一致，当选定的多个表的 schema 不一致时，将以选中的第一个表的 schema 为准。
表主键	支持选择值、输入值/表达式两种方式： • 选择值时：可通过下拉列表选择。 • 输入值/表达式时：可手动输入。 ⚠ **注意：** **分库分表模式下默认表 schema 一致，当选定的多个表的 schema 不一致时，系统将使用拉取的第一张表的主键。**
读取模式	• 全量：将同步库内历史数据，全量同步结束后，会继续同步增量数据。 • 增量：仅从任务启动后的 binlog cdc 位点开始同步数据。
过滤操作	支持插入、更新和删除三种操作，设置后将不同步指定操作类型的数据。
高级设置	当前算子的运行参数，具体参数说明请参见 实时节点高级参数 。

配置 Doris 信息

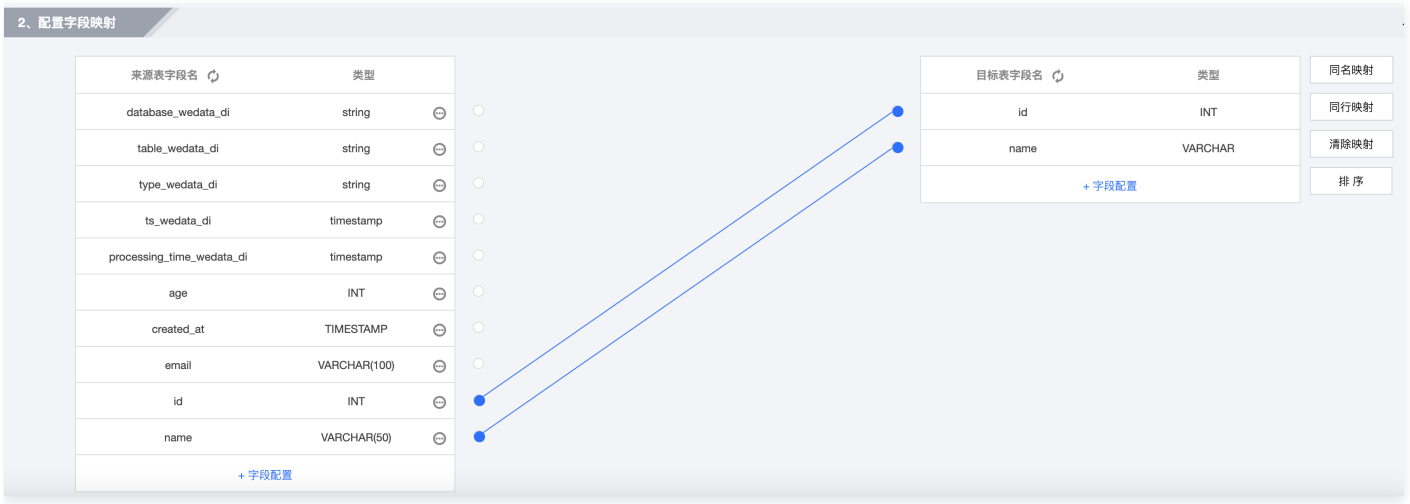


参数	说明
数据源类型	可支持多种，此处选择 Doris。
数据源	可选择上文“数据源管理”中已注册好的 Doris 数据源。
库	单选，支持选择值、输入值/表达式两种方式： • 选择值时：可通过下拉列表选择在“数据源管理”中注册的数据库。 • 输入值/表达式时：可手动输入选定数据源下的已有库。
表	单选，支持选择值、输入值/表达式两种方式： • 选择值时：可通过下拉列表选择。 • 输入值/表达式时：可手动输入。
过滤操作	支持插入、更新和删除三种操作，设置后将不同步指定操作类型的数据。

高级设置	当前算子的运行参数，具体参数说明请参见 实时节点高级参数 。

配置表字段映射

此处可设置来源和目标端数据对应关系，后续任务仅同步具有映射关系的字段之间的数据。



参数	说明
同名映射	即建立源端表、目标表同名字段的映射关系。
同行映射	即根据相同的行号建立源端表、目标表字段的映射关系。
手动映射	除同名映射、同行映射的快捷方式外，还支持手动连线的方式进行字段映射。
清除映射	即清除当前已创建好的映射关系。
排序	对当前字段映射进行格式化显示，点击后具有连线关系字段将显示为一行。 说明：此排序并不变更实际表内字段顺序。
字段配置	可单击**字段配置**手动添加字段名称及类型。 **说明：** 1. MySQL / Doris 已提供直接获取数据表结构能力，您可以直接使用或查看界面内已展示出的字段。 2. 源端提供 flink 函数对字段进行转换，可在添加字段内选择“函数”类型增加转换字段写入结果中。

注意：

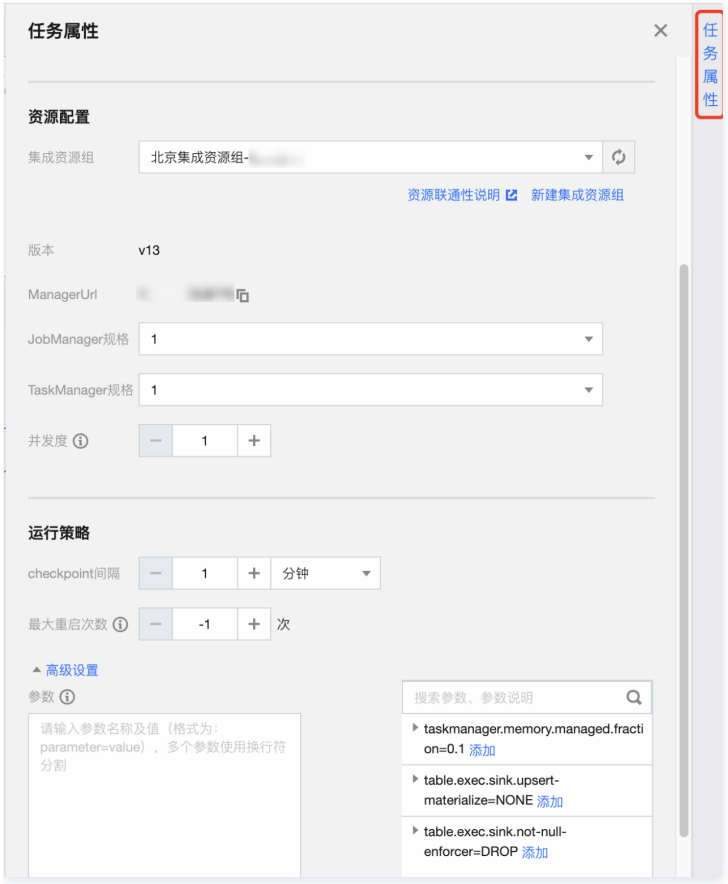
- 未配置映射关系的目标字段内容将为空。
- 若来源字段类型与目标字段类型间无法转换时，可能会导致任务失败。

MySQL > Doris字段映射说明如下：

类型	MySQL 数据类型	建议转化 Doris 数据类型	补充说明
	BOOLEAN	BOOLEAN	-
数值类型	TINYINT	TINYINT	-
	SMALLINT	SMALLINT	-
	MEDIUMINT	INT	-
	INT	INT	-

	BIGINT	BIGINT	-
	UNSIGNED TINYINT	SMALLINT	-
	UNSIGNED MEDIUMINT	INT	-
	UNSIGNED INT	BIGINT	-
	UNSIGNED BIGINT	LARGEINT	-
	FLOAT	FLOAT	-
	DOUBLE	DOUBLE	-
	DECIMAL	DECIMALV3	-
日期时间类型	YEAR	SMALLINT	-
	TIME	STRING	-
	DATE	DATEV2	-
	DATETIME	DATETIMEV2	-
	TIMESTAMP	DATETIMEV2	TIMESTAMP 字段数据会随着系统时区而改变但 DATETIME 字段数据不会，建议根据业务场景进行时区转化
字符串类型	CHAR	CHAR	-
	VARCHAR	VARCHAR	如果 MySQL 字段长度超过65533，建议转化为 string
	TINYTEXT、TEXT	STRING	-
	MEDIUMTEXT、LONGTEXT	STRING	MySQL 字段长度超过1048576 字节时可能精度丢失
二进制字符串	TINYBLOB、BLOB	STRING	-
	MEDIUMBLOB、LONGBLOB	STRING	MySQL 字段长度超过1048576 字节时可能精度丢失
	BINARY、VARBINARY	STRING	-
其他	JSON	STRING	MySQL 字段大小超过1M时可能精度丢失
	SET、BIT	STRING	-
	ENUM	UNSUPPORTED	暂不支持

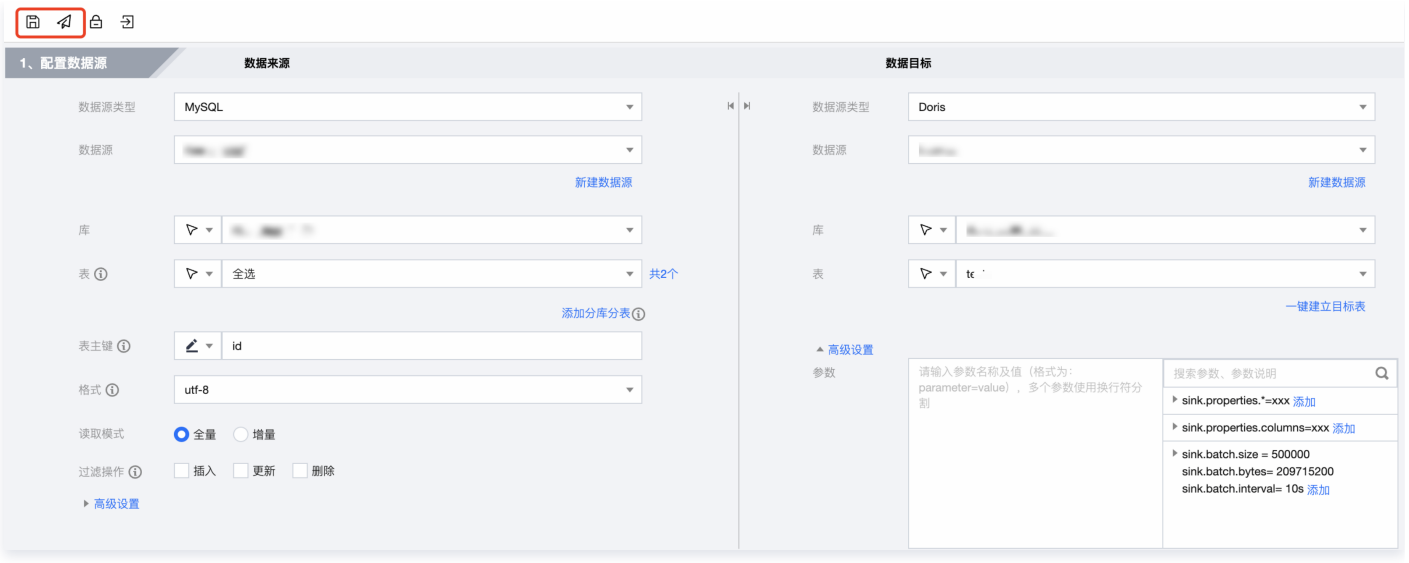
配置任务属性



参数	说明
集成资源组	可选择绑定此项目的集成资源组，一个任务仅可绑定一个资源组。 若未购买资源组或未绑定资源组，请先进行绑定操作。
JobManager 规格	支持0.25、0.5、1、2CU，设置后任务将默认占用此规格。 CU 任务实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度。
TaskManager 规格	支持0.25、0.5、1、2CU，设置后任务将默认占用此规格 。 CU 任务实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度.
并发度	每个算子的默认并行度，默认1。 任务实际占用CU数= JobManager 规格 + TaskManager 规格 × 并行度.
checkpoint 间隔	当前任务提交的最大 checkpoint 间隔.
最大重启次数	设置在执行过程中发生故障时任务最大的重启阈值，若运行中重启次数超过此阈值，任务状态将置为“失败”。设置范围为[−1,100]， 阈值为0表示不重启，−1 表示不限制最大重启次数.
高级设置	设置任务级别运行参数，具体参数说明请参见 实时节点高级参数 。

任务保存与提交

1. 配置完成后，单击页面左上角的**保存**按钮完成配置保存，再单击**提交**按钮完成作业启动。

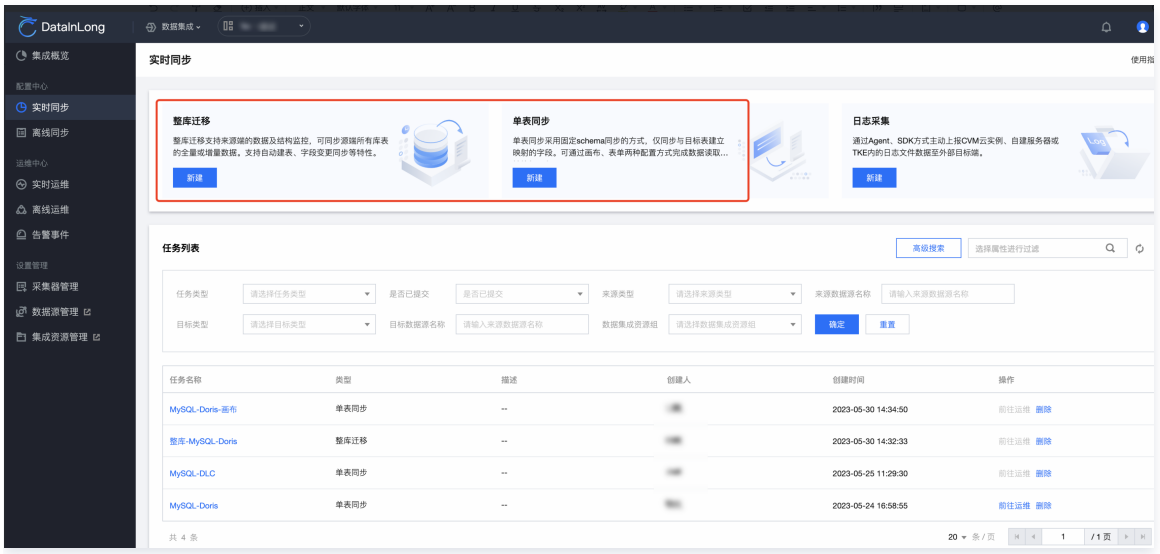


2. 作业启动前，会对必要配置进行校验，请确认无误后再提交。

配置整库实时迁移任务

最近更新时间：2024-08-06 18:01:21

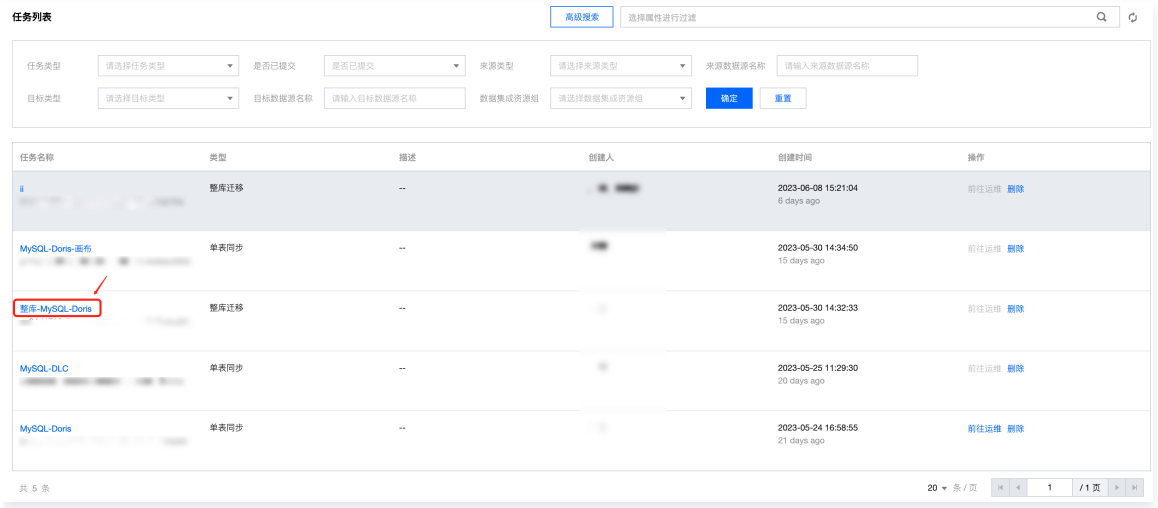
数据集成支持单表同步、整库迁移两种数据同步方式，下文将介绍整库实时迁移的操作方法。



创建整库同步任务

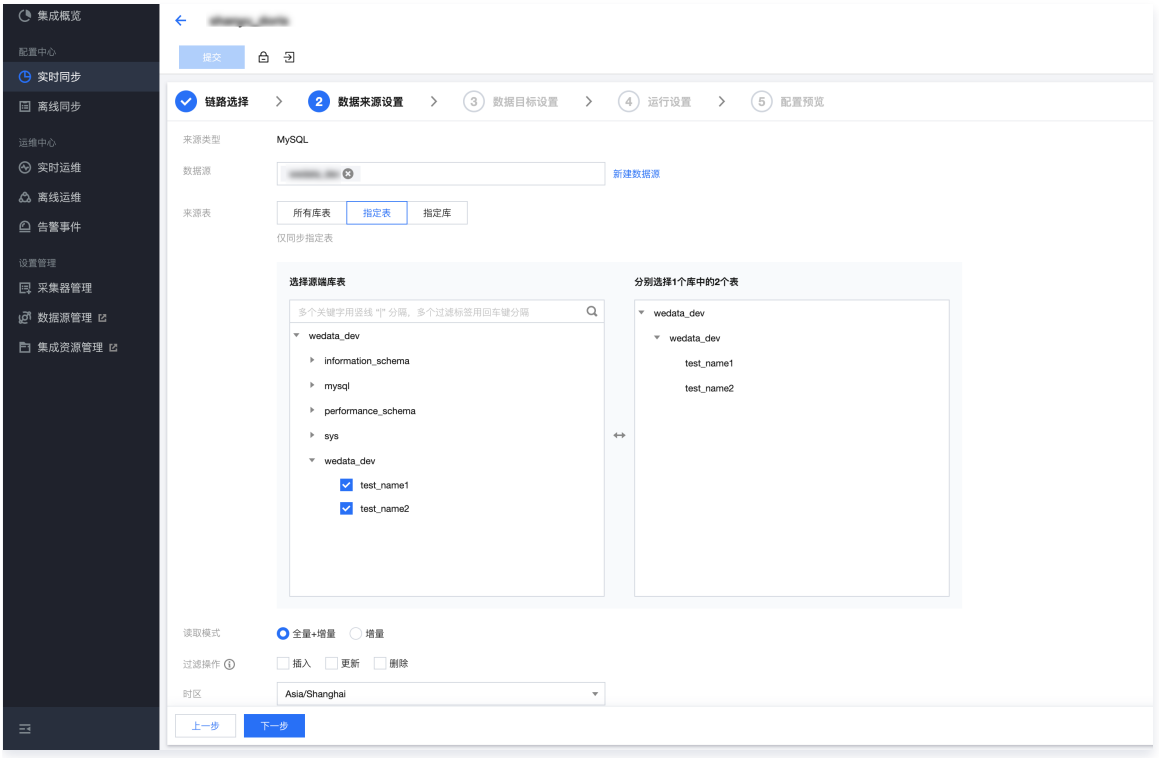
1. 进入数据集成 > 实时同步页面，单击新建创建整库迁移任务。

2. 创建完毕后，单击任务列表中的任务名称，即可进行具体配置。



选择同步至 Doris 目标端的链路

配置 MySQL 源端读取多张表



参数	说明
数据源	选择需要同步的已配置好的 MySQL 数据源。
来源表	• 所有库表：监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端。 • 指定表：此选项下需指定到具体表名称，设置后任务仅同步指定表。 • 指定库：此选项下需指定具体库名、以表名正则表达式，设置后，任务运行期间符合表名表达式的新增表默认将同步至目标端。示例如下： ○ 多表使用"，" 多表匹配 ○ ".*" 表示全表匹配 ○ "(a|b)_.*" 表示a_或者b_开头的表
读取模式	• 全量 + 增量：数据同步分为全量和增量同步阶段，全量阶段完成后任务进入增量阶段。全量阶段将同步库内历史数据，增量阶段从任务启动后 binlog cdc 的位点开始同步。 • 增量：仅从任务启动后的 binlog cdc 位点开始同步数据。
过滤操作	支持插入、更新和删除三种操作，设置后将不同步指定操作类型的数据。
时区	设置日志时间所属时区，默认上海。

设置 Doris 目标写入方式

参数	说明
数据源	选择已经创建的 Doris 数据源。
库/表匹配策略	设置任务运行时 Doris 中数据库以及数据表对象的名称匹配规则： • 与来源库/表同名：任务运行时系统将默认在目标数据源内匹配与来源库/表同名对象。 • 自定义：自定义规则支持设置来源与目标之间特殊关系，例如，统一将源端库名或表名加上统一固定前缀或者后缀在写入目标库或表任务运行时。此策略下，任务运行时系统将默认根据命名规则匹配目标对象。
高级设置 – 参数	设置 Doris 写入端的运行参数，此参数可根据业务需求配置。 Doris 端已支持参数详情请参见实时节点高级参数。

配置运行资源和策略

- **集成资源配置**

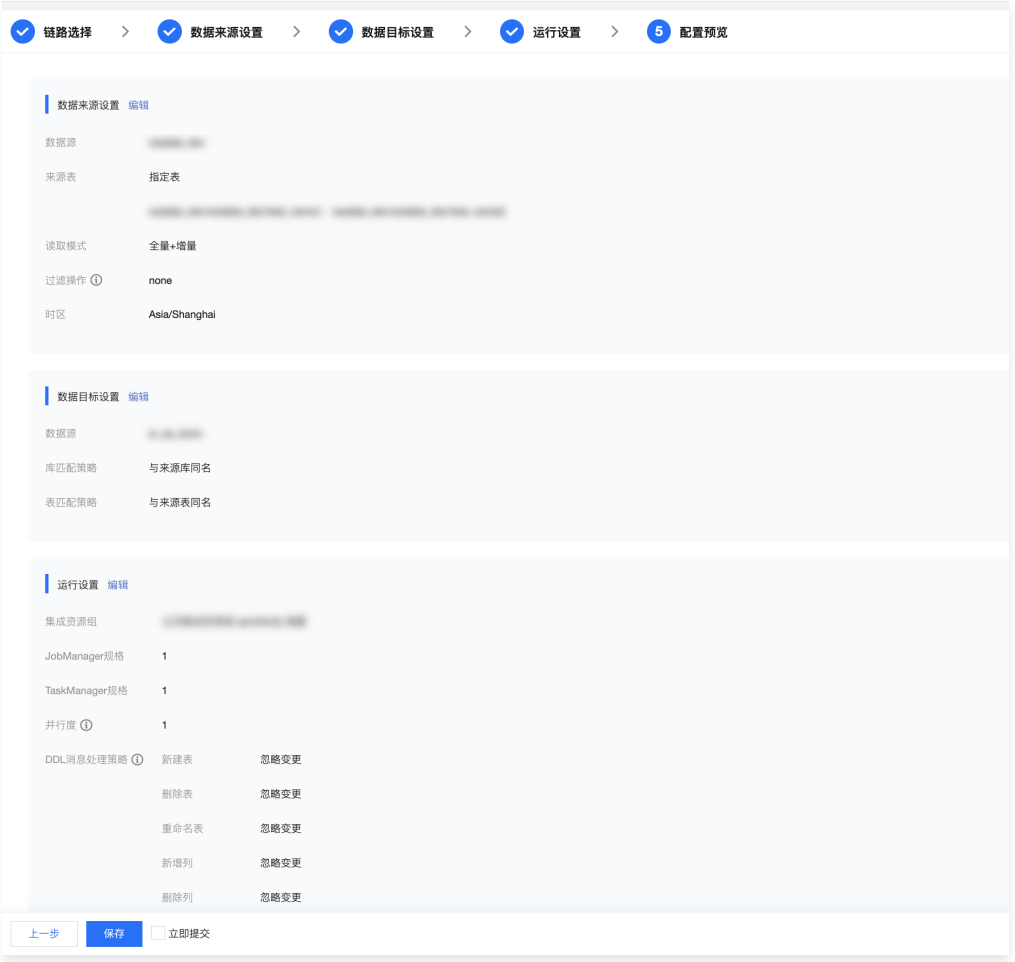
  为当前任务关联的集成资源组，同时设定运行时 JM、TM 规格以及任务运行并行度。其中，当前任务实际运行时实际占用 CU 数= JobManager 规格 + TaskManager 规格 × 并行度。
- **消息处理策略**

参数	策略名	策略说明
DDL 消息处理	新建表	1. 自动建表：当来源端被监控的库中出现新建表时，Doris 端将自动创建同结构的表及字段： ○ 若来源端表包含主键，任务默认创建 Unique key 模型表。 ○ 若来源端表包含主键，任务默认创建 Duplicate 模型表。 2. 忽略变更：目标端忽略来源端的产生的 DDL 变更消息，Doris 端及日志不做任何响应或消息提醒。 3. 日志告警：目标端仅接收 DDL 变更消息，并在日志内打印消息内容，不触发新建表操作。 4. 任务出错：目标端接收 DDL 变更消息并持续重启任务，重启过程中任务日志报错并出现数据写入异常。
	新增列	1. 新增列：当来源端被监控的库中出现表增加字段时，Doris 端将自动同步新增同名字段。 2. 忽略变更：目标端忽略来源端的产生的 DDL 变更消息，Doris 端及日志不做任何响应或消息提醒。 3. 日志告警：目标端仅接收 DDL 变更消息，并在日志内打印消息内容。此策略并不触发新增列操作。 4. 任务出错：目标端接收 DDL 变更消息并持续重启任务，重启过程中任务日志报错并出现数据写入异常。
	删除表 重命名表 删除列	除新建表、新增字段外其他 DDL 变更消息不支持自动响应，目前提供忽略变更、日志告警、任务出错三种策略选择： 1. 忽略变更：目标端忽略来源端的产生的 DDL 变更消息，Doris 端及日志不做任何响应或消息提醒。 2. 日志告警：目标端仅接收 DDL 变更消息，并在日志内打印消息内容。此策略并不触发新建表操作。

	重命名列	3. 任务出错：目标端接收 DDL 变更消息并持续重启任务，重启过程中任务日志报错并出现数据写入异常。
	修改列	
	删除列	
写入异常	部分停止	数据无法写入目标表时丢弃数据，后续该异常表对应的数据自动丢弃不再同步。
	异常重启	任意表数据写入异常后任务将异常退出并自动重启。重启后任务将持续尝试写入，直到所有表均可正常同步。重启期间可能导致部分表数据重复写入。
	忽略异常	忽略表内无法写入的异常数据并标记为脏数据，任务继续读取并写入剩下的数据。
脏数据	COS 归档	写入异常策略配置为 忽略异常 时，将未写入至目标端的数据同步写入到指定的 COS 桶及文件内。
	不归档	不归档保存未写入的异常的数据。

配置预览及提交

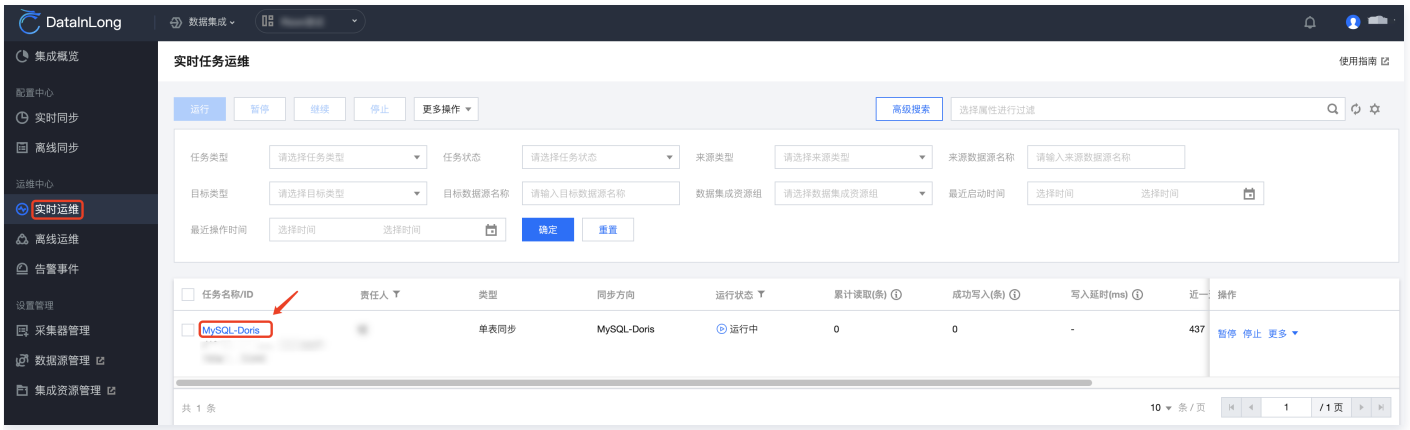
配置完成后可进行预览，确认后单击**保存**。



任务运维

最近更新时间：2024-07-25 11:24:21

提交任务以后，可进入数据集成 > 实时运维页面查看并监控当前任务状态、读写指标统计、日志及配置当前任务监控规则。

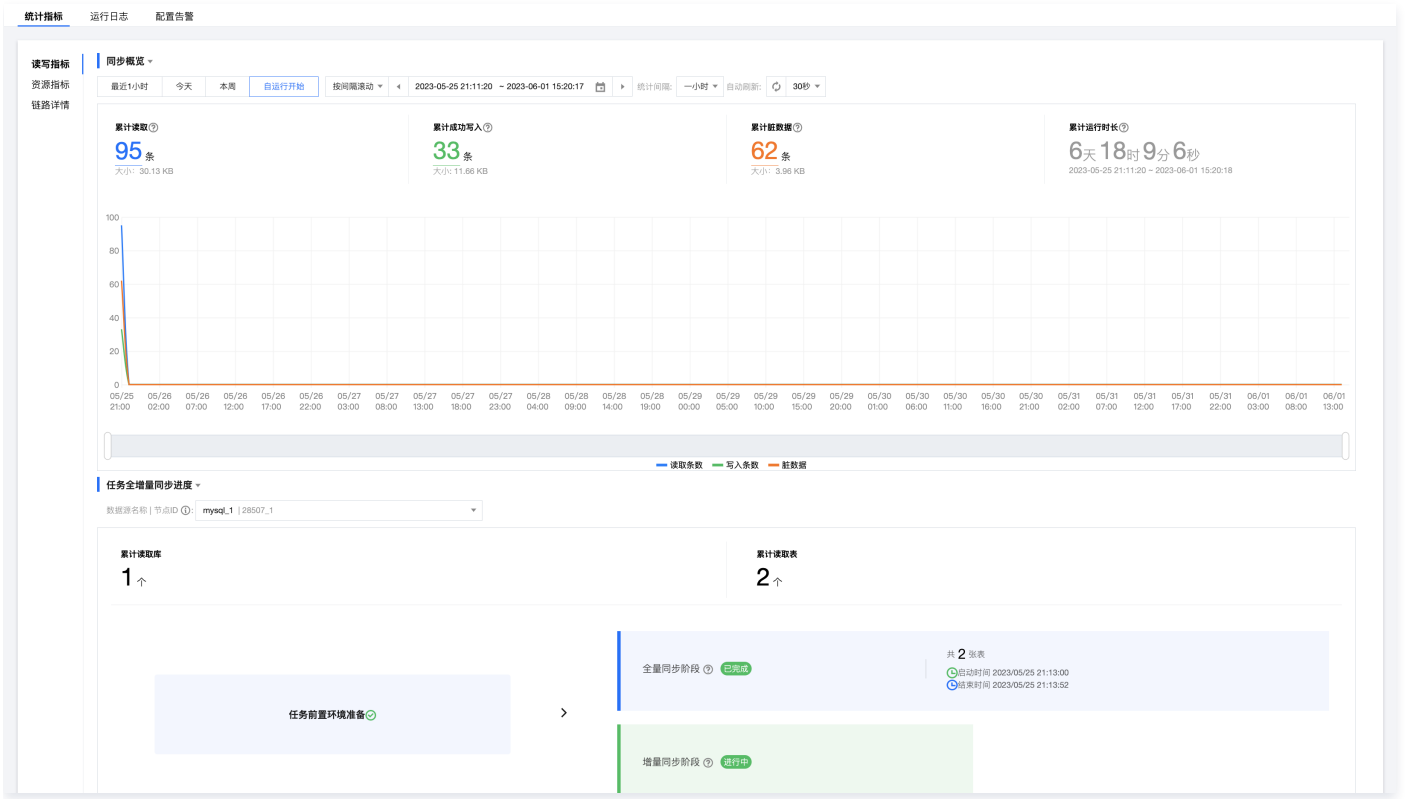


统计指标

统计指标页面展示了任务内读写及资源运行情况。

读写指标

展示当前任务读写整体条数、全增量同步阶段、以及读写速度等。

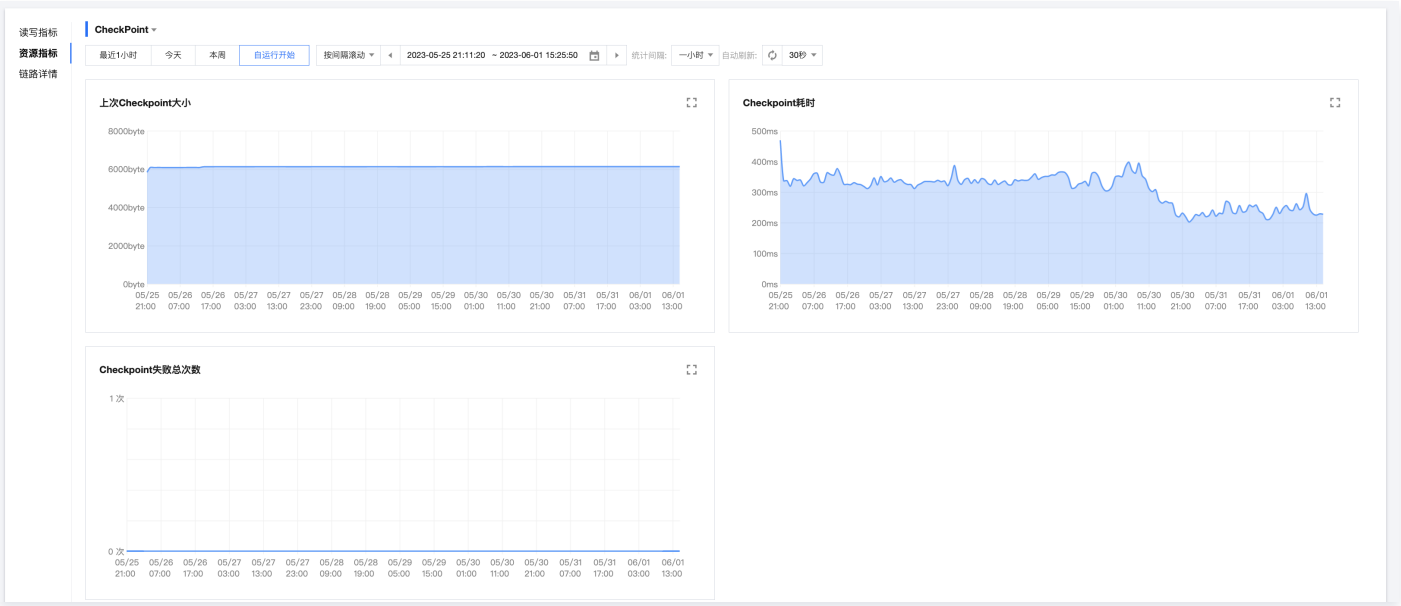


	指标参数	说明
同步概览	累计读取	本次任务运行期间，从来源端实际读取数据条数。此指标不包含筛选过滤等方式剔除的数据总量。
	累计成功写入	本次任务运行期间，已读取的数据中成功写入到目标端的数据总量。
	累计脏数据	本次任务运行期间，已读取的数据中异常写入失败的数据总量。此指标不包含任务配置中主动忽略/过滤而导

		致未写入的数据，包括指定部分停止、异常重启等运行策略，以及数据过滤等。
	累计运行时长	本次任务启动后，累计总运行时长（包含暂停时间）。
	累计读取库	本次任务运行期间，从来源端实际读取数据库数量。
	累计读取表	本次任务运行期间，从来源端实际读取数据表数量，并且分别全量同步阶段和增量同步阶段数量。
全增量同步进度	全量/增量状态	提供未启动、进行中和已完成三种状态。
	全量同步阶段	读取源端库表中的所有记录，本阶段内仅统计读取成功且有存量业务数据的表，并且同步展示增量启动时间、统计时间、全量结束时间。
	增量同步阶段	从 binlog 消费变更数据，本阶段内仅统计读取成功且有新增业务数据的表，并且同步展示增量启动时间。
	读取速度	读取速度 = 统计间隔内总读取条数 / 统计间隔。
	读取吞吐	读取吞吐 = 统计间隔内总读取总量 / 统计间隔。
	写入速度	写入速度 = 统计间隔内成功写入条数 / 统计间隔。
读写详情	写入吞吐	写入吞吐 = 统计间隔内成功写入总量 / 统计间隔。
	写入延时	来源 Source 端至写入 Sink 端之间的链路延迟，写入延时 = 系统时间 − 记录读取时间（读取端 LatencyMarker 时间戳）。
	作业重启次数	统计间隔内当前任务重启次数。

资源指标

展示当前任务使用资源情况。



	指标参数	说明
CheckPoint	上次 Checkpoint 大小	当前作业最近一次的 Checkpoint 大小。
	Checkpoint 耗时	当前作业的 Checkpoint 耗时。
	Checkpoint 失败总次数	当前作业的 Checkpoint 的失败总次数。
TaskManager	TaskManager CPU 使用率	当前作业 TaskManager 的 CPU 使用率。
	TaskManager 堆内存使用量	当前作业 TaskManager 堆内存的用量。
	TaskManager 老年代总 GC 次	当前作业 TaskManager 老年代 GC 次数。

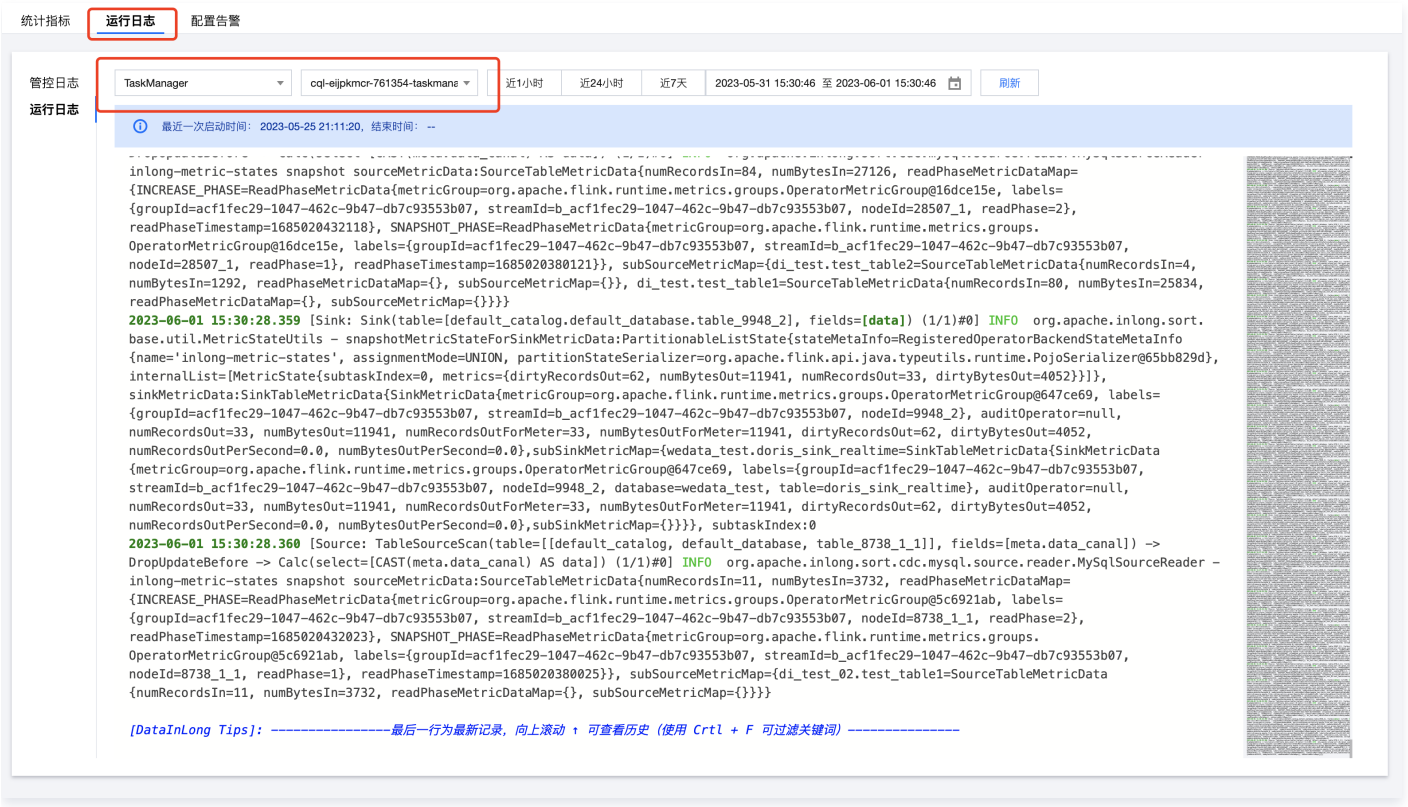
	数	
	TaskManager 老年代总 GC 时间	当前作业 TaskManager 老年代 GC 时间。
	TaskManager 物理内存用量	当前作业 TaskManager 所在的 JVM 的物理内存用量（RSS），包括堆内、堆外、Native 等所有区域的总内存用量。
JobManager	JM CPU Load	TaskManager 维度的 JVM 最近 CPU 利用率。
	JM Head Memory	TaskManager 维度的堆内存使用情况。
	JM GC Count	TaskManager 维度的 Status.JVM.GarbageCollector.<GarbageCollector>.Count，GC（垃圾回收）次数。
	JM GC Time	TaskManager 维度的 Status.JVM.GarbageCollector.<GarbageCollector>.Time，GC（垃圾回收）时间。

链路详情

展示整库任务下每张表的读写情况（仅整库同步时会展示此页面）。

运行日志

展示运行日志。



配置告警

配置告警页面支持对实时任务创建监控规则及告警渠道。

常见问题

最近更新时间：2024-07-25 11:24:21

Doris 规格如何选型及调优?

请参考 Doris 资源规格选型及调优建议 。

导入任务过多，新导入任务提交报错 “current running txns on db xxx is xx, larger than limit xx ”？

调整 fe 参数：max_running_txn_num_per_db，默认100，可适当调大，建议控制在500以内。

导入频率太快出现 err=[E-235] 错误?

- 参数调优建议：可通过适当调大 max_tablet_version_num 参数暂时解决，此参数默认200，建议控制在2000以内。
- 业务调优建议：降低导入频率才能根本解决这个问题。

导入文件过大，被参数限制。报错 “The size of this batch exceed the max size ”？

调整 be 参数：streaming_load_max_mb，建议超过需要导入的文件大小。

导入数据报错：“[-238]”？

- 原因：-238错误通常出现在同一批导入数据量过大的情况，从而导致某一个 tablet 的 Segment 文件过多（由 ）。
- 参数调优建议：可适当调大 BE 参数 max_segment_num_per_rowset ，此参数默认值200，可按倍数调大（如400、800），建议控制在2000以内；
- 业务调优建议：建议减少一批次导入的数据量。

导入失败，报错：“too many filtered rows xxx, "ErrorURL":"或 Insert has filtered data in strict mode, tracking url=xxxx.”？

原因：表的 schema、分区等与导入的数据不匹配。可在 CDW Studio 或客户端执行 doris 命令查看具体原因：show load warnings on `<tracking url>`，<tracking url> 即为报错信息中返回的 error url。