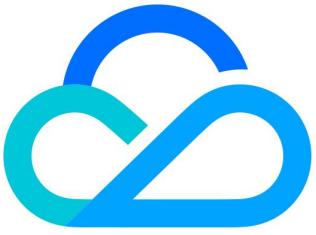
数据集成

操作指南



腾讯云

文档目录

操作指南

概述

最近更新时间：2024-07-18 17:43:21

数据集成通过快速连接和融合云上或云下自建的各种数据，解决数据平台构建、数据库迁移备份，以及业务升级、整合，数据访问加速、全文检索等多个场景中数据整合和同步问题。腾讯云产品 数据集成（DataInLong）为 WeData 提供数据集成能力，支持离线同步、实时数据库监控、数据上报等同步能力。

使用限制

1. **数据同步**：数据集成仅支持传输能够抽象为逻辑二维表的数据对象，如结构化、半结构化、无结构化（COS 等，要求具体同步数据必须抽象为结构化数据）的数据内容同步。FTP 方式支持将完全非结构化的文件（例如气象文件）同步至 HDFS，但此传输方式不支持数据内容提取。
2. **网络连通**：支持单地域内及部分跨地域的数据存储相互同步、交换的数据同步需求。 部分地域之间可以通过经典网络传输，但不能保证其连通性。如果测试经典网络不通，建议您使用公网方式进行连接。
3. **任务运行**：运行数据集成任务需要使用数据集成资源组，在使用数据集成功能前请先完成集成资源组创建。集成资源组包含离线包、实时包等，可根据需运行任务类型按需购买。
4. **数据一致性**：数据集成同步支持 at least once，不完全保证支持 exactly once (即不能保证数据完全不重复)。数据完全不重复需依赖主键 + 目的端能力来保证。
5. **数据类型及精度**：离线或者实时同步中，同步任务源端和目标端字段需要注意类型匹配及精度转换。若来源与目标端类型无法兼容，或目标端字段类型最大值小于源端最大值（或最小值大于源端最小值，或精度低于源端精度），可能会导致写入失败或精度被截断的风险。

离线同步

数据集成提供离线数据同步能力，该能力通过定期运行方式批量读取来源库表中数据并同步写入至目标端。详情请参考 离线同步 。

实时同步

数据集成提供实时数据同步能力，该能力支持流式数据传输。实时同步下支持单表、分库分表、多库多表粒度的实时数据消费，任务类型包括单表同步、整库同步、以及日志采集。

- **单表同步**：来源端为单张表或分库分表，目标端仅支持一张表。单表同步采用固定schema搭配的方式，需在任务中指定来源与目标表之间的字段映射关系，任务运行时仅将指定来源字段内容写入至目标字段。详情请参考 单表同步任务配置 。
- **整库同步**：整库同步支持将来源端整个实例、或者指定的多个库表对象内的全部数据同步至目标端的多张表中。此任务无需指定来源与目标端之间的字段映射关系，默认所有来源表字段全部读取，且表与表之间字段默认同名匹配。详

情请参考 整库同步任务配置 。

- **日志采集**：日志采集通过 Agent、SDK 方式主动上报 CVM 云实例、自建服务器或 TKE 内的日志文件数据至外部目标端。详情请参考 日志采集任务配置 。

基本概念

- **数据源**

  数据集成过程中使用数据源作为读取/写入的目标对象，数据源可以是一个数据库或者是一个数据仓库（EMR 引擎实例等）。在数据集成同步任务配置前，您需要在数据源管理页面配置好需要同步的源端和目标端数据库或数据仓库的相关信息，配置好后可在同步任务中通过选择数据源名称来控制同步读取和写入的数据库或数据仓库。
- **网络联通性**

  使用数据集成同步任务之前，需要保证数据源网络（包括读端、写端）与数据集成资源组之间网络互通，且资源不可因为白名单限制等原因被拒绝访问，否则无法完成数据传输同步。详情参考 集成连通性与使用规划 。
  - 若数据源开通公网：需要购买并创建 NAT 网关，允许集成资源通过网关连通数据源所在 VPC，详细操作请参见 NAT 网关 相关文档。
  - 若数据源处于 VPC 内：
    - 若与集成资源位于同一 VPC：可直接使用。
    - 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：需购买 VPN 或 专线网关 打通集成与数据源所在 VPC。
- **限速**

  限速是数据集成同步任务允许达到的最大传输速度限制。
- **并发数**

  并发数是数据同步任务中，最大并行读取或并行写入数据数。并发数影响数据同步的效率，并发设置越高对应资源消耗也越多，由于资源原因或者任务本身特性等原因，实际执行时并发数可能小于等于此值。
- **脏数据**

  脏数据是指在同步过程中由于字段类型不匹配、或者写入目标数据源发生了异常等情况导致写入失败的数据。 所有写入失败的数据均被归类于脏数据。例如，源端是 String 类型的数据写到 INT 类型的目标字段中，因为类型转换不合理而无法写入的数据。

  离线同步中，您可以在任务中配置脏数据阈值以控制同步过程中最大脏数据条数。当任务超过此阈值以后，任务将被中断运行。

  实时同步中，您可以配置脏数据归档方式，将写入失败的脏数据统一写入归档存储中以保证实时数据流不中断。

数据源配置与管理
支持的数据源与读写能力
读写端 DML 及 DDL 操作

最近更新时间：2024-06-12 10:03:31

DML 及 DDL 支持情况概览

说明：

1. 读：“是” 表示读端可感知对应 DDL 操作，传递变更给下游。
2. 写：“是” 表示写端可跟随响应对应 DDL 变更。

类型	名称	DML			DDL							
		插入	删除	更新	新增列	删除列	更新列名	修改列类型	新增表	修改表名	删除表	清空表
读	MySQL	是	是	是	是	是	是	是	是	是	是	是
写	DLC	是	是	是	是	否	否	否	是	否	否	否
	Iceberg	是	是	是	是	否	否	否	是	否	否	否
	Doris	是	是	是	是	否	否	否	是	否	否	否
	Kafka	是	否	否	NA	NA	NA	NA	NA	NA	NA	NA

DDL 解析语法要求

注意：

> 仅符合 DDL 类型及语法要求的 SQL 可被源端解析并响应，不支持的 SQL 语句可能会导致任务异常重启等情况。请评估业务表结构变更常用语法是否符合要求，并为任务配置符合业务要求的 DDL 变更响应及写入异常处理策略。

MySQL：

1. 重命名表（rename table）支持下列语法：

```
RENAME TABLE tbl_name TO new_tbl_name
```

2. 新建表（create table）支持下列语法：

```
CREATE TABLE [IF NOT EXISTS] tbl_name (create_definition,...)
CREATE TABLE new_tbl LIKE orig_tbl;
CREATE TABLE new_tbl AS SELECT * FROM orig_tbl;
```

> **警告：**
> 不支持 CHECK \ TemporaryTable 等约束。

> **说明：**
> 创建表时支持建表语句中设置字符集，例如 utf8 等。

3. 修改列（modify \ change column）与 重命名列（rename column）支持下列语法：

```
ALTER TABLE t1 MODIFY b INT NOT NULL;
ALTER TABLE t1 CHANGE b a INT NOT NULL;
ALTER TABLE t1 CHANGE COLUMN b a INT NOT NULL;
```

4. 增加列（add column）/ 重命名列 支持下列语法：

```
ALTER TABLE table_name
  ADD new_column_name column_definition
    [ FIRST | AFTER column_name ];
```

5. 删除列（drop column）支持下列语法：

```
Alter table t2 DROP [COLUMN] col_name
```

6. 清空表（truncate table）支持下列语法：

```
truncate table a;
```

7. 删除表（drop table）支持下列语法：

```
drop table a;
```

支持数据源读写情况

最近更新时间：2024-07-12 11:38:31

数据集成包括离线与实时数据链路，目前支持关系型数据库、大数据存储、半结构化数据、NoSQL、消息队列等各类数据源。您可以根据各个模块对数据源的支持情况，选择对应的功能模块进行同步任务的配置。

数据源		离线同步		实时同步					版本
		离线读取	离线写入	单表读取	单表写入	整库读取	整库写入	日志采集	
关系型数据库	MySQL	✓	✓	✓	✓	✓	-	✓	5.6，5.7，8.0.x
	TDSQL-C MySQL	✓	✓	-	✓	✓	-	✓	-
	PostgreSQL	✓	✓	✓	✓	✓	✓	✓	9.6，10，11，12
	TCHouse-P	✓	✓	-	✓	-	✓	✓	
	SQL Server	✓	✓	✓	✓	✓	-	✓	2012，2014，2016，2017，2019
	Oracle	✓	✓	✓	✓	✓	-	✓	11，12，19
	DB2	✓	✓	-	-	-	-	-	-
	SAP HANA	✓	✓	-	-	-	-	-	-
	DM	✓	✓	-	-	-	-	-	-
	SAP IQ(Sybase)	✓	-	-	-	-	-	-	-
大数据	Hive	✓	✓	-	✓	-	✓	✓	1.x，2.x，3.x
	DLC	✓	✓	-	✓	-	✓	✓	-

	Doris	✓	✓	–	✓	–	✓	✓	0.15+
	StarRocks	✓	✓	–	–	–	✓	–	–
	ClickHouse	✓	✓	–	✓	–	–	✓	20.7+
	Iceberg	✓	✓	–	✓	–	✓	✓	0.13.1+
	HBase	✓	✓	–	✓	–	–	✓	2.2.x
	HDFS	✓	✓	–	✓	–	–	✓	2.x，3.x
	Kudu	✓	✓	–	–	–	–	–	–
	TBase	✓	✓	–	✓	–	–	✓	–
	GaussDB	✓	✓	–	–	–	–	–	–
	GBase	✓	✓	–	–	–	–	–	–
	Greenplum	✓	✓	–	✓	–	–	✓	4.x，5.x，6.x
半结构化	COS	✓	✓	–	–	–	–	–	–
	Rest API	✓	–	–	–	–	–	–	–
	FTP	✓	✓	–	–	–	–	–	–
	SFTP	✓	✓	–	–	–	–	–	–
NoSQL	Redis	–	✓	–	–	–	–	–	–
	Elasticsearch	✓	✓	–	✓	–	–	✓	6.x，7.x
	Mongo	✓	✓	✓	–	✓	–	–	–
消息队列	Kafka	✓	✓	✓	✓	✓	✓	✓	–
	TiDB-Kafka	–	–	✓	–	–	–	–	–
	DTS-kafka	–	–	✓	–	–	–	–	–

数据上报	SDK	–	–	–	–	–	–	✓	–
	TKE	–	–	–	–	–	–	✓	–
	CVM	–	–	–	–	–	–	✓	–

说明

- 实时数据链路读写端，需满足对应数据源版本。
- 蓝色 √ 为可跳转的超链接，您可单击蓝色 √，进入对应文档页面进行查看。

数据源配置与数据库环境准备

数据库环境概述

最近更新时间：2024-06-18 14:47:41

在配置数据集成同步任务前，您需要配置好需要同步的数据源端和目标端数据库相关信息，以便在配置同步任务时，您可以通过选择数据源名称来确定同步任务读取和写入的数据库。

例如，同步任务配置过程中在目标端数据库创建与源端同名的 schema 时，需要数据源配置的账号拥有创建 schema 的权限，每个数据库对应操作需要的授权命令不同，具体可以参考对应数据库语法，避免任务配置或运行过程中出现 Access denied 或 permission denied 之类的报错。

数据集成统一使用**项目管理**模块内配置的数据源信息，您可以在**数据源管理**页面内配置数据源名称、链接、分享权限等。具体操作方式及使用说明请参见 数据源管理 。

数据安全与使用

数据源包含数据 JDBC 连接、数据库、账号名及密码等安全敏感信息，项目管理内仅支持项目管理员等角色创建数据源。请谨慎修改数据源名称、链接、数据库及账号密码等敏感信息，以防止引起关联该数据源的数据集成任务运行失败。

数据默认数据对象使用与配置

大部分自定义数据源（例如 MySQL 等）在配置数据源对象时候需指定数据库及账号密码，该数据库为数据集成内默认拉取以及同步的数据库对象。

数据源所属及授权项目

如上图所示，数据集成支持使用归属或授权本项目空间/当前用户的数据源。

各类型数据库环境准备操作不同，请根据您同步任务的来源端与目的端，选择对应的文档进行授权。

- MySQL 环境准备与数据库配置

MySQL 环境准备与数据库配置

最近更新时间：2024-04-18 17:34:21

数据集成提供了 MySQL 的读取和写入能力，本文为您介绍使用 MySQL 进行实时数据同步的前置环境配置以及当前能力支持情况。

支持版本

目前数据集成已支持 MySQL 单表及整库级实时读取，使用实时读取能力需遵循以下版本限制：

类型	版本	Driver
MySQL	5.6，5.7，8.0.x	JDBC Driver：8.0.21
RDS MySQL	5.6，5.7， 8.0.x	
PolarDB MySQL	5.6，5.7，8.0.x	
Aurora MySQL	5.6，5.7，8.0.x	
MariaDB	10.x	
PolarDB X	2.0.1	

使用限制

- 需要开启 Binlog 日志，仅支持同步 MySQL 服务器 Binlog 配置格式为 ROW。
- 无主键的表由于无法保证 exactly once 可能会有数据重复，因此实时同步任务最好保证有主键。
- 不支持 XA ROLLBACK，实时同步的任务不会针对 XA PREPARE 的数据进行回滚的操作，若要处理 XA ROLLBACK 场景，需要手动将 XA ROLLBACK 的表从实时同步任务中移除，再添加表后重新进行同步。

数据库环境准备

确认 MySQL 版本

数据集成对 MySQL 版本有要求，查看当前待同步的 MySQL 是否符合版本要求。您可以在 MySQL 数据库通过如下语句查看当前 MySQL 数据库版本。

```
select version();
```

设置 MySQL 服务器权限

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。

1. **创建 MySQL 用户（可选）：**

```
mysql> CREATE USER 'user'@'localhost' IDENTIFIED BY 'password';
```

2. 向用户授予所需的权限:

在实时数据同步的情况下，该账号必须拥有数据库的 SELECT、REPLICATION SLAVE 和 REPLICATION CLIENT 权限。执行命令可以参考下面:

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

> ⚠ **注意:**
> 启用 scan.incremental.snapshot.enabled 时不再需要 RELOAD 权限（默认启用）。

3. 刷新用户的权限:

```
mysql>FLUSH PRIVILEGES;
```

查看更多关于 权限说明 。

开启 MySQL Binlog

1. 检查 Binlog 是否开启

```
show variables like "log_bin"
```

返回结果为 ON 时，表示已经开启 Binlog， 如果为备库，使用如下语句:

```
show variables like "log_slave_updates";
```

如果返回为 ON 时，表示已经开启 Binlog，如果已经开启 Binlog，可跳过下面流程。

2. 开启 Binlog

如果确认没有开启 Binlog，则需要进行以下操作:

- 对于腾讯云实例 MySQL / TDSQL-C MySQL，默认开启了 Binlog。
- 对于开源 MySQL，参考官方文档开启 Binlog。

3. 修改 Binlog 格式为 Row

实时同步仅支持同步 MySQL 服务器 Binlog 配置格式为 ROW，使用如下语句查询 Binlog 的使用格式。

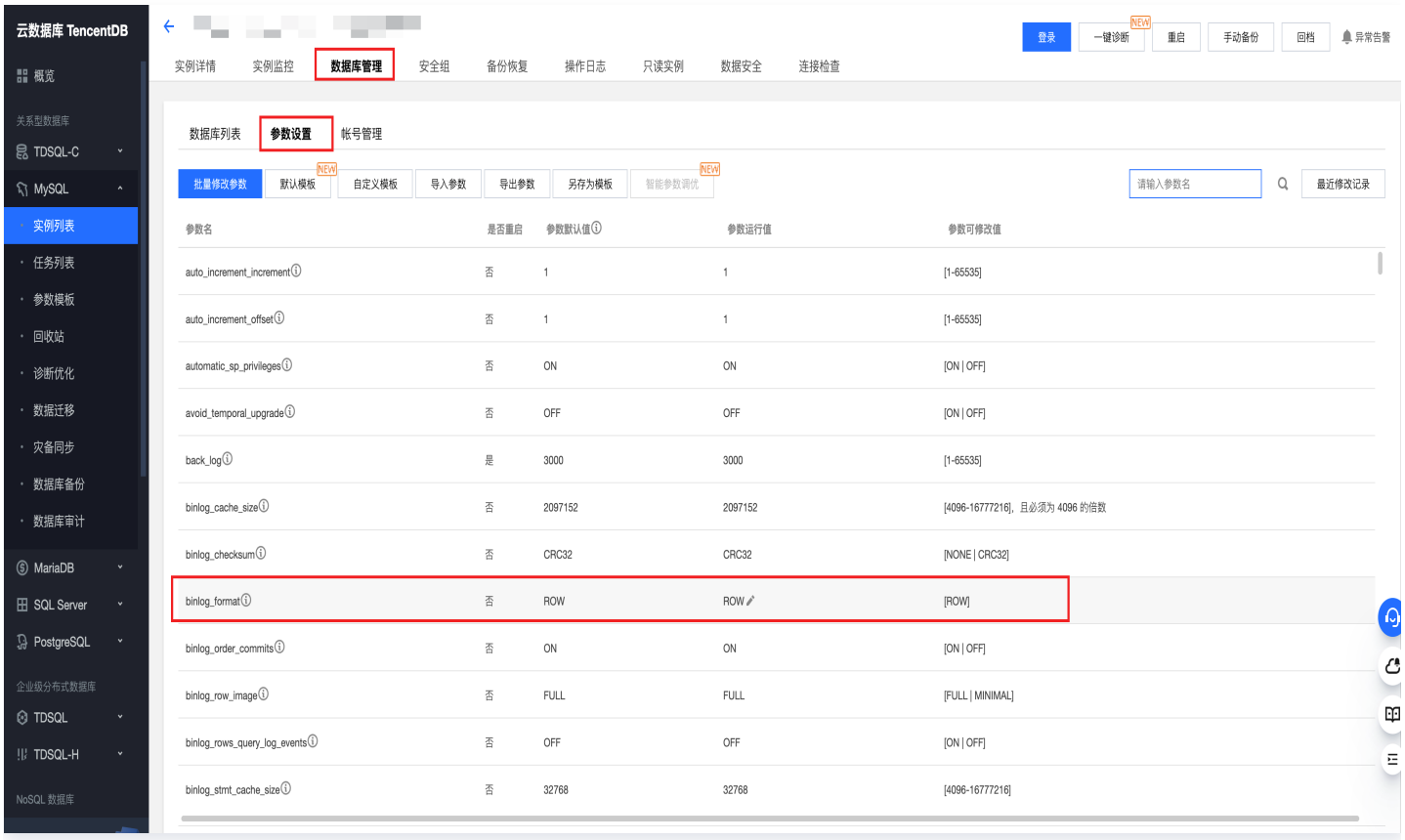
```
show variables like "binlog_format";
```

如果返回非 ROW 请修改 Binlog Format。

- 对于开源 MySQL，参考官方文档：

MySQL :: MySQL 8.0 Reference Manual :: 17.1.6.4 Binary Logging Options and Variables

- 对于腾讯云实例 MySQL / TDSQL-C MySQL ：
  - 登录腾讯云 MySQL / TDSQL-C MySQL 控制台，找到要开启 Binlog 的实例，点击进入该实例的详细信息页面。
  - 在上面选项卡中选择**数据库管理**，找到**参数设置标签页**。
  - 在**参数设置**选项卡中，找到 **binlog_format** 参数，将其设置为 “ROW” 。



4. binlog_row_image

实时同步仅支持同步 MySQL 服务器 binlog_row_image 配置格式为 FULL or full。
使用如下语句查询 binlog_row_image 的使用格式。

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

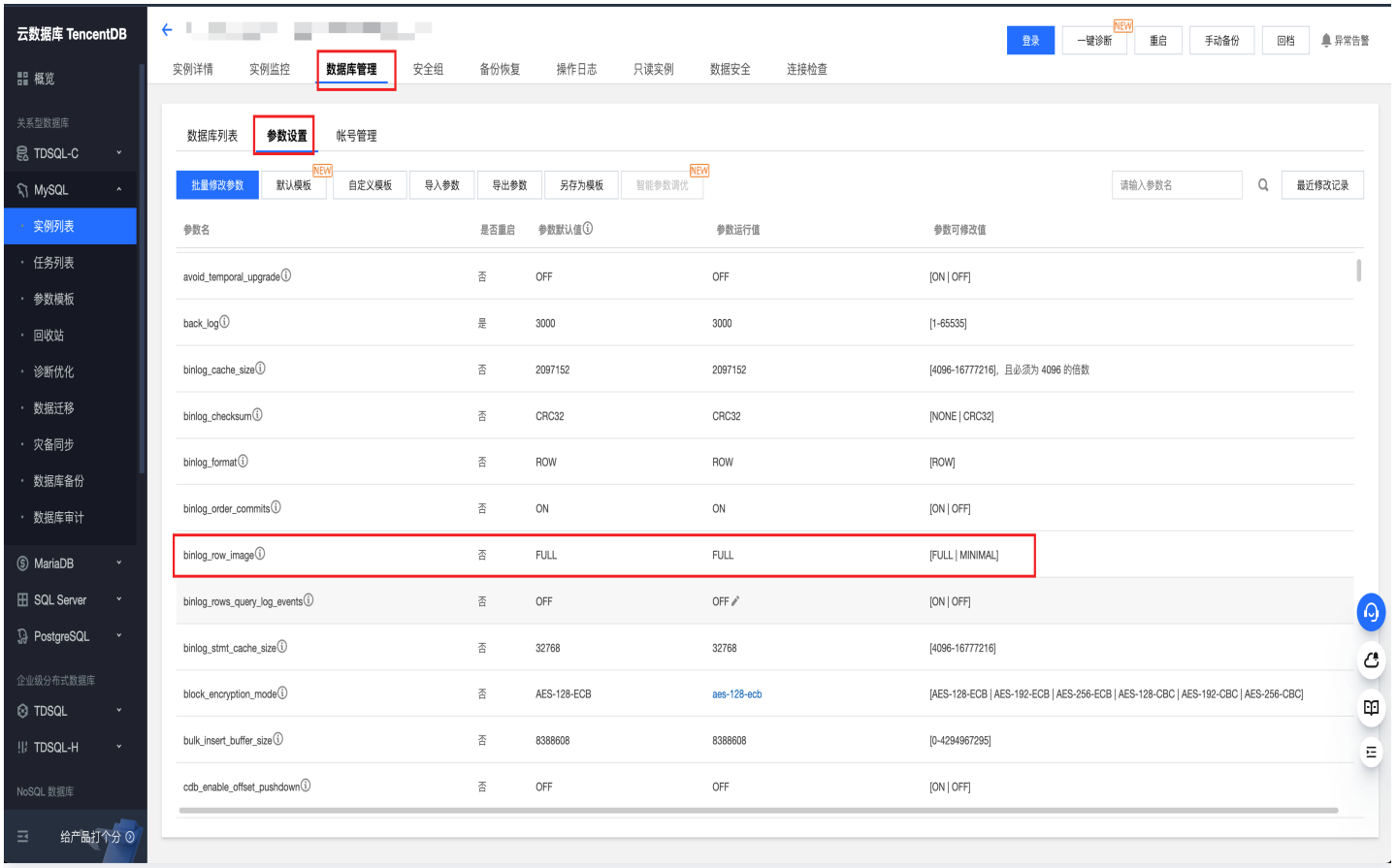
如果返回非 FULL/full 请修改 binlog_row_image：

- 对于开源 mysql，参考官方文档：

MySQL :: MySQL 8.0 Reference Manual :: 17.1.6.4 Binary Logging Options and Variables

- 对于腾讯云实例 MySQL / TDSQL-C MySQL ：

- 登录腾讯云 MySQL / TDSQL-C MySQL 控制台，找到要开启 Binlog 的实例，点击进入该实例的详细信息页面。
- 在上面选项卡中选择**数据库管理**，找到**参数设置**标签页。
- 在**参数设置**选项卡中，找到 **binlog_row_image** 参数，将其设置为“FULL”。



开启 GTIDs（可选）

GTID（Global Transaction Identifier， 全局事务标识），用于在 binlog 中唯一标识一个事务，使用 GTID 可以避免事务重复执行导致数据混乱或者主从不一致。

开启流程

1. 检查是否开启了 GTID。

```
show global variables like '%GTID%';
```

返回结果类似如下，证明已经开启 GTID。

```
+--------------------------+-------+
| Variable_name                    | Value |
+--------------------------+-------+
```

```
| enforce_gtid_consistency   | ON     |
| gtid_mode                          | ON     |
+---------------------------+-------+
```

2. 开启 GTID

- 对于开源 MySQL 参考官方文档 MySQL :: MySQL 8.0 Reference Manual :: 17.1.4.2 Enabling GTID Transactions Online 。
- 对于腾讯云实例 MySQL / TDSQL-C MySQL，默认为开启，不支持关闭。

数据源配置

进入配置数据源界面，MySQL 数据源支持云实例和连接串两种连接方式。

单击 **项目管理 > 数据源管理 > 新建数据源 > 选择 MySQL 数据源**。

- 通过云实例创建数据源。

参数	说明
连接类型	选择云实例或连接串的数据源连接形式
所属项目	当前数据源创建时的归属项目
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改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	需要连接的数据库名称
用户名	连接数据库的用户名称
密码	连接数据库的密码
数据连通性	测试是否能够连通所配置的数据库 说明： 若连通性测试不通过，数据源仍可保存。连通性测试未通过而保存但数据源不可使用 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

- **通过连接串创建数据源。**

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改数据源仅创建人和项目管理员可用。

部署方式	支持 CDB、自建实例、公网实例三种部署方式。
区域与网络	数据源所在地域与 VPCid。
JDBC URL	用于连接 MySQL 数据库的连接串信息。
数据库名称	需要连接的数据库名称。
用户名	连接数据库的用户名称。
密码	连接数据库的密码。
数据连通性	测试是否能够连通所配置的数据库。 **说明：** • 若连通性测试不通过，数据源仍可保存。连通性测试未通过而保存但数据源不可使用。 • 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 。

支持数据类型

- 实时数据类型转换

其他参考文档

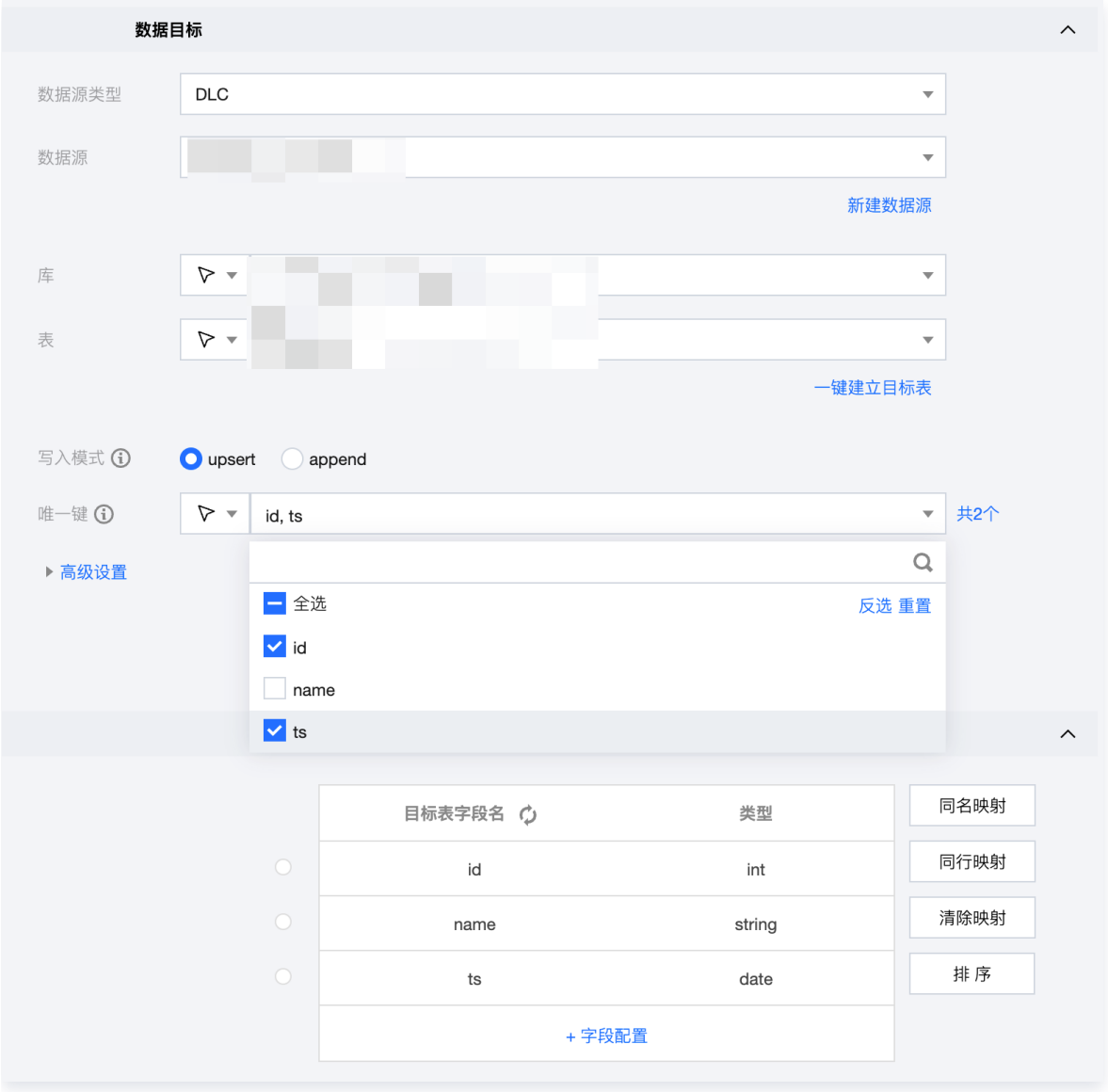
- 实时单表同步任务配置
- 实时整库同步任务配置
- MySQL 高级参数
- MySQL FAQ

DLC 环境准备与数据库配置

最近更新时间：2024-04-07 16:11:21

使用限制

- 选择 upsert 模式同步时，DLC 表的表必须是 V2表，并且开启设置 write.upsert.enabled=true。
- 选择 upsert 模式同步时，必须指定主键，如果是一个分区表，分区字段也必须添加到主键中。

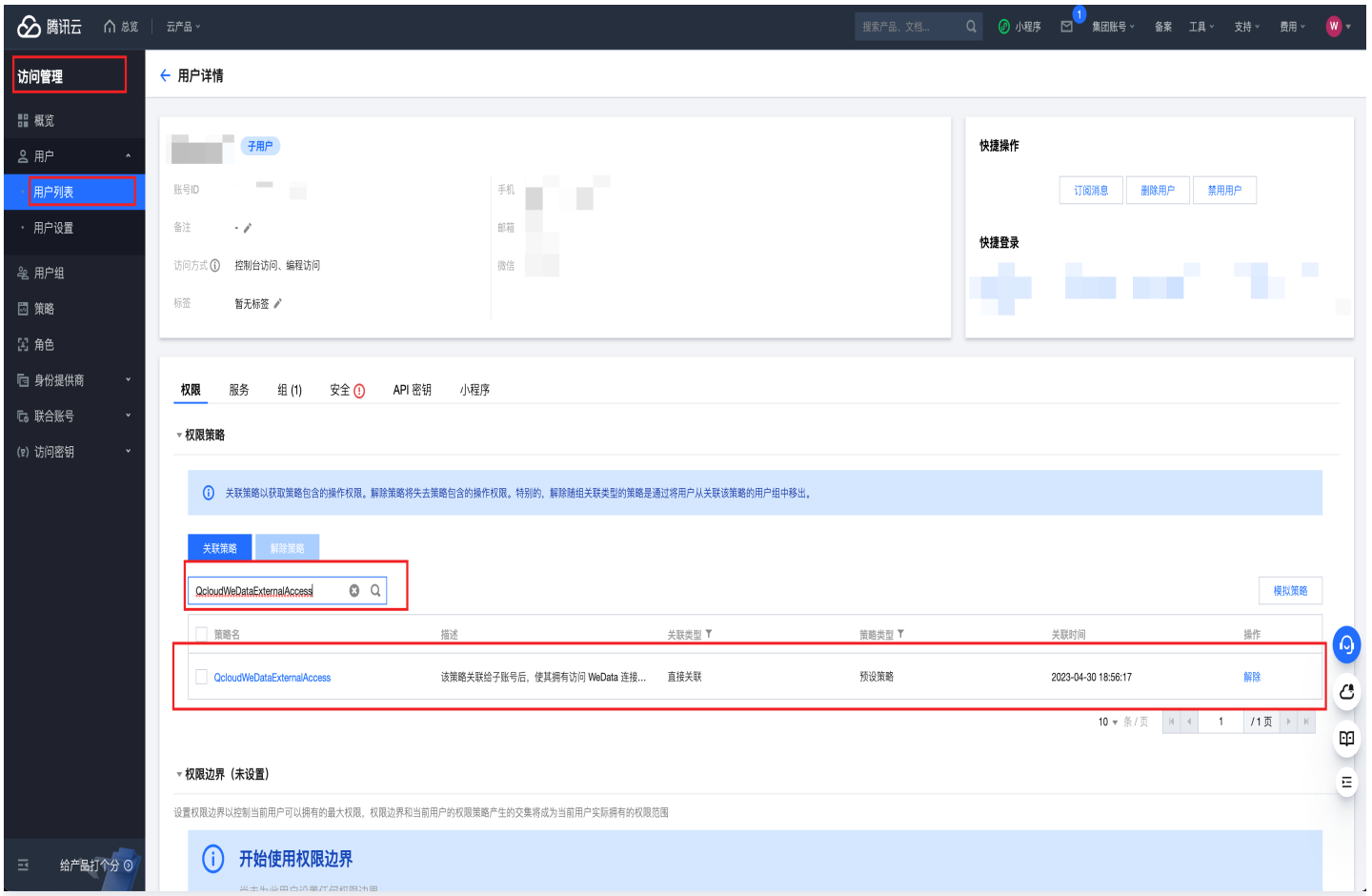


数据库环境准备

1.检查 WeData 是否有访问 DLC 的权限

1.登录腾讯云控制台，进入 访问管理（CAM）页面。
2.在左侧菜单中选择**用户**，找到您需要绑定策略的用户。

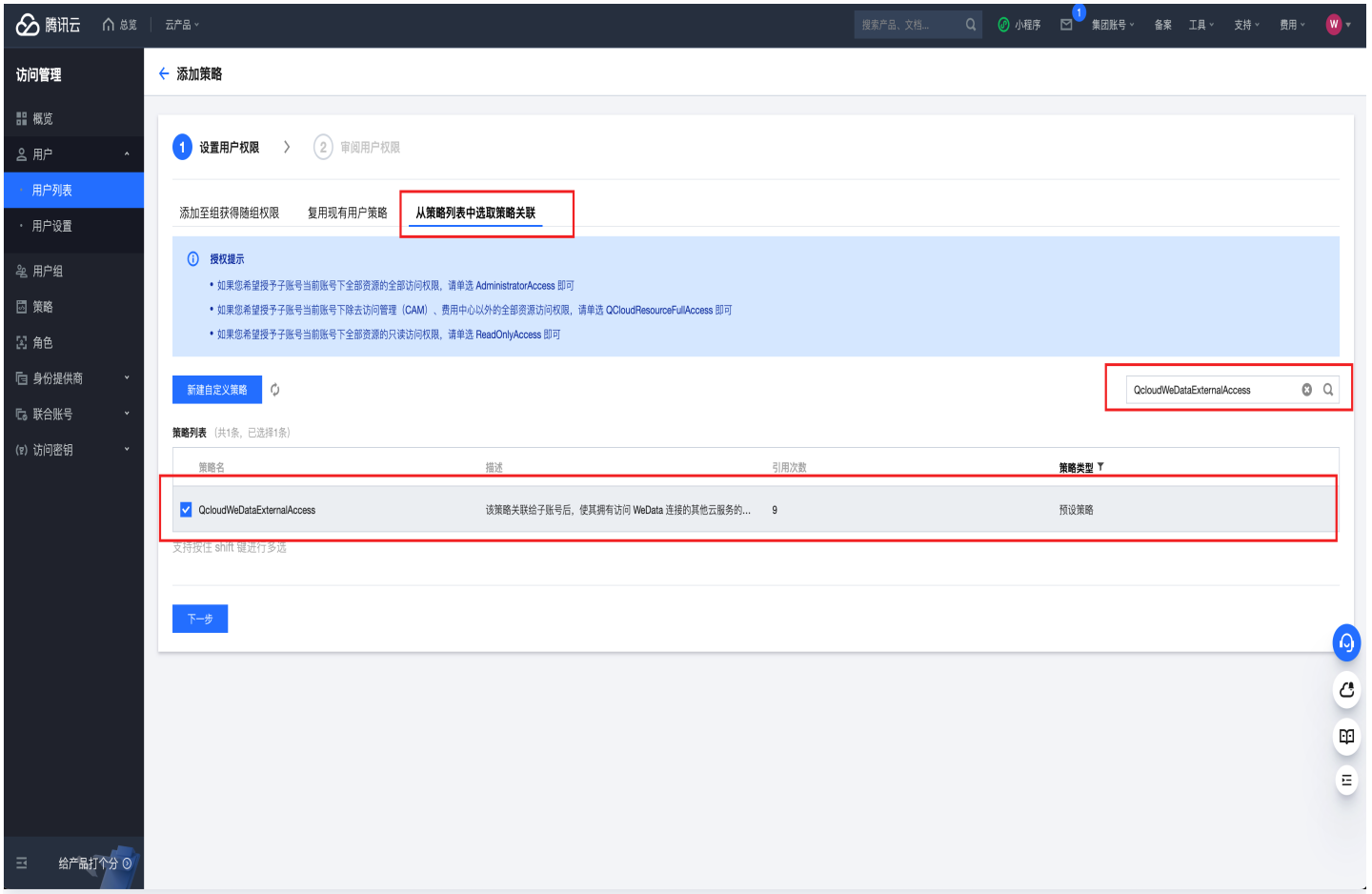
3.单击该用户的名称，进入用户详情页。在权限栏搜索 QcloudWeDataExternalAccess，如果已经关联了该策略，下面步骤可以跳过。



4.在用户详情页中，找到**权限管理**模块下的**关联策略** 。

5.单击**关联策略**中的**新建关联**按钮。选择**从策略列表中选取策略关联**，并输入 QcloudWeDataExternalAccess。

6.勾选 QcloudWeDataExternalAccess，单击**下一步**按钮，完成策略绑定。



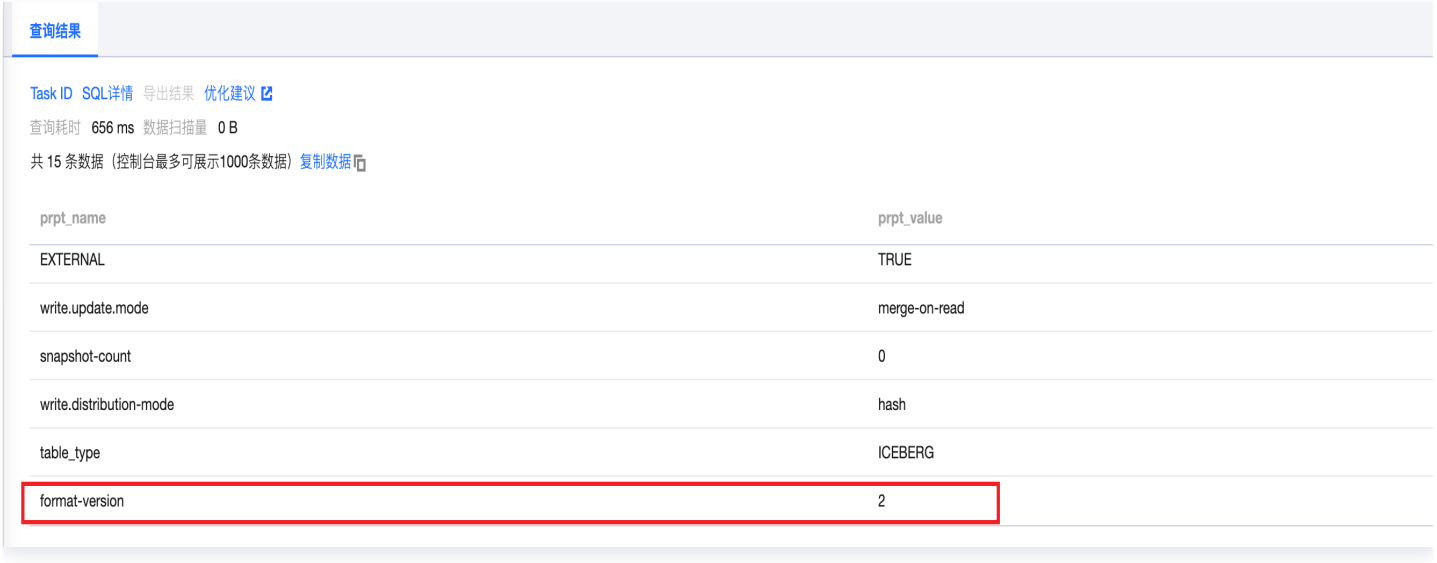
绑定成功后，该用户即可使用 QcloudWeDataExternalAccess 策略中定义的权限。

2.检查表是 V1表还是 V2表

实时同步到 DLC，支持两种模式：**Append 模式** 和 **Upsert 模式**。其中 Upsert 模式必须为 V2表。使用下面命令查看创建的表属性。

```
SHOW TBLPROPERTIES `DataLakeCatalog`.`databases_name`.`table_name`;
```

返回结果如下，format-version = 2 表示创建的是 V2表。如果您创建的 V1表，想改成 V2表，可参考 修改表属性 。


查询结果

Task ID SQL详情 导出结果 优化建议

查询耗时 656 ms 数据扫描量 0 B

共 15 条数据（控制台最多可展示1000条数据）复制数据

prpt_name	prpt_value
EXTERNAL	TRUE
write.update.mode	merge-on-read
snapshot-count	0
write.distribution-mode	hash
table_type	ICEBERG
format-version	2

3.创建 DLC 表

3.1 创建 V1表

DLC 默认创建的表为 V1表，V1表不支持 Upsert 模式。建表语句参考下面：

```
CREATE TABLE IF NOT EXISTS `DataLakeCatalog`.`dbname`.`test_v1` (`id` int,
`name` string, `ts` date) PARTITIONED BY (`ts`);
```

3.2 创建 V2表

使用 upsert 模式的写入 DLC，需要创建 V2表，需要在建立表的时候指定，建表语句参考下面：
创建 V2表，也一定要设置 'write.upsert.enabled' = 'true' 否则仍然是 Append 模式。

```
CREATE TABLE IF NOT EXISTS `DataLakeCatalog`.`dbname`.`test_v2` (`id` int,
`name` string, `ts` date) PARTITIONED BY (`ts`) TBLPROPERTIES (
  'format-version' = '2', -- 创建 V2 表
  'write.upsert.enabled' = 'true', -- 写入时做 upsert 操作，只支持 V2 表
  'write.distribution-mode' = 'hash', -- 定义写入数据的分布，设置为hash，支持多并
发写入
  'write.update.mode' = 'merge-on-read' -- 写入更新模式，在写入的时候做merge操
作，只支持 V2 表
)
```

3.3 修改表属性（可选）

对于已经创建了的表，需要修改其属性，可以参考下面语法：

```
SHOW TBLPROPERTIES table_name [('property_name')]
```

更多 DLC SQL 语法，可以参考 数据湖计算 DLC SQL 语法概览-SQL 语法-文档中心-腾讯云 。
下面是将已经存在的 V1表改成 V2表的示例：

```
ALTER TABLE
  `DataLakeCatalog`.`database_name`.`table_name`
SET
  TBLPROPERTIES (
    'format-version' = '2',
    'write.upsert.enabled' = 'true'
  )
```

数据源配置

WeData 目前支持通过连接串方式引入 DLC 数据源。
单击**项目管理 > 数据源管理 > 新建数据源 > 选择 DLC 数据源。**

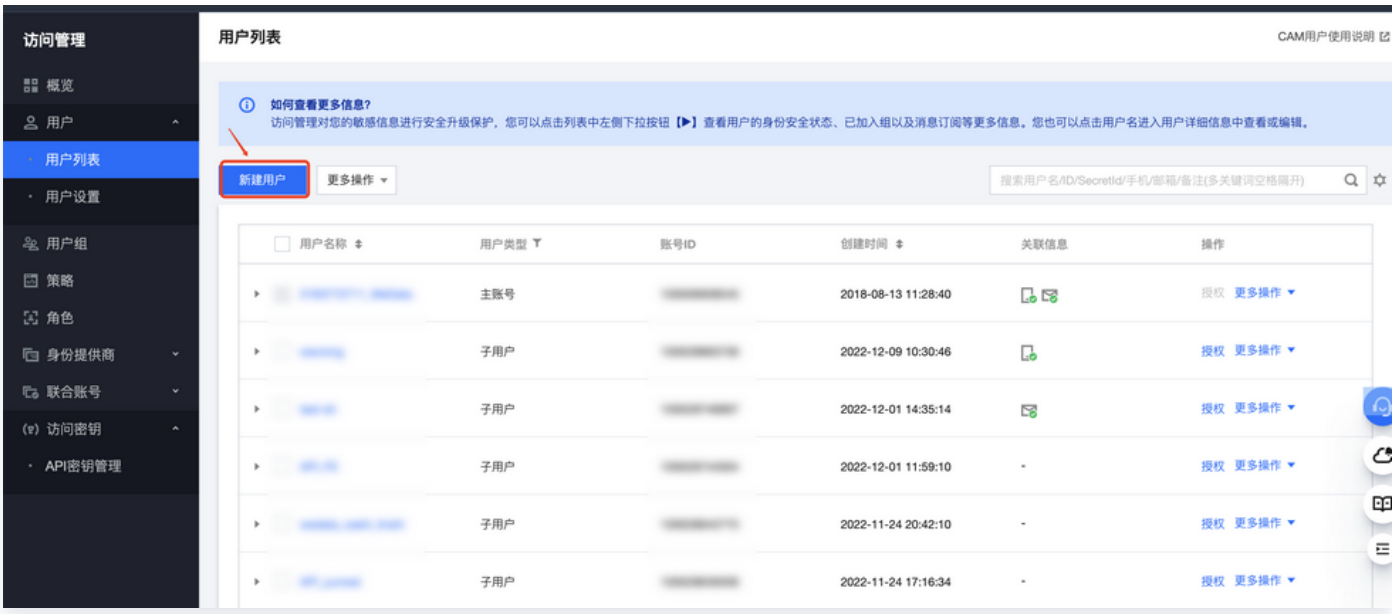
参数说明如下：

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改数据源仅创建人和项目管理员可用
JDBC URL	用于连接 DLC 数据源的连接串信息
用户名	连接数据源的用户名称
SecretId	连接数据源的用户名称

SecretKey	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

- **DLC 权限范围操作指南**

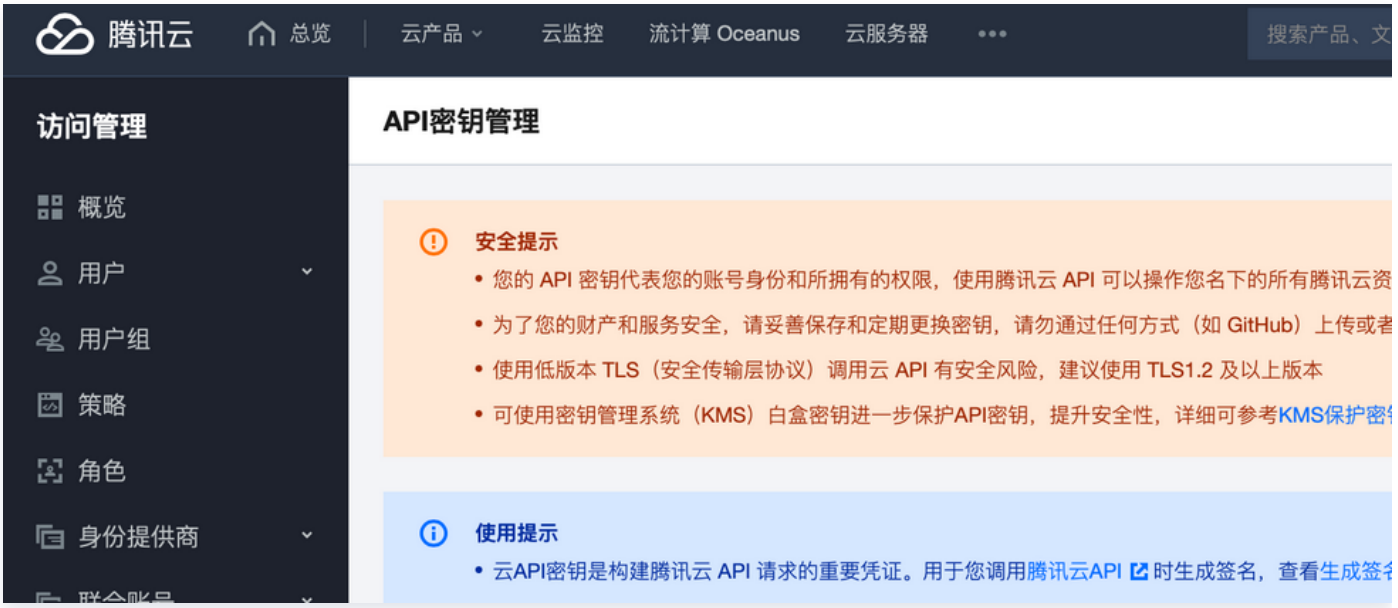
1. 请在**访问管理 > 用户 > 用户列表**页面中，单击**新建用户**并开始创建新用户（即子账号）。



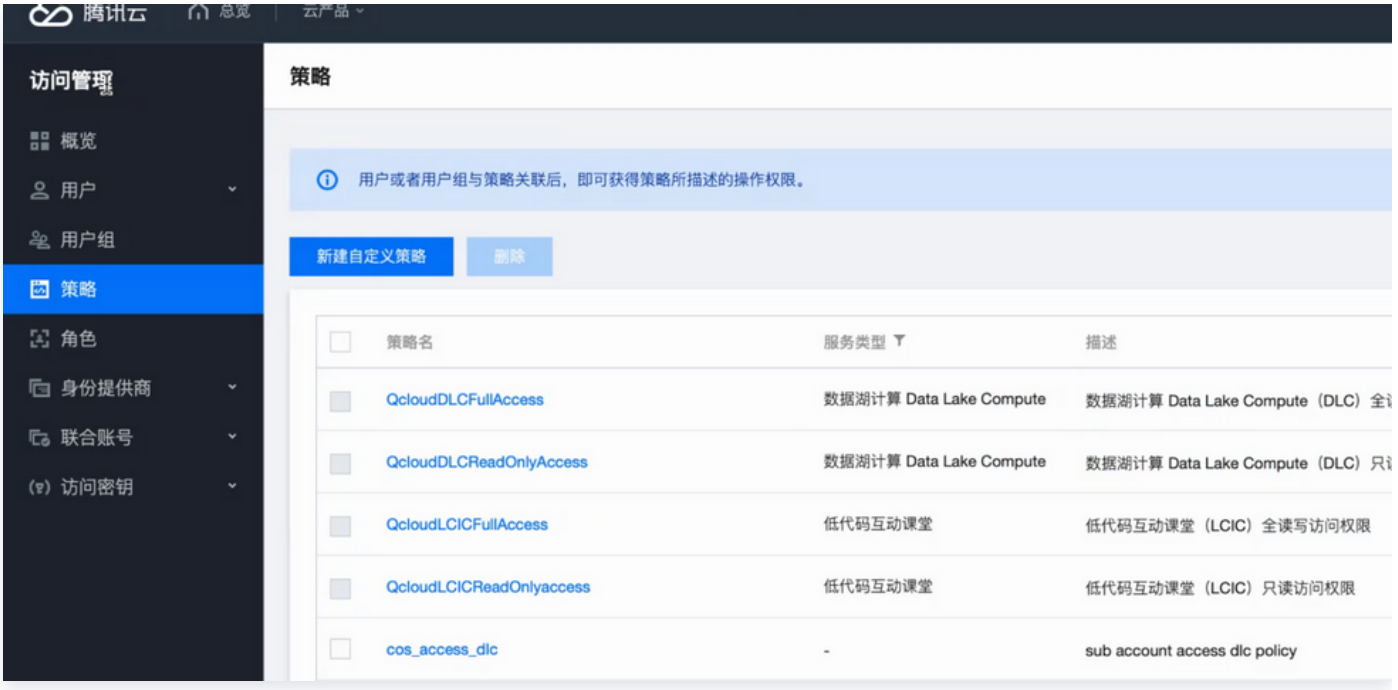
2. 按创建用户页面流程创建用户：填写用户名、选择标签键与标签值，即可完成新用户的创建。

3. 在**访问管理 > 策略**页面中，给制定的角色或子账号设置对应 DLC 数据策略赋予权限。

4. 申请密钥：若当无子账号时，可引导并跳转到创建子账号对应 API 密钥页面进行创建密钥，如下图：



5. DLC 平台 CAM 权限增加 DLC 策略权限，详情请参见 访问管理 CAM 服务 。



支持数据类型

- DLC - iceberg / Iceberg 数据类型转换

其他参考文档

- 实时单表同步任务配置
- 实时整库同步任务配置
- DLC高级参数

- DLC FAQ
- DLC 子账号权限管理

Kafka 环境准备与数据库配置

最近更新时间：2024-04-07 16:11:21

前提条件

1. **版本**：Kafka 版本0.11及以上。
2. **自建 Kafka**：自建 Kafka 和集成资源组的网络需要打通。

使用限制

Kafka 版本0.11及以上。

注意事项

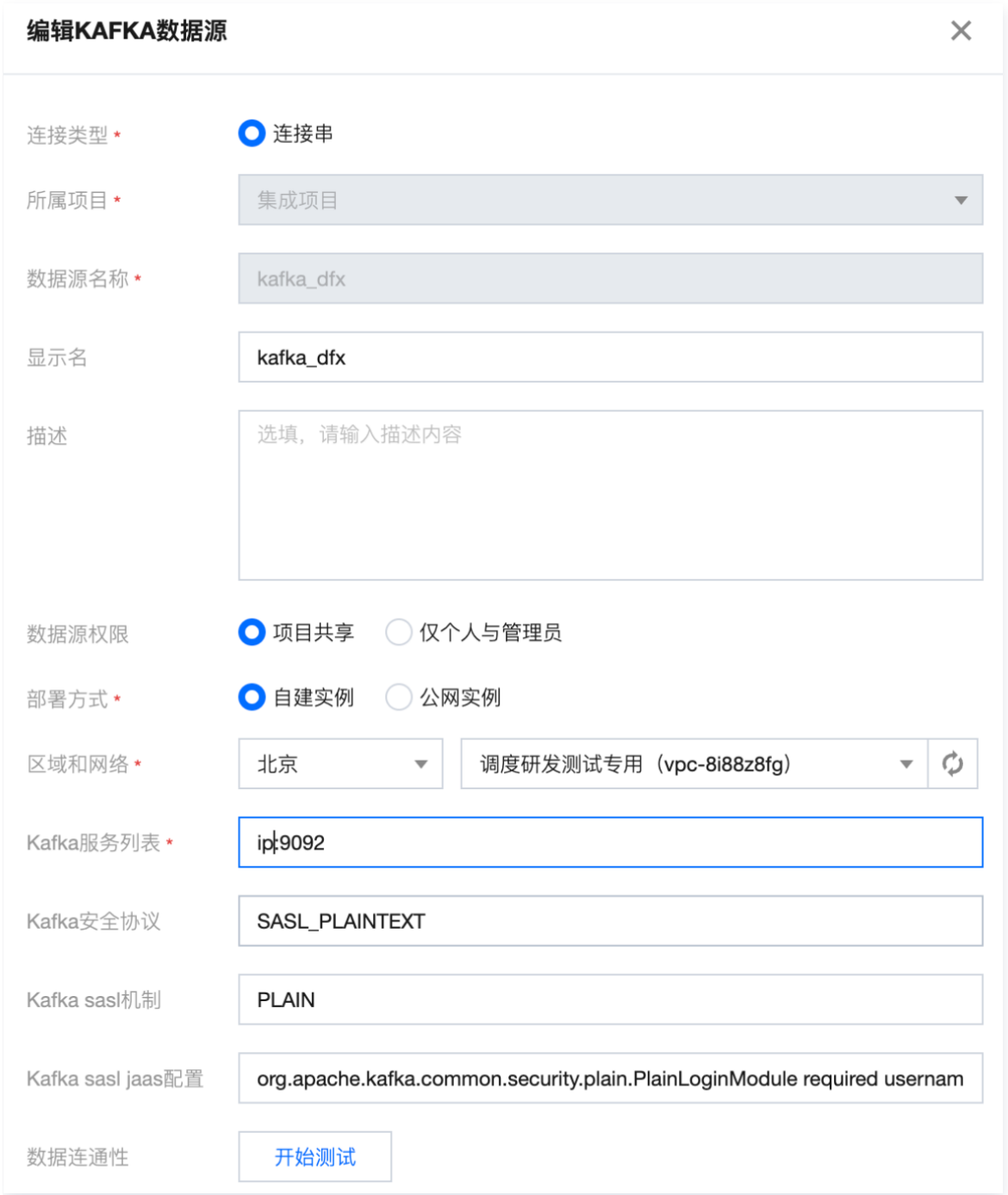
Kafka 客户端与服务端建立连接的过程如下所示：

1. 客户端使用您指定的 bootstrap.servers 地址连接 Kafka 服务端，Kafka 服务端根据配置向客户端返回集群中各台 broker 的元信息，包括各台 broker 的连接地址。
2. 客户端使用第一步 broker 返回的连接地址连接各台 broker 进行读取或写入：

我们需要注意当 bootstrap.servers 地址可以连通时，还是报网络问题联通性问题，可以排查下 kafka 服务端返回的 broker 连接地址是否连通性存在问题。可以检查 kafka broker 配置文件 server.properties中listeners 和 advertised.listeners 的地址是否可以和集成资源组网络连通。

数据源配置

目前支持通过连接串方式引入 Kafka 类型数据。



参数说明如下：

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	选择项目共享或仅个人与管理官可使用
部署方式	数据来源于自建实例或公网实例

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
Kafka 服务列表	请输入服务列表，例如 ip1:9092，ip2:9092
Kafka 安全协议	请输入 Kafka 安全协议
Kafka sasl 机制	请输入 Kafka sasl 机制
Kafka sasl jaas 配置	请输入 Kafka sasl jaas 配置
数据连通性	测试是否能够连通所配置的数据库（若连通性测试不通过，会给出错误提示供排查原因，同时，数据源仍可保存，但该数据源使用时会发生异常）

其他参考问答文档

- 实时单表同步任务配置
- 整库同步至 Kafka 配置详情

Doris 环境准备与数据库配置

最近更新时间：2025-04-25 15:47:12

前提条件

1. 支持 Doris 0.15+。
2. 若为自建 Doris 数据库，需要和集成资源组的网络打通。

使用限制

1. Doris 版本 0.15+。
2. 修改和删除只支持在 Unique Key 模型上。
3. 数据同步到 Doris 过程中的 DDL 响应。
   - 仅在整库同步场景下支持 DDL 响应。
   - 目前仅支持在 Doris 端添加列、添加表两种 DDL 。

注意事项

1. Doris 版本必须为0.15+。
2. Doris 支持 [DUPLICATE KEY|UNIQUE KEY|AGGREGATE KEY] 三种数据模型，若需以 Upsert 方式写入 Doris，需要确保数据模型为 UNIQUE KEY，详情可见：数据模型 - Apache Doris 。

数据库环境配置

1. 检查 Doris 数据库网络和数据集成资源组网络连通性。

参考**数据源配置**中的连通性测试，确保数据集成资源组可正常访问 Doris 数据库。

2. 检查待写入 Doris 表的数据模型。

Doris 写入支持 Append/Upsert 两种模式，若想以 Upsert 模式写入，需要确保 Doris 数据模型为 Unique key。具体可在 SQL 客户端执行如下 SQL：

```
show create table example_db.table_hash;
// 执行上诉 SQL 后的结果
CREATE TABLE example_db.table_hash
(
    k1 BIGINT,
    k2 LARGEINT,
    v1 VARCHAR(2048),
    v2 SMALLINT DEFAULT "10"
)
```

```
UNIQUE KEY(k1, k2)
DISTRIBUTED BY HASH (k1, k2)
// 若为 Unique key 模型，会在结果中出现关键词：UNIQUE KEY
```

3. 创建 Doris 表。

- 创建明细模型表。

```
CREATE TABLE example_db.table_hash
(
    k1 TINYINT,
    k2 DECIMAL(10, 2) DEFAULT "10.5",
    k3 CHAR(10) COMMENT "string column",
    k4 INT NOT NULL DEFAULT "1" COMMENT "int column"
)
COMMENT "my first table"
DISTRIBUTED BY HASH(k1)
```

- 创建主键唯一模型表。

```
CREATE TABLE example_db.table_hash
(
    k1 BIGINT,
    k2 LARGEINT,
    v1 VARCHAR(2048),
    v2 SMALLINT DEFAULT "10"
)
UNIQUE KEY(k1, k2)
DISTRIBUTED BY HASH (k1, k2)
```

详情可参考 CREATE-TABLE - Apache Doris

数据源配置

支持通过连接串方式引入 Doris 数据源。

新建DorisDB数据源

选择类型 > 2 配置数据源

连接类型 * 连接串

所属项目 * WeData产品运营测试_1129

数据源名称 * 请输入数据源名称

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

部署方式 * 自建实例 公网实例

区域和网络 * 请选择地域 请选择vpcId

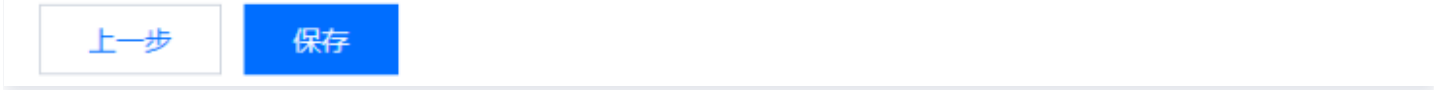
JDBC URL * [redacted]

FE URL * ip:http_port

用户名 * tiyan1

密码 ••••••••••••••

数据连通性 开始测试



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	可选择自建实例或公网实例，用户自建实例需要输入区域和网络，公网实例无需区域和网络信息
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
FE URL	输入 fe http 地址，格式为：IP地址:http端口，多个地址之间使用逗号（,）分隔，例如：172.17.16.3:8030,172.17.16.4:8030 **注意：** 如果直接从 Doris 中将地址复制过来，会自动添加前缀 https:// 或 http://，在数据源这里填写的时候需要去掉这些前缀，只保留 iP:PORT。
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

其他参考问答文档

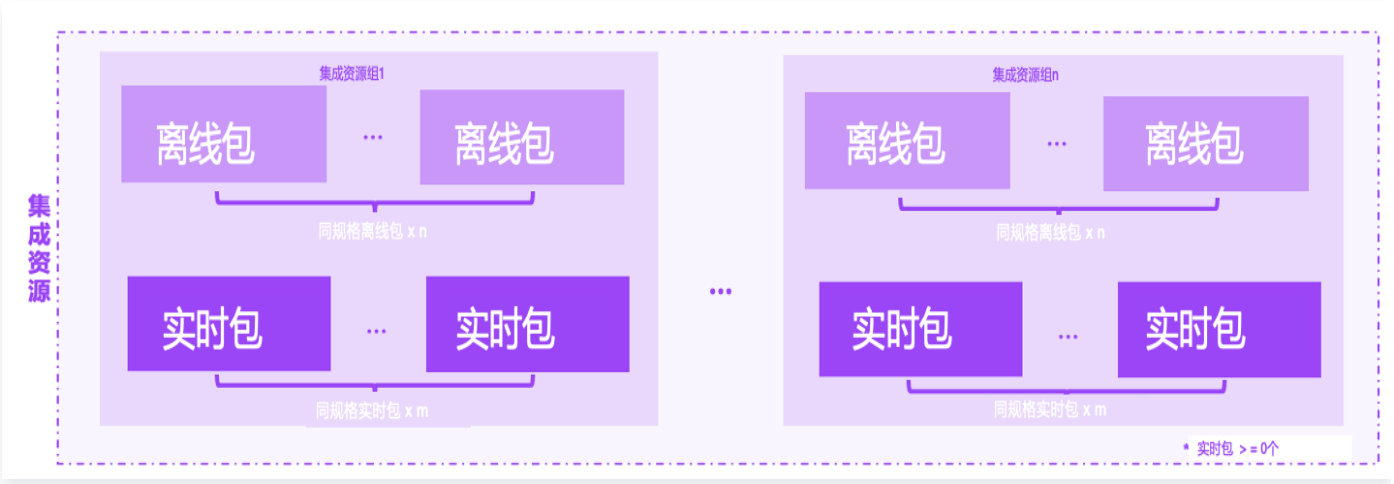
- 单表任务配置概览
- 整库同步至 Doris 配置详情
- 实时节点高级参数

集成资源配置与管理
集成资源概述

最近更新时间：2024-04-02 16:17:11

集成资源组是指在运行数据集成任务时专享使用到的计算资源，本资源主要以资源组形式展现。资源组由“离线包”和“实时包”两部分组成：

- 离线包：必选。主要用于构建集成资源环境及运行离线任务，可根据离线集成任务量选择对应规格的资源包以及资源包数量。
- 实时包：非必选，可根据是否需要实时集成按需配置。实时包用于运行实时同步任务，实时包必须配合离线包一起使用，不可单独分开使用。



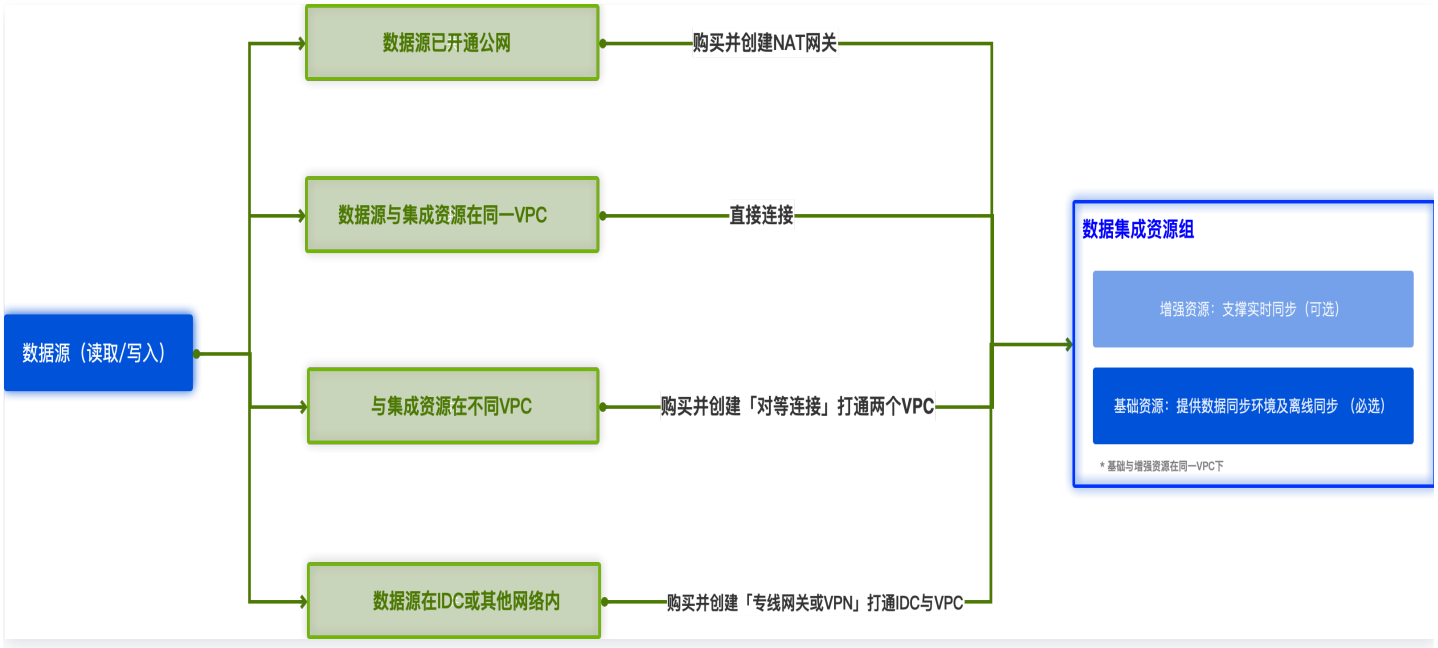
注意事项：

- 资源组为最小任务绑定单位，其中单个资源包仅可归属于一个资源组。
- 集成资源内离线包必选，实时包为实时集成可选项。
- 实时包不支持脱离离线包单独使用。
- 可单独销毁实时包；若销毁离线包，实时包也将一并销毁，计费周期内未使用基础和实时包资源均进行退费。
- 离线包和实时包必须为同一网络环境。

集成连通性与使用规划

最近更新时间：2024-07-09 22:01:41

使用数据集成同步任务前，需要保证数据源网络（包括读端、写端）与数据集成资源组之间网络互通，且资源不可因为白名单限制等原因被拒绝访问，否则无法完成数据传输同步。
数据集成资源组内包含的机器资源默认需处于同一 VPC 网络环境下：



- 若数据源开通公网：需要购买并创建 NAT 网关，允许集成资源通过网关连通数据源所在 VPC，详细操作请参见 NAT 网关 相关文档。
- 若数据源处于 VPC 内：
  - 若与集成资源位于同一 VPC：可直接使用。
  - 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：需购买 VPN 或 专线网关 打通集成与数据源所在 VPC。

集成资源列表及配置

最近更新时间：2024-04-07 17:44:51

进入 数据集成 DataInLong 控制台，单击**集成资源**，进入资源组列表页面。资源组列表展示了当前主账号下所有购买的资源组及状态信息，列表内详细参数及说明如下：

参数	说明
资源组名称	资源组显示名称
地域	购买资源组所归属的物理地域
网络	当前资源组内配置绑定的网络信息
绑定项目	当前资源组绑定归属的项目空间名称及 ID
状态	当前资源组运行状态： • 运行中：资源正常运转 • 创建中：订单已下达，当前资源内资源包正处于创建及配置流程中，此过程可能需要1 – 2分钟 • 创建失败：资源分配或组建失败，此状态下可单击**重试** • 初始化中：资源配置初始化中 • 运行异常：资源组运行过程中出现异常 • 更新中：资源组变更或其他操作后，正在更新当前资源组的配置 • 更新失败：资源组配置变更失败，此状态下可单击重试 • 已到期：资源有效时间到达，冻结中 • 释放中：资源有效时间到达超过7天或用户已触发资源销毁
资源包规格/数量	当前资源组内包含的单个资源包的规格以及总的资源包数量
到期时间	资源组到期时间，到期后资源组将被冻结影响任务运行
操作	针对当前资源组进行项目关联或资源变配等操作，允许的操作与资源状态关联

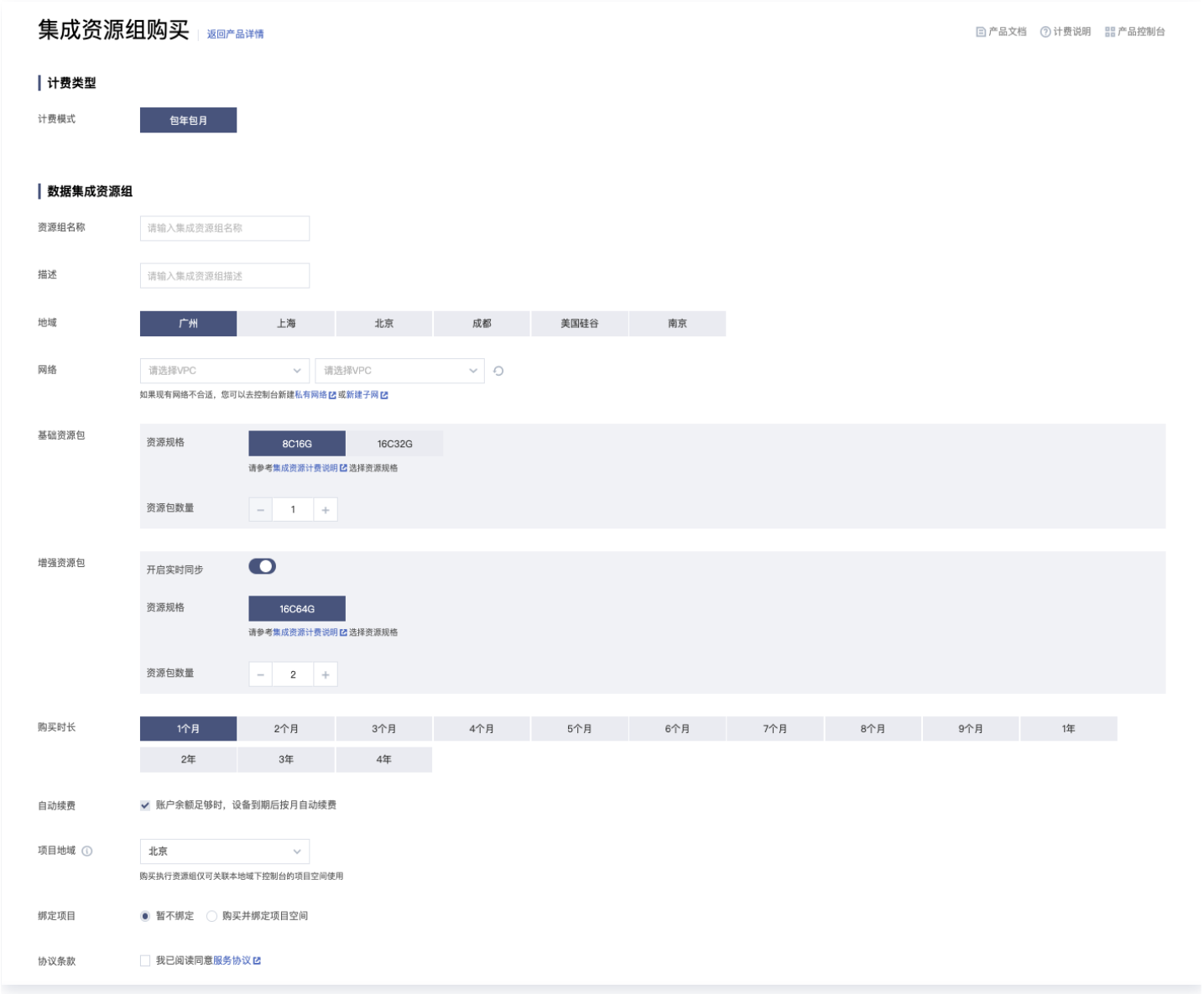
状态与操作状态关联状态集说明：

资源组状态 | 操作	绑定项目	解除绑定	销毁/删除	续费	调整配置
运行中	✓	✓	✓	✓	✓
创建中	–	–	–	–	–

创建失败	✓	✓	此状态下，操作栏显示删除操作。单击删除后，默认删除未创建成功的资源并退费	–	–
初始化中	–	–	–	–	–
运行异常	✓	✓	✓	✓	✓
更新中	–	–	–	–	–
更新失败	✓	✓	✓	✓	✓
已到期	–	–	–	✓	–
释放中	–	–	–	–	–
已释放	–	–	–	–	–

创建

1. 在选定集成资源组 Tab 页面下，单击**创建**，即可进入资源配置页面。



2. 您可以参考下表进行资源参数配置。

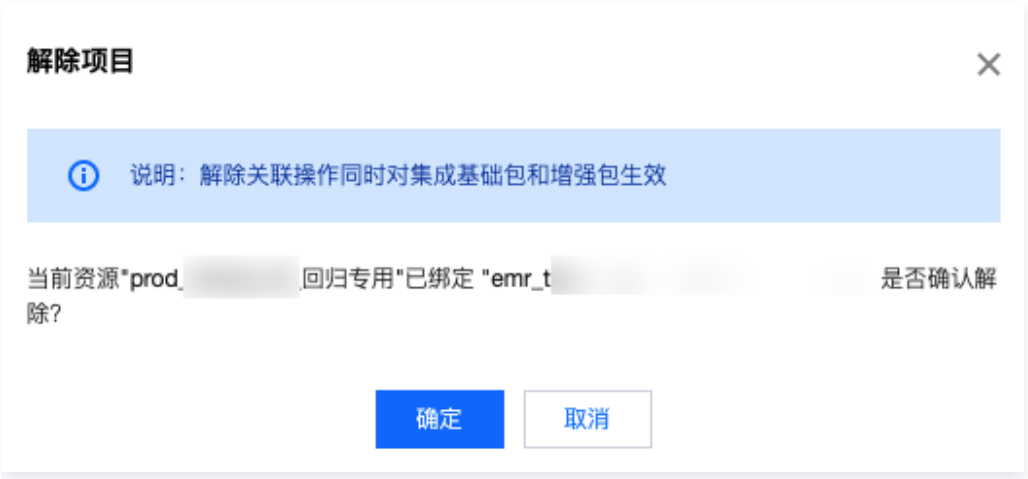
参数	说明
资源组名称	设置当前资源组显示名称，命名不可为空支持中文、英文、数字、-、_
描述	设置当前资源组备注信息
地域	购买资源组所归属的物理地域
网络	指定当前资源组需配置的 VPC 和子网信息。数据集成内，离线包和实时包必须为统一网络环境
离线资源包	选择离线资源包的规格及数量

实时资源包	• 开启：非必选，可根据是否需要实时集成按需配置。开启后需配置实时资源包规格 • 选择实时资源包的规格及数量
购买时长	指定当前资源包购买总时间
项目地域	指定当前所购买的产品版本地域信息，指定后当前资源组仅可关联本地域内的项目空间使用
绑定项目	指定当前资源组关联的项目

3. 配置完成后可单击**创建**，并支付完成后，即可前往控制台页面查看资源组信息。

关联项目/解除关联

关联项目指定当前资源组关联的项目空间信息，指定后该项目空间下的任务可使用本资源组。目前一个资源组仅可关联一个项目空间。解除关联是将资源组从当前空间可使用的资源中移除。解除关联需当前项目空间下不存在正在运行的任务，否则将解除失败。
数据集成资源组内所包含的离线包和实时包必须关联同一项目空间，关联或者解除关联操作都将同时对资源组生效。



调整配置

变更当前资源组内所包含资源包的基础规格或者资源包的数量等操作，包括变更资源规格、增加资源包数量、以及减少资源包数量三种。

- 变更资源规格：扩大或者缩小资源组内离线资源包的规格，资源包数量不变。
- 增加资源包数量：选择后，仅可增加当前资源组下所包含的资源包总数。
- 减少资源包数量：选择后，仅可调整当前资源组下所包含的资源包总数。

> **注意**
> 减少资源包数量可能会影响资源组关联的任务运行，请谨慎操作。

续费

对当前资源进行续期操作，当资源组内包含多个资源包续费操作将对所有资源包生效。

销毁

销毁是指强制销毁当前资源组，销毁后资源组将无法恢复。若当前资源组有关联的正在运行的任务，此操作会直接任务运行与数据产出，请谨慎操作！

> **说明**
> 若销毁离线包，实时包将同时一并销毁。

资源组配置公网

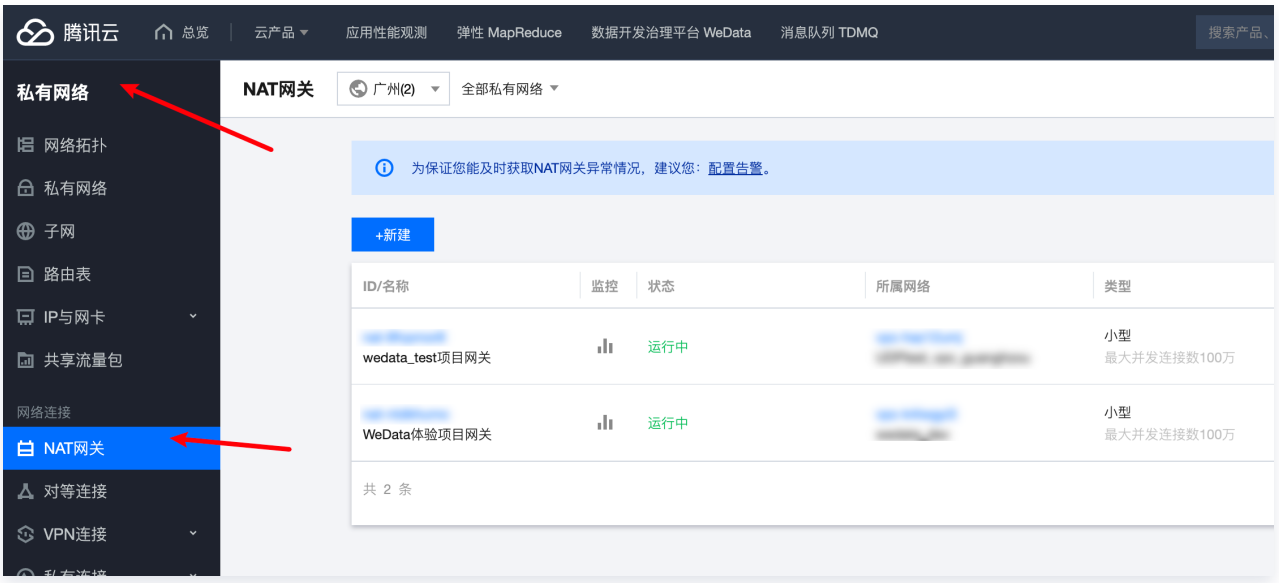
最近更新时间：2024-04-07 17:44:51

操作场景

适用于为执行资源组的 VPC 配置访问公网的能力。

操作步骤

1. 在对应地域购买创建 NAT 网关。进入腾讯云私有网络 > NAT 网关：

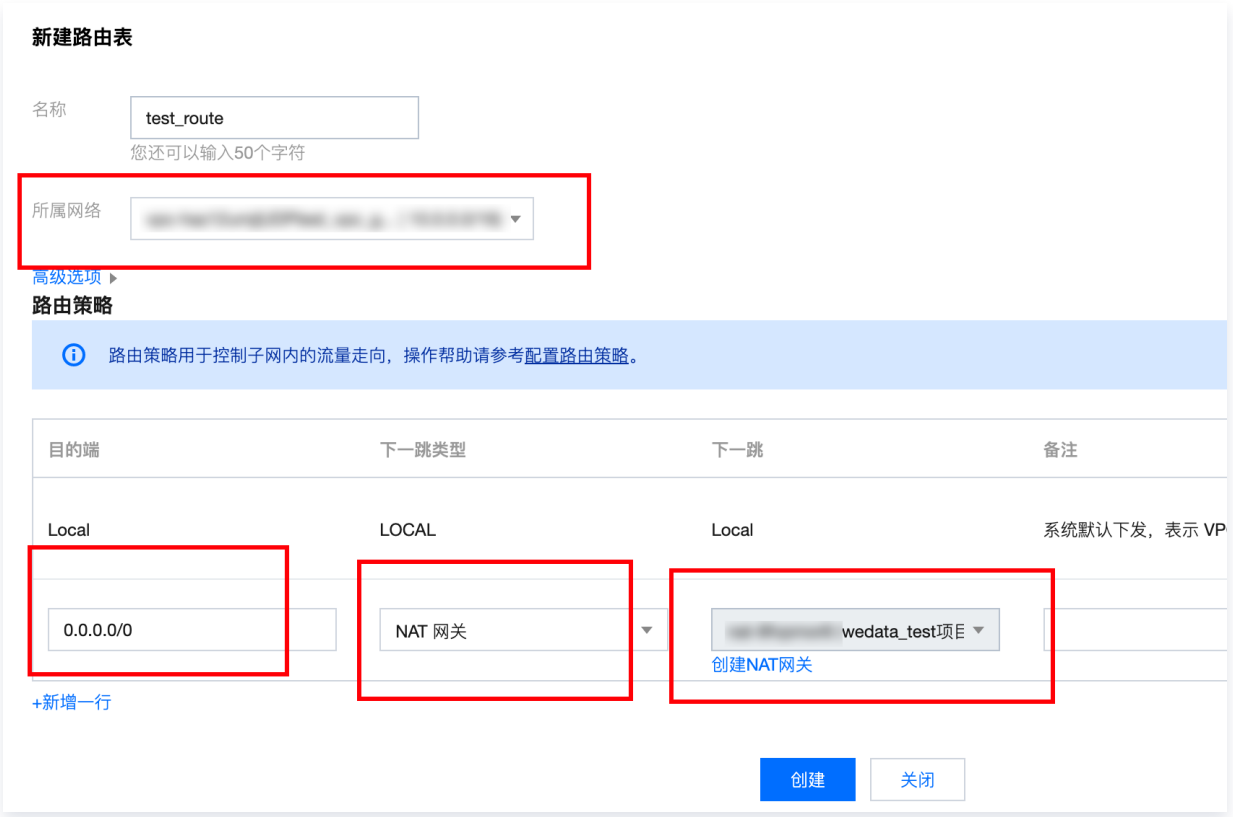


2. 购买 NAT 网关：


NAT 网关
返回产品详情
网关配置
网关名称
你还可以输入60个字符
地域
广州
私有网络
去创建
网关类型
小型（最大并发连接数100万）
出带宽上限
10Mbps
访问公网流量同时受到 NAT 网关和弹性公网 IP 的带宽上限限制，最终以较小上限值为准
NAT 网关入带宽暂不支持调整上限
弹性公网 IP 配置
弹性公网 IP
已有弹性公网 IP
新建弹性公网 IP
选择 IP

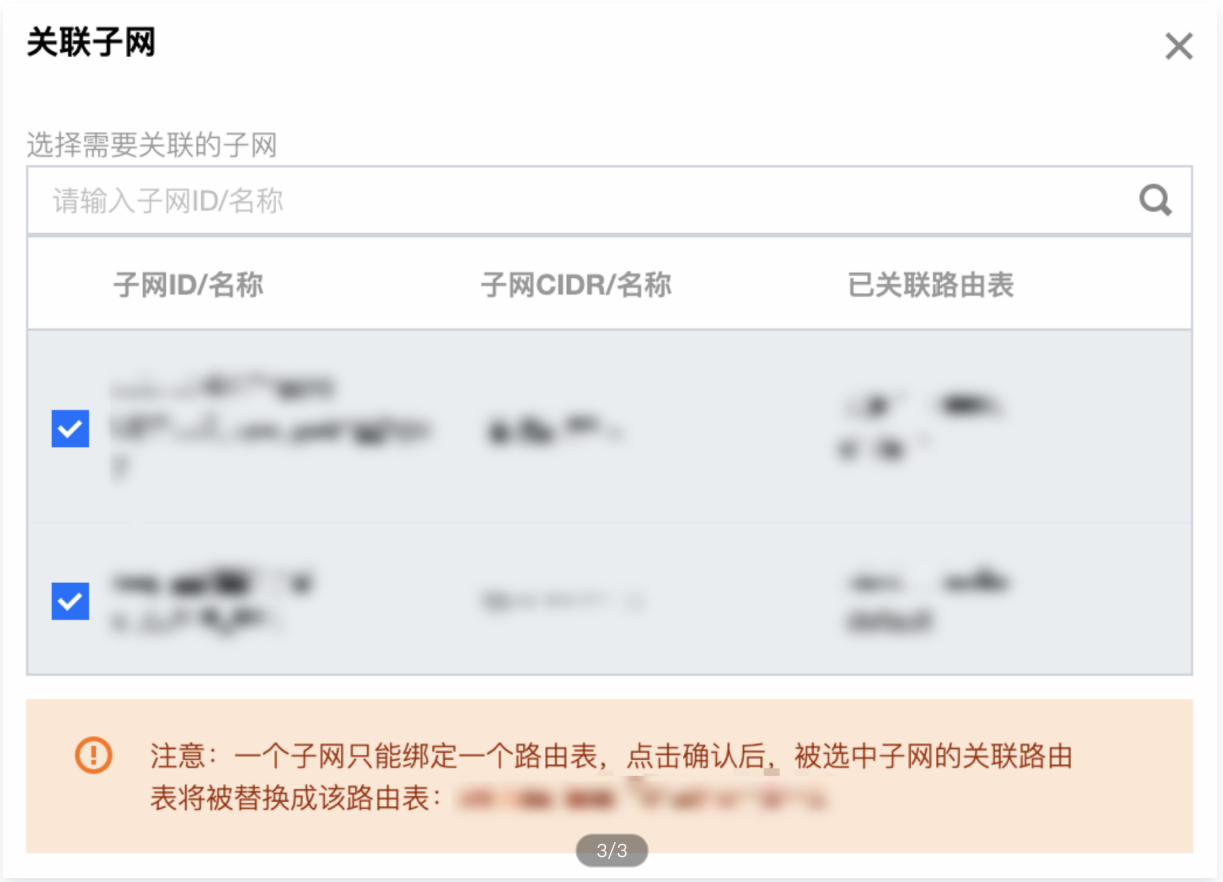
3. 创建并绑定路由表指向 NAT 网关，进入腾讯云私有网络 > 路由表：

4. 创建自定义路由表：选择所属的 VPC 网络，目的端填写0.0.0.0/0或固定路由，下一跳类型选择 NAT 网关，下一跳选择刚才创建的 NAT 网关。



5. 路由表创建完成后， 进入路由表详情，关联绑定子网：

6. 选择需要关联的子网并绑定：



实时同步配置与运维
实时同步支持的数据源

最近更新时间：2024-09-06 16:40:21

背景信息

实时同步支持多种输入及输出数据源搭配组成同步链路进行同步，您可以根据数据集成支持的数据源，配置实时同步任务。

支持的数据源

1. 单表实时同步支持的数据源

数据源		实时同步	
		单表读取	单表写入
关系型数据库	MySQL	✓	✓
	TDSQL-C MySQL	-	✓
	PostgreSQL	✓	✓
	TCHouse-P	-	✔
	SQL Server	✓	✓
	Oracle	✓	✓
大数据	Hive	-	✓
	DLC	-	✓
	Doris	-	✓
	ClickHouse	-	✓
	Iceberg	-	✓
	HBase	-	✓
	HDFS	-	✓
	TBase	-	✓
	Greenplum	-	✓

NoSQL	Elasticsearch	–	✓
	Mongo	✓	–
消息队列	Kafka	✓	✓
	TiDB-Kafka	✓	–
	DTS-kafka	✓	–
图数据库	GDB	✓	

2. 整库实时同步支持的数据源

数据源		实时同步	
		整库读取	整库写入
关系型数据库	MySQL	✓	–
	TDSQL-C MySQL	✓	–
	TCHouse-P	–	✔
	SQL Server	✔	–
	PostgreSQL	✓	✓
	Oracle	✓	–
大数据	Hive	–	✓
	DLC	–	✓
	Doris	–	✓
	Iceberg	–	✓
	StarRocks	–	✓
NoSQL	Mongo	✓	–
消息队列	Kafka	✓	✓

3. 日志采集支持的数据源

数据	日志采集		
	日志读取	日志写入	

关系型数据库	MySQL	-	✓	
	TDSQL-C MySQL	-	✓	
	PostgreSQL	-	✓	
	TCHouse-P	-	✓	
	SQL Server	-	✓	
	Oracle	-	✓	
大数据	Hive	-	✓	
	DLC	-	✓	
	Doris	-	✓	
	ClickHouse	-	✓	
	Iceberg	-	✓	
	HBase	-	✓	
	HDFS	-	✓	
	TBase	-	✓	
	Greenplum	-	✓	
NoSQL	Elasticsearch	-	✓	
消息队列	Kafka	-	✓	
数据上报	SDK	✓	-	
	TKE	✓	-	
	CVM	✓	-	

整库同步任务配置
整库同步支持链路类型

最近更新时间：2024-04-02 16:17:11

本产品支持将 MySQL、TDSQL-C MySQL、Kafka 等数据源进行实例级同步搬迁，目前已支持 Doris、DLC、Kafka 等目标端。整库链路支持详情如下：

数据来源	数据目标
MySQL	DLC
TDSQL-C MySQL	
Kafka	
MySQL	Doris
TDSQL-C MySQL	
PostgreSQL	
MySQL	Kafka
TDSQL-C MySQL	
Mongo	
PostgreSQL	
MySQL	Iceberg
TDSQL-C MySQL	
Kafka	
Oracle	
Oracle	HIVE
PostgreSQL	StarRocks
Mongo	PostgreSQL

整库任务配置概览

最近更新时间：2024-12-10 11:41:52

背景信息

整库迁移支持来源端的数据及结构监控，可将源端所有库表下的全量或增量数据实时同步至目标端，同时支持目标端自动建表、字段变更同步等特性。支持 MySQL、Doris、DLC、Kafka 等数据源。

前提条件

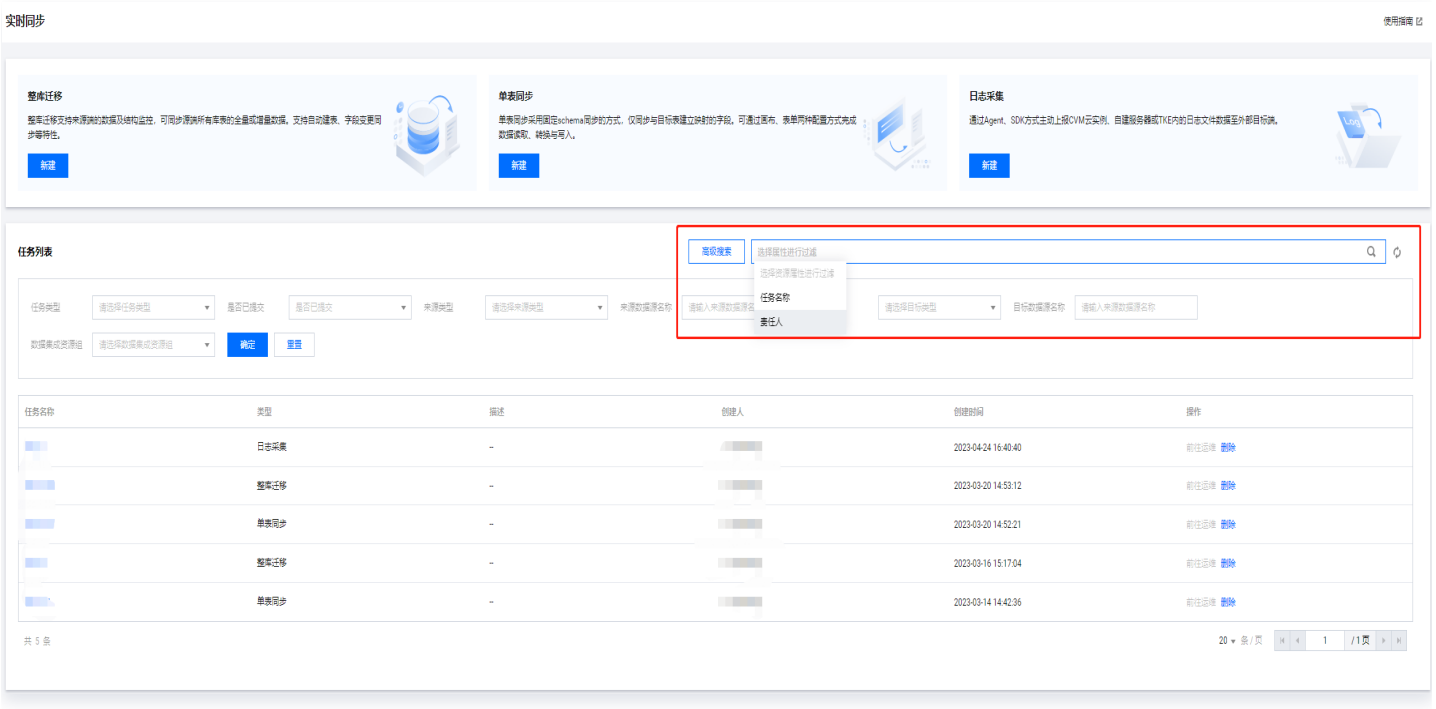
1. 已配置好来源及目标端的数据源以备后续任务使用。详情请参见 数据源管理与配置方式 。
2. 已购买数据集成资源组。详情请参见 配置集成资源组 。
3. 已完成数据集成资源组与数据源的网络连通。详情请参见 集成连通性与使用规划 。
4. 已完成数据源环境准备。您可以基于您需要进行的同步配置，在同步任务执行前，授予数据源配置的账号在数据库进行相应操作的权限。详情参见 数据源配置与数据库环境准备 。

操作步骤

步骤一：创建整库同步任务

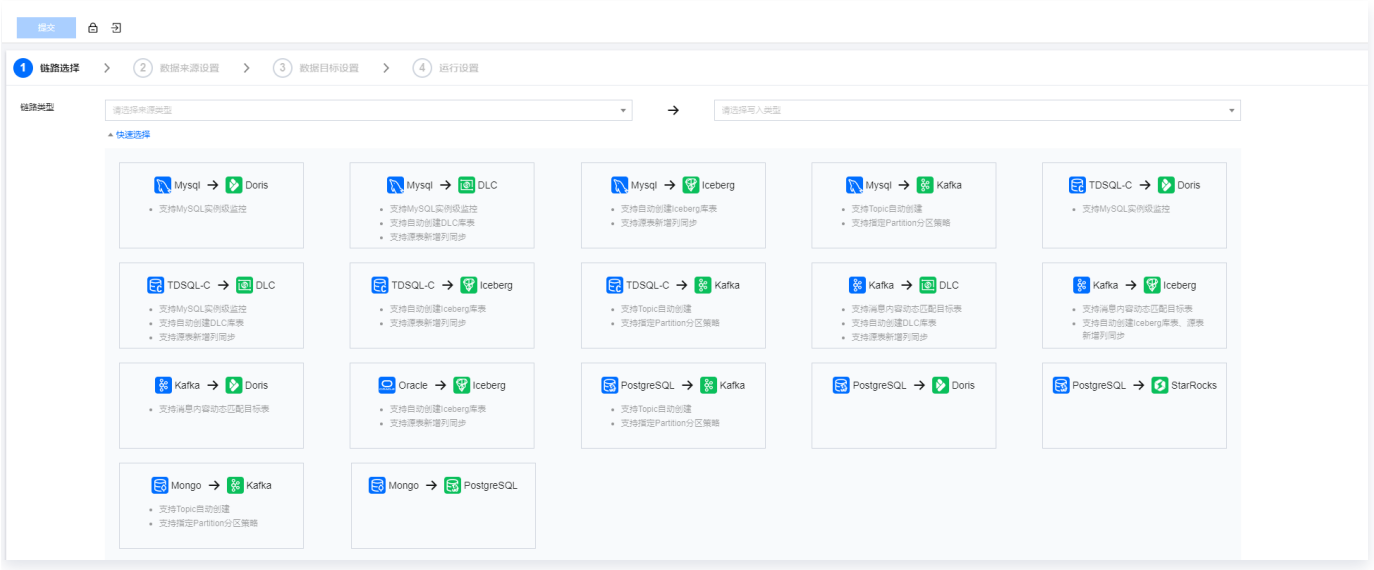
进入**配置中心 > 实时同步**任务页面后，单击**新建**整库迁移任务。

筛选：实时任务配置列表优化增加多种搜索及筛选条件。



步骤二：链路选择

您可以根据实际业务选择需要同步的来源与去向数据类型。选择后，后续步骤将基于此步选择的类型展示对应来源及目标端配置参数。请保证所选择数据源类型与实际配置数据源类型保持一致。



步骤三：数据来源设置

在此步骤中，您可以选择源端数据源中需要同步的库和表，配置来源端读取方式、时区等。

> **说明：**

此处的配置项根据来源端数据源类型的不同而存在一定的差异，具体以各方案实际配置界面为准。

步骤四：数据目标设置

在此步骤中，您可以定义写入目标端数据源及库表等相关属性。例如，写入模式，库、表名称匹配规则等，以及设定目标对象与来源对象之间的名称映射规则。目前整库目标对象匹配策略支持**与来源同名**、以及**自定义**两类：

- **与来源库/来源表同名**

  默认情况下，整库同步任务中源端数据库、数据表将写入目标端同名schema或同名表中。此策略下，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。

> **说明：**任务计划同步将源端数据库 DB1下的 TableA 和 TableB 同步至 Doris 数据源，配置**与来源库同名**以及**与来源表同名**策略以后，任务运行时候将默认在数据源连接内分别匹配 “DB1.TableA” 和 “DB1.TableB”。

- **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。

自定义方式下，系统针对整库场景提供了内置系统参数。内置参数主要覆盖源端数据源名称、源端库名、源端表名等。其中 Kafka 类型，支持动态匹配消息内数据字段 value 值。自定义方式下内置参数说明如下：

参数名称	参数说明
来源数据源名	${datasource_name_di_src}
来源库名	${db_name_di_src}
来源表名	${table_name_di_src}
来源 schema 名	${schema_name_di_src} **说明：** 适用于 pg 等。
来源 Topic 名	${topic_di_src} **说明：** 仅适用于 kafka 类型。
数据字段	${key} **说明：** 仅适用于 kafka 类型，请替换 key 为具体字段名称。

说明：

- 示例1：如来源表名称为 table1，映射规则为 ${table_name_di_src}_inlong，则 table1的数据将被最终映射写入至 table1_inlong 中。
- 示例2：如果来源 kafka 消息格式为 {“name”：“inlong”；“age”:12}，映射规则为 ${name}_di，则此条消息将被最终映射写入至 inlong_di 表中。
- 支持的配置项根据目标端数据源类型的不同而存在一定的差异，具体以各方案实际配置界面为准。

步骤五（可选）：批量手动创建目标表

针对来源表与目标表之间名称匹配策略，系统将自动转换异构数据源之间DDL结构，为您提供手动批量快速修改、及创建目标表的能力。

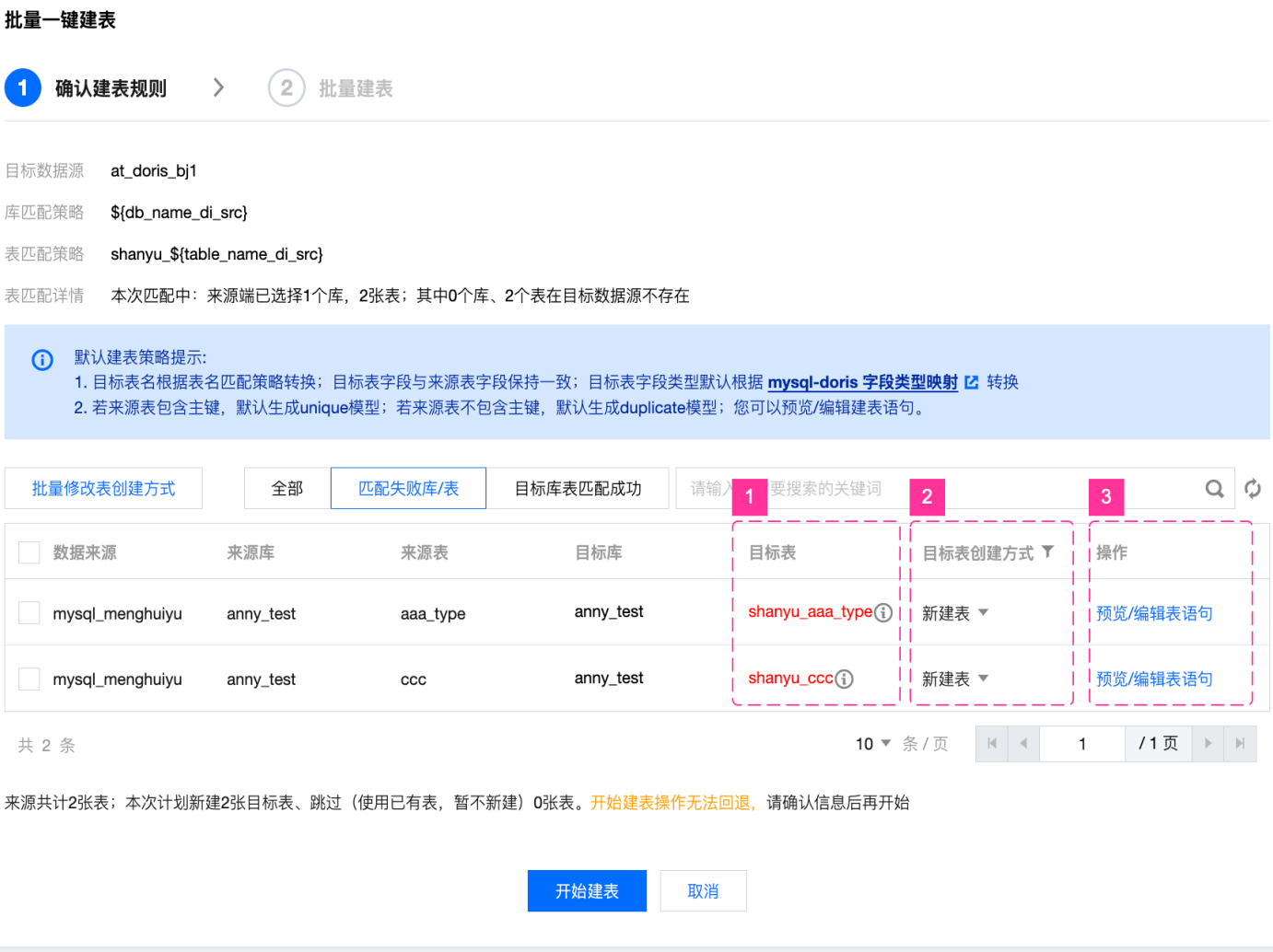
说明：

1. 批量建表是已经配置完目标库、表的名称匹配策略，操作步骤可参考本文档 步骤四 。
2. 批量建表仅支持部分目标数据源及同步链路。

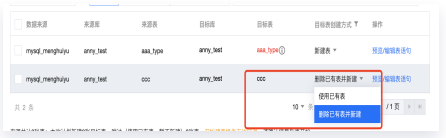
1. 检查库表名称匹配策略，并单击**批量建表**。

2. 确定建表规则

本步骤中，系统将根据库及表匹配策略扫描目标对象，并自动生成目标表 DDL 语句。您可以在此步骤中检查目标表是否匹配正确、配置目标表创建方式、及编辑建表语句等。重点功能及对应说明如下：



重点功能点		功能说明	功能示例
1	目标库/表名称	系统将默认根据匹配规则预设生成目标库/表名称，同时扫描确认该库/表在目标数据源中是否存在： • 匹配失败的库/表：未在目标数据源中发现符合匹配规则的库/表对象。列表默认高亮展示匹配失败的库表。 • 匹配成功的库/表：目标数据源中存在符合匹配规则的库/表对象。	
2	目标表建立方式	针对目标数据源中库/表对象，系统提供多种建表策略： • 匹配失败的库/表：支持**新建表、暂不新建**两种方式	1. 匹配失败的库/表策略配置： 

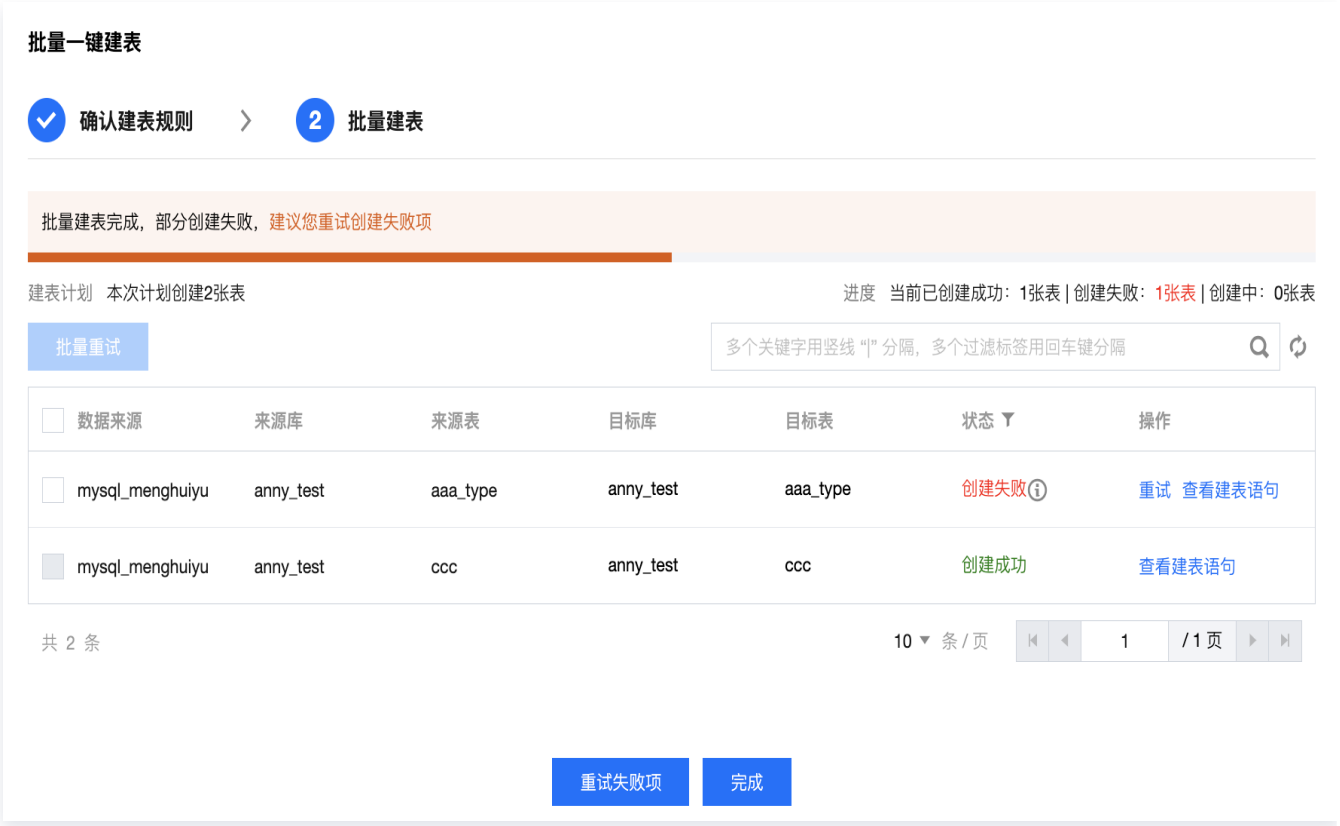
		○ 新建表：此次批量创建中，根据转换生成的目标表DDL 自动创建。 ○ 暂不新建：本次操作暂时忽略此表。 • 匹配成功的库/表：支持**使用已有表、删除已有表并新建**两种方式 ○ 使用已有表：本次操作暂时忽略此表。 ○ 删除已有表并新建：本次操作中删除已有表，并根据转换生成的目标表 DDL 重新创建同名表。	2. 匹配成功的库/表策略配置： 
3	预览/编辑表语句	针对指定的目标表创建方式，系统将自动生成 DDL 样例。您可查看、编辑建表语句： 1. 目标表名称默认根据来源表名、目标表匹配策略自动生成。 2. 目标表字段默认与来源表名称保持一致。 3. 部分具有多个数据模型的目标数据源（如 doris），系统将根据默认策略指定相应的模型。您可手动修改 DDL 语句以符合业务特性。	1. 新建表 DDL 语句：  2. 删除已有表并新建DDL语句： 

3. 批量创建目标表

完成确认目标表创建规则后，可在本步骤中一次性创建目标表。

○ 创建成功后，您可单击**完成**关闭弹窗，并继续配置后续任务。

- 若存在创建失败的表，您可查看创建失败的原因或单击**重试**重新创建。

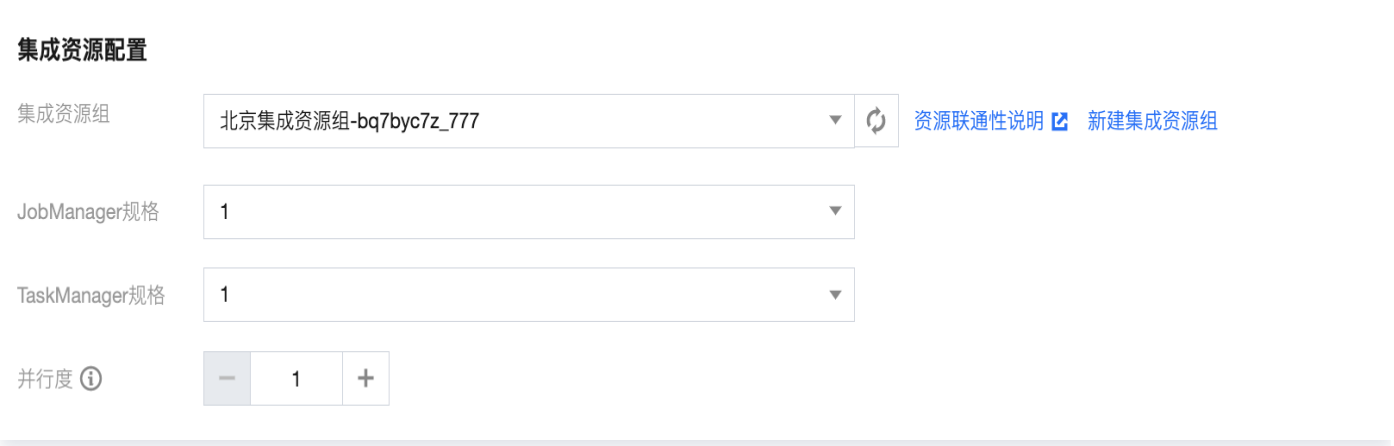


步骤六：配置运行资源和策略

当前步骤主要是为任务配置资源、DDL 和异常数据处理策略、以及高级运行参数等。

1. **集成资源配置**

   为当前任务关联对应的集成资源组，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。



2. **消息处理策略：**

   为当前任务配置包括写入 DDL 消息响应策略、数据异常写入策略、脏数据归档策略等。

> **说明：**
> 此处的配置项根据目标端数据源类型的不同而存在一定的差异，具体以各方案实际配置界面为准。

参数说明：

参数	说明
DDL 消息	该参数主要用于设置任务同步过程中，当来源端将捕获得 DDL 变更消息捕获后传递至下游，下游如何响应对应消息。目标端针对 DDL 消息提供了以下响应策略： 1. 自动响应：此策略下针对来源端捕获的消息，目标端将自动跟随响应源端的结构变化，包括自动建表、自动新增列等。 2. 忽略变更：此策略下，目标端将忽略 DDL 变更消息不做任何响应、以及消息通知等。 3. 日志告警：此策略下，目标端忽略 DDL 变更消息，但是日志中将提醒 DDL 变更消息详情。 4. 任务出错：此策略下，一旦源端出现 DDL 变更，整个任务将出现异常持续重启并报异常。

	说明： 不同来源及目标端支持 DDL 类型及消息处理存在差异，请参考不同链路支持配置策略为准。数据源支持 DDL 类型详情请参见 支持的 DML 及 DDL 操作 。
写入异常	该参数用于设置在同步过程中由于表段结构不匹配、字段类型不匹配等各种原因导致数据写入失败时，任务如何处理该异常写入数据、以及是否中断数据流。整体写入异常策略包含： 1. 部分停止：部分表写入异常时，仅停止该表数据写入，其他表正常同步。已停止的表不可在本次任务运行期间恢复写入。 2. 异常重启：部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，重启期间可能导致部分表数据重复写入。 3. 忽略异常：忽略表内无法写入的异常数据并标记为脏数据。该表的其他数据、以及任务内的其他表正常同步。脏数据提供COS 归档和不归档两种方案。
脏数据	当写入异常配置为**忽略异常**时，可选择针对忽略的数据是否进行归档： 1. COS 归档：统一将写入异常的数据归档至 COS 文件中保存，此方式下可避免异常写入数据丢失，可在后续场景中进行异常写入原因分析、数据补录等。 2. 不归档：任务完全忽略并丢弃写入异常的数据。

3. 任务运行策略：

为当前任务设置提交间隔、最大重启次数以及任务级运行参数等。

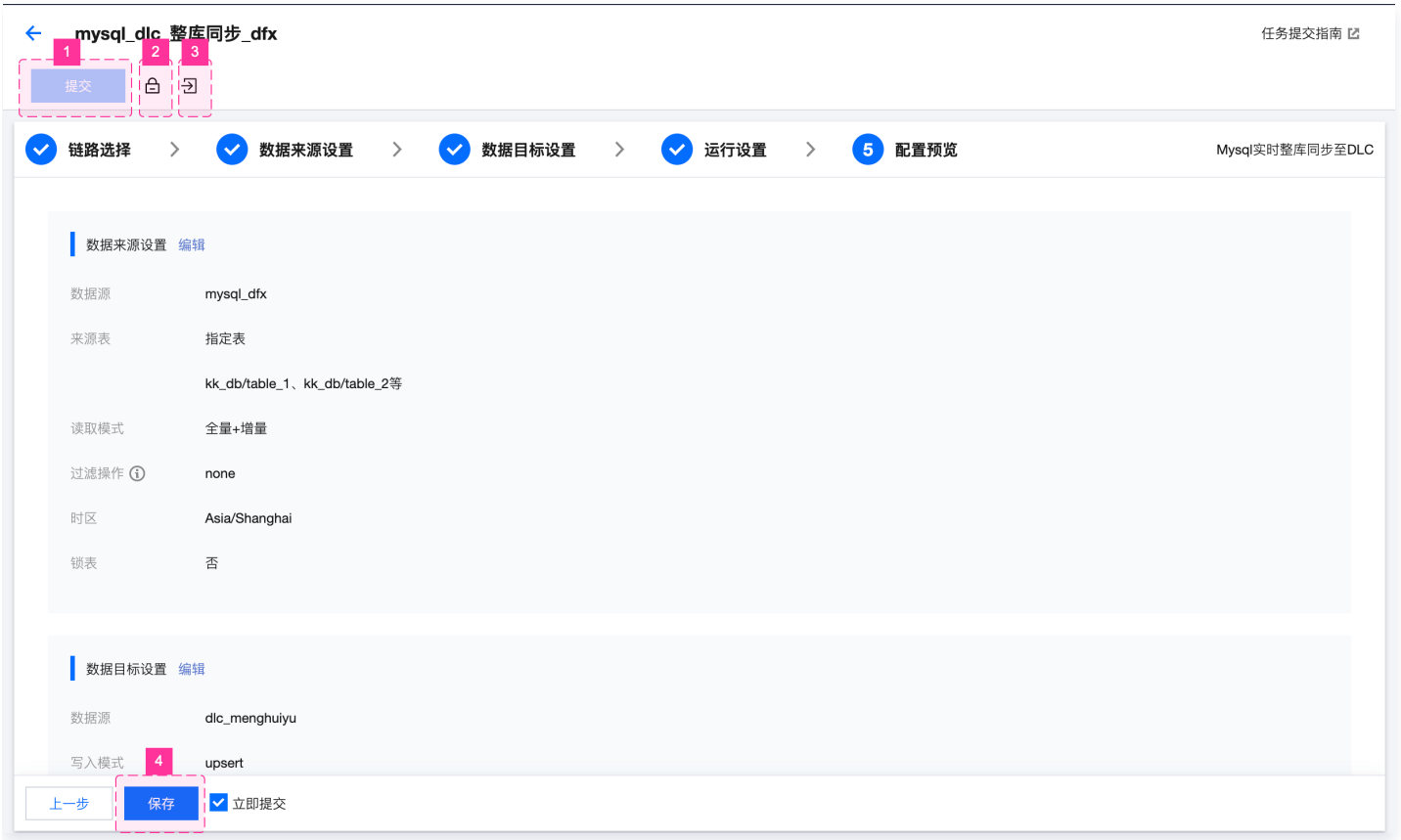
参数说明：

参数	说明
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]。其中 • 阈值为0表示不重启；

	• -1 表示不限制最大重启次数。
参数	设置任务级别运行参数。不同来源及目标端支持的任务级参数存在差异，详情请参见 实时节点高级参数 。

步骤七：配置预览及任务提交

1. 单击**提交**。



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时可选择不同提交策略及版本描述。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可单击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

2. 任务配置检测及提交。

步骤	步骤说明	

任务配置检测	本步骤将针对任务内读端、写端以及资源进行检测： • 检测通过：配置无误。 • 检测失败：配置存在问题，需修复以进行后续配置。 • 检测告警：此检测为系统建议修改项，修改完成后可单击**重试**重新检测；或者，您可以单击**忽略异常**进入下一步骤不阻塞后续配置。	

提交策略选择	本步骤中可选择本次任务提交策略：
	• **首次提交：首次提交任务支持从默认或指定点位同步数据**
	  ○ **立即启动，从默认点位开始同步**：若源端配置为“全量 + 增量”读取方式，则默认先同步存量数据（全量阶段），完成后消费 binlog 获取变更数据（增量阶段）；若源端配置为“仅增量”读取，则默认使用 binlog 最新位点开始读取。
	  ○ **立即启动，从指定时间点开始同步**：任务将根据配置的时间及时区同步数据。若未找到指定的时间位点，任务将默认从 binlog 最早位点开始同步；若源端读取方式为“全量 + 增量”，任务将默认跳过全量阶段从增量的指定时间位点开始同步。
	  ○ **暂不启动**：提交后暂不启动运行任务，后续可在运维列表内手动启动任务。
	• **非首次提交：支持带运行状态启动或继续运行任务**
	  ○ **继续运行**：此策略下新版本任务提交后，将从上次同步最后位点继续运行。
	  ○ **重新启动，从指定位点开始**：此策略下您可指定重新启动读取的位点，任务将忽略老版本从指定位点重新开始读取。若未找到指定的时间位点任务将默认从 binlog 最早位点开始同步。
	  ○ **重新启动，从默认位点开始运行**：此策略下将根据源端配置从默认位点开
		1. 首次提交
		 
		2. 非首次提交
		 

	始读取。若源端配置为“全量 + 增量”读取方式，则默认先同步存量数据（全量阶段），完成后即可消费 binlog 获取变更数据（增量阶段）；若源端配置为 “仅增量”读取，则默认使用 binlog 最新位点开始读取。 同时，每次提交都将新生成一个实时任务版本，您可在对话框内配置版本描述。	
任务提交	提交成功后，您可单击**前往运维**查看任务运行情况。	

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 DLC 配置详情

最近更新时间：2024-04-18 17:34:21

背景信息及特性支持

支持 MySQL、TDSQL-C MySQL、Kafka 内整个实例或库表数据实时同步至 DLC 原生表或外部表中：

- MySQL、TDSQL-C MySQL 支持 DDL 变更监控，MySQL 数据源支持实例、库、表级数据变更监控。
- Kafka 支持拆分消息内容动态匹配目标 DLC 表。
- DLC 目标端支持根据源端表、及字段变更自动创建库、自动创建表、自动创建字段。
- DLC 目标端支持小文件合并。

条件与限制

1. 仅支持同步 DLC iceberg 表，需使用 v2表设置表属性：{ "format-version", "2"}。
2. DLC 对来源端的 DDL 变更响应：
   - 仅支持同步新增列操作，且暂不支持在中间增加列。
   - 暂不支持同步删除列、列名变更、列类型变更、列位置变更等操作。
   - 自动建表时，对于 Mysql 历史表中的 bool 会映射成 int 类型，对于新增表中的 bool 会映射成 bool 类型。
3. 实时同步任务开启 DLC 小文件合并以后，将会消耗 DLC 计算资源。

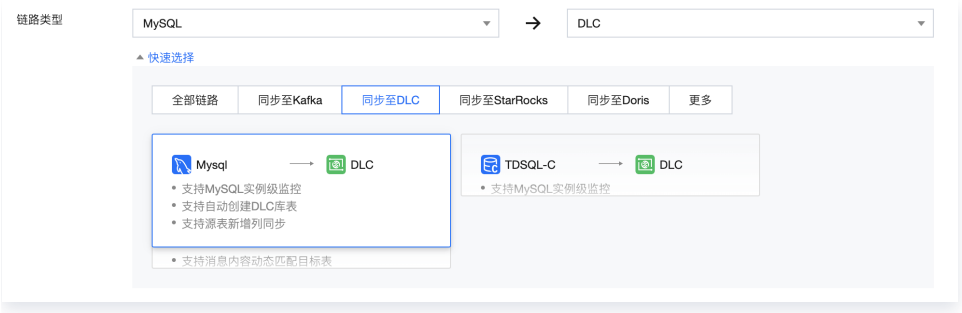
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

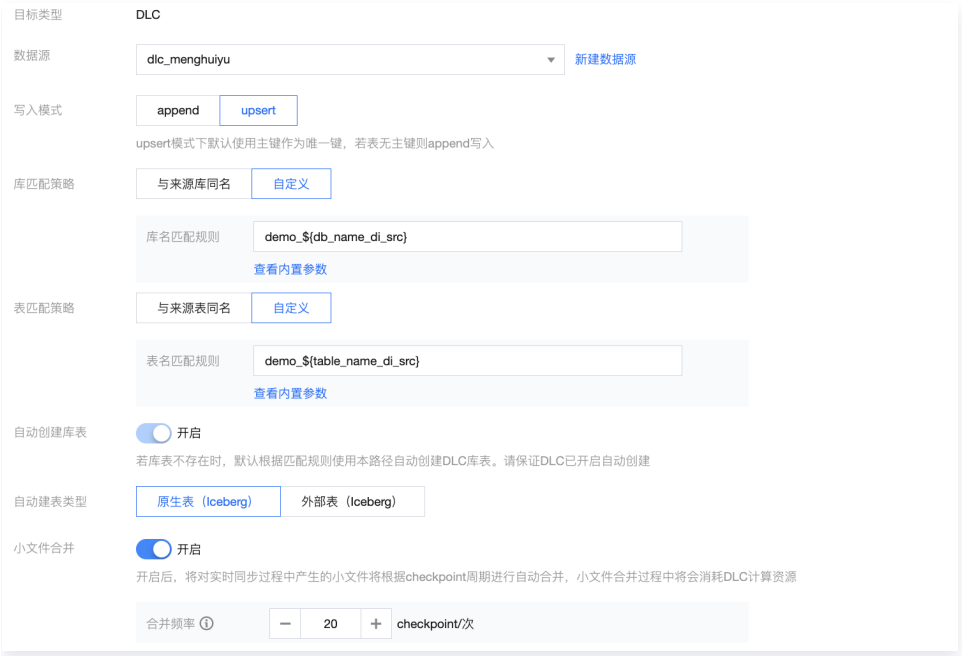
在首页卡片中选择同步至 DLC 目标端的链路。



步骤三：数据来源设置

- MySQL/TDSQL-C MySQL 来源
- Kafka 来源

步骤四：数据目标设置



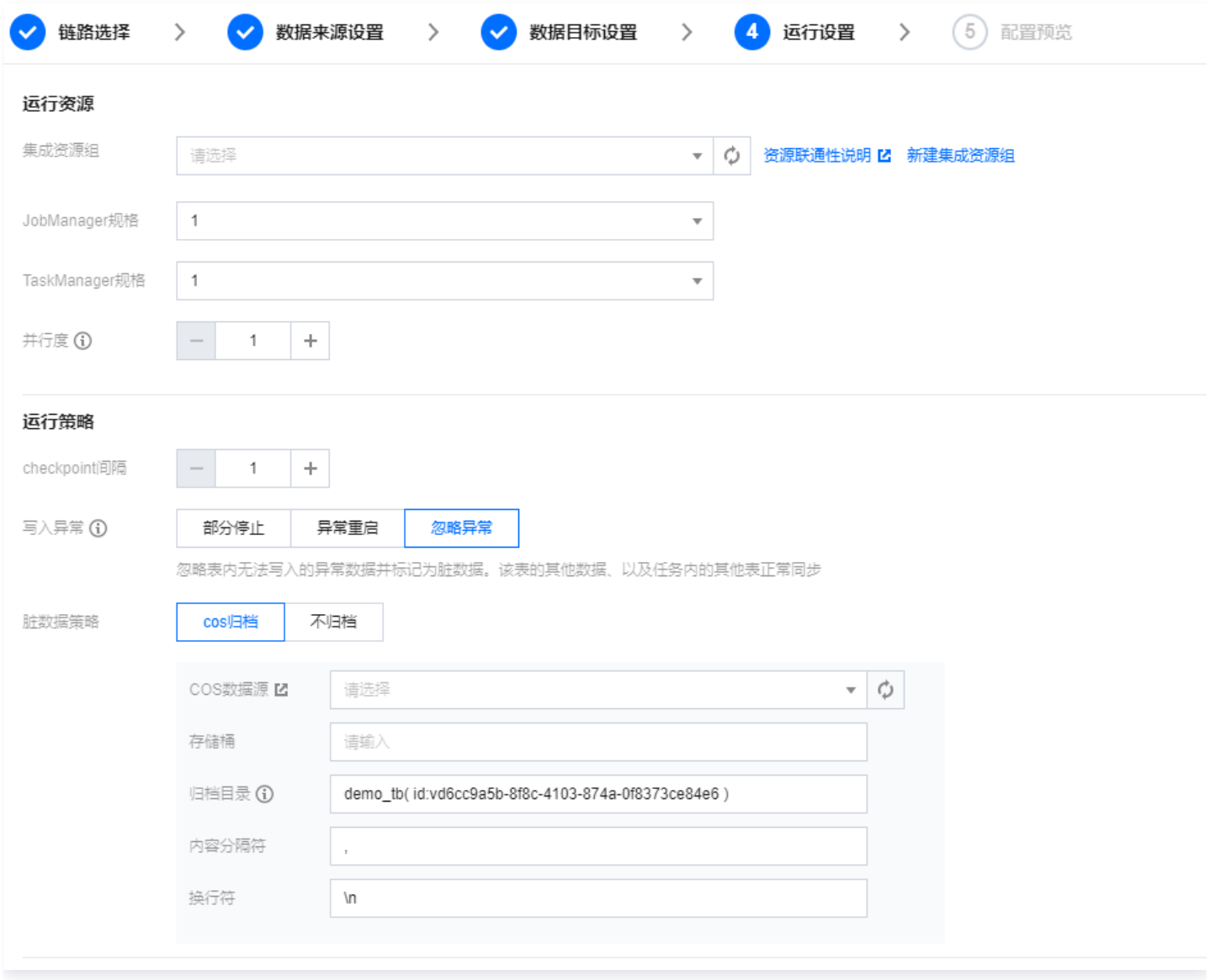
参数	说明
数据源	选择需要同步的目标数据源。

写入模式	• upsert：更新方式写入目标表。此方式下要求目标表中已设置主键：
	  ○ 任务将默认使用主键作为唯一键进行记录更新。
	  ○ 若表无主键则 append 写入。
	• append：追加模式写入数据表。

库/表匹配策略	DLC 中数据库以及数据表对象的名称匹配规则。
自动创建库表	默认开启。当目标 DLC 不存在符合**库/表匹配策略**的目标对象时，系统将自动在 DLC 内创建库表以接收来源库表的数据。 **说明：** 此能力要求 DLC 已开启自动创建库表能力。
自动建表类型	• 原生表：即 DLC 内部表（Iceberg）。 • 外部表：外部表需手动指定建 COS 路径，COS 路径需以 **cosn://** 开头。
小文件合并	默认关闭。开启后，将对实时同步过程中产生的小文件将根据 checkpoint 周期进行自动合并。小文件合并过程中将会消耗 DLC 计算资源，并需要输入合并频率。
高级设置	可根据业务需求配置参数。

步骤五：配置运行资源和策略

DLC 整库同步任务提供任务级运行资源及数据失败写入处理策略。其中数据写入失败处理策略支持三种：



策略名称	策略说明
默认策略	任意表写入异常时，所有表终止写入，任务将失败
部分停止	部分表写入异常时，仅停止该表数据写入，其他表正常同步。已停止的表不可在本次任务运行期间恢复写入。
异常重启	部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，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。该表的其他数据、以及任务内的其他表正常同步。脏数据提供 COS 归档和不归档两种方案。

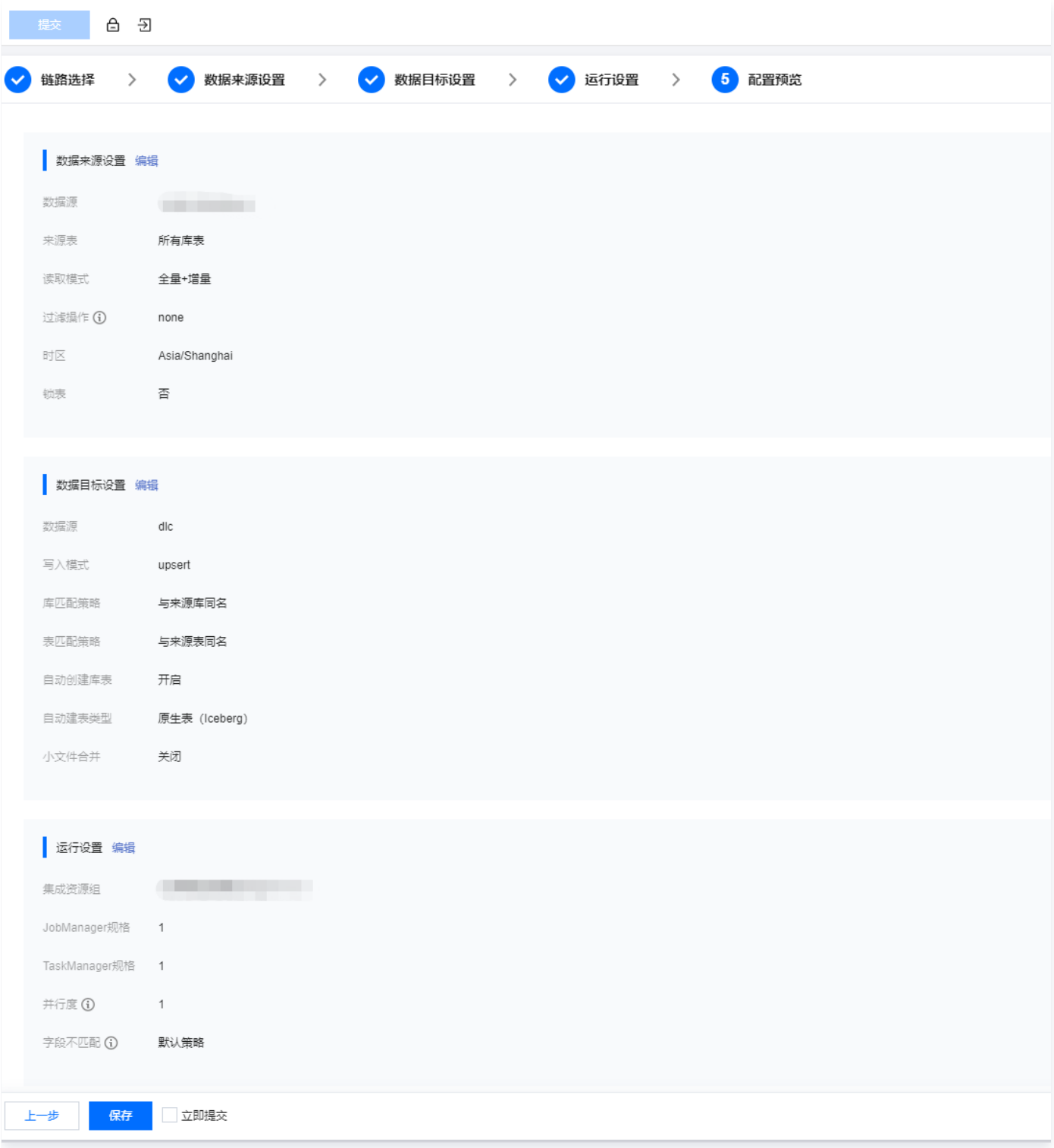
	• COS 归档：将无法写入的脏数据进行归档，需要配置 COS 数据源、存储桶、存储目录、内容分隔符及换行符。 • 不归档：不需要做其他操作。

场景示例：

任务 Task1下计划同步50张表，任务运行过程中表 A 内出现新增字段或字段类型变更：

- 默认策略：**表 A 任务运行后新增了一个字段 "DEMO" 。**此策略下，任务将在 DLC 端的目标表 A 内新建字段 "DEMO" 后同步数据。期间，其余49张表数据正常同步。
- 部分停止：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下，任务将在停止源端表A的数据读取，后续任务仅同步其余49张表至目标端。
- 异常重启：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将在持续重启，期间任务内配置的所有50张表将暂停数据写入，直到表 A 字段纠正。
- 忽略异常：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将忽略无法写入的异常数据，并标记为脏数据，表内其他数据正常同步。

步骤六：配置预览及任务提交



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

说明：
单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。

2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交
检测仅存在警告及以下	可直接提交

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 Doris 配置详情

最近更新时间：2024-04-07 17:44:51

背景信息及特性支持

支持 MySQL、TDSQL-C MySQL、Kafka、PostgreSQL 内整个实例或库表数据实时同步至 Doris 中：

- MySQL、TDSQL-C MySQL 支持 DDL 变更监控，MySQL 数据源支持实例、库、表级数据变更监控。
- Kafka 支持拆分消息内容动态匹配目标表。

条件与限制

- 全量同步阶段，Doris 仅支持同步至已有库表，任务运行前需保证目标库表已存在。
- 增量同步阶段，Doris 对于源端新建表、新建字段操作支持自动响应写入。具体数据源已支持 DDL 变更类型详情参见 支持的 DML 及 DDL 操作 。

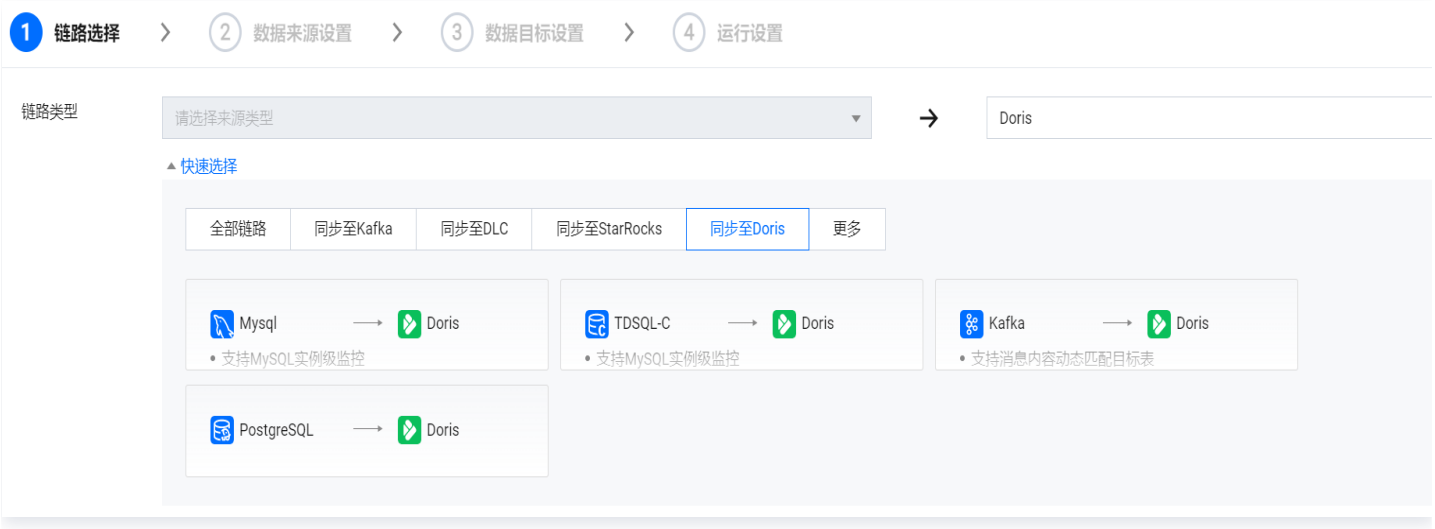
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

在首页卡片中选择同步至 Doris 目标端的链路。



步骤三：数据来源设置

- MySQL/TDSQL-C MySQL 来源
- Kafka 来源

- PostgreSQL 来源

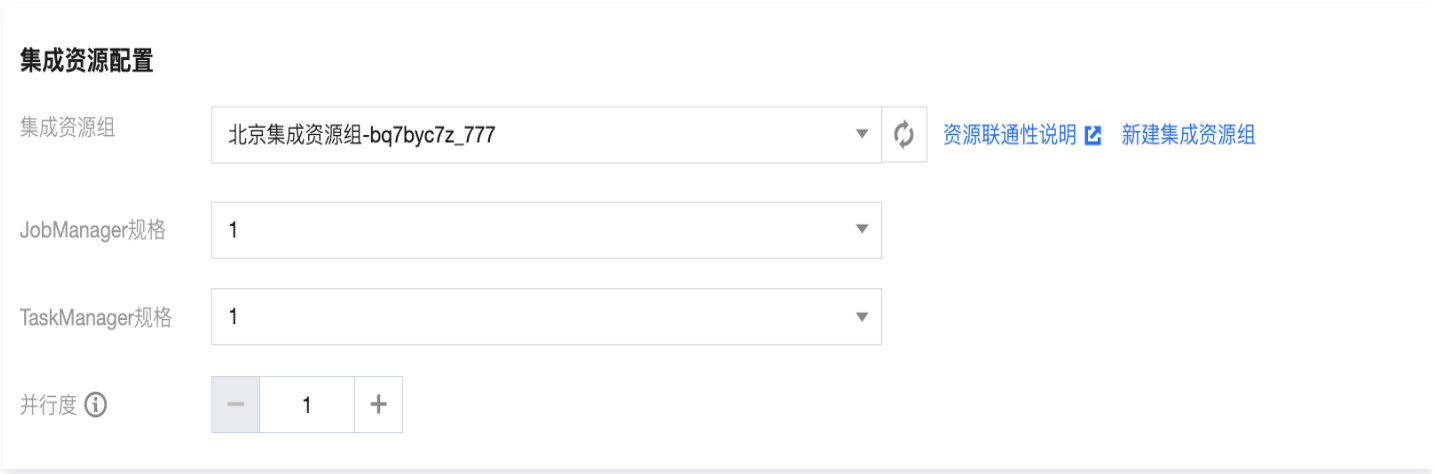
步骤四：数据目标设置

参数	说明
数据源	选择需要同步的目标数据源。
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。 **说明：** • 同步至 Doris 的整库任务全量同步阶段暂不支持自动建库/表，请提前在 Doris 端构建存量库、表对象以保证任务正常运行 。 • 增量数据同步阶段，任务支持自动创建表对象。
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 。

步骤五：配置运行资源和策略

- **集成资源配置**

  为当前任务关联对应的集成资源组，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**

> **注意：**
> 整库同步至 Doris 仅支持对 MySQL、TDSQL-C MySQL 支持 DDL 消息变更响应，具体数据源已支持 DDL 变更类型详情参见 支持的 DML 及 DDL 操作 。

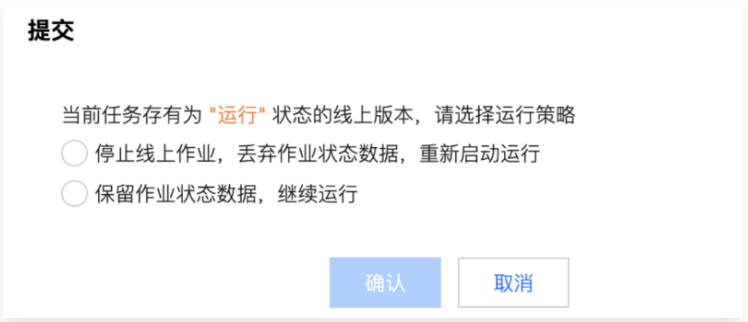
参数	策略名	策略说明
DDL 消息处理	新建表	1. 自动建表：当来源端被监控的库中出现新建表时，Doris 端将**自动创建同结构的表及字段**。针对 Doris 的不同数据模型： ○ 若来源端表包含主键，任务默认创建 Unique key 模型表。 ○ 若来源端表包含主键，任务默认创建 Duplicate 模型表。 2. 忽略变更：目标端忽略来源端的产生的 DDL 变更消息，**Doris 端及日志不做任何响应或消息提醒**。 3. 日志告警：目标端仅接收 DDL 变更消息，并在**日志内打印消息内容**，不触发新建表操作。

		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. 自动建表：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 重命名列 修改列 清空表	除新建表、新增字段外其他 DDL 变更消息不支持自动响应，目前提供了 **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。 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。
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。 **说明：** **示例：**若任务中总共包含50张表，其中表 A 由于字段类型与目标端对应字段类型无法匹配导致无法写入，**部分停止策略** 下任务将在后续同步过程中自动停止读取表 A 中的数据，其余**49张表将继续读写不受影响**。
	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。 **说明：**

		示例：若任务中总共包含50张表，其中表 A 由于字段类型与目标端对应字段类型无法匹配导致无法写入，**异常重启** 策略下任务将持续处于异常重启状态，**重启过程中、异常被更正前所有表写入均将受到影响**。
	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。 **说明：** **示例：**若任务中总共包含50张表，其中表 A 由于字段类型与目标端对应字段类型无法匹配导致无法写入，**忽略异常** 策略下该异常将被忽略，同时未写入数据将被标记统计为脏数据，**所有50张表后续读写均不受影响。**
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。

步骤六：配置预览及任务提交

序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略： • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

		 **说明：** **单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。**
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可单击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。

- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 Kafka 配置详情

最近更新时间：2024-04-07 17:44:51

背景信息及特性支持

支持 MySQL、TDSQL-C MySQL、PostgreSQL、Mongo 内整个实例或库表数据实时同步至 Kafka 中：

- MySQL、TDSQL-C MySQL、PostgreSQL 等来源端支持 DDL 变更监控，支持实例、库、表级数据变更监控。
- Kafka 目标端支持 Topic 自动创建，支持指定 Partition 分区策略。

条件与限制

- 如需使用自动创建 Topic 能力，请提前在 Kafka 服务端设置：

```
auto.create.topics.enable=true
```

- 开启自动创建 Topic 功能后，目标 Topic 需遵守 CKafka/kafka Topic 命名规则，以防止任务运行时 Topic 创建失败。
- Kafka 开启自动创建 Topic 时，请合理配置好分区数，避免造成性能问题。

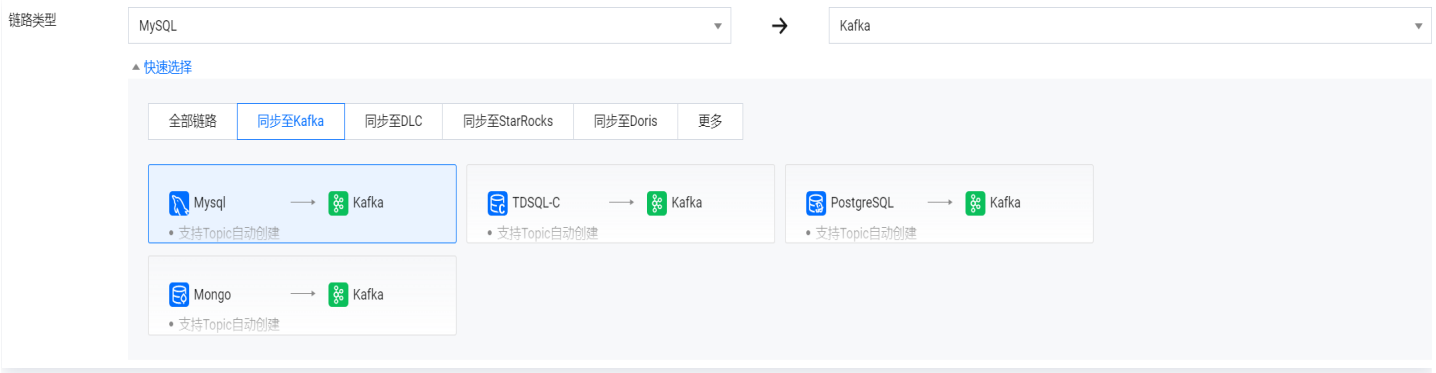
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

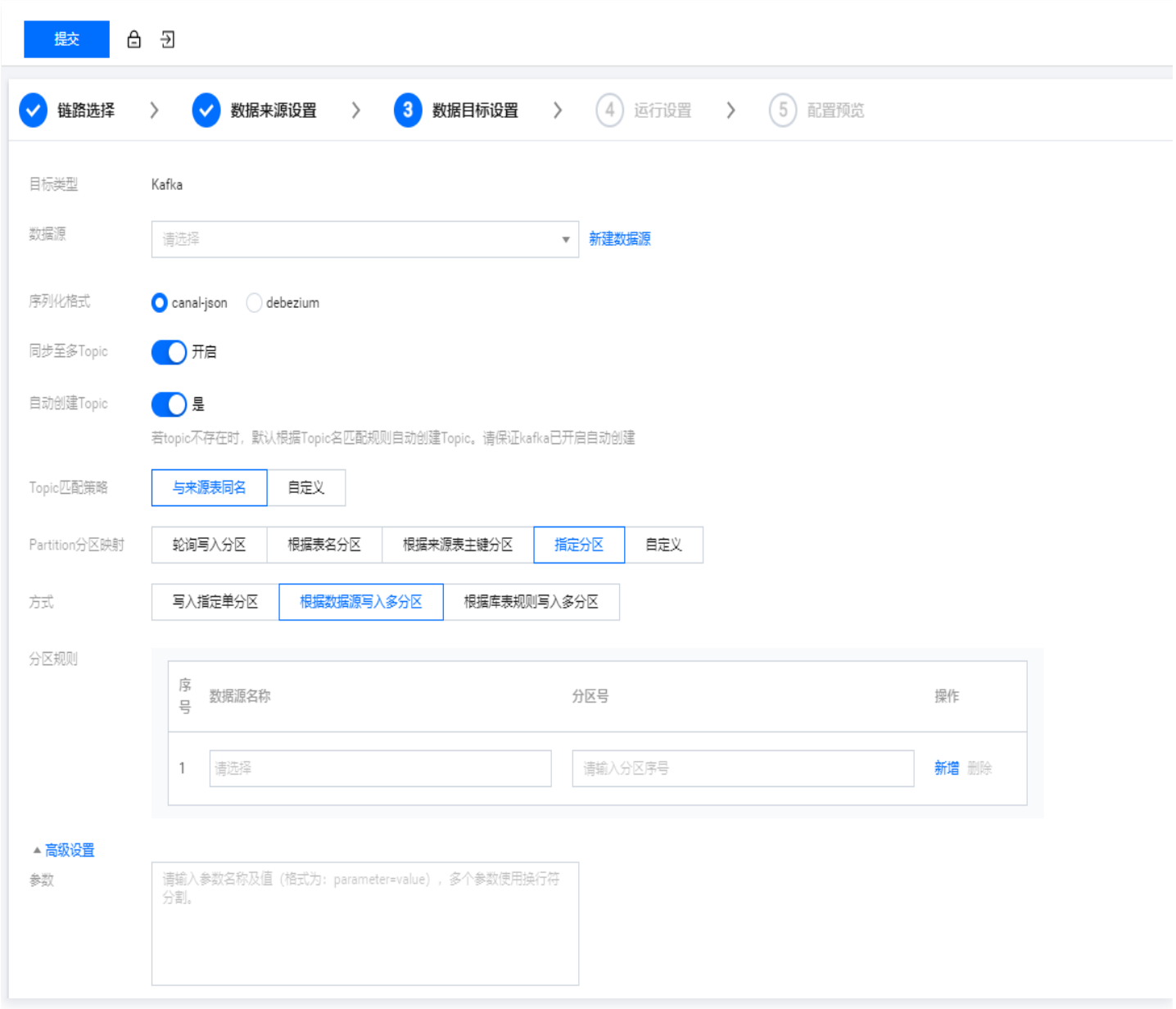
在首页卡片中选择同步至 Kafka 目标端的链路。



步骤三：数据来源设置

- MySQL/TDSQL-C MySQL 来源
- Mongo 来源
- PostgreSQL 来源

步骤四：数据目标设置

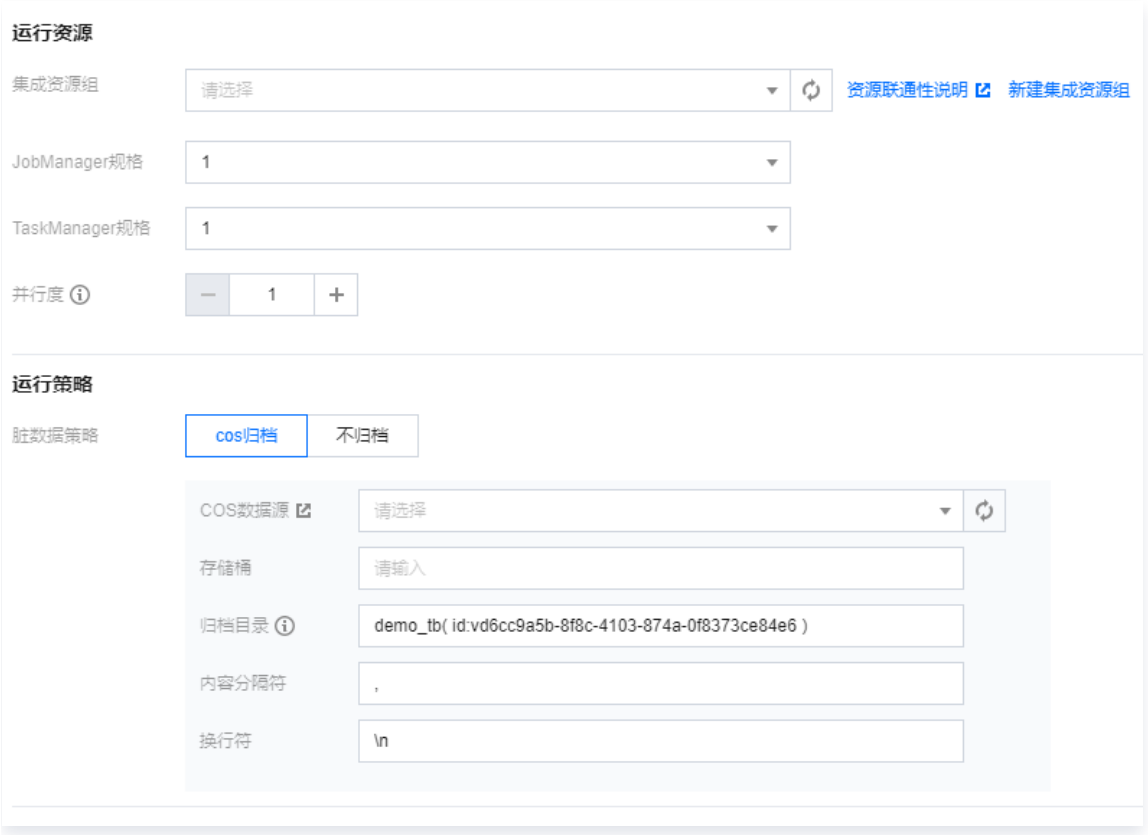


参数	说明
数据源	选择需要同步的目标数据源。
序列化格式	支持 canal-json 和 debezium 两种格式。

同步至多 Topic	• 默认打开：此选项下可实现来源数据与目标 Topic 多对多映射，任务执行过程中将根据策略匹配对应 Topic 的名称。 • 关闭：手动输入或者选择目标 Topic 名称，后续所有数据将统一写入该 Topic 内。
支持自动创建 Topic	• 打开后，若 Topic 不存在时系统将根据 Topic 名匹配规则自动创建 Topic。 **⚠ 注意：** 此功能需保证 Kafka 已开启自动创建，请提前在 Kafka 服务端设定： `auto.create.topics.enable=true`
Topic 匹配策略	• 与来源表同名：默认使用与来源表同名的 Topic 。 • 自定义：根据定义策略规则匹配 Topic 。

分区规则	配置 topic partition 分区映射（轮询写入分区、根据表名分区、根据来源表主键分区）：
	• 轮询写入分区：轮询（Round Robin）上游数据写入到每个 partition。
	• 根据表名写入分区：根据上游数据中的表名hash映射写入每个 partition。
	• 根据来源表主键分区：根据上游数据中的主键数据内容 hash 映射写入每个 partition。
	• 指定分区：
	  ○ 写入指定单分区：输入分区序号，所有消息仅写入到固定分区。
	  ○ 根据数据源写入多分区：同一行设置的数据源将统一写入对应分区：
	  ○ 数据源：数据源范围为所有来源端配置的数据源名称，支持多选，同行内，已选数据源不可重复选择。
	  ○ 分区号：输入分区序号。
	  ○ 支持新增/删除管理。
	  ○ 根据表规则写入多分区：支持输入库、表正则进行对象匹配，符合匹配规则的对象写入到指定分区中，规则之间顺序执行，已匹配库表不参与后续规则匹配。
	• 自定义：支持使用 “内置参数” 拼接写入分区规则，设定后将根据分区规则对应的值对消息进行 hash 分区。

步骤五：配置运行资源和策略



脏数据策略：提供 cos 归档和不归档两种方案。

- cos 归档：将无法写入的脏数据进行归档，需要配置 cos 数据源、存储桶、存储目录、内容分隔符及换行符。
- 不归档：不需要做其他操作。

步骤六：配置预览及任务提交

序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略： • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

> **说明：**
> 单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。

2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明

检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

附录：Canal-json/Debezium 数据格式样例

- Canal-json

```
{
  "data": [
    {
      "id": "2",
      "name": "scooter33",
      "description": "Big 2-wheel scooter233",
      "weight": "5.11"
    }
  ],
  "database": "pacino99",
```

```
  "es": 1589373560000,
  "id": 9,
  "isDdl": false,
  "mysqlType": {
    "id": "INTEGER",
    "name": "VARCHAR(255)",
    "description": "VARCHAR(512)",
    "weight": "FLOAT"
  },
  "old": [
    {
      "weight": "5.12"
    }
  ],
  "pkNames": [
    "id"
  ],
  "sql": "",
  "sqlType": {
    "id": 4,
    "name": 12,
    "description": 12,
    "weight": 7
  },
  "table": "products999",
  "ts": 1589373560798,
  "type": "UPDATE"
}
```

- **Debezium**

```
{
  "schema": {
    "type": "struct",
    "fields": [
      {
        "type": "struct",
        "fields": [
          {
            "type": "int32",
            "optional": false,
            "field": "id"
          },
          {
```

```
            "type": "string",
            "optional": false,
            "field": "first_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "last_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "email"
          }
        ],
        "optional": true,
        "name": "mysql-server-1.inventory2.customers2.Value",
        "field": "before"
      },
      {
        "type": "struct",
        "fields": [
          {
            "type": "int32",
            "optional": false,
            "field": "id"
          },
          {
            "type": "string",
            "optional": false,
            "field": "first_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "last_name"
          },
          {
            "type": "string",
            "optional": false,
            "field": "email"
          }
        ],
        "optional": true,
        "name": "mysql-server-1.inventory2.customers2.Value",
```

```
        "field": "after"
      },
      {
        "type": "struct",
        "fields": [
          {
            "type": "string",
            "optional": false,
            "field": "version"
          },
          {
            "type": "string",
            "optional": false,
            "field": "connector"
          },
          {
            "type": "string",
            "optional": false,
            "field": "name"
          },
          {
            "type": "int64",
            "optional": false,
            "field": "ts_ms"
          },
          {
            "type": "boolean",
            "optional": true,
            "default": false,
            "field": "snapshot"
          },
          {
            "type": "string",
            "optional": false,
            "field": "db"
          },
          {
            "type": "string",
            "optional": true,
            "field": "table"
          },
          {
            "type": "int64",
            "optional": false,
            "field": "server_id"
```

```
        },
        {
          "type": "string",
          "optional": true,
          "field": "gtid"
        },
        {
          "type": "string",
          "optional": false,
          "field": "file"
        },
        {
          "type": "int64",
          "optional": false,
          "field": "pos"
        },
        {
          "type": "int32",
          "optional": false,
          "field": "row"
        },
        {
          "type": "int64",
          "optional": true,
          "field": "thread"
        },
        {
          "type": "string",
          "optional": true,
          "field": "query"
        }
      ],
      "optional": false,
      "name": "io.debezium.connector.mysql.Source",
      "field": "source"
    },
    {
      "type": "string",
      "optional": false,
      "field": "op"
    },
    {
      "type": "int64",
      "optional": true,
      "field": "ts_ms"
```

```
        }
      ],
      "optional": false,
      "name": "mysql-server-1.inventory.customers.Envelope"
    },
    "payload": {
      "op": "c",
      "ts_ms": 1465491411815,
      "before": null,
      "after": {
        "id": 12003,
        "first_name": "Anne322",
        "last_name": "Kretchmar3222",
        "email": "annek@noanswer.org3222"
      },
      "source": {
        "version": "1.9.6.Final",
        "connector": "mysql",
        "name": "mysql-server-1",
        "ts_ms": 0,
        "snapshot": false,
        "db": "inventory333",
        "table": "customers433",
        "server_id": 0,
        "gtid": null,
        "file": "mysql-bin.000003",
        "pos": 154,
        "row": 0,
        "thread": 7,
        "query": ""
      }
    }
}
```

整库同步至 Iceberg 配置详情

最近更新时间：2024-04-07 17:44:51

背景信息及特性支持

支持 MySQL、TDSQL-C MySQL、Kafka、oracle 内整个实例或库表数据实时同步至 Iceberg 中：

- MySQL、TDSQL-C MySQL 支持 DDL 变更监控，MySQL 数据源支持实例、库、表级数据变更监控。
- Kafka 支持拆分消息内容动态匹配目标 Iceberg 表。
- Iceberg 目标端支持根据源端表、及字段变更自动创建库、自动创建表、自动创建字段。

条件与限制

- 使用 Iceberg 自动建表能力需设置表属性：

```
{
    "format-version", "2",
    "write.upsert.enabled", "true",
    "engine.hive.enabled", "true"
}
```

- 暂不支持在中间增加列。
- 仅支持同步新增列操作，暂不支持同步删除列、列名变更、列类型变更、列位置变更等操作。
- 自动建表时，对于 Mysql 历史表中的 bool 会映射成 int 类型，对于新增表中的 bool 会映射成 bool 类型。

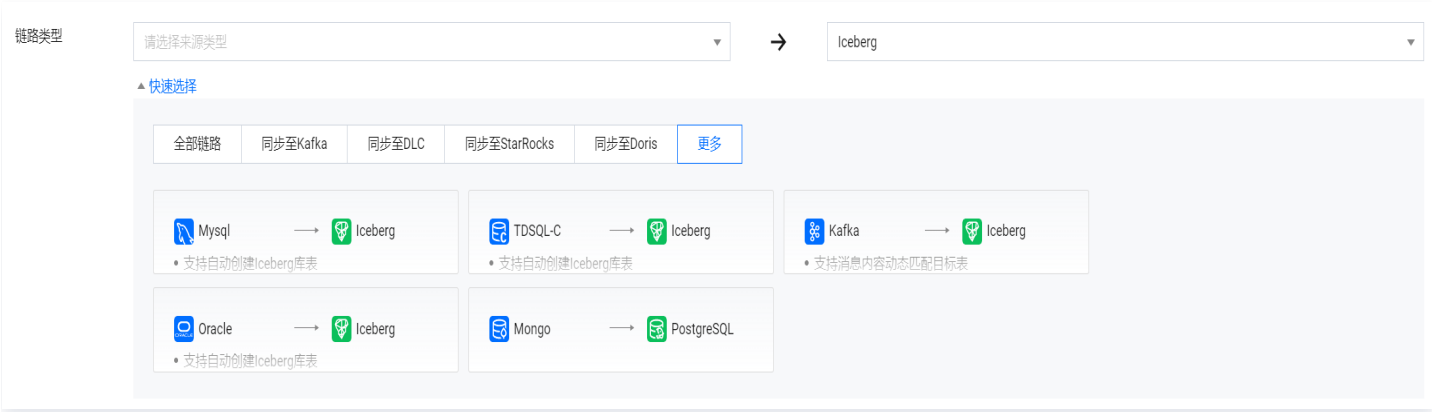
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

在首页卡片中选择同步至 Iceberg 目标端的链路。



步骤三：数据来源设置

- MySQL/TDSQL-C MySQL 来源
- Kafka 来源
- Oracle 来源

步骤四：数据目标设置

参数	说明
数据源	选择需要同步的目标数据源。

写入模式	• upsert：更新方式写入目标表。此方式下要求目标表中已设置主键：
	  ○ 任务将默认使用主键作为唯一键进行记录更新。
	  ○ 若表无主键则 append 写入。
	• append：追加模式写入数据表。

库表匹配策略	Iceberg 中数据库以及数据表对象的名称匹配规则。
高级设置	可根据业务需求配置参数。

步骤五：配置运行资源和策略

Iceberg 整库同步任务提供任务级运行资源及数据失败写入处理策略。其中数据写入失败处理策略支持四种：

策略名称	策略说明
默认策略	根据来源端字段变化自动添加目标端字段并同步数据
部分停止	部分表写入异常时，仅停止该表数据写入，其他表正常同步。已停止的表不可在本次任务运行期间恢复写入。

异常重启	部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，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。该表的其他数据、以及任务内的其他表正常同步。脏数据提供 COS 归档和不归档两种方案。 • COS 归档：将无法写入的脏数据进行归档，需要配置 COS 数据源、存储桶、存储目录、内容分隔符及换行符。 • 不归档：不需要做其他操作。

场景示例：

任务 Task1下计划同步50张表，任务运行过程中表 A 内出现新增字段或字段类型变更：

- 默认策略：**表 A 任务运行后新增了一个字段 "DEMO" 。**此策略下，任务将在 iceberg 端的目标表 A 内新建字段 "DEMO" 后同步数据。期间，其余49张表数据正常同步。
- 部分停止：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下，任务将在停止源端表 A 的数据读取，后续任务仅同步其余49张表至目标端。
- 异常重启：**表 A 任务运行后将字段 "DEMO "的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将在持续重启，期间任务内配置的所有50张表将暂停数据写入，直到表A字段纠正。
- 忽略异常：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将忽略无法写入的异常数据，并标记为脏数据，表内其他数据正常同步。

步骤六：配置预览及任务提交

序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略： • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

说明：
单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。

2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 PostgreSQL 配置详情

最近更新时间：2024-04-07 17:44:51

背景信息及特性支持

支持 Mongo 内整个实例或库表数据实时同步至 PG 中：

- Mongo 支持实例、库、表级数据变更监控。

条件与限制

- PG 仅支持同步至已有库表，任务运行前需保证目标库表已存在。
- PG 写入暂不支持自动建库、表、字段能力。

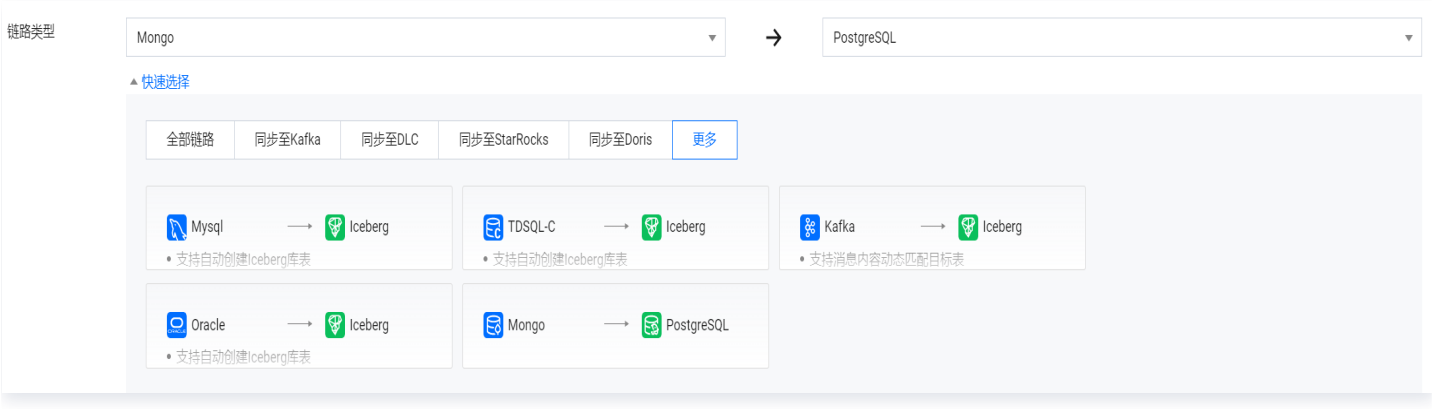
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

在首页卡片中选择同步至 PostgreSQL 目标端的链路。



步骤三：数据来源设置

- Mongo 来源

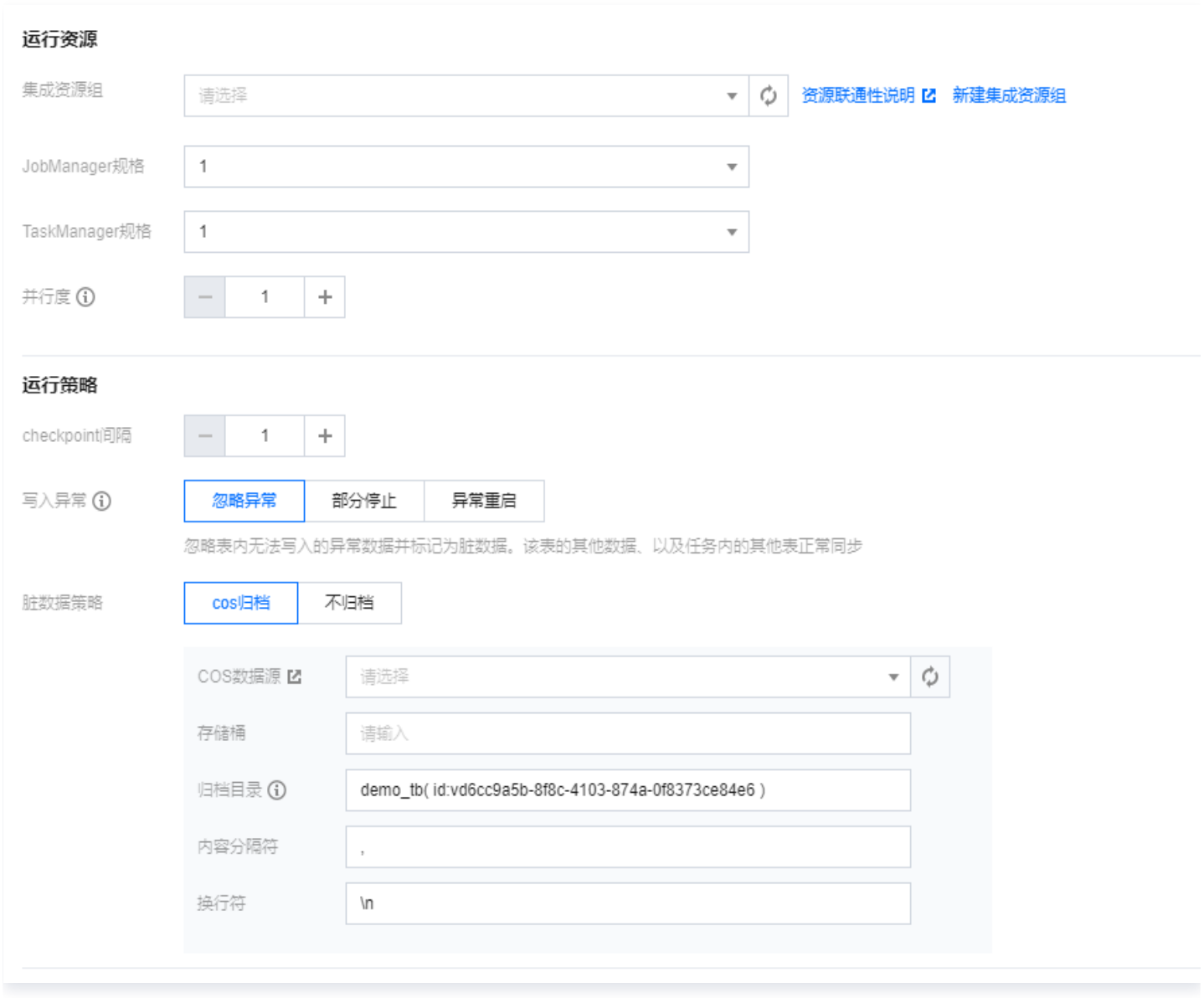
步骤四：数据目标设置

参数	说明
数据源	选择需要同步的 PostgreSQL 目标数据源。

写入模式	• upsert：更新方式写入目标表。此方式下要求目标表中已设置主键：
	  ○ 任务将默认使用主键作为唯一键进行记录更新。
	  ○ 若表无主键则 append 写入。
	• append：追加模式写入数据表。

库/Schema/表匹配策略	默认与来源库/表同名，也可以自定义配置，输入匹配规则即可（鼠标放至“查看内置参数”即可查看匹配规则）。

步骤五：配置运行资源和策略



PostgreSQL 整库同步任务提供任务级运行资源及数据失败写入处理策略。其中数据写入失败处理策略支持三种：

策略名称	策略说明
忽略异常	忽略表内无法写入的异常数据并标记为脏数据。该表的其他数据、以及任务内的其他表正常同步。脏数据提供 COS 归档和不归档两种方案。 • COS 归档：将无法写入的脏数据进行归档，需要配置 COS 数据源、存储桶、存储目录、内容分隔符及换行符。 • 不归档：不需要做其他操作。
部分停止	部分表写入异常时，仅停止该表数据写入，其他表正常同步。已停止的表不可在本次任务运行期间恢复写入。
异常重启	部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，重启期间

	可能导致部分表数据重复写入。

场景示例：

任务 Task1下计划同步50张表，任务运行过程中表 A 内出现新增字段或字段类型变更：

- 忽略异常：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将忽略无法写入的异常数据，并标记为脏数据，表内其他数据正常同步。
- 部分停止：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下，任务将在停止源端表A的数据读取，后续任务仅同步其余49张表至目标端。
- 异常重启：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将在持续重启，期间任务内配置的所有50张表将暂停数据写入，直到 schema 匹配。

步骤六：配置预览及任务提交

序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略： • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

说明：
单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。

2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 StarRocks 配置详情

最近更新时间：2024-08-15 21:18:51

StarRocks 背景信息及特性支持

支持 PostgreSQL 内整个实例或库表数据实时同步至 StarRocks 中。

条件与限制

- StarRocks 仅支持同步至已有库表，任务运行前需保证目标库表已存在。
- StarRocks 写入暂不支持自动建库、表、字段能力。

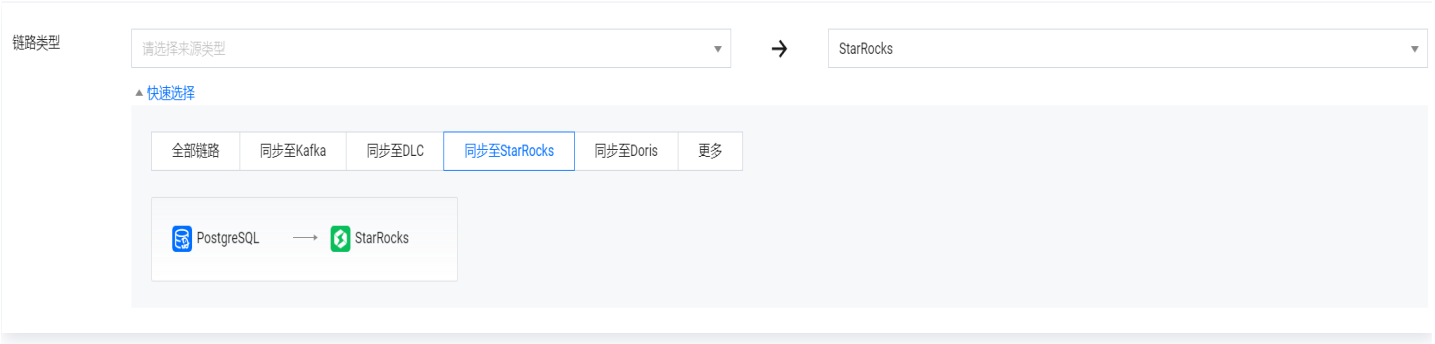
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

在首页卡片中选择同步至 StarRocks 目标端的链路。



步骤三：数据来源设置

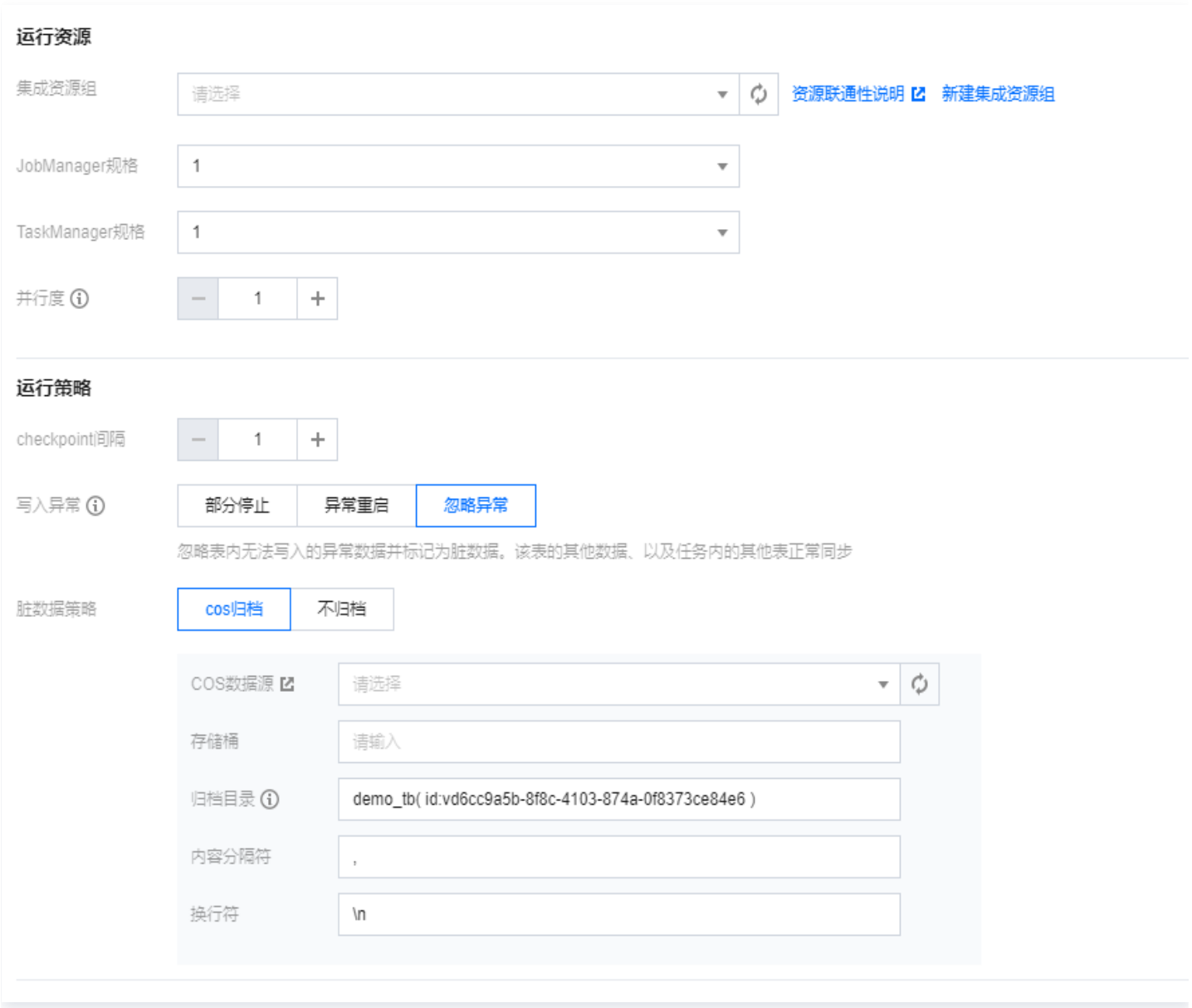
PostgreSQL 来源

步骤四：数据目标设置

参数	说明
数据源	选择需要同步的目标数据源。
库/表匹配策略	StarRocks 中数据库以及数据表对象的名称匹配规则。 **注意：** 请提前创建 StarRocks 库表，当前暂不支持自动创建 StarRocks 库表。

步骤五：配置运行资源和策略

StarRocks 整库同步任务提供任务级运行资源及数据失败写入处理策略。其中数据写入失败处理策略支持三种：



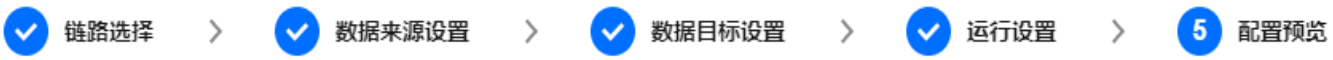
策略名称	策略说明
部分停止	数据无法写入目标表时丢弃数据，后续该异常表对应的数据自动丢弃不再同步。
异常重启	部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，重启期间可能导致部分表数据重复写入。
忽略异常	忽略表内无法写入的异常数据并标记为脏数据。该表的其他数据、以及任务内的其他表正常同步。脏数据提供 COS 归档和不归档两种方案。 COS 归档：将无法写入的脏数据进行归档，需要配置 COS 数据源、存储桶、存储目录、内容分隔符及换行符。 不归档：不需要做其他操作。

场景示例：

任务 Task1下计划同步50张表，任务运行过程中表 A 内出现新增字段或字段类型变更：

- **部分停止：表 A任务运行后新增了一个字段 "DEMO" 。**此策略下，任务将在 StarRocks 端的目标表 A 内新建字段 "DEMO" 后同步数据。期间，其余49张表数据正常同步。
- **异常重启：表 A 任务运行后将字段 "DEMO" 进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将在持续重启，期间任务内配置的所有50张表将暂停数据写入，直到表 A 字段纠正。
- **忽略异常：表 A 任务运行后将字段 "DEMO" 进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将忽略无法写入的异常数据，并标记为脏数据，表内其他数据正常同步。

步骤六：配置预览及任务提交



数据来源设置 编辑

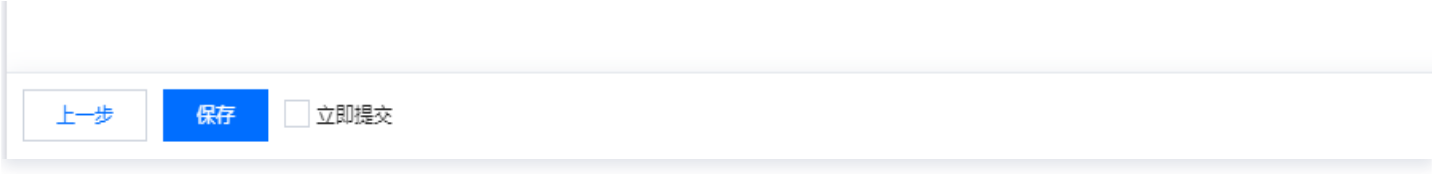
数据源	9700
来源表	所有库表
读取模式	全量+增量

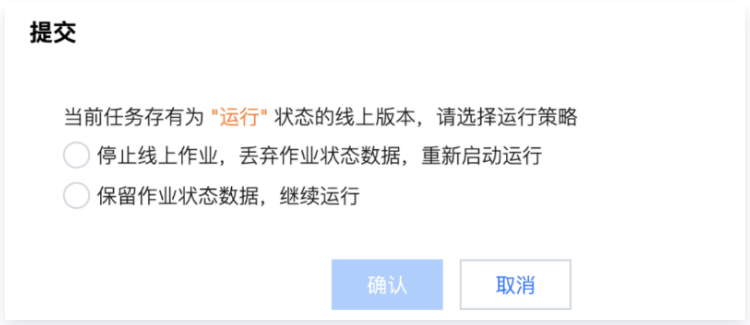
数据目标设置 编辑

数据源	9735
库匹配策略	与来源schema同名
表匹配策略	与来源表同名

运行设置 编辑

集成资源组	20230113214325476961
JobManager规格	1
TaskManager规格	1
并行度	1
字段不匹配	忽略异常



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略： • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。  **说明：** 单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。

- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

整库同步至 HIVE 配置详情

最近更新时间：2024-09-06 16:40:21

条件与限制

- HIVE 仅支持同步至已有库表，任务运行前需保证目标库表已存在。
- HIVE 写入暂不支持自动建库、表、字段能力。

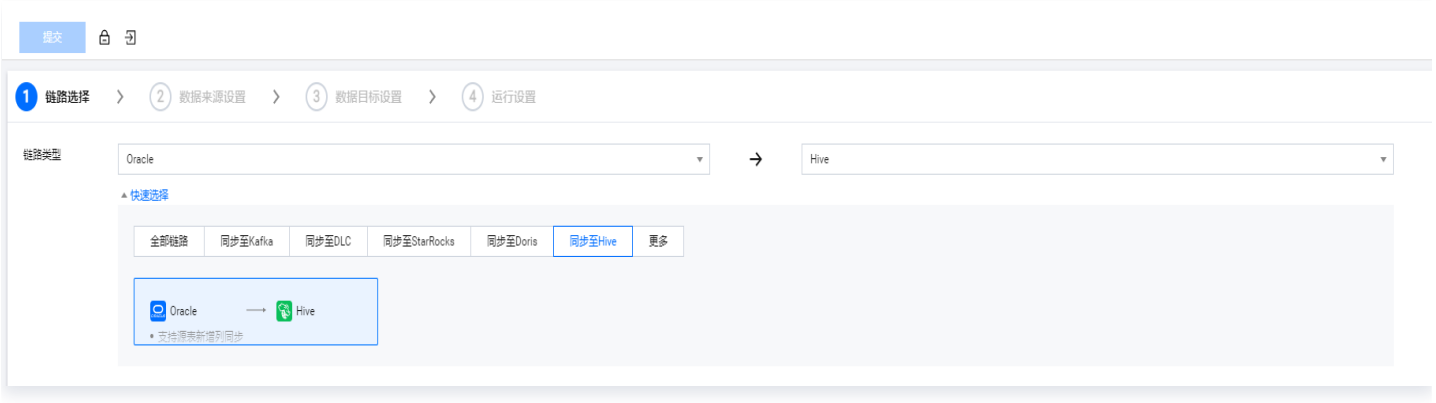
操作步骤

步骤一：创建整库同步任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**整库迁移任务。

步骤二：链路选择

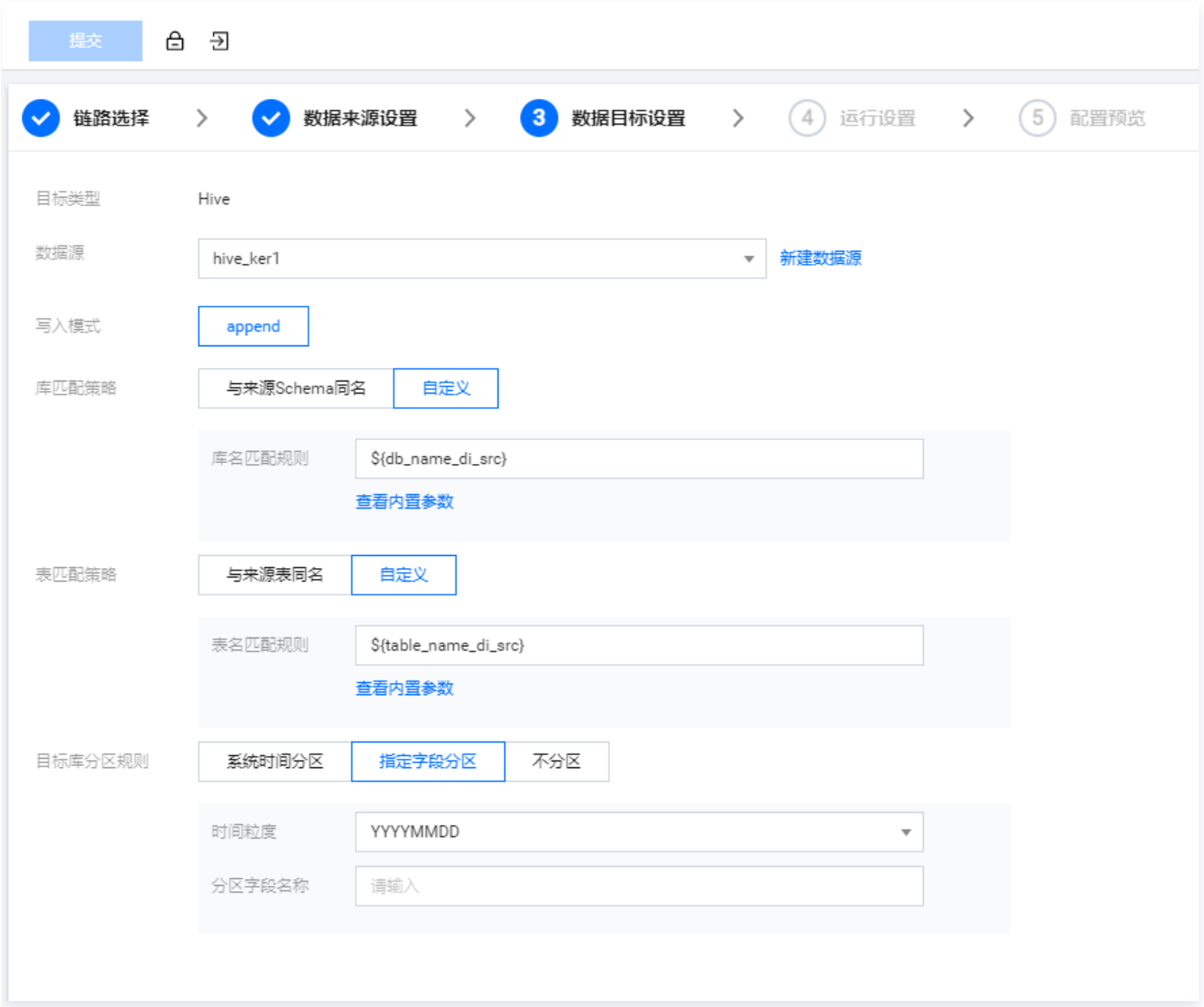
在首页卡片中选择同步至 HIVE 目标端的链路。



步骤三：数据来源设置

- Oracle 来源

步骤四：数据目标设置

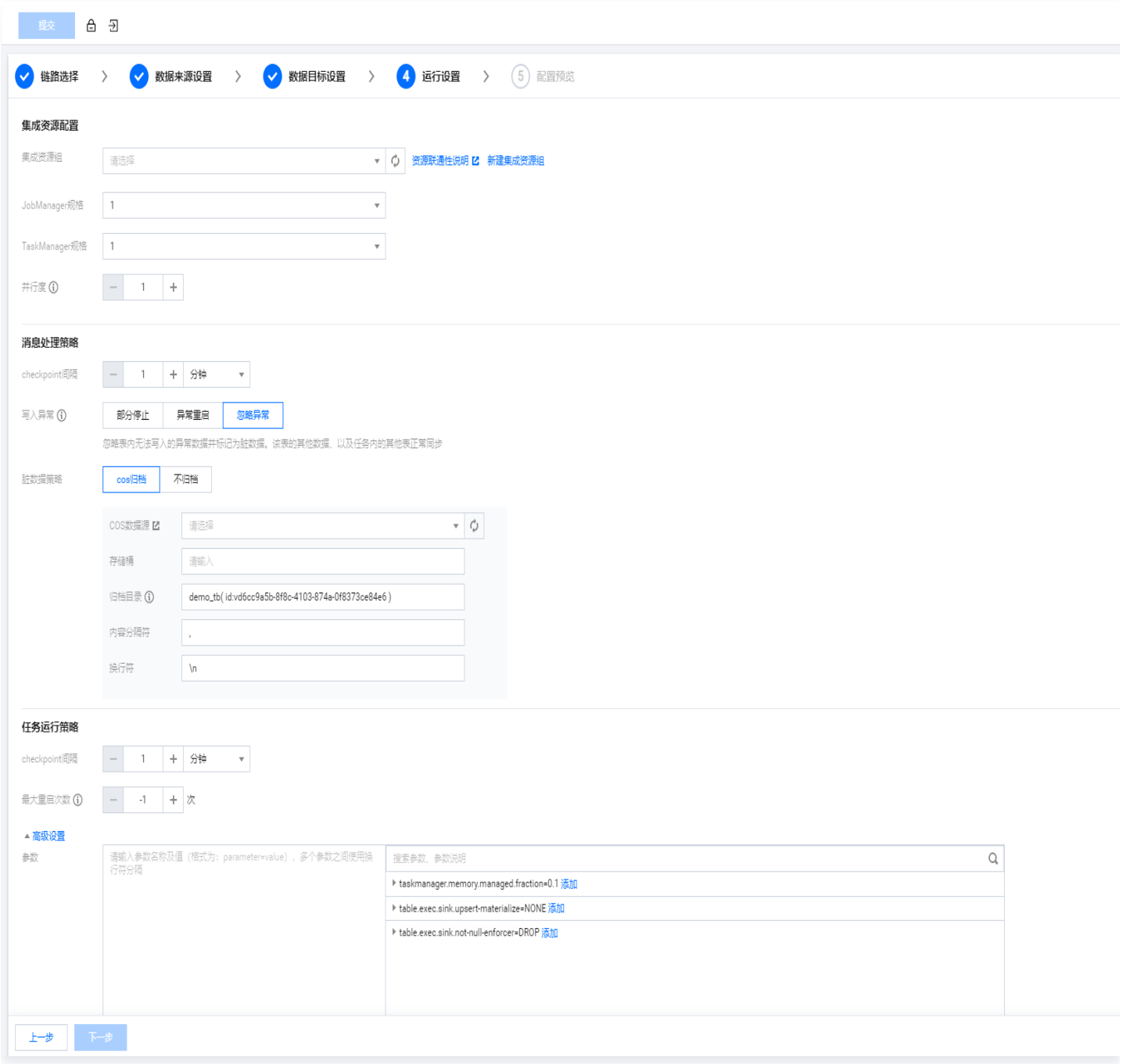


参数	说明
数据源	选择需要同步的目标数据源。
写入模式	append：追加模式写入数据表（当前仅支持此模式）。
库/表匹配策略	DLC 中数据库以及数据表对象的名称匹配规则： • 默认与来源 schema/来源表同名。 • 自定义：支持使用内置参数和字符串组合生成目标库表名称。 **说明：** 示例：如来源表名称为 table1，映射规则为 ${table_name_di_src}_inlong，则 table1 的数据将被最终映射写入至 table1_inlong 中。

目标库分区规则	系统时间分区： • 时间粒度：单选，用户可选择四种时间格式：YYYYMMDD、YYYYMM、YYYY、YYYY-MM-DD HH。 指定字段分区： • 时间粒度：单选，用户可选择四种时间格式：YYYYMMDD、YYYYMM、YYYY、YYYY-MM-DD HH。 • 分区字段名称：输入框，用户可输入指定分区字段的字段名称。 不分区：不做分区处理。

步骤五：配置运行资源和策略

HIVE 整库同步任务提供任务级运行资源及数据失败写入处理策略。其中数据写入失败处理策略支持三种：



策略名称	策略说明
默认策略	任意表写入异常时，所有表终止写入，任务将失败。
部分停止	部分表写入异常时，仅停止该表数据写入，其他表正常同步。已停止的表不可在本次任务运行期间恢复写入。
异常重启	部分表写入异常时，所有表均暂停写入。此策略下任务将持续重启直到所有表正常同步，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。该表的其他数据、以及任务内的其他表正常同步。脏数据提供 COS 归档和不归档两种方案。 • COS 归档：将无法写入的脏数据进行归档，需要配置 COS 数据源、存储桶、存储目录、内容分隔符及换行符。 • 不归档：不需要做其他操作。
checkpoint 间隔	可选择间隔：分钟/秒。
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。

场景示例：

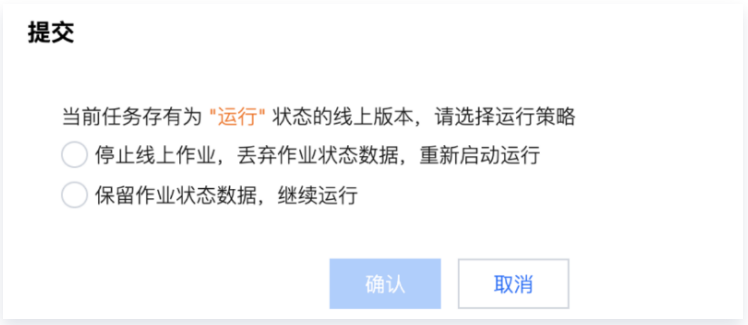
任务 Task1下计划同步50张表，任务运行过程中表 A 内出现新增字段或字段类型变更：

- 部分停止：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下，任务将在停止源端表 A 的数据读取，后续任务仅同步其余49张表至目标端。
- 异常重启：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将在持续重启，期间任务内配置的所有50张表将暂停数据写入，直到表 A 字段纠正。
- 忽略异常：**表 A 任务运行后将字段 "DEMO" 的进行了字段类型变更，并且变更后字段类型与目标端字段类型无法匹配写入。**此策略下任务将忽略无法写入的异常数据，并标记为脏数据，表内其他数据正常同步。

步骤六：配置预览及任务提交

序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继

续运行。



说明：
单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。

2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交
检测仅存在警告及以下	可直接提交

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。

- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

单表同步任务配置
单表任务配置概览

最近更新时间：2024-04-02 16:17:11

背景信息

实时单表同步通过 binlog 日志实时监控及同步源端数据。单表同步采用固定字段同步的方式，仅将在任务配置中指定映射关系的来源字段数据同步至目标端。单表任务支持画布、表单两种配置模式，覆盖 MySQL、Kafka、Mongo、SqlSever、Hive、DLC 等数据源。

条件与限制

1. 已配置好来源及目标端的数据源以备后续任务使用。详情请参见 数据源管理与配置方式 。
2. 已购买数据集成资源组。详情请参见 配置集成资源组 。
3. 已完成数据集成资源组与数据源的网络连通。详情请参见 集成连通性与使用规划 。
4. 已完成数据源环境准备。您可以基于您需要进行的同步配置，在同步任务执行前，授予数据源配置的账号在数据库进行相应操作的权限。
5. 若数据源配置的数据库账号不具备读写权限将导致任务运行失败，请根据实际读写场景配置具备相应权限的账号。

操作步骤

步骤一：创建实时同步任务

进入**配置中心 > 实时同步**任务页面后，单击**新建**单表同步任务。
在弹窗中配置任务基本信息，单击**确定**后即可进入任务配置页面。任务创建参数及说明如下：

参数	说明
任务名称	• 必填项 • 命名规则：仅支持中文、英文、数字和下划线 • 长度：不可超过100字符
任务模式	• 画布模式：主要采用可视化拖拽方式，适用于包含清洗环节、多对多数据链路。默认项 • 表单模式：适用于单表至单表离线同步，适用于 ODS 层无需数据清洗环节的数据同步
描述	选填项

说明：
- 当前版本，画布模式仅支持包含一个写入节点；单个读取节点仅支持连接一个下游节点。
- 任务中不可存在未连线的孤立节点，否则任务将提交失败。

- 数据节点默认命名规则为：${节点类型}_${编码}。

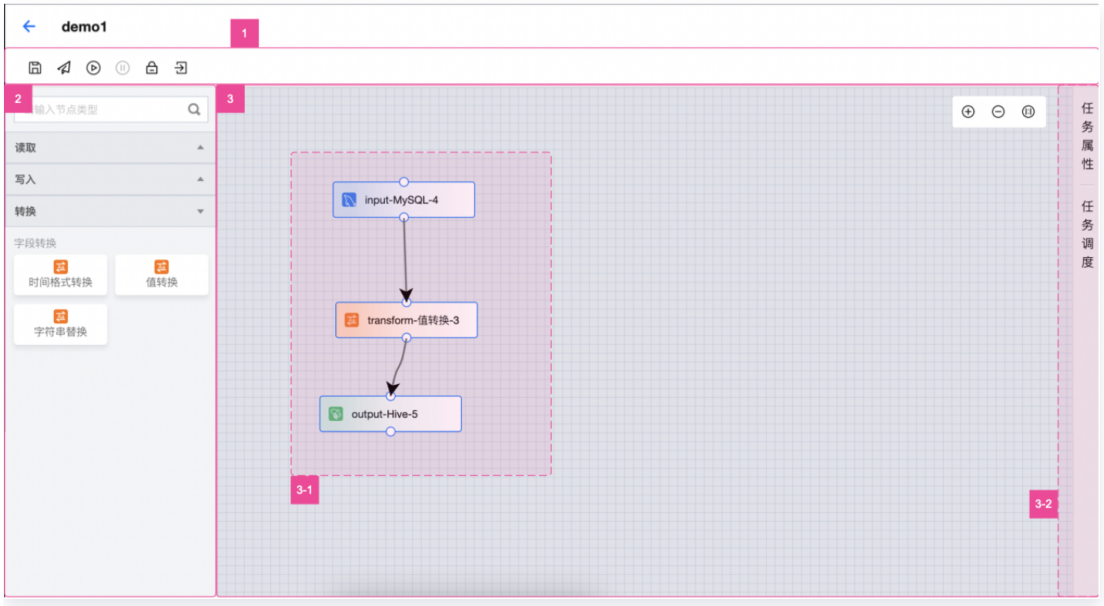
步骤二：选择任务配置模式

任务配置目前提供了表单和画布两种配置模式：

- 表单模式适用于贴源层数据同步，仅支持使用源端函数进行数据转换。
- 画布模式提供转换节点，支持在数据同步过程加入定制化的复杂数据转换。

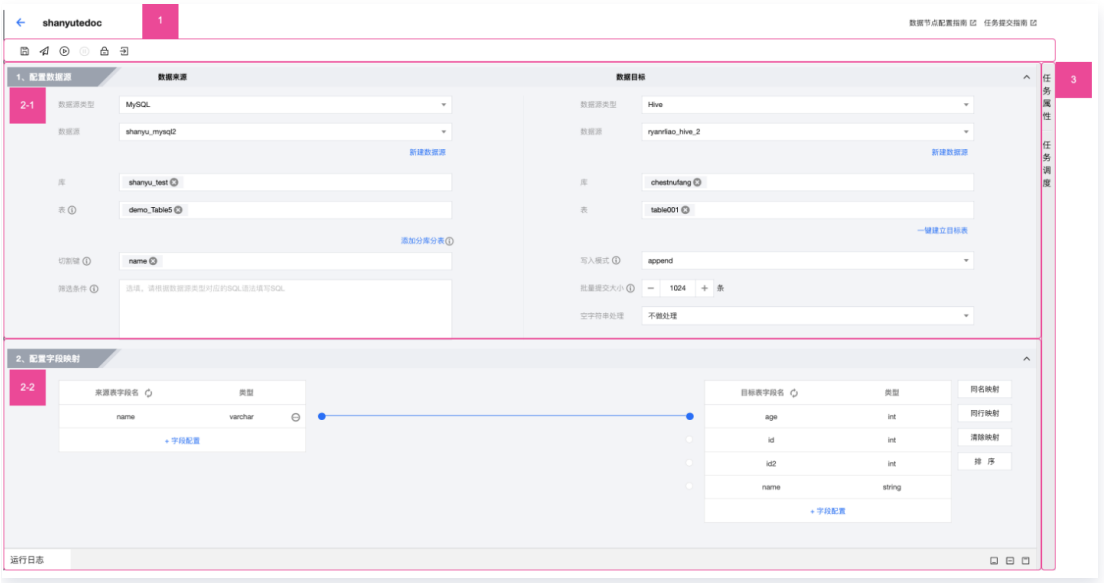
配置模式1：画布模式

在列表中单击**任务名称**即可进入任务配置页面，置界面总体包含任务操作栏、数据节点菜单、链路配置区三个部分：



序号		参数说明
1		任务操作栏。对整个任务生效的操作，包括保存、提交、测试运行、停止、解锁、前往运维等。
2		数据节点菜单。根据链路对象分类为读取、写入、转换节点，支持拖拽方式直接添加节点至画布。
3	3-1	数据链路。由读取、写入、转换节点及节点间连线构成的数据链路，代表了同步任务内数据流向。
	3-2	任务配置，此配置信息对全局任务生效： • 离线任务包括任务属性和调度配置两类，涵盖了基本信息、任务使用资源、数据通道控制等，详情请参见 离线同步 。 • 实时任务包括任务属性配置一类，详情请参见 实时同步 。

配置模式2：表单配置



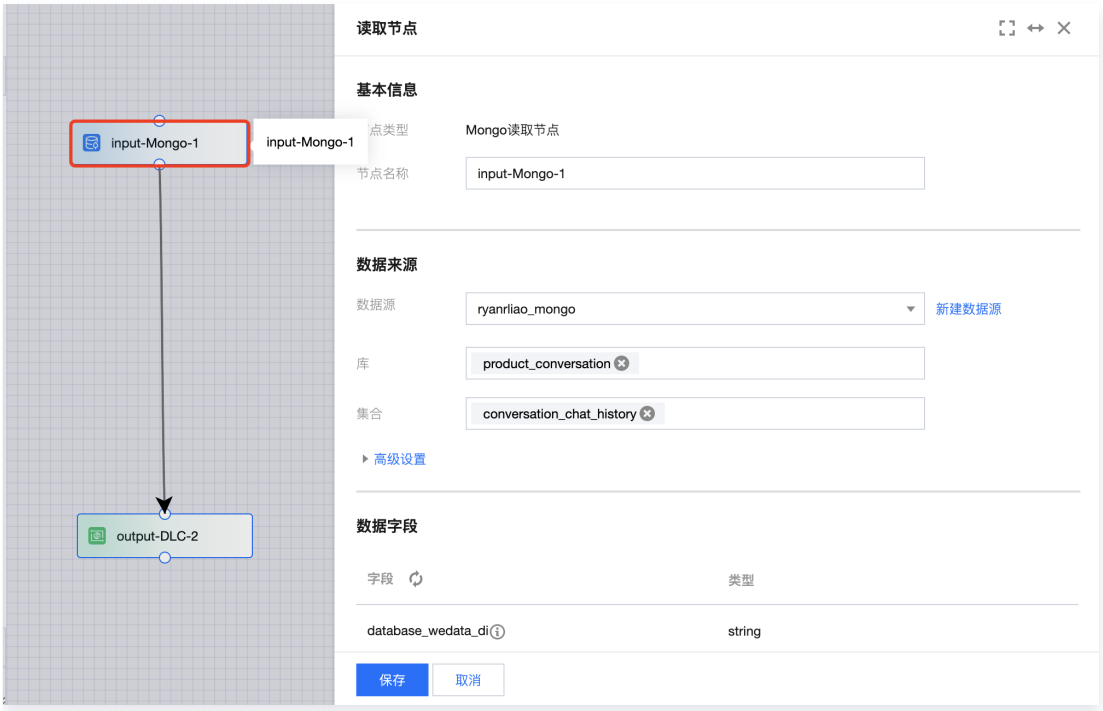
序号		参数说明
1		任务操作栏。对整个任务生效的操作，包括保存、提交、测试运行、停止、解锁、前往运维等。
2	2-1	数据来源及目标：配置任务读取和写入的数据源、库、表以及读写方式。
	2-2	字段映射：设置来源和目标端数据对应关系，后续任务仅同步具有映射关系的字段之间的数据。
3		任务配置，此配置信息对全局任务生效： • 离线任务包括任务属性和调度配置两类，涵盖了基本信息、任务使用资源、数据通道控制等。 • 实时任务包括任务属性配置，提供任务资源并发度、CU 用量配置等。

步骤三：数据节点配置

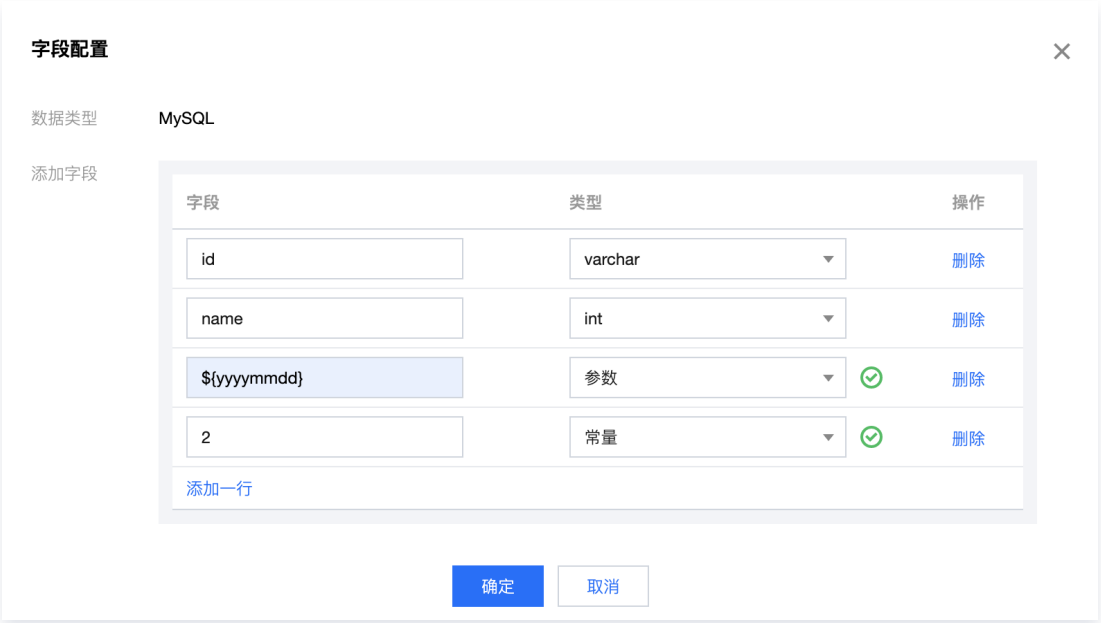
创建的新任务后在任务列表中单击**任务名称**即可进入任务画布或表单界面。画布模式下，可从右侧节点菜单直接拖拽数据节点及连线，系统将根据节点间连线关系自动创建数据流链路。

配置读取节点

读取节点配置包括基本信息、数据来源、数据字段三部分。



- **基本信息**
  节点名称不可为空，且单个任务内不可存在同名的数据节点。
- **数据来源**
  配置需要读取的库表对象以及同步方式等信息。
- **数据字段**
  根据配置的数据表对象，系统支持默认拉取字段元数据信息以及手动配置字段两种方式。
  - 默认拉取：针对 MySQL、Hive、PostgreSQL 等类型，系统已支持根据其库表信息自动拉取元数据字段及类型，无需手动编辑。
  - 手动配置：文件（如 HDFS、COS）以及列式存储数据源（如 HBase、Mongo）等数据源系统不支持自动拉取元数据，可单击**字段配置**手动添加字段名称及类型。读取节点还额外支持配置时间参数以及常量。



> **说明**
>
> - 时间参数字段：仅离线任务的读取节点支持配置时间参数字段，常用将实例运行时间值写入表的一级或多级分区 。
> - 常量字段：仅读取节点支持配置常量字段。常量字段可在来源与目标表字段个数不一致的情况下固定将某个常量值写入目标表。

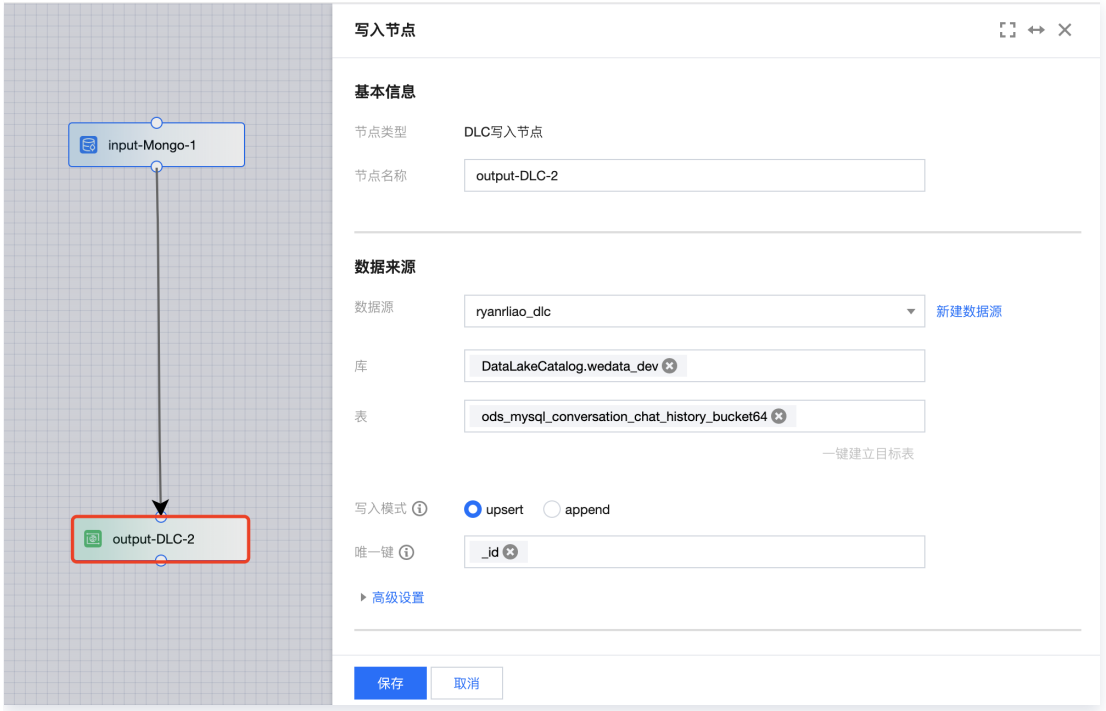
配置转换节点（可跳过）

转换节点配置包括基本信息、转换规则、数据字段三部分。其中，转换转换节点必须作为读取节点下游，在创建与读取节点连线后系统将自动获取上游节点内字段信息，同时根据转换规则完成数据转换。

- 基本信息
  配置节点名称信息。节点名称不可为空，且单个任务内不可存在同名的数据节点。
- 转换规则
  配置字段或数据级转换规则，其中字段信息继承自上游节点，在与上游节点连线后系统将自动获取上游节点内字段信息。
- 数据字段
  默认拉取上游节点全部数据字段用于后续写入节点映射。

配置写入节点

写入节点配置包括基本信息、数据来源、数据字段、字段映射四部分。写入节点将根据连线关系，将上游数据内容写入目标对象内。



- 基本信息
  节点名称不可为空，且单个任务内不可存在同名的数据节点。

- 数据来源
  配置需要读取的库表对象以及同步方式等信息。
- 数据字段
  根据配置的数据表对象，系统支持默认拉取字段元数据信息以及手动配置字段两种方式。
  - 默认拉取：针对 MySQL、Hive、PostgreSQL 等类型，系统已支持根据其库表信息自动拉取元数据字段及类型，无需手动编辑。
  - 手动配置：文件（如 HDFS、COS）以及列式存储数据源（如 HBase、Mongo）等数据源系统不支持自动拉取元数据，可单击**字段配置**手动添加字段名称及类型。
- 字段映射
  写入节点相对于读取节点需额外配置字段映射关系。字段映射关系旨在通过连线的方式指定目标字段内容的来源，支持同名映射、同行映射、以及手动连线三种方式配置来源与目标节点间关系。

> **说明**
> - 配置字段映射的前提为当前写入节点有已连线的来源（读取节点或转换节点）。
> - 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，可能会导致任务失败。

步骤四：实时任务属性配置

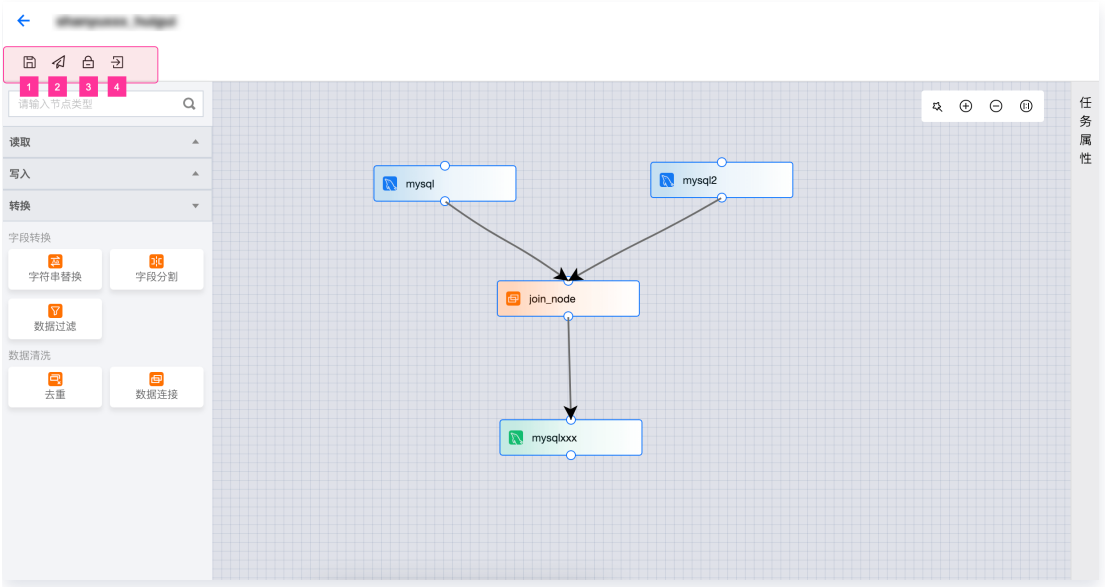
实时任务属性配置包括**基本属性**和**资源配置**两部分：


任务属性
基本属性
任务名称
mysql_hive_DML
任务类型
实时同步
责任人
描述
资源配置
集成资源组
上海集成资源组-ihytz1px | 实时包剩余CU:4.0/16
资源联通性说明
新建集成资源组
版本
v11
ManagerUrl
172.16.0.12:8083
资源分配方式
使用固定资源
按同步阶段分配
JobManager规格
1
TaskManager规格
1
并行度
1
运行策略
checkpoint间隔
1
分钟
最大重启次数
-1
次
任务属性

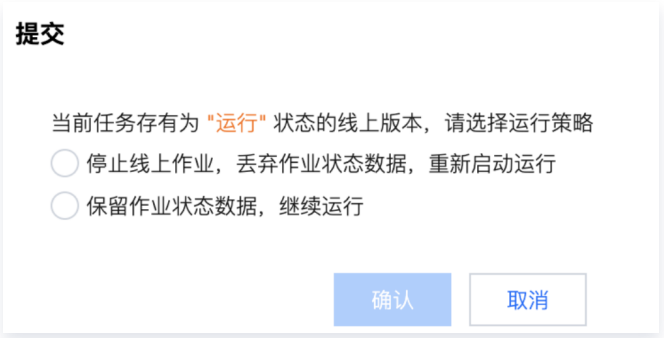
类别	参数	说明

任务属性	任务名称/类型	展示当前任务名称及类型基本信息 。
	责任人	对此任务负责的一个或多个空间成员名称，默认为任务创建者。
	描述	展示当前任务备注信息。
资源配置	集成资源组	指定当前任务使用的集成资源组名称，一个任务仅可绑定一个资源组。
	资源分配方式	集成资源支持多种分配方式： • 固定分配：此方式下不区分任务同步阶段，全量及增量同步过程中始终为当前任务分配固定资源量。此方式可避免任务间资源抢占，适用于任务运行过程中数据可能存在较大变动的场景。 • 按同步阶段分配：按全量和增量不同同步阶段分配计划的资源使用量，以节约整体资源用量。 
	JobManager	支持0.25、0.5、1、2C，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	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。
	并行度	每个算子的默认并行度。
运行策略	checkpoint 间隔	设置 checkpoint 间隔时间，支持分钟/秒级。
	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，任务状态将变为**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。

步骤五：任务提交



实时同步任务在配置完成后可配置运行策略并提交到生产环境中运行。目前可在任务配置页面支持保存、提交、锁定/解锁以及前往运维操作。

序号	参数	说明
1	保存	保存当前任务配置信息，包括数据节点配置、节点连线、任务属性和任务调度配置。
2	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。  **说明：** **单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。**
3	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。

4	前往运维	根据当前任务名称快捷跳转至任务运维页面。

任务提交检测

参数	说明
检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时运维 。

节点配置及参数说明

最近更新时间：2024-07-09 22:01:41

单表实时同步支持输入、输出和转换三种类型的数据节点，本文将重点说明各读、写及转换节点配置及底层数据类型映射。

读写节点配置

数据库类型	读取节点	写入节点
关系型数据库	MySQL 读取节点配置	MySQL 写入节点配置
	–	TDSQL-C MySQL 单表写入节点
	PostgreSQL 单表读取节点	PostgreSQL 单表写入节点
	SQL Server 单表读取节点	SQL Server 单表写入节点
	Oracle 单表读取节点	Oracle 单表写入节点
大数据	–	Hive 写入节点配置
	–	Greenplum 单表写入节点
	–	Tbase 单表写入节点
	–	DLC 单表写入节点
	–	Hbase 单表写入节点
	–	Iceberg 单表写入节点
	–	HDFS 单表写入节点
	–	Doris 单表写入节点
	–	ClickHouse 写入节点配置
消息队列	Kafka 读取节点配置	kafka 写入节点配置
	TiDB-kafka 单表读取节点	–
NoSQL	MongoDB 读取节点配置	–

	–	Elasticsearch 写入节点配置

转换节点配置

转换类型	节点
字段转换	字符串替换
	字段分割
数据清洗	数据过滤
	去重
	数据连接（join）

日志采集任务配置
日志采集配置概览

最近更新时间：2024-07-09 22:01:41

背景信息：

日志采集通过 Agent、SDK 方式主动上报 CVM 云实例、自建服务器或TKE内的日志文件数据至外部目标端。Agent 是 InLong 提供的轻量型日志采集器，可自动安装并运行于腾讯云 TKE、CVM 等云集群服务内，主动上报提供指定文件数据并实时同步到目标端。同时，inlong 还提供了 Java 以及 C++ SDK 进行数据上报。

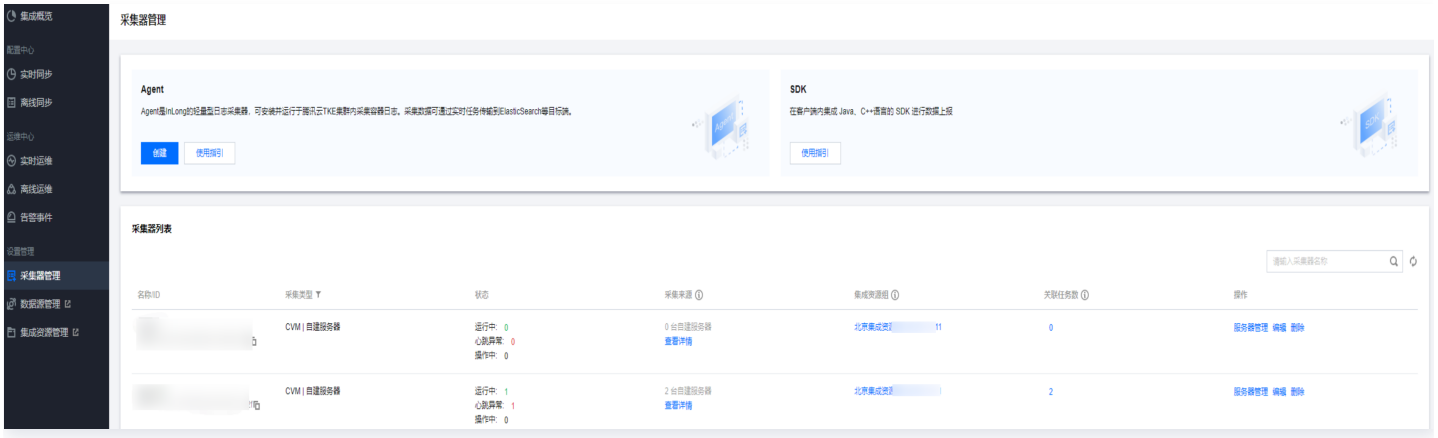
条件与限制：

1. 已配置好来源及目标端的数据源以备后续任务使用。详情请参见 数据源管理与配置方式。
2. 已购买数据集成资源组。详情请参见配置集成资源组。
3. 已完成数据集成资源组与数据源的网络连通。详情请参见 集成连通性与使用规划。
4. 已完成数据源环境准备。您可以基于您需要进行的同步配置，在同步任务执行前，授予数据源配置的账号在数据库进行相应操作的权限。
5. 若数据源配置的数据库账号不具备读写权限将导致任务运行失败，请根据实际读写场景配置具备相应权限的账号。

操作步骤

步骤一：创建采集器

采集器是 TKE 和 CVM 两种数据源类型的前置条件，在创建采集任务之前，需要提前创建好可用的采集器，用户可以在**设置管理** > 采集器管理 中进行创建和查看。



步骤二：创建同步任务

进入**配置中心** > **实时同步任务**页面后，单击**新建**日志采集任务。输入任务名称并选择配置模式。

任务配置目前提供了表单和画布两种配置模式：

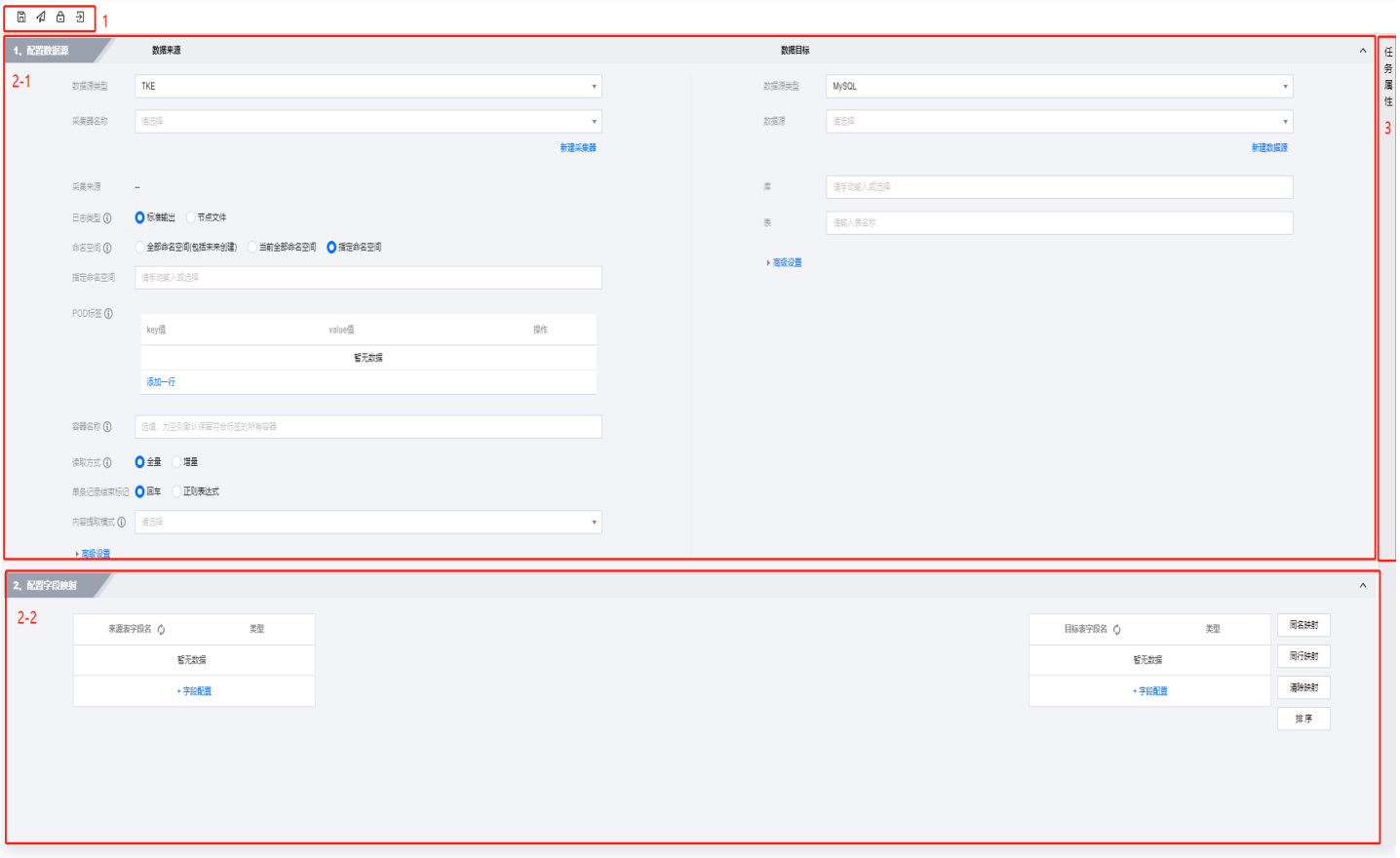
- 表单模式适用于贴源层数据同步，仅支持使用源端函数进行数据转换。
- 画布模式提供转换节点，支持在数据同步过程加入定制化的复杂数据转换。

配置模式1：画布模式

在列表中单击**任务名称**，即可进入任务配置页面，置界面总体包含任务操作栏、数据节点菜单、链路配置区三个部分：

序号		参数说明
1		任务操作栏。对整个任务生效的操作，包括保存、提交、测试运行、停止、解锁、前往运维等。
2		数据节点菜单。根据链路对象分类为读取、写入、转换节点，支持拖拽方式直接添加节点至画布。
3	3-1	数据链路。由读取、写入、转换节点及节点间连线构成的数据链路，代表了同步任务内数据流向。
	3-2	任务属性配置，此配置信息对全局任务生效，主要包含基本属性和资源配置。

配置模式2：表单配置



序号		参数说明
1		任务操作栏。对整个任务生效的操作，包括保存、提交、测试运行、停止、解锁、前往运维等。
2	2-1	数据来源及目标：配置任务读取和写入的数据源、库、表以及读写方式。
	2-2	字段映射：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。
3		任务属性配置，此配置信息对全局任务生效，主要包含基本属性和资源配置。

步骤三：配置数据来源

不同数据来源配置步骤略有差异，详情如下：

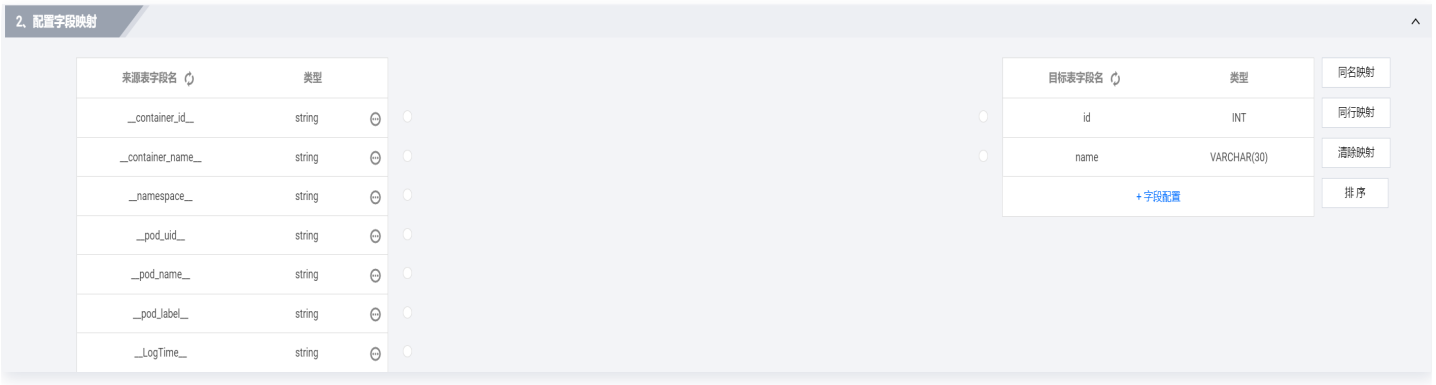
- TKE 来源采集任务配置
- SDK 来源采集任务配置
- CVM 来源采集任务配置

步骤四：配置数据目标

数据来源	已支持目标数据源
TKE / SDK / CVM	MySQL 写入节点
	TDSQL-C MySQL 写入节点
	PostgreSQL 写入节点
	SQL Server 写入节点
	Oracle 写入节点
	HIVE 写入节点
	Clickhouse 写入节点
	Greenplum 写入节点
	TBase 写入节点
	DLC 写入节点
	HBase 写入节点
	Iceberg 写入节点
	HDFS 写入节点
	Doris 写入节点
	Elasticsearch 写入节点
	Kafka 写入节点

步骤五：配置字段映射

1. 配置好数据来源和数据目标后，则会展示来源表和目标表的字段信息，我们需要对字段进行映射，支持同名映射和同行映射两种映射方式，并可以对字段进行排序和配置。



2. 单击**字段配置**则对字段进行配置：

- **表单配置**：可以对字段名称、类型进行更改，也可以删除字段和新增字段。

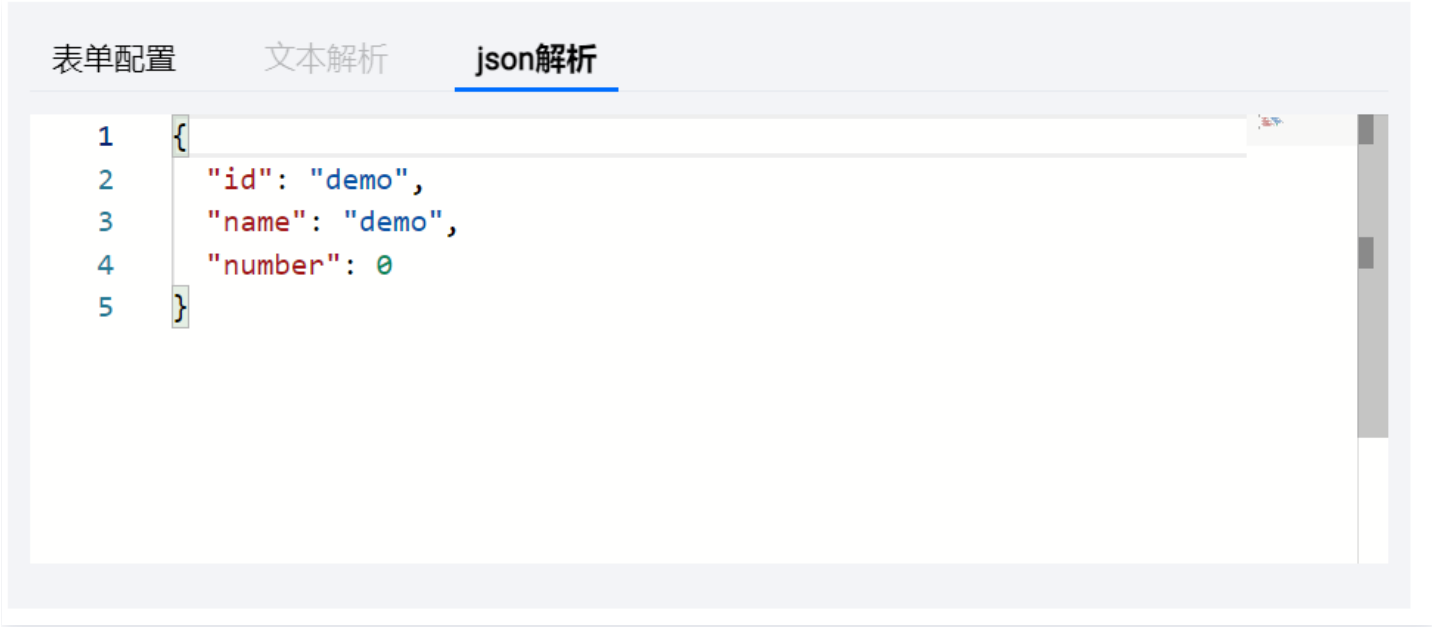
> **注意：**
> 文本或 JSON 提取内容将覆盖除内置元数据字段外的数据字段。内置元数据字段不支持修改和编辑。

- **文本解析**：对已有字段进行解析，也可以手动新增字段。

> **注意：**
> 一行默认为一个字段及类型，字段名称和类型使用设定的分割符号分割，如 age int。提醒：首尾空行会被撷取，空行会被忽略。

- **JSON 解析**：数据为 JSON 格式时（如{"age":10,"name":"demo"}），系统将自动提取字段名并解析值类型。

> **注意：**
> 当前系统仅支持解析部分类型，可在表单模式下确认并调整解析结果。重复字段保留最后一条。



步骤六：配置任务属性

单击右侧**任务属性**，进入基本属性界面，配置正确的基本属性和集成资源组即可。

任务属性

基本属性

任务名称

任务类型 实时同步

责任人

描述

资源配置

集成资源组 请选择

资源联通性说明 新建集成资源组

版本 --

ManagerUrl --

JobManager规格 1

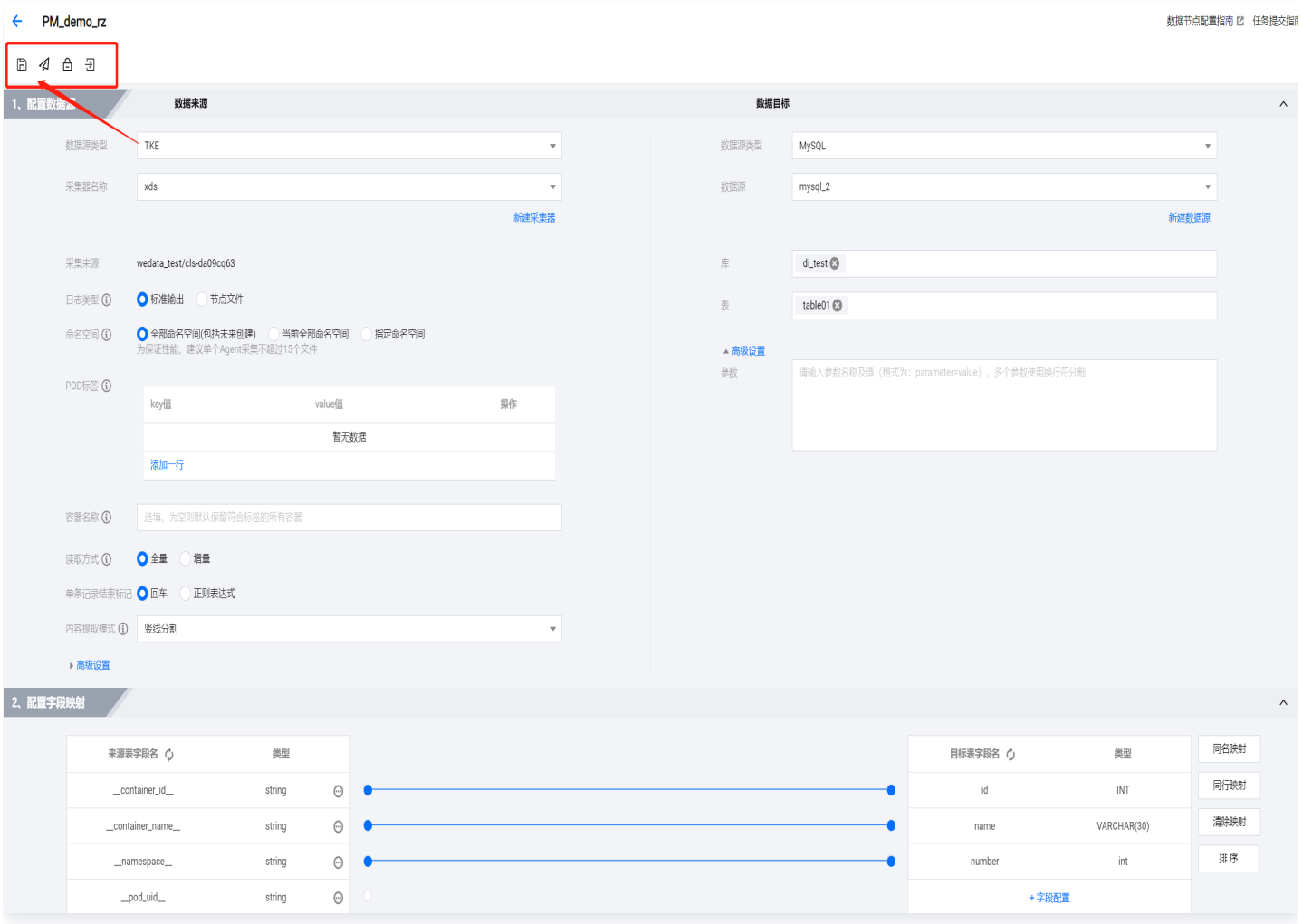
TaskManager规格 1

并发度 1

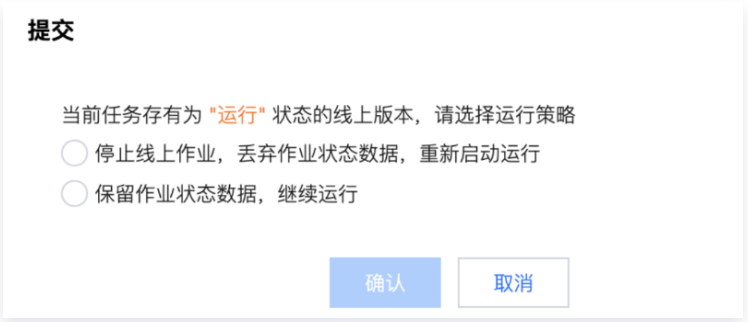
高级参数

请输入参数名称及值, 多个参数使用换行符分割
参数示意：inlong.task.group.id=xxxxxx
inlong.task.stream.id=xxxxxxxx

步骤七：任务提交



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。

		 **说明：** 单击**立即启动**，任务将在提交后立即开始运行，否则需要手动触发才会正式运行。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮，保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

后续步骤

完成任务配置后，您可以对已创建的采集任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时任务运维 。

TKE 来源采集任务配置

最近更新时间：2024-08-15 21:18:51

背景信息

Agent 是数据集成提供的一种轻量日志采集器，它可以通过产品界面化方式无代码完成安装、管理以及运维等全生命周期管理。当使用腾讯云 TKE 进行业务部署时，可通过配置 Agent 来采集 TKE 集群内各个 pod 的标准输出日志并投递到下游目标端。

操作步骤

步骤一：创建采集任务

进入**配置中心** > **实时同步任务**页面后，单击新建日志采集任务。输入任务名称并选择配置模式，支持**表单**和**画布**两种模式。

步骤二：配置数据来源

在数据源类型中选择 TKE 类型，并配置该数据源参数信息。

参数	说明
采集器名称	选择当前项目可用的采集器。
采集来源	选择已经配置的 TKE Agent 名称， TKE Agent 配置方式查看
日志类型	支持标准输出和节点文件两种类型： • 标准输出：容器的标准输出的日志，STDOUT 、STDERR。 • 节点文件：容器内部指定路径的文本日志，选择节点文件则需要手动输入日志文件路径。
命名空间	选择该采集任务的命名空间： • 全部命名空间：包含当前以及未来创建的所有命名空间。 • 当前全部命名空间：选中当前所有命名空间，不包含未来创建的命名空间。 • 指定命名空间：手动输入或者选择合适的命名空间。 **⚠ 注意：**

	为保障使用性能，建议单个 Agent 采集命名空间不超过15个文件。若需采集文件数超过15个，可创建多个采集器。
POD 标签	通过标签值筛选出指定的 pod 对象进行数据采集。一个 key 可对应多个 value，value 间请使用逗号分割。如“ environment=production,qa ”，系统将筛选出 environment = production 或者 environment = qa 的 pod 对象。
容器名称	选填，为空则默认保留符合标签的所有容器。
读取方式	支持全量和增量两种读取方式： • 全量：从日志文件内容第一行开始读取。 • 增量：从日志末尾开始读取最新内容。
单条记录结束标记	默认回车选项，若选择正则表达式，则需要手动输入正确的正则表达式。
内容提取模式	选择需要的内容提取模式，支持以下三种模式： • 全内容：每条日志记录内容被解析为键值为 __CONTENT__ 的一行完全字符串。 • JSON： 每条日志记录内容解析为 json 键值对，键值需在数据字段内定义。 • 分割：根据指定的分割符解析日志内容，键值需在数据字段内定义（支持竖线、逗号、分号分割）。
高级设置（可选）	可根据业务需求配置参数。

步骤三：配置数据目标

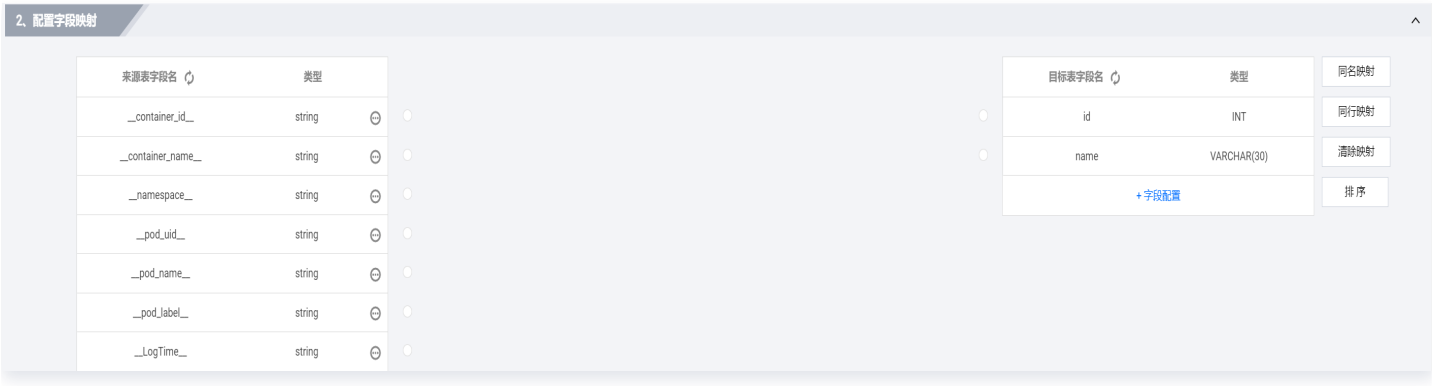
日志采集目前已支持大部分主流数据库连接。

数据来源	已支持目标数据源
TKE	MySQL 写入节点
	TDSQL-C MySQL 写入节点
	PostgreSQL 写入节点
	SQL Server 写入节点
	Oracle 写入节点
	HIVE 写入节点
	Clickhouse 写入节点
	Greenplum 写入节点

	TBase 写入节点
	DLC 写入节点
	HBase 写入节点
	Iceberg 写入节点
	HDFS 写入节点
	Doris 写入节点
	Elasticsearch 写入节点
	Kafka 写入节点

步骤四：配置字段映射

配置好数据来源和数据目标后，则会展示来源表和目标表的字段信息，我们需要对字段进行映射，支持同名映射和同行映射两种映射方式，并可以对字段进行排序和配置。



单击**字段配置**则对字段进行配置：

- 表单配置：可以对字段名称、类型进行更改，也可以删除字段和新增字段。

> **注意：**
> 文本或 json 提取内容将覆盖除内置元数据字段外的数据字段。内置元数据字段不支持修改和编辑。

- 文本解析：对已有字段进行解析，也可以手动新增字段。

> **注意：**
> 一行默认为一个字段及类型，字段名称和类型使用设定的分割符号分割，如 age int。提醒：首尾空行会被撷取，空行会被忽略。

- json 解析：数据为 JSON 格式时（如{"age":10,"name":"demo"}），系统将自动提取字段名并解析值类型。

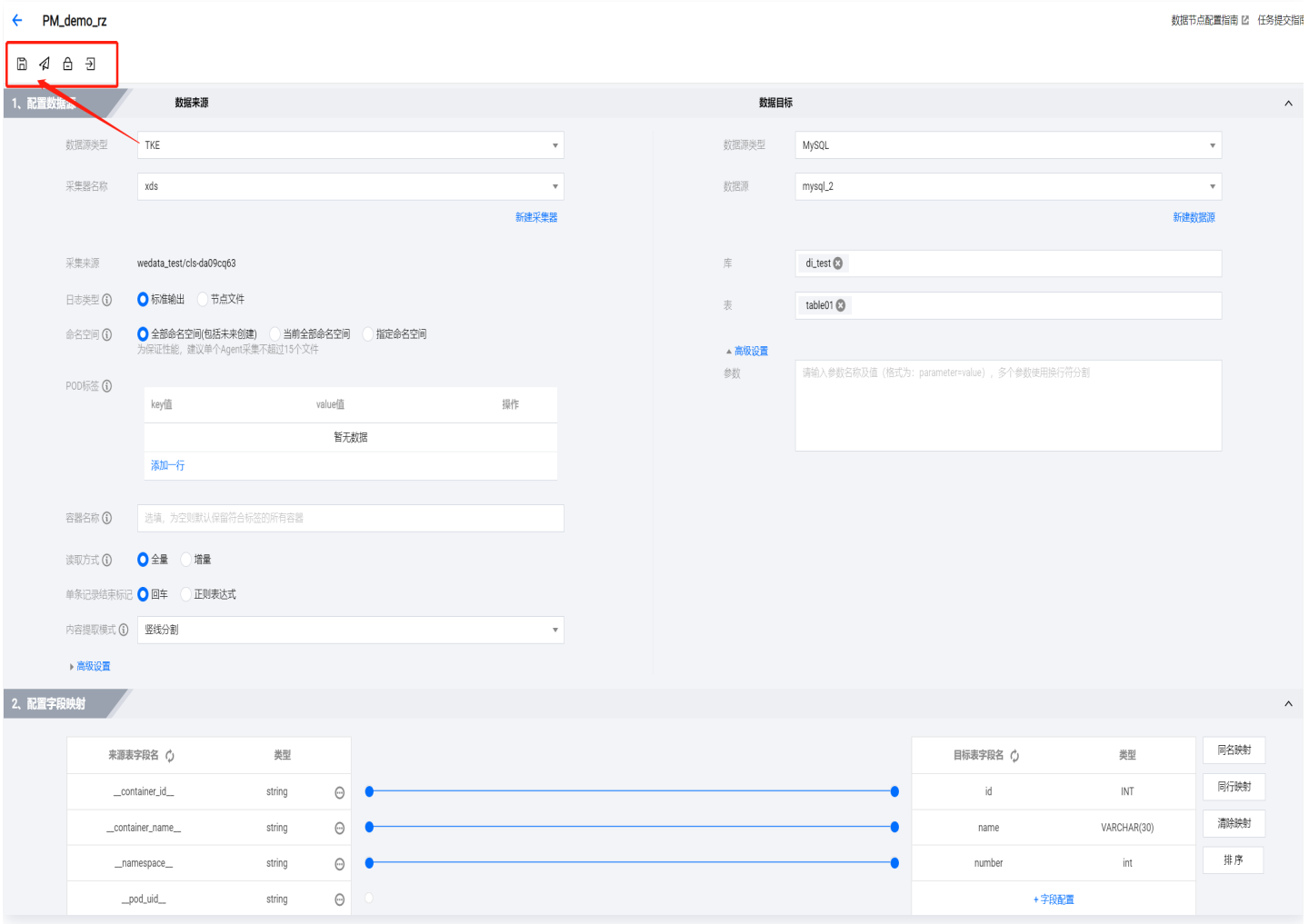
> **注意：**
> 当前系统仅支持解析部分类型，可在表单模式下确认并调整解析结果。重复字段保留最后一条。

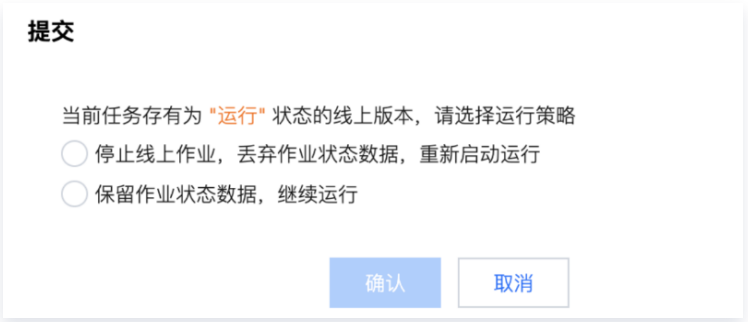
步骤五：配置任务属性

单击右侧**任务属性**进入，配置正确的基本属性和集成资源组即可。


任务属性
基本属性
任务名称
任务类型
实时同步
责任人
描述
资源配置
集成资源组
请选择
资源联通性说明
新建集成资源组
版本
--
ManagerUrl
--
JobManager规格
1
TaskManager规格
1
并发度
1
高级参数
请输入参数名称及值，多个参数使用换行符分割
参数示意：inlong.task.group.id=xxxxxx
inlong.task.stream.id=xxxxxxxx

步骤六：任务提交



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。 

		说明： 单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

后续步骤

完成任务配置后，您可以对已创建的采集任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时任务运维 。

SDK 来源采集任务配置

最近更新时间：2024-07-18 17:43:21

背景信息

数据集成提供 Java、C++ SDK，支持在系统内通过内置 SDK 方式上报业务数据。

操作步骤

步骤一：创建采集任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**日志采集任务。输入任务名称并选择配置模式，支持**表单**和**画布**两种模式。

步骤二：配置数据来源

在数据源类型中选择 SDK 类型，并配置该数据源参数信息。

参数	说明
内容提取模式	SDK 支持三种读取模式： • 全内容：每条日志记录内容被解析为键值为 ___CONTENT___ 的一行完全字符串。 • JSON：每条日志记录内容解析为 JSON 键值对，键值需在数据字段内定义。 • 符号分割：根据指定的分割符解析日志内容，键值需在数据字段内定义。

步骤三：配置数据目标

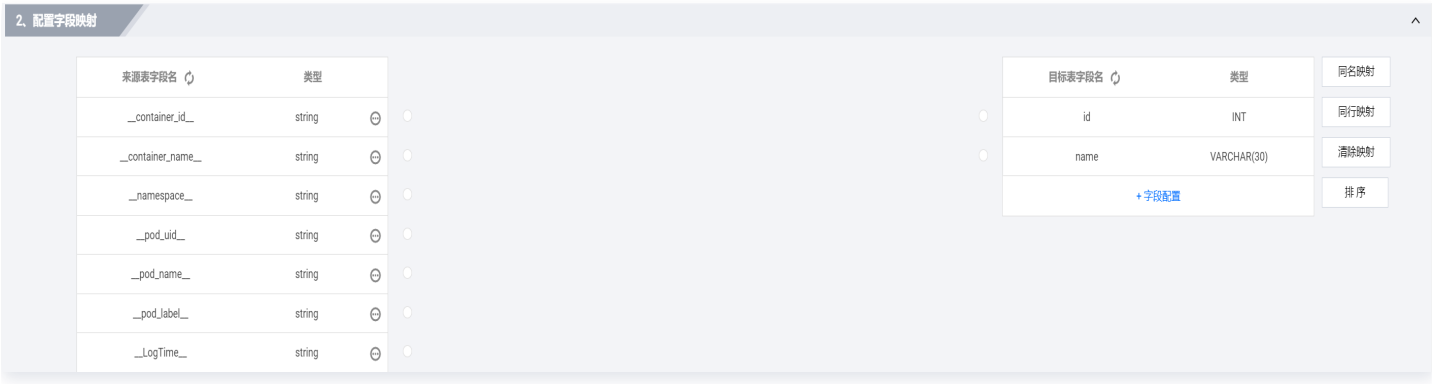
日志采集目前已支持大部分主流数据库连接。

数据来源	已支持目标数据源
TKE	MySQL 写入节点
	TDSQL-C MySQL 写入节点
	PostgreSQL 写入节点
	SQL Server 写入节点
	Oracle 写入节点
	HIVE 写入节点
	Clickhouse 写入节点
	Greenplum 写入节点

	TBase 写入节点
	DLC 写入节点
	HBase 写入节点
	Iceberg 写入节点
	HDFS 写入节点
	Doris 写入节点
	Elasticsearch 写入节点
	Kafka 写入节点

步骤四：配置字段映射

1. 配置好数据来源和数据目标后，则会展示来源表和目标表的字段信息，我们需要对字段进行映射，支持同名映射和同行映射两种映射方式，并可以对字段进行排序和配置。



2. 单击**字段配置**则对字段进行配置：
   - **表单配置**：可以对字段名称、类型进行更改，也可以删除字段和新增字段。

> **注意：**
> 文本或 JSON 提取内容将覆盖除内置元数据字段外的数据字段。内置元数据字段不支持修改和编辑。

- **文本解析**：对已有字段进行解析，也可以手动新增字段。

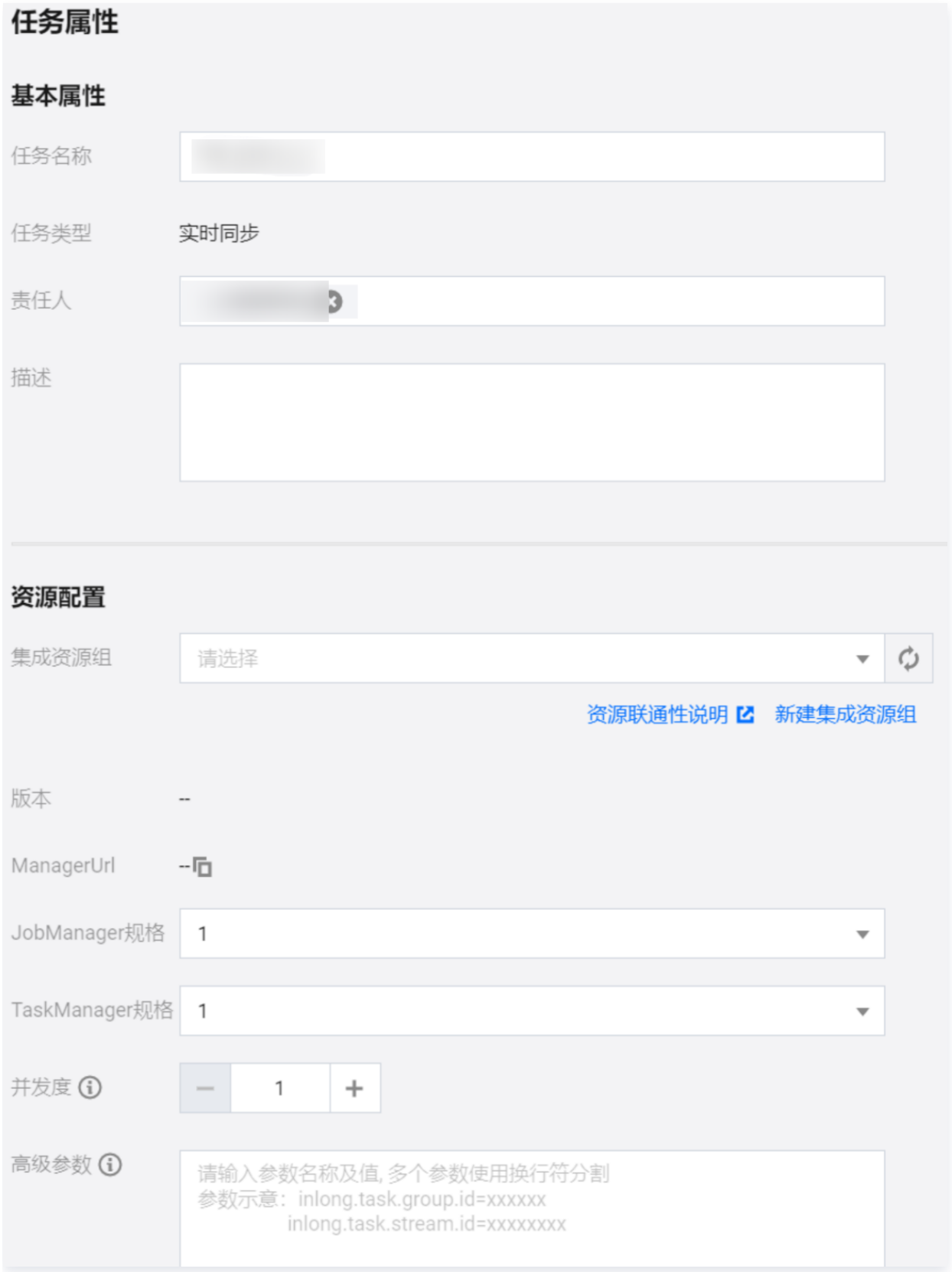
> **注意：**
> 一行默认为一个字段及类型，字段名称和类型使用设定的分割符号分割，如 age int。提醒：首尾空行会被撷取，空行会被忽略。

- **JSON 解析**：数据为 JSON 格式时（如{"age":10,"name":"demo"}），系统将自动提取字段名并解析值类型。

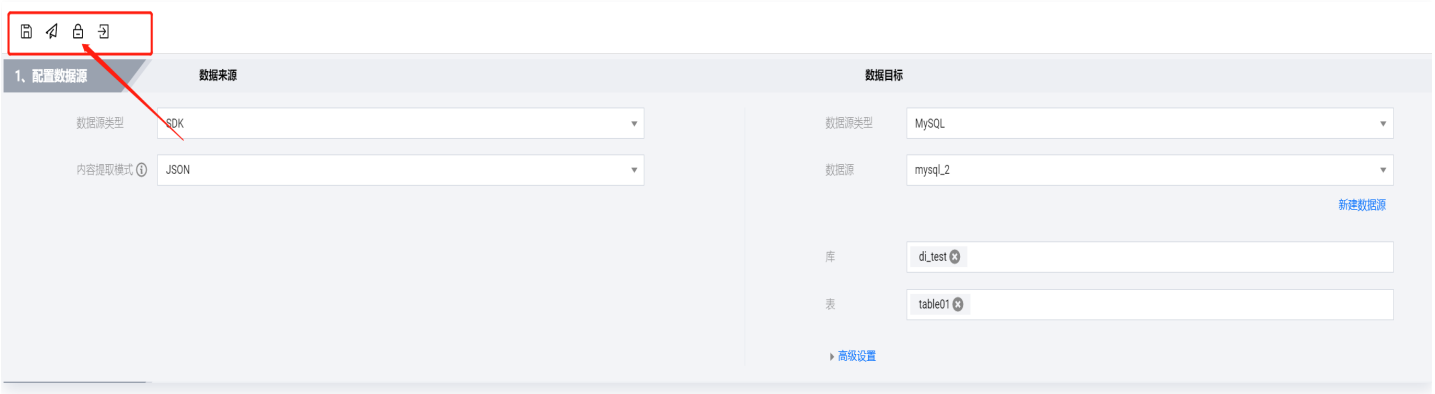
> **注意：**
> 当前系统仅支持解析部分类型，可在表单模式下确认并调整解析结果。重复字段保留最后一条。

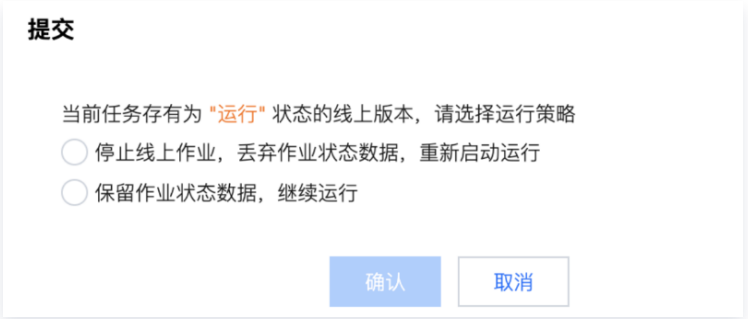
步骤五：配置任务属性

单击右侧**任务属性**进入，配置正确的基本属性和集成资源组即可。


任务属性
基本属性
任务名称
任务类型
实时同步
责任人
描述
资源配置
集成资源组
请选择
资源联通性说明
新建集成资源组
版本
--
ManagerUrl
--
JobManager规格
1
TaskManager规格
1
并发度
1
高级参数
请输入参数名称及值, 多个参数使用换行符分割
参数示意：inlong.task.group.id=xxxxxx
inlong.task.stream.id=xxxxxxxx

步骤六：任务提交



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略。 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接**提交**。 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。  **注意：** 单击**立即启动**，任务将在提交后立即开始运行，否则需要手动触发才会正式运行。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作。
3	前往运维	根据当前任务名称快捷跳转至任务运维页面。
4	保存	预览完成后，可单击**保存**按钮，保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**。

后续步骤

完成任务配置后，您可以对已创建的采集任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时任务运维 。

CVM 来源采集任务配置

最近更新时间：2024-04-02 16:17:11

背景信息

Agent 是数据集成提供的一种轻量日志采集器，它可以通过产品界面化方式无代码完成安装、管理以及运维等全生命周期管理。当使用腾讯云 CVM 或者自建服务器 进行业务部署时，可通过配置 Agent 来采集服务器内日志及文件信息投递到下游目标端。

操作步骤

步骤一：创建采集任务

进入**配置中心 > 实时同步任务**页面后，单击**新建**日志采集任务。输入任务名称并选择配置模式，支持**表单**和**画布**两种模式。

步骤二：配置数据来源

在数据源类型中选择 CVM 类型，并配置该数据源参数信息。

参数	说明
采集器组	选择当前项目可用的采集器，若没有可以单击**新建采集器**创建。
服务器分类	选择服务器所归属分类，选择后任务将采集该分类下所有服务器。
文件路径	手动输入数据来源的文件路径。
黑名单	默认关闭，开启后配置的黑名单文件路径默认不采集。
读取方式	CVM 来源支持两种读取方式： • 全量：从日志文件内容第一行开始读取。 • 增量：从日志末尾开始读取最新内容。

单条记录结束标记	默认回车选项，若选择正则表达式，则需要手动输入正确的正则表达式。
内容提取模式	支持三种内容提取模式： • 全内容：每条日志记录内容被解析为键值为 CONTENT 的一行完全字符串。 • JSON：每条日志记录内容解析为json键值对，键值需在数据字段内定义。 • 分割：根据指定的分割符解析日志内容，键值需在数据字段内定义（支持竖线、逗号、分号分割）。

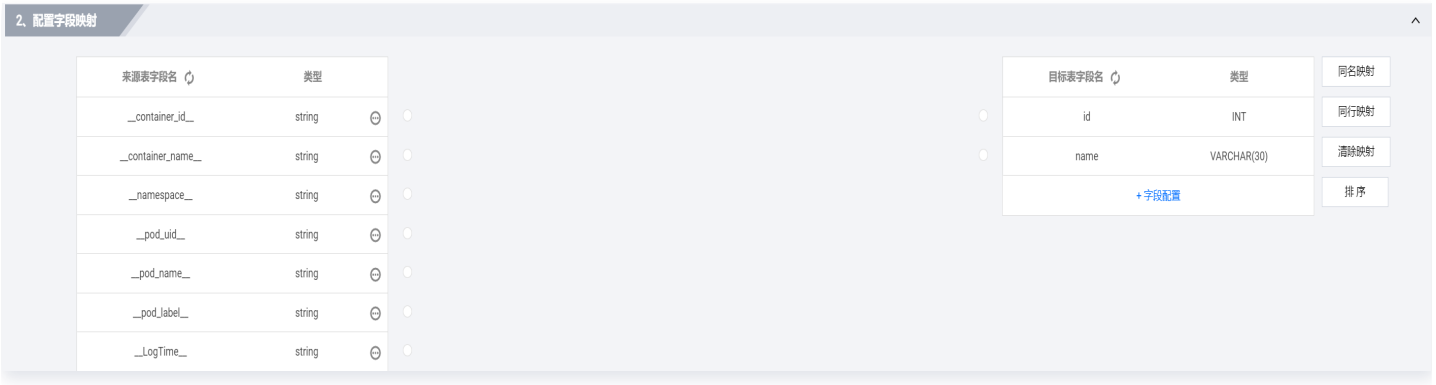
步骤三：配置数据目标

日志采集目前已支持大部分主流数据库连接

数据来源	已支持目标数据源
TKE	MySQL 写入节点
	TDSQL-C MySQL 写入节点
	PostgreSQL 写入节点
	SQL Server 写入节点
	Oracle 写入节点
	HIVE 写入节点
	Clickhouse 写入节点
	Greenplum 写入节点
	TBase 写入节点
	DLC 写入节点
	HBase 写入节点
	Iceberg 写入节点
	HDFS 写入节点
	Doris 写入节点
	Elasticsearch 写入节点
	Kafka 写入节点

步骤四：配置字段映射

配置好数据来源和数据目标后，则会展示来源表和目标表的字段信息，我们需要对字段进行映射，支持同名映射和同行映射两种映射方式，并可以对字段进行排序和配置。



单击**字段配置**则对字段进行配置：

- 表单配置：可以对字段名称、类型进行更改，也可以删除字段和新增字段。

> **注意：**
> 文本或 json 提取内容将覆盖除内置元数据字段外的数据字段。内置元数据字段不支持修改和编辑。

- 文本解析：对已有字段进行解析，也可以手动新增字段。

注意：
一行默认为一个字段及类型，字段名称和类型使用设定的分割符号分割，例如 age int。提醒：首尾空行会被撷取，空行会被忽略。

- json 解析：数据为 JSON 格式时（如{"age":10,"name":"demo"}），系统将自动提取字段名并解析值类型。

> **注意：**
> 当前系统仅支持解析部分类型，可在表单模式下确认并调整解析结果。重复字段保留最后一条。

步骤五：配置任务属性

单击右侧**任务属性**进入，配置正确的基本属性和集成资源组即可。

任务属性

基本属性

任务名称

任务类型 实时同步

责任人

描述

资源配置

集成资源组 请选择

资源联通性说明 新建集成资源组

版本 --

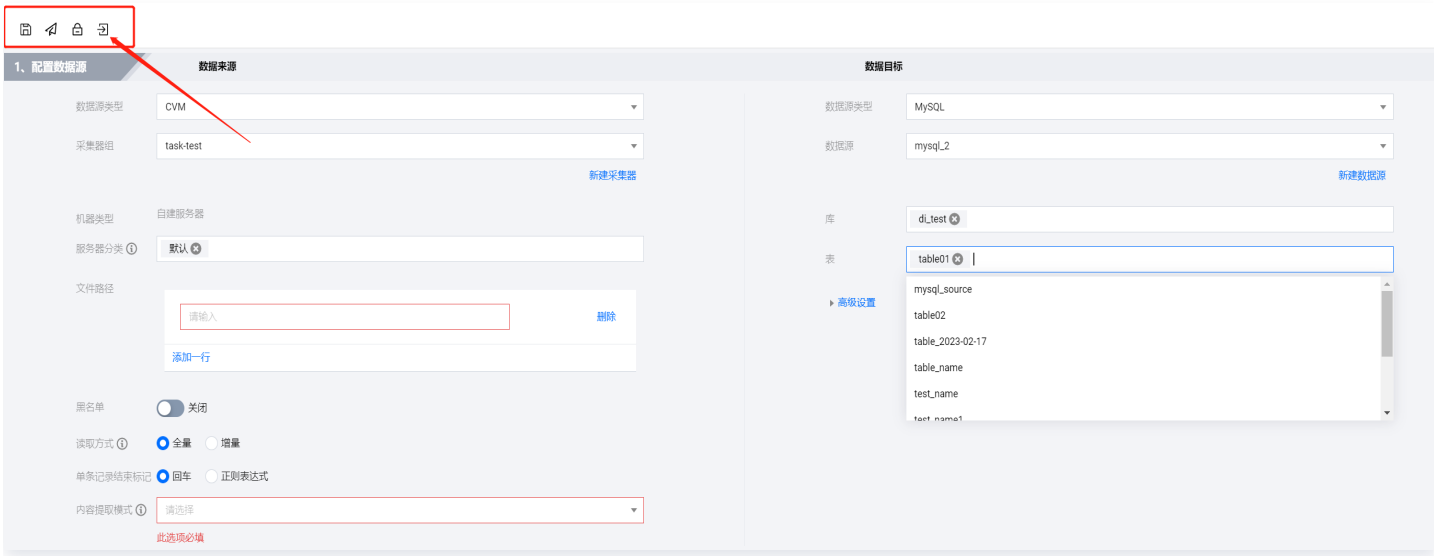
ManagerUrl --

JobManager规格 1

TaskManager规格 1

并发度 1

步骤六：任务提交



序号	参数	说明
1	提交	将当前任务提交至生产环境，提交时根据当前任务是否有生产态任务可选择不同运行策略 • 若当前任务无生效的线上任务，即首次提交或线上任务处于“失败”状态，可直接提交 • 若当前任务存在“运行中”或“暂停”状态的线上任务需选择不同策略。停止线上作业将抛弃之前任务运行位点，从头开始消费数据，保留作业状态将在重启后从之前最后消费位点继续运行。  **说明：**

		单击立即启动任务将在提交后立即开始运行，否则需要手动触发才会正式运行。
2	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可点击图标抢锁，抢锁成功可进行编辑操作
3	前往运维	根据当前任务名称快捷跳转至任务运维页面
4	保存	预览完成后，可单击**保存**按钮保存整库任务配置。仅保存的情况下，任务将不会提交至**运维中心**

后续步骤

完成任务配置后，您可以对已创建的采集任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 实时任务运维 。

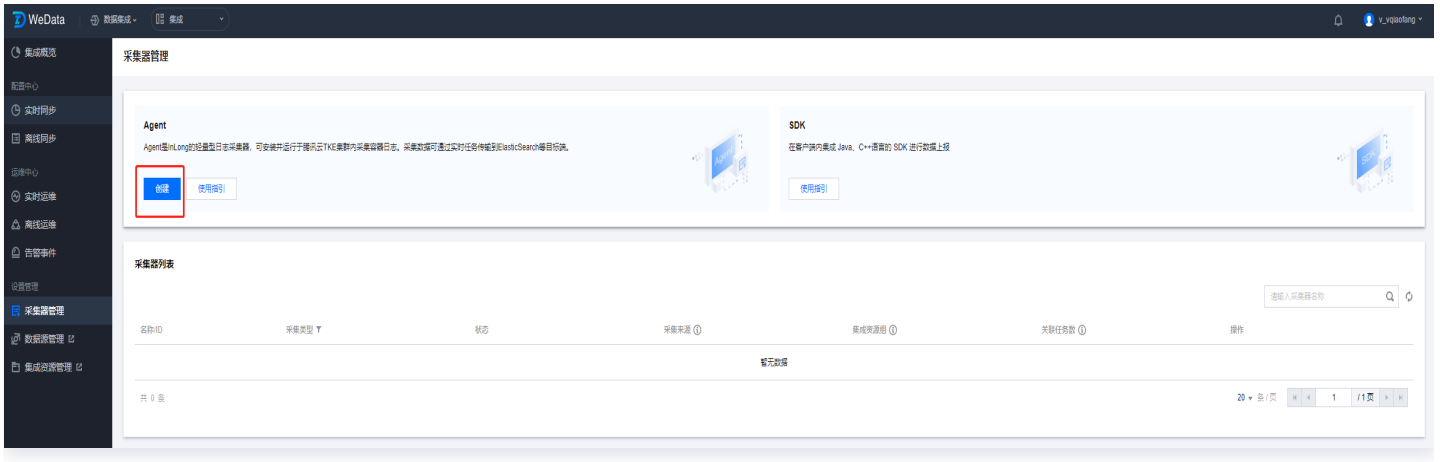
采集器管理

最近更新时间：2024-08-15 21:18:51

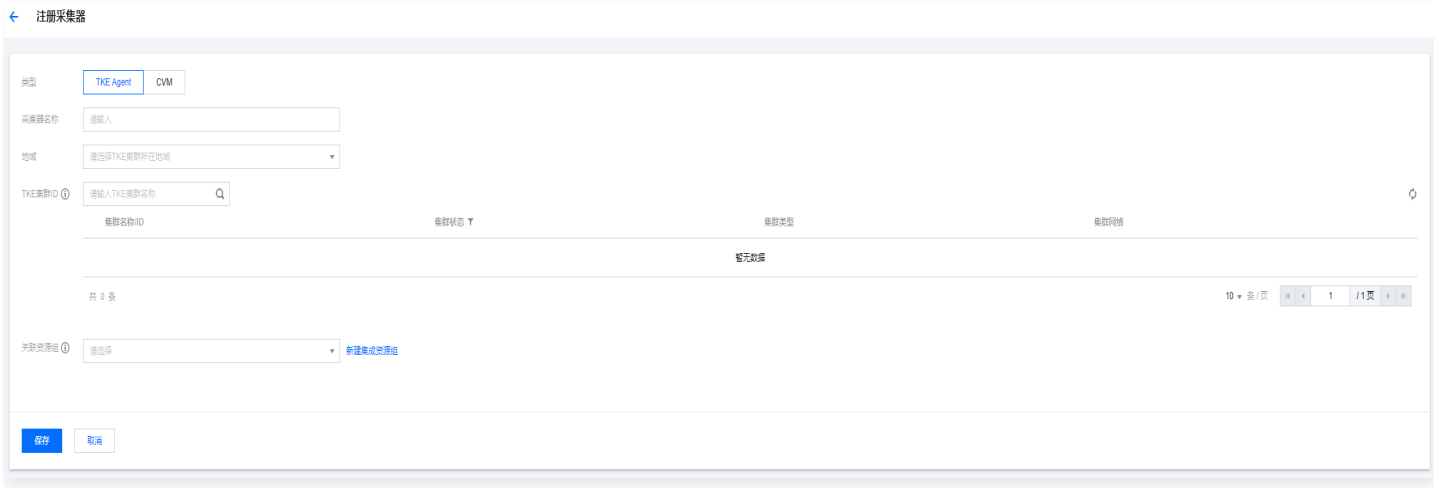
进入 **wedata 数据集成 > 设置管理 > 采集器管理**界面，页面展示 Agent 和 SDK 两种途径。
Agent 是 InLong 的轻量型日志采集器，可安装并运行于腾讯云 TKE 集群内采集容器日志。采集数据可通过实时任务传输到 ElasticSearch 等目标端。

创建 Agent 采集器

1. 进入数据集成模块，单击**采集器管理 > Agent > 创建**。



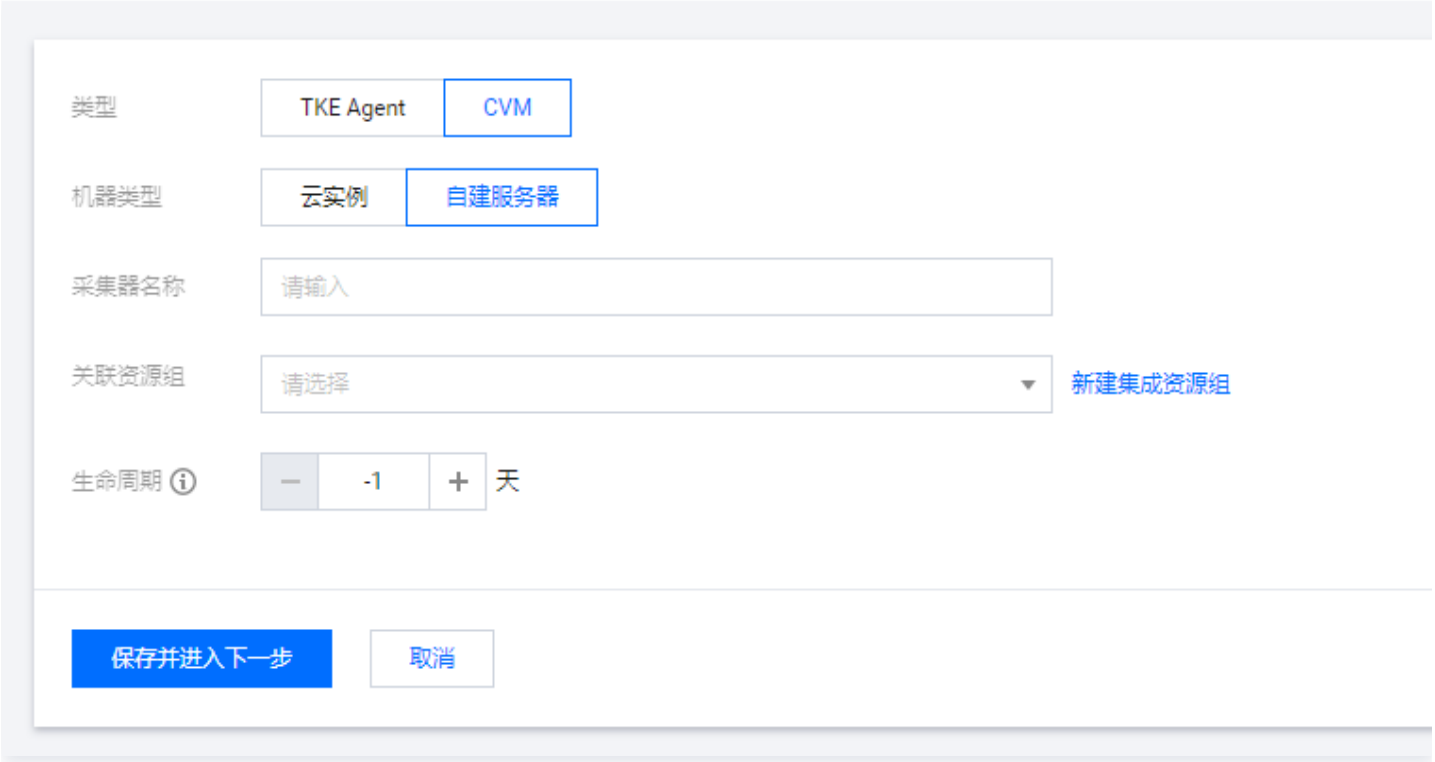
2. 在创建 Agent 采集器的弹窗中，配置 TKE Agent /CVM 的参数信息。
   - **TKE Agent**



参数	参数说明

类型	请选择 TKE Agent。
地域	选择需要采集的 TKE 集群所属地域，可登录前往登录 腾讯云 查看集群信息。
TKE 集群 ID	选择一个需安装 TKE 集群信息。 提示： 1. 仅支持在“运行中”状态的 TKE 集群安装 Agent。 2. 一个 agent 将占用集群1C512M规格。
关联资源组	将 Agent 与具体执行资源组进行绑定，Agent 将使用资源组中 manager url 进行数据上报。 **注意：** TKE 集群（Agent）需与集成资源组位于同一个 VPC，或对应 VPC 已配置公网的情况可同步数据。资源组公网配置流程请参见 数据集成 资源组配置公网-操作指南-文档中心-腾讯云 。

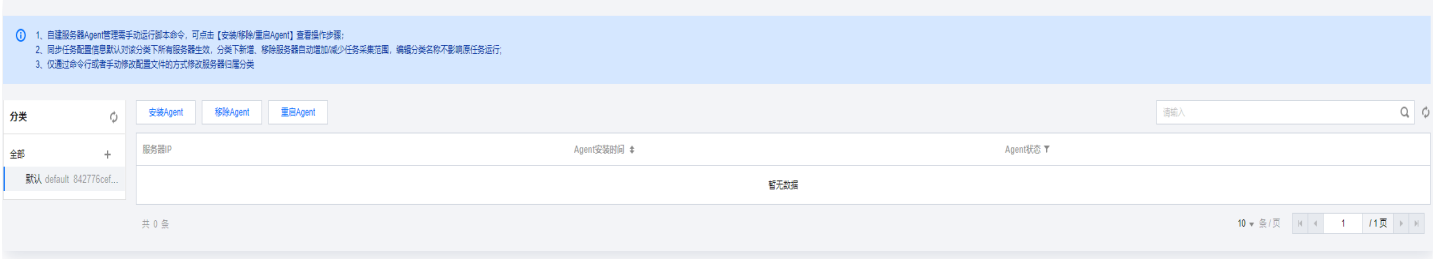
- **CVM**



参数	参数说明
机器类型	默认自建服务器。

采集器名称	自定义采集器名称。
关联资源组	选择当前项目中可以用的资源组。
生命周期	设置当前采集器组下的 Agent 心跳超时的最长时间。超过此时间后，该 Agent 对应服务器 IP 将被自动回收。默认为-1不回收，可以在采集器列表中进行编辑修改。

单击保存并进入下一步，即可进入到服务器管理。



> **注意：**
>
> 1. 自建服务器 Agent 管理需手动运行脚本命令，可单击 **安装/移除/重启 Agent** 查看操作步骤。
> 2. 同步任务配置信息默认对该分类下所有服务器生效，分类下新增、移除服务器自动增加/减少任务采集范围，编辑分类名称不影响原任务运行。
> 3. 仅通过命令行或者手动修改配置文件的方式修改服务器归属分类。

配置完成后，Agent 将作为 TKE 上的日志采集器，后续支持在多个实时任务中同时使用一个 Agent 用于 POD 日志提取。

采集器列表

创建好的采集器在采集器列表中可以进行管理查看，列表展示内容包括：名称 ID/状态/采集来源/集成资源组/关联任务数及相关操作。

实时同步运维
实时任务运维

最近更新时间：2024-08-15 21:18:52

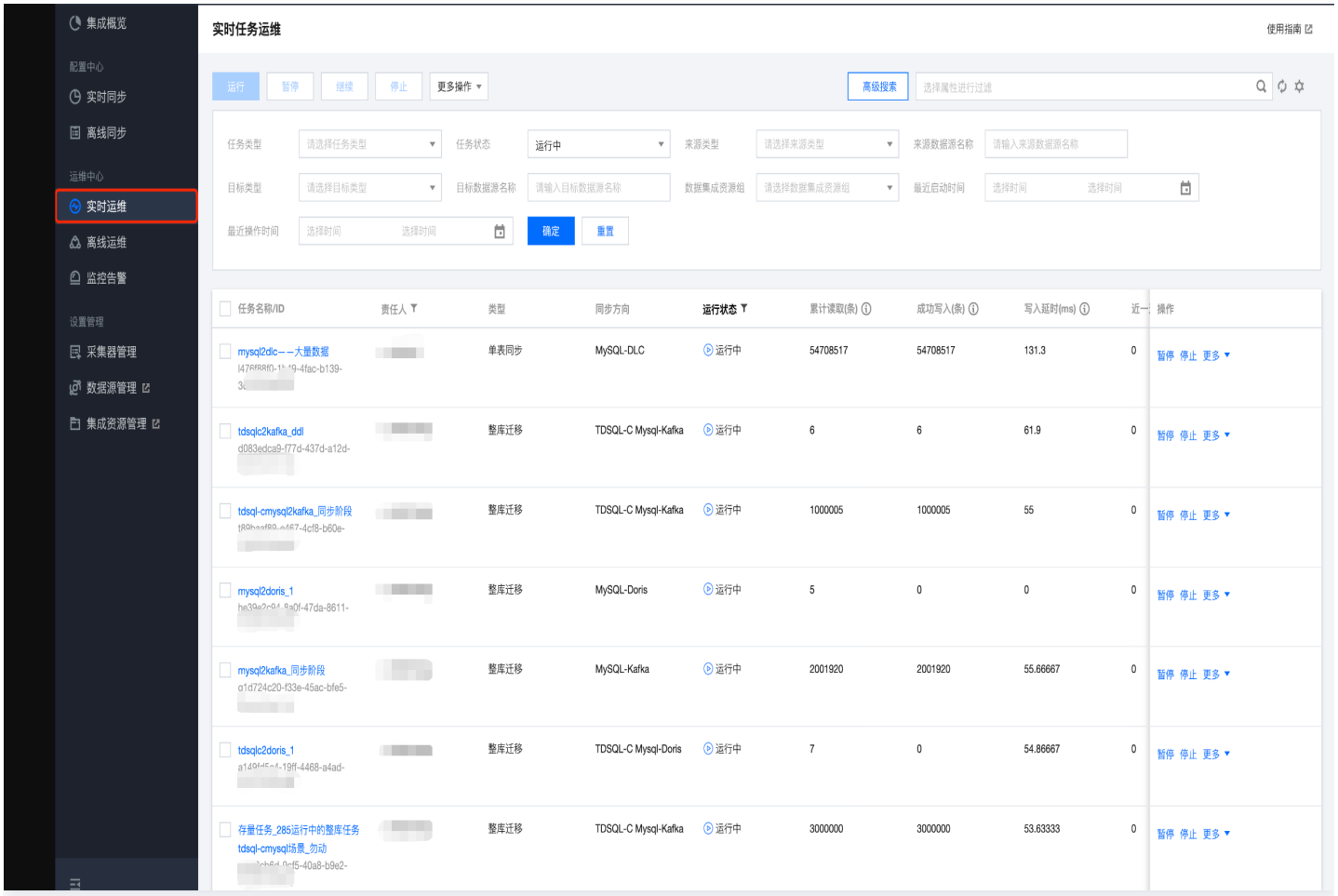
对于在完成配置并提交的实时任务，您可以进入**实时运维**内查看并操作对应的任务。您可以在实时运维内查看任务运行的基本运行状态、任务运行指标统计、运行日志、以及配置任务异常告警等。本文列举实时同步任务的常见运维操作。

前提条件

运维页面仅进行支持并展示已完成提交的实时任务，详情请参见 整库任务配置概览 、单表任务配置概览 、日志采集配置概览 。

运维列表

运维列表支持对所有已提交任务进行管理和启停等操作。



实时任务状态及含义：

状态类型	说明
初始化	任务首次提交到运行后暂未启动运行。
操作中	状态扭转中。此状态常出现在刚完成任务操作任务后，任务正在进行状态扭转。
运行中	任务正在运行中。
已暂停	当前任务已暂停运行，任务状态及读端位点被保留。
已停止	任务已通过手动方式停止运行。 **说明：**用户手动操作并停止任务，任务将扭转至此状态。其他非异常情况停止

实时任务支持操作及含义：

操作类型	说明
运行	运行当前任务启动读写
停止	手动停止当前任务。停止后，将不保留当前任务运行状态及位点
暂停	暂停运行当前任务。暂停后，保留任务状态及位点，后续支持从已完成位点继续读取。
继续	从上次暂停位点继续运行。 **说明：** 建议任务勿暂停过长时间，否则再次运行时可能由于源端日志过期、或位点丢失等导致重新启动失败。

实时任务状态及对应允许操作：

状态类型	操作			
	运行	暂停	继续	停止
初始化	✓	–	–	–
操作中	–	–	–	–
运行中	–	✓	–	✓

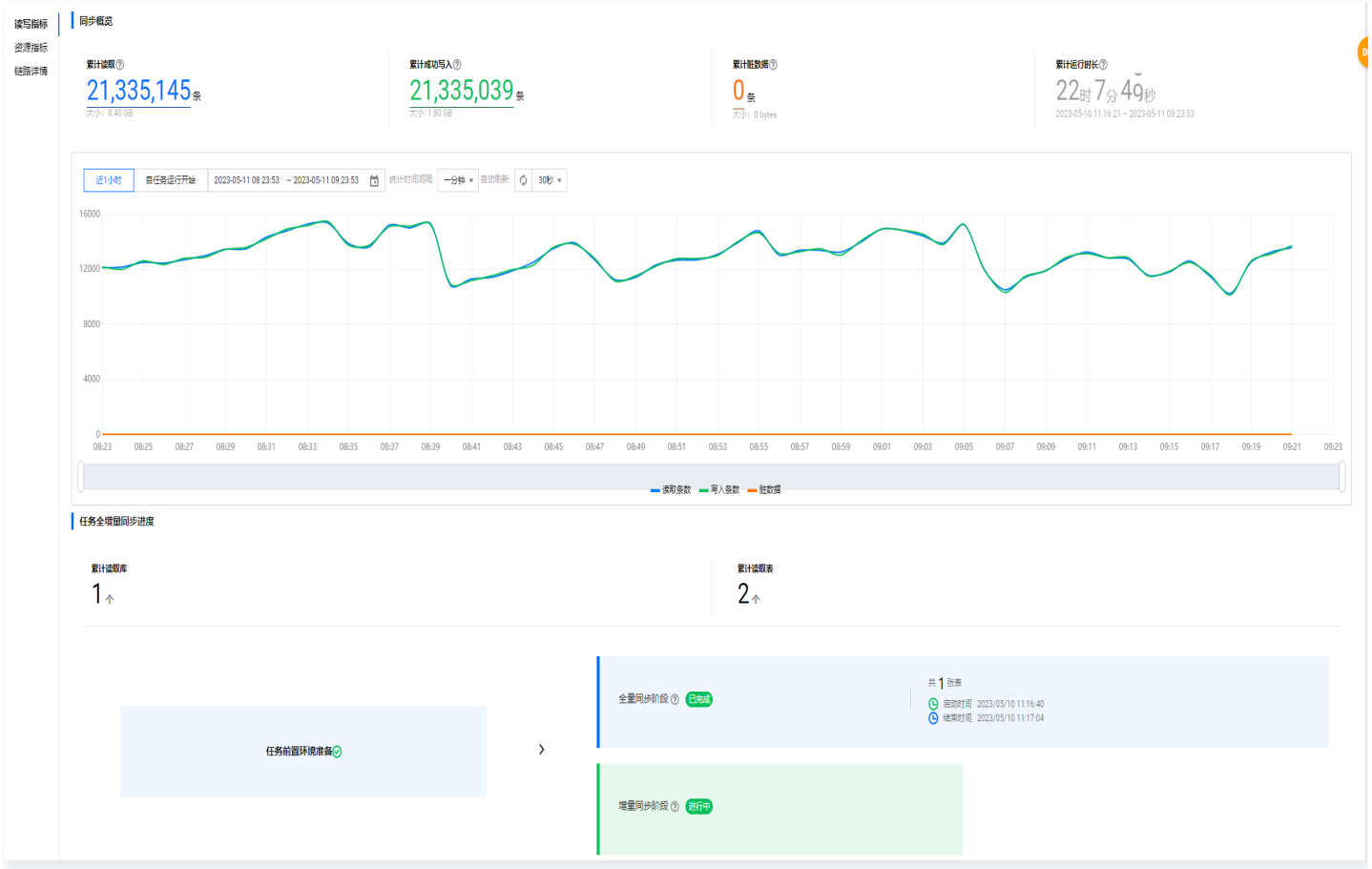
已暂停	–	–	✓	✓
已停止	✓	–	–	–

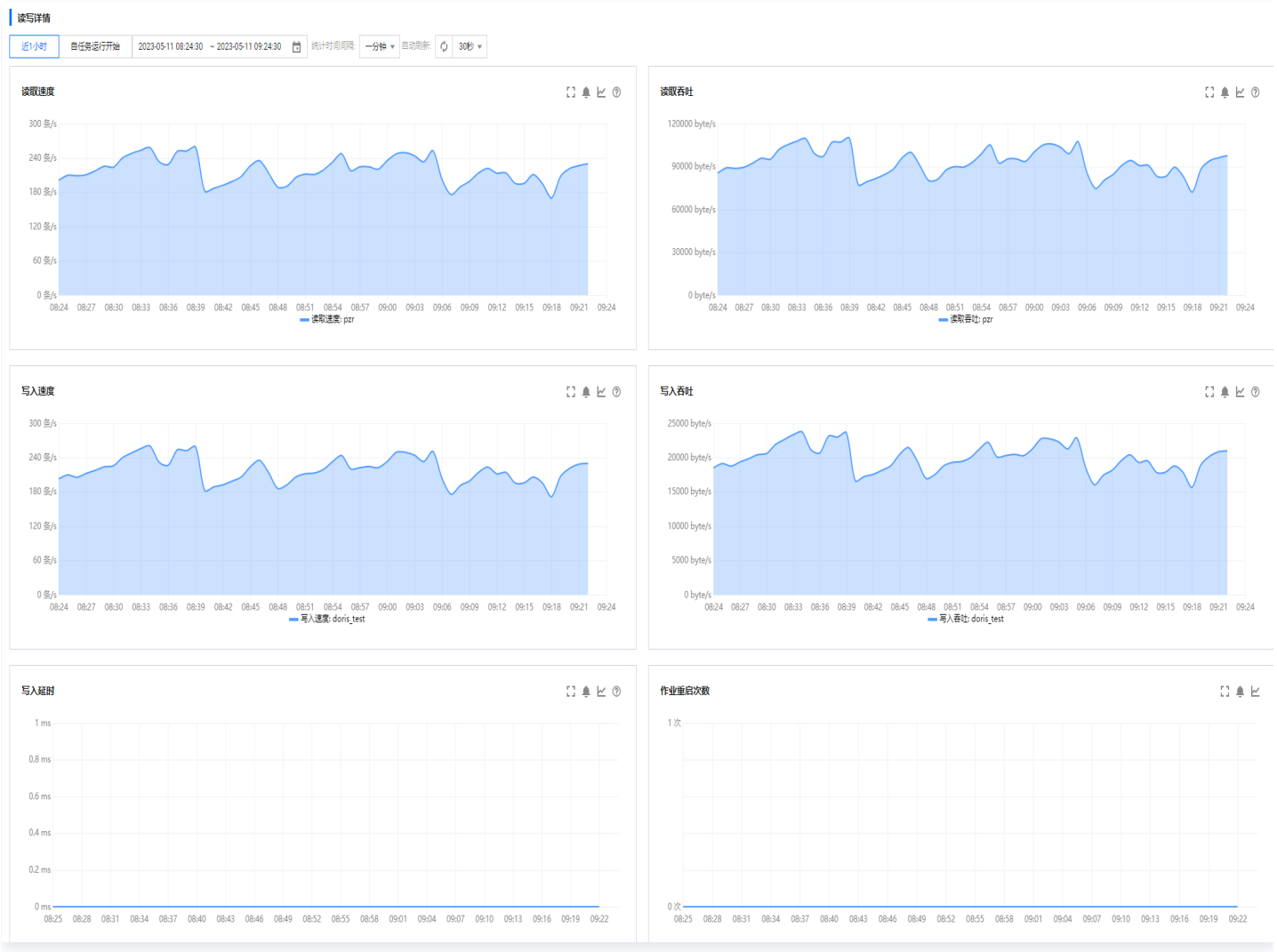
实时同步指标统计

最近更新时间：2024-08-15 21:18:52

您可以通过单击实时运维 > 操作 > 更多 > 指标统计或实时运维 > 任务名称 > 指标统计，进入实时同步指标统计整体界面。指标统计分别展示三个不同模块的内容。

一、读写指标





概览模块参数说明

指标参数	说明
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）

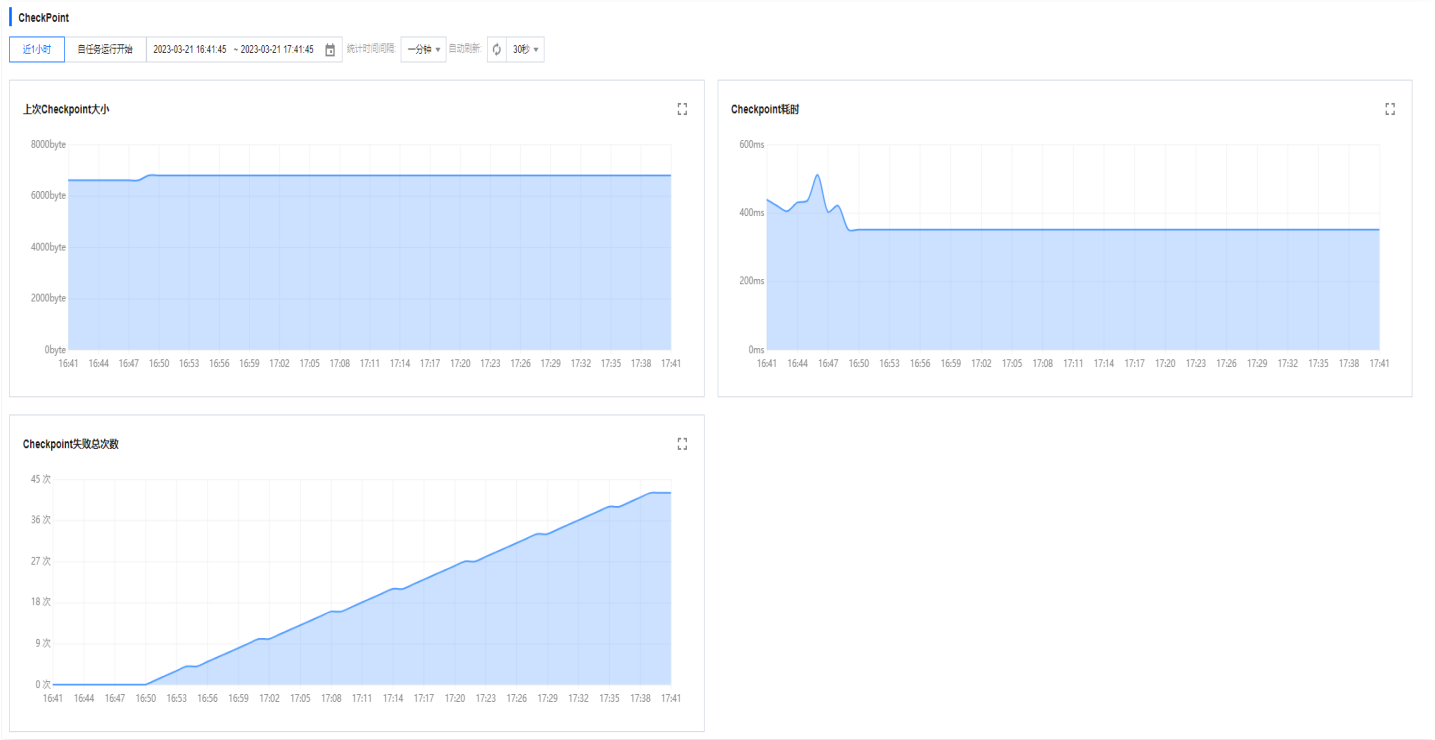
任务全增量同步进度

指标参数	说明
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

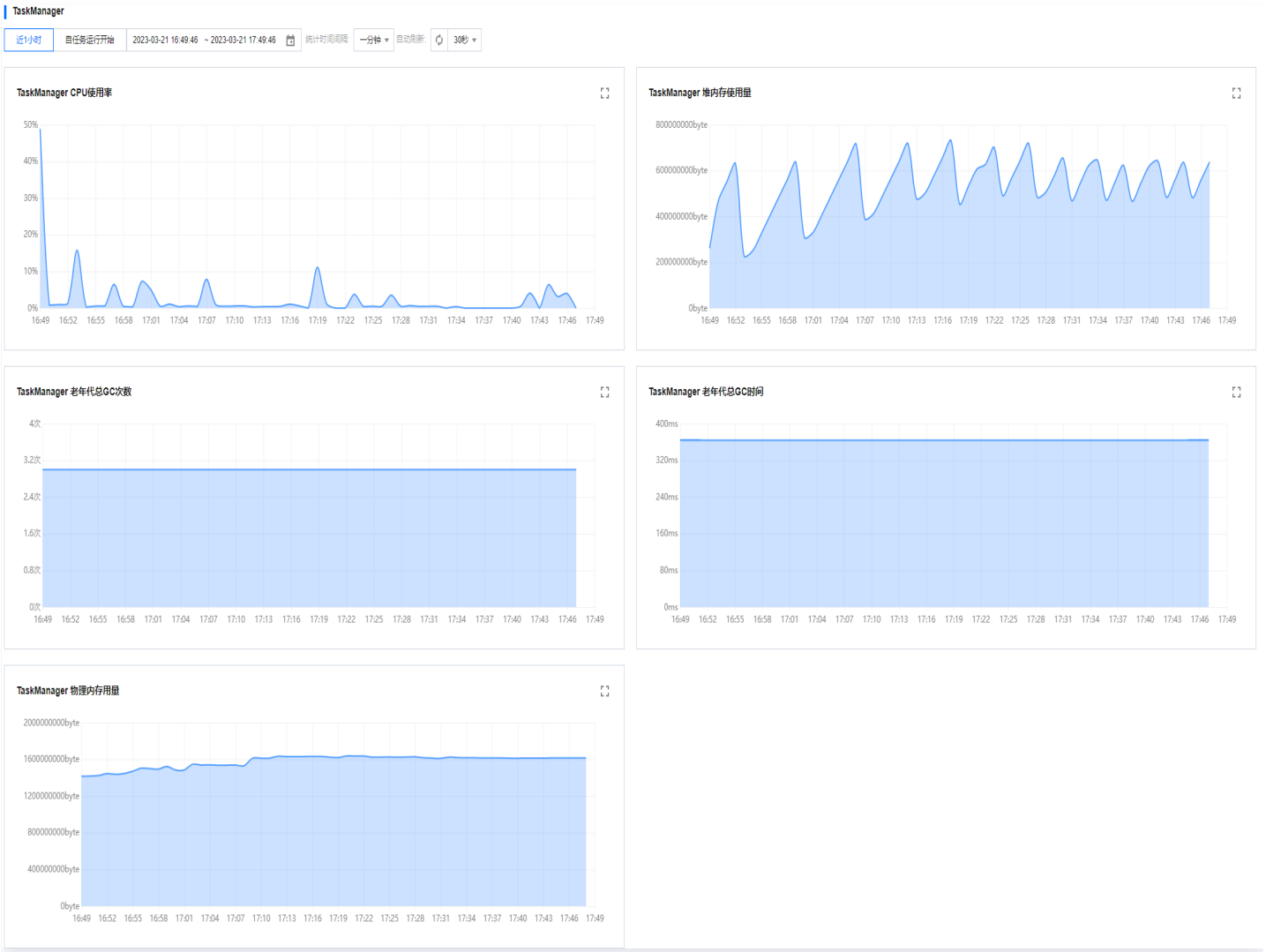
读写详情

指标参数	说明
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
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

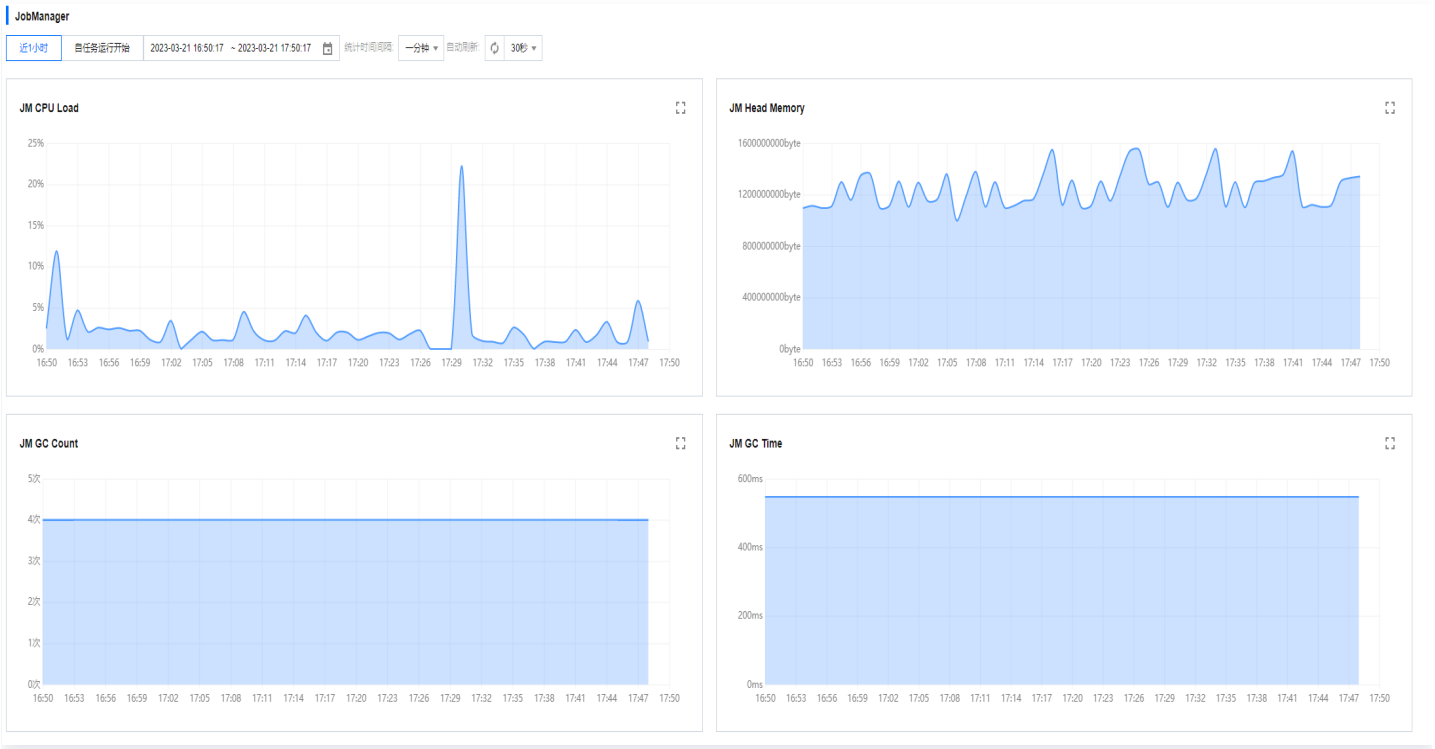
二、资源指标



指标参数	说明	示例值
上次 Checkpoint 大小	当前作业最近一次的 Checkpoint 大小	751321.00 Bytes
Checkpoint 耗时	当前作业的 Checkpoint 耗时	723.00 ms
Checkpoint 失败总次数	当前作业的 Checkpoint 的失败总次数	8次

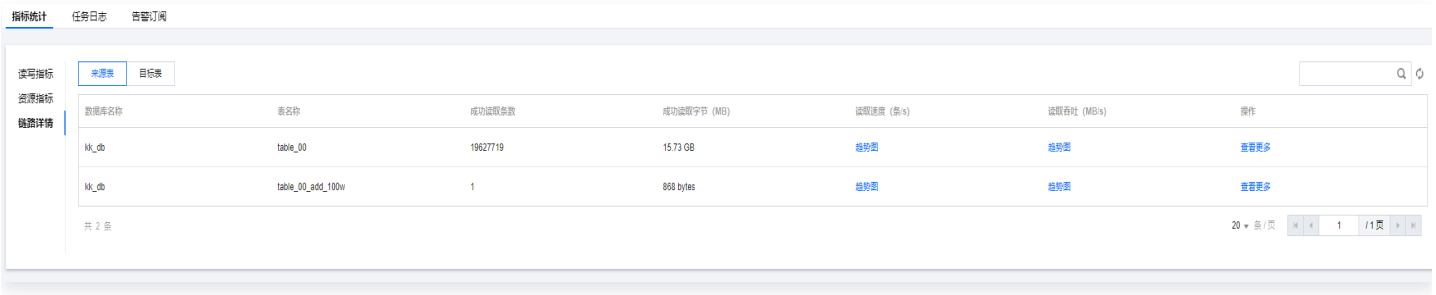


指标参数	说明	示例值
TaskManager CPU 使用率	当前作业 TaskManager 的 CPU 使用率	7.12%
TaskManager 堆内存使用量	当前作业 TaskManager 堆内存的用量	1040001560.00Bytes
TaskManager 老年代总 GC 次数	当前作业 TaskManager 老年代 GC 次数	3次
TaskManager 老年代总 GC 时间	当前作业 TaskManager 老年代 GC 时间	701.00ms
TaskManager 物理内存用量	当前作业 TaskManager 所在的 JVM 的物理内存用量（RSS），包括堆内、堆外、Native 等所有区域的总内存用量	3597035110.00Bytes



指标参数	说明	示例值
JM CPU Load	TaskManager 维度的 JVM 最近 CPU 利用率	12%
JM Head Memory	TaskManager 维度的堆内存使用情况	1次
JM GC Count	TaskManager 维度的 Status.JVM.GarbageCollector.<GarbageCollector>.Count，GC（垃圾回收）次数	5次
JM GC Time	TaskManager 维度的 Status.JVM.GarbageCollector.<GarbageCollector>.Time，GC（垃圾回收）时间	65ms

三、链路详情



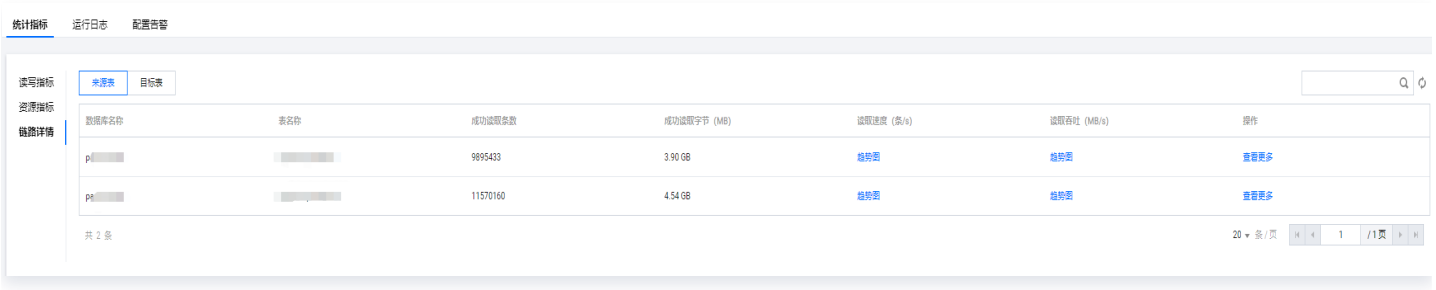
指标参数	说明
来源表	展示整库同步任务中成功读取表明细统计，包括库名、表名、读取条数/字节、读取速度/吞吐，趋势图支持对任务运行期间单表的详细指标进行查看，支持根据表名、库名搜索
目标表	展示整库任务写入的表统计，包括库名、表名、写入条数/字节、写入速度/吞吐、脏数据，趋势图支持对任务运行期间单表的详细指标进行查看，支持根据表名、库名搜索
趋势图	来源和目标表均支持单表趋势图，默认展开当前近一小时内各个指标的趋势图，弹窗内数据1分钟自动刷新，支持手动刷新

链路详情（整库）

最近更新时间：2024-08-06 18:01:21

整库同步运维支持查看同步链路详情，单击**实时运维 > 运行详情 > 指标统计 > 链路详情**，即可进入链路详情运维页面。

读取端

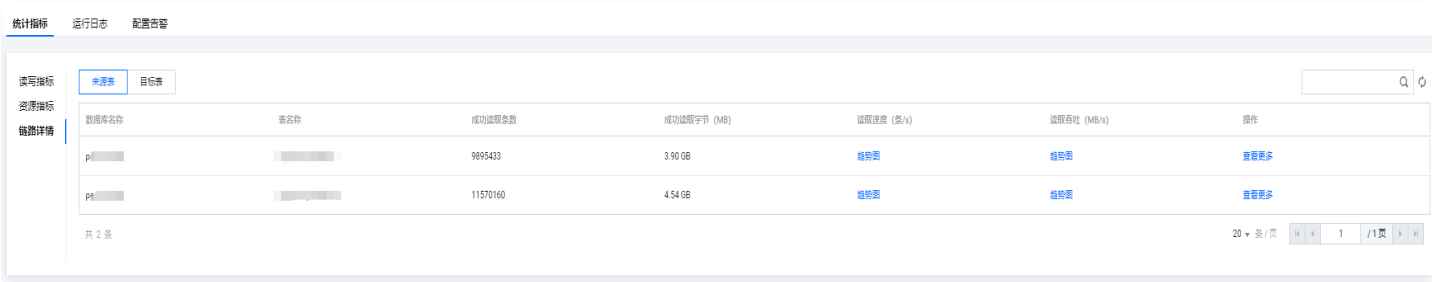


来源表指标：展示整库同步任务中成功读取表明细统计：

- 包括库名、表名、读取条数/字节、读取速度/吞吐。
- 趋势图支持对任务运行期间单表的详细指标进行查看。
- 支持根据表名、库名搜索。

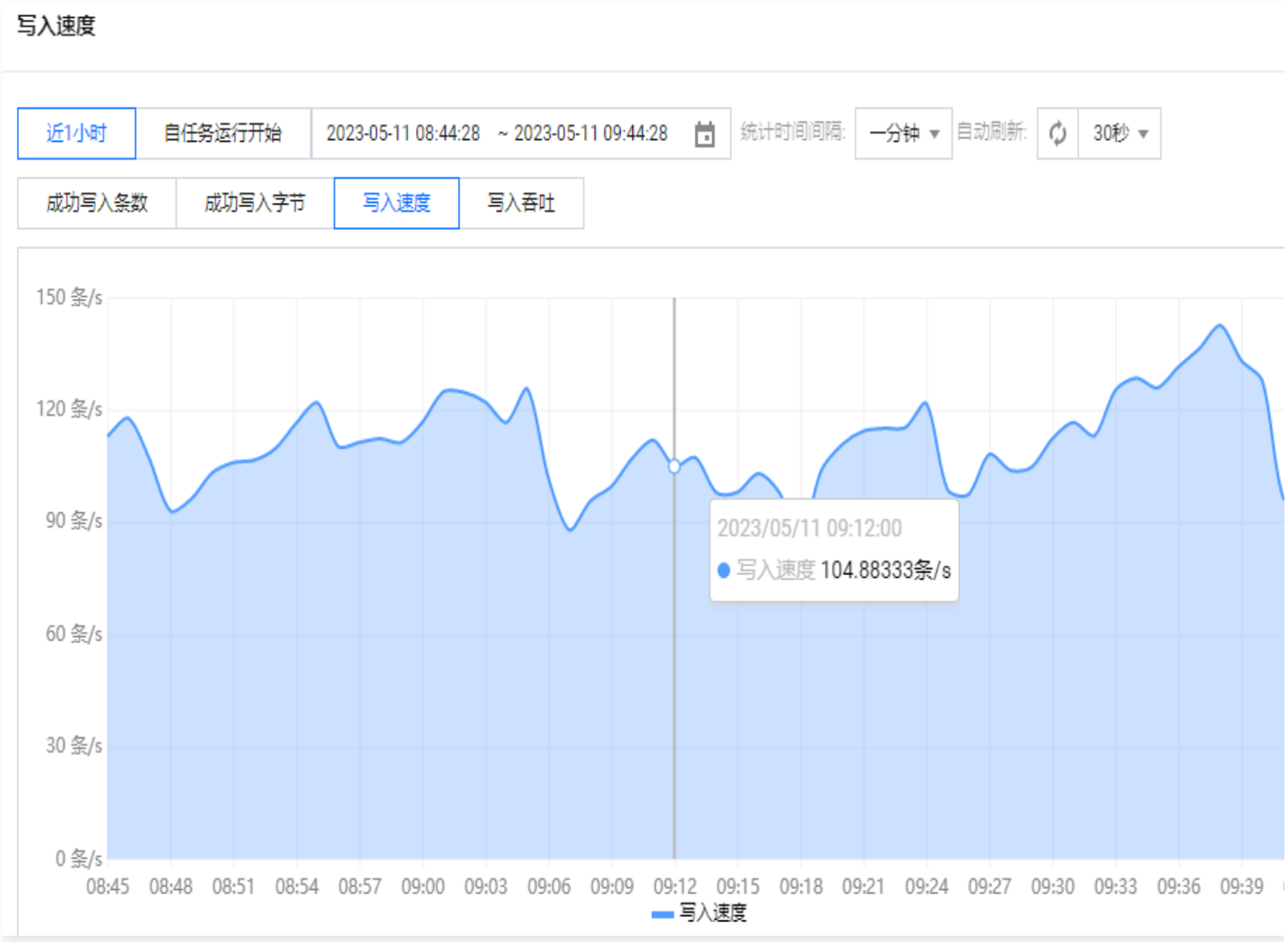
- 趋势图
  - 默认展开当前近1小时内各个指标的趋势图，支持手动调节展示时间段。
  - 弹窗内数据1分钟自动刷新，支持手动刷新。

写入端



目标表指标：展示整库任务写入的表统计。

- 包括库名、表名、写入条数/字节、写入速度/吞吐、脏数据。
- 趋势图支持对任务运行期间单表的详细指标进行查看。
- 支持根据表名、库名搜索。



- 趋势图
  - 默认展开当前近1小时内各个指标的趋势图，支持手动调节展示时间段。
  - 弹窗内数据1分钟自动刷新，支持手动刷新。

配置告警

最近更新时间：2024-04-02 16:17:11

配置告警为每个同步任务提供基于不同指标及告警阈值的创建任务告警规则，一个任务支持创建多个不同告警级别、不同告警规则的告警监控。支持通过规则状态、告警级别及告警指标进行筛选，并且支持通过规则名称搜索告警事件。
实时运维界面单击任务名 > 配置告警界面进入。

参数	说明
规则ID	系统默认生成的当前告警规则 ID
规则名称	用户指定配置的告警规则名称
规则状态	• 开启：开启当前告警规则，任务运行时将根据告警规则及阈值触发告警 • 关闭：关闭告警验证
告警指标	告警指标是对任务运行的监控，展示任务失败、异常、暂停、重启等状态出现的原因，支持通过告警指标筛选运维任务 以下情况会触发告警： • 任务暂停：任务暂停超过累计时长后告警 • 任务停止：任务手动或被动停止后立即触发告警 • 任务失败：任务运行失败后立即触发告警 • 任务异常/任务检测异常：任务在运行过程中出现异常立即触发告警 • 任务重启：任务重启超过累计次数后立即触发告警 • 读取速度、写入速度、读取吞吐、写入吞吐、任务写入延时、脏数据字节数、脏数据条数超过定义的值立即触发告警
告警阈值	设置的告警指标临界值，阈值范围内不会触发告警
接收人	接收任务告警的空间成员名称

告警级别	当前触发告警的重要程度，根据不同指标的告警级别，区分告警信息发送内容
告警方式	多选，支持电话、短信、微信、企业微信、邮件、HTTP 及企业微信群告警方式
操作	• 编辑：编辑当前告警规则，支持修改除告警指标外属性 • 删除：删除规则并停止告警

创建告警规则

用户可以通过自定义告警规则对当前任务进行运维监控。

1. 进入实时运维界面，单击**任务名 > 配置告警 > 创建告警规则**，即可进入告警规则配置界面。
2. 告警规则配置可参考下表进行配置：

基本信息	• 规则名称：用户可以自定义规则名称

	• 开关：用于控制当前规则的开关
告警指标	支持一个规则配置多个告警指标 • 状态指标：任务运行过程中出现失败、停止、异常、暂停、重启等情况达到一定的数值时触发告警。例如：任务暂停1分钟触发告警 • 读写指标：支持通过读取/写入速度、任务写入延时、读取/写入吞吐、脏数据字节/条数触发告警
告警通知	• 告警级别：根据不同指标的告警级别，区分告警信息发送内容，提供普通、重要及紧急三种级别 • 告警方式：支持邮件、短信、微信、电话、企业微信、HTTP 及企业微信群告警方式 • 接收人：支持指定当前空间人员及任务责任人接受告警信息

任务日志

最近更新时间：2024-07-18 17:43:21

在任务日志界面中，您可以查看间隔时间段内的管控日志和运行日志。同时支持自定义时间段和手动刷新。

告警事件

最近更新时间：2024-06-18 14:47:41

告警事件页面展示了已触发告警通知的所有告警规则信息，包括告警规则、告警级别、告警指标、告警阈值以及触达详情等。
支持按照日期（今天/昨天/近7天/近30天/自定义日期）快速检索，并支持任务/规则/接收人名称全局检索。

参数	说明
告警时间	本次触发告警的具体时间
告警级别	触发告警规则对应的级别
任务名称	触发告警的任务对象名称，可单击跳转到任务运维页面 **说明：** 如果任务类别为离线任务，则具体到触发告警的实例 ID。
规则名称	触发告警的规则名称，可点击跳转到任务下对应告警规则订阅页面
告警指标	当前规则内配置的告警指标类别
告警阈值	对告警指标设定的触发临界值
接收人	规则设定的默认接收人信息
触达详情	默认接收人对告警信息的触达接受情况，包括全部发送成功、部分发送成功、全部发送失败，单击**查看详情**，可在弹窗中具体查看每位用户接收情况

：表示该用户在该方式下通知发送成功

--：表示当前用户未配置对应消息通知方式，具体操作可在 腾讯云 账号信息 > **登录方式**中关联对应类别

：表示该用户在该方式下通知未发送成功

实时节点高级参数

最近更新时间：2024-04-18 17:34:21

参数说明

类型	参数级别	读/写	适用场景	配置内容：	描述
Mysql / tdsql-c Mysql	节点级别	读	单表+整库	scan.newly-added-table.enabled=true	**参数描述：** 设置这个参数，在**暂停** > **继续**后可以感知新增的表。默认是 false 1. 全增量同步时使用该参数，新增的表会读取存量数据后再读取增量数据 2. 增量同步时使用该参数，新增的表只会读取增量数据
		读	单表+整库	scan.incremental.snapshot.chunk.size=20000	**参数描述：** 对于数据分布均匀的任务，这个参数代表一个 chunk 内大约的条数，可以用总的数据量除以 chunk size 估算任务有多少个 chunk 数，chunk 数的多少影响了 jobmanager 是否 oom，目前2CU的情况下可以支持10w多 chunk，如果数据量太大，我们可以调大 chunk size 来减少 chunk 数量 **注意事项：** 大数据量任务（例如，总数据量1个亿以上，单条记录大于0.1M）一般建议设置20000
		读	单表+整库	split-key.even-distribution.factor.upper-bound=10.0d	**参数描述：** mysql 存量数据读取阶段，如果数据比较离散、主键字段的最大值超大，可以修改这个参数，来使用非均匀分割，减少由于主键值超大情况下导致 chunk 数量太大从而 jm oom 的问题 **注意事项：** 默认值为10.0d一般不用修改
		读	单表+整库	debezium.query.fetch.size=0	**参数描述：** 代表每次读取从数据库拉取的数据条数，默认是0代表jdbc 默认的 fetch size **注意事项：**

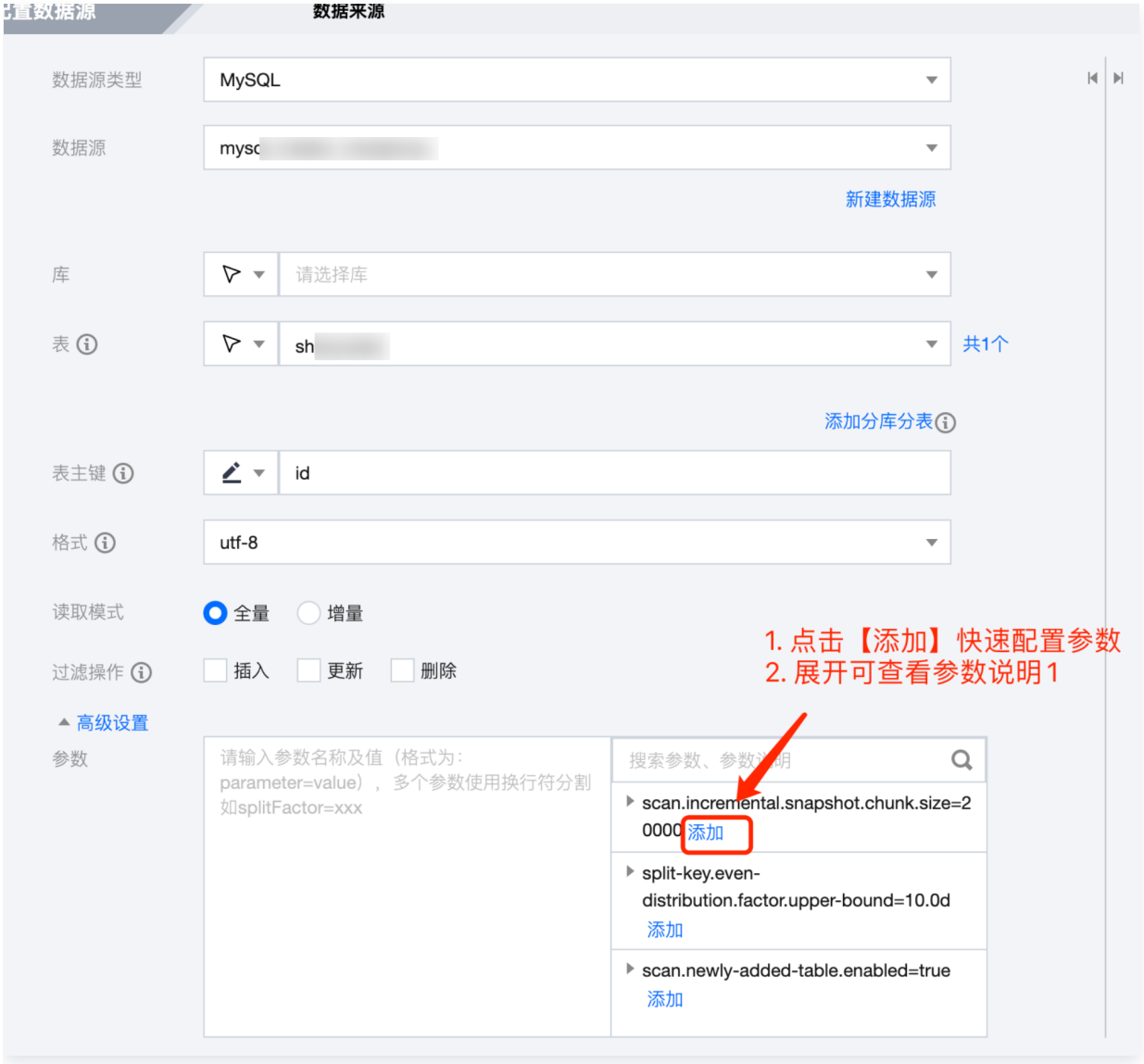
					1. 大任务（例如，总数据量1个亿以上，单条记录大于0.1M）只有一个读取实例时建议取1024条 2. 任务存在多个读取实例时建议降低这个值，减少内存消耗，建议取512条
		读	单表+整库	debezium.max.queue.size=8192	**参数描述：** 属性定义了内部队列中存储的最大事件数。如果达到此限制，Debzium 将暂停读取新事件，直到处理和提交尚未处理的事件。这个属性可以帮助避免过多事件积压在队列中，导致内存耗尽和性能下降。默认是8192 **注意事项：** 1. 大任务（例如，总数据量1个亿以上，单条记录大于0.1M）只有一个读取实例建议取4096 2. 任务存在多个读取实例时建议降低这个值，减少内存消耗，建议取1024
	作业级别	–	–	taskmanager.memory.managed.fraction=0.1	**参数描述：** 调整flink 程序 taskmanager 托管内存比例
		–	–	table.exec.sink.upsert-materialize=NONE	**参数描述：** 由于分布式系统中的 shuffle 会造成 Changelog 数据的乱序，所以 sink 接收到的数据可能在全局的 upsert 中乱序，所以要在 upsert sink 之前添加一个 upsert 物化算子。该算子接收上游 changelog 数据，并且给下游生成一个 upsert 视图。这个参数用于控制物化算子的添加 **注意事项：** 1. 默认情况下，在唯一 key 遇到分布式乱序时，该物化算子会被添加，也可以选择不物化（NONE），或者是强制物化（FORCE） 2. 可选值有：NONE、AUTO、FORCE
		–	–	table.exec.sink.not-null-enforcer=DROP	**参数描述：** 决定当 NOT NULL 字段遇到 null 值时任务如何处理 **建议值及作用：**

					1. ERROR：NOT NULL 字段遇到 null 值时抛出运行时异常。 2. DROP：NOT NULL 字段遇到 null 值时直接丢弃数据
dlc	节点级别	写	单表+整库	write.distribution-mode=hash	**参数描述：** dlc 并发写入，支持参数 none|hash（默认）|range： 1. none：存在主键则根据主键进行并发写入，否则单并发写入 2. hash：存在分区字段，根据分区字段并发写入，否则根据参数 none 的策略写入 3. range：暂不支持，策略和 none 一致
	节点级别	写	单表+整库	write.compact.enable=true	**参数描述：** 控制 dlc 是否开启小文件合并： 1. false：默认不开启整库同步界面有选项开启 2. true: 设置为小文件合并开启 **注意事项：** 单表同步需要手动配置该参数
	节点级别	写	单表+整库	write.compact.snapshot.interval=20	**参数描述：** DLC 合并小文件频率，单位为一个 checkpoint。默认是20个 checkpoint 的时候做一次合并， **注意事项：** 需要配置 write.compact.enable = true 才生效
doris	节点级别	写	仅单表	sink.properties.*=xxx	**参数描述：** Stream Load 的导入参数。 例如 'sink.properties.column_separator' = ', ' **详细配置参考：** https://doris.apache.org/zh-CN/docs/ecosystem/flink-doris-connector/
	节点级别	写	仅单表	sink.properties.columns=xxx	**参数描述：** 配置 columns 的函数映射关系。 例如 'sink.properties.columns' = 'dt,page,user_id,user_id=to_bitmap(user_id)'

					参考： https://doris.apache.org/zh-CN/docs/data-operate/import/import-way/stream-load-manual/
	节点级别	写	单表+整库	• sink.batch.size = 100000 • sink.batch.bytes= 83886080 • sink.batch.interval= 10s	**参数描述：** 提高写入 doris 效率 **注意事项：** tm cu 建议设置为2CU，避免 tm oom
oracle	节点级别	读	单表+整库	• 'debezium.log.mining.strategy' = 'online_catalog' • 'debezium.log.mining.continuous.mine' = 'true'	**参数描述：** 开启这个参数后，可以减少数据同步的延迟和减少 redo 日志的存储，适用于单表同步+整库同步（指定表） **注意事项：** 1. 设置后无法感知新增表，如果配置同步所有库表 / 指定库 将无法读取新增表数据 2. 不适用于 oracle19 版本不支持这个参数，因此需要设为 false，否则将导致任务失败。 **参考：** https://github.com/ververica/flink-cdc-connectors/wiki/FAQ(ZH)
	节点级别	读	单表+整库	debezium.lob.enabled=false	**参数描述：** 是否同步 blob 类型数据，默认是 false **注意事项：** 1. 如果设置会 true，可能会影响同步性能 2. oracle 默认推荐配置为 false
mongodb	节点级别	读	仅单表	scan.incremental.snapshot.enabled=true	**参数描述：** 开启并发读，默认为 false **注意事项：** mongodb4.0 版本以上才支持
	节点级别	读	仅单表	copy.existing=false	**参数描述：** 是否从源集合复制现有数据: 1. 默认是 true，表示从全量开始读取数据 2. false 表示从增量开始读取数据
	节点级别	读	仅单表	poll.await.time.ms	**参数描述：** 变更事件拉取时间间隔，默认为1500ms

					注意事项： 1. 对于变更频繁的集合，可以适当调小拉取间隔，提升处理时效 2. 对于变更缓慢的集合，可以适当调大拉取时间间隔，减轻数据库压力
	节点级别	读	仅单表	poll.max.batch.size	**参数描述：** 每一批次拉取变更事件的最大条数，默认为1000条 **注意事项：** 调大改参数会加快从 Cursor 中拉取变更事件的速度，但会提升内存的开销
	节点级别	读	仅单表	scan.incremental.snapshot.chunk.size.mb	**参数描述：** 增量快照的块大小，单位是mb ，默认大小为64mb
	节点级别	读	仅单表	changelog.normalize.enabled	**参数描述：** 是否开启 changelogNormalize 算子，默认为 true，表示开启 **注意事项：** mongodb 缺乏 -u 消息，开启该算子会补齐 -u 消息，但是会消耗一定性能。关闭该算子会提高传输速度，但是 delete 操作无法同步到下游，其他操作不影响

节点级配置方式


数据来源
数据源类型
MySQL
数据源
新建数据源
库
请选择库
表
共1个
添加分库分表
表主键
id
格式
utf-8
读取模式
全量
增量
过滤操作
插入
更新
删除
高级设置
参数
请输入参数名称及值（格式为：parameter=value），多个参数使用换行符分割
如splitFactor=xxx
搜索参数、参数说明
scan.incremental.snapshot.chunk.size=20000
添加
split-key.even-distribution.factor.upper-bound=10.0d
添加
scan.newly-added-table.enabled=true
添加
1. 点击【添加】快速配置参数
2. 展开可查看参数说明1

任务级配置方式

说明：

1. 一个参数一行；若需配合使用的参数写在一行内。
2. 每个参数带默认值。

离线同步任务配置与运维
离线同步支持的数据源

最近更新时间：2024-04-18 17:34:21

背景信息

数据集成提供离线数据同步能力，该能力通过定期运行方式批量读取来源库表中数据，并同步写入至目标端。

支持的数据源

数据源		离线同步	
		离线读取	离线写入
关系型数据库	MySQL	✓	✓
	TDSQL-C MySQL	✓	✓
	PostgreSQL	✓	✓
	TCHouse-P	✔	✔
	SQL Server	✓	✓
	Oracle	✓	✓
	DB2	✓	✓
	SAP HANA	✓	✓
	DM	✓	✓
	SAP IQ(Sybase)	✓	-
大数据	Hive	✓	✓
	DLC	✓	✓
	Doris	✔	✓
	ClickHouse	✓	✓
	Iceberg	✓	✓
	HBase	✓	✓
	HDFS	✓	✓

	Kudu	✓	✓
	TBase	✓	✓
	GaussDB	✓	✓
	GBase	✓	✓
	Greenplum	✓	✓
半结构化	COS	✓	✓
	Rest API	✓	–
	FTP	✓	✓
	SFTP	✓	✓
NoSQL	Redis	–	✓
	Elasticsearch	✓	✓
	Mongo	✓	✔
消息队列	Kafka	✓	✔

离线任务配置概览

最近更新时间：2024-04-02 16:17:11

背景信息

单表同步采用固定字段同步的方式，仅将在任务配置中指定映射关系的来源字段数据同步至目标端。单表任务支持画布、表单两种配置模式，覆盖 MySQL、Hive、DLC、Doris 等数据源。

条件与限制

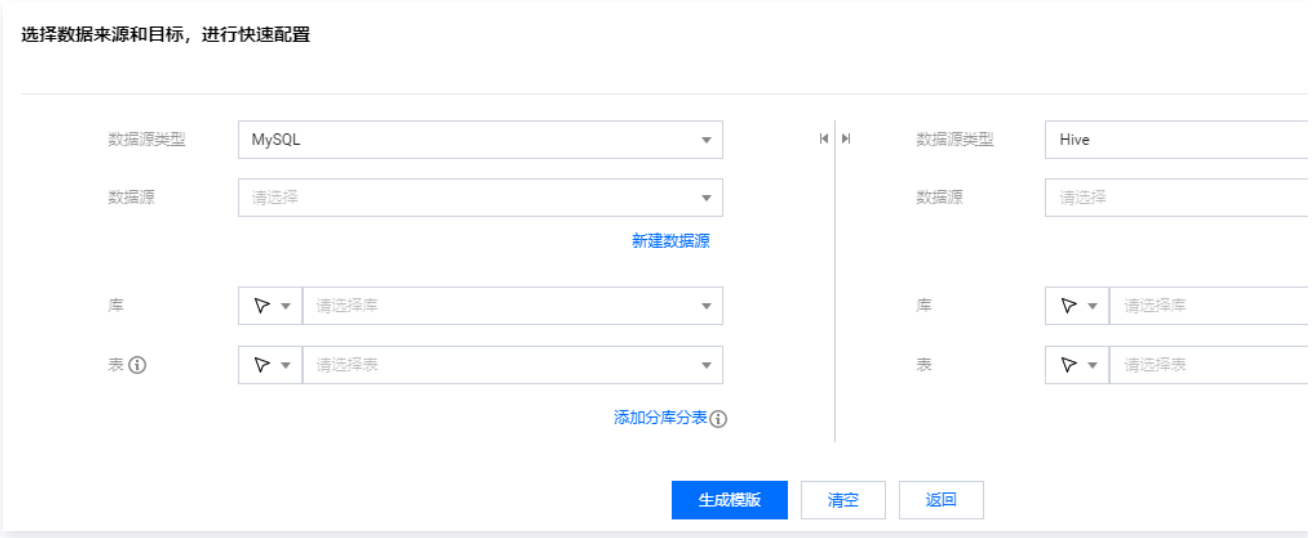
1. 已配置好来源及目标端的数据源以备后续任务使用。详情请参见 数据源管理与配置方式 。
2. 已购买数据集成资源组。详情请参见 配置集成资源组 。
3. 已完成数据集成资源组与数据源的网络连通。详情请参见 集成连通性与使用规划 。
4. 已完成数据源环境准备。您可以基于您需要进行的同步配置，在同步任务执行前，授予数据源配置的账号在数据库进行相应操作的权限。
5. 若数据源配置的数据库账号不具备读写权限将导致任务运行失败，请根据实际读写场景配置具备相应权限的账号。

操作步骤

步骤一：新建离线同步任务并选择配置模式

在数据集成页面下，单击**同步链路 > 离线同步**即可进入同步任务列表。在弹窗中配置任务基本信息，单击**确定**后即可进入任务配置页面。

参	说明

数	
任务名称	必填项。
任务模式	• 表单模式：仅提供读取、写入节点，适用于单表至单表固定字段同步。适用于 ODS 层无需数据清洗环节的 • 画布模式：提供读取、写入、转换三类节点。适用于包含清洗环节、多对多数据链路。 • 脚本模式：支持初始化的脚本模式配置页面，支持用户选择不同的数据来源、数据目标，展示对应的脚本模 ○ 用户需要先选择数据来源与数据目标，未选择的状态下不允许编辑。 ○ 本期优先支持 mysql > hive 、hive > mysql、mysql > DLC、mysql > doris。 ○ 选择后，展示对应的脚本模块。 ○ 在脚本中，用户可以手动编写数据源、连接信息等参数。 ○ 支持在脚本中写 sql 语句，将 querysql 写到 connection 中。 
描述	选填项。

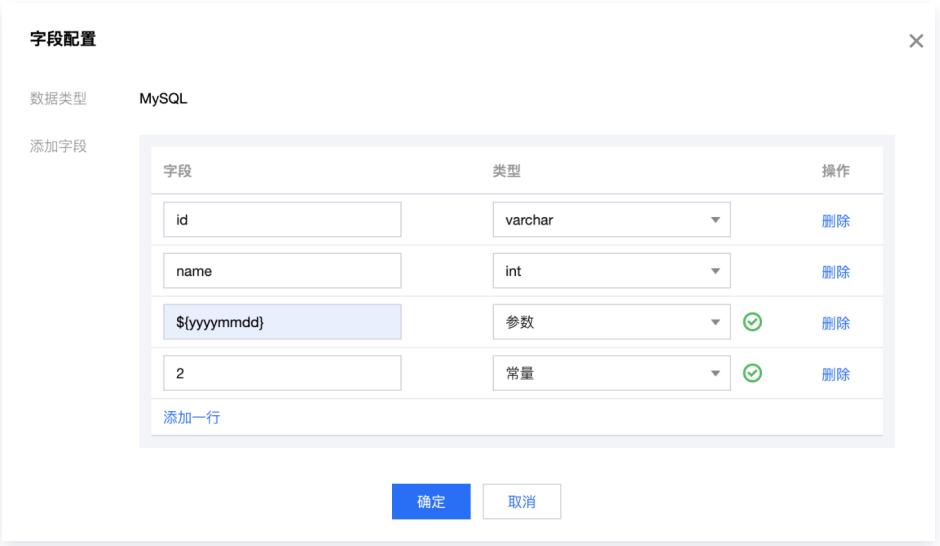
说明：
脚本模式目前支持以下数据源：
- 读取：MySQL、HIVE。
- 写入：HIVE、Clickhouse、DLC、Doris。

步骤二：数据节点配置

配置读取节点

读取节点配置包括基本信息、数据来源、数据字段三部分。

- 基本信息
  节点名称不可为空，且单个任务内不可存在同名的数据节点。
- 数据来源
  配置需要读取的库表对象以及同步方式等信息。
- 数据字段
  根据配置的数据表对象，系统支持默认拉取字段元数据信息以及手动配置字段两种方式。
  - 默认拉取：针对 MySQL、Hive、PostgreSQL 等类型，系统已支持根据其库表信息自动拉取元数据字段及类型，无需手动编辑。
  - 手动配置：文件（如 HDFS、COS）以及列式存储数据源（如 HBase、Mongo）等数据源系统不支持自动拉取元数据，可单击**字段配置**手动添加字段名称及类型。读取节点还额外支持配置时间参数以及常量。



> **说明**
> - 时间参数字段：仅离线任务的读取节点支持配置时间参数字段，常用将实例运行时间值写入表的一级或多级分区。
> - 常量字段：仅读取节点支持配置常量字段。常量字段可在来源与目标表字段个数不一致的情况下固定将某个常量值写入目标表。

配置转换节点

转换节点配置包括基本信息、转换规则、数据字段三部分。其中，转换转换节点必须作为读取节点下游，在创建与读取节点连线后系统将自动获取上游节点内字段信息，同时根据转换规则完成数据转换。

- 基本信息
  配置节点名称信息。节点名称不可为空，且单个任务内不可存在同名的数据节点。
- 转换规则
  配置字段或数据级转换规则，其中字段信息继承自上游节点，在与上游节点连线后系统将自动获取上游节点内字段信息。不同转化节点规则及参数说明请参见 转换节点 。

- 数据字段
  默认拉取上游节点全部数据字段用于后续写入节点映射。

配置写入节点

写入节点配置包括基本信息、数据来源、数据字段、字段映射四部分。写入节点将根据连线关系，将上游数据内容写入目标对象内。

- 基本信息
  节点名称不可为空，且单个任务内不可存在同名的数据节点。
- 数据来源
  配置需要读取的库表对象以及同步方式等信息。
- 数据字段
  根据配置的数据表对象，系统支持默认拉取字段元数据信息以及手动配置字段两种方式。
  - 默认拉取：针对 MySQL、Hive、PostgreSQL 等类型，系统已支持根据其库表信息自动拉取元数据字段及类型，无需手动编辑。
  - 手动配置：文件（如 HDFS、COS）以及列式存储数据源（如 HBase、Mongo）等数据源系统不支持自动拉取元数据，可单击**字段配置**手动添加字段名称及类型。
- 字段映射
  写入节点相对于读取节点需额外配置字段映射关系。字段映射关系旨在通过连线的方式指定目标字段内容的来源，支持同名映射、同行映射、以及手动连线三种方式配置来源与目标节点间关系。

> **说明**
> - 配置字段映射的前提为当前写入节点有已连线的来源（读取节点或转换节点）。
> - 未配置映射关系的目标字段内容将为空或保持不变。
> - 若来源字段类型与目标字段类型间无法转换，可能会导致任务失败。

步骤三：离线任务属性配置

离线任务属性配置包括**基本属性、任务调度、和资源配置**三部分：

基本属性

设置任务基本属性、使用资源以及数据链路通道信息。

类别	参数	说明
基本属性	任务名称/类型	展示当前任务名称及类型基本信息。
	责任人	对此任务负责的一个或多个空间成员名称，默认为任务创建者。
	描述	展示当前任务备注信息。
资源配置	集成资源组	指定当前任务使用的集成资源组名称，一个任务仅可绑定一个资源组。
通道设置	脏数据阈值	脏数据是指同步过程中写入失败的数据脏数据阈值是指同步中可容忍的最大脏数据条或字节数，一旦超过该阈值，任务将自动结束。默认阈值为0，即不容忍脏数据。
	并发数	实际执行时期望任务的最大并发数，实际执行时由于资源、数据源类型和任务优化结果等原因并发数可能小于等于此值。该值越大，预分配执行机资源越多。
	同步速率限制	按照流量或记录条数限制同步速率以保护数据来源端或者数据去向端的读写压力。该值为最大运行速率，默认-1表示不限制速率。

任务调度

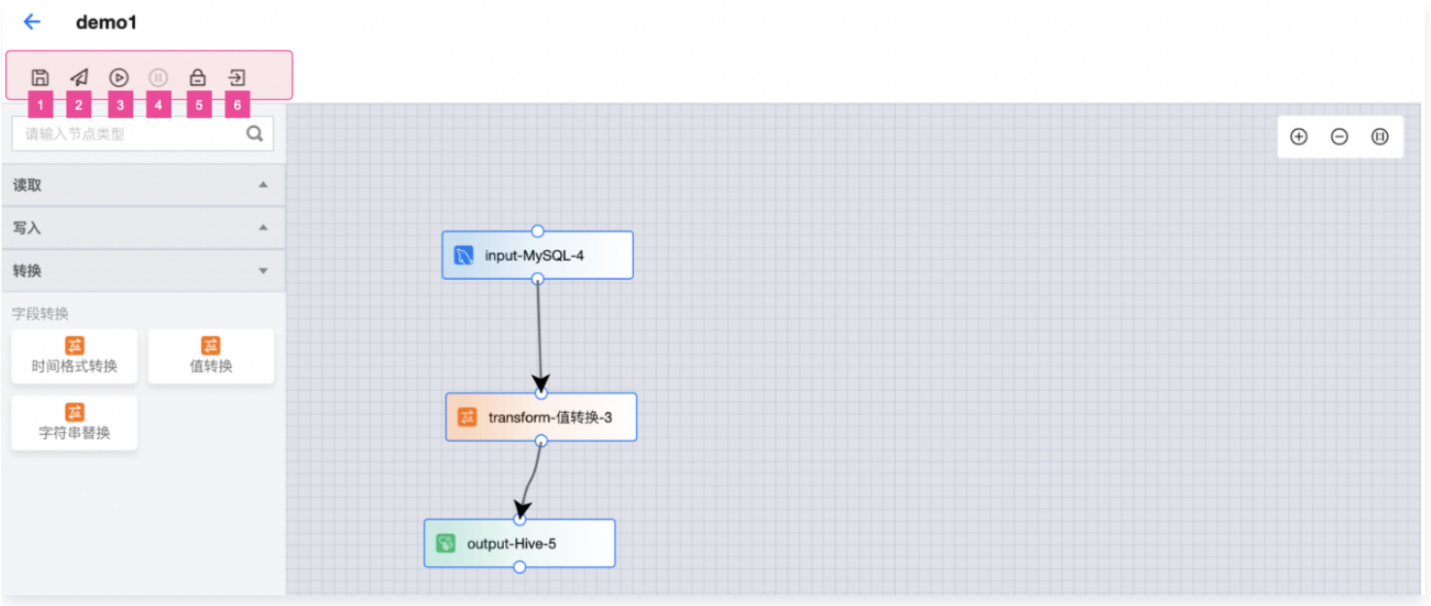
设置当前任务周期运行计划，包括调度时间及依赖属性。

类别	参数	说明
调度时间	调度方式	周期调度：任务根据配置调度计划周期运行。 一次性执行：任务仅在指定时间运行一次。
	生效日期	调度时间配置的有效时间段，系统会在该时间范围内按照时间配置自动调度，超过有效期将不会再自动调度。
	调度周期	调度计划间隔步长单位，支持年、月、周、天、小时、分钟： • 分钟：需指定具体执行开始时间及间隔，任务将从每小时执行分钟开始，按时间间隔周期运行。如执行时间为02:00～23:59，间隔为5分钟，则任务将从02:00开始每隔5分钟运行一次实例。 • 小时：需指定具体执行开始、结束时间及间隔。如执行时间为02:20～05:00，间隔为1小时，则任务将在02:20、03:20、04:20分别运行一次。 • 天：需指定每天具体执行时刻，任务每天仅在该时刻运行。 • 周：需执行每周固定运行的天数（支持多选）以及时间。任务仅在指定当天的该时刻运行。 • 月：指定每月固定运行的号数及时间。若选择月末，将根据不同的月份取最后一天运行。

		• 年：指定每年固定运行日期及时间。
依赖属性	自依赖	自依赖是指同一任务中不同实例之间的依赖关系： • 有序串行：当前实例依赖前一个周期实例的状态。 • 无序串行：当前实例和前一个周期实例没有依赖关系，如果一个任务同时存在多个实例，系统随机选取一个实例运行。同时只有一个实例是运行状态。 • 并行：前一个周期实例和后一个周期实例之间没有依赖关系，如果一个任务同时存在多个实例，多个实例会同时运行。
	重试等待时间	实例运行失败后，每次重试运行的最大等待时间间隔。若超过此值实例仍未重试运行，实例将被置为失败。
	失败重试次数	实例运行失败后，最大重试次数。若过超过此值，任务将被置为失败。

步骤四：任务测试运行与提交

离线同步任务在配置完成后可进行在线测试运行或提交到生产调度环境中，目前可在任务配置页面支持保存、提交、测试运行、调试停止、锁定/解锁以及前往运维操作。



序号	参数	说明
1	保存	保存当前任务配置信息，包括数据节点配置、节点连线、任务属性和任务调度配置。
2	提交	将当前任务提交至生产环境，提交后任务将按调度属性周期运行，同时提交任务将在**任务运维 > 离线运维**生成任务及实例记录。 **说明：**

		• 提交前任务将默认保存最新配置。 • 提交前任务将进行必要性检测，包括任务节点配置、任务连线、资源组等。若必要性检测不通过，任务将提交失败并提示。
3	测试运行	调试运行当前任务。
4	调试停止	终止当前正在测试运行中的任务。
5	锁定/解锁	默认创建者为首个持锁者，仅允许持锁者编辑任务配置及运行任务。若锁定者5分钟内没有编辑操作，其他人可单击图标抢锁，抢锁成功可进行编辑操作。
6	前往运维	根据当前任务名称快捷跳转至任务运维页面。

任务提交检测

参数	说明

检测存在异常	支持跳过异常直接提交，或者终止提交。
检测仅存在警告及以下	可直接提交。

提交结果

- 任务提交中：
  - 展示提交进度百分比。
  - 提示用户勿刷新/关闭页面，文案：当前任务已提交成功，可前往运维进行任务状态及数据管理。
- 任务提交结果-成功：
  - 展示任务提交成功结果。
  - 提示成功及后续跳转：文案 “提交成功，10秒后将跳转至当前任务运维详情页面” “当前任务已提交成功，可前往运维进行任务状态及数据管理”。
- 展示任务提交失败原因：
  - 失败原因返回。

后续步骤

完成任务配置后，您可以对已创建的任务进行运维及监控告警，如对任务配置监控报警，并查看任务运行的关键指标等。详情请参见 离线任务运维 。

节点配置参数及说明

最近更新时间：2024-07-18 17:43:21

离线实时同步支持输入、输出和转换三种类型的数据节点，本文将重点说明各读、写及转换节点配置及底层数据类型映射。

读写节点配置

数据库类型	读取节点	写入节点
关系型数据库	MySQL 离线节点（读取）	MySQL 离线节点（写入）
	TDSQL-C Mysql 离线节点（读取）	TDSQL-C Mysql 离线节点（写入）
	PostgreSQL 离线节点（读取）	PostgreSQL 离线节点（写入）
	SQL Server 离线节点（读取）	SQL Server 离线节点（写入）
	Oracle 离线节点（读取）	Oracle 离线节点（写入）
	DB2 离线节点（读取）	DB2 离线节点（写入）
	DM 离线节点（读取）	DM 离线节点（写入）
	SAP HANA 离线节点（读取）	SAP HANA 离线节点（写入）
	SAP IQ（sybaseIQ）离线节点（读取）	-
大数据	HIVE 离线节点（读取）	Hive 离线节点（写入）
	HBase 离线节点（读取）	HBase 离线节点（写入）
	Clickhouse 离线节点（读取）	Clickhouse 离线节点（写入）
	DLC 离线节点（读取）	DLC 离线节点（写入）
	Kudu 离线节点（读取）	Kudu 离线节点（写入）
	HDFS 离线节点（读取）	HDFS 离线节点（写入）
	Greenplum 离线节点（读取）	Greenplum 离线节点（写入）
	GaussDB 离线节点（读取）	GaussDB 离线节点（写入）
	Gbase 离线节点（读取）	Gbase 离线节点（写入）
	TBase 离线节点（读取）	TBase 离线节点（写入）

	Iceberg 离线节点（读取）	Iceberg 离线节点（写入）
	–	Doris 离线节点（写入）
消息列表	Kafka 离线节点（读取）	–
NoSQL	Mongo 离线节点（读取）	–
	Elasticsearch 离线节点（读取）	Elasticsearch 离线节点（写入）
	–	Redis 离线节点（写入）
半结构化	COS 离线节点（读取）	COS 离线节点（写入）
	FTP 离线节点（读取）	FTP 离线节点（写入）
	SFTP 离线节点（读取）	SFTP 离线节点（写入）
	Rest API 离线节点（读取）	–

转换节点配置

转换类型	节点
字段转换	字符串替换
	字段分割
数据清洗	数据过滤
	去重
	数据连接（join）

时间参数说明

最近更新时间：2024-07-09 22:01:41

对于周期性运行的离线任务而言，系统支持通过使用时间参数自动获取周期任务实例的数据时间。
时间参数使用 ${…} 进行自定义，支持例如${yyyyMMdd}、${yyyy-MM-dd}、${HH：mm：ss}和${yyyyMMddHHmmss}等。其中，yyyy 表示4位的年份，yy 表示2位的年份，MM 表示月，dd 表示天，HH 表示小时，mm 表示分钟，ss 表示秒。各部分之间支持灵活组合，如：
系统的内置参数 ${timestamp} 作为调度时间对应的10位时间戳，精度到秒级。
示例如下：
以 20210710080000时间为基准：

时间	时间参数格式	备注
后 N 年	${yyyyMMdd+Ny}	若引用 dt=${yyyyMMdd-1M}，将执行替换：dt=20210610 若引用 dt=${yyyyMMdd-1d}，将执行替换：dt=20210709 若引用 time=${yyyyMMddHHmmss-3h}，将执行替换：time=20210710050000 若引用 ti=${yyyyMMddHHmmss-25m}，将执行替换：ti=20210710073500
前 N 年	${yyyyMMdd-Ny}	
后 N 月	${yyyyMMdd+NM}	
前 N 月	${yyyyMMdd-NM}	
后 N 周	${yyyyMMdd+Nw}	
前 N 周	${yyyyMMdd-Nw}	
后 N 天	${yyyyMMdd+Nd}	
前 N 天	${yyyyMMdd-Nd}	
后 N 小时	${yyyyMMddHHmmss+NH}	
前 N 小时	${yyyyMMddHHmmss-NH}	
后 N 分钟	${yyyyMMddHHmmss+Nm}	

前 N 分钟	${yyyyMMddHHmmss-Nm}	
调度时间戳	${timestamp}	${timestamp}=1625875200 该参数为固定参数，暂不支持使用 “+” 、 “-” 等运算符处理。

对于常见的日期提供了快捷的转换表达式，如下：

序号	时间参数格式	说明
1	${yyyyMMdd+TE}	TENDAY END 数据日期对应旬末
2	${yyyyMMdd+ME}	MONTH END 数据日期对应月末
3	${yyyyMMdd+QE}	QUARTER END 数据日期对应季末
4	${yyyyMMdd+HYE}	HALF YEAR END 数据日期对应半年末
5	${yyyyMMdd+YE}	YEAR END 数据日期对应年末
6	${yyyyMMdd+TS}	TENDAY START 数据日期对应旬初
7	${yyyyMMdd+MS}	MONTH START 数据日期对应月初
8	${yyyyMMdd+HYS}	HALFYEAR START 数据日期对应半年初
9	${yyyyMMdd+YS}	YEAR START 数据日期对应年初
10	${yyyyMMdd+PME}	PRI MONTH END 数据日期对应上月月末
11	${yyyyMMdd+PYE}	PRI YEAR END 数据日期上年年末

离线同步运维
离线任务运维

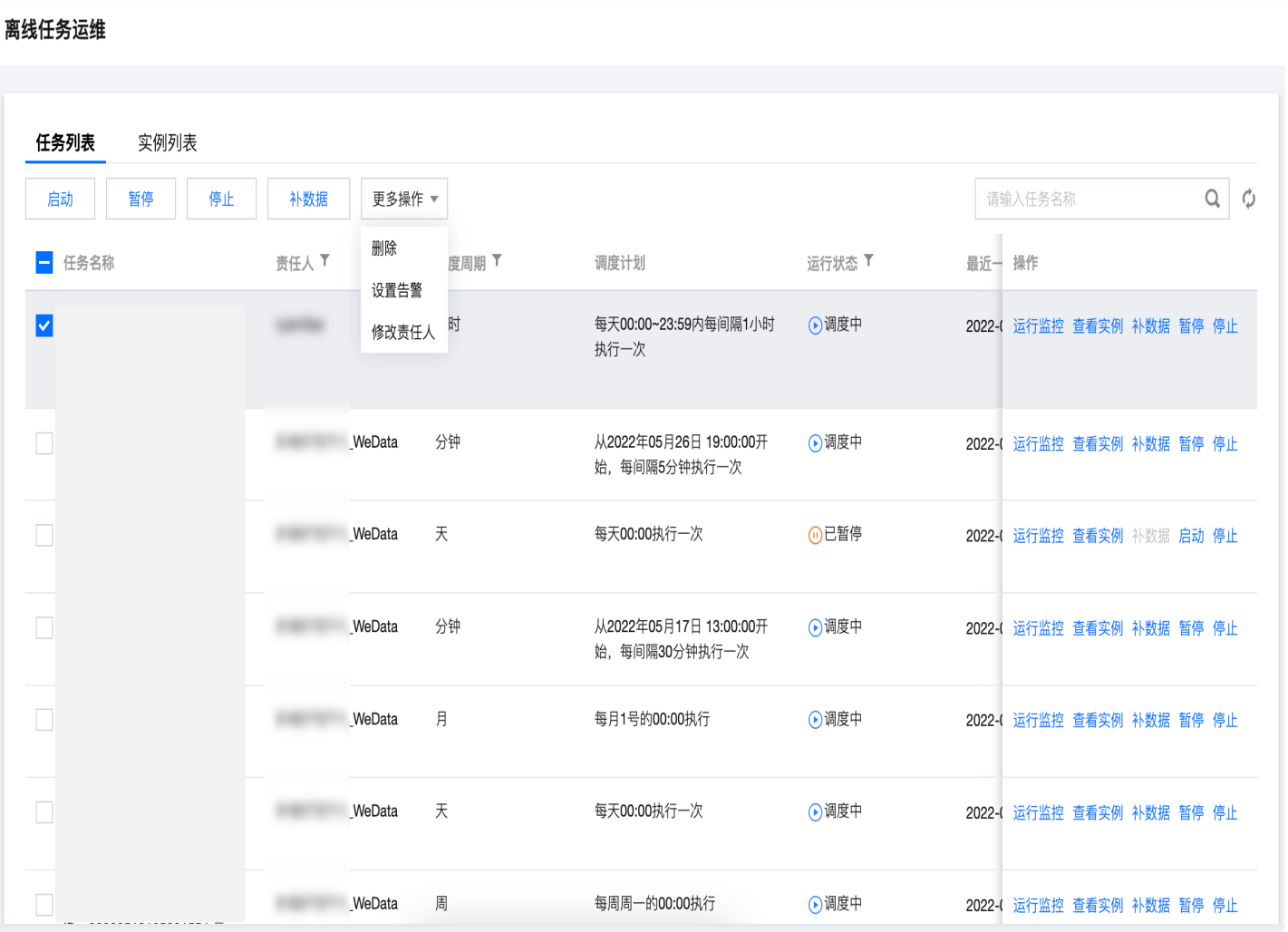
最近更新时间：2024-07-09 22:01:41

离线运维页面是对离线同步内所有已提交离线同步任务的统一运维中心，包括周期任务运维和实例运维两部分。

任务运维

任务列表页面以列表形式默认展示当前账号下所有提交到调度系统中的周期运行任务。

离线任务运维



参数	说明
任务名称	当前记录归属对任务名称
责任人	当前任务创建时配置的责任人名称
调度周期	当前任务配置的周期调度频率

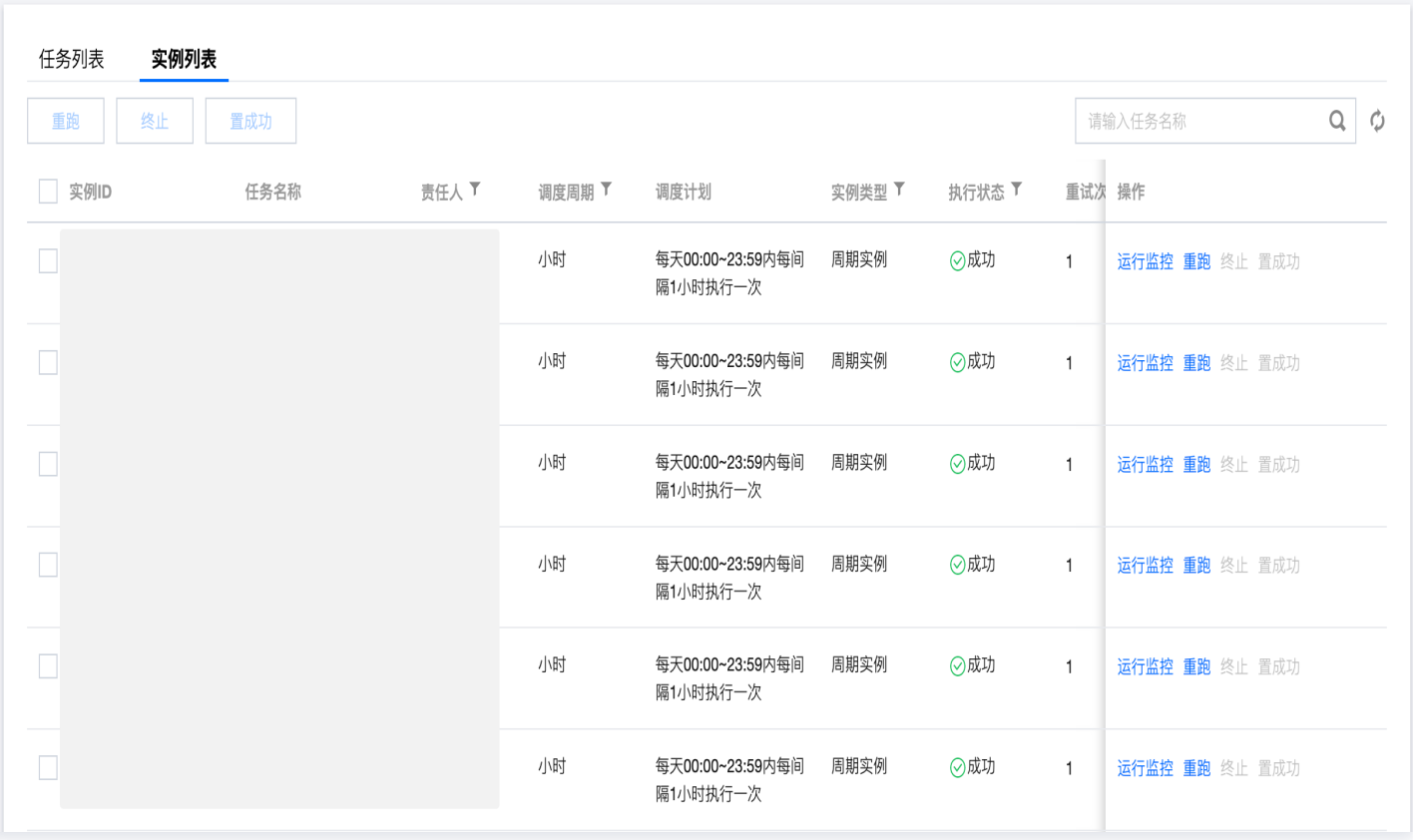
调度计划		当前任务详细调度计划
运行状态		当前任务的调度运行状态 调度中：任务已提交调度，处于正常调度中 已暂停：暂时中断当前任务调度，后续可重启 已停止：当前任务调度被终止 停止中：已对当前进行停止操作，状态扭转中
最近一次提交时间		任务最近一次提交至调度系统的时间
操作	运行监控	包括任务指标统计、监控规则配置等说明：指标统计详见 离线同步指标统计 ，监控规则配置详见 告警订阅
	查看实例	单击跳转至该任务的任务实例信息列表
	补数据	对该任务进行批量补数据，仅对“**调度中**”的任务有效
	启动	启动节点的调度任务，仅对“**已暂停**”和“已停止”的有效
	暂停	暂停节点的调度任务，仅对“**调度中**”的有效暂停后任务将不会再生成新的实例，已生成实例将继续运行
	删除	将该任务及任务下的所有实例一起删除，仅对“**已停止**”的任务有效
	停止	对该节点任务所有“**等待运行**”和“**运行中**”的实例进行终止，并不再产生新的实例
	告警设置	设置任务的告警信息，支持批量操作
	修改责任人	修改任务的责任人，支持批量操作

实例运维

最近更新时间：2024-07-09 22:01:41

任务实例是周期运行的任务按照调度配置周期性生成的实例快照，在实例运维页面，以实例 ID 为对象展示了不同任务下关联周期实例运维状态、同时支持实例包括运行监控、重跑等操作。

离线任务运维



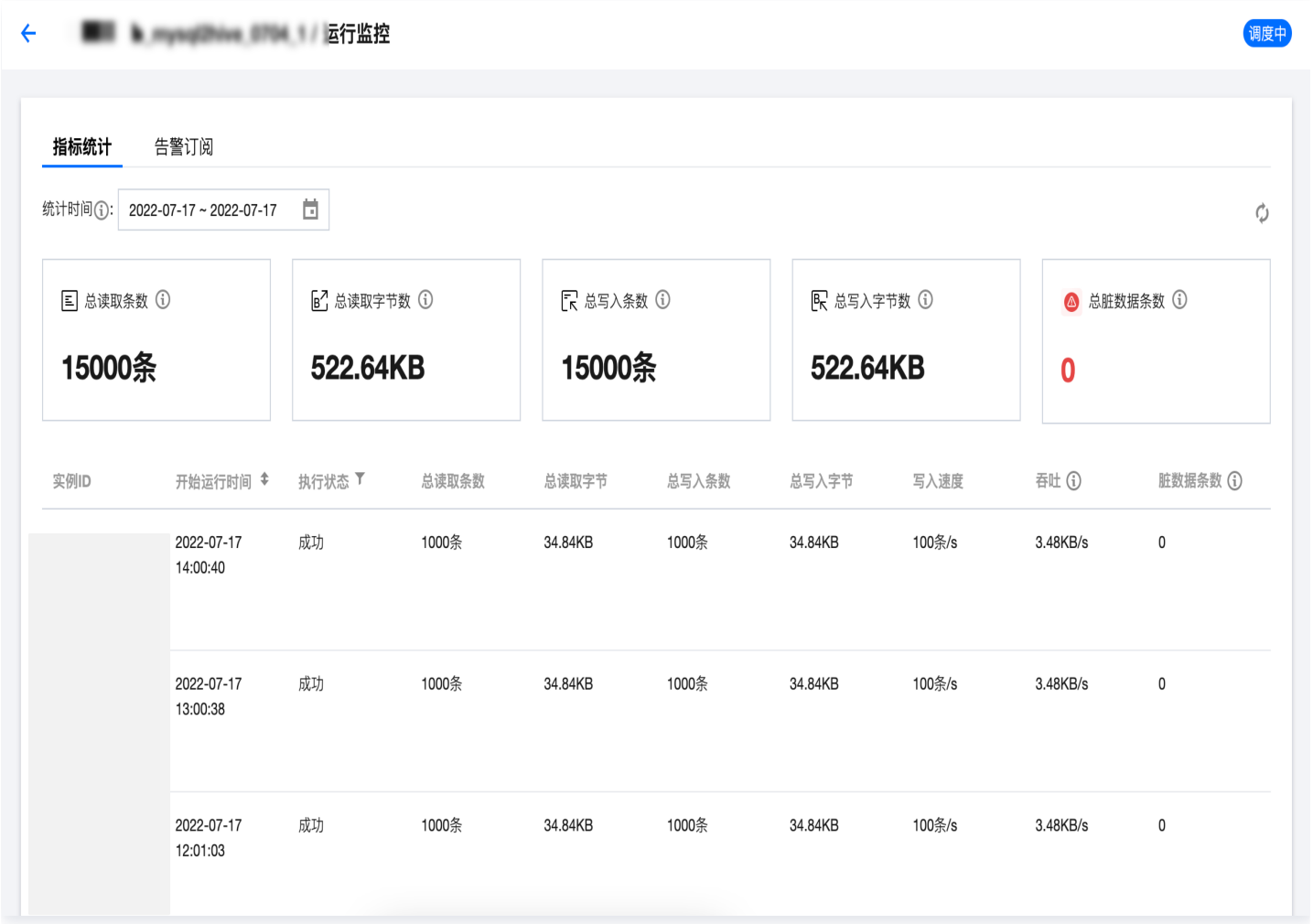
参数	说明
实例ID	系统为任务每次周期运行生成的系统 ID 及实例数据时间
任务名称	当前实例归属的任务名称
责任人	任务配置的责任人名称
调度计划	当前实例所属任务配置的周期运行计划详情
实例类型	周期实例：根据任务运行计划自动生成、自动执行的实例 补数实例：由用户手动单击补数操作生成的实例
执行状态	当前实例的运行状态： 成功：实例已运行成功

		失败：实例已运行失败 执行中：实例运行中 等待执行：包括等待时间、等待自依赖、等待调度资源、等待计算资源等情况 正在终止：刚单击终止操作的实例 失败重试：实例上一次运行失败正在准备重试
重试次数		当前实例已重试运行的次数
开始运行时间		最近一次开始运行时间
运行耗时		总共运行时长，若任务未运行完成显示为“-”
操作	运行监控	包括实例指标统计、日志等说明：指标统计详见 离线同步指标统计
	重跑	对“成功”或“失败”状态的实例进行再执行
	终止	对“等待运行”或“运行中”的实例进行强制终止
	置成功	对“等待运行”或“失败”的实例设置成“成功”状态。若任务存在串行自依赖，当前实例置成功后可触发后续实例运行

离线同步指标统计

最近更新时间：2024-07-09 22:01:41

离线同步运行数据指标支持从任务及实例两个维度查看同步运行情况。离线运维界面支持从任务或实例后记录后，单击运行监控进入到指标统计界面。



概览模块参数说明：

指标参数	说明
总读取条数	任务在当前统计时间范围内所有实例总共读取的记录条数
总读取字节数	任务在当前统计时间范围内所有实例总共读取的记录大小
总写入条数	任务在当前统计时间范围内所有实例总成功写入的记录条数
总写入字节数	任务在当前统计时间范围内所有实例总成功写入的记录大小
总脏数据条数	任务在当前统计时间范围内所有实例写入失败的记录条数

实例列表指标参数说明：

指标参数	说明
实例 ID	展示当前实例的 ID 及数据时间
开始运行时间	当前实例实际开始运行的时刻
执行状态	当前实例实际开始运行状态，仅成功/失败的实例有具体的统计数据 • 成功：实例已运行成功 • 失败：实例已运行失败 • 执行中：实例运行中 • 等待执行：包括等待时间、等待自依赖、等待调度资源、等待计算资源等情况 • 正在终止：刚单击**终止**操作的实例 • 失败重试：实例上一次运行失败正在准备重试
总读取条数	当前实例在运行期间总读取记录条数
总读取字节数	当前实例在运行期间总读取记录字节数
总写入条数	当前实例在运行期间成功写入的记录条数
总写入字节数	当前实例在运行期间成功写入的记录字节数
写入速度	当前实例务在写入时候的速率，单位条/秒
吞吐	当前实例在写入时候的字节吞吐量，单位 kb/秒
脏数据	当前实例在运行期间写入失败的记录条数

告警订阅

最近更新时间：2024-07-09 22:01:41

告警订阅为每个同步任务提供基于不同指标及告警阈值的创建任务告警规则，一个任务支持创建多个不同告警级别、不同告警规则的告警监控。
离线运维界面支持从任务记录中，单击运行监控 > 告警订阅，即可进入告警订阅界面管理规则。

参数	说明
规则ID	系统默认生成的当前告警规则 ID
规则名称	用户指定配置的告警规则名称
规则状态	• 开启：开启当前告警规则，任务运行时将根据告警规则及阈值触发告警 • 关闭：关闭告警验证
告警指标	离线同步任务目前支持配置任务失败和任务超时告警： • 任务失败：当任务超过指定的失败次数后触发告警 • 任务超时：当任务运行超过指定时长（分钟）后触发告警
告警阈值	设置的告警指标临界值，阈值范围内不会触发告警

接收人	接收任务告警的空间成员名称
告警级别	当前触发告警的重要程度，根据不同指标的告警级别，区分告警信息发送内容
告警方式	多选，支持电话、短信、微信、企业微信、邮件告警方式

创建告警规则


告警规则
基本信息
规则名称
请输入
开关
打开
告警指标
状态指标
任务失败
任务停止
任务异常
任务检测异常
任务暂停
阈值
请输入
分钟
累计暂停超过 --分钟 后触发告警
任务重启
阈值
请输入
次
累计重启超过 --次 后触发告警
读写指标
读取速度
写入速度
任务写入延时
读取吞吐
写入吞吐
脏数据字节
阈值
大于
请输入
字节
脏数据字节超过 --字节 后触发告警
脏数据条数
阈值
大于
请输入
条
脏数据条数超过 --条 后触发告警
告警通知
告警级别
普通
重要
紧急


告警方式
邮件
短信
微信
电话
企业微信
HTTP
企业微信群
接收人
指定人员
任务责任人
接收人
请选择
保存
取消

告警事件

最近更新时间：2024-07-09 22:01:41

告警事件页面展示了已触发告警通知的所有告警规则信息，展示了包括告警规则、告警级别、告警指标及阈值以及触达情况等。
支持按照日期（今天/昨天/近7天/近30天/自定义日期）快速检索，并支持任务/规则/接收人名称全局检索。

参数	说明
告警时间	本次触发告警的具体时间。
告警级别	触发告警规则对应的级别。
任务名称	触发告警的任务对象名称，可单击跳转到任务运维页面。 **说明：** 如果任务类别为离线任务，则具体到触发告警的实例 ID。
规则名称	触发告警的规则名称，可点击跳转到任务下对应告警规则订阅页面。
告警指标	当前规则内配置的告警指标类别。
告警阈值	对告警指标设定的触发临界值。
接收人	规则设定的默认接收人信息。
触达详情	默认接收人对告警信息的触达接受情况，包括全部发送成功、部分发送成功、全部发送失败，单击**查看详情**，可在弹窗中具体查看每位用户接收情况。

：表示该用户在该方式下通知发送成功。

--：表示当前用户未配置对应消息通知方式，具体操作可在 腾讯云 账号信息 > **登录方式**中关联对应类别。

：表示该用户在该方式下通知未发送成功

离线节点高级参数

最近更新时间：2024-04-02 16:17:11

参数说明

离线类型	读/写	配置内容	适用场景	描述
es	写		单表	-
mysql	读	splitFactor=5	单表	-
cos	写	splitFileSize=134217728	单表	• 单文件切分大小 • 针对 hive on cos 不生效 • 支持 text、orc、parquet 类型的文件
HDFS	写	splitFileSize=134217728	单表	• 单文件切分大小 • hive on hdfs 不生效 • 支持 text、orc、parquet 类型的文件
hive	写	compress=none/snappy/lz4/bzip2/gzip/deflate	单表	默认为 none。只对 textfile 格式有效，对 orc/parquet 无效（orc/parquet 需要在建表语句指定压缩）
hive	写	format=orc/parquet	单表	hdfs 临时文件的格式，默认为 orc，跟最终 hive 表格式无关
doris	写	sameNameWildcardColumn=true	单表	mysql-doris 配置* 支持同名字段映射

元数据字段	读/写	配置内容

kafka	读	• __key__ 表示消息的 key • __value__ 表示消息的完整内容 • __partition__ 表示当前消息所在分区 • __headers__ 表示当前消息 headers 信息 • __offset__ 表示当前消息的偏移量 • __timestamp__ 表示当前消息的时间戳

配置方式


数据来源
数据源类型
MySQL
数据源
mysql_master_menghuiyu
新建数据源
库
请选择库
表
shanyutest
共1个
添加分库分表
表主键
id
格式
utf-8
读取模式
全量
增量
过滤操作
插入
更新
删除
高级设置
参数
请输入参数名称及值（格式为：parameter=value），多个参数使用换行符分割
如splitFactor=xxx
搜索参数、参数说明
scan.incremental.snapshot.chunk.size=20000
添加
split-key.even-distribution.factor.upper-bound=10.0d
添加
scan.newly-added-table.enabled=true
添加
1. 点击【添加】快速配置参数
2. 展开可查看参数说明1

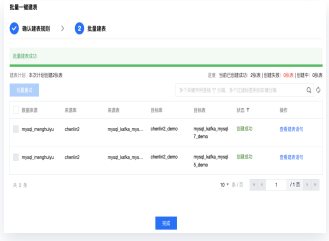
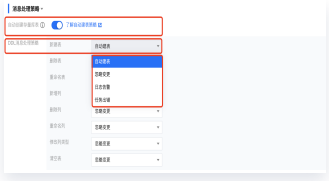
同步任务自动建表能力

最近更新时间：2024-03-28 17:06:01

数据集成在离线及实时场景下提供一键建表、自动建表等功能，以帮助您在同步前或同步中快速构建目标数据表，从而提高数据链路使用效率。本文主要介绍对应功能使用方式、异构数据源间类型转换关系等。

功能介绍及使用场景

数据集成提供了一键建表、批量建表、以及根据 DDL 消息自动建表三种建表方式：

功能项	功能示意	功能介绍	适用场景
一键建表		此功能可根据来源端指定的多个表对象自动完成来源至目标端的 DDL 转换，支持用户手动修改转换后的 DDL 内容，并一次性批量创建。**功能关键词：异构 DDL 自动转换、批量建表、手动建表。**	本功能适合于针对来源端的存量，快速构建与来源表结构近似的目标表。针对表模型复杂，需要业务自定义的情况，用户可根据业务特性编辑自动生成的 DDL 语句，提高异构 DDL 编写效率。
批量建表			
根据源端 DDL消息自动建表		此功能可实时监控来源端是否存在新建表 DDL，一旦源端产生可识别新建表消息，目标端可自动根据来源表结构自动生成目标端表对象。**功能关键词：DDL 自动感知与响应、自动建表。**	本功能适合在实时整库搬迁时，期望目标端与来源端实时保证结构一致的数据同步场景。

异构数据源建表类型转换关系：

MySQL 至 Doris 建表字段类型转换：

	MySQL（源端）	DORIS（目标端）	补充说明

	数据类型	精度说明	数据类型	精度说明	
	BOOLEAN	0代表 false，1代表 true	BOOLEAN	0代表 false，1代表 true	-
数值类型	TINYINT	范围[-128, 127]	TINYINT	范围[-128, 127]	-
	SMALLINT	范围[-32768, 32767]	SMALLINT	范围[-32768, 32767]	-
	MEDIUMINT	范围[-8388608，8388607]	INT	范围[-2147483648, 2147483647]	-
	INT	范围[-2147483648, 2147483647]	INT	范围[-2147483648, 2147483647]	-
	BIGINT	范围[-9223372036854775808, 9223372036854775807]	BIGINT	范围[-9223372036854775808, 9223372036854775807]	-
	UNSIGNED TINYINT	范围[0, 255]	SMALLINT	范围[-32768, 32767]	-
	UNSIGNED MEDIUMINT	范围[0, 16777215]	INT	范围[-2147483648, 2147483647]	-
	UNSIGNED INT	范围[0, 4294967295]	BIGINT	范围[-9223372036854775808, 9223372036854775807]	-
	UNSIGNED BIGINT	范围[0, 18,446,744,073,709,551,615]	LARGEINT	范围[-2^127 + 1 ~ 2^127 - 1]	-
	FLOAT	4 字节浮点数	FLOAT	4字节浮点数	-
	DOUBLE	8 字节浮点数	DOUBLE	8字节浮点数	-

	DECIMAL	DECIMAL(M,D)，M 范围 [1, 65]， D 范围是[0, 30]	DECIMALV3	DECIMAL(M,D)，M 范围 [1, 38]， D 范围是 [0, precision]	-
日期时间类型	YEAR	范围：1901 到 2155显示格式：YYYY	SMALLINT	范围[-32768, 32767]	-
	TIME	范围：-838:59:59 到 838:59:59显示格式：hh:mm:ss 或 hh:mm:ss	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	-
	DATE	范围：1000-01-01 到 9999-12-31显示格式：YYYY-MM-DD	DATEV2	范围：0000-01-01 到 9999-12-31显示格式：YYYY-MM-DD	-
	DATETIME	1000-01-01 00:00:00 到 9999-12-31 23:59:59显示格式：YYYY-MM-DD HH:mm:ss	DATETIMEV2	0000-01-01 00:00:00 到 9999-12-31 23:59:59打印格式：YYYY-MM-dd HH:mm:ss.SSSSSS，可不选时间精度。	-
	TIMESTAMP	UTC 1970-01-01 00:00:01 到 2038-01-19 03:14:07显示格式：YYYY-MM-DD HH:mm:ss	DATETIMEV2	显示格式：YYYY-MM-DD HH:mm:ss	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	0 到 255 字符	CHAR	定长字符串，范围是1-255	-
	VARCHAR	0 到 65,535 字符	VARCHAR	变长字符串，范围是1-65533	如果 MySQL 字段长度超过65533，建议转化为string
	TINYTEXT、TEXT	0 到 255 字符	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	-
	MEDIUMTEXT、	0 到 65535 字符	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	MySQL 字段长度超过1048576

	LONGTEXT				字节时可能精度丢失
二进制字符串	TINYBLOB、BLOB	二进制字符串，0 到 255 字节	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	-
	MEDIUMBLOB、LONGBLOB	二进制字符串，0 到 16,777,215 字节，最大16M	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	MySQL 字段长度超过1048576字节时可能精度丢失
	BINARY、VARBINARY	固定长度二进制数据，最多 255 字节	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	-
其他	JSON	JSON 数据，最大存储大小为 1GB	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	MySQL 字段大小超过1M时可能精度丢失
	SET、BIT	字符串集合，最多 64 个成员	STRING	变长字符串，最大（默认）支持1048576 字节（1MB）	-
	ENUM	枚举对象，最多65535个成员	UNSUPPORTED	-	暂不支持

MySQL 至 DLC iceberg表建表字段类型转换：

Mysql 类型（源端）	DLC Iceberg 表（目标端）	说明
tinyint(1)	int	-
smallint	smallint	-
int	int	-
mediumint	int	-
bigint	bigint	-
float	float	-
double	double	-

decimal	decimal	-
datetime	timestamp	-
timestamp	timestamp	-
date	date	-
time	time	-
tinytext	string	-
text	string	-
mediumtext	string	-
longtext	string	-
varchar	string	-
char	string	-
bool	boolean	-
tinyblob	binary	-
mediumblob	binary	-
blob	binary	-
longblob	binary	-
varbinary	binary	-
binary	binary	-
decimal unsigned(p,x)	decimal(p+1, x)	说明：decimal(p+1, x)/string(超长后改为string)
decimal unsigned(p,x)	decimal(p+1, x)	说明：decimal(p+1, x)/string(超长后改为string)
int unsigned	bigint	-
int unsigned zerofill	bigint	-
smallint unsigned	int	-

smallint unsigned zerofill	int	-
mediumint unsigned	bigint	-
mediumint unsigned zerofill	bigint	-
float unsigned	double	-
float unsigned zerofill	double	-
double unsigned	decimal(20,0)	-
double unsigned zerofill	decimal(20,0)	-
bigint unsigned	decimal(20,0)	-