



PieDataCS 数据
集成使用文档draft

内容

1 数据集成简介	4
核心特性	4
前期准备	5
离线同步任务使用流程	5
实时同步任务使用流程	6
2 授予用户数据集成功能权限	6
3 信息总览	7
Source 维度	8
Table 维度	9
4 配置调度系统	11
添加 S3 配置	11
添加调度器配置	12
5 创建数据源	13
创建数据源并配置连接	13
管理数据源	16
6 配置表的映射关系	17
同步表：从真实数据源获取表信息并与之关联	17
更新表：与真实数据源中数据表的结构做同步更新	18
关联表：将通用表模型关联到当前数据源	19
解除关联：将当前数据源与指定表解除关联	20
查看字段信息	20
7 创建和执行 DDL 任务	20
8 创建离线同步任务	22
创建表到表的离线同步任务	23
创建库到库的离线同步任务	27
离线同步任务运行配置项说明	30
离线同步任务失败问题排查	31
清理源数据	32
清理导出文件	34
9 创建实时同步任务	35
创建表到表的实时同步任务	35
创建库到库的实时同步任务	39
查看实时同步任务详情	42
实时同步任务运行配置项说明	44
实时同步常见报错与解决方案	45
10 管理元数据	47

关于通用表模型.....	47
创建一个新的通用表模型.....	48
从数据源导入一个通用表模型.....	49
管理通用表模型的字段.....	51

draft

1 数据集成简介

PieCloudDB 的数据集成（Dataflow）功能支持从不同的数据库抽取数据，并将其加载到同构或异构的目标数据库中。这一数据迁移任务管理服务依托于 PieCloudDB 管控平台实现。用户只需在平台上配置数据源、数据流和作业调度，即可实现数据全量迁移与备份的可视化和精细化管理。

数据集成功能主要适用于将多源数据库的历史数据进行全量或者增量迁移至中央历史数仓，并进行汇聚处理的场景。

在当前版本中，数据集成支持两种数据迁移任务类型：离线同步和实时同步。对于非周期性作业，系统会创建一个单次任务，这可以是离线同步任务或实时同步任务；对于周期性作业，系统会创建一个定期执行的任务并按照预设周期运行。但请注意，周期性作业目前仅适用于表到表的离线同步任务。

注意：

数据集成功能属于 PieCloudDB 企业版的定制化功能，需要进行独立部署。有关部署方式，请联系 OpenPie 技术支持团队。

核心特性

数据集成的核心特性包括但不限于：

- 支持多种数据源

能够从多种数据源抽取数据并加载到目标数据库，实现了对异构数据源（例如 MySQL、PostgreSQL 等）的自动化配置。同时，提供目标数据库 Schema 创建语法的自动化支持，并内置及自定义字段改写规则。

- 支持多种内置校验算法

在数据迁移过程中，提供数据一致性校验，包括全量目标数据表校验和单次作业范围校验，确保数据的全面和高效准确性。

- 任务调度精细化管理

用户可以按需选择单表、多表或全库的方式进行历史数据迁移，并支持灵活的单次或周期性迁移作业策略。

- 多线程高效运行

数据集成过程采用多线程技术，高效处理数据的导出、压缩、导入以及数据校验等任务，显著提升数据处理效率。

- 可视化监控

基于 PieCloudDB 管控平台，实现数据集成流程的可视化监控。例如，实时展示数据导出和导入的进度条；通过数据集成总览面板全局查看数据集成任务的状态。

- 国产化部署

支持部署在 x86 和 ARM CPU 架构的服务器，并适配主流国产操作系统，例如 Kylin OS 等。

前期准备

在执行数据集成之前，请确保完成以下准备工作：

- 源数据库准备工作

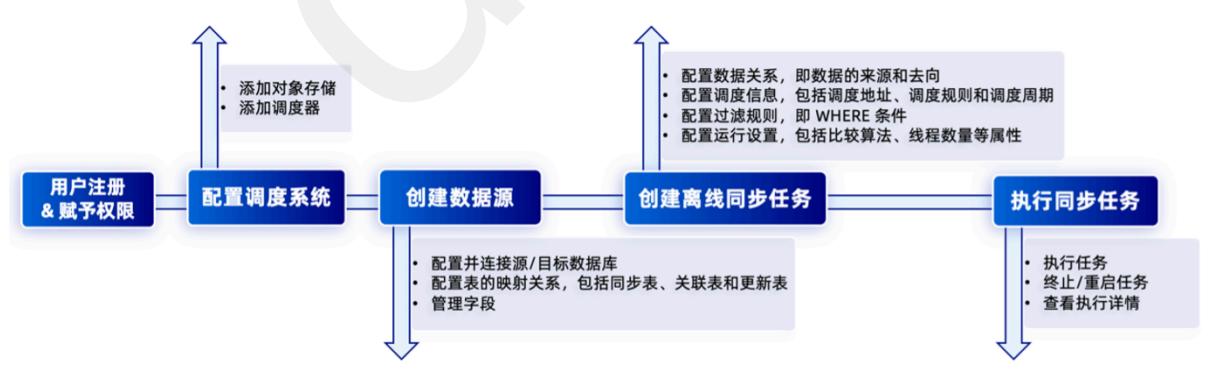
确保源数据库提供了一个具有所需迁移库表读取权限的用户。此外，该用户需要能够从运行数据集成的服务器或容器访问源数据库。

- 目标数据库准备工作

预先在目标数据库中创建好需要导入数据的表，并确保有一个用户具备对这些表的写入权限。同样，该用户需要能够从运行数据集成的服务器或容器访问目标数据库。

离线同步任务使用流程

数据集功能支持创建离线同步任务，适用于批处理数据的场景。对于离线同步任务，建议使用流程如下：

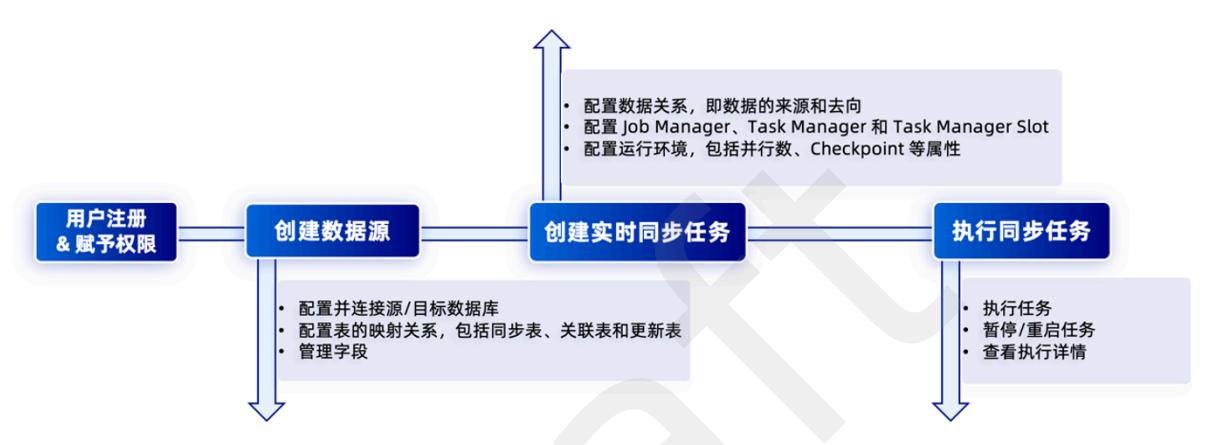


- 授予用户数据集功能权限
- 配置调度系统
- 创建源数据源并配置连接
- 配置源数据源的表的映射关系

5. 创建和执行 DDL 任务为目标数据源创建新的表
6. 配置目标数据源的表的映射关系
7. 创建离线同步任务

实时同步任务使用流程

数据集成功能支持创建实时同步任务，适用于流处理数据的场景。对于实时同步任务，建议使用流程如下：



1. 授予用户数据集成功能权限
2. 创建源数据源并配置连接
3. 配置源数据源的表的映射关系
4. 创建和执行 DDL 任务为目标数据源创建新的表
5. 配置目标数据源的表的映射关系
6. 创建实时同步任务

2 授予用户数据集成功能权限

数据集成功能具有两类权限：只读和管理，属于系统权限范畴。具有“只读”权限的用户仅能在总览、数据源、数据流和作业调度页面执行一些查询操作；除了包含只读权限所开放的功能外，具有“管理”权限的用户还可以对各个页面上的功能执行创建、删除和修改等操作。

数仓预设角色中，accountmanager 角色作为数仓管理员，默认同时具有数据集成功能的管理权限（权限名称为“数据集成功能管理”）和只读权限（权限名称为“数据集成功能只读”）。

本文通过具体示例来介绍如何在 PieDataCS 云原生平台对一位新的数仓用户授予数据集成功能的管理权限。

- 以管理员身份登录云原生平台并创建自定义角色（例如 dataflow-manager），并将数据集成的管理权限“数据集成功能管理”授予该角色。



如果当前其他角色已具有“数据集成功能管理”的权限，在创建自定义角色时也可以通过直接继承角色的方式获得该权限。

有关创建自定义角色的详细信息，请参见 [管理计算空间角色-添加自定义角色](#)。

- 以管理员身份在「**用户**」页面创建一个数仓用户，例如 dataflow-user。

有关新建用户的详细信息，请参见 [管理计算空间用户-新建用户](#)。

- 在「**角色**」页面，以管理员身份为新数仓用户 dataflow-user 赋予步骤 1 中所创建的角色 dataflow-manager，则该用户会继承该角色的数据集成功能管理权限。

相关操作信息请参见 [管理计算空间角色-授予角色给用户](#)。

- 以步骤 3 所创建的数仓用户 dataflow-user 登录 PieDataCS 云原生平台，即可开始使用数据集成的功能。

3 信息总览

数据集成的「**总览**」页面展示指定时间内，离线同步的数据导入和导出的历史统计信息，包括执行次数、导入记录数、平均执行时间和总执行时间。

该页面上方提供时间筛选控件，以方便用户根据时间筛选信息：

- 使用**时间范围**控件可以筛选指定起止时间范围内的数据导入和导出的信息，单位为天。
- 使用**看本周**控件可以筛选显示本周内的数据导入和导出的相关信息。
- 使用**看本月**控件可以筛选显示本月内的数据导入和导出的相关信息。

Source 维度

在「总览」页面，**Source 维度** 基于数据源的维度展示数据集成的统计信息。

数据源	操作类型	↑ 执行次数	记录数	平均执行时间(秒)	总执行时间(秒)	操作
pdb_pg_nodeport	数据导入	1	13	1.76	1.76	查看 table 统计
pdb_mysql	数据导入	1	119,994,608	895.04	895.04	查看 table 统计
pdb_pg	数据导入	20	84	1.89	37.77	查看 table 统计
pg12	数据导出	23	112	0.39	9.07	查看 table 统计

Source 维度的信息列表显示如下信息：

- 数据源：数据集成所使用的数据源的名称。
- 操作类型：数据集成操作类型，包括“数据导入”和“数据导出”。
- 执行次数：数据集成操作被执行的次数。

- 记录数：数据集成操作所输出数据的行数。
- 平均执行时间(秒)：该数据源的数据集成操作的平均执行时间，该值的计算公式为“总执行时间/执行次数”。单位为秒。
- 总执行时间(秒)：该数据源的数据集成操作的总执行时间。单位为秒。
- 操作：相关扩展操作。当前版本支持“查看table统计”，点击该按钮即可跳转至当前数据源所对应的表维度页面，详细信息请参见本章节 **Table 维度** 的内容。

表名	所属数据源	操作类型	↑ 执行次数	记录数	平均执行时间(秒)	总执行时间(秒)
tb	pdb_pg_nodeport	数据导入	1	13	1.76	1.76

Source 维度的信息列表提供如下快捷操作：

- 根据操作类型筛选信息：使用**操作类型**控件可以筛选指定操作类型的数据集成信息。默认展示全部操作类型的信息。
- 指定时间范围筛选信息：使用**开始时间-结束时间**控件可以筛选显示某个时间范围（单位为天）内的信息。
- 根据数据源名称搜索信息：使用**输入数据源名称搜索**的搜索框可以通过数据源的名称搜索相关信息。
- 按照执行次数排序：点击表头的“执行次数”字段，可以按照该操作的执行次数的大小顺序显示列表信息。
- 按照记录数排序：点击表头的“记录数”字段，可以按照该操作所执行的记录数的大小顺序显示列表信息。
- 按照平均执行时间排序：点击表头的“平均执行时间”字段，可以按照该操作的平均执行时间的大小顺序显示列表信息。
- 按照总执行时间排序：点击表头的“总执行时间”字段，可以按照该操作的总执行时间的大小顺序显示列表信息。
- 立即刷新：点击立即刷新图标可以即刻同步当前的列表信息为最新。

Table 维度

在「总览」页面，**Table 维度** 基于数据源中表的维度展示数据集成的统计信息。

表名	所属数据源	操作类型	↑ 执行次数	记录数	平均执行时间(秒)	总执行时间(秒)
part	pdb_oracle	数据导入	1	2,000,000	67.73	67.73
supplier	pdb_oracle	数据导入	1	100,000	5.10	5.10
customer	pdb_oracle	数据导入	1	1,500,000	68.58	68.58
orders	pdb_oracle	数据导入	1	15,000,000	229.77	229.77

Table 维度的信息列表显示如下字段：

- 表名：表的名称。
- 所属数据源：该表所属的数据源。
- 操作类型：数据集成的操作类型，包括“数据导入”和“数据导出”。
- 执行次数：该表被执行的次数。
- 记录数：数据集成操作所流转的记录数量。
- 平均执行时间(秒)：该表的数据集成操作的平均执行时间，该值的计算公式为“总执行时间/执行次数”。单位为秒。
- 总执行时间(秒)：该表的数据集成操作的总执行时间。单位为秒。

Table 维度的信息列表提供如下快捷操作：

- 根据操作类型筛选信息：使用**操作类型**控件可以筛选指定操作类型的数据集成信息。默认展示全部操作类型的信息。
- 指定时间范围筛选信息：使用**开始时间-结束时间**控件可以筛选显示某个时间范围（单位为天）内的信息。
- 根据表名称搜索信息：使用**输入表名搜索**的搜索框可以通过表的名称搜索相关信息。
- 根据数据源名称搜索信息：使用**输入数据源名称搜索**的搜索框可以通过数据源的名称搜索相关信息。
- 按照执行次数排序：点击表头的“执行次数”字段，可以按照该表被执行次数的大小顺序显示列表信息。
- 按照记录数排序：点击表头的“记录数”字段，可以按照该表所执行的记录数的大小顺序显示列表信息。
- 按照平均执行时间排序：点击表头的“平均执行时间”字段，可以按照该表的平均执行时间的大小顺序显示列表信息。
- 按照总执行时间排序：点击表头的“总执行时间”字段，可以按照该表的总执行时间的大小顺序显示列表信息。
- 立即刷新：点击立即刷新图标可以即刻同步当前的列表信息为最新。

4 配置调度系统

系统配置涵盖 S3 配置和调度器设置。在数据的离线同步过程中，调度器（dataflow-worker）负责管理数据的导出、导入和校验任务。数据源中导出的数据将被保存到 S3，作为备份存储。

提示：

S3 配置和调度器配置仅适用于批处理数据的场景，该配置信息会应用于创建离线同步任务。

数据集成的「总览」页面提供了一个进入系统配置的入口。用户在该页面点击 **系统配置** 即可进入系统配置页面。

添加 S3 配置

S3 对象存储作为数据流转的中间站，负责存储从不同的数据源抽取的数据，然后加载数据到目标数据库（例如 PieCloudDB）。

添加 S3 配置的步骤如下：

1. 在系统配置的「**S3 配置**」页面，点击 **添加 S3 配置** 即可进入创建 S3 配置的页面。
2. 在创建 S3 配置页面，输入如下 S3 配置信息：
 - 名称：S3 配置的自定义名称。
 - Endpoint：与 S3 服务通信的终端 URL 地址。
 - Region：S3 服务所在的区域，每个区域都有一个唯一的标识符。
 - Bucket Name：存放备份文件的 S3 Bucket 名称，需要提前创建。
 - S3 Access Key：访问密钥。
 - S3 Secret Key：私有访问密钥。
3. 点击 **完成**。操作成功后，新添加的 S3 配置会自动同步到列表中。

系统配置的「**S3 配置**」页面以列表形式展示当前已有的 S3 配置信息。该页面还支持添加 S3 配置，以及对 S3 配置的修改和删除操作。

The screenshot shows a table with one row of data. The columns are: ID, Name, Endpoint, Region, Bucket Name, S3 Access Key, S3 Secret Key, and Set as Default Configuration. The data row contains: gxtan_S3, http://[REDACTED], gxtan-type-demo, minioadmin, a toggle switch (which is turned on), and three vertical dots. Above the table is a navigation bar with tabs for 'S3 Configuration' and 'Scheduler Configuration'. Below the table is a blue button labeled 'Add S3 Configuration'.

ID	名称	Endpoint	Region	Bucket Name	S3 Access Key	S3 Secret Key	设为默认配置
gxtan_S3	http://[REDACTED]			gxtan-type-demo	minioadmin	<input checked="" type="checkbox"/>	⋮

针对于添加后的 S3 配置，用户可以执行如下操作：

- 修改/删除：指定某条 S3 配置，用户点击隐藏的列表并选择对应选项来“修改”或者“删除”该 S3 配置。
- 设为默认配置：如果系统配置了多个 S3 对象存储，第一个 S3 配置会自动成为默认配置。用户可以滑动开关指定默认的 S3 配置。

该设置会影响离线同步任务的运行配置中所涉及的 S3_config_id 字段值，如果在创建离线同步任务时未手动配置 S3_config_id，系统就会使用默认的 S3 配置。有关离线同步任务的详细信息，请参见 [创建离线同步任务](#)。

添加调度器配置

在数据集成过程中，调度器负责发起数据的导出、导入和校验任务。调度系统会根据执行周期、调度环境、运行配置等因素，对数据的离线同步任务进行协调和管理，这称为作业调度。

添加调度器配置的步骤如下：

- 在系统配置的「调度器配置」页面，点击 **添加调度器** 即可进入创建调度器的页面。
- 在创建调度器配置页面，输入如下调度器信息：
 - 名称：调度器的自定义名称。
 - 主机：Piescheduler 所在宿主机的 IP 地址。
 - 端口：Piescheduler 的 nodePort 暴露出的端口号。
 - Workflow Names：工作流的名称。
 - Task Queues：任务队列的名称。

提示：

基于客户真实环境中所部署的 LoadBalancer，主机和端口号可以为 LoadBalancer 的 IP 和其所暴露的端口号。

- 点击 **完成**。操作成功后，新添加的调度器配置会自动同步到列表中。

系统配置的「调度器配置」页面以列表形式展示当前已有的调度器配置信息。在目标调度器配置所在行，启用默认配置开关，可以将目标调度器设置为默认调度器。

端口	Workflow Names	Task Queues	设为默认配置	创建时间	更新时间	更多操作
PyFileBatchWorkflow	PYTHON_FILE_BATCH_QUEUE	<input checked="" type="checkbox"/>	2024-05-21 09:36:28	2025-01-16 10:34:02	⋮	
PyFileBatchWorkflow,FlinkFileBatchWorkflow	PYTHON_FILE_BATCH_QUEUE,FLINK_FILE_BATCH_QUEUE	<input type="checkbox"/>	2024-05-07 10:33:03	2024-05-07 10:33:03	⋮	

在列表中，用户点击目标调度器配置所在行的隐藏的下拉列表并选择对应选项来 **修改** 或者 **删除** 该调度器配置。

5 创建数据源

数据集功能支持与多种不同数据源的连接，当前版本支持连接到 PostgreSQL、PieCloudDB TP、PieCloudDB XP、PieCloudDB、MySQL 以及 Oracle 等数据源类型。用户通过配置相应的连接参数并进行连通性测试后，即可成功连接到数据源，为数据集成做好准备。

创建数据源并配置连接

用户必须首先创建数据源，然后配置相应的连接参数，之后才能执行数据集成操作。

在数据集成的「数据源」页面，创建数据源并配置连接的步骤如下：

1. 点击 **创建数据源** 即可进入创建数据源的页面。
2. 在“类型”下拉列表中选择一种数据源类型。当前版本支持 Oracle、MySQL、PostgreSQL、PieCloudDB TP、PieCloudDB XP 和 PieCloudDB。
3. 输入数据源名称和对该数据源的描述（选填）。
4. 点击 **完成**。操作成功后，新创建的数据源会自动同步到列表中。
5. 在数据源列表中的新创建数据源的“操作”栏下，点击 **连接配置** 即可进入配置页面。
6. 在数据源连接配置页面，分别输入所选数据源的连接信息（各数据源的连接信息参考下表），并点击 **完成** 以确认更改。

字段	含义	适用的数据源类型
host (必填)	要连接数据源的主机 IP 地址	<ul style="list-style-type: none"> Oracle MySQL PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
port (必填)	要连接数据源的主机端口	<ul style="list-style-type: none"> Oracle MySQL PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
database (必填)	要连接数据源的目标数据库的名称	<ul style="list-style-type: none"> MySQL PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
user (必填)	要连接数据源的数据库的用户名	<ul style="list-style-type: none"> Oracle MySQL PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
password (必填)	要连接数据源的数据库的密码	<ul style="list-style-type: none"> Oracle MySQL PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
warehouse (选填)	如果连接的外部数据源为 PieCloudDB 虚拟数仓且使用 PieProxy 外部接入方式，则须填入外部数据源的虚拟数仓 ID 信息	PieCloudDB
schema (选填)	要连接数据源的目标 Schema 名称	<ul style="list-style-type: none"> PostgreSQL PieCloudDB TP PieCloudDB XP PieCloudDB
servertimezone (选填)	要连接数据源的服务器所在的时区	MySQL
service (必填)	要连接数据源的目标数据库服务	Oracle

字段	含义	适用的数据源类型
sid (必填)	要连接数据源的目标数据库实例的系统标识符	Oracle
connparam (选填)	连接参数	Oracle
extparam (选填)	外部参数	<ul style="list-style-type: none"> • PostgreSQL • PieCloudDB TP • PieCloudDB XP • PieCloudDB

7. 点击 **连通性测试**，如果显示“连通性测试成功”的信息，则说明已与所配置的数据源建立了连接。下图以 PieCloudDB 数据源为例。



8. 点击页面的图标“X”，即可退出配置页面并返回到数据源主页面。

提示：

如果需要修改一个数据源的连接配置，请重复执行上述步骤 5 ~ 8。

创建数据源之后，还需要配置表的映射关系，详细信息请参见 [配置表的映射关系](#)。

管理数据源

在数据源列表中，用户可以对目标数据源进行修改（仅限于修改数据源的名称和描述，数据源类型无法更改）和删除操作。这些功能都可以通过数据源列表每行旁边的隐藏菜单「...」来快速访问。

用户也可以点击目标数据源名称以进入数据源详情页面，该页面支持执行修改数据源的基本信息和配置信息等操作。



The screenshot shows a table of data sources with the following columns: Data Source Name, Data Source Type, Description, Creation Time, and Operations. The operations column contains links for 'Connection Configuration' and 'Associated Tables' along with a three-dot menu icon.

数据源名称	数据源类型	描述	创建时间	操作
pdb_mysql	pieclouddb		2024年11月20日星期三 14:11	连接配置 关联的表 ...
pdb_pg_nodeport	pieclouddb		2024年11月19日星期二 15:11	连接配置 关联的表 ...
pg_test	postgres		2024年11月15日星期五 11:15	连接配置 关联的表 ...
pdb_up_low	pieclouddb		2024年11月14日星期四 19:48	连接配置 关联的表 ...
mysql_up_low	mysql		2024年11月14日星期四 19:46	连接配置 关联的表 ...
pdb_oracle	pieclouddb		2024年11月14日星期四 17:07	连接配置 关联的表 ...
oracle_stage	oracle		2024年11月14日星期四 17:03	连接配置 关联的表 ...

数据源信息列表提供如下快捷操作：

- 根据状态筛选信息：使用**状态**控件可以筛选指定状态的数据源信息，包括“正常”和“回收站”两种状态。默认展示“正常”状态的数据源信息。当数据源被删除后，它将被移至回收站。

如果数据源被删除，则会被标记为“回收站”状态，点击**恢复**快捷键即可将其返回到“正常”状态。



The screenshot shows a table of data sources with the following columns: Data Source Name, Data Source Type, Description, Creation Time, and Operations. The operations column contains links for 'Connection Configuration' and 'Associated Tables' along with a three-dot menu icon. A red box highlights the '恢复' (Restore) button next to the 'demo' entry.

数据源名称	数据源类型	描述	创建时间	操作
demo	oracle		2024年11月21日星期四 15:27	连接配置 关联的表 ...
pdb_mysql2	mysql		2024年11月20日星期三 14:10	连接配置 关联的表 恢复

- 根据数据源名称搜索信息：使用**输入名称查询**的搜索框可以基于数据源的名称搜索相关信息。

- 根据描述搜索信息：使用**输入描述查询**的搜索框可以基于数据源的描述信息搜索相关信息。

6 配置表的映射关系

数据源在初始创建状态下不包含任何表，因此需要配置表的映射关系。

在数据集成的「**数据源**」页面，用户点击目标数据源所在行的“操作”栏下的**关联的表**选项，即可进入目标数据源的详情页面。

The screenshot shows the 'Data Source / Data Source' details page. At the top, there are filters for 'Status: Normal', 'Input Name Search', and 'Input Description Search'. Below the header, a table lists data sources. One row is selected for 'pdbtpt', which has a 'local' type, was created on '2024-12-16星期一 11:15', and has 'Connection Configuration' and 'Associated Tables' operations. The 'Associated Tables' button is highlighted with a red box.

在数据源详情页面中，“**关联的表**” 区域会展示与当前数据源相关联的所有通用表（如果存在），并提供配置表映射关系的相关操作功能。

The screenshot shows the 'Associated Tables' details page. It includes search bars for 'Input schema search' and 'Input table name search'. A table lists associated tables with columns: 'Table Name', 'Schema', 'Title', 'Table name case sensitivity', and 'Operations'. Two entries are shown: '交易数据_2020_2023' (schema: 服装销售数据, title: 服装销售数据.交易数据_2020_2023, case sensitivity: No, operations: Remove Association, View Fields) and '顾客数据_2020_2023' (schema: 顾客数据, title: 顾客数据.顾客数据_2020_2023, case sensitivity: No, operations: Remove Association, View Fields). Buttons for 'C', 'Associate Table', 'Sync Table', and 'Update Table' are at the top right.

提示：

在数据集成的「**数据源**」页面，用户也可以点击目标数据源名称来进入数据源详情页面。

同步表：从真实数据源获取表信息并与之关联

如果需要从当前实际的数据源中获取表信息并建立关联，可以使用同步表功能，该功能支持整个数据库的同步以及部分表的同步。从当前数据源获取的全量或者增量的表信息会被自动维护到通用表模型中，以方便其他数据源与之关联。

同步表的操作步骤如下：

- 在数据源详情页面中的“**关联的表**”区域，点击**同步表**即可弹出关联窗口。
- 选择同步数据源中表的方式。

- 如果需要与当前数据源中全量表相关联，则勾选 **整库同步**。
 - 如果需要与当前数据源中指定表相关联，则勾选 **选择部分表进行同步**，之后在“要同步的表”下拉列表中选择需要同步的表（至少选择一个表）。
3. 点击 **完成**。如果关联成功，相关同步信息会显示在同步结果的窗口中。



4. 点击 **X**，即可关闭同步表的窗口。

上述操作完成后，目标数据源即可与从当前数据源所选的表相关联并将其同步到列表中。

更新表：与真实数据源中数据表的结构做同步更新

从真实数据源获取表信息并同步表的操作仅涉及增量同步表的信息。如果数据源的表结构发生变化，则需要使用更新表的功能来与真实数据源的数据表结构进行同步更新，这一操作支持全量更新和部分更新。

与真实数据源的指定表的结构做同步更新后，最新的表信息会被自动维护到通用表模型，用于构建数据同步任务。

更新表的操作步骤如下：

- 在数据源详情页面中的“关联的表”区域，点击 **更新表** 即可弹出更新窗口。
- 选择同步更新数据源中表的方式。
 - 如果需要与当前数据源中全量表的结构做同步更新，则勾选 **更新所有表**。
 - 如果需要与当前数据源中指定表的结构做同步更新，则勾选 **选择部分表进行更新**，之后在“要更新的表”下拉列表中选择要同步结构的表（至少选择一个表）。
- 点击 **更新**。如果操作成功，相关更新信息会显示在更新结果的窗口中。



4. 点击 X，即可关闭更新表的窗口。

关联表：将通用表模型关联到当前数据源

数据源在初始创建状态下，通常不与任何通用表模型关联。如果数据集成模块中存在可用的通用表模型，用户可以点击 **关联表**，并从下拉框中选择要关联的表。



然后点击 **完成**，即可将选定的通用表与目标数据源关联，并将表信息同步到列表中。

This screenshot shows a table listing associated tables. The 'CUSTOMER' and 'ORDERS' tables are highlighted with red boxes. The table has columns: 表名 (Table Name), schema, 标题 (Title), 表名是否大小写敏感 (Case-sensitive), and 操作 (Operations). The 'CUSTOMER' table is associated with the 'ORDERS' schema, while the 'ORDERS' table is associated with the 'public' schema. Both are case-insensitive.

表名	schema	标题	表名是否大小写敏感	操作
CUSTOMER	CUSTOMER	CUSTOMER	否	解除关联 查看字段
ORDERS	ORDERS	ORDERS	否	解除关联 查看字段
交易数据_2020_2023	服装销售数据	服装销售数据.交易数据_2020_2023	否	解除关联 查看字段
顾客数据_2020_2023	顾客数据	顾客数据.顾客数据_2020_2023	否	解除关联 查看字段

解除关联：将当前数据源与指定表解除关联

如果需要将当前数据源与指定表解除关联，在目标表所在行的“操作”列中，点击**解除关联**，确认后即可该表与当前数据源解除关联，同时该表也会从关联列表中移除。



关联的表 查看全部数据表		操作		
表名	schema	标题	表名是否大小写敏感	操作
CUSTOMER		CUSTOMER	否	解除关联 查看字段
ORDERS		ORDERS	否	解除关联 查看字段
交易数据_2020_2023	服装销售数据	服装销售数据.交易数据_2020_2023	否	解除关联 查看字段
顾客数据_2020_2023	顾客数据	顾客数据.顾客数据_2020_2023	否	解除关联 查看字段

查看字段信息

在目标数据源详情页面的“关联的表”列表中，在目标表所在行的“操作”列中，点击**查看字段**，即可查看该表的所有字段信息。



关联的表 查看全部数据表		操作		
表名	schema	标题	表名是否大小写敏感	操作
CUSTOMER		CUSTOMER	否	解除关联 查看字段
ORDERS		ORDERS	否	解除关联 查看字段
交易数据_2020_2023	服装销售数据	服装销售数据.交易数据_2020_2023	否	解除关联 查看字段
顾客数据_2020_2023	顾客数据	顾客数据.顾客数据_2020_2023	否	解除关联 查看字段

7 创建和执行 DDL 任务

在数据仓库或数据库迁移过程中，DDL 任务可以将源数据源的表结构迁移到目标数据源下并创建新的表。

提示：

请确保在创建和执行 DDL 任务之前，源数据源的连接配置和表的映射关系均已正确设置。

在数据集成的「**数据源**」页面，点击**DDL 任务**即可进入相应的功能页面。

在「**DDL 任务**」页面，创建 DDL 任务的步骤如下：

1. 点击**创建任务**即可进入创建任务页面。

2. 在“源数据源”下拉列表中选择源数据源，并根据实际需要填写 Schema。如果没有指定 Schema，则系统会迁移源数据源的默认 Schema 下的表。

注意：

对于源数据源，在多个 Schema 的场景下建议指定 Schema。

3. 在“目标数据源”下拉列表中选择目标数据源，并根据实际需要填写 Schema。如果没有指定 Schema，则会保存在目标数据源的默认 Schema 下。

注意：

对于目标数据源，在多个 Schema 的场景下需要指定 Schema。

4. 点击 **添加**，分别添加源数据源中需要迁移的表和需要排除的表。
 5.（可选）设置执行该任务的并行数。默认情况下，使用系统预定义的并行数。
 6.（可选）根据实际使用需求开启相关任务配置。在默认情况下，如下功能选项是关闭的。

- 在提交后立刻执行：如果没有开启该选项，则 DDL 任务在创建后为“初始”状态并不会执行。
- 若表不存在，则创建表：如果在目标数据源中找不到与源数据源同名的表，则系统将在指定的 Schema 下自动创建一个具有相同结构的表。
- 若表已存在，则删除表重新创建：如果在目标数据源中存在与源数据源同名的表，则系统将删除该表并重新创建一个具有相同结构的表。
- 是否同步索引：在执行 DDL 任务时同步源数据源中表的索引。
- 是否同步主键：在执行 DDL 任务时同步源数据源中表的主键。

7. 点击 **完成**。DDL 任务创建成功后，其信息会同步到列表中。

如果在步骤 6 中开启“在提交后立刻执行”，则该 DDL 任务会在创建成功后就开始执行，否则，需要用户点击操作栏的隐藏菜单「...」下的**执行**选项来执行该任务。

源数据源	源数据源Schema	目标数据源	目标数据源Schema	状态	执行信息	执行时间	表的同步情况	操作
pdb-pdb		pdb-np		● 初始		开始时间： 结束时间：	成功：0 失败：0	查看 ...
oracle_stage		pdb_oracle		● 已完成	finished	开始时间：2024-11-21 19:05:56 结束时间：2024-11-21 19:06:04	成功：19 失败：0	查看 修改 执行
mysql		pdb_mysql		● 已完成	finished	开始时间：2024-11-20 15:33:06 结束时间：2024-11-20 15:33:11	成功：13 失败：0	查看 修改 删除

在 DDL 任务执行成功后，用户点击目标 DDL 任务所在行的“表的同步情况”栏下的**查看**，即可查看已成功同步的表和执行 SQL 明细等信息，以及同步失败的表信息。

表的同步情况				
表名	状态	执行信息	SQL 明细	
TB	● 成功	success		查看
lower_tb	● 成功	success		查看
UP_TB	● 失败	pg: relation "up_tb" already exists		无
TEST_NUMERIC	● 失败	pg: relation "test_numeric" already exists		无
PT_LIST_TEST	● 成功	success		查看
PART_HASH_T1	● 成功	success		查看
MID_tb	● 成功	success		查看
TRANS_SMT_ATP_T_REPORT	● 失败	pg: relation "trans_smt_atp_t_report" already exists		无
TEST_NUMERIC_2	● 失败	pg: relation "test_numeric_2" already exists		无
PT_RANGE_TEST1	● 成功	success		查看

DDL 任务列表显示已创建的 DDL 任务信息，包括源数据源、目标数据源、表的同步情况等。

数据集成 / 数据源 / DDL 任务							
源数据源	源数据源Schema	目标数据源	目标数据源Schema	状态	执行信息	执行时间	表的同步情况
oracle_stage		pdb_oracle		● 已完成	finished	开始时间: 2024-11-21 19:05:56 结束时间: 2024-11-21 19:06:04	成功: 19 查看 失败: 0 查看
mysql		pdb_mysql		● 已完成	finished	开始时间: 2024-11-20 15:33:06 结束时间: 2024-11-20 15:33:11	成功: 13 查看 失败: 0 查看
mysql		pdb_mysql		● 已完成	finished	开始时间: 2024-11-20 14:06:38 结束时间: 2024-11-20 14:06:42	成功: 12 查看 失败: 0 查看
pg12		pdb_pg		● 已完成	finished	开始时间: 2024-11-20 11:45:43 结束时间: 2024-11-20 11:45:48	成功: 11 查看 失败: 0 查看

用户也可以点击目标 DDL 任务的操作栏的隐藏菜单「...」下的 **修改** 或者 **删除** 选项来执行相应的操作。

需要注意的是，在 DDL 任务执行完成后，目标数据源也需要配置表的映射关系，详细操作信息请参见 [配置表的映射关系-同步表](#)。

8 创建离线同步任务

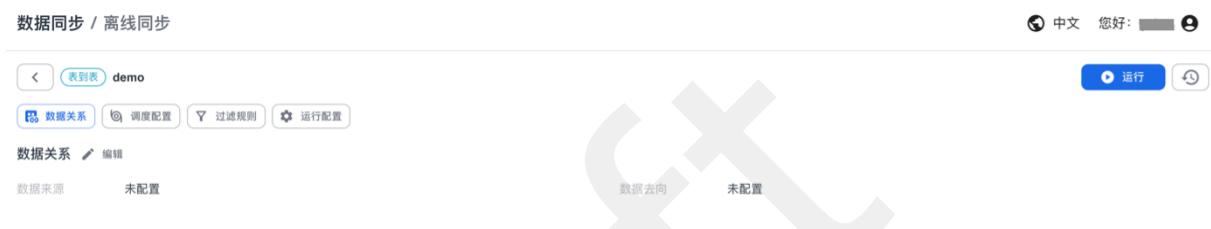
离线同步适用于批处理场景，支持以下类型的数据同步：

- 单表的全量同步：针对单个表进行完整的数据同步。
- 整库批量表全量同步：对整个数据库中的多个表进行批量的全量数据同步。
- 数据库指定表列表的全量同步：根据指定的表列表，对选定的表进行全量数据同步。

创建表到表的离线同步任务

在数据集成的「**数据同步**」页面，用户可以根据业务需求分别设置数据关系、调度配置、过滤规则和运行配置以创建表到表的离线同步任务。参考步骤如下：

1. 切换到「**离线同步**」页面，点击 **新建离线同步** 即可进入创建页面。
2. 输入该离线同步任务的名称。
3. 从“类型”下拉列表中选择离线同步的类型为“表到表”。
- 4.（可选）输入对该任务的描述。
5. 点击 **下一步** 以进入离线同步任务的配置详情页面。



6. 设置“数据关系”以定义数据流向。

在「**数据关系**」页面，首先点击 **编辑**，然后分别选择源数据表和目标数据表。完成选择后，点击 **保存** 以确认更改。

7. 添加调度器配置。

切换到「**调度配置**」页面，点击 **编辑** 并执行如下操作：

1. 在“调度地址”下拉列表中选择运行该数据流的调度器。

有关配置调度器的详细信息，请参见 [配置调度系统-添加调度器配置](#)。

提示：

通常情况下，如果没有特殊环境需求，设置一个调度地址即可。

2. 在“工作流名称”下拉框中选择“PyFileBatchWorkflow”。工作流的作用是将作业任务按照指定顺序组织起来。
3. 在“任务队列”下拉框中选择“PYTHON_FILE_BATCH_QUEUE”。任务队列是轻量级的、动态分配的队列，用于轮询任务。

注意：

系统初始化时应该已经配置了调度器信息。在当前版本的系统中，调度器的工作流名称和任务队列都是固定配置。在选择调度地址后，系统会自动协助用户完成选择。如果在使用过程中发现缺少这类信息，请联系 OpenPie 技术支持团队获取帮助。

4. 设置任务执行周期为“立即执行”或者“周期”。

“立即执行”表示只会创建一个任务并执行一次；“周期”表示会按照周期表达式（Cron 表达式）的设置来创建周期任务并定时执行。

对于周期性的任务，可以通过输入 cron 表达式来定义执行的周期。点击“？”可以显示 cron 表达式的格式说明。例如，“*/60 20 1-30 1-12 6”表示“对于 1 月到 12 月，每月的 1 号到 30 号，每个星期六晚上八点，每隔 60 分钟执行一次作业”。

5. 点击保存**以确认更改。****8. (可选) 添加过滤规则。**

切换到「**过滤规则**」页面，首先点击**编辑**，之后输入 where 条件用于过滤源数据表和目标数据表中的数据。

过滤规则分为如下三类：

- 添加普通条件：需要输入字段名和字段值，并指定逻辑关系。例如“l_quantity > 100”。
- 添加时间范围条件：需要输入字段名和满足时间范围（年/月）的字段值。例如，字段“l_shipdate”满足时间范围过去“12 月”。
- 自定义过滤条件：点击**切换为自定义**，分别输入源表和目标表的自定义 where 表达式（请注意，不要在表达式末尾加分号“;”）。例如，“where update_time > '2024-01-01'"。

在所有过滤条件添加完成后，点击**保存**以确认更改。

9. 修改运行配置。

“运行配置”展示数据同步任务的系统信息。由于离线同步任务需要做数据校验和文件导出的场景，由此运行配置包括比较算法、差异行数量限制、S3 桶信息、导出文件格式等配置项。

切换到「**运行配置**」页面，点击**修改**，并执行如下操作：

- 在运行配置的编辑页面修改 `s3_config_id` 字段以设置 S3 连接配置信息，之后点击 **完成**。对于其他配置项，可以保持它们的默认设置不变。

提示：

对于批处理类型的导出文件，当前版本默认将其存储路径设置为 S3。用户可以通过数据集成的「总览」页面来添加 S3 配置信息。一旦配置完成，用户就可以在运行配置中通过 `s3_config_id` 字段选择对应的 S3 配置 ID。如果需要指定除默认配置之外的存储桶，用户还可以通过配置 `s3_config_bucket` 字段来指定 S3 存储桶，从而覆盖默认的存储桶信息配置。

- 有关运行配置项的详细信息，请参见 [离线同步任务运行配置项说明](#)。
- 运行配置修改完成后，点击 **X** 即可返回运行配置页面。
 - 点击 **测试 S3 连接**，如果配置信息无误，系统会显示“连接测试成功”的信息。
 - 在当前离线同步任务的配置详情页面，点击 **运行** 即可开始运行表到表的离线同步任务。



如果在点击 **运行** 后就发生配置相关报错，用户可以点击 **返回配置** 并根据报错信息修改相应的配置，之后可以重新运行该任务。在任务运行期间，如果需要强制停止运行该任务，点击 **终止** 即可，相应的调度任务也会被同步删除。

提示：

如果终止了一个正在运行的任务，它可能会在数据导出导入过程中的任意点停止。如果需要在终止后重新启动任务，建议先检查目标表，确认数据是否需要被清空，然后重新创建并启动任务。

在表到表的离线同步任务的详情页面，运行结果区域会动态展示离线同步任务在 `ExportTable`、`ImportTable` 和 `TableDataDiff` 各个阶段的执行结果、响应消息、执行时间，以及是否执行成功。



在离线同步任务执行完成后，用户可以点击 **查看调度详情** 来查看执行该任务时详细的调度情况。

数据集成 / 同步详情 / 调度详情 / 1c8773ac-1434-4738-9613-0c4edf94e495

1c8773ac-1434-4738-9613-0c4edf94e495 ~ Running

时间	事件类型	事件描述	更多信息
47 2024-11-25 17:20:18.619	ActivityTaskScheduled	Activity Type export_table_task	更多信息
46 2024-11-25 17:20:18.619	WorkflowTaskCompleted	Scheduled Event ID 44	更多信息
45 2024-11-25 17:20:18.608	WorkflowTaskStarted	Scheduled Event ID 44	更多信息
44 2024-11-25 17:20:18.602	WorkflowTaskScheduled	Task Queue Name 8@dataflow-pyworker-857c99c44f-tlj4p-91c55b04... 更多信息	更多信息
43 2024-11-25 17:20:18.602	ActivityTaskCompleted	Result {"heartbeat_timeout":60}	更多信息

对于周期性的表到表的离线同步任务，要查看当前的调度信息，只需切换到「**调度配置**」页面，并点击图标 **>>**，即可显示详细的调度信息。

调度配置

工作流名称	PyFileBatchWorkflow
任务队列	PYTHON_FILE_BATCH_QUEUE
调度地址	gx_scheduler(temporal-frontend.openpie-infra:7233)
周期设置	周期
任务执行周期	*/60 20 1-30 1-12 6

当前调度

调度状态	● 运行中 暂停调度
最近执行时间	2024-11-25 16:43:01
下次执行时间	2024-11-30 20:00:00 2024-12-07 20:00:00 2024-12-14 20:00:00 2024-12-21 20:00:00 2024-12-28 20:00:00 2025-01-04 20:00:00 2025-01-11 20:00:00 2025-01-18 20:00:00 2025-01-25 20:00:00 2025-02-01 20:00:00

对于周期性的表到表的离线同步任务，如果需要暂停未来的调度周期，只需点击 **暂停调度**，之后可以随时重启暂停的调度。请注意，暂停操作并不会影响当前正在运行的任务。如果决定取消未来的所有调度周期，可以点击 **删除**。这一操作同样不会删除任何正在执行的任务的调度信息。

创建库到库的离线同步任务

在数据集成的「**数据同步**」页面，用户可以根据业务需求分别设置数据关系、调度配置、过滤规则和运行配置以创建库到库的离线同步任务。参考步骤如下：

1. 切换到「**离线同步**」页面，点击 **新建离线同步** 即可进入创建页面。
2. 输入该离线同步任务的名称。
3. 从“类型”下拉列表中选择离线同步的类型为“库到库”。
- 4.（可选）输入对该任务的描述。
5. 点击 **下一步** 以进入离线同步任务的配置详情页面。



6. 设置“数据关系”以定义数据流向。

在「**数据关系**」页面，点击 **编辑** 并执行如下操作：

1. 分别选择源数据源和目标数据源。
2. 选择数据复制方式为“整库复制”或者“多表复制”。

提示：

整库复制方式仅适用于源数据库和目标数据库中表名完全相同的情况。如果表名不一致，请选择多表复制方式。

- 如果选择“多表复制”，点击 **添加表** 并根据需要选择源表和对应的目标表。
3. 点击 **保存** 以确认更改。
 7. 添加调度器配置。

切换到「**调度配置**」页面，点击 **编辑** 并执行如下操作：

1. 在“调度地址”下拉列表中选择运行该数据流的调度器。

有关配置调度器的详细信息，请参见 [配置调度系统-添加调度器配置](#)。

提示:

通常情况下，如果没有特殊环境需求，设置一个调度地址即可。

2. 在“工作流名称”下拉框中选择“PyFileBatchWorkflow”。工作流的作用是将作业任务按照指定顺序组织起来。
3. 在“任务队列”下拉框中选择“PYTHON_FILE_BATCH_QUEUE”。任务队列是轻量级的、动态分配的队列，用于轮询任务。

注意:

系统初始化时应该已经配置了调度器信息。在当前版本的系统中，调度器的工作流名称和任务队列都是固定配置。在选择调度地址后，系统会自动协助用户完成选择。如果在使用过程中发现缺少这类信息，请联系 OpenPie 技术支持团队获取帮助。

4. 设置任务执行周期。库到库的离线同步仅支持“立即执行”，即只会创建一个任务并执行一次。
5. 点击**保存**以确认更改。
8. (可选) 添加过滤规则。

切换到「**过滤规则**」页面，首先点击**编辑**，之后输入 where 条件以用于过滤源数据表和目标数据表中的数据。

过滤规则分为如下三类：

- 添加普通条件：需要输入字段名和字段值，并指定逻辑关系。例如“l_quantity > 100”。
- 添加时间范围条件：需要输入字段名和满足时间范围（年/月）的字段值。例如，字段“l_shipdate”满足时间范围过去“12 月”。
- 自定义过滤条件：点击**切换为自定义**，分别输入源表和目标表的自定义 where 表达式（请注意，不要在表达式末尾加分号“;”）。例如，“where update_time > '2024-01-01'"。

在所有过滤条件都添加完成后，点击**保存**以确认更改。

9. 修改运行配置。

“运行配置”展示数据同步任务的系统信息。由于离线同步任务需要做数据校验和文件导出的场景，由此运行配置包括比较算法、差异行数量限制、S3 桶信息、导出文件格式等配置项。

切换到「**运行配置**」页面，点击**修改**，并执行如下操作：

- 在运行配置的编辑页面修改 `s3_config_id` 字段以设置 S3 连接配置信息，之后点击 **完成**。对于其他配置项，可以保持它们的默认设置不变。

提示：

对于批处理类型的导出文件，当前版本默认将其存储路径设置为 S3。用户可以通过数据集成的「总览」页面来添加 S3 配置信息。一旦配置完成，用户就可以在运行配置中通过 `s3_config_id` 字段选择对应的 S3 配置 ID。如果需要指定除默认配置之外的存储桶，用户还可以通过配置 `s3_config_bucket` 字段来指定 S3 存储桶，从而覆盖默认的存储桶信息配置。

- 有关运行配置项的详细信息，请参见 [离线同步任务运行配置项说明](#)。
- 运行配置修改完成后，点击 **X** 即可返回运行配置页面。
 - 点击 **测试 S3 连接**，如果配置信息无误，系统会显示“连接测试成功”的信息。
 - 在当前离线同步任务的配置详情页面，点击 **运行** 即可开始运行库到库的离线同步任务。

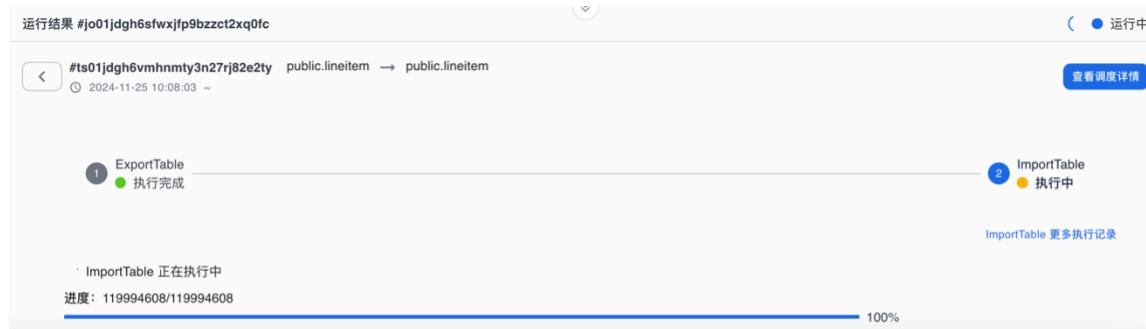
与表到表的离线同步任务不同，库到库的离线同步任务一般会以表为单位拆分成多个子任务，子任务会自动同步到任务列表中并实时显示子任务的执行状态，可能为如下之一：

- 已创建
- 正在运行
- 已完成
- 正在终止
- 正在重启
- 已终止
- 失败
- 数据不一致

运行结果 #jo01jdgh6sfwxjf9bzczl2xq0fc			
任务ID	任务内容	任务状态	运行时间
ts01jdgh6vppp5eyyzv9k2v31jq	public.orders → public.orders	已完成	开始时间: 2024-11-25 10:08:03 结束时间: 2024-11-25 10:37:32
ts01jdgh6vndsn92vcn9b50n6yr	public.supplier → public.supplier	已完成	开始时间: 2024-11-25 10:08:03 结束时间: 2024-11-25 10:08:28
ts01jdgh6vnn7e0rhkhjp7h3j9s	public.customer → public.customer	已完成	开始时间: 2024-11-25 10:08:03 结束时间: 2024-11-25 10:14:06
ts01jdgh6vmhnmt3n27rj82e2ty	public.lineitem → public.lineitem	正在运行	开始时间: 2024-11-25 10:08:03 结束时间: -

用户选中库到库的离线同步任务的子任务并单击，即可进入目标子任务的执行详情页面。

对于正在运行中的离线同步子任务，运行结果区域会动态展示离线同步任务在 ExportTable、ImportTable 和 TableDataDiff 各个阶段的执行进度。



与表到表的离线同步任务相同，各个子任务的运行结果区域会展示该任务在 ExportTable、ImportTable 和 TableDataDiff 各个阶段的执行结果、响应消息、执行时间，以及是否执行成功。

在离线同步任务执行完成后，用户可以点击 **查看调度详情** 来查看执行该任务时详细的调度情况。

离线同步任务运行配置项说明

配置项	说明	默认值
key_columns	表的主键	无
extra_columns	用于比较的非主键列	无
algorithm	比较算法。取值范围： • md5diff：直接计算 MD5 值，且只能确定文件或数据之间是否存在差异。 • hashdiff：递归计算 md5 值，并找出具体不同的行。 • countdiff：通过计算记录数量判断是否有差异。	md5diff
limit	最多能返回的有差异的行的数量。取值范围 [1,1000]。	1000
threads	执行比较的线程数量	1
bisection_threshold	在 diff 操作期间每个段中的行数阈值。如果段内的行数低于此阈值，则将在内存中直接进行比较。	无
bisection_factor	在每次比较迭代中数据集被划分为校验和片段的数量。该数值应小于 bisection_threshold。	无

配置项	说明	默认值
verbose	校验时是否输出详细日志	false
s3_config_id	s3 连接配置信息	无
s3_config_bucket	s3 桶的名称。如果为空，则使用默认值。	空
mssql_server_name	mssql 服务的名称，如果用 Parquet 就是必填项。	无
export_batch_rows	导出文件被分割成多个文件，每个文件的行数。	100000
batch_parallel_size	批作业并行运行子作业的个数	5
batch_file_format	批作业的文件格式，支持 csv 和 parquet 两种。	csv
heartbeat_timeout	activity 的心跳超时时间。单位：秒。	60
export_compress_parallel	压缩上传任务最大并行数量	5
enable_pk_duplicate_detect	是否开启检测主键重复。开启后，在导入数据前会根据主键进行查重。	True
source_timezone	源库时区。连接源数据库和目标数据库时都使用该时区。	Asia/Shanghai
import_order_by	对于大表，将该字段设置为常用于查询过滤条件的字段（例如日期），可以显著提高数据库查询对数据文件的过滤效率，从而提升查询性能。语法和 SQL 的 ORDER BY 一致。	无

离线同步任务失败问题排查

在执行同步任务过程中，可能遇到的执行失败的情况，在任务详情页面的运行结果区域，用户可以通过查看任务的状态和执行详情来了解失败的原因。

通常情况下，库到库的离线同步任务很少需要再次执行，除非失败原因是由于资源、环境等不可控因素引起的，在这种情况下，任务可以被重新执行。

如果任务执行失败是由于系统稳定性、数据库状态等服务质量问题引起的，用户可以在任务详情页面的运行结果区域，找到当前执行的任务记录以及执行失败的具体步骤，并点击该步骤下方的 **重试** 来重新执行该步骤。

The screenshot shows a task execution history with three steps:

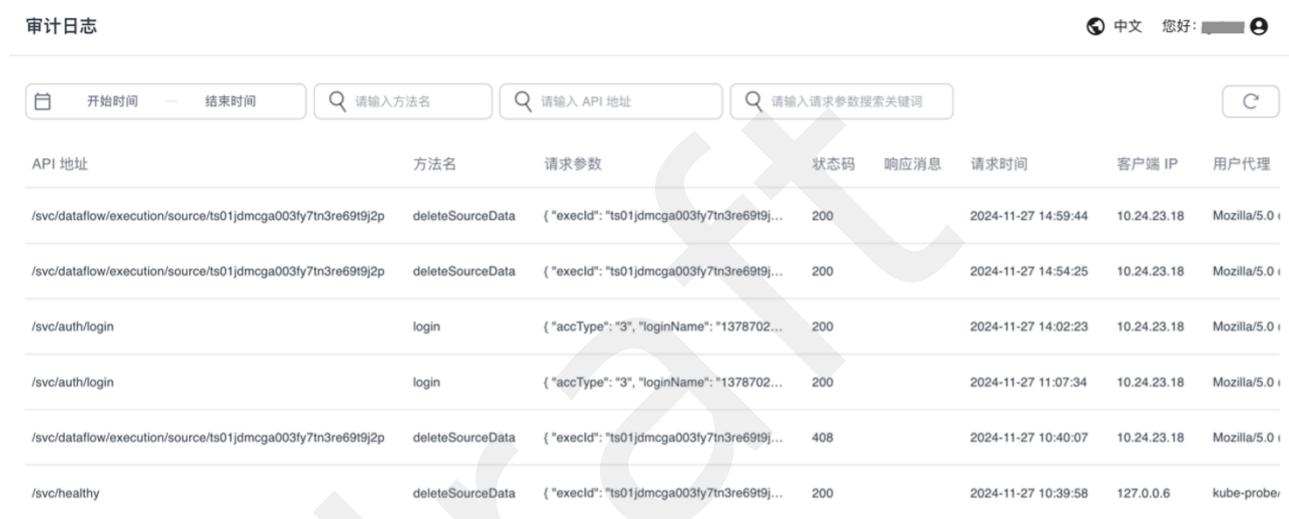
- 1 ExportTable**: Status: 执行完成 (Completed), Step icon: 1 (green dot)
- 2 ImportTable**: Status: 执行完成 (Completed), Step icon: 2 (green dot)
- 3 TableDataDiff**: Status: 执行失败 (Failed), Step icon: 3 (red dot). A red box highlights this step.

Below the steps, there is a message: "错误信息: ReadTimeout: HTTPConnectionPool(host='pie-data-diff.openpie-connect.svc.cluster.local', port=8000): Read timed out. (read timeout=30)" and "响应消息: datadiff task failed: Activity task failed". At the bottom right, there is a button labeled "TableDataDiff 更多执行记录" and a "重试" (Retry) button.

如果任务失败是由于配置问题引起的，那么需要重新修改任务配置。用户可以点击 [返回配置](#) 并根据报错信息修改相应的配置，之后可以重新运行该任务。

在执行同步任务时，通常不会出现数据不一致的情况。目前，数据不一致的最主要原因是重复导入数据，而次要原因可能是数据导出、导入或数据校验流程中出现了问题。如果任务状态显示“数据不一致”，建议首先清理目标数据源的数据，然后确认环境配置等的正确性，最后重新执行任务。如果执行上述操作后数据仍然不一致，请联系 OpenPie 技术支持团队。

此外，用户可以通过审计日志来了解任务的方法调用情况。在 PieDataCS 云原生平台的数仓操作界面，点击菜单栏「[审计日志](#)」即可进入该功能页面。



The screenshot shows the 'Audit Log' page with the following interface elements:

- Header:** 中文 您好: [User Icon]
- Search and Filter:**
 - 开始时间 - 结束时间
 - 请输入方法名
 - 请输入 API 地址
 - 请输入请求参数搜索关键词
 - 刷新图标
- Table Headers:** API 地址, 方法名, 请求参数, 状态码, 响应消息, 请求时间, 客户端 IP, 用户代理
- Table Data:** A list of API calls with the following details:

API 地址	方法名	请求参数	状态码	响应消息	请求时间	客户端 IP	用户代理
/svc/dataflow/execution/source/ts01jdmcga003fy7tn3re69t9j2p	deleteSourceData	{ "execId": "ts01jdmcga003fy7tn3re69t9j... }	200		2024-11-27 14:59:44	10.24.23.18	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/115.0.5790.102 Safari/537.36
/svc/dataflow/execution/source/ts01jdmcga003fy7tn3re69t9j2p	deleteSourceData	{ "execId": "ts01jdmcga003fy7tn3re69t9j... }	200		2024-11-27 14:54:25	10.24.23.18	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/115.0.5790.102 Safari/537.36
/svc/auth/login	login	{ "accType": "3", "loginName": "1378702... }	200		2024-11-27 14:02:23	10.24.23.18	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/115.0.5790.102 Safari/537.36
/svc/auth/login	login	{ "accType": "3", "loginName": "1378702... }	200		2024-11-27 11:07:34	10.24.23.18	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/115.0.5790.102 Safari/537.36
/svc/dataflow/execution/source/ts01jdmcga003fy7tn3re69t9j2p	deleteSourceData	{ "execId": "ts01jdmcga003fy7tn3re69t9j... }	408		2024-11-27 10:40:07	10.24.23.18	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/115.0.5790.102 Safari/537.36
/svc/healthy	deleteSourceData	{ "execId": "ts01jdmcga003fy7tn3re69t9j... }	200		2024-11-27 10:39:58	127.0.0.6	kube-probe/

清理源数据

对于离线同步任务，用户可以根据设定的周期和 where 条件来实现定时定量的数据迁移。数据迁移完成后，如果用户希望清理源数据库的表中的特定数据，可以通过执行清理源数据的操作来实现。

在数据集成的「[数据同步](#)」页面，用户点击 [数据清理](#) 即可进入清理数据的功能页面。该页面以列表形式展示了当前所有的离线同步任务。

数据同步 / 离线同步 / 清理数据

任务 ID	状态	同步任务	运行时间	导出文件状态	源数据状态
ts01jd96jv3qrqsx55tvzn26jtr7	● 失败	oracle_lineitem	开始时间: 2024-11-22 14:29:43 结束时间: 2024-11-22 14:30:24	存在 清理	未清理
ts01jd968p8ytng0xzpg4ysmmskv	● 已完成	oracle_test	开始时间: 2024-11-22 13:42:08 结束时间: 2024-11-22 13:42:16	存在 清理	未清理 清理源数据 清理历史
ts01jd965x3e9rtxt633syc0fg4s	● 数据不一致	mysql_test	开始时间: 2024-11-22 13:40:37 结束时间: 2024-11-22 13:40:44	存在 清理	未清理
ts01jd963jx2w551k61pfa73r2f6	● 已完成	mysql_up	开始时间: 2024-11-22 13:39:21 结束时间: 2024-11-22 13:39:28	存在 清理	未清理 清理源数据 清理历史
ts01jd43k0y1hzexsf2mv2gez981	● 已完成	mysql_up	开始时间: 2024-11-20 15:30:41 结束时间: 2024-11-20 15:30:46	存在 清理	已清理 清理历史
ts01jd3qjvgrc2ap0x8dcnknqr0	● 已完成	mysql_lineitem	开始时间: 2024-11-20 10:49:20 结束时间: 2024-11-20 12:30:31	已删除	清理失败 清理历史

在「清理数据」页下的离线任务列表中的目标任务的“源数据状态”栏下，用户可以点击 **清理源数据**，并基于 where 条件来按需清理目标源数据，如下图所示。



对于数据量较大的场景，为了保证源数据库的稳定性，可以实现源数据的分批多轮自动清理。默认情况下，每轮删除的数据量为 10000 条，但用户也可以输入自定义的整数值来设定每轮删除的数据量。

如果未设置 where 条件，源数据清理操作将会清空整张表的所有数据。

源数据清理完成后，系统会显示“已清理”。如果用户执行了清理数据源的操作，可以点击 **查看清理历史** 来查看历史清理信息，其中包括调度器的 IP 地址、执行的 SQL 语句、当前用户名、清理时间以及本次清理所删除的记录条数等详细信息。



清理导出文件

对于离线同步任务，数据迁移完成后或在迁移过程中，如果通过数据校验失败等方式发现了导出数据的问题，用户可以清理这些有问题的导出文件。

在数据集成的「**数据同步**」页面，用户点击 **数据清理** 即可进入清理数据的功能页面。该页面以列表形式展示了当前所有的离线同步任务。

在「**清理数据**」页下的离线任务列表中的目标任务的“导出文件状态”栏下，用户可以点击 **清理**，确认后即可清理目标任务。

文件清理成功后，系统会显示“已删除”。如果导出文件状态栏显示为“不存在”，则表明该任务没有生成导出文件。

9 创建实时同步任务

实时同步任务实现了利用 PieCloudDB 的 Flink 动态执行器中间件与 Flink 实时流计算引擎进行对接，进而将计算结果数据写入 PieCloudDB 的虚拟数据仓库中。

实时同步适用于流处理场景，支持以下数据同步类型：

- 单表全量与 CDC (Change Data Capture) 增量同步
- 整库全量与 CDC 增量同步
- 整库指定表的全量与 CDC 增量同步

创建表到表的实时同步任务

在数据集成的「数据同步」页面，用户可以根据业务需求分别设置数据关系、机器配置和运行配置以创建表到表的实时同步任务。参考步骤如下：

1. 切换到「实时同步」页面，点击 **新建实时同步** 即可进入创建页面。
2. 输入该实时同步任务的名称。
3. 从“类型”下拉列表中选择同步类型为“表到表”。
4. (可选) 输入对该任务的描述。
5. 点击 **下一步** 以进入该实时同步任务的配置详情页面。



6. 设置“数据关系”以定义数据流向。

由于实时同步任务是用于处理实时数据的，为了避免在实时处理过程中因源端和目标端的表结构变化而导致的问题，用户需要手动配置字段之间的映射关系。

注意:

- 当前版本的实时表同步功能仅支持单个源表到单个目标表的同步，尚不支持涉及多个源表或目标表的同步。关于此类定制化需求，请联系 OpenPie 技术支持团队。
- 为保证数据一致性和完整性，在进行实时同步时，源表必须包含主键。主键能够唯一标识每条记录，防止因故障导致数据丢失或重复。如果进行部分字段关联而未将源表的主键与目标表的对应字段关联，系统将自动寻找与源表主键字段名相同的目标字段以进行关联和同步。如果未找到相应的目标字段，同步将失败。

在「数据关系」页面，点击 **编辑** 并执行如下操作：

1. 点击 **添加源表** 选择源数据源类型、源数据源和源表，点击 **确定** 后，系统将自动将其添加到编辑区，用于配置映射关系。

注意:

对于源数据源，当前仅支持 MySQL、PostgreSQL 和 PieCloudDB TP。

用户也可以输入源表的别名，该选项同 SQL 中的 `AS` 语法，例如 “`table1 AS t1`”，则表别名可输入 `t1`。

2. 点击 **添加目标表** 选择目标数据源类型、目标数据源和目标表，点击 **确定** 后，系统将自动将其添加到编辑区，用于配置映射关系。

注意:

对于目标数据源的类型，当前仅支持 PieCloudDB。而且目标表不支持多任务共享同步。

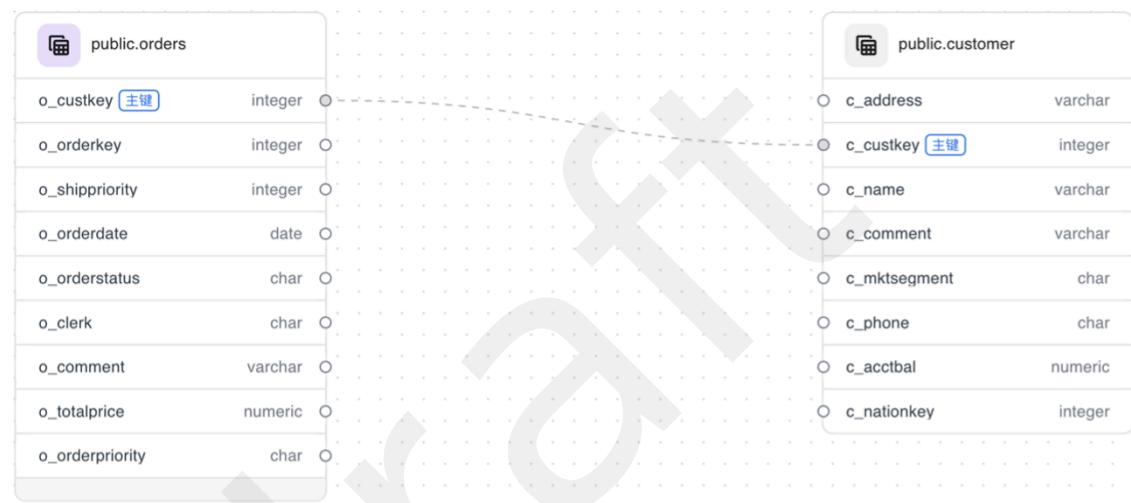
用户也可以输入目标表的别名，该选项同 SQL 中的 `AS` 语法，例如 “`table1 AS t1`”，则表别名可输入 `t1`。

3. 在编辑区域配置源表和目标表之间的字段关联。

提示:

在配置数据关系时，如果进行的是整表实时同步，并且源表与目标表的表结构（即字段）完全一致，那么就无需手动进行字段关联。反之，如果源表和目标表的表结构不完全一致，或者只同步部分字段，那么就需要手动连接相关字段以建立关联。

选择指定源表（或目标表）的字段从原点进行拖拽到目标表（或源表）的对应字段生成关系线，多个字段映射可以重复关联操作，直到源表和目标表及其字段之间的所有关系都配置完成。如下图所示。



7. 修改机器配置。

切换到「**机器配置**」页面，点击**编辑**以修改机器配置。如下机器配置设置完成后，点击**保存**以确认更改。

机器配置项说明如下：

- Job Manager CPU 核数：Job Manager 所能使用的 CPU 总数。默认值：1 Core。
- Job Manager CPU 内存：Job Manager 的总进程内存大小。默认值：1024 Mi。
- Task Manager CPU 核数：Task Manager 所能使用的 CPU 总数。默认值：1 Core。
- Task Manager CPU 内存：Task Manager 的总进程内存大小。默认值：1024 Mi。
- Task Manager Slot：Task Manager 的任务槽数量。默认值：16。

其中，Job Manager 是实时同步作业的管理中心，负责接收用户提交的作业，并进行作业的调度和管理；Task Manager 是实际负责执行计算的 Worker，在其上执行

实时同步作业的一组 Task；Task Manager Slot 是 Task Manager 中的最小资源调度单位，用于分配和托管任务的内存空间。

注意：

为避免资源浪费，当 Task Manager Slot 数量超过作业执行的默认并行度 (parallelism.default) 时，系统会自动将 Slot 数量调整为与并行度相等。也就是说，仅当并行度大于 Slot 数量时，系统会按照用户实际设定的 Slot 数量执行作业；反之，如果并行度小于 Slot 数量，系统会按照并行度执行作业。

8. 修改运行配置。

“运行配置”展示数据同步任务的系统信息。对于实时同步任务，运行配置包括 flink checkpoint 触发间隔、超时时长、重启策略、作业并行度等配置项。

切换到「运行配置」页面，首先点击 **修改** 以根据实际需求修改相应的配置信息，之后点击 **完成**。

有关运行配置项的详细信息，请参见 [实时同步任务运行配置项说明](#)。

点击 **X** 即可退出并返回配置页面。

9. 在实时同步任务的详情页面面，点击 **运行** 即可开始运行表到表的实时同步任务。

如果在点击 **运行** 后就发生配置相关报错，用户可以点击 **返回配置** 并根据报错信息修改相应的配置，之后可以重新运行该任务。

在任务运行期间，如果需要强制停止运行该任务，点击 **终止** 即可。如需重新运行该任务，用户可以点击 **重启作业** 或者 **重启任务**，重启任务将从上一个 Savepoint 开始执行。

表到表 tp-pdb #2 #jo01jf99fy24mjk3m3echfxgx064 ● 已关闭

数据关系 **机器配置** **运行配置**

public.customer

c_acctbal	numeric
c_nationkey	integer
c_address	varchar
c_custkey	integer
c_mktsegment	char
c_name	varchar

public.customer

c_name	varchar
c_address	varchar
c_nationkey	integer
c_custkey	integer
c_phone	char
c_acctbal	numeric

运行结果 #jo01jf99fy24mjk3m3echfxgx064

#ts01jf99fy2h3n4a0tk2qvkgzgz3
① 2024-12-17 11:09:55 ~ 2024-12-17 11:11:56

● 已终止 重启任务

结果

execution terminated gracefully

查看详情 复制文本

创建库到库的实时同步任务

在数据集成的「**数据同步**」页面，用户可以根据业务需求分别设置数据关系、机器配置和运行配置以创建库到库的实时同步任务。参考步骤如下：

1. 切换到「**实时同步**」页面，点击**新建实时同步**即可进入创建页面。
2. 输入该实时同步任务的名称。
3. 从“类型”下拉列表中选择同步类型为“库到库”。
- 4.（可选）输入对该任务的描述。
5. 点击**下一步**以进入该实时同步任务的配置详情页面。

库到库 tp-pdb

数据关系 **编辑**

数据来源

数据源类型	未配置
源数据源	未配置

数据去向

数据源类型	未配置
目标数据源	未配置

表映射

未配置

运行

6. 设置“**数据关系**”以定义数据流向。

在「**数据关系**」页面，点击**编辑**并执行如下操作：

1. 分别选择源数据源和目标数据源。

2. 选择数据复制方式为“整库复制”或者“多表复制”。

提示:

整库复制方式仅适用于源数据库和目标数据库中表名完全相同的情况。如果表名不一致，请选择多表复制方式。

如果选择“多表复制”，则点击**添加表**，并根据需要选择源表和对应的目标表。在库到库的多表映射中，系统提供了同名表的自动匹配和填充功能。



3. 点击**保存**以确认更改。

7. 修改机器配置。

切换到「**机器配置**」页面，点击**编辑**以修改机器配置。如下机器配置设置完成后，点击**保存**以确认更改。

机器配置项说明如下：

- Job Manager CPU 核数：Job Manager 所能使用的 CPU 总数。默认值：1 Core。
- Job Manager CPU 内存：Job Manager 的总进程内存大小。默认值：1024 Mi。
- Task Manager CPU 核数：Task Manager 所能使用的 CPU 总数。默认值：1 Core。
- Task Manager CPU 内存：Task Manager 的总进程内存大小。默认值：1024 Mi。
- Task Manager Slot：Task Manager 的任务槽数量。默认值：16。

其中，Job Manager 是实时同步作业的管理中心，负责接收用户提交的作业，并进行作业的调度和管理；Task Manager 是实际负责执行计算的 Worker，在其上执行

实时同步作业的一组 Task；Task Manager Slot 是 Task Manager 中的最小资源调度单位，用于分配和托管任务的内存空间。

注意：

为避免资源浪费，当 Task Manager Slot 数量超过作业执行的默认并行度 (parallelism.default) 时，系统会自动将 Slot 数量调整为与并行度相等。也就是说，仅当并行度大于 Slot 数量时，系统会按照用户实际设定的 Slot 数量执行作业；反之，如果并行度小于 Slot 数量，系统会按照并行度执行作业。

8. 修改运行配置。

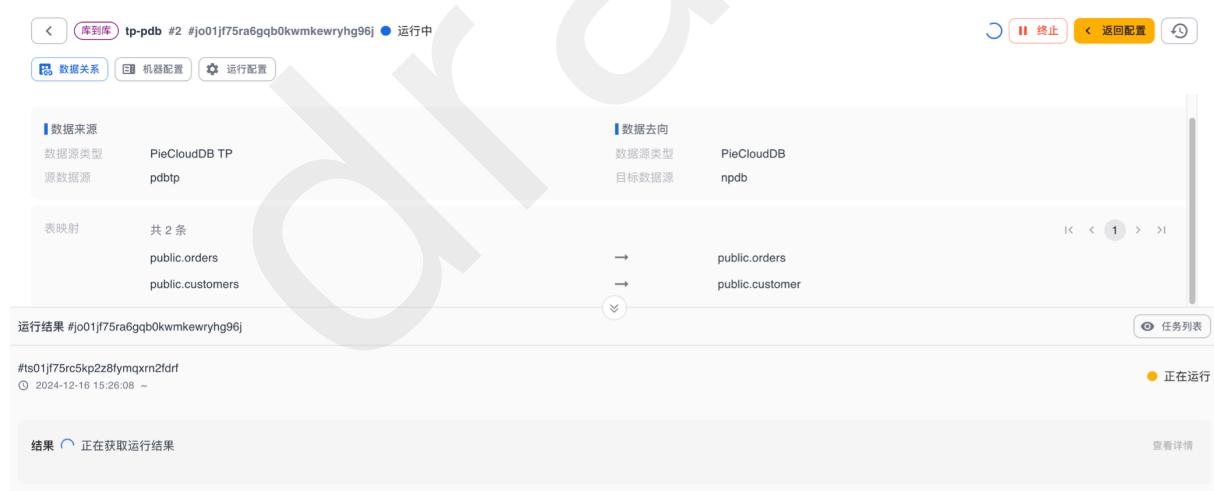
“运行配置”展示数据同步任务的系统信息。对于实时同步任务，运行配置包括 flink checkpoint 触发间隔、超时时长、重启策略、作业并行度等配置项。

切换到「**运行配置**」页面，点击 **修改** 以根据实际需求修改相应的配置信息，之后点击 **完成**。

有关运行配置项的详细信息，请参见 **实时同步任务运行配置项说明**。

点击 **X** 即可退出并返回配置页面。

9. 在实时同步任务的详情页面，点击 **运行** 即可开始运行库到库的实时同步任务。



如果在点击 **运行** 后就发生配置相关报错，用户可以点击 **返回配置** 并根据报错信息修改相应的配置，之后可以重新运行该任务。

与表到表的实时同步任务相同，在任务运行期间，如果需要强制停止运行该任务，点击 **终止** 即可。如需重新运行该任务，用户可以点击 **重启作业** 或者 **重启任务**，重启任务将从上一个 Savepoint 开始执行。

查看实时同步任务详情

对于执行成功的实时同步任务，用户在运行结果区域点击 **查看详情** 即可查看 flink 算子信息和运行日志等信息。

The screenshot shows the 'View Task Details' section for a successful sync task. At the top, there's a header with a back arrow, a 'Table to Table' button, the task ID 'tp-pdb #3 #jo01jf953z0t38gbn5nt6zk11agd', and a status indicator 'Running'. Below the header are three tabs: 'Data Relationship' (selected), 'Machine Configuration', and 'Run Configuration'. The main area displays two database schemas: 'pieclouddb_tp pdbtp' and 'pieclouddb npdb'. The 'public.customer' table is shown in both schemas with columns: c_acctbal (numeric), c_nationkey (integer), and c_address (varchar). Below the schemas is a preview of the Flink operator graph. A 'Run Result' section shows the command '#ts01jf953z1mt7f6yhc26p4frv95' and the start time '2024-12-17 09:53:28'. To the right, a 'Task List' button is visible. At the bottom, a 'Result' section says 'Execution successful! Click the top-right corner to view details.' and a 'View Details' link.

- 算子列表：算子列表页面展示运行中的 flink 作业的概览信息和算子详情。概览信息包括作业的运行时长、状态、开始时间等；算子列表以图形化方式展示了作业的逻辑执行图，可以查看作业中每个算子的名称、并行度、反压等信息。有关各个算子的更多信息展示在下方列表中。

The screenshot shows the 'Operator List' page for a running job. The top navigation bar includes tabs for 'Operator List', 'Exception', 'Checkpoints', 'Job Configuration', 'Cluster Configuration', and 'Run Log'. The job details at the top show the ID 'c6365268914519f388ad8b44cf1cc8cd', a green 'RUNNING' status with a '2' badge, and a start time of '2024-12-16 14:13:18.952'. Below this, the execution graph shows a 'Source: PieCloudDB-TP Parallel Source -> Map' node and a 'Sink: Writer' node. The 'Source' node has a parallelism of 1 and shows metrics like 'Backpressured (max):0%' and 'busyPercentage (max):0%'. The 'Sink' node also shows metrics like 'Backpressured (max):0%' and 'busyPercentage (max):0%'. The graph is labeled 'FORWARD'.

- 异常：异常页面展示 flink 作业执行与重试过程中发生的所有异常堆栈，以协助排查问题。

算子列表 异常 Checkpoints Job 配置 集群配置

Root 异常

```

1 2024-05-24 13:44:52
2 org.apache.flink.runtime.JobException: Recovery is suppressed by NoRestartBackoffTimeStrategy
3 at org.apache.flink.runtime.executiongraph.failover.flip1.ExecutionFailureHandler.handleFailure(ExecutionFailureHandler.java:176)
4 at org.apache.flink.runtime.executiongraph.failover.flip1.ExecutionFailureHandler.getFailureHandlingResult(ExecutionFailureHandler.java:107)
5 at org.apache.flink.runtime.scheduler.DefaultScheduler.recordTaskFailure(DefaultScheduler.java:285)
6 at org.apache.flink.runtime.scheduler.DefaultScheduler.handleTaskFailure(DefaultScheduler.java:276)
7 at org.apache.flink.runtime.scheduler.DefaultScheduler.onTaskFailed(DefaultScheduler.java:269)
8 at org.apache.flink.runtime.scheduler.SchedulerBase.onTaskExecutionStateUpdate(SchedulerBase.java:764)
9 at org.apache.flink.runtime.scheduler.SchedulerBase.updateTaskExecutionState(SchedulerBase.java:741)
10 at org.apache.flink.runtime.scheduler.SchedulerNG.updateTaskExecutionState(SchedulerNG.java:83)
11 at org.apache.flink.runtime.jobmaster.JobMaster.updateTaskExecutionState(JobMaster.java:488)
12 at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
13 at java.base/jdk.internal.reflect.NativeMethodAccessorImpl.invoke(Unknown Source)
14 at java.base/jdk.internal.reflect.DelegatingMethodAccessorImpl.invoke(Unknown Source)
15 at java.base/java.lang.reflect.Method.invoke(Unknown Source)
16 at org.apache.flink.runtime.rpc.pekko.PekkoRpcActor.lambda$handleRpcInvocation$1(PekkoRpcActor.java:309)
17 at org.apache.flink.runtime.concurrent.ClassLoadingUtils.runWithContext(ClassLoadingUtils.java:83)
18 at org.apache.flink.runtime.rpc.pekko.PekkoRpcActor.handleRpcInvocation(PekkoRpcActor.java:307)
19 at org.apache.flink.runtime.rpc.pekko.PekkoRpcActor.handleRpcMessage(PekkoRpcActor.java:222)
20 at org.apache.flink.runtime.rpc.pekko.FencedPekkoRpcActor.handleRpcMessage(FencedPekkoRpcActor.java:85)
21 at org.apache.flink.runtime.rpc.pekko.PekkoRpcActor.handleMessage(PekkoRpcActor.java:168)
22 at org.apache.pekko.japi.pf.UnitCaseStatement.apply(CaseStatements.scala:33)
23 at org.apache.pekko.japi.pf.UnitCaseStatement.apply(CaseStatements.scala:29)
24 at scala.PartialFunction.applyOrElse$(PartialFunction.scala:127)
25 at scala.PartialFunction.applyOrElse$(PartialFunction.scala:126)
26 at org.apache.pekko.japi.pf.UnitCaseStatement.applyOrElse(CaseStatements.scala:29)
27 at scala.PartialFunction$OrElse.applyOrElse$(PartialFunction.scala:175)

```

异常历史

时间	异常	任务名称	地址	异常信息
2024-05-24 13:44:52	org.apache.flink.runtime.JobException	(global failure)	(未分配)	查看

- Checkpoints: Checkpoints 页面展示 flink 作业执行过程中所有的 checkpoint 详情、统计信息和配置信息。

算子列表 异常 Checkpoints Job 配置 集群配置 运行日志

总览 历史 统计 配置

Checkpoint 数量 Triggered:1104 In Progress:0 Completed:1091 Failed:13 Restored:0

最新完成的 Checkpoint ID:1091 | Completion Time:2024-12-16 14:31:50.992 | End to End Duration:20073d 6h 31m 50s | Checkpointed Data Size:3.32 KB | Full Checkpoint Data Size:3.32 KB

Checkpoint 详情: Path:/file/flink/its01/f71jev1p3emew2kerz598px/checkpoints/c6365268914519f388ad8b44cf1cc8cd/chk-1091 Discarded:- Checkpoint Type:aligned checkpoint

算子:

Name	Acknowledged	Latest Acknowledgment	End to End Duration	Checkpointed Data Size	Full Checkpoint Data Size	Processed (per)
Source: PieCloudDB-TP Parallel Source -> Map	1 / 1 (100 %)	2024-12-16 14:31:50.992	14ms	3.09 KB	3.09 KB	3.09 KB
Sink: Writer	1 / 1 (100 %)	2024-12-16 14:31:50.992	14ms	234 B	234 B	234 B

最新失败的 Checkpoint 无

最新保存的 Checkpoint 无

最新保存的 Checkpoint 无

- Job 配置: Job 配置页面展示 flink 作业的执行配置。

算子列表 异常 Checkpoints Job 配置 集群配置 运行日志

执行配置

执行模式 PIPELINED 最大重试次数 Restart deactivated. 任务并行数 1
对象重用模式 false

用户配置

暂无数据

- 集群配置：集群配置页面展示 flink 作业所属集群的配置详情、JVM 启动参数和 classpath。

The screenshot shows the 'Cluster Configuration' tab selected in a navigation bar. Below it, two sections are displayed: 'Job Manager 配置' and 'JVM'.

Job Manager 配置

\$internal.application.program-args	--jobId:j01 f71jetqa7n82h3gm8gpew9--serverAddr:pie-dataflowserver-grpc.openpie-connect.svc.cluster.local:8099
\$internal.flink.version	v1_18
\$internal.pipeline.job-id	c6365268914519f388ad8b44cf1cc8cd
blob.server.port	6124
execution.checkpointing.externalized-checkpoint-retention	RETAIN_ON_CANCELLATION
execution.checkpointing.interval	5 min
execution.shutdown-on-application-finish	false
execution.submit-failed-job-on-application-error	true

JVM

version	OpenJDK 64-Bit Server VM - Eclipse Adoptium - 11/11.0.22+7
arch	amd64
options	-Xmx1530082096 -Xms1530082096 -XX:MaxMetaspaceSize=268435456 -XX:+IgnoreUnrecognizedVMOptions -Dlog.file=/opt/flink/log/flink--kubernetes-application-0-ts01jf71jev1p3emew2kerz598px-676db68df9-jq748.log

- 运行日志：运行日志页面展示 flink 作业所属集群的 Job manager 和 Task manager 的运行日志。

The screenshot shows the 'Run Log' tab selected in a navigation bar. It displays a list of log files under the 'Job Manager' section.

最后更新时间	文件大小
2024-12-16 14:37:24	1.03 MB

Log file details:

- flink--kubernetes-application-0-ts01jf71jev1p3emew2kerz598px-676db68df9-jq748.log

实时同步任务运行配置项说明

配置项	说明	默认值
execution.checkpointing.interval	触发检查点（Checkpoint）的时间间隔。它指定了在连续两个检查点之间应有的最长时间（毫秒）。如果设置为大于 0 的值，系统将在这个时间间隔后自动触发检查点。	1000
execution.checkpointing.min-pause	这是在连续两个检查点之间的最小暂停时间（以毫秒为单位）。如果任务处理速度落后于数据生成速度，系统会暂停发送数据以等待任务赶上。该参数定义了暂停操作的最小时长。	500

配置项	说明	默认值
execution.checkpointing.timeout	检查点操作的超时时间（以毫秒为单位）。如果在指定的超时时间内检查点未能完成，系统将取消当前检查点，并尝试执行下一次检查点。	600000
parallelism.default	作业执行的默认并行度。它决定了作业中所有算子的默认并行实例数。	1
state.checkpoints.num-retained	保留的检查点（Checkpoint）数量。当启用了基于时间的触发检查点时，这个参数决定了在磁盘上保留多少个最近的检查点。	1
restart-strategy.type	作业失败时的重启策略类型。可能的值包括： <ul style="list-style-type: none">• none：无重启策略。• fixed-delay：固定延迟启动策略。• failure-rate：故障率重启策略。	none
restart-strategy.fixed-delay.delay	当重启策略类型为 fixed-delay 时，该参数设置了两次连续重启尝试之间的延迟。单位：秒（s）。	1
restart-strategy.fixed-delay.attempts	当重启策略类型为 fixed-delay 时，该参数设置了将作业声明为失败之前重试执行的次数。-1 表示次数无上限。	-1
restart-strategy.failure-rate.delay	当重启策略类型为 failure-rate 时，该参数设置了两次连续重启尝试之间的延迟。单位：秒（s）。	1
restart-strategy.failure-rate.failure-rate-interval	当重启策略类型为 failure-rate 时，该参数设置了测量重新启动策略故障率的时间间隔。单位：秒（s）。	60
restart-strategy.failure-rate.max-failures-per-interval	当重启策略类型为 failure-rate 时，该参数设置了在 failure-rate.delay 指定的时间窗口内允许的作业失败之前的最大重启次数。	1
state.backend.dir	状态后端存储的目录路径用于保存实时同步作业的状态信息，包括检查点和保存点。	默认使用 flink 集群默认配置

实时同步常见报错与解决方案

在执行同步任务过程中，可能遇到的执行失败的情况，在任务详情页面的运行结果区域，用户可以通过查看任务的状态和执行详情来了解失败的原因。

如果任务失败是由于配置问题引起的，那么需要重新修改任务配置。用户可以点击 **返回配置** 并根据报错信息修改相应的配置，之后可以重新运行该任务。如果实时同步任务执行失败，用户可以在运行结果区域点击 **查看详情** 来查看 flink 算子信息和运行日志等信息，以辅助排查问题。

表到表 tp-pdb #3 #jo01jf953z0t38gbn5nt6zk11agd ● 失败

数据关系 机器配置 运行配置

运行结果 #jo01jf953z0t38gbn5nt6zk11agd

#ts01jf953z1mt7f6yhc26p4frv5
① 2024-12-17 09:53:28 ~ 2024-12-17 09:54:27 ● 失败

结果

错误信息：

```
java.io.IOException: Could not perform checkpoint 5 for operator Sink: Writer (1/1)#0.
    java.lang.RuntimeException: org.postgresql.util.PSQLException: ERROR: tuple concurrently deleted (mstoream.c:5739)
        org.postgresql.util.PSQLException: ERROR: tuple concurrently deleted (mstoream.c:5739)
```

查看详情 复制文本

在执行同步任务时，如果发生如下报错，可以参考相应的解决方案。

- 报错信息:** `java.lang.IllegalArgumentException: source table [database_name.table_name] do not contain any primary key.`

原因: 源数据表未包含主键。

解决方案: 为源数据表指定一个主键。同时，如果源数据库中实际的表没有主键，而用户仅在关联表模型的字段管理中修改“主键”的信息，实时同步任务在校验时同样会报错。

- 报错信息:** `java.lang.IllegalArgumentException: source[database_name.table_name] wal level is replica, should be set to logical.`

原因: 日志级别未设置为“replica”。

解决方案: 对于源数据库（通常是 PostgreSQL 系列），将其日志级别应设置为“replica”。为了启用逻辑复制（logical replication）功能，这是实时同步功能的底层依赖，需要将数据库的日志级别设置为 `wal_level=logical`。对于 PostgreSQL 系列数据库，可以通过修改配置文件并重启数据库实例来实现这一设置；而对于 PieCloudDB TP 数据库，除了直接修改配置文件外，还可以通过在计算空间平台上设置实例的 GUC 模板来进行配置。

- 报错信息:** `java.lang.IllegalArgumentException: table [table_name] replication identity is [d], should be set to full.`

原因：源表的 *REPLICA IDENTITY* 属性未设置为 *FULL*。

解决方案：修改源表的 *REPLICA IDENTITY* 属性为 *FULL*，参考 SQL 命令如下：

```
ALTER TABLE table_name REPLICA IDENTITY FULL;
```

在 PostgreSQL 中，*REPLICA IDENTITY* 属性用于确定在逻辑复制过程中，如何识别被更新或删除的行。这个属性对于确保变更数据流（Change Data Capture, CDC）的准确性至关重要。为了在实时同步中准确地识别更新（UPDATE）和删除（DELETE）操作的前后状态，需要将 *REPLICA IDENTITY* 设置为 *FULL*。这样，PostgreSQL 会在 WAL（Write-Ahead Logging）中记录足够的信息来支持逻辑复制。

10 管理元数据

在数据集成模块中，元数据构成了一个统一的、可配置的通用表模型，包含表信息、字段信息和字段排序等属性。

通用表模型可以关联多个数据源，使不同的数据源能够共享该模型，从而减少元数据冗余，增加系统的灵活性和可扩展性。

关于通用表模型

在数据集成的「数据源」页面的数据源列表中，用户点击目标数据源名称即可进入数据源详情页面。

进入数据源详情页面后，在数据源“关联的表”区域，用户点击 **查看全部数据表** 即可进入数据集成的「全部数据表」页面。该页面以列表形式显示所有的通用表模型，并支持创建通用表模型和从数据源导入通用表模型。目标通用表模型所在行的隐藏菜单「...」还提供修改和删除通用表模型以及字段管理的选项。



数据集成 / 全部数据表							
表名	schema	表名是否大小写敏感	标题	数据源	数据源类型	描述	创建时间
tb	public	否	public.tb	adb_pg	pieclouddb	2024年11月25日星期一 10:07	...
region	public	否	public.region	adb_pg	pieclouddb	2024年11月25日星期一 10:07	...
tb3	public	否	public.tb3	adb_pg	pieclouddb	2024年11月25日星期一 10:07	...

通用表模型的列表包含如下字段信息：

- 表名：通用表模型的表名称。
- Schema（选填）：Schema 名称。如果未指定，则默认为“public”。
- 表名是否大小写敏感：表名是否开启了大小写敏感。
- 标题（选填）：自定义表的标题。
- 数据源：如果通用表是“从数据源导入”的并绑定到了源数据源，则会显示该数据源的名称。
- 数据源类型：如果通用表是“从数据源导入”的并绑定到了源数据源，则会显示该数据源的类型。
- 描述（选填）：对通用表的描述。
- 创建时间：创建通用表的时间。

通用表模型的列表提供如下快捷操作：

- 根据数据源类型筛选信息：使用**数据源类型**控件可以通过数据源的类型筛选相关的通用表信息。
- 根据 Schema 名称搜索信息：使用**输入 Schema 查询**的搜索框可以通过 Schema 的名称搜索相关的通用表信息。
- 根据表名搜索信息：使用**输入表名查询**的搜索框可以通过表的名称搜索相关的通用表信息。
- 立即刷新：点击立即刷新图标可以即刻同步当前的列表信息为最新。
- 修改通用表：在通用表模型列表的行隐藏菜单下提供了修改的快捷键，可以用于修改指定的表信息。
- 删除通用表：在通用表模型列表的行隐藏菜单下提供了删除的快捷键，可以用于删除指定的表。

创建一个新的通用表模型

在数据集成的「**全部数据表**」页面，如果需要创建一个新的通用表模型，操作步骤如下：

1. 点击**创建**即可进入创建表的页面。
2. 输入该通用表的表名。
3. （可选）指定该通用表所属的 Schema。
4. 指定是否启用表名的大小写敏感。默认设置为不启用，勾选该选项后即可启用。
5. （可选）分别输入该通用表的标题和描述信息。
6. 点击**完成**。通用表模型创建成功后，其信息会自动同步到列表中。

通过上述步骤创建的通用表模型，只是创建了一张表并无字段信息。在新创建的通用表模型所在行的隐藏菜单「...」下，用户选择**字段管理**选项即可进入目标通用表的**「字段管理」**页面，并根据实际需求进行新增字段、修改字段属性和删除字段等管理操作，详细信息请参见本章节的**管理通用表模型的字段**。

从数据源导入一个通用表模型

在数据集成的**「全部数据表」**页面，如果需要从数据源导入一个通用表模型，操作步骤如下：

1. 点击**从数据源导入**即可进入从数据源导入表的页面。
2. 在“选择数据源”的下拉列表中选择一个目标数据源，之后页面会显示“选择原始表”字段。
3. 在“选择原始表”下拉列表中选择目标数据源的一张表。如果需要查看该原始表的字段信息，点击**查看字段**即可显示所选原始表的字段名、字段类型、字段精度、主键等信息。

指定原始表后，页面会随即展开需要配置的字段信息。



4. 指定所导入的原始表是否绑定到数据源。该原始表会默认被绑定到所选的目标数据源，取消勾选即可解除绑定。

提示：

解除绑定后，该表在通用表模型列表的“数据源”和“数据源类型”字段不会显示相应的信息。

5. (可选) 如果需要重新命名要导入的表，可以输入一个新的表名。
6. 指定是否启用表名的大小写敏感。默认设置为不启用，勾选该选项后即可启用。
7. (可选) 分别输入该通用表的标题和描述信息。
8. 点击 **完成**。通用表模型导入成功后，其信息会自动同步到列表中。

在所导入的通用表模型所在行的隐藏菜单「...」下，用户选择 **字段管理** 选项即可进入目标通用表的「**字段管理**」页面，并进行新增字段、修改字段属性和删除字段等管理操作，详细信息请参见本章节的 **管理通用表模型的字段**。

管理通用表模型的字段

注意:

建议用户谨慎使用字段管理功能。

用户可以根据不同的数据源，灵活地定义表结构以适应不同的业务场景，包括新增字段、自定义已有字段的名称和字段数据类型等属性信息。

在通用表模型列表的行隐藏菜单下，用户点击 **字段管理** 选项即可进入指定通用表的「**字段管理**」页面。该页面支持如下操作：

- 新增字段

在指定通用表的字段列表下方，点击 **新增** 来添加新的自定义字段。分别输入字段的属性信息，包括字段名、字段类型、字段长度（如有）、字段精度（如有）、是否主键和注释等。

完成每个字段的操作后，点击 **完成** 按钮，该字段就会被添加到字段列表中。

- 修改字段属性

在字段列表的“操作”栏下，点击 **修改** 来修改指定字段的属性信息，包括字段名、字段类型、字段长度（如有）、字段精度（如有）、是否主键和注释等。完成每个字段的操作后，点击 **完成** 即可保存对目标字段的更改。

- 删除字段

在字段列表的隐藏菜单下，点击 **删除**，确认后即可删除指定字段并将其从列表中移除。

- 字段排序

字段排序管理功能使用户能够根据个人喜好或业务需求定义字段的显示顺序。用户可以通过拖拽或手动调整字段顺序，来合理地安排字段的排序。