大数据智能体工作台 DataBuddy

数据接入

腾讯云

文档目录

数据接入
- 支持的接入方案
- 支持的数据源与读取能力
- 数据源管理
- 数据源列表
  - 数据源概览
  - Kafka 数据源
  - MongoDB 数据源
  - MySQL 数据源
  - Oracle 数据源
  - PostgreSQL 数据源
  - SQL Server 数据源
  - TDSQL-C MySQL 数据源
- 离线数据同步
  - 创建单表同步任务
    - 表到表同步
    - 表到 Volume 同步
    - 文件到表同步
    - 文件到 Volume 同步
  - 批量创建单表同步任务
  - 离线同步任务运维监控
- 实时数据同步
  - 实时整库多表任务配置
  - 实时分库分表任务配置
  - 实时同步任务运维监控
  - 实时同步任务数据对账
- 文件上传
  - 本地上传到表
    - 本地文件上传到表
    - 支持的文件类型及约束
    - 文件解析与推断
    - 预览与 Schema 编辑
  - COS 文件上传到表
  - 本地文件上传到卷

数据接入
支持的接入方案

最近更新时间：2026-09-11 20:57:30

概述

DataBuddy 数据接入提供从外部系统将数据导入数据湖仓的统一入口，覆盖**离线同步** 、**实时同步** 、**文件上传** 三类场景，支持单表、批量单表、整库多表、分库分表等多种粒度。无论数据源类型与同步频率如何，统一遵循「连接数据源 → 选择来源对象 → 配置目标 → 设置运行策略 → 发布运行」的主体流程。

接入方案总览

DataBuddy 提供以下 5 类接入方案，您可以根据数据源类型、数据规模与时效性要求选择合适的方案：

接入方案	典型场景	入口	配套文档
离线数据同步	周期性批量同步关系型数据库、大数据存储、消息队列等数据到数据湖仓	数据接入 → 任务管理 → 创建同步任务 → 离线单表同步 / 离线批量同步	创建单表同步任务 / 批量创建单表同步任务
实时数据同步	实时同步关系型数据库 CDC、消息队列数据，支持 DDL 变更同步	数据接入 → 任务管理 → 创建同步任务 → 实时整库多表同步 / 实时分库分表同步	实时整库多表任务配置 / 实时分库分表任务配置
文件上传到表	将本地或 COS 文件解析后落入结构化表	数据接入 → 文件上传 → 文件到表 / COS 文件到表	本地上传到表 / COS 文件上传到表
文件上传到卷	将文件以原始格式上传到 Volume，保留原文件结构	数据接入 → 文件上传 → 文件到卷	本地文件上传到卷
文件批量同步到卷	在离线同步任务中，将结构化数据按指定格式写入 Volume；或将非结构化文件批量同步到 Volume	数据接入 → 任务管理 → 创建同步任务（目标端选择 Volume 类型 Catalog）	文件到 Volume 同步

通用接入流程

无论选择哪种接入方案，您都需要完成以下步骤：

1. 准备数据源连接（数据源配置）
   ↓
2. 准备计算资源（实时接入用实时集群，离线接入用平台集群）

↓
3. 创建同步任务并选择同步方式
↓
4. 配置来源（选库 / 选表 / 选文件路径，必要时配置筛选条件、过滤规则）
↓
5. 配置目标（Catalog / Schema / Table / Volume，必要时配置自动建表或手动建表）
↓
6. 配置字段映射与运行参数（脏数据阈值、并发数、限速、checkpoint 间隔等）
↓
7. 配置告警与运行策略
↓
8. 保存 / 发布 / 启动任务

离线同步与实时同步对比

维度	离线同步	实时同步
同步粒度	单表、批量单表	整库多表、分库分表
触发方式	由工作流调度或手动调试运行	启动后持续运行，捕获源端 CDC 变更
来源端读取模式	全量读取、增量读取（按筛选条件）	全量 + 增量、仅增量
写入模式	overwrite / append / upsert	append/upsert
自动建表	支持手动建表	支持自动建表（Kafka 等部分来源端除外）
Schema 变更同步	不涉及（按任务执行时刻的 Schema）	支持 DDL 变更同步（MySQL / 腾讯云 MySQL / TDSQL-C MySQL 完整支持）
数据对账	支持全量、增量数据对账，按条数对账	支持全量、增量数据对账，按条数 / 内容对账
适用场景	定时数仓刷新、ODS 层加载、报表数据准备	数据库实时入湖、CDC 流式分析

文件接入能力

文件接入为非结构化或半结构化数据进入湖仓提供入口，覆盖以下 3 类用法：

入口	说明	适用文件类型

文件到表	上传文件后自动解析为结构化表，落入指定 Catalog 下的 Schema	.csv / .tsv / .json / .jsonl / .avro / .parquet / .txt / .xml
文件到卷	文件以原格式落入 Volume，保留二进制 / 原始结构	任意类型
COS 文件到表	从腾讯云对象存储 COS 选择文件，解析为结构化表	.csv / .json / .xml / .avro / .parquet / .txt

文件上传支持文件级解析参数（行、列分隔符、表头、转义符、JSON 嵌套、XML 行标签等），上传后表会被自动注册到 Catalog 中。

接入方式选型建议

您的需求	推荐方案
业务库每日定时同步到数据湖仓	离线单表同步 + Workflow 调度
多张同源表批量同步到湖仓	离线批量同步任务
业务库 CDC 实时入湖	实时整库多表同步任务
多个分库分表合并到一张目标表	实时分库分表同步任务
一次性导入历史 CSV / Parquet 文件做分析	文件到表
文件原样归档到湖仓	文件到卷

前提条件

在使用任何接入方案前，请确保已经完成以下准备：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关权限**（详见 成员与权限管理）。
- 已创建用于数据接入的计算资源（详见 创建计算资源）。
- 已根据来源端类型创建并测试通过对应的数据源连接（详见 添加数据源连接）。
- 目标端已规划好 Catalog 与 Schema（详见 统一元数据管理架构）。

使用限制

限制项	说明
计算资源	实时接入任务仅支持选择 **实时集群** 计算资源；离线接入任务仅支持选择 **平台集群** 计算资源
来源端类型	完整数据源支持范围参见 支持的数据源与读写能力

目标端类型	当前结构化任务的目标端固定为 catalog 数据表（基于 Iceberg 原生表）；非结构化数据写入支持 catalog Volume 卷
文件上传容量	文件到表：单次最多 10 个文件，总容量 ≤ 2 GB；文件到卷：单次最多 100 个文件，单文件 ≤ 5 GB

相关文档

- 支持的数据源与读写能力
- 数据源配置
- 创建单表同步任务
- 实时整库多表任务配置
- 本地上传到表
- 创建计算资源

支持的数据源与读取能力

最近更新时间：2026-09-11 19:09:00

概述

DataBuddy 数据接入持续扩展所支持的数据源类型，覆盖**关系型数据库** 、**大数据存储** 、**半结构化数据** 、**NoSQL 数据库** 、**消息队列** 等类别。本文按类别列出当前支持的数据源以及在不同接入链路（离线同步、实时整库同步、分库分表同步）中的读取能力支持情况。
数据源以「数据源连接」的方式注册到 DataBuddy，详见 添加数据源连接 。同步任务的目标端固定为 Catalog 数据表（基于 Iceberg 原生表）或 Volume。

读取能力图例

标记	含义
✓	支持
–	当前不支持

实时同步链路对来源端版本通常有更严格的要求，详细版本支持范围以「支持版本」列为准。

关系型数据库

数据源	离线读	实时整库读	分库分表读	支持版本
MySQL	✓	✓	✓	5.6 / 5.7 / 8.0.x
腾讯云 MySQL	✓	✓	✓	同 MySQL
TDSQL-C MySQL	✓	✓	✓	5.7 / 8.0
TDSQL MySQL	✓	✓	–	InnoDB 类型
PostgreSQL	✓	✓	✓	9.6 / 10 / 11 / 12
SQL Server	✓	✓	✓	2008 / 2008R2 / 2012 / 2014 / 2016 / 2017 / 2019
Oracle	✓	✓	✓	11 / 12 / 19
OceanBase	✓	✓	–	3.1.4 / 4.2.1.3

IBM DB2	✓	–	–	—
达梦 DM	✓	–	–	7 / 8
GBase	✓	–	–	—
GaussDB	✓	–	–	—
Greenplum	✓	–	–	4.x / 5.x / 6.x
TiDB	✓	–	–	—

大数据存储

数据源	离线读	实时整库读	分库分表读	支持版本
Hive	✓	–	–	1.x / 2.x / 3.x
HDFS	✓	–	–	2.x / 3.x
HBase	✓	–	–	2.2.x
TCHouse-D	✓	–	–	—
TCHouse-C	✓	–	–	—
TCHouse-X	✓	–	–	—
TCHouse-P	✓	–	–	—
Doris	✓	–	–	0.15 / 1.x
Iceberg	✓	–	–	0.13.1+
StarRocks	✓	–	–	—
ClickHouse	✓	–	–	20.7+
Kudu	✓	–	–	—
Hudi	✓	–	–	—
TDengine	✓	–	–	—
BigQuery	✓	–	–	—

半结构化数据

半结构化数据源主要承担非结构化文件接入与离线读取场景。结合 文件到表 与目标端为 Volume 的同步任务，可以实现「原样归档」「解析入表」「格式转换」等多种用法。

数据源	离线读	实时读	说明
COS	✓	–	腾讯云对象存储；支持作为离线同步源端，也作为「COS 文件到表」入口
Amazon S3	✓	–	海外对象存储；支持文件路径正则匹配与全量 / 增量同步
Azure Blob	✓	–	通过容器路径访问
FTP	✓	–	仅离线同步链路
SFTP	✓	–	支持密码与 SSH 私钥认证
File System	✓	–	自有文件系统
Rest API	✓	–	支持 GET / POST、Header、单次 / 多次请求；支持单条 JSON 与数组 JSON

NoSQL 数据库

数据源	离线读	实时整库读	支持版本
MongoDB	✓	✓	4.x
腾讯云 MongoDB	✓	✓	同 MongoDB
Elasticsearch	✓	–	6.x / 7.x
腾讯云 Elasticsearch	✓	–	同 Elasticsearch
CTSDB / InfluxDB	✓	–	—

消息队列

数据源	离线读	实时整库读	支持版本
Kafka	✓	✓	2.4.1 / 2.7.1 / 2.8.1 / 2.8.2
CKafka	✓	✓	同 Kafka

查看实际可用的数据源类型

以上列表为参考。**进入数据源管理或在数据接入任务中创建数据源连接时，可在「数据源类型」下拉列表中查看当前版本实际支持的完整列表** 。各数据源的具体连接配置项详见 数据源列表 中对应章节。

相关文档

- 支持的接入方案
- 数据源配置
- 添加数据源连接
- 实时整库多表任务配置
- 创建单表同步任务

数据源管理

最近更新时间：2026-09-11 20:42:30

概述

数据源（Data Source）是数据接入任务的连接入口。在 DataBuddy 中，数据源以**全局对象** 的形式由平台统一管理，不绑定到某一个具体的接入任务——一份数据源定义可以被多个离线同步、实时同步、文件上传任务、以及外部 Catalog 复用。

为避免数据源相关内容在数据接入模块与平台管理模块中重复维护，DataBuddy 把数据源的**创建、编辑、连通性测试、权限管理** 等通用能力统一收敛到 数据源管理 章节，本节作为「数据接入」视角下的入口指引，引导您到平台管理中完成具体操作。

在数据接入中如何用到数据源

接入场景	数据源在其中的角色	相关文档
离线数据同步（单表/批量/整库）	同步任务的**来源端** 或**目标端** 连接	离线数据同步
实时数据同步（整库/分库分表）	同步任务的**来源端** 连接（目标端一般为内置 DataBuddy 数据源）	实时整库多表任务配置
文件上传到表 / 卷	COS 类型上传需要 COS 数据源；本地上传不需要预先注册数据源	文件上传
数据源类型与读写能力	在任务配置前确认所选数据源类型是否支持作为来源 / 目标	支持的数据源与读写能力

数据源的核心操作（在平台管理中完成）

数据源的所有管理操作均在「平台管理 > 数据源管理」中进行，具体能力如下：

能力	入口
创建 / 查看 / 编辑 / 删除数据源，管理 ACL 权限与关联任务	全局数据源配置
配置数据源连接（数据源类型、连接参数、认证方式、SSM 凭据托管）	添加修改数据源连接
通过指定计算资源对数据源做连通性测试，验证网络可达性与凭据有效性	数据源连通性测试
内置 DataBuddy 数据源（默认目标端，不可编辑）	数据源管理概述
计算资源网络打通（数据源位于私有 VPC 时）	计算资源网络设置

各类型数据源的连接字段

在创建数据源时，**数据源类型** 下拉列表展示当前版本实际支持的完整类型清单。各数据源类型在「数据接入」视角下的字段说明与读写能力，详见：

- 支持的数据源与读写能力
- 数据源列表 及其下各具体数据源文档（MySQL / TDSQL-C MySQL / Oracle / PostgreSQL / SQL Server / MongoDB / Kafka 等）

推荐操作流程

首次接入一个外部数据源，建议按以下顺序：

1. **类型评估**：在 支持的数据源与读写能力 中确认所需的数据源类型是否支持您计划的同步方向（作为来源、作为目标）。
2. **网络打通**（若数据源位于您的私有 VPC）：先完成 计算资源网络设置 。
3. **创建数据源**：在 全局数据源配置 中创建数据源连接。
4. **连通性测试**：在创建、编辑数据源页面指定计算资源并 完成连通性测试 。
5. **在数据接入任务中引用**：根据接入需求选择 离线数据同步 或 实时数据同步 中的对应任务类型。

相关文档

- 平台管理：数据源管理概述
- 数据接入概览：支持的接入方案
- 数据源类型与读写能力：支持的数据源与读写能力
- 计算资源网络设置：计算资源网络设置
- 成员与权限管理：成员与权限管理

数据源列表
数据源概览

最近更新时间：2026-09-11 20:42:32

概述

DataBuddy 数据接入持续扩展所支持的数据源类型，覆盖**关系型数据库**、**大数据存储**、**半结构化数据**、**NoSQL 数据库**、**消息队列**五大类别。本文汇总当前支持的数据源类型，以及各数据源在不同数据接入链路中的支持情况，帮助您快速判断目标数据源是否可用、并选择合适的接入方式。
数据源以「数据源连接」的方式注册到 DataBuddy，详见 数据源管理 。各数据源的具体连接配置项，参见 数据源列表 中对应章节。
DataBuddy 数据源提供**离线同步**、**实时同步**、**文件上传**、**直连分析** 四类接入场景。其中离线同步、实时同步按「读（来源端接入）」「写（目标端写入）」两个方向区分，DataBuddy 以读为主，目标端固定为内置 DataBuddy TCLake。

数据源类型总览

下表按数据源类型汇总各数据源的接入能力。**离线集成**、**实时集成**按「读」「写」两个方向区分：读表示该数据源可作为来源端接入，写表示可作为目标端写入；DataBuddy 以读为主，目标端固定为内置 DataBuddy TCLake。**直连分析**表示该数据源支持 Linked Catalog 直连查询（不落湖、直接分析源数据）。

数据源类型	数据源	离线集成读	离线集成写	实时集成读	实时集成写	直连分析
关系型数据库	MySQL	✓	✓	✓	–	✓
关系型数据库	腾讯云 MySQL	✓	–	✓	–	✓
关系型数据库	TDSQL-C MySQL	✓	–	✓	–	✓
关系型数据库	PostgreSQL	✓	–	✓	–	✓
关系型数据库	Oracle	✓	–	✓	–	–
关系型数据库	SQL Server	✓	–	✓	–	–

关系型数据库	TDSQL MySQL	✓	–	–	–	✓
关系型数据库	TDSQL PostgreSQL	✓	–	–	–	–
关系型数据库	IBM DB2	✓	–	–	–	–
关系型数据库	达梦 DM	✓	–	–	–	–
关系型数据库	Greenplum	✓	–	–	–	–
关系型数据库	GBase	✓	–	–	–	–
关系型数据库	GaussDB	✓	–	–	–	✓
关系型数据库	TiDB	✓	–	–	–	–
关系型数据库	OceanBase	✓	–	–	–	–
大数据存储	HDFS	✓	–	–	–	–
大数据存储	HBase	✓	✓	–	–	–
大数据存储	Hive	✓	–	–	–	–
大数据存储	Kudu	✓	–	–	–	–
大数据存储	ClickHouse	✓	–	–	–	–
大数据存储	TCHouse-C	✓	–	–	–	–
大数据存储	TCHouse-D	✓	–	–	–	✓
大数据存储	TCHouse-X	✓	–	–	–	–
大数据存储	TCHouse-P	✓	–	–	–	–
大数据存储	Doris	✓	–	–	–	✓
大数据存储	StarRocks	✓	–	–	–	✓

大数据存储	Iceberg	✓	–	–	–	–
大数据存储	Hudi	✓	–	–	–	–
大数据存储	TDengine	✓	–	–	–	–
大数据存储	BigQuery（海外）	✓	–	–	–	–
半结构化数据	FTP	✓	–	–	–	–
半结构化数据	SFTP	✓	–	–	–	–
半结构化数据	COS	✓	✓	–	–	–
半结构化数据	REST API	✓	–	–	–	–
半结构化数据	File System（海外）	✓	–	–	–	–
半结构化数据	Amazon S3（海外）	✓	–	–	–	–
NoSQL 数据库	MongoDB	✓	–	✓	–	–
NoSQL 数据库	腾讯云 MongoDB	✓	–	✓	–	–
NoSQL 数据库	Elasticsearch	✓	–	–	–	–
NoSQL 数据库	腾讯云 Elasticsearch	✓	–	–	–	–
NoSQL 数据库	CTSDB / InfluxDB	✓	–	–	–	–
消息队列	Kafka	✓	–	✓	–	–
消息队列	CKafka	✓	–	✓	–	–

相关文档

- 支持的接入方案
- 支持的数据源与读取能力
- 数据源管理
- 实时整库多表任务配置
- 实时分库分表任务配置

Kafka 数据源

最近更新时间：2026-09-11 20:42:31

概述

Kafka 是分布式流处理平台。DataBuddy 数据接入支持将 Kafka（含腾讯云 CKafka）作为来源端读取消息，覆盖离线读取与实时整库场景。本文按「建数据源 → 配置同步任务 → 处理常见问题」的顺序介绍其使用方式。

支持的版本

类型	支持版本
自建 Kafka	2.x / 3.x
腾讯云 CKafka	2.4.1 / 2.8.1 / 3.2.3

读取能力

能力	离线读	实时整库读
支持	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

序列化格式支持

场景	读取格式
整库同步	`canal-JSON` / `debezium` / `ogg-json` / `csv` / `json` / `raw`

> 整库同步的 DDL 变更仅支持**新增列、删除列、新增表** 。

创建数据源

操作步骤

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入** > **任务管理** > **数据源管理** （或 **平台管理** > **数据源管理** ）。
4. 点击 **添加数据源** ，选择 **Kafka** 。

5. 填写连接配置与认证信息，详见 参数说明 。
6. 点击 **测试连接**（参见 数据源连通性测试 ）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务** 。

参数说明

参数	说明	是否必填
数据源名称	数据源在工作空间内的唯一标识	是
Kafka 类型	**开源自建 Kafka / 腾讯云 CKafka**	是
Bootstrap Servers	`host1:port1,host2:port2`（如 `10.0.0.1:9092,10.0.0.2:9092` ）	是
认证方式	见 认证方式	否
用户名 / 密码	选择 SASL 认证时必填（密码支持 SSM 凭证托管）	视认证

认证方式

方式	说明	必填字段
无认证（PLAINTEXT）	不开启认证	—
SASL_PLAINTEXT	明文 SASL 认证	用户名、密码
SASL_GSSAPI（Kerberos）	Kerberos 认证	keytab 文件、conf 文件、principal、Kafka 服务名称
SASL_SSL	SASL + SSL 双重认证	Truststore 认证文件、Keystore 密码、用户名、密码
SSL	TLS 双向认证	Truststore 认证文件、Truststore 密码、Keystore 认证文件、Keystore 密码、私钥口令

网络与服务端预备

- 集成资源组到所有 broker 必须**双向连通** 。
- 服务端 `server.properties` 中的 `listeners` / `advertised.listeners` 必须填**对外可达** 地址（很多连接失败的根因在此）。

在数据接入任务中使用

实时同步：来源端

参数	说明
数据源	已创建的 Kafka 数据源,支持连通性测试
Topic	要消费的 Topic，支持下拉多选
序列化格式	见 序列化格式支持
读取位置	**最早**（earliest）/ **最新**（latest）/ **指定时间**
订阅类型	可勾选 INSERT / UPDATE / DELETE，至少勾选一种

嵌套 JSON

默认不支持嵌套 JSON，开启需在高级参数中加：

```
source.nested.json.enabled=true
```

> **提示**
> 该参数同时作用于 Kafka **key 和 value** 。如果 key 不是 JSON（普通字符串、数字主键），需要再追加 `source.key.nested.json.enabled=false`，否则 key 解析会报错。

引用嵌套字段时，使用 `父字段->子字段` 形式声明，多层嵌套用多个 `->` 连接。在函数中引用必须用反引号包裹：

```
TO_TIMESTAMP(`book->published`, 'yyyyMMdd')
```

高级参数

除下表所列参数外，所有 Kafka 原生客户端参数加 `properties.` 前缀均可透传，例如 `properties.fetch.max.wait.ms=500` 。

参数	说明	默认
`source.nested.json.enabled`	启用嵌套 JSON（同时作用于 key 和 value）	`false`
`source.key.nested.json.en`	是否对 key 启用嵌套 JSON	跟随 `source.nested.j`

abled		son.enabled
source.ignore-parse-errors	解析失败时跳过错误记录	false
scan.topic-partition-discovery.interval	周期性发现新 partition / 新 Topic 的间隔；填 0 关闭	5min
timestamp.format	时间字段解析格式	ISO8601

数据格式样例

canal-json

```
{
  "data": [{ "id": "1", "name": "张三", "age": "25" }],
  "database": "test_db",
  "table": "user_info",
  "type": "INSERT",
  "es": 1589373560000,
  "ts": 1589373560798,
  "isDdl": false,
  "mysqlType": { "id": "int", "name": "varchar(50)", "age": "int" },
  "sqlType": { "id": 4, "name": 12, "age": 4 },
  "pkNames": ["id"],
  "old": null,
  "sql": ""
}
```

debezium

```
{
  "before": null,
  "after": { "id": 1, "name": "张三", "age": 25 },
  "source": {
    "version": "1.5.0.Final",
    "connector": "mysql",
    "name": "mysql_server",
    "ts_ms": 1589373560000,
    "db": "test_db",
```

```
    "table": "user_info"
  },
  "op": "c",
  "ts_ms": 1589373560798
}
```

常见问题

Q：连接失败、测试连接超时?

A：按以下顺序排查：

1. 集成资源组到 broker 网络是否双向通（VPC、白名单、安全组）。
2. `advertised.listeners` 返回的地址在资源组所在网络是否可达。
3. SASL 认证下用户名密码、SASL 机制是否与服务端一致。

Q：消费不到数据?

可能原因	怎么确认
读取位置选了 latest 但 Topic 没新消息	改成 earliest 或指定时间点重启
Topic 实际无数据	用 Kafka 控制台直接 consume 一次
Consumer group 在外部已有持久化 offset	换一个 group ID 重启
「过滤操作」/「库表范围」把数据全过滤掉	检查同步任务的对应配置

Q：序列化格式解析失败、字段全为 null?

A：按以下顺序排查：

1. 用消费工具拉一条原始消息，对比配置的序列化格式。
2. 消息为嵌套 JSON 但**未开启** `source.nested.json.enabled=true` 。
3. key 不是 JSON 但启用了嵌套 JSON，需补 `source.key.nested.json.enabled=false` 。

Q：整库报 mysqlType / sqlType / pkNames 缺失?

A： `canal-json` / `debezium` 整库需消息携带这些元字段，确认上游 Canal / Debezium 输出包含。用工具抓一条消息核对即可。

Q：嵌套字段函数处理报 SQL 语法错误?

A： `-` 与 `>` 不是合法 SQL 标识符。在函数中引用嵌套字段必须用反引号包裹，并确认字段已通过「添加字段」声明：

```
TO_TIMESTAMP(`book->published`, 'yyyyMMdd')
```

相关文档

- 添加数据源连接
- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

MongoDB 数据源

最近更新时间：2026-09-11 20:42:30

概述

MongoDB 是一款高性能、开源的 NoSQL 文档数据库，以灵活的文档模型与强大的横向扩展能力著称，广泛应用于 Web 应用、移动应用与大数据分析场景。DataBuddy 数据接入支持将 MongoDB 作为来源端进行实时整库同步与离线同步。本文介绍其连接配置、能力支持范围、前提条件与常见问题。

支持的版本

类型	支持版本
自建 MongoDB	3.6 / 4.0 / 4.2 / 4.4 / 5.0 / 6.0 / 7.0 / 8.0
腾讯云数据库 MongoDB	3.6 / 4.0 / 4.2 / 4.4 / 5.0 / 6.0 / 7.0 / 8.0

实时同步（Change Streams）要求 MongoDB 版本 ⩾ 3.6。

读取能力

能力	离线读	实时整库读
支持	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制

限制项	说明
版本要求	MongoDB 版本必须 ⩾ 3.6
副本集	需配置为副本集（Replica Set）模式才能使用 Change Streams
oplog 保留	需要足够的 oplog 保留时间以支持增量同步
系统集合	不支持 `system.*` 等系统集合的同步

离线同步限制

限制项	说明

单文档大小	不超过 16 MB（MongoDB 限制）
嵌套深度	建议控制文档嵌套深度，过深可能影响性能

支持的字段类型

完整 BSON 类型请参见 MongoDB 官方文档 。

字段类型	离线读	实时读
Int32 / Int64 / Long / Double / Decimal128	✓	✓
Boolean / String / ObjectId / UUID / Symbol / Regex	✓	✓
JavaScript / BinData	✓	✓
Date / Timestamp	✓	✓
Object / Array / Null / Undefined / MinKey / MaxKey / DBPointer	✓	✓

MongoDB 为 Schema-free 数据库，同一字段可能存在多种类型，建议在源端统一字段类型以避免同步异常。

实时同步支持的 DML / DDL

DML

操作	是否支持	说明
INSERT / UPDATE / REPLACE / DELETE	✓	文档级变更操作

DDL

操作	是否支持
DROP COLLECTION / DROP DATABASE / RENAME COLLECTION	✗

MongoDB 基于 Change Streams 进行变更捕获，不支持 DDL 变更的自动同步。集合的删除、重命名等操作需要在目标端手动处理。

前提条件

1. 确认副本集配置（仅实时同步）

```
rs.status()
```

2. 配置账号权限

离线同步：

```
use admin
db.createUser({
  user: "wedata_user",
  pwd: "your_password",
  roles: [{ role: "read", db: "your_database" }]
})
```

实时同步：

```
use admin
db.createUser({
  user: "wedata_cdc_user",
  pwd: "your_password",
  roles: [
    { role: "read", db: "your_database" },
    { role: "read", db: "local" }    // 读取 oplog 必需
  ]
})
```

3. 配置 oplog 大小（仅实时同步）

```
use local
db.oplog.rs.stats()                                    // 查看当前 oplog 配
置
db.adminCommand({ replSetResizeOplog: 1, size: 10240 })    // 单位 MB
```

> 修改 oplog 大小需要在副本集每个节点执行。

创建数据源

操作步骤

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入 > 任务管理 > 数据源管理**（或 **平台管理 > 数据源管理** ）。

4. 点击 **添加数据源** ，选择 **MongoDB** 。
5. 填写连接配置与认证信息，详见 **参数说明** 。
6. 点击 **测试连接** （参见 **数据源连通性测试** ）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务** 。

参数说明

参数	说明	是否必填
数据源名称	数据源在工作空间内的唯一标识	是
连接地址	MongoDB 连接地址（见下方格式）	是
数据库	默认连接的数据库	是
用户名	数据库连接账号（如开启认证）	视认证情况
密码	数据库连接密码（如开启认证；支持 SSM 凭证托管）	视认证情况
数据源版本	见 **支持的版本**	是

连接地址格式

部署方式	格式示例
单节点	`mongodb://host:port`
副本集	`mongodb://host1:port1,host2:port2,host3:port3/?replicaSet=rsName`
分片集群	`mongodb://mongos1:port1,mongos2:port2`

完整示例：

```
mongodb://10.0.0.1:27017,10.0.0.2:27017,10.0.0.3:27017/?replicaSet=rs0
mongodb://10.0.0.1:27017/mydb?authSource=admin
```

在数据接入任务中使用

实时整库

配置项	说明
来源表范围	**指定集合**（仅同步选中的几个，新增集合需重启）/ **正则匹配**（指定库名 + 集合名正则，符合规则的新增集合自动接入）

读取模式	**全量 + 增量 / 仅增量**
订阅类型	可勾选 INSERT / UPDATE / DELETE，至少勾选一种

离线单表

参数	说明
库 / 集合	必填；支持选择或手动输入
筛选条件	MongoDB 查询条件（JSON 格式）

筛选条件示例：

```
{"age": {"$gte": 18}, "status": "active"}
{"create_time": {"$gte": {"$date": "2024-01-01T00:00:00Z"}}}
```

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

- **账号权限最小化：** 使用专用账号，避免 root 权限。
- **oplog 保留时间：** 建议 ⩾ 72 小时，防止任务重启后数据丢失。
- **索引优化：** 离线同步时在查询字段上建索引以提升读取效率。
- **网络打通：** 优先使用内网连接。

常见问题

Q：连接失败?

A：检查网络连通性、确认用户名密码正确、检查 `authSource` 配置。

Q：报错 Change Streams 不可用?

A：MongoDB 未配置为副本集模式或版本 < 3.6。请将 MongoDB 配置为副本集，并升级到 3.6 及以上版本。

Q：增量同步报错找不到 oplog?

A：oplog 保留时间太短，已被覆盖。增加 oplog 大小：

```
db.adminCommand({ replSetResizeOplog: 1, size: 20480 })
```

Q：同步过程中报错类型不匹配?

A：MongoDB 是 Schema-free 的，同一字段可能存在不同类型。建议在源端统一数据类型，或在写入端配置类型转换。

Q：某些文档同步失败?

A：单文档超过 16 MB 限制。请拆分大文档，或使用 GridFS 存储大文件。

相关文档

- 添加数据源连接
- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

MySQL 数据源

最近更新时间：2026-09-11 20:42:31

概述

MySQL 是广泛使用的开源关系型数据库。DataBuddy 数据接入支持将 MySQL 作为来源端，覆盖离线同步、实时整库同步、分库分表同步等多种场景。本文介绍 MySQL 数据源在 DataBuddy 中的连接配置、能力支持范围、前提条件与常见问题。

支持的版本

类型	支持版本
自建 MySQL	5.6.x / 5.7.x / 8.0.x / 8.1.x / 8.2.x / 8.3.x / 8.4.x
腾讯云数据库 MySQL（CDB）	5.6 / 5.7 / 8.0
JDBC Driver	8.0.21

读取能力

能力	离线读	实时整库读	分库分表读
支持	✓	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制

限制项	说明
Binlog 格式	仅支持 ROW 格式
`binlog_row_image`	必须设置为 FULL
Binlog 保留时间	建议至少保留 72 小时；全量数据量越大需保留越长
无主键表	无主键表无法保证 exactly-once 语义，可能存在数据重复
XA 事务	不支持 XA ROLLBACK，需手动处理相关表
只读库	不支持 5.6.x 以下版本的只读库实例

Functional Index	不支持包含 Functional Index 的表（MySQL 8.0 新特性）
级联删除	不支持同步级联删除的关联表记录
存储过程、物化视图	不支持读取
SET 数据类型	实时同步不支持
在线 DDL 工具	仅支持 gh-ost，不支持其他 DDL 在线变更工具

离线同步限制

限制项	说明
分表同步并发	多表同步切分单表时，任务并发数需大于表个数
存储过程	不支持读取

支持的字段类型

完整字段类型请参见 MySQL 官方文档 。以下是 DataBuddy 在不同链路上的支持情况（节选 MySQL 8.0.x）。

字段类型	离线读	实时读
TINYINT / TINYINT UNSIGNED	✓	✓
SMALLINT / MEDIUMINT / INT / BIGINT（含 UNSIGNED）	✓	✓
FLOAT / DOUBLE / DECIMAL / NUMERIC	✓	✓
REAL	✗	✓
CHAR / VARCHAR / TEXT 系列	✓	✓
JSON	✓	✓
BINARY / VARBINARY / BLOB 系列	✓	✓
DATE / TIME / DATETIME / TIMESTAMP / YEAR	✓	✓
BIT / BOOL / BOOLEAN / ENUM	✓	✓
SET	✓	✗
GEOMETRY / POINT	✗	✓

LINESTRING / POLYGON / MULTIPOINT 等空间类型	✕	✕

空间类型（GEOMETRY 系列）支持有限，建议转换为字符串类型后再同步。

实时同步支持的 DML / DDL

DML

操作	是否支持
INSERT / DELETE / UPDATE	✓

DDL

操作	是否支持	支持的语法示例
新增列（ADD COLUMN）	✓	`ALTER TABLE t ADD col TYPE [FIRST | AFTER col]`
删除列（DROP COLUMN）	✓	`ALTER TABLE t DROP [COLUMN] col`
重命名列（RENAME / CHANGE COLUMN）	✓	`ALTER TABLE t CHANGE old new TYPE`
修改列类型（MODIFY COLUMN）	✓	`ALTER TABLE t MODIFY col TYPE`
新增表（CREATE TABLE）	✓	不支持 CHECK、Temporary Table
重命名表 / 删除表 / 清空表	✕	

DDL 变更响应**仅在增量阶段** 生效；**删除表、重命名表** 这类高危操作目标端通常不支持自动响应。
详细的 DDL 变更响应策略（自动变更、忽略、日志告警）参见 **实时整库多表任务配置** 。

前提条件

1. 配置账号权限

离线同步：

```
CREATE USER 'wedata_user'@'%' IDENTIFIED BY 'your_password';
GRANT SELECT ON your_database.* TO 'wedata_user'@'%';
FLUSH PRIVILEGES;
```

实时同步：

```
CREATE USER 'wedata_cdc_user'@'%' IDENTIFIED BY 'your_password';
GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT
    ON *.* TO 'wedata_cdc_user'@'%';
FLUSH PRIVILEGES;
```

> 建议为同步任务创建专用账号，避免与业务账号混用。

2. 开启 Binlog（仅实时同步）

```
SHOW VARIABLES LIKE 'log_bin';    -- ON 表示已开启
```

腾讯云 MySQL 默认开启 Binlog。自建实例需在 `my.cnf` 中配置：

```
[mysqld]
log-bin=mysql-bin
server-id=1
binlog_format=ROW
binlog_row_image=FULL
```

修改后需重启数据库使其生效。

3. 设置会话超时（推荐）

针对大库全量同步，建议放开会话超时：

```
SET GLOBAL interactive_timeout = 28800;
SET GLOBAL wait_timeout = 28800;
```

创建数据源

操作步骤

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入** > **任务管理** > **数据源管理** （或 **平台管理** > **数据源管理** ）。
4. 点击 **添加数据源** ，选择 **MySQL** 。
5. 填写连接配置与认证信息，详见 参数说明 。

6. 点击 **测试连接**（参见 数据源连通性测试）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务**。

参数说明

参数	说明	是否必填	默认值
数据源名称	数据源在工作空间内的唯一标识	是	—
连接方式	**串连接（JDBC URL）** / **腾讯云实例**	是	串连接
JDBC URL	串连接方式下的连接串，格式 `jdbc:mysql://host:port/database`	串连接必填	—
地域、实例	腾讯云实例方式下选择实例	腾讯云实例必填	—
用户名	数据库连接账号	是	—
密码	数据库连接密码（支持 SSM 凭证托管）	是	—
数据源版本	5.6 / 5.7 / 8.0.x	是	—
高级参数	连接级扩展参数，如 `useUnicode`、`characterEncoding`、`serverTimezone`、`connectTimeout`	否	—

JDBC URL 示例：

```
jdbc:mysql://10.0.0.1:3306/mydb?
useUnicode=true&characterEncoding=utf8&serverTimezone=Asia/Shanghai
```

在数据接入任务中使用

实时整库、分库分表

配置项	说明
来源表范围	**指定表**（仅同步选中的表，新增表需重启）/ **正则匹配**（指定库名 + 表名正则，符合规则的新增表自动接入）
读取模式	**全量 + 增量** / **仅增量**

订阅类型	可勾选 INSERT / UPDATE / DELETE，至少保留一种
gh-ost 同步	是否同步 gh-ost 在线 DDL 产生的临时表

离线单表

配置项	说明
数据来源	已配置的 MySQL 数据源
库 / 表	支持指定与正则匹配；表名支持范围如 `table_[0-99]`
切割键	仅支持整型字段；建议选择主键或带索引的列；避免 COLLATE 列
筛选条件	标准 WHERE 语法，支持时间参数（如 `gmt_create > $bizdate`）

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

实践	建议
账号权限最小化	为同步任务创建专用账号，仅授必要权限，不使用 root
Binlog 保留时间	至少 72 小时；大表全量同步建议更长
切割键选择	主键或索引列；避免 COLLATE 列以防数据重复
网络打通	优先使用内网连接，提升传输性能

主备切换处理（实时同步）

实时同步任务在 MySQL 主备切换场景下能否平滑恢复，取决于地址类型与是否开启 GTID：

场景	处理方式
腾讯云数据库 MySQL（实例地址不变）	闪断秒级，任务自动重连新主，从最近快照位点继续；故障 failover 自动完成，无需人工干预
自建 MySQL + VIP / DNS	VIP / DNS 漂移到新主后任务自动重连；建议任务重试次数 ⩾ 3
自建 MySQL + 固定 IP	切换前暂停任务，切换后修改数据源 IP，再继续运行

> **强烈建议开启 GTID**：主备切换场景下基于 GTID 恢复才能保证不重复、不丢数据。
> 异步复制下主库故障可能丢失最后几个未同步的事务，这是 MySQL 复制层本身的限制；如需零丢失，建议使

> 用强同步或半同步复制。

常见问题

Q：报错 `A slave with the same server_uuid/server_id as this slave has connected to the master`？

A：多个同步任务使用了相同的 server-id。系统已优化为随机生成 server-id，如旧任务在高级参数中显式配置了 `server-id`，请删除该参数。

Q：报错 `Cannot replicate because the master purged required binary logs`？

A：MySQL 上的 Binlog 文件已被清理。请增加 Binlog 保留时间，或优化作业并发以加速消费。

Q：报错 `Connection reset by peer` 或连接超时？

A：网络问题或作业反压导致空闲超时。可调大 `slave_net_timeout`（如 120），或优化作业并行度与内存。

Q：JobManager OOM？

A：数据量大或主键范围分布稀疏导致切分过多。可：

- 调大 JobManager CU。
- 调整高级参数 `scan.incremental.snapshot.chunk.size`（默认 8096）。
- 调整 `split-key.even-distribution.factor.upper-bound`。

Q：实时任务运行期间，将 MySQL 的 `binlog_format` 从 ROW 改为 STATEMENT 会怎样？

A：实时任务仅在启动时校验 `binlog_format=ROW`，运行期间不再检查。切换为 STATEMENT 后，所有 DML 都会被解析器视为 DDL，下游无法感知数据变更，造成静默数据缺失。**请在任务运行期间不要修改 `binlog_format`。**

Q：使用 COLLATE 列作为切割键，部分数据重复？

A：COLLATE 影响排序、筛选与分组结果。设置切割键时请选择非 COLLATE 列。

Q：是否支持 gh-ost 在线 DDL？

A：支持。开启 gh-ost 同步开关后，会监控并同步 `*_gho` 临时表的变更。前提条件：gh-ost 能正常访问 MySQL；腾讯云 CDB 需在 gh-ost 命令中追加 `--aliyun-rds` 参数。

相关文档

- 添加数据源连接

- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

Oracle 数据源

最近更新时间：2026-09-11 20:42:31

概述

Oracle 是企业级关系型数据库，广泛应用于金融、电信、政府等行业的核心业务系统。DataBuddy 数据接入支持将 Oracle 作为来源端，覆盖离线同步、实时整库同步、分库分表同步等场景。本文介绍其连接配置、能力支持范围、前提条件与常见问题。

支持的版本

类型	支持版本
Oracle	11g / 12c / 19c
JDBC Driver	ojdbc6

> 不支持 Oracle 19c 的新特性。

读取能力

能力	离线读	实时整库读	分库分表读
支持	✓	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制（基于 LogMiner）

限制项	说明
日志保留时间	增量日志至少保留 72 小时
备库支持	仅支持 logical standby，不支持 physical standby
无主键表	无主键表无法保证 exactly-once 语义
账号限制	严禁使用 system 或 sys 账号配置数据源
DDL 风险	加列带默认值可能导致解析异常；LogMiner 可能导致 DML 与 DDL 顺序错乱
性能瓶颈	LogMiner 读取超过 5k/s 时可能出现较大延迟

RAC 模式风险	日志文件切换可能导致 LogMiner 不感知，造成数据缺失

> 若无法接受上述限制，建议使用 OGG (Oracle GoldenGate) 方案。

离线同步限制

限制项	说明
视图	支持读取视图
编码格式	仅支持 `UTF8` / `AL32UTF8` / `AL16UTF16` / `ZHS16GBK`

支持的字段类型

完整字段类型请参见 Oracle 官方文档 。

字段类型	离线读	实时读
NUMBER / INTEGER / INT / SMALLINT / NUMERIC / DECIMAL	✓	✓
FLOAT / BINARY_FLOAT / DOUBLE PRECISION / BINARY_DOUBLE / REAL	✓	✓
CHAR / VARCHAR / VARCHAR2 / NCHAR / NVARCHAR2	✓	✓
CLOB / NCLOB	✓	✓
DATE / TIMESTAMP / TIMESTAMP WITH TIME ZONE / TIMESTAMP WITH LOCAL TIME ZONE	✓	✓
BLOB / RAW / LONG RAW	✓	✓
ROWID	×	✓
INTERVAL YEAR TO MONTH / INTERVAL DAY TO SECOND	×	✓
XMLTYPE / BFILE	×	×

> NUMBER 类型根据精度自动映射为 TINYINT / SMALLINT / INT / BIGINT 或 DECIMAL；精度未指定时默认映射为 DECIMAL(38, 18)。

实时同步支持的 DML / DDL

DML

操作	是否支持
INSERT / DELETE / UPDATE	✓

DDL

操作	是否支持
新增表 / 新增列 / 删除列 / 重命名列 / 修改列类型	✓
删除表 / 清空表 / 重命名表（RENAME TABLE）	✗

> DDL 变更响应**仅在增量阶段** 生效；加列带默认值可能导致解析异常，LogMiner 可能导致 DML 与 DDL 顺序错乱。

前提条件

1. 配置账号权限

离线同步：

```
CREATE USER wedata_user IDENTIFIED BY your_password;
GRANT CREATE SESSION TO wedata_user;
GRANT SELECT ANY TABLE TO wedata_user;
```

实时同步：

```
CREATE USER wedata_cdc_user IDENTIFIED BY your_password DEFAULT
TABLESPACE USERS QUOTA UNLIMITED ON USERS;
GRANT CREATE SESSION, SELECT ANY TABLE, SELECT_CATALOG_ROLE TO
wedata_cdc_user;
GRANT ANALYZE ANY TO wedata_cdc_user;
GRANT EXECUTE_CATALOG_ROLE TO wedata_cdc_user;
GRANT LOGMINING TO wedata_cdc_user;                -- 12c 及以上版本需要，11g
不需要
GRANT EXECUTE ON DBMS_LOGMNR TO wedata_cdc_user;
GRANT EXECUTE ON DBMS_LOGMNR_D TO wedata_cdc_user;

-- 系统视图查询权限
GRANT SELECT ON V_$DATABASE TO wedata_cdc_user;
GRANT SELECT ON V_$LOG TO wedata_cdc_user;
```

```
GRANT SELECT ON V_$LOG_HISTORY TO wedata_cdc_user;
GRANT SELECT ON V_$LOGMNR_CONTENTS TO wedata_cdc_user;
GRANT SELECT ON V_$LOGMNR_LOGS TO wedata_cdc_user;
GRANT SELECT ON V_$ARCHIVED_LOG TO wedata_cdc_user;
GRANT SELECT ON V_$ARCHIVE_DEST_STATUS TO wedata_cdc_user;
```

> ⓘ **注意**
> 监控表字段变更时，**严禁使用** `system` **或** `sys` **账号配置数据源。**

2. 开启日志归档（仅实时同步）

需要重启数据库，并占用磁盘空间。

```
sqlplus / as sysdba
ALTER SYSTEM SET db_recovery_file_dest_size = 10G;
ALTER SYSTEM SET db_recovery_file_dest =
'/opt/oracle/oradata/recovery_area' scope=spfile;
SHUTDOWN IMMEDIATE;
STARTUP MOUNT;
ALTER DATABASE ARCHIVELOG;
ALTER DATABASE OPEN;
ARCHIVE LOG LIST;      -- 验证归档模式
```

3. 开启补充日志（仅实时同步）

未开启会导致 UPDATE 操作下游其他列为空。

```
ALTER DATABASE ADD SUPPLEMENTAL LOG DATA;
ALTER TABLE schema_name.table_name ADD SUPPLEMENTAL LOG DATA (ALL)
COLUMNS;
```

创建数据源

操作步骤

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入 > 任务管理 > 数据源管理**（或 **平台管理 > 数据源管理** ）。
4. 点击 **添加数据源** ，选择 **Oracle** 。

5. 填写连接配置与认证信息，详见 **参数说明** 。
6. 点击 **测试连接** （参见 **数据源连通性测试** ）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务** 。

参数说明

参数	说明	是否必填
数据源名称	数据源在工作空间内的唯一标识	是
JDBC URL	`jdbc:oracle:thin:@host:port:sid` 或 `jdbc:oracle:thin:@//host:port/service_name`	是
用户名	数据库连接账号（不允许使用 system / sys ）	是
密码	数据库连接密码（支持 SSM 凭证托管 ）	是
数据源版本	11g / 12c / 19c	是

JDBC URL 示例：

```
jdbc:oracle:thin:@10.0.0.1:1521:ORCL
jdbc:oracle:thin:@//10.0.0.1:1521/ORCLPDB
```

在数据接入任务中使用

实时整库

配置项	说明
来源表范围	**指定表** （仅同步选中的表，新增表需重启）/ **正则匹配** （指定库名 + 表名正则，符合规则的新增表自动接入 ）
读取模式	**全量 + 增量** / **仅增量**

离线单表

参数	说明
库 / Schema / 表	必填；支持指定与正则匹配

切割键	建议主键或索引列；用于并行同步
筛选条件	标准 WHERE 语法

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

- **账号权限最小化**： 使用专用账号，避免 system / sys。
- **日志保留时间**： 归档日志至少保留 72 小时。
- **切割键选择**： 主键或索引列。
- **性能考量**： LogMiner 超过 5k/s 时建议使用 OGG 方案替代。

常见问题

Q：实时同步延迟较大?

A：LogMiner 读取超过 5k/s 时存在性能瓶颈。可优化作业并行度、减少同步表数量；高吞吐场景建议 OGG。

Q：加列带默认值后任务异常?

A：LogMiner 对带默认值的加列解析存在问题。建议先加列、再 UPDATE 默认值。

Q：RAC 环境下部分数据未同步?

A：日志文件切换可能导致 LogMiner 不感知。请增加日志保留时间，优化归档日志配置。

Q：UPDATE 后目标端其他列为空?

A：未开启补充日志。执行：

```
ALTER TABLE schema_name.table_name ADD SUPPLEMENTAL LOG DATA (ALL)
COLUMNS;
```

Q：连接超时或被拒绝?

A：检查网络连通性、监听器状态、防火墙端口。

相关文档

- 添加数据源连接
- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务

- 实时整库多表任务配置

PostgreSQL 数据源

最近更新时间：2026-09-11 20:42:31

概述

PostgreSQL 是功能强大的开源对象关系型数据库系统。DataBuddy 数据接入支持将 PostgreSQL 作为来源端，覆盖离线同步、实时整库同步与分库分表同步等场景。本文介绍其连接配置、能力支持范围、前提条件与常见问题。

支持的版本

类型	支持版本
自建 PostgreSQL	9.6 / 10.x / 11.x / 12.x / 13.x / 14.x / 15.x / 16.x
腾讯云 PostgreSQL	11.x / 12.x / 13.x / 14.x / 15.x / 16.x / 17.x / 18.x
JDBC Driver	postgresql-42.2.26

读取能力

能力	离线读	实时整库读	分库分表读
支持	✓	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制

限制项	说明
逻辑复制版本要求	PostgreSQL 10+ 才支持原生逻辑复制
`wal_level`	必须设置为 **logical**
复制槽	需要创建复制槽（Replication Slot）
无主键表	需设置 `REPLICA IDENTITY FULL`
发布（Publication）	需要创建发布来指定同步的表

离线同步限制

限制项	说明
视图	支持读取视图
切割键	仅支持整型字段
特殊字符表名	表名 / 字段名不支持数字开头或包含大小写字母、中划线

支持的字段类型

完整字段类型请参见 PostgreSQL 官方文档 。

字段类型	离线读	实时读
SMALLINT / SMALLSERIAL / INTEGER / SERIAL / BIGINT / BIGSERIAL	✓	✓
REAL / DOUBLE PRECISION / NUMERIC / DECIMAL / MONEY	✓	✓
BOOLEAN	✓	✓
CHAR / VARCHAR / TEXT / UUID	✓	✓
JSON / JSONB / ARRAY	✓	✓
DATE / TIME / TIMETZ / TIMESTAMP / TIMESTAMPTZ	✓	✓
INTERVAL	×	✓
BYTEA / BIT / BIT VARYING	✓	✓
CIDR / INET / MACADDR	×	×
POINT / LINE / LSEG / BOX / PATH / POLYGON / CIRCLE	×	×

实时同步支持的 DML / DDL

DML

操作	是否支持
INSERT / DELETE / UPDATE	✓

DDL

操作	是否支持
新增表 / 新增列 / 删除列 / 修改列类型	✓

删除表 / 清空表 / 重命名表（RENAME TABLE）	×

> PostgreSQL 逻辑复制对 DDL 变更的支持有限，Schema 变更后可能需要重新配置同步任务与发布。

前提条件

1. 配置账号权限

离线同步：

```
CREATE USER wedata_user WITH PASSWORD 'your_password';
GRANT CONNECT ON DATABASE your_database TO wedata_user;
GRANT USAGE ON SCHEMA public TO wedata_user;
GRANT SELECT ON ALL TABLES IN SCHEMA public TO wedata_user;
```

实时同步：

```
CREATE USER wedata_cdc_user WITH REPLICATION PASSWORD 'your_password';
GRANT CONNECT ON DATABASE your_database TO wedata_cdc_user;
GRANT USAGE ON SCHEMA public TO wedata_cdc_user;
GRANT SELECT ON ALL TABLES IN SCHEMA public TO wedata_cdc_user;
```

2. 配置 WAL 级别（仅实时同步）

```
SHOW wal_level;
```

腾讯云 PostgreSQL 在控制台「参数设置」中将 `wal_level` 设置为 `logical`。自建 PostgreSQL 修改 `postgresql.conf`：

```
wal_level = logical
max_replication_slots = 10
max_wal_senders = 10
```

修改后需要重启数据库才能生效。

3. 配置 pg_hba.conf（仅实时同步）

```
# TYPE  DATABASE        USER            ADDRESS                 METHOD
```

```
host    replication     wedata_cdc_user 10.0.0.0/8              md5
```

4. 创建发布（仅实时同步）

```
CREATE PUBLICATION wedata_pub FOR TABLE table1, table2;
-- 或：CREATE PUBLICATION wedata_pub FOR ALL TABLES;
```

5. 设置 REPLICA IDENTITY（无主键表）

```
ALTER TABLE your_table REPLICA IDENTITY FULL;
```

创建数据源

操作步骤

1. 登录 DataBuddy 控制台。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入 > 任务管理 > 数据源管理**（或 **平台管理 > 数据源管理**）。
4. 点击 **添加数据源**，选择 **PostgreSQL**。
5. 填写连接配置与认证信息，详见 **参数说明**。
6. 点击 **测试连接**（参见 **数据源连通性测试**）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务**。

参数说明

参数	说明	是否必填
数据源名称	数据源在工作空间内的唯一标识	是
JDBC URL	`jdbc:postgresql://host:port/database`	是
用户名	数据库连接账号	是
密码	数据库连接密码（支持 SSM 凭证托管）	是
数据源版本	12 / 13 / 14 / 15 / 16	是

JDBC URL 示例：

```
jdbc:postgresql://10.0.0.1:5432/mydb
```

```
jdbc:postgresql://10.0.0.1:5432/mydb?reWriteBatchedInserts=true
```

常用 URL 参数：

参数	说明
`reWriteBatchedInserts`	优化批量插入性能（推荐 true）
`ssl` / `sslmode`	SSL 连接配置
`connectTimeout`	连接超时时间（秒）

在数据接入任务中使用

实时整库

配置项	说明
来源表范围	**指定表**（仅同步选中的表，新增表需重启）/ **正则匹配**（指定库名 + 表名正则，符合规则的新增表自动接入）
读取模式	**全量 + 增量 / 仅增量**

离线单表

参数	说明
库 / Schema / 表	必填，Schema 默认 `public`；表支持范围 `table_[0-99]`
切割键	仅支持整型字段；建议主键或索引列
筛选条件	标准 WHERE 语法

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

- **账号权限最小化：** 使用专用账号，仅授必要权限。
- **WAL 日志保留：** `ALTER SYSTEM SET wal_keep_size = '1GB';`
- **切割键选择：** 主键或索引列。
- **网络打通：** 优先使用内网连接。

常见问题

Q：实时同步报错 `wal_level` 不支持？

A：未将 `wal_level` 设置为 `logical`。

```
ALTER SYSTEM SET wal_level = 'logical';
-- 修改后需重启数据库
```

Q：报错无法创建复制槽？

A：`max_replication_slots` 太小：

```
ALTER SYSTEM SET max_replication_slots = 20;
```

Q：无主键表 UPDATE / DELETE 无法正确同步？

A：未设置 `REPLICA IDENTITY`：

```
ALTER TABLE your_table REPLICA IDENTITY FULL;
```

Q：连接被拒绝？

A：`pg_hba.conf` 未允许 DataBuddy 网段连接。在文件中添加：

```
host    all             wedata_user     10.0.0.0/8      md5
host    replication     wedata_cdc_user 10.0.0.0/8      md5
```

Q：包含特殊字符的表名无法同步？

A：不支持表名、字段名以数字开头或包含大小写字母、中划线

相关文档

- 添加数据源连接
- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

SQL Server 数据源

最近更新时间：2026-09-11 20:42:30

概述

SQL Server 是 Microsoft 推出的关系型数据库管理系统，广泛应用于企业级应用与数据管理。DataBuddy 数据接入支持将 SQL Server 作为来源端，覆盖离线同步、实时整库同步与分库分表同步等场景。本文介绍其连接配置、能力支持范围、前提条件与常见问题。

支持的版本

类型	支持版本
自建 SQL Server	2012 / 2014 / 2016 / 2017 / 2019 / 2022
云数据库 SQL Server	2012 / 2014 / 2016 / 2017 / 2019 / 2022
JDBC Driver	mssql-jdbc-7.4.1.jre8

读取能力

能力	离线读	实时整库读	分库分表读
支持	✓	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制

限制项	说明
CDC 功能	需要在数据库与表上启用 CDC（Change Data Capture）
日志保留时间	CDC 日志至少保留 72 小时
无主键表	无主键表无法保证 exactly-once 语义
版本要求	CDC 仅在 Enterprise / Developer / Standard 版本可用
锁争用	建议开启快照隔离避免锁争用

离线同步限制

限制项	说明
视图	支持读取视图
切割键	仅支持整型字段

支持的字段类型

完整字段类型请参见 SQL Server 官方文档 。

字段类型	离线读	实时读
BIT / TINYINT / SMALLINT / INT / INTEGER / BIGINT	✓	✓
FLOAT / REAL / DECIMAL / NUMERIC / MONEY / SMALLMONEY	✓	✓
CHAR / NCHAR / VARCHAR / NVARCHAR / TEXT / NTEXT / XML	✓	✓
DATE / TIME / DATETIME / DATETIME2 / SMALLDATETIME / DATETIMEOFFSET	✓	✓
BINARY / VARBINARY / IMAGE / UNIQUEIDENTIFIER	✓	✓
GEOGRAPHY / GEOMETRY / HIERARCHYID / SQL_VARIANT	×	×

实时同步支持的 DML / DDL

DML

操作	是否支持
INSERT / DELETE / UPDATE	✓

DDL

操作	是否支持
新增列 / 删除列 / 重命名列 / 修改列类型	×
新增表 / 重命名表 / 删除表 / 清空表	×

SQL Server 基于 CDC 机制进行变更捕获，不支持 DDL 变更的自动同步。Schema 变更后需要重新启用 CDC，并可能需要重新配置同步任务。建议在业务低峰期进行 Schema 变更。

前提条件

1. 配置账号权限

离线同步：

```
CREATE LOGIN wedata_user WITH PASSWORD = 'your_password';
USE your_database;
CREATE USER wedata_user FOR LOGIN wedata_user;
GRANT SELECT ON SCHEMA::dbo TO wedata_user;
```

实时同步：

```
CREATE LOGIN wedata_cdc_user WITH PASSWORD = 'your_password';
USE your_database;
CREATE USER wedata_cdc_user FOR LOGIN wedata_cdc_user;
GRANT SELECT ON SCHEMA::dbo TO wedata_cdc_user;
GRANT SELECT ON SCHEMA::cdc TO wedata_cdc_user;
ALTER ROLE db_datareader ADD MEMBER wedata_cdc_user;
```

2. 开启数据库与表 CDC（仅实时同步）

```
USE your_database;
EXEC sys.sp_cdc_enable_db;

EXEC sys.sp_cdc_enable_table
    @source_schema = N'dbo',
    @source_name = N'your_table',
    @role_name = NULL,
    @supports_net_changes = 1;
```

3. 确保 SQL Server Agent 服务正在运行

CDC 依赖 SQL Server Agent Job 来捕获变更：

```
EXEC sys.sp_cdc_help_jobs;
```

4. 开启快照隔离（推荐）

```
ALTER DATABASE your_database SET ALLOW_SNAPSHOT_ISOLATION ON;
```

```
ALTER DATABASE your_database SET READ_COMMITTED_SNAPSHOT ON;
```

创建数据源

操作步骤

1. 登录 DataBuddy 控制台。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入 > 任务管理 > 数据源管理**（或 **平台管理 > 数据源管理**）。
4. 点击 **添加数据源**，选择 **SQL Server**。
5. 填写连接配置与认证信息，详见 参数说明。
6. 点击 **测试连接**（参见 数据源连通性测试）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务**。

参数说明

参数	说明	是否必填
数据源名称	数据源在工作空间内的唯一标识	是
JDBC URL	`jdbc:sqlserver://host:port;DatabaseName=database`	是
用户名	数据库连接账号	是
密码	数据库连接密码（支持 SSM 凭证托管）	是
数据源版本	见 支持的版本	是

JDBC URL 示例：

```
jdbc:sqlserver://10.0.0.1:1433;databaseName=mydb
jdbc:sqlserver://10.0.0.1:1433;databaseName=mydb;encrypt=true;trustServe
rCertificate=true
```

常用 URL 参数：

参数	说明
`databaseName`	数据库名称
`encrypt` / `trustServerCertificate`	加密连接相关

`loginTimeout`	登录超时时间（秒）

在数据接入任务中使用

实时整库

配置项	说明
来源表范围	**指定表**（仅同步选中的表，新增表需重启）/ **正则匹配**（指定库名 + 表名正则，符合规则的新增表自动接入）
读取模式	**全量 + 增量 / 仅增量**

离线单表

参数	说明
库 / Schema / 表	必填，Schema 默认 `dbo`
切割键	仅支持整型字段
筛选条件	标准 WHERE 语法

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

- **账号权限最小化：** 使用专用账号，避免 `sa` 。
- **CDC 日志保留：** 建议至少 72 小时：

```
EXEC sys.sp_cdc_change_job
    @job_type = 'cleanup',
    @retention = 4320;  -- 单位：分钟（72 小时）
```

- **切割键选择：** 主键或索引列。
- **网络打通：** 优先使用内网连接。

常见问题

Q：报错 CDC 未启用或找不到 CDC 表?

A：在数据库与表上启用 CDC：

```
USE your_database;
EXEC sys.sp_cdc_enable_db;
EXEC sys.sp_cdc_enable_table
    @source_schema = N'dbo',
    @source_name = N'your_table',
    @role_name = NULL;
```

Q：CDC 变更未被捕获?

A：SQL Server Agent 服务未启动。请启动 Agent 服务，并通过 `EXEC sys.sp_cdc_help_jobs;` 检查 CDC 作业状态。

Q：报错找不到 CDC 变更数据?

A：CDC 日志保留时间太短，已被清理。请增加 CDC 保留时间（见 最佳实践 ）。

Q：连接超时或被拒绝?

A：检查网络连通性、确认 SQL Server 允许远程连接、检查防火墙端口（默认 1433）。

Q：同步过程中出现锁等待或死锁?

A：未开启快照隔离：

```
ALTER DATABASE your_database SET ALLOW_SNAPSHOT_ISOLATION ON;
ALTER DATABASE your_database SET READ_COMMITTED_SNAPSHOT ON;
```

相关文档

- 添加数据源连接
- 数据源连通性测试
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

TDSQL-C MySQL 数据源

最近更新时间：2026-09-11 20:42:32

概述

TDSQL-C MySQL 是腾讯云自研的云原生关系型数据库，100% 兼容 MySQL，提供高吞吐与海量分布式存储。DataBuddy 数据接入支持将 TDSQL-C MySQL 作为来源端进行实时同步与离线同步。本文介绍其连接配置、能力支持范围与常见问题。

支持的版本

类型	支持版本
TDSQL-C MySQL	5.7、8.0
JDBC Driver	8.0.21

读取能力

能力	离线读	实时整库读	分库分表读
支持	✓	✓	✓

完整数据源能力对照参见 支持的数据源与读取能力 。

使用限制

实时同步限制

限制项	说明
Binlog 格式	仅支持 **ROW**
`binlog_row_image`	必须设置为 **FULL**
Binlog 保留时间	建议至少保留 72 小时
无主键表	无主键表无法保证 exactly-once 语义
XA 事务	不支持 XA ROLLBACK
只读备库	不支持读取（存在稳定性与兼容性问题）
存储过程、物化视图	不支持读取

SET 类型	实时同步不支持

离线同步限制

限制项	说明
分表同步并发	多表同步切分单表时，任务并发数需大于表个数
存储过程	不支持读取

支持的字段类型

完整字段类型请参见 MySQL 官方文档 。

字段类型	离线读	实时读
TINYINT / TINYINT UNSIGNED	✓	✓
SMALLINT / MEDIUMINT / INT / BIGINT（含 UNSIGNED）	✓	✓
FLOAT / DOUBLE / DECIMAL / NUMERIC	✓	✓
REAL	×	✓
CHAR / VARCHAR / TEXT 系列	✓	✓
JSON	✓	✓
BINARY / VARBINARY / BLOB 系列	✓	✓
DATE / TIME / DATETIME / TIMESTAMP / YEAR	✓	✓
BIT / BOOL / BOOLEAN / ENUM	✓	✓
SET	✓	×
GEOMETRY / POINT	×	✓
LINESTRING / POLYGON / MULTIPOINT 等空间类型	×	×

`TINYINT(1)` 默认映射为 BOOLEAN，可在高级设置中改为 TINYINT。

实时同步支持的 DML / DDL

DML

操作	是否支持

INSERT / DELETE / UPDATE	✓

DDL

操作	是否支持	支持的语法示例
新增列（ADD COLUMN）	✓	`ALTER TABLE t ADD col TYPE [FIRST | AFTER col]`
删除列（DROP COLUMN）	✓	`ALTER TABLE t DROP [COLUMN] col`
重命名列（RENAME / CHANGE COLUMN）	✓	`ALTER TABLE t CHANGE old new TYPE`
修改列类型（MODIFY COLUMN）	✓	`ALTER TABLE t MODIFY col TYPE`
新增表（CREATE TABLE）	✓	不支持 CHECK、Temporary Table
重命名表 / 删除表 / 清空表	✗	

DDL 变更响应**仅在增量阶段** 生效；删除表、重命名表等高危操作目标端通常不支持自动响应。

前提条件

1. 配置账号权限

离线同步：

```
CREATE USER 'wedata_user'@'%' IDENTIFIED BY 'your_password';
GRANT SELECT ON your_database.* TO 'wedata_user'@'%';
FLUSH PRIVILEGES;
```

实时同步：

```
CREATE USER 'wedata_cdc_user'@'%' IDENTIFIED BY 'your_password';
GRANT SELECT, SHOW DATABASES, REPLICATION SLAVE, REPLICATION CLIENT
    ON *.* TO 'wedata_cdc_user'@'%';
FLUSH PRIVILEGES;
```

启用 `scan.incremental.snapshot.enabled` 后不再需要 RELOAD 权限。

2. Binlog 配置（仅实时同步）

TDSQL-C MySQL 默认开启 Binlog。在控制台「参数设置」中确认：

- `binlog_format` = `ROW`
- `binlog_row_image` = `FULL`
- GTID 默认开启且不支持关闭。

修改参数后需要按控制台提示进行实例重启。

创建数据源

操作步骤

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入** > **任务管理** > **数据源管理** （或 **平台管理** > **数据源管理** ）。
4. 点击 **添加数据源** ，选择 **TDSQL-C MySQL** 。
5. 填写连接配置与认证信息，详见 参数说明 。
6. 点击 **测试连接** （参见 数据源连通性测试 ）。
7. 测试通过后点击 **保存** 或 **保存 & 创建任务** 。

参数说明

参数	说明	是否必填	默认值
数据源名称	数据源在工作空间内的唯一标识	是	—
连接方式	**串连接（JDBC URL）** / **腾讯云实例**	是	腾讯云实例
JDBC URL	串连接方式下的连接串，格式 `jdbc:mysql://host:port/database`	串连接必填	—
地域、实例	腾讯云实例方式必填	腾讯云实例必填	—
用户名	数据库连接账号	是	—
密码	数据库连接密码（支持 SSM 凭证托管）	是	—
数据源版本	5.7	是	—
高级参数	连接级扩展参数	否	—

在数据接入任务中使用

实时整库、分库分表

配置项	说明
来源表范围	**指定表**（仅同步选中的表，新增表需重启）/ **正则匹配**（指定库名 + 表名正则，符合规则的新增表自动接入）
读取模式	**全量 + 增量 / 仅增量**
订阅类型	可勾选 INSERT / UPDATE / DELETE，至少保留一种
gh-ost 同步	是否同步 gh-ost 在线 DDL 产生的临时表

离线单表

配置项	说明
数据来源	已配置的 MySQL 数据源
库 / 表	支持指定与正则匹配；表名支持范围如 `table_[0-99]`
切割键	仅支持整型字段；建议选择主键或带索引的列；避免 COLLATE 列
筛选条件	标准 WHERE 语法，支持时间参数（如 `gmt_create > $bizdate`）

详细任务配置参见 创建单表同步任务 与 实时整库多表任务配置 。

最佳实践

- **账号权限最小化**：使用专用账号，仅授必要权限。
- **Binlog 保留时间**：至少 72 小时；大表全量同步建议更长。
- **切割键选择**：优先主键或索引列。
- **网络打通**：优先使用内网连接。

常见问题

Q：报错 server-id 冲突?

A：多个同步任务使用了相同的 server-id。系统已优化为随机生成；如旧任务在高级参数中显式配置了 `server-id`，请删除。

Q：报错 Binlog 文件被清理?

A：增加 Binlog 保留时间，或优化作业性能加速消费。

Q：报错 Connection reset?

A：网络问题或作业反压。可调大 `slave_net_timeout = 120`，并优化并行度与内存。

Q：JobManager OOM?

A：调大 JobManager CU；调整高级参数 `scan.incremental.snapshot.chunk.size`（默认 8096）。

Q：是否支持 gh-ost 在线 DDL?

A：支持，开启 gh-ost 同步开关后会监控 `*_gho` 临时表变更，支持配置影子表正则。

相关文档

- MySQL 数据源
- 添加数据源连接
- 支持的数据源与读取能力
- 创建单表同步任务
- 实时整库多表任务配置

离线数据同步
创建单表同步任务
表到表同步

最近更新时间：2026-09-11 21:25:00

概述

表到表同步是最常见的离线同步链路：将结构化数据源（如 MySQL、PostgreSQL、Oracle、SQL Server 等）的表数据按调度计划同步到 DataBuddy 数据湖仓的目标表。本文以 MySQL 同步到 DataBuddy Catalog 表为例介绍配置流程。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关角色**（详见 成员与权限管理 ）。
- 已创建用于数据接入的计算资源（数据接入型）。
- 已创建并测试通过来源端数据源连接（详见 添加数据源连接 ）。
- 目标端 Catalog 已规划好；如目标 Schema 或表不存在，可通过本任务的「一键建表」自动创建（需要 Catalog 创建权限）。

使用限制

限制项	说明
计算资源	仅支持「平台集群」计算资源
来源端类型	见 支持的数据源与读写能力
目标端类型	DataBuddy 数据湖仓 Catalog 下的表（Iceberg 原生表）
切割键	仅整型字段；建议主键或索引列
并发数	> 1 时必须配置切割键

操作步骤

步骤 1：进入任务创建页

1. 登录 DataBuddy 控制台 。

2. 进入 **数据接入 > 任务管理** ，点击 **创建同步任务** 。

步骤 2：填写基础信息

填写任务名称、同步方式、责任人、描述（详见 创建单表同步任务 通用配置）。点击 **下一步** 。

步骤 3：配置来源端

参数	说明	是否必填
数据源类型	选择来源数据库类型，如 MySQL	是
数据源连接	选择已注册的数据源；下拉列表内可点击 **新增数据连接** 跳转创建	是
来源端数据库	支持 **选择数据库** 与 **输入数据库名称** 两种模式；可插入参数变量	是
来源端数据表	支持 **选择数据表** 与 **输入数据表名称** 两种模式；可插入参数变量	是
添加分库分表	同源类型可添加多条数据源 + 库表	否
切割键	整型字段；并发 > 1 时必填，否则保存时将提示补全	视并发数
筛选条件	标准 WHERE 条件（无需 `where` 关键字）；支持时间参数与调度参数	否
高级设置	数据源专属参数（详见对应数据源章节）	否

> **提示**
> **数据源连通性测试**：选择数据源连接后，展示「待测试」，可点击 **测试** 发起连通性测试。详见 数据源连通性测试 。

库表选择模式

模式	行为	适用场景
选择数据库 / 数据表	字面量精确选择	同步固定的库表
输入数据库名称 / 数据表名称	可通过规则表达式匹配，也可用参数变量	多分表（如 `table_[0-99]` ）或者按照读取当日分区数据进行同步的场景

分库分表配置

点击 **添加分库分表** 可叠加多条同源类型的数据源，每条独立配置：数据源连接、来源端数据库、来源端数据表。已添加的源支持删除。
要求：所有源的表 Schema 一致。

步骤 4：配置目标端

参数	说明	是否必填	默认值
目标端 Catalog	选择目标 Catalog	是	-
目标端 Schema	在所选 Catalog 下指定 Schema	是	—
目标端 Table	指定目标表；不存在时点击 **创建目标表** 来即时创建目标表	是	—
写入模式	详见 写入模式	是	append
高级设置	目标端扩展参数	否	—

创建目标表

点击下拉列表，点击 **创建目标表** 进入建表弹窗：

1. 选择用于执行建表语句的计算资源（交互式集群类型）；资源未启动时提交后会自动启动。
2. 平台基于来源端 Schema 生成默认建表 SQL，允许在编辑器中调整。
3. 点击 **提交**：
   - 成功 → 弹窗关闭，提示「目标表 ${目标表名称} 创建成功」。
   - 失败 → 弹窗保留，提示具体失败原因，可调整 SQL 后重试。

注意：

- 部分数据源不支持自动生成建表 DDL，可手动写入建表 SQL。

写入模式

模式	说明
append	插入模式；主键冲突时写入失败
overwrite	替换模式；主键冲突时先删除原行再插入新行
upsert	更新模式；主键冲突时更新指定字段

步骤 5：配置字段映射

字段映射默认按同名、同行规则填充，支持以下操作：

操作	说明
同名映射	来源 / 目标同名字段自动对齐
同行映射	按行号顺序对齐

清除	清空所有映射
置顶已映射	已映射字段排在前面
移动	拖动字段前面的移动图标，可以移动行
映射开关	点击「映射」列的箭头开关，可以切换该行的映射关系

字段操作：

操作	说明
添加字段	手动新增一行字段映射
删除字段	手动删除一行字段映射
字段类型调整	调整字段类型，类型支持：字段、常量、函数、参数；当选择字段，则可以选择来源端字段类型如「string」

各种字段类型的设置：

- 字段：可选择一个来源端字段类型，如「string」、「int」等；
- 常量：可输入常量值作为字段名，如「1」、「true」等；
- 函数：可输入数据库函数作为字段名，如「concat(first_name,last_name)」、「sum(amount)」等；
- 参数：可输入参数变量作为字段名，如「${plan_date}」,参数的定义可在 创建单表同步任务 》任务参数 中设置；

步骤 6：配置运行设置

详见：创建单表同步任务 》运行设置。

步骤 7：调试、保存、发布

详见：创建单表同步任务 》调试、保存、发布。

常见问题

Q：保存任务时提示「需要先设置切割键」？

A：并发数 > 1 时必须配置切割键。请在 **来源 & 目标** 步骤补充切割键，或将并发数调整为 1。

Q：建表失败，提示权限不足?

A：当前账号没有目标 Catalog 的建表权限。请联系空间管理员或 Catalog 所有者授予 **可管理** 或建表所需权限（详见 成员与权限管理 ）。

Q：分库分表场景如何确保所有源连接都通过连通性测试?

A：选择多个数据源时，需要分别测试每个连接。

Q：调试运行会真实写入数据吗?

A：**是** 。调试与发布运行使用相同的搬运链路，会按当前配置真实写入目标端。请在调试前确认目标表与写入模式，避免覆盖业务数据。

Q：任务编辑后多久生效?

A：保存只生成新版本，不影响调度。需要点击 **发布** 将当前版本设为发布版本，下一次工作流调度才会使用新版本。

相关文档

- 表到 Volume 同步
- 文件到表同步
- 批量创建单表同步任务
- 离线同步任务运维监控
- 工作流概述

表到 Volume 同步

最近更新时间：2026-09-11 20:42:32

概述

表到 Volume 同步将结构化数据源（如 MySQL、PostgreSQL、Oracle、SQL Server 等）的表数据按指定文件格式写入 DataBuddy Volume，常用于将加工后的数据导出为 .ok / .ctl 文件给下游系统消费。本文介绍该链路的特有配置，通用步骤参见 创建单表同步任务 。

适用场景

- 周期性导出加工结果文件给短信、营销、外部系统等下游业务（如 .txt / .csv / .json）。
- 数据交付：将湖仓数据按约定格式导出到指定 Volume 路径。
- 文件归档：将业务库数据按时间分区落盘到 Volume，方便长期存储。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关角色**（详见 成员与权限管理 ）。
- 已创建用于数据接入的计算资源（平台集群型）。
- 已创建并测试通过来源端数据源连接（详见 添加数据源连接 ）。
- 目标 Catalog 下已存在 Volume 类型的 Schema，或当前账号有权限创建。

使用限制

限制项	说明
来源端类型	仅支持结构化数据源（参见 支持的数据源与读写能力 ）
目标端类型	DataBuddy Catalog 下的 Volume 类型 Catalog
字段映射	来源端可设置；目标端不可设置（Volume 不感知字段结构）
文件类型	仅支持 txt / csv / json
.ok 与 .ctl	不同文件类型支持的标记文件不同

操作步骤

通用流程同 表到表同步 。本节仅说明**目标端** 与**字段映射** 的差异。

步骤 1：进入任务创建页

参见 表到表同步 入口。

步骤 2：填写基础信息、来源端

来源端字段（数据源类型、连接、来源端数据库、数据表、切割键、筛选条件、分库分表、高级设置）与表到表同步一致，参见 表到表同步 > 步骤 3 。

步骤 3：配置目标端

当目标端选择为 Volume 类型 Catalog 时，配置项如下：

参数	说明	是否必填	默认值
Catalog	选择有权限的 Volume 类型 Catalog	是	—
Schema	Catalog 下的 Schema	是	—
Volume	选择已存在的 Volume；不存在时可在 Catalog 中创建	是	—
Volume 目录	文件写入的 Volume 子路径；可选择、可输入、可插入正则、可插入参数；不建议你把根路径「/」作为子路径	是	—
文件名称	写入的文件名；可输入、可插入正则、可插入参数	是	—
文件格式	txt / csv / json；不同类型对应的解析配置不同	是	—
格式设置	见 文件解析与推断	是	—
写入模式	Append / Overwrite / Truncate / NonConflict	是	Append
标记完成文件	当文件写入完成后生成.ok 文件	是	不生成
ctl 文件	当文件写入完成后生成.ctl 文件	视文件类型	不生成

步骤 4：格式设置

点击「格式设置」弹出弹窗，按所选文件类型配置写入规则：

TXT / CSV

配置项	说明
压缩格式	gzip / 无压缩等

列分隔符	字段分隔符（如 `,` `;` `
行分隔符	行结束符（如 `\n` `\r\n` 自定义）
编码	UTF-8 / GBK 等
控制转换	是否对特殊字符做转义
Quote Character	字符串包裹符（如 `"` `'`）
包含表头	是否在第一行记录写入字段名

JSON

配置项	说明
压缩格式	gzip / 无压缩等
编码	UTF-8 / GBK 等
空值转换	是否将 null 转换为指定字符串
包含表头	是否在第一行记录写入字段名

步骤 5：字段映射

链路	来源端字段映射	目标端字段映射
结构化数据源 → Volume	可设置	不可设置

- 如果你不希望某个字段被写入 Volume，可以通过删除字段映射行来实现。
- 添加字段映射行时，字段类型必须为「字段」，否则会报错。

步骤 6：运行设置、调试、保存、发布

参见 表到表同步 > 步骤 4 / 5 。

文件类型与标记文件支持

文件类型	.ok 文件	.ctl 文件
Text	有	有
CSV	有	有
JSON	有	无

> **提示**
> `.ok` 文件用于通知下游「数据已写入完成」；`.ctl` 文件用于记录数据描述（如行数、文件清单）。

验证结果

任务保存并调试运行后，您可以通过以下方式确认：

1. 在 数据目录 中进入目标 Volume，看到已写入的文件（如 `output_20260101.txt` 与 `output_20260101.ok`）。
2. 下载、预览文件内容，确认字段顺序、分隔符、编码与配置一致。
3. 在 离线同步任务运维监控 中查看任务日志确认无错误。

常见问题

Q：目标端 Volume 路径或文件名支持哪些参数?

A：支持系统时间参数（如 `${yyyyMMdd}`）和工作流调度参数。手动调试运行时如未设置参数值，系统使用默认、缺省值；建议在「运行参数」中提前定义，确保调试可重复。

Q：写入模式选择 Overwrite 时，目标 Volume 路径下的旧文件会被删除吗?

A：Overwrite 仅覆盖同名文件，不会删除路径下的其他文件；Append 在同名文件后追加内容（取决于文件类型）；

Q：JSON 文件的「空值转换」如何配置?

A：开启后，所有 NULL 字段会被转换为指定字符串（如空字符串 `""` 或字面量 `NULL`）。关闭时按 JSON 标准输出 `null` 字面量。

Q：是否支持 .parquet / .avro 等列式格式?

A：当前仅支持 TEXT / CSV / JSON 三种格式。

相关文档

- 表到表同步
- 文件到 Volume 同步
- 批量创建单表同步任务
- 离线同步任务运维监控
- 非结构化数据 Catalog（文件、Volume 管理）

文件到表同步

最近更新时间：2026-09-11 20:42:32

概述

文件到表同步将半结构化文件（FTP / SFTP / COS / HDFS / REST API 等）按解析规则读取并写入 DataBuddy Catalog 下的目标表。本文介绍该链路的特有配置，通用步骤参见 批量创建单表同步任务 。

适用场景

- 周期性接入业务伙伴上传到 FTP / SFTP / COS 的批量文件，落入数据湖仓供下游分析。
- 接入第三方 SaaS 系统通过 REST API 提供的数据。
- 增量同步对象存储中按日期分区的文件。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关角色**（详见 成员与权限管理 ）。
- 已创建用于数据接入的计算资源（数据接入型）。
- 已创建并测试通过来源端数据源连接（详见 添加数据源连接 ）。
- 目标 Catalog 与 Schema 已规划好；如目标表不存在，可由本任务一键建表（需要建表权限）。

使用限制

限制项	说明
来源端类型	半结构化数据源（FTP / SFTP / COS / HDFS / REST API 等）
目标端类型	DataBuddy Catalog 下的表（Iceberg 原生表）
读取方式	仅 **解析数据**（解析后入表）；如需保留原始文件结构请使用 文件到 Volume 同步
文件路径	路径下无文件或仅有空文件时任务正常运行成功，但目标端无新增数据
字段映射	来源端不读取字段元数据，需要用户手动添加字段并指定类型

操作步骤

通用流程同 表到表同步 。本节仅说明**来源端** 与**字段映射** 的差异。

步骤 1：进入任务创建页、填写基础信息

参见 表到表同步 。

步骤 2：配置来源端

按所选数据源类型，来源端配置项略有差异。

通用字段（FTP / SFTP / COS / HDFS）

参数	说明	是否必填
数据源类型	选择来源数据源类型	是
源连接	选择已注册的数据源连接	是
文件路径	文件或文件夹路径；点击文件夹图标可弹出路径选择器逐层下钻；支持插入参数与正则；	是
读取方式	**解析数据**：解析文件入表；**原始文件**：原样落 Volume（属于 文件到 Volume 同步 ）	是
文件格式	数据同步模式下选择 Text/CSV/JSON/ORC/Parquet	是
格式设置	按文件类型展示对应字段；	视文件类型
高级设置	数据源专属扩展参数	否

REST API

参数	说明	是否必填
数据源类型	REST API	是
数据源连接	选择已注册的 API 数据源	是
读取方式	**解析数据**：解析结构化数据内容，按字段关系进行数据内容映射与同步; **原始文件**：不做内容解析传输整个文件，可应用于非结构化数据同步	是
请求方式	GET / POST	是
数据路径	接口返回中数据所在 JSON 路径（如 `data.list` ）	否
返回结构	单条 JSON / 数组 JSON	是

请求 Header	自定义请求头	否
请求参数	GET 查询参数或 POST 请求体参数	否
第一行作为字段	当返回为表格类结构时是否将首行作为字段名	否
高级设置	**请求方式**：单次请求 / 多次请求；**参数**：分页参数等	否

步骤 3：解析配置（按文件类型）

文件类型	关键配置
Text	列分隔符（自定义 / `;` / `|` / Tab）、行分隔符（自定义 / 回车 / 断行）、第一行作为字段行、转义字符（默认 `\`）
CSV	列分隔符（自定义 / `,` / `;` / Tab）、转义符（`"` / `\`）、第一行作为字段名、行跨行
JSON	将时间文本转换为时间类型、允许使用单引号、过滤注释、行跨行 当存在 JSON 嵌套时，嵌套的值会被解析为一个文本对象（不会自动展开为多列）。
ORC / Parquet	无需额外配置（自描述格式）

步骤 4：配置目标端

目标端配置项与 表到表同步 完全一致：

- 目标 Catalog / Schema / Table。
- 写入模式：append / overwrite / upsert。

步骤 5：配置字段映射

链路	来源端字段映射	目标端字段映射
文件（解析）→ 表	可设置（需手动添加字段）	可设置

文件类型的来源端**不读取字段元数据**，需要用户手动添加字段：

1. 点击 **添加字段**，输入字段索引值，如输入 0，表示读取出来的第一个元素，注意是从 0 开始检索的。
2. 选择字段类型（支持 STRING / INT / BIGINT / FLOAT / DOUBLE / DECIMAL / DATE / TIMESTAMP / BOOLEAN 等）。
3. 与目标端字段建立映射。

操作支持：清除、置顶已映射、刷新字段。

步骤 6：运行设置、调试、保存、发布

参见 表到表同步 。

验证结果

任务保存并调试运行后，您可以通过以下方式确认：

1. 在数据接入任务列表中可以看到刚创建的任务。
2. 调试运行成功后，目标 Catalog 下对应表已包含解析后的数据；可在 数据目录 或 SQL 探索中查询验证。
3. 如需更细粒度的字段类型校验，可在 SQL 探索中运行 `DESCRIBE` 或简单查询观察实际写入字段。

常见问题

Q：上传的 CSV 文件包含表头，但映射结果将表头当作了数据行？

A：来源端的 CSV / TXT 解析配置中需勾选 **第一行作为字段名** 。

Q：JSON 文件含嵌套对象，如何展开为多列？

A：当前 JSON 嵌套字段会整体解析为文本对象，不自动展开。如需展开，建议在 SQL 任务中使用 `get_json_object` / `from_json` 等函数后处理。

Q：XML 行标签自动推断错误，如何手动指定？

A：在解析设置中，行标签字段输入正确的行标签名即可。例如 `<root><record>...</record></root>` 中行标签应为 `record` 。

Q：路径下没有文件，任务还会运行吗？

A：会，**任务运行成功但目标端无新增数据** 。

相关文档

- 表到表同步
- 文件到 Volume 同步
- 批量创建单表同步任务
- 离线同步任务运维监控
- 本地上传到表（一次性手动上传场景）

文件到 Volume 同步

最近更新时间：2026-09-11 20:57:30

概述

文件到 Volume 同步将半结构化文件（FTP / SFTP / COS / S3 / Azure Blob / HDFS 等）写入 DataBuddy Volume，支持两种用法：

- **原始文件传输**：保留文件原始格式，直接落入 Volume，常用于归档与备份。
- **解析后转换格式**：读取源文件并按目标文件类型重新输出，常用于格式转换（如 TXT 转 JSON）。

本文介绍该链路的特有配置，通用步骤参见 创建单表同步任务 。

适用场景

- 文件归档：将业务伙伴上传到 SFTP 的原始文件按日期落入 Volume。
- 格式转换：将上游 TXT 文件读出后以 JSON 格式重新落到 Volume，方便下游消费。
- 多文件合并：将多个源路径下的文件读取后按统一格式集中输出。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关角色**（详见 成员与权限管理 ）。
- 已创建用于数据接入的计算资源（数据接入型）。
- 已创建并测试通过来源端数据源连接（详见 添加数据源连接 ）。
- 目标 Catalog 下已存在 Volume 类型 Schema，或当前账号有创建权限。

使用限制

限制项	说明
来源端类型	半结构化数据源（FTP / SFTP / COS / S3 / Azure Blob / HDFS 等）
目标端类型	DataBuddy Catalog 下的 Volume 类型 Catalog
读取方式	**原始文件**（原样落 Volume）/ **解析数据**（解析后按目标文件类型重新输出）
字段映射	原始文件模式：来源端 / 目标端均不可设置；解析模式：来源端可设置（手动添加字段），目标端不可设置
输出文件类型	仅 txt / csv / dat / json

操作步骤

步骤 1：进入任务创建页、填写基础信息

参见 创建单表同步任务 。

步骤 2：配置来源端

参数	说明	是否必填
数据源类型	选择来源数据源类型（如 SFTP）	是
数据源连接	选择已注册的数据源连接	是
文件路径	文件或文件夹路径，点击文件夹图标可弹出路径选择器；支持插入参数与正则；S3 需带桶名；Azure Blob 需带容器名	是
读取方式	**原始文件** 或 **解析数据**（决定是否走解析）	是
文件格式	数据同步模式必填：.csv / .avro / .json / .parquet / .txt	视模式
解析配置	数据同步模式必填，按文件类型配置	视模式
同步范围（仅 S3）	全量 / 增量	视场景
高级设置	数据源专属扩展参数	否

具体来源端字段差异请参考 文件到表同步 ，本链路与文件到表的来源端配置完全一致。

步骤 3：配置目标端

当目标端选择为 **Volume 类型 Catalog** 时：

参数	说明	是否必填
目标 Catalog	选择有权限的 Volume 类型 Catalog；不存在时可点击「新建」创建	是
目标 Schema	Catalog 下的 Schema	是
Volume	选择已存在的 Volume	是
Volume 文件目录	Volume 内子路径；为空表示落到 Volume 根目录	否（默认根目录）

文件名称	输出文件名（仅数据同步模式必填）	视模式
文件格式	数据同步模式必填：txt / csv / dat / json	视模式
格式设置	数据同步模式必填，按目标类型配置	视模式
写入模式	overwrite / append / upsert	是

数据同步模式下的输出格式配置

文件类型	配置项
TXT / CSV / DAT	压缩格式、列分隔符、行分隔符、编码、控制转换、Quote Character、包含表头
JSON	压缩格式、编码、空值转换、包含表头

标记文件支持

文件类型	.ok 文件	.ctl 文件
txt	有	有
csv	有	有
dat	有	无
json	有	无

步骤 4：字段映射

模式	来源端字段映射	目标端字段映射
文件同步（原始）	✕ 不可设置	✕ 不可设置
数据同步（解析）	✓ 可设置（手动添加字段）	✕ 不可设置

步骤 5：运行设置、调试、保存、发布

参见 创建单表同步任务 。

验证结果

任务保存并调试运行后，您可以通过以下方式确认：

1. 进入 数据目录 中目标 Volume，看到已写入的文件。
2. 文件同步模式下文件名与原文件一致；数据同步模式下文件名按目标端配置。
3. 下载或预览文件确认内容；如有 .ok 文件，确认内容标记符合下游约定。

常见问题

Q：选错了读取方式，能改吗？

A：可以。在任务编辑页将「读取方式」从「原始文件」切到「解析数据」，并补充必填配置项后保存即可。注意：切换会重置已有的解析配置与字段映射。

Q：读取方式选择「原始文件」时，多个源文件会合并为一个文件吗？

A：不会。文件同步模式下源文件按原结构 1:1 落入 Volume（保留原文件名），不做合并。

Q：数据同步模式下输出文件名相同会被覆盖吗？

A：取决于写入模式：

- **overwrite**：覆盖同名文件。
- **append**：在同名文件后追加内容。
- **upsert**：行为与 overwrite 类似（Volume 不感知主键）。

Q：源端选择了不同路径下的同名文件，将会如何？

A：将会报错，逻辑上在读取时，会将所有文件提到一个 list，如果有重名文件，会报错。

Q：能否在同一个任务中既保留原始文件又生成解析后文件？

A：不支持。一个任务只能选择一种同步模式。如需同时归档与转换，建议创建两个任务并通过 Workflow 编排。

相关文档

- 创建单表同步任务
- 文件到表同步
- 表到 Volume 同步
- 批量创建单表同步任务
- 本地文件上传到卷

批量创建单表同步任务

最近更新时间：2026-09-11 21:16:00

概述

批量创建单表同步任务允许一次性为多张来源表生成多个独立的单表同步任务，避免逐个创建的重复劳动。常用于一次性入湖大量结构化表，或周期性接入多个非结构化文件路径到目标 Catalog / Volume。
每个生成出来的单表任务结构与 创建单表同步任务 完全相同，可独立编辑、调试、发布、调度。批量任务的本质是「一次配置、批量生成 N 个单表任务」。

适用场景

- 业务库批量入湖：MySQL 中 100 张业务表一次性同步到湖仓。
- 多分库同源同步：多个分库（如 db_0 / db_1 / ... / db_n）的同名表合并到目标表。
- 多文件路径批量同步：FTP 上多个目录下的同结构文件批量落入 Volume。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有 **数据接入相关角色**（详见 成员与权限管理）。
- 已创建用于数据接入的计算资源（平台集群型）。
- 已创建并测试通过来源端数据源连接（详见 添加数据源连接）。
- 目标 Catalog 已规划好；如目标 Schema 或表不存在，可由批量任务在创建过程中自动建表（需要 Catalog 创建权限）。

使用限制

限制项	说明
来源端类型	结构化数据源（与单表任务一致）+ 非结构化文件（FTP / SFTP / COS / S3 / Azure Blob / HDFS / Rest API）
目标端类型	表（结构化场景）/ Volume（文件场景）
目标端禁用部分配置	选择目标 Catalog 后，目标 Schema、Table（或 Volume / 路径 / 文件名称）由表映射逐项指定，目标端表单中显示「请在映射中设置」
单任务并发	通道设置作用于每个单表任务，相当于批量设置每个单表任务属性

操作步骤

步骤 1：进入任务创建页

1. 登录 DataBuddy 控制台 。
2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入 > 任务管理** ，点击 **创建同步任务** 。
4. 选择 **同步方式： 离线批量同步** 。

步骤 2：填写基础信息

参数	说明	是否必填	默认值
任务前缀	用于生成单表任务名称的前缀；最终单表任务名格式 `${任务前缀}_${目标表名称}`	是	—
接入方式	批量离线单表接入	是	—
责任人	默认应用到生成的所有单表任务	是	当前创建用户
任务描述	默认应用到生成的所有单表任务	否	—

步骤 3：配置来源端

结构化场景

参数	说明	是否必填
源连接	与单表任务一致	是
选择源端表	支持选择多张表；支持按库 / 表检索；展示已选库表结构；可逐项删除	是
切割键	默认应用到每个单表任务	否
筛选条件	默认应用到每个单表任务	否
高级设置	默认应用到每个单表任务	否

非结构化文件场景

参数	说明
源连接	与单表任务一致

读取方式	数据同步 / 文件同步（与单表一致）
文件格式	将以指定的文件格式解析文件
解析配置	与单表一致；默认应用到每个单表任务
高级设置	默认应用到每个单表任务

步骤 4：配置目标端

目标 Catalog 可选表类型或 Volume 类型，根据目标 Catalog 类型进行配置。

选择目标 Catalog 为表类型

参数	说明	是否必填
Schema 匹配策略	选择同步目标端 Schema 匹配策略，可选与来源库一致 / 自定义，自定义可通过匹配规则来进行配置	是
表匹配策略	选择同步目标端表匹配策略，可选与来源表一致 / 自定义，自定义可通过匹配规则来进行配置	是
写入模式	append / overwrite / upsert	是
高级设置	默认应用到每个单表任务	否

选择目标 Catalog 为 volume 类型

参数	说明	是否必填
写入模式	append / overwrite / upsert	是
文件格式	将以指定的文件格式写入文件	是
格式设置	写入文件将以设置的格式组装文件	是
标记完成文件	当文件写入完成后生成.ok 文件	是
CTL 文件	当文件写入完成后生成.ctl 文件	视文件类型
高级设置	默认应用到每个单表任务	否

步骤 5：配置表映射

点击「刷新源表和目标表映射」可以刷新源表和目标表映射关系。稍后将以该映射关系生成单表任务。

表映射区根据目标 Catalog 类型动态展示。

目标端为表类型 Catalog

字段	说明
任务名称 / 源库 / 源表	不可编辑，跟随来源端设置
目标 catalog	在 Catalog 下指定
目标 Schema	根据目标端设置的 schema 匹配规则，显示目标 Schema 名称，如果该 Schema 不存在，显示红色"!"提示, 可以选择换一个或创建一个
目标 Table	根据目标端设置的表匹配规则，显示目标表名词，如果该表不存在，显示红色"!"提示, 可以选择换一个或创建一个
字段映射规则	可选同名 / 同行映射
计算资源	默认任务统一资源

目标端为 Volume 类型 Catalog

字段	说明
来源	可以通过点击「添加一行」生成一条单表任务
目标 Schema	在 Catalog 下指定, 可以选择换一个或创建一个
目标 Volume	在 Schema 下指定, 可以选择换一个或创建一个
Volume 路径	Volume 内子路径
文件名称	路径端点为文件时自动默认展示该文件名变量

表映射操作

操作	说明
批量操作	配置字段映射规则、配置计算资源、移除选中行
仅展示未配置	过滤出尚未配置目标的行
批量配置目标 schema	批量配置目标 schema
批量配置目标 volume	批量配置目标 volume
列表操作 – 编辑	弹出单表任务编辑（详见 单表覆写 ）
列表操作 – 移除	移除该映射条目

步骤 6：单表覆写

设置同单表任务的设置，详见 创建单表任务，部分设置不支持修改。

步骤 7：配置运行设置

详见 创建单表同步任务 - 步骤 4。

步骤 8：保存并创建单表任务

点击 **开始批量创建** 。系统返回任务列表页，在页面右下角弹出**进度弹层**：

字段	说明
批量任务状态	**创建中 / 已完成 / 正在终止 / 已终止**
批量任务名称	显示批量任务前缀
进度指标	成功数 / 异常数 / 总任务数
删除	删除批量任务，则可删除批量任务，但是已生成的单表任务不受影响

点击「任务列表」区域可查看具体的单表任务创建详情：

字段	说明
任务名称	如果任务创建成功，则显示单表任务名；否则显示「-」
来源 / 目标	来源库表 / 目标库表
状态	全部 tab 下展示；按 全部 / 成功 / 失败 切换

验证结果

批量任务执行完成后，您可以通过以下方式确认：

1. 进度弹层显示 **已完成** 且 成功数 = 总任务数。
2. 进入 **数据接入** > **任务管理** 任务列表，按任务前缀搜索，看到生成的多个单表任务。
3. 任意打开一个生成的单表任务，配置项与批量任务设置一致；可独立调试、发布、调度。

常见问题

Q：批量任务创建后，单表任务是否仍然可以独立编辑?

A：可以。生成的单表任务与手动创建的单表任务地位相同，可独立编辑、调试、发布、删除。修改单表任务不会影响批量任务记录。

Q：单表覆写后，再次回到批量任务修改全局设置会覆盖吗?

A：批量任务保存后即生效，已生成的单表任务不会被批量任务的后续配置覆盖。如需统一更新，需对每个单表任务单独编辑。

Q：批量任务支持哪些数据源?

A：与单表任务一致。结构化数据源支持表批量同步；非结构化数据源支持多路径批量同步（参见 使用限制 ）。

Q：批量任务的字段映射如何处理多张异构表?

A：批量任务统一不在批量层级配置字段映射，每个单表任务独立配置。如表结构差异较大，建议拆分为多个批量任务，分别处理同结构表组。

相关文档

- 表到表同步
- 表到 Volume 同步
- 文件到表同步
- 文件到 Volume 同步
- 离线同步任务运维监控

离线同步任务运维监控

最近更新时间：2026-09-11 20:42:30

概述

离线同步任务的运维监控围绕「任务列表 → 版本管理 → 发布 → 权限授权 → 运行实例查看」展开。本文介绍各项运维操作的入口与流程，方便您在日常使用中查看任务状态、回滚版本、控制发布节奏与诊断运行问题。

> **提示**
> 离线同步任务的实际**运行实例** 由 工作流 调度产生，运行详情、日志可在 Workflow 任务节点的实例视图中查看。

入口

入口	路径
任务列表	**数据集成** > **接入任务** > **离线**
任务编辑页	任务列表页操作列点击**编辑**，或点击任务名称
任务实例（运行详情）	**Workflow** > 工作流详情 > 任务节点

任务列表

任务列表展示当前用户**有权限** 的所有离线同步任务。

默认展示字段

- 任务名称（点击进入编辑页）
- 发布版本
- 创建人
- 最新修改人
- 最近修改时间

完整可选字段

支持自定义展示字段（点击列表头右侧的设置图标弹出字段勾选弹层）。完整字段如下表所示：

字段	说明
任务名称	单表同步任务名称

任务 ID	任务的唯一 ID
发布版本	当前已发布的版本号
发布时间	最近一次发布时间
创建人	创建者用户名
创建时间	任务创建时间
负责人	任务实际运行身份
最近修改人	最近一次保存任务的用户
最近修改时间	最近一次保存时间

> **提示**
> 自定义展示字段的设置保存到本地浏览器，下次进入仍生效。

列表操作

操作	说明
发布	触发任务发布检测，
编辑	进入任务编辑页
权限	配置任务级 ACL 权限，详见 成员与权限管理
删除	删除任务（不可恢复）

批量操作

任务列表支持多选后批量发布：

1. 复选 1 个或多个任务。
2. 点击 **批量发布** ，弹出发布检测弹层。
3. 切换查看每个任务的检测结果。
4. 选择 **重新检测 / 继续发布 / 忽略异常并发布** 。

版本管理

历史版本

用户每次点击 **保存** 都会生成一个历史版本，最多保留 200 个；超过则自动从最早的非发布版本开始删除。
进入任务编辑页，点击右上角 **版本管理** 图标可查看版本列表：

字段	说明
版本名称	由保存时间构成，格式「月日 时分秒」
创建人	保存的用户
标识	发布版本显示「已发布」

创建版本

操作	行为
保存	生成新历史版本；提示「任务保存成功」/「任务保存失败：${失败原因}」
仅保存	同上
创建并发布 / 保存并发布	同上

删除版本

操作选项：

操作	说明
单个版本删除	鼠标悬停历史版本，点击 **删除** 移除；至少保留 1 个版本
清空	二次确认后清空所有版本（**保留发布版本与最近版本**）

> **警告**
> **删除发布版本** 会触发额外二次确认：「当前删除已发布版本，该操作可能导致关联本任务的模块产生异常」。继续删除后任务发布状态变为 **未发布**，引用该任务的工作流执行时会失败，提示「任务未发布」。

版本回滚

版本回滚是「复制版本、编辑、发布」的过程：

1. 在历史版本列表中点击某个版本并回到此版本，编辑页切换为该版本配置（基于该版本生成临时编辑版本）。
2. 点击 **保存**：生成最新的保存版本。
3. 点击 **发布**：保存并发布，生成新版本并将发布版本指向该版本。

发布流程

单表任务发布

入口：任务列表、任务编辑页操作区点击 **发布**。
发布流程：

```
点击发布 → 发布检测弹层
        ├─ 全部通过：选择「重新检测」或「继续发布」
        └─ 部分异常：选择「重新检测」或「忽略异常并发布」
```

发布检测项与调试一致（详见 创建单表同步任务 – 调试运行 ）：

阶段	检测项
任务保存	序列化与持久化
配置检测	来源 / 目标 / 映射
数据源检测	来源 / 目标连通性、来源权限
资源检测	资源状态

发布完成提示：

- 成功 → 提示「任务发布成功」。
- 失败 → 提示「任务发布失败：${失败原因}」。

发布版本规则：

- 发布操作将当前编辑版本设置为发布版本。
- 一个任务**只保留一个发布版本**：再次发布会以新版本替换旧的发布版本。
- 列表「已发布」标识跟随当前发布版本移动。

批量任务发布

任务列表多选后点击 **批量发布**：

1. 弹出发布检测弹层，列表中展示每个任务的检测情况。
2. 支持按状态查看：通过、异常、进行中。
3. 进度条展示通过数与异常数。
4. 选择 **重新检测 / 继续发布 / 忽略异常并发布** 。

权限管理

任务级 ACL 权限基于平台统一权限模型。详见 成员与权限管理 。

权限项

权限	操作范围
可管理	查看 / 编辑 / 运行 / 删除 / 授权

可运行	查看 / 编辑 / 运行
可编辑	查看 / 编辑
可查看	仅查看

> **提示**
> **可运行** 权限同时控制调试运行：仅有 **可运行** 或更高权限的用户才能调试任务。

默认权限

- 工作空间管理员默认拥有 **可管理** 权限，且不可取消。
- 任务负责人默认拥有所有权限，且不可取消。

操作入口

入口	路径
任务列表	操作列点击 **更多 > 权限**

运行实例查看

> **提示**
> 离线同步任务通过 工作流 触发运行。任务实例的查看入口为：工作流 > 工作流详情 > 任务节点。

数据概览

实例运行详情提供以下指标：

类型	字段
读取	读取大小、记录数
写入	写入大小、记录数
耗时	执行时长、速度、脏数据

读取节点明细

字段	说明
数据来源	`数据库.数据表`
总条数	读取总条数

总字节数	读取总字节数
速度	行 / 秒
吞吐	字节 / 秒
等待读取时间	阻塞等待时间
脏数据条数	解析失败或超过阈值的脏数据数量

写入节点明细

字段与读取节点相同，但描述对象为目标表：目标表、总条数、总字节数、速度、吞吐、等待读取时间、脏数据条数。

实例操作

在 工作流 任务实例详情中支持以下操作：

操作	触发条件
运行	实例尚未运行
终止	实例正在运行；二次确认后停止
重试	实例已失败

运行日志

入口与运行实例一致：在 Workflow 任务实例详情中切换到**运行日志** 标签。

功能	说明
日志切换	实例多次执行（如重试）会生成多条日志记录，名称为「开始 ~ 结束时间」，可逐条查看
日志内容	文本形式展示完整运行日志
下载	点击右上角下载，导出为 .log 文件

删除任务

入口

入口	路径
任务列表	操作列点击 **更多>删除**

删除前提

当前账号拥有任务的 **可管理** 权限或为任务负责人。

删除行为

1. 点击删除后弹出二次确认。
2. 确认后任务被删除（实际为假删除，便于异常恢复）。
3. 已通过该任务完成的资产（如已同步的数据、已建的目标表）**不受影响** 。
4. 引用该任务的实体（如工作流）在执行时会报错「任务已删除」。
5. 删除操作记录在操作审计日志中（待补充），便于追查。

常见问题

Q：列表中看不到我创建的任务？

A：检查以下几点：

1. 当前选择的 Workspace 是否为创建任务的 Workspace。
2. 过滤条件（如创建人、最近修改时间）是否限制了显示范围。
3. 您的账号是否有该任务的查看权限（详见 成员与管理权限 ）。

Q：发布版本后，调度的工作流何时使用新版本？

A：工作流下次调度时使用最新发布版本；正在运行的实例不会切换到新版本。

Q：删除任务后能恢复吗？

A：平台采用假删除便于异常恢复，但**不开放对外恢复入口** 。如需恢复请提交工单联系技术支持。

Q：批量发布时部分任务失败，是否影响其他任务？

A：不影响。每个任务独立检测与发布；可在进度弹层中针对失败任务点击 **重新检测** 或 **忽略异常并发布** 。

Q：实例运行失败，如何快速定位？

A：在 Workflow 任务实例详情中：

1. 数据概览中查看「脏数据量」是否超阈值。
2. 切换到 **运行日志** 标签查看详细错误。
3. 点击 **下载日志** 导出 .log 进一步分析或反馈给技术支持。

相关文档

- 创建单表同步任务
- 批量创建单表同步任务

- 工作流概述
- 工作流与任务的运行监控
- 成员与权限管理

实时数据同步
实时整库多表任务配置

最近更新时间：2026-09-11 18:35:00

概述

实时整库多表任务用于将来源端数据库中的**一张或多张表** 通过 CDC 链路实时同步到 DataBuddy 数据湖仓。任务一次配置即可覆盖整库范围或多张指定表，支持自动建表、字段变更（DDL）同步与运行中自动加减表。本文介绍任务的完整配置流程。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有数据接入相关权限（详见 成员与权限管理 ）。
- 已创建并通过连通性测试的来源端 **数据源连接**（详见 添加数据源连接 ）。
- 已创建用于实时同步的 **实时集群** 计算资源（详见 创建计算资源 ）。
- 来源端已开启 CDC 能力（如 MySQL Binlog 设置为 ROW、Oracle LogMiner、SQL Server CDC 等），具体要求参见各数据源章节。
- 目标 Catalog 已规划好；如需自动建表，当前账号需要拥有目标 Catalog 的写入与建表权限。

使用限制

限制项	说明
来源端类型	MySQL、腾讯云 MySQL、TDSQL-C MySQL、PostgreSQL、Oracle、SQL Server、Kafka、CKafka、MongoDB、腾讯云 MongoDB
目标端类型	DataBuddy 数据湖仓 Catalog 下的 Iceberg 原生表
计算资源	仅支持 **实时集群** 计算资源
资源占用	单任务范围 1 ~ 256 CU，步长 1 CU
来源表数量上限	单任务最多选择 5000 张表
自动建表	Kafka 作为来源端时不支持自动建表，需在目标端预先建表
映射匹配	Kafka、MongoDB 不支持映射匹配预览

预览	
库表自动加减	仅 **正则匹配** 模式支持运行中自动接入新匹配的表；**指定表** 模式需暂停任务后手动加减

操作步骤

实时整库多表任务采用向导式配置，包含 **基本信息** → **来源** → **目标** → **数据对账** → **运行设置** 五个步骤。

步骤 1：进入任务创建页

1. 登录 DataBuddy 控制台 。
2. 从工作空间列表进入目标工作空间。
3. 进入 **数据集成** > **接入任务** ，点击 **创建任务** 。
4. 选择数据源类型后，在接入方式中选择 **实时整库多表接入** 。

步骤 2：基本信息设置

填写任务基本信息后点击 **下一步** 。参数详见 基本设置 。

步骤 3：来源设置

1. 选择数据源连接，可进行连通性测试。失败时点击 **重试** ，可查看具体原因。
2. 选择 **来源表** ：
   - **指定表** ： 通过库名、表名过滤后勾选具体表。
   - **正则匹配表** ： 填写库名、表名正则表达式，运行期内匹配的新增表会自动接入。
1. 选择 **读取模式** ： **全量、增量** （默认）或 **仅增量** 。
2. 在 **订阅类型** 中至少保留一个：插入、更新、删除。
3. 如需设置分片参数、控制元数据字段写入等高级行为，展开 **高级设置** 配置。

参数详见 来源设置 。

> **提示**
> 不同数据源类型支持的配置有差异，具体内容见数据源列表下的各种数据源类型文档，如 Kafka 数据源

步骤 4：目标设置

1. **基本配置** ： 选择目标 Catalog 与系统字段。
2. **建表配置** ： 当前仅支持 Iceberg 原生表。Kafka 来源时本节置灰，需手动建表。
3. **映射匹配预览** ： 检查每张来源表与目标表的映射关系。可点击 **编辑** 调整字段映射、目标 Schema/Table、同步主键。

参数详见 目标设置 。

> ⓘ **注意**
> **修改字段映射后，DDL 变更不再支持新增列、删除列、重命名列、修改列类型** 。如果来源表发生此类变更，任务可能报错。仅在确实需要差异化映射时使用。

步骤 5：数据对账（可选）

按需配置数据对账规则。详细配置项与限制参见 实时同步任务数据对账 。

步骤 6：运行设置

1. 配置 **DDL 消息处理策略**： 决定收到来源端 DDL 时是否自动同步、忽略或停止任务。（部分数据源无此配置，具体以实际页面为准。）
2. 配置 **任务运行策略**： Checkpoint 间隔、写入异常策略、最大重启次数。
3. 配置 **告警通知**： 选择已有告警渠道并启用监控指标。详细规则参见 告警通知 。
4. 如需调整透传参数，展开 **高级设置** 。

步骤 7：保存或发布

操作	行为	任务状态
仅保存	保存配置但不发布	未发布
保存并发布	保存并提交发布	待启动

操作完成后页面回到任务列表。后续启动、重启、停止等运维操作详见 实时同步任务运维监控 。

参数说明

基本设置

参数	说明	是否必填	默认值
接入类型	选择来源端类型	是	—
接入方式	**实时整库多表接入 / 实时分库分表接入**	是	实时整库多表接入
任务名称	大小写字母、中文、数字、下划线，长度 ≤ 256	是	`${数据源类型}_YYYYMMDD_HHmmss`
计算资源	实时同步使用的 **实时集群** 计算资源	是	—

资源占用	CU 数量，范围 1 ~ 256，步长 1	是	—
负责人	任务运行账号；需具备目标端访问权限	是	当前创建用户
任务描述	≤ 200 字符	否	—

来源设置

参数	说明	是否必填
数据来源	当前 Workspace 下匹配类型的数据源连接	是
来源表	见 来源表的两种选择模式	是
读取模式	**全量+增量 / 仅增量**	是
订阅类型	**插入 / 更新 / 删除** ，至少保留一项	是
高级设置	数据源专属高级参数	否

来源表的两种选择模式

模式	行为	运行中自动加减表
指定表	按库名、表名过滤后勾选；最多 5000 张	✕ 需暂停任务后手动编辑
正则匹配表	填写库表正则；运行期内匹配的新增表自动接入	✓ 自动加减，加减的表必须在正则范围内

目标设置

基本配置

参数	说明	是否必填
数据位置	选择目标 Catalog；支持本地新建（需 Catalog 创建权限）	是
Schema 匹配策略	控制目标 Schema 的命名规则	是
表匹配策略	控制目标表的命名规则	是
系统字段	见 系统字段（元数据字段）	否

系统字段（元数据字段）

可勾选写入到目标表的内置字段：

字段名	类型	说明
`datasource_wedata_di`	string	数据源名称
`database_wedata_di`	string	数据库名称
`table_wedata_di`	string	数据表名
`sequence_id_wedata_di`	string	增量事件记录 ID，唯一且递增
`type_wedata_di`	string	操作类型
`ts_wedata_di`	timestamp	日志时间
`processing_time_wedata_di`	timestamp	处理时间（机器本地）
`before_image_wedata_di`	string	是否更新前的记录，取值 `Y` / `N`
`after_image_wedata_di`	string	是否更新后的记录，取值 `Y` / `N`

> **注意**
> 任务发布运行后再取消勾选某个系统字段，**目标表中的字段不会被删除**，但后续数据将写为空。请在保存前再次确认。

建表配置

参数	说明
建表类型	当前仅支持 **原生表（Iceberg）**
字段映射	平台按默认规则填充字段类型映射，支持按需修改

映射匹配预览

预览每张来源表与目标表的映射关系，列含序号、源库、源 Schema（如有）、源表、目标 Schema、目标表、同步主键、表创建方式、字段映射、操作。

项	说明
同步主键	自动填充：使用已有表填该表主键；自动建表填待建表主键。无主键时需选择 1+ 列作为联合替代主键
字段映射	默认 **默认映射**；点击 **编辑** 修改后变为 **自定义映射**（DDL 变更能力受限）

操作（移除）	仅 **指定表** 模式可用；正则匹配模式下移除按钮置灰
操作（编辑）	打开右侧抽屉，按目标 DataBuddy 类型展示并允许调整字段映射

数据量过大时不支持预览，可在任务发布运行后通过运维详情查看实际链路明细。

运行设置

DDL 消息处理策略

收到来源端 DDL 变更时的处理方式（不同链路支持的 DDL 类型见各数据源章节，如 MySQL 数据源 等）：

策略	行为
自动同步	在目标端自动执行对应的 DDL
忽略	跳过 DDL；后续数据可能因 Schema 不匹配失败

任务运行策略

参数	说明
Checkpoint 间隔	状态快照写入间隔
写入异常策略	单条写入失败时的处理方式
最大重启次数	任务异常自动重启的次数上限

告警通知

支持对接 通知渠道 。基础平台新增渠道后会自动出现在选项中。

告警类型	说明	关键阈值
任务失败	任务进入失败状态时告警	–
业务延迟	同步链路延迟超过阈值	延迟阈值 (秒)、持续时间 (分)
Failover 次数	观测窗口内重启次数超过阈值	观测窗口 (分)、重启次数 (次)
DDL 通知	发生选定的 DDL 类型时通知	类型多选：新增列、删除列、重命名列、修改列类型、删除表、清空表、重命名表

参数	含义	取值范围	默认	提示
延迟阈值(秒)	业务延迟告警的延迟下限	1 ~ 86400，步长 1	120	建议 ≥ Checkpoint 间隔的 2 倍
持续时间(分)	持续超过延迟阈值多久后告警	1 ~ 60，步长 1	5	设置过短可能因网络抖动产生误报
观测窗口(分)	Failover 统计窗口	5 ~ 60，步长 1	10	建议 10 ~ 30 分钟
重启次数(次)	观测窗口内重启次数告警阈值	1 ~ 20，步长 1	3	建议小于自动重启次数上限

数据对账相关告警参见 实时同步任务数据对账 。

任务版本管理

任务编辑页右上角支持查看版本：

- **保存** 即生成版本，记录保存时间与保存人。
- 点击具体版本进入只读态；点击 **回到此版本** 转为编辑态，编辑保存后产生新版本。
- 任务运行中的版本附 **生产** 标识。
- 任务列表中通过「修改负责人」「修改告警」入口的变更不会生成新版本。

验证结果

操作完成后，您可以通过以下方式确认：

1. 在 **数据集成** > **接入任务** > **实时** 列表中可看到任务，状态符合预期：
   - **未发布**： 仅保存。
   - **待启动**： 已发布，可点击 **启动** 进入运行。
1. 启动后状态依次为 **操作中** → **运行中** 。点击任务名称可进入运维详情查看运行进度（详见 实时同步任务运维监控 ）。
2. 在目标 Catalog 下查看自动创建的目标表，全量阶段完成后会有数据写入；增量阶段持续追加变更。

常见问题

Q：来源端新增了一张表，运行中的任务没自动接入？

A：仅在 **正则匹配表** 模式下，新增表落入正则范围才会自动接入。**指定表** 模式需要暂停任务，编辑添加新表后重新发布、启动。

Q：自动建表失败？

A：以下情况会导致自动建表失败：

- 当前账号没有目标 Catalog 的建表权限——请联系管理员授权。
- 来源表字段类型在目标端没有对应类型——可在 **映射匹配预览** 中点击 **编辑** ，手动调整字段类型后重试。
- Kafka 作为来源端时不支持自动建表——需要预先在目标端建表后再选择 **使用已有表** 。

Q：取消勾选系统字段后目标表的对应列还在?

A：是。系统字段在目标表中创建后不会因取消勾选而删除，仅停止写入新数据。如需彻底移除，请在数据接入任务外手动调整目标表 Schema。

Q：编辑过字段映射后再发生 DDL 变更，任务报错?

A：自定义字段映射模式下，DDL 变更（新增列、删除列、重命名列、修改列类型）不再自动同步。建议：

- 评估是否可以恢复为 **默认映射** 。
- 如必须保留自定义映射，则在来源端发生变更时同步手动调整目标表 Schema 与任务字段映射。

相关文档

- 实时分库分表任务配置
- 实时同步任务运维监控
- 实时同步任务数据对账
- 支持的接入方案
- 支持的数据源与读写能力
- Kafka 数据源
- MySQL 数据源
- 设置通知渠道
- 创建计算资源

实时分库分表任务配置

最近更新时间：2026-09-11 18:35:00

概述

实时分库分表任务用于把**多个数据源连接、多个数据库、多张物理分表**的实时变更聚合到 DataBuddy 数据湖仓的同一张目标逻辑表。常见场景如：业务采用了水平分库分表（按 user_id 哈希分布到 8 库 32 表），希望在数据湖仓中合并为单张大表用于分析。本文介绍任务的完整配置流程。

前提条件

在开始本操作前，请确保满足以下条件：

- 已开通 DataBuddy 服务并完成空间初始化。
- 当前账号在目标 Workspace 中拥有数据接入相关权限（详见 成员与权限管理 ）。
- 已为**所有**待接入的物理库创建数据源连接，并通过连通性测试（详见 添加数据源连接 ）。
- 已创建 **实时集群** 计算资源（详见 创建计算资源 ）。
- 在来源数据库中已开启 CDC 能力。
- 所有参与合并的物理分表 Schema 必须**一致**——字段名、类型、主键约束相同。

使用限制

限制项	说明
来源端类型	MySQL、腾讯云 MySQL、TDSQL-C MySQL、PostgreSQL、Oracle、SQL Server
目标端类型	DataBuddy 数据湖仓 Catalog 下的 Iceberg 原生表
计算资源	仅支持 **实时集群** 计算资源
资源占用	单任务范围 1 ~ 256 CU，步长 1 CU
物理表 Schema	同一张逻辑表对应的所有物理分表 Schema 必须一致
不支持的来源端	Kafka、CKafka、MongoDB（如需将其作为实时来源端，请使用 实时整库多表任务 ）

操作步骤

实时分库分表任务采用向导式配置，包含 **基本信息** → **来源** → **目标** → **数据对账** → **运行设置** 五个步骤，**与整库任务流程一致**，本文重点描述差异步骤；与整库任务一致的部分通过链接复用 实时整库多表任务配置 。

步骤 1：进入任务创建页

1. 登录 DataBuddy 控制台 。
2. 进入工作空间。
3. 进入 **数据集成** > 接入任务，点击**创建任务** 。
4. 选择需要接入的数据源类型后在接入方式中选择 **实时分库分表接入** 。

步骤 2：基本信息设置

填写任务基本信息后点击 **下一步** 。参数详见 参数说明 > 基本信息设置 。

步骤 3：来源设置

1. **数据来源**： 在下拉中**选择多个** 同类型数据源连接。
2. 配置 **逻辑表** 映射规则：将多个物理库、物理表声明为同一张逻辑表，作为目标端的合并维度（具体规则参见页面提示）。
3. 选择 **读取模式**： **全量+增量** 或 **仅增量** 。
4. 选择 **订阅类型**： 插入、更新、删除（至少一项）。
5. 如需设置高级参数，展开 **高级设置** 。

步骤 4：目标设置

目标端配置项与运行规则与 实时整库多表任务的目标设置 完全一致：

- **基本配置**： 选择目标 Catalog、Schema 匹配策略、表匹配策略、系统字段。
- **建表配置**： 当前仅支持 Iceberg 原生表，平台按默认字段映射规则生成建表语句，支持用户按需修改字段映射。
- **映射匹配预览**： 每条记录的「源库、源表」会显示逻辑库、逻辑表，其余列与整库任务一致。

步骤 5：数据对账

按需配置数据对账规则。详见 实时同步任务数据对账 。

> **提示**
> 分库分表场景下，对账会**汇总所有源端物理分表的条数、主键** 与目标逻辑表进行比对。如源端 8 库 32 表合并到目标 1 张表，会用 8×32 张物理表的总条数、主键去对比目标 1 张表。

步骤 6：运行设置

DDL 处理策略、任务运行策略、告警通知、高级设置与整库任务一致，详见 运行设置 。

步骤 7：保存或发布

操作	行为	任务状态

仅创建	保存配置但不发布	未发布
创建并发布	保存并提交发布	待启动

后续启动、重启、停止等运维操作详见 实时同步任务运维监控 。

参数说明

基本信息设置

参数	说明	是否必填	默认值
接入类型	选择来源端类型	是	目标端固定为 DataBuddy
接入方式	**实时分库分表接入**	是	实时整库多表接入
任务名称	大小写字母、中文、数字、下划线，长度 ≤ 256	是	`${数据源类型}_YYYYMMDD_HHmmss`
计算资源	实时同步使用的 **实时集群** 计算资源	是	—
资源占用	CU 数量，范围 1 ~ 256，步长 1	是	—
负责人	任务运行账号；需具备目标端访问权限	是	当前创建用户
任务描述	≤ 200 字符	否	—

来源设置

参数	说明	是否必填
数据来源	当前 Workspace 下匹配类型的数据源连接，**支持多选**	是
逻辑表	将多个物理库 + 物理表合并为一个逻辑表	是
读取模式	**全量+增量 / 仅增量**	是
订阅类型	插入、更新、删除，至少保留一项	是
高级设置	数据源专属透传参数	否

> **提示**
> 分库分表任务的 **运行身份** 必须对所有选定数据源连接都有访问权限。任意一个数据源访问失败都会导致任务失败。

目标设置

参数完全继承 实时整库多表任务的目标设置 。

运行设置

参数完全继承 实时整库多表任务的运行设置，包含 DDL 处理策略、Checkpoint 间隔、最大重启次数、告警渠道与告警类型。

验证结果

操作完成后，您可以通过以下方式确认：

1. 在 **数据集成** > **接入任务** > **实时同步** 列表中可看到任务，**任务类型** 列显示 **分库分表** 。
2. 启动任务后，进入运维详情可看到链路详情列表中每个 **物理库、物理表** 都对应一行映射记录，目标列均指向同一张逻辑表。
3. 数据对账概览中按目标端逻辑表统计 **数据一致表数量** ；详情中可查看源端总条数、主键与目标条数、主键的差异。

常见问题

Q：分库分表场景下能合并成多张目标表吗?

A：当前任务以 **逻辑表** 为合并单位，单个任务可定义多个逻辑表，每个逻辑表对应一组分库分表 → 一张目标表。如需将不同业务模块拆到多个目标表，可以在同一任务中定义多个逻辑表，或拆为多个任务。

Q：源端物理分表的 Schema 出现差异?（个别表多了字段）

A：当前任务建议所有参与合并的物理分表 Schema **完全一致** 。如个别表存在微小差异，如某个表多个字段，当前会将按源表字段的并集创建目标表。如果源表的 Schema 差异较大，当前建议：

- 在源端补齐差异，使所有分表对齐。
- 或将差异表拆出单独配置一条任务。

Q：源端某个数据源的实例做主备切换会怎样?

A：参考对应数据源章节的「主备切换」说明（如 MySQL 数据源 ）。整体规则：CDC 链路优先按 GTID 恢复；不支持 GTID 时，可能需要重启任务并指定恢复位点。

Q：分库分表任务支持运行中加表吗?

A：实时分库分表任务的「逻辑表」配置在保存后即固化。如需变更范围，需要暂停任务，编辑后重新发布、启动。

相关文档

- 实时整库多表任务配置
- 实时同步任务运维监控
- 实时同步任务数据对账
- 支持的接入方案
- 支持的数据源与读写能力
- MySQL 数据源
- Oracle 数据源
- 设置通知渠道

实时同步任务运维监控

最近更新时间：2026-09-11 18:35:00

概述

实时同步任务的运维监控覆盖任务从发布、启动、运行到停止的完整生命周期。本文介绍任务列表、运维详情页、运行日志、链路详情与 DDL 变更记录的查看方式，以及常用运维操作（启动、重启、停止、修改负责人、修改告警等）。

前提条件

- 已创建实时同步任务，详见 实时整库多表任务配置 或 实时分库分表任务配置 。
- 当前账号在目标 Workspace 中拥有任务的相应权限：**可查看 / 可编辑 / 可运行 / 可管理**（详见 成员与权限管理 ）。

任务状态

状态	说明	可执行操作（主操作 + 更多）
未发布	任务新建后未发布	发布、编辑、权限、删除
待启动	任务已发布尚未启动	启动、编辑、权限、删除（编辑保存后主操作变为「发布」）
操作中	启动、停止、重启的中间态	—
运行中	启动成功后的稳定运行态	停止、强制停止、重启、编辑、权限
运行中（有更新）	运行中的任务被重新编辑发布	重启（主操作）、停止、强制停止、编辑、权限
运行中（异常）	10 分钟内重启次数 > 1 次；状态后会附带异常原因	停止、强制停止、重启、编辑、权限
已停止	停止操作完成	启动、编辑、权限、删除
失败	启动失败（如启动超时）或运行失败（如达到最大重启次数）；状态后附带异常原因	启动、编辑、权限、删除

操作步骤

步骤 1：打开任务列表

1. 登录 DataBuddy 控制台 。

2. 在顶部切换到目标地域和 Workspace。
3. 进入 **数据接入** > **任务管理** > **实时同步** 。

任务列表展示任务名 / ID、任务类型、状态、数据来源、负责人、创建人、更新时间。支持按任务名称 / ID 模糊搜索，按来源类型、数据源连接、目标 Catalog、计算资源筛选。

步骤 2：发布任务

1. 选择 **未发布** 任务，点击 **发布** 。
2. 在发布弹窗中选择：
   - **仅发布** ： 发布后任务进入 **待启动** 状态。
   - **发布并启动** ： 发布后直接进入启动流程（详见 步骤 3 ）。

步骤 3：启动任务

启动分为三步：检测 → 选择启动策略 → 查看进度。

1. **任务启动检测** ： 平台检查任务配置完整性等。
2. **选择启动策略** ： 根据任务当前状态可选不同策略，详见 启动、重启策略 。
3. **查看进度与结果** ： 可点击 **查看运行详情** 进入任务运维详情页，或点击 **关闭** 回到列表。

启动、重启策略

当前状态	可选启动策略	备注
待启动	立即启动，从默认位点同步；立即启动，指定位点同步；立即启动，指定时间点同步	仅 MySQL / 腾讯云 MySQL / TDSQL-C MySQL 数据源支持指定位点和指定时间点
运行中 / 运行中（有更新） / 运行中（异常）	继续运行，保留作业状态从最后位点继续；重新启动，从指定时间点 / 指定位点继续；重新启动，丢弃作业状态从默认位点开始	「继续运行」适用于配置变更后无需丢弃状态的场景
已停止	同上四种	—
失败	从上次运行失败的 Checkpoint 位点恢复；重新启动，从默认位点开始	—

步骤 4：查看运维详情

点击任务名称进入运维详情页：

- **任务名称导航区** ： 返回列表、复制任务名称、收藏。
- **任务操作区** ： 启动、停止、编辑、删除、权限（删除与权限折叠到「更多」中）。
- **任务基本信息区** ： 任务名称、ID、类型、数据源、计算资源、运行身份、启动时间、创建人、创建时间。数据源与计算资源支持点击跳转。

- **运行结果**：见 运行结果 。
- **运行日志**：见 运行日志 。
- **数据对账**：见 实时同步任务数据对账 。
- **配置详情**：展示最新发布版本的任务配置。

步骤 5：常用运维操作

操作	入口	行为
启动	列表「启动」/ 详情页操作区	见 步骤 3
停止	列表「停止」/ 详情页操作区	优雅停止
强制停止	列表「强制停止」/ 详情页操作区	跳过优雅流程立刻终止
重启	列表「重启」/ 详情页操作区	等价于停止 + 启动
编辑	列表「编辑」/ 详情页操作区	进入任务编辑页
权限	列表「权限」/ 详情页「更多」	见 权限管理
修改负责人	列表「更多 → 修改负责人」	见 修改负责人
告警	列表「更多 → 告警」	直接修改告警渠道 / 告警类型 / 阈值，无需重新发布
删除	列表「更多 → 删除」	二次确认后删除；运行中的任务需先停止

权限管理

任务权限按 ACL 实体访问控制，可对成员分配 **可管理 / 可运行 / 可编辑 / 可查看** 等权限。具体权限语义参见 成员与权限管理 。

修改负责人

负责人是任务实际的运行账号。运行中的任务修改负责人后**需要重启任务才生效** ——重启之前列表负责人列后会出现 info 图标提示「负责人已被修改，需要重启后生效」。从编辑任务修改负责人时还需要先发布再启动。

运行结果

任务执行状态与进度

实时任务按 **环境准备 → 全量同步 → 增量同步** 三个阶段展示进度：

阶段	含义	进度展示
环境准	任务初始化（资源就绪、连接建立	一次性阶段

备	等）	
全量同步	将源表所有存量数据同步到目标表	进度百分比 + 已完成 / 总库表数
增量同步	持续不间断同步源端实时变更	当前延迟（如「延迟 2 秒」），停止时显示「已停止」

链路详情

以列表形式展示每张表的同步情况：来源库、来源 Schema（如有）/ 来源表、目标 Catalog / 目标 Schema / 目标表、目标对接方式、建表 DDL / 读取条数、写入条数、更新条数、删除条数、状态。

行为	说明
过滤搜索	来源 / 目标的库、Schema、表、目标对接方式、状态均可过滤
排序	读取条数 / 写入条数 支持排序
状态	**正常** / **异常**（运行异常会展示具体原因）
跳转	点击目标表新窗口进入 Catalog 表详情页
建表 DDL	目标对接方式为「自动建表」时显示完整建表 DDL；为「使用已有表」时显示 —

DDL 变更记录

仅来源类型为 **MySQL / 腾讯云 MySQL / TDSQL-C MySQL** 时支持。其他来源类型显示「当前链路暂不支持 DDL 变更记录」。
记录字段：来源库、来源 Schema（如有）/ 来源表、目标 Catalog / 目标 Schema / 目标表、变更时间、变更策略、来源表 DDL 语句、目标表 DDL 语句、变更状态。

变更状态	出现条件
成功	变更策略为「自动变更」且执行成功
失败	变更策略为「自动变更」但执行失败；行末 info 图标可查看失败原因，任务默认忽略此次变更
忽略	变更策略为「忽略」

运行日志

能力	说明
历史日志	任意状态（包括停止后）都支持查看历史日志

快捷诊断	文本匹配 + 关键词高亮，可快速定位匹配内容
自动刷新	默认关闭；如需查看最新日志请手动点击刷新
日志查询	支持按关键词搜索；按日志级别和时间过滤

数据对账

数据对账结果展示在 **数据对账** 中。详细配置规则、告警与失败处理参见 实时同步任务数据对账 。

验证结果

完成运维操作后，您可以通过以下方式确认：

1. **任务列表状态** 已变更为目标状态（如启动后变为 **运行中** ，停止后变为 **已停止** ）。
2. **任务详情页** 中环境准备、全量同步、增量同步阶段的进度更新符合预期。
3. **链路详情** 列表中的每张表读取、写入条数持续增长（增量同步阶段）。
4. **运行日志** 没有异常事件。

常见问题

Q：任务停留在「操作中」迟迟无变化?

A：操作中是启动、停止、重启的中间态，通常会在分钟级内自动转入下一个状态。如长时间停留：

1. 查看运行日志是否有错误信息。
2. 检查计算资源是否处于异常状态（详见 计算资源用量监控 ）。
3. 如确认卡死，可联系技术支持处理。

Q：「运行中（异常）」状态如何排查?

A：运行中（异常）通常因 10 分钟内重启次数 > 1 次触发。请按以下顺序排查：

1. 状态后的异常原因提示。
2. 运行日志中最近一段错误堆栈。
3. 链路详情中过滤异常表，查看原因。
4. DDL 变更记录中是否存在「失败」的变更（变更策略为「自动变更」且执行失败时，任务默认忽略，但可能导致后续数据写入异常）。
5. 查看告警通知，确认是否已超出 Failover 重启次数阈值。

Q：修改负责人后任务还在用旧账号运行?

A：负责人修改后**需要重启任务才生效** 。请在合适时机点击 **重启** ，否则继续以原负责人身份运行。如果是从编辑任务页修改的，还需要先发布再启动。

Q：删除任务后能恢复吗?

A：删除是终态操作，请通过「停止 → 确认无依赖 → 删除」的顺序谨慎操作。运行中的任务无法直接删除，需要先停止后再删除。

Q：能否一次启动多个任务?

A：当前任务列表不提供批量启动入口，每个任务需要单独启动。

相关文档

- 实时整库多表任务配置
- 实时分库分表任务配置
- 实时同步任务数据对账
- 离线同步任务运维监控
- 设置通知渠道
- 成员与权限管理
- 计算资源用量监控

实时同步任务数据对账

最近更新时间：2026-09-11 18:35:00

概述

数据对账用于在来源端与目标端之间周期性比对数据，及时发现数据不一致问题。DataBuddy 实时同步任务支持基于 **数据条数** 与 **数据内容（主键）** 的两种对账规则，覆盖全量与增量两种范围，并提供告警通知与差异详情查看。本文介绍数据对账的支持范围、配置方式与结果查看。

前提条件

- 已创建实时整库或实时分库分表同步任务。
- 来源、目标端在 支持的链路 范围内。
- 如使用「校验数据内容」规则，源表必须包含主键或唯一索引。

使用限制

限制项	说明
任务类型	实时整库迁移任务 / 实时分库分表任务支持
主键	「校验数据内容」仅支持基于主键的对账，无主键且无唯一索引的表不支持
MongoDB	MongoDB 来源端**不支持增量对账**（仅支持全量对账）
单任务库表数量	任务涉及库表数量超过 5000 时，无法按单表配置数据对账（不影响整体同步任务运行）
对账周期	最小对账间隔为 60 分钟
差异详情预览	单次最多预览 10000 条差异数据，且仅展示主键字段
资源占用	对账任务与同步任务**复用同一计算资源**，会占用资源组资源
目标端筛选条件	增量对账的筛选 SQL 在 Flink SQL 中执行，仅支持 Flink SQL 语法（如当前日期写作 `${YYYY-MM-dd}` 而不是 `CURRENT_DATE()`，不等于使用 `<>` 而不是 `!=`）

支持的链路

任务类型	来源端
整库迁移任务	MySQL、TDSQL-C MySQL、Oracle、MongoDB、Kafka

分库分表任务	MySQL、TDSQL-C MySQL、Oracle

操作步骤

步骤 1：开启数据对账

数据对账在任务配置流程的 **数据对账** 步骤中配置：

1. 在 实时整库多表任务配置 或 实时分库分表任务配置 流程中进入 **数据对账** 步骤。
2. 打开 **数据对账** 开关。关闭时所有表都不进行对账；开启后按全局对账策略统一对账。

步骤 2：配置全局对账策略

参数	说明	是否必填
对账规则	**校验数据条数** / **校验数据内容**，详见下方说明	是
对账字段	仅在选「校验数据内容」时显示，**当前仅支持主键对账**	视规则
对账范围	**全量对账** / **增量对账**，详见 对账范围	是
对账时间	对账周期，**最小 60 分钟**	是

对账规则

规则	行为	适用场景
校验数据条数	通过数据源连接直接 `count(*)` 比较来源表与目标表条数	数据量不大，关注总量一致
校验数据内容	比对来源端与目标端的主键集合，主键不一致即认为数据不一致	主键稳定且关注每条记录是否落库；要求来源 / 目标端有主键或唯一索引

对账范围

范围	行为	用户输入
全量对账	对源 / 目标表全量数据对账	—
增量对账	通过过滤条件筛选后的数据进行对账，适合按时间分区	来源端筛选条件 + 目标端筛选条件

增量对账提示：

- 时间字段筛选支持 系统内置参数。
- 支持细粒度对账（如按小时分区），可使用分组语法。
- 来源、目标端筛选条件需各自符合对应数据源的 SQL 语法。

- 如筛选条件中字段在来源、目标端不存在，需要在「单表对账」中按表修改，否则该表对账会失败。

示例（按小时分区对账）：

```
select count(1), partition from table where partition >
${yyyyMMddHHmmss-8H} group by partition
```

> **提示**
> **目标端的筛选 SQL 在 Flink SQL 中执行**，仅支持 Flink SQL 语法：
> - 当前日期使用 `${YYYY-MM-dd}` 表达，**不要写** `CURRENT_DATE()` 。
> - 不等于使用 `<>`，**不要写** `!=` 。
>
> 不符合 Flink SQL 语法时，对账会失败。

步骤 3：（可选）按单表自定义对账

全局策略对任务下所有表生效。如需差异化处理：

1. 在配置流程的对账步骤中点击需要差异化的单表。
2. 可单表开启、关闭数据对账，或修改单表的增量对账条件。

> **提示**
> 单表对账修改在任务发布运行后生效。如某张表需要不同的筛选条件或不参与对账，建议在保存前调整完整。

步骤 4：（可选）配置对账告警

实时同步任务的告警规则页面支持配置数据对账告警。**仅对开启了数据对账的任务生效** 。

告警指标	说明	阈值
数据对账 / 任务失败	数据对账常驻任务运行失败	失败时按 X 分钟 / 小时 / 天告警 N 次
数据对账 / 对账失败	对账明细中某些表为对账失败状态	失败时按 X 分钟 / 小时 / 天告警 N 次
数据对账 / 对账时间	单次数据对账时间超过阈值	对账时间 > X 分钟 / 小时 / 天
数据对账 / 数据不一致条数	对账明细中某些表的数据差异条数不为 0	数据不一致条数 > N 条

> 「数据不一致条数」按**表粒度**告警——任意一张表的不一致条数超阈值即触发。

告警渠道与基础规则参见 设置通知渠道 。

查看对账结果

实时同步任务按设置的首次对账时间和对账间隔进行周期持续对账。

入口

进入任务运维详情页（见 实时同步任务运维监控 ），切换到 **数据对账** 页面。

对账概览指标

指标	说明
表总数	同步任务涉及的表数量；分库分表场景下按目标端表数量统计
数据一致表数量	来源 / 目标条数或主键完全一致的表数量
数据不一致表数量	来源 / 目标存在差异（差异数 ≠ 0）的表数量
对账失败数量	对账过程中失败的表数量
未对账表数量	未开启数据对账的表数量

对账明细

明细按表逐行展示对账结果。基于主键对账且差异数 ≠ 0 的记录支持 **查看差异详情**：

项	说明
差异详情预览上限	10000 条
展示字段	仅展示主键字段

对账任务状态与可执行操作

对账任务是常驻任务，按设置的时间周期自动运行：

状态	说明	可执行操作
对账任务未启动	尚未启动	运行
对账任务运行中	正在执行	停止
对账任务已停止	人为停止	运行
对账任务失败	启动失败或运行失败	运行
对账任务操作中	中间态	—

验证结果

操作完成后，您可以通过以下方式确认：

1. 在 **数据对账** 概览看到 **表总数** 与各分类计数符合预期，**未对账表数量** 等于未配置对账的表数。
2. 「对账明细」按周期产生新的对账记录。
3. 已配置对账告警时，触发条件下能在告警渠道收到通知。
4. 单表自定义对账配置生效——查看明细时该表的对账规则与全局不同。

常见问题

Q：开启数据对账后会拖慢任务性能吗？

A：会有一定影响。数据对账需要拉取来源端与目标端数据，**会对数据源有一定负担**，并且对账任务与同步任务**复用同一计算资源**，会占用资源组资源。建议：

- 把对账周期设置得不要过短（最小 60 分钟）。
- 对大表使用 **校验数据条数**，对中小表再使用 **校验数据内容**。
- 资源组监控值会包含数据对账实例占用，请按实际负载评估资源组规格。

Q：单表对账失败，怎么排查？

A：常见原因：

1. **筛选条件中字段在来源端或目标端不存在** —— 在单表对账配置中调整该表的增量对账条件，使字段名匹配。
2. **目标端筛选 SQL 使用了非 Flink SQL 语法** —— 如 `CURRENT_DATE()`、`!=`。改为 `${YYYY-MM-dd}`、`<>`。
3. **MongoDB 配置了增量对账** —— MongoDB 不支持增量对账，请改为全量对账或在该表关闭对账。
4. **任务涉及库表 > 5000，无法按单表配置** —— 当前规模无法支持单表对账，请拆分为多个任务。

Q：基于主键对账，差异条数为 0 但内容仍可能不一致？

A：是。当前 **校验数据内容** 仅比对主键集合，不比对其他字段值。如果业务需要比对全字段一致性，建议：

- 通过数据质量任务增加字段值校验（详见 数据质量）。
- 或通过 SQL 任务自定义比对脚本（详见 SQL 任务）。

Q：差异详情只显示 10000 条，怎么处理超出部分？

A：差异详情仅用于快速定位差异样本。**超出 10000 条** 时建议：

1. 通过差异样本判断不一致原因（如 DDL 漏同步、过滤条件遗漏、写入失败等）。
2. 修复同步任务后重新启动一次对账。
3. 如需全量比对，建议另起一个 SQL 任务，将来源、目标主键集合分别落到中间表后做差集。

Q：对账任务能单独停止，不影响数据同步吗？

A：可以。对账任务是常驻任务，可以在 **数据对账** 单独点击 **停止** ，**不影响数据同步主任务** 。停止后再点击 **运行** 即可恢复。

相关文档

- 实时整库多表任务配置
- 实时分库分表任务配置
- 实时同步任务运维监控
- 创建计算资源
- 设置通知渠道
- 系统内置参数

文件上传
本地上传到表
本地文件上传到表

最近更新时间：2026-09-11 20:42:32

概述

本地文件上传到表用于把本地的结构化、半结构化文件解析为结构化数据表，并落入 Catalog（数据目录）下指定的 Schema。上传完成后，表在 数据目录 中可被检索，可直接用于 SQL 查询、仪表盘与下游任务加工。
控制台的入口卡片名称为 **文件上传到表** ，交互形式为右侧抽屉，整体流程为「选择文件 → 预览数据并确认 Schema → 存储设置 → 建表」。

适用场景

- **一次性数据分析**： 把手工整理的销售表、库存表、对账文件等直接建表，用于临时分析。
- **样本数据快速入湖**： 把小规模样例数据入表，用于验证 SQL 逻辑或搭建仪表盘原型。
- **无需建模的散落文件**： 源数据只有文件、没有可连接的数据库时，直接由文件生成表。

如果需要**周期性**把新增文件持续拉取入表，请改用 文件到表同步 。如果文件需要保留原始格式、不做解析，请使用 本地文件上传到卷 。

前提条件

条件	说明
数据集成权限	当前账号在所属 Workspace（工作空间）中拥有数据集成模块的操作权限，详见 成员与权限管理
计算资源	当前 Workspace 下存在可用的交互式集群计算资源，用于解析文件与建表，详见 创建计算资源
目标 Catalog 与 Schema	拥有目标 Catalog、Schema 的写入权限；如没有可用 Schema，可在存储设置步骤中即时创建
待上传文件	文件类型与容量符合 支持的文件类型及约束

使用限制

限制项	说明

文件类型	CSV、JSON、Avro、Parquet、Text、XML，对应扩展名 `.csv` 、 `.json` 、 `.avro` 、 `.parquet` 、 `.txt` 、 `.xml`
文件数量	单次至少 1 个，最多 10 个，允许多种类型混合
文件容量	所有文件大小之和不超过 2 GB
预览行数	数据预览最多展示 50 行

各类型的解析行为、字段命名规则与完整错误提示，详见 支持的文件类型及约束 。

操作步骤

步骤 1：进入文件上传到表

1. 登录 DataBuddy 控制台，切换到目标 Workspace。
2. 在左侧导航栏选择 **数据集成** ，确认当前位于「 接入方式 」Tab。
3. 在「 文件 」区域点击 **文件上传到表** 卡片，系统在右侧打开「 文件上传到表 」抽屉。

步骤 2：选择本地文件

支持两种方式添加文件：

- **拖拽** ： 把文件直接拖到抽屉中的文件上传区。
- **点击选择** ： 点击 **选择文件** ，在系统文件浏览器中单选或多选文件。

> ⓘ **注意**
> 文件数量或总容量超出限制时，页面会给出非阻断式提示并在数秒后自动消失，超限文件不会被加入列表。请先拆分或精简文件后重试。

步骤 3：选择服务资源

服务资源用于解析上传文件并执行建表，为必选项。

参数	说明	是否必填	默认值
服务资源	单选。下拉列表仅展示当前账号有使用权限的交互式集群计算资源，并显示资源的当前状态	是	—

- 选中资源后，页面展示该资源的创建人与资源类型；点击 **详情** 可在新标签页打开计算资源详情页。
- 如所选资源当前未启动，系统会在解析文件前自动启动该资源。

> **提示**
> 计算资源按实际使用量计费，资源规格与自动启停策略详见 计算资源概述 。

步骤 4：预览数据与确认 Schema

上传文件以后自动进入「数据预览」步骤。系统先把文件写入湖仓再解析，解析完成后，预览表格展示数据样本，各项操作恢复可用。
在本步骤中，您需要确认以下内容：

1. 确认表名：表名不能跟现有表重复。
2. 确认字段：勾选需要落库的字段，并按需修改字段名、字段类型与 DECIMAL 精度。至少需要保留 1 个字段。
3. 校正解析规则：点击 **高级设置** ，按文件类型调整分隔符、行标签解析参数。

字段编辑的完整说明详见 预览与 Schema 编辑；各文件类型可调整的解析参数详见 文件解析与推断 。

> **注意**
> 修改「高级设置」会触发重新解析，已手动调整的字段名与字段类型可能恢复为系统推断结果。建议先把解析规则调整稳定，再编辑字段。

步骤 5：存储设置

确认预览结果后进入「存储设置」步骤，指定表的落库位置与同名处理方式。

参数	说明	是否必填	默认值
目标位置	在目录树中选择 Catalog 与 Schema，表将创建在所选 Schema 下。仅展示您有权限的目录	是	—
处理方式	**同名覆盖**：用本次数据覆盖已存在的同名表；**新建副本**：另建一张新表，名称为「表名_1」	是	—

- 目录树支持按关键字搜索 Catalog 与 Schema。
- 鼠标悬停在目录上时，右侧出现 **+ Schema** ，点击可直接创建 Schema，无需退出当前流程。
- 点击 **数据目录** 可在新标签页浏览完整的数据目录。

> **警告**
> 选择「同名覆盖」会用本次上传的数据替换目标表中的原有数据，覆盖后数据不可恢复。如需保留原表数据，请选择「新建副本」。

步骤 6：建表并验证结果

1. 点击 **建表** 。建表期间当前步骤的所有选项不可操作。
2. 建表成功后，页面自动跳转到 数据目录 中该表所在的位置。
3. 在数据目录中确认表已生成、字段与预期一致，随后即可通过 SQL 查询该表，表的元数据管理详见 数据目录概述 。

复用最近一次上传

打开「文件上传到表」抽屉时，如果您在 24 小时内有成功提交过的上传记录，页面会提示最近一次上传的文件，点击 **预览文件** 可直接沿用该批文件继续后续步骤，无需重新选择本地文件。

特性	说明
有效期	24 小时，超期后记录不再展示
可见范围	仅当前账号自己的上传记录，其他成员的上传记录不会展示

常见问题

Q：单次上传超过 10 个文件或总容量超过 2 GB 怎么办？

A：分批多次上传；如果数据量较大或需要周期性接入，请改用 文件到表同步 通过离线同步任务处理。

Q：多个文件字段不完全一致，能合并成一张表吗？

A：可以。合表上传时字段取各文件字段的并集，缺少该字段的数据行填充为 null。若各文件结构差异较大，建议改用分表上传，避免产生大量空值字段。

Q：JSON、XML 中的嵌套结构会展开成多列吗？

A：不会。嵌套内容整体解析为一个文本字段。如需拆解，可在入表后通过 SQL 函数处理，详见 文件解析与推断 。

Q：源文件字段名是中文，建表后字段名会变化吗？

A：会。系统会自动把字段名规范化，中文转为拼音、非法字符替换为下划线。您可以在数据预览步骤手动改为需要的名称，规则详见 支持的文件类型及约束 。

Q：上传的表结构后续还能调整吗？

A：建表后的字段调整需要在数据目录或通过 SQL 任务完成。如果只是本次上传解析有误，建议在数据预览步骤修正后重新建表。

Q：提示「无权限执行上传服务」怎么办？

A：通常是数据集成模块权限或计算资源使用权限发生变更。请联系管理员确认权限配置，详见 成员与权限管理 。

相关文档

- 支持的文件类型及约束
- 文件解析与推断
- 预览与 Schema 编辑
- COS 文件上传到表
- 本地文件上传到卷
- 文件到表同步
- 计算资源概述
- 数据目录概述

支持的文件类型及约束

最近更新时间：2026-09-11 20:42:32

概述

本文列出 本地上传到表 支持的文件类型、单次上传约束以及常见错误提示，帮助您在上传前对文件做必要的预处理与拆分。

支持的文件类型

文件类型	扩展名	说明
CSV	`.csv`	逗号分隔的结构化文本，常见的导出格式
JSON	`.json`	JSON 文档；支持单条对象、数组、多对象（每行一条）三种排列方式
Avro	`.avro`	自描述的二进制行存格式，无需额外指定字段
Parquet	`.parquet`	自描述的二进制列存格式，无需额外指定字段
Text	`.txt`	自定义分隔符的纯文本文件
XML	`.xml`	XML 文档，需指定行标签（系统会自动推断）

各类型的解析行为与可调整参数详见 文件解析与推断 。

> **提示**
> 单次上传支持**多种类型混合**，例如可以一次上传 1 个 CSV + 2 个 JSON。但建议同一批次上传**结构相似**的文件，否则在合表模式下会产生大量空值字段。

单次上传约束

约束项	限制值	校验提示
文件数量	至少 1 个，最多 10 个	「最多可上传 10 个文件」
文件总容量	所有文件大小之和 ≤ 2 GB	「最多可上传 2GB 文件」

字段命名规则

源文件字段可能存在中文、非法字符等不符合作为建表信息。因此 DataBuddy 在文件上传到表流程中会**自动对字段名进行规范化**，您也可在「数据预览」阶段手动调整：

原始字段	处理后字段	规则
空格	清空	字段名仅含空格时清空；如全部字段名为空，依次赋予 `c_1` 、 `c_2`
中文	拼音	例如「名称」→ `mingcheng`
非合法字符	`_`	例如 `a&b` → `a_b`

合法字符的定义：字母、数字、下划线、 `/` 、 `=` 、 `-` 。
字段名输入框校验规则：

规则	说明
起始字符	必须是字母、数字或下划线
字符范围	仅允许字母、数字、下划线、 `/` 、 `=` 、 `-`
长度	不超过 64 个字符

表名也按相同规则处理：原始文件名解析后自动输入表名。

常见错误提示

场景	提示	处理建议
超出文件数量	「最多可上传 10 个文件！」	删除部分文件后重试
总容量超出	「最多可上传 2GB 文件」	拆分文件或仅保留必要的子集
文件类型不支持	无法选择文件	检查文件类型是否在支持列表中或者先转换下类型
权限不足	「无权限执行上传服务」	检查模块权限或服务资源权限是否变更，详见 成员与权限管理
兜底错误	「上传失败 — {错误详细信息}」	根据具体错误描述排查

相关文档

- 本地上传到表
- 文件解析与推断
- 预览与 Schema 编辑
- COS 文件上传到表

- 文件到表同步

文件解析与推断

最近更新时间：2026-09-11 20:57:30

概述

DataBuddy 在「数据预览」阶段会**自动推断**文件的结构与字段类型，您可以通过**转换设置**入口对推断结果进行校正。本文按文件类型逐一说明可调整的参数。
转换设置的入口位于「数据预览」页面字段表上方，点击**高级设置**弹出参数面板。

CSV

配置	默认值	说明
列分隔符	自动推断；无法推断时回退为 `,`	可选：自定义、`,`、`;`、Tab
转义符	`\`	可选：`"` 或 `\`。用于对值中出现的非法字符进行转义存储
第一行作为字段行	勾选	取消勾选时字段名自动生成为 `t__c0` / `t__c1` …，原始第一行作为数据行写入
行跨行	不勾选	是否允许单条记录跨越多行（用于值中含换行的情况）

> **提示**
> **例：源文件部门字段值为 `<script://alert('abc')>`。当转义符为 `"` 时，存储为 `<script:"/"/alert("'abc"')>`；当转义符为 `\` 时，存储为 `<script:\/\/alert(\'abc\')>`。**

Text

配置	默认值	说明
列分隔符	自动推断；无法推断时回退为 `|`	可选：自定义、`;`、`|`、Tab
行分隔符	自动推断；无法推断时回退为「断行 `\n`」	可选：自定义、回车 `\r`、断行 `\n`
第一行作为字段行	勾选	取消勾选时字段名自动生成为 `t__c0` / `t__c1` …，原始第一行作为数据行写入
转义字符	`\`	用于转义非法字符

JSON

配置	默认值	说明
将时间文本转换为时间类型	勾选	当字段值符合时间特征时，自动识别为时间类型；否则解析为文本
允许使用单引号	勾选	勾选时把单引号当作双引号解析；取消勾选且文本含单引号时会提示「无法推断出 JSON 格式」
过滤注释	勾选	仅支持单行 `//` 与多行 `/* */` 标准注释；不支持 `?` 注释、约定 key 等非标准注释
行跨行	勾选	勾选时只关注完整 JSON 体，忽略不完整的行；取消勾选时不完整行解析为空行（所有字段为 null）

JSON 推荐格式：

- 多 JSON 体形式：每行一条独立 JSON 对象。
- 数组形式：根元素为 JSON 数组，每个数组元素是一条记录。

嵌套处理： 当存在 JSON 嵌套时，嵌套字段整体解析为一个文本对象，**不会自动展开为多列**。如需拆解，建议在入表后通过 SQL 任务用 `get_json_object` / `from_json` 等函数后处理。

XML

配置	默认值	说明
行标签	自动推断（取最外层一级标签）	决定哪个标签作为一行数据的分隔；如自动推断错误可手动指定
属性作为字段	勾选	勾选时把 XML 元素的属性作为单独字段；取消勾选时属性被忽略
属性前缀	`_`	「属性作为字段」勾选时，字段名形式为 `属性前缀 + 属性名`

> **提示**
> **行标签嵌套行为：** 若 XML 中行标签出现在嵌套结构中，仅以行标签定义的层级作为数据行；该层级的父级数据视为无意义数据，不会写入字段。

属性前缀冲突处理： 若「属性前缀、属性名」与已有字段名重名，自动在末尾追加 `_序号`（如 `a_name_1`），以避免覆盖。

Avro / Parquet

Avro 与 Parquet 是**自描述格式**，文件中已包含 Schema 信息。系统直接读取并展示推断结果，无需额外转换设置。
如需调整字段类型或字段名，可直接在「数据预览」页面的字段表上修改，详见 预览与 Schema 编辑 。

字段类型自动推断

预览阶段，系统会基于前若干行数据为每个字段自动推断类型。当前支持的字段类型包括（在字段类型下拉菜单中可选）：

类型族	说明
字符串	STRING / VARCHAR
整数	INT / BIGINT
浮点	FLOAT / DOUBLE
高精度数	DECIMAL（需指定精度与小数位）
布尔	BOOLEAN
时间	DATE / TIMESTAMP

DECIMAL 精度设置：当字段类型设为 DECIMAL 时，悬停在类型上会显示编辑图标；点击后可设置精度与小数位。

参数	取值范围	说明
精度（precision）	1 ~ 38	小数点左侧的最大位数
小数位（scale）	0 ~ 精度	小数点右侧的最大位数。超过精度上限时会以红色错误提示

「高级设置」与字段编辑的关系

- 「高级设置」修改的是**解析规则**（如分隔符、行标签）。修改后系统会重新解析文件，预览数据与字段列表会刷新。
- 「字段名、字段类型、建表字段」修改的是**目标表的 Schema** 。这些修改不会再次触发解析，仅影响最终建表的字段定义。

如果在调整「高级设置」前已手动改过字段名或类型，重新解析后字段列表可能恢复为系统默认推断结果。建议**先把转换规则调整稳定，再编辑字段** 。

相关文档

- 本地上传到表
- 支持的文件类型及约束

- 预览与 Schema 编辑
- 文件到表同步 – 解析配置

预览与 Schema 编辑

最近更新时间：2026-09-11 20:57:30

概述

「数据预览」是 本地上传到表 流程中介于「上传文件」与「存储设置」之间的关键步骤。在此页面您可以：

- 调整**字段名、字段类型、分表字段**，最终决定落库 Schema。
- 查看前 50 行数据样本以验证解析结果。

建表方式

解析文件后，可以选择如何建表：

方式	行为	适用场景
创建新表	在目标 catalog 里创建新表	散落文件首次上传
覆盖现有表	选择现有表，覆盖其数据	更新现有表数据

存储位置

建表的目标位置设置，包含两个设置项：

- catalog 和 schema：可选择 table 类型的 catalog，如果没有合适的 catalog 或 schema，则可以点击**创建数据目录** 或者「+」创建 schema;
- 表名：如果选择覆盖现有表，则需要选择一个现有表；如果选择新建表，则需要输入表名；

字段操作

预览表的字段表头支持以下操作：

操作	入口	说明
修改字段名	点击表头字段名	字段名不可为空，不可重名；默认是解析推断后的名称
修改字段类型	点击字段名旁的类型图标	弹出类型下拉菜单，可选 STRING / INT / BIGINT / FLOAT / DOUBLE / DECIMAL / BOOLEAN / DATE / TIMESTAMP 等
设置 DECIMAL 精度	字段类型选 DECIMAL → 悬停显示编辑图标 → 点击	设置精度（1-38）与小数位（0-精度），详见 文件解析与推断 – DECIMAL 精度设置
选择建表字段	选择字段复选框	仅勾选的字段会落入目标表，其他字段不入库；至少需选择 1 个字段（仅剩一个时不允许取消）

数据预览表

特性	说明
预览行数	最多展示 50 行
数据行/列数	实际文件读取的行列数
数据预览表	根据选择的字段展示预览数据

高级设置入口

预览表上方提供 **高级设置** 入口，可针对当前文件类型调整解析规则。详见 文件解析与推断 。

> ⓘ **注意**
> 修改「高级设置」后，系统会重新解析文件并刷新预览数据。如果您之前手动改过字段名或字段类型，请在「转换规则稳定」之后再编辑字段，避免重复修改。

进入下一步

确认预览数据、字段名、字段类型与存档字段正确后，点击 **建表** 进入存储设置（选择落库的 Catalog / Schema 与处理方式），详见 本地上传到表 – 步骤 5：存储设置 。

相关文档

- 本地上传到表
- 支持的文件类型及约束
- 文件解析与推断

COS 文件上传到表

最近更新时间：2026-09-11 20:57:30

概述

腾讯云 COS 文件上传到表用于把已经存放在 COS 桶里的结构化或半结构化文件，**直接** 解析后落入 Catalog 下的目标表，省去先把文件下载到本地再上传的步骤。
整体交互与 本地上传到表 一致，主要差异在于「来源」步骤：从 COS 桶中浏览并选择文件，而不是浏览器本地文件。

适用场景

- **数据已沉淀在 COS**：例如埋点日志、对账文件、第三方 SaaS 导出文件等已经定时落入 COS，希望快速建表做一次性分析。
- **跨账号、跨地域数据接入**：将其他 COS 桶（具备访问权限）中的文件接入当前 Workspace 的湖仓。
- **Avoid 大文件本地中转**：避免把数 GB 文件先下载再上传，直接在云上完成解析与建表。

如需**周期性** 把 COS 中按日期分区的新文件持续拉取入表，请改用 文件到表同步 。

前提条件

条件	说明
数据集成权限	当前账号在所属 Workspace 中拥有数据集成模块的操作权限
COS 访问授权	当前腾讯云账号已开通本模块所需的 COS 访问权限（CAM 授权）
计算资源	已创建数据**交互式集群** 计算资源，用于解析与建表

操作步骤

步骤 1：进入 COS 文件上传到表

在控制台进入 **数据集成** 模块，确认在「接入方式」Tab，点击 **腾讯云 COS 文件上传到表** 卡片。系统在右侧打开「腾讯云 COS 文件上传到表」抽屉。

步骤 2：选择 COS 桶

参数	说明
选择 COS 桶	下拉列表展示当前账号有权限访问的 COS 桶

下拉列表的桶来自当前腾讯云账号的 COS 资源。如果列表为空，请参考 COS 授权流程 完成授权或创建桶。

步骤 3：选择文件

选定桶后，列表展示桶根目录下的文件与文件夹。

行为	说明
浏览	点击文件夹名进入下一层；空文件夹不展示子项
选择	在文件名前的复选框中勾选；支持多选；至少选择 1 个
限制	仅可选择文件，**不可选择文件夹**；文件类型与数量约束同 本地文件上传到表（最多 10 个、总大小 ≤ 2 GB）

如果列表显示「暂未拉取到 bucket 地址」或某桶显示「暂无数据」，请先到 COS 控制台确认桶与文件确实存在。

步骤 4：选择计算资源

计算资源为必选，单选，仅展示当前账号有使用权限的「交互式集群」计算资源。如选中的资源未启动，系统会在解析前自动启动。详见 本地文件上传到表 – 步骤 3：选择服务资源 。

步骤 5：数据预览

点击 **预览表** 进入数据预览。后续行为（解析推断、建表方式、字段编辑、高级设置）与本地上传到表完全一致：

- 解析与推断逻辑详见 文件解析与推断 。
- 字段编辑、表名修改详见 预览与 Schema 编辑 。

步骤 6：存储设置与建表

进入「存储设置」步骤，选择落库 Catalog / Schema 与处理方式（同名覆盖、创建新表），点击 **立即建表** 完成建表。详见 本地文件上传到表 – 步骤 5：存储设置 。

COS 授权流程

DataBuddy 通过腾讯云 CAM 体系访问 COS。如果当前账号尚未为 DataBuddy 开通 COS 访问权限，进入「选择 COS 桶」时会显示授权引导：

1. 页面展示当前腾讯云账号信息：
   - 账户名称：当前腾讯云账户的用户名。
   - 账户 ID：当前账号的 UIN。
   - 需开通：本模块需要的 COS 权限项。
1. 点击 **复制** 图标可复制以上信息。

2. 点击 **腾讯云访问管理** ，在新标签页跳转到 腾讯云 CAM 控制台 。
3. 在 CAM 中按需为账号、角色添加 COS 相关权限策略后，回到 DataBuddy 页面点击 **刷新**：
   - 授权成功 → 「选择 COS 桶」展示可访问的桶列表。
   - 仍无权限 → 仍显示引导页，请检查 CAM 策略是否生效。

常见问题

Q：选择 COS 桶后，文件列表显示「暂无数据」？

A：可能是该桶的根目录下确实为空，也可能是您选择的桶位于不同地域、不同子账号之下。请到 COS 控制台确认桶的实际情况。

Q：能否选择 COS 文件夹批量入表?

A：当前不支持选择文件夹，仅支持选择文件。如需把整个目录的文件批量入表，请使用 文件到表同步 。

Q：COS 文件总容量超过 2 GB 怎么处理?

A：与本地上传相同，单次仍受 2 GB 限制。建议拆分多次上传，或改用 文件到表同步 通过离线同步任务接入。

Q：是否支持跨地域 COS?

A：取决于当前 Workspace 与 COS 桶的网络可达性。具体可达地域请参见 DataBuddy 控制台中显示的支持地域列表。

相关文档

- 本地文件上传到表
- 本地文件上传到卷
- 文件到表同步
- 对象存储 COS
- 访问管理 CAM

本地文件上传到卷

最近更新时间：2026-09-11 20:57:30

概述

本地文件上传到卷用于把本地的**任意格式** 文件（图片、音视频、PDF、模型权重、原始日志等）以**原始字节** 上传到 Catalog 下的 Volume，保留文件本来的结构。上传完成后，文件以 Volume 资源形式在 数据目录 中可见，可被 Notebook、Python 任务、Agent 等下游消费。
与 本地上传到表 不同，本流程**不解析文件、不生成结构化表**：

维度	本地上传到表	本地上传到卷
文件处理方式	按文件类型解析为结构化字段	原始字节存储
文件类型	仅部分类型文件	任意格式
是否需要计算资源	是（用于解析）	否
单次文件数量上限	10 个	100 个
单文件容量上限	（总大小限制 2 GB）	5 GB（按单文件计算）
目标位置	Catalog → Schema → Table	Catalog → Schema → Volume

适用场景

- **机器学习样本灌注**：把带标签的图片、音频、文本语料以原文件上传到 Volume，供 Notebook 直接读取训练。
- **模型与配置文件**：上传模型权重（`.pt` / `.pkl`）、配置文件（`.yaml`）等模型物料。
- **暂存原始日志**：把待解析的原始日志先放到 Volume，后续在 Workflow 中通过 Python 任务解析入表。
- **图片、文档归档**：批量上传 PDF、合同、扫描件等非结构化资产。

操作步骤

步骤 1：进入文件上传到卷

在控制台进入 **数据集成** 模块，确认在「接入方式」Tab，点击 **文件上传到卷** 卡片。系统弹出「文件上传到卷」对话框。

步骤 2：选择文件

支持两种方式：

- **拖拽**：将文件直接拖到上传区。

- **点击选择**：点击 **选择文件** 按钮，在系统文件浏览器中单选或多选文件。

约束项	限制	校验提示
文件数量	至少 1 个，最多 100 个	「最多可上传 100 个文件」
单文件容量	≤ 5 GB	「单个文件不超过 5GB」
文件类型	任意格式	不做扩展名校验

完整错误清单：

场景	提示
超出文件数量	上传失败 — 最多可上传 100 个文件!
单文件超出容量	上传失败 — 单个文件不超过 5GB
网络断开	上传失败 — 网络连接断开
权限变更	上传失败 — 无权限执行上传服务
兜底	上传失败 — {错误详情}

文件添加到列表后，可以：

- 点击 **X** 移除单个文件。
- 再次点击 **选择文件** 追加新文件。

步骤 3：选择目标路径

参数	说明	是否必填
目标路径	选择目标 Volume。仅可选 Volume 类型目录；当前 Catalog / Schema 下没有 Volume 时，可在树形目录中即时创建	是

如需新建 Volume：

- 鼠标悬停在 Schema 上方，右侧出现 **+**，点击弹出创建面板。
- Volume 创建详见 非结构化数据 Catalog 。

> **提示**
> Volume 创建动作**脱离当前上传流程** 。即使您在创建 Volume 后取消了本次上传，已创建的 Volume 仍会保留在 Catalog 中。

参数	说明	是否必填
同名处理	**自动重命名**（默认）：生成新文件名 `文件名_1`；**替换现有文件**：用最新数据覆盖已存在的同名文件	是

仅展示当前账号有写入权限的 Catalog / Schema / Volume。可使用搜索框模糊匹配目录名。

步骤 4：上传

点击 **上传**：

- 系统跳转回数据集成首页，并在右下角弹出**上传进度列表** 。
- 列表中以背景条展示每个文件的进度百分比。

进度列表支持的操作：

操作	说明
跳转到资源	上传完成后点击文件后的「链接」图标，跳转到该文件所在 Volume 详情页；未完成时不可跳转
隐藏 / 展开列表	点击右上角关闭图标，列表隐藏为页面顶部的缩略进度标识；点击标识可重新展开列表

步骤 5：验证结果

上传完成后，您可以：

1. 在 数据目录 中导航到目标 Volume，确认文件已存在。
2. 在 Notebook 中通过文件路径读取该 Volume 文件。
3. 在 Workflow 中以 Python File 任务读取并处理 Volume 中的文件。

常见问题

Q：单次上传超过 100 个文件或超过 5 GB 怎么办？

A：分批多次上传；如有大量原始数据沉淀在对象存储或文件系统中，建议改用 文件到 Volume 同步 通过周期性离线同步任务接入。

Q：上传到卷的文件能否反向解析为结构化表？

A：可以。上传后您可以创建 SQL 任务或 Python 任务读取 Volume 文件，并写入到结构化表；也可使用 文件到表同步 把 Volume 中的文件批量入表。

Q：上传过程中关闭浏览器会怎样？

A：未完成的上传会被中断。再次进入页面时，已成功上传到 Volume 的文件不受影响；中断的文件需要重新上传。

Q：上传时网络中断后能续传吗?

A：当前不支持断点续传。出现「网络连接断开」提示时，请等待网络恢复后重新上传该文件。

Q：为什么目标路径无法选中?

A：目标 Volume 路径，必须至少要到 volume 节点（完整路径为 catalog 》 schema 》 volume 》 文件目录），可以点击文件树的左侧小三角形，来展开文件树，直到 volume 节点或文件目录节点。

相关文档

- 文件上传
- 本地上传到表
- COS 文件上传到表
- 文件到 Volume 同步（周期性接入场景）
- 非结构化数据 Catalog
- 数据目录概述