

流计算 Oceanus

CDC 开发指南



腾讯云

【 版权声明 】

©2013–2026 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

CDC 开发指南

概述

CDC 作业术语表

CDC 开发指南

概述

最近更新时间：2026-09-15 10:33:00

Flink CDC Pipeline 是基于 [Apache Flink](#) 的数据集成能力，允许用户通过一份 YAML 配置文件来声明一个数据同步作业，而无需编写任何 Java 代码。用户只需描述数据从何处来（source）、要写入何处（sink），以及（可选）如何对数据进行路由（route）与转换（transform），Flink CDC 便会在作业提交时自动将这些声明编译为一组 Flink 算子，并部署为一个 Flink 作业运行。

Pipeline 作业的组成

一个 Pipeline 作业由以下部分组成：

组成部分	是否必填	说明
source	是	定义数据源，负责访问元数据并从外部系统读取变更数据。
sink	是	定义数据目标，负责应用表结构变更并向外部系统写入变更数据。
pipeline	是	定义作业级配置，例如作业名称、并行度、时区、表结构变更行为等。
route	否	定义源表到目标表的路由规则，支持分库分表合并同步。
transform	否	定义字段投影、过滤、主键/分区键重设等数据转换规则。

其中，source、sink 和 pipeline 是一个作业必需的三个部分，route 和 transform 为可选部分。

核心特性

- 整库同步：通过库表匹配规则，在一个作业中同步一个数据库实例下的多张表。
- 表结构演进（Schema Evolution）：自动将上游 DDL 变更同步到下游系统。
- 数据转换（Transform）：支持字段投影、过滤、计算列以及内置与自定义函数。
- 精确一次（Exactly-Once）：作业失败重启后仍能保证端到端的数据不重不丢。

示例

下面是一个将 MySQL app_db 库下的所有表同步到 Doris 的 Pipeline 作业：

```
source:
  type: mysql
  hostname: localhost
```

```
port: 3306
username: root
password: 123456
tables: app_db.\.*

sink:
  type: doris
  fenodes: 127.0.0.1:8030
  username: root
  password: ""

pipeline:
  name: Sync MySQL Database to Doris
  parallelism: 2
```

CDC 作业术语表

最近更新时间：2026-09-15 10:33:00

下表列出了 Pipeline 作业开发中常用的术语及其含义，供快速查阅。

术语	英文	说明
Pipeline 作业	Data Pipeline	在 Flink CDC 中，数据从上游流转到下游遵循 Pipeline 模式，整个 ETL 作业被称为 Data Pipeline。一个 Pipeline 作业由 source、sink、pipeline 以及可选的 route、transform 组成，在 Flink 集群中对应一个作业。
数据源	Data Source	用于访问元数据并从外部系统读取变更数据的组件。一个数据源可以同时读取多个表。
数据目标	Data Sink	用于应用表结构变更并向外部系统写入变更数据的组件。一个数据目标可以同时写入多个表。
表标识	Table ID	用于建立连接器与外部系统存储对象（如表）之间映射关系的标识，由命名空间、模式名、表名三部分组成。
路由	Route	用于匹配一个或多个源表并映射到目标表的路由规则，常用于分库分表的合并同步。
数据转换	Transform	对数据进行字段投影、过滤、主键/分区键重设等操作的规则。
表结构演进	Schema Evolution	将上游的 DDL 变更（建表、加列、改列类型等）自动同步到下游的能力。
数据变更事件	DataChangeEvent	数据同步过程中承载行级数据变更的事件，如插入、更新、删除。
表结构变更事件	SchemaChangeEvent	承载表结构变更的事件，如 create.table、add.column 等。
元数据字段	Metadata Field	在 transform 规则中可引用的隐藏列，如 __namespace_name__、__schema_name__、__table_name__、__data_event_type__。
用户自定义函数	UDF	由用户自定义、可在 transform 表达式中使用的函数。
精确一次	Exactly-Once	作业失败重启后仍能保证端到端的数据不重不丢的处理语义。
整库同步	Full Database Synchronization	通过配置库表匹配规则，在一个作业中同步一个数据库实例下的多张表。

