# 数据集成

# 附录

腾讯云

# 文档目录

- TDSQL-C Mysql 离线写入
- PostgreSQL 离线写入
- SQL Server 离线写入
- Oracle 离线写入
- DB2 离线写入
- DM 离线写入
- SAP HANA 离线写入
- Hive 离线写入
- HBase 离线写入
- Clickhouse 离线写入
- DLC 离线写入
- Kudu 离线写入
- HDFS 离线写入
- Greenplum 离线写入
- GaussDB 离线写入
- Gbase 离线写入
- TBase 离线写入
- COS 离线写入
- FTP 离线写入
- SFTP 离线写入
- Elasticsearch 离线写入
- Redis 离线写入
- Iceberg 离线写入
- Doris 离线写入

转换节点

# 附录
# 任务类型转换
# 实时任务
# MySQL 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 读取

| 字段类型 | 是否支持 | 内部映射字段 | 备注 |
| --- | --- | --- | --- |
| TINYINT | 是 | TINYINT | TINYINT(1)映射到 BOOLEAN<br>需要增加选项支持 TINYINT(1)可以映射到 bool 或者 tinyint |
| SMALLINT | 是 | SMALLINT | - |
| TINYINT_UNSIGNED | 是 | SMALLINT | - |
| TINYINT_UNSIGNED_ZEROFILL | 是 | SMALLINT | - |
| INT | 是 | INT | - |
| INTEGER | 是 | INT | - |
| YEAR | 是 | INT | - |
| MEDIUMINT | 是 | INT | - |
| SMALLINT_UNSIGNED | 是 | INT | - |
| SMALLINT_UNSIGNED_ZEROFILL | 是 | INT | - |
| BIGINT | 是 | LONG | - |
| INT_UNSIGNED | 是 | LONG | - |
| MEDIUMINT_UNSIGNED | 是 | LONG | - |
| MEDIUMINT_UNSIGNED_ZEROFILL | 是 | LONG | - |
| INT_UNSIGNED_ZEROFILL | 是 | LONG | - |
| BIGINT_UNSIGNED | 是 | DECIMAL | DECIMAL(20,0) |
| BIGINT_UNSIGNED_ZEROFILL | 是 | DECIMAL | DECIMAL(20,0) |
| SERIAL | 是 | DECIMAL | DECIMAL(20,0) |
| FLOAT | 是 | FLOAT | - |
| FLOAT_UNSIGNED | 是 | FLOAT | - |
| FLOAT_UNSIGNED_ZEROFILL | 是 | FLOAT | - |
| DOUBLE | 是 | DOUBLE | - |
| DOUBLE_UNSIGNED | 是 | DOUBLE | - |
| DOUBLE_UNSIGNED_ZEROFILL | 是 | DOUBLE | - |
| DOUBLE_PRECISION | 是 | DOUBLE | - |
| DOUBLE_PRECISION_UNSIGNED | 是 | DOUBLE | - |

| | | | |
|---|---|---|---|
| ZEROFILL | 是 | DOUBLE | - |
| REAL | 是 | DOUBLE | - |
| REAL_UNSIGNED | 是 | DOUBLE | - |
| REAL_UNSIGNED_ZEROFILL | 是 | DOUBLE | - |
| NUMERIC | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| NUMERIC_UNSIGNED | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| NUMERIC_UNSIGNED_ZEROFILL | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| DECIMAL | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| DECIMAL_UNSIGNED | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| DECIMAL_UNSIGNED_ZEROFILL | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| FIXED | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| FIXED_UNSIGNED | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| FIXED_UNSIGNED_ZEROFILL | 是 | DECIMAL | 采用用户数据库实际的精度 p<=38 映射到 DECIMAL 38 < p <= 65 时映射到 String |
| BOOLEAN | 是 | BOOLEAN | - |
| DATE | 是 | DATE | - |
| TIME | 是 | TIME | - |
| DATETIME | 是 | TIMESTAMP | - |
| TIMESTAMP | 是 | TIMESTAMP | - |
| CHAR | 是 | STRING | - |
| JSON | 是 | STRING | - |
| BIT | 是 | STRING | BIT(1) 映射到 BOOLEAN |
| VARCHAR | 是 | STRING | - |
| TEXT | 是 | STRING | - |
| BLOB | 是 | STRING | - |
| TINYBLOB | 是 | STRING | - |
| TINYTEXT | 是 | STRING | - |
| MEDIUMBLOB | 是 | STRING | - |
| MEDIUMTEXT | 是 | STRING | - |
| LONGBLOB | 是 | STRING | - |

| | | | |
|---|---|---|---|
| LONGTEXT | 是 | STRING | - |
| VARBINARY | 是 | STRING | - |
| GEOMETRY | 是 | STRING | - |
| POINT | 是 | STRING | - |
| LINESTRING | 是 | STRING | - |
| POLYGON | 是 | STRING | - |
| MULTIPOINT | 是 | STRING | - |
| MULTILINESTRING | 是 | STRING | - |
| MULTIPOLYGON | 是 | STRING | - |
| GEOMETRYCOLLECTION | 是 | STRING | - |
| ENUM | 是 | STRING | - |
| BINARY | 是 | BINARY | BINARY(1) |
| SET | 否 | | - |

## 写入

| 内部类型 | MySQL 类型 |
|---|---|
| TINYINT | TINYINT |
| SMALLINT | SMALLINT，TINYINT UNSIGNED |
| INT | INT，MEDIUMINT，SMALLINT UNSIGNED |
| BIGINT | BIGINT，INT UNSIGNED |
| DECIMAL(20, 0) | BIGINT UNSIGNED |
| FLOAT | FLOAT |
| DOUBLE | DOUBLE，DOUBLE PRECISION |
| DECIMAL(p, s) | NUMERIC(p, s)，DECIMAL(p, s) |
| BOOLEAN | BOOLEAN，TINYINT(1) |
| DATE | DATE |
| TIME [(p)][WITHOUT TIMEZONE] | TIME [(p)] |
| TIMESTAMP [(p)][WITHOUT TIMEZONE] | DATETIME [(p)] |
| STRING | CHAR(n)，VARCHAR(n)，TEXT |
| BYTES | BINARY，VARBINARY，BLOB |
| ARRAY | - |

# Mongo 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 读取

| BSON 类型 | 内部类型 |
| --- | --- |
| – | TINYINT |
| – | SMALLINT |
| Int | INT |
| Long | BIGINT |
| – | FLOAT |
| Double | DOUBLE |
| Decimal128 | DECIMAL(p, s) |
| Boolean | BOOLEAN |
| Date Timestamp | DATE |
| Date Timestamp | TIME |
| Date | TIMESTAMP(3) TIMESTAMP_LTZ(3) |
| Timestamp | TIMESTAMP(0) TIMESTAMP_LTZ(0) |
| String，ObjectId，UUID，Symbol，MD5，JavaScript，Regex | STRING |
| BinData | BYTES |
| Object | ROW |
| Array | ARRAY |
| DBPointer | ROW&ltSTRING, STRING> |

# Oracle 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 读取

| Oracle 类型 | 内部类型 |
| --- | --- |
| NUMBER(p, s <= 0), p - s < 3 | TINYINT |
| NUMBER(p, s <= 0), p - s < 5 | SMALLINT |
| NUMBER(p, s <= 0), p - s < 10 | INT |
| NUMBER(p, s <= 0), p - s < 19 | BIGINT |
| NUMBER(p, s <= 0), 19 <= p - s <= 38 | DECIMAL(p - s, 0) |
| NUMBER(p, s > 0) | DECIMAL(p, s) |
| NUMBER(p, s <= 0), p - s > 38 | STRING |
| FLOAT，BINARY_FLOAT | FLOAT |
| DOUBLE PRECISION，BINARY_DOUBLE | DOUBLE |
| NUMBER(1) | BOOLEAN |
| DATE，TIMESTAMP [(p)] | TIMESTAMP [(p)] [WITHOUT TIMEZONE] |
| TIMESTAMP [(p)] WITH TIME ZONE | TIMESTAMP [(p)] WITH TIME ZONE |
| TIMESTAMP [(p)] WITH LOCAL TIME ZONE | TIMESTAMP_LTZ [(p)] |
| CHAR(n)，NCHAR(n)，NVARCHAR2(n)，VARCHAR(n)，VARCHAR2(n)，CLOB，NCLOB，XML 类型 | STRING |
| BLOB，ROWID | BYTES |
| INTERVAL DAY TO SECOND，INTERVAL YEAR TO MONTH | BIGINT |

## 写入

| 内部类型 | Oracle 类型 |
| --- | --- |
| FLOAT | BINARY_FLOAT |
| DOUBLE | BINARY_DOUBLE |
| DECIMAL(p, s) | SMALLINT，FLOAT(s)，DOUBLE PRECISION，REAL，NUMBER(p, s) |
| DATE | DATE |
| DECIMAL(20, 0) | - |
| FLOAT | REAL，FLOAT4 |
| DOUBLE | FLOAT8，DOUBLE PRECISION |
| DECIMAL(p, s) | NUMERIC(p, s)，DECIMAL(p, s) |
| BOOLEAN | BOOLEAN |
| DATE | DATE |

| | |
|---|---|
| TIMESTAMP [(p)][WITHOUT TIMEZONE] | TIMESTAMP [(p)]WITHOUT TIMEZONE |
| STRING | CHAR(n)，VARCHAR(n)，CLOB(n) |
| BYTES | RAW(s)，BLOB |
| ARRAY | – |

# PostgreSQL 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 读取

| PostgreSQL 类型 | 内部类型 |
| --- | --- |
| - | TINYINT |
| SMALLINT，INT2，SMALLSERIAL，SERIAL2 | SMALLINT |
| INTEGER，SERIAL | INT |
| BIGINT，BGSERIAL | BIGINT |
| - | DECIMAL(20, 0) |
| REAL，FLOAT4 | FLOAT |
| FLOAT8，DOUBLE PRECISION | DOUBLE |
| NUMERIC(p, s)，DECIMAL(p, s) | DECIMAL(p, s) |
| BOOLEAN | BOOLEAN |
| DATE | DATE |
| TIME [(p)][WITHOUT TIMEZONE] | TIME [(p)][WITHOUT TIMEZONE] |
| TIMESTAMP [(p)]WITHOUT TIMEZONE | TIMESTAMP [(p)][WITHOUT TIMEZONE] |
| CHAR(n)，CHARACTER(n)，VARCHAR(n)，CHARACTER ，VARYING(n)，TEXT | STRING |
| BYTEA | BYTES |

## 写入

| 内部类型 | PostgreSQL 类型 |
| --- | --- |
| TINYINT | - |
| SMALLINT | SMALLINT，INT2，SMALLSERIAL，SERIAL2 |
| INT | INTEGER，SERIAL |
| BIGINT | BIGINT，BIGSERIAL |
| DECIMAL(20, 0) | - |
| FLOAT | REAL，FLOAT4 |
| DOUBLE | FLOAT8，DOUBLE PRECISION |
| DECIMAL(p, s) | NUMERIC(p, s)，DECIMAL(p, s) |
| BOOLEAN | BOOLEAN |
| DATE | DATE |
| TIME [(p)][WITHOUT TIMEZONE] | TIME [(p)][WITHOUT TIMEZONE] |
| TIMESTAMP [(p)][WITHOUT TIMEZONE] | TIMESTAMP [(p)]WITHOUT TIMEZONE |
| STRING | CHAR(n)，CHARACTER(n)，VARCHAR(n)，CHARACTER VARYING(n)，TEXT |

| | |
|---|---|
| BYTES | BYTEA |

# SQLserver 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 读取

| SQLServer 类型 | 内部类型 |
| --- | --- |
| char(n) | CHAR(n) |
| varchar(n)，nvarchar(n)，nchar(n) | VARCHAR(n) |
| text，ntext，xml | STRING |
| decimal(p, s)，money，smallmoney | DECIMAL(p, s) |
| numeric | NUMERIC |
| REAL，FLOAT | FLOAT |
| bit | BOOLEAN |
| int | INT |
| tinyint | TINYINT |
| smallint | SMALLINT |
| time (n) | TIME (n) |
| bigint | BIGINT |
| date | DATE |
| datetime2，datetime，smalldatetime | TIMESTAMP(n) |
| datetimeoffset | TIMESTAMP_LTZ(3) |

## 写入

| 内部类型 | SQLServer 类型 |
| --- | --- |
| CHAR(n) | char(n) |
| VARCHAR(n) | varchar(n)，nvarchar(n)，nchar(n) |
| STRING | text，ntext，xml |
| BIGINT | BIGINT，BIGSERIAL |
| DECIMAL(p, s) | decimal(p, s)，money，smallmoney |
| NUMERIC | numeric |
| FLOAT | float，real |
| BOOLEAN | bit |
| INT | int |
| TINYINT | tinyint |
| SMALLINT | smallint |
| BIGINT | bigint |
| TIME(n) | time(n) |
| TIMESTAMP(n) | datetime2，datetime，smalldatetime |

| | |
|---|---|
| TIMESTAMP_LTZ(3) | datetimeoffset |

# ClickHouse 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | ClickHouse 类型 |
| --- | --- |
| CHAR | String |
| VARCHAR | String，IP，UUID |
| STRING | String，EnumL |
| BOOLEAN | UInt8 |
| BYTES | FixedString |
| DECIMAL | Decimal，Int128，Int256，UInt64，UInt128，UInt256 |
| TINYINT | Int8 |
| SMALLINT | Int16，UInt8 |
| INTEGER | Int32，UInt16，Interval |
| BIGINT | Int64，UInt32 |
| FLOAT | Float32 |
| DATE | Date |
| TIME | DateTime |
| TIMESTAMP | DateTime |
| TIMESTAMP_LTZ | DateTime |
| INTERVAL_YEAR_MONTH | Int32 |
| INTERVAL_DAY_TIME | Int64 |
| ARRAY | Array |
| MAP | Map |

# ElasticSearch 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | JSON 类型 |
| --- | --- |
| CHAR / VARCHAR / STRING | string |
| BOOLEAN | boolean |
| BINARY / VARBINARY | string with encoding: base64 |
| DECIMAL | number |
| TINYINT | number |
| SMALLINT | number |
| INT | number |
| BIGINT | number |
| FLOAT | number |
| DOUBLE | number |
| DATE | string with format: date |
| TIME | string with format: time |
| TIMESTAMP | string with format: date-time |
| TIMESTAMP_WITH_LOCAL_TIME_ZONE | string with format: date-time (with UTC time zone) |
| INTERVAL | number |
| ARRAY | array |
| MAP / MULTISET | object |
| ROW | object |

# Greenplum 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | Greenplum 类型 |
| --- | --- |
| TINYINT | - |
| SMALLINT | SMALLINT，INT2，SMALLSERIAL，SERIAL2 |
| INT | INTEGER，SERIAL |
| BIGINT | BIGINT，BIGSERIAL |
| DECIMAL(20, 0) | - |
| FLOAT | REAL，FLOAT4 |
| DOUBLE | FLOAT8，DOUBLE PRECISION |
| DECIMAL(p, s) | NUMERIC(p, s)，DECIMAL(p, s) |
| BOOLEAN | BOOLEAN |
| DATE | DATE |
| TIME [(p)][WITHOUT TIMEZONE] | TIME [(p)][WITHOUT TIMEZONE] |
| TIMESTAMP [(p)][WITHOUT TIMEZONE] | TIMESTAMP [(p)]WITHOUT TIMEZONE |
| STRING | CHAR(n)，CHARACTER(n)，VARCHAR(n)，CHARACTER VARYING(n)，TEXT |
| BYTES | BYTEA |

# HBase 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | HBase 类型 |
| --- | --- |
| CHAR，VARCHAR，STRING | byte[] toBytes(String s)，String toString(byte[] b) |
| BOOLEAN | byte[] toBytes(boolean b)，boolean toBoolean(byte[] b) |
| BINARY VARBINARY | Returns byte[] as is. |
| DECIMAL | byte[] toBytes(BigDecimal v)，BigDecimal toBigDecimal(byte[] b) |
| TINYINT | new byte[] { val }，bytes[0] // returns first and only byte from bytes |
| SMALLINT | byte[] toBytes(short val)，short toShort(byte[] bytes) |
| INT | byte[] toBytes(int val)，int toInt(byte[] bytes) |
| BIGINT | byte[] toBytes(long val)，long toLong(byte[] bytes) |
| FLOAT | byte[] toBytes(float val)，float toFloat(byte[] bytes) |
| DOUBLE | byte[] toBytes(double val)，double toDouble(byte[] bytes) |
| DATE | Stores the number of days since epoch as int value. |
| TIME | Stores the number of milliseconds of the day as int value. |
| TIMESTAMP | Stores the milliseconds since epoch as long value. |

# Hive 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | Hive 类型 |
| --- | --- |
| CHAR(p) | char(p) |
| VARCHAR(p) | varchar(p) |
| STRING | string |
| BOOLEAN | boolean |
| TINYINT | tinyint |
| SMALLINT | smallint |
| INT | int |
| BIGINT | bigint |
| FLOAT | float |
| DOUBLE | double |
| DECIMAL(p, s) | decimal(p, s) |
| DATE | date |
| TIMESTAMP | timestamp(9) |
| BINARY | bytes |
| LIST | array |
| MAP | map |
| STRUCT | row |

# DLC - iceberg/Iceberg 数据类型转换（实时）

最近更新时间：2024-09-06 16:40:21

## 写入

| 内部类型 | Iceberg 类型 |
| --- | --- |
| CHAR | STRING |
| VARCHAR | STRING |
| STRING | STRING |
| BOOLEAN | BOOLEAN |
| BINARY | FIXED(L) |
| VARBINARY | BINARY |
| DECIMAL | DECIMAL(P,S) |
| TINYINT | INT |
| SMALLINT | INT |
| INTEGER | INT |
| BIGINT | LONG |
| FLOAT | FLOAT |
| DOUBLE | DOUBLE |
| DATE | DATE |
| TIME | TIME |
| TIMESTAMP | TIMESTAMP |
| TIMESTAMP_LTZ | TIMESTAMPTZ |
| INTERVAL | - |
| ARRAY | LIST |
| MULTISET | MAP |
| MAP | MAP |
| ROW | STRUCT |
| RAW | - |

# 离线任务
# MySQL/TDSQL-C MySQL 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| Mysql 数据类型 | 内部类型 |
| --- | --- |
| int，tinyint，smallint，mediumint，int，bigint | Long |
| float，double，decimal | Double |
| varchar，char，tinytext，text，mediumtext，longtext，year | String |
| date，datetime，timestamp，time | Date |
| bit，bool | Boolean |
| tinyblob，mediumblob，blob，longblob，varbinary | Bytes |

## 写入

| 内部类型 | Mysql 数据类型 |
| --- | --- |
| Long | int，tinyint，smallint，mediumint，int，bigint，year |
| Double | float，double，decimal |
| String | varchar，char，tinytext，text，mediumtext，longtext |
| Date | date，datetime，timestamp，time |
| Boolean | bit，bool |
| Bytes | tinyblob，mediumblob，blob，longblob，varbinary |

# PostgreSQL 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| PostgreSQL 数据类型 | 内部类型 |
| --- | --- |
| bigint，bigserial，integer，smallint，serial | Long |
| double precision，money，numeric，real | Double |
| varchar，char，text，bit，inet | String |
| date，time，timestamp | Date |
| bool | Boolean |
| bytea | Bytes |

## 写入

| 内部类型 | PostgreSQL 数据类型 |
| --- | --- |
| Long | bigint，bigserial，integer，smallint，serial |
| Double | double precision，money，numeric，real |
| String | varchar，char，text，bit |
| Date | date，time，timestamp |
| Boolean | bool |
| Bytes | bytes |

# SQLServer 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| SqlServer 数据类型 | 内部类型 |
| --- | --- |
| bigint，int，smallint，tinyint | Long |
| float，decimal，real，numeric | Double |
| char，nchar，ntext，nvarchar，text，varchar，nvarchar(MAX)，varchar(MAX) | String |
| date，datetime，time | Date |
| bit | Boolean |
| binary，varbinary，varbinary(MAX)，timestamp | Bytes |

## 写入

| 内部类型 | SqlServer 数据类型 |
| --- | --- |
| Long | bigint，int，smallint，tinyint |
| Double | float，decimal，real，numeric |
| String | char，nchar，ntext，nvarchar，text，varchar，nvarchar(MAX)，varchar(MAX) |
| Date | date，datetime，time |
| Boolean | bit |
| Bytes | binary，varbinary，varbinary(MAX)，timestamp |

# Oracle 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| Oracle 数据类型 | 内部类型 |
| --- | --- |
| NUMBER，INTEGER，INT，SMALLINT | Long |
| NUMERIC，DECIMAL，FLOAT，DOUBLE PRECISION，REAL | Double |
| LONG，CHAR，NCHAR，VARCHAR，VARCHAR2，NVARCHAR2，CLOB，NCLOB，CHARACTER，CHARACTER VARYING，CHAR VARYING，NATIONAL CHARACTER，NATIONAL CHAR，NATIONAL CHARACTER VARYING，NATIONAL CHAR VARYING，NCHAR VARYING | String |
| TIMESTAMP，DATE | Date |
| bit，bool | Boolean |
| BLOB，BFILE，RAW，LONG RAW | Bytes |

## 写入

| 内部类型 | Oracle 数据类型 |
| --- | --- |
| Long | NUMBER，INTEGER，INT，SMALLINT |
| Double | NUMERIC，DECIMAL，FLOAT，DOUBLE PRECISION，REAL |
| String | LONG，CHAR，NCHAR，VARCHAR，VARCHAR2，NVARCHAR2，CLOB，NCLOB，CHARACTER，CHARACTER VARYING，CHAR VARYING，NATIONAL CHARACTER，NATIONAL CHAR，NATIONAL CHARACTER VARYING，NATIONAL CHAR VARYING，NCHAR VARYING |
| Date | TIMESTAMP，DATE |
| Boolean | bit，bool |
| Bytes | BLOB，BFILE，RAW，LONG RAW |

# Hive 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| Hive 数据类型 | 内部类型 |
| --- | --- |
| TINYINT，SMALLINT，INT，BIGINT | Long |
| FLOAT，DOUBLE | Double |
| String，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY | String |
| BOOLEAN | Boolean |
| Date，TIMESTAMP | Date |

## 写入

| 内部类型 | Hive 数据类型 |
| --- | --- |
| Long | TINYINT，SMALLINT，INT，BIGINT |
| Double | FLOAT，DOUBLE |
| String | String，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY |
| Boolean | BOOLEAN |
| Date | Date，TIMESTAMP |

# HBase 数据类型转换（离线）

最近更新时间：2024-12-10 11:41:52

## 读取

| HBase 数据类型 | 内部类型 |
| --- | --- |
| int，short，long | Long |
| float，double | Double |
| string，binary string | String |
| date | Date |
| boolean | Boolean |

## 写入

| 内部类型 | HBase 数据类型 |
| --- | --- |
| Long | int，short，long |
| Double | float，double |
| String | string，binary string |
| Date | date |
| Boolean | boolean |

# HDFS 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| HDFS（Hive 表）数据类型 | 内部类型 |
| --- | --- |
| TINYINT，SMALLINT，INT，BIGINT | Long |
| FLOAT，DOUBLE | Double |
| String，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY | String |
| BOOLEAN | Boolean |
| Date，TIMESTAMP | Date |

## 写入

| 内部类型 | HDFS（Hive 表 ）数据类型 |
| --- | --- |
| Long | TINYINT，SMALLINT，INT，BIGINT |
| Double | FLOAT，DOUBLE |
| String | String，CHAR，VARCHAR，STRUCT，MAP，ARRAY，UNION，BINARY |
| Boolean | BOOLEAN |
| Date | Date，TIMESTAMP |

# Mongo 数据类型转换（离线）

最近更新时间：2024-07-09 22:01:41

## 读取

| Mongo 数据类型 | 内部类型 |
| --- | --- |
| int，Long | Long |
| double | Double |
| string，array | String |
| date | Date |
| boolean | Boolean |
| bytes | Bytes |

## 写入

| 内部类型 | Mongo 数据类型 |
| --- | --- |
| Long | int，Long |
| Double | double |
| String | string，array |
| Date | date |
| Boolean | boolean |
| Bytes | bytes |

# 实时单表
# 读取节点
# Kafka 单表读取

最近更新时间：2024-07-09 22:01:41

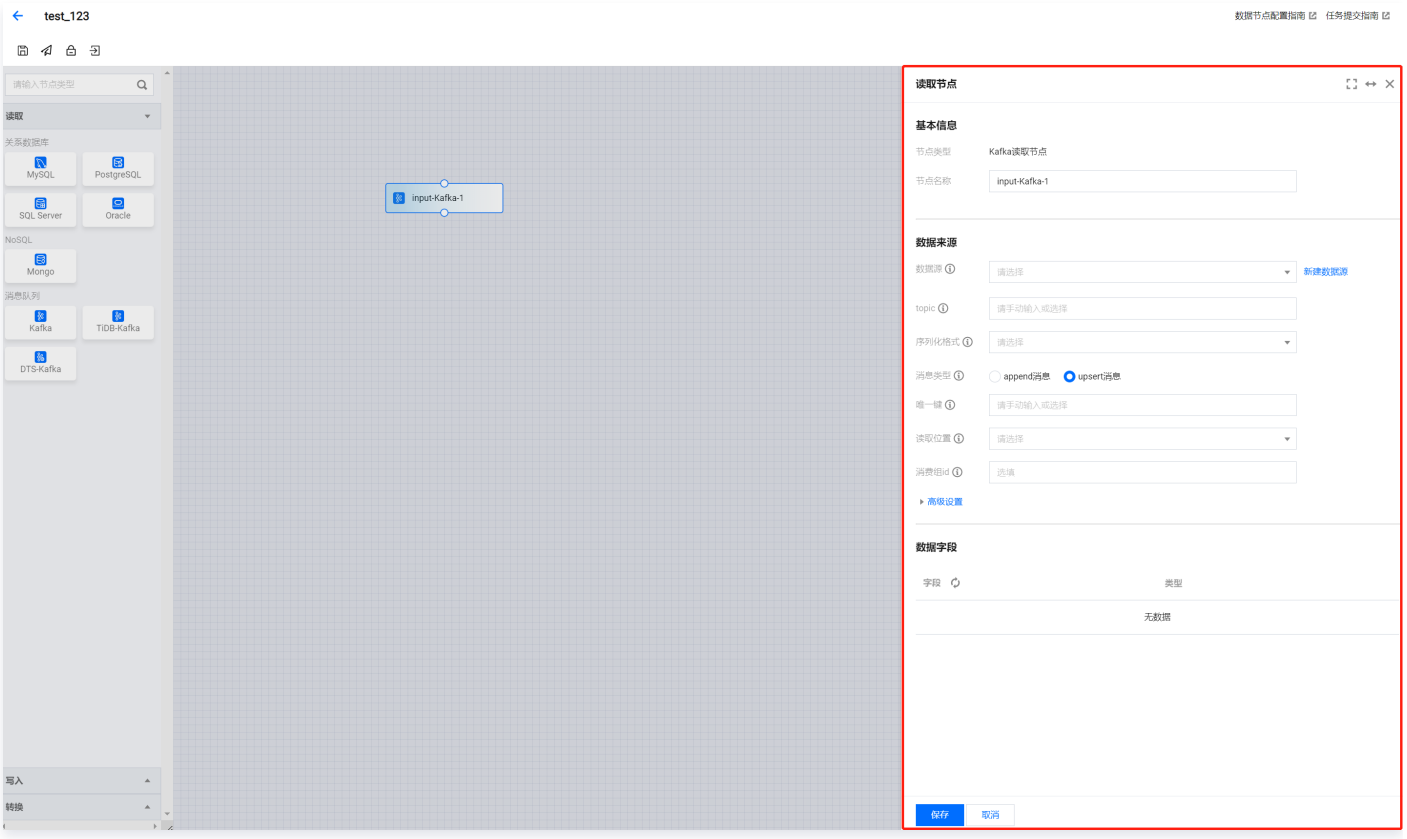
## 条件及限制

- Kafka 支持版本：

| 节点 | 版本 |
| --- | --- |
| Kafka | 0.10+ |

## 创建 Kafka 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 Kafka 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 Kafka 节点名称 |
| 数据源 | Kafka 读取端数据源类型支持 Kafka、CKafka |
| Topic | Kafka 数据源中的 Topic |
| 序列化格式 | Kafka 消息序列化格式类型，支持：canal-json、json、avro、csv |
| 消息类型 | • append 消息：Kafka 内消息来源于 append 消息流，通常消息中不携带唯一键。写入节点建议搭配 append 写入模式 |

| | |
|---|---|
| | • upsert 消息：Kafka 内消息来源于 upsert 消息流，通常消息中携带唯一键，设置后消息可保证 Exactly-Once。写入节点建议搭配 upsert 写入模式 |
| 读取位置 | 启动同步任务时开始同步数据的起始位点 |
| 消费组 ID | 请避免该参数与其他消费进程重复，以保证消费位点的正确性。如果不指定该参数，默认设定 group.id=WeData_ group_ $ {任务 id} |

5. 预览字段，单击**保存**。

# MySQL 单表读取

最近更新时间：2024-04-07 11:31:04

## 条件及限制

- MySQL 支持版本：

| 节点 | 版本 | Driver |
| --- | --- | --- |
| MySQL-CDC | MySQL：5.6，5.7，8.0.x<br>RDS MySQL：5.6，5.7，8.0.x<br>PolarDB MySQL：5.6，5.7，8.0.x<br>Aurora MySQL：5.6，5.7，8.0.x<br>MariaDB：10.x<br>PolarDB X：2.0.1 | JDBC Driver：8.0.21 |

- 每个 MySQL 数据库客户端需设置一个不同的 SERVER ID。
  每一个读取 Binlog 的 MySQL 数据库客户端都应该有一个唯一的 ID，称为 SERVER ID。MySQL 服务器将使用此 ID 来维护网络连接和 Binlog 位置。因此，如果不同的作业共享相同的服务器 ID，可能会导致从错误的 Binlog 位置读取。因此，建议通过 SQL Hints ，例如假设源并行度为4，那么我们可以使用 `SELECT * FROM source_table /*+ OPTIONS('server-id'='5401-5404') */` ；为4个 Source Reader 中的每一个分配唯一的服务器 ID。
- 设置 MySQL 会话超时：
  当为大型数据库创建初始一致快照时，您建立的连接可能会在读取表时超时。您可以通过在 MySQL 配置文件中配置 interactive_timeout 和 wait_timeout 来防止这种行为。
  - interactive_timeout：服务器在关闭交互式连接之前等待其活动的秒数。请参阅 MySQL :: MySQL 8.0 Reference Manual :: 5.1.8 Server System Variables 。
  - wait_timeout：服务器在关闭非交互式连接之前等待其活动的秒数。请参阅 MySQL :: MySQL 8.0 Reference Manual :: 5.1.8 Server System Variables 。

## 设置 MySQL 服务器权限

您必须定义一个对 Debezium MySQL 连接器监控的所有数据库具有适当权限的 MySQL 用户。

1. 创建 MySQL 用户：

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

2. 向用户授予所需的权限：

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

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

3. 刷新用户的权限：

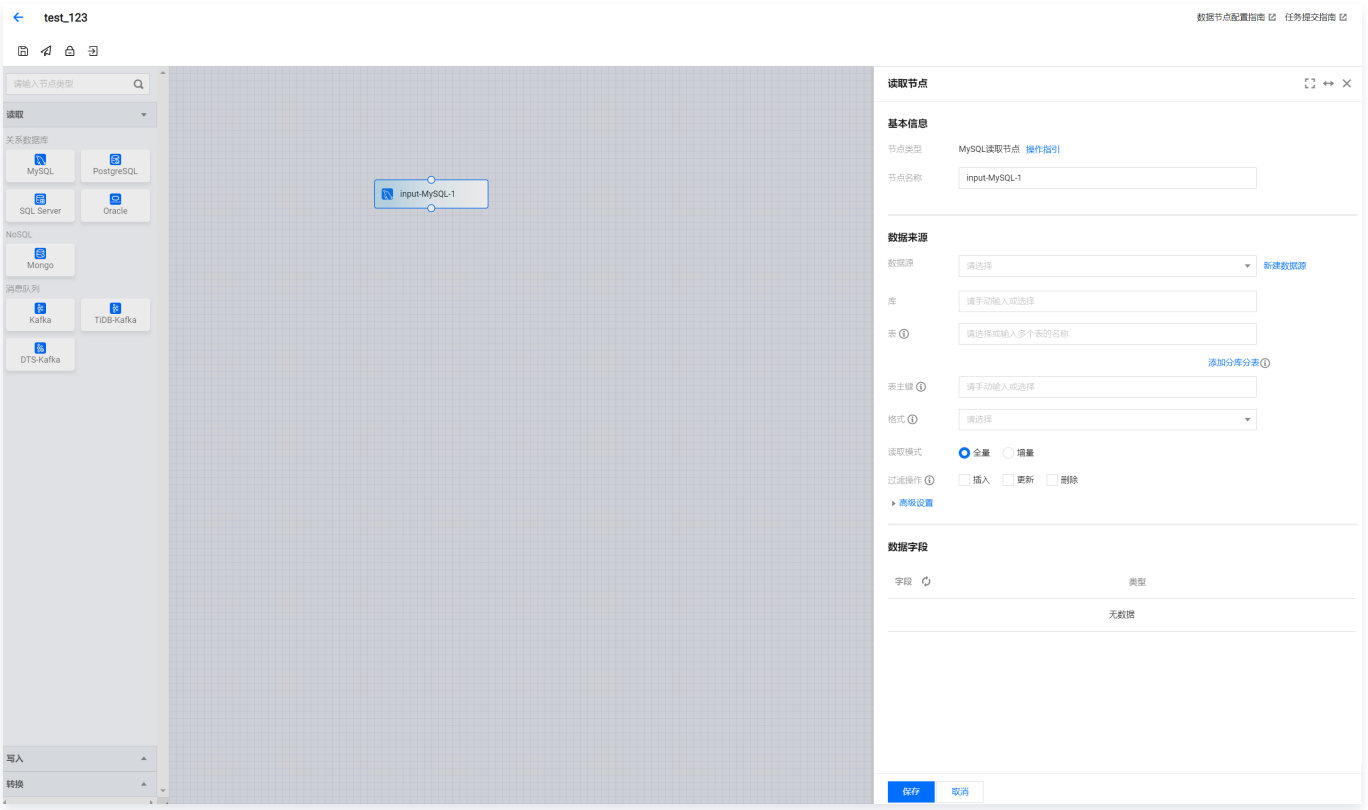
```
mysql> FLUSH PRIVILEGES;
```

查看更多关于 权限说明 。

## MySQL 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。

3. 单击左侧**读取**，单击选择 MySQL 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 MySQL 节点名称。 |
| 数据源 | 可用的 MySQL 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。<br>• 分表情况下，可在 MySQL 源端支持选择或输入多个表名称，多个表需保证结构一致。<br>• 分表情况下，支持配置表序号区间。例如'table_[0-99]'表示读取'table_0'、'table_1'、'table_2'直到'table_99'；如果您的表数字后缀的长度一致，例如'table_000'、'table_001'、'table_002'直到'table_999'，您可以配置为'"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]'。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 添加分库分表 | 适用于分库场景，点击后可配置多个数据源、库及表信息。分库分表场景下需保证所有表结构一致，任务配置将默认展示并使用第一个表结构进行数据获取。 |
| 表主键 | 分库分表模式下默认表 schema 一致。系统将使用拉去第一张表的主键，请选择或输入表主键字段名称。 |
| 格式 | 指定 MySQL 日志编码格式（utf-8、gbk、Latin1、utf8mb4）。 |
| 读取模式 | 支持全量和增量两种模式。 |
| 过滤操作 | 设置后将不同步指定操作类型的数据，支持插入、更新和删除。 |

5. 预览数据字段，单击**保存**。

## 附录

- MySQL 实时任务数据类型转换。

# MongoDB 单表读取

最近更新时间：2024-07-09 22:01:41

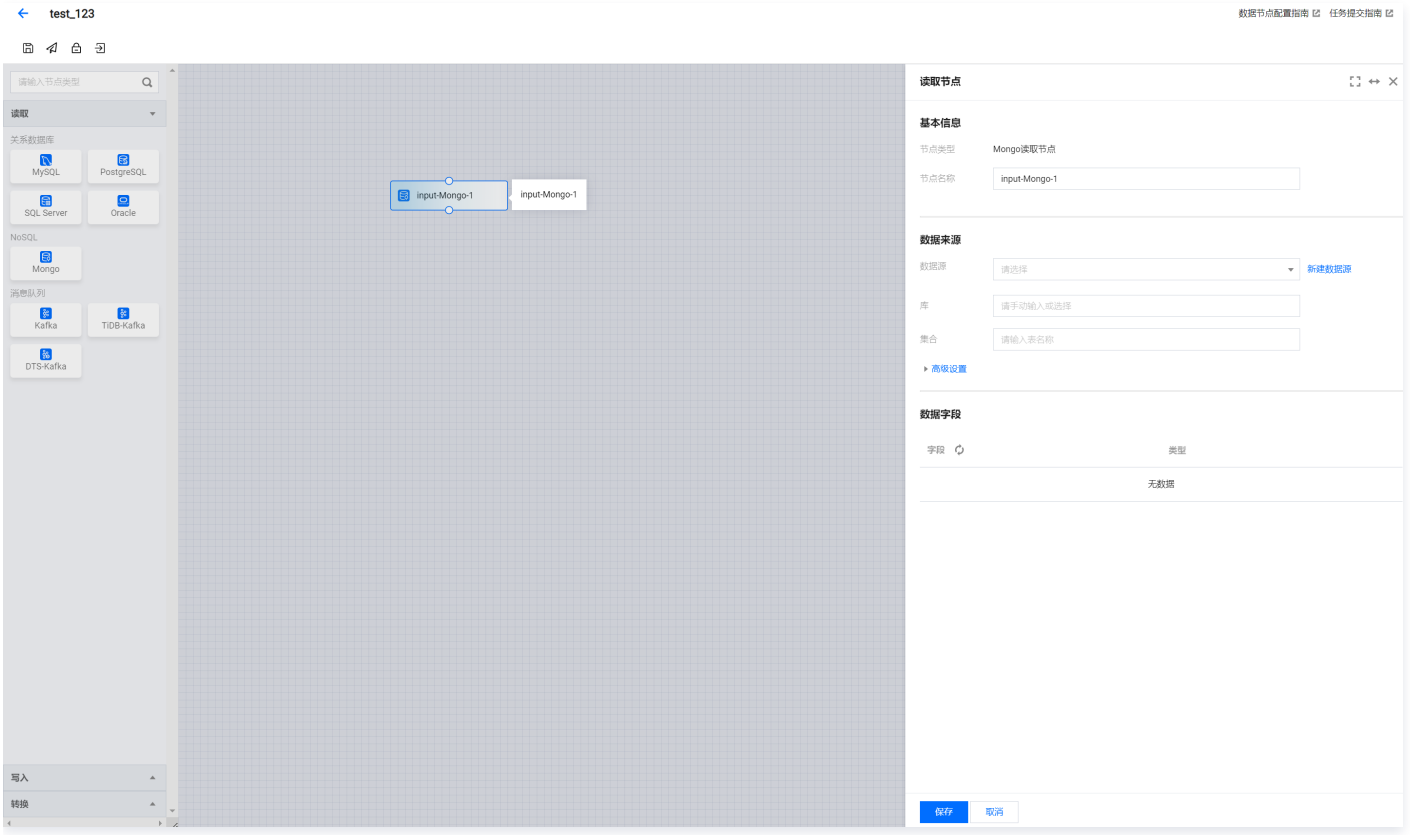
## 条件及限制

支持 Mongo 版本详情：

| 节点 | 版本 |
| --- | --- |
| MongoDB-CDC | MongoDB>=3.6 |

# MongoDB 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 Mongo 节点并配置节点信息。




4. 预览数据字段，单击**保存**。

## 附录

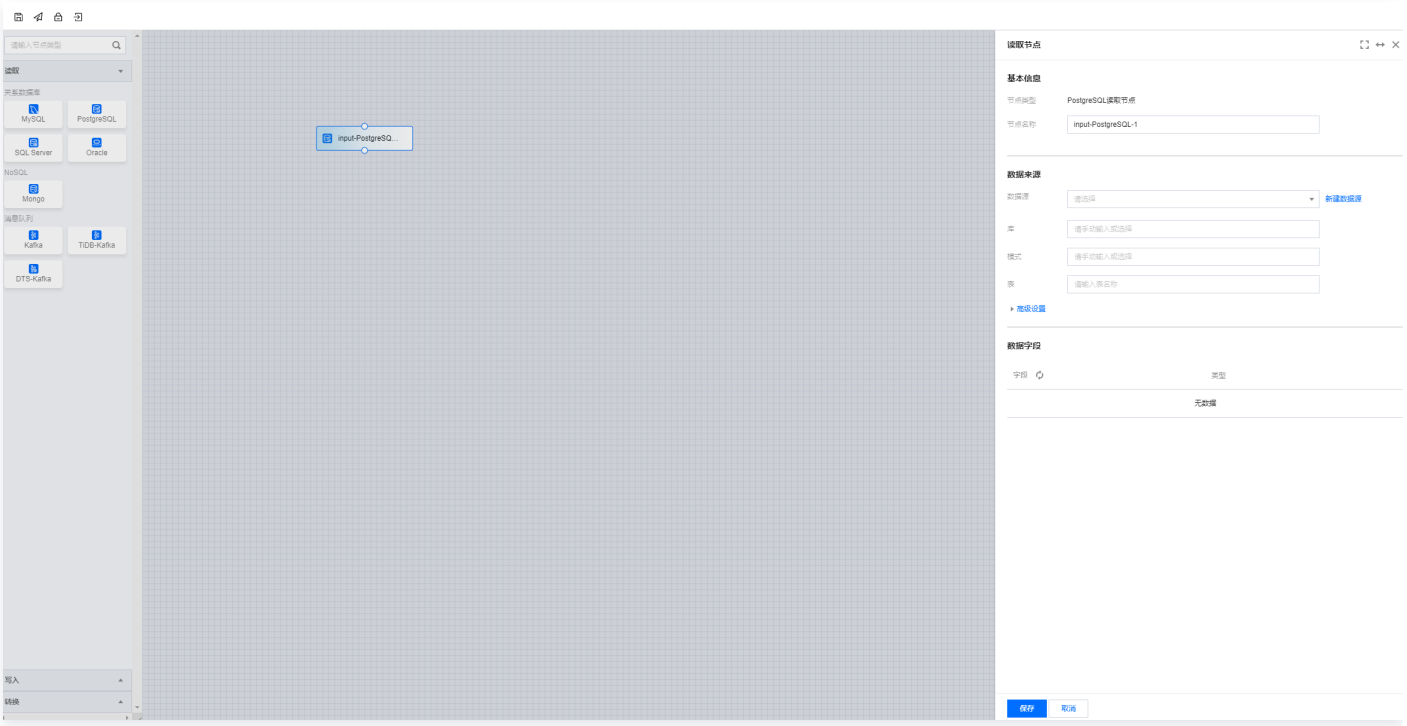
- Mongo 实时任务数据类型转换

# PostgreSQL 单表读取

最近更新时间：2024-09-06 16:40:21

## PostgreSQL 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 PostgreSQL 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 PostgreSQL 节点名称。 |
| 数据源 | 选择需要同步的表所在数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段，单击**保存**。

## 附录

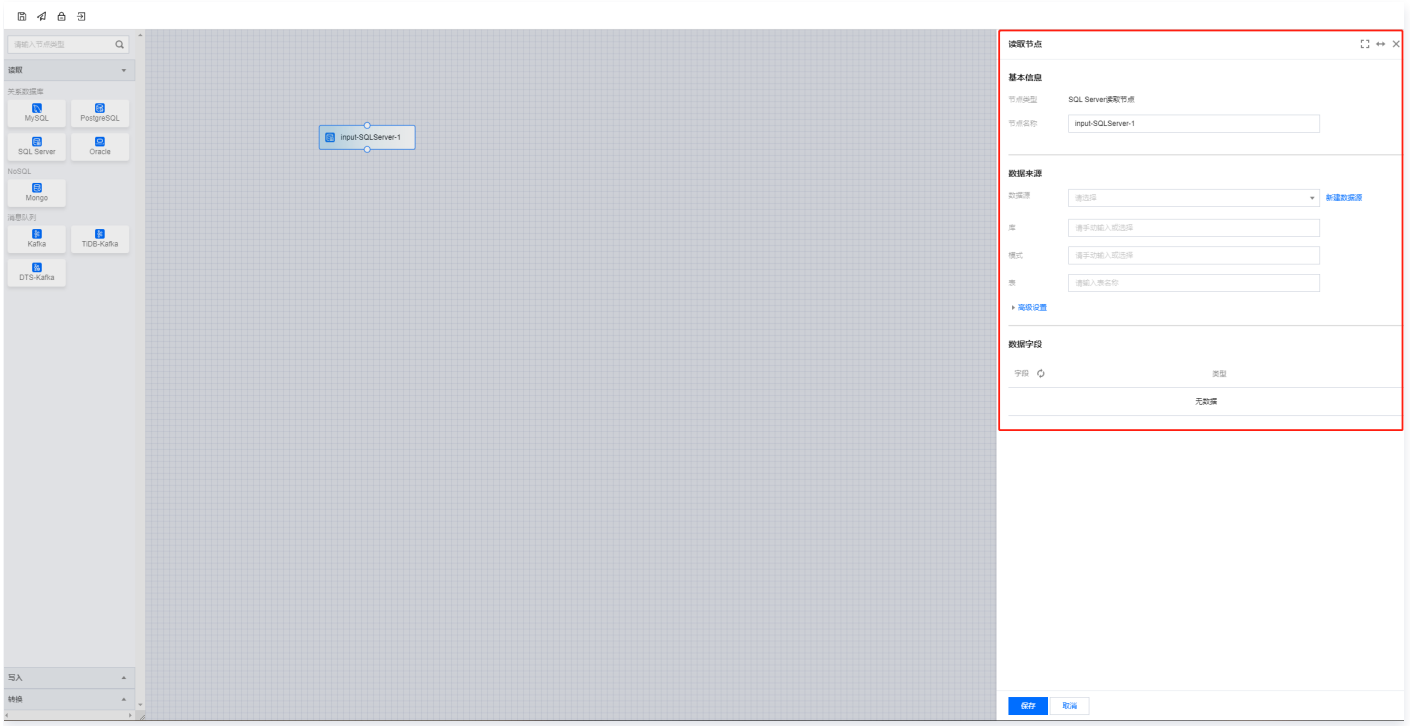
PostgreSQL 实时任务数据类型转换

# SQL Server 单表读取

最近更新时间：2024-08-08 11:53:41

## SQL Server 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 SQL Server 节点并配置节点信息。




4. 参数信息

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 SQL Server 节点名称。 |
| 数据源 | 选择需要同步的表所在数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段，单击**保存**。

## 附录

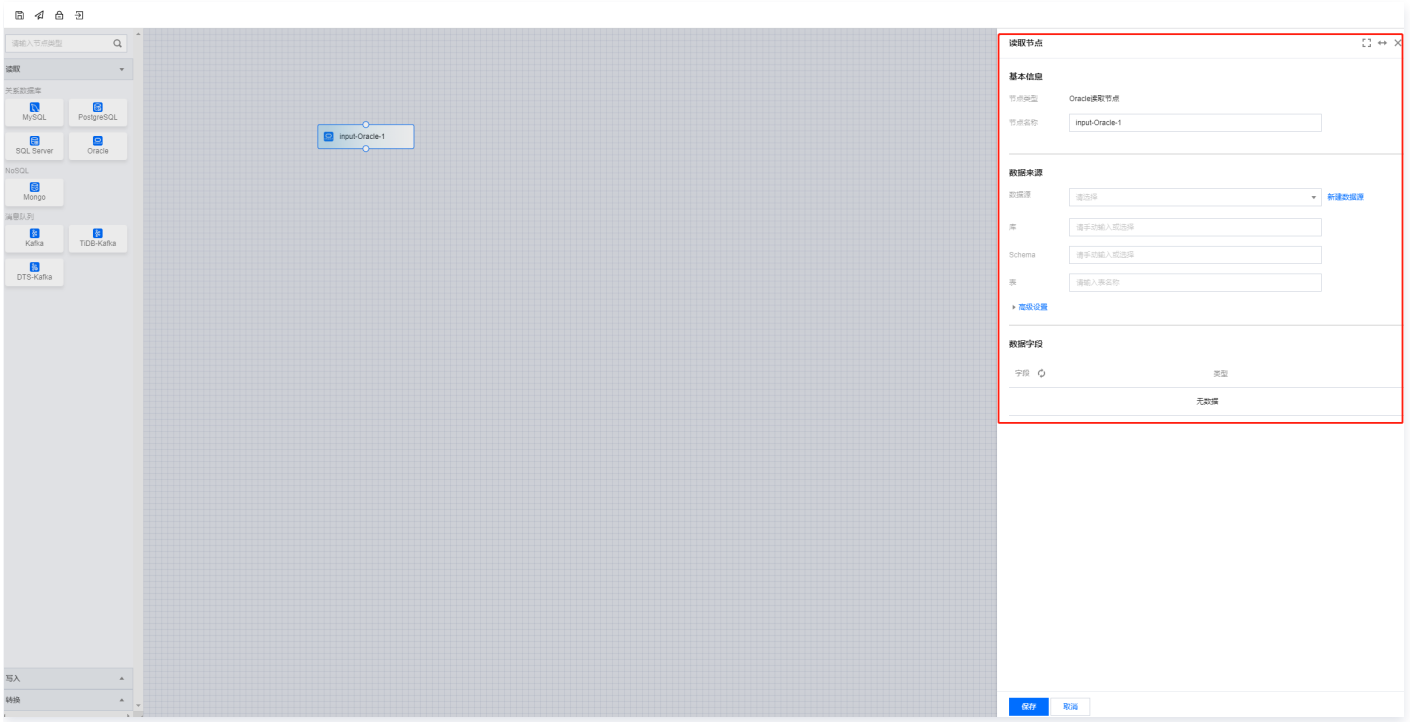
SQLserver 数据类型转换（实时任务）

# Oracle 单表读取

最近更新时间：2024-08-08 11:53:41

## Oracle 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 Oracle 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 Oracle 节点名称。 |
| 数据源 | 选择需要同步的表所在数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段，单击**保存**。

## 附录

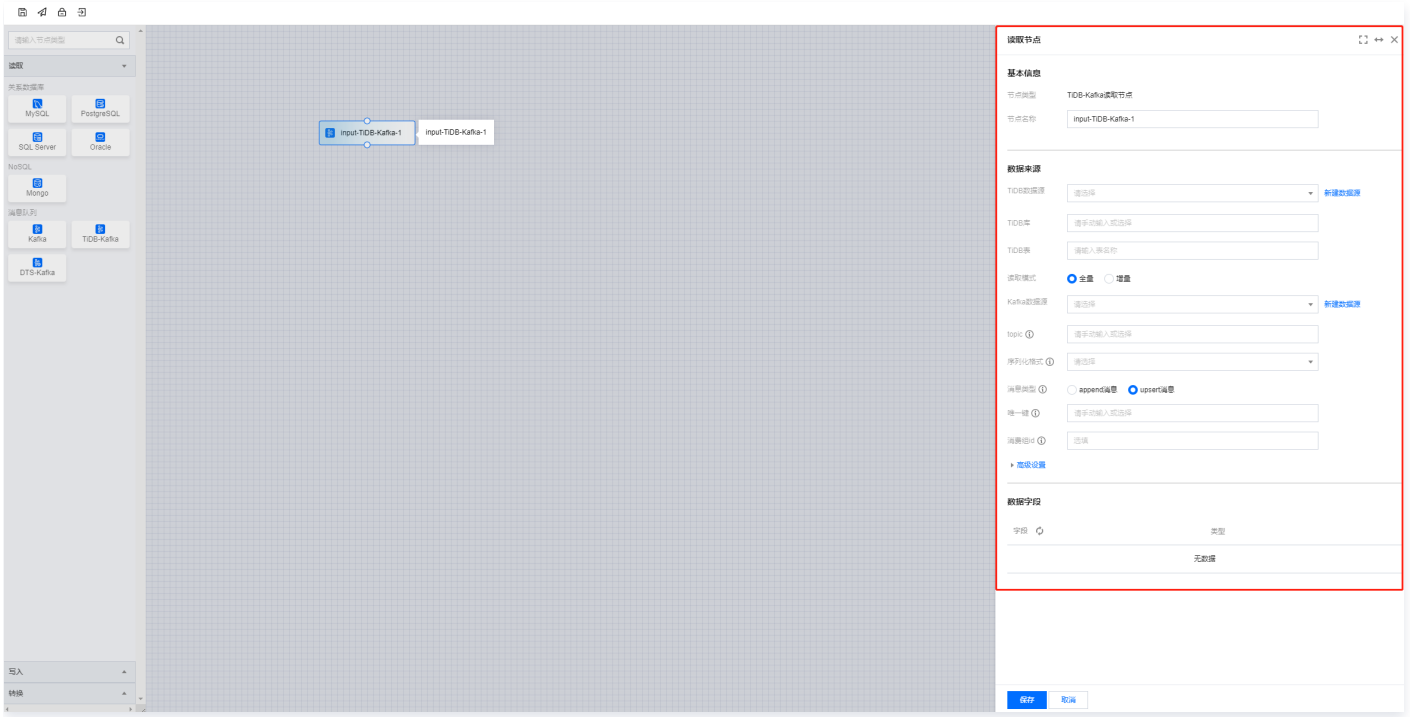
Oracle 实时任务数据类型转换

# TiDB-kafka 单表读取

最近更新时间：2024-08-06 18:01:21

## 条件与 Oracle 读取配置参数说明

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**读取**，单击选择 Oracle 节点并配置节点信息。




4. 参数信息

| 参数 | 描述 |
| --- | --- |
| TiDB 数据源 | 选择需要同步的表所在数据源。 |
| TiDB 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| TiDB 表 | 支持选择、或者手动输入需读取的表名称。 |
| 读取模式 | TiDB 数据源目前支持全量和增量两种读取模式。 |
| Kafka 数据源 | kafka 读取端数据源类型支持 kafka、Ckafka。 |
| topic | Kafka 的Topic，是Kafka处理资源的消息源（feeds of messages）的聚合。 |
| 序列化格式 | Kafka消息序列化格式类型，目前支持 JSON 和 craft 两种类型 。 |
| 消息类型 | kafka 读取端数据源支持两种消息类型：<br>• append 消息：kafka内消息来源于 append 消息流，通常消息中不携带唯一键。写入节点建议搭配 append 写入模式。<br>• upsert 消息：kafka内消息来源于 upsert 消息流，通常消息中携带唯一键，设置后消息可保证 Exactly-Once。写入节点建议搭配 upsert 写入模式。 |
| 唯一键 | Upsert写入模式下，需设置唯一键保证数据有序性。 |
| 消费组 ID | 请避免该参数与其他消费进程重复，以保证消费位点的正确性。如果不指定该参数，默认设定 group.id=WeData_ group_ ${任务id}。 |

| | |
|---|---|
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段，单击**保存**。

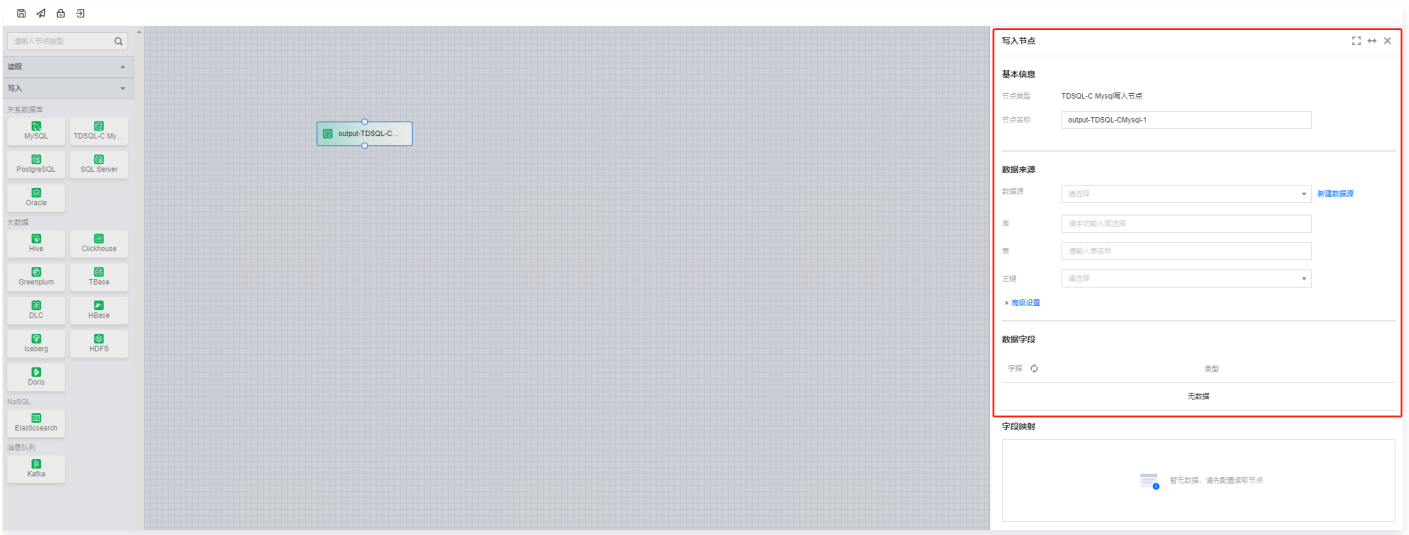
# 写入节点

# TDSQL-C MySQL 单表写入

最近更新时间：2024-07-25 11:24:21

## 配置 TDSQL-C MySQL 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 **TDSQL-C MySQL 节点**并配置节点信息。




4. 您可以参考下表进行参数配置。

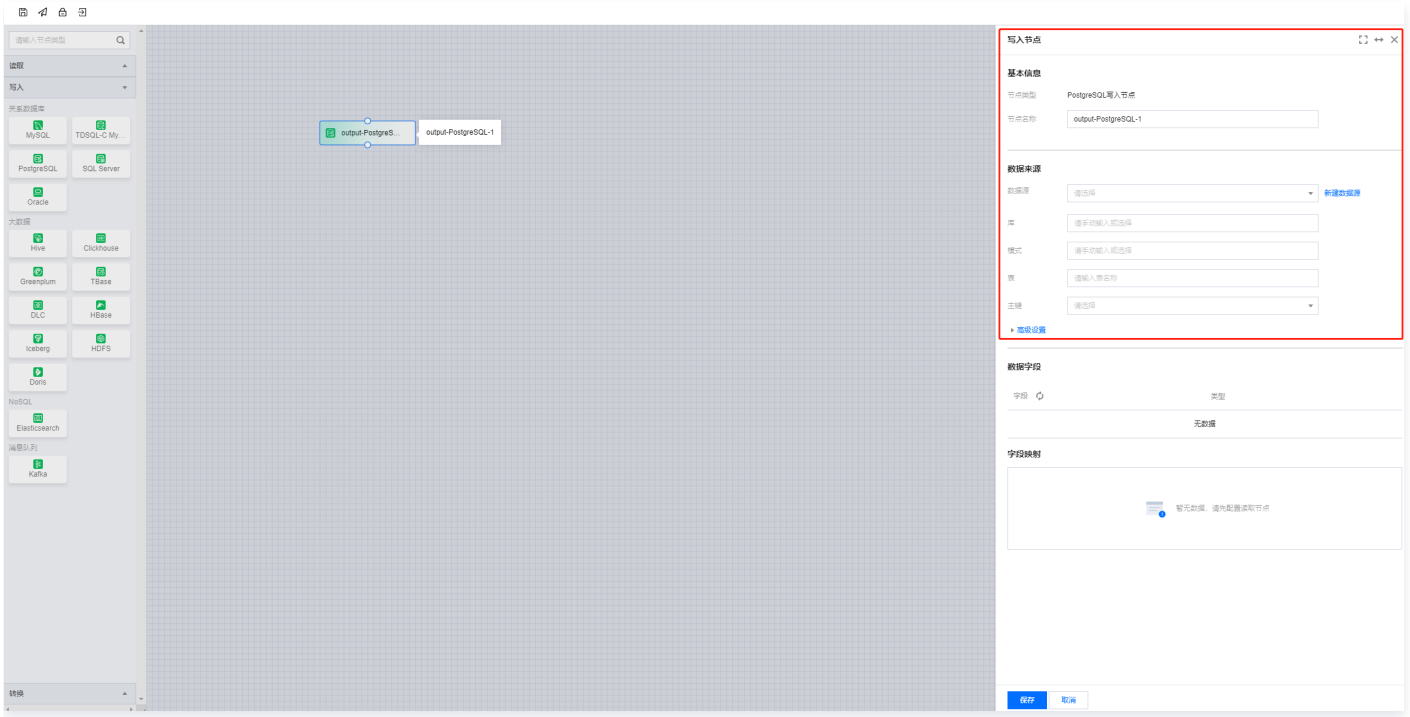
| 参数 | 说明 |
| --- | --- |
| 节点名称 | 输入 TDSQL-C Mysql 节点名称。 |
| 数据源 | 需要写入的 TDSQL-C Mysql 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# PostgreSQL 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 PostgreSQL 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击**选择 PostgreSQL 节点**并配置节点信息。




4. 您可以参考下表进行参数信息配置。

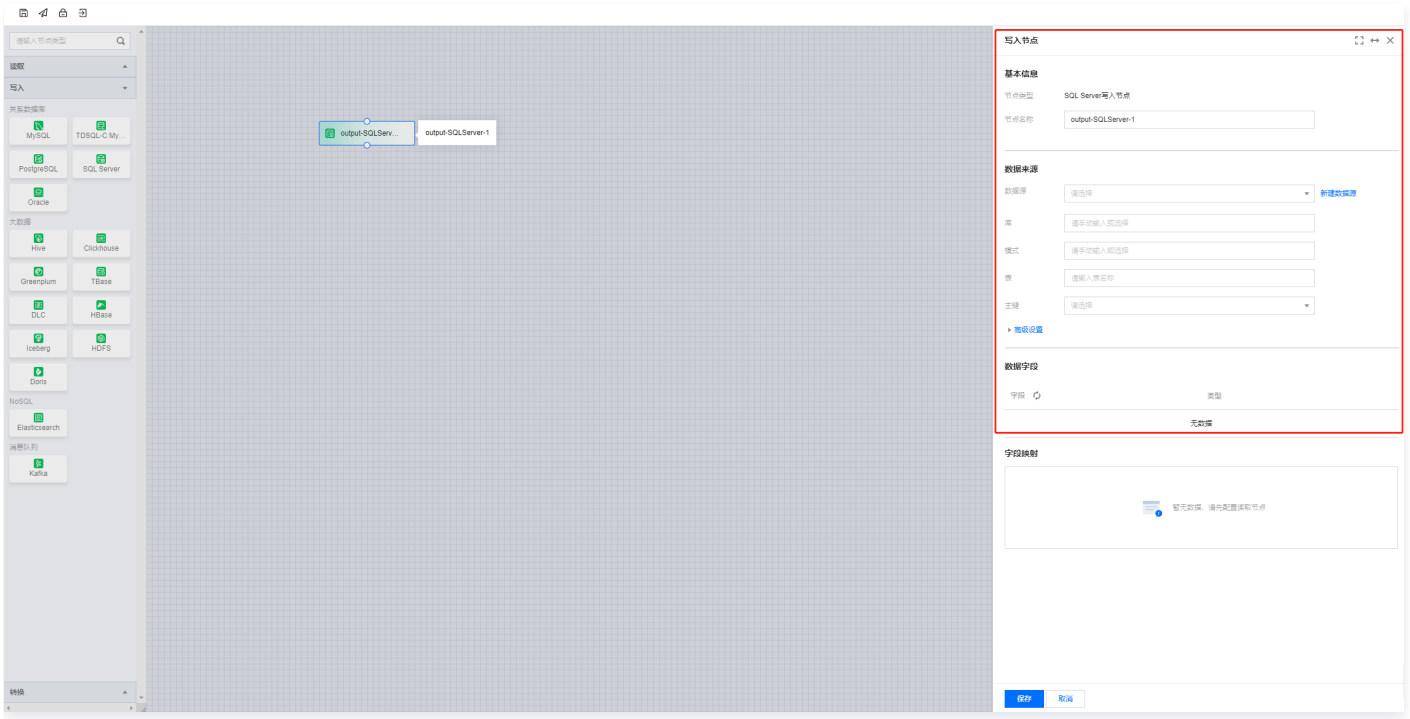
| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 PostgreSQL 节点名称。 |
| 数据源 | 选择需要写入的表所在数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需写入该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# SQL Server 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 SQL Server 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 **SQL Server 节点**并配置节点信息。




4. 您可以参考下表进行参数信息配置。

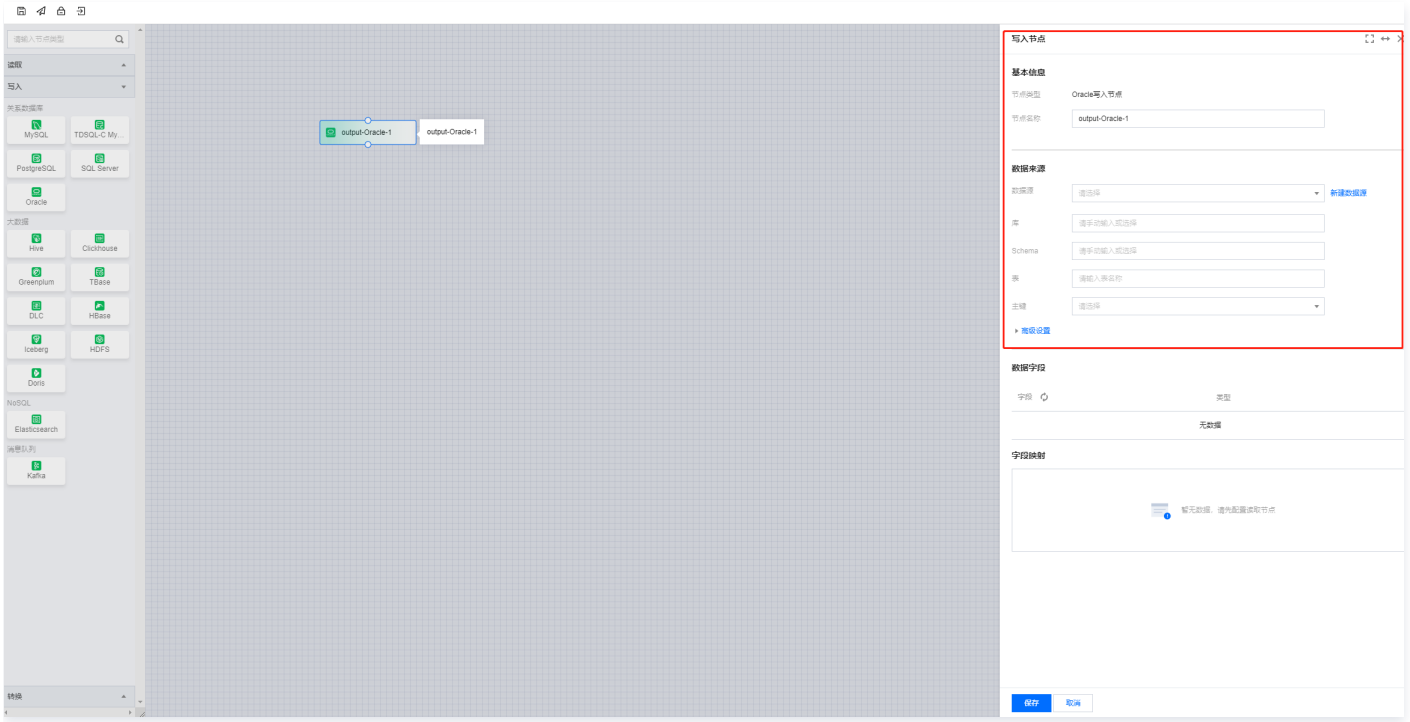
| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 SQL Server 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 SQL Server 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# Oracle 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 Oracle 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 **Oracle 节点**并配置节点信息。




4. 您可以参考下表进行参数信息配置。

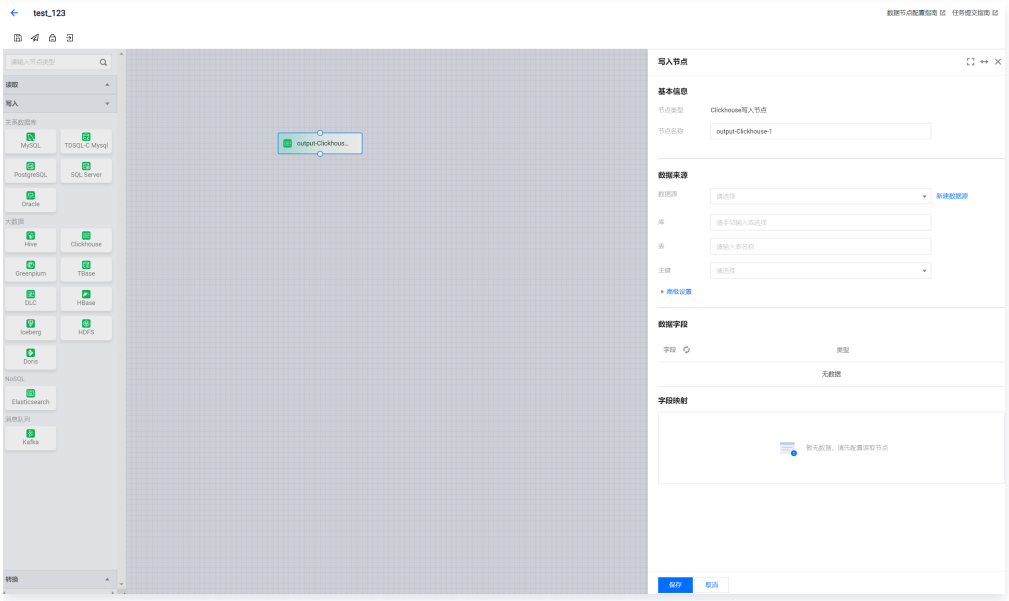
| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Oracle 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| Schema | 支持选择、或者手动输入需要写入的 Oracle 数据模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# ClickHouse 单表写入

最近更新时间：2024-07-09 22:01:41

## 配置 ClickHouse 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 ClickHouse 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 ClickHouse  节点名称。 |
| 数据源 | 选择需要写入的 ClickHouse 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通，导致无法直接拉取库信息时，可手动输入数据库名称。<br>• 在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通，导致无法直接拉取表信息时，可手动输入表名称。<br>• 在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 分区字段 | 选择一个或多个字段作为分区字段。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览字段并与写入节点配置字段映射，单击**保存**。

## 附录

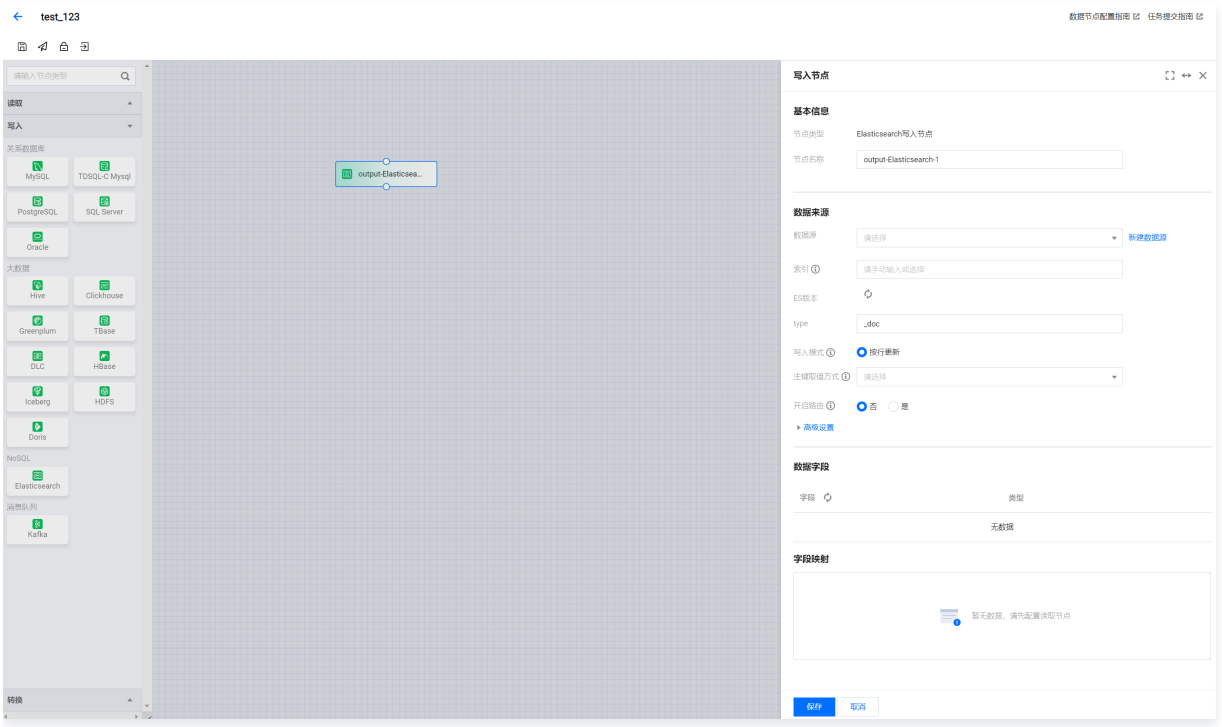
ClickHouse 数据类型转换（实时任务）

# Elasticsearch 单表写入

最近更新时间：2024-07-09 22:01:41

## 配置 Elasticsearch 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Elasticsearch 节点并配置节点信息。




4. 您可以参考下表进行参数配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 Elasticsearch 节点名称 |
| 数据源 | 选择需要写入的 Elasticsearch 数据源 |
| 索引 | ElasticSearch 中的索引名称 |
| 写入模式 | 更新每行记录所有字段（目前支持按行更新） |
| 主键取值方式 | 源表主键：document 的 ID 使用源表的主键联合主键：document 的 ID 使用源表的多个列共同确定无主键：默认生成_id 值 |
| 开启路由 | Elasticsearch 是否开启路由分区索引数据。开启路由功能后，可控制在 ElasticSearch 中使用哪个分区来存储文档 |

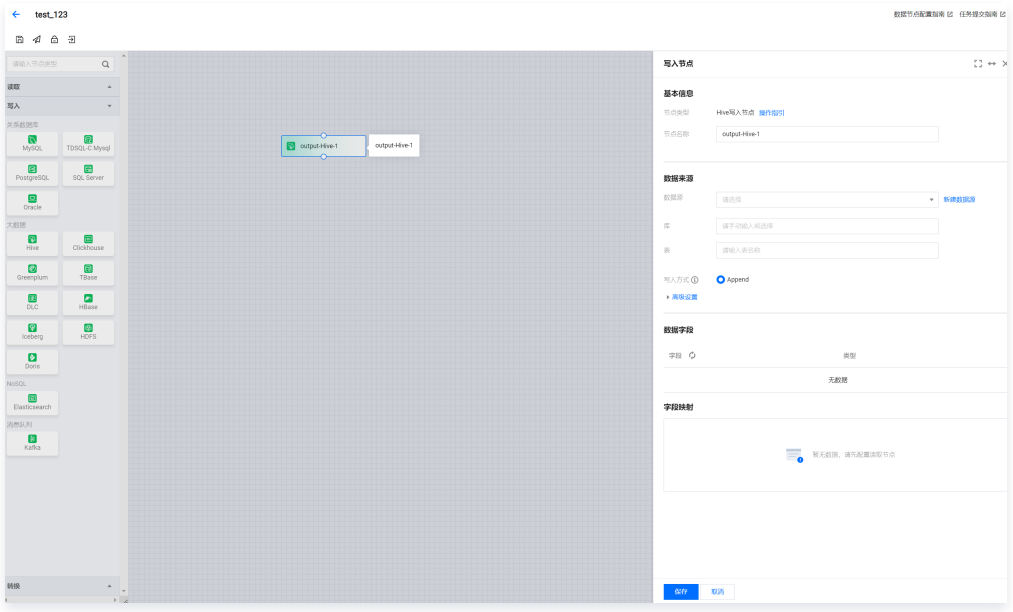
5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

# Hive 单表写入

最近更新时间：2024-08-06 18:01:21

## 配置 Hive 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Hive 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Hive 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 写入方式 | Hive 实时同步仅支持 Append 写入。<br>• Append：保留原始数据，新行追加写入 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

## 附录

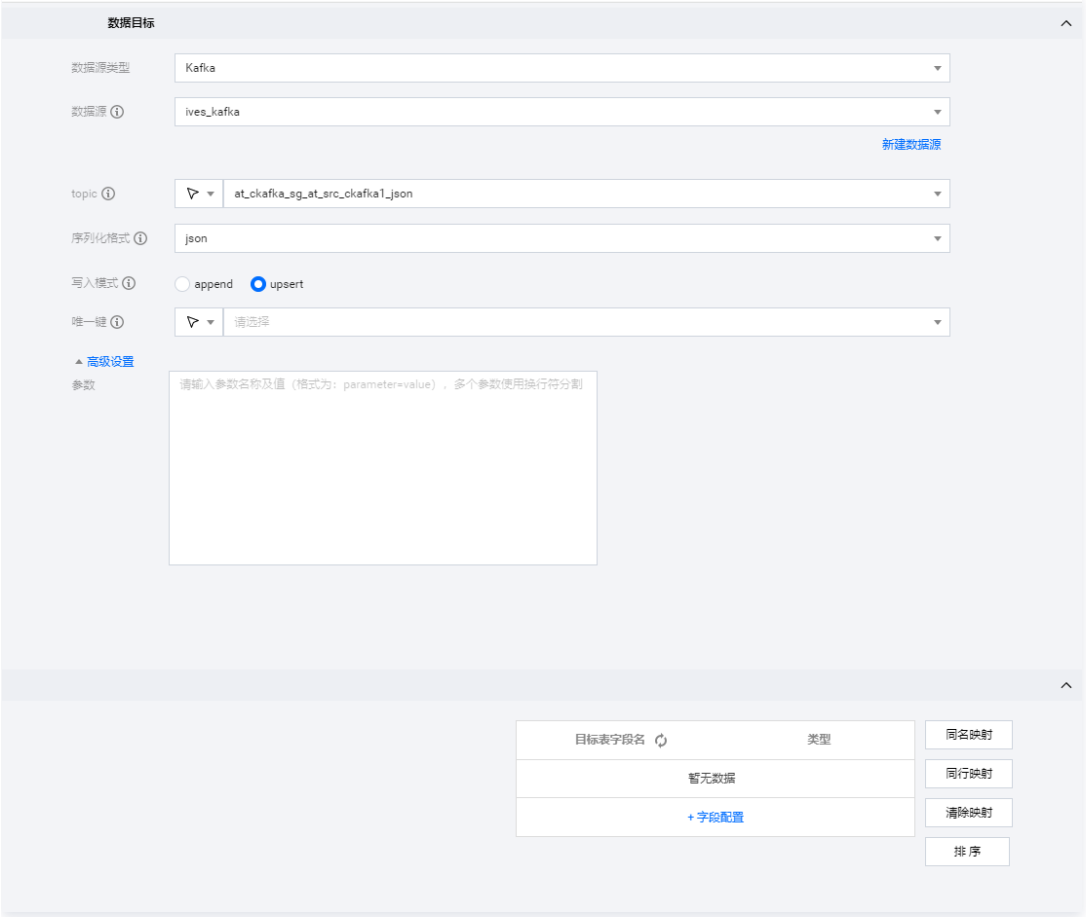
Hive 数据类型转换（实时任务）

# Kafka 单表写入

最近更新时间：2024-07-09 22:01:41

## 创建 Kafka 节点

1. 在数据集成页面左侧目录栏单击 **实时同步**。
2. 在实时同步页面上方选择 **单表同步** 新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧 **写入**，单击选择 Kafka 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 Kafka 节点名称。 |
| 数据源 | Kafka 写入端数据源类型支持 Kafka、Ckafka。 |
| topic | Kafka 数据源中的 Topic。 |
| 序列化格式 | Kafka 消息序列化格式类型，支持：canal-json、json、avro、csv。 |
| 写入模式 | • Append：追加写入。<br>• Upsert：以 Upsert 方式插入消息，设置后消息仅只能被消息端处理一次以保证 Exactly-Once。 |
| 唯一键 | Upsert 写入模式下，需设置唯一键保证数据有序性 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

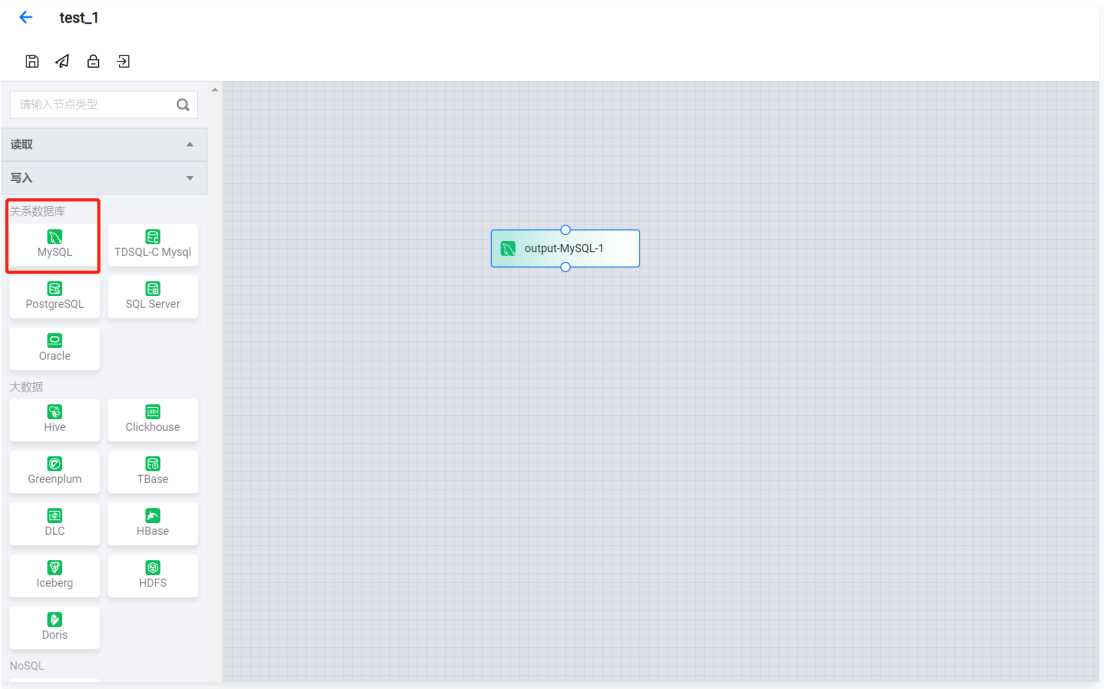
5. 预览字段，单击 **保存**。

# MySQL 单表写入

最近更新时间：2024-07-09 22:01:41

## 创建 MySQL 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 MySQL 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 描述 |
| --- | --- |
| 节点名称 | 输入 MySQL 节点名称。 |
| 数据源 | 选择需要同步的表所在数据源。 |
| 库 | 选择需要同步的表所在数据库。 |
| 表 | 支持选择多个表，请保证多表 schema 一致。 |
| 表主键 | 分库分表模式下默认表 schema 一致。系统将使用拉去第一张表的主键，请选择或输入表主键字段名称。 |
| 格式 | 指定 MySQL 日志编码格式（utf-8、gbk、Latin1、utf8mb4）。 |
| 读取模式 | 支持全量和增量两种模式。 |
| 过滤操作 | 设置后将不同步指定操作类型的数据，支持插入、更新和删除。 |

5. 预览数据字段，单击**保存**。

## 注意事项

1. 为每个 Reader 设置一个不同的 SERVER ID。
   每一个读取 Binlog 的 MySQL 数据库客户端都应该有一个唯一的 ID，称为 SERVER ID。MySQL 服务器将使用此 ID 来维护网络连接和 Binlog 位置。因此，如果不同的作业共享相同的服务器 ID，可能会导致从错误的 Binlog 位置读取。因此，建议通过 SQL Hints ，例如假设源并行度为4，那么我们可以使用 `SELECT * FROM source_table /*+ OPTIONS('server-id'='5401-5404') */` ；为 4个 Source Reader 中的每一个分配唯一的服务器 ID。

2. 设置 MySQL 会话超时。
   当为大型数据库制作初始一致快照时，您建立的连接可能会在读取表时超时。您可以通过在 MySQL 配置文件中配置 interactive_timeout 和 wait_timeout 来防止这种行为。
   - interactive_timeout：服务器在关闭交互式连接之前等待其活动的秒数。请参阅 MySQL :: MySQL 8.0 Reference Manual :: 5.1.8 Server System Variables 。
   - wait_timeout：服务器在关闭非交互式连接之前等待其活动的秒数。请参阅 MySQL :: MySQL 8.0 Reference Manual :: 5.1.8 Server System Variables 。

## 附录

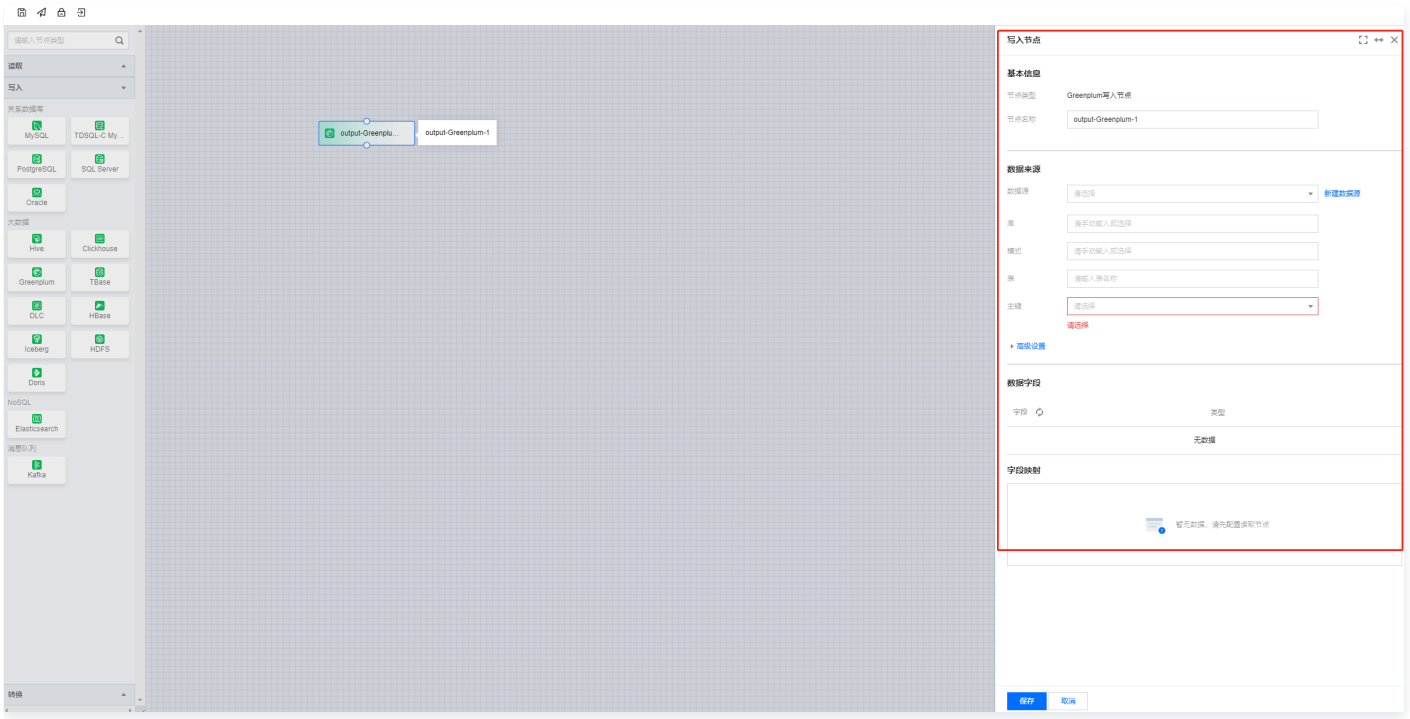
MySQL 实时任务数据类型转换

# Greenplum 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 Greenplum 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Greenplum 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Greenplum 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 Greenplum 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

## 附录

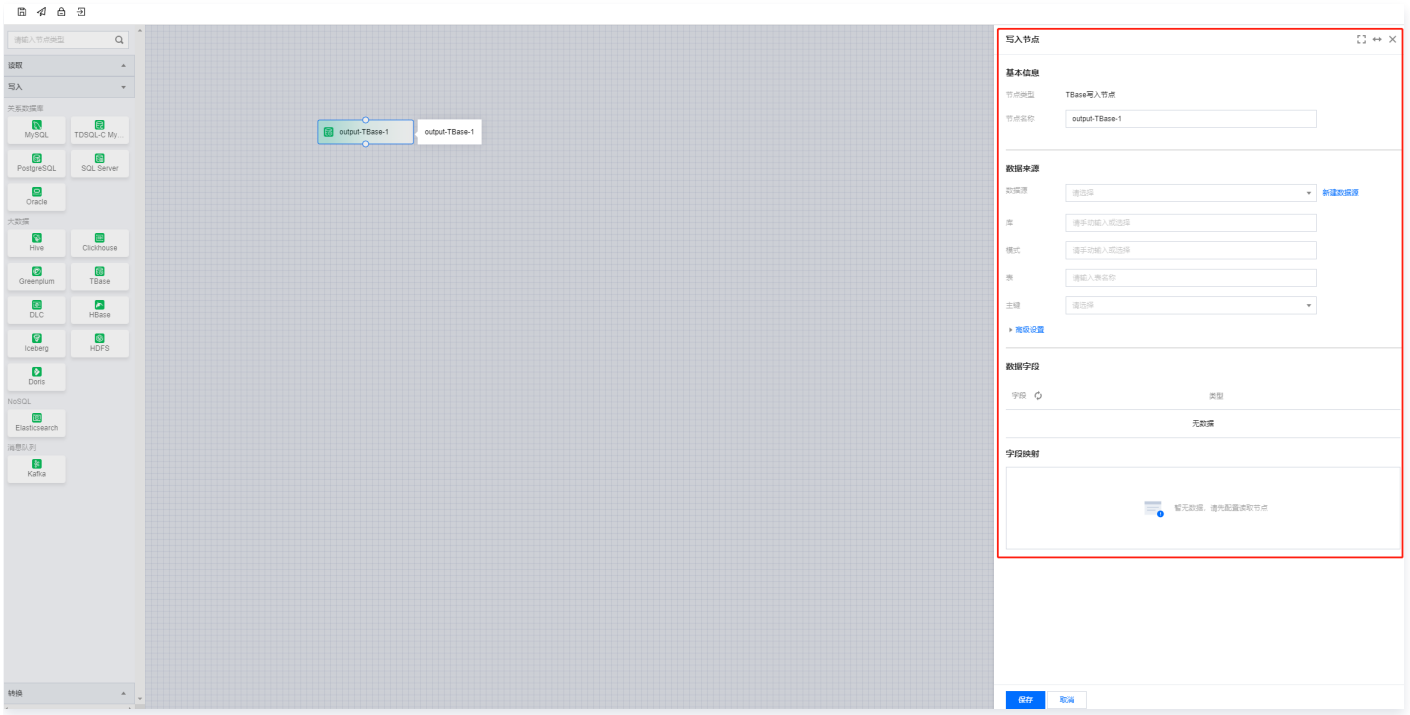
Greenplum 数据类型转换（实时任务）

# Tbase 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 Tbase 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Hive 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Tbase 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 Tbase 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 主键 | 选择一个字段作为写入表的主键。 |
| 高级设置 | 可根据业务需求配置参数。 |

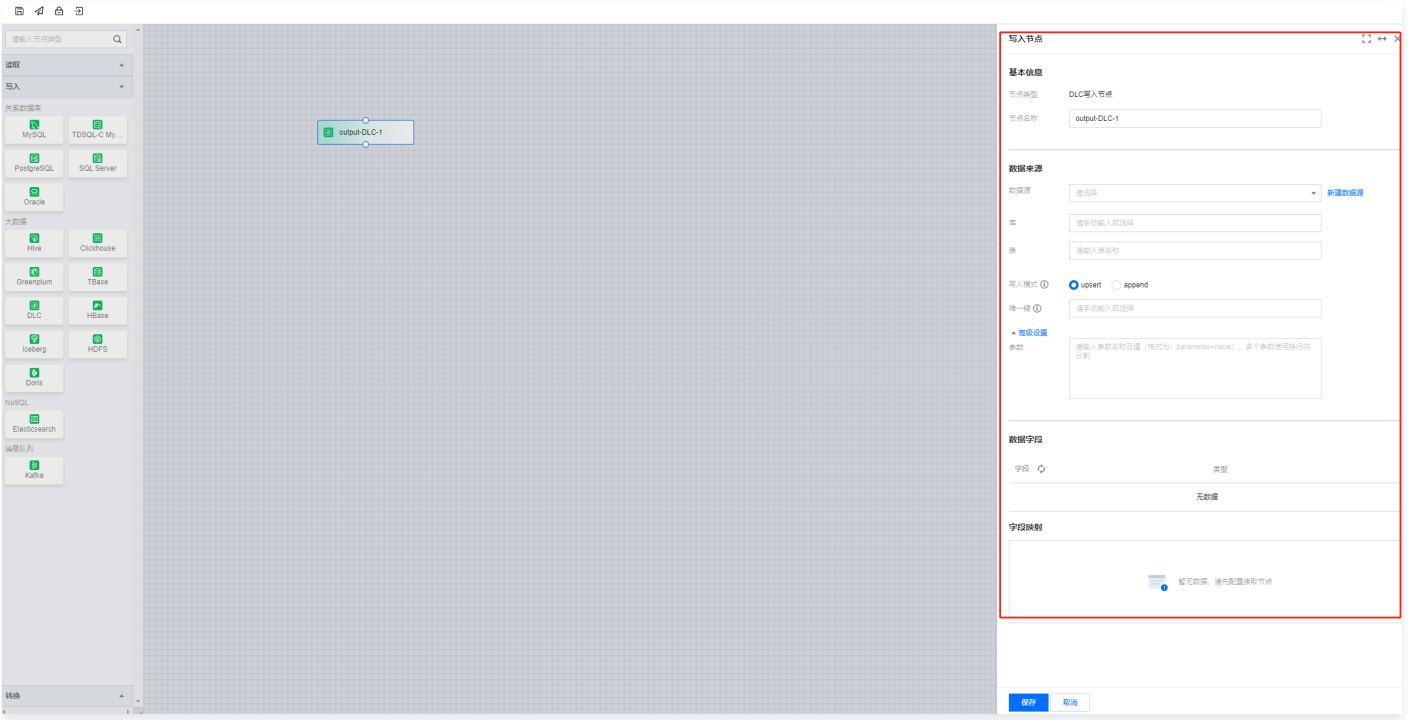
5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

# DLC 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 DLC 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 DLC 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 DLC 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 写入模式 | DLC 实时同步写入支持两种模式：<br>append：追加写入，主键冲突时报错。<br>upsert：更新写入，主键冲突时更新数据。 |
| 唯一键 | Upsert 写入模式下，需设置唯一键保证数据有序性，支持多选。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

## 附录

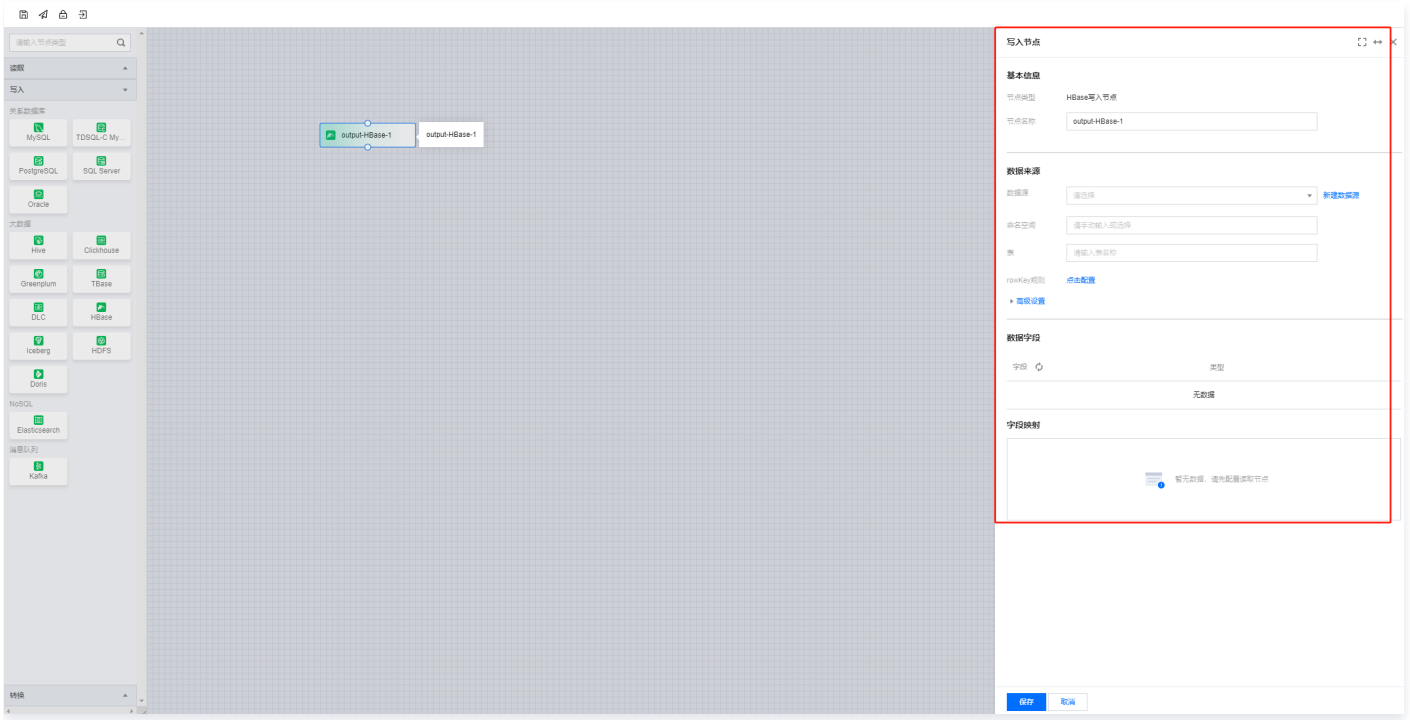
DLC - iceberg / Iceberg 数据类型转换（实时任务）

# Hbase 单表写入

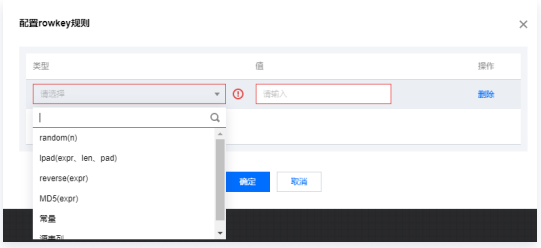
最近更新时间：2024-06-18 14:47:41

## 配置 Hbase 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Hbase 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 HBase 数据源。 |
| 命名空间 | 支持选择、或者手动输入需写入的空间。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| rowkey 规则 | 单击配置即可进入 rowkey 的配置页面，配置类型和对应的值即可。<br>  |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

## 附录

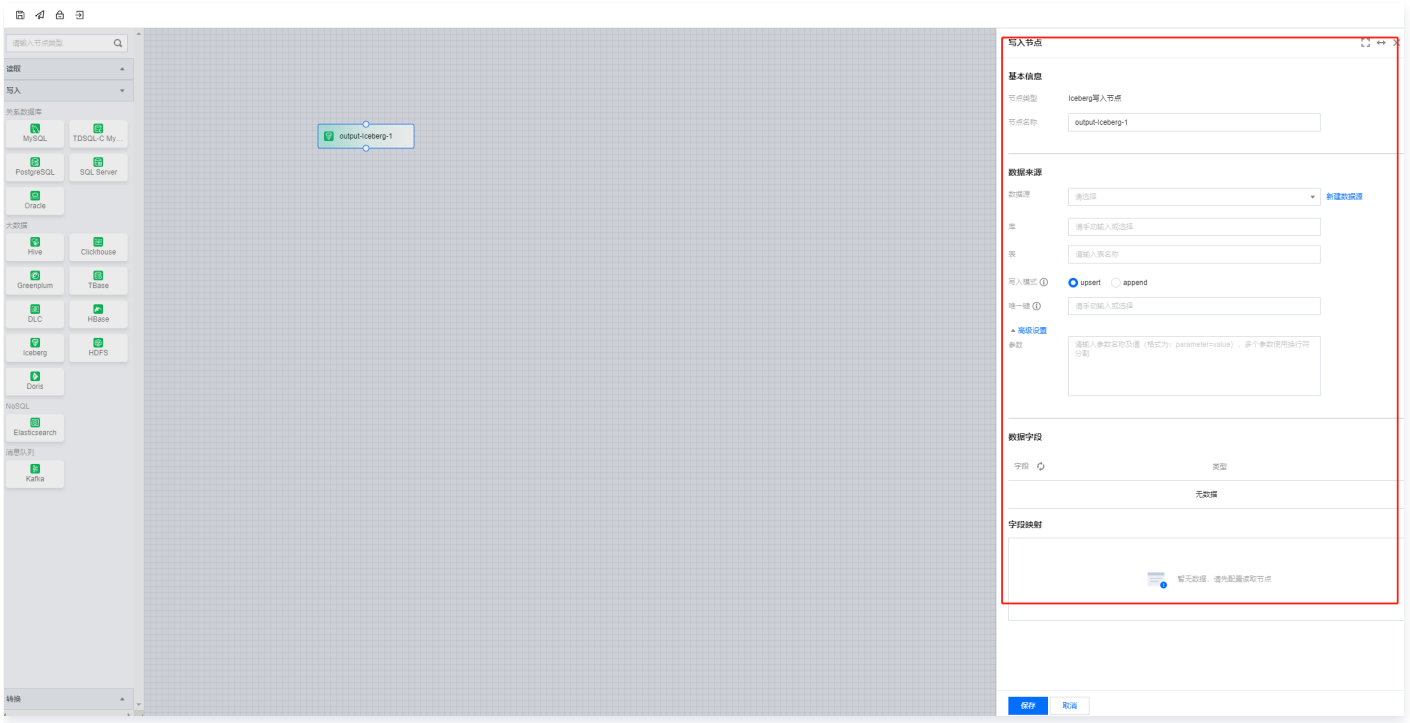
HBase 数据类型转换（实时任务）

# Iceberg 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 Iceberg 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Iceberg 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Iceberg 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 写入模式 | Iceberg 实时同步写入支持两种模式：<br>append：追加写入，主键冲突时报错。<br>upsert：更新写入，主键冲突时更新数据。 |
| 唯一键 | 选择一个字段作为写入表的主键。 |
| 高级设置 | 可根据业务需求配置参数。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

## 附录

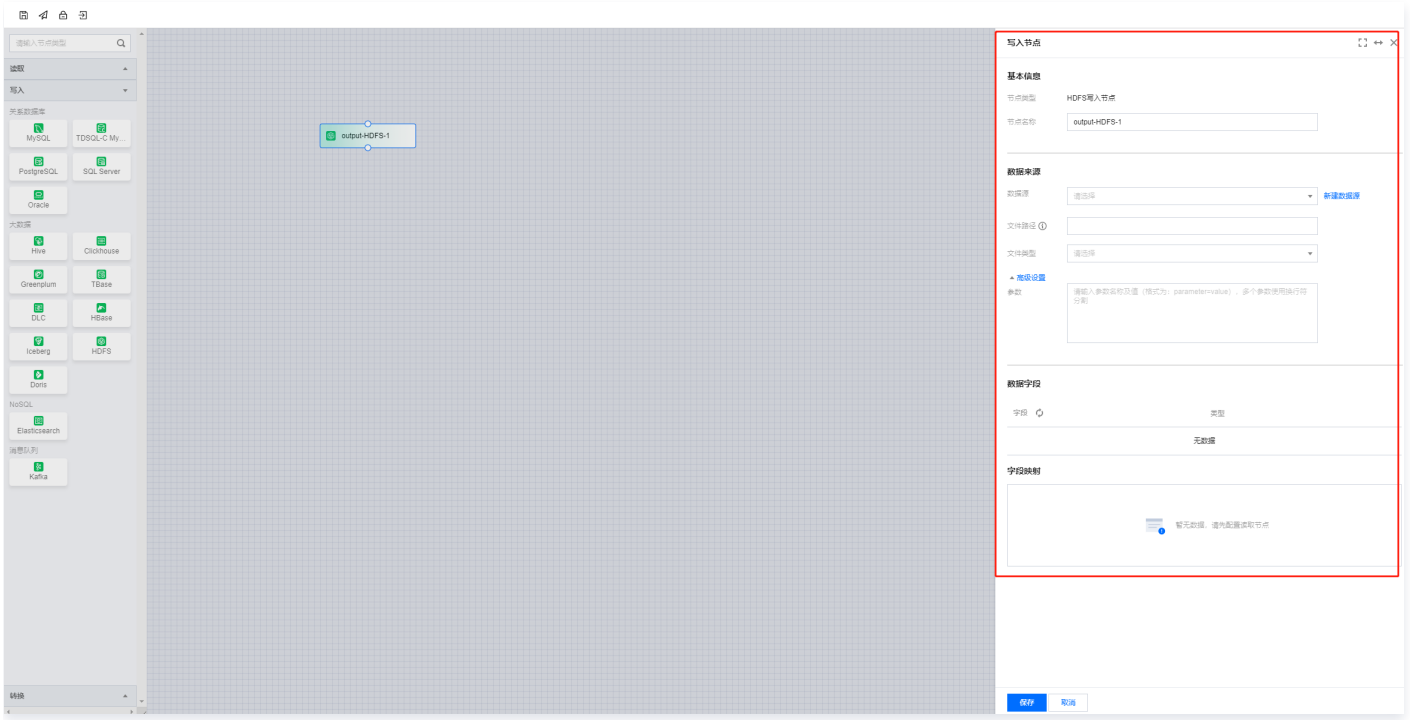
DLC – iceberg / Iceberg 数据类型转换（实时任务）

# HDFS 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 HDFS 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 HDFS 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 HDFS 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用‘*’作为通配符，指定通配符后将遍历多个文件信息。 |
| 文件类型 | HDFS 支持四种文件类型：txt 、orc 、parquet 、csv。<br>• txt：表示 TextFile 文件格式。<br>• orc：表示 ORCFile 文件格式。<br>• parquet：表示普通 Parquet 文件格式。<br>• csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 高级设置 | 可根据业务需求配置参数。 |

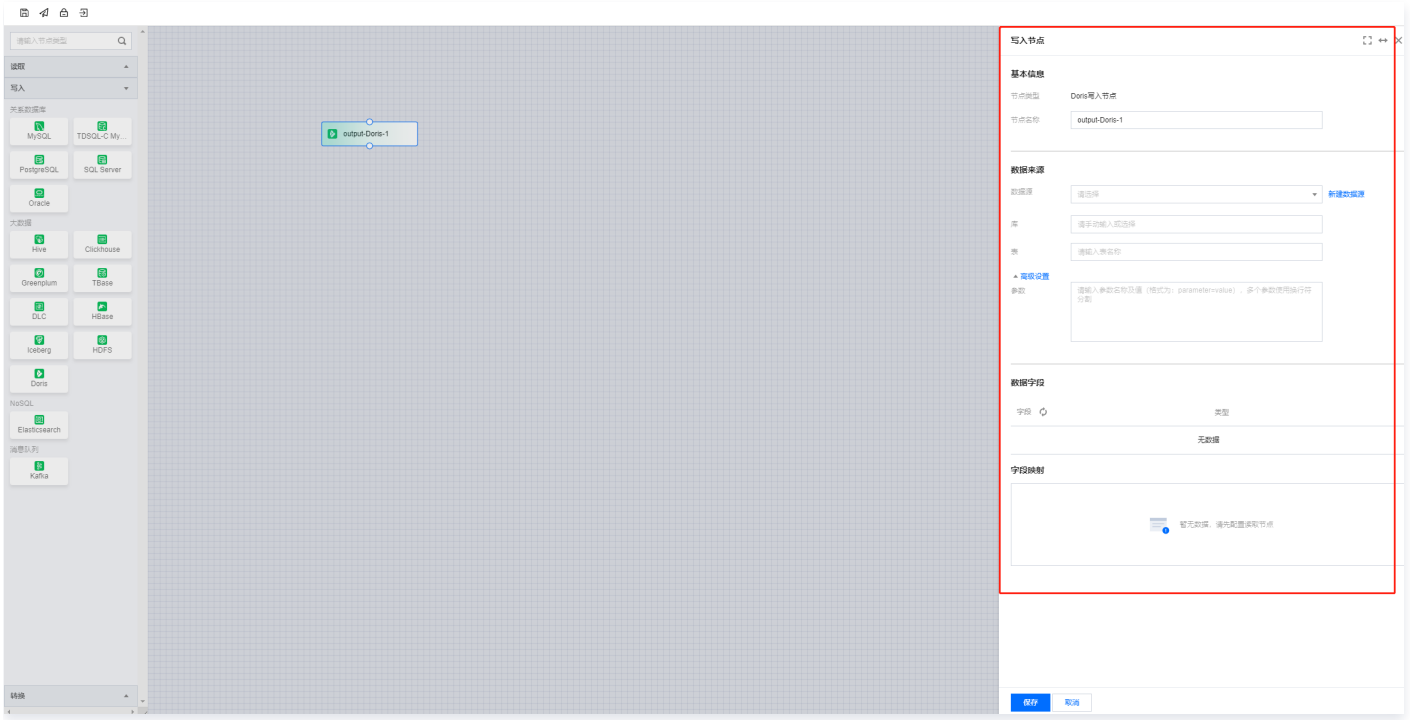
5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

# Doris 单表写入

最近更新时间：2024-06-18 14:47:41

## 配置 Tbase 节点

1. 在数据集成页面左侧目录栏单击**实时同步**。
2. 在实时同步页面上方选择**单表同步**新建（可选择表单和画布模式）并进入配置页面。
3. 单击左侧**写入**，单击选择 Doris 节点并配置节点信息。




4. 您可以参考下表进行参数信息配置。

| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Doris 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 高级设置（可选） | 可根据业务需求配置参数。要求如下：<br>1. 一个参数一行；若需配合使用的参数写在一行内；<br>2. 每个参数带默认值。 |

5. 预览数据字段并与读取节点配置字段映射，单击**保存**。

# 实时整库
# MySQL/TDSQL-C MySQL 整库来源配置详情

最近更新时间：2024-08-15 21:18:52

## 使用限制

MySQL 整库同步过程中，任务使用表主键或第一个字段的 type（当表无主键时）作为后续切分表的 type 的依据。当前仅支持以下类型：

> **主键支持范围类型：**
> TINYINT、TINYINT_UNSIGNED、SMALLINT、SMALLINT_UNSIGNED、INT、MEDIUMINT、INT_UNSIGNED、MEDIUMINT_UNSIGNED、BIGINT、BIGINT_UNSIGNED、FLOAT、DOUBLE、DECIMAL、TIME、DATE、DATETIME、TIMESTAMP、CHAR、VARCHAR、TEXT、BINARY、VARBINARY、BLOB.

支持数据库版本详情：

| 节点 | 版本 | Driver |
| --- | --- | --- |
| MySQL-CDC | MySQL：5.6，5.7，8.0.x<br>RDS MySQL：5.6，5.7，8.0.x<br>PolarDB MySQL：5.6，5.7，8.0.x<br>Aurora MySQL：5.6，5.7，8.0.x<br>MariaDB：10.x<br>PolarDB X：2.0.1 | JDBC Driver：8.0.21 |

## 设置 MySQL 服务器权限

您必须定义一个对 Debezium MySQL 连接器监控的所有数据库具有适当权限的 MySQL 用户。

1. 创建 MySQL 用户：

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

2. 向用户授予所需的权限：

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

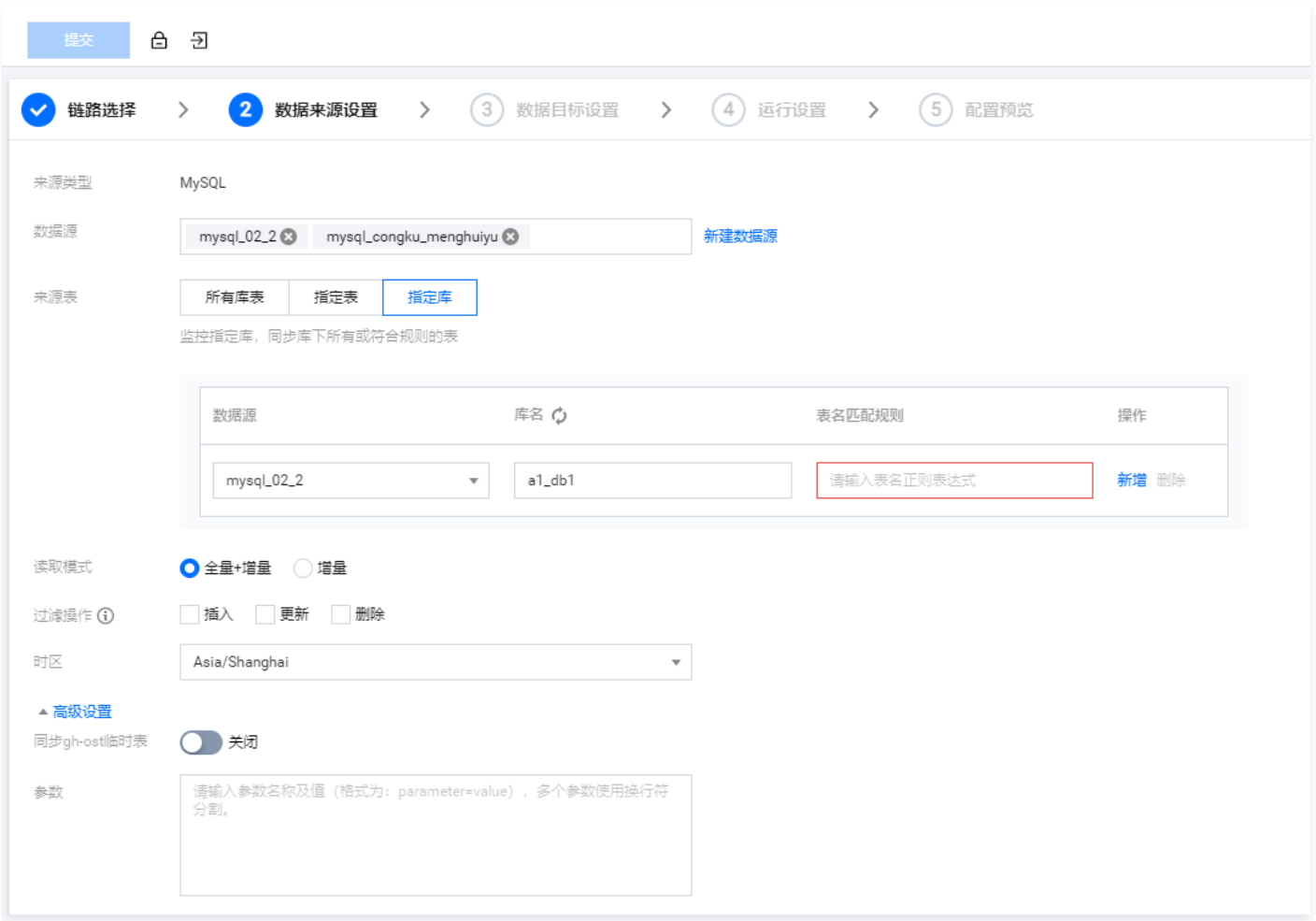
> **注意：**
> 启用 `scan.incremental.snapshot.enabled` 时，不再需要 RELOAD 权限（默认启用）。

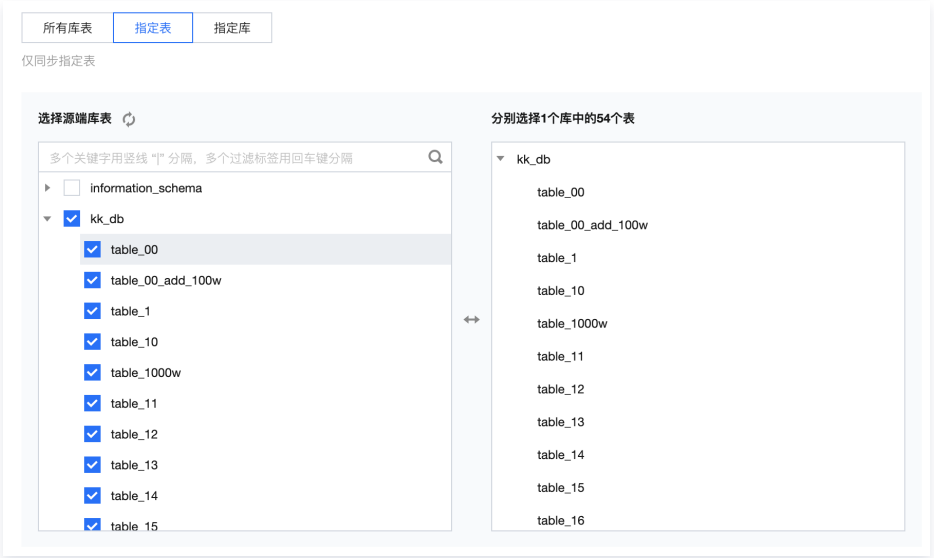
3. 刷新用户的权限：

```
mysql> FLUSH PRIVILEGES;
```

查看更多关于 权限说明 。

## MySQL 读取配置参数说明




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择需要同步的 MYSQL/TDSQL-C MySQL 数据源。 |
| 来源表 | • 所有库表：监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端。<br>• 指定表：此选项下需指定到具体表名称，设置后任务仅同步指定表；若需要新增同步表需停止并重启任务。<br>  <br>• 指定库：此选项下需指定具体库名、以表名正则表达式。设置后，任务运行期间符合表名表达式的新增表默认将同步至目标端。 |

| | |
|---|---|
| |  |
| 读取模式 | • 全量 + 增量：数据同步分为全量和增量同步阶段，全量阶段完成后任务进入增量阶段。全量阶段将同步库内历史数据，增量阶段从任务启动后 binlog cdc 的位点开始同步。<br>• 增量：仅从任务启动后的 binlog cdc 位点开始同步数据。 |
| 过滤操作 | 支持插入、更新和删除三种操作，设置后将不同步指定操作类型的数据。 |
| 锁表 | 开启后系统将在启动和全量同步期间锁定来源表，请确保当前数据库账户已具备锁表权限。 |
| 时区 | 设置日志时间所属时区，默认上海。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

## 附录

- MySQL 实时任务数据类型转换

# Kafka 整库来源配置详情

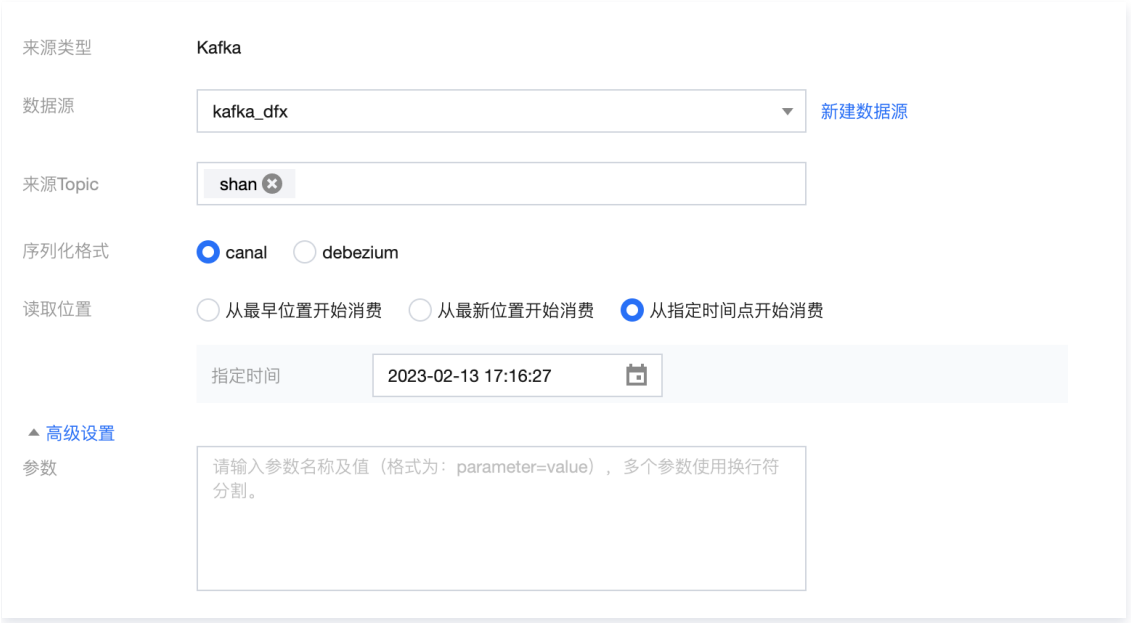
最近更新时间：2024-07-09 22:01:41

## 条件及限制

**支持 Kafka 版本详情：**

| 节点 | 版本 |
| --- | --- |
| Kafka | 0.10+ |

## Kafka 读取配置参数说明




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择需要同步的 Kafka 数据源。 |
| 来源 Topic | 选择或输入任务计划消费的 Topic 名称。 |
| 序列化格式 | 设置 Kafka 内原始消息格式，目前支持解析 canal 和 debezium。<br>**说明：**<br>设置格式需与消息实际格式保持一致。 |
| 读取位置 | 设置 Kafka 数据读取位点：<br>• 从最早开始：earlist。<br>• 从最新开始：latest。<br>• 从指定时间开始：设定具体任务启动时间位点。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# PostgreSQL 整库来源配置详情

最近更新时间：2024-07-09 22:01:41


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择需要同步的 PostgreSQL 数据源。 |
| 来源表 | 根据业务需求，选择“所有库表”、“指定表”、“指定库”。<br>• 所有库表：监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端。<br>• 指定表：仅同步指定表。<br>• 指定库：监控指定库，同步库下所有或符合规则的表。 |
| 读取模式 | 全量 + 增量、增量。 |
| 过滤操作 | 提供多种过滤操作，包括插入、更新、删除、删除集合、删除数据库、重命名集合，设置后将不同步指定操作类型的数据。 |

# Mongo 整库来源配置详情

最近更新时间：2024-07-09 22:01:41

## 条件及限制

- 支持 Mongo 版本详情：

| 节点 | 版本 |
| --- | --- |
| MongoDB-CDC | MongoDB>=3.6 |

## Mongo 读取配置参数说明


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择需要同步的 Mongo 数据源。 |
| 来源表 | 根据业务需求，选择“所有库表”、“指定表”、“指定库”：<br>• 所有库表：监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端。<br>• 指定表：仅同步指定表。<br>• 指定库：监控指定库，同步库下所有或符合规则的表。 |
| 读取模式 | 全量 + 增量、增量。 |
| 过滤操作 | 提供多种过滤操作，包括插入、更新、删除、删除集合、删除数据库、重命名集合，设置后将不同步指定操作类型的数据。 |

## 附录

- Mongo 实时任务数据类型转换

# Oracle 整库来源配置详情

最近更新时间：2024-07-25 11:24:21

**若要监控 Oracle 端表字段变更，数据源请勿配置 system/sys 两个账户，否则所有表（包括新增表）都需要开启日志才能进行同步。开启命令 "ALTER TABLE SCHEMA_NAME.TABLE_NAME ADD SUPPLEMENTAL LOG DATA (ALL) COLUMNS" 。**


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择需要同步的 Oracle 数据源 |
| 来源表 | 根据业务需求，选择“所有库表”、“指定表”、“指定库” |
| | • 所有库表：监控数据源下所有库。任务运行期间新增库、表默认将同步至目标端<br>• 指定表：仅同步指定表<br>• 指定库：监控指定库和 schema，同步 schema 下所有或符合规则的表 |
| 读取模式 | 全量 + 增量、增量 |
| 锁表 | 开启后系统将在启动和全量同步期间锁定来源表，请确保当前数据库账户已具备锁表权限 |
| 高级设置（可选） | 可根据业务需求配置参数 |

# 日志采集
# 写入节点
# MySQL 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 MySQL 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

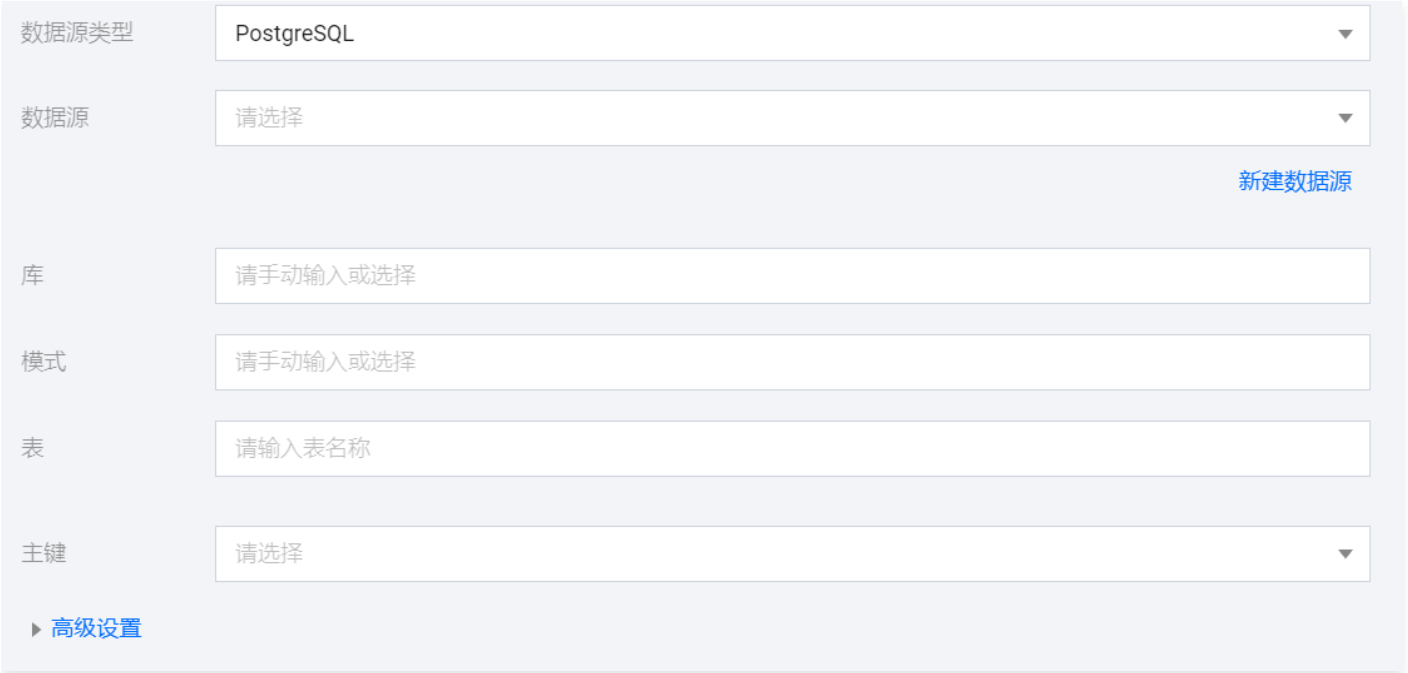
# TDSQL-C MySQL 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 TDSQL-C MySQL 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# PostgreSQL 日志采集

最近更新时间：2024-07-18 17:43:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 PostgreSQL 数据源。 |
| 库/表 | 选择 PostgreSQL 数据源中对应的库表。 |
| 模式 | 选择 PostgreSQL 数据源中的模式。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# SQL Server 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 SQL Server 数据源。 |
| 库/表 | 选择 SQL Server 数据源中对应的库表。 |
| 模式 | 选择 SQL Server 数据源中的模式。 |
| 主键 | 选择一个字段作为数据表主键 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# Oracle 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Oracle 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# HIVE 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Hive 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 写入模式 | Hive 仅支持 Append 写入。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# Clickhouse 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Clickhouse 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# Greenplum 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Greenplum 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 模式 | 选择 Greenplum 数据源中的模式。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

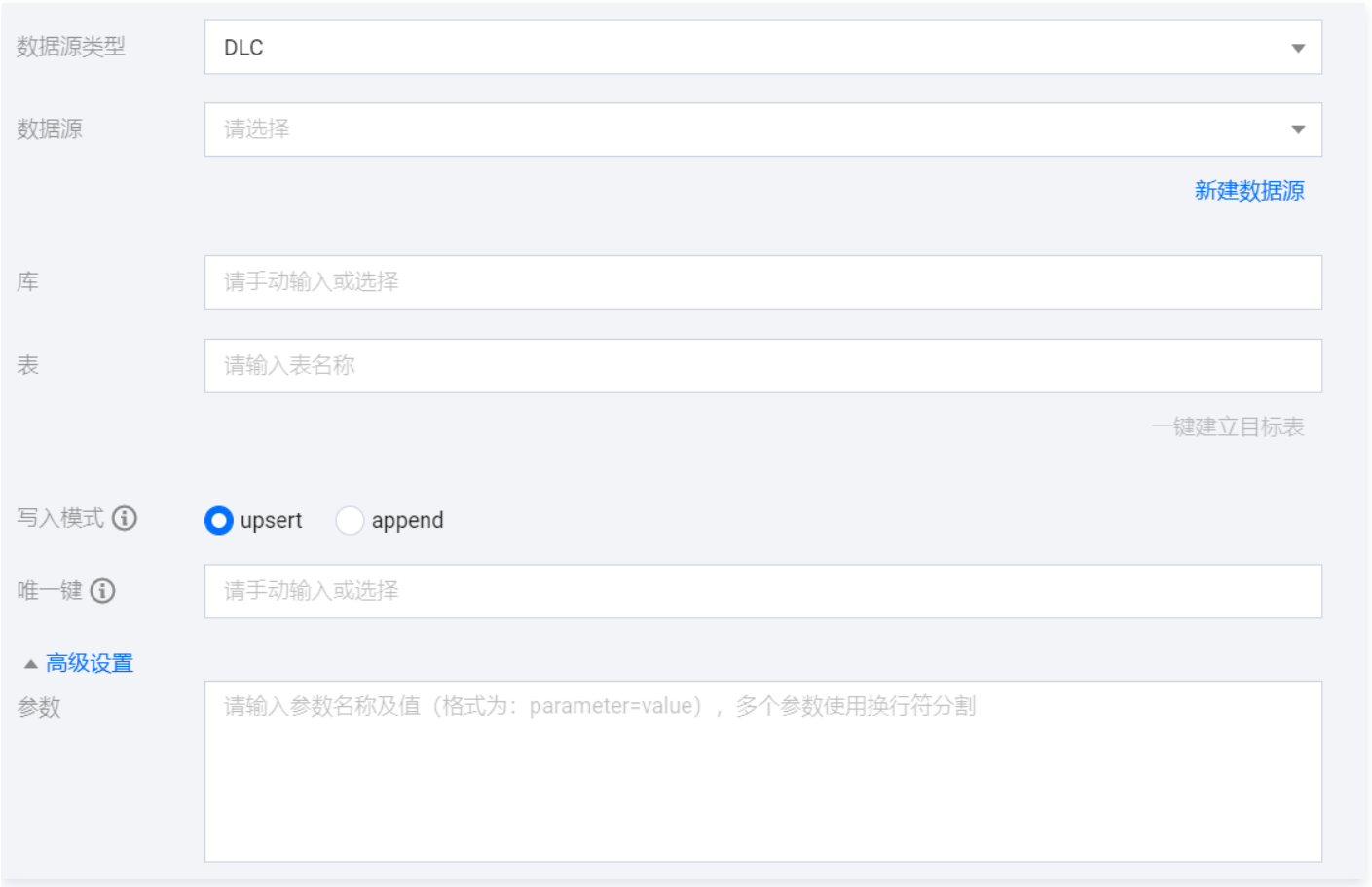
# TBase 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 TBase 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 模式 | 选择 TBase 数据源中的模式。 |
| 主键 | 选择一个字段作为数据表主键。 |
| 高级设置（可选） | 可根据您的业务需求配置参数。 |

# DLC 日志采集

最近更新时间：2024-07-18 17:43:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 DLC 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 写入模式 | DLC 支持两种写入模式：<br>• Append：追加写入。<br>• Upsert：以 Upsert 方式插入消息，设置后消息仅只能被消费端处理一次以保证 Exactly-Once。 |
| 唯一键 | Upsert 写入模式下，需设置唯一键保证数据有序性，支持多选，Append 模式则不需要设置唯一键。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# HBase 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 HBase 数据源。 |
| 命名空间 | 手动输入或者选择命名空间。 |
| 表 | 选择该数据源中对应的表。 |
| rowkey 规则 | HBase 数据源需要配置 rowkey 规则。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

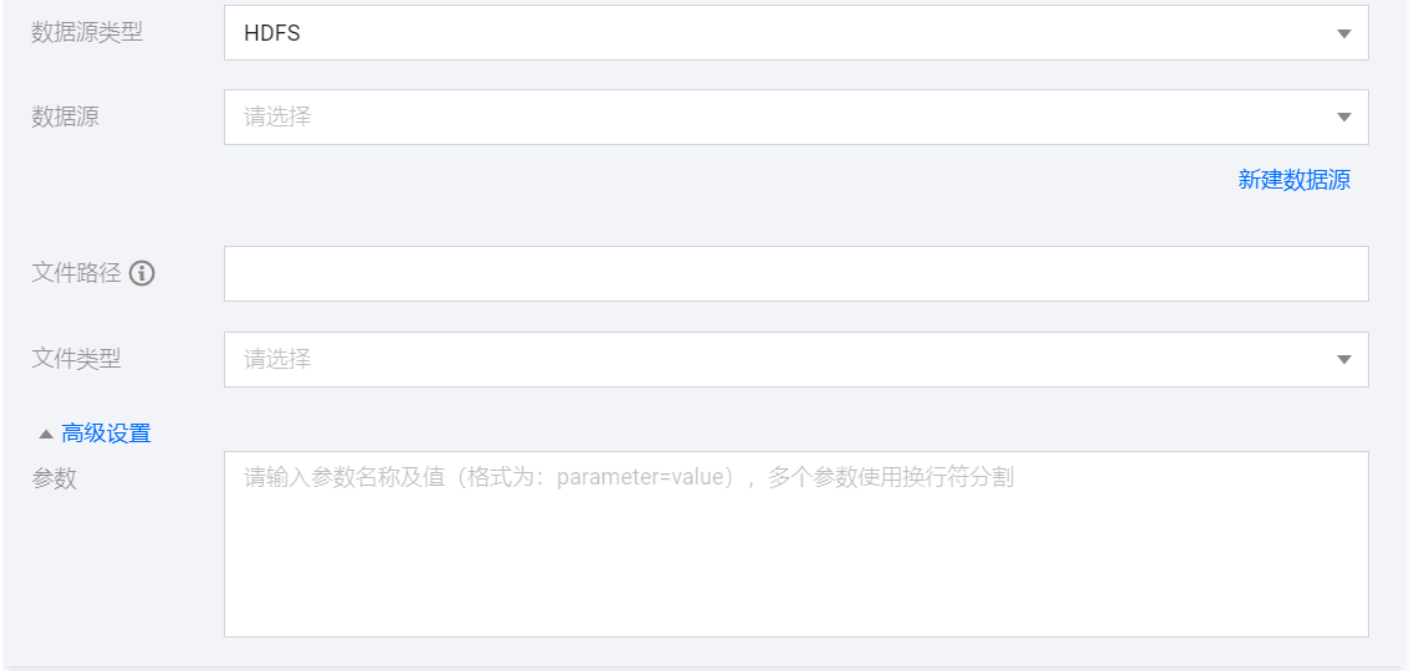
# Iceberg 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Iceberg 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 写入模式 | Iceberg 支持两种写入模式：<br>• Append：追加写入。<br>• Upsert：以 Upsert 方式插入消息，设置后消息仅只能被消费端处理一次以保证 Exactly-Once。 |
| 唯一键 | Upsert 写入模式下，需设置唯一键保证数据有序性，支持多选，Append 模式则不需要设置唯一键。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# HDFS 日志采集

最近更新时间：2024-07-18 17:43:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 HDFS 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用‘*’作为通配符，指定通配符后将遍历多个文件信息。 |
| 文件类型 | HDFS 支持四种文件类型：txt 、orc 、parquet 、csv。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

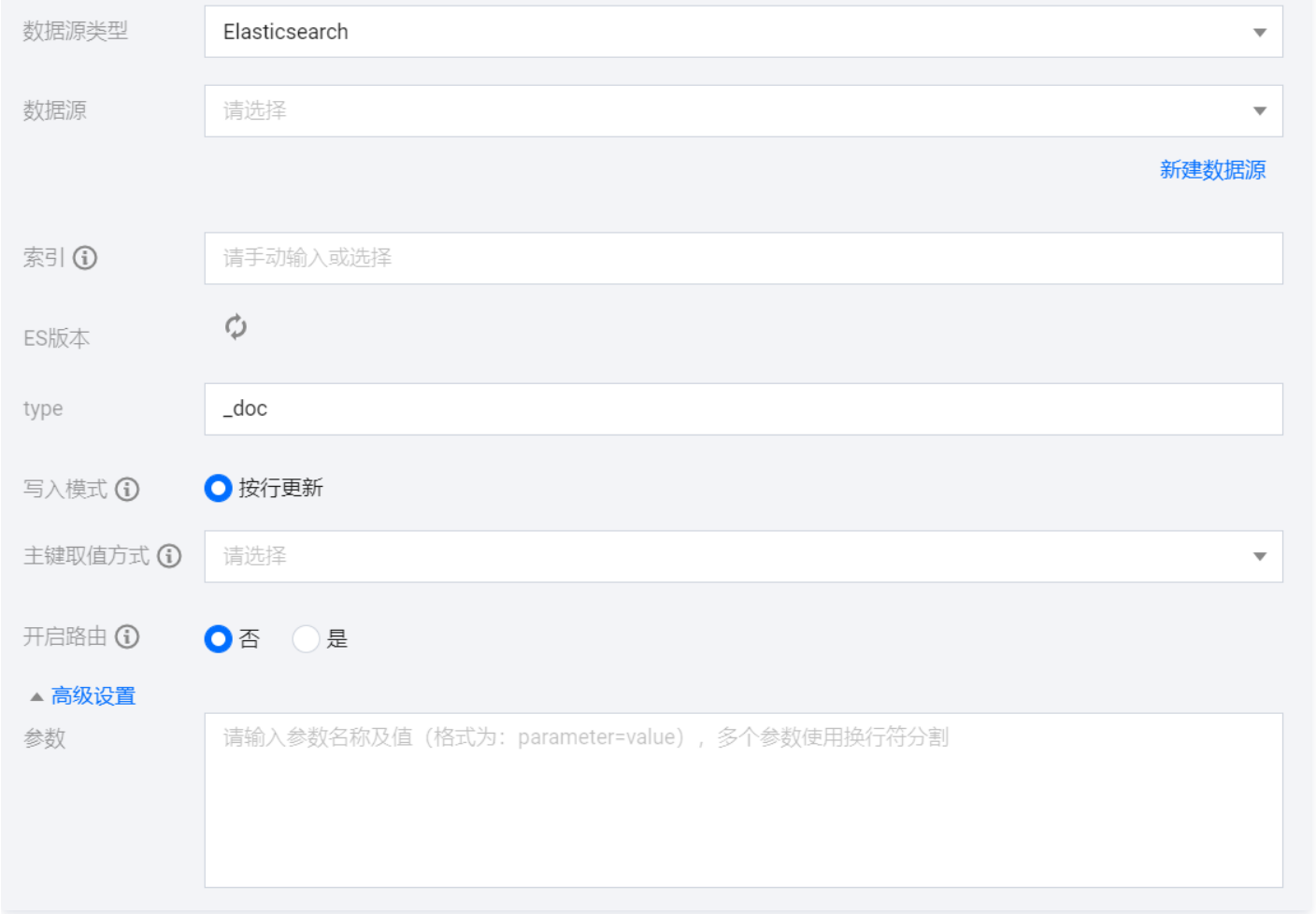
# Doris 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Doris 数据源。 |
| 库/表 | 选择该数据源中对应的库表。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# Elasticsearch 日志采集

最近更新时间：2024-07-18 17:43:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Elasticsearch 数据源。 |
| 索引 | Elasticsearch 数据源中的索引名称。 |
| type | 根据索引自动识别，7.X版本的 ElasticSearch 默认 type 为 _doc。 |
| 写入模式 | ElasticSearch 仅支持按行更像，更新每行记录所有字段。 |
| 主键取值方式 | 支持三种取值方式：<br>• 源表主键：document 的 id 使用源表的主键。<br>• 联合主键：document 的 id 使用源表的多个列共同确定。<br>• 无主键：默认生成 _id 值。 |
| 开启路由 | Elasticsearch 是否开启路由分区索引数据。开启路由功能后，可控制在 ElasticSearch 中使用哪个分区来存储文档。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# Kafka 日志采集

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Kafka 数据源，Kafka 写入端数据源类型支持 Kafka、Ckafka 。 |
| Topic | Kafka 数据源中的 Topic。 |
| 序列化格式 | Kafka 消息序列化格式类型，支持三种类型：<br>• canal-json<br>• json<br>• avro |
| 写入模式 | Kafka 支持两种写入模式：<br>• append：追加写入。<br>• upsert：以 upsert 方式插入消息，设置后消息仅只能被消息端处理一次以保证 Exactly-Once。 |
| 唯一键 | Upsert 写入模式下，需设置唯一键保证数据有序性，支持多选，Append 模式则不需要设置唯一键。 |
| 高级设置（可选） | 可根据业务需求配置参数。 |

# 离线任务
# 读取节点
# MySQL 离线读取

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 MySQL 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称<br>• 分表情况下，可在 MySQL 源端支持选择或输入多个表名称，多个表需保证结构一致。<br>• 分表情况下，支持配置表序号区间。例如 'table_[0-99]' 表示读取 'table_0'、'table_1'、'table_2' 直到 'table_99'； 如果您的表数字后缀的长度一致，例如 'table_000'、'table_001'、'table_002' 直到 'table_999'，您可以配置为 '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]' 。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 添加分库分表 | 适用于分库场景，点击后可配置多个数据源、库及表信息。分库分表场景下需保证所有表结构一致，任务配置将默认展示并使用第一个表结构进行数据获取。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步，提升数据同步效率。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create>$bizdate` 。where 条件可以有效地进行业务增量同步。<br>• 如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

| | |
|---|---|
| | • 不可以将 where 条件指定为 limit10，这不符合 MySQL WHERE 子句约束。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# TDSQL-C Mysql 离线读取

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 TDSQL-C Mysql 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称<br>• 分表情况下，可在 TDSQL-C Mysql 源端支持选择或输入多个表名称，多个表需保证结构一致。<br>• 分表情况下，支持配置表序号区间。例如 'table_[0-99]' 表示读取 'table_0'、'table_1'、'table_2' 直到 'table_99'； 如果您的表数字后缀的长度一致，例如 'table_000'、'table_001'、'table_002' 直到 'table_999'，您可以配置为 '"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]' 。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 添加分库分表 | 适用于分库场景，点击后可配置多个数据源、库及表信息。分库分表场景下需保证所有表结构一致，任务配置将默认展示并使用第一个表结构进行数据获取。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步，提升数据同步效率。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create>$bizdate` 。where 条件可以有效地进行业务增量同步。<br>• 如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。<br>• 不可以将 where 条件指定为 limit 10，这不符合 TDSQL-C Mysql WHERE 子句约束。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# PostgreSQL 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 PostgreSQL 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称<br>• 分表情况下，可在 PostgreSQL 源端支持选择或输入多个表名称，多个表需保证结构一致。<br>• 分表情况下，支持配置表序号区间。例如'table_[0-99]'表示读取'table_0'、'table_1'、'table_2'直到'table_99'；如果您的表数字后缀的长度一致，例如'table_000'、'table_001'、'table_002'直到'table_999'，您可以配置为'"table": ["table_00[0-9]", "table_0[10-99]", "table_[100-999]"]' 。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 添加分库分表 | 适用于分库场景，点击后可配置多个数据源、库及表信息。分库分表场景下需保证所有表结构一致，任务配置将默认展示并使用第一个表结构进行数据获取。 |
| 切割键 | 您可以将源数据表中某一列作为切割键，建议使用主键或有索引的列作为切割键，仅支持类型为整型的字段。读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 |
| 筛选条件（选填） | PostgreSQL 根据指定的 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如，测试时，可以将 where 条件指定实际业务场景，往往会选择当天的数据进行同步，将 where 条件指定为 id>2 and sex=1。 |

# SQL Server 离线读取

最近更新时间：2024-07-18 17:43:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 SQL Server 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称，一个任务仅支持一个表同步。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件。<br>SQL Server 根据指定的 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如在测试时，可以将 where 条件指定为 limit 10。在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create > $bizdate` 。<br>• where 条件可以有效地进行业务增量同步。<br>• where 条件为空，视作同步全表所有的信息。 |

# Oracle 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Oracle 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| Schema | 支持选择、或者手动输入需读取的 Schema。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | • 指定用于数据分片的字段，指定后将启动并发任务进行数据同步，可以提升数据同步效率。<br>• 您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键， 如果需要配置为字符串，浮点和日期等其它类型的字段，请手动输入即可。 |
| 筛选条件（选填） | Oracle 根据指定 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如，在测试时指定 where 条件为 row_number()。<br>• where 条件可以有效地进行业务增量同步 。<br>• where 条件不配置或为空时，将视作全表同步数据 。 |

# DB2 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 DB2 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。 |
| 筛选条件（选填） | DB2 根据指定的 where 条件拼接 SQL，并根据该 SQL 抽取数据。在实际业务场景中，通常会选择当天的数据进行同步，可以指定 where 条件为 gmt_create>$bizdate。where 条件可以有效地进行业务增量同步。如果该值为空，代表同步全表所有的信息。 |

# DM 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 DM 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 您可以将源数据表中某一列作为切割键，建议使用主键或有索引的列作为切割键，仅支持类型为整型的字段。<br>读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件。<br>DM 根据指定的 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如在测试时，可以将 where 条件指定为limit 10。在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 gmt_create > $bizdate。<br>• where 条件可以有效地进行业务增量同步。<br>• where 条件为空，视作同步全表所有的信息。 |

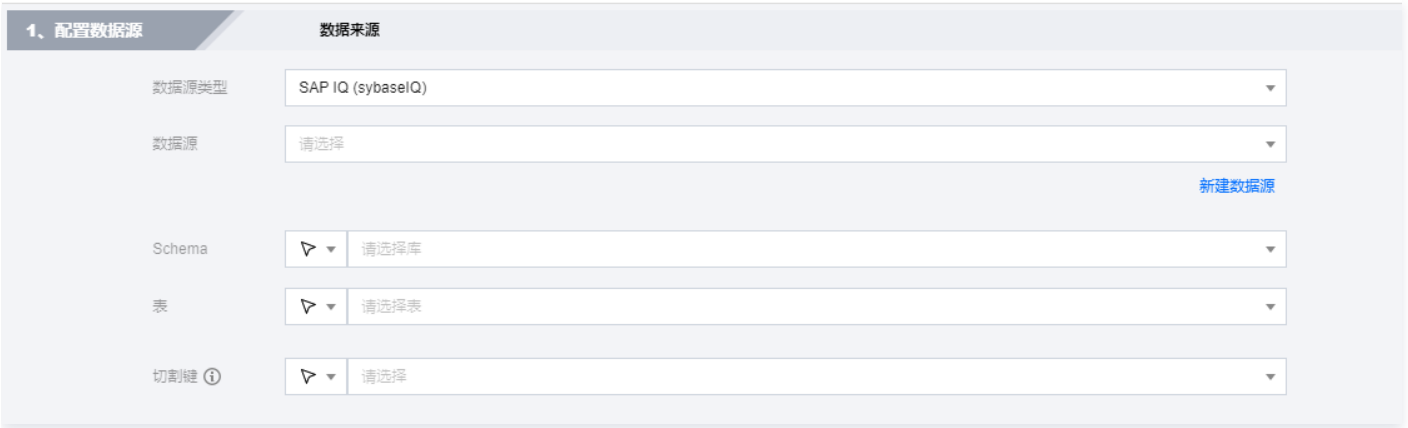
# SAP HANA 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 SAP HANA 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | • 您可以将源数据表中某一列作为切割键，建议使用主键或有索引的列作为切割键，仅支持类型为整型的字段。<br>• 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件，暂时不支持 limit 关键字过滤，SQL 语法与选择的数据源一致。 |

# SAP IQ（sybaseIQ）离线读取

最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 SAP IQ 数据源 |
| Schema | 支持选择、或者手动输入需读取的数据模式 |
| 表 | 支持选择、或者手动输入需读取的表名称 |
| 切割键 | • 您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键，仅支持类型为整型的字段<br>• 读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件，SQL 语法与选择的数据源一致 |

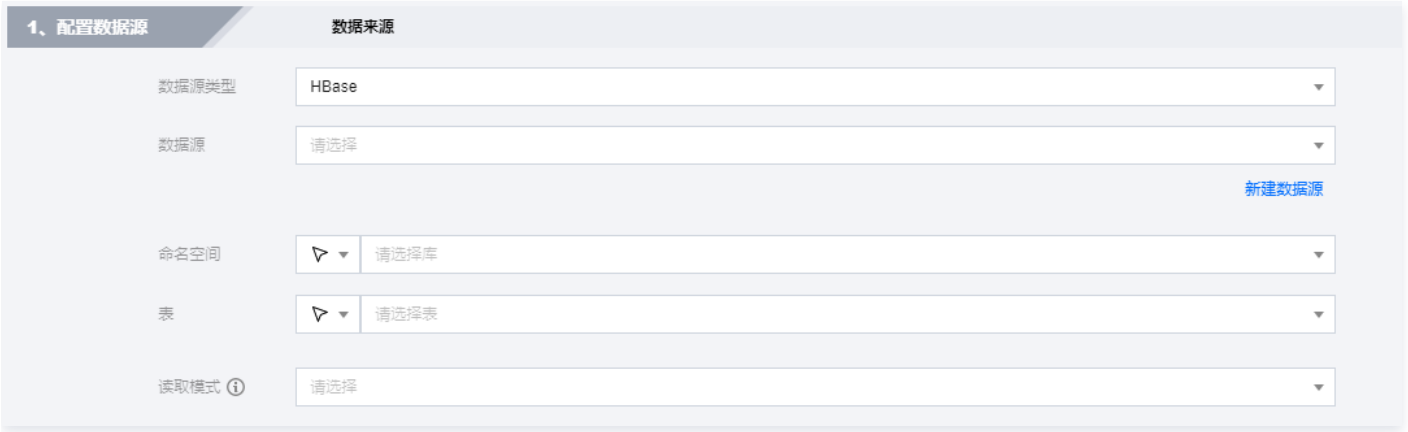
# HIVE 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 HIVE 数据源 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动数据库名称。在数据集成网络联通的情况下，仍可进行数据同步 |
| 表 | 支持选择、或者手动输入需读取的表名称 |
| 读取方式 | 仅支持 JDBC 读取方式 |
| 筛选条件（选填） | 基于 Hive JDBC 方式读取数据时，支持使用 Where 条件做数据过滤，但是此场景下，Hive 引擎底层可能会生成 MapReduce 任务，效率较慢 |

# HBase 离线读取

最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 HBase 数据源 |
| 命名空间 | 选择该数据源下可用的空间 |
| 表 | 支持选择、或者手动输入需读取的表名称 |
| 读取模式 | 仅支持横表读取模式， 将 Hbase 表当成普通二维表（横表）进行读取，读取最新版本数据 |

# Clickhouse 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Clickhouse 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | • ClickHouse 进行数据抽取时，如果指定 splitPk，表示您希望使用 splitPk 代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。<br>• 您可以将源数据表中某一列作为切割键，建议使用主键或有索引的列作为切割键，仅支持类型为整型的字段。<br>读取数据时，根据配置的字段进行数据分片，实现并发读取，可以提升数据同步效率。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 gmt_create>$bizdate。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

# DLC 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 DLC 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 gmt_create>$bizdate。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

# Kudu 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Kudu 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切割键，建议使用主键或有索引的列作为切割键。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 gmt_create>$bizdate。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |
| upperBound | 分区上限。<br>• 若 sql 建表语句中 partition “5” <= values <= “10”，则 lowerbound 为 “5” ，upperbound 为 “10”；<br>• 若 sql 建表语句中 partition value = “x”，则 lowerbound为 “x”，upperbound 为 “x\000”。 |
| lowerBound | 分区下限。<br>• 若 sql 建表语句中 partition “5” <= values <= “10”，则 lowerbound 为 “5” ，upperbound 为 “10”；<br>• 若 sql 建表语句中 partition value = “x”，则 lowerbound 为 “x”，upperbound 为 “x\000”。 |

# HDFS 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 HDFS 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用‘*’作为通配符，指定通配符后将遍历多个文件信息。例如指定/代表读取/目录下所有的文件，指定 /bazhen/ 代表读取 bazhen 目录下游所有的文件。HDFS 目前只支持*和?作为文件通配符，语法类似于通常的 Linux 命令行文件通配符。 |
| 文件类型 | HDFS 支持四种文件类型：txt 、orc 、parquet 、csv。<br>• txt：表示 TextFile 文件格式。<br>• orc：表示 ORCFile 文件格式。<br>• parquet：表示普通 Parquet 文件格式。<br>• csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 压缩格式 | 当 fileType（文件类型）为 csv 下的文件压缩方式，目前仅支持：none、deflate、gzip、bzip2、lz4、snappy。<br>• 由于 snappy 目前没有统一的 stream format，数据集成目前仅支持最主流的 hadoop-snappy（hadoop 上的 snappy stream format）和 framing-snappy（google 建议的 snappy stream format）。<br>• ORC 文件类型下无需填写。 |
| 字段分隔符 | 读取的字段分隔符，HDFS 在读取 TextFile 数据时，需要指定字段分隔符，如果不指定默认为逗号（,）。HDFS 在读取 ORC File时，您无需指定字段分隔符。<br>• 其他可用分隔符：'\t'、'\u001'、'\|'、'空格'、';'、','。<br>• 如果您想将每一行作为目的端的一列，分隔符请使用行内容不存在的字符。例如，不可见字符\u0001。 |
| 编码 | 读取文件的编码配置。支持 utf8 和 gbk 两种编码。 |

# Greenplum 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Greenplum 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create>$bizdate` 。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

# GaussDB 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 GaussDB 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create>$bizdate` 。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

# Gbase 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Gbase 数据源 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。支持同时读取多张表。当配置为多张表时，您需要保证多张表的 schema 结构一致，Gbase 不检查表的逻辑是否统一。 |
| 切割键 | Gbase 进行数据抽取时，如果指定 splitPk，表示您希望使用 splitPk 代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。<br>• 推荐 splitPk 用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。<br>• 目前 splitPk 仅支持整型数据切分，不支持字符串、浮点和日期等其它类型 。如果您指定其它非支持类型，则忽略 splitPk 功能，使用单通道进行同步。<br>• 如果设置 splitPk 值为空，底层将视作您不允许对单表进行切分，因此使用单通道进行抽取。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件。<br>Gbase 根据指定的 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如在测试时，可以将 where 条件指定为 limit 10。在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create > $bizdate` 。<br>• where 条件可以有效地进行业务增量同步。<br>• where 条件为空，视作同步全表所有的信息。 |

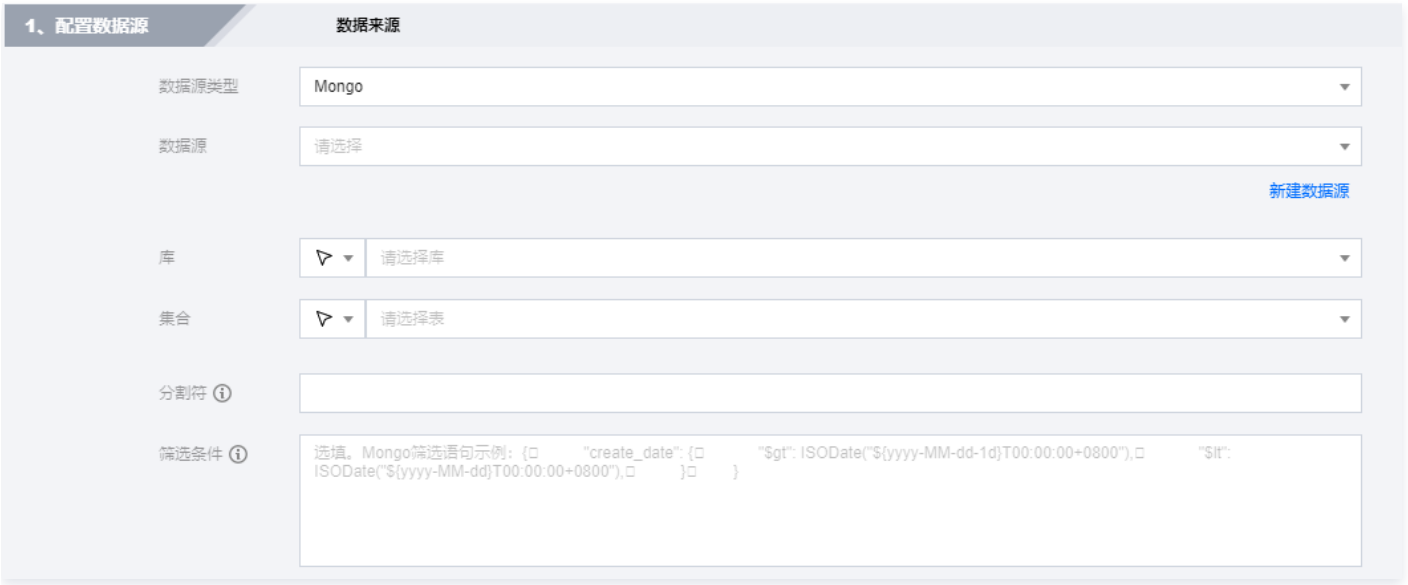
# TBase 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 TBase 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需读取该数据源下可用的模式。 |
| 表 | 支持选择、或者手动输入需读取的表名称。支持同时读取多张表。当配置为多张表时，您需要保证多张表的 schema 结构一致。 |
| 切割键 | Tbase 进行数据抽取时，如果指定 splitPk，表示您希望使用 splitPk 代表的字段进行数据分片，数据同步因此会启动并发任务进行数据同步，提高数据同步的效能。<br>• 推荐 splitPk 用户使用表主键，因为表主键通常情况下比较均匀，因此切分出来的分片也不容易出现数据热点。<br>• 目前 splitPk 仅支持整型数据切分，不支持字符串、浮点和日期等其它类型 。如果您指定其它非支持类型，则忽略 splitPk 功能，使用单通道进行同步。<br>• 如果设置 splitPk 值为空，底层将视作您不允许对单表进行切分，因此使用单通道进行抽取。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件。<br>Gbase 根据指定的 where 条件拼接 SQL，并根据该 SQL 进行数据抽取。例如在测试时，可以将 where 条件指定为 limit10。在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create > $bizdate` 。<br>• where 条件可以有效地进行业务增量同步。<br>• where 条件为空，视作同步全表所有的信息。 |

# Mongo 离线读取

最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Mongo 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 集合 | 支持选择、或者手动输入需读取的集合。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 根据数据类型填写对应筛选语句，该语句会作为将要同步数据的筛选条件。 |

# COS 离线读取

最近更新时间：2024-12-10 11:41:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 COS 数据源。 |
| 文件路径 | COS文件路径需带上桶名称，如 cosn://bucket_name。 |
| 文件类型 | COS 支持四种文件类型：txt 、orc 、parquet 、csv。<br>txt：表示 TextFile 文件格式。<br>orc：表示 ORCFile 文件格式。<br>parquet：表示普通 Parquet 文件格式。<br>csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 压缩格式 | 当 fileType（文件类型）为 csv 下的文件压缩方式，目前仅支持：none、deflate、gzip、bzip2、lz4、snappy。<br>由于 snappy 目前没有统一的 stream format，数据集成目前仅支持最主流的 hadoop-snappy（hadoop 上的 snappy stream format）和 framing-snappy（google 建议的 snappy stream format）。<br>ORC文件类型下无需填写。 |
| 字段分隔符 | 读取的字段分隔符，COS 在读取 TextFile 数据时，需要指定字段分割符，如果不指定默认为逗号（,）。COS 在读取 ORC File 时，您无需指定字段分割符。<br>其他可用分隔符：'\t'、'\u001'、'\|'、'空格'、';'、','。<br>如果您想将每一行作为目的端的一列，分隔符请使用行内容不存在的字符。例如，不可见字符 \u0001。 |
| 编码 | 读取文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 读取时，将指定字符串转为 null。 |

# FTP 离线读取

最近更新时间：2024-03-28 17:13:51


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 FTP 数据源。 |
| 同步方式 | FTP 支持两种同步方式：<br>• 数据同步：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。<br>• 文件传输：不做内容解析传输整个文件，可应用于非结构化数据同步。 |
| 文件路径 | 远程 FTP 文件系统的路径和文件名信息，需要填写包含路径和文件后缀的完整文件路径和文件名。这里可以支持填写多个路径。<br>• 当指定单个远程 FTP 文件，FTP 暂时只能使用单线程进行数据抽取。后期会在非压缩文件情况下针对单个 File 进行多线程并发读取。<br>• 当指定多个远程 FTP 文件，FTP 支持使用多线程进行数据抽取。线程并发数通过通道数指定。<br>• 当指定通配符，FTP 尝试遍历出多个文件信息。例如，指定/代表读取/目录下所有的文件，指定 /bazhen/ 代表读取 bazhen 目录下所有的文件。FTP 目前仅支持星号（*）作为文件通配符，并支持使用调度参数配合调度，灵活配置文件名与文件路径。 |
| 文件类型 | FTP 支持 text 类型，此类型适用于所有文本形式文件内容。 |
| 字段分隔符 | 读取的字段分隔符，FTP 在读取数据时，需要指定字段分隔符，如果不指定会默认为（,），界面配置也会默认填写（,）。 |
| 编码 | 读取文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 读取时，将指定字符串转为 null。 |

**关于文件路径说明：**

- 通常不建议您使用星号（*），易导致任务运行报 JVM 内存溢出的错误。
- 数据同步会将一个作业下同步的所有 Text File 视作同一张数据表。您必须自己保证所有的 File 能够适配同一套 Schema 信息。
- 您必须保证读取文件为类 CSV 格式，并且提供给数据同步系统权限可读。
- 如果 Path 指定的路径下没有符合匹配的文件抽取，同步任务将报错。

# SFTP 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 SFTP 数据源。 |
| 文件路径 | SFTP 文件系统的路径和文件名信息，需要填写包含路径和文件后缀的完整文件路径和文件名。这里可以支持填写多个路径。<br>• 当指定单个远程 SFTP 文件，SFTP 暂时只能使用单线程进行数据抽取。后期会在非压缩文件情况下针对单个 File 进行多线程并发读取。<br>• 当指定多个远程 SFTP 文件，SFTP 支持使用多线程进行数据抽取。线程并发数通过通道数指定。<br>• 当指定通配符，SFTP 尝试遍历出多个文件信息。例如，指定/代表读取/目录下所有的文件，指定 /bazhen/ 代表读取 bazhen 目录下所有的文件。SFTP 目前仅支持星号（*）作为文件通配符，并支持使用调度参数配合调度，灵活配置文件名与文件路径。 |
| 文件类型 | SFTP 支持四种文件类型：txt 、orc 、parquet 、csv。<br>• txt：表示 TextFile 文件格式。<br>• orc：表示 ORCFile 文件格式。<br>• parquet：表示普通 Parquet 文件格式。<br>• csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 字段分隔符 | 读取的字段分隔符，SFTP 在读取数据时，需要指定字段分隔符，如果不指定会默认为（,），界面配置也会默认填写（,）。 |
| 编码 | 读取文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 读取时，将指定字符串转为 null。 |

# Rest API 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Rest API 数据源。 |
| 同步方式 | 支持数据同步和文件传输两种同步方式：<br>• 数据同步：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。<br>• 文件传输：不做内容解析传输整个文件，可应用于非结构化数据同步。 |
| 请求方式 | 支持 POST 和 GET 两种请求方式 。<br>• GET 方法填入 abc=1&def=1。<br>• POST 方法填入 JSON 类型参数。 |
| 返回数据类型 | 返回数据的格式，目前仅支持 JSON 数据。 |
| JSON 数据路径（选填） | 从返回结果中查询单个 JSON 对象或者 JSON 数组的路径。 |
| 返回数据结构 | 支持单条 JSON 数据、JSON 数组数据。 |
| 请求 Header（选填） | 传递给 RESTful 接口的 Header 信息。 |
| 请求参数（选填） | • GET 方法填入abc=1&def=1。<br>• POST 方法填入 JSON 类型参数。 |
| 请求方式 | 单次运行同步任务时是否多次发起请求，多次请求需配置相应参数及起始 index 值。 |

# Elasticsearch 离线读取

最近更新时间：2024-08-15 21:18:52


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Elasticsearch 数据源 |
| 索引 | 支持多个索引名称或正则表达式。索引名称正则表达式请使用通配符(*)，如 index_* |
| ES版本 | 根据数据源和索引确定 ES 版本 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键 |
| 检索条件（选填） | 使用 JSON 格式进行检索 |

# kafka 离线读取

最近更新时间：2024-09-06 16:40:21


**参数详情：**

| 参数 | 说明 |
| --- | --- |
| 数据源 | Kafka 读取端数据源类型支持 Kafka、Ckafka |
| topic | Kafka 的 Topic，是 Kafka 处理资源的消息源（feeds of messages）的聚合 |
| 序列化格式 | 需要读取的 Kafka 数据，支持常量列、数据列和属性列：<br>• 常量列：使用单引号包裹的列为常量列，例如["'abc'", "'123'"]<br>• 数据列<br>  ○ 如果您的数据是一个 JSON，支持获取 JSON 的属性，例如["event_id"]<br>  ○ 如果您的数据是一个 JSON，支持获取 JSON 的嵌套子属性，例如["tag.desc"]<br>• 属性列<br>  ○ __key__表示消息的 key<br>  ○ __value__表示消息的完整内容<br>  ○ __partition__表示当前消息所在分区 |

| | |
|---|---|
| | ○ __headers__表示当前消息 headers 信息<br>○ __offset__表示当前消息的偏移量<br>○ __timestamp__表示当前消息的时间戳 |
| 消费组id | 避免该参数与其他消费进程重复，以保证消费位点的正确性。如果不指定该参数，默认设定 group.id=WeData_group_${任务id} |
| 周期起始位点 | 任务周期运行时，每次读取 kafka 的开始位点。默认上周期计划调度时间，可选：分区起始位点、消费组当前位点、指定位点、指定时间<br>• 指定时间：数据写入 Kafka 的时候自动生成一个 unixtime 时间戳作为该数据的时间记录。同步任务通过获取用户配置的 yyyymmddhhmmss 数值，将该值转成 unixtimestamp 后从 kafka 中读取相应数据。例如，"beginDateTime": "20210125000000"<br>• 分区起始位点：从 kafka topic 每个分区没有删除的位点最小的数据开始抽取数据<br>• 消费组当前位点：从任务配置上面指定的消费群组 ID 保存的位点开始读取数据，一般是使用这个群组 ID 读数据的进程上次停止的位点（最好确保使用这个群组 ID 的进程只有配置的这个数据集成任务，避免共用群组 ID 造成数据丢失），如果使用群组当前位点，一定要配置消费群组 ID，否则数据集成任务会随机生成一个群组ID，而新的群组 ID 因为没有保存过位点，根据位点重置策略的不同会引起任务报错或从开始或结束位点开始读取数据。另外群组位点在客户端会定时自动提交到 Kafka 服务端，所以在任务失败后，如果重跑任务时，可能有数据重复或者丢失，另外向导模式下会自动丢弃读到的超过结束位点的记录，而这些丢弃数据的群组位点已经提交到服务端，在下一个周期任务运行时将无法读到这些丢弃的数据 |
| 周期结束位点 | 任务周期运行时，每次读取 Kafka 的结束位点。默认本周期计划调度时间。当 keyType 或 valueType 配置为 STRING 时，将使用该配置项指定的编码解析字符串 |
| 位点读取模式 | 手动运行同步任务时开始同步数据的起始位点。提供两种读取模式：<br>• latest：从上次偏移位置读取<br>• earlist：从开始位点读取 |

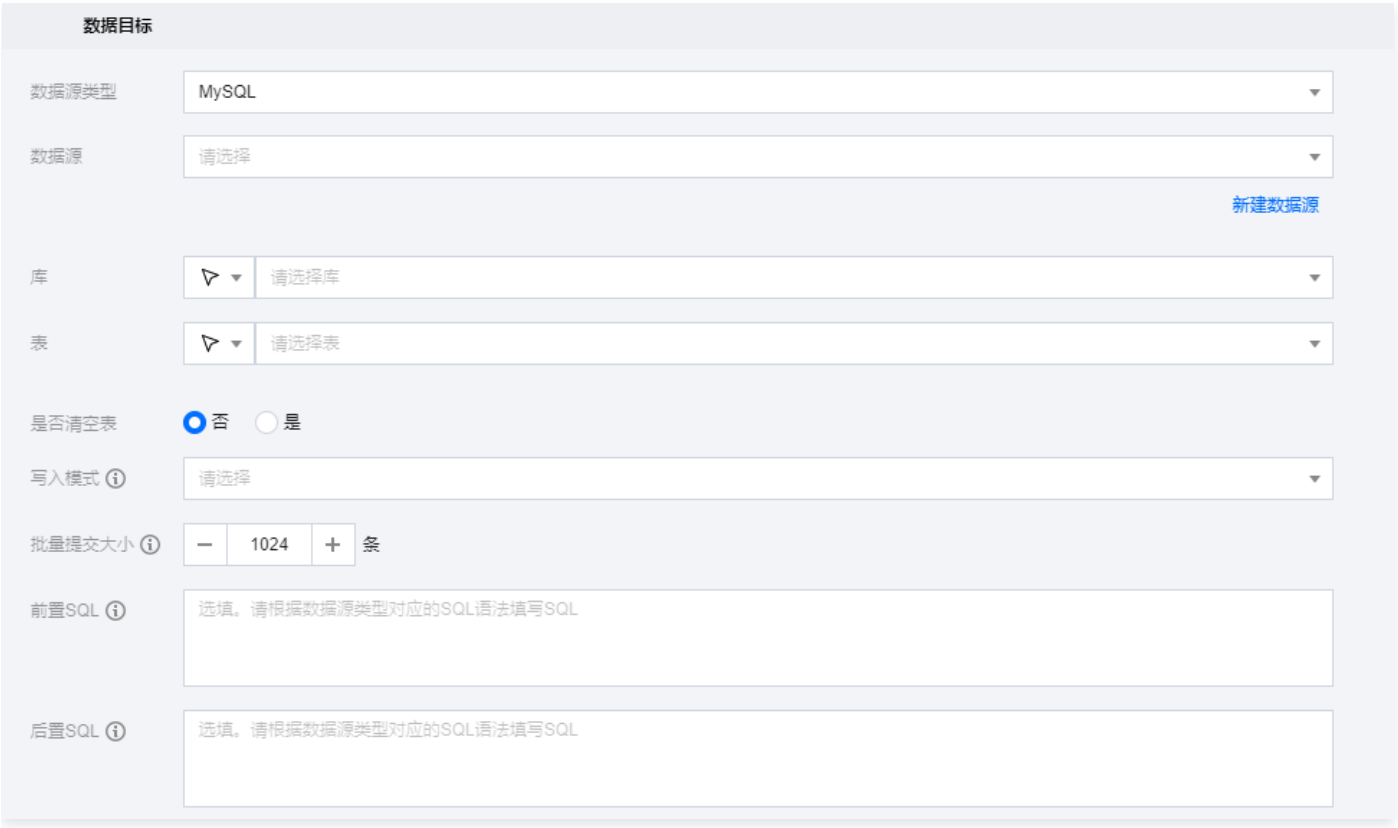
# Iceberg 离线读取

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 可用的 Iceberg 数据源。 |
| 库 | 支持选择、或者手动输入需读取的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需读取的表名称。 |
| 切割键 | 指定用于数据分片的字段，指定后将启动并发任务进行数据同步。您可以将源数据表中某一列作为切分键，建议使用主键或有索引的列作为切分键。 |
| 筛选条件（选填） | 在实际业务场景中，往往会选择当天的数据进行同步，将 where 条件指定为 `gmt_create>$bizdate` 。where 条件可以有效地进行业务增量同步。如果不填写 where 语句，包括不提供 where 的 key 或 value，数据同步均视作同步全量数据。 |

# 写入节点
# MySQL 离线写入

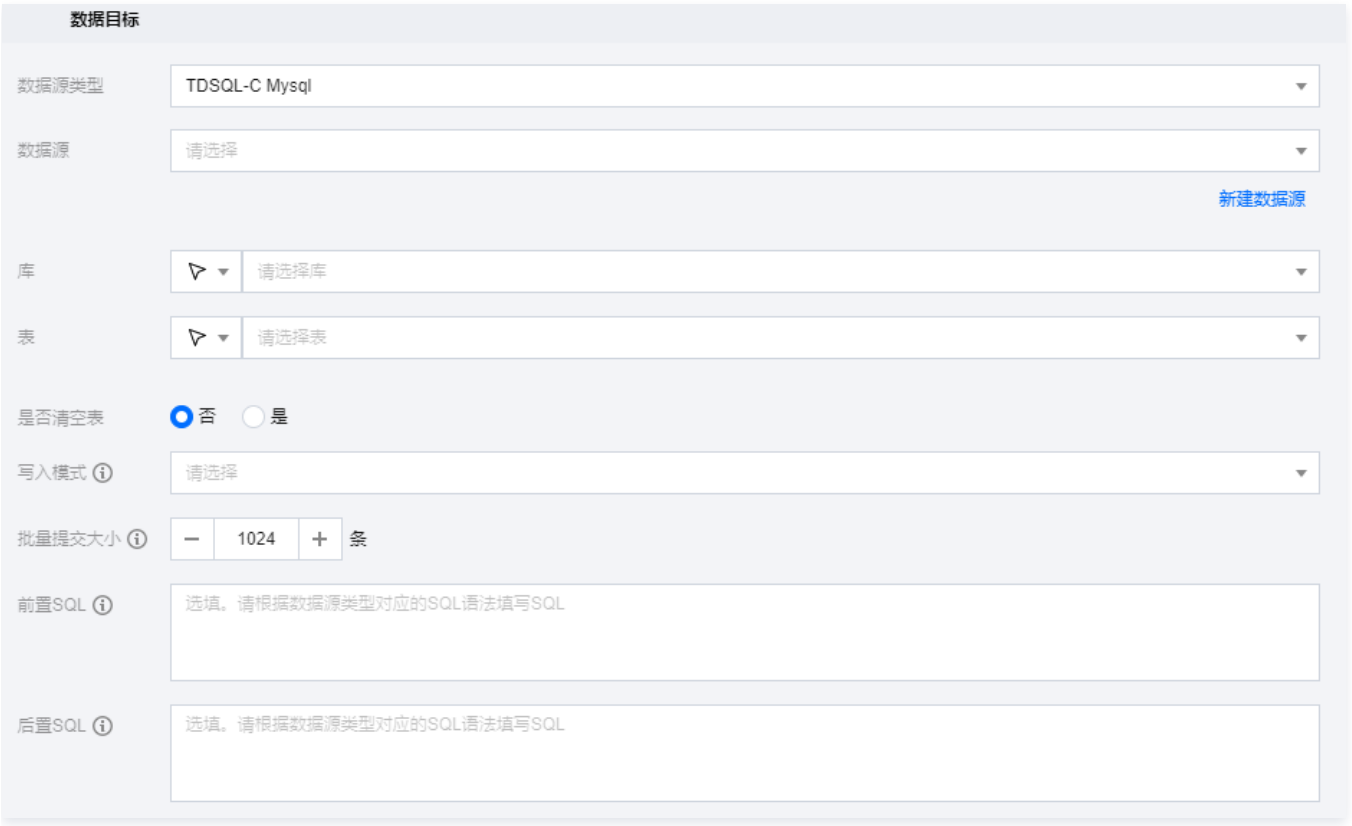
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 MySQL 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该mysql数据表前可以手动选择是否清空该数据表。 |
| 写入模式 | MySQL 写入支持三种模式：<br>• Append： 当主键/唯一性索引冲突时，冲突行无法写入 。<br>• Overwrite：主键/唯一性索引冲突时,会先删除原有行，再插入新行 。<br>• On duplicate key：主键/唯一性索引冲突时,新行会替换已指定的字段的语句。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 MySQL 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# TDSQL-C Mysql 离线写入

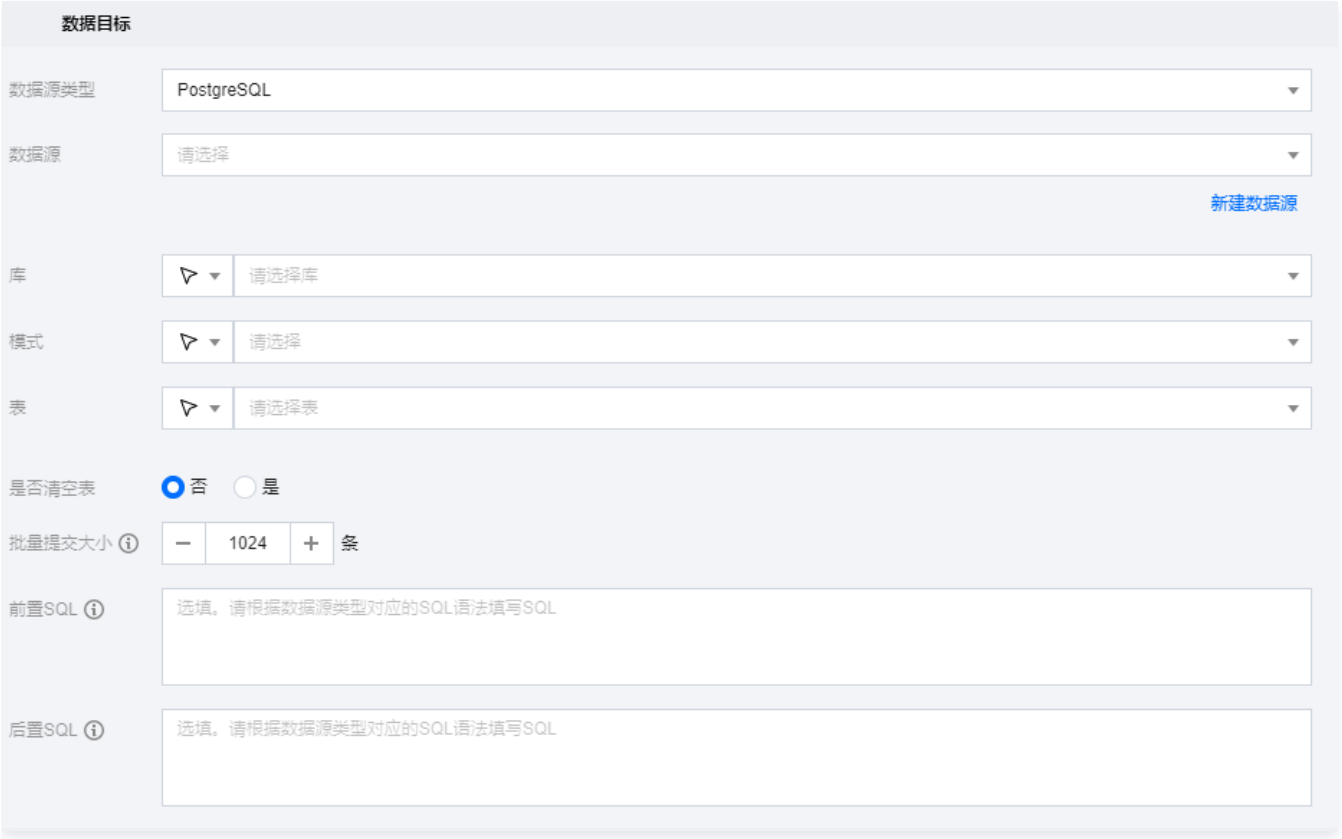
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 TDSQL-C Mysql 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 TDSQL-C Mysql 数据表前可以手动选择是否清空该数据表。 |
| 写入模式 | TDSQL-C Mysql 写入支持三种模式：<br>• Append：当主键/唯一性索引冲突时，冲突行无法写入 。<br>• Overwrite：主键/唯一性索引冲突时，会先删除原有行，再插入新行 。<br>• On duplicate key：主键/唯一性索引冲突时，新行会替换已指定的字段的语句。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 TDSQL 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# PostgreSQL 离线写入

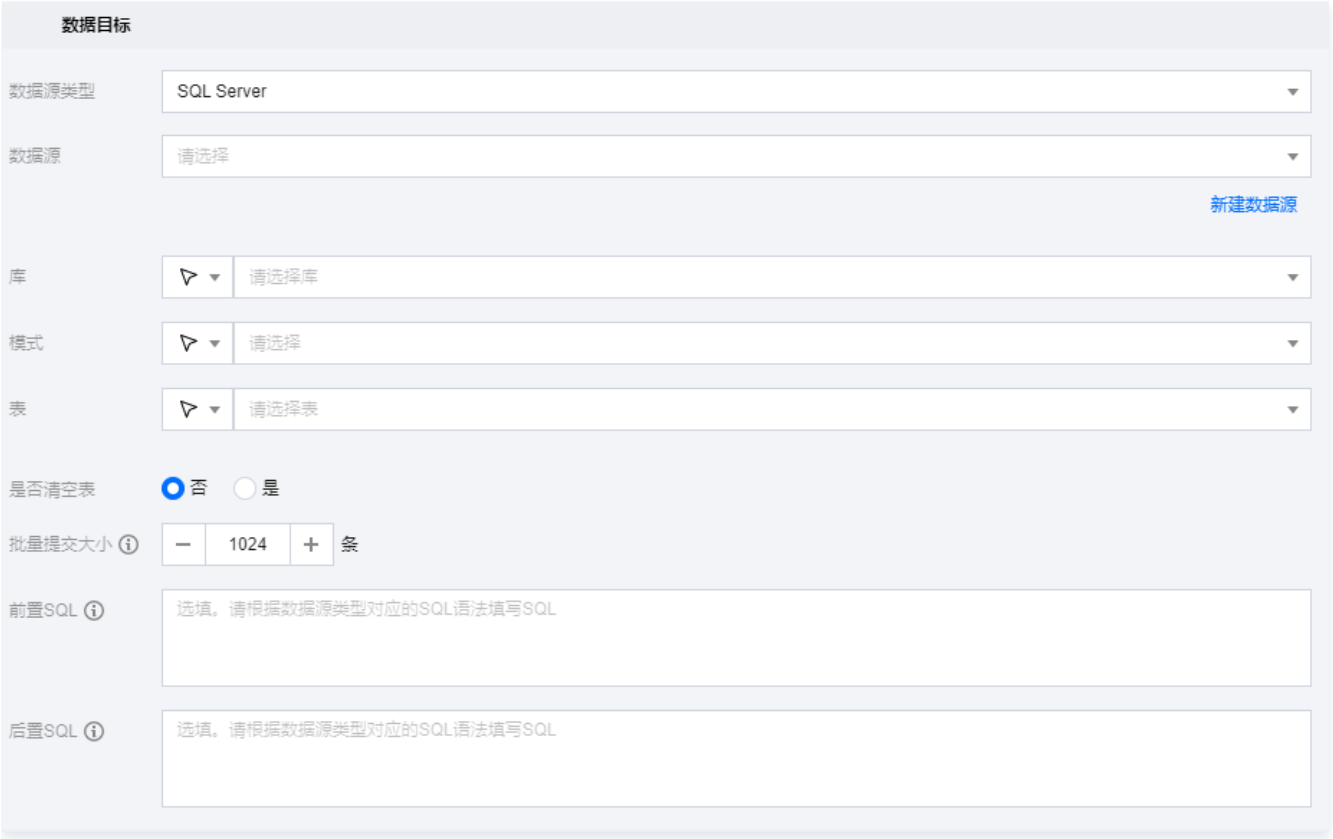
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 PostgreSQL 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 PostgreSQL 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 PostgreSQL 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 PostgreSQL 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# SQL Server 离线写入

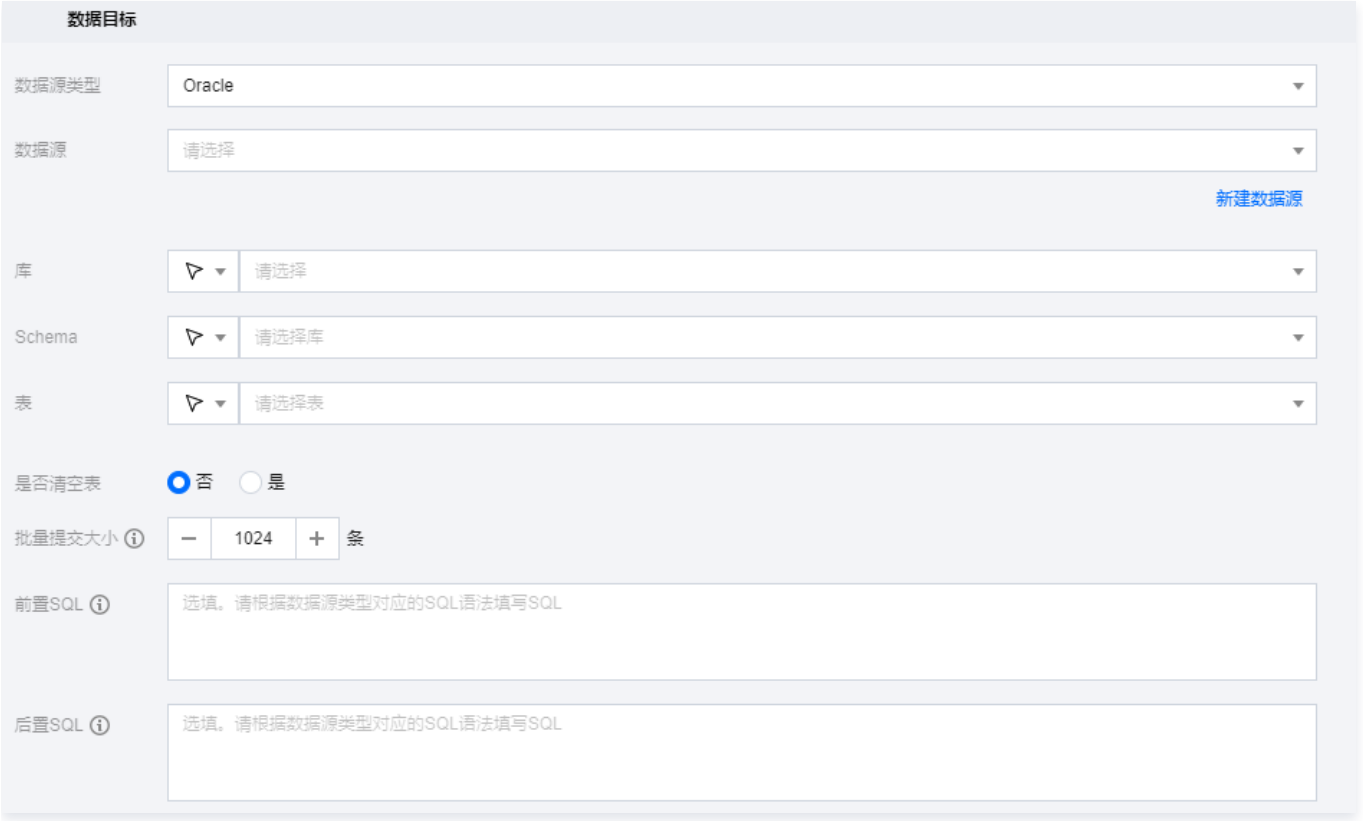
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 SQL Server 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 SQL Server 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 SQL Server 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 SQL Server 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# Oracle 离线写入

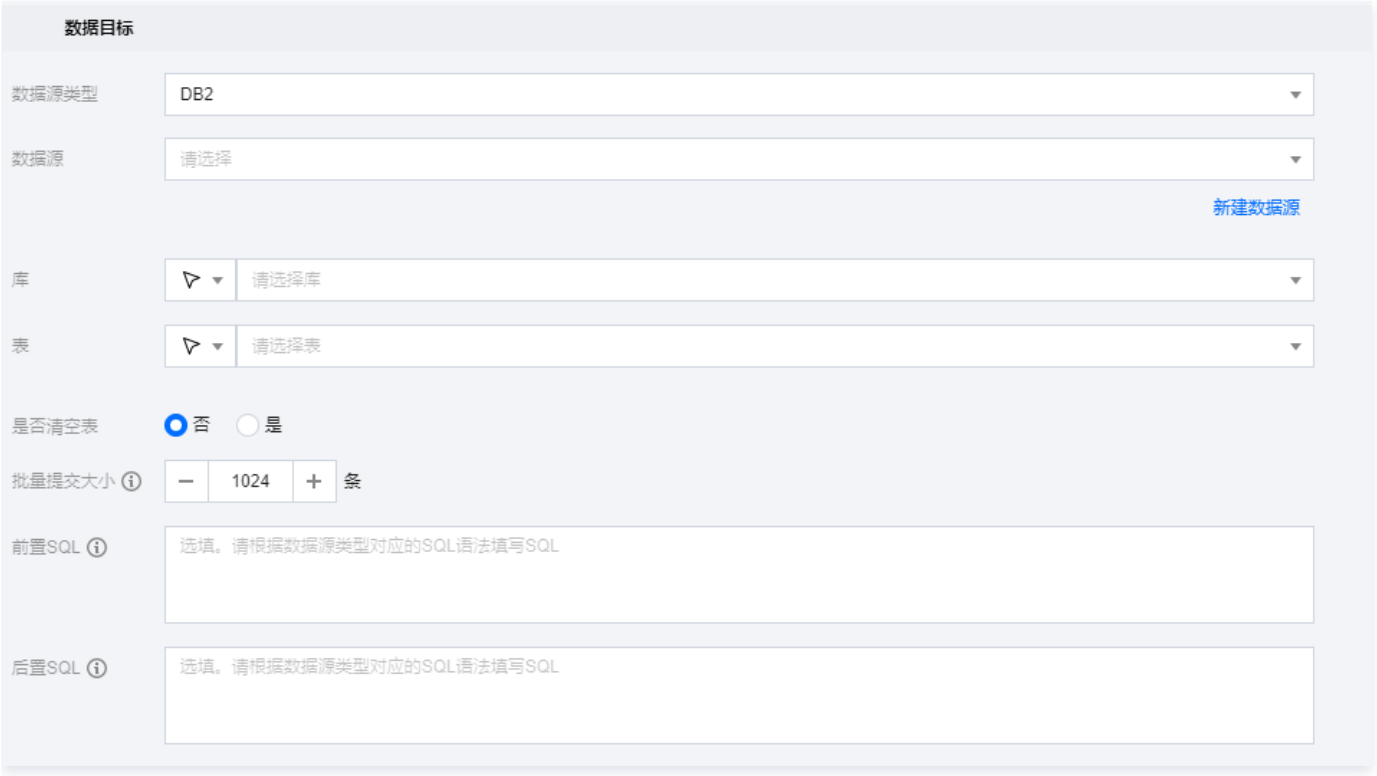
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Oracle 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| Schema | 支持选择、或者手动输入需要写入的 Oracle 数据模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 Oracle 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Oracle 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# DB2 离线写入

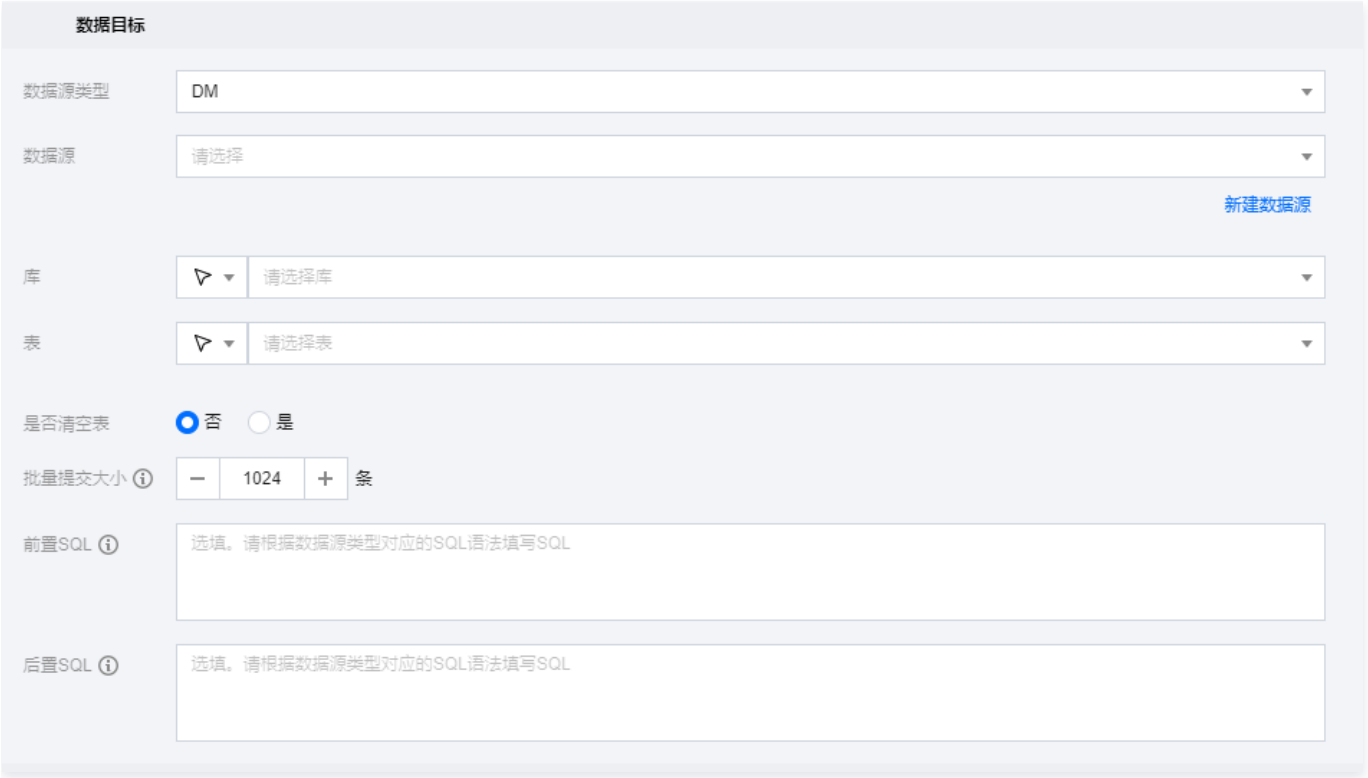
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 DB2 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 DB2 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 DB2 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# DM 离线写入

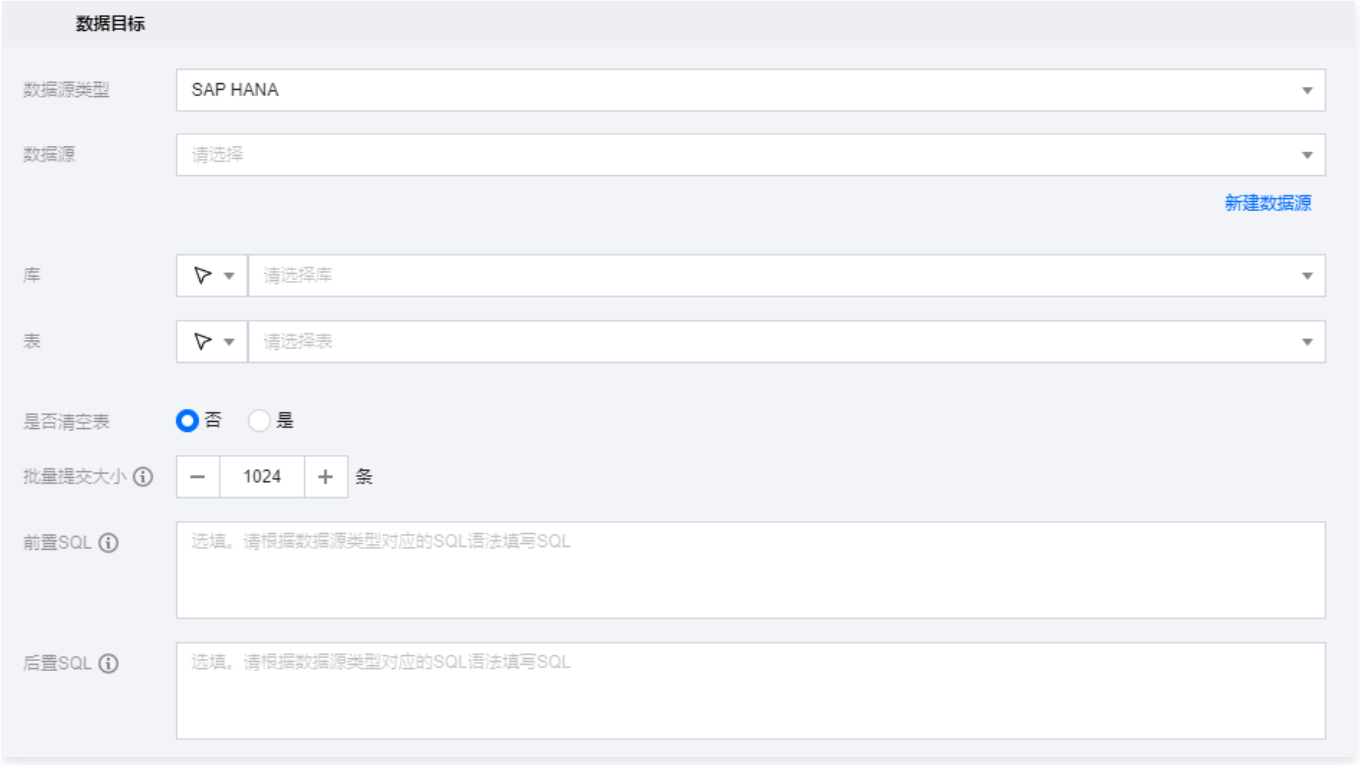
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 DM 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 DM 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 DM 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# SAP HANA 离线写入

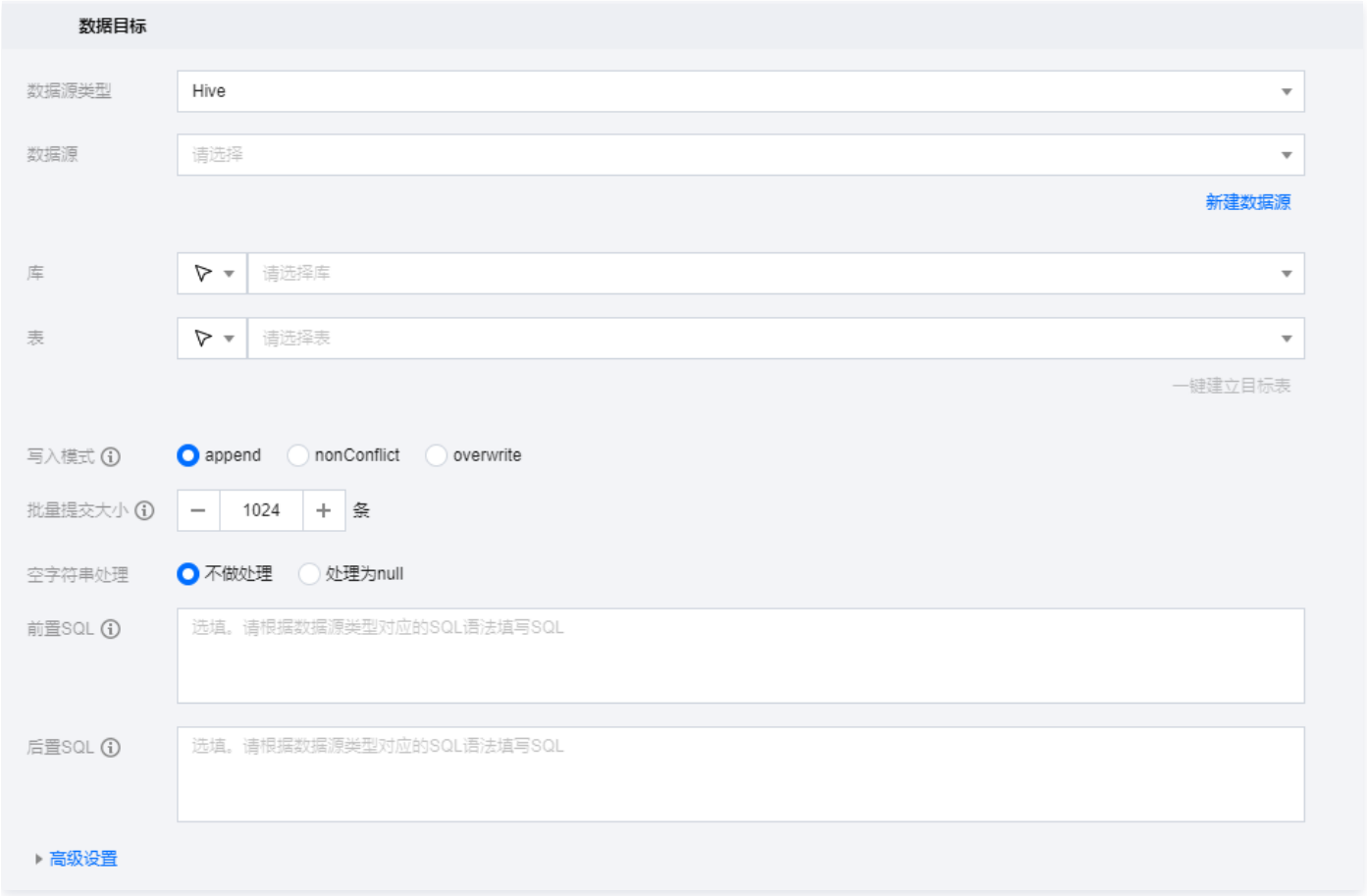
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 SAP HANA 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 SAP HANA 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 SAP HANA 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# Hive 离线写入

最近更新时间：2024-07-18 18:06:21

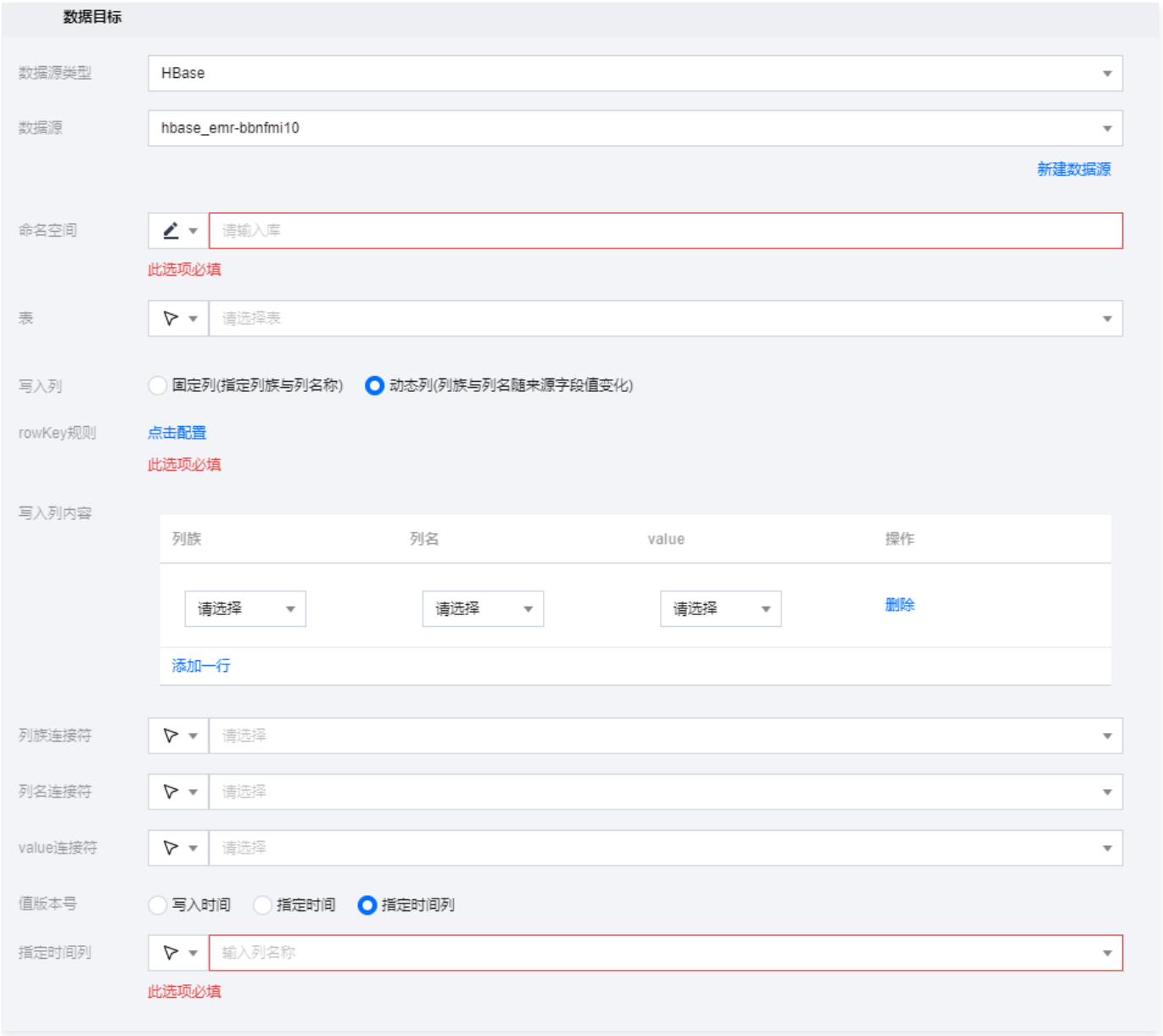



| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Hive 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称。<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 Hive 数据表前可以手动选择是否清空该数据表。 |
| 写入模式 | Hive 写入支持三种模式：<br>• Append：保留原始数据，新行追加写入<br>• nonConflict：数据冲突时报错<br>• Overwrite：删除原有数据重新写入<br>writeMode 是高危参数，请您注意数据的写出目录和写入模式，避免误删数据。加载数据行为需要配合 hiveConfig 使用，请注意您的配置。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Hive 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |

| | |
|---|---|
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# HBase 离线写入

最近更新时间：2024-08-15 21:18:52

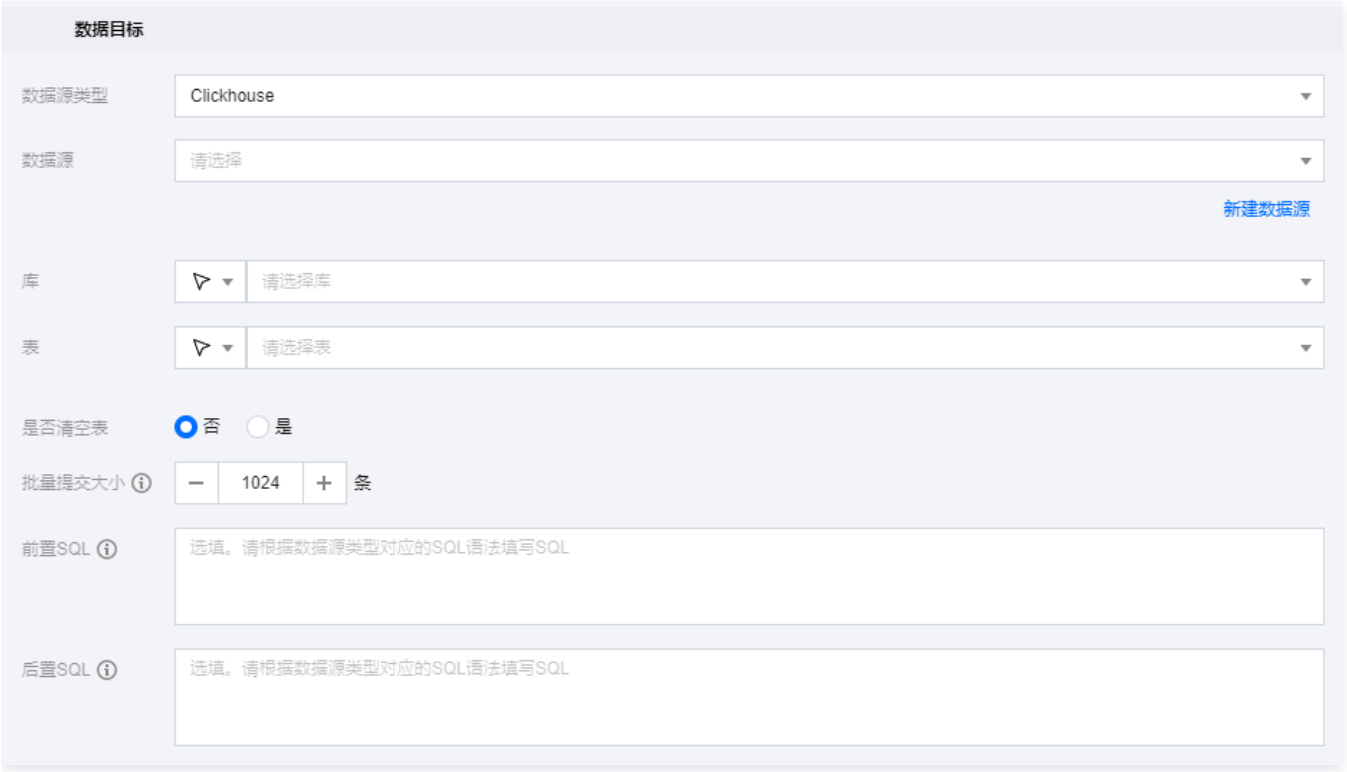



| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 HBase 数据源。 |
| 命名空间 | 支持选择、或者手动输入需写入的空间。 |
| 表 | 支持选择、或者手动输入需写入的表名称：<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 写入列 | HBase 支持两种方式写入列：<br>• 固定列（指定列族与列名称）。<br>• 动态列（列族与列名随来源字段值变化），需要手动配置列内容、列族连接符、列名连接符和 value 连接符（可选：'\u001' 、'\|' 、':' 、',' ）。 |
| 值版本号 | 指定写入 HBase 的时间戳。支持当前时间、指定时间列或指定时间（三者选一），如果不配置则表示用当前时间：<br>• index：指定对应 Reader 端 column 的索引，从0开始，需保证能转换为 LONG。 |

| | |
|---|---|
| | • type：如果是 Date 类型，会尝试用 yyyy-MM-dd HH:mm:ss 和 yyyy-MM-dd HH:mm:ssSss 解析。如果是指定时间，则 index 为-1。<br>• value：指定时间的值，LONG 类型。 |

# Clickhouse 离线写入

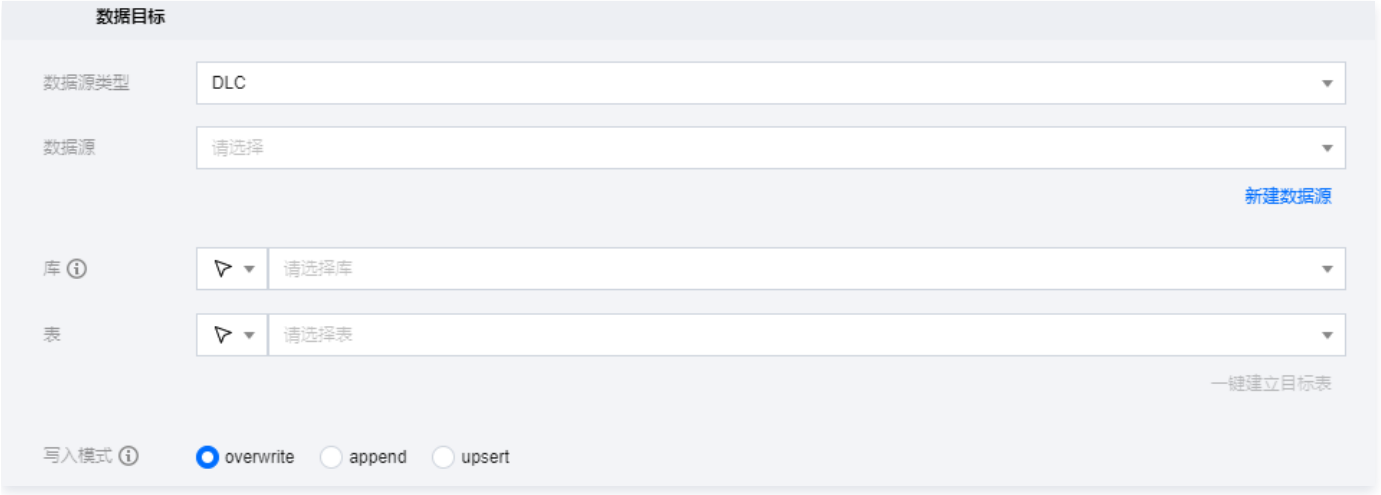
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Clickhouse 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 Clickhouse 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Clickhouse 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# DLC 离线写入

最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 DLC 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。<br>• 当来源表为 MySQL、ES、Kafka 类型，支持一键建立目标表。 |
| 写入模式 | DLC 写入支持三种模式：<br>• overwrite：覆盖原有数据写入<br>• append：追加写入，主键冲突时报错<br>• upsert：更新写入，主键冲突时更新数据 |

# Kudu 离线写入

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Kudu 数据源 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称。<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 Kudu 数据表前可以手动选择是否清空该数据表。 |
| 写入模式 | Kudu 写入支持三种模式：<br>• Append： 当主键/唯一性索引冲突时，冲突行无法写入 。<br>• Overwrite：主键/唯一性索引冲突时，会先删除原有行，再插入新行 。<br>• On duplicate key：主键/唯一性索引冲突时，新行会替换已指定的字段的语句。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Kudu 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |

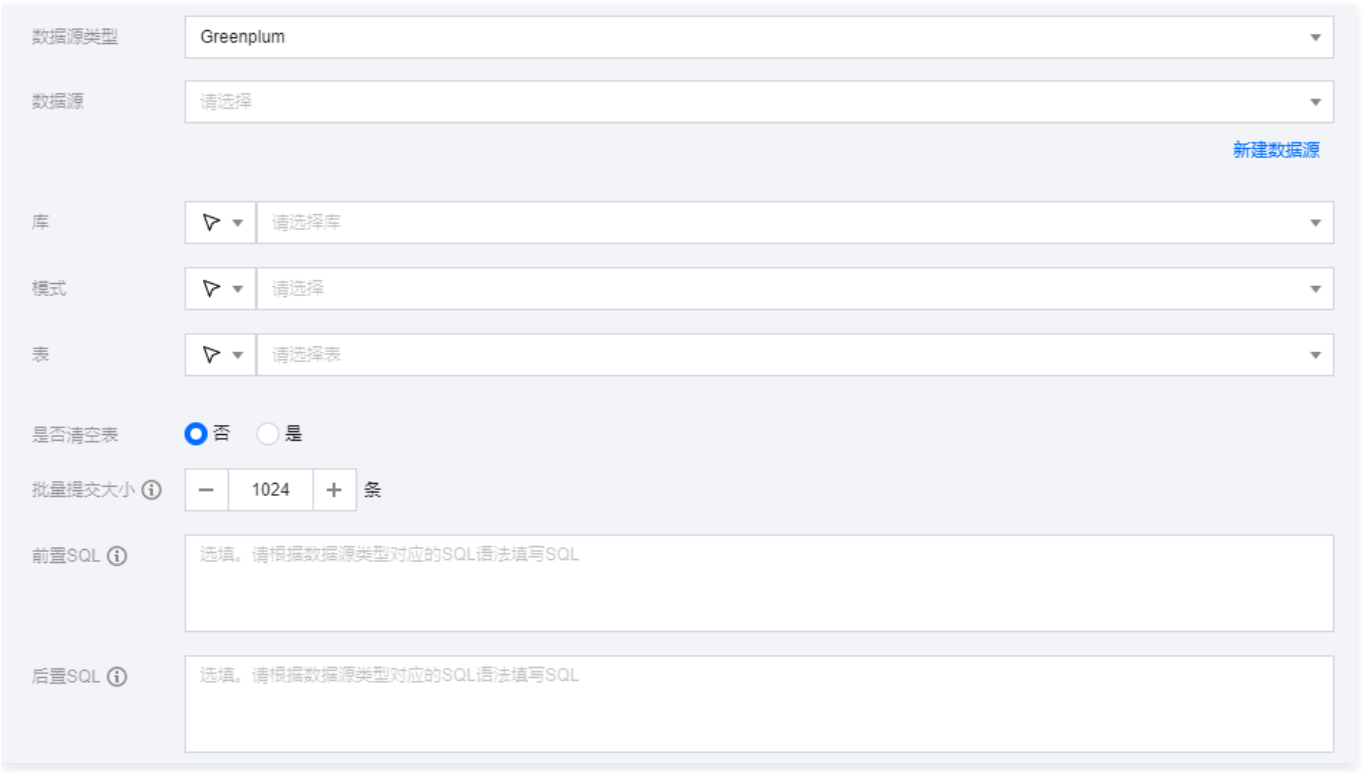
# HDFS 离线写入

最近更新时间：2024-07-18 18:06:21


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 HDFS 数据源。 |
| 同步方式 | HDFS 支持两种同步方式：<br>• 数据同步：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。<br>• 文件传输：不做内容解析传输整个文件，可应用于非结构化数据同步。 |
| 文件路径 | 文件系统的路径信息。路径支持使用‘*’作为通配符，指定通配符后将遍历多个文件信息。 |
| 写入模式 | HDFS 支持三种写入模式：<br>• append：写入前不做任何处理，直接使用 filename 写入，保证文件名不冲突<br>• nonConflict：文件名重复时报错<br>• overwrite：写入前清理以文件名为前缀的所有文件，例如，"fileName": "abc"，将清理对应目录所有 abc 开头的文件。 |
| 文件类型 | HDFS支持四种文件类型：txt 、orc 、parquet 、csv。<br>txt：表示 TextFile 文件格式。<br>orc：表示 ORCFile 文件格式。<br>parquet：表示普通 Parquet 文件格式。<br>csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 压缩格式 | 当 fileType（文件类型）为 csv 下的文件压缩方式，目前仅支持：none、deflate、gzip、bzip2、lz4、snappy。<br>由于 snappy 目前没有统一的 stream format，数据集成目前仅支持最主流的 hadoop-snappy（hadoop 上的 snappy stream format）和 framing-snappy（google 建议的 snappy stream format）。<br>ORC 文件类型下无需填写。 |
| 字段分隔符 | HDFS 写入时的字段分隔符，需要您保证与创建的 HDFS 表的字段分隔符一致，否则无法在 HDFS 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# Greenplum 离线写入

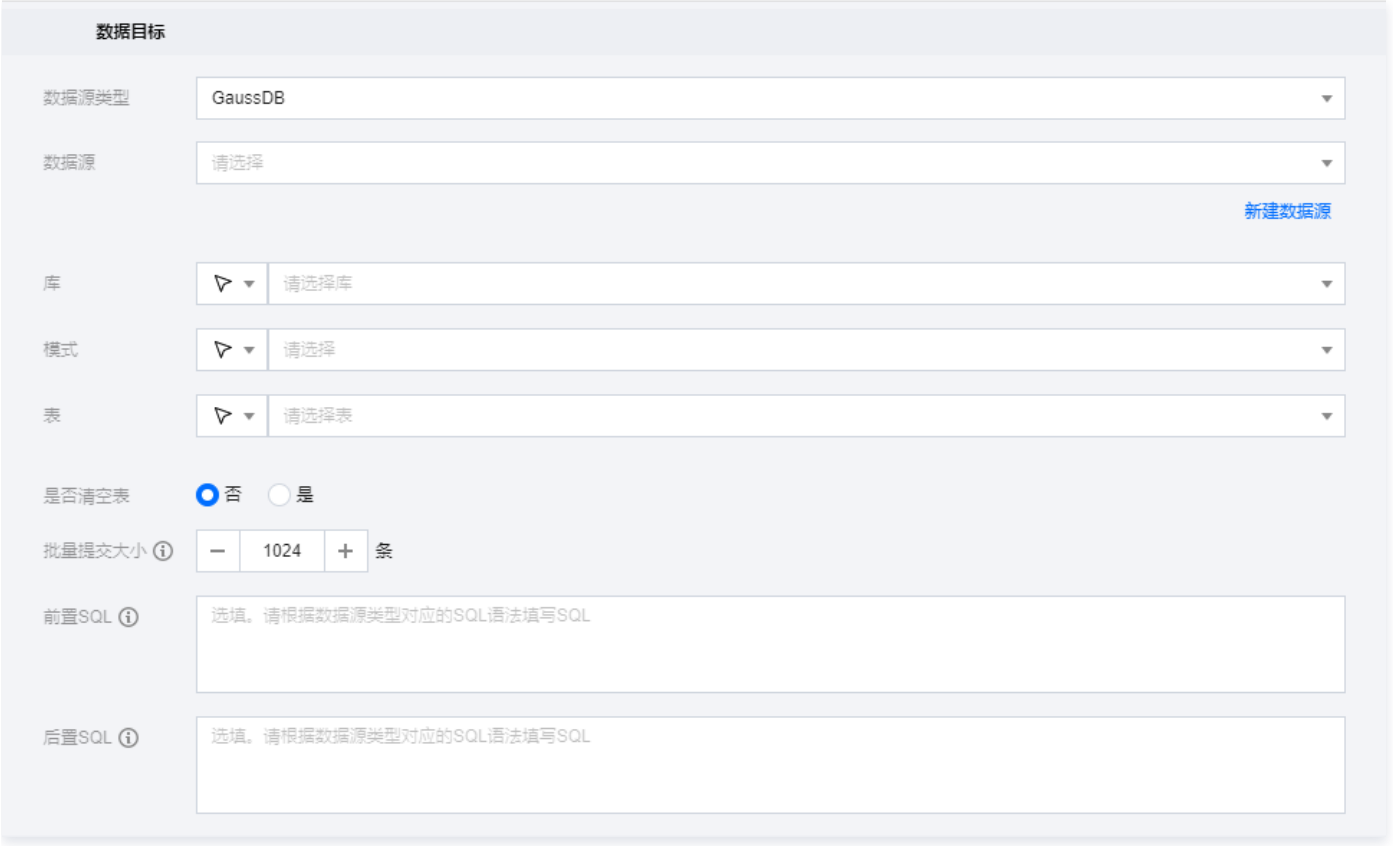
最近更新时间：2024-07-18 18:06:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Greenplum 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 Greenplum 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 Greenplum 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Greenplum 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# GaussDB 离线写入

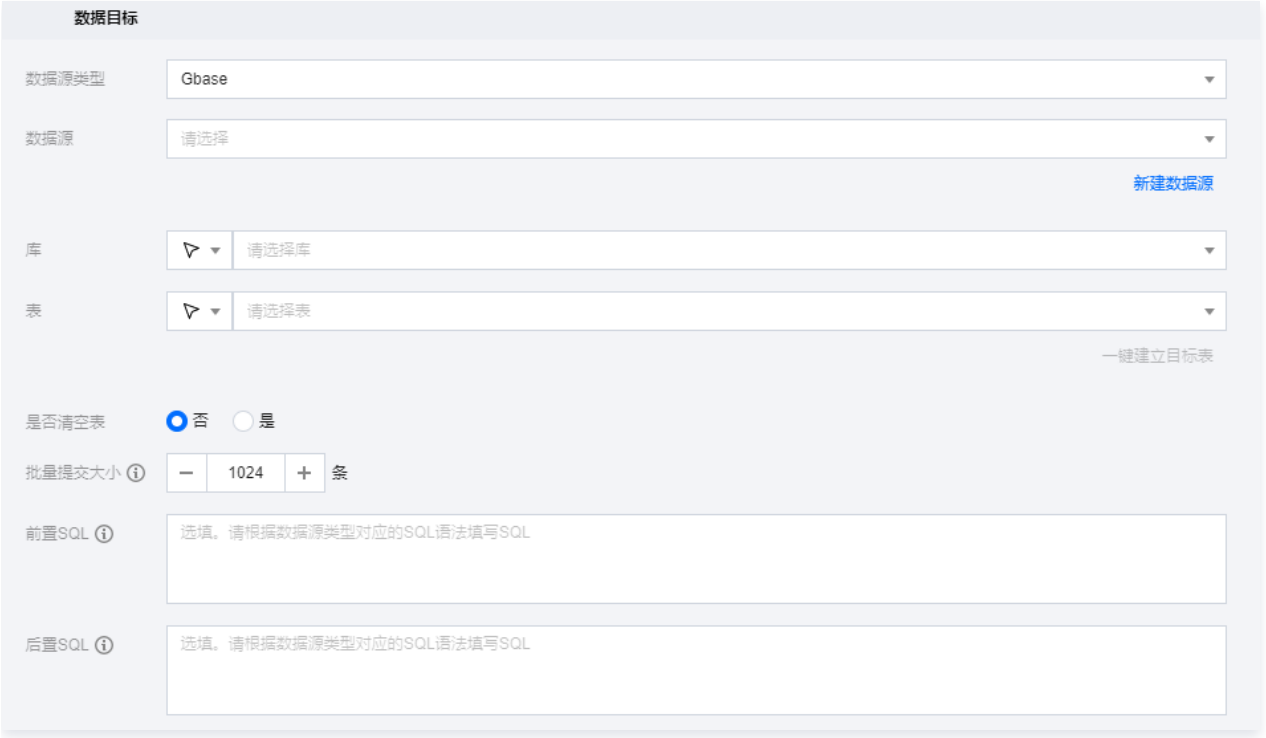
最近更新时间：2024-07-25 11:24:21




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 GaussDB 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 GaussDB 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 GaussDB 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 GaussDB 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# Gbase 离线写入

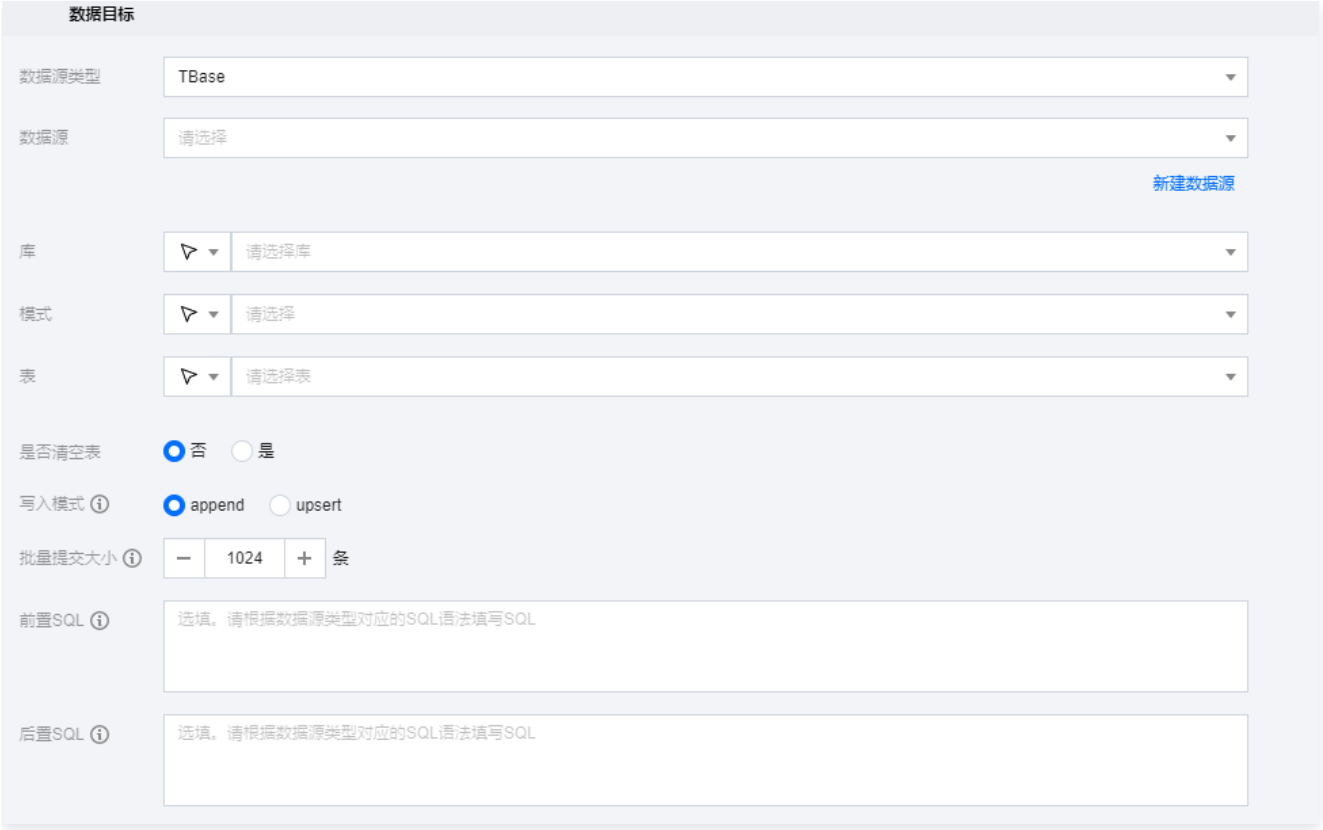
最近更新时间：2024-08-15 21:18:52




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Gbase 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。<br>• 当来源表为 Oracle 类型时，Gbase 支持一键建立目标表。 |
| 是否清空表 | 在写入该 Gbase 数据表前可以手动选择是否清空该数据表。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Gbase 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# TBase 离线写入

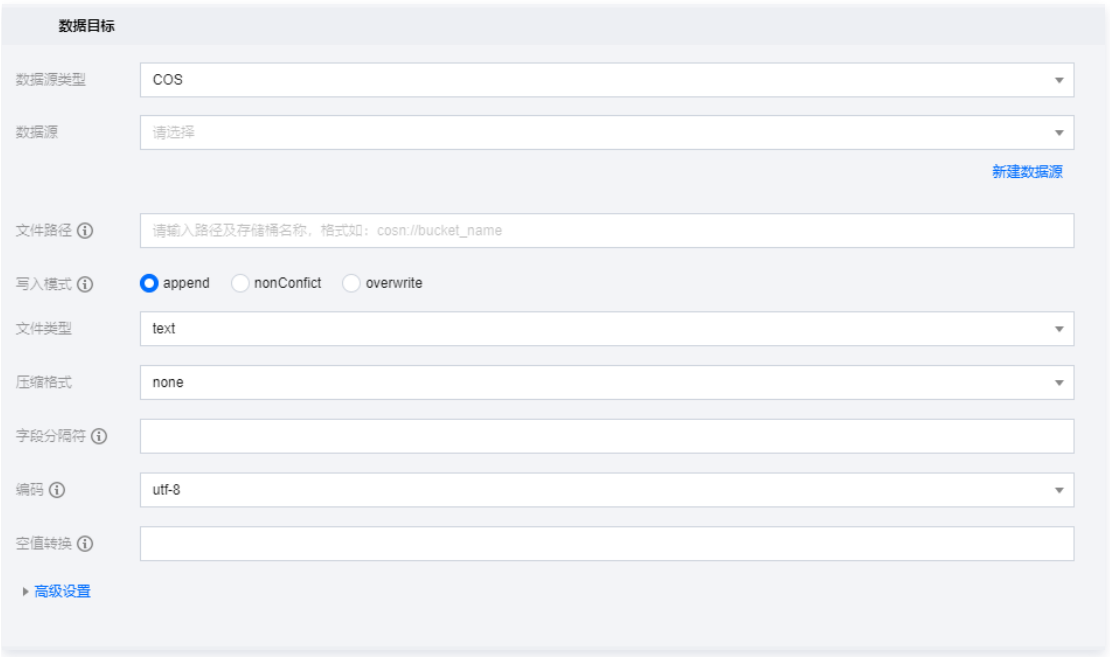
最近更新时间：2024-08-08 11:53:41




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 TBase 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 模式 | 支持选择、或者手动输入需要写入的 TBase 模式。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 是否清空表 | 在写入该 TBase 数据表前可以手动选择是否清空该数据表。 |
| 写入模式 | TBase 支持两种写入模式：<br>• append：追加写入，主键冲突时报错。<br>• upsert：更新写入，主键冲突时更新数据。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 TBase 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 前置 SQL（选填） | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL（选填） | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳 alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |

# COS 离线写入

最近更新时间：2024-06-19 10:18:01




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 COS 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用‘*’作为通配符，指定通配符后将遍历多个文件信息。 |
| 写入模式 | COS 支持三种写入模式：<br>• append：写入前不做任何处理，直接使用 filename 写入，保证文件名不冲突 。<br>• nonConflict：文件名重复时报错 。<br>• overwrite：写入前清理以文件名为前缀的所有文件，例如，"fileName": "abc"，将清理对应目录所有 abc 开头的文件。 |
| 文件类型 | COS 支持两种文件类型：txt 、csv。<br>• txt：表示 TextFile 文件格式。<br>• csv：表示普通 HDFS 文件格式（逻辑二维表）。 |
| 压缩格式 | 当 fileType（文件类型）为 csv 下的文件压缩方式，目前仅支持：none、deflate、gzip、bzip2、lz4、snappy。<br>由于 snappy 目前没有统一的 stream format，数据集成目前仅支持最主流的 hadoop-snappy（hadoop 上的 snappy stream format）和 framing-snappy（google 建议的 snappy stream format）。<br>ORC 文件类型下无需填写。 |
| 字段分隔符 | 写入的字段分隔符。COS 写入时的字段分隔符，需要您保证与创建的 Hive 表的字段分隔符一致，否则无法在 Hive 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 编码 | 写入文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 写入时，将 null 转为指定字符串。 |
| 高级设置（选填） | 可根据业务需求配置参数。 |

# FTP 离线写入

最近更新时间：2024-08-08 11:53:41


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 FTP 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用“*”作为通配符，指定通配符后将遍历多个文件信息。 |
| 写入模式 | FTP 支持三种写入模式：<br>• append：写入前不做任何处理，直接使用 filename 写入，保证文件名不冲突 。<br>• nonConflict：文件名重复时报错 。<br>• overwrite：写入前清理以文件名为前缀的所有文件。 |
| 字段分隔符 | 写入的字段分隔符。FTP 写入时的字段分隔符，需要您保证与创建的 FTP 表的字段分隔符一致，否则无法在 FTP 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 编码 | 写入文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 写入时，将 null 转为指定字符串。 |

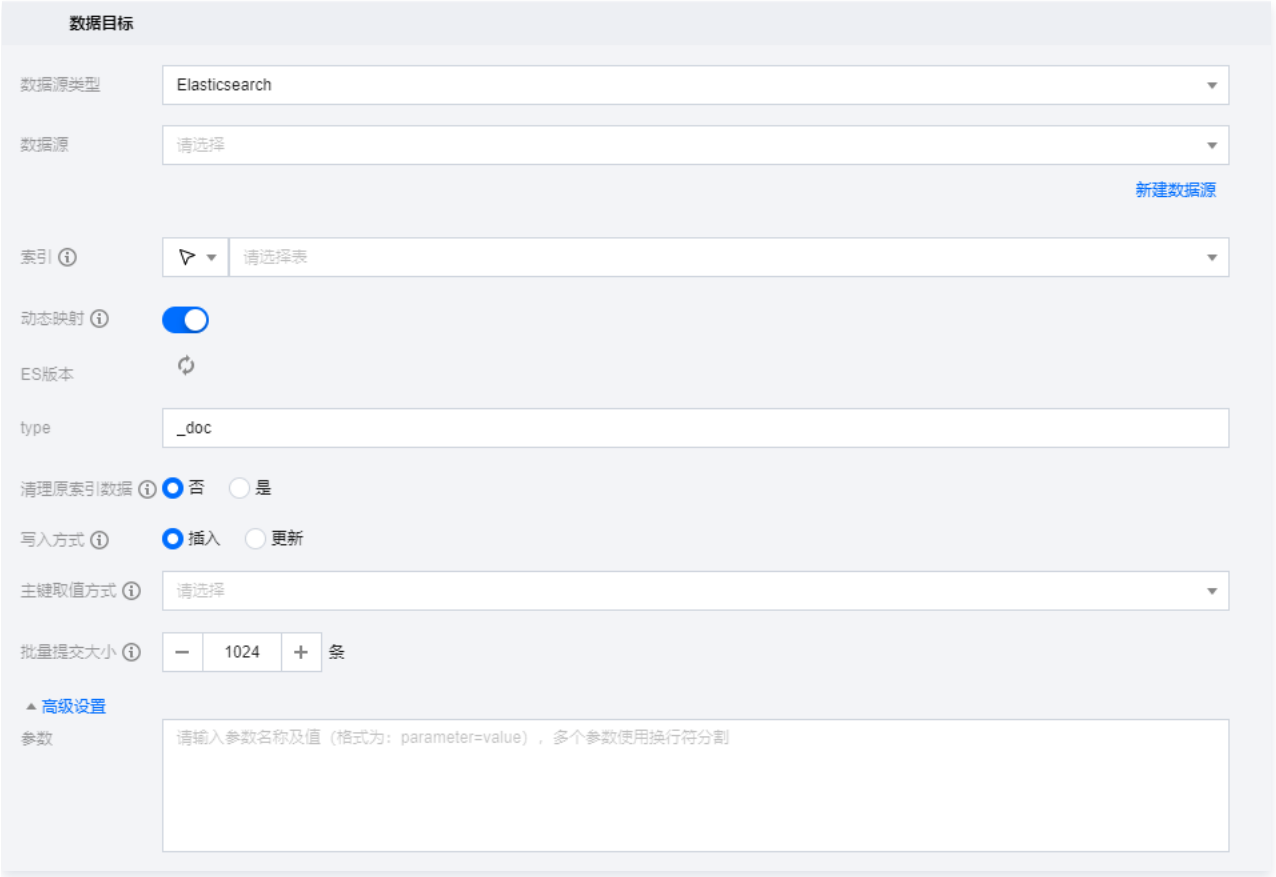
# SFTP 离线写入

最近更新时间：2024-06-18 14:47:41


| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 SFTP 数据源。 |
| 文件路径 | 文件系统的路径信息。路径支持使用“*”作为通配符，指定通配符后将遍历多个文件信息。 |
| 写入模式 | SFTP 支持三种写入模式：<br>• append：写入前不做任何处理，保证文件名不冲突 。<br>• nonConflict：文件名重复时报错 。<br>• overwrite：写入前清理以文件名为前缀的所有文件。 |
| 字段分隔符 | 写入的字段分隔符。SFTP 写入时的字段分隔符，需要您保证与创建的 SFTP 表的字段分隔符一致，否则无法在 SFTP 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 编码 | 写入文件的编码配置。支持 utf8 和 gbk 两种编码。 |
| 空值转换 | 写入时，将 null 转为指定字符串。 |

# Elasticsearch 离线写入

最近更新时间：2024-08-08 11:53:41

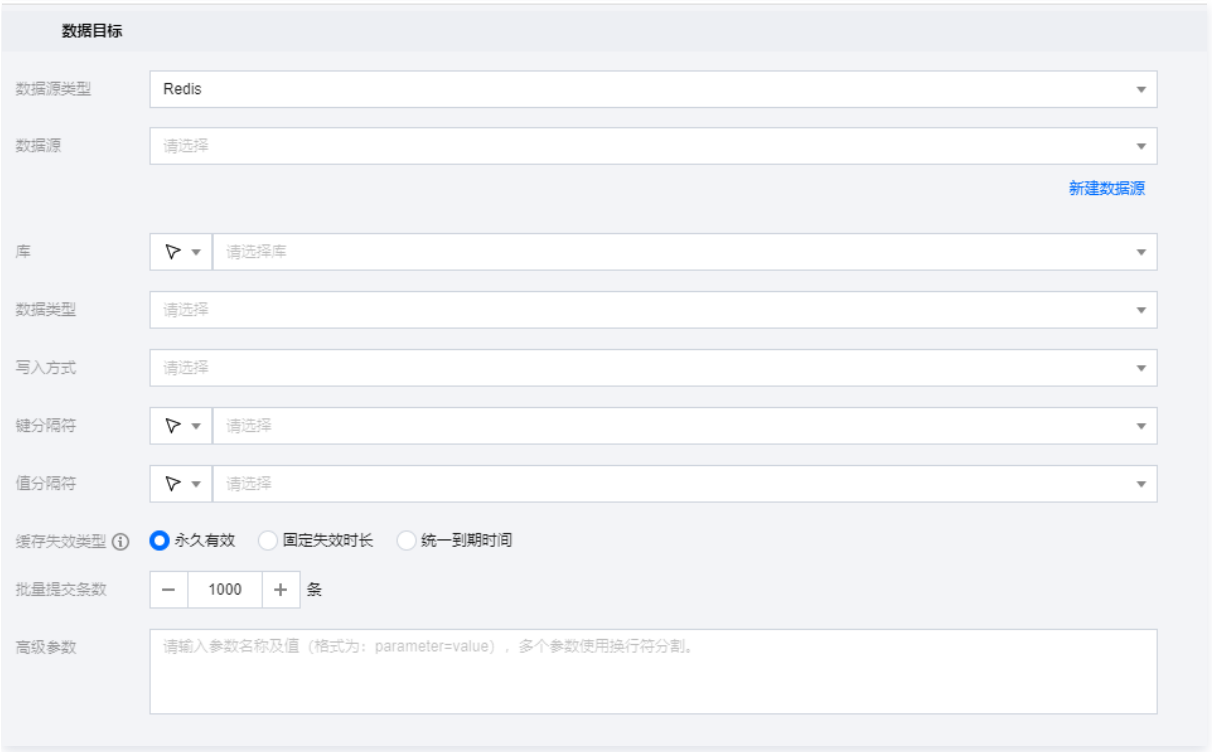



| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Elasticsearch 数据源。 |
| 索引 | ElasticSearch 中的索引名称。 |
| 动态映射 | 定义当在文档中发现未存在的字段时，同步任务是否通过 Elasticsearch 动态映射机制为字段添加映射。<br>• 打开：保留 Elasticsearch 的自动 mappings 映射。<br>• 关闭：默认关闭，根据同步任务配置的 column 生成并更新 Elasticsearch 的 mappings 映射。<br>• Elasticsearch 7.x 版本的默认 type 为_doc。使用 Elasticsearch 的自动 mappings 时，请配置 _doc 和 esVersion 为7。 |
| 清理原索引数据 | 手动选择是否清理原索引数据：<br>• 否：导入数据前保留索引中已存在的数据。<br>• 是：导入数据前删除原来的索引并重建同名索引，此操作会删除该索引下的数据。 |
| 写入方式 | 支持插入和更新两种写入方式：<br>• 插入：所有数据直接插入。<br>• 更新：存在相同主键时更新数据，否则插入。 |
| 主键取值方式 | 支持三种取值方式：<br>• 源表主键: document 的 id 使用源表的主键。<br>• 联合主键: document 的 id 使用源表的多个列共同确定。<br>• 无主键: 默认生成_id 值。 |
| 批量提交大小 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 ElasticSearch 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |

| | |
|---|---|
| 高级设置（选填） | 可根据业务需求配置参数。 |

# Redis 离线写入

最近更新时间：2024-06-18 14:47:41




| 参数 | 说明 |
| --- | --- |
| 数据源 | 选择当前项目中可用的 Redis 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 数据类型 | Redis 写入 Redis 的 value 类型包含以下5种：<br>• 字符串（string）<br>• 字符串列表（list）<br>• 字符串集合（set）<br>• 有序字符串集合（zset）<br>• 哈希（hash）<br>不同的 value 类型，数据类型配置会略有差异。 |
| 写入方式 | 根据写入类型自动填入对应的写入方式。 |
| 键分隔符 | Redis 写入的键分隔符，需要您保证与创建的 Redis 表的字段分隔符一致，否则无法在 Redis 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 值分隔符 | Redis 写入的值分隔符，需要您保证与创建的 Redis 表的值分隔符一致，否则无法在 Redis 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 缓存失效类型 | 支持三种失效类型：<br>• 永久有效：key 值不设定失效时间，永久有效。<br>• 固定失效时长：数据以分批实际写入时间为起点，经过固定设置时间长度后失效。<br>• 统一到期时间：所有写入数据均在指定时间一起失效。 |
| 批量提交条数 | 一次性批量提交的记录数大小，该值可以极大减少数据同步系统与 Redis 的网络交互次数，并提升整体吞吐量。如果该值设置过大，会导致数据同步运行进程 OOM 异常。 |
| 高级参数 | 可根据业务需求配置参数。 |

**数据类型详解**

| value 类型 | type 参数（必选） | mode 参数（必选） | valueFieldDelimiter 参数（非必选） | writeMode 配置样例 |
| --- | --- | --- | --- | --- |
| 字符串（string） | type 需配置为 string。 | mode 为写入模式参数，value 为字符串（string）时：mode 需配置为set。如果需存储的数据已经存在，则覆盖原有的数据。 | valueFieldDelimiter 为 value 之间的分隔符，默认值为 \u0001。该配置项主要用于源数据每行超过两列的情况，例如有三列时，各列通过分隔符分割样例为 value1\u0001value2\u0001value3。如果源数据只有两列（即 key 和 value）时，则无需配置。 | `"writeMode":{ "type": "string", "mode": "set", "valueFieldDelimiter": "\u0001" }` |
| 字符串列表（list） | type 需配置为 list。 | mode 为写入模式参数，value 为字符串列表（list）时，可配置为：<br>• lpush，表示在 list 最左边存储数据。<br>• rpush，表示在 list 最右边存储数据。 | | `"writeMode":{ "type": "list", "mode": "lpush\|rpush", "valueFieldDelimiter": "\u0001" }` |
| 字符串集合（set） | type 需配置为 set。 | mode 为写入模式参数，value 为字符串集合（set）时：mode 需配置为 sadd，表示向 set 集合中存储数据。如果需存储的数据已经存在，则覆盖原有的数据。 | | `"writeMode":{ "type": "set", "mode": "sadd", "valueFieldDelimiter": "\u0001" }` |
| 有序字符串集合（zset） | type 需配置为 zset。 | mode 为写入模式参数，value 为有序字符串集合（zset）时：mode 需配置为 zadd，表示向 zset 有序集合中存储数据。如果需存储的数据已经存在，则覆盖原有的数据。 | 无需配置此参数。 | `"writeMode":{ "type": "zset", "mode": "zadd"` |

| | | | | |
|---|---|---|---|---|
| | | | | <pre>}</pre><br>• 当 value 类型为zset时，数据源的每行记录均需遵循相应的规范。即每行记录除 key 外，只能有1对 score 和 value， 并且 score 必须在 value 前面，Redis Writer 方可解析出 column 对应的是 score 或 value。 |
| 哈希（hash） | type 需配置为 hash。 | mode 为写入模式参数，value 为哈希（hash）时：mode 需配置为 hset，表示向 hash 有序集合中存储数据。如果需存储的数据已经存在，则覆盖原有的数据。 | 无需配置此参数。 | <pre>"writeMode":{<br>"type": "hash"<br>,<br>"mode": "hset"<br>}</pre><br>• 当 value 类型为 hash 时，数据源的每行记录都需遵循相应的规范。即每行记录除 key 外，只能有1对 attribute 和 value，并且 attribute 必须在 value 前面，Redis Writer 方可解析出 column 对应的是 attribute 或 value。 |

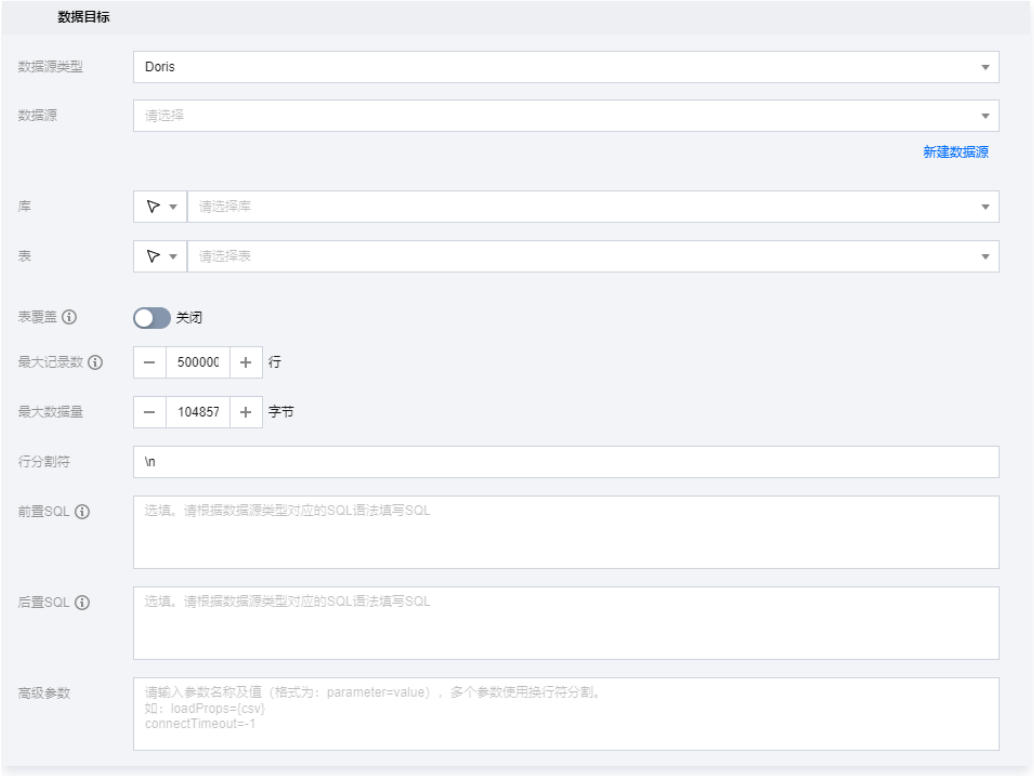
# Iceberg 离线写入

最近更新时间：2024-06-18 14:47:41


| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Iceberg 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 写入模式 | Iceberg 写入支持三种模式：<br>• overwrite：覆盖写入。<br>• append：追加写入。<br>• upsert：根据设置主键字段进行数据更新写入。 |

# Doris 离线写入

最近更新时间：2024-06-18 14:47:41




| 参数 | 说明 |
| --- | --- |
| 数据源 | 需要写入的 Doris 数据源。 |
| 库 | 支持选择、或者手动输入需写入的库名称<br>• 默认将数据源绑定的数据库作为默认库，其他数据库需手动输入库名称。<br>• 当数据源网络不联通导致无法直接拉取库信息时，可手动输入数据库名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表 | 支持选择、或者手动输入需写入的表名称<br>• 当数据源网络不联通导致无法直接拉取表信息时，可手动输入表名称。在数据集成网络连通的情况下，仍可进行数据同步。 |
| 表覆盖 | 开启后，Doris 将支持表级别的原子覆盖写操作。写入数据前会先使用 CREATE TABLE LIKE 语句创建一个相同结构的新表，将新的数据导入到新表后，通过 swap 方式原子的替换旧表，以达到表覆盖目的。 |
| 最大记录数 | 一次性批量提交的记录数大小。 |
| 最大数据量 | 一次性批量提交的最大数据量。 |
| 行分隔符（选填） | Doris 写入的键分隔符，需要您保证与创建的 Doris 表的字段分隔符一致，否则无法在 Doris 表中查到数据。可选：'\t'、'\u001'、'\|'、'空格'、';'、','。 |
| 前置 SQL | 执行同步任务之前执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，执行前清空表中的旧数据（truncate table tablename）。 |
| 后置 SQL | 执行同步任务之后执行的 SQL 语句，根据数据源类型对应的正确 SQL 语法填写 SQL，例如，加上某一个时间戳<br>alter table tablename add colname timestamp DEFAULT CURRENT_TIMESTAMP ON UPDATE CURRENT_TIMESTAMP。 |
| 高级参数 | 可根据业务需求配置参数。 |

# 转换节点

最近更新时间：2024-04-02 16:17:11

转换节点主要用于在同步过程中进行数据内容或格式处理，目前实时同步任务支持字段转换和数据清理两类转换节点，离线同步任务仅支持字段转换节点。

## 字符串替换

字符串替换节点主要使用 Java 正则表达式对字符串字段的内容进行匹配与替换，本节点将在字段内容全部或部分匹配字符串时执行替换操作。
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 字段 | 选择需要进行字符串匹配的字段。本节点处理的字段必须为字符串类型。<br>说明：本转换节点中，可对同一个字段配置多个转换规则。多个规则之间将根据配置顺序串行。 |
| 正则表达式 | 用于匹配原字段内容的 Java 正则表达式。 |
| 新字符串 | 对命中的字段内容替换成指定新字符串。 |
| 替换方式 | 支持仅替换第一个匹配结果或全部替换。 |
| 删除 | 删除本行规则。 |

## 时间格式转换

时间格式转换节点主要用于对表中时间字段转换为目标格式，如将 yyyymmdd 时间格式转换为 yyyymmdd hh:mm:ss 格式。
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 字段 | 选择需要进行时间格式匹配的字段，字段需为 datetime、time、timestamp，date 等类型。<br>说明：本转换节点中，可对同一个字段配置多个转换规则。 |
| 源时间格式 | 指定本字段时间格式，若指定格式与实际格式不符合将不进行转换，数据处理为脏数据。 |
| 目标时间格式 | 对于每条记录，若当前字段与指定的源目标格式一致，系统将会把源时间格式转为目标格式。 |
| 删除 | 删除本行规则。 |

> **说明**
> 目前本节点仅支持离线同步任务。

## 值转换

值转换节点主要用于对字段的内容进行标准化处理，本节点将在字段内容与指定内容完全相等时执行替换操作，替换后原字段内容将被更新。如将“CHINA”统一替换成“CHN”。
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 字段 | 选择需要进行时间格式匹配的字段，字段默认来源于上游字段，支持对所有上游字段进行处理。<br>说明：本转换节点中，可对同一个字段配置多个转换规则。 |
| 匹配值 | 用于判断是否原字段是否等于该指定值。若相等，将执行替换行为替换值。 |
| 替换值 | 值转化目标，不可为空。 |
| 删除 | 删除本行规则。 |

> **说明**
> 目前本节点仅支持离线同步任务。

## 字段分割

字段分割节点主要用于原始字段内容使用固定符号进行内容切割，切割后的内容将写入新字段内，原始内容保持不变。本节点通常用于字段内容进行一对多映射的场景，如将字段 A 中"Tencent-cloud"用"-"切割并分别写入两个新字段 B、C 中。
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 字段 | 选择需要进行内容切割的字段。<br>说明：本转换节点中，可对同一个字段配置多个转换规则。 |
| 分割符 | 填写分割标识。<br>说明：分割符支持通过下拉或者手动输入的方式填充,可以从下拉菜单中选择内置的分割符号或者手动填充。 |
| 结果字段名称 | 输入用于接收分割结果到字段名称，多个字段名称之间使用逗号分割。分割结果默认将依次填充进目标字段中，若定义字段名称多于分割结果则多余的字段内容将为默认空值，反之多出的分割内容将被舍弃。<br>说明：如字段 A（内容为"Tencent-cloud"）使用"-"切割，写入 B、C、D 三个字段，则 BC 字段会被分别填充 Tencent、cloud，D 字段内容将为默认空值。 |
| 删除 | 删除本行规则。 |

> **说明**
> 目前本节点仅支持实时同步任务。

## 数据过滤

对表中每行内容根据过滤规则进行筛选和匹配，对于匹配的数据行支持保留或者去除。
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 过滤动作 | • 保留：将命中过滤规则的数据写入到目标表中。<br>• 去除：将命中的规则的数据不写入到目标表中。 |
| 字段 | 选择需要进行内容切割的字段。<br>说明：本转换节点中，可对同一个字段配置多个转换规则 |
| 逻辑运算符 | 支持 AND、OR |
| 运算符 | 目前支持 >、<、<=、>=、=、!=、为空以及不为空 |
| 类型 | 用于与字段内容比较的值的类型：<br>• 字段：使用指定字段的内容过滤字段比较，通常字段内容随不同数据行变化。<br>• 自定义值：使用固定的常量与过滤字段内容比较。 |
| 比较值 | 选择字段或者输入自定义值。 |
| 删除 | 删除本行规则。 |

> **说明**
> 目前本节点仅支持实时同步任务。

## 去重

根据实时数据的处理时间（process_time）对窗口内数据进行内容去重复
转换规则参数说明如下：

| 参数 | 说明 |
| --- | --- |
| 去重动作 | • 保留第一条：当时间窗口内存在重复数据时，保留时间顺序排名的第一条数据。<br>• 保留最后一条：当时间窗口内存在重复数据时，保留时间顺序排名的最后一条数据。 |

| | |
|---|---|
| 去重字段 | 指定用于判定字段内容是否重复的字段，若指定默认使用全部字段。 |
| 删除 | 删除本行规则。 |

> **说明**
> 目前本节点仅支持实时同步任务。

## 数据连接（join）

实时数据流 join，目前支持常规 regular join 。常规 regular join 适用于有界的输入流，默认保存所有 state。join 节点默认需要配置两个输入节点。
转换规则参数说明如下：

| 参数 | 说明 |
|---|---|
| 左/右表 | 选择 join 上游输入节点中作为左表对象的节点名称 |
| 左/右表关联键 | 左右表用于 join 关联的字段名称 |
| 连接方式 | 支持左连接（left join）、右连接（right join）、全连接（full join） |