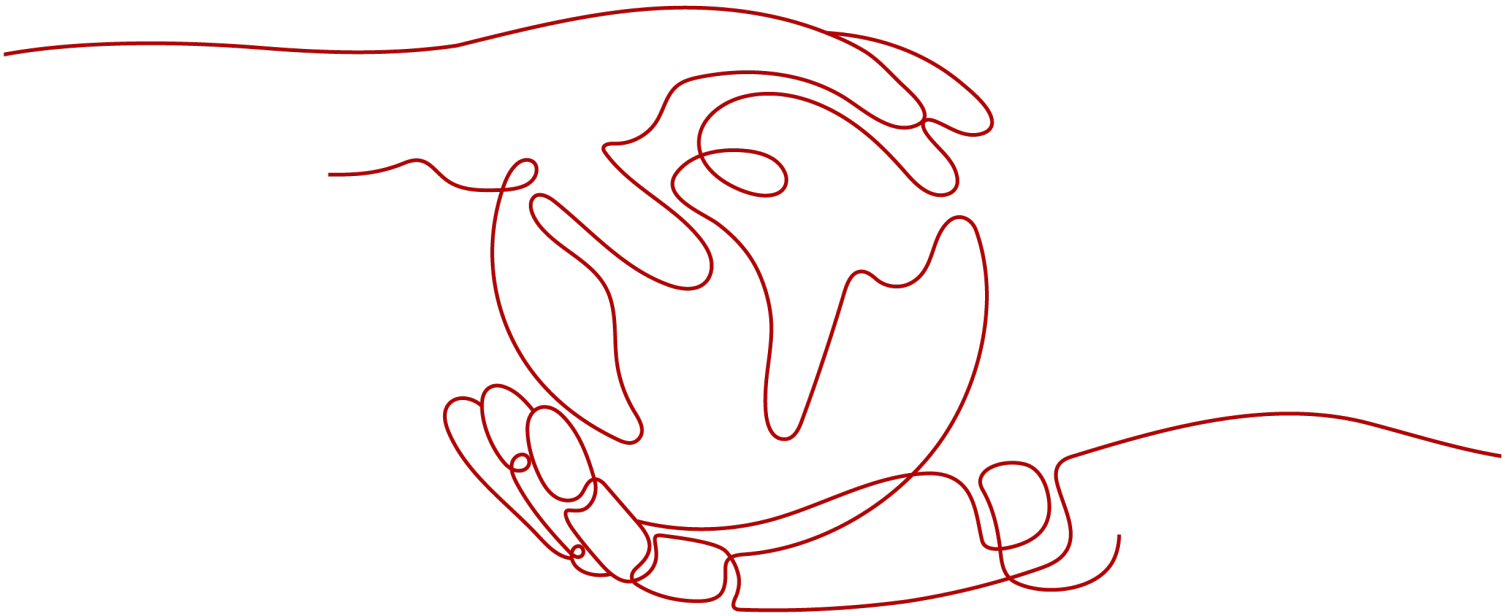
表格存储服务

最佳实践

文档版本 01

发布日期 2025-01-15



华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 数据导入

1.1 使用 DLI Flink 作业实时同步 MRS Kafka 数据至 CloudTable HBase 集群

此章节为您介绍数据实时同步的最佳实践，通过数据湖探索服务DLI Flink作业将MRS kafka数据实时同步给HBase，实现Kafka实时入库到HBase的过程。

- 了解DLI请参见**数据湖探索产品介绍**。
- 了解Kafka请参见**MRS产品介绍**。

图 1-1 数据同步流程图

使用限制

- MRS集群未开启Kerberos认证。
- 为了确保网络连通，MRS集群必须与CloudTable集群的安全组、区域、VPC、子网保持一致。
- MRS与CloudTable安全组入方向添加DLI队列弹性资源网段，建立跨源连接，请参见**创建增强型跨源连接**。
- 必须打通DLI上下游的网络连通性，请参考**测试地址连通性**。

操作流程

基本流程如下：

1. **步骤一：创建CloudTable HBase集群**
2. **步骤二：MRS集群中创建Flink作业制造数据**
3. **步骤三：创建DLI Flink作业进行数据同步**
4. **步骤四：结果验证**

准备工作

- 已注册华为账号并开通华为云，具体请参见**注册华为账号并开通华为云**，且在使用CloudTable前检查账号状态，账号不能处于欠费或冻结状态。
- 已创建虚拟私有云和子网，参见创建**虚拟私有云和子网**。

步骤一：创建 CloudTable HBase 集群

1. 登录表格存储服务控制台，**创建CloudTable HBase集群**。
2. 创建ECS，**请参考准备弹性云服务**。
3. **安装客户端**。
4. 启动Shell访问集群。执行“bin/hbase shell”，启动Shell访问集群。
5. 创建order表。

```
create 'order', {NAME => 'detail'}
```

步骤二：MRS 集群中创建 Flink 作业制造数据

1. 创建**MRS集群**。
2. 登录Manager，选择“集群 > Flink > 概览”，进入概览页面。
3. 单击“Flink WebUI”右侧的链接，访问Flink WebUI。
4. 在MRS Flink WebUI中创建Flink任务产生数据。
   a. 单击作业管理中的“新建作业”，弹出新建作业页面。
   b. 填写参数，单击“确定”，建立Flink SQL作业。如果修改SQL，单击操作列的“开发”，进入SQL页面添加以下命令。

   说明

   ip:port获取ip地址和端口。

   - ip地址获取：进入集群的Manager页面，单击“集群 > Kafka > 实例 > 管理IP（Broker）”，可获取ip地址。
   - port获取：单击配置，进入配置页面，搜索“port”，获取端口（该port是Broker服务监听的PLAINTEXT协议端口号）。
   - 建议properties.bootstrap.servers参数添加多个ip:port，防止kafka实例网络不稳定或其他原因宕机，导致作业运行失败。

   SQL语句示例：

```
CREATE TABLE IF NOT EXISTS `lineorder_hbase` (
`order_id` string,
`order_channel` string,
`order_time` string,
`pay_amount` double,
`real_pay` double,
`pay_time` string,
`user_id` string,
`user_name` string,
`area_id` string
) WITH (
'connector' = 'kafka',
'topic' = 'test_flink',
'properties.bootstrap.servers' = 'ip:port',
'value.format' = 'json',
'properties.sasl.kerberos.service.name' = 'kafka'
);
CREATE TABLE lineorder_datagen (
`order_id` string,
`order_channel` string,
```

```
`order_time` string,
`pay_amount` double,
`real_pay` double,
`pay_time` string,
`user_id` string,
`user_name` string,
`area_id` string
) WITH (
'connector' = 'datagen',
'rows-per-second' = '1000'
);
INSERT INTO
lineorder_hbase
SELECT
*
FROM
lineorder_datagen;
```

c. 回到作业管理界面，单击操作列的“启动”。作业状态为“运行中”表示作业运行成功。

步骤三：创建 DLI Flink 作业进行数据同步

1. 创建弹性资源和队列，请参见“**创建弹性资源池并添加队列**”章节。
2. 创建跨源连接，请参见**创建增强型跨源连接**。
3. 分别测试DLI与上游MRS Kafka和下游CloudTable HBase的连通性。
   a. 弹性资源和队列创建后，单击“资源管理 > 队列管理”，进入队列管理界面测试地址连通性，请参见**测试地址连通性**。
   b. 获取上游IP地址和端口：进入集群的Manager页面，单击“集群 > Kafka > 实例 > 管理IP（Broker）”，可获取IP地址。单击配置，进入配置页面，搜索“port”，获取端口（该port是Broker服务监听的PLAINTEXT协议端口号）。
   c. 获取下游ip地址和端口。
      i. 获取ip：进入集群详情页 > 集群信息 > ZK链接地址（内网）获取域名，执行以下命令解析ip地址。
      ```
      ping 访问域名
      ```
      ii. 获取端口：进入集群详情页 > 集群信息 > ZK链接地址（内网）获取端口。
4. 创建Flink作业，请参见**使用DLI提交作业Flink作业**。
5. 选择**1**中创建的Flink作业，单击操作列的“编辑”，添加SQL进行数据同步。

```
CREATE TABLE orders (
order_id string,
order_channel string,
order_time string,
pay_amount double,
real_pay double,
pay_time string,
user_id string,
user_name string,
area_id string
) WITH (
'connector' = 'kafka',
'topic' = 'test_flink',
'properties.bootstrap.servers' = 'ip:port',
'properties.group.id' = 'testGroup_1',
'scan.startup.mode' = 'latest-offset',
'format' = 'json'
);
create table hbaseSink(
order_id string,
```



```
detail Row(
order_channel string,
order_time string,
pay_amount double,
real_pay double,
pay_time string,
user_id string,
user_name string,
area_id string)
) with (
'connector' = 'hbase-2.2',
'table-name' = 'order',
'zookeeper.quorum' = 'ip:port',
'sink.buffer-flush.max-rows' = '1'
);
insert into hbaseSink select order_id,
Row(order_channel,order_time,pay_amount,real_pay,pay_time,user_id,user_name,area_id) from orders;
```

6. 单击“格式化”，再单击“保存”。

须知

请务必先单击“格式化”将SQL代码进行格式化处理，否则可能会因为代码复制和粘贴操作过程中引入新的空字符，而导致作业执行失败。

7. 回到DLI控制台首页，单击左侧“作业管理 > Flink作业”。
8. 启动1中创建的作业，单击操作列的“启动 > 立即启动”。作业状态为“运行中”表示作业运行成功。

步骤四：结果验证

1. 待MRS Flink任务和DLI任务运行成功后，回到HBase集群运行命令的窗口，启动下游HBase shell客户端。

```
scan 'order'
```

2. 可以看到数据源持续更新。

1.2 使用 DLI Flink 作业实时同步 MRS Kafka 数据至 CloudTable ClickHouse 集群

此章节为您介绍数据实时同步的最佳实践，通过数据湖探索服务DLI Flink作业将MRS kafka任务制造数据实时同步给ClickHouse，实现Kafka实时入库到ClickHouse的过程。

- 了解DLI请参见**数据湖探索产品介绍**。
- 了解Kafka请参见**MRS产品介绍**。

图 1-2 数据同步流程图

使用限制

- MRS集群未开启Kerberos认证。

- 为了确保网络连通，MRS集群必须与CloudTable集群的安全组、区域、VPC、子网保持一致。
- MRS与CloudTable安全组入方向添加DLI队列弹性资源网段，建立跨源连接，请参见**创建增强型跨源连接**。
- 必须打通DLI上下游的网络连通性，请参考**测试地址连通性**。

操作流程

基本流程如下：

1. **步骤一：创建CloudTable ClickHouse集群**
2. **步骤二：MRS集群中创建Flink作业制造数据**
3. **步骤三：创建DLI Flink任务进行数据同步**
4. **步骤四：结果验证**

准备工作

- 已注册华为账号并开通华为云，具体请参见**注册华为账号并开通华为云**，且在使用CloudTable前检查账号状态，账号不能处于欠费或冻结状态。
- 已创建虚拟私有云和子网，参见创建**虚拟私有云和子网**。

步骤一：创建 CloudTable ClickHouse 集群

1. 登录表格存储服务控制台，**创建非安全ClickHouse集群**。
2. 下载**客户端和客户端校验文件**。
3. **准备弹性云服务**。
4. **安装客户端并校验客户端**。
5. 建立flink数据库。

   ```
   create database flink;
   ```

   使用flink数据库。

   ```
   use flink;
   ```

6. 创建flink.order表。

   ```
   create table flink.order(order_id String,order_channel String,order_time String,pay_amount Float64,real_pay Float64,pay_time String,user_id String,user_name String,area_id String) ENGINE = ReplicatedMergeTree('/clickhouse/tables/{shard}/flink/order', '{replica}')ORDER BY order_id;
   ```

7. 查看表是否创建成功。

   ```
   select * from flink.order;
   ```

步骤二：MRS 集群中创建 Flink 作业制造数据

1. 创建**MRS集群**。
2. 登录Manager，选择“集群 > Flink > 概览”，进入概览页面。
3. 单击“Flink WebUI”右侧的链接，访问Flink WebUI。
4. 在MRS Flink WebUI中创建Flink任务产生数据。
   a. 单击作业管理中的“新建作业”，弹出新建作业页面。
   b. 填写参数，单击“确定”，建立Flink SQL作业。如果修改SQL，单击操作列的“开发”，进入SQL页面添加以下命令。

说明

ip:port获取ip地址和端口：

- ip地址获取：进入集群的Manager页面，单击“集群 > Kafka > 实例 > 管理IP（Broker）”，可获取ip地址。
- port获取：单击配置，进入配置页面，搜索“port”，获取端口（该port是Broker服务监听的PLAINTEXT协议端口号）。
- 建议properties.bootstrap.servers参数添加多个ip:port，防止kafka实例网络不稳定或其他原因宕机，导致作业运行失败。

SQL语句示例：

```
CREATE TABLE IF NOT EXISTS `lineorder_ck` (
`order_id` string,
`order_channel` string,
`order_time` string,
`pay_amount` double,
`real_pay` double,
`pay_time` string,
`user_id` string,
`user_name` string,
`area_id` string
) WITH (
'connector' = 'kafka',
'topic' = 'test_flink',
'properties.bootstrap.servers' = 'ip:port',
'value.format' = 'json',
'properties.sasl.kerberos.service.name' = 'kafka'
);
CREATE TABLE lineorder_datagen (
`order_id` string,
`order_channel` string,
`order_time` string,
`pay_amount` double,
`real_pay` double,
`pay_time` string,
`user_id` string,
`user_name` string,
`area_id` string
) WITH (
'connector' = 'datagen',
'rows-per-second' = '1000'
);
INSERT INTO
lineorder_ck
SELECT
*
FROM
lineorder_datagen;
```

c. 回到作业管理界面，单击操作列的“启动”。作业状态为“运行中”表示作业运行成功。

步骤三：创建 DLI Flink 任务进行数据同步

1. 创建弹性资源和队列，请参见“**创建弹性资源池并添加队列**”章节。
2. 创建跨源连接，请参见**创建增强型跨源连接**。
3. 分别测试DLI与上游MRS Kafka和下游CloudTable HBase的连通性。
   a. 弹性资源和队列创建后，单击“资源管理 > 队列管理”，进入队列管理界面测试地址连通性，请参见**测试地址连通性**。
   b. 获取上游IP地址和端口：进入集群的Manager页面，单击“集群 > Kafka > 实例 > 管理IP（Broker）”，可获取IP地址。单击配置，进入配置页面，搜

索“port”，获取端口（该port是Broker服务监听的PLAINTEXT协议端口号）。

c. 获取下游ip地址和端口：进入集群详情页可查看节点ip和端口。

4. 创建Flink作业，请参见**使用DLI提交作业Flink作业**。
5. 选择**1**中创建的Flink作业，单击操作列的“编辑”，添加SQL进行数据同步。

```
create table orders (
order_id string,
order_channel string,
order_time string,
pay_amount double,
real_pay double,
pay_time string,
user_id string,
user_name string,
area_id string
) WITH (
'connector' = 'kafka',
'topic' = 'test_flink',
'properties.bootstrap.servers' = 'ip:port',
'properties.group.id' = 'testGroup_1',
'scan.startup.mode' = 'latest-offset',
'format' = 'json'
);
create table clickhouseSink(
order_id string,
order_channel string,
order_time string,
pay_amount double,
real_pay double,
pay_time string,
user_id string,
user_name string,
area_id string
) with (
'connector' = 'clickhouse',
'url' = 'jdbc:clickhouse://ip:port/flink',
'username' = 'admin',
'password' = '****',
'table-name' = 'order',
'sink.buffer-flush.max-rows' = '10',
'sink.buffer-flush.interval' = '3s'
);
insert into clickhouseSink select * from orders;
```

6. 单击“格式化”，再单击“保存”。

须知

请务必先单击“格式化”将SQL代码进行格式化处理，否则可能会因为代码复制和粘贴操作过程中引入新的空字符，而导致作业执行失败。

7. 回到DLI控制台首页，单击左侧“作业管理 > Flink作业”。
8. 启动**1**中创建的作业，单击操作列的“启动 > 立即启动”。作业状态为“运行中”表示作业运行成功。

步骤四：结果验证

1. 待MRS Flink任务和DLI任务运行成功后，回到ClickHouse集群运行命令的窗口，进入集群客户端。
2. 查看数据库。

```
show databases;
```

3. 使用数据库。
   ```
   use databases;
   ```
4. 查看数据表。
   ```
   show tables;
   ```
5. 查看同步数据。
   ```
   select * from order limit 10;
   ```

图 1-3 查看同步数据