ClickHouse->Apacche Doris(Local File)

1、环境配置

0、前置工作

- 搭建Doris
- 准备好HDFS/对象存储
- 在Doris建好对应的表（字段名称大小写需和ClickHouse相同）

1、下载ClickHouse导出工具

```
git clone https://github.com/LOVEGISER/clickhouse_export.git
```

2、检查环境是否有python3

2、配置工具

```
#进入目录
cd clickhouse_export/clickhouse_python_sink/
#打开配置文件
vim config.py
```

```
# -*- encoding=utf8 -*-

"""
-------------------------------------------------
@author: "wanglei@flywheels.com"
@file: config.py
@time: 2022-07-27
@desc: clickhouse_python_sink Server Config
-------------------------------------------------
"""
from log_utils import logger
#1. which table should want to been export
export_table_list = [
    {
      "db":"ssb", #指定数据库
```

```
      "table":"lineorder", #指定表
      "format":"Parquet", #Parquet/CSVWithNames #指定导出格式
      "filenameExtension":"Parquet", #Parquet/csv #指定导出文件后缀名
      "mode":"partition",#导出模式，可选按照partition分区导出或者all全量导出
      "partition_expr":"toYear(LO_ORDERDATE)",#指定的分区字段
      "upper_condition":"toYear(LO_ORDERDATE)<=999912",#结束分区
      "lower_condition":"toYear(LO_ORDERDATE)>=000001",#开始分区
      "partition_split_filed": "LO_TAX",#对数据量过大的partition按照split_filed再次切
      "partition_split_filed_model": "continuous",#continuous:连续型（数据按照字段线
      "partition_split_filed_type": "long" #datetime/long/date
    }
    # 可以同时指定多个表
   # ,{
   #    "db":"default",
   #    "table":"trips_np",
   #    "format": "Parquet",
   #    "mode":"all",
   #    "partition_expr":"",
   #    "upper_condition":"",
   #    "lower_condition":"",
   #    "partition_split_filed": "",
   #    "partition_split_filed_model": "",
   #    "partition_split_filed_type": ""
   #  }
]
#例子
'''
example:
export_table_list = [
    {
      "db":"default",
      "table":"trips",
      "mode":"all/partition",
      "partition_expr":"toYYYYMM(pickup_date)",
      "upper_condition":"toYYYYMM(pickup_date)<=201508",
      "lower_condition":"toYYYYMM(pickup_date)>=201506",
    }
]
'''
#example :./clickhouse-client --host=<host> --port=<port> --user=<user> --passwor
clickhouse_connect_command = "clickhouse-client --host=172.16.70.243 --port=9000"
#2.thread number use
process_number = 10
sub_partition_max_size=200000
#指定clickhouse的userfiles路径，导出的文件在此
user_files_path = "/mnt/data/clickhouse/user_files"
```

3、运行导出

```
python3 scheduler.py
```

4、数据上传HDFS/对象存储

由clickhouse-client导出的Parquet文件，并不能直接导入到Doris中，需要上传到HDFS/对象存储再由Doris进行Broker Load/S3 Load导入

上传HDFS可用HDFS命令

1、上传对象存储

对象存储上传文件参考如下

腾讯COS：https://cloud.tencent.com/document/product/436/10976
阿里OSS：https://help.aliyun.com/document_detail/50451.html

此处以OSS为例：

```
#先找到clickhouse的user_file目录或者在python脚本中自定义的数据文件目录
#用ossutil64上传整个目录到存储桶
./ossutil64 cp /mnt/data/clickhouse/user_files/ssb/lineorder oss://ck-doris/lineorder
```

2、上传HDFS

5、Doris S3 Load/Broker Load 数据

1、S3 Load

S3 Load针对ck导出的数据存到对象存储上

参考：https://doris.apache.org/zh-CN/docs/dev/data-operate/import/import-way/s3-load-manual

```
LOAD LABEL ssb.lineorder2
    (
        DATA INFILE("s3://ck-doris/lineorder/*")
        INTO TABLE lineorder
        FORMAT AS "parquet"
    )
WITH S3
    (
```

```
        "AWS_ENDPOINT" = "*",
        "AWS_ACCESS_KEY" = "*",
        "AWS_SECRET_KEY"="*",
        "AWS_REGION" = "*"
    )
PROPERTIES
    (
        "timeout" = "3600"
    );
```

2、Broker Load

Broker Load 针对ck导出的数据存到HDFS里

6、数据验证

可以跑一下count(*) 或者常用分析SQL对比CK验证下结果