## 前提

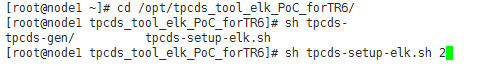
* + - 1. 集群必须安装Hive服务和Elk服务
      2. Tpcds工具包上传节点必须有集群客户端
      3. 本节点可以使用beeline连接hive

## 生成数据

1.进入工具包目录，运行生成数据脚本，2代表生成2G的tpcds测试数据

cd /opt/tpcds\_tool\_elk\_PoC\_forTR6/

sh tpcds-setup-elk.sh 2



生成数据脚本的作用有两个，一是使用hive生成tpcds的测试数据，二是将tpcds中24个表的数据转化为orc格式。

hive生成数据目录为hdfs上的/tmp/tpcds-generate/2（2为数据量）



转换为orc格式的tpcds的数据存储路径为hdfs上的/user/hive/warehouse/tpcds\_bin\_partitioned\_orc\_2.db/（2为数据量，orc格式会将源数据压缩）



## 建立elk外表

1.取消ACL权限

hdfs dfs -setfacl -R -b /user/hive/warehouse/tpcds\_bin\_partitioned\_orc\_2.db/



2.修改create\_elk\_for\_table\_replication\_bf.sql

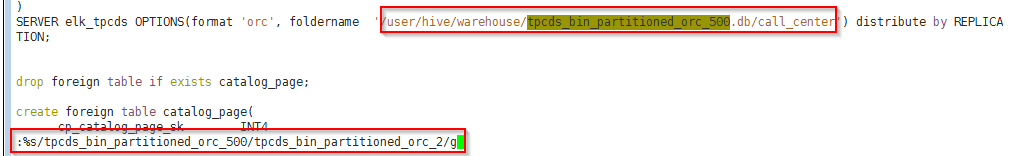
vim ddl-tpcds/elk\_hdfs\_table/create\_elk\_for\_table\_replication\_bf.sql

修改address中ip地址，改为HDFS的namenode所在的节点ip（端口号不需改变）



运行命令批量修改数据存储路径，使其与orc格式数据的存储路径保持一致

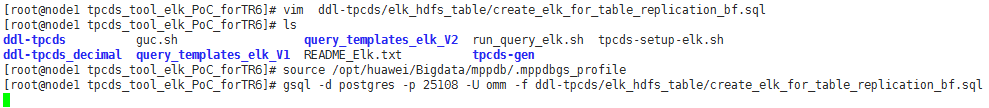
:%s/tpcds\_bin\_partitioned\_orc\_500/tpcds\_bin\_partitioned\_orc\_2/g



3.生成外表

source /opt/huawei/Bigdata/mppdb/.mppdbgs\_profile

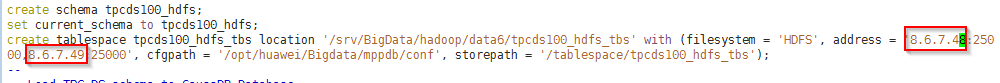
gsql -d postgres -p 25108 -U omm -f ddl-tpcds/elk\_hdfs\_table/create\_elk\_for\_table\_replication\_bf.sql



## 建立elk表，导入数据

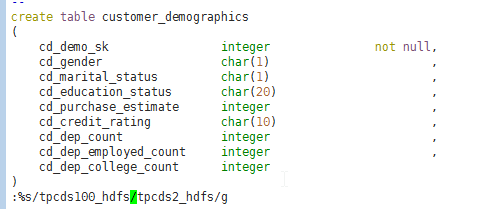
vim ddl-tpcds/elk\_hdfs\_table/tpcds\_hdfs\_table\_mid\_new.sql

修改address中ip地址，改为HDFS的namenode所在的节点ip（端口号不需改变）



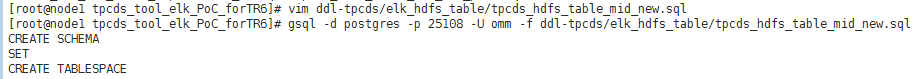
运行命令批量修改表空间名

:%s/tpcds100\_hdfs/tpcds2\_hdfs/g



生成elk表

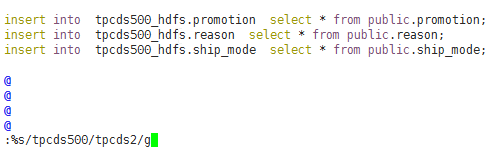
gsql -d postgres -p 25108 -U omm -f ddl-tpcds/elk\_hdfs\_table/tpcds\_hdfs\_table\_mid\_new.sql



vim ddl-tpcds/elk\_hdfs\_table/bulkload.sql

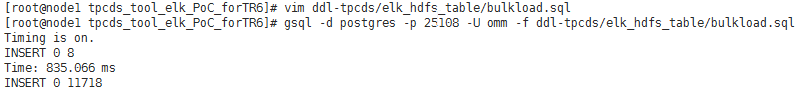
运行命令批量修改表空间名

:%s/tpcds500/tpcds2/g



将外表数据导入elk表

gsql -d postgres -p 25108 -U omm -f ddl-tpcds/elk\_hdfs\_table/bulkload.sql



## 运行查询脚本

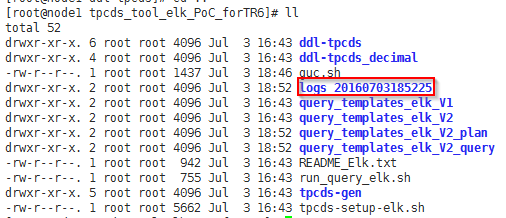
使用外表数据运行tpcds查询

sh run\_query\_elk.sh postgres public

使用elk表数据运行tpcds查询

sh run\_query\_elk.sh postgres tpcds2\_hdfs(tpcds2\_hdfs为在tpcds\_hdfs\_table\_mid\_new.sql 中设置的schema名)

查询结果在最新生成的日志文件夹里面



进入日志文件可以运行下面命令，查看总耗时（单位为ms）

grep "total time" \* | awk '{sum+=$3} END {print sum} '

