Практика. Работа с HIVE

1. <u>Развернем контейнер с HIVE</u>

Перейду с папку с контейнером

cd Downloads/ docker-hive-master

и подниму контейнер

docker-compose up -d

```
max@max:~/Downloads/docker-hive-master$ docker-compose up -d
Starting docker-hive-master_datanode_1 ... done
Starting docker-hive-master_hive-server_1 ... done
Starting docker-hive-master_hive-metastore-postgresql_1 ... done
Starting docker-hive-master_namenode_1 ... done
Starting docker-hive-master_presto-coordinator_1 ... done
Starting docker-hive-master_hive-metastore_1 ... done
```

2. Идем в бобра. -> Запускаем DBeaver

Скачать датасет с каггла: https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city

uber-raw-data-janjune-15.csv

3. Закину его на namenod:

docker cp "/home/max/Downloads/hive_db/uber-raw-data-janjune-15.csv" docker-hive-master_namenode_1:

```
max@max:~/Downloads/docker-hive-master$ docker cp "/home/max/Downloads/hive_db/uber-raw-data-janjun
e-15.csv" docker-hive-master_namenode_1:/
max@max:~/Downloads/docker-hive-master$
```

- 4. Запускаю namenod, куда скопировали файл docker exec -it docker-hive-master_namenode_1 bash
- ls увидим, что uber-raw-data-janjune-15.csv лежит в папке

```
root@5d6869/0546f:/# Ls
airports.csv
               entrypoint.sh
                               home
                                                                                      var
                                       mnt
                                             run
                                                      SYS
bin
               etc
                               lib
                                      opt
                                             run.sh
                                                     tmp
                               lib64
                                                      uber-raw-data-janjune-15.csv
boot
               hadoop
                                      ргос
                                             sbin
               hadoon-data
```

В контейнере создам папку «/my_testdata»

hdfs dfs -mkdir /my_testdata

и закину туда файл uber-raw-data-janjune-15.csv

hdfs dfs -put uber-raw-data-janjune-15.csv /my_testdata

```
root@5d686970546f:/# hdfs dfs -put uber-raw-data-janjune-15.csv /my_testdata
root@5d686970546f:/# hdfs dfs -ls /my_testdata
Found 1 items
-rw-r--r-- 3 root supergroup 551672691 2021-12-19 09:04 /my_testdata/uber-raw-data-janjune-15.cs
```

Удалю файл файл uber-raw-data-janjune-15.csv из контейнера (он уже внутри hdfs)

ВЗЯТЬ файл из контейнера и закинуть его в нашу систему:

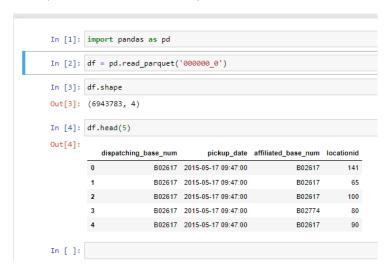
Сохранили в контейнере файл

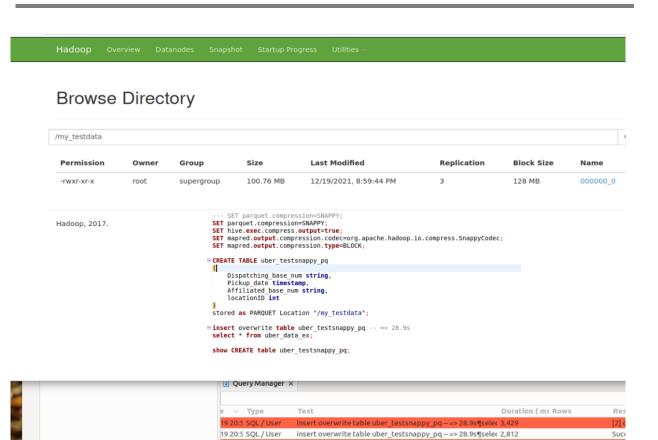
hdfs dfs -get /user/hive/warehouse/mydb.db/uber_data_ex_pq/000000_0

Скопировал в локалку (в Downloads)

docker cp 5d686970546f:/000000_0 ~/Downloads

Прочитал Pandas`ом. Попробовал вычисления всякие – все работает





insert overwrite table uber_testsnappy_pq - => 28.9s¶selec 4,785

insert overwrite table uber_testsnappy_pq - => 28.9s¶selec 3,867

insert overwrite table uber testsnappy pg -=> 28.9s¶selec 4,726

19 20:5 SQL / User

19 20:5 SQL / User

19 20:5 SOL / User

ORC c orc.compress"="gzip" не записалась

