Практика. Работа с HIVE

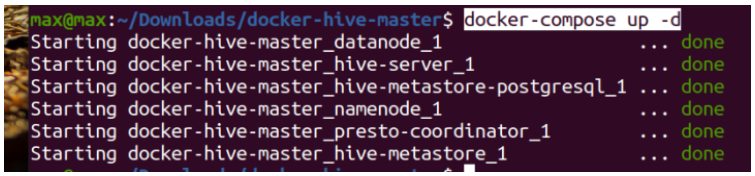
1. Развернем контейнер с HIVE

Перейду с папку с контейнером

cd Downloads/ docker-hive-master

и подниму контейнер

docker-compose up -d



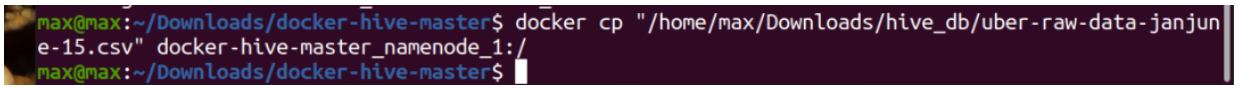
2. Идем в бобра. -> Запускаем DBeaver

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

uber-raw-data-janjune-15.csv

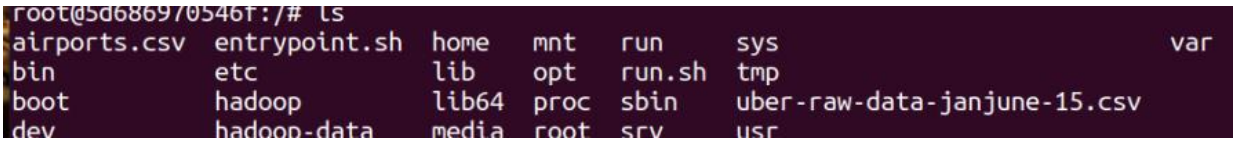
3. Закину его на namenod:

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



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

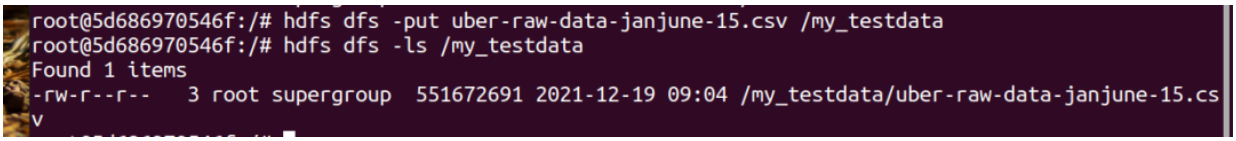


В контейнере создам папку «/my_testdata»

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



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)

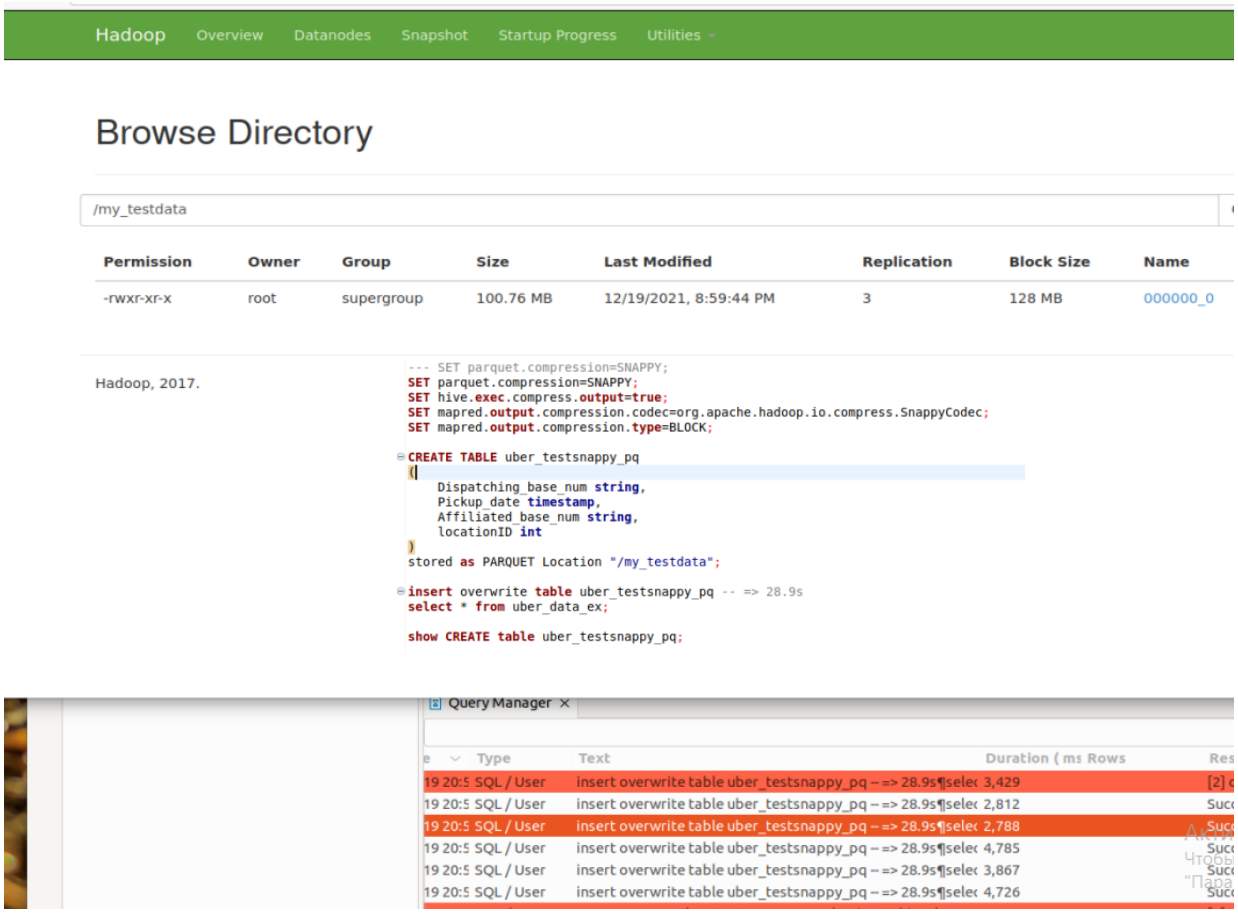
docker cp 5d686970546f:/000000_0 ~/Downloads

Прочитал Pandas`ом. Попробовал вычисления всякие – все работает

```
In [1]: import pandas as pd

In [2]: df = pd.read_parquet('000000_0')

In [3]: df.shape
Out[3]: (6943783, 4)

In [4]: df.head(5)
Out[4]:
```

	dispatching_base_num	pickup_date	affiliated_base_num	locationid
0	B02617	2015-05-17 09:47:00	B02617	141
1	B02617	2015-05-17 09:47:00	B02617	65
2	B02617	2015-05-17 09:47:00	B02617	100
3	B02617	2015-05-17 09:47:00	B02774	80
4	B02617	2015-05-17 09:47:00	B02617	90

```
In [ ]:
```



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