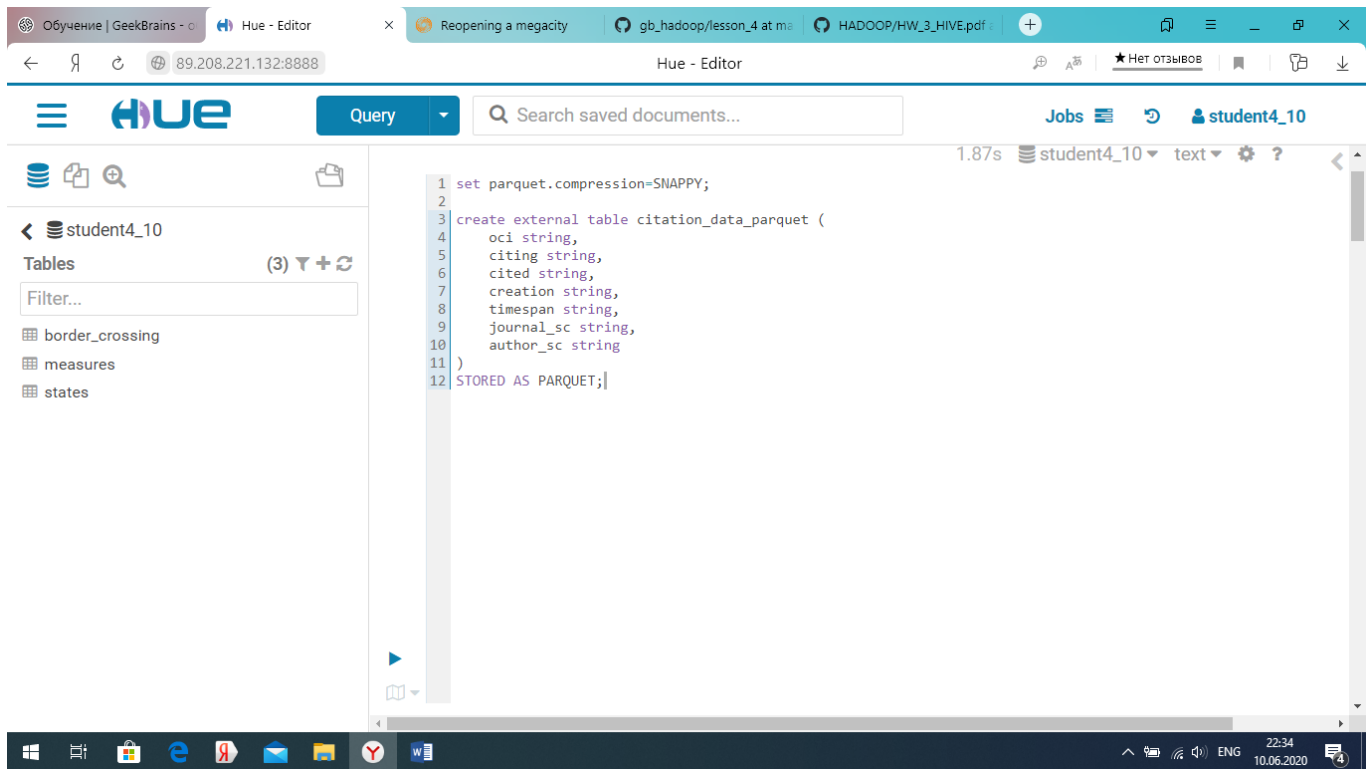


1. Создать таблицы в форматах PARQUET/ORC/AVRO с компрессией и без оной. (Выберите один вариант, например ORC с компрессией).

Выберем вариант PARQUET с компрессией.



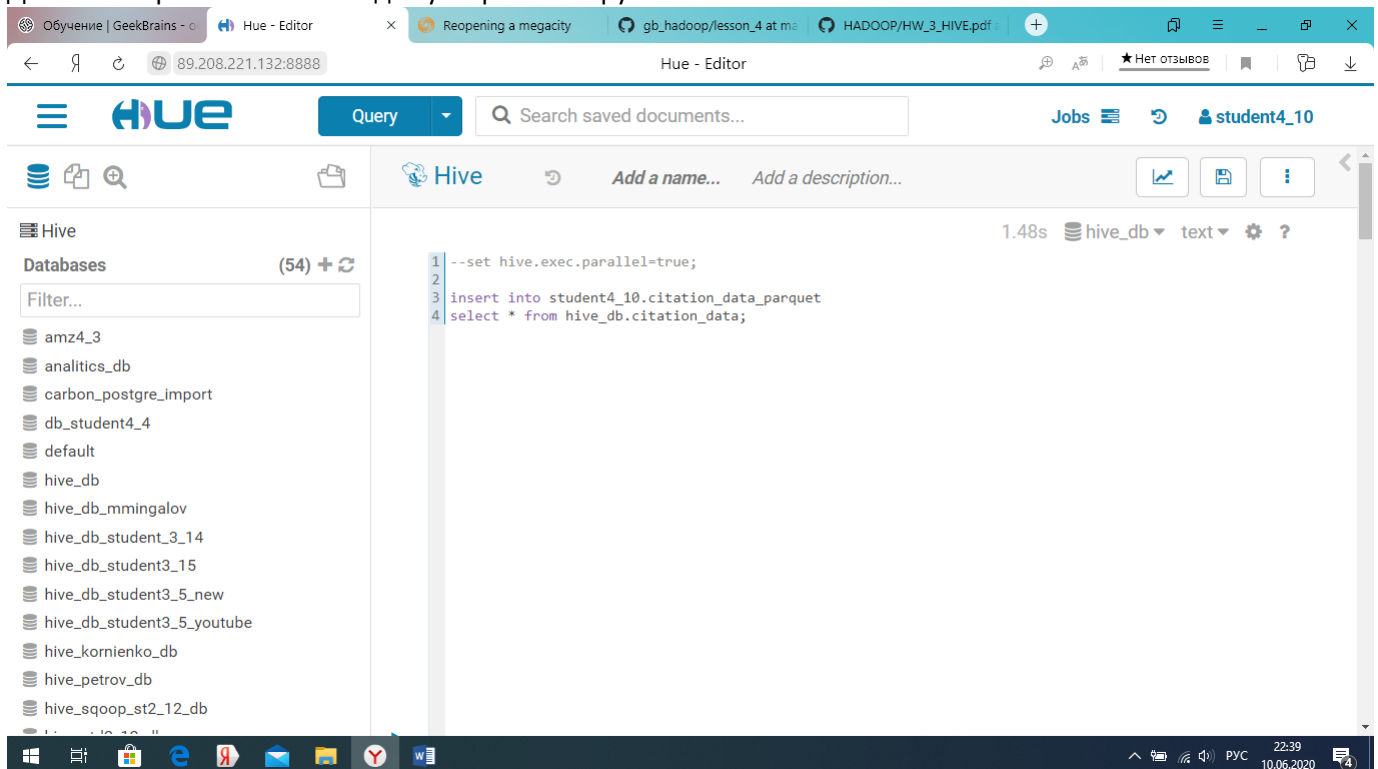
```
1 set parquet.compression=SNAPPY;
2
3 create external table citation_data_parquet (
4   oci string,
5   citing string,
6   cited string,
7   creation string,
8   timespan string,
9   journal_sc string,
10  author_sc string
11 )
12 STORED AS PARQUET;
```

Try to do it with the location

2. Заполнить данными из большой таблицы hive_db.citation_data

set hive.exec.parallel=true;

Добавляем parallel execution для ускорения загрузки



```
1 --set hive.exec.parallel=true;
2
3 insert into student4_10.citation_data_parquet
4 select * from hive_db.citation_data;
```

3-4. Посмотреть на получившийся размер данных.

```
student4_10@manager:~$ ssh -i /home/student4_10/.ssh/authorized_keys student4_10@manager
Using username "student4_10".
Authenticating with public key "imported-openssh-key"
Passphrase for key "imported-openssh-key":
Last login: Thu Jun 11 10:01:22 2020 from 2.92.115.137
[student4_10@manager ~]$ hdfs dfs -du -h -s /user/hive/warehouse/student4_10.db/citation_data_parquet
22.7 G  68.2 G  /user/hive/warehouse/student4_10.db/citation_data_parquet
[student4_10@manager ~]$ hdfs dfs -du -h -s /test_datasets/citation
97.2 G  194.4 G  /test_datasets/citation
[student4_10@manager ~]$
```

Размер citation_data_parquet как мин в 4 раза меньше исходного файла CSV citation