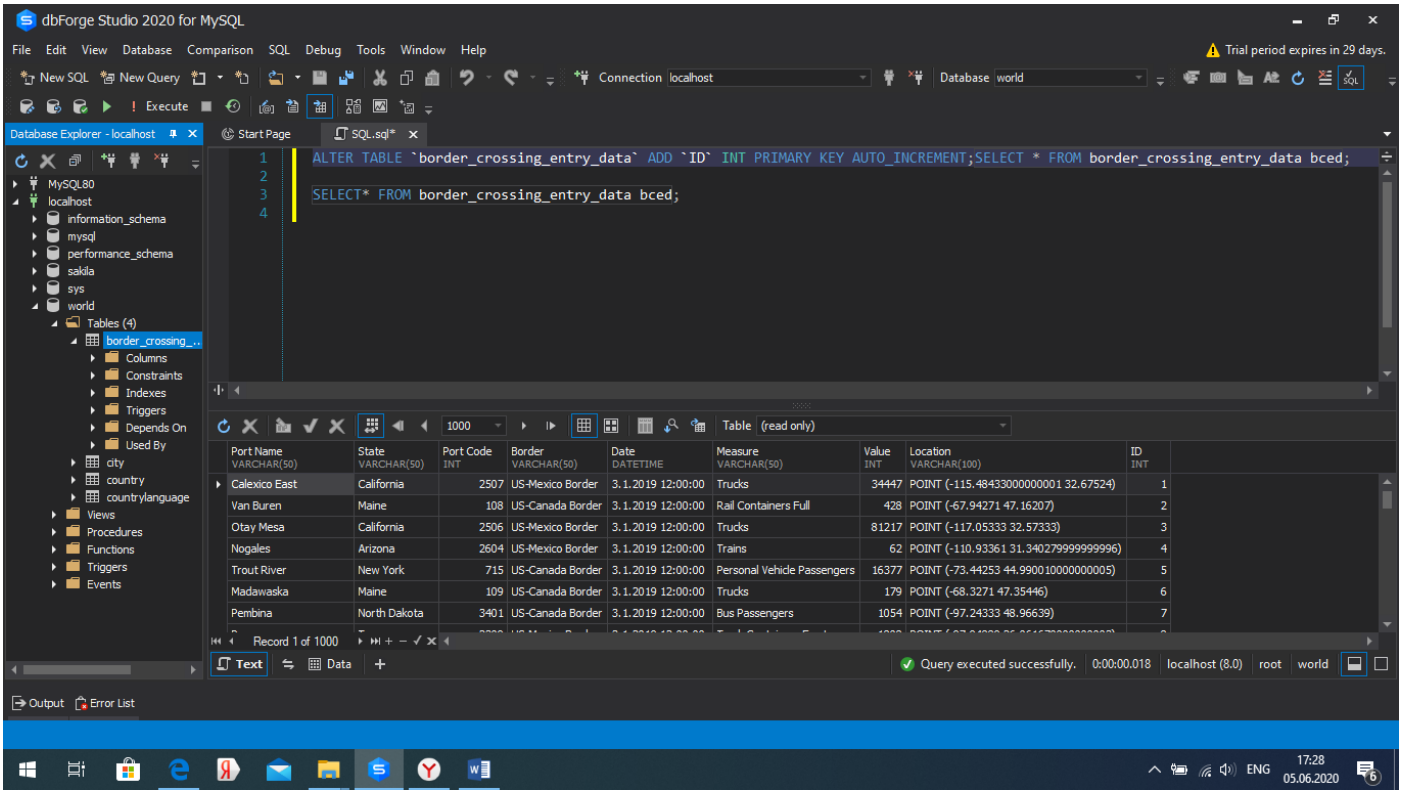
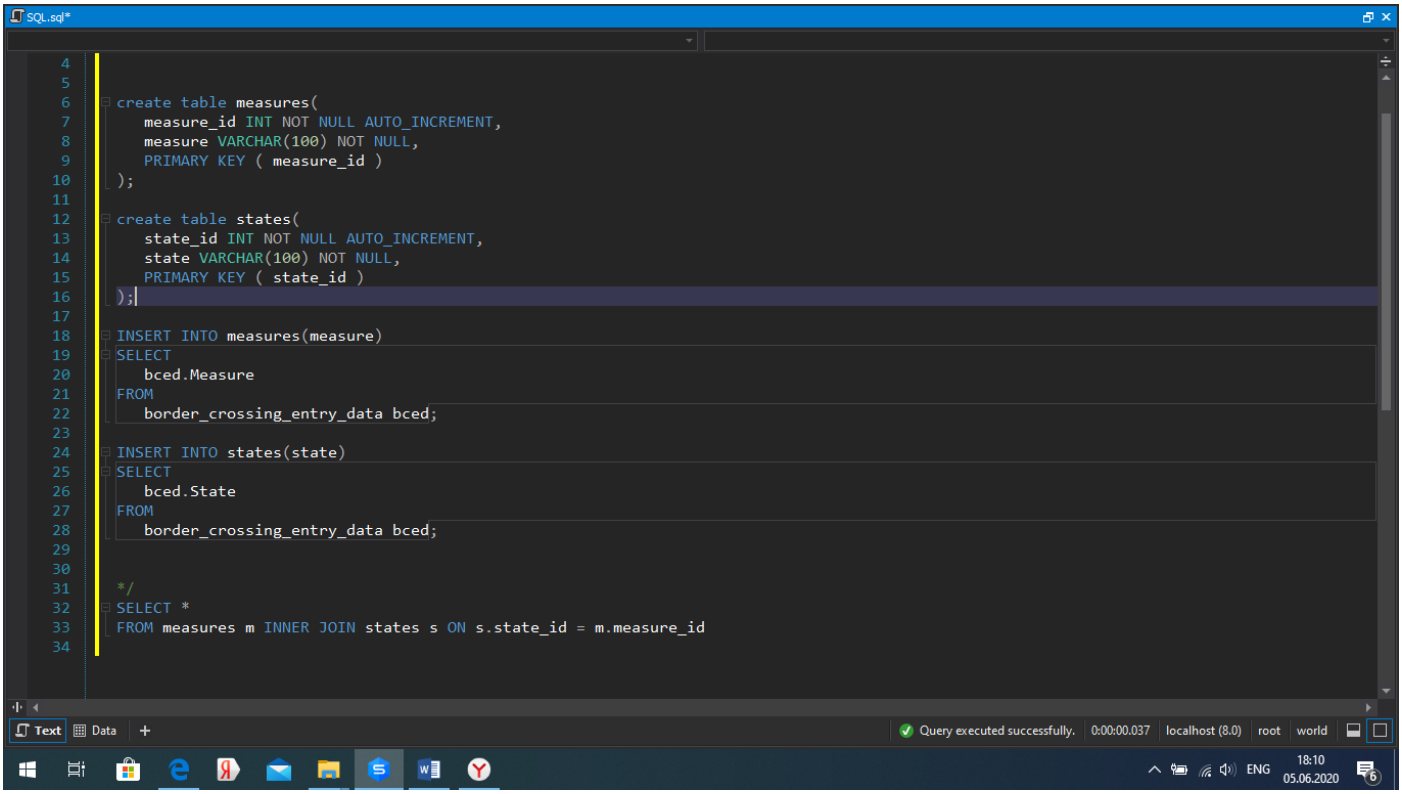
2. Скачать любой датасет из списка ниже.

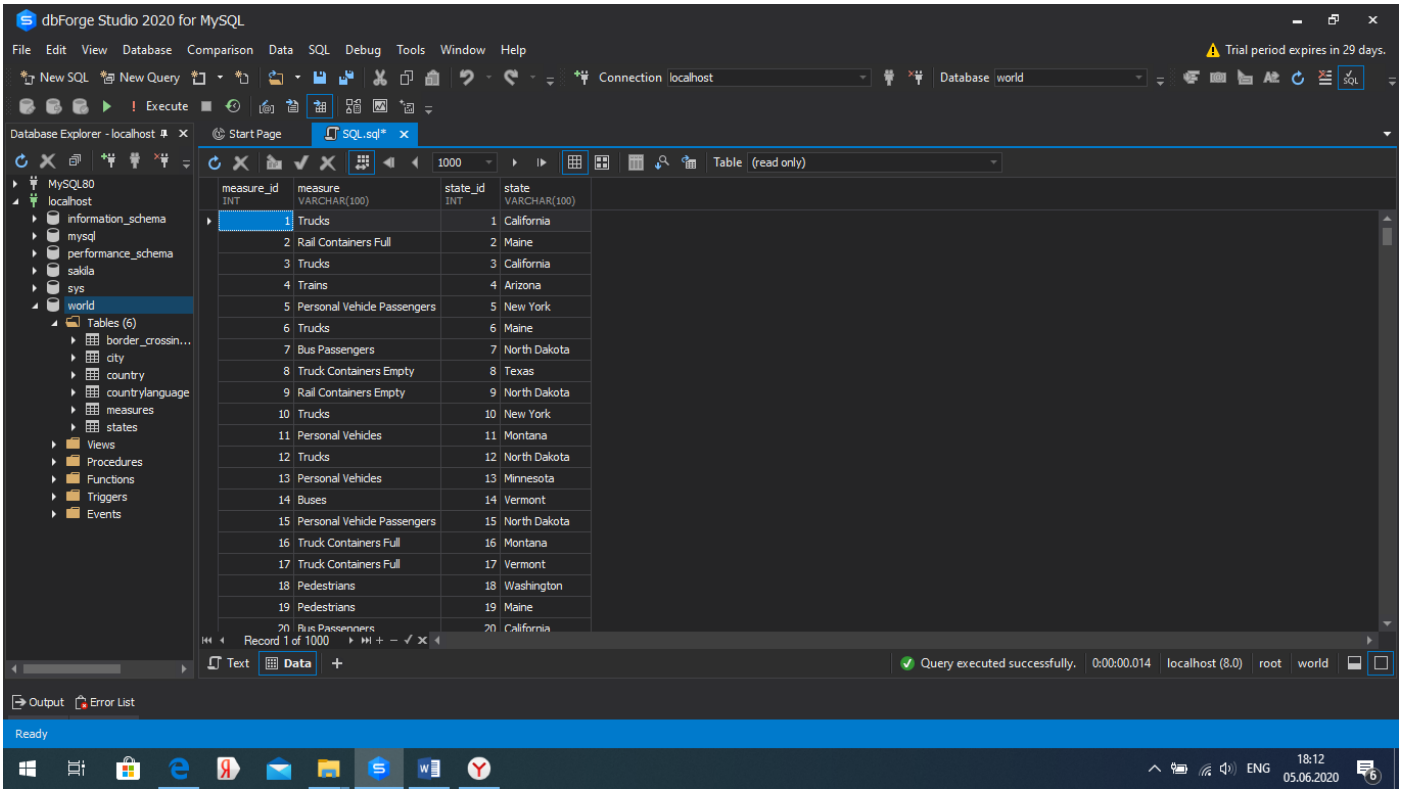
https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions https://www.kaggle.com/datasnaek/youtube-new https://www.kaggle.com/akhilv11/border-crossing-entry-data https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil

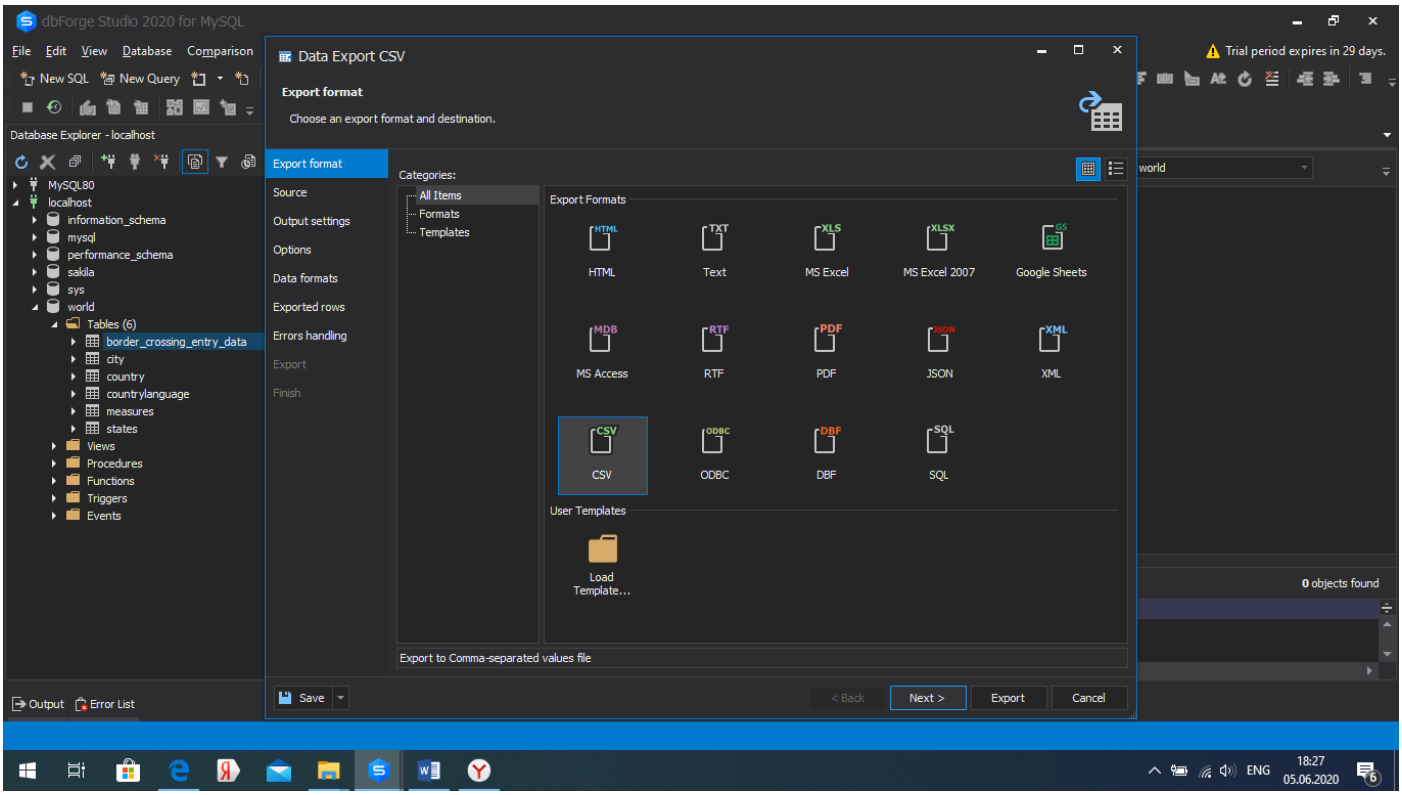


Скопировали с Сайта KAGGLE csv file border_crossing_entry_data и загрузили его в базу. Потом добавил колонку id



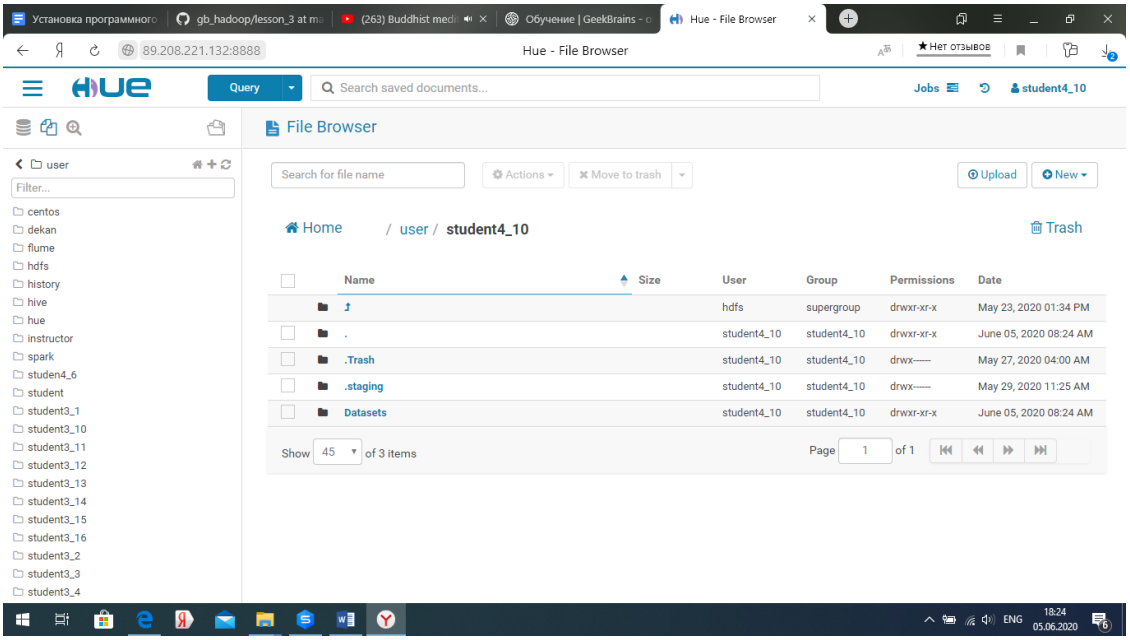
Создаем таблицы для таблиц measures и states, потом заливаем данные в эти таблицы. Сделаем select запрос для проверки данных



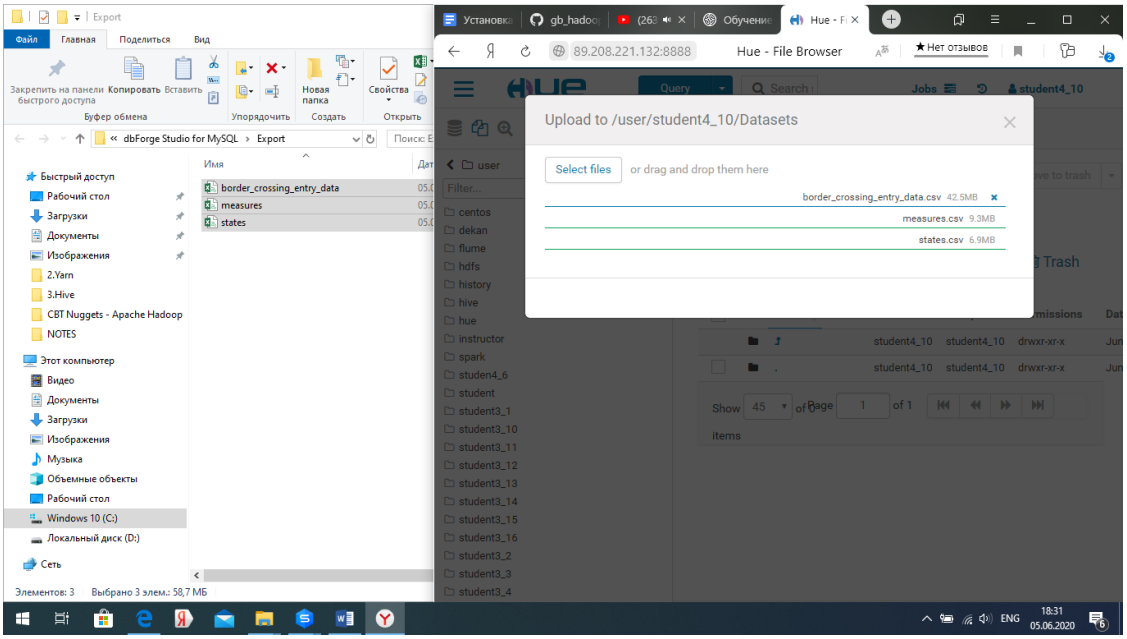


Экспортируем данные как CSV

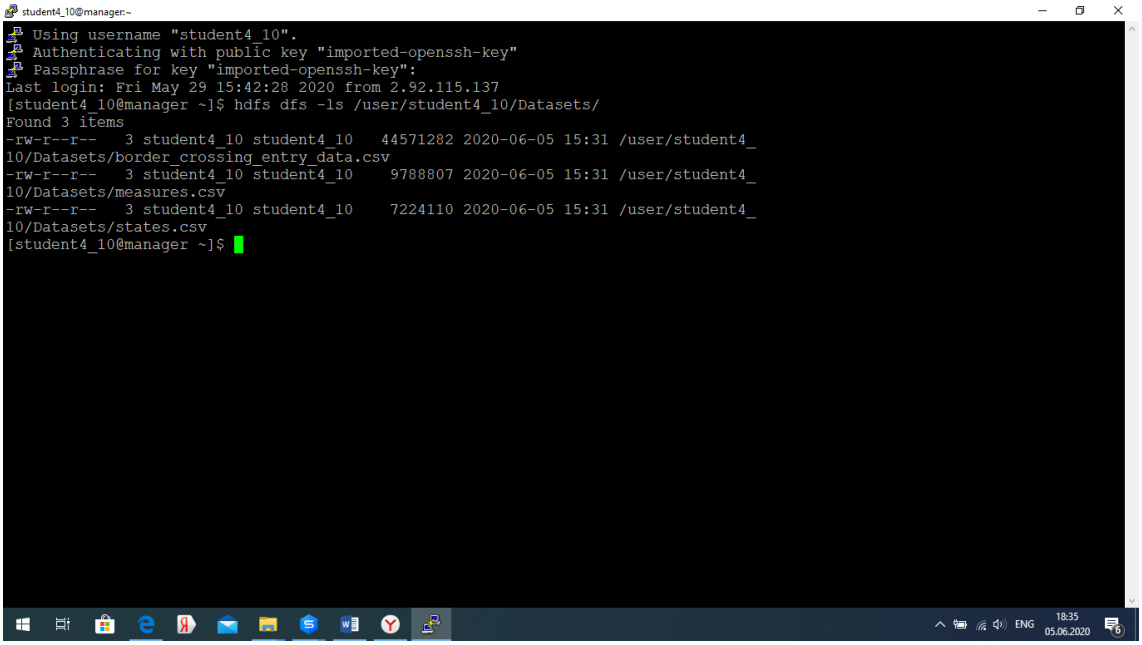
3. Загрузить этот датасет в HDFS в свою домашнюю папку.



Создаем Папку DATASET в HUE

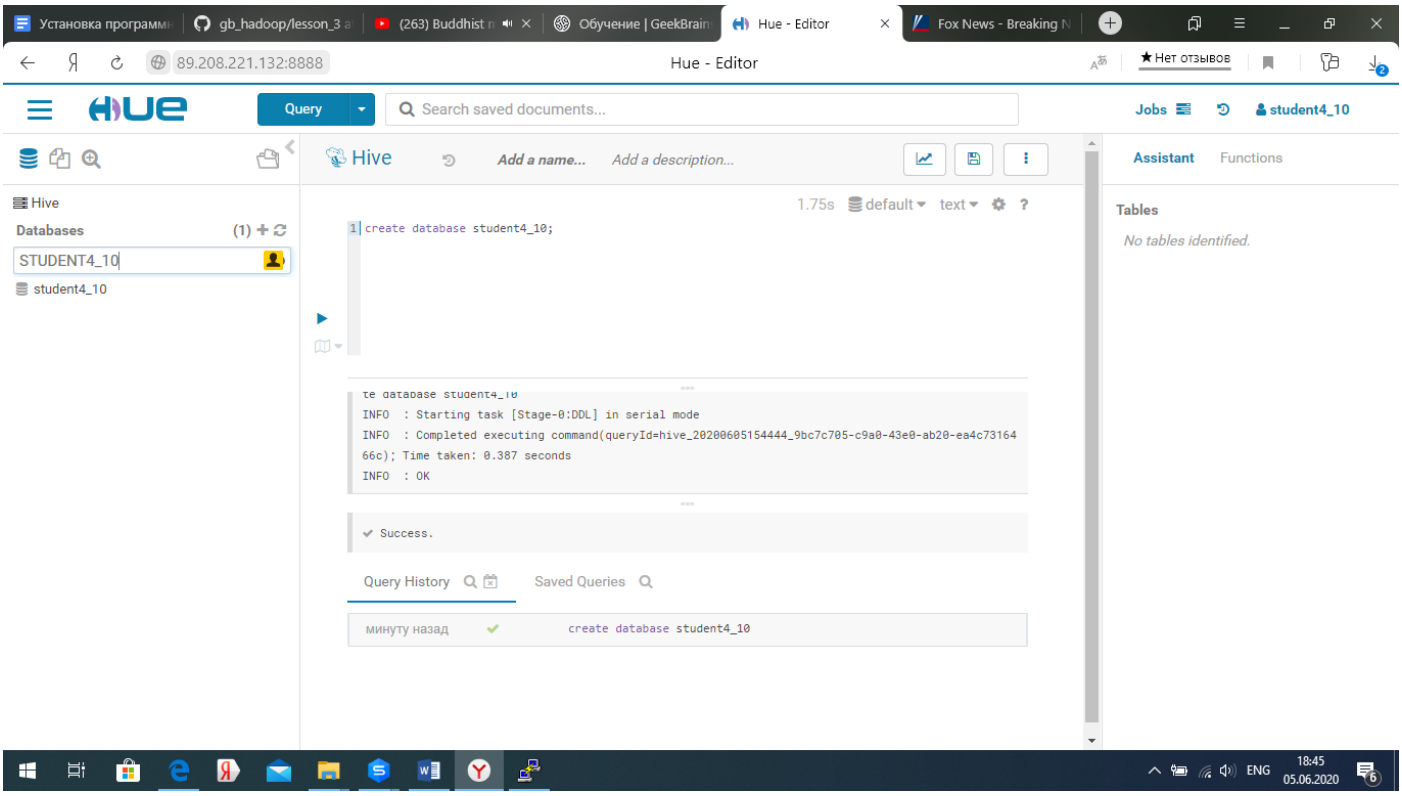


Загружаем файлы в HUE

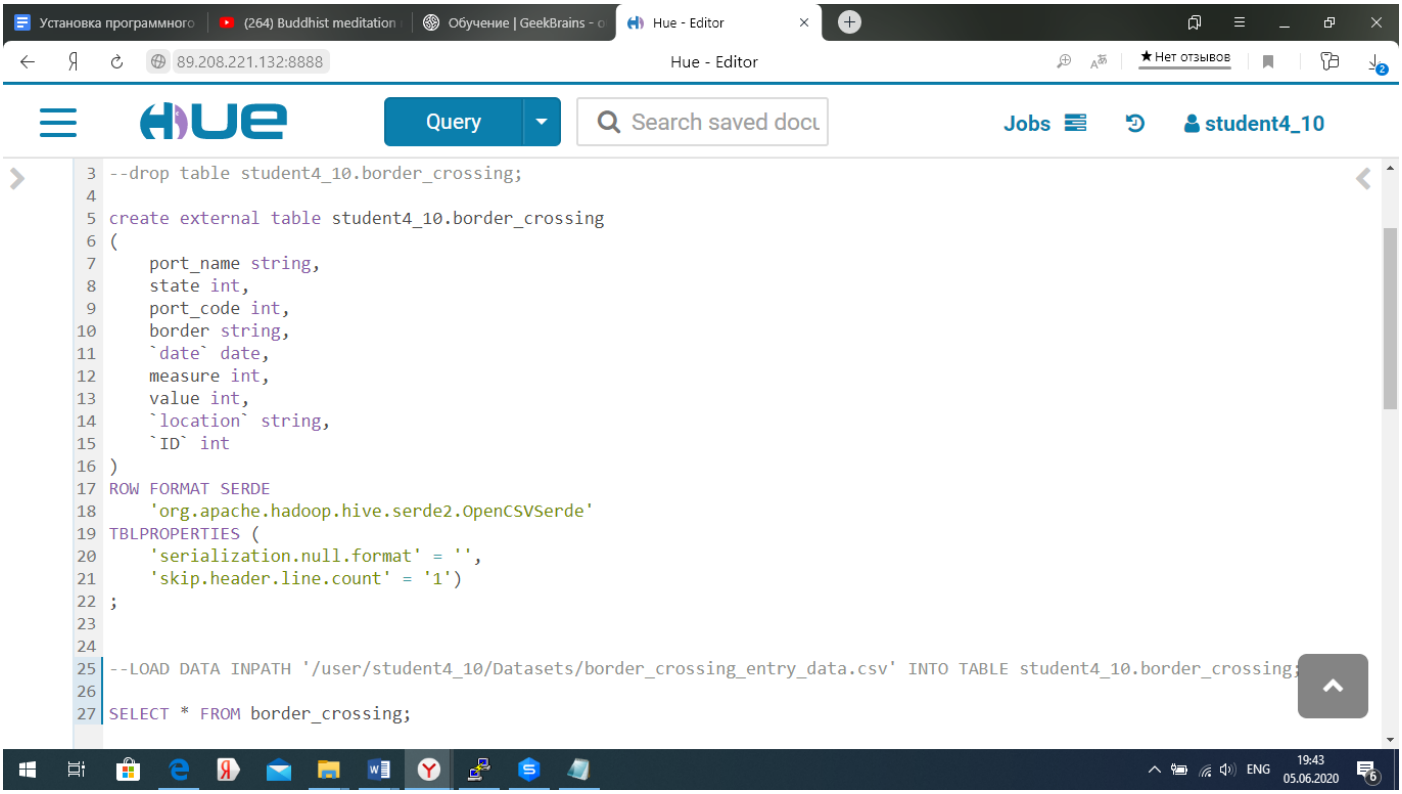


Через Putty перепроверяем что данные загружены в папку DATASET

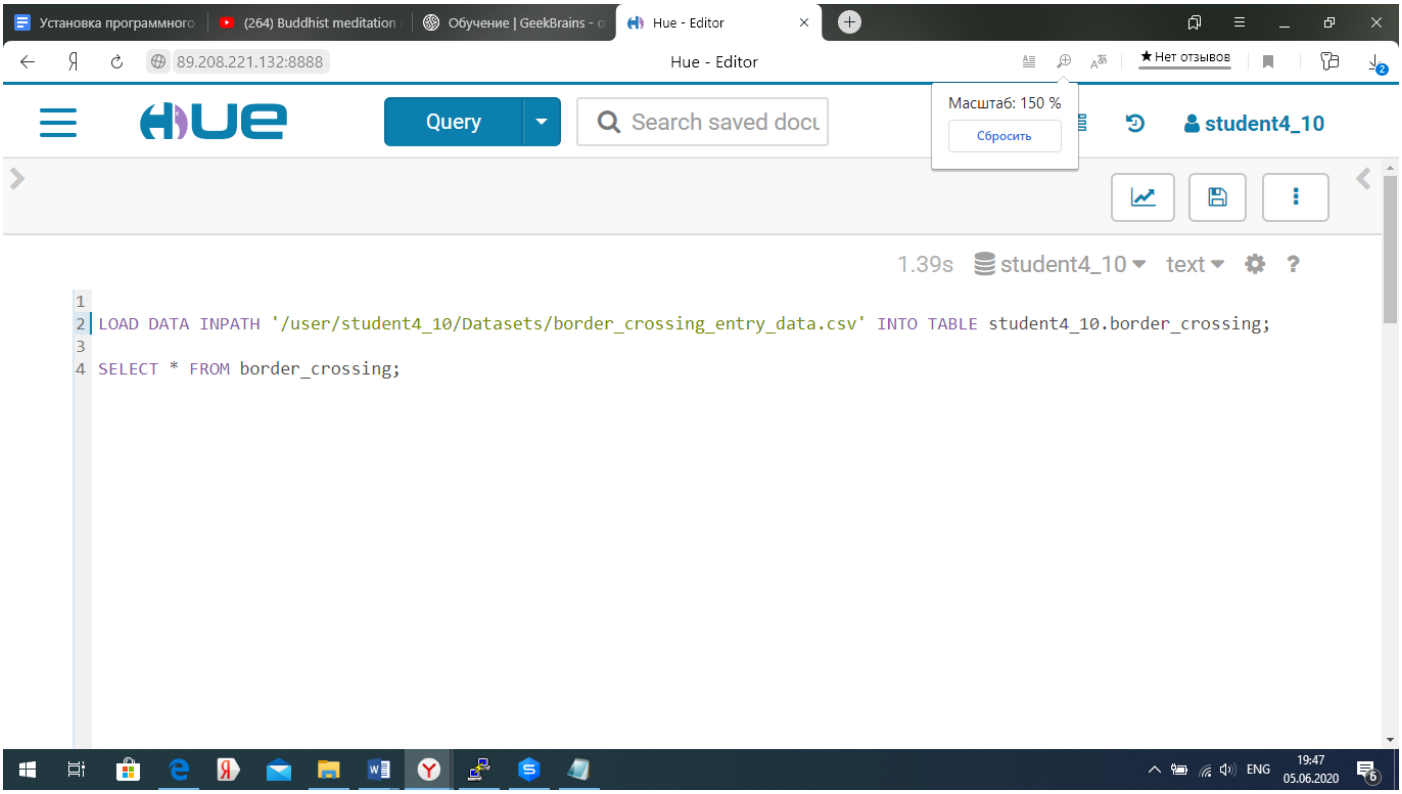
4. Создать собственную базу данных в HIVE



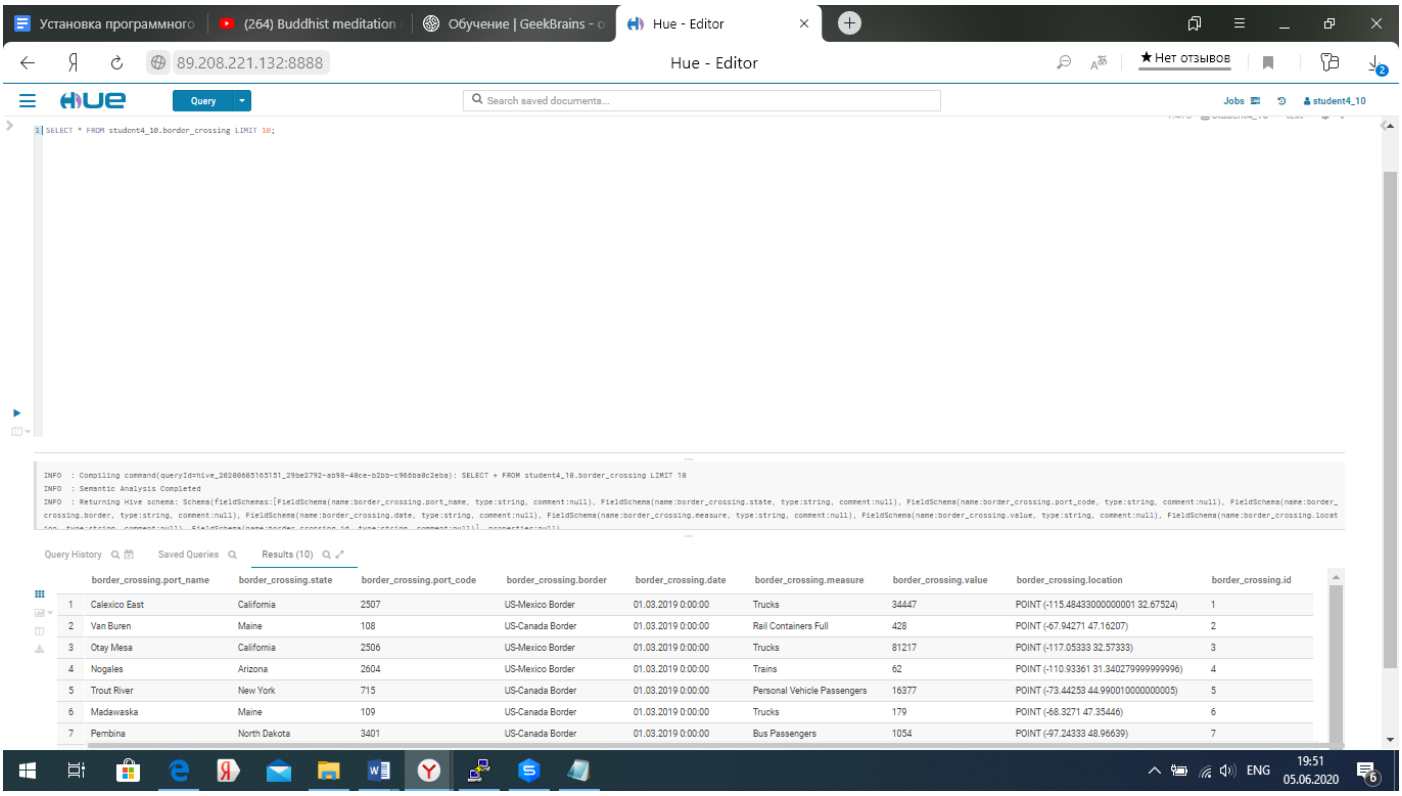
5. Создать EXTERNAL таблицы внутри базы данных с использованием всех загруженных файлов. Один файл – одна таблица.



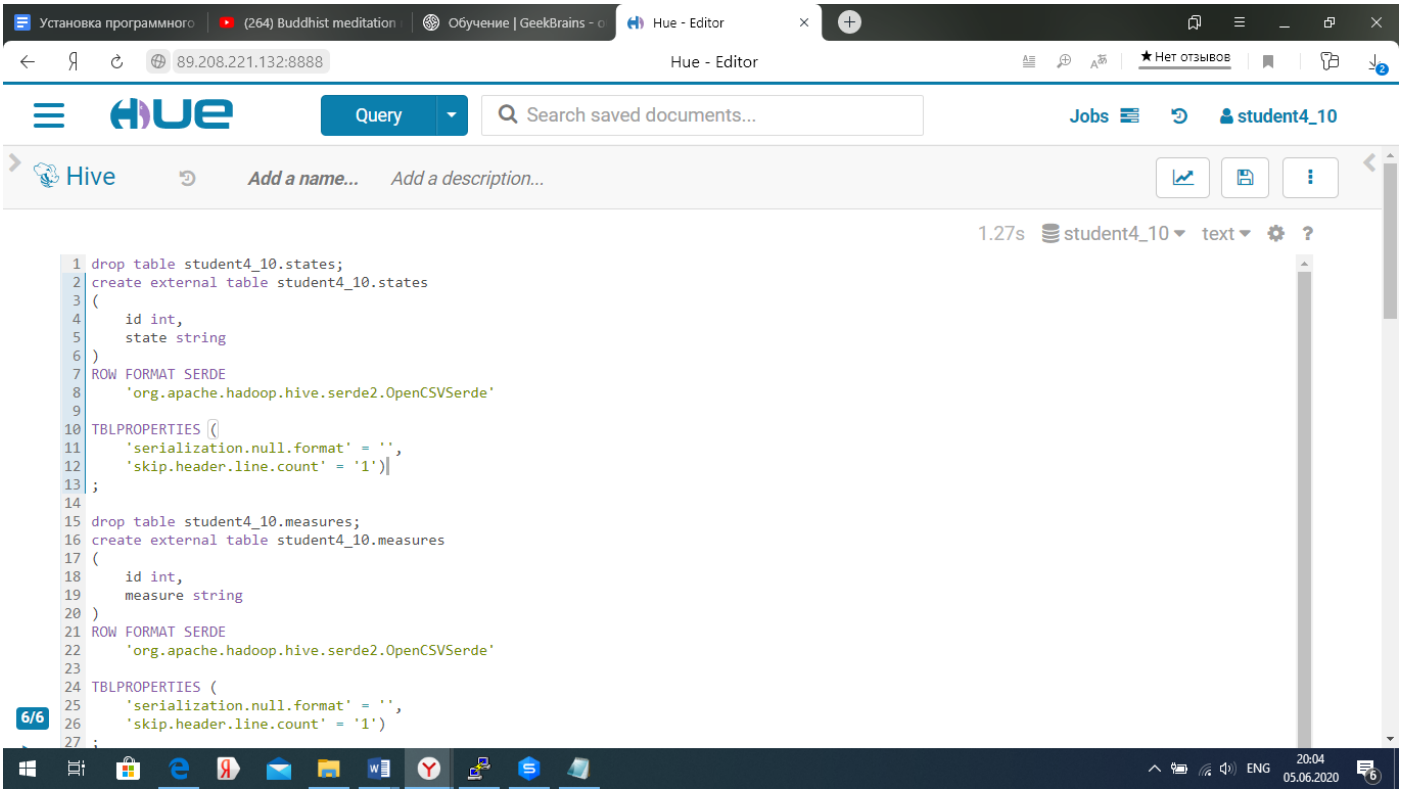
Создаем extertnal table под именем border_crossing с CSV чтением



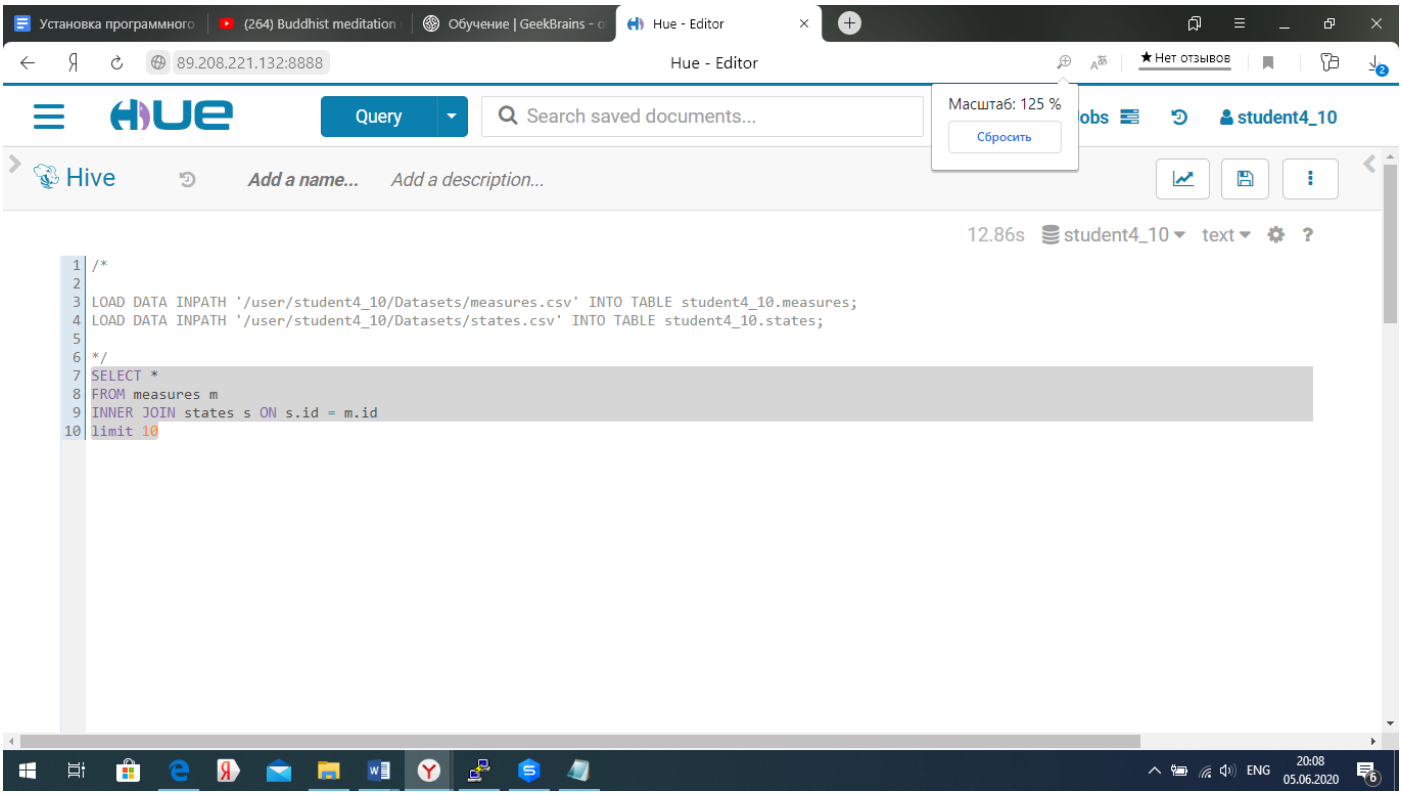
Загружаем данные из csv файла в базовую таблицу border_crossing



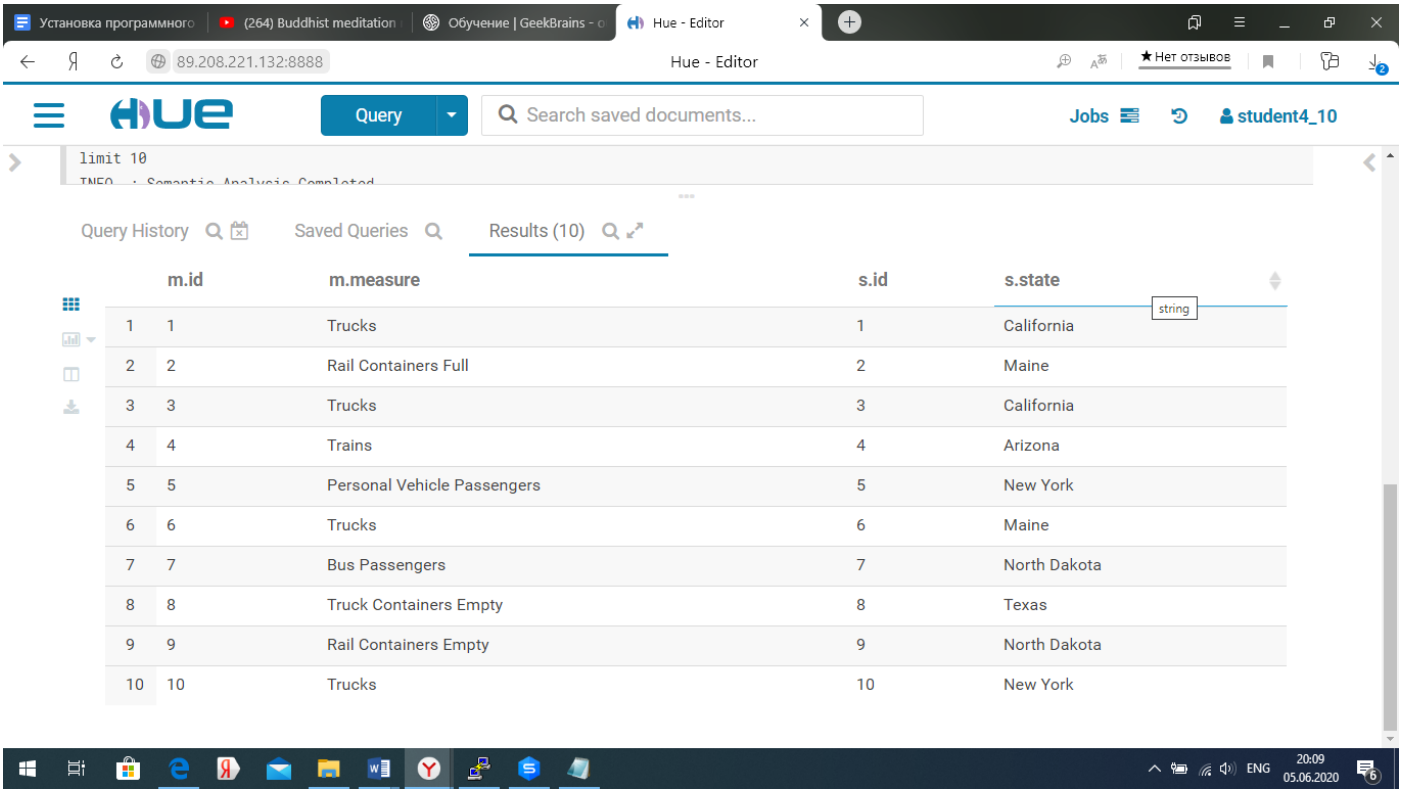
Проверяем данные



создаем внешние таблицы для states и measures



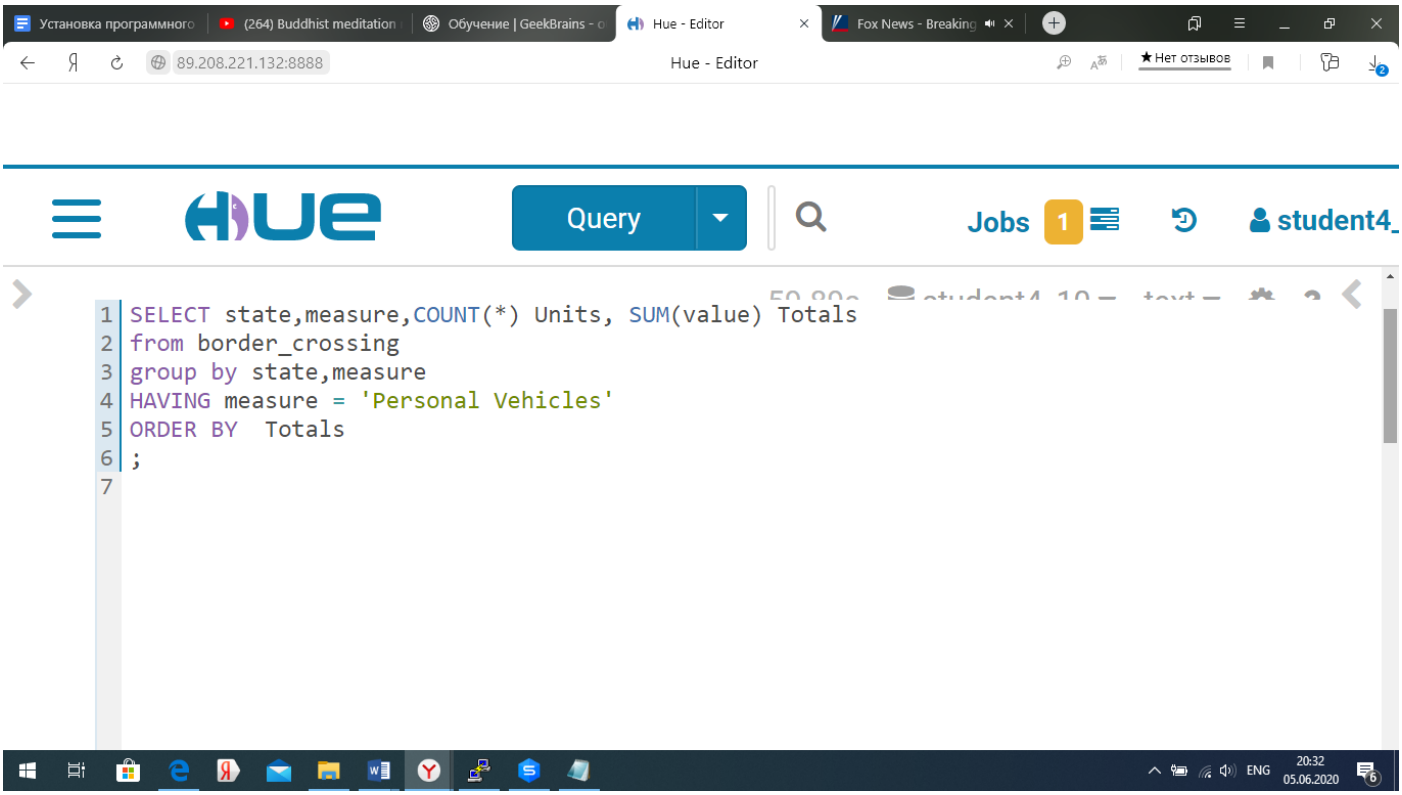
Загружаем данные из csv файла в базовую таблицы states и measures, потом проверяем данные

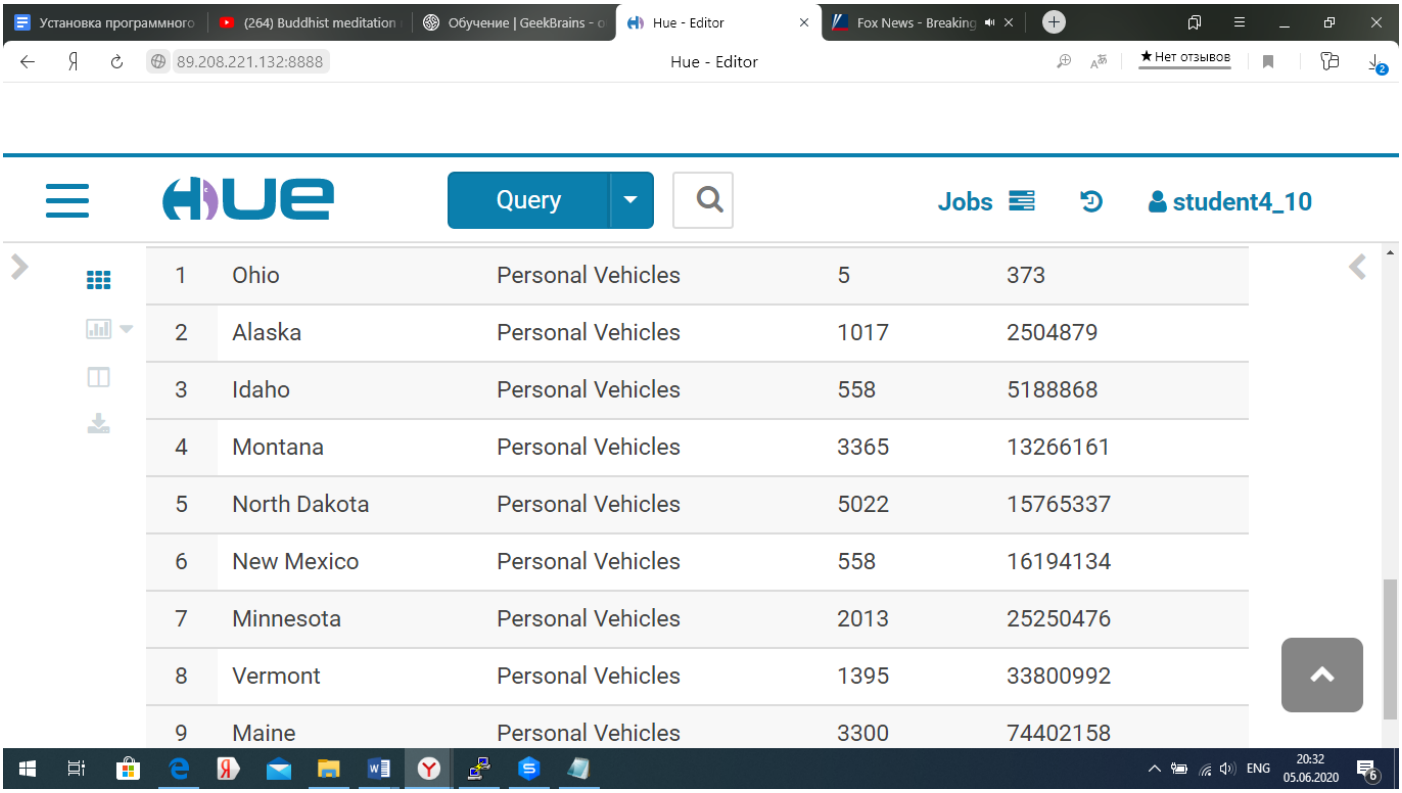


	m.id	m.measure	s.id	s.state
1	1	Trucks	1	California
2	2	Rail Containers Full	2	Maine
3	3	Trucks	3	California
4	4	Trains	4	Arizona
5	5	Personal Vehicle Passengers	5	New York
6	6	Trucks	6	Maine
7	7	Bus Passengers	7	North Dakota
8	8	Truck Containers Empty	8	Texas
9	9	Rail Containers Empty	9	North Dakota
10	10	Trucks	10	New York

результат JOIN двух таблиц

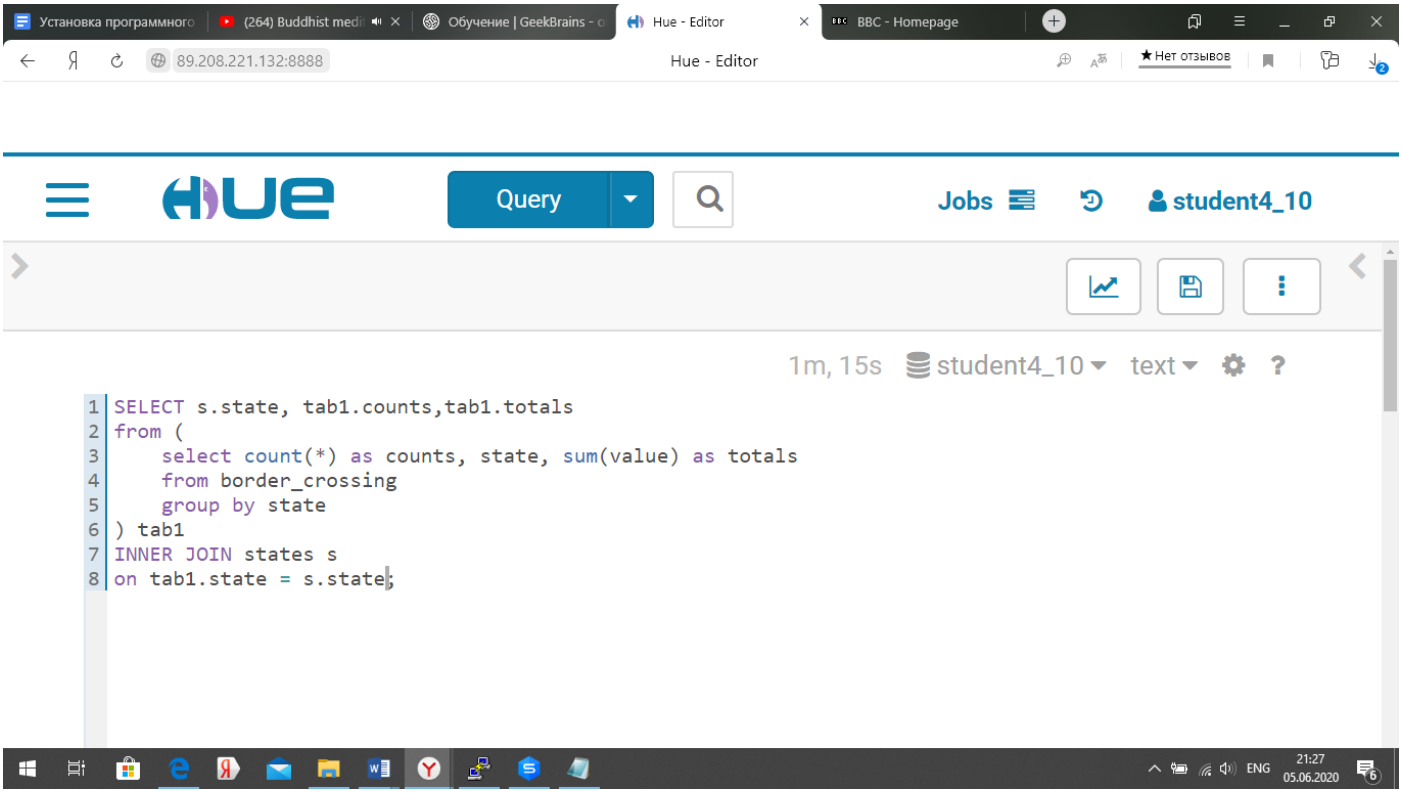
6.Сделать любой запрос по загруженным данным используя груповые и агрегатные функции.





1	Ohio	Personal Vehicles	5	373
2	Alaska	Personal Vehicles	1017	2504879
3	Idaho	Personal Vehicles	558	5188868
4	Montana	Personal Vehicles	3365	13266161
5	North Dakota	Personal Vehicles	5022	15765337
6	New Mexico	Personal Vehicles	558	16194134
7	Minnesota	Personal Vehicles	2013	25250476
8	Vermont	Personal Vehicles	1395	33800992
9	Maine	Personal Vehicles	3300	74402158

7. Сделать любой запрос по загруженным данным используя JOIN.




```
SELECT s.state, tab1.counts,tab1.totals
from (
    select count(*) as counts, state, sum(value) as totals
    from border_crossing
    group by state
) tab1
INNER JOIN states s
on tab1.state = s.state;
```