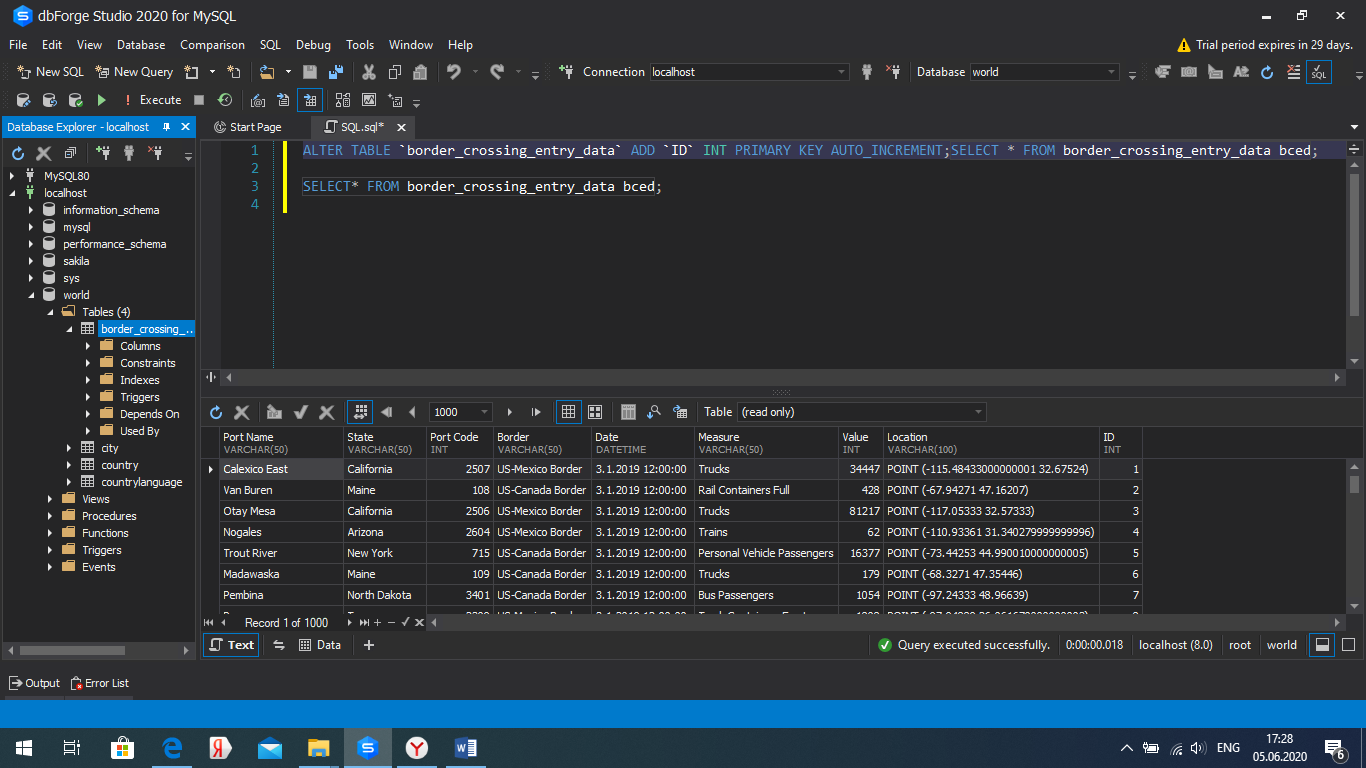
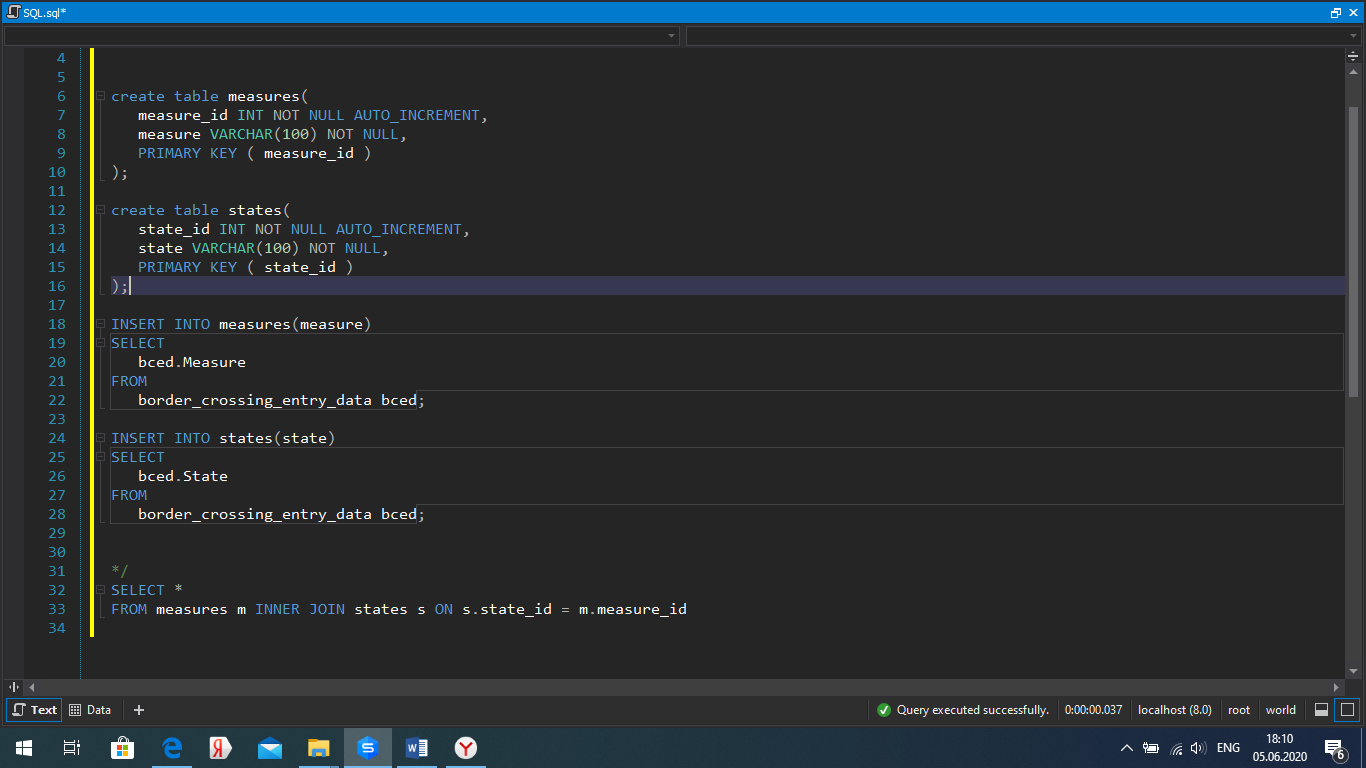
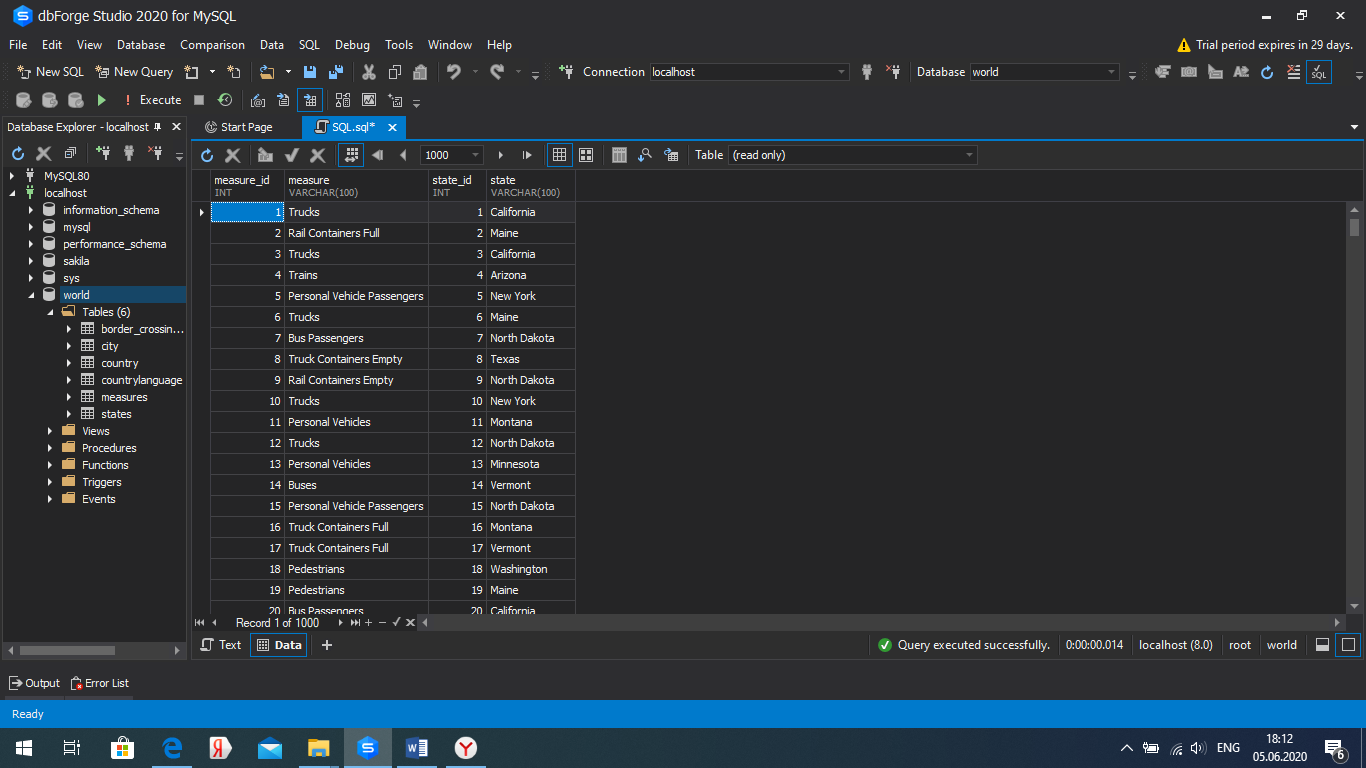
**2. Скачать любой датасет из списка ниже.**

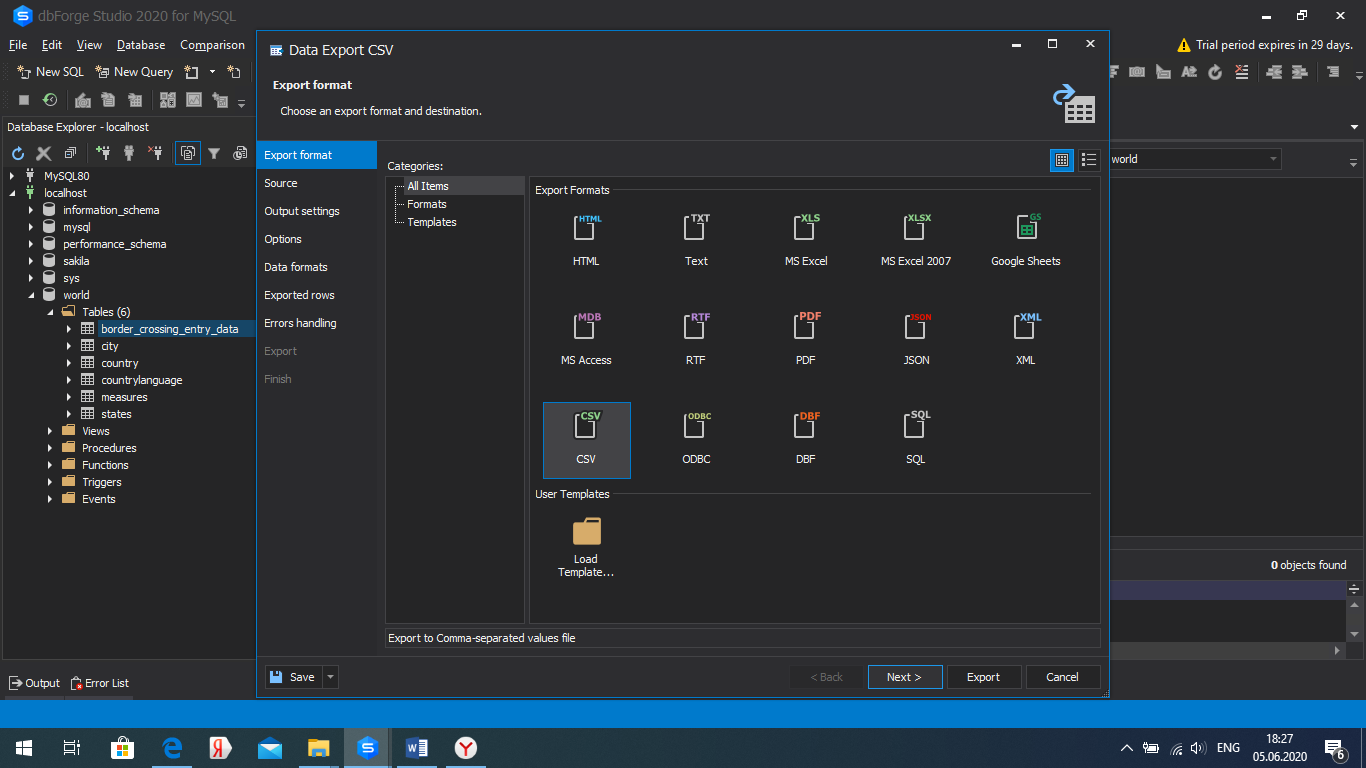
<https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions> <https://www.kaggle.com/datasnaek/youtube-new> <https://www.kaggle.com/akhilv11/border-crossing-entry-data> <https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games> <https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil>

Скопировали с Сайта KAGGLE csv file border\_crossing\_entry\_data и загрузили его в базу. Потом добавил колонку id



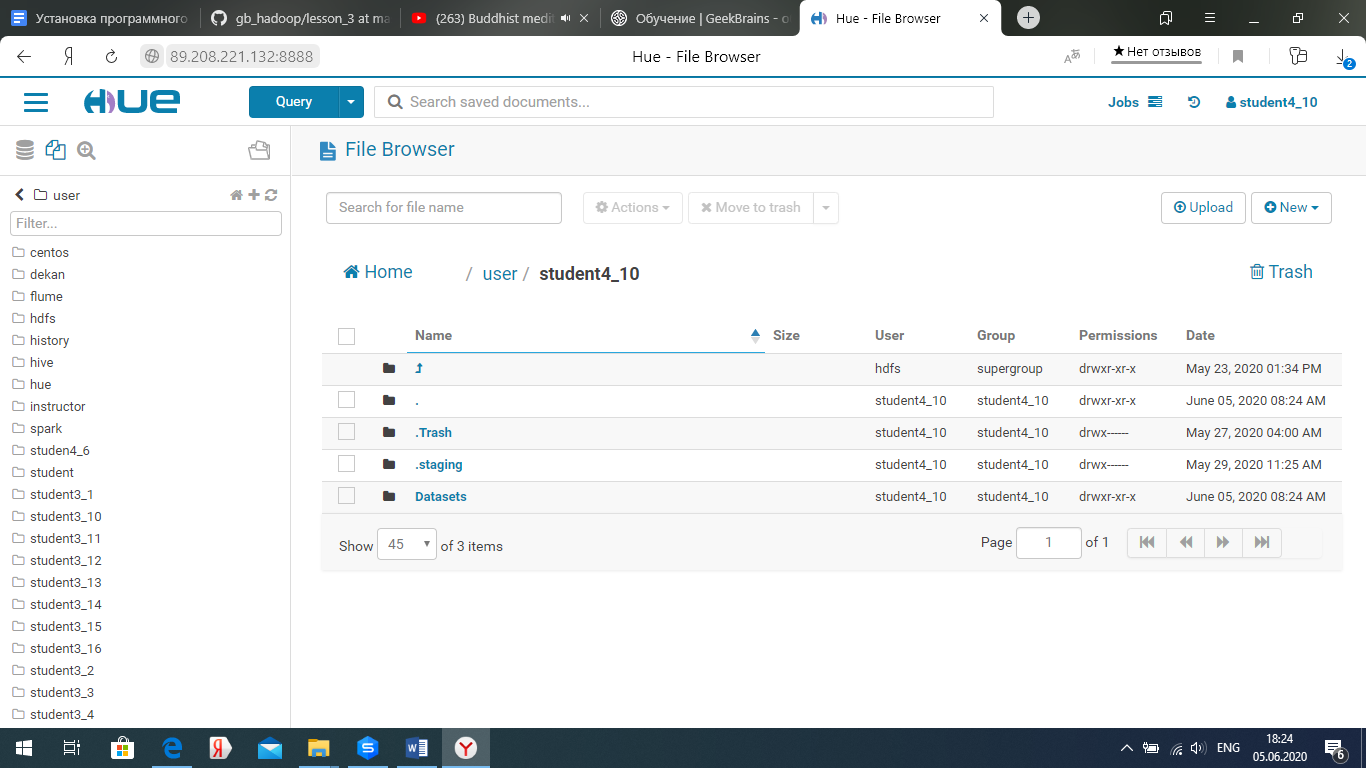
Создаем таблицы для таблиц measures и states, потом заливаем данные в эти таблицы. Сделаем select запрос для проверки данных



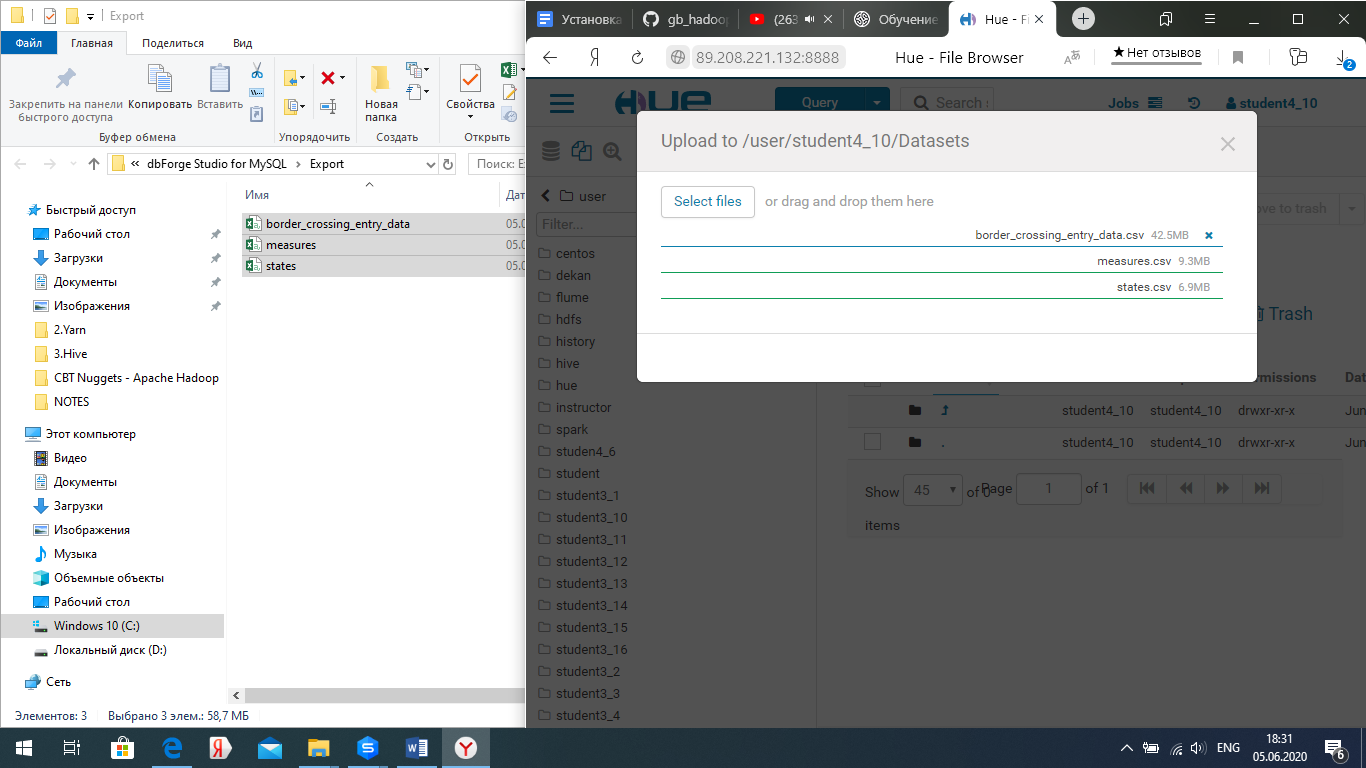


Экспортируем данные как CSV

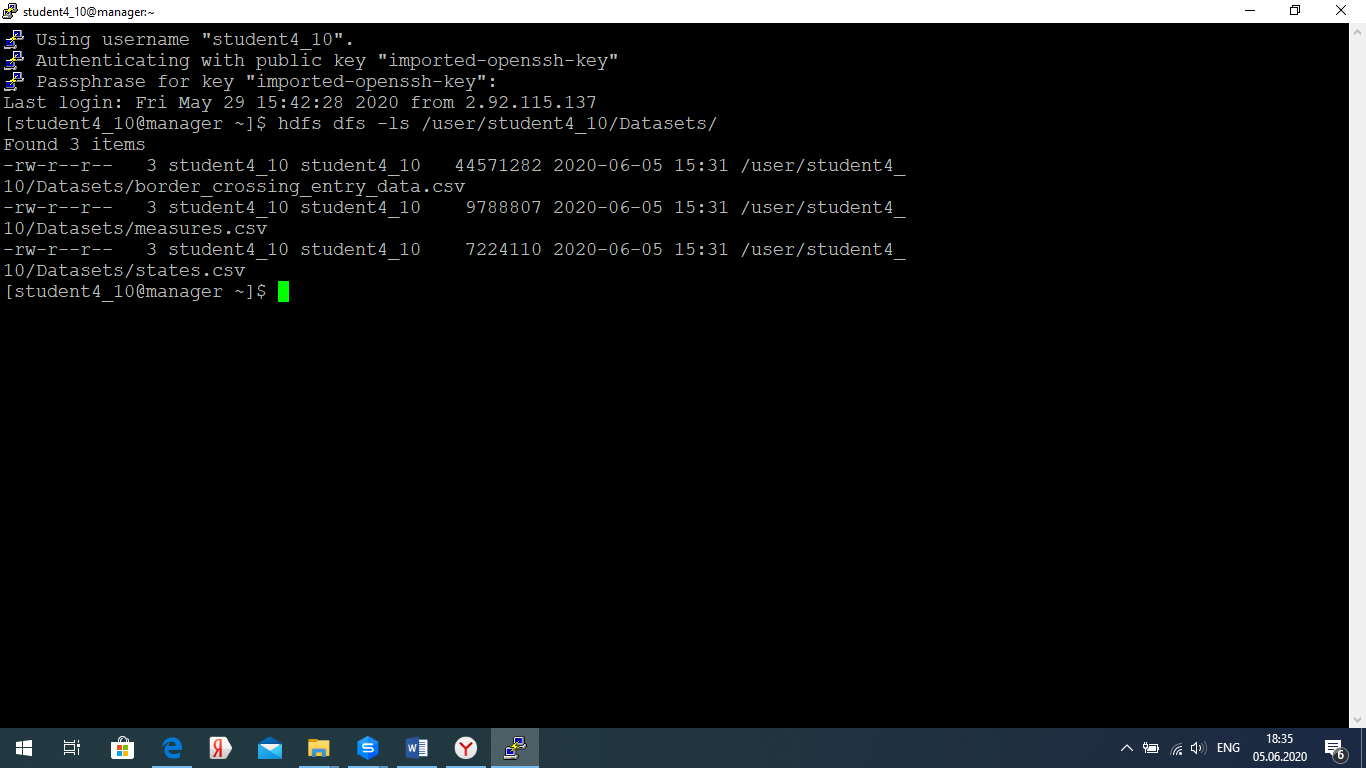
**3. Загрузить этот датасет в HDFS в свою домашнюю папку.**



Создаем Папку DATASET в HUE

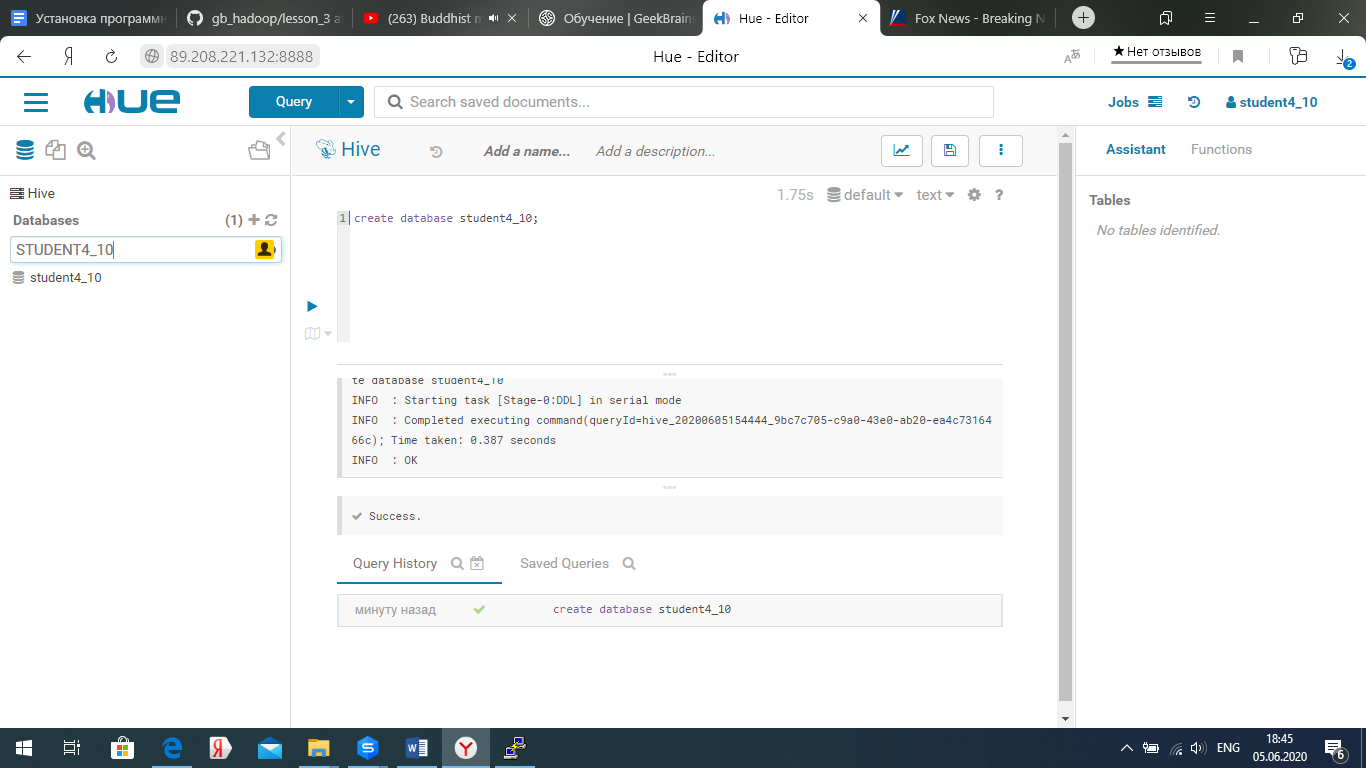


Загружаем файлы в HUE

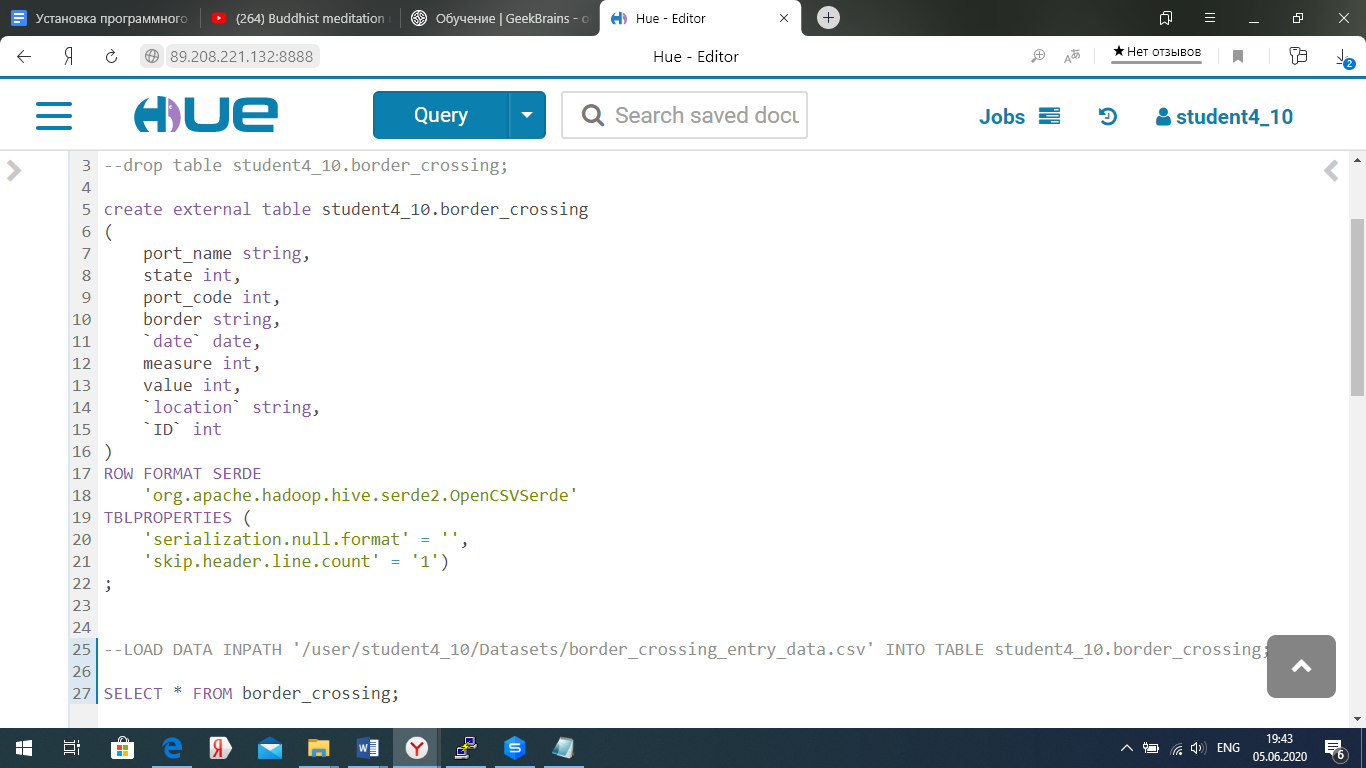


Через Putty перепроверяем что данные загружены в папку DATASET

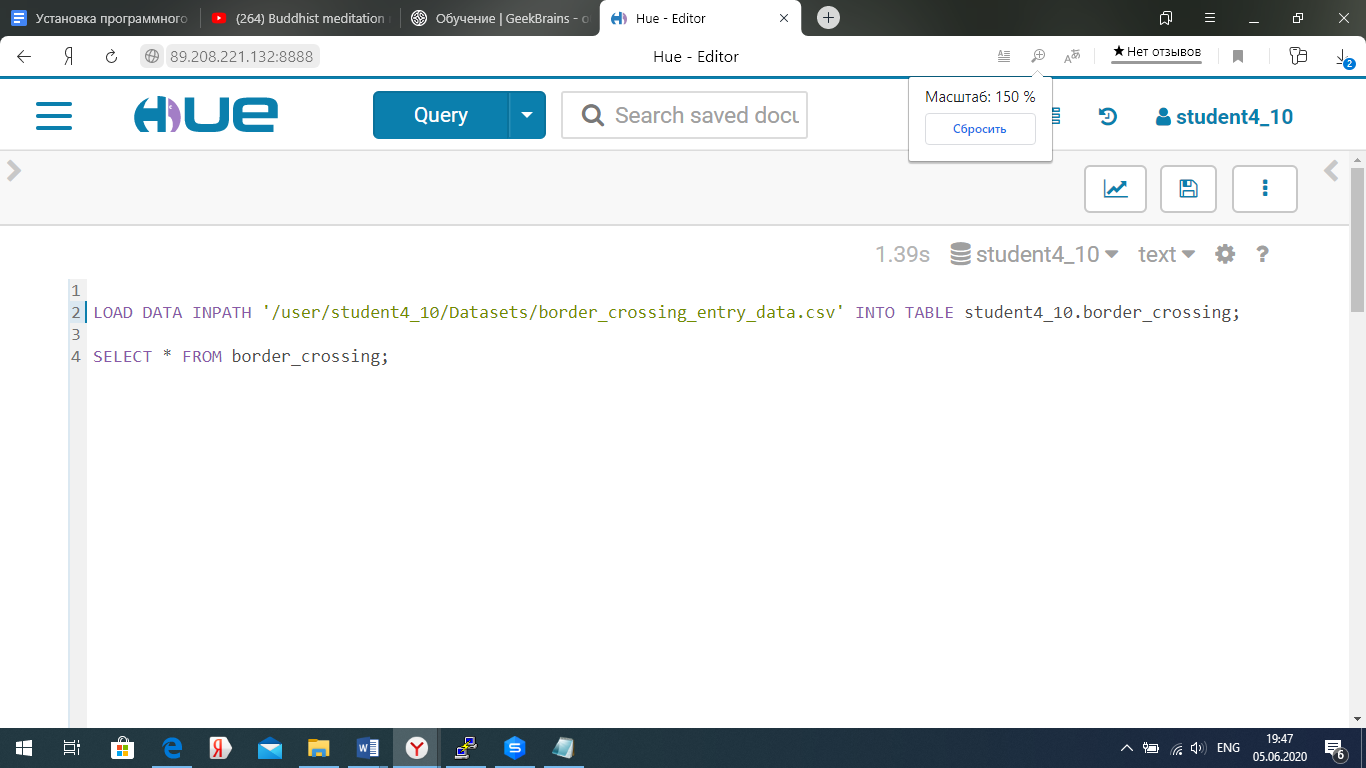
4. Создать собственную базу данных в HIVE



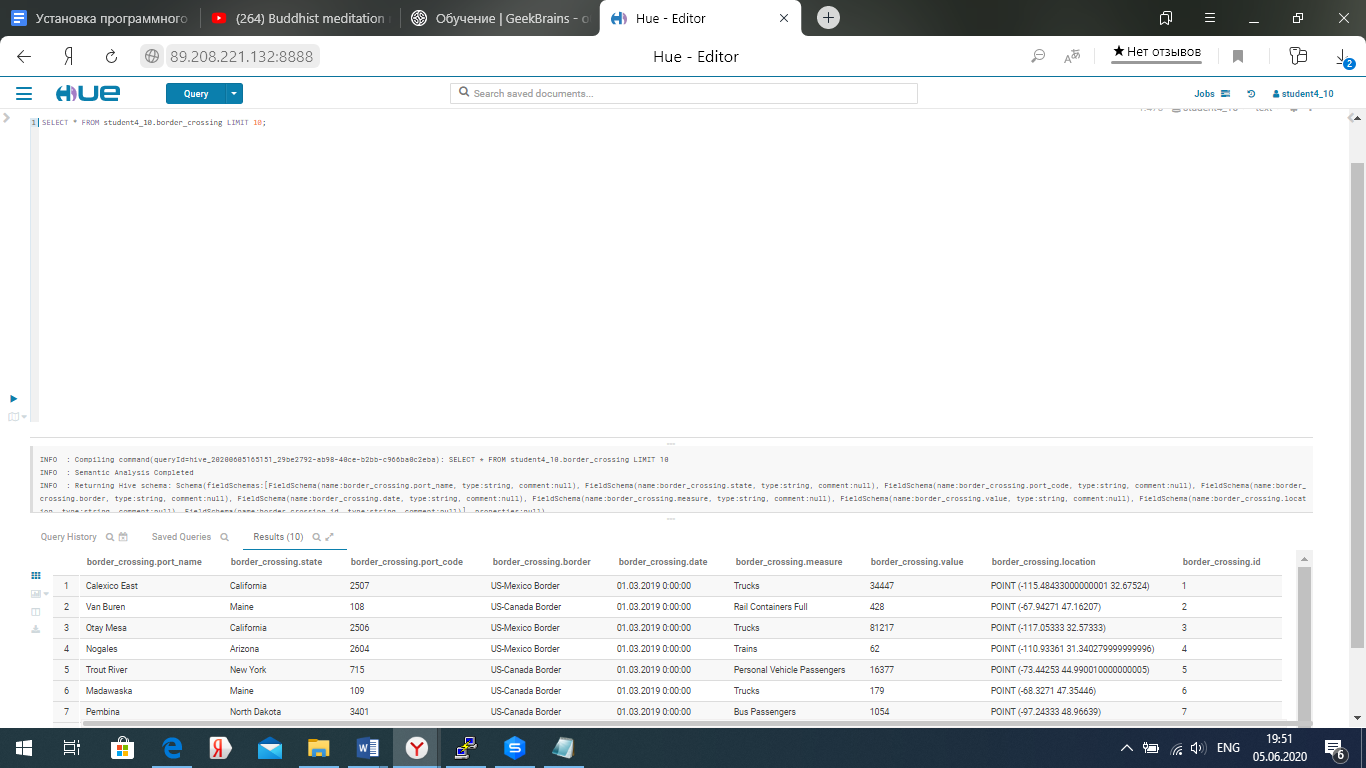
5. Создать EXTERNAL таблицы внутри базы данных с использованием всех загруженных файлов. Один файл – одна таблица.



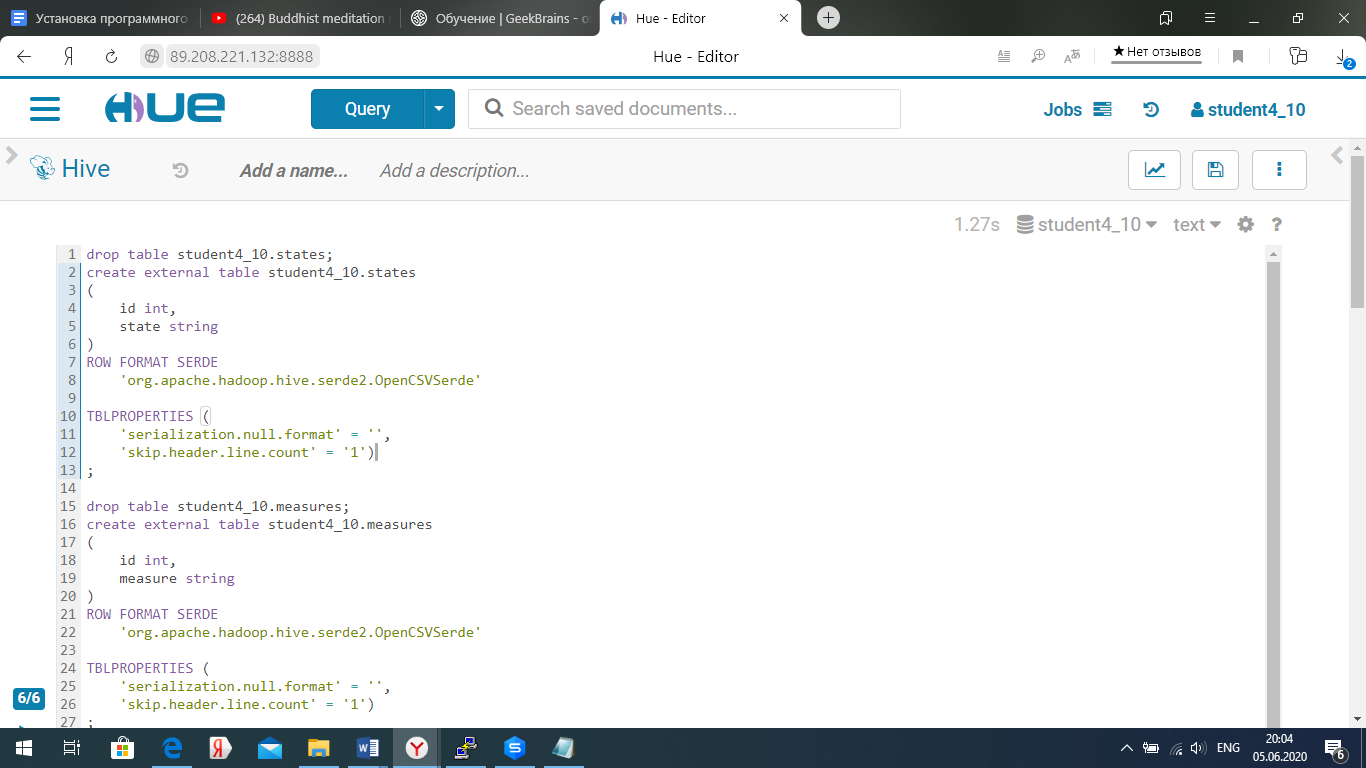
Создаем extertnal table под именем border\_crossing с CSV чтением

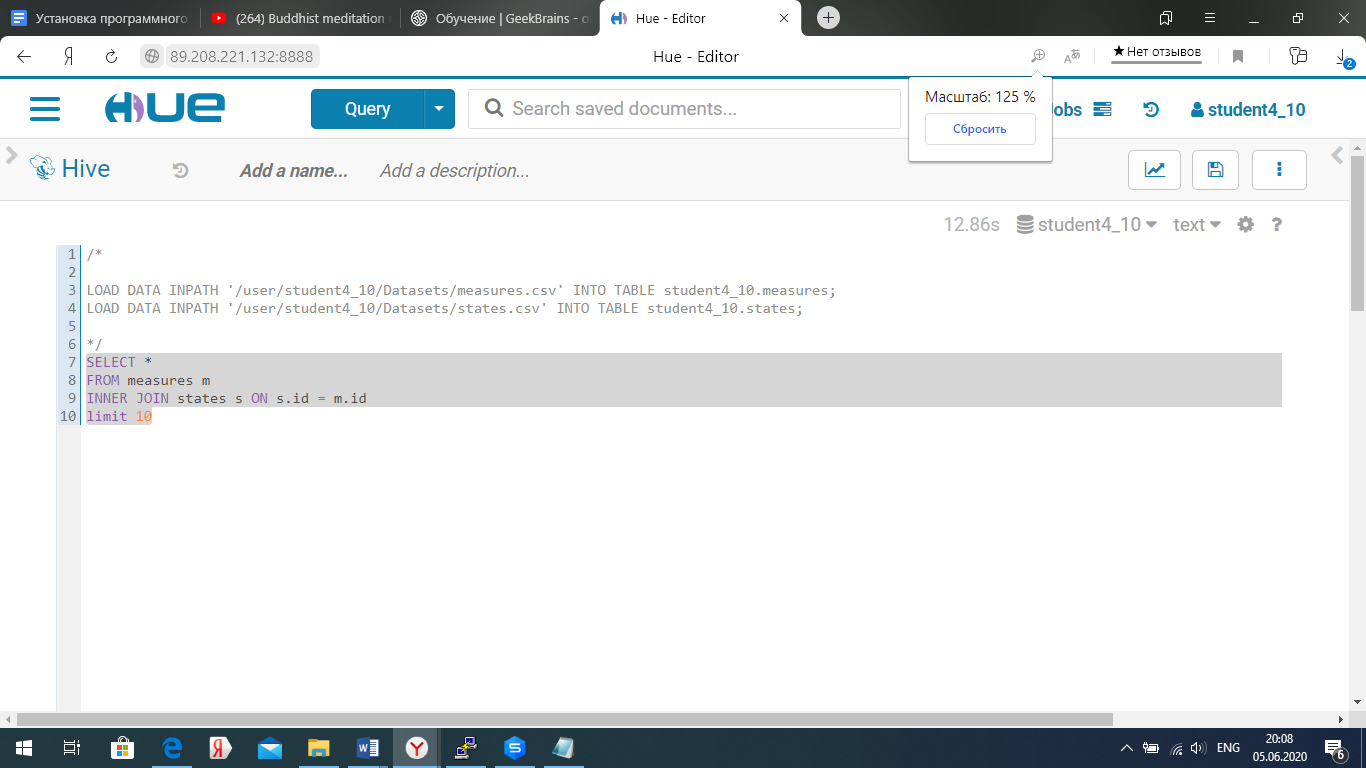


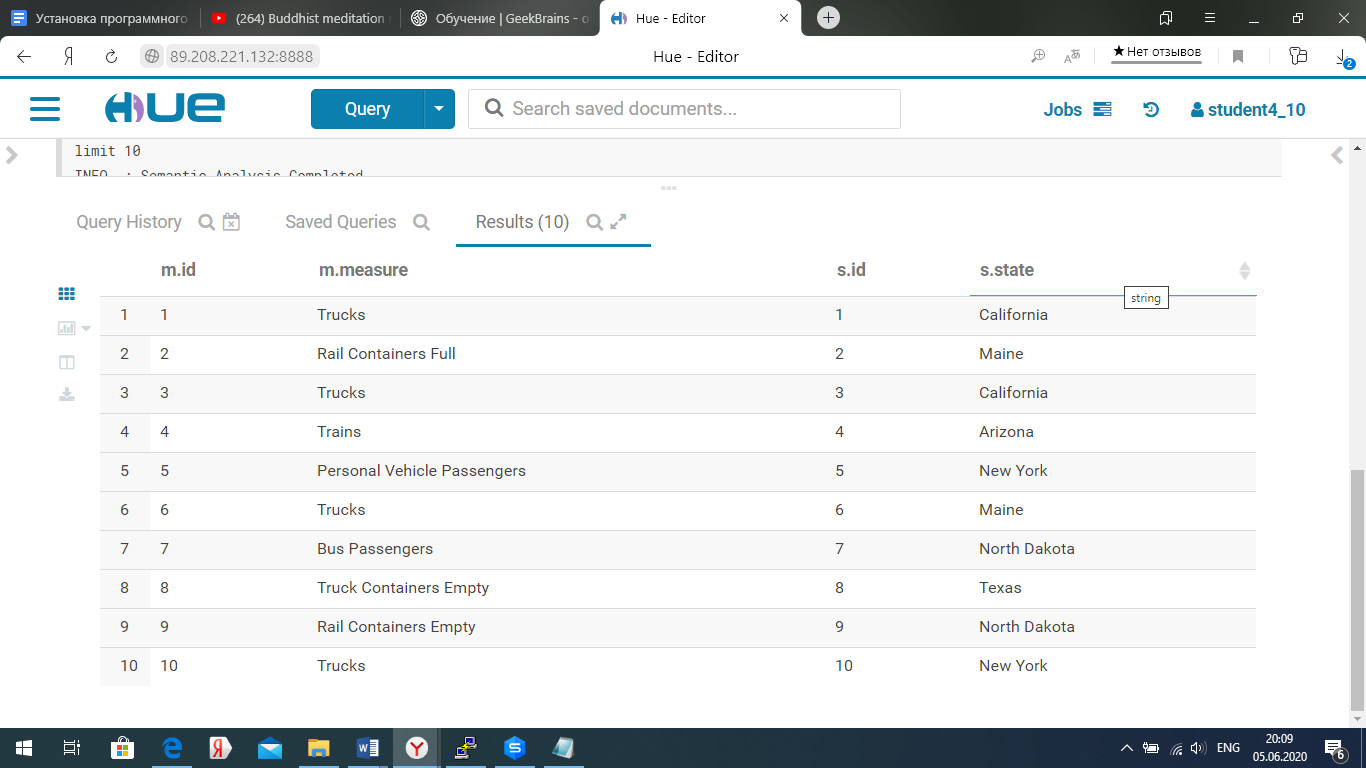
Загружаем данные из csv файла в базовую таблицу border\_crossing



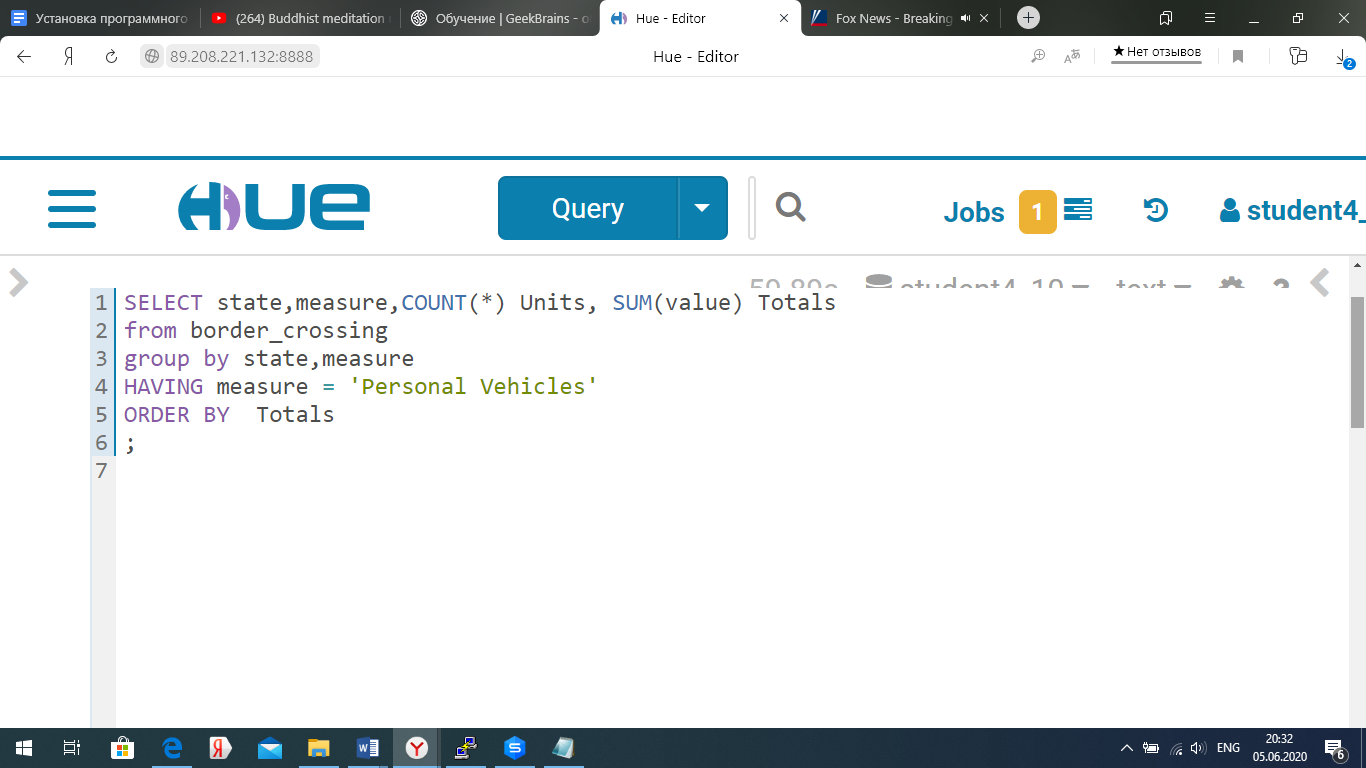
Проверяем данные

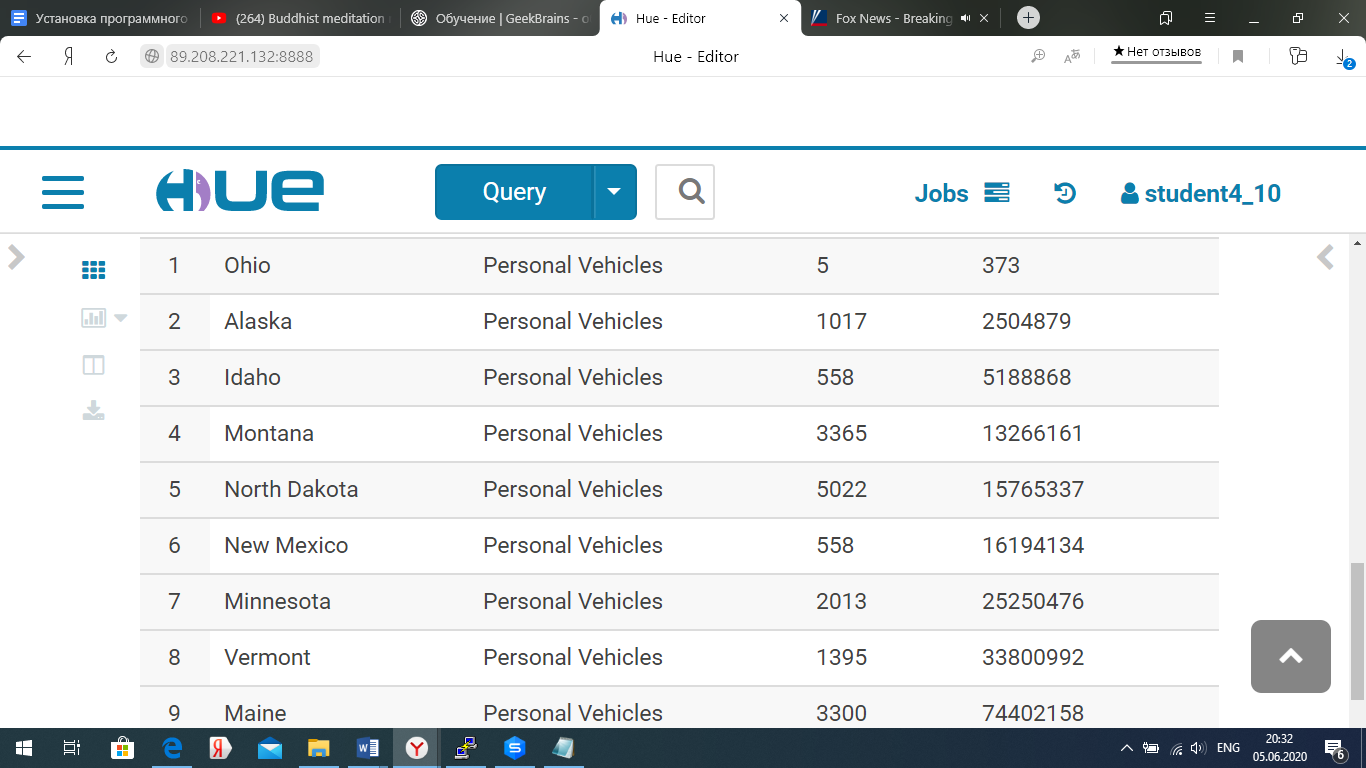
создаем внешние таблицы для states и measures

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

результат JOIN двух таблиц

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





7. Сделать любой запрос по загруженным данным используя JOIN.

