

## 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>

The screenshot shows the dbForge Studio 2020 for MySQL interface. The SQL editor contains the following query:

```
ALTER TABLE `border_crossing_entry_data` ADD `ID` INT PRIMARY KEY AUTO_INCREMENT; SELECT * FROM border_crossing_entry_data bced; SELECT * FROM border_crossing_entry_data bced;
```

The Database Explorer on the left shows the 'world' database with a table named 'border\_crossing\_entry\_data'. The results pane shows the following data:

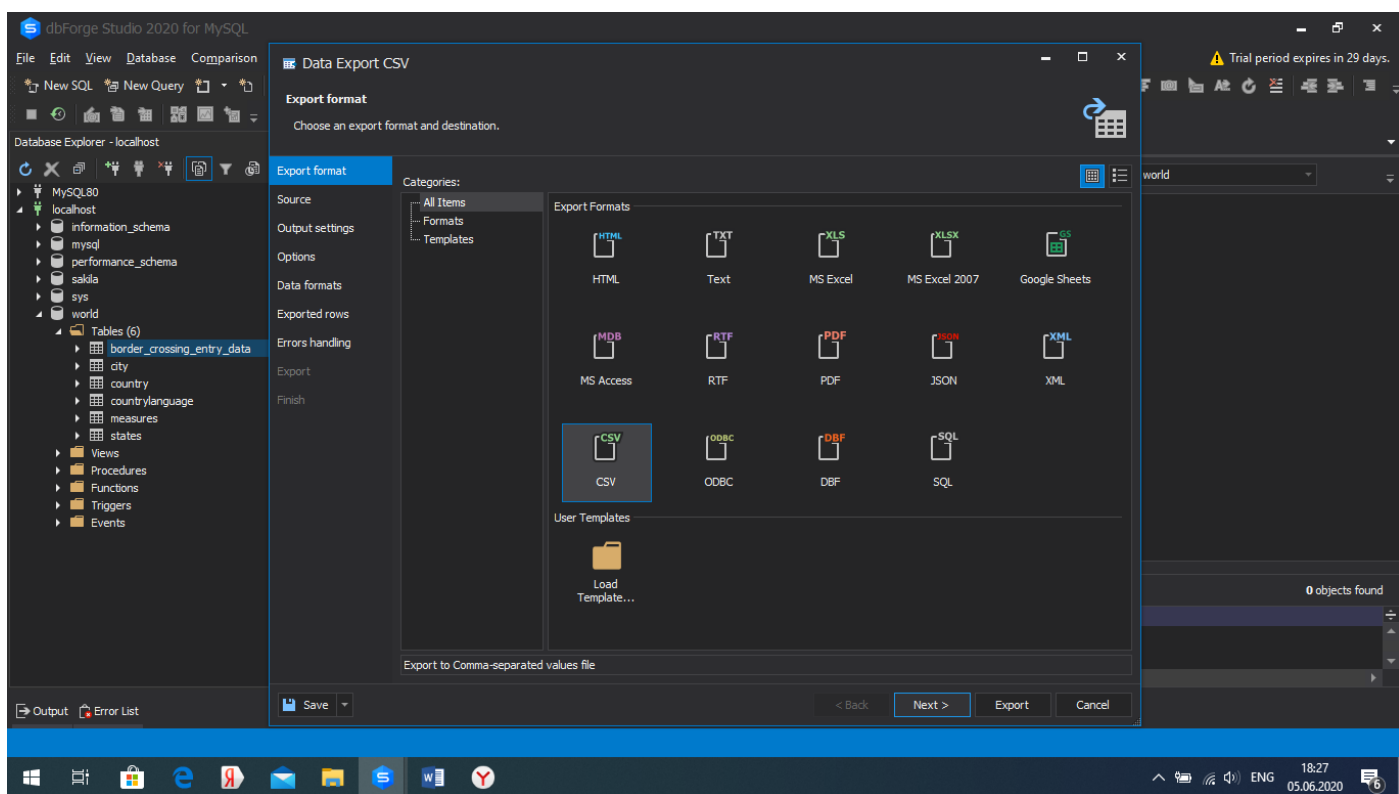
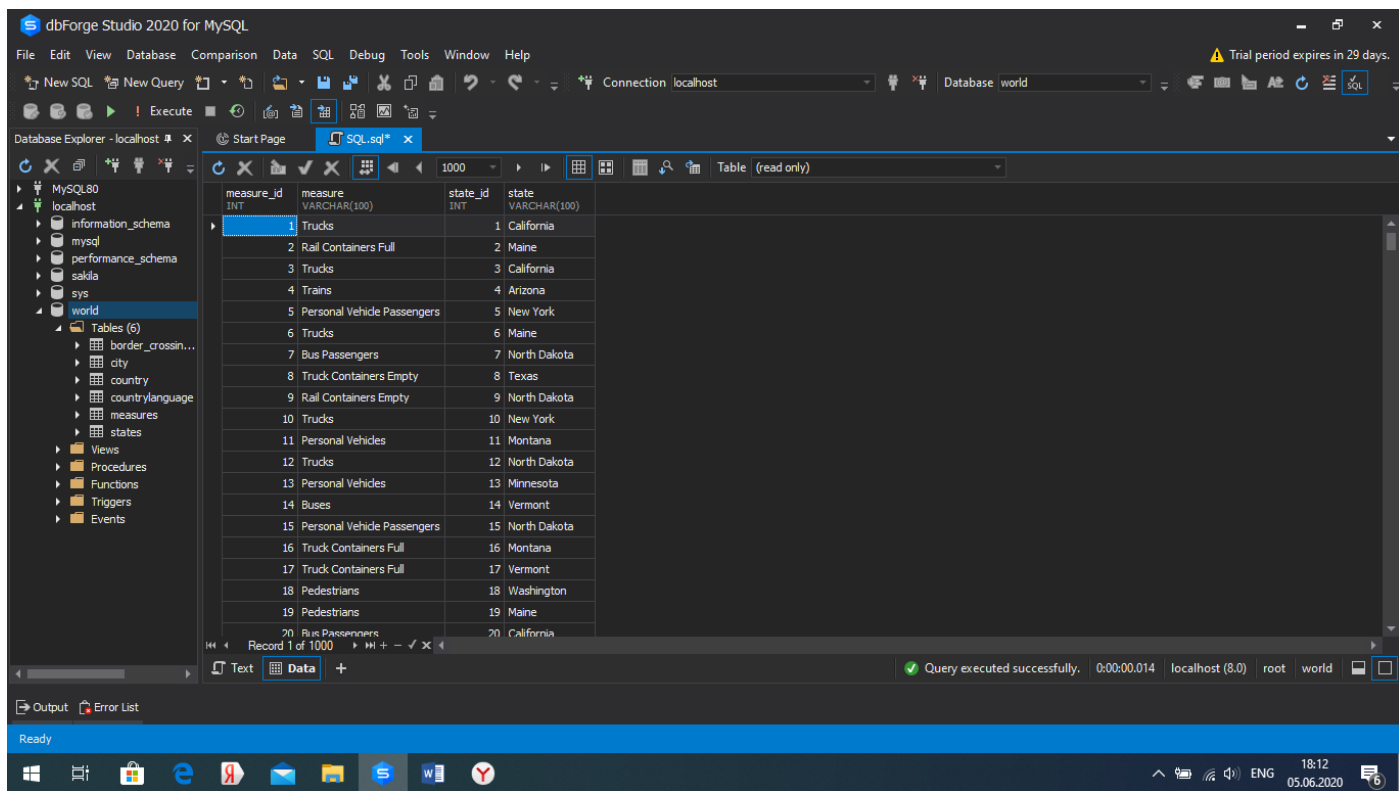
Port Name	State	Port Code	Border	Date	Measure	Value	Location	ID
Calxico East	California	2507	US-Mexico Border	3.1.2019 12:00:00	Trucks	34447	POINT (-115.484330000000001 32.67524)	1
Van Buren	Maine	108	US-Canada Border	3.1.2019 12:00:00	Rail Containers Full	428	POINT (-67.94271 47.16207)	2
Otay Mesa	California	2506	US-Mexico Border	3.1.2019 12:00:00	Trucks	81217	POINT (-117.05333 32.57333)	3
Nogales	Arizona	2604	US-Mexico Border	3.1.2019 12:00:00	Trains	62	POINT (-110.93361 31.340279999999996)	4
Trout River	New York	715	US-Canada Border	3.1.2019 12:00:00	Personal Vehicle Passengers	16377	POINT (-73.44253 44.9900100000000005)	5
Madawaska	Maine	109	US-Canada Border	3.1.2019 12:00:00	Trucks	179	POINT (-68.3271 47.35446)	6
Pembina	North Dakota	3401	US-Canada Border	3.1.2019 12:00:00	Bus Passengers	1054	POINT (-97.24333 48.96639)	7

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

The screenshot shows the dbForge Studio 2020 for MySQL interface with the following SQL queries:

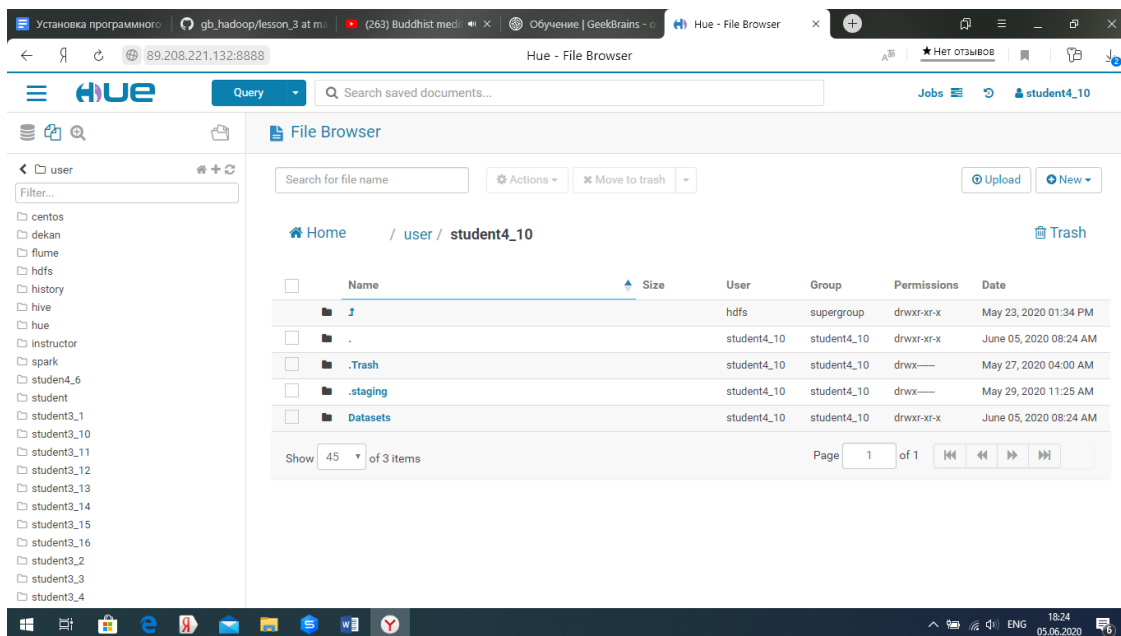
```
create table measures(  
    measure_id INT NOT NULL AUTO_INCREMENT,  
    measure VARCHAR(100) NOT NULL,  
    PRIMARY KEY ( measure_id )  
);  
  
create table states(  
    state_id INT NOT NULL AUTO_INCREMENT,  
    state VARCHAR(100) NOT NULL,  
    PRIMARY KEY ( state_id )  
);  
  
INSERT INTO measures(measure)  
SELECT  
    bced.Measure  
FROM  
    border_crossing_entry_data bced;  
  
INSERT INTO states(state)  
SELECT  
    bced.State  
FROM  
    border_crossing_entry_data bced;  
  
/*  
SELECT *  
FROM measures m INNER JOIN states s ON s.state_id = m.measure_id
```

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

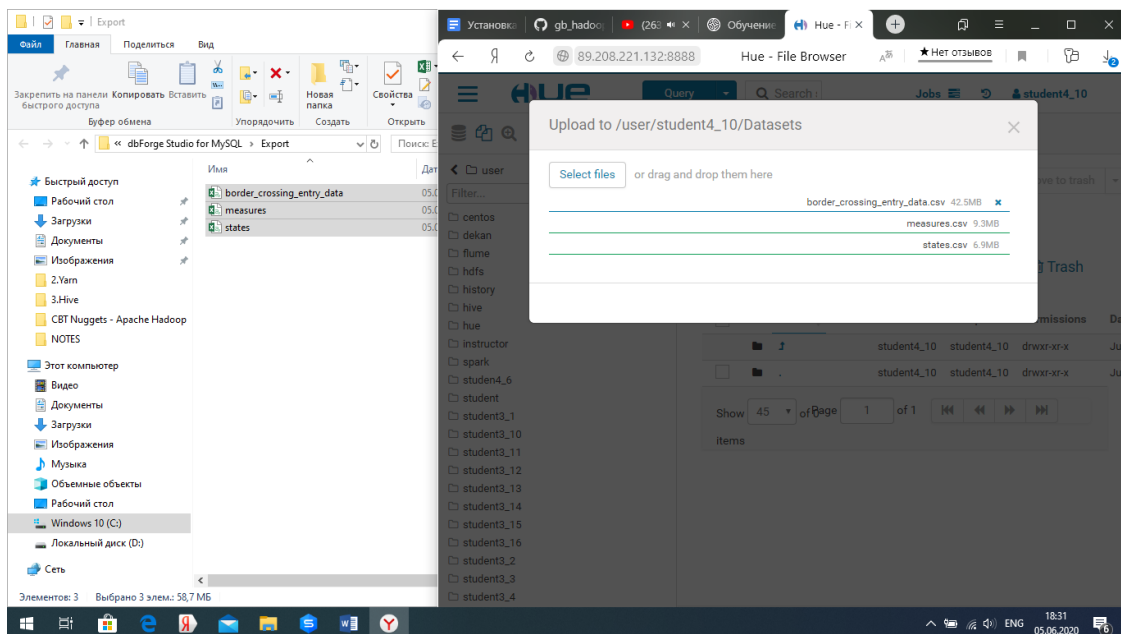


Экспортируем данные как CSV

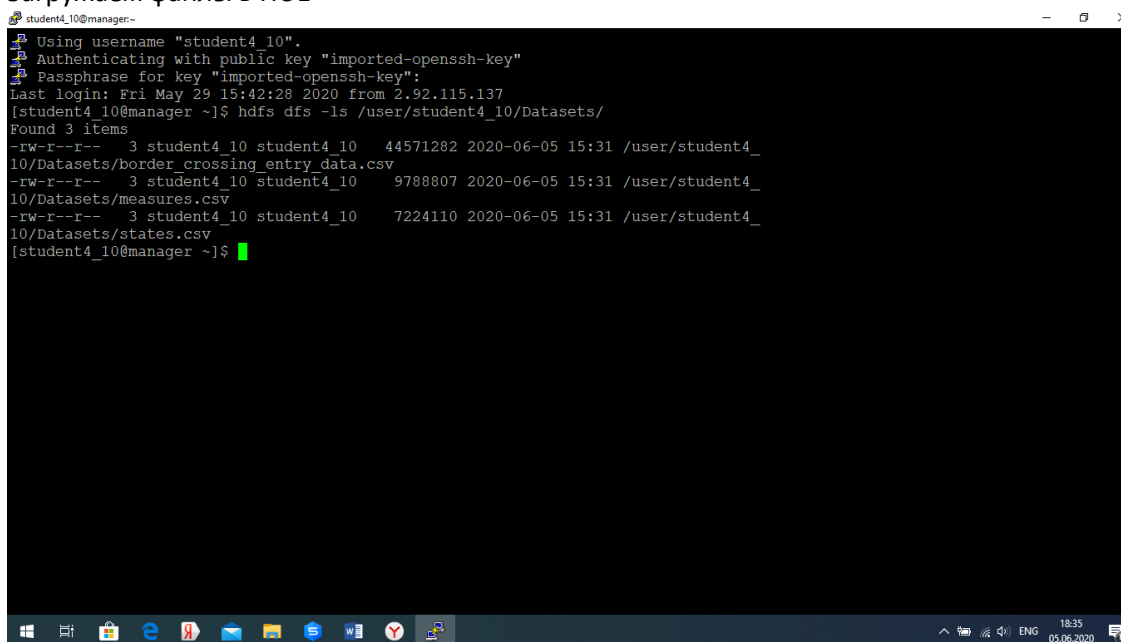
3. Загрузить этот датасет в HDFS в свою домашнюю папку.



## 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

The screenshot shows the Hue web interface with the URL `89.208.221.132:8888`. The user is logged in as `student4_10`. In the left sidebar, under 'Databases', the database `student4_10` is listed. The main editor area shows a query `create database student4_10;` which has been executed successfully. The execution log displays the following information:

```
te database student4_10
INFO : Starting task [Stage-0:DOL] in serial mode
INFO : Completed executing command(queryId=hive_20200605154444_9bc7c705-c9a0-43e0-ab20-ea4c7316466c); Time taken: 0.387 seconds
INFO : OK
```

Below the log, a 'Success.' message is shown. The 'Query History' section at the bottom lists the executed query: `create database student4_10` from 'минуту назад'.

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

The screenshot shows the Hue web interface with the URL `89.208.221.132:8888`. The user is logged in as `student4_10`. The main editor area shows a query that creates an external table and loads data from a CSV file:

```
3 --drop table student4_10.border_crossing;
4
5 create external table student4_10.border_crossing
6 (
7     port_name string,
8     state int,
9     port_code int,
10    border string,
11    `date` date,
12    measure int,
13    value int,
14    `location` string,
15    `ID` int
16 )
17 ROW FORMAT SERDE
18 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
19 TBLPROPERTIES (
20     'serialization.null.format' = '',
21     'skip.header.line.count' = '1')
22 ;
23
24
25 --LOAD DATA INPATH '/user/student4_10/Datasets/border_crossing_entry_data.csv' INTO TABLE student4_10.border_crossing;
26
27 SELECT * FROM border_crossing;
```

Создаем external table под именем border\_crossing с CSV чтением

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

Масштаб: 150 % Сбросить

student4\_10

1.39s student4\_10 text ?

```
1
2 LOAD DATA INPATH '/user/student4_10/Datasets/border_crossing_entry_data.csv' INTO TABLE student4_10.border_crossing;
3
4 SELECT * FROM border_crossing;
```

19:47 05.06.2020

Загружаем данные из csv файла в базовую таблицу border\_crossing

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

student4\_10

```
SELECT * FROM student4_10.border_crossing LIMIT 10;
```

INFO : Compiling command(queryIDhive\_38288685163151\_298e2792-a098-48ce-b2b0-c96d0a82ee8): SELECT \* FROM student4\_10.border\_crossing LIMIT 10  
INFO : Semantic Analysis Completed  
INFO : Returning hive schema: Schema(fieldsSchemas:[FieldSchema(name:border\_crossing.port\_name, type:string, comment:null), FieldSchema(name:border\_crossing.state, type:string, comment:null), FieldSchema(name:border\_crossing.port\_code, type:string, comment:null), FieldSchema(name:border\_crossing.border, type:string, comment:null), FieldSchema(name:border\_crossing.date, type:string, comment:null), FieldSchema(name:border\_crossing.measure, type:string, comment:null), FieldSchema(name:border\_crossing.value, type:string, comment:null), FieldSchema(name:border\_crossing.location, type:string, comment:null), FieldSchema(name:border\_crossing.id, type:string, comment:null)])

Query History Saved Queries Results (10) Q ✓

	border_crossing.port_name	border_crossing.state	border_crossing.port_code	border_crossing.border	border_crossing.date	border_crossing.measure	border_crossing.value	border_crossing.location	border_crossing.id
1	Calexico East	California	2507	US-Mexico Border	01.03.2019 0:00:00	Trucks	34447	POINT (-115.48433000000001 32.67524)	1
2	Van Buren	Maine	108	US-Canada Border	01.03.2019 0:00:00	Rail Containers Full	428	POINT (-67.94271 47.16207)	2
3	Otay Mesa	California	2506	US-Mexico Border	01.03.2019 0:00:00	Trucks	81217	POINT (-117.05333 32.57333)	3
4	Nogales	Arizona	2604	US-Mexico Border	01.03.2019 0:00:00	Trains	62	POINT (-110.93361 31.340279999999996)	4
5	Trout River	New York	715	US-Canada Border	01.03.2019 0:00:00	Personal Vehicle Passengers	16377	POINT (-73.44253 44.990010000000005)	5
6	Madawaska	Maine	109	US-Canada Border	01.03.2019 0:00:00	Trucks	179	POINT (-68.3271 47.35446)	6
7	Pembina	North Dakota	3401	US-Canada Border	01.03.2019 0:00:00	Bus Passengers	1054	POINT (-97.24333 48.96639)	7

19:51 05.06.2020

Проверяем данные

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

Jobs student4\_10

Hive Add a name... Add a description...

1.27s student4\_10 text ?

```
1 drop table student4_10.states;
2 create external table student4_10.states
3 (
4     id int,
5     state string
6 )
7 ROW FORMAT SERDE
8     'org.apache.hadoop.hive.serde2.OpenCSVSerde'
9
10 TBLPROPERTIES (
11     'serialization.null.format' = '',
12     'skip.header.line.count' = '1')
13 ;
14
15 drop table student4_10.measures;
16 create external table student4_10.measures
17 (
18     id int,
19     measure string
20 )
21 ROW FORMAT SERDE
22     'org.apache.hadoop.hive.serde2.OpenCSVSerde'
23
24 TBLPROPERTIES (
25     'serialization.null.format' = '',
26     'skip.header.line.count' = '1')
27 ;
```

6/6

создаем внешние таблицы для states и measures

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

Jobs student4\_10

Hive Add a name... Add a description...

12.86s student4\_10 text ?

```
1 /*
2
3 LOAD DATA INPATH '/user/student4_10/Datasets/measures.csv' INTO TABLE student4_10.measures;
4 LOAD DATA INPATH '/user/student4_10/Datasets/states.csv' INTO TABLE student4_10.states;
5
6 */
7 SELECT *
8 FROM measures m
9 INNER JOIN states s ON s.id = m.id
10 limit 10
```

Масштаб: 125 %  
Сбросить

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

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

Jobs student4\_10

limit 10

Query History Saved Queries Results (10)

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

Windows taskbar: 20:09 05.06.2020

результат JOIN двух таблиц

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

Установка программного (264) Buddhist meditation Обучение | GeekBrains - Hue - Editor

89.208.221.132:8888 Hue - Editor

Jobs 1 student4\_

```

1 SELECT state,measure,COUNT(*) Units, SUM(value) Totals
2 from border_crossing
3 group by state,measure
4 HAVING measure = 'Personal Vehicles'
5 ORDER BY Totals
6 ;
7

```

Windows taskbar: 20:32 05.06.2020

Установка программного (264) Buddhist meditation Обучение | GeekBrains Hue - Editor Fox News - Breaking

89.208.221.132:8888 Hue - Editor Нет отзывов

Menu Hue Query Jobs student4\_10

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

Windows taskbar: 20:32 05.06.2020

7. Сделать любой запрос по загруженным данным используя JOIN.

Установка программного (264) Buddhist meditation Обучение | GeekBrains Hue - Editor BBC - Homepage

89.208.221.132:8888 Hue - Editor Нет отзывов

Menu Hue Query Jobs student4\_10

1m, 15s student4\_10 text

```

1 SELECT s.state, tab1.counts, tab1.totals
2 from (
3     select count(*) as counts, state, sum(value) as totals
4     from border_crossing
5     group by state
6 ) tab1
7 INNER JOIN states s
8 on tab1.state = s.state;
```

Windows taskbar: 21:27 05.06.2020