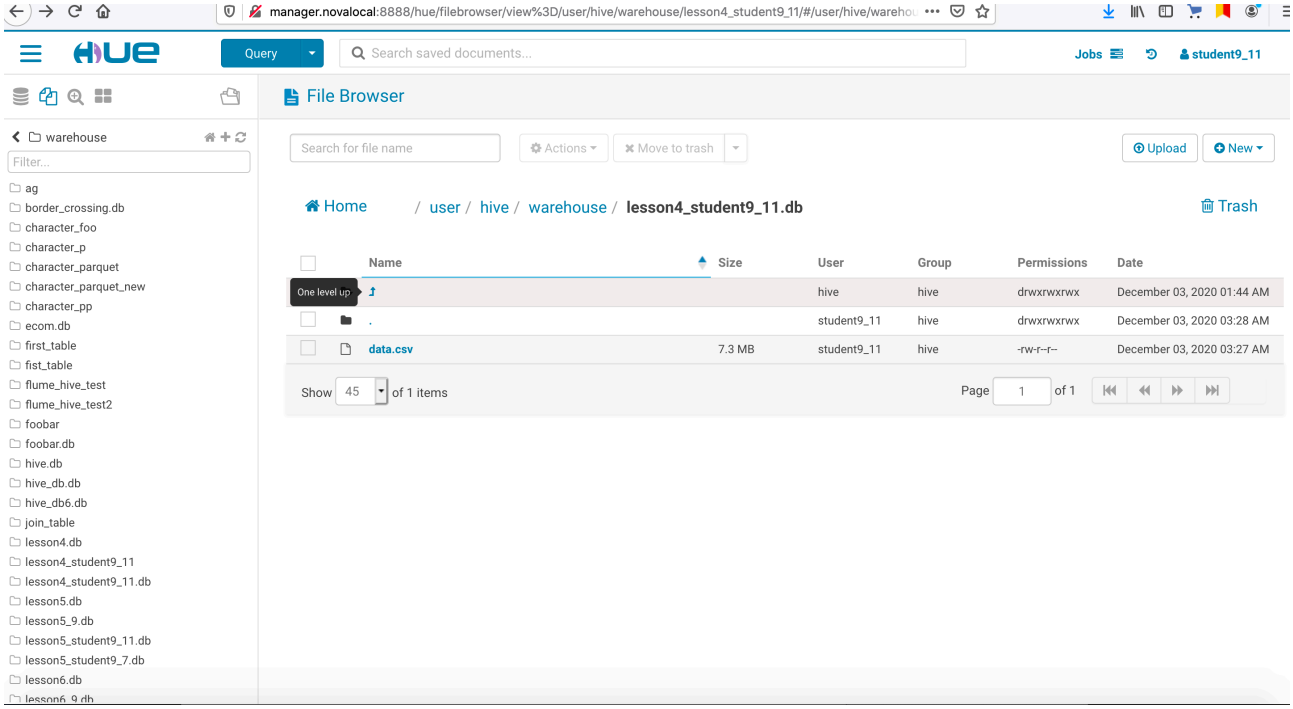
Скачать любой датасет из списка ниже с сайта Kaggle.com (достаточно большой)
Загрузить этот датасет в HDFS в свою домашнюю папку



Создать собственную базу данных в HIVE

```
create external table lesson4_student9_11.engagement
(source_id int,
source_name string,
author string,
title string,
description string,
url string,
url_to_image string,
published_at date,
content string,
top_article boolean,
engagement_reaction_count int,
engagement_comment_count int,
engagement_share_count int,
engagement_comment_plugin_count int
)

ROW FORMAT serde 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'

WITH SERDEPROPERTIES ('field.delim' = ',')

location "/user/hive/warehouse/lesson4_student9_11.db/"

tblproperties("skip.header.line.count"="1")
;

--
-- drop table lessons9_4.customers;

select * from lesson4_student9_11.engagement limit 100;
```

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

```
INFO  : Executing command(queryId=hive_20201203112929_81773760-c65f-4be4-b1c2-a817ca9ae817): select * from lesson4
_student9_11.engagement limit 100
INFO  : Completed executing command(queryId=hive_20201203112929_81773760-c65f-4be4-b1c2-a817ca9ae817); Time taken:
0.001 seconds
INFO  : OK
```

Query History | Saved Queries | Results (100+)

	engagement.source_id	engagement.source_name
1	0	reuters
2	1	the-irish-times
3	NULL	0.0
4	2	the-irish-times
5	NULL	including the designers close frie... [+2156 chars]"
6	3	al-jazeera-english
7	NULL	0.0

Сделать любой отчет по загруженным данным используя групповые и агрегатные функции.

- среднее по количество пересолов новостей

SELECT avg(engagement_share_count) from lesson4_student9_11.engagement;

```
32 SELECT avg(engagement_share_count) from lesson4_student9_11.engagement;
```

```
INFO  : The url to track the job: http://manager.novalocal:8088/proxy/application_1604591611666_0904/
INFO  : Starting Job = job_1604591611666_0904, Tracking URL = http://manager.novaloca l:8088/proxy/application_1604
591611666_0904/
INFO  : Kill Command = /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop/bin/hadoop job  -kill job_1604
591611666_0904
```

job_1604591611666_0904

Query History | Saved Queries | Results (1)

	_c0
1	112.90034533793784

Сделать любой отчет по загруженным данным используя JOIN.

create table lesson4_student9_11.engagement_new AS

select source_id, source_name from lesson4_student9_11.engagement;

create table lesson4_student9_11.engagement_new_2 AS

select source_id, author from lesson4_student9_11.engagement;

SELECT t1.source_id, t1.source_name, t2.author from lesson4_student9_11.engagement_new t1

JOIN lesson4_student9_11.engagement_new_2 t2

ON t1.source_id = t2.source_id

```
create table lesson4_student9_11.engagement_new AS
select source_id, source_name from lesson4_student9_11.engagement;

create table lesson4_student9_11.engagement_new_2 AS
select source_id, author from lesson4_student9_11.engagement;

SELECT t1.source_id, t1.source_name, t2.author from lesson4_student9_11.engagement_new t1
JOIN lesson4_student9_11.engagement_new_2 t2
ON t1.source_id = t2.source_id
```

```
INFO  : Compiling command(queryId=hive_20201203115454_c54dd36f-9414-42a6-ae1e-ef1196d3ae67): SELECT t1.source_id,
t1.source_name, t2.author from lesson4_student9_11.engagement_new t1
JOIN lesson4_student9_11.engagement_new_2 t2
ON t1.source_id = t2.source_id
INFO  : Semantic Analysis Completed
```

job_1604591611666_0907

Query History | Saved Queries | Results (100+)

	t1.source_id	t1.source_name	t2.author
1	0	reuters	Reuters
2	1	the-irish-times	The Irish Times
3	2	the-irish-times	The Irish Times