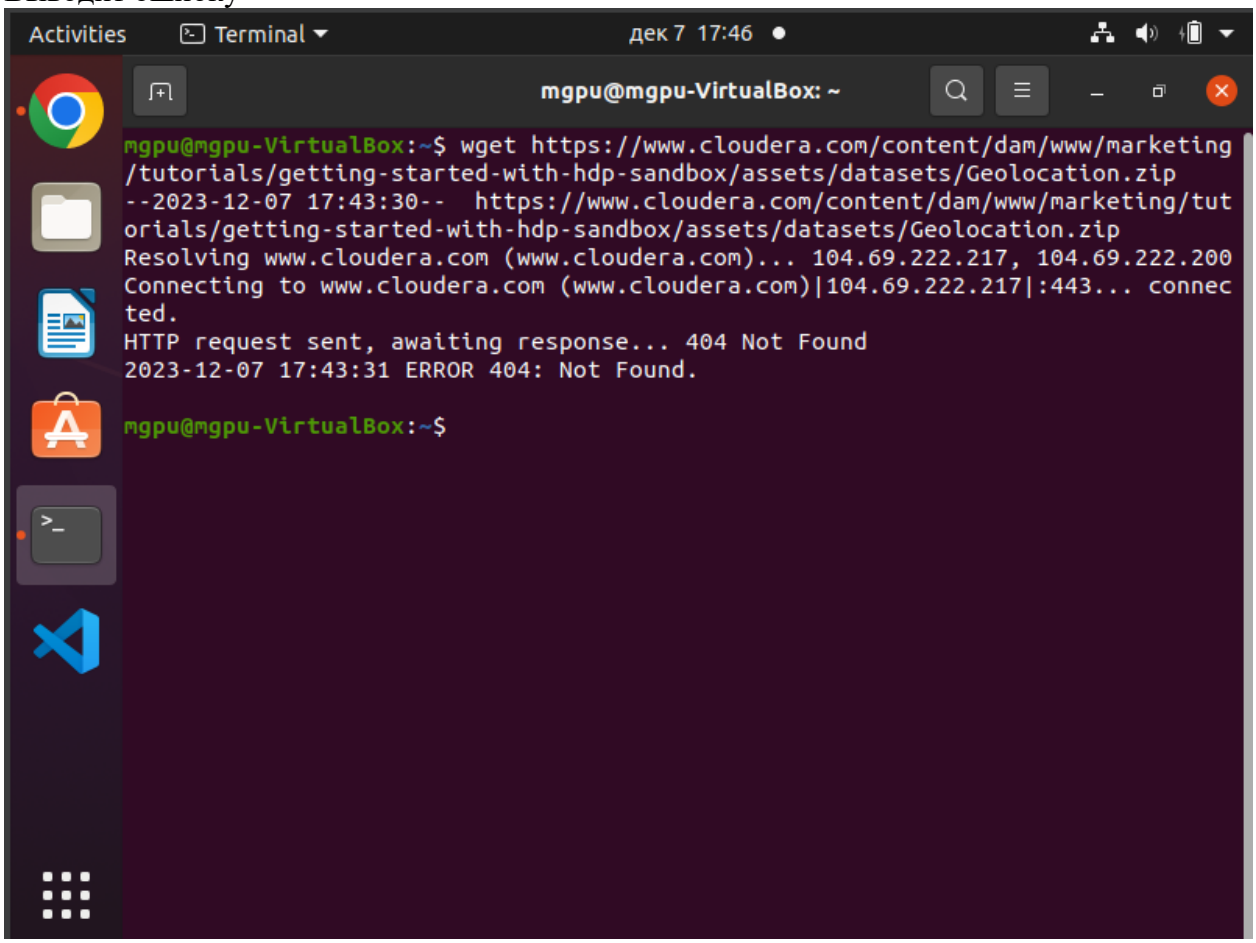
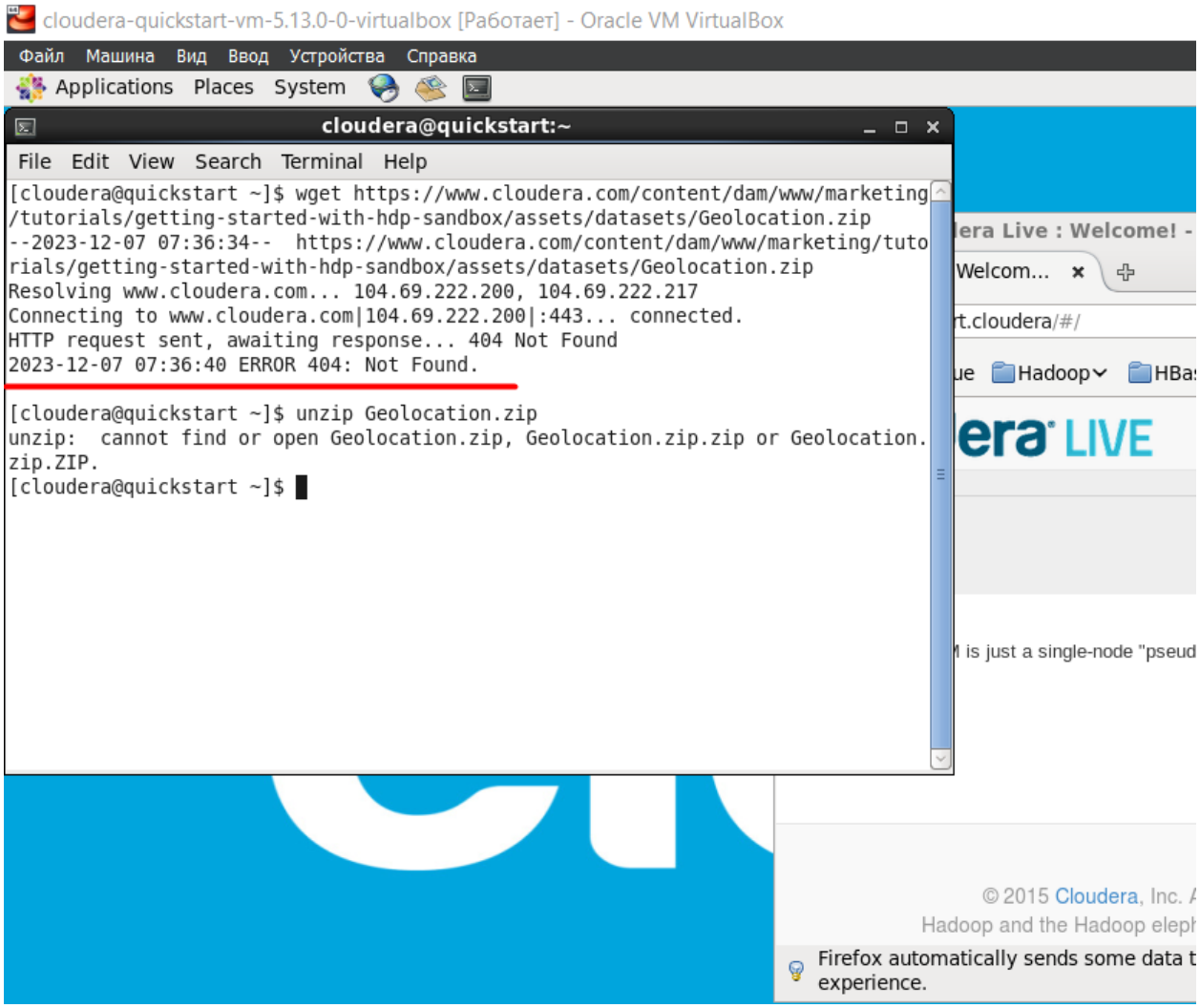
Установка cloudera
Выводит ошибку

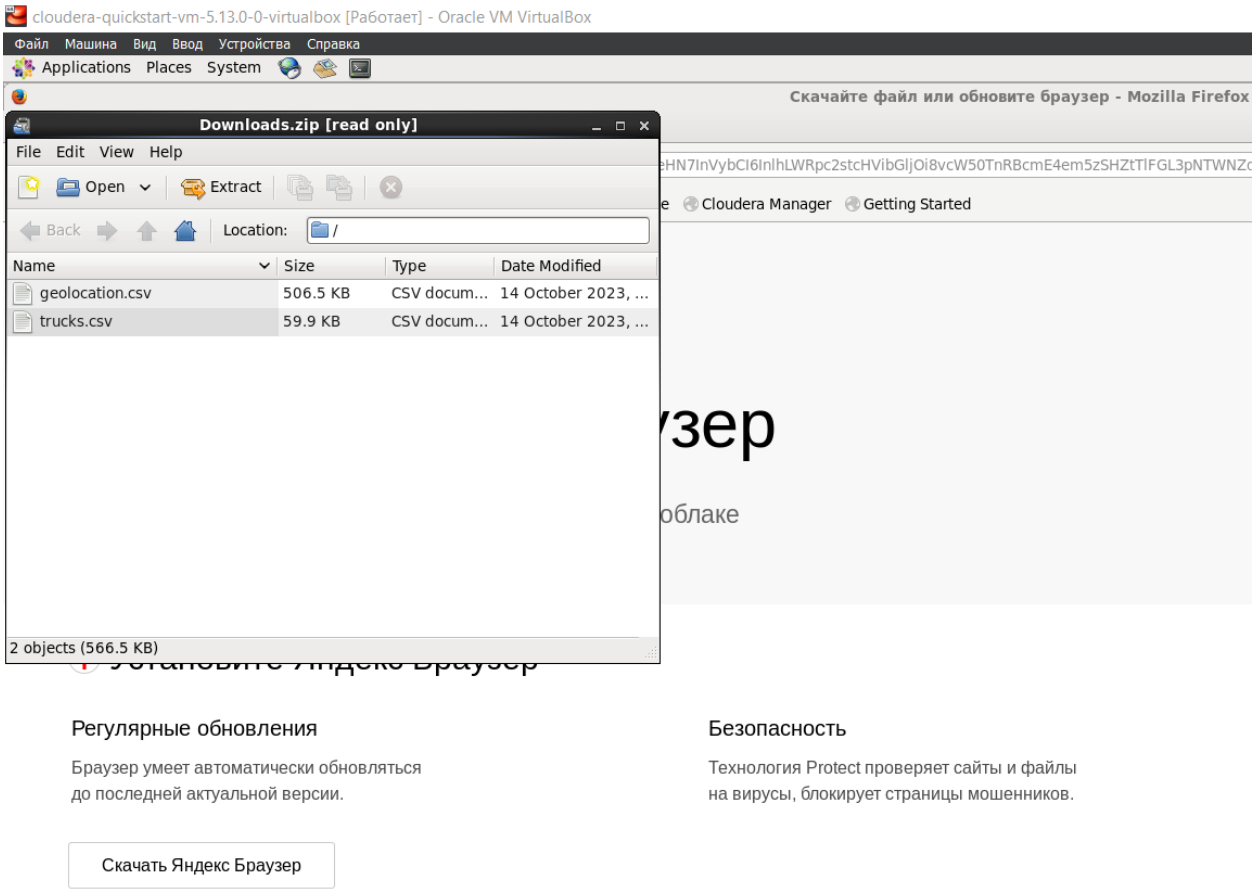


Пробуем установить по-другому(на другой вм)

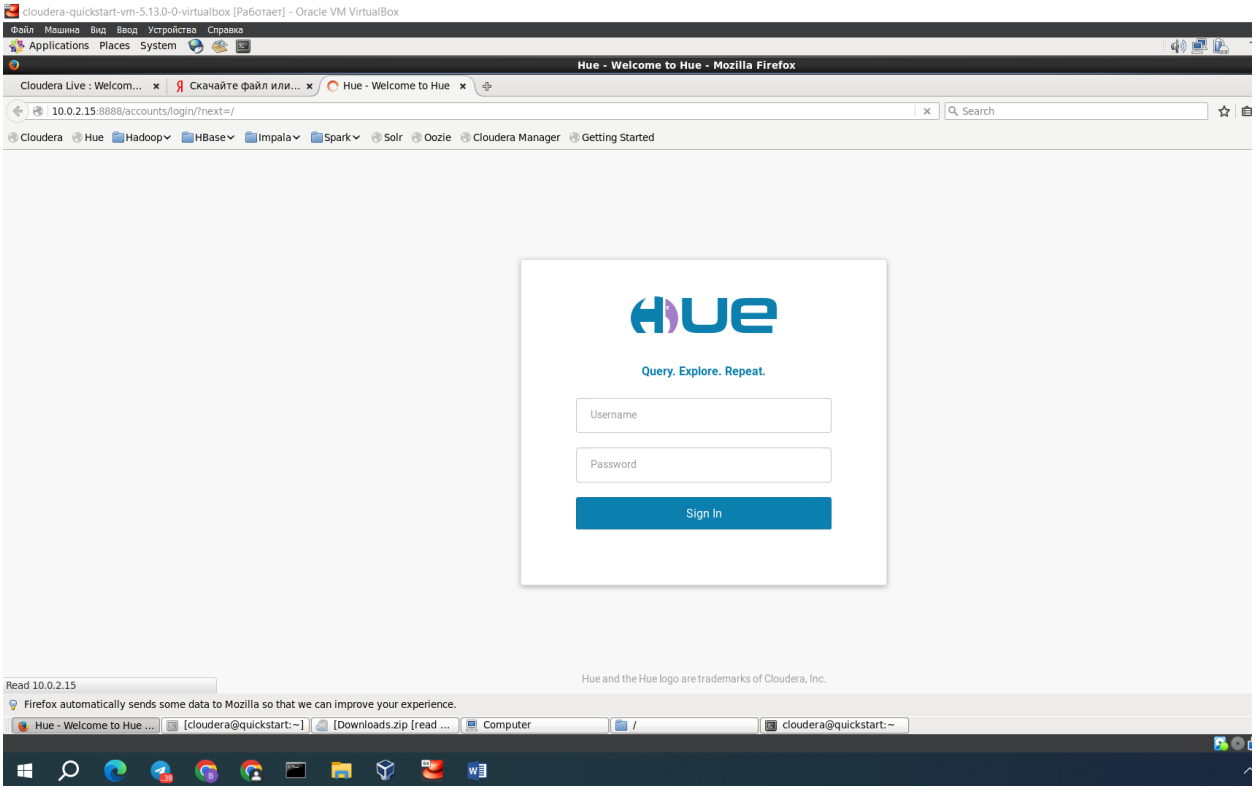


Все равно та же ошибка

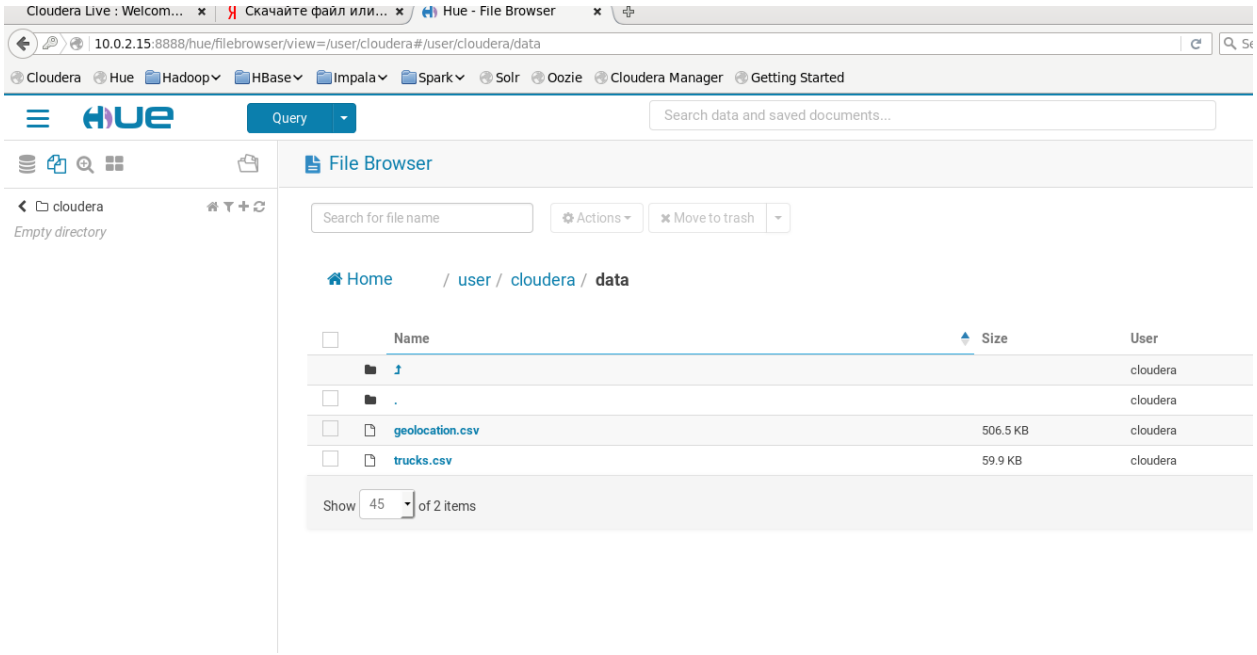
Скачиваем из облака



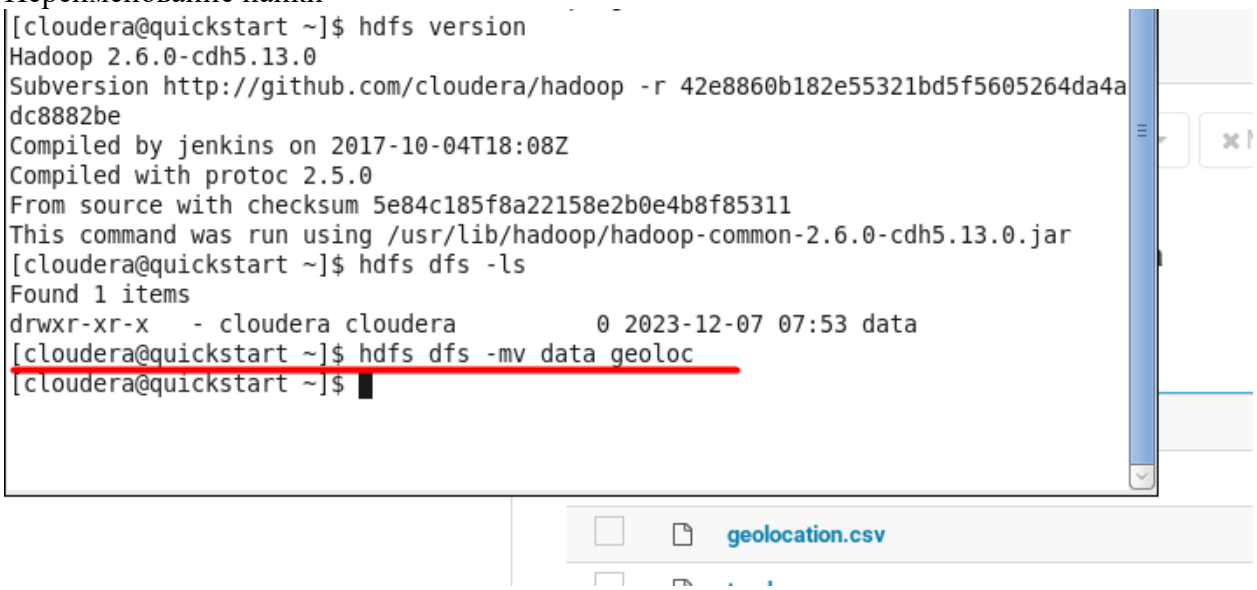
Узнаем свой айпи и заходим

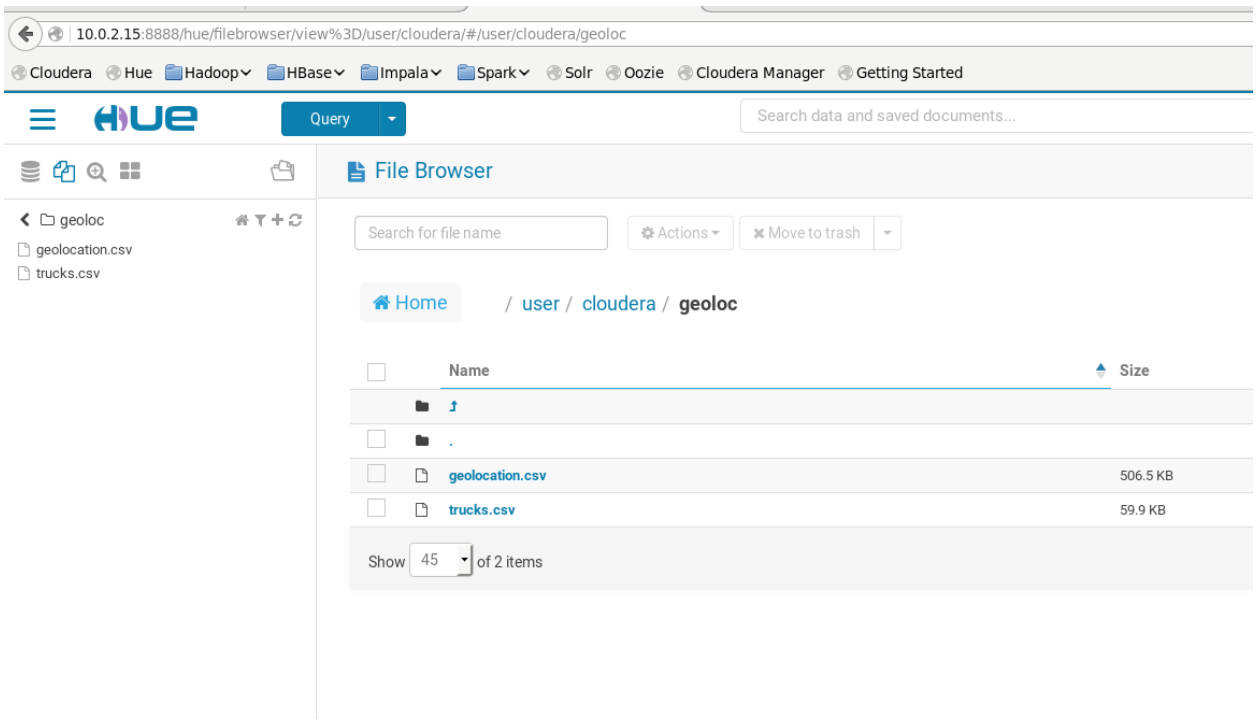


Входим по логину паролю
Cloudera
Cloudera
Загружаем таблицы

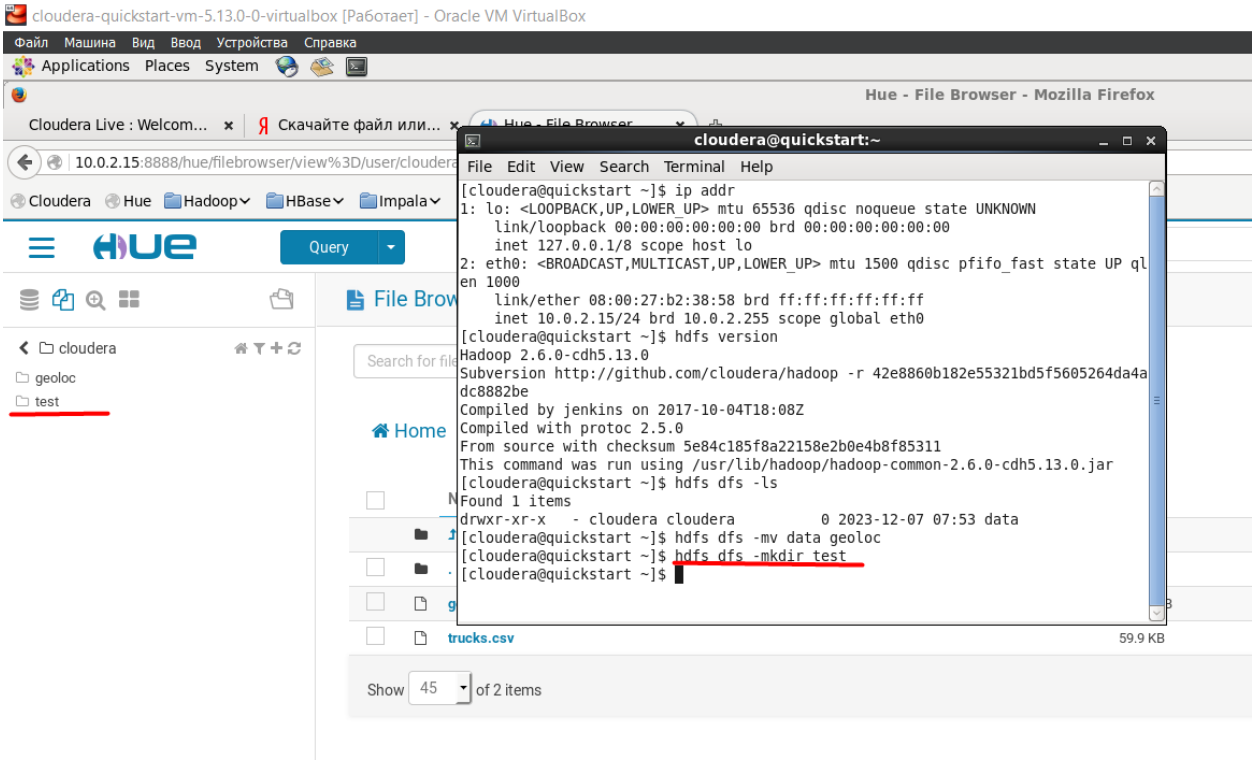


Переименование папки

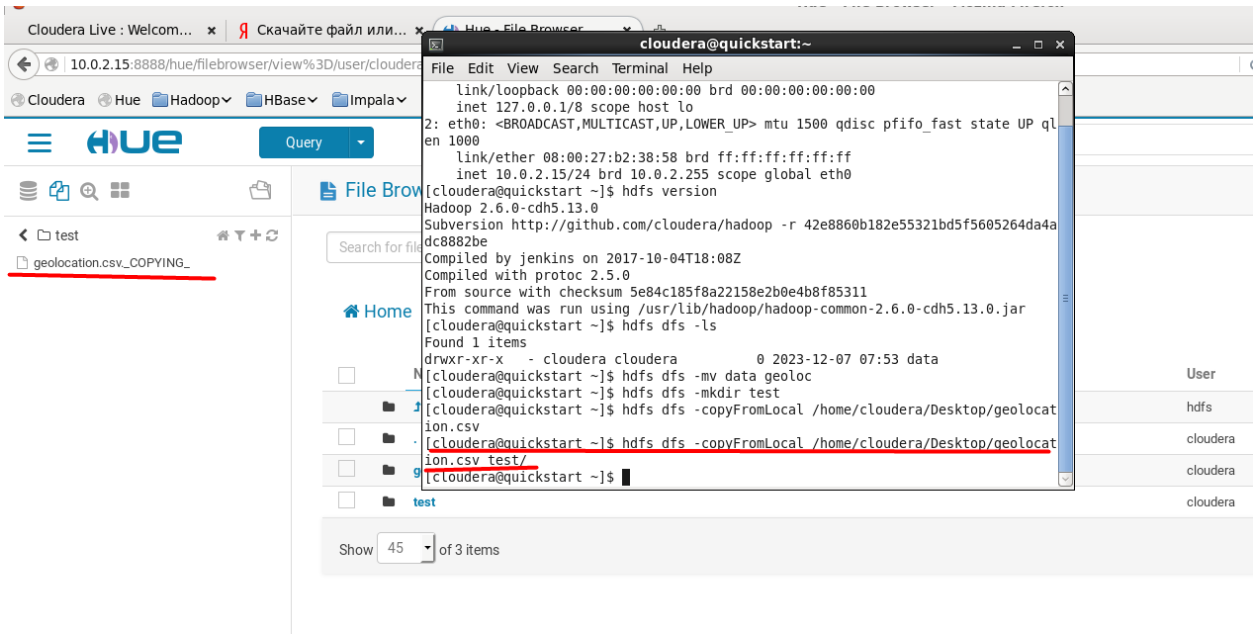




Создание папки тест



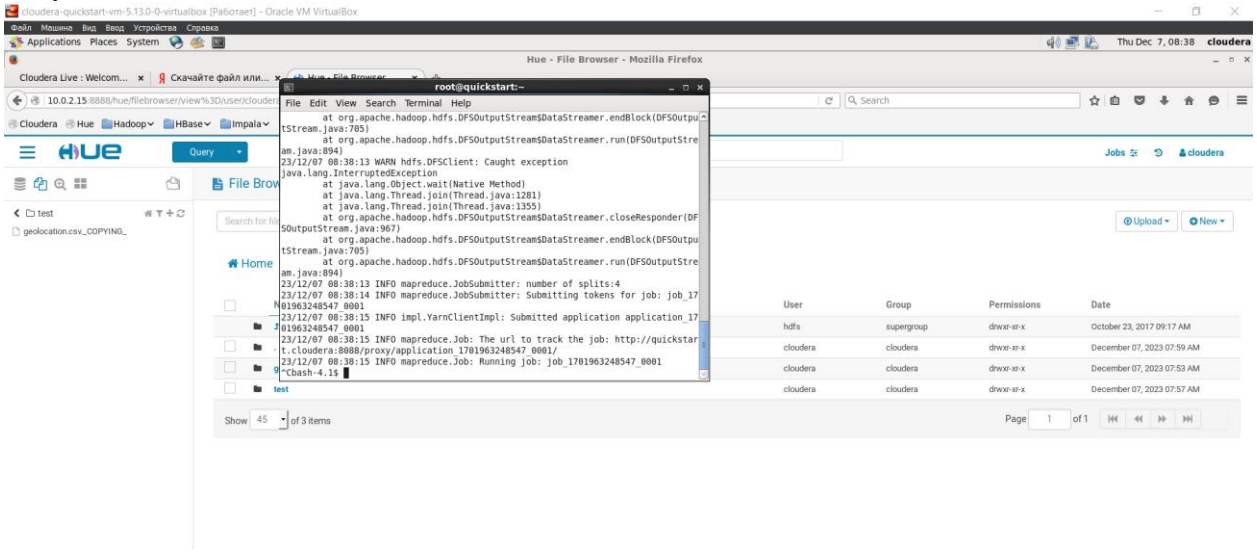
Копируем файл в тест



Запускаем скрипт, предложенный в репозитории

```
[cloudera@quickstart ~]$ sudo su -
[root@quickstart ~]# su hdfs
bash-4.1$ hdfs dfs -chmod -R 777 /tmp
bash-4.1$ ^C
```

Запускаем Pi



Запускаем

```
root@quickstart:~
File  Edit  View  Search  Terminal  Help

23/12/07 08:39:24 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/12/07 08:39:25 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp
/hadoop-yarn/staging/hdfs/.staging/job_1701963248547_0002
23/12/07 08:39:25 WARN security.UserGroupInformation: PriviledgedActionException
 as:hdfs (auth:SIMPLE) cause:org.apache.hadoop.mapreduce.lib.input.InvalidInputE
xception: Input path does not exist: hdfs://quickstart.cloudera:8020/user/hdfs/g
eoloc/geolocation.csv
org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not
 exist: hdfs://quickstart.cloudera:8020/user/hdfs/geoloc/geolocation.csv
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedL
istStatus(FileInputFormat.java:323)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(File
InputFormat.java:265)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileI
nputFormat.java:387)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.
java:305)
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.jav
a:322)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitt
er.java:200)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:415)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInforma
tion.java:1917)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
        at org.apache.hadoop.examples.WordCount.main(WordCount.java:87)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(Progra
mDriver.java:71)
        at org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
        at org.apache.hadoop.examples.ExampleDriver.main(ExampleDriver.java:74)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
```

Немного меняем команду на другой путь

yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/cloudera/geoloc/geolocation.csv output
получили success и открываем

```
root@quickstart:~
File  Edit  View  Search  Terminal  Help
distance,38.693723,-119.779683,Markleeville,California,53,1,0   1
distance,38.716687,-121.364422,Antelope,California,0,1,1        2
distance,38.716687,-121.364422,Antelope,California,15,1,0       1
distance,38.716687,-121.364422,Antelope,California,21,1,0       1
distance,38.716687,-121.364422,Antelope,California,31,1,0       2
distance,38.716687,-121.364422,Antelope,California,43,1,0       1
distance,38.716687,-121.364422,Antelope,California,45,1,0       1
distance,38.716687,-121.364422,Antelope,California,53,1,0       1
distance,38.716687,-121.364422,Antelope,California,62,1,0       1
distance,38.716687,-121.364422,Antelope,California,63,1,0       2
distance,38.752124,-121.288006,Roseville,California,21,1,0      1
distance,38.752124,-121.288006,Roseville,California,23,1,0      1
distance,38.752124,-121.288006,Roseville,California,41,1,0      1
distance,38.799703,-121.720574,Knights  1
distance,38.805462,-123.017223,Cloverdale,California,0,1,1      1
distance,39.017396,-122.057748,Arbuckle,California,0,1,1        1
distance,39.017396,-122.057748,Arbuckle,California,23,1,0       1
distance,39.017396,-122.057748,Arbuckle,California,31,1,0       3
distance,39.017396,-122.057748,Arbuckle,California,54,1,0       1
distance,39.145725,-121.591355,Marysville,California,28,1,0     1
distance,39.145725,-121.591355,Marysville,California,68,1,0     2
distance,39.150171,-123.207783,Ukiah,California,31,1,0  1
distance,39.150171,-123.207783,Ukiah,California,62,1,0  1
distance,39.230151,-121.136075,Rough    2
distance,39.409608,-123.355566,Willits,California,21,1,0        2
distance,39.409608,-123.355566,Willits,California,23,1,0        2
distance,39.409608,-123.355566,Willits,California,26,1,0        1
distance,39.409608,-123.355566,Willits,California,28,1,0        3
distance,39.409608,-123.355566,Willits,California,33,1,0        1
distance,39.409608,-123.355566,Willits,California,38,1,0        1
distance,39.409608,-123.355566,Willits,California,45,1,0        1
distance,39.409608,-123.355566,Willits,California,54,1,0        2
distance,39.409608,-123.355566,Willits,California,55,1,0        1
distance,39.409608,-123.355566,Willits,California,70,1,0        1
distance,40.563763,-122.238892,Palo     5
distance,40.58654,-122.391675,Redding,California,0,1,1  2
distance,40.58654,-122.391675,Redding,California,53,1,0 1
distance,40.761243,-123.994778,Kneeland,California,46,1,0       1
distance,41.526509,-124.038407,Klamath,California,0,1,1 1
distance,41.526509,-124.038407,Klamath,California,24,1,0        1
distance,41.526509,-124.038407,Klamath,California,31,1,0        1
distance,41.526509,-124.038407,Klamath,California,33,1,0        1
distance,41.526509,-124.038407,Klamath,California,43,1,0        1
following       150
tail    65
truckid,driverid,event,latitude,longitude,city,state,velocity,event_ind,idling_ind      1
bash-4.1$
```

Приступаем к самостоятельной работе 3.1

Виртуальное окружение уже запущено, вм уже открыта и работает

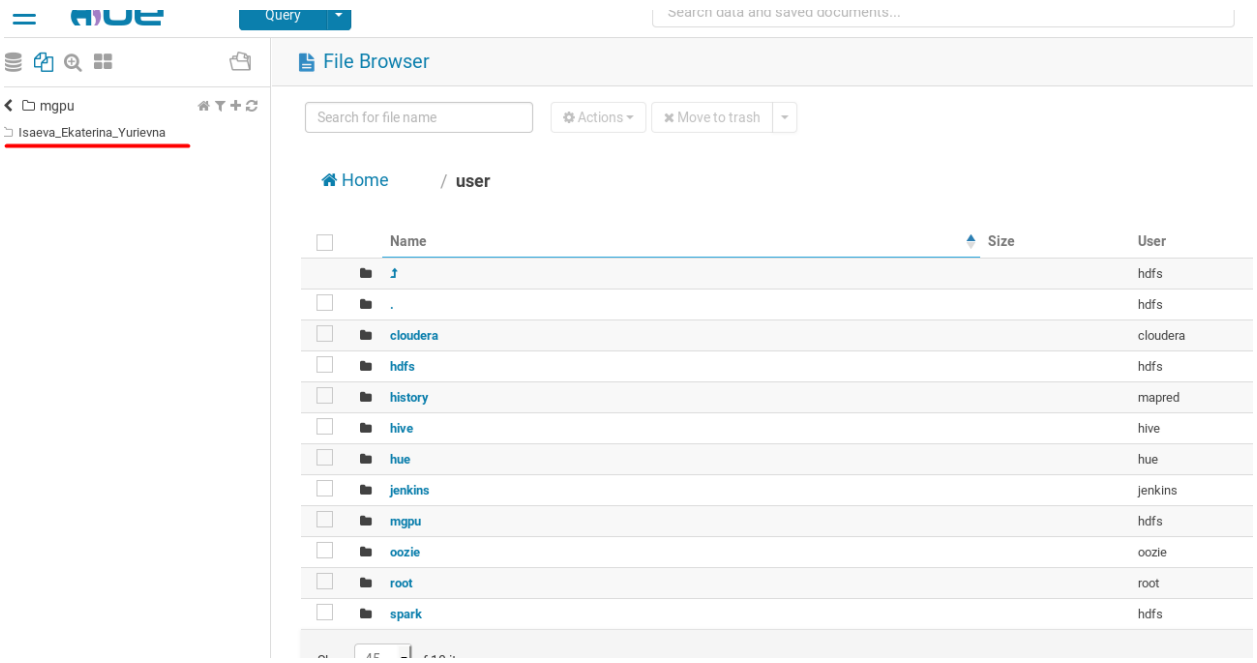
Открываем help

```
root@quickstart:~
File  Edit  View  Search  Terminal  Help
        permissions numberOfReplicas userId groupId sizeOfFile(in bytes)
  modificationDate(yyyy-MM-dd HH:mm) fileName

    -C  Display the paths of files and directories only.
    -d  Directories are listed as plain files.
    -h  Formats the sizes of files in a human-readable fashion
        rather than a number of bytes.
    -q  Print ? instead of non-printable characters.
    -R  Recursively list the contents of directories.
    -t  Sort files by modification time (most recent first).
    -S  Sort files by size.
    -r  Reverse the order of the sort.
    -u  Use time of last access instead of modification for
        display and sorting.

-mkdir [-p] <path> ... :
  Create a directory in specified location.

  -p  Do not fail if the directory already exists

-moveFromLocal <localsrc> ... <dst> :
  Same as -put, except that the source is deleted after it's copied.

-moveToLocal <src> <localdst> :
  Not implemented yet

-mv <src> ... <dst> :
  Move files that match the specified file pattern <src> to a destination <dst>.
  When moving multiple files, the destination must be a directory.

-put [-f] [-p] [-l] <localsrc> ... <dst> :
  Copy files from the local file system into fs. Copying fails if the file already
  exists, unless the -f flag is given.
  Flags:

  -p  Preserves access and modification times, ownership and the mode.
  -f  Overwrites the destination if it already exists.
  -l  Allow DataNode to lazily persist the file to disk. Forces
         replication factor of 1. This flag will result in reduced
         durability. Use with care.

-renameSnapshot <snapshotDir> <oldName> <newName> :
  Rename a snapshot from oldName to newName

-rm [-f] [-r|-R] [-skipTrash] <src> ... :
  Delete all files that match the specified file pattern. Equivalent to the Unix
  command "rm <src>"
/
```

Просматриваем корневую папку

```
bash-4.1$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2023-12-07 07:35 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2023-12-07 07:35 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2023-12-07 08:38 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /var
bash-4.1$
root@quickstart:~
```

Создать в HDFS в директории /user/mgpu поддиректорию ваше_фио



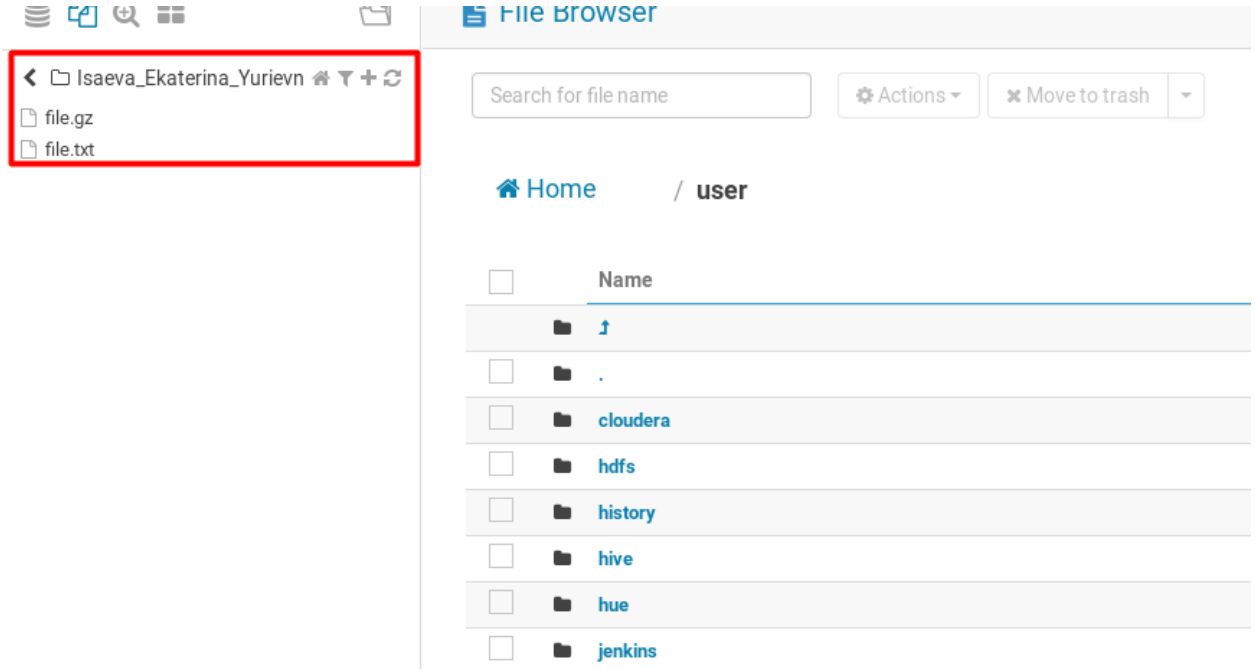
Создать в локальной файловой системе случайный текстовый файл размером 10 Mb с именем, образованным вашими инициалами base64 /dev/urandom | head -c 10000000 > file.txt

```
bash-4.1$ hdfs dfs -mkdir -p /user/mgpu/Isaeva_Ekaterina_Yurievna
bash-4.1$ echo -n "EY" | base64 --decode | dd bs=1024 count=10240 > file.txt
base64: invalid input
0+1 records in
0+1 records out
1 byte (1 B) copied, 0.0117094 s, 0.1 kB/s
bash-4.1$
```

Заархивировать созданный текстовый файл gzip -c file.txt > file.gz

```
mkdir: `/user/mgpu/Isaeva_Ekaterina_Yurievna': No such file or directory
bash-4.1$ ^C
bash-4.1$ hdfs dfs -mkdir -p /user/mgpu/Isaeva_Ekaterina_Yurievna
bash-4.1$ echo -n "EY" | base64 --decode | dd bs=1024 count=10240 > file.txt
base64: invalid input
0+1 records in
0+1 records out
1 byte (1 B) copied, 0.0117094 s, 0.1 kB/s
bash-4.1$ gzip -c file.txt > file.gz
bash-4.1$
```

Скопировать текстовый файл и архив в директорию /user/mgpu/fio HDFS виртуальной машины



Просмотреть файл и архив с помощью утилит cat, text в комбинации с каналами и утилитами head, tail -- привести не менее 3 вариантов команд и просмотра файла

Через cat

```
bash-4.1$ gzip -c file.txt > file.gz
bash-4.1$ hdfs dfs -copyFromLocal file.txt /user/mgpu/Isaeva_Ekaterina_Yurievna/
bash-4.1$ hdfs dfs -copyFromLocal file.gz /user/mgpu/Isaeva_Ekaterina_Yurievna/
bash-4.1$ hdfs dfs -cat /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt
bash-4.1$ hdfs dfs -cat /user/mgpu/Isaeva_Ekaterina_Yurievna/file.gz
qefile.txtbash-4.1$
```

Text + head

```
qefile.txtbash-4.1$ hdfs dfs -text /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt | head
bash-4.1$ hdfs dfs -text /user/mgpu/Isaeva_Ekaterina_Yurievna/file.gz | head
bash-4.1$
```

Text + tail

```
bash-4.1$ hdfs dfs -text /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt | tail
bash-4.1$ hdfs dfs -text /user/mgpu/Isaeva_Ekaterina_Yurievna/file.gz | tail
bash-4.1$
```

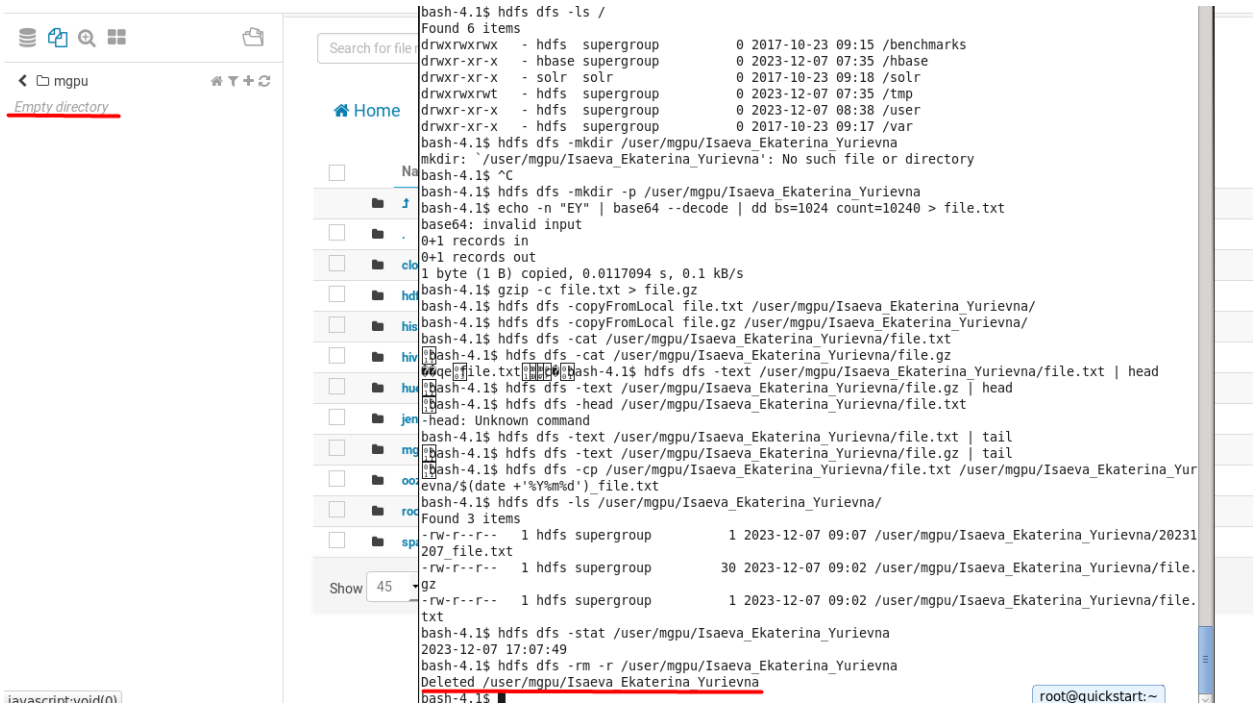
Создать копию файла file.txt вида date_file.txt, где в начале имени файла-копии указана текущая дата. Вывести листинг

```
bash-4.1$ hdfs dfs -cp /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt /user/mgpu/Isaeva_Ekaterina_Yur
evna/$(date +'%Y%m%d')_file.txt
bash-4.1$ hdfs dfs -ls /user/mgpu/Isaeva_Ekaterina_Yurievna/
Found 3 items
-rw-r--r--   1 hdfs supergroup          1 2023-12-07 09:07 /user/mgpu/Isaeva_Ekaterina_Yurievna/20231
207_file.txt
-rw-r--r--   1 hdfs supergroup         30 2023-12-07 09:02 /user/mgpu/Isaeva_Ekaterina_Yurievna/file.
gz
-rw-r--r--   1 hdfs supergroup          1 2023-12-07 09:02 /user/mgpu/Isaeva_Ekaterina_Yurievna/file.
txt
bash-4.1$
```

Вывести статистику по директории /user/mgpu/fio виртуальной машины

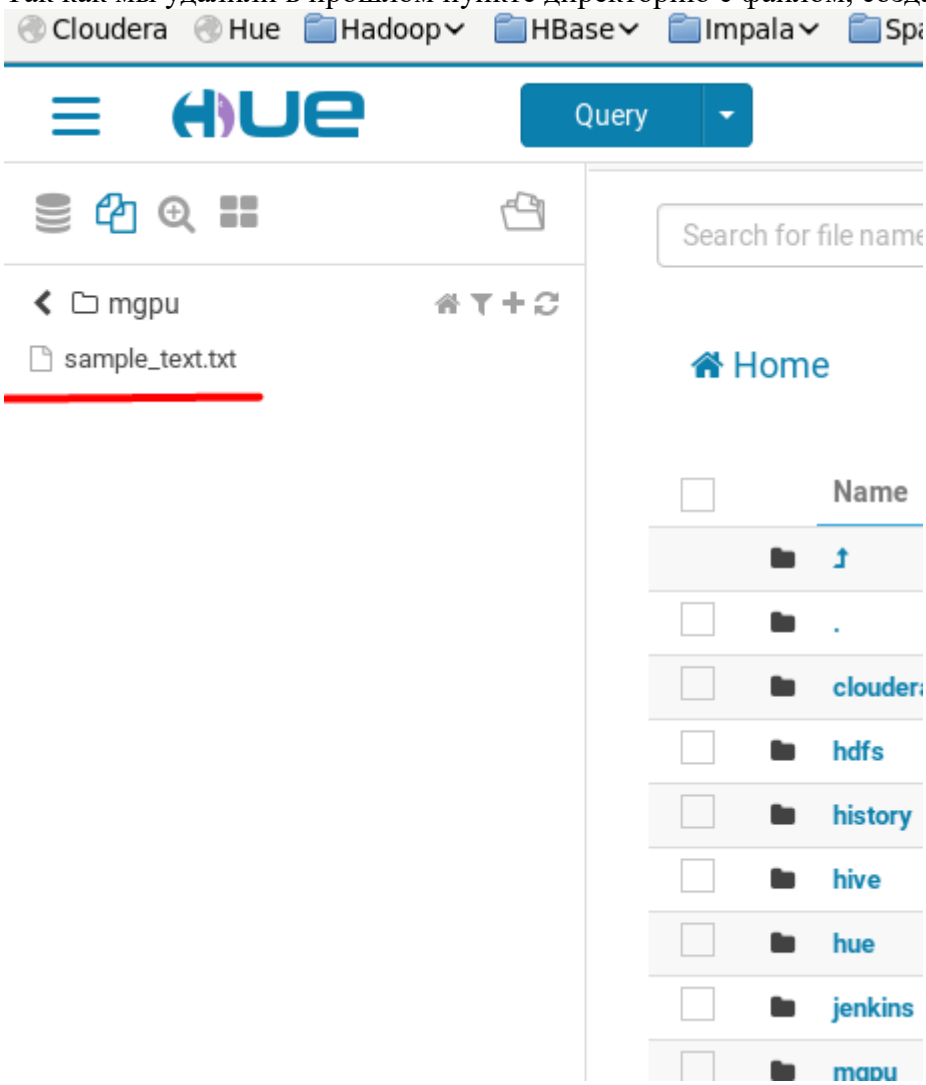
```
bash-4.1$ hdfs dfs -stat /user/mgpu/Isaeva_Ekaterina_Yurievna
2023-12-07 17:07:49
bash-4.1$
```

Удалить поддиректорию /fio со всем содержимым



Подсчитать количество слов в файле внутри HDFS с помощью методологии Map Reduce (размер файла не менее 128 Мб).

Так как мы удалили в прошлом пункте директорию с файлом, создадим новый файл



root@quickstart:~

File Edit View Search Terminal Help

```
bash-4.1$ hdfs dfs -ls /user/mgpu/Isaeva_Ekaterina_Yurievna/
Found 3 items
-rw-r--r--   1 hdfs supergroup          1 2023-12-07 09:07 /user/mgpu/Isaeva_Ekaterina_Yurievna/20231
207_file.txt
-rw-r--r--   1 hdfs supergroup         30 2023-12-07 09:02 /user/mgpu/Isaeva_Ekaterina_Yurievna/file.
gz
-rw-r--r--   1 hdfs supergroup          1 2023-12-07 09:02 /user/mgpu/Isaeva_Ekaterina_Yurievna/file.
txt
bash-4.1$ hdfs dfs -stat /user/mgpu/Isaeva_Ekaterina_Yurievna
2023-12-07 17:07:49
bash-4.1$ hdfs dfs -rm -r /user/mgpu/Isaeva_Ekaterina_Yurievna
Deleted /user/mgpu/Isaeva_Ekaterina_Yurievna
bash-4.1$ hdfs dfs -copyFromLocal date_file.txt /user/mgpu/Isaeva_Ekaterina_Yurievna/
copyFromLocal: `/user/mgpu/Isaeva_Ekaterina_Yurievna/': No such file or directory
bash-4.1$ hdfs dfs -copyFromLocal date_file.txt /user/mgpu/
copyFromLocal: `date_file.txt': No such file or directory
bash-4.1$ base64 /dev/urandom | head -c 150M > sample_text.txt
bash-4.1$ hdfs dfs -copyFromLocal sample_text.txt /user/mgpu/
bash-4.1$ yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/mgpu/sampl
e_text.txt /user/mgpu/output_unique
23/12/07 09:19:36 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/12/07 09:19:37 INFO input.FileInputFormat: Total input paths to process : 1
23/12/07 09:19:37 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:96
7)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
23/12/07 09:19:37 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:96
7)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
23/12/07 09:19:37 INFO mapreduce.JobSubmitter: number of splits:2
23/12/07 09:19:37 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1701963248547_0005
23/12/07 09:19:37 INFO impl.YarnClientImpl: Submitted application application_1701963248547_0005
23/12/07 09:19:38 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy
/application_1701963248547_0005/
23/12/07 09:19:38 INFO mapreduce.Job: Running job: job_1701963248547_0005
```

```
root@quickstart:~
File  Edit  View  Search  Terminal  Help
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=95387
                Total time spent by all reduces in occupied slots (ms)=33994
                Total time spent by all map tasks (ms)=95387
                Total time spent by all reduce tasks (ms)=33994
                Total vcore-milliseconds taken by all map tasks=95387
                Total vcore-milliseconds taken by all reduce tasks=33994
                Total megabyte-milliseconds taken by all map tasks=97676288
                Total megabyte-milliseconds taken by all reduce tasks=34809856
        Map-Reduce Framework
                Map input records=2042681
                Map output records=2042681
                Map output bytes=165457125
                Map output materialized bytes=169542499
                Input split bytes=244
                Combine input records=3785769
                Combine output records=3785769
                Reduce input groups=2042681
                Reduce shuffle bytes=169542499
                Reduce input records=2042681
                Reduce output records=2042681
                Spilled Records=5828450
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=1068
                CPU time spent (ms)=22710
                Physical memory (bytes) snapshot=711589888
                Virtual memory (bytes) snapshot=4525752320
                Total committed heap usage (bytes)=552845312
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=157290496
        File Output Format Counters
                Bytes Written=161371763
bash-4.1$
```

Приступаем к самостоятельной работе 3.2

В интерактивном режиме через Терминал, запустив оболочку Pig с помощью pig и выполняя команды одну за другой.

```
root@quickstart:~
File  Edit  View  Search  Terminal  Help
2023-12-07 09:34:54,788 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /var/lib/ha
doop-hdfs/.pigbootup not found
2023-12-07 09:34:55,960 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:55,961 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:55,961 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine -
 Connecting to hadoop file system at: hdfs://quickstart.cloudera:8020
2023-12-07 09:34:58,403 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:58,403 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine -
 Connecting to map-reduce job tracker at: localhost:8021
2023-12-07 09:34:58,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:58,526 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:58,527 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:58,667 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:58,668 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:58,822 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:58,827 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:58,986 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:58,989 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:59,141 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:59,143 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:59,284 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:59,287 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:59,445 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:59,445 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
2023-12-07 09:34:59,565 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.na
me is deprecated. Instead, use fs.defaultFS
2023-12-07 09:34:59,567 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tr
acker is deprecated. Instead, use mapreduce.jobtracker.address
grunt>
root@quickstart:~
```

В Hueможно перейти в редактор Pig через Query > Editor > Pig. Это предпочтительный метод, если хотим запускать полные сценарии, но выполняется намного дольше, чем оболочка Pig.

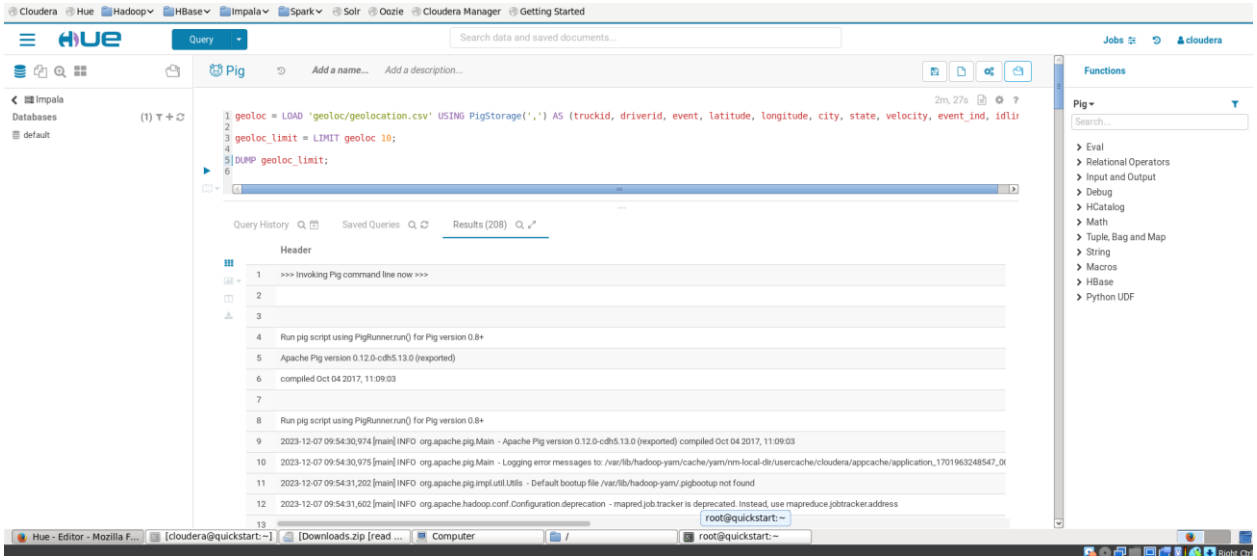
```
root@quickstart:~
File  Edit  View  Search  Terminal  Help
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
        at org.apache.hadoop.mapreduce.lib.jobcontrol.ControlledJob.submit(ControlledJob.java:335)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.pig.backend.hadoop23.PigJobControl.submit(PigJobControl.java:128)
        at org.apache.pig.backend.hadoop23.PigJobControl.run(PigJobControl.java:191)
        at java.lang.Thread.run(Thread.java:745)
        at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher$1.run(MapRe
duceLauncher.java:270)
Caused by: org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hd
fs://quickstart.cloudera:8020/user/hdfs/geoloc/geolocation.csv
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedListStatus(FileInputFo
rmat.java:323)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
        at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigTextInputFormat.listStatus
(PigTextInputFormat.java:36)
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileInputFormat.java:387)
        at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigInputFormat.getSplits(PigI
nputFormat.java:274)
        ... 18 more


Input(s):
Failed to read data from "hdfs://quickstart.cloudera:8020/user/hdfs/geoloc/geolocation.csv"

Output(s):

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1701963248547_0006  ->      null,
null


2023-12-07 09:51:18,498 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Map
ReduceLauncher - Failed!
2023-12-07 09:51:18,500 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open it
erator for alias geoloc_limit
Details at logfile: /var/lib/hadoop-hdfs/pig_1701970494712.log
grunt>
```

56.71s

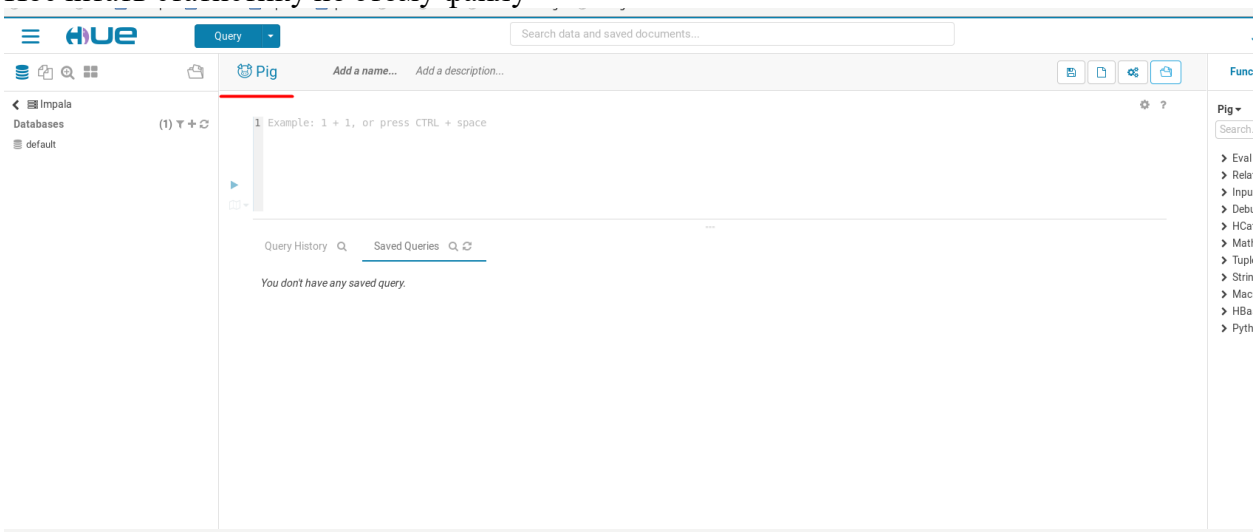
```
geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid, driverid, event, latitude, longitude, city, state, velocity, event_ind, idling

geoloc_limit = LIMIT geoloc 10;

DUMP geoloc_limit;
```

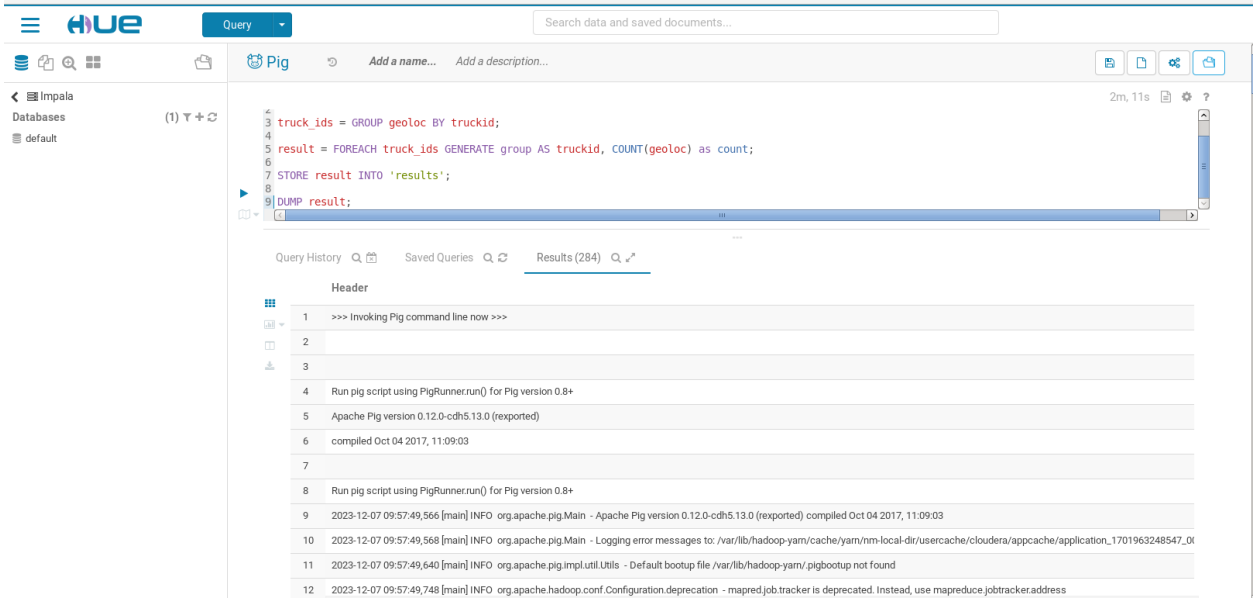
Query History Saved Queries

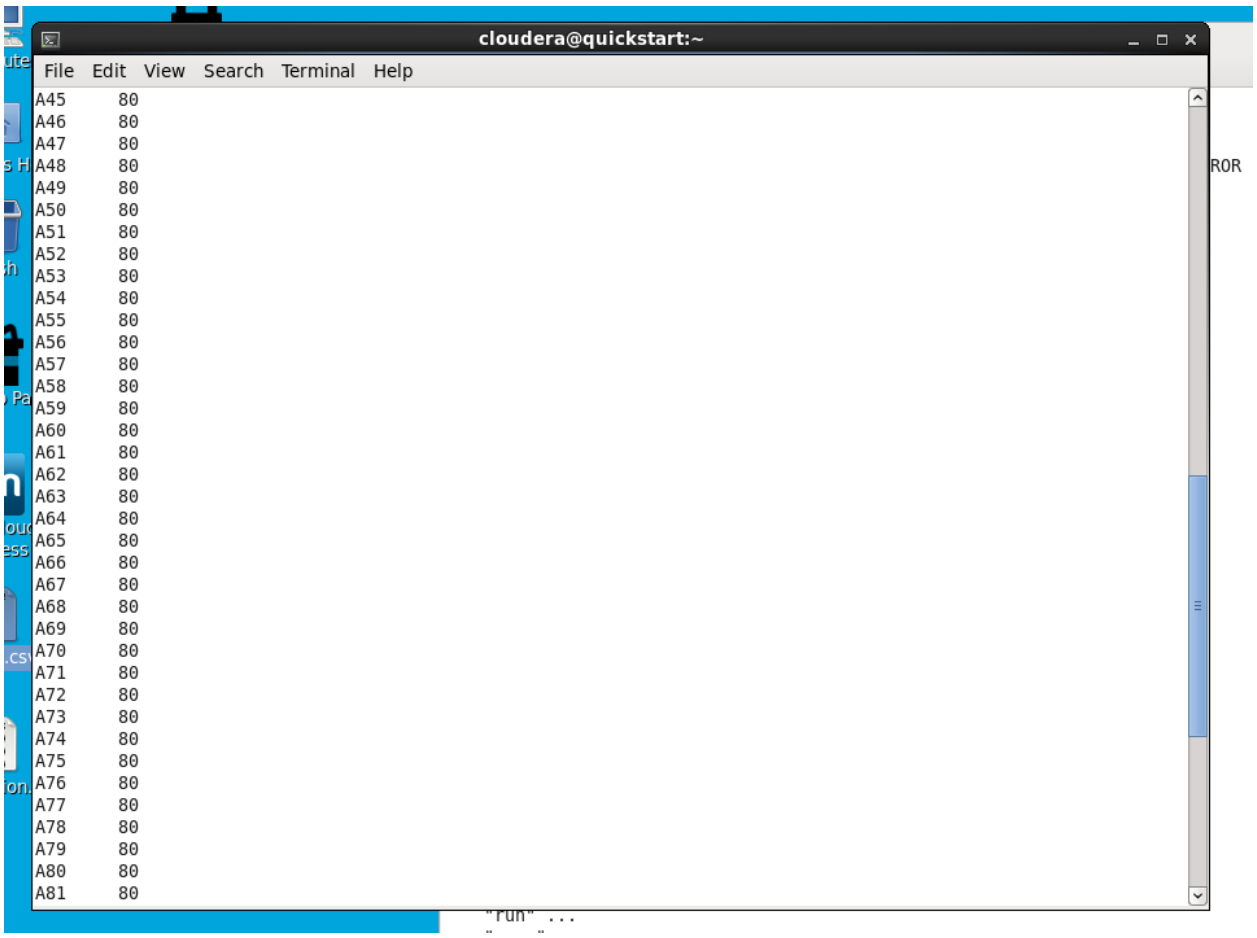
a minute ago geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid, driverid, event, latitude, longitude, city, state, velocity, event_ind, idling_ind);
geoloc_limit = LIMIT geoloc 10; DUMP geoloc_limit;



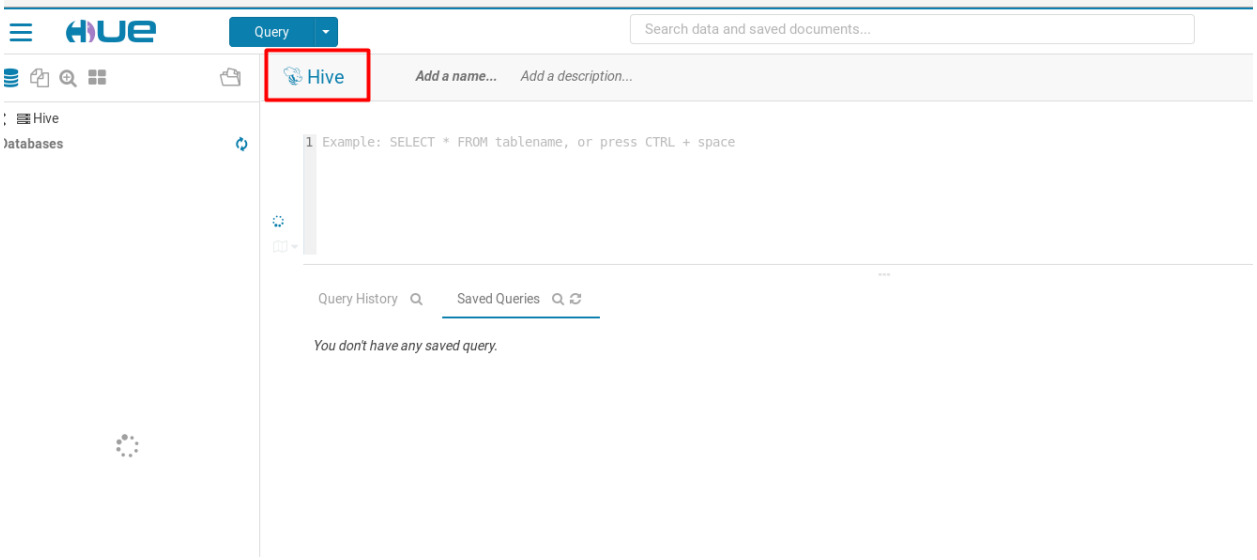
Посчитать статистику по этому файлу







Переключаемся на HIVE



```sql
CREATE EXTERNAL TABLE geolocation (
   truckid STRING,
   driverid STRING,
   event STRING,
   latitude DOUBLE,
   longitude DOUBLE,
   city STRING,
   state STRING,
   velocity DOUBLE,
   event_ind BIGINT,
   idling_ind BIGINT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LOCATION '/user/cloudera/geoloc';
```

Success.

Query History | Saved Queries

a few seconds ago — `CREATE EXTERNAL TABLE geolocation ( truckid STRING, driverid STRING, event STRING, latitude DOUBLE, longitude DOUBLE, city STRING, state STRING, velocity DOUBLE, event_ind BIGINT, idling_ind BIGINT ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/cloudera/geoloc'`

Первые 10 строк

```sql
SELECT truckid FROM geolocation LIMIT 10;
```

Query History | Saved Queries | Results (10)

	truckid
1	truckid
2	A54
3	A20
4	A40
5	A31
6	A71
7	A50
8	A51
9	A19
10	A77

Среднее для городов

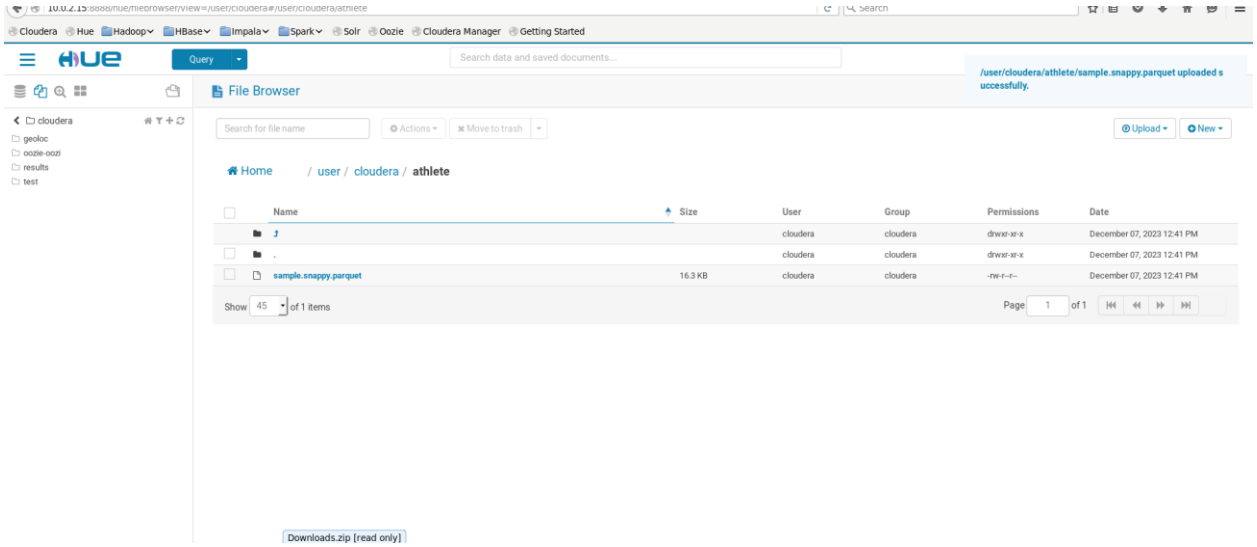
Hive — Add a name... — Add a description...

```sql
SELECT truckid, city, AVG(velocity) AS avg_speed
FROM geolocation
GROUP BY truckid, city;
```

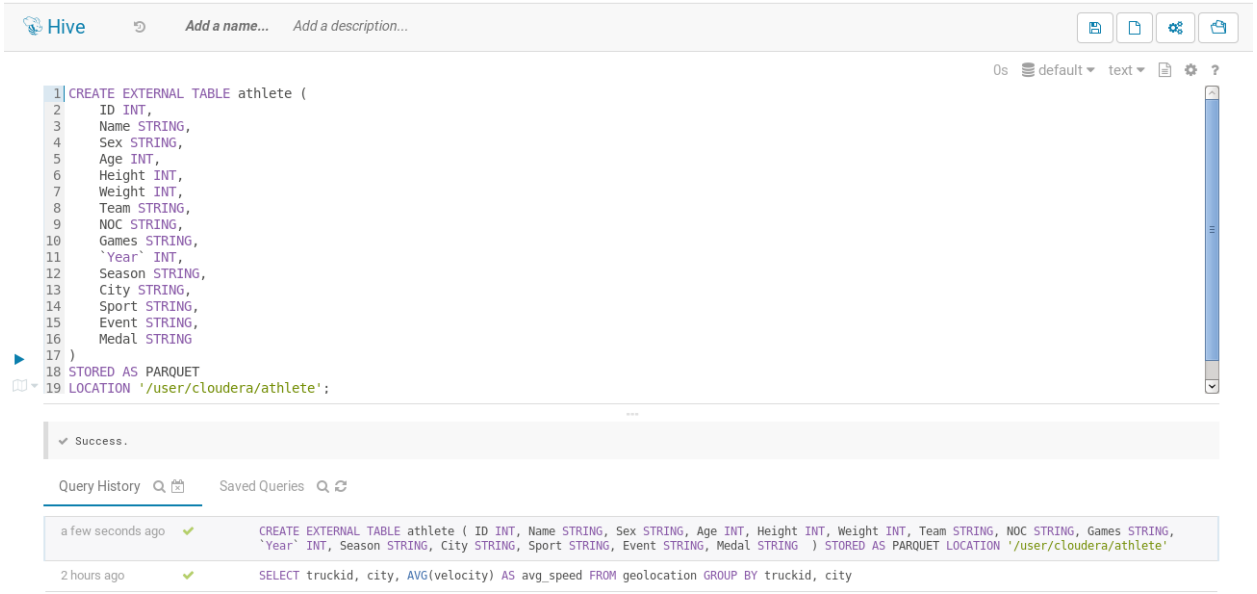
Query History | Saved Queries | Results (1,353)

	truckid	city	avg_speed
1	A1	8769	14234
2	A1	Antelope	27
3	A1	Aptos	45.600000000000001
4	A1	Arbuckle	46.600000000000001
5	A1	Gilroy	31.714285714285715
6	A1	Klamath	66
7	A1	Lodi	36.571428571428569
8	A1	Markleeville	35
9	A1	Modesto	52.571428571428569
10	A1	Napa	37.5
11	A1	Oceano	47.142857142857146
12	A1	Palmdale	26.600000000000001

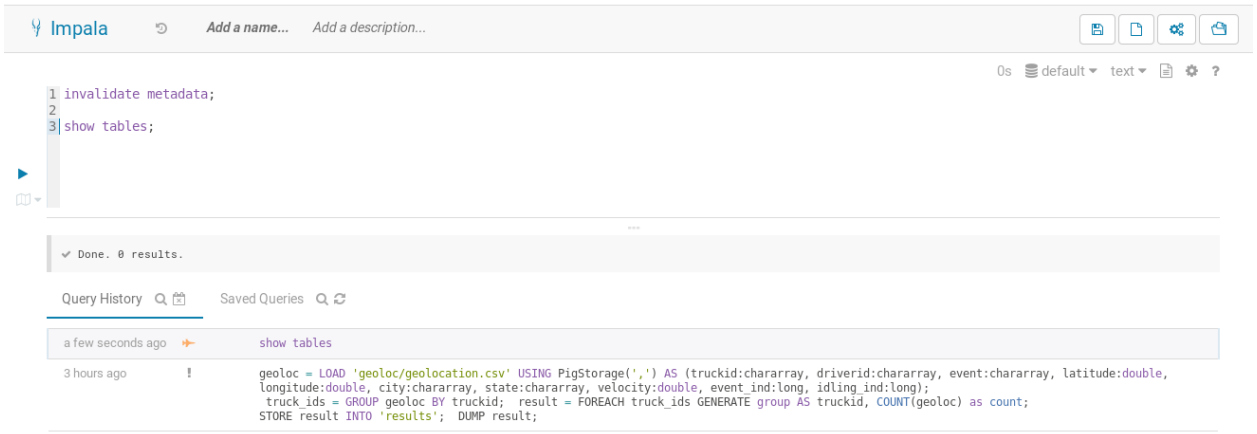
Скачиваем файл



Выполняем запрос



Переходим в Impala



Переходим в Hive и выполняем задание

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

19.25s default text

```
  truckid,
  COUNT(*) AS location_count
FROM
  geolocation
GROUP BY
  truckid;
```

Query History Saved Queries

	truckid	location_count
1	A1	81
2	A10	81
3	A100	81
4	A11	81
5	A12	81
6	A13	81
7	A14	81
8	A15	81
9	A16	81
10	A17	81
11	A18	81
12	A19	81
13	A2	81
14	A20	81
15	A21	81
16	A22	81
17	A23	81

root@quickstart:~

```
SELECT
  latitude,
  longitude
FROM
  geolocation
WHERE
  truckid = 'A80';
```

Tables

Search...

No tables

Query History Saved Queries Results (81)

	latitude	longitude
1	37.774929999999998	-122.419416
2	37.774929999999998	-122.419416
3	36.471865000000001	-6.1965950000000003
4	34.448050000000002	-119.24288900000001
5	38.440466999999998	-122.714431
6	39.409607999999999	-123.355566
7	36.471865000000001	-6.1965950000000003
8	38.161861000000002	-121.611621
9	38.440466999999998	-122.714431
10	38.019365999999998	-122.13413199999999
11	40.563763000000002	-122.23889200000001
12	38.019365999999998	-122.13413199999999
13	36.471865000000001	-6.1965950000000

root@quickstart:~