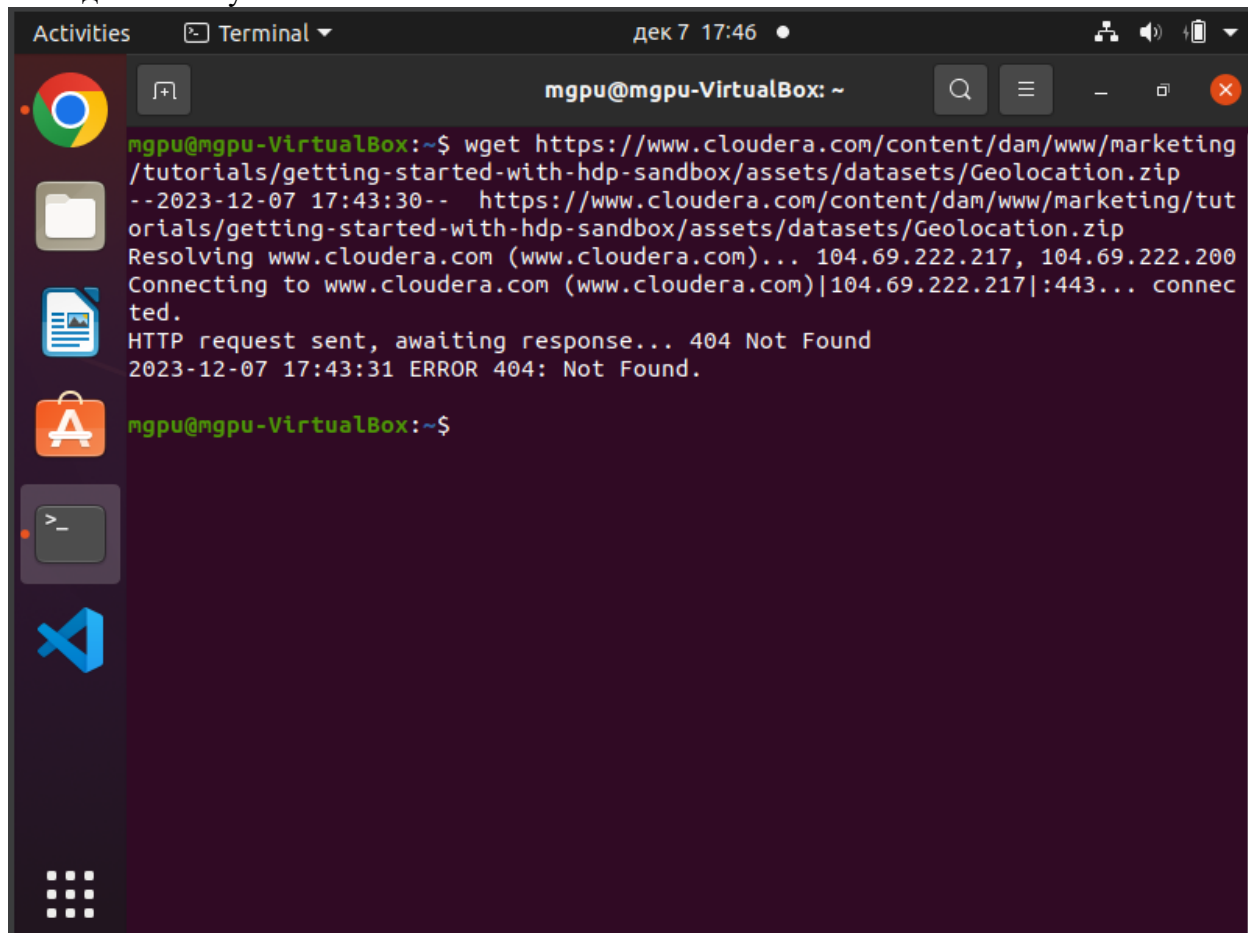


Установка cloudera

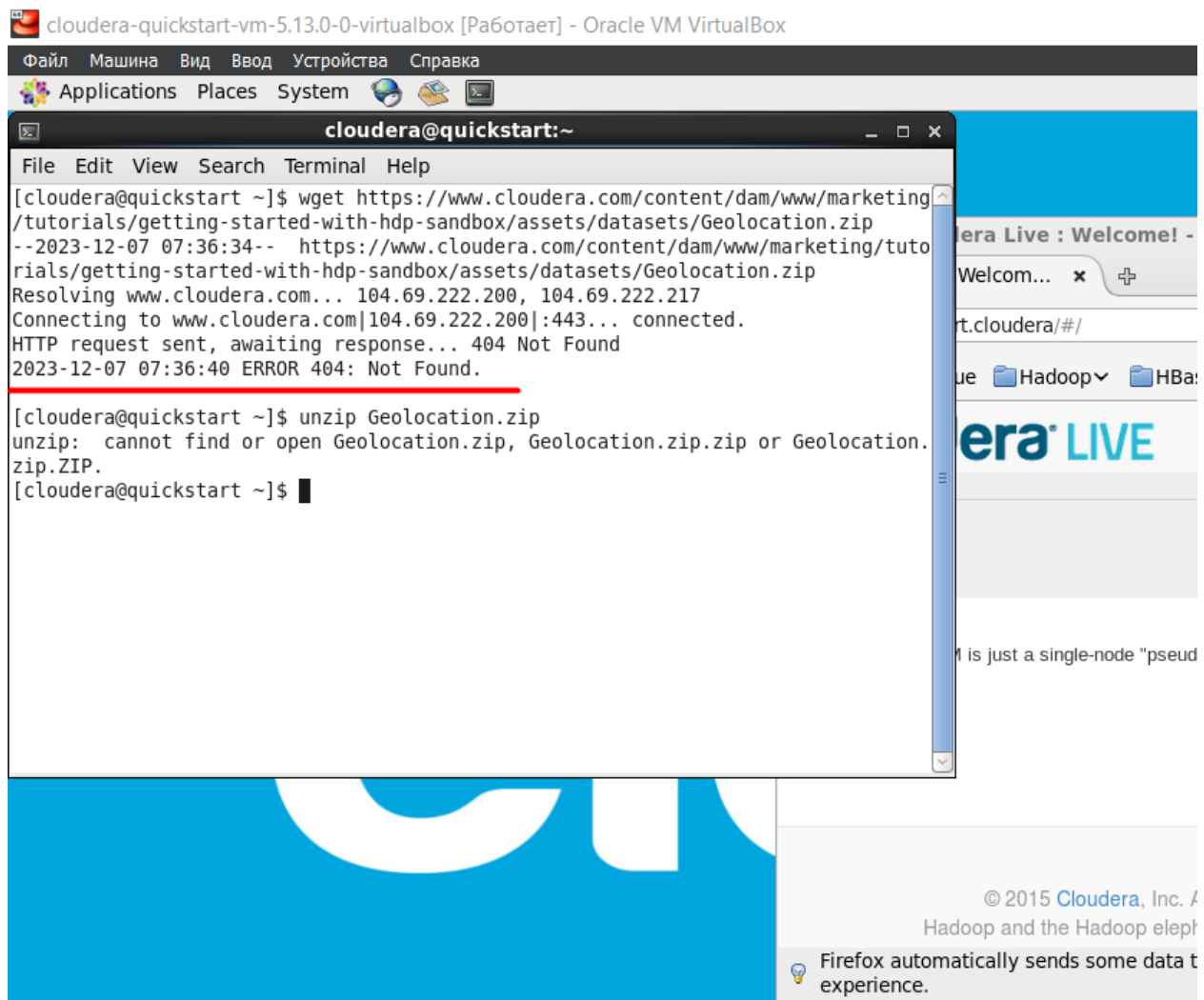
Выводит ошибку

A terminal window titled 'mgpu@mgpu-VirtualBox: ~' with a search bar and window controls. The terminal shows a command to download a file from Cloudera, which fails with a 404 error. The left sidebar of the desktop environment shows icons for Chrome, Files, Documents, Applications, and a Dash button.

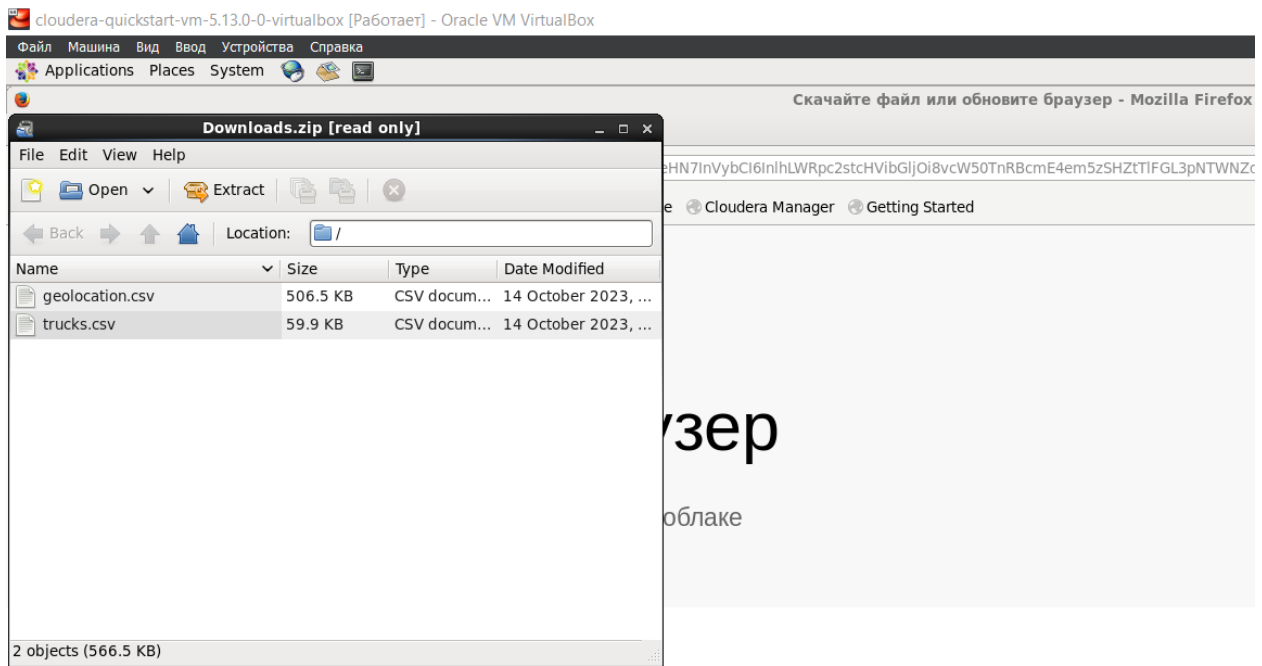
```
mgpu@mgpu-VirtualBox:~$ wget https://www.cloudera.com/content/dam/www/marketing/tutorials/getting-started-with-hdp-sandbox/assets/datasets/Geolocation.zip
--2023-12-07 17:43:30-- https://www.cloudera.com/content/dam/www/marketing/tutorials/getting-started-with-hdp-sandbox/assets/datasets/Geolocation.zip
Resolving www.cloudera.com (www.cloudera.com)... 104.69.222.217, 104.69.222.200
Connecting to www.cloudera.com (www.cloudera.com)|104.69.222.217|:443... connected.
HTTP request sent, awaiting response... 404 Not Found
2023-12-07 17:43:31 ERROR 404: Not Found.

mgpu@mgpu-VirtualBox:~$
```

Пробуем установить по-другому(на другой вм)



Все равно та же ошибка
Скачиваем из облака



Регулярные обновления

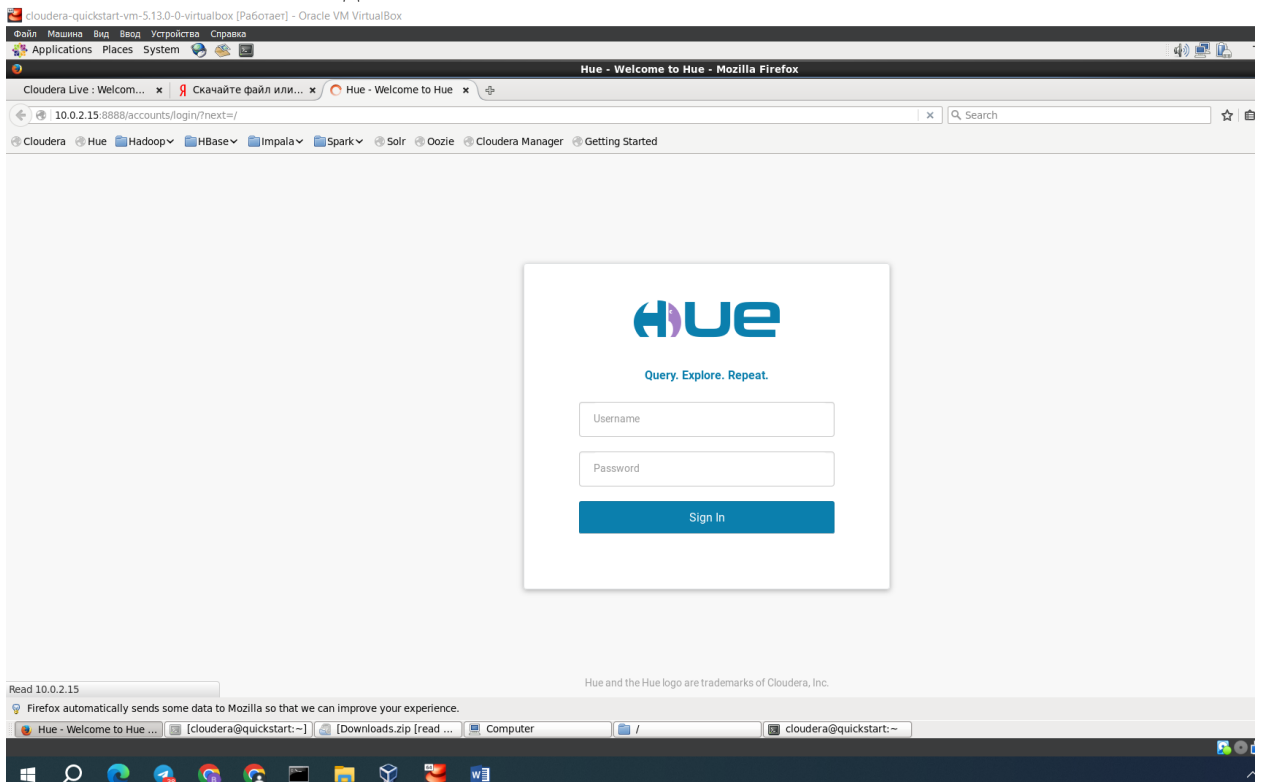
Браузер умеет автоматически обновляться до последней актуальной версии.

Безопасность

Технология Protect проверяет сайты и файлы на вирусы, блокирует страницы мошенников.

Скачать Яндекс Браузер

Узнаем свой айпи и заходим



Входим по логину паролю

Cloudera

Cloudera

Загружаем таблицы

Cloudera Live : Welcom... x | Скачайте файл или... x | Hue - File Browser x

10.0.2.15:8888/hue/filebrowser/view=/user/cloudera#/user/cloudera/data

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE Query Search data and saved documents...

File Browser

cloudera
Empty directory

Search for file name Actions Move to trash

Home / user / cloudera / data

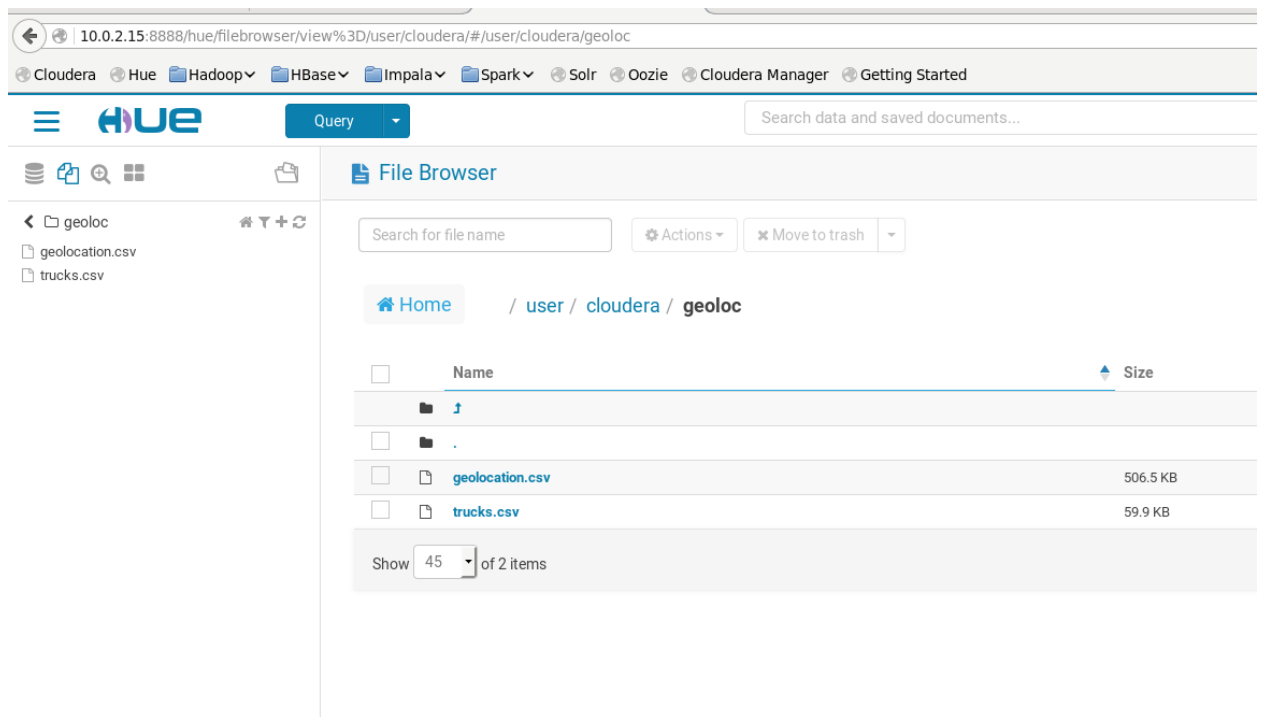
Name	Size	User
.		cloudera
geolocation.csv	506.5 KB	cloudera
trucks.csv	59.9 KB	cloudera

Show 45 of 2 items

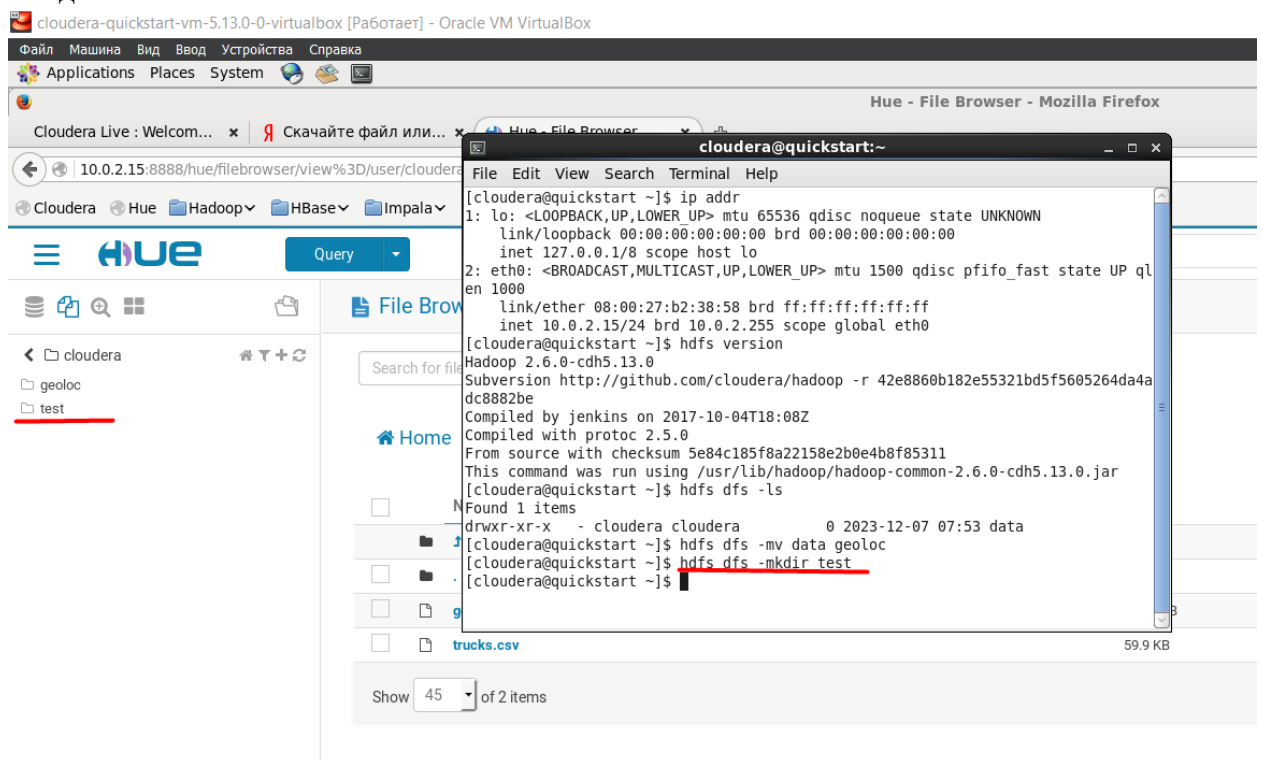
Переименование папки

```
[cloudera@quickstart ~]$ hdfs version
Hadoop 2.6.0-cdh5.13.0
Subversion http://github.com/cloudera/hadoop -r 42e8860b182e55321bd5f5605264da4a
dc8882be
Compiled by jenkins on 2017-10-04T18:08Z
Compiled with protoc 2.5.0
From source with checksum 5e84c185f8a22158e2b0e4b8f85311
This command was run using /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x - cloudera cloudera 0 2023-12-07 07:53 data
[cloudera@quickstart ~]$ hdfs dfs -mv data geoloc
[cloudera@quickstart ~]$
```

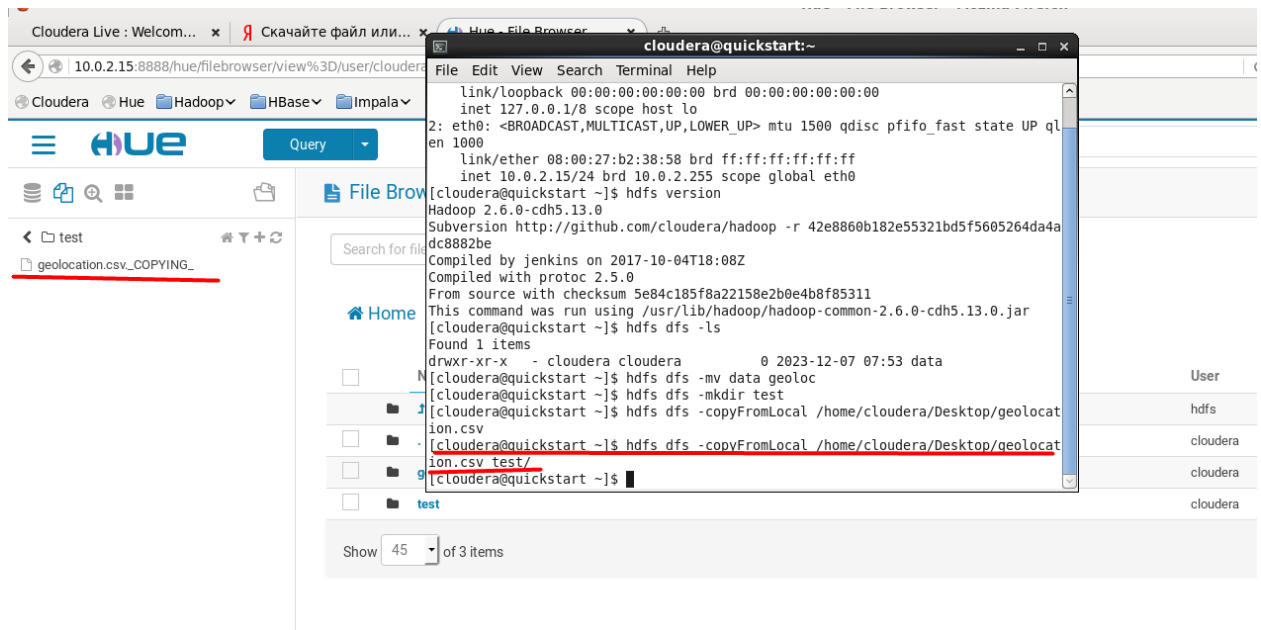
geolocation.csv



Создание папки тест



Копируем файл в тест



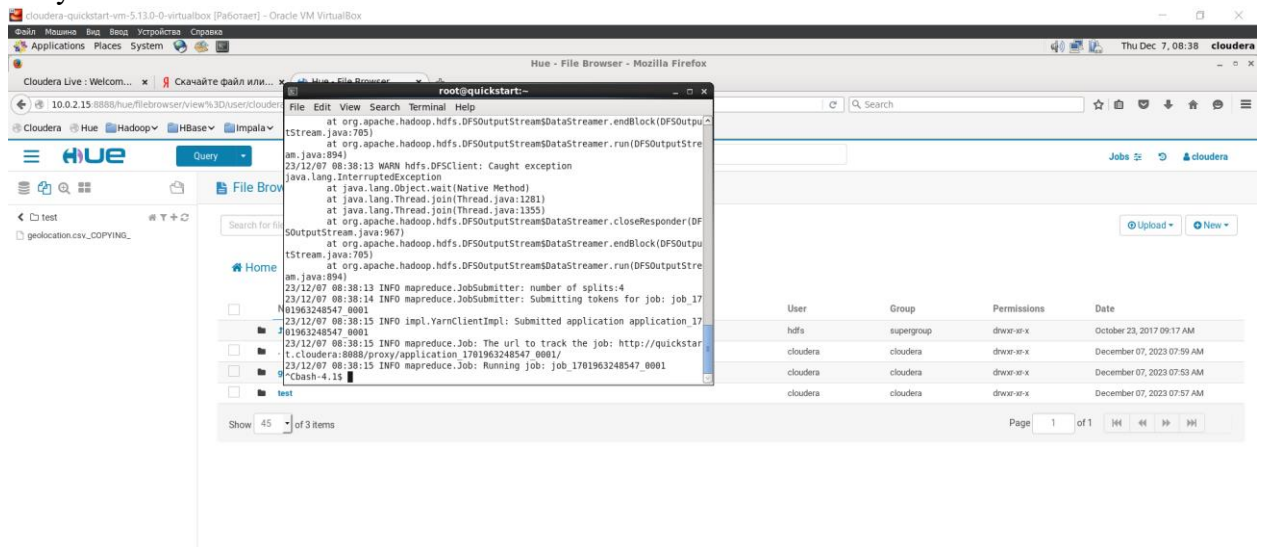
Запускаем скрипт, предложенный в репозитории

```

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

```

Запускаем Pi



Запускаем

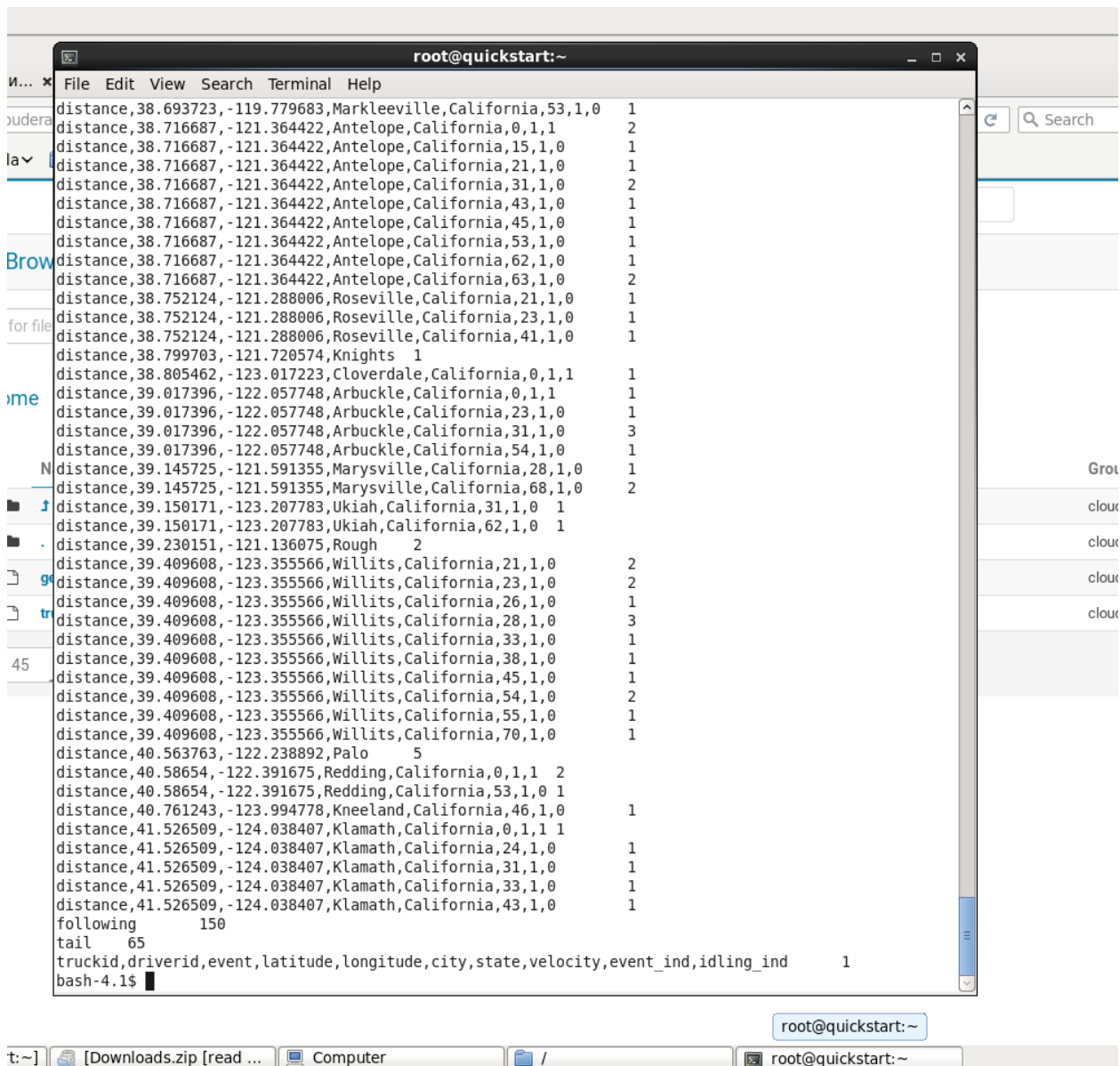
```
root@quickstart:~
File Edit View Search Terminal Help

23/12/07 08:39:24 INFO client.RMPProxy: 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: PrivilegedActionException as:hdfs (auth:SIMPLE) cause:org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path does not exist: hdfs://quickstart.cloudera:8020/user/hdfs/geoloc/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.singleThreadedListStatus(FileInputFormat.java:323)
    at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileInputFormat.java:265)
    at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.get_splits(FileInputFormat.java:387)
    at org.apache.hadoop.mapreduce.JobSubmitter.writeNew_splits(JobSubmitter.java:305)
    at org.apache.hadoop.mapreduce.JobSubmitter.write_splits(JobSubmitter.java:322)
    at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:200)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
    at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
    at java.security.AccessController.doPrivileged(Native Method)
    at javax.security.auth.Subject.doAs(Subject.java:415)
    at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)
    at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
    at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
    at org.apache.hadoop.examples.WordCount.main(WordCount.java:87)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
    at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:606)
    at org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(ProgramDriver.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(DelegatingMethodAccessorImpl.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



Приступаем к самостоятельной работе 3.1

Виртуальное окружение уже запущено, vm 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

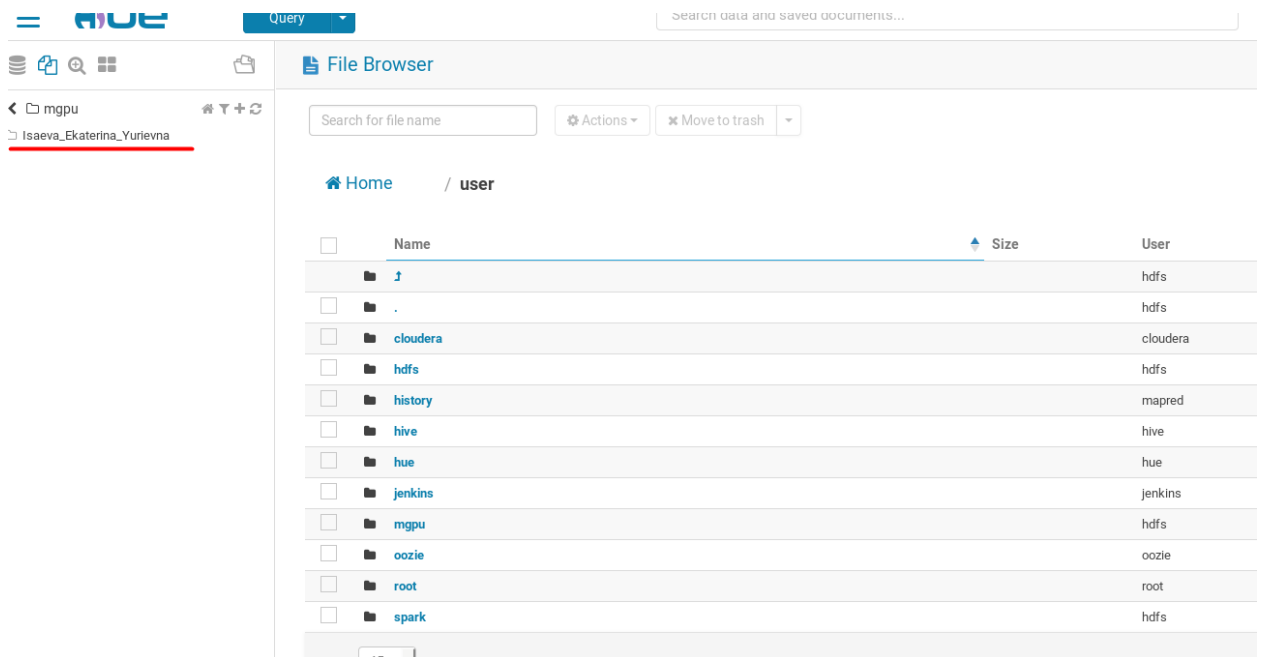
-rm [-f] [-r|-R] [-skipTrash] <src> ... :
  Delete all files that match the specified file pattern. Equivalent to the Unix
  command "rm <src>"
```

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
```

root@quickstart:~

Создать в HDFS в директории /user/mgpi 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$
```

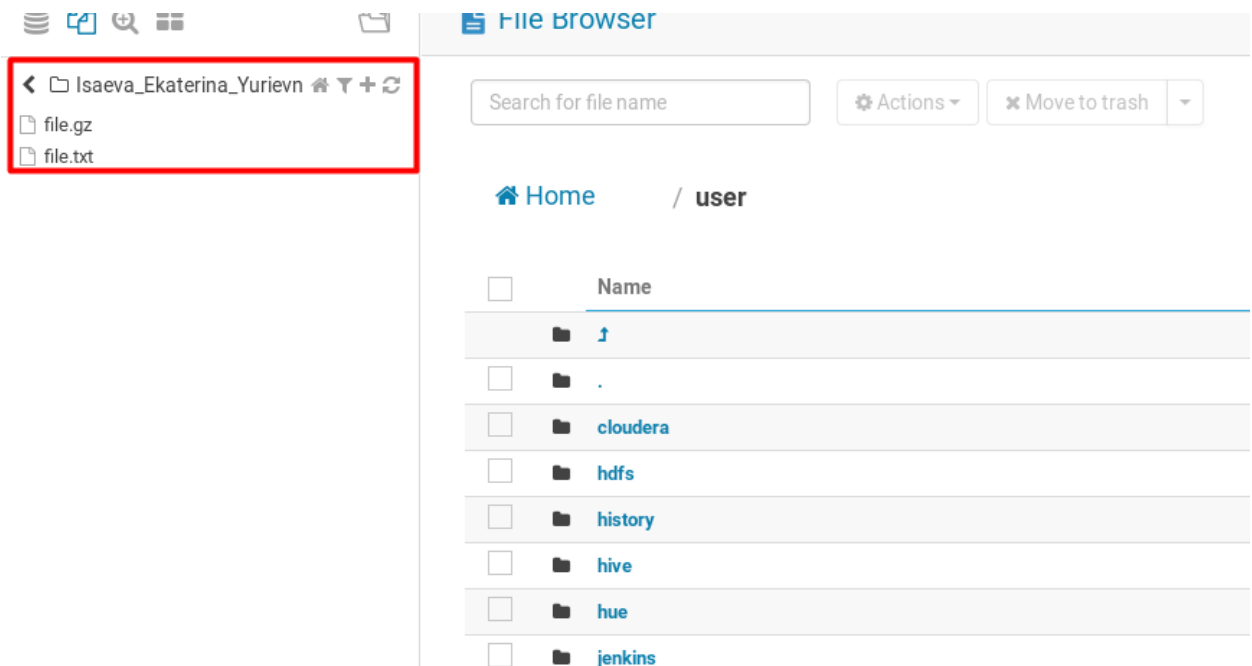
Computer

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

```
root@q bash-4.1$ hdfs dfs -mkdir -p /user/mgpu/Isaeva_Ekaterina_Yurievna
bash-4.1$ ^C
sp 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$
```

root@q

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
cat file.txt bash-4.1$
```

Text + head

```
cat file.txt bash-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_Yurievna/$(date +%Y%m%d')_file.txt
bash-4.1$ 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/20231207_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

mgpu

Empty directory

Search for file

Home

Name

↑

.

cloudera

hdfs

history

hive

hue

jenkins

mgpu

oozie

rook

spool

Show 45

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\$ hdfs dfs -mkdir /user/mgpu/Isaeva_Ekaterina_Yurievna

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\$ 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

bash-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\$ hdfs dfs -head /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt

-head: Unknown command

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\$ hdfs dfs -cp /user/mgpu/Isaeva_Ekaterina_Yurievna/file.txt /user/mgpu/Isaeva_Ekaterina_Yurievna/\$(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/20231207_file.txt

-rw-r--r-- 1 hdfs supergroup 30 2023-12-07 09:02 /user/mgpu/Isaeva_Ekaterina_Yurievna/file.

-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\$

root@quickstart:~

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

Cloudera Hue Hadoop HBase Impala Spool

mgpu

sample_text.txt

Search for file name

Home

Name

↑

.

cloudera

hdfs

history

hive

hue

jenkins

madu

```
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.RMPProxy: 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$ █ Computer
```

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/hadoop-hdfs/.pigbootup not found  
2023-12-07 09:34:55,960 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:55,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name 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.tracker 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.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:58,526 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:58,527 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:58,667 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:58,668 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:58,822 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:58,827 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:58,986 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:58,989 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:59,141 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:59,143 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:59,284 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:59,287 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:59,445 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:59,445 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2023-12-07 09:34:59,565 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2023-12-07 09:34:59,567 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
brunt> ■ root@quickstart:~
```

В Неможно 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(MapReduceLauncher.java:270)  
Caused by: 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.singleThreadedListStatus(FileInputFormat.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.get_splits(FileInputFormat.java:387)  
at org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigInputFormat.get_splits(PigInputFormat.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.MapReduceLauncher - Failed!  
2023-12-07 09:51:18,500 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1066: Unable to open iterator for alias geoloc_limit  
Details at logfile: /var/lib/hadoop-hdfs/pig_1701970494712.log  
grunt> /
```

```
56.71s [?] ?  
1 geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid, driverid, event, latitude, longitude, city, state, velocity, event_ind, idling_ind);  
2  
3 geoloc_limit = LIMIT geoloc 10;  
4  
5 DUMP geoloc_limit;  
6  
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;
```


Cloudera Hue interface showing a Pig script execution. The script reads a CSV file and calculates the limit of geolocations per truck. The results are displayed in a table with 208 rows.

```
1 geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid, driverid, event, latitude, longitude, city, state, velocity, event_ind, idlis);
2
3 geoloc_limit = LIMIT geoloc 10;
4
5 DUMP geoloc_limit;
```

Header
>>> Invoking Pig command line now >>>
1
2
3
4 Run pig script using PigRunner.run() for Pig version 0.8+
5 Apache Pig version 0.12.0-cdh5.13.0 (exported)
6 compiled Oct 04 2017, 11:09:03
7
8 Run pig script using PigRunner.run() for Pig version 0.8+
9 2023-12-07 09:54:30,974 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (exported) compiled Oct 04 2017, 11:09:03
10 2023-12-07 09:54:30,975 [main] INFO org.apache.pig.Main - Logging error messages to: /var/lib/hadoop-yarn/cache/yarn-rm-local-dir/usercache/cloudera/appcache/application_1701963248547_01
11 2023-12-07 09:54:31,202 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /var/lib/hadoop-yarn/pigbootstrap not found
12 2023-12-07 09:54:31,602 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
13 root@quickstart:~

Посчитать статистику по этому файлу

Cloudera Hue interface showing a Pig script execution. The script calculates the count of geolocations per truck. The results are displayed in a table with 284 rows.

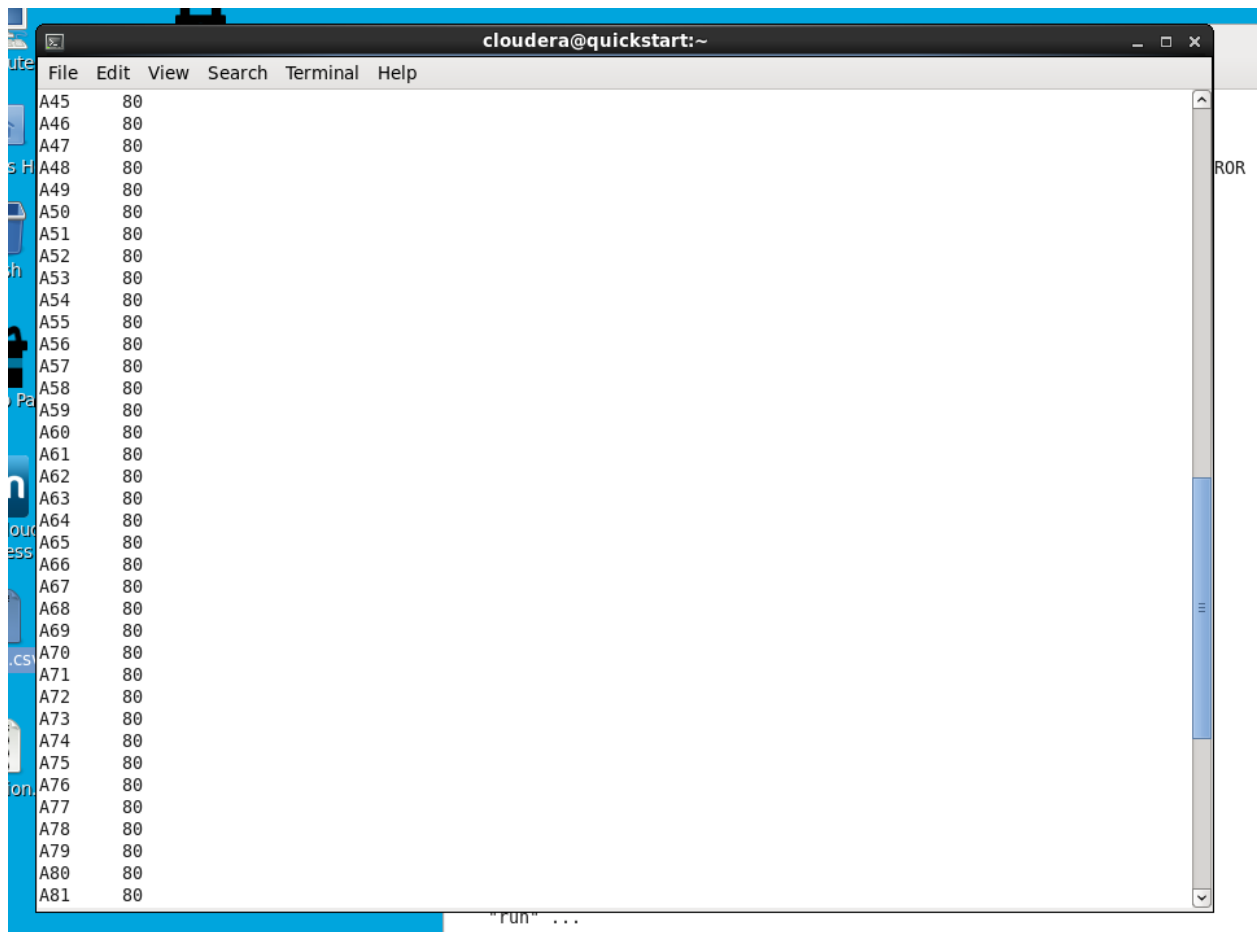
```
1 Example: 1 + 1, or press CTRL + space
```

Header
>>> Invoking Pig command line now >>>
1
2
3
4 Run pig script using PigRunner.run() for Pig version 0.8+
5 Apache Pig version 0.12.0-cdh5.13.0 (exported)
6 compiled Oct 04 2017, 11:09:03
7
8 Run pig script using PigRunner.run() for Pig version 0.8+
9 2023-12-07 09:57:49,566 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (exported) compiled Oct 04 2017, 11:09:03
10 2023-12-07 09:57:49,568 [main] INFO org.apache.pig.Main - Logging error messages to: /var/lib/hadoop-yarn/cache/yarn-rm-local-dir/usercache/cloudera/appcache/application_1701963248547_01
11 2023-12-07 09:57:49,640 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /var/lib/hadoop-yarn/pigbootstrap not found
12 2023-12-07 09:57:49,748 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address

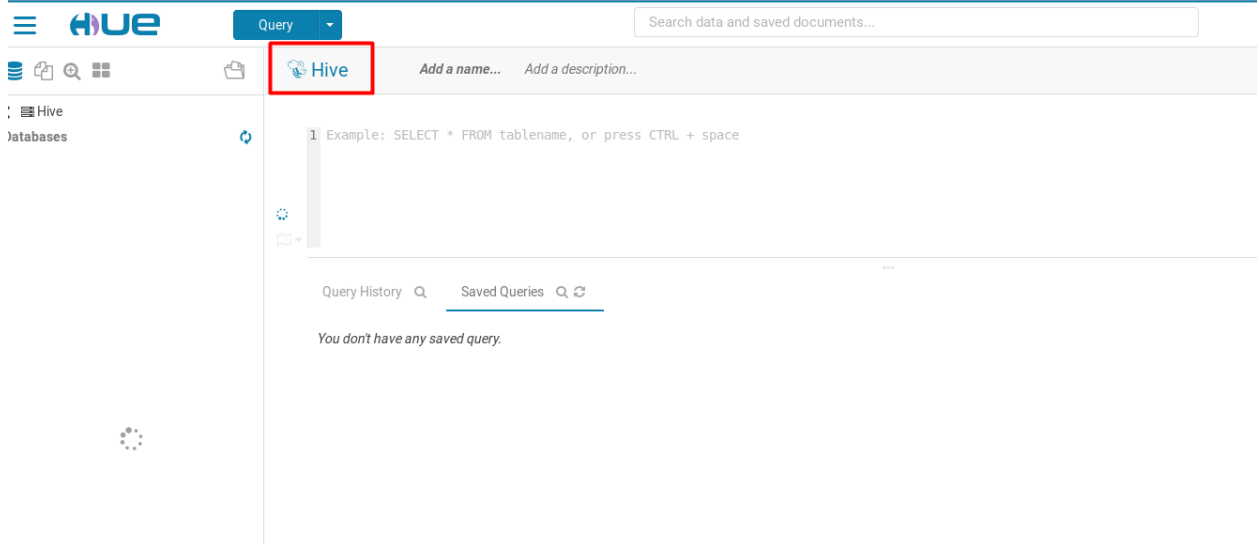
Cloudera Hue interface showing a Pig script execution. The script calculates the count of geolocations per truck. The results are displayed in a table with 284 rows.

```
1 truck_ids = GROUP geoloc BY truckid;
2
3 result = FOREACH truck_ids GENERATE group AS truckid, COUNT(geoloc) as count;
4
5 STORE result INTO 'results';
6
7 DUMP result;
```

Header
>>> Invoking Pig command line now >>>
1
2
3
4 Run pig script using PigRunner.run() for Pig version 0.8+
5 Apache Pig version 0.12.0-cdh5.13.0 (exported)
6 compiled Oct 04 2017, 11:09:03
7
8 Run pig script using PigRunner.run() for Pig version 0.8+
9 2023-12-07 09:57:49,566 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (exported) compiled Oct 04 2017, 11:09:03
10 2023-12-07 09:57:49,568 [main] INFO org.apache.pig.Main - Logging error messages to: /var/lib/hadoop-yarn/cache/yarn-rm-local-dir/usercache/cloudera/appcache/application_1701963248547_01
11 2023-12-07 09:57:49,640 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /var/lib/hadoop-yarn/pigbootstrap not found
12 2023-12-07 09:57:49,748 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address



Переключаемся на HIVE



0sdefaulttext?

```
1 CREATE EXTERNAL TABLE geolocation (  
2   truckid STRING,  
3   driverid STRING,  
4   event STRING,  
5   latitude DOUBLE,  
6   longitude DOUBLE,  
7   city STRING,  
8   state STRING,  
9   velocity DOUBLE,  
10  event_ind BIGINT,  
11  idling_ind BIGINT  
12 )  
13 ROW FORMAT DELIMITED  
14 FIELDS TERMINATED BY ','  
15 LOCATION '/user/cloudera/geoloc';
```

Success.

Query HistorySaved 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

0sdefaulttext?

```
1 SELECT truckid FROM geolocation LIMIT 10;  
2
```

Query HistorySaved QueriesResults (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

HiveAdd a name...Add a description...

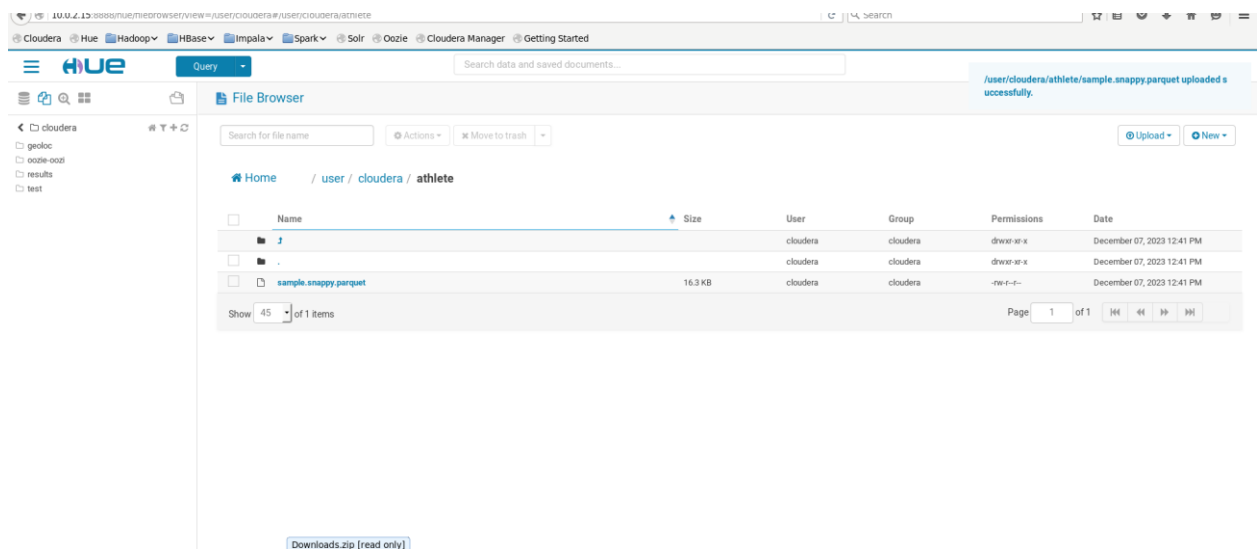
37.86sdefaulttext?

```
1 SELECT truckid, city, AVG(velocity) AS avg_speed  
2 FROM geolocation  
3 GROUP BY truckid, city;  
4
```

Query HistorySaved QueriesResults (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	Oceanos	47.142857142857146
12 A1	Palmdale	26.600000000000001

Скачиваем файл



Выполняем запрос

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

```
1 CREATE EXTERNAL TABLE athlete (  
2   ID INT,  
3   Name STRING,  
4   Sex STRING,  
5   Age INT,  
6   Height INT,  
7   Weight INT,  
8   Team STRING,  
9   NOC STRING,  
10  Games STRING,  
11  `Year` INT,  
12  Season STRING,  
13  City STRING,  
14  Sport STRING,  
15  Event STRING,  
16  Medal STRING  
17 )  
18 STORED AS PARQUET  
19 LOCATION '/user/cloudera/athlete';
```

▼ Success.

Query History Saved Queries

a few seconds ago	✓	CREATE EXTERNAL TABLE athlete (ID INT, Name STRING, Sex STRING, Age INT, Height INT, Weight INT, Team STRING, NOC STRING, Games STRING, `Year` INT, Season STRING, City STRING, Sport STRING, Event STRING, Medal STRING) STORED AS PARQUET LOCATION '/user/cloudera/athlete'
2 hours ago	✓	SELECT truckid, city, AVG(velocity) AS avg_speed FROM geolocation GROUP BY truckid, city

Переходим в Impala

Impala Add a name... Add a description...

```
1 invalidate metadata;  
2  
3 show tables;
```

▼ Done. 0 results.

Query History Saved Queries

a few seconds ago	✗	show tables
3 hours ago	!	geoloc = LOAD 'geoloc/geolocation.csv' USING PigStorage(',') AS (truckid:chararray, driverid:chararray, event:chararray, latitude:double, longitude:double, city:chararray, state:chararray, velocity:double, event_ind:long, idling_ind:long); truck_ids = GROUP geoloc BY truckid; result = FOREACH truck_ids GENERATE group AS truckid, COUNT(geoloc) as count; STORE result INTO 'results'; DUMP result;

Переходим в Hive и выполняем задание

```
2 truckid,
3 COUNT(*) AS location_count
4 FROM
5 geolocation
6 GROUP BY
7 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

```
1 SELECT
2 latitude,
3 longitude
4 FROM
5 geolocation
6 WHERE
7 truckid = 'A80';
```

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	-119.242889000000001
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	-122.238892000000001
12	38.019365999999998	-122.13413199999999
13	36.471865000000001	-6.1965950000000003

Tables

Search...

No tables