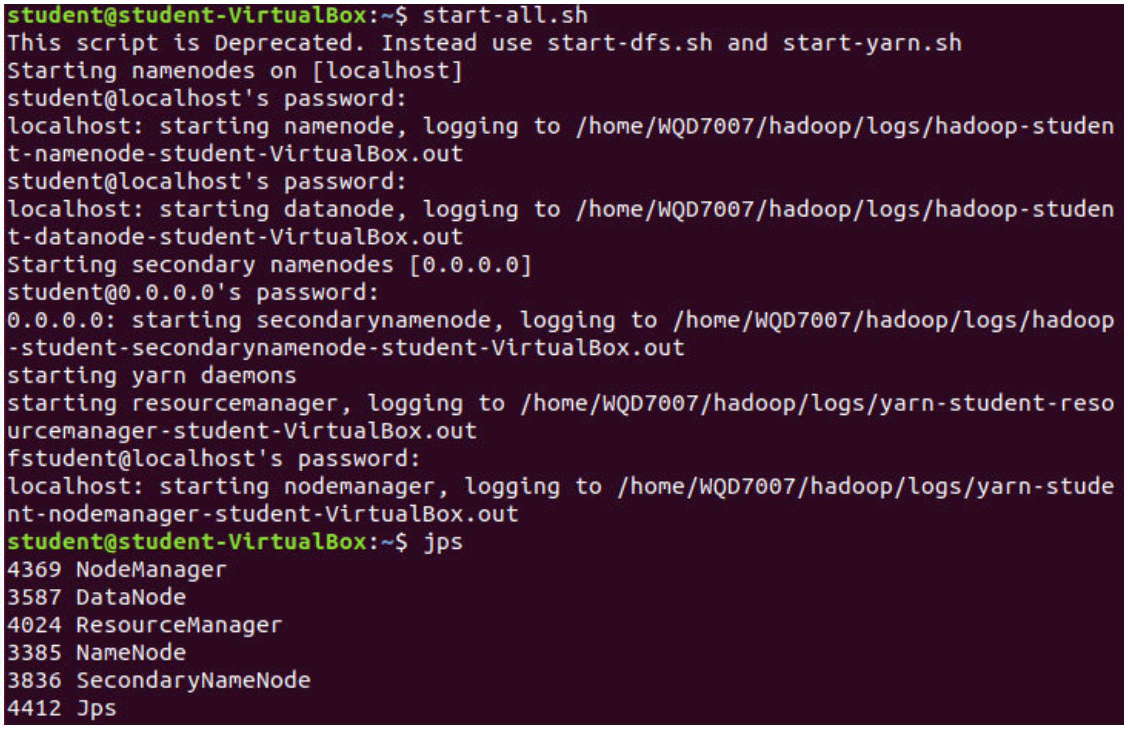
Author: LIU,HONGYANG

Matrix Number: 17201091/1

Before the Lab Test:

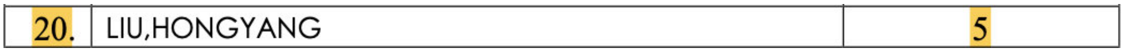
Start the Hadoop and check the daemons:

```
start-all.sh
jps
```



```
student@student-VirtualBox:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
student@localhost's password:
localhost: starting namenode, logging to /home/WQD7007/hadoop/logs/hadoop-studen
t-namenode-student-VirtualBox.out
student@localhost's password:
localhost: starting datanode, logging to /home/WQD7007/hadoop/logs/hadoop-studen
t-datanode-student-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
student@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/WQD7007/hadoop/logs/hadoop
-student-secondarynamenode-student-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /home/WQD7007/hadoop/logs/yarn-student-reso
urcemanager-student-VirtualBox.out
fstudent@localhost's password:
localhost: starting nodemanager, logging to /home/WQD7007/hadoop/logs/yarn-stude
nt-nodemanager-student-VirtualBox.out
student@student-VirtualBox:~$ jps
4369 NodeManager
3587 DataNode
4024 ResourceManager
3385 NameNode
3836 SecondaryNameNode
4412 Jps
```

DownLoad dabasets:



20.	LIU,HONGYANG	5

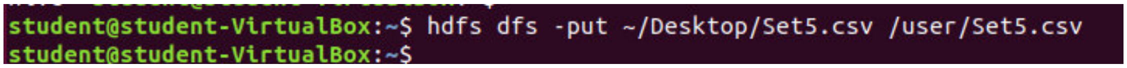
LabTest:

Part 1:

1.Import the downloaded dataset to HDFS

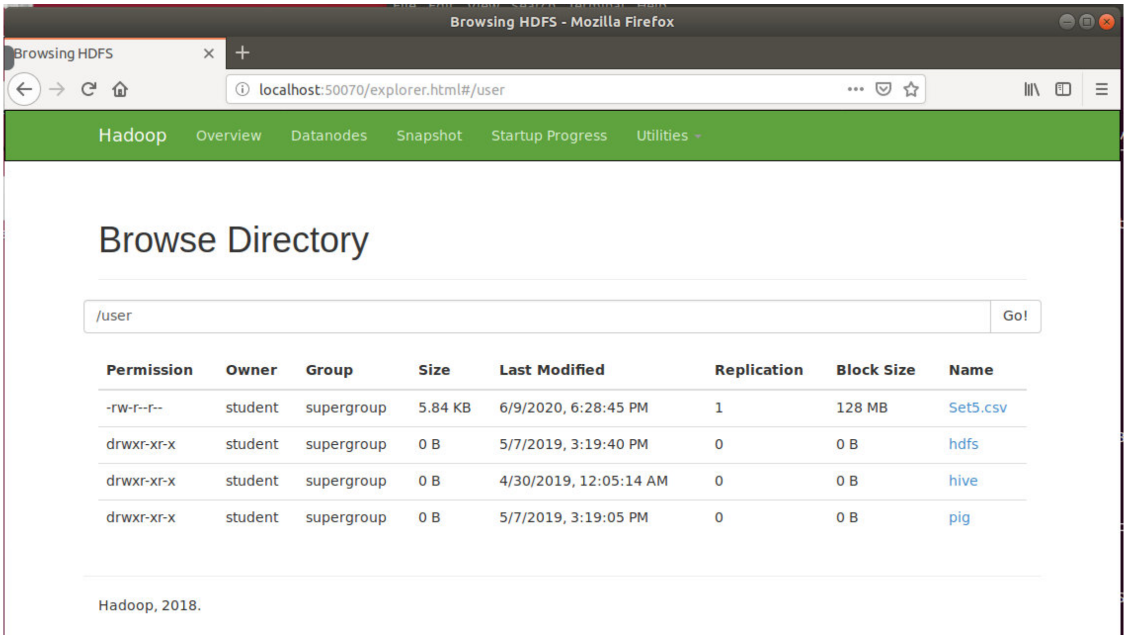
code:

```
hdfs dfs -put ~/Desktop/Set5.csv /user/Set5.csv
```



```
student@student-VirtualBox:~$ hdfs dfs -put ~/Desktop/Set5.csv /user/Set5.csv
student@student-VirtualBox:~$
```

Results:



Browse Directory

/user Go!

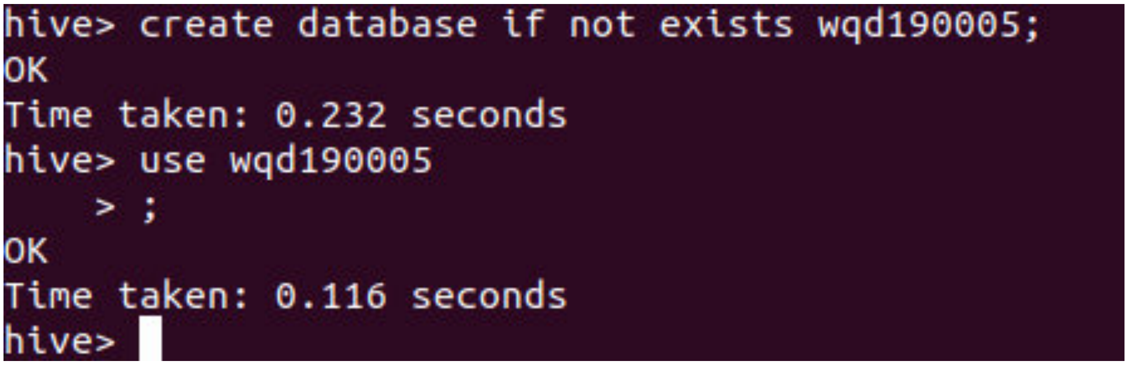
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	student	supergroup	5.84 KB	6/9/2020, 6:28:45 PM	1	128 MB	Set5.csv
drwxr-xr-x	student	supergroup	0 B	5/7/2019, 3:19:40 PM	0	0 B	hdfs
drwxr-xr-x	student	supergroup	0 B	4/30/2019, 12:05:14 AM	0	0 B	hive
drwxr-xr-x	student	supergroup	0 B	5/7/2019, 3:19:05 PM	0	0 B	pig

Hadoop, 2018.

2.By using Hive or Pig, identify 5 rows of data that have the

1. highest reading score.
2. lowest CGPA.

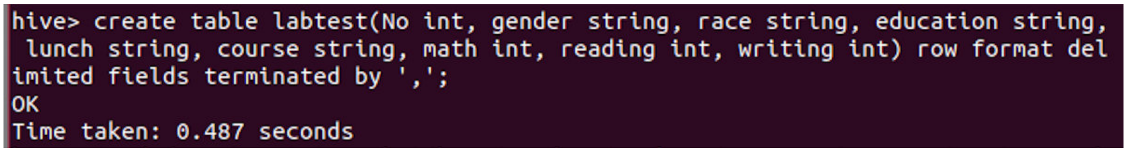
Create database:



```
hive> create database if not exists wqd190005;
OK
Time taken: 0.232 seconds
hive> use wqd190005
    > ;
OK
Time taken: 0.116 seconds
hive> 
```

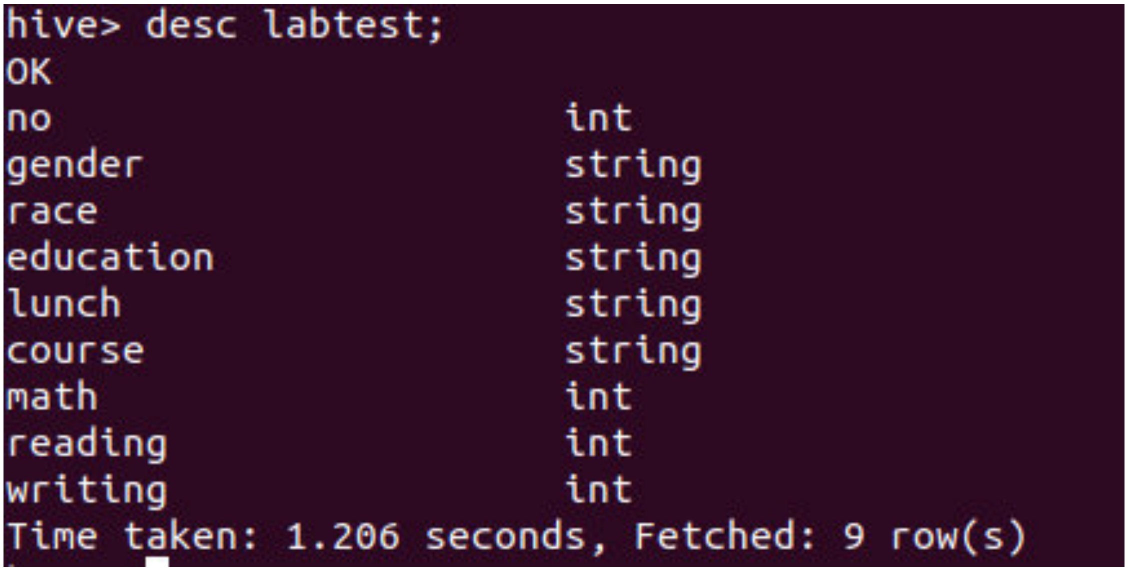
Create table:

```
create table labtest(No int, gender string, race string, education string, lunch string, course string, math int, reading int, writing int) row format delimited fields terminated by ',';
```



```
hive> create table labtest(No int, gender string, race string, education string,
 lunch string, course string, math int, reading int, writing int) row format del
imited fields terminated by ',';
OK
Time taken: 0.487 seconds
```

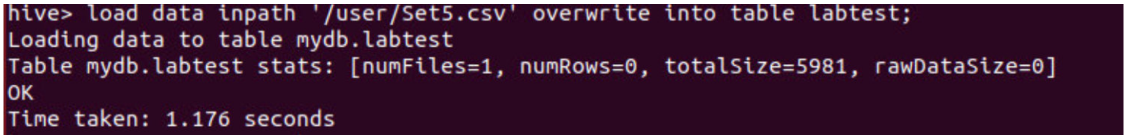
```
desc labtest;
```



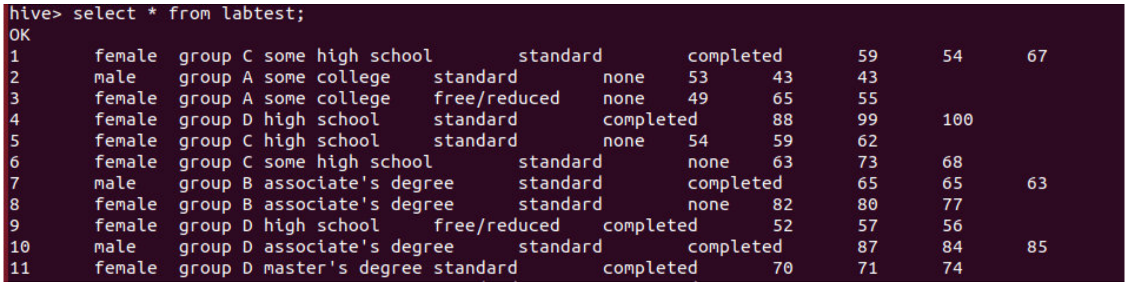
```
hive> desc labtest;
OK
no                      int
gender                  string
race                    string
education               string
lunch                   string
course                  string
math                    int
reading                 int
writing                 int
Time taken: 1.206 seconds, Fetched: 9 row(s)
```

load data from hdfs to hive:

```
load data inpath '/user/Set5.csv' overwrite into table labtest;
```



```
hive> load data inpath '/user/Set5.csv' overwrite into table labtest;
Loading data to table mydb.labtest
Table mydb.labtest stats: [numFiles=1, numRows=0, totalSize=5981, rawDataSize=0]
OK
Time taken: 1.176 seconds
```

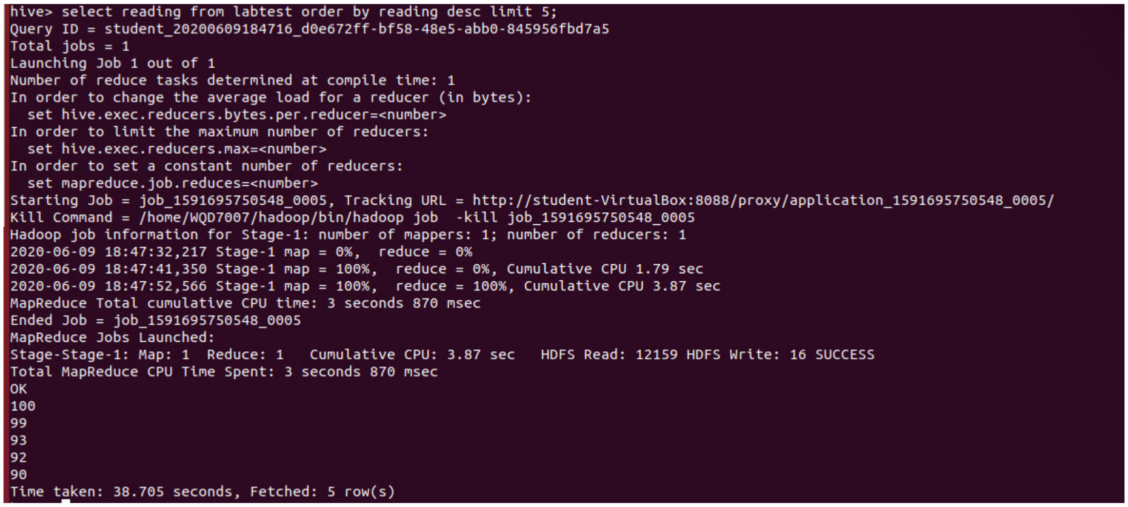


```
hive> select * from labtest;
OK
1       female  group C some high school        standard        completed       59      54      67
2       male    group A some college    standard        none    53      43      43
3       female  group A some college    free/reduced    none    49      65      55
4       female  group D high school     standard        completed       88      99      100
5       female  group C high school     standard        none    54      59      62
6       female  group C some high school        standard        none    63      73      68
7       male    group B associate's degree      standard        completed       65      65      63
8       female  group B associate's degree      standard        none    82      80      77
9       female  group D high school     free/reduced    completed       52      57      56
10      male    group D associate's degree      standard        completed       87      84      85
11      female  group D master's degree standard        completed       70      71      74
```

Answer 1:

highest reading score(5 rows):

```
select reading from labtest order by reading desc limit 5;
```



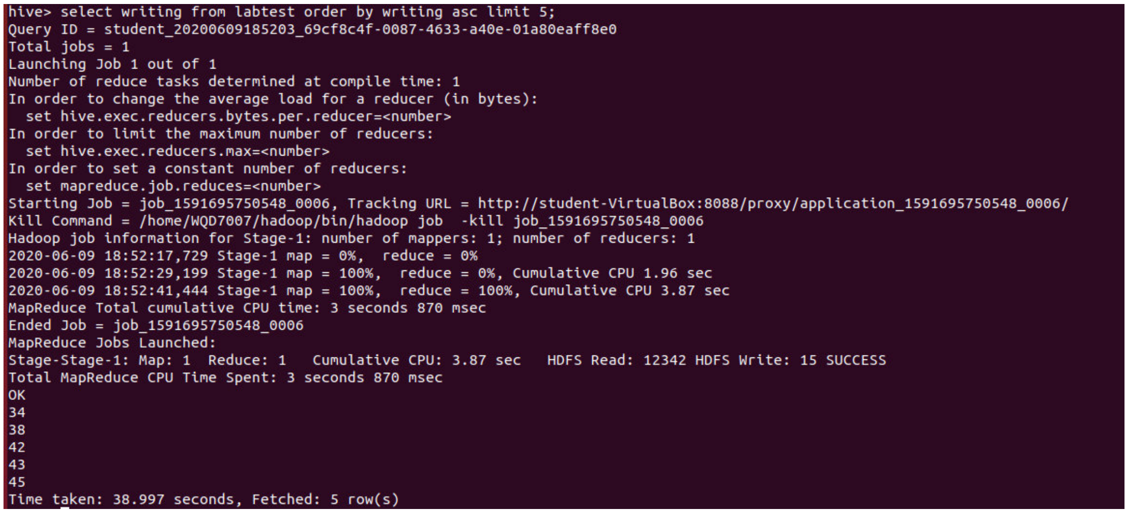
```
hive> select reading from labtest order by reading desc limit 5;
Query ID = student_20200609184716_d0e672ff-bf58-48e5-abb0-845956fbd7a5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1591695750548_0005, Tracking URL = http://student-VirtualBox:8088/proxy/application_1591695750548_0005/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1591695750548_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-09 18:47:32,217 Stage-1 map = 0%,  reduce = 0%
2020-06-09 18:47:41,350 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.79 sec
2020-06-09 18:47:52,566 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.87 sec
MapReduce Total cumulative CPU time: 3 seconds 870 msec
Ended Job = job_1591695750548_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.87 sec   HDFS Read: 12159 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 870 msec
OK
100
99
93
92
90
Time taken: 38.705 seconds, Fetched: 5 row(s)
```

They are 100, 99, 93, 92, 90

Answer 2:

lowest writing score(5 rows):

```
select writing from labtest order by writing asc limit 5;
```



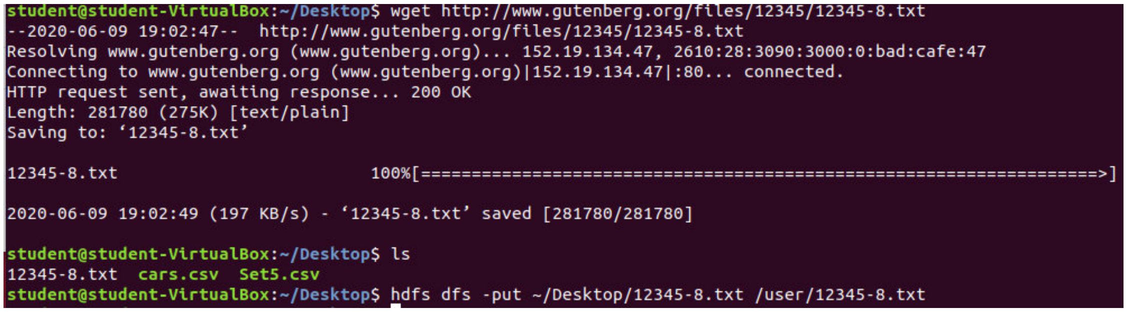
```
hive> select writing from labtest order by writing asc limit 5;
Query ID = student_20200609185203_69cf8c4f-0087-4633-a40e-01a80eaff8e0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1591695750548_0006, Tracking URL = http://student-VirtualBox:8088/proxy/application_1591695750548_0006/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1591695750548_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-09 18:52:17,729 Stage-1 map = 0%,  reduce = 0%
2020-06-09 18:52:29,199 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.96 sec
2020-06-09 18:52:41,444 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.87 sec
MapReduce Total cumulative CPU time: 3 seconds 870 msec
Ended Job = job_1591695750548_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.87 sec   HDFS Read: 12342 HDFS Write: 15 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 870 msec
OK
34
38
42
43
45
Time taken: 38.997 seconds, Fetched: 5 row(s)
```

They are 34, 38, 42,43,45

Part 2:

download file and upload it to hdfs

```
wget http://www.gutenberg.org/files/12345/12345-8.txt
hdfs dfs -put ~/Desktop/12345-8.txt /user/12345-8.txt
```



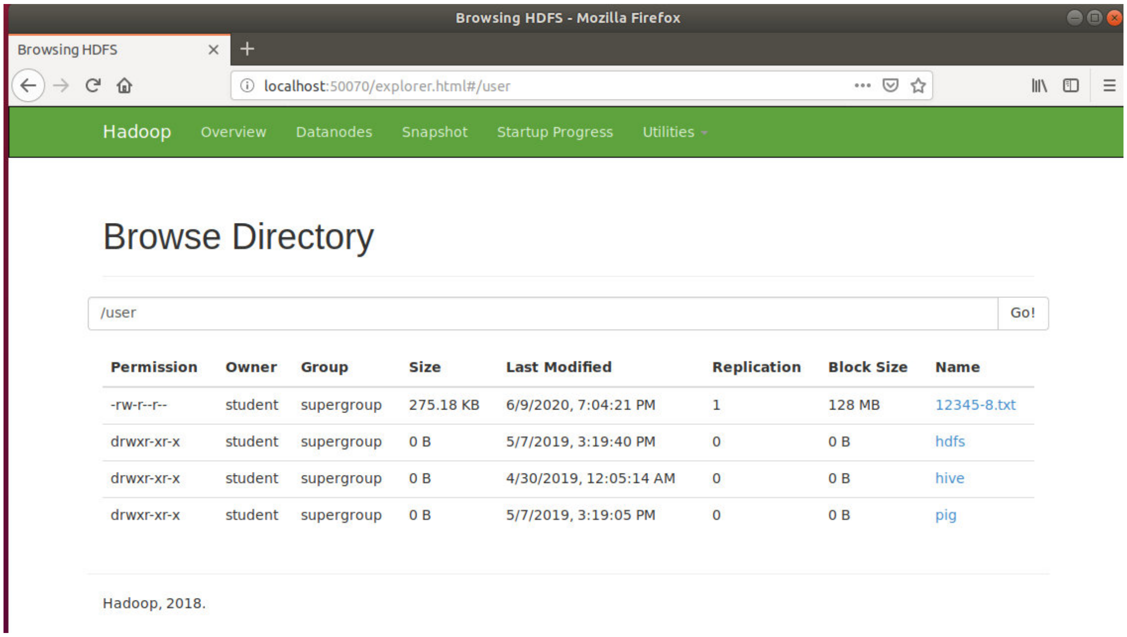
```
student@student-VirtualBox:~/Desktop$ wget http://www.gutenberg.org/files/12345/12345-8.txt
--2020-06-09 19:02:47--  http://www.gutenberg.org/files/12345/12345-8.txt
Resolving www.gutenberg.org (www.gutenberg.org)... 152.19.134.47, 2610:28:3090:3000:0:bad:cafe:47
Connecting to www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 281780 (275K) [text/plain]
Saving to: ‘12345-8.txt’

12345-8.txt                      100%[===================================================================>]

2020-06-09 19:02:49 (197 KB/s) - ‘12345-8.txt’ saved [281780/281780]

student@student-VirtualBox:~/Desktop$ ls
12345-8.txt  cars.csv  Set5.csv
student@student-VirtualBox:~/Desktop$ hdfs dfs -put ~/Desktop/12345-8.txt /user/12345-8.txt
```

Answer 1:



Browse Directory

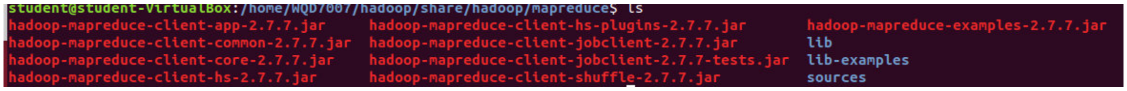
/user

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	student	supergroup	275.18 KB	6/9/2020, 7:04:21 PM	1	128 MB	12345-8.txt
drwxr-xr-x	student	supergroup	0 B	5/7/2019, 3:19:40 PM	0	0 B	hdfs
drwxr-xr-x	student	supergroup	0 B	4/30/2019, 12:05:14 AM	0	0 B	hive
drwxr-xr-x	student	supergroup	0 B	5/7/2019, 3:19:05 PM	0	0 B	pig

Hadoop, 2018.

Answer 2:

enter the file example file: hadoop-mapreduce-examples-2.7.7.jar



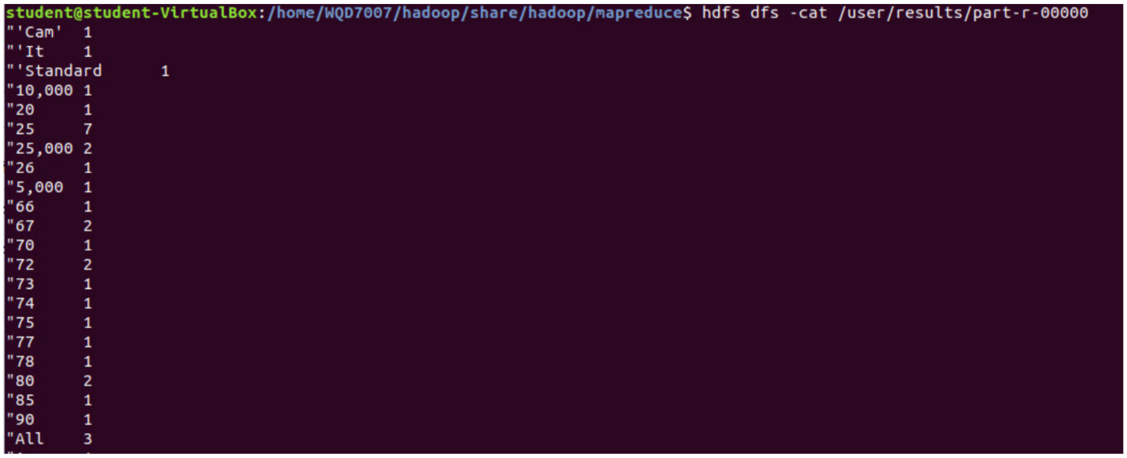
```
student@student-VirtualBox:/home/WQD7007/hadoop/share/hadoop/mapreduce$ ls
hadoop-mapreduce-client-app-2.7.7.jar     hadoop-mapreduce-client-hs-plugins-2.7.7.jar       hadoop-mapreduce-examples-2.7.7.jar
hadoop-mapreduce-client-common-2.7.7.jar  hadoop-mapreduce-client-jobclient-2.7.7.jar        lib
hadoop-mapreduce-client-core-2.7.7.jar    hadoop-mapreduce-client-jobclient-2.7.7-tests.jar  lib-examples
hadoop-mapreduce-client-hs-2.7.7.jar      hadoop-mapreduce-client-shuffle-2.7.7.jar          sources
```

run shell:

```
hadoop jar hadoop-mapreduce-examples-2.7.7.jar wordcount /user/12345-8.txt
/user/results
```

check results:

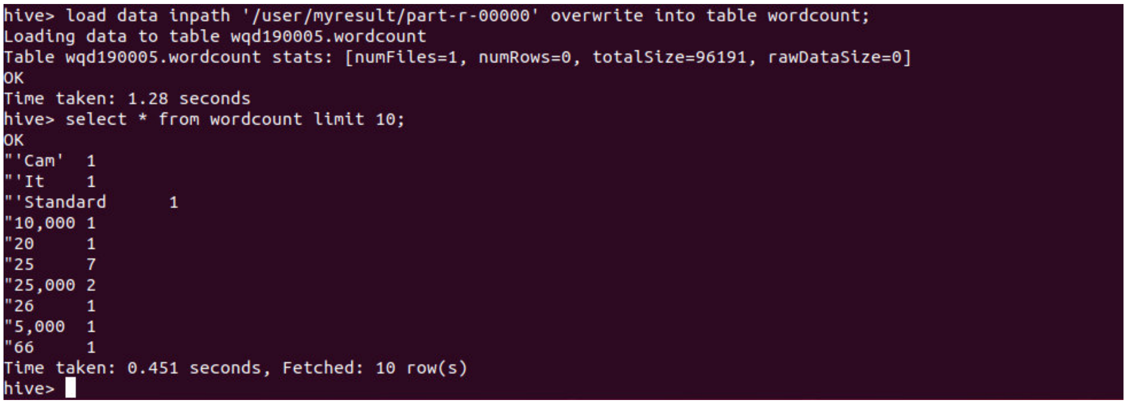
```
hdfs dfs -cat /user/results/part-r-0000
```



```
student@student-VirtualBox:/home/WQD7007/hadoop/share/hadoop/mapreduce$ hdfs dfs -cat /user/results/part-r-00000
"'Cam'  1
"'It    1
"'Standard      1
"10,000 1
"20     1
"25     7
"25,000 2
"26     1
"5,000  1
"66     1
"67     2
"70     1
"72     2
"73     1
"74     1
"75     1
"77     1
"78     1
"80     2
"85     1
"90     1
"All    3
```

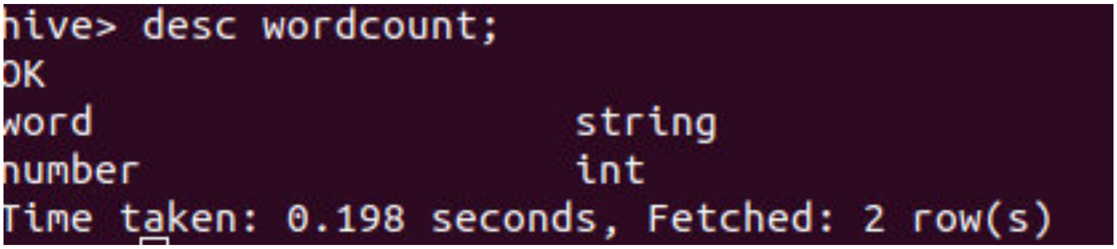
Answer 3:

```
create table wordcount(word string, number int) row format delimited fields
terminated by '\t';
```



```
hive> load data inpath '/user/myresult/part-r-00000' overwrite into table wordcount;
Loading data to table wqd190005.wordcount
Table wqd190005.wordcount stats: [numFiles=1, numRows=0, totalSize=96191, rawDataSize=0]
OK
Time taken: 1.28 seconds
hive> select * from wordcount limit 10;
OK
"'Cam'  1
"'It    1
"'Standard      1
"10,000 1
"20     1
"25     7
"25,000 2
"26     1
"5,000  1
"66     1
Time taken: 0.451 seconds, Fetched: 10 row(s)
hive>
```

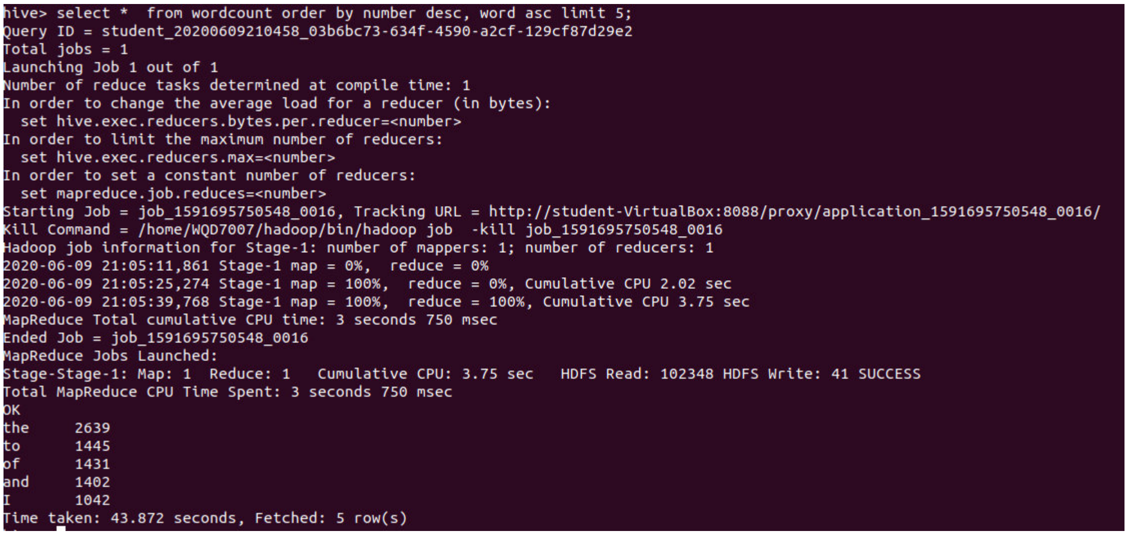
desc wordcount;



```
hive> desc wordcount;
OK
word                    string
number                  int
Time taken: 0.198 seconds, Fetched: 2 row(s)
```

a

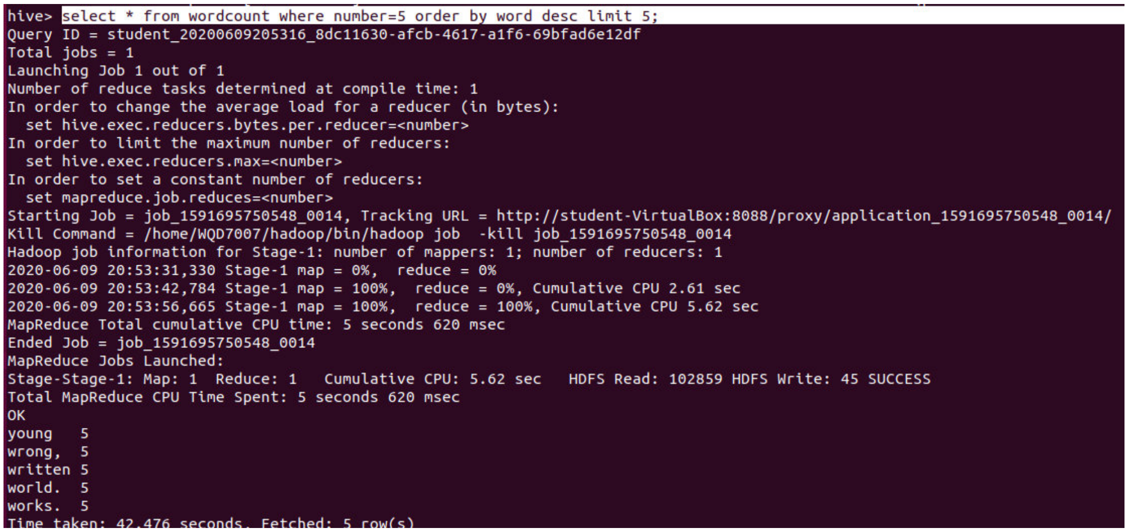
select * from wordcount order by number desc, word asc limit 5;



```
hive> select *  from wordcount order by number desc, word asc limit 5;
Query ID = student_20200609210458_03b6bc73-634f-4590-a2cf-129cf87d29e2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1591695750548_0016, Tracking URL = http://student-VirtualBox:8088/proxy/application_1591695750548_0016/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1591695750548_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-09 21:05:11,861 Stage-1 map = 0%,  reduce = 0%
2020-06-09 21:05:25,274 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.02 sec
2020-06-09 21:05:39,768 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.75 sec
MapReduce Total cumulative CPU time: 3 seconds 750 msec
Ended Job = job_1591695750548_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.75 sec   HDFS Read: 102348 HDFS Write: 41 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 750 msec
OK
the     2639
to      1445
of      1431
and     1402
I       1042
Time taken: 43.872 seconds, Fetched: 5 row(s)
```

b

select * from wordcount where number=5 order by word desc limit 5;



```
hive> select * from wordcount where number=5 order by word desc limit 5;
Query ID = student_20200609205316_8dc11630-afcb-4617-a1f6-69bfad6e12df
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1591695750548_0014, Tracking URL = http://student-VirtualBox:8088/proxy/application_1591695750548_0014/
Kill Command = /home/WQD7007/hadoop/bin/hadoop job  -kill job_1591695750548_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-06-09 20:53:31,330 Stage-1 map = 0%,  reduce = 0%
2020-06-09 20:53:42,784 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.61 sec
2020-06-09 20:53:56,665 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.62 sec
MapReduce Total cumulative CPU time: 5 seconds 620 msec
Ended Job = job_1591695750548_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.62 sec   HDFS Read: 102859 HDFS Write: 45 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 620 msec
OK
young   5
wrong,  5
written 5
world.  5
works.  5
Time taken: 42.476 seconds, Fetched: 5 row(s)
```