# BIGDATA ASSIGNMENT – HDFS,HIVE

## Task 1: HADOOP, HDFS, PIG and HIVE

## 1. Import RStudio Log Files from one week in Jun 2021 into HDFS

Step-1

Download one week Logs from http://cran-logs.rstudio.com/

Step-2

Create a new folder to store downloaded logs files.mkdir RLogFiles

Step-3 Verify files in RStudioLogs folder

ls

Step-3 Unzip all *.zip files

gunzip *.csv.gz

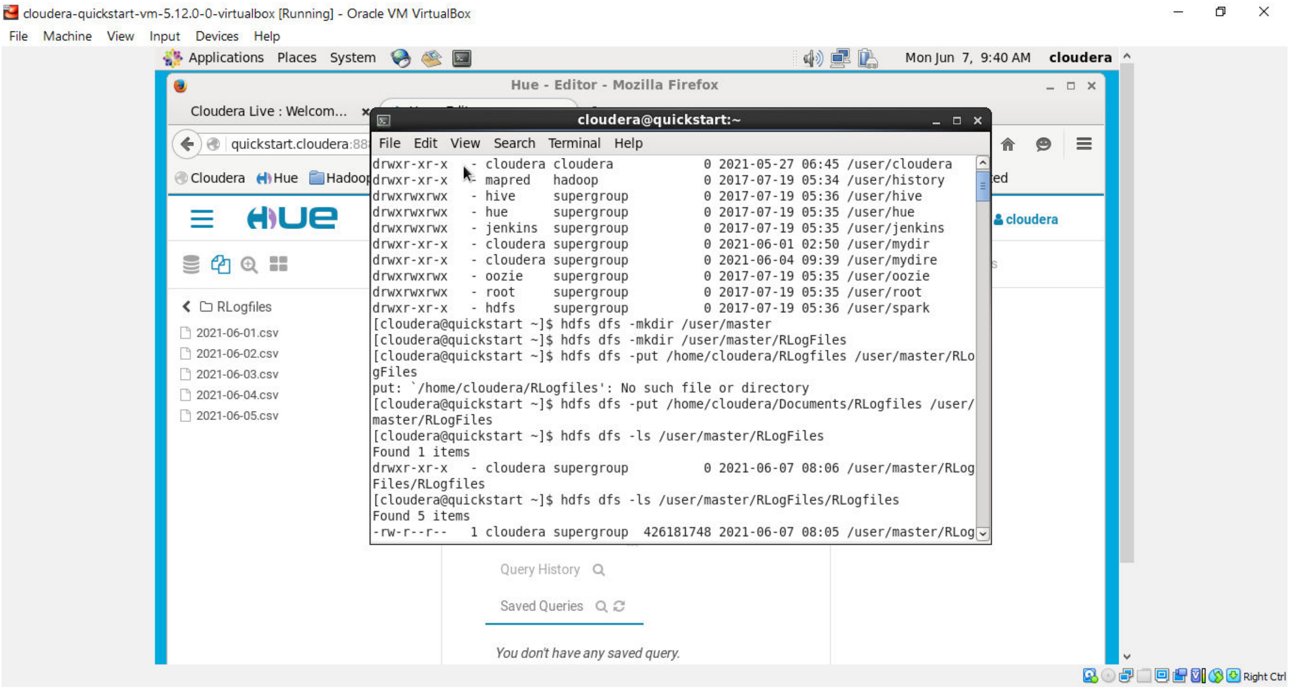
Step-4 Put on ?HDFS and verify

hdfs dfs -mkdir /user/master

hdfs dfs -mkdir /user/master/RLogFiles/

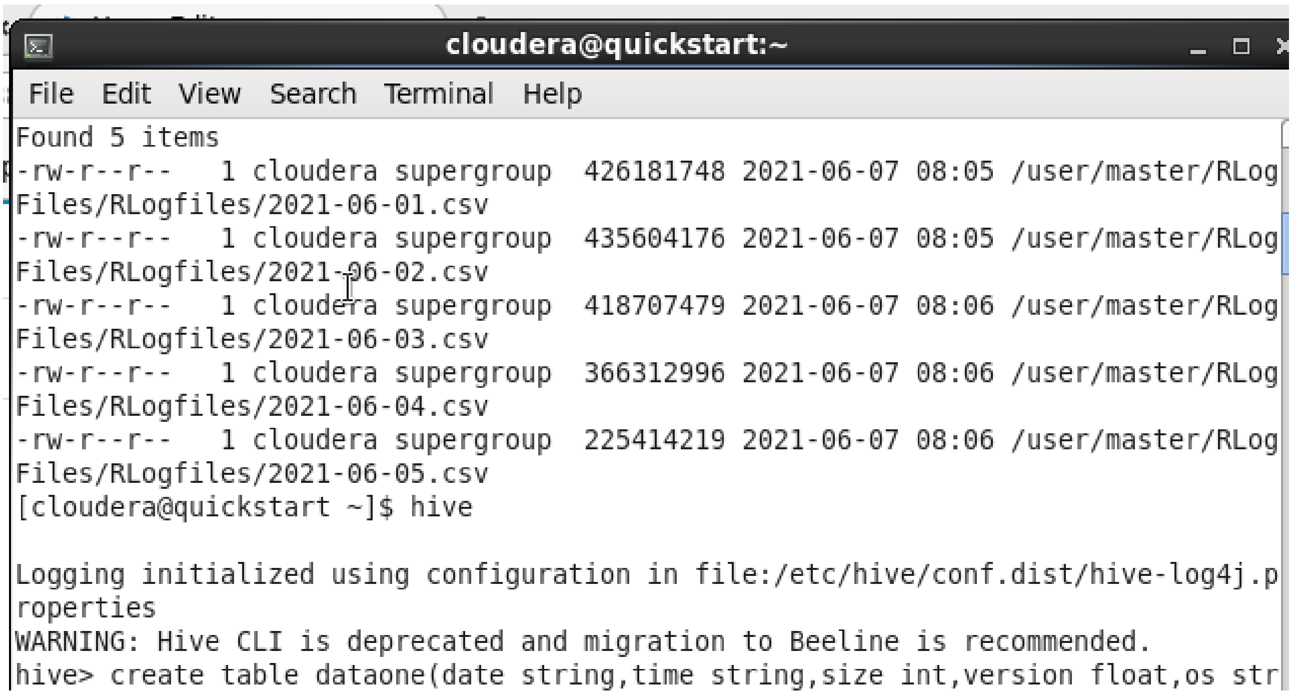
hdfs dfs -put *.csv /user/master/RLogFiles

hdfs dfs -ls /user/master/RlogFiles/RLogfile



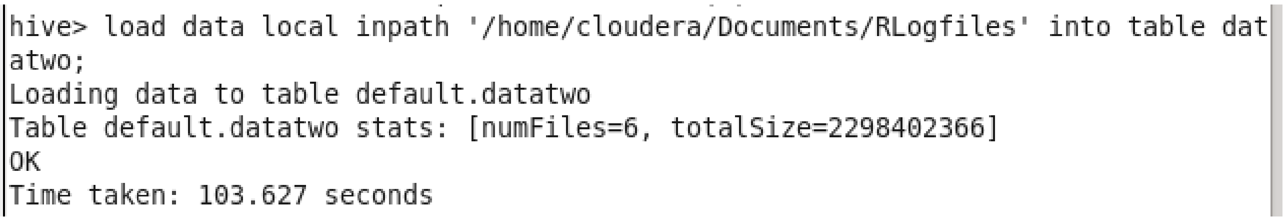
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Jun 7, 9:40 AM cloudera
Hue - Editor - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera:88
Cloudera Hue Hadoop
cloudera
RLogfiles
2021-06-01.csv
2021-06-02.csv
2021-06-03.csv
2021-06-04.csv
2021-06-05.csv
Query History
Saved Queries
You don't have any saved query.
cloudera@quickstart:~
File Edit View Search Terminal Help
Right Ctrl

```
drwxr-xr-x   - cloudera cloudera            0 2021-05-27 06:45 /user/cloudera
drwxr-xr-x   - mapred   hadoop              0 2017-07-19 05:34 /user/history
drwxrwxrwx   - hive     supergroup          0 2017-07-19 05:36 /user/hive
drwxrwxrwx   - hue      supergroup          0 2017-07-19 05:35 /user/hue
drwxrwxrwx   - jenkins  supergroup          0 2017-07-19 05:35 /user/jenkins
drwxr-xr-x   - cloudera supergroup          0 2021-06-01 02:50 /user/mydir
drwxr-xr-x   - cloudera supergroup          0 2021-06-04 09:39 /user/mydire
drwxrwxrwx   - oozie    supergroup          0 2017-07-19 05:35 /user/oozie
drwxrwxrwx   - root     supergroup          0 2017-07-19 05:35 /user/root
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 05:36 /user/spark
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/RLogfiles /user/master/RLo
gFiles
put: `/home/cloudera/RLogfiles': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Documents/RLogfiles /user/
master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -ls /user/master/RLogFiles
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2021-06-07 08:06 /user/master/RLog
Files/RLogfiles
[cloudera@quickstart ~]$ hdfs dfs -ls /user/master/RLogFiles/RLogfiles
Found 5 items
-rw-r--r--   1 cloudera supergroup  426181748 2021-06-07 08:05 /user/master/RLog
```



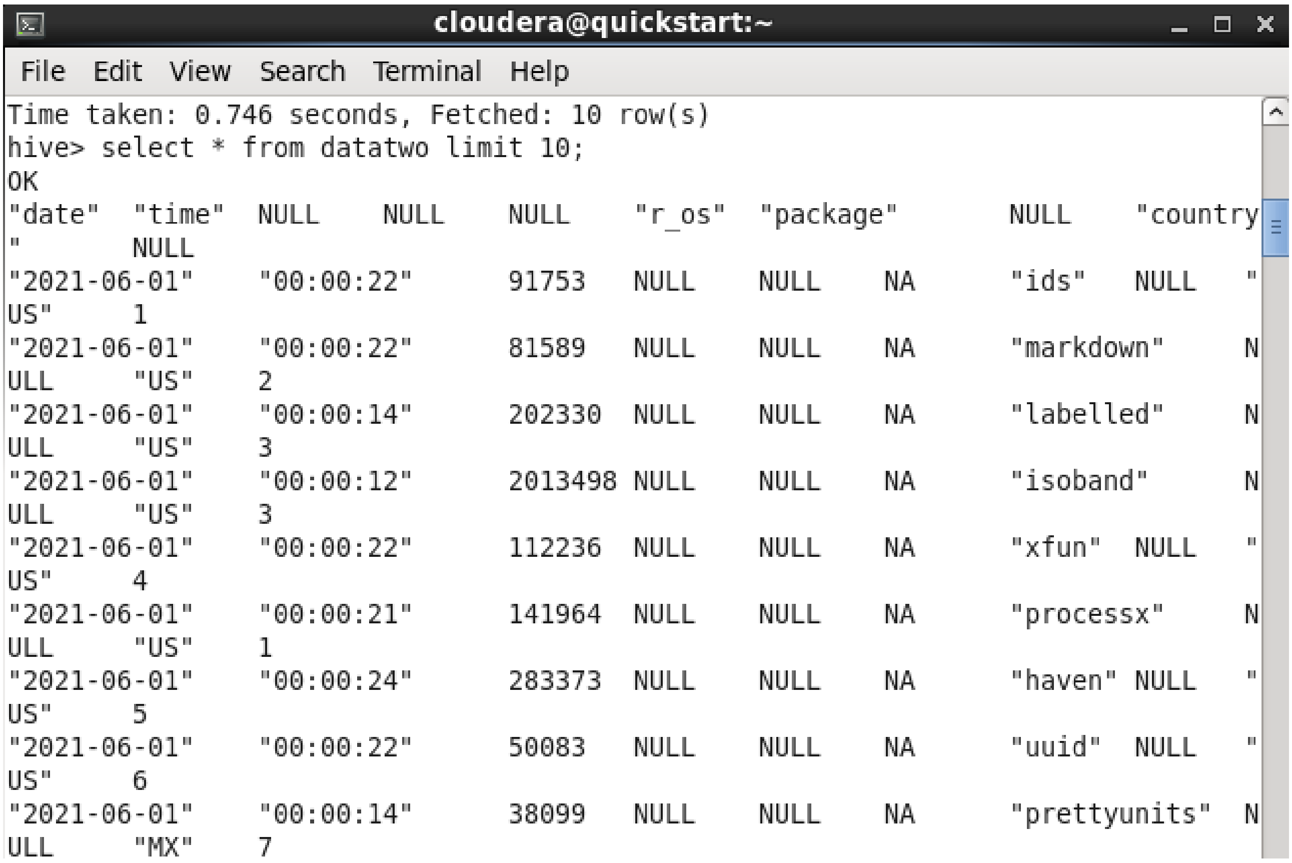
```
Found 5 items
-rw-r--r--   1 cloudera supergroup  426181748 2021-06-07 08:05 /user/master/RLog
Files/RLogfiles/2021-06-01.csv
-rw-r--r--   1 cloudera supergroup  435604176 2021-06-07 08:05 /user/master/RLog
Files/RLogfiles/2021-06-02.csv
-rw-r--r--   1 cloudera supergroup  418707479 2021-06-07 08:06 /user/master/RLog
Files/RLogfiles/2021-06-03.csv
-rw-r--r--   1 cloudera supergroup  366312996 2021-06-07 08:06 /user/master/RLog
Files/RLogfiles/2021-06-04.csv
-rw-r--r--   1 cloudera supergroup  225414219 2021-06-07 08:06 /user/master/RLog
Files/RLogfiles/2021-06-05.csv
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table dataone(date string,time string,size int,version float,os str
```



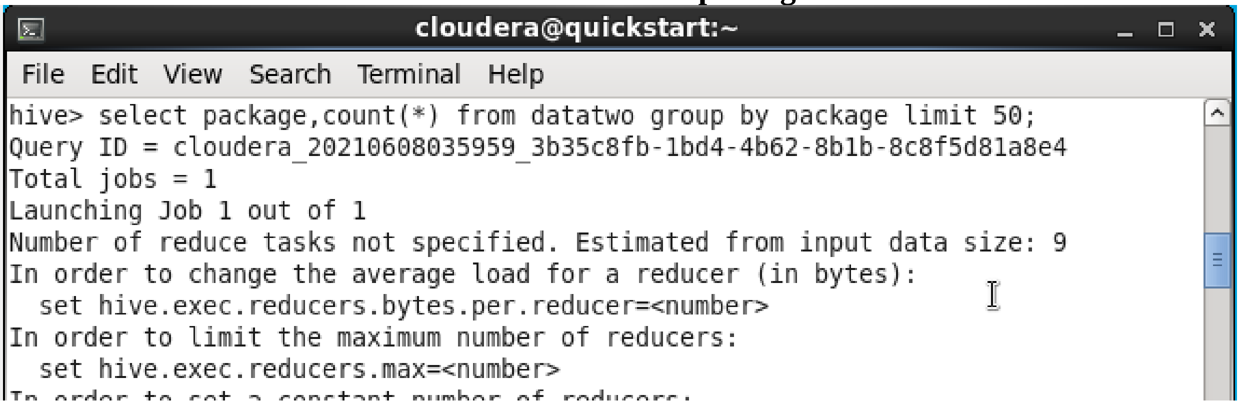
```
hive> load data local inpath '/home/cloudera/Documents/RLogfiles' into table dat
atwo;
Loading data to table default.datatwo
Table default.datatwo stats: [numFiles=6, totalSize=2298402366]
OK
Time taken: 103.627 seconds
```

**a)Load log-file of one day b. Dump the first 10 entries on screen (attach a screen shot into your report) to check if it works or not**

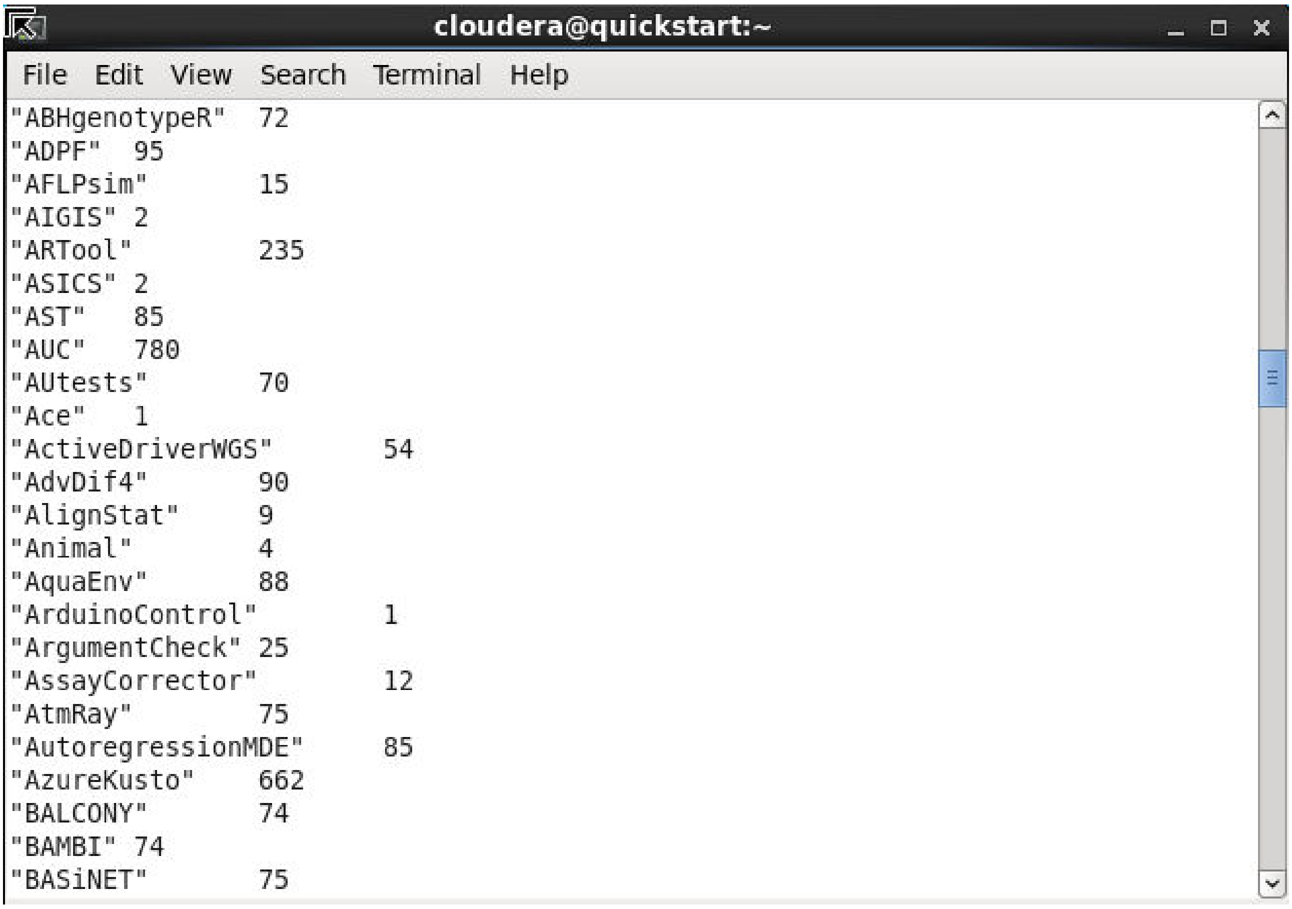


```
Time taken: 0.746 seconds, Fetched: 10 row(s)
hive> select * from datatwo limit 10;
OK
"date"  "time"  NULL    NULL    NULL    "r_os"  "package"       NULL    "country
"       NULL
"2021-06-01"    "00:00:22"      91753   NULL    NULL    NA      "ids"   NULL    "
US"     1
"2021-06-01"    "00:00:22"      81589   NULL    NULL    NA      "markdown"      N
ULL     "US"    2
"2021-06-01"    "00:00:14"      202330  NULL    NULL    NA      "labelled"      N
ULL     "US"    3
"2021-06-01"    "00:00:12"      2013498 NULL    NULL    NA      "isoband"       N
ULL     "US"    3
"2021-06-01"    "00:00:22"      112236  NULL    NULL    NA      "xfun"  NULL    "
US"     4
"2021-06-01"    "00:00:21"      141964  NULL    NULL    NA      "processx"      N
ULL     "US"    1
"2021-06-01"    "00:00:24"      283373  NULL    NULL    NA      "haven" NULL    "
US"     5
"2021-06-01"    "00:00:22"      50083   NULL    NULL    NA      "uuid"  NULL    "
US"     6
"2021-06-01"    "00:00:14"      38099   NULL    NULL    NA      "prettyunits"   N
ULL     "MX"    7
```

c. **Count the number of occurrences of different packages**

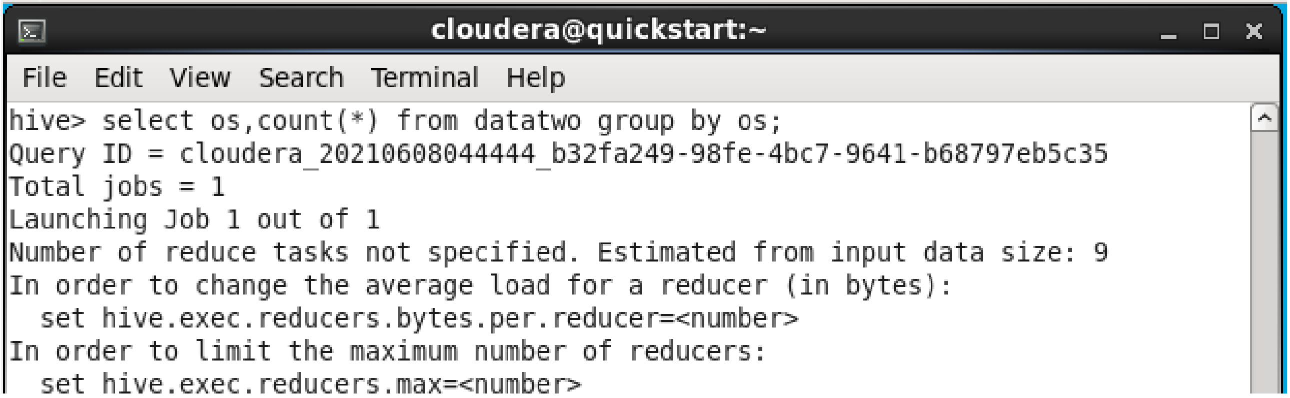


```
hive> select package,count(*) from datatwo group by package limit 50;
Query ID = cloudera_20210608035959_3b35c8fb-1bd4-4b62-8b1b-8c8f5d81a8e4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 9
In order to change the average load for a reducer (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:
```

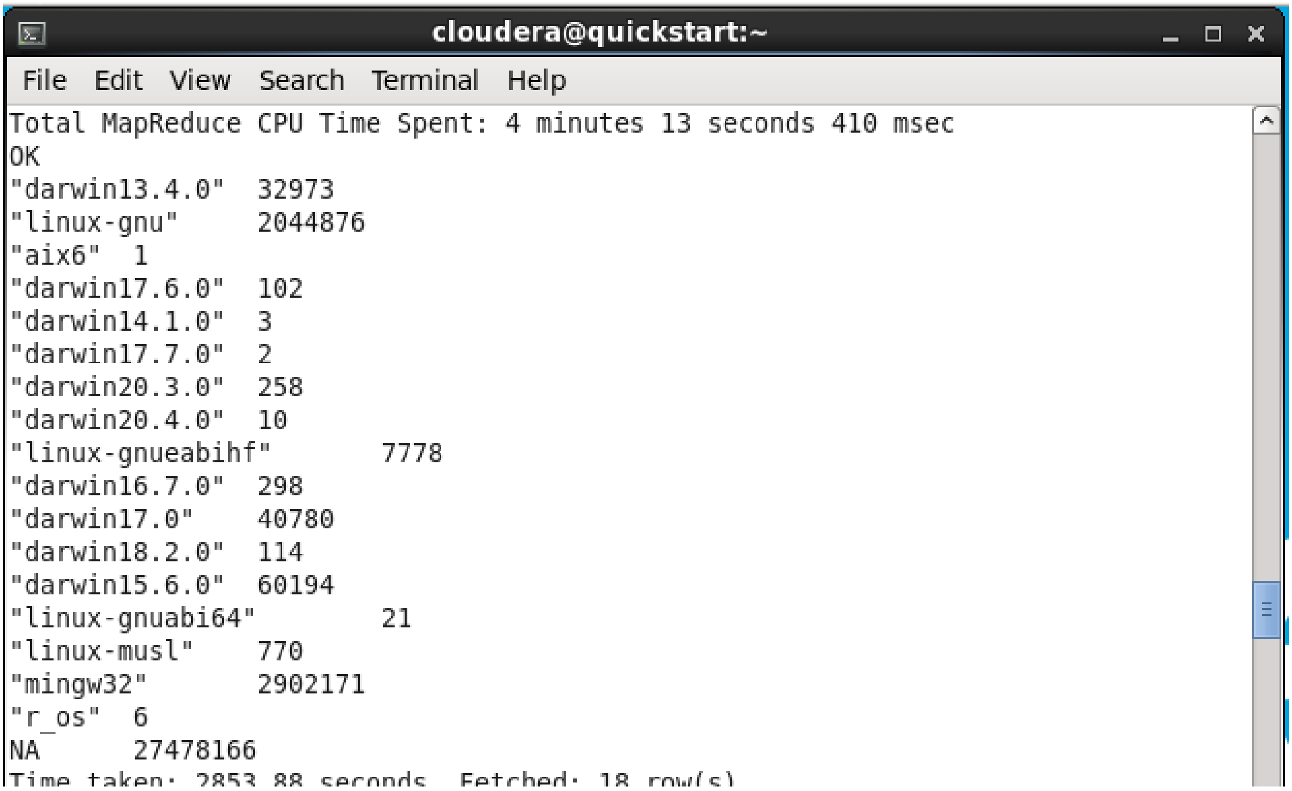


```
"ABHgenotypeR"  72
"ADPF"  95
"AFLPsim"       15
"AIGIS" 2
"ARTool"        235
"ASICS" 2
"AST"   85
"AUC"   780
"AUtests"       70
"Ace"   1
"ActiveDriverWGS"       54
"AdvDif4"       90
"AlignStat"     9
"Animal"        4
"AquaEnv"       88
"ArduinoControl"        1
"ArgumentCheck" 25
"AssayCorrector"        12
"AtmRay"        75
"AutoregressionMDE"     85
"AzureKusto"    662
"BALCONY"       74
"BAMBI" 74
"BASiNET"       75
```

d. **Count the number of occurrences of different packages by operating system;**

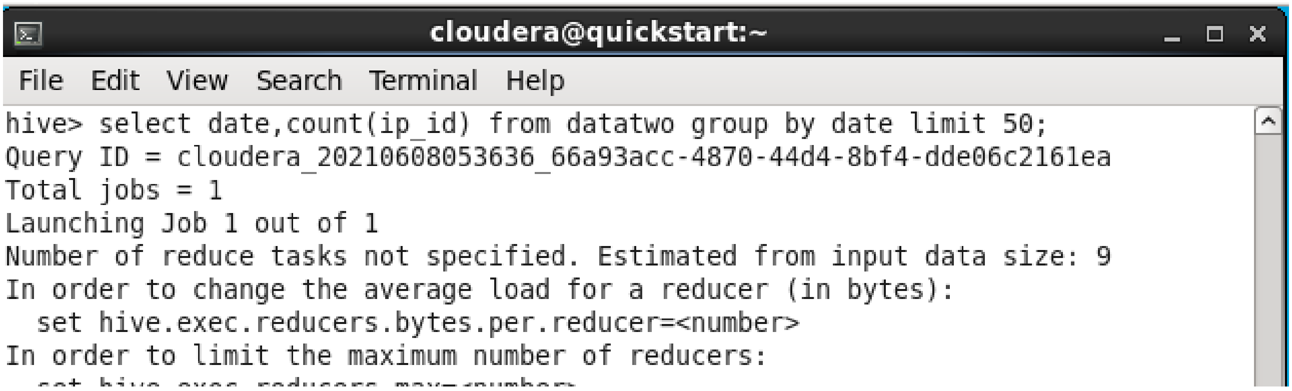


```
hive> select os,count(*) from datatwo group by os;
Query ID = cloudera_20210608044444_b32fa249-98fe-4bc7-9641-b68797eb5c35
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 9
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
```

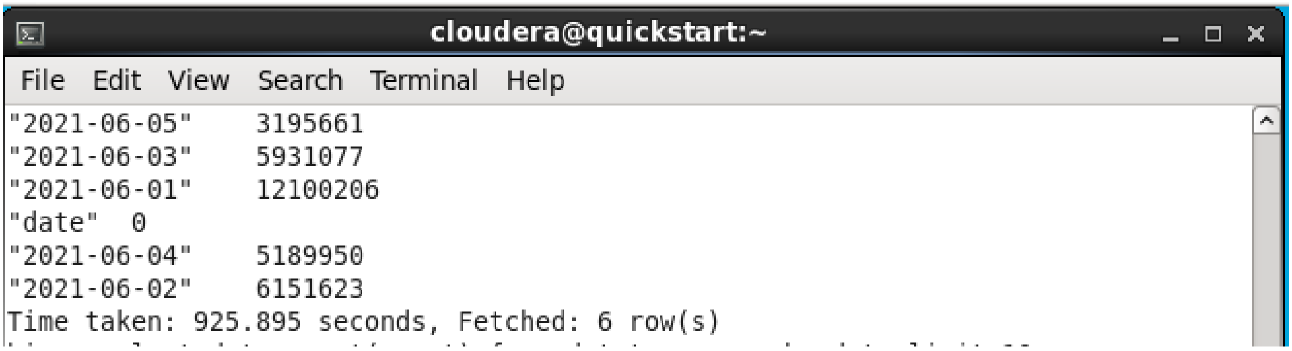


```
Total MapReduce CPU Time Spent: 4 minutes 13 seconds 410 msec
OK
"darwin13.4.0"  32973
"linux-gnu"     2044876
"aix6"  1
"darwin17.6.0"  102
"darwin14.1.0"  3
"darwin17.7.0"  2
"darwin20.3.0"  258
"darwin20.4.0"  10
"linux-gnueabihf"       7778
"darwin16.7.0"  298
"darwin17.0"    40780
"darwin18.2.0"  114
"darwin15.6.0"  60194
"linux-gnuabi64"        21
"linux-musl"    770
"mingw32"       2902171
"r_os"  6
NA      27478166
Time taken: 2853.88 seconds, Fetched: 18 row(s)
```

**Count the number of distinct users each day(with one week data)**



```
hive> select date,count(ip_id) from datatwo group by date limit 50;
Query ID = cloudera_20210608053636_66a93acc-4870-44d4-8bf4-dde06c2161ea
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 9
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>
```



```
"2021-06-05"    3195661
"2021-06-03"    5931077
"2021-06-01"    12100206
"date"  0
"2021-06-04"    5189950
"2021-06-02"    6151623
Time taken: 925.895 seconds, Fetched: 6 row(s)
```