

1. Import RStudio Log Files from one week in February 2019 into HDFS

Step-1 Download one week Logs from <http://cran-logs.rstudio.com/>

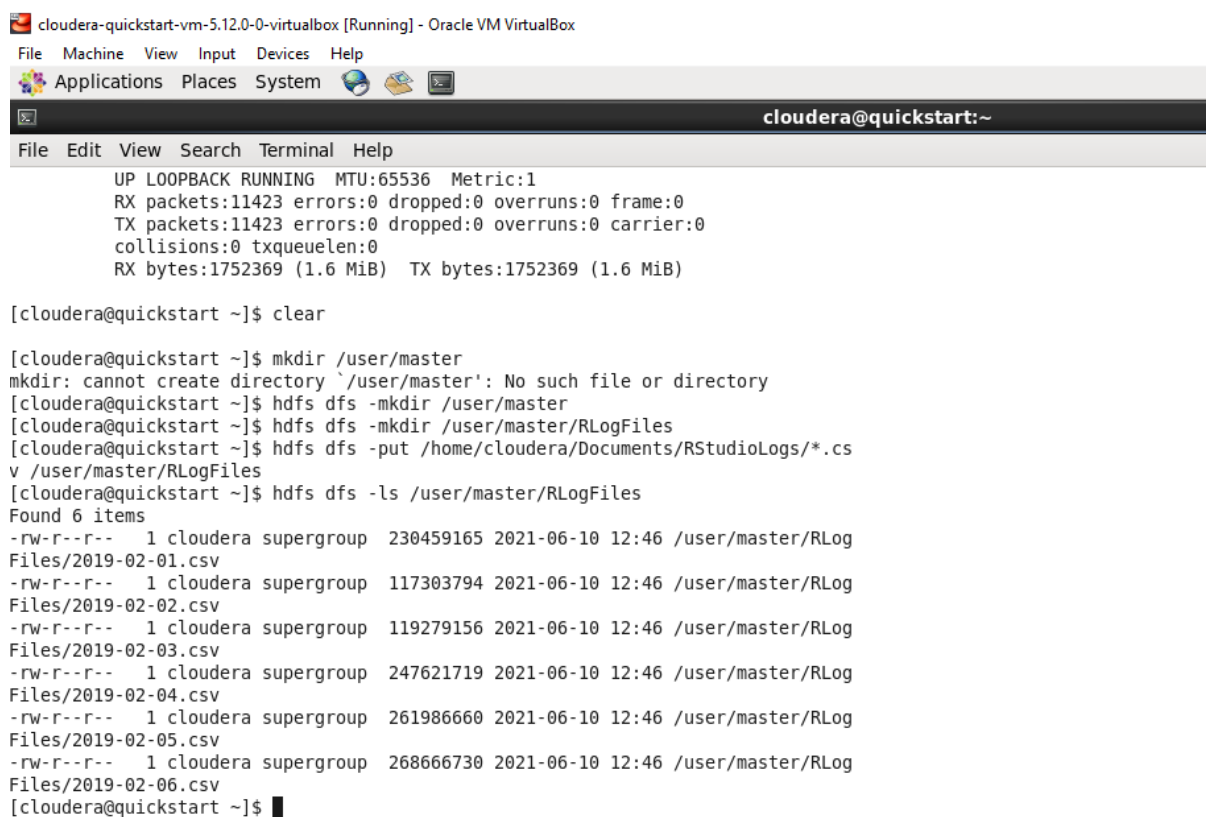
Step-2 Create a new folder to store downloaded logs files.

```
mkdir RStudioLogs
```

Step-3 Unzip all *.zip files

Step-4 Put on HDFS and verify

```
hdfs dfs -put /home/cloudera/Documents/RStudioLogs/*.csv /user/master/RLogFiles
```



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
UP LOOPBACK RUNNING MTU:65536 Metric:1
RX packets:11423 errors:0 dropped:0 overruns:0 frame:0
TX packets:11423 errors:0 dropped:0 overruns:0 carrier:0
collisions:0 txqueuelen:0
RX bytes:1752369 (1.6 MiB) TX bytes:1752369 (1.6 MiB)

[cloudera@quickstart ~]$ clear

[cloudera@quickstart ~]$ mkdir /user/master
mkdir: cannot create directory '/user/master': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Documents/RStudioLogs/*.csv /user/master/RLogFiles
[cloudera@quickstart ~]$ hdfs dfs -ls /user/master/RLogFiles
Found 6 items
-rw-r--r-- 1 cloudera supergroup 230459165 2021-06-10 12:46 /user/master/RLogFiles/2019-02-01.csv
-rw-r--r-- 1 cloudera supergroup 117303794 2021-06-10 12:46 /user/master/RLogFiles/2019-02-02.csv
-rw-r--r-- 1 cloudera supergroup 119279156 2021-06-10 12:46 /user/master/RLogFiles/2019-02-03.csv
-rw-r--r-- 1 cloudera supergroup 247621719 2021-06-10 12:46 /user/master/RLogFiles/2019-02-04.csv
-rw-r--r-- 1 cloudera supergroup 261986660 2021-06-10 12:46 /user/master/RLogFiles/2019-02-05.csv
-rw-r--r-- 1 cloudera supergroup 268666730 2021-06-10 12:46 /user/master/RLogFiles/2019-02-06.csv
[cloudera@quickstart ~]$
```

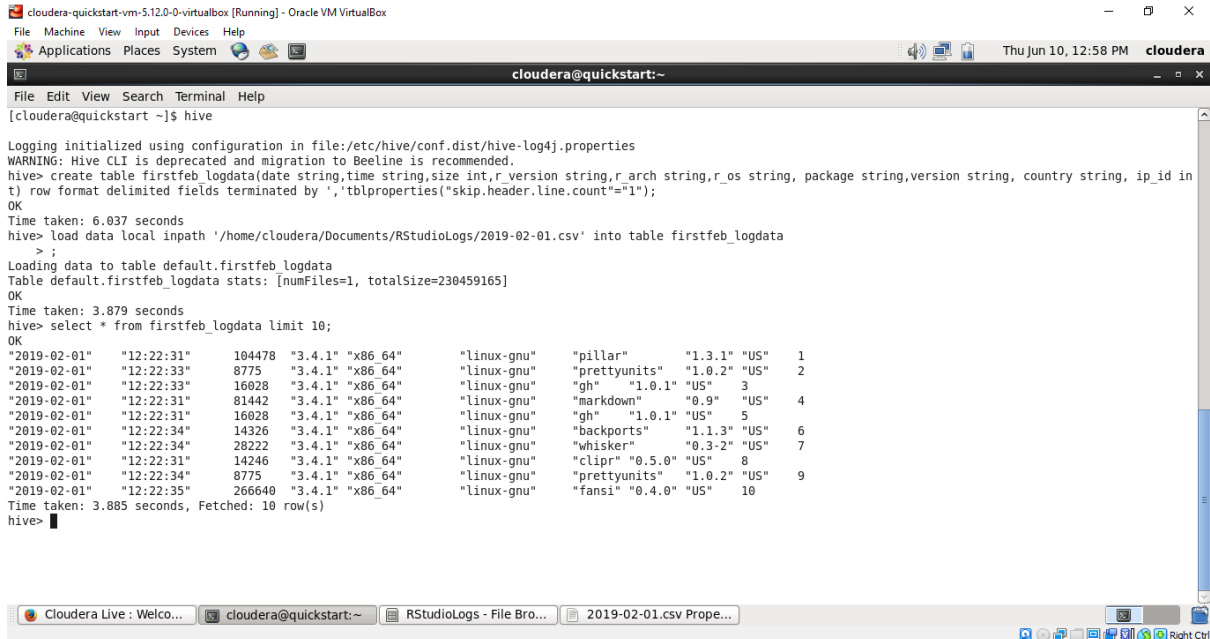
a. Load log-file of one day (e.g., 1st of February 2019)

create table firstfeb_logdata(date string,time string,size int,r_version string,r_arch string,r_os string, package string,version string, country string, ip_id int) row format delimited fields terminated by ',' tblproperties("skip.header.line.count"="1");

LOAD DATA LOCAL INPATH '/home/cloudera/Documents/RStudioLogs/2019-02-01.csv' into table firstfeb_logdata;

- b. b. Dump the first 10 entries on screen (attach a screen shot into your report) to check if it works or not

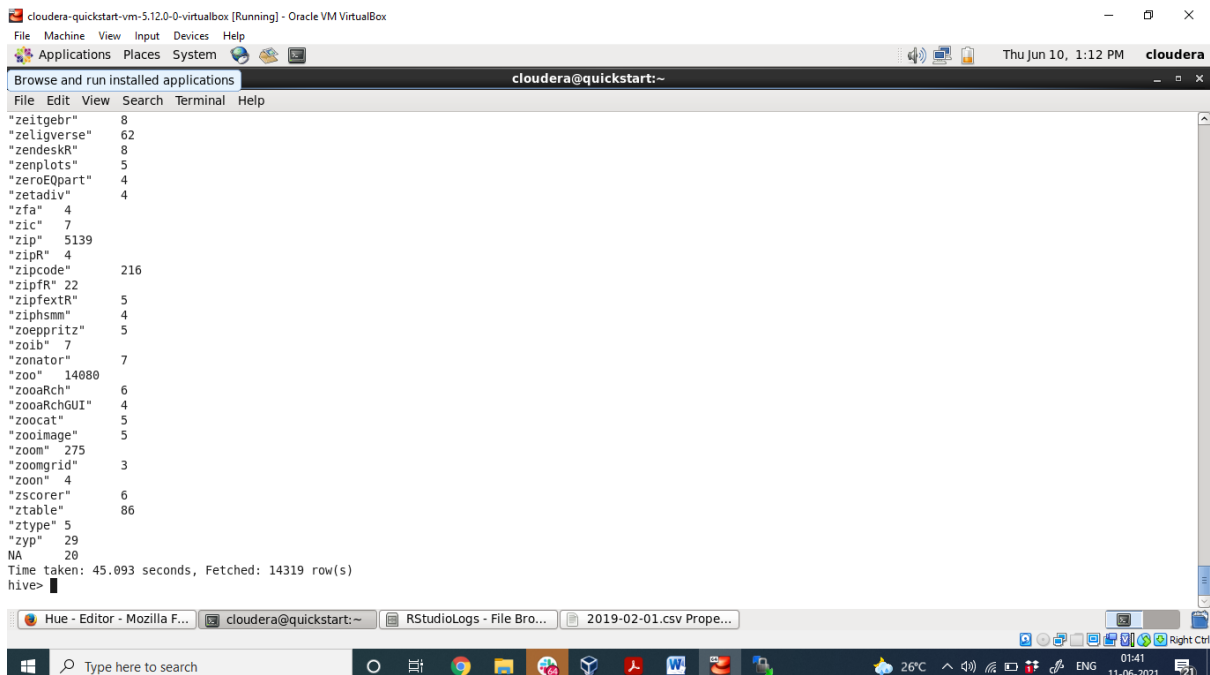
code: select * from firstfeb_logdata limit 10;



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table firstfeb_logdata(date string,time string,size int,r_version string,r_arch string,r_os string, package string,version string, country string, ip_id in
t) row format delimited fields terminated by ','tblproperties("skip.header.line.count"="1");
OK
Time taken: 6.037 seconds
hive> load data local inpath '/home/cloudera/Documents/RStudioLogs/2019-02-01.csv' into table firstfeb_logdata
> ;
Loading data to table default.firstfeb_logdata
Table default.firstfeb_logdata stats: [numFiles=1, totalSize=230459165]
OK
Time taken: 3.879 seconds
hive> select * from firstfeb_logdata limit 10;
OK
"2019-02-01" "12:22:31" 104478 "3.4.1" "x86_64" "linux-gnu" "pillar" "1.3.1" "US" 1
"2019-02-01" "12:22:33" 8775 "3.4.1" "x86_64" "linux-gnu" "prettyunits" "1.0.2" "US" 2
"2019-02-01" "12:22:33" 16028 "3.4.1" "x86_64" "linux-gnu" "gh" "1.0.1" "US" 3
"2019-02-01" "12:22:31" 81442 "3.4.1" "x86_64" "linux-gnu" "markdown" "0.9" "US" 4
"2019-02-01" "12:22:31" 16028 "3.4.1" "x86_64" "linux-gnu" "gh" "1.0.1" "US" 5
"2019-02-01" "12:22:34" 14326 "3.4.1" "x86_64" "linux-gnu" "backports" "1.1.3" "US" 6
"2019-02-01" "12:22:34" 28222 "3.4.1" "x86_64" "linux-gnu" "whisker" "0.3.2" "US" 7
"2019-02-01" "12:22:31" 14246 "3.4.1" "x86_64" "linux-gnu" "clipr" "0.5.0" "US" 8
"2019-02-01" "12:22:34" 8775 "3.4.1" "x86_64" "linux-gnu" "prettyunits" "1.0.2" "US" 9
"2019-02-01" "12:22:35" 266640 "3.4.1" "x86_64" "linux-gnu" "fansi" "0.4.0" "US" 10
Time taken: 3.885 seconds, Fetched: 10 row(s)
hive>
```

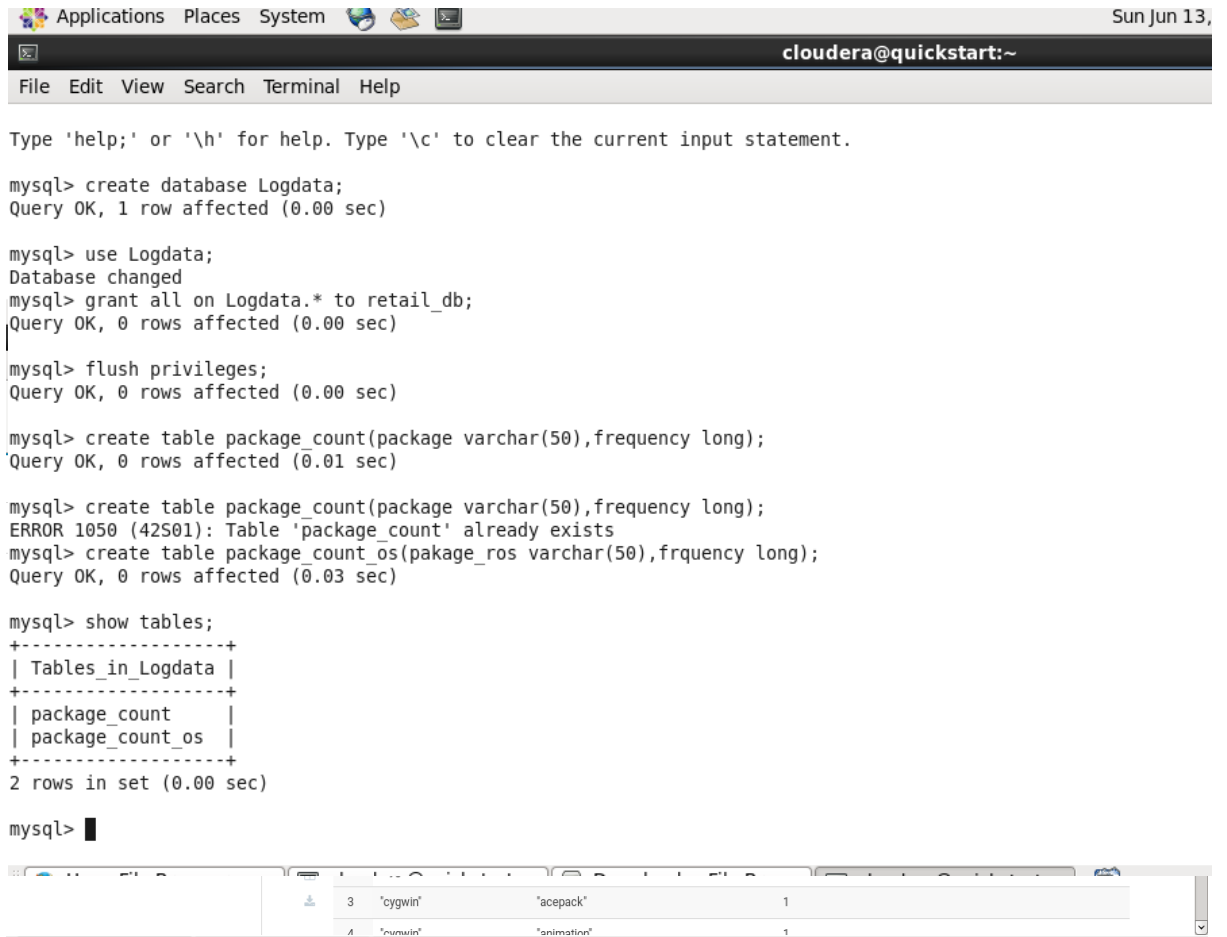
- c. Count the number of occurrences of different packages.

code: select package ,count(*) as no_of_occurences from firstfeb_logdata group by package;



```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
Browse and run installed applications
File Edit View Search Terminal Help
"zeitgebr" 8
"zeligverse" 62
"zendeskR" 8
"zenplots" 5
"zeroEQpart" 4
"zetadiv" 4
"zfa" 4
"zic" 7
"zip" 5139
"zipR" 4
"zipcode" 216
"zipfR" 22
"zipfextR" 5
"ziphsmm" 4
"zoeppritz" 5
"zoib" 7
"zonator" 7
"zoo" 14080
"zooaRch" 6
"zooaRchGUI" 4
"zoocat" 5
"zoimage" 5
"zoom" 275
"zoomgrid" 3
"zoon" 4
"zscorer" 6
"ztable" 86
"ztype" 5
"zyp" 29
NA 20
Time taken: 45.093 seconds, Fetched: 14319 row(s)
hive>
```

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



The screenshot shows a terminal window with the following MySQL commands and output:

```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database Logdata;
Query OK, 1 row affected (0.00 sec)

mysql> use Logdata;
Database changed
mysql> grant all on Logdata.* to retail_db;
Query OK, 0 rows affected (0.00 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.00 sec)

mysql> create table package_count(package varchar(50),frequency long);
Query OK, 0 rows affected (0.01 sec)

mysql> create table package_count(package varchar(50),frequency long);
ERROR 1050 (42S01): Table 'package_count' already exists
mysql> create table package_count_os(package_os varchar(50),frequency long);
Query OK, 0 rows affected (0.03 sec)

mysql> show tables;
+-----+
| Tables_in_Logdata |
+-----+
| package_count      |
| package_count_os   |
+-----+
2 rows in set (0.00 sec)

mysql>
```

Below the terminal window, a file manager window is visible, showing a list of files:

Icon	Name	Type	Size
3	"cygwin"	"acepack"	1
A	"mumaxin"	"animation"	1

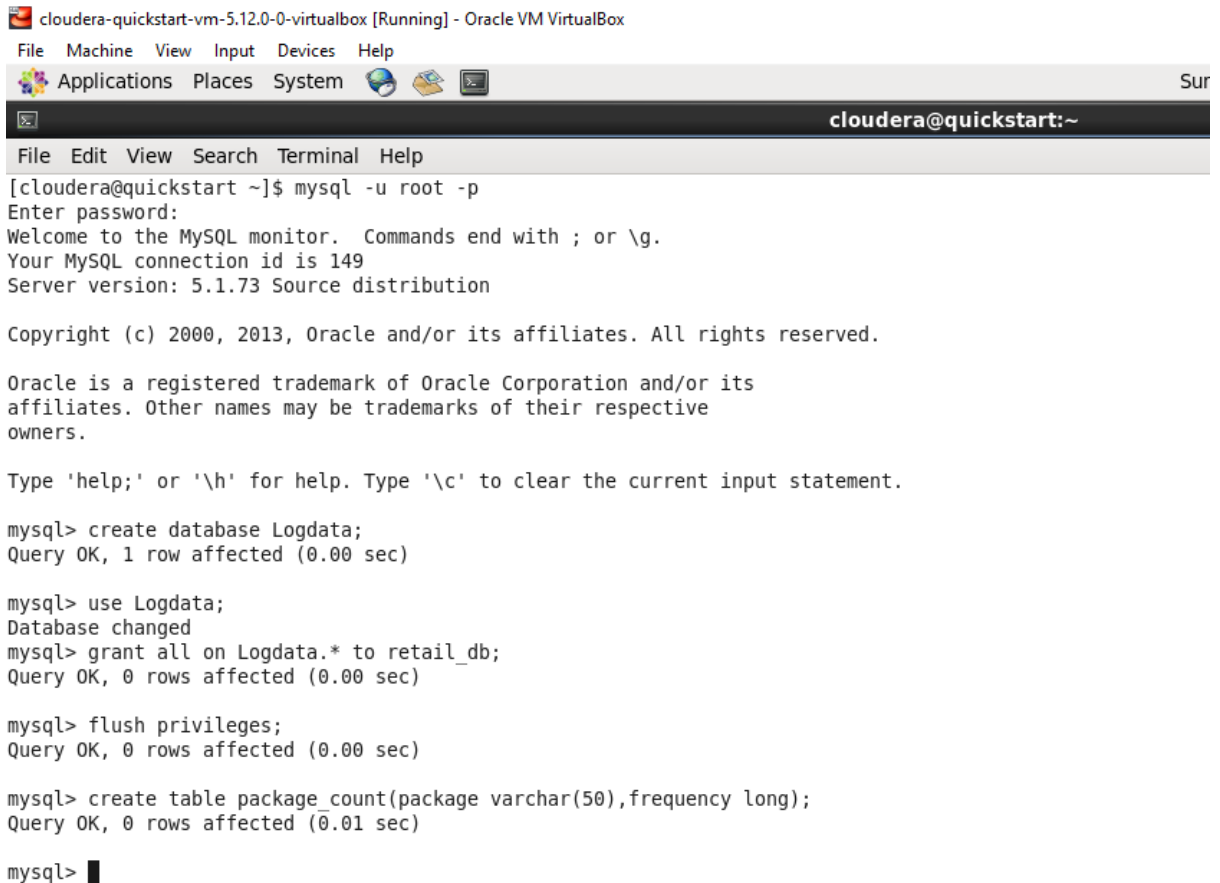
code: `Select concat(package,"",r_os) as package_ros,count(package) as
no_of_occurences from firstfeb_logdata group by r_os,package;`

e. e. Store the results of both operations in HDFS;

code: `hdfs dfs -put /home/cloudera/Downloads/*.csv /user/master/RLogFiles`

3. sqoop, MySQL and R/Python:

**a. Export the results of both operations (package frequencies and package frequencies by
operating systems) via sqoop into MySQL;**



```

cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 149
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database Logdata;
Query OK, 1 row affected (0.00 sec)

mysql> use Logdata;
Database changed
mysql> grant all on Logdata.* to retail_db;
Query OK, 0 rows affected (0.00 sec)

mysql> flush privileges;
Query OK, 0 rows affected (0.00 sec)

mysql> create table package_count(package varchar(50),frequency long);
Query OK, 0 rows affected (0.01 sec)

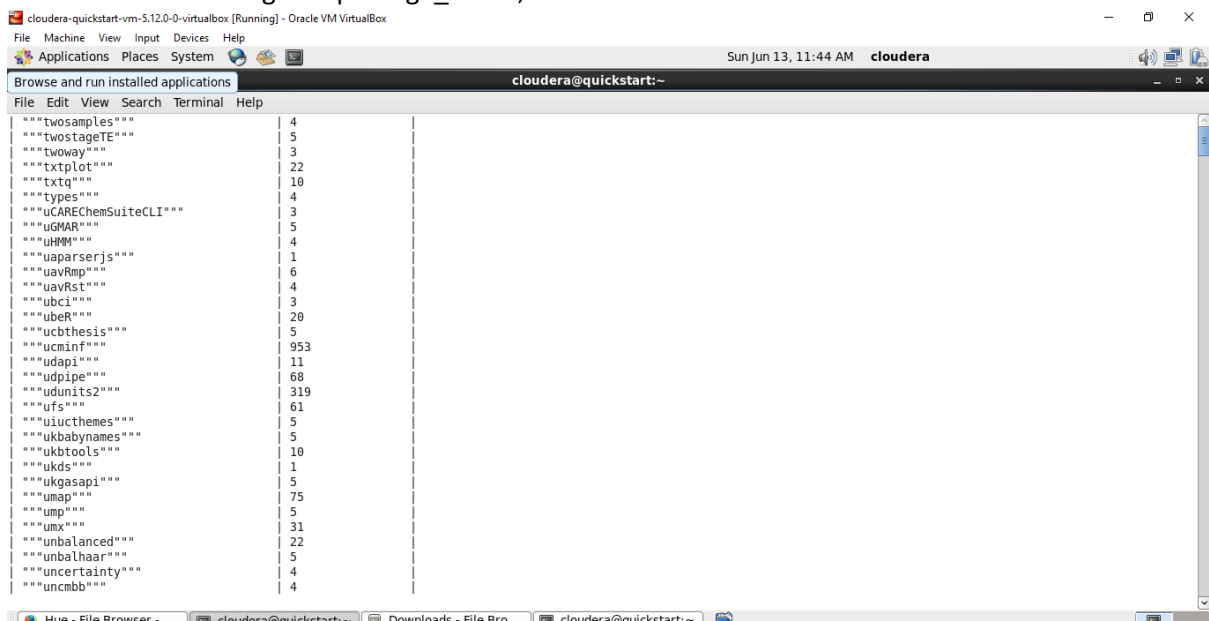
mysql>

```

a. Export the results of both operations (package frequencies and package frequencies by operating systems) via sqoop into MySQL;

code: `sqoop export --connect "jdbc:mysql://quickstart.cloudera:3306/Logdata" --username root --password cloudera --table package_count --export-dir / user/ master/ RLogFiles/ package.csv --input-fields-terminated-by ',' --input-lines-terminated-by '\n' --num-mappers 2 --batch`

`select * from Logdata.package_count;`



```

cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
Browse and run installed applications
File Edit View Search Terminal Help
Sun Jun 13, 11:44 AM cloudera

***twosamples***      4
***twostageTE***      5
***two-way***         3
***txtploT***         22
***txtq***            10
***types***           4
***uCARChemSuiteCLI*** 3
***uGWAR***           5
***uHMM***            4
***uaparserjs***      1
***uavRmp***          6
***uavRst***          4
***ubci***            3
***uberR***           20
***ucbthesis***       5
***ucminf***          953
***udapi***           11
***udpipe***          68
***udunits2***        319
***ufs***             61
***uIucthemes***      5
***ukbabynames***     5
***ukbttools***       10
***ukds***            1
***ukgasapi***        5
***umap***            75
***ump***             5
***umx***             31
***unbalanced***      22
***unbalhaar***       5
***uncertainty***     4
***uncmbb***          4

```