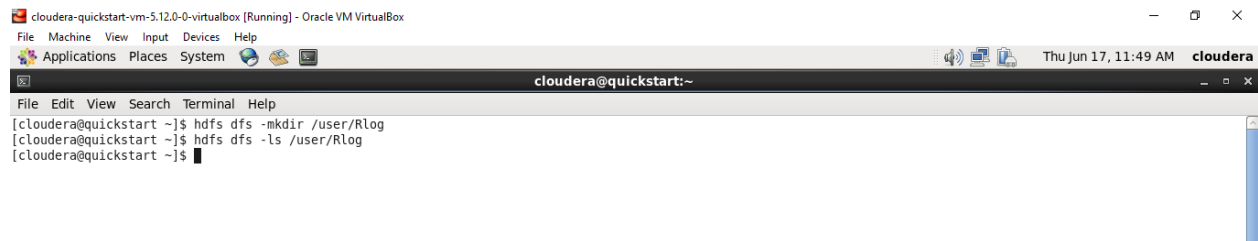


HIVE Task-2

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

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

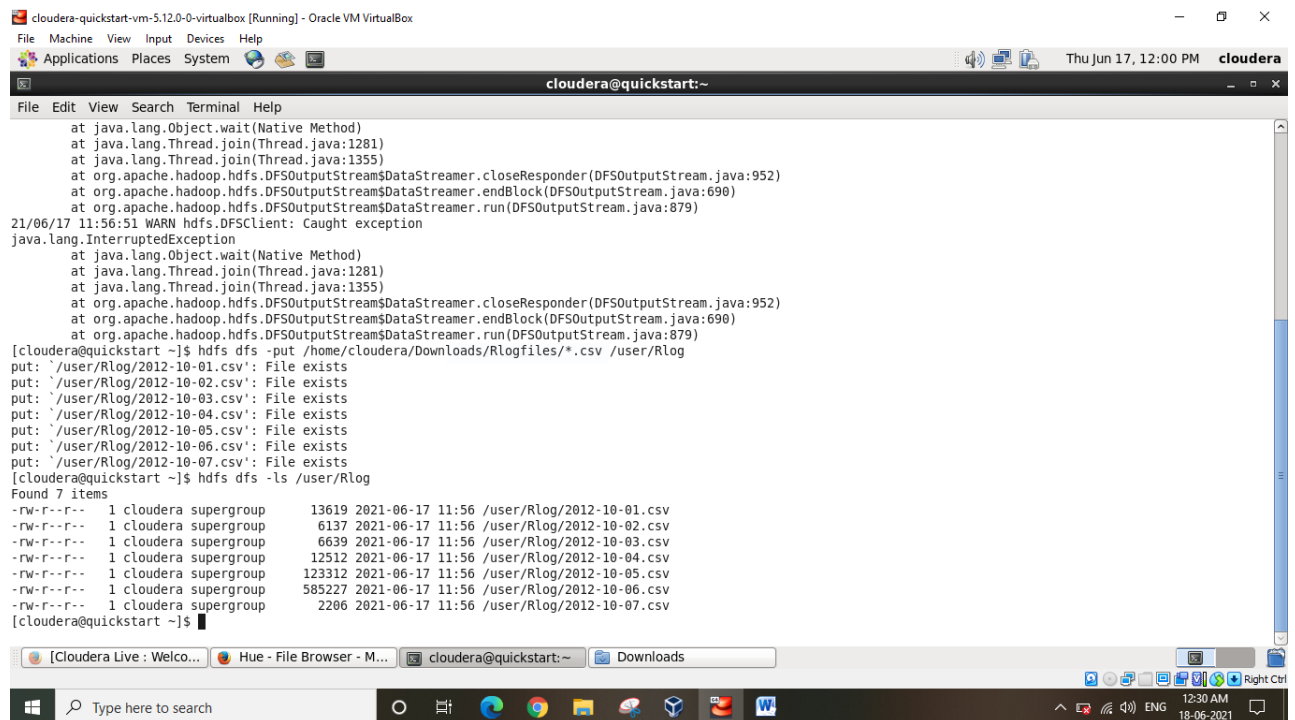
Step-2 Create a new folder to store downloaded logs files



```
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 ~]$ hdfs dfs -mkdir /user/Rlog
[cloudera@quickstart ~]$ hdfs dfs -ls /user/Rlog
[cloudera@quickstart ~]$
```

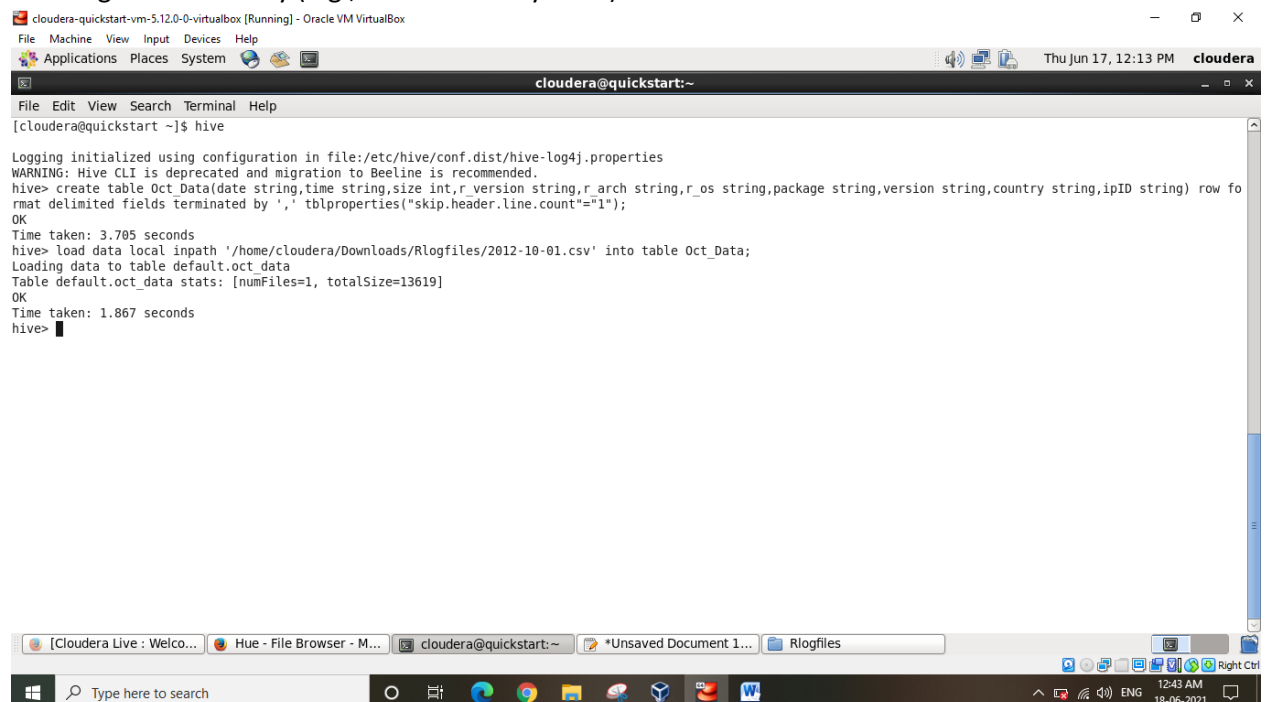
Step-3 Unzip all *.zip files

Step-4 Put on ?HDFS and verify



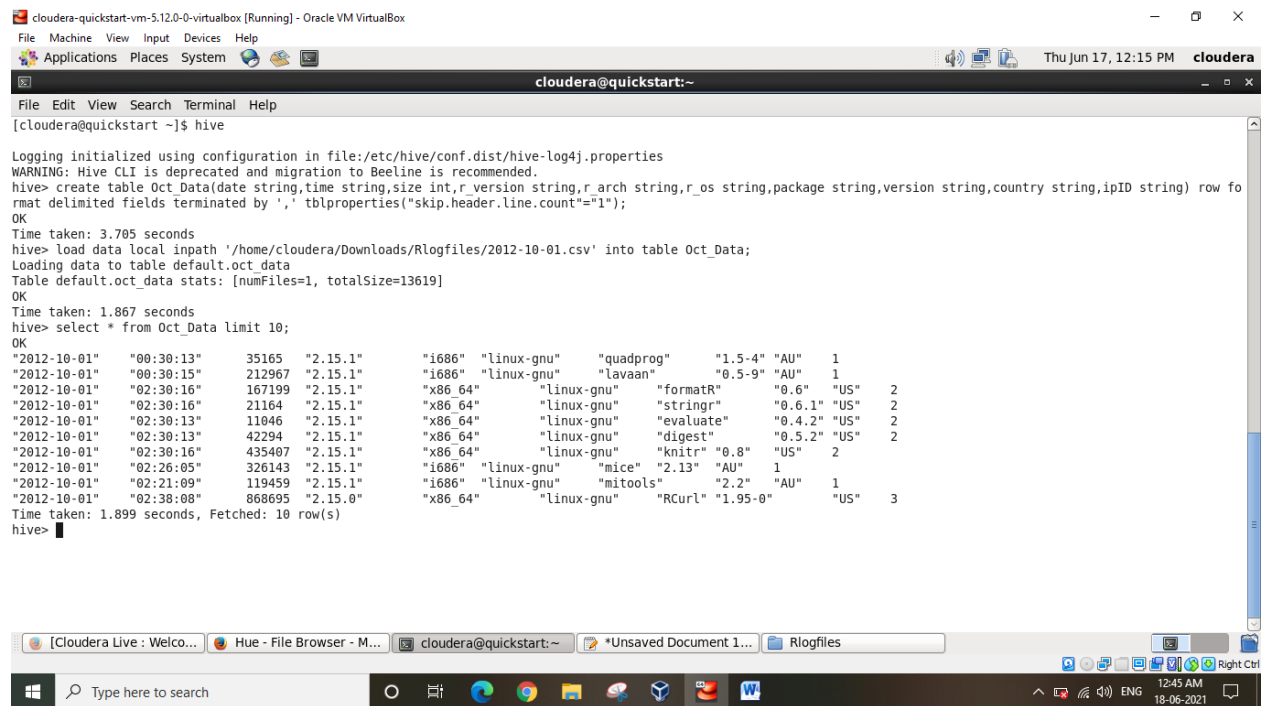
```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
at java.lang.Object.wait(Native Method)
at java.lang.Thread.join(Thread.java:1281)
at java.lang.Thread.join(Thread.java:1355)
at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:952)
at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:690)
at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:879)
21/06/17 11:56:51 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.DFSOutputStreamDataStreamer.closeResponder(DFSOutputStream.java:952)
at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.endBlock(DFSOutputStream.java:690)
at org.apache.hadoop.hdfs.DFSOutputStreamDataStreamer.run(DFSOutputStream.java:879)
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/Downloads/Rlogfiles/*.csv /user/Rlog
put: /user/Rlog/2012-10-01.csv: File exists
put: /user/Rlog/2012-10-02.csv: File exists
put: /user/Rlog/2012-10-03.csv: File exists
put: /user/Rlog/2012-10-04.csv: File exists
put: /user/Rlog/2012-10-05.csv: File exists
put: /user/Rlog/2012-10-06.csv: File exists
put: /user/Rlog/2012-10-07.csv: File exists
[cloudera@quickstart ~]$ hdfs dfs -ls /user/Rlog
Found 7 items
-rw-r--r-- 1 cloudera supergroup 13619 2021-06-17 11:56 /user/Rlog/2012-10-01.csv
-rw-r--r-- 1 cloudera supergroup 6137 2021-06-17 11:56 /user/Rlog/2012-10-02.csv
-rw-r--r-- 1 cloudera supergroup 6639 2021-06-17 11:56 /user/Rlog/2012-10-03.csv
-rw-r--r-- 1 cloudera supergroup 12512 2021-06-17 11:56 /user/Rlog/2012-10-04.csv
-rw-r--r-- 1 cloudera supergroup 123312 2021-06-17 11:56 /user/Rlog/2012-10-05.csv
-rw-r--r-- 1 cloudera supergroup 585227 2021-06-17 11:56 /user/Rlog/2012-10-06.csv
-rw-r--r-- 1 cloudera supergroup 2206 2021-06-17 11:56 /user/Rlog/2012-10-07.csv
[cloudera@quickstart ~]$
```

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



```
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 oct_data(date string,time string,size int,r_version string,r_arch string,r_os string,package string,version string,country string,iPId string) row fo
rmat delimited fields terminated by ',' tblproperties('skip.header.line.count'='1');
OK
Time taken: 3.705 seconds
hive> load data local inpath '/home/cloudera/Downloads/Rlogfiles/2012-10-01.csv' into table oct_data;
Loading data to table default.oct_data
Table default.oct_data stats: [numFiles=1, totalSize=13619]
OK
Time taken: 1.867 seconds
hive>
```

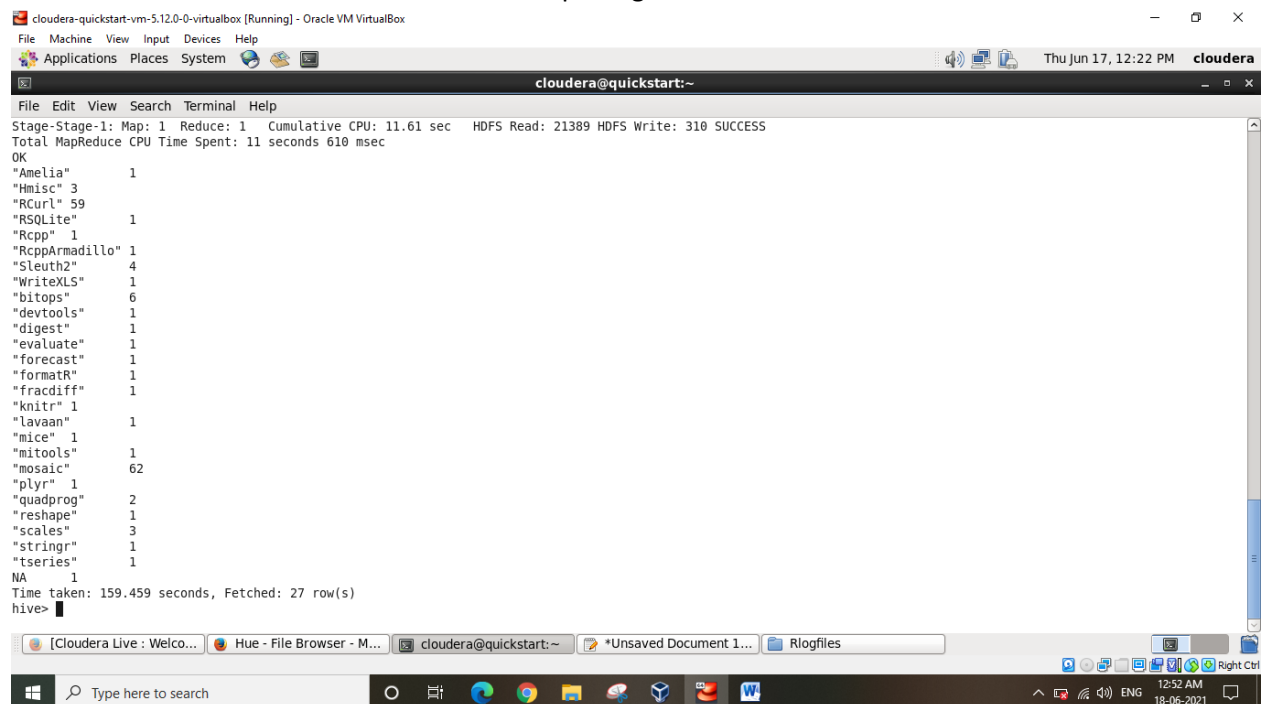
- b. Dump the first 10 entries on screen (attach a screen shot into your report) to check if it works or not



```
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 Oct_Data(date string,time string,size int,r_version string,r_arch string,r_os string,package string,version string,country string,ipID string) row fo
rmat delimited fields terminated by ',' tblproperties("skip.header.line.count"="1");
OK
Time taken: 3.705 seconds
hive> load data local inpath '/home/cloudera/Downloads/Rlogfiles/2012-10-01.csv' into table Oct_Data;
Loading data to table default.oct_data
Table default.oct_data stats: [numFiles=1, totalSize=13619]
OK
Time taken: 1.867 seconds
hive> select * from Oct_Data limit 10;
OK
"2012-10-01" "00:30:13" 35165 "2.15.1" "i686" "linux-gnu" "quadprog" "1.5-4" "AU" 1
"2012-10-01" "00:30:15" 212967 "2.15.1" "i686" "linux-gnu" "lavaan" "0.5-9" "AU" 1
"2012-10-01" "02:30:16" 167199 "2.15.1" "x86_64" "linux-gnu" "formatR" "0.6" "US" 2
"2012-10-01" "02:30:16" 21164 "2.15.1" "x86_64" "linux-gnu" "stringr" "0.6.1" "US" 2
"2012-10-01" "02:30:13" 11046 "2.15.1" "x86_64" "linux-gnu" "evaluate" "0.4.2" "US" 2
"2012-10-01" "02:30:13" 42294 "2.15.1" "x86_64" "linux-gnu" "digest" "0.5.2" "US" 2
"2012-10-01" "02:30:16" 435407 "2.15.1" "x86_64" "linux-gnu" "knitr" "0.8" "US" 2
"2012-10-01" "02:26:05" 326143 "2.15.1" "i686" "linux-gnu" "mice" "2.13" "AU" 1
"2012-10-01" "02:21:09" 119459 "2.15.1" "i686" "linux-gnu" "mitools" "2.2" "AU" 1
"2012-10-01" "02:38:08" 868695 "2.15.0" "x86_64" "linux-gnu" "RCurl" "1.95-0" "US" 3
Time taken: 1.899 seconds, Fetched: 10 row(s)
hive>
```

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



```
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
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.61 sec HDFS Read: 21389 HDFS Write: 310 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 610 msec
OK
"Amelia" 1
"Hmisc" 3
"RCurl" 59
"RSQLite" 1
"Rcpp" 1
"RcppArmadillo" 1
"Sleuth2" 4
"WriteXLS" 1
"bitops" 6
"devtools" 1
"digest" 1
"evaluate" 1
"forecast" 1
"formatR" 1
"fracdiff" 1
"knitr" 1
"lavaan" 1
"mice" 1
"mitools" 1
"mosaic" 62
"plyr" 1
"quadprog" 2
"reshape" 1
"scales" 3
"stringr" 1
"tseries" 1
NA 1
Time taken: 159.459 seconds, Fetched: 27 row(s)
hive>
```

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

cloudera-quickstart-vm-5.12.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Sat Jun 19, 10:20 AM cloudera

cloudera@quickstart:~

File Edit View Search Terminal Help

```
country          string
ipid             string
Time taken: 0.58 seconds, Fetched: 10 row(s)
hive> select r_os, count(*) from oct_data group by r_os;
Query ID = cloudera_20210619101717_a44cfd39-6e33-4ab6-82e3-eb016cfc8188
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1624113472164_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1624113472164_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1624113472164_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-06-19 10:19:02,507 Stage-1 map = 0%, reduce = 0%
2021-06-19 10:19:56,950 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.87 sec
2021-06-19 10:20:37,629 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.86 sec
MapReduce Total cumulative CPU time: 11 seconds 860 msec
Ended Job = job_1624113472164_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.86 sec HDFS Read: 21384 HDFS Write: 49 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 860 msec
OK
"darwin9.8.0" 6
"linux-gnu" 149
"mingw32" 2
NA 2
Time taken: 169.233 seconds, Fetched: 4 row(s)
hive>
```

Hue - File Browser - M... cloudera@quickstart:~

Type here to search

10:50 PM 19-06-2021