Task 1: HADOOP, HDFS, PIG and HIVE

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 Verify files in RStudioLogs folder

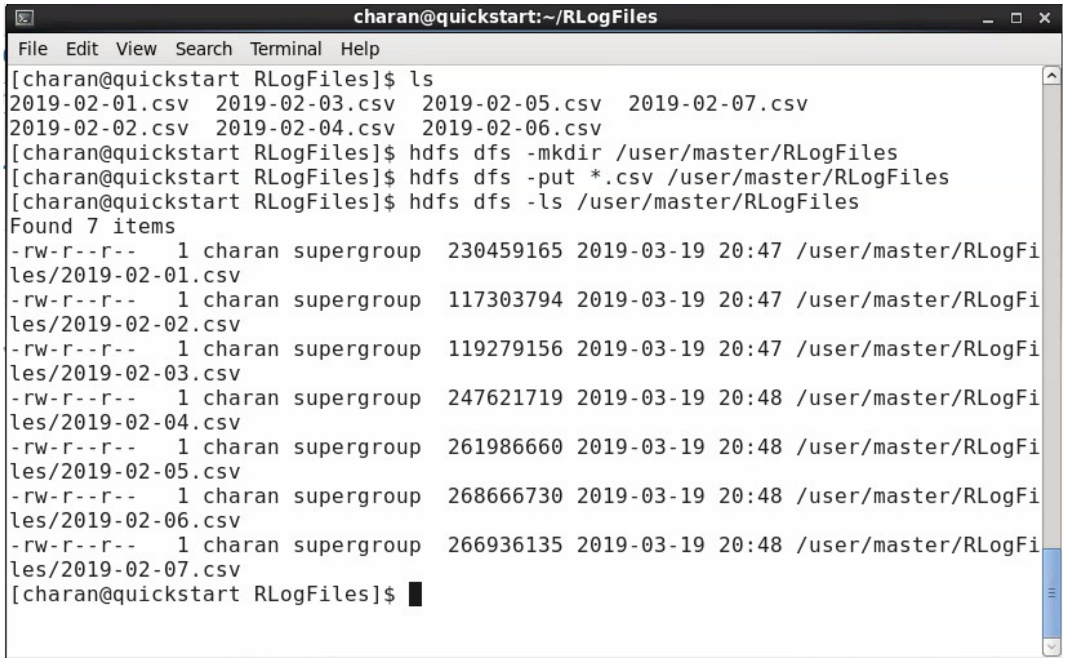
```
ls
```

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
```



charan@quickstart:~/RLogFiles
File Edit View Search Terminal Help

```
[charan@quickstart RLogFiles]$ ls
2019-02-01.csv  2019-02-03.csv  2019-02-05.csv  2019-02-07.csv
2019-02-02.csv  2019-02-04.csv  2019-02-06.csv
[charan@quickstart RLogFiles]$ hdfs dfs -mkdir /user/master/RLogFiles
[charan@quickstart RLogFiles]$ hdfs dfs -put *.csv /user/master/RLogFiles
[charan@quickstart RLogFiles]$ hdfs dfs -ls /user/master/RLogFiles
Found 7 items
-rw-r--r--   1 charan supergroup  230459165 2019-03-19 20:47 /user/master/RLogFi
les/2019-02-01.csv
-rw-r--r--   1 charan supergroup  117303794 2019-03-19 20:47 /user/master/RLogFi
les/2019-02-02.csv
-rw-r--r--   1 charan supergroup  119279156 2019-03-19 20:47 /user/master/RLogFi
les/2019-02-03.csv
-rw-r--r--   1 charan supergroup  247621719 2019-03-19 20:48 /user/master/RLogFi
les/2019-02-04.csv
-rw-r--r--   1 charan supergroup  261986660 2019-03-19 20:48 /user/master/RLogFi
les/2019-02-05.csv
-rw-r--r--   1 charan supergroup  268666730 2019-03-19 20:48 /user/master/RLogFi
les/2019-02-06.csv
-rw-r--r--   1 charan supergroup  266936135 2019-03-19 20:48 /user/master/RLogFi
les/2019-02-07.csv
[charan@quickstart RLogFiles]$ █
```

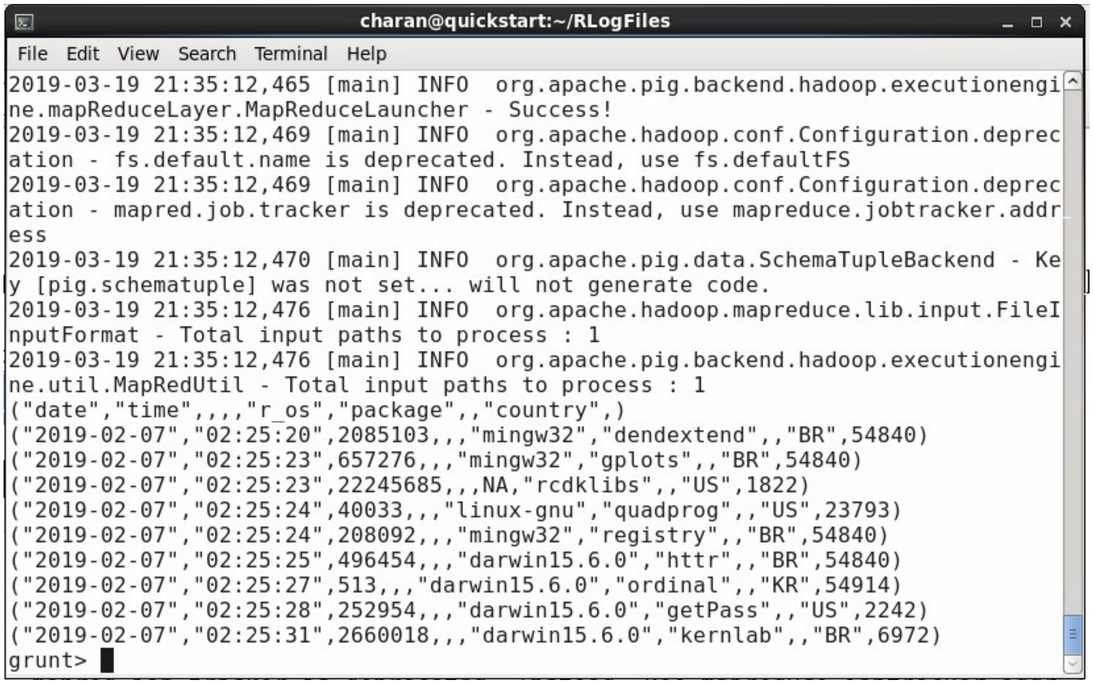
2. Pig Latin: Top-100-packages (by operating system)

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

```
logfile = LOAD '/user/master/RLogFiles/2019-02-07.csv' USING PigStorage(',') AS
(date:chararray,time : chararray,size: int,r_version: float,r_arch: int,r_os: chararray, package : chararray,
version : float, country : chararray, ip_id : int);
```

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

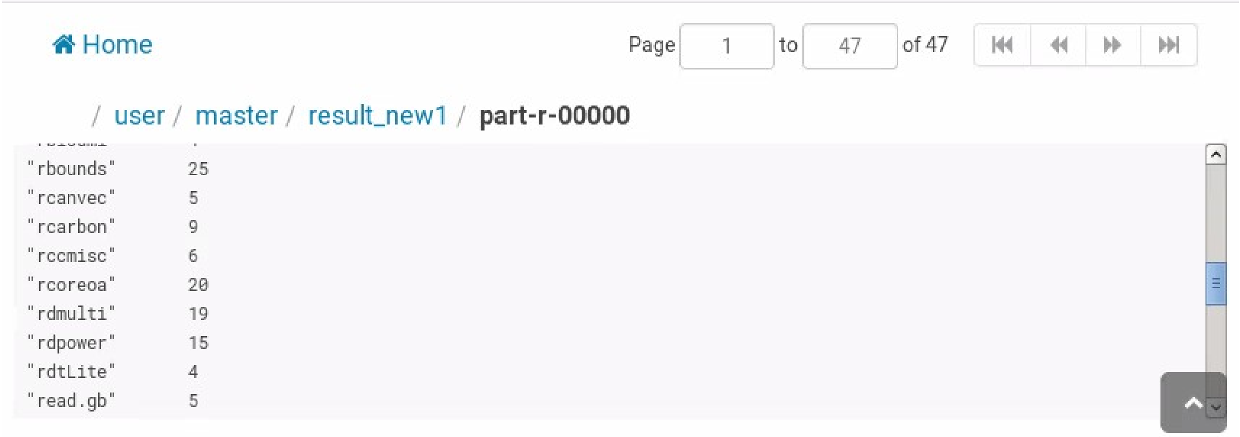
grunt> FIRST_10 = LIMIT logfile 10;
grunt> DUMP FIRST_10;


charan@quickstart:~/RLogFiles
File Edit View Search Terminal Help

```
2019-03-19 21:35:12,465 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-03-19 21:35:12,469 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-19 21:35:12,469 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2019-03-19 21:35:12,470 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2019-03-19 21:35:12,476 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2019-03-19 21:35:12,476 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
("date","time",,,,"r_os","package",,"country",)
("2019-02-07","02:25:20",2085103,,,"mingw32","dendextend",,"BR",54840)
("2019-02-07","02:25:23",657276,,,"mingw32","gplots",,"BR",54840)
("2019-02-07","02:25:23",22245685,,,NA,"rcdklibs",,"US",1822)
("2019-02-07","02:25:24",40033,,,"linux-gnu","quadprog",,"US",23793)
("2019-02-07","02:25:24",208092,,,"mingw32","registry",,"BR",54840)
("2019-02-07","02:25:25",496454,,,"darwin15.6.0","httr",,"BR",54840)
("2019-02-07","02:25:27",513,,,"darwin15.6.0","ordinal",,"KR",54914)
("2019-02-07","02:25:28",252954,,,"darwin15.6.0","getPass",,"US",2242)
("2019-02-07","02:25:31",2660018,,,"darwin15.6.0","kernlab",,"BR",6972)
grunt> █
```

c. Count the number of occurrences of different packages;

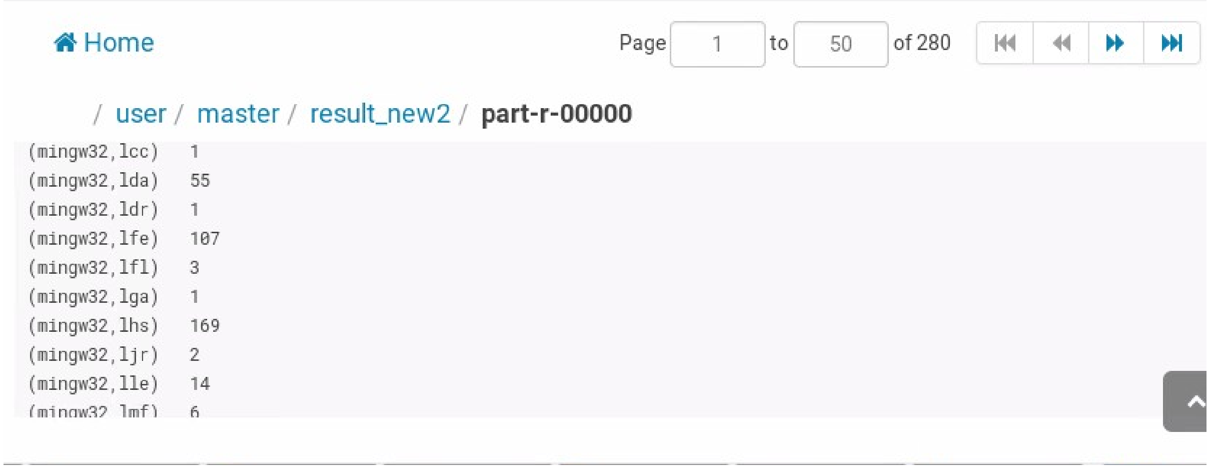
pack_R = FOREACH logfile GENERATE $6 as pack_R;
grouped = GROUP pack_R BY $0;
count = FOREACH grouped GENERATE group, COUNT(pack_R);
STORE count INTO '/user/master/result_new1';



```
"rbounds"     25
"rcanvec"     5
"rcarbon"     9
"rccmisc"     6
"rcoreoa"     20
"rdmulti"     19
"rdpower"     15
"rdtLite"     4
"read.gb"     5
```

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

```
FEB07= LOAD '/user/master/RLogFiles/2019-02-07.csv' USING PigStorage(',') AS (date:chararray,
time:chararray, size:int, rversion:chararray, arch:chararray, os:chararray, pkg:chararray,
version:chararray, country:chararray, ipid:int);

pack_R_OS = FOREACH FEB07 GENERATE REPLACE(os, '[{(")}]', "), REPLACE(pkg,'[{(")}]', ");

grouped_os = GROUP pack_R_OS BY ($0,$1);
count = FOREACH grouped_os GENERATE group, COUNT(pack_R_OS);
STORE count INTO '/user/master/result_new2';
```



```
(mingw32,lcc)   1
(mingw32,lda)   55
(mingw32,ldr)   1
(mingw32,lfe)   107
(mingw32,lfl)   3
(mingw32,lga)   1
(mingw32,lhs)   169
(mingw32,ljr)   2
(mingw32,lle)   14
(mingw32,lmf)   6
```

e. Store the results of both operations in HDFS;

```
a = LOAD '/user/master/result_new2/part-r-00000' using PigStorage('\t') AS
(os_pkg:chararray,count:chararray);
b = FOREACH a GENERATE REPLACE(os_pkg, '[{(")}]', "), REPLACE(count,'[{(")}]', ");
STORE b INTO '/user/master/result_new3';
```

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;

Step-1 Create new database in mysql or use existing one, we are using 'retail_db' with required permissions. Create tables 'packeage_count' and 'package_count_os'.

```
use retail_db;
create table package_count(package_r varchar(50), counting long);
create table package_count_os(package_r varchar(50), counting long);
```

Step-2 Use sqoop for exporting data from HDFS to MySql tables package_count and package_count_os respectivly.

```
sqoop export --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username root --password
cloudera --table package_count --export-dir /user/master/result_new1/part-r-00000 --input-fields-
terminated-by '\t' --input-lines-terminated-by '\n' --num-mappers 2 --batch --outdir java_files
```

```
sqoop export --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username root --password
cloudera --table package_count_os --export-dir /user/master/result_new3/part-m-00000 --input-fields-
terminated-by '\t' --input-lines-terminated-by '\n' --num-mappers 2 --batch --outdir java_files
```

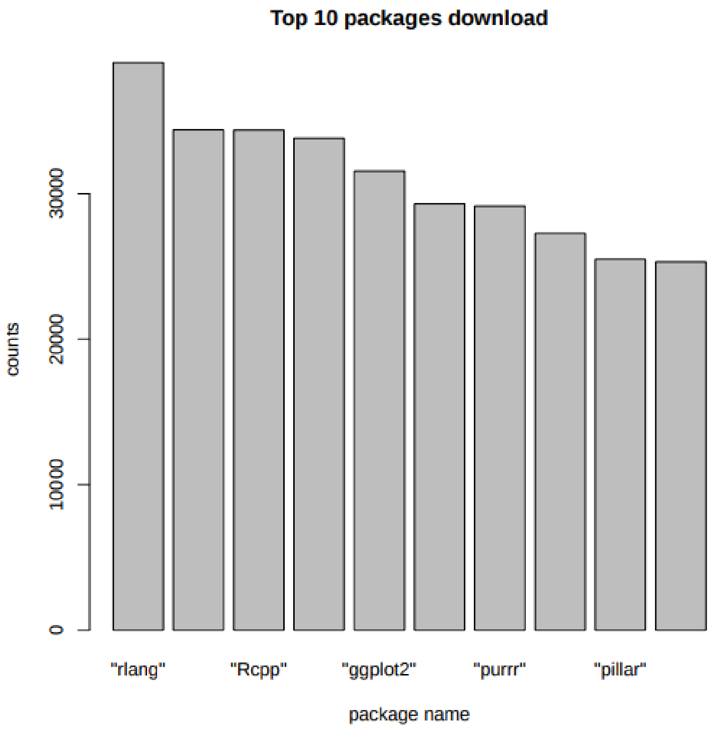
b. Access the tables by R/RStudio or Python and display the results (Top-10-results in bar charts)

Step-1 Install required R packages.

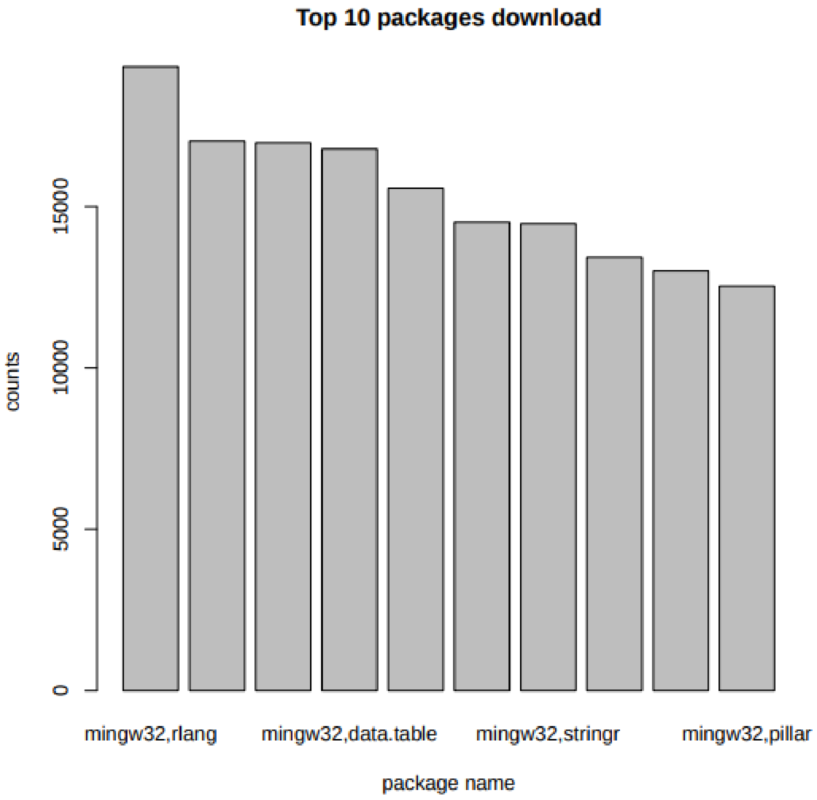
```
install.packages("https://cloud.r-project.org/src/contrib/DBI_1.0.0.tar.gz")
install.packages("http://cloud.r-project.org/src/contrib/RMySQL_0.10.16.tar.gz")
install.packages("http://cloud.r-project.org/src/contrib/Rcpp_1.0.0.tar.gz")
install.packages("http://cloud.r-project.org/src/contrib/plyr_1.8.4.tar.gz")
```

Step-2 Create Script

```
library("RMySQL")
library("plyr")
drv <- dbDriver("MySQL")
con <- dbConnect(drv, user='root', password='cloudera',
host='quickstart.cloudera',dbname="retail_db", port=3306)
dbListConnections(MySQL())
res <- dbGetQuery(con, "SELECT package_r, counting from package_count order by CAST(counting
AS DECIMAL) desc limit 10")
t1 <- unlist(res[1], use.names = FALSE)
c2 <- res[2]
c3 <- as.numeric(unlist(c2))
barplot(c3, names.arg = t1, xlab = "package name", ylab = "counts",
main = "Top 10 packages download")
```



```
library("RMySQL")
library("plyr")
drv <- dbDriver("MySQL")
con <- dbConnect(drv, user='root', password='cloudera',
host='quickstart.cloudera',dbname="retail_db", port=3306)
dbListConnections(MySQL())
res <- dbGetQuery(con, "SELECT package_r, counting from package_count_os order by
CAST(counting AS DECIMAL) desc limit 10")
t1 <- unlist(res[1], use.names = FALSE)
c2 <- res[2]
c3 <- as.numeric(unlist(c2))

barplot(c3, names.arg = t1, xlab = "package name", ylab = "counts",
main = "Top 10 packages download")
```



4. Pig Latin and HIVE: Number of individual users each day

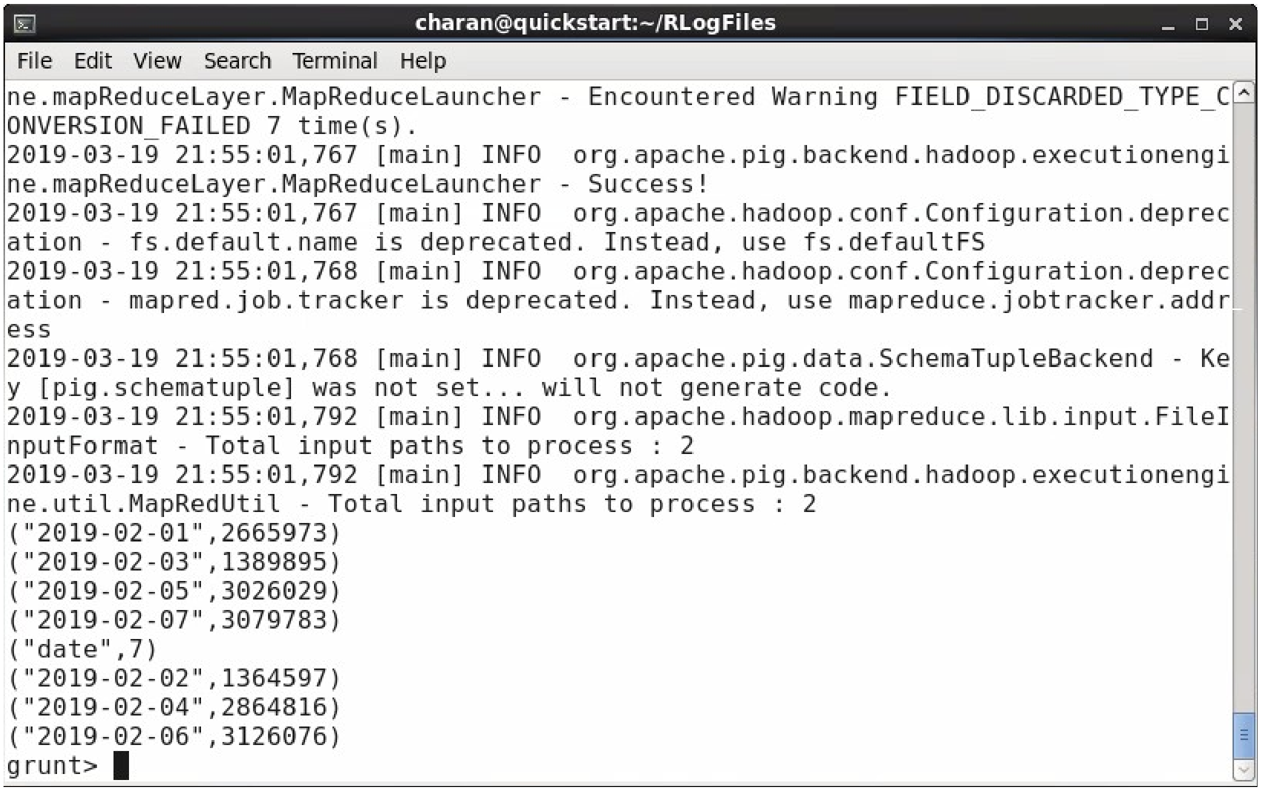
a. Load the log-files into HDFS

As per 1 solution, data is already present into HDFS.

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

```
week_logfile = LOAD '/user/master/RLogFiles/*.csv' USING PigStorage(',') AS (date:chararray,time :
chararray,size: int,r_version: float,r_arch: int,r_os: chararray, package : chararray, version : float,
country : chararray, ip_id : int);
```

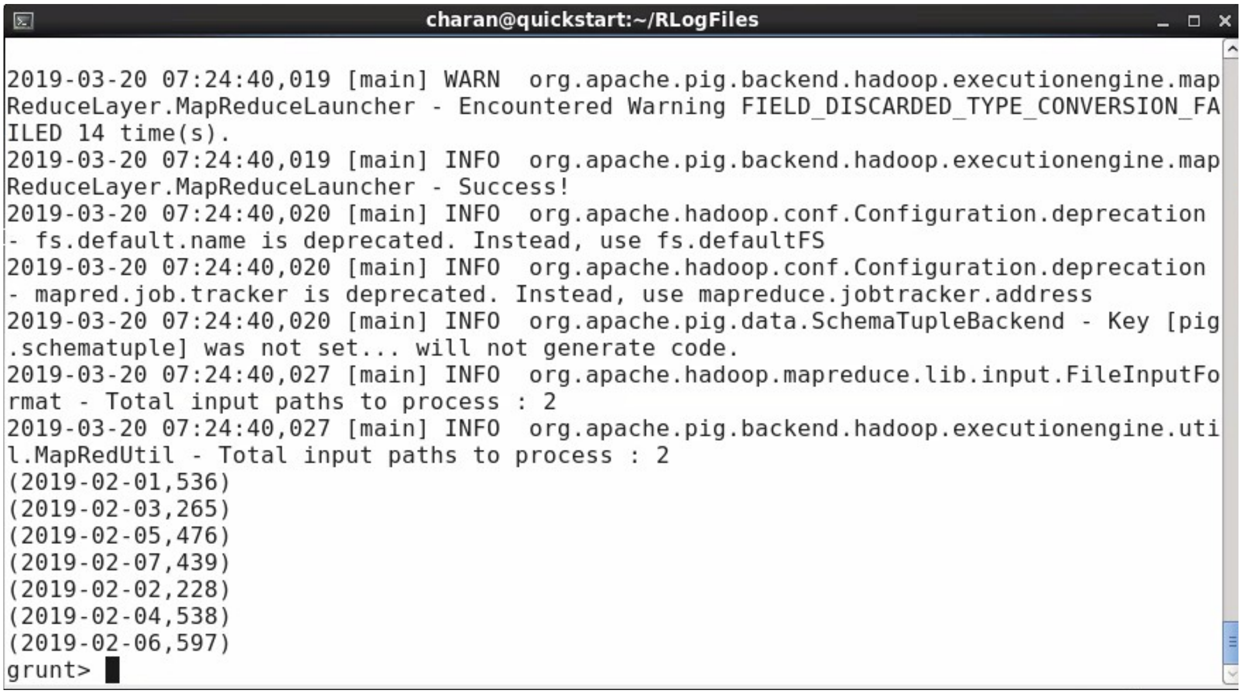
user_ip = FOREACH week_logfile GENERATE $0,$9 as user_ip;
grouped = GROUP user_ip BY $0;
count = FOREACH grouped GENERATE group, COUNT(user_ip);
STORE count INTO '/user/master/result_new6';


charan@quickstart:~/RLogFiles
File Edit View Search Terminal Help

```
ne.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_C
ONVERSION_FAILED 7 time(s).
2019-03-19 21:55:01,767 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2019-03-19 21:55:01,767 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-19 21:55:01,768 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2019-03-19 21:55:01,768 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2019-03-19 21:55:01,792 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 2
2019-03-19 21:55:01,792 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 2
("2019-02-01",2665973)
("2019-02-03",1389895)
("2019-02-05",3026029)
("2019-02-07",3079783)
("date",7)
("2019-02-02",1364597)
("2019-02-04",2864816)
("2019-02-06",3126076)
grunt> 
```

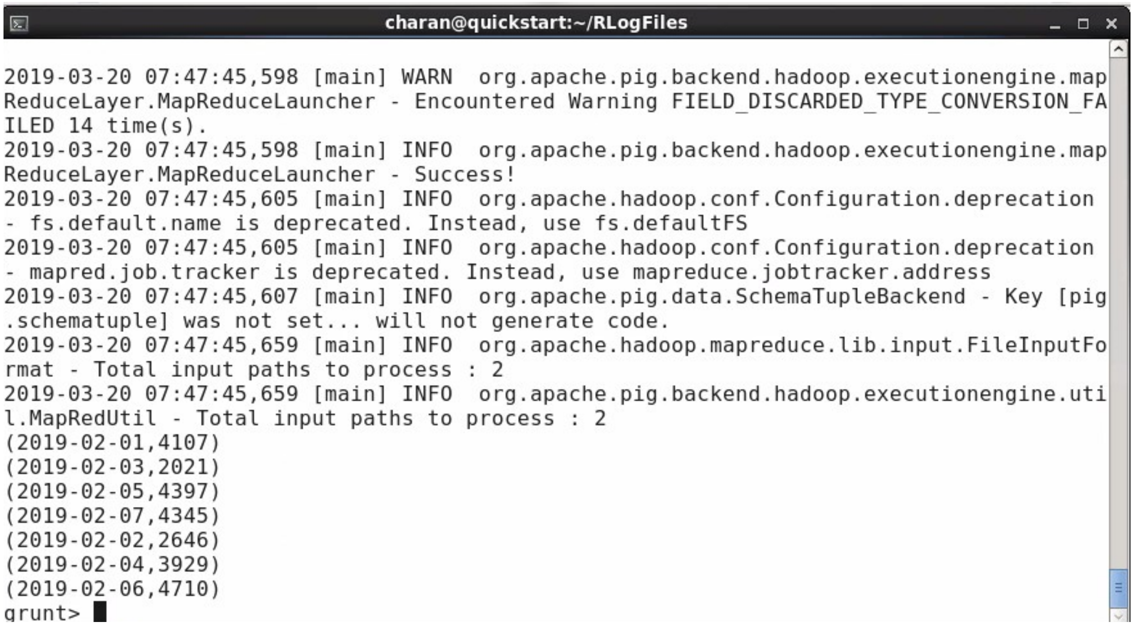
5. Pig Latin and HIVE: Machine Learning Frameworks
a. There are two important machine learning frameworks in R available: caret and mlr. Caret is widely used and mlr provides a second approach (please check the official CRAN webpage)
b. We are interested how frequent both frameworks are used: count the number of package caret and package mlr downloads each day)

register '/usr/lib/pig/piggybank.jar' ;
define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();
logsAll= LOAD '/user/master/RLogFiles/2019-02*.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE') AS
(date:chararray, time:chararray, size:long, r_version:chararray, r_arch:chararray, r_os:chararray,
package:chararray, version: chararray, country:chararray, ip_id : int);
mlrpkg = FILTER logsAll BY NOT(($6 != 'mlr'));
mlrgrp = GROUP mlrpkg BY date;
count = FOREACH mlrgrp GENERATE group as date,COUNT(mlrpkg) as cnt;
STORE count INTO '/user/master/result_new17' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE',
'UNIX','WRITE_OUTPUT_HEADER');



```
2019-03-20 07:24:40,019 [main] WARN  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FA
ILED 14 time(s).
2019-03-20 07:24:40,019 [main] INFO  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Success!
2019-03-20 07:24:40,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-20 07:24:40,020 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2019-03-20 07:24:40,020 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
.schematuple] was not set... will not generate code.
2019-03-20 07:24:40,027 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFo
rmat - Total input paths to process : 2
2019-03-20 07:24:40,027 [main] INFO  org.apache.pig.backend.hadoop.executionengine.uti
l.MapRedUtil - Total input paths to process : 2
(2019-02-01,536)
(2019-02-03,265)
(2019-02-05,476)
(2019-02-07,439)
(2019-02-02,228)
(2019-02-04,538)
(2019-02-06,597)
grunt> 
```

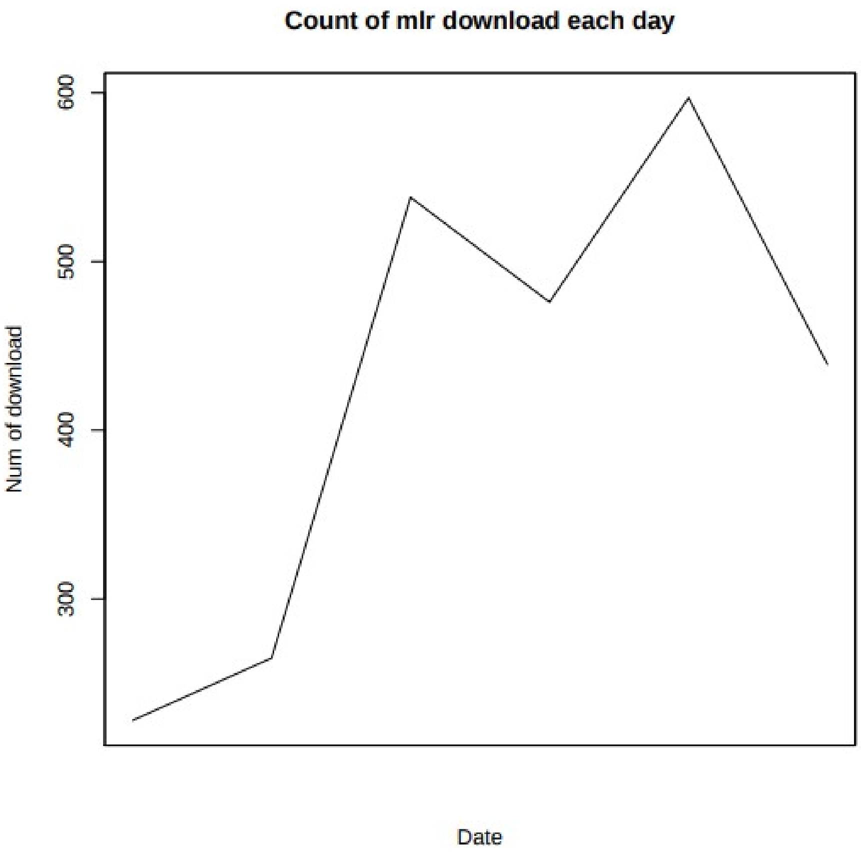
```
register '/usr/lib/pig/piggybank.jar' ;
define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();
logsAll= LOAD '/user/master/RLogFiles/2019-02*.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE') AS
(date:chararray, time:chararray, size:long, r_version:chararray, r_arch:chararray, r_os:chararray,
package:chararray, version: chararray, country:chararray, ip_id : int);
mlrpkg = FILTER logsAll BY NOT(($6 != 'caret'));
mlrgrp = GROUP mlrpkg BY date;
count = FOREACH mlrgrp GENERATE group as date,COUNT(mlrpkg) as cnt;
STORE count INTO '/user/master/result_new18' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE',
'UNIX','WRITE_OUTPUT_HEADER');
```

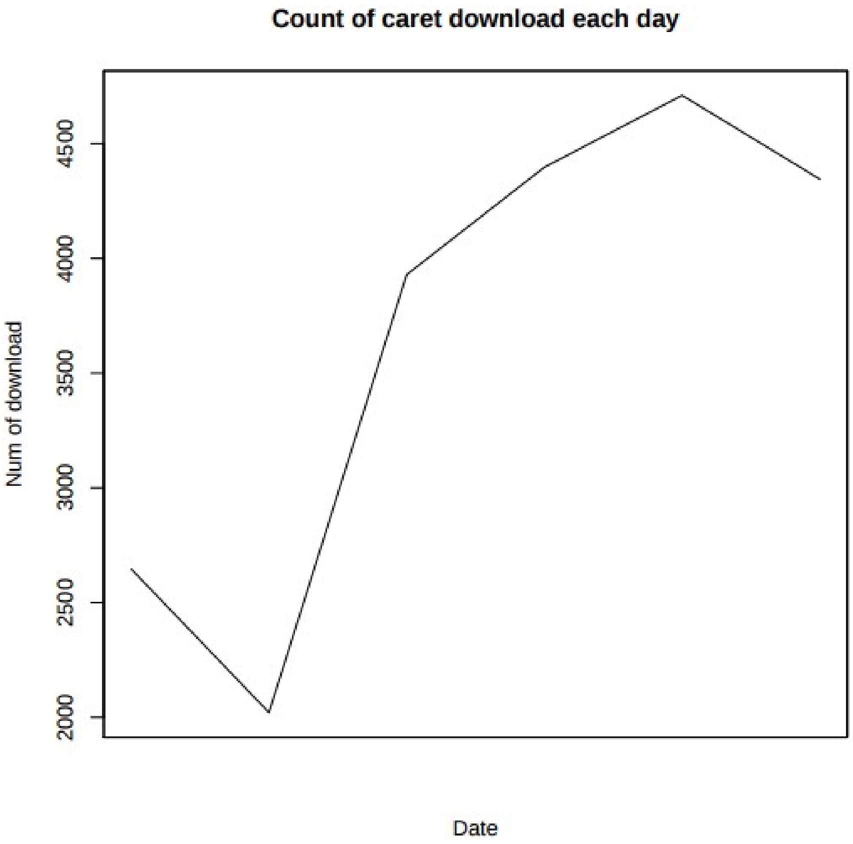


```
2019-03-20 07:47:45,598 [main] WARN  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FA
ILED 14 time(s).
2019-03-20 07:47:45,598 [main] INFO  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Success!
2019-03-20 07:47:45,605 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-20 07:47:45,605 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2019-03-20 07:47:45,607 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
.schematuple] was not set... will not generate code.
2019-03-20 07:47:45,659 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFo
rmat - Total input paths to process : 2
2019-03-20 07:47:45,659 [main] INFO  org.apache.pig.backend.hadoop.executionengine.uti
l.MapRedUtil - Total input paths to process : 2
(2019-02-01,4107)
(2019-02-03,2021)
(2019-02-05,4397)
(2019-02-07,4345)
(2019-02-02,2646)
(2019-02-04,3929)
(2019-02-06,4710)
grunt> 
```

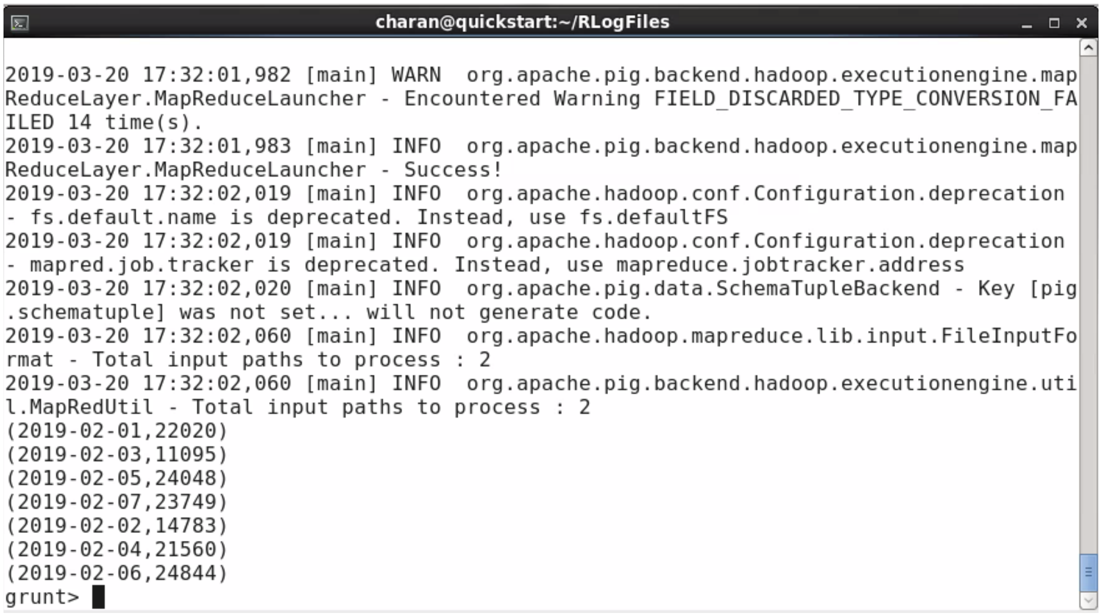
c. Visualize the results in R (line chart) (follow the step in No. 3 or import the results directly into R)

```
install.packages("https://cloud.r-project.org/src/contrib/csvread_1.2.1.tar.gz")
install.packages("https://cloud.r-project.org/src/contrib/zoo_1.8-4.tar.gz")

library('csvread')
library('zoo')
z<- read.zoo("mlr_count.csv", sep = ',', header=TRUE)
plot(z, xaxt="n", type = "l", xlab = "Date", ylab = "Num of download" )
title(main = "Count of ctv download each day")

library('csvread')
library('zoo')
z<- read.zoo("caret_count.csv", sep = ',', header=TRUE)
plot(z, xaxt="n", type = "l", xlab = "Date", ylab = "Num of download" )
title(main = "Count of ctv download each day")
```


Count of mlr download each day
Num of download
600
500
400
300
Date

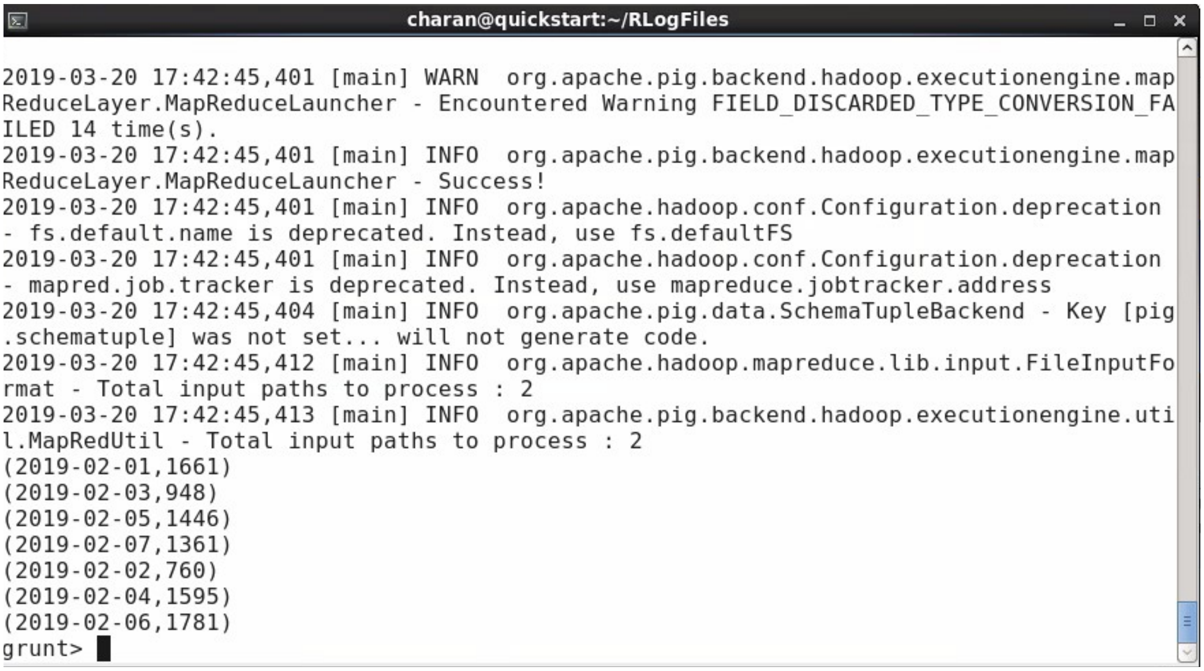

Count of caret download each day
Num of download
4500
4000
3500
3000
2500
2000
Date

6. Pig Latin and HIVE: Download volume (in MB) of caret and mlr packages
a. Use CRAN to find out the package size of the caret and mlr packages (use Windows-Package file size) in MB. Round to 1 decimal place.



```
2019-03-20 17:32:01,982 [main] WARN  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FA
ILED 14 time(s).
2019-03-20 17:32:01,983 [main] INFO  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Success!
2019-03-20 17:32:02,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-20 17:32:02,019 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2019-03-20 17:32:02,020 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
.schematuple] was not set... will not generate code.
2019-03-20 17:32:02,060 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFo
rmat - Total input paths to process : 2
2019-03-20 17:32:02,060 [main] INFO  org.apache.pig.backend.hadoop.executionengine.uti
l.MapRedUtil - Total input paths to process : 2
(2019-02-01,22020)
(2019-02-03,11095)
(2019-02-05,24048)
(2019-02-07,23749)
(2019-02-02,14783)
(2019-02-04,21560)
(2019-02-06,24844)
grunt> █
```

b. Enter this information into a text file together with the name (should be the same as in log-files)
c. Import this file into HDFS
d. Load the file in Pig (I assume that the RStudio CRAN Log Files are available already)



```
2019-03-20 17:42:45,401 [main] WARN  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FA
ILED 14 time(s).
2019-03-20 17:42:45,401 [main] INFO  org.apache.pig.backend.hadoop.executionengine.map
ReduceLayer.MapReduceLauncher - Success!
2019-03-20 17:42:45,401 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2019-03-20 17:42:45,401 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2019-03-20 17:42:45,404 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
.schematuple] was not set... will not generate code.
2019-03-20 17:42:45,412 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFo
rmat - Total input paths to process : 2
2019-03-20 17:42:45,413 [main] INFO  org.apache.pig.backend.hadoop.executionengine.uti
l.MapRedUtil - Total input paths to process : 2
(2019-02-01,1661)
(2019-02-03,948)
(2019-02-05,1446)
(2019-02-07,1361)
(2019-02-02,760)
(2019-02-04,1595)
(2019-02-06,1781)
grunt> █
```

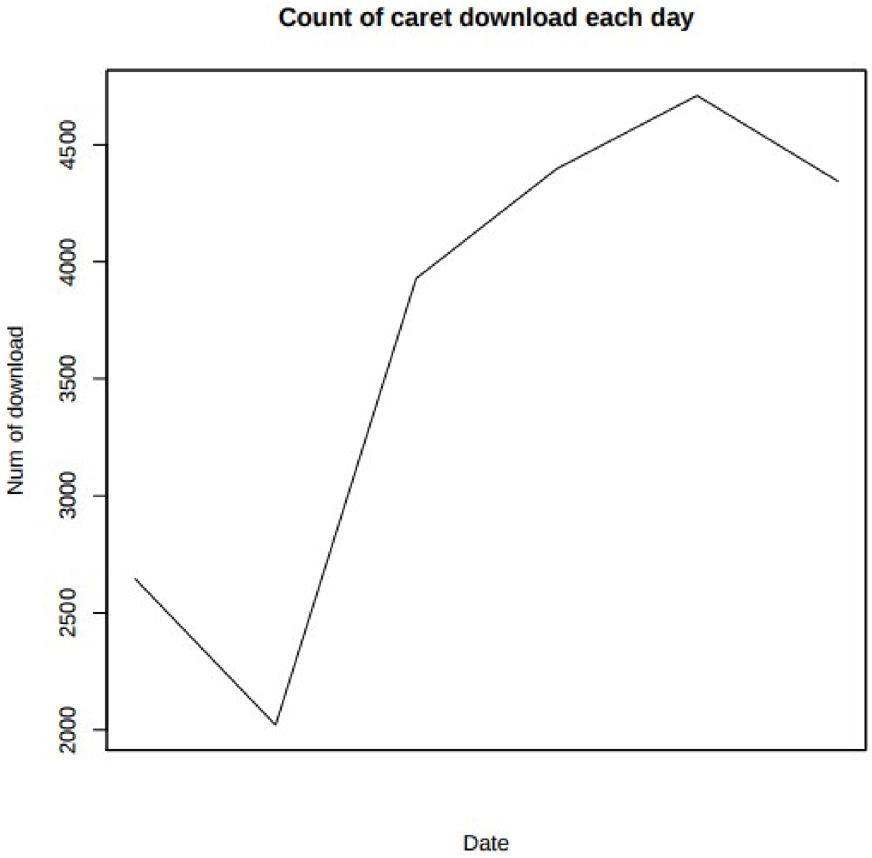
e. Filter out the caret and mlr packages in Pig
f. Add the size information

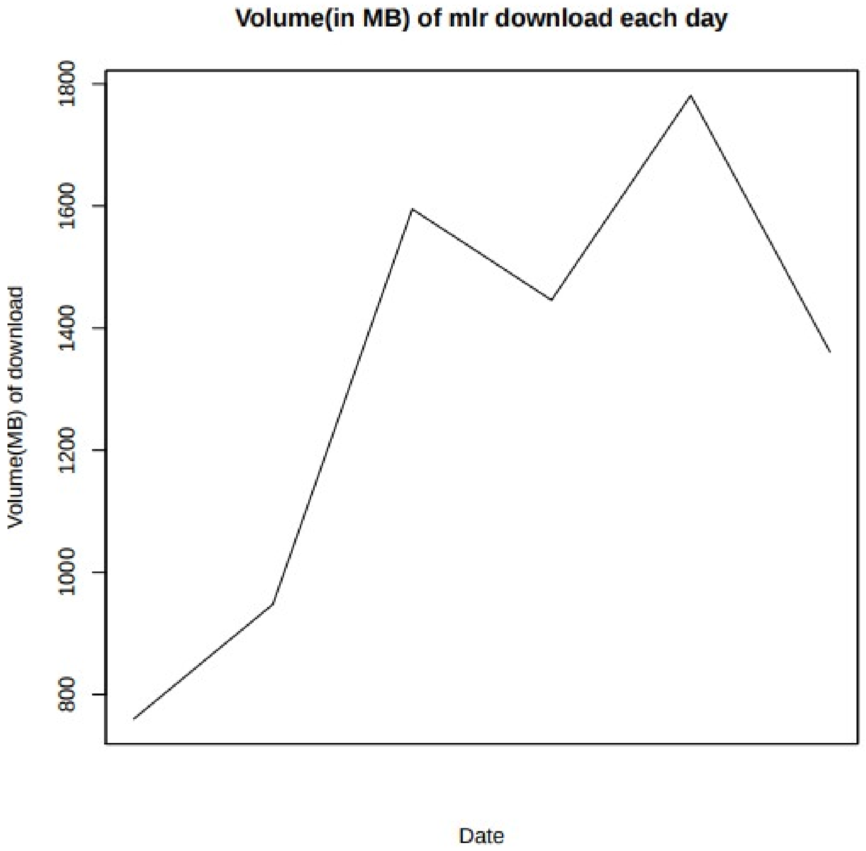
g. Calculate the download volume of caret and mlr packages by day
h. Export the results and display the results in R or Python

```
register '/usr/lib/pig/piggybank.jar' ;
define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();
logsAll= LOAD '/user/master/RLogFiles/2019-02*.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE') AS
(date:chararray, time:chararray, size:long, r_version:chararray, r_arch:chararray, r_os:chararray,
package:chararray, version: chararray, country:chararray, ip_id : int);
crtpkg = FILTER logsAll BY NOT(($6 != 'caret'));
crtgrp = GROUP crtpkg BY date;
volume = FOREACH crtgrp GENERATE group as date,ROUND(SUM(crtpkg.$2)*0.000001) as sum;
STORE volume INTO '/user/master/result_new19' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE',
'UNIX','WRITE_OUTPUT_HEADER');
```

```
register '/usr/lib/pig/piggybank.jar' ;
define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();
logsAll= LOAD '/user/master/RLogFiles/2019-02*.csv' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE') AS
(date:chararray, time:chararray, size:long, r_version:chararray, r_arch:chararray, r_os:chararray,
package:chararray, version: chararray, country:chararray, ip_id : int);
mlrpkg = FILTER logsAll BY NOT(($6 != 'mlr'));
mlrgrp = GROUP mlrpkg BY date;
volume = FOREACH mlrgrp GENERATE group as date,ROUND(SUM(mlrpkg.$2)*0.000001) as sum;
STORE volume INTO '/user/master/result_new20' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO_MULTILINE',
'UNIX','WRITE_OUTPUT_HEADER');
```

f. Add the size information
g. Calculate the download volume of each of the 10 packages by day
h. Export the results and display the results in R or Pythoncreate table package_size(package_r varchar(50), counting long);


Count of caret download each day
Num of download
2000
2500
3000
3500
4000
4500
Date


Volume(in MB) of mlr download each day
Volume(MB) of download
800
1000
1200
1400
1600
1800
Date