Task 1- Apache PIG – Analyzing Log-Files

1. Import Rstudio LogFiles from one month (October 2017) into HDFS

a) Download from R studio CRAN log files page,

In R console,

```
start <- as.Date(‘2017-10-01’)
end <- as.Date(‘2017-10-31’)
all_days <- seq(start, end, by = ‘day’)
year <- as.POSIXlt(all_days)$year + 1900
urls <- paste0(‘http://cran-logs.rstudio.com/’,year,’/’,all_days, ‘.csv.gz’)
filenames <- paste0(‘~/Downloads/’, “,’’,c(1:31),’.csv.gz’)
download.file(url = urls[1], destfile = filenames[1])
for (i in 1:31) download.file(url=urls[i], destfile = filenames[i])
```

b) unzip the files

In the command line,

```
gunzip -dk *.gz
```

c. Import the complete directory into HDFS into folder RLogFiles

In the command line,

```
hdfs dfs -put ~/Downloads/RLogFiles/
```

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

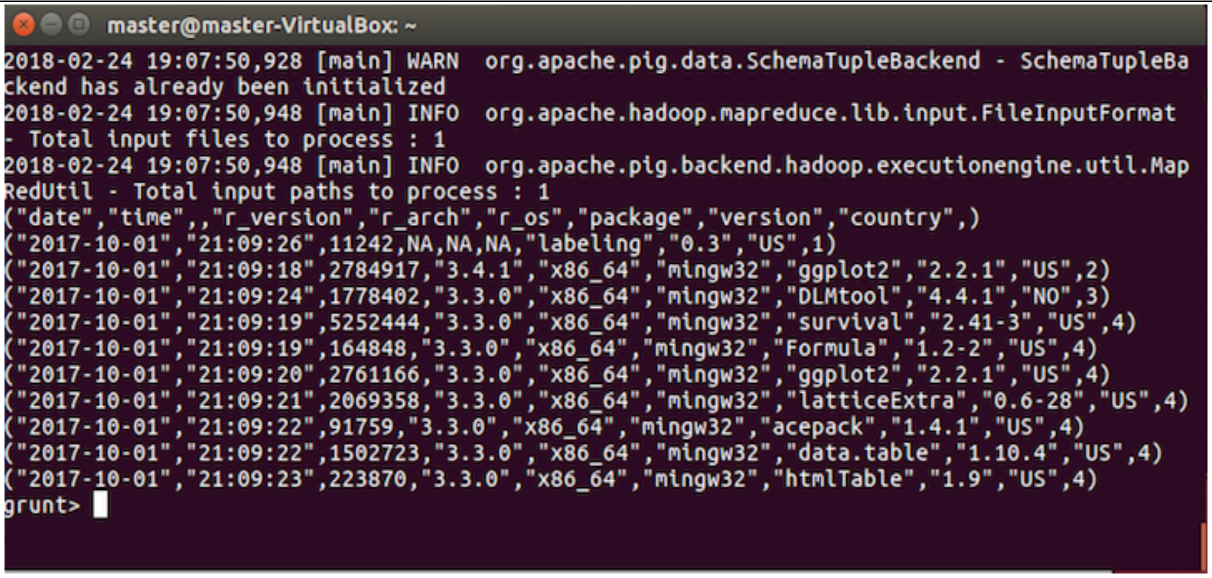
a. Load log-file of one day (1st of October)

In Pig,

```
A = LOAD '/user/master/RLogFiles/1.csv' USING PigStorage(',') AS ( date:chararray,
time:chararray, size:int, r_version:chararray, r_arch:chararray, r_os:chararray,
package:chararray, version:chararray, country:chararray, ip_id:int );
```

b. Dump the first 10 entries on the screen to check if it works

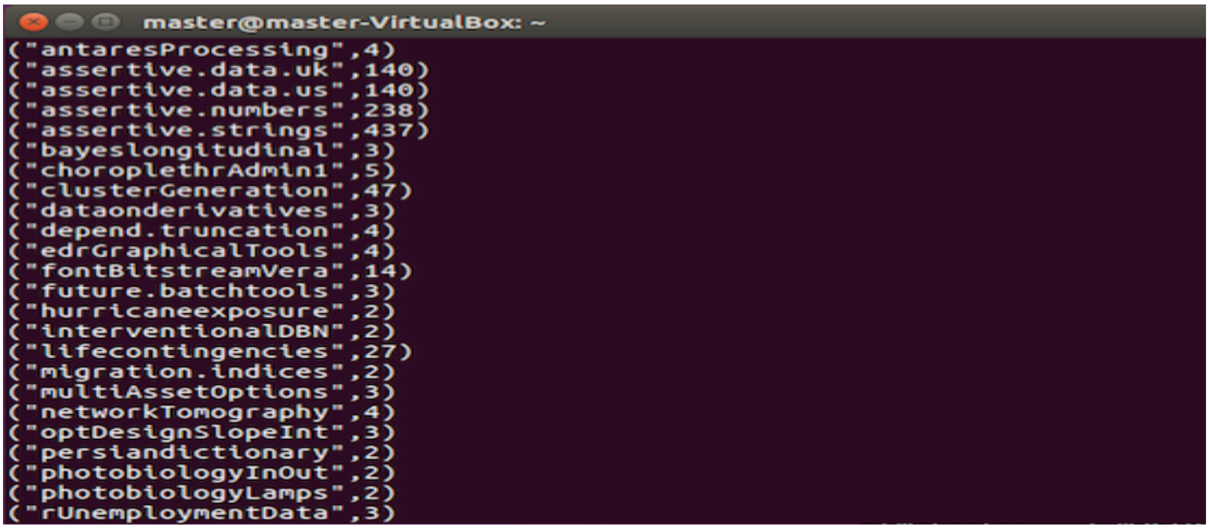
```
B = LIMIT A 10;
```



```
2018-02-24 19:07:50,928 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBa
ckend has already been initialized
2018-02-24 19:07:50,948 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat
- Total input files to process : 1
2018-02-24 19:07:50,948 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.Map
RedUtil - Total input paths to process : 1
("date","time",,"r_version","r_arch","r_os","package","version","country",)
("2017-10-01","21:09:26",11242,NA,NA,NA,"labeling","0.3","US",1)
("2017-10-01","21:09:18",2784917,"3.4.1","x86_64","mingw32","ggplot2","2.2.1","US",2)
("2017-10-01","21:09:24",1778402,"3.3.0","x86_64","mingw32","DLMtool","4.4.1","NO",3)
("2017-10-01","21:09:19",5252444,"3.3.0","x86_64","mingw32","survival","2.41-3","US",4)
("2017-10-01","21:09:19",164848,"3.3.0","x86_64","mingw32","Formula","1.2-2","US",4)
("2017-10-01","21:09:20",2761166,"3.3.0","x86_64","mingw32","ggplot2","2.2.1","US",4)
("2017-10-01","21:09:21",2069358,"3.3.0","x86_64","mingw32","latticeExtra","0.6-28","US",4)
("2017-10-01","21:09:22",91759,"3.3.0","x86_64","mingw32","acepack","1.4.1","US",4)
("2017-10-01","21:09:22",1502723,"3.3.0","x86_64","mingw32","data.table","1.10.4","US",4)
("2017-10-01","21:09:23",223870,"3.3.0","x86_64","mingw32","htmlTable","1.9","US",4)
grunt>
```

c. Count the number of occurrences of different packages

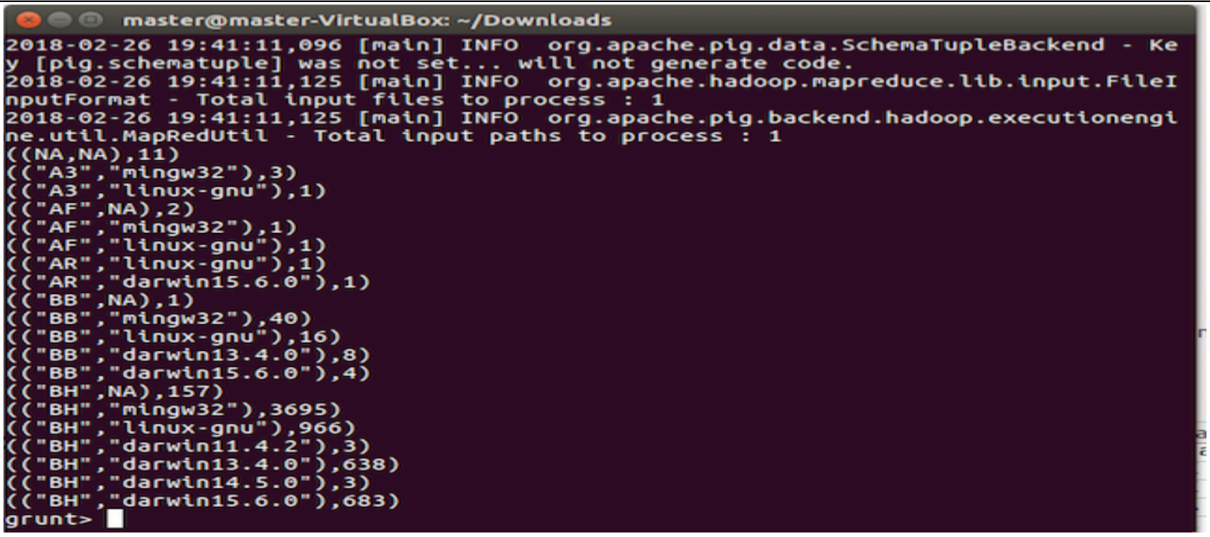
C = GROUP A by package;
D = FOREACH C GENERATE group as (package), COUNT(A) as (count);
E = ORDER D BY count DESC;
F = LIMIT E 100;
DUMP F;



```
("antaresProcessing",4)
("assertive.data.uk",140)
("assertive.data.us",140)
("assertive.numbers",238)
("assertive.strings",437)
("bayeslongitudinal",3)
("choroplethrAdmin1",5)
("clusterGeneration",47)
("dataonderivatives",3)
("depend.truncation",4)
("edrGraphicalTools",4)
("fontBitstreamVera",14)
("future.batchtools",3)
("hurricaneexposure",2)
("interventionalDBN",2)
("lifecontingencies",27)
("migration.indices",2)
("multiAssetOptions",3)
("networkTomography",4)
("optDesignSlopeInt",3)
("persiandictionary",2)
("photobiologyInOut",2)
("photobiologyLamps",2)
("rUnemploymentData",3)
```

d. Count the number of occurances of different package by os

G = GROUP A by (package,r_os);
H = FOREACH G GENERATE group as (packagewos), COUNT(A) as (count);
I = ORDER H BY count DESC;
J = LIMIT I 100;
DUMP J;



```
2018-02-26 19:41:11,096 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-02-26 19:41:11,125 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2018-02-26 19:41:11,125 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
((NA,NA),11)
(("A3","mingw32"),3)
(("A3","linux-gnu"),1)
(("AF",NA),2)
(("AF","mingw32"),1)
(("AF","linux-gnu"),1)
(("AR","linux-gnu"),1)
(("AR","darwin15.6.0"),1)
(("BB",NA),1)
(("BB","mingw32"),40)
(("BB","linux-gnu"),16)
(("BB","darwin13.4.0"),8)
(("BB","darwin15.6.0"),4)
(("BH",NA),157)
(("BH","mingw32"),3695)
(("BH","linux-gnu"),966)
(("BH","darwin11.4.2"),3)
(("BH","darwin13.4.0"),638)
(("BH","darwin14.5.0"),3)
(("BH","darwin15.6.0"),683)
grunt> 
```

e. Store the results of both operations in HDFS

STORE F INTO '/user/master/RLogFiles/output/' USING PigStorage(',', '-schema');
STORE J INTO '/user/master/RLogFiles/output1/' USING PigStorage(',', '-schema');

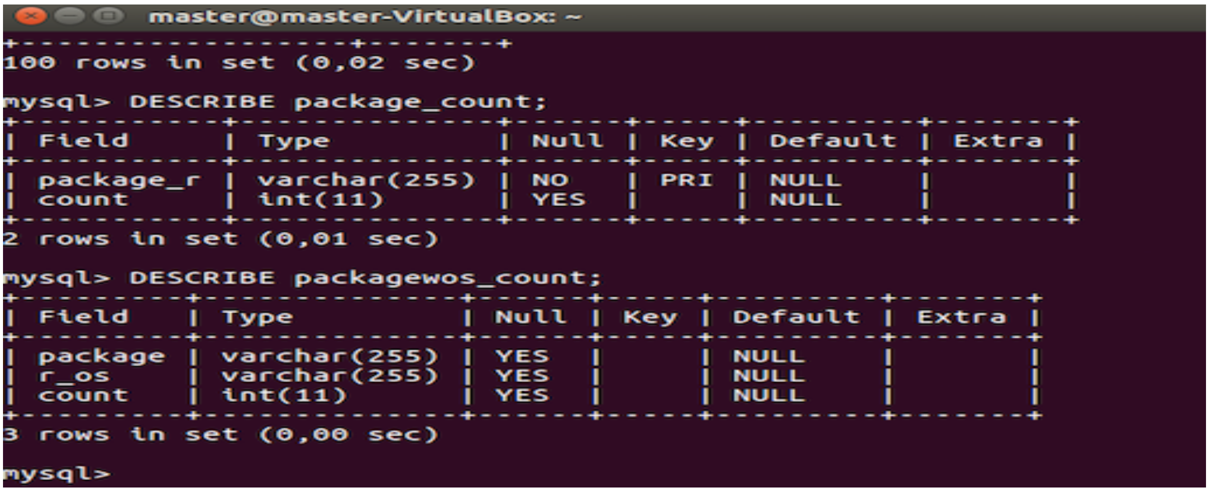
3. Sqoop, MySQL and R studio

a. Export the results of both operations via sqoop into MySQL

In MySQL command line,

CREATE DATABASE assignment;
USE assignment;

CREATE TABLE package_count (package_r varchar(255) NOT NULL PRIMARY KEY, count int);
CREATE TABLE packagewos_count (package varchar(255), r_os varchar(255), count int);



```
+-------------------+-------+
100 rows in set (0,02 sec)

mysql> DESCRIBE package_count;
+-----------+--------------+------+-----+---------+-------+
| Field     | Type         | Null | Key | Default | Extra |
+-----------+--------------+------+-----+---------+-------+
| package_r | varchar(255) | NO   | PRI | NULL    |       |
| count     | int(11)      | YES  |     | NULL    |       |
+-----------+--------------+------+-----+---------+-------+
2 rows in set (0,01 sec)

mysql> DESCRIBE packagewos_count;
+---------+--------------+------+-----+---------+-------+
| Field   | Type         | Null | Key | Default | Extra |
+---------+--------------+------+-----+---------+-------+
| package | varchar(255) | YES  |     | NULL    |       |
| r_os    | varchar(255) | YES  |     | NULL    |       |
| count   | int(11)      | YES  |     | NULL    |       |
+---------+--------------+------+-----+---------+-------+
3 rows in set (0,00 sec)

mysql>
```

In command line,

```
sqoop export --connect "jdbc:mysql://localhost/assignment" --username root --password 123456789 --table package_count --export-dir /user/master/RLogFiles/output/part-r-00000 -m 1

sqoop export --connect "jdbc:mysql://localhost/assignment" --username root --password 123456789 --table package_count --export-dir /user/master/RLogFiles/output1/part-r-00000 -m 1
```

Can check if the import is complete using,

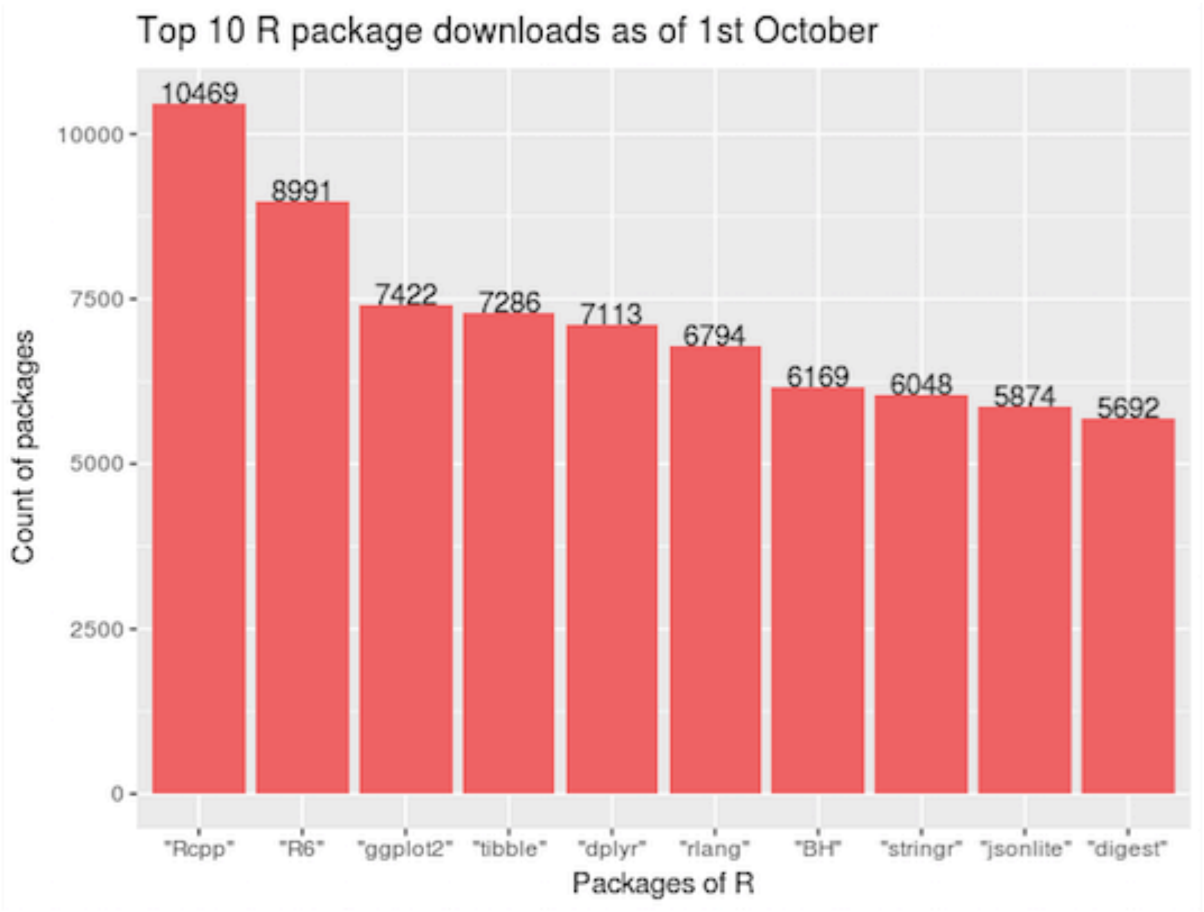
```
SELECT * FROM package_count;
SELECT * FROM packagewos_count;
```

b. Access the tables by R-studio and display the results (Top-10 in bar chart)

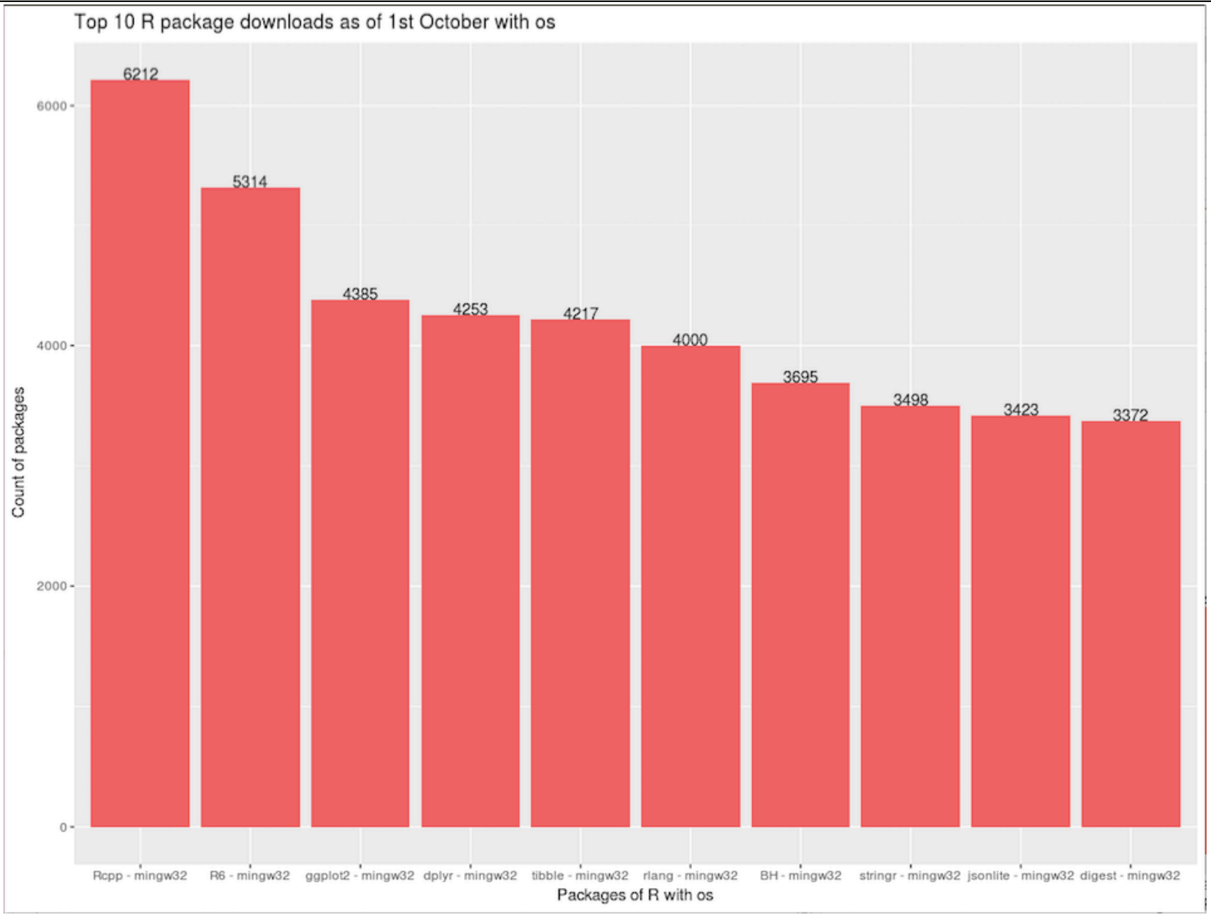
Open R studio,

```
install.packages("RMySQL", dependencies = TRUE)
install.packages("dbConnect")
library(RMySQL)
library(dbConnect)

drv = dbDriver("MySQL")
con <-dbConnect(drv = drv,
            user = 'root',
            password = '123456789',
            host = '127.0.0.1',
            dbname = 'assignment',
            port = 3306)

dbGetInfo (con)
dbListTables(con)

package_count <- dbGetQuery(con, "SELECT * FROM package_count")
packagewos_count <- dbGetQuery(con, "SELECT * FROM packagewos_count")

packagewos_count$package<- gsub("[[:punct:]]", "", packagewos_count$package)
packagewos_count$r_os<- gsub("[[:punct:]]", "", packagewos_count$r_os)
# to remove punctuations from the variable.

plot1 <- package_count %>% arrange(desc(count))

plot2 <- plot1[c(1:10),]

bar <- ggplot(plot2,aes(x = reorder(package_r, -count), y = count))
```

```
bar+geom_bar(stat = "identity", fill = "#FF6666") +
geom_text(aes(label= count, vjust=0)) +
xlab("Packages of R") + ylab("Count of packages") +
ggtitle("Top 10 R package downloads as of 1st October")
```



```
# For second graph,

plot1_os <- packagewos_count %>% arrange(desc(count))

plot2_os <- plot1_os[c(1:10),]

plot2_os$combined <- paste(plot2_os$package, "-", plot2_os$r_os)

bar_both <- ggplot(plot2_os,aes(x = reorder(combined, -count), y = count))
bar_both+geom_bar(stat = "identity", fill = "#FF6666")+
geom_text(aes(label= count, vjust=0)) +
xlab("Packages of R with os") + ylab("Count of packages") +
ggtitle("Top 10 R package downloads as of 1st October with os")
```



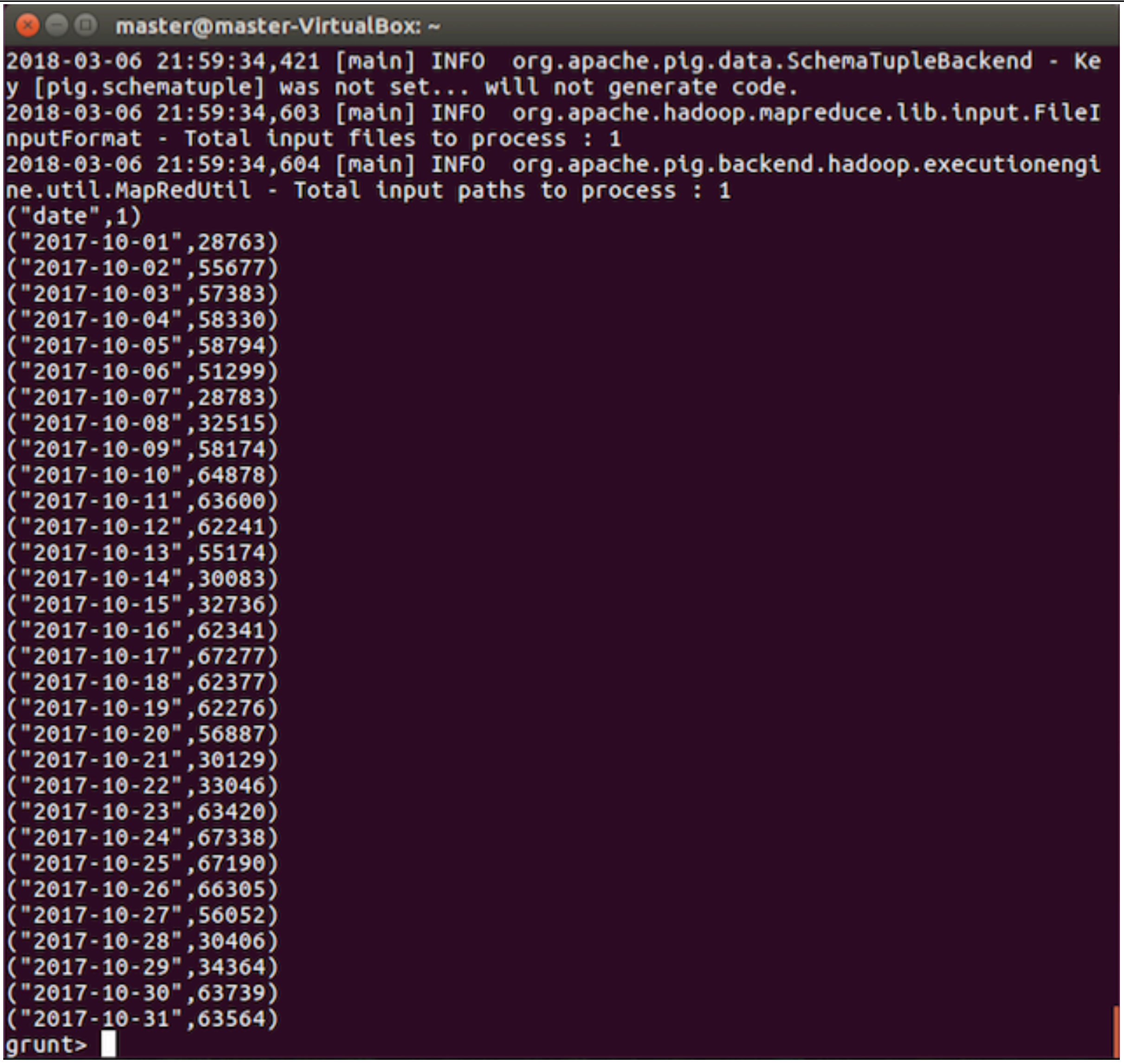
4. Pig Latin: Number of individual users each day

a. Load the log-files into HDFS

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

b. Count the number of distinct users each day

```
B = GROUP A BY (date,ip_id);

C = FOREACH B GENERATE $0, COUNT($1);

D = FOREACH C GENERATE FLATTEN ($0);

E = GROUP D BY date;

final = FOREACH E GENERATE $0, COUNT($1);

DUMP final;
```



```
2018-03-06 21:59:34,421 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2018-03-06 21:59:34,603 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input files to process : 1
2018-03-06 21:59:34,604 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
("date",1)
("2017-10-01",28763)
("2017-10-02",55677)
("2017-10-03",57383)
("2017-10-04",58330)
("2017-10-05",58794)
("2017-10-06",51299)
("2017-10-07",28783)
("2017-10-08",32515)
("2017-10-09",58174)
("2017-10-10",64878)
("2017-10-11",63600)
("2017-10-12",62241)
("2017-10-13",55174)
("2017-10-14",30083)
("2017-10-15",32736)
("2017-10-16",62341)
("2017-10-17",67277)
("2017-10-18",62377)
("2017-10-19",62276)
("2017-10-20",56887)
("2017-10-21",30129)
("2017-10-22",33046)
("2017-10-23",63420)
("2017-10-24",67338)
("2017-10-25",67190)
("2017-10-26",66305)
("2017-10-27",56052)
("2017-10-28",30406)
("2017-10-29",34364)
("2017-10-30",63739)
("2017-10-31",63564)
grunt> 
```

5. Average number of packages downloaded by an individual user each day

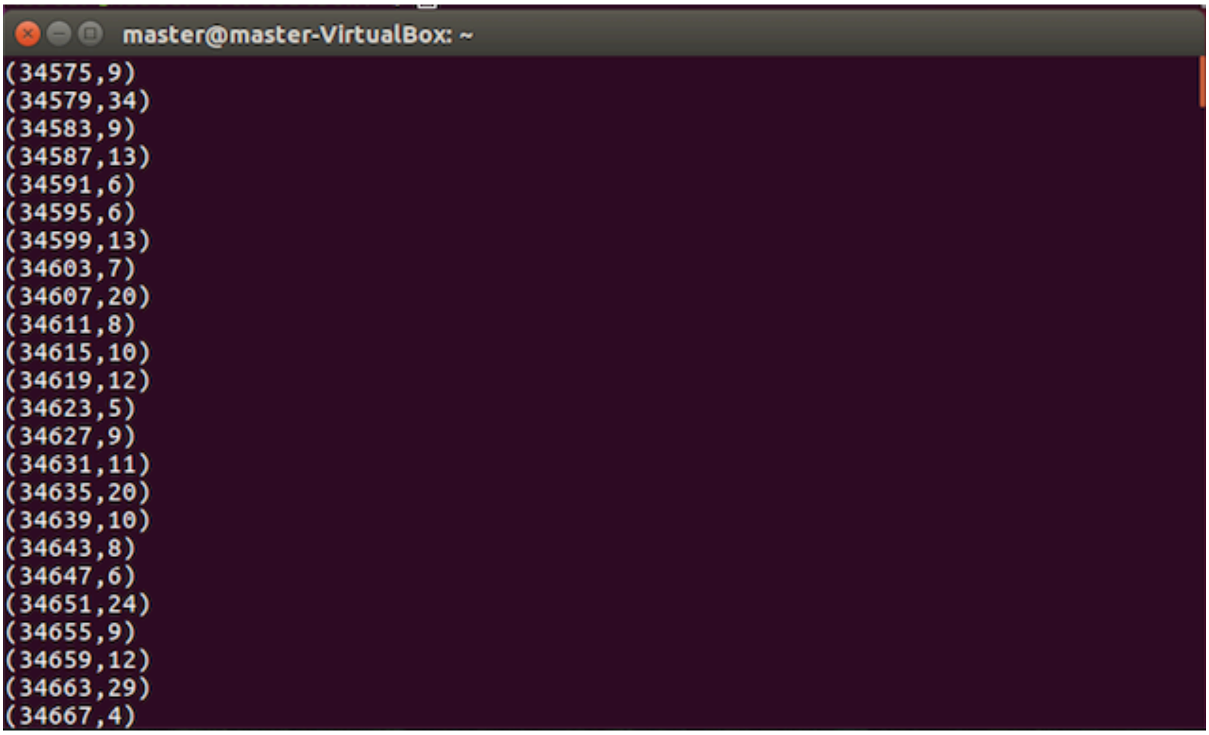
a. Load the log-files into HDFS

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

b. Average number of packages download by an individual user each day

B = GROUP A by ip_id;

C = FOREACH B GENERATE group, (COUNT(A)/31) as avg;



```
(34575,9)
(34579,34)
(34583,9)
(34587,13)
(34591,6)
(34595,6)
(34599,13)
(34603,7)
(34607,20)
(34611,8)
(34615,10)
(34619,12)
(34623,5)
(34627,9)
(34631,11)
(34635,20)
(34639,10)
(34643,8)
(34647,6)
(34651,24)
(34655,9)
(34659,12)
(34663,29)
(34667,4)
```

6. Pig Latin: Task Views

a. Task views are collections of R packages of a certain topic

b. Check if Task Views are used by R-users (package ctv)

In Pig,

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

B = FILTER A by package =='"ctv"';

C = GROUP B by date;

D = FOREACH C GENERATE $0, COUNT_STAR($1) AS cnt;

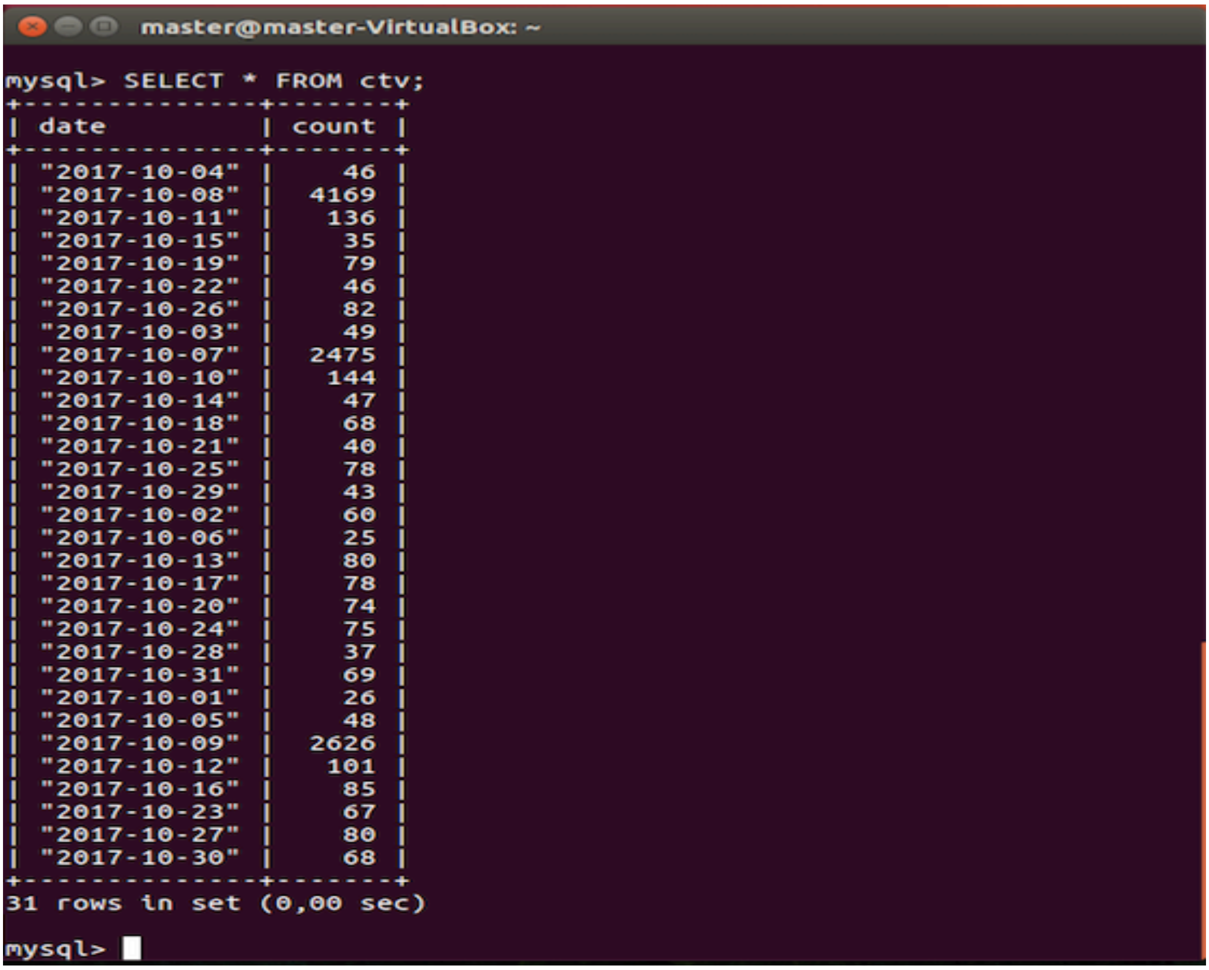
STORE D INTO '/user/master/RLogFiles/ctvfinal/' USING PigStorage(',', '-schema');

In MySQL,

CREATE TABLE ctv (date varchar(255), count int);

Sqoop,

```
sqoop export --connect "jdbc:mysql://localhost/assignment" --username root --password 123456789 --table ctv --export-dir /user/master/RLogFiles/ctvfinal/part-r-00000 -m 1
```

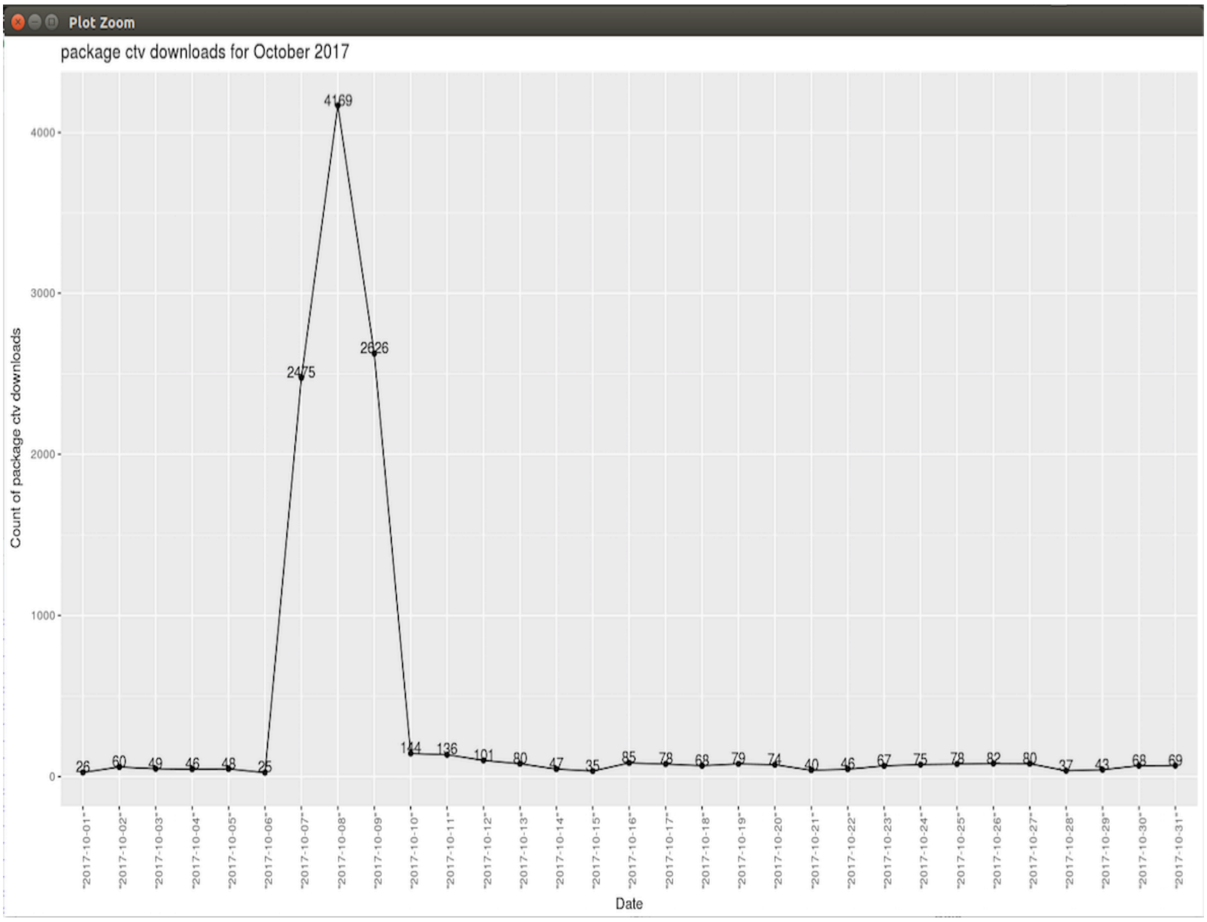


```
mysql> SELECT * FROM ctv;
+--------------+-------+
| date         | count |
+--------------+-------+
| "2017-10-04" |    46 |
| "2017-10-08" |  4169 |
| "2017-10-11" |   136 |
| "2017-10-15" |    35 |
| "2017-10-19" |    79 |
| "2017-10-22" |    46 |
| "2017-10-26" |    82 |
| "2017-10-03" |    49 |
| "2017-10-07" |  2475 |
| "2017-10-10" |   144 |
| "2017-10-14" |    47 |
| "2017-10-18" |    68 |
| "2017-10-21" |    40 |
| "2017-10-25" |    78 |
| "2017-10-29" |    43 |
| "2017-10-02" |    60 |
| "2017-10-06" |    25 |
| "2017-10-13" |    80 |
| "2017-10-17" |    78 |
| "2017-10-20" |    74 |
| "2017-10-24" |    75 |
| "2017-10-28" |    37 |
| "2017-10-31" |    69 |
| "2017-10-01" |    26 |
| "2017-10-05" |    48 |
| "2017-10-09" |  2626 |
| "2017-10-12" |   101 |
| "2017-10-16" |    85 |
| "2017-10-23" |    67 |
| "2017-10-27" |    80 |
| "2017-10-30" |    68 |
+--------------+-------+
31 rows in set (0,00 sec)

mysql> 
```

c. Visualize the results in R studio: line chart

R studio,

```
install.packages("RMySQL", dependencies = TRUE)
install.packages("dbConnect")

library(RMySQL)
library(dbConnect)
library(ggplot2)

drv = dbDriver("MySQL")
con <-dbConnect(drv = drv,
                user = 'root',
                password = '123456789',
                host = '127.0.0.1',
          dbname = 'assignment',
          port = 3306)

dbGetInfo (con)
dbListTables(con)

ctv <- dbGetQuery(con, "SELECT * FROM ctv")

line <- ggplot(ctv,aes(x = date, y = count, group = 1)) + geom_line() + geom_point() +
geom_text(aes(label= count, vjust=0)) +
  xlab("Date") + ylab("Count of package ctv downloads") +
  ggtitle("package ctv downloads for 5 days") +
  theme(axis.text.x=element_text(angle = 90, hjust = 0))

line
```


Plot Zoom
package ctv downloads for October 2017
Count of package ctv downloads
Date
"2017-10-01"
"2017-10-02"
"2017-10-03"
"2017-10-04"
"2017-10-05"
"2017-10-06"
"2017-10-07"
"2017-10-08"
"2017-10-09"
"2017-10-10"
"2017-10-11"
"2017-10-12"
"2017-10-13"
"2017-10-14"
"2017-10-15"
"2017-10-16"
"2017-10-17"
"2017-10-18"
"2017-10-19"
"2017-10-20"
"2017-10-21"
"2017-10-22"
"2017-10-23"
"2017-10-24"
"2017-10-25"
"2017-10-26"
"2017-10-27"
"2017-10-28"
"2017-10-29"
"2017-10-30"
"2017-10-31"