

1. Import RStudio Log Files from one month (e.g., December 2018) into HDFS

a. Download from RStudio CRAN log files page

- i) 'cd' into Downloads directory. (cd Downloads)
- ii) create a txt file 'download_file_list.txt' using cat command and list all 31 http links files for Dec 2014 R log files.
- iii) enter `command 'wget -i download_file_list.txt'`
- iv) this command will download all 31 files listed inside the txt file.
- v) create directory 'RLogFiles' and move all 31 zip files into it.
- vi) `mkdir RLogFiles; mv 2014* RLogFiles`

b. Unzip the files

- i) 'cd' into RLogFiles. (cd RLogFiles)
- ii) decompress all files using `command 'gzip -d *.gz'`

c. Import the complete directory into HDFS into a folder RLogFiles

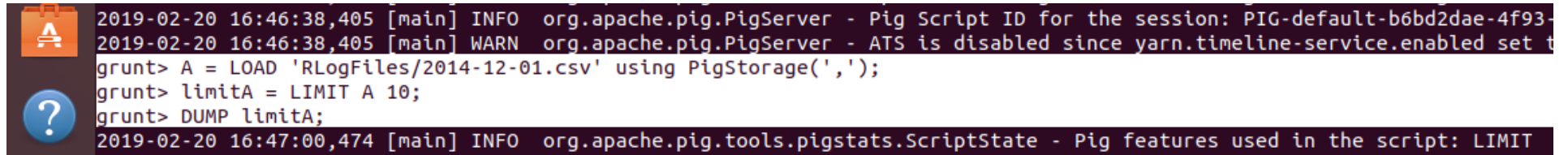
- i) copy directory RLogFiles into HDFS folder RLogFiles using command

```
'hdfs dfs -put ~/Downloads/RLogFiles /user/master/'
```

```
File Edit View Search Terminal Help
master@master:~$ hdfs dfs -put RLogFiles/
put: 'RLogFiles/': No such file or directory
master@master:~$ cd Downloads/
master@master:~/Downloads$ hdfs dfs -put RLogFiles/
master@master:~/Downloads$ hdfs dfs -ls RLogFiles
Found 31 items
-rw-r--r-- 1 master supergroup 26577453 2019-02-20 16:43 RLogFiles/2014-12-01.csv
-rw-r--r-- 1 master supergroup 28911187 2019-02-20 16:43 RLogFiles/2014-12-02.csv
-rw-r--r-- 1 master supergroup 29525861 2019-02-20 16:43 RLogFiles/2014-12-03.csv
-rw-r--r-- 1 master supergroup 26053851 2019-02-20 16:43 RLogFiles/2014-12-04.csv
-rw-r--r-- 1 master supergroup 22446071 2019-02-20 16:43 RLogFiles/2014-12-05.csv
-rw-r--r-- 1 master supergroup 16799310 2019-02-20 16:43 RLogFiles/2014-12-06.csv
-rw-r--r-- 1 master supergroup 16624680 2019-02-20 16:43 RLogFiles/2014-12-07.csv
-rw-r--r-- 1 master supergroup 20910127 2019-02-20 16:43 RLogFiles/2014-12-08.csv
-rw-r--r-- 1 master supergroup 26288696 2019-02-20 16:43 RLogFiles/2014-12-09.csv
-rw-r--r-- 1 master supergroup 26866418 2019-02-20 16:43 RLogFiles/2014-12-10.csv
-rw-r--r-- 1 master supergroup 23887181 2019-02-20 16:43 RLogFiles/2014-12-11.csv
-rw-r--r-- 1 master supergroup 19083753 2019-02-20 16:43 RLogFiles/2014-12-12.csv
-rw-r--r-- 1 master supergroup 17258463 2019-02-20 16:43 RLogFiles/2014-12-13.csv
-rw-r--r-- 1 master supergroup 13078614 2019-02-20 16:43 RLogFiles/2014-12-14.csv
-rw-r--r-- 1 master supergroup 21220829 2019-02-20 16:43 RLogFiles/2014-12-15.csv
-rw-r--r-- 1 master supergroup 22924397 2019-02-20 16:44 RLogFiles/2014-12-16.csv
-rw-r--r-- 1 master supergroup 22801191 2019-02-20 16:44 RLogFiles/2014-12-17.csv
-rw-r--r-- 1 master supergroup 21306109 2019-02-20 16:44 RLogFiles/2014-12-18.csv
-rw-r--r-- 1 master supergroup 20057996 2019-02-20 16:44 RLogFiles/2014-12-19.csv
-rw-r--r-- 1 master supergroup 16177906 2019-02-20 16:44 RLogFiles/2014-12-20.csv
-rw-r--r-- 1 master supergroup 12549729 2019-02-20 16:44 RLogFiles/2014-12-21.csv
-rw-r--r-- 1 master supergroup 16993142 2019-02-20 16:44 RLogFiles/2014-12-22.csv
-rw-r--r-- 1 master supergroup 14078294 2019-02-20 16:44 RLogFiles/2014-12-23.csv
-rw-r--r-- 1 master supergroup 13627108 2019-02-20 16:44 RLogFiles/2014-12-24.csv
-rw-r--r-- 1 master supergroup 8808043 2019-02-20 16:44 RLogFiles/2014-12-25.csv
-rw-r--r-- 1 master supergroup 10860639 2019-02-20 16:44 RLogFiles/2014-12-26.csv
-rw-r--r-- 1 master supergroup 12823138 2019-02-20 16:44 RLogFiles/2014-12-27.csv
-rw-r--r-- 1 master supergroup 10561456 2019-02-20 16:44 RLogFiles/2014-12-28.csv
-rw-r--r-- 1 master supergroup 14979841 2019-02-20 16:44 RLogFiles/2014-12-29.csv
-rw-r--r-- 1 master supergroup 13528508 2019-02-20 16:44 RLogFiles/2014-12-30.csv
-rw-r--r-- 1 master supergroup 10926780 2019-02-20 16:44 RLogFiles/2014-12-31.csv
master@master:~/Downloads$ hdfs dfs -ls
Found 1 items
drwxr-xr-x - master supergroup 0 2019-02-20 16:44 RLogFiles
```

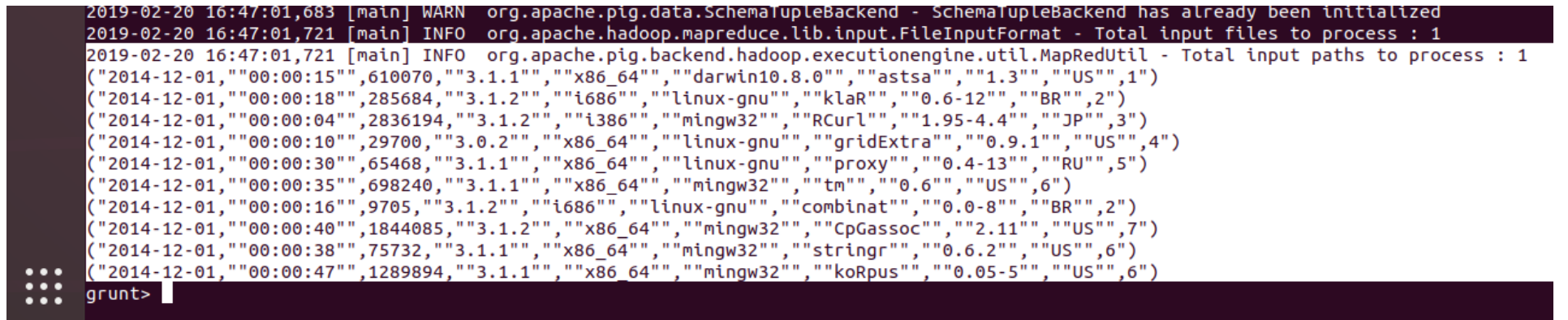
2. Pig Latin: Top-100-packages (by operating system)

a. Load log-file of one day (e.g., 1st of December 2018)



```
2019-02-20 16:46:38,405 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-b6bd2dae-4f93-
2019-02-20 16:46:38,405 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set t
grunt> A = LOAD 'RLogFiles/2014-12-01.csv' using PigStorage(',');
grunt> limita = LIMIT A 10;
grunt> DUMP limita;
2019-02-20 16:47:00,474 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
```

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



```
2019-02-20 16:47:01,683 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2019-02-20 16:47:01,721 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2019-02-20 16:47:01,721 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
("2014-12-01","00:00:15",610070,"3.1.1","x86_64","darwin10.8.0","astsa","1.3","US",1)
("2014-12-01","00:00:18",285684,"3.1.2","i686","linux-gnu","klaR","0.6-12","BR",2)
("2014-12-01","00:00:04",2836194,"3.1.2","i386","mingw32","RCurl","1.95-4.4","JP",3)
("2014-12-01","00:00:10",29700,"3.0.2","x86_64","linux-gnu","gridExtra","0.9.1","US",4)
("2014-12-01","00:00:30",65468,"3.1.1","x86_64","linux-gnu","proxy","0.4-13","RU",5)
("2014-12-01","00:00:35",698240,"3.1.1","x86_64","mingw32","tm","0.6","US",6)
("2014-12-01","00:00:16",9705,"3.1.2","i686","linux-gnu","combinat","0.0-8","BR",2)
("2014-12-01","00:00:40",1844085,"3.1.2","x86_64","mingw32","CpGassoc","2.11","US",7)
("2014-12-01","00:00:38",75732,"3.1.1","x86_64","mingw32","stringr","0.6.2","US",6)
("2014-12-01","00:00:47",1289894,"3.1.1","x86_64","mingw32","koRpus","0.05-5","US",6)
grunt>
```

c. Count the number of occurrences of different packages;

```
("2014-12-01","00:00:47",1289894,"3.1.1","x86_64","mingw32","koRpus","0.05-5","US",6)
```

```
grunt> A = LOAD 'RLogFiles/2014-12-01.csv' using PigStorage(',');
```

```
grunt> B = GROUP A BY $6;
```

```
grunt> C = FOREACH B generate group, COUNT(A.$6) as pkcount;
```

```
grunt> STORE C into 'pkcount_dec1_2014';
```

```
2019-02-20 16:56:02,673 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. I  
extoutputformat.separator
```

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime	Alias	Feature
job_local374249666_0001	1	1	n/a	n/a	n/a	n/a	n/a	n/a	A,B,C	GROUP_BY,COMBINER	hdfs://localhost:8020/user/master/pkcount_dec1_2014,	

Input(s):

Successfully read 264991 records (64676892 bytes) from: "hdfs://localhost:8020/user/master/RLogFiles/2014-12-01.csv"

Output(s):

Successfully stored 6286 records (11609902 bytes) in: "hdfs://localhost:8020/user/master/pkcount_dec1_2014"

Counters:

Total records written : 6286

Total bytes written : 11609902

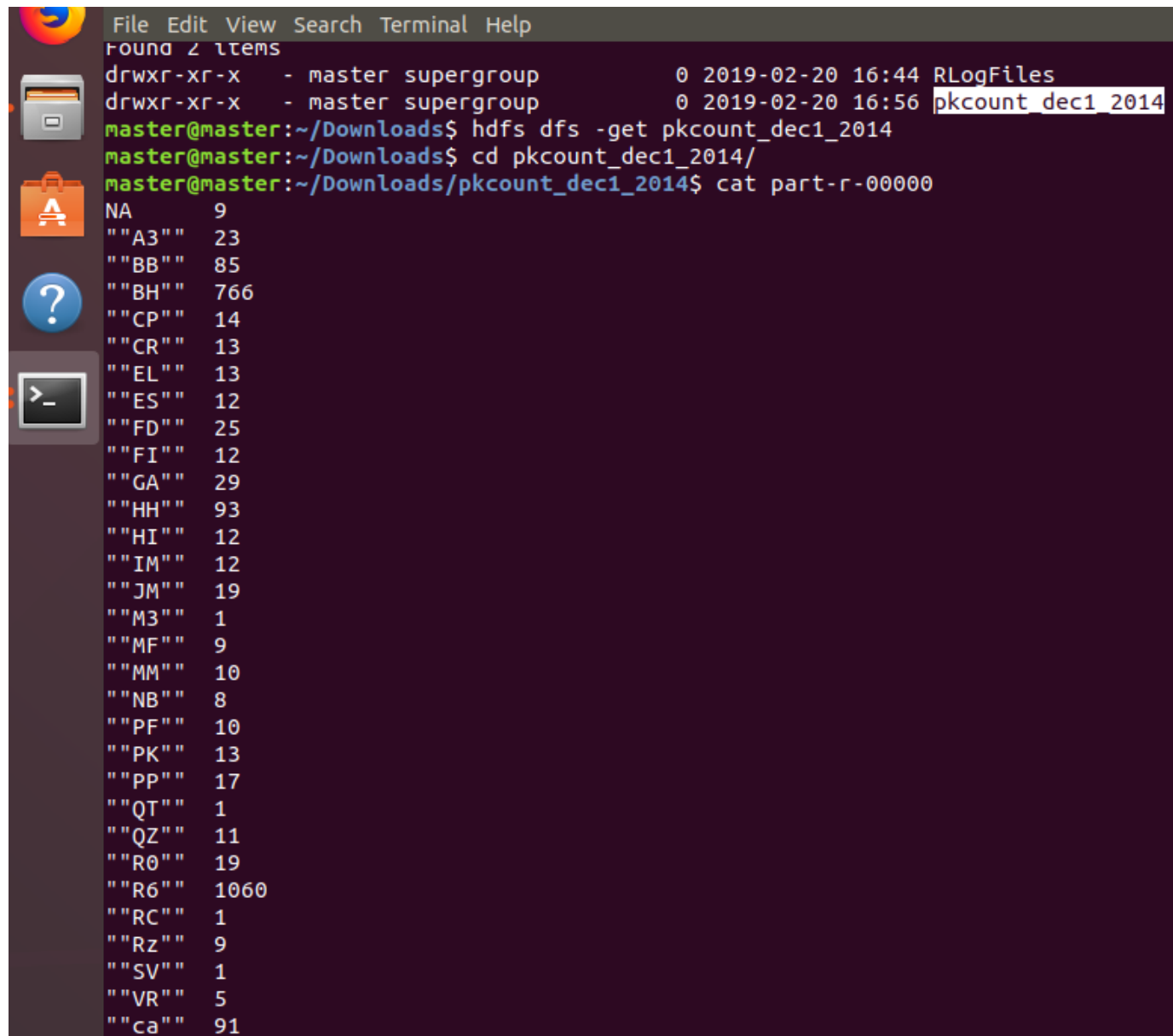
Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_local374249666_0001



A terminal window with a dark purple background and a sidebar on the left containing icons for a file manager, a folder, a question mark, and a terminal. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The output shows the results of a search for 'pkcount_dec1_2014' in the 'Downloads' directory, listing two files: 'RLogFiles' and 'pkcount_dec1_2014'. The user then navigates to the 'pkcount_dec1_2014' directory and runs 'cat part-r-00000', which displays a list of items with their names and counts.

```
File Edit View Search Terminal Help
Found 2 items
drwxr-xr-x - master supergroup      0 2019-02-20 16:44 RLogFiles
drwxr-xr-x - master supergroup      0 2019-02-20 16:56 pkcount_dec1_2014
master@master:~/Downloads$ hdfs dfs -get pkcount_dec1_2014
master@master:~/Downloads$ cd pkcount_dec1_2014/
master@master:~/Downloads/pkcount_dec1_2014$ cat part-r-00000
NA          9
""A3""     23
""BB""     85
""BH""    766
""CP""     14
""CR""     13
""EL""     13
""ES""     12
""FD""     25
""FI""     12
""GA""     29
""HH""     93
""HI""     12
""IM""     12
""JM""     19
""M3""      1
""MF""      9
""MM""     10
""NB""      8
""PF""     10
""PK""     13
""PP""     17
""QT""      1
""QZ""     11
""R0""     19
""R6""    1060
""RC""      1
""Rz""      9
""SV""      1
""VR""      5
""Ca""     91
```

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

```
2019-02-20 17:09:47,411 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = LOAD 'RLogFiles/2014-12-01.csv' using PigStorage(',');
grunt> B = FOREACH A generate $5 as os, $6 as package;
grunt> C = GROUP B by os;
grunt> D = FOREACH C generate group as os, COUNT(B);
grunt> STORE D into 'package_count_by_os';
2019-02-20 17:12:27,770 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated.

Job Stats (time in seconds):
JobId  Maps  Reduces MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReductime  Alias  Feature  Outputs
job_local571176766_0001 1 1 n/a n/a n/a n/a n/a n/a A,B,C,D GROUP_BY,COMBINER hdfs://localhost:8020/user/master/package_count_by_os,

Input(s):
Successfully read 264991 records (64666470 bytes) from: "hdfs://localhost:8020/user/master/RLogFiles/2014-12-01.csv"

Output(s):
Successfully stored 19 records (11511935 bytes) in: "hdfs://localhost:8020/user/master/package_count_by_os"

Counters:
Total records written : 19
Total bytes written : 11511935
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local571176766_0001
```


```
master@master:~/Downloads/pkcount_dec1_2014$ cd ..
master@master:~/Downloads$ hdfs dfs -ls
Found 3 items
drwxr-xr-x - master supergroup          0 2019-02-20 16:44 RLogFiles
drwxr-xr-x - master supergroup          0 2019-02-20 17:12 package_count_by_os
drwxr-xr-x - master supergroup          0 2019-02-20 16:56 pkcount_dec1_2014
master@master:~/Downloads$ hdfs dfs -get package_count_by_os
master@master:~/Downloads$ cd package_count_by_os/
master@master:~/Downloads/package_count_by_os$ cat part-r-00000
NA      19589
"cygwin"      70
"mingw32"     167364
"linux-gnu"   38903
"aix6.1.0.0"  1
"darwin9.8.0" 535
"darwin10.5.0" 14
"darwin10.8.0" 16368
"darwin12.4.0" 33
"darwin12.5.0" 50
"darwin13.0.0" 16
"darwin13.1.0" 9536
"darwin13.2.0" 83
"darwin13.3.0" 126
"darwin13.4.0" 11641
"darwin14.0.0" 587
"darwin8.10.1" 12
"darwin8.11.1" 54
"linux-gnueabi" 9
master@master:~/Downloads/package_count_by_os$
```

e. Store the results of both operations in HDFS;

```
master@master:~/Downloads/package_count_by_os$ cd ..
master@master:~/Downloads$ hdfs dfs -ls
Found 3 items
drwxr-xr-x - master supergroup          0 2019-02-20 16:44 RLogFiles
drwxr-xr-x - master supergroup          0 2019-02-20 17:12 package_count_by_os
drwxr-xr-x - master supergroup          0 2019-02-20 16:56 pkcount_dec1_2014
master@master:~/Downloads$
```

3. sqoop, MySQL and R/Python:

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



```
mysql> create DATABASE rlogs;
Query OK, 1 row affected (0.00 sec)


mysql> use rlogs;
Database changed
mysql> create table pkg_count(pk_name varchar(40), count int(10));
Query OK, 0 rows affected (0.17 sec)

mysql> create table pkg_count_by_os(os_name varchar(40), pk_count int(10));
Query OK, 0 rows affected (0.10 sec)

mysql> describe pkg_count;
+-----+-----+-----+-----+-----+-----+
| Field  | Type          | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| pk_name | varchar(40)   | YES  |     | NULL    |       |
| count   | int(10)       | YES  |     | NULL    |       |
+-----+-----+-----+-----+-----+-----+
2 rows in set (0.01 sec)

mysql> describe pkg_count_by_os;
+-----+-----+-----+-----+-----+-----+
| Field    | Type          | Null | Key | Default | Extra |
+-----+-----+-----+-----+-----+-----+
| os_name  | varchar(40)   | YES  |     | NULL    |       |
| pk_count | int(10)       | YES  |     | NULL    |       |
+-----+-----+-----+-----+-----+-----+
2 rows in set (0.01 sec)

mysql>
```



try --help for usage instructions.

```
master@master:/usr/local/sqoop$ sqoop export --connect jdbc:mysql://localhost:3306/rlogs --username root --password 123 --table pkg_count --export-dir "pkcount_dec1_2014/part-r-00000" --input-fields-terminated-by '\t'
```



```
Bytes Written=0
19/02/20 18:05:41 INFO mapred.LocalJobRunner: Finishing task: attempt_local1042702955_0001_m_000003_0
19/02/20 18:05:41 INFO mapred.LocalJobRunner: map task executor complete.
19/02/20 18:05:41 INFO mapreduce.Job: map 100% reduce 0%
19/02/20 18:05:41 INFO mapreduce.Job: Job job_local1042702955_0001 completed successfully
19/02/20 18:05:41 INFO mapreduce.Job: Counters: 20
  File System Counters
    FILE: Number of bytes read=81483586
    FILE: Number of bytes written=83700392
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=277176
    HDFS: Number of bytes written=0
    HDFS: Number of read operations=54
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=0
  Map-Reduce Framework
    Map input records=6286
    Map output records=6286
    Input split bytes=580
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=206
    Total committed heap usage (bytes)=1023410176
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=0
19/02/20 18:05:41 INFO mapreduce.ExportJobBase: Transferred 270.6797 KB in 6.4497 seconds (41.968 KB/sec)
19/02/20 18:05:41 INFO mapreduce.ExportJobBase: Exported 6286 records.
master@master:/usr/local/sqoop$
```

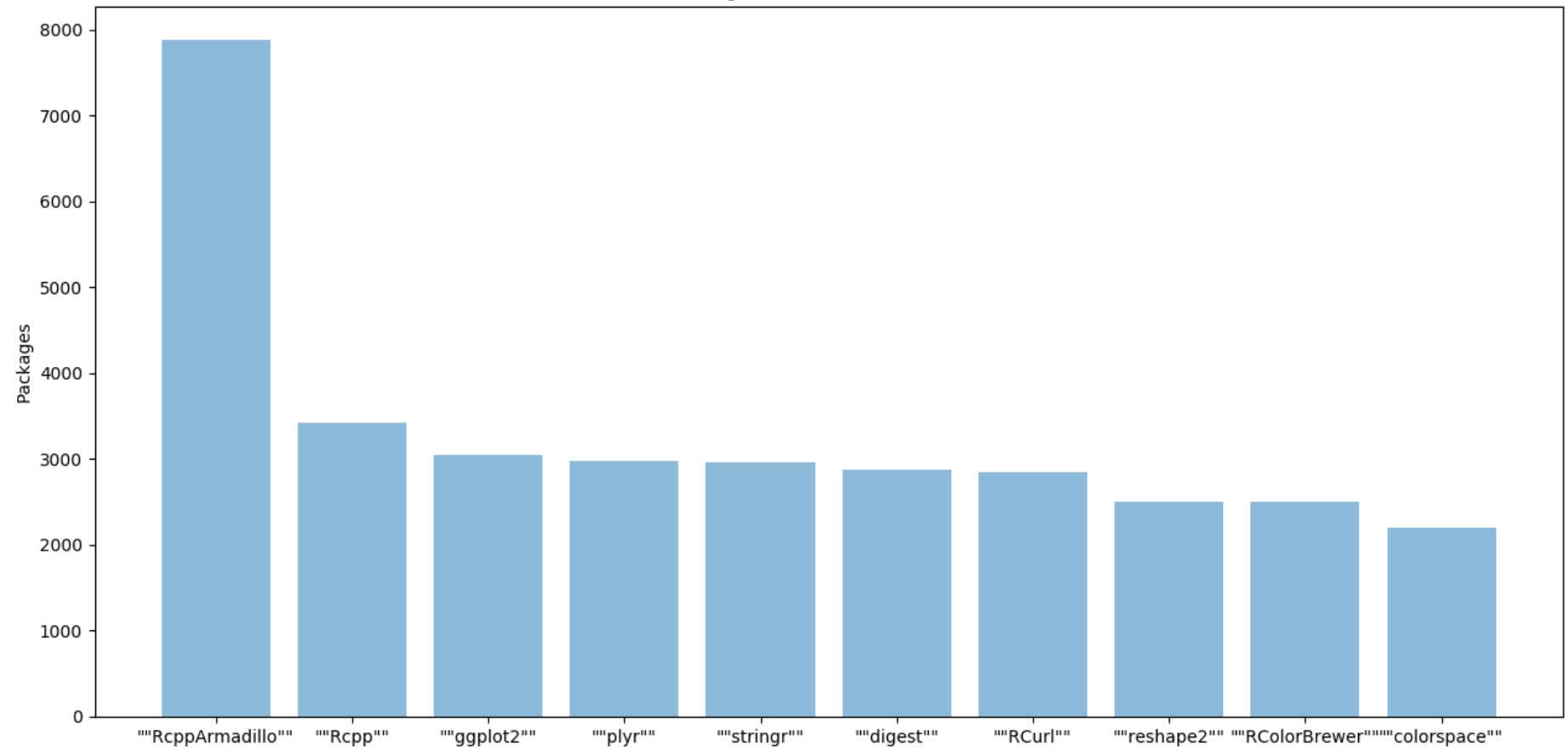


```
master@master:/usr/local/sqoop$ sqoop export --connect jdbc:mysql://localhost:3306/rlogs --username root --password 123 --table pkg_count_by_os --export-dir "package_count_by_os/part-r-00000" --input-fields-terminated-by '\t'
```

```
?
19/02/20 18:09:40 INFO mapred.LocalJobRunner: Finishing task: attempt_local597954113_0001_m_000003_0
19/02/20 18:09:40 INFO mapred.LocalJobRunner: map task executor complete.
19/02/20 18:09:41 INFO mapreduce.Job: map 100% reduce 0%
19/02/20 18:09:41 INFO mapreduce.Job: Job job_local597954113_0001 completed successfully
19/02/20 18:09:41 INFO mapreduce.Job: Counters: 20
    File System Counters
        FILE: Number of bytes read=81484264
        FILE: Number of bytes written=83693940
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=2484
        HDFS: Number of bytes written=0
        HDFS: Number of read operations=66
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=0
    Map-Reduce Framework
        Map input records=19
        Map output records=19
        Input split bytes=671
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=185
        Total committed heap usage (bytes)=1050673152
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=0
19/02/20 18:09:41 INFO mapreduce.ExportJobBase: Transferred 2.4258 KB in 6.0799 seconds (408.5584 bytes/sec)
19/02/20 18:09:41 INFO mapreduce.ExportJobBase: Exported 19 records.
master@master:/usr/local/sqoop$
```

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

Packages download 2014-12-01



Package Count

```
Traceback (most recent call last):
  File "pkg.py", line 18, in <module>
    import matplotlib.pyplot as plt; plt.rcParams['figure.figsize'] = (10, 6)
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    _backend_mod, new_figure_manager, draw_if_interactive, _initialize()
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    [backend_name], 0)
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    from . import tkagg # Paint image to Tk photo buffer
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    from six.moves import tkinter as Tk
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    mod = mod._resolve()
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    return _import_module(self.mod)
  File "/home/master/.local/lib/python2.7/site-packages/matplotlib/figure.py", line 101, in __init__
    __import__(name)
  File "/usr/lib/python2.7/lib-tk/Tkinter.py", line 177, in <module>
    raise ImportError, str(msg) + ', please install the Tkinter package'
ImportError: No module named _tkinter, please install the Tkinter package
master@master:~/tflow$ python pkg.py
Package_Name Count
0 "RcppArmadillo" 7881
1 "Rcpp" 3425
2 "ggplot2" 3044
3 "plyr" 2972
4 "stringr" 2961
5 "digest" 2881
6 "RCurl" 2847
7 "reshape2" 2506
8 "RColorBrewer" 2504
9 "colorspace" 2192
master@master:~/tflow$ gedit pkg.py
master@master:~/tflow$ gedit pkg.py
```

```
Open [icon] pkg.py ~/tflow

import mysql.connector
import pandas as pd

mysqlconnection = mysql.connector.connect(host='localhost',
                                         database='rlogs',
                                         user='root',
                                         password='123')

sql_select_Query = "select * from pkg_count order by count desc"
cursor = mysqlconnection.cursor()
cursor.execute(sql_select_Query)
records = cursor.fetchall()
df = pd.DataFrame(records, columns=['Package_Name', 'Count'])
cursor.close()

df10 = df.head(10);
print(df10);

import matplotlib.pyplot as plt; plt.rcParams['figure.figsize'] = (10, 6)
import numpy as np
import matplotlib.pyplot as plt

y = df10["Package_Name"]
y_pos = np.arange(len(y))
x = df10["Count"]

plt.bar(y_pos, x, align='center', alpha=0.5)
plt.xticks(y_pos, y)
plt.ylabel('Packages')
plt.title('Packages download 2014-12-01')

plt.show()
```

Package Count By OS

```
1  "linux-gnu" 38903
2  NA 19589
3  "darwin10.8.0" 16368
4  "darwin13.4.0" 11641
5  "darwin13.1.0" 9536
6  "darwin14.0.0" 587
7  "darwin9.8.0" 535
8  "darwin13.3.0" 126
9  "darwin13.2.0" 83
Traceback (most recent call last):
  File "pkg_os.py", line 24, in <module>
    x = df210["Count"]
  File "/home/master/.local/lib/python2.7/site-packages/pandas/indexer.py", line 108, in pandas._i
    indexer = self.columns.get_loc(key)
  File "/home/master/.local/lib/python2.7/site-packages/pandas/indexer.py", line 132, in pandas._i
    return self._engine.get_loc(self._maybe_cast_indexer
  File "pandas/_libs/index.pyx", line 108, in pandas._l
  File "pandas/_libs/index.pyx", line 132, in pandas._l
  File "pandas/_libs/hashtable_class_helper.pxi", line
  File "pandas/_libs/hashtable_class_helper.pxi", line
KeyError: 'Count'
master@master:~/tflow$ gedit pkg_os.py
master@master:~/tflow$ python pkg_os.py
  OS_Name  Pkg_Count
0  "mingw32" 167364
1  "linux-gnu" 38903
2  NA 19589
3  "darwin10.8.0" 16368
4  "darwin13.4.0" 11641
5  "darwin13.1.0" 9536
6  "darwin14.0.0" 587
7  "darwin9.8.0" 535
8  "darwin13.3.0" 126
9  "darwin13.2.0" 83
master@master:~/tflow$ gedit pkg_os.py
```

```
pkg_os.py
~/tflow
import mysql.connector
import pandas as pd

mysqlconnection = mysql.connector.connect(host='localhost',
                                         database='rlogs',
                                         user='root',
                                         password='123')

sql_select_Query = "select * from pkg_count_by_os order by pk_count desc"
cursor = mysqlconnection.cursor()
cursor.execute(sql_select_Query)
records2 = cursor.fetchall()
df2 = pd.DataFrame(records2, columns=['OS_Name', 'Pkg_Count'])
cursor.close()

df210 = df2.head(10);
print(df210);

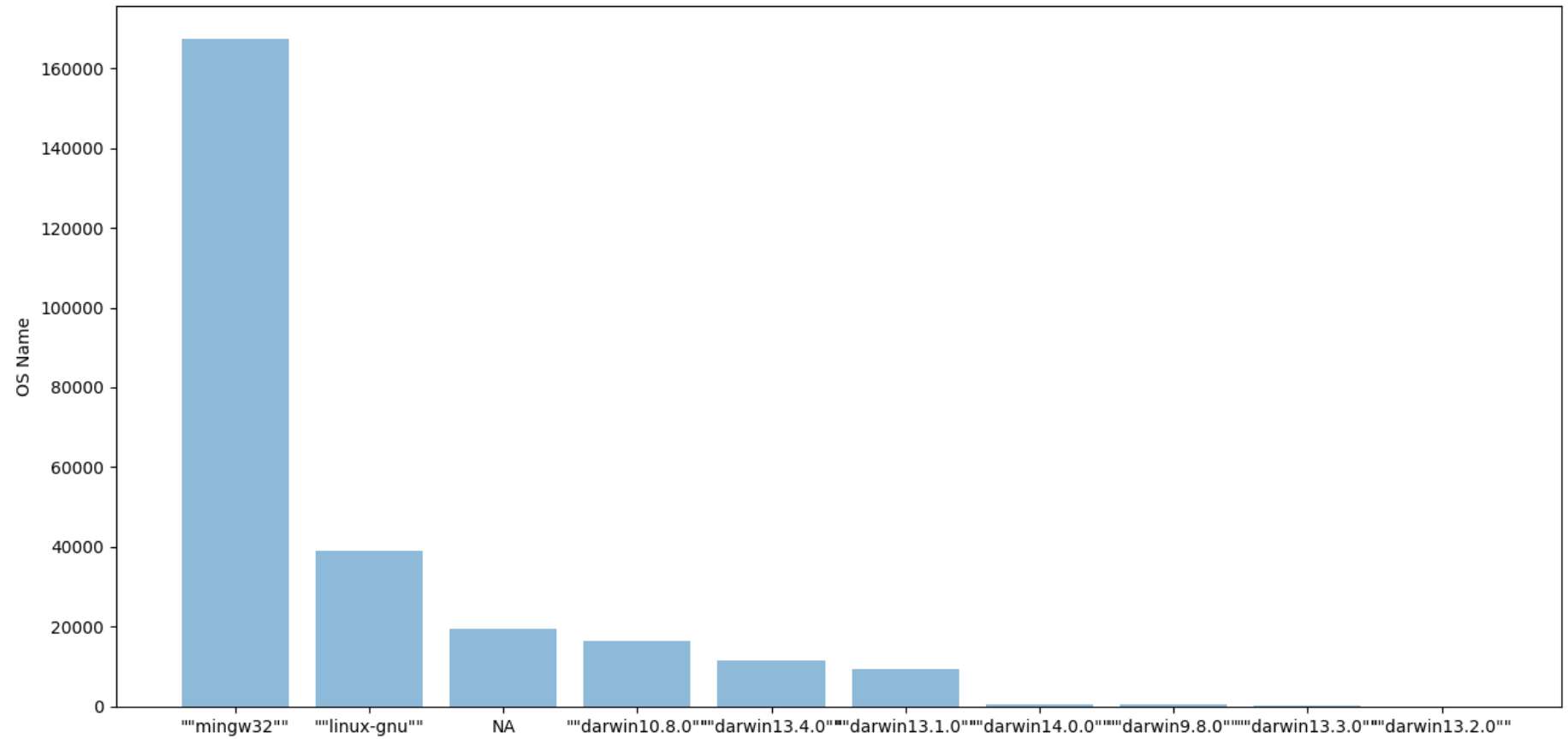
import matplotlib.pyplot as plt; plt.rcdefaults()
import numpy as np
import matplotlib.pyplot as plt

y = df210["OS_Name"]
y_pos = np.arange(len(y))
x = df210["Pkg_Count"]

plt.bar(y_pos, x, align='center', alpha=0.5)
plt.xticks(y_pos, y)
plt.ylabel('OS Name')
plt.title('Packages download by OS 2014-12-01')

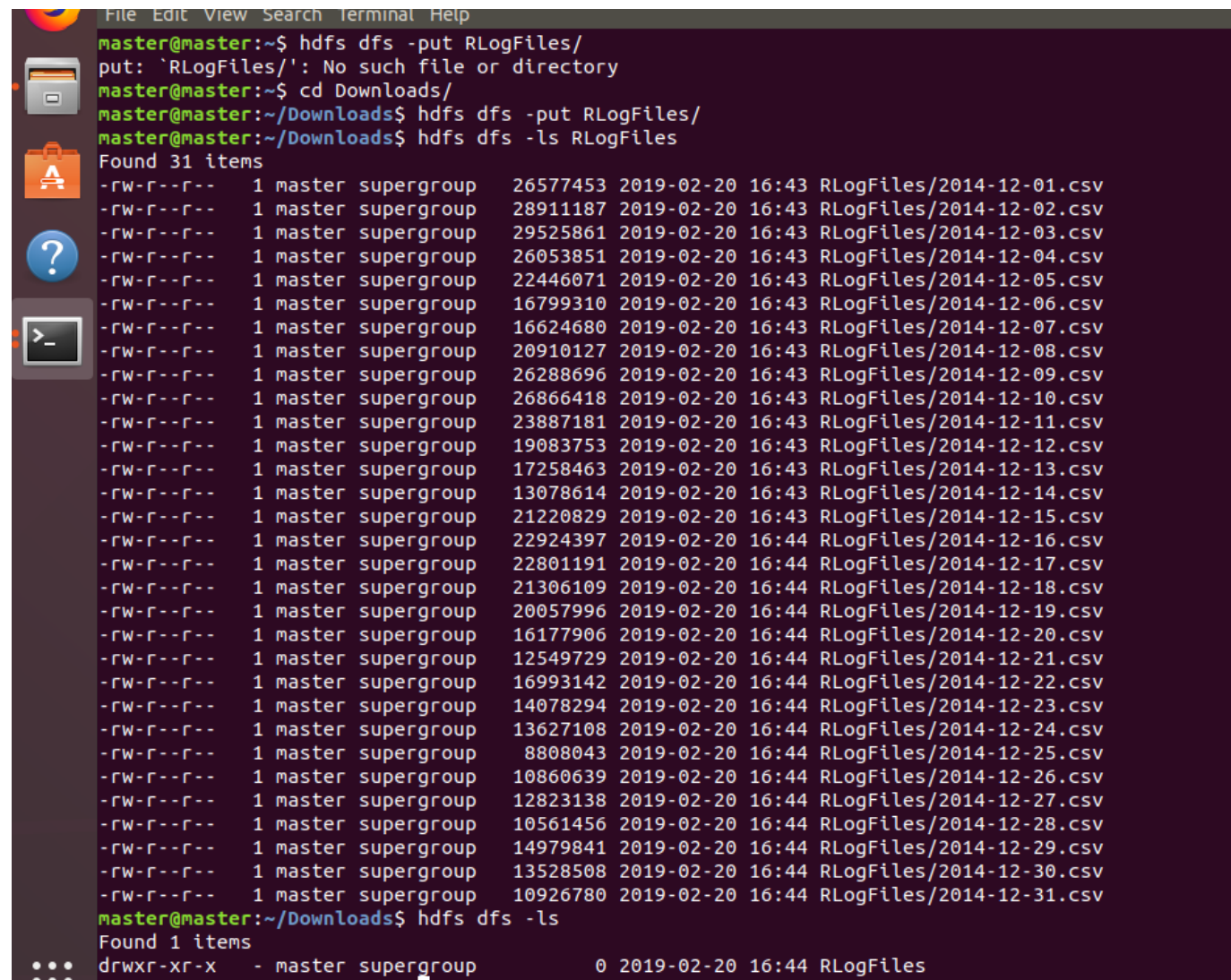
plt.show()
```

Packages download by OS 2014-12-01



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

a. Load the log-files into HDFS

A terminal window with a dark background and light text. The window title bar shows 'File Edit View Search Terminal Help'. The terminal content shows a series of commands and their outputs. The first command is 'hdfs dfs -put RLogFiles/' which fails with 'No such file or directory'. The second command is 'cd Downloads/'. The third command is 'hdfs dfs -put RLogFiles/' which succeeds. The fourth command is 'hdfs dfs -ls RLogFiles' which lists 31 CSV files. The fifth command is 'hdfs dfs -ls' which lists the 'RLogFiles' directory.

```
File Edit View Search Terminal Help
master@master:~$ hdfs dfs -put RLogFiles/
put: 'RLogFiles/': No such file or directory
master@master:~$ cd Downloads/
master@master:~/Downloads$ hdfs dfs -put RLogFiles/
master@master:~/Downloads$ hdfs dfs -ls RLogFiles
Found 31 items
-rw-r--r-- 1 master supergroup 26577453 2019-02-20 16:43 RLogFiles/2014-12-01.csv
-rw-r--r-- 1 master supergroup 28911187 2019-02-20 16:43 RLogFiles/2014-12-02.csv
-rw-r--r-- 1 master supergroup 29525861 2019-02-20 16:43 RLogFiles/2014-12-03.csv
-rw-r--r-- 1 master supergroup 26053851 2019-02-20 16:43 RLogFiles/2014-12-04.csv
-rw-r--r-- 1 master supergroup 22446071 2019-02-20 16:43 RLogFiles/2014-12-05.csv
-rw-r--r-- 1 master supergroup 16799310 2019-02-20 16:43 RLogFiles/2014-12-06.csv
-rw-r--r-- 1 master supergroup 16624680 2019-02-20 16:43 RLogFiles/2014-12-07.csv
-rw-r--r-- 1 master supergroup 20910127 2019-02-20 16:43 RLogFiles/2014-12-08.csv
-rw-r--r-- 1 master supergroup 26288696 2019-02-20 16:43 RLogFiles/2014-12-09.csv
-rw-r--r-- 1 master supergroup 26866418 2019-02-20 16:43 RLogFiles/2014-12-10.csv
-rw-r--r-- 1 master supergroup 23887181 2019-02-20 16:43 RLogFiles/2014-12-11.csv
-rw-r--r-- 1 master supergroup 19083753 2019-02-20 16:43 RLogFiles/2014-12-12.csv
-rw-r--r-- 1 master supergroup 17258463 2019-02-20 16:43 RLogFiles/2014-12-13.csv
-rw-r--r-- 1 master supergroup 13078614 2019-02-20 16:43 RLogFiles/2014-12-14.csv
-rw-r--r-- 1 master supergroup 21220829 2019-02-20 16:43 RLogFiles/2014-12-15.csv
-rw-r--r-- 1 master supergroup 22924397 2019-02-20 16:44 RLogFiles/2014-12-16.csv
-rw-r--r-- 1 master supergroup 22801191 2019-02-20 16:44 RLogFiles/2014-12-17.csv
-rw-r--r-- 1 master supergroup 21306109 2019-02-20 16:44 RLogFiles/2014-12-18.csv
-rw-r--r-- 1 master supergroup 20057996 2019-02-20 16:44 RLogFiles/2014-12-19.csv
-rw-r--r-- 1 master supergroup 16177906 2019-02-20 16:44 RLogFiles/2014-12-20.csv
-rw-r--r-- 1 master supergroup 12549729 2019-02-20 16:44 RLogFiles/2014-12-21.csv
-rw-r--r-- 1 master supergroup 16993142 2019-02-20 16:44 RLogFiles/2014-12-22.csv
-rw-r--r-- 1 master supergroup 14078294 2019-02-20 16:44 RLogFiles/2014-12-23.csv
-rw-r--r-- 1 master supergroup 13627108 2019-02-20 16:44 RLogFiles/2014-12-24.csv
-rw-r--r-- 1 master supergroup 8808043 2019-02-20 16:44 RLogFiles/2014-12-25.csv
-rw-r--r-- 1 master supergroup 10860639 2019-02-20 16:44 RLogFiles/2014-12-26.csv
-rw-r--r-- 1 master supergroup 12823138 2019-02-20 16:44 RLogFiles/2014-12-27.csv
-rw-r--r-- 1 master supergroup 10561456 2019-02-20 16:44 RLogFiles/2014-12-28.csv
-rw-r--r-- 1 master supergroup 14979841 2019-02-20 16:44 RLogFiles/2014-12-29.csv
-rw-r--r-- 1 master supergroup 13528508 2019-02-20 16:44 RLogFiles/2014-12-30.csv
-rw-r--r-- 1 master supergroup 10926780 2019-02-20 16:44 RLogFiles/2014-12-31.csv
master@master:~/Downloads$ hdfs dfs -ls
Found 1 items
drwxr-xr-x - master supergroup 0 2019-02-20 16:44 RLogFiles
```


b. Count the number of distinct users each day – PIG LATIN

```
master@master: /usr/local/pig$ pig -x mapreduce users_pday.pig
```

```
2014-12-23 17.825987
2014-12-24 17.950096
2014-12-26 15.117096
2014-12-28 15.850888
2014-12-29 16.016605
2014-12-30 14.612893
2014-12-31 15.466631
2014-12-08 12.634688
2014-12-12 13.185286
2014-12-14 12.912641
2014-12-25 17.504742
2014-12-27 24.501846
master@master:~/Downloads/AVGPKGPDAY$ cd
master@master:~$ cd /usr/local/pig
master@master: /usr/local/pig$ gedit users_pday.pig
```

```
Open users_pday.pig /usr/local/pig Save
A = LOAD 'RLogFiles' USING PigStorage(',') as (date:chararray, time:chararray, size:float,
r_version:chararray, r_arch:chararray, r_os:chararray, package:chararray, version:chararray,
country:chararray, ip_id:chararray);
B = foreach A generate date, package;
C = DISTINCT(foreach A generate date, ip_id);
D = group B by date;
E = group C by date;
F = foreach D generate group, COUNT(B.package);
G = foreach E generate group, COUNT(C.ip_id);
store F into 'PKCOUNT';
store G into 'USERCOUNT';
```

```
Success!

Job Stats (time in seconds):
JobId  Maps  Reduces  MaxMapTime  MinMapTime  AvgMapTime  MedianMapTime  MaxReduceTime  MinReduceTime  AvgReduceTime  MedianReducetime  Alias  F
eature  Outputs
job_local1087782189_0001  5  1  n/a  n/a  n/a  n/a  n/a  n/a  n/a  A,B,D,F MULTI_QUERY,COMBINER  hdfs://localhost:8020/user/master/PKCOUN
T,
job_local191584641_0003  1  1  n/a  n/a  n/a  n/a  n/a  n/a  n/a  E,G GROUP_BY,COMBINER  hdfs://localhost:8020/user/maste
r/USERCOUNT,
job_local2015708803_0002  2  1  n/a  n/a  n/a  n/a  n/a  n/a  n/a  1-1,A DISTINCT

Input(s):
Successfully read 5984786 records (2528096082 bytes) from: "hdfs://localhost:8020/user/master/RLogFiles"

Output(s):
Successfully stored 32 records (603 bytes) in: "hdfs://localhost:8020/user/master/PKCOUN
T"
Successfully stored 32 records (531495844 bytes) in: "hdfs://localhost:8020/user/master/USERCOUNT"

Counters:
Total records written : 64
Total bytes written : 531496447
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1087782189_0001 -> job_local2015708803_0002,
job_local2015708803_0002 -> job_local191584641_0003,
job_local191584641_0003
```

```
master@master:~/Downloads/PKCOUNT$ cd ..
master@master:~/Downloads$ cd USERCOUNT/
master@master:~/Downloads/USERCOUNT$ cat part-r-00000
"date" 1
"2014-12-01" 20321
"2014-12-02" 22290
"2014-12-03" 21735
"2014-12-04" 20977
"2014-12-05" 17747
"2014-12-06" 11112
"2014-12-07" 12475
"2014-12-09" 20980
"2014-12-10" 19915
"2014-12-11" 19845
"2014-12-13" 10574
"2014-12-15" 17575
"2014-12-16" 17233
"2014-12-17" 16483
"2014-12-18" 15601
"2014-12-19" 13648
"2014-12-20" 7733
"2014-12-21" 8730
"2014-12-22" 12096
"2014-12-23" 11013
"2014-12-24" 7755
"2014-12-26" 7259
"2014-12-28" 6921
"2014-12-29" 9635
"2014-12-30" 9478
"2014-12-31" 7372
"2014-12-08" 19430
"2014-12-12" 16990
"2014-12-14" 12088
"2014-12-25" 5906
"2014-12-27" 6502
```

b. Count the number of distinct users each day – HIVE

```
hive> CREATE DATABASE rlogs;
OK
Time taken: 1.776 seconds
hive> use rlogs;
OK
Time taken: 0.051 seconds
hive> CREATE TABLE RLogFiles (datum STRING, time STRING, size FLOAT, r_version STRING, r_arch STRING, r_os STRING, package STRING, country STRING, ip_id STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 1.206 seconds
hive> CREATE TABLE userpday (datum STRING, userspday STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
OK
Time taken: 0.594 seconds
hive> CREATE TABLE pkgpday (datum STRING, pkgspday STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
OK
Time taken: 0.567 seconds
hive> LOAD DATA INPATH 'RLogFiles/*' INTO TABLE RLogFiles;
Loading data to table rlogs.rlogfiles
Table rlogs.rlogfiles stats: [numFiles=31, totalSize=578536771]
OK
Time taken: 3.57 seconds
hive> SELECT COUNT(*) FROM RLogFiles;
Time taken: 21.055 seconds, Fetched: 1 row(s)
hive> INSERT OVERWRITE TABLE userpday SELECT datum, COUNT(datum) FROM (SELECT datum, ip_id FROM RLogFiles GROUP BY datum, ip_id) dt GROUP BY datum;
```

Ended Job = job_local1125397010_0001

MapReduce Jobs Launched:

Stage-Stage-1: HDFS Read: 2096 HDFS Write: 0 SUCCESS

Total MapReduce CPU Time Spent: 0 msec

OK

2014-12-01	20321
2014-12-02	22290
2014-12-03	21735
2014-12-04	20977
2014-12-05	17747
2014-12-06	11112
2014-12-07	12475
2014-12-08	19430
2014-12-09	20980
2014-12-10	19915
2014-12-11	19845
2014-12-12	16990
2014-12-13	10574
2014-12-14	12088
2014-12-15	17575
2014-12-16	17233
2014-12-17	16483
2014-12-18	15601
2014-12-19	13648
2014-12-20	7733
2014-12-21	8730
2014-12-22	12096
2014-12-23	11013
2014-12-24	7755
2014-12-25	5906
2014-12-26	7259
2014-12-27	6502
2014-12-28	6921
2014-12-29	9635
2014-12-30	9478
2014-12-31	7372

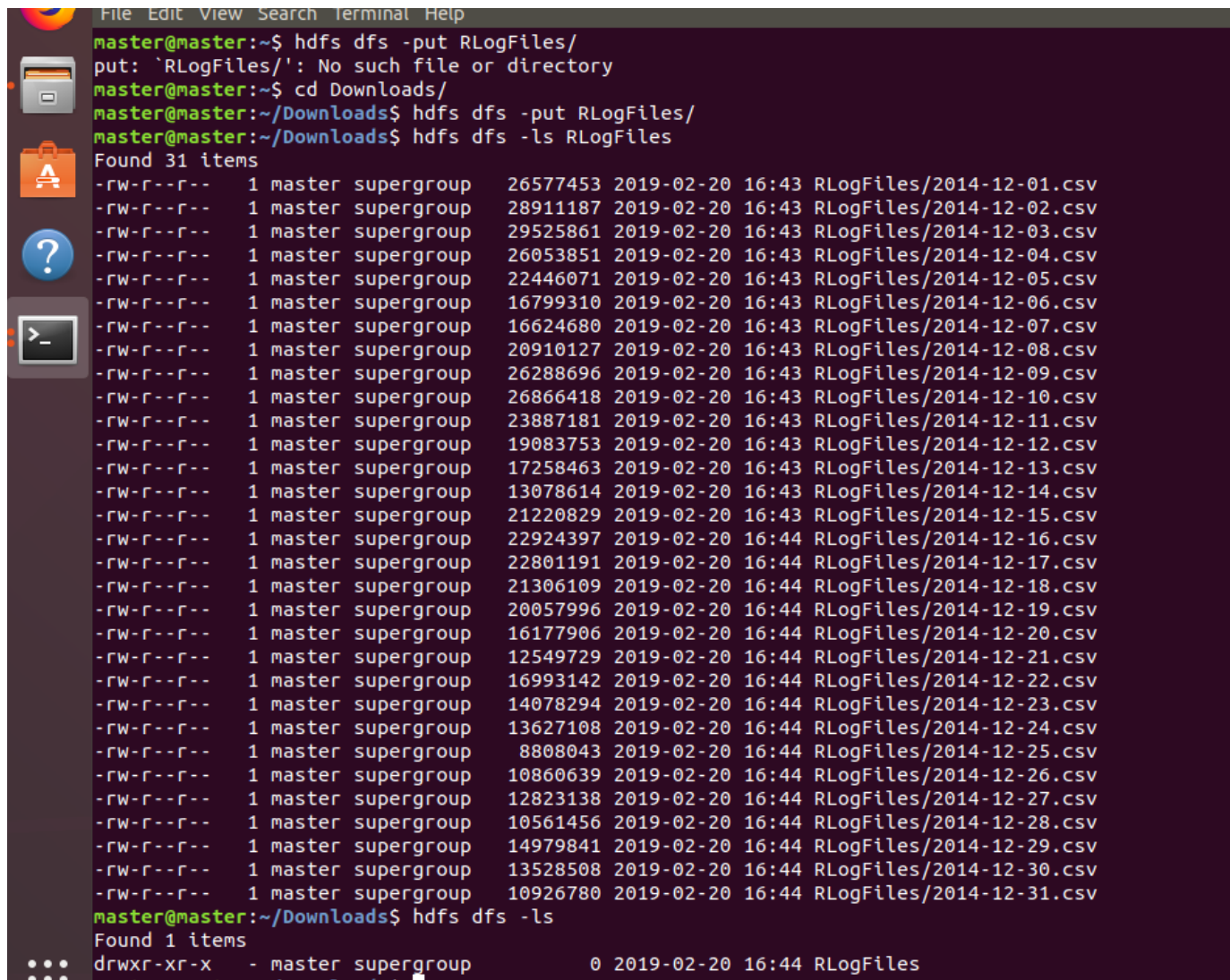
date 1

Time taken: 3.861 seconds, Fetched: 32 row(s)

hive> select * from userpday order by datum;

5. Pig Latin and HIVE: Average Number of packages downloaded by an individual user each day.

a. load the log files into HDFS

A terminal window with a dark background and light-colored text. The window has a menu bar at the top with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. On the left side, there is a vertical toolbar with icons for file operations. The terminal shows a series of commands and their outputs. The first command is 'hdfs dfs -put RLogFiles/' which results in an error: 'put: 'RLogFiles/': No such file or directory'. The second command is 'cd Downloads/'. The third command is 'hdfs dfs -put RLogFiles/' which also results in an error. The fourth command is 'hdfs dfs -ls RLogFiles', which lists 31 items. Each item is a CSV file named 'RLogFiles/2014-12-01.csv' through 'RLogFiles/2014-12-31.csv'. The listing shows permissions, file size, modification time, owner, and group for each file. The final command is 'hdfs dfs -ls', which lists 1 item: 'RLogFiles'.

b. Calculate the number average number of packages downloaded by an individual user each day – PIG LATIN.

```
master@master: /usr/local/pig$ pig -x mapreduce pkgs_pday.pig
```

```
"2014-12-08" 12.634688
"2014-12-12" 13.185286
"2014-12-14" 12.912641
"2014-12-25" 17.504742
"2014-12-27" 24.501846
```

```
master@master:~/Downloads/AVGPKGPDAY$ cd
master@master:~$ cd /usr/local/pig
master@master: /usr/local/pig$ gedit users_pday.pig
master@master: /usr/local/pig$ gedit pkgs_pday.pig
```

```
pkgs_pday.pig
/usr/local/pig
Q = LOAD 'PKCOUNT/part-r-00000' USING PigStorage('\t') as (date:chararray, pkcount:float);
R = LOAD 'USERCOUNT/part-r-00000' USING PigStorage('\t') as (date:chararray, usercount:float);
S = JOIN Q by date, R by date;
T = foreach S GENERATE $0, (float)($1/$3);
store T into 'AVGPKGPDAY';
```

Success!

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReducetime	Alias	Feature
job_local32292517_0001	2	1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Q,R,S,T HASH_JOIN	hdfs://localhost:8020/user/master/AVGPKGPDAY,

Input(s):

Successfully read 32 records from: "hdfs://localhost:8020/user/master/PKCOUNTPart-r-00000"

Successfully read 32 records from: "hdfs://localhost:8020/user/master/USERCOUNT/part-r-00000"

Output(s):

Successfully stored 32 records (17268041 bytes) in: "hdfs://localhost:8020/user/master/AVGPKGPDAY"

Counters:

Total records written : 32

Total bytes written : 17268041

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job DAG:

job_local32292517_0001

```
master@master:~/Downloads$ hdfs dfs -get AVGPKGPDAY
19/02/20 19:56:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for optimization
master@master:~/Downloads$ cd AVGPKGPDAY/
master@master:~/Downloads/AVGPKGPDAY$ cat part-r-00000
"date" 5.0
"2014-12-01" 13.040254
"2014-12-02" 12.958323
"2014-12-03" 13.667817
"2014-12-04" 12.339276
"2014-12-05" 12.481546
"2014-12-06" 15.835313
"2014-12-07" 13.670781
"2014-12-09" 12.477836
"2014-12-10" 13.439769
"2014-12-11" 12.065961
"2014-12-13" 17.40836
"2014-12-15" 12.039488
"2014-12-16" 13.23751
"2014-12-17" 13.829218
"2014-12-18" 13.721621
"2014-12-19" 14.75718
"2014-12-20" 22.433855
"2014-12-21" 14.747538
"2014-12-22" 14.171544
"2014-12-23" 13.029964
"2014-12-24" 17.950096
"2014-12-26" 15.117096
"2014-12-28" 15.850888
"2014-12-29" 16.016605
"2014-12-30" 14.612893
"2014-12-31" 15.466631
"2014-12-08" 12.634688
"2014-12-12" 13.185286
"2014-12-14" 12.912641
"2014-12-25" 17.504742
"2014-12-27" 24.501846
master@master:~/Downloads/AVGPKGPDAY$
```


HIVE:

```
2014-12-01 00:00:38 75732.0 3.1.1 x86_64 mingw32 stringr 0.6.2 US 6
2014-12-01 00:00:47 1289894.0 3.1.1 x86_64 mingw32 koRpus 0.05-5 US 6
Time taken: 0.138 seconds, Fetched: 10 row(s)
hive> INSERT OVERWRITE TABLE userpday SELECT datum, COUNT(datum) FROM (SELECT datum, ip_id FROM RLogFiles GROUP BY datum, ip_id) dt GROUP BY datum;
Query ID = master_20190221184736_093d833f-ec16-45a7-8119-dadbc5131ca7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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>
Job running in-process (local Hadoop)
2019-02-21 18:47:37,823 Stage-1 map = 0%, reduce = 0%
2019-02-21 18:47:47,090 Stage-1 map = 100%, reduce = 0%
2019-02-21 18:47:53,126 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local291689039_0005
Loading data to table rlogs.userpday
Table rlogs.userpday stats: [numFiles=2, numRows=32, totalSize=524, rawDataSize=492]
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 10146765462 HDFS Write: 7138 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 17.4 seconds
hive> INSERT OVERWRITE TABLE pkgpday SELECT datum, COUNT(datum) FROM (SELECT datum, ip_id, package FROM RLogFiles) dt2 GROUP BY datum;
Query ID = master_20190221184819_56d13860-975c-41a5-b240-c31e78729b61
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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: 0 msec

OK

2014-12-02	12.958322117541499
2014-12-04	12.339276350288412
2014-12-06	15.835313174946004
2014-12-08	12.634688625836336
2014-12-11	12.065961199294533
2014-12-13	17.408360128617364
2014-12-15	12.039487908961593
2014-12-17	13.829217982163442
2014-12-19	14.757180539273154
2014-12-20	22.433854907539118
2014-12-22	14.171544312169312
2014-12-24	17.95009671179884
2014-12-26	15.117096018735364
2014-12-28	15.850888599913308
2014-12-31	15.466630493760174

date 5.0

2014-12-01	13.040253924511589
2014-12-03	13.66781688520819
2014-12-05	12.481546176818616
2014-12-07	13.670781563126253
2014-12-09	12.477836034318399
2014-12-10	13.439769018327894
2014-12-12	13.185285462036491
2014-12-14	12.912640635340834
2014-12-16	13.237509429582778
2014-12-18	13.721620408948144
2014-12-21	14.747537227949598
2014-12-23	13.029964587305912
2014-12-25	17.50474094141551
2014-12-27	24.50184558597355
2014-12-29	16.016606123508044
2014-12-30	14.612893015404094

Time taken: 10.32 seconds, Fetched: 32 row(s)

hive> SELECT userpday.datum, (pkgpday.pkgspday/userpday.userspday) AS avgpday FROM userpday JOIN pkgpday on (userpday.datum = pkgpday.datum);

6. Pig Latin and HIVE: Task Views.

- a. Task Views are collections of R packages of a certain topic (check the CRAN webpage)
- b. We are interested if these Task Views are used by R users: count the number of package ctv downloaded each day)

PIG:

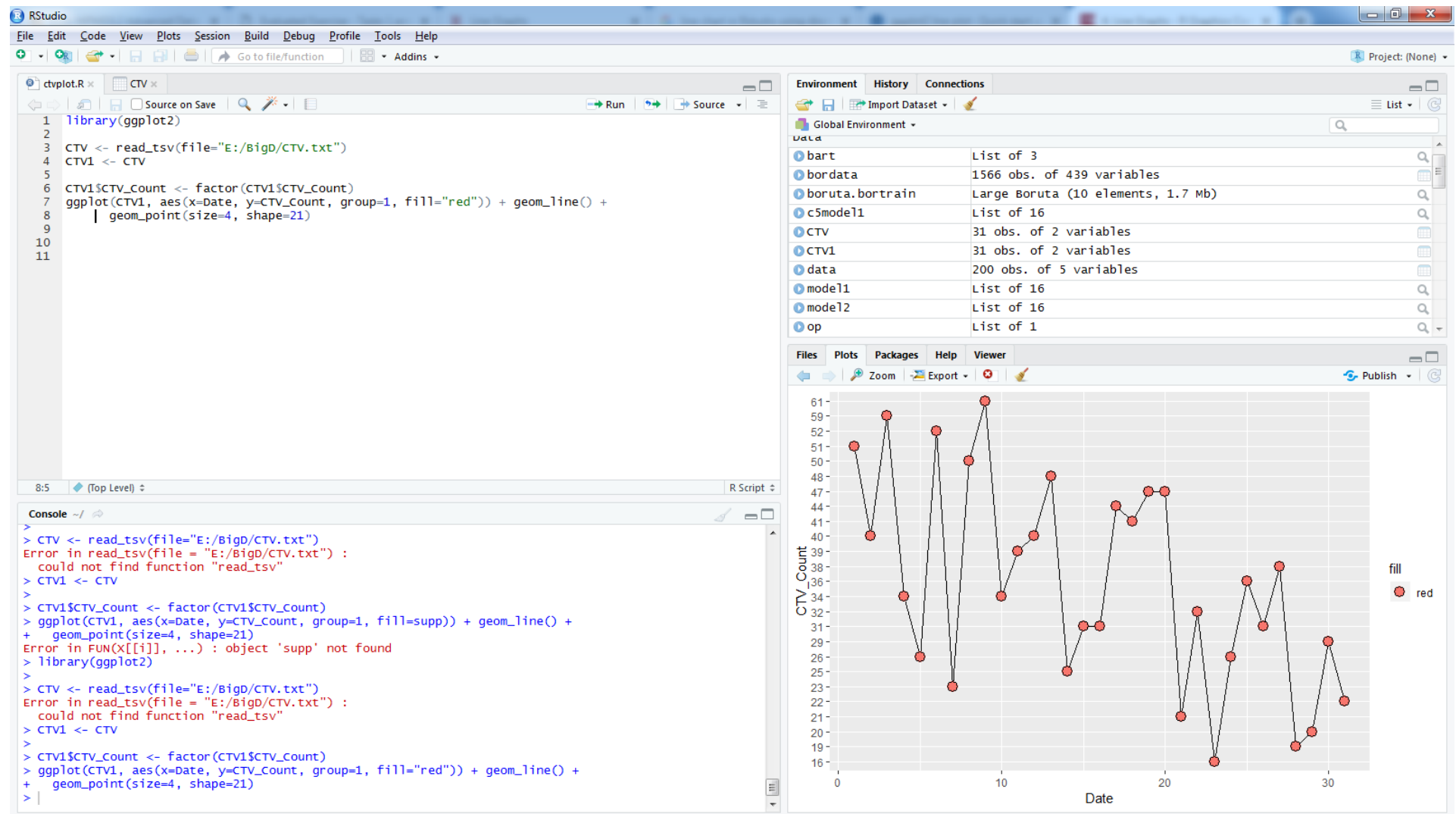
```
2019-02-22 14:39:43,986 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> A = LOAD 'RLogFiles' USING PigStorage(',') as (date:chararray, time:chararray, size:float, r_version:chararray, r_arch:chararray, r_os:chararray, package:chararray, version:chararray, country:chararray, ip_id:chararray);
grunt> B = FILTER A BY package == 'ctv';
grunt> C = foreach B generate date, package;
grunt> D = group C by date;
grunt> E = foreach D generate group, COUNT(C.package);
grunt> F = ORDER E BY $0;
grunt> STORE F INTO 'ctv';
2019-02-22 14:42:42,462 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY ORDER BY FILTER
```

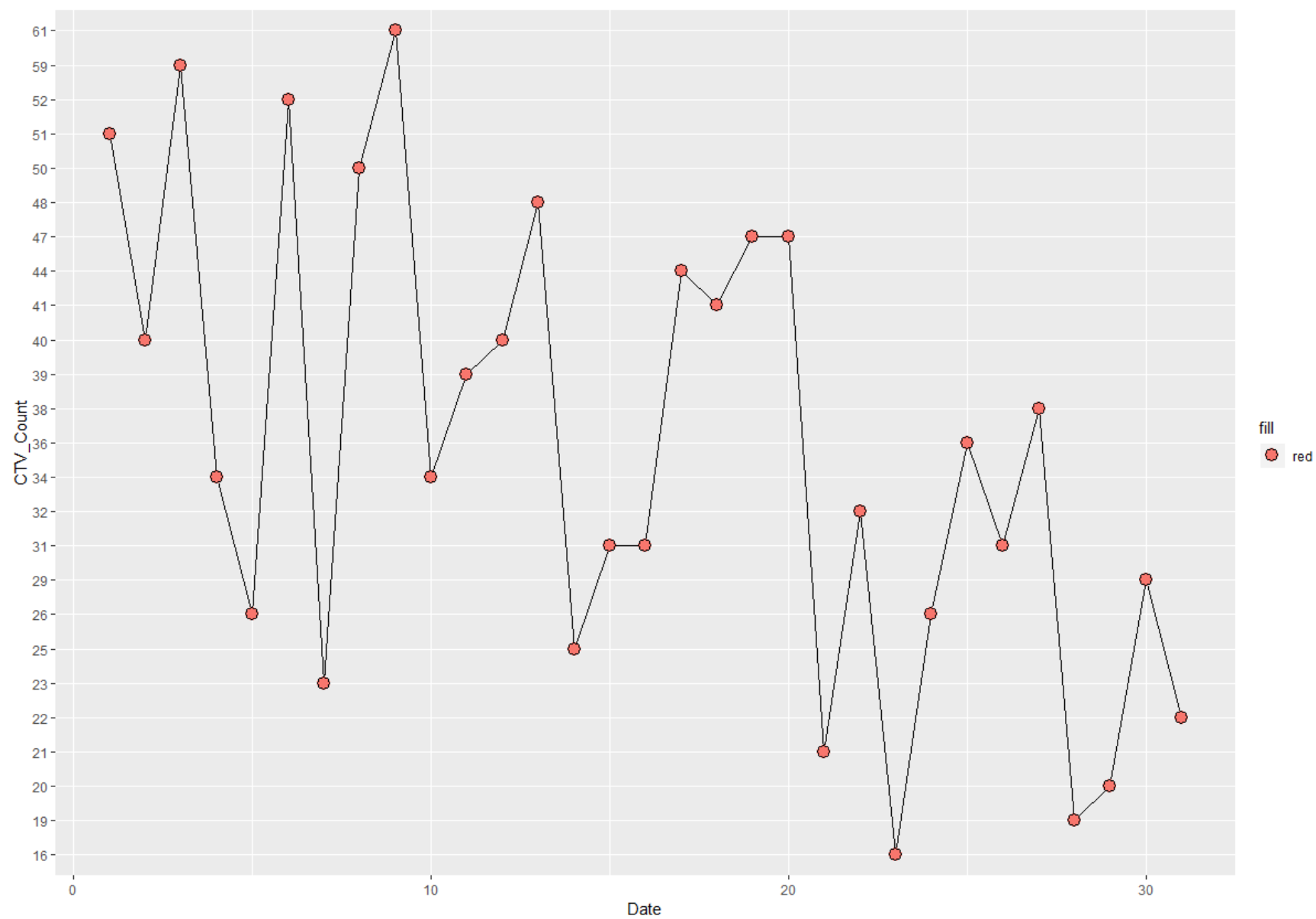
```
master@master:~/Downloads$ hdfs dfs -get ctv
19/02/22 14:43:39 WARN util.NativeCodeLoader: Unable
master@master:~/Downloads$ cd ctv
master@master:~/Downloads/ctv$ cat part-r-00000
2014-12-01      51
2014-12-02      40
2014-12-03      59
2014-12-04      34
2014-12-05      26
2014-12-06      52
2014-12-07      23
2014-12-08      50
2014-12-09      61
2014-12-10      34
2014-12-11      39
2014-12-12      40
2014-12-13      48
2014-12-14      25
2014-12-15      31
2014-12-16      31
2014-12-17      44
2014-12-18      41
2014-12-19      47
2014-12-20      47
2014-12-21      21
2014-12-22      32
2014-12-23      16
2014-12-24      26
2014-12-25      36
2014-12-26      31
2014-12-27      38
2014-12-28      19
2014-12-29      20
2014-12-30      29
2014-12-31      22
master@master:~/Downloads/ctv$
```

HIVE:

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/usr/local/hive/ctv' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' SELECT datum, COUNT(datum) FROM (SELECT datum, ip_id, packag
e FROM RLogFiles WHERE package = 'ctv') dt3 GROUP BY datum ORDER BY datum;
FAILED: SemanticException [Error 10001]: Line 1:169 Table not found 'RLogFiles'
hive> INSERT OVERWRITE LOCAL DIRECTORY '/usr/local/hive/ctv' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' SELECT datum, COUNT(datum) FROM (SELECT datum, ip_id, packag
e FROM rlogs.RLogFiles WHERE package = 'ctv') dt3 GROUP BY datum ORDER BY datum;
Query ID = master_20190221192239_7c6d7f8c-9992-47b0-90dc-2dceab5fa2cc
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
```

```
-rw-r--r-- 1 master master 28 Feb 21 15:27 user.txt
master@master:/usr/local/hive$ cat ctv.txt
master@master:/usr/local/hive$ cd ctv/
master@master:/usr/local/hive/ctv$ ls -l
total 4
-rw-r--r-- 1 master master 434 Feb 21 19:22 000000_0
master@master:/usr/local/hive/ctv$ cat 000000_0
2014-12-01      51
2014-12-02      40
2014-12-03      59
2014-12-04      34
2014-12-05      26
2014-12-06      52
2014-12-07      23
2014-12-08      50
2014-12-09      61
2014-12-10      34
2014-12-11      39
2014-12-12      40
2014-12-13      48
2014-12-14      25
2014-12-15      31
2014-12-16      31
2014-12-17      44
2014-12-18      41
2014-12-19      47
2014-12-20      47
2014-12-21      21
2014-12-22      32
2014-12-23      16
2014-12-24      26
2014-12-25      36
2014-12-26      31
2014-12-27      38
2014-12-28      19
2014-12-29      20
2014-12-30      29
2014-12-31      22
master@master:/usr/local/hive/ctv$
```

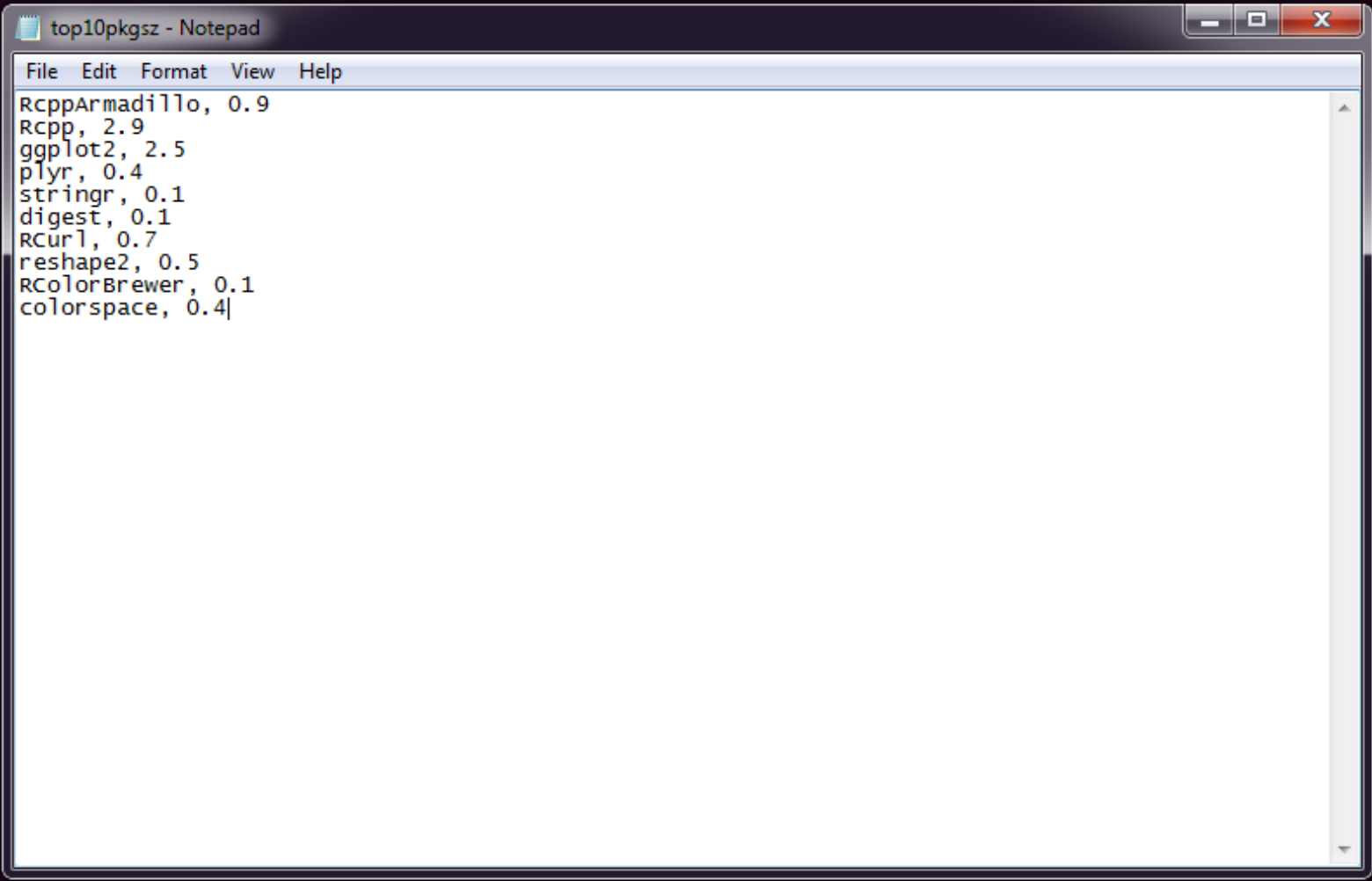




7. Pig Latin and HIVE: Download volume (in MB) of Top-10-packages

- a. Use CRAN to find out the package size of the Top-10-packages (use WindowsPackage file size) in MB. Round to 1 decimal place.
- 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)
- e. Filter out the Top-10-packages in Pig
- f. Add the size information

g. Calculate the download volume of each of the 10 packages by day



A screenshot of a Notepad window titled "top10pkgasz - Notepad". The window has a menu bar with "File", "Edit", "Format", "View", and "Help". The text inside the window lists 10 R packages and their versions, each on a new line:

```
RcppArmadillo, 0.9  
Rcpp, 2.9  
ggplot2, 2.5  
plyr, 0.4  
stringr, 0.1  
digest, 0.1  
RCurl, 0.7  
reshape2, 0.5  
RColorBrewer, 0.1  
colorspace, 0.4|
```

```

19/02/21 20:13:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...
Found 8 items
drwxr-xr-x - master supergroup 0 2019-02-20 19:56 AVGPKGPDAY
drwxr-xr-x - master supergroup 0 2019-02-20 19:51 PKCOUNT
drwxr-xr-x - master supergroup 0 2019-02-21 18:46 RLogFiles
drwxr-xr-x - master supergroup 0 2019-02-20 19:52 USERCOUNT
drwxr-xr-x - master supergroup 0 2019-02-20 17:12 package_count_by_os
drwxr-xr-x - master supergroup 0 2019-02-20 16:56 pkcount_dec1_2014
drwxr-xr-x - master supergroup 0 2019-02-21 13:27 rlogs
-rw-r--r-- 1 master supergroup 144 2019-02-21 20:13 top10pkgsz.txt
master@master:~/Downloads$

```

PIG:

```

2019-02-23 10:06:27,059 [main] WARN org.apache.pig.PigServer - AFS is disabled since yarn.timeline-service.enabled set to false
grunt> A = LOAD 'RLogFiles' USING PigStorage(',') as (date:chararray, time:chararray, size:float, r_version:chararray, r_arch:chararray, r_os:chararray, package:chararray, version:chararray, country:chararray, ip_id:chararray);
grunt> B = FILTER A BY (package == 'RcppArmadillo' OR package == 'Rcpp' OR package == 'ggplot2' OR package == 'plyr' OR package == 'stringr' OR package == 'digest' OR package == 'RCurl' OR package == 'reshape2' OR package == 'RColorBrewer' OR package == 'colorspace');
grunt> C = foreach B generate date, package;
grunt> D = LOAD 'top10pkgsz.txt' USING PigStorage(',') as (pkgname:chararray, size:float);
grunt> E = group C by (date,package);
grunt> F = FOREACH E generate FLATTEN(group) AS (data,package), COUNT(C);
grunt> G = JOIN F by package, D by pkgname;
grunt> H = foreach G generate $0 AS date, $1 AS pkgname, (float)($2 * $4) AS szvol;
2019-02-23 10:09:45,038 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
grunt> STORE H INTO 'szvolpg';
2019-02-23 10:10:21,507 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
2019-02-23 10:10:21,582 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).

```

```
master@master:~/Downloads$ cd szvolpg/
master@master:~/Downloads/szvolpg$ cat part-r-00000
2014-12-01      Rcpp      9932.5
2014-12-03      Rcpp     12026.301
2014-12-19      Rcpp     7296.4004
2014-12-17      Rcpp     9053.801
2014-12-18      Rcpp     8204.101
2014-12-30      Rcpp     4773.4004
2014-12-04      Rcpp     11069.301
2014-12-29      Rcpp      5402.7
2014-12-05      Rcpp     10431.301
2014-12-02      Rcpp     11663.801
2014-12-28      Rcpp     3213.2002
2014-12-06      Rcpp     5875.4004
2014-12-07      Rcpp      6075.5
2014-12-27      Rcpp      3625.0
2014-12-08      Rcpp     9929.601
2014-12-26      Rcpp     3662.7002
2014-12-09      Rcpp     11272.301
2014-12-25      Rcpp      3045.0
2014-12-10      Rcpp     10190.601
2014-12-24      Rcpp     3946.9001
2014-12-11      Rcpp     10579.2
2014-12-12      Rcpp      8665.2
2014-12-31      Rcpp      3630.8
2014-12-23      Rcpp      5727.5
2014-12-13      Rcpp     5309.9004
2014-12-22      Rcpp     6817.9004
2014-12-14      Rcpp      5736.2
2014-12-21      Rcpp      4178.9
2014-12-15      Rcpp      8920.4
2014-12-20      Rcpp      3970.1
2014-12-16      Rcpp      8926.2
2014-12-06      plyr       760.0
2014-12-09      plyr      1400.8
2014-12-08      plyr      1279.6
2014-12-17      plyr      1116.4
2014-12-13      plyr       646.8
2014-12-30      plyr       608.4
2014-12-25      plyr       427.6
```

HIVE:

```
Time taken: 1.28 seconds
hive> CREATE TABLE top10pkgsz (pkname STRING, pkgsz FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 0.206 seconds
hive> LOAD DATA INPATH 'top10pkgsz.txt' INTO TABLE top10pkgsz;
Loading data to table rlogs.top10pkgsz
Table rlogs.top10pkgsz stats: [numFiles=1, totalSize=144]
OK
Time taken: 0.962 seconds
hive> SELECT * FROM top10pkgsz;
OK
RcppArmadillo    0.9
Rcpp             2.9
ggplot2          2.5
plyr             0.4
stringr          0.1
digest           0.1
RCurl            0.7
reshape2         0.5
RColorBrewer     0.1
colorspace       0.4
Time taken: 0.165 seconds, Fetched: 10 row(s)
hive> █
```

```

Time taken: 0.437 seconds
hive> CREATE TABLE top10pkgsvol (datum STRING, pkgname STRING, pkcount DOUBLE) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
OK
Time taken: 0.141 seconds
hive> INSERT OVERWRITE TABLE top10pkgsvol SELECT datum, package, COUNT(package) FROM (SELECT datum, package FROM RLogFiles WHERE package in ('RcppArmadillo', 'Rcpp', 'ggplot2', 'plyr', 'stringr', 'digest', 'RCurl',
reshape2', 'RColorBrewer', 'colorspace')) dt4 GROUP BY datum, package;
Query ID = master_20190221203800_3330bdb3-9b72-4797-b894-5b82cf97b1c2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 2
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>
Job running in-process (local Hadoop)
2019-02-21 20:38:01,589 Stage-1 map = 0%,   reduce = 0%
2019-02-21 20:38:06,609 Stage-1 map = 100%,   reduce = 0%
2019-02-21 20:38:10,658 Stage-1 map = 100%,   reduce = 100%
Ended Job = job_local1510728303_0008
Loading data to table rlogs.top10pkgsvol
Table rlogs.top10pkgsvol stats: [numFiles=2, numRows=310, totalSize=8214, rawDataSize=7904]
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 13294793394 HDFS Write: 39070 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 11.181 seconds
hive> select * from top10pkgsvol limit 100;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'top10pkgsvol'
hive> select * from top10pkgsvol limit 100;
OK

```

```

Time taken: 9.8 seconds, Fetched: 310 row(s)
hive> INSERT OVERWRITE LOCAL DIRECTORY '/usr/local/hive/pkgsvol' ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' SELECT top10pkgsvol.datum, top10pkgsvol.pkgname, (top
10pkgsvol.pkcount * top10pkgsvol.pkgsz) AS pkgvolpday FROM top10pkgsvol JOIN top10pkgsvol on (top10pkgsvol.pkgname = top10pkgsvol.pkgname);
Query ID = master_20190221205200_ba6f53c9-a65e-41b9-ae41-58698d88b843
Total jobs = 1
19/02/21 20:52:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Execution log at: /tmp/master/master_20190221205200_ba6f53c9-a65e-41b9-ae41-58698d88b843_log

```

```

master@master:/usr/local/hive$ cd pkgsvol/
master@master:/usr/local/hive/pkgsvol$ ls -l
total 12
-rw-r--r-- 1 master master 11057 Feb 21 20:52 000000_0
master@master:/usr/local/hive/pkgsvol$ cat 000000_0
2014-12-01      Rcpp      9932.500326633453
2014-12-01      colorspace      876.8000130653381
2014-12-01      ggplot2 7610.0
2014-12-01      plyr      1188.8000177145004
2014-12-01      stringr 296.10000441223383
2014-12-02      RColorBrewer      283.5000042244792
2014-12-02      RCurl      2209.199962377548
2014-12-02      RcppArmadillo      7011.899814248085
2014-12-02      digest      343.600005120039
2014-12-02      reshape2      1465.0
2014-12-03      Rcpp      12026.300395488739
2014-12-03      colorspace      996.400014847517
2014-12-03      ggplot2 8312.5
2014-12-03      plyr      1336.4000199139118
2014-12-03      stringr 352.3000052496791
2014-12-04      RColorBrewer      267.5000039860606
2014-12-04      RCurl      1719.8999707102776
2014-12-04      RcppArmadillo      6832.799818992615
2014-12-04      digest      285.5000042542815
2014-12-04      reshape2      1317.0
2014-12-05      Rcpp      10431.300343036652
2014-12-05      colorspace      863.2000128626823
2014-12-05      ggplot2 7332.5
2014-12-05      plyr      1216.800018131733
2014-12-05      stringr 280.9000041857362
2014-12-06      RColorBrewer      144.90000215917826
2014-12-06      RCurl      1803.899969279766
2014-12-06      RcppArmadillo      6620.399824619293
2014-12-06      digest      167.70000249892473
2014-12-06      reshape2      889.5
2014-12-07      Rcpp      6075.500199794769
2014-12-07      colorspace      501.200007468462
2014-12-07      ggplot2 4270.0

```

