Question 1

In [13]:
```
from pyspark.streaming import StreamingContext
from pyspark.sql import Row
import time
```

In [14]:
```
def save(time, rdd):
    try:
        df = spark.createDataFrame(rdd.map(\
        lambda row: Row(time=time, package=row[0], count=row[1])))
        df.show(5)
        df.write.format("org.apache.spark.sql.cassandra")\
        .options(table="cran_counts", keyspace="streamingbdm")\
        .save(mode="append")
    except:
        pass
```

In [15]:
```
ssc = StreamingContext(sc, 3)
```

In [*]:
```
cran_data = ssc.textFileStream("file:///home/victorbdm/assignStreaming/")
cran_data = cran_data.map(lambda x: x.split(','))

def remove_quotation(x):
    return([xx.replace('"', '') for xx in x])
cran_data = cran_data.map(remove_quotation)

package_download_count = cran_data.map(lambda package:(package[6], 1))
package_download_count = package_download_count.reduceByKey(lambda a,b: a+b)
package_download_count.pprint()
package_download_count.foreachRDD(save)
ssc.start()
time.sleep(120)
ssc.stop(stopSparkContext=False)
```

victorbdm@bdm: ~

```
cqlsh:streamingbdm> drop table word_counts;
cqlsh:streamingbdm> cls
cqlsh:streamingbdm> CREATE TABLE cran_counts( time text, package
text, count int, PRIMARY KEY(time, package));
cqlsh:streamingbdm> select * from cran_counts;

 time | package | count
------+---------+-------

(0 rows)
cqlsh:streamingbdm> select * from cran_counts;

 time                     | package                | count
--------------------------+------------------------+-------
 2021-12-21 23:02:00+0000 |              AutoDeskR |     1
 2021-12-21 23:02:00+0000 |                     BH |     1
 2021-12-21 23:02:00+0000 |                    C50 |     1
 2021-12-21 23:02:00+0000 |                    DBI |     3
 2021-12-21 23:02:00+0000 |                     DT |     1
 2021-12-21 23:02:00+0000 |              DescTools |     1
 2021-12-21 23:02:00+0000 |               EnvStats |     1
 2021-12-21 23:02:00+0000 |                  Exact |     1
 2021-12-21 23:02:00+0000 |                    FNN |     1
 2021-12-21 23:02:00+0000 |                Formula |     2
 2021-12-21 23:02:00+0000 |              GADMTools |     1
 2021-12-21 23:02:00+0000 |                 GGally |     1
 2021-12-21 23:02:00+0000 |            GPArotation |     2
 2021-12-21 23:02:00+0000 |                  GPfit |     1
 2021-12-21 23:02:00+0000 |                  Hmisc |     8
 2021-12-21 23:02:00+0000 |              JSconsole |     1
 2021-12-21 23:02:00+0000 |                 KMsurv |     1
 2021-12-21 23:02:00+0000 |               MLEcens |     1
 2021-12-21 23:02:00+0000 |           MatrixModels |     3
 2021-12-21 23:02:00+0000 |           ModelMetrics |     4
 2021-12-21 23:02:00+0000 |                 NlinTS |     1
 2021-12-21 23:02:00+0000 |                     R6 |    10
 2021-12-21 23:02:00+0000 |           RColorBrewer |     3
 2021-12-21 23:02:00+0000 |                  RCurl |     1
 2021-12-21 23:02:00+0000 |                  RMySQL |     1
 2021-12-21 23:02:00+0000 |                 RSQLite |     1
 2021-12-21 23:02:00+0000 |                  Rcgmin |     1
```

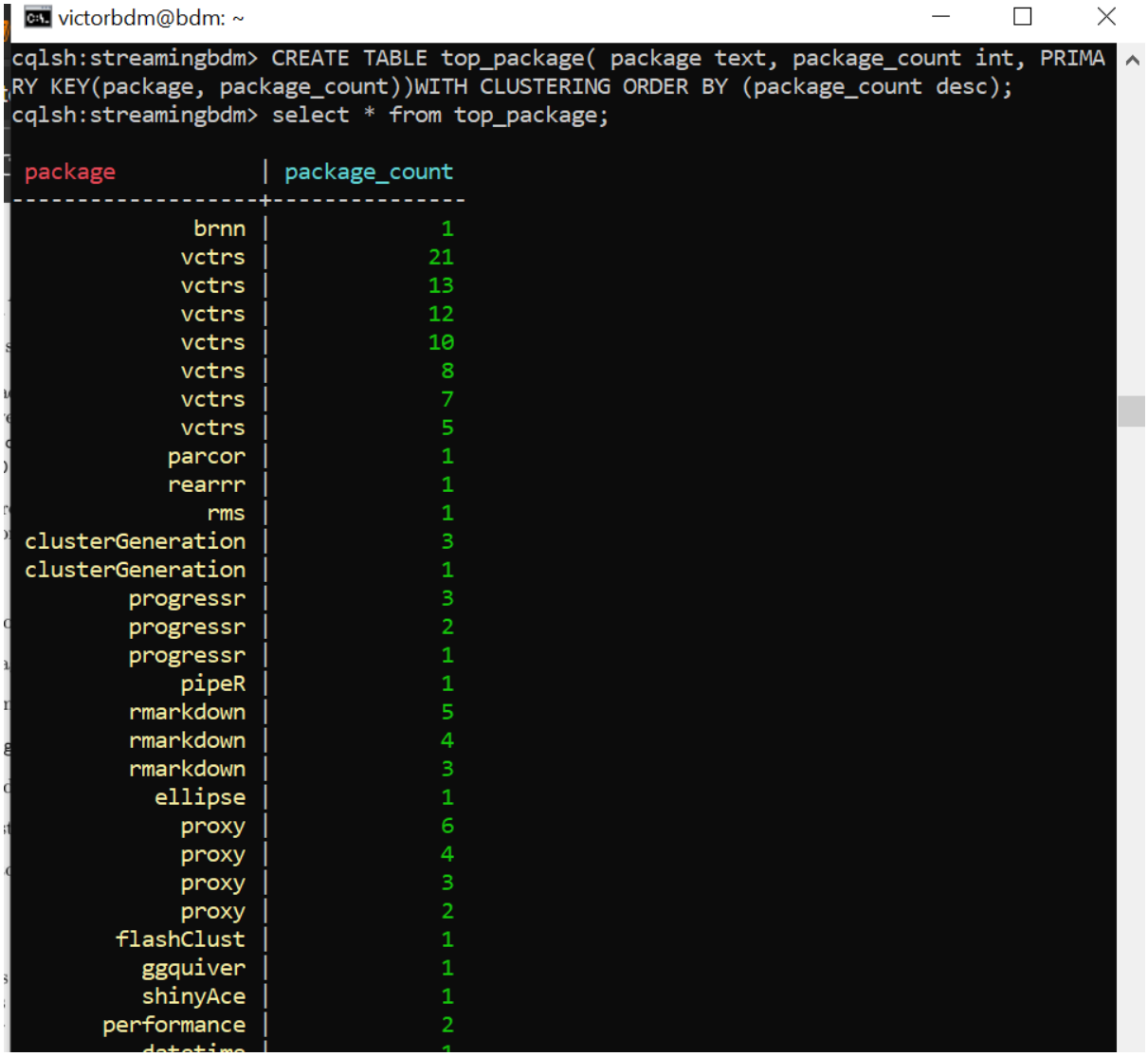
```
2021-12-21 23:02:00+0000 |                 zip |     3
2021-12-21 23:02:00+0000 |                 zoo |     1
2021-12-21 23:01:54+0000 |       ATAforecasting |     1
2021-12-21 23:01:54+0000 |                 ATE |     1
2021-12-21 23:01:54+0000 |                 ATR |     1
2021-12-21 23:01:54+0000 |               ATmet |     1
2021-12-21 23:01:54+0000 |          AsynchLong |     1
2021-12-21 23:01:54+0000 |            AtmChile |     1
2021-12-21 23:01:54+0000 |              AtmRay |     1
2021-12-21 23:01:54+0000 |                  BH |     1
2021-12-21 23:01:54+0000 |            BioStatR |     1
2021-12-21 23:01:54+0000 |                 DBI |     2
2021-12-21 23:01:54+0000 |            DEoptimR |     2
2021-12-21 23:01:54+0000 |           DescTools |     1
2021-12-21 23:01:54+0000 |    DistributionUtils |     1
2021-12-21 23:01:54+0000 |                 FNN |     1
2021-12-21 23:01:54+0000 |          GPArotation |     1
2021-12-21 23:01:54+0000 | GeneralizedHyperbolic |     1
2021-12-21 23:01:54+0000 |            Gmedian |     1
2021-12-21 23:01:54+0000 |           HaploSim |     1
2021-12-21 23:01:54+0000 |           HistData |     1
2021-12-21 23:01:54+0000 |              Hmisc |     6
2021-12-21 23:01:54+0000 |               ISLR |     3
2021-12-21 23:01:54+0000 |             KFKSDS |     1
2021-12-21 23:01:54+0000 |             Matrix |     1
2021-12-21 23:01:54+0000 |         MatrixModels |     4
2021-12-21 23:01:54+0000 |         ModelMetrics |     3
2021-12-21 23:01:54+0000 |             NGSSEML |     1
```

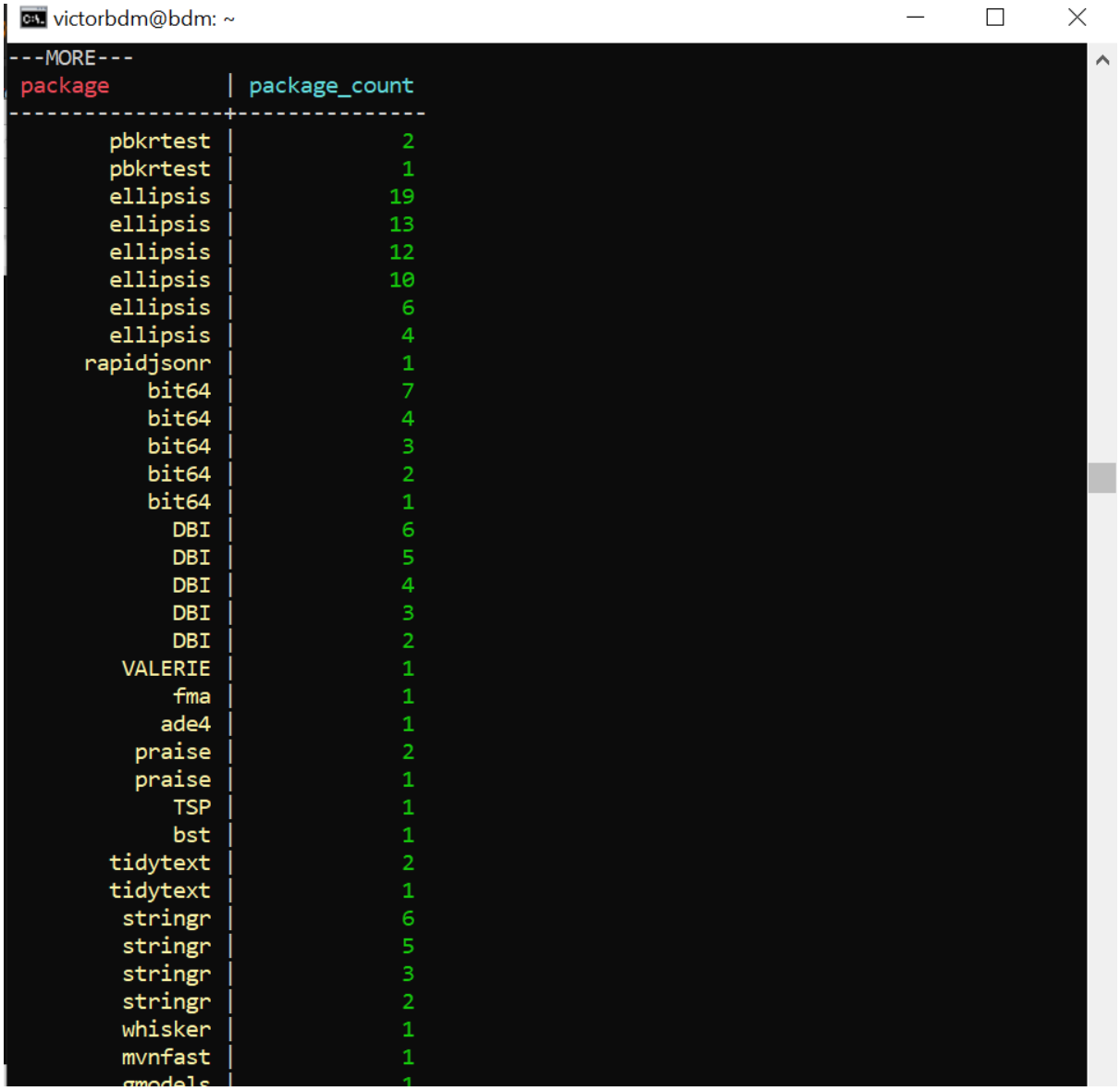
## Question 2

**question 2**

In [15]:
```
def save(time, rdd):

    try:
        df = spark.createDataFrame(rdd.map(\
        lambda row: Row(timenew=time, date=row[0], time=row[1], size=row[2], r_version=row[3], r_arch=row[4], r_os=row[5], packag
        df.createOrReplaceTempView('cran_data')
        df = spark.sql('SELECT package, COUNT(cran_data.package) AS package_count FROM cran_data GROUP BY package ORDER BY packag
        df.show()
        df.select("package", "package_count")\
        .write.format("org.apache.spark.sql.cassandra")\
        .options(table="top_package", keyspace="streamingbdm")\
        .save(mode="append")
    except:
        pass
```

In [16]:
```
ssc = StreamingContext(sc, 3)

cran_data = ssc.textFileStream("file:///home/victorbdm/assignStreaming/")
cran_data = cran_data.map(lambda x: x.split(','))

def remove_quotation(x):
    return([xx.replace('"', '') for xx in x])
cran_data = cran_data.map(remove_quotation)
cran_data.foreachRDD(save)
ssc.start()
time.sleep(120)
ssc.stop(stopSparkContext=False)
```


victorbdm@bdm: ~
cqlsh:streamingbdm> CREATE TABLE top_package( package text, package_count int, PRIMARY KEY(package, package_count))WITH CLUSTERING ORDER BY (package_count desc);
cqlsh:streamingbdm> select * from top_package;
package | package_count
brnn | 1
vctrs | 21
vctrs | 13
vctrs | 12
vctrs | 10
vctrs | 8
vctrs | 7
vctrs | 5
parcor | 1
rearrr | 1
rms | 1
clusterGeneration | 3
clusterGeneration | 1
progressr | 3
progressr | 2
progressr | 1
pipeR | 1
rmarkdown | 5
rmarkdown | 4
rmarkdown | 3
ellipse | 1
proxy | 6
proxy | 4
proxy | 3
proxy | 2
flashClust | 1
ggquiver | 1
shinyAce | 1
performance | 2




Question 3

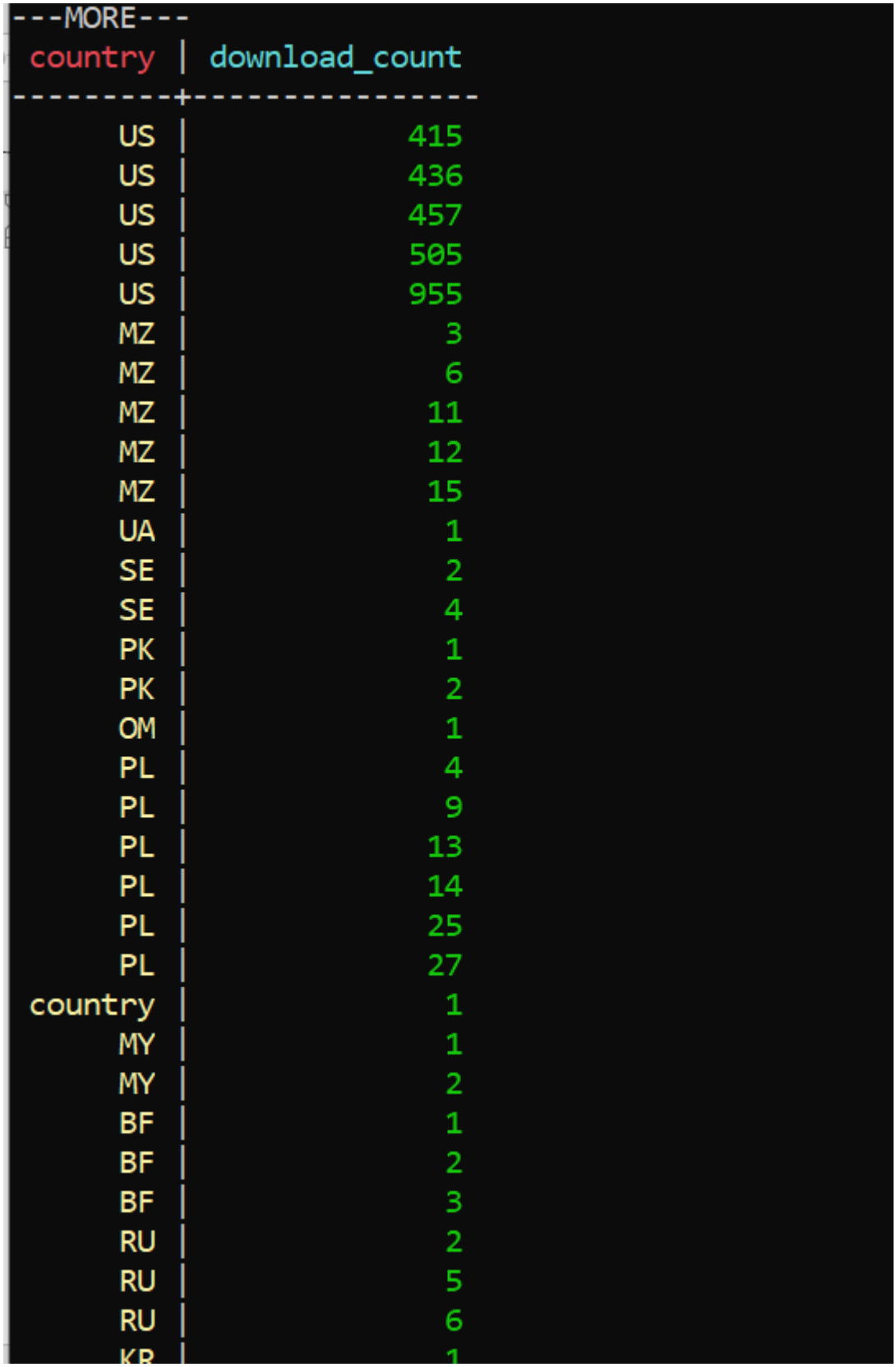
**question 3**

In [11]:
```
save(time, rdd):
try:
    df = spark.createDataFrame(rdd.map(\
    lambda row: Row(timenew=time, date=row[0], time=row[1], size=row[2], r_version=row[3], r_arch=row[4], r_os=row[5], package=ro
    df.createOrReplaceTempView('cran_data')
    df = spark.sql('SELECT country, COUNT(package) AS download_count FROM cran_data GROUP BY country ORDER BY download_count DESC
    df.show(5)
    df.write.format("org.apache.spark.sql.cassandra")\
    .options(table="country_counts", keyspace="streamingbdm")\
    .save(mode="append")
except:
    pass
```

In [12]:
```
ssc = StreamingContext(sc, 3)
cran_data = ssc.textFileStream("file:///home/victorbdm/assignStreaming/")
cran_data = cran_data.map(lambda x: x.split(','))

def remove_quotation(x):
    return([xx.replace('"', '') for xx in x])
cran_data = cran_data.map(remove_quotation)
cran_data.foreachRDD(save)
ssc.start()
time.sleep(120)
ssc.stop(stopSparkContext=False)
```

```
+-------+--------------+
```

```
cqlsh:streamingbdm> select * from country_counts;

 country | download_count
---------+----------------
      IN |              1
      IN |              2
      IN |              5
      IN |             11
      PE |              2
      PE |              4
      PE |              8
      PE |             13
      PH |              5
      AT |              3
      AT |              5
      JP |              1
      JP |              2
      JP |              9
      JP |             10
      JP |             13
      HK |              2
      HK |              5
      HK |              7
      HK |             36
      HK |             47
      FR |              2
      FR |              3
      FR |              6
      FR |              7
      NA |             99
      NA |            115
      NA |            125
```



```
---MORE---
 country | download_count
---------+----------------
      US |            415
      US |            436
      US |            457
      US |            505
      US |            955
      MZ |              3
      MZ |              6
      MZ |             11
      MZ |             12
      MZ |             15
      UA |              1
      SE |              2
      SE |              4
      PK |              1
      PK |              2
      OM |              1
      PL |              4
      PL |              9
      PL |             13
      PL |             14
      PL |             25
      PL |             27
 country |              1
      MY |              1
      MY |              2
      BF |              1
      BF |              2
      BF |              3
      RU |              2
      RU |              5
      RU |              6
      KR |              1
```

Question 4

## question 4

In [2]:
```
def save(time, rdd):
    try:
        df = spark.createDataFrame(rdd.map(\
        lambda row: Row(time=time, package=row[0], gp_count=row[1])))
        df.show(5)
        df.write.format("org.apache.spark.sql.cassandra")\
        .options(table="gplot_counts", keyspace="streamingbdm")\
        .save(mode="append")
    except:
        pass
```

In [3]:
```
ssc = StreamingContext(sc, 3)
cran_data = ssc.textFileStream("file:///home/victorbdm/assignStreaming/
cran_data = cran_data.map(lambda x: x.split(','))

def remove_quotation(x):
    return([xx.replace('"', '') for xx in x])
cran_data = cran_data.map(remove_quotation)
ggplot2_package = cran_data.map(lambda package:(package[6], 1))
ggplot2_package = ggplot2_package.reduceByKey(lambda a,b: a+b)
ggplot2_package= ggplot2_package.filter(lambda a: 'ggplot2' in a)
ggplot2_package.foreachRDD(save)

ssc.start()
time.sleep(120)
ssc.stop(stopSparkContext=False)
```

victorbdm@bdm: ~

```
(0 rows)
cqlsh:streamingbdm> select * from gplot_counts;

 time                     | package | gp_count
--------------------------+---------+----------
 2021-12-22 02:17:18+0000 | ggplot2 |       22
 2021-12-22 02:17:09+0000 | ggplot2 |       54
 2021-12-22 02:17:15+0000 | ggplot2 |       34

(3 rows)
cqlsh:streamingbdm> select * from gplot_counts;

 time                     | package | gp_count
--------------------------+---------+----------
 2021-12-22 02:17:18+0000 | ggplot2 |       22
 2021-12-22 02:17:09+0000 | ggplot2 |       54
 2021-12-22 02:17:24+0000 | ggplot2 |       15
 2021-12-22 02:17:21+0000 | ggplot2 |       17
 2021-12-22 02:17:15+0000 | ggplot2 |       34

(5 rows)
cqlsh:streamingbdm> select * from gplot_counts;

 time                     | package | gp_count
--------------------------+---------+----------
 2021-12-22 02:17:18+0000 | ggplot2 |       22
 2021-12-22 02:17:36+0000 | ggplot2 |       17
 2021-12-22 02:17:09+0000 | ggplot2 |       54
 2021-12-22 02:17:39+0000 | ggplot2 |       17
 2021-12-22 02:17:27+0000 | ggplot2 |       23
 2021-12-22 02:17:24+0000 | ggplot2 |       15
 2021-12-22 02:17:45+0000 | ggplot2 |       27
 2021-12-22 02:17:33+0000 | ggplot2 |       28
 2021-12-22 02:17:21+0000 | ggplot2 |       17
 2021-12-22 02:17:30+0000 | ggplot2 |       19
 2021-12-22 02:17:42+0000 | ggplot2 |       19
 2021-12-22 02:17:15+0000 | ggplot2 |       34
```

victorbdm@bdm: ~

```
(26 rows)
cqlsh:streamingbdm> select * from gplot_counts;

 time                     | package | gp_count
--------------------------+---------+----------
 2021-12-22 02:18:30+0000 | ggplot2 |       28
 2021-12-22 02:17:54+0000 | ggplot2 |       20
 2021-12-22 02:18:21+0000 | ggplot2 |       27
 2021-12-22 02:17:18+0000 | ggplot2 |       22
 2021-12-22 02:17:57+0000 | ggplot2 |       23
 2021-12-22 02:18:24+0000 | ggplot2 |       19
 2021-12-22 02:17:36+0000 | ggplot2 |       17
 2021-12-22 02:17:09+0000 | ggplot2 |       54
 2021-12-22 02:17:39+0000 | ggplot2 |       17
 2021-12-22 02:17:27+0000 | ggplot2 |       23
 2021-12-22 02:17:24+0000 | ggplot2 |       15
 2021-12-22 02:18:03+0000 | ggplot2 |       14
 2021-12-22 02:17:45+0000 | ggplot2 |       27
 2021-12-22 02:18:15+0000 | ggplot2 |       17
 2021-12-22 02:17:51+0000 | ggplot2 |       13
 2021-12-22 02:18:27+0000 | ggplot2 |       28
 2021-12-22 02:18:48+0000 | ggplot2 |       30
 2021-12-22 02:18:42+0000 | ggplot2 |       22
 2021-12-22 02:18:39+0000 | ggplot2 |       27
 2021-12-22 02:19:03+0000 | ggplot2 |       14
 2021-12-22 02:18:36+0000 | ggplot2 |       33
 2021-12-22 02:17:33+0000 | ggplot2 |       28
 2021-12-22 02:17:21+0000 | ggplot2 |       17
 2021-12-22 02:18:00+0000 | ggplot2 |       18
 2021-12-22 02:17:30+0000 | ggplot2 |       19
 2021-12-22 02:18:54+0000 | ggplot2 |       27
 2021-12-22 02:18:51+0000 | ggplot2 |       27
 2021-12-22 02:17:48+0000 | ggplot2 |       20
 2021-12-22 02:17:42+0000 | ggplot2 |       19
 2021-12-22 02:18:09+0000 | ggplot2 |       16
 2021-12-22 02:18:57+0000 | ggplot2 |       16
 2021-12-22 02:19:00+0000 | ggplot2 |       14
```