

# GRIFFITH COLLEGE DUBLIN

## BIG DATA MANAGEMENT AND ANALYTICS

### BIG DATA MANAGEMENT ASSIGNMENT 2

ARAVINDAN SRINIVASAN(2981707)

28/03/2019

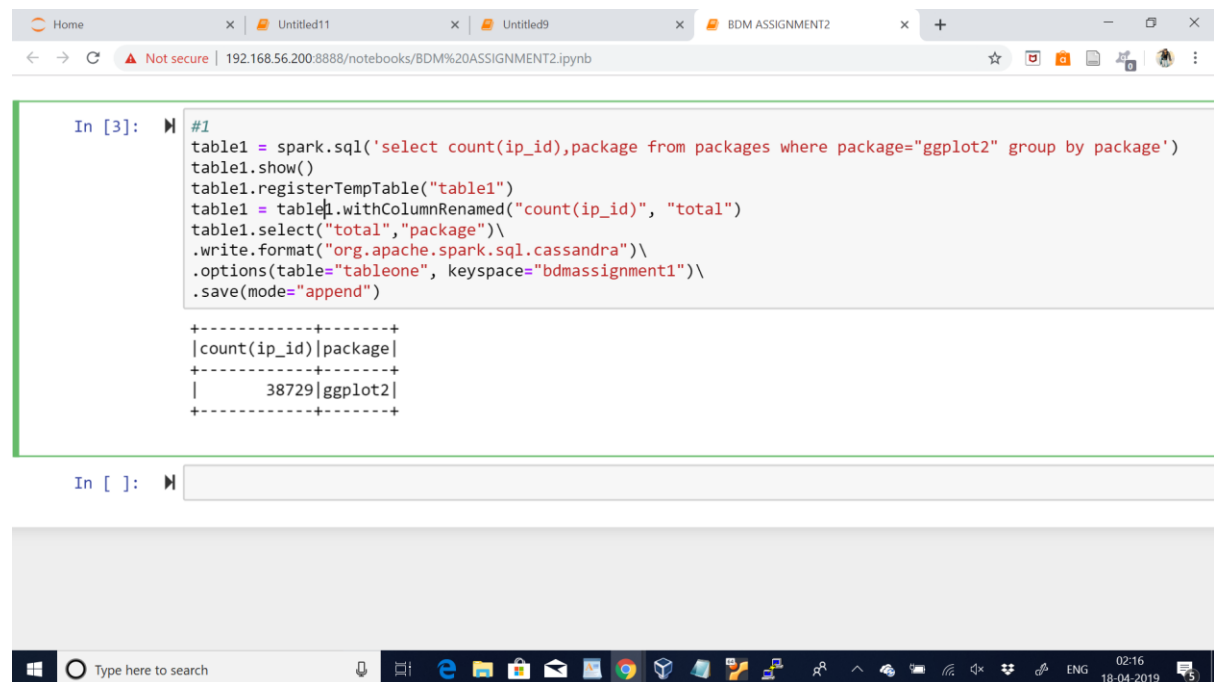
[aravindan.srinivasan@student.griffith.ie](mailto:aravindan.srinivasan@student.griffith.ie)

#### INTRODUCTION:

Our task is to prepare batch layer of lambda architecture. We have to do some basic analytics and display the output as well as transfer the output into Cassandra tables.

#### **#1. Show number of downloads for package ggplot2**

#### **CODING:**



The screenshot shows a Jupyter Notebook window titled 'BDM ASSIGNMENT2'. The browser address bar indicates the URL '192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb'. The notebook contains a code cell with the following Python code:

```
In [3]: #1
table1 = spark.sql('select count(ip_id),package from packages where package="ggplot2" group by package')
table1.show()
table1.registerTempTable("table1")
table1 = table1.withColumnRenamed("count(ip_id)", "total")
table1.select("total", "package")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tableone", keyspace="bdmassignment1")\
.save(mode="append")
```

The output of the code is a table with two columns: 'count(ip\_id)' and 'package'. The output is displayed as follows:

count(ip_id)	package
38729	ggplot2

The bottom of the screenshot shows the Windows taskbar with the search bar and various application icons. The system clock indicates the time is 02:16 on 18-04-2019.

#### **O/P:**







```
aravindan81222@ubuntu: ~
cqlsh:bdmassignment1> create table tablefour (total int ,package text,primary key(package));
cqlsh:bdmassignment1> select * from tablefour;

package | total
-----+-----
ggplot2 | 38729
rlang | 55592
fansi | 37598
dplyr | 39443
stringr | 39439
yaml | 38422
pillar | 40948
Rcpp | 50448
tibble | 45020
R6 | 39063

(10 rows)
cqlsh:bdmassignment1>
```

**#5. In both days, at what specific hour there are most of the download hits?**

**CODING:**

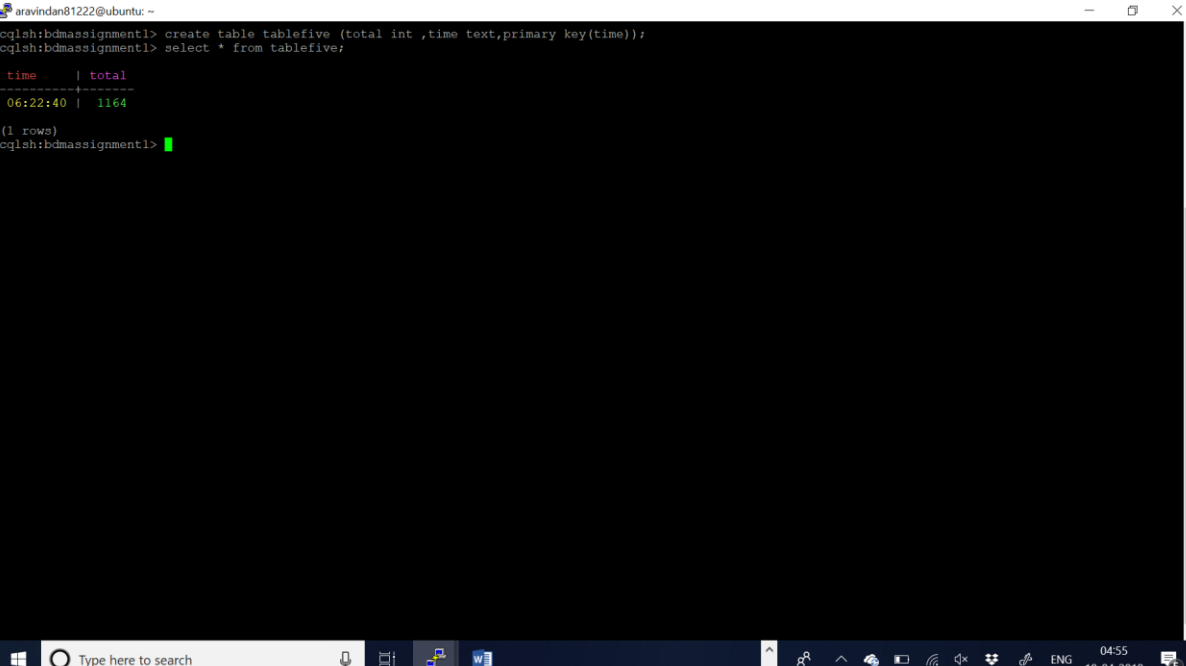
```
Home x | Untitled11 x | Untitled9 x | BDM ASSIGNMENT2 x | +
192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb
Jupyter BDM ASSIGNMENT2 Last Checkpoint: 28 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
Run Code

In [14]: #5
tables5 = spark.sql('select count(ip_id),time from packages group by time order by count(ip_id) desc limit 1')
tables5.show()
tables5.registerTempTable("tables5")
tables5 = tables5.withColumnRenamed("count(ip_id)", "total")
tables5.select("total","time")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tablefive", keyspace="bdmassignment1")\
.save(mode="append")

+-----+-----+
|count(ip_id)| time|
+-----+-----+
| 1164|06:22:40|
+-----+-----+

In [ ]:
In [ ]:
```

**O/P:**



```

aravindan81222@ubuntu: ~
cqlsh:bdmassignment1> create table tablefive (total int ,time text,primary key(time));
cqlsh:bdmassignment1> select * from tablefive;

time      | total
-----+-----
06:22:40  | 1164

(1 rows)
cqlsh:bdmassignment1>
  
```

### #6. What are the 5 most popular packages in UK?

**CODING:**

The screenshot shows a Jupyter Notebook in a web browser. The browser's address bar indicates the URL is `192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb`. The notebook interface includes a toolbar with icons for saving, running, and other actions. The code cell, labeled [17], contains the following Spark SQL query:

```
#6
table6 = spark.sql('select count(ip_id),package from packages where country="GB" group by package limit 5')
table6.show()
table6.registerTempTable("table6")
table6 = table6.withColumnRenamed("count(ip_id)", "total")
table6.select("total", "package")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tablesix", keyspace="bdmassignment1")\
.save(mode="append")
```

The output of the query is displayed as a table:

count(ip_id)	package
167	TH.data
120	RcppParallel
1	RSSL
1	rtson
1	RYandexTranslate

The bottom of the image shows the Windows taskbar with various application icons and the system clock indicating 05:45 on 18-04-2019.

**O/P:**

```
aravindan81222@ubuntu: ~
cqlsh:bdmassignment1> create table tablesix (total int ,package text,primary key(package));
cqlsh:bdmassignment1> select * from tablesix;

package | total
-----+-----
RYandexTranslate | 1
RoppParallel | 120
TH_data | 167
RSSL | 1
rtson | 1

(5 rows)
cqlsh:bdmassignment1>
```

**#7. Show total number of downloads by (each of the) top five machines?**

**CODING:**

```
Home x | Untitled11 x | Untitled9 x | BDM ASSIGNMENT2 x | +
192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb
+-----+
In [19]: #7
table7 = spark.sql('select count(ip_id),ip_id from packages group by ip_id order by count(ip_id) desc limit 5')
table7.show()
table7.registerTempTable("table7")
table7 = table7.withColumnRenamed("count(ip_id)", "total")
table7.select("total","ip_id")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tableseven", keyspace="bdmassignment1")\
.save(mode="append")

+-----+-----+
|count(ip_id)|ip_id|
+-----+-----+
|228763|8|
|187698|18|
|164673|38|
|19272|3007|
|18803|1|
+-----+-----+

In [18]: downloadsDF.persist()
```

**O/P:**

```
aravindan81222@ubuntu: ~
cqlsh:bdmassignment1> create table tableseven (total int ,ip_id text,primary key(ip_id));
cqlsh:bdmassignment1> select * from tableseven;

 ip_id | total
-----+-----
    18 | 187698
     8 | 228763
    38 | 164673
   3007 | 19272
     1 | 18803

(5 rows)
cqlsh:bdmassignment1>
```

**#8. Show top three OSs that are most popular among the R programmers?**

**CODING:**

```
Home x | Untitled11 x | Untitled9 x | BDM ASSIGNMENT2 x | +
< -> C Not secure | 192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb ☆ ⓘ 📄 📁 📧 📧 📧 ⋮
R DataFrame[date: date, time: string, size: int, r_version: string, r_arch: string, os: string, package: string, country: string, ip_id: string]

#8
table8 = spark.sql('select count(ip_id),os from packages group by os order by count(ip_id) desc limit 3')
table8.show()
table8.registerTempTable("table8")
table8 = table8.withColumnRenamed("count(ip_id)", "total")
table8.select("total","os")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tableeight", keyspace="bdmassignment1")\
.save(mode="append")

+-----+-----+
|count(ip_id)|      os|
+-----+-----+
|    2000498|  mingw32|
|    1581058| linux-gnu|
|    454098 |darwin15.6.0|
+-----+-----+

#9
table9 = spark.sql('select count(ip_id),os from packages group by os order by count(ip_id) desc')
```

**O/P:**



```
aravindan81222@ubuntu: ~
cqlsh:bdmassignment1> create table tableeight (total int ,os text,primary key(os));
cqlsh:bdmassignment1> select * from tableeight;

os | total
-----
darwin15.6.0 | 454098
mingw32 | 2000498
linux-gnu | 1581058

(3 rows)
cqlsh:bdmassignment1>
```

**#9. Show total number of downloads by each OS type?**

**CODING:**

```
Home x | Untitled11 x | Untitled9 x | BDM ASSIGNMENT2 x | +
192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb
File Edit View Insert Cell Kernel Widgets Help Trusted Python
Run Code

In [21]: #9
table9 = spark.sql('select count(ip_id),os from packages group by os order by count(ip_id) desc')
table9.show()
table9.registerTempTable("table9")
table9 = table9.withColumnRenamed("count(ip_id)", "total")
table9.select("total","os")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tablenine", keyspace="bdmassignment1")\
.save(mode="append")

+-----+
|count(ip_id)| os |
+-----+
|2000498| mingw32 |
|1581058| linux-gnu |
|454098| darwin15.6.0 |
|120588| darwin13.4.0 |
|120413| NA |
|9407| darwin18.2.0 |
|2410| darwin17.7.0 |
|1883| darwin16.7.0 |
|1871| linux-gnueabi |
|1661| darwin17.6.0 |
|861| darwin17.4.0 |
|641| darwin18.0.0 |
|338| darwin11.4.2 |
|222| darwin10.8.0 |
```

**O/P:**

```
aravindan81222@ubuntu: ~
oqlsh:bdmassignment1> create table tablenine (total int ,os text,primary key(os));
oqlsh:bdmassignment1> select * from tablenine;

os | total
-----
linux-gnueabi | 1871
darwin17.4.0 | 861
darwin16.6.0 | 44
darwin15.2.0 | 4
darwin18.5.0 | 3
NA | 120413
darwin15.6.0 | 454098
solaris2.10 | 12
darwin15.5.0 | 264
darwin16.0.0 | 2
darwin11.4.2 | 338
darwin18.0.0 | 641
freebsd11.2 | 35
darwin13.4.0 | 120588
darwin17.2.0 | 29
mingw32 | 2000498
darwin17.6.0 | 1661
darwin14.5.0 | 147
darwin17.3.0 | 165
linux-gnu | 1591058
darwin16.4.0 | 3
darwin16.7.0 | 1883
darwin17.0.0 | 18
darwin10.8.0 | 322
darwin19.2.0 | 9407
darwin17.7.0 | 2410
darwin17.5.0 | 222
darwin16.1.0 | 198
darwin13.1.0 | 5

(29 rows)
oqlsh:bdmassignment1>
```

#10. Show total number of downloads by each country?

**CODING:**

```
Home x | Untitled11 x | Untitled9 x | BDM ASSIGNMENT2 x | +
192.168.56.200:8888/notebooks/BDM%20ASSIGNMENT2.ipynb
jupyter BDM ASSIGNMENT2 Last Checkpoint: 44 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3
In [22]: #10
table10 = spark.sql('select count(ip_id),country from packages group by country order by count(ip_id) desc')
table10.show()
table10.registerTempTable("table10")
table10 = table10.withColumnRenamed("count(ip_id)", "total")
table10.select("total","country")\
.write.format("org.apache.spark.sql.cassandra")\
.options(table="tablenine", keyspace="bdmassignment1")\
.save(mode="append")

+-----+-----+
|count(ip_id)|country|
+-----+-----+
|1776597|US|
|352318|NA|
|324601|CA|
|152275|DE|
|146479|GB|
|127671|IN|
|111794|HK|
|103439|JP|
|101058|CN|
|83768|FR|
|81804|ES|
|67838|NL|
|63396|AU|
|56621|CO|
|54846|CH|
|46861|MX|
|44648|IT|
+-----+-----+
```

**O/P:**

```
aravindan81222@ubuntu: ~  
country | total  
-----  
AZ | 38  
VI | 23  
HR | 1234  
IN | 127671  
TW | 16486  
EU | 3958  
FE | 10708  
PR | 7748  
NP | 649  
AT | 16532  
PG | 19  
JP | 103439  
IR | 5153  
KE | 6454  
KW | 611  
NE | 295  
CU | 139  
CD | 80  
UY | 1938  
HR | 111784  
BW | 602  
CM | 413  
FR | 83768  
MD | 125  
CG | 72  
UZ | 41  
NA | 352318  
HT | 14  
KZ | 756  
RE | 197  
AO | 195  
SV | 973  
LK | 839  
JO | 237  
SO | 152  
BE | 13576  
AZ | 383  
HU | 4541  
IT | 44640  
CN | 101058  
ET | 719  
FR | 1068  
  
SK | 1451  
BR | 39300  
MB | 28  
IS | 1486  
LA | 91  
CL | 10736  
DK | 16369  
MG | 230  
GN | 318  
KG | 33  
GR | 5019  
SZ | 35  
GY | 7  
FJ | 33  
ZM | 457  
LU | 1170  
GB | 276  
BO | 993  
SN | 1901  
SL | 167  
DE | 152275  
QA | 339  
NG | 1  
AB | 2891  
DZ | 846  
AP | 624  
NG | 2649  
GI | 13  
FI | 9840  
RS | 2608  
CR | 80  
HN | 539  
LV | 1289  
GB | 146479  
MK | 331  
MW | 419  
NI | 7  
ZA | 20407  
TM | 1338  
VE | 1169  
EG | 1640  
CR | 5971  
GT | 2407  
MU | 1274
```

```
aravindan81222@ubuntu: ~  
GB | 276  
BO | 993  
SN | 1901  
SL | 167  
DE | 152275  
QA | 339  
NG | 1  
AE | 2891  
DZ | 846  
AP | 624  
NG | 2649  
GI | 13  
FI | 9840  
RS | 2608  
CR | 80  
HN | 539  
LV | 1289  
GB | 146479  
MK | 331  
MW | 419  
NI | 7  
ZA | 20407  
TN | 1398  
VE | 1169  
EG | 1640  
CR | 5971  
GT | 2407  
MU | 1274  
NZ | 9819  
SG | 19354  
AR | 13587  
AU | 63396  
MO | 570  
IE | 10541  
CO | 56621  
VN | 5749  
BY | 183  
US | 1776597  
MZ | 704  
BN | 10  
SA | 2230  
BS | 3  
---MORE---
```

## CONCLUSION:

Thus, the key space having ten tables which fetched the values from spark sql. All the information and values are successfully loaded into Cassandra tables.

```
aravindan81222@ubuntu: ~  
cqlsh:bdmassignment1> describe tables;  
tablefour  tablesix  tableeight  mike  tableone  tablethree  
tabletwo   tableseven  tableten    tablefive  tablenine  
cqlsh:bdmassignment1>
```