Subject : BDM
Student Number : 2982006
Name : Siddhi Kate
Topic : Assignment 02

siddhi123@ubuntu: ~

```
login as: siddhi123
siddhi123@192.168.56.200's password:
Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-131-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

67 packages can be updated.
47 updates are security updates.

New release '18.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


Last login: Thu Apr 18 19:38:23 2019 from 192.168.56.1
siddhi123@ubuntu:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.4 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
cqlsh> use assignment_02
   ... ;
```

Q.1 Show number of downloads for package ggplot2.

```
#Question 1
query_results = spark.sql('SELECT package, COUNT(package) AS count FROM packages WHERE package="ggplot2" GROUP BY package')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question01",keyspace="assignment_02")\
    .save(mode="append")
```

```
query_results = spark.sql('SELECT package, COUNT(package) AS count FROM packages WHERE package="ggplot2" GROUP BY package')\
query_results.show()
```

```
+-------+--------------+
|package|count(package)|
+-------+--------------+
|ggplot2|         38729|
+-------+--------------+
```

```
cqlsh:assignment_02> select * from question01;

 package | count
---------+-------
 ggplot2 | 38729

(1 rows)
```

Q.2 Highest number of downloads by a country and Operating System.

```
# Question 2
query_results = spark.sql('SELECT country,COUNT(package) AS count FROM packages GROUP BY country ORDER BY COUNT(package)\
               DESC LIMIT 10' )\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question02",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question02( country text primary key, count in
t);
cqlsh:assignment_02> select * from question02;

 country | count
---------+---------
      IN |  127671
      JP |  103439
      HK |  111794
      FR |   83768
      NA |  352318
      CN |  101058
      DE |  152275
      GB |  146479
      US | 1776597
      CA |  324601

(10 rows)
```

Q.3 Top 10 (distinct) largest sized packages.

```
#Question 3
query_results = spark.sql('select distinct package, size FROM packages order by size desc limit 10')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question03",keyspace="assignment_02")\
    .save(mode="append")
```

```
query_results = spark.sql('select distinct package, size FROM packages order by size desc limit 10')
query_results.show()
```

```
+-------+---------+
|package|     size|
+-------+---------+
|    h2o|122267950|
|    h2o|122267949|
|    h2o|122267948|
|    h2o|122267946|
|    h2o|122267940|
|    h2o|122133439|
|    h2o|122133438|
|    h2o|122133437|
|    h2o|122133435|
|    h2o|122133429|
+-------+---------+
```

```
cqlsh:assignment_02> select * from questionss;

 size      | package
-----------+---------
 122133438 |     h2o
 122267940 |     h2o
 122267948 |     h2o
 122267949 |     h2o
 122133429 |     h2o
 122267946 |     h2o
 122133437 |     h2o
 122267950 |     h2o
 122133435 |     h2o
 122133439 |     h2o

(10 rows)
```

Q.4 What are the top 10 most popular (distinct) packages?

```
#Question 4
query_results = spark.sql('SELECT package, COUNT(package) AS count FROM packages GROUP BY package ORDER BY COUNT(package)\
                DESC limit 10')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question04",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question04( package text primary key, count in
t);
cqlsh:assignment_02> select * from question04;

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

(10 rows)
```

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

```
#Question 5
query_results = spark.sql('SELECT time, COUNT(time) AS count  FROM packages GROUP BY time ORDER BY COUNT(time) DESC limit 4 ')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question05",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question05( time text primary key, count int);
cqlsh:assignment_02> select * from question05;

 time     | count
----------+-------
 06:20:08 |   715
 06:22:39 |   682
 06:22:40 |  1164
 06:30:15 |   521

(4 rows)
```

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

```
#Question 6
query_results = spark.sql('SELECT package,COUNT(package) AS count FROM packages WHERE country="GB" GROUP BY package\
                \ORDER BY COUNT(package) DESC limit 5')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question06",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question06(
                ... package text primary key,
                ... count int);
cqlsh:assignment_02> select * from question06;

 package | count
---------+-------
 ggplot2 |  1516
   rlang |  1839
   dplyr |  1486
    Rcpp |  1488
  tibble |  1486

(5 rows)
```

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

```
#Question 7
query_results = spark.sql('SELECT ip_id, COUNT(ip_id) AS count  FROM packages GROUP BY ip_id ORDER BY COUNT(ip_id) DESC LIMIT 5')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question07",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> select * from question07;

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

(5 rows)
```

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

```
#Question 8
query_results = spark.sql('SELECT r_os, COUNT(r_os) AS count FROM packages GROUP BY r_os ORDER BY COUNT(r_os) DESC limit 3')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question08",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question08( r_os text primary key, count int);
cqlsh:assignment_02> select * from question08;

 r_os         | count
--------------+---------
 darwin15.6.0 |  454098
      mingw32 | 2000498
    linux-gnu | 1581058

(3 rows)
```

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

```
#Question 9
query_results = spark.sql('SELECT r_os, COUNT(r_os) AS count FROM packages GROUP BY r_os ')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question09",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question09( r_os text primary key, count int);
cqlsh:assignment_02> select * from question09;

 r_os            | count
-----------------+---------
 linux-gnueabihf |    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 | 1581058
    darwin16.4.0 |       3
    darwin16.7.0 |    1883
    darwin17.0.0 |      18
    darwin10.8.0 |     322
    darwin18.2.0 |    9407
    darwin17.7.0 |    2410
    darwin17.5.0 |     222
    darwin16.1.0 |     198
    darwin13.1.0 |       5

(29 rows)
```

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

```
#Question 10
query_results = spark.sql('SELECT country, COUNT(package) AS count  FROM packages GROUP BY country ')\
    .write.format("org.apache.spark.sql.cassandra")\
    .options(table="question10",keyspace="assignment_02")\
    .save(mode="append")
```

```
cqlsh:assignment_02> create table question10( country text primary key, count in
t);
cqlsh:assignment_02> select * from question10;

 country | count
---------+---------
      A2 |      38
      VI |      23
      HR |    1234
      IN |  127671
      TW |   16486
      EU |    3958
      PE |   10708
      PH |    7748
      NP |     649
      AT |   16532
      PG |      19
      JP |  103439
      IR |    5193
      KE |    6454
      KW |     611
      NE |     295
      CU |     139
      CD |      80
      UY |    1938
      HK |  111794
      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
```