Project-1

Analysis of Facebook Dataset by using Apache Hive

SUBMITTED BY:- URVASHI UPADHYAY

Insights Needed

- Find out the total number of users in this data sets
- Find out the number of Facebook users above the age of 25
- Do male Facebook users tend to have more friends , or female users?
- How many likes do young people receive on Facebook opposed to older members?
- Find out the count of Facebook users for each birthday month?
- Do young members use mobile phones or computer for Facebook browsing?
- Do adult members use mobile phone or computers for facebook browsing?

List of files inside HDFS

```
16194 ResourceManager
12901 ApplicationHistoryServer
17735 HistoryServer
17031 RunJar
14665 RunJar
10377 Bootstrap
8876 RunJar
7117 ZeppelinServer
7438 TagSynchronizer
9358 UnixAuthenticationService
22958 RunJar
7535 QuorumPeerMain
15379 Kafka
16373 SecondaryNameNode
13589 JobHistoryServer
24727 Jps
13816 AmbariServer
10458 NameNode
10172 DataNode
3069 LivyServer
6942 JournalNode
8959 EmbeddedServer
23455 DAGAppMaster
[root@sandbox-hdp ~]# hdfs dfs -ls /
Found 12 items
drwxr-xr-x   - admin  hdfs            0 2021-08-20 09:13 /Data
drwxrwxrwx   - yarn   hadoop          0 2021-08-20 08:43 /app-logs
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:13 /apps
drwxr-xr-x   - yarn   hadoop          0 2018-06-18 14:52 /ats
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 14:52 /hdp
drwx------   - livy   hdfs            0 2018-06-18 15:11 /livy2-recovery
drwxr-xr-x   - mapred hdfs            0 2018-06-18 14:52 /mapred
drwxrwxrwx   - mapred hadoop          0 2018-06-18 14:52 /mr-history
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 15:59 /ranger
drwxrwxrwx   - spark  hadoop          0 2021-08-20 09:14 /spark2-history
drwxrwxrwx   - hdfs   hdfs            0 2018-06-18 16:06 /tmp
drwxr-xr-x   - hdfs   hdfs            0 2018-06-18 16:08 /user
[root@sandbox-hdp ~]# hdfs dfs -ls /Data/
Found 1 items
-rw-r--r--   1 admin hdfs    1048838 2021-08-20 09:13 /Data/weather_data1.csv
[root@sandbox-hdp ~]#
```

Input data in hive

```
CREATE TABLE FB(ID INT, AGE INT, DAY INT, YEAR INT, MONTH
INT, GENDER STRING, TENURE INT, FRIENDS INT, FRIEND_INIT
INT, LIKES INT, LIKES_REC
D INT, MLIKES INT, MLIKES_RECD INT, WLIKES INT,
WLIKES_RECD INT) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ',' STORED AS TEXTFILE LOCATION
'/PROJECT/FACEBOOK_DATA';
```

```
------------+
| fb.id | fb.age | fb.day | fb.year | fb.month | fb.gender | fb.tenure | fb.friends | fb.friend_init | fb.likes | fb.likes_recd | fb.mlikes | fb.mlikes_recd | fb.wlik
.wlikes_recd |
+--------+---------+---------+----------+-----------+------------+------------+-------------+----------------+-----------+---------------+-----------+----------------+-----------+
--------------+
+--------+--------
```

Find out the total number of users in this data sets

select count(userid) AS Total_Users from fb;

```
hive> SELECT count(userid) AS Total_Users FROM fb;
Query ID = root_20210812184526_d4917bd3-5ada-4f19-a9ec-c31f45892269
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1628785648210_0003)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.69 s
--------------------------------------------------------------------------------
OK
99003
Time taken: 26.01 seconds, Fetched: 1 row(s)
hive>
```

Find out the number of Facebook users above the age of 25

select count(userid) AS totalUsers from fb where age>25;

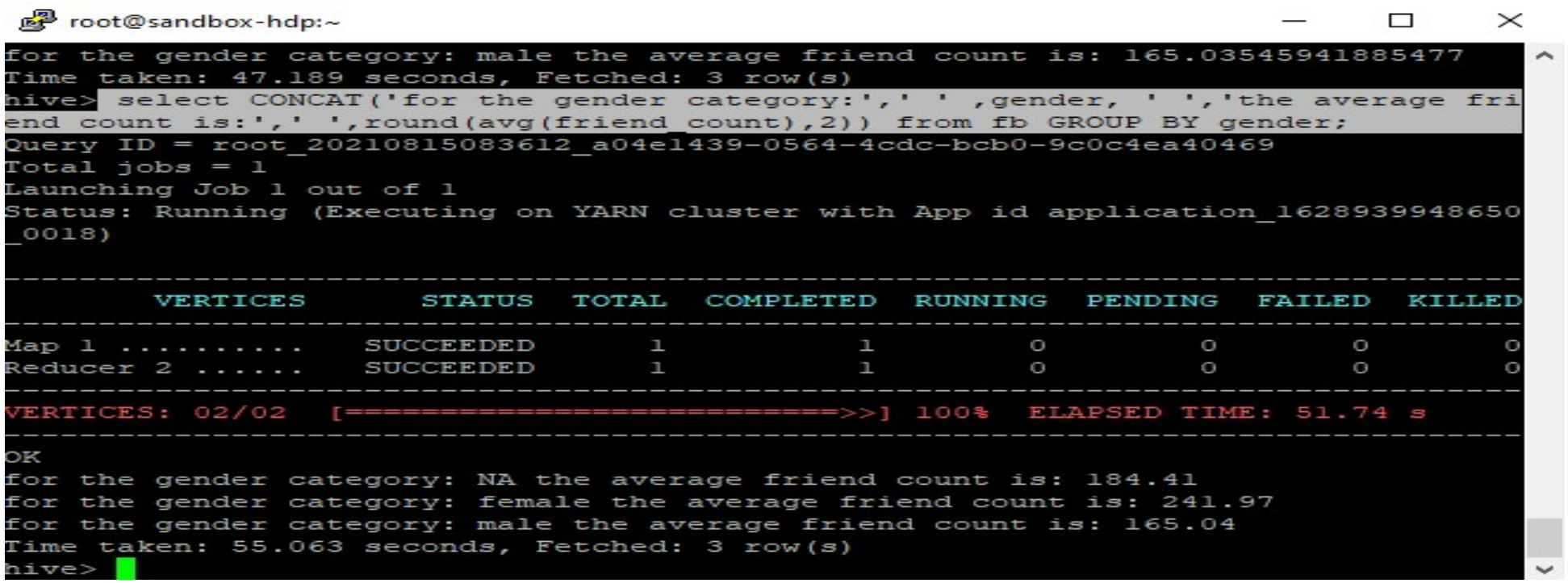
```
hive> SELECT COUNT(userid) AS totalUsers FROM fb WHERE age>25;
Query ID = root_20210812185612_db2d1586-a64b-44b8-9d23-dfc9e9e84bc2
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1628785648210_0003)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 24.77 s
----------------------------------------------------------------------------------------------
OK
totalusers
56676
Time taken: 27.924 seconds, Fetched: 1 row(s)
hive>
```

Do male Facebook users tend to have more friends , or female users?

```
select avg(friends) from fb where gender= 'male' ;
Select avg(friends) from fb where gender = 'female' ;
Select gender , avg(friends) from fb group by gender;
```



How many likes do young people receive on Facebook opposed to older members

select avg(likes_received) AS avg_likes from fb where age>25 ;

```
hive> select avg(likes_received) AS avg_likes_received from fb where age>25;
Query ID = root_20210812192629_1083ae60-6e88-4cb1-8738-ace2f499858c
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1628785648210_0003)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.19 s
----------------------------------------------------------------------------------------------
OK
avg_likes_received
99.67402427835415
Time taken: 24.758 seconds, Fetched: 1 row(s)
hive>
```

Find out the count of Facebook users for each birthday month

select month, count(*) from fb group by month;

```
--------------------------------------------------------------------------------
OK
200.29834384671722
Time taken: 11.018 seconds, Fetched: 1 row(s)
hive> select month, count(*) from fb group by month;
Query ID = root_20210823175611_141cc80a-274e-4805-9b11-cb315a21a4cc
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629707043374_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.10 s
--------------------------------------------------------------------------------
OK
1       11772
2       7632
3       8110
4       7810
5       8271
7       8021

8       8266
9       7939
10      8476
11      7205
12      7894
Time taken: 9.56 seconds, Fetched: 12 row(s)
hive> select avg(mlikes) from fb  where age>=13 AND age<=25;
Query ID = root_20210823175737_5cbba1d1-2c2b-4c53-bfba-9602d7f326b4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1629707043374_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.65 s
--------------------------------------------------------------------------------
OK
124.01923122356888
Time taken: 9.031 seconds, Fetched: 1 row(s)
hive>                 Select avg(wlikes) from fb where age>=13 AND age<=25;
```

Do young members use mobile phones or computer for Facebook browsing

```
OK
age       mobile_likes     computer_likes
13        137.13   63.93
14        117.7    62.86
15        150.69   75.48
16        143.86   65.41
17        147.37   84.32
18        152.26   69.17
19        132.6    68.35
20        127.8    69.94
21        108.95   44.37
22        99.31    41.07
23        96.58    37.78
24        102.4    25.98
25        99.41    19.37
Time taken: 27.359 seconds, Fetched: 13 row(s)
hive>
```

root@s...

```
hive> set hive.cli.print.header=true;
hive> SELECT age,round(avg(mobile_likes),2) AS mobile_likes, round(avg(computer_
likes),2) AS computer_likes FROM fb WHERE age<=25 AND age>=13 group by age;
Query ID = root_20210818172037_2a2cd72a-fe78-4e4b-8f2a-c6871ab897d8
```