

FACEBOOK DATA ANALYSIS

1. Uploading Facebook dataset from my local system into HDFS which is Hadoop Distributed File System.

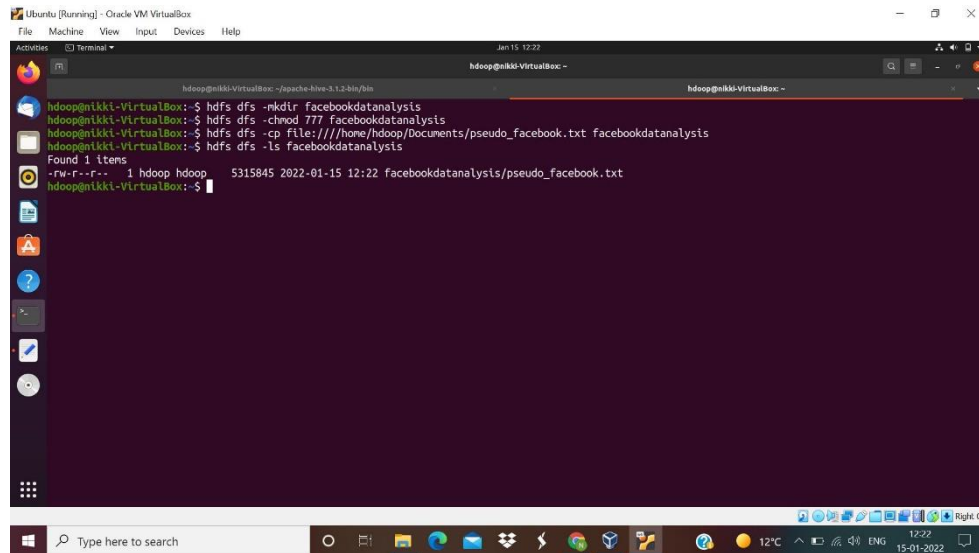
Command:

```
hdfs dfs -mkdir facebookdatanalysis
```

```
hdfs dfs -chmod 777 facebookdatanalysis
```

```
hdfs dfs -cp file:///home/hadoop/Documents/pseudo\_facebook.txt facebookdatanalysis
```

```
hdfs dfs -ls facebookdatanalysis
```



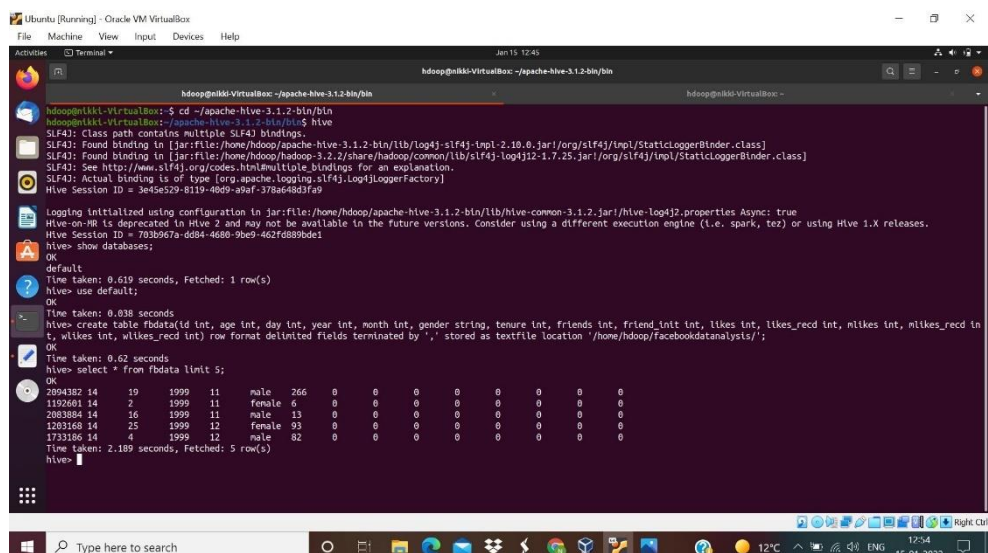
```
hadoop@hadoop-VirtualBox: ~$ hdfs dfs -mkdir facebookdatanalysis
hadoop@hadoop-VirtualBox: ~$ hdfs dfs -chmod 777 facebookdatanalysis
hadoop@hadoop-VirtualBox: ~$ hdfs dfs -cp file:///home/hadoop/Documents/pseudo_facebook.txt facebookdatanalysis
hadoop@hadoop-VirtualBox: ~$ hdfs dfs -ls facebookdatanalysis
Found 1 items
-rw-r--r-- 1 hadoop hadoop 5315845 2022-01-15 12:22 facebookdatanalysis/pseudo_facebook.txt
```

2. Starting hive in my system

Creating table inside default database to store the dataset (i.e Facebook data set) inside that table

Command to create table inside hive:

```
create table fbdata(id int, age int, day int, year int, month int, gender string, tenure int, friends int, friend_init int, likes int, likes_recd int, mlikes int, mlikes_recd int, wlikes int, wlikes_recd int) row format delimited fields terminated by ',' stored as textfile location '/home/hadoop/facebookdatanalysis/';
```



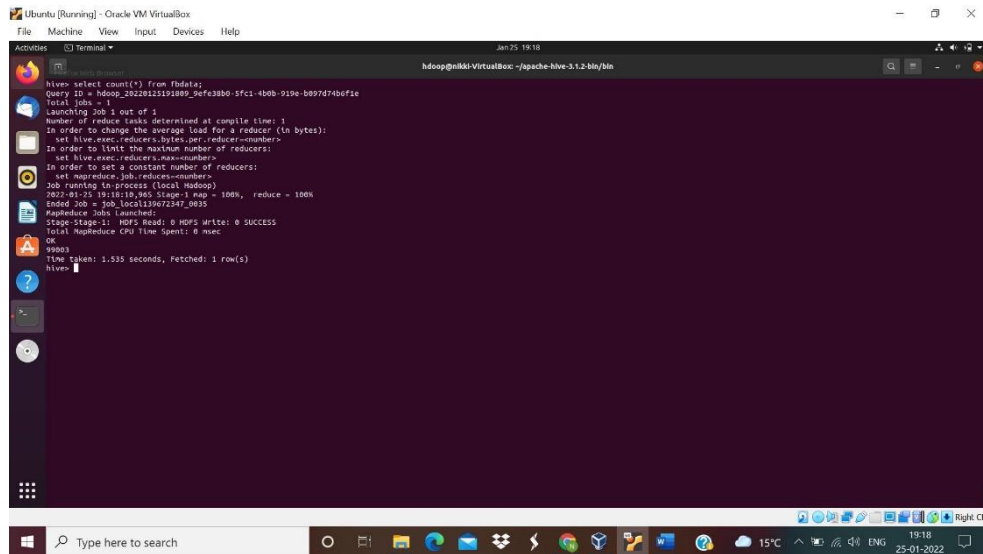
```
hadoop@hadoop-VirtualBox: ~$ cd ~/apache-hive-3.12.0-bin/bin
hadoop@hadoop-VirtualBox: ~/apache-hive-3.12.0-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/apache-hive-3.12.0-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop-3.2.2/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.log4j.Log4jLoggerFactory]
Hive Session ID = 3e45e529-8119-48d9-a9af-378a648d3fa9
Logging initialized using configuration in jar:file:/home/hadoop/apache-hive-3.12.0-bin/lib/hive-common-3.12.0.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Hive Session ID = 703b967a-d884-4688-9be9-462f889bde1
hive> show databases;
OK
default
hive> use default;
OK
Time taken: 0.619 seconds, Fetched: 1 row(s)
hive> create table fbdata(id int, age int, day int, year int, month int, gender string, tenure int, friends int, friend_init int, likes int, likes_recd int, mlikes int, mlikes_recd int, wlikes int, wlikes_recd int) row format delimited fields terminated by ',' stored as textfile location '/home/hadoop/facebookdatanalysis/';
OK
Time taken: 0.62 seconds
hive> select * from fbdata limit 5;
OK
2094382 14 19 1999 11 male 265 0 0 0 0 0 0 0 0 0
1192681 14 2 1999 11 female 6 0 0 0 0 0 0 0 0 0
2083884 14 16 1999 11 male 13 0 0 0 0 0 0 0 0 0
1203168 14 25 1999 12 female 93 0 0 0 0 0 0 0 0 0
1732186 14 4 1999 12 male 82 0 0 0 0 0 0 0 0 0
Time taken: 2.189 seconds, Fetched: 5 row(s)
hive>
```

HIVE QUERY

1. Total number of users in his dataset

Query:

→ Select count(*) from fbdata;



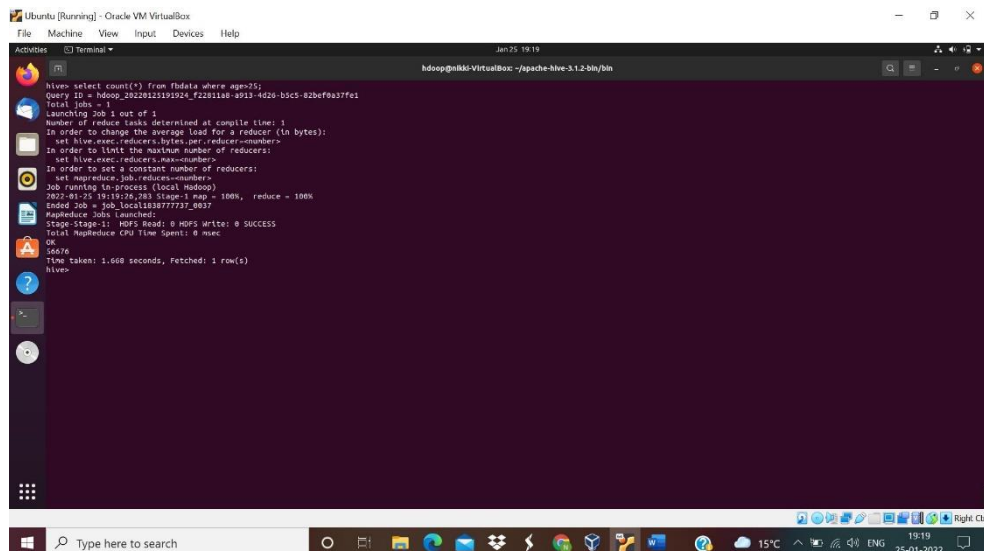
```
hadoop@nllki-VirtualBox: ~/apache-hive-3.12.0-bin$
hive> select count(*) from fbdata;
Query ID = hdoop_20220125191924_f2281188-4913-4d26-b3c5-82bef9a37fe1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (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)
2022-01-25 19:18:18,945 Stage-1 map = 100%, reduce = 100%
Ended job = job_local18071347_0035
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
99003
Time taken: 1.535 seconds, Fetched: 1 row(s)
hive>
```

Result: There are total 99003 users in our dataset

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

Query:

→ Select count(*) from fbdata where age>25;



```
hadoop@nllki-VirtualBox: ~/apache-hive-3.12.0-bin$
hive> select count(*) from fbdata where age>25;
Query ID = hdoop_20220125191924_f2281188-4913-4d26-b3c5-82bef9a37fe1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
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)
2022-01-25 19:19:26,283 Stage-1 map = 100%, reduce = 100%
Ended job = job_local1838777737_0037
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
56676
Time taken: 1.668 seconds, Fetched: 1 row(s)
hive>
```

Result: There are 56676 users that are above the age 25

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

Query:

→ Select avg(friends) from fbdata where gender='male';

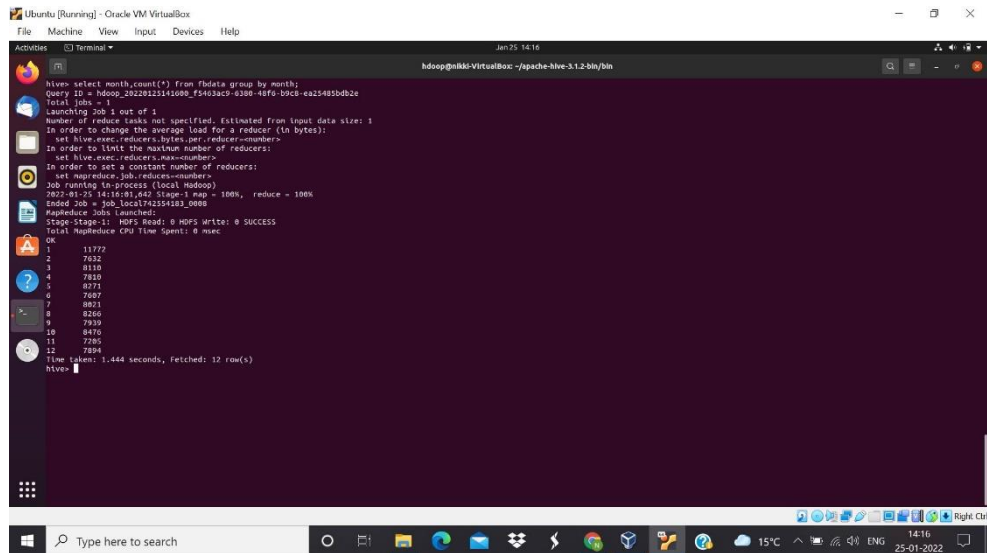
→ Select avg(friends) from fbdata where gender=' female';

Result: Young people receive more like than older people

5. Find out the count of Facebook users for each birthday month

Query:

→ select month, count(*) from fbdata group by month;

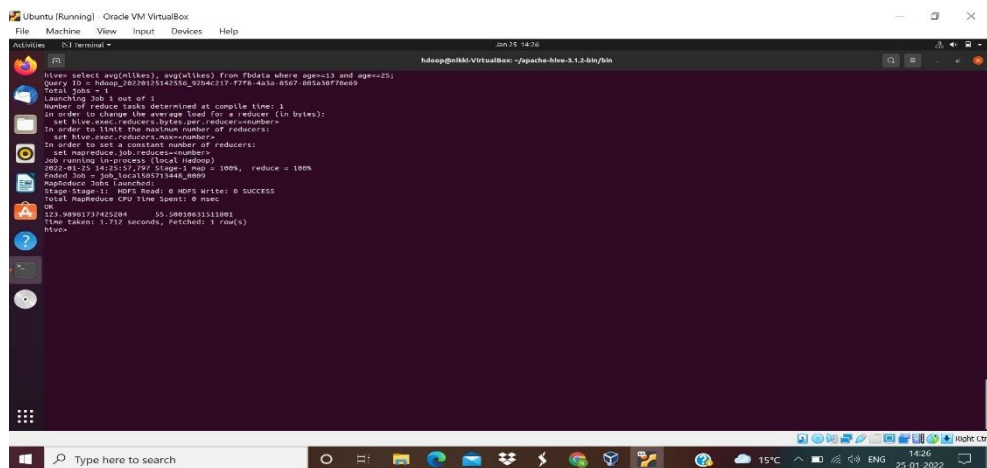


```
hadoop@hadoop-VirtualBox: ~$ hivesql
hive> select month, count(*) from fbdata group by month;
Query ID = h00p_20220125141608_f5463ac9-6380-48f6-b9cd-ca25483db2e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average 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.reducers=number
Ended Job = job_local742554283_0000
MapReduce jobs launched:
Stage-1: Map: 0, Reduce: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1 11772
2 7632
3 8110
4 7810
5 8271
6 7687
7 8821
8 8269
9 7939
10 8470
11 7285
12 7894
Time taken: 1.444 seconds, Fetched: 12 row(s)
hive>
```

6. Do young members use mobile phone or computers for Facebook browsing?

Query:

→ select avg(mlikes), avg(wlikes) from fbdata where age>=13 and age<=25;



```
hadoop@hadoop-VirtualBox: ~$ hivesql
hive> select avg(mlikes), avg(wlikes) from fbdata where age>=13 and age<=25;
Query ID = h00p_20220125141615_516_9784c17f-77f8-4a1a-a167-8d1a3877e0d9
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
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.reducers=number
Ended Job = job_local568713448_0000
MapReduce jobs launched:
Stage-1: Map: 0, Reduce: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
123.98901737425204 55.58018031511801
Time taken: 1.717 seconds, Fetched: 1 row(s)
hive>
```

Result: From the above observation young people use mobile phone for Facebook browsing than computers;

7. Do older members use mobile phone or computer for facebook browsing?

Query:

→ select avg(mlikes), avg(wlikes) from fbdata where age>=35;

```

hadoop@hadoop:~/hadoop$ hivesql
hive> select avg(likes), avg(likes) from fbdata where age=35;
Query ID = hadoop_20220125142737_f6299064-60e8-4019-b80a-41410805f093
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
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)
2022-01-25 14:27:39,111: Stage-1 map = 100%, reduce = 100%
Ended Job = job_local115181279_0010
Mapreduce jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total Mapreduce CPU Time Spent: 0 msec
OK
1
1
Time taken: 1.407 seconds, fetched: 1 row(s)
hive>

```

Result: From the above observation older people use mobile phone for Facebook browsing than computers;

8. Find out number of people in each age group

Query:

→ select age, count(*) from fbdata group by age;

```

hadoop@hadoop:~/hadoop$ hivesql
hive> select age, count(*) from fbdata group by age;
Query ID = hadoop_20220125143310_f037230f-1442-410f-954f-7836757a325e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified, estimated from input data size: 1
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)
2022-01-25 14:33:17,978: Stage-1 map = 100%, reduce = 100%
Ended Job = job_local1776240072_0011
Mapreduce jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total Mapreduce CPU Time Spent: 0 msec
OK
age      count(*)
13       486
14       1925
15       2618
16       3880
17       3283
18       5190
19       4191
20       3769
21       3671
22       3932
23       4404
24       2827
25       3641
26       2815
27       2260
28       2364
29       1936
30       1710
31       1694
32       1443
33       1999
34       1127
35       1175
36       1118
37       968
38       1899
39       962
40       815
41       883

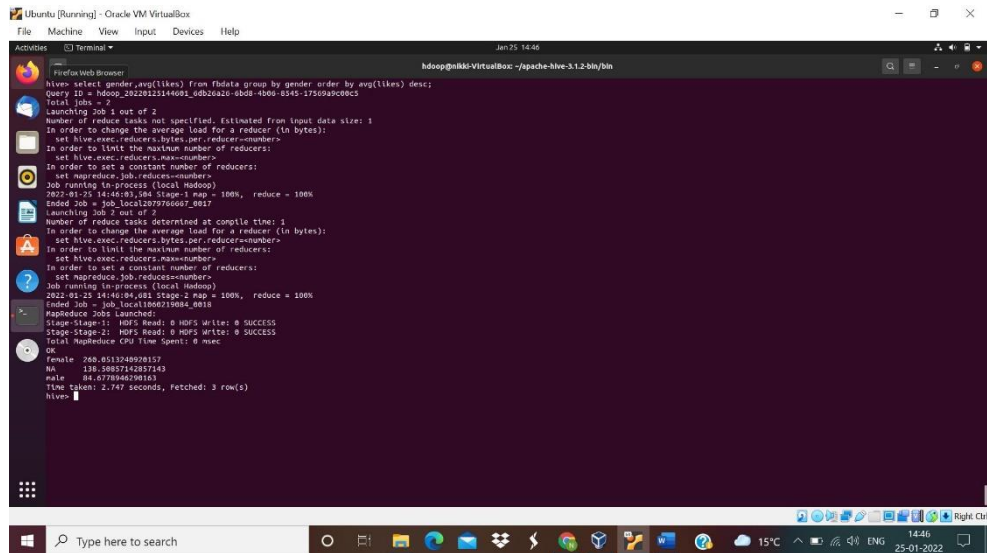
```

Result: Facebook is most popular between age groups 16 and 26.

9. Who give more likes male or female?

Query:

→ select gender, avg(likes) from fbdata group by gender order by avg(likes) desc;



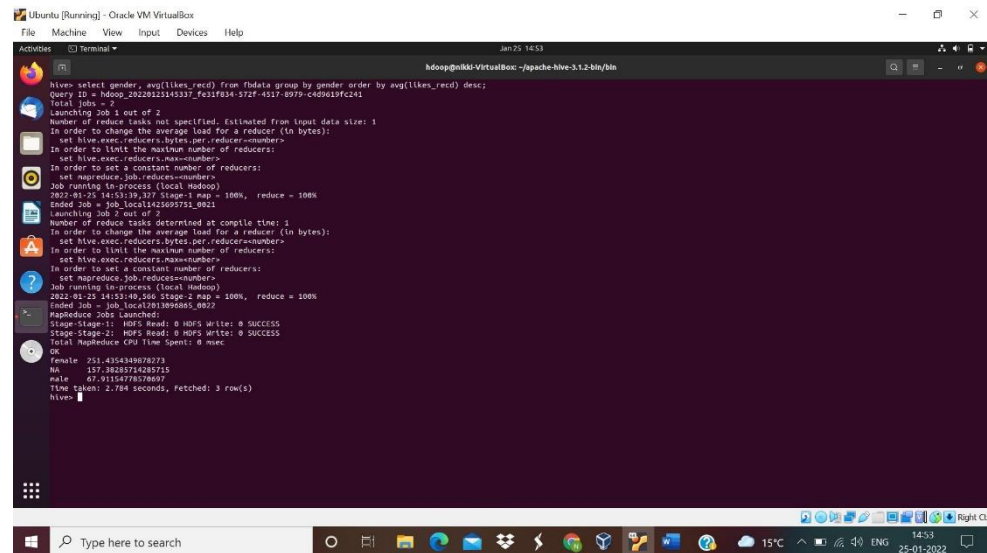
```
hadoop@nikki-VirtualBox: ~/apache-hive-3.12-bin/bin
hive> select gender, avg(likes) from fbdata group by gender order by avg(likes) desc;
Query ID = hdoop_20220125144001_0db26a26-4bd8-4b06-8545-175699c08c5
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
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)
2022-01-25 14:40:03,646 Stage:1 Map = 100%, reduce = 100%
Ended Job = Job_local2079760607_0017
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
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)
2022-01-25 14:40:04,061 Stage:2 Map = 100%, reduce = 100%
Ended Job = Job_local106070984_0018
MapReduce Jobs Launched:
Stage:Stage=1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Stage:Stage=2: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Female 250.0513709070157
Male 138.50857442057143
Time taken: 2.747 seconds, Fetched: 3 row(s)
hive>
```

Result: Female give more likes then men.

10. Who receive more likes?

Query:

➔ select gender, avg(likes_recd) from fbdata group by gender order by avg(likes_recd) desc;



```
hadoop@nikki-VirtualBox: ~/apache-hive-3.12-bin/bin
hive> select gender, avg(likes_recd) from fbdata group by gender order by avg(likes_recd) desc;
Query ID = hdoop_20220125145337_f631f034-572f-4217-b979-cd0621fc241
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
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)
2022-01-25 14:53:39,327 Stage:1 Map = 100%, reduce = 100%
Ended Job = Job_local162499731_0021
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
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)
2022-01-25 14:53:40,566 Stage:2 Map = 100%, reduce = 100%
Ended Job = Job_local199389685_0022
MapReduce Jobs Launched:
Stage:Stage=1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Stage:Stage=2: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Female 251.4354349870273
Male 137.2628714205715
Time taken: 2.784 seconds, Fetched: 3 row(s)
hive>
```

Result: Females receive more likes than men.

11. Gender count

Query:

➔ select gender, count(*) from fbdata group by gender order by count(*) desc;


```

hadoop@hadoop-VirtualBox: ~/apache-hive-3.12.0/bin
hive> select gender, count(*) from fbdata group by gender order by count(*) desc;
Query ID = hdoop_20220121170847_e4b01717-7bdc-432a-b0ff-4c946cccfcb1
Total Jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
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.reducers=<number>
Job running in-process (local Hadoop)
2022-01-25 17:08:49,216: Stage-1 map = 100%, reduce = 100%
Ended Job = job_local22981007_0025
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
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.reducers=<number>
Job running in-process (local Hadoop)
2022-01-25 17:08:59,430: Stage-2 map = 100%, reduce = 100%
Ended Job = job_local22981007_0026
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Stage-Stage-2: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
male      175
female    48254
NA         175
Time taken: 3.172 seconds, Fetched: 3 row(s)
hive>

```

Result: There are more male users than female.

12. Likes split up

Query:

→ select gender, avg(mlikes), avg(mlikes_recd), avg(wlikes), avg(wlikes_recd) from fbdata where gender <> 'NA' group by gender;

```

hadoop@hadoop-VirtualBox: ~/apache-hive-3.12.0/bin
hive> select gender, avg(mlikes), avg(mlikes_recd), avg(wlikes), avg(wlikes_recd) from fbdata where gender <> 'NA' group by gender;
Query ID = hdoop_202201211712309_1f9a6c4b-50b4-437b-814b-2e18f26c202
Total Jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
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.reducers=<number>
Job running in-process (local Hadoop)
2022-01-25 17:12:31,440: Stage-1 map = 100%, reduce = 100%
Ended Job = job_local166127589_0033
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
male      177.91292790776556      147.1888883041606      87.13829681122733      184.3144112465991
female    48.2517751547735      48.8339146491374      24.416508895121728      27.07953311759957
Time taken: 1.492 seconds, Fetched: 2 row(s)
hive>

```

Result: Interesting observation for gender specific interaction with Facebook: women likes are a lot more than man.

13. Friend counts and Friendship initiated

Query:

→ select gender, avg(friends), avg(friend_init) from fbdata group by gender;

```
hadoop@hdfs-VirtualBox: ~/apache-hive-3.12.0/bin
hive> select gender, avg(friends), avg(friend_init) from fbdata group by gender;
Query ID = hdoop_20220125172009_474ba3ce-e107-46d3-99cf-346c3f6e9db
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
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.reducers=<number>
Job running in-process (local Hadoop)
3027-9125 17:20:11,248: User-1 map = 100%, reduce = 100%
Ended Job = job_local308783388_0030
Mapreduce jobs Launched:
Stage-Stage-1: HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK

```

gender	avg(friends)	avg(friend_init)
MA	184.41142857142856	92.57142857142857
Female	245.56944444444445	111.89090909090909
male	185.63545454545454	101.66659955114324

Time taken: 1.402 seconds, fetched: 3 row(s)
hive>

Result: Women have more friends than men on facebook, the friendships initiated in proportion to friend count are more in case of men than women.