Proof of Concept

On

Title: FaceBook Data Analysis

Submitted for the requirement of

Big Data Engineering Course

BACHELOR OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING
(Big Data and Analytics)
CSC-334

Semester-6

Submitted to:
Ms. Gurpreet Kaur
Project Supervisor
Submitted by:
Arman Gupta(18BCS3759)
Vishavdeep Singh Aulakh(18BCS3767)

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we give to our 5th semester B.D.E project supervisor, Ms. Gurpreet Kaur, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project and especially in writing this report. Furthermore, we would also like to acknowledge with much appreciation her crucial role, in giving the permission to use all required equipment and the necessary materials to complete the task 'Analysis of Facebook Dataset' using Apache Hadoop (Hive), IBM Cognos Insight and Python Programming, and gave suggestion about the task.

OVERVIEW

Facebook has not only changed how we communicate to each other, but how we collect data for the benefit of our business. As opposed to big budget ad campaigns that often become ineffective due to no direction, Facebook has refined its advertising mechanism so that target users will see your product and enjoy it. These advancements in online marketing have made it possible to interact when more data is collected from the users, which is opposed to the days of the user data being stored provided little to no avenues of strategy for the marketer. With this being said, here's a brief look at how Facebook Data Analytics benefit not only how a company invests into marketing… but the effectiveness of their marketing strategies in relations to the customer

OBJECTIVES

Brand awareness

Increase overall awareness for your brand by showing ads to people who are more likely to pay attention to them.

Works well with: ad recall lift

Reach

Show ads to the maximum number of people in your audience while staying within your budget. You can also choose to reach only people who are near your business locations.

COLUMNS AND DATA TYPE :

Age int
Id 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

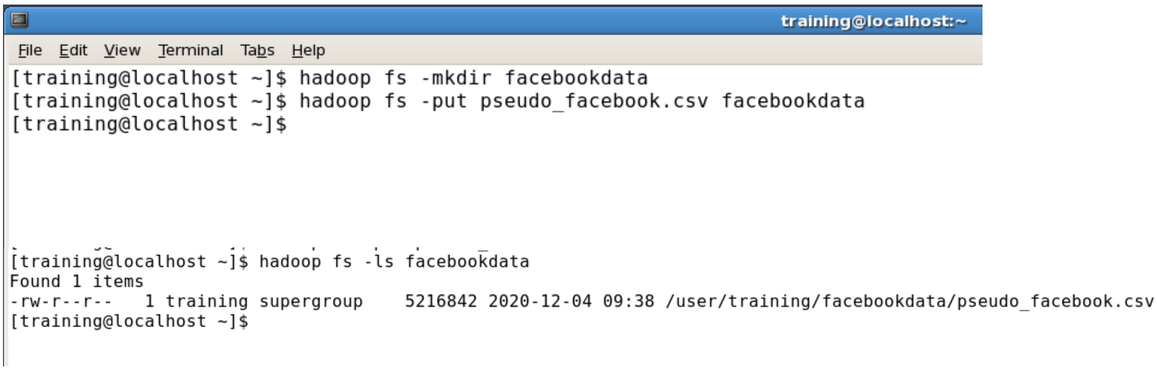
PROBLEM STATEMENTS:

1. Find the total number of users in this dataset.
2. Find out the number of Facebook users above the age of 25.
3. Do male Facebook users tend to have more friends ,or female users?
4. How many likes do young people receive on Facebook opposed to older members
5. Find out the count of Facebook users for each birthday month.
6. Do young members use mobile phones or computers for Facebook browsing?
7. Do adult members use mobile phones or computers for Facebook browsing
8. Visualisation graph for the age wise number of people on Facebook.

9. Visualisation for the number of likes which was received by male and female.
10. Visualisation for the likes received for the age of the users (Male or Female)

HIVE QUERIES

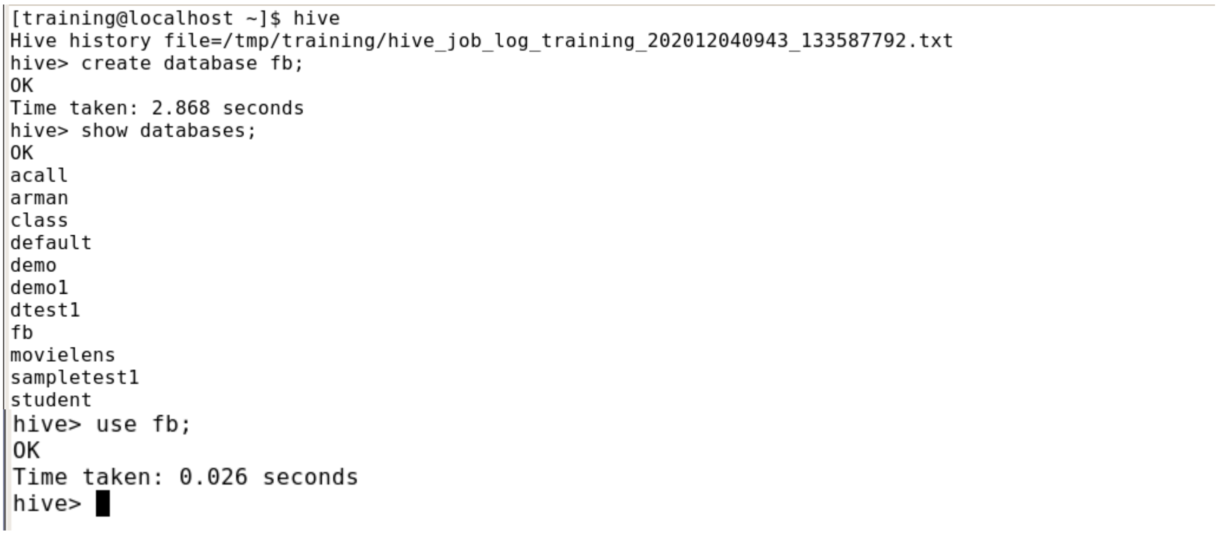
1. Create a directory and copy the data in it.



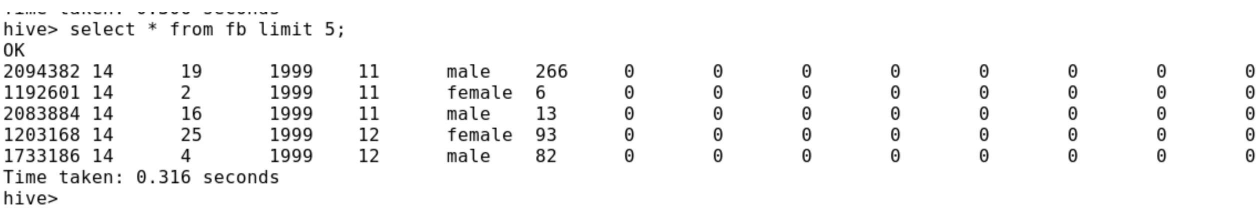
```
[training@localhost ~]$ hadoop fs -mkdir facebookdata
[training@localhost ~]$ hadoop fs -put pseudo_facebook.csv facebookdata
[training@localhost ~]$
```

```
[training@localhost ~]$ hadoop fs -ls facebookdata
Found 1 items
-rw-r--r--   1 training supergroup    5216842 2020-12-04 09:38 /user/training/facebookdata/pseudo_facebook.csv
[training@localhost ~]$
```

2. Creating a database and use it.



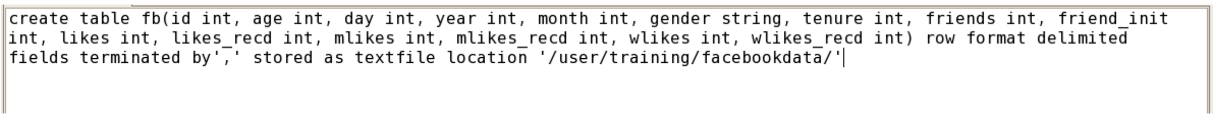
```
[training@localhost ~]$ hive
Hive history file=/tmp/training/hive_job_log_training_202012040943_133587792.txt
hive> create database fb;
OK
Time taken: 2.868 seconds
hive> show databases;
OK
acall
arman
class
default
demo
demo1
dtest1
fb
movielens
sampletest1
student
hive> use fb;
OK
Time taken: 0.026 seconds
hive> 
```



```
hive> select * from fb limit 5;
OK
2094382 14      19      1999    11      male    266     0       0       0       0       0       0       0       0
1192601 14      2       1999    11      female  6       0       0       0       0       0       0       0       0
2083884 14      16      1999    11      male    13      0       0       0       0       0       0       0       0
1203168 14      25      1999    12      female  93      0       0       0       0       0       0       0       0
1733186 14      4       1999    12      male    82      0       0       0       0       0       0       0       0
Time taken: 0.316 seconds
hive>
```

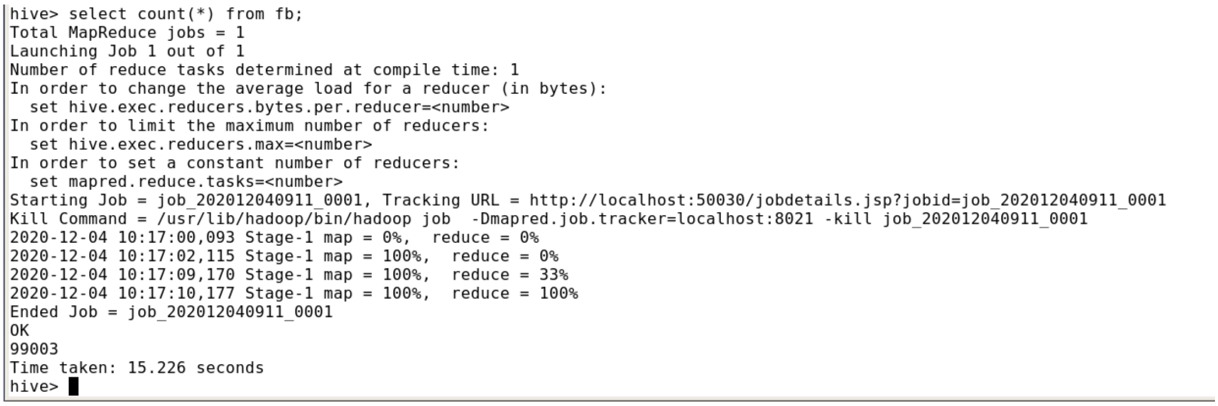
3. Displaying the top 5 rows of uploaded data:

4. Creating a hive table:



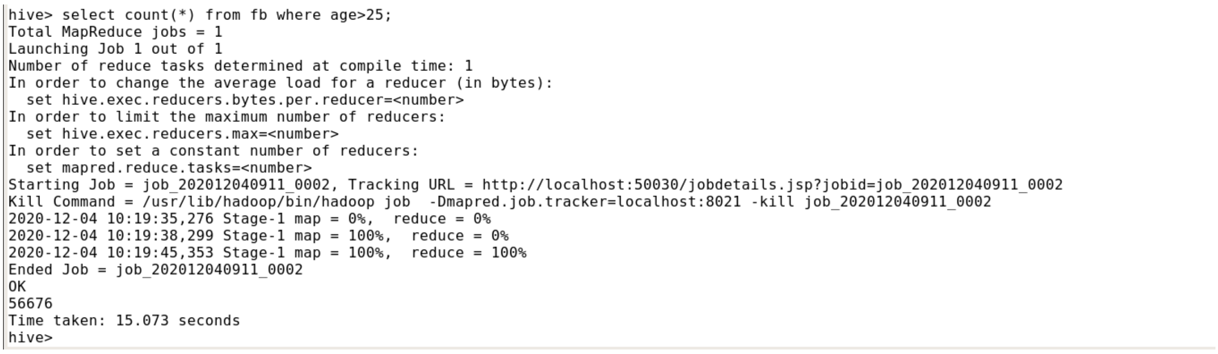
```
create table fb(id int, age int, day int, year int, month int, gender string, tenure int, friends int, friend_init
int, likes int, likes_recd int, mlikes int, mlikes_recd int, wlikes int, wlikes_recd int) row format delimited
fields terminated by',' stored as textfile location '/user/training/facebookdata/'
```

PROBLEM STATEMENT 1: Find the total number of users in this dataset.



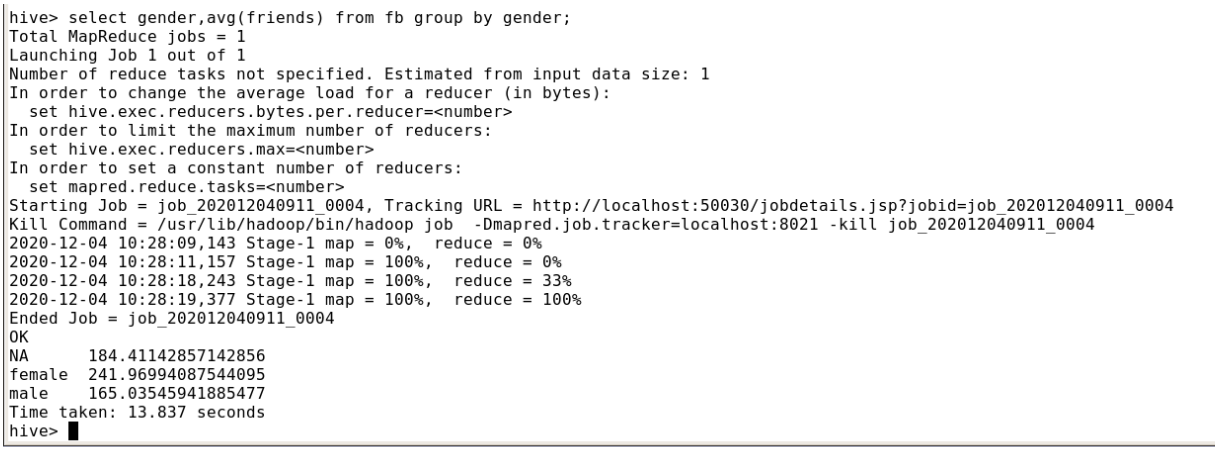
```
hive> select count(*) from fb;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0001, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0001
2020-12-04 10:17:00,093 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:17:02,115 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:17:09,170 Stage-1 map = 100%,  reduce = 33%
2020-12-04 10:17:10,177 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0001
OK
99003
Time taken: 15.226 seconds
hive> █
```

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



```
hive> select count(*) from fb where age>25;
Total MapReduce 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 mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0002, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0002
2020-12-04 10:19:35,276 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:19:38,299 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:19:45,353 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0002
OK
56676
Time taken: 15.073 seconds
hive>
```

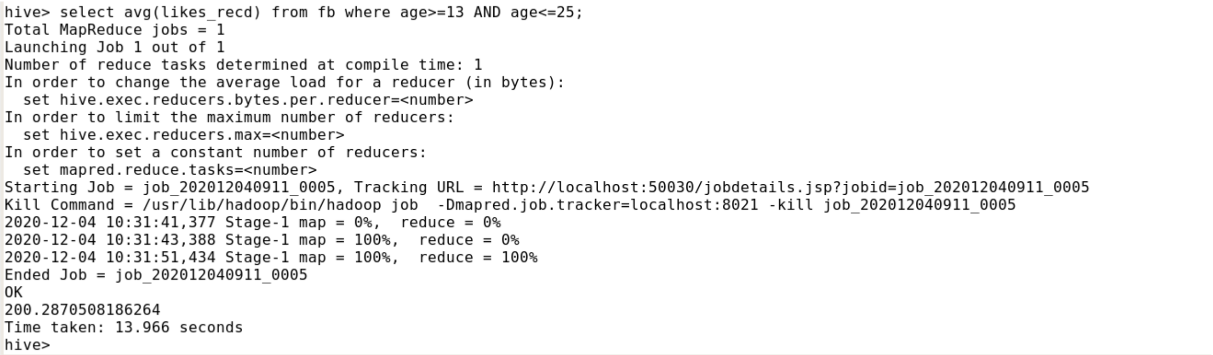
PROBLEM STATEMENT 3:Do male Facebook users tend to have more friends ,or female users?



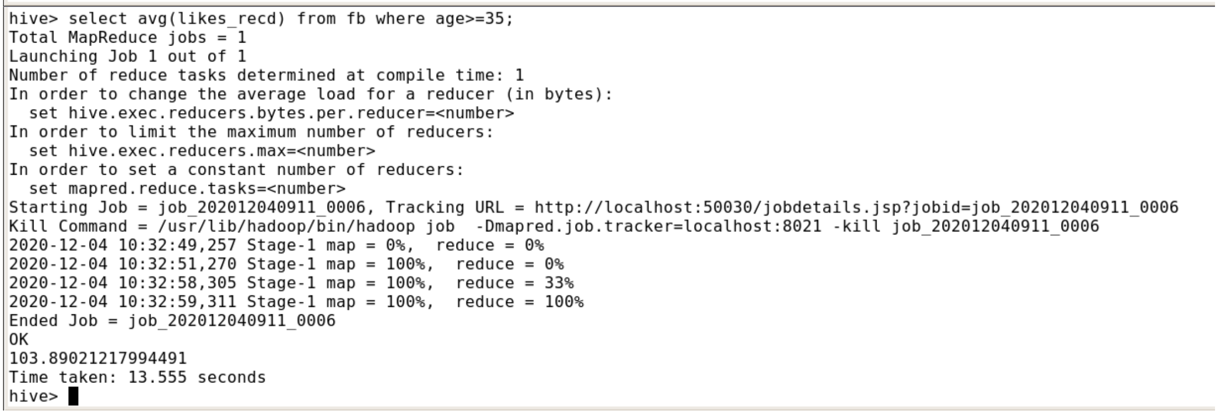
```
hive> select gender,avg(friends) from fb group by gender;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0004, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0004
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0004
2020-12-04 10:28:09,143 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:28:11,157 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:28:18,243 Stage-1 map = 100%,  reduce = 33%
2020-12-04 10:28:19,377 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0004
OK
NA      184.41142857142856
female  241.96994087544095
male    165.03545941885477
Time taken: 13.837 seconds
hive> █
```

This result is as expected and quite obvious female receives more like then male .So brand or product can select the girl or lady who received most of the likes or more socially active on Facebook for brand promotion.

PROBLEM STATEMENT 4:How many likes do young people receive on Facebook opposed to older members ?



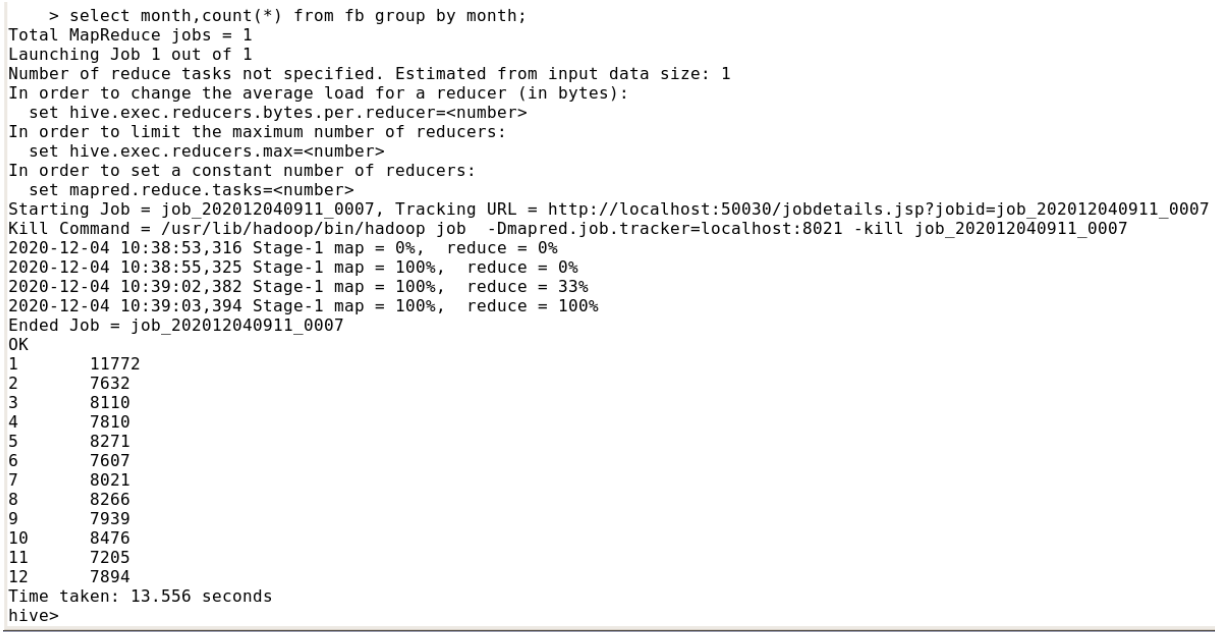
```
hive> select avg(likes_recd) from fb where age>=13 AND age<=25;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0005, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0005
2020-12-04 10:31:41,377 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:31:43,388 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:31:51,434 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0005
OK
200.2870508186264
Time taken: 13.966 seconds
hive>
```



```
hive> select avg(likes_recd) from fb where age>=35;
Total MapReduce 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 mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0006, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0006
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0006
2020-12-04 10:32:49,257 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:32:51,270 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:32:58,305 Stage-1 map = 100%,  reduce = 33%
2020-12-04 10:32:59,311 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0006
OK
103.89021217994491
Time taken: 13.555 seconds
hive> █
```

We’ve use average function as we use in sql. We can also take sum but if any outliers is present in dataset so error can be occur . So the result clearly shows that number of young people is more than the old people .

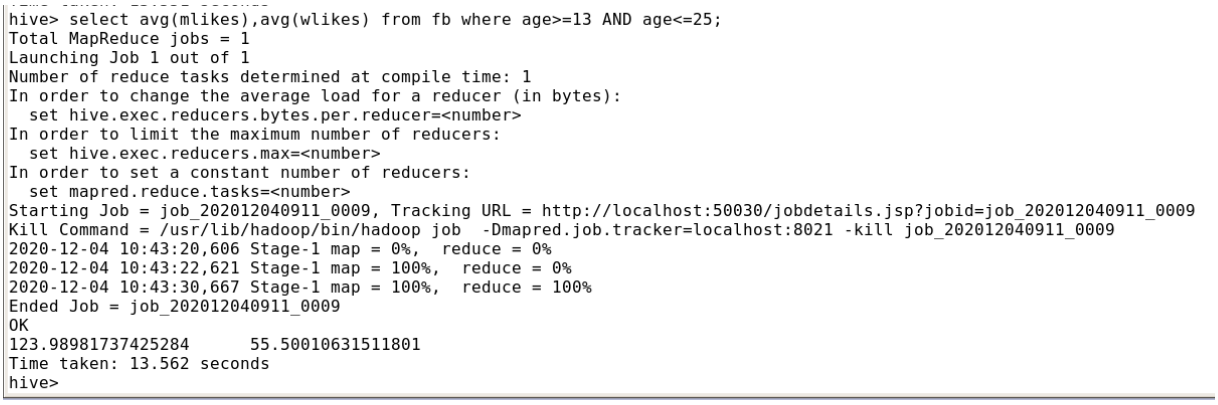
PROBLEM STATEMENT 5:Find out the count of Facebook users for each birthday month.



```
    > select month,count(*) from fb group by month;
Total MapReduce 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 mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0007, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0007
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0007
2020-12-04 10:38:53,316 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:38:55,325 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:39:02,382 Stage-1 map = 100%,  reduce = 33%
2020-12-04 10:39:03,394 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0007
OK
1       11772
2       7632
3       8110
4       7810
5       8271
6       7607
7       8021
8       8266
9       7939
10      8476
11      7205
12      7894
Time taken: 13.556 seconds
hive>
```

From the above result we can say that most number of users created their account in the month of January so the best time for brand promotion can be considered as January .

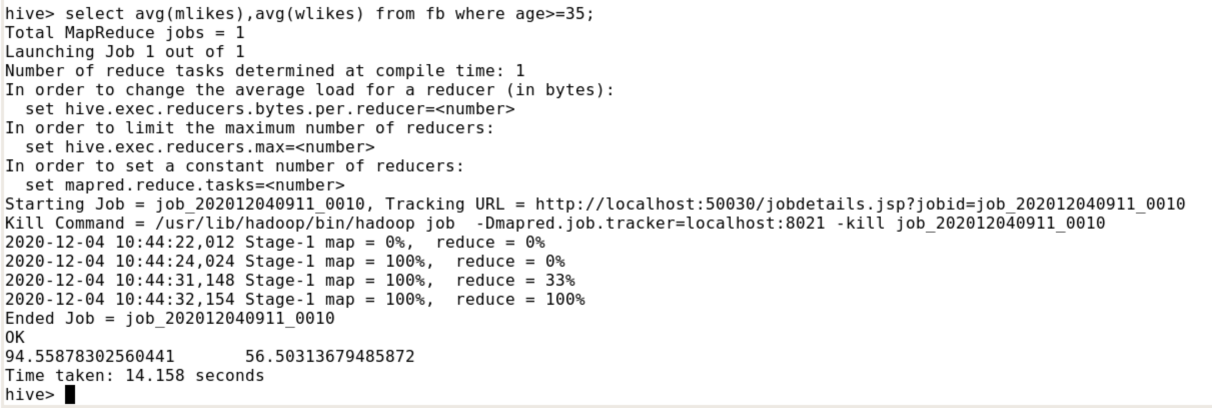
PROBLEM STATEMENT 6.Do young members use mobile phones or computers for Facebook browsing?



```
hive> select avg(mlikes),avg(wlikes) from fb where age>=13 AND age<=25;
Total MapReduce 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 mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0009, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0009
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0009
2020-12-04 10:43:20,606 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:43:22,621 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:43:30,667 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0009
OK
123.98981737425284      55.50010631511801
Time taken: 13.562 seconds
hive>
```

Thus, it can be seen that young members use mobile phones for using Facebook instead of using computers so we can display our ads on mobile phones.

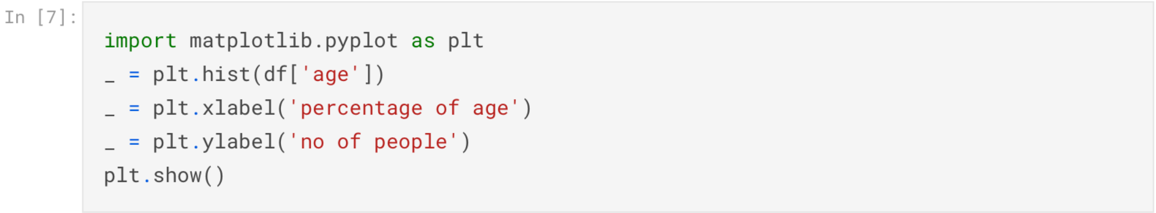
PROBLEM STATEMENT 7:Do adult members use mobile phones or computers for Facebook browsing?



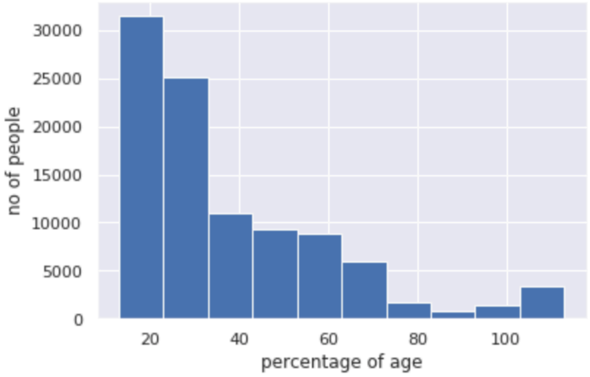
```
hive> select avg(mlikes),avg(wlikes) from fb where age>=35;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202012040911_0010, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202012040911_0010
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_202012040911_0010
2020-12-04 10:44:22,012 Stage-1 map = 0%,  reduce = 0%
2020-12-04 10:44:24,024 Stage-1 map = 100%,  reduce = 0%
2020-12-04 10:44:31,148 Stage-1 map = 100%,  reduce = 33%
2020-12-04 10:44:32,154 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202012040911_0010
OK
94.55878302560441       56.50313679485872
Time taken: 14.158 seconds
hive> 
```

I thought that adult users that is above the age of 35 mostly prefer computers for Facebook use, but the result is shocking as we can say adult also prefer mobile phones for use of Facebook but the number of adult users are less then young users .

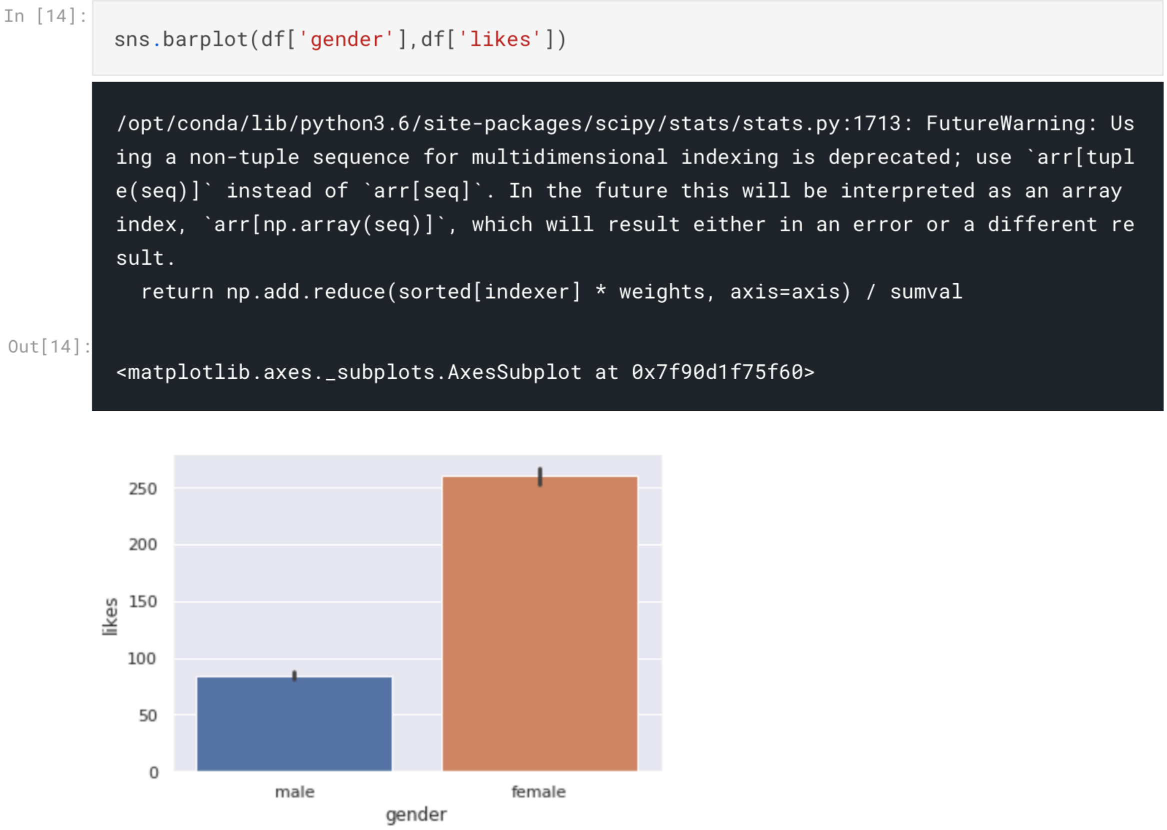
PROBLEM STATEMENT 8:Visualisation graph for the age wise number of people on Facebook :



```
In [7]:
import matplotlib.pyplot as plt
_ = plt.hist(df['age'])
_ = plt.xlabel('percentage of age')
_ = plt.ylabel('no of people')
plt.show()
```



PROBLEM STATEMENT 9:Visualisation for the number of likes which was received by male and female:



In [14]:

```
sns.barplot(df['gender'],df['likes'])
```

```
/opt/conda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Us
ing a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tupl
e(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array
index, `arr[np.array(seq)]`, which will result either in an error or a different re
sult.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f90d1f75f60>
```

PROBLEM STATEMENT 10:Visualisation for the likes received for the age of the users (Male or Female):

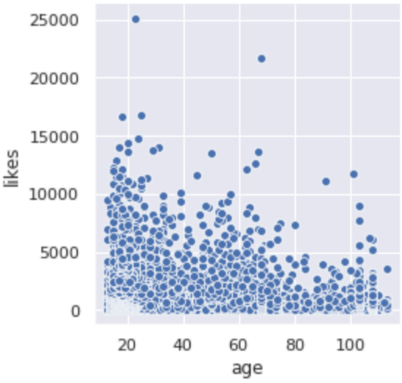
In [11]:

```
sns.pairplot(df,x_vars=["age"],y_vars="likes",size=4)
```

```
/opt/conda/lib/python3.6/site-packages/seaborn/axisgrid.py:2065: UserWarning: The `
size` parameter has been renamed to `height`; pleaes update your code.
  warnings.warn(msg, UserWarning)
```

Out[11]:

```
<seaborn.axisgrid.PairGrid at 0x7f90c9219dd8>
```


likes
age