PROOF OF CONCEPT

ON

TITLE: INSTAGRAM DATA ANALYSIS

Submitted for the requirement of

Big Data Engineering Course

BACHELOR OF ENGINEERING

(Big Data and Analytics)

SHRI RAMSWAROOP
MEMORIAL COLLEGE OF ENGINEERING & MANAGEMENT

IT-4B

Semester-5

Submitted to:

Ms. Gurpreet Kaur
Project Supervisor

Submitted by:

Rishi Naudiyal(2001220130088)
Satish Maurya(2001220130094)

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 ac knowledge with much appreciation her crucial role, in giving the permission to use all required equipment and the necessary materials to complete the task sight and Python Programming, and gave suggestion about the task.

OVERVIEW

Instagram 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, Instagram 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 Instagram 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.

DATA-SET USED

https://drive.google.com/file/d/1NX2RO3-jWK6jyk_WxDQOWIRgBAwmFpFO/view?usp=sharing

PROBLEM STATEMENTS

***1.**Find the total number of users in this dataset.*

***2.**Find out the number of Instagram users above the age of 25.*

***3.**Do male Instagram users tend to have more followers ,or female users?*

***4.**How many likes do young people receive on Instagram opposed to older members*

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

6.*How many young members use Instagram?*

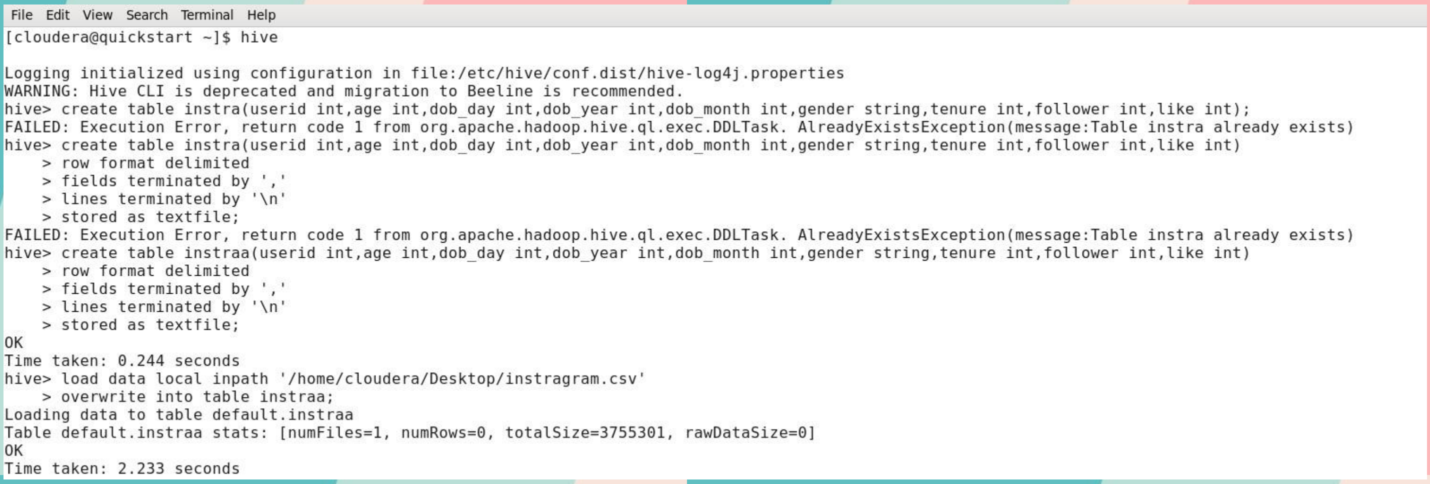
7.*How many adult members use Instagram ?*

8. *Visualization graph for the age wise number of people on Instagram.*

9.*Visualization for the number of likes which was received by male and female.*

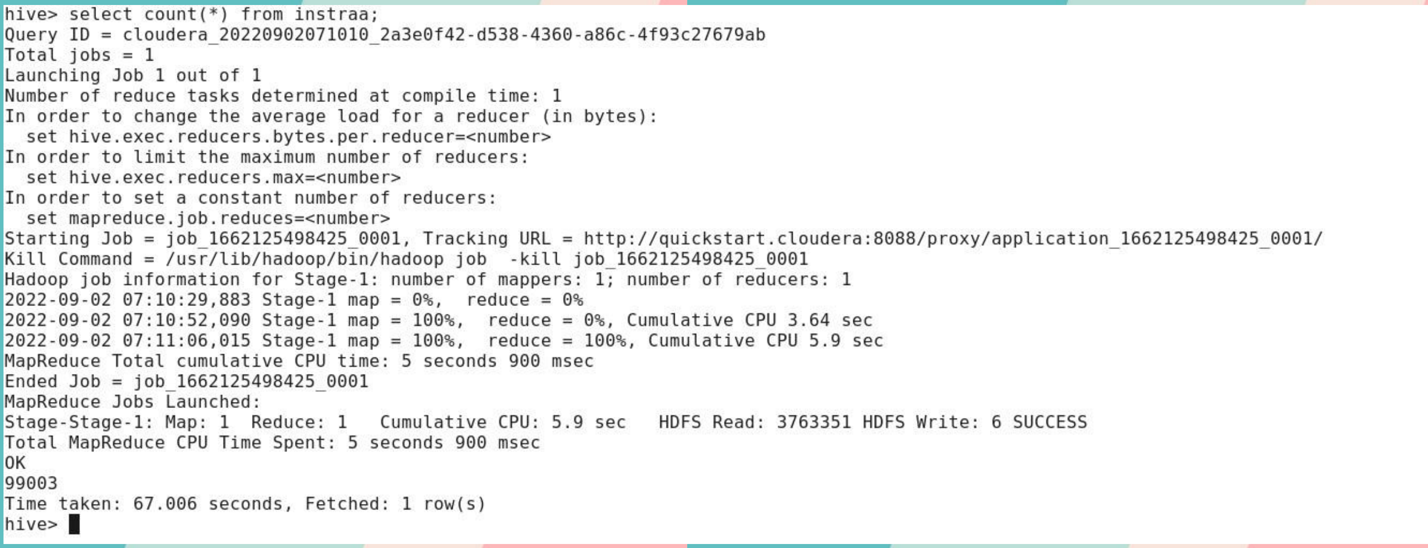
10. Visualization for the likes received for the age of the users (Male or Female)

CREATING A TABLE IN HIVE AND LOADING DATA INTO IT



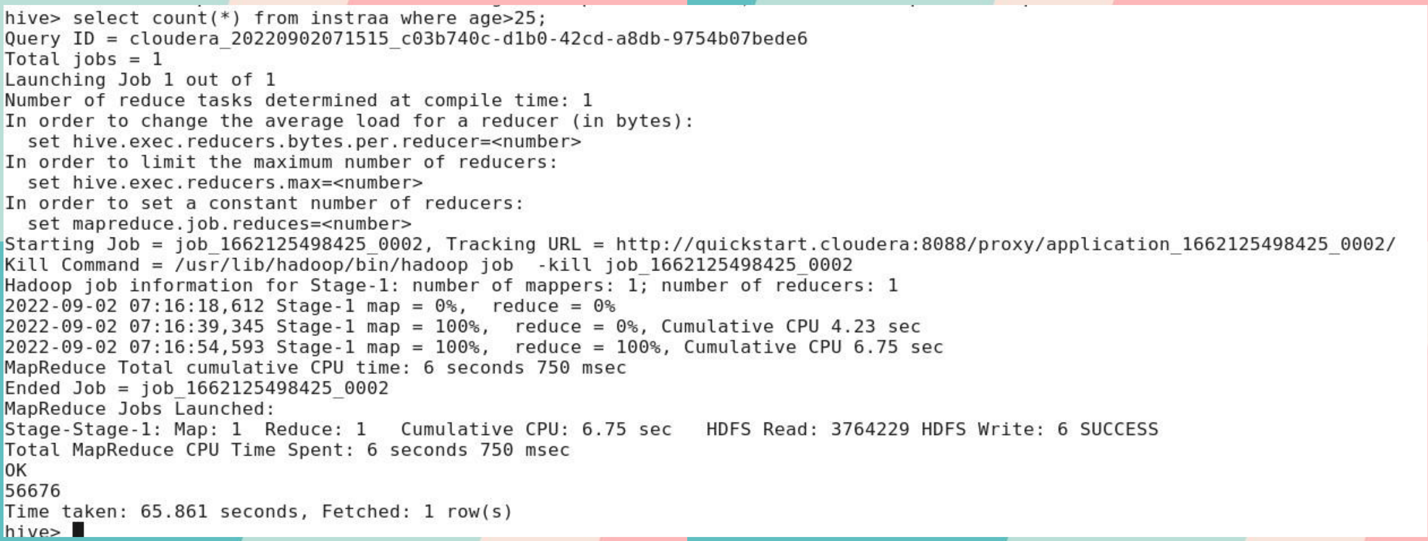
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table instra(userid int,age int,dob_day int,dob_year int,dob_month int,gender string,tenure int,follower int,like int);
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table instra already exists)
hive> create table instra(userid int,age int,dob_day int,dob_year int,dob_month int,gender string,tenure int,follower int,like int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
    > stored as textfile;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table instra already exists)
hive> create table instraa(userid int,age int,dob_day int,dob_year int,dob_month int,gender string,tenure int,follower int,like int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
    > stored as textfile;
OK
Time taken: 0.244 seconds
hive> load data local inpath '/home/cloudera/Desktop/instragram.csv'
    > overwrite into table instraa;
Loading data to table default.instraa
Table default.instraa stats: [numFiles=1, numRows=0, totalSize=3755301, rawDataSize=0]
OK
Time taken: 2.233 seconds
```

PROBLEM STATEMENT 1: FIND THE TOTAL NUMBER OF USERS IN THIS DATASET.



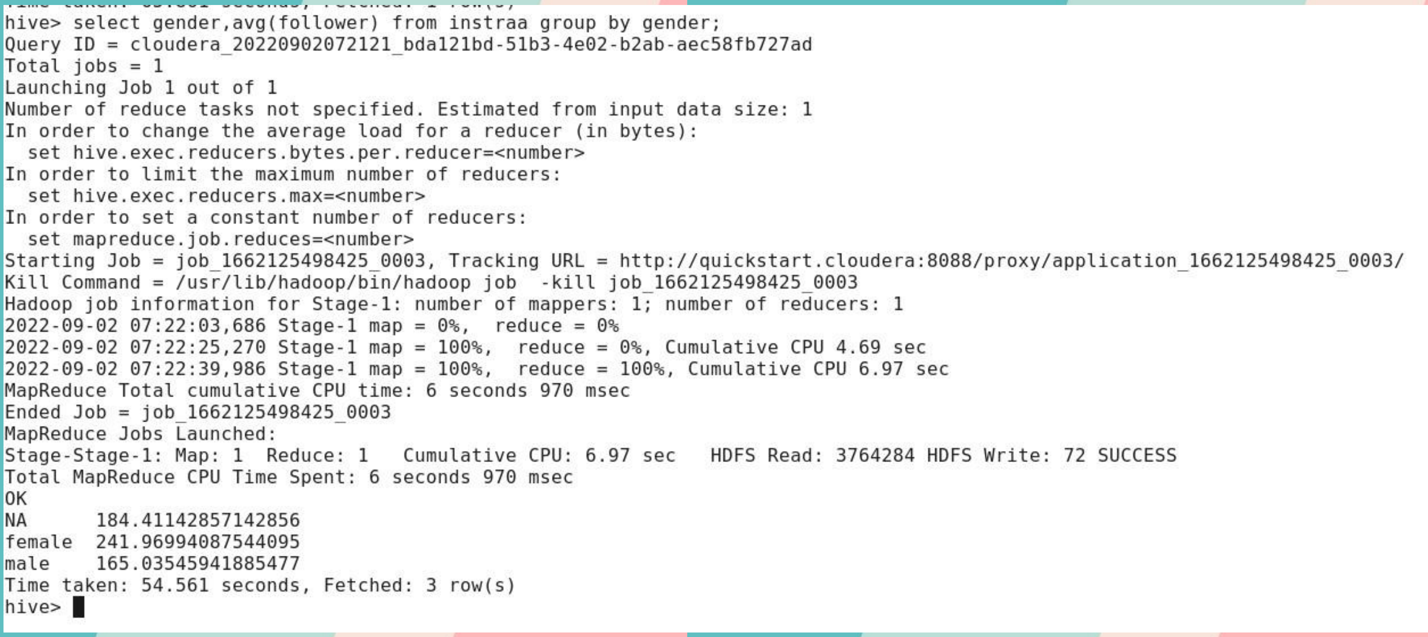
```
hive> select count(*) from instraa;
Query ID = cloudera_20220902071010_2a3e0f42-d538-4360-a86c-4f93c27679ab
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1662125498425_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:10:29,883 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:10:52,090 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.64 sec
2022-09-02 07:11:06,015 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.9 sec
MapReduce Total cumulative CPU time: 5 seconds 900 msec
Ended Job = job_1662125498425_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.9 sec   HDFS Read: 3763351 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 900 msec
OK
99003
Time taken: 67.006 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 2: FIND OUT THE NUMBER OF INSTAGRAM USERS ABOVE THE AGE OF 25.



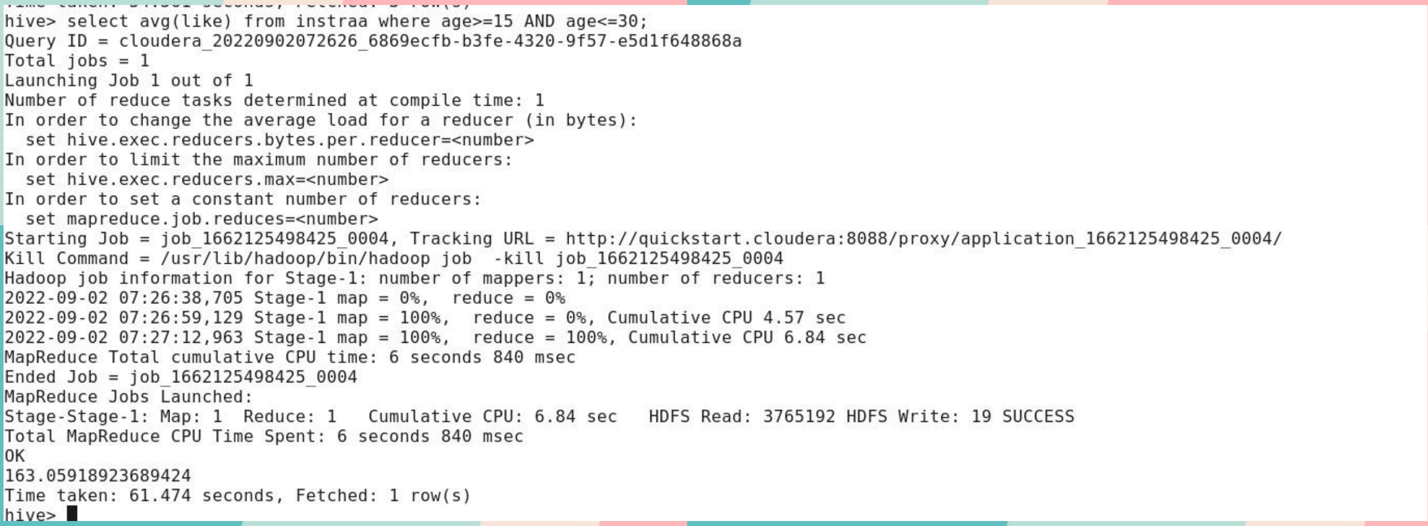
```
hive> select count(*) from instraa where age>25;
Query ID = cloudera_20220902071515_c03b740c-d1b0-42cd-a8db-9754b07bede6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1662125498425_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:16:18,612 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:16:39,345 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.23 sec
2022-09-02 07:16:54,593 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.75 sec
MapReduce Total cumulative CPU time: 6 seconds 750 msec
Ended Job = job_1662125498425_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.75 sec   HDFS Read: 3764229 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 750 msec
OK
56676
Time taken: 65.861 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 3:DO MALE INSTAGRAM USERS TEND TO HAVE MORE FOLLOWERS ,OR FEMALE USERS?



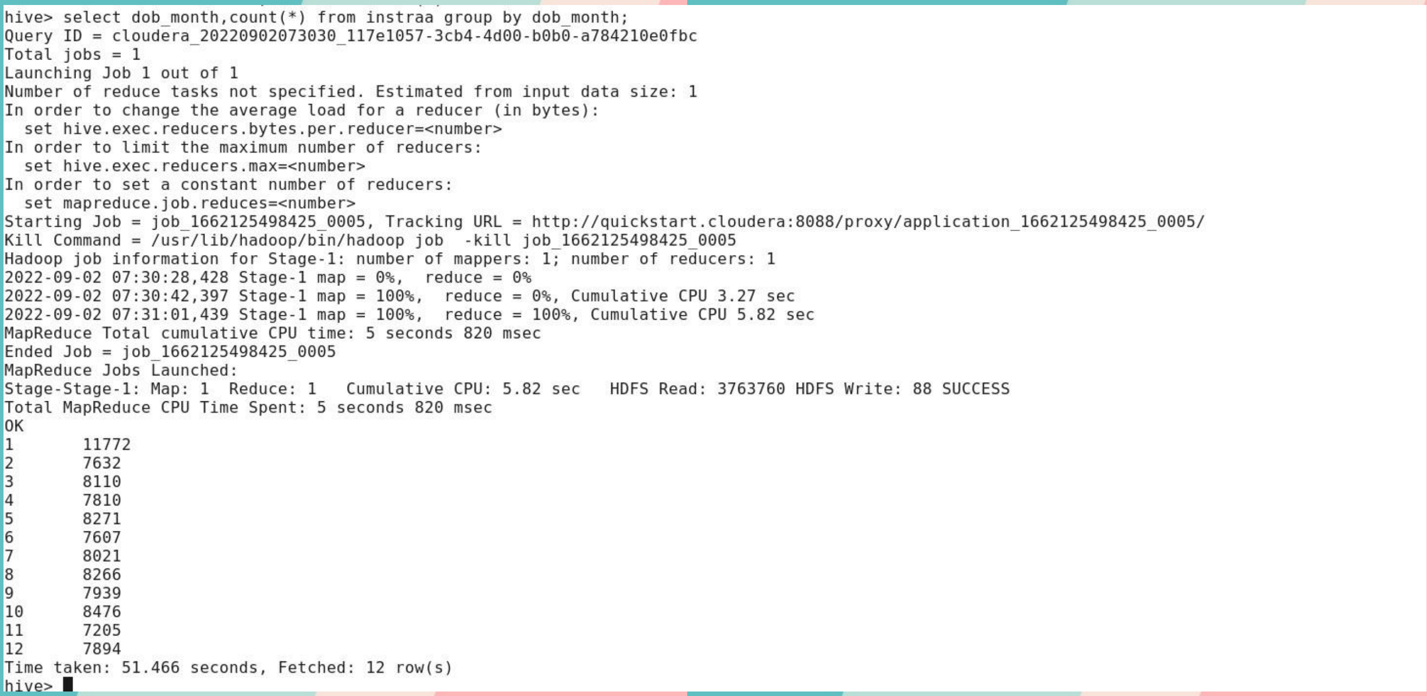
```
hive> select gender,avg(follower) from instraa group by gender;
Query ID = cloudera_20220902072121_bda121bd-51b3-4e02-b2ab-aec58fb727ad
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (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>
Starting Job = job_1662125498425_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:22:03,686 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:22:25,270 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.69 sec
2022-09-02 07:22:39,986 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.97 sec
MapReduce Total cumulative CPU time: 6 seconds 970 msec
Ended Job = job_1662125498425_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.97 sec   HDFS Read: 3764284 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 970 msec
OK
NA      184.41142857142856
female  241.96994087544095
male    165.03545941885477
Time taken: 54.561 seconds, Fetched: 3 row(s)
hive> █
```

PROBLEM STATEMENT 4:HOW MANY LIKES DO YOUNG PEOPLE RECEIVE ON INSTAGRAM OPPOSED TO OLDER MEMBERS ?



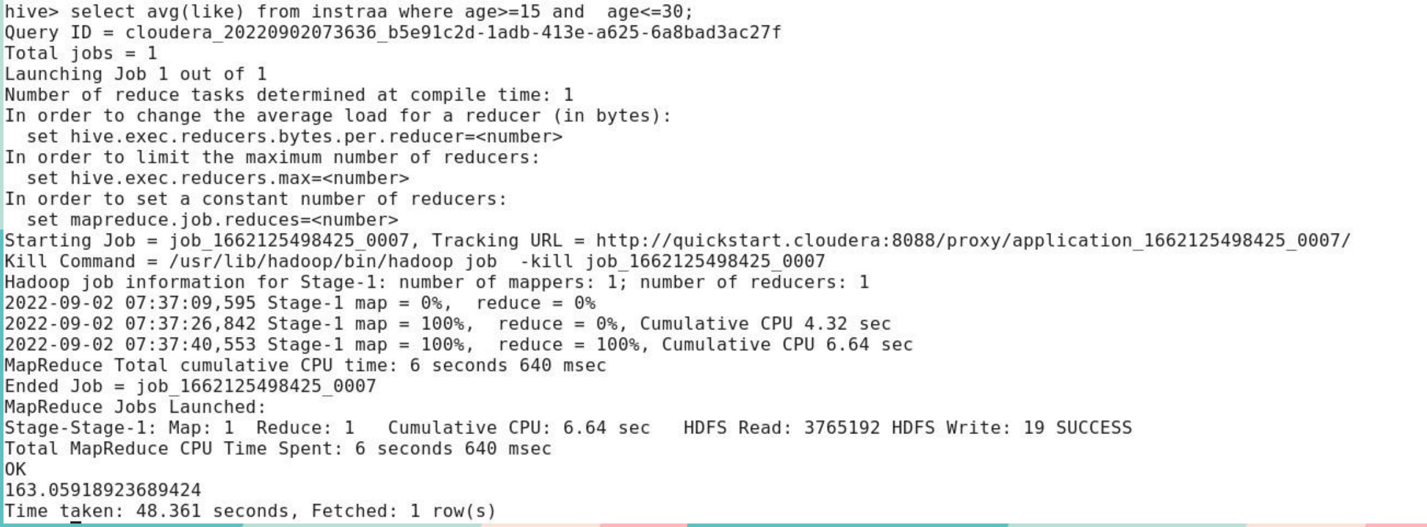
```
hive> select avg(like) from instraa where age>=15 AND age<=30;
Query ID = cloudera_20220902072626_6869ecfb-b3fe-4320-9f57-e5d1f648868a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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>
Starting Job = job_1662125498425_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:26:38,705 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:26:59,129 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.57 sec
2022-09-02 07:27:12,963 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.84 sec
MapReduce Total cumulative CPU time: 6 seconds 840 msec
Ended Job = job_1662125498425_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.84 sec   HDFS Read: 3765192 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 840 msec
OK
163.05918923689424
Time taken: 61.474 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 5:FIND OUT THE COUNT OF INSTAGRAM USERS FOR EACH BIRTHDAY MONTH.



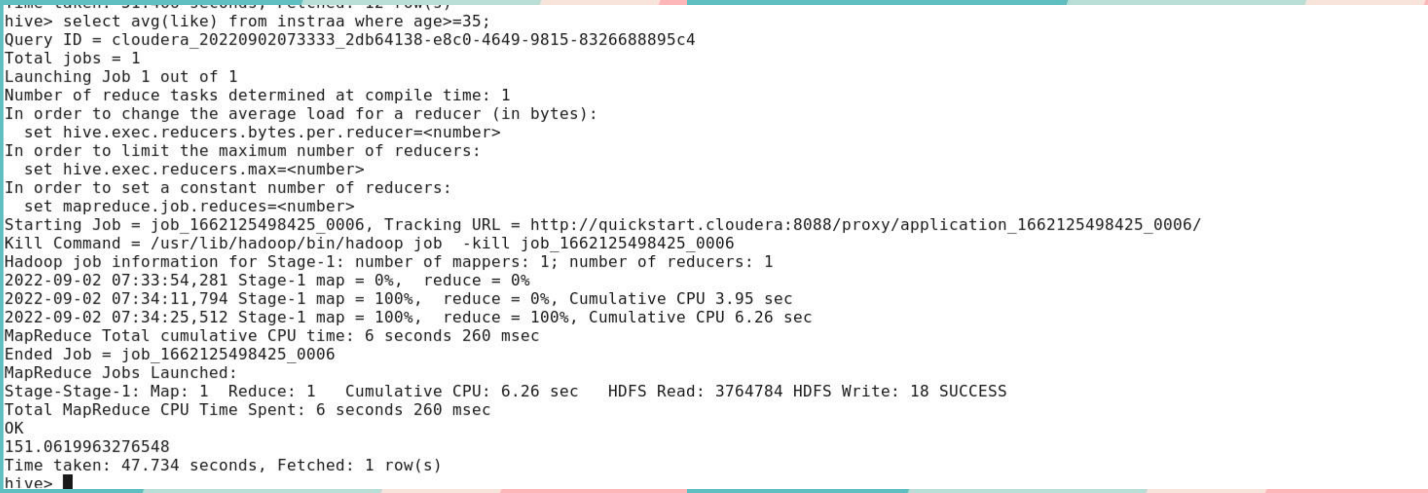
```
hive> select dob_month,count(*) from instraa group by dob_month;
Query ID = cloudera_20220902073030_117e1057-3cb4-4d00-b0b0-a784210e0fbc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (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>
Starting Job = job_1662125498425_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:30:28,428 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:30:42,397 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.27 sec
2022-09-02 07:31:01,439 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.82 sec
MapReduce Total cumulative CPU time: 5 seconds 820 msec
Ended Job = job_1662125498425_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.82 sec   HDFS Read: 3763760 HDFS Write: 88 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 820 msec
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: 51.466 seconds, Fetched: 12 row(s)
hive> █
```

PROBLEM STATEMENT 6.HOW MANY YOUNG MEMBERS USE INSTAGRAM ?



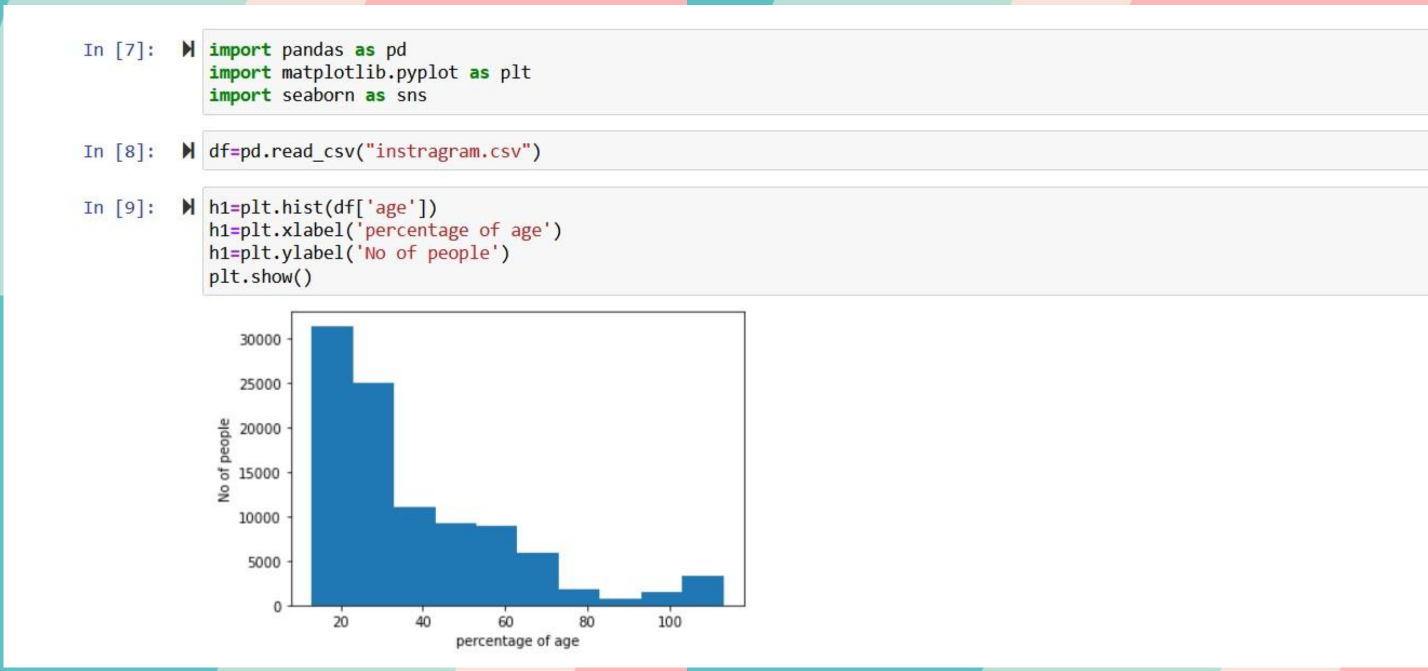
```
hive> select avg(like) from instraa where age>=15 and  age<=30;
Query ID = cloudera_20220902073636_b5e91c2d-1adb-413e-a625-6a8bad3ac27f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set 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>
Starting Job = job_1662125498425_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:37:09,595 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:37:26,842 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.32 sec
2022-09-02 07:37:40,553 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.64 sec
MapReduce Total cumulative CPU time: 6 seconds 640 msec
Ended Job = job_1662125498425_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.64 sec   HDFS Read: 3765192 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 640 msec
OK
163.05918923689424
Time taken: 48.361 seconds, Fetched: 1 row(s)
```

PROBLEM STATEMENT 7:-HOW MANY ADULT MEMBERS USE INSTAGRAM ?



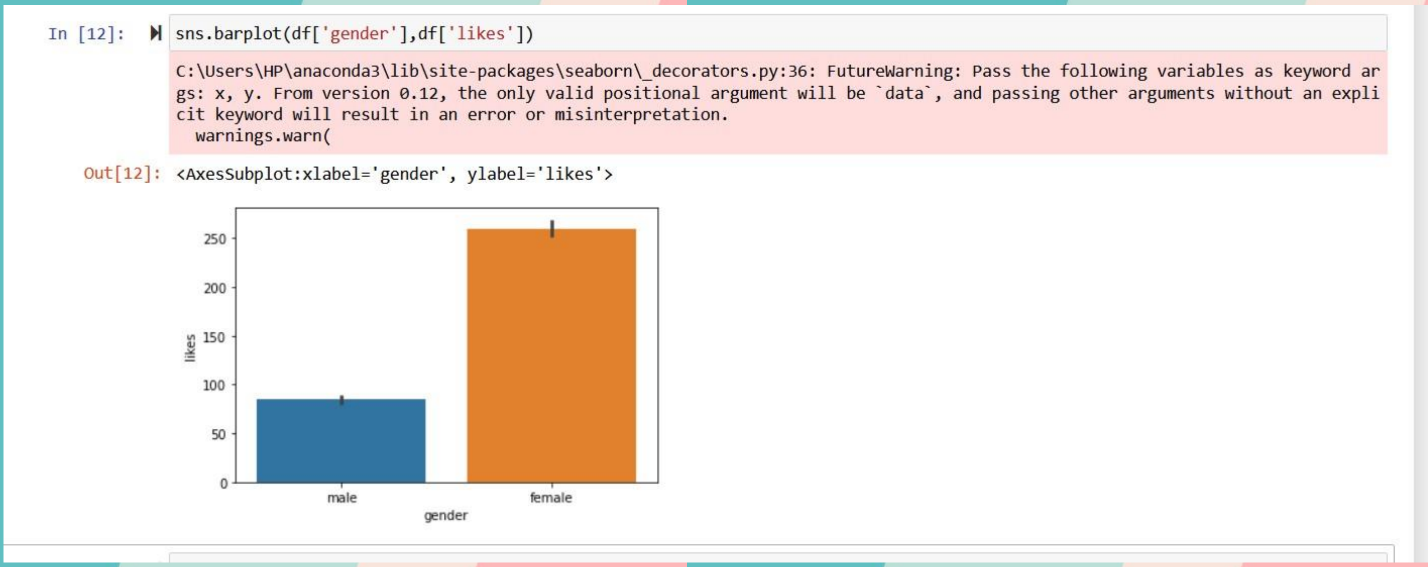
```
Time taken: 31.408 seconds, Fetched: 12 row(s)
hive> select avg(like) from instraa where age>=35;
Query ID = cloudera_20220902073333_2db64138-e8c0-4649-9815-8326688895c4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load 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>
Starting Job = job_1662125498425_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662125498425_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662125498425_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 07:33:54,281 Stage-1 map = 0%,  reduce = 0%
2022-09-02 07:34:11,794 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.95 sec
2022-09-02 07:34:25,512 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.26 sec
MapReduce Total cumulative CPU time: 6 seconds 260 msec
Ended Job = job_1662125498425_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.26 sec   HDFS Read: 3764784 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 260 msec
OK
151.0619963276548
Time taken: 47.734 seconds, Fetched: 1 row(s)
hive> 
```

PROBLEM STATEMENT 8:VISUALISATION GRAPH FOR THE AGE WISE NUMBER OF PEOPLE ON INSTAGRAM :



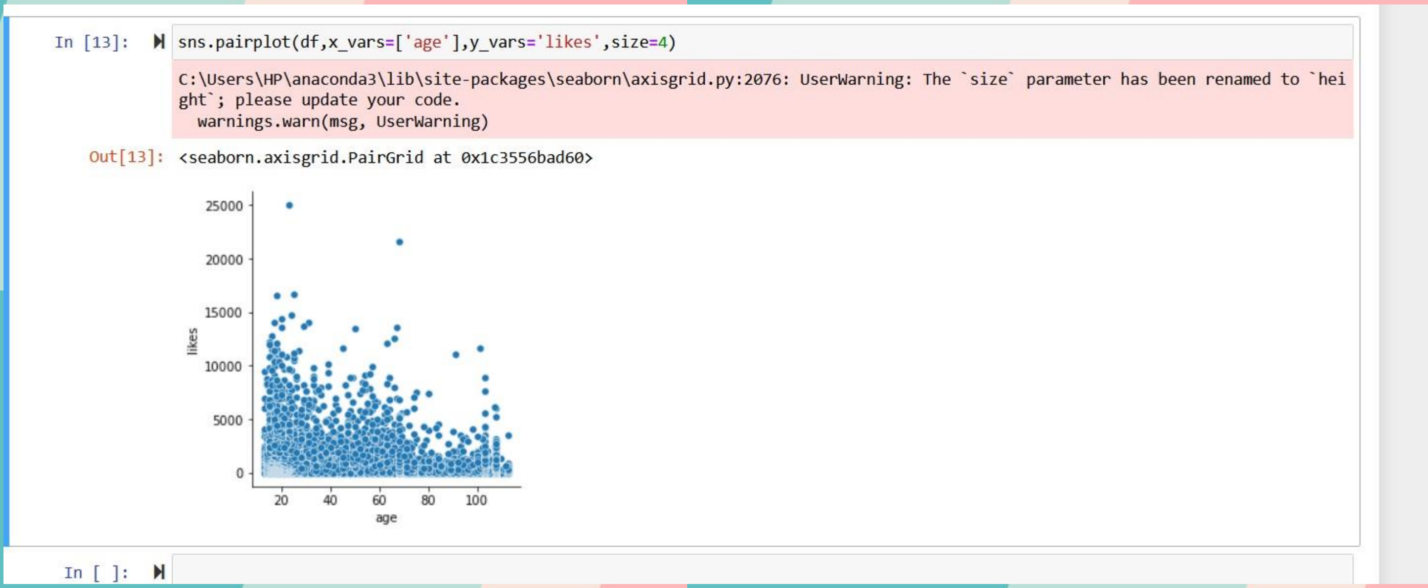
```
In [7]:  import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

In [8]:  df=pd.read_csv("instragram.csv")

In [9]:  h1=plt.hist(df['age'])
         h1=plt.xlabel('percentage of age')
         h1=plt.ylabel('No of people')
         plt.show()
```

PROBLEM STATEMENT 9:VISUALISATION FOR THE NUMBER OF LIKES WHICH WAS RECEIVED BY MALE AND FEMALE:



```
In [12]:  sns.barplot(df['gender'],df['likes'])

          C:\Users\HP\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword ar
          gs: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an expli
          cit keyword will result in an error or misinterpretation.
            warnings.warn(

Out[12]:  <AxesSubplot:xlabel='gender', ylabel='likes'>
```

PROBLEM STATEMENT 10:VISUALISATION FOR THE LIKES RECEIVED FOR THE AGE OF THE USERS (MALE OR FEMALE):



```
In [13]:  sns.pairplot(df,x_vars=['age'],y_vars='likes',size=4)

          C:\Users\HP\anaconda3\lib\site-packages\seaborn\axisgrid.py:2076: UserWarning: The `size` parameter has been renamed to `hei
          ght`; please update your code.
            warnings.warn(msg, UserWarning)

Out[13]:  <seaborn.axisgrid.PairGrid at 0x1c3556bad60>

In [ ]:
```

THANK YOU