Facebook Data Analysis using Hadoop

Submitted in partial fulfillment of the requirements
of the degree of

B. E. Computer Engineering

By

Daniel Lobo	**07**	**172028**
Jenny Dcruz	**16**	**172047**
Smita Deulkar	**19**	**172113**
Leander Fernandes	**20**	**172061**

Guide:

Ms. Anuradha Srinivasaraghavan
Associate Professor

Department of Computer Engineering
St. Francis Institute of Technology
(Engineering College)

University of Mumbai
2020-2021

Contents

Chapter	Contents	Page No.
1	PROBLEM STATEMENT	1
2	OBJECTIVE	2
3	EXPERIMENTAL SETUP	2
4	IMPLEMENTATION CODE	3
5	RESULT AND ANALYSIS	28
6	CONCLUSIONS	33

1. Problem Statement

Facebook is a social networking site that makes it easy for you to connect and share media with family and friends online. Facebook can be accessed from devices with Internet connectivity, such as personal computers, tablets and smartphones. Facebook allows companies/ entrepreneurs/ individuals to post advertisements about their services/ product's online on it's platform. It's ads are targeted to users based on their location, activities, profile, etc. These ad sales and promotions are the primary source of Facebook's revenue. In today's day and time with digital revolution at its peak, Facebook is experiencing an ever increasing demand for advertising amid an acceleration of the shift to online commerce which is spurred on by the COVID-19 pandemic. Analysis of the user demographic is crucial for any advertisement agency. With the rise of Big Data upon us, we can use Facebook's user data and generate intelligible insights for enhancing an organization's decision-making process so that they can advertise their services/ products accordingly in an effective manner and have users interact with them. For example, analyzing the age group of the users can help the agency decide whether they must target the younger or older age groups and accordingly push forward their products on the platform. Engagement is one of the most important Facebook metrics you can track. Analysing social media data is a very complex and challenging task as Facebook has over 2.7 billion active users. Hence, to store, analyze or process even a small portion of this data we cannot use a simple RDBMS.

Hadoop helps us tackle this issue. Hadoop is a distributed computing framework which has two core components – Hadoop Distributed File System (HDFS) for storing data and MapReduce for processing data. As Hadoop uses MapReduce it is especially effective at analyzing and processing petabytes of data. Therefore, it is used in big data analysis. Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. It takes very less time to write a Hive query in comparison to MapReduce code.

While data analysis helps us dig through data, bring order and structure to data, data visualization represents it in a discernable context by making explicit the trends, patterns immanent in the data. Here's where Microsoft Power BI comes into play. The 'BI' in Power BI stands for Business Intelligence. It is a data visualization tool that allows you to quickly connect your data, transform it, and visualize it as you like. It gives you the power to transform all your data into live interactive visuals, create customized real time business view dashboards, thus extracting insights for enhanced decision making. You can visualize data and share insights.

Therefore, over the course of this project we shall use Hadoop and Power BI to analyse and visualize a dataset on Facebook users activities. The data contains 99003 entries with 15 columns of varying fields.

2. Objective

Data helps us understand and improve business processes so we can reduce wasted money and time. Every organization feels the effects of waste. It depletes resources, squanders time, and ultimately impacts the bottom line. For example, bad advertising decisions can be one of the greatest wastes of resources in a company.

Therefore our objective is to analyze the Facebook data using Hadoop in order to gain a certain understanding about Facebook users which can be used for the purpose of better decision making. The dataset gives us insights of which platform (Mobile or PC) is more dominant and in turn lets us know the user base's preferred device. This will help advertisers to target their advertisements to the right audience.

By analysing the user base's age, we get a better understanding of whether the users are of younger or older demographic. This would help in improving the platform and promote new features towards the dominant demographic.

3. Experimental Setup

Software Requirements

- Windows OS
- VirtualBox
- Cloudera CDH

Hardware Requirements

- CPU: Intel i5/Ryzen 5 or above
- Clock speed: 2.5 GHz or above
- RAM required: 8GB or above
- Hard Disk capacity: 64GB or above

Technology Used

- Hadoop
- HIVE
- Power BI

Dataset

- Facebook Data: https://www.kaggle.com/sheenabatra/facebook-data

4. Implementation Code:

Importing the dataset into Hive and loading it into a table:

```
[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> use fbdata;
OK
Time taken: 0.448 seconds

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_recvd int, mlikes int, mlikes_recvd int, wlikes int,
wlikes_recvd int) row format delimited fields terminated by ','
stored as textfile;
OK
Time taken: 0.427 seconds

hive> load data local
inpath'/home/cloudera/Desktop/Project/pseudo_facebook.csv' into
table fb;
Loading data to table fbdata.fb
Table fbdata.fb stats: [numFiles=1, totalSize=5315845]
OK
Time taken: 1.229 seconds
```

Viewing the dataset

```
hive> select * from fb limit 5;
OK
209438214    19    1999  11    male  266   0     0     0     0     0
0  0     0
119260114    2     1999  11    female      6     0     0     0     0
0  0     0     0
208388414    16    1999  11    male  13    0     0     0     0     0
0  0     0
120316814    25    1999  12    female      93    0     0     0     0
0  0     0     0
173318614    4     1999  12    male  82    0     0     0     0     0
0  0     0
Time taken: 0.126 seconds, Fetched: 5 row(s)
```

Queries:

A. Total number of users in the dataset.

```
hive> select count(*) from fb;
Query ID =
cloudera_20201128034040_d29138da-17d4-46a0-8c29-897be68dcc6c
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_1606559954037_0001, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0001
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 03:40:40,387 Stage-1 map = 0%,  reduce = 0%
2020-11-28 03:40:51,125 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.53 sec
2020-11-28 03:41:03,449 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.29 sec
MapReduce Total cumulative CPU time: 3 seconds 290 msec
Ended Job = job_1606559954037_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.29 sec
HDFS Read: 5323616 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 290 msec
OK
99003
Time taken: 42.29 seconds, Fetched: 1 row(s)
```

B. Number of users whose age is greater than or equal to 18 years.

```
hive> select count(*) from fb where age>=18;
Query ID =
cloudera_20201128034747_2fe58dd7-3b56-4946-8e60-1bd17d5a086a
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_1606559954037_0002, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0002
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 03:47:26,354 Stage-1 map = 0%,  reduce = 0%
2020-11-28 03:47:36,622 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.07 sec
2020-11-28 03:47:47,556 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.57 sec
MapReduce Total cumulative CPU time: 3 seconds 570 msec
Ended Job = job_1606559954037_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.57 sec
HDFS Read: 5324473 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 570 msec
OK
87607
Time taken: 34.89 seconds, Fetched: 1 row(s)
```

C. Number of users whose age is less than 18 years.

```
hive> select count(*) from fb where age<18;
Query ID =
cloudera_20201128034848_d06ebe69-d225-44fb-9419-edc5671e5b41
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_1606559954037_0003, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0003/
```

```
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0003
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 03:48:44,555 Stage-1 map = 0%,  reduce = 0%
2020-11-28 03:48:54,699 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.04 sec
2020-11-28 03:49:05,994 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.9 sec
MapReduce Total cumulative CPU time: 3 seconds 900 msec
Ended Job = job_1606559954037_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.9 sec
HDFS Read: 5324425 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 900 msec
OK
11396
Time taken: 34.95 seconds, Fetched: 1 row(s)
```

D. Which gender has more number of friends on average.
Command for average friends of male users

```
hive> select avg(friends) from fb where gender='male';
Query ID =
cloudera_20201128035353_c10ed159-a381-4eb9-b9d5-452260c53b3a
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_1606559954037_0005, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0005
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 03:53:42,448 Stage-1 map = 0%,  reduce = 0%
2020-11-28 03:53:53,339 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.07 sec
```

```
2020-11-28 03:54:03,328 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.66 sec
MapReduce Total cumulative CPU time: 3 seconds 660 msec
Ended Job = job_1606559954037_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.66 sec
HDFS Read: 5324799 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 660 msec
OK
165.03545941885477
Time taken: 35.055 seconds, Fetched: 1 row(s)
```

Command for average friends of female users

```
hive> select avg(friends) from fb where gender='female';
Query ID =
cloudera_20201128035454_f54b9365-934f-46b6-8d1b-ffcc20742be6
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_1606559954037_0006, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0006
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 03:55:12,869 Stage-1 map = 0%,  reduce = 0%
2020-11-28 03:55:24,030 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.42 sec
2020-11-28 03:55:35,077 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 4.17 sec
MapReduce Total cumulative CPU time: 4 seconds 170 msec
Ended Job = job_1606559954037_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.17 sec
HDFS Read: 5324812 HDFS Write: 19 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 170 msec
OK
```

```
241.96994087544095
Time taken: 36.557 seconds, Fetched: 1 row(s)
```

Command for average friends of each genders

```
hive> select gender,avg(friends) from fb group by gender;
Query ID =
cloudera_20201128040000_b60c5f72-5abf-455b-99f9-bdf294f825f6
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_1606559954037_0007, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0007
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:01:12,021 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:01:22,377 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.68 sec
2020-11-28 04:01:32,683 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.15 sec
MapReduce Total cumulative CPU time: 3 seconds 150 msec
Ended Job = job_1606559954037_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.15 sec
HDFS Read: 5324323 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 150 msec
OK
NA        184.41142857142856
female    241.96994087544095
male      165.03545941885477
Time taken: 39.107 seconds, Fetched: 3 row(s)
```

E. Which age group (younger, older) receives more number of likes on average.
Command for average likes received by young users (<=25)

```
hive> select avg(likes_recvd) from fb where age<=25;
Query ID =
cloudera_20201128040404_18c8139b-c00b-4476-b206-ac27053f3527
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_1606559954037_0008, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0008
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:04:47,141 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:04:58,531 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.82 sec
2020-11-28 04:05:10,748 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 4.65 sec
MapReduce Total cumulative CPU time: 4 seconds 650 msec
Ended Job = job_1606559954037_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.65 sec
HDFS Read: 5324843 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 650 msec
OK
200.2870508186264
Time taken: 37.092 seconds, Fetched: 1 row(s)
```

Command for average likes received by older users (>25)

```
hive> select avg(likes_recvd) from fb where age>25;
Query ID =
cloudera_20201128040707_e6c1cc01-7b65-4c2d-9b40-4b9311df8023
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_1606559954037_0009, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0009
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:07:21,372 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:07:32,599 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.29 sec
2020-11-28 04:07:43,716 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.98 sec
MapReduce Total cumulative CPU time: 3 seconds 980 msec
Ended Job = job_1606559954037_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.98 sec
HDFS Read: 5324824 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 980 msec
OK
99.67402427835415
Time taken: 35.971 seconds, Fetched: 1
```

F. What age receives the maximum number of likes.
Command for likes received by different ages grouped by number of likes

```
hive> select age,avg(likes_recvd) from fb group by age;
Query ID =
cloudera_20201128040808_d7a5fa7e-622c-47fc-8146-7a81452c949a
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_1606559954037_0010, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0010
```

```
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:09:11,721 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:09:21,863 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.91 sec
2020-11-28 04:09:34,198 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.8 sec
MapReduce Total cumulative CPU time: 3 seconds 800 msec
Ended Job = job_1606559954037_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.8 sec
HDFS Read: 5324334 HDFS Write: 2136 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 800 msec
OK
13  157.3904958677686
14  146.94545454545454
15  241.5305576776165
16  208.57939079714842
17  313.39415169052694
18  240.00577367205543
19  236.73104076520156
20  307.8355001326612
21  140.2402615091256
22  144.90468337730871
23  167.9786557674841
24  120.23134064379201
25  91.45674265311727
26  96.43374777975133
27  80.97857142857143
28  102.43020304568527
29  84.35640495867769
30  80.23193473193473
31  96.39020070838252
32  76.54885654885655
33  95.52276138069034
34  91.24661893396977
35  77.77361702127659
36  88.16905187835421
37  113.91847265221878
38  92.86442220200182
39  129.4068736141907
40  78.1808383233533
41  91.86766541822722
42  100.30419161676647
43  113.32925682031986
44  107.78677309007982
```

```
45  115.20338983050847
46  98.78965922444183
47  129.26940133037695
48  108.09821428571429
49  96.1955835962145
50  122.9047619047619
51  104.3501544799176
52  101.73869346733669
53  103.06879194630872
54  106.63406940063092
55  102.02594594594595
56  99.79612756264237
57  106.10729613733906
58  112.71892497200447
59  93.80935251798562
60  87.66577540106952
61  88.72836538461539
62  125.38315217391305
63  81.25137816979051
64  92.09026798307475
65  92.64357864357865
66  94.86628733997155
67  70.70860927152317
68  154.29432624113474
69  78.30869565217391
70  59.15068493150685
71  58.15625
72  49.695945945945944
73  44.472727272727276
74  196.39857651245552
75  51.66094420600859
76  31.146067415730336
77  66.66863905325444
78  113.51851851851852
79  30.339285714285715
80  62.419117647058826
81  55.657407407407405
82  56.65384615384615
83  83.63815789473684
84  140.51162790697674
85  95.4578313253012
86  61.421052631578945
87  89.33333333333333
88  215.98360655737704
89  161.85
90  127.77464788732394
```

```
91  110.76315789473684
92  129.3846153846154
93  129.22660098522167
94  128.2173913043478
95  101.35064935064935
96  269.62857142857143
97  124.67857142857143
98  107.82795698924731
99  59.373493975903614
100 138.09230769230768
101 189.9235668789809
102 185.76470588235293
103 149.44252873563218
104 152.5205479452055
105 69.1
106 84.256
107 293.35714285714283
108 113.09993979530404
109 28.77777777777778
110 75.53333333333333
111 62.55555555555556
112 67.27777777777777
113 103.06930693069307
Time taken: 37.225 seconds, Fetched: 101 row(s)
```

Command for likes received by different ages ordered by number of likes

```
hive> select age,avg(likes_recvd) as LikesRCVD from fb group by
age order by LikesRCVD desc;
Query ID =
cloudera_20201128041414_d0d1e0a0-42cb-4f7c-854b-ccc1f20d3d66
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>
Starting Job = job_1606559954037_0011, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0011
```

```
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:14:55,674 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:15:05,680 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.82 sec
2020-11-28 04:15:16,817 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.52 sec
MapReduce Total cumulative CPU time: 3 seconds 520 msec
Ended Job = job_1606559954037_0011
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>
Starting Job = job_1606559954037_0012, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0012
Hadoop job information for Stage-2: number of mappers: 1;
number of reducers: 1
2020-11-28 04:15:30,115 Stage-2 map = 0%,  reduce = 0%
2020-11-28 04:15:39,151 Stage-2 map = 100%,  reduce = 0%,
Cumulative CPU 1.14 sec
2020-11-28 04:15:51,133 Stage-2 map = 100%,  reduce = 100%,
Cumulative CPU 2.71 sec
MapReduce Total cumulative CPU time: 2 seconds 710 msec
Ended Job = job_1606559954037_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.52 sec
HDFS Read: 5323596 HDFS Write: 2742 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.71 sec
HDFS Read: 7087 HDFS Write: 2136 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 230 msec
OK
17  313.39415169052694
20  307.8355001326612
107 293.35714285714283
96  269.62857142857143
15  241.5305576776165
18  240.00577367205543
19  236.73104076520156
88  215.98360655737704
```

```
16  208.57939079714842
74  196.39857651245552
101 189.9235668789809
102 185.76470588235293
23  167.9786557674841
89  161.85
13  157.3904958677686
68  154.29432624113474
104 152.5205479452055
103 149.44252873563218
14  146.94545454545454
22  144.90468337730871
84  140.51162790697674
21  140.2402615091256
100 138.09230769230768
39  129.4068736141907
92  129.3846153846154
47  129.26940133037695
93  129.22660098522167
94  128.2173913043478
90  127.77464788732394
62  125.38315217391305
97  124.67857142857143
50  122.9047619047619
24  120.23134064379201
45  115.20338983050847
37  113.91847265221878
78  113.51851851851852
43  113.32925682031986
108 113.09993979530404
58  112.71892497200447
91  110.76315789473684
48  108.09821428571429
98  107.82795698924731
44  107.78677309007982
54  106.63406940063092
57  106.10729613733906
51  104.3501544799176
113 103.06930693069307
53  103.06879194630872
28  102.43020304568527
55  102.02594594594595
52  101.73869346733669
95  101.35064935064935
42  100.30419161676647
56  99.79612756264237
```

```
46  98.78965922444183
26  96.43374777975133
31  96.39020070838252
49  96.1955835962145
33  95.52276138069034
85  95.4578313253012
66  94.86628733997155
59  93.80935251798562
38  92.86442220200182
65  92.64357864357865
64  92.09026798307475
41  91.86766541822722
25  91.45674265311727
34  91.24661893396977
87  89.33333333333333
61  88.72836538461539
36  88.16905187835421
60  87.66577540106952
29  84.35640495867769
106 84.256
83  83.63815789473684
63  81.25137816979051
27  80.97857142857143
30  80.23193473193473
69  78.30869565217391
40  78.1808383233533
35  77.77361702127659
32  76.54885654885655
110 75.53333333333333
67  70.70860927152317
105 69.1
112 67.27777777777777
77  66.66863905325444
111 62.55555555555556
80  62.419117647058826
86  61.421052631578945
99  59.373493975903614
70  59.15068493150685
71  58.15625
82  56.65384615384615
81  55.657407407407405
75  51.66094420600859
72  49.695945945945944
73  44.472727272727276
76  31.146067415730336
79  30.339285714285715
```

```
109 28.77777777777778
Time taken: 68.366 seconds, Fetched: 101 row(s)
```

Command for count of the users of top five ages with highest number of likes received.

```
hive> select age,count(*) from fb where age=17 or age=20 or
age=107 or age=96 or age=15 group by age;
Query ID =
cloudera_20201128042626_823dad3e-7dd1-4b6d-89df-ae4b6367db2b
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_1606559954037_0015, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0015
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:26:37,054 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:26:47,436 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.19 sec
2020-11-28 04:26:58,574 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.84 sec
MapReduce Total cumulative CPU time: 3 seconds 840 msec
Ended Job = job_1606559954037_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.84 sec
HDFS Read: 5324416 HDFS Write: 37 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 840 msec
OK
15  2618
17  3283
20  3769
96  70
107 98
Time taken: 35.876 seconds, Fetched: 5 row(s)
```

G. Which user group (mobile , PC) receives more likes in total

Command for total likes received.

```
hive> select sum(likes_recvd) from fb;
Query ID =
cloudera_20201128043535_96437d22-c7b7-415c-9d37-e3206f3f82c8
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_1606559954037_0017, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0017
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:35:20,712 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:35:30,776 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.79 sec
2020-11-28 04:35:42,701 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.41 sec
MapReduce Total cumulative CPU time: 3 seconds 410 msec
Ended Job = job_1606559954037_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.41 sec
HDFS Read: 5323662 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 410 msec
OK
14126675
Time taken: 37.097 seconds, Fetched: 1 row(s)
```

Command for total likes received by mobile users.

```
hive> select sum(mlikes_recvd) from fb;
Query ID =
cloudera_20201128043636_b75a6083-4eef-4d39-a53d-c62066d7b137
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_1606559954037_0018, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0018
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:36:52,813 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:37:04,866 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.9 sec
2020-11-28 04:37:18,202 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.77 sec
MapReduce Total cumulative CPU time: 3 seconds 770 msec
Ended Job = job_1606559954037_0018
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.77 sec
HDFS Read: 5323662 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 770 msec
OK
8328181
Time taken: 37.501 seconds, Fetched: 1 row(s)
```

Command for total likes received by PC users.

```
hive> select sum(wlikes_recvd) from fb;
Query ID =
cloudera_20201128043838_cf911a43-4611-4fba-9ad5-938a6ddc9162
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_1606559954037_0019, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0019/
```

```
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0019
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:39:09,356 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:39:18,257 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.61 sec
2020-11-28 04:39:30,206 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.19 sec
MapReduce Total cumulative CPU time: 3 seconds 190 msec
Ended Job = job_1606559954037_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.19 sec
HDFS Read: 5323662 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 190 msec
OK
5798490
Time taken: 33.856 seconds, Fetched: 1 row(s)
```

H. Which user group (mobile , PC) gives more likes in total.
Command for total likes given.

```
hive> select sum(likes) from fb;
Query ID =
cloudera_20201128044141_59fce2ba-933d-4f60-bcc7-838a5363b2fa
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_1606559954037_0020, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0020
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:41:16,250 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:41:25,345 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.59 sec
```

```
2020-11-28 04:41:38,888 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.51 sec
MapReduce Total cumulative CPU time: 3 seconds 510 msec
Ended Job = job_1606559954037_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.51 sec
HDFS Read: 5323662 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 510 msec
OK
15452268
Time taken: 34.695 seconds, Fetched: 1 row(s)
```

Command for total likes given by mobile users.

```
hive> select sum(mlikes) from fb;
Query ID =
cloudera_20201128044444_18205917-3437-44ab-ba28-9fca0305d097
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_1606559954037_0021, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0021
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:45:03,141 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:45:14,572 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.78 sec
2020-11-28 04:45:26,920 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.38 sec
MapReduce Total cumulative CPU time: 3 seconds 380 msec
Ended Job = job_1606559954037_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.38 sec
HDFS Read: 5323628 HDFS Write: 9 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 380 msec
OK
10505832
```

```
Time taken: 42.995 seconds, Fetched: 1 row(s)
```

Command for total likes given by PC users.

```
hive> select sum(wlikes) from fb;
Query ID =
cloudera_20201128044545_4b443022-5e4a-48fc-97ca-c18bc0404641
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_1606559954037_0022, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0022/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0022
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:46:05,584 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:46:15,632 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.83 sec
2020-11-28 04:46:25,398 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.42 sec
MapReduce Total cumulative CPU time: 3 seconds 420 msec
Ended Job = job_1606559954037_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.42 sec
HDFS Read: 5323662 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 420 msec
OK
4946430
Time taken: 32.31 seconds, Fetched: 1 row(s)
```

I. Count of users for each birthday month.

```
hive> select month,count(*) from fb group by month;
Query ID =
cloudera_20201128044949_619cba31-27e0-4d6f-bf40-26ab6832093b
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_1606559954037_0023, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0023
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:49:50,473 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:50:00,522 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 2.2 sec
2020-11-28 04:50:11,425 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.65 sec
MapReduce Total cumulative CPU time: 3 seconds 650 msec
Ended Job = job_1606559954037_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.65 sec
HDFS Read: 5323924 HDFS Write: 88 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 650 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: 33.0 seconds, Fetched: 12 row(s)
```

J. How old is each age group's user account on average.

```
hive> select age,avg(tenure) as fbaccage from fb group by age
order by fbaccage desc;
Query ID =
cloudera_20201128045656_ccddd590-9970-4617-84b5-7e8099e7b32a
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>
Starting Job = job_1606559954037_0025, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0025
Hadoop job information for Stage-1: number of mappers: 1;
number of reducers: 1
2020-11-28 04:56:17,840 Stage-1 map = 0%,  reduce = 0%
2020-11-28 04:56:27,749 Stage-1 map = 100%,  reduce = 0%,
Cumulative CPU 1.79 sec
2020-11-28 04:56:38,754 Stage-1 map = 100%,  reduce = 100%,
Cumulative CPU 3.2 sec
MapReduce Total cumulative CPU time: 3 seconds 200 msec
Ended Job = job_1606559954037_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.reduces=<number>
Starting Job = job_1606559954037_0026, Tracking URL =
http://quickstart.cloudera:8088/proxy/application_160655995403
7_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill
job_1606559954037_0026
Hadoop job information for Stage-2: number of mappers: 1;
number of reducers: 1
2020-11-28 04:56:52,572 Stage-2 map = 0%,  reduce = 0%
2020-11-28 04:57:00,430 Stage-2 map = 100%,  reduce = 0%,
Cumulative CPU 1.03 sec
```

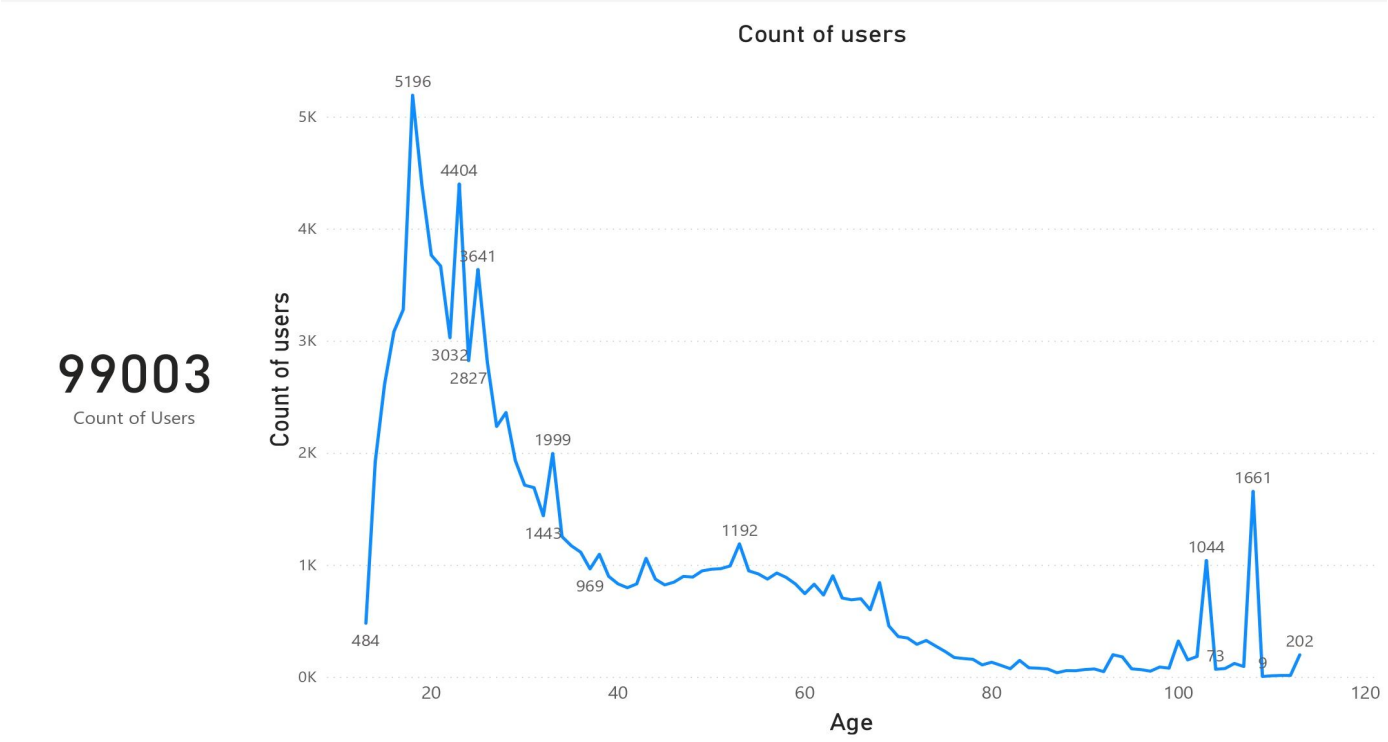
```
2020-11-28 04:57:10,402 Stage-2 map = 100%,  reduce = 100%,
Cumulative CPU 2.45 sec
MapReduce Total cumulative CPU time: 2 seconds 450 msec
Ended Job = job_1606559954037_0026
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.2 sec
HDFS Read: 5323596 HDFS Write: 2742 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.45 sec
HDFS Read: 7086 HDFS Write: 2136 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 650 msec
OK
111 1903.2777777777778
110 1834.2
109 1689.4444444444443
113 1557.7970297029703
112 1483.1666666666667
89  1295.7333333333333
102 1214.0855614973261
92  1194.1153846153845
97  1192.267857142857
105 1188.175
100 1179.2707692307692
96  1179.1142857142856
88  1170.5833333333333
95  1170.051948051948
86  1159.2631578947369
106 1136.616
90  1101.338028169014
84  1100.2558139534883
104 1091.4109589041095
85  1084.2289156626507
101 1083.5286624203823
99  1073.710843373494
108 1072.0752558699578
107 1053.1632653061224
98  1047.021505376344
83  1042.8486842105262
93  1026.1330049261085
79  1015.6428571428571
75  1008.0515021459228
82  1003.1666666666666
78  996.7098765432099
94  992.1902173913044
103 991.9702780441036
70  986.172602739726
81  965.8796296296297
```

```
74  962.5622775800712
72  957.0371621621622
77  950.2307692307693
80  947.4264705882352
73  935.9757575757576
87  932.3095238095239
64  925.3102961918195
66  924.1237553342817
76  922.1910112359551
67  922.1506622516556
71  916.5284090909091
65  888.9422799422799
69  885.5152173913043
62  877.2635869565217
91  871.4605263157895
68  861.3096926713948
60  853.1457219251337
63  841.8346196251379
61  836.6322115384615
59  792.7961630695444
58  785.5576707726764
57  777.9045064377682
55  744.9394594594595
56  736.5706150341686
54  706.8937960042061
53  676.51677852349
50  661.1314699792961
51  646.2193614830072
49  645.1997896950578
52  635.7577889447236
47  612.3913525498891
48  606.4720982142857
46  571.6028202115159
45  550.0811138014528
44  541.9350057012542
42  495.9077844311377
43  490.032925682032
38  481.01182893539584
41  477.1498127340824
37  477.02579979360166
39  476.519955654102
40  456.4479041916168
36  454.01788908765656
30  453.52331002331005
32  451.75121275121273
35  450.76851063829787
```

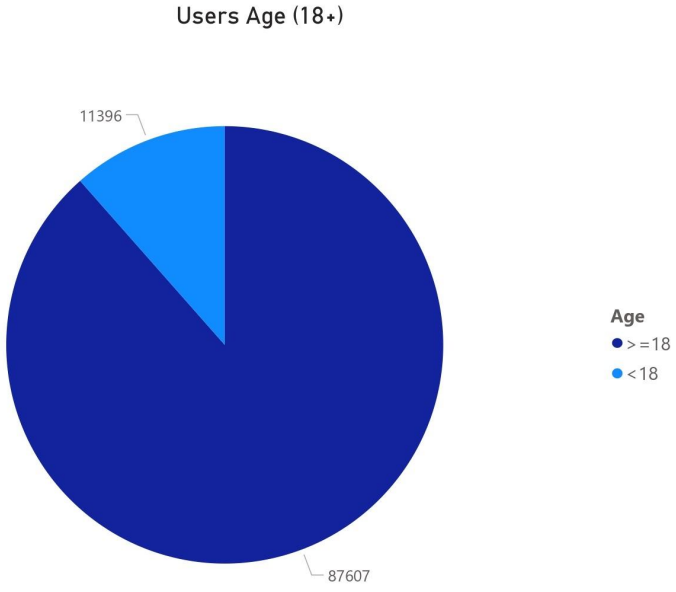
```
34  447.5067621320605
31  445.04545454545456
29  442.08677685950414
33  439.8934467233617
28  438.4039763113367
27  433.9486607142857
20  429.36879808967893
24  427.8394057304563
21  425.32797602833017
26  424.6287744227353
22  421.34630606860156
23  421.2488646684832
19  410.48235026189934
17  381.8540968626256
16  367.850615683733
18  362.28849114703615
25  350.3457841252403
15  341.4556913674561
14  259.68103896103895
13  192.0103305785124
```

5. Result and Analysis:

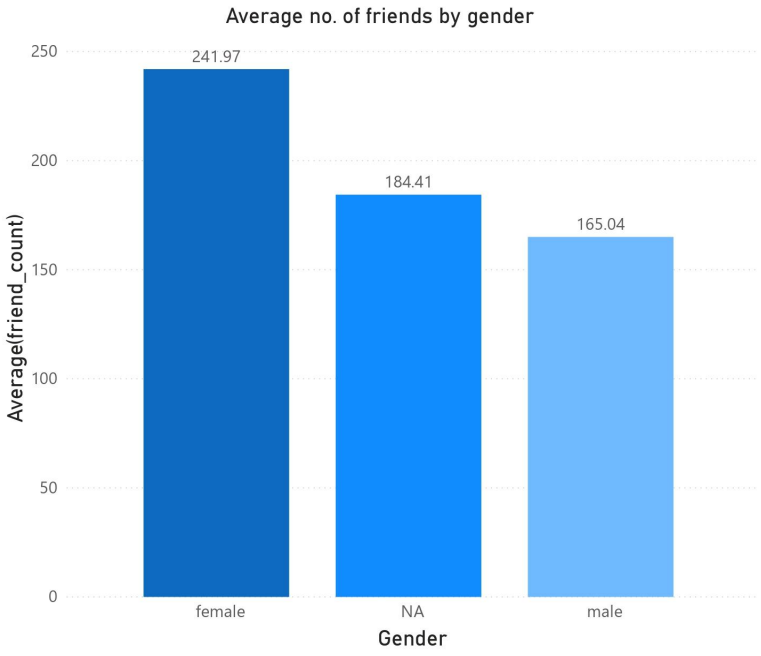
- First we loaded Facebook data into a table. Following which we performed a few queries to analyze the user's activities beginning with the total number of users in the dataset which was found to be 99003.



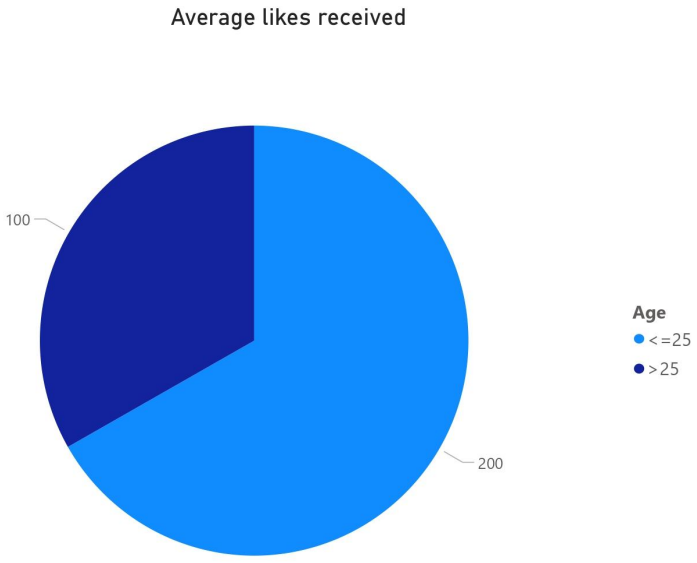
- In order to analyze the age groups we counted the number of users whose age was greater than equal to and less than 18 years. We found the values to be 87607 and 11396 respectively. It was observed that the majority of the users were 18 years of age or above.



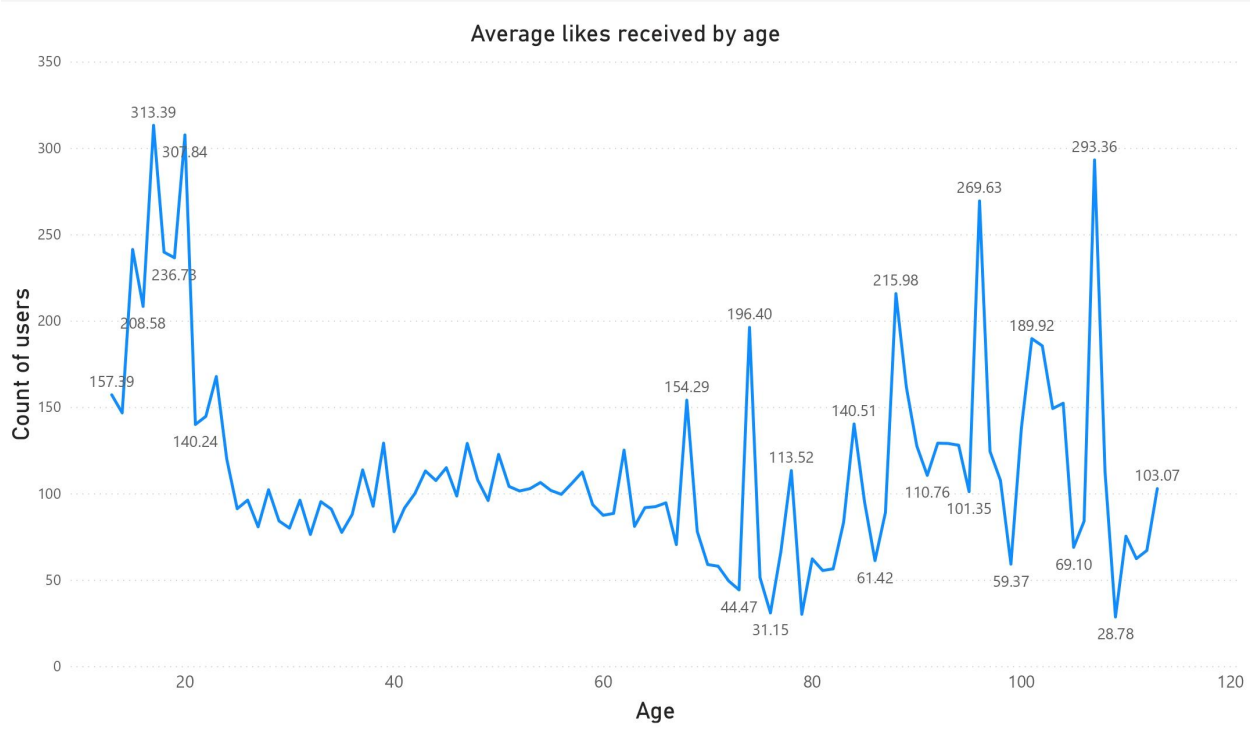
- We calculated the average number of friends based on their sex: male, female, N/A. The results showed that on an average female users had substantially more friends (242) as compared to male users (165), while the users whose gender was not disclosed showed an average of 184 friends.



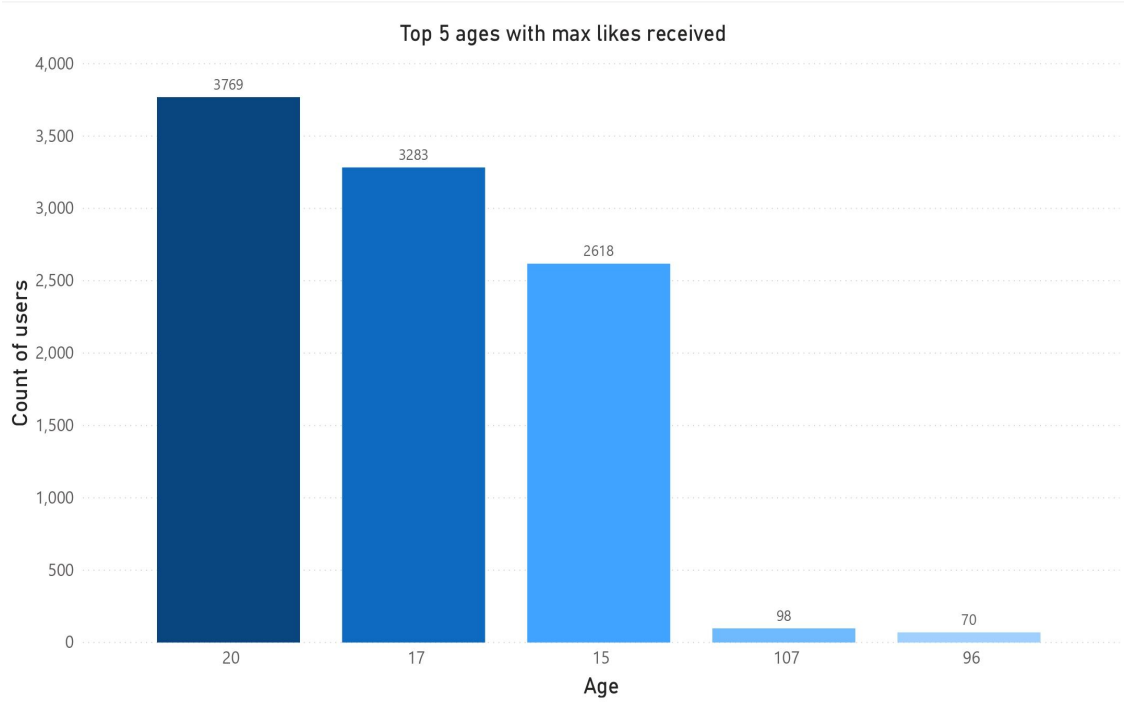
- Our next task was to compare the number of likes based on the users age groups with 25 as the threshold value. The numbers showed that on an average the younger group had twice as many likes (200) as that of the older group (100).



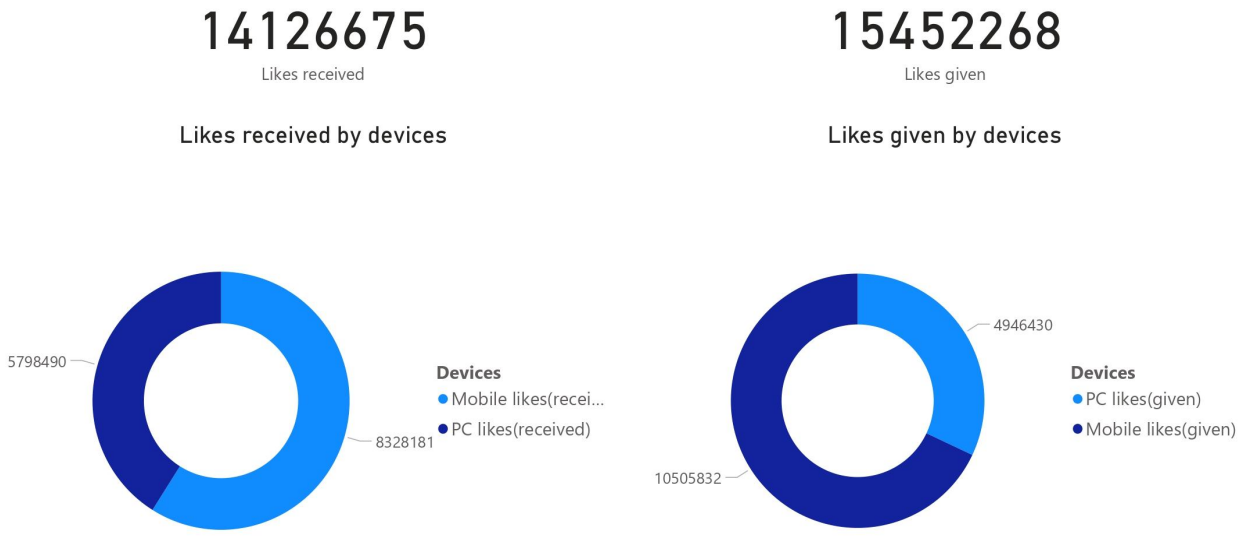
- Moreover, the average number of likes received by people of all ages and found that 17 year olds led with the maximum number of likes on an average (313 likes), followed by 20 (308 likes), 107 (293 likes), 96 (270), and 15 year olds (242 likes).



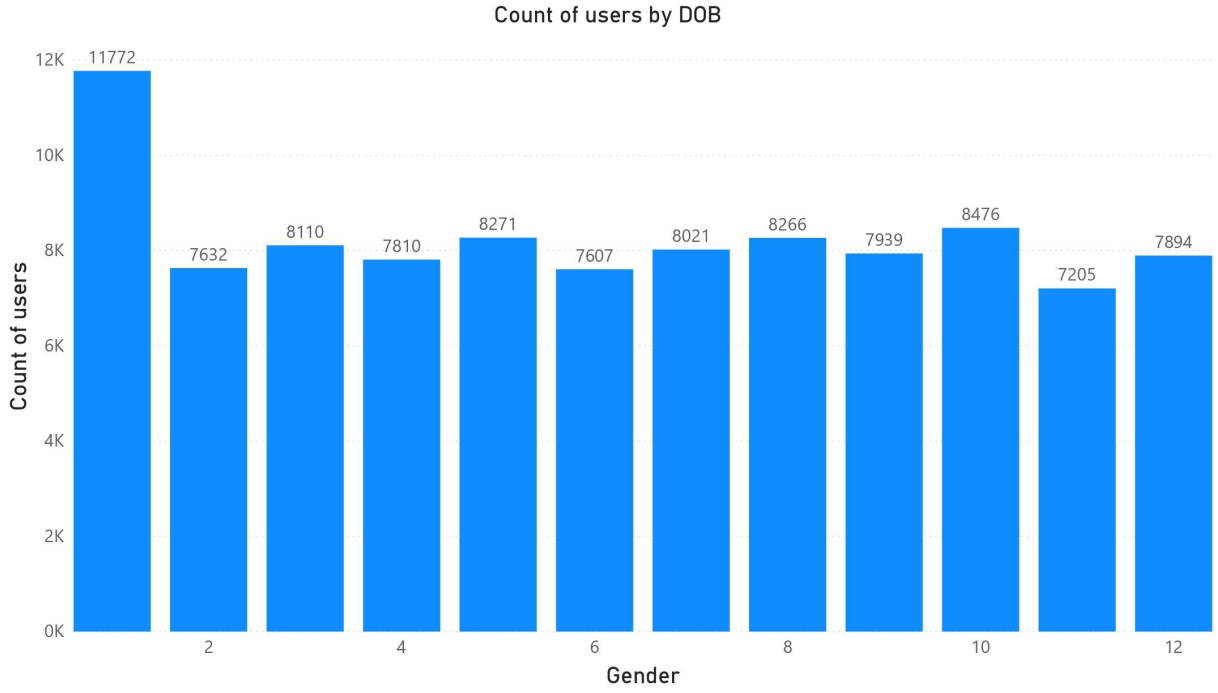
- One would expect that on an average, younger users would receive a higher number of likes. However, an outlier to this is the appearance of people in the age groups of 96 and 107. To delve into this, we analyzed the number of users in this age group. The number of users in the age group of 15, 17, 20 are around 50 times greater than the users in the 96+ age groups. Hence, we can conclude that the average likes received can be very easily skewed by a few number of users in this age group receiving a very high number of likes (famous personalities).



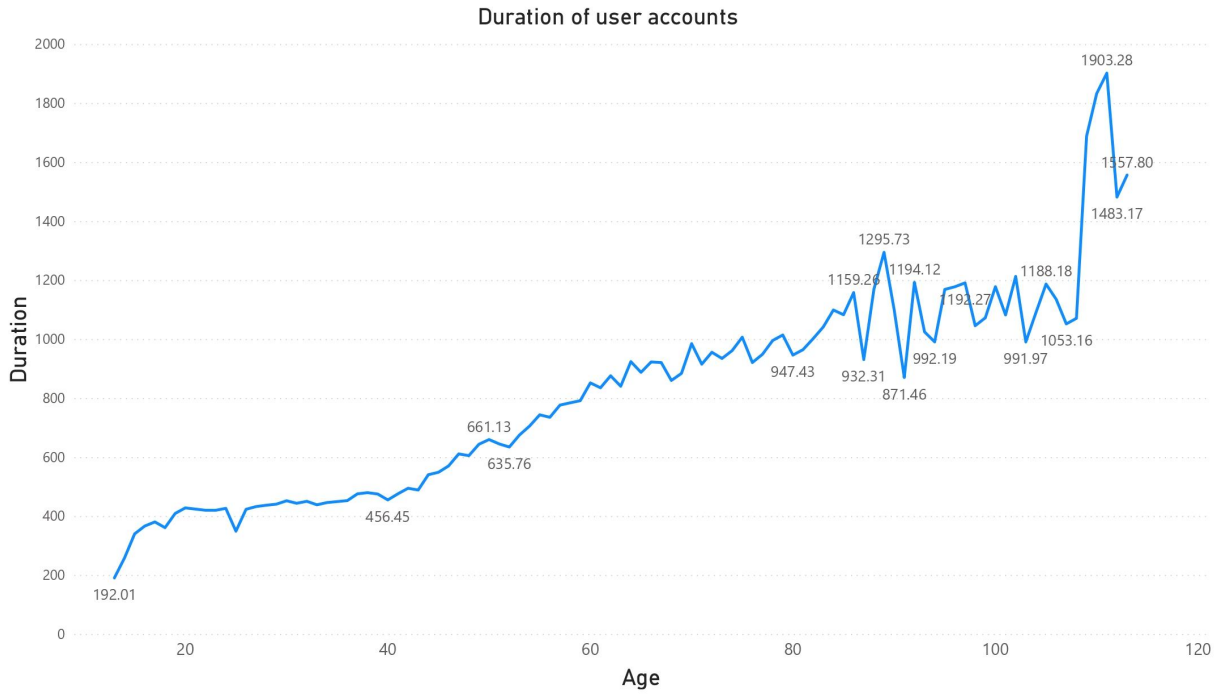
- The number of likes received/given indicates the interest/activeness of a user on Facebook. Based on this, we found that the total number of likes received by PC users and mobile users was 5798490 and 8328181 respectively. In addition, the total number of likes given by PC and mobile users were found to be 4946430 and 10505832 respectively. From these outputs, it is evident that mobile users are more active in posting content.



- Additionally we wrote queries for obtaining the count of users according to the month they were born in. The findings indicated that the maximum number of users were born in January with the count being 11772 and the least in November with 7205 users.



- Lastly, we calculated the longest duration of a user account on an average on Facebook and it resulted in 111 years with 1903 days of activity.



6. Conclusion:

Using the traditional RDBMS we cannot perform analysis on such large and complex data sets. But using Hadoop's tool: Hive, an intense analysis of the data can be done. The Hive data warehouse infrastructure tool is used to process structured data in Hadoop and we can use it for analyzing massive amounts of data. Here we performed data analysis on Facebook's data because it gives different, unique insights on several fields, and is a perfect demonstration of how we can use user data for decision making. In the future the same analysis can be done on other data as well using any of the above Hadoop technologies.