Proof of Concept

On

Title: Instagram Data Analysis

Submitted for the requirement of Big Data Engineering Course

BACHELOR OF ENGINEERING

Information Technology

(Big Data and Analytics)

Semester 5

Submitted to:

Ms. Gurpreet Kaur

Project Supervisor

Submitted by:

Anshika Singh(2001220130019)

Mohammad Monis Khan(2001220130060)

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 Instagram Dataset" using Apache Hadoop (Hive), IBM Cognos Insight and Python Programming, and gave suggestions about the task.

OVERVIEW

Instagram is a popular social mobile app that is focused on sharing images and video. For companies, you can share your day-to-day happenings and behind-the-scenes moments of running your business. Instagram is mostly about sharing your own content although you can “regram” or repost other people’s images as well.

Instagram now offers business profiles, and you can easily convert your current personal account to business in Settings on the mobile app. Setting up a business profile is advantageous for company’s because you can include a direct way for people to reach you by adding your phone number, email, or location. You can also access analytics to better measure your Instagram marketing results.

When you set up your business profile, Instagram pulls some information directly from your company’s Facebook page – both should be linked. This includes your business profile category so if you need to change it, do so in the Settings of your Facebook Page. Double check your business profile Contact button to make sure you’re providing the best way for people to contact your company.

COLUMNS AND DATA TYPE :

userid int

likes int

days_passed_from int

likes_score int

type string

no_of_tags int

no_of_comments int

date_posted int

year int

month int

day int

minute int

PROBLEM STATEMENTS:

1)Find the count of users.

2)Determine the average likes on posts of users.

3) Determine the average comments on posts of users.

4)Analyse the count of users who spent less than 10 minutes after post.

5)Find out the difference of likes and active users between any two consecutive years.

6)Determine the post with more than 3 tags.

7)Determine average like scores in a month when the app was usage by the users was more but likes were low compared to the usage.

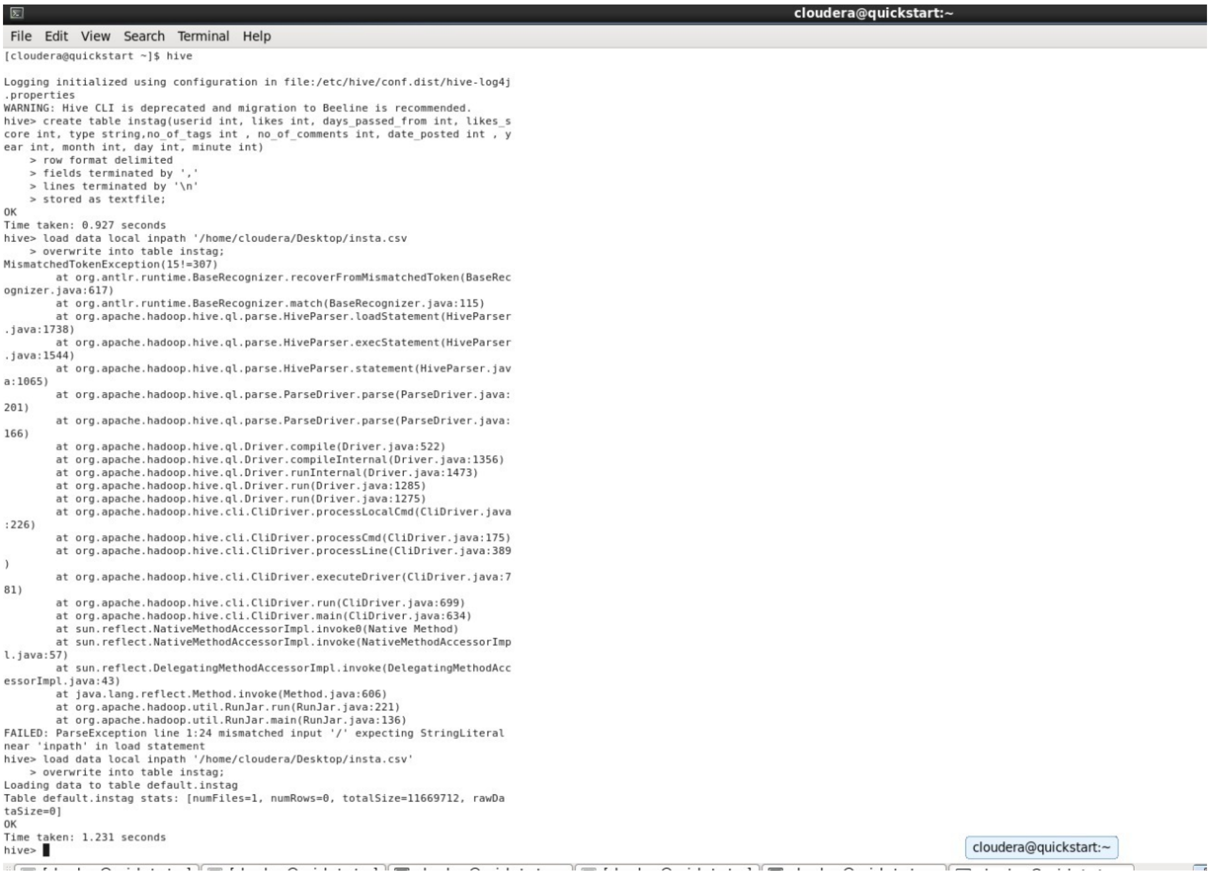
8)Visualisation graph for number of likes in a year.

9)Visualisation graph for number of tags and comments.

10) Visualisation graph for like score in a year.

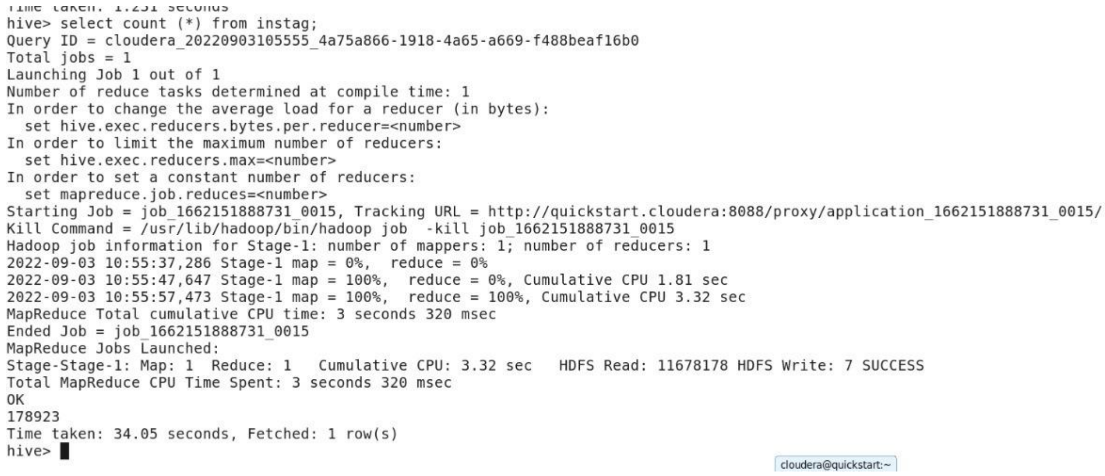
HIVE QUERIES

Data Loading



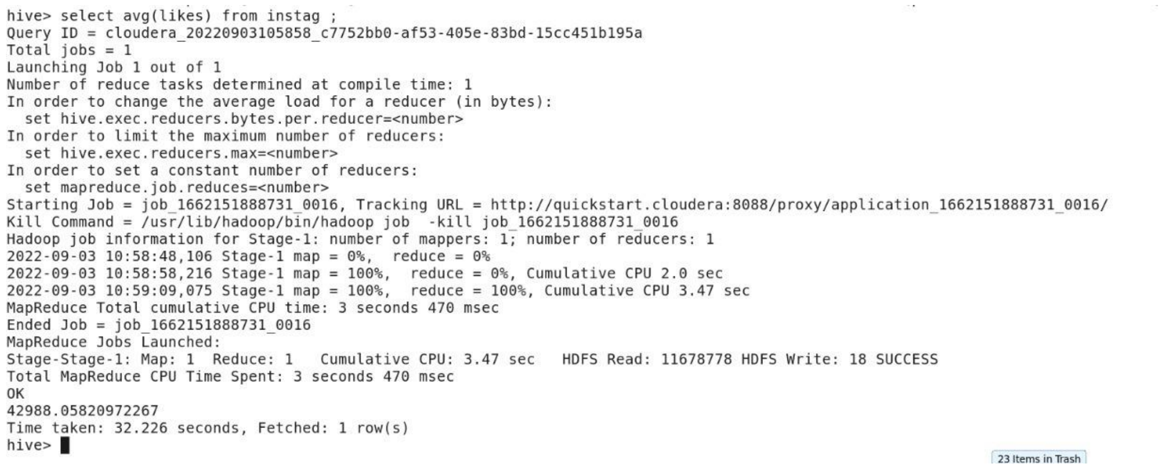
```
[cloudera@quickstart ~]$ hive

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 instag(userid int, likes int, days_passed_from int, likes_s
core int, type string,no_of_tags int , no_of_comments int, date_posted int , y
ear int, month int, day int, minute int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
    > stored as textfile;
OK
Time taken: 0.927 seconds
hive> load data local inpath '/home/cloudera/Desktop/insta.csv
    > overwrite into table instag;
MismatchedTokenException(15!=307)
        at org.antlr.runtime.BaseRecognizer.recoverFromMismatchedToken(BaseRec
ognizer.java:617)
        at org.antlr.runtime.BaseRecognizer.match(BaseRecognizer.java:115)
        at org.apache.hadoop.hive.ql.parse.HiveParser.loadStatement(HiveParser
.java:1738)
        at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser
.java:1544)
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.jav
a:1065)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:
201)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:
166)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:522)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1356)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1473)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1285)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1275)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java
:226)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:175)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:389
)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:7
81)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:699)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:634)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImp
l.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcc
essorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:24 mismatched input '/' expecting StringLiteral
near 'inpath' in load statement
hive> load data local inpath '/home/cloudera/Desktop/insta.csv'
    > overwrite into table instag;
Loading data to table default.instag
Table default.instag stats: [numFiles=1, numRows=0, totalSize=11669712, rawDa
taSize=0]
OK
Time taken: 1.231 seconds
hive> 
```

PROBLEM STATEMENT 1: Find the count of users.



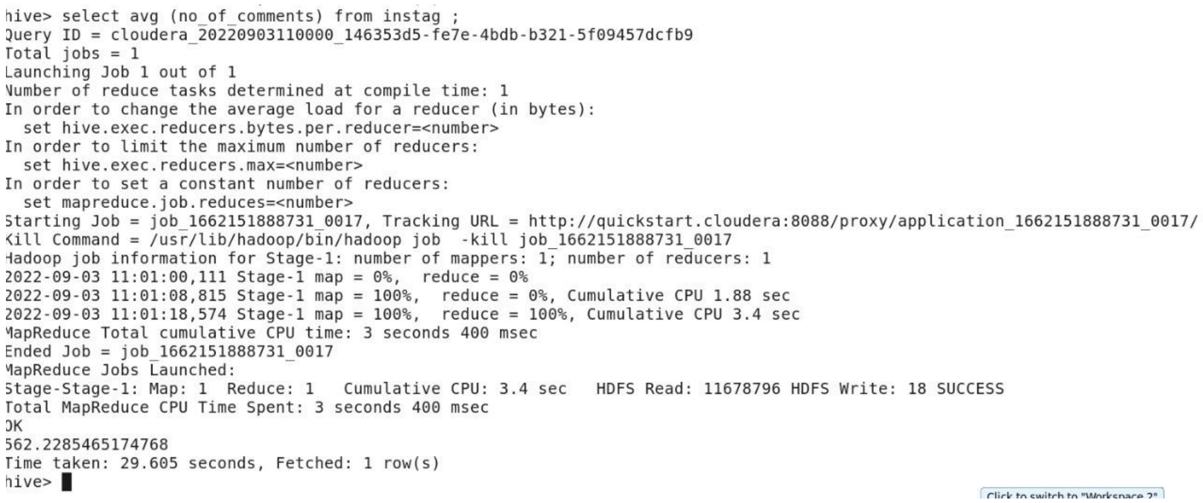
```
Time taken: 1.231 seconds
hive> select count (*) from instag;
Query ID = cloudera_20220903105555_4a75a866-1918-4a65-a669-f488beaf16b0
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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_1662151888731_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 10:55:37,286 Stage-1 map = 0%,  reduce = 0%
2022-09-03 10:55:47,647 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.81 sec
2022-09-03 10:55:57,473 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.32 sec
MapReduce Total cumulative CPU time: 3 seconds 320 msec
Ended Job = job_1662151888731_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.32 sec   HDFS Read: 11678178 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 320 msec
OK
178923
Time taken: 34.05 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 2: Determine the average likes on posts of users.



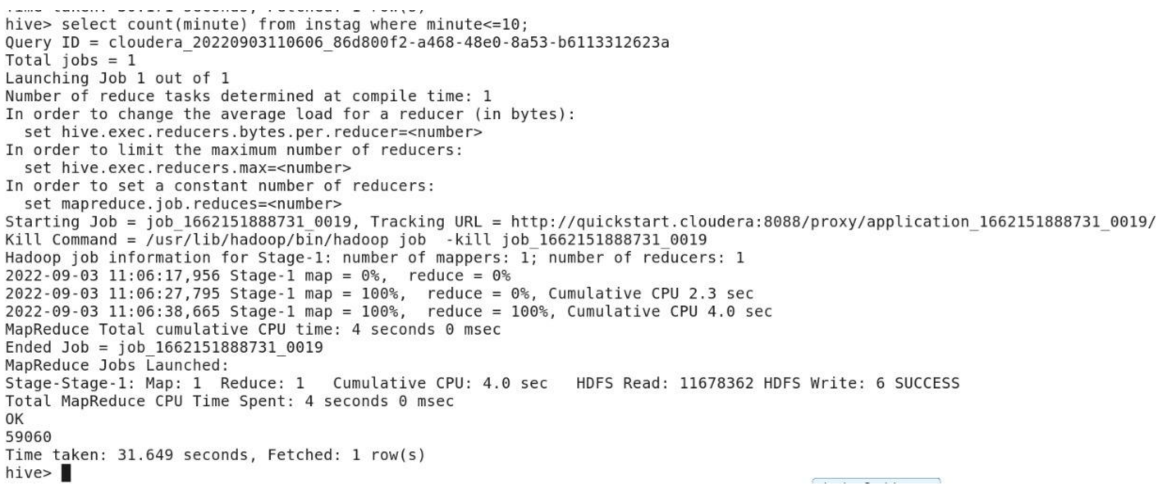
```
hive> select avg(likes) from instag ;
Query ID = cloudera_20220903105858_c7752bb0-af53-405e-83bd-15cc451b195a
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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_1662151888731_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 10:58:48,106 Stage-1 map = 0%,  reduce = 0%
2022-09-03 10:58:58,216 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.0 sec
2022-09-03 10:59:09,075 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.47 sec
MapReduce Total cumulative CPU time: 3 seconds 470 msec
Ended Job = job_1662151888731_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.47 sec   HDFS Read: 11678778 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 470 msec
OK
42988.05820972267
Time taken: 32.226 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 3: Determine the average comments on posts of users.



```
hive> select avg (no_of_comments) from instag ;
Query ID = cloudera_20220903110000_146353d5-fe7e-4bdb-b321-5f09457dcfb9
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_1662151888731_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:01:00,111 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:01:08,815 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.88 sec
2022-09-03 11:01:18,574 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.4 sec
MapReduce Total cumulative CPU time: 3 seconds 400 msec
Ended Job = job_1662151888731_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.4 sec   HDFS Read: 11678796 HDFS Write: 18 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 400 msec
OK
562.2285465174768
Time taken: 29.605 seconds, Fetched: 1 row(s)
hive> █
```

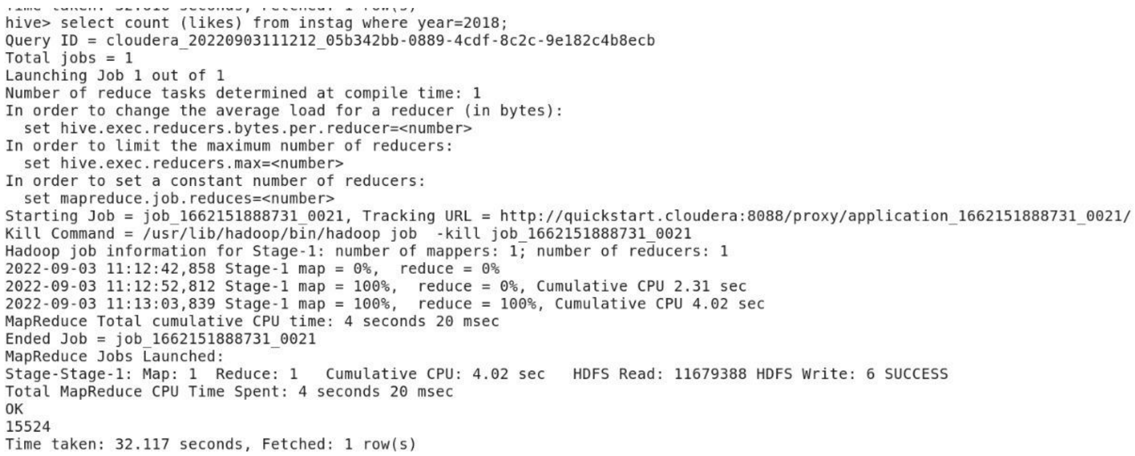
PROBLEM STATEMENT 4: Analyse the count of users who spent less than 10 minutes after post.



```
hive> select count(minute) from instag where minute<=10;
Query ID = cloudera_20220903110606_86d800f2-a468-48e0-8a53-b6113312623a
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_1662151888731_0019, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:06:17,956 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:06:27,795 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.3 sec
2022-09-03 11:06:38,665 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.0 sec
MapReduce Total cumulative CPU time: 4 seconds 0 msec
Ended Job = job_1662151888731_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.0 sec   HDFS Read: 11678362 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 0 msec
OK
59060
Time taken: 31.649 seconds, Fetched: 1 row(s)
hive> █
```

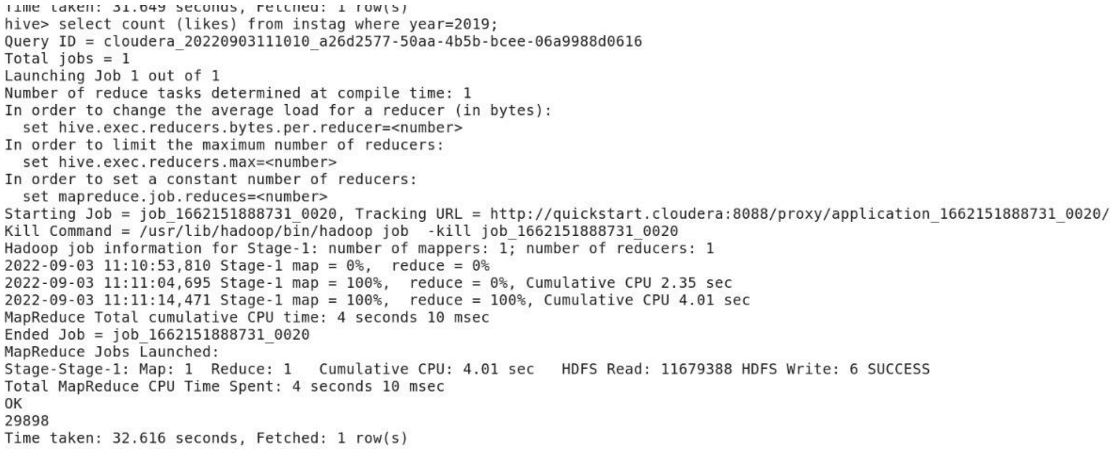
PROBLEM STATEMENT 5: Find out the difference of likes and active users between any two consecutive years.

2018



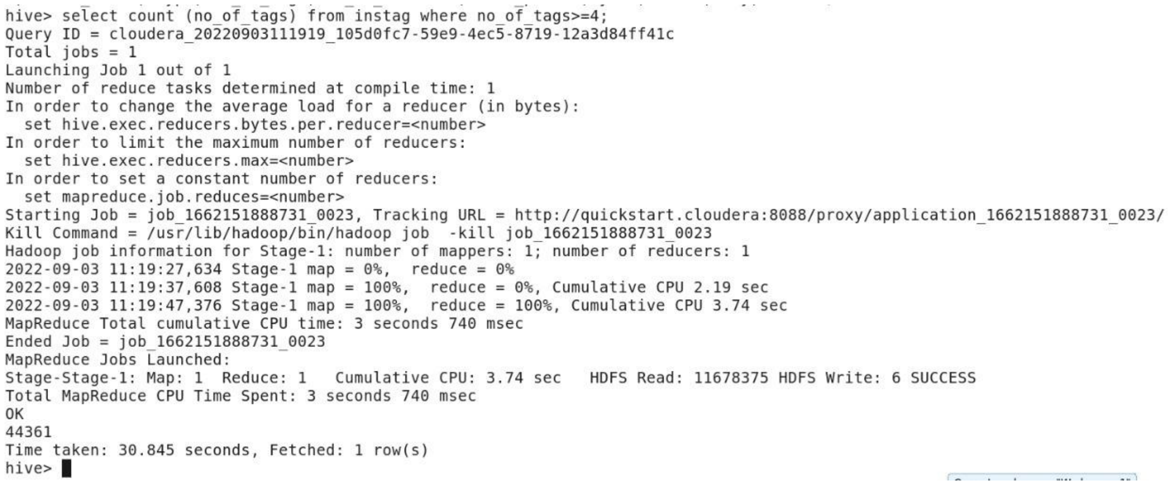
```
Time taken: 32.616 seconds, Fetched: 1 row(s)
hive> select count (likes) from instag where year=2018;
Query ID = cloudera_20220903111212_05b342bb-0889-4cdf-8c2c-9e182c4b8ecb
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average 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_1662151888731_0021, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:12:42,858 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:12:52,812 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.31 sec
2022-09-03 11:13:03,839 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.02 sec
MapReduce Total cumulative CPU time: 4 seconds 20 msec
Ended Job = job_1662151888731_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.02 sec   HDFS Read: 11679388 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 20 msec
OK
15524
Time taken: 32.117 seconds, Fetched: 1 row(s)
```

2019



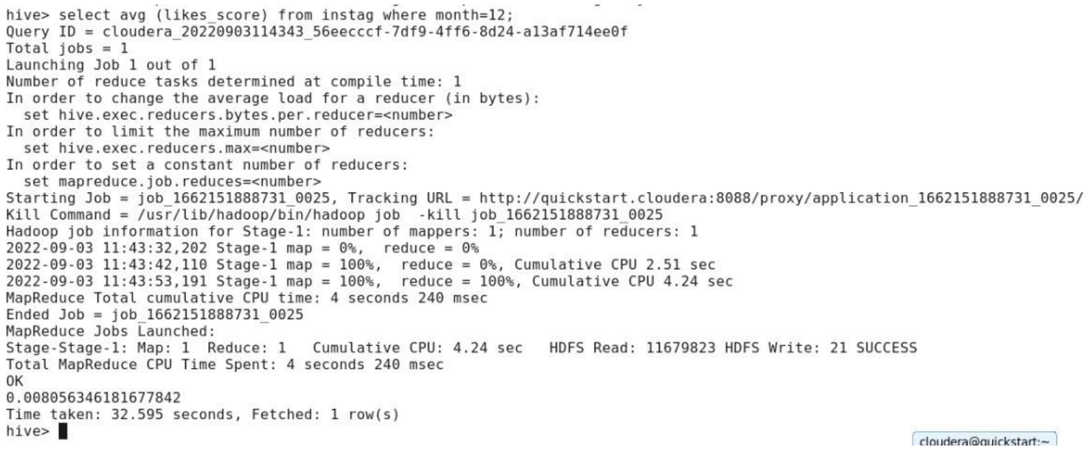
```
Time taken: 31.649 seconds, Fetched: 1 row(s)
hive> select count (likes) from instag where year=2019;
Query ID = cloudera_20220903111010_a26d2577-50aa-4b5b-bcee-06a9988d0616
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average 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_1662151888731_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:10:53,810 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:11:04,695 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.35 sec
2022-09-03 11:11:14,471 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.01 sec
MapReduce Total cumulative CPU time: 4 seconds 10 msec
Ended Job = job_1662151888731_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.01 sec   HDFS Read: 11679388 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 10 msec
OK
29898
Time taken: 32.616 seconds, Fetched: 1 row(s)
```

PROBLEM STATEMENT 6: Determine the post with more than 3 tags.



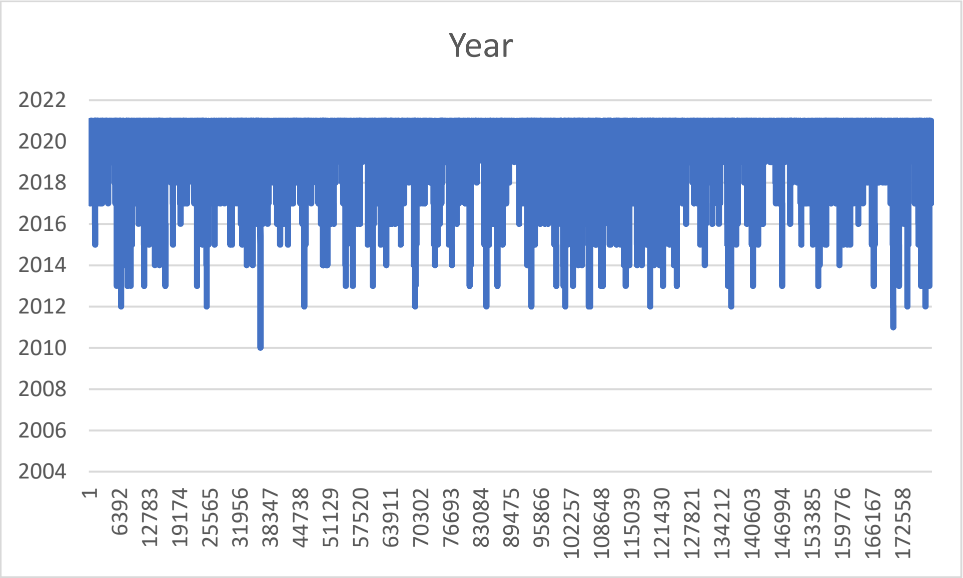
```
hive> select count (no_of_tags) from instag where no_of_tags>=4;
Query ID = cloudera_20220903111919_105d0fc7-59e9-4ec5-8719-12a3d84ff41c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set 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_1662151888731_0023, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0023/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:19:27,634 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:19:37,608 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.19 sec
2022-09-03 11:19:47,376 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.74 sec
MapReduce Total cumulative CPU time: 3 seconds 740 msec
Ended Job = job_1662151888731_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.74 sec   HDFS Read: 11678375 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 740 msec
OK
44361
Time taken: 30.845 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 7: Determine average like scores in a month when the app was usage by the users was more but likes were low compared to the usage.

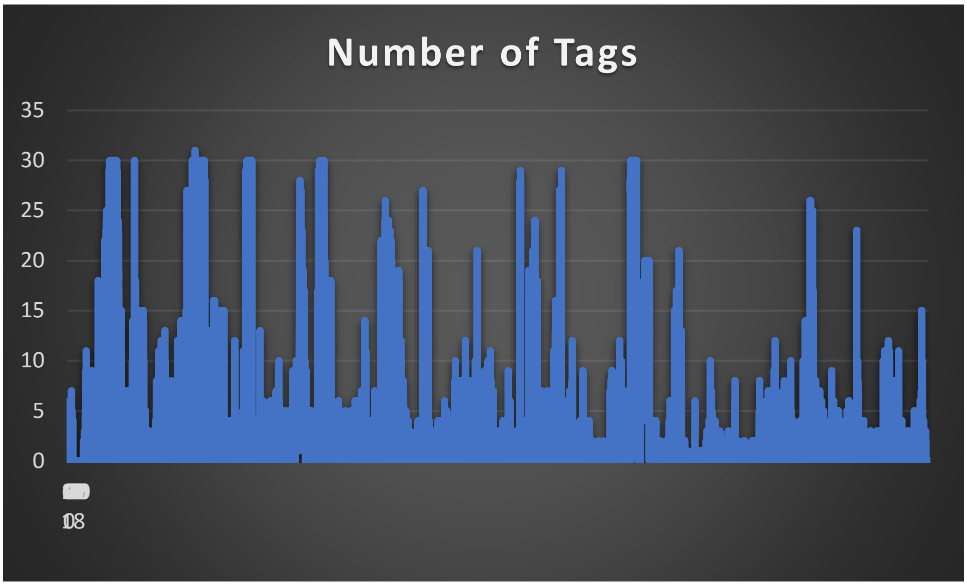


```
hive> select avg (likes_score) from instag where month=12;
Query ID = cloudera_20220903114343_56eecccf-7df9-4ff6-8d24-a13af714ee0f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<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_1662151888731_0025, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1662151888731_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1662151888731_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 11:43:32,202 Stage-1 map = 0%,  reduce = 0%
2022-09-03 11:43:42,110 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.51 sec
2022-09-03 11:43:53,191 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.24 sec
MapReduce Total cumulative CPU time: 4 seconds 240 msec
Ended Job = job_1662151888731_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.24 sec   HDFS Read: 11679823 HDFS Write: 21 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 240 msec
OK
0.008056346181677842
Time taken: 32.595 seconds, Fetched: 1 row(s)
hive> █
```

PROBLEM STATEMENT 8: Visualisation graph for number of likes in a year.



PROBLEM STATEMENT 9: Visualisation graph for number of tags and comments.



PROBLEM STATEMENT 10: Visualisation graph for like score in a year.

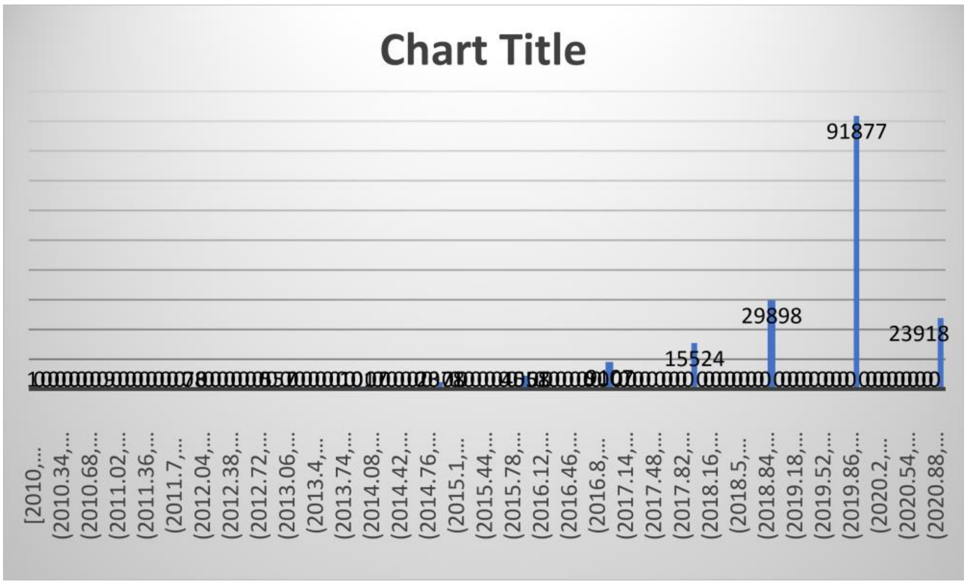


Chart Title
[2010,...
(2010.34,...
(2010.68,...
(2011.02,...
(2011.36,...
(2011.7,...
(2012.04,...
(2012.38,...
(2012.72,...
(2013.06,...
(2013.4,...
(2013.74,...
(2014.08,...
(2014.42,...
(2014.76,...
(2015.1,...
(2015.44,...
(2015.78,...
(2016.12,...
(2016.46,...
(2016.8,...
(2017.14,...
(2017.48,...
(2017.82,...
(2018.16,...
(2018.5,...
(2018.84,...
(2019.18,...
(2019.52,...
(2019.86,...
(2020.2,...
(2020.54,...
(2020.88,...