SMIT R PATEL

19162121031

SEM 5

PRACTICAL 13

HIVE

BIG DATA AND ANALYTICS

AIM- To work with Hive in Hadoop.

Exercise: You work as a data analyst for a bank, which now needs you to analyse a few things mentioned below in order to launch a new scheme.

Tasks:

1. Create Table for the Bank Dataset using following columns:

   1 - age (numeric)
   2 - job: type of job
   ("admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
   3 - marital : marital status ("married","divorced","single"; note: "divorced" means divorced or widowed)
   4 - education ( "unknown","secondary","primary","tertiary")
   5 - default: has credit in default? ("yes","no")
   6 - balance: average yearly balance, in euros (numeric)
   7 - housing: has housing loan? ("yes","no")
   8 - loan: has personal loan? ("yes","no")
   9 - contact: contact communication type ("unknown","telephone","cellular")
   10 - day: last contact day of the month (numeric)
   11 - month: last contact month of year "jan", "feb", "mar", ..., "nov", "dec")
   12 - duration: last contact duration, in seconds (numeric)
   13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
   14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
   15 - previous: number of contacts performed before this campaign and for this client (numeric)
   16 - poutcome: outcome of the previous marketing campaign ( "unknown","other","failure","success")
   17 - y - has the client subscribed a term deposit? ("yes","no")

Command:-

Hive>create database bank;

Hive>create table bankdata;

>(age INT, job STRING, marital STRING, education STRING, default STRING, balance INT, housing STRING, loan STRING, contact STRING, day int, month STRING, duration INT, compaign INT, pdays INT, previous INT, poutcome STRING, y STRING)

>row format delimited fields terminated by ',' ;

```
hive> create database bank;
OK
Time taken: 0.078 seconds
hive> create table bankdata
    > (age INT, job STRING, marital STRING, education STRING, default STRING, ba
lance INT, housing STRING, loan STRING, contact STRING, day int, month STRING, d
uration INT, compaign INT, pdays INT, previous INT, poutcome STRING, y STRING)
    > row format delimited fields terminated by ',';
OK
Time taken: 0.177 seconds
hive>
```

2 understand the schema, LOAD file in to table

command :-

Hive>LOAD DATA LOCAL INPATH 'bankfull.csv' INTO TABLE bankdata;

Hive>describe bank;

```
hive> LOAD DATA LOCAL INPATH 'bankfull.csv' INTO TABLE bankdata;
Loading data to table default.bankdata
Table default.bankdata stats: [numFiles=1, totalSize=3751306]
OK
Time taken: 0.965 seconds
hive> describe bankdata;
OK
age                     int
job                     string
marital                 string
education               string
default                 string
balance                 int
housing                 string
loan                    string
contact                 string
day                     int
month                   string
duration                int
compaign                int
pdays                   int
previous                int
poutcome                string
y                       string
Time taken: 0.433 seconds, Fetched: 17 row(s)
hive>
```

3 with the use of AND Logical operator get the records who has age greater than 25 and married.

Command :-

select * from bankdata where age>25 and marital='married' limit 50;

```
hive> select * from bankdata where age>25 and marital='married' limit 50;
OK
58      management      married tertiary        no      2143    yes     no      u
nknown  5       may     261     1       -1      0       unknown no
33      entrepreneur    married secondary       no      2       yes     yes     u
nknown  5       may     76      1       -1      0       unknown no
47      blue-collar     married unknown no      1506    yes     no      unknown5
may     92      1       -1      0       unknown no
35      management      married tertiary        no      231     yes     no      u
nknown  5       may     139     1       -1      0       unknown no
58      retired married primary no      121     yes     no      unknown 5       m
ay      50      1       -1      0       unknown no
53      technician      married secondary       no      6       yes     no      u
nknown  5       may     517     1       -1      0       unknown no
58      technician      married unknown no      71      yes     no      unknown5
may     71      1       -1      0       unknown no
57      services        married secondary       no      162     yes     no      u
nknown  5       may     174     1       -1      0       unknown no
51      retired married primary no      229     yes     no      unknown 5       m
ay      353     1       -1      0       unknown no
57      blue-collar     married primary no      52      yes     no      unknown5
may     38      1       -1      0       unknown no
60      retired married primary no      60      yes     no      unknown 5       m
ay      219     1       -1      0       unknown no
33      services        married secondary       no      0       yes     no      u
nknown  5       may     54      1       -1      0       unknown no
28      blue-collar     married secondary       no      723     yes     yes     u
nknown  5       may     262     1       -1      0       unknown no
56      management      married tertiary        no      779     yes     no      u
nknown  5       may     164     1       -1      0       unknown no
40      retired married primary no      0       yes     yes     unknown 5       m
ay      181     1       -1      0       unknown no
44      admin.  married secondary       no      -372    yes     no      unknown5
may     172     1       -1      0       unknown no
52      entrepreneur    married secondary       no      113     yes     yes     u
nknown  5       may     127     1       -1      0       unknown no
57      technician      married secondary       no      839     no      yes     u
nknown  5       may     225     1       -1      0       unknown no
49      management      married tertiary        no      378     yes     no      u
nknown  5       may     230     1       -1      0       unknown no
60      admin.  married secondary       no      39      yes     yes     unknown5
may     208     1       -1      0       unknown no
59      blue-collar     married secondary       no      0       yes     no      u
nknown  5       may     226     1       -1      0       unknown no
51      management      married tertiary        no      10635   yes     no      u
nknown  5       may     336     1       -1      0       unknown no
cloudera@quickstart:~...
```

Home | cloudera-quickstart-vm-5.1...

Applications Places System Fri Sep 24, 12:45 AM cloudera

cloudera@quickstart:~/Desktop

File Edit View Search Terminal Help

```
44      technician      married secondary       no      0       yes     no      u
nknown  5       may     225     2       -1      0       unknown no
54      blue-collar     married secondary       no      1291    yes     no      u
nknown  5       may     266     1       -1      0       unknown no
32      management      married tertiary        no      0       yes     no      u
nknown  5       may     179     1       -1      0       unknown no
47      blue-collar     married unknown no      306     yes     no      unknown5
may     13      1       -1      0       unknown no
46      services        married primary no      179     yes     no      unknown5
may     1778    1       -1      0       unknown no
32      admin.  married tertiary        no      0       yes     no      unknown5
may     138     1       -1      0       unknown no
57      blue-collar     married primary no      249     yes     no      unknown5
may     164     1       -1      0       unknown no
33      services        married secondary       no      790     yes     no      u
nknown  5       may     391     1       -1      0       unknown no
49      blue-collar     married unknown no      154     yes     no      unknown5
may     357     1       -1      0       unknown no
51      management      married tertiary        no      6530    yes     no      u
nknown  5       may     91      1       -1      0       unknown no
60      retired married tertiary        no      100     no      no      unknown5
may     528     1       -1      0       unknown no
55      technician      married secondary       no      1205    yes     no      u
nknown  5       may     158     2       -1      0       unknown no
57      blue-collar     married secondary       no      5935    yes     yes     u
nknown  5       may     258     1       -1      0       unknown no
31      services        married secondary       no      25      yes     yes     u
nknown  5       may     172     1       -1      0       unknown no
54      management      married secondary       no      282     yes     yes     u
nknown  5       may     154     1       -1      0       unknown no
55      blue-collar     married primary no      23      yes     no      unknown5
may     291     1       -1      0       unknown no
43      technician      married secondary       no      1937    yes     no      u
nknown  5       may     181     1       -1      0       unknown no
53      technician      married secondary       no      384     yes     no      u
nknown  5       may     176     1       -1      0       unknown no
44      blue-collar     married secondary       no      582     no      yes     u
nknown  5       may     211     1       -1      0       unknown no
59      admin.  married secondary       no      2343    yes     no      unknown5
may     1042    1       -1      0       unknown yes
46      self-employed   married tertiary        no      137     yes     yes     u
nknown  5       may     246     1       -1      0       unknown no
51      blue-collar     married primary no      173     yes     no      unknown5
may     529     2       -1      0       unknown no
Time taken: 1.181 seconds, Fetched: 50 row(s)
hive>
```

cloudera@quickstart:~...

4. Who has not subscribed to a term deposit (column: y)

Command:-
Hive> select * from bank where y='no' limit 50;

```
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help
hive> select * from bankdata where y='no' limit 50;
OK
58      management      married tertiary        no      2143    yes     no      unknown 5  m
ay      261     1       -1      0       unknown no
44      technician      single  secondary       no      29      yes     no      unknown 5  m
ay      151     1       -1      0       unknown no
33      entrepreneur    married secondary       no      2       yes     yes     unknown 5  m
ay      76      1       -1      0       unknown no
47      blue-collar     married unknown no      1506    yes     no      unknown 5       may9
2       1       -1      0       unknown no
33      unknown single  unknown no      1       no      no      unknown 5       may     1981
-1      0       unknown no
35      management      married tertiary        no      231     yes     no      unknown 5  m
ay      139     1       -1      0       unknown no
28      management      single  tertiary        no      447     yes     yes     unknown 5  m
ay      217     1       -1      0       unknown no
42      entrepreneur    divorced        tertiary        yes     2       yes     no      unkn
own     5       may     380     1       -1      0       unknown no
58      retired married primary no      121     yes     no      unknown 5       may     50 1
-1      0       unknown no
43      technician      single  secondary       no      593     yes     no      unknown 5  m
ay      55      1       -1      0       unknown no
41      admin.  divorced        secondary       no      270     yes     no      unknown 5  m
ay      222     1       -1      0       unknown no
29      admin.  single  secondary       no      390     yes     no      unknown 5       may1
37      1       -1      0       unknown no
53      technician      married secondary       no      6       yes     no      unknown 5  m
ay      517     1       -1      0       unknown no
58      technician      married unknown no      71      yes     no      unknown 5       may7
1       1       -1      0       unknown no
57      services        married secondary       no      162     yes     no      unknown 5  m
ay      174     1       -1      0       unknown no
51      retired married primary no      229     yes     no      unknown 5       may     3531
-1      0       unknown no
45      admin.  single  unknown no      13      yes     no      unknown 5       may     98 1
-1      0       unknown no
57      blue-collar     married primary no      52      yes     no      unknown 5       may3
8       1       -1      0       unknown no
60      retired married primary no      60      yes     no      unknown 5       may     2191
-1      0       unknown no
33      services        married secondary       no      0       yes     no      unknown 5  m
ay      54      1       -1      0       unknown no
28      blue-collar     married secondary       no      723     yes     yes     unknown 5  m
ay      262     1       -1      0       unknown no
56      management      married tertiary        no      779     yes     no      unknown 5  m
ay      164     1       -1      0       unknown no
cloudera@quickstart:~...
ve the mouse pointer inside or press Ctrl+G.
```

cloudera@quickstart:~/Desktop

File Edit View Search Terminal Help

```
46      management      single  secondary       no      -246    yes     no      unknown 5  m
ay      255     2       -1      0       unknown no
36      technician      single  secondary       no      265     yes     yes     unknown 5  m
ay      348     1       -1      0       unknown no
57      technician      married secondary       no      839     no      yes     unknown 5  m
ay      225     1       -1      0       unknown no
49      management      married tertiary        no      378     yes     no      unknown 5  m
ay      230     1       -1      0       unknown no
60      admin.  married secondary       no      39      yes     yes     unknown 5       may2
08      1       -1      0       unknown no
59      blue-collar     married secondary       no      0       yes     no      unknown 5  m
ay      226     1       -1      0       unknown no
51      management      married tertiary        no      10635   yes     no      unknown 5  m
ay      336     1       -1      0       unknown no
57      technician      divorced        secondary       no      63      yes     no      unkn
own     5       may     242     1       -1      0       unknown no
25      blue-collar     married secondary       no      -7      yes     no      unknown 5  m
ay      365     1       -1      0       unknown no
53      technician      married secondary       no      -3      no      no      unknown 5  m
ay      1666    1       -1      0       unknown no
36      admin.  divorced        secondary       no      506     yes     no      unknown 5  m
ay      577     1       -1      0       unknown no
37      admin.  single  secondary       no      0       yes     no      unknown 5       may1
37      1       -1      0       unknown no
44      services        divorced        secondary       no      2586    yes     no      unkn
own     5       may     160     1       -1      0       unknown no
50      management      married secondary       no      49      yes     no      unknown 5  m
ay      180     2       -1      0       unknown no
60      blue-collar     married unknown no      104     yes     no      unknown 5       may2
2       1       -1      0       unknown no
54      retired married secondary       no      529     yes     no      unknown 5       may1
492     1       -1      0       unknown no
58      retired married unknown no      96      yes     no      unknown 5       may     6161
-1      0       unknown no
36      admin.  single  primary no      -171    yes     no      unknown 5       may     2421
-1      0       unknown no
58      self-employed   married tertiary        no      -364    yes     no      unknown 5  m
ay      355     1       -1      0       unknown no
44      technician      married secondary       no      0       yes     no      unknown 5  m
ay      225     2       -1      0       unknown no
55      technician      divorced        secondary       no      0       no      no      unkn
own     5       may     160     1       -1      0       unknown no
29      management      single  tertiary        no      0       yes     no      unknown 5  m
ay      363     1       -1      0       unknown no
Time taken: 0.268 seconds, Fetched: 50 row(s)
hive>
```

cloudera@quickstart:~...

5. List the age group of the people who have subscribed to a term deposit.
Command:-
Hive> select distinct(age) from bank where y='yes' limit 50;

```
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help
hive> select distinct(age) from bankdata where y='yes' limit 50;
Query ID = cloudera_20210924005151_1110e23b-fcd0-4751-b5b1-c44b812966eb
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_1632468285119_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1632468285119_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632468285119_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-24 00:51:54,737 Stage-1 map = 0%,  reduce = 0%
2021-09-24 00:52:15,174 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.81 sec
2021-09-24 00:52:35,225 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.41 sec
MapReduce Total cumulative CPU time: 8 seconds 410 msec
Ended Job = job_1632468285119_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.41 sec   HDFS Read: 3759971 HDFS Write:
 150 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 410 msec
OK
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
cloudera@quickstart:~...
```

Home | cloudera-quickstart-vm-5.1...

Applications Places System Fri Sep 24, 12:53 AM cloudera

cloudera@quickstart:~/Desktop

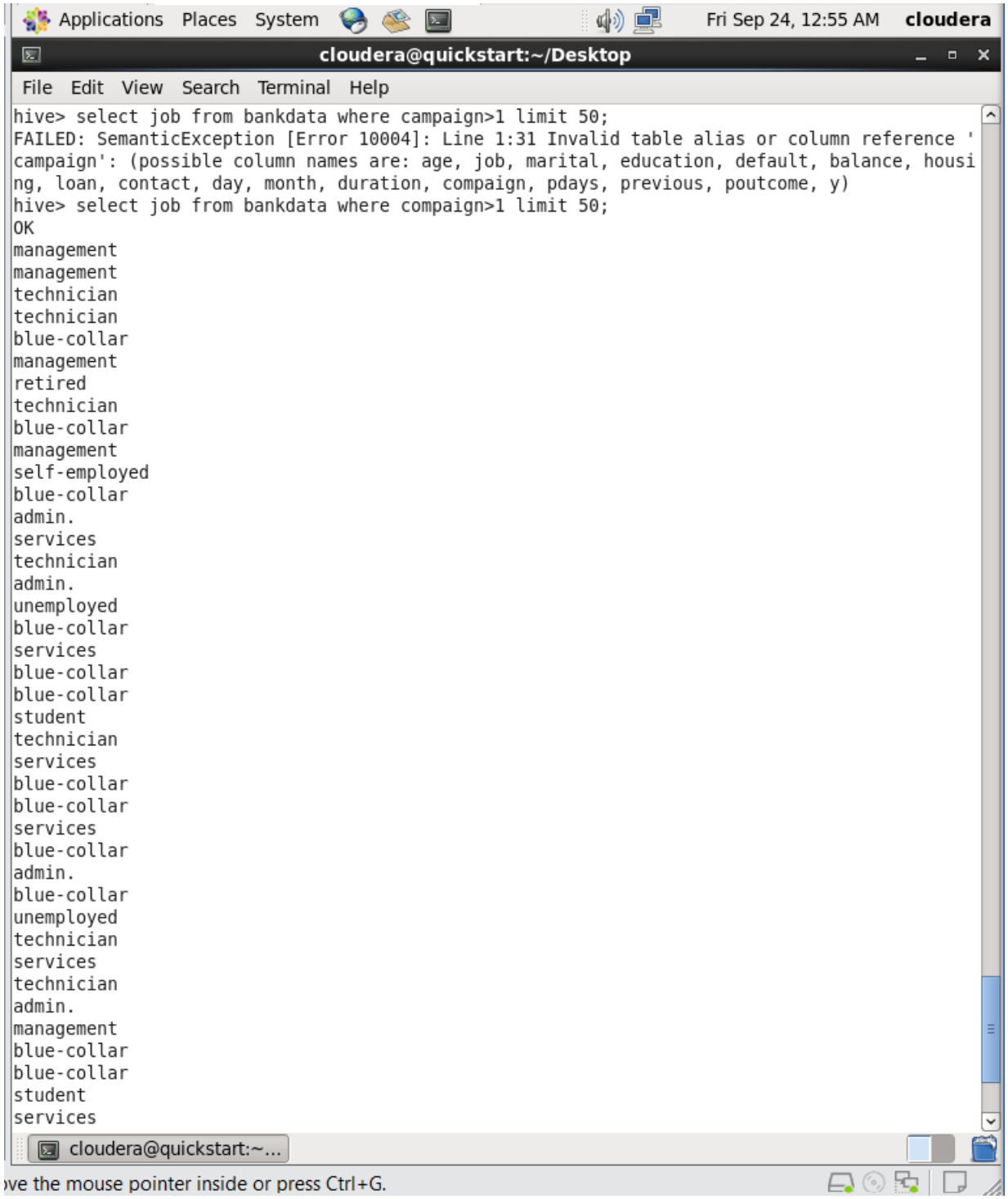
File Edit View Search Terminal Help

```
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
Time taken: 76.247 seconds, Fetched: 50 row(s)
hive>
```

cloudera@quickstart:~...

6. Find the job status for the people who has been contacted more than once.

Command:-
Hive> select job from bank where compaign>1 limit 50;



```
hive> select job from bankdata where campaign>1 limit 50;
FAILED: SemanticException [Error 10004]: Line 1:31 Invalid table alias or column reference '
campaign': (possible column names are: age, job, marital, education, default, balance, housi
ng, loan, contact, day, month, duration, compaign, pdays, previous, poutcome, y)
hive> select job from bankdata where compaign>1 limit 50;
OK
management
management
technician
technician
blue-collar
management
retired
technician
blue-collar
management
self-employed
blue-collar
admin.
services
technician
admin.
unemployed
blue-collar
services
blue-collar
blue-collar
student
technician
services
blue-collar
blue-collar
services
blue-collar
admin.
blue-collar
unemployed
technician
services
technician
admin.
management
blue-collar
blue-collar
student
services
```

cloudera@quickstart:~/Desktop

File Edit View Search Terminal Help

```
retired
technician
blue-collar
management
self-employed
blue-collar
admin.
services
technician
admin.
unemployed
blue-collar
services
blue-collar
blue-collar
student
technician
services
blue-collar
blue-collar
services
blue-collar
admin.
blue-collar
unemployed
technician
services
technician
admin.
management
blue-collar
blue-collar
student
services
retired
housemaid
unknown
unemployed
management
blue-collar
management
blue-collar
services
entrepreneur
Time taken: 0.134 seconds, Fetched: 50 row(s)
hive>
```

cloudera@quickstart:~...

7. Find the number of single people contacted and sort by age.

Command :-

Hive> select age,count(*) from bankdata where marital='single' group by age LIMIT 10;

```
cloudera@quickstart:~/Desktop
File  Edit  View  Search  Terminal  Help
hive> select age,count(*) from bankdata where marital= 'single' group by age LIMIT 10;
Query ID = cloudera_20210924005656_9ca9ef24-2e2c-4bef-874a-8563f8e2a101
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_1632468285119_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1632468285119_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632468285119_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-24 00:56:51,166 Stage-1 map = 0%,  reduce = 0%
2021-09-24 00:57:10,983 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.0 sec
2021-09-24 00:57:29,768 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.61 sec
MapReduce Total cumulative CPU time: 8 seconds 610 msec
Ended Job = job_1632468285119_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.61 sec   HDFS Read: 3760682 HDFS Write:
 66 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 610 msec
OK
18      12
19      35
20      47
21      74
22      120
23      175
24      248
25      423
26      615
27      658
Time taken: 64.15 seconds, Fetched: 10 row(s)
```

8. Calculate the average balance in the month of June and July.

Command:-

Hive> select AVG(balance) from bankdata where month='jun';

```
hive> select AVG(balance) from bankdata where month='jun';
Query ID = cloudera_20210924005858_ef441928-ad35-4d38-b103-e12c3d2d91d7
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_1632468285119_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1632468285119_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632468285119_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-24 00:59:12,084 Stage-1 map = 0%,  reduce = 0%
2021-09-24 00:59:33,273 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.05 sec
2021-09-24 00:59:54,056 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.43 sec
MapReduce Total cumulative CPU time: 8 seconds 430 msec
Ended Job = job_1632468285119_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.43 sec   HDFS Read: 3760460 HDFS Write:
 19 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 430 msec
OK
1608.2222430256506
Time taken: 68.451 seconds, Fetched: 1 row(s)
```

Hive> select AVG(balance) from bankdata where month='jul';

```
hive> select AVG(balance) from bankdata where month='jul';
Query ID = cloudera_20210924010000_35576157-515c-439d-b69c-d7479eea6e8a
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_1632468285119_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1632468285119_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1632468285119_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-09-24 01:00:29,619 Stage-1 map = 0%,  reduce = 0%
2021-09-24 01:00:49,374 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.16 sec
2021-09-24 01:01:08,791 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.52 sec
MapReduce Total cumulative CPU time: 8 seconds 520 msec
Ended Job = job_1632468285119_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.52 sec   HDFS Read: 3760446 HDFS Write:
 18 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 520 msec
OK
900.0255257432923
Time taken: 65.824 seconds, Fetched: 1 row(s)
hive>
cloudera@quickstart:~...
```