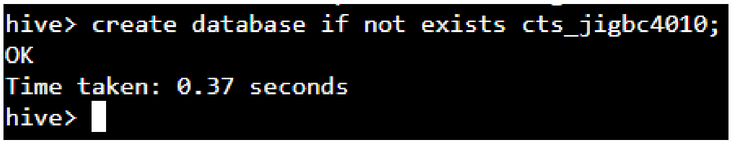
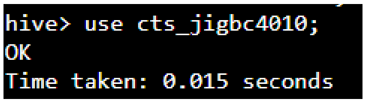
EPBA4 – Graded Assignment06 (HIVE)

NOTE: Due to memory issues in Jigsaw Lab, I have used "LIMIT 10"parameter in some of the queries.

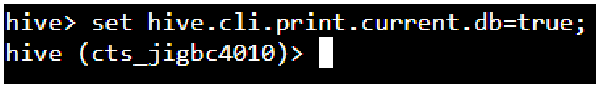
Create a database cts_<id> in your respective HDFS home location. Use this database for creating all the tables.



```
hive> create database if not exists cts_jigbc4010;
OK
Time taken: 0.37 seconds
hive>
```

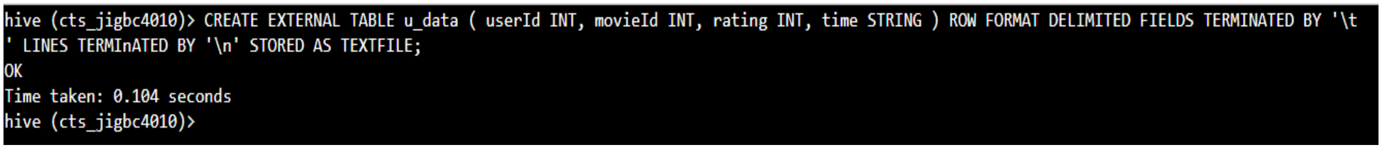


```
hive> use cts_jigbc4010;
OK
Time taken: 0.015 seconds
```



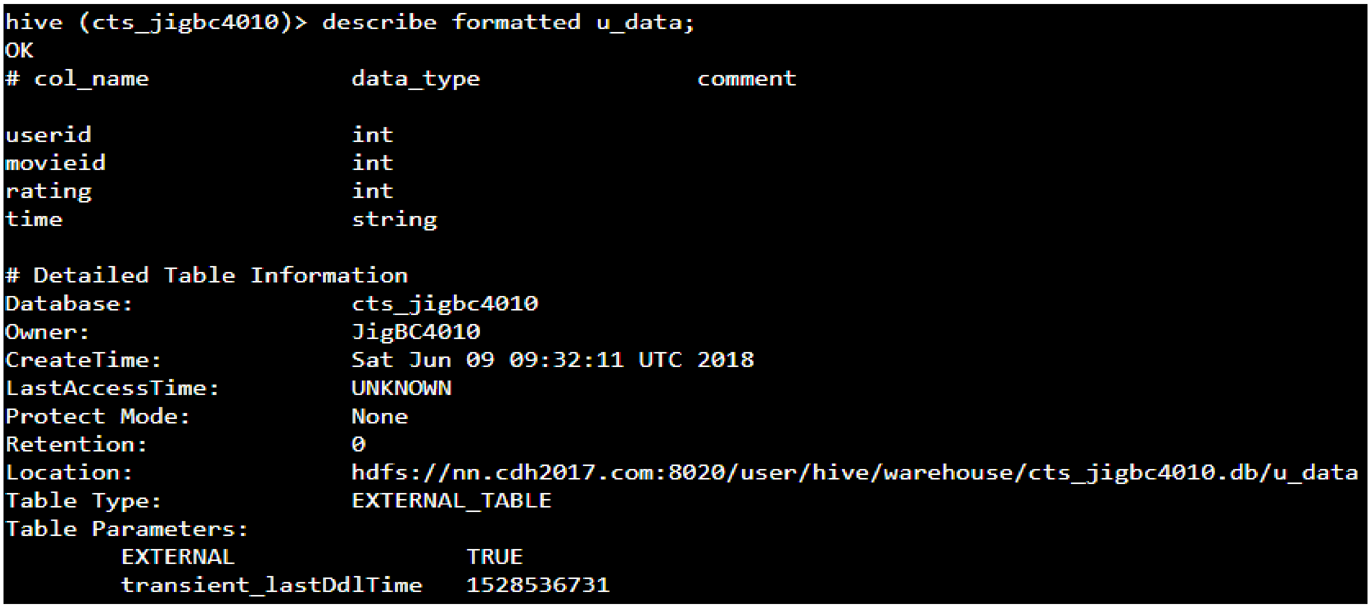
```
hive> set hive.cli.print.current.db=true;
hive (cts_jigbc4010)>
```

1. Create an external table u_data for u.data in HDFS.

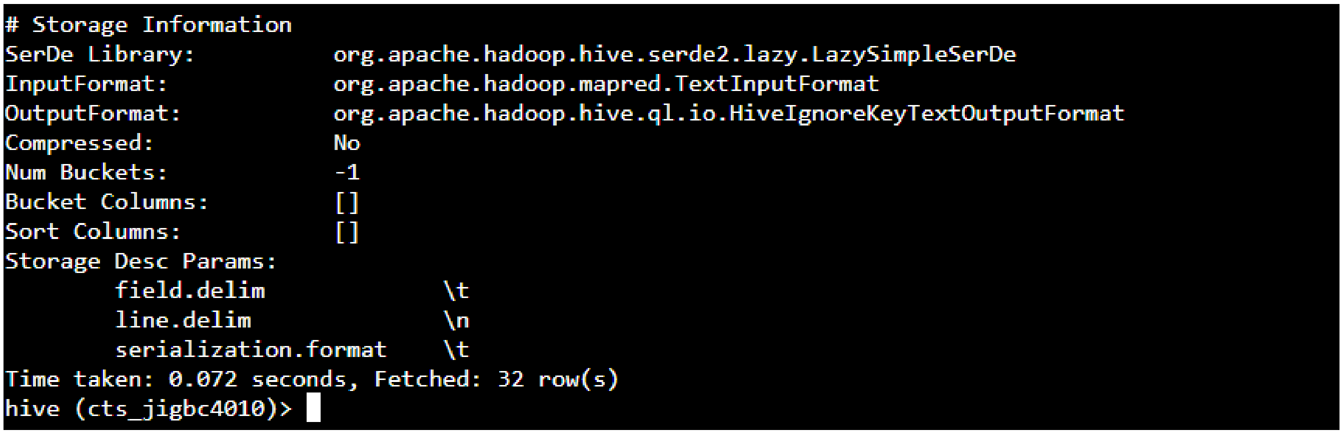


```
hive (cts_jigbc4010)> CREATE EXTERNAL TABLE u_data ( userId INT, movieId INT, rating INT, time STRING ) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t
' LINES TERMInATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.104 seconds
hive (cts_jigbc4010)>
```

2. See the field descriptions of u_data table

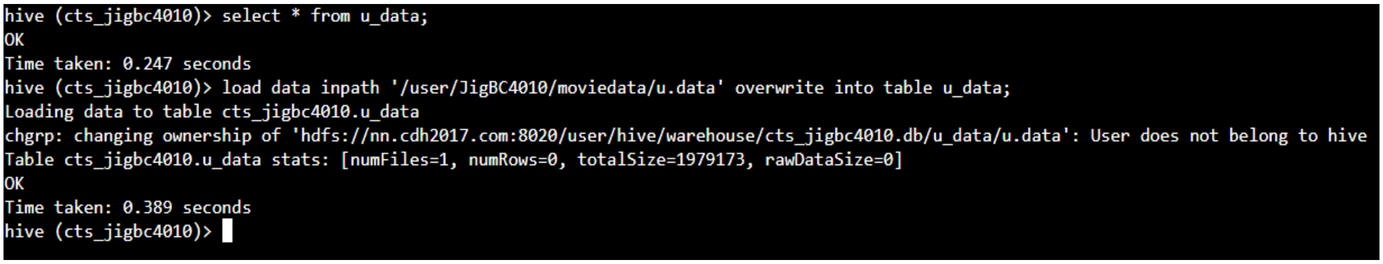


```
hive (cts_jigbc4010)> describe formatted u_data;
OK
# col_name              data_type               comment

userid                  int
movieid                 int
rating                  int
time                    string

# Detailed Table Information
Database:               cts_jigbc4010
Owner:                  JigBC4010
CreateTime:             Sat Jun 09 09:32:11 UTC 2018
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://nn.cdh2017.com:8020/user/hive/warehouse/cts_jigbc4010.db/u_data
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        transient_lastDdlTime   1528536731
```



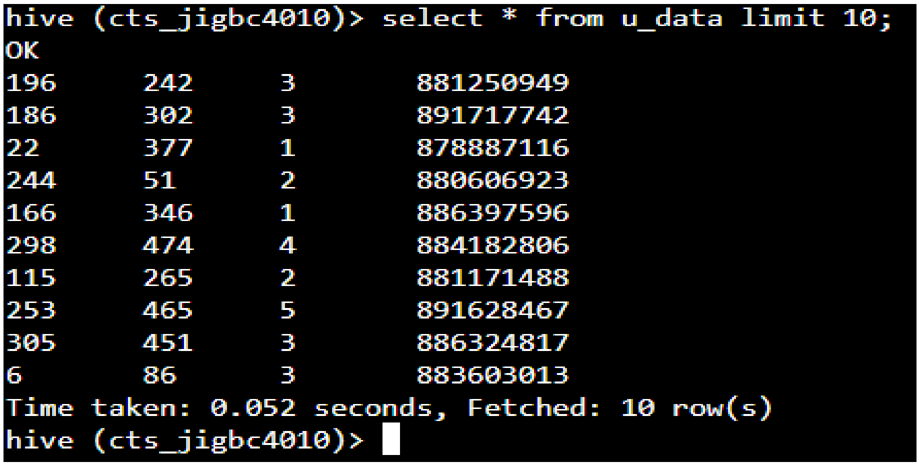
```
# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             \t
        line.delim              \n
        serialization.format    \t
Time taken: 0.072 seconds, Fetched: 32 row(s)
hive (cts_jigbc4010)>
```

3. Load the data into the table.



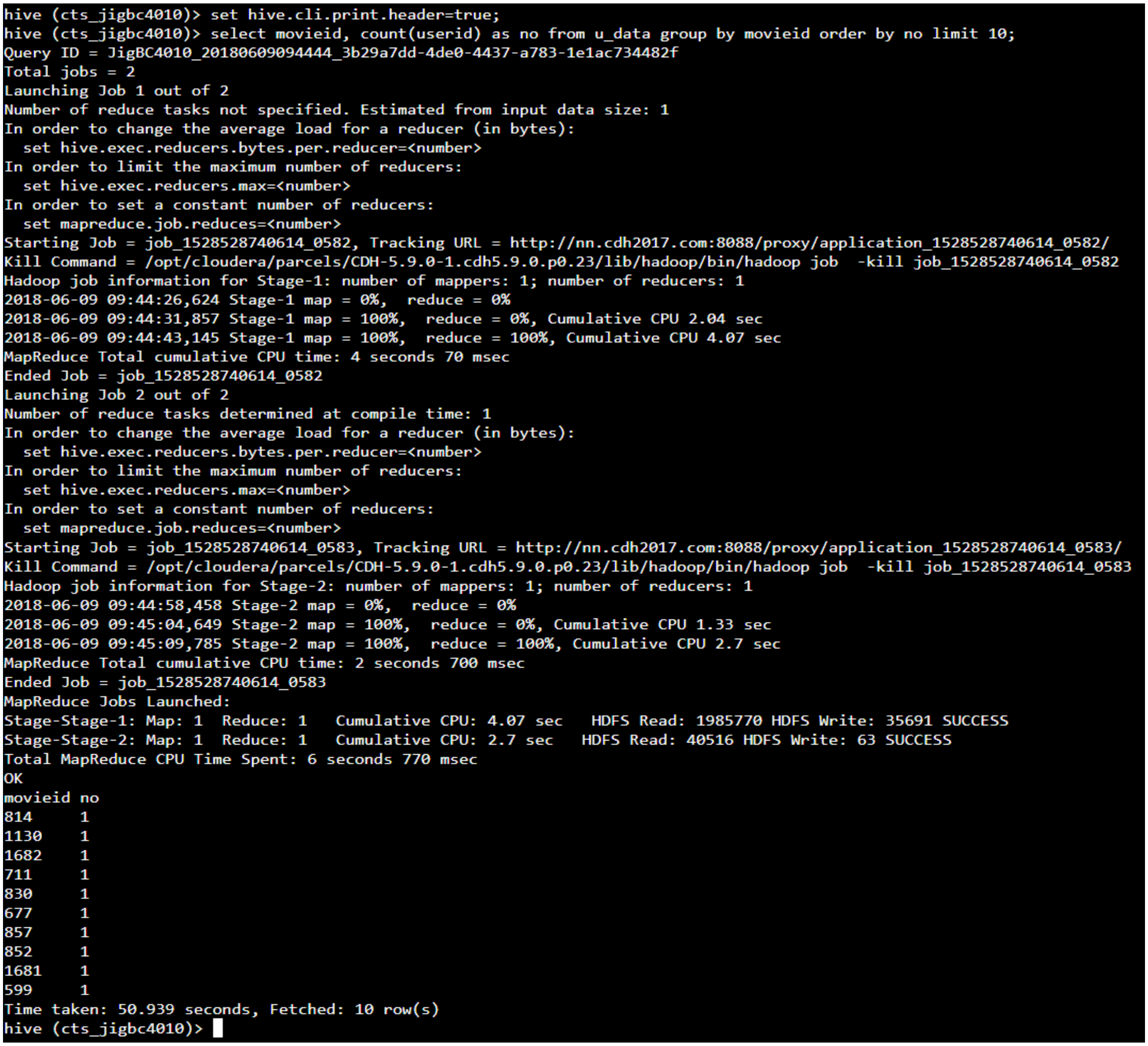
```
hive (cts_jigbc4010)> select * from u_data;
OK
Time taken: 0.247 seconds
hive (cts_jigbc4010)> load data inpath '/user/JigBC4010/moviedata/u.data' overwrite into table u_data;
Loading data to table cts_jigbc4010.u_data
chgrp: changing ownership of 'hdfs://nn.cdh2017.com:8020/user/hive/warehouse/cts_jigbc4010.db/u_data/u.data': User does not belong to hive
Table cts_jigbc4010.u_data stats: [numFiles=1, numRows=0, totalSize=1979173, rawDataSize=0]
OK
Time taken: 0.389 seconds
hive (cts_jigbc4010)>
```

4. Show all the data in the newly created u_data table



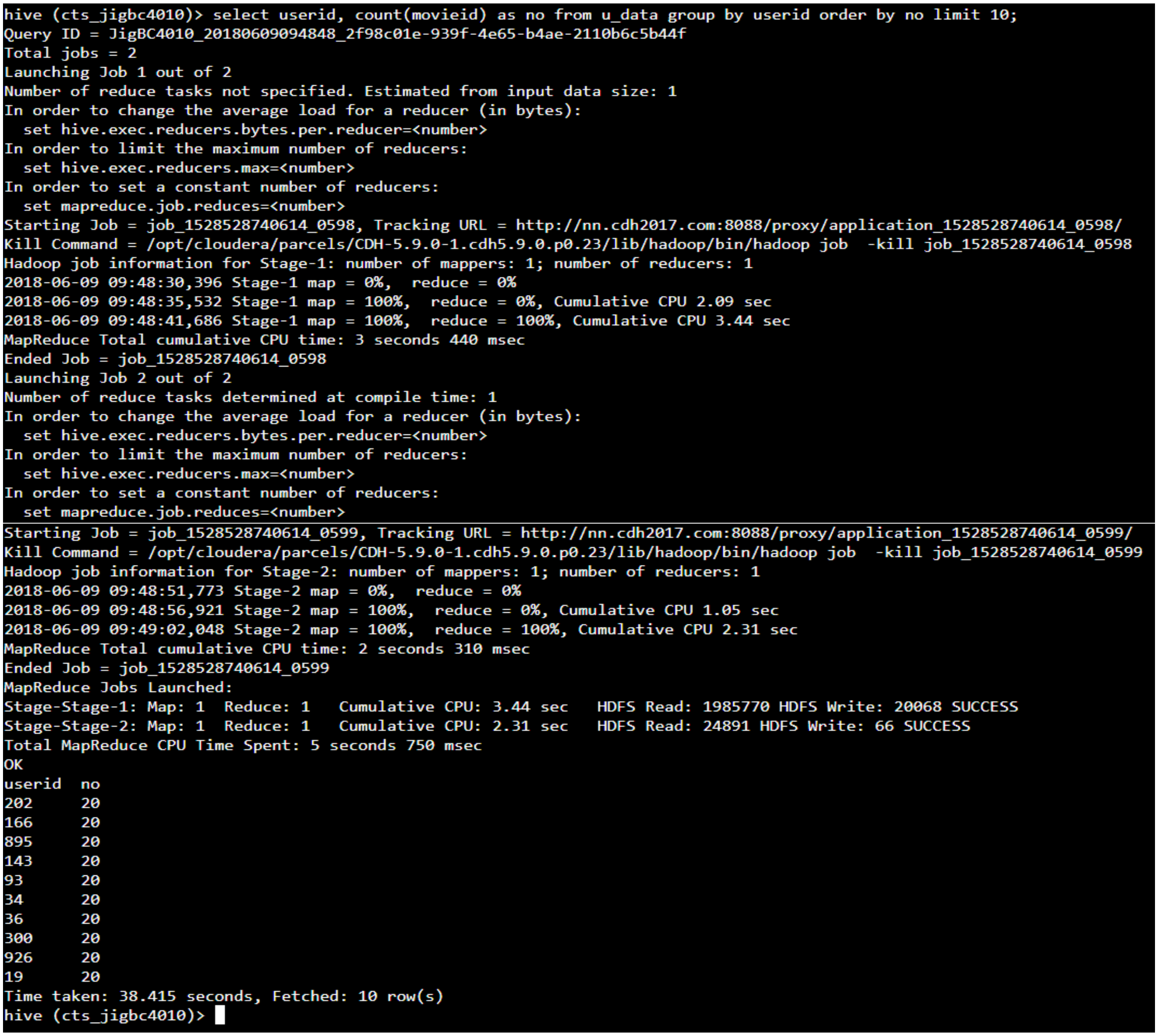
```
hive (cts_jigbc4010)> select * from u_data limit 10;
OK
196     242     3       881250949
186     302     3       891717742
22      377     1       878887116
244     51      2       880606923
166     346     1       886397596
298     474     4       884182806
115     265     2       881171488
253     465     5       891628467
305     451     3       886324817
6       86      3       883603013
Time taken: 0.052 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

5. Show the numbers of item reviewed by each user in the newly created u_data table



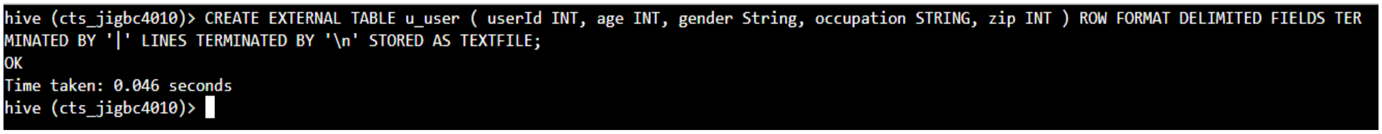
```
hive (cts_jigbc4010)> set hive.cli.print.header=true;
hive (cts_jigbc4010)> select movieid, count(userid) as no from u_data group by movieid order by no limit 10;
Query ID = JigBC4010_20180609094444_3b29a7dd-4de0-4437-a783-1e1ac734482f
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks 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_1528528740614_0582, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0582/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0582
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 09:44:26,624 Stage-1 map = 0%,  reduce = 0%
2018-06-09 09:44:31,857 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.04 sec
2018-06-09 09:44:43,145 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.07 sec
MapReduce Total cumulative CPU time: 4 seconds 70 msec
Ended Job = job_1528528740614_0582
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_1528528740614_0583, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0583/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0583
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-09 09:44:58,458 Stage-2 map = 0%,  reduce = 0%
2018-06-09 09:45:04,649 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.33 sec
2018-06-09 09:45:09,785 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.7 sec
MapReduce Total cumulative CPU time: 2 seconds 700 msec
Ended Job = job_1528528740614_0583
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.07 sec   HDFS Read: 1985770 HDFS Write: 35691 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.7 sec   HDFS Read: 40516 HDFS Write: 63 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 770 msec
OK
movieid no
814     1
1130    1
1682    1
711     1
830     1
677     1
857     1
852     1
1681    1
599     1
Time taken: 50.939 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

6. Show the numbers of users reviewed each item in the newly created u_data table



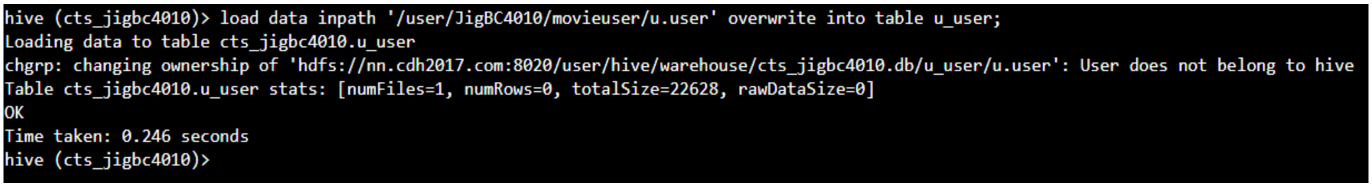
```
hive (cts_jigbc4010)> select userid, count(movieid) as no from u_data group by userid order by no limit 10;
Query ID = JigBC4010_20180609094848_2f98c01e-939f-4e65-b4ae-2110b6c5b44f
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order 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_1528528740614_0598, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0598/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0598
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 09:48:30,396 Stage-1 map = 0%,  reduce = 0%
2018-06-09 09:48:35,532 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.09 sec
2018-06-09 09:48:41,686 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.44 sec
MapReduce Total cumulative CPU time: 3 seconds 440 msec
Ended Job = job_1528528740614_0598
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_1528528740614_0599, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0599/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0599
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-06-09 09:48:51,773 Stage-2 map = 0%,  reduce = 0%
2018-06-09 09:48:56,921 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.05 sec
2018-06-09 09:49:02,048 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.31 sec
MapReduce Total cumulative CPU time: 2 seconds 310 msec
Ended Job = job_1528528740614_0599
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.44 sec   HDFS Read: 1985770 HDFS Write: 20068 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.31 sec   HDFS Read: 24891 HDFS Write: 66 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 750 msec
OK
userid  no
202     20
166     20
895     20
143     20
93      20
34      20
36      20
300     20
926     20
19      20
Time taken: 38.415 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

7. Create an external table u_user for u.user in HDFS .



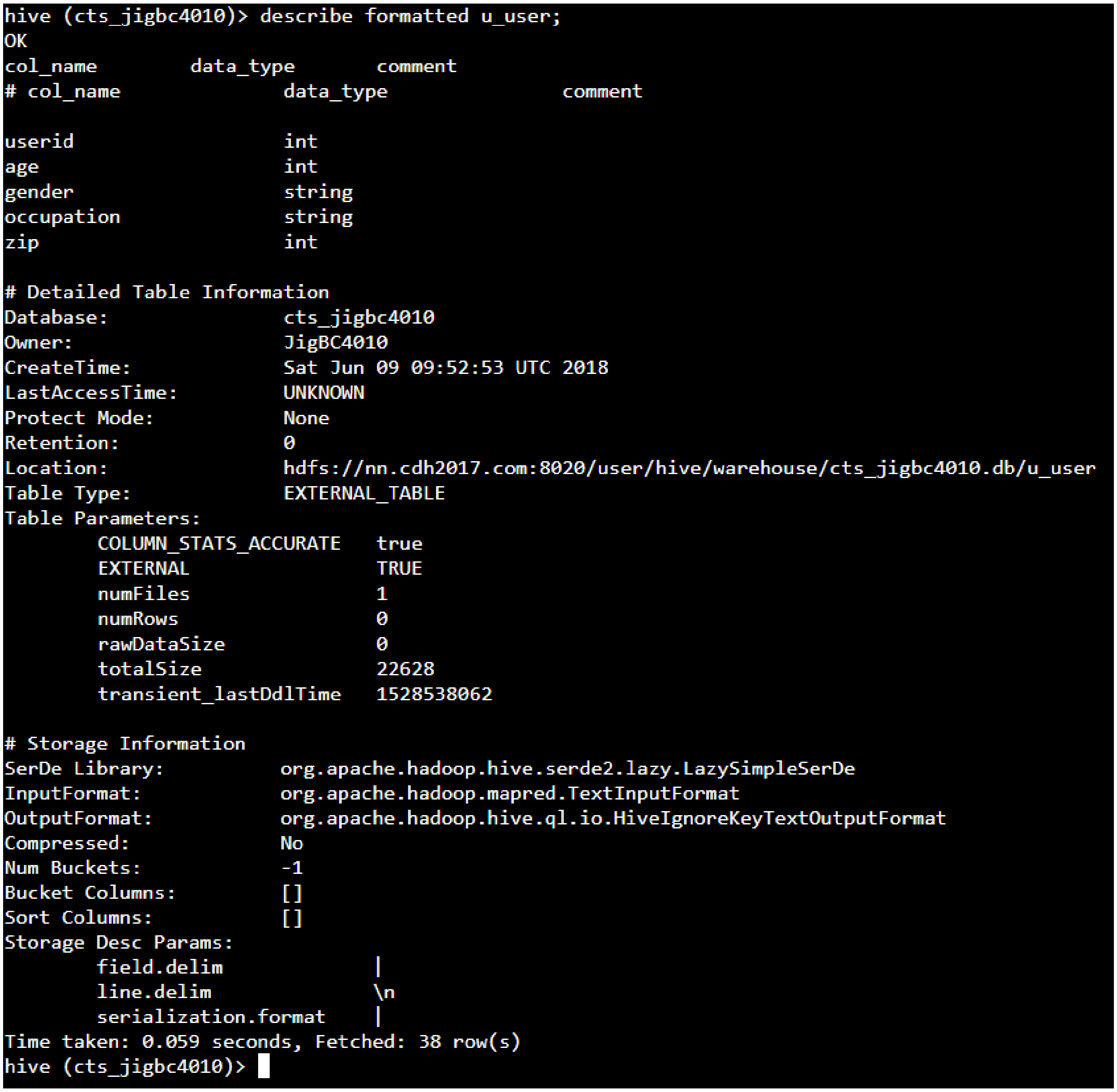
```
hive (cts_jigbc4010)> CREATE EXTERNAL TABLE u_user ( userId INT, age INT, gender String, occupation STRING, zip INT ) ROW FORMAT DELIMITED FIELDS TER
MINATED BY '|' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.046 seconds
hive (cts_jigbc4010)>
```

8. Load the data into the table.



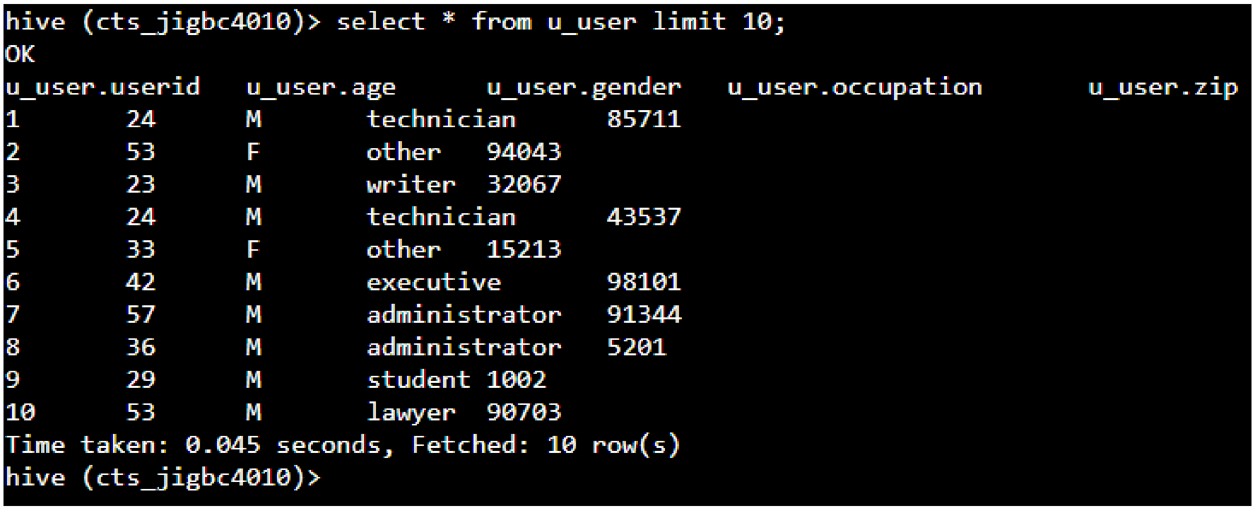
```
hive (cts_jigbc4010)> load data inpath '/user/JigBC4010/movieuser/u.user' overwrite into table u_user;
Loading data to table cts_jigbc4010.u_user
chgrp: changing ownership of 'hdfs://nn.cdh2017.com:8020/user/hive/warehouse/cts_jigbc4010.db/u_user/u.user': User does not belong to hive
Table cts_jigbc4010.u_user stats: [numFiles=1, numRows=0, totalSize=22628, rawDataSize=0]
OK
Time taken: 0.246 seconds
hive (cts_jigbc4010)>
```

9. See the field descriptions of u_user table



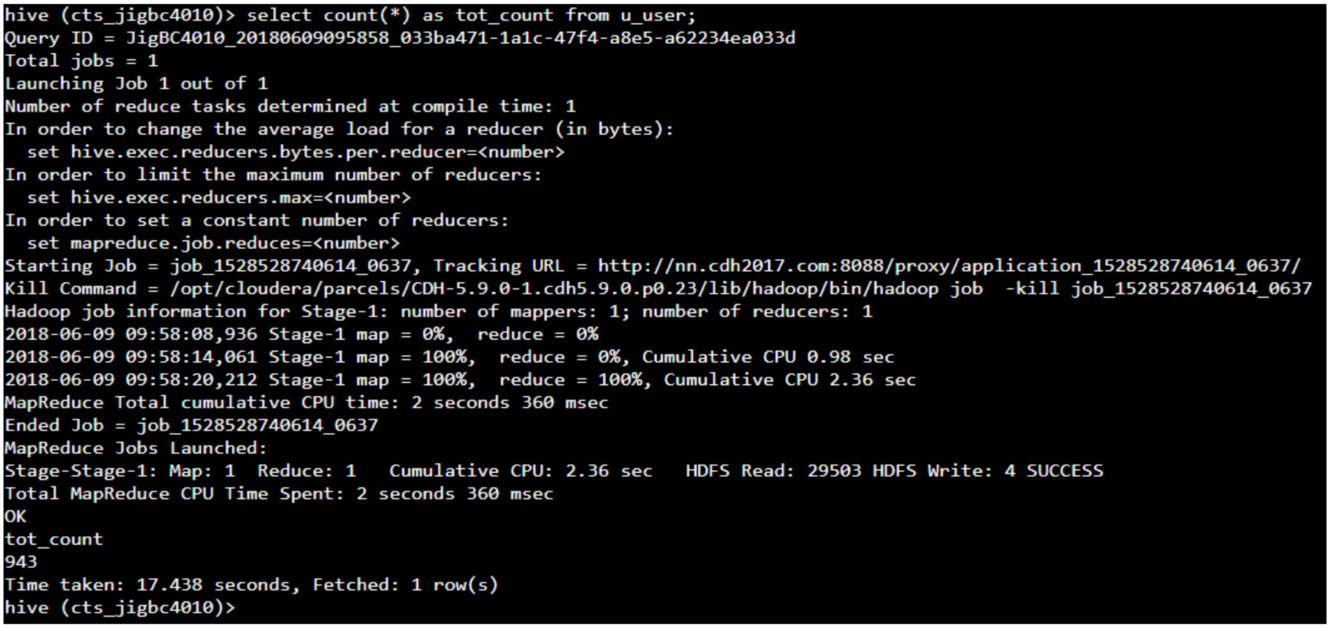
```
hive (cts_jigbc4010)> describe formatted u_user;
OK
col_name        data_type       comment
# col_name              data_type               comment

userid                  int
age                     int
gender                  string
occupation              string
zip                     int

# Detailed Table Information
Database:               cts_jigbc4010
Owner:                  JigBC4010
CreateTime:             Sat Jun 09 09:52:53 UTC 2018
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               hdfs://nn.cdh2017.com:8020/user/hive/warehouse/cts_jigbc4010.db/u_user
Table Type:             EXTERNAL_TABLE
Table Parameters:
        COLUMN_STATS_ACCURATE   true
        EXTERNAL                TRUE
        numFiles                1
        numRows                 0
        rawDataSize             0
        totalSize               22628
        transient_lastDdlTime   1528538062

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             |
        line.delim              \n
        serialization.format    |
Time taken: 0.059 seconds, Fetched: 38 row(s)
hive (cts_jigbc4010)>
```

10. Show all the data in the newly created user table



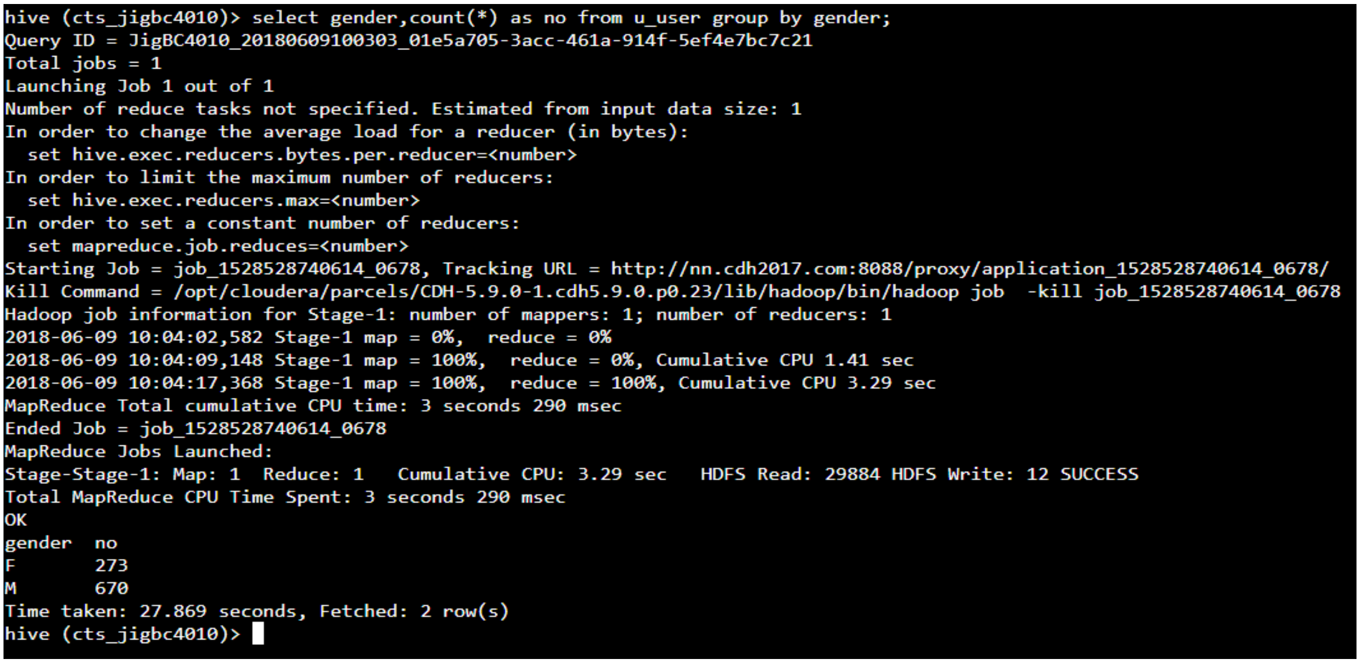
```
hive (cts_jigbc4010)> select * from u_user limit 10;
OK
u_user.userid   u_user.age      u_user.gender   u_user.occupation       u_user.zip
1       24      M       technician      85711
2       53      F       other   94043
3       23      M       writer  32067
4       24      M       technician      43537
5       33      F       other   15213
6       42      M       executive       98101
7       57      M       administrator   91344
8       36      M       administrator   5201
9       29      M       student 1002
10      53      M       lawyer  90703
Time taken: 0.045 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

11. Count the number of data in the u_user table



```
hive (cts_jigbc4010)> select count(*) as tot_count from u_user;
Query ID = JigBC4010_20180609095858_033ba471-1a1c-47f4-a8e5-a62234ea033d
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_1528528740614_0637, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0637/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0637
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 09:58:08,936 Stage-1 map = 0%,  reduce = 0%
2018-06-09 09:58:14,061 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.98 sec
2018-06-09 09:58:20,212 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.36 sec
MapReduce Total cumulative CPU time: 2 seconds 360 msec
Ended Job = job_1528528740614_0637
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.36 sec   HDFS Read: 29503 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 360 msec
OK
tot_count
943
Time taken: 17.438 seconds, Fetched: 1 row(s)
hive (cts_jigbc4010)>
```

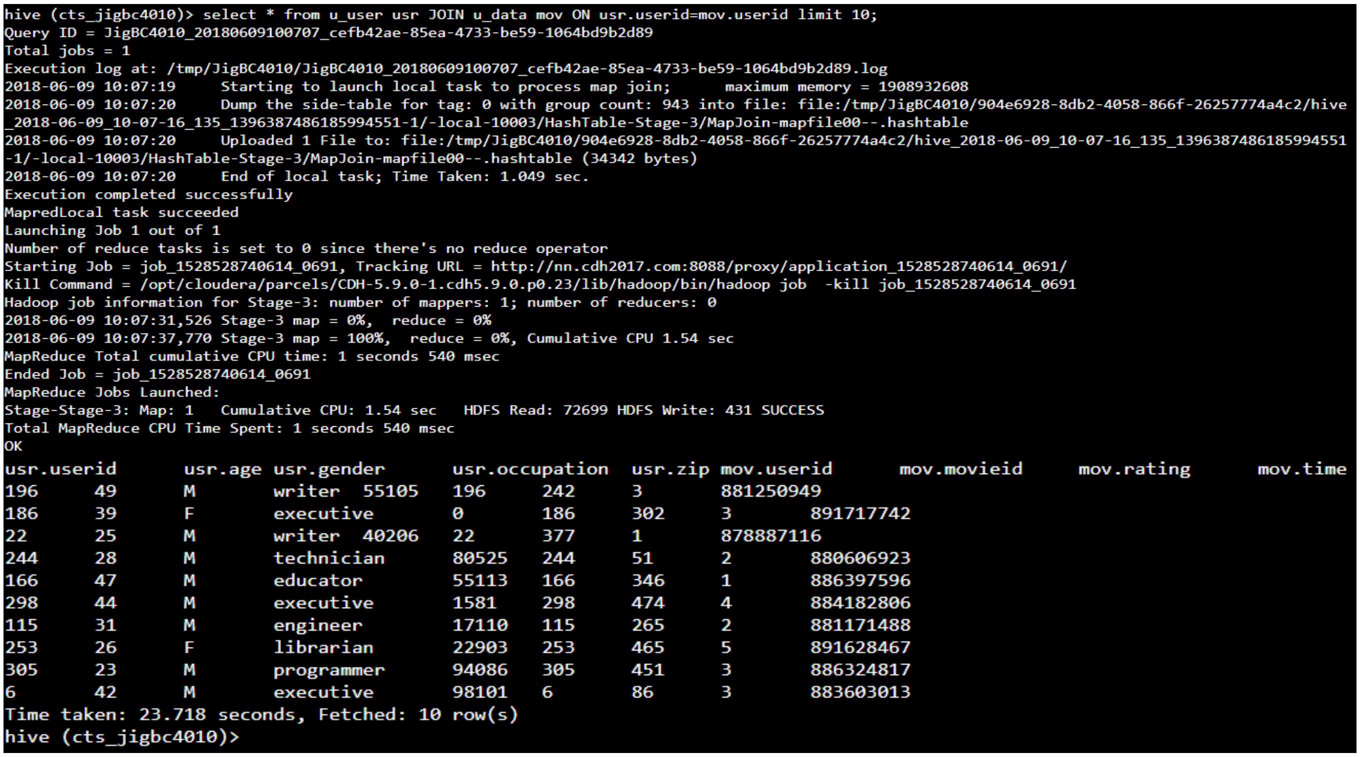
12. Count the number of user in the u_user table genderwise



```
hive (cts_jigbc4010)> select gender,count(*) as no from u_user group by gender;
Query ID = JigBC4010_20180609100303_01e5a705-3acc-461a-914f-5ef4e7bc7c21
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_1528528740614_0678, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0678/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0678
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 10:04:02,582 Stage-1 map = 0%,  reduce = 0%
2018-06-09 10:04:09,148 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.41 sec
2018-06-09 10:04:17,368 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.29 sec
MapReduce Total cumulative CPU time: 3 seconds 290 msec
Ended Job = job_1528528740614_0678
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.29 sec   HDFS Read: 29884 HDFS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 290 msec
OK
gender  no
F       273
M       670
Time taken: 27.869 seconds, Fetched: 2 row(s)
hive (cts_jigbc4010)>
```

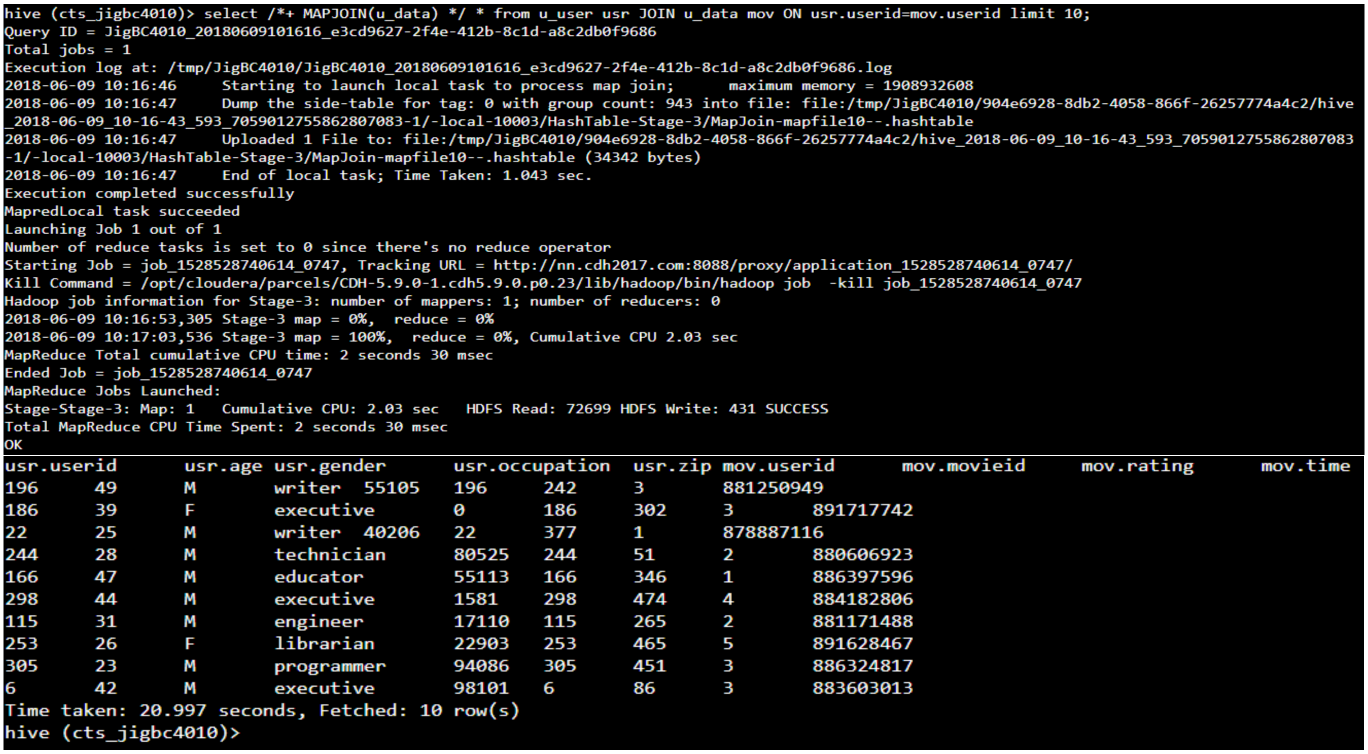
13. Join u_data table and u_user tables based on userid - Perform a reduce side join and map side join for the same and compare the time taken in both cases.

Reduce side join (Time Taken : 23.718 seconds)



```
hive (cts_jigbc4010)> select * from u_user usr JOIN u_data mov ON usr.userid=mov.userid limit 10;
Query ID = JigBC4010_20180609100707_cefb42ae-85ea-4733-be59-1064bd9b2d89
Total jobs = 1
Execution log at: /tmp/JigBC4010/JigBC4010_20180609100707_cefb42ae-85ea-4733-be59-1064bd9b2d89.log
2018-06-09 10:07:19     Starting to launch local task to process map join;      maximum memory = 1908932608
2018-06-09 10:07:20     Dump the side-table for tag: 0 with group count: 943 into file: file:/tmp/JigBC4010/904e6928-8db2-4058-866f-26257774a4c2/hive
_2018-06-09_10-07-16_135_1396387486185994551-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.hashtable
2018-06-09 10:07:20     Uploaded 1 File to: file:/tmp/JigBC4010/904e6928-8db2-4058-866f-26257774a4c2/hive_2018-06-09_10-07-16_135_1396387486185994551
-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile00--.hashtable (34342 bytes)
2018-06-09 10:07:20     End of local task; Time Taken: 1.049 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1528528740614_0691, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0691/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0691
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-06-09 10:07:31,526 Stage-3 map = 0%,  reduce = 0%
2018-06-09 10:07:37,770 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 1.54 sec
MapReduce Total cumulative CPU time: 1 seconds 540 msec
Ended Job = job_1528528740614_0691
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 1.54 sec   HDFS Read: 72699 HDFS Write: 431 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 540 msec
OK
usr.userid      usr.age usr.gender      usr.occupation  usr.zip mov.userid      mov.movieid     mov.rating      mov.time
196     49      M       writer  55105   196     242     3       881250949
186     39      F       executive       0       186     302     3       891717742
22      25      M       writer  40206   22      377     1       878887116
244     28      M       technician      80525   244     51      2       880606923
166     47      M       educator        55113   166     346     1       886397596
298     44      M       executive       1581    298     474     4       884182806
115     31      M       engineer        17110   115     265     2       881171488
253     26      F       librarian       22903   253     465     5       891628467
305     23      M       programmer      94086   305     451     3       886324817
6       42      M       executive       98101   6       86      3       883603013
Time taken: 23.718 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

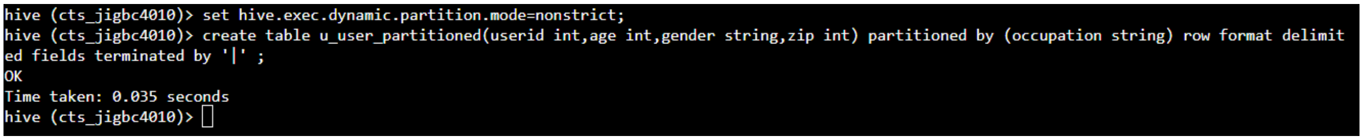
Map side Join (Time Taken : 20.997 seconds)



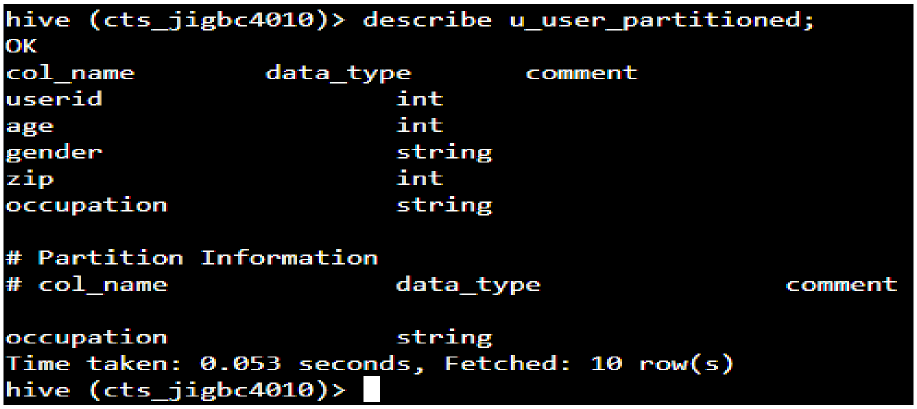
```
hive (cts_jigbc4010)> select /*+ MAPJOIN(u_data) */ * from u_user usr JOIN u_data mov ON usr.userid=mov.userid limit 10;
Query ID = JigBC4010_20180609101616_e3cd9627-2f4e-412b-8c1d-a8c2db0f9686
Total jobs = 1
Execution log at: /tmp/JigBC4010/JigBC4010_20180609101616_e3cd9627-2f4e-412b-8c1d-a8c2db0f9686.log
2018-06-09 10:16:46     Starting to launch local task to process map join;      maximum memory = 1908932608
2018-06-09 10:16:47     Dump the side-table for tag: 0 with group count: 943 into file: file:/tmp/JigBC4010/904e6928-8db2-4058-866f-26257774a4c2/hive
_2018-06-09_10-16-43_593_7059012755862807083-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable
2018-06-09 10:16:47     Uploaded 1 File to: file:/tmp/JigBC4010/904e6928-8db2-4058-866f-26257774a4c2/hive_2018-06-09_10-16-43_593_7059012755862807083
-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable (34342 bytes)
2018-06-09 10:16:47     End of local task; Time Taken: 1.043 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1528528740614_0747, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_0747/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_0747
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2018-06-09 10:16:53,305 Stage-3 map = 0%,  reduce = 0%
2018-06-09 10:17:03,536 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.03 sec
MapReduce Total cumulative CPU time: 2 seconds 30 msec
Ended Job = job_1528528740614_0747
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1   Cumulative CPU: 2.03 sec   HDFS Read: 72699 HDFS Write: 431 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 30 msec
OK
usr.userid      usr.age usr.gender      usr.occupation  usr.zip mov.userid      mov.movieid     mov.rating      mov.time
196     49      M       writer  55105   196     242     3       881250949
186     39      F       executive       0       186     302     3       891717742
22      25      M       writer  40206   22      377     1       878887116
244     28      M       technician      80525   244     51      2       880606923
166     47      M       educator        55113   166     346     1       886397596
298     44      M       executive       1581    298     474     4       884182806
115     31      M       engineer        17110   115     265     2       881171488
253     26      F       librarian       22903   253     465     5       891628467
305     23      M       programmer      94086   305     451     3       886324817
6       42      M       executive       98101   6       86      3       883603013
Time taken: 20.997 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

Map Side join took 2.721 sec less time compared to Reduce side join for the same query.

14. Create a partitioned table u_user_partitioned, partitioned by occupation column

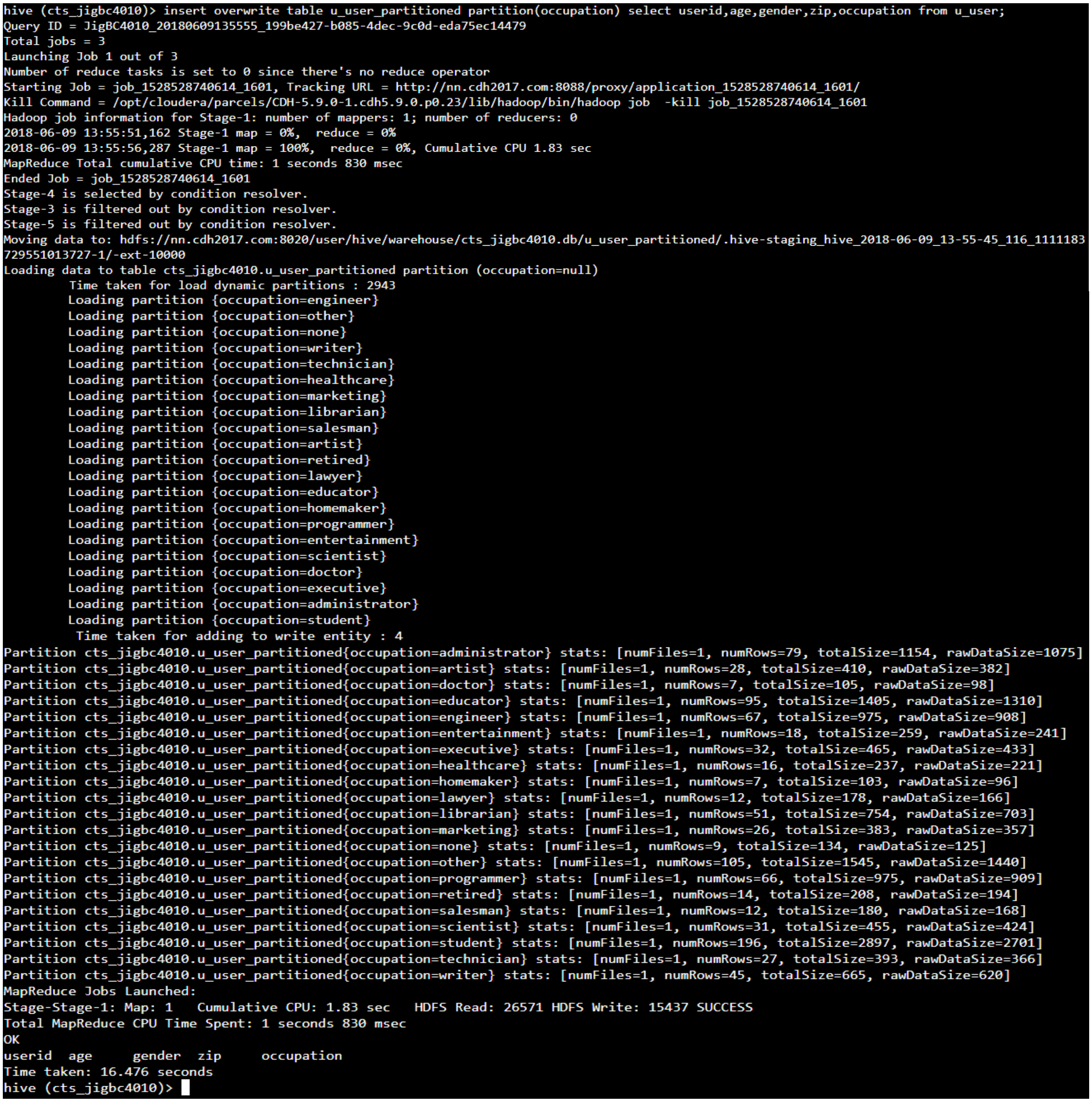


```
hive (cts_jigbc4010)> set hive.exec.dynamic.partition.mode=nonstrict;
hive (cts_jigbc4010)> create table u_user_partitioned(userid int,age int,gender string,zip int) partitioned by (occupation string) row format delimit
ed fields terminated by '|' ;
OK
Time taken: 0.035 seconds
hive (cts_jigbc4010)>
```



```
hive (cts_jigbc4010)> describe u_user_partitioned;
OK
col_name                data_type               comment
userid                  int
age                     int
gender                  string
zip                     int
occupation              string

# Partition Information
# col_name              data_type               comment

occupation              string
Time taken: 0.053 seconds, Fetched: 10 row(s)
hive (cts_jigbc4010)>
```

Loading/inserting data into partitioned table:

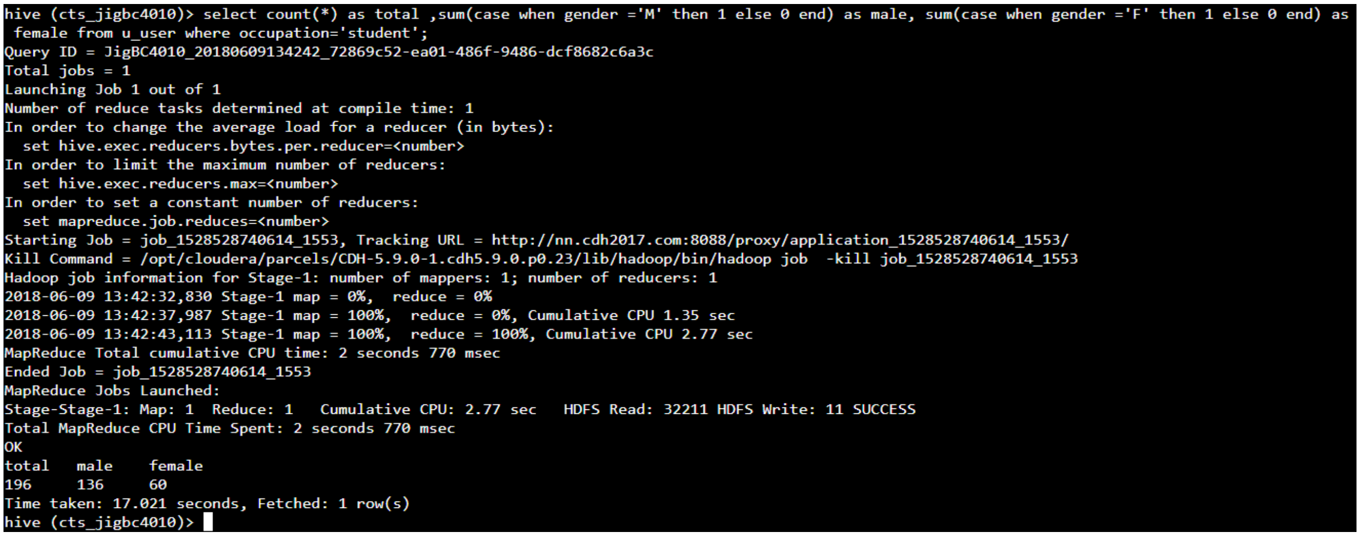


```
hive (cts_jigbc4010)> insert overwrite table u_user_partitioned partition(occupation) select userid,age,gender,zip,occupation from u_user;
Query ID = JigBC4010_20180609135555_199be427-b085-4dec-9c0d-eda75ec14479
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1528528740614_1601, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_1601/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_1601
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-06-09 13:55:51,162 Stage-1 map = 0%,  reduce = 0%
2018-06-09 13:55:56,287 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.83 sec
MapReduce Total cumulative CPU time: 1 seconds 830 msec
Ended Job = job_1528528740614_1601
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://nn.cdh2017.com:8020/user/hive/warehouse/cts_jigbc4010.db/u_user_partitioned/.hive-staging_hive_2018-06-09_13-55-45_116_1111183
729551013727-1/-ext-10000
Loading data to table cts_jigbc4010.u_user_partitioned partition (occupation=null)
         Time taken for load dynamic partitions : 2943
        Loading partition {occupation=engineer}
        Loading partition {occupation=other}
        Loading partition {occupation=none}
        Loading partition {occupation=writer}
        Loading partition {occupation=technician}
        Loading partition {occupation=healthcare}
        Loading partition {occupation=marketing}
        Loading partition {occupation=librarian}
        Loading partition {occupation=salesman}
        Loading partition {occupation=artist}
        Loading partition {occupation=retired}
        Loading partition {occupation=lawyer}
        Loading partition {occupation=educator}
        Loading partition {occupation=homemaker}
        Loading partition {occupation=programmer}
        Loading partition {occupation=entertainment}
        Loading partition {occupation=scientist}
        Loading partition {occupation=doctor}
        Loading partition {occupation=executive}
        Loading partition {occupation=administrator}
        Loading partition {occupation=student}
         Time taken for adding to write entity : 4
Partition cts_jigbc4010.u_user_partitioned{occupation=administrator} stats: [numFiles=1, numRows=79, totalSize=1154, rawDataSize=1075]
Partition cts_jigbc4010.u_user_partitioned{occupation=artist} stats: [numFiles=1, numRows=28, totalSize=410, rawDataSize=382]
Partition cts_jigbc4010.u_user_partitioned{occupation=doctor} stats: [numFiles=1, numRows=7, totalSize=105, rawDataSize=98]
Partition cts_jigbc4010.u_user_partitioned{occupation=educator} stats: [numFiles=1, numRows=95, totalSize=1405, rawDataSize=1310]
Partition cts_jigbc4010.u_user_partitioned{occupation=engineer} stats: [numFiles=1, numRows=67, totalSize=975, rawDataSize=908]
Partition cts_jigbc4010.u_user_partitioned{occupation=entertainment} stats: [numFiles=1, numRows=18, totalSize=259, rawDataSize=241]
Partition cts_jigbc4010.u_user_partitioned{occupation=executive} stats: [numFiles=1, numRows=32, totalSize=465, rawDataSize=433]
Partition cts_jigbc4010.u_user_partitioned{occupation=healthcare} stats: [numFiles=1, numRows=16, totalSize=237, rawDataSize=221]
Partition cts_jigbc4010.u_user_partitioned{occupation=homemaker} stats: [numFiles=1, numRows=7, totalSize=103, rawDataSize=96]
Partition cts_jigbc4010.u_user_partitioned{occupation=lawyer} stats: [numFiles=1, numRows=12, totalSize=178, rawDataSize=166]
Partition cts_jigbc4010.u_user_partitioned{occupation=librarian} stats: [numFiles=1, numRows=51, totalSize=754, rawDataSize=703]
Partition cts_jigbc4010.u_user_partitioned{occupation=marketing} stats: [numFiles=1, numRows=26, totalSize=383, rawDataSize=357]
Partition cts_jigbc4010.u_user_partitioned{occupation=none} stats: [numFiles=1, numRows=9, totalSize=134, rawDataSize=125]
Partition cts_jigbc4010.u_user_partitioned{occupation=other} stats: [numFiles=1, numRows=105, totalSize=1545, rawDataSize=1440]
Partition cts_jigbc4010.u_user_partitioned{occupation=programmer} stats: [numFiles=1, numRows=66, totalSize=975, rawDataSize=909]
Partition cts_jigbc4010.u_user_partitioned{occupation=retired} stats: [numFiles=1, numRows=14, totalSize=208, rawDataSize=194]
Partition cts_jigbc4010.u_user_partitioned{occupation=salesman} stats: [numFiles=1, numRows=12, totalSize=180, rawDataSize=168]
Partition cts_jigbc4010.u_user_partitioned{occupation=scientist} stats: [numFiles=1, numRows=31, totalSize=455, rawDataSize=424]
Partition cts_jigbc4010.u_user_partitioned{occupation=student} stats: [numFiles=1, numRows=196, totalSize=2897, rawDataSize=2701]
Partition cts_jigbc4010.u_user_partitioned{occupation=technician} stats: [numFiles=1, numRows=27, totalSize=393, rawDataSize=366]
Partition cts_jigbc4010.u_user_partitioned{occupation=writer} stats: [numFiles=1, numRows=45, totalSize=665, rawDataSize=620]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 1.83 sec   HDFS Read: 26571 HDFS Write: 15437 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 830 msec
OK
userid  age     gender  zip     occupation
Time taken: 16.476 seconds
hive (cts_jigbc4010)>
```

13. Find out the total number of male and total number of female only for the most common occupation – you can hard code the occupation/ use subqueries. - Perform the query on both un-partitioned table and partitioned table. - Compare and report the performance differences.

NOTE: Here I am using "student" as most common occupation.

Un-partitioned table:

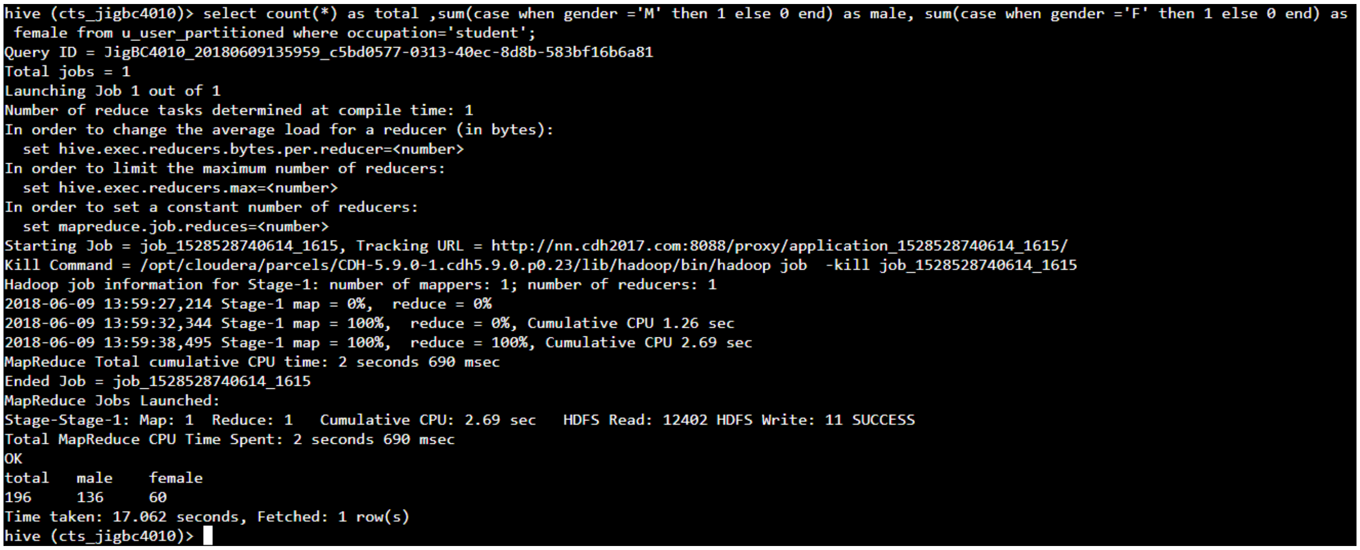


```
hive (cts_jigbc4010)> select count(*) as total ,sum(case when gender ='M' then 1 else 0 end) as male, sum(case when gender ='F' then 1 else 0 end) as
 female from u_user where occupation='student';
Query ID = JigBC4010_20180609134242_72869c52-ea01-486f-9486-dcf8682c6a3c
Total jobs = 1
Launching Job 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_1528528740614_1553, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_1553/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_1553
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 13:42:32,830 Stage-1 map = 0%,  reduce = 0%
2018-06-09 13:42:37,987 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.35 sec
2018-06-09 13:42:43,113 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.77 sec
MapReduce Total cumulative CPU time: 2 seconds 770 msec
Ended Job = job_1528528740614_1553
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.77 sec   HDFS Read: 32211 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 770 msec
OK
total   male    female
196     136     60
Time taken: 17.021 seconds, Fetched: 1 row(s)
hive (cts_jigbc4010)>
```

MapReduce CPU Time : 2 sec 770 msec.

TimeTaken : 17.021 sec.

Partitioned table:



```
hive (cts_jigbc4010)> select count(*) as total ,sum(case when gender ='M' then 1 else 0 end) as male, sum(case when gender ='F' then 1 else 0 end) as
 female from u_user_partitioned where occupation='student';
Query ID = JigBC4010_20180609135959_c5bd0577-0313-40ec-8d8b-583bf16b6a81
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_1528528740614_1615, Tracking URL = http://nn.cdh2017.com:8088/proxy/application_1528528740614_1615/
Kill Command = /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/bin/hadoop job  -kill job_1528528740614_1615
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-09 13:59:27,214 Stage-1 map = 0%,  reduce = 0%
2018-06-09 13:59:32,344 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.26 sec
2018-06-09 13:59:38,495 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.69 sec
MapReduce Total cumulative CPU time: 2 seconds 690 msec
Ended Job = job_1528528740614_1615
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.69 sec   HDFS Read: 12402 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 690 msec
OK
total   male    female
196     136     60
Time taken: 17.062 seconds, Fetched: 1 row(s)
hive (cts_jigbc4010)>
```

MapReduce CPU Time	**: 2 sec 690 msec.**
TimeTaken	**: 17.062 sec.**

- **The MapReduce CPU time for partitioned table was less by 0.08 seconds compared to that of un-partitioned table.**
- **However, the total time taken for complete execution of query over partitioned table was slightly more by 0.041 seconds than that of un-partitioned table.**