Q1 The dataset provided - MovieLens data sets are collected by the GroupLens Research Project at the University of Minnesota. It represents users' reviews of movies. This data set consists of: * 100,000 ratings (1-5) from 943 users on 1682 movies. * Each user has rated at least 20 movies. * Simple demographic info for the users (age, gender, occupation, zip) u.data -- The full u data set, 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1.

The data is randomly ordered. This is a tab separated list of user id item id rating timestamp The time stamps are Unix seconds since 1/1/1970 UTC u.user -- Demographic information about the users; this is a pipe (|) separated list of user id | age | gender | occupation | zip code The user ids are the ones used in the u.data data set. Paste the code for each step and the output of the Query Queries to be performed

1. Create an external table u_data for u.data in HDFS.
2. See the field descriptions of u_data table
3. Show all the data in the newly created u_data table
4. Show the numbers of item reviewed by each user in the newly created u_data table
5. Show the numbers of users reviewed each item in the newly created u_data table
6. Create an external table u_user for u.user in HDFS .
7. See the field descriptions of u_user table
8. Show all the data in the newly created user table
9. Count the number of data in the u_user table
10. Count the number of user in the u_user table genderwise
11. 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.
12. Create a partitioned table u_user_partitioned, partitioned by occupation column

13.Join u_data table and u_user tables based on userid

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

15. 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.

Solutions

Question 1:

```
File  Edit  Tabs  Help
hive> create database cts_Jig14696;
OK
Time taken: 0.032 seconds
hive> use cts_Jig14696;
OK
Time taken: 0.012 seconds
hive> create table u_data (userid int, movieid int, rating int, timestamp string)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.06 seconds
hive>
```

Question 2:

```
File  Edit  Tabs  Help
hduser@vin...  ×  hduser@vin...  ×
hive> describe u_data;
OK
userid                  int
movieid                 int
rating                  int
timestamp               string
Time taken: 0.067 seconds, Fetched: 4 row(s)
hive>
```

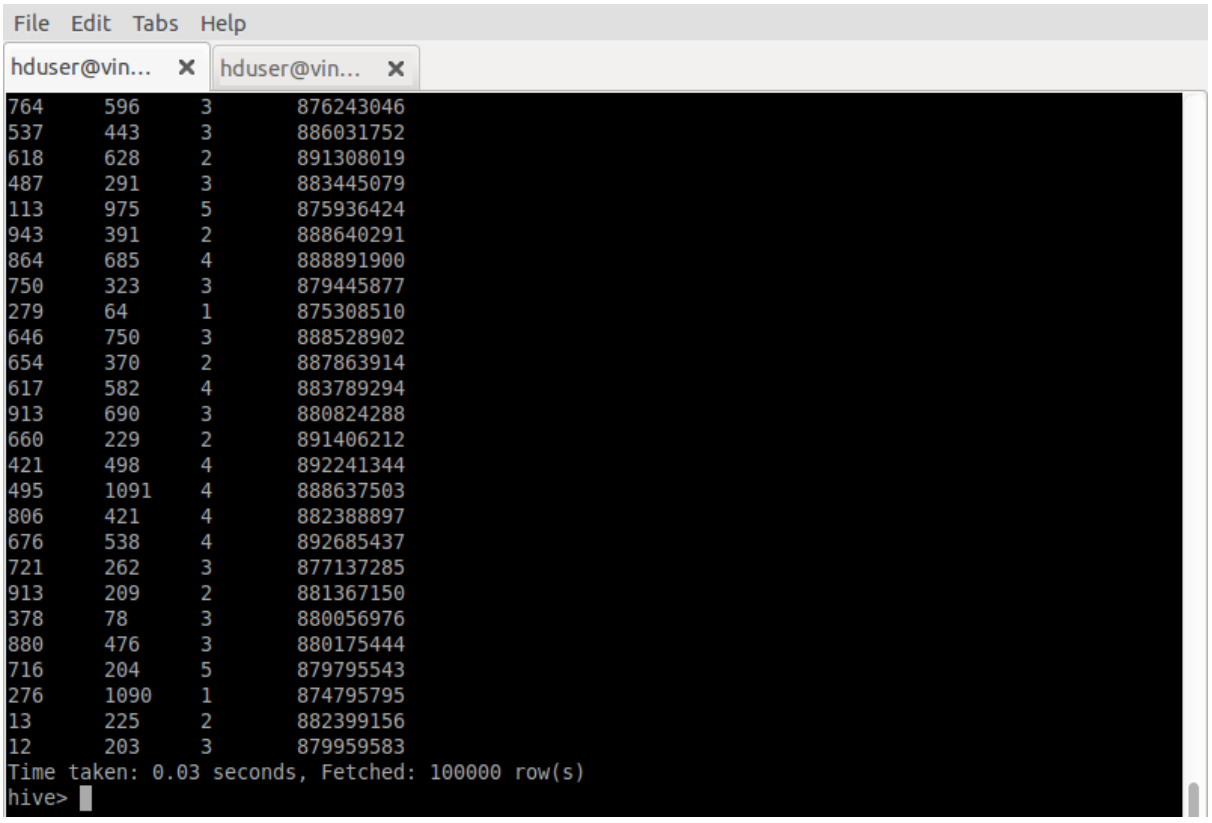
Question 3:

```
File  Edit  Tabs  Help
hduser@vin...  ×  hduser@vin...  ×
hduser@vinod-virtual-machine:~$ hadoop fs -mkdir cts
hduser@vinod-virtual-machine:~$ hadoop fs -put /home/hduser/u.data /user/hduser/cts
hduser@vinod-virtual-machine:~$ hadoop fs -put /home/hduser/u.user /user/hduser/cts
hduser@vinod-virtual-machine:~$ hadoop fs -ls /home/hduser/u.user /user/hduser/cts
ls: Cannot access /home/hduser/u.user: No such file or directory.
Found 2 items
-rw-r--r--   1 hduser supergroup    1979173 2017-11-29 15:36 /user/hduser/cts/u.data
-rw-r--r--   1 hduser supergroup      22628 2017-11-29 15:36 /user/hduser/cts/u.user
hduser@vinod-virtual-machine:~$
```

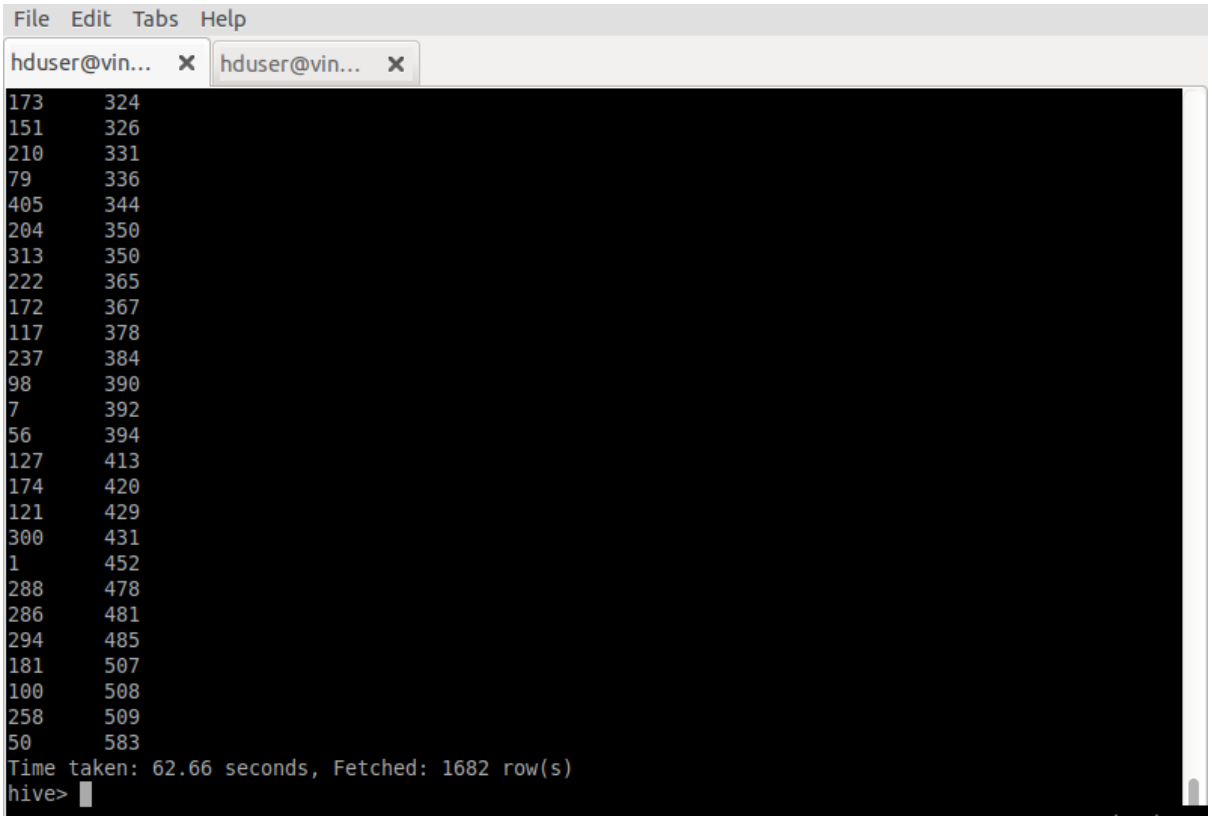
```
File  Edit  Tabs  Help
hduser@vin...  ×  hduser@vin...  ×
hive> use cts_Jig14696;
OK
Time taken: 0.013 seconds
hive> create table u_data (userid int, movieid int, rating int, timestamp string)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.145 seconds
hive> describe u_data;
OK
userid                  int
movieid                 int
rating                  int
timestamp               string
Time taken: 0.072 seconds, Fetched: 4 row(s)
hive> LOAD DATA INPATH '/user/hduser/cts/u.data' OVERWRITE INTO TABLE u_data;
```

```
File  Edit  Tabs  Help
hduser@vin...  ×  hduser@vin...  ×
hive> LOAD DATA INPATH '/user/hduser/cts/u.data' OVERWRITE INTO TABLE u_data;
Loading data to table cts_jig14696.u_data
Deleted hdfs://localhost:54310/user/hive/warehouse/cts_jig14696.db/u_data
Table cts_jig14696.u_data stats: [numFiles=1, numRows=0, totalSize=1979173, rawDataSize=0]
OK
Time taken: 0.257 seconds
hive> select * from u_data;
```

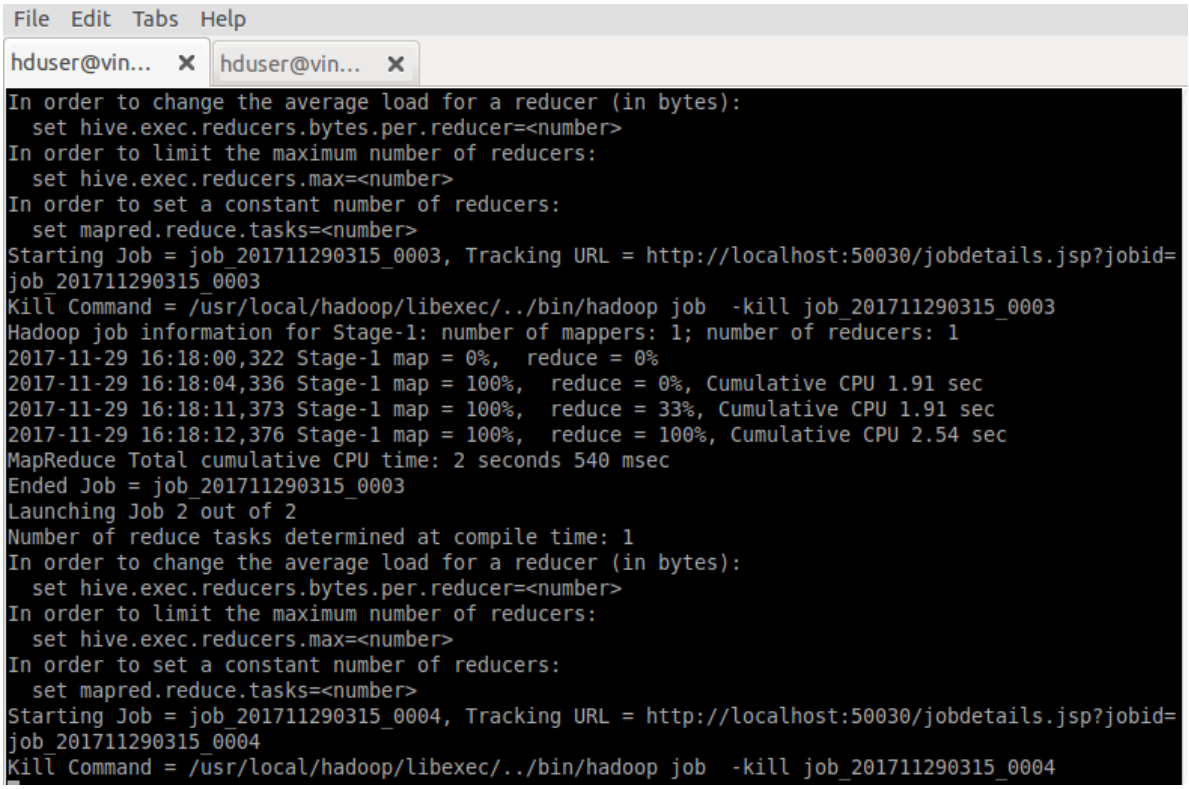
Question 4:

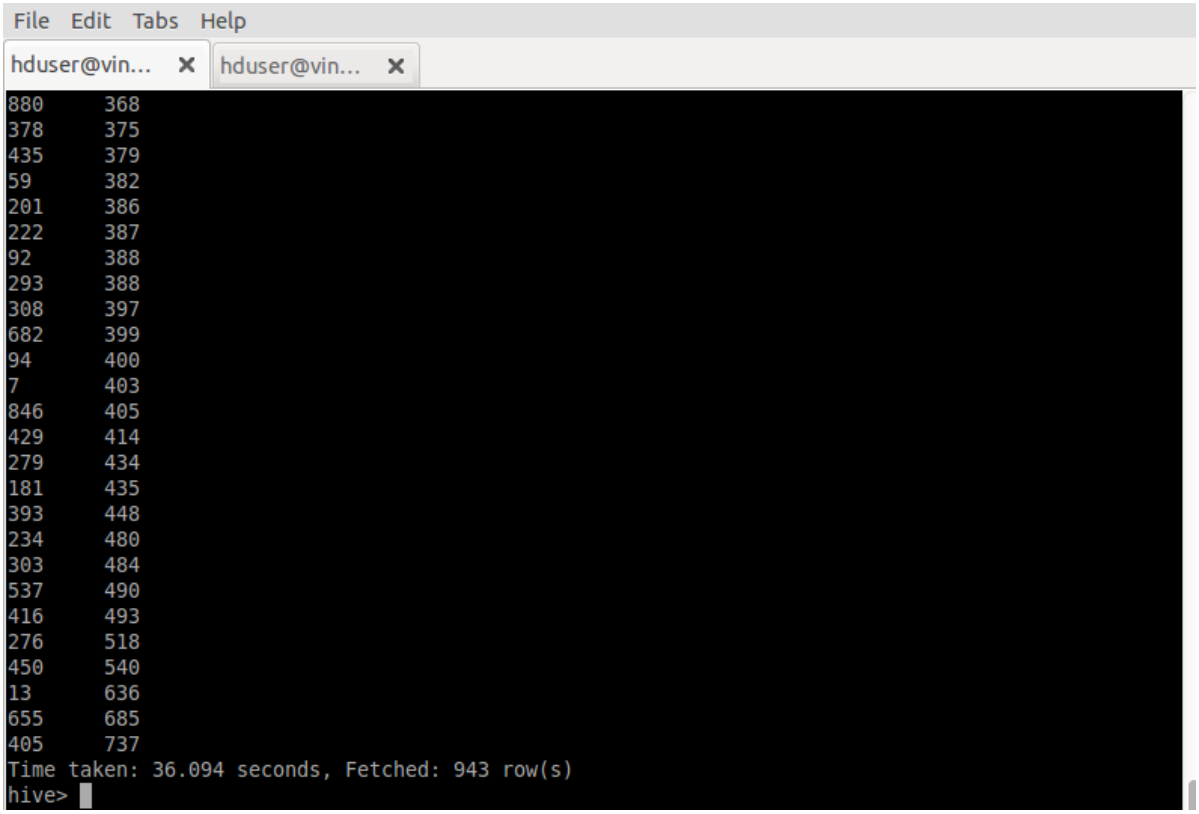


Question 5:



Question 6:


File Edit Tabs Help
hduser@vin...
hduser@vin...
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 mapred.reduce.tasks=<number>
Starting Job = job_201711290315_0003, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=
job_201711290315_0003
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job -kill job_201711290315_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-29 16:18:00,322 Stage-1 map = 0%, reduce = 0%
2017-11-29 16:18:04,336 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.91 sec
2017-11-29 16:18:11,373 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 1.91 sec
2017-11-29 16:18:12,376 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.54 sec
MapReduce Total cumulative CPU time: 2 seconds 540 msec
Ended Job = job_201711290315_0003
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 mapred.reduce.tasks=<number>
Starting Job = job_201711290315_0004, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=
job_201711290315_0004
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job -kill job_201711290315_0004


File Edit Tabs Help
hduser@vin...
hduser@vin...
880 368
378 375
435 379
59 382
201 386
222 387
92 388
293 388
308 397
682 399
94 400
7 403
846 405
429 414
279 434
181 435
393 448
234 480
303 484
537 490
416 493
276 518
450 540
13 636
655 685
405 737
Time taken: 36.094 seconds, Fetched: 943 row(s)
hive>

Question 7:

```
File  Edit  Tabs  Help
hive> create database cts_Jig14696;
OK
Time taken: 0.032 seconds
hive> use cts_Jig14696;
OK
Time taken: 0.012 seconds
```

```
File  Edit  Tabs  Help
hduser@vin...  ×   hduser@vin...  ×
hive> create table u_user (userid int, age int, gender string, occupation string, zipcode int)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '|'
    > STORED AS TEXTFILE;
OK
Time taken: 0.242 seconds
hive>
```

Question 8:

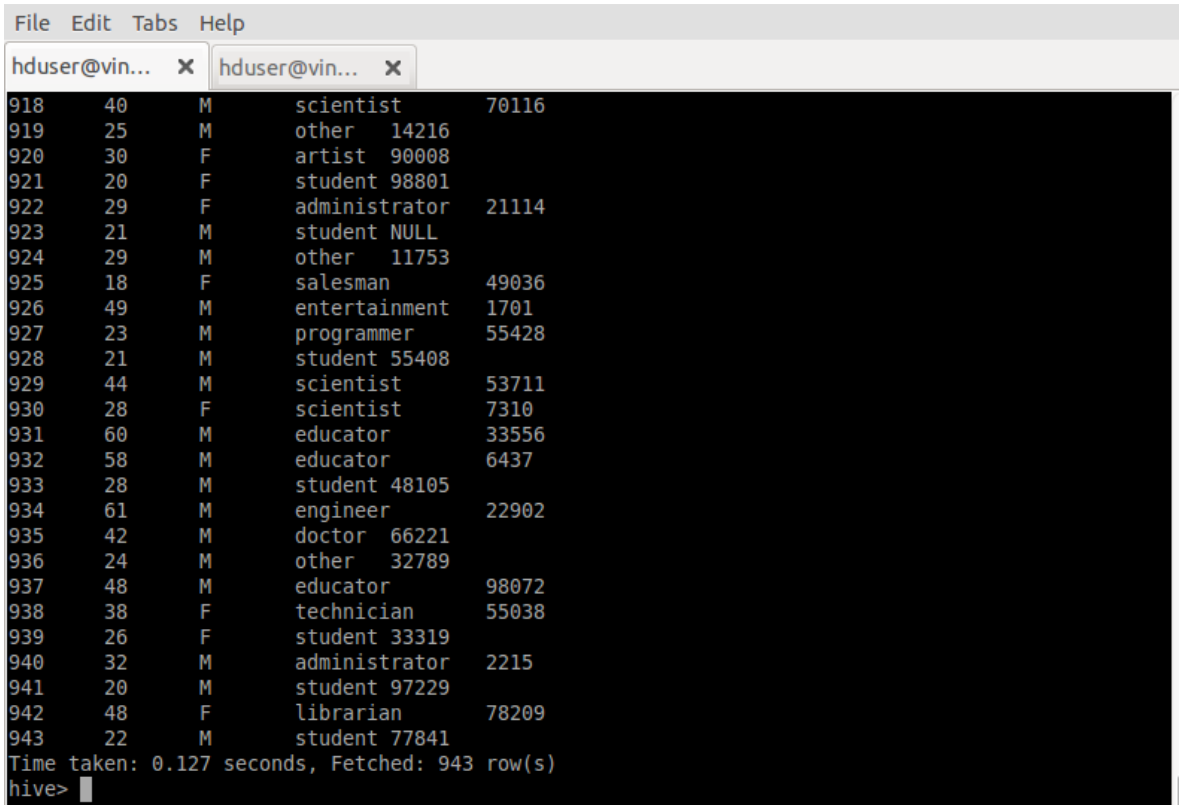
```
hive> LOAD DATA INPATH '/user/hduser/cts/u.user' OVERWRITE INTO TABLE u_user;
Loading data to table cts_jig14696.u_user
Deleted hdfs://localhost:54310/user/hive/warehouse/cts_jig14696.db/u_user
Table cts_jig14696.u_user stats: [numFiles=1, numRows=0, totalSize=22628, rawDataSize=0]
OK
Time taken: 0.302 seconds
hive>
```

Question 9:

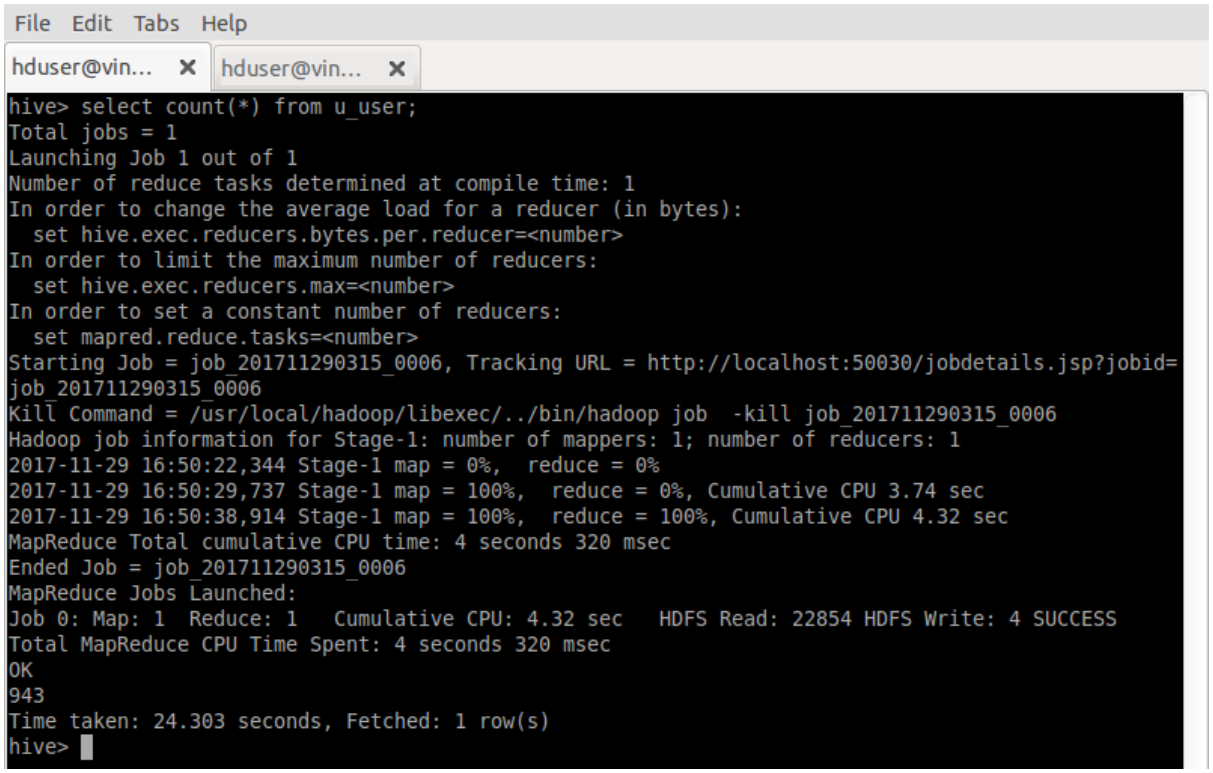
```
hive> describe u_user
    > ;
OK
userid                  int
age                     int
gender                  string
occupation              string
zipcode                 int
Time taken: 0.192 seconds, Fetched: 5 row(s)
hive>
```

Question 10:

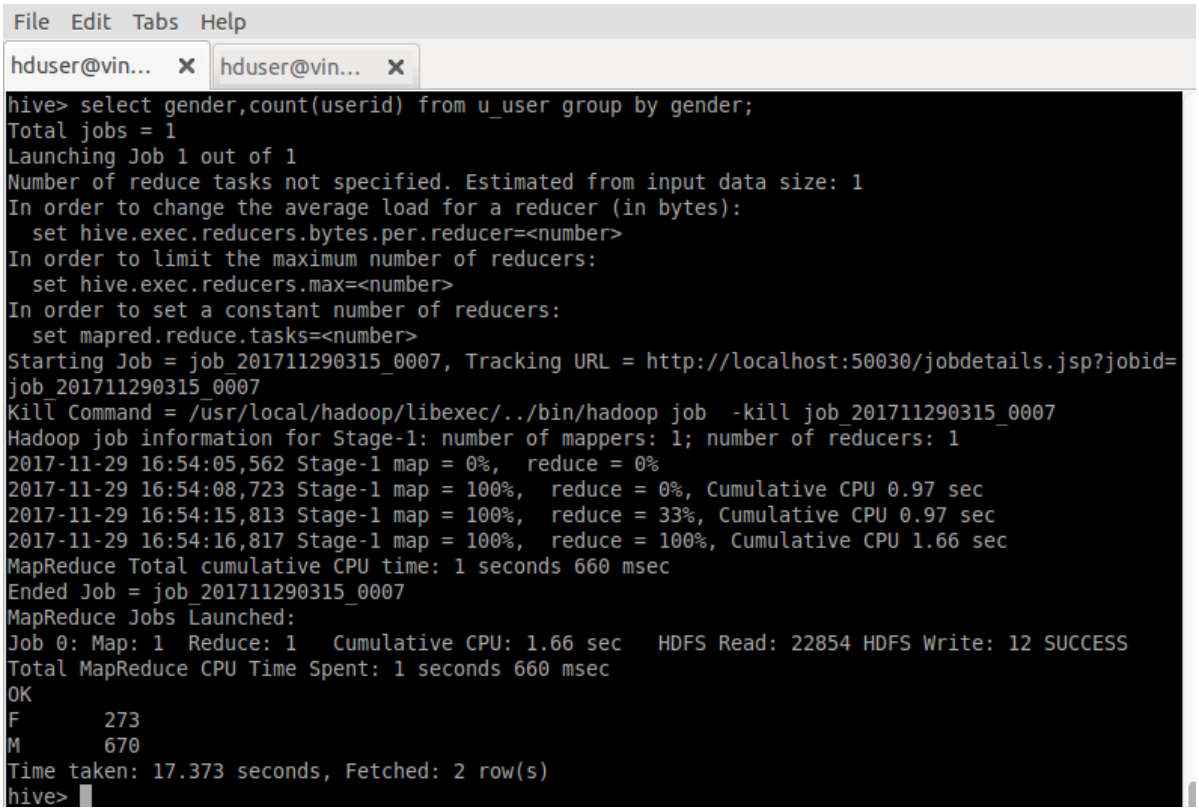
Select * from u_user;



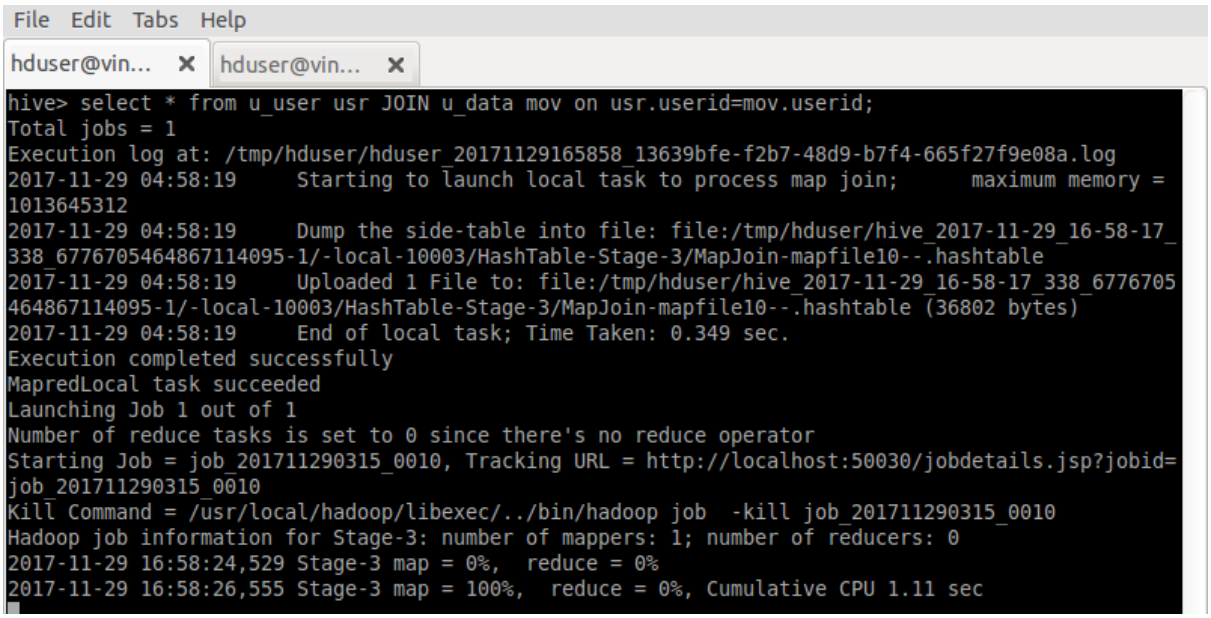
Question 11:

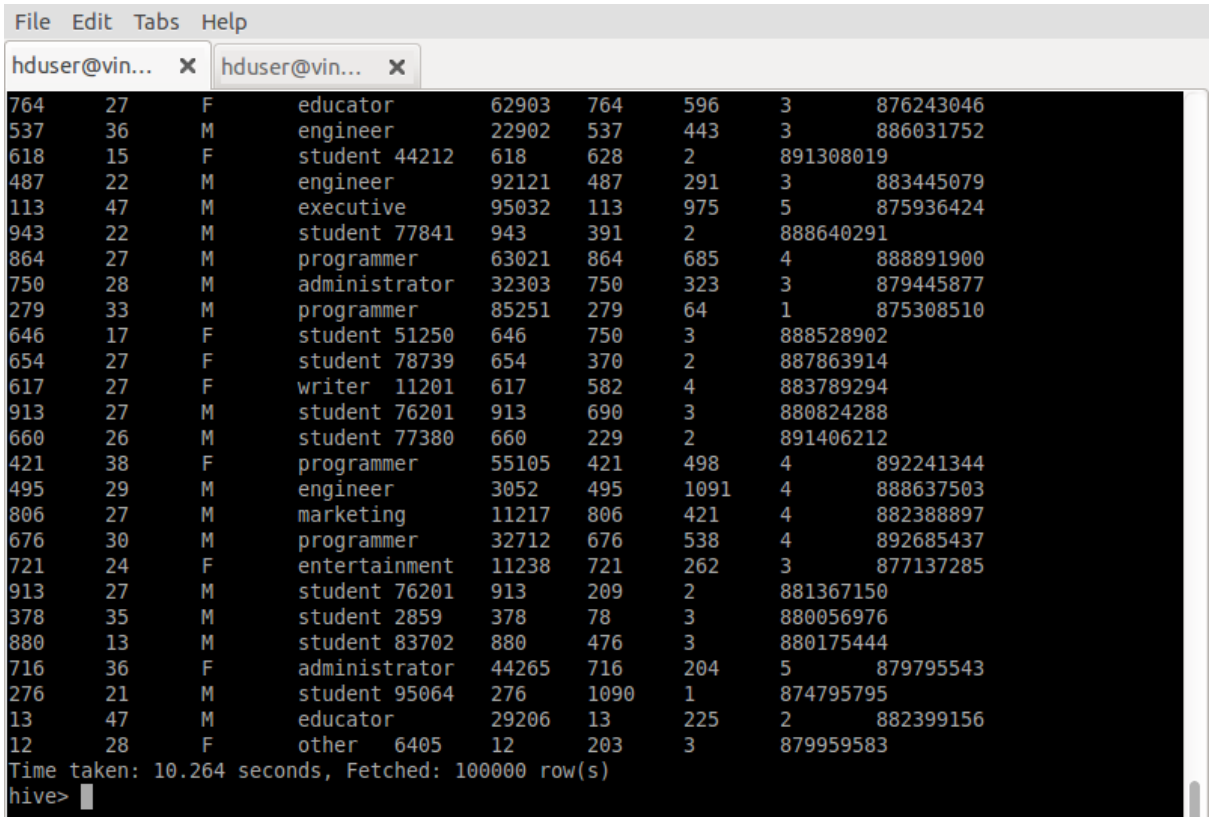


Question 12:

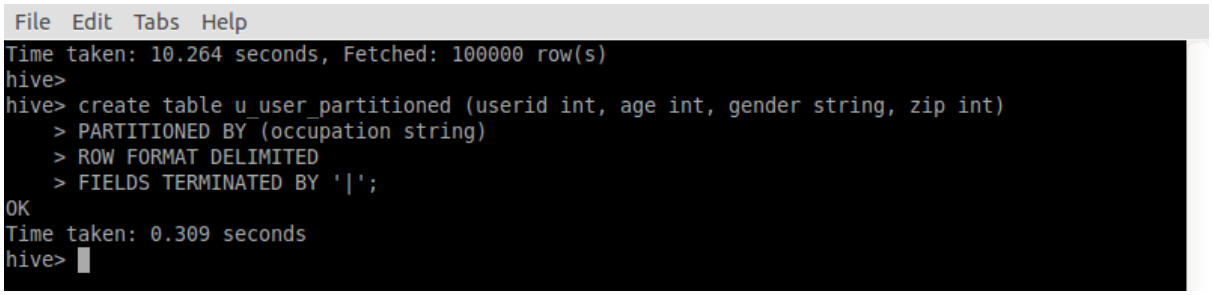

File Edit Tabs Help
hduser@vin...
hduser@vin...
hive> select gender,count(userid) from u_user group by gender;
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<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 mapred.reduce.tasks=<number>
Starting Job = job_201711290315_0007, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201711290315_0007
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job -kill job_201711290315_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-29 16:54:05,562 Stage-1 map = 0%, reduce = 0%
2017-11-29 16:54:08,723 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 0.97 sec
2017-11-29 16:54:15,813 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 0.97 sec
2017-11-29 16:54:16,817 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 1.66 sec
MapReduce Total cumulative CPU time: 1 seconds 660 msec
Ended Job = job_201711290315_0007
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Cumulative CPU: 1.66 sec HDFS Read: 22854 HDFS Write: 12 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 660 msec
OK
F 273
M 670
Time taken: 17.373 seconds, Fetched: 2 row(s)
hive>

Question 13:


File Edit Tabs Help
hduser@vin...
hduser@vin...
hive> select * from u_user usr JOIN u_data mov on usr.userid=mov.userid;
Total jobs = 1
Execution log at: /tmp/hduser/hduser_20171129165858_13639bfe-f2b7-48d9-b7f4-665f27f9e08a.log
2017-11-29 04:58:19 Starting to launch local task to process map join; maximum memory = 1013645312
2017-11-29 04:58:19 Dump the side-table into file: file:/tmp/hduser/hive_2017-11-29_16-58-17_338_6776705464867114095-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable
2017-11-29 04:58:19 Uploaded 1 File to: file:/tmp/hduser/hive_2017-11-29_16-58-17_338_6776705464867114095-1/-local-10003/HashTable-Stage-3/MapJoin-mapfile10--.hashtable (36802 bytes)
2017-11-29 04:58:19 End of local task; Time Taken: 0.349 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_201711290315_0010, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201711290315_0010
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job -kill job_201711290315_0010
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2017-11-29 16:58:24,529 Stage-3 map = 0%, reduce = 0%
2017-11-29 16:58:26,555 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 1.11 sec



Question 14:



Question 15:

```
hive> select occupation,gender,count(occupation) from u_user group by occupation,gender;
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 mapred.reduce.tasks=<number>
Starting Job = job_201711290315_0017, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=
job_201711290315_0017
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job  -kill job_201711290315_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-29 18:50:50,231 Stage-1 map = 0%,  reduce = 0%
2017-11-29 18:50:52,244 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.84 sec
2017-11-29 18:51:00,303 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 0.84 sec
2017-11-29 18:51:01,309 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 1.87 sec
MapReduce Total cumulative CPU time: 1 seconds 870 msec
Ended Job = job_201711290315_0017
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.87 sec   HDFS Read: 22854 HDFS Write: 569 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 870 msec
OK
administrator   F       36
administrator   M       43
artist  F       13
artist  M       15
doctor  M       7
```

File Edit Tabs Help

```
healthcare      F       11
healthcare      M       5
homemaker       F       6
homemaker       M       1
lawyer  F       2
lawyer  M       10
librarian       F       29
librarian       M       22
marketing       F       10
marketing       M       16
none    F       4
none    M       5
other   F       36
other   M       69
programmer      F       6
programmer      M       60
retired F       1
retired M       13
salesman        F       3
salesman        M       9
scientist       F       3
scientist       M       28
student F       60
student M       136
technician      F       1
technician      M       26
writer  F       19
writer  M       26
Time taken: 19.591 seconds, Fetched: 41 row(s)
hive>
```