

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... x hduser@vin... x
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... x hduser@vin... x
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... x hduser@vin... x
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... x hduser@vin... x
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:

```
File Edit Tabs Help
hduser@vin... x hduser@vin... x
764 596 3 876243046
537 443 3 886031752
618 628 2 891308019
487 291 3 883445079
113 975 5 875936424
943 391 2 888640291
864 685 4 888891900
750 323 3 879445877
279 64 1 875308510
646 750 3 888528902
654 370 2 887863914
617 582 4 883789294
913 690 3 880824288
660 229 2 891406212
421 498 4 892241344
495 1091 4 888637503
806 421 4 882388897
676 538 4 892685437
721 262 3 877137285
913 209 2 881367150
378 78 3 880056976
880 476 3 880175444
716 204 5 879795543
276 1090 1 874795795
13 225 2 882399156
12 203 3 879959583
Time taken: 0.03 seconds, Fetched: 100000 row(s)
hive>
```

Question 5:

```
File Edit Tabs Help
hduser@vin... x hduser@vin... x
173 324
151 326
210 331
79 336
405 344
204 350
313 350
222 365
172 367
117 378
237 384
98 390
7 392
56 394
127 413
174 420
121 429
300 431
1 452
288 478
286 481
294 485
181 507
100 508
258 509
50 583
Time taken: 62.66 seconds, Fetched: 1682 row(s)
hive>
```

Question 6:

```
File Edit Tabs Help
hduser@vin... X hduser@vin... X
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... X hduser@vin... X
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... x hduser@vin... x
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;

```
File Edit Tabs Help
hduser@vin... x hduser@vin... x
918 40 M scientist 70116
919 25 M other 14216
920 30 F artist 90008
921 20 F student 98801
922 29 F administrator 21114
923 21 M student NULL
924 29 M other 11753
925 18 F salesman 49036
926 49 M entertainment 1701
927 23 M programmer 55428
928 21 M student 55408
929 44 M scientist 53711
930 28 F scientist 7310
931 60 M educator 33556
932 58 M educator 6437
933 28 M student 48105
934 61 M engineer 22902
935 42 M doctor 66221
936 24 M other 32789
937 48 M educator 98072
938 38 F technician 55038
939 26 F student 33319
940 32 M administrator 2215
941 20 M student 97229
942 48 F librarian 78209
943 22 M student 77841
Time taken: 0.127 seconds, Fetched: 943 row(s)
hive>
```

Question 11:

```
File Edit Tabs Help
hduser@vin... x hduser@vin... x
hive> select count(*) from u_user;
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 mapred.reduce.tasks=<number>
Starting Job = job_201711290315_0006, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=
job_201711290315_0006
Kill Command = /usr/local/hadoop/libexec/./bin/hadoop job -kill job_201711290315_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-29 16:50:22,344 Stage-1 map = 0%, reduce = 0%
2017-11-29 16:50:29,737 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.74 sec
2017-11-29 16:50:38,914 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.32 sec
MapReduce Total cumulative CPU time: 4 seconds 320 msec
Ended Job = job_201711290315_0006
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Cumulative CPU: 4.32 sec HDFS Read: 22854 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 320 msec
OK
943
Time taken: 24.303 seconds, Fetched: 1 row(s)
hive>
```

Question 12:

```
File Edit Tabs Help
hduser@vin... X hduser@vin... X
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... X hduser@vin... X
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_6776705
464867114095-1/-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
```



```
File Edit Tabs Help
hduser@vin... x hduser@vin... x
764 27 F educator 62903 764 596 3 876243046
537 36 M engineer 22902 537 443 3 886031752
618 15 F student 44212 618 628 2 891308019
487 22 M engineer 92121 487 291 3 883445079
113 47 M executive 95032 113 975 5 875936424
943 22 M student 77841 943 391 2 888640291
864 27 M programmer 63021 864 685 4 888891900
750 28 M administrator 32303 750 323 3 879445877
279 33 M programmer 85251 279 64 1 875308510
646 17 F student 51250 646 750 3 888528902
654 27 F student 78739 654 370 2 887863914
617 27 F writer 11201 617 582 4 883789294
913 27 M student 76201 913 690 3 880824288
660 26 M student 77380 660 229 2 891406212
421 38 F programmer 55105 421 498 4 892241344
495 29 M engineer 3052 495 1091 4 888637503
806 27 M marketing 11217 806 421 4 882388897
676 30 M programmer 32712 676 538 4 892685437
721 24 F entertainment 11238 721 262 3 877137285
913 27 M student 76201 913 209 2 881367150
378 35 M student 2859 378 78 3 880056976
880 13 M student 83702 880 476 3 880175444
716 36 F administrator 44265 716 204 5 879795543
276 21 M student 95064 276 1090 1 874795795
13 47 M educator 29206 13 225 2 882399156
12 28 F other 6405 12 203 3 879959583
Time taken: 10.264 seconds, Fetched: 100000 row(s)
hive>
```

Question 14:

```
File Edit Tabs Help
Time taken: 10.264 seconds, Fetched: 100000 row(s)
hive>
hive> create table u_user_partitioned (userid int, age int, gender string, zip int)
> PARTITIONED BY (occupation string)
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY '|';
OK
Time taken: 0.309 seconds
hive>
```

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>
```