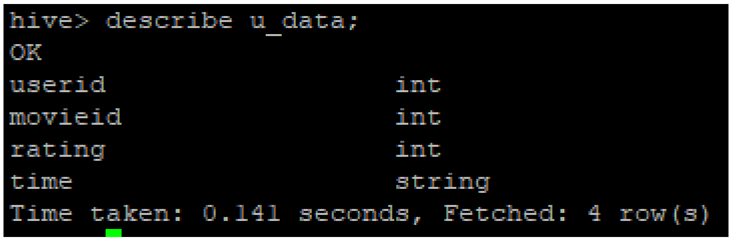
Database - cts_jig13302

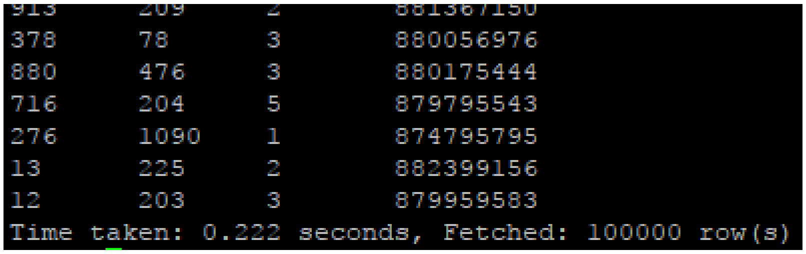
1. CREATE EXTERNAL TABLE u_data (userId INT, movieId INT, rating INT, time STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;

2. Describe u_data;



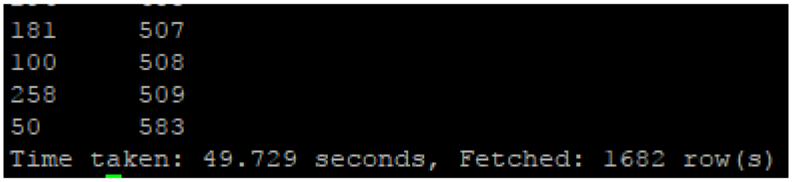
```
hive> describe u_data;
OK
userid                  int
movieid                 int
rating                  int
time                    string
Time taken: 0.141 seconds, Fetched: 4 row(s)
```

3. LOAD DATA LOCAL INPATH '/home/data/cts/u.data' OVERWRITE INTO TABLE u_data;
SELECT * FROM u_data;



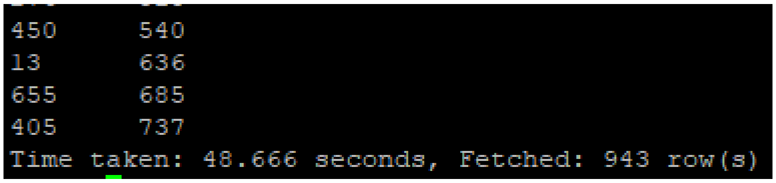
```
913     209     2       881367150
378     78      3       880056976
880     476     3       880175444
716     204     5       879795543
276     1090    1       874795795
13      225     2       882399156
12      203     3       879959583
Time taken: 0.222 seconds, Fetched: 100000 row(s)
```

4. SELECT movieid, COUNT(userid) AS no from u_data GROUP BY movieid ORDER BY no;



```
181     507
100     508
258     509
50      583
Time taken: 49.729 seconds, Fetched: 1682 row(s)
```

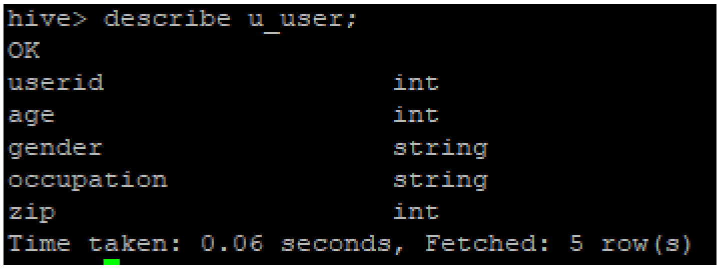
5. SELECT userid, COUNT(movieid) AS no from u_data GROUP BY userid ORDER BY no;



```
450     540
13      636
655     685
405     737
Time taken: 48.666 seconds, Fetched: 943 row(s)
```

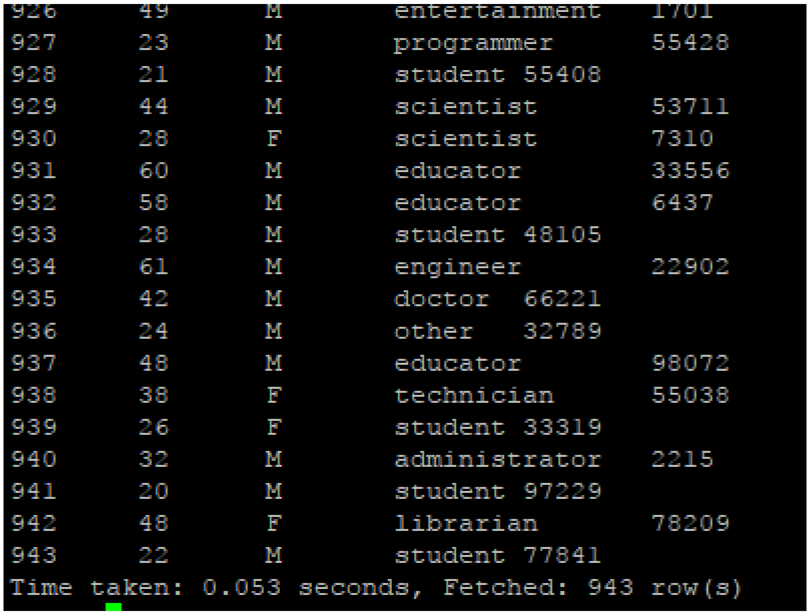
6. CREATE EXTERNAL TABLE u_user (userid INT, age INT, gender STRING, occupation STRING, zip INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE;

7. DESCRIBE u_user;



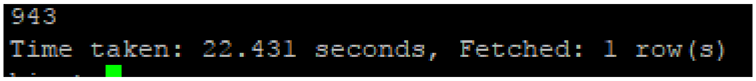
```
hive> describe u_user;
OK
userid                  int
age                     int
gender                  string
occupation              string
zip                     int
Time taken: 0.06 seconds, Fetched: 5 row(s)
```

8. LOAD DATA LOCAL INPATH 'home/data/cts/u.user' OVERWRITE INTO TABLE u_user;
SELECT * from u_user;



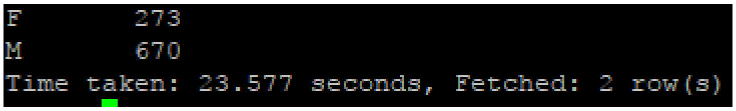
```
926     49      M       entertainment   1701
927     23      M       programmer      55428
928     21      M       student 55408
929     44      M       scientist       53711
930     28      F       scientist       7310
931     60      M       educator        33556
932     58      M       educator        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: 0.053 seconds, Fetched: 943 row(s)
```

9. SELECT COUNT(*) from u_user;



```
943
Time taken: 22.431 seconds, Fetched: 1 row(s)
```

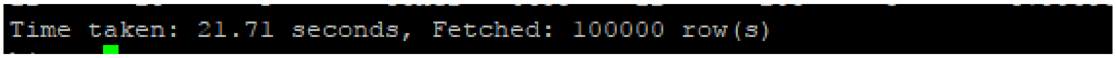
10. SELECT gender, COUNT(*) from u_user GROUP BY gender;



```
F       273
M       670
Time taken: 23.577 seconds, Fetched: 2 row(s)
```

11. (a) Reduce Side Join

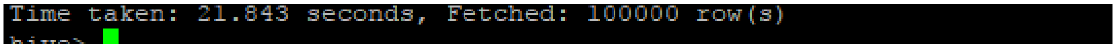
SELECT * from u_user usr JOIN u_data mov ON usr.userid=mov.userid;



```
Time taken: 21.71 seconds, Fetched: 100000 row(s)
```

(b) Map-side Join

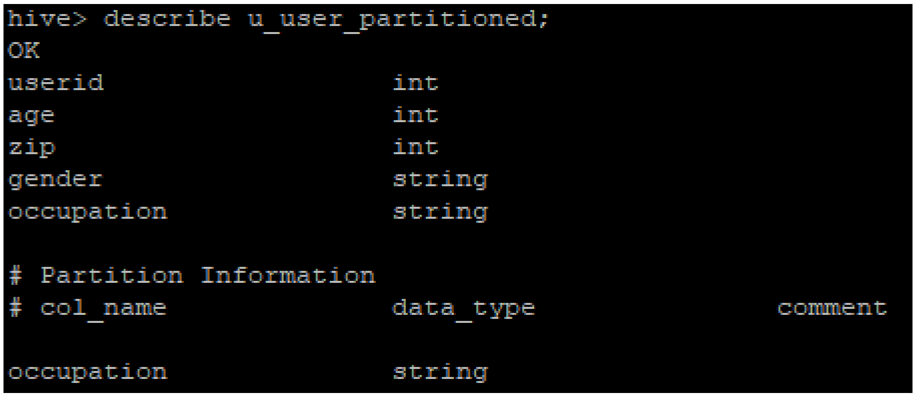
SELECT /*+ MAPJOIN(usr) */ * from u_user usr JOIN u_data mov ON usr.userid=mov.userid;



```
Time taken: 21.843 seconds, Fetched: 100000 row(s)
```

Reduce join is faster when compared to Map-side join. In local VM the difference is much more when compared to AWS cluster.

12. CREATE TABLE u_user_partitioned (userId INT, age INT , zip INT, gender STRING) PARTITIONED BY (occupation STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS SEQUENCEFILE;

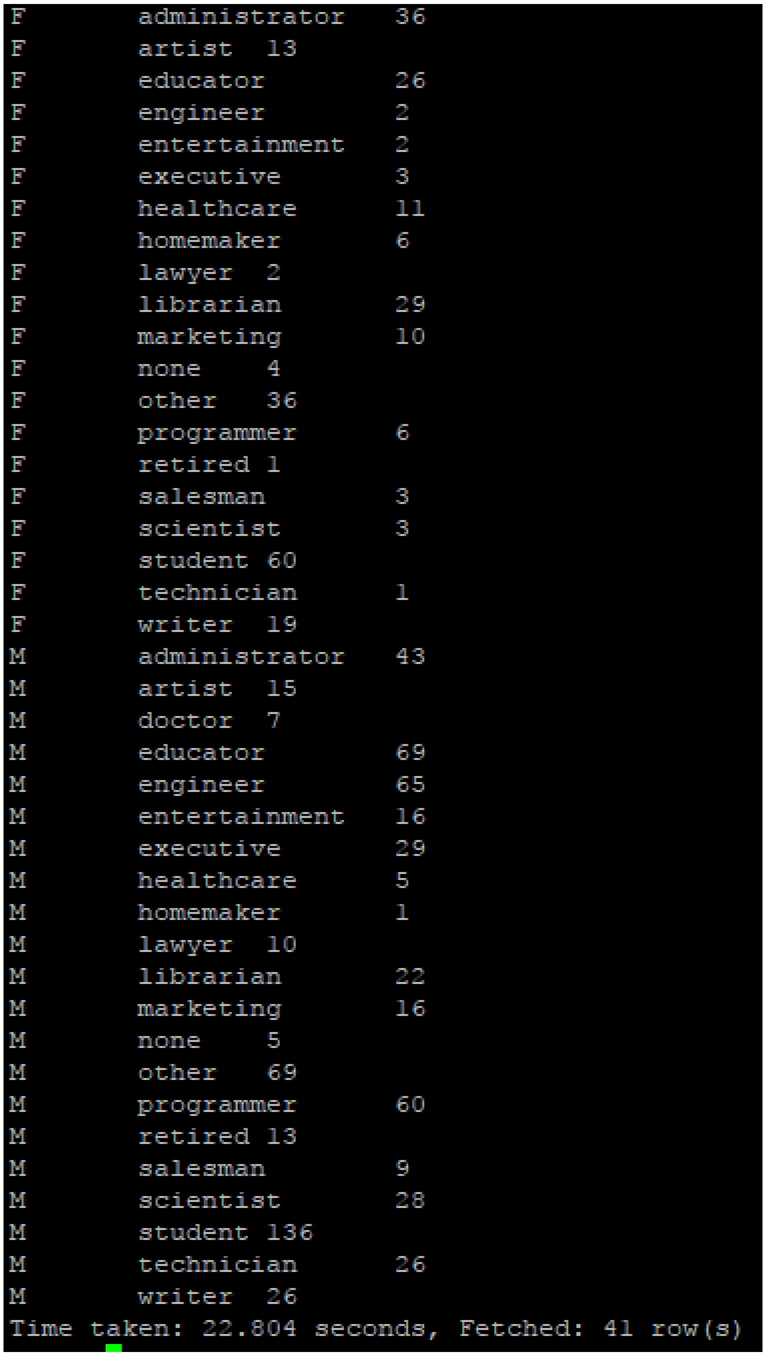


```
hive> describe u_user_partitioned;
OK
userid                  int
age                     int
zip                     int
gender                  string
occupation              string

# Partition Information
# col_name              data_type               comment

occupation              string
```

INSERT INTO TABLE u_user_partitioned PARTITION(occupation) SELECT userid, age, zip, gender, occupation from u_user;

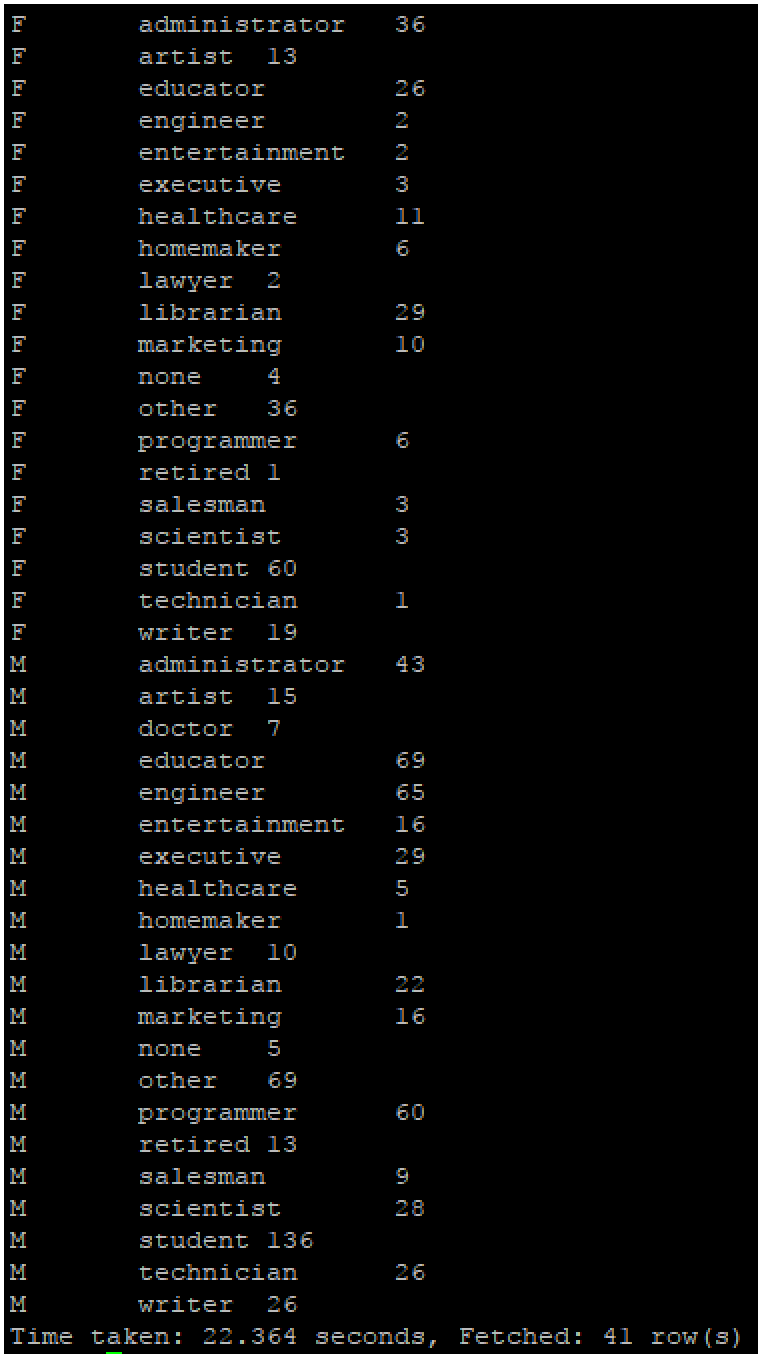
13. (a) With Partition

SELECT gender, occupation , COUNT(*) from u_user_partitioned GROUP BY gender, occupation;



```
F       administrator   36
F       artist  13
F       educator        26
F       engineer        2
F       entertainment   2
F       executive       3
F       healthcare      11
F       homemaker       6
F       lawyer  2
F       librarian       29
F       marketing       10
F       none    4
F       other   36
F       programmer      6
F       retired 1
F       salesman        3
F       scientist       3
F       student 60
F       technician      1
F       writer  19
M       administrator   43
M       artist  15
M       doctor  7
M       educator        69
M       engineer        65
M       entertainment   16
M       executive       29
M       healthcare      5
M       homemaker       1
M       lawyer  10
M       librarian       22
M       marketing       16
M       none    5
M       other   69
M       programmer      60
M       retired 13
M       salesman        9
M       scientist       28
M       student 136
M       technician      26
M       writer  26
Time taken: 22.804 seconds, Fetched: 41 row(s)
```

(b) Without Partition



```
F       administrator   36
F       artist  13
F       educator        26
F       engineer        2
F       entertainment   2
F       executive       3
F       healthcare      11
F       homemaker       6
F       lawyer  2
F       librarian       29
F       marketing       10
F       none    4
F       other   36
F       programmer      6
F       retired 1
F       salesman        3
F       scientist       3
F       student 60
F       technician      1
F       writer  19
M       administrator   43
M       artist  15
M       doctor  7
M       educator        69
M       engineer        65
M       entertainment   16
M       executive       29
M       healthcare      5
M       homemaker       1
M       lawyer  10
M       librarian       22
M       marketing       16
M       none    5
M       other   69
M       programmer      60
M       retired 13
M       salesman        9
M       scientist       28
M       student 136
M       technician      26
M       writer  26
Time taken: 22.364 seconds, Fetched: 41 row(s)
```

Performance of without partition is more than the one with the partitioned table.