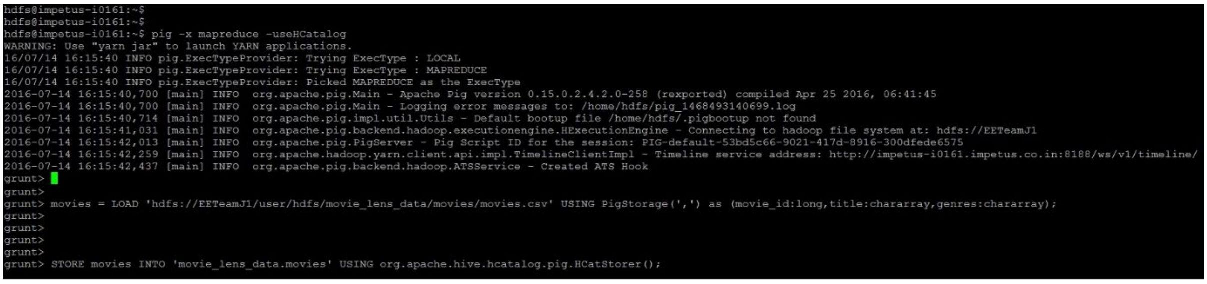
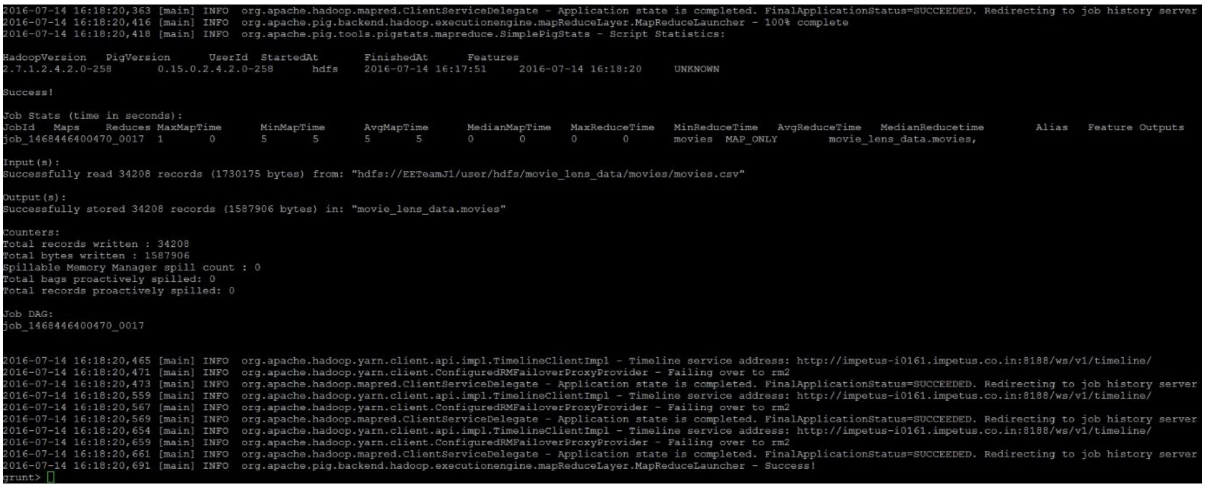
Data upload using PIG

Load Movies data into HIVE table:



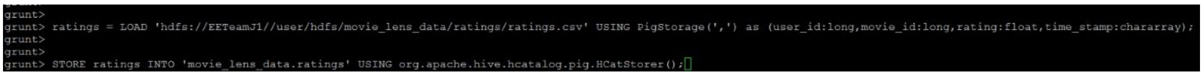
```
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ pig -x mapreduce -useHCatalog
WARNING: Use "yarn jar" to launch YARN applications.
16/07/14 16:15:40 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/07/14 16:15:40 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/07/14 16:15:40 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-07-14 16:15:40,700 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0.2.4.2.0-258 (rexported) compiled Apr 25 2016, 06:41:45
2016-07-14 16:15:40,700 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hdfs/pig_1468493140699.log
2016-07-14 16:15:40,714 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hdfs/.pigbootup not found
2016-07-14 16:15:41,031 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://EETeamJ1
2016-07-14 16:15:42,013 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-53bd5c66-9021-417d-8916-300dfede6575
2016-07-14 16:15:42,259 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:15:42,437 [main] INFO  org.apache.pig.backend.hadoop.ATSService - Created ATS Hook
grunt>
grunt>
grunt> movies = LOAD 'hdfs://EETeamJ1/user/hdfs/movie_lens_data/movies/movies.csv' USING PigStorage(',') as (movie_id:long,title:chararray,genres:chararray);
grunt>
grunt>
grunt>
grunt>
grunt> STORE movies INTO 'movie_lens_data.movies' USING org.apache.hive.hcatalog.pig.HCatStorer();
```

Movie data upload result:



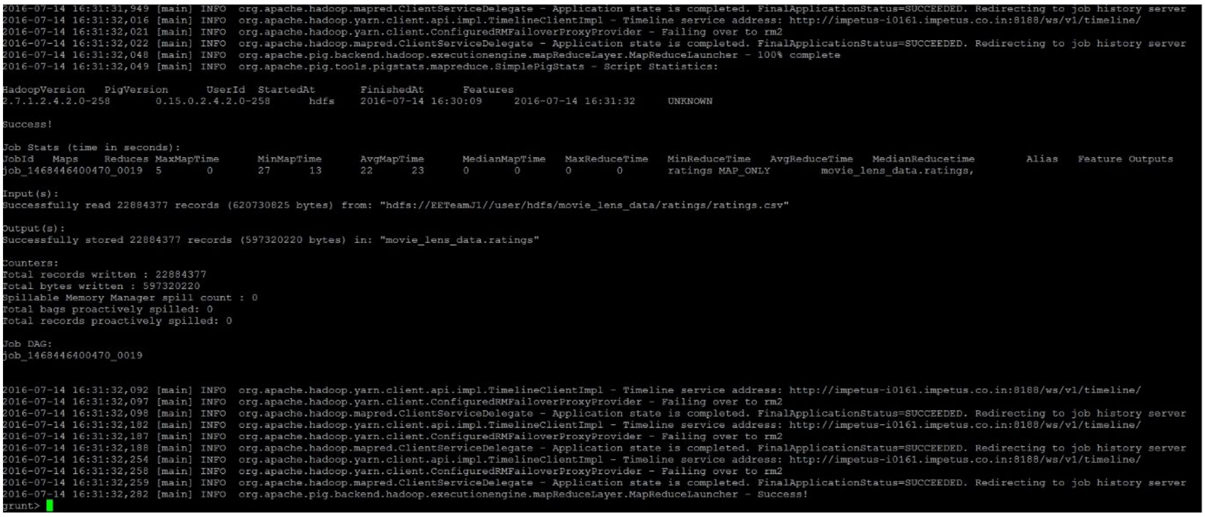
```
2016-07-14 16:18:20,363 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:18:20,416 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-14 16:18:20,418 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.7.1.2.4.2.0-258       0.15.0.2.4.2.0-258      hdfs    2016-07-14 16:17:51     2016-07-14 16:18:20     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1468446400470_0017  1       0       5       5       5       5       0       0       0       0       movies  MAP_ONLY        movie_lens_data.movies,

Input(s):
Successfully read 34208 records (1730175 bytes) from: "hdfs://EETeamJ1/user/hdfs/movie_lens_data/movies/movies.csv"

Output(s):
Successfully stored 34208 records (1587906 bytes) in: "movie_lens_data.movies"

Counters:
Total records written : 34208
Total bytes written : 1587906
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1468446400470_0017


2016-07-14 16:18:20,465 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:18:20,471 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:18:20,473 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:18:20,559 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:18:20,567 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:18:20,569 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:18:20,654 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:18:20,659 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:18:20,661 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:18:20,691 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Load Ratings data into HIVE table:



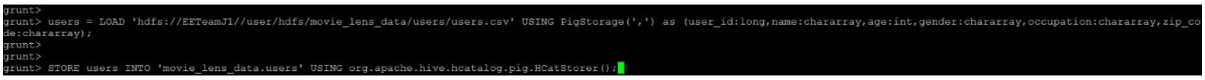
```
grunt>
grunt> ratings = LOAD 'hdfs://EETeamJ1//user/hdfs/movie_lens_data/ratings/ratings.csv' USING PigStorage(',') as (user_id:long,movie_id:long,rating:float,time_stamp:chararray);
grunt>
grunt>
grunt> STORE ratings INTO 'movie_lens_data.ratings' USING org.apache.hive.hcatalog.pig.HCatStorer();
```

Ratings data upload result:



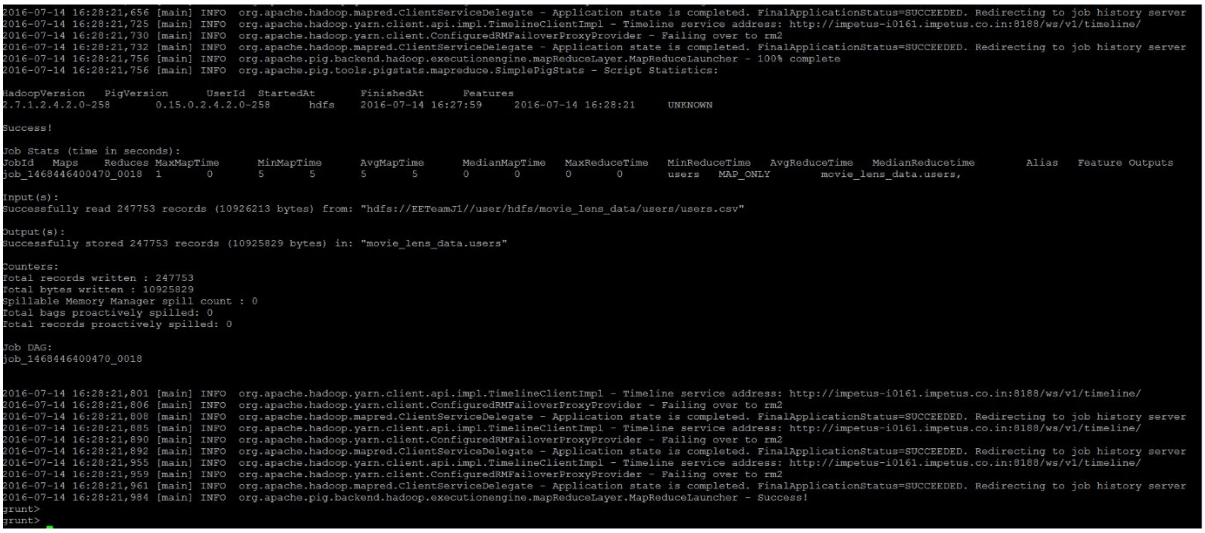
```
2016-07-14 16:31:31,949 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:31:32,016 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:31:32,021 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:31:32,022 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:31:32,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-14 16:31:32,049 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.7.1.2.4.2.0-258       0.15.0.2.4.2.0-258      hdfs    2016-07-14 16:30:09     2016-07-14 16:31:32     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1468446400470_0019  5       0       27      13      22      23      0       0       0       0       ratings MAP_ONLY        movie_lens_data.ratings,

Input(s):
Successfully read 22884377 records (620730825 bytes) from: "hdfs://EETeamJ1//user/hdfs/movie_lens_data/ratings/ratings.csv"

Output(s):
Successfully stored 22884377 records (597320220 bytes) in: "movie_lens_data.ratings"

Counters:
Total records written : 22884377
Total bytes written : 597320220
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1468446400470_0019


2016-07-14 16:31:32,092 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:31:32,097 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:31:32,098 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:31:32,182 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:31:32,187 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:31:32,188 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:31:32,254 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:31:32,258 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:31:32,259 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:31:32,282 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Load Users data into HIVE table:



```
grunt>
grunt> users = LOAD 'hdfs://EETeamJ1//user/hdfs/movie_lens_data/users/users.csv' USING PigStorage(',') as (user_id:long,name:chararray,age:int,gender:chararray,occupation:chararray,zip_co
de:chararray);
grunt>
grunt>
grunt> STORE users INTO 'movie_lens_data.users' USING org.apache.hive.hcatalog.pig.HCatStorer();
```

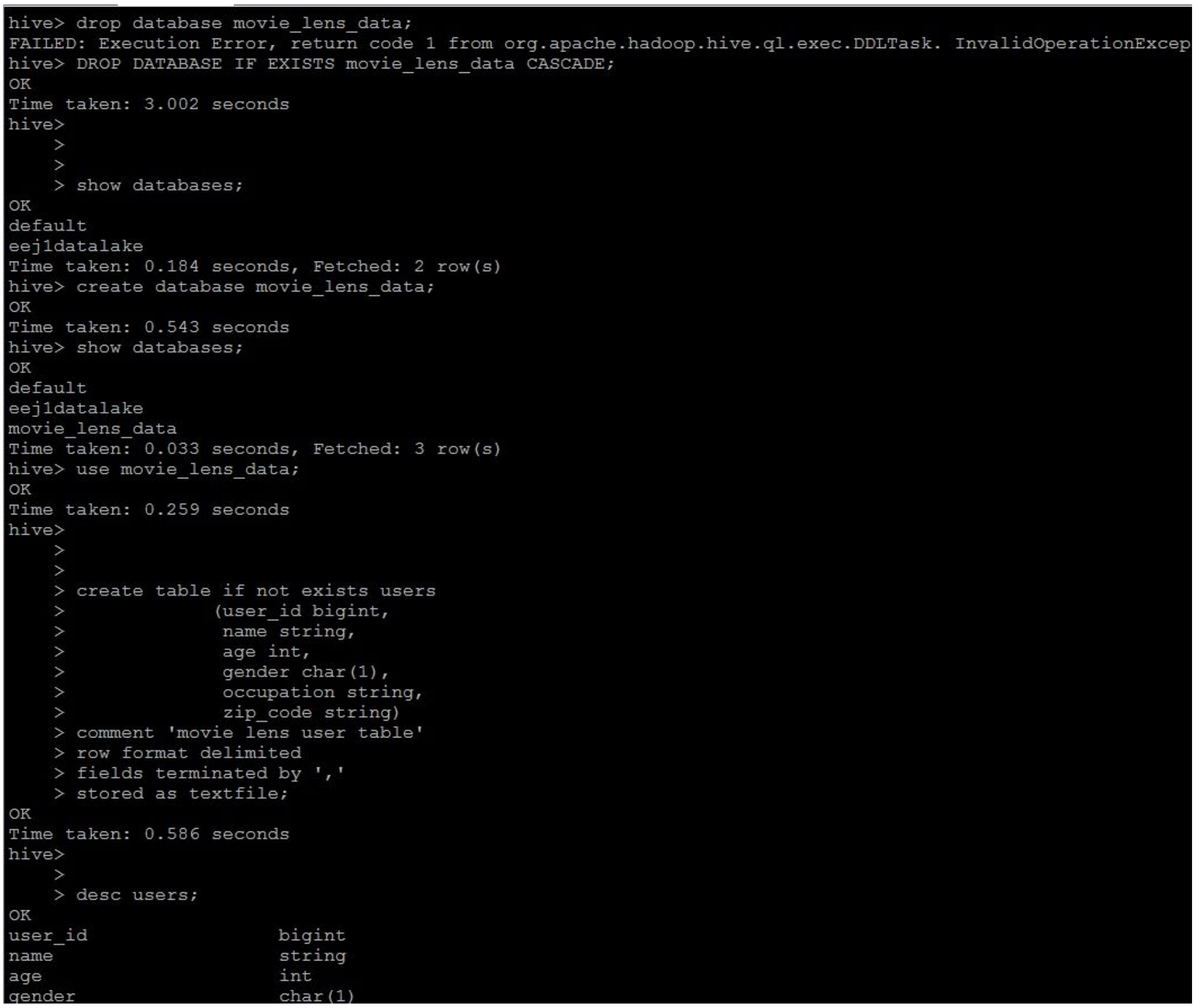
Users data upload result:



```
2016-07-14 16:28:21,656 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:28:21,725 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:28:21,730 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:28:21,732 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:28:21,756 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-14 16:28:21,756 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.7.1.2.4.2.0-258       0.15.0.2.4.2.0-258      hdfs    2016-07-14 16:27:59     2016-07-14 16:28:21     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1468446400470_0018  1       0       5       5       5       5       0       0       0       0       users   MAP_ONLY        movie_lens_data.users,

Input(s):
Successfully read 247753 records (10926213 bytes) from: "hdfs://EETeamJ1//user/hdfs/movie_lens_data/users/users.csv"

Output(s):
Successfully stored 247753 records (10925829 bytes) in: "movie_lens_data.users"

Counters:
Total records written : 247753
Total bytes written : 10925829
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1468446400470_0018


2016-07-14 16:28:21,801 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:28:21,806 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:28:21,808 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:28:21,885 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:28:21,890 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:28:21,892 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:28:21,955 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: http://impetus-i0161.impetus.co.in:8188/ws/v1/timeline/
2016-07-14 16:28:21,959 [main] INFO  org.apache.hadoop.yarn.client.ConfiguredRMFailoverProxyProvider - Failing over to rm2
2016-07-14 16:28:21,961 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2016-07-14 16:28:21,984 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
grunt>
```

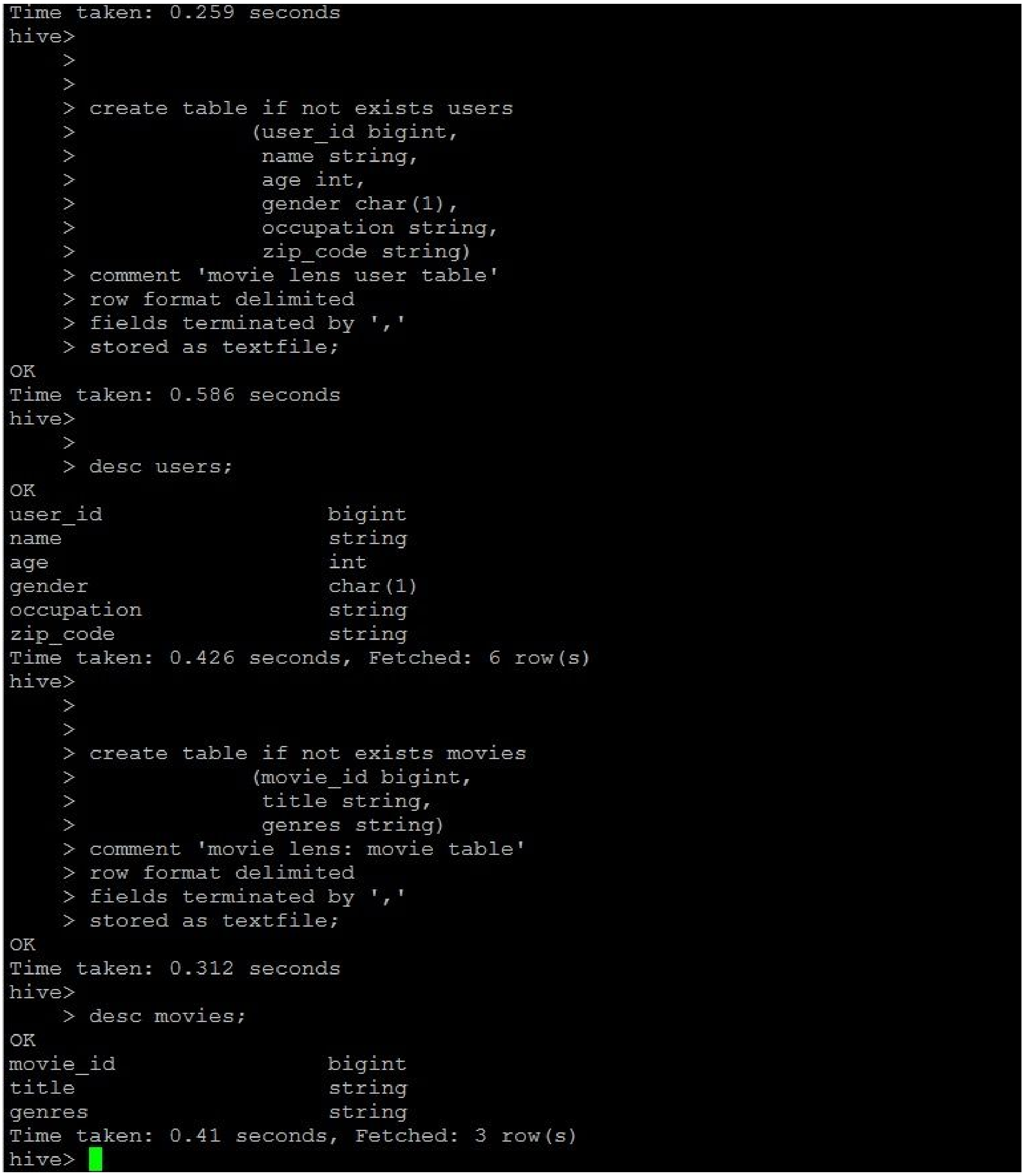
HIVE Queries execution:

Create Database & Users table:



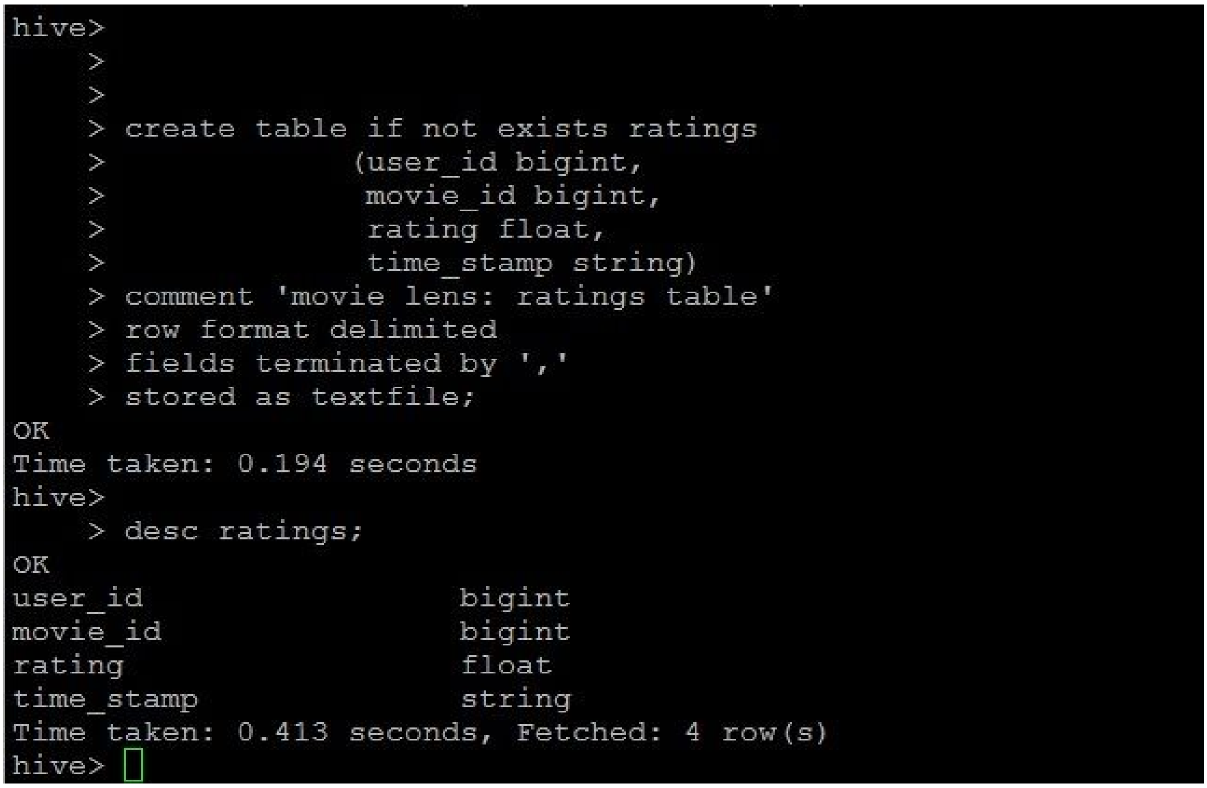
```
hive> drop database movie_lens_data;
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. InvalidOperationExcep
hive> DROP DATABASE IF EXISTS movie_lens_data CASCADE;
OK
Time taken: 3.002 seconds
hive>
    >
    >
    > show databases;
OK
default
eej1datalake
Time taken: 0.184 seconds, Fetched: 2 row(s)
hive> create database movie_lens_data;
OK
Time taken: 0.543 seconds
hive> show databases;
OK
default
eej1datalake
movie_lens_data
Time taken: 0.033 seconds, Fetched: 3 row(s)
hive> use movie_lens_data;
OK
Time taken: 0.259 seconds
hive>
    >
    >
    > create table if not exists users
    >             (user_id bigint,
    >              name string,
    >              age int,
    >              gender char(1),
    >              occupation string,
    >              zip_code string)
    > comment 'movie lens user table'
    > row format delimited
    > fields terminated by ','
    > stored as textfile;
OK
Time taken: 0.586 seconds
hive>
    >
    > desc users;
OK
user_id                 bigint
name                    string
age                     int
gender                  char(1)
```

Create Movies Table:



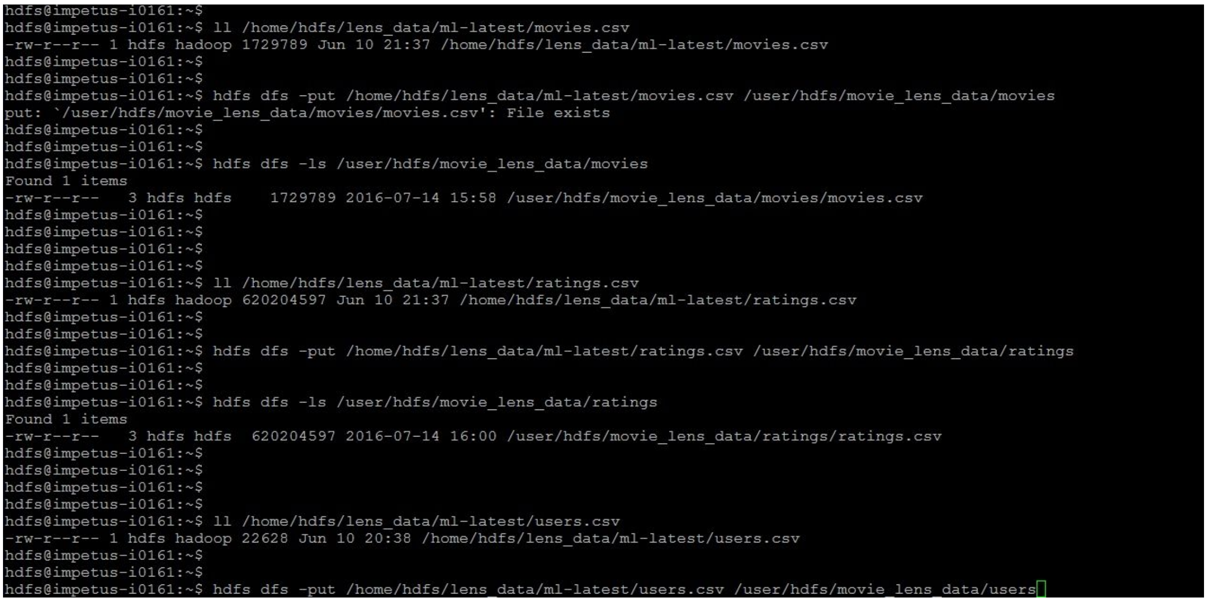
```
Time taken: 0.259 seconds
hive>
    >
    >
    > create table if not exists users
    >             (user_id bigint,
    >              name string,
    >              age int,
    >              gender char(1),
    >              occupation string,
    >              zip_code string)
    > comment 'movie lens user table'
    > row format delimited
    > fields terminated by ','
    > stored as textfile;
OK
Time taken: 0.586 seconds
hive>
    >
    > desc users;
OK
user_id                 bigint
name                    string
age                     int
gender                  char(1)
occupation              string
zip_code                string
Time taken: 0.426 seconds, Fetched: 6 row(s)
hive>
    >
    >
    > create table if not exists movies
    >             (movie_id bigint,
    >              title string,
    >              genres string)
    > comment 'movie lens: movie table'
    > row format delimited
    > fields terminated by ','
    > stored as textfile;
OK
Time taken: 0.312 seconds
hive>
    > desc movies;
OK
movie_id                bigint
title                   string
genres                  string
Time taken: 0.41 seconds, Fetched: 3 row(s)
hive> 
```

Create Ratings Table:



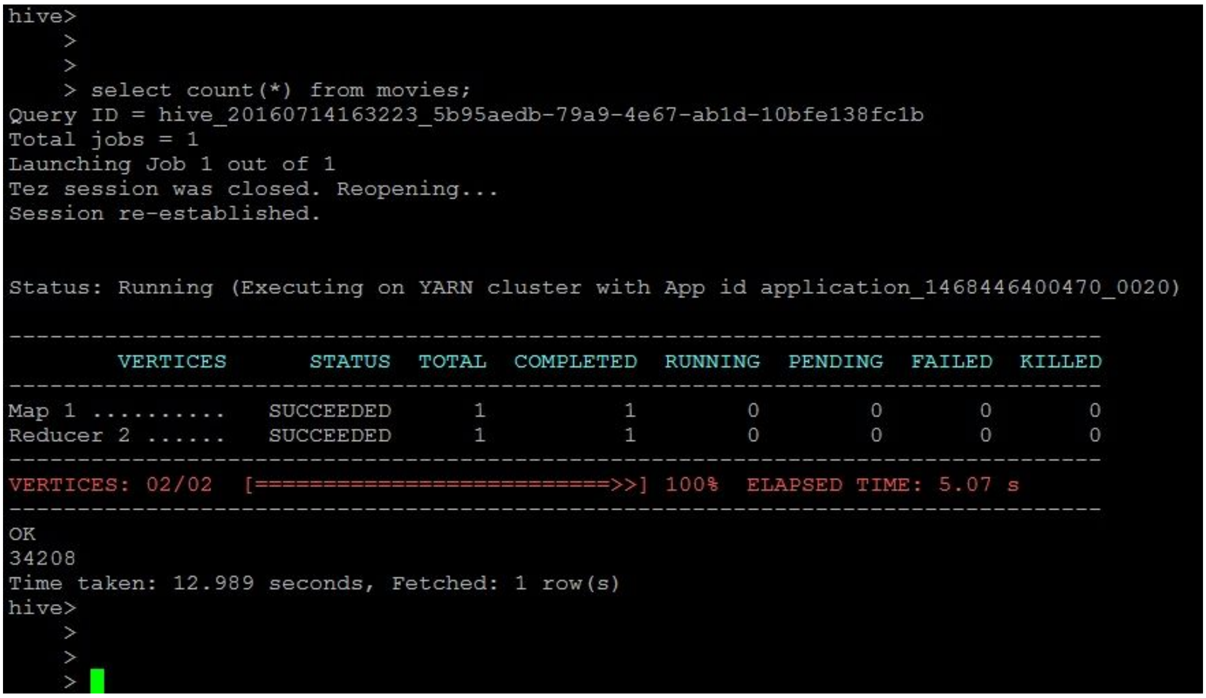
```
hive>
    >
    >
    > create table if not exists ratings
    >             (user_id bigint,
    >              movie_id bigint,
    >              rating float,
    >              time_stamp string)
    > comment 'movie lens: ratings table'
    > row format delimited
    > fields terminated by ','
    > stored as textfile;
OK
Time taken: 0.194 seconds
hive>
    > desc ratings;
OK
user_id                 bigint
movie_id                bigint
rating                  float
time_stamp              string
Time taken: 0.413 seconds, Fetched: 4 row(s)
hive> 
```

Load Data to HDFS:



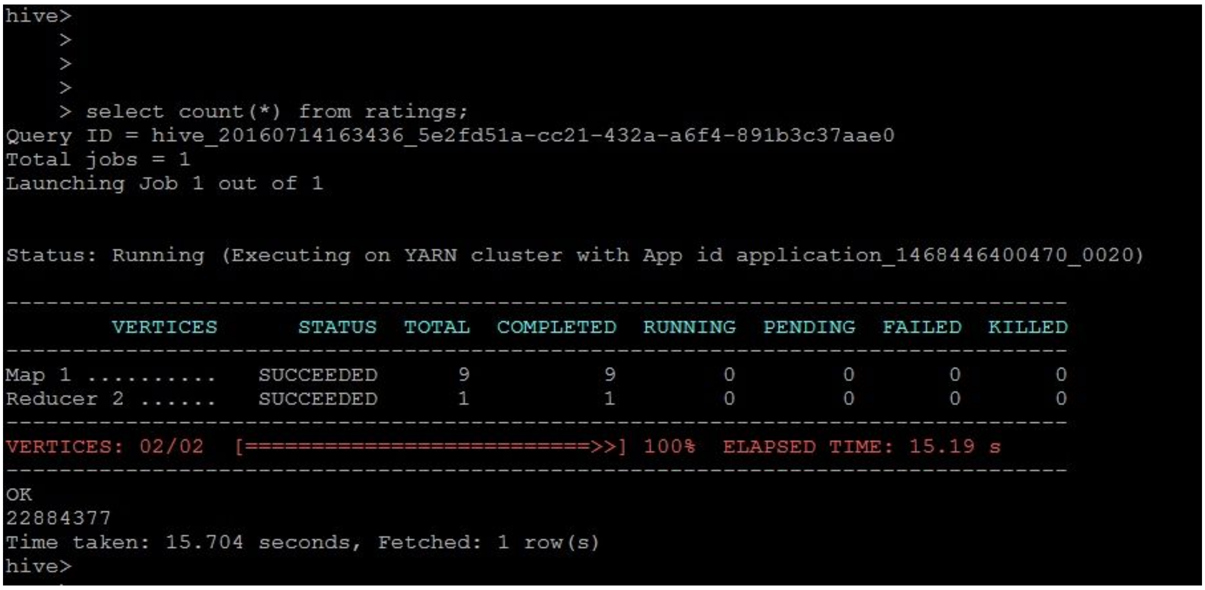
```
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ ll /home/hdfs/lens_data/ml-latest/movies.csv
-rw-r--r-- 1 hdfs hadoop 1729789 Jun 10 21:37 /home/hdfs/lens_data/ml-latest/movies.csv
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ hdfs dfs -put /home/hdfs/lens_data/ml-latest/movies.csv /user/hdfs/movie_lens_data/movies
put: `/user/hdfs/movie_lens_data/movies/movies.csv': File exists
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ hdfs dfs -ls /user/hdfs/movie_lens_data/movies
Found 1 items
-rw-r--r--   3 hdfs hdfs    1729789 2016-07-14 15:58 /user/hdfs/movie_lens_data/movies/movies.csv
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ ll /home/hdfs/lens_data/ml-latest/ratings.csv
-rw-r--r-- 1 hdfs hadoop 620204597 Jun 10 21:37 /home/hdfs/lens_data/ml-latest/ratings.csv
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ hdfs dfs -put /home/hdfs/lens_data/ml-latest/ratings.csv /user/hdfs/movie_lens_data/ratings
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ hdfs dfs -ls /user/hdfs/movie_lens_data/ratings
Found 1 items
-rw-r--r--   3 hdfs hdfs  620204597 2016-07-14 16:00 /user/hdfs/movie_lens_data/ratings/ratings.csv
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ ll /home/hdfs/lens_data/ml-latest/users.csv
-rw-r--r-- 1 hdfs hadoop 22628 Jun 10 20:38 /home/hdfs/lens_data/ml-latest/users.csv
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$
hdfs@impetus-i0161:~$ hdfs dfs -put /home/hdfs/lens_data/ml-latest/users.csv /user/hdfs/movie_lens_data/users
```

Movies Count Verification:



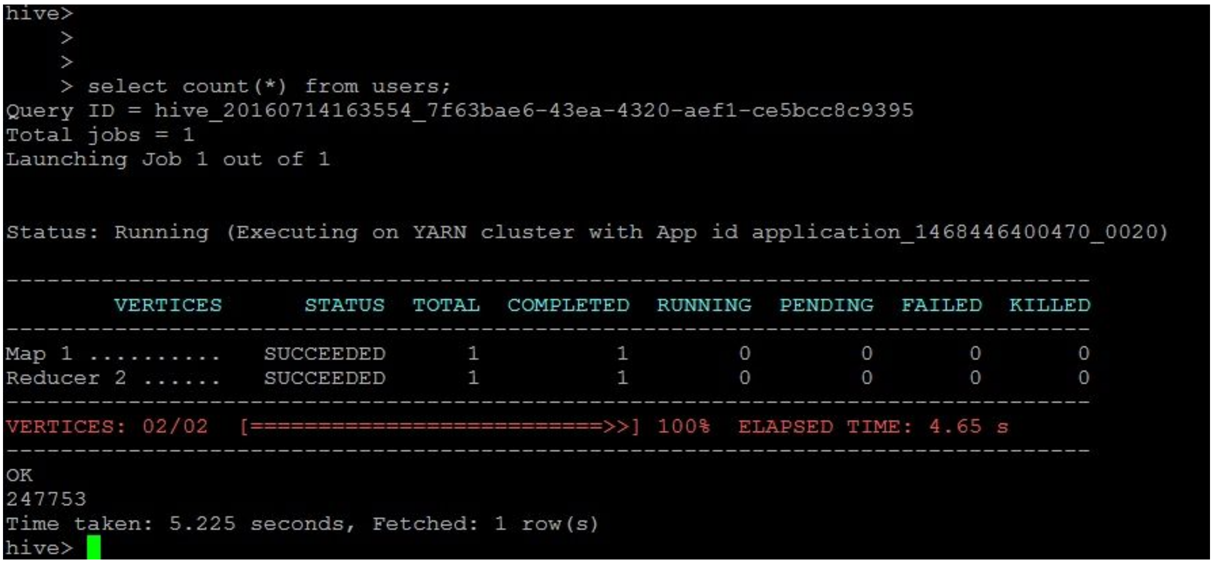
```
hive>
    >
    >
    > select count(*) from movies;
Query ID = hive_20160714163223_5b95aedb-79a9-4e67-ab1d-10bfe138fc1b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1468446400470_0020)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.07 s
--------------------------------------------------------------------------------
OK
34208
Time taken: 12.989 seconds, Fetched: 1 row(s)
hive>
    >
    >
    >
```

Ratings Count Verification:



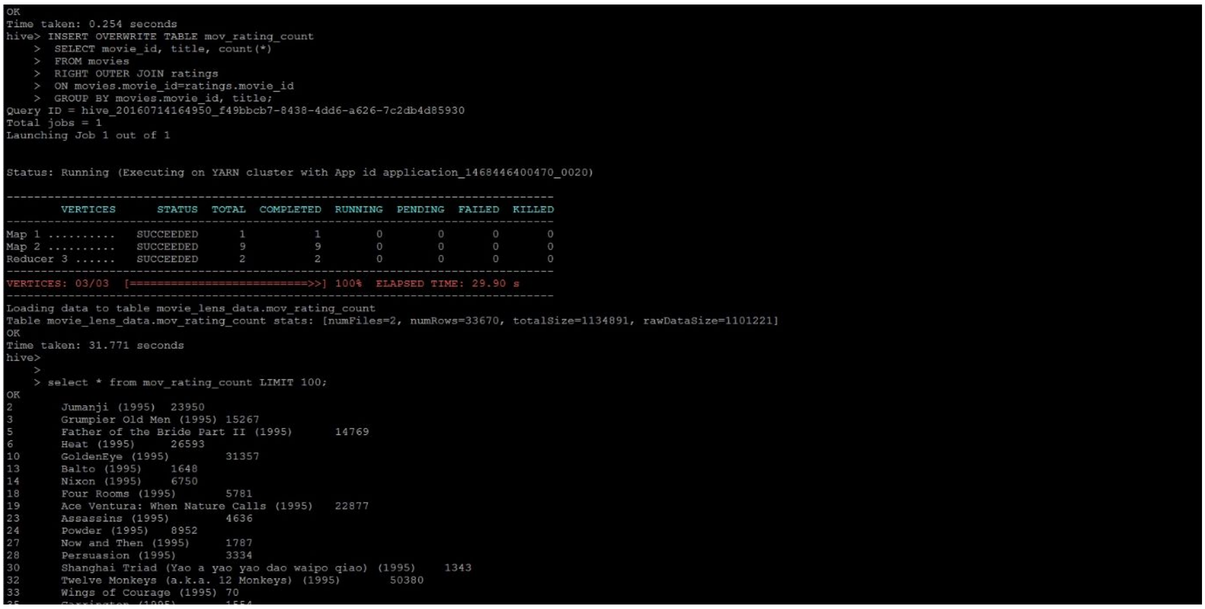
```
hive>
    >
    >
    >
    > select count(*) from ratings;
Query ID = hive_20160714163436_5e2fd51a-cc21-432a-a6f4-891b3c37aae0
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468446400470_0020)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      9          9        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 15.19 s
--------------------------------------------------------------------------------
OK
22884377
Time taken: 15.704 seconds, Fetched: 1 row(s)
hive>
```

Users Count Verification:



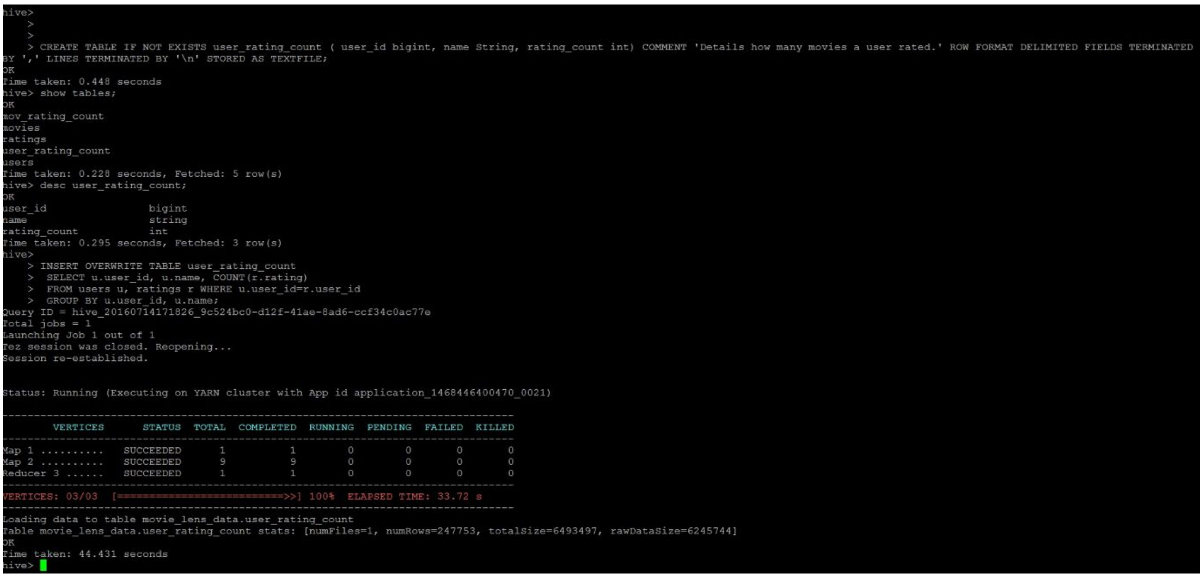
```
hive>
    >
    >
    > select count(*) from users;
Query ID = hive_20160714163554_7f63bae6-43ea-4320-aef1-ce5bcc8c9395
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468446400470_0020)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.65 s
--------------------------------------------------------------------------------
OK
247753
Time taken: 5.225 seconds, Fetched: 1 row(s)
hive>
```

Use Case #1: List all the movies and the number of ratings

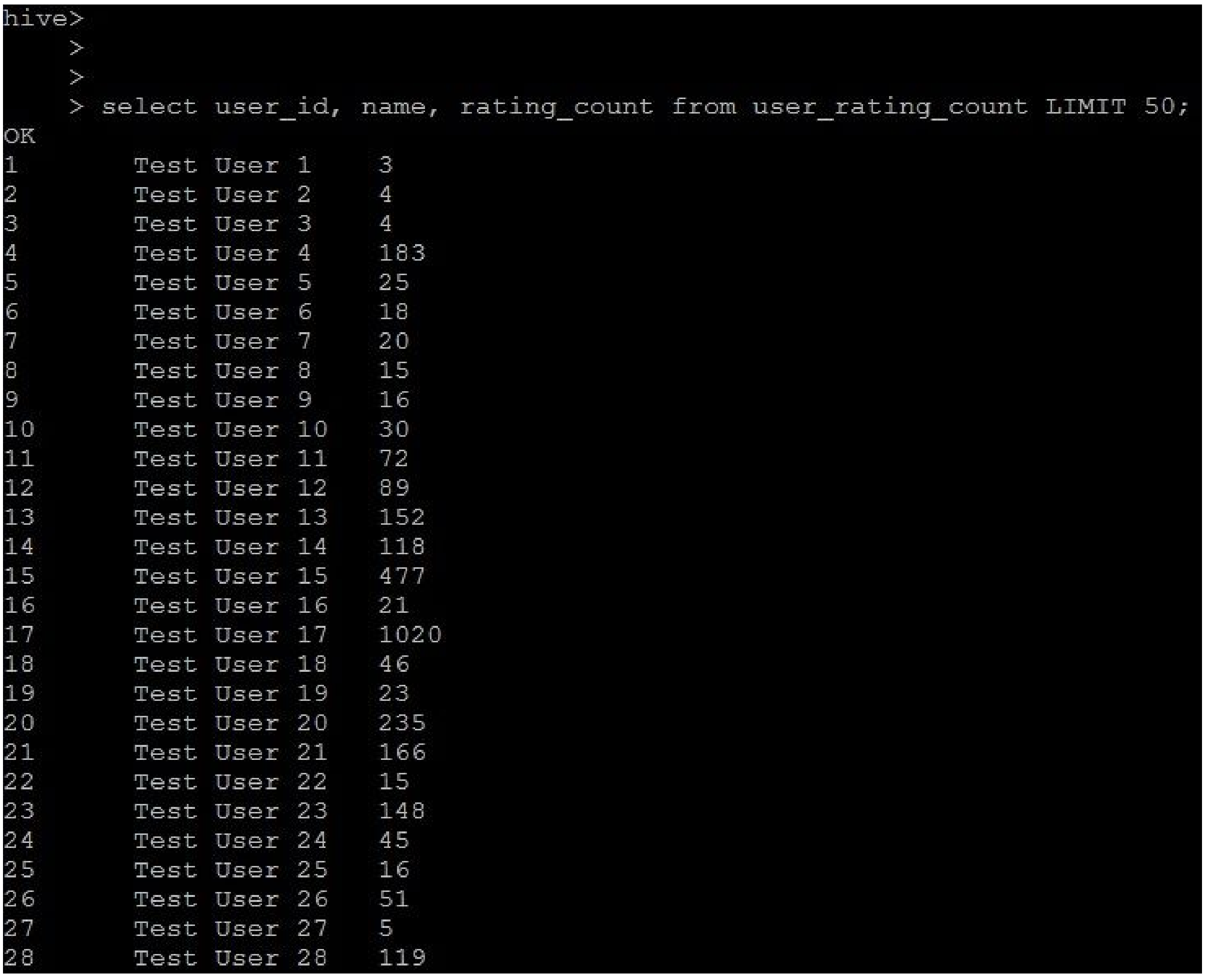


```
OK
Time taken: 0.254 seconds
hive> INSERT OVERWRITE TABLE mov_rating_count
    >  SELECT movie_id, title, count(*)
    >  FROM movies
    >  RIGHT OUTER JOIN ratings
    >  ON movies.movie_id=ratings.movie_id
    >  GROUP BY movies.movie_id, title;
Query ID = hive_20160714164950_f49bbcb7-8438-4dd6-a626-7c2db4d85930
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468446400470_0020)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   SUCCEEDED      9          9        0        0       0       0
Reducer 3 ......   SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 29.90 s
--------------------------------------------------------------------------------
Loading data to table movie_lens_data.mov_rating_count
Table movie_lens_data.mov_rating_count stats: [numFiles=2, numRows=33670, totalSize=1134891, rawDataSize=1101221]
OK
Time taken: 31.771 seconds
hive>
    >
    > select * from mov_rating_count LIMIT 100;
OK
2       Jumanji (1995)  23950
3       Grumpier Old Men (1995) 15267
5       Father of the Bride Part II (1995)      14769
6       Heat (1995)     26593
10      GoldenEye (1995)        31357
13      Balto (1995)    1648
14      Nixon (1995)    6750
18      Four Rooms (1995)       5781
19      Ace Ventura: When Nature Calls (1995)   22877
23      Assassins (1995)        4636
24      Powder (1995)   8952
27      Now and Then (1995)     1787
28      Persuasion (1995)       3334
30      Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)    1343
32      Twelve Monkeys (a.k.a. 12 Monkeys) (1995)       50380
33      Wings of Courage (1995) 70
35      Carrington (1995)       1554
```

Use Case #2: List all the users and the number of ratings they have done for a movie

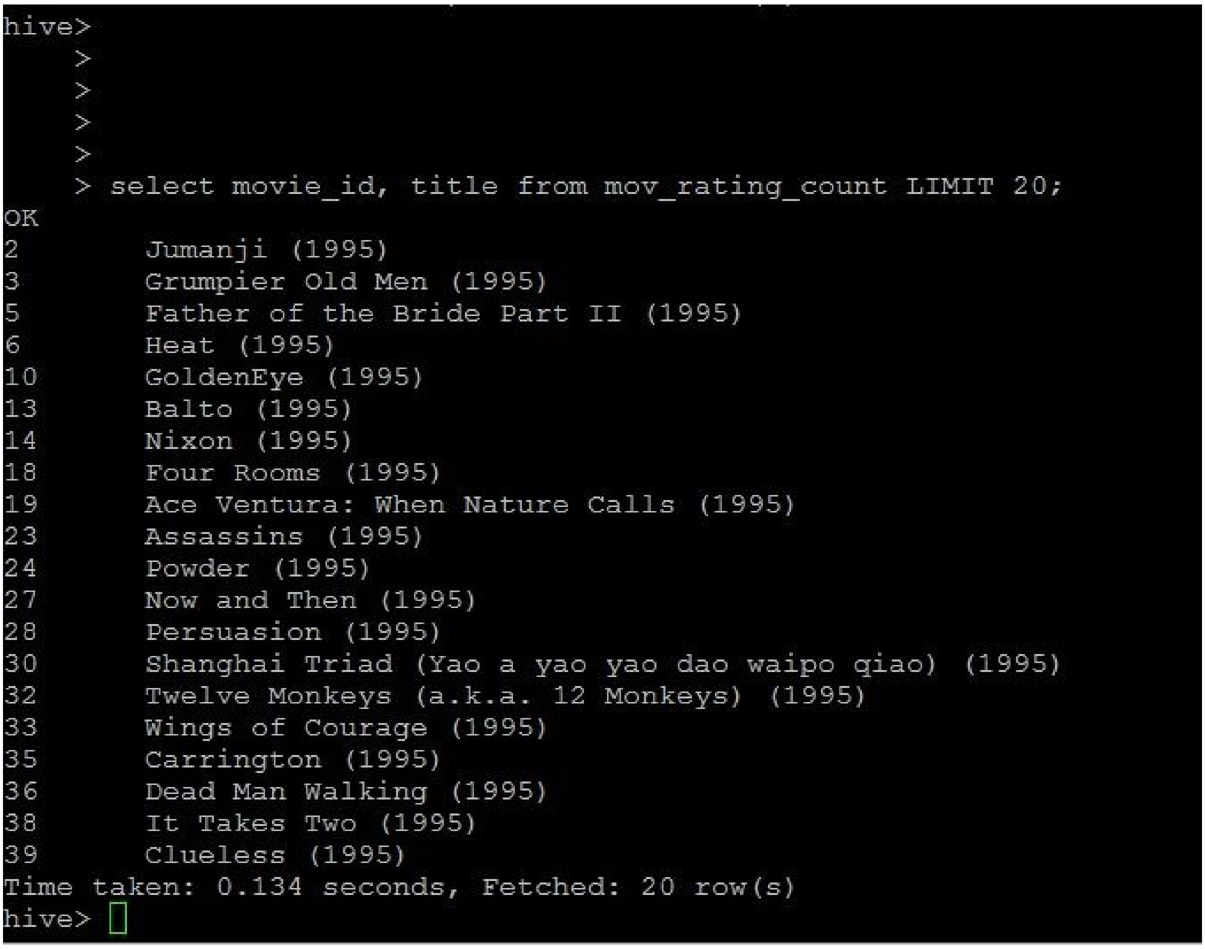


```
hive>
    >
    >
    > CREATE TABLE IF NOT EXISTS user_rating_count ( user_id bigint, name String, rating_count int) COMMENT 'Details how many movies a user rated.' ROW FORMAT DELIMITED FIELDS TERMINATED
BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.448 seconds
hive> show tables;
OK
mov_rating_count
movies
ratings
user_rating_count
users
Time taken: 0.228 seconds, Fetched: 5 row(s)
hive> desc user_rating_count;
OK
user_id                 bigint
name                    string
rating_count            int
Time taken: 0.295 seconds, Fetched: 3 row(s)
hive>
    > INSERT OVERWRITE TABLE user_rating_count
    >  SELECT u.user_id, u.name, COUNT(r.rating)
    >  FROM users u, ratings r WHERE u.user_id=r.user_id
    >  GROUP BY u.user_id, u.name;
Query ID = hive_20160714171826_9c524bc0-d12f-41ae-8ad6-ccf34c0ac77e
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1468446400470_0021)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   SUCCEEDED      9          9        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 33.72 s
--------------------------------------------------------------------------------
Loading data to table movie_lens_data.user_rating_count
Table movie_lens_data.user_rating_count stats: [numFiles=1, numRows=247753, totalSize=6493497, rawDataSize=6245744]
OK
Time taken: 44.431 seconds
hive>
```



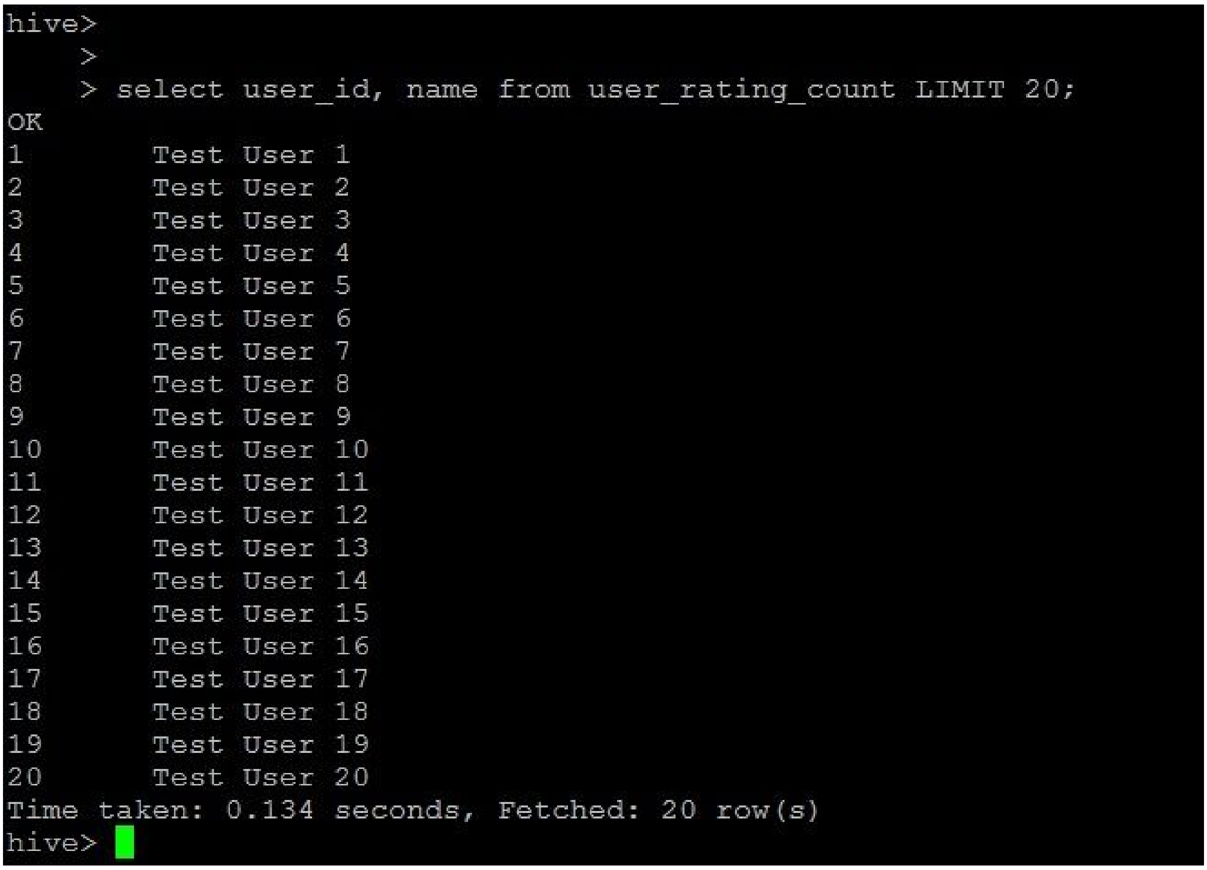
```
hive>
    >
    >
    > select user_id, name, rating_count from user_rating_count LIMIT 50;
OK
1       Test User 1     3
2       Test User 2     4
3       Test User 3     4
4       Test User 4     183
5       Test User 5     25
6       Test User 6     18
7       Test User 7     20
8       Test User 8     15
9       Test User 9     16
10      Test User 10    30
11      Test User 11    72
12      Test User 12    89
13      Test User 13    152
14      Test User 14    118
15      Test User 15    477
16      Test User 16    21
17      Test User 17    1020
18      Test User 18    46
19      Test User 19    23
20      Test User 20    235
21      Test User 21    166
22      Test User 22    15
23      Test User 23    148
24      Test User 24    45
25      Test User 25    16
26      Test User 26    51
27      Test User 27    5
28      Test User 28    119
```

Use Case #3: List all the Movie IDs which have been rated (Movie Id with atleast one user rating it)



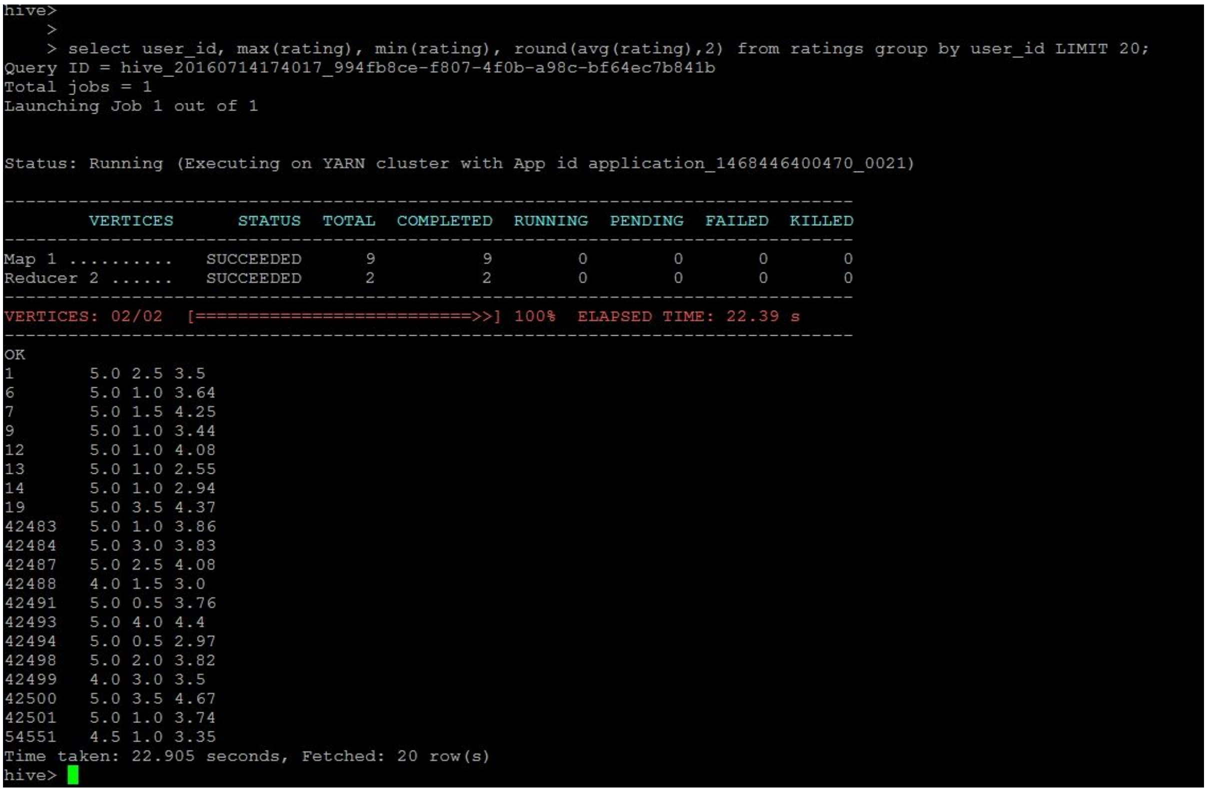
```
hive>
    >
    >
    >
    >
    > select movie_id, title from mov_rating_count LIMIT 20;
OK
2       Jumanji (1995)
3       Grumpier Old Men (1995)
5       Father of the Bride Part II (1995)
6       Heat (1995)
10      GoldenEye (1995)
13      Balto (1995)
14      Nixon (1995)
18      Four Rooms (1995)
19      Ace Ventura: When Nature Calls (1995)
23      Assassins (1995)
24      Powder (1995)
27      Now and Then (1995)
28      Persuasion (1995)
30      Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)
32      Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
33      Wings of Courage (1995)
35      Carrington (1995)
36      Dead Man Walking (1995)
38      It Takes Two (1995)
39      Clueless (1995)
Time taken: 0.134 seconds, Fetched: 20 row(s)
hive> 
```

Use Case #4: List all the Users who have rated the movies (Users who have rated atleast one movie)



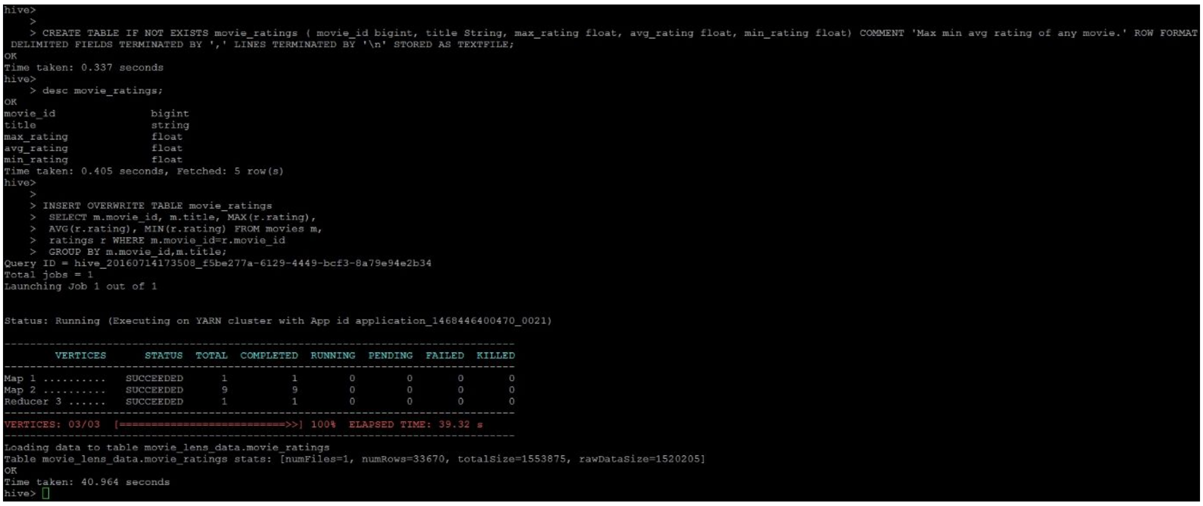
```
hive>
    >
    > select user_id, name from user_rating_count LIMIT 20;
OK
1       Test User 1
2       Test User 2
3       Test User 3
4       Test User 4
5       Test User 5
6       Test User 6
7       Test User 7
8       Test User 8
9       Test User 9
10      Test User 10
11      Test User 11
12      Test User 12
13      Test User 13
14      Test User 14
15      Test User 15
16      Test User 16
17      Test User 17
18      Test User 18
19      Test User 19
20      Test User 20
Time taken: 0.134 seconds, Fetched: 20 row(s)
hive>
```

Use Case #5: List of all the User with the max,min,average ratings they have given against any movie

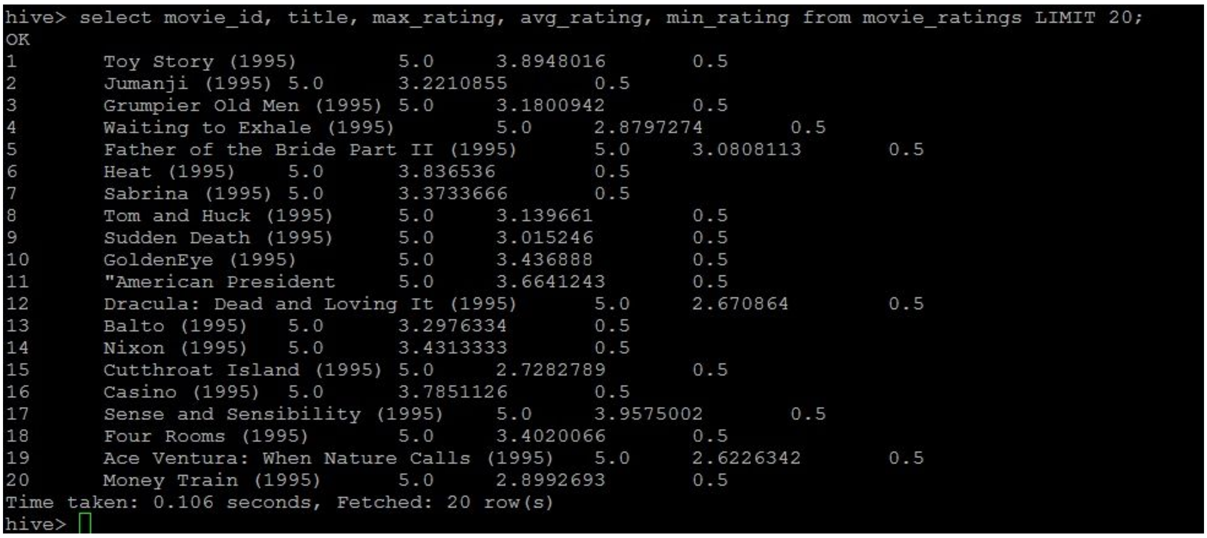


```
hive>
    >
    > select user_id, max(rating), min(rating), round(avg(rating),2) from ratings group by user_id LIMIT 20;
Query ID = hive_20160714174017_994fb8ce-f807-4f0b-a98c-bf64ec7b841b
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468446400470_0021)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      9          9        0        0       0       0
Reducer 2 ......   SUCCEEDED      2          2        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 22.39 s
--------------------------------------------------------------------------------
OK
1       5.0 2.5 3.5
6       5.0 1.0 3.64
7       5.0 1.5 4.25
9       5.0 1.0 3.44
12      5.0 1.0 4.08
13      5.0 1.0 2.55
14      5.0 1.0 2.94
19      5.0 3.5 4.37
42483   5.0 1.0 3.86
42484   5.0 3.0 3.83
42487   5.0 2.5 4.08
42488   4.0 1.5 3.0
42491   5.0 0.5 3.76
42493   5.0 4.0 4.4
42494   5.0 0.5 2.97
42498   5.0 2.0 3.82
42499   4.0 3.0 3.5
42500   5.0 3.5 4.67
42501   5.0 1.0 3.74
54551   4.5 1.0 3.35
Time taken: 22.905 seconds, Fetched: 20 row(s)
hive>
```

Use Case #6: List all the Movies with the max, min, average ratings given by any user



```
hive>
    >
    > CREATE TABLE IF NOT EXISTS movie_ratings ( movie_id bigint, title String, max_rating float, avg_rating float, min_rating float) COMMENT 'Max min avg rating of any movie.' ROW FORMAT
 DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.337 seconds
hive>
    > desc movie_ratings;
OK
movie_id                bigint
title                   string
max_rating              float
avg_rating              float
min_rating              float
Time taken: 0.405 seconds, Fetched: 5 row(s)
hive>
    >
    > INSERT OVERWRITE TABLE movie_ratings
    >  SELECT m.movie_id, m.title, MAX(r.rating),
    >  AVG(r.rating), MIN(r.rating) FROM movies m,
    >  ratings r WHERE m.movie_id=r.movie_id
    >  GROUP BY m.movie_id,m.title;
Query ID = hive_20160714173508_f5be277a-6129-4449-bcf3-8a79e94e2b34
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468446400470_0021)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   SUCCEEDED      9          9        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 39.32 s
--------------------------------------------------------------------------------
Loading data to table movie_lens_data.movie_ratings
Table movie_lens_data.movie_ratings stats: [numFiles=1, numRows=33670, totalSize=1553875, rawDataSize=1520205]
OK
Time taken: 40.964 seconds
hive>
```



```
hive> select movie_id, title, max_rating, avg_rating, min_rating from movie_ratings LIMIT 20;
OK
1       Toy Story (1995)        5.0     3.8948016       0.5
2       Jumanji (1995) 5.0      3.2210855       0.5
3       Grumpier Old Men (1995) 5.0     3.1800942       0.5
4       Waiting to Exhale (1995)        5.0     2.8797274       0.5
5       Father of the Bride Part II (1995)      5.0     3.0808113       0.5
6       Heat (1995)    5.0      3.836536        0.5
7       Sabrina (1995) 5.0      3.3733666       0.5
8       Tom and Huck (1995)     5.0     3.139661        0.5
9       Sudden Death (1995)     5.0     3.015246        0.5
10      GoldenEye (1995)        5.0     3.436888        0.5
11      "American President     5.0     3.6641243       0.5
12      Dracula: Dead and Loving It (1995)      5.0     2.670864        0.5
13      Balto (1995)   5.0      3.2976334       0.5
14      Nixon (1995)   5.0      3.4313333       0.5
15      Cutthroat Island (1995) 5.0     2.7282789       0.5
16      Casino (1995)  5.0      3.7851126       0.5
17      Sense and Sensibility (1995)    5.0     3.9575002       0.5
18      Four Rooms (1995)       5.0     3.4020066       0.5
19      Ace Ventura: When Nature Calls (1995)   5.0     2.6226342       0.5
20      Money Train (1995)      5.0     2.8992693       0.5
Time taken: 0.106 seconds, Fetched: 20 row(s)
hive>
```