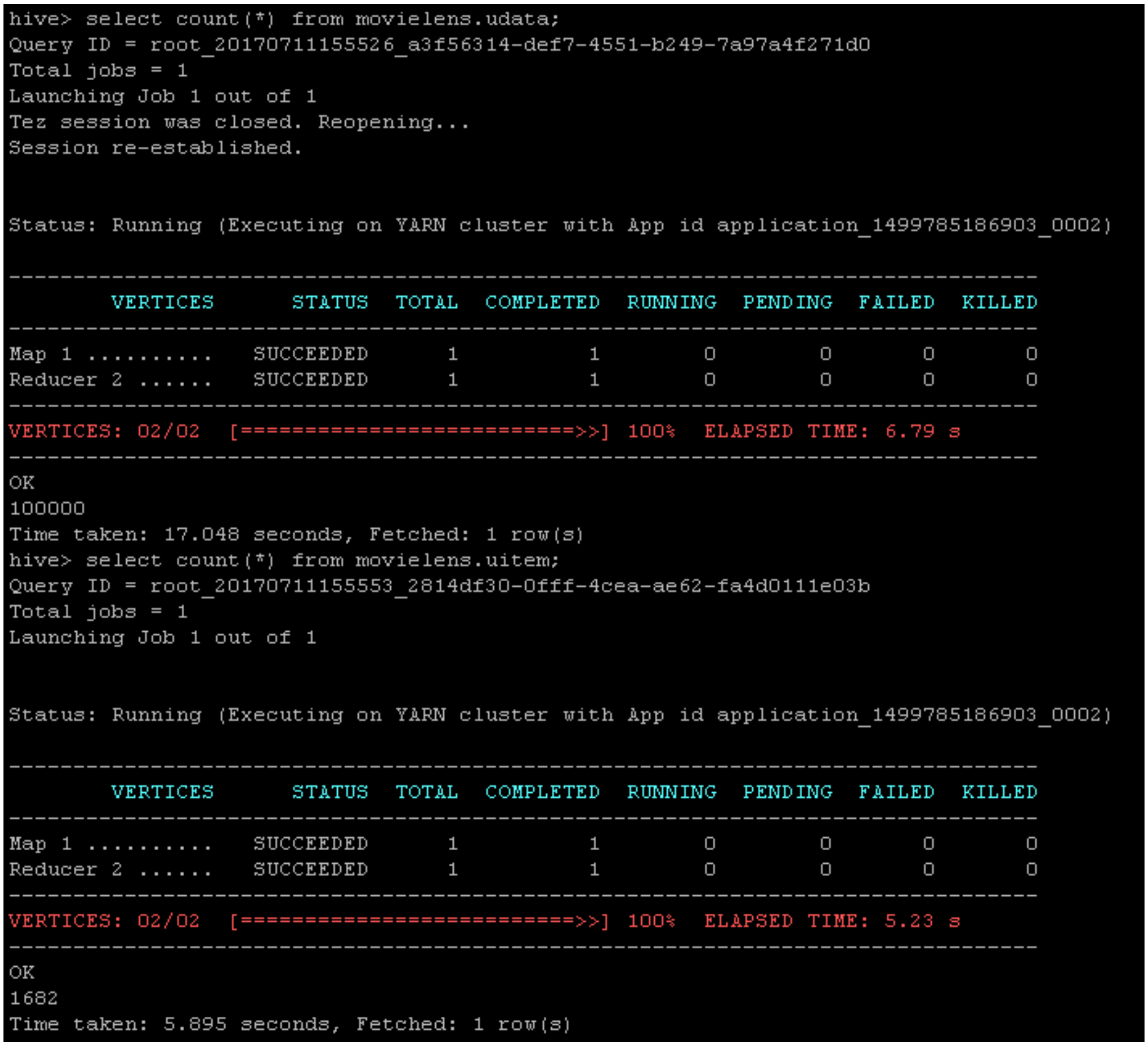
Assignment 02

1. How many records are there in both tables? Please specify separately for each table.

Answer:

```
select count(*) from movielens.udata;
select count(*) from movielens.uitem;
```

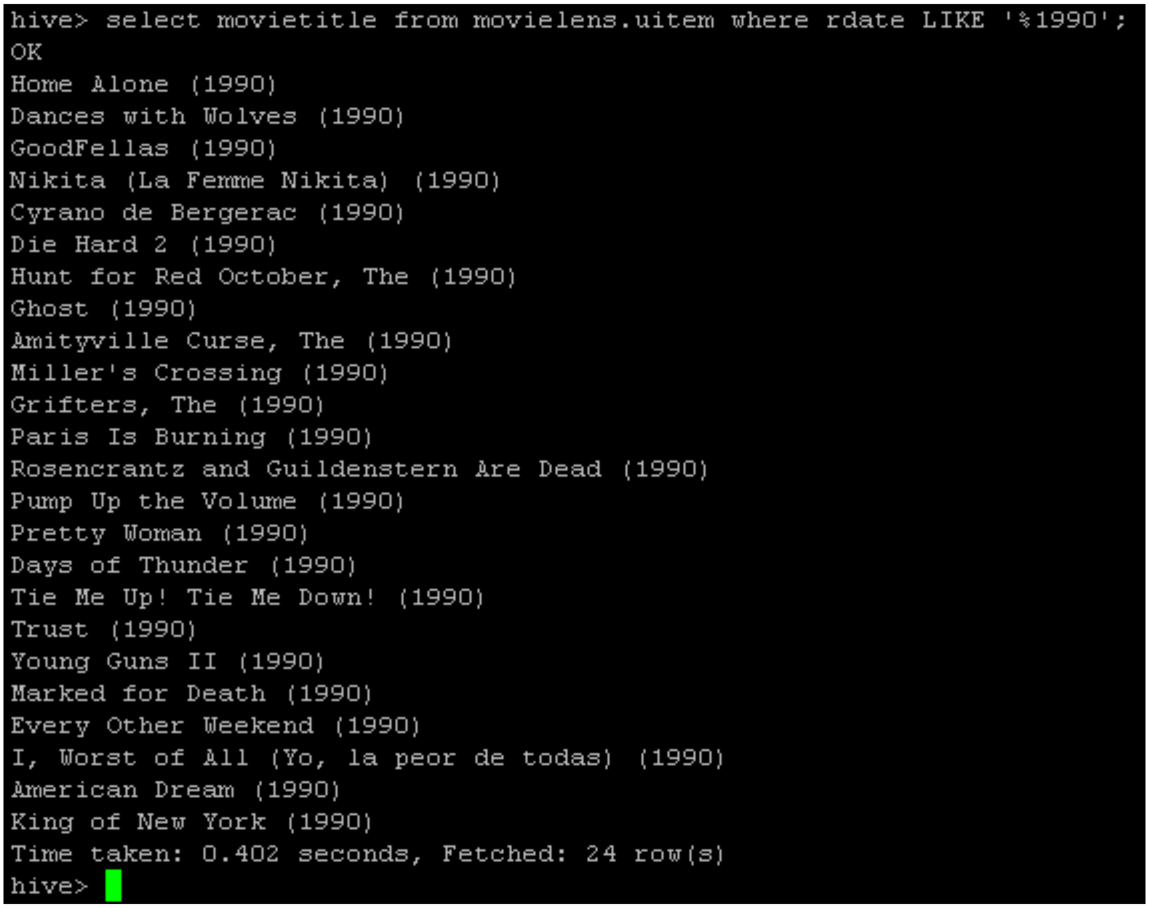


```
hive> select count(*) from movielens.udata;
Query ID = root_20170711155526_a3f56314-def7-4551-b249-7a97a4f271d0
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1499785186903_0002)

--------------------------------------------------------------------------------
        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: 6.79 s
--------------------------------------------------------------------------------
OK
100000
Time taken: 17.048 seconds, Fetched: 1 row(s)
hive> select count(*) from movielens.uitem;
Query ID = root_20170711155553_2814df30-0fff-4cea-ae62-fa4d0111e03b
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1499785186903_0002)

--------------------------------------------------------------------------------
        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: 5.23 s
--------------------------------------------------------------------------------
OK
1682
Time taken: 5.895 seconds, Fetched: 1 row(s)
```

2. Find the name of all movies released in 1990.

Answer:

```
select movietitle from movielens.uitem where rdate LIKE '%1990'
```



```
hive> select movietitle from movielens.uitem where rdate LIKE '%1990';
OK
Home Alone (1990)
Dances with Wolves (1990)
GoodFellas (1990)
Nikita (La Femme Nikita) (1990)
Cyrano de Bergerac (1990)
Die Hard 2 (1990)
Hunt for Red October, The (1990)
Ghost (1990)
Amityville Curse, The (1990)
Miller's Crossing (1990)
Grifters, The (1990)
Paris Is Burning (1990)
Rosencrantz and Guildenstern Are Dead (1990)
Pump Up the Volume (1990)
Pretty Woman (1990)
Days of Thunder (1990)
Tie Me Up! Tie Me Down! (1990)
Trust (1990)
Young Guns II (1990)
Marked for Death (1990)
Every Other Weekend (1990)
I, Worst of All (Yo, la peor de todas) (1990)
American Dream (1990)
King of New York (1990)
Time taken: 0.402 seconds, Fetched: 24 row(s)
hive>
```

3. List the movieid of the 10 films that received the most ratings (not necessarily highest rating) in the table you created from u.data.

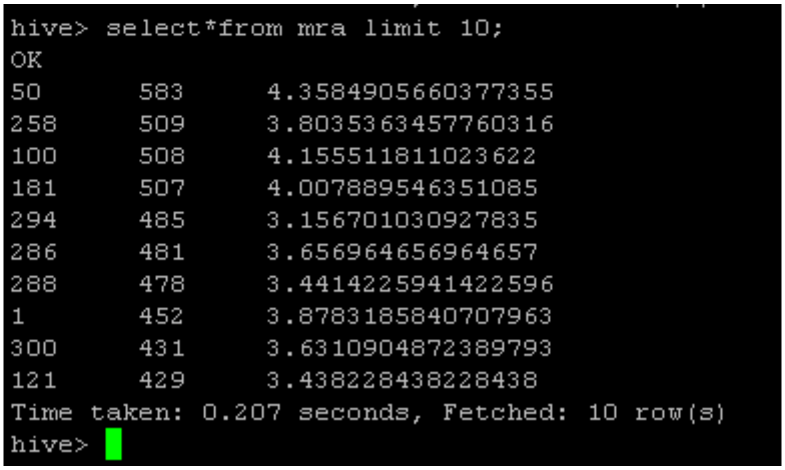
Answer:

First: Create a new table from udatatable

```
create table mra as select movieid, count(rating) as a, avg(rating) as b from udata
group by movieid sort by a desc  limit 10 ;
```

Second: Retrieve the list the 10 movies using

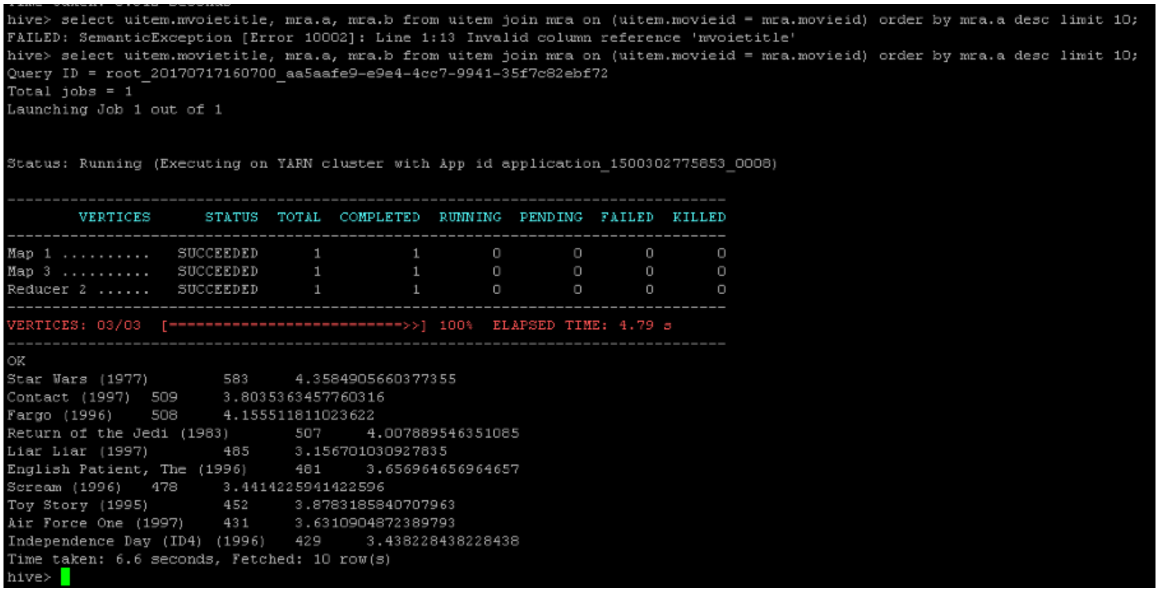
```
select *from mra limit 10;
```



```
hive> select*from mra limit 10;
OK
50      583     4.3584905660377355
258     509     3.803536345776016
100     508     4.155511811023622
181     507     4.007889546351085
294     485     3.156701030927835
286     481     3.656964656964657
288     478     3.4414225941422596
1       452     3.8783185840707963
300     431     3.6310904872389793
121     429     3.438228438228438
Time taken: 0.207 seconds, Fetched: 10 row(s)
hive> 
```

4. Use a join to list the titles of the movies you found in step Answer

select uitem.mvoietitle, mra.a, mra.b from uitem join mra on (uitem.movieid = mra.movieid) order by mra.a desc limit 10;

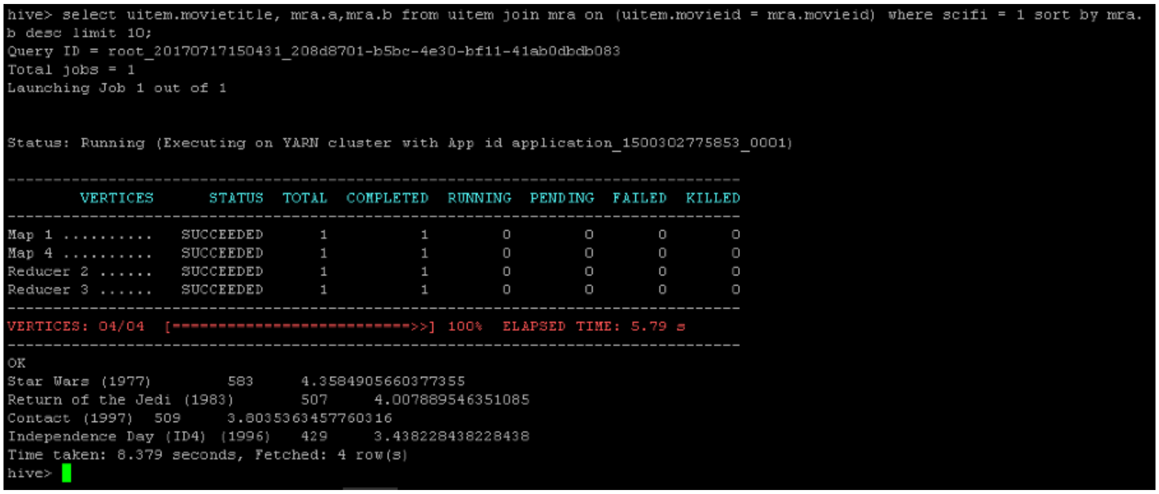


```
hive> select uitem.mvoietitle, mra.a, mra.b from uitem join mra on (uitem.movieid = mra.movieid) order by mra.a desc limit 10;
FAILED: SemanticException [Error 10002]: Line 1:13 Invalid column reference 'mvoietitle'
hive> select uitem.movietitle, mra.a, mra.b from uitem join mra on (uitem.movieid = mra.movieid) order by mra.a desc limit 10;
Query ID = root_20170717160700_aa5aafe9-e9e4-4cc7-9941-35f7c82ebf72
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1500302775853_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 3 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 4.79 s
--------------------------------------------------------------------------------
OK
Star Wars (1977)        583     4.3584905660377355
Contact (1997)  509     3.803536345776016
Fargo (1996)    508     4.155511811023622
Return of the Jedi (1983)       507     4.007889546351085
Liar Liar (1997)        485     3.156701030927835
English Patient, The (1996)     481     3.656964656964657
Scream (1996)   478     3.4414225941422596
Toy Story (1995)        452     3.8783185840707963
Air Force One (1997)    431     3.6310904872389793
Independence Day (ID4) (1996)   429     3.438228438228438
Time taken: 6.6 seconds, Fetched: 10 row(s)
hive> 
```

5. Find the highest rated sci_fi movie. Explain how you define "highest rating”.

Answer: Highest rating is defined as the highest value in the ratings column. So a movie with highest value in the rating column with be highest rated

select uitem.movietitle, mra.a ,mra.b from uitem join mra on (uitem.movieid = mra.movieid) where scifi = 1 sort by mra.b desc limit 10;



```
hive> select uitem.movietitle, mra.a,mra.b from uitem join mra on (uitem.movieid = mra.movieid) where scifi = 1 sort by mra.
b desc limit 10;
Query ID = root_20170717150431_208d8701-b5bc-4e30-bf11-41ab0dbdb083
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1500302775853_0001)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 4 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 5.79 s
--------------------------------------------------------------------------------
OK
Star Wars (1977)        583     4.3584905660377355
Return of the Jedi (1983)       507     4.007889546351085
Contact (1997)  509     3.8035363457760316
Independence Day (ID4) (1996)   429     3.438228438228438
Time taken: 8.379 seconds, Fetched: 4 row(s)
hive>
```