Project - Hive

 # The following analysis is done using the u.data and u.item files from the folder movielens.zip from http://grouplens.org/datasets/movielens/

Summary: from the datasets in the u.data and u.item files, the data has been filtered to yield certain results - finding the number of records; filtering the data to show movies from a certain year; filtering the data to display only movies with the highest ratings; expanding the filter to show the movie titles with the corresponding year (from the previous query); searching only for sci-fi movies with more than 250 ratings; checking whether there were any movies with no ratings.

 # Finding the number of records in both the files (u.data and u.item):

In the table u.data: there are 100,000 records (as shown below)

```
hive> select count (*) from udata;
Query ID = root_20160713165511_fcc15435-4c39-4470-bcce-0ba12cd18ed6
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_1468423699620_0004)

--------------------------------------------------------------------------------
        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: 5.13 s
--------------------------------------------------------------------------------
OK
100000
Time taken: 22.707 seconds, Fetched: 1 row(s)
```

In the table: u.item: there are 1,682 records (as shown below)

```
hive> select count (*) from uitem;
Query ID = root_20160713165613_89af501f-d228-43da-886c-f76db8de2cfd
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468423699620_0004)

--------------------------------------------------------------------------------
        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: 3.68 s
--------------------------------------------------------------------------------
OK
1682
Time taken: 5.881 seconds, Fetched: 1 row(s)
```

 # Below is the script and result - to find a list of movies that were released in the year 1990 from the u.item dataset:

```
hive> select movie_title from uitem where substr(release_date,8) = "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: 1.487 seconds, Fetched: 24 row(s)
```

 # Below is the script and the result - to display a list of the top 10 movies that have received the most ratings from the u.data file:

```
hive> select movieid, count(rating) as cnt from udata group by movieid order by cnt desc limit 10;
Query ID = root_20160713183236_bfb6feb6-c5a0-4cb1-91e3-311e5703f48a
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_1468423699620_0008)

--------------------------------------------------------------------------------
        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
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.34 s
--------------------------------------------------------------------------------
OK
50      583
258     509
100     508
181     507
294     485
286     481
288     478
1       452
300     431
121     429
Time taken: 13.405 seconds, Fetched: 10 row(s)
```

 # The following shows the script and the result - which includes joining the title of the top 10 movies with the most ratings (found above):

```
hive> select ui.movie_title, count(rating) as cnt from udata as ud JOIN uitem as ui ON ud.movieid=ui.movie_id group by ud.movieid,ui.movie_title order by cnt desc limit 10;
Query ID = root_20160713190552_d0889f1b-6461-45c5-9112-707518fe6a49
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_1468423699620_0009)

--------------------------------------------------------------------------------
        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: 7.37 s
--------------------------------------------------------------------------------
OK
Star Wars (1977)        583
Contact (1997)  509
Fargo (1996)    508
Return of the Jedi (1983)       507
Liar Liar (1997)        485
English Patient, The (1996)     481
Scream (1996)   478
Toy Story (1995)        452
Air Force One (1997)    431
Independence Day (ID4) (1996)   429
Time taken: 16.131 seconds, Fetched: 10 row(s)
```

 # Below is the script and result – to finding the top 10 highest rate sci-fi movies. There is a filter applied where movie titles are only shown where they have more than 250 ratings. Also, the highest rated movie is when the average rating is the highest for each movie.

```
hive> select ui.movie_title, count(ud.rating) as cnt, avg(ud.rating) as avg1 from udata as ud JOIN uitem as ui ON ud.movieid=ui.movie_id where ui.scifi="1" group by ud.movieid,ui.movie_titl
e having cnt>250 order by avg1 desc limit 10;
Query ID = root_20160713195708_28ca5cec-0452-48df-b05c-283d8268d758
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_1468423699620_0012)

--------------------------------------------------------------------------------
        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: 7.15 s
--------------------------------------------------------------------------------
OK
Star Wars (1977)        583     4.3584905660377355
Empire Strikes Back, The (1980) 367     4.204359673024523
Blade Runner (1982)     275     4.138181818181818
Alien (1979)    291     4.034364261168385
Return of the Jedi (1983)       507     4.007889546351085
Terminator 2: Judgment Day (1991)       295     4.0067796610169495
2001: A Space Odyssey (1968)    259     3.969111969111969
Aliens (1986)   284     3.9471830985915495
Terminator, The (1984)  301     3.9335548172757475
Back to the Future (1985)       350     3.834285714285714
Time taken: 16.408 seconds, Fetched: 10 row(s)
```

 # Below is the script to check whether there are any movies with no ratings:

```
hive> select count(ui.movie_id) from uitem as ui full outer join udata as ud on ud.movieid=ui.movie_id where ud.rating is null;
Query ID = root_20160713201159_e3d1bc41-87dd-4d0a-9ddc-8d1ed49bbf97
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1468423699620_0013)

--------------------------------------------------------------------------------
        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: 6.71 s
--------------------------------------------------------------------------------
OK
0
Time taken: 9.851 seconds, Fetched: 1 row(s)
```