Movie data Set Analysis using Pig

C.V.Raman Global University
Bhubaneswar,Odisha

Problem Statement: Find the movie with avg rating >4.0 from u.data dataset.

Step 1: Copy the files form local system Download folder to root directory:

```
[root@quickstart cloudera]#  hdfs dfs -copyFromLocal
/home/cloudera/Downloads/u.data  /
```

```
[root@quickstart cloudera]#  hdfs dfs -copyFromLocal
/home/cloudera/Downloads/u.item  /
```

And to display do

```
[root@quickstart cloudera]#  hdfs dfs -ls /
```

Step 2. Load the movie data into pig

Note: Now, open the grunt shell of Apache pig

```
ratings = LOAD '/u.data' AS (userID:int, movieID:int, rating:int,
ratingTime:int);
```

```
dump ratings
```

Step 3: Find the movie with avg rating >4.0 from u.data dataset.

To find we follow the following steps

Step 3a) Group the data according to movieID:

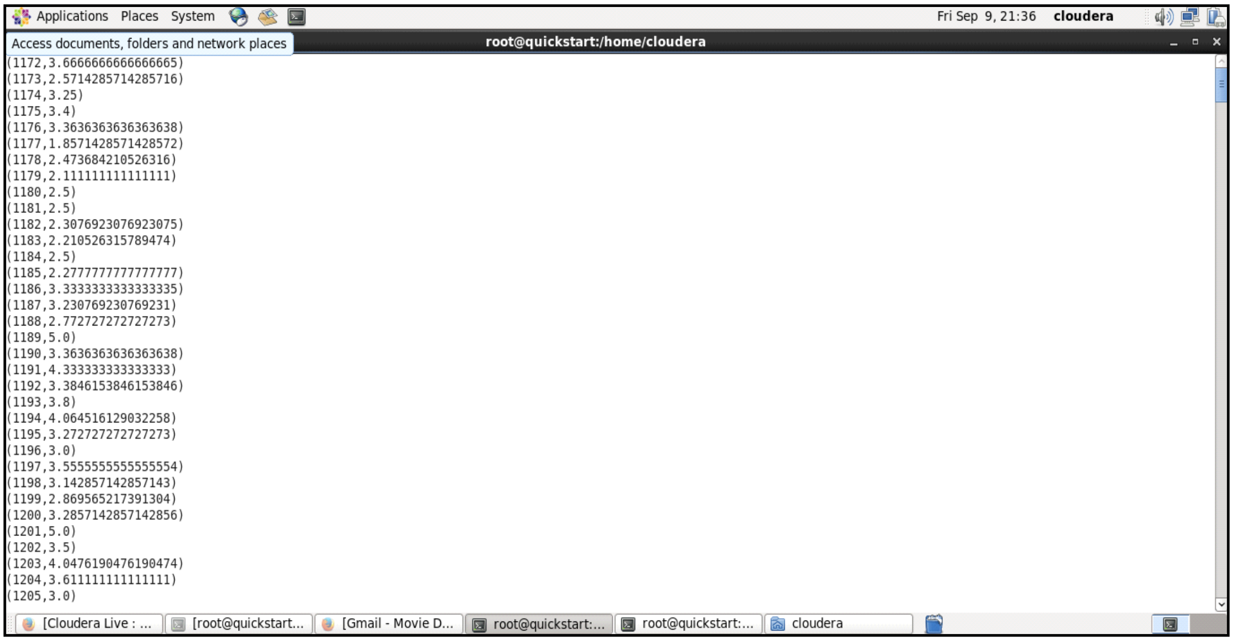
ratingsByMovie = GROUP ratings BY movieID;

dump ratingsByMovie;

Step 4) Find average ratings for the grouped data:

avgRatings = FOREACH ratingsByMovie GENERATE group AS movieID, AVG(ratings.rating) AS avgRating;

dump avgRatings;



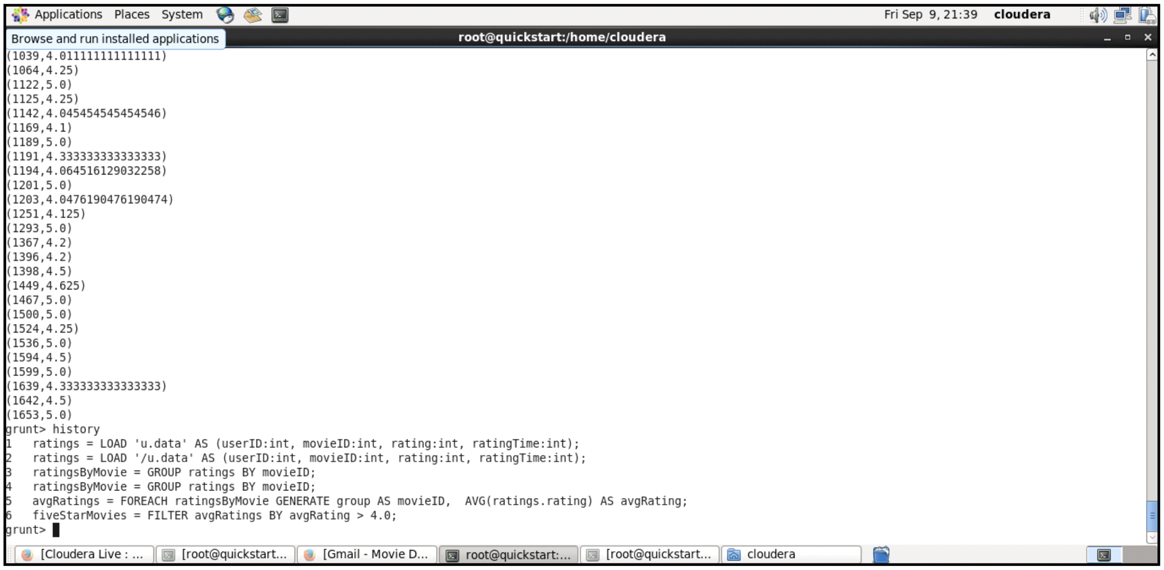
```
(1172,3.6666666666666665)
(1173,2.5714285714285716)
(1174,3.25)
(1175,3.4)
(1176,3.3636363636363638)
(1177,1.8571428571428572)
(1178,2.473684210526316)
(1179,2.111111111111111)
(1180,2.5)
(1181,2.5)
(1182,2.3076923076923075)
(1183,2.210526315789474)
(1184,2.5)
(1185,2.2777777777777777)
(1186,3.3333333333333335)
(1187,3.230769230769231)
(1188,2.772727272727273)
(1189,5.0)
(1190,3.3636363636363638)
(1191,4.333333333333333)
(1192,3.3846153846153846)
(1193,3.8)
(1194,4.064516129032258)
(1195,3.272727272727273)
(1196,3.0)
(1197,3.5555555555555554)
(1198,3.142857142857143)
(1199,2.869565217391304)
(1200,3.2857142857142856)
(1201,5.0)
(1202,3.5)
(1203,4.0476190476190474)
(1204,3.611111111111111)
(1205,3.0)
```

Step 5: Filter the required results:

fiveStarMovies = FILTER avgRatings BY avgRating > 4.0;

To print execute:

dump fiveStarMovies;



```
(1039,4.011111111111111)
(1064,4.25)
(1122,5.0)
(1125,4.25)
(1142,4.045454545454546)
(1169,4.1)
(1189,5.0)
(1191,4.333333333333333)
(1194,4.064516129032258)
(1201,5.0)
(1203,4.0476190476190474)
(1251,4.125)
(1293,5.0)
(1367,4.2)
(1396,4.2)
(1398,4.5)
(1449,4.625)
(1467,5.0)
(1500,5.0)
(1524,4.25)
(1536,5.0)
(1594,4.5)
(1599,5.0)
(1639,4.333333333333333)
(1642,4.5)
(1653,5.0)
grunt> history
1   ratings = LOAD 'u.data' AS (userID:int, movieID:int, rating:int, ratingTime:int);
2   ratings = LOAD '/u.data' AS (userID:int, movieID:int, rating:int, ratingTime:int);
3   ratingsByMovie = GROUP ratings BY movieID;
4   ratingsByMovie = GROUP ratings BY movieID;
5   avgRatings = FOREACH ratingsByMovie GENERATE group AS movieID,  AVG(ratings.rating) AS avgRating;
6   fiveStarMovies = FILTER avgRatings BY avgRating > 4.0;
grunt> 
```

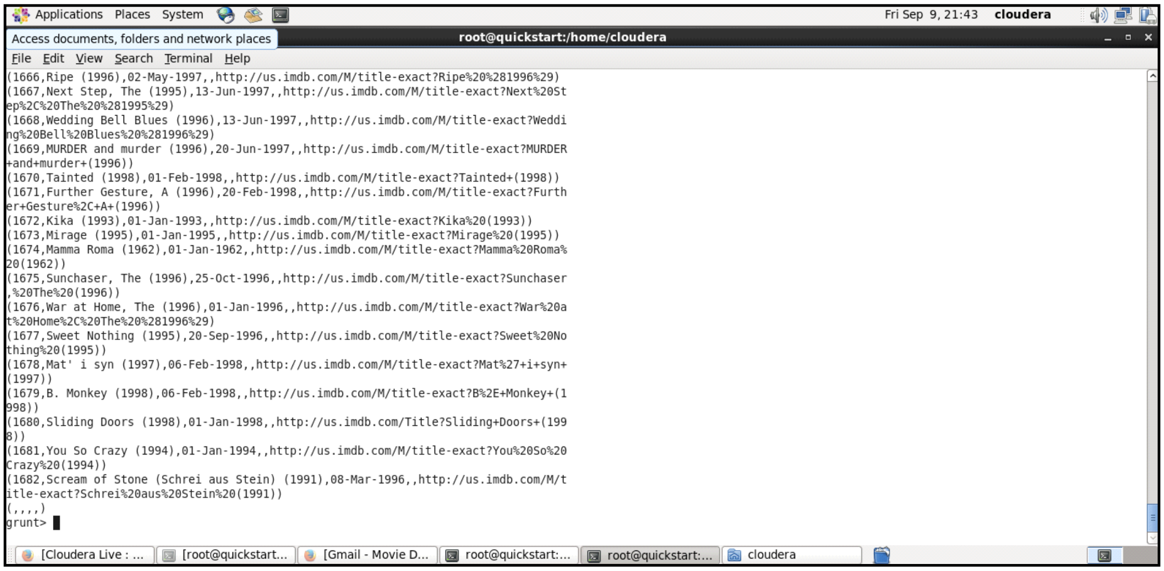
………
………………

Problem 2: Find the oldest 5-star movies from u.data and u.item datasets.

- Load the u.item in pig:

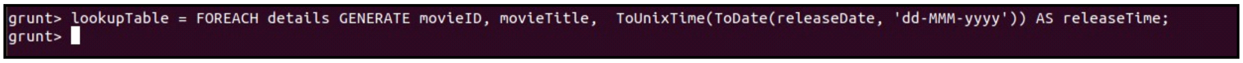
  details = LOAD '/u.item' USING PigStorage('|') AS (movieID:int, movieTitle:chararray, releaseDate:chararray, videoRelease:chararray, imdbLink:chararray);

  dump details


Applications Places System
Fri Sep 9, 21:43 cloudera
Access documents, folders and network places
root@quickstart:/home/cloudera
File Edit View Search Terminal Help
[Cloudera Live : ...
[root@quickstart...
[Gmail - Movie D...
root@quickstart:...
root@quickstart:...
cloudera

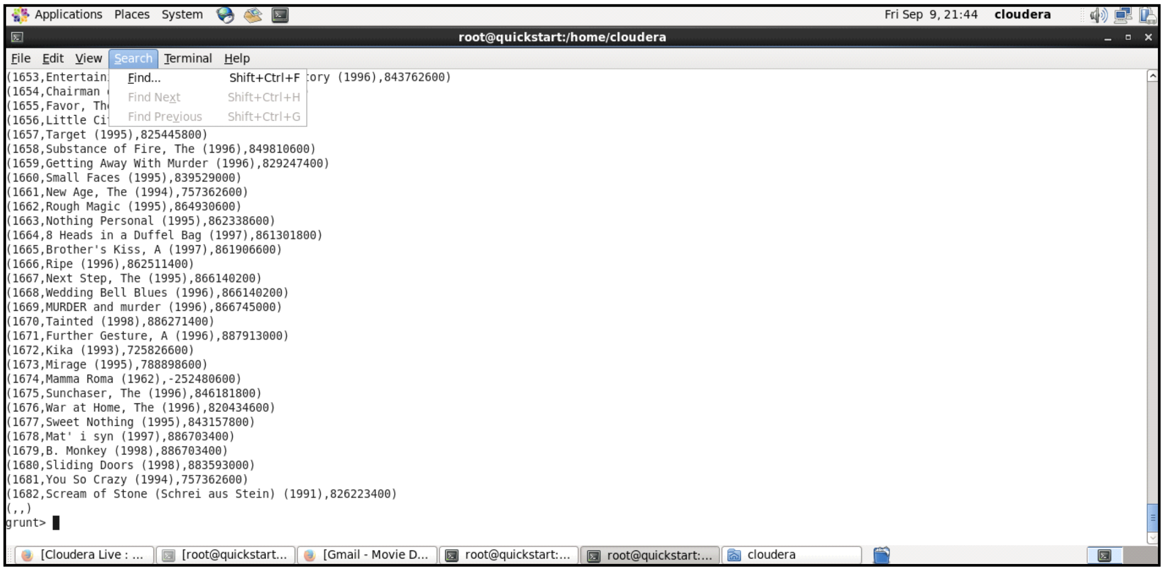
```
(1666,Ripe (1996),02-May-1997,,http://us.imdb.com/M/title-exact?Ripe%20%281996%29)
(1667,Next Step, The (1995),13-Jun-1997,,http://us.imdb.com/M/title-exact?Next%20St
ep%2C%20The%20%281995%29)
(1668,Wedding Bell Blues (1996),13-Jun-1997,,http://us.imdb.com/M/title-exact?Weddi
ng%20Bell%20Blues%20%281996%29)
(1669,MURDER and murder (1996),20-Jun-1997,,http://us.imdb.com/M/title-exact?MURDER
+and+murder+(1996))
(1670,Tainted (1998),01-Feb-1998,,http://us.imdb.com/M/title-exact?Tainted+(1998))
(1671,Further Gesture, A (1996),20-Feb-1998,,http://us.imdb.com/M/title-exact?Furth
er+Gesture%2C+A+(1996))
(1672,Kika (1993),01-Jan-1993,,http://us.imdb.com/M/title-exact?Kika%20(1993))
(1673,Mirage (1995),01-Jan-1995,,http://us.imdb.com/M/title-exact?Mirage%20(1995))
(1674,Mamma Roma (1962),01-Jan-1962,,http://us.imdb.com/M/title-exact?Mamma%20Roma%
20(1962))
(1675,Sunchaser, The (1996),25-Oct-1996,,http://us.imdb.com/M/title-exact?Sunchaser
,%20The%20(1996))
(1676,War at Home, The (1996),01-Jan-1996,,http://us.imdb.com/M/title-exact?War%20a
t%20Home%2C%20The%20%281996%29)
(1677,Sweet Nothing (1995),20-Sep-1996,,http://us.imdb.com/M/title-exact?Sweet%20No
thing%20(1995))
(1678,Mat' i syn (1997),06-Feb-1998,,http://us.imdb.com/M/title-exact?Mat%27+i+syn+
(1997))
(1679,B. Monkey (1998),06-Feb-1998,,http://us.imdb.com/M/title-exact?B%2E+Monkey+(1
998))
(1680,Sliding Doors (1998),01-Jan-1998,,http://us.imdb.com/Title?Sliding+Doors+(199
8))
(1681,You So Crazy (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?You%20So%20
Crazy%20(1994))
(1682,Scream of Stone (Schrei aus Stein) (1991),08-Mar-1996,,http://us.imdb.com/M/t
itle-exact?Schrei%20aus%20Stein%20(1991))
(,,,,)
grunt> 
```

- Create a scalable timestamp column to compare the time:
  lookupTable = FOREACH details GENERATE movieID, movieTitle, ToUnixTime(ToDate(releaseDate, 'dd-MMM-yyyy')) AS releaseTime;



```
grunt> lookupTable = FOREACH details GENERATE movieID, movieTitle,  ToUnixTime(ToDate(releaseDate, 'dd-MMM-yyyy')) AS releaseTime;
grunt> 
```

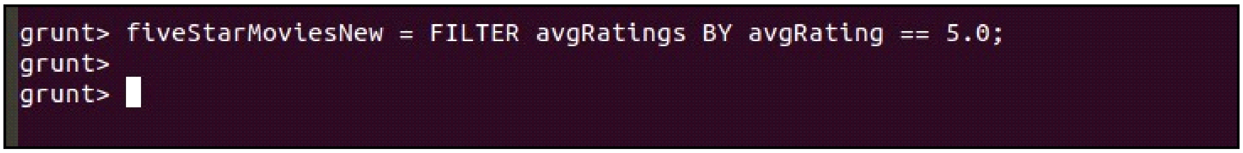
dump lookupTable;



```
(1653,Entertain                   :ory (1996),843762600)
(1654,Chairman (
(1655,Favor, Th
(1656,Little Ci
(1657,Target (1995),825445800)
(1658,Substance of Fire, The (1996),849810600)
(1659,Getting Away With Murder (1996),829247400)
(1660,Small Faces (1995),839529000)
(1661,New Age, The (1994),757362600)
(1662,Rough Magic (1995),864930600)
(1663,Nothing Personal (1995),862338600)
(1664,8 Heads in a Duffel Bag (1997),861301800)
(1665,Brother's Kiss, A (1997),861906600)
(1666,Ripe (1996),862511400)
(1667,Next Step, The (1995),866140200)
(1668,Wedding Bell Blues (1996),866140200)
(1669,MURDER and murder (1996),866745000)
(1670,Tainted (1998),886271400)
(1671,Further Gesture, A (1996),887913000)
(1672,Kika (1993),725826600)
(1673,Mirage (1995),788898600)
(1674,Mamma Roma (1962),-252480600)
(1675,Sunchaser, The (1996),846181800)
(1676,War at Home, The (1996),820434600)
(1677,Sweet Nothing (1995),843157800)
(1678,Mat' i syn (1997),886703400)
(1679,B. Monkey (1998),886703400)
(1680,Sliding Doors (1998),883593000)
(1681,You So Crazy (1994),757362600)
(1682,Scream of Stone (Schrei aus Stein) (1991),826223400)
(,,)
grunt> 
```

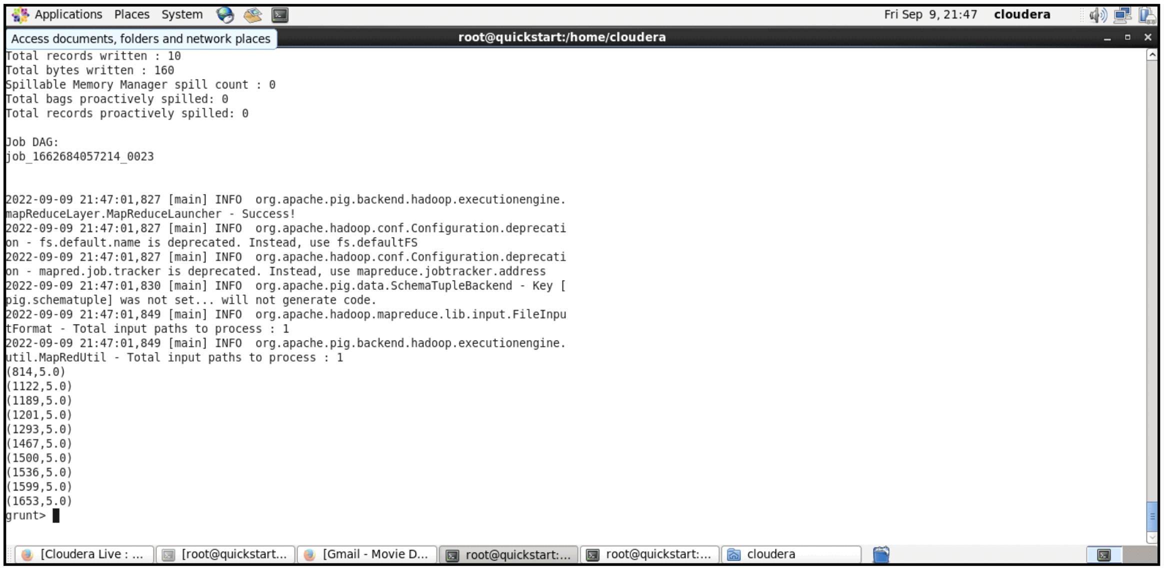
- Filter the movies with average rating as 5:

  fiveStarMoviesNew = FILTER avgRatings BY avgRating == 5.0;



```
grunt> fiveStarMoviesNew = FILTER avgRatings BY avgRating == 5.0;
grunt>
grunt>
```

  dump fiveStarMoviesNew;



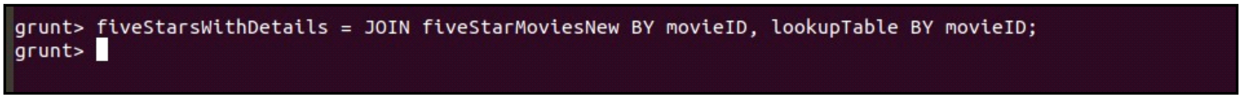
```
Total records written : 10
Total bytes written : 160
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1662684057214_0023


2022-09-09 21:47:01,827 [main] INFO  org.apache.pig.backend.hadoop.executionengine.
mapReduceLayer.MapReduceLauncher - Success!
2022-09-09 21:47:01,827 [main] INFO  org.apache.hadoop.conf.Configuration.deprecati
on - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-09-09 21:47:01,827 [main] INFO  org.apache.hadoop.conf.Configuration.deprecati
on - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-09-09 21:47:01,830 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [
pig.schematuple] was not set... will not generate code.
2022-09-09 21:47:01,849 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInpu
tFormat - Total input paths to process : 1
2022-09-09 21:47:01,849 [main] INFO  org.apache.pig.backend.hadoop.executionengine.
util.MapRedUtil - Total input paths to process : 1
(814,5.0)
(1122,5.0)
(1189,5.0)
(1201,5.0)
(1293,5.0)
(1467,5.0)
(1500,5.0)
(1536,5.0)
(1599,5.0)
(1653,5.0)
grunt>
```

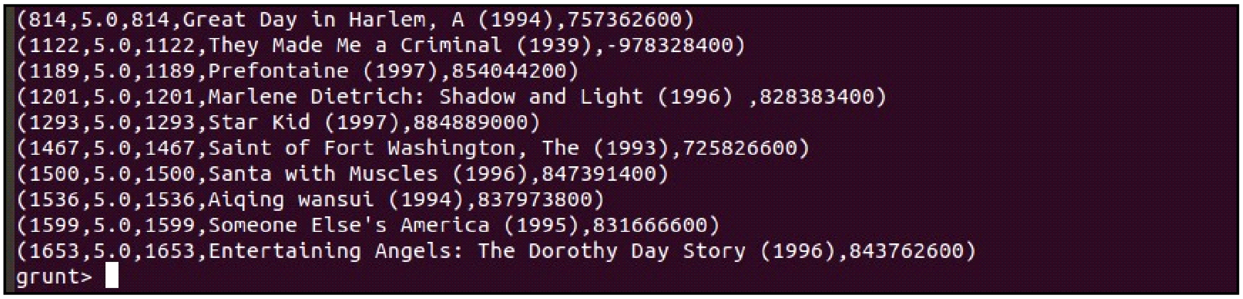
- Join this table with the lookup table:

  fiveStarsWithDetails = JOIN fiveStarMoviesNew BY movieID, lookupTable BY movieID;



```
grunt> fiveStarsWithDetails = JOIN fiveStarMoviesNew BY movieID, lookupTable BY movieID;
grunt>
```

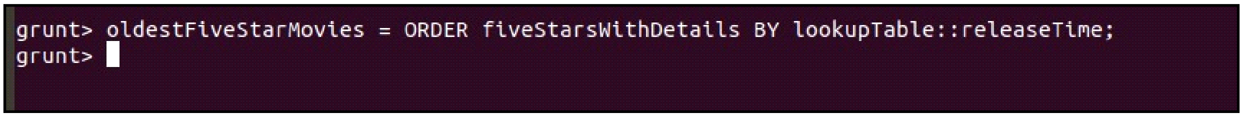
  dump fiveStarsWithDetails;



```
(814,5.0,814,Great Day in Harlem, A (1994),757362600)
(1122,5.0,1122,They Made Me a Criminal (1939),-978328400)
(1189,5.0,1189,Prefontaine (1997),854044200)
(1201,5.0,1201,Marlene Dietrich: Shadow and Light (1996) ,828383400)
(1293,5.0,1293,Star Kid (1997),884889000)
(1467,5.0,1467,Saint of Fort Washington, The (1993),725826600)
(1500,5.0,1500,Santa with Muscles (1996),847391400)
(1536,5.0,1536,Aiqing wansui (1994),837973800)
(1599,5.0,1599,Someone Else's America (1995),831666600)
(1653,5.0,1653,Entertaining Angels: The Dorothy Day Story (1996),843762600)
grunt>
```

- Order the results by time (year):

oldestFiveStarMovies = ORDER fiveStarsWithDetails BY lookupTable::releaseTime;



```
grunt> oldestFiveStarMovies = ORDER fiveStarsWithDetails BY lookupTable::releaseTime;
grunt>
```

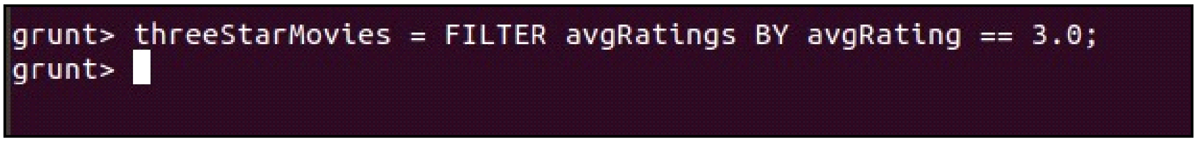
dump oldestFiveStarMovies;

c). Find the oldest 3-star movies from u.data and u.item datasets.

To do follow the following steps

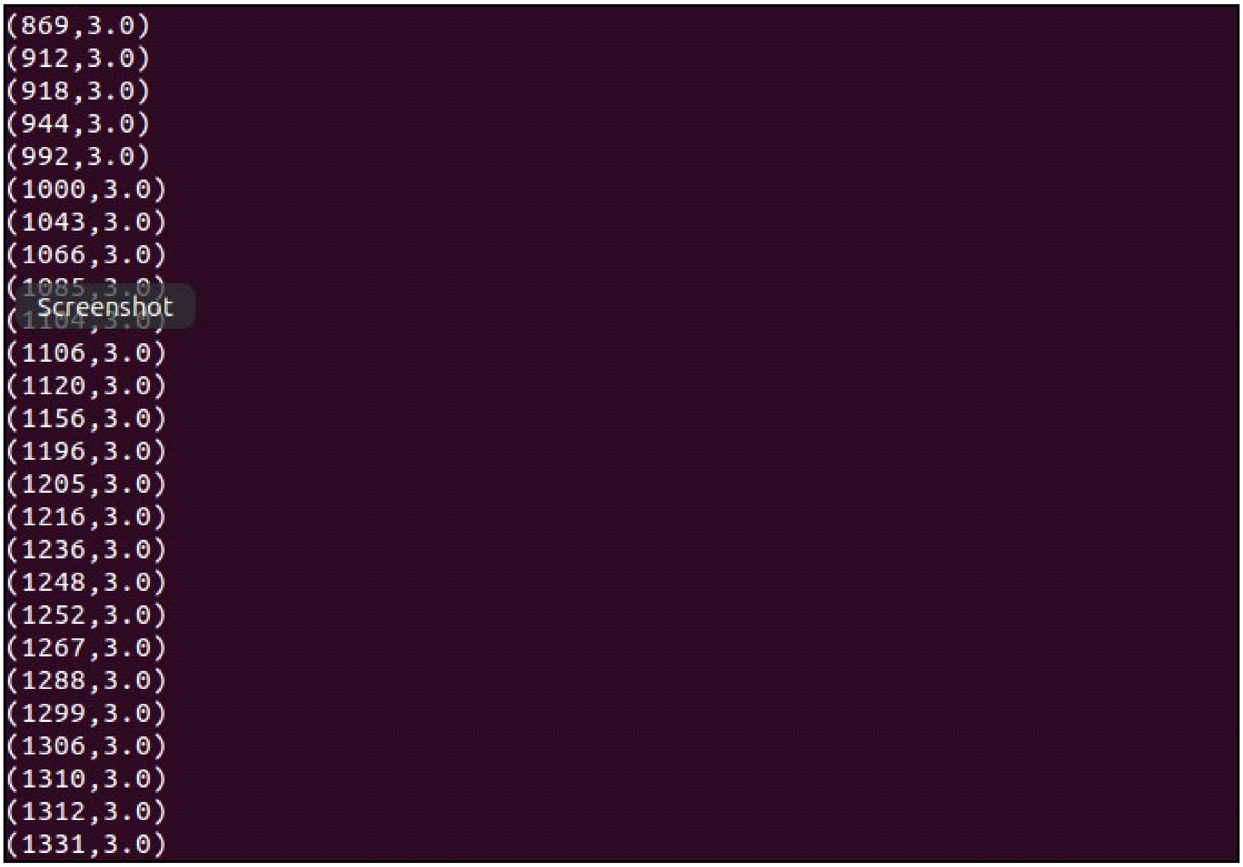
Filter the movies with average rating 3:

threeStarMovies = FILTER avgRatings BY avgRating == 3.0;



```
grunt> threeStarMovies = FILTER avgRatings BY avgRating == 3.0;
grunt>
```

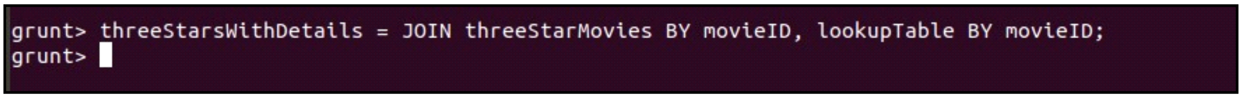
dump threeStarMovies;



```
(869,3.0)
(912,3.0)
(918,3.0)
(944,3.0)
(992,3.0)
(1000,3.0)
(1043,3.0)
(1066,3.0)
(1085,3.0)
(1104,3.0)
(1106,3.0)
(1120,3.0)
(1156,3.0)
(1196,3.0)
(1205,3.0)
(1216,3.0)
(1236,3.0)
(1248,3.0)
(1252,3.0)
(1267,3.0)
(1288,3.0)
(1299,3.0)
(1306,3.0)
(1310,3.0)
(1312,3.0)
(1331,3.0)
```

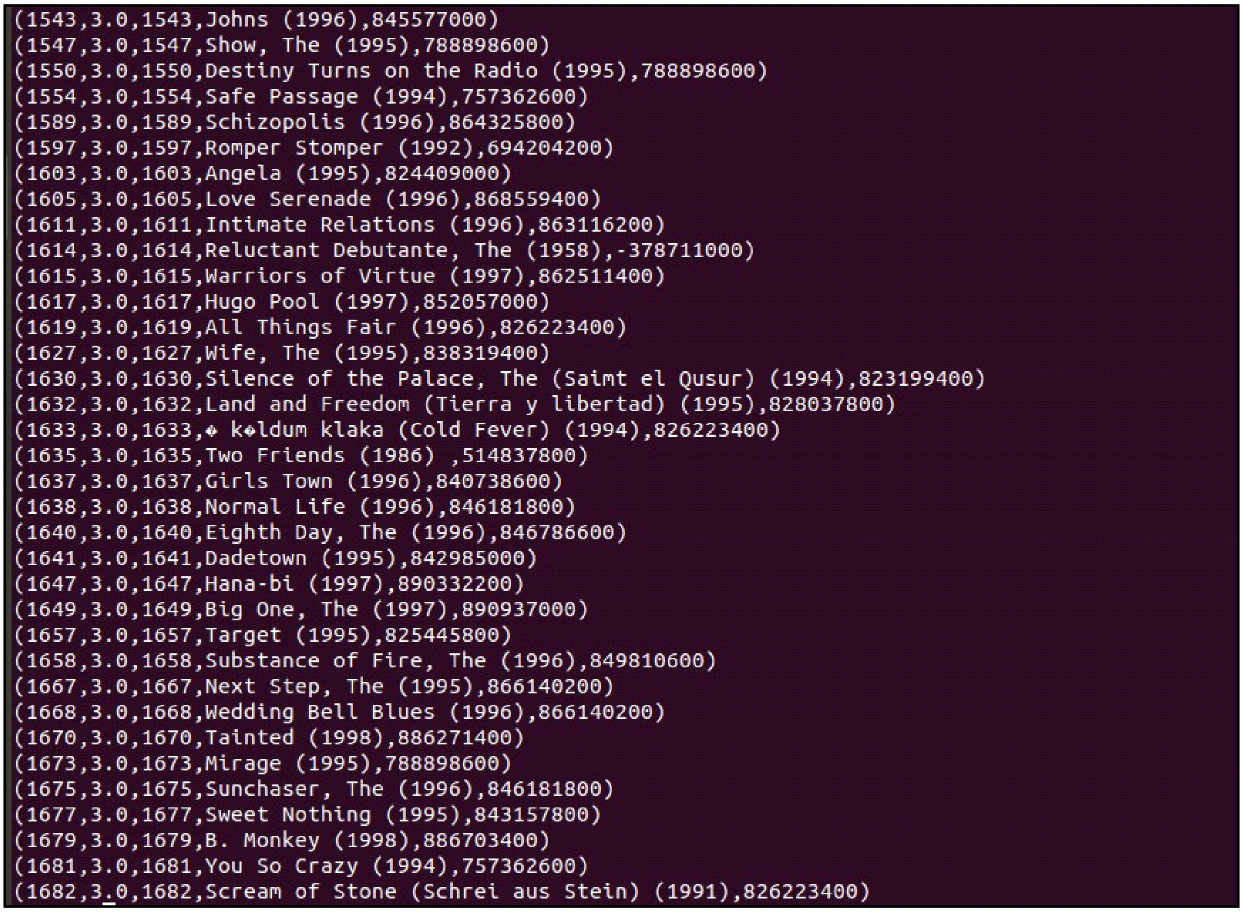
Join this table with the lookup table found in part (b):

threeStarsWithDetails = JOIN threeStarMovies BY movieID, lookupTable BY movieID;



```
grunt> threeStarsWithDetails = JOIN threeStarMovies BY movieID, lookupTable BY movieID;
grunt> 
```

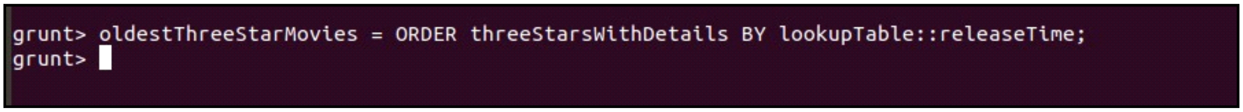
dump threeStarsWithDetails;



```
(1543,3.0,1543,Johns (1996),845577000)
(1547,3.0,1547,Show, The (1995),788898600)
(1550,3.0,1550,Destiny Turns on the Radio (1995),788898600)
(1554,3.0,1554,Safe Passage (1994),757362600)
(1589,3.0,1589,Schizopolis (1996),864325800)
(1597,3.0,1597,Romper Stomper (1992),694204200)
(1603,3.0,1603,Angela (1995),824409000)
(1605,3.0,1605,Love Serenade (1996),868559400)
(1611,3.0,1611,Intimate Relations (1996),863116200)
(1614,3.0,1614,Reluctant Debutante, The (1958),-378711000)
(1615,3.0,1615,Warriors of Virtue (1997),862511400)
(1617,3.0,1617,Hugo Pool (1997),852057000)
(1619,3.0,1619,All Things Fair (1996),826223400)
(1627,3.0,1627,Wife, The (1995),838319400)
(1630,3.0,1630,Silence of the Palace, The (Saimt el Qusur) (1994),823199400)
(1632,3.0,1632,Land and Freedom (Tierra y libertad) (1995),828037800)
(1633,3.0,1633,◆ k◆ldum klaka (Cold Fever) (1994),826223400)
(1635,3.0,1635,Two Friends (1986) ,514837800)
(1637,3.0,1637,Girls Town (1996),840738600)
(1638,3.0,1638,Normal Life (1996),846181800)
(1640,3.0,1640,Eighth Day, The (1996),846786600)
(1641,3.0,1641,Dadetown (1995),842985000)
(1647,3.0,1647,Hana-bi (1997),890332200)
(1649,3.0,1649,Big One, The (1997),890937000)
(1657,3.0,1657,Target (1995),825445800)
(1658,3.0,1658,Substance of Fire, The (1996),849810600)
(1667,3.0,1667,Next Step, The (1995),866140200)
(1668,3.0,1668,Wedding Bell Blues (1996),866140200)
(1670,3.0,1670,Tainted (1998),886271400)
(1673,3.0,1673,Mirage (1995),788898600)
(1675,3.0,1675,Sunchaser, The (1996),846181800)
(1677,3.0,1677,Sweet Nothing (1995),843157800)
(1679,3.0,1679,B. Monkey (1998),886703400)
(1681,3.0,1681,You So Crazy (1994),757362600)
(1682,3.0,1682,Scream of Stone (Schrei aus Stein) (1991),826223400)
```

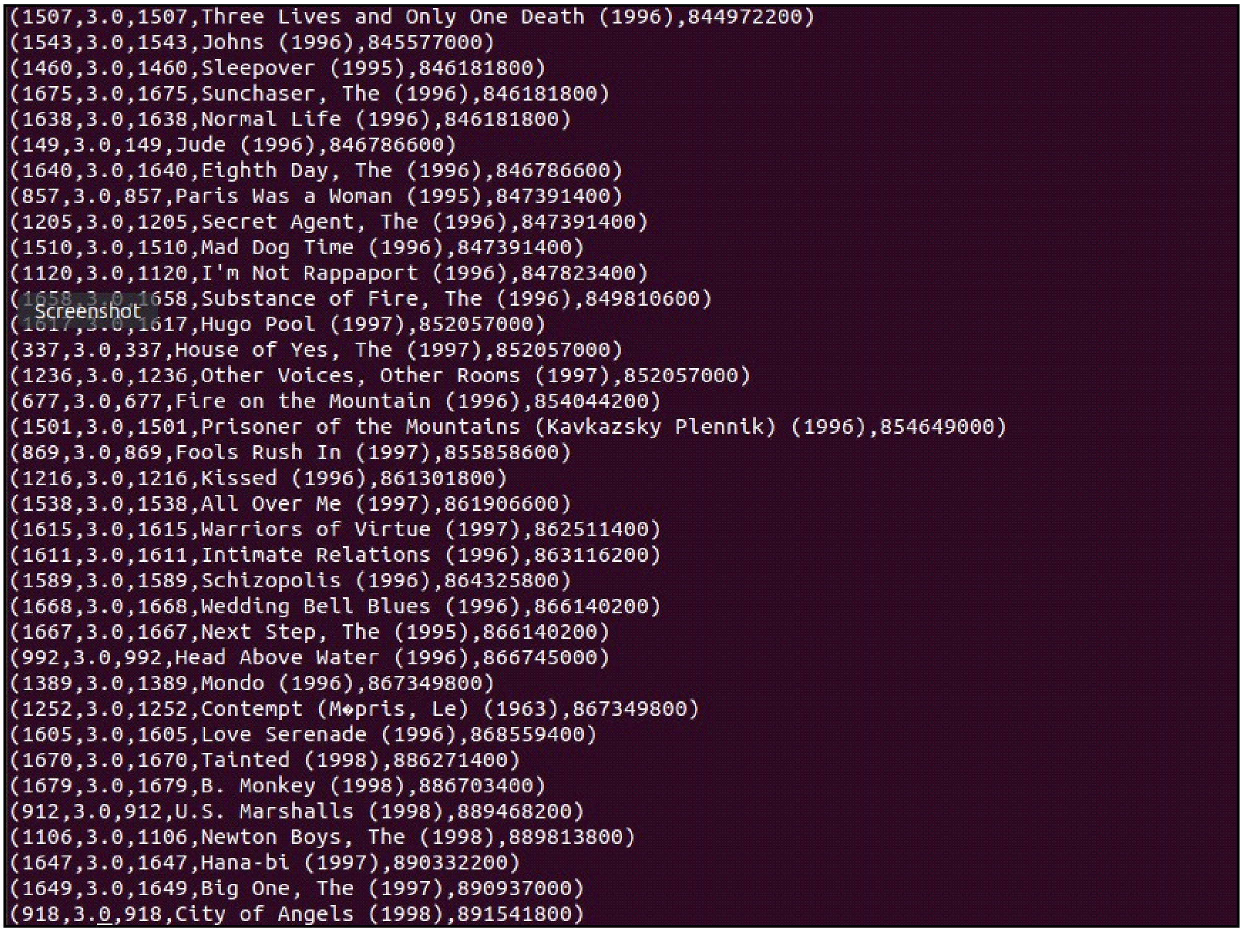
3) Order the results by time (year):

oldestThreeStarMovies = ORDER threeStarsWithDetails BY lookupTable::releaseTime;



```
grunt> oldestThreeStarMovies = ORDER threeStarsWithDetails BY lookupTable::releaseTime;
grunt>
```

dump oldestThreeStarMovies;

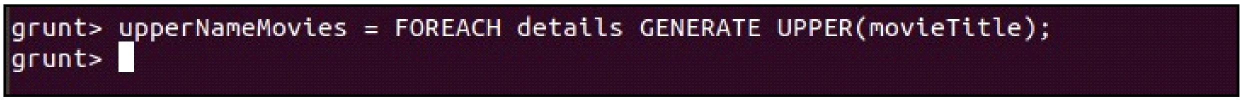


```
(1507,3.0,1507,Three Lives and Only One Death (1996),844972200)
(1543,3.0,1543,Johns (1996),845577000)
(1460,3.0,1460,Sleepover (1995),846181800)
(1675,3.0,1675,Sunchaser, The (1996),846181800)
(1638,3.0,1638,Normal Life (1996),846181800)
(149,3.0,149,Jude (1996),846786600)
(1640,3.0,1640,Eighth Day, The (1996),846786600)
(857,3.0,857,Paris Was a Woman (1995),847391400)
(1205,3.0,1205,Secret Agent, The (1996),847391400)
(1510,3.0,1510,Mad Dog Time (1996),847391400)
(1120,3.0,1120,I'm Not Rappaport (1996),847823400)
(1658,3.0,1658,Substance of Fire, The (1996),849810600)
(1617,3.0,1617,Hugo Pool (1997),852057000)
(337,3.0,337,House of Yes, The (1997),852057000)
(1236,3.0,1236,Other Voices, Other Rooms (1997),852057000)
(677,3.0,677,Fire on the Mountain (1996),854044200)
(1501,3.0,1501,Prisoner of the Mountains (Kavkazsky Plennik) (1996),854649000)
(869,3.0,869,Fools Rush In (1997),855858600)
(1216,3.0,1216,Kissed (1996),861301800)
(1538,3.0,1538,All Over Me (1997),861906600)
(1615,3.0,1615,Warriors of Virtue (1997),862511400)
(1611,3.0,1611,Intimate Relations (1996),863116200)
(1589,3.0,1589,Schizopolis (1996),864325800)
(1668,3.0,1668,Wedding Bell Blues (1996),866140200)
(1667,3.0,1667,Next Step, The (1995),866140200)
(992,3.0,992,Head Above Water (1996),866745000)
(1389,3.0,1389,Mondo (1996),867349800)
(1252,3.0,1252,Contempt (M�pris, Le) (1963),867349800)
(1605,3.0,1605,Love Serenade (1996),868559400)
(1670,3.0,1670,Tainted (1998),886271400)
(1679,3.0,1679,B. Monkey (1998),886703400)
(912,3.0,912,U.S. Marshalls (1998),889468200)
(1106,3.0,1106,Newton Boys, The (1998),889813800)
(1647,3.0,1647,Hana-bi (1997),890332200)
(1649,3.0,1649,Big One, The (1997),890937000)
(918,3.0,918,City of Angels (1998),891541800)
```

d). Display name of all movies in uppercase.

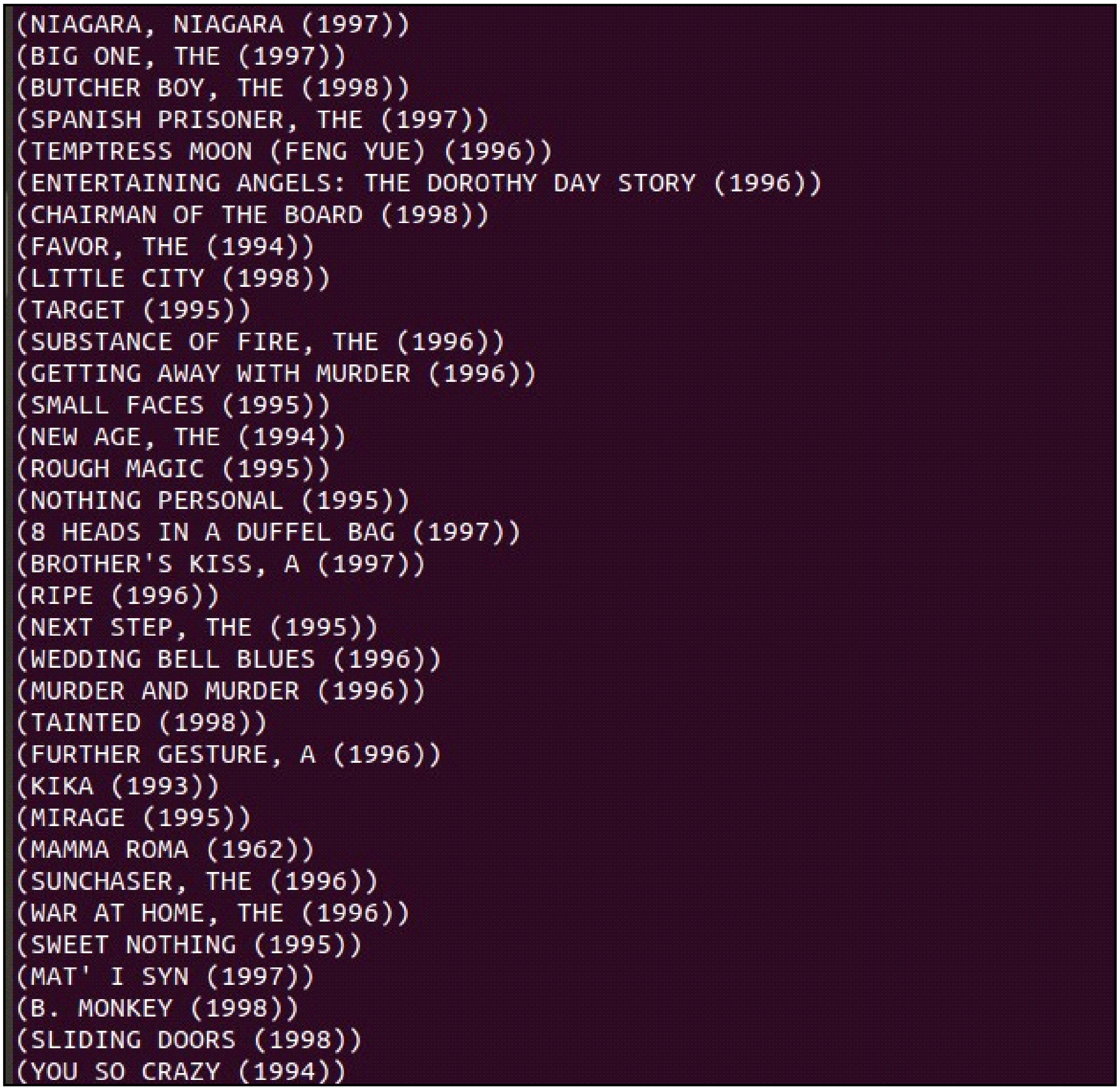
If while loading the data we don’t use USING PigStorage(‘,) then while executing the following commands entire columns would be uppercase as it is reqd for a delimiter!

upperNameMovies = FOREACH details GENERATE UPPER(movieTitle);



```
grunt> upperNameMovies = FOREACH details GENERATE UPPER(movieTitle);
grunt> 
```

dump upperNameMovies;



```
(NIAGARA, NIAGARA (1997))
(BIG ONE, THE (1997))
(BUTCHER BOY, THE (1998))
(SPANISH PRISONER, THE (1997))
(TEMPTRESS MOON (FENG YUE) (1996))
(ENTERTAINING ANGELS: THE DOROTHY DAY STORY (1996))
(CHAIRMAN OF THE BOARD (1998))
(FAVOR, THE (1994))
(LITTLE CITY (1998))
(TARGET (1995))
(SUBSTANCE OF FIRE, THE (1996))
(GETTING AWAY WITH MURDER (1996))
(SMALL FACES (1995))
(NEW AGE, THE (1994))
(ROUGH MAGIC (1995))
(NOTHING PERSONAL (1995))
(8 HEADS IN A DUFFEL BAG (1997))
(BROTHER'S KISS, A (1997))
(RIPE (1996))
(NEXT STEP, THE (1995))
(WEDDING BELL BLUES (1996))
(MURDER AND MURDER (1996))
(TAINTED (1998))
(FURTHER GESTURE, A (1996))
(KIKA (1993))
(MIRAGE (1995))
(MAMMA ROMA (1962))
(SUNCHASER, THE (1996))
(WAR AT HOME, THE (1996))
(SWEET NOTHING (1995))
(MAT' I SYN (1997))
(B. MONKEY (1998))
(SLIDING DOORS (1998))
(YOU SO CRAZY (1994))
```