

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

EXPERIMENT-7

AIM: Movie data Set Analysis using Pig, Database Operations in PIG Using CLOUDER

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

```
[cloudera@quickstart LAB 7]$ hdfs dfs -copyFromLocal u.item/
copyFromLocal: u.item/: No such file or directory
[cloudera@quickstart LAB 7]$ hdfs dfs -copyFromLocal u.data/
copyFromLocal: `u.data': File exists
```

- Load the movie data into pig

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

```
dump ratings;
```

1. Find the movie with avg rating >4.0 from u.data dataset.

```
(1635,{(655,1635,3,887432079)})
(1636,{(655,1636,4,887473570)})
(1637,{(655,1637,3,888984255)})
(1638,{(655,1638,3,887488947)})
(1639,{(655,1639,4,887650483),(918,1639,5,891987571),(840,1639,4,891211447)})
(1640,{(655,1640,3,888474646)})
(1641,{(655,1641,3,887427810)})
(1642,{(707,1642,5,886286491),(655,1642,4,888474934)})
(1643,{(699,1643,3,879147169),(655,1643,5,887611511),(782,1643,2,891499321),(901,1643,5,877130473)})
(1644,{(782,1644,2,891500110),(655,1644,1,888474327)})
(1645,{(655,1645,4,892871225)})
(1646,{(655,1646,3,891913577),(828,1646,4,893186124)})
(1647,{(655,1647,3,891817435)})
(1648,{(655,1648,2,891817435)})
(1649,{(655,1649,3,892333993)})
(1650,{(655,1650,4,892871225)})
(1651,{(655,1651,4,891913500)})
(1652,{(782,1652,1,891500230),(662,1652,3,880570909),(756,1652,1,874828198)})
```

2.

```
grunt> ratingsByMovie = GROUP ratings BY movieID;
grunt> dump ratingsByMovie;
```

Group the data according to movieID:

```
(1172,3.6666666666666665)
(1173,2.5714285714285716)
(1174,3.25)
(1175,3.4)
(1176,3.3636363636363638)
(1177,1.8571428571428572)
(1178,2.473684210526316)
(1179,2.1111111111111111)
(1180,2.5)
(1181,2.5)
(1182,2.3076923076923075)
(1183,2.210526315789474)
(1184,2.5)
(1185,2.2777777777777777)
(1186,3.3333333333333335)
```

3.

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

Find average ratings for the

grouped data:

```
(12,4.385767790262173)
(22,4.151515151515151)
(23,4.1208791208791204)
(45,4.05)
(48,4.094017094017094)
(50,4.359589041095891)
(56,4.060913705583756)
(59,4.0602409638554215)
(60,4.015625)
(64,4.445229681978798)
(79,4.044642857142857)
(83,4.0625)
(89,4.138181818181818)
```

```
grunt> fiveStarMovies = FILTER avgRatings BY avgRating> 4.0;
grunt> dump fiveStarMovies;
```

Filter the required results:

5. Find the oldest 5-star movies from u.data and u.item datasets.
6. Load the u.item in pig:

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

```
(1173,Bliss (1997),06-Jun-1997,,http://us.imdb.com/M/title-exact?Bliss%20(1997))
(1174,Caught (1996),27-Sep-1996,,http://us.imdb.com/M/title-exact?Caught%20(1996))
(1175,Hugo Pool (1997),01-Jan-1997,,http://us.imdb.com/M/title-exact?Hugo+Pool+(1997))
(1176>Welcome To Sarajevo (1997),01-Jan-1997,,http://us.imdb.com/M/title-exact?Welcome+To+Sarajevo+(1997))
(1177,Dunston Checks In (1996),01-Jan-1996,,http://us.imdb.com/M/title-exact?Dunston%20Checks%20In%20(1996))
(1178,Major Payne (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Major%20Payne%20(1994))
(1179,Man of the House (1995),01-Jan-1995,,http://us.imdb.com/M/title-exact?Man%20of%20the%20House%20(1995))
(1180,I Love Trouble (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?I%20Love%20Trouble%20(1994))
(1181,Low Down Dirty Shame, A (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Low%20Down%20Dirty%20Shame,%20A%20(1994))
(1182,Cops and Robbersons (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Cops%20and%20Robbersons%20(1994))
(1183,Cowboy Way, The (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Cowboy%20Way,%20The%20(1994))
(1184,Endless Summer 2, The (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Endless%20Summer%202,%20The%20(1994))
(1185,In the Army Now (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?In%20the%20Army%20Now%20(1994))
(1186,Inkwell, The (1994),01-Jan-1994,,http://us.imdb.com/M/title-exact?Inkwell,%20The%20(1994))
(1187,Switchblade Sisters (1975),17-May-1975,,http://us.imdb.com/M/title-exact?Switchblade%20Sisters%20(1975))
(1188,Young Guns II (1990),01-Jan-1990,,http://us.imdb.com/M/title-exact?Young%20Guns%20II%20(1990))
(1189,Prefontaine (1997),24-Jan-1997,,http://us.imdb.com/M/title-exact?Prefontaine%20(1997))
(1190,That Old Feeling (1997),04-Apr-1997,,http://us.imdb.com/M/title-exact?That%20Old%20Feeling%20(1997))
(1191,Letter From Death Row, A (1998),01-Feb-1998,,http://us.imdb.com/M/title-exact?Letter+From+Death+Row%2C+A+(1998))
(1192,Boys of St. Vincent, The (1993),01-Jan-1993,,http://us.imdb.com/M/title-exact?Boys%20of%20St.%20Vincent,%20The%20(1993))
dump details;
```

```
(1636,Brothers in Trouble (1995),830502000)
(1637,Girls Town (1996),840783600)
(1638,Normal Life (1996),846226800)
(1639,Bitter Sugar (Azucar Amargo) (1996),848649600)
(1640,Eighth Day, The (1996),846835200)
(1641,Dadetown (1995),843030000)
1. (1642,Some Mother's Son (1996),851673600)

grunt> lookupTable = FOREACH details GENERATE movieID, movieTitle, ToUnixTime(ToDate(releaseDate, 'dd-MMM-yyyy')) AS
2023-03-19 03:29:24,256 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated.
2023-03-19 03:29:24,256 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
grunt> dump lookupTable;
```

Create a scalable timestamp column to compare the time:

```
(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)
2. (1653,5.0) Filter the movies with average rating as 5:
```

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

```
(814,5.0,814,Great Day in Harlem, A (1994),757411200)
(1122,5.0,1122,They Made Me a Criminal (1939),-978278400)
(1189,5.0,1189,Prefontaine (1997),854092800)
(1201,5.0,1201,Marlene Dietrich: Shadow and Light (1996),828432000)
(1293,5.0,1293,Star Kid (1997),884937600)
(1467,5.0,1467,Saint of Fort Washington, The (1993),725875200)
(1500,5.0,1500,Santa with Muscles (1996),847440000)
(1536,5.0,1536,Aiqing wansui (1994),838018800)
(1599,5.0,1599,Someone Else's America (1995),831711600)
1. (1653,5.0,1653,Entertaining Angels: The Dorothy Day Story (1996),843807600)
```

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

Join this table with the lookup table:

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

Order the results by time (year):

```
(1122,5.0,1122,They Made Me a Criminal (1939),-978278400)
(1467,5.0,1467,Saint of Fort Washington, The (1993),725875200)
(814,5.0,814,Great Day in Harlem, A (1994),757411200)
(1201,5.0,1201,Marlene Dietrich: Shadow and Light (1996),828432000)
(1599,5.0,1599,Someone Else's America (1995),831711600)
(1536,5.0,1536,Aiqing wansui (1994),838018800)
(1653,5.0,1653,Entertaining Angels: The Dorothy Day Story (1996),843807600)
(1500,5.0,1500,Santa with Muscles (1996),847440000)
(1189,5.0,1189,Prefontaine (1997),854092800)
(1293,5.0,1293,Star Kid (1997),884937600)
```

(43,3.0)
 (74,3.0)
 (75,3.0)
 (84,3.0)
 (149,3.0)
 (337,3.0)
 (460,3.0)
 (583,3.0)
 (600,3.0)
 (643,3.0)
 (670,3.0)
 (677,3.0)
 (698,3.0)
 (780,3.0)
 (822,3.0)
 (843,3.0)
 (857,3.0)

- grunt> threeStarMovies = FILTER avgRatings BY avgRating == 3.0;
 1. grunt> dump threestarMovies;

Find the oldest 3-star movies from u.data and

u.item datasets.

```
(43,3.0,43,Disclosure (1994),757411200)
(74,3.0,74,Faster Pussycat! Kill! Kill! (1965),-157737600)
(75,3.0,75,Brother Minister: The Assassination of Malcolm X (1994),757411200)
(84,3.0,84,Robert A. Heinlein's The Puppet Masters (1994),757411200)
(149,3.0,149,Jude (1996),846835200)
(337,3.0,337,House of Yes, The (1997),852105600)
(460,3.0,460,Crossing Guard, The (1995),788947200)
(583,3.0,583,Romeo Is Bleeding (1993),725875200)
(600,3.0,600,Daniel Defoe's Robinson Crusoe (1996),820483200)
(643,3.0,643,The Innocent (1994),757411200)
(670,3.0,670,Body Snatchers (1993),725875200)
(677,3.0,677,Fire on the Mountain (1996),854092800)
(698,3.0,698,Browning Version, The (1994),757411200)
(780,3.0,780,Dumb & Dumber (1994),757411200)
(822,3.0,822,Faces (1968),-63129600)
(843,3.0,843,Shaggy Dog, The (1959),-347126400)
(857,3.0,857,Paris Was a Woman (1995),847440000)
(869,3.0,869,Fools Rush In (1997),855907200)
(912,3.0,912,U.S. Marshalls (1998),889516800)
```

Join this table with the lookup table found in part

- (b):
 (1542,3.0,1542,Scarlet Letter, The (1926),-1388505600)
 (1299,3.0,1299,Penny Serenade (1941),-915120000)
 (1310,3.0,1310,Walk in the Sun, A (1945),-788893200)
 (1454,3.0,1454,Angel and the Badman (1947),-725817600)
 (1614,3.0,1614,Reluctant Debutante, The (1958),-378662400)
 (843,3.0,843,Shaggy Dog, The (1959),-347126400)
 (74,3.0,74,Faster Pussycat! Kill! Kill! (1965),-157737600)
 (822,3.0,822,Faces (1968),-63129600)
 (1635,3.0,1635,Two Friends (1986),514886400)
 (1196,3.0,1196,Savage Nights (Nuits fauves, Les) (1992),694252800)
 (1104,3.0,1104,C'est arriv   pr  s de chez vous (1992),694252800)
 (1597,3.0,1597,Romper Stomper (1992),694252800)
 (1521,3.0,1521,Mr. Wonderful (1993),725875200)
 3. (670,3.0,670,Body Snatchers (1993),725875200)

Order the results by time (year):

4. Display name of all movies in uppercase.

```
grunt> upperNameMovies = FOREACH details GENERATE UPPER(movieTitle);
grunt> dump upperNameMovies;
grunt> threeStarsWithDetails = JOIN threeStarMovies BY movieID, lookupTable BY movieID;
grunt> dump threestarsWithDetails;
```

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!
 grunt> oldestThreeStarMovies = ORDER threeStarsWithDetails BY lookupTable::releaseTime;
 grunt> dump oldestThreestarMovies;

(FULL SPEED (1996))
(CUTTHROAT ISLAND (1995))
(GHOST IN THE SHELL (KOKAKU KIDOTAI) (1995))
(VAN, THE (1996))
(OLD LADY WHO WALKED IN THE SEA, THE (VIEILLE QUI MARCHAIT DANS LA MER, LA) (1991))
(NIGHT FLIER (1997))
(METRO (1997))
(GRIDLOCK'D (1997))
(BUSHWHACKED (1995))
(BAD GIRLS (1994))
(BLINK (1994))
(FOR LOVE OR MONEY (1993))
(BEST OF THE BEST 3: NO TURNING BACK (1995))
(A CHEF IN LOVE (1996))
(CONTEMPT (MÉPRIS, LE) (1963))
(TIE THAT BINDS, THE (1995))
(GONE FISHIN' (1997))
(BROKEN ENGLISH (1996))

Submitted By:

Name: Soumya Ranjan Maharana

Regd. No.: 20010414

Branch: CSE (AI & ML)

Group: 4 (H)