Assignment Submission-Case Study 1 (Movie Rating)

Problem Statement:

What are the movie titles that the user has rated?
How many times a movie has been rated by the user?
In question 2 above, what is the average rating given for a movie?

Solution: As per the output generated by MR job that is provided for reference, there are basically 3 columns: "Title Of Movie", "Number of times it has been rated(Count)" "Average rating provided".
I have attained the same output after running the "pig" script written for this.
Input Files involved:
1. Movies.csv (movieId, title, genres)
2. Ratings.csv(userid, movieid, rating,timestamp)

Explanation of script:
1. First of all "piggybank.jar" is registered to use **CSVExcelStorage**.
2. Loading the dataset movies.csv using **CSVExcelStorageData** to get rid of the header rows.
3. Getting required fields and explicitly typecasting each of them.
4. Filtering data to avoid entries where **movieID** OR Title is null.
5. Loading the dataset ratings.csv using **CSVExcelStorageData** to get rid of the header rows.
6. Getting all fields and explicitly typecasting each of them.
7. Filtering data to avoid entries where movieID OR rating is null.
8. Since rating is present only in "ratings.csv" and as mentioned in the provided "readme file→ movieId is consistent between the ratings.csv and movies.csv", grouping the content of ratings.csv by 'movieId'.
9. Using foreach, got the count(number of times movie is rated by users) and average(rating) .
10. Now joined the output generated with "movies.csv" to fetch the title for movies. Joined on the basis of movieid.
11. Created another relation to get the data in the desired format.
12. Stored the data in the file.

Going to the directory where casestudy pig script is placed. The input file location needs to be given in the script, Running in local mode:
Running as below screenshot:

```
grunt> run CaseStudy_MovieRating.pig
2018-11-26 23:40:05,341 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-26 23:40:05,341 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.defaul
t.name is deprecated. Instead, use fs.defaultFS
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-11-26 23:40:05,541 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-26 23:40:05,541 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.defaul
t.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> loadMovieList = LOAD '/home/acadgild/sumanth/CASE_STUDY_PIG_IMPLEMENTATION/movies.csv' USI
NG org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER'
);
2018-11-26 23:40:06,506 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-26 23:40:06,506 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.defaul
t.name is deprecated. Instead, use fs.defaultFS
grunt>
grunt> movieList = FOREACH loadMovieList GENERATE (int)$0 as movieId,(chararray)$1 as title,(char
array)$2 as genres;
grunt>
grunt> finalMovieList = FILTER movieList BY (movieId IS NOT NULL) AND (title IS NOT NULL);
grunt>
grunt> loadRatingList = LOAD '/home/acadgild/sumanth/CASE_STUDY_PIG_IMPLEMENTATION/ratings.csv' U
SING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADE
R');
2018-11-26 23:40:07,159 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-11-26 23:40:07,159 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.defaul
t.name is deprecated. Instead, use fs.defaultFS
```



```
grunt>
grunt> ratingList = FOREACH loadRatingList GENERATE (int)$0 as userId,(int)$1 as movieId,(double)
$2 as rating, (long)$3 as timestamp;
grunt>
grunt> finalRatingList = FILTER ratingList BY (movieId IS NOT NULL) AND (rating IS NOT NULL);
grunt>
grunt> groupingById = GROUP finalRatingList BY (movieId);
grunt>
grunt> requiredDetails = FOREACH groupingById GENERATE group, COUNT(finalRatingList.rating) AS Nu
mOfRatings, AVG(finalRatingList.rating) AS AvgRating;
grunt>
grunt> joinToGetTitle = JOIN finalMovieList BY movieId, requiredDetails BY group;
grunt>
grunt> finalData = FOREACH joinToGetTitle GENERATE finalMovieList::title,requiredDetails::NumOfRa
tings,AvgRating;
grunt>
grunt> STORE finalData INTO '/home/acadgild/sumanth/CASE_STUDY_PIG_IMPLEMENTATION/Output';
2018-11-26 23:40:08,042 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

Please refer script and output file from running the script for further details.