CA4022 - MovieLens dataset analysis using Apache Pig and Hive

Adam Tegart - 19327493

October 2022

Github Repo here

1 Introduction

Apache Hadoop is a framework for processing large scale data in a distributed way across multiple computers. This framework is built on top of MapReduce, a processing technique with a divide-and-conquer paradigm that allows big data to be processed in parallel. Apache have other Hadoop-related projects that can make use of MapReduce within Hadoop. Apache Pig is a high-level language for parallel computation and can be used to clean and process big data. Apache Hive is a data warehousing infrastructure that allows for ad-hoc querying.

This assignment was carried out using Apache Pig and Hive. The files used and created in this assignment can be found here along with the code used to generate them. The MovieLens dataset was cleaned and queried in Pig and then more complex queries were carried out using Hive. The following is an overview of the steps taken at each stage and an analysis of the results from queries to derive insight from the MovieLens dataset.

2 Cleaning the Data

Before carrying out any sort of analysis, the dataset had to be cleaned and processed into a more suitable format for querying. This was carried out using Apache Pig locally. The general steps taken and issues addressed are mentioned below briefly, the cleaning.pig script contains the code used and comments for more detail.

- CSVLoader to ensure the columns weren't split on quoted commas.
- Headers removed.
- Year was separated from the title.
- Genres were split (This was done after the join as Hive can read the string version in easier).

- Convert the timestamps into YYYY-MM-DD HH:MM:SS UTC format.

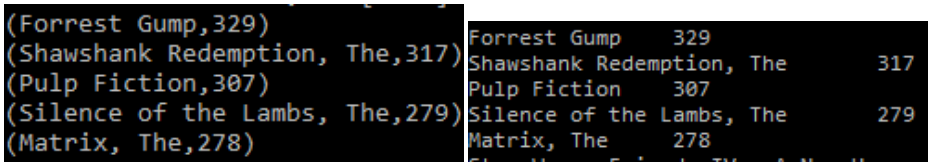
Once all of the files had been cleaned (excluding links due to needing to web-scrape to derive insight) I joined the 3 relations and saved the new relation in outputs/joined. I then noticed that the ratings were duplicated due to the nature of tags.csv. I was performing a left join of ratings with tags to ensure that tags were only added if a rating was present. I had not realised that several tags could relate to one rating. So I dropped tags and only used movies and ratings. After this I carried out the genre split as mentioned previously and then it was time to query the data.

3 Querying the Data

The simple queries were carried out in both Pig and Hive. The results for both will be shown together for a sanity check, there will be some analysis though the majority is saved for the more advanced queries.

3.1 Movies with the most ratings

This query was quite simple, the data had to be grouped by movie and the ratings counted.



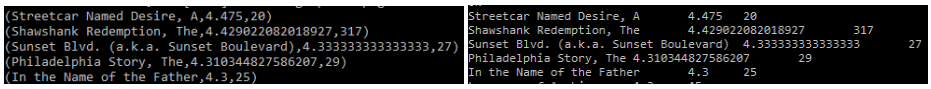
(a) Pig query results (b) Hive query results

Figure 1: Pig and Hive results for movies with most ratings

It can be seen that the results match in both cases, which is a good sanity check of the final results. Forrest Gump can be seen to have the most ratings, followed closely by Shawshank Redemption and Pulp Fiction..

3.2 Most liked movies

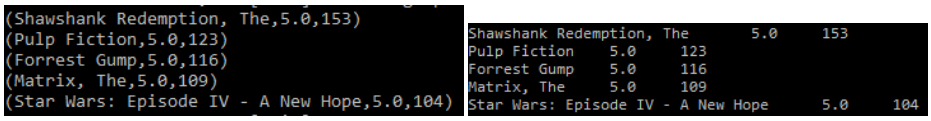
This query could have been interpreted several ways, I have decided to look at 2 of these. The first is the movies with the highest average rating where the number of reviews is significant (> 20). The second is the movies with the most 5 star ratings.



(a) Pig query results (b) Hive query results

Figure 2: Pig and Hive results for movies with the highest average rating

We can see that Streetcar Named Desire has the highest average rating where the number of ratings is over 20. We can see Shawshank Redemption has about 0.045 less in average rating but 293 more ratings, which may mean that the average for Shawshank Redemption is a better representation of the true ratings from the population.



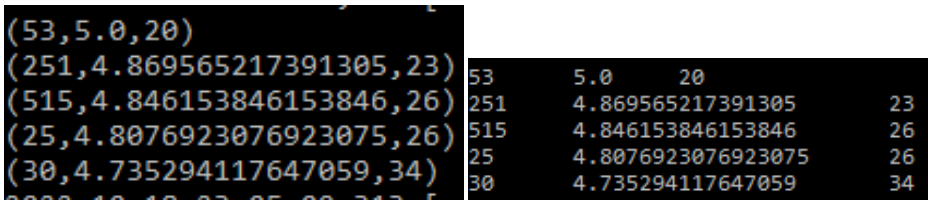
(a) Pig query results (b) Hive query results

Figure 3: Pig and Hive results for movies with the most 5 star ratings

We can see that Shawshank Redemption has 30 more 5 star ratings than Pulp Fiction, the next highest movie on the list. Taking both results into account it may be best to consider Shawshank Redemption the most liked movie, due to the number or ratings and the average rating it seems as though it is highly liked among the users leaving reviews (and hopefully the users who don't review too, assuming that the sample represents the population).

3.3 Users with the highest average rating

This query meant the data had to be grouped on the users and the ratings averaged for each user. We can see that the user with the highest average rating had slightly less ratings than the other users in the results.



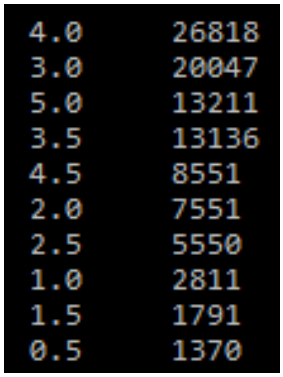
(a) Pig query results (b) Hive query results

Figure 4: Pig and Hive results for users with the highest average ratings

Now to look at the more advanced queries carried out through Hive.

3.4 Count of each rating and most popular rating

This query required the data to be grouped on the rating and the occurrences of each rating counted. The total counts and the most popular rating can be seen below. We have the group/rating first, the average of the ratings (to sanity check that we only counted the correct ratings) and the count of each rating.



(a) Count of each rating

(b) Most popular rating

Figure 5: Hive results for ratings count and most popular rating

So we can say that 4 stars is the most popular rating given by users. A couple of interesting insights we get from these results are users prefer using integer ratings and rating generously. Firstly, we can say that users prefer to use whole numbers as a rating as each whole number is most used than the half star counterpart, e.g. 3.0 has more than 3.5 etc. I believe that this is the case due to whole numbers being more simple and natural to use. We can also see that users tend to be generous with ratings as the average rating overall is 3.502 and we can see this reflected in our counts where 4 star + makes up just less than half the data (48,580 ratings 4+ star to 52,256 under 4 star).

3.5 Count of ratings by genre

This query required the genres to be read into Hive as a collection/array. This could then be exploded out so that each element of genres was a row with a copy of all other fields. The genres could then be grouped and the ratings counted for each. This gives us the following results.

```
Drama   41928
Comedy  39053
Action  30635
Thriller        26452
Adventure       24161
Romance 18124
Sci-Fi  17243
Crime   16681
Fantasy 11834
Children        9208
Mystery 7674
Horror  7291
Animation       6988
War     4859
IMAX    4145
Musical 4138
Western 1930
Documentary     1219
Film-Noir       870
(no genres listed)      47
```

Figure 6: Count of ratings in each genre

We can see that Drama, Comedy and Action are the top 3 most rated genres. It can be seen that there are ratings against a movies with no genre listed. Some of these are Pirates of the Caribbean: Dead Men Tell No Tales, Let It Be Me and The Adventures of Sherlock Holmes and Doctor to name a few. Now we can break this down further and look at the specific ratings within each genre.

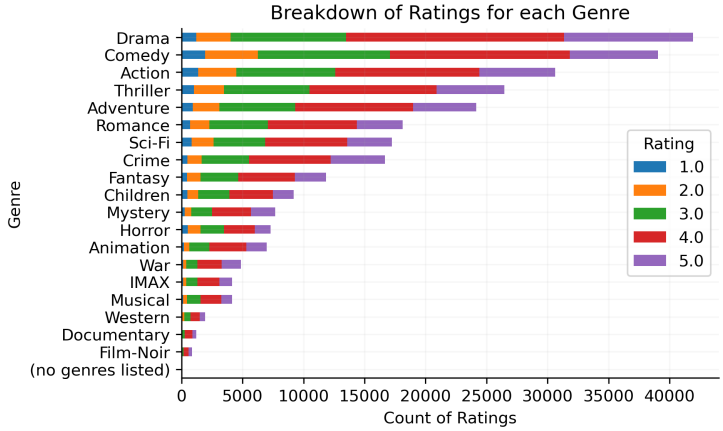


Figure 7: Breakdown of each rating per genre

We can see from Fig.7 above that Drama is rated better than Comedy as there are more 4 and 5 star ratings. It is quite difficult to compare the genres due to the size differences, so it may be best to look at the breakdown of the proportion each rating has within genre.

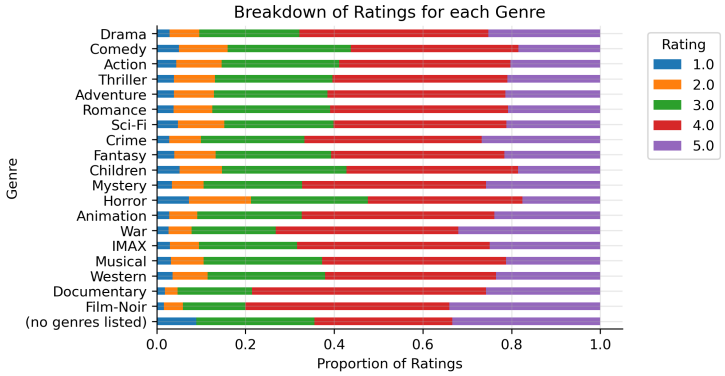


Figure 8: Proportion of each rating per genre

This makes it much easier to compare the proportions within each genre. We can see that the no genre listed group had the most ratings at either end of the scale, with the proportion of 1 and 5 star ratings. We can see that Documentary and Film-Noir had very few 1 star ratings as a proportion of their total ratings. Finally, we can see that Horror seems to be the least liked genre as they have the smallest proportion of 5 star ratings and half of the ratings are 3 star and below which is the worst of all groups as see in Fig.7(b).

We will now take a look at the distribution of ratings for popular movies over time. We will look at Pulp Fiction, Star Wars: Episode IV - A New Hope and Forrest Gump.

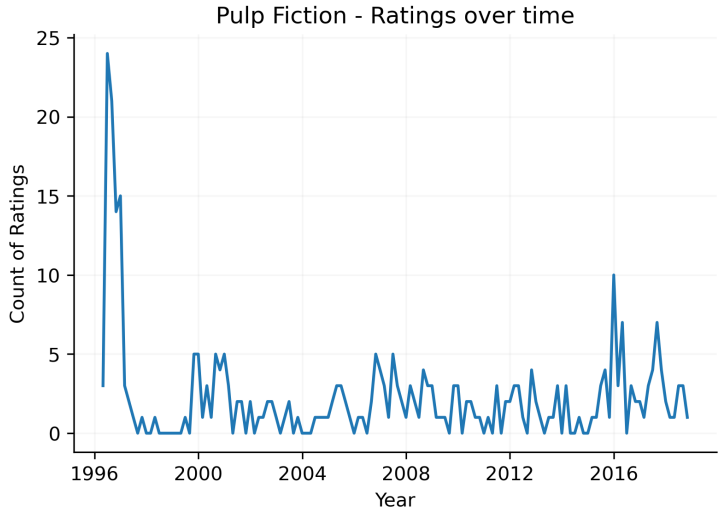


Figure 9: Ratings over time for Pulp Fiction

We can see that the ratings start in 1996, but Pulp Fiction was released in 1994. This tells us that the ratings were collected from 1996 onward. We can see that initially there is a spike in ratings when they began collecting ratings. This peaters off and has a small spike in 2016. These datapoints are aggregations over 2 month periods, so a small jump is not unlikely. From doing some research I can't find anything related to the movie that may have caused this so it seems to be a natural occurance by chance.

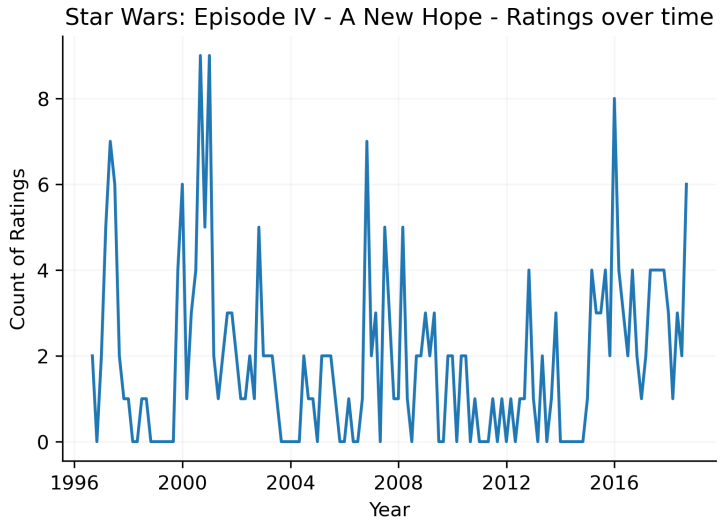


Figure 10: Ratings over time for Star Wars: Episode IV - A New Hope

The above plot of Star Wars: Episode 4 ratings over time can be seen to have several spikes over the years. The movie was originally released in 1977 which is before the ratings had begun to be collected, so the plot starts in 1996. We can possibly attribute these spikes to the release of subsequent Star Wars movies which lead to people re-watching Episode 4 and rating the movie. The spikes in 2002 and 2016 seem to coincide with the release of Attack of the Clones and Rogue One respectively. The spike in 2007 doesn't seem to coincide with any movie releases.

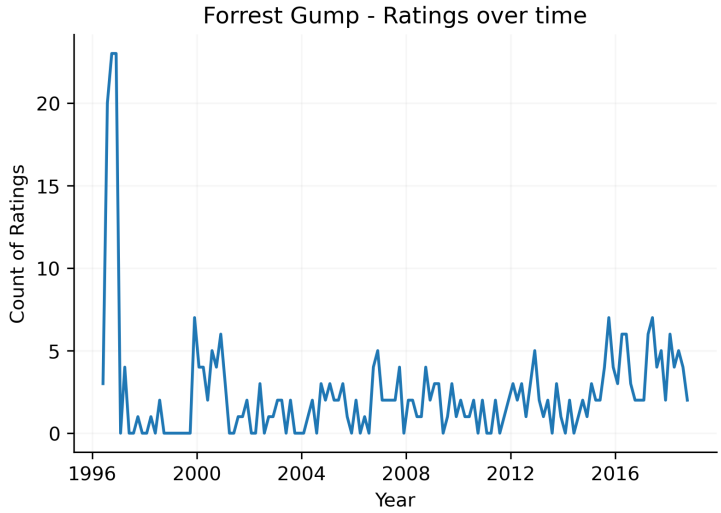


Figure 11: Ratings over time for Forrest Gump

With Forrest Gump we can see what we would typically expect with regards to ratings for movies. Forrest Gump was released in 1994 and we can see a large number of ratings as early after the release as ratings are collected. After this the curve drops and doesn't fluctuate much.

4 Conclusion

Overall, this assignment has showcased how Apache Pig and Hive can be used in the Hadoop ecosystem for processing big data. Once the syntax is understood and the Hadoop cluster is set up, running Pig and Hive is very easy, powerful and can utilize the distributed clusters using more readable languages than Java, which the default MapReduce uses. Developing skills to clean, process and query data can only done through carrying out the steps repeatedly. This assignment has helped me to brush up on these skills and get experience using a distributed framework for parallel processing that is essential for processing the large volumes of data that exist today.

From our analysis we have learnt that users tend to rate with whole numbers more frequently and are likely to give movies a better than average rating. We also looked at the ratings in each genre and have discovered that Horror has lower ratings than the other genres on average. The ratings for each movie can be plotted over time to see if an increase in ratings, due to the movie being relevant, coincides with news relating to the movie or the franchise. Lastly, we discovered some smaller insights about the dataset with regard to users and movies average ratings.