Analysis of movie data to determine the importance of directors to the success of the movie

Data Intensive Architecture

Idris-Animashaun Ayoola - x20103689

MSc Data Analytics – Jan 2021

National College of Ireland

I. ABSTRACT

This paper highlights the use of map reduce paradigm to filter large datasets to prepare them for analysis. The paper involves the IMDB dataset which consists of millions of movie titles dating as far back as 1800 to show how the map reduce paradigm can be used in big data analytics.

In this paper, the reader will see how chaining map reduce jobs helps in processing and filtering large datasets and how java data structures like tree map can assist map reduce in delivering on processes.

Lastly, the reader will see the answers to the research question posed for study under this project.

Keyword: Map reduce, java, tree map, IMDB.

II. MOTIVATION

A movie ends, the credits roll up, and the audience streams out of the theatre or nowadays, grope blindly to the light switch. Everyone is talking about how the protagonist was passionate about a cause or how the movie would have ended. This is usually the features of an almost perfectly executed script directed and acted to a world class performance.

Behind this glamour of awe, there is a matrix of decisions and long hours planning what direction a title should take. What colour to set the mood, who to cast and what to budget for a particular scene. According to Studiobinder, a studio software service,

"A film director is a person who directs the making of a film by visualizing the script while guiding the actors and technical crew to capture the vision for the screen. They control the film's dramatic and artistic aspects." – Studiobinder, [1]

Among other things like word of mouth [8] that has an impact on how titles could fare in the market, creative telling of a script is critical to the success of any title. A study by Dhir & Raj lists factors like actors/actresses involved in the film title, director, time of release of the film title, background story as issues movie success relies on.[2]. The motivation of this study is to understand if truly directors have any correlation to the success of film titles as stated in [3] by Gao et al.

This is a very important study as more investors are looking to the entertainment sector to invest into due to the recent rise in streaming platforms[5][6][7]. Before the pandemic of 2020, the US box office recorded a whooping spend of $100billion, while the UK entertainment scene had its highest year on record [4]. As the world returns to normalcy post-COVID-19 and as more streaming platforms come out, investors are beginning to look at investing in the sector once again [5][6][7], there will be a need for more and fresh content to offer viewers. Even Netflix made sure to cater for this market by offering its Netflix Originals and securing its future as a content owner[19]. The entertainment sector is booming, therefore it is important for any investor to do due diligence into the proposal of any film title that will be invested in. research question

This study will take a look at the film title data put together by IMDB, a global movie site with film titles dating as far back as 1800s, and find out who the top directors are over the many years of film title making. By combing through this data, it is the hope of this study to find traits or patterns that come to light about top rated film titles and their directors. By knowing this, investors can know what to look for when doing due diligence into the directors of the film title proposals they get. The study does not say directors alone account for the success or failure of a film title, it simply wants to help investors and film enthusiasts appreciate the role of directors and put emphasis on what kinds of director to look at based on the success of previous directors. A further study into the directors that will come out of this study can be done to understand deeper the traits that have helped them achieve high rated film titles. The specific questions that will be answered by this study are;

1) **Who are the top directors in the last 70 years?**

By comparing the highest rated film titles with their directors, we will be able to group the directors with the highest rated film titles for the years involved

2) **What genres i.e. action, romance, drama, comedy are they from?** Some genres perform better than others, our analysis will show what genres the top directors work in.

3) **What types of film title do they produce i.e. series vs movies?** Our data set contains movies and series, we want to know the top directors in series and movies.

III. DATA

Our data is taken from IMDB. IMDB is a project started by Col Needham with the aim to store movies he had watched since his teenage years. This project soon got the attention of others and is currently storing over 5 million film titles from around the world. The company was acquired by Amazon in 1998 for $55million and is said to be the largest film title website on the internet with over 190 million monthly users. [18]

The IMDB data collection we will be working with has seven folders in it, each containing data about film titles like their market title name and other title names, what year it was released, when it stopped running (if it was a series), who was on it, who worked on it (which we are interested in), what the rating is and more. Each folder contains a single data TSV file with over one million rows. The primary key connecting the data files is the *'tconst*' key which is the alphanumeric unique identifier of the title. This key will help us join the various datasets together to run our analysis. Another key of note is the *'nconst'* key which is the alphanumeric unique identifier for each person involved with

a title which will help us tie high rated titles to their directors.

IV. METHODOLOGY

The data in this project is stored across seven TSV files with unique keys connecting the records. To get our insight and answer our research questions, we will carry out a map reduce process to join the documents logically, and filter our answers.

From observation, we can see that there are numerous empty fields which can impact our map reduce process and analysis. We will have to take care of these issues as we write our map-reduce process.

Our map reduce process will happen over a single virtual machine Hadoop installation running on Linux Ubuntu OS platform. To get started, we will make sure our project environment has the latest Ubuntu updates installed. Our local Hadoop environment will be started and our written code to filter through the data will be carried out.

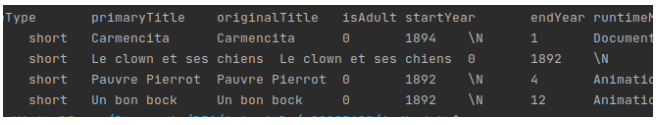
Special if-statements within our code will handle data quality for us by escaping rows that have missing values.

V. IMPLEMENTATION AND ARCHITECTURE

In this previous section, the map reduce paradigm was stated as the process through which we will find the answers to our research questions. In this section, we will describe how we have handled the process.

Our first responsibility was in securing the data from it source, which is freely hosted on IMDB's website for public access. The link provides us with seven download links for each of the TSV files that contain the full movie record of the website.

On downloading the files, the TSV files were extracted and placed in appropriate folders. On visual inspection via the terminal on Linux, (as other TSV opening applications could not process the size), we discover '\N' values within the rows as seen in Figure 1. We can expect that the dataset is dirty and needs cleaning before analysis can be carried out. However, loading the dataset to data cleaning programs like Jupyter notebook (Python), MS Excel proved impossible as the program would crash when loading. This led to using the terminal to view the first 50 lines of each dataset to see its content.



Type	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeM
short	Carmencita	Carmencita	0	1894	\N	1	Document
short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	\N	
short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animatic
short	Un bon bock	Un bon bock	0	1892	\N	12	Animatic

Figure 1 – Missing values within TSV files

Missing values are always expected so since the data could not be cleaned by loading on Jupyter, we would have to manage this within our map reduce code. By taking note of the missing values, it is possible to carefully filter them from the bulk of the dataset.

By having a way to deal with our missing values, we can continue to implement map reduce on our data. The next step involved loading the dataset to the local HDFS. The Hadoop map reduce process runs on HDFS which has been optimized as a distributed file system that can run on multiple commodity hardware systems. The HDFS gives a high through put access to application data and is very suitable for large datasets scenario like ours.

When uploading our data, we will establish an input folder that we will share with our map reduce driver. The map reduce driver is simply a program that co-ordinates the different steps our map reduce program will run on.

For our research, we will have 3 map reduce processes, chained together to help us acquire our answers.

1. **Job1 (Enrich the data)**
   The first job that will run on our map reduce process is a join process that will join together three of the seven tables in our dataset. By this join, we aim to get a broader picture of our movie titles and fill in necessary information like who directed what title? How is the rating for each title? How many people voted for a title and more importantly, allow us calculate the product rating for each title.

   The product rating is calculated by multiplying the ratings allocated to a title by the total number of people who voted for the title. This is to find a uniform way to access the rating for each movie. Finger 2 shows that while some movie titles have high ratings, they have low number of votes. A way to look at this is to ask if a movie title with a high rating of ten and less than twenty votes has performed better than a movie title with a rating of six with over a thousand number of votes. By combining both measures, we hope to consider the average rating of each title with the number of vote it got to get a more robust measure into the performance of that title.

   Our Job1 consists of three mapper classes and one reducer class. Each of the mapper class is linked to a dataset which it pulls values from. Each mapper class has *'tconst'* as its key which makes it easy to reducer the individual mapping process into a single data output stream.

   The mappers contained in Job1 are listed below; *getRatingsMapper, getBasicsMapper, getTitleDirectorsMapper*. The *getRatingsMapper* has the *'titleratings.tsv'* file as its input. The mapper takes each line as input and outputs the *tconst* value as key and a product value which is the product of the average rating column and the number of votes column. The *getBasicsMapper* takes the *'titleBasics.tsv'* document as input and outputs the following values for each movie title; the *titleType*, *primaryTitle*, *startYear*, *runtime*, *genres*. The *getTitleDirectorsMapper* takes the *'titlecrew.tsv'*

file and outputs the *'nconst'* number associated with each title. This number will be cross referenced with the *'namebasics.tsv'* file to get the names of the directors in another job stage.

The output of all three mapper stages are joined within the reducer in a simple reducer join to get a bigger picture into each movie title. Each of the output is joined as a new line by the reducer process in the form;

key ratings output:basics output:title directors

output. Each mapper output is separated from the others by attaching a tracker unique to each mapper process. The reducer process finds this tracker to know what each value for each key it receives is.i.e. if 'R' is the tracker for this value, it is the ratings data, if 'B' is the tracker for this value, it is the basics data and lastly if 'D' is the tracker for this value, it is the title directors data. The trackers are attached during the mapper process for this purpose.

In our driver class, we tell the node manager to store the value of the join process in a folder which we give as input to the next job to work with.

2. Job 2 (Who is the director?)

In Job 2 of our map reduce process, we take the output of Job 1 and enrich it even more by adding the names of the directors (found in the nconst id) for each movie title. In this process, we run another multiple input job that takes values from two files. One is the output file from Job1 and the second is the *'namebasics.tsv'* file. The *'namebasics.tsv'* file contains a nconst id as key and first name, birth year, death year, crew position and array of movie titles. The two mappers involved in this stage is *getdirectorid* mapper and *getdirectorname* mapper. The former mapper takes the output of Job1 which has the director ID and makes that value the key while the remaining values become the value for each line. The *getdirectorname* mapper uses the *nconst* value as key and outputs the *firstName* value belonging to each key. The combined output is reduced with the *nconst* value as key and is given as output by the reducer for this stage. The result is shown in Figure 2 where we have the *nconst* value as key and the *firstName* from the *getdirectorname* separated by a colon from the output of Job 1.

```
nconst  primaryName:
nm0000001       Fred Astaire:
nm0000002       Lauren Bacall:
nm0000003       Brigitte Bardot:
nm0000004       John Belushi:
nm0000005       Ingmar Bergman:nm0000005:I:nm0000005:tt8413310
nm0000006       Ingrid Bergman:
nm0000007       Humphrey Bogart:
```

Figure 2 – Output of getdirectorid and getdirectirname reducer phase

Note the empty columns, they are *nconst* values that are not directors. Our map reduce program is only getting the first name for directors only. The output of this job phase is stored in yet another folder called *'getdirectoroutput'*, this will be the input folder for yet another third job that helps us get closer to our research questions.

3. Job 3 (Get the top 100 movies only for past 70 years ONLY)

In this Job, we do the following activities;

a. Filter the data by getting the rows for movies for the past seventy years only i.e.movieyear > 1950
b. We remove rows with '\N' values
c. We remove rows without any ratings value
d. And we get the top 1000 movies by their product rating from all movies within the past seventy years filtered in (a).

We employ the use of a single mapper called *top100rated* and single reducer called *top100ratedreducer* in this Job to accomplish this task.

Fom the past two jobs we have added and built a composite view of our data. We have joined ratings with directors with titles and their genres, start year and more. The mapper in this job simply filters out what we don't need to answer our research questions and the output sent to the reducer phase contains only the top 100 movie titles arranged according to their product ratings.

To get our 100 values,we make use of a Tree map data structure in Java. The tree map structure allows us create a 100 limit which is sorted on input of any new values. What we do is send our product ratings value as the key and the movie details as values into the tree map. The tree map helps us sort the product values according to the top 100 which is the size we want. By using our product ratings as key, we can get the top 100 top rated movie titles to work with and thus, we have filtered our composite data joins from the previous two jobs.

The output of the third job will go into a Jupyter notebook where the answers we seek will be revealed for our enlightment.

VI. RESULTS

The result to each of the research question we set out to answer is as follows.

A. *The top ten directors in the last 70 years according to product rating*

```
In [52]:
#Get top ten directors based on product rating
imdbdatatmp.sort_values(by=['Product Rating'], inplace = True, ascending = False)
imdbdatatmp.head(10)
```

Out[52]:

	Title ID	Title Type	Title Name	Start Year	Title Duration	Genre	Product Rating	Director ID	Directors First Name
99	tt1375666	movie	Inception	2010	148	Action,Adventure,Sci-Fi	18266170	nm0634240	Christopher Nolan
98	tt0434409	movie	V for Vendetta	2005	132	Action,Drama,Sci-Fi	8491739	nm0574625	James McTeigue
97	tt0110357	movie	The Lion King	1994	88	Adventure,Animation,Dramaloop	8039529	nm0021249	Roger Allers
96	tt1431045	movie	Deadpool	2016	108	Action,Adventure,Comedy	7246544	nm1783265	Tim Miller
95	tt1454468	movie	Gravity	2013	91	Drama,Sci-Fi,Thriller	5935814	nm0190859	Alfonso Cuaron
94	tt0796366	movie	Star Trek	2009	127	Action,Adventure,Sci-Fi	4566642	nm0009190	J.J. Abrams
93	tt5052448	movie	Get Out	2017	104	Horror,Mystery,Thriller	3816366	nm1443502	Jordan Peele
92	tt0234215	movie	The Matrix Reloaded	2003	138	Action,Sci-Filoop	3725366	nm0905152	Lilly Wachowski
91	tt2306299	tvSeries	Vikings	2013	44	Action,Adventure,Dramaloop	3683738	nm0935395	Katheryn Winnick
90	tt0945513	movie	Source Code	2011	93	Action,Drama,Mystery	3643710	nm1512910	Duncan Jones

Fig 4 – top 10 directors

B. *What type of genres do they work on?*

	Title ID	Title Type	Title Name	Start Year	Title Duration	Genre	Product Rating	Director ID	Directors First Name
99	tt1375666	movie	Inception	2010	148	Action,Adventure,Sci-Fi	18266170	nm0634240	Christopher Nolan
98	tt0434409	movie	V for Vendetta	2005	132	Action,Drama,Sci-Fi	8491739	nm0574625	James McTeigue
97	tt0110357	movie	The Lion King	1994	88	Adventure,Animation,Dramaloop	8039529	nm0021249	Roger Allers
96	tt1431045	movie	Deadpool	2016	108	Action,Adventure,Comedy	7246544	nm1783265	Tim Miller
95	tt1454468	movie	Gravity	2013	91	Drama,Sci-Fi,Thriller	5935814	nm0190859	Alfonso Cuaron
94	tt0796366	movie	Star Trek	2009	127	Action,Adventure,Sci-Fi	4566642	nm0009190	J.J. Abrams
93	tt5052448	movie	Get Out	2017	104	Horror,Mystery,Thriller	3816366	nm1443502	Jordan Peele
92	tt0234215	movie	The Matrix Reloaded	2003	138	Action,Sci-Filoop	3725366	nm0905152	Lilly Wachowski

Figure 5 – Top genres

It would seem that a blend of genres make for the top titles and not one genre can be attributed but common among the top ten are adventure, sci-fi, and action

C. *What type of titles do they work on?*

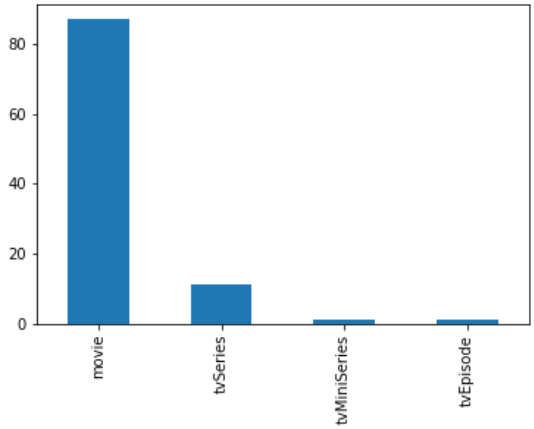


Figure 6 – Top title types

From the above, movie is clearly the favored title type by the top directors

VII. CONCLUSIONS AND FUTURE WORK

In this project, we set out to analyse the movie records archieved by IMDB, a global movie database, for insight into the top performing movies on the globe. The database kept by IMDB covers movie titles as far back as 1800s. We hoped that by knowing what kinds of movies succeed, we will be able to advise investment into movies.

To achieve our goals, we set out to use the map reduce paradigm as it was an efficient way to look through the million rows contained in this archieve. Our map reduce implementation consisted of three jobs with a map reduce process in each to reach our objectives. The output of our map reduce was introduced to a Python environment for the complete analysis of the IMDB data.

The map reduce approach to handling the data was a welcome one as it was difficult to load the datasets to other data processing environments. The Hadoop map reduce environment was able to collect our datasets and process it easily and in a short while even on a single node Hadoop set up. In the future, this single node implementation can be moved to a stack or cloud environment to improve more on the execution time of the process.

More work that can be done involves adding a spark layer to query the hdfs in near real time using a Hive or Pig layer. Lastly, a batch by batch enrichment/fixing of the missing values within the dataset will go a long way to restoring the quality integrity of the results of this data. Due to the missing values, some rows were skipped in order to have successful map reduce runs across the nodes.

REFERENCES

The following papers are referenced in this proposal.

[1] Studiobinder, Producer vs Director: The Roles & Responsibilities Explained (studiobinder.com), Studiobinder, August 29, 2019. [Online]. Available: https://www.studiobinder.com/blog/producer-vs-director/. [Accessed March 1st, 2021]

[2] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2018, pp. 385-390. doi: 10.1109/ICSCCC.2018.8703320.

[3] Gao Z., Malic V., Ma S., Shih P. (2019) How to Make a Successful Movie: Factor Analysis from both Financial and Critical Perspectives. In: Taylor N., Christian-Lamb C., Martin M., Nardi B. (eds) Information in Contemporary Society. iConference 2019. Lecture Notes in Computer Science, vol 11420. Springer, Cham. https://doi.org/10.1007/978-3-030-15742-5_63.

[4] BFI, BFI statistics for 2019 show film and high-end TV generates £30% uplift for UK economy | BFI,BFI, Jaunary 30, 2020.[Online]. Available:https://www2.bfi.org.uk/news-opinion/news-bfi/announcements/bfi-statistics-2019. [Accessed: March 1st, 2021]

[5] Jean Noh, Netflix to invest $500m in Korean content in 2021 | News | Screen (screendaily.com), Screen Daily, February 25, 2021. [Online]. Available: https://www.screendaily.com/news/netflix-to-invest-500m-in-korean-content-in-2021/5157431.article?share=1. [Accessed: March 1st, 2021]

[6] Michael Rosser, UK's Twickenham Film Studios plans £15m refurbishment and expansion | News | Screen (screendaily.com), Sreen Daily. February 4, 2021. [Online]. Available: https://www.screendaily.com/news/uks-twickenham-film-studios-plans-15m-refurbishment-and-expansion/5157063.article?share=1. [Accessed: March 1st, 20201]

[7] Michael Rosser, UK inward investment rallies to near record level in last quarter of 2020 | News | Screen (screendaily.com), Screen Daily, February 4, 2021. [Online]. Available: https://www.screendaily.com/news/uk-inward-investment-rallies-to-near-record-level-in-last-quarter-of-2020/5156825.article. [Accessed: March 1st, 2021]

[8] Chiu, Ya-Ling & Chen, Ku-Hsieh & Wang, Jying-Nan & Hsu, Yuan-Teng. (2019). The impact of online movie word-of-mouth on consumer choice: A comparison of American and Chinese consumers. International Marketing Review. ahead-of-print. 10.1108/IMR-06-2018-0190.

[9] Thottathyl, H., Pavan, K.K., Panchadula, R.P. (2020). Microarray breast cancer data clustering using map reduce based K-means algorithm. Revue d'Intelligence Artificielle, Vol. 34, No. 6, pp. 763-769. https://doi.org/10.18280/ria.340610

[10] D. P. Balasaheb, A. Vikram Kishor and T. A. Sudhir, "Performance Improvement of Parallel Programming Model Based on Parameterized Pipelined Map Reduce Approach," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2019, pp. 1-5, doi: 10.1109/ICECCT.2019.8869321.

[11] Hung-chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker. 2007. Map-reduce-merge: simplified relational data processing on large clusters. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (SIGMOD '07). Association for Computing Machinery, New York, NY, USA, 1029–1040. DOI:https://doi.org/10.1145/1247480.1247602

[12] S. Hemalatha and S. Valarmathi, "Efficient Hybrid framework for parallel Resource and task scheduling in the Map reduce programming," 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2016, pp. 1-7, doi: 10.1109/ICCCI.2016.7479961.

[13] R. Bhardwaj, N. Mishra and R. Kumar, "Data analyzing using Map-Join-Reduce in cloud storage," 2014 International Conference on Parallel, Distributed and Grid Computing, Solan, India, 2014, pp. 370-373. doi: 10.1109/PDGC.2014.7030773

[14] S. Saravanan and V. Venkatachalam, "Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture," 2014 Sixth International Conference on Advanced Computing (ICoAC), Chennai, India, 2014, pp. 21-25. doi: 10.1109/ICoAC.2014.7229736

[15] D. Buono, M. Danelutto, S. Lametti, Map, reduce and mapreduce, the skeleton way, Procedia Computer Science, Volume 1, Issue 1, 2010, Pages 2095-2103, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2010.04.234.

[16] González-Vélez, H., & Kontagora, M. (2011). Performance evaluation of MapReduce using full virtualisation on a departmental cloud, International Journal of Applied Mathematics and Computer Science, 21(2), 275-284. doi: https://doi.org/10.2478/v10006-011-0020-3

[17] Astha S, Why was Hadoop written in Java?|LinkedIn, Astha S, Linkedin. [Online]. January 2, 2017. Available: https://www.linkedin.com/pulse/why-hadoop-written-java-astha-srivastava/. [Accessed: March 2nd,2021]

[18] Buster Coen, 7 of Amazon's Biggest Acquisitions Before The Blockbuster Whole Foods Deal, Buster Coen, TheStreet. [Online]. Available: https://www.thestreet.com/technology/as-a-slack-quisition-looms-here-are-amazon-s-7-biggest-purchases-html. June, 16, 2027. [Accessed:March 2nd, 2021]

[19] Felix Salmon, Why Netflix is producing original content, Reuters. [Online]. Available: http://blogs.reuters.com/felix-salmon/2013/06/13/why-netflix-is-producing-original-content/. June 13, 2013.[Accessed: March 2nd, 2021]

[20] New

[21] Heterogeneous Data and Big Data Analytics (sciepub.com)