Analyze Movie Big Data Set for Crowdfunding

Swathi Badicole
Big Data Tech and App,Department of Data Analytics
San Jose State University
ORCID:015258360
California, United States of America
Email ID :swathi.badicole@sjsu.edu

Smita Kulkarni
Big Data Tech and App, Department of Data Analytics
San Jose State University
ORCID:0152625
California, United States of America
Email ID : smita.kulkarn@sjsu.edu

Aishwarya Mohan Iyengar
Big Data Tech and App, Department of Data Analytics
San Jose State University
ORCID:015269371
California, United States of America
aishwarya.mohaniyengar@sjsu.edu

***Abstract*—Crowdfunding is a practise of funding a project by raising small amounts from a large number of people which is usually via the internet. In order to get such required crowdfunding, communicating about the project with some facts about the success of the project is important. In our project we are focussing on analysing Big Data sets of movies which helps for crowdfunding of movies. Starting from collecting the raw data files for the movie and IMDB rating of each movie, we will be doing 'Extraction Transformation and Loading' using the AWS services. Final stage would be getting the data stored in the AWS into the Microsoft Power BI and showing the analysis visually. This will be helpful for the people who are interested in movie crowdfunding to make a decision about their investment keeping in mind the trend of previous movies.**

Key words: Crowdfunding, Power BI Analysis, Amazon Prime.

I. INTRODUCTION

"The Unique value of crowdfunding is not money, its COMMUNITY". Crowdfunding is a happening kind of practise for any kind of project which involves a large amount of people but a small contribution of money from them. The primacy about crowdfunding is creating a new public market which connects people with investment opportunities that used to be accessible only by the very wealthy people.

Film making is also a business in which a lot of industrialists would want to invest into. Nowadays, there are many movies which are produced by crowdfunding. In order to get a crowdfund for a business, showing the success of the previous trend is very important. Here we are making an attempt to analyse the Big data set of Movies on Amazon prime with the IMDB ratings in different genres, its director, actors etcAmazon Prime video is the stand alone service which is provided by amazon which is dedicated for movies of different languages and genres. On the other hand, ratings by IMDB (Internet movie database) is also one of the trusted ratings by the audience for any of the movies. As we have targeted the people who are interested in movie crowdfunding, people raising the crowdfunding campaign and general people who are interested as our audience, we have collected different files such as IMDB dataset, ratings dataset, movie industry dataset and so on.

People who have a good idea of a movie but not budget to produce a movie will always try to raise a campaign for crowdfunding, for which they will have to show some statistics of previous success to the public while raising the funds. Here comes the analysis of data being critical. Many investors who fund for a movie need data on the directors or actors or genre past performance. They need data to understand what kind of movie is doing really well and what kind of movie is not. This kind of analysis on any movie dataset can help with crowdfunding. As per the research until now regarding the topic, there is no good project today supporting these kinds of analysis which will help both the movie maker to set up a campaign based on his past records and investors.

Our targeted audience are those who generally have no idea about the technology, they would just expect to see the final result after the analysis. For anyone, seeing the final analysed data in the form of visuals would be understood much better. As there is data from different sources collected, we will have to do an Extraction,Transformation and Loading of the data before doing the analysis. Visualization helps the audience to understand the analysed data quickly and concisely.

Suppose a movie has got a very good review, the credit will be taken by all the crew members of that movie. This becomes a plus point for the director, hero and heroine. Mainly we get their next projects.When the producers are searching for the next project, they see these previous achievements before investing. So here, we will be able to analyse based on the ratings which director, hero the movie is in. Some people would have gotten great responses for their comedy genre but not in the horror genre. We will be able to help them do this by just looking at our end visual.

While we are raising some funds in the crowd, we will have to target a larger number of people for which we will have to satisfy the opinions of larger people which is a critical task. Some people get satisfaction to invest with just the trend from past two years but some will want to explore entire trend from start of career of person in film industry to till date. Analysis becomes a challenge at this point to satisfy all the audience who will be using our analysis in their investment or crowdfunding decision making.

In this project, we are giving an analysis of which movie has got the best rating in which genre and many more. So, the

audience will be able to judge before they are investing. Our analysis can also be a proof support for the people who are planning to start campaigns for a movie crowdfunding. This will also be interesting for the people who watch a lot of movies and are curious to see the trends in different movies and its ratings compared to some of them released at same time.

II. RELATED WORK

Nada Elgendy and Ahmed [1] worked on Big Data Analytics: A Literature review Paper in the year 2014 which describes the basics of Big Data, characteristics of Big Data i.e., Velocity, Variety and Volume. There are many tools which are used to analyze Big data. Here in this paper, we understood the tools that can be used and got an idea about what all methods can be implemented on Big Data. As the project is mainly related to Big Data, this literature paper gave an idea for us from the basics to advance methods that can be implemented. We also got an idea about the different places where the big data can be stored and managed. The processing of Data is also explained well here by stressing the HDFS and MapReduce. After the analysis of Big Data, Decision making is also explained with the Customer Intelligence, Supply chain management, Quality Management, Risk management and fraud Detection from which we got an idea who exactly we can use the analysis for Decision making.

In 2017 [2] , Two authors namely Sonali Vyas and Pragya Vaishnav worked on a paper called A comparative study of various ETL processes and their testing techniques in Data Warehouse. As we had referred and got an idea about the basics of Big data from previous papers. This paper helped us to get information regarding the ETL (Extraction, Transformation and loading) process. In detail we were able to understand the Different extraction processes like Logical Extraction process and Physical Extraction Process. Post the ETL process, checking if the ETL process has been done as per requirement is also important. So, in this paper, we also got to know the different testing techniques and also challenges which need to be faced by testing. From this paper we understood the ETL process and different tools which support it.

Three authors Ruhaab Markas, Yisha Wang and John Tseng wrote the paper called [3] Dare to Venture : Data Science Perspective on Crowdfunding in the year 2019. In this paper, we got an idea of Crowdfunding. The audience which we have targeted are the people who are interested in the trends of movies, crowdfunding campaigns for movie production and also the people who are interested to contribute for crowdfunding. In this paper, we got an idea of crowdfunding by explaining the parts of crowdfunding, comparison between normal funding and crowdfunding and also different business models. This paper explained who data science was used in the analysis of the dataset for crowdfunding. From this paper, we got a clear idea about the importance of crowdfunding and also the perspective of data science on it.

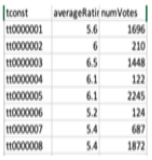
In 2020 [4] , Theo Lynn, Pierangelo Rosati, Binesh Nair and Ciáran Mac an Bhaird have written an article called An Exploratory Data Analysis of the Crowdfunding Network on Twitter. The article was very close to the project we have thought of. This article contains details about crowdfunding and its background. They have also done a small analysis of crowdfunding networks on Twitter. Different analysis was done like the spatial analysis, text analysis, Centrality analysis etc., From this we have got an idea of Exploratory Data Analysis which we could do on our Movie Big Data in the project which helps for the crowdfunding movies.

III. DATA SET

Initially to analyse any information, raw data is very important. Similarly for our analysis which helps the crowdfunding of movies, we needed the raw data. Detail of the movie, Rating for those movies, crew in the movie information was not available at a single source. So, we have got many files from different sources. In order to make our analysis more accurate we decided to take the raw data for ratings of movies from IMDB official website. From the official website of IMDB we have taken 5 different files namely title.basics.tsv.gz,title.rating.tsv.gz,itle.crew.tsv.gz, title.principals.tsv.gz,name.basics.tsv.gz.

tconst	titleType	primaryTitle	originalTitle	isAdult	startYear	endYear	runtimeMinutes	genres
tt0000001	short	Carmencita	Carmencita	0	1894	\N	1	Documentary,Short
tt0000002	short	Le clown et ses chiens	Le clown et ses chiens	0	1892	\N	5	Animation,Short
tt0000003	short	Pauvre Pierrot	Pauvre Pierrot	0	1892	\N	4	Animation,Comedy,Romance
tt0000004	short	Un bon bock	Un bon bock	0	1892	\N	12	Animation,Short
tt0000005	short	Blacksmith Scene	Blacksmith Scene	0	1893	\N	1	Comedy,Short
tt0000006	short	Chinese Opium Den	Chinese Opium Den	0	1894	\N	1	Short
tt0000007	short	Corbett and Courtney Before the Kinet	Corbett and Courtney Before the Kine	0	1894	\N	1	Short,Sport

title.basics.tsv.gz



tconst	averageRatir	numVotes
tt0000001	5.6	1696
tt0000002	6	210
tt0000003	6.5	1448
tt0000004	6.1	122
tt0000005	6.1	2245
tt0000006	5.2	124
tt0000007	5.4	687
tt0000008	5.4	1872

title.rating.tsv.gz

tconst	ordering	nconst	category	job	characters
tt0000001	1	nm1588970	self	\N	["Self"]
tt0000001	2	nm0005690	director	\N	\N
tt0000001	3	nm0374658	cinematographer	director of photography	\N
tt0000002	1	nm0721526	director	\N	\N
tt0000002	2	nm1335271	composer	\N	\N
tt0000003	1	nm0721526	director	\N	\N
tt0000003	2	nm1770680	producer	producer	\N
tt0000003	3	nm1335271	composer	\N	\N
tt0000003	4	nm5442200	editor	\N	\N

title.principals.tsv.gz

nconst	primaryName	birthYear	deathYear	primaryProfession	knownForTitles
nm0000001	Fred Astaire	1899	1987	soundtrack,actor,miscellaneous	tt0072308,tt0031983,tt0050419,tt0053137
nm0000002	Lauren Bacall	1924	2014	actress,soundtrack	tt0037382,tt0117057,tt0071877,tt0038355
nm0000003	Brigitte Bardot	1934	\N	actress,soundtrack,music_department	tt0057345,tt0056404,tt0054452,tt0049189
nm0000004	John Belushi	1949	1982	actor,soundtrack,writer	tt0078723,tt0077975,tt0080455,tt0072562
nm0000005	Ingmar Bergman	1918	2007	writer,director,actor	tt0050986,tt0050976,tt0069467,tt0060827
nm0000006	Ingrid Bergman	1915	1982	actress,soundtrack,producer	tt0077711,tt0038109,tt0036855,tt0034583
nm0000007	Humphrey Bogart	1899	1957	actor,soundtrack,producer	tt0042593,tt0043265,tt0034583,tt0033870
nm0000008	Marlon Brando	1924	2004	actor,soundtrack,director	tt0070849,tt0047296,tt0068646,tt0078788
nm0000009	Richard Burton	1925	1984	actor,soundtrack,producer	tt0059749,tt0057877,tt0087803,tt0061184

name.basics.tsv.gz

After searching a lot of sites for the suitable movie dataset, we found a dataset on Kaggle which was on csv form.

The movies.csv file looks like below which contains 15 columns in it.

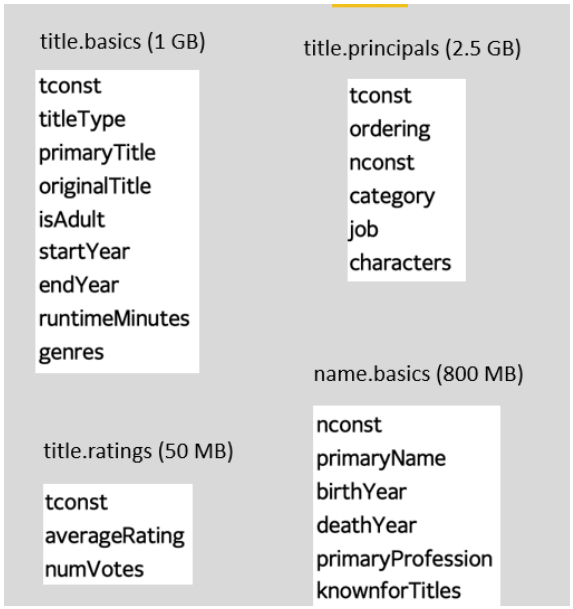
Movies.csv

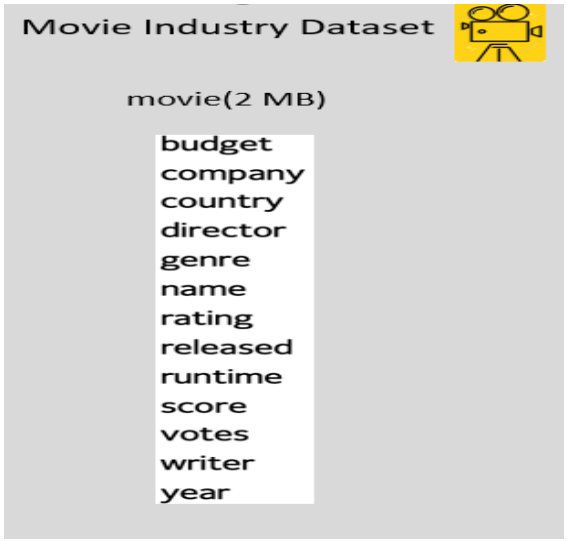
We have used all these 6 files from 2 different sources as a part of raw data in our project.

IV. METHODS

A. *Data Collection:*

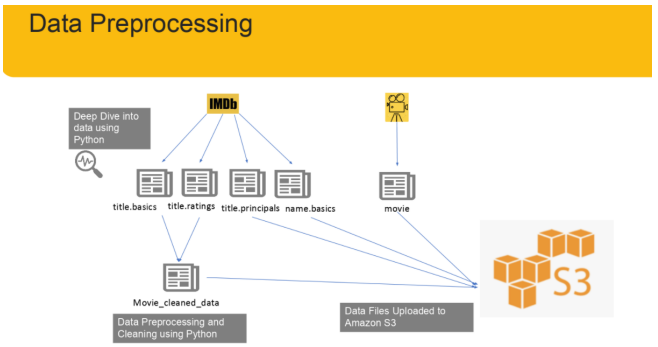
Dataset obtained from IMDB contains Movie review data divided into 7 different files. Out of those 7 files we took 4 files which will help for crowdfunding.Data from kaggle contains Movie Industry data. Considering data which is important we merged all these different files using AWS glue. We created a data frame using Jupyter Notebook with python code. The initial data was in tsv form and we had to join the movie title file and movie ratings file from Kaggle using python script as having two separate files was not useful.





B. *Data PreProcessing:*

The data frame has been checked for null values, non printable values and special characters. Using Jupyter Notebook and python code we removed all the null values. After this cleaning we exported this movie dataset. The data set acquired from Kaggle is a movie industry dataset for which we need to add movie ids, actor ids, director ids, and writer ids to existing data set, this is required as the data warehouse has a fact movie which has two foreign keys: Movie ID and Name ID, and in order to fit movie industry data set into our movie dataset it needs to have unique ids. So, using python script we have successfully added all the ids that are needed and after initial steps our dataset is all set to upload to AWS s3.



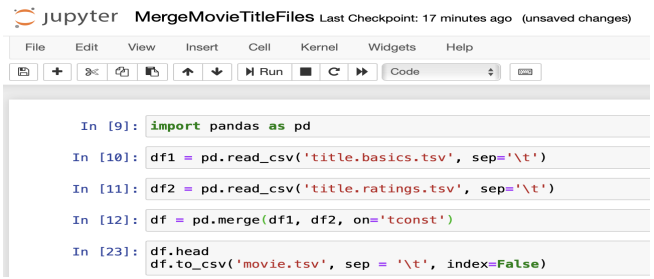


Fig 1: Merging files

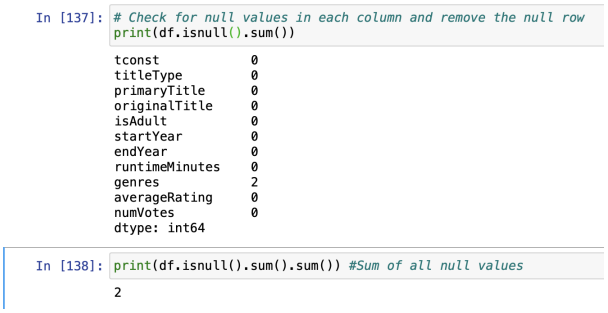


Fig 2 : Python Script(null values)

```
import pandas as pd

df = pd.read_csv('movies.csv',encoding='latin-1')

df.insert(7, 'movie_id', range(1, 1 + len(df))) # Add director Ids as an autoincrement value

df.insert(4, 'director_id', range(900000, 900000 + len(df))) # Add director Ids as an autoincrement value

df.insert(14, 'actor_id', range(100000, 100000 + len(df))) # Add actor Ids as an autoincrement value

df.insert(17, 'writer_id', range(600000, 600000 + len(df))) # Add writer Ids as an autoincrement value

df.to_csv('movie_industry_cleaned.tsv', sep = '\t', index=False) #Export the file as tsv file

df['director_id'].min(), df['director_id'].max()
(900000, 906819)

df['actor_id'].min(), df['actor_id'].max()
(100000, 106819)

df['writer_id'].min(), df['writer_id'].max()
(600000, 606819)
```

Fig 3: Python Script(Adding ids)

budget	company	country	director	director_id	genre	gross	name	movie_id	rating	released	runtime	score	star	actor_id	votes	writer	writer_id	year
8000000	Columbia Pic	USA	Rob Reiner	900000	Adventure	52287414	Stand by Me	1	R	8/22/86	89	8.1	Wil Wheaton	100000	299174	Stephen King	600000	1986
6000000	Paramount P	USA	John Hughes	900001	Comedy	70136369	Ferris Buelle	2	PG-13	6/11/86	103	7.8	Matthew Bro	100001	264740	John Hughes	600001	1986
15000000	Paramount P	USA	Tony Scott	900002	Action	179800601	Top Gun	3	PG	5/16/86	110	6.9	Tom Cruise	100002	236909	Jim Cash	600002	1986
18500000	Twentieth Ce	USA	James Came	900003	Action	85160248	Aliens	4	R	7/18/86	137	8.4	Sigourney W	100003	540152	James Came	600003	1986
9000000	Walt Disney	USA	Randal Kleis	900004	Adventure	18566613	Flight of the	5	PG	8/1/86	90	6.9	Joey Cramer	100004	36636	Mark H. Bake	600004	1986
6000000	Hemdale	UK	Oliver Stone	900005	Drama	138530565	Platoon	6	R	2/6/87	120	8.1	Charlie Shee	100005	317585	Oliver Stone	600005	1986
25000000	Henson Asso	UK	Jim Henson	900006	Adventure	12729917	Labyrinth	7	PG	6/27/86	101	7.4	David Bowie	100006	102879	Dennis Lee	600006	1986
6000000	De Laurentiis	USA	David Lynch	900007	Drama	8551228	Blue Velvet	8	R	10/23/86	120	7.8	Isabella Ross	100007	146768	David Lynch	600007	1986
9000000	Paramount P	USA	Howard Deut	900008	Comedy	40471663	Pretty in Pink	9	PG-13	2/28/86	96	6.8	Molly Ringw	100008	60565	John Hughes	600008	1986
15000000	SLM Product	USA	David Croner	900009	Drama	40456565	The Fly	10	R	8/15/86	96	7.5	Jeff Goldblu	100009	129698	George Lang	600009	1986
8800000	Rimfire Film	Australia	Peter Faimar	900010	Adventure	174650000	Crocodile Du	11	PG-13	9/26/86	97	6.5	Paul Hogan	100010	79465	Ken Shadie	600010	1986
16000000	Thorn EMI Sc	UK	Russell Mulc	900011	Action	5900000	Highlander	12	R	3/7/86	116	7.2	Christopher L	100011	104860	Gregory Wid	600011	1986
6000000	Twentieth Ce	USA	David Seltzer	900012	Comedy	8200000	Lucas	13	PG-13	3/28/86	100	6.8	Corey Haim	100012	12228	David Seltzer	600012	1986
25000000	Twentieth Ce	USA	John Carpent	900013	Action	11100000	Big Trouble i	14	PG-13	7/2/86	99	7.3	Kurt Russell	100013	101678	Gary Goldmi	600013	1986
15000000	De Laurentiis	USA	Michael Mar	900014	Crime	8620929	Manhunter	15	R	8/15/86	120	7.2	William Pete	100014	54000	Thomas Harr	600014	1986

Fig 4: Data set

C. AWS Usage:

We decided to work with Amazon Web Services which is the largest public cloud computing service provider. We used AWS to store the data and merge files which are from multiple sources. AWS is a secured cloud services platform that offers compute power, database storage, content delivery and various other functionalities. It is a large bundle of cloud based services.

D. Amazon S3:

Amazon Management Console is a simple and intuitive web interface. We can accomplish data storage through this AWS management console. We have used Amazon s3 bucket to store and retrieve our data anytime and anywhere. This Amazon Simple Storage Service is storage for the internet. We uploaded our huge data set file to this Amazon s3 bucket which is named as project-input-files.

E. Amazon Redshift:

It is a massively parallel, column-oriented database deployed on the AWS platform that makes it simple and cost efficient to analyse our data all across our data warehouse and data lake. Group of nodes on Amazon Redshift is called cluster. We have created a cluster on Redshift. Used star schema design for data warehouse as shown below.

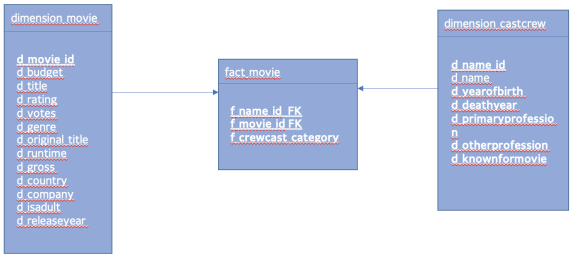


Fig 5 : Star Schema

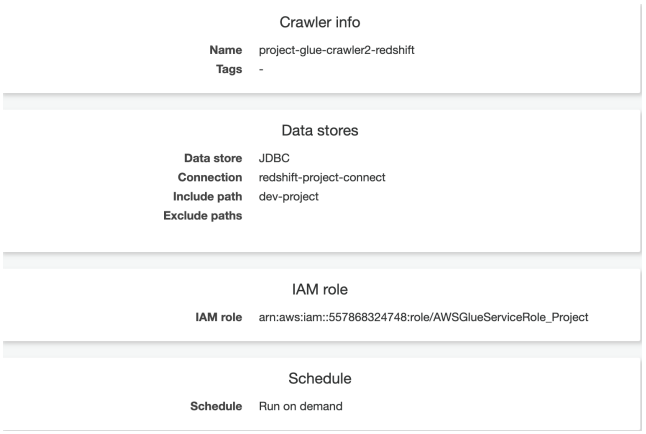
We have created an IAM role which allows us to assign access with specified permissions and trusted entities. This IAM role granted permission to access all s3 and redshift policies.

F. Amazon Glue:

Amazon Glue is a fully managed Extract, Transform and Load which makes it simple and cost efficient to categorize all our data , clean it , enrich it and move it reliably between data stores. There is no set up or management as AWS Glue is serverless. ETL engine automatically generates python code or scala code and a flexible scheduler that handles dependency resolution and job monitoring.

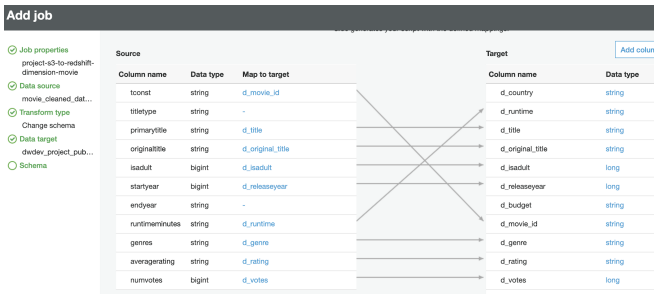
The first step is using the IAM role to create a Glue classifier. This classifier helped us in determining the schema of our data. AWS Glue provides classifiers for common files such as csv, json etc. and also a relational database management system using JDBC systems. We also created a Crawler named project-glue-crawler2-redshift which connects to the data store. This creates metadata tables in the data catalog. Then created a Glue connection to Redshift. We created an Endpoints which helps if the VPC fails and has security issues.

The second step is to run the crawlers. We ran the crawler first for s3 so that database got populated automatically with tables. Then the crawler has been run for redshift and the database got created with redshift tables automatically.

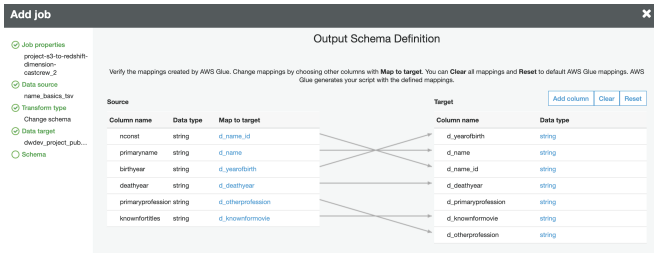


To populate dimension_movie, dimension_castcrew, fact_movie table by mapping the source data attributes to the target attributes we created a job in Glue and named it as project-s3-to-redshift-dimension-movie, project-s3-to-redshift-dimension-castcrew, project-s3-to-redshift-fact-movie alternatively. After running the jobs on AWS Glue our jobs were successfully given results as shown below.

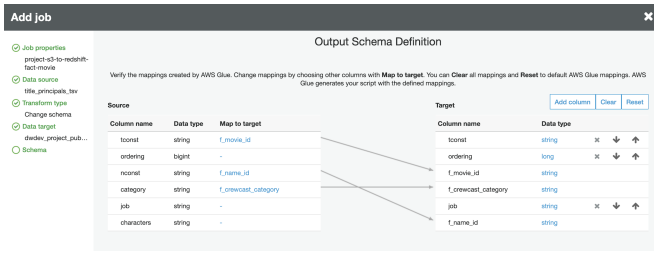
- Mapping for Dimension movie using Glue Job:
  Source S3 movie.tsv
  Target Redshift table dimension_movie



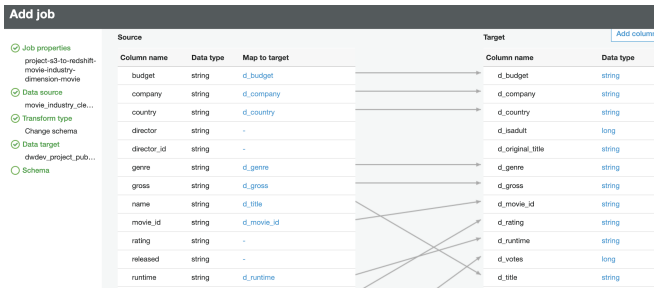
- Mapping for Dimension castcrew using Glue Job
  Source S3 name.tsv
  Target Redshift dimension_castcrew table

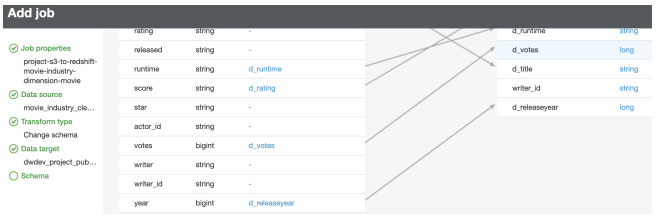


- Mapping for fact movie using Glue Job:
  Source S3 titleprinciple.tsv
  Target Redshift table fact_movie

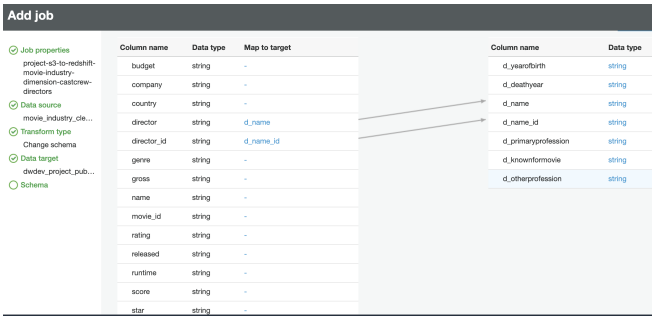


- Mapping for dimension movie using Glue Job:
  Source S3 movie_industry_cleaned.tsv
  Target Redshift table dimension_movie

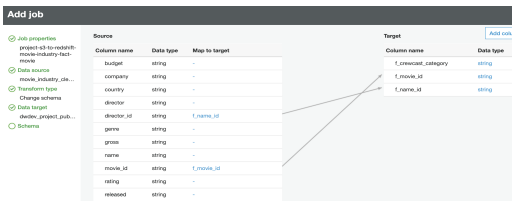


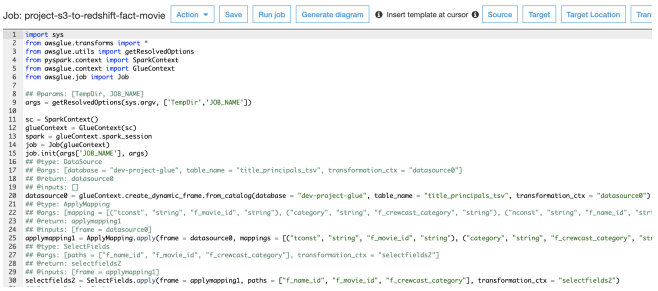


- Mapping for dimension castcrew using Glue Job:
  Source S3 movie_industry_cleaned.tsv
  Target Redshift table dimension_castcrew



- Mapping for fact movie using Glue Job:
  Source S3 movie_industry_cleaned.tsv
  Target Redshift table fact_movie





The Python script got generated automatically when we ran jobs on ETL. We made a few modifications in the script as required for our project. This script contains the code that performs extract, transfer and load work. We can create our own scripts as per requirement in this AWS glue.

Github Link :

https://github.com/Smitashri/DataWarriors

G. *Query Used to Merge:*

The Query to merge from Movie Industry data to IMDB data on redshift is:

```
update dimension_castcrew
set d_primaryprofession = 'director'
where d_name_id >= '900000' and d_name_id
<= '906819'

update dimension_castcrew
set d_primaryprofession = 'actor'
where d_name_id >= '100000' and d_name_id
<= '106819'

update dimension_castcrew
set d_primaryprofession = 'writer'
where d_name_id >= '600000' and d_name_id
<= '606819'


update fact_movie
set f_crewcast_category = 'writer'
where f_name_id >= '600000' and f_name_id
<= '606819'

update fact_movie
set f_crewcast_category = 'actor'
where f_name_id >= '100000' and f_name_id
<= '106819'

update fact_movie
set f_crewcast_category = 'director'
where f_name_id >= '900000' and f_name_id
<= '906819'
```

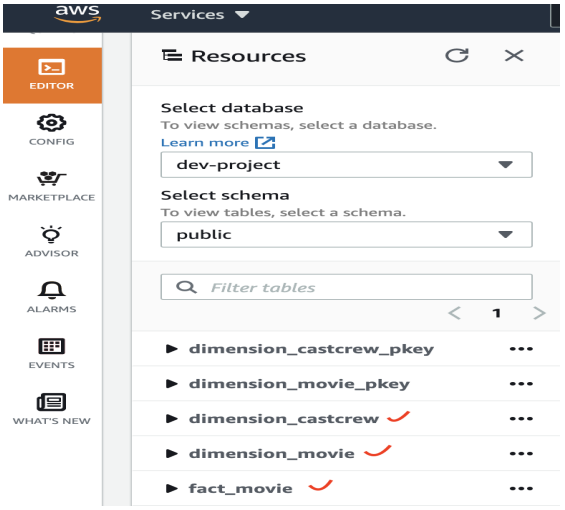
H. *Create tables on redshift:*

We have created a database named dev-project on the redshift and created tables. Populated data into the tables in redshift directly through ETL. Using Redshift we can run most complex queries. For query optimization redshift maintains high performance.

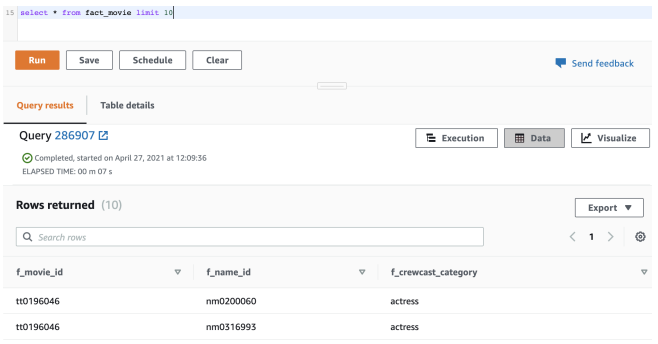
```
CREATE TABLE dimension_movie
(
  d_movie_id VARCHAR(30) PRIMARY KEY,
  d_budget VARCHAR(22),
  d_title VARCHAR(255),
  d_rating VARCHAR(100)
  d_votes BIGINT,
  d_genre VARCHAR(200),
  d_original_title VARCHAR(255),
  d_runtime VARCHAR(40),
  d_gross VARCHAR(50),
  d_country VARCHAR(50),
  d_company VARCHAR(225),
  d_isadult BIGINT,
  d_releaseyear BIGINT
);

CREATE TABLE dimension_castcrew
(
  d_name_id VARCHAR(22) NOT NULL PRIMARY KEY,
  d_name VARCHAR(200),
  d_yearofbirth VARCHAR(20),
  d_deathyear VARCHAR(20),
  d_primaryprofession VARCHAR(200),
  d_otherprofession VARCHAR(200),
  d_knownformovie VARCHAR(225)
);

CREATE TABLE fact_movie
(
  f_movie_id VARCHAR(30) NOT NULL,
  f_name_id VARCHAR(22),
  f_crewcast_category VARCHAR(100)
  foreign key(f_movie_id) references dimension_movie(d_movie_id),
  foreign key(f_name_id) references dimension_castcrew(d_name_id)
);
```

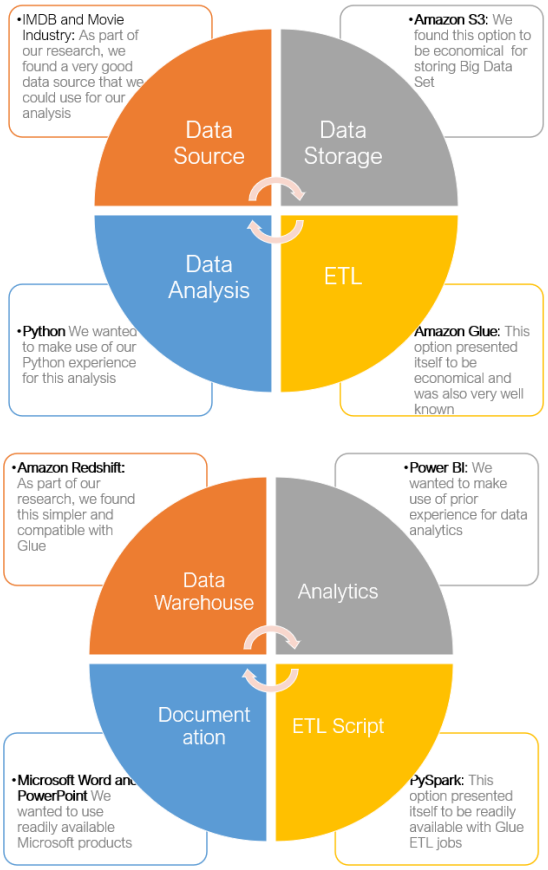


We ran query on redshift.The sample result is shown below



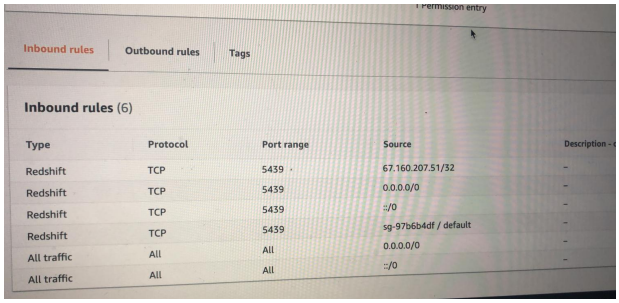
I. *Overview*

As shown below this is the overview of what we have done using amazon AWS for our project.



J. Connecting Redshift to Power BI:

For better insights and analytics we decided to connect the Redshift database to Power BI. Amazon Redshift is a fast, fully managed, cloud-native data warehouse that makes it simple and cost effective to analyze all our data using standard SQL and Business Intelligence tools(BI). To connect our Redshift database to the Power BI tool we have installed Power BI desktop tool. As we have to get data from the redshift database we have to select amazon redshift as a source in BI tool to import data directly into the tool. Then it will ask for server and port details along with the database name. Initially we faced errors while connecting but later after changing Inbound rules ie., we gave access to all the ip addresses as shown below



Later after giving access to all types and all ip addresses we could connect our redshift database to power BI tools for better insights and analytics.

V. RESULTS

Once the data has been loaded into the redshift in the star schema, we wanted to perform some of the visual analysis by creating a dashboard. As this is a crowdfunding based business-related analysis, we choose the Microsoft power BI which is said to be the best for the business analytics. As our database was in redshift, PBI was also compatible with it. Hence, we connected the redshift to PBI for our visual analysis.

We have done different charts, tables, score cards and drop downs. We have come up with two interactive dashboards.

The first dashboard helps the user to select the country, year, rating from drop down which are multiple select. Then the list of movies as per the selection will be displayed in the movie name table. To go even more depth, if people want to see the artists in that particular movie, then the movie name table is also interactive which helps in selecting a particular movie name and displays all the artists in that movie in table two below it.

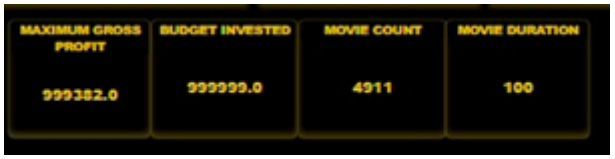
We also have a score card which displays max gross profit of selected movie, budget, movie duration. As we select countries in the dropdown, we will be able to see two score card values, movie count in that country and max rating received among those movies.

These are the dropdowns used in our dashboards. As the dashboard is interactive, users will have to select at least one among these dropdowns to see the results in the tables.

We have also included a geographical map that locates the country once it's selected by the user.

There are four Score cards in the first dashboard which shows the gross profit, budget, movie count, movie

duration. It displays the movie duration of the movie which is selected in the table.



We have wrote DAX (Data Analysis Expressions) for the score cards like finding the maximum, count of movie, movie durations, max rating, max rating with votes etc., We can see max rating which is just the max rating of a movie and max rating with votes is the max rating of movie with particular number of people voting for it. Below is the SS of those score cards and the card which displays the production house for the selected movie.

If the end user wants to select the artist category and then see their performances in different movies and rating of those movies, we have a second dashboard containing the dropdown for choosing the artist(actor,director,writer). We also have a country wise slicer with the artist dropdown. Then it displays all the movies directed by that artist selected in the table and also displays the movie count in the scorecard. After it shows all the artists as per selection, once those details populate in one table, we will be able to select a single name from the table and post selection, we have all the movies directed by that actor in the second table.

The image of two dashboards are below.

VII. DISCUSSION

We acquired the data from two multiple sources ie., from Kaggle and IMDB. After preprocessing our data as it is a huge dataset we thought of storing and performing ETL operation on AWS cloud computing service.

We used Amazon S3 to store our data. Using Amazon Glue we have extracted and Transformed our data and Loaded into the Amazon Redshift. As our database is already in Redshift for better insights and to analyse the data we connected our database in redshift with power BI tool. Our goal is to give a clear picture and make the audience understand our aim and as it should be useful for the audience to invest in movies and help them in decision making we created an interactive dashboard.

Connecting the redshift database to power BI was initially a challenge for us. We tried to connect with the endpoint server to power BI but it threw an error. Initially finding out the error and solving that was tough as nowhere online we were able to figure out why it is throwing connection and what are the ways to fix. Finally after changing VPC inbound rules by giving access to all the IP addresses we were able to connect our redshift database to power BI successfully. Apart from that connection issue as this AWS is all new for all of us we first had to learn from scratch and then started working on a project. Although it was a time taking process we have enjoyed working on this project and learnt new things.

Our project's main aim is to help crowdfunding. People who are interested in investing in movies can get an idea of which movies or directors they can trust by our visualization analysis. As we have created an interactive dashboard it is easy to analyze and get to know about the movie they want to fund in.

VII. CONCLUSION & FUTURE WORK

We have the data of movies from 1874 - 2021 across various countries. If you are interested in crowdfunding you are at the right place. This project helps people to understand which is the right place to invest their money. By Tracking the trending movies and rating people can make their decision of investing in movie crowdfunding.

Our data from IMDB and Kaggle is around 5GB. For this project we decided to work on this dataset and as a future work we will add data from multiple sources like Amazon Prime, Rotten Tomatoes and Netflix.

VIII. REFERENCES

[1]“An Exploratory Data Analysis of the #Crowdfunding ... - EconPapers.” econ papers. https://econpapers.repec.org/RePEc:gam:joitmc:v:6:y:2020:i:3:p:80-:d:412088 (accessed: May 19, 2021).

[2] “Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi, A. van den Hengel *SIGIR*, 2015”

[3] “Inferring networks of substitutable and complementary products J. McAuley, R. Pandey, J. Leskovec Knowledge Discovery and Data Mining, 2015”.

[4]:https://scholar.smu.edu/cgi/viewcontent.cgi?article=1080&context=data science review - ‘Study of Crowdfunding’.

[5]:https://www.researchgate.net/publication/264555968_Big_Data_Analytics_A_Literature_Review_Paper?_sg=QdVKPLFc0L3Slw2bLshmh7KpuzjnnA7q7NuqHvzW4P3JkxiGs020OZZvvQwEVAUI8gNb56s5gyPxnIk - “Comparative study of Various ETL processes”.