Internet Movies Database (IMDb)

**By: Wizards at Work**

Yogita Suryavanshi (015274077)

Preeti Khatri(015349321)

Saroj Saran(015276014)

Hrushikesh Pokala(015261259)

**Abstract:**

The Internet Movie Database (IMDb) is a website that acts as a global film database. This website provides a wealth of public information about films, including the title, the year of release, the genre, the audience, critics' ratings, the length of the film, the synopsis of the film, actors, directors, and much more. Given the vast amount of information available on this site, we thought it would be interesting to examine the data on movies on the IMDb website from 2000 to 2017.

Our project focuses on such IMDb datasets, with one dataset focusing on 45,000 movies and containing 26 million ratings from 270,000 users on a scale of 1-5, and the other dataset focusing on 45,000 movies and containing 26 million ratings from 270,000 users. This dataset includes films that were released on or before July 1, 2017. This dataset has been normalized and added to the Ohio S3 bucket. We also took another dataset with 6 million records, which we put in the Oregon S3 bucket, and created a single cluster in Ohio where the tables for all the datasets were created.

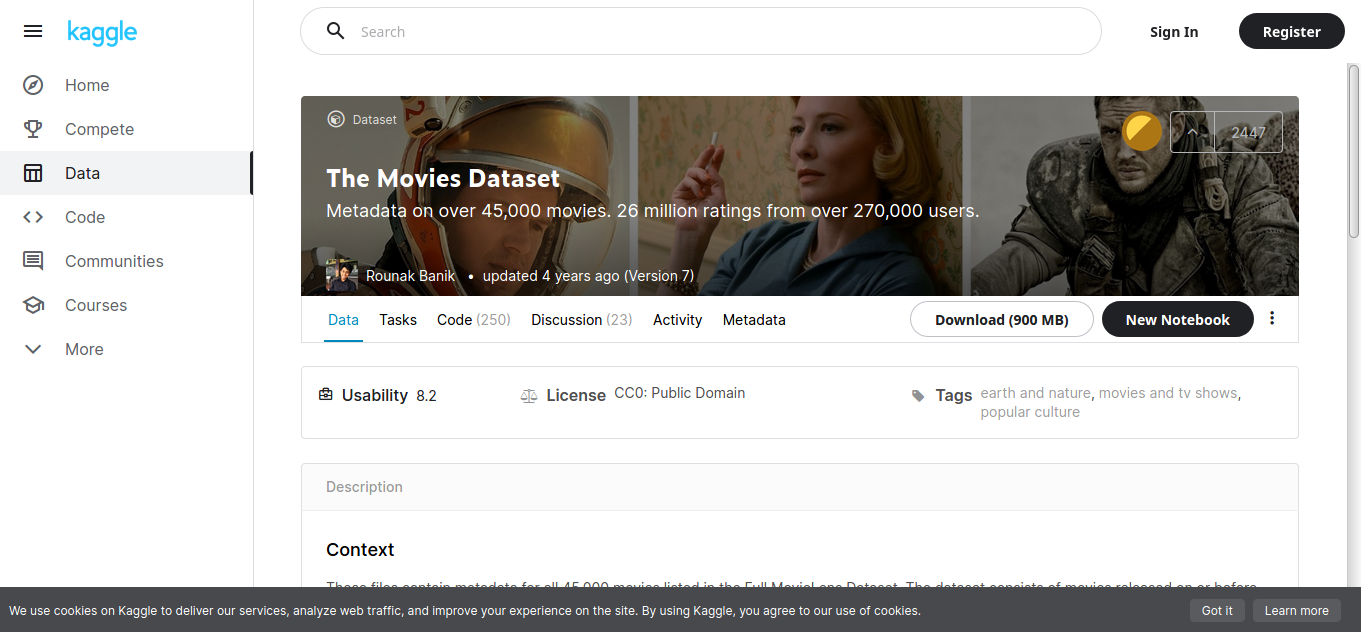
We have done cross-region replication i.e. when any file is changed or loaded with additional information, be it the Ohio dataset or the Oregon dataset it automatically updates the data in the Ohio cluster, which is used for further processing and analysis.

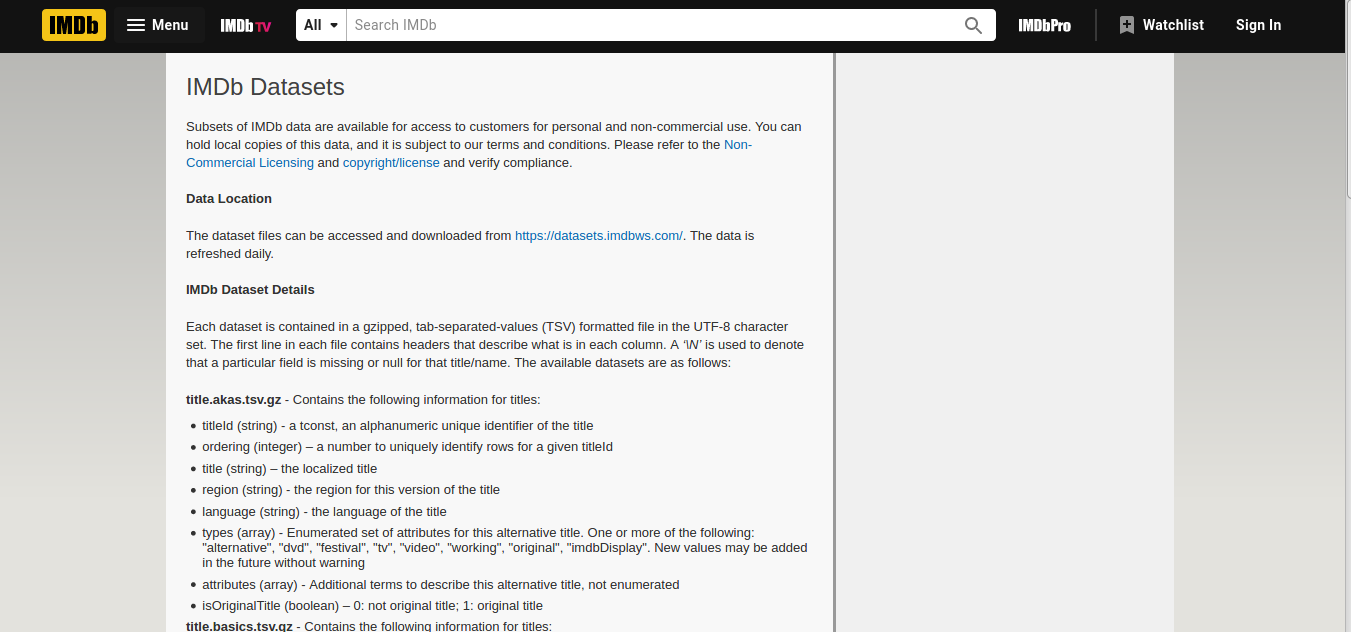
**Objective:**

Our project aims to analyse the movies data for two different data sources, one with a data length of 6 million IMDB data and the second data source of 45,000 movies with various graphical representations to give an interpretation of these movie datasets.

**Data sourcing:**

We have sourced our data from two different datasets. The first dataset is from Kaggle which involves metadata of 45000 movies. This data set is normalized. The other dataset is from IMDb which involves a data length of 6 million and it is a denormalized dataset.

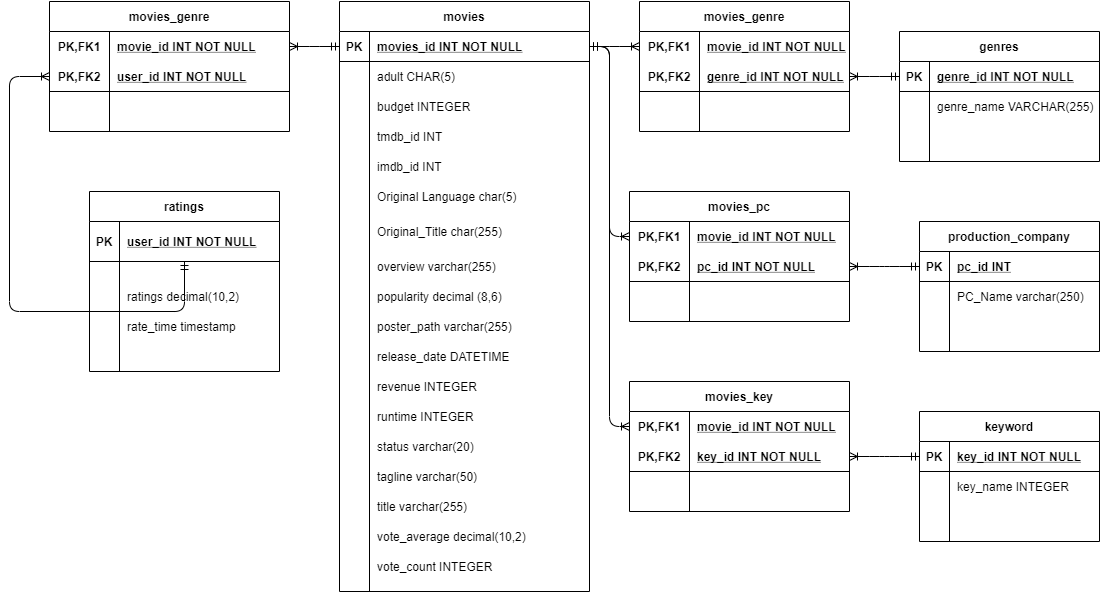




Few data files in data source 1 was in JSON format and the rest was in csv format.

**Data modelling:**

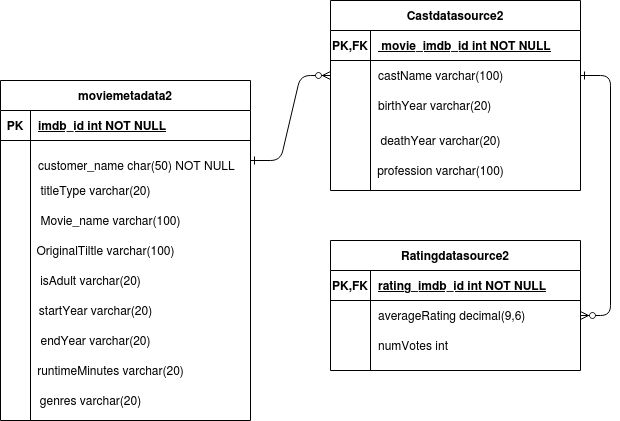
**ER Diagram For First Datasource:**

****

We have the following major entities in our first data source:

1. **Movies,**
2. **Ratings**
3. **Plot Keywords**
4. **Movie Genre**
5. **Movie**
6. **production Company**

**ER Diagram For Second Datasource:**

****

We have the following major entities in the second data source:

**1.Moviemetada**

**2.Castdata**

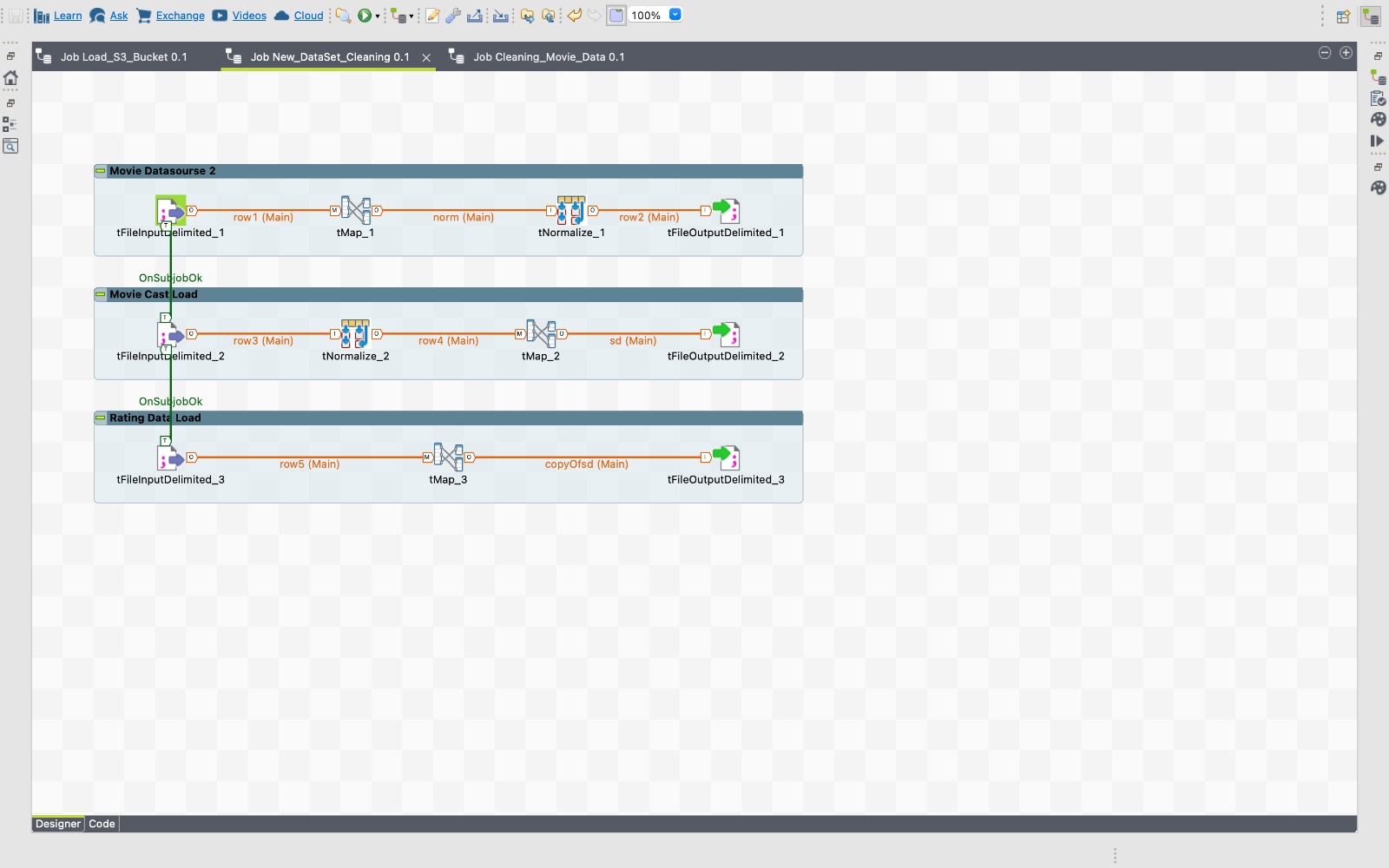
**3.Ratingdata**

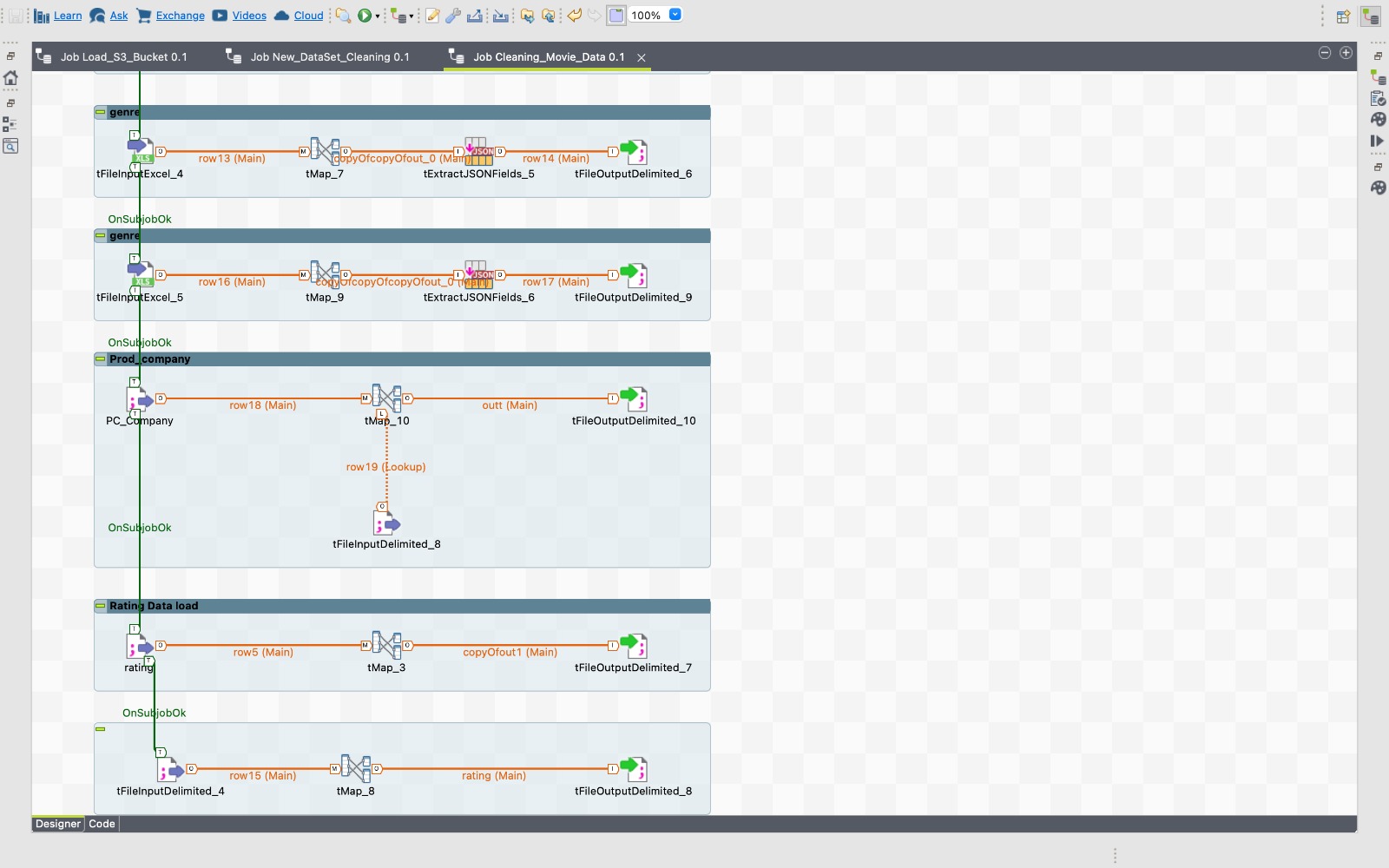
**Data cleaning:**

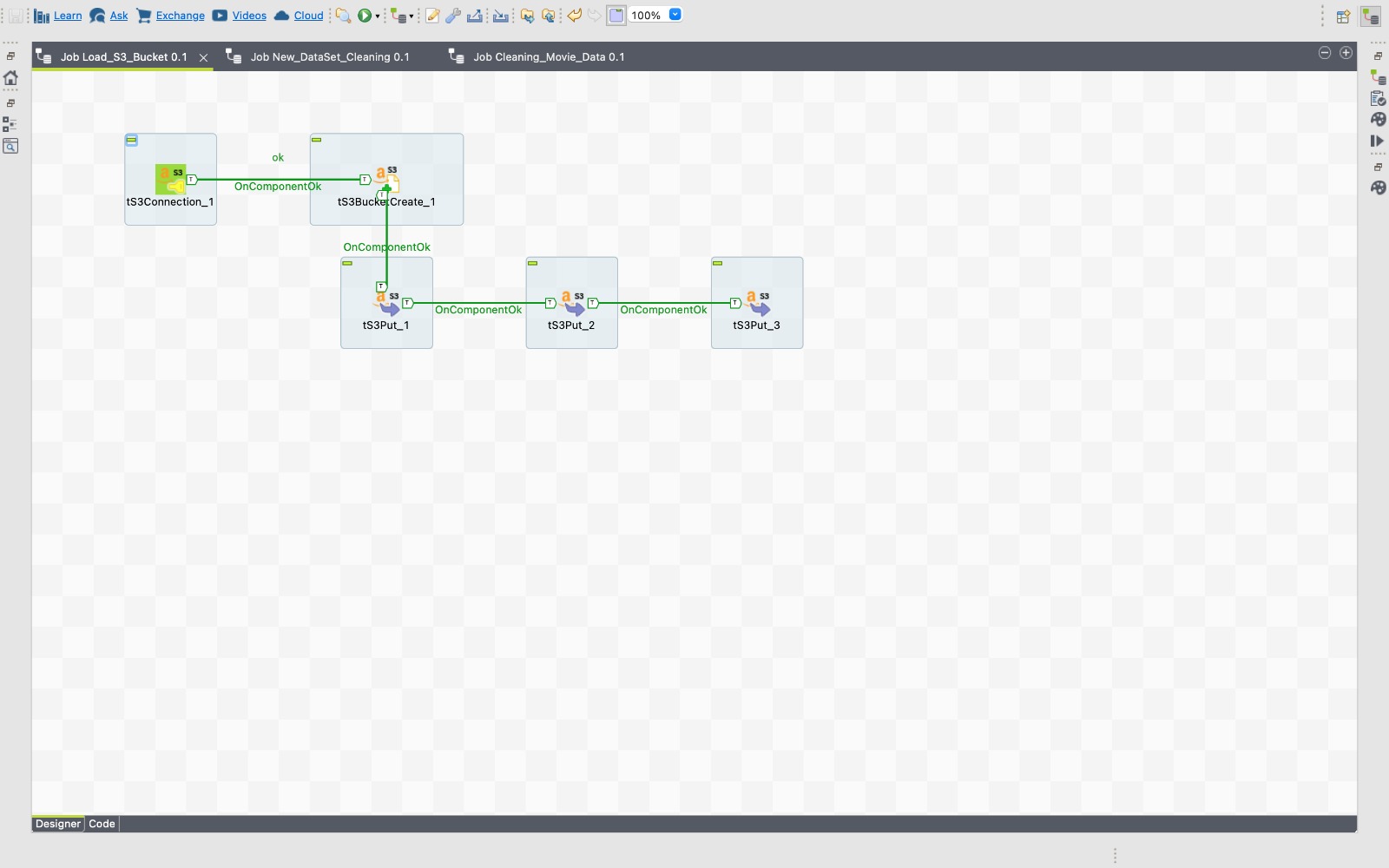
As we have two different data sources, data source 1 was of 45000 movie records which were cleaned and transformed using ETL and then generated CSV was loaded to S3 bucket to create tables in Clusters.

Whereas data source 2 consists of millions of IMDb records that were cleaned, transformed, and loaded to the S3 bucket directly.

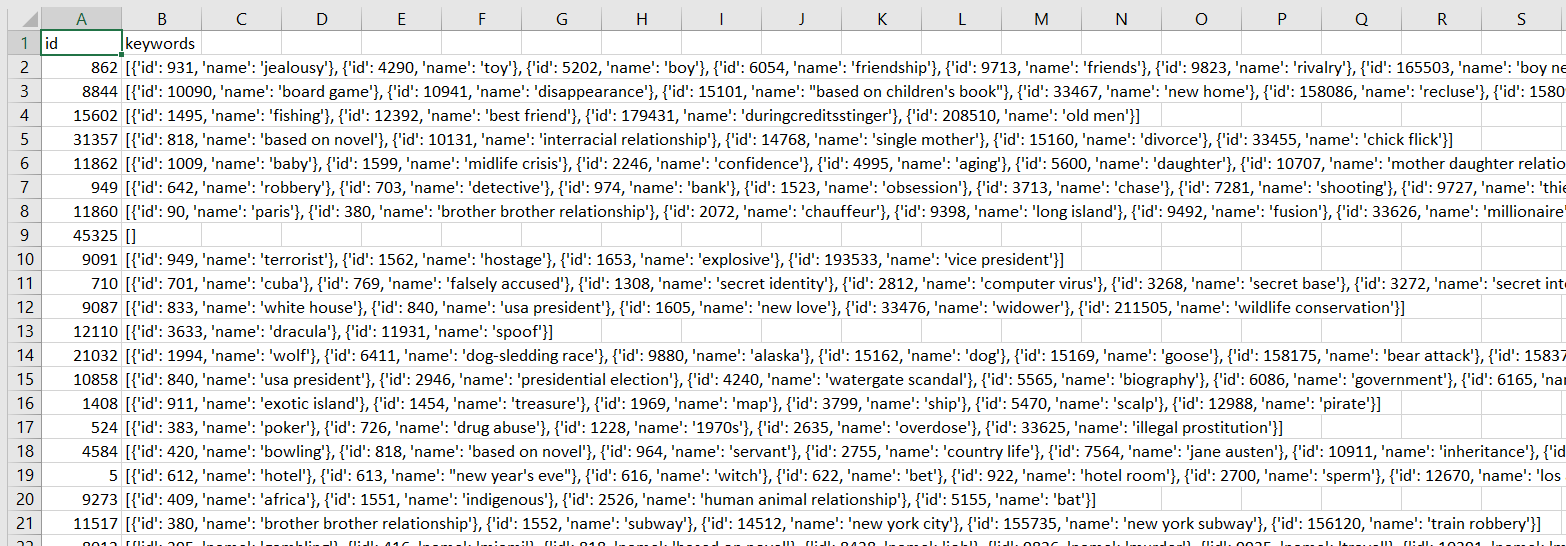
Talend has been used for data extraction and transformation. Some of the data cleanings have also been done with Talend jobs. Job has been created in such a way that it takes all the data CSV file as input and transforms it according to requirement and loads it in the Amazon S3. Talend job has been loaded in the repository.

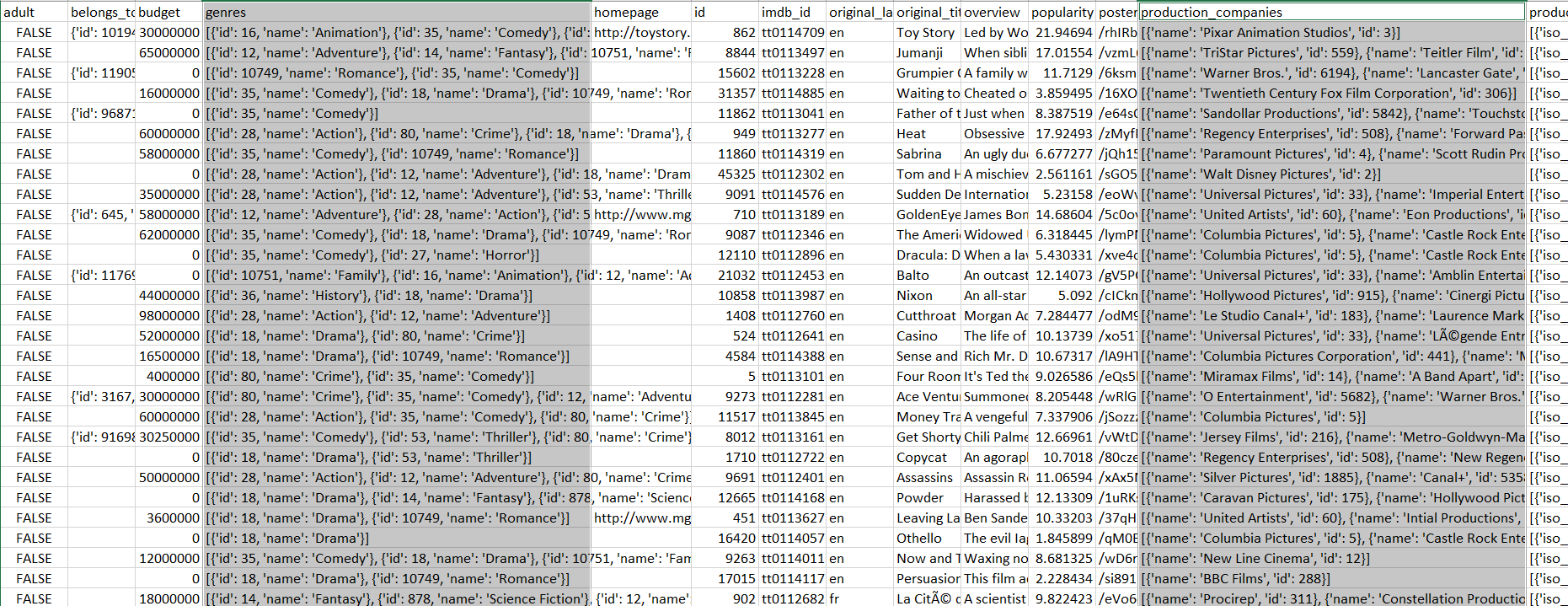






Few of the files in data source 1 had data in **json format**, which we had to handle as part of data cleaning.

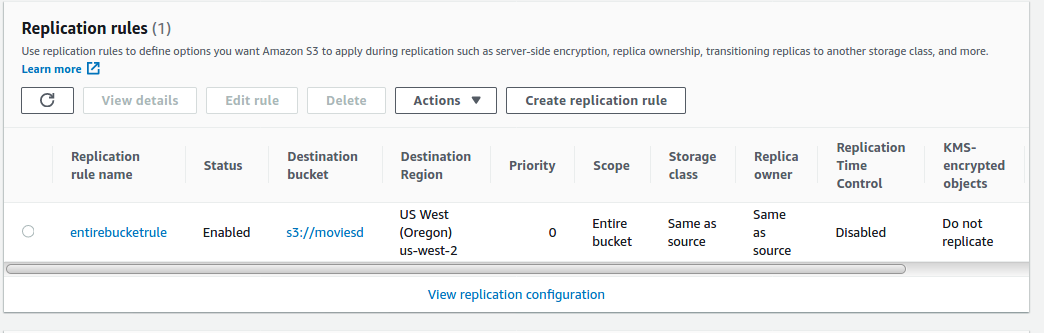
****

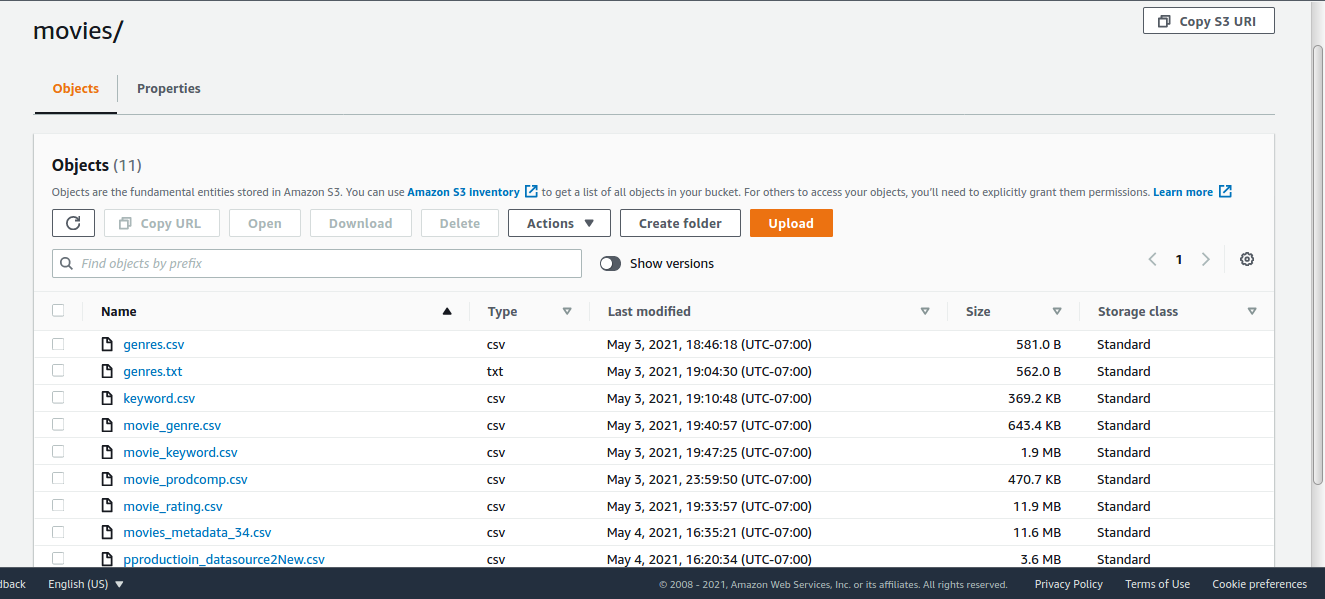
****

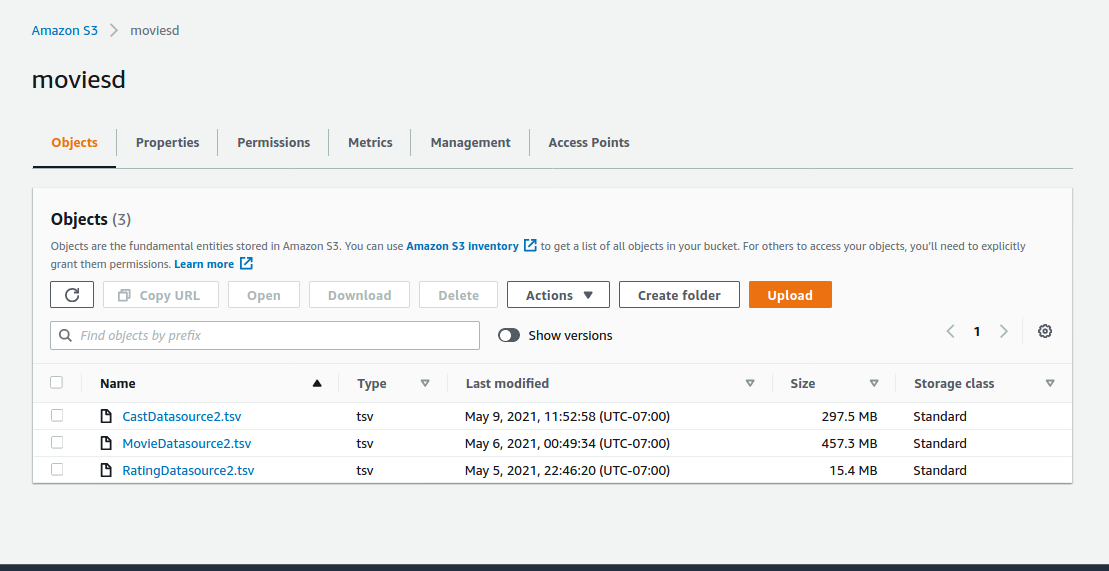
**Data loading and Processing :**

We have loaded the data files from different sources into two different s3 buckets in different regions via Talend Etl Tool. This is done by enabling cross-region replication which allows you to replicate data between distant regions to satisfy requirements and minimize latency. Cross-region replication allows you to set up a replication bucket, where when an object is uploaded to a specific S3 bucket it automatically gets updated into the replication bucket that is in a different AWS Region.

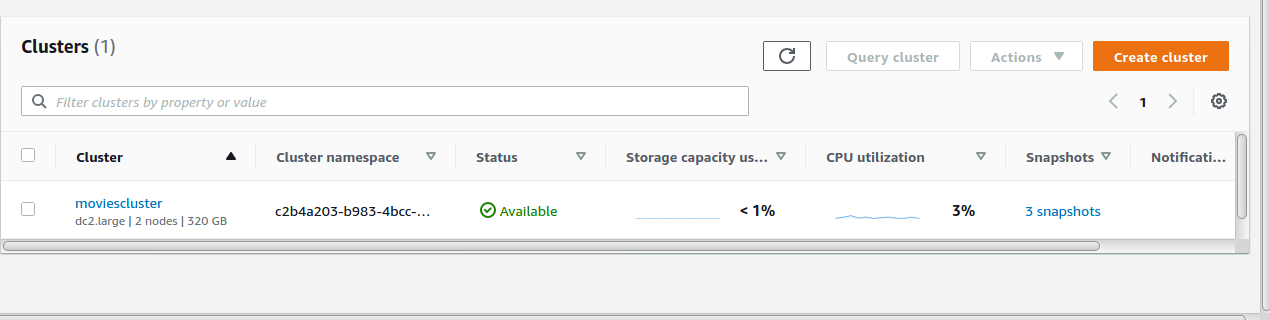


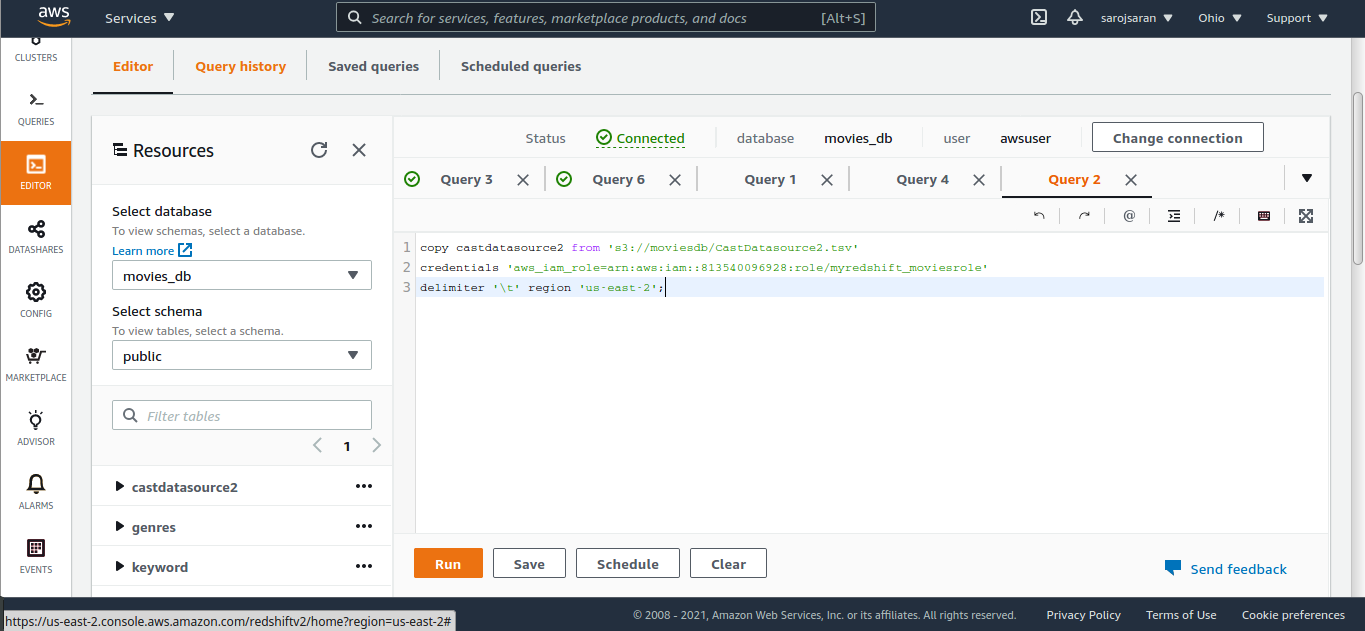






By using Redshift Cluster we have created the tables and have loaded the files into them using the Redshift query editor.





**Data Analysis and Visualization :**

Now that we have cleaned and loaded the data to clusters. Next, we have extracted the data for analysis and visualization.

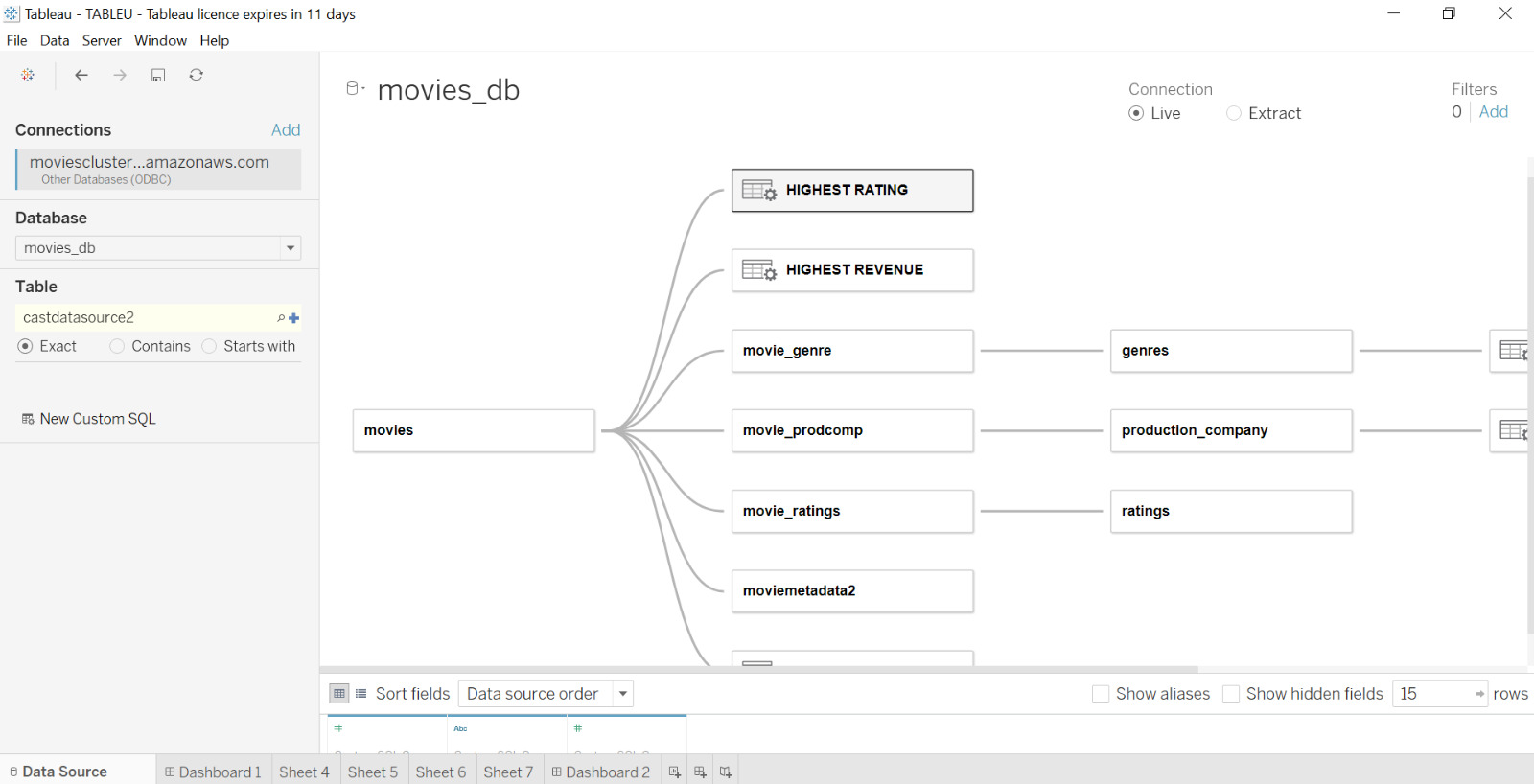
We have tried two platforms :

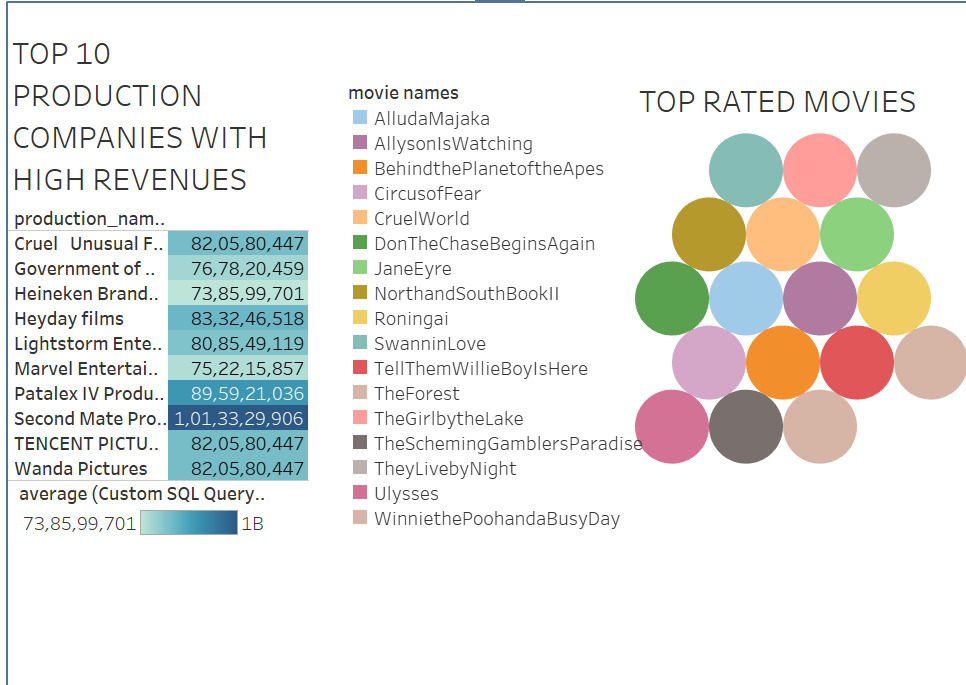
* Tableau
* Python

1. **Tableau**

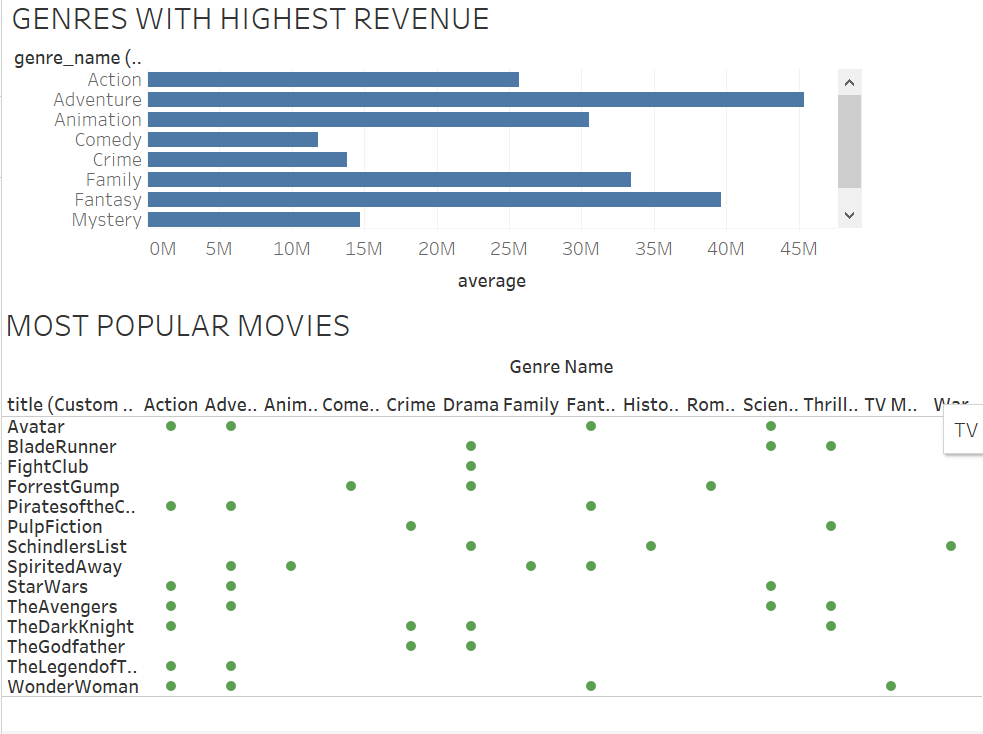
We have connected Tableau to our Redshift Cluster and have conducted the analysis. We have connected Tableau to Redshift Cluster with the help of the ODBC Driver.

**Visuals are as below :**





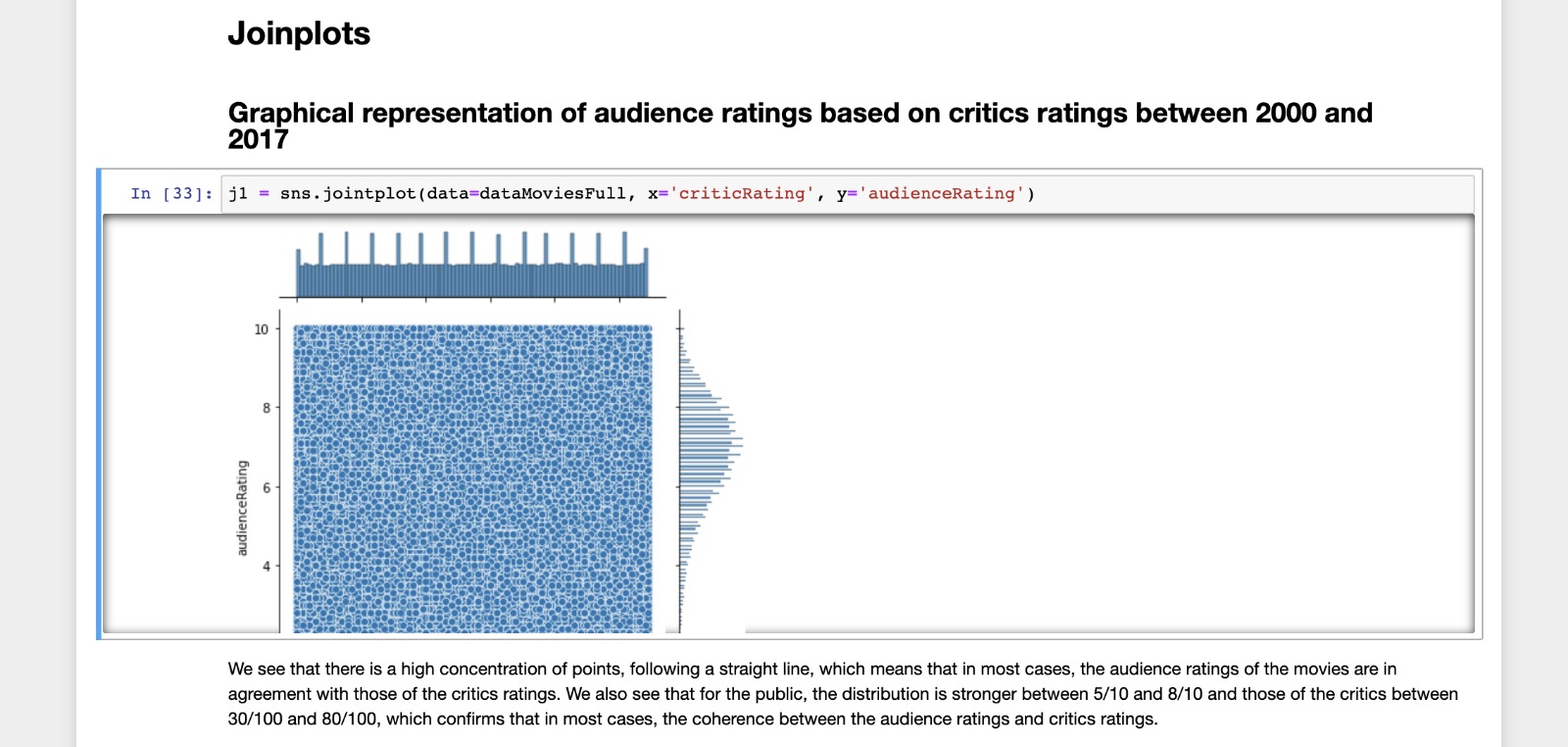
We made an analysis of the top 10 production companies with high revenues and we have found that Second Mate Productions have made approximately around 1 Billion. We have also found the top-rated movies where the rating scale is from 1-5 and we have chosen the movies above rating 4.5



We have found the most popular movies with respect to their genres and also the genres that have made the highest revenues. This can be used for predictive analysis where users can see what type of users are more likely to watch what kind of genre movies, this will help production companies to make such genre-related movies.

1. **Python :**

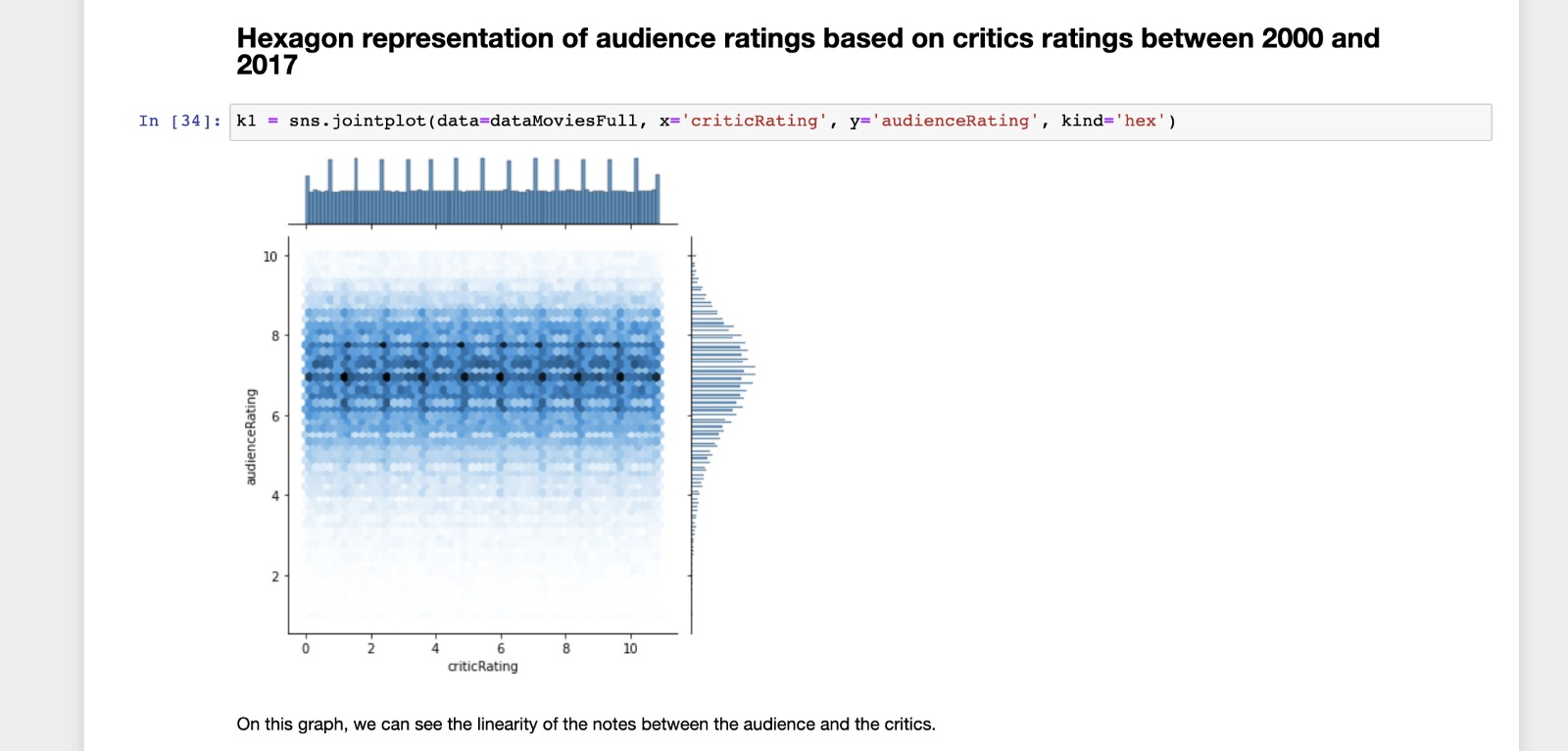
Python stack: Numpy, Matplotlib, Pandas and Seaborn



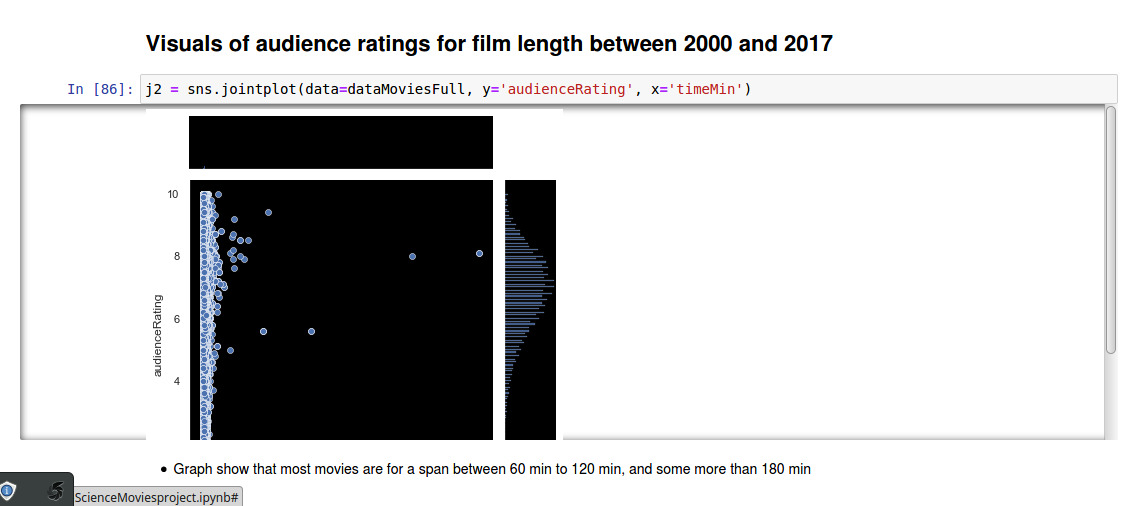
We see that there is a high concentration of points, following a straight line, which means that in most cases, the audience ratings of the movies are in agreement with those of the critic’s ratings. We also see that for the public, the distribution is stronger between 5/10 and 8/10 and those of the critics between 30/100 and 80/100, which confirms that in most cases, the coherence between the audience ratings and critics ratings.



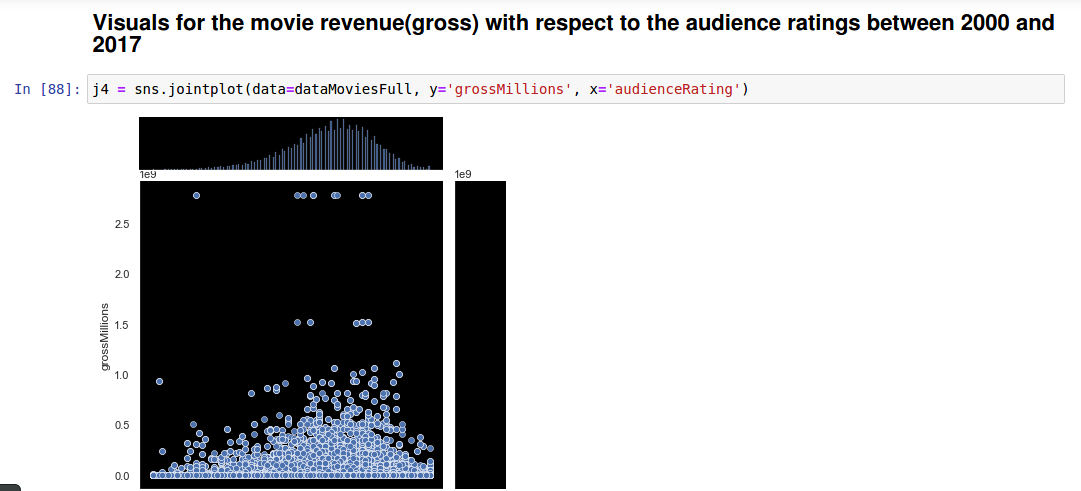
For the duration of a 60 min to 120 min movie, the critic’s ratings are varying from 10 to 0 to 10.



On this graph, we can see the linearity of the notes between the audience and the critics.



From this graph, it is shown that most movies are for a span between 60 min to 120 min, and some more than 180 mins.

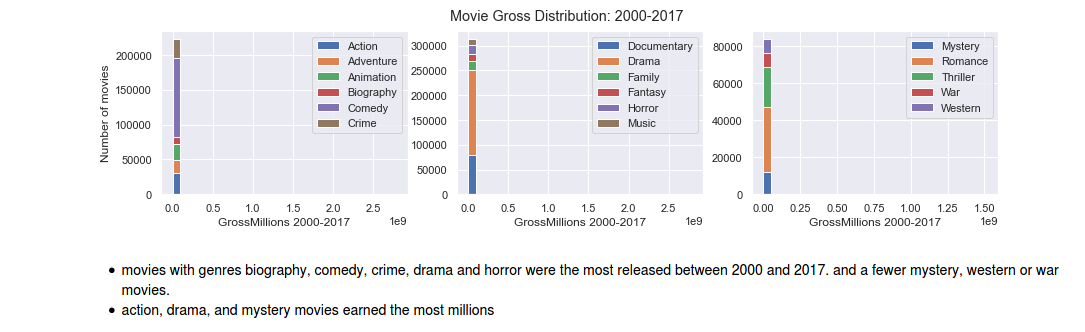


Movies that were well rated(ratings concentrated between 5 to 10) have generated the most million dollars.

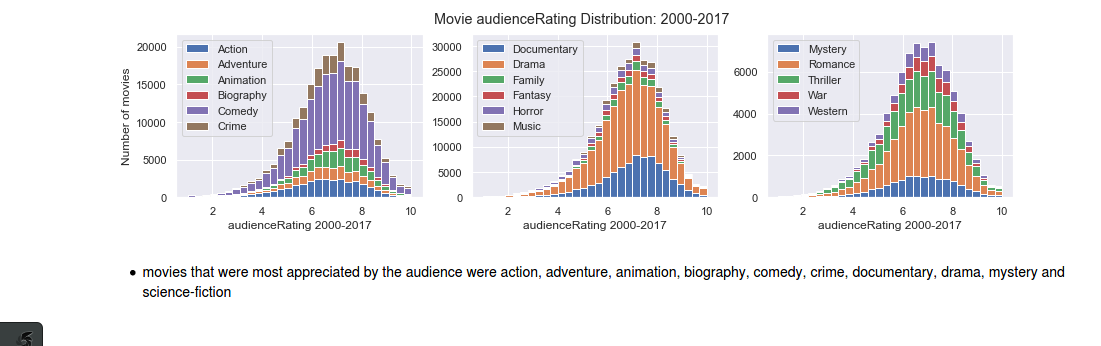




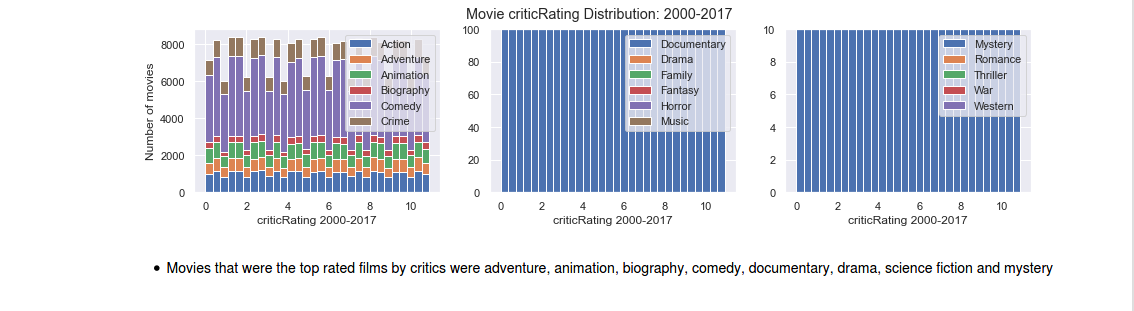
We note that the High revenue for movies with higher critics ratings, whereas concentration is mostly stable between .5 to1.5 millions.



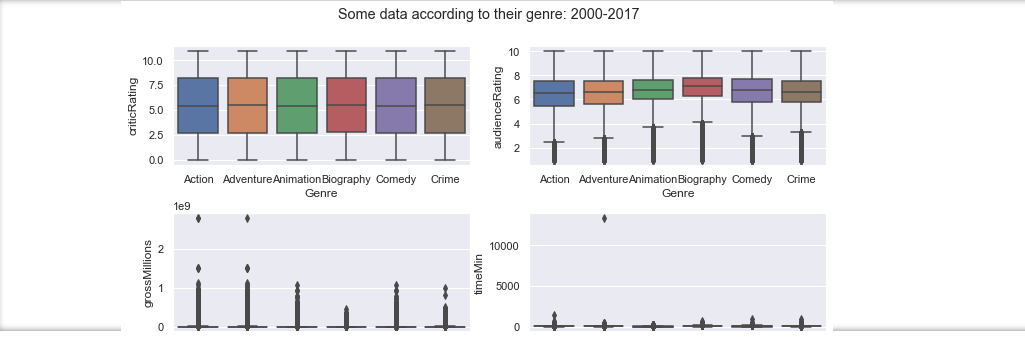
Movies with genres such as biography, comedy, crime, drama and horror were the most released between 2000 to 2017, and fewer mystery, western or war movies. Action, drama and mystery movies earned the most millions.



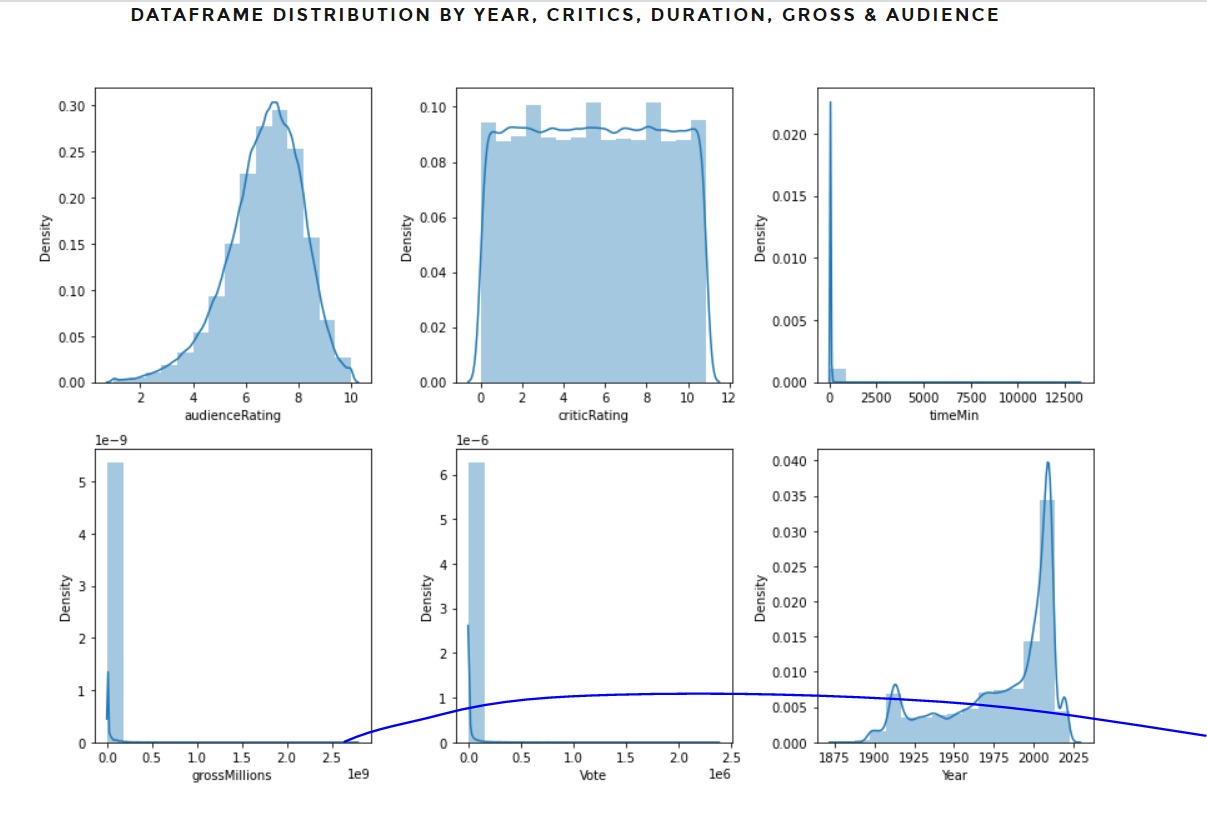
Movies that were most appreciated by the audience were action, adventure, animation, biography, comedy, documentary, drama, mystery and science-fiction.

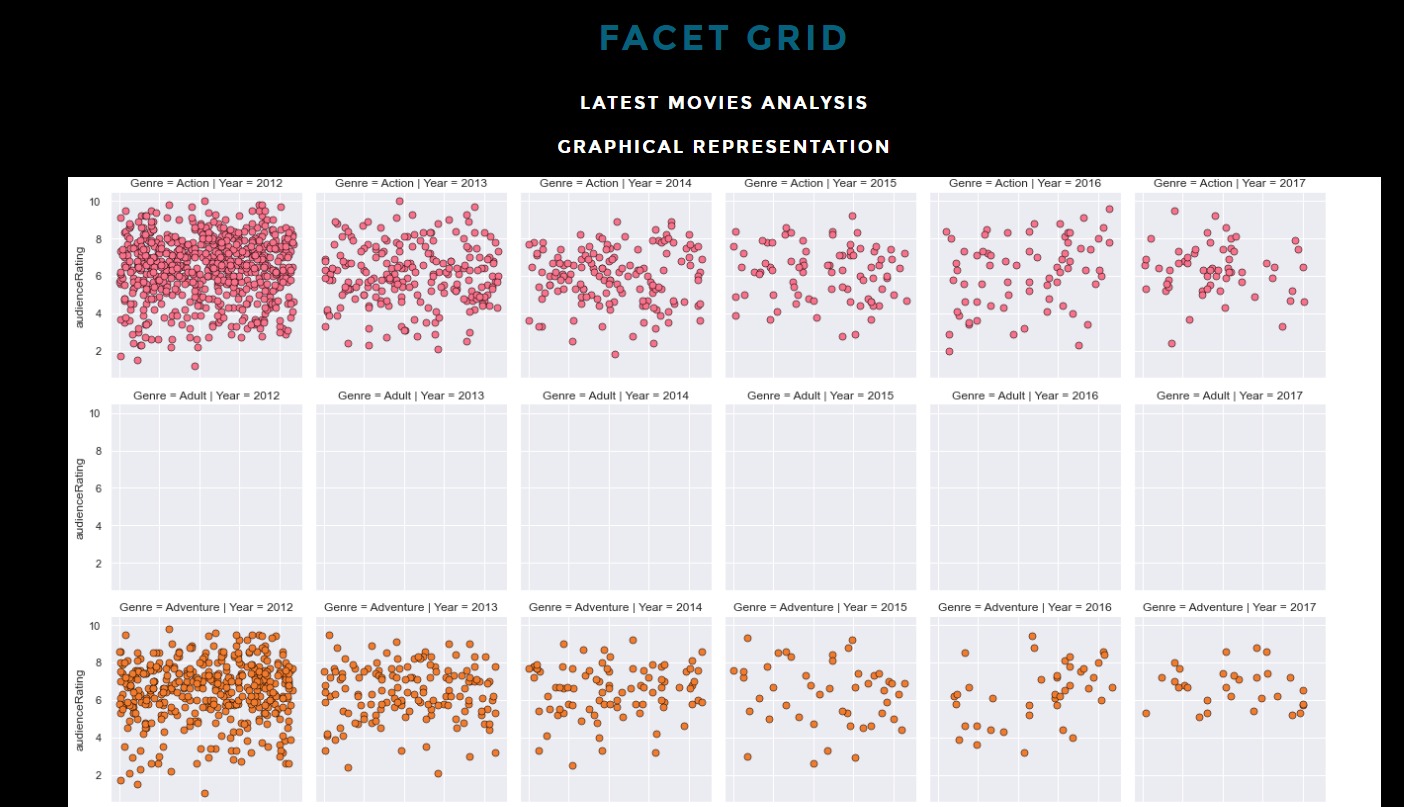


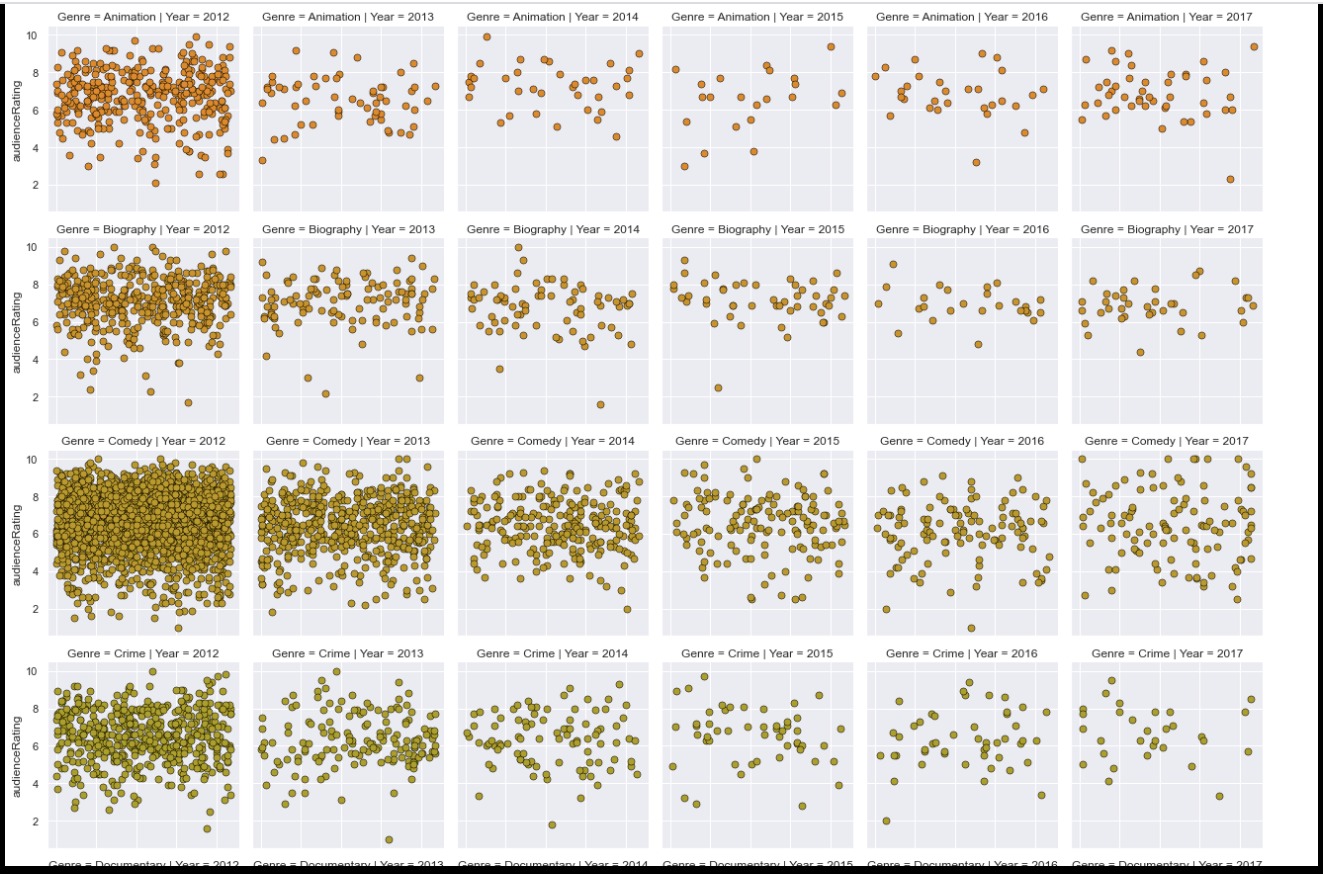
Movies that were the top-rated films by critics were adventure, animation, biography, comedy, documentary, drama, science fiction and mystery.



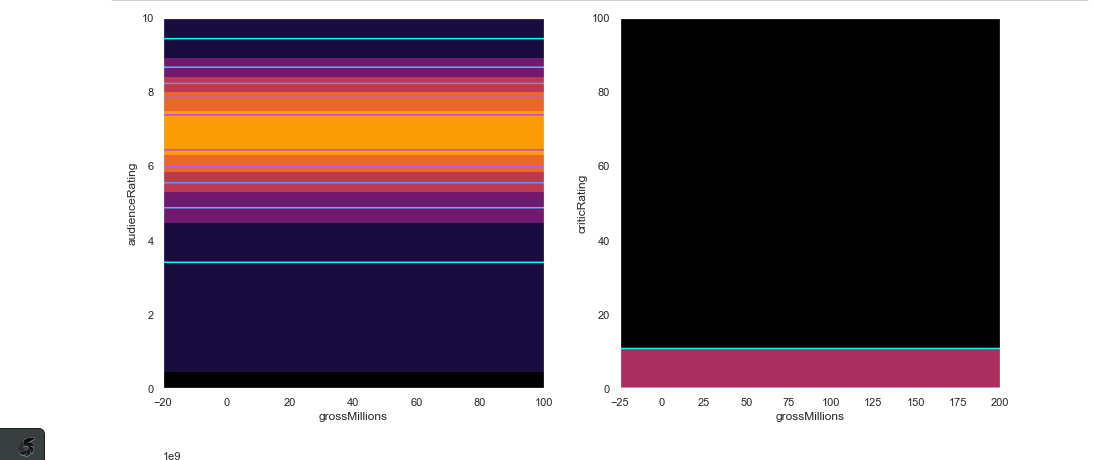
This is the boxplot representation of critics ratings and audience ratings wrt to the genre, which shows that the critics ratings are strict and can vary with a range from 2 to 8 whereas the audience ratings are mostly in the range of 6 to 8

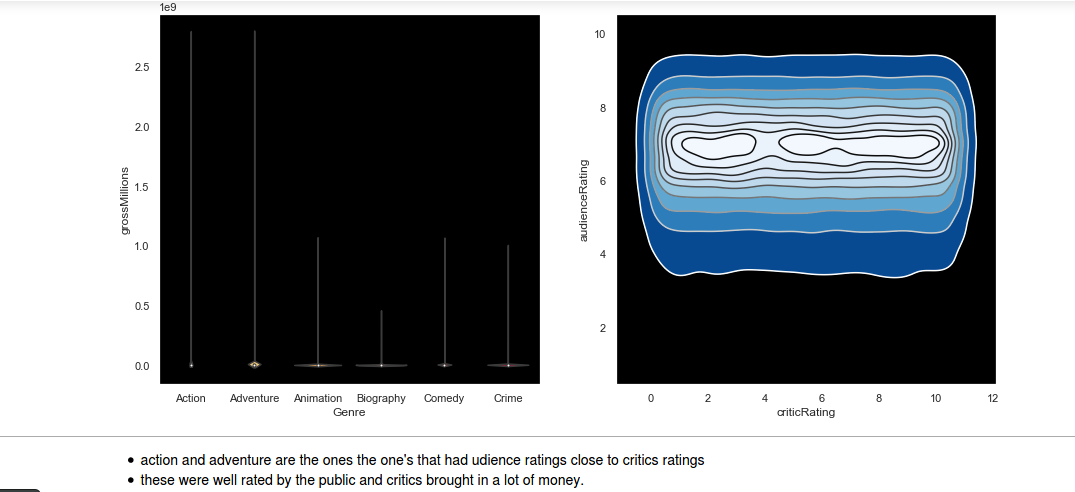






Below are the Histogram representation of the above Facet graph, We can understand that the action and adventure movies highly earned genres





Action and adventure are the ones that had audience ratings close to critics ratings. These were well rated by the critics and brought in a lot of money.

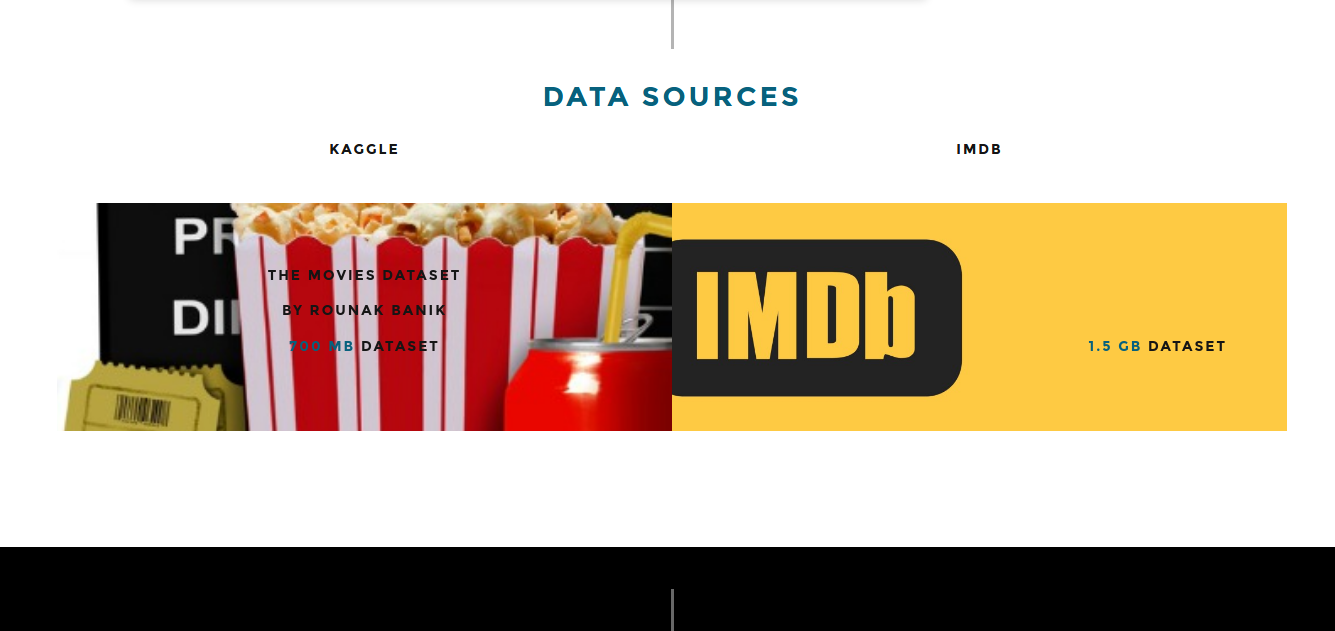
**Static Website Deployment**

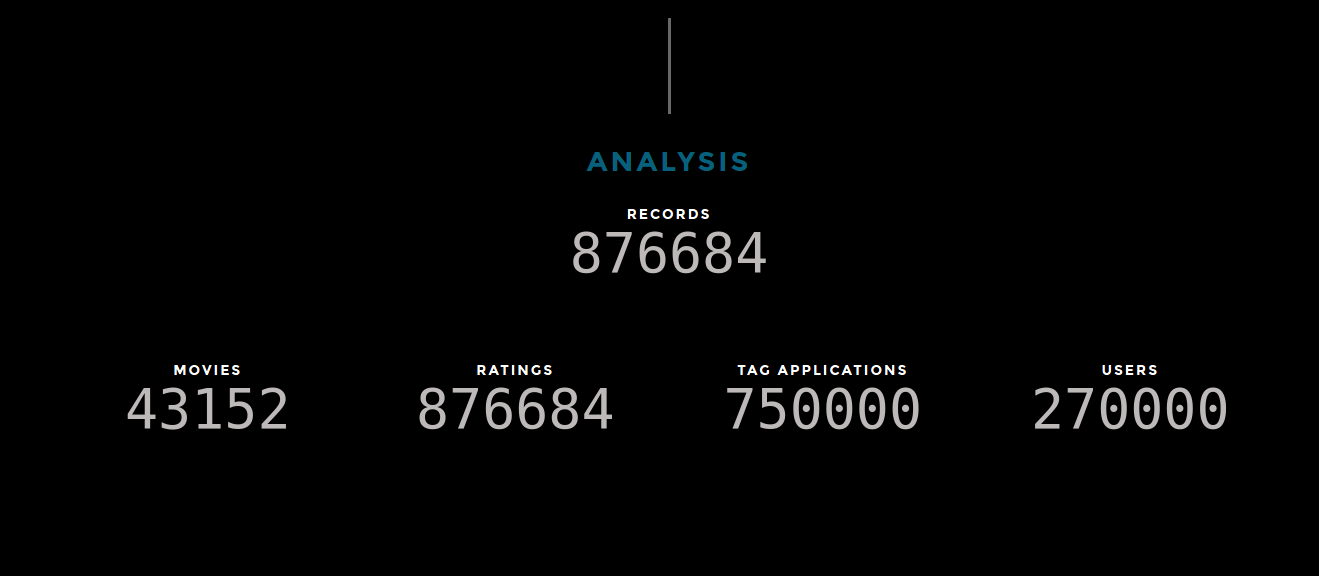
We have created a static webpage using technologies like HTML, CSS, Javascript to host this from the S3 bucket to represent all the visuals and analysis.

The link is as below :

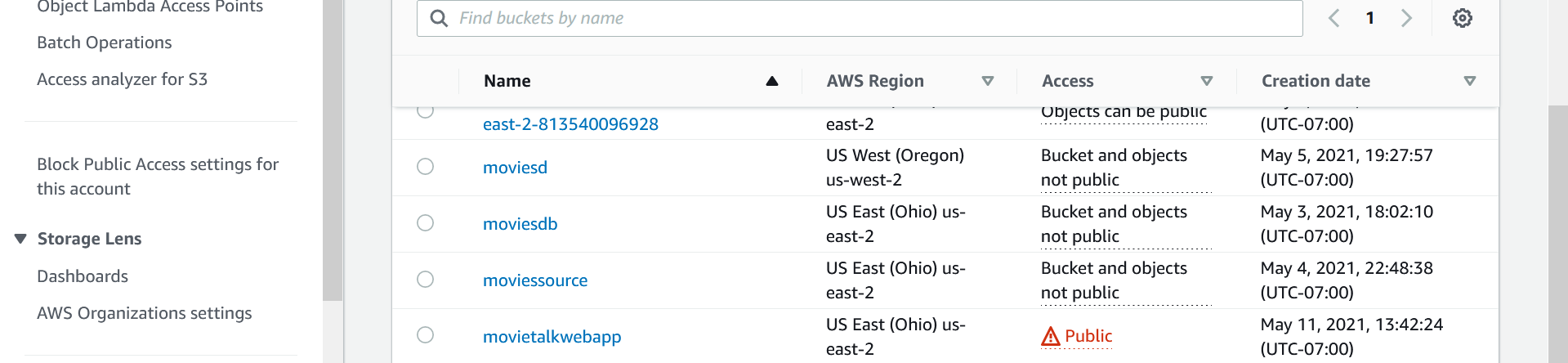
<https://movietalkwebapp.s3.us-east-2.amazonaws.com/index.html>

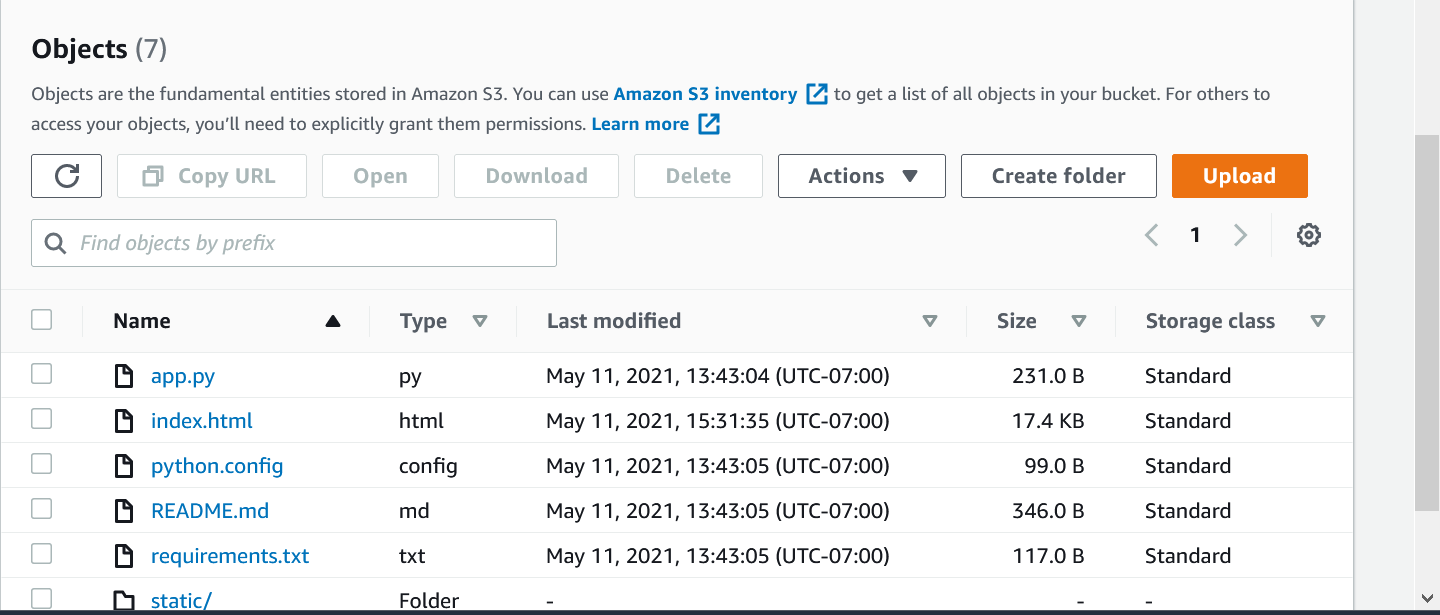
****

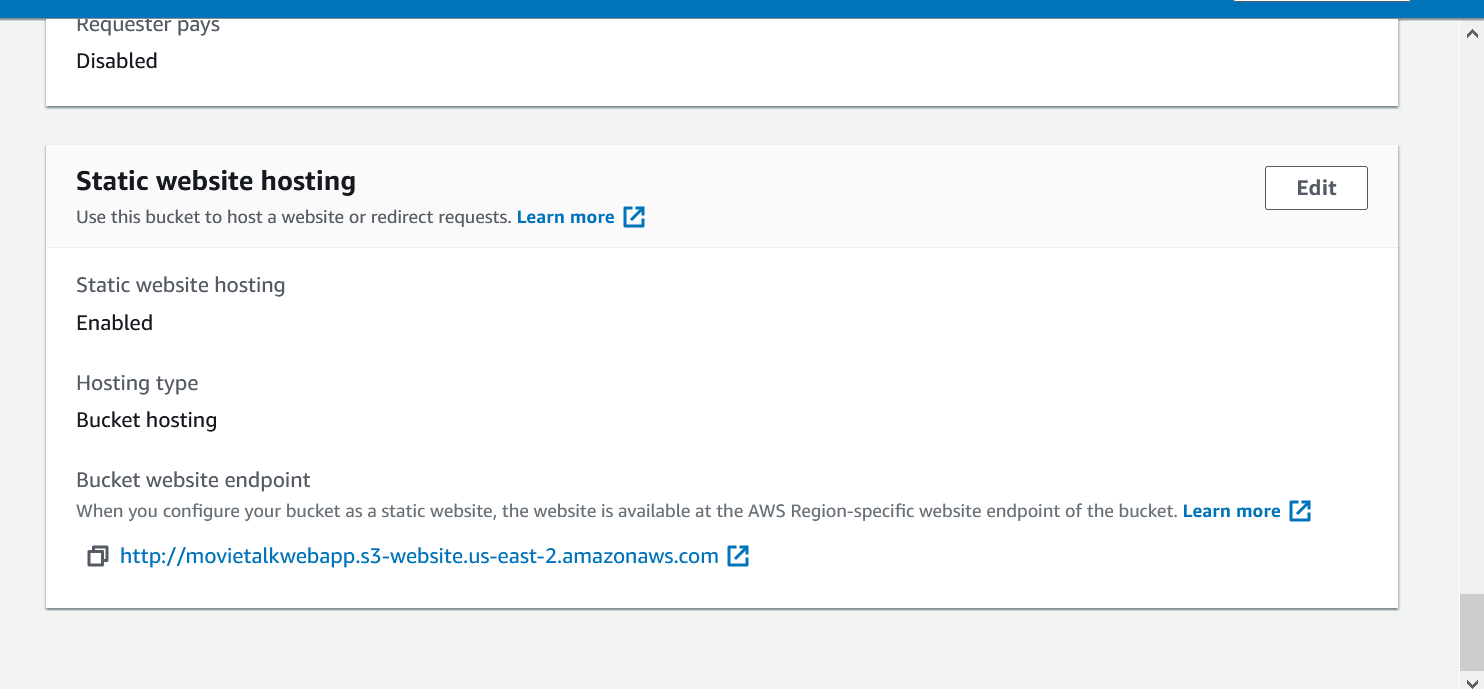
****

****

This is our S3 bucket for our web application which is publicly accessible. Our S3 bucket is called movietalk webapp.







**Findings:**

* **Tableau**

1. Top 10 production companies with the highest revenue
2. Top Rated Movies
3. Genres with the highest revenue
4. Most popular movies

* **Python**

1. Critics ratings are more severe than the public ratings.
2. Audience ratings of the movies were quite close to those of the critic’s ratings.
3. Average Movie duration is mostly ranging from 60 minutes and 120 minutes.
4. Movies make the most Millions which are rated well by critics and audiences.
5. Movies makes the most millions if movie duration fall under 60 minutes and 150 minutes (2h30) make the most money
6. Movies are expected to make less money if they exceed 3 hours duration.
7. Action, Drama and Mystery movies make the most money compared to other genres
8. Movies under Animation, adventure, biography, crime, documentary, mystery and science-fiction were highest rated by the public.

**Conclusion:**

We were able to study this large volume of movies dataset released between 2000 and 2017:

* **Data sourcing and feeding:** two different datasets from two different sources
* **Data preparation and cleaning through python pandas and Talend ETL tool:** to handle the missing data without data loss and formatting the dataset files in a readable format, including conversion of Json file to csv.
* **Modelling of the data:** processed and cleaned data than loading to Amazon s3 bucket and then to amazon’s Redshift cluster.
* **Analysis and Visualization:** data collected is then analysed to get meaningful information for further predictions, analysis and visualization was achieved through two different means by python and Tableau.

**References:**

<https://github.com/preetikhatrisjsu/Data228_Project>

<https://movietalkwebapp.s3.us-east-2.amazonaws.com/index.html>

<https://www.kaggle.com/rounakbanik/the-movies-dataset>

<https://www.imdb.com/interfaces/>

<https://www.pinterest.com/pin/243335186087373891/>

<https://www.lamag.com/culturefiles/covid-19-movie-theaters/>

<https://www.pinterest.com/pin/316448311291200389/>

<https://www.pinterest.com/pin/547398529686485966/>