Amrita Vishwa Vidyapeetham
Centre for Excellence in Computational Engineering and Networking
Amrita School of Engineering, Coimbatore

Movies Analysis & Recommendation System

Prepared By:
J. Ajay Surya
Mohitha.V
M. Prasanna Teja
P. Sai Ravula

Supervised by:
Dr. Sanjana Sri JP
Asst. Professor

An End Semester Project submitted to the CEN department as a part of course evaluations of “**21AIE304 - Big Data and Database Management”** for B. tech in **Computer Science Engineering** – **Artificial Intelligence.**

ACKNOWLEDGMENT

We are deeply thankful to **Centre for Excellence in Computational Engineering and Networking (CEN)** at Amrita Vishwa Vidyapeetham, Coimbatore for providing us such a wonderful environment to pursue our research. We would like to express our sincere gratitude to **Dr. Sanjana Sri JP, Asst, Professor, Department of CEN**, Amrita Vishwa Vidyapeetham. We have completed our research under her guidance. We found the research area, topic, and problem with her suggestions. She guided us with our study and supplied us with many research papers and academic sources in this area. She is patient and responsible when we had questions and needed her help. She would always find time to meet and discuss with us no matter how busy she was. We deeply appreciate your contributions to our team's journey.

We would also like to extend our gratitude to **Dr. K P Soman, Head of the Department, School of Artificial Intelligence**, at CEN, Amrita Vishwa Vidyapeetham, who provided us with the base knowledge of this field of research. Without him this research work would not exist in its present form. We would like to acknowledge our team members for supporting each other and be grateful to our university for providing this opportunity for us.

TABLE OF CONTENTS

Abstract:

This report presents a comprehensive analysis of IMDB Movies dataset utilizinga data driven approach, integrating technologies such as Spark, MySQL, and PySpark. The project encompasses data collection, integration with Spark fordistributed processing, and subsequent storage in Data Frames. Results obtained from analysis are saved to csv files and then imported into Tableau foreffective visualization.

The report begins with an introduction outlining the project's objectives and scope. It details the data collection process, defining schema in MySQL andthen integration with Spark is explained. Data Frame creation in Spark isdiscussed, providing insights into the structured representation of the data.

A Movie Recommendation system was developed in PySpark, which recommends movies from 1980 – 2000 to the user visiting the website developed using Streamlit based on genre, star in the movie and productionhouse that brough the movie.

Key findings from the analysis are summarized, providing insights gained from both visualizations and movie recommendation system. Challenges and limitations are addressed, paving the way for future improvements. The report concludes by emphasizing the project's significance and proposing avenues forfuture research.

Introduction:

This project undertakes a comprehensive analysis of IMDB Movies Dataset consisting of movies details from 1980-2000, leveraging a multifaceted approach integrating Spark, MySQL, and PySpark. The primary objective is togain insights into the trends in the film industry in that period by employing big data analytics and database management. The scope encompasses data collection, integration with Spark for efficient processing, recommendation system development for movie suggestion, and the creation of an interactivedashboard in Tableau based on the analysis performed in Apache Spark for visualization. By amalgamating these technologies, the project aims to offer a holistic understanding of movie trends and user's choice on movies.

Motivation:

Embarking on the journey of analyzing the IMDB movies dataset from 1980 to 2000 using Apache Spark and MySQL was driven by a deep-seated passion forcinema and a curiosity to understand the evolution of the film industry over two pivotal decades. The utilization of cutting-edge technologies like Apache Sparkand MySQL reflects our commitment to harnessing the power of big data toolsfor effective data processing and storage. As the project unfolded, the desire tocreate something impactful led us to venture into building a movie recommendation system in PySpark.

Data Collection:

The primary source of our IMDB Movies data is Kaggle, offering a comprehensive dataset spanning from 1980 to December 2000. Kaggle provides a reliable and up-to-date repository of various datasets, ensuring theinclusion of relevant data points crucial for our analysis. Features in our datasetare title, rating, genre, release date, IMDB score, people's vote, director, star,country, budget, gross, company and runtime.

Data loading into MySQL:

Before integrating with Spark, the Movies data underwent a crucial phase ofloading into a MySQL table. This process involved several key steps to ensureseamless storage and retrieval:

i. **Database Schema Design:** A well-defined database schema was crafted to accommodate the specific attributes and structure of the Movies dataset. Thisschema served as the blueprint for organizing and storing data in the MySQLdatabase.

ii. **Data Transformation:** Data transformation steps were implemented to align the raw data with the predefined database schema. This included handling datatypes, ensuring consistency, and preparing the dataset for efficient storage in arelational database.

iii. **Loading Data into MySQL:** Using MySQL's data import tools, the dataset was loaded into the designated table. This step involved mapping the transformed data to the corresponding fields in the MySQL table, facilitating a seamless transfer of information.

Database Integration with Spark:

Spark was instrumental in handling the voluminous Movies dataset, providing ascalable and efficient framework for distributed processing. Leveraging Spark'scapabilities, the data integration process involved parallel computation across multiple nodes, significantly reducing processing time.

Connecting MySQL with Spark through JDBC facilitated seamless data integrationand analysis. The integration process involved several key steps:

1. **Configuration Setup:**
   JDBC drivers for MySQL were configured within the Spark environment. Thisentailed specifying the driver class, connection URL, and authentication credentials to establish a secure link between Spark and the MySQL database.

2. **Establishing Connection:**
   A connection was established using Spark's JDBC API. This connection servedas the bridge between the Spark application and the MySQL database, allowingfor the efficient exchange of data between the two environments.

3. **Data Extraction:**
   Spark SQL queries were employed to extract data from the MySQL databaseinto Spark Data Frames. This step involved crafting SQL queries to retrieve specific data relevant to the analysis, leveraging the power of Spark's distributed computing capabilities.

Data Pre-Processing in Spark:

1. Identification of Null Values:

The first critical step involved identifying attributes containing null values.We systematically iterated through each column of our dataset and compiled a list of attributes that exhibited the presence of null values.

2. Removal of null values in Specific attributes:

A targeted approach was adopted to handle null values in key attributes essential for our analysis. Leveraging the capabilities of Spark, we efficiently removed samples with null values in specific attributes, thereby enhancing the completeness and reliability of our dataset.

3. Detection and Handling of duplicate values:

Through a meticulous comparison of records, we calculated the number ofduplicate values. If any duplicates were found, the count was reported, demonstrating our dedication to ensuring the uniqueness and accuracy ofour movie data.

4. Identification of attributes with zero values:

To further refine our dataset, we proactively sought attributes containingzero values. This was particularly crucial, as zero values in certain fields could indicate missing or inaccurate data.

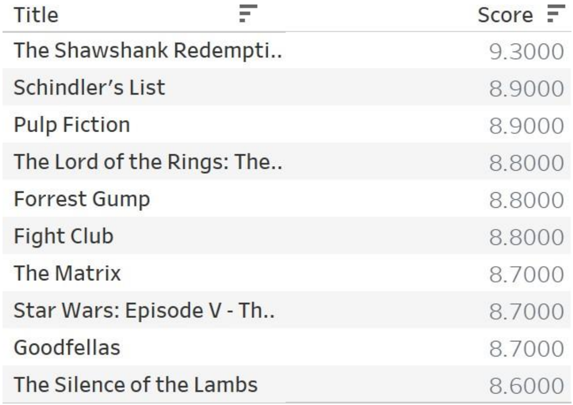
5. Removal of samples with zero values in identified attributes:

Building upon the identification of zero values, we systematically eliminatedsamples where these values were present in the identified attributes. Through an iterative process, we carefully filtered out rows containing zero values, ensuring that our dataset was free from potentially misleading orerroneous information.

Main Analysis in Spark:

i. Top 10 Movies based on IMDb Scores:

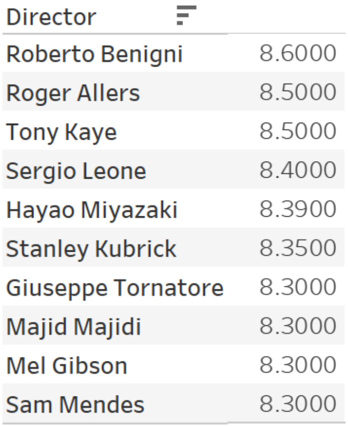
Identified and showcased the top 10 movies with IMDb scores greater than or equal to 8.0, providing insights into the highest-rated films in the period 1980 – 2000.



Title	Score
The Shawshank Redempti..	9.3000
Schindler's List	8.9000
Pulp Fiction	8.9000
The Lord of the Rings: The..	8.8000
Forrest Gump	8.8000
Fight Club	8.8000
The Matrix	8.7000
Star Wars: Episode V - Th..	8.7000
Goodfellas	8.7000
The Silence of the Lambs	8.6000

ii. Top 10 Successful Directors basedon the Average of IMDb Scores:

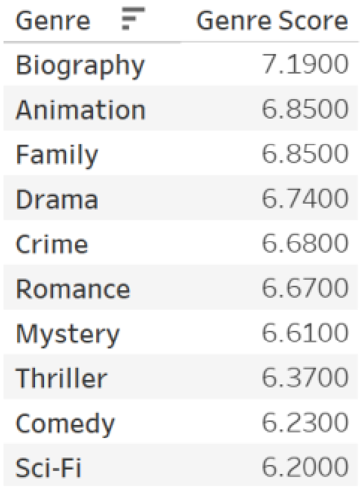
Identified and presented the top directors based on the average IMDb scores of their movies, showcasing directorial success.



Director	
Roberto Benigni	8.6000
Roger Allers	8.5000
Tony Kaye	8.5000
Sergio Leone	8.4000
Hayao Miyazaki	8.3900
Stanley Kubrick	8.3500
Giuseppe Tornatore	8.3000
Majid Majidi	8.3000
Mel Gibson	8.3000
Sam Mendes	8.3000

iii. Top 10 Genres based on average of IMDB score votes:

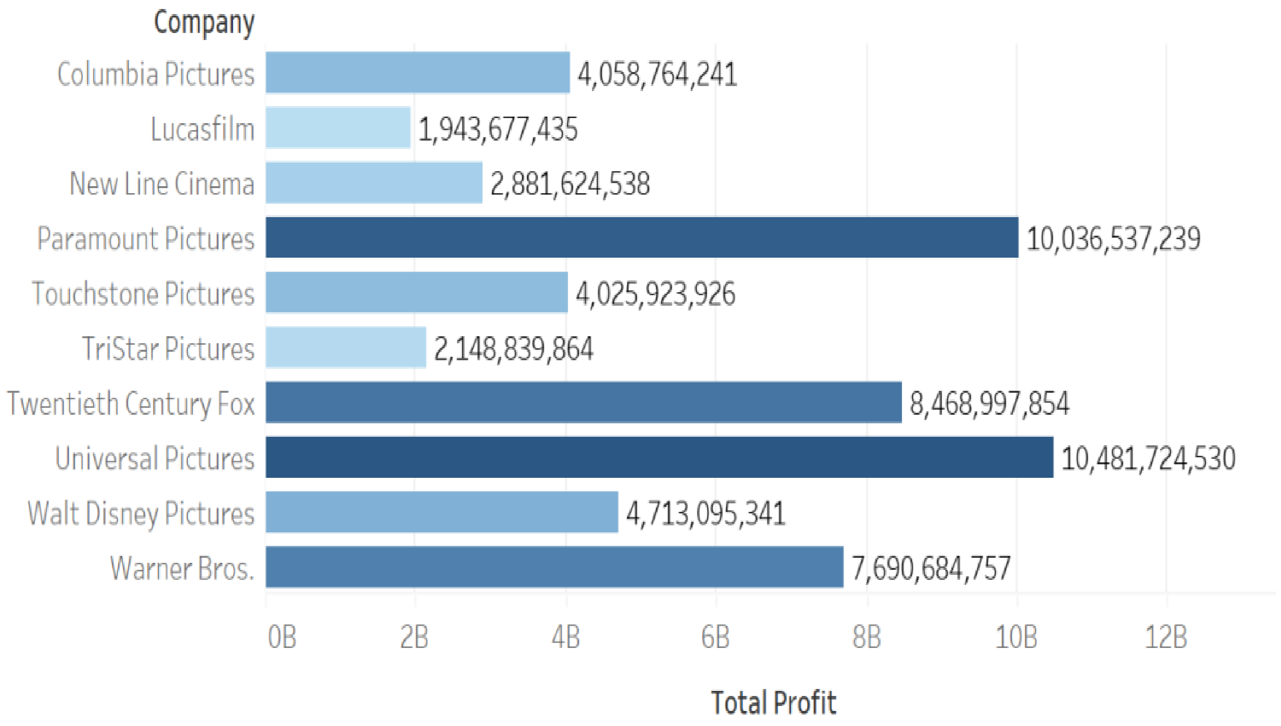
Identified and presented the top genres based on the average IMDb scores of their movies, showcasing genre preferences.



Genre	Genre Score
Biography	7.1900
Animation	6.8500
Family	6.8500
Drama	6.7400
Crime	6.6800
Romance	6.6700
Mystery	6.6100
Thriller	6.3700
Comedy	6.2300
Sci-Fi	6.2000

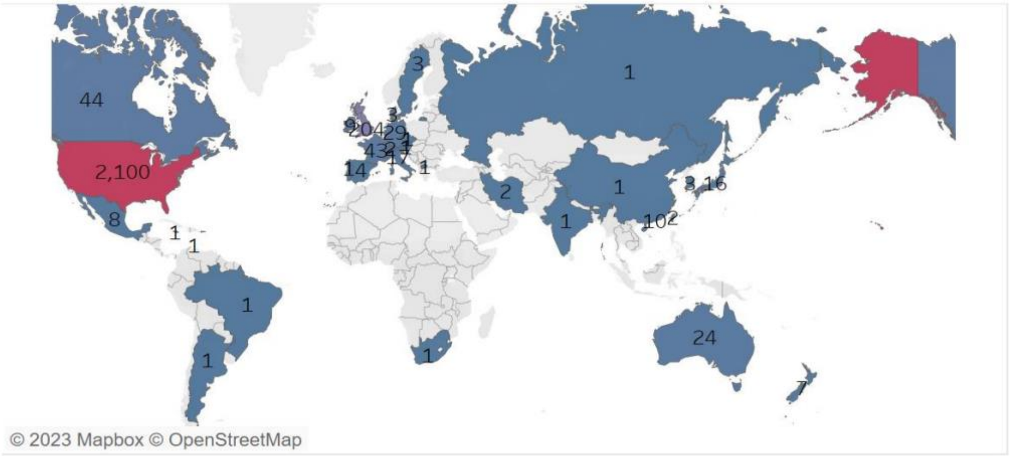
iv. Companies which made the Most Profit Over the Span of 20 Years:

Explored and showcased companies with the highest total profit, calculatedby subtracting budgets from gross earnings.



v. Movies count released by each country:

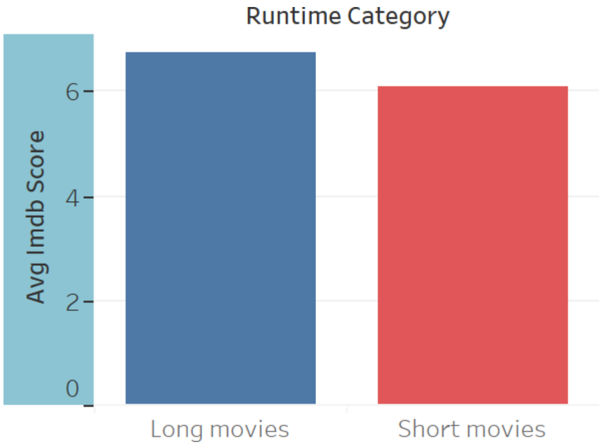
This analysis focuses on determining the count of movies released by each country in the dataset. It enables a quick comparison of the contribution ofeach country to the overall dataset, shedding light on the global distributionof film production.



© 2023 Mapbox © OpenStreetMap

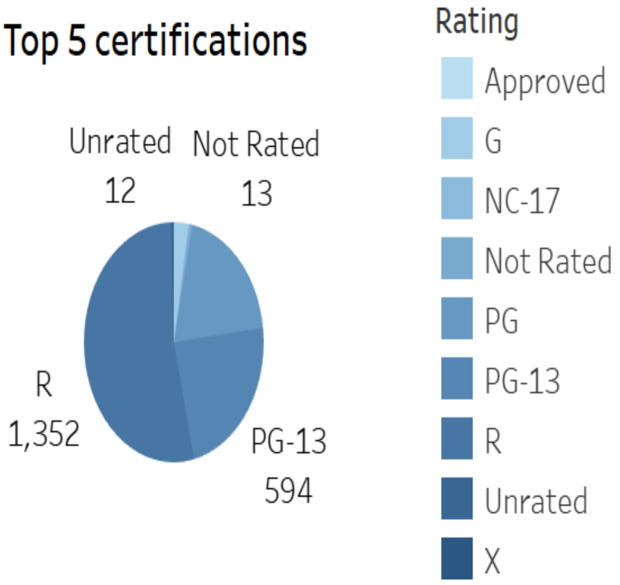
vi. Average scores of longer movies:

This analysis aims to assess the average IMDb scores of movies categorized by their runtime concerning the overall avg.runtime. By categorizing movies as “Shorter movies” or “Longer movies,” this analysis offers a nuanced preview of how audience perceptions of movies may vary based on their duration, contributing to a more informed perspective on movie preferences.



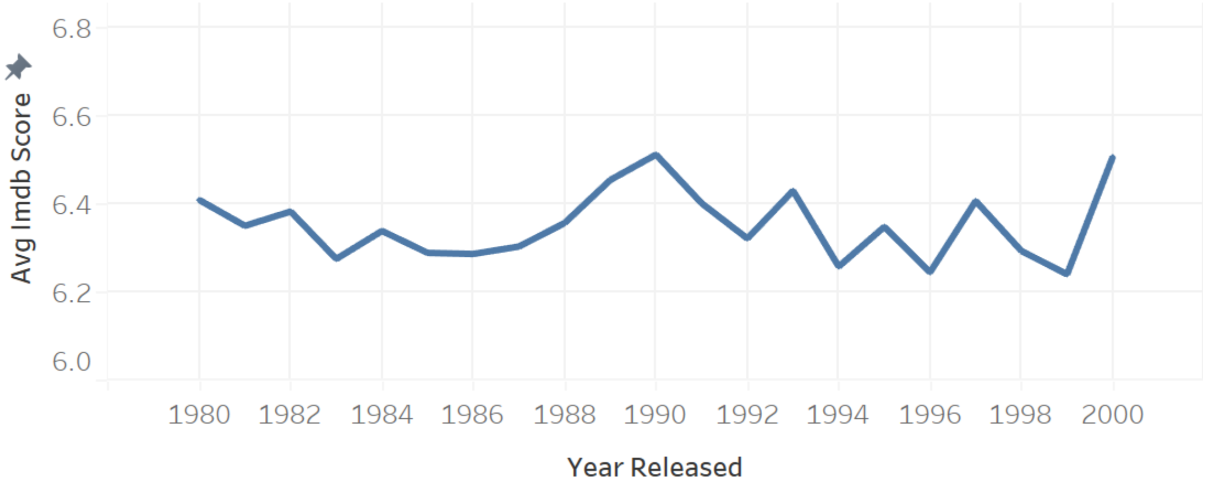
vii. Average Scores of Movies with Different Age Certificates:

This analysis focuses on calculating the number of movies released grouped by different age certificates or ratings. This information is key for understanding the perceived quality of movies across various content rating categories, offering a glimpse into people preferences within each age group.



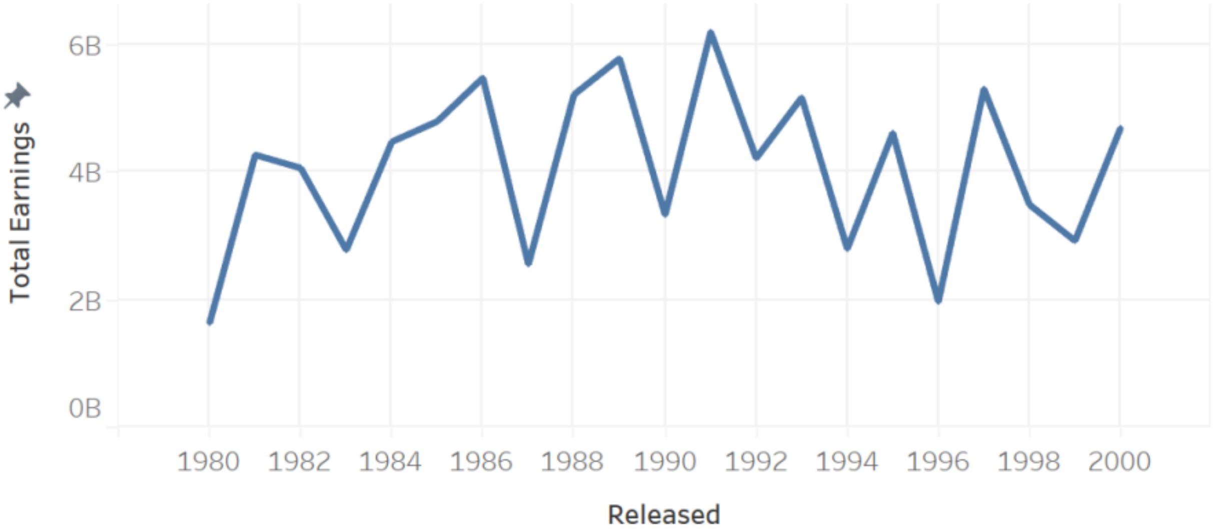
viii. Average scores of movies over years:

This analysis aims to assess the quality of movies released over the span of two decades (1980-2000). This analysis is valuable for understanding trends in audience reception and critical acclaim across different eras in the realm of cinema.



ix. Trends of profits earned per year:

This analysis focuses on examining the trend of profits earned by movies peryear. It is essential for understanding the financial performance of the filmindustry over the specified time frame, highlighting years of significant profitability or challenges.



Additional analysis:

x. Directors preferred my most successful companies:

This analysis focuses on identifying directors preferred by successful movie production companies. It contributes to understanding the collaborative success of specific directors and production companies within the film industry.

xi. Actors preferred my most successful companies:

This analysis focuses on identifying actors preferred by successful movie production companies. It contributes to understanding the collaborative success of specific actors and production companies within the film industry.

xii. Average Runtime of Movies Directed by the Most Successful Directors:

This analysis focuses on determining the average runtime of movies directed by the most successful directors, offering a holistic perspectiveon the artistic and temporal aspects of their work.

xiii. Countries That Produced the Best Films Based on Average IMDb Score:

This analysis aims to identify the countries that have produced the best films based on the average IMDb score. It contributes to understanding the global distribution of high-quality films and can be indicative of thecinematic excellence associated with specific regions.

Movie Recommendation System in PySpark:

Objective: The aim of this part of the project is to develop a Movie Recommendation System using PySpark for data analysis and Streamlit for interactive web application deployment.

Application Structure: The Streamlit app is structured to allow users to choose between three recommendation types: Movie, Star, and Production House.

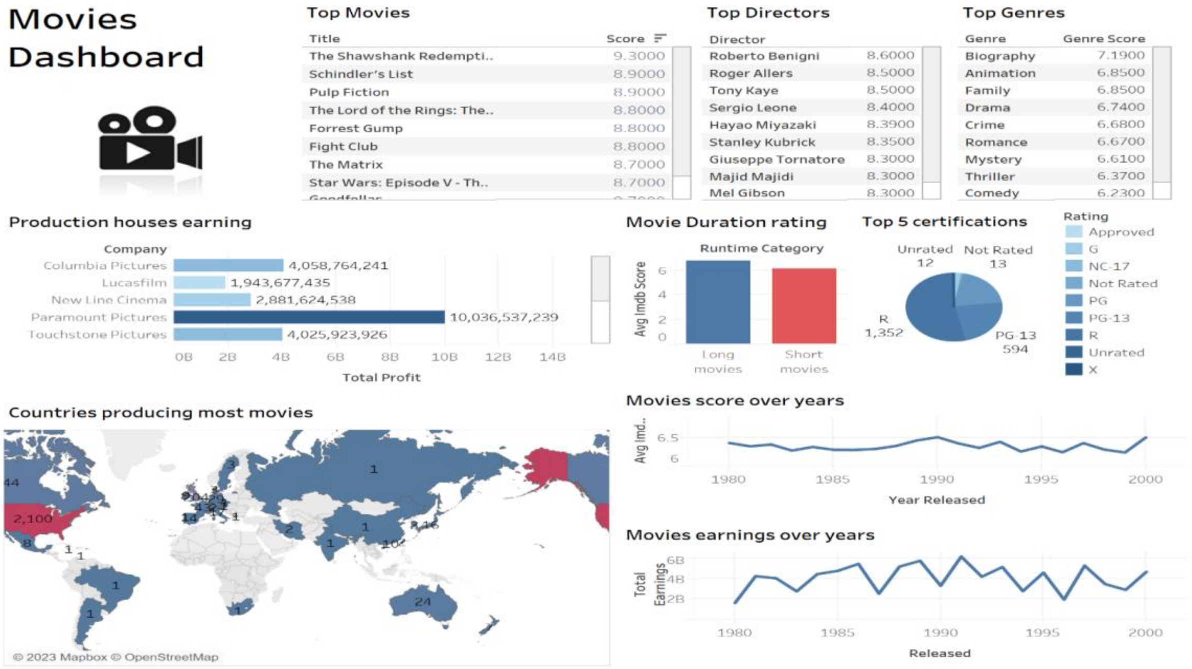
Recommendation based on Genre: When a user chooses a specific movie, the system first extracts the genre of that movie. It then filters the dataset toinclude only movies with the same genre, ensuring thematic similarity. The filtered data is subsequently sorted in descending order of IMDb ratings, andthe top 5 movies are recommended to the user. This approach ensures thatusers receive recommendations closely aligned with the genre of the movie they initially selected.

Recommendation based on Star: When a user selects a star (actor), the system filters the dataset to include only movies featuring the chosen star. Thefiltered data is then sorted based on IMDb ratings in descending order. The top 5 movies featuring the selected star are presented as recommendations. This mechanism allows users to explore movies associated with their favorite actors and discover highly rated performances.

Recommendation based on Production House: In the case of selecting a production house, the system filters the dataset to include only movies produced by the chosen production house. The dataset is then sorted based onIMDb ratings in descending order, and the top 5 movies produced by the selected entity are recommended. This functionality provides users with insights into the success and quality of movies associated with specific production houses.

Results:

Using Tableau for creating Movie Dashboard serves as a powerful tool for visualizing and understanding the results of the Apache Spark analysis. Its interactive features empower users to explore movie data dynamically, uncover insights, and make data-driven decisions in the realm of the film industry. So upon completing the analysis using Apache Spark and storing theresults in CSV files, we created an interactive and visually appealing Movie Dashboard in Tableau. The goal was to provide a user-friendly interface for exploring and understanding the insights derived from the Spark analysis.



Movies Dashboard

Top Movies

Title	Score
The Shawshank Redempti..	9.3000
Schindler's List	8.9000
Pulp Fiction	8.9000
The Lord of the Rings: The..	8.8000
Forrest Gump	8.8000
Fight Club	8.8000
The Matrix	8.7000
Star Wars: Episode V - Th..	8.7000

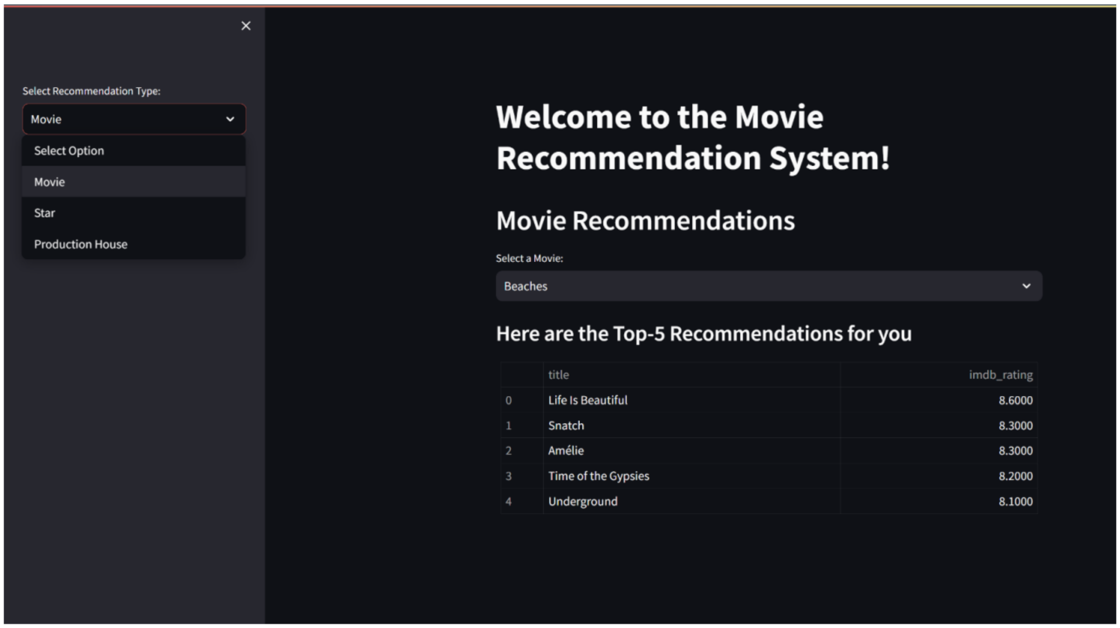
Top Directors

Director	
Roberto Benigni	8.6000
Roger Allers	8.5000
Tony Kaye	8.5000
Sergio Leone	8.4000
Hayao Miyazaki	8.3900
Stanley Kubrick	8.3500
Giuseppe Tornatore	8.3000
Majid Majidi	8.3000
Mel Gibson	8.3000

Top Genres

Genre	Genre Score
Biography	7.1900
Animation	6.8500
Family	6.8500
Drama	6.7400
Crime	6.6800
Romance	6.6700
Mystery	6.6100
Thriller	6.3700
Comedy	6.2300

For movie recommendation system, we utilized Streamlit to create an interactive web application that enables system users to explore movies, stars, and production houses while receiving personalized recommendations basedon their preferences.



Select Recommendation Type: Movie

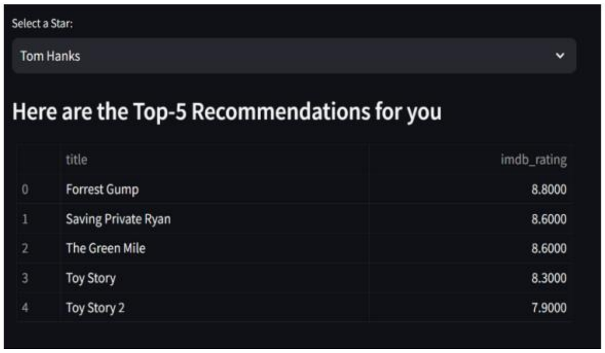
Welcome to the Movie Recommendation System!

Movie Recommendations

Select a Movie: Beaches

Here are the Top-5 Recommendations for you

	title	imdb_rating
0	Life Is Beautiful	8.6000
1	Snatch	8.3000
2	Amélie	8.3000
3	Time of the Gypsies	8.2000
4	Underground	8.1000

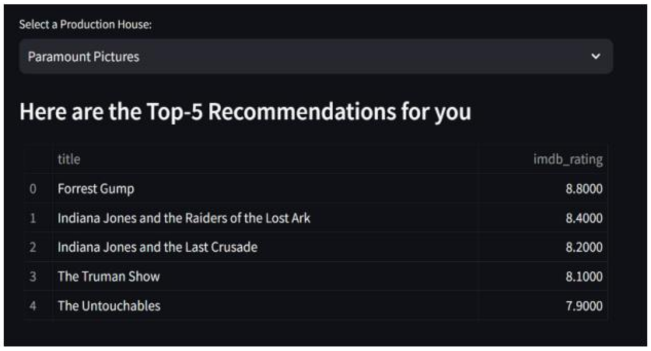


Select a Star:

Tom Hanks

Here are the Top-5 Recommendations for you

	title	imdb_rating
0	Forrest Gump	8.8000
1	Saving Private Ryan	8.6000
2	The Green Mile	8.6000
3	Toy Story	8.3000
4	Toy Story 2	7.9000



Select a Production House:

Paramount Pictures

Here are the Top-5 Recommendations for you

	title	imdb_rating
0	Forrest Gump	8.8000
1	Indiana Jones and the Raiders of the Lost Ark	8.4000
2	Indiana Jones and the Last Crusade	8.2000
3	The Truman Show	8.1000
4	The Untouchables	7.9000

Conclusion:

In conclusion, the Movie Analysis Project successfully navigates the intricacies of the film industry, from data preprocessing and analysis using Apache Spark to the creation of an interactive Tableau Movie Dashboard. The project not only provides insights into historical movie trends but also sets thestage for future advancements in recommendation algorithms and real-time data integration. This comprehensive approach ensures the project's relevance, impact, and potential for continuous exploration within the dynamic landscape of the film industry.

Future Works:

Some of the future works that can be incorporated into our research are,

Advanced Algorithms: The project lays the foundation for future enhancements, including the integration of advanced recommendation algorithms for more accurate and personalized suggestions.

Real-Time Updates: Exploring possibilities for real-time data integration could further enhance the project's relevance and timeliness.

References:

https://kontext.tech/article/610/spark-scala-load-data-from-mysql

https://spark.apache.org/docs/latest/

https://dev.mysql.com/doc/

https://help.tableau.com/current/pro/desktop/enus/default.htm