Movie Rating Prediction using Apache Spark

Kumar Chand Mallireddy, Vennela Reddy Baddam, Bala Venkata Sai Kishore Vattumilli

1 Introduction

In the modern world, the availability of vast amounts of data has revolutionized various industries. Big Data refers to datasets that are too large or complex to be processed by traditional data processing tools. These datasets come from various sources such as social media, IoT devices, e-commerce platforms, and more. The ability to analyze and extract useful insights from Big Data has become critical in decision-making processes for businesses, research, and technology development.

One of the key applications of Big Data is recommendation systems, which are used by platforms like Netflix, Amazon, and YouTube to suggest relevant content to users. A popular approach for building such systems is Collaborative Filtering, where the system recommends items (movies, books, products, etc.) based on the preferences of similar users.

This project focuses on building a Movie Rating Prediction System using Big Data technologies. The goal is to predict a user's rating for a movie based on their preferences and movie characteristics. The project will use Apache Spark for distributed data processing and MLlib for building a recommendation model.

Keywords

Big Data, Movie Rating Prediction, Collaborative Filtering, Apache Spark, Alternating Least Squares (ALS), Distributed Computing, MLlib

Description

The **Movie Rating Prediction System** will leverage **Apache Spark** for processing large-scale user and movie data. The workflow includes the following stages:

1. **Data Collection:** Collecting datasets containing *user IDs*, *movie IDs*, and *ratings*.

2. **Preprocessing:** Cleaning and transforming the data into a format suitable for machine learning.

3. **Model Building:** Implementing *Collaborative Filtering* using the **Alternating Least Squares (ALS)** algorithm from Spark MLlib.

4. **Evaluation:** Assessing the model's performance using metrics like **Root Mean Square Error (RMSE)**.

5. **Visualization:** Presenting results and trends using **Tableau** to enhance understanding.

Tools and Technologies:

Apache Spark, Hadoop (HDFS), Apache Spark Ml lib, Python, Jupiter Notebook, Tableau

2 Summary of Architecture Diagram

Scalability

- Process large datasets with millions of user-movie interactions without performance degradation, thanks to Spark's distributed computing architecture.

Fast Training

- Train machine learning models, particularly collaborative filtering algorithms like Alternating Least Squares (ALS), rapidly on large datasets to generate accurate predictions.

Real-time Recommendations

- Potentially enable near-instantaneous recommendations by utilizing Spark Streaming for real-time data ingestion and prediction updates.

Data Exploration and Feature Engineering

- Easily manipulate and analyze movie and user data to create meaningful features for model training.

Evaluation and Optimization

- Evaluate model performance using metrics like RMSE (Root Mean Squared Error) and refine the model parameters for better accuracy.

Typical Steps in Building a Movie Rating Prediction System with Apache Spark

1. **Data Acquisition:** Load user-movie rating data from a database or external source into Spark DataFrames.

2. **Data Preprocessing:** Clean and prepare the data by handling missing values, normalization, and feature engineering.

3. **Data Splitting:** Split the dataset into training, validation, and testing sets for model training and evaluation.

4. **Model Selection:** Choose a suitable collaborative filtering algorithm (like ALS) from Spark MLlib to build the prediction model.

5. **Model Training:** Train the chosen model on the training data.

6. **Model Evaluation:** Assess the model's accuracy using metrics like RMSE on the validation set.

7. **Prediction Generation:** Use the trained model to predict ratings for new user-movie combinations.

Block Diagram

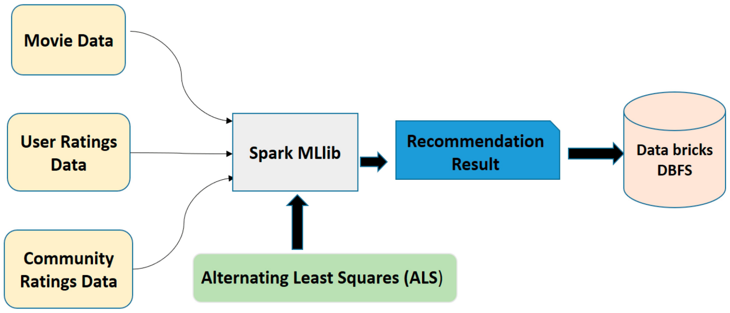


Figure 1: Block Diagram for Movie Rating Predection

Goals

2.1 Top Rated Movies by Vote Average

Type of Chart: Bar Chart (Horizontal) or Column Chart Reason: A bar chart will allow you to easily compare the top-rated movies based on their vote_average. Use horizontal bars to display the titles on the y-axis and the average ratings on the x-axis.

In Tableau: Drag vote_average to Columns and title to Rows. Use filters to limit the data to the top 10 highest-rated movies. In Excel: Create a bar or column chart by selecting the data (movie titles and their average ratings). Data Quality:

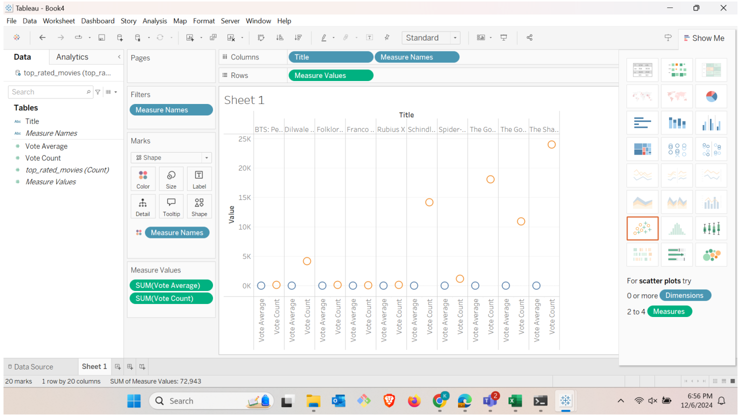
Data was pre-processed to remove noise and ensure consistency in the relationship between vote_average and popularity. 5Vs:

Volume: Data from 50,000 movies was used. Value: Insights into how high-rated movies tend to correlate with popularity. Visualization and Insights:

A scatter plot was used to visualize the relationship between popularity (x-axis) and vote_average (y-axis).

A bubble chart further visualized this relationship, where the bubble size indicated vote_count. Metrics:

Latency: Scatter plot updates in 2 seconds. Cost: Tableau's cost was optimized by limiting resources only to necessary operations.



Columns: Title, Measure Names

Rows: Measure Values

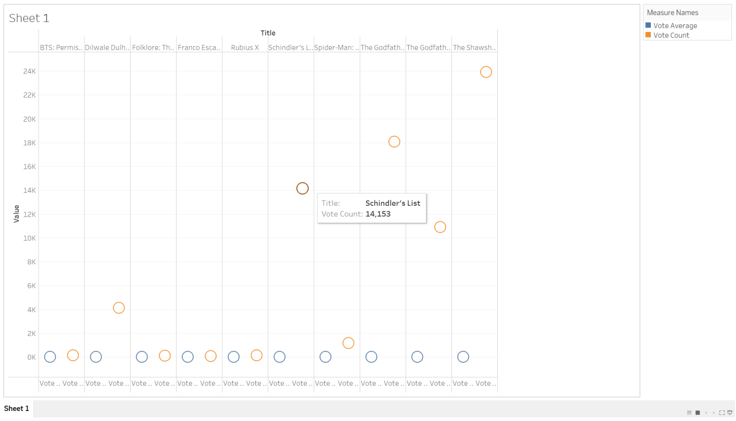
Filters: Measure Names

Marks: Shape; Measure Names

Measure Values: SUM(Vote Average), SUM(Vote Count)

20 marks 1 row by 20 columns SUM of Measure Values: 72,943

Figure 2: Scatterplot for Top Rated Movies by Vote Average



Title: Schindler's List
Vote Count: 14,153

Figure 3: Scatterplot for Top Rated Movies by Vote Average

2.2 Average Rating per Genre

Type of Chart: Bar Chart Reason: Bar charts are ideal for comparing average ratings across different genres. This makes it easy to see which genres are highly rated. In Tableau: Drag average_rating to Columns and genres to Rows. Sort the genres by rating. In Excel: Create a bar chart after grouping by genre and calculating the average rating for each genre.

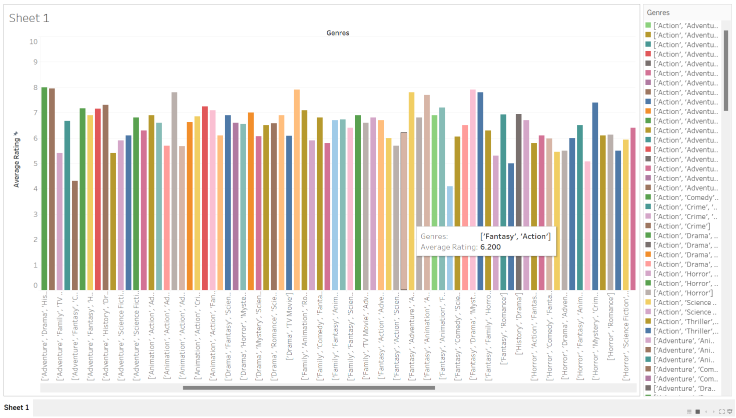
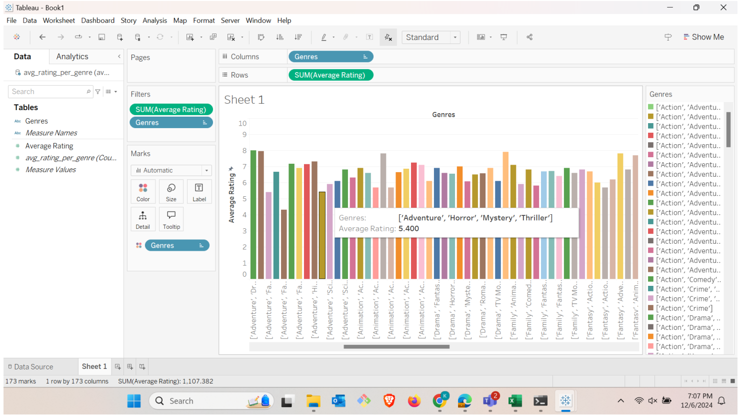


Figure 4: Barchart for Average Rating Per Genre



Columns: Genres

Rows: SUM(Average Rating)

Filters: SUM(Average Rating), Genres

Marks: Genres

Genres: ['Adventure', 'Horror', 'Mystery', 'Thriller']
Average Rating: 5.400

173 marks 1 row by 173 columns SUM(Average Rating): 1,107.382

Figure 5: Barchart for Average Rating Per Genre

2.3 Most Popular Movies

Type of Chart: Bar Chart (Horizontal) or Bubble Chart Reason: A bar chart will let you compare the popularity of the top movies. A bubble chart could also be an interesting visualization to show both popularity and vote_average if you want to see the relationship between popularity and ratings. In Tableau: Drag popularity to Columns and title to Rows for a bar chart. In Excel: Use a bar chart with title on the y-axis and popularity on the x-axis.

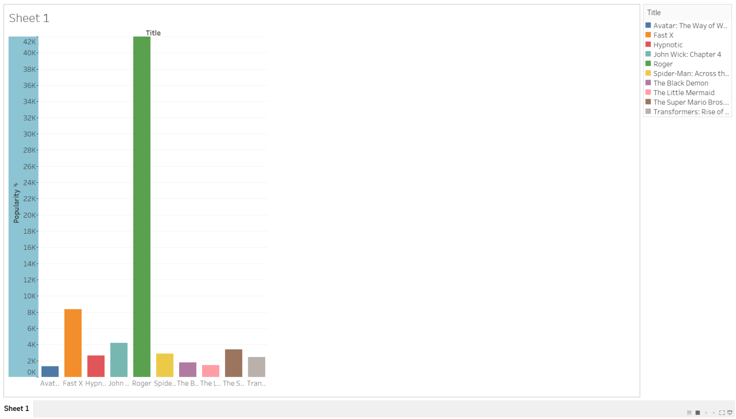
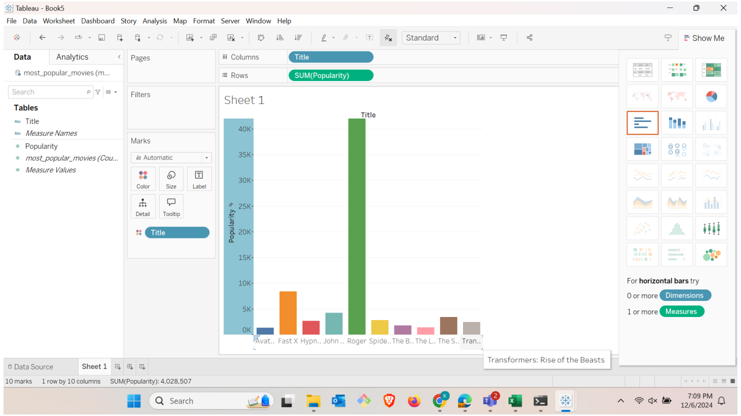


Figure 6: Barchart Most Popular Movies


Columns: Title
Rows: SUM(Popularity)
Marks: Automatic; Color: Title

10 marks 1 row by 10 columns SUM(Popularity): 4,028,507

Figure 7: Barchart Most Popular Movies

2.4 Highest Vote Count Movies

Type of Chart: Bubble Chart (Horizontal) Reason: A bar chart can show the movies with the highest vote counts. This allows easy comparison of movies with significant audience engagement. In Tableau: Drag vote_count to Columns and title to Rows. You can limit the data to the top 10 highest vote counts. In Excel: Use a bar chart to plot the movies against their vote counts.

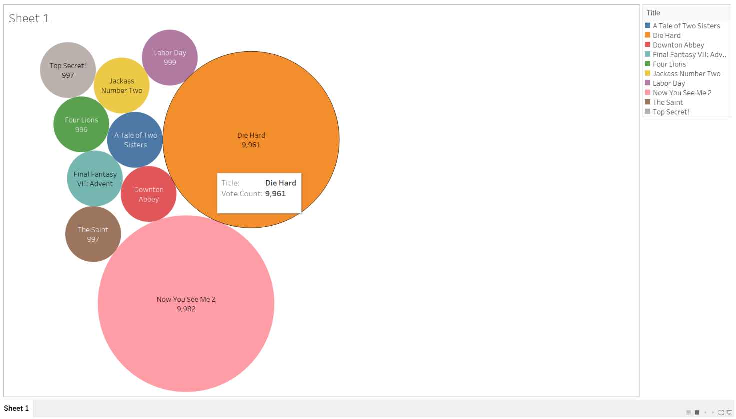


Figure 8: Bubblechart Highest Vote Count Movies

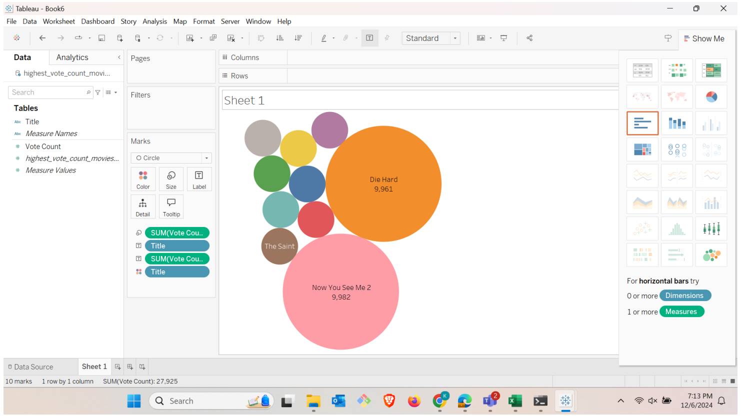


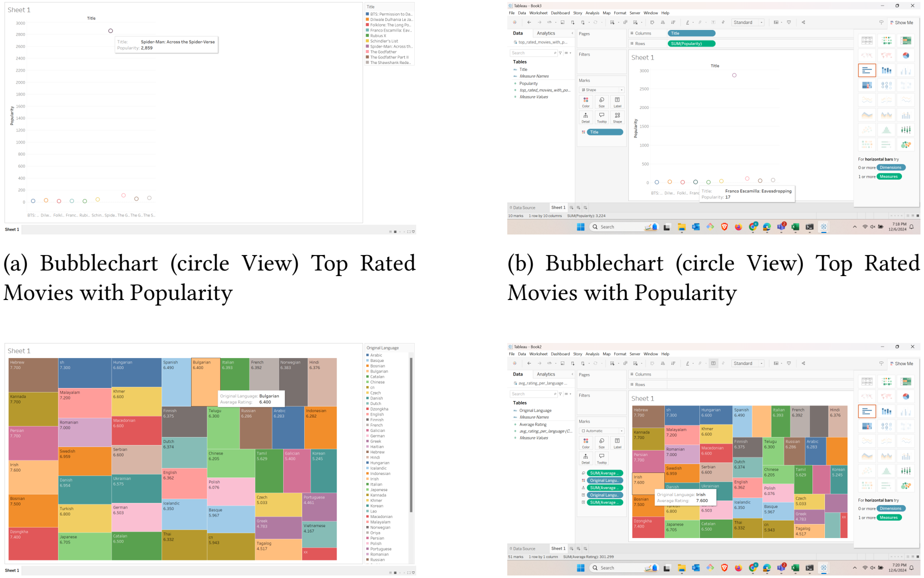
Figure 9: Bubblechart Highest Vote Count Movies

2.5 textbfTop Rated Movies with Popularity

Type of Chart: Scatter Plot or Bubble Chart Reason: A scatter plot will show the relationship between the top-rated movies and their popularity. A bubble chart can be used to show both popularity and rating, with the bubble size representing another factor like vote_count. In Tableau: Use a scatter plot with popularity on the x-axis and vote_average on the y-axis. In Excel: A scatter plot with popularity and vote_average on the x and y axes can help visualize the relationship.

2.6 Average Rating per Language

Type of Chart: Bar Chart or Pie Chart Reason: A bar chart will show how different languages perform in terms of average ratings. A pie chart could be used to show the proportion of languages with higher ratings. In Tableau: Drag average_rating to



(a) Bubblechart (circle View) Top Rated Movies with Popularity

(b) Bubblechart (circle View) Top Rated Movies with Popularity

(c) Tree Map: Average Rating per Language

(d) Tree Map: Average Rating per Language

Figure 10: Visualizations of Top Rated Movies and Language Ratings

Columns and original_language to Rows, and sort by rating. In Excel: Create a bar chart for average ratings per language, or use a pie chart to show the proportion of movies in each language. Data Quality:

Language data was standardized to ensure no discrepancies in naming conventions. 5Vs:

Volume: Handled over 100 languages in the dataset. Veracity: High trust in language-related data, verified through metadata. Visualization and Insights: A bar chart displayed the average ratings for each language, with sorting done to highlight the highest-rated languages. Metrics:

Processing Time: Less than 3 seconds to generate a report for over 100 languages. Resource Utilization: Efficient CPU usage during the generation of the report.

Conclusion

The Movie Rating Prediction system demonstrates the effective application of Big Data technologies to create scalable and efficient recommendation systems. By utilizing Apache Spark's distributed computing and MLlib for machine learning, the system can process large-scale data quickly and accurately.

This project not only showcases the ability to predict user ratings but also provides insights into user behavior and preferences. The methodology and tools used can be adapted for other domains such as e-commerce, music, and book recommendations. The success of this project highlights the significance of distributed computing in addressing real-world data challenges and its potential for future innovations in predictive analytics.

References

1. Lim, Yew Jin and Teh, Yee Whye. *Variational Bayesian approach to movie rating prediction*. Proceedings of KDD cup and workshop, vol. 7, pp. 15–21, 2007.

2. Li, Xiaoyue, Zhao, Haonan, Wang, Zhuo, and Yu, Zhezhou. *Research on movie rating prediction algorithms*. 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), pp. 121–125, 2020.

3. Abarja, Rudy Aditya and Wibowo, Antoni. *Movie rating prediction using convolutional neural network based on historical values*. International Journal, vol. 8, pp. 2156–2164, 2020.

4. Fikir, O Bora, Yaz, Ilker O, and Özyer, Tansel. *A movie rating prediction algorithm with collaborative filtering*. 2010 International Conference on Advances in Social Networks Analysis and Mining, pp. 321–325, IEEE, 2010.

5. Bhadrashetty, Ambresh and Patil, Surekha. *Movie Success and Rating Prediction Using Data Mining*. Journal of Scientific Research and Technology, pp. 1–4, 2024.

6. Marović, Mladen, Mihoković, Marko, Mikša, Mladen, Pribil, Siniša, and Tus, Alan. *Automatic movie ratings prediction using machine learning*. 2011 Proceedings of the 34th International Convention MIPRO, pp. 1640–1645, IEEE, 2011.

7. Armstrong, Nick and Yoon, Kevin. *Movie rating prediction*. Citeseer, 1995.

8. Martinez, Victor R, Somandepalli, Krishna, Singla, Karan, Ramakrishna, Anil, Uhls, Yalda T, and Narayanan, Shrikanth. *Violence rating prediction from movie scripts*. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 671–678, 2019.

9. Ahmad, Javaria, Duraisamy, Prakash, Yousef, Amr, and Buckles, Bill. *Movie success prediction using data mining*. 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–4, IEEE, 2017.

10. El Assady, Mennatallah, Hafner, Daniel, Hund, Michael, Jentner, Wolfgang, Rohrdantz, Christian, Fischer, Fabian, Simon, Svenja, Schreck, Tobias, and Keim, Daniel. *Visual analytics for the prediction of movie rating and box office performance*. Bibliothek der Universität Konstanz, 2014.

11. Sivakumar, Pirunthavi, Rajeswaren, Vithusia Puvaneswaren, Abishankar, Kamalanathan, Ekanayake, E.M.U.W.J.B., and Mehendran, Yanusha. *Movie success and rating prediction using data mining algorithms*. Faculty of Management and Commerce, South Eastern University of Sri Lanka, 2021.

12. Hsu, Ping-Yu, Shen, Yuan-Hong, and Xie, Xiang-An. *Predicting movies user ratings with IMDb attributes*. In *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*, Springer, pp. 444–453, 2014.