Analysis of IMDB Movie Data

Manikanta Gaddameda(S546827)
Shashank pasumarthi(s547089)
Prudhvidhar Reddy Gopireddy(s545253)

1 Project Idea

Our project aims to analyze the IMDB movie dataset in CSV format to gain insights into the movie industry. To achieve this, we will utilize MapReduce and Hadoop for data analysis, along with tools like Kafka, Eclipse, and JupyterLab for development and analysis. Our six primary goals are to identify the top-rated movies of the last 20 years, analyze the trend of movie ratings over time, identify the most popular genres of the last 20 years, analyze the relationship between budget and gross revenue, identify the most successful actors and directors of the last 20 years, and analyze the relationship between runtime and IMDB rating.

2 Tools and Technologies

We plan to use the following tools and technologies for our project:

1. MapReduce and Hadoop for distributed processing of large datasets.
2. Kafka for real-time data streaming.
3. Eclipse for Java development.
4. JupyterLab for data analysis and visualization.
5. CSV for data storage and manipulation.
6. Pandas library for data manipulation and analysis.

3 High-Level Architecture or Methodology

Our high-level architecture for the project is as follows:

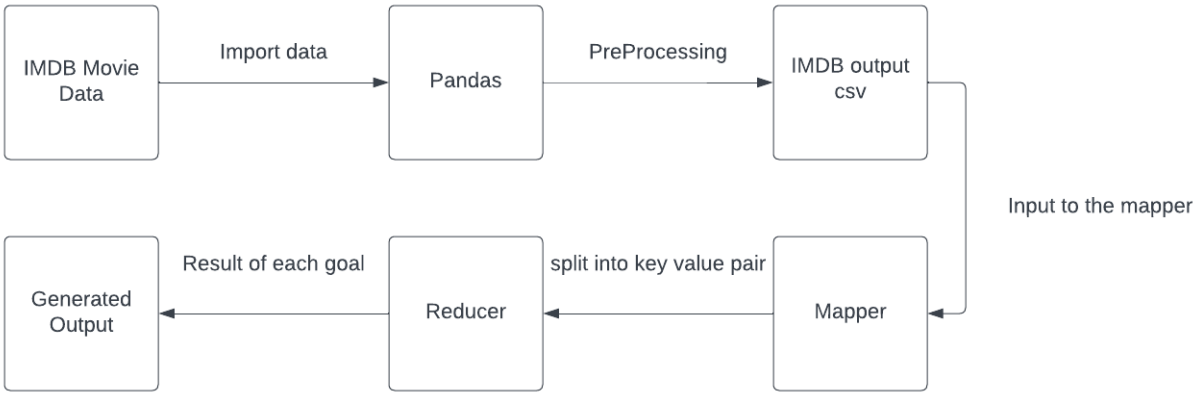

Figure 1: High-level architecture of the project.

4 Explanation of the Diagram

1. **Import data:** The IMDB movie dataset is imported into the system.
2. **Pandas:** The dataset is processed and cleaned using the Pandas library.
3. **Preprocessing:** The dataset is preprocessed to include only movies released in the last 20 years, and the values in the "genres" column are split into multiple columns.
4. **IMDB output csv:** The preprocessed dataset is stored in a CSV file.
5. **Mapper:** The Mapper takes the input data and splits it into key-value pairs.
6. **Reducer:** The Reducer takes the output of the Mapper and combines the data based on the keys.
7. **Result of each goal:** The results of each goal are generated using MapReduce and Hadoop.
8. **Generated Output:** The final output is generated and stored in a CSV file.

5 Goals of the project

Our team wants to investigate the following goals:

1. **To identify the top-rated movies from the dataset of the past 20 years:**

We can use Hadoop and Python to process the dataset and extract the ratings of all movies released in the last 20 years. Then, we can sort the ratings in descending order to determine the top-rated movies.

2. **To analyze the trends in movie ratings over time and gain insights into the factors that affect them:**
   Using Hadoop and Python, we can extract the ratings and release dates of all movies in the dataset. Then, we can create a graph or chart to visualize the trend of movie ratings over time. By analyzing this graph, we can gain insights into the factors that influence movie ratings.

3. **To determine the most popular movie genres in the past 20 years by analyzing the dataset:**
   We can use Hadoop and Python to process the dataset and extract the genres of all movies released in the last 20 years. Then, we can calculate the frequency of each genre to determine the most popular ones.

4. **To analyze the relationship between the movie's budget and its gross revenue to understand the factors that impact profitability:**
   Using Hadoop and Python, we can extract the budget and gross revenue of all movies in the dataset. Then, we can plot a graph or chart to visualize the relationship between the two variables. By analyzing this graph, we can understand the factors that affect the profitability of movies.

5. **To filter the dataset to include only movies released within the past 20 years for further analysis:**
   Using Hadoop and Python, we can filter the dataset to include only movies released in the last 20 years. This will allow us to perform further analysis on the subset of the dataset that is most relevant to our project goals.

6. **To identify the most successful actors and directors based on the movies they have worked on in the past 20 years:**
   Using Hadoop and Python, we can extract the names of all actors and directors who have worked on movies released in the last 20 years. Then, we can calculate their total box office earnings and the number of movies they have worked on in the past 20 years. Using this information, we can determine the most successful actors and directors.