Movie Popularity and Revenue Analysis Data using PySpark

December 5, 2024

1 Project Idea

This project's objective is to use PySpark to evaluate a dataset that includes comprehensive movie information, such as budget, release date, revenue, genres, popularity, and average votes. In order to gain a better knowledge of the trends and metrics that affect a film's performance, this analysis attempts to reveal insights into the elements that affect film success.

The dataset makes it possible to investigate the connections between variables like budget, genres, and production languages and results like popularity and revenue. We can effectively handle and analyze this dataset by utilizing PySpark, a distributed data processing framework, especially when working with large-scale data operations.

In order to assist those involved in the film industry in making wise decisions, the analysis attempts to offer important insights into film production and marketing tactics. Identifying the genres with the highest average revenue or knowing how release dates affect performance, for instance, might help guide future film planning and marketing initiatives.

PySpark is the best option for managing this dataset because of its features, which enable high-speed processing and scalability.

2 Tools and Technologies

1. PySpark - for data processing and analysis
2. Pandas - for data cleaning and manipulation
3. Power BI or Tableau - for data visualization
4. Jupyter Notebook - for coding and analysis
5. Kaggle - for accessing and downloading the dataset
6. Git - for code management and collaboration

3 High-Level Architecture and Methodology

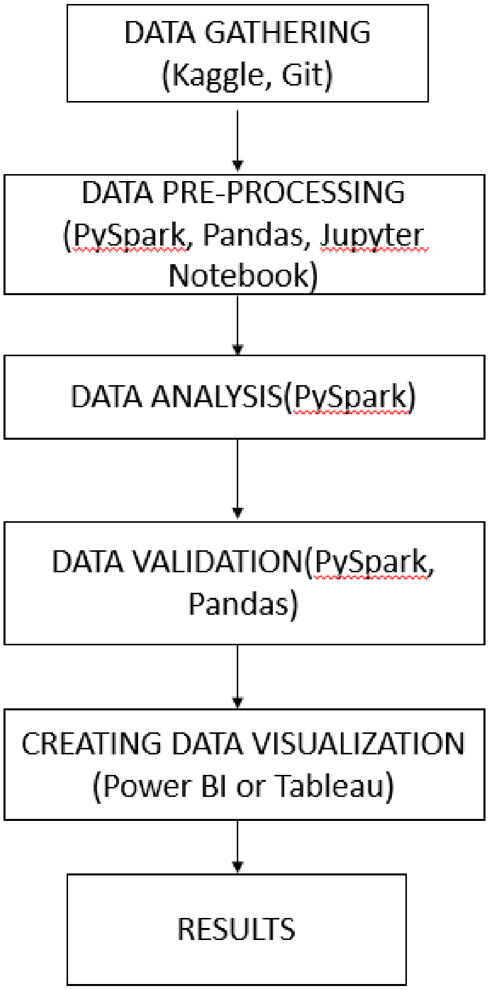

Figure 1: High-Level Architecture of the Project

4 Explanation of the block diagram

The methodology of the project is as follows:

1. **Data Gathering**
   - Access and download the movie dataset from Kaggle.
   - Use Git for version control and collaboration during development.
2. **Data Pre-Processing**
   - Load the dataset into Jupyter Notebook.

- Utilize PySpark and Pandas to clean and manipulate the data.
- Address missing values, standardize formats, and prepare the data for analysis.

3. **Data Analysis**
   - Leverage PySpark for processing large-scale data efficiently.
   - Extract insights like top-grossing movies, popularity trends, and average revenue by genre.
4. **Data Validation**
   - Ensure data consistency and accuracy with PySpark and Pandas validation checks.
   - Validate the correctness of the extracted insights.
5. **Data Visualization**
   - Use tools like Power BI or Tableau to create visual representations of the findings.
   - Generate charts to illustrate trends in revenue, popularity, and other metrics.
6. **Results**
   - Marks the conclusion of the project with all objectives met.

5 Goals

1. List the top ten highest-grossing films together with their titles and earnings.
2. Display movies in descending order by popularity score.
3. Calculate the total number of movies produced in each language.
4. Determine the average revenue generated by movies in each genre.
5. Find the most popular movies in a specific genre, such as "Action" or "Comedy."
6. Fetch movies' names with the longest runtime.
7. Retrieve all movies released in a given year."
8. Count how many movies are in each release status category.