**Netflix Data Analysis Using PySpark**

**Introduction**

In the ag of data-driven decision-making, streaming services like Netflix rely heavily on data analysis for content management and strategic planning. This report outlines a data analysis project using Apache PySpark to explore the Netflix dataset. The project aims to perform exploratory data analysis (EDA) by examining various content attributes such as directors, genres, releas years, and content durations.

**Project Objectives**

The primary objective of this project is to derive meaningful insights from the Netflix dataset. Specific goals include:

1. Understanding the dataset structure.
2. Identifying top directors by the number of titles.
3. Analyzing average release years by content type.
4. Examining content duration distributions.
5. Discovering countries with diverse genres.
6. Identifying the longest titles by word count.
7. Analyzing the distribution of content ratings.

**Data Source and Tools**

The Netflix dataset used in this project contains information about movies and TV shows, including attributes such as title, director, cast, country, release year, rating, and listed genres. Key tools used include:

* **Apache PySpark:** F or scalable data processing.
* **Jupyter Notebook:** For an interactive coding environment.
* **Python Libraries:** Including PySpark functions such as count, avg, max, min, desc, length, and split.

**Methodology**

1. **Data Loading:** The dataset is loaded into a Spark DataFrame using PySpark’s CSV reader with headers and inferred schema.
2. **Data Exploration:** The data schema and record count are displayed to understand the dataset structure.
3. **Analysis Tasks:**
   * **Top Directors:** Grouping by the director field and counting titles.
   * **Average Release Year:** Aggregating average release years by content type.
   * **Content Duration Analysis:** Extracting numerical values from the duration field to calculate average, maximum, and minimum durations.
   * **Diverse Genres by Country:** Counting unique genres listed per country.
   * **Longest Titles:** Using the title length to find records with the longest titles.
   * **Content Rating Distribution:** Grouping by ratings and counting occurrences.

**Key Findings**

1. **Prolific Directors:** The most prolific directors were identified, indicating individuals frequently collaborating with Netflix.
2. **Release Year Trends:** Average release years provided insight into content creation trends.
3. **Duration Patterns:** Movies generally had longer durations compared to TV shows.
4. **Genre Diversity:** The United States and India showed the most diverse content offerings.
5. **Longest Titles:** Certain shows and movies had exceptionally long titles.
6. **Rating Distribution:** The most common ratings were TV-MA and TV-14.

**Challenges and Solutions**

* **Data Inconsistencies:** Missing values in the dataset required filtering and cleaning.
* **Schema Complexity:** Proper schema inference and type casting were implemented to handle mixed data types.

**Conclusion**

The Netflix EDA project successfully demonstrated PySpark's capabilities in processing large datasets efficiently. The analysis provided actionable insights into content trends, enabling informed decision-making for similar data-driven projects.

**Future Work**

* Expanding analysis with more granular data.
* Incorporating sentiment analysis on user reviews.
* Applying machine learning models for content recommendation.

**Acknowledgments**

* Netflix for the dataset.
* Apache Spark community for maintaining a robust big data framework.

This project underscores the importance of large-scale data analysis in today’s streaming-driven media landscape.

[Github link](https://github.com/Mujtabakhakwani/DS-Assignment.git)