Google Playstore Data Analytics using PySpark

Abstract

The rapid growth of mobile applications has made the Google Play Store a central hub for developers and users alike. This study conducts an extensive data analysis of over 10,000 apps from the Google Play Store using PySpark—a big data framework—to uncover trends influencing app popularity, monetization, and user engagement. The study emphasizes data cleaning, transformation, and visualization techniques that extract valuable business insights from large datasets.

Introduction

Mobile applications have transformed the way users interact with technology, offering services across communication, entertainment, education, and productivity. Understanding app characteristics that contribute to success is essential for developers and business analysts. The Google Playstore Dataset provides detailed attributes like app category, ratings, installs, content rating, price, and size, offering an ideal opportunity for Big Data Analytics using PySpark. This research aims to perform data cleaning, explore trends, and identify key features impacting app performance.

Dataset Description

The dataset contains 10,837 records and 13 attributes including app name, category, rating, reviews, installs, size, type, price, content rating, and update information. It provides a broad view of app distribution across multiple genres and helps identify user engagement trends.

Methodology

Tools: Python, PySpark, Matplotlib, Seaborn. Data Cleaning: Handled missing values, converted data types, filtered invalid entries. Exploratory Analysis: Computed app counts, free vs paid ratio, average installs, and top-performing categories. Visualization: Bar charts, histograms, and heatmaps were used to represent app distribution and relationships.

Results and Discussion

Key Findings: • Total Apps: 10,837 • Unique Categories: 33 • Free Apps: 92% • Paid Apps: 8% Top categories: Family, Games, Tools. Most installed categories: Communication, Social, Video

Players. Paid apps have higher ratings but fewer installs. User ratings cluster between 4.0 and 4.7, reflecting strong satisfaction.

Conclusion

The analysis reveals that category and pricing strategies strongly influence app success. PySpark efficiently handled large-scale data, proving its effectiveness in big data analytics. Future work can involve predictive modeling to forecast app success using advanced machine learning algorithms.

References

1. Google Play Store Dataset (Kaggle, 2021) 2. Apache Spark Documentation 3. Han, J., & Kamber, M. (2012). Data Mining: Concepts and Techniques 4. Research on Mobile App Market Trends, IEEE Access, 2020