Research Paper on Google Playstore Data Analytics using PySpark*

1. 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.

---

2. 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:

1. Perform data cleaning and preprocessing on inconsistent app data.
2. Explore data-driven insights using descriptive statistics and visualization.
3. Identify key features impacting app performance on the Play Store.

---

3. Dataset Description

The dataset contains *10,837 records* and *13 attributes*, including:

* *App Information:* Name, Category, Genres
* *User Metrics:* Rating, Reviews, Installs
* *App Properties:* Size, Type (Free/Paid), Price
* *Technical Metadata:* Content Rating, Last Updated, Android Version

Data source: googleplaystore.csv

---

4. Methodology

4.1 Tools and Framework

* *Programming Language:* Python
* *Framework:* Apache Spark (PySpark)
* *Visualization:* Matplotlib, Seaborn
* *Environment:* Jupyter Notebook

4.2 Data Cleaning and Preprocessing

* Removed corrupted entries (e.g., invalid app rows).
* Converted numeric attributes (Installs, Price, Size, Reviews) to appropriate data types.
* Handled missing values using mean imputation for continuous variables.
* Filtered non-relevant or malformed rows (e.g., apps with invalid types).

4.3 Exploratory Data Analysis (EDA)

* Computed total number of apps, unique categories, and distribution of free vs paid apps.
* Identified most popular app categories.
* Analyzed rating distributions and install patterns.

* Computed average installs per category and visualized the relationship between price and rating.

4.4 Visualization

Visualizations include:

* *Bar charts* for category-wise app distribution.
* *Pie charts* showing Free vs Paid apps.
* *Heatmaps* to examine correlations.
* *Histograms* for rating and install distributions.

---

5. Results and Discussion

5.1 Dataset Summary

* *Total Apps:* 10,837
* *Unique Categories:* 33
* *Free Apps:* 92% (≈10,037)
* *Paid Apps:* 8% (≈800)

5.2 Insights

* *Top 3 Categories by Volume:* Family, Games, Tools
* *Most Installed Categories:* Communication, Social, Video Players
* *Highest Average Rating:* Education and Art & Design
* *Monetization Trend:* Paid apps are relatively few but often have higher ratings and smaller install counts.
* *App Size Trend:* No strong correlation between app size and installs, but very large apps tend to have fewer users.

5.3 Interpretation

Free apps dominate the Play Store, but high-quality paid apps retain loyal user bases. User ratings tend to cluster between 4.0 and 4.7, indicating generally high satisfaction levels. The categories Communication and Social lead in average installs, reflecting user dependence on connectivity and social interaction platforms.

---

6. Conclusion

The analysis demonstrates that:

* *Category and pricing models* strongly influence app installs and ratings.

* *Data preprocessing* is vital in transforming raw data into meaningful insights.

* *PySpark* effectively handles large-scale datasets, making it ideal for big data analytics in app markets.

Future work could include predictive modeling using *machine learning techniques* to estimate app success based on features like category, size, and type.

---

7. References

1. Google Play Store Dataset (Kaggle, 2021).

2. Apache Spark Documentation.

3. Python Data Analysis Library (pandas, matplotlib, seaborn).

4. Han, J., & Kamber, M. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.

5. Research on Mobile App Market Trends, IEEE Access, 2020.