High Level Design (HLD)

Analysing Google Apps Store dataset

Presented By:

Ishita Shetty

Document Version Control

Date Issued	Version	Description	Author
07th June 2022	1.0	Abstract, Introduction, General Description	Ishita Shetty
08th June 2022	1.1	Design Detail, KPI, Deployment	Ishita Shetty
22nd June 2022	1.2	Final Revision	Ishita Shetty

Contents

Abstract

Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created.

This project aims to predict the ratings of Google Play Store apps using Python libraries. I have performed Data Analysis into the Google Play store application dataset that has been collected from Kaggle. I discovered the relationships among various attributes present in my dataset such as which application is free or paid, about the user reviews, rating of the application, etc.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:

  Security, Reliability, Maintainability, Portability, Reusability, Application compatibility, Resource utilization and Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

2 General Description

2.1 Product Perspective & Problem Statement

Android is expanding as an operating system. It has captured around 74% of the total market which is a true indicator of the huge amount of population using android. The goal is to help android developers to know what is the motivating factor for people to download an app. It will also help to find out the factors that affect someone's decision to download an app. I would like to analyse category, reviews, price, ratings and installs for this purpose and find out how they are inter related.

The aim of my analysis is to provide insights about android applications and their categories. I want to dive deep in data for the factors of influences on an application, to know why and how certain applications succeed others. Also, what is required for an application to be considered as successfully topping the charts.

2.2 Tools used

Business Intelligence tools and Python libraries works such as Pandas, NumPy, Matplotlib, MS- Excel, Tableau are used to build the whole framework.

Python: It is a popular multi-purpose programming language widely used for its flexibility, as well as its extensive collection of libraries, which are valuable for analytics and complex calculations.

Python's extensibility means that it has thousands of libraries dedicated to analytics, including the widely used Python Data Analysis Library (also known as Pandas).

For the most part, data analytics libraries in Python are at least somewhat derived from the NumPy library, which includes hundreds of mathematical calculations, operations, and functions. Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Tableau: It is used to turn unrelated sources of data into coherent, visually immersive, and interactive insights.

3 Design Details

3.1 Functional Architecture

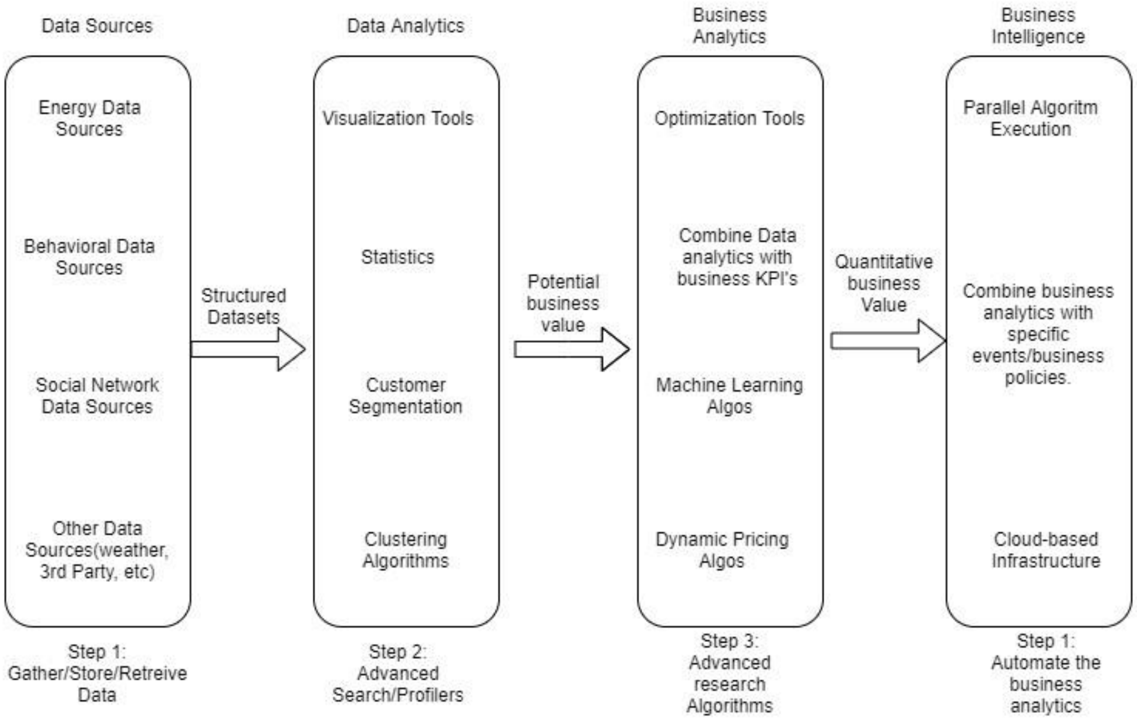
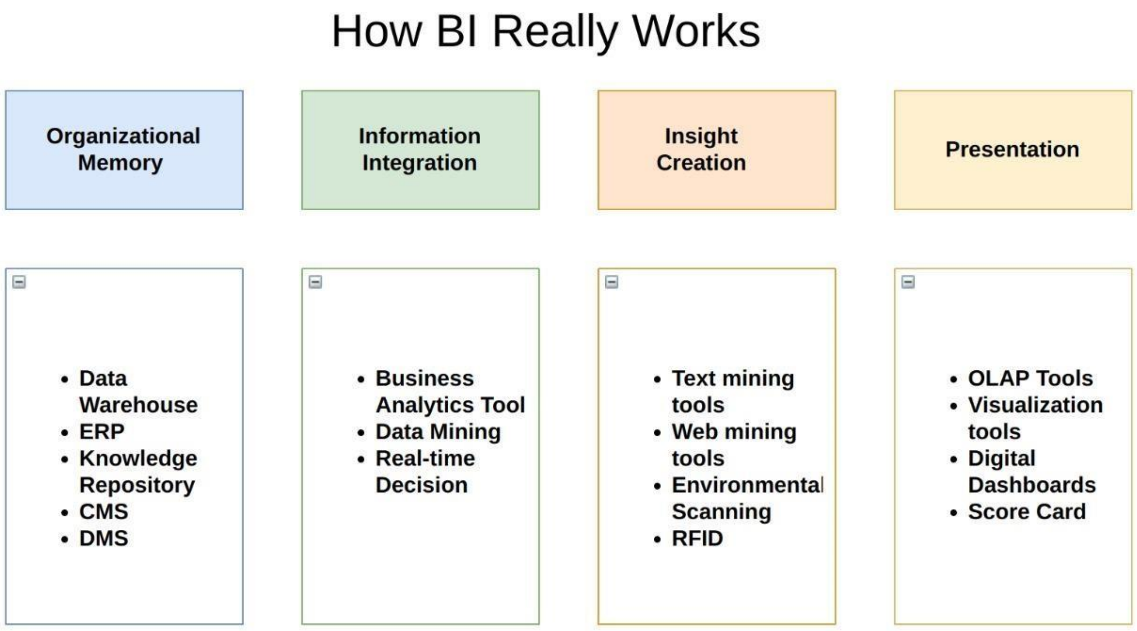

Data Sources
Energy Data Sources
Behavioral Data Sources
Social Network Data Sources
Other Data Sources(weather, 3rd Party, etc)
Step 1: Gather/Store/Retreive Data
Structured Datasets
Data Analytics
Visualization Tools
Statistics
Customer Segmentation
Clustering Algorithms
Step 2: Advanced Search/Profilers
Potential business value
Business Analytics
Optimization Tools
Combine Data analytics with business KPI's
Machine Learning Algos
Dynamic Pricing Algos
Step 3: Advanced research Algorithms
Quantitative business Value
Business Intelligence
Parallel Algoritm Execution
Combine business analytics with specific events/business policies.
Cloud-based Infrastructure
Step 1: Automate the business analytics

Figure 1: Functional Architecture of Business Intelligence



3.2 Optimization

Your data strategy drives performance

- Minimize the number of fields
- Minimize the number of records
- Optimize extracts to speed up future queries by materializing calculations, removing columns and the use of accelerated views

Reduce the marks (data points) in your view

- Practice guided analytics. There's no need to fit everything you plan to show in a single view. Compile related views and connect them with action filters to travel from overview to highly-granular views at the speed of thought.
- Remove unneeded dimensions from the detail shelf.
- Explore. Try displaying your data in different types of views. **Limit your filters by number and type**
- Reduce the number of filters in use. Excessive filters on a view will create a more complex query, which takes longer to return results. Double-check your filters and remove any that aren't necessary.
- Use an include filter. Exclude filters load the entire domain of a dimension, while include filters do not. An include filter runs much faster than an exclude filter, especially for dimensions with many members.
- Use a continuous date filter. Continuous date filters (relative and range-of-date filters) can take advantage of the indexing properties in your database and are faster than discrete date filters.
- Use Boolean or numeric filters. Computers process integers and Booleans (t/f) much faster than strings.
- Use parameters and action filters. These reduce the query load (and work across data sources).

Optimize and materialize your calculations

- Perform calculations in the database
- Reduce the number of nested calculations.
- Reduce the granularity of LOD or table calculations in the view. The more granular the calculation, the longer it takes.
  - LODs - Look at the number of unique dimension members in the calculation.
  - Table Calculations - the more marks in the view, the longer it will take to calculate.
- Where possible, use MIN or MAX instead of AVG. AVG requires more processing than MIN or MAX. Often rows will be duplicated and display the same result with MIN, MAX, or AVG.

- Make groups with calculations. Like include filters, calculated groups load only named members of the domain, whereas Tableau’s group function loads the entire domain.
- Use Booleans or numeric calculations instead of string calculations. Computers can process integers and Booleans (t/f) much faster than strings. Boolean>Int>Float>Date>DateTime>String

4 KPIs

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the apps.

Power BI

tableau

As and when, the system starts to capture the historical/periodic data for a user, the dashboards will be included to display charts over time with progress on various indicators or factors.

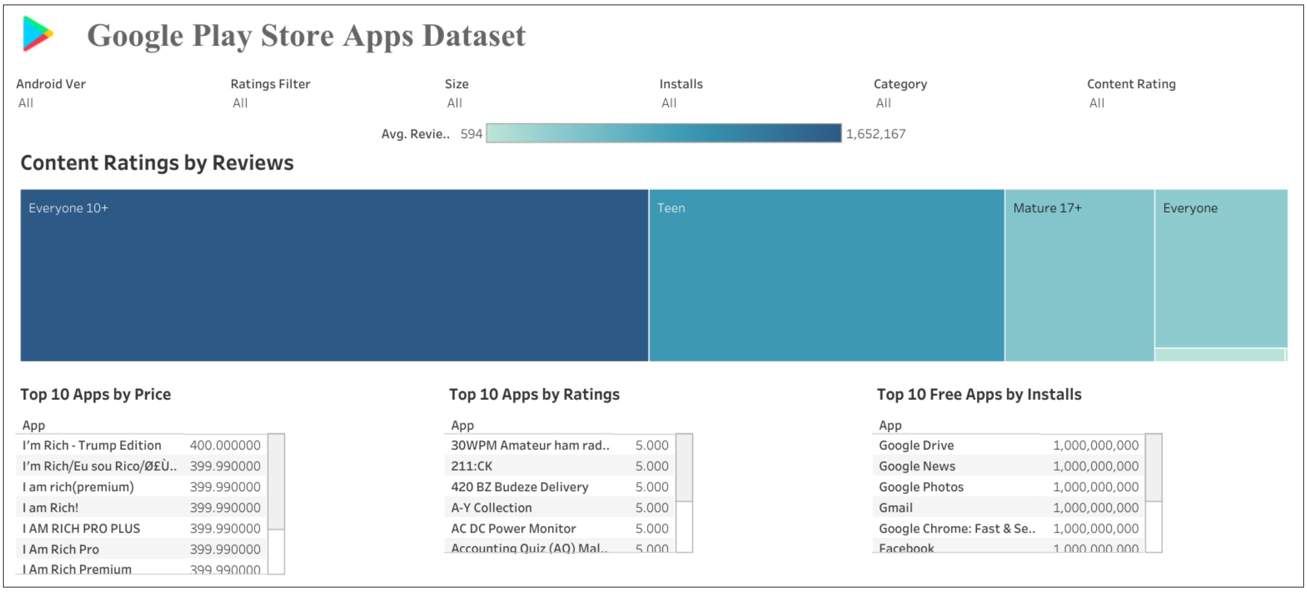
4.1 KPIs (Key Performance Indicators)

Key indicators displaying a summary of the the apps in Google Play Store and its relationship with different metrics

1. Content Ratings by Reviews
2. Top 10 Apps by Ratings
3. Top earning apps
4. Top 10 Apps by Installs
5. Highest and Lowest rated genres
6. Count of application in each category differentiated by their type
7. Percentage of Review Sentiments

5 Deployment

Prioritizing data and analytics couldn’t come at a better time. A company, no matter what size, is already collecting data and most likely analysing just a portion of it to solve business problems, gain competitive advantages, and drive enterprise transformation. With the explosive growth of enterprise data, database technologies, and the high demand for analytical skills, today’s most effective IT organizations have shifted their focus to enabling self-service by deploying and operating Tableau at scale, as well as organizing, orchestrating, and unifying disparate sources of data for business users and experts alike to author and consume content.



Google Play Store Apps Dataset

Android Ver	Ratings Filter	Size	Installs	Category	Content Rating
All	All	All	All	All	All

Top 10 Apps by Price

App	
I’m Rich - Trump Edition	400.000000
I’m Rich/Eu sou Rico/Ø£Ù..	399.990000
I am rich(premium)	399.990000
I am Rich!	399.990000
I AM RICH PRO PLUS	399.990000
I Am Rich Pro	399.990000
I Am Rich Premium	399.990000

Top 10 Apps by Ratings

App	
30WPM Amateur ham rad..	5.000
211:CK	5.000
420 BZ Budeze Delivery	5.000
A-Y Collection	5.000
AC DC Power Monitor	5.000
Accounting Quiz (AQ) Mal..	5.000

Top 10 Free Apps by Installs

App	
Google Drive	1,000,000,000
Google News	1,000,000,000
Google Photos	1,000,000,000
Gmail	1,000,000,000
Google Chrome: Fast & Se..	1,000,000,000
Facebook	1,000,000,000

Tableau prioritizes choice in flexibility to fit, rather than dictate, your enterprise architecture. Tableau Server and Tableau Online leverage your existing technology investments and integrate into your IT infrastructure to provide a self-service, modern analytics platform for your users. With on-premises, cloud, and hosted options, there is a version of Tableau to match your requirements.

.