HIGH LEVEL DESIGN (HLD)

|  |  |
| --- | --- |
| **Written By** | **Saud Qureshi** |
| **Document Version** | **1.0** |
| **Last Revised Data** | **18-Dec-2024** |

Document Version Control:

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | 18-Dec-2024 | Saud Qureshi |  |
|  |  |  |  |

Contents

[1 Abstract 3](#_Toc188366555)

[2 Introduction 3](#_Toc188366556)

[2.1 Why High-level Design? 3](#_Toc188366557)

[3 General Description 4](#_Toc188366558)

[3.1 Product Perspective 4](#_Toc188366559)

[3.2 Problem Statement 4](#_Toc188366560)

[3.3 Proposed Solution 4](#_Toc188366561)

[3.4 Technical Requirements 4](#_Toc188366562)

[3.5 Data Requirements 4](#_Toc188366563)

[3.6 Tools Used 4](#_Toc188366564)

[3.7 Constraints 5](#_Toc188366565)

[3.8 Assumptions 5](#_Toc188366566)

[4 Design Details 5](#_Toc188366567)

[4.1 For Preprocessing and Transformation 5](#_Toc188366568)

[4.2 Deployment Process 5](#_Toc188366569)

[4.3 Event Log 5](#_Toc188366570)

[4.4 Error Handling 5](#_Toc188366571)

[5 Performance 6](#_Toc188366572)

[5.1 Reusability 6](#_Toc188366573)

[5.2 Application Compatibility 6](#_Toc188366574)

[5.3 Deployment 6](#_Toc188366575)

[5.4 User Interface 6](#_Toc188366576)

# Abstract

This project demonstrates the implementation of an end-to-end real-time data analytics solution using PySpark, a powerful distributed computing framework. The project focuses on processing and analyzing prescription data, encompassing aspects such as data ingestion, transformation, and visualization to draw meaningful insights.

The dataset includes key features like prescriber details, prescription counts, drug costs, and city-wise demographics, providing a comprehensive base for exploration and analysis. The project follows a structured pipeline:

1. **Data Ingestion**: Real-world prescription and city demographic data are ingested from multiple sources into a distributed environment for scalability.
2. **Data Processing**: Utilizing PySpark, the data is cleaned, transformed, and enriched to ensure accuracy and consistency for downstream analysis.
3. **Exploratory Data Analysis (EDA)**: Insightful dashboards and visualizations are created to reveal patterns, such as trends in drug costs, prescription counts by region, and the impact of population on prescription activities.
4. **Interactive Dashboards**: With Streamlit, dynamic dashboards are built to showcase visualizations like bar charts, scatter plots, and distribution graphs, providing stakeholders with actionable insights.
5. **Business Insights**: The project highlights key metrics, such as top prescribers, cost patterns, and geographic prescription distributions, supporting decision-making in healthcare and pharmaceutical domains.

This real-time project equips learners with practical skills in PySpark for big data processing, alongside modern tools like Streamlit for interactive visualization. It bridges the gap between data engineering and analytics, empowering professionals to handle large-scale datasets effectively and deliver impactful solutions in a real-world context.

# Introduction

## Why High-level Design?

The purpose of this High Level Design (HLD) Document is to add the necessary detail description to represent a suitable model. This document is designed to help in operational requirement and can be used as a reference manual for how the modules interact. Basically, HLD is a technical representation of functional requirements and flow of information across assets or components.

# General Description

## Product Perspective

This PySpark-based project enables real-time analysis of large healthcare datasets, offering insights into prescription patterns, drug costs, and regional trends. Using Hadoop for data management and Streamlit for interactive dashboards, it delivers scalable, user-friendly solutions for data-driven decision-making in healthcare and beyond.

## Problem Statement

Healthcare and pharmaceutical organizations face challenges in processing and analyzing vast amounts of data in real-time to identify trends, optimize costs, and make informed decisions. Traditional analytics tools struggle with scalability and efficiency, hindering timely insights into prescription patterns, drug usage, and geographic variations. This project addresses these challenges by leveraging big data technologies to provide a scalable, real-time solution for analyzing healthcare datasets, enabling actionable insights for stakeholders.

## Proposed Solution

The proposed solution leverages PySpark for distributed data processing, Hadoop for scalable data storage, and Streamlit for interactive dashboards to enable real-time analysis of healthcare datasets. This system processes large-scale data efficiently, identifies trends in prescription patterns and drug costs, and visualizes insights through user-friendly dashboards. By integrating big data technologies with advanced analytics, the solution empowers healthcare organizations to make data-driven decisions, optimize costs, and improve service delivery.

## Technical Requirements

In this project we are having a set of requirements and they are given below

a) Model should be exposed through API or User Interface, so that anyone can test model.

b) Model should be deployed on streamlit.

## Data Requirements

Data Requirement completely depends on our problem.

a) We are using the data from Kaggle (csv and parquet file).

## Tools Used

* PyCharm is used as IDE.
* For visualization of the plots, Matplotlib, Seaborn are used.
* Streamlit is used for UI.

## Constraints

Dataset is small.

## Assumptions

Data acquired is authentic.

# Design Details

## For Preprocessing and Transformation

The preprocessing and transformation design involves several key steps:

1. **Data Ingestion**: Raw data is loaded into Hadoop Distributed File System (HDFS) or directly into PySpark from various sources like CSV or databases.
2. **Data Cleaning**: Missing values, duplicates, and inconsistencies are handled using PySpark DataFrames to ensure data quality.
3. **Feature Engineering**: Relevant columns are selected, new features are derived, and data is formatted for analysis (e.g., encoding categorical variables).
4. **Data Transformation**: Data is aggregated, filtered, and normalized to prepare it for visualization and analytics, ensuring scalability and efficiency.
5. **Storage**: Cleaned and transformed data is stored in HDFS or passed to the visualization layer.

This pipeline ensures large datasets are efficiently processed for real-time insights.

## Deployment Process

It was deployed to Streamlit.

## Event Log

In this Project we are logging every process so that the user will know what process is running internally. We have designed logging in such a way that debugging will be easy task.

## Error Handling

We have designed this project in such a way that, at any step if error occur then our application should not terminate rather it should catch that error and display that error with proper explanation as to what went wrong during process flow.

# Performance

Solution of Flight fare prediction is used to predict the flight fare in advance, so it should be as accurate as possible so that it should give as much as possible accurate price prediction.

## Reusability

We have done programming of this project in such a way that it should be reusable. So that anyone can add and contribute without facing any problems.

## Application Compatibility

The different module of this project is using Python as an interface between them. Each module performs its own job to perform as per defined by user.

## Deployment

We have chosen Streamlit for deployment of this project.

## User Interface

Below is how user interface looks like to end user.

