# Creating Data Pipeline for Healthcare Insurance Analysis

## Requirements Specifications Document

## 1. Introduction

This document outlines the requirements for the Health Care Data Pipeline Project, which aims to enhance revenue and understand customer behavior through the analysis of various data sources. This introduction sets expectations that we will refer back to throughout the SRS.

### a. Purpose

The purpose of this document is to define the requirements for developing data pipelines that will enable a Health Care insurance company to analyze customer behavior, send customized offers, calculate and distribute royalties, and develop business strategies to enhance revenue.

### b. Intended Audience and Use

The intended audience for this document includes developers, testers, project managers, and business analysts. Each group will use the SRS as follows: developers will use it to understand the system requirements and implement the data pipelines; testers will create and execute test cases based on the defined requirements; project managers will plan, track progress, and ensure the project meets the business objectives; and business analysts will ensure the requirements align with business goals and validate the final solution.

### c. Product Scope

The product aims to leverage the Big Data Ecosystem to analyze data received from various sources such as web scraping and third-party vendors. The objectives include analyzing customer behaviors, sending customized offers to customers, calculating and distributing royalties to past customers, and developing appropriate business strategies to enhance revenue.

### d. Definitions and Acronyms

This section provides definitions for key terms, acronyms, and abbreviations used in the SRS, such as SRS (Software Requirements Specification), AWS (Amazon Web Services), EMR (Elastic MapReduce), ETL (Extract, Transform, Load), PySpark (Python API for Spark), and

Databricks (a unified analytics platform). Clear definitions help eliminate ambiguity and ensure all parties can easily understand the document.

# 2. Overall Description

## a. User Needs

The Health Care Data Pipeline Project is needed to help the insurance company understand customer behaviors and conditions better, enabling them to customize offers and calculate royalties effectively. The primary users include data analysts, who will perform data analysis and generate insights; the marketing team, who will create and send customized offers based on insights; the finance team, who will calculate and distribute royalties; and management, who will develop and implement business strategies based on data insights.

## b. Assumptions and Dependencies

We assume that the data sources will be accessible and provide the necessary data for analysis. The system will be built on AWS infrastructure, utilizing services such as S3, Redshift, EMR, and Databricks. The project will use PySpark for data processing, and the implementation will be compatible with existing systems and technologies used by the company.

# 3. System Features and Requirements

## a. Functional Requirements

- The system should be able to identify the disease with the maximum number of claims by analyzing the claims data.
- It should find subscribers under the age of 30 who subscribe to any subgroup.
- The system should determine which group has the maximum number of subgroups by analyzing the data.
- It should identify the hospital that serves the most number of patients.
- The system should find out which subgroup is subscribed to the most.
- It should calculate the total number of claims that were rejected.
- The system should determine the city from which most claims are coming.
- It should identify whether subscribers mostly subscribe to government or private policies.
- The system should calculate the average monthly premium paid by subscribers to the insurance company.
- It should find out which group is most profitable.
- The system should list all patients below age 18 who are admitted for cancer.
- It should list patients who have cashless insurance and total charges greater than or equal to Rs. 50,000.
- The system should list female patients over the age of 40 that have undergone knee surgery in the past year.

## b. External Interface Requirements

The system should provide a user-friendly interface for data analysts and business users, enabling them to run analyses, view results, and generate reports. It should be compatible with the company's existing hardware, ensuring efficient operation on current systems. The system should integrate seamlessly with existing software systems and tools, and enable secure communication between different system components using secure protocols for data transmission and communication between services.

## c. System Features

The system should include features such as automated data ingestion from multiple sources, efficient data processing using PySpark, secure data storage in AWS S3 and Redshift, and the ability to generate reports and visualizations using Databricks.

## d. Nonfunctional Requirements

The system should handle large volumes of data efficiently, ensuring it processes data within acceptable timeframes. It should ensure data integrity and prevent data loss through data backup and recovery mechanisms. The system must ensure data security and compliance with relevant regulations by implementing access controls, encryption, and secure data transmission. Usability is crucial, so the system should be easy to use and navigate, providing a user-friendly interface and documentation. Finally, the system should be scalable to accommodate increasing data volumes and users, ensuring both data processing and storage capabilities can grow as needed.