Requirements Specifications Document

# Introduction - *This introduction is very important as it sets expectations that we will come back to throughout the SRS.*

## Purpose - *The purpose of this project is to develop a Big Data analytics pipeline to assist a health care insurance company in improving revenue and understanding customer behavior. This involves analyzing competitor data from various sources, enabling the company to track customer behavior, customize insurance offers, and calculate royalties for past policyholders.*

## Intended Audience and Use - *This Software Requirements Specification (SRS) document is intended for use by:*

## **Developers**: For building and implementing the data pipeline and analytics modules.

## **Testers**: For validating data accuracy, integrity, and output.

## **Project Managers**: For tracking progress, resource allocation, and project deliverables.

## **Stakeholders**: For understanding the project scope, business value, and outcomes.

## Product Scope - *This project aims to:*

## Implement a Big Data pipeline for data ingestion, cleaning, and analysis.

## Enhance revenue by analyzing customer data and providing personalized offers.

## Store and process large datasets using AWS infrastructure (S3, Redshift, EMR).

## Improve decision-making by providing insights through analytics dashboards.

## Definitions and Acronyms –

## AWS: Amazon Web Services

## S3: Amazon Simple Storage Service

## EMR: Elastic MapReduce

## PySpark: Python API for Apache Spark

## SRS: Software Requirements Specification

# Overall Description -

## User Needs –

## ***Insurance Analysts****: Need insights into customer behavior and claim patterns.*

## ***Marketing Teams****: Require data to design personalized offers.*

## ***Management****: Requires analytics for strategic decision-making and revenue growth*

## Assumptions and Dependencies –

## *Data is received via web scraping and third-party APIs.*

## *AWS infrastructure (S3, Redshift, EMR) will be available.*

## *Users have access to Databricks for visualization and analysis.*

# System Features and Requirements -

## Functional Requirements –

## *Identify the disease with the maximum number of claims.*

## *Find subscribers under 30 years old who subscribe to any subgroup.*

## *Determine which group has the maximum number of subgroups.*

## *Identify the hospital serving the highest number of patients.*

## *Discover the most subscribed subgroup.*

## *Calculate the total number of rejected claims.*

## *Identify the city generating the most claims.*

## *Analyze whether government or private policy groups are more popular.*

## *Calculate the average monthly premium paid by subscribers.*

## *Identify the most profitable group.*

## *List patients below 18 years old admitted for cancer.*

## *Find patients with cashless insurance and charges ≥ $. 50,000.*

## *List female patients over 40 who underwent knee surgery in the past year.*

## External Interface Requirements –

## *User Interfaces: Visualizations on Databricks.*

## *Software Interfaces: Integration with AWS services (S3, Redshift, EMR).*

## *Communication Interfaces: Data pipeline leveraging AWS and PySpark.*

## System Features –

## *Data ingestion from multiple sources.*

## *Data cleaning and transformation.*

## *Storing results in AWS Redshift.*

## *Generating insights and visualizations.*

## *.*

## Nonfunctional Requirements –

## *Performance: Process large datasets efficiently.*

## *Scalability: Scale with increasing data volume.*

## *Security: Ensure data privacy and secure access.*

## *Usability: Provide actionable insights via Databricks dashboards.*