Software Requirements Specification (SRS)

# 1. Introduction

## a. Purpose

The purpose of this document is to outline the software requirements for a big data analytics solution aimed at enhancing revenue and understanding customer behavior for a healthcare insurance company. The system will involve creating data pipelines using AWS S3, AWS Redshift, Databricks, and PySpark.

## b. Intended Audience and Use

This document is intended for:  
- Developers: To guide the implementation of the system.  
- Testers: To ensure all functional and non-functional requirements are met.  
- Project Managers: To track the project's progress and ensure alignment with business goals.  
- Stakeholders: To understand the project's scope and deliverables.

## c. Product Scope

The primary goal is to design and implement a scalable data pipeline solution that aggregates, cleans, and processes large volumes of customer data to generate actionable insights. The system will improve decision-making, customer targeting, and overall business performance.

## d. Definitions and Acronyms

- AWS S3: Amazon Web Services Simple Storage Service  
- AWS Redshift: Amazon Web Services Redshift, a data warehousing service  
- Databricks: A cloud-based data engineering platform  
- PySpark: The Python API for Apache Spark  
- EMR: Elastic MapReduce, a cloud-based big data platform  
- JIRA: A project management tool used to track sprints and tasks

# 2. Overall Description

## a. Product Perspective

This product is a new data pipeline system designed to integrate with the healthcare insurance company’s existing infrastructure. It will interact with various AWS services and Databricks for data processing and storage.

## b. User Needs

- Data Engineers: Need to efficiently clean and process raw data into structured formats.  
- Data Scientists: Require easy access to processed data for analysis and model building.  
- Business Analysts: Need to generate reports and dashboards to gain insights into customer behavior.

## c. Assumptions and Dependencies

The project assumes the availability of AWS and Databricks infrastructure.  
Dependencies include existing data sources and tools like AWS S3, AWS Redshift, Databricks, and PySpark.

# 3. System Features and Requirements

## a. Functional Requirements

- Data Ingestion: The system shall ingest raw data from AWS S3 and Azure Data Lake.  
- Data Processing: The system shall clean and transform data using PySpark in Databricks.  
- Data Storage: The processed data shall be stored in AWS Redshift for further analysis.  
- Data Access: The system shall provide mechanisms for data scientists and analysts to access the processed data.  
- Data Output: The system shall generate output reports in HTML format.

## b. External Interface Requirements

- User Interface: A web-based dashboard for accessing reports and monitoring pipeline status.  
- Hardware Interface: Integration with existing cloud infrastructure on AWS.  
- Software Interface: Integration with AWS services like S3, Redshift, and Databricks.

## c. System Features

- Data Pipeline Orchestration: The system will use Databricks workflows to orchestrate the data pipeline.  
- Error Handling: The system shall log errors and provide notifications for failures.  
- Scalability: The system shall scale to handle large volumes of data.

## d. Nonfunctional Requirements

- Performance Requirements: The system should process data within a defined SLA (e.g., within 4 hours for a full data pipeline run).  
- Security Requirements: The system shall comply with healthcare data security standards like HIPAA.  
- Usability Requirements: The system shall have an intuitive user interface for non-technical users.  
- Scalability Requirements: The system shall scale seamlessly with increased data volumes.