Template - Requirements Specifications Document

# Introduction - This section will explain the purpose of analyzing healthcare data to boost revenue and gain deeper insights into customer behavior.

## Purpose -The purpose is to outline the requirements for building data pipelines to analyze healthcare insurance data. The goal is to help the company increase revenue by personalizing offers and calculating customer royalties.

## Intended Audience and Use - The intended audience includes data engineers, developers, testers, and project managers involved in this project. The document will guide the development, testing, and implementation of the data pipelines.

* Data Engineers: They will design, build, and maintain the data pipelines. Their tasks include data cleaning, transforming data, and ensuring data flows smoothly through the system.
* Developers: They will write the code needed to implement the data pipelines and integrate different components of the system. They’ll also work on any custom tools or applications required for the project.
* Testers: They will test the data pipelines and related systems to ensure everything works correctly. This includes verifying data accuracy, performance testing, and identifying any bugs or issues.
* Project Managers: They will oversee the project, ensuring it stays on track and meets deadlines. They’ll coordinate between different teams, manage resources, and ensure that the project aligns with the overall business goals.

## Product Scope - The product aims to analyze healthcare data to identify patterns and insights that can drive business decisions. The scope includes setting up the data environment, cleaning the data, and implementing data pipelines for analysis.

Definitions and Acronyms –

 **AWS S3:** Cloud storage for saving and retrieving large amounts of data.

 **AWS Redshift:** A cloud-based data warehouse for analyzing big data using SQL.

 **AWS EMR:** A service for processing big data using tools like Spark and Hadoop.

 **AWS Glue:** A tool for preparing and moving data for analysis.

 **AWS Lambda:** Runs code automatically in response to events without needing servers.  **AWS IAM:** Manages user access and permissions for AWS resources.

Overall Description - The project focuses on analyzing data from different sources to help a healthcare insurance company better understand customer behavior and increase revenue. It's essential for the company to make informed decisions and tailor their services to meet customer needs.

1. User Needs - The product will be used by business analysts to gain insights from the data and by data engineers to build and manage the data pipelines.

## Assumptions and Dependencies - We assume that the project will utilize AWS services, that the necessary datasets will be readily available, and that some of the data will come from third-party sources, which we will rely on for comprehensive analysis.

# System Features and Requirements -

## Functional Requirements - The system should analyze data, clean datasets, create data pipelines, and generate insights.

## External Interface Requirements - Use AWS S3 for storing data, Redshift for managing databases, and Databricks for processing data.

## System Features - The system must support data cleaning, aggregation, and analysis

## Nonfunctional Requirements - The system should perform well and scale efficiently to handle large datasets.