Requirements Specifications Document

# Introduction

This SRS is to dive deep into the behavioral analysis of our current customers and answer questions related to enhance revenue of our company.

## Purpose

The purpose of these requirements is to understand the customers and drive down the customized offers for them to potentially increase the revenue cycle.

## Intended Audience and Use

This SRS will be referenced by fellow DEs, Business Analysts and Data scientists.

## Product Scope

The scope of deploying this data pipeline would help in creating business strategies that would augment the revenue for meeting the business goals.

## Definitions and Acronyms

The following Acronyms will be used often throughout the documentation.

SRS: Software Requirements Specification

AWS: Amazon Web Services

S3: Simple Storage Service

EMR: Elastic MapReduce

Other than this I would use the following tools and it is defined below:

Redshift – a data warehouse in AWS environment

Databricks – a platform where I would perform data processing and I am using Pyspark to perform data processing.

# Overall Description

1. User Needs

This data pipeline will be used by various stakeholders within our Health Care insurance company, including software developers to understand and build new features or tools to boost the customer retention, fellow Data Engineers as a reference or someone who wants to work on the given data as core table to answer other set of use cases, and Data Scientists to predict and forecast the customer behavior over the years with the tables and pipelines created.

1. Assumptions

The assumption that I make during this project are technological assumptions, that is assuming that the pipeline I create and deploy are compatible with AWS and Databricks environments.

# System Features and Requirements

## Functional Requirements

The functional requirements include that the system should involve data cleaning modules, data upload modules to load into Redshift for the following datasets: Patients, Subscriber, Claims, Group\_subgroup.The system must include result generation modules for each use case output.

## External Interface Requirements

### User Interface:

The data pipeline will interact with users through PyCharm and Databricks for implementation purposes.

### Hardware Interface:

The project relies on AWS infrastructure, including S3 for data storage, Redshift for data warehousing, and EMR/Databricks for data processing.

### Software Interface:

The system will interface with PySpark, Spark-SQL for data processing and analysis. The pipeline will be pushed to GitHub and stored in code repository.

## System Features

The system features must integrate well with AWS S3, Redshift and EMR. Since the tasks are going to be tracked via JIRA, the system must integrate well for tracking the progress and identify any blockers that’s coming in the way.

## Nonfunctional Requirements

### Performance Requirements:

The whole project and answering use cases should be completed within a reasonable time frames and work as intended.

### Safety Requirements:

Data safety measures and privacy must be maintained throughout the data cleaning, exploration and implementation stages by sticking to company’s norms and policies.

### Security Requirements:

Access controls and authentication like AWS IAM role and Access keys would be in place to ensure data is secured and connections are encrypted when connecting to Databricks or similar processing tools.

### Scalability Requirements:

The system that’s existing should be able to scale with an increase in data volume, ensuring continued performance capability.