Requirement Specification Document

**Introduction**

**1.Objectives**:

The main objective of this project is to incorporate the big data ecosystem into the company’s operations to help the company grow the revenue and understand its customers. In addition, the project seeks to adeptly analyze the data pertaining to competitor companies, received from diverse sources and formats. This analysis will help track behaviors and condition of customers there by enabling the precise customized offers for target customer. This approach aims to simulate the acquisition of insurance policies and accurately calculate royalties to those customers who have previously brought policies and subsequently help to enhance the revenue.

**2.Goal**  
The main goal of this project is to help the company to create a robust and scalable data pipeline that will help to extract all the raw data that can be helpful for the analysis of various historical events and capture some insightful information to help for strategic business decision.

**3.Project Scope**

As this project deals with vast amounts of data related to diseases, various consumer groups and subgroups, insurance claims, hospitals, patients the main scope of the project would be to analyze how this data is affecting on the main challenges faced by the company such as with growth of the revenue and understanding the customers, creating a pipeline that will help to extract, transform and load the raw data in such a way that it can provide various important insight relevant to the problem and help the company with some helpful business decision. This will allow us to utilize a wide area of tools and technologies.

4.Definition and Acronyms

**Overall Description**

The result of this project will help in the strategic business decision-making process.

* 1. User Needs:

This project will help various endpoint user such as data scientist and data analyst to further create a predictive model, Marketing teams to build on marketing strategies for customer retention, policy managers to help create and edit the policy related to data decision. It will help by providing insights on various historical information related to data and by providing quality data to the end users.

* 1. Assumption and Dependencies:
* The success of the project hinges on the availability and accessibility of relevant data from both internal and external sources.
* It is assumed that the data obtained will be of sufficient quality and consistency for meaningful analysis.
* The project depends on the accurate collection and reliability of competitor data from various sources and formats.
* The assumption is that relevant teams will receive training to effectively utilize the data pipeline and interpret analytical insights.

**System Features and Requirements**

**1.Environment**

● AWS S3

● AWS Redshift

● Databricks

● AWS EMR Studio

● Pyspark

● Jira

● GitHub

**2.Functional Requirements**

● Which disease has a maximum number of claims.

● Find those Subscribers having age less than 30 and they subscribe any subgroup

● Find out which group has maximum subgroups.

● Find out hospital which serve most number of patients

● Find out which subgroups subscribe most number of times

● Find out total number of claims which were rejected

● From where most claims are coming (city)

● Which groups of policies subscriber subscribe mostly Government or private

● Average monthly premium subscriber pay to insurance company.

● Find out Which group is most profitable

● List all the patients below age of 18 who admit for cancer

● List patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.

● List female patients over the age of 40 that have undergone knee surgery in the past year

**3.System Features**