Requirements Specifications Document

# Introduction – *The document specifies the data pipeline, as the product, to be used to enhance revenue and understanding customer behavior. We will be preparing the data for the data analysts/scientists of the organization to enable better decision making.*

## Purpose – *The purpose of the product is to prepare a ready-to-analyze dataset(s) to track the customer behavior, condition of customer so that the Health Care Insurance (HCI) company can customize offers for them to buy insurance policies and also calculate royalties to the existing customers or the customer who have purchased policies in the past.*

## Intended Audience and Use - *The data scientist(s), data analysts, project manager. Developers, product owner and related subject matter experts (SMEs)*

## Product Scope -

## *Benefit- The benefits of the product is to reduce ETL processing time of the data and enhance cost effectiveness,*

## *Objectives are to extract the data from local file system and store it in a Datalake (AWS S3). Transform, clean and validate, the data from the datalake, using Databricks, and load into Datawarehouse (AWS Redshift) for further analysis*

## *Goals intended is to maximum query performance and optimize other processing task related to ETL and present a validated and cleaned dataset for analysis. This should relate to overall business goals, especially if teams outside of development will have access to the SRS.*

## Definitions and Acronyms - *ETL 🡪 Extract, Transform and Load.*

# Overall Description - *Your next step is to give a description of what you’re going to build. Why is this product needed? 🡪 To ensure low latency, scalable and secure data pipeline.*

# *Who is it for? 🡪 For the client & product owner to maximize revenue*

# *Is it a new product? 🡪 Yes*

# *Is it an add-on to a product you’ve already created? 🡪 No*

# *Is this going to integrate with another product? 🡪 Sometime in the future maybe*

# *Understanding and getting your team aligned on the answers to these questions on the front end makes creating the product much easier and more efficient for everyone involved.*

## User Needs –

## *Describe who will use the product and how 🡪 Data Analyst/Scientist and might be released as a ETL tool.*

## *Understanding the various users of the product and their needs is a critical part of the SRS writing process.*

## Assumptions and Dependencies –

## *What are we assuming will be true? Data availability and requirements are fixed. We will also be fixed on the platforms used and the use cases.*

## *Understating and laying out these assumptions ahead of time will help with headaches later.*

## *Are we assuming current technology? 🡪 Yes mostly related to Apache frameworks and AWS cloud technologies.*

## *Are we basing this on a Windows framework? 🡪 No, Ubuntu Linux framework.*

## *We need to take stock of these technical assumptions to better understand where our product might fail or not operate perfectly.*

# System Features and Requirements -*In order for your development team to meet the requirements properly, we must include as much detail as possible. This can feel overwhelming but becomes easier as you break down your requirements into categories.*

## Functional Requirements – *Cloud services for deployment from any region (AWS S3, Redshift). Scalability, and security of the codes. The following datasets need to be analyzed, cleaned and validated:*

Claim Analysis

Subscriber Analysis

Group Analysis

Hospital Analysis

Subgroup Subscription Analysis

Claims Rejection Analysis

City-wise Claims Analysis

Policy Type Analysis

Premium Analysis

Profitable Group Analysis

Patient Demographics Analysis

High-Value Insurance Coverage Analysis

## External Interface Requirements -

### User 🡪 Product Owner, Analysts.

### Hardware 🡪 CPU

### Software 🡪 Distributed File System, S3 for Storage, Query, Python 3, Pyspark, Databricks/AWS EMR, GitHub (for Version Control)

### Communications 🡪 Jira and Teams.

## System Features – Databricks 12.2 LTS(Apache Spark 3.3.2, Scala 2.12)

## Nonfunctional Requirements -

### Performance requirements 🡪 Optimized query performance, low latency network, Optimized processing time, automated scripts,

### Safety requirements

### Security requirements

### Usability requirements

### Scalability requirements

Code needs to be secured and optimized.

5mins vs 2mins

Function needs to be optimal. Function needs to be able to work in a distributed system. 🡪 non-functional requirements.

Link: Link: <https://github.com/ShouravDey96/Health_Insurance_data_pipeline>