Requirements Specifications Document

# Introduction – *Data pipeline for healthcare company to help optimize product offering.*

## Purpose –

## *The purpose of the project is to create data pipelines for the Health Care insurance company which will make the company make appropriate business strategies to enhance their revenue by analysing customers behaviours and send offers and royalties to customers respectively*

## Intended Audience and Use –

## *The intended audiences/team with direct impact are Sales, product, corporate treasure as well as customer services as they will likely see uptrend in new customer creation and monetary generation. Other teams may be indirectly impacted based on how closely they work with above mentioned audiences/team.*

## Product Scope –

## *The objective is to attract more customers through better product offering and there will be direct monetary benefits if implemented strategically and methodically. This document will be accessed by anyone that are directly impacted by it and those that management will decide to give access to.*

## Definitions and Acronyms –

## *here are* *few acronyms we will be using:*

* + 1. *ETL-stands for extraction, transformation, load. It’s used in data context to load data from one platform to another after making necessary changes to the initial dataset.*
    2. *EMR- Elastic map reduce. Its container within AWS where ETL will be created.*
    3. *AWS-Amazon Web Services-cloud platform by amazon.*

# Overall Description –

# *This new product/ETL will be very helpful in quickly analyzing the large volume of dataset related to customer trends from competitors and help us customize our product accordingly. It will be a value-add for product, sales and customer services as it will drive up their volume.*

## User Needs –

## *The users will be anyone in Product, Sales, Customer Services and below are the different use cases that will be answered by the project:*

1. Disease that has a maximum number of claims.
2. Subscribers having age less than 30 and they subscribe any subgroup
3. Group has maximum subgroups.
4. Hospital which serve most number of patients
5. Subgroups subscribe most number of times
6. Total number of claims which were rejected
7. From where most claims are coming (city)
8. Groups of policies subscriber subscribe mostly Government or private
9. Average monthly premium subscriber pay to insurance company.
10. Group that is most profitable
11. all the patients below age of 18 who admit for cancer.
12. patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.
13. female patients over the age of 40 that have undergone knee surgery in the past year.

## Assumptions and Dependencies – *we are assuming all of these data to be depended to each other.*

## *.*

# 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 –

## *below are functional requirements for this project:*

1. Answer which disease has a maximum number of claims.
2. Find those Subscribers having age less than 30 and they subscribe any subgroup
3. Find out which group has maximum subgroups.
4. Find out hospital which serve most number of patients
5. Find out which subgroups subscribe most number of times
6. Find out total number of claims which were rejected
7. From where most claims are coming (city)
8. Answer which groups of policies subscriber subscribe mostly Government or private
9. Average monthly premium subscriber pay to insurance company.
10. Find out Which group is most profitable
11. List all the patients below age of 18 who admit for cancer
12. List patients who have cashless insurance and have total charges greater than or equal for Rs. 50,000.
13. List female patients over the age of 40 that have undergone knee surgery in the past year

## External Interface Requirements:

### User: Users will be able to see the result in tabular form and have capability to download the data in .csv for further visualization

### Hardware: hardware used will be physical devices I will use to create the pipeline

### Software – AWS S3, EMR, Jupyter Notebook

### Communications: communications will be done via emails, phone call and teams video conferencing system

## System Features - *System should have basic features so, users can see final output. The development team will have advanced system features such as access to necessary software and devices with larger ram and storage for smooth data processing.*

## Nonfunctional Requirements -*:*

### Performance requirements: Final ETL pipeline should be optimal and efficient so that relevant data is available to users.

### Security requirements: There are multiple safety protocol followed that comes with the software we are using as well as ones that our company enforces.

### Usability requirements: The pipeline will be usable by anyone with proper access, and they can further analyze the data to suit their needs.

### Scalability requirements: Pipeline will be scalable and will be able to handle large amounts of data in future without compromising the efficiency and throughput.

## 