Template - Requirements Specifications Document

# Introduction –

To be competitive in the industry and to increase the revenue stream, understanding customer behavior is critical. Now to understand the customer data, what can give us more leverage than the integration of big data tools and the data collected from sources such as competitors, market trends, and third-party companies. By analyzing these data, we can customize offers for new users, calculate royalties for current users and retain the customers. This project targets in enhancing the company’s revenue and increase customer engagement by leveraging the power of big data, compute, and analysis.

## Purpose –

## *The goal of the Project is to design and implement data pipelines for the Health Care Insurance Company which will make the company make appropriate business decisions. There will be in-depth study of customer behavior and market trends and based on these studies decisions are made. Here, we mainly focus on 4 points*

* + 1. Study customer behavior
    2. Analyze market trends based on the customers
    3. Calculate royalties for current customers
    4. Increase revenue and customer retention.

## Intended Audience and Use – *There is diverse range of intended audience and users*

* + 1. Stakeholders: The stakeholders and project managers will use this project and the documentation to keep track of the project, align the project with the company’s goal and to monitor progress
    2. Business Analysts: The analysts align the insights obtained with strategic decisions to start new initiatives.

## Product Scope - *Benefits: increase the revenues and attract more customers by providing them customized offers, Objective: We should analyze the data such that we can provide answers for the requirement that stakeholders can make business decisions, Goals: Increase Revenues by acquiring new customers, Increase the retention rate of the existing customers*

The scope of the project consists of achieving several decisions based the insights obtained from the areas below

1. The diseases with maximum number of claims => (change the policy and pricing strategy for diseases with maximum number of claims)
2. Categorizing subscribers based on age and disease => Focusing on specific policies and pricings for specific set of people to improve acquiring new customers and maintaining customer retention.
3. Hospital with maximum number of patients => Focusing in maintaining better relation with the hospital that server more number of patient to identify patient’s needs that helps make business decisions

## Definitions and Acronyms –

1. Subscriber: the user that has held the insurance policy

# Overall Description *The project aims to develop a data analytics pipeline solution that will process and analyze large volume of health care data from various sources, including competitor information and customer records. The system will enable the company to gain deeper insights into customer behavior, optimize policy offering, and enhance revenue through data-driven decision-making. By implementing this solution, the company expects to improve customer satisfaction, increase policy sales, and streamline its operations in response to market trends and customer needs.*

product will be viewed by business stakeholders to make data-driven decisions

## User Needs – The health care insurance company requires

1. Improved understanding of customer behavior and health conditions
2. Ability to have insights to customize insurance policy
3. Insights into competitor strategies and market trends
4. Enhanced revenue through data driven decisions.

## Assumptions and Dependencies –

1. Data is legally obtained and is available
2. AWS infrastructure is used in the company
3. Business analytics is compatible with the tools being used
4. Timely approval and allocation of resources for the projects

# System Features and Requirements -*.*

## Functional Requirements –

* + 1. 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. 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

1. Stakeholders and analysts have access to AWS Redshift
2. Role based access and IAM roles has to be assigned

### Hardware

### AWS EC2 instances for running EMR clusters

1. AWS S3 for data storage
2. AWS Redshift for data warehousing

### Software

1. Databricks => data processing and analytics
2. PySpark => Data manipulation and transformation
3. AWS EMR => Development env for data manipulation and transformation
4. Jira => Project Management
5. GitHub => Version control and collaboration

### Communications

1. Jira for User stories and tasks
2. Github for collaboration and version control
3. API integrations for data exchange

## System Features

These features are required for the automation and optimization of the pipeline such that it is able to produce accurate and insightful results

1. Data Pipelines for ingesting and processing healthcare data
2. Automated reporting and alert systems
3. Redshift access for analysis

## Nonfunctional Requirements -

### Performance requirements

* Cuncurrent access for 10 users
* Processing should handle 10 GB of the data per day

### Safety requirements

* Regular backup and disaster recovery

### Security requirements

* MFA for system access
* Regular security audits

### Scalability requirements

* Ability to scale processes based on data volume
* Support increasing number of users
* Flexibility to add new data sources with the use of SCD

## 