Specifications Document for Capstone Project

# Introduction - *A health insurance company is having trouble increasing its revenue and better understanding its clientele. As a result, it is looking to the Big Data Ecosystem for assistance in analyzing competitor firm data that it has obtained from a variety of sources, including third-party sources, and scraping. Through this analysis, the company will be able to monitor customer behavior and conditions and tailor offers to encourage clients to purchase insurance policies. The insurance company will research on the data and will forecast the possible solution through the help of data analyst.*

## Purpose – *The project aims to build data pipelines for the health care insurance firm so that it may analyze customer behavior, deliver offers and royalties to clients, and make suitable business decisions to increase income.*

## Intended Audience and Use – *The audience for the project are developers, tester, project manage and data analyst.*

## Product Scope - A Health Care insurance company

# Overall Description –

# *The project will stats as follows,*

* + - 1. *Dataset creation*
      2. *Data cleaning using pyspark in Databriks*
      3. *Table creation using AWS Redshift*
      4. *Mapping multiple relational tables into dimensional tables using schemas design doc*
      5. *Creation of final dimensional dataset using Databriks*
      6. *Publishing the dimensional tables into AWS Redshift*
      7. *Uploading all the documents related to the project into GitHub*
      8. *Prepare a sprint for the project*

# System Features and Requirements -*.*

## Functional Requirements – *The functional requirements are:*

## *find out the disease has a maximum number of claims, those subscribers having age less than 30 and they subscribe any subgroup, which group has maximum subgroups, hospital which serve most number of patients, which subgroups subscribe most number of times, 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, 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.*

## External Interface Requirements -

### Software

### AWS S3, AWS Redshift, Databricks, AWS EMR Studio, Pyspark, Jira, GitHub.

## Nonfunctional Requirements -

### Security requirements

*Security requirements are IAM role, IAM Access and Security keys.*