## **Requirement**

A Health Care insurance company is facing challenges in enhancing its revenue and understanding the customers so it wants to take help from Big Data Ecosystem to analyse the Competitors company data received from varieties of sources, namely through scrapping and third-party sources. This analysis will help them to track the behaviour, condition of customers so that they can customise offers for them to buy insurance policies and also calculate royalties to those customers who buy policies in the past, this in turn will enhance their revenues.

**List Of 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

## **The goal of the project**

The goal 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.

## **Environment**

* AWS S3
* AWS Redshift
* Databricks
* AWS EMR Studio
* Pyspark
* Jira
* GitHub

### **DATASET CREATION**

* + Upload the given sample data on AWS s3 in a folder named input-data.

### **DATA CLEANING**

* + Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.
  + Cleaning Activity
    - First check if there are null values in dataset
    - Count the total Null values for each column
    - And then replace the null values for specific columns by NA
    - Check the If three are duplicates records
    - If there are duplicates then drop duplicates
  + Clean data for at least for following datasets
    - Patients
    - Subscriber
    - Claims
    - Group\_subgroup
  + Upload cleaned data corresponding to each data set into a redshift table.
  + Please create a schema design doc for target tables.

### **RESULT CREATION ON REDSHIFT**

### Create a separate redshift table for each use case output in a redshift schema

* + - Schema = Project-Output

***Note****-*

* *For implementation and testing purposes you can use Databricks community edition.*
* *When your code is ready to use you can push this code to the github repo with the help of pycharm.*
* *At the end this code will be deployed on AWS EMR or Databricks with the help of Github.*

**Submission - You are supposed to submit followings in github repos**

* **Documents Repo**
  + Requirement specification documents
  + Solution Design Documents
  + Create documents for followings
    - Create a 2 week sprint for this project in jira.
    - First week for documentations and solution design.
    - Second week for implementation and testing.
    - List of jira user stories and tasks which you have created for individual use cases in a scrum type project.
    - List of jira user stories and tasks created for test cases Or a single test cases xls sheet.
  + A database schema design doc for final redshift tables with primary and foreign key combinations.
* **Pyspark Source Code Repo - This should include following**
  + Data cleansing modules
  + Result generations modules
* You are supposed to present your solution to an available audience with a **live class presentation** with use cases explanation and solution design with required screen-shots.
  + Recommendation
    - *You can include databricks visualisations for results snapshots.*