Solution Design Document

# Solution

The solution will be approached as step by step as follows:

Step1: The datasets will be created from the given raw data and loaded into AWS S3 in a folder named as Input\_data for reference.

Step2: The loaded data will be cleansed and enriched for the following datasets namely Patient\_records, group, grpsubgroup, hospital, Subscriber, Claims. The cleaning of data will include removing null values, counting the null values, replacing the null values for specific columns with NA, dropping duplicates.

Step3: After the data is cleaned, a schema is created with the Entity- Relationship diagram.

Step4: Once the schema is created, structure of the data including Metadata, Datatypes, PK and FK are defined in this step.

Step5: This step we implement our project by shuffling the data using Join, Group by clause to solve the usecases. This might require me to join 2 or more tables as per requirements.

Step6: Separate Redshift tables will be created for each use case and the output table will be stored under the schema name project-output for future reference.

# Usecases

* Disease with the maximum number of claims.
* Subscribers below 30 who subscribe to any subgroup.
* Group with the maximum subgroups.
* Hospital serving the most number of patients.
* Subgroups subscribing the most number of times.
* Total number of rejected claims.
* City from which most claims originate.
* Policy groups subscribed to mostly (Government or private).
* Average monthly premium paid by subscribers.
* Most profitable group.
* List of patients below 18 admitted for cancer.
* List of cashless-insured female patients over 40 undergoing knee surgery in the past year.

# 3.**Database Design:**

## a. Tables Metadata Info with Pk/FK Relationship:

**Patients\_records Table:**

Columns: patient\_id (PK), patient\_name, patient\_gender, patient\_birth\_date, patient\_phone, disease\_name, city, hospital\_id (FK)

**Hospital Table:**

Columns: hospital\_id (PK), hospital\_name, city, state, country

**Disease Table:**

Columns: SubGrpID (FK), Disease\_ID (PK), Disease\_name

**Group Table:**

Columns: Grp\_Id (PK), Country, premium\_written, zipcode, Grp\_Name, Grp\_Type, city, year

**Group\_subGroup Table:**

Columns: SubGrp\_ID, Grp\_Id

**Subgroup Table:**

Columns: SubGrp\_id (PK), SubGrp\_Name, Monthly\_Premium

**Subscriber Table:**

Columns: sub\_id (PK), first\_name, last\_name, Street, Birth\_date, Gender, Phone, Country, City, Zip Code, Subgrp\_id (FK), Elig\_ind, eff\_date, term\_date

**Claims Table:**

Columns: claim\_id (PK), patient\_id (FK), disease\_name, SUB\_ID, Claim\_Or\_Rejected, claim\_type, claim\_amount, claim\_date

## b. ER Diagram

A screenshot of a computer

Description automatically generated

# 4. Technologies and Platforms to be used in this solution

* **AWS S3:** For storing the uncleaned data.
* **AWS Redshift:** For the data warehouse to store cleaned and processed tables or usecases.
* **Databricks:** For data processing using PySpark by coonecting with AWS S3.
* **PySpark:** For data cleaning and processing data.
* **Jira:** For project management and tracking the whole project.
* **GitHub:** For version control and pushing code.