# Solution - *Explain your solution here in a step by step manner.*

Phase 1: Data Collection

* Gather all the data.
* Then Create a new bucket and folder.
* Upload all the data into that folder.

Phase 2: Data Cleaning and Preparation

* After loading the data into S3. Use databricks to clean and prepare the data.
* Connect S3 with databricks notebook.
* Create a data frame and read the S3 data from there.
* Check for any null values in the tables and replace with ‘NA’.
* Check for duplicate values.
* Drop duplicate values, if any.
* Now, the data is all cleaned.

Phase 3: Data Processing and Analysis

* Connect databrick notebook with AWS Redshift.
* Write the clean data into the redshift table.
* Use the Redshift query editor for data processing.

Phase 4: Result Storage and Visualization

* Create a different schema and tables for uploading output.
* Load all the query output in the tables

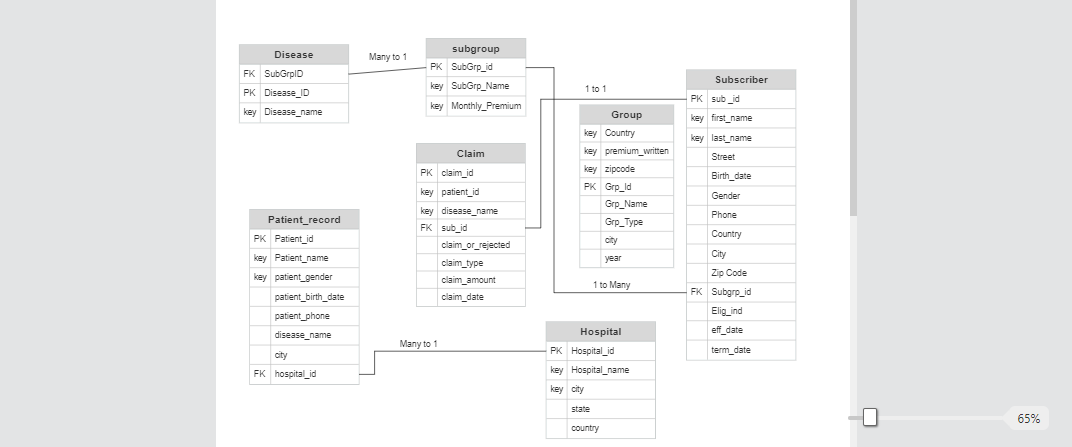
# Use Cases - *List down all the use cases on which this solution will be applicable.*

* 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

1. Database Design - List down all possible db(Redshift) tables here

## Tables Metadata Info with Pk/FK relationship -

## ER diagram – *Optional*



# Technologies and Platforms to be used in this solution -*List down list of technologies like spark, aws and databricks etc.*

* 1. AWS S3
  2. AWS Redshift
  3. Databricks
  4. AWS EMR
  5. Github
  6. Pyspark