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

* Upload all the required data files on AWS s3 named “input-data” from the local device.
* Load the Data files from S3 to EMR and perform data cleaning according to the requirements:
  1. Check if there are null values in the dataset.
  2. Count the total Null values for specific columns by NA.
  3. And replace the null values for specific columns by NA.
  4. Check if there are duplicate records, if yes, drop them.
* Load the cleaned data corresponding to each redshift table.
* Create a separate redshift table for each use case output in a redshift Shema as project -output.

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

* 1. Which disease has a maximum number of claims?
  2. Find out which group has the maximum subgroups.
  3. Find out the hospital which serves the most number of patients
  4. Find out which subgroups subscribe the most number of times
  5. Find out the total number of claims which were rejected
  6. From where most claims are coming (city)
  7. Which groups of policies do subscribers subscribe mostly Government or private
  8. The average monthly premium subscriber pays to the insurance company.
  9. Find out Which group is most profitable
  10. List all the patients below the age of 18 who admit for cancer
  11. List patients who have cashless insurance and have total charges greater than or equal to Rs. 50,000.
  12. 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 –

## Disease, Metadata- Disease\_ID(PK), SubGrpID (FK), Disease\_name.

* + - 1. Group, Metadata- Country, premium\_written, zipcode, GrpID (PK), Grp\_type, City, Year.
      2. Hospital, Metadata- Hospital\_id (PK), Hospital\_name, City, state, country.
      3. Patient\_records, Metadata- Patient\_id (PK), patient\_name, Patient\_gender, Patient\_birth\_date, Patient\_phone, disease\_name (FK), city , hospital\_id (FK).
      4. Subgroup, Metadata- SubGrp\_id (PK), SubGrp\_name, Monthly\_premium.
      5. Subscriber. Sub\_id (PK), first\_name, last\_name, Street, Birth\_date, Gender, Phone, Country, City, Zipcode, SubGrp\_id (FK), Elig\_ind, eff\_date, term\_date

## ER diagram - *Optional*

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

* PySpark
* SQL
* AWS S3, EMR, Redshift
* Databricks
* Jira
* GitHub