# Solution

* Data Upload- Created Folder to upload data files in AWS s3
* Data Ingestion- Used Databrick to read the files from AWS S3.
* Data Cleaning- Cleaned the data using Databrick by sorting null values and dropDuplicates.
* Data Loading- Load the data into Redshift and create tables to view.

# Use Cases

* 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:

1. Group\_Subgroup

Group\_id

Group\_name

Subgroup\_id

Subgroup\_name

1. Claims

Claim\_id

Subscriber\_id

Disease

Claim\_amount

Claim\_status

Claim\_date

city

1. Patients

Patient\_id

Name

Age

Disease

Hospital

Admission\_date

Discharge\_date

1. Subscribes

Subscriber\_id

Name

Age

Gender

Policy\_id

Subscription\_date

Premium\_amount

Cashies\_insurance

# Technologies and Platforms to be used in this solution

AWS S3

Databricks

AWS Redshift

PySpark

Python

Github

Jira