# Solution – *Our solution will be to first upload all the data onto AWS S3 as storage source, from here we will extract the data onto databricks where we plan to perform transformations that include: checking and counting the amount of missing data, then replacing or deleting it, checking for duplicate data and removing it. Afterwards we will upload our cleaned data onto AWS Redshift. All of these tasks will be tracked on Jira.*

# Use Cases –

This project will allow our BI analysts to understand more of our customers:

* 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 –

Technology we will be using include :

AWS S3

AWS Redshift

Databricks

Jira

Pyspark

Github

## 