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

*We will be creating a ETL Pipeline.*

1. We will first ingest the data from different CSV files to AWS S3
2. For each file we will do data cleaning which involves
   1. Checking for null values
   2. Count the total Null Values for each columns
   3. Replace the Null values by NA
   4. Deduplication
   5. Check the labels
   6. Join different tables to get the corresponding datasets
3. After the cleaning, we will upload each data set into redshift table
4. We should create Schema design for target tables
5. We should create separate redshift table for each use case output in a redshift schema

This a completely new product developed to understand the customer behavior so that we can make a decision to increase company’s revenue by acquiring more customers.

The final usecase of the product will be viewed by business stakeholders to make data-driven decisions

The final use case will be the following:

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

* 1. The stakeholders can look into the following functionalities. The business stakeholders use these use cases to make different decisions
     1. Find those Subscribers having age less than 30 and they subscribe any subgroup
     2. Find out which group has maximum subgroups.
     3. Find out hospital which serve most number of patients
     4. Find out which subgroups subscribe most number of times
     5. Find out total number of claims which were rejected
     6. From where most claims are coming (city)
     7. Which groups of policies subscriber subscribe mostly Government or private
     8. Average monthly premium subscriber pay to insurance company.
     9. Find out Which group is most profitable
     10. List all the patients below age of 18 who admit for cancer
     11. List patients who have cashless insurance and have total charges greater than or equal for 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 – (create the diagram)

A diagram of a computer

Description automatically generated

## ER diagram - *Optional*

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

* Technologies we are going to use

## AWS S3: For data storage and as data lake

1. AWS Redshift: For data warehousing and analytics
2. AWS EMR: For big data processing
3. PySpark: For data processing and transformation within EMR
4. GitHub: For version control
5. Jiira: For project management and tracking
6. AWS IAM: For access Management and Security