1. Solution

- Upload the given sample data on AWS s3 in a folder named input-data
- Cleaning Activity
    i. First check if there are null values in dataset
    ii. Count the total Null values for each column
    iii. And then replace the null values for specific columns by NA
    iv. Check the If three are duplicates records
    v. If there are duplicates then drop duplicates

- Clean data for at least for following datasets
    i. Patients
    ii. Subscriber
    iii. Claims
    iv. Group_subgroup

- Upload cleaned data corresponding to each data set into a redshift table.
- Create a schema design doc for target tables.
- Create a separate redshift table for each use case output in a redshift schema

2. 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

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

- Tables Metadata Info with Pk/FK relationship -
    i. Claim
        1. Claim_Or_Rejected: string,
        2. SUB_ID: string,
        3. claim_amount: string,
        4. claim_date: string,

5. claim_id: bigint,
6. claim_type: string,
7. disease_name: string,
8. patient_id: bigint

ii. Disease
    1. SubGrpID: string,
    2. Disease_ID: int,
    3. Disease_name: string

iii. Group
    1. Country: string,
    2. premium_written: int,
    3. zipcode: int,
    4. Grp_Id: string,
    5. Grp_Name: string,
    6. Grp_Type: string,
    7. city: string,
    8. year: int

iv. SubGroup

    1. SubGrp_id: string,
    2. SubGrp_Name: string,
    3. Monthly_Premium: int

v. Hospital
    1. Hospital_id: string,
    2. Hospital_name: string,
    3. city: string,
    4. state: string,
    5. country: string

vi. Patient_Records
    1. Patient_id: int
    2. Patient_name: string
    3. patient_gender: string
    4. patient_birth_date: timestamp
    5. patient_phone: string
    6. disease_name: string
    7. city: string
    8. hospital_id: string

vii. Subscriber
    1. [sub _id: string
    2. first_name: string
    3. last_name: string
    4. Street: string
    5. Birth_date: timestamp

6. Gender: string
7. Phone: string
8. Country: string
9. City: string
10. Zip Code: int
11. Subgrp_id: string
12. Elig_ind: string
13. eff_date: timestamp
14. term_date: timestamp

viii. GroupSubgroup

1. SubGrp_ID: string,
2. Grp_Id: string

- ER diagram - *Optional*

4. Technologies and Platforms to be used in this solution -

- *Databrick*
- *AWS S3*
- *Redshift*
- *Pyspark*
- *SQL*