Solution - The solution involves creating data pipelines to process, clean, and analyze healthcare insurance data. This will be done step by step as follows:

* **Step 1:** Upload the given sample data to AWS S3.
* **Step 2:** Perform data cleaning tasks such as handling null values, removing duplicates, and formatting data correctly.
* **Step 3:** Load the cleaned data into AWS Redshift tables.
* **Step 4:** Create specific Redshift tables to store the results of various analyses.
* **Step 5:** Analyze the data to generate insights, such as identifying trends in claims, subscriber behavior, and the profitability of different groups.
* **Step 6:** Use AWS EMR and Databricks for processing and AWS Lambda for automating specific tasks.
* **Step 7:** Present the final results through visualizations and documentation.

Use Cases -  **Use Case 1:** Analyze which disease has the maximum number of claims.

 **Use Case 2:** Identify subscribers under 30 years old who subscribe to any subgroup.

 **Use Case 3:** Determine which group has the maximum subgroups.

 **Use Case 4:** Find out which hospital serves the most patients.

 **Use Case 5:** Analyze which subgroups are subscribed to the most.

 **Use Case 6:** Calculate the total number of claims rejected.

 **Use Case 7:** Identify the city from where most claims are coming.

 **Use Case 8:** Determine if subscribers prefer government or private policies.

 **Use Case 9:** Calculate the average monthly premium paid by subscribers.

 **Use Case 10:** Identify the most profitable group.

 **Use Case 11:** List patients below the age of 18 who were admitted for cancer.

 **Use Case 12:** Identify patients with cashless insurance and charges greater than or equal to Rs. 50,000.

1. Database Design – The database design includes the following Redshift tables:

* **Patients**: Stores patient demographic and medical information.
* **Subscribers**: Contains details about insurance subscribers.
* **Claims**: Records insurance claim information.
* **Group\_Subgroup**: Contains details on insurance policy groups and subgroups.

# Technologies and Platforms to be used in this solution

 **AWS S3**: For storing the initial and cleaned datasets.

 **AWS Redshift**: For creating and managing the database tables and running queries.

 **AWS EMR**: For processing large datasets using Spark and Hadoop.

 **Databricks**: For data processing and analysis using PySpark.

 **AWS Lambda**: For automating specific tasks within the data pipeline.

 **GitHub**: For version control and collaboration.