# Solution –

**Overview:** The objective is to design and implement data pipelines for a healthcare insurance company, enabling them to enhance their revenue by analyzing customer behavior, competitors’ data, and other relevant metrics. This will be achieved by ingesting data from various sources into AWS S3, cleaning the data using PySpark on Databricks, storing the cleaned data in AWS Redshift, and creating analytical outputs that the company can use to develop business strategies.

**Data Flow:**

1. **Data Ingestion:**
   * **Data is collected from various sources, including web scraping and third-party data providers.**
   * **The data is stored in an S3 bucket in a folder named input-data.**
2. **Data Cleaning:**
   * **Clean the data using PySpark. This includes handling null values, removing duplicates, and formatting data correctly.**
   * **The cleaned data is then uploaded into AWS Redshift tables.**
3. **Data Analysis:**
   * **Run SQL queries on AWS Redshift to generate the required insights.**
   * **Store the results of these queries in separate tables within the Redshift schema Project-Output.**
4. **Output & Visualization:**
   * **Use Databricks to visualize the results of the queries for reporting purposes.**
   * **The final code is pushed to a GitHub repository and will be deployed on AWS EMR or Databricks.**

# 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 –

 **Patients**

 **Subscribers**

 **Claims**

 **Groups\_Subgroups**

 **Hospitals**

 **Output Tables** (multiple tables for specific use cases)

## ER diagram - *Optional*

# Technologies and Platforms to be used in this solution –

* AWS S3
* AWS Redshift
* Databricks
* AWS EMR Studio
* Pyspark
* Jira
* GitHub

## 