Solution Design Document

# Solution

This section outlines the solution to achieve the project objectives. The solution involves the creation of data pipelines to process and analyze big data for the healthcare insurance company using AWS and Databricks.

# Use Cases

1. Identify salesmen and customers who reside in the same city and publish the results to an AWS Redshift table.  
2. Analyze customer behavior to predict future insurance product sales.  
3. Provide insights into customer data to increase revenue by targeting high-potential clients.

# Database Design

The solution will involve the following Redshift tables:  
- Customers: This table will store customer details including customer ID, name, address, city, etc.  
- Salesmen: This table will store salesman details including salesman ID, name, city, etc.  
- Transactions: This table will store transaction details including transaction ID, customer ID, salesman ID, and sales amount.

# Tables Metadata Info with PK/FK Relationship

The tables will have the following relationships:  
- Customers (PK: Customer\_ID)  
- Salesmen (PK: Salesman\_ID)  
- Transactions (PK: Transaction\_ID, FK: Customer\_ID, Salesman\_ID)  
These relationships will enable the analysis of sales data to determine the performance of salesmen in different cities.

# ER Diagram

An ER Diagram can be included here to visually represent the database schema.

# Technologies and Platforms

The following technologies and platforms will be used in this solution:  
- AWS S3 for data storage.  
- Databricks for data processing.  
- AWS Redshift for data warehousing.  
- PySpark for data transformation.  
- AWS EMR for managing the data processing environment.