**SOLUTION DESIGN DOCUMENT  
- Requirements Specifications Document**

**A. Solution**

- The solution involves developing a comprehensive data analysis system that utilizes Big Data technologies to enhance revenue and customer understanding in the Health Care insurance company. The solution encompasses the following steps:

Step 1: Data Collection and Integration

- Ingest data from various sources

- Integrate the collected data into a central repository for analysis.

Step 2: Data Cleaning and Preprocessing

- Detect and handle missing values, null values, duplicates, and inconsistencies in the data.

- Standardize and transform the data into a consistent format.

Step 3: Data Transformation and Enrichment

- Apply calculations, aggregations, and mappings to derive meaningful insights.

- Enrich the data with relevant information to support analysis.

Step 4: Data Storage and Management

- Store the processed data in AWS S3 and utilize AWS Redshift for data warehousing.

- Implement data partitioning and indexing for efficient querying.

Step 5: Exploratory Data Analysis (EDA)

- Utilize tools like Databricks and Pyspark for in-depth exploratory analysis.

- Create visualizations and reports to uncover patterns and trends in the data (Tableau, PowerBI or Excel)

Step 6: Querying and Reporting

- Provide querying capabilities to retrieve specific insights from the data.

- Generate reports and visualizations for different user roles.

**B. Use Cases**

- The solution will be applicable to various use cases, including:

1. Executives making informed business decisions based on revenue insights.

2. Marketing teams tailoring campaigns using customer behavior analysis.

3. Product development teams creating customized insurance policies.

4. Customer service representatives providing personalized assistance.

5. Data analysts deriving insights for strategic planning.

**C. Database Design**

a. Tables Metadata Info

Patients – patient\_id (pk), Claim – claim\_id (PK), Sub\_id (FK), Disease – subgrp\_id (FK), Hospitals – hospital\_id (pk), subscriber – sub\_id (pk), subgrp\_id(fk), subgroup – subgroup\_id(pk), group – group\_id (pk)

b. ER Diagram (Optional)

- A visual representation of the database schema.A diagram of a medical organization

Description automatically generated with medium confidence

D. Technologies and Platforms to be Used in This Solution

- The solution will leverage the following technologies and platforms:

- AWS S3 for data storage

- AWS Redshift for data warehousing

- Databricks for data analysis

- Pyspark for distributed data processing

- Amazon EMR for big data processing

- Jira and GitHub for collaboration and version control

- Draw.io – For ERD

- Tableau – Data visualization