**Solution Design Document**

**1. Solution**

The proposed solution involves the creation of data pipelines and analysis mechanisms using various technologies to enhance the revenue generation and customer understanding of the Health Care insurance company. The solution will follow these steps:

1. **Data Collection:** Competitor company data will be scraped from various sources and collected in the AWS S3 bucket named "input-data".
2. **Data Cleaning:** Data cleaning modules will be developed to handle null values, duplicates, and data inconsistencies.
3. **Data Pre-processing:** The cleaned data for each dataset (Patients, Subscriber, Claims, Group subgroup) will be uploaded into corresponding Redshift tables.
4. **Data Analysis:** Data analysis modules will be implemented using Pyspark and Databricks to perform various analyses on the datasets.
5. **Result Generation and Visualization:** Results of analyses will be stored in separate Redshift tables within the "Project-Output" schema. Databricks will be used to visualize the results snapshots.
6. **Data Deployment:** The code will be deployed on AWS EMR with the help of GitHub.

**Use Case Customization:**

The solution will address various use cases by generating insights based on customer behavior, policy preferences, and claim trends.

**Business Strategies:**

The insights generated will help the insurance company customize offers and calculate royalties for customers, thereby enhancing revenue.

**2. Use Cases**

The solution will be applicable to the following use cases:

* Identify disease with maximum claims.
* Find subscribers below 30 years subscribing to any subgroup.
* Determine the group with the most subgroups.
* Identify the hospital serving the most patients.
* Determine the subgroups with the highest subscription frequency.
* Count total number of rejected claims.
* Identify the city with the highest claim count.
* Determine whether subscribers prefer government or private policies.
* Calculate average monthly premium paid by subscribers.
* Identify the most profitable group.
* List patients under 18 admitted for cancer.
* List patients with cashless insurance and charges ≥ Rs. 50,000.
* List female patients over 40 who had knee surgery in the past year.

**3. Database Design**

**Tables Metadata Info:**

* **Patients:** Patient information including patient ID (PK), patient\_name, patient\_gender, patient\_birth\_date, patient\_phone, disease\_name, city, hospital\_id (FK).
* **Disease:** Details of diseases with Subgrp\_ID (FK), ???, and disease\_name.
* **Claims:** Claims information including claim ID (PK), patient ID (FK), disease\_name, SUB\_ID (FK), claim\_Or\_Rejected, claim\_type, claim\_amount and claim\_date.
* **Group**: Information about policy groups and subgroups with Country, ???, ??? group\_ID(PK), policy\_name, policy\_type, City, policy\_date.
* **Hospitals:** Hospital information including Hospital\_id (PK), Hospital\_name, city, state, country.
* **Subgroup:** Information regarding patients including Subgrp\_ID (PK), disease???, ???, group\_ID(FK)
* **Subscriber**: Subscriber information including sub\_ID (PK), first\_name, last\_name, Street, Birth\_date, Gender, Phone, Country, City, Zip Code, Subgrp\_id (FK), Elig\_ind, eff\_date, term\_date

**ER Diagram: (Optional)**

A black screen with white text

Description automatically generated

**4. Technologies and Platforms**

The solution will utilize the following technologies and platforms:

AWS Services: S3 for data storage, Redshift for data warehousing, EMR for data processing.

Databricks: Unified analytics platform for data processing, analysis and data visualization.

Pyspark: Python-based framework for data manipulation and analysis.

Jira: Project management and issue tracking.

GitHub: Version control and code repository.

Gmail: To communicate to team members and managers.

Microsoft Teams : To have meetings about the project.

Draw.io : To create the ER Class Diagram