Solution Design Document

1. Solution

The solution involves creating robust data pipelines using the Big Data Ecosystem to analyze data from various sources. Here is a step-by-step explanation:

1. **Data Ingestion**:
    - Data will be ingested from multiple sources such as web scraping, third-party APIs, and internal databases.
    - AWS S3 will be used for initial data storage.
2. **Data Processing**:
    - Databricks and PySpark will be utilized for data processing.
    - Data will be cleansed, transformed, and loaded into AWS Redshift.
3. **Data Analysis**:
    - Various analyses will be performed on the processed data to derive insights about customer behaviors, claims, subscriptions, and other relevant metrics.
4. **Data Storage**:
    - Processed data will be stored in AWS Redshift.
    - Appropriate table structures with primary and foreign keys will be created for efficient querying.
5. **Deployment**:
    - The entire pipeline will be automated and scheduled using AWS EMR and Databricks.
    - Code and configuration will be managed and deployed via GitHub.

2. Use Cases

The solution will be applicable to the following use cases:

- **Disease Analysis**: The solution will leverage data pipelines to analyze disease patterns among policyholders, identifying prevalent conditions, treatment outcomes, and associated costs. Insights from this analysis will inform proactive healthcare interventions and resource allocation strategies.
- **Subscriber Analysis**: By integrating demographic, behavioral, and health data, the solution will enable detailed subscriber segmentation. This will facilitate personalized service offerings, targeted marketing campaigns, and enhanced customer satisfaction.
- **Group Analysis**: Utilizing data analytics, the solution will analyze group policyholder demographics, claims histories, and healthcare utilization patterns. This analysis will

optimize group insurance offerings and tailor benefit packages to meet diverse group needs effectively.

- **Hospital Analysis**: Data pipelines will facilitate comprehensive hospital performance analysis, including service quality metrics, cost efficiency, and patient outcomes. Insights will aid in optimizing provider networks and enhancing healthcare service delivery.
- **Subgroup Analysis**: The solution will enable detailed analysis of smaller policyholder subgroups based on specific criteria such as age, location, or health conditions. This will support targeted healthcare interventions and personalized member engagement strategies.
- **Claims Rejection Analysis**: Through advanced analytics, the solution will identify patterns contributing to claims rejections. Insights will streamline claims processing workflows, reduce rejections, and improve operational efficiency.
- **Claims Origin Analysis**: By analyzing the origin and types of claims filed, the solution will provide insights into healthcare utilization trends and cost drivers. This analysis will inform strategic decision-making and policy design.
- **Policy Type Analysis**: Data-driven analysis will evaluate the performance and profitability of different policy types. Insights will guide product development, pricing strategies, and customer retention initiatives.
- **Premium Analysis**: The solution will analyze premium payment trends, payment behaviors, and factors influencing premium adjustments. This analysis will optimize pricing strategies and financial forecasting.
- **Profitability Analysis**: Leveraging financial and operational data, the solution will conduct profitability analysis across various segments and policies. Insights will guide resource allocation and business strategy formulation.
- **Pediatric Cancer Patients**: Dedicated analysis will focus on pediatric cancer patient demographics, treatment effectiveness, and healthcare utilization patterns. This will support specialized care management and treatment planning.
- **High-Value Cashless Insurance Patients**: The solution will analyze high-value cashless insurance claims to optimize reimbursement processes, reduce administrative costs, and enhance customer experience.
- **Female Knee Surgery Patients**: Tailored analysis will focus on female knee surgery patients, evaluating treatment outcomes, recovery times, and patient satisfaction. Insights will inform personalized care protocols and surgical interventions.

3. Database Design

This part outlines the schema design for the healthcare insurance company's data analysis project. It includes details about each table, their columns, data types, nullable properties, and the primary and foreign key relationships.

1. Subscriber Table: Contains information about the subscribers.

Column Name	Data Type	Constraints	Description
sub_id	**STRING**	**PRIMARY KEY**	**Subscriber ID**
first_name	**STRING**	**NOT NULL**	**Subscriber's first name**
last_name	**STRING**	**NOT NULL**	**Subscriber's last name**
street	**STRING**		**Street address**
birth_date	**DATE**		**Subscriber's birth date**
gender	**STRING**		**Subscriber's gender**
phone	**STRING**		**Subscriber's phone number**
country	**STRING**		**Subscriber's country**
city	**STRING**		**Subscriber's city**
zip_code	**INTEGER**		**Subscriber's zip code**
subgrp_id	**STRING**	**FOREIGN KEY**	**Foreign key to subgroup.subgrp_id**
elig_ind	**STRING**		**Eligibility indicator**
eff_date	**DATE**		**Effective date**
term_date	**DATE**		**Termination date**

2. **Subgroup Table:** Contains information about subgroups of subscribers.

Column Name	Data Type	Constraints	Description
subgrp_id	**STRING**	**PRIMARY KEY**	**Subgroup ID**
subgrp_name	**STRING**	**NOT NULL**	**Subgroup name**

monthly_premium	**INTEGER**		**Monthly premium amount**

3**. Group Table:** Contains information about the groups.

Column Name	Data Type	Constraints	Description
grp_id	**STRING**	**PRIMARY KEY**	**Group ID**
grp_name	**STRING**	**NOT NULL**	**Group name**
grp_type	**STRING**		**Group type**
country	**STRING**		**Country**
city	**STRING**		**City**
zipcode	**INTEGER**		**Zipcode**
premium_written	**INTEGER**		**Premium written**
year	**INTEGER**		**Year**

4. **GroupSubgroup Table**: Contains mappings between groups and subgroups.

Column Name	Data Type	Constraints	Description
subgrp_id	**STRING**	**PRIMARY KEY**	**Foreign key to subgroup.subgrp_id**
grp_id	**STRING**	**PRIMARY KEY**	**Foreign key to group.grp_id**

5. **Disease Table: Contains information about diseases.**

Column Name	Data Type	Constraints	Description
disease_id	INTEGER	PRIMARY KEY	Disease ID
disease_name	STRING		Disease name
subgrp_id	STRING	FOREIGN KEY	Foreign key to subgroup.subgrp_id

6. **Hospital Table**:Contains information about hospitals.

Column Name	Data Type	Constraints	Description
hospital_id	STRING	PRIMARY KEY	Hospital ID
hospital_name	STRING	NOT NULL	Hospital name
city	STRING		City
state	STRING		State
country	STRING		Country

7. **Claims Table:** Contains information about claims.

Column Name	Data Type	Constraints	Description
claim_id	LONG	PRIMARY KEY	Claim ID
sub_id	STRING	FOREIGN KEY	Foreign key to subscriber.sub_id
claim_amount	DECIMAL		Claim amount
claim_date	DATE		Claim date

claim_type	**STRING**		**Type of claim**
disease_id	**INTEGER**	**FOREIGN KEY**	**Foreign key to disease.disease_id**
hospital_id	**STRING**	**FOREIGN KEY**	**Foreign key to hospital.hospital_id**
claim_status	**STRING**		**Claim status**

8. Patient Table: Contains information about the patients.

Column Name	Data Type	Constraints	Description
patient_id	**STRING**	**PRIMARY KEY**	**Patient ID**
sub_id	**STRING**	**FOREIGN KEY**	**Foreign key to subscriber.sub_id**
first_name	**STRING**	**NOT NULL**	**Patient's first name**
last_name	**STRING**	**NOT NULL**	**Patient's last name**
birth_date	**DATE**		**Patient's birth date**
gender	**STRING**		**Patient's gender**
address	**STRING**		**Patient's address**
phone	**STRING**		**Patient's phone number**
email	**STRING**		**Patient's email**

Relationships

1. subscriber to subgroup:
    - subscriber.subgrp_id is a foreign key referencing subgroup.subgrp_id.
2. group_subgroup to subgroup:
    - group_subgroup.subgrp_id is a foreign key referencing subgroup.subgrp_id.
3. group_subgroup to group:
    - group_subgroup.grp_id is a foreign key referencing group.grp_id.

4. disease to subgroup:
    - disease.subgrp_id is a foreign key referencing subgroup.subgrp_id.
5. claims to subscriber:
    - claims.sub_id is a foreign key referencing subscriber.sub_id.
6. claims to disease:
    - claims.disease_id is a foreign key referencing disease.disease_id.
7. claims to hospital:
    - claims.hospital_id is a foreign key referencing hospital.hospital_id.
8. patient to subscriber:
    - patient.sub_id is a foreign key referencing subscriber.sub_id.

4. Technologies and Platforms to be used in this solution

- **Data Storage**: AWS S3
- **Data Processing**: Databricks, PySpark
- **Data Warehousing**: AWS Redshift
- **Data Visualization and Reporting**: Databricks
- **Version Control and Deployment**: GitHub
- **Project Management**: Jira