**Requirements Specifications Document: Identification of Disease with Maximum Insurance Claims**

1. Introduction

a. Purpose

The purpose of this requirement is to identify the disease with the highest number of insurance claims. Understanding this can help the healthcare insurance company:

* Gain insights into prevalent health issues among policyholders.
* Adjust policy coverage and develop preventive healthcare incentives.
* Foster strategic partnerships with healthcare providers.
* Improve cost management and customer satisfaction.

b. Intended Audience and Use

This requirement is intended for the following groups within the organization:

* Data Analysts: To analyze high-claim diseases and provide insights on prevalent health trends among customers.
* Developers: To build and validate the data pipeline, ensuring accurate data processing and efficient querying.
* Project Managers: To track progress and confirm alignment with the organization’s goals of enhancing revenue through customer insights.

c. Product Scope

This requirement will help identify high-claim diseases, enabling the healthcare insurance company to:

* Optimize policy offerings.
* Set competitive premiums.
* Enhance customer satisfaction through targeted health initiatives. The resulting insights will support multiple teams, including strategy, marketing, and finance, in shaping effective customer engagement and product development strategies.

d. Definitions and Acronyms

* Claim: A request by policyholders for insurance coverage of specific healthcare expenses.
* Premium: Regular payment by a subscriber to maintain insurance coverage.
* PII (Personally Identifiable Information): Data that can identify an individual, requiring special protection.
* SRS (Software Requirements Specification): Document outlining software functionality and requirements.
* AWS (Amazon Web Services): Cloud platform offering on-demand infrastructure and services.
* S3 (Simple Storage Service): AWS’s object storage service.
* Redshift: AWS’s managed data warehouse service.
* Databricks: Cloud platform for big data processing and machine learning.
* ETL (Extract, Transform, Load): Process of data extraction, transformation, and loading.

2. Overall Description

a. User Needs

* Data Analysts: Require efficient data processing and easy access to insights for report generation.
* Business Strategists and Decision-Makers: Need actionable summaries to support insurance offerings and marketing strategies.
* Pricing Teams: Require claim frequency and cost data to develop accurate pricing models.
* Customer Service and Sales Teams: Need insights on customer behavior and claims trends for targeted marketing.
* IT and Engineering Teams: Need a stable, scalable, and maintainable system that integrates well with AWS, Databricks, etc.

b. Assumptions and Dependencies

* Assumptions:
  + Clean Data: The claims, patients, and subscriber data are mostly clean, structured, and up-to-date.
  + Tools and Technologies: Team members are proficient with the required technologies (e.g., PySpark, AWS, Databricks).
  + Security Compliance: Necessary protocols are in place for PII data protection.
  + Stakeholder Availability: Key stakeholders are available for collaboration.
* Dependencies:
  + Data Sources: Access to competitor and third-party data may impact analysis comprehensiveness.
  + AWS Services: Correct configuration of AWS services (S3, Redshift, EMR) is required.
  + Jira and GitHub: For tracking tasks and version control.
  + Team Expertise: Requires team skills in AWS, Databricks, and PySpark.
  + Performance and Scalability: Efficient handling of large data volumes across platforms is critical.

3. System Features and Requirements

a. Functional Requirements

* Data Ingestion: Ingest data from internal claims databases and external sources.
* Data Storage: Store raw input files in AWS S3 for organized and efficient access.
* Data Cleaning: Handle missing values, duplicates, and incorrect data formats.
* Data Transformation: Standardize data format and load it into Redshift tables.
* Data Processing: Generate insights on disease trends, demographics, and claims patterns using PySpark on Databricks.
* Query Execution: Run Redshift queries to identify diseases with maximum claims.
* Result Storage: Store analysis results in dedicated Redshift tables for easy stakeholder access.
* Regular Updates: Ensure results reflect the latest data.

b. External Interface Requirements

* System Integration: Must integrate with AWS S3, Redshift, and Databricks.
* Notification Access: Stakeholders should receive alerts via Jira, with downloadable reports for review.

c. System Features

* Automated Data Pipeline: An ETL pipeline to manage data ingestion, processing, and loading.
* Error Handling and Logging: Capture and log processing issues and alert the IT team.
* Data Quality Checks: Regular checks to ensure claims data accuracy and completeness.

d. Nonfunctional Requirements

1. Performance Requirements
   * Handle large data volumes efficiently, with tasks completing within SLAs.
2. Security Requirements
   * Implement access controls and comply with data privacy regulations (e.g., HIPAA).
3. Usability Requirements
   * Provide an intuitive dashboard with clear visualizations and downloadable reports (PDF/Excel).
4. Scalability Requirements
   * Ensure scalability to manage growing data volumes and maintain Redshift and Databricks performance.