Requirements Specifications Document

1. Introduction

1.1 Purpose

The purpose of this project is to outline the necessary steps and specifications for creating data pipelines and analytical solutions for Health Care insurance company, where the company aims to enhance its revenue by understanding customer behaviours, customizing offers, and optimizing business strategies.

1.2 Intended Audiences and Use

The intended audience for this document includes developers, data engineers, data scientists, testers, project managers, and stakeholders involved in the implementation and management of the analytics solution.

Developers will use this document as a guide for implementing the required functionalities, while testers will refer to it for creating test cases.

Project managers will utilize it for tracking progress and ensuring alignment with business objectives.

1.3 Product Scope

The scope of this project is the development of data pipelines and analytics using AWS S3, Redshift, Databricks, and Pyspark.

1.4 Definitions and Acronyms

Listed are the technologies and tools used and their definitions.

- AWS: Amazon Web Services is a cloud computing platform.
- S3: It is an object storage service provided by AWS.
- Redshift: It is a data warehouse services managed by AWS.
- Databricks: It is a unified data analytical platform for big data and AI.
- PySpark: Apache spark is a python API used for processing large datasets.
- Jira: It is a project management tool.
- GitHub: It is a platform for hosting and collaborating on Git repositories.

2. Overall Description

2.1 User Needs

The primary users of the system include data analysts, data engineers, and business stakeholders within the Health Care insurance company. These users require a robust data analytics solution that can:

- Collect and integrate competitor data from various sources.
- Cleanse and preprocess the data to ensure accuracy and consistency.
- Analyze customer behavior and preferences to tailor insurance offers.
- Generate actionable insights and reports for strategic decision-making.
- Provide an intuitive interface for accessing and visualizing analytics results.

2.2 Assumptions and Dependencies

- The availability and reliability of data sources from competitors.
- Access to S3, Redshift and Databricks platform.
- Availability of skilled personnel for implementing and maintaining the solution.
- Compliance with data privacy and security regulations.
- Integration with existing systems and databases within the organization.

3. System Features and Dependencies

3.1 Functional Requirements

- Data ingestion and integration from multiple sources.
- Data cleaning and preprocessing workflows.
- Customer behavior analysis and segmentation.
- Royalty calculation algorithms.
- Reporting and visualization tools for generating insights.

3.2 External Interface Requirements

i. User

- The system should provide a user-friendly interface for accessing analytics results and reports.
- Integration with Jira for project management and task tracking.
- Integration with GitHub for version control and code management

ii. Hardware

The system relies on cloud infrastructure provided by AWS (S3, Redshift, EMR) for data storage and processing.

iii. Software

The system utilizes Databricks platform for data analytics and processing.
Pyspark is used for implementing data cleansing and analysis algorithms.

iv. Communications

- Communication between different components of the system occurs over secure channels using HTTPS protocols.
- Integration with external APIs for data ingestion and third-party services**.**

3.3 System Features

- Data pipelines for automated data ingestion and processing.
- Data cleansing modules for identifying and handling missing or erroneous data.
- Analytics algorithms for customer behavior analysis and segmentation.
- Royalty calculation modules based on historical policyholder data.
- Reporting and visualization tools for presenting insights and recommendations.

3.4 Nonfunctional Requirements

i. Performance Requirements

- The system should be capable of processing large volumes of data efficiently.
- Response times for analytics queries should be within acceptable limits.
- Scalability to handle increasing data volumes and user loads.

ii. Safety Requirements

- Data encryption and secure access controls to protect sensitive information.
- Regular backups and disaster recovery mechanisms to prevent data loss.

iii. Security Requirements

- Compliance with data privacy regulations.
- Role-based access control to restrict unauthorized access to sensitive data.
- Monitoring and auditing of system activities for security compliance.

iv. Usability Requirements

- Intuitive user interface for easy navigation and interaction.
- Clear documentation and training materials for users and administrators.

v. Scalability Requirements

- The system should be designed to scale horizontally to accommodate growing data volumes and user demands.
- Automated provisioning and scaling of resources in response to workload fluctuations.

These requirements ensure that the system meets the functional and nonfunctional needs of users while maintaining high performance, security, and usability standards.