**Introduction**

**Purpose**

This project aims to leverage Big Data technologies to aid a healthcare insurance company in increasing its revenue and gaining insights into its customers' behaviors. By analyzing competitor data collected through web scraping and from third-party sources, we plan to offer actionable insights that will help the company in tailoring offers, calculating royalties to customers who bought policies in the past, and enhancing its overall business tactics. This approach will allow the company to customize its insurance policies to meet customer needs better, thereby enhancing customer satisfaction and increasing revenue growth.

**Intended Audience and Use**

This document is intended for developers, testers, project managers, and other key participants in the project. It details the objectives and requirements, establishing a definitive guide for the creation and execution of data pipelines and analytical procedures. Its purpose is to steer the project team toward constructing a solid system that meets the specified requirements and secures the targeted business results.

**Product Scope**

This project's scope includes the creation of robust data pipelines that will handle the processing and analysis of data related to customers and competitors. Through this analysis, we aim to uncover significant trends and patterns, including the most claimed diseases, the demographics of subscribers, and the profitability metrics. The valuable insights derived from this project will enable the company to make decisions based on data, ultimately leading to higher revenue and improved customer satisfaction. This initiative aligns with the company's overall strategic goals and will be smoothly incorporated with the current systems to guarantee efficient functionality and the ability to scale.

**Definitions and Acronyms**

**Definitions**

**Data Pipeline:** Series of data processing steps that include data collection, cleaning, analyzing, storing, and efficiently handling and transforming data.

**Redshift:** A fully managed data warehouse service provided by AWS (Amazon Web Services) that allows for scalable data storage and analysis.

**Databricks:** A cloud-based platform that provides tools for data engineering, machine learning, and analytics, built on top of Apache Spark.

**AWS EMR Studio:** A web-based, integrated development environment for Amazon EMR that allows data engineers and data scientists to develop, visualize, and debug data processing applications.

**Pyspark:** The Python API for Apache Spark, used for large-scale data processing and analytics.

**Jira:** A project management tool used for bug tracking, issue tracking, and project management.

**Acronyms**

**AWS:** Amazon Web Services

**S3:** Simple Storage Service

**EMR:** Elastic Map Reduce

**Overall Description**

**User Needs**

The product is intended for use by data analysts, business strategists, and marketing teams at the healthcare insurance company. Identifying the diverse user base and their specific needs is an essential element of the SRS development process.

**Assumptions and Dependencies**

The project is predicated on using existing technological infrastructures such as AWS, Databricks, and Redshift. It further presupposes the ability to obtain precise and up-to-date information through web scraping and external providers. The project's success hinges on having dependable internet access and the necessary permissions to employ these platforms effectively.

**System Features and Requirements**

**Functional Requirements**

Analyze competitor data to determine the disease with the maximum number of claims.

Identify subscribers under 30 who subscribe to any subgroup.

Determine the group with the maximum subgroups.

Identify the hospital serving the most patients.

Determine the subgroup subscribed to the most.

Count the total number of rejected claims.

Identify the city where most claims originate.

Compare subscription rates between government and private policies.

Calculate the average monthly premium paid by subscribers.

Identify the most profitable group.

List patients under 18 admitted for cancer.

List patients with cashless insurance and total charges ≥ Rs. 50,000.

List female patients over 40 who had knee surgery in the past year.

**External Interface Requirements**

**User Interfaces**

The system shall provide user-friendly dashboards and reports tailored for data analysts and business strategists, ensuring easy access to actionable insights.

**Hardware Interfaces**

The system shall integrate seamlessly with the existing company hardware infrastructure, facilitating efficient data storage and processing.

**Software Interfaces**

The system shall support integration with various AWS services, including but not limited to AWS S3, Databricks, and Redshift, as well as internal databases to ensure a cohesive and efficient data processing environment.

**Communication Interfaces**

The system shall utilize secure APIs for data exchange and notifications, ensuring reliable communication between system components and external entities.

**System Features**

Data storage on AWS S3.

Data processing using Databricks and Pyspark.

Data warehousing on AWS Redshift.

Reporting and visualization tools for data insights.

**Nonfunctional Requirements**

**Performance Requirements**

Efficient data processing and storage to handle large datasets.

**Safety Requirements**

Secure handling and storage of sensitive customer data.

**Security Requirements**

Robust data security measures to protect data from unauthorized access.

**Usability Requirements**

User-friendly interfaces and documentation for ease of use.

**Scalability Requirements**

The system is designed to accommodate future data growth and additional use cases.