Requirements Specifications Document

1. **Introduction:**

The goal of this project is to use Big Data analytics to assist a healthcare insurance firm in improving customer insights and revenue. The business can track claims, evaluate consumer behavior, and customize policy offerings by combining data from multiple sources (web scraping, third-party providers).

1. **Purpose:** The main purpose of this document is to establish the requirements for a data pipeline solution that allows a healthcare insurance company to examine competition data from several sources via a Big Data Ecosystem. The information acquired from this analysis will aid in tracking consumer behavior, tailoring insurance offerings, and computing royalties for previous policyholders, resulting in increased revenue.
2. **Intended Audience and Use:** This document is intended for:

* The ETL pipelines are being implemented by data engineers.
* Data analysts are examining the findings.
* Project managers are in charge of the undertaking.
* The solution is being tested by quality assurance teams.

1. **Product Scope:** The major goal of this project is to build and execute data pipelines that will assist the healthcare insurance firm in making strategic business choices.

The primary advantages include:

* Looking out patterns in insurance claims.
* Understanding the characteristics and habits of subscribers.
* Increasing customer satisfaction with customized offerings.
* Finding valuable sectors to increase revenue.
* Increasing productivity through the automation of data analysis and processing.

1. **Definitions and Acronyms:**

* ETL (Extract, Transform, Load) is a procedure for integrating data from numerous sources.
* Amazon Simple Storage Service (S3) is used for data storage.
* Redshift is an AWS Data Warehouse designed for processing massive datasets.
* EMR: AWS Elastic MapReduce - used for huge data processing with Spark.
* PySpark is a Python API for Apache Spark.
* JIRA: A project management and issue-tracking application.
* Big data refers to large, complicated data sets that must be processed using expert technologies.
* Subscribers are individuals who have acquired insurance plans.

1. Overall Description: The suggested solution is a Big Data Analytics Pipeline that aims to help a healthcare insurance company increase revenue and better understand consumer behavior. This system will use data analytics to examine competition data obtained through web scraping and third-party sources. Processing this data allows the corporation to acquire insights into customer trends, identify high-risk patients, personalize insurance offers, and compute royalty advantages for loyal clients.
2. **User Needs:** A data analytics pipeline is necessary for the healthcare insurance firm to process competition data and derive insightful information. Among the principal users are:

* Business analysts study trends and make data-driven judgments.
* Data engineers are responsible for implementing and maintaining data pipelines.
* Executives: Develop company plans based on the insights.

1. **Assumptions and Dependencies:**

* **AWS** Cloud will host the solution's deployment.
* Third-party providers, web scraping, and competition analysis will be the sources of the data.
* PySpark on AWS EMR will be used to create the data pipeline.
* Analytical processing and data storage will be done with Redshift.
* Visualization and development will be done with Databricks.
* The project will be managed using JIRA and adhere to an agile methodology.
* GitHub will be utilized for collaboration and version control.

1. **System Features and Requirements:**
2. **Functional Requirements:** The data pipeline should be able to:

* Import and store data in AWS S3.
* Clean and preprocess data (handle null values, remove duplicates, and simplify formats).
* Load the processed data into Redshift tables.
* Execute analytical queries to gain insights.
* Use Databricks, that can create reports and visualize data

1. **User Interface**: Databricks visualization dashboards and notebooks.

* **Communication Interface**: Integration with APIs for third-party data sources.

1. **System Features:** The system will provide insights into the following use cases:

* Determine the illness that has the most claims.
* Locate subscribers who are under 30 and have joined any subgroup.
* Find the group with the greatest number of subgroups.
* Determine which hospital sees the greatest number of patients.
* Identify which subgroup has the most subscribers.
* Determine how many claims were denied overall.
* Find the city where the most claims have been filed.
* Examine whether people prefer private or public policies.
* Determine the typical monthly premium that subscribers pay.
* Determine which policy group is the most profitable.
* Enumerate cancer sufferers who are younger than 18.
* Find cashless insurance patients whose bills total more than Rs. 50,000.
* Enumerate women over 40 who have had knee surgery during the last 12 months.

1. **Nonfunctional Requirements:**
2. **Performance:** The system should be able to efficiently process huge datasets using AWS EMR.
3. **Security:** Data should be safely stored in Redshift and AWS S3 and encrypted.
4. **Scalability:** The pipeline should be able to manage increased data volumes while maintaining performance.
5. **Usability:** Business analysts should be able to interpret outcomes using Databricks dashboards.
6. **Reliability:** The pipeline should handle failures carefully and retry operations as needed.