**Introduction:**

Purpose: This requirements specification document outlines the Health Care Insurance Data Analysis Project's functional and non-functional needs. To ensure a successful project completion, developers, testers, and project managers will utilize this document as a guide to understand the project scope, user needs, system features, and other important factors.

Intended Audience and Use: This paper is designed for the project development team, which includes data engineers, data scientists, project managers, and stakeholders. It will direct the system's development, testing, and implementation.

Product Scope: By creating a Big Data ecosystem, the project hopes to follow customer activity, analyze competitor data from several sources, and increase revenue through targeted offers and royalties.

**Overall Description:**

The project aims to create data pipelines for cleaning and analyzing health insurance data, enabling the insurance firm to boost client engagement and revenue.

Product Features: The product will support data ingestion, cleaning, analysis, and visualization via tools such as AWS S3, Redshift, Databricks, and PySpark.

User Classes and Characteristics: The key users will be data analysts, data scientists, and business strategists from the Health Care Insurance Company. They will use the system to learn about customer behavior and rival activity.

**User needs:**

Users require a comprehensive system to examine massive datasets such as health insurance claims, customer demographics, and competitor behaviors.

The system should facilitate effective data cleansing, storage, and retrieval to enable fast analysis and decision-making.

**Assumptions and  Dependencies:**The project assumes the use of AWS services and Databricks for data processing.  
The project's success is contingent upon the availability of precise and uncontaminated data.  
 Technologies Used: PySpark, Jira, GitHub, Databricks, AWS EMR Studio, AWS Redshift, and AWS S3.  
Operating Environment: The system will be deployed on AWS EMR or Databricks for final production.

**System Features and Requirements:**

**Functional requirements:**

Data Ingestion: Save the provided example data to AWS S3 in a folder called input-data.

Data Cleaning: Remove null values, duplicates, and other irregularities from datasets (Patients, Subscribers, Claims, Group\_subgroups).

Data Storage: Upload cleansed data to Redshift tables that correspond with the schema design.

Result Generation: Create and store analysis results in Redshift tables under the Project-Output schema.

**Non-functional Requirements:**

Performance: The system should be able to process huge datasets efficiently using AWS and Databricks infrastructure.

Security: Data must be securely stored and accessible to ensure compliance with health-care privacy legislation.

Scalability: The system should be able to manage growing amounts of data as the project progresses.

**Requirements for External Interfaces**  
User Interfaces: The system will have user interfaces for uploading data, visualizing analysis results, and creating reports.  
Software Interfaces: The system will process and store data using AWS services (S3, Redshift), Databricks, and PySpark.  
**Data requirements:**  
The system will process and save data on insurance claims, patient records, hospitals, diseases, and subscriber information.  
Data cleaning and validation are crucial for ensuring the correctness of analysis results.