Introduction

Our mission to leverage big data analytics to enhance healthcare insurance company revenue, to help them customers better and stay competitive in the market. Through the strategic use of data engineering and analytics, we aim to unlock valuable insights that drive informed decision-making and personalized offerings for their customers

**Objectives**

1. **Enhance Revenue:** Increase the company's revenue through better data-driven strategies.
2. **Better Customer Understanding**: Gain deeper insights into customer behavior, preferences, and conditions.
3. **Customer Behavior Tracking**: Monitor and analyze customer behavior and conditions.
4. **Custom Offer Development:** Develop personalized insurance offers for customers based on the analysis.
5. **Royalty Calculation:** Calculate royalties for customers who have purchased policies in the past.

**Goals**

1. **Improve Customer Retention**: Increase customer retention by offering tailored insurance policies.
2. **Optimize Marketing Strategies:** Use insights gained from data analysis to refine marketing strategies.
3. **Data-Driven Decision Making:** Enable data-driven decision making across various departments.
4. **Efficiency in Royalty Payments:** Streamline the process of calculating and paying out royalties to past customers.

**Scope**

1. **Data Processing**: Clean and integrate data from various sources into a unified database.
2. **Data Analysis:** Analyze customer behavior and competitor data using advanced analytics tools.
3. **Implementation of Big Data Tools:** Use tools like PySpark, Databricks, and AWS services (EMR Studio, S3, Redshift) for data processing and analysis.
4. **Develop Custom Offers:** Create personalized insurance offers based on customer data analysis.
5. **Royalty Calculation:** Develop an automated system to calculate royalties for past customers.
6. **Visualization and Reporting:** Develop interactive dashboards and generate detailed reports.
7. **Project Management:** Use tools like Jira for managing tasks and GitHub for version control.

**Intended Audience and Use:** The intended audience for this Software Requirements Specification (SRS) includes:

* **Developers:** Who will use the SRS as a guide for designing and implementing the software system.
* **Testers:** Who will refer to the SRS to understand the system functionalities and develop test cases for validation and verification.
* **Project Managers:** Who will use the SRS to track project progress, manage resources, and ensure that the project stays within scope and meets stakeholders' expectations.

**Key Terms**

1. **Analyse**: To examine data methodically and in detail to understand its structure, identify patterns, and draw conclusions.
2. **Pyspark**: A Python API for Apache Spark, enabling large-scale data processing and analysis.
3. **Jira**: A tool for project management and issue tracking used primarily in software development.
4. **GitHub**: A web-based platform for version control and collaboration, allowing multiple people to work on projects simultaneously.
5. **Big Data Ecosystem**: A comprehensive framework of tools, technologies, and processes used to collect, store, process, and analyze large and complex data sets.
6. **Data pipelines**: Automated workflows that move data from one system to another, typically involving data extraction, transformation, and loading (ETL).
7. **Databricks**: A cloud-based platform founded by the creators of Apache Spark, used for big data processing and machine learning.
8. **DATASET**: A collection of data, often presented in tabular form, used for analysis and research.
9. **DATA CLEANING:** The process of detecting and correcting (or removing) errors and inconsistencies in data to improve its quality.
10. **Schema**: The structure of a database that defines how data is organized, including the tables, fields, and relationships between them.
11. **xls:** A file extension for Excel spreadsheets, commonly used for storing and manipulating tabular data.
12. **Visualisations**: Graphical representations of data designed to make complex data more understandable and accessible.
13. **AWS EMR Studio**: An integrated development environment (IDE) on Amazon Web Services (AWS) for data scientists and engineers to build, visualize, and debug data applications.
14. **AWS S3**: Amazon Simple Storage Service, a scalable object storage service used for storing and retrieving any amount of data at any time.
15. **AWS Redshift:** A fully managed data warehouse service on AWS that allows for fast querying and analysis of large datasets.

**User Needs:** The primary users of the product will include:

* **Healthcare Insurance Company Executives:** They will utilize the insights derived from big data analytics to make strategic decisions related to revenue enhancement, customer engagement, and competitive positioning.
* **Data Analysts and Scientists:** Responsible for collecting, processing, and analyzing data to extract actionable insights and develop personalized insurance offers.
* **Marketing and Sales Teams:** Will leverage the generated insights to tailor marketing campaigns and sales strategies to specific customer segments.
* **Customers:** Will benefit from personalized insurance offers that better meet their needs and preferences.

**Assumptions and Dependencies:**

* **Technological Assumptions:** We assume the availability and functionality of necessary technologies such as big data analytics tools (e.g., PySpark, Databricks), cloud computing services (e.g., AWS EMR Studio, S3, Redshift), and web scraping tools.
* **Data Accessibility:** We assume access to relevant and reliable data sources, including competitor data and customer information, for analysis.
* **Compliance:** The project assumes compliance with relevant data privacy regulations (e.g., GDPR, HIPAA) and adherence to internal data governance policies.
* **Integration with Existing Systems:** The product may need to integrate with existing systems within the company, such as CRM platforms or internal databases, to access necessary data for analysis.

**Functional Requirements:** Functional requirements describe the specific functionalities or capabilities that the software must provide. These include:

* **Data Collection:** The system should be able to collect data from various sources, including competitor data and customer information.
* **Data Processing:** It should process and clean the collected data to ensure accuracy and consistency.
* **Data Analysis:** The system should analyze data to identify customer behavior patterns, trends, and insights.
* **Custom Offer Generation**: It should generate personalized insurance offers based on the analysis of customer data.
* **Royalty Calculation:** The system should calculate royalties for customers who have purchased policies in the past.
* **Reporting:** It should generate reports and visualizations to present the analyzed data and insights to stakeholders.
* **Integration:** The system should integrate with existing software systems and tools within the organization.

**External Interface Requirements:**

i. **User Interface:** The system should have a user-friendly interface for data analysts, executives, and other stakeholders to interact with.

ii. **Hardware Interface:** It should be compatible with standard hardware configurations used within the organization.

iii. **Software Interface:** The system should be compatible with relevant software tools and platforms, such as big data analytics tools and CRM systems.

iv. **Communication Interface:** It should support communication protocols for data exchange with external systems and services.

**System Features:** System features are essential functionalities required for the system to operate effectively. These include:

* Big Data Ecosystem Integration
* Web Scraping Functionality
* Data Cleaning and Preprocessing
* Advanced Analytics and Machine Learning Algorithms
* Personalized Offer Generation Engine
* Automated Royalty Calculation System
* Interactive Dashboards and Reporting Tools

**Nonfunctional Requirements:**

i. **Performance Requirements:** The system should be able to handle large volumes of data efficiently and provide timely insights.

ii. **Safety Requirements:** It should adhere to data privacy and security regulations to protect sensitive customer information.

iii. **Security Requirements:** The system should have robust security measures in place to prevent unauthorized access and data breaches.

iv. **Usability Requirements:** It should be intuitive and easy to use, requiring minimal training for users.

v. **Scalability Requirements:** The system should be scalable to accommodate future growth in data volume and user base.

## 