Requirements Specifications Document

# Introduction- Enhancing Revenue and Customer Understanding in Health Care Insurance using Big Data Analysis

## Purpose - The purpose of the project outlined above is to address the challenges faced by a Health Care insurance company in enhancing its revenue and understanding its customers.

## Intended Audience and Use - The project and its outcomes are designed to cater to a diverse range of stakeholders within the Health Care insurance company, as well as those involved in its operations, decision-making, and strategic planning. The intended audience are:

* **Executives and Decision makers:** They can use the information to formulate business strategies, allocate resources, and make informed decisions that impact the company's growth and direction.
* **Marketing Sales Team:** This information helps them tailor their approaches, design targeted campaigns, and engage with potential policyholders more effectively.
* **Product Development teams:** The project's data-driven insights can guide product development teams in creating insurance policies that cater to specific customer needs and preferences. This results in products that are aligned with market demands.
* **Customer service representative**: The insights gained from the project can assist customer service representatives in understanding customer behaviors and preferences. This enables them to provide more relevant assistance and personalized recommendations.
* **Data Analyst and Scientist:** Professionals involved in data analysis and data science can use the project as a reference for understanding best practices in data cleaning, data processing, and result generation.
* **Technical Teams:** The technical teams responsible for maintaining and optimizing the data pipeline, as well as the chosen technology stack (AWS S3, Redshift, Databricks, etc.), can gain insights into best practices, architectural design, and implementation details.
* **Project Managers:** Project managers can find value in the project's documentation, sprint planning, and task tracking methods, which are presented using tools like Jira. They can adapt these practices to other projects within the company.

## Product Scope – The product scope outlines the boundaries and objectives of the data analysis system that we are developing for the Health Care insurance company. It defines what the system will include and what it will accomplish. Here's a breakdown of the benefits, objectives, and goals for this product:

**Benefits:**

**1. Enhanced Revenue Generation:** The primary benefit is to enhance the company's revenue by gaining insights into customer behavior, preferences, and needs. This will enable the customization of insurance offers and royalty calculations, leading to higher policy sales and customer satisfaction.

**2. Data-Driven Decision Making: The** system will empower the company to make informed business decisions by analyzing patterns, trends, and correlations in the data. This will result in more strategic and targeted marketing and operational strategies.

**3. Customer Satisfaction:** By offering personalized insurance policies based on customer behaviors and preferences, the company can improve customer satisfaction and retention rates.

**4. Competitive Advantage:** With a better understanding of the market and customer needs, the company can stay ahead of competitors by tailoring offerings to match changing trends.

**5. Efficient Operations:** The insights gained from the system will optimize operational efficiency, allowing the company to allocate resources more effectively and streamline processes.

**Objectives:**

**1. Data Collection and Integration:** The system's objective is to gather data from various sources, including scraping and third-party providers, and integrate it into a unified repository for analysis.

**2. Data Cleaning and Transformation:** Ensure that the data is cleansed, transformed, and standardized, enabling accurate and reliable analysis.

**3. Customer Behavior Analysis:** Analyze customer behaviors, patterns, and preferences to identify opportunities for offering customized insurance policies.

**4. Royalty Calculation:** Develop mechanisms to calculate royalties for customers who have previously bought insurance policies, incentivizing repeat business.

**5. Insights Generation:** Generate meaningful insights through data analysis to guide business strategies, marketing campaigns, and policy customization.

**6. Performance and Scalability:** Ensure that the system can handle growing data volumes, increasing user loads, and still maintain acceptable performance levels.

**7. Data Privacy and Security:** Implement robust security measures to protect sensitive customer data and comply with data protection regulations.

**Goals:**

**1. Increase Policy Sales:** The goal is to increase policy sales by 20% within the first year of system deployment through targeted marketing based on customer insights.

**2. Enhance Customer Retention:** Improve customer retention by 15% within the first year by offering tailored policies and rewards to loyal customers.

**3. Reduce Rejected Claims:** Aim to reduce the number of rejected claims by 10% through better risk assessment and customer profiling.

**4. Optimize Premium Pricing:** Utilize data analysis to optimize premium pricing, ensuring competitiveness while maintaining profitability.

**5. Real-time Insights:** Provide real-time insights to business stakeholders, allowing them to make timely decisions and react to market changes swiftly.

**6. Compliance and Data Privacy:** Ensure compliance with healthcare data regulations, avoiding legal issues and building trust with customers.

**7. Seamless Integration:** Facilitate seamless integration with existing company systems and tools, ensuring cross-functional access and usability.

## Definitions and Acronyms

Definitions:

1. **Health Care Insurance Company**: The company that provides insurance coverage for medical expenses and healthcare services to its policyholders.
2. **Big Data Ecosystem:** A collection of technologies and tools used to store, process, analyze, and extract insights from large and complex datasets.
3. **Competitors' Data:** Information related to other insurance companies operating in the same market, including their policies, customer demographics, claims data, and business strategies.
4. **Data Scraping:** The process of extracting data from websites and online sources using automated methods.
5. **Third-Party Sources:** External data providers that offer information, datasets, and analytics services that can be integrated into the company's data analysis processes.
6. **Royalties:** Payments made to customers as a form of reward or incentive for purchasing insurance policies in the past.
7. **Data Pipelines:** A series of data processing steps that transform raw data into meaningful and usable information.
8. **AWS S3:** Amazon Simple Storage Service, a scalable object storage service provided by Amazon Web Services.
9. **AWS Redshift:** Amazon Redshift, a fully managed data warehousing service provided by Amazon Web Services.
10. **Databricks:** A cloud-based platform for big data analytics and collaboration, often used for Apache Spark-based processing.
11. **AWS EMR:** Amazon Elastic MapReduce, a cloud service used for processing large amounts of data using frameworks like Apache Spark and Hadoop.
12. **Pyspark:** The Python library for Apache Spark, used for distributed data processing and analysis.
13. **Jira**: A popular project management and issue tracking tool used to manage tasks, sprints, and projects.
14. **GitHub**: A web-based platform for version control and collaboration on software development projects.

Acronyms:

1. EMR: Elastic MapReduce
2. S3: Simple Storage Service
3. AWS: Amazon Web Services
4. QA: Quality Assurance
5. ETL: Extract, Transform, Load
6. SQL: Structured Query Language
7. API: Application Programming Interface
8. UI: User Interface
9. UX: User Experience
10. BI: Business Intelligence
11. KPI: Key Performance Indicator
12. PDF: Portable Document Format
13. CSV: Comma-Separated Values
14. JSON: JavaScript Object Notation

# Overall Description – A Health Care insurance company aims to boost revenue and improve customer understanding using Big Data. They plan to analyze competitor data from various sources via scraping and third-party channels. The analysis intends to monitor customer behavior, enabling customized insurance offers and royalties for policyholders, ultimately driving increased revenue. The project goal involves constructing data pipelines to facilitate strategic decision-making, leading to enhanced revenue through customer behavior analysis, personalized offers, and royalty distributions.

## User Needs –

* **Executives and Decision Makers:**

Needs: Executives require high-level insights to make informed business decisions. They need concise summaries of customer behaviors, profitable groups, and overall revenue trends.

Product Use: They will access dashboards and reports generated by the product to understand the market landscape and identify strategic opportunities.

* **Marketing and Sales Teams:**

Needs: Marketing and sales teams need customer segmentation and personalized insights to design effective campaigns and offers.

Product Use: They will utilize visualizations and reports to tailor marketing strategies based on customer demographics, preferences, and historical behaviors.

* **Product Development Teams:**

Needs: Product teams need customer preferences and data on policy subscriptions to design insurance products that align with market demand.

Product Use: They will analyze subscriber preferences and behaviors to develop insurance policies that cater to specific customer needs.

* **Customer Service Representatives:**

Needs: Customer service representatives require insights into customer behaviors to provide personalized assistance and recommendations.

Product Use: They will access customer profiles and preferences to enhance customer interactions and address queries effectively.

* **Data Analysts and Scientists:**

Needs: Data analysts need access to clean, organized data for in-depth analysis and modeling.

Product Use: They will utilize the clean data stored in the Redshift database to conduct advanced analyses and derive insights.

* **Technical Teams:**

Needs: Technical teams need clear documentation and well-structured code to understand the data pipeline and solution architecture.

Product Use: They will refer to documentation and code repositories for maintenance, troubleshooting, and potential improvements.

* **Project Managers:**

Needs: Project managers require well-defined tasks and sprint planning for effective project management.

Product Use: They will use the Jira board to track progress, allocate resources, and ensure that tasks are completed on schedule.

* **Training and Learning Groups:**

Needs: Training teams need educational materials to onboard new team members and enhance their skills in data analysis.

Product Use: They will utilize documentation, training materials, and the live presentation to educate team members about the project's processes and outcomes.

* **Data Compliance and Security Teams:**

Needs: Compliance teams need to ensure that data handling and storage adhere to regulatory standards.

Product Use: They will review documentation and the solution architecture to ensure data security and compliance.

* **End Users (Viewers of Presentations):**

Needs: End users attending presentations need a clear understanding of the project's purpose, process, and outcomes.

Product Use: They will engage with the live class presentation and visualizations to grasp the project's value and potential impact.

## Assumptions and Dependencies –

* **Data Quality and Availability:**

Assumption: The provided sample data on AWS S3 is accurate, relevant, and representative of the actual data.

Dependency: The accuracy and quality of the analysis heavily rely on the accuracy and completeness of the data provided.

* **Data Privacy and Compliance:**

Assumption: The data used for analysis is compliant with relevant data protection laws and regulations.

Dependency: Ensuring data privacy and compliance with regulations is essential to avoid legal and ethical issues.

* **AWS Services and Infrastructure:**

Assumption: The project will be executed within the AWS ecosystem, utilizing services like S3, Redshift, EMR Studio, and Databricks.

Dependency: The availability and proper functioning of AWS services are crucial for successful data processing and analysis.

* **Data Cleaning Accuracy:**

Assumption: The data cleaning process effectively identifies and addresses missing values, duplicates, and inconsistencies.

Dependency: Accurate data cleaning is necessary to ensure the validity and reliability of the analysis results.

* **Domain Knowledge:**

Assumption: The team working on the project has a solid understanding of health insurance, healthcare domain, and relevant business rules.

Dependency: Accurate interpretation of data and creation of meaningful insights requires domain expertise.

* **Technical Expertise:**

Assumption: The team possesses the necessary technical skills to implement the data pipelines, perform data cleaning, and generate analysis results.

Dependency: Adequate technical knowledge and skills are essential for successful project execution.

* **Tool Proficiency:**

Assumption: The team is proficient in using tools like Pyspark, AWS services, Databricks, Jira, and GitHub.

Dependency: Proper utilization of these tools is essential for efficient development, collaboration, and deployment.

* **Project Timeline and Planning:**

Assumption: The project timeline, as outlined in the two-week sprint plan, is feasible and accounts for potential challenges.

Dependency: Adhering to the planned timeline is important for meeting project goals and expectations.

* **Testing and Quality Assurance:**

Assumption: Proper testing methodologies will be employed to validate the accuracy and reliability of the analysis results.

Dependency: Thorough testing is necessary to ensure that the insights and outcomes are valid and actionable.

* **Availability of Resources:**

Assumption: Required resources, such as computing power, storage, and technical personnel, will be available as needed.

Dependency: Sufficient resources are essential for efficient execution and completion of the project.

* **Documentation and Communication:**

Assumption: Clear and detailed documentation of requirements, design, implementation, and testing will be maintained throughout the project.

Dependency: Comprehensive documentation aids in knowledge transfer, collaboration, and project continuity.

* **Stakeholder Collaboration:**

Assumption: Effective collaboration and communication with stakeholders, including the Health Care insurance company, are established.

Dependency: Stakeholder input and feedback are critical for aligning the project with business goals and expectations.

* **Scaling and Performance:**

Assumption: The designed data pipelines and solutions are scalable to handle larger datasets and increased user demand.

Dependency: Scalability ensures that the solution remains effective as the volume of data and user interactions grow.

* **Data Accuracy and Relevance:**

Assumption: The analysis results accurately reflect real-world scenarios and provide actionable insights.

Dependency: Accurate analysis is crucial for making informed business decisions and strategies.

* **Change Management:**

Assumption: Changes in requirements or unforeseen challenges will be addressed with effective change management processes.

Dependency: Flexible project management allows for adjustments while minimizing disruptions.

# System Features and Requirements -

## Functional Requirements –

● Identify the disease with the highest number of claims.

● Locate subscribers under 30 who have joined any subgroup.

● Determine the group with the most subgroups.

● Identify the hospital treating the highest number of patients.

● Find the subgroup with the highest subscription frequency.

● Calculate the total count of rejected claims.

● Discover the city with the most incoming claims.

● Analyze whether subscribers prefer Government or private policy groups.

● Calculate the average monthly premium paid by subscribers.

● Find the most profitable policy group.

● List patients under 18 admitted for cancer.

● Display patients with cashless insurance and charges ≥ Rs. 50,000.

● Show females above 40 who had knee surgery in the last year.

## External Interface Requirements -

● AWS S3

● AWS Redshift

● Databricks

● AWS EMR Studio

● Pyspark

● Jira

● GitHub

## System Features –

**1. Data Ingestion and Integration:**

- Ability to ingest and integrate data from various sources including scraping and third-party sources into a unified data repository.

**2. Data Cleaning and Preprocessing:**

- Modules for detecting and handling missing values, duplicates, and inconsistencies in the data to ensure its quality and reliability.

3**. Data Transformation and Enrichment:**

- Tools to transform and enrich the data by applying necessary calculations, aggregations, and mappings to derive meaningful insights.

**4. Data Storage and Management:**

- Storage solutions, such as AWS S3 and Redshift, to securely store and manage large volumes of data for efficient querying and analysis.

**5. Data Pipeline Orchestration:**

- Framework to automate and orchestrate end-to-end data pipelines, including scheduling and monitoring tasks for regular updates.

**6. Exploratory Data Analysis (EDA):**

- Visualization and analytics tools for performing exploratory data analysis to uncover patterns, trends, and correlations within the data.

**7. Advanced Analytics and ML Models:**

- Integration of machine learning models and advanced analytics techniques to predict customer behaviors, identify profitable groups, and optimize offerings.

**8. Querying and Reporting:**

- Querying capabilities to retrieve specific insights from the data, generate reports, and visualize analysis results.

**9. Performance Optimization:**

- Techniques to optimize query performance for quick and efficient retrieval of insights from large datasets.

**10. Data Privacy and Security:**

- Implementing data encryption, access controls, and compliance measures to ensure data privacy and security.

**11. User Authentication and Access Control:**

- User management system to control access to different parts of the system based on user roles and permissions.

**12. Collaboration and Workflow:**

- Integration with tools like Jira and GitHub to support collaboration among team members, manage tasks, and track progress.

13. **Deployment and Version Control:**

- Process for deploying code changes to production environment and maintaining version control for reproducibility.

**14. Error Handling and Logging:**

- Mechanisms for logging errors, exceptions, and system activities to facilitate debugging and monitoring.

**15. Scalability and Resource Management:**

- Designing the system to handle increasing data volumes and user interactions by scaling resources as needed.

**16. Real-time Processing:**

- Real-time data processing capabilities for immediate insights and responsiveness to changing data.

**17. Integration with External Systems:**

- Integration with external systems and APIs for fetching additional data or triggering actions based on analysis results.

**18. Monitoring and Performance Metrics:**

- Monitoring tools to track system performance, resource utilization, and user interactions.

**19. Documentation and User Guides:**

- Providing comprehensive documentation and user guides to ensure proper understanding and usage of the system.

**20. Backup and Disaster Recovery:**

- Implementing backup strategies and disaster recovery plans to ensure data availability in case of failures.

## Nonfunctional Requirements –

### 

### **1. Performance Requirements:**

* Requirement: The system should provide query response times for common analytical tasks, such as retrieving insights and generating reports, within 3 seconds.
* Details: Ensuring fast query performance enhances user experience and productivity.

**2. Safety Requirements:**

* Requirement: The system must handle sensitive healthcare data in compliance with relevant data protection laws (e.g., HIPAA), ensuring patient privacy and security.
* Details: Safeguarding patient data is critical to maintain trust and comply with legal requirements.

**3. Security Requirements:**

* Requirement: User authentication and authorization mechanisms should be implemented to ensure that only authorized users can access and manipulate data.
* Details: Security measures should include data encryption, secure connections, and role-based access controls.

**4. Usability Requirements:**

* Requirement: The user interfaces should be intuitive, easy to navigate, and designed to accommodate users with varying levels of technical expertise.
* Details: Good usability contributes to user adoption and efficient interaction with the system.

**5. Scalability Requirements:**

* Requirement: The system should scale to accommodate a potential 50% increase in data volume and user traffic over the next two years.
* Details: Scalability ensures that the system can handle growing demands without compromising performance.

**6. Availability Requirements:**

* Requirement: The system should have an uptime of at least 99.9% to ensure that users can access and use the system reliably.
* Details: High availability is essential for uninterrupted data analysis and decision-making.

**7. Reliability Requirements:**

* Requirement: The system should handle unexpected errors gracefully and recover from failures without data loss or corruption.
* Details: Ensuring reliable operation reduces the risk of downtime and data loss.

**8. Data Integrity Requirements:**

* Requirement: Data stored and processed by the system must maintain its accuracy and integrity throughout the analysis process.
* Details: Data accuracy is essential for making informed business decisions.

**9. Compliance Requirements:**

* Requirement: The system should adhere to relevant industry regulations and standards, such as healthcare data compliance (e.g., HIPAA) and data protection regulations (e.g., GDPR).
* Details: Compliance ensures legal and ethical handling of sensitive data.

**10. Disaster Recovery Requirements:**

* Requirement: The system should have a disaster recovery plan in place to quickly restore functionality in the event of data loss or system failure.
* Details: Preparedness for potential disasters minimizes downtime and data loss.

**11. Compatibility Requirements:**

* Requirement: The system should be compatible with major web browsers (e.g., Chrome, Firefox, Safari) and mobile devices (iOS, Android) to ensure widespread access.
* Details: Compatibility enhances user reach and accessibility.

**12. Maintainability Requirements:**

* Requirement: The codebase should be well-documented and structured, making it easier for developers to maintain and update the system in the future.
* Details: Maintainable code reduces technical debt and supports future enhancements.

**13. Performance Testing Requirements:**

* Requirement: Performance testing should be conducted to verify that the system meets the defined performance requirements under expected user loads.
* Details: Performance testing ensures that the system can handle real-world usage scenarios.

## 