Requirements Specifications Document (SRS)

**Project Name:** Healthcare Insurance Competitor Data Analysis Pipeline for Customer Attraction and Revenue Growth

**Version:** 1.0

**Prepared By:** Bikash Lama Bamjan

**Date:** 08/20/2024

# Introduction

A healthcare insurance company is facing challenges in improving its revenue and gaining deeper insights into customer behavior. To address this, the company plans to leverage a Big Data ecosystem to analyze competitor data gathered from various sources, including web scraping and third-party providers. By performing this analysis, the company aims to better understand customer patterns, health conditions, and preferences. This will enable them to create tailored insurance offers, attract new customers, and reward loyal policyholders, ultimately driving revenue growth.

The goal of this project is to develop data pipelines that will enable the company to analyze customer behaviors and formulate effective business strategies. These insights will empower the company to send personalized offers and loyalty rewards to customers, further enhancing customer engagement and boosting overall revenue.

### Requirements

1. **Identify Top Diseases by Claim Volume**: Determine which disease has the highest number of insurance claims.
2. **Young Subscribers in Subgroups**: Find all subscribers under the age of 30 who are subscribed to any subgroup.
3. **Group with Maximum Subgroups**: Identify the insurance group that contains the highest number of subgroups.
4. **Hospital with Most Patients**: Find the hospital that serves the largest number of patients.
5. **Most Subscribed Subgroups**: Identify the subgroups that have been subscribed to the most times.
6. **Rejected Claims**: Calculate the total number of insurance claims that were rejected.
7. **City with Most Claims**: Determine the city from which the highest number of claims originates.
8. **Popular Insurance Type**: Analyze whether subscribers prefer government or private insurance policies.
9. **Average Monthly Premium**: Calculate the average monthly premium that subscribers pay to the insurance company.
10. **Most Profitable Group**: Identify the insurance group that generates the highest profitability.
11. **Patients under 18 with Cancer**: List all patients under the age of 18 who were admitted for cancer treatment.
12. **High-Cost Cashless Claims**: Identify patients with cashless insurance whose total charges are greater than or equal to Rs. 50,000.
13. **Female Patients Over 40 with Knee Surgery**: List all female patients over the age of 40 who underwent knee surgery in the past year.

### 1.2 Purpose

The purpose of this project is to empower a healthcare insurance company to enhance its business strategies by leveraging data-driven insights. Through the analysis of competitor and customer data, the company aims to optimize its revenue streams, improve customer satisfaction, and gain a competitive edge in the market. The pipeline will facilitate the extraction, transformation, and loading of large volumes of data to generate actionable insights. These insights will enable the company to:

* Identify key trends in disease claims and customer behavior.
* Personalize insurance offers and improve engagement with different customer segments.
* Streamline operations by identifying high-performing hospitals and insurance groups.
* Make informed decisions regarding policy pricing and profitability.
* Address customer needs more effectively by understanding their preferences and health conditions.

By achieving these goals, the company will be able to attract new customers, retain existing ones, and increase overall profitability through targeted offerings and loyalty programs.

### 1.3 Intended Audience and Use

#### Intended Audience:

* **Data Engineers**: Responsible for designing, building, and maintaining the ETL pipelines and data infrastructure. They will use this document to understand the data requirements, pipeline architecture, and data transformation needs to ensure efficient and accurate data processing.
* **Data Scientists**: Engaged in analyzing and interpreting the processed data to generate insights and develop predictive models. This document will guide them in understanding the data sources, transformations, and the type of data available for analysis.
* **Data Analysts**: Tasked with generating reports and visualizations based on the processed data. They will use this document to familiarize themselves with the data structure, key metrics, and the specific queries required to derive actionable business insights.
* **Business Users**: Includes stakeholders such as product managers, marketing teams, and executives who will leverage the insights generated from the data to make strategic business decisions. This document will help them understand the data's relevance to their business objectives and how it can be used to inform their strategies.

#### Use:

* **Data Engineers**: To guide the development and implementation of the ETL pipelines, ensuring data is accurately extracted, transformed, and loaded as per the requirements specified. They will use this document to align technical solutions with business needs and data specifications.
* **Data Scientists**: To comprehend the data processing workflows and access the cleaned and transformed data required for building models and performing advanced analytics. This document provides the necessary context for interpreting data and applying analytical techniques.
* **Data Analysts**: To design and execute queries and reports based on the specified requirements. They will use this document to identify the key data points, metrics, and insights needed for reporting and decision-making.
* **Business Users**: To understand how data-driven insights will impact business operations and strategy. This document will help them grasp the scope of the data analysis, the key findings, and how these insights align with their business objectives.

Overall, this document serves as a comprehensive guide to ensure all stakeholders are aligned on the objectives, data requirements, and usage of the ETL pipeline and the resulting data insights.

### Product Scope

The output from the ETL pipeline should help the company to optimize its revenue streams, improve customer satisfaction, and gain a competitive edge in the market by analyzing the competitor and customer data. The key objectives of this product are outlined as below:

#### Key Objectives of the Data Pipeline:

* **Revenue Optimization:**

By analyzing sales trends, pricing strategies, and market demand, the pipeline will generate insights to the business users that should help the company identify new revenue opportunities and maximize existing ones. This could involve dynamic pricing, targeted promotions, or identifying cross-sell and up-sell opportunities.

* **Improved Customer Satisfaction:**

Understanding customer behavior through data analysis is critical for enhancing customer experiences. Data analysts can analyze the pipeline for trends in customer feedback, purchasing patterns, and service usage to provide actionable insights for improving products, services, and customer interactions.

* **Gaining a Competitive Edge:**

Competitor data analysis will allow the company to benchmark its performance against industry leaders. By using the pipeline data on competitors' strategies, market positioning, and strengths, data scientist team can build models that can predict market opportunities and customer needs should help the company to stay ahead in the market.

### 1.5 Definitions and Acronyms

#### Definitions

1. **Insurance Claims**

* Insurance claims are formal requests made by policyholders to their insurance provider for payment or reimbursement for services or coverage as specified by their insurance policy. This could include medical expenses, damages, or losses covered by the insurance plan.

**2. Customers**

* In the context of healthcare insurance, customers are individuals or entities who purchase insurance policies from the company. They pay premiums and expect coverage for medical costs, depending on the terms of their insurance plan.

**3. Revenue**

* Revenue is the total amount of money earned by the healthcare insurance company through premiums paid by customers, investment income, and any other sources of income. It is a key financial metric indicating the company’s overall earnings before expenses are deducted.

**4. Claim Rejection**

* Claim rejection occurs when an insurance provider denies a policyholder’s request for reimbursement or payment for a specific claim. Rejections can occur for various reasons, such as incomplete information, non-coverage under the policy, or fraudulent claims.

**5. Pricing Strategies**

* Pricing strategies in healthcare insurance involve determining the premium rates that customers will pay for insurance coverage. These strategies are often based on factors such as risk assessment, competition, regulatory requirements, and customer demographics.

**6. Profitability Metrics**

* Profitability metrics are financial indicators used to assess the financial health of a company. In the context of healthcare insurance, these metrics can include measures like gross profit, net income, return on investment (ROI), and profit margins. They help evaluate the company's ability to generate profit relative to its costs and revenues.

**7. Big Data**

* Big Data refers to extremely large and complex data sets that are difficult to process and analyze using traditional data-processing methods. In healthcare insurance, Big Data can include vast amounts of customer information, claims data, and market trends that require advanced analytics and tools to extract valuable insights.

**8. Big Data Ecosystem**

* A Big Data ecosystem is a collection of tools, technologies, and frameworks that work together to collect, store, process, and analyze large-scale data sets. In the healthcare insurance industry, a Big Data ecosystem typically includes cloud platforms (e.g., AWS), data lakes (e.g., AWS S3), distributed computing frameworks (e.g., Apache Spark), and data visualization tools (e.g., Tableau, Power BI, Amazon QuickSight).

**9. Cloud Services**

* Cloud services refer to a wide range of services delivered over the internet, including computing power, storage, and applications. In this project, cloud services provided by AWS (such as S3, Redshift, and EMR) enable the scalable processing, storage, and management of Big Data.

**10. ETL Pipeline**

* An ETL (Extract, Transform, Load) pipeline is a series of processes that extract data from multiple sources, transform it into a usable format (e.g., cleaning, aggregating, or enriching the data), and load it into a data warehouse or database for analysis. ETL pipelines are essential for preparing data for insights and reporting.

**11. Web Scraping**

* Web scraping is the process of automatically extracting data from websites. It involves using software tools to gather information from web pages, such as competitor pricing, customer reviews, or market trends, which can then be analyzed to support business decisions.

**12. Third-Party Providers**

* Third-party providers are external entities or vendors that supply additional data or services to support the company's operations. In healthcare insurance, these providers might offer demographic data, medical records, or market analysis reports that the company can use to enhance its decision-making and customer strategies.

**13. Predictive Models**

* Predictive models use historical data and statistical algorithms, often enhanced by machine learning, to forecast future outcomes. In healthcare insurance, predictive models can be used to estimate future claims, identify high-risk customers, and optimize pricing strategies.

**14. Machine Learning Algorithms**

* Machine learning algorithms are a subset of artificial intelligence (AI) that allow systems to learn from data and improve their performance over time without being explicitly programmed. In this project, machine learning algorithms can be used to identify patterns in customer behavior, predict trends, and automate decision-making processes.

**15. Advanced Analytics**

* Advanced analytics refers to the application of sophisticated techniques, such as machine learning, predictive modeling, and statistical analysis, to extract deeper insights from data. This goes beyond basic reporting and includes actionable insights that can drive strategic business decisions.

**16. Platform-Independent**

* Platform-independent refers to software or systems that can operate on multiple computing platforms without modification. In the context of this project, platform-independent tools and systems can be used across various environments (e.g., Windows, Linux, cloud) without compatibility issues.

#### Acronyms

1. **SRS**

* **Software Requirements Specification**: A detailed document that describes the functional and non-functional requirements of a software system, including how the system should behave, its interactions with other systems, and any constraints.

1. **CRM**

* **Customer Relationship Management**: A system used by companies to manage and analyze customer interactions and data throughout the customer lifecycle. CRM systems help improve customer retention and drive sales growth by enhancing customer relationships.

1. **KPI**

* **Key Performance Indicator**: A measurable value that demonstrates how effectively a company or a specific business process is achieving its objectives. KPIs are used to track performance and identify areas for improvement.

1. **HIPAA**

* **Health Insurance Portability and Accountability Act**: A US law that mandates the protection of patient data and privacy. It establishes regulations for the handling of healthcare information and imposes penalties for non-compliance.

1. **GDPR**

* **General Data Protection Regulation**: A regulation in the European Union (EU) that governs data protection and privacy for individuals within the EU. GDPR sets strict rules for data processing and grants individuals more control over their personal information.

1. **ETL**

* **Extract, Transform, Load**: A process in data warehousing that involves extracting data from various sources, transforming it into a usable format, and loading it into a target database or data warehouse. It's essential for data integration and analysis.

1. **AWS**

* **Amazon Web Services**: A comprehensive cloud platform that offers various computing services, including storage, processing, analytics, and machine learning. AWS is widely used for scalable cloud infrastructure in data-driven projects.

1. **EMR**

* **Elastic MapReduce**: A cloud-based big data platform provided by AWS that simplifies processing large data sets using distributed computing frameworks like Apache Hadoop and Apache Spark. EMR is commonly used for large-scale data processing tasks.

1. **S3**

* **Simple Storage Service**: An object storage service provided by AWS that offers scalable storage solutions for data, supporting a variety of use cases, including data lakes, backups, and archival storage.

1. **CI/CD**

* **Continuous Integration/Continuous Deployment**: A software development practice where code changes are automatically built, tested, and deployed. CI/CD aims to minimize manual intervention, ensuring rapid and reliable software updates.

1. **ORC**

* **Optimized Row Columnar**: A columnar storage format for Hadoop and big data workloads. ORC files are optimized for high-performance data processing and are commonly used in data lakes and analytics projects.

1. **CSV**

* **Comma-Separated Values**: A widely used plain text format for representing tabular data, where each line corresponds to a row and each field is separated by a comma. CSV files are often used for data exchange between different systems.

1. **HTTPS**

* **Hypertext Transfer Protocol Secure**: An extension of HTTP that provides secure communication over a computer network by encrypting the data in transit using SSL/TLS protocols. It ensures that sensitive information is protected during transmission.

1. **SSL/TLS**

* **Secure Sockets Layer / Transport Layer Security**: Cryptographic protocols designed to provide secure communication over a network. SSL is the predecessor of TLS, and both are used to encrypt data transmitted over the internet, including HTTPS connections.

1. **SQS**

* **Simple Queue Service**: A fully managed message queuing service provided by AWS that enables decoupling and scaling of microservices, distributed systems, and serverless applications by sending, storing, and receiving messages between components.

1. **SNS**

* **Simple Notification Service**: A fully managed messaging service provided by AWS that enables the sending of notifications to distributed systems, microservices, and serverless applications. SNS supports various protocols like email, SMS, and HTTP/S endpoints.

# Overall Description

This project named “Healthcare Insurance Competitor Data Analysis Pipeline for Customer Attraction and Revenue Growth” should act as the data pipeline which will be essential for the company to overcome challenges related to stagnant revenue growth and limited insights into customer behavior. The company needs to better understand health trends, customer preferences, and market competition in order to create tailored insurance offers that meet customer needs. Additionally, the product should enable the company to identify high-value customers and reward loyalty, further enhancing customer retention. By leveraging Big Data analytics, the company can unlock hidden opportunities within its vast data repositories to drive revenue growth, optimize pricing strategies, and improve overall customer satisfaction.

This product is designed for the company’s decision-makers, including executives, data analysts, data scientists and marketing teams, who require deeper insights into customer behavior and market trends. It should also benefit customers, as the company will be able to offer more personalized and relevant insurance products that cater to their specific needs.

This will be a new product using Big Data ecosystem that has not been previously implemented. The data pipelines that are being developed as part of this project are new capabilities that will allow the company to harness the power of large-scale data analytics to gain actionable insights and drive business growth. It is a standalone project that will serve as a foundational component of the company’s future data-driven decision-making strategies. While it might interact with existing systems such as customer databases and CRM tools, the data pipelines and Big Data ecosystem are being built from scratch. Moreover, this product will integrate with the company’s existing IT infrastructure, including its CRM system, insurance claim databases, and customer data platforms. The insights generated from the data pipelines will feed into other tools used by the marketing, finance, and customer service departments, helping them execute personalized marketing campaigns, optimize pricing strategies, and improve customer service. Furthermore, the data pipeline will pull in data from external sources, including web scraping tools and third-party data providers, to create a comprehensive view of the market landscape.

### 2.1 User Needs

The "Healthcare Insurance Competitor Data Analysis Pipeline for Customer Attraction and Revenue Growth" is designed to meet the needs of various stakeholders within the healthcare insurance company. Each group of users will interact with the product in unique ways to achieve the company's overarching goals of improving revenue and customer satisfaction.

#### 1. Executives

* **Needs:** Executives require high-level insights that drive strategic decision-making. They need to understand broad trends in customer behavior, health patterns, and competitive positioning to make informed decisions about new product offerings, pricing strategies, and market expansion.
* **How They Will Use the Product:** Executives will use dashboards and reports generated from the data pipeline to review key performance indicators (KPIs) such as revenue growth, customer retention rates, and competitive standing. They will use these insights to make data-driven decisions that align with the company’s long-term goals.

#### 2. Data Analysts

* **Needs:** Data analysts need detailed and granular data to identify patterns, trends, and anomalies. They require a system that enables them to manipulate large datasets, perform complex queries, and generate reports that support business strategy.
* **How They Will Use the Product:** Data analysts will interact directly with the data pipeline to extract and transform data from various sources. They will run analyses to identify top diseases by claim volume, the most profitable insurance groups, and other key metrics. Their reports will be used to guide marketing, sales, and product development teams.

#### 3. Data Scientists

* **Needs:** Data scientists require access to large datasets to build predictive models and machine learning algorithms. They need a robust pipeline that can efficiently handle the processing of complex data from multiple sources.
* **How They Will Use the Product:** Data scientists will use the data pipeline to train models that predict customer behavior, assess risk, and optimize pricing strategies. They will also use the data to develop personalized insurance offerings and loyalty programs tailored to specific customer segments.

#### 4. Marketing Teams

* **Needs:** Marketing teams need insights into customer preferences and behaviors to design effective campaigns and offers. They require segmentation data, loyalty metrics, and competitive analysis to attract and retain customers.
* **How They Will Use the Product:** Marketing teams will use insights from the pipeline to create personalized marketing campaigns. For example, they might identify young subscribers under 30 and design specific offers for them or target regions with high claim volumes with tailored insurance packages. The data will also help them understand which types of insurance (government vs. private) are more popular and adjust marketing strategies accordingly.

#### 5. Finance Teams

* **Needs:** Finance teams require insights into profitability and cost-effectiveness. They need data on revenue drivers, claim rejection rates, and high-cost claims to optimize financial performance.
* **How They Will Use the Product:** Finance teams will use the data pipeline to track profitability metrics, such as identifying the most profitable insurance group or analyzing trends in average monthly premiums. This information will help them adjust pricing models and make budgetary recommendations to the executive team.

#### 6. Customer Service Teams

* **Needs:** Customer service teams need real-time insights into customer profiles and claim histories to provide personalized and effective service. They require data on customer engagement and satisfaction levels to improve service quality.
* **How They Will Use the Product:** Customer service teams will access the data pipeline to quickly retrieve information about a customer’s insurance plan, claim history, and previous interactions with the company. This will enable them to offer tailored solutions and resolve customer issues more efficiently.

### 2.2 Assumptions and Dependencies

When designing the "Healthcare Insurance Competitor Data Analysis Pipeline for Customer Attraction and Revenue Growth," it's crucial to outline the assumptions and dependencies that will impact the project's implementation and performance. Below are the key considerations:

#### 1. Technical Assumptions:

* **Current Technology Stack:** We assume that the technologies specified for the project—PySpark, Databricks, AWS S3, Amazon Redshift, AWS EMR Studio, Jira, and GitHub—are suitable and sufficient to meet the project’s data processing, storage, and orchestration needs. We are not planning to introduce new, experimental technologies during the project’s lifecycle.
* **Scalability:** The Big Data ecosystem (PySpark, AWS S3, Amazon Redshift, etc.) will be scalable enough to handle increasing volumes of data as the company expands its customer base and data sources. We assume that the cloud infrastructure will allow dynamic scaling without significant performance degradation.
* **Data Consistency:** We assume that data coming from different sources, including web scraping and third-party providers, will be relatively clean, structured, and consistent enough to enable seamless integration into the pipeline without excessive preprocessing requirements.
* **Security Compliance:** We assume that the cloud services (AWS S3, Redshift, etc.) meet the necessary security and compliance standards (e.g., HIPAA, GDPR) for handling sensitive healthcare data. Additionally, we assume that all data handling processes comply with industry-specific regulations.

#### 2. Framework and Environment Assumptions:

* **Cloud-Based Framework:** We assume that the core of the pipeline will be cloud-based, leveraging services such as AWS S3, Amazon Redshift, and AWS EMR. On-premises infrastructure will not be required for this implementation, and any dependencies on on-premise systems will be minimal.
* **Cross-Platform Compatibility:** While most of the development and data processing will occur in the cloud, we assume that the tools and code will be platform-independent and can run in different environments (e.g., Windows, macOS, Linux) without significant modifications.
* **Development and Collaboration Tools:** We assume that Jira will be used for project management and issue tracking, and GitHub will be used for version control and code collaboration. Both tools are expected to integrate seamlessly with the development environment and support CI/CD workflows.

#### 3. Data Assumptions:

* **Data Availability:** We assume that all necessary data sources, including competitor data, customer data, and third-party data, will be accessible and available on time for pipeline development and testing. Delays or inconsistencies in data availability could impact the project timeline.
* **Data Growth:** We assume that the volume of data will grow over time, but this growth will remain manageable within the scalability limits of the chosen cloud infrastructure (AWS). Sudden, unanticipated spikes in data volume could strain the system.

#### 4. Performance Assumptions:

* **Low Latency:** We assume that the cloud infrastructure (AWS) will provide low-latency access to data and computing resources, ensuring that the pipeline processes data efficiently. Delays in cloud resource allocation or network issues could impact pipeline performance.
* **High Availability:** We assume that the cloud services used (e.g., AWS S3, Amazon Redshift) will have high availability, with minimal downtime. Any unexpected service outages could disrupt data processing and analysis.

#### 5. Integration Assumptions:

* **Seamless Integration:** We assume that the new Big Data ecosystem will integrate smoothly with existing IT infrastructure, including CRM systems, insurance claim databases, and customer data platforms. Compatibility issues with legacy systems could hinder the flow of data between the new pipeline and existing systems.
* **Third-Party Data Integration:** We assume that the integration of third-party data sources (e.g., competitor data providers) will not require extensive reformatting or transformation. If data formats vary widely, additional data engineering efforts may be needed to standardize the incoming data.

#### 6. Resource Assumptions:

* **Skilled Personnel:** We assume that the development team has the necessary expertise in PySpark, Databricks, AWS services, and other relevant technologies. If the team lacks expertise in any critical areas, additional training or external hiring may be required, potentially delaying the project.
* **Adequate Budget and Timeline:** We assume that the project is adequately funded and that the timeline allocated for development, testing, and deployment is sufficient to meet the project's requirements without cutting corners. Budgetary constraints could force compromises on technology choices or reduce the scope of the project.

#### 7. Risks and Potential Failures:

* **Data Quality Issues:** If data from various sources is inconsistent, incomplete, or of poor quality, it could lead to inaccurate insights and affect the success of the pipeline.
* **Cloud Service Limitations:** AWS services may have limitations on throughput, storage capacity, or performance that could hinder the pipeline's ability to process large datasets efficiently.
* **Security Breaches:** If the system is not properly secured, there is a risk of data breaches, particularly given the sensitive nature of healthcare data.
* **Integration Failures:** If the pipeline fails to integrate smoothly with existing systems, it could disrupt business operations and lead to project delays.

By documenting and addressing these assumptions and dependencies, the project team can better anticipate challenges, mitigate risks, and ensure that the "Healthcare Insurance Competitor Data Analysis Pipeline for Customer Attraction and Revenue Growth" operates effectively and meets its objectives.

# 3. System Features and Requirements

This section outlines the key requirements for the system, categorized into functional requirements, external interface requirements, and non-functional requirements.

### 3.1 Functional Requirements

Functional requirements describe the specific functions the data pipeline must perform. These requirements focus on the features that add value to the system and provide specific functionalities.

#### 1. Data Ingestion and Processing

* The system must support data ingestion from various sources, including web scraping tools, third-party data providers, and existing internal data repositories (e.g., customer databases, insurance claim systems).
* The system must be capable of processing large volumes of data using PySpark and Databricks to clean, transform, and format the data for analysis.
* The system must support distributed data processing using AWS EMR and Databricks for scalability and efficient handling of Big Data workloads.

#### 2. Data Storage

* The system must store processed data in AWS S3 for long-term storage and in Amazon Redshift for structured querying and analysis.
* The system must ensure data is stored in a format that supports efficient querying, reporting, and analytics (e.g., Parquet, ORC, CSV).

#### 3. Data Analytics and Reporting

* The system must provide capabilities for data analytics, including statistical analysis, trend identification, and predictive modeling using PySpark and other tools integrated with Databricks.
* The system must generate reports on key business metrics, such as customer behavior patterns, claim volumes, and profitability analysis, and make them accessible through dashboards.

#### 4. Integration with Existing Tools

* The system must integrate with the company’s existing CRM system, insurance claim databases, and other internal tools to pull and push data.
* The system must provide APIs or other data interfaces to allow data sharing and integration with external tools used by marketing, finance, and customer service teams.

#### 5. Data Pipeline Orchestration

* The system must support workflow orchestration using AWS EMR Studio to manage the end-to-end data pipeline, including scheduling, monitoring, and logging.
* The system must integrate with Jira for issue tracking and project management, ensuring that data pipeline development and maintenance tasks are properly managed.

#### 6. Version Control

* The system must use GitHub for version control and collaboration among developers, ensuring that all code changes are tracked and managed in a collaborative manner.

### 3.2 External Interface Requirements

External interface requirements describe how the system interacts with other tools, hardware, software, and users. These requirements ensure that the system integrates seamlessly into the broader IT ecosystem.

#### 1. User Interface

* The system must provide a user interface for data analysts, data scientists, and marketing teams to access reports, dashboards, and analytics results.
* The user interface must support role-based access control, ensuring that different user roles have appropriate levels of access to data and functionalities.

#### 2. Hardware

* The system must be compatible with standard cloud infrastructure hardware provided by AWS. There should be no dependency on specific on-premises hardware.

#### 3. Software

* The system must integrate with AWS services (S3, Redshift, EMR) and third-party software tools (e.g., Jira, GitHub) for seamless operation.
* The system must support API integration with external tools for data ingestion and export.

#### 4. Communication

* The system must support secure communication protocols (e.g., HTTPS, SSL/TLS) to ensure data security during transmission between the system and external/internal tools.
* The system must support message-based communication between components using cloud-native messaging services (e.g., Amazon SQS, SNS) as needed.

### 3.3 Non-Functional Requirements

Non-functional requirements describe the qualities and attributes the system must have, including performance, safety, security, usability, and scalability.

#### 1. Performance

* The system must handle large-scale data processing efficiently, ensuring that data ingestion, processing, and storage can occur within defined SLAs (e.g., processing 1 TB of data within 2 hours).
* The system must support real-time or near-real-time data processing where applicable, particularly for time-sensitive business insights.

#### 2. Safety

* The system must ensure data integrity, preventing any data corruption or loss during processing and storage.
* The system must have failover and disaster recovery mechanisms in place, ensuring business continuity in the event of system failures.

#### 3 Security

* The system must comply with industry-specific security standards (e.g., HIPAA for healthcare data) to protect sensitive data, including encryption at rest and in transit.
* The system must enforce strict access controls and audit logging to track user actions and ensure that only authorized personnel can access and modify data.
* The system must undergo regular security assessments and vulnerability testing to identify and address potential security risks.

#### 4. Usability

* The system must provide a user-friendly interface that allows non-technical users (e.g., marketing teams) to access data insights without needing to write code or perform complex queries.
* The system must provide comprehensive documentation and training materials to support users in understanding how to interact with the data pipeline, analytics tools, and reporting features.

#### 5. Scalability

* The system must scale horizontally to handle increasing data volumes as the company grows. This includes scaling both data storage (e.g., AWS S3, Redshift) and data processing (e.g., EMR, Databricks) resources dynamically based on demand.
* The system architecture must be designed to accommodate additional data sources and new types of analysis as business needs evolve, without requiring major overhauls to the core infrastructure.

By outlining these requirements, the project team can ensure that the project is designed to meet the needs of all stakeholders while also addressing critical technical, security, and performance concerns.

# 