Template - Requirements Specifications Document

# Introduction

**a. Purpose**

The purpose of this Software Requirements Specification (SRS) document is to outline the functional and non-functional requirements for developing data pipelines for a healthcare insurance company using Azure and Databricks. The goal of these pipelines is to analyze customer behavior, calculate royalties, and provide insights to enhance the company’s revenue. This document will serve as a reference throughout the project lifecycle to ensure that all requirements are clearly defined, understood, and met.

**b. Intended Audience and Use**

This SRS is intended for the following audiences within the organization:

* **Developers:** To guide the development of the data pipelines by providing a clear set of functional and non-functional requirements.
* **Testers:** To develop test cases and validate the correctness and performance of the data pipelines based on the defined requirements.
* **Project Managers:** To monitor the progress, ensure alignment with business goals, and manage resources effectively.
* **Data Analysts:** To understand the expected outcomes from the data pipelines and how the data will be processed for analysis.
* **Stakeholders (Business Analysts, Product Owners):** To ensure that the project aligns with business objectives and provides the desired outcomes.

**c. Product Scope**

The scope of this project includes the development of a comprehensive data pipeline system to process and analyze healthcare insurance data. The key benefits, objectives, and goals include:

* **Benefits:**
  + Improved customer insights to offer tailored insurance policies.
  + Enhanced decision-making through detailed analysis of customer behavior.
  + Increased revenue by identifying and targeting profitable customer segments.
* **Objectives:**
  + To create reliable and scalable data pipelines using Azure and Databricks.
  + To clean, transform, and analyze large volumes of healthcare data from multiple sources.
  + To generate actionable insights that align with the company's revenue-enhancement strategies.
* **Goals:**
  + Implement automated data processing workflows.
  + Deliver accurate reports and dashboards on customer behavior and claims data.
  + Facilitate quick and informed decision-making by providing real-time data insights.

**d. Definitions and Acronyms**

* **SRS:** Software Requirements Specification
* **Azure:** A cloud computing service created by Microsoft for building, testing, deploying, and managing applications and services through Microsoft-managed data centers.
* **Databricks:** A cloud-based data engineering and data science platform, which provides tools for processing and analyzing large datasets using Apache Spark.
* **ETL:** Extract, Transform, Load - a data integration process that involves extracting data from source systems, transforming it into a usable format, and loading it into a data warehouse.
* **Healthcare Insurance:** A type of insurance coverage that pays for medical and surgical expenses incurred by the insured.
* **Pipeline:** A set of data processing steps organized to take raw data from sources and transform it into usable insights.
* **ROI:** Return on Investment - a performance measure used to evaluate the efficiency or profitability of an investment.
* **Claims Data:** Information related to the requests made by insured individuals to the insurance company for payment of healthcare services.
* **Subscriber:** An individual who is enrolled in a health insurance plan.
* **PBM:** Pharmacy Benefit Management - a service that manages prescription drug benefits on behalf of health insurers.
* **API:** Application Programming Interface - a set of routines, protocols, and tools for building software and applications.
* **KPI:** Key Performance Indicator - a measurable value that demonstrates how effectively a company is achieving key business objectives.

# Overall Description –

The overall description section provides a detailed understanding of the product being developed, its necessity, and its intended use. It ensures that all team members are aligned with the project’s objectives, making the development process more efficient.

**Why is this Product Needed?**

The healthcare insurance company is facing challenges in enhancing its revenue and understanding customer behavior. To address these challenges, the company needs a robust data processing system that can analyze large volumes of healthcare data from various sources. The insights derived from this analysis will help the company customize insurance offers, calculate royalties, and develop targeted marketing strategies to improve customer engagement and revenue.

**Who is this Product For?**

This product is designed for the healthcare insurance company's internal teams, including data analysts, business strategists, and marketing professionals. It will also be used by IT teams responsible for maintaining and scaling the data infrastructure. The product will enable these teams to derive actionable insights from data, allowing them to make informed decisions that align with the company’s revenue goals.

**Is it a New Product?**

Yes, this is a new product being developed to address specific business needs of the healthcare insurance company. It involves building data pipelines from scratch using Azure and Databricks, tailored to process and analyze healthcare data.

**Is it an Add-On to a Product You’ve Already Created?**

No, this is not an add-on but a standalone system designed to integrate with the company’s existing data infrastructure. It will work alongside other tools and systems within the company’s ecosystem but is a distinct project with its own scope and objectives.

**Is this Going to Integrate with Another Product?**

Yes, the data pipelines will integrate with existing data storage, reporting, and business intelligence tools used by the company. This integration ensures that the processed data and insights are accessible to relevant stakeholders across different platforms.

**a. User Needs**

The primary users of this product are:

* **Data Analysts:** They need the system to process raw healthcare data into clean, structured datasets for analysis. The system must allow them to generate reports, visualize data, and identify trends related to customer behavior, claims, and insurance policies.
* **Business Strategists:** They require insights derived from data to craft targeted marketing campaigns, develop new insurance products, and optimize existing offerings. The system should provide actionable insights on customer demographics, claims patterns, and profitability.
* **Marketing Professionals:** They need data-driven insights to design personalized insurance offers and loyalty programs. The system should help them identify the most profitable customer segments and tailor marketing efforts accordingly.
* **IT and Data Engineering Teams:** They are responsible for maintaining and scaling the data pipelines. The system should be reliable, scalable, and easy to manage, with clear documentation and robust monitoring tools.

**b. Assumptions and Dependencies**

**Assumptions:**

* **Data Availability:** It is assumed that the required healthcare data from various sources (scraping, third-party) will be readily available and accessible for ingestion into the data pipelines.
* **Cloud Infrastructure:** The development assumes that the healthcare insurance company has a functioning Azure environment where Databricks can be deployed.
* **Data Security:** The system will comply with all relevant healthcare data privacy regulations, including HIPAA, and it is assumed that the company’s existing security measures are adequate for protecting the processed data.
* **User Proficiency:** It is assumed that users (data analysts, business strategists) have a basic understanding of data analysis tools and are capable of working with the outputs generated by the data pipelines.

**Dependencies:**

* **Azure Services:** The system depends on Azure services such as Azure Data Lake Storage, Azure Databricks, and Azure SQL Database for data storage, processing, and analysis.
* **Third-Party Data Providers:** The project is dependent on the timely and accurate delivery of data from third-party sources.
* **Existing IT Infrastructure:** The new data pipelines will need to integrate smoothly with the existing IT infrastructure, including data warehouses, business intelligence tools, and reporting systems.
* **Regulatory Compliance:** The project’s success depends on adherence to healthcare regulations and standards, which may impact the design and operation of the data pipelines.

# System Features and Requirements

In this section, we will break down the specific features and requirements necessary for the successful development of the healthcare data pipeline system. These requirements will guide the development team in creating a robust and functional product that meets business goals and user needs.

**a. Functional Requirements**

Functional requirements define the specific behavior and functions that the system must perform. For this healthcare data pipeline project, the key functional requirements include:

* **Data Ingestion:**
  + The system must ingest data from various sources, including third-party APIs, web scraping, and internal databases.
  + Data ingestion should support batch processing and real-time streaming to handle different data arrival patterns.
* **Data Cleaning and Transformation:**
  + The system must clean and transform raw data into a structured format suitable for analysis.
  + Data cleaning should include removing duplicates, handling missing values, and normalizing data formats.
  + Transformation tasks should include aggregating data, calculating new fields, and converting data types as required.
* **Data Storage:**
  + The cleaned and transformed data should be stored in a scalable data lake (Azure Data Lake Storage) for further processing and analysis.
  + The system should also support the creation of data warehouses or databases for reporting and business intelligence (e.g., Azure SQL Database).
* **Data Analysis and Reporting:**
  + The system must support the execution of analytical queries to extract insights, such as identifying trends in customer behavior and calculating profitability.
  + The system should integrate with BI tools (e.g., Power BI) to generate reports and dashboards based on the processed data.
* **Pipeline Orchestration:**
  + The system should automate the orchestration of data pipelines using Azure Data Factory or Databricks Workflows.
  + Pipelines must be scheduled to run at predefined intervals or triggered by specific events.
* **Security and Compliance:**
  + The system must ensure data privacy and comply with healthcare regulations such as HIPAA.
  + Access controls and encryption must be implemented to protect sensitive data.

**b. External Interface Requirements**

These requirements outline how the system will interact with external tools and systems. This includes interactions with users, hardware, software, and communication interfaces.

* **i. User Interface:**
  + The system must provide a user-friendly interface for data analysts and business users to access reports and dashboards.
  + The UI should support querying datasets, visualizing trends, and exporting reports.
* **ii. Hardware Interface:**
  + The system must be compatible with the company’s existing hardware infrastructure, including on-premises servers and cloud-based resources.
  + The system should efficiently utilize available computing resources to ensure optimal performance.
* **iii. Software Interface:**
  + The system should integrate with existing software tools used by the company, such as Azure services, BI tools, and databases.
  + APIs must be developed to facilitate data exchange between the data pipeline system and other software components.
* **iv. Communication Interface:**
  + The system must support secure communication protocols (e.g., HTTPS, SSL/TLS) for data transfer between different components.
  + Notifications and alerts should be sent via email or messaging services to inform stakeholders of pipeline status, errors, or completion.

**c. System Features**

System features are a type of functional requirement essential for the system's operation. The key features for this healthcare data pipeline project include:

* **Automated Data Pipelines:**
  + The system must automate the entire ETL (Extract, Transform, Load) process, from data ingestion to storage and analysis.
  + Pipelines should be resilient to failures and capable of automatic recovery and retries.
* **Real-Time Data Processing:**
  + The system must support real-time data processing capabilities to handle streaming data from sources such as IoT devices or real-time APIs.
  + Real-time analytics should be available for immediate decision-making.
* **Scalable Data Storage:**
  + The system must provide scalable storage solutions capable of handling large volumes of healthcare data.
  + Storage solutions should support both structured and unstructured data formats.
* **Customizable Reporting:**
  + The system must enable users to create custom reports and dashboards tailored to specific business needs.
  + Reports should be exportable in various formats (e.g., PDF, Excel).
* **Compliance and Audit Trail:**
  + The system must maintain an audit trail of data processing activities to ensure compliance with regulatory requirements.
  + All actions performed on data should be logged and traceable.

**d. Nonfunctional Requirements**

Nonfunctional requirements ensure that the system will work as expected by users and stakeholders. These include performance, safety, security, usability, and scalability requirements.

* **i. Performance Requirements:**
  + The system must process large datasets efficiently, with minimal latency, especially during peak data ingestion periods.
  + Data queries should return results within acceptable timeframes, typically within seconds for most queries.
* **ii. Safety Requirements:**
  + The system must include fail-safe mechanisms to prevent data loss or corruption in the event of hardware or software failures.
  + Regular backups and data replication strategies should be implemented to ensure data integrity.
* **iii. Security Requirements:**
  + The system must enforce strict access controls, ensuring that only authorized users can access sensitive data.
  + All data in transit and at rest must be encrypted to protect against unauthorized access.
* **iv. Usability Requirements:**
  + The system’s user interface must be intuitive and easy to navigate, allowing users to access data and reports with minimal training.
  + User documentation and help resources should be provided to assist users in troubleshooting common issues.
* **v. Scalability Requirements:**
  + The system must be able to scale horizontally and vertically to accommodate increasing data volumes and user demands.
  + The architecture should support the addition of new data sources and processing capabilities without significant rework.