Template - Requirements Specifications Document

# **Introduction** -

## **Purpose** - The purpose of this project is to design and implement a scalable, data-driven analytics pipeline for a health care insurance company using the AWS Big Data ecosystem. The goal is to analyze competitor and customer data to generate actionable insights that will enable the company to Enhance Revenue, Understand Customer Behavior, Improve Decision, Optimize Operations, Enhance Data Accuracy.

* 1. **Intended Audience and Use** - This document is intended for the following audience within the health care insurance company:

1. ***Developers*:**
   * Responsible for implementing the data ingestion, cleaning, transformation, and analysis pipelines.
   * Will use this document to understand the system requirements and functional specifications.
2. ***Data Engineers*:**
   * Responsible for managing the data infrastructure on AWS, such as S3, Redshift, and EMR.
   * Will refer to this document to ensure the proper setup of data ingestion, transformation, and loading processes.
3. ***Data Scientists/Analysts*:**
   * Will analyze the data to derive actionable insights related to claims, policies, and customer behavior.
   * This document will help them understand the data flow and structure to access the necessary datasets.
4. ***Project Managers*:**
   * Overseeing the project implementation and ensuring that deadlines and business objectives are met.
   * This document serves as a guide for tracking deliverables and project progress.
5. ***Testers*:**
   * Responsible for validating the functionalities and ensuring that the system meets the defined requirements.
   * This document provides a detailed understanding of expected outcomes and use cases to ensure comprehensive testing.

## **Product Scope** - The scope of this project is to build a comprehensive data pipeline and analysis system that will enable a health care insurance company to Enhance Revenue, Customer Behavior Insights, Data Management and Scalability, Business Intelligence Reporting

* 1. **Definitions and Acronyms** - 
* **AWS (Amazon Web Services)**: A comprehensive and widely adopted cloud platform that offers various services such as computing power, storage, and analytics.

 **S3 (Simple Storage Service)**: AWS storage service used to store and retrieve any amount of data from anywhere on the web. Used to store raw, processed, and cleaned data.

 **EMR (Elastic MapReduce)**: A cloud-based big data platform on AWS for processing large amounts of data using Apache Hadoop, Apache Spark, and other big data frameworks.

 **Redshift**: A fully managed, petabyte-scale data warehouse service in the cloud. Used for analytics and query processing.

 **Databricks**: A cloud platform built for big data and AI workloads, often used in conjunction with Apache Spark for data processing and machine learning.

 **Jira**: A project management tool used to track tasks, bugs, and progress, commonly used for managing software development projects.

 **GitHub**: A web-based platform used for version control and collaboration, allowing multiple users to work on projects simultaneously.

 **ETL (Extract, Transform, Load)**: A process used in data warehousing where data is extracted from source systems, transformed into a format suitable for analysis, and loaded into a data warehouse.

 **Data Pipeline**: A series of processes that move data from one system to another, often used to ingest, clean, and process data.

 **Schema**: The structure that defines the organization of data within a database. It specifies tables, fields, and relationships between fields in a relational database.

## **Overall Description** - This project aims to build a data pipeline and analytics solution for a Health Care Insurance company using the AWS ecosystem. The company seeks to enhance its revenue by analyzing customer behavior and competitor data, which it receives from a variety of sources such as web scraping and third-party vendors. The insights gathered from this analysis will help the company optimize its offers, increase policy sales, and calculate royalties for past policyholders.

* 1. **User Needs** - The users of this system will include:
* **Business Analysts**: They require easy access to analytics-ready data stored in Redshift to generate reports on customer behavior, revenue trends, and policy performance.
* **Data Scientists**: They will utilize cleaned and pre-processed data for advanced modeling and predictions to further personalize customer insurance offers.
* **Data Engineers**: They are responsible for building and maintaining the data pipeline, ensuring that raw data from different sources is ingested, cleaned, and made available for analysis.
  1. **Assumptions and Dependencies** –

 **Assumptions**:

* The data sources (competitor data, customer behavior data, etc.) are regularly updated and accessible via APIs or third-party services.
* The AWS environment will have the necessary permissions and resources (e.g., S3 buckets, Redshift cluster, EMR) to support data ingestion, storage, and processing.
* The data schema is relatively stable, and any significant changes will be communicated to the data engineering team in advance.

 **Dependencies**:

* **AWS Services**: This solution depends heavily on AWS services like S3, Redshift, EMR, and IAM for access management.
* **Databricks**: Used for developing and testing the data processing pipeline. If there are issues with the Databricks platform or integrations, they could impact the overall data pipeline.
* ***Pyspark***: Used for data cleaning and processing. Any issues in the Pyspark environment or code execution can affect the data quality.
* ***Jira and GitHub***: Tools for project management and version control will be necessary to ensure smooth collaboration and codebase tracking.

## **System Features and Requirements** - The system for the Health Care Insurance Company needs to meet a variety of functional and non-functional requirements to help the organization analyze competitor data and customer behavior. These requirements focus on building a robust, scalable, and secure data pipeline that allows for seamless data processing, analysis, and reporting.

* 1. **Functional Requirements** –

 **Data Ingestion**

* **Requirement**: The system should ingest raw data from various sources (competitor data, customer behavior data, etc.) into AWS S3.
* **Source**: Data could come from APIs, CSV files, or third-party data providers.
* **Tool**: AWS S3 for raw data storage, integrated with Databricks for data ingestion.

 **Data Cleaning and Transformation**

* **Requirement**: The system must clean and transform raw data into a format suitable for analysis.
* **Action**: Remove null values, eliminate duplicates, normalize data formats, and categorize information.
* **Tool**: Pyspark running on AWS EMR or Databricks.

 **Data Storage**

* **Requirement**: The system should store cleaned and processed data in Redshift for querying and analysis.
* **Structure**: Data should be organized into Redshift tables with appropriate schemas for easy querying.
* **Tool**: AWS Redshift for storage.

 **Data Analysis and Queries**

* **Requirement**: Business analysts and data scientists should be able to run SQL queries on the data stored in Redshift.
* **Action**: Provide predefined queries to retrieve insights such as customer trends, most profitable groups, rejected claims, and policy performance.
* **Tool**: AWS Redshift and tools like AWS QuickSight for dashboarding.

 **Data Pipeline Automation**

* **Requirement**: The data pipeline must be automated to run at regular intervals or trigger-based events.
* **Action**: The system should support the automatic ingestion, cleaning, and transformation of data using workflows and triggers.
* **Tool**: AWS Lambda, Step Functions, and Databricks jobs for pipeline orchestration.

 **Data Security**

* **Requirement**: All data stored in S3 and Redshift should be encrypted, and user access should be restricted based on roles.
* **Tool**: AWS IAM for access control, KMS for encryption, and VPC to secure the environment.
  1. **External Interface Requirements** –

 **User** :

* **Tool**: AWS Management Console and Databricks Notebooks will serve as the primary interfaces for data engineers and developers.
* **Access**: Users should be able to interact with these tools for monitoring, debugging, and managing the data pipeline.

 **Hardware** :

* **AWS Infrastructure**: The system will leverage AWS cloud resources (S3, Redshift, EMR) with no direct hardware dependency on the user side.

 **Software** :

* **Integration**: The system should integrate with third-party APIs to ingest competitor data and customer data. It should also interact with Jira for project management and GitHub for version control.
* **Data Connectors**: Use AWS Glue for data discovery and cataloging and JDBC for connecting to Redshift from Databricks.

 **Communication** :

* **Tool**: Secure data transfers between S3, EMR, and Redshift using HTTPS and AWS SDKs.
* **Monitoring**: Use AWS CloudWatch for tracking pipeline health and performance.
  1. **System Features** -  **Competitor Data Analysis**
* **Feature**: Analyze competitor data to track customer behavior and recommend personalized insurance offers.

 **Customer Behavior Analysis**

* **Feature**: Analyze customer data to identify trends in policy subscriptions, claims, and revenue generation.

 **Royalties Calculation**

* **Feature**: Calculate royalties for customers who bought insurance policies in the past based on predefined rules.
  1. **Nonfunctional Requirements** -

 **Performance Requirements**:

* **Scalability**: The system must handle large datasets (both historical and real-time) without performance degradation.
* **Latency**: The end-to-end data pipeline should complete processing within an acceptable timeframe (e.g., less than 30 minutes per batch).

 **Safety Requirements**:

* **Data Protection**: Ensure that all sensitive data (e.g., customer information, claims) is encrypted in transit and at rest to comply with healthcare data regulations like HIPAA.

 **Security Requirements**:

* **Access Control**: Implement AWS IAM roles and policies to restrict access to different components of the system.
* **Audit Logs**: All data access and changes must be logged and auditable using AWS CloudTrail.

 **Usability Requirements**:

* **Ease of Use**: Business analysts and data scientists should easily query and retrieve data from Redshift using standard SQL interfaces.
* **Documentation**: Comprehensive documentation should be provided for users to interact with the system effectively.

 **Scalability Requirements**:

* **Elasticity**: The system should scale up and down based on the volume of data being processed. AWS services like EMR and Redshift should automatically adjust their resources as needed.
* **Modular Architecture**: The system should be designed in a modular fashion, so individual components (e.g., data ingestion, transformation) can be scaled independently.