**Contents**

[a) Purpose 1](#_Toc175085751)

[b) Intended Audience and Use 1](#_Toc175085752)

[c) Product Scope 3](#_Toc175085753)

[d) Definitions and Acronyms 4](#_Toc175085754)

[Overall Description 4](#_Toc175085755)

[User Needs 4](#_Toc175085756)

[Assumptions and Dependencies 5](#_Toc175085757)

[System Features and Requirements 5](#_Toc175085758)

[Functional Requirements 5](#_Toc175085759)

[External Interface Requirements 5](#_Toc175085760)

[ User 5](#_Toc175085761)

[ Hardware 5](#_Toc175085762)

[ Software 5](#_Toc175085763)

[ Communications 5](#_Toc175085764)

[System Features 5](#_Toc175085765)

[Nonfunctional Requirements 5](#_Toc175085766)

[ Performance requirements 5](#_Toc175085767)

[ Safety requirements 5](#_Toc175085768)

[ Security requirements 5](#_Toc175085769)

[ Usability requirements 5](#_Toc175085770)

[ Scalability requirements 5](#_Toc175085771)

# Introduction:

In the highly competitive health care insurance industry, understanding customer behavior and optimizing revenue streams are critical to maintaining a competitive edge. The integration of Big Data analytics provides an opportunity to gain valuable insights into customer behavior, competitor strategies, and market trends. By analyzing data from various sources, including competitor data, the company can develop customized insurance offers, optimize existing policies, and accurately calculate royalties for past customers. This project aims to enhance the company’s revenue and customer engagement by leveraging the power of Big Data.

## Purpose

The primary purpose of this project is to design and implement data pipelines that will enable the company to develop effective business strategies. These strategies will be informed by in-depth analysis of customer behavior, optimization of insurance offers, and accurate calculation of royalties. The project seeks to harness the potential of Big Data analytics to:

* Track customer behavior and conditions to tailor insurance policies.
* Analyze competitor data to identify trends and opportunities.
* Calculate royalties for past customers based on historical data.
* Ultimately, enhance revenue through data-driven decision-making.

## Intended Audience and Use

This document is intended for a diverse audience, including:

* **Stakeholders:** Executives, project sponsors, and decision-makers who need a clear understanding of the project's objectives, scope, and expected outcomes.
* **Data Analysts:** Professionals responsible for designing, implementing, and maintaining the data pipelines and analytics processes.
* **Business Strategists:** Individuals tasked with interpreting the data insights to formulate and execute business strategies aimed at revenue enhancement and customer satisfaction.

**How the Audience Will Use This Document:**

* **Stakeholders** will use the document to align the project goals with the company’s broader strategic objectives and to monitor progress.
* **Data Analysts** will reference the document to ensure their work aligns with the project’s goals and scope, particularly in developing data models and algorithms.
* **Business Strategists** will leverage the insights provided to shape marketing campaigns, policy offers, and other revenue-generating initiatives.

## Product Scope

The scope of this project encompasses several key areas critical to achieving the desired outcomes. These areas include:

1. **Identifying the Disease with the Maximum Number of Claims:**
   * Analyze claims data to determine which diseases are most frequently claimed.
   * Use insights to adjust policy offerings and pricing strategies.
2. **Analyzing Subscriber Demographics:**
   * Segment the subscriber base by age, location, income, and other relevant demographics.
   * Understand the needs and behaviors of different demographic groups to tailor insurance products accordingly.
3. **Evaluating the Profitability of Different Policy Groups:**
   * Assess the financial performance of various insurance policy groups.
   * Identify underperforming policies and recommend adjustments to improve profitability.
4. **Tracking Claims Based on Various Criteria:**
   * Monitor claims data based on factors such as time, location, and disease type.
   * Use this information to detect patterns, predict future claims, and optimize risk management strategies.

## Definitions and Acronyms

This section provides definitions and explanations for key terms and acronyms used throughout the document:

* **Big Data Ecosystem:** The complex infrastructure of data collection, storage, processing, and analysis that supports large-scale data analytics.
* **ROI (Return on Investment):** A performance measure used to evaluate the efficiency of an investment or compare the efficiency of several investments.
* **Subscriber:** An individual who holds an insurance policy with the company.
* **Subgroup:** A specific segment of the subscriber base, categorized by factors such as demographics or policy type.
* **Claim:** A request made by the policyholder for payment of the benefits provided by the insurance policy.
* **Data Pipeline:** A series of data processing steps, often involving extraction, transformation, and loading (ETL) of data from various sources into a central repository for analysis.

# Overall Description

This project aims to build a sophisticated Big Data analytics platform specifically designed for a health care insurance company. The product will be a comprehensive solution that leverages large-scale data analysis to understand customer behavior, monitor market trends, and optimize insurance offerings. It will also facilitate the accurate calculation of royalties for past customers.

The platform is not a standalone product but an add-on to the company's existing data infrastructure. It will integrate seamlessly with existing tools and databases, enriching the company’s analytical capabilities and allowing for more data-driven decision-making. This product is necessary because the company needs a more advanced way to analyze customer data, stay competitive in the insurance market, and boost revenue by offering more personalized and strategically priced insurance policies.

## User Needs

The product will be used by various stakeholders within the company, including:

1. Data Analysts:

* Need: Tools to ingest, clean, and analyze large datasets from multiple sources, including competitor data.
* Use: Data analysts will use the platform to develop models that predict customer behavior, identify market trends, and optimize pricing strategies.

1. Business Strategists:

* Need: Actionable insights derived from data analysis to inform decision-making.
* Use: Strategists will use the analytics platform to refine marketing strategies, tailor insurance policies, and enhance customer engagement.

1. Stakeholders (Executives and Managers):

* Need: High-level dashboards and reports summarizing key metrics and trends.
* Use: Stakeholders will utilize the platform to monitor the company’s performance, track ROI, and guide the overall direction of business initiatives.

## Assumptions and Dependencies

**Assumptions:**

* The platform will operate on the company’s existing IT infrastructure, which includes a mix of on-premise and cloud-based solutions.
* The data sources, including competitor data, are assumed to be accessible, and legal considerations for data usage have been addressed.
* The end-users are familiar with the basic principles of data analysis and interpretation, reducing the need for extensive training on the new platform.
* Current data processing technologies, such as Hadoop and Spark, are assumed to be sufficient for handling the scale and complexity of the data.

**Dependencies:**

* Technology: The project depends on the availability and performance of the company’s current technology stack, including databases, data warehouses, and analytical tools.
* Data Sources: Reliable and continuous access to both internal and external data sources is crucial for the platform’s success.
* Third-Party Tools: The project may rely on third-party tools or services for data scraping, storage, or processing, and any changes in these services could impact the platform’s functionality.
* Compliance and Legal Factors: The platform must comply with relevant data protection regulations and insurance industry standards.

# System Features and Requirements

## Functional Requirements

**Data Ingestion:**

* The platform must be capable of ingesting data from various sources, including competitor data, customer records, and claims databases.
* It should support real-time data processing and batch processing to accommodate different types of data.

**Data Processing and Analysis:**

* The system should offer tools for cleaning, transforming, and analyzing data.
* It must support predictive modeling to anticipate customer behavior and market trends.
* The platform should allow for segmentation of data by various criteria (e.g., demographics, claim history).

**Reporting and Visualization:**

* The platform should include a robust reporting module that allows users to generate customized reports.
* Visualization tools should be available to help users understand complex data patterns and trends.

**Royalty Calculation Module:**

* The system must include a module for calculating royalties for past customers based on historical data and predefined rules.
* This module should be flexible to accommodate different calculation methods as needed by the business.

## External Interface Requirements

### User

* The platform should have an intuitive user interface that caters to both technical and non-technical users.
* Customizable dashboards should be available to meet the specific needs of different user groups.

### Hardware

* The platform should be compatible with the company’s existing servers and storage systems.
* It must be scalable to handle increased data volume without requiring significant hardware upgrades.

### Software

* The system must integrate with existing data management tools (e.g., SQL databases, data lakes).
* It should support APIs for data import/export with other systems within the organization.

### Communications

* The platform must support secure data transfer protocols to ensure data integrity and compliance with regulations.
* It should be capable of sending automated alerts and notifications based on predefined triggers (e.g., threshold breaches, anomaly detection).

## System Features

* **Customizable Data Pipelines:**
  + Users should be able to define and modify data pipelines based on specific business needs, allowing for flexibility in data processing.
* **Real-time Analytics:**
  + The platform should support real-time analytics, enabling the company to respond quickly to emerging trends or issues.
* **Scalability:**
  + The system must be designed to scale seamlessly with the growing volume of data and users, ensuring consistent performance.

## Nonfunctional Requirements

### Performance requirements

* + The platform should process large datasets efficiently, with minimal latency in data ingestion and analysis.
  + It must handle concurrent users without performance degradation.

### Safety requirements

* + The system should include failover mechanisms to ensure continuous operation in case of hardware or software failures.

### Security requirements

* + Data encryption should be employed both at rest and in transit to protect sensitive information.
  + Access control mechanisms must be in place to ensure that only authorized users can access or modify data.

### Usability requirements

* + The user interface should be intuitive and easy to navigate, reducing the learning curve for new users.
  + Context-sensitive help and documentation should be available within the platform.

### Scalability requirements

* + The platform architecture should support horizontal scaling to accommodate increasing data volume and processing demands.
  + It should be easy to add new data sources or processing nodes without significant reconfiguration.