Health Care Insurance Data Analysis (Requirements Specifications Document)

Introduction

Purpose:

The purpose of the document is to define the requirements and scope of implementing a Big Data solution for a Health Care insurance company. It outlines the analysis and data pipeline creation to enhance revenue through customer behavior analysis and personalized offers.

Intended Audience and Use:

The document is intended for data engineers, analysts, and project managers who will use Databricks and PySpark to develop and deploy data pipelines. It guides them in understanding project requirements and scope for effective implementation.

Product Scope:

The project aims to use Databricks and PySpark for analyzing competitor data sourced through scraping and third-party sources. The objectives include tracking customer behavior, personalizing insurance offers, and calculating royalties to enhance revenue. In addition, the benefits include improved customer insights, operational efficiency through automation, and optimized insurance offerings.

Definitions and Acronyms:

- Databricks: Unified analytics platform for big data and AI
- PySpark: Python API for Spark
- GitHub: Version control platform

Overall Description

User Needs:

The primary users include data engineers, analysts, and business stakeholders who require insights into customer behavior and revenue enhancement strategies. The system will integrate data from various sources, perform analytics, and generate actionable insights.

Assumptions and Dependencies:

<u>Assumptions</u>

- Availability of necessary datasets
- Use of Databricks Community Edition for development and testing.
- Integration with Databricks runtime environment for data processing and analysis.

<u>Dependencies</u>

- Timely availability of sample data from data sources
- Compatibility with Databricks platform updates.

System Features and Requirements

Functional Requirements

- Data cleaning modules: Implement PySpark scripts to handle null values, duplicates, and inconsistencies in datasets (Patients-records.csv, subscriber.csv, claims.json, grpsubgrp.csv).
- Result generation modules: Develop PySpark queries to fulfill analytical requirements (e.g., disease with maximum claims, subscribers below 30 subscribing any subgroup).

External Interface Requirements

- User interfaces: Develop Databricks notebooks for data input, processing, and result visualization.
- Software Interfaces: Utilize PySpark for data processing and integrate with GitHub for version control and collaboration.

System Features

- Data pipelines: Design and implement end-to-end data pipelines using PySpark within Databricks to transform raw data into actionable insights.
- Data visualization: Utilize Databricks visualizations for presenting analytical results to stakeholders.

Nonfunctional Requirements

- Performance requirements: Optimize PySpark scripts and queries for efficient data processing and analysis.
- Security requirements: Implement data security measures within Databricks to ensure compliance with General Data Protection Regulation and other regulatory standards.
- Scalability Requirements: Design data pipelines to scale with increasing volumes of data processed by Databricks.

Conclusion

The Requirements Specifications Document outlines the framework for implementing a Big Data solution using Databricks and PySpark to enhance revenue and customer understanding for a Health Care insurance company. It ensures alignment with business objectives and technical requirements for successful project execution.