**Product Requirements Document**

**Project Name: Data Pipeline Application**

**Overview**

Company X, a conglomerate, is seeking to invest in a new application that can be marketed by the company. The application should be able to maintain a set of users and demonstrate key functionalities. The goal of this project is to build a data pipeline application that enables data storage, data cleaning, data warehouse management, data upload handling, and data visualization through a dashboard. This document outlines the requirements for the application.

**Application Description**

The data pipeline application will consist of the following components:

**2.1 Data Storage**

* The application will utilize an S3 Bucket as the data lake for storing raw data.

**2.2 Data Cleaning**

* The data cleaning can be performed using either a local Jupyter Notebook or a Colab notebook.
* The cleaning process should be clearly documented, including the steps taken and any **visualizations created.**

**2.3 Data Warehouse**

* The application will facilitate the pushing of cleaned data into a relational database (RDBMS) of choice.
* The data warehouse will act as a storage solution for processed and structured data.
* The steps for pushing and pulling data from the data warehouse should be demonstrated.

**2.4 Data Upload Handling**

* The application will include AWS Lambda functions to handle data upload.
* Option 1: A Lambda function that accepts a data payload and adds it to the S3 Bucket.
* Option 2: A Lambda function that cleans the data from the S3 bucket and adds it to the RDBMS.
* The structure of the data payload for the Lambda function can be assumed to be uniform.

**2.5 Data Visualization**

* The application will utilize Tableau or Microsoft Power BI to create a dashboard for displaying insights derived from the dataset.
* The dashboard should showcase meaningful visualizations and insights.
  1. **Website functionality**
* Communicate the data to the website.
* Allow to create instances into the database.
* Allow for querying data based on id.
* Integrate the Tableau or Microsoft Power BI dashboard into a page on our site.
* Have a signup and login page.

**Target Audience**

The target audience for this application includes data engineers, data analysts, and data scientists within Company X. The application aims to streamline data processing and analysis tasks for these professionals.

**Usefulness**

**The application provides the following benefits:**

* Efficient storage and retrieval of raw data using the S3 Bucket as a data lake.
* Streamlined data cleaning process through Jupyter Notebook or Colab.
* Centralized data warehouse management using a relational database.
* Automated data upload handling with AWS Lambda functions.
* Insightful data visualization and exploration through Tableau or Microsoft Power BI.

**Project Management**

To ensure effective project management, the following practices will be implemented:

**5.1 Daily Standup Meetings and Agile Approach**

* Daily standup meetings will be held to provide updates, discuss progress, and address any challenges.
* The project will follow an Agile methodology, with iterative development and regular feedback cycles.
* A designated scrum master will be responsible for facilitating the Agile process.

**5.2 Kanban Board**

* A kanban board will be created using free services like Jira.
* The kanban board will help visualize and track the project's progress, tasks, and priorities.

**Presentation**

A presentation on the project will be prepared to showcase the application and its key features to the board of directors. The presentation should cover the application's purpose, functionalities, and demonstrate the insights derived from the data through the dashboard.

Please review and approve this Product Requirements Document before proceeding with the development of the Data Pipeline Application.