Real-time Monitoring and Analytics System for Smart Greenhouses

1. Project Description

This project involves the development of a comprehensive, real-time data pipeline to monitor, analyze, and visualize critical environmental data from a smart greenhouse. The system will leverage a sophisticated Python-based simulator to generate high-fidelity sensor data (temperature, humidity, pH, etc.) tailored to the Egyptian climate. This data will be ingested in real-time using **Apache Kafka**. The pipeline will then process, store, and serve this data to an interactive dashboard, enabling stakeholders to make data-driven decisions for optimizing crop yield, reducing resource consumption, and implementing predictive alerts for potential issues.

2. Project Team & Role Distribution

- **Mostafa Mohamed Abdelzaher: Team Leader & Data Pipeline Architect**
    - **Responsibilities** included leading the project, coordinating tasks, and ensuring milestones were met. Responsibilities included designing the end-to-end data pipeline architecture, developing the core Kafka integration, and defining the high-level design and key metrics for the final dashboard.
- **Mahmoud Allam: Data Producer & Application Developer**
    - **Responsibilities** included developing the Python-based data producer to simulate real-time IoT sensor data. Additional responsibilities included creating a Graphical User Interface (GUI) to manage and control the data generation process.
- **Mohamed Fathallah: Python & Visualization Developer**
    - **Responsibilities** included developing core Python components for the data source. The role also involved collaborating on the development of the interactive dashboard, implementing key visualizations based on the defined design.
- **Axanii: Dat: Data Processing & ETL Engineer**
    - **Responsibilities**: The Data Processing & ETL Engineer was tasked with designing and implementing the data transformation logic. This involved

developing the real-time ETL process using Apache Spark to consume data from Kafka, apply business rules, and prepare data for the data warehouse.

- **Fatma: Database & Visualization Specialist**
    - **Responsibilities** include collaboration on database development and visualization. Responsibilities included contributing to the database schema design and **working with the team to develop the final interactive dashboard** to present real-time data and key insights.
- **Sama: Cloud Database Specialist**
    - **Responsibilities**: The Cloud Database Specialist focused on the implementation and management of the data warehouse on the cloud. This role was responsible for setting up the Azure SQL Database, implementing the database schema, ensuring data integrity, and optimizing the cloud database for performance.

3. Objectives

1. To deploy and manage a reliable data simulation script that generates realistic greenhouse sensor data and streams it to an Apache Kafka topic with a 5-second interval.
2. To implement a unified real-time data pipeline using **Apache Spark Streaming** for continuous data processing, real-time analytics, and immediate alerting.
3. To design an optimized database schema on **Azure SQL Database** to store both raw and aggregated sensor data for long-term analysis.
4. To develop an interactive **Power BI dashboard** capable of visualizing key metrics in real-time and displaying alerts for conditions that breach predefined thresholds (e.g., soil humidity < 40%).
5. To deliver a comprehensive final report detailing the system architecture, performance, and key analytical findings.

4. Tools & Technologies

- **Data Simulation & Ingestion:** Python (Pandas,Datetime,numpy,matplotlib,pyarrow), **Apache Kafka**.

- **Real-time Processing:** Apache Spark Streaming.
- **Batch Processing (ETL):** Apache Spark.
- **Data Storage: Azure SQL Database**
- **Data Visualization: Microsoft Power BI or streamlit**
- **Version Control:** Git & GitHub.

5. KPIs (Key Performance Indicators)

1. **Data Ingestion & Preprocessing (Kafka Stream & Python):**
    a. **100%** of generated messages successfully published to the Kafka topic.
    b. Data latency (from generation to Kafka topic) maintained **under 500 milliseconds**.
2. **SQL Integration & real time processing (Azure SQL):**
    a. Query accuracy: **≥ 98%** for historical trend analysis queries.
    b. Query performance: Average execution time for daily summary queries **is under 3 seconds**.
3. **Visualization & Real-time Analytics (Power BI, Apache Spark Streaming):**
    a. Dashboard data refresh latency: **< 5 seconds** from event occurrence.
    b. **100%** of critical metrics visualized (soil/air temperature, soil/air humidity, water level, pH).
4. **Presentation (Report, slide deck):**
    a. Report completeness: **100%** of all required sections delivered.
    b. Stakeholder clarity/feedback score: Achieve a score of **≥ 4.5/5** on the project's technical depth and clarity.