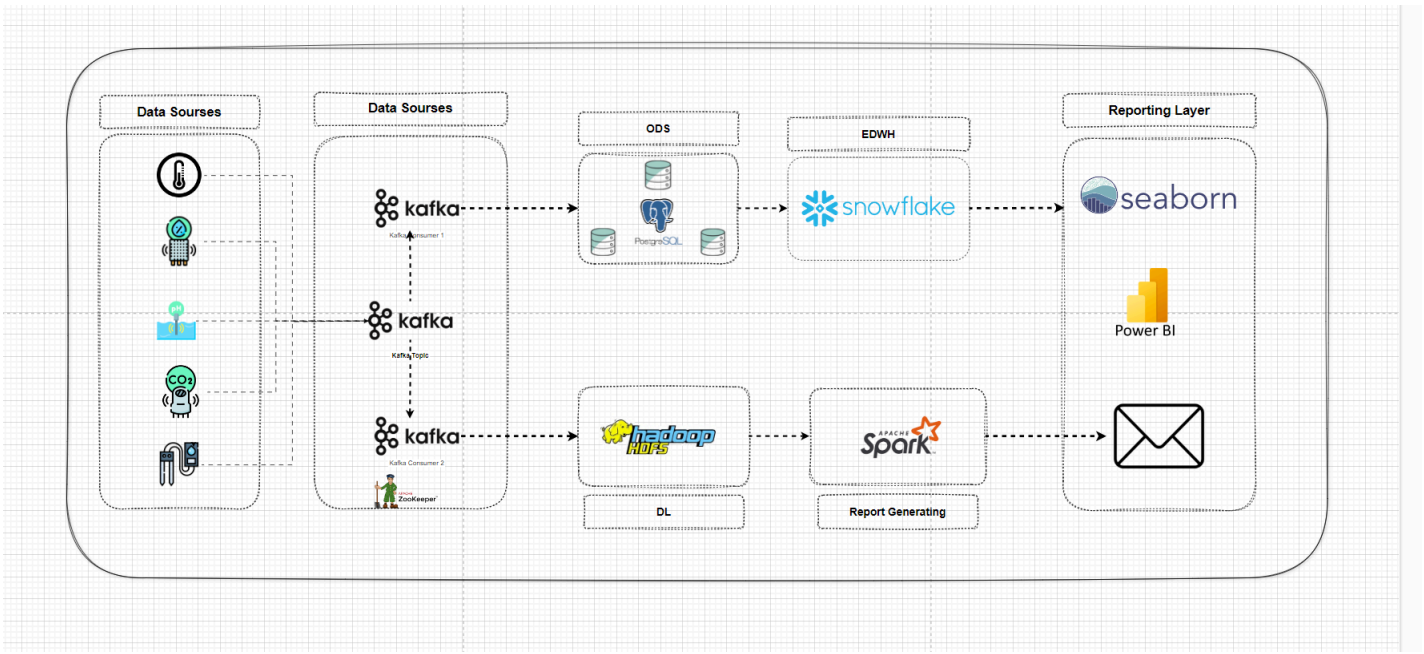
Real-Time AgriTech Data Pipeline for Smart Greenhouses

Description:

Real-Time AgriTech Data Pipeline for Smart Greenhouses powered by a real-time, end-to-end **data pipeline**.

The system simulates IoT sensor data including temperature, humidity, air quality, and soil moisture using **Python** to reflect realistic greenhouse conditions. Each sensor streams data into a unified **Apache Kafka** topic, with producers and consumers handling ingestion and initial processing. Incoming data is **validated and enriched** with metadata before being stored in an **Operational Data Store (ODS) using** PostgerSQL for real-time monitoring. For historical insights and strategic decision-making, processed data is modeled in a **star schema** and loaded into **Snowflake** (EDWH). Meanwhile, raw data is archived in **HDFS** to support batch processing. Using **Apache Spark**, I run nightly aggregation jobs that analyze daily sensor trends and generate **automated email reports** for stakeholders.

System architecture



System architecture logical components:

1. Data Sources

IoT Sensors: Devices that generate real-time data (temperature, humidity, traffic, air quality) from various city locations.

- **Input format**: JSON

```
{
  "sensor_id": string,
  "timestamp": string,
  "sensor_type": string,
  "value": float,
  "location": string
}
```

---

2. Ingestion Layer

Apache Kafka:

- Acts as the real-time data ingestion and messaging layer.
- Each sensor type writes to a dedicated Kafka topic (e.g., `temperature`, `traffic`, `air_quality`).

3. Processing Layer

Real-Time Processing:

- **Python and Kafka Consumer**:
  - Reads data from Kafka topics.
  - **Data validated and enriched** with metadata

Batch Processing:

- **Apache Spark on HDFS**:
  - Runs nightly batch jobs to aggregate data by region, sensor type, and time intervals (e.g., daily, weekly).

---

4. Storage Layer

Data Lake (Raw Data):

- **HDFS**: Stores raw IoT data for archival and future processing needs.

Data Warehouse (Processed Data):

- SnowFlake: Stores cleaned and aggregated data for analytics and dashboarding.
- Optimized for fast query performance and analytics.

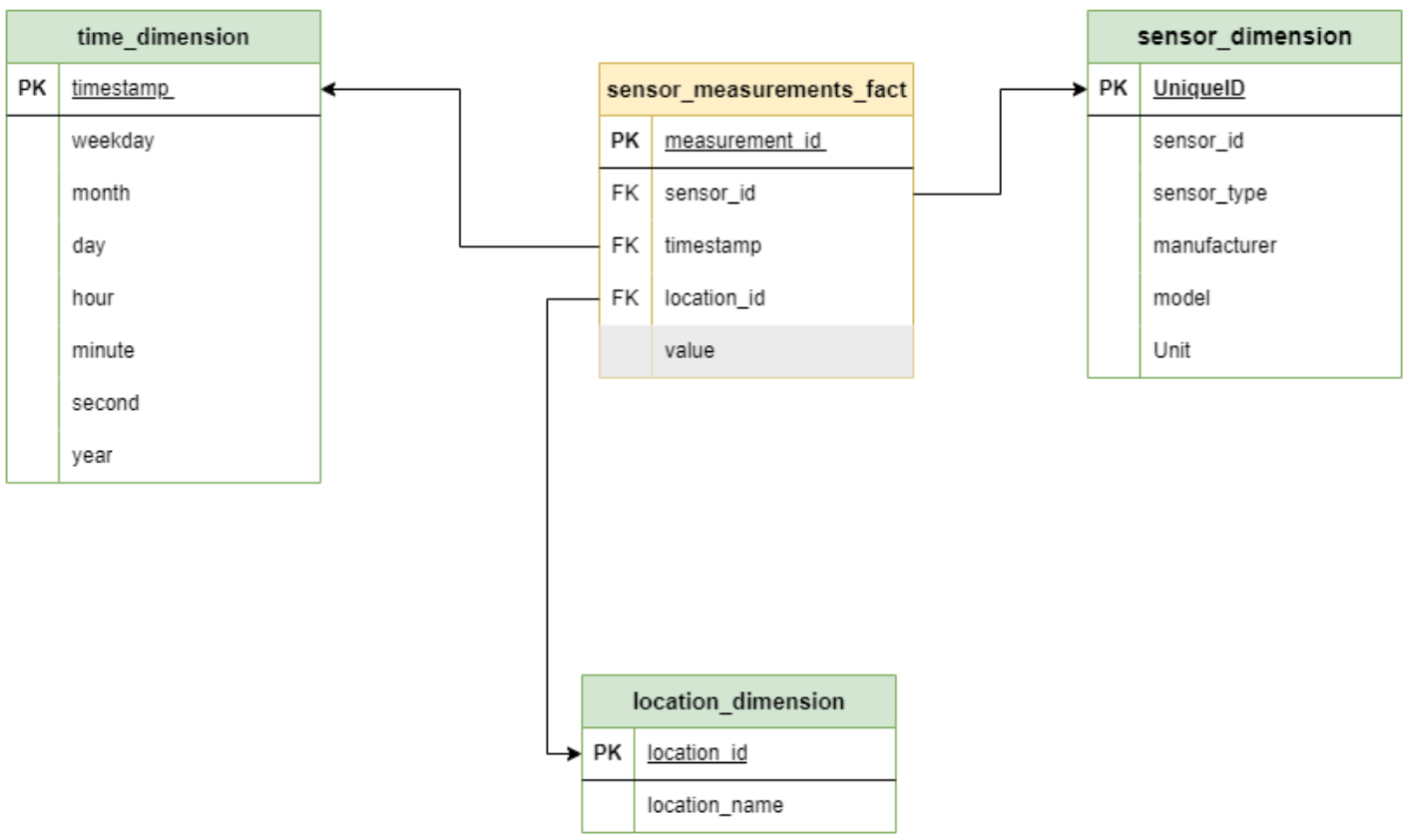
---

5. Analytics ,Visualization Layer & Reporting

- Displays actionable insights:
  - Real-time KPIs like active sensor
  - Historical trends for air quality and other metrics.

---

6. Schema



---

7. Future Work

- **Docker** based containerization of all services to simplify deployment and scalability
- Full **workflow orchestration with Apache Airflow** to automate Spark jobs, report generation, and ETL processes from ODS to the Snowflake EDWH.

High-Level Data Flow

1. **Data Generation**: IoT sensors send data to Kafka.
2. **Ingestion**: Kafka distributes data to real-time and batch processing layers.
3. **Processing**: Kafka consumer and custom transformations using python for both real-time insights and spark batch aggregation.
4. **Storage**: Processed data is stored in a ODS, DWH, while raw data is archived in HDFS.
5. **Visualization**: Power BI to fetch data from the warehouse to create dashboards and analytics.