Real-time IoT Data Pipeline – Final Documentation

TEAM GIRLS

2025 NOVAMBER 30

Cover Page

Project Title: Real-time IoT Data Pipeline

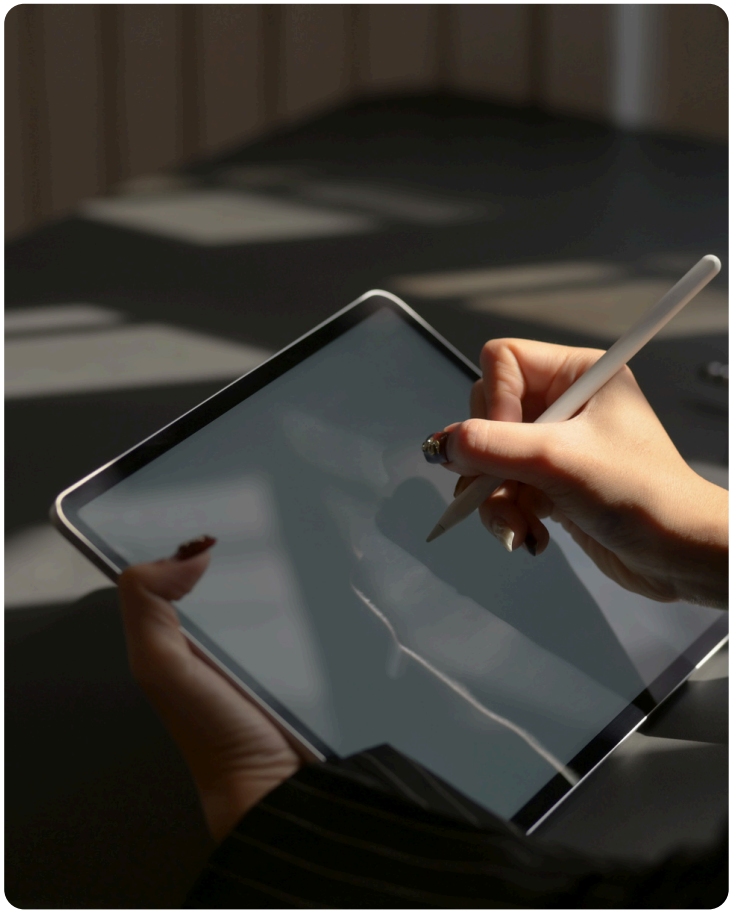
Team Leader: Samaa Sobhy

Email: samasobhy363@gmail.com

Date: 29/11/2025

Project Overview

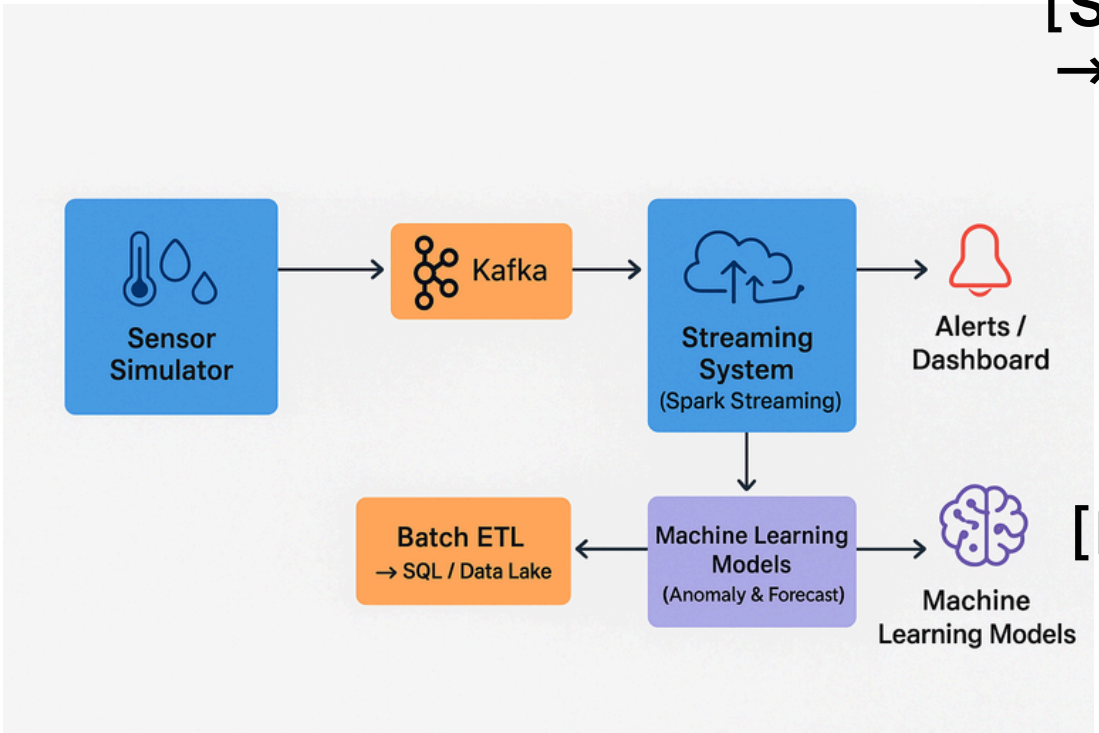
- **Problem: Managing large-scale IoT sensor data in real-time while detecting anomalies and forecasting temperature.**
- **Solution: A unified pipeline using Python, Apache Kafka, Spark Streaming, batch ETL, SQL Server, and machine learning models for anomaly detection and forecasting.**
- **Goal: Provide real-time insights and alerts to stakeholders via dashboard integration.**



Abstract

This project simulates IoT sensor data (temperature and humidity) and processes it using both batch and real-time streaming pipelines. The solution integrates anomaly detection, temperature forecasting, and dashboard visualization to enable continuous monitoring and alerting for operational decisions.

. Architecture Diagram



[Sensor Simulator] → [Kafka Topic] → [Streaming System (Spark Streaming)] → [Alerts / Dashboard]
↓
[Batch ETL → SQL / Data Lake]
↓
[Machine Learning Models (Anomaly & Forecast)]

Colors & Symbols:
Blue → Cloud / Data
Orange → ETL
Red → Alerts
Green → Dashboard
Purple → Machine Learning
Output: PNG (high-res), PDF (vector-based), embedded in documentation

Components:

- **Sensor Simulator: Generates temperature and humidity data every 5 seconds.**
- **Kafka: Real-time message broker for transporting sensor data.**
- **Spark Streaming: Processes data from Kafka in real-time, triggers alerts, and writes CSV files.**
- **Batch ETL (Spark/Pandas): Processes historical CSV data, transforms it, and loads into SQL/Data Lake.**
- **SQL / Data Lake: Stores Dim + Fact tables for historical analysis.**
- **Machine Learning Models:**
  - **Anomaly Detection (Isolation Forest): Flags abnormal sensor readings.**
  - **Temperature Forecasting (Linear Regression): Predicts future temperature values.**
- **Dashboard / Alerts: Visualizes real-time and historical metrics with notifications.**

. Data Flow

Sensor Simulator

Simulator generates data every 5s

Kafka

receives messages in sensor_stream topic

Spark Streaming

1reads from Kafka:
- o Parses JSON
- o Writes CSV output
- o Triggers alerts based on thresholds
- o Sends real-time results to dashboard

Batch ETL

reads CSV or database:
- o Cleans and transforms data
- o Loads Dim + Fact tables in SQL/Data Lake
- o Exports final warehouse data

Machine Learning:

- o Anomaly detection on temperature/humidity
- o Temperature forecasting based on humidity

Dashboard / Alerts:

Visualizes real-time and historical data, anomalies, and predictions

Stakeholder Analysis

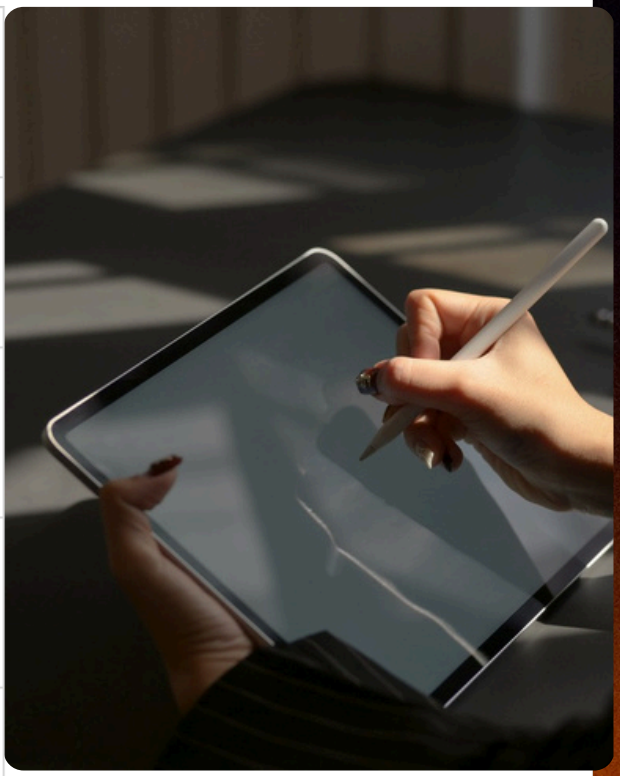
Stakeholder	Role
Data Analysts	Review historical
Operations	Monitor real-time alerts
Admins	Manage pipeline,
Team Members	See roles in Proposal

Tools & Technologies

- Python: Data simulation, ETL, ML models
- Pandas / Spark: Data processing
- Kafka / Docker: Real-time messaging and containerization
- SQL Server / Data Lake: Storage
- Power BI: Real-time dashboard
- Scikit-learn: ML models (Isolation Forest, Linear Regression)

9. Milestones Summary

Milestone	Objective	Deliverables
Data Simulation	Generate sensor data and send to Kafka	Python generator script, sample logs
Batch ETL	Extract, transform, load	ETL script, processed CSV
Streaming Analytics	Real-time alerts	Spark Streaming pipeline, alert
Machine Learning	Anomaly detection & forecasting	Isolation Forest & Linear Regression
Dashboard & Report	Visualize live & historical data	Streamlit/Power BI dashboard,



Code Explanation

Producer (kafka_producer.py)

- Generates sensor data and sends to Kafka topic every 5 seconds

Streaming (spark_kafka_to_csv.py)

- Reads Kafka topic in real-time
- Writes CSV output for downstream ETL and alerts

Alerts (spark_alerts_final.py)

- Checks temperature/humidity thresholds
- Generates alert flags and messages
- Writes console output and CSV

Batch ETL (etl_batch.py)

- Cleans CSV data
- Computes anomalies & quality scores
- Loads Dim + Fact tables into SQL Server
- Exports final CSV snapshot

Machine Learning (train_test.py, prepare_data.py, batch_test.py, stream_test.py)

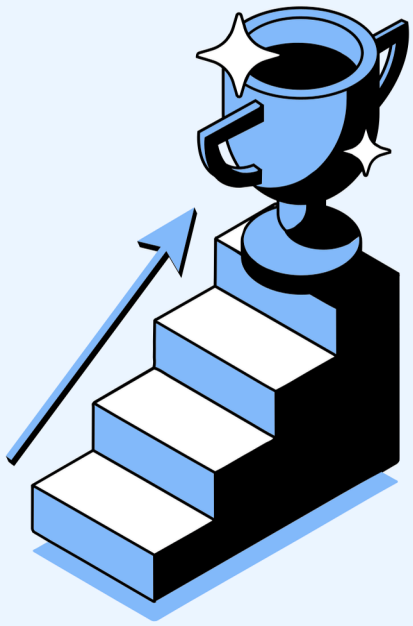
- Trains Isolation Forest for anomaly detection
- Trains Linear Regression for temperature forecasting
- Integrates predictions into batch and streaming pipelines
- Saves models (.pkl) and evaluation reports (.txt)
-

Database (database_setup.sql)

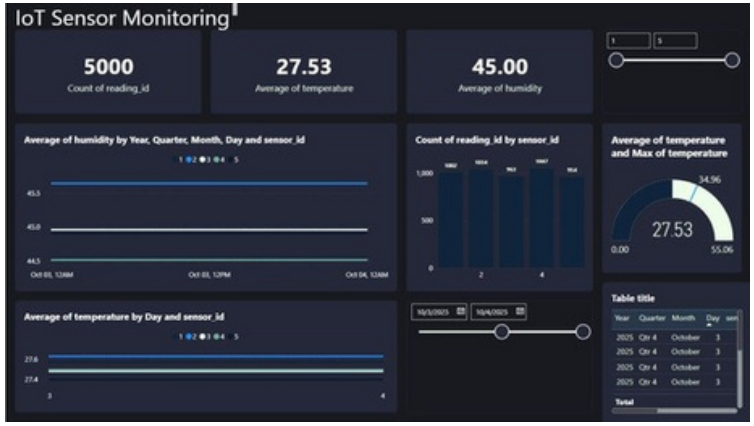
- Creates Star Schema (Dim + Fact tables)
- Inserts initial data for sensors and locations

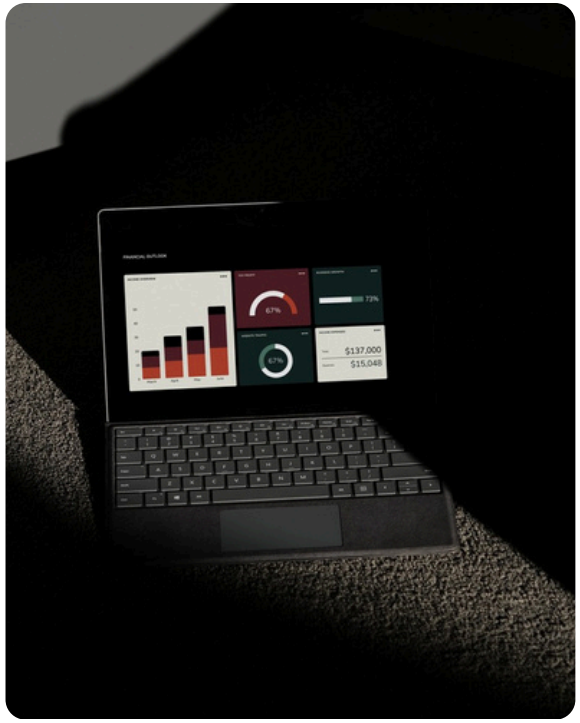
Results / Machine Learning

- Anomaly Detection: Flags abnormal sensor readings
- Temperature Forecasting: Predicts temperature based on humidity
- CSV outputs: anomaly_results.csv, batch_results.csv, stream_results.csv
- Plots: anomaly_plot.png shows anomalies visually
- Evaluation Report: evaluation_report.txt summarizes regression & anomaly metrics



Dashboard

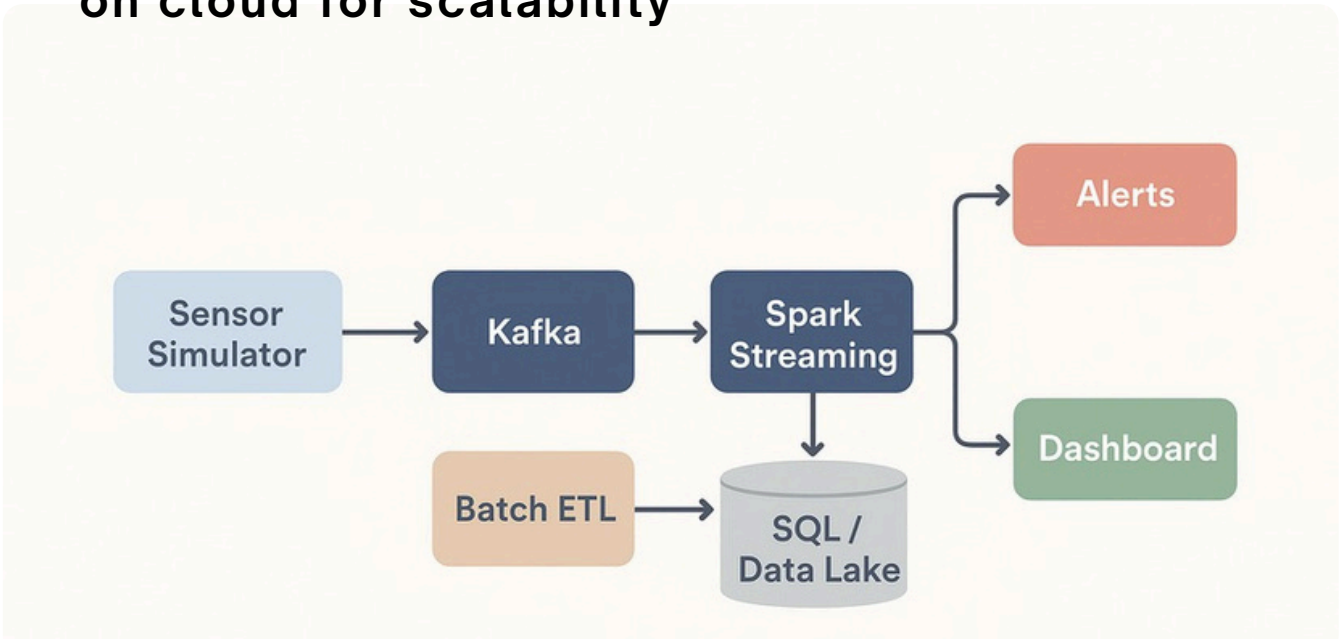




Developed an interactive Power BI dashboard for real-time IoT sensor monitoring. It displays key metrics such as total readings, average temperature & humidity, alerts, and anomalies, with dynamic date filtering. Visualizations include line charts, bar charts, gauges, and summary tables, providing a clean interface for quick insights and improved sensor monitoring.

Conclusion

- Successfully implemented end-to-end IoT pipeline with batch & streaming processing
- Integrated anomaly detection and temperature forecasting models
- Provides real-time monitoring and alerting
- Future Work: Add predictive maintenance alerts, deploy on cloud for scalability



We hope this project meets your expectations and showcases our work effectively.

Team members:

1. Samaa Sobhy
2. Shahd Osama
3. Nada Khaled
4. Basmalla Eid,
5. Menna Saber
6. Habiba Elnamnky

Thank you for reviewing our work, and we hope you enjoyed it

TEAM GIRLS

2025 NOVAMBER 29