Harnessing the Data Flood: A Real-Time Pipeline for Smart Building Management

A Data Engineering Project for the Digital Egypt Pioneers Initiative

Date : 12/11/2025

Presenters:

Mohamed Ahmed Abdelaziz

Rodina Amr

Hanin Salem

Youssef Tamer

Ganna Ehab

Mohamed Ibrahim

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

The Modern Building's Dilemma: Drowning in Data,Starving forInsight

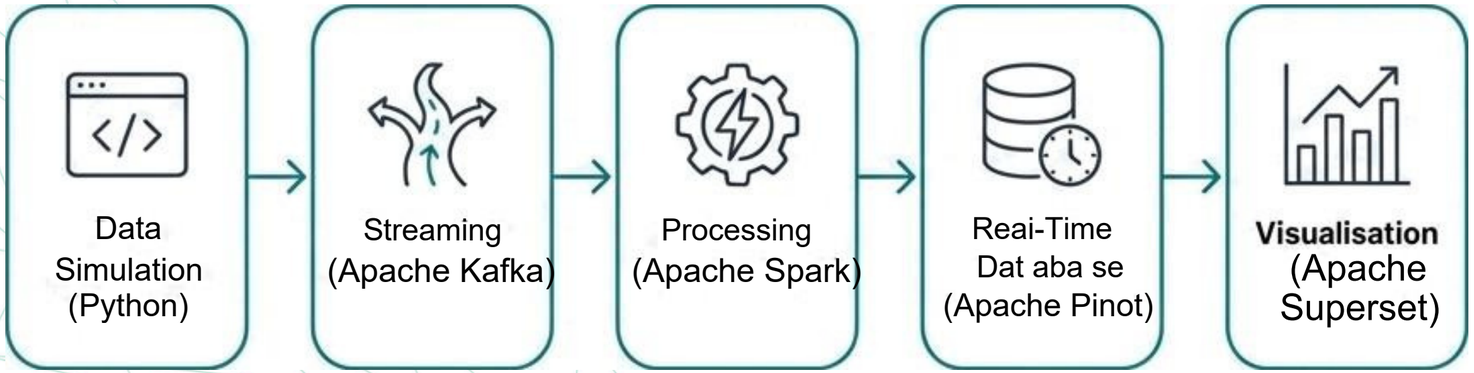
Modern buildings generate a massive, continuous streamof sensor data, including temperature, humidity, CO2, motion, and energy consumption.

Without a properdatapipeline,thisvaluable information becomes messy, unorganised, and difficult to analyse in real time.

This directly leads to significant operational inefficiency, unnecessary energy waste, and poor performance decisions by facility managers.

eyouth

Our Solution: An End-to-End, Real-Time Analytics Pipeline



End-to-End Real-Time Flow: Continuous ingestion, processing, and visual1sation with no batch de la ys.

Low-Latency Analytics: A powerful combination of Apache Pinot and Superset enables immediate querying and insights.

Scalable & Flexible: The architecture is designed Io handlebothstreamingandbatch workloads effectively.

Actionable Insights: Unlocks critical smart—building metrics for energy, air quality, and occupancy management.

The Blueprint: Our Technology Stack

Data Streaming

Apache Kafka

For high—throu9hput,fault—tolerant real-time data i nge s tio n.

Data Processing

Apache Spark

For scalable batch and structured sireaming data processing, cleaning, and aggregation,

Real-Time Analytics Database

Apache Pinot

A real—time, distributed OLAP datastore,purpose-buiitforlow-latency queries on fresh data.

Trino

A high-performance query engine that acts as the vital conneciorbetweenPinotand Superset.

Data Visualisation

Apache Superset

For creating interactive, real—time dashboards and visualising key performance indicators.

Languages & Supporting Tools

Python

Used for data simulation and processing logic.

SQL

The language for querying data in Pinoi via Trino.

GitHub

For version control and collaborative development.

Anatomy of the Pipeline: Data Flow & Architecture

Detailed Data Flow

1. APythonscriptsimulates and generates real-time sensor data.
2. Data is published to the Kafka topic bms_data .
3. An Apache Spark streaming job consumes from bms data‘, performs cleaning and aggregation, and publishes the processed data to a new Kafka topic, ”spark_data”.
4. The processed data is ingested from ‘sparldata‘ into Apache Pinot columnar tables.
5. Trino queries Pinot's tables, exposing the data for visualisation.
6. Superset dashboards query Trino to display real-time KPIs.

Database Architecture Rationale

Apache Pinot Rationale

Chosen forits columnar storage and optimisation for real-time OLAP queries, handling high-throughput streaming ingestion with powerful aggregation and filtering capabilities.

Key Entities

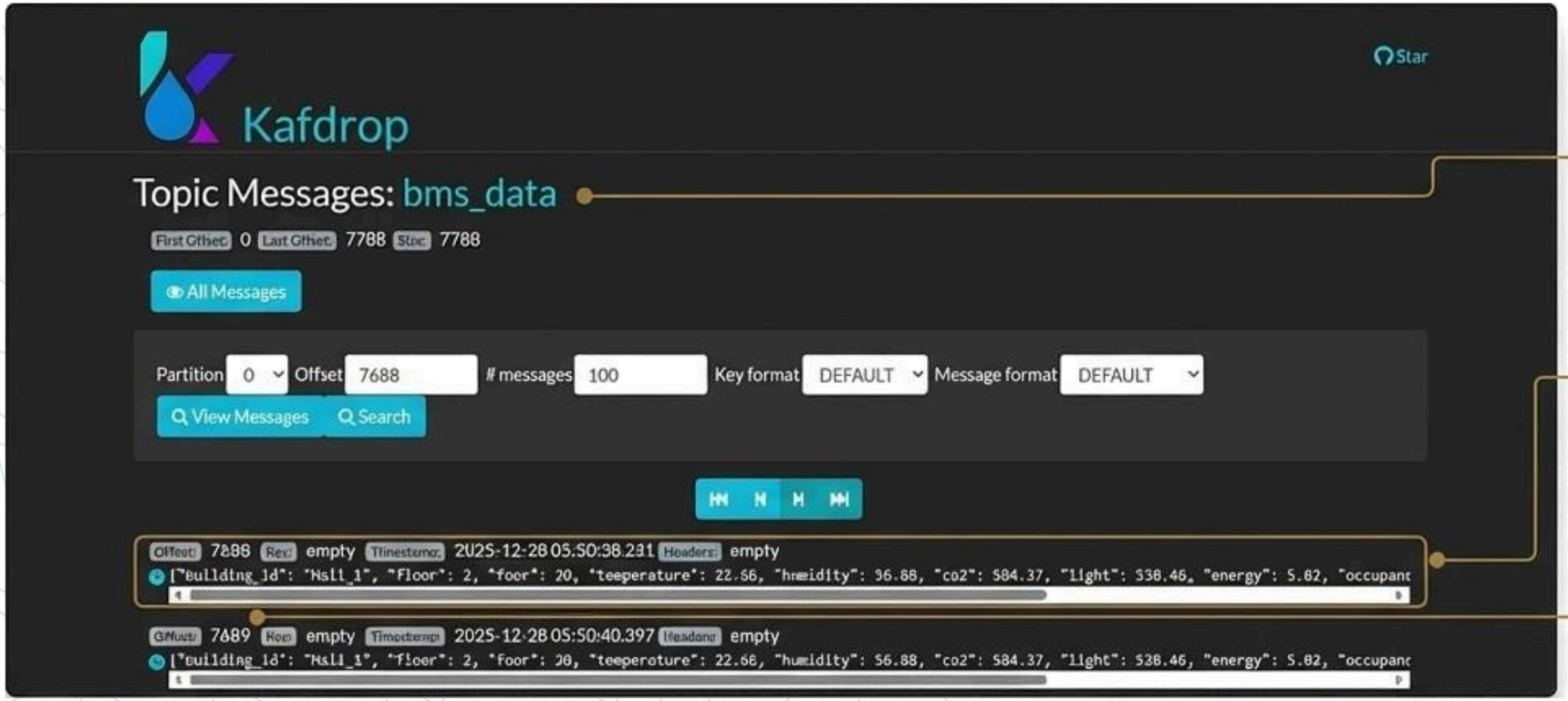
Thedata model centres on Sensors (temperature, humidity, CO2, etc.) which produce timestamped Readings .

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

The Data's Journey, Part Real-Time Ingestion

Thepipeline begins with a Python simulator generatingcontinuous sensor readings.These raw data points are immediately published as messagesto the Apache Kafkatopic ' bos_data". Thescreenshot below shows a live view of these messages arriving in the topic, confirming successful data capture at the entry point of our system.



The topic name: bns data".

The message payload, showing fields like bu11ding_id", t 1mestarip", temperature', co2*,and energy'.

The constantly increasing offset number, indicating a live stream.

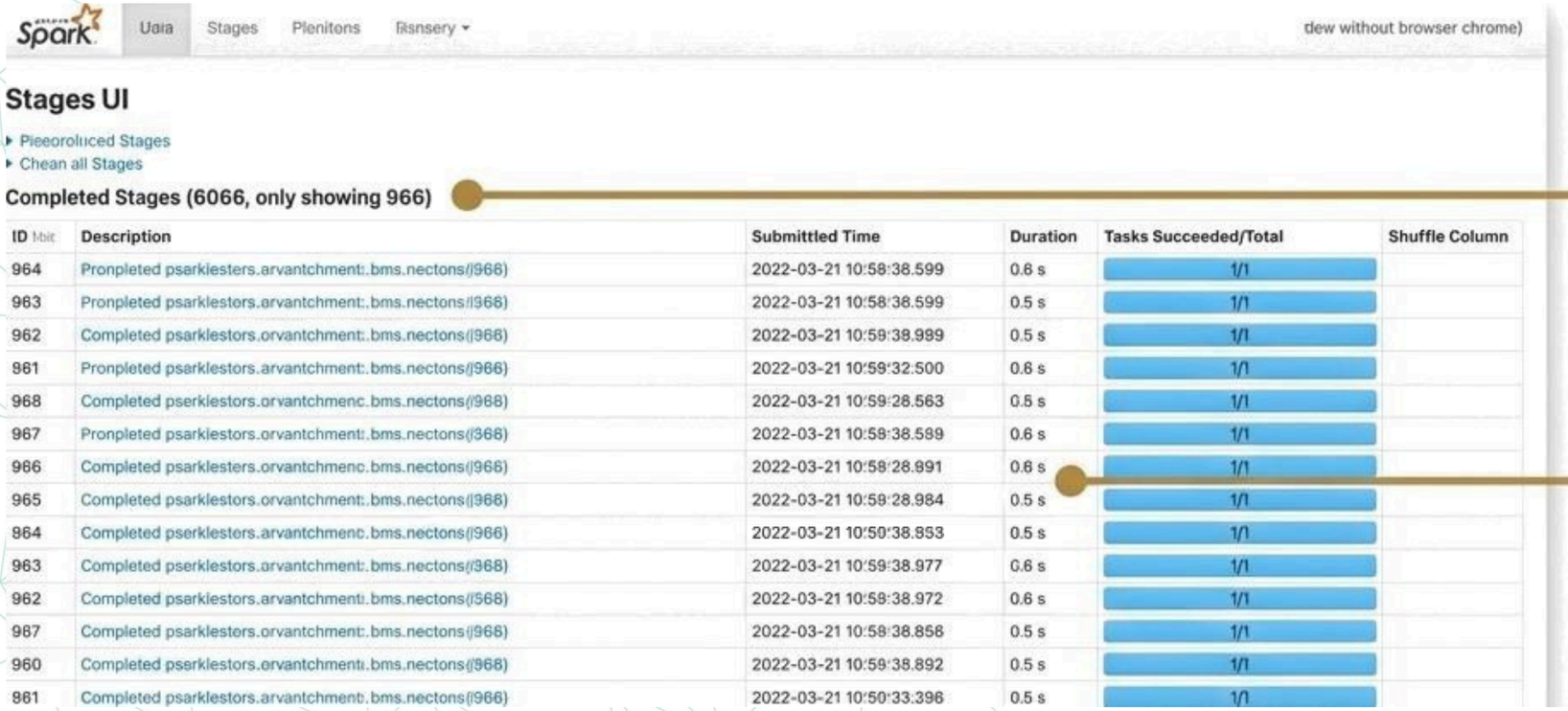
eyouth

The Data's Journey, Part 2: Stream Processing with Spark

AnApache Sparkstreaming application continuously consumestheraw datafromthe”bms_data”Kafka topic.

Spark jobs perform essential cleaning, transformation, and aggregation logic.

The Spark UI confirms that these processing stages are completing successfully and efficiently in near real—time, preparing the data for storage.

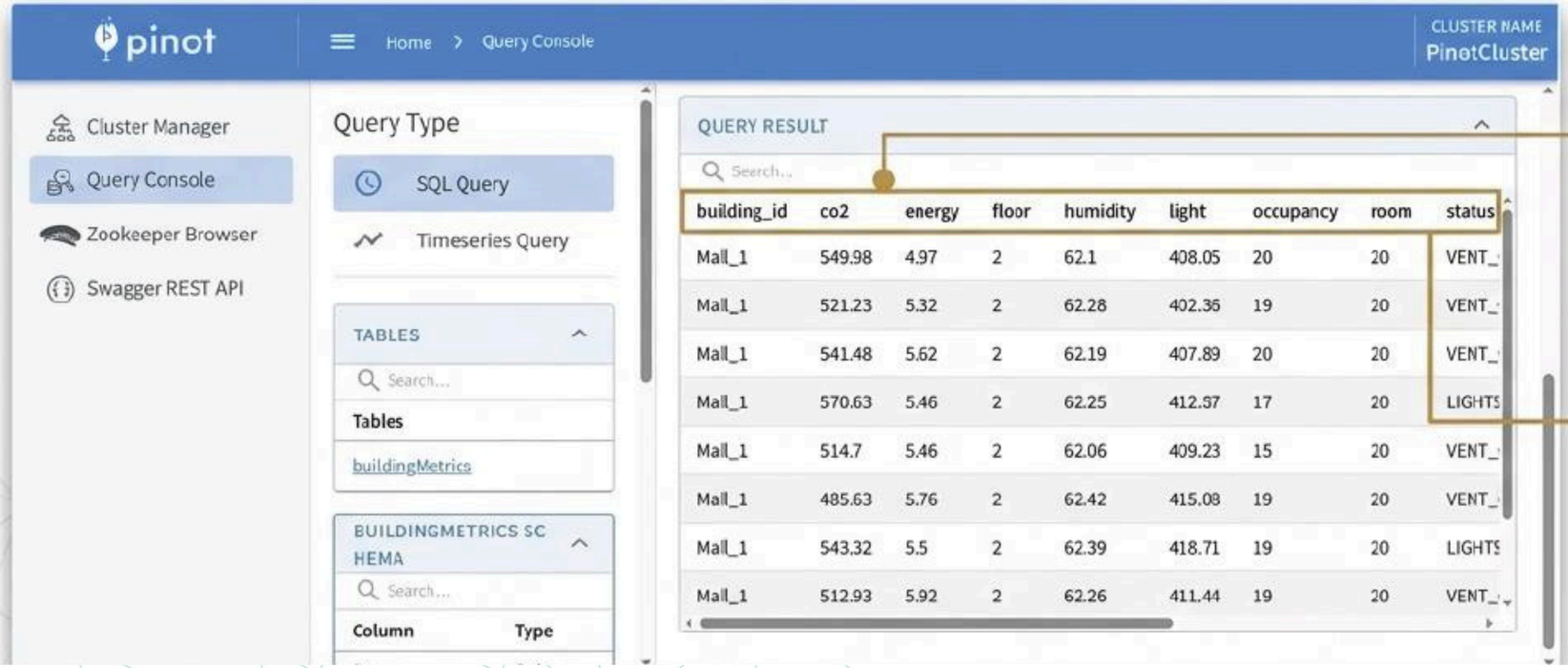


A list of thousands of successfully executed micro-batches, proving continuous processing.

The low duration of each stage underscores the high performance of the processing layer.

The Data's Journey, Part 3: Storing for Low-Latency Analytics

Cleaned and aggregated data from Spark is ingested into Apache Pinot. Pinot stores the data in a columnar format, highly optimised for the fast analytical queries required by our dashboards. The Pinot Query Console below shows the structured `buildingMetrics` table, confirming that processed data is available and ready for immediate querying.



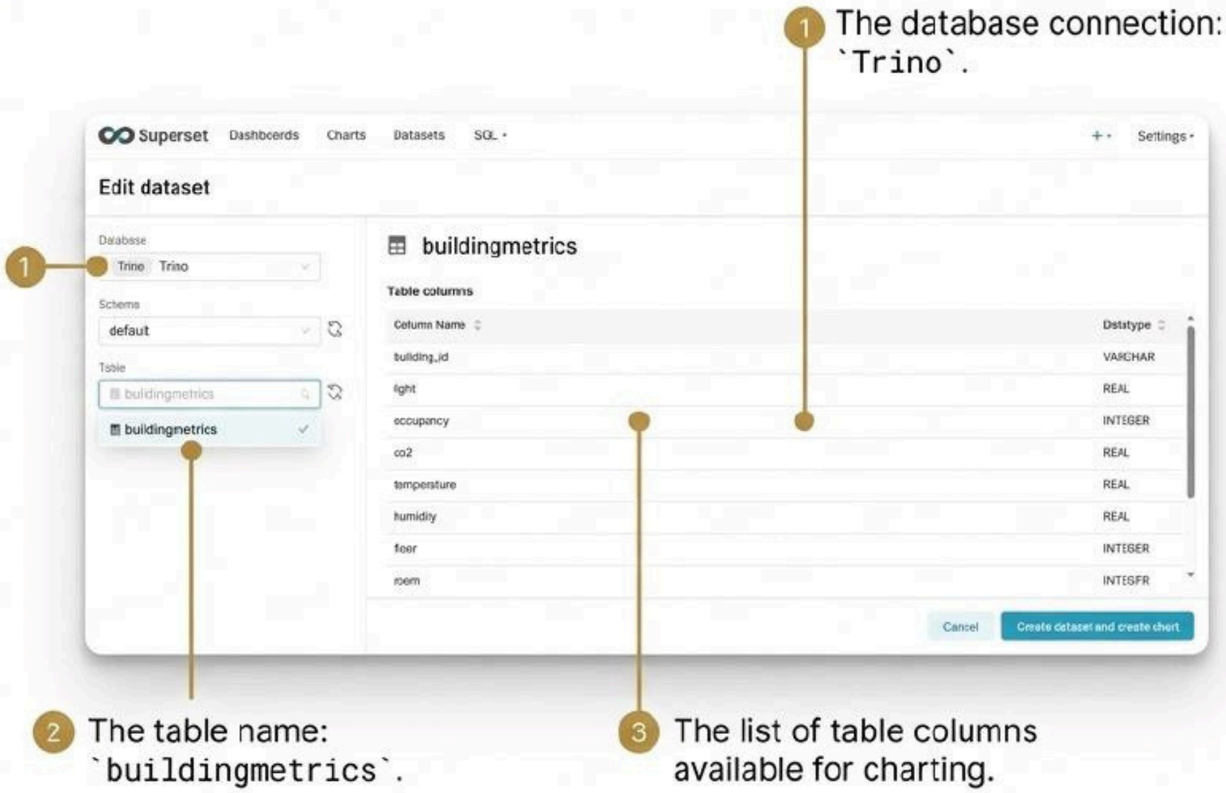
The selected table: `buildingMetrics`.

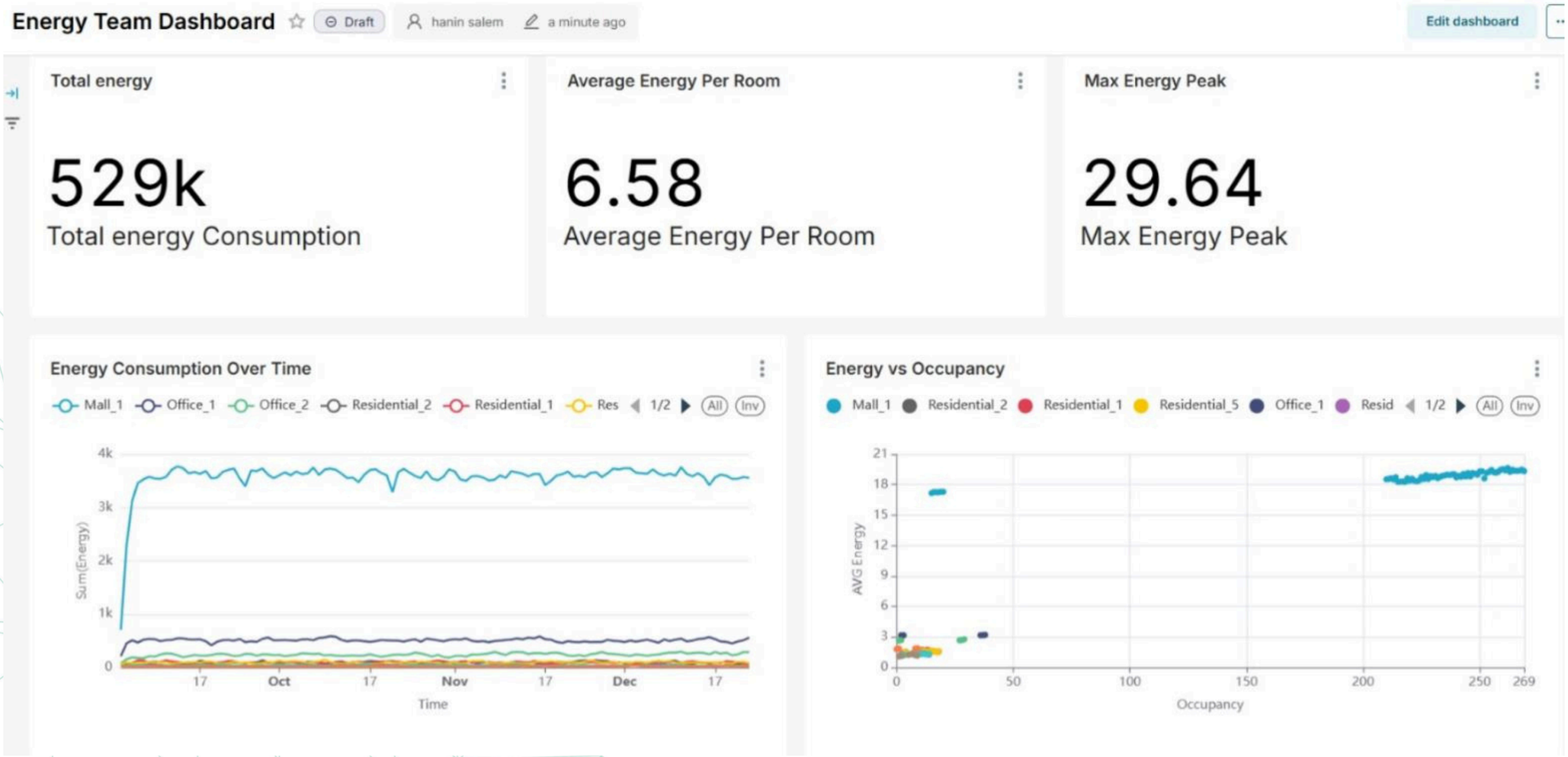
The query result, showing structured columns like `building_id`, `co2`, `energy`, `occupancy`, etc.

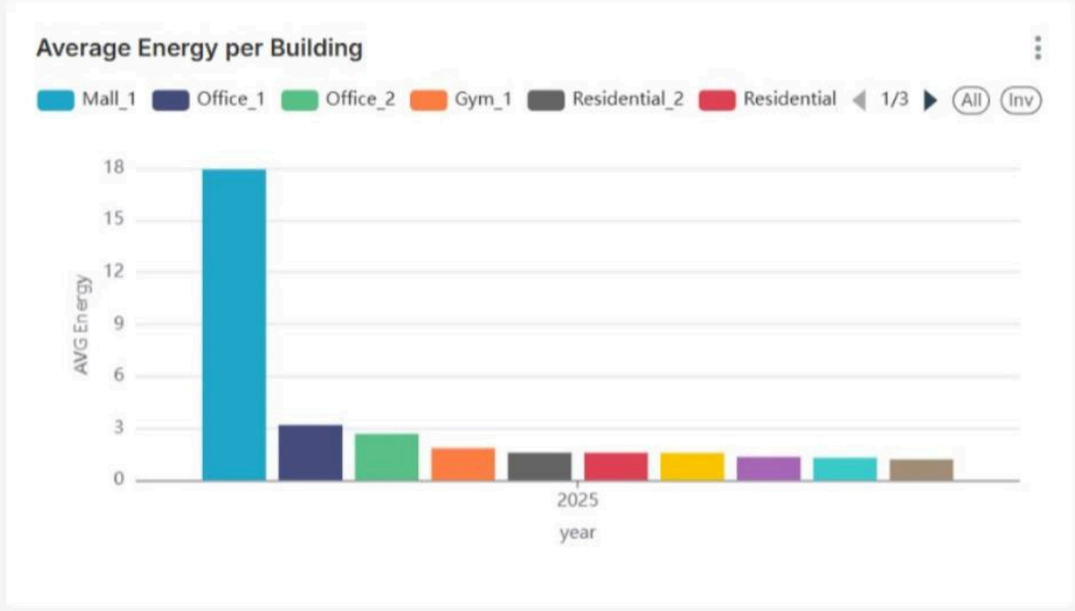
The Data's Journey, Part 4: Connecting to Superset via Trino

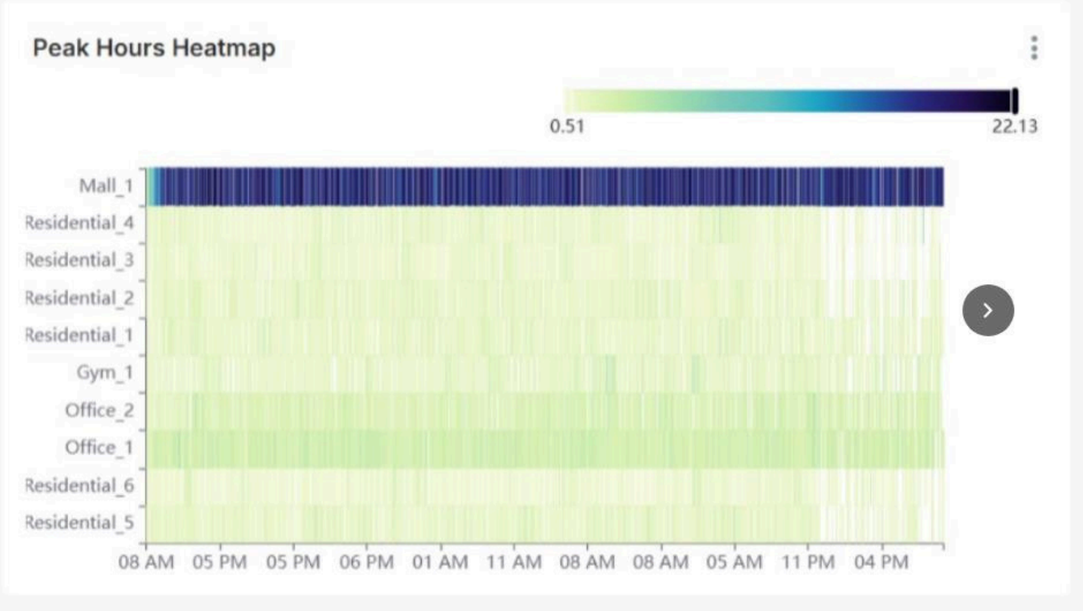
To make the data in Pinot available for visualisation, we use Trino as a federated query engine. Apache Superset connects to our Trino server, which in turn queries the Pinot database. This architecture allows Superset to leverage Pinot's speed while using standard SQL.

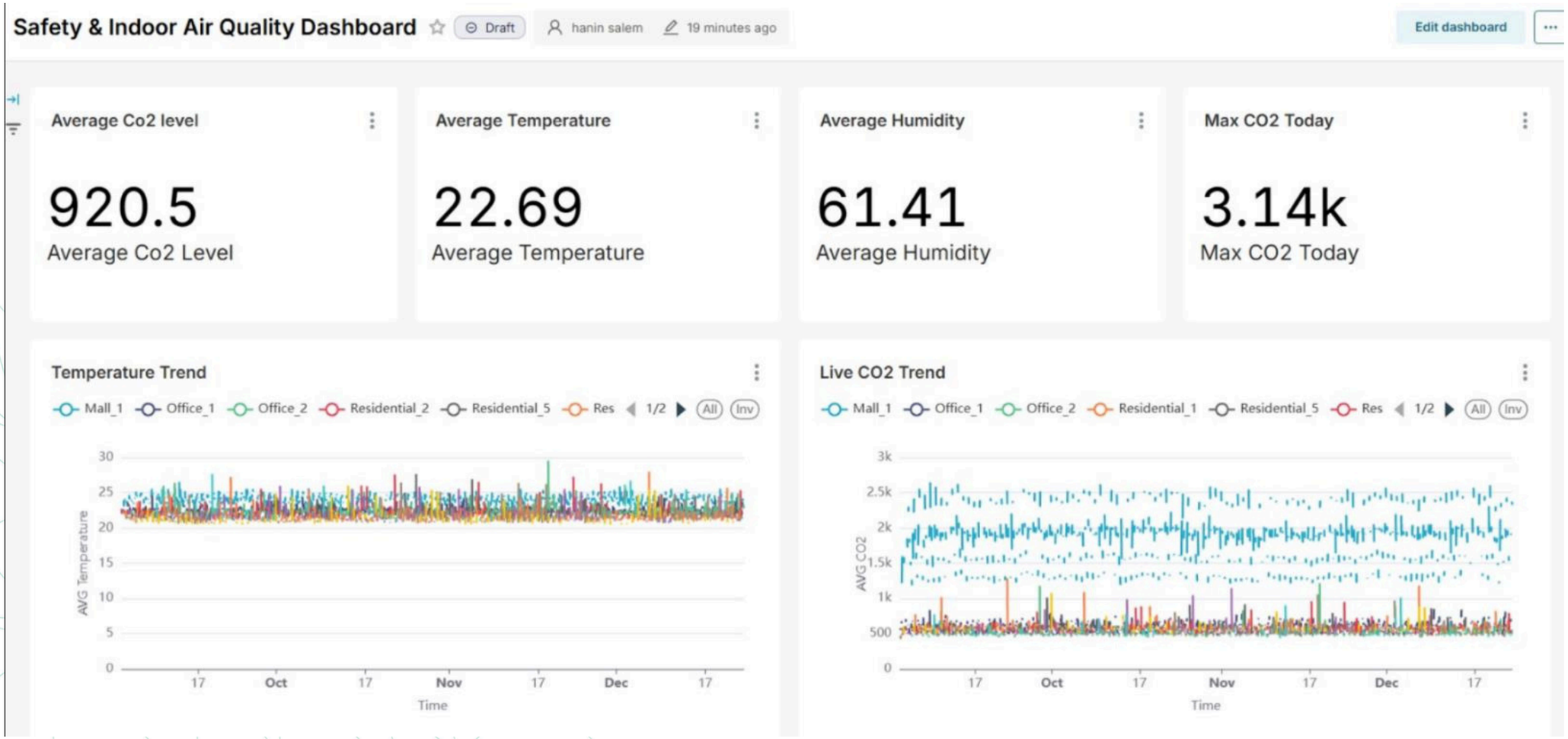
The screenshot shows the `buildingmetrics` dataset successfully configured within Superset, ready for dashboard creation.

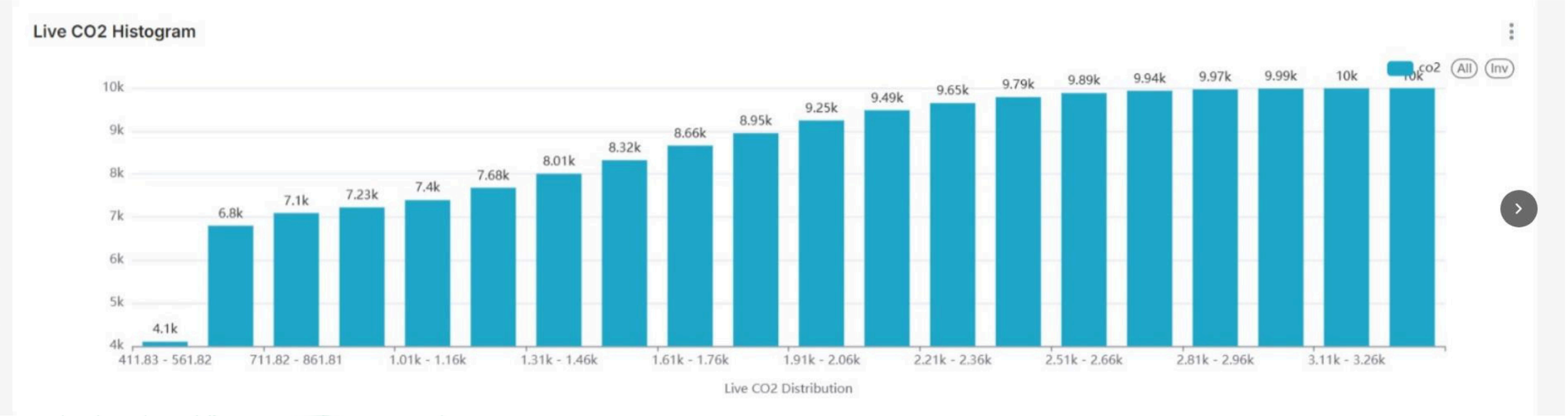


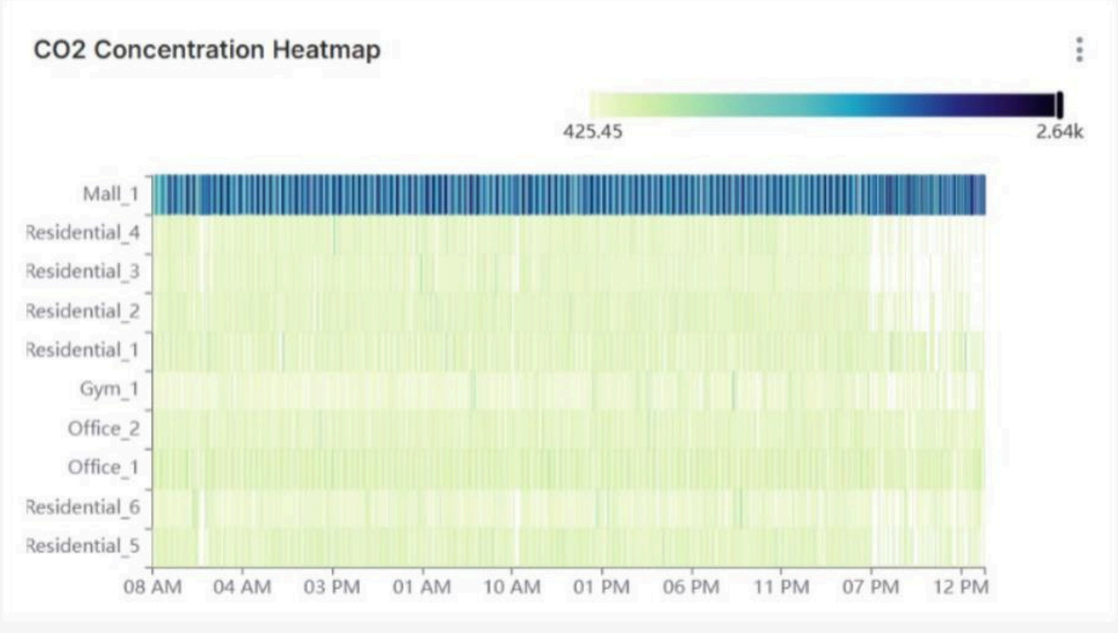

Energy Team Dashboard
Draft
hanin salem
a minute ago
Edit dashboard
Total energy
529k
Total energy Consumption
Average Energy Per Room
6.58
Average Energy Per Room
Max Energy Peak
29.64
Max Energy Peak
Energy Consumption Over Time
Mall_1
Office_1
Office_2
Residential_2
Residential_1
Res
1/2
All
Inv
4k
3k
2k
1k
0
Sum(Energy)
17
Oct
17
Nov
17
Dec
17
Time
Energy vs Occupancy
Mall_1
Residential_2
Residential_1
Residential_5
Office_1
Resid
1/2
All
Inv
21
18
15
12
9
6
3
0
AVG Energy
0
50
100
150
200
250
269
Occupancy

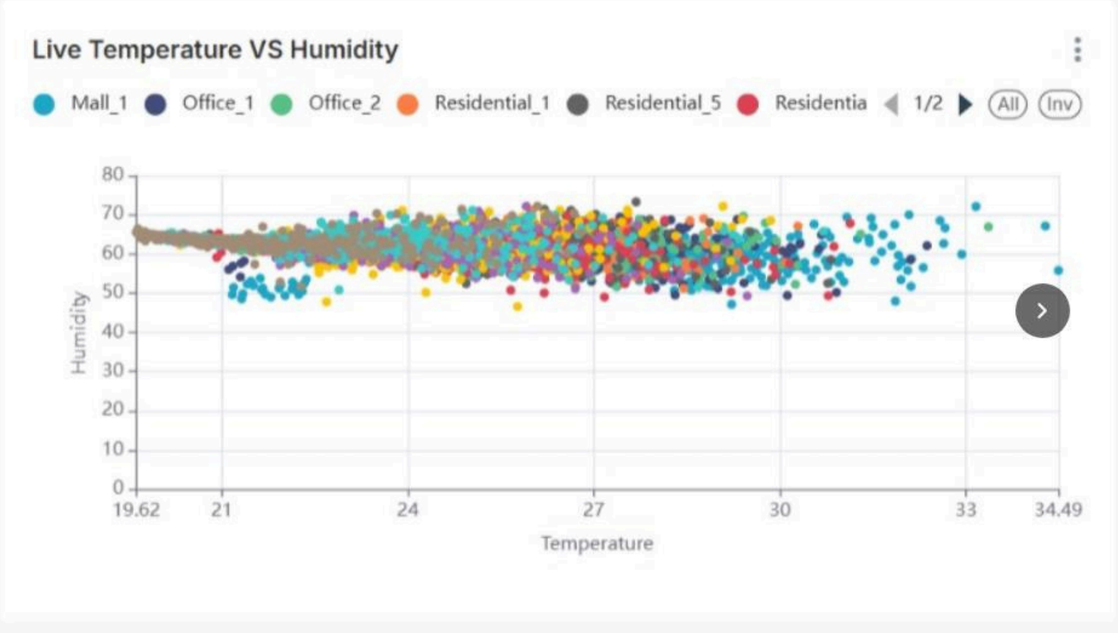

Average Energy per Building
Mall_1
Office_1
Office_2
Gym_1
Residential_2
Residential
1/3
All
Inv
18
15
12
9
6
3
0
AVG Energy
2025
year


Peak Hours Heatmap
0.51
22.13
Mall_1
Residential_4
Residential_3
Residential_2
Residential_1
Gym_1
Office_2
Office_1
Residential_6
Residential_5
08 AM
05 PM
05 PM
06 PM
01 AM
11 AM
08 AM
08 AM
05 AM
11 PM
04 PM


Safety & Indoor Air Quality Dashboard
Draft
hanin salem
19 minutes ago
Edit dashboard
Average Co2 level
920.5
Average Co2 Level
Average Temperature
22.69
Average Temperature
Average Humidity
61.41
Average Humidity
Max CO2 Today
3.14k
Max CO2 Today
Temperature Trend
Mall_1
Office_1
Office_2
Residential_2
Residential_5
Res
1/2
All
Inv
AVG Temperature
30
25
20
15
10
5
0
17
Oct
17
Nov
17
Dec
17
Time
Live CO2 Trend
Mall_1
Office_1
Office_2
Residential_1
Residential_5
Res
1/2
All
Inv
AVG CO2
3k
2.5k
2k
1.5k
1k
500
0
17
Oct
17
Nov
17
Dec
17
Time


Live CO2 Histogram
co2
All
Inv
10k
9k
8k
7k
6k
5k
4k
4.1k
6.8k
7.1k
7.23k
7.4k
7.68k
8.01k
8.32k
8.66k
8.95k
9.25k
9.49k
9.65k
9.79k
9.89k
9.94k
9.97k
9.99k
10k
411.83 - 561.82
711.82 - 861.81
1.01k - 1.16k
1.31k - 1.46k
1.61k - 1.76k
1.91k - 2.06k
2.21k - 2.36k
2.51k - 2.66k
2.81k - 2.96k
3.11k - 3.26k
Live CO2 Distribution


CO2 Concentration Heatmap
425.45
2.64k
Mall_1
Residential_4
Residential_3
Residential_2
Residential_1
Gym_1
Office_2
Office_1
Residential_6
Residential_5
08 AM
04 AM
03 PM
01 AM
10 AM
01 PM
06 PM
11 PM
07 PM
12 PM


Live Temperature VS Humidity
Mall_1
Office_1
Office_2
Residential_1
Residential_5
Residentia
1/2
All
Inv
Humidity
80
70
60
50
40
30
20
10
0
19.62
21
24
27
30
33
34.49
Temperature

رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Empowering the Facility Manager: From Data to Decisions

Feature	Benefit	Outcome
Real-time monitoring of energy, COC, and oc cup anc y.	Faster, data-driven operational decisions.	Optimised energy cOnsumption and reducedoperational costs.
Automated alerts for abnormal readings (e.g., high CO_2).	Proactive air quality management and improved occupant comfort.	Healthier and more productive building environments.
Htstortcal trend analysis for all metrics.	Root cause analysis of inefficiencies.	Prevention of systemic waste and equipment malfunction.

eyouth

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Project Execution: Status, Testing, and Quality Assurance

Current Status

Pipeline is fully functionalin beta.

Data flows correctly from Python simulation through to the Superset dashboards.

Rigorous Testing Phases

Unit Testing: Individual Spark processing functions were thoroughly tested for correctness.

Integration Testing: The end-to-end connection between Kafka, Spark, Pinot, and Superset was validated.

FunctionalTesting:Dashboards were verified to ensure they displayaccurate,real-timeKPIs that match the source data.

QA Feedback Highlights

- Dashboardsare responsive with low-latency updates.
- Displayed KPIs are accurate and reliable.

eyouth

The Engineering Team and Collaborative Process

Team Members & Roles

Mohamed Ahmed: Team Leader / DataEngineer (Pipeline Architecture & DataGeneration)

Hanin Mohamed: StreamingSpecialist (Kafka Operations)

Rodina Amr: Data ProcessingSpecialist (Spark Batch 8 St rea m ing)

Youssef Tamer: Database Specialist (Pinot Setup & Performance)

Ganna Ehab: Visualisation Specialist (Superset Dashboards)

Mohamed Ibrahim: Documentation & QA (Testing & Reporting)

Collaboration Methods

Process: Agile methodologywith regular team syncs.

Tools: GitHub for source control,VS Code for development, and Slack/Teams for com m u nica tio n.

Note: The team used Tailscale VPN to create a secure network, allowing all comoonents to connect seamlessly without public internet exposure.

Conclusion: Transforming Raw Data into a Powerful Decision-Making Tool

This project successfully demonstrates the design and implementation of an end-to-end, real-time data pipeline for smart building management. We have built a robust, scalable system that converts a chaotic stream of sensor data into clear, actionable insights, empowering facility managers to optimise energy, improve occupant comfort, and reduce waste.

For Further Exploration:

The complete source code and technical documentation are available for review.

GitHub Repository: github.com//MAAbdelaziz22 DEPI Project

Questions?