

# Final Group Project Report

## Team

Yergazy Adil 22B22B1519

Kairov Danial 23B031324

Babayev Yerassyl 23B031232

## 1. Goal

Implement a complete pipeline for frequently updating real-world data: pseudo-streaming ingestion (API to Kafka), hourly batch cleaning + storage (Kafka to SQLite), and daily analytics (SQLite to summary table).

## 2. Data Source (API) and Justification

API: OpenAQ v3 (air-quality measurements).

- Real values from sensors (e.g., PM2.5).
- It was in the Allowed API categories.
- Structured JSON, stable/documented REST API.
- New/updated measurements typically appear hourly or more (depends on sensor).

Endpoints used:

- GET /v3/locations/{location\_id}/sensors
- GET /v3/locations/{location\_id}/latest

## 3. Architecture & Airflow DAGs

Flow: OpenAQ API -> Job 1 -> Kafka (raw\_events) -> Job 2 -> SQLite (events) -> Job 3 -> SQLite (daily\_summary).

Job	Schedule	What it does
DAG 1	*/10 * * * *	Poll API every POLL_SECONDS (~60s) and send raw JSON to Kafka (RUN_SECONDS=540 per run).
DAG 2	@hourly	Read new Kafka messages,

DAG 3

@daily

clean with pandas, append  
to SQLite table events.

Aggregate events per day  
and write to SQLite table  
daily\_summary  
(count/avg/min/max).

## 4. Kafka Topic Schema

Topic: raw\_events (default; env KAFKA\_TOPIC). Each message is one measurement with fields:

fetch\_at\_utc (text), location\_id (int), sensor\_id (int), parameter (text), unit (text), parameter\_display (text), datetime\_utc (text), datetime\_local (text), value (real), latitude (real), longitude (real).

## 5. Cleaning Rules (Kafka -> SQLite)

- Convert value to numeric; convert datetime\_utc to UTC datetime.
- Drop rows with missing sensor\_id/parameter/unit/datetime\_utc/value.
- Filter invalid measurements: value >= 0.
- Drop duplicates by (sensor\_id, parameter, datetime\_utc).
- Add ingested\_at\_utc = current UTC time (when storing).

## 6. SQLite Schema

Database: SQLite file (default data/app.db; env SQLITE\_PATH).

Table: events (cleaned measurements)

Columns: id (PK), ingested\_at\_utc, fetch\_at\_utc, location\_id, sensor\_id, parameter, unit, parameter\_display, datetime\_utc, datetime\_local, value, latitude, longitude.

Table: daily\_summary (daily aggregates)

Columns: id (PK), created\_at\_utc, date\_utc, parameter, unit, measurements\_count, value\_avg, value\_min, value\_max.

7. Daily analytics logic

Group by (date\_utc, parameter, unit) and compute: measurements\_count, value\_avg, value\_min, value\_max. Before inserting, delete existing daily\_summary rows for the same date\_utc to keep reruns consistent.

8. Screenshots



produce\_to\_kafka

success

BashOperator

clean\_kafka\_to\_sqlite

success

BashOperator

compute\_daily\_summary

success

BashOperator

```
3 ✓ select * from events;
```

id	ingestion_at_utc	received_at_utc	location_id	sensor_id	param.	unit	parameter_desc.	datetime_utc	datetime_local	value	latitude	longitude
1	2025-12-17T18:45:16.411715+00:00	2025-12-17T18:46:12.408074+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-17 17:35:00+00:00	2025-12-17T17:35:00+05:30	254	28.43576	77.22445
2	2025-12-17T19:00:10.761062+00:00	2025-12-17T19:00:08.707999+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-17 18:10:00+00:00	2025-12-18T00:00:00+05:30	289	28.43576	77.22445
3	2025-12-17T19:00:10.761062+00:00	2025-12-17T19:00:11.355364+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-17 17:35:00+00:00	2025-12-17T17:35:00+05:30	254	28.43576	77.22445
4	2025-12-18T07:35:06.362759+00:00	2025-12-18T07:36:17.622891+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-18 06:30:00+00:00	2025-12-18T12:00:00+05:30	300	28.43576	77.22445
5	2025-12-18T07:38:43.355874+00:00	2025-12-18T07:39:17.762225+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-18 06:30:00+00:00	2025-12-18T12:00:00+05:30	300	28.43576	77.22445
6	2025-12-18T08:10:21.664735+00:00	2025-12-18T07:39:00.793324+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-18 06:30:00+00:00	2025-12-18T12:00:00+05:30	300	28.43576	77.22445
7	2025-12-18T08:10:21.664735+00:00	2025-12-18T07:40:17.530684+00:00	8118	23524	pm25	µg/m³	PM2.5	2025-12-18 07:35:00+00:00	2025-12-18T13:00:00+05:30	258	28.43576	77.22445

```
5 ✓ select * from daily_summary;
```

id	created_at_utc	date_utc	parameter	unit	measurements_count	value_avg	value_min	value_max
3	2025-12-18T07:38:47.875326+00:00	2025-12-17	pm25	µg/m³	3	265.6666666666667	254	289
4	2025-12-18T07:38:47.875326+00:00	2025-12-18	pm25	µg/m³	2	300	300	300