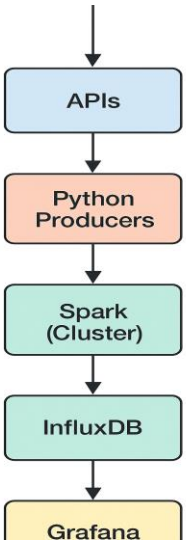
Project Title & Team

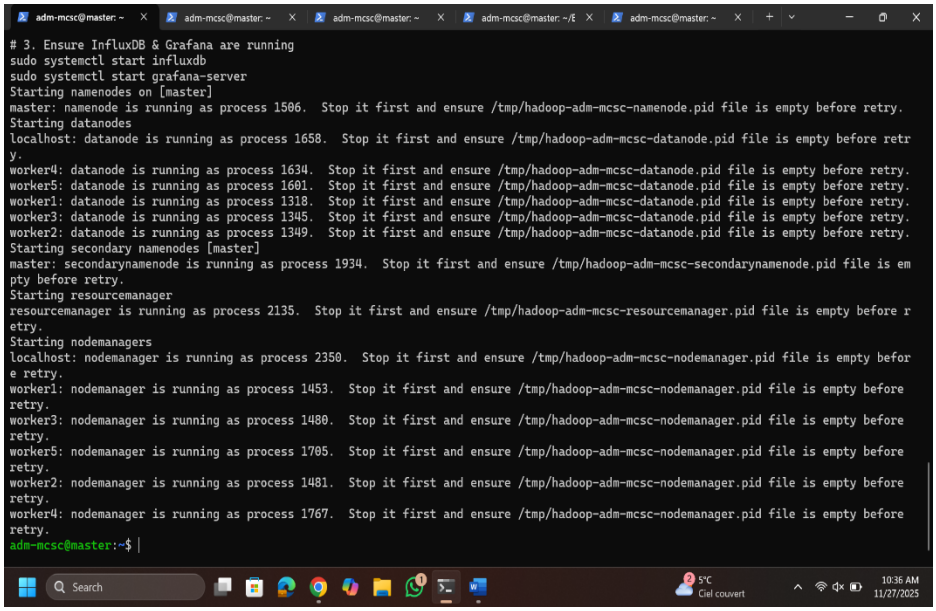
- **Project Name:** EnviroHealth-Monitor
- **Checkpoint:** Milestone 2 - Implementation & Demonstration
- **Team Members:**
- Hussnain Amanat Ali
- Ayman Berri
- Muhammad Irfan

2. System Architecture

- **Description:** Briefly explain the pipeline you built.
    - *We implemented a real-time Lambda Architecture pipeline. Data is ingested from OpenAQ and Open-Meteo APIs via Kafka, processed by Apache Spark Structured Streaming on a 6-node cluster, and stored in both HDFS (Data Lake) and InfluxDB (Real-time Database) for visualization in Grafana.*
- **Diagram:**
- **API -> Python Producers -> Kafka -> Spark (Cluster) -> InfluxDB -> Grafana**
- 

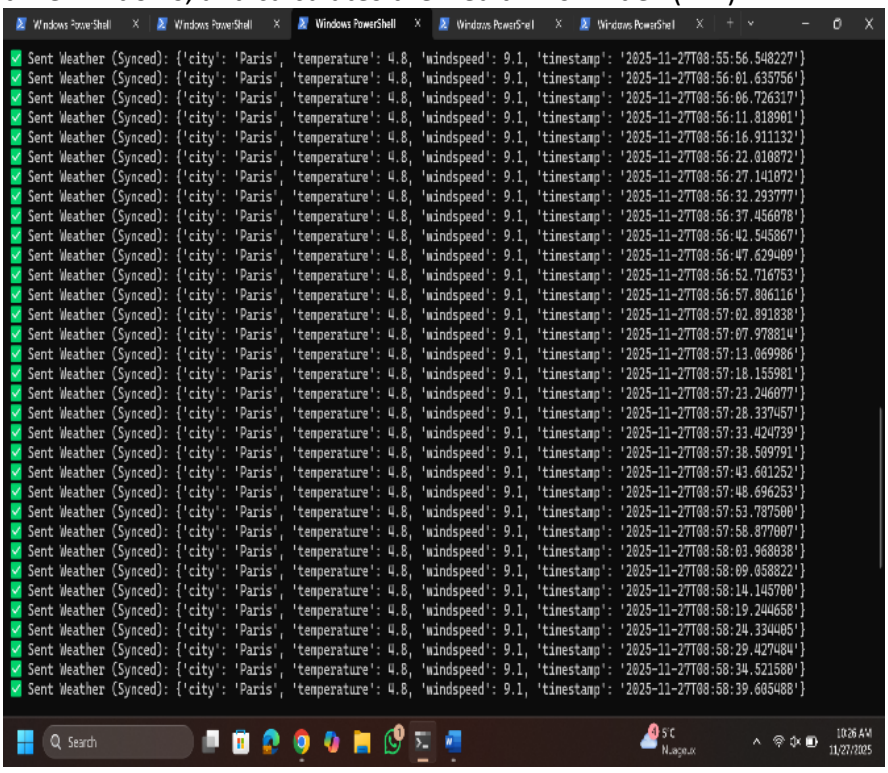
3. Infrastructure (The 6-Node Cluster)

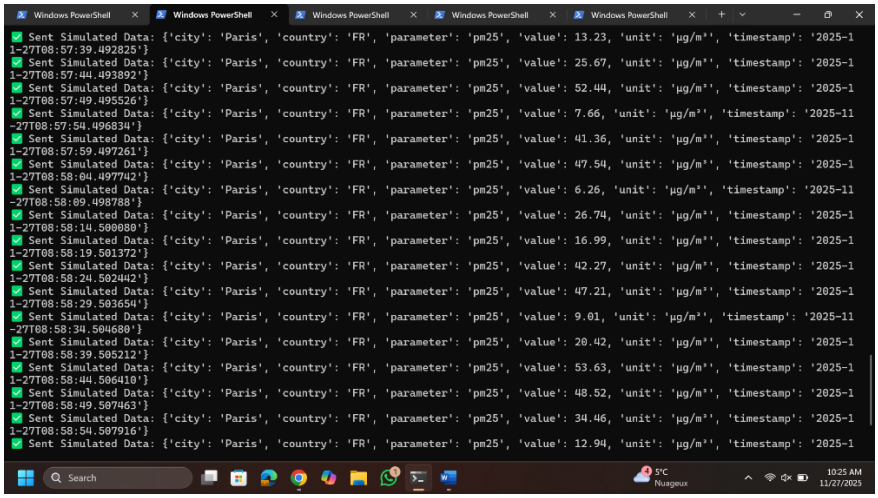
- **Description:** *We successfully configured a 6-node distributed Hadoop cluster on Azure, consisting of 1 Master node and 5 Worker nodes.*

- **Evidence: [Screenshot of hdfs dfsadmin -report]**
    - HDFS Admin Report showing 6 Live Datanodes connected to the Master.
    - 

4. Data Ingestion & Processing

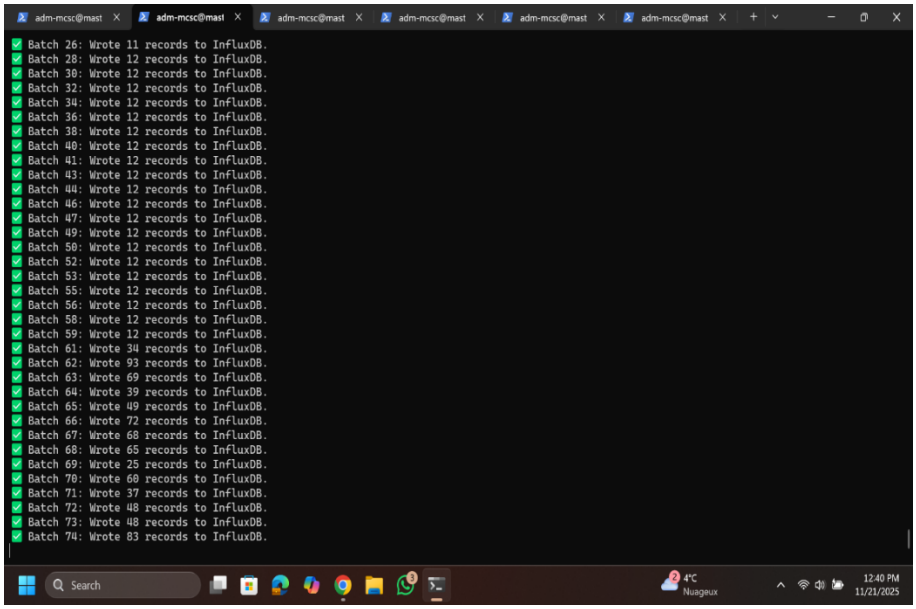
Description: Two Python producers fetch live Air Quality and Weather data and publish to Kafka. A Spark Structured Streaming job consumes these topics, joins the streams based on time windows, and calculates the Health Risk Index (HRI).





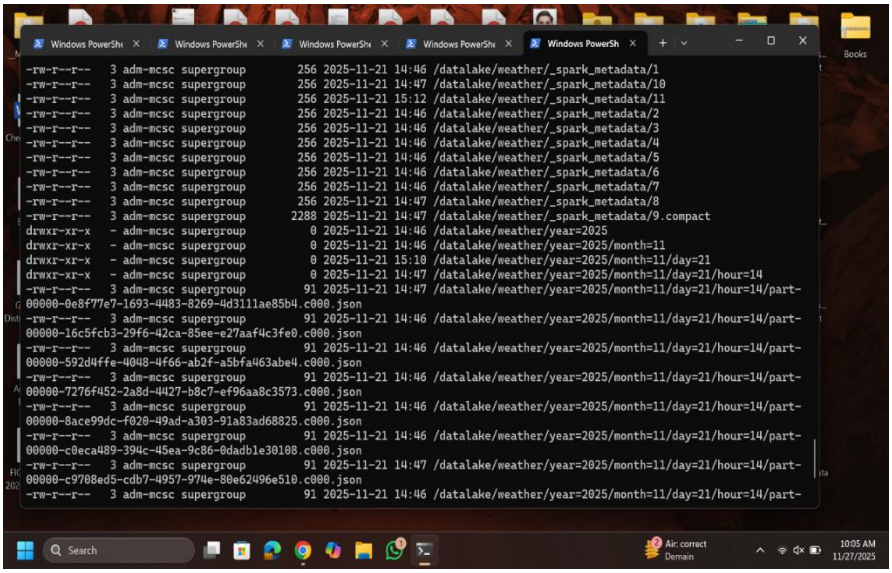
- **Evidence: [Screenshot of Spark Terminal]**

Spark Structured Streaming logs confirming successful processing of micro-batches and writing to InfluxDB.

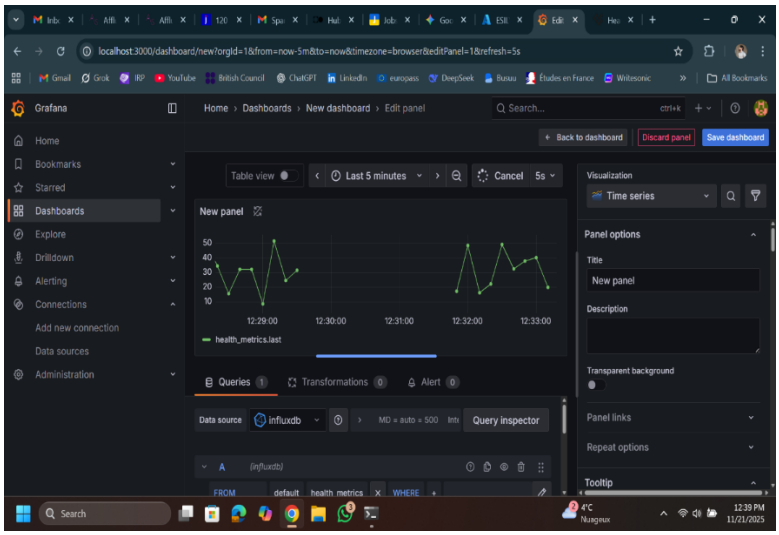


5. Data Lake Storage (The "High Quality" Feature)

- **Description:** *To ensure data reproducibility, raw events are partitioned and stored in HDFS as a Data Lake."*
- **Evidence: [Screenshot of hdfs dfs -ls -R /datalake]**
- HDFS Data Lake structure showing raw JSON events partitioned by year, month, day, and hour.



6. Visualization (The Dashboard)

- **Description:** *Real-time health risks are visualized on a Grafana dashboard connected to InfluxDB.*
- **Evidence: [Screenshot of Grafana]**
    - Real-time Grafana Dashboard displaying the calculated Health Risk Index (Gauge) and live Air Quality trends.
    - 

7. Conclusion

- *We have successfully implemented an end-to-end Big Data pipeline that ingests, processes, stores, and visualizes environmental health data in real-time across a distributed cloud cluster.*