EnviroHealth-Monitor – Milestone 1: Data & Service Brief

Group Members:

- Muhammad Irfan
- Hussnain Amanat Ali
- Ayman

Project Overview

EnviroHealth-Monitor is a **real-time monitoring system** designed to track air quality and assess health risks in major French cities. The system ingests live air pollution and weather data, computes a **Health Risk Index (HRI)** based on PM2.5, PM10, temperature, and humidity, and visualizes trends on an interactive **Grafana dashboard**.

Big Data Components:

- **Ingestion:** Apache Kafka (producers + topics)
- **Storage:** HDFS (raw data lake)
- **Processing:** Apache Spark (Structured Streaming)
- **Time-series Database:** InfluxDB
- **Visualization:** Grafana

Data Sources:

Source	**Type**	**API Endpoint / URL**	**Sample Fields**	**Notes / Limits**
OpenAQ	REST API	https://docs.openaq.org/	city, location, parameter, value, timestamp	Free, rate-limited, attribution required
Open-Meteo	REST API	https://open-meteo.com/en/docs	city, temperature, humidity, timestamp	Free, no API key required

Event Format: JSON
Expected Throughput: ~1–10 events/sec per city
Time Zone: Europe/Paris

Problem & Key Metrics

Problem: Urban air pollution impacts public health, and existing monitoring systems often lack **real-time analysis**. EnviroHealth-Monitor addresses this gap by providing live insights and computing a **Health Risk Index (HRI)**.

Key Metrics / KPIs:

- PM2.5 and PM10 levels
- Temperature and humidity
- Health Risk Index (HRI)
- Dashboard metrics: hourly averages, city comparisons, alerts

Initial Streaming Design

- **Kafka Topics:**
    - openaq_air_quality
    - weather_forecast
- **Message Schema (JSON):**

```
{
  "city": "Paris",
  "parameter": "PM2.5",
  "value": 42.5,
  "timestamp": "2025-10-22T12:00:00+02:00"
}
```

HDFS Landing Plan:

- Raw events stored in hourly folders for reproducibility

Spark Structured Streaming:

- Consume from Kafka, compute sliding-window aggregates, join pollution and weather data, calculate HRI

Repo Skeleton (Initial Setup)

```
EnviroHealth-Monitor/
│
├─ README.md
├─ .gitignore
├─ .gitattributes
├─ producers/
│   ├─ README.md
│   └─ sample_producer.py
├─ schemas/
│   ├─ README.md
│   └─ sample_schema.json
└─ docs/
    ├─ Report1.pdf
    └─ README.md
```

Notes & Assumptions

- Expected throughput is low (~1–10 events/sec per city), both APIs can run simultaneously.
- API rate limits handled via batching or retries.
- Each group member can take ownership of one component for Milestone 2 (Kafka, Spark/HDFS, InfluxDB/Grafana).