

Concept note on system setup for streaming environmental metrics data

<https://github.com/jurdabos/vlc>

For the course DLBDSEDE02 – tutor: Dr. Sahar Qaadan

STUDENT: TORDA BALÁZS, 92128244 (Data Science Bsc.)

Finalized for submission on 26 Oct 2025

Table of Contents

Abbreviations and definitions	ii
Stream processing pipeline – conception phase	1
1. Data source.....	1
2. Overall goal	1
3. Why Kafka.....	1
4. Prototype.....	1
Bibliography	2
APPENDIX.....	3
Weather stations	3
Weather metrics	3
Air stations	3
Air metrics	3
Diagram illustrating the pipeline.....	4

Abbreviations and definitions

Abbrev	Longform (= translation)
al.	alii = others
AV	avenida = avenue
c	Celsius
calidad_am	calidad ambiental = air quality
cf.	cōfer = compare
CO	carbon monoxide
deg	degree
DLBDSEDE02	Distance Learning Bachelor Data Science Elective Data Engineering – course #2
DR	doctor
DT	Delegación Territorial (regional office of the State Meteorological Agency)
e. g.	exemplī grātiā = for example
etc.	et cetera
fecha_carg	fecha de carga = upload date
FIWARE	Future Internet WARE
hpa	hPa = hectopascal
IU	International University of Applied Sciences
JDBC	Java Database Connectivity
m	meter
mm	millimeter
ms	millisecond
NO2	nitrogen dioxide
O3	ozone
p./pp.	page/pages
pct	percentage
pm	particulate matter
SO2	sulphur dioxide
UPV	Universitat Politècnica de València = Polytechnic University of Valencia
W	weather

Stream processing pipeline – conception phase

1. Data source

The data sets the city of Valencia provides expose **environmental metrics** from 11 and **weather metrics** from 5 fixed stations via public APIs (cf. Felici-Castell et al., 2023). Our planned system will ingest, store, and serve pertaining data for more front-end use. Data consumption is to be set up from <https://valencia.opendatasoft.com/api/explore/v2.1/catalog/datasets> for dataset IDs estaciones-atmosferiques-estaciones-atmosfericas and estaciones-contaminacion-atmosferiques-estaciones-contaminacion-atmosfericas, live snapshots offering the latest readings per station.

2. Overall goal

Load the historic data (currently available from 2014 to mid-2025) to a database. Create Python-based puller scripts, flipped to become producers to Kafka, setting up streams that emit each station's data when a snapshot timestamp advances. Build a pipeline that ingests sensor updates, preserves history, and provides fast queries and safe rollups for analytics and possible alerts.

3. Why Kafka

Kafka's replicated log decouples producers & consumers (e. g. Narkhede et al., 2017, Chapter 1), enables replay, and scales via partitions. Offsets, idempotent writes, and topic compaction make operations simple and reliable for demos. Everything can be shipped wrapped as containers for portability.

4. Prototype

Stack: Dockerized Kafka, Schema Registry, Kafka-UI, TimescaleDB, Kafka Connect, Python scripts.

Ingestion: scripts poll Explore v2.1 with offset chaining (`where=fecha_carg>date'${offset}'`), and produce to topic `vlc.air/vlc.weather`.

Storage: TimescaleDB hypertable `air.hyper + weather.hyper` (fiwareid, ts, etc.) with compression, retention, geospatial index, and continuous aggregates for daily/weekly means.

Loading: JDBC Sink does idempotent upserts on (fiwareid, ts).

Observability: Kafka-UI + optional Grafana; CI/CD stubs for reproducibility.

Bibliography

- Felici-Castell, S., Segura-Garcia, J., Perez-Solano, J. J., Fayos-Jordan, R., Soriano-Asensi, A., & Alcaraz-Calero, J. M. (2023). AI-IoT Low-Cost Pollution-Monitoring Sensor Network to Assist Citizens with Respiratory Problems. *Sensors*, 23(23), 9585.
<https://doi.org/10.3390/s23239585>
- Narkhede, N., Shapira, G., & Palino, T. (2017). *Kafka - the definitive guide [electronic resource]: Real-time data and stream processing at scale* (iuo.oai.edge.iu.folio.ebsco.com.fs00001148.0df0ce53.6c47.50e9.a2b0.5af2449cc637).
<https://research.ebsco.com/linkprocessor/plink?id=716f2bfe-f2d6-36ab-97d1-59e7bdca9645>

APPENDIX

Weather stations

W01_AVFRANCIA_10m, W02_NAZARET_10m, W03_VALENCIAEROPUERTO_10m,
W04_VALENCIADT_10m, W05_VALENCIA_UPV_10m

Weather metrics

air_temperature_c = air temperature in °C

humedad_re = % relative humidity

precipitac = precipitation in mm over the reporting interval

presion_ba = barometric surface pressure in hPa

viento_dir = wind direction in degrees (0–360, meteorological convention)

viento_vel = wind speed in m/s

Expected volume for weather metrics ≈ a few hundred readings/day → cc. 10,000/month

Air stations

A01_AVFRANCIA_60m, A02_BULEVARDSUD_60m, A03_MOLISOL_60m,
A04_PISTASILLA_60m, A05_POLITECNIC_60m, A06_VIVERS_60m,
A07_VALENCIACENTRE_60m, A08_DR_LLUCH_60m, A09_CABANYAL_60m,
A10 OLIVERETA_60m, A11_PATRAIX_60m

Air metrics

CO, NO2, O3, PM2.5, PM10, SO2

Expected volume for air metrics ≈ 1000 measurements per day → cc. 30,000/month.

Diagram illustrating the pipeline

