



# Walid EL KHOUKH

Cloud Data Engineer

4+ years of experience in data engineering, cloud, lakehouse

✉ [walidelkhouch99@gmail.com](mailto:walidelkhouch99@gmail.com)

☎ +33 6 50 04 37 94

📍 Île-de-France, France

🌐 [LinkedIn](#)

🐙 [GitHub](#)

## PROFILE

Cloud Data Engineer with 4+ years of experience designing, migrating and industrializing data platforms on AWS and GCP. Led the migration of 50+ data pipelines to Databricks, reducing average runtime by 50% and cloud costs by 25%.

## TECHNICAL SKILLS

**Languages:** Python, SQL, Bash

**Data Engineering:** Apache Spark, Apache Airflow, Delta Lake, Delta Live Tables, BigQuery

**Cloud / Platforms:** AWS, GCP, Databricks, dbt

**Ingestion / Streaming:** Fivetran, Kafka, Redpanda, Debezium, APIs, CDC

**DevOps / Industrialization:** Terraform, Git, GitLab CI/CD, GitHub Actions, Docker, Kubernetes

**Databases:** PostgreSQL, pgvector, MySQL, MongoDB, DuckDB, Oracle

**BI / Dataviz:** Power BI, Apache Superset

**Data Quality / Governance:** dbt Tests, Data Contracts, Data Governance (RLS, Unity Catalog)

**GenAI / LLM Engineering:** LangChain, RAG pipelines, pgvector, MLflow, Embeddings, Ollama, Mistral, Claude Code

## EXPERIENCE

**Socotec Monitoring France, Île-de-France**

02/2023 – Present

*Cloud Data Engineer · Partial Technical Lead*

06/2025 – Present

*APOLLO — ERP Migration from Atlas to SAP / S/4HANA*

- Built **Databricks** pipelines/jobs to ingest and transform SAP data exposed through Datasphere.
- Implemented a **Kafka** flow to publish APOLLO lead messages.
- Implemented data quality controls to identify discrepancies, inconsistencies and missing data.
- Developed production **Databricks Apps**, deployed through **Asset Bundles**, with RLS applied on views.
- **Results:** delivered a shared data foundation enabling Atlas/SAP and APOLLO coexistence.

*Tools: AWS, Databricks, PySpark, Spark SQL, Delta Live Tables, Auto Loader, Kafka, Terraform, Power BI*

*Cloud Data Engineer · Partial Technical Lead*

02/2024 – 06/2025

*On-Premise Data Platform Migration to AWS / Databricks*

- Migrated **Airflow / Spark / S3 MinIO** workloads from the on-premise platform to **AWS / Databricks**.
- Refactored Airflow DAGs into **Databricks Jobs/Workflows** and developed **Delta Live Tables** pipelines.
- Implemented data cataloging with Unity Catalog, AI-generated comments and system tables for lineage.
- Implemented Databricks governance: Unity Catalog, permissions, RLS, secrets and SCIM via **Terraform**.
- **Results:** industrialized the Databricks platform, supported SOCOTEC US/UK teams and reduced average processing runtime.

*Tools: AWS, Databricks, PySpark, Spark SQL, Delta Live Tables, Airflow, Fivetran, Terraform*

*Data Engineer · Data Project Lead*

02/2023 – 02/2024

*Finance Analytics International Program*

- Developed **Airflow DAGs** and PySpark/SQL/dbt transformations to unify group finance reporting.
- Integrated ATLAS ERP, Oracle DWH and API data, with FX rates ingested through the **Kyriba API**.
- Built **Power BI** dashboards with DAX measures, business visuals and RLS by department/country.
- **Results:** unified finance reporting across **18 countries / 12 currencies** and mentored **2 apprentices**.

*Tools: Apache Airflow, PySpark, SQL, dbt, S3 MinIO, Delta Lake, Kyriba API, Power BI, DAX, Terraform, CI/CD*

*Administrative Performance Cockpit*

- Designed and delivered a cockpit to monitor the *lead* → prospecting → order cycle, reduce delays, control backlog and standardize administrative KPIs.
- Developed an end-to-end data pipeline with **Apache Airflow**, PySpark, Spark SQL, S3 and Delta Lake.
- Computed business indicators (**DSO**, overdue rate, **D+1 SLA**, backlog) and exposed them in **Power BI**.
- **Results:** improved operational monitoring of the administrative cycle through **Terraform** / **GitLab CI/CD** industrialization and proactive alerts.

*Tools: Terraform, GitLab CI/CD, Apache Airflow, PySpark, Spark SQL, Delta Lake, S3, Power BI***Data Engineer — xHub, Casablanca, Morocco (client: M2T / Banque Populaire)**

01/2022 – 07/2022

*Client Scoring / Decision Data Foundation*

- Implemented a data system for client scoring to improve the reliability of credit decisioning.
- Designed a data lake feeding pipeline: collection, normalization and automation of processing workflows.
- Developed a **Flask API** to expose KPIs to scoring tools, with integration on **MongoDB**.
- **Results:** centralized client KPIs, removed manual handling and provided near real-time indicator access for business teams.

*Tools: Hadoop, PySpark, Apache Airflow, MongoDB, PostgreSQL, Docker, Flask, Git, Python***Data Engineer — DeliverMe, Tangier, Morocco**

03/2021 – 09/2021

- Built an analytics foundation for route analysis and logistics optimization.
- Migrated **MySQL** / **Cloud SQL** data to **BigQuery** through **Apache Airflow**, structured on **GCS** and productionized in a cloud analytics architecture.
- Monitored DAGs with email / Slack alerting and exposed data in **Apache Superset**.
- **Result:** deployed self-hosted Apache Airflow through **Docker** on **Compute Engine**, reducing costs by about **80%** compared with a managed solution.

*Environment: Apache Airflow, GCP (BigQuery, GCS, Cloud SQL, Compute Engine), Superset, Docker, Git, Python*PERSONAL PROJECTS

---

**MonÉlu — Civic Transparency Platform**

2025 – Present

*Civic-tech platform based on open data from the French National Assembly*

- Built an end-to-end civic platform: ingestion of public parliamentary APIs and a chatbot to query each elected official's activity in natural language.
- Built a real-time streaming pipeline with **Redpanda** (Kafka-compatible) and **Spark Structured Streaming** for vote ingestion and alerts under 5 minutes, with user-configurable rules and data quality using **dbt Core + Soda**.
- Built a **RAG** pipeline: legislative text chunking, embeddings, vector storage in **pgvector** and a **LangChain** query chain using local **Mistral 7B via Ollama**.

*Environment: LangChain, pgvector, MLflow, Ollama, Mistral 7B, Soda, Apache Airflow, Spark, dbt Core, Redpanda, Docker, GitHub Actions, Claude Code*EDUCATION

---

**ENSMR – National School of Mines of Rabat**

2019 – 2022

State Engineering Degree – Management of Information Systems

**CPGE – Lycée Mohammed V, Beni Mellal, Morocco**

2017 – 2019

Preparatory Classes for Engineering Schools – Mathematics and Physics

CERTIFICATIONS

---

**Obtained:** Astronomer Certification – Apache Airflow Fundamentals**In progress:** Databricks Certified Data Engineer Associate · AWS Certified Solutions Architect – AssociateLANGUAGES

---

French – fluent · English – fluent