



# Walid EL KHOUKH

Cloud Data Engineer

4+ ans d'expérience en data engineering, cloud, lakehouse

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

☎ +33 6 50 04 37 94

📍 Île-de-France, France

🌐 [LinkedIn](#)

🐙 [GitHub](#)

## PROFIL

Cloud Data Engineer avec 4+ ans d'expérience dans la conception, la migration et l'industrialisation de plateformes data sur AWS et GCP. A piloté la migration de 50+ pipelines data vers Databricks avec une réduction de 50% du runtime et 25% des coûts cloud.

## COMPÉTENCES TECHNIQUES

**Langages :** Python, SQL, Bash

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

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

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

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

**Bases de données :** PostgreSQL, pgvector, MySQL, MongoDB, DuckDB, Oracle

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

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

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

## EXPÉRIENCES

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

02/2023 – Présent

*Cloud Data Engineer · Lead technique partiel*

06/2025 – Présent

*APOLLO — Migration ERP Atlas vers SAP / S/4HANA*

- Construction de pipelines/jobs **Databricks** pour l'ingestion et la transformation des données SAP via Datasphere.
- Mise en place d'un flux **Kafka** pour publier des messages de leads APOLLO.
- Mise en place de contrôles de qualité pour identifier les écarts, incohérences et données manquantes.
- Développement de **Databricks Apps** en production, déployées via **Asset Bundles**, avec RLS sur les vues.
- **Résultats :** Mise à disposition d'un socle data commun permettant la coexistence Atlas/SAP et APOLLO.

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

*Cloud Data Engineer · Lead technique partiel*

02/2024 – 06/2025

*Migration de la plateforme data on-premise vers AWS / Databricks*

- Migration de traitements **Airflow / Spark / S3 MinIO** depuis la plateforme on-premise vers **AWS / Databricks**.
- Refonte de DAGs Airflow en **Databricks Jobs/Workflows** et développement de pipelines **Delta Live Tables**.
- Mettre en place le catalogage des données avec Unity Catalog, génération des commentaires avec l'IA et tables system pour le lineage.
- Mise en place de la gouvernance Databricks : Unity Catalog, permissions, RLS, secrets et SCIM via **Terraform**.
- **Résultats :** Plateforme Databricks industrialisée, Support aux équipes SOCOTEC US/UK et réduction du runtime moyen des traitements.

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

*Data Engineer · Chef de projet data*

02/2023 – 02/2024

*Programme Finance Analytics International*

- Développement de **DAGs Airflow** et traitements PySpark/SQL/dbt pour unifier le reporting finance groupe.
- Intégration des données ATLAS ERP, DWH Oracle et API, avec ingestion des taux de change via **API Kyriba**.
- Développement de dashboards **Power BI** avec mesures DAX, visuels métier et RLS par direction/pays.
- **Résultats :** reporting finance unifié sur **18 pays / 12 devises** et encadrement de **2 alternants**.

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

- Conception et livraison d'un cockpit de pilotage du cycle *lead* → prospection → commande afin de réduire les délais, maîtriser le backlog et standardiser les KPI administratifs.
- Développement d'un pipeline data end-to-end avec **Apache Airflow**, PySpark, Spark SQL, S3 et Delta Lake.
- Calcul des indicateurs métier (**DSO**, taux d'échu, **SLA J+1**, backlog) et restitution dans **Power BI**.
- **Résultats** : amélioration du pilotage opérationnel du cycle administratif avec industrialisation **Terraform** / **GitLab CI/CD** et alertes proactives.

*Environnement* : Terraform, GitLab CI/CD, Apache Airflow, PySpark, Spark SQL, Delta Lake, S3, Power BI

- Mise en place d'un système data pour le scoring client afin de fiabiliser la décision d'octroi de crédit.
- Conception d'un pipeline d'alimentation du datalake : collecte, normalisation et automatisation des traitements.
- Développement d'une **API Flask** pour l'exposition des KPI aux outils de scoring, avec intégration sur **MongoDB**.
- **Résultats** : centralisation des KPI clients, suppression des manipulations manuelles et accès quasi temps réel aux indicateurs pour les équipes métier.

*Environnement* : Hadoop, PySpark, Apache Airflow, MongoDB, PostgreSQL, Docker, Flask, Git, Python

- Mise en place d'un socle analytique pour l'analyse et l'optimisation des trajets logistiques.
- Migration des données **MySQL** / **Cloud SQL** vers **BigQuery** via **Apache Airflow**, avec structuration sur **GCS** et mise en production dans une architecture cloud orientée analytics.
- Supervision des DAGs avec alerting email / Slack et exposition des données dans **Apache Superset**.
- **Résultat** : déploiement d'Apache Airflow auto-hébergé via **Docker** sur **Compute Engine**, réduction des coûts d'environ **80%** par rapport à une solution managée.

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

## PROJETS PERSONNELS

- Création d'une plateforme citoyenne de bout en bout : ingestion des APIs parlementaires publiques, et mise en place d'un chatbot permettant d'interroger l'activité de chaque élu en langage naturel.
- Pipeline de streaming temps réel avec **Redpanda** (Kafka-compatible) et **Spark Structured Streaming** pour l'ingestion et l'alerting sur les votes en moins de 5 minutes, avec moteur de règles configurable par utilisateur, en assurant la qualité des données avec **dbt Core** + **Soda**.
- Pipeline **RAG** : chunking des textes législatifs, embeddings, stockage vectoriel **pgvector** et chaîne **LangChain** pour l'interrogation avec **Mistral 7B** via **Ollama** comme LLM local.

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

## FORMATION

## CERTIFICATIONS

## LANGUES