Yassine OUZAINA

Paris, France • work@youzaina.fr
youzaina.fr • linkedin.com/in/youzaina001 • github.com/youzaina001

Data Engineer with 3+ years building and operating ETL pipelines and self-service data products.
Python, SQL, Airflow and Databricks: metadata and quality across biomedical and product data.
Authorized to work in France (Passeport Talent).

EXPERIENCE

Owkin Paris, France
Data Engineer, Metadata & Interoperability Jul 2025 to Present

- Co-designed a dimensional model and built daily Polars/DuckDB catalog pipelines (7 sources, about 26 PostgreSQL fact/dimension tables), enabling about 35 research, partnership and finance users to query cohorts, QC and costs via Metabase and Metabot/Bedrock.
- Built a Python ontology client and CI-driven importer to synchronize published definitions into a shared data-contract registry (921 canonical variables, 24 modalities), replacing manual update pull requests.
- Migrate legacy pipelines and analytics workloads to Databricks on AWS (Unity Catalog, Delta Lake, Workflows, Asset Bundles) to reduce custom Python clients and simplify client data integration; deliver recurring datasets for Owkin K Pro.
- Operate idempotent Airflow/MWAA ETL with schema, completeness and referential-integrity checks; broadened date-format support, filtering unparseable records with quality warnings.
- Manage AWS infrastructure through Terraform, contribute to Airflow 3 migrations and resolve production incidents using Grafana, fixes and rollbacks.

re:mind Paris, France
Data Engineer Sep 2023 to May 2025

- Built PySpark pipelines on Azure Databricks to transform thousands of product manuals and repair guides into structured attributes and searchable datasets.
- Co-developed and deployed an LLM extraction workflow with an ML engineer, validating and prefilling fields from invoices, technical sheets and receipts for operator verification.
- Converted data and ML notebooks into tested Python modules deployed through CI/CD.

Eviden Bordeaux, France
HPC Research Engineer Intern Feb 2023 to Aug 2023

- Profiled CPU, memory and I/O bottlenecks in MPI/OpenMP workloads.
- Automated benchmark collection using Python, C and Bash on Linux.

TECHNICAL SKILLS

Languages & data: Python, SQL, PostgreSQL, DuckDB, Polars, PySpark, SQLAlchemy, Kimball modeling.

Cloud & delivery: AWS (MWAA, S3, RDS, ECS, Glue, CloudWatch, CodeArtifact), Azure Databricks, Terraform, Docker, Kubernetes, CI/CD, Grafana.

Quality & analytics: pytest, Pydantic, Pandera, data contracts, REST APIs, Metabase, LLM workflows.

PERSONAL PROJECTS

airflow-dq-agent: Airflow 3 data-quality remediation engine with allow-listed actions, human approval, and SERIALIZABLE target-fingerprint rechecks (zero model-written SQL).

LedgerLens: FastAPI/PostgreSQL 16 invoice triage workbench with 4 least-privilege database roles, row-level security, and offline evaluation replay.

geocrawl: Python and DuckDB crawl-freshness pipeline comparing archived web snapshots with visible-prose extraction, three-word run deduplication, and SHA-256 manifest provenance.

EDUCATION & CERTIFICATION

ENSEIRB-MATMECA
Engineering degree in Applied Mathematics (Diplôme d'ingénieur, Master's equivalent) • 2023

Astronomer
DAG Authoring for Apache Airflow 3 • Nov 2025

LANGUAGES

French (C2, bilingual) • English (C2, bilingual) • Arabic (native)