Yassine OUZAINA

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

Data Engineer avec plus de 3 ans d'expérience en ETL et produits de données en libre-service.
Python, SQL, Airflow et Databricks : données biomédicales, données produit, métadonnées et qualité.
Autorisé à travailler en France (Passeport Talent).

EXPÉRIENCE PROFESSIONNELLE

Owkin Paris, France
Data Engineer, Metadata & Interoperability Juil. 2025 à Présent

- Co-conçu le modèle dimensionnel et développé les ETL quotidiens Polars/DuckDB (7 sources, environ 26 tables PostgreSQL) : cohortes, qualité et coûts en libre-service pour environ 35 collaborateurs (recherche, partenariats, finance) via Metabase et Metabot/Bedrock.
- Développé le client Python du service ontologique et l'importeur de contrats, synchronisant par CI les définitions de référence (921 variables canoniques, 24 modalités), remplaçant les PR manuelles.
- Migration en cours de pipelines et analyses vers Databricks sur AWS (Unity Catalog, Delta Lake, Workflows, Asset Bundles) pour réduire les clients Python spécifiques et faciliter l'intégration des données clients ; livraisons récurrentes pour Owkin K Pro.
- Exploité des ETL Airflow/MWAA idempotents : contrôles de schéma, complétude et intégrité référentielle ; étendu les formats de dates et filtré les lignes aux dates non reconnues avec alertes qualité.
- Géré l'infrastructure AWS via Terraform, contribué aux migrations Airflow 3 et résolu les incidents de production avec Grafana, correctifs et rollbacks.

re:mind Paris, France
Data Engineer Sept. 2023 à Mai 2025

- Conçu des pipelines PySpark sur Azure Databricks transformant des milliers de manuels produits et guides de réparation en attributs structurés et données indexables.
- Co-développé et déployé avec un ingénieur ML un workflow LLM d'extraction documentaire (factures, fiches techniques, reçus), avec validation des champs et préremplissage pour vérification par les opérateurs.
- Converti des notebooks data et ML en modules Python testés et déployés par CI/CD.

Eviden Bordeaux, France
Ingénieur recherche HPC (stage) Févr. 2023 à Août 2023

- Analysé les limites CPU, mémoire et I/O de charges MPI/OpenMP.
- Automatisé la collecte de benchmarks sous Linux avec Python, C et Bash.

COMPÉTENCES TECHNIQUES

Données : Python, SQL, PostgreSQL, DuckDB, Polars, PySpark, SQLAlchemy, modélisation Kimball.

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

Qualité & analytique : pytest, Pydantic, Pandera, contrats de données, API REST, Metabase, workflows LLM.

PROJETS PERSONNELS

airflow-dq-agent : remédiation Airflow 3 avec actions autorisées, validation humaine et revérification SERIALIZABLE (zéro SQL écrit par le modèle).

LedgerLens : triage de factures FastAPI/PostgreSQL 16 avec 4 rôles SQL cloisonnés, RLS et évaluations rejouables hors ligne.

geocrawl : pipeline Python et DuckDB comparant deux captures web archivées par extraction du texte visible, déduplication sur séquences de trois mots et provenance SHA-256.

FORMATION & CERTIFICATION

ENSEIRB-MATMECA
Diplôme d'ingénieur en mathématiques appliquées (grade de Master, Bac+5) • 2023

Astronomer
DAG Authoring for Apache Airflow 3 • Nov. 2025

LANGUES

Français (C2, bilingue) • Anglais (C2, bilingue) • Arabe (langue maternelle)