

# **Djalil Salah-bey**

### Data Engineer | Expert ETL/ELT (AirFlow, dbt, BigQuery, Databricks)

Ingénieur Data spécialisé dans la modélisation analytique, l'orchestration et la mise en production de pipelines de données en intégrant les bonnes pratiques de DataOps, CI/CD et observabilité.

salahbeydjalil@gmail.com

Numéro de téléphone +33 6 11 27 91 53

Site web

https://djo-project-portfolio.vercel.app/

#### Langues

Français Natif

**Anglais** 

Courant

# Soft Skills

Autonome Curieux Proactif Pédagogue

#### Compétences **Techniques**

- Langages: Python, SQL, PySpark, Bash
- Intégration : Airflow, dbt Cloud, Talend
- DataOps: Azure, GCP, Terraform, Docker, GitHub Actions, Git
- Visualisation: Power Bl, Looker Studio, Metabase
- Stockage Cloud: Databricks, GCS, ADLS, S3, BigQuery
- Bases de données & Moteurs analytiques BigQuery, PostgreSQL, ClickHouse, DuckDB, MongoDB, InfluxDB

### Diplômes et **formations**

- IA School Diplôme RNCP Niveau 7 Expert Data & Ingénierie de l'Intelligence Artificielle (2023-2025)
- IA School Bachelor 3 **Expert Data &** Ingénierie de l'Intelligence Artificielle (2022-2023)
- Université de Lille Licence Banque, Finance & Analyse Quantitative (2019-2022)

### Expériences professionnelles

PowerUp Technology

Freelance Data/Analytics Engineer

De mars 2025 à octobre 2025

Paris

Refonte d'une plateforme BI vers ClickHouse pour des heatmaps 2D/3D thermiques performantes et une navigation hiérarchique fluide (cellule → pack → site → client).

- Définition du grain d'analyse, des faits/dimensions et des hiérarchies spatiales et temporelles avec règles d'agrégation adaptées à ClickHouse.
- Intégration des données depuis Azure Data Lake Storage, gestion des chargements incrémentaux, backfills contrôlés et idempotence des pipelines.
- Conception d'un système de discrétisation spatiale pour agréger les températures et mesures dans des cartes 2D/3D homogènes, garantissant cohérence et précision sur de larges volumétries.
- Génération automatisée de DAGs Airflow via templates Jinja, avec dépendances, SLA, alertes et exécutions multi-clients.
- Création et optimisation de tables (partitionnement, compression, indexation) et de materialized views garantissant des latences <100 ms sur les requêtes analytiques.
- Mise en place d'un modèle analytique dbt avec documentation automatisée, tests (unicité, fraîcheur, cohérence) et CI GitHub Actions pour validation continue.
- Création d'un monorepo Databricks Repos pour standardiser le développement (Docker, conventions, packaging, linting) et uniformiser le déploiement.
- Publication des heatmaps 2D/3D complètes avec toutes les fonctionnalités métiers demandées (zoom, filtres, agrégation dynamique, hiérarchies interactives).
- Mise en place de data contracts, tests SQL et monitoring Grafana pour garantir la fiabilité opérationnelle.
- Stack: ClickHouse, SQL, Python, dbt, Airflow, Databricks, Docker, ADLS, Grafana, Power Bl, Git. Résultat: Migration end-to-end livrée: modèle OLAP ClickHouse performant et gouverné, heatmaps 2D/3D interactives conformes aux exigences métier, pipelines automatisés et industrialisés (Airflow + CI).

Koacher De décembre 2022 à décembre 2024

Data Engineer

- Développement de pipelines PySpark orchestrés sous Airflow, intégrant des flux issus des backoffices applicatifs et d'APIs clients, avec stockage sur Azure Data Lake.
- Implémentation d'une architecture Medallion sur Databricks avec Delta Lake, structurant l'ensemble des traitements d'ingestion, de nettoyage et de transformation des données.
- Préparation et normalisation des données dans la couche Silver, incluant les contrôles de qualité, filtrage et formatage afin de garantir la conformité des jeux de données.
- Calcul et agrégation des KPI de performance, d'usage et de sécurité dans la couche Gold à l'aide de PySpark, assurant la standardisation des indicateurs clés exposés aux équipes Produit et BI.
- Automatisation et orchestration des pipelines sous Airflow, avec gestion des dépendances, des SLA et des notifications, assurant une exécution fiable et planifiée des flux de données.
- Validation systématique des jeux de données par tests automatisés (Great Expectations, Pytest) et suivi des exécutions dans Grafana et Prometheus, garantissant la fiabilité opérationnelle du pipeline.
- Mise en place d'une chaîne CI/CD complète sous GitHub Actions, connectée à Databricks Repos, et déploiement de l'infrastructure via Terraform avec cohérence et reproductibilité des environnements.
- Préparation, versionnage et exposition des datasets d'entraînement dans le Feature Store Databricks assurant la reproductibilité et la traçabilité des modèles IA avec les équipes Data Science.
- Exposition des tables Gold à l'équipe BI pour la création de tableaux de bord Power BI (KPI : rétention, churn, activité utilisateur) et harmonisation des indicateurs avec les équipes Produit.
- Stack: Python, PySpark, SQL, DBT, Databricks, Delta Lake, Azure Data Lake, MongoDB, Airflow, Terraform, GitHub Actions, Great Expectations, Grafana, Prometheus, MLflow, Power Bl.