Djalil Salah-bey

Data Engineer | Recherche CDI

Data Engineer spécialisé Azure & Databricks, passionné par la conception de plateformes Lakehouse et l'industrialisation de pipelines Data & ML.
Mon approche combine fiabilité, performance et automatisation, avec un focus constant sur la qualité, la traçabilité et la mise en producti

Email
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

Passion

Sport
Programmation

Compétences Techniques

Langes
Python · PySpark · SQL
Cloud & Plateformes
Azure Databricks · Azure Data Lake Storage · Azure Synapse · Azure Machine Learning · Azure Purview · Snowflake · Delta Lake · Databricks Workflows
Orchestration & CI/CD
Airflow (AKS) · Azure Data Factory · GitHub Actions · Azure DevOps · Docker · Kubernetes · dbt Core
Machine Learning & MLOps
MLflow · scikit-learn · XGBoost · LightGBM · Great Expectations
Visualisation & Gouvernance
Power BI · Marquez · Unity Catalog

Diplômes et formations

- **IA School —** Diplôme RNCP Niveau 7 Expert Data & Ingénierie de l'Intelligence Artificielle (2022-2024)
- **IA School —** Bachelor 3 Expert Data & Ingénierie de l'Intelligence Artificielle (2021-2022)
- **Université de Lille —** Licence Banque, Finance (2018-2021)

Soft Skills

- Curieux
- Proactif
- Rigoureux

Expériences professionnelles

PowerUp Technology — **Paris**
Freelance Data & AI Engineer — De mars 2025 à octobre 2025

Renforcer l'équipe Data & AI pour concevoir et industrialiser une **plateforme Lakehouse Databricks / GCP** dédiée à la détection d'anomalies et à la prédiction de dégradation sur les batteries industrielles (BESS).

- Conception d'un **Lakehouse Databricks** reposant sur **Azure Data Lake Storage (ADLS)** et **Delta Lake**, structuré selon une architecture **Bronze / Silver / Gold** pour assurer fiabilité et gouvernance des données.
- Développement de **pipelines de préparation et de feature engineering** en **PySpark / SQL**, orchestrés sous **Airflow (AKS)** et **Databricks Workflows**.
- Entraînement et suivi de **modèles prédictifs** (scikit-learn, XGBoost) avec **MLflow**, incluant le tracking des métriques, le versioning et la gestion des artefacts modèles.
- Déploiement du **pipeline MLOps complet** : ingestion, entraînement, validation et déploiement automatisés sur **Azure Machine Learning**.
- Intégration de la **CI/CD (GitHub Actions)** pour le déploiement des notebooks, configurations et modèles.
- Mise en place de tests de **qualité et de reproductibilité** (Great Expectations, PyTest) et optimisation des jobs Spark (partitionnement, caching, broadcast join).

Stack : **Azure Databricks (PySpark, Delta Lake, MLflow)** · **Azure Data Lake Storage (ADLS)** · **Azure Machine Learning** ·**Airflow (AKS)** · **Python (scikit-learn, XGBoost)** · **SQL** · **GitHub Actions (CI/CD)** · **Great Expectations** · **Docker** · **Kubernetes**

Koacher — **Lyon**
Alternant Data Engineer — De décembre 2022 à décembre 2024

- Conception et exploitation d'un **Customer Data Lake (CDL)** sur **Azure Data Lake Storage**, structuré en **Bronze / Silver / Gold** pour assurer la fiabilité et la scalabilité des flux.
- Intégration du **catalogue Purview** et mise en conformité **RGPD**, incluant la gestion des **PII** et la traçabilité complète du **lineage** via **YAML + Marquez**.
- Mise en place de **standards de qualité et de nomenclature** avec **dbt tests** et documentation automatisée, accompagnés d'une gestion des accès sécurisée (**RBAC**).
- Déploiement d'**Airflow sur AKS (Azure Kubernetes Service)** pour orchestrer les pipelines de données à grande échelle : monitoring, alerting, reprise automatique et gestion des dépendances.
- Développement de **pipelines ETL/ELT hybrides** sous **Airflow** et **Azure Data Factory**, automatisant l'ingestion depuis les **APIs applicatives**, **Hubspot**, **Firebase** et **Stripe**, avec logs centralisés et reprise sur incident.
- Industrialisation des modèles dbt (SQL Jinja) avec **logique incrémentale, historisation SCD2** et **optimisation du lineage** afin d'assurer des transformations performantes et auditées sur Snowflake.
- Optimisation de **Snowflake** (warehouse sizing, clustering keys, micro-partitioning, Time Travel) pour améliorer les performances et réduire les coûts d'exécution.
- Automatisation de la **CI/CD dbt + Power BI** via **Azure DevOps** : exécution des tests, déploiement des modèles et génération continue de la documentation.
- Développement de **dashboards Power BI** connectés en **Direct Query à Snowflake**, reposant sur un **modèle sémantique centralisé** et des **mesures DAX** normalisées.
- Construction d'un **catalogue KPI** (rétention, churn, LTV, engagement, conversion) partagé et validé avec les équipes produit et marketing.
- Refonte du **process de refresh Power BI** (sécurité, dépendances, performance) garantissant des tableaux de bord actualisés en moins de 15 minutes.
- Mise en place d'un **système de suivi qualité automatisé** (ADF + Airflow + Slack + Power BI) assurant la supervision proactive des pipelines.
- Documentation complète du **lineage** et des dépendances sous **Marquez**, pour une visibilité totale sur les flux.
- **Réduction de 35 % des coûts Snowflake** et **-50 % d'incidents d'ingestion** grâce à l'optimisation des pipelines et au refactoring SQL.

Stack : Azure Data Lake Storage, **Azure Data Factory**, **Airflow (AKS)**, **Snowflake**, **dbt Core**, **Power BI (DAX)**, **SQL avancé**, **Python (ETL)**, **Azure DevOps**, **Purview**, **Marquez**, **GitHub Actions.**