

Big Data Engineer - Azure Fabric / Spark / DBT / Snowflake

Expériences professionnelles

Avril 2024 –  **France TV Pub**, Paris, Data Engineer.
En poste


- ETL : Conception de workflows Spark évolutifs (Bronze, Silver, Gold) sur Microsoft Fabric (Azure Data factory, Spark)
- Traitement et analyses distribuées big data (python, pySpark)
- POC - Streaming de données (Kafka, Airflow, Azure et Docker-Compose).
- POC - Conception d'un pipeline ELT avec DBT (modèles SQL, tests de données, documentation) sur Snowflake
- Monitoring : Création de dashboards de suivi de la consommation des capacités Unity (Power BI).
- Data science : Développement de modèles de recommandation pour optimiser les propositions commerciales (filtrage collaboratif, content-based, factorisation matricielle).

JUL 2023 –  **Sopra Steria**, Lille, DataOps.
Mars 2024

- intégration IA : Conception d'une application web avec Flask et intégration de LLM via les APIs OpenAI et AWS Bedrock.
- CI/CD : Configuration de pipelines Jenkins CI/CD pour le déploiement continu sur des clusters Kubernetes, avec intégration de tests automatisés.
- Monitoring : Déploiement d'une stack ELK (Elasticsearch, Logstash, Kibana) sur Kubernetes, mise en place d'un système de monitoring et d'alertes avec Grafana.
- Culture DevOps : Suivi des bonnes pratiques d'intégration applicative, fiabilisation des déploiements et amélioration continue de l'environnement technique.

Nov 2022 – **KANTAR**
Mars 2023 **Kantar Worldpanel**, Paris, Data scientist.

- Analyse de données comportementales : Statistiques descriptives, segmentation (K-means), détection d'anomalies et analyse de corrélations/règles d'association (R, arules, Tidy Anomaly Detection)
- Data visualisation et reporting : Création de visualisations interactives pour l'analyse de tendances (Python, Matplotlib, Seaborn, PowerBI).

Mai 2022 –  **Mutualité Sociale Agricole**, Lille, Stage.
Oct 2022

- Conception et mise en place d'une base de connaissances provenant du big data à l'aide de scripts SQL élaborés (Hive, HDFS, SQL).
- Réalisation d'analyses en composantes principales (ACP) sur des données volumineuses (R, Python).
- Clustering des entreprises selon les niveaux de risque avec K-means (Python / Apache Spark Scala).
- Application de modèles ML pour la détection des salariés en situation de désinsertion professionnelle (Régression logistique, SVM, Gradient Boosting).

Compétences techniques

Ingenierie des données

- **ELT**: dbt, Fabric (Azure data factory), Airflow, dbt
- **Cloud & Stockage** : AWS (S3, Glue), Azure (SQL database,), Oracle
- **Modélisation & DWH** : Etoile / flocon, Lakehouse (Fabric, Delta Lake), Snowflake
- **Bases de données** : PostgreSQL, MongoDB, MySQL, DynamoDB

Big Data & Traitement

- **Frameworks** : Apache Spark (PySpark), Hadoop, Kafka
- **Streaming & Batch** : Kafka Streams, Spark Structured Streaming, Pub/Sub
- **Data Formats** : Parquet, Avro, ORC, JSON

Développement & Ops

- **Langages** : Python, SQL, Bash
- **DevOps & CI/CD** : Docker, Kubernetes, Github, Jenkins
- **Tests & Qualité** : Pytest, Great Expectations, dbt tests

Formation

- 2024–2025 **Mastère Spécialisé Expert Big Data Engineer**, *Université de technologie de Troyes*, France
- 2019–2022 **Master 2 Mathématique et informatique appliqués parcours Data Science**, *Université de Lille*, France
- 2016–2019 **Licence Mathématiques et Informatiques**, *UT*, Sénégal

Langues

- Français Niveau C2
- Anglais Niveau C1