

DATA ENGINEER +4 ans d'expérience

Data Engineer spécialisé dans la conception et l'industrialisation de plateformes data cloud sur GCP, les architectures Lakehouse / Delta Lake.

Spécialisé dans la construction de pipelines de données robustes et scalables avec Python, Spark, SQL et Airflow, j'interviens sur l'ensemble de la chaîne de valeur data : ingestion, transformation, orchestration, stockage, scalabilité et mise à disposition des données.

J'ai également une expérience significative dans la mise en place d'environnements industrialisés sur Google Cloud Platform (BigQuery, Cloud Storage, Pub/Sub, Dataproc), l'automatisation CI/CD et l'Infrastructure as Code avec Terraform.

FORMATIONS & CERTIFICATIONS

2020 : INSA - Master, Lyon

2018 : Diplôme d'ingénieur, Ecole Mohammadia des Ingénieurs

COMPETENCES FONCTIONNELLES & TECHNIQUES

Data engineering : Python, PySpark, Spark Streaming, SQL, Airflow, DBT

Cloud & Big Data : GCP : BigQuery, Dataproc, Cloud Storage, Pub/Sub, Cloud Run

Data Platform & Stockage : Architecture Lakehouse, Delta Lake, PostgreSQL, Snowflake, BigQuery

DevOps & Infrastructure : Terraform, Terragrunt, Docker, Kubernetes, GitHub, Gitlab CI/CD

Data Quality & Gouvernance : Data Quality, monitoring pipelines, documentation technique

Visualisation : Power BI, Looker Studio

Machine Learning : Scikit-Learn, TensorFlow, MLflow

Langues : Anglais courant

Secteurs d'activités : Industrie, Transport / Logistique

EXPERIENCES PROFESSIONNELLES

MAMATA (éditeur de logiciel), de Juin 2024 à Aujourd'hui
Data Engineer

Contexte du projet :

- Optimisation et industrialisation des pipelines d'ingestion de données time series provenant de différentes APIs avec réduction de 30% des temps d'exécution.
- Industrialisation des workflows data
- Mise en place des pipelines CI/CD avec GitHub Actions et Terraform.
- Création de fonctions d'extraction et d'applications Python Spark déployées sur Cloud Run / Cloud Functions de GCP
- Analyse des besoins métiers pour la modélisation des données et définition des KPIs de production issus de la transformation des données
- Gouvernance des données : sécurité et de conformité des données à l'échelle de l'organisation. En mettant en place une stratégie de gouvernance efficace, les entreprises peuvent mettre leurs données au service d'une prise de décision axée sur les données, tout en les protégeant des accès non autorisés et en assurant leur conformité réglementaire.

- Conception et orchestration de pipelines ELT sur GCP avec Airflow pour l'ingestion, la transformation et le stockage des données dans BigQuery et Google Cloud Storage (GCS).
- Élaboration de dashboards de suivi des KPIs avec Looker Studio.

Environnement technique :

PYTHON, SPARK, GITHUB ACTIONS CI/CD, AIRFLOW, COGNITE DATA FUSION, PYDANTIC, GRAPHQL, BIGQUERY, DBT, CLOUD RUN, LOOKER STUDIO, GOOGLE CLOUD STORAGE (GCS), PUB/SUB, DOCKER, TERRAFORM, KAFKA

SETEC - ANDRA, de Mai 2022 à Mai 2024

Data Engineer

Contexte du projet :

- Conception de l'architecture Lakehouse / Delta Lake de collecte des données avec PostgreSQL et automatisation du processus d'acquisition depuis différentes sources.
- Mise en place de pipelines ETL avec Airflow, DBT et Databricks pour transformer les données brutes et calculer les KPIs définis par les besoins métiers.
- Réalisation de dashboards interactifs Power BI intégrant les visualisations par batchs temporels des données issues des capteurs dans le data warehouse.
- Analyse statistique des relations entre paramètres métiers et mise en place de modèles de Machine Learning, notamment régression et Random Forest Regressor.
- Conception et développement d'APIs avec FastAPI
- Intégration et livraison continues pour l'industrialisation et le déploiement des modèles.

Environnement technique :

PYTHON, PANDAS, MLFLOW, SQL, GITHUB ACTIONS CI/CD, GITLAB CI/CD, DOCKER, POWER BI, AIRFLOW, DBT, DATABRICKS, DELTA LAKE, SNOWFLAKE

GEOS, de Décembre 2020 à Avril 2022

Data Analyst

Contexte du projet :

Création d'une cellule de pilotage multi-client incluant un acteur majeur du secteur assurance (Direct Assurance).

- Automatisation du processus d'acquisition des données depuis différentes sources et prétraitement des données dans une logique de Data Quality.
- Analyse statistique des données issues des résultats d'essais géotechniques.
- Établissement de modèles géologiques à partir des données et automatisation des calculs géomécaniques aux éléments finis selon les modèles établis.

Environnement technique :

POSTGRES SQL, POWER BI, GATEWAY, MACHINE LEARNING