PRÉNOM NOM

DATA ENGINEER SENIOR, PLATEFORMES LAKEHOUSE, SPARK ET DBT

9 ans Expérience data engineering	**12 To/j** Volume traité quotidiennement	**380** Modèles dbt industrialisés
62 % Réduction coûts cloud	**99,8 %** Disponibilité des pipelines	**26** Data engineers formés

PROFIL

Je construis et j'exploite des plateformes de données depuis 9 ans, du flux d'ingestion jusqu'au modèle servi en production. Mon terrain de jeu : les volumétries qui font craquer les outils classiques, entre 5 et 15 To par jour, sur des architectures Spark et Databricks. Je modélise avec dbt des socles que les équipes métier comprennent sans traducteur, pas seulement des tables techniques. Je regarde aussi la facture : sur mes 3 dernières missions, j'ai divisé par 2 le coût de calcul du socle data sans dégrader la fraîcheur. Ce qui m'intéresse, c'est la donnée qui sert une décision réelle, tarification, prévision de stock, détection de fraude. Je travaille en squad agile, je documente, je transmets. Et je reste joignable quand le pipeline casse à 3 heures du matin.

COMPÉTENCES CLÉS

Apache Spark (PySpark, Spark SQL, tuning de jobs) · dbt (modélisation, tests, macros, exposures) · SQL avancé et optimisation de requêtes · Python data (pandas, PyArrow, Polars) · Airflow et orchestration de pipelines · Kafka et ingestion temps réel

Databricks, Delta Lake, Apache Iceberg · Snowflake et BigQuery · Modélisation dimensionnelle (Kimball, Data Vault) · Qualité de donnée et contrats de données · Terraform et infrastructure as code · Docker, Kubernetes, CI/CD GitLab

Machine learning supervisé (scikit-learn, XGBoost) · MLOps (MLflow, suivi de dérive, réentraînement) · FinOps data et réduction des coûts de calcul · Observabilité des pipelines et gestion d'astreinte · RGPD, anonymisation et cloisonnement des données · Animation d'ateliers métier et transfert de compétences

EXPÉRIENCE PROFESSIONNELLE

Data Engineer senior, socle temps réel et migration Lakehouse

janvier 2023 à aujourd'hui

[Nom du client] est un acteur de la distribution spécialisée : 1 200 magasins, 18 millions de clients encartés, 4 milliards d'euros de chiffre d'affaires. La direction data compte 40 personnes réparties en 6 squads, la mienne couvre le socle d'ingestion et le référentiel client. L'existant reposait sur un entrepôt on premise saturé, avec des batchs de nuit qui débordaient sur les heures ouvrées 3 jours par semaine. L'enjeu : basculer en Lakehouse cloud et servir la donnée magasin en moins de 5 minutes, sans couper les 200 rapports déjà en production.

- Concevoir l'architecture d'ingestion temps réel de 140 flux magasins et e-commerce, de Kafka jusqu'aux tables Delta, pour alimenter le pilotage des ventes à J+0
- Migrer 380 traitements Spark et 2 400 tables depuis l'entrepôt historique vers Databricks, par lots de 30 traitements, avec une phase de double run de 6 semaines
- Industrialiser la qualité de donnée avec 620 tests automatisés sur les tables critiques, alerte en moins de 2 minutes et gel du pipeline en cas de rupture de contrat
- Optimiser le coût de calcul en retravaillant partitionnement, formats de fichiers et dimensionnement des clusters sur les 25 jobs les plus consommateurs
- Encadrer 4 data engineers et 2 alternants, revues de code hebdomadaires et standard de développement Spark adopté par les 6 squads

Résultats : Fraîcheur de la donnée magasin ramenée de 14 heures à 4 minutes sur 100 % des 1 200 points de vente · Coût de calcul réduit de 62 %, soit 480 000 euros par an, pour une volumétrie multipliée par 1,8

Environnement technique : Apache Spark · Databricks · Delta Lake · dbt · Airflow · Kafka · Terraform · Azure · GitLab CI

Analytics Engineer, modélisation décisionnelle et déploiement dbt

mars 2020 à décembre 2022

[Nom du client], mutuelle santé de 900 salariés et 1,4 million d'adhérents, voulait sortir de 15 ans de reporting fabriqué sous Excel. 3 directions métier, actuariat, souscription et relation client, produisaient chacune ses indicateurs, avec 4 définitions concurrentes du taux de résiliation. J'ai rejoint une équipe data de 8 personnes, en Scrum, sprints de 2 semaines. Le sujet n'était pas technique au départ : il fallait poser un langage commun avant de poser des tables.

- Modéliser le socle décisionnel en étoile, 22 tables de faits et 60 dimensions, à partir de 6 systèmes sources dont un mainframe de 1998
- Déployer dbt de zéro : structure de projet, conventions de nommage, 340 modèles, 900 tests et documentation regénérée à chaque merge
- Animer 14 ateliers de définition avec les métiers jusqu'à figer 45 indicateurs dans un dictionnaire versionné
- Automatiser l'orchestration sous Airflow, 70 DAG, avec reprise sur incident et engagement de service affiché par domaine
- Former 12 analystes à SQL et dbt, 6 sessions de 3 heures, puis accompagnement en binôme pendant 2 mois

Résultats : Production du reporting mensuel ramenée de 9 jours à 6 heures · Écarts de chiffres entre directions supprimés : 45 indicateurs, 1 seule définition validée en comité

Environnement technique : BigQuery · dbt · Airflow · Python · SQL · Looker · Git · Docker

Data Scientist, prévision de la demande et scoring transporteurs

septembre 2017 à février 2020

[Nom du client] opérait une place de marché logistique : 30 000 transporteurs référencés, 6 millions de commandes par an. L'équipe data comptait 5 personnes rattachées au produit, avec un cycle de livraison de 3 semaines. Les prix étaient fixés à la main par 20 chargés d'affaires, avec une marge très variable selon les corridors. Ma mission portait sur 2 sujets : prévoir la demande à 7 jours et scorer la fiabilité des transporteurs.

- Construire un modèle de prévision de la demande à 7 jours sur 400 corridors, en gradient boosting, réentraîné chaque nuit sur 4 ans d'historique
- Développer un score de fiabilité transporteur à partir de 60 variables, exposé en API et interrogé 12 000 fois par jour
- Préparer les jeux d'entraînement avec Spark, 1,2 milliard de lignes d'événements nettoyées, dédoublonnées et agrégées
- Mettre en place le suivi de dérive : contrôle hebdomadaire et alerte au-delà de 5 points d'écart sur l'erreur moyenne
- Présenter les résultats aux équipes commerciales chaque mois, en cadrant ce que le modèle sait faire et ce qu'il ne sait pas faire

Résultats : Erreur de prévision réduite de 31 % face à la référence historique, sur les 400 corridors · Taux de litige en baisse de 18 % en 12 mois grâce au score de fiabilité

Environnement technique : Python · scikit-learn · XGBoost · Apache Spark · MLflow · Airflow · FastAPI · PostgreSQL

FORMATION

Diplôme d'ingénieur, spécialité informatique et science des données · [Nom de l'école] 2012 à 2015

Master 2 statistique et apprentissage automatique · [Nom de l'université] 2015 à 2016

CERTIFICATIONS

- Databricks Certified Data Engineer Professional
- Google Cloud Professional Data Engineer
- dbt Analytics Engineering Certification

LANGUES

Français, langue maternelle · Anglais, courant (C1), missions menées en équipe internationale

CV réalisé avec Missionlib · missionlib.fr