Antoine Gélin

Data Engineer — Modern Data Stack · 5 ans d'expérience

Paris antoineg@alt144.net 07 86 83 13 75
antoinegelin.alt144.net antoinegelin gelinant

Synthèse

Data Engineer avec 5 ans d'expérience — ingestion, modélisation, entrepôts de données, couches BI. Background Java/Spark sur de gros volumes en production 24/7 dans des groupes bancaires, aujourd'hui centré sur la Modern Data Stack (Python, dbt, Redshift, BigQuery, AWS, GCP).

Mon terrain : reprendre des pipelines legacy fragiles et les reconstruire en modèles modulaires, testés et documentés, exploitables par les équipes BI et métier — jusqu'à la transmission à l'équipe interne.

Domaines d'intervention

Refonte et modernisation de plateformes data (ETL legacy → ELT / Modern Data Stack)

Conception d'architectures en couches raw / staging / marts, raw rejouable

Modélisation de données (dimensionnelle, SCD, marts métier) et industrialisation dbt (tests, documentation, CI)

Pipelines distribués Spark sur gros volumes, optimisation et support de production

Qualité, gouvernance et sécurité des données (tests, monitoring, RBAC, masking RGPD)

Préparation du self-service BI, documentation et montée en compétences des équipes

Secteurs d'activité

Banque & Finance: Calcul réglementaire (risque), moteurs de règles métier — Natixis (Groupe BPCE), BNP Paribas

Fintech / Services: Plateforme data et BI — Keobiz

Mobilité: Open data Vélib Métropole — projet open source

Compétences techniques

Langages: Python, SQL avancé (fenêtrage, CTE, optimisation), Java, Bash

Data Engineering: dbt (modélisation, tests, documentation, CI), Apache Spark, Hadoop/Hive, Airflow, ingestion Python (sinks modulaires, stateful), stratégies incrémentales, rejeu et idempotence

Modélisation: Modélisation dimensionnelle (faits, dimensions, grain, historisation SCD, star schema), couches raw / staging / intermediate / marts, choix des clés de partitionnement et de jointure

Cloud AWS: Redshift Serverless, S3, Fargate, EventBridge, Secrets Manager, QuickSight, IAM

Cloud GCP: BigQuery, GCS, Cloud Run Jobs, Cloud Scheduler, Cloud Workflows, Secret Manager, Artifact Registry, IAM

Bases de données: Redshift, BigQuery, PostgreSQL / RDS, SQL Server, Hive / ORC

BI & restitution: QuickSight, Power BI, Streamlit

DevOps / DataOps: Terraform, Docker, CI/CD (Jenkins, GitHub Actions, Forgejo Actions), Git, Linux, Cloudera

Qualité: Tests dbt (fraîcheur, unicité, contraintes métier), pytest, ruff, SQLFluff, typage statique, alerting Slack, revue de code, TDD

Méthodes: Agile / Scrum, TDD, cadrage avec le métier, recette iso-fonctionnelle, documentation et handover

Formation

Bac+5	**ESIEE Paris**, Diplôme d'ingénieur — Data Science et Intelligence Artificielle	2016 - 2021
	Université de Sherbrooke, Computer Science — échange universitaire de 8 mois	Québec, Canada 2019 - 2020
	UPEM — Université Paris-Est Marne-la-Vallée, Double licence en Informatique	2016 - 2019

Langues

Français: Langue maternelle

Anglais: Courant — TOEIC 990

Matrice de compétences

Compétence	Niveau	Années
Langages		
Python	●●●●● Expert	5+
SQL	●●●●● Expert	5+
Java	●●●●● Expert	5+
Bash	●●●●○ Avancé	5
Data Engineering		
Pipelines ELT / ETL	●●●●● Expert	5
Apache Spark	●●●●● Expert	5+
Hadoop / Hive / ORC	●●●●● Expert	5
dbt	●●●●● Expert	2
Airflow	●●●●○ Avancé	2
Qualité & monitoring	●●●●● Expert	5
Modélisation		
Modél. dimensionnelle	●●●●● Expert	5
Archi. raw/staging/marts	●●●●● Expert	3
Optimisation Spark / SQL	●●●●● Expert	4

Compétence	Niveau	Années
Cloud & entrepôts		
AWS	●●●●○ Avancé	3
Redshift Serverless	●●●●○ Avancé	2
GCP	●●●●○ Avancé	2
BigQuery	●●●●○ Avancé	2
PostgreSQL / SQL Server	●●●●○ Avancé	4
DevOps / DataOps		
Git	●●●●● Expert	5+
CI/CD	●●●●○ Avancé	5
Docker	●●●●○ Avancé	4
Terraform	●●●●○ Avancé	2
Linux	●●●●○ Avancé	5+
BI & restitution		
QuickSight	●●●●○ Avancé	2
Power BI	●●●●○ Avancé	2
Streamlit	●●●●○ Avancé	2
Méthodes		
Agile / Scrum	●●●●● Expert	5
TDD & revue de code	●●●●● Expert	5

Expériences professionnelles

Keobiz, Lead Data Engineer — Refonte Modern Data Stack — Paris

Mar 2026 – Juin 2026

Contexte — Owner côté data d'une mission flash de 3-4 mois : migration d'un ETL Python monolithique (extraction, chargement SQL inline et marts mêlés) vers une Modern Data Stack livrée en production. Pipeline nocturne de 3-4 h, incrémental mal maîtrisé, RDS saturée, self-service BI impossible. Périmètre : ~10 sources, ~90 modèles staging, ~10 marts métier, ~300 Go.

- **Architecture** — conception de la cible en couches raw / staging / marts : raw append-only et rejouable sur S3, structuration du monorepo data, séparation nette extract / load / transform / BI.
- **Ingestion** — réécriture de l'extraction en sinks Python modulaires, conteneurisés sur AWS Fargate et ordonnancés via EventBridge.
- **Transformation** — modèles dbt staging et marts sur Redshift Serverless, stratégies incrémentales, tests qualité, alertes Slack, CI sur chaque PR.
- **Gouvernance** — contribution au RBAC et aux masking policies RGPD côté entrepôt.
- **Bascule** — validation de l'iso-fonctionnalité entre l'ancien ETL et la nouvelle plate-forme, source par source, avant mise en production.
- **Métier & transmission** — cadrage avec les data analysts (3 nouveaux marts métier), documentation et équipe Data rendue autonome sur dbt.
- **Résultats** — pipeline nocturne ramené de 3-4 h à moins d'1 h ; plateforme prête pour l'ouverture du self-service BI (~80 utilisateurs).
- **Environnement technique** — Python, dbt, SQL, AWS (S3, Redshift Serverless, Fargate, EventBridge, Secrets Manager, QuickSight), Docker, CI/CD, Git, PostgreSQL/RDS, SQLFluff, Slack.

Natixis (Groupe BPCE), Data Engineer — Modélisation & pipelines distribués Java/Spark — Paris
Contexte — Moteur de calcul réglementaire (risque) sur plateforme Cloudera en production 24/7, avec runs quotidiens et échéances d'arrêtés trimestriels. — Fév 2024 – Mar 2026

- **Pipelines** — développement et maintenance de traitements ETL Java & Spark sur données Hive/ORC, orchestrés sous Airflow (DAGs quotidiens, dépendances, reprises).
- **Modélisation** — modélisation dimensionnelle (faits, dimensions, historisation SCD) ; clés de partitionnement et de jointure selon volumes et usages.
- **Interface métier** — traduction des spécifications des Business Analysts en traitements, validation fonctionnelle, challenge des exigences coûteuses (tables ~500 colonnes, jointures en cascade).
- **Optimisation** — analyse des plans d'exécution et de la Spark UI, optimisation Hive et SQL Server.
- **Production** — support lors des runs et des arrêtés trimestriels, revues de code, CI/CD Jenkins.
- **Environnement technique** — Java, Apache Spark, Hive/ORC, Cloudera, Airflow, SQL Server, SQL, Jenkins, Git, Linux.

BNP Paribas, Data Engineer — Java & Spark — Paris
Contexte — Traitement de données sur plateforme Apache Spark et développement sur un moteur de règles métier (BRMS IBM ODM) en production. — Déc 2021 – Aoû 2023

- **Traitement de données** — développement et analyse de traitements sur plateforme distribuée Apache Spark.
- **Règles métier** — développement Java sur IBM ODM, de la spécification fonctionnelle au code en production.
- **Méthode** — Agile / Scrum, TDD, collaboration régulière avec les équipes métier.
- **Environnement technique** — Java, Apache Spark, SQL, IBM ODM, Git, Agile/Scrum, TDD.

Projet open source

Velib Data Platform — ELT serverless de bout en bout sur GCP
Contexte — Plateforme conçue et réalisée seul, de l'infrastructure à la restitution : snapshot complet du parc Vélib toutes les 5 min (~1 400 stations, ~20 000 vélos, ~200M observations/an).

- **Ingestion & orchestration** — extraction Python sur Cloud Run Jobs, déclenchée par Cloud Scheduler et orchestrée par Cloud Workflows ; raw JSON append-only et rejouable dans GCS.
- **Modélisation dbt** — couches staging / intermediate / mart / analytics sur BigQuery ; modèles incrémentaux (merge, insert_overwrite avec lookback) ; détection de trajets et de passages atelier à partir des snapshots.
- **Infrastructure** — 100% Terraform, Docker multi-stage, IAM au moindre privilège par compte de service, serverless scale-to-zero.
- **Qualité** — CI/CD avec pytest, ruff, SQLFluff, typage statique ; dashboard Streamlit sur la couche analytics.
- **Environnement technique** — Python 3.12, uv, dbt-core, dbt-bigquery, BigQuery, GCS, Cloud Run, Cloud Workflows, Cloud Scheduler, Terraform, Docker, Streamlit, GitHub / Forgejo Actions.