Data Lakes & Data Integration - Projet Final

Cours EFREI 2024-2025

Yvann VINCENT
Machine Learning Engineer

Introduction

1 Objectifs du projet

Ce projet final vise à mettre en pratique l'ensemble des concepts que nous avons explorés durant ce cours de Data Lakes & Data Integration. Au-delà d'une simple validation des acquis, il représente une opportunité de créer un projet concret et potentiellement utile pour votre portfolio.

L'objectif est de concevoir et d'implémenter un data lake de A à Z, en partant de l'ingestion des données jusqu'à leur exposition via une API. Ce projet vous permettra de démontrer votre compréhension des architectures de données modernes et votre capacité à mettre en place des pipelines de données robustes et efficaces.

2 Exigences de base

2.1 Architecture du Data Lake

Le cœur de votre projet reposera sur un data lake structuré en trois couches distinctes, suivant les bonnes pratiques que nous avons étudiées en cours. Je vous laisse libre d'utiliser les technologies qui vous semblent les plus pertinentes parmi celles que nous avons vues en cours pour stocker les données dans chaque couche.

- **La couche Raw** : Cette première couche accueillera vos données dans leur état original, sans aucune transformation. Elle servira d'archive immuable de vos données sources, permettant de revenir aux données originales en cas de besoin.
- **La couche Staging** : Cette couche intermédiaire contiendra vos données après une première phase de nettoyage et de structuration. C'est ici que vous effectuerez vos premières transformations pour rendre les données plus exploitables, tout en conservant un maximum d'informations de la source.
- **La couche Curated** : La dernière couche hébergera vos données dans leur forme la plus raffinée. Les données y seront enrichies, optimisées et prêtes pour l'analyse ou la consommation par des applications tierces.

2.2 Source de données

Je vous laisse une liberté totale sur le choix de votre source de données. Vous pouvez opter pour :

- Des datasets disponibles sur Hugging Face
- Des APIs publiques de votre choix
- Vos propres datasets d'images, de vidéos, ou tout autre type de données qui vous intéresse
- En bref, ce qui vous semble pertinent pour votre portfolio personnel

Cette flexibilité vous permet de travailler sur un sujet qui vous passionne ou qui pourrait être utile dans vos projets futurs. Par exemple, si vous êtes intéressé par le traitement d'images, vous pourriez construire un data lake pour gérer une collection de photos avec leurs métadonnées associées.

2.3 Scripts de transformation

Trois scripts distincts sont requis :

1. Script d'insertion dans la couche raw
2. Script de transformation raw vers staging
3. Script de transformation staging vers curated

2.4 Pipeline d'intégration

Implémentez une pipeline d'intégration continue utilisant au choix :

- DVC
- Apache Airflow

La pipeline doit assurer la reproductibilité des transformations et l'alimentation automatisée du data lake.

2.5 API Gateway

L'API doit fournir les endpoints suivants :

- `/raw` : Accès aux données brutes
- `/staging` : Accès aux données intermédiaires
- `/curated` : Accès aux données finales
- `/health` : Vérification de l'état des services
- `/stats` : Métriques sur le remplissage des buckets et bases de données

3 Niveau Avancé

Pour ceux qui souhaitent se démarquer et relever un défi supplémentaire, j'ai créé un niveau avancé. Je tiens à préciser d'emblée que ce niveau est totalement optionnel - vous pouvez obtenir une excellente note allant de 16 à 20-20 sans vous y attaquer. Ce niveau a été conçu pour ceux qui veulent aller plus loin et explorer des concepts plus avancés.

3.1 Récompenses et Motivation

Pour encourager l'innovation et récompenser les efforts supplémentaires, j'ai mis en place plusieurs récompenses pour le niveau avancé :

- Pour les trois meilleurs projets réalisés en solo : des recommandations LinkedIn personnalisées et détaillées, mettant en avant vos compétences techniques et votre capacité à mener un projet complexe de bout en bout.
- Pour la meilleure équipe : un mois d'abonnement à Google Colab Pro, vous permettant d'expérimenter avec des ressources plus conséquentes et d'approfondir vos projets personnels.

- Pour tous les participants au niveau avancé : un t-shirt exclusif ***"Swimming through Streams and Data Lakes"***. Pour en bénéficier, il vous suffira de m'indiquer lors du TP7 que vous participez au niveau avancé, en précisant votre taille de t-shirt. Cela me permettra de préparer les commandes pour une distribution lors de notre dernière séance du 25 février.

3.2 Endpoints Avancés

3.2.1 L'endpoint /ingest

L'une des contraintes principales du niveau avancé concerne l'implémentation d'un endpoint d'ingestion de données depuis l'API Gateway. Votre API gateway devra exposer un endpoint `/ingest` capable d'accepter des données pertinentes à votre data source au format JSON et de les propager à travers votre pipeline d'ingestion.

Par exemple, si vous travaillez avec des données textuelles, votre endpoint pourrait accepter une structure JSON comme :

```
{
    "data": {
        "texts": ["Premier texte à analyser", "Deuxième texte à traiter", ...]
    }
}
```

Un aspect crucial de cet endpoint sera la mesure de ses performances. Vous devrez chronométrer et documenter le temps d'exécution de votre pipeline pour :

- Un batch contenant un seul élément
- Un batch de 10 éléments

Ces mesures serviront de base de comparaison pour le second endpoint optimisé.

3.2.2 L'endpoint /ingest_fast

Le véritable défi technique réside dans l'implémentation de l'endpoint `/ingest_fast`. Dans un environnement de production, la performance est cruciale - un data lake lent peut rapidement devenir un goulot d'étranglement pour toute l'infrastructure.

Cet endpoint doit offrir une amélioration de temps d'exécution d'au moins 30% par rapport à `/ingest`. Pour atteindre cet objectif, vous devrez réfléchir à des stratégies d'optimisation comme :

- L'utilisation de Numba pour accélérer les calculs numériques
- La vectorisation des opérations avec NumPy
- La parallélisation intelligente des traitements
- L'optimisation des requêtes et des accès aux données
- La mise en cache stratégique

La comparaison des performances devra être documentée, avec des mesures précises pour les mêmes tailles de batch que `/ingest`.

4 Critères d'Évaluation

4.1 Évaluation du niveau standard

Les critères d'évaluation de base incluront :

- La qualité de l'implémentation des trois couches du data lake
- La robustesse des scripts de transformation
- La fiabilité de la pipeline d'intégration
- La qualité et la complétude de l'API Gateway
- La documentation du projet
- La clarté du code et son organisation

4.2 Évaluation du niveau avancé

Pour le niveau avancé, je regarderai également :

- Les performances comparées des endpoints `/ingest` et `/ingest_fast`
- La créativité dans les solutions d'optimisation
- La qualité de la documentation technique des optimisations
- Bonus : L'originalité du choix des données et du domaine d'application

5 Livrables Attendus

Pour valider votre projet, vous devrez fournir :

- Un dépôt GitHub contenant l'intégralité du code source
- Une documentation technique détaillée expliquant :
  - L'architecture de votre solution
  - Les choix techniques effectués
  - Les procédures d'installation et de build de votre projet
  - Les résultats des tests de performance (pour le niveau avancé)
- Soyez exhaustifs dans votre README. Pas besoin de faire du Shakespeare, j'ai simplement besoin de savoir comment builder votre solution et l'utiliser en contrepartie de la liberté techique que je vous laisse pour réaliser le projet.

6 Planning et Deadline

Le projet devra être finalisé pour le 11 Mars 2025. Je vous conseille de :

- Commencer par définir l'architecture globale de votre solution
- Implémenter progressivement les différentes couches
- Garder du temps pour les optimisations et les tests
- Préparer une documentation au fur et à mesure

Pour ceux qui le souhaitent, je serai là pour review votre solution et vous donner des conseils lors de la séance du TP7 et du TP8.

Je vous souhaite bon courage, et j'espère que ce projet sera utile à votre portfolio.