- Déploiement de l'infrastructure (Docker Compose) : **MinIO + PostgreSQL + Airflow + Container Python pour exécuter les scripts**
- Ingestion des données (MinIO) : Déposer une version **étendue** du dataset Diabète (100k–2M lignes ou ~200–500MB).

Pipeline Airflow (ETL + ML)

Votre DAG doit contenir au minimum 4 tâches :

🔹 **(T1) Extraction**

- Télécharger le fichier CSV depuis MinIO via `boto3`.

🔹 **(T2) Nettoyage / Préparation**

- Gestion des valeurs manquantes
- Normalisation / scaling des variables
- Encodage (si nécessaire)
- Génération d'un dataset propre pour le ML

🔹 **(T3) Stockage**

- Chargement du dataset nettoyé dans PostgreSQL
- Création d'une table normalisée (types → INTEGER, FLOAT, BOOLEAN…)

🔹 **(T4) Modélisation ML simple**

- Séparation train/test
- Entraînement d'un modèle simple (Logistic Regression ou RandomForest)
- Évaluation (accuracy, f1-score)
- Sauvegarde du modèle ou des métriques dans MinIO ou PostgreSQL

Analyse SQL (obligatoire)

Créer **5 requêtes SQL analytiques** sur la base nettoyée.
Exemples :

- moyenne des niveaux de glucose par âge

- distribution des BMI par classe diabétique
- corrélations simples (via SQL)
- pourcentage de personnes diabétiques par tranche d'âge
- valeurs extrêmes/exceptions

Rapport technique

À intégrer :

- Architecture Docker + schéma
- Explication du DAG Airflow
- Description du dataset (taille, variables…)
- Détails du nettoyage / transformation
- Choix du modèle ML et justification
- Résultats + interprétation