Énoncé – Exercice pratique : Data Lake vs Data Warehouse (version locale)

Contexte :

Votre entreprise collecte des données de ventes journalières sous forme de fichiers CSV déposés dans un dossier local.
Vous devez analyser ces données en adoptant deux approches distinctes :

1. **Approche Data Lake** : stockage brut sans transformation préalable.
2. **Approche Data Warehouse** : structuration et chargement dans une base relationnelle locale.

Objectifs pédagogiques :

- Comprendre la différence entre **stockage brut (Data Lake)** et **stockage structuré (Data Warehouse)**.
- Manipuler des fichiers de données réelles (CSV).
- Effectuer des opérations de transformation minimale.
- Charger et interroger une base relationnelle.

Instructions :

Étape 1 – Mise en place du "Data Lake" local

1. Crée un dossier `data_lake/` avec la structure suivante :

```
data_lake/
├── raw/
├── transformed/
└── analytics/
```

2. Dans `data_lake/raw/`, place un fichier `ventes_2024.csv` contenant :

```
Date,Client,Produit,Quantite,PrixUnitaire
2024-01-01,Jean,PC,2,1200
2024-01-02,Marie,Téléphone,1,700
2024-01-03,Paul,Écran,3,300
```

Étape 2 – Simulation d'un Data Warehouse local

1. Installe ou utilise un moteur de base relationnelle local : **PostgreSQL, SQLite ou autre**.

2. Crée une base nommée `entreprise_dw`.
3. Crée une table `ventes` avec la structure suivante :

```
CREATE TABLE ventes (
    id SERIAL PRIMARY KEY,
    date DATE,
    client TEXT,
    produit TEXT,
    quantite INT,
    prix_unitaire NUMERIC,
    total NUMERIC
);
```

4. Écris un petit script Python (ou utilise un outil ETL) pour :
    - Lire le fichier CSV depuis `data_lake/raw/`
    - Calculer une colonne `total` = `quantite` × `prix_unitaire`
    - Charger les données dans la base `entreprise_dw` dans la table `ventes`

Étape 3 – Analyse et réflexion

1. Compare les deux approches (Data Lake vs Data Warehouse) à travers ces questions :
    - Quels types de requêtes sont faciles ou difficiles à faire dans chaque approche ?
    - Quel est le niveau de structuration des données dans chaque cas ?
    - Quelles sont les implications en termes de gouvernance, qualité, et sécurité ?

Livrables attendus

- Affichage de la table `ventes` remplie.
- Réponses à la comparaison Data Lake vs Data Warehouse.
- Scripts utilisés pour le traitement.