Pipeline OLTP → ETL Pentaho PDI → Data Warehouse → Reporting Power BI

Rapport de TP Data Engineering

Réalisé par : **Aymane EL MKADMI**

Année Universitaire : 2025-2026

Introduction

Ce rapport présente la construction complète d'un pipeline décisionnel pour l'entreprise fictive TechStore, allant d'une base de données opérationnelle (OLTP) à la création d'un Data Warehouse alimenté par un processus ETL réalisé dans Pentaho PDI, puis analysé dans Power BI.

OLTP vs OLAP

OLTP (Online Transaction Processing) : Gère les opérations quotidiennes (clients, commandes, produits). Fortement normalisé pour éviter la redondance. **OLAP (Online Analytical Processing)** : Utilisé pour l'analyse (tendances, performance). S'appuie sur un Data Warehouse en schéma en étoile.

Rôles des composants

• **Pentaho PDI** : Extraction, nettoyage, transformation et chargement des données OLTP vers le DWH.

• **Data Warehouse** : Données intégrées, nettoyées, historisées et optimisées pour l'analyse.

• **Power BI** : Connexion au DWH, création de tableaux de bord et visualisation dynamique des indicateurs.

1. Modèle OLTP

TechStore, entreprise spécialisée dans la vente de matériel électronique, utilise une base OLTP (ventes_oltp) contenant quatre tables principales : clients, produits, commandes et lignes_commandes. Ces tables sont reliées par des clés étrangères formant un modèle relationnel normalisé.

Figure 1 – Tables de la base OLTP

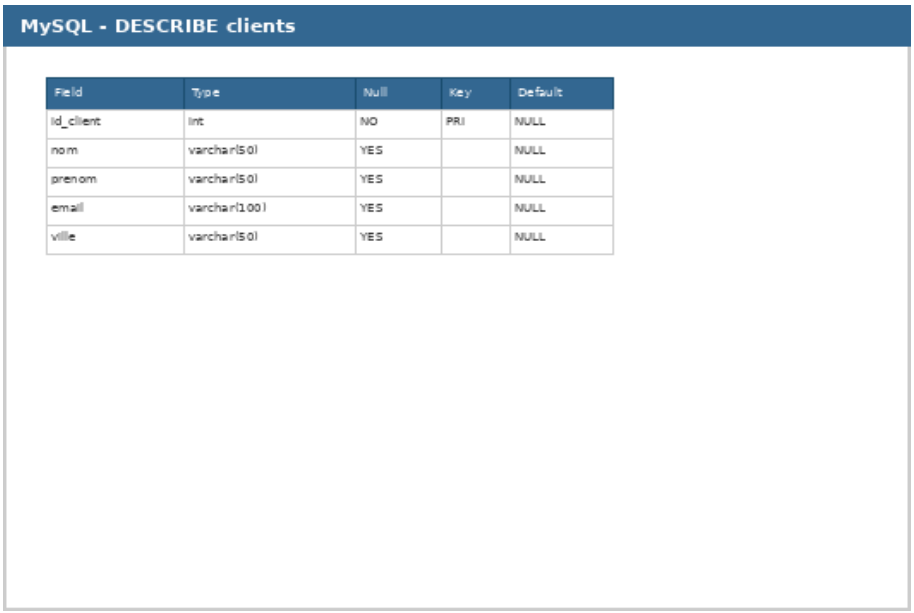

Figure 2 – Structure de la table clients

Génération et vérification des données

Après création de la structure, un script Python génère des données de test réalistes (10,000 clients, 500 produits, 50,000 commandes) et les exporte en CSV. Les fichiers sont ensuite importés dans MySQL.

Figure 3 – Fichiers CSV générés

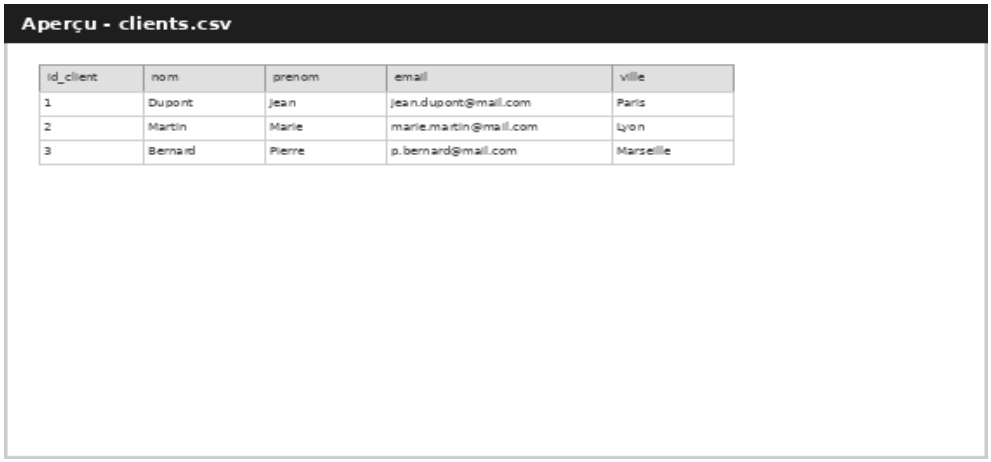

Figure 4 – Aperçu des données clients

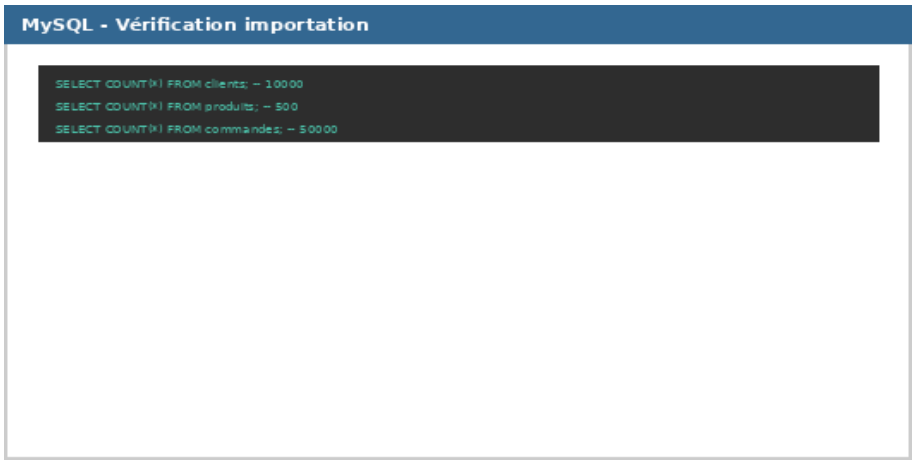

Figure 5 – Vérification de l'importation MySQL

2. Création du Data Warehouse

La base ventes_dwh est créée avec un schéma en étoile comprenant trois dimensions (DimClient, DimProduit, DimDate) et une table de faits (FactVentes). Les clés surrogates permettent de gérer l'historique et les relations.

Figure 6 – Tables du Data Warehouse

3. Pentaho PDI – ETL et Transformations

Pentaho Data Integration (PDI) est configuré avec deux connexions : OLTP_MySQL (source) et DWH_MySQL (destination). Quatre transformations sont créées pour charger les dimensions et la table de faits.

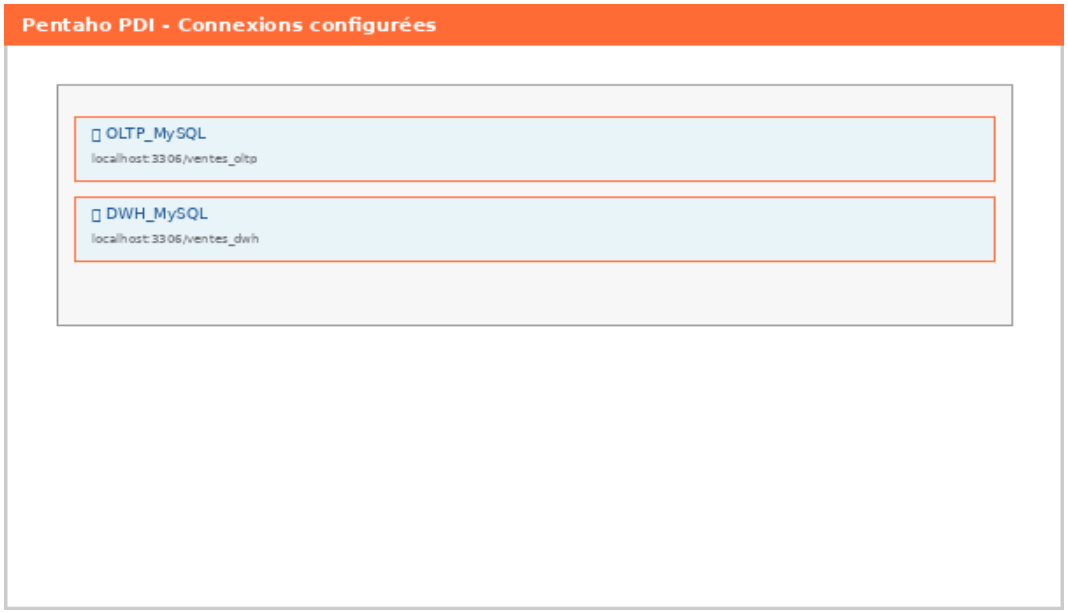

Figure 7 – Connexions configurées dans Pentaho

Transformations ETL

• **dim_client.ktr** : Extraction clients OLTP → Renommage colonnes → Chargement DimClient (10,000 lignes)

• **dim_produit.ktr** : Extraction produits → Renommage → Chargement DimProduit (500 lignes)

• **dim_date.ktr** : Génération artificielle dates 2022-2024 avec composantes temporelles (1,096 lignes)

• **fact_ventes.ktr** : Jointures OLTP + Lookups dimensions + Calculs → FactVentes (100,000 lignes)

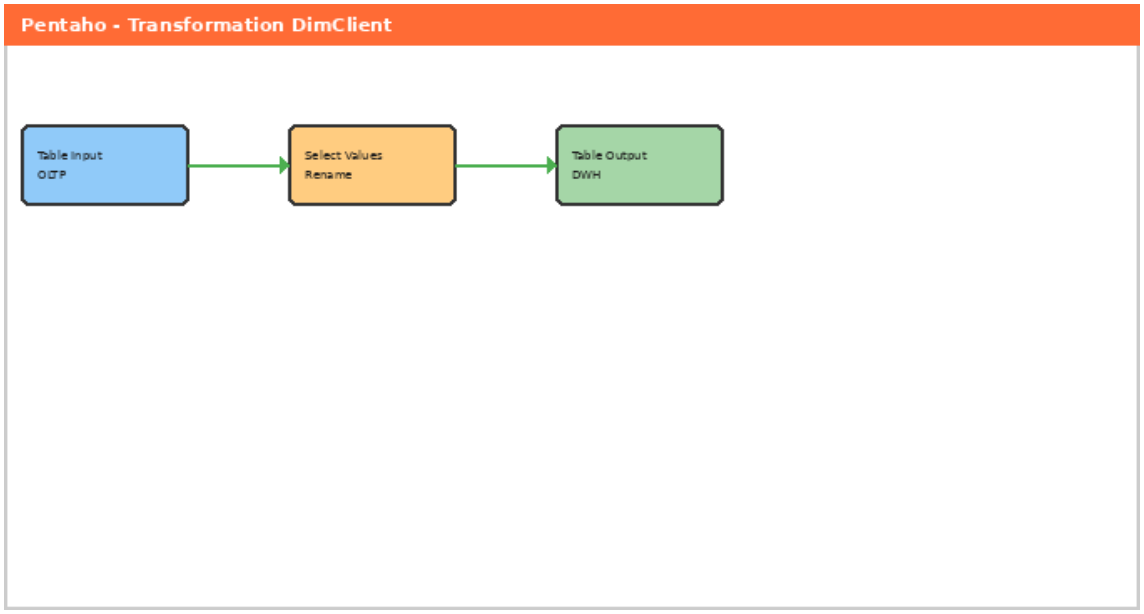

Figure 8 – Transformation DimClient dans Pentaho

Job d'orchestration

Un Job Pentaho (job_etl_complet.kjb) orchestre l'exécution séquentielle des quatre transformations avec gestion des dépendances et contrôle d'erreur global.

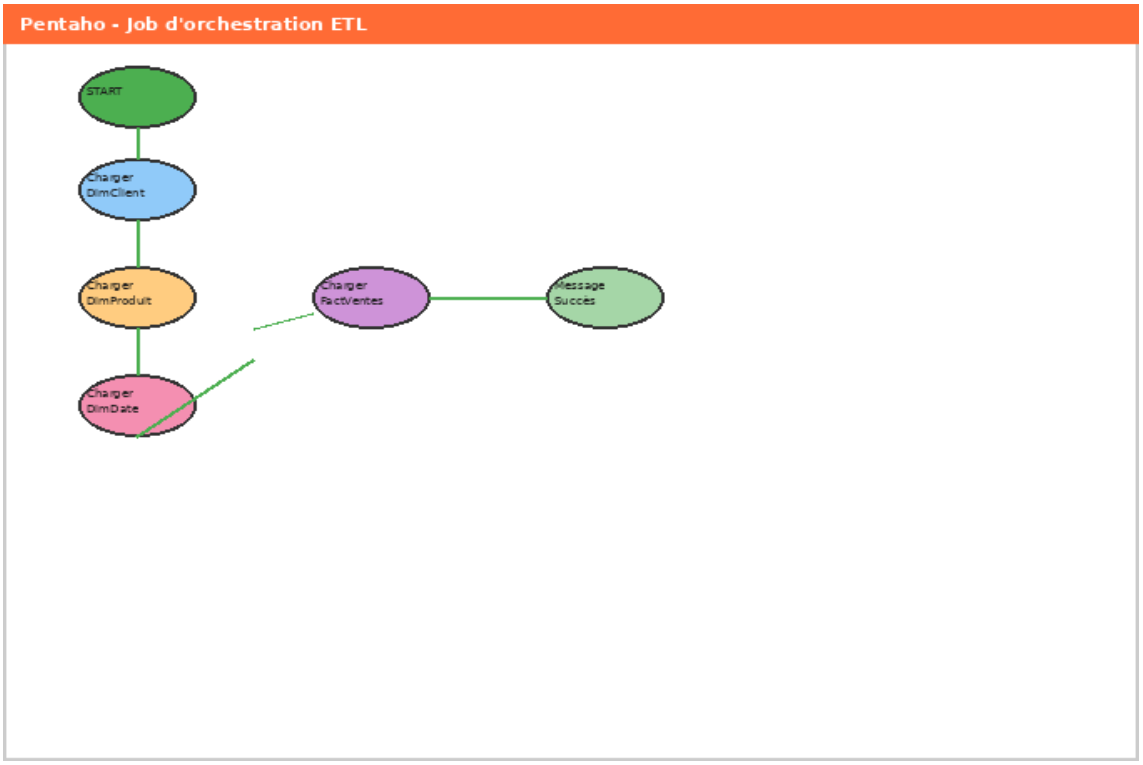

Figure 9 – Job d'orchestration ETL complet

4. Requêtes OLAP sur le Data Warehouse

Après chargement, plusieurs requêtes analytiques sont exécutées pour analyser les ventes selon différentes dimensions. Le schéma en étoile permet des requêtes rapides avec peu de jointures.

• CA par ville : Identification des zones géographiques les plus rentables

• CA par catégorie : Performance des gammes de produits

• Évolution mensuelle : Analyse de la saisonnalité

• Top 10 produits : Produits leaders à maintenir en stock

• Analyse trimestre-catégorie : Performance croisée temps/catégorie

MySQL - Requête OLAP CA par ville

```
SELECT c.ville, SUM(f.montant_total) AS ca_total
FROM FactVentes f JOIN DimClient c ON f.id_client_dim = c.id_client_dim
GROUP BY c.ville ORDER BY ca_total DESC LIMIT 5;
```

ville	ca_total
Paris	2,450,000
Lyon	1,890,000
Marseille	1,620,000

Figure 10 – Requête OLAP : CA par ville

5. Power BI - Reporting et Visualisation

Power BI Desktop se connecte au Data Warehouse MySQL et importe les quatre tables. Le modèle en étoile est automatiquement reconnu et les relations sont configurées.

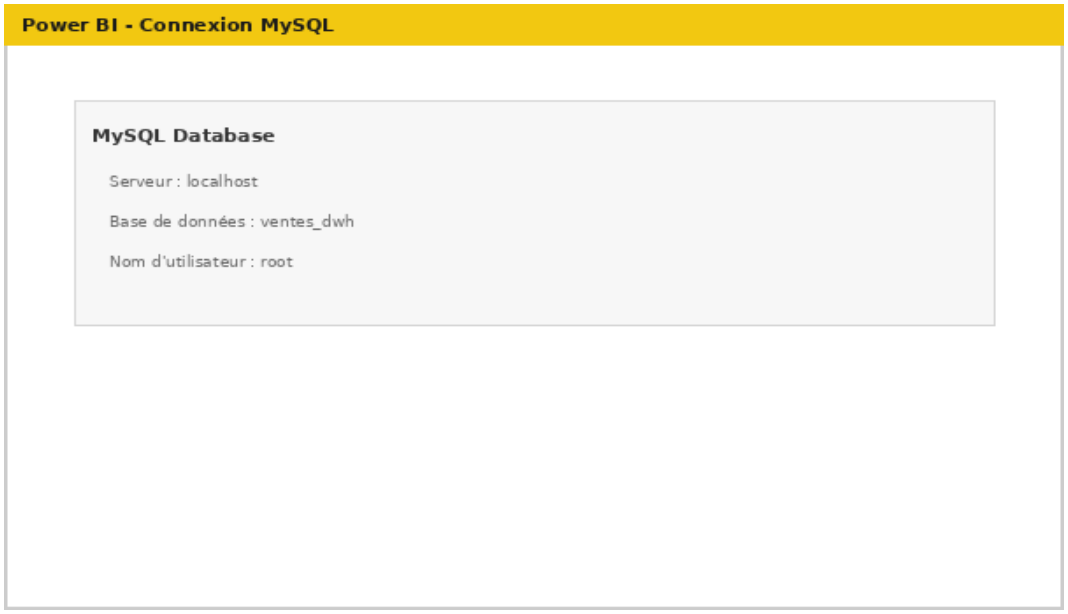

Figure 11 – Connexion à MySQL depuis Power BI

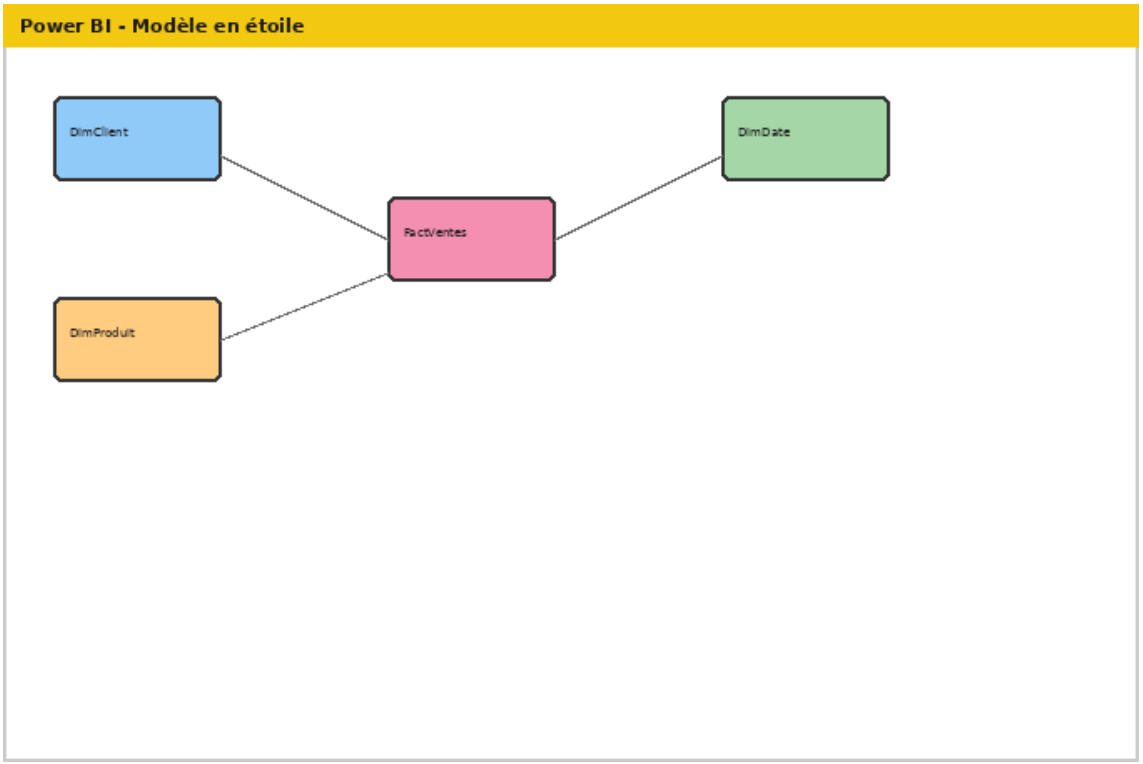

Figure 12 – Modèle de données en étoile dans Power BI

Visualisations créées

• CA par ville (barres) : Identification des zones performantes

• CA par catégorie (secteurs) : Contribution relative de chaque catégorie

• Évolution mensuelle (courbes) : Tendances et saisonnalité

• Top 10 produits (barres horizontales) : Produits phares

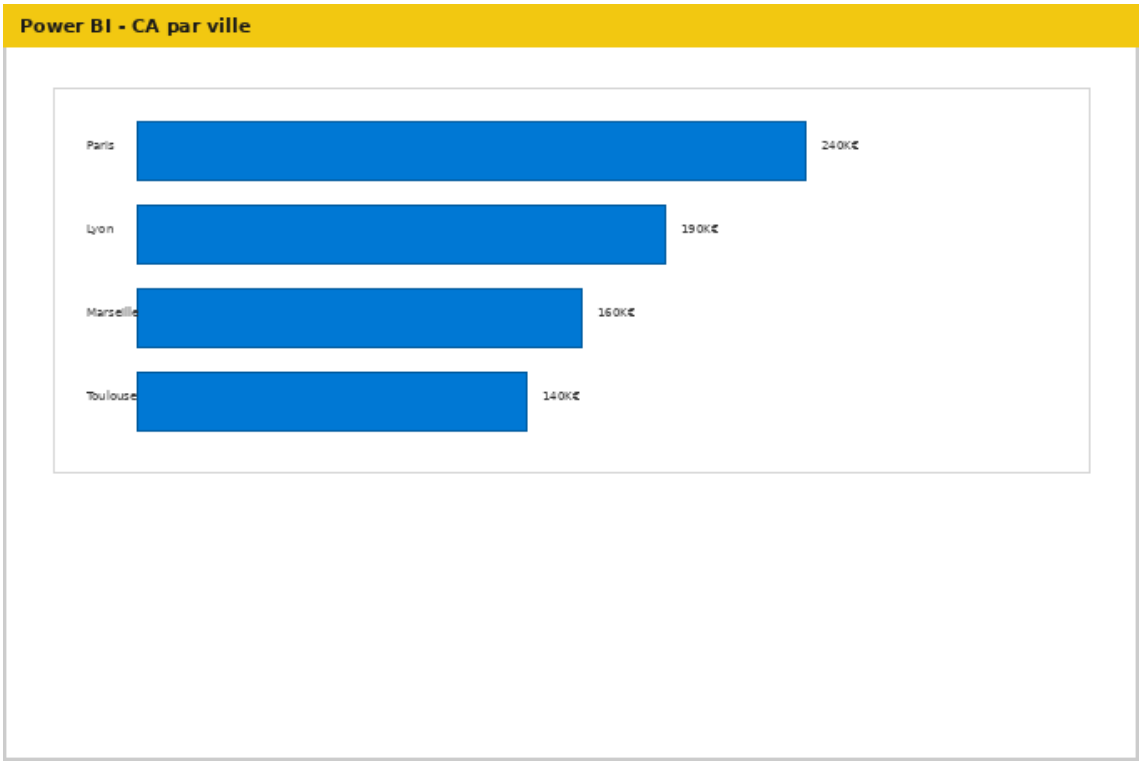

Figure 13 – Graphique CA par ville dans Power BI

Interactivité et mesures DAX

Des segments (slicers) Année et Catégorie permettent le filtrage dynamique de tous les visuels. Des mesures DAX sont créées : CA Total, Nombre de ventes, Panier Moyen, Clients Uniques. Une page de drill-through permet l'analyse détaillée par produit.

6. Conclusion

Ce projet a permis de construire un pipeline décisionnel complet de bout en bout, démontrant la complémentarité entre systèmes OLTP (opérationnel) et OLAP (analytique). Les compétences acquises incluent :

• Conception et implémentation de schémas OLTP normalisés et DWH en étoile

• Maîtrise de Pentaho PDI pour l'extraction, transformation et chargement de données

• Orchestration de processus ETL complexes avec gestion des dépendances

• Requêtage OLAP optimisé pour l'analyse multidimensionnelle

• Création de dashboards interactifs avec Power BI et mesures DAX

• Compréhension globale d'une architecture décisionnelle moderne

Ce pipeline constitue une base solide pour des analyses décisionnelles en entreprise, permettant de transformer des données opérationnelles brutes en insights actionnables pour la prise de décision stratégique.