Rapport de Projet de Business Intelligence
Pipeline complet : OLTP → ETL Pentaho → Data Warehouse → Reporting Power BI

FORMATION : 4e année d'ingénierie Big Data – UEMF

Promo : BD 2025-2026

Réalisée par :
Ahlam OUBOUAZZA

Supervisé par :
Ahmed AMAMOU

Sommaire

1. Introduction
- Contexte du projet
- Objectifs et buts du projet

2. Technologies utilisées

3. Implémentation
- Backend (FastAPI)
- Frontend (React)
- Structure du projet

4. Résultats

4.1. Évaluation du modèle

4.2. Prédictions de prix

5. Discussion

5.1. Analyse des résultats obtenus

5.2. Limites du modèle

6. Conclusion
- Résumé des conclusions
- Impact des résultats obtenus

1. Introduction

TechStore est une entreprise e-commerce fictive spécialisée dans les produits électroniques. Ce projet répond au besoin de transformer ses données transactionnelles brutes en informations analysables pour la prise de décision.

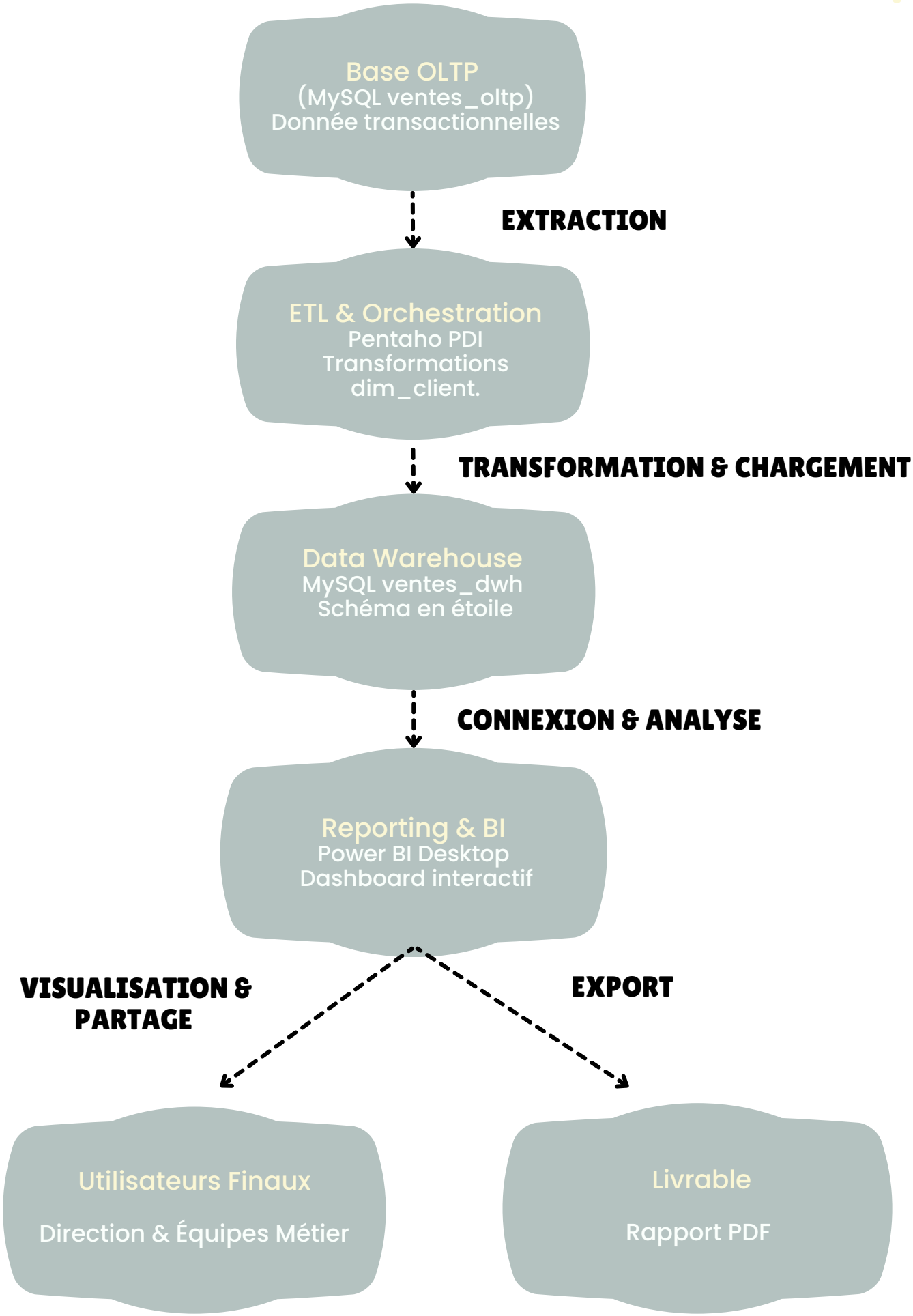
Objectifs :

-Modéliser une base OLTP et générer des données réalistes.
-Concevoir un Data Warehouse avec un schéma en étoile.
-Automatiser le flux ETL avec Pentaho PDI.
-Créer un tableau de bord interactif avec Power BI.

2. Architecture du Projet

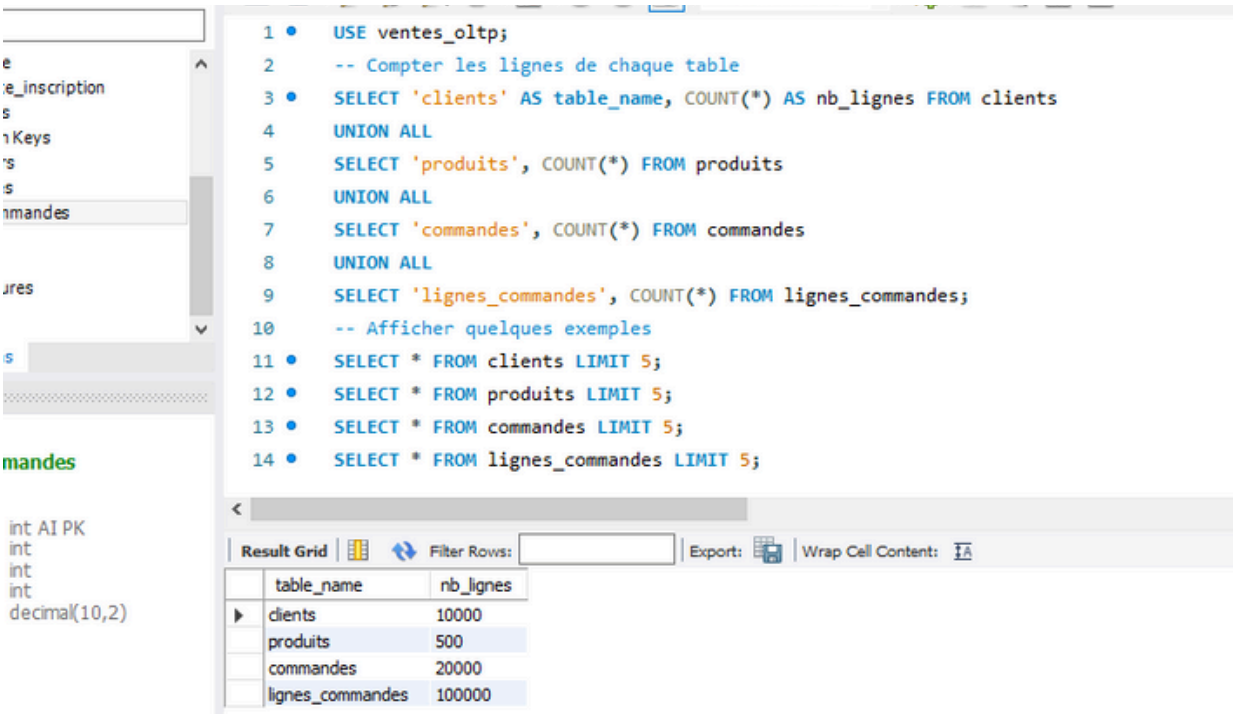
Technologies utilisées :

- SGBD : MySQL 8.0
- ETL : Pentaho Data Integration (PDI / Kettle) 9.x
- Visualisation : Microsoft Power BI Desktop
- Langages : SQL, Python (pour la génération de données)
- Système : Windows


Base OLTP
(MySQL ventes_oltp)
Donnée transactionnelles
EXTRACTION
ETL & Orchestration
Pentaho PDI
Transformations
dim_client.
TRANSFORMATION & CHARGEMENT
Data Warehouse
MySQL ventes_dwh
Schéma en étoile
CONNEXION & ANALYSE
Reporting & BI
Power BI Desktop
Dashboard interactif
VISUALISATION & PARTAGE
EXPORT
Utilisateurs Finaux
Direction & Équipes Métier
Livrable
Rapport PDF

3. Modélisation OLTP

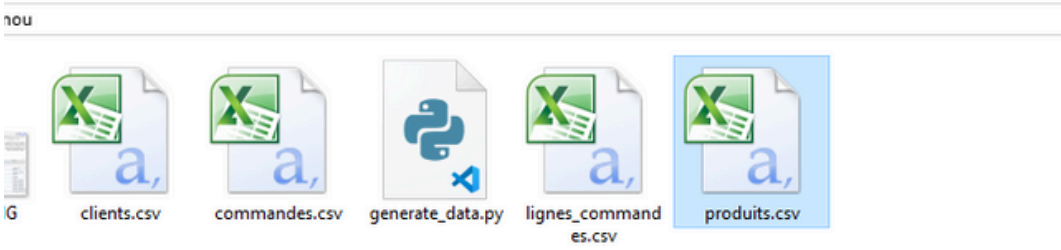
La base **ventes_oltp** contient 4 tables normalisées :
-clients : 10 000 clients
-produits : 500 produits
-commandes : 20 000 commandes
-lignes_commandes : 100 000 lignes

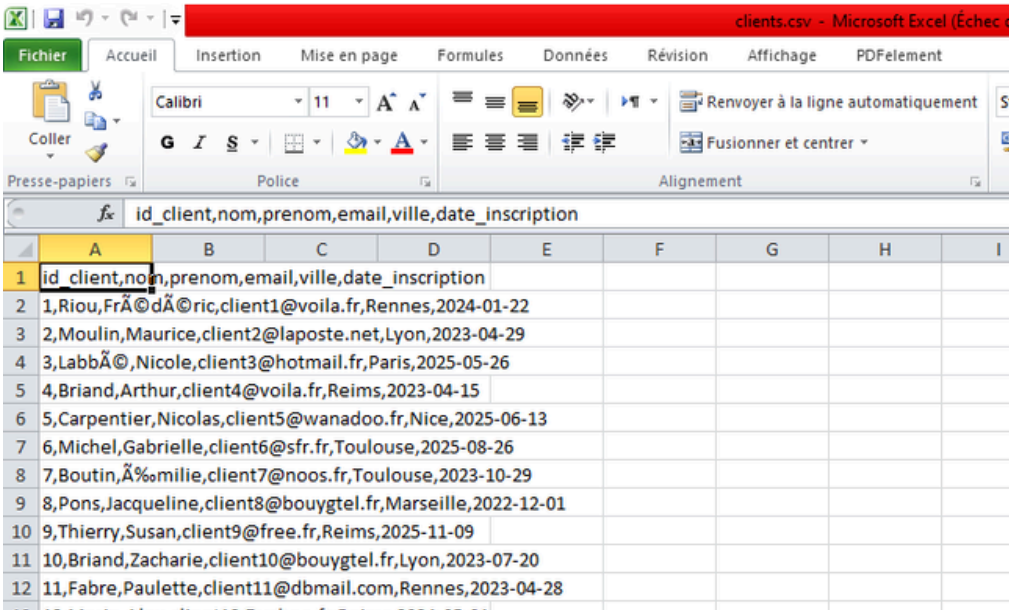


Justification des choix : Le modèle relationnel (3FN) a été choisi pour son efficacité dans la gestion des transactions (INSERT, UPDATE, DELETE), la minimisation des redondances et le maintien de l'intégrité référentielle via des clés étrangères.

4. Génération des Données

Pour simuler une activité réaliste, un jeu de données conséquent a été généré via un script Python personnalisé.
Méthodologie : Utilisation des bibliothèques Faker (pour des données réalistes) et pandas. Les données ont été générées en respectant les contraintes d'intégrité référentielle (ex: un id_client dans la table commandes doit exister dans la table clients).
Volumes et Répartition :
-Période couverte : 3 ans (2022-2024)
-Clients : 10 000, répartis sur 12 villes françaises.
-Produits : 500, répartis en 5 catégories (Ordinateurs, Téléphones, Tablettes, Accessoires, Montres).
-Commandes : 20 000, générées aléatoirement sur la période.
L-ignes de commande : 100 000, assurant un lien cohérent entre commandes et produits.





5. Modélisation du Data Warehouse (DWH)

Pour l'analyse, un schéma en étoile dénormalisé a été conçu dans la base ventes_dwh.

Structure du Schéma en Étoile :

-Table de Faits : FactVentes (100 000 lignes). Contient les mesures (quantité, prix, montant) et des clés étrangères vers les dimensions.

Dimensions :

-DimClient : Dérivée de la table clients.

-DimProduit : Dérivée de la table produits.

-DimDate : Table de dimension temporelle artificielle et pré-calculée (1 096 jours).

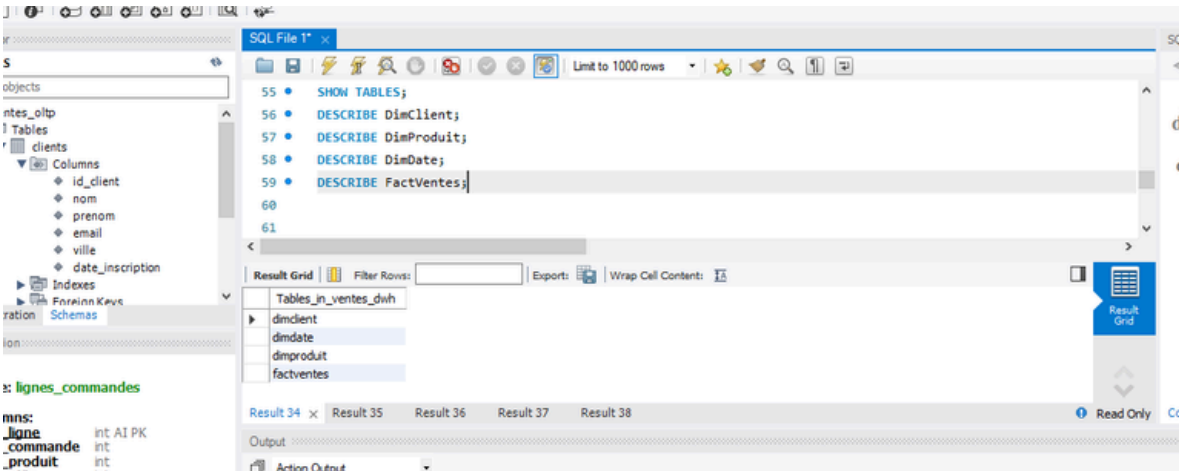
Justification et Avantages :

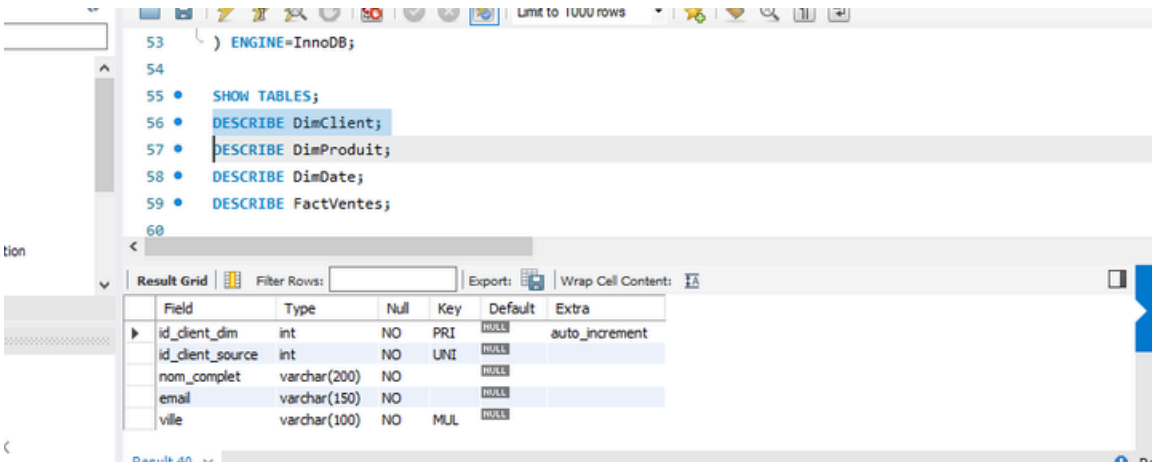
Ce modèle a été choisi pour :

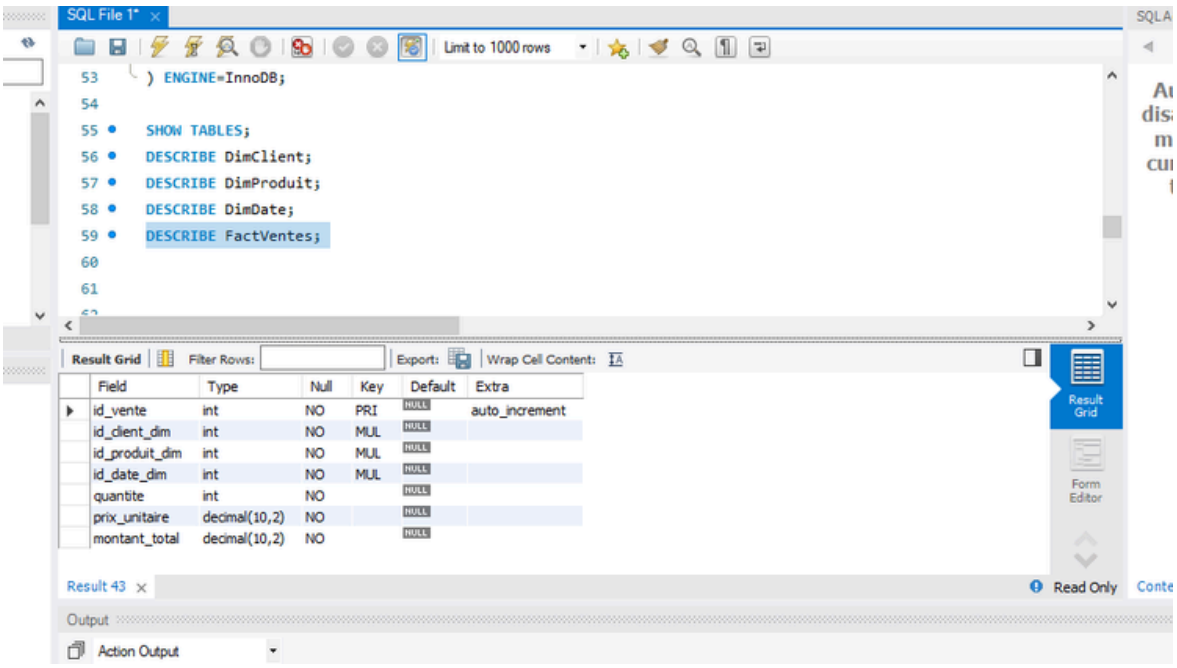
-Performance : Réduction du nombre de jointures pour les requêtes analytiques.

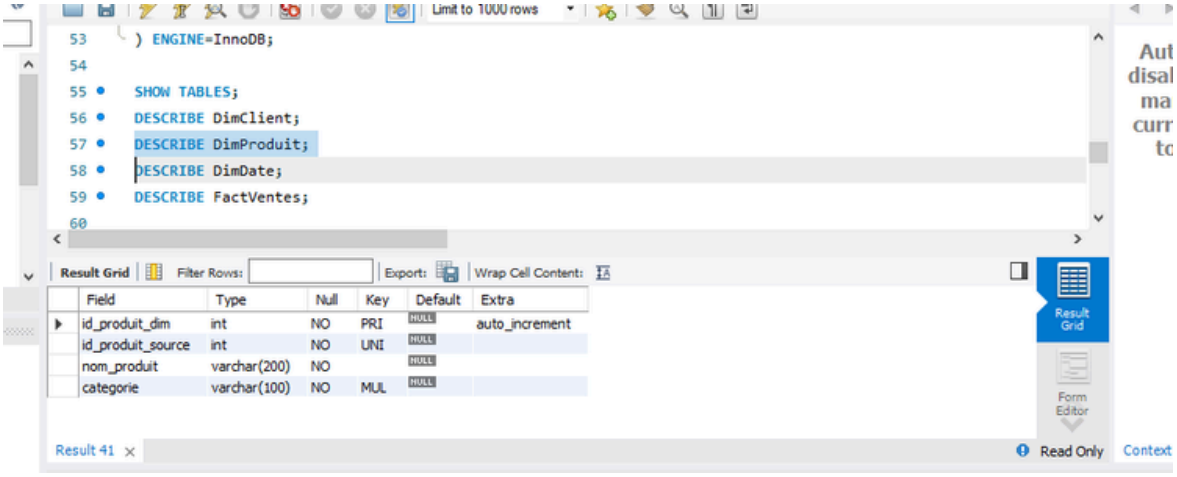
-Simplicité : Structure intuitive pour les utilisateurs métier et les outils de BI.

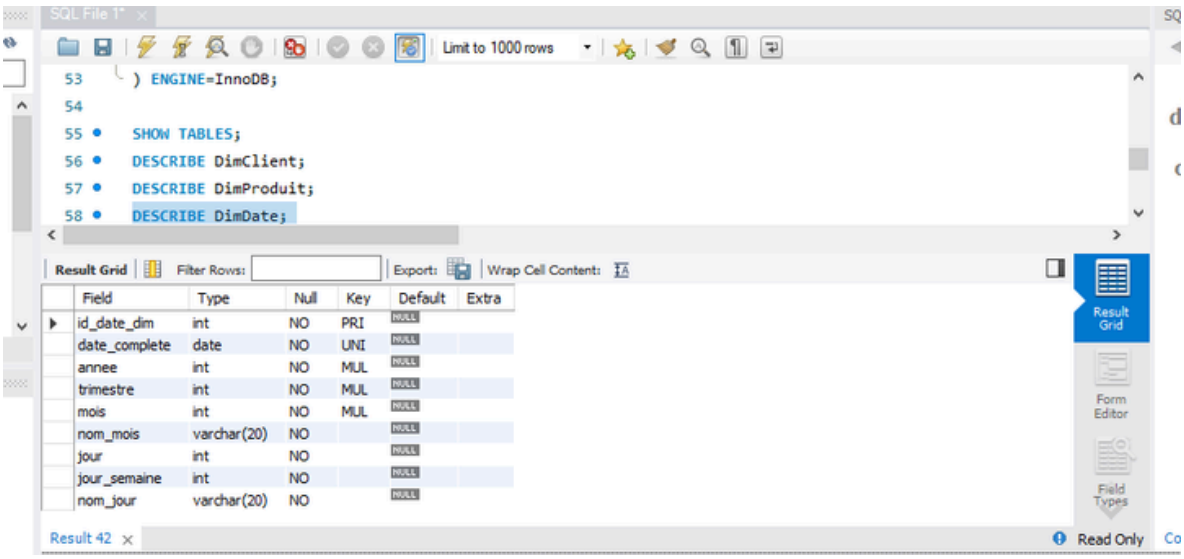
-Intégration : Optimisation pour les requêtes d'agrégation et de filtrage multidimensionnel typiques des outils comme Power BI.






SQL File 1*
Limit to 1000 rows
53) ENGINE=InnoDB;
54
55 SHOW TABLES;
56 DESCRIBE DimClient;
57 DESCRIBE DimProduit;
58 DESCRIBE DimDate;
59 DESCRIBE FactVentes;
60
61
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Field | Type | Null | Key | Default | Extra
id_vente | int | NO | PRI | NULL | auto_increment
id_client_dim | int | NO | MUL | NULL |
id_produit_dim | int | NO | MUL | NULL |
id_date_dim | int | NO | MUL | NULL |
quantite | int | NO | | NULL |
prix_unitaire | decimal(10,2) | NO | | NULL |
montant_total | decimal(10,2) | NO | | NULL |
Result Grid
Form Editor
Result 43
Read Only
Output
Action Output


Limit to 1000 rows
53) ENGINE=InnoDB;
54
55 SHOW TABLES;
56 DESCRIBE DimClient;
57 DESCRIBE DimProduit;
58 DESCRIBE DimDate;
59 DESCRIBE FactVentes;
60
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Field | Type | Null | Key | Default | Extra
id_produit_dim | int | NO | PRI | NULL | auto_increment
id_produit_source | int | NO | UNI | NULL |
nom_produit | varchar(200) | NO | | NULL |
categorie | varchar(100) | NO | MUL | NULL |
Result Grid
Form Editor
Result 41
Read Only

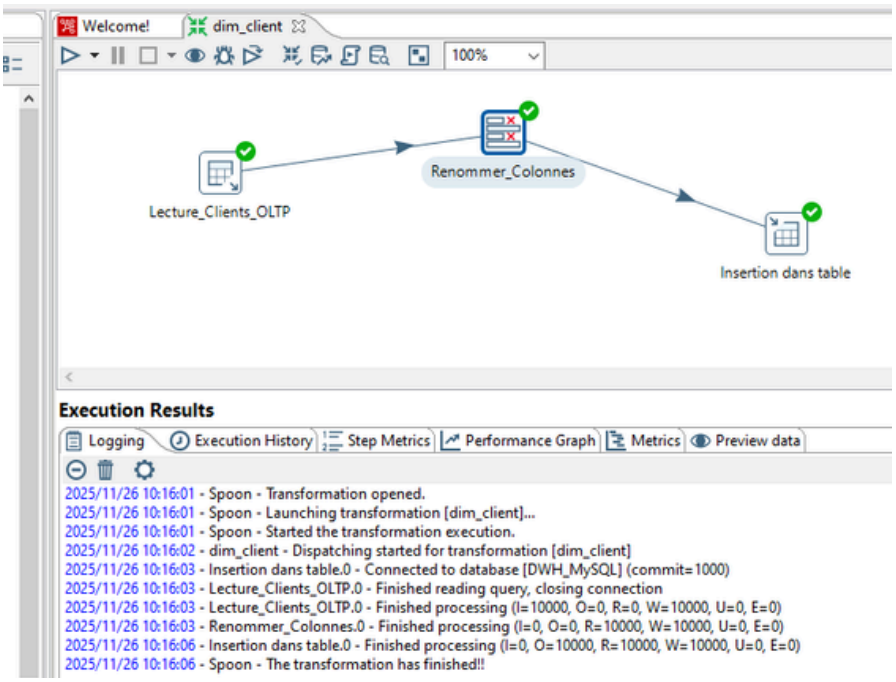

SQL File 1*
Limit to 1000 rows
53) ENGINE=InnoDB;
54
55 SHOW TABLES;
56 DESCRIBE DimClient;
57 DESCRIBE DimProduit;
58 DESCRIBE DimDate;
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Field | Type | Null | Key | Default | Extra
id_date_dim | int | NO | PRI | NULL |
date_complete | date | NO | UNI | NULL |
annee | int | NO | MUL | NULL |
trimestre | int | NO | MUL | NULL |
mois | int | NO | MUL | NULL |
nom_mois | varchar(20) | NO | | NULL |
jour | int | NO | | NULL |
jour_semaine | int | NO | | NULL |
nom_jour | varchar(20) | NO | | NULL |
Result Grid
Form Editor
Field Types
Result 42
Read Only

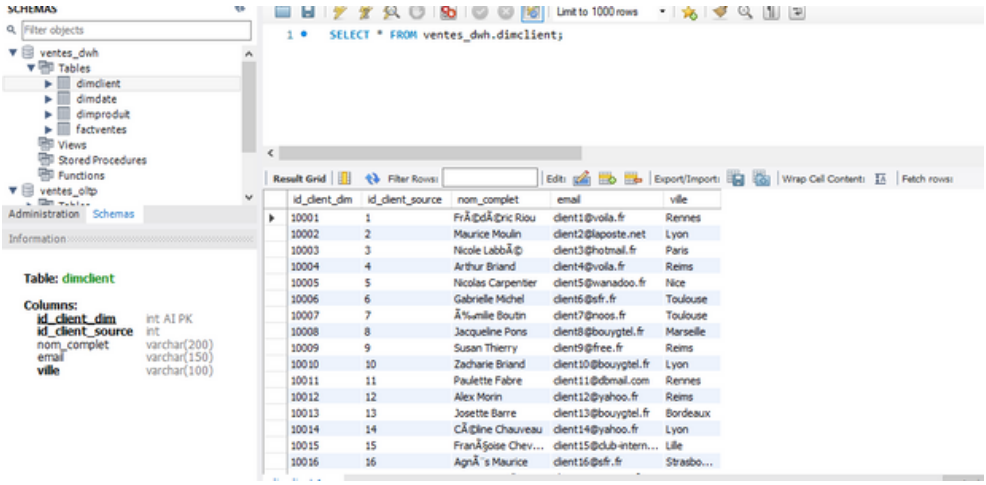
6. Processus ETL avec Pentaho

Le processus ETL a été conçu dans Pentaho Data Integration (Spoon) pour migrer et transformer les données de l'OLTP vers le DWH de manière automatisée et reproductible.

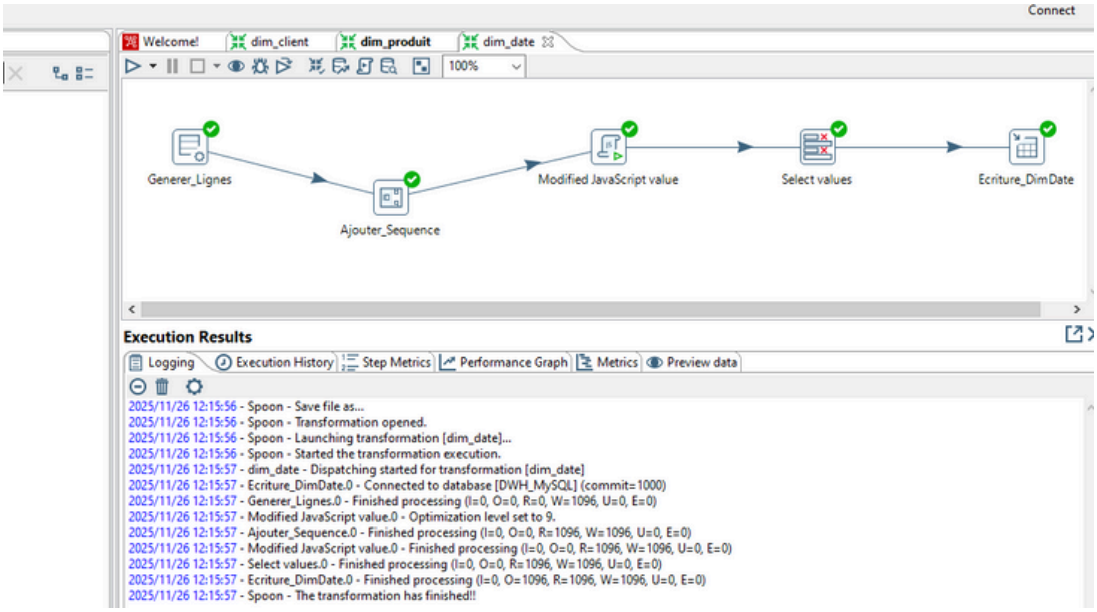
Transformations développées :

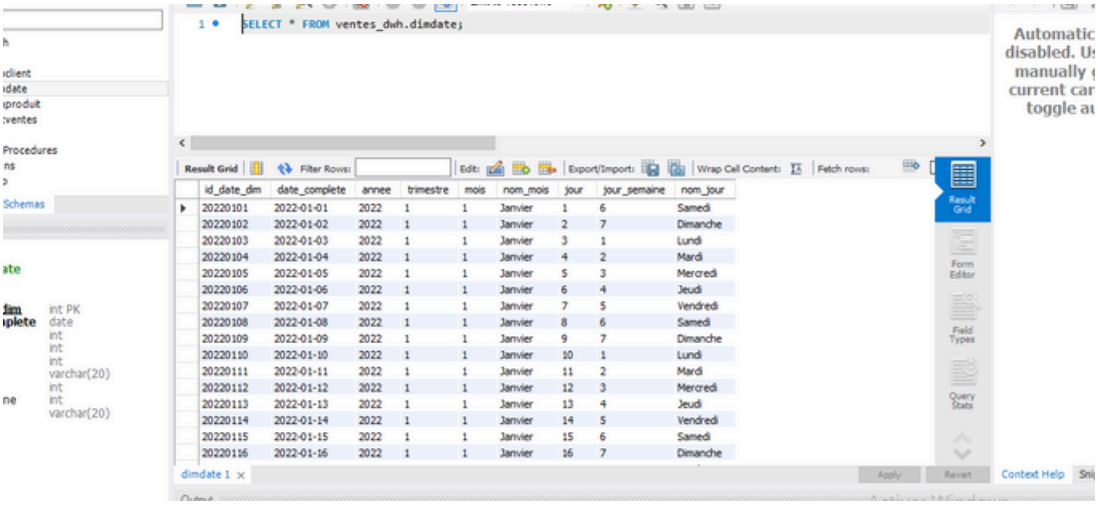
1-dim_client.ktr : Lit la table clients, concatène le nom et prénom, et charge DimClient.



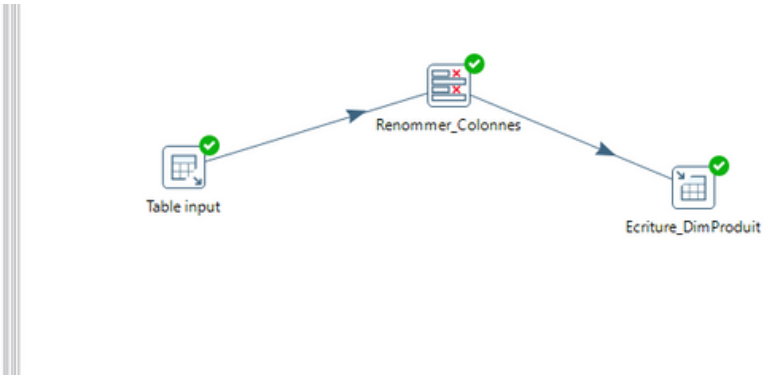


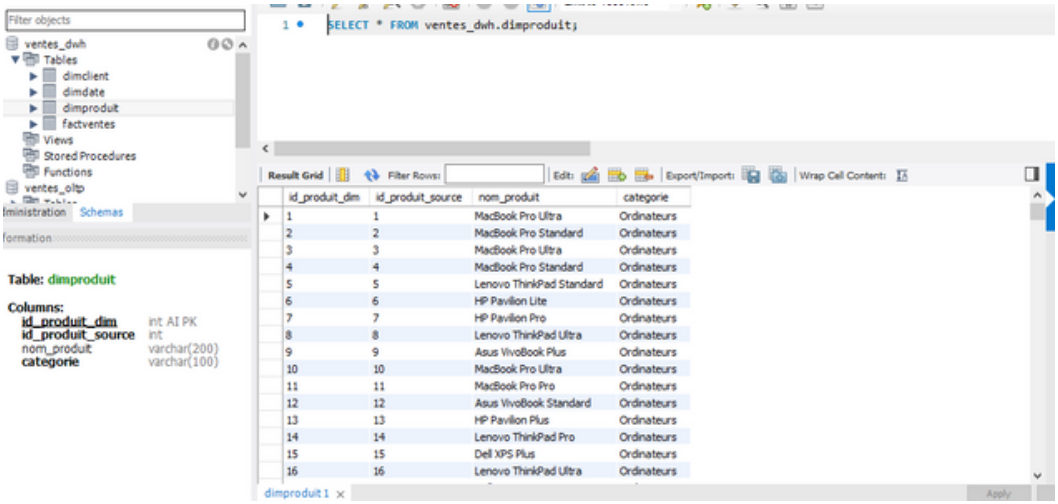
2-dim_date.ktr : Génère dynamiquement la table de temps DimDate pour la période 2022-2024.





3-dim_produit.ktr : Lit la table produits et charge DimProduit.





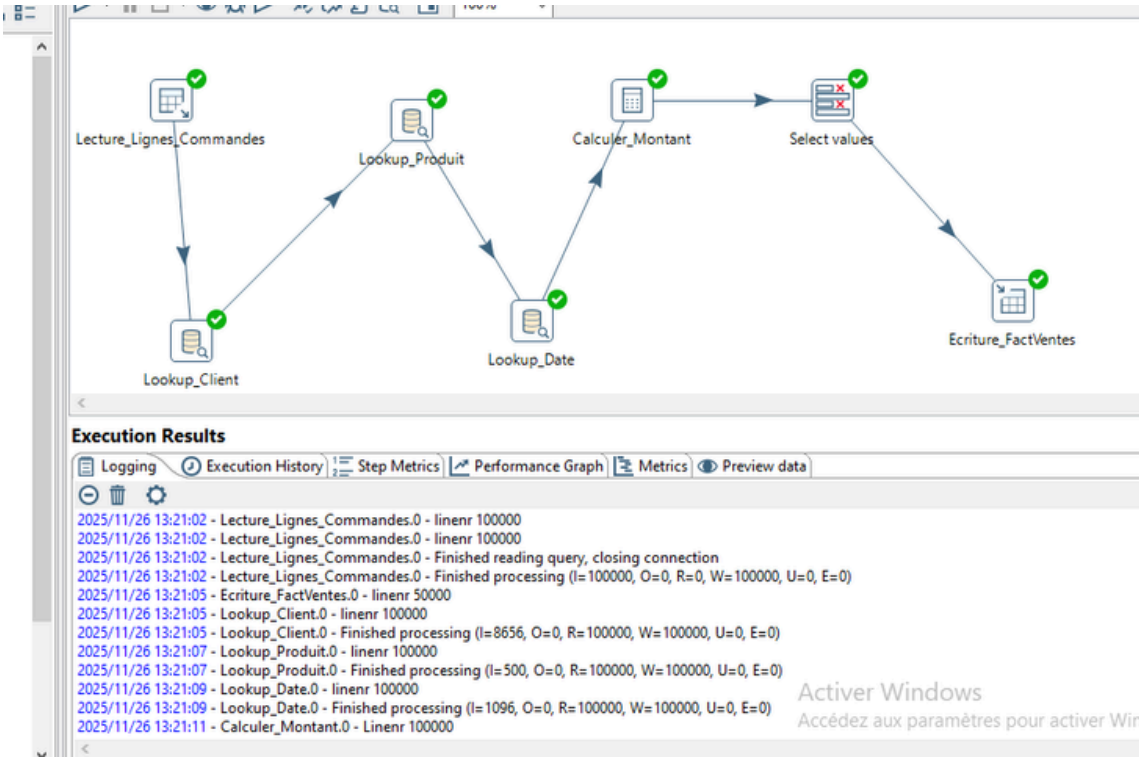
4-fact_ventes.ktr : La transformation principale. Elle :

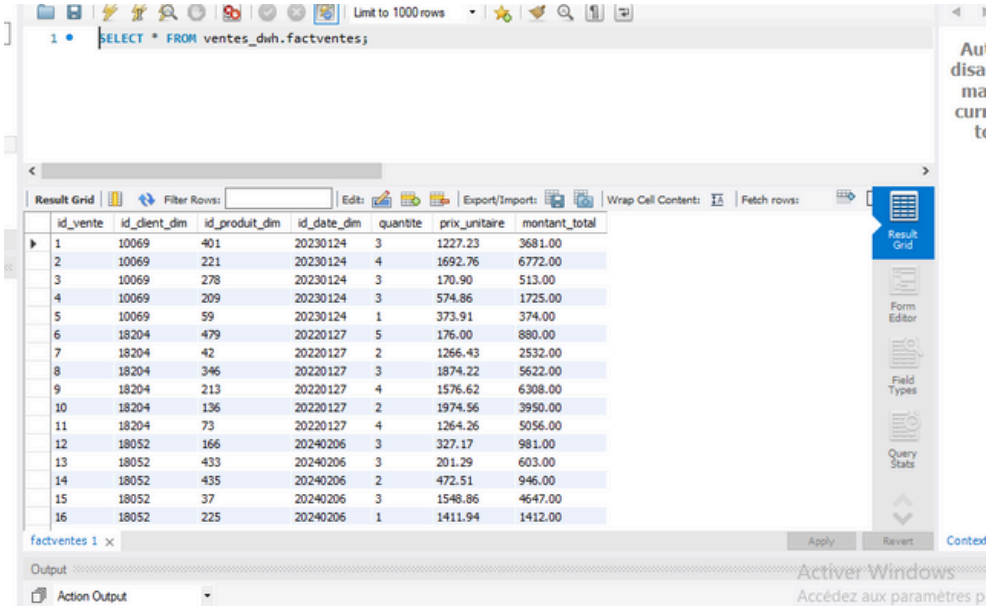
-Joint les tables lignes_commandes et commandes.

-Effectue trois "lookups" pour récupérer les clés de substitution (id_client_dim, id_produit_dim, id_date_dim) depuis les dimensions du DWH.

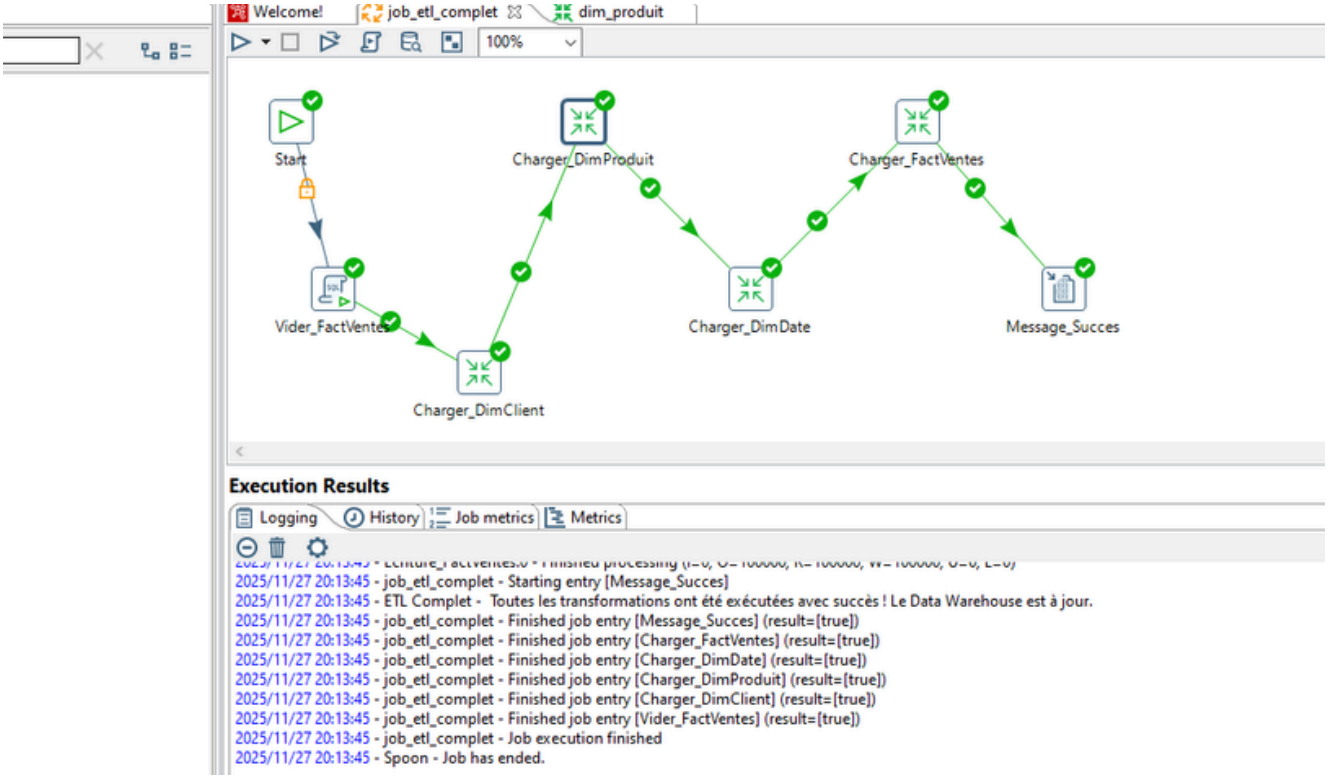
-Calcule le montant_total (quantité * prix).

-Charge les faits dans la table FactVentes.





Orchestration : Un job principal (job_etl_complet.kjb) a été créé pour exécuter les transformations dans l'ordre logique (dimensions d'abord, faits ensuite) et gérer les dépendances.

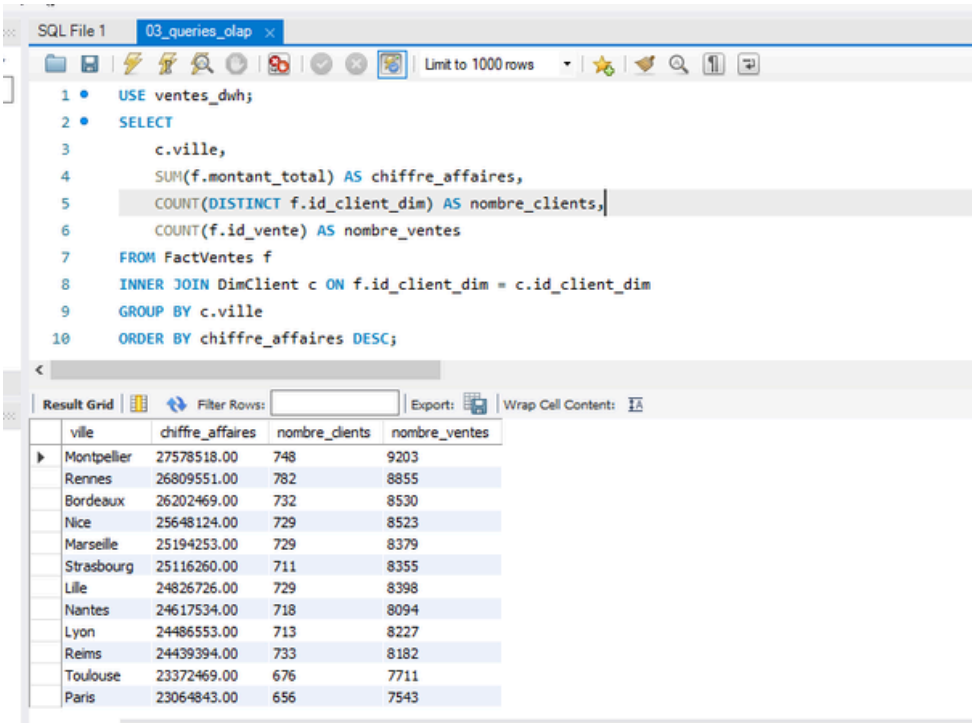


7- Analyses OLAP

Avec le DWH en place, des requêtes analytiques complexes deviennent simples et rapides.

Requêtes et Interprétation Métier :

-CA par Ville : Montpellier, Rennes et Bordeaux arrivent en tête. Cela valide l'opportunité de cibler les campagnes marketing sur ces zones à fort potentiel.



-CA par Catégorie : La catégorie "Ordinateurs" génère à elle seule près de 45% du chiffre d'affaires, confirmant son statut de moteur principal des ventes.

SQL File 1 | 03_queries_olap

Limit to 1000 rows

```sql
SELECT
    p.categorie,
    SUM(f.montant_total) AS chiffre_affaires,
    SUM(f.quantite) AS quantite_vendue,
    COUNT(DISTINCT f.id_produit_dim) AS nombre_produits_distincts,
    ROUND(AVG(f.prix_unitaire), 2) AS prix_moyen
FROM FactVentes f
INNER JOIN DimProduit p ON f.id_produit_dim = p.id_produit_dim
GROUP BY p.categorie
ORDER BY chiffre_affaires DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

categorie	chiffre_affaires	quantite_vendue	nombre_produits_distincts	prix_moyen
Ordinateurs	64825326.00	60514	100	1070.22
TÃ©lÃ©phones	64526264.00	60122	100	1072.42
Accessoires	60129964.00	60127	100	1000.07
Tablettes	56774487.00	59296	100	953.92
Montres	55100653.00	59737	100	919.41

Évolution Mensuelle : Une saisonnalité marquée est observée avec des pics systématiques en novembre-décembre (période des fêtes). Une recommandation serait d'anticiper ces pics par des préparatifs logistiques.

SQL File 1 | 03_queries_olap

Limit to 1000 rows

```sql
SELECT
    d.annee,
    d.mois,
    d.nom_mois,
    SUM(f.montant_total) AS chiffre_affaires,
    COUNT(f.id_vente) AS nombre_ventes,
    ROUND(AVG(f.montant_total), 2) AS panier_moyen
FROM FactVentes f
INNER JOIN DimDate d ON f.id_date_dim = d.id_date_dim
GROUP BY d.annee, d.mois, d.nom_mois
ORDER BY d.annee, d.mois;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

annee	mois	nom_mois	chiffre_affaires	nombre_ventes	panier_moyen
2022	1	Janvier	8897711.00	3003	2962.94
2022	2	Février	8141180.00	2655	3066.36
2022	3	Mars	8611136.00	2892	2977.57
2022	4	Avril	8364191.00	2767	3022.84
2022	5	Mai	7941259.00	2705	2935.77
2022	6	Juin	8000701.00	2602	3074.83
2022	7	Juillet	8454878.00	2887	2928.60
2022	8	Août	8361499.00	2740	3051.64
2022	9	Septembre	8268176.00	2717	3043.13
2022	10	Octobre	8026616.00	2688	2986.09
2022	11	Novembre	8328418.00	2760	3017.54
2022	12	Décembre	8354504.00	2770	3016.07

-Top 10 Produits : 70% des ventes proviennent de seulement 10 références. Ces produits "stars" doivent être particulièrement surveillés en termes de stock et de mise en avant.

SQL File 1 | 03_queries_olap

Limit to 1000 rows

```sql
SELECT
    p.nom_produit,
    p.categorie,
    SUM(f.quantite) AS quantite_totale_vendue,
    SUM(f.montant_total) AS chiffre_affaires,
    ROUND(AVG(f.prix_unitaire), 2) AS prix_moyen
FROM FactVentes f
INNER JOIN DimProduit p ON f.id_produit_dim = p.id_produit_dim
GROUP BY p.id_produit_dim, p.nom_produit, p.categorie
ORDER BY quantite_totale_vendue DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

nom_produit	categorie	quantite_totale_vendue	chiffre_affaires	prix_moyen
Apple Watch Plus	Montres	751	1334527.00	1776.81
Clavier mÃ©canique Ultra	Accessoires	740	1277240.00	1726.43
Huawei MatePad Lite	Tablettes	734	136524.00	186.37
Xiaomi 13 Plus	TÃ©lÃ©phones	725	1046175.00	1442.98
HP Pavilion Plus	Ordinateurs	714	369852.00	518.16
iPhone 14 Standard	TÃ©lÃ©phones	707	1260581.00	1782.59
Asus VivoBook Ultra	Ordinateurs	704	1265792.00	1797.82
Garmin Forerunner Lite	Montres	704	772992.00	1097.54
Webcam HD Ultra	Accessoires	703	1120582.00	1594.23
OnePlus 11 Standard	TÃ©lÃ©phones	702	819234.00	1167.47

Result 4

Limit to 1000 rows

```sql
ORDER BY quantite_totale_vendue DESC
LIMIT 10;
SELECT
    d.annee,
    d.trimestre,
    p.categorie,
    SUM(f.montant_total) AS chiffre_affaires,
    SUM(f.quantite) AS quantite_vendue
FROM FactVentes f
INNER JOIN DimDate d ON f.id_date_dim = d.id_date_dim
INNER JOIN DimProduit p ON f.id_produit_dim = p.id_produit_dim
GROUP BY d.annee, d.trimestre, p.categorie
ORDER BY d.annee, d.trimestre, chiffre_affaires DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

annee	trimestre	categorie	chiffre_affaires	quantite_vendue
2022	1	TÃ©lÃ©phones	5474825.00	5004
2022	1	Ordinateurs	5381872.00	5081
2022	1	Accessoires	5157471.00	5135
2022	1	Tablettes	4935860.00	5287
2022	1	Montres	4699999.00	5208
2022	2	Ordinateurs	5573268.00	5183
2022	2	TÃ©lÃ©phones	5493563.00	5091
2022	2	Accessoires	4660607.00	4617
2022	2	Tablettes	4392008.00	4516

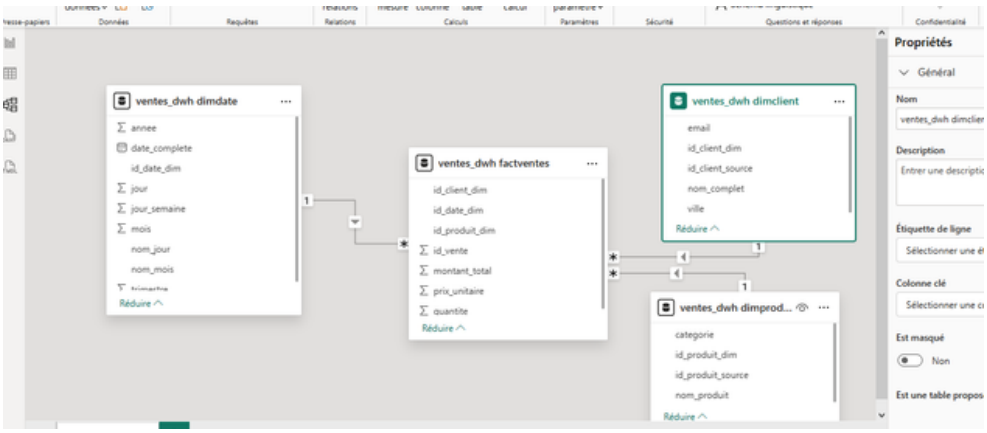
8. Visualisations Power BI

8. Visualisations Power BI

Le tableau de bord interactif synthétise l'ensemble des analyses pour une prise de décision rapide.

Description du Dashboard :

Le rapport Power BI se connecte directement au DWH (ventes_dwh) .



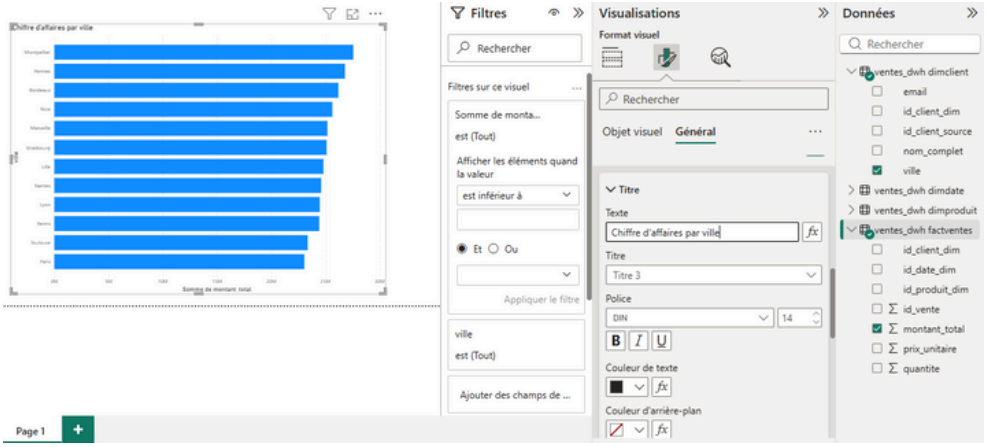
et présente :

-Des indicateurs clés (KPI) : CA Total, Panier Moyen, Nombre de Ventes, Clients Uniques.

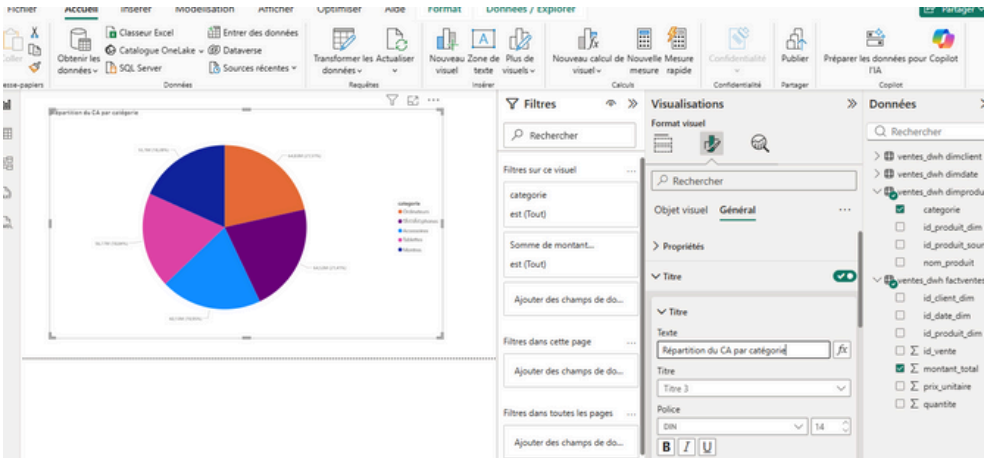
-Des segments (slicers) interactifs : Filtrage par Année et Catégorie.

-Quatre visualisations principales :

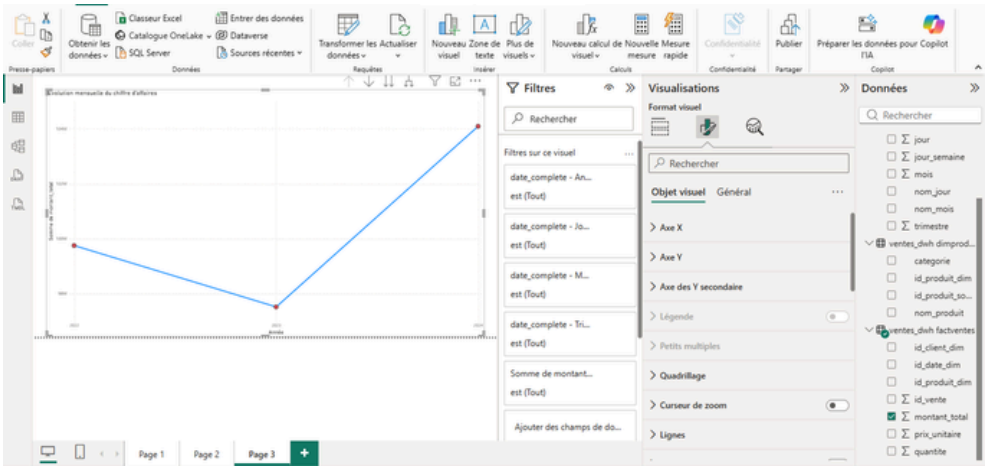
-*CA par Ville (Graphique à barres) :* Permet d'identifier les marchés géographiques les plus rentables.



Répartition du CA par Catégorie (Graphique en secteurs) : Montre la contribution de chaque famille de produits.



-Évolution Mensuelle du CA (Graphique en courbes) : Visualise les tendances et la saisonnalité.



Insights Métier Tirés :
Le dashboard permet de constater en un clin d'œil que la performance de TechStore est tirée par les ventes d'ordinateurs à Montpellier pendant le dernier trimestre. Il facilite ainsi des décisions telles que l'ajustement des stocks, le ciblage des campagnes publicitaires, ou l'évaluation de l'opportunité d'ouvrir un point de vente physique.

9. Conclusions et Recommandations

Synthèse : Ce projet a permis de construire avec succès un pipeline BI opérationnel pour TechStore. L'architecture mise en place répond aux problématiques initiales en fournissant des insights rapides, fiables et actionnables à partir des données transactionnelles.
Apprentissages :

- Maîtrise de la différence fondamentale entre modèles OLTP (normalisé) et OLAP (dénormalisé).
- Expérience pratique sur l'outil ETL Pentaho pour l'automatisation des flux.
- Conception d'un schéma en étoile adapté aux besoins analytiques.
- Création d'un tableau de bord interactif orienté métier avec Power BI.

Améliorations Possibles :

1. Industrialisation : Planifier l'exécution du job Pentaho via Kitchen et un planificateur de tâches (cron/Windows Scheduler) pour un rafraîchissement quotidien automatique.
2. Enrichissement des Données : Intégrer des données externes (météo, jours fériés) dans la dimension DimDate pour affiner l'analyse des ventes.
3. Power BI Cloud : Publier le rapport sur le service Power BI pour le partage en ligne et l'accès mobile.
4. Gestion des Dimensions Lentes (SCD) : Implémenter dans Pentaho la gestion des changements historiques dans les dimensions (ex: suivi des changements d'adresse client).