Rapport de Fin de Projet BI

Mise en place d'un système décisionnel complet

Étude de Cas : Northwind Traders

Étudiant :
Djebbar Seddik Adel

Enseignant :
Mekahlia Fatma Zohra

Année Universitaire 2024-2025

Table des matières

Chapitre 1

Introduction Générale

1.1 Contexte du Projet

Dans le paysage économique actuel, la donnée est devenue l'actif le plus précieux pour une entreprise. *Northwind Traders*, une entreprise internationale spécialisée dans l'importation et l'exportation de produits alimentaires, ne fait pas exception à cette règle. Malgré un volume de transactions important, l'entreprise souffrait d'une fragmentation de ses systèmes d'information. Les données de ventes historiques étaient stockées sur un serveur **SQL Server**, tandis que les nouvelles commandes et les données de terrain étaient saisies dans des bases de données locales **MS Access**.

1.2 Problématique

L'absence d'une source de vérité unique (Single Source of Truth) posait plusieurs problèmes majeurs :

- **Incohérence des données** : Des doublons apparaissaient entre les deux systèmes.
- **Difficulté de reporting** : Il était impossible d'obtenir une vision consolidée du chiffre d'affaires en temps réel.
- **Analyse limitée** : Les outils de visualisation ne pouvaient pas se connecter efficacement à des sources hétérogènes sans risquer des erreurs de calcul.

1.3 Objectifs et Périmètre

L'objectif principal de ce projet est la mise en place d'un **système décisionnel (Business Intelligence)** complet. Ce système doit être capable d'extraire, transformer et charger les données dans un entrepôt de données (Data Warehouse) structuré en schéma en étoile. Les objectifs spécifiques incluent :

- La consolidation de plus de 800 commandes issues de deux sources différentes.
- Le nettoyage automatisé des données (déduplication, gestion des valeurs nulles).
- La création de KPIs stratégiques tels que le *SalesAmount* et le *Taux de livraison*.

1.4 Méthodologie de Travail

Pour mener à bien ce projet, nous avons adopté une démarche itérative :

1. **Phase d'Analyse** : Audit des tables SQL Server et Access.
2. **Phase ETL** : Développement de scripts Python pour le traitement des données.
3. **Phase de Modélisation** : Conception physique de la base de données *Northwind_DW*.
4. **Phase de Visualisation** : Création de graphiques d'analyse de tendance et de performance.

1.5 Organisation du Rapport

Le présent rapport est structuré comme suit : le deuxième chapitre présente l'analyse détaillée des sources de données. Le troisième chapitre décrit le processus technique de transformation (ETL). Le quatrième chapitre est dédié à l'architecture de l'entrepôt de données, et enfin, le cinquième chapitre présente l'analyse des résultats à travers des tableaux de bord.

Chapitre 2

Analyse des Sources de Données

2.1 Présentation de l'Environnement de Stockage

La réussite d'un projet de Business Intelligence repose sur une compréhension fine des données sources. Dans notre cas, nous avons dû composer avec deux technologies de stockage différentes, chacune possédant ses propres contraintes de types de données, de formats et de relations. Cette hétérogénéité est le défi majeur de la phase d'intégration.

2.2 Source Primaire : Microsoft SQL Server

Le serveur de production SQL Server héberge l'essentiel de l'activité historique de l'entreprise. C'est une base de données relationnelle robuste dont nous avons extrait les dimensions et les faits principaux.

2.2.1 Structure des tables extraites

- **Table** `Orders` : Contient l'en-tête des transactions (dates, employé responsable, identifiant client). Elle sert de pivot pour la chronologie des ventes.
- **Table** `Order Details` : Contient les lignes de commande détaillées. C'est ici que se trouvent les quantités et les prix unitaires, données essentielles pour le calcul du chiffre d'affaires ($Prix \times Quantit$).
- **Table** `Customers` : Un référentiel complet contenant les informations sur les clients, incluant les noms de sociétés et les données géographiques.

2.2.2 Analyse de la volumétrie et Qualité

À l'issue de l'audit, la source SQL Server présentait environ **830 commandes**. La qualité des données y est globalement élevée, bénéficiant de contraintes d'intégrité référentielle (clés primaires et étrangères) déjà établies en amont.

2.3 Source Secondaire : Microsoft Access

La base Access représente les données de "terrain", souvent saisies manuellement pour compléter les lacunes du système principal ou enregistrer les ventes les plus récentes non encore synchronisées.

2.3.1 Défis liés à la source Access

— **Incohérence des formats** : Les noms de colonnes contiennent souvent des espaces (ex : "Order Date" au lieu de "OrderDate"), ce qui complique les scripts de traitement automatique.
— **Doublons potentiels** : Certaines commandes d'Access peuvent être des doublons ou des corrections de commandes déjà existantes dans SQL Server.
— **Données éparses** : Présence importante de champs vides (`NULL` ou `NaN`) dans les colonnes de livraison (*Shipped Date*).

2.3.2 Inventaire des données Access

Nous avons extrait environ **48 commandes supplémentaires**. Bien que le volume soit plus faible, ces données sont critiques car elles portent sur l'année 1998, période clé pour l'analyse de la tendance actuelle de l'entreprise.

2.4 Cartographie des Correspondances (Mapping)

Le tableau suivant résume la matrice de passage entre les champs hétérogènes des deux sources vers la structure cible du Data Warehouse :

Entité	**Champ SQL Server**	**Champ Access**	**Cible Warehouse**
Client	CompanyName	Company	Company
Date	OrderDate	Order Date	OrderDateKey
Employé	FirstName + LastName	First Name / Last Name	FullName
Prix	UnitPrice	Unit Price	SalesAmount

Table 2.1 – Matrice de correspondance entre les sources hétérogènes

2.5 Synthèse de la Phase d'Analyse

L'analyse montre que la simple fusion des deux sources produirait des erreurs de calcul. Une étape de transformation est nécessaire pour normaliser les noms, supprimer les doublons et traiter les dates manquantes afin d'obtenir une base de données analytique fiable.

Chapitre 3

Processus ETL (Extraction, Transformation, Load)

3.1 Introduction au Pipeline de Données

Le cœur technique de ce projet réside dans la mise en place d'un pipeline ETL robuste. Le but est de transformer des données brutes, parfois incohérentes, en une structure optimisée pour l'analyse. Nous avons utilisé le langage **Python** avec la bibliothèque **Pandas** pour sa puissance de manipulation de données, ainsi que **SQLAlchemy** pour assurer une communication fluide avec l'entrepôt de données final.

3.2 Phase d'Extraction (E)

L'extraction a consisté à récupérer les données de deux sources techniquement distinctes :

- **SQL Server** : Connexion via le driver `ODBC 17` pour extraire les tables *Orders*, *Order_Details*, et *Employees*.
- **MS Access** : Chargement des fichiers `.mdb` ou `.accdb` contenant les données de saisie complémentaire.

3.3 Phase de Transformation (T)

C'est l'étape la plus critique du projet, visant à garantir la "propreté" (Data Cleaning) du Data Warehouse.

3.3.1 Standardisation et Nettoyage

Les données provenant d'Access présentaient des noms de colonnes avec des espaces. Nous avons automatisé leur normalisation pour éviter les erreurs de syntaxe dans les requêtes SQL :

```
# Normalisation des colonnes (Suppression des espaces)
for df in [customers, orders_access]:
    df.columns = df.columns.str.replace(' ', '', regex=False)
```

3.3.2 Gestion de l'Identité et Fusion

Pour les dimensions **Client** et **Employé**, nous avons fusionné les noms et prénoms pour créer des champs uniques plus lisibles (ex : `FullName`, `ContactName`) et supprimé les colonnes inutiles pour l'analyse comme les numéros de fax ou les adresses détaillées.

3.3.3 Déduplication et Intégrité

Certaines commandes se trouvaient à la fois dans SQL Server et Access. Nous avons appliqué une règle de déduplication basée sur les clés primaires (`OrderID`) pour ne conserver que la version la plus complète.

3.3.4 Traitement des Commandes Non Livrées (Clé Sentinelle)

L'un des défis majeurs était la présence de valeurs nulles dans les dates de livraison (*ShippedDate*). Dans un environnement BI, les valeurs NULL perturbent les jointures avec la dimension Temps.

> **Solution technique :** Nous avons remplacé toutes les dates de livraison manquantes par la clé technique **1011900**. Dans notre dimension *DimTime*, cette clé correspond à un enregistrement fictif désignant un état "En attente de livraison".

3.4 Phase de Chargement (L)

Le chargement dans la base `Northwind_DW` a été réalisé selon une stratégie de rafraîchissement total (*Full Load*).

3.4.1 Chargement de la Table de Faits

La table de faits (*FactSales*) a été construite en joignant les tables de commandes avec les dimensions déjà nettoyées. Le calcul du montant total par ligne a été effectué au préalable :

$$\text{SalesAmount} = (\text{UnitPrice} \times \text{Quantity}) \times (1 - \text{Discount}) \tag{3.1}$$

3.5 Validation et Contrôle Qualité

À l'issue du processus, nous avons vérifié que le nombre total de lignes dans le Data Warehouse correspondait à la somme des sources moins les doublons identifiés (soit environ 878 enregistrements uniques). Cette validation garantit qu'aucune donnée n'a été perdue lors du transfert.

Chapitre 4

Architecture et Modélisation du Data Warehouse

4.1 Le Choix du Modèle Dimensionnel

Pour répondre aux besoins analytiques de *Northwind Traders*, nous avons opté pour une modélisation multidimensionnelle. Contrairement aux bases de données transactionnelles (OLTP) qui sont normalisées pour éviter la redondance, un Data Warehouse (OLAP) privilégie la performance des requêtes et la facilité de lecture pour les utilisateurs finaux.

4.2 Schéma en Étoile (Star Schema)

Le cœur de notre entrepôt de données repose sur un **Schéma en Étoile**. Ce choix se justifie par la simplicité des jointures, permettant d'obtenir des résultats rapides lors de l'agrégation de gros volumes de données.

Notre schéma se compose d'une table de faits centrale entourée de quatre tables de dimensions :

4.2.1 La Table de Faits : `FactSales`

C'est la table centrale qui stocke les indicateurs quantitatifs (mesures) associés aux événements métiers (les ventes). Chaque ligne représente une transaction unique.

— **Mesures** : `Quantity`, `UnitPrice`, `Discount`, et la mesure calculée `SalesAmount`.
— **Clés Étrangères** : `CustomerKey`, `EmployeeKey`, `ProductKey`, `OrderDateKey`, `ShippedDateKey`.

4.2.2 Les Tables de Dimensions

Les dimensions permettent de filtrer et de regrouper les données de la table de faits :

1. **DimCustomers** : Contient les attributs descriptifs des clients (Société, Ville, Pays).
2. **DimEmployees** : Stocke les informations sur le personnel de vente, notamment le champ fusionné `FullName`.
3. **DimProducts** : Répertorie les articles vendus, leurs catégories et fournisseurs.
4. **DimDate** : Dimension cruciale qui permet une analyse temporelle par année, trimestre et mois.

4.3 Intégrité Référentielle et Clés Techniques

Pour assurer la robustesse du système, nous avons remplacé les identifiants métier (Business Keys) par des **clés substitutives (Surrogate Keys)**.

L'un des points techniques les plus importants de notre modélisation est la gestion des faits orphelins :

— Les commandes non encore expédiées pointent vers l'enregistrement **1011900** dans la table `DimDate`.

— Cet enregistrement possède des attributs textuels tels que "Date non définie" ou "En attente", évitant ainsi l'utilisation de valeurs nulles qui pourraient fausser les calculs de moyennes ou de totaux.

4.4 Matrice de Granularité

La granularité choisie pour ce Data Warehouse est la **ligne de commande (Order Line)**. Ce niveau de détail le plus fin permet de reconstruire n'importe quel agrégat supérieur (par jour, par pays ou par catégorie de produit) sans perte d'information.

Chapitre 5

Analyse des Résultats et Visualisation

5.1 Introduction à l'Analyse Décisionnelle

Une fois l'entrepôt de données constitué et alimenté, la dernière phase consiste à extraire de la valeur de ces données. Cette section présente les indicateurs clés de performance (KPI) et les visualisations générées pour faciliter la prise de décision au sein de *Northwind Traders*.

5.2 Analyse de la Performance Commerciale

L'un des premiers objectifs était d'étudier la santé financière de l'entreprise à travers l'évolution de son chiffre d'affaires.

5.2.1 Évolution Temporelle des Ventes

Le graphique de tendance mensuelle montre une activité stable en 1996 et 1997, suivie d'une accélération brutale au début de l'année 1998.

- **Observation :** Le pic de revenus enregistré en avril 1998 (dépassant les 120 000 $) coïncide avec l'intégration réussie des données provenant de la source MS Access.
- **Interprétation :** Cette hausse n'est pas seulement due à une augmentation des ventes, mais aussi à une meilleure visibilité sur les données "terrain" qui étaient auparavant invisibles dans le système SQL Server seul.

5.3 Analyse de l'Efficacité Logistique

Le tableau de bord logistique nous permet d'évaluer la réactivité de l'entreprise face aux commandes clients.

5.3.1 Répartition des Expéditions

Grâce à l'utilisation de la clé sentinelle `1011900`, nous avons pu quantifier précisément le retard de livraison :

- **Taux de service :** 94,5 % des commandes sont expédiées.
- **Commandes en attente :** Environ 5,5 % des transactions n'ont pas encore quitté l'entrepôt.

5.3.2 Analyse Géographique des Retards

Une analyse par pays révèle une concentration anormale des retards aux **États-Unis (USA)** avec 27 commandes en attente, suivis du Venezuela (8) et de l'Autriche (7). *Recommandation stratégique :* L'entreprise doit auditer ses partenaires logistiques sur le territoire nord-américain pour identifier les causes de ces goulots d'étranglement.

5.4 Performance des Ressources Humaines

Le croisement entre les faits de vente et la dimension employés permet d'identifier les piliers de l'entreprise.

- **Volume de ventes :** Margaret Peacock se distingue comme l'employée la plus productive avec le plus grand nombre de dossiers traités.
- **Gestion des retards :** L'analyse montre que la charge de travail est inégalement répartie, certains employés gérant un volume de commandes en attente plus élevé, ce qui pourrait nécessiter une redistribution des zones géographiques.

5.5 Synthèse Décisionnelle

Grâce à ce système BI, *Northwind Traders* dispose désormais d'un outil capable de :

1. Surveiller en temps réel la croissance des revenus.
2. Identifier les zones géographiques prioritaires pour l'amélioration logistique.
3. Évaluer objectivement la contribution de chaque membre de l'équipe commerciale.

Conclusion Générale

Ce projet de Business Intelligence a permis de franchir une étape cruciale dans la gestion des données de l'entreprise *Northwind Traders*. En partant de sources fragmentées et hétérogènes (SQL Server et MS Access), nous avons réussi à construire un pipeline ETL automatisé et un entrepôt de données structuré.

La mise en place du schéma en étoile a prouvé son efficacité en offrant des temps de réponse rapides pour les analyses complexes. Au-delà de l'aspect technique, les visualisations obtenues apportent une valeur métier immédiate, notamment en mettant en lumière les retards logistiques spécifiques au marché américain.

En perspective, ce travail pourrait être étendu par l'ajout de données de prédiction (Machine Learning) pour anticiper les pics de commandes mensuels, ou par l'intégration d'une dimension "Transporteurs" pour analyser l'impact des coûts d'expédition sur la marge nette de l'entreprise.