République Algérienne Démocratique Populaire
Université des Sciences et de la Technologie Houari Boumediene
Faculté d'Informatique

Projet Business Intelligence : Analyse de la base de données Northwind

-Réalisé par:

BOUCHAREB.
Anfal.
232331550809.
Section A 3eme ING Cybersécurité.

-Sommaire:

1-Introduction:

La Business Intelligence (BI) regroupe l'ensemble des méthodes, outils et technologies permettant de collecter, transformer, analyser et visualiser les données d'une organisation dans le but de soutenir la prise de décision. Elle s'appuie sur l'analytique métier, le data mining, la visualisation des données ainsi que sur des infrastructures de gestion des données et des bonnes pratiques associées. Grâce à la BI, les entreprises disposent d'une vision globale et cohérente de leurs données, leur permettant d'identifier des tendances, d'optimiser leurs performances et de s'adapter efficacement aux évolutions du marché.

Dans ce cadre, le présent projet porte sur l'exploitation de la base de données Northwind, une base de données décisionnelle couramment utilisée à des fins pédagogiques, représentant les activités commerciales d'une entreprise (clients, employés, commandes, produits, etc.). L'objectif principal du projet est de mettre en place un processus ETL (Extract, Transform, Load) complet afin de centraliser, nettoyer et intégrer des données provenant de plusieurs sources hétérogènes, notamment des fichiers Microsoft Access et des tables SQL Server.

Les données transformées sont ensuite chargées dans un Data Warehouse, conçu pour faciliter l'analyse décisionnelle. Sur la base de ce Data Warehouse, plusieurs indicateurs clés de performance (KPI) ont été définis et calculés, puis visualisés à travers un dashboard interactif, permettant une lecture claire et synthétique des performances de l'entreprise et facilitant ainsi la prise de décision stratégique.

2-Concepts Théoriques:

2.1- BI : Business Intelligence:

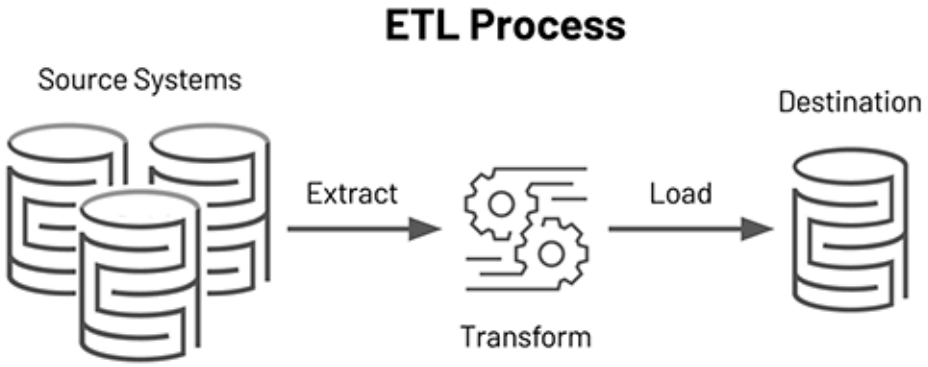
La Business Intelligence (BI) est un processus technologique d'analyse des données et de présentation d'informations pour aider les dirigeants, managers et autres utilisateurs finaux de l'entreprise à prendre des décisions business éclairées. La Business Intelligence englobe une grande variété d'outils, d'applications et de méthodologies qui permettent aux organisations de collecter des données à partir de systèmes internes et de sources externes. Ces données sont ensuite préparées pour l'analyse afin de créer des rapports, tableaux de bord et autres outils pour rendre les résultats analytiques disponibles aux décideurs et aux opérations.

2.2-Data Warehouse:

Un data warehouse est un type de système de gestion de données conçu pour permettre et faciliter les activités de business intelligence (BI), en particulier l'analytique. Les data warehouses sont uniquement destinés à effectuer des requêtes et des analyses. Ils contiennent souvent de grandes quantités de données historiques. Les données contenues dans un data warehouse proviennent généralement d'un large éventail de sources

2.3-ETL:

L'ETL (Extract, Transform, Load) est un processus fondamental en Business Intelligence qui consiste à extraire des données à partir de différentes sources, à les transformer afin d'assurer leur qualité, leur cohérence et leur compatibilité, puis à les charger dans un système de stockage cible, généralement un Data Warehouse. L'objectif de l'ETL est de rendre les données fiables, structurées et exploitables pour l'analyse décisionnelle et la production de rapports.



1.Extract (Extraire):

La phase initiale du processus vise à récupérer les données à partir de différentes sources, souvent de nature variée. Ces sources peuvent inclure des systèmes opérationnels, des bases de données transactionnelles, des interfaces de programmation (API)

2.Transform (Transformer):

La deuxième étape consiste à transformer les données brutes extraites des sources sous une forme utilisable par différentes applications. À ce stade, les données sont nettoyées, mappées et transformées, souvent selon un schéma spécifique pour répondre aux besoins opérationnels. Ce processus implique plusieurs types de transformation qui assurent la qualité et l'intégrité des données. Généralement, les données ne sont pas directement chargées dans la source de destination : il est courant de les charger dans une base de données de préparation. Cette étape permet d'effectuer une annulation rapide en cas de problème imprévu. Pendant cette phase, vous aurez la possibilité de générer des rapports d'audit à des fins de conformité réglementaire, ou de diagnostiquer et réparer les problèmes de données.

3.Load (Charger):

Enfin, la fonction de chargement consiste à écrire les données converties et stockées dans un emplacement préparatoire dans une base de données de destination, existante ou non. Selon les exigences de l'application, ce processus peut être très simple ou plus complexe. Chacune de ces étapes peut être réalisée à l'aide d'outils ETL ou de code personnalisé.

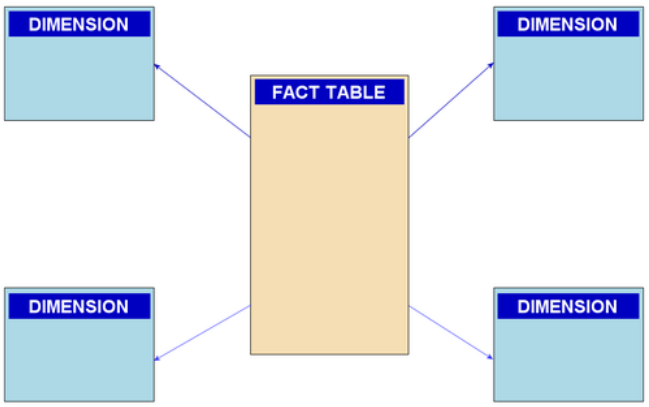
2.3-Les Indicateurs Clés de Performance KPI:

Le terme *KPI* (Key Performance Indicator), ou indicateur clé de performance, désigne une donnée chiffrée utilisée pour mesurer l'atteinte d'un objectif. Qu'il s'agisse de suivre l'avancement d'un projet, d'évaluer l'efficacité d'un processus ou d'orienter une stratégie, la mise en place de KPI pertinents permet de piloter avec méthode.
Bien plus qu'un simple chiffre, un KPI aide à prendre des décisions stratégiques. Il offre une visibilité claire sur les résultats, alerte en cas d'écart et favorise un alignement durable entre les objectifs et les actions des équipes.

2.4-Modèle de Données en Étoile :

Un schéma en étoile, ou modèle de données en étoile, est une structure multidimensionnelle stockant des données atomiques ou agrégées, typiquement dans des datawarehouse ou datamart. Souvent considéré comme un modèle dénormalisé[1], le modèle en étoile permet une économie de jointures à l'interrogation, ce qui le rend optimisé pour les requêtes d'analyse.

Le modèle en étoile est implémenté sur un SGBD relationnel classique tel que Oracle, IBM DB2, SQL Server, Teradata, MySQL ou encore PostgreSQL.



Un modèle en étoile est constitué d'une table centrale, dite table des faits, et de nombreuses tables dimensionnelles autour.

La table située au centre de l'étoile, table des faits ou mesures (ou encore métriques), contient les éléments mesurés dans l'analyse comme les montants, les quantités, les taux, etc.

Les tables situées aux extrémités de l'étoile, tables de dimensions, ou encore axes d'analyse ; niveaux de suivi sont les dimensions explorées dans l'analyse comme le temps (jour, mois, période, etc.), la nomenclature des produits (référence, famille, etc.), la segmentation clientèle (tranche d'âge, marché, etc.), etc.

2.5-Northwind:

La base de données Northwind est une base de données relationnelle utilisée principalement à des fins pédagogiques pour l'apprentissage et la démonstration des concepts de gestion et d'analyse des données. Elle représente le système d'information d'une entreprise commerciale fictive spécialisée dans la vente et la distribution de produits. La base Northwind contient des informations structurées sur les clients, les employés, les produits, les fournisseurs, les commandes et les détails de commandes, ce qui en fait un support idéal pour l'étude des bases de données, du SQL et de la Business Intelligence.

3-Présentation des outils utilisés:

3.1-Visual Studio Code (VS Code) :

c'est un éditeur de code léger et puissant développé par Microsoft, utilisé pour le développement de logiciels dans plusieurs langages de programmation. Dans le cadre de ce projet, VS Code a été utilisé pour écrire et exécuter le code Python dédié au processus ETL, au calcul des indicateurs clés de performance (KPI), ainsi qu'à la réalisation du dashboard de visualisation.

3.2-SQL Server Management Studio (SSMS) :

c'est un environnement graphique développé par Microsoft pour gérer les bases de données SQL Server. Il permet de créer, administrer et interroger des bases de données de manière efficace. Dans ce projet, SSMS a été utilisé pour concevoir et exécuter les scripts SQL nécessaires à la création du Data Warehouse, garantissant ainsi la structuration et l'organisation des données pour l'analyse décisionnelle.

3.3-Microsoft Access :

c'est un système de gestion de base de données relationnelle destiné à la création et à la manipulation de bases de données locales. Dans ce projet, Access a été utilisé pour extraire les données de la base Northwind, afin de les intégrer ensuite dans le processus ETL et les charger dans le Data Warehouse pour l'analyse décisionnelle.

3.4-Language Python:

c'est un langage de programmation polyvalent, simple à apprendre et largement utilisé pour l'analyse de données et le développement d'applications. Dans ce projet, plusieurs bibliothèques Python ont été choisies pour leurs fonctionnalités spécifiques :

-Pandas : pour la manipulation, le nettoyage et l'analyse des données sous forme de tableaux (DataFrames).
-Streamlit : pour créer facilement des dashboards interactifs et visualiser les KPI.
-Jupyter Notebook : pour développer et tester le code de manière interactive, étape par étape.
-PyODBC : pour se connecter et interagir avec des bases de données SQL Server depuis Python.
-OpenPyXL : pour lire et écrire des fichiers Excel, notamment pour extraire les données de sources Access converties en Excel.
-SQLAlchemy : pour gérer les connexions aux bases de données et exécuter des requêtes

SQL de manière plus flexible et sécurisée.

Ces outils ont été sélectionnés pour faciliter le processus ETL, le calcul des KPI et la réalisation du dashboard, tout en assurant efficacité et compatibilité avec les différentes sources de données.

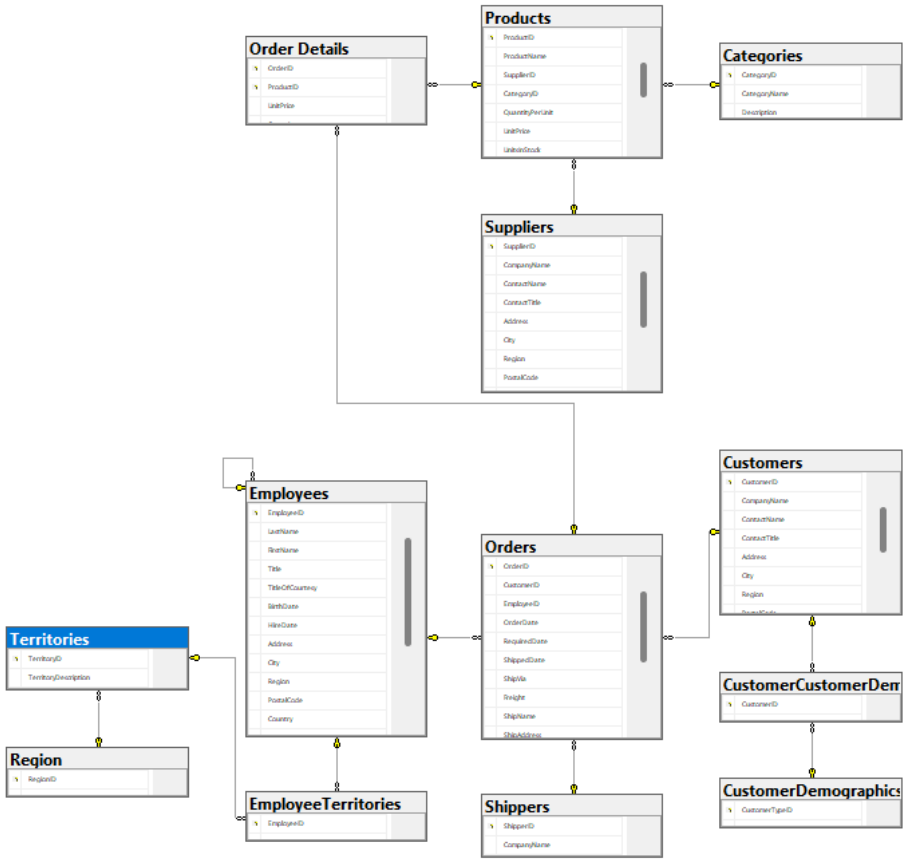
3.5-Le langage SQL:

a été utilisé pour concevoir et construire le Data Warehouse du projet. Les différentes tables ont été organisées selon le schéma en étoile, avec des tables de dimensions (DIM) pour les informations descriptives et une table de faits pour les données transactionnelles. Cette structure facilite l'analyse des données et le calcul des indicateurs clés de performance (KPI).

4-Réalisation de la solution Business Intelligence:

4.1- Exploration de la base de données:

L'exploration des bases de données sur SQL serveur et Microsoft Access de Northwind a constitué la première étape du projet. Elle a permis de comprendre la structure des tables, la nature des données et les relations existantes entre elles. Les tables principales identifiées étaient Customers, Employees et Orders. Cette étape est cruciale pour évaluer la qualité des données, identifier les doublons ou valeurs manquantes, et préparer les transformations nécessaires. Les outils utilisés incluent pyodbc pour se connecter à SQL Server et récupérer les métadonnées, ainsi que pandas pour examiner le contenu des tables et générer des analyses préliminaires.



4.2-Détermination des KPI:

La définition des KPI (Key Performance Indicators) a consisté à sélectionner les indicateurs essentiels pour suivre la performance commerciale. Les mesures retenues comprenaient le nombre de commandes livrées et non livrées par: client, employé et date. L'objectif était de fournir des indicateurs pertinents pour la prise de décision. Cette étape a été réalisée à l'aide de pandas et d'Excel, permettant de calculer des valeurs agrégées et d'identifier les tendances avant la construction du Data Warehouse.

4.3-Détermination des dimensions:

Les dimensions représentent les entités descriptives permettant d'analyser les faits. Les dimensions identifiées étaient les clients, les employés et le temps. Chaque dimension a été nettoyée et normalisée pour garantir l'unicité des identifiants et la cohérence des données entre différentes sources. Par exemple, les prénoms et noms des clients ou employés ont été combinés en un champ fullname pour faciliter le mapping. Cette étape a été réalisée principalement avec Python et pandas, en appliquant des transformations et des fusions entre les données Access et SQL Server.

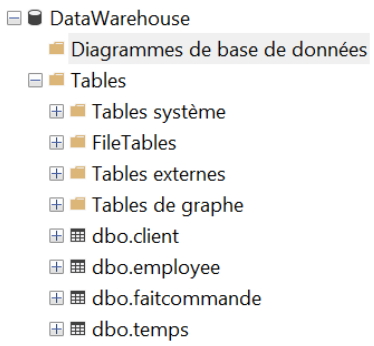
4.4-Détermination de la table des faits:

La table des faits centralise les mesures quantitatives pour l'analyse. Elle a été construite à partir des commandes (Orders) et contient des informations clés telles que order_id, customer_id, employee_id, order_status, ship_address et ship_country. L'objectif était de regrouper toutes les transactions commerciales dans une seule table, afin de faciliter l'agrégation et le calcul des KPI. Cette consolidation a été réalisée après avoir harmonisé les données des différentes sources et résolu les doublons ou incohérences.

4.5-Construction du schéma en étoile:

Le schéma en étoile est une architecture classique pour les Data Warehouses. Il consiste à placer la table des faits au centre, entourée des tables de dimensions qui décrivent les entités. Cette structure simplifie l'écriture de requêtes analytiques et améliore la performance des analyses OLAP. Les relations entre la table des faits et les dimensions ont été définies via des clés primaires et étrangères dans SQL Server, assurant la cohérence et l'intégrité référentielle des données.

4.6-Réalisation du Data Warehouse:



Le Data Warehouse a été créé pour centraliser les données transformées et servir de source unique pour l'analyse. Les tables de dimensions et la table des faits ont été implémentées dans SQL Server, en respectant les contraintes d'intégrité et en normalisant les colonnes pour uniformiser les formats. Cette étape garantit que toutes les données sont cohérentes, fiables et prêtes à être exploitées pour le calcul des KPI et la visualisation.

4.7-Réalisation de l'ETL:

L'ETL (Extract, Transform, Load) a été réalisé en trois phases distinctes:

1- Extraction des données :

La première étape consiste à extraire les données de la base SQL Server. Pour cela, la bibliothèque pyodbc a été utilisée pour établir une connexion sécurisée via Trusted_Connection. Cette méthode permet d'éviter de stocker les identifiants dans le code tout en garantissant un accès direct aux tables SQL. Une fois la connexion établie, toutes les tables de type "BASE TABLE" ont été récupérées via une requête SQL sur INFORMATION_SCHEMA.TABLES. Chaque table a ensuite été exportée vers un fichier Excel grâce à pandas. Cette étape permet de travailler sur les données hors ligne et de les fusionner facilement avec les données Access. Le choix d'Excel est justifié par sa compatibilité universelle avec les systèmes Access et SQL, ainsi que par sa facilité de manipulation pour le nettoyage et la vérification des données.

2-Transformation des données:

Les tables clients et employés provenant d'Access et de SQL Server ont été normalisées pour harmoniser les noms de colonnes et faciliter la fusion. Pour les clients, firstname et lastname ont été combinés en fullname, tandis que pour les employés, de nouveaux employeeid séquentiels ont été générés pour éviter les conflits. Les doublons ont été supprimés et les fichiers finaux enregistrés sous dbo.client.xlsx et dbo.employee.xlsx.

La table de faits Orders a été transformée en normalisant les colonnes Les deux sources ont été fusionnées, les doublons supprimés, et le résultat final a été enregistré sous dbo.faitcommande2.xlsx.

3-Chargement des données:

Les données transformées ont été chargées dans le Data Warehouse SQL Server à l'aide de pandas.to_sql et SQLAlchemy. Les clés primaires et étrangères ont été correctement mappées

pour maintenir l'intégrité référentielle. Cette étape a permis de rendre les données immédiatement exploitables pour l'analyse.

4.8-Calcul des KPI:

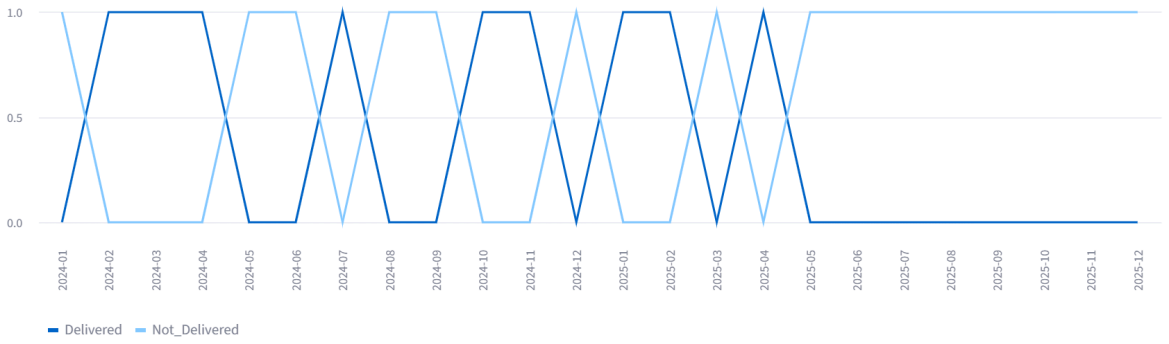
Après le chargement des données dans le Data Warehouse, les KPI ont été calculés à partir de la table des faits et des dimensions. Les indicateurs incluent nombre de commandes livrées et non livrées par: client, employé et date. Ces calculs ont été réalisés avec pandas et SQL, permettant d'agréger les données, d'obtenir des métriques précises et de préparer les éléments pour la visualisation

4.9-Construction du Dashboard et visualisation:

Pour rendre les résultats accessibles et interactifs, un tableau de bord a été construit avec Streamlit. Il permet de visualiser les KPI sous forme de graphiques et de tableaux, de filtrer les données par périodes ou catégories, et d'explorer les tendances commerciales. Cette interface interactive facilite la prise de décision et offre une vue complète sur les performances des clients et employés, ainsi que sur l'activité commerciale globale.

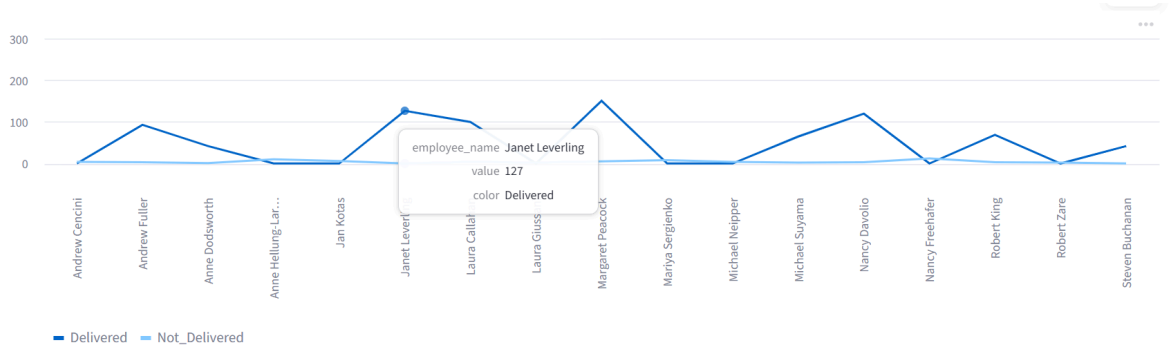
Voici quelques résultats issus de notre tableau de bord interactif:Le graphique en courbes (line chart) illustre l'évolution du nombre de commandes livrées et non livrées selon le temps (par date), ainsi que leur répartition par employé.De son côté, le graphique en barres (bar chart) présente la répartition des commandes livrées et non livrées par client, et les tables représente les données en détails.

L'ensemble du tableau de bord est dynamique et permet d'affiner l'analyse grâce à des filtres par employé, intervalle de dates ou client, offrant ainsi une meilleure compréhension de la performance commerciale



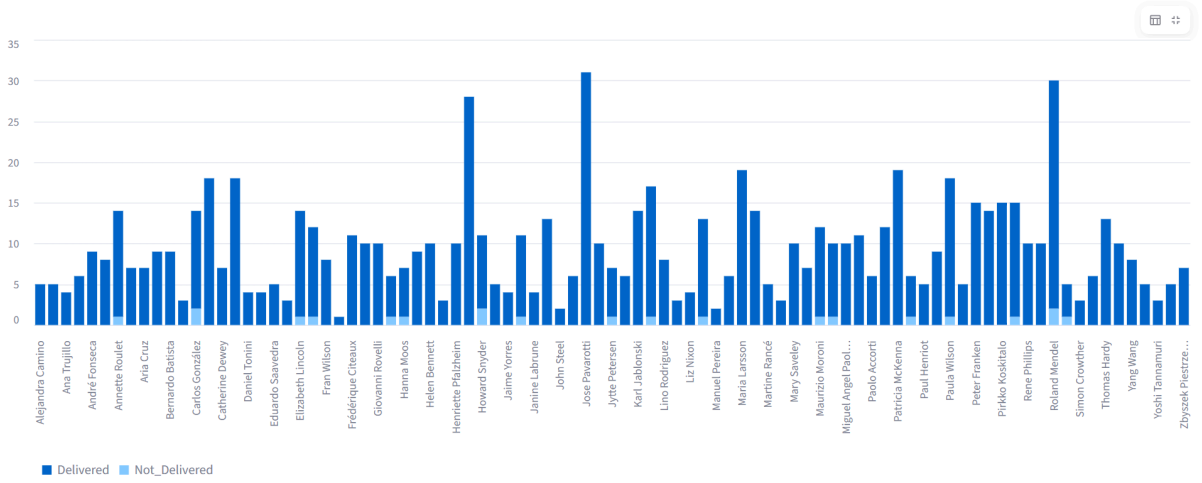
KPI by Month

	year_month	Delivered	Not_Delivered
0	2024-01	0	1
1	2024-02	1	0
2	2024-03	1	0
3	2024-04	1	0
4	2024-05	0	1
5	2024-06	0	1
6	2024-07	1	0
7	2024-08	0	1
8	2024-09	0	1
9	2024-10	1	0



KPI by Employee

	employee_name	Delivered	Not_Delivered
8	Margaret Peacock	151	5
5	Janet Leverling	127	0
12	Nancy Davolio	120	3
6	Laura Callahan	100	4
1	Andrew Fuller	93	3
14	Robert King	69	3
11	Michael Suyama	65	2
16	Steven Buchanan	42	0
2	Anne Dodsworth	42	1
7	Laura Giussani	0	2



KPI by Client

	client_name	Delivered	Not_Delivered
42	Jose Pavarotti	31	0
33	Horst Kloss	28	0
78	Roland Mendel	28	2
54	Maria Larsson	19	0
66	Patricia McKenna	19	0
13	Carlos Hernández	18	0
15	Christina Berglund	18	0
70	Paula Wilson	17	1
47	Laurence Lebihan	16	1
72	Peter Franken	15	0

5-Conclusion:

Le projet a permis de mettre en place un processus complet de Business Intelligence, allant de l'exploration des données à la visualisation des indicateurs clés de performance. Grâce à l'extraction et à la normalisation des données provenant de sources hétérogènes (SQL Server et Access), nous avons pu construire des tables de dimensions cohérentes et une table de faits centralisée, prêtes à être exploitées dans un Data Warehouse.

La mise en place d'un schéma en étoile a facilité l'organisation des données et le calcul des KPI, offrant une vue structurée et intuitive des informations commerciales. Le processus ETL développé avec Python (pandas, SQLAlchemy) a permis d'automatiser la transformation et le chargement des données tout en garantissant leur intégrité et leur homogénéité.

Enfin, la réalisation d'un tableau de bord interactif avec Streamlit a permis de visualiser et d'analyser les KPI en temps réel, offrant un outil pratique pour la prise de décision et le suivi de la performance commerciale. Ce projet constitue ainsi une solution BI complète, fiable et évolutive, pouvant être enrichie à l'avenir avec de nouveaux indicateurs, d'autres sources de données ou des visualisations plus avancées.