Rapport projet Bi

Djahid Aoudia
212131096417
M2 Big Data

Introduction :

Dans un contexte où les entreprises doivent exploiter efficacement leurs données pour améliorer leur prise de décision, la Business Intelligence (BI) occupe une place centrale dans les systèmes d'information modernes. Elle permet de collecter, transformer et analyser les données afin de fournir aux décideurs des informations fiables et pertinentes.

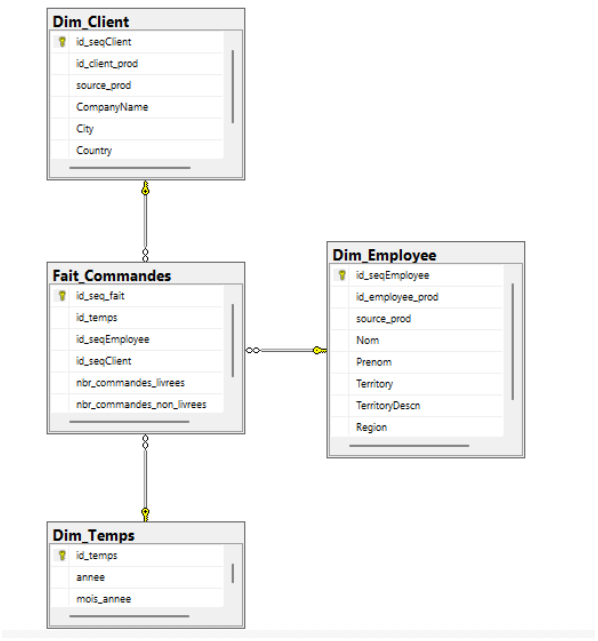
Dans le cadre de ce projet, nous avons mis en œuvre un processus complet de Business Intelligence basé sur la célèbre base de données **Northwind**. Celle-ci contient un ensemble de données représentatives du fonctionnement d'une entreprise commerciale : clients, produits, commandes, livraisons et employés. Les données Northwind ont été extraites à partir de **différentes sources**, notamment **Microsoft Access** et **SQL Server**, afin d'illustrer la diversité des environnements possibles dans un projet décisionnel réel.

Pour réaliser ce travail, deux outils complémentaires ont été utilisés :

- **Talend Open Studio**, un outil d'ETL (Extract, Transform, Load) open source, utilisé pour l'extraction des données sources, leur transformation selon le modèle décisionnel, puis leur chargement dans une base dédiée. Talend permet d'automatiser et d'industrialiser l'intégration des données tout en assurant leur qualité, leur cohérence et leur traçabilité.
- **Microsoft Power BI**, un outil de visualisation et d'analyse dynamique, intégrant également un moteur ETL via **Power Query**. Ce dernier permet de préparer, nettoyer et transformer les données directement dans Power BI avant l'étape de modélisation. Power BI est ensuite utilisé pour concevoir des tableaux de bord interactifs, créer des indicateurs de performance (KPI) et offrir une exploration intuitive des données, facilitant ainsi l'identification de tendances et la prise de décision.

Architecture Dimensionnel :

Le modèle décisionnel a été conçu sous la **forme d'un schéma en étoile**, qui est particulièrement adapté aux analyses de type **statistiques et reporting**. L'objectif principal est de déterminer le nombre de **commandes livrées ou non livrées** afin d'améliorer le suivi des performances commerciales et logistiques.



Dimension	Description / Rôle
Dim_Temps	Contient les informations temporelles (annee, mois_annee). Permet d'analyser les commandes selon la période.
Dim_Client	Contient les informations sur les clients (CompanyName, City, Country …). Permet d'analyser les commandes par client, ville ou pays.
Dim_Employee	Contient les informations sur les employés responsables (Nom, Prenom, Territory, Region). Permet d'analyser les performances des employés sur les commandes.

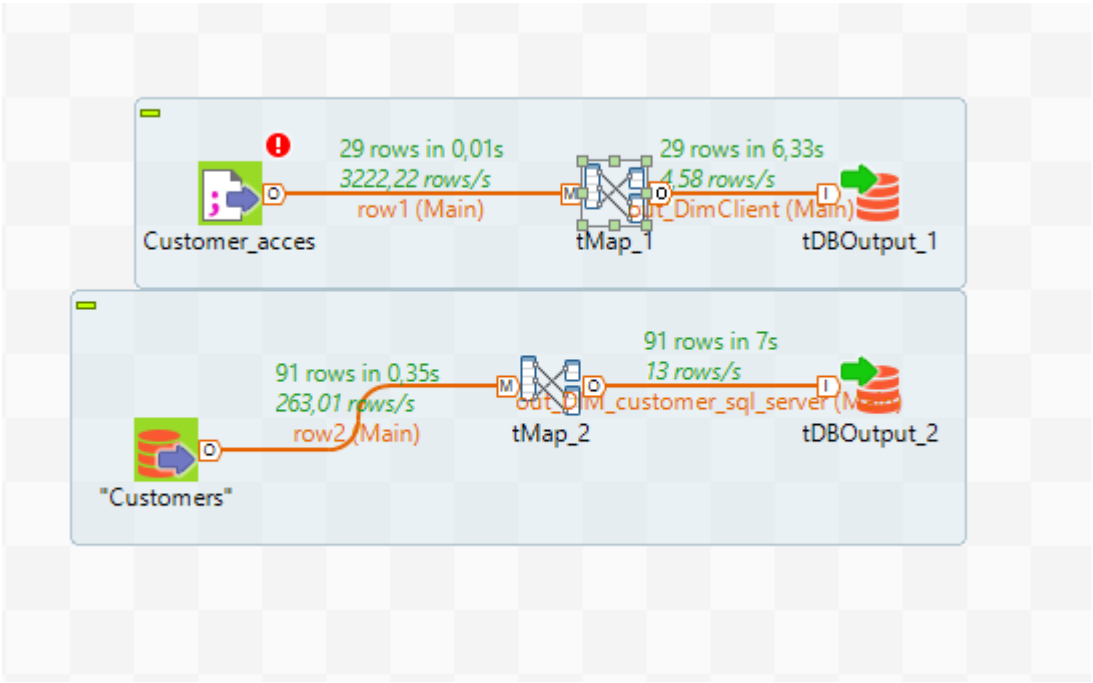
Présentation générale de la démarche Etl :

Réalisation de l'ETL avec Talend :

Dans le cadre de la construction du Data Warehouse *Northwind_DW*, un ensemble de jobs Talend a été développé afin d'alimenter les différentes dimensions du modèle en étoile. Chaque job est dédié au traitement d'une dimension spécifique, garantissant ainsi une meilleure lisibilité, une maintenance facilitée et une séparation cohérente des flux de données provenant de plusieurs sources (SQL Server et Microsoft Access):

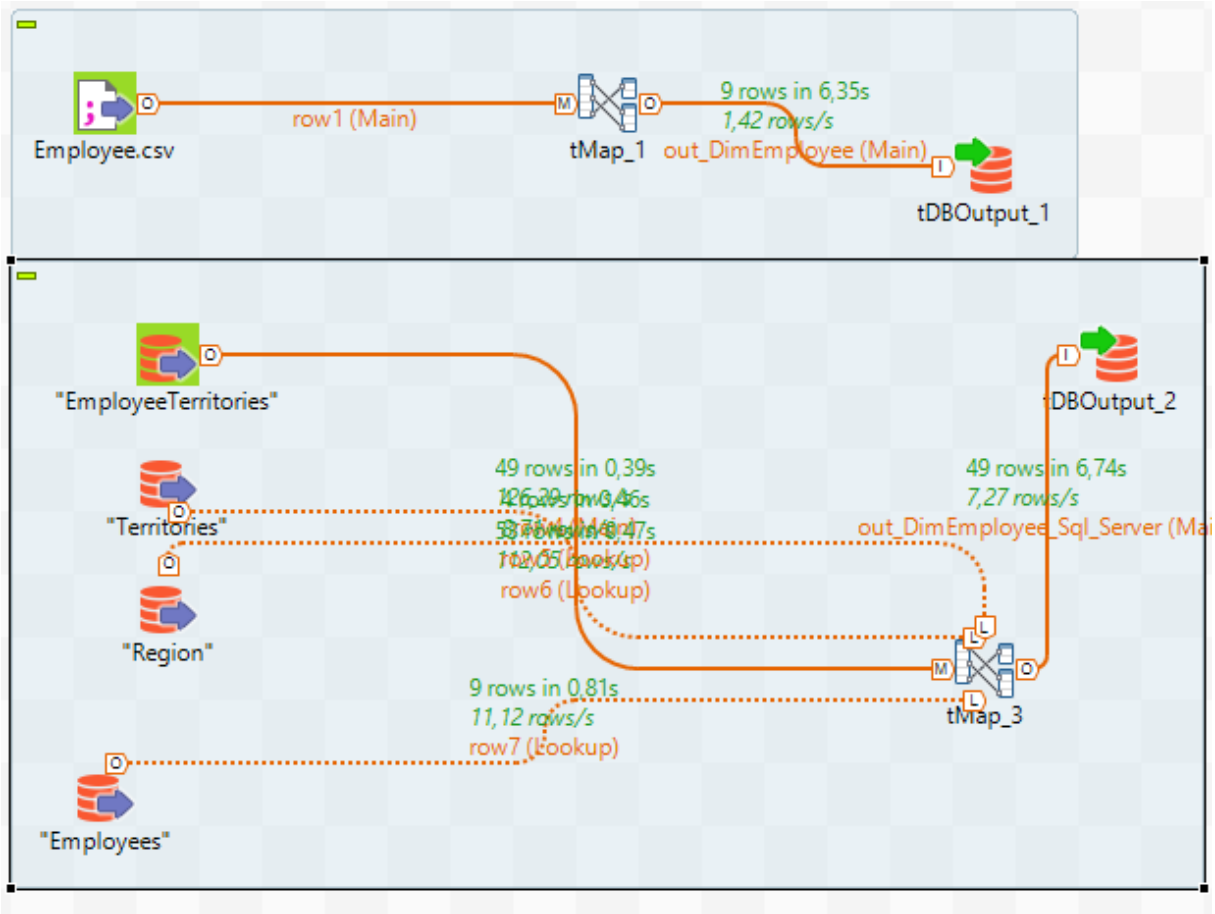
Remarque : j'ai séparé le traitement des données de sql server et acces pour une meilleure clarification :

- ***Job Dim_Client :***



Le traitement consiste essentiellement à mapper les champs correspondants dans le composant tMap, sans nécessiter de jointures supplémentaires et d'ajouté simplement le champ source pour identifier les informations qui vienne de SQL server ou Acess

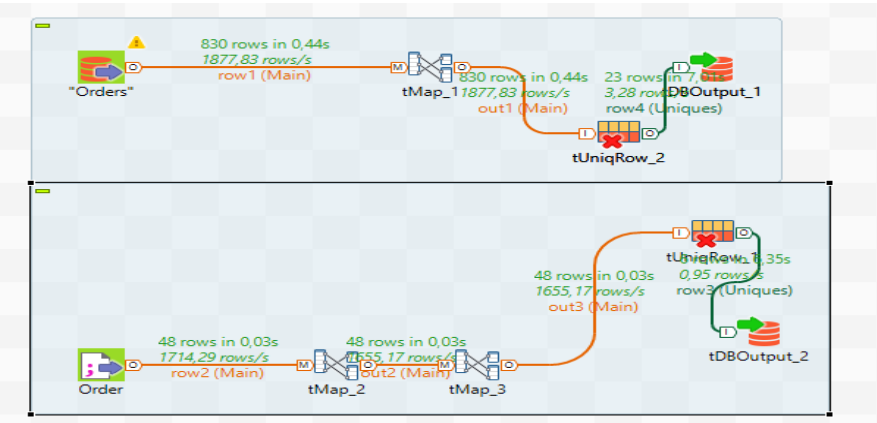
- ***Job Dim_Employee :***



Pour les données provenant de **SQL Server**, une jointure entre les tables *EmployeeTerritories*, *Territories*, *Region* et *Employees* a été réalisée afin de reconstituer l'ensemble des informations nécessaires, notamment le territoire et la région d'affectation de chaque employé.

En revanche, dans la source **Access**, certaines informations telles que le territoire et la région ne sont pas disponibles. Afin de garantir l'homogénéité du schéma dimensionnel, ces attributs ont été renseignés avec des valeurs *NULL*, traduisant explicitement l'absence de données dans la source d'origine.

- ***Job Dim_Temp :***



La construction de la dimension Temps repose sur l'extraction de l'attribut OrderDate disponible dans la table Orders.
À partir de cette date, deux transformations principales sont appliquées :

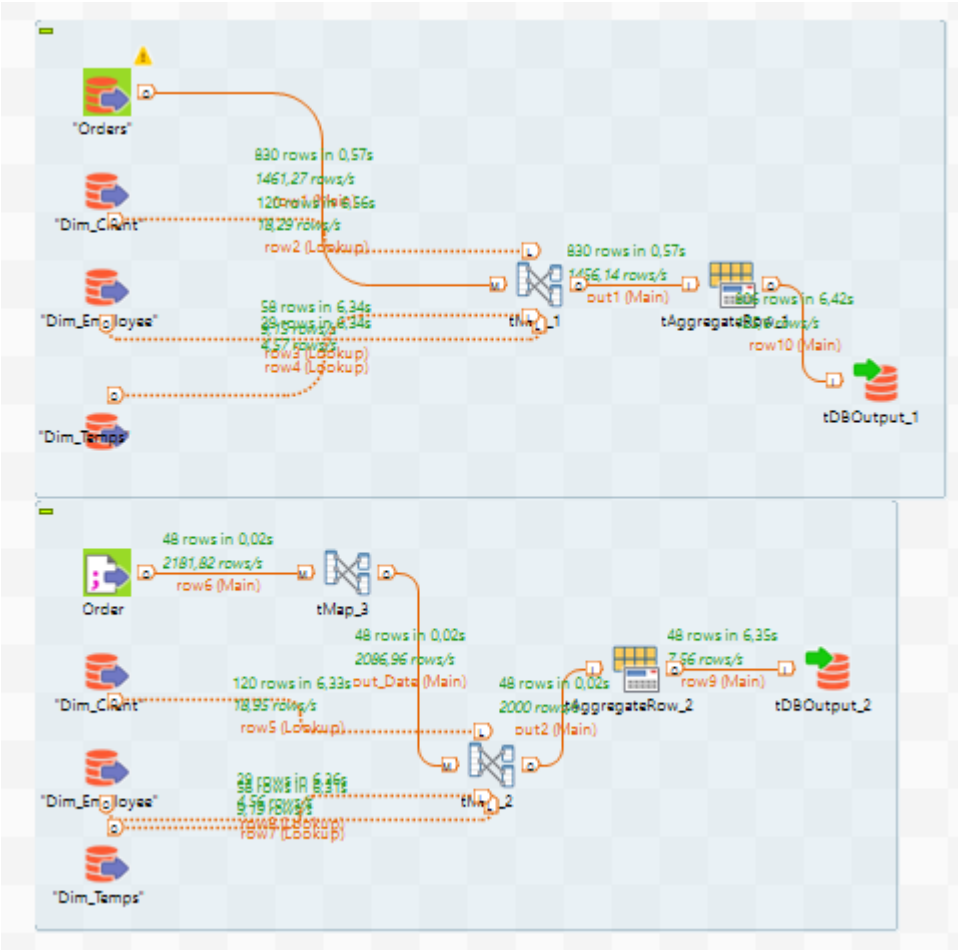
- Extraction de l'année grâce à l'expression :
  Integer.valueOf(new java.text.SimpleDateFormat("yyyy").format(row1.OrderDate))
- Génération du couple mois–année au moyen de :
  new SimpleDateFormat("MMMM yyyy").format(row1.OrderDate)

Dans le cas des données issues de Microsoft Access, deux composants tMap ont été utilisés successivement :

1. Conversion du champ OrderDate en un format Date compatible avec Talend.
2. Extraction de l'année et du mois–année à partir de cette date convertie.

Enfin, le composant tUniqueRow est appliqué afin d'éliminer les doublons et de structurer la dimension sous forme **d'un calendrier** unique contenant chaque combinaison année et mois–année.

- ***Job la table Fait*** :



La construction de la table de faits repose sur l'intégration des informations issues de la table *Orders*, enrichies par les clés substituts provenant des différentes dimensions.

Pour cela, une **jointure** est effectuée entre la table *Orders* et les dimensions **Temps**, **Client** et **Employee**, permettant de récupérer les identifiants dimensionnels nécessaires.

Une logique conditionnelle est ensuite appliquée pour déterminer le statut de livraison de chaque commande :

- si *ShippedDate* est **NULL**, la commande est considérée comme **non livrée** (commande non livrée valeur 1 sinon 0),
- sinon, elle est considérée comme **livrée** (commande livrée valeur 1 sinon 0).

Ces indicateurs (commande livrée / non livrée) sont ensuite agrégés à l'aide du composant **tAggregateRow**, qui calcule la somme totale de commandes livrées et non livrées, groupées par les clés des dimensions (Temps, Client, Employé).
Le résultat final constitue ainsi les mesures de la table de faits **Fait_Commandes**, structurées selon le modèle en étoile conçu.

Réalisation de l'ETL avec Power bi :

Pour mettre en œuvre un processus ETL complémentaire, reposant principalement sur **Power Query**, le moteur d'extraction, de transformation et de chargement intégré à l'outil Power Bi.
L'objectif était d'assurer une préparation efficace des données issues de Northwind, qu'elles proviennent de **SQL Server** et de **Microsoft Access**, avant leur exploitation dans un tableau de bord décisionnel.

Construction de la Dimension Client

La dimension Client a été préparée en appliquant les opérations suivantes :

- Sélection des colonnes pertinentes (identifiant, nom de l'entreprise, ville, pays, etc.).
- Uniformisation des formats des champs provenant de SQL Server et Access.
- Ajout d'un champ indiquant la source d'origine des données (SQL Server / Access).

Cette étape aboutit à une table **Dim_Client** prête à être intégrée dans le modèle en étoile.

Construction de la Dimension Employee

Pour cette dimension, les données proviennent de plusieurs tables.

Données SQL Server

Une jointure a été effectuée dans Power Query entre :

- *Employees*,
- *EmployeeTerritories*,
- *Territories*,
- *Region*.

Cette combinaison permet d'enrichir chaque employé avec les informations relatives à son territoire et à sa région de travail.

- Sélection des colonnes correspondante

Données Access

- Sélection des colonnes correspondante

Certaines informations étant absentes dans la version Access, notamment *territory* et *region*, des valeurs **null** ont été attribuées pour représenter ce manque.

Les deux sources ont ensuite été combinées et nettoyées pour produire la table **Dim_Employee**.

Construction de la Dimension Temps

La dimension temps est générée à partir du champ *OrderDate* présent dans la table Orders.

Les transformations effectuées dans Power Query sont :

- Extraction de l'année via la fonction *Year()*,
- Création d'un champ *Mois_Année* (format : « Mois Année ») à l'aide des fonctions de formatage de dates,
- Suppression des doublons afin de constituer une dimension temps propre et complète qui aura la forme **d'un calendrier**.

Construction de la Table de Faits :

La table de faits a été élaborée en combinant *Orders* avec les dimensions préparées.

Les étapes réalisées :

- Jointure des données *Orders* avec les dimensions Temps, Client et Employee afin de récupérer les clés substituts.
- Création d'un indicateur de livraison :
  - si *ShippedDate* est null → commande non livrée (=1),
  - sinon → commande livrée (=1).
- Agrégation des données à l'aide du groupe de lignes (Group By) de Power Query :
  - nombre total de commandes livrées,
  - nombre total de commandes non livrées,
    regroupés par les identifiants dimensionnels.

Le résultat final constitue la table Fait_Commandes, prête à être utilisée dans les visualisations Power BI.

Présentation et Interprétation du Tableau de Bord BI :

Pour que power bi permet la réalisation de le tableau de bord en dois crée la vue model d'abord :

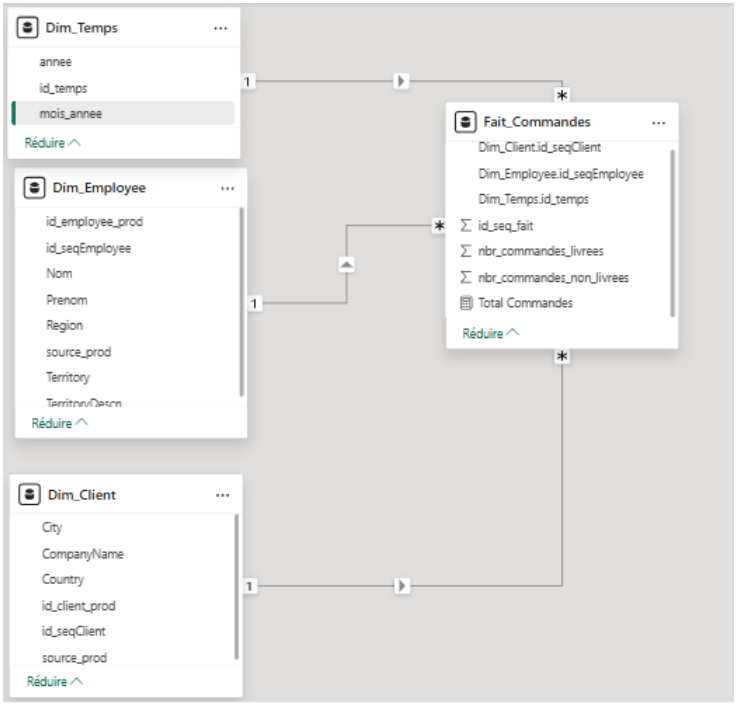
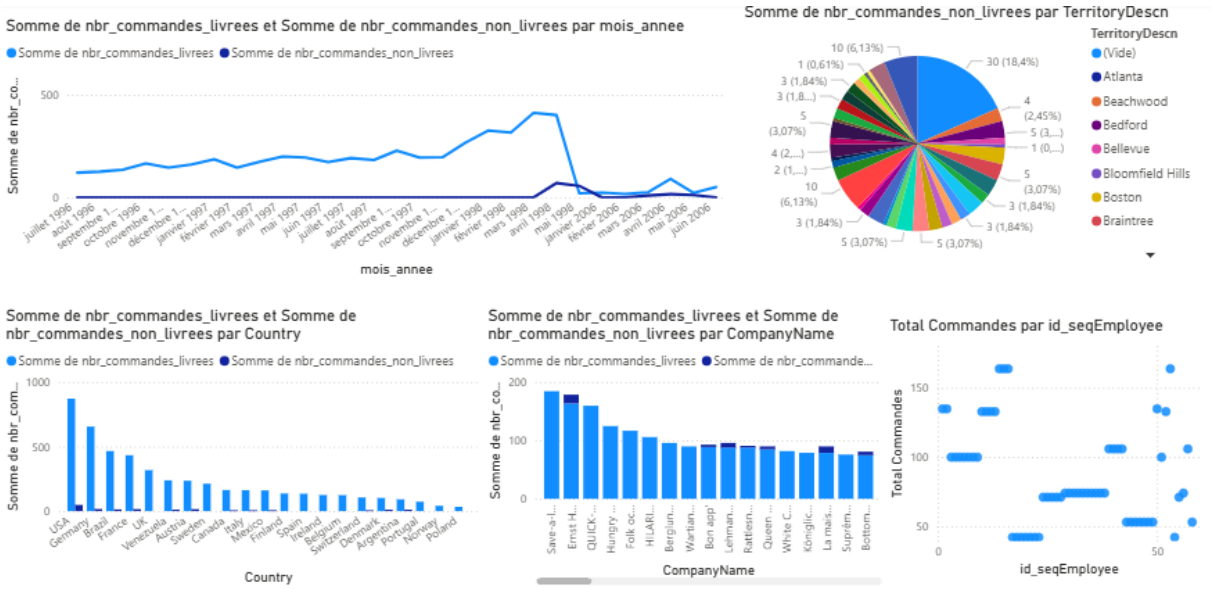
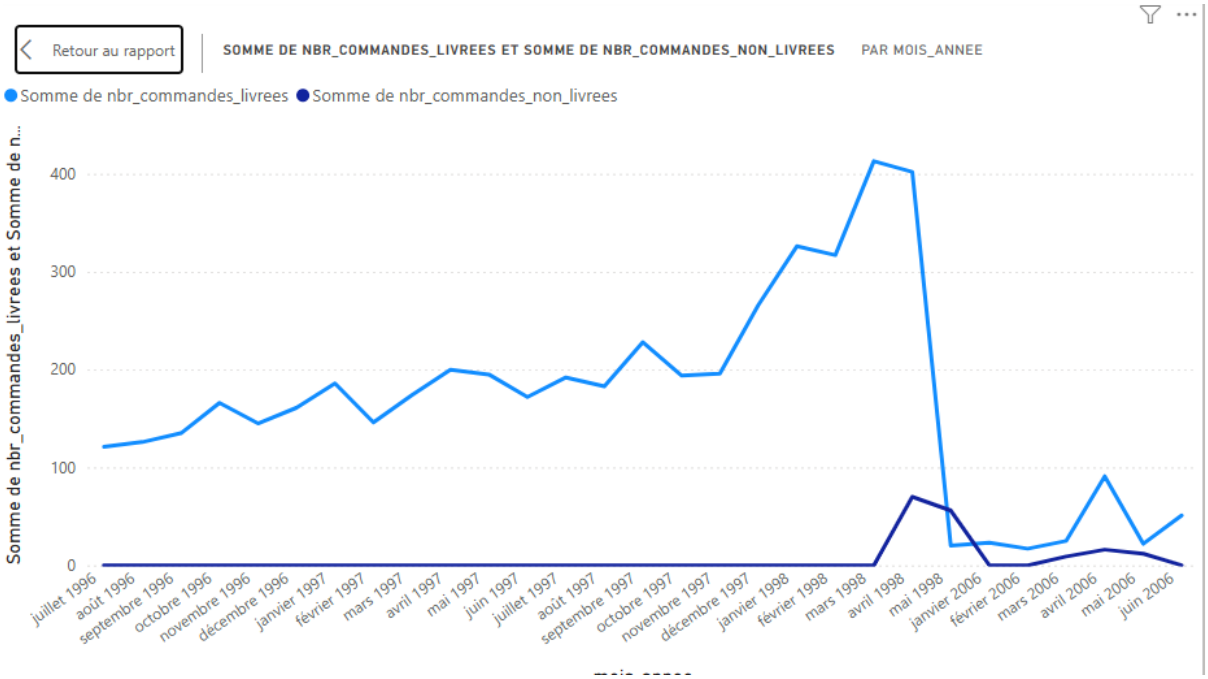


Tableau de Bord :



Interprétation :

Performance des Commandes au Fil du Temps (Graphique en Ligne) :



Commandes Livrées (Bleu clair) : Elles suivent une tendance croissante et relativement stable de juillet 1996 jusqu'au pic de début 1998, signalant une bonne exécution opérationnelle durant cette phase.

***Commandes Non Livrées (Bleu foncé)* :** Le volume est généralement faible jusqu'au pic. L'augmentation spectaculaire des commandes non livrées en **avril 1998** est la **cause principale du pic total**, indiquant un **problème opérationnel majeur** ou un **goulot d'étranglement** critique sur cette période. La proportion de commandes non livrées est particulièrement élevée à ce moment, après la forte baisse, les volumes semblent se stabiliser à un niveau beaucoup plus bas, avec une proportion de commandes non livrées plus faible mais jusque a un autre pic **avril 2006**.

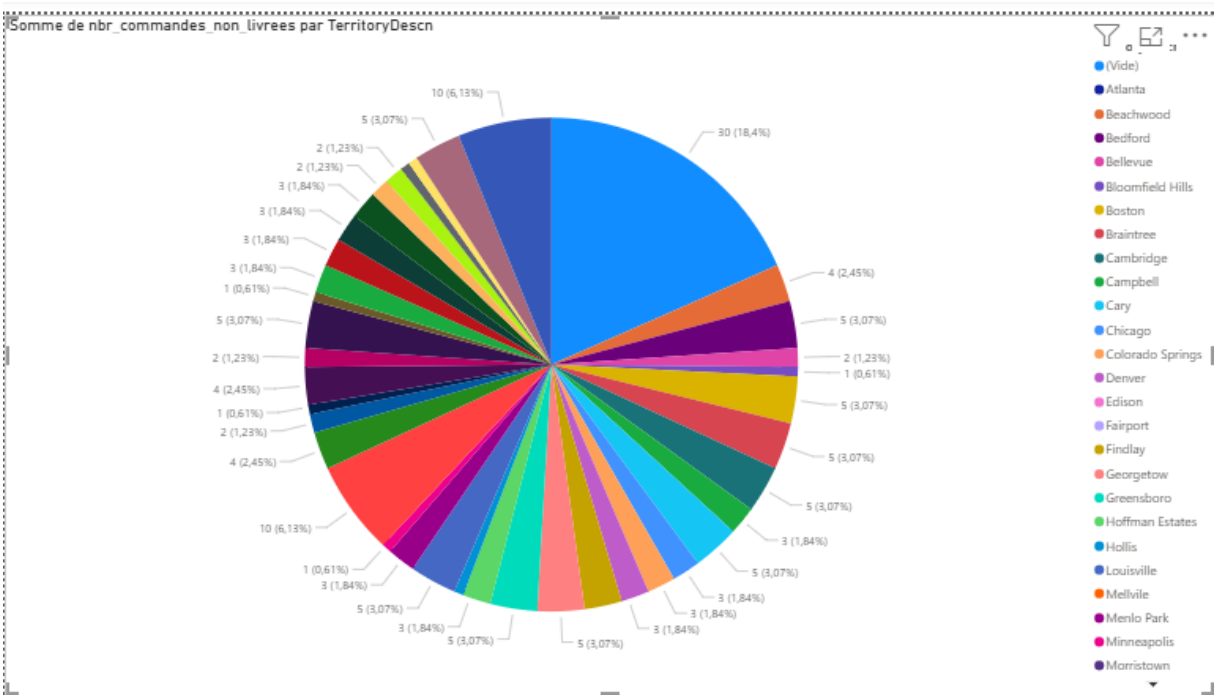
Observation des tendances mensuelles

Les mois d'avril et de mai se distinguent par un **nombre particulièrement élevé de commandes**. On note également une chute notable des commandes livrées après avril 1998 : les commandes sont passées de 402 livrées à seulement 20 (mai 1998), probablement en raison du **pic de commandes non livrées** observé à cette même période.

Recommandations

- Investiguer les causes des pics observés en **avril 1998 et avril 2006** (campagne marketing, événement exceptionnel, contraintes logistiques…).
- Se préparer à une **augmentation du volume de commandes** pendant les mois d'avril et mai afin d'optimiser les ressources et la capacité opérationnelle.

Distribution des Commandes Non Livrées par Territoire (Graphique Circulaire) :



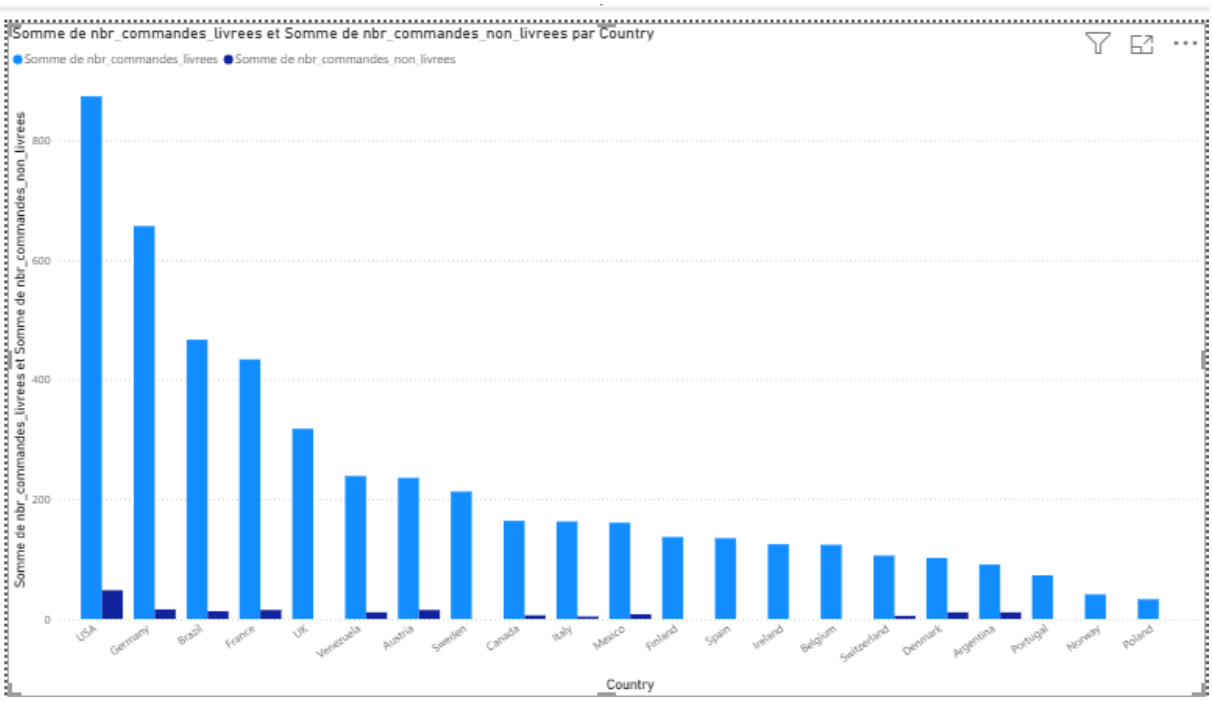
Territoires à Risque : **Wilton** et **Neward (6.13%)** sont les deux territoires nommés qui contribuent le plus aux commandes non livrées, nécessitant une analyse approfondie de leurs processus locaux.

Faible Concentration : Les non-livraisons sont très **dispersées** sur de nombreux territoires, chacun contribuant à un faible pourcentage (la plupart sont autour de 1.84% à 3.07%).

Recommandations

- Prioriser la correction des données pour les territoires (Vide)
- Mener une enquête sur les processus logistiques et/ou de gestion des stocks à **: Wilton et Neward**
- Développer les territoires sous-performant
- Équilibrer la charge géographique pour réduire les risques

Performance des Commandes par Pays (Graphique à Barres) :

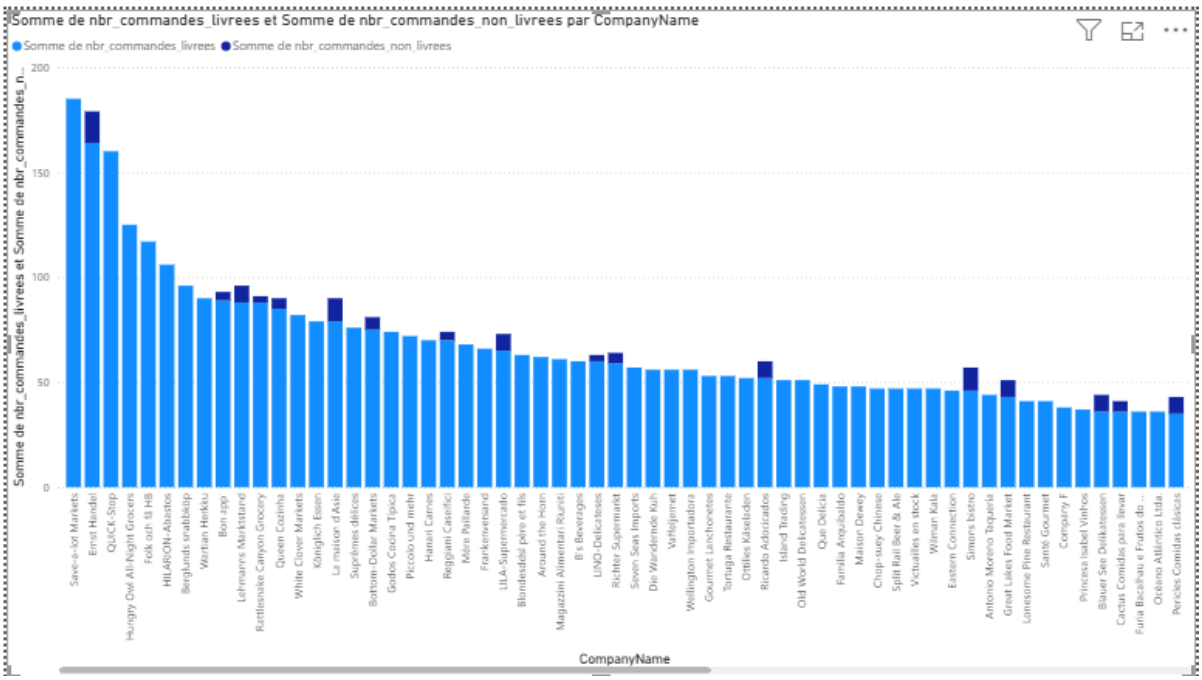


1. **USA** : Leader incontesté (~874 commandes livrées, ~48 non livrées)
2. **Allemagne** : 2ème position (~657 commandes livrées)
3. **France, Brésil** : Positions intermédiaires
4. **UK** : future client important
5. **Longue traîne** : 15+ pays avec faible volume

Recommandations

- **Anomalie États-Unis :** Le volume élevé de commandes non livrées est préoccupant et pourrait indiquer des **problèmes logistiques ou de gestion** à investiguer.
- **Opportunités de croissance :** Les marchés émergents tels que le **Mexique, la Belgique, la Suisse et l'Espagne** présentent un potentiel de développement important et méritent une attention particulière.
- **Concentration du chiffre d'affaires :** Les **quatre principaux pays** (USA, Allemagne, France, Brésil) représentent environ **60 % du volume total des commandes**, suggérant la nécessité d'optimiser la gestion de ces marchés tout en diversifiant l'activité vers les pays à faible volume.

Performance des Commandes par Entreprise (Graphique à Barres)



- ***Top performeur :*** Le client **"Save-a-Lot"** se distingue avec environ **185 commandes livrées**, représentant un volume significatif de l'activité commerciale.
- ***Concentration des non-livraisons* :** L'analyse montre que certains clients présentent un **nombre élevé de commandes non livrées**, notamment :
    - **Ernst Handel** (Australie) : principal client australien avec le plus grand nombre de commandes non livrées.
    - **La Maison d'Aise** (France) : deuxième plus grand client français qui a le plus nombre commandes non livrées.
    - **Simon Bistro** (Danemark) : premier client danois, a nombre très élevé de commandes non livrées.

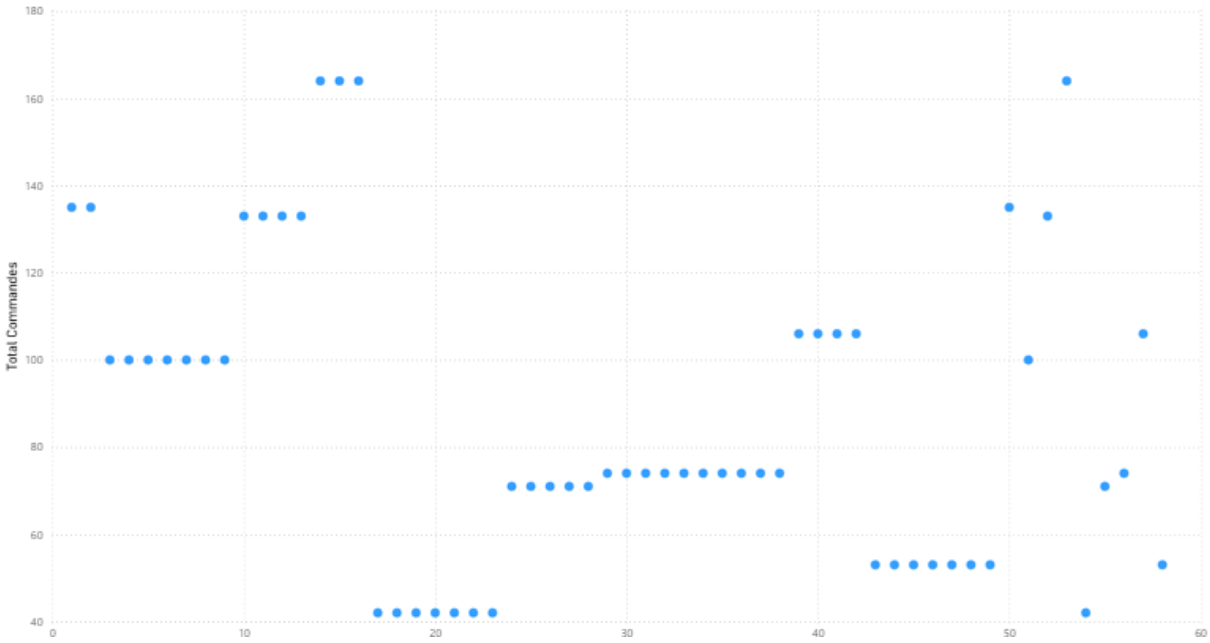
La hauteur de la barre correspondant aux commandes non livrées (bleu foncé) indique un problème récurrent pour ces clients stratégiques.

Recommandations

- **Suivi proactif des commandes à risque** Implémenter des indicateurs de suivi en temps réel pour détecter rapidement les commandes susceptibles de ne pas être livrées et intervenir avant l'échec de livraison.
- **Communication client renforcée** Informer les clients stratégiques de manière proactive en cas de retard ou d'incident afin de maintenir leur confiance et leur satisfaction.
- **Optimisation logistique par région** Analyser les zones géographiques où les non-livraisons sont fréquentes et ajuster les ressources logistiques pour améliorer la fiabilité des livraisons.

- **Segmentation des clients** Prioriser les ressources pour les **tops clients**, tout en maintenant un suivi régulier pour les clients à volume intermédiaire afin d'éviter la perte de potentiel commercial.

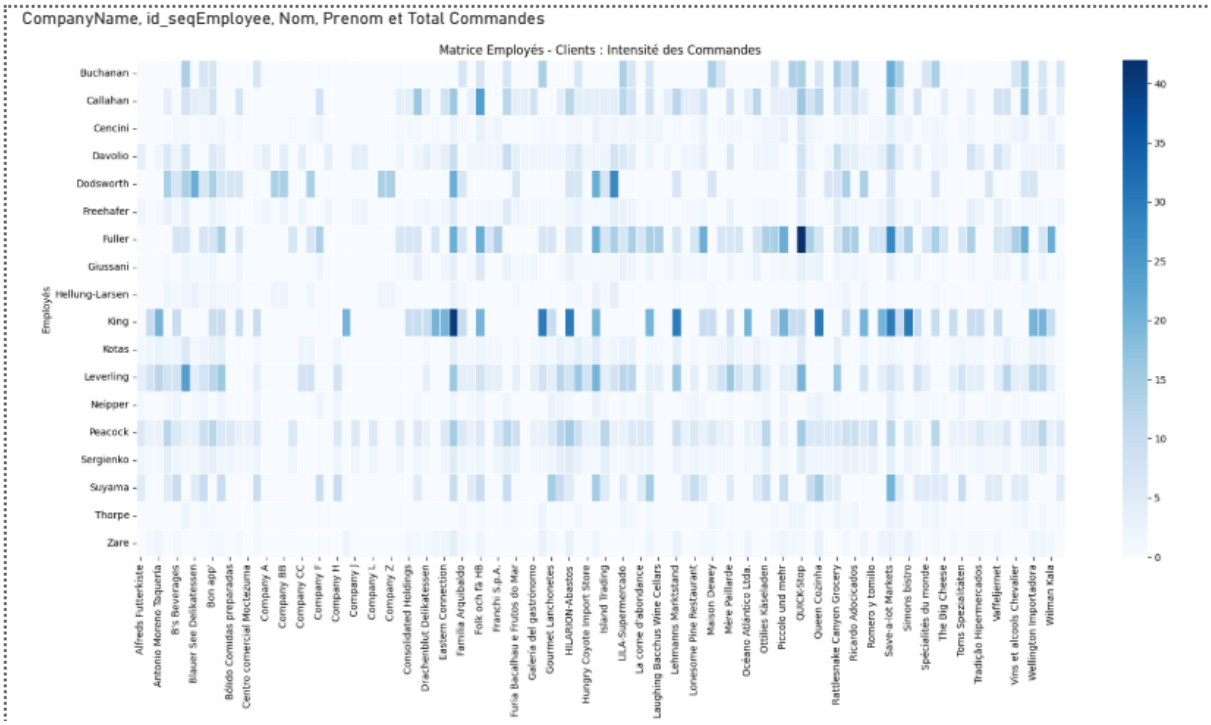
Distribution des Commandes par Employé (Nuage de Points - Bas Droite)



- ***Clusters de performance*** *:*
    - Groupe à ~100 commandes : Employés standards
    - Groupe à ~45-50 commandes : Employés en développement
    - Top performers : 4 employés à 150+ commandes
    - Quelques outliers à 60+ employés avec très faible volume

Analyse de la Matrice Employés : Clients - Intensité des Commandes (Hash map)

Ce graphe a été générer par le code python :



Le graphique est une carte thermique où l'axe Y représente les Employés (Employees) et l'axe X représente les Clients (Clients/CompanyName). L'intensité de la couleur bleue (voir l'échelle à droite) indique le nombre de commandes passées par un client spécifique et traitées par un employé spécifique.

Intensité des Commandes par Combinaison Employé-Client

L'analyse se concentre sur les zones de couleur bleu foncé, qui représentent les volumes de commandes les plus élevés (jusqu'à 40 commandes ou plus).

Pôles de Forte Intensité (Bleu Foncé) :

- Employé King : Présente une très forte concentration de commandes avec le client 'Ernstein Connection' (le bloc le plus foncé et potentiellement le plus haut en volume). Il a également des relations fortes avec 'Consolidated Holdings', 'Familia Arquibaldo', et 'Furia Bacalhau'.
- Employé Dodsworth : Montre une relation très intense avec le client 'Galleira de gastromonie'.
- Employé Callahan : A une forte activité avec le client 'Furia Bacalhau e Frutos do Mar'.
- Employé Fuller : Montre une forte intensité avec 'LILA-Supermercado'.
- Employé Peacock : Présente un volume élevé avec 'Oceano Atlantico Ltda'.

Distribution Moyenne (Bleu Moyen) :

- Employé Fuller : Est l'un des employés avec la distribution la plus large et constante de commandes, montrant des volumes moyens (autour de 15-25) avec de nombreux clients sur tout le spectre.

- Employé Dodsworth, King et Leverling : Semblent gérer un portefeuille client assez diversifié avec des pics d'intensité marqués.

Observation des Tendances Employé

- Employés les Plus Performants / Demandés (Volume Total) :
  - Fuller et King se distinguent par la fréquence et l'intensité de leurs commandes.
  - Dodsworth et Leverling affichent également une performance très solide, souvent avec des pics de volume importants avec quelques clients clés.
- Employés avec un Portefeuille Client Restreint ou à Faible Volume :
  - Zare, Thorpe, Suyama, Serglenko, Nepper, Koch, Hellung-Larsen, Davollo, Cencini, Buchanan, et Callahan (mis à part son pic avec 'Furia Bacalhau') montrent des volumes de commandes globalement plus faibles ou se concentrent sur un très petit nombre de clients. Par exemple, Zare n'a pratiquement aucune activité visible.

Observation des Tendances Client

- Clients les Plus Fidèles / Grands Commanditaires :
  - Ernste in Connection est le client qui génère le volume de commandes le plus élevé avec un seul employé (King).
  - Des clients comme LILA-Supermercado, Galleira de gastromonie, Oceano Atlantico Ltda, et Furia Bacalhau e Frutos do Mar sont également de grands commanditaires, chacun ayant une relation privilégiée avec un employé principal.
- Clients Nouveaux / à Faible Volume :
  - De nombreux clients à l'extrémité gauche et droite du graphique (par exemple, 'Alfreds Futterkiste', 'Bon app', 'Vins et alcools Chevalier') ne présentent aucune zone bleu foncé, indiquant soit de faibles volumes de commandes, soit une répartition sur de très nombreux employés de manière insignifiante.

Recommandations Basées sur la Matrice

1. Identifier et Analyser les Relations Clés (Risque de Concentration) :
   - Investiguer la dépendance : La relation King - Ernste in Connection est critique. Si l'employé King est absent ou quitte l'entreprise, le client risque d'être mal desservi, ce qui peut entraîner une perte de revenus majeure.
   - Action : Mettre en place un plan de redondance et de transfert de connaissances pour les clients clés gérés par un seul employé (e.g., Dodsworth/Galleira de gastromonie, Peacock/Oceano Atlantico).
2. Optimiser l'Allocation des Tâches :

- Équilibrage de la Charge : Les employés comme Fuller et King gèrent des volumes très élevés. Il pourrait être judicieux de réaffecter des clients à faible volume ou moyennement actifs vers des employés moins chargés (e.g., Zare, Thorpe, Suyama) afin d'éviter l'épuisement professionnel (burnout) chez les plus performants et d'augmenter le volume des autres.

3. Encourager la Diversification :

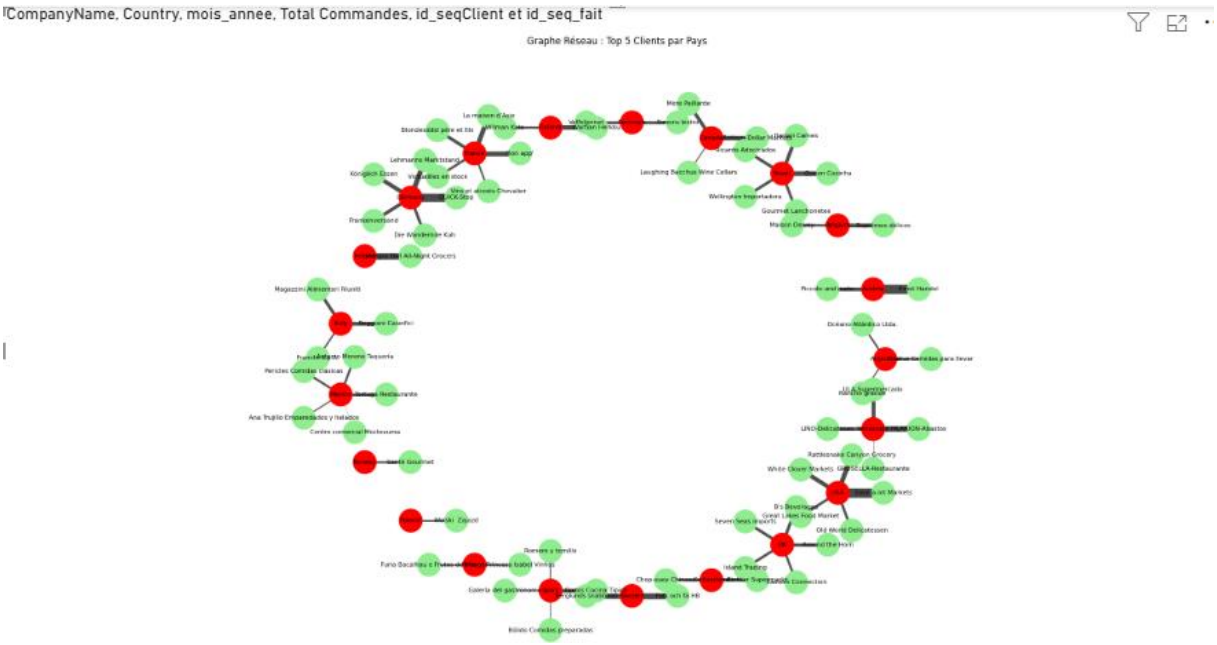
    - Développer les Relations Secondaires : Les employés avec de faibles volumes peuvent être encouragés à développer des relations plus fortes avec les clients qui commandent peu ou à cibler des clients qui sont actuellement desservis de manière non exclusive par d'autres.

4. Récompenser les Performeuses :

    - Reconnaissance : Les performances exceptionnelles de Fuller, King, Dodsworth, et Leverling, en termes de volume de commandes traitées, devraient être reconnues et récompensées.

Top 5 Clients par Pays (graphe réseau) :

Ce graphe a été gênerai par code python



L'analyse du "Top 5 Clients par Pays" est une méthode de segmentation et de hiérarchisation du portefeuille client. Elle consiste à identifier, pour chaque marché géographique (pays), les cinq clients qui génèrent la plus grande valeur .

Recommandation :

- Identifier les leviers de croissance :

Mettre en lumière les clients qui contribuent de manière disproportionnée au succès de l'entreprise dans une région donnée. Ces clients sont les modèles de succès à répliquer.

- Gestion des risques (Concentration) :

Mesurer la **dépendance** de l'entreprise à l'égard de quelques grands acheteurs. Si un seul client représente 80% des commandes dans un pays, son départ est un risque majeur. L'analyse quantifie ce risque.

- Allocation des ressources :

Permettre aux équipes de vente et marketing d'allouer leurs ressources (temps, budget, personnel dédié) en priorité sur les clients qui offrent le meilleur retour sur investissement (ROI), assurant ainsi leur rétention et leur croissance.

- Personnalisation de la relation :

Faciliter la création de stratégies de fidélisation et de négociation sur mesure (tarification, conditions logistiques) adaptées aux besoins et à l'importance de chaque top client.

RECOMMANDATIONS STRATÉGIQUES

Court terme (0-3 mois)

1. **Urgence USA** : Task force pour résoudre le problème de livraison
2. **Top client** : Plan d'action dédié pour les clients stratégiques
3. **Formation** : Programme de montée en compétences des employés

Moyen terme (3-12 mois)

1. **Optimisation logistique** : Système prédictif de gestion des stocks
2. **Transformation digitale** : Tableau de bord temps réel

Long terme (1-3 ans)

1. **Excellence opérationnelle** : Viser 98% de taux de livraison
2. **Croissance équilibrée** : Pénétrer 5 nouveaux marchés majeurs

Tableau comparative Talend et Power bi

Critère	**Talend Open Studio**	**Power BI (Power Query)**
Type d'outil	ETL complet (Extract, Transform, Load) open source	Outil de Business Intelligence avec moteur ETL intégré (Power Query)
Objectif principal	Intégration, nettoyage et transformation des données avant le chargement dans le Data Warehouse	Préparation, transformation et analyse des données pour créer des rapports et dashboards interactifs
Sources de données	Très large : SQL Server, Oracle, Access, fichiers plats, API, web services…	SQL Server, Access, Excel, fichiers plats, API, services en ligne Microsoft (SharePoint, Azure)

Transformation des données	Très avancée : jointures complexes, calculs, conditionnelles, flux multiples, routines personnalisées (Java)	Transformations simples à intermédiaires : nettoyage, fusion, agrégation, colonnes calculées via M ou DAX
Chargement des données	Vers bases de données relationnelles, fichiers, entrepôts de données	Chargement vers le modèle Power BI interne (Data Model) pour visualisation et analyse
Automatisation	Oui, planification via Talend Administration Center ou scripts	Limité : rafraîchissement planifié via Power BI Service
Complexité / Courbe d'apprentissage	Moyenne à élever, nécessite connaissance ETL et concepts BI	Moyenne, interface plus intuitive pour les analystes
Visualisation des données	Non (se concentre sur l'ETL)	Oui, tableaux de bord interactifs, graphiques, KPI
Maintenance / Traçabilité	Forte capacité de journalisation, suivi des erreurs et versioning des jobs	Limitée à Power Query et logs de rafraîchissement
Usage recommandé	Projets nécessitant un ETL robuste, intégration multi-sources, traitement complexe	Analyse rapide, reporting interactif, tableaux de bord décisionnels