

# Amélioration de la gestion des données d'une plateforme e-commerce

## I. Introduction

Dans un secteur e-commerce de plus en plus compétitif, l'exploitation efficace des données est essentielle pour stimuler les ventes, optimiser la gestion des stocks et améliorer la satisfaction client. Ce projet vise à concevoir et à mettre en œuvre un pipeline complet de gestion et d'analyse des données pour une plateforme e-commerce.

En utilisant des outils avancés tels que Talend pour les opérations ETL, SQL Server pour l'entrepôt de données et Power BI pour la visualisation et l'analyse, ce projet garantit un traitement fiable des données, des insights approfondis et une conformité avec les réglementations (comme le RGPD).

L'objectif est de fournir des analyses exploitables aux décideurs tout en maintenant la qualité, la sécurité et la scalabilité des données.

## II. Traitement des données

### 1. Présentation des données

Les données de la plateforme e-commerce sont réparties en quatre catégories principales :

- **Produits** : Contient des informations sur les articles en vente, y compris le nom du produit, la catégorie, la sous-catégorie, le prix, la quantité disponible et les soldes.
- **Clients** : Comprend des informations sur les clients, notamment leur niveau (Gold, Silver, Bronze) ainsi que des données personnelles comme le numéro de téléphone, l'email et la localisation.
- **Fournisseurs** : Regroupe des données sur les fournisseurs, incluant la méthode de livraison, le nom du fournisseur et d'autres détails pertinents.
- **Expéditeurs** : Contient des informations sur les stocks, y compris la quantité disponible, la quantité vendue, les articles en main et ceux encore en stock.

### 2. Le prétraitement des données

Le prétraitement des données a consisté en plusieurs étapes clés pour assurer leur qualité et leur fiabilité :

1. **Correction des dates** : Les dates étaient stockées dans des formats incohérents. Nous avons standardisé toutes les dates au format *MM-DD-YYYY* afin d'assurer une uniformité dans l'ensemble des données.
2. **Correction des prix** : Certains prix étaient incohérents. Pour les corriger, nous avons exploité les colonnes *total\_amount* et *quantity\_vendue* afin de recalculer un prix correct pour chaque produit.
3. **Correction des catégories** : Certaines lignes contenaient des valeurs invalides dans la colonne *categorie*. Pour résoudre ce problème, nous avons utilisé la colonne *sub\_categorie* et remplacé les valeurs invalides par la

catégorie la plus fréquente (mode). Cette correction a été facilitée par le fait qu'il n'existe que cinq catégories distinctes.

4. **Correction des noms de produits** : Pour traiter les incohérences dans la colonne *product\_name*, nous avons utilisé les informations des colonnes *categorie* et *sub\_categorie*. Cependant, cela ne suffisait pas toujours, car plusieurs produits pouvaient appartenir aux mêmes catégories et sous-catégories tout en ayant des noms différents. Pour affiner la correction, nous avons ajouté le prix comme indicateur en utilisant un intervalle défini par la moyenne  $\pm$  l'écart-type des prix des produits similaires.

Ces étapes ont permis d'améliorer la qualité des données avant leur exploitation pour l'analyse et la visualisation.

### III. Base de données

Dans la gestion des données pour la plateforme e-commerce, l'utilisation de Talend pour le prétraitement et l'intégration des données est un point clé pour assurer la qualité des informations avant leur transfert vers SQL Server. Voici un détail plus approfondi des étapes liées à la création de la base de données et de la mise en œuvre de l'architecture.

#### 1. Prétraitement avec Talend

Talend a été utilisé pour les opérations ETL (Extract, Transform, Load), garantissant que les données sont nettoyées, enrichies et prêtes à être intégrées dans le système de gestion des bases de données. Voici les actions principales réalisées :

- **Extraction des données** des différentes sources (base de données existantes, fichiers CSV, etc.)
- **Transformation des données** pour corriger les incohérences, les erreurs et les doublons.
- **Chargement des données** dans la base de données SQL Server selon un modèle structuré et optimisé.

#### 2. Schéma étoile (Star Schema)

Pour la structure de la base de données, nous avons adopté un modèle **schéma étoile**, qui est optimal pour l'analyse des données volumineuses et complexes. Ce modèle comporte des **tables de faits** et des **tables de dimensions** :

- **Tables de Dimensions** :
  - **Produit** : Contient des informations sur chaque produit (ID, nom, catégorie, prix, etc.).
  - **Client** : Contient des informations personnelles sur chaque client (ID, niveau, email, téléphone, etc.).
  - **Fournisseur** : Stocke les informations des fournisseurs (ID, nom, mode de livraison, etc.).
  - **Expéditeur** : Détaille les informations sur la logistique (ID, mode de livraison, quantité, etc.).

- **Tables de Faits :**

- Les **tables de faits** centralisent les informations transactionnelles, comme les ventes effectuées, les quantités vendues, et les informations de paiement. Ces tables sont reliées aux dimensions via des **clés étrangères** (foreign keys), facilitant les jointures et les analyses.

### 3. Sécurité au niveau de la base de données

Pour garantir la sécurité et la conformité avec les réglementations (par exemple, le RGPD), des mesures de sécurité ont été appliquées à plusieurs niveaux :

- **Contrôle d'accès** : Des rôles et permissions ont été définis pour chaque utilisateur, permettant de restreindre l'accès aux données sensibles. Par exemple, seuls les administrateurs peuvent modifier les données personnelles des clients ou des fournisseurs.
- **Chiffrement** : Les données sensibles (comme les informations bancaires et les adresses) ont été cryptées lors de leur stockage dans la base de données.
- **Journalisation** : Un système de journalisation a été mis en place pour suivre toutes les modifications effectuées dans la base de données, permettant de garantir la traçabilité des opérations.

### 4. SCD (Slowly Changing Dimensions)

Le concept des **dimensions à évolution lente (SCD)** a été appliqué pour suivre les changements dans les données des dimensions au fil du temps. Par exemple, si un client change de niveau (de Bronze à Argent), ce changement est enregistré, permettant de suivre l'historique des évolutions dans les dimensions :

- **Type 1 (SCD1)** : Les anciennes données sont écrasées par les nouvelles (pas d'historique).
- **Type 2 (SCD2)** : Les changements sont enregistrés avec une nouvelle ligne, permettant de garder l'historique (ex. : un client ayant plusieurs niveaux au fil du temps).
- **Type 3 (SCD3)** : Seuls les changements récents sont enregistrés, limitant la quantité de données historiques.

Cette gestion des dimensions à évolution lente permet de conserver l'historique des clients, des produits, et d'autres entités importantes pour une analyse chronologique et une étude des tendances.

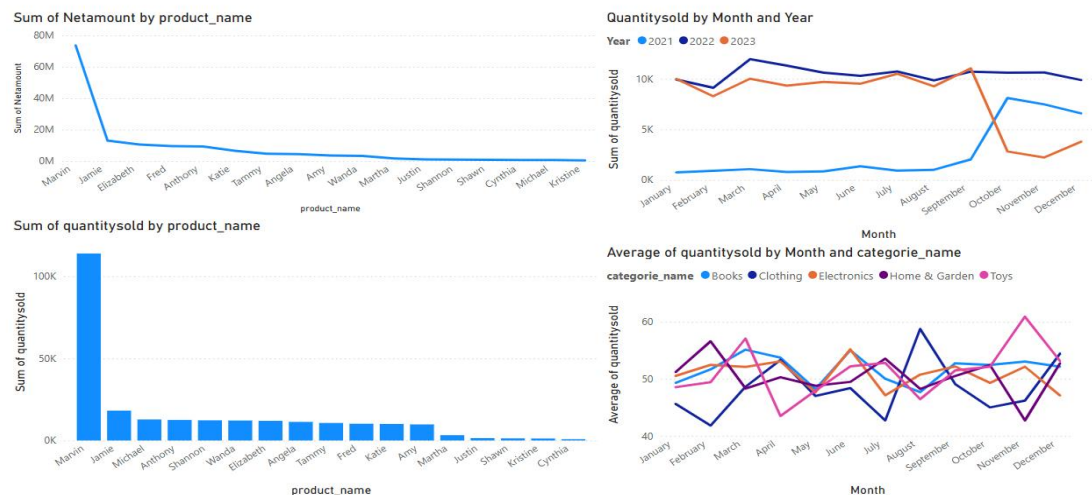
### 5. Intégration et Synchronisation

L'intégration des données depuis Talend vers SQL Server a été réalisée avec un flux continu et automatisé. Cela permet de garantir que les données sont régulièrement mises à jour et cohérentes entre les différentes sources de données.

## IV. Analyses et Visualisations.

### 1. Analyse de la Demande et des Tendances

- **Top Produits** : Certains produits enregistrent des ventes nettement supérieures aux autres (ex : Kevin et Jamie en quantité vendue).
- **Évolution Temporelle** : Les ventes fluctuent tout au long de l'année avec des pics à certaines périodes (ex : septembre et décembre montrent des hausses pour certaines catégories).
- **Catégories de Produits** : Différentes catégories montrent des tendances distinctes, par exemple, les jouets et l'électronique semblent avoir des ventes irrégulières.



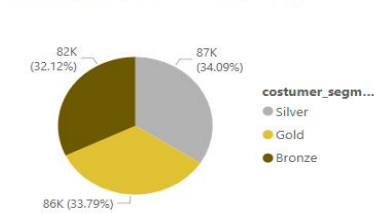
### ❖ Recommandations :

- Se concentrer sur les produits à forte demande pour maximiser les profits.
- Adapter le stock en fonction des tendances mensuelles pour éviter les ruptures ou les surplus.
- Investir dans la publicité pendant les mois à faible demande.

### 2. Analyse des Clients et Segmentation

- **Segments de Clients** : La clientèle est divisée en trois segments (Or, Argent, Bronze) avec une répartition relativement équilibrée.
- **Montant Total des Achats** : Les clients "Gold" et "Silver" contribuent fortement au chiffre d'affaires.
- **Réductions Appliquées** : Les segments Gold et Bronze bénéficient des plus fortes remises.

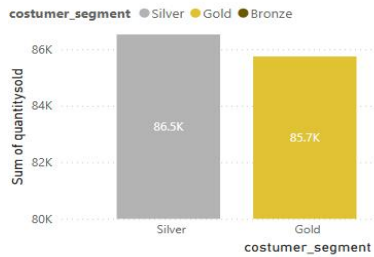
Sum of quantitysold by costumer\_segment



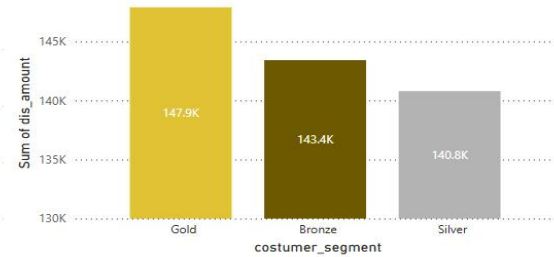
Sum of totalamount by costumer\_location and costumer\_segment



Sum of quantitysold by costumer\_segment and costumer\_segment



Sum of dis\_amount by costumer\_segment



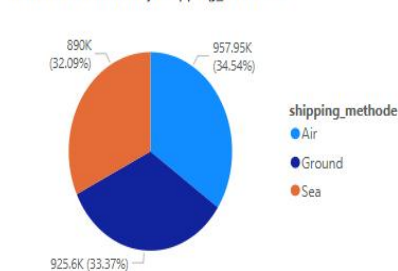
## ❖ Recommandations :

- Fidéliser les clients "Gold" et "Silver" avec des offres exclusives.
- Encourager les clients "Bronze" à monter en gamme via des programmes de fidélité.
- Offrir des promotions ciblées aux segments les moins actifs.

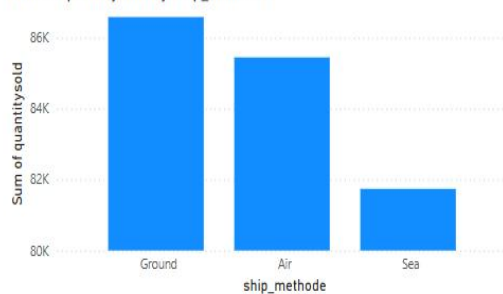
## 3. Analyse Concurrentielle et Positionnement

- **Modes de Livraison :** Les ventes via la livraison terrestre dominent (Ground), tandis que la livraison maritime est moins populaire.
- **Fournisseurs Clés :** Quelques fournisseurs dominent l'approvisionnement, ce qui peut créer une dépendance.
- **Corrélation Prix-Vente :** Une relation positive entre le prix moyen et la quantité vendue suggère que des produits premium trouvent preneur.

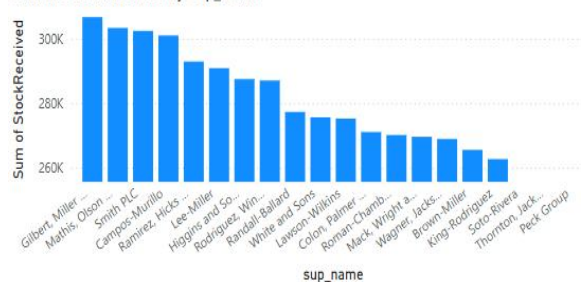
Sum of StockSold by shipping\_methode



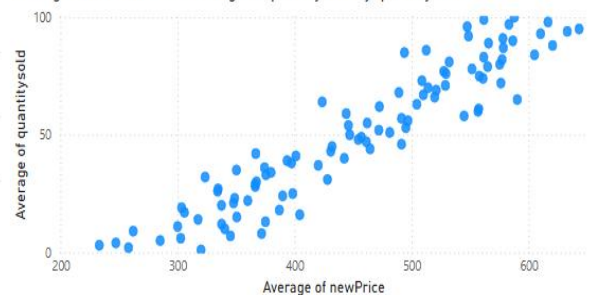
Sum of quantitysold by ship\_methode



Sum of StockReceived by sup\_name



Average of newPrice and Average of quantitysold by quantitysold



### ❖ **Recommandations :**

- Optimiser la logistique en renforçant l'efficacité des livraisons terrestres.
- Diversifier les fournisseurs pour minimiser les risques d'approvisionnement.
- Tester des hausses de prix sur les produits haut de gamme pour maximiser la rentabilité.

## **V. Conclusion**

Ce projet a permis de mettre en place une gestion des données efficace et optimisée pour une plateforme e-commerce, en assurant un traitement rigoureux des informations à l'aide d'outils performants comme Talend, SQL Server, et Power BI. Grâce à des opérations ETL bien structurées et à l'adoption d'un schéma de base de données étoile, les données ont été nettoyées, enrichies, et prêtes pour une analyse approfondie.

Les recommandations issues des analyses, telles que l'optimisation des stocks en fonction des tendances de vente et l'amélioration de la fidélisation des clients, offrent des leviers concrets pour stimuler la croissance et améliorer la rentabilité. De plus, l'intégration de processus de sécurité rigoureux permet de garantir la conformité avec les réglementations en matière de protection des données, tout en assurant la traçabilité et l'intégrité des informations.

En somme, cette approche globale et méthodique offre à la plateforme e-commerce une visibilité accrue sur ses opérations et son marché, facilitant ainsi la prise de décisions stratégiques pour rester compétitive dans un environnement dynamique.