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

Introduction :

Présentation du projet :

- Le projet a pour objectif d'optimiser la gestion des données d'une plateforme e-commerce via un pipeline complet d'extraction, transformation et chargement (ETL) des données. Ce projet vise à renforcer l'analyse des ventes et des inventaires tout en assurant une prise de décision efficace grâce à des outils d'analyse visuelle comme Power BI.

Objectifs :

1. Exploiter efficacement les données de ventes et d'inventaire pour fournir des insights décisionnels.
2. Implémenter des processus ETL robustes à l'aide de Talend et structurer les données dans SQL Server.
3. Fournir des analyses visuelles via Power BI pour soutenir la prise de décision.
4. Garantir la sécurité et la conformité aux normes RGPD.

Public cible :

- Équipe dirigeante, responsables des ventes et de la logistique.

Modèles de données :

Structure de la base de données :

- La base de données est structurée pour permettre une gestion optimale des données relatives aux ventes, à l'inventaire et à la logistique. Elle utilise des tables de faits et de dimensions afin d'assurer des requêtes efficaces et une analyse approfondie.

Description des tables principales et de leurs champs :

Tables de dimensions :

- **DateDimension** : Contient des informations temporelles (jour, mois, année..).
- **CustomerDimension** : Informations sur les clients (nom, adresse, email).
- **ProductDimension** : Détails des produits (catégorie, prix, etc.).
- **SupplierDimension** : Informations sur les fournisseurs.
- **ShipperDimension** : Détails des transporteurs.

Tables de faits :

- **SalesFact** : Quantité vendue, prix unitaire, total des ventes.
- **InventoryFact** : Quantité en stock.....

Contraintes appliquées :

- **Clés primaires** pour garantir l'unicité des enregistrements.
- **Clés étrangères** pour assurer l'intégrité des données entre les tables.
- **Index** pour optimiser la vitesse des requêtes sur les données.

Processus ETL :

Extraction :

- **Sources des données** : Les données sont fournies principalement sous les formats JSON et CSV.
- **Outils utilisés pour l'extraction** : Talend, qui permet d'extraire efficacement les données à partir de différentes sources et formats.

Transformation :

- **Nettoyage des données** : Traitement des valeurs manquantes et harmonisation des types de données (par exemple, dates et nombres).
- **Calculs dérivés** : Ajout de nouvelles colonnes, telles que le chiffre d'affaires (calculé par quantité * prix).
- **Gestion des Slowly Changing Dimensions (SCD)** : Mise en œuvre des processus permettant de suivre l'évolution des dimensions au fil du temps, en gérant à la fois les changements mineurs et majeurs, tout en garantissant l'intégrité des données historiques.

Chargement :

- **Stockage des données transformées dans SQL Server** pour garantir une gestion centralisée et sécurisée des données.
- **Automatisation du processus avec Talend** pour la planification des tâches ETL et le suivi des exécutions.

Optimisation de la base de données :

Pour améliorer la performance de la gestion des données dans SQL Server, plusieurs optimisations ont été appliquées :

- **Index** : Création d'index sur les colonnes fréquemment utilisées dans les requêtes pour accélérer l'accès aux données et réduire le temps de réponse.
- **Partitionnement** : Les tables volumineuses, telles que les tables de faits (SalesFact, InventoryFact), ont été partitionnées pour optimiser l'efficacité des requêtes et réduire les coûts de gestion des données.
- **Mise à jour automatique des statistiques (Auto Update Stats)** : Configuration des statistiques pour s'assurer que les données de la base de données restent actualisées et optimisées pour le plan d'exécution des requêtes.

Ces optimisations permettent de réduire les temps de réponse des requêtes complexes et assurent la scalabilité du système à mesure que les volumes de données augmentent.

Validation des données :

- Afin de garantir que les données extraites et transformées respectent les règles de qualité avant leur chargement dans la base de données, un processus de validation a été mis en place. Ce processus inclut des vérifications pour s'assurer que les données sont complètes, cohérentes, et bien formatées. Des contrôles sont également effectués pour valider l'intégrité des données et éviter les erreurs lors de l'insertion dans la base.

Mesures de sécurité et conformité :

Sécurité des données :

- **Cryptage des données sensibles** pour garantir leur confidentialité.
- **Contrôle d'accès basé sur les rôles** pour limiter l'accès aux données sensibles en fonction des rôles des utilisateurs.

Conformité RGPD :

- Respect des exigences du RGPD en matière de protection des données personnelles.
- **Masquage dynamique des données sensibles** dans Power BI pour garantir que les données sensibles ne soient pas exposées dans les rapports et analyses.

Annexes :

1. Liste des scripts SQL utilisés pour créer la base de données

- Vous trouverez les scripts SQL nécessaires pour créer la base de données, définir les relations, les contraintes et les index dans le dossier SQL server/ du projet. Ces scripts permettent de configurer la structure de la base de données SQL Server.

2. Liste des workflows ETL dans Talend

- Les détails des workflows ETL créés avec Talend, y compris les étapes d'extraction, de transformation et de chargement des données, sont disponibles dans le dossier `talend/`. Ce dossier contient les fichiers des processus ETL.

3. Liens vers les ressources

- Pour plus d'informations et des ressources d'apprentissage, veuillez consulter les liens suivants :
  - Documentation Talend
  - Apprentissage Power BI
  - SQL Server Indexing Best Practices

4. Tests de validation des données

- Les exemples de tests de validation des données (contrôle des valeurs manquantes, vérification de la cohérence des données, etc.) peuvent être trouvés dans le dossier SQL server/.