

**Les solutions et les préconisations d’utilisation de la base de données**

**Date** : 21/10/2024

* **Introduction**
  + **Objectif du rapport**
* Ce rapport vise à analyser les différentes solutions pour extraire, traiter et visualiser les données de afin d’optimiser leur utilisation dans un environnement de data visualisation.
* Les solutions explorées incluent une connexion directe à la base de données, une extraction en format CSV et l'utilisation d'outils d'ETL (Extraction, Transformation, Loading) tels que Power Query ou Knime. Nous évaluerons chaque option en fonction de la structure de la base de données, de l’efficacité d’utilisation des données et de la cohérence avec nos besoins métier.
  + **Contexte de l'analyse**
* La base de données contient des informations essentielles sur les produits, les ventes, les promotions et la finance, structurées autour de quatre tables principales : Web, Finance, Promo et Sales. Ces données couvrent la période du 1er octobre 2022 au 30 septembre 2023, et l’objectif est de les rendre exploitables dans des outils de data visualisation pour le suivi des performances et la prise de décision stratégique.
* **État des lieux et évaluation de la pertinence** 
  + **2.1 Description de la situation actuelle**
* Actuellement, notre base de données est stockée sous format SQLite, regroupant des données cruciales comme les prix, les ventes mensuelles, les stocks, et les promotions. Le dictionnaire de données présente une bonne structure, avec des clés primaires identifiant de manière unique chaque enregistrement.
  + **2.2 Évaluation de la pertinence des données existantes**
* Les données présentes sont pertinentes pour des analyses approfondies sur les performances de vente, les effets des promotions, et les fluctuations des stocks. Elles sont bien structurées, avec des types de données variés (entiers, flottants, chaînes de caractères et dates) permettant une analyse multidimensionnelle.
* Cependant, il existe des limitations potentielles en termes de qualité des données, comme les incohérences entre les systèmes, les informations spécifiques à WordPress (Comment\_status, Ping\_status), qui peuvent être non essentielles, et les colonnes non pertinentes pour l'analyse, comme les colonnes liées à la publication des articles (post\_author, post\_date, etc.)
  + **2.3 Identification des principaux problèmes ou enjeux**

Les principaux défis concernent :

* **Redondance des données WordPress** : certaines colonnes spécifiques à WordPress ne semblent pas essentielles pour l’analyse des performances produits.
* **Qualité des données** : Colonnes vides ou inutiles pour les analyses.
* **La table Finance** n’a pas de distinction claire entre les produits en stock et ceux en rupture, ce qui pourrait limiter l'analyse des tendances de disponibilité.
* **L'actualisation et la synchronisation des données** : Si les données changent régulièrement, une solution de connexion directe pourrait être préférée pour garantir des données à jour.
* **Le traitement de grandes quantités de données** : Certaines analyses peuvent nécessiter de manipuler d’importants volumes de données, et un outil ETL pourrait être plus adapté pour préparer les données avant leur exploitation dans les visualisations.
* **Besoin d'outils** 
  + **3.1 Identification des outils nécessaires pour collecter les données**
* **Extraction directe via SQL** : Une solution de connexion directe permet d'accéder en temps réel à la base de données SQLite, idéale pour les analyses en direct.
* **Extraction CSV** : Pour des analyses plus ponctuelles, l’extraction en CSV peut être suffisante.
  + **3.2 Identification des outils nécessaires pour traiter les données**
* **ETL (Power Query, Knime)** : Ces outils permettent d'extraire, transformer et charger les données, facilitant la préparation des données pour une analyse approfondie et l'automatisation des transformations répétitives.
* **SQL** : pour effectuer des requêtes et des jointures complexes entre les tables.
  + **3.3 Identification des outils nécessaires pour analyser les données**
* **Outils de visualisation** : Power BI ou Tableau peuvent se connecter aux bases de données et aux fichiers CSV transformés, permettant de créer des tableaux de bord interactifs et la visualisation des KPIs.
* **Excel :** pour des analyses plus simples ou des rapports ponctuels.
* **Solutions d'extraction, de traitement et de visualisation** 
  + **4.1 Solutions proposées pour l'extraction des données**
* **Connexion directe** : extraction directe des données via SQL permet d'accéder aux données en temps réel pour des visualisations dynamiques.
* **Extraction CSV** : Solution simple pour des analyses ponctuelles.
* **ETL** : Les outils comme Power Query ou Knime permettent d'automatiser l'extraction, tout en transformant les données pour les rendre prêtes à l'emploi.
  + **4.2 Solutions proposées pour le traitement des données**
* **ETL (Power Query)** : Outil de Microsoft Excel et Power BI, idéal pour manipuler les données en amont. Permet de transformer facilement les données, d’ajouter des colonnes calculées et de filtrer les données non pertinentes.
  + **4.3 Solutions proposées pour la visualisation des données**

**Power BI** : Solution puissante pour créer des rapports interactifs et des tableaux de bord. Compatible avec les connexions directes et les fichiers CSV, et peut facilement intégrer les transformations de Power Query.

* **Cohérence des solutions avec le besoin** 
  + **5.1 Explication détaillée de chaque solution proposée**

**Connexion directe** : La connexion directe permet à l’outil de visualisation d’accéder en temps réel aux données depuis leur source (base de données, entrepôt de données, etc.). Cela signifie que chaque fois qu’un rapport est consulté ou actualisé, les données sont récupérées directement depuis la source sans passer par une étape de stockage intermédiaire.

Cette solution est idéale lorsque les données doivent être consultées en temps réel, par exemple dans le cas de tableaux de bord opérationnels utilisés pour le suivi quotidien des indicateurs de performance (KPIs). Elle est souvent privilégiée dans des environnements où les informations évoluent rapidement.

**Extraction CSV** L’extraction CSV consiste à exporter les données depuis leur source sous forme de fichier CSV, puis de charger ce fichier dans l’outil de visualisation. Cette méthode est adaptée pour des analyses ponctuelles ou lorsque les données ne changent pas fréquemment.

L’extraction en CSV convient aux analyses ponctuelles, aux rapports qui ne sont actualisés qu’occasionnellement ou lorsque la simplicité et la rapidité sont les principaux critères. Elle peut aussi être utile dans un contexte de partage de données limité.

**ETL avec Power Query** : ETL (Extract, Transform, Load) avec Power Query consiste à extraire les données de leur source, les transformer en fonction des besoins analytiques (nettoyage, formatage, filtrage), puis les charger dans l’outil de visualisation. Power Query permet d’automatiser ces étapes, ce qui en fait un choix performant pour des analyses régulières.

L’ETL avec Power Query est adapté aux analyses régulières où les données nécessitent une transformation répétée avant de pouvoir être analysées. Elle est particulièrement utile dans des environnements où la qualité et la cohérence des données sont essentielles, et où les rapports sont actualisés fréquemment.

* + **5.2 Alignement des solutions avec les besoins identifiés**
* **Pour les analyses en temps réel** : La connexion directe est la plus appropriée.
* **Pour les analyses périodiques ou ponctuelles** : L'extraction en CSV avec traitement préalable peut suffire.
* **Pour une préparation récurrente et automatisée des données** : Un ETL est recommandé.
  + **5.3 Avantages et limites de chaque approche**

| **Solution** | **Avantages** | **Limites** |
| --- | --- | --- |
| **Connexion directe** | - Accès en temps réel - Réduction de la redondance - Flexibilité pour analyses avancées. | - Configuration complexe - Charge sur la base de données - Problèmes de sécurité et de stabilité réseau. |
| **Extraction CSV** | - Simple à mettre en œuvre - Compatible avec de nombreux outils - Moins de charge sur la base - Contrôle manuel de la qualité des données. | - Données statiques - Désynchronisation fréquente - Limité aux petits volumes de données - Risques d'erreurs humaines. |
| **ETL (Power Query...)** | - Automatisation des transformations - Intégration de sources multiples - Préparation avancée des données - Suivi et transparence des étapes. | - Configuration initiale chronophage - Nécessite des compétences techniques - Performance limitée avec de grands volumes - Dépendance à l'outil et coûts élevés. |

En résumé, la **connexion directe** est privilégiée pour des besoins en temps réel, tandis que l’**extraction en CSV** est idéale pour des analyses ponctuelles de faible volume.

L’**ETL** est particulièrement adapté pour des analyses récurrentes nécessitant des données transformées ou consolidées provenant de multiples sources.

**Conclusion**

* + **Synthèse des principales conclusions de l'analyse**
* Pour une utilisation optimisée de notre base de données dans un outil de data visualisation, les solutions de connexion directe et d’ETL avec Power Query sont recommandées selon les besoins en actualisation et en traitement des données.
  + **Recommandations pour la mise en œuvre des solutions proposées**
* **Pour les analyses en temps réel** : Mettre en place une connexion directe à la base de données.
* **Pour les analyses périodiques et automatisées** : Utiliser un ETL comme Power Query pour extraire, transformer et charger les données vers Power BI ou Tableau.

**Annexes**

* **Liste des sources de données utilisées**
* Table Web
* Finance
* Promo
* Sales.
* **Description détaillée des outils recommandés**
  + Power Query pour la préparation des données.
  + Power BI pour la visualisation dynamique.
* **Exemples de visualisations de données proposées.**
  + Graphiques sur les ventes mensuelles par produit.
  + Tableaux comparatifs des prix avant et après promotion.
  + Visualisation des stocks restants en fin de mois.