# Rapport d'Analyse sur les différentes solutions de mise à disposition des données Bottleneck dans un outil de data visualisation

Date : 29 janvier 2025

## 1. Introduction

### 1.1 Objectif du rapport

Ce rapport a pour objectif de comparer les différentes solutions de data visualisation à envisager, afin d'identifier la solution la plus pertinente pour la visualisation et l'analyse des données de l'entreprise Bottleneck.

### 1.2 Contexte de l'analyse

Florian, le PDG de Bottleneck, a décidé de rendre les données de l'entreprise accessibles à tous ses collaborateurs, à travers une solution de data visualisation interactive, afin de les aider à suivre les performances de l'entreprise, gérer les variations économiques, et prendre des décisions éclairées au quotidien.

## 2. État des lieux et évaluation de la pertinence

### 2.1 Description de la situation actuelle

L'entreprise a du retard sur l'utilisation de ses données, mais celles-ci sont maintenant à jour, nettoyées et accessibles à partir d'une seule et unique base de données SQLite, ce qui va permettre de les exploiter avec une solution de data visualisation. Ce tableau de bord sera utilisé par Florian et les chefs de produits dans le but d'améliorer la compétitivité, l'organisation et la croissance de Bottleneck.

## 2.2 Évaluation de la pertinence des données existantes

Les données existantes sont pertinentes car à jour, nettoyées et centralisées dans une BDD, cependant il faudra :

Effectuer des traitements sur le format des données, notamment au niveau des colonnes :

- SKU : Format Texte à convertir en nombre entier (Integer) + supprimer les espaces
- Prix : Montants à convertir en devise
- Contenance : Remplacer '0' dans la colonne et régler le problème des 100cl
- 'Post_title' et 'Post_excerpt' : retirer les guillemets
- Supprimer des colonnes inutiles pour cet exercice…

Créer des nouvelles colonnes ou mesures pour calculer :

- La marge commerciale
- Les prix hors taxe
- La rotation des stocks...

## 2.3 Identification des principaux problèmes ou enjeux

Les principaux problèmes ou enjeux sont :

- L'**automatisation** des processus de mise à jour : Les données doivent être mises à jour de manière régulière sans nécessiter d'interventions manuelles fréquentes.
- La **qualité des données** : Malgré le nettoyage effectué, il est essentiel de maintenir un processus continu de gestion de la qualité des données afin d'éviter toute dérive.
- Le choix de l'**Outil de data visualisation** : Il faudra s'assurer que l'outil choisi permette à l'équipe de facilement visualiser et analyser les données tout en garantissant des performances optimales.

# 3. Besoin d'outils et solutions identifiées

## 3.1 Identification des outils nécessaires pour collecter les données

Pour collecter et centraliser les données dans une base de données unique, l'entreprise doit veiller à ce que les flux de données issus des différentes sources (Web, Finance, Promo et Sales) soient intégrés efficacement dans une base SQLite. Pour ce faire, nous utiliserons **SQLite ODBC Driver**, qui permettra d'établir une connexion entre l'outil ETL (Extract, Transform, Load) Power Query et la base de données SQLite.

La démarche consistera à utiliser Power Query pour automatiser le processus d'extraction des données depuis les différentes sources, leur transformation pour assurer la cohérence et la qualité des informations, puis leur chargement dans la base SQLite via l'ODBC Driver. Cette approche garantira une intégration fluide et efficace des données, facilitant ainsi leur exploitation et leur analyse.

## 3.2 Identification des outils nécessaires pour traiter les données

Les données doivent être nettoyées, transformées et structurées correctement avant toute analyse ou visualisation. Des outils ETL comme **PowerQuery** sont appropriés pour cette tâche.

Ils permettront de :

- Nettoyer les données en éliminant les valeurs manquantes ou aberrantes,
- Transformer les données sous des formats et structures adaptés aux analyses,
- Effectuer des calculs ou des agrégations pour enrichir les datasets.

## 3.3 Identification des outils nécessaires pour analyser les données

L'analyse des données peut être effectuée à l'aide de plateformes de **business intelligence** (BI) comme **Power BI**, **Tableau**, ou **Google Data Studio**. Nous vous proposons Power BI qui permettra de :

- Explorer les données à travers des filtres, des tableaux croisés dynamiques et des graphiques,
- Mettre en place des **tableaux de bord interactifs** et des rapports personnalisés pour différents utilisateurs.

# 4. Solutions d'extraction, de traitement et de visualisation (retenues)

Les solutions décrites ci-après ont été étudiées avant de procéder aux choix d'outils précédemment mentionnés.

## 4.1 Solutions proposées pour l'extraction des données

Les solutions envisagées pour l'extraction des données étaient les suivantes :

- **Connexion directe à la base de données SQLite** : Permet d'interroger directement la base de données à l'aide d'un outil de visualisation de données comme Tableau ou Power BI.
- **Extraction manuelle en CSV** : Simple mais peu adaptée pour des mises à jour fréquentes, car elle nécessite une gestion manuelle des fichiers.
- **Utilisation d'un ETL (Power Query, Knime, etc.)** : Permet l'automatisation de l'extraction et de la transformation des données avant leur chargement dans l'outil de visualisation.

## 4.2 Solutions proposées pour le traitement des données

Les solutions de traitement comprenaient :

- **Power Query**

  - 🔹 **Idéal pour** : Manipulations simples et interactives des données
  - 🔹 **Cas d'usage** : Nettoyage des données, transformation de colonnes, fusion de sources de données
  - 🔹 **Avantages** :
  - ✔️ Interface intuitive (drag-and-drop)
  - ✔️ Intégré à Excel et Power BI
  - ✔️ Accessible aux utilisateurs non techniques
  - 🔹 **Limites** :
  - ❌ Moins efficace pour les gros volumes de données
  - ❌ Pas conçu pour des analyses avancées ou l'automatisation complète

- **Knime**

  - 🔹 **Idéal pour** : Analyses complexes et traitement de gros volumes de données
  - 🔹 **Cas d'usage** : Machine learning, analyse statistique, workflows avancés
  - 🔹 **Avantages** :
  - ✔️ Open-source et flexible
  - ✔️ Interface visuelle avec une bibliothèque de nœuds pour les transformations avancées
  - ✔️ Compatible avec Python, R et d'autres outils analytiques
  - 🔹 **Limites** :
  - ❌ Courbe d'apprentissage plus élevée
  - ❌ Peut nécessiter des ressources importantes pour le traitement de très grands jeux de données

- **Automatisation via un ETL**

  - 🔹 **Idéal pour** : Intégration et transformation de données de manière automatisée
  - 🔹 **Cas d'usage** : Extraction de données de différentes sources, transformation récurrente, chargement dans un Data Warehouse
  - 🔹 **Avantages** :
  - ✔️ Exécution planifiée sans intervention manuelle
  - ✔️ Gestion efficace de grandes volumétries de données
  - ✔️ Fiabilité et suivi des flux
  - 🔹 **Limites** :
  - ❌ Configuration initiale plus complexe
  - ❌ Nécessite souvent des compétences techniques

  **-> Nos experts prenant en charge la configuration initiale, ces limites ne sont donc pas à prendre en considération.**

### 4.3 Solutions proposées pour la visualisation des données

Les solutions de visualisation proposées étaient :

- **Power BI** : Outil puissant pour créer des tableaux de bord interactifs et des analyses ad hoc, permettant de connecter facilement à SQLite et d'offrir une personnalisation avancée.
- **Tableau** : Idéal pour des visualisations complexes et une interface utilisateur fluide, avec des analyses avancées et des tableaux de bord dynamiques.
- **Google Data Studio** : Solution gratuite et simple, adaptée à des besoins moins complexes et permettant un partage facile via le cloud.

## 5. Cohérence des solutions avec le besoin

### 5.1 Explication de chaque solution proposée

Chaque solution a été choisie en fonction des besoins de l'entreprise, en tenant compte de la simplicité d'utilisation, des coûts et de l'intégration avec les données existantes. L'utilisation d'une **connexion directe à la base de données** et d'un **ETL pour automatiser les processus** garantit une mise à jour continue des données. **L'outil de visualisation Power BI** offre des interfaces interactives, adaptées aux utilisateurs non techniques, avec des fonctionnalités avancées pour l'analyse des données.

### 5.2 Alignement des solutions avec les besoins identifiés

Les solutions proposées répondent aux besoins identifiés de manière équilibrée :

- **Accessibilité** des données : Les outils de visualisation se connectent directement à la base de données, garantissant une mise à jour en temps réel.
- **Automatisation** : L'utilisation d'un ETL pour le traitement des données et l'exportation automatisée dans des formats visuels répond au besoin de réduire la charge manuelle.
- **Simplicité d'utilisation** : Power BI et Tableau sont des outils conviviaux, adaptés à des utilisateurs non techniques.

## 5.3 Avantages et limites de chaque approche

1. **Connexion directe à la base de données**

Avantages :

- **Temps réel** : Permet de travailler avec des données à jour, sans avoir à les extraire manuellement.
- **Automatisation** : Les outils de visualisation (comme Tableau, Power BI, etc.) peuvent se connecter directement à la base de données et récupérer les données à la volée, ce qui facilite les mises à jour automatiques des rapports.
- **Flexibilité** : Pratique pour exécuter des requêtes complexes ou dynamiques sur les données.

Limites :

- **Performances** : Si la base de données est volumineuse ou mal optimisée, cela peut entraîner des ralentissements ou des problèmes de performance.
- **Dépendance à l'infrastructure** : La connexion à la base de données nécessite une connexion continue et peut poser des problèmes en cas de coupures réseau ou de défaillances du serveur.
- **Sécurité** : Il faut garantir un accès sécurisé, en particulier si des données sensibles sont impliquées.

2. **Extraction des données en CSV de la base de données**

Avantages :

- **Simplicité** : Cette méthode est très simple et ne nécessite pas de connaissances avancées en outils de data visualisation ou en programmation.
- **Portabilité** : Le fichier CSV peut être partagé facilement entre différents utilisateurs et applications.
- **Indépendance** : Une fois les données extraites, la base de données n'est plus nécessaire pour travailler dessus, ce qui peut être pratique pour des analyses ponctuelles.

Limites :

- **Données statiques** : Les données deviennent rapidement obsolètes si elles ne sont pas régulièrement mises à jour. Cela implique de refaire l'extraction chaque fois que vous aurez besoin de nouvelles données.
- **Volume de données** : Le format CSV peut devenir difficile à gérer avec de très grands ensembles de données (manque de flexibilité pour les grands volumes).
- **Limitation fonctionnelle** : Certaines informations comme les relations entre les tables peuvent être perdues dans l'exportation.

### 3. Utilisation d'un ETL (PowerQuery, Knime, Talend, etc.)

Avantages :

- **Automatisation et transformation** : Permet de manipuler et transformer les données avant de les importer dans l'outil de visualisation. Cela offre une flexibilité supérieure pour préparer des données complexes.
- **Gestion de flux de données** : Les ETL permettent d'automatiser le processus d'extraction, de transformation et de chargement (ETL) sur une base régulière, ce qui facilite la gestion des données à long terme.
- **Nettoyage des données** : Très utile pour nettoyer ou transformer les données avant de les analyser.

Limites :

- **Coût** : Certains outils ETL peuvent être coûteux, cependant des versions gratuites ou open-source existent, ce que nous utiliserons.
- **Complexité** : Les outils ETL nécessitent une certaine courbe d'apprentissage.
- **Temps de mise en place** : Il peut y avoir un investissement initial en temps pour configurer correctement l'ETL.

**Ces dernières limites ne s'appliquent pas à la situation de Bottelneck, nos équipes étant missionnées pour mettre en place le système.**

## Conclusion

### Synthèse des principales conclusions de l'analyse

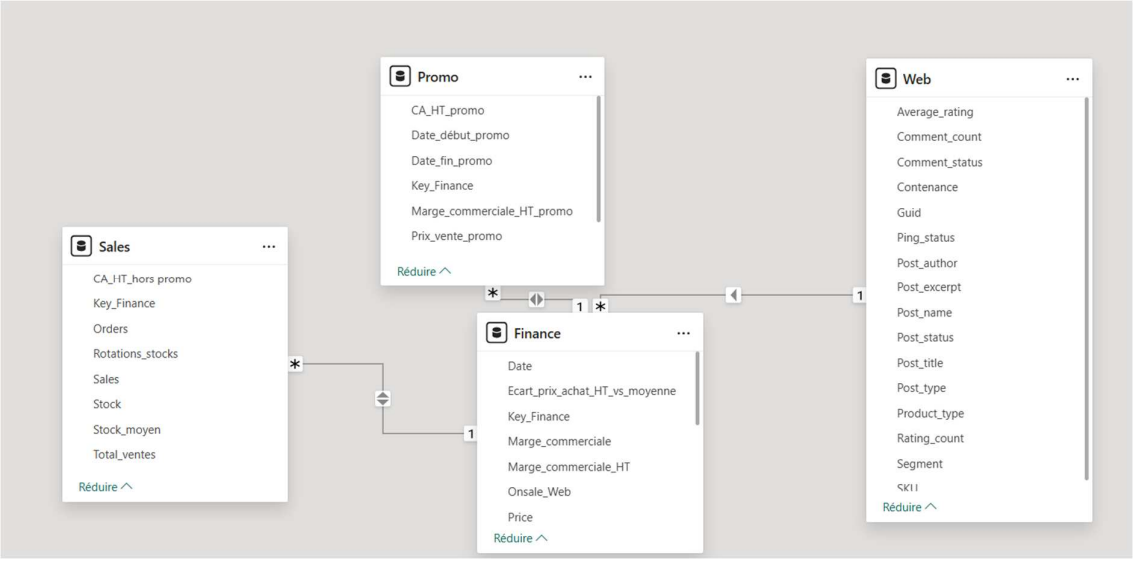
Bottleneck dispose maintenant de données centralisées et nettoyées, prêtes à être exploitées dans un cadre de visualisation. La mise en place d'une solution automatisée pour l'extraction, le traitement et la visualisation des données est primordiale pour gagner en efficacité. **L'utilisation d'outils BI comme Power BI, couplée avec un ETL comme Power Query pour l'automatisation des processus, semble être la solution la plus adaptée.**

### Recommandations pour la mise en œuvre des solutions proposées

1. Mettre en place un flux automatisé d'extraction et de transformation des données via un ETL (PowerQuery).
2. Connecter les outils de visualisation (Power BI) à la base de données pour garantir une accessibilité continue aux données actualisées.
3. Former les utilisateurs aux outils de visualisation pour s'assurer de leur adoption et de leur efficacité.

## [Annexes] Liste des sources de données utilisées

Schéma relationnel




Dictionnaire de données

| | Nom des colonnes | Type de données | Taille | Clé | Description |
|---|---|---|---|---|---|
| Web | SKU | INT | | Clé primaire | Référence |
| | Virtual | INT | | | Informations Wordpress |
| | Downloadable | INT | | | Informations Wordpress |
| | Rating_count | FLOAT | | | Nombre de notes |
| | Average_rating | FLOAT | | | Note moyenne des clients |
| | Post_author | VARCHAR | | | Informations sur la création de l'article sur le site |
| | Post_date | DATE | | | Date de création de l'article sur le site |
| | Post_date_gmt | DATE | | | Date de création de l'article sur le site |
| | Product_type | VARCHAR | | | Type de produits (cognac, vin, champagne, etc.) |
| | Segment | VARCHAR | | | Segments (spiritueux, vin, sans alcool, etc.) |
| | Post_title | VARCHAR | | | Nom du Vin |
| | Post_excerpt | VARCHAR | | | Description du vin |
| | Contenance | VARCHAR | | | Contenance |
| | Post_status | VARCHAR | | | Articles publiés (publish) ou en attente (wait) |
| | Comment_status | VARCHAR | | | Informations Wordpress |
| | Ping_status | VARCHAR | | | Informations Wordpress |
| | Post_name | VARCHAR | | | Lien Wordpress |
| | Post_modified | DATE | | | Date de modification de l'article |
| | Post_modified_gmt | DATE | | | Date de modification de l'article |
| | Post_parent | VARCHAR | | | Informations Wordpress |
| | Guid | VARCHAR | | | URL vers les produits |
| | Post_type | VARCHAR | | | Type de produits (produits ou images) |
| | Comment_count | VARCHAR | | | Nombre de commentaires |
| Finance | Key_Finance | VARCHAR | | Clé primaire | Identifiant unique (concaténation de la date (année-mois et de la référence)) |
| | Date | DATE | | | Date du mois (exemple : 01/10/2022 = données d'octobre 2022) |
| | SKU | VARCHAR | | | Référence |
| | Taxe_Status | VARCHAR | | | Produit vendable (taxable) ou non (vide) |
| | Onsale_Web | BOOLEAN | | | Présent sur le site (1) ou non (0) |
| | Price | FLOAT | | | Prix de vente |
| | Purchase_Price | FLOAT | | | Prix d'achat |
| Promo | Key_Finance | VARCHAR | | Clé primaire | Identifiant unique (concaténation de la date (année-mois et de la référence)) |
| | Start_Date | DATE | | | Date de début de promotion |
| | End_Date | DATE | | | Date de fin de promotion |
| | Sales_Offer | INT | | | Prix de vente en promotion |
| | Price_Offer | FLOAT | | | Nombre de produits vendus pendant la promotion |
| Sales | Key_Finance | VARCHAR | | Clé primaire | Identifiant unique (concaténation de la date (année-mois et de la référence)) |
| | Sales | INT | | | Vente sur le mois (hors promotions) |
| | Orders | INT | | | Commandes fournisseurs sur le mois |
| | Stock | INT | | | Stock de l'entreprise au dernier jour du mois |

Export SQLite

Les différentes étapes d'extraction des données se trouvent sur les slides de présentation.