



---

# CONCEVOIR UN SYSTEME DE RECOMMANDATION DES LIVRES

---

Projet de validation du matière : NoSQL avancé et Big data



**ÉLABORÉ PAR :**

**GAMATI WIDED  
SOUID ANISSA**

**ENCADRÉ PAR :**

**Dr. Omheni Nizar**

# TABLE DE MATIERE

<b>Introduction générale.....</b>	<b>1</b>
<b>Chapitre 1 : Présentation du cadre du projet.....</b>	<b>2</b>
1. Introduction .....	2
2. Contexte générale et problématique .....	2
3. Choix du dataset utilisé .....	3
4. Conclusion.....	5
<b>Chapitre 2: Prétraitement et stockage du données .....</b>	<b>6</b>
1. Introduction .....	6
2. Environnement de travail .....	6
2.1. Apache Spark .....	6
2.2. MongoDB .....	6
3. Phase de prétraitement .....	7
4. Phase de stockage.....	7
5. Conclusion.....	8
<b>Chapitre 3 : Phase de modélisation .....</b>	<b>9</b>
1. Introduction .....	9
2. Modélisation.....	9
2.1. Modèle 1 : KnnWithMeans.....	9
2.2. Modèle 2 : ALS.....	10
3. Evaluation des modèles .....	12
4. Conclusion.....	12
<b>Chapitre 4 : Phase de visualisation des données .....</b>	<b>13</b>
1. Introduction .....	13
2. Environnement de travail .....	13

2.1.	MongoDB .....	13
2.2.	Power BI .....	13
2.3.	Connection entre eux .....	13
3.	Génération du rapport.....	15
4.	Conclusion.....	16

# LISTE DES TABLEAU

Tableau 1: Description des attributs du premier fichier .....	3
Tableau 2: Description des attributs du second fichier .....	4

# LISTE DES FIGURES

Figure 1: Diagramme ERD .....	3
Figure 2: La base de donnée.....	8
Figure 3: Terminale de connexion entre la BD et le Power BI.....	14
Figure 4: connexion établit.....	14
Figure 5: Rapport / Tableau de bord Power BI.....	15

# Introduction générale

Dans le contexte évolutif du monde numérique actuel, la gestion et l'exploitation des vastes quantités de données sont devenues des enjeux cruciaux pour de nombreuses industries. Le secteur de la recommandation, en particulier, a bénéficié de manière significative des avancées en matière de bases de données NoSQL, de technologies Big Data et d'algorithmes de Machine Learning (ML).

Notre projet de fin de semestre s'inscrit dans cette dynamique, visant à mettre en œuvre ces techniques de pointe pour concevoir un système de recommandation de livres innovant tout en exploitant une dataset massive concernant les livres publiés sur Amazon.

Le présent rapport sera structuré en quatre chapitres comme suit : Le premier chapitre sera dédié à la présentation du cadre du projet, le deuxième abordera la phase de prétraitement des données, le troisième couvrira la phase de modélisation, et enfin, le dernier chapitre traitera de la visualisation des données.

# Chapitre 1 : Présentation du cadre du projet

## 1. Introduction

L'étude et la présentation du cadre de projet constituent une étape préliminaire pour la réalisation d'une solution informatique. Ainsi, ce chapitre sera consacré à la présentation du contexte et du cadre du projet.

## 2. Contexte générale et problématique

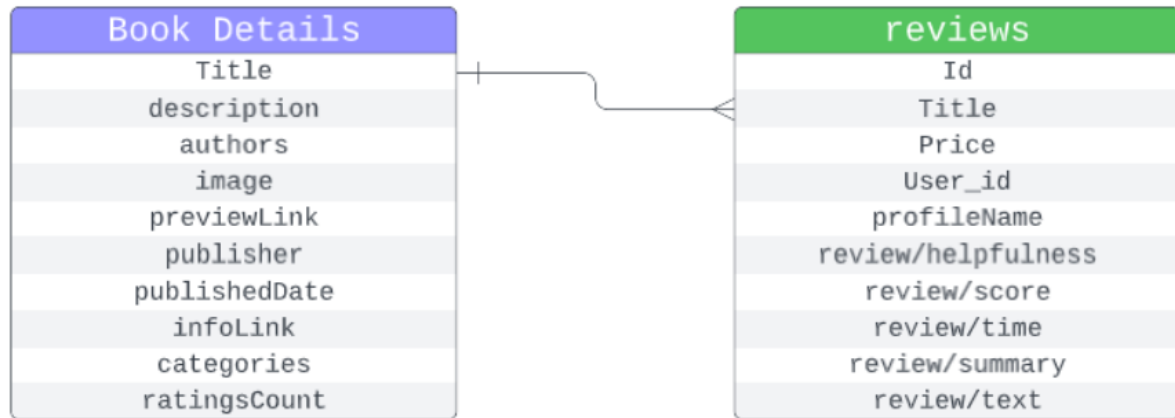
À l'ère de l'information numérique, la quantité exponentielle de données générées par les utilisateurs crée des opportunités considérables mais également des défis, notamment dans le domaine de la recommandation de livres en ligne. Face à la surabondance d'informations et de choix, les utilisateurs se trouvent confrontés à la difficulté de découvrir des livres correspondant véritablement à leurs goûts.

C'est dans ce contexte que notre projet prend tout son sens. En utilisant des technologies avancées telles que les bases de données NoSQL, le Big Data et l'apprentissage automatique (ML), nous cherchons à concevoir un système de recommandation innovant capable de surmonter les défis liés à la gestion de données massives, à la complexité des préférences de lecture et à la nécessité d'une personnalisation accrue.

Notre objectif est de répondre efficacement à cette problématique en offrant une solution évolutive et précise qui exploite pleinement le potentiel de ces technologies, fournissant ainsi des recommandations de livres personnalisées et pertinentes pour les utilisateurs avides de lecture.

### 3. Choix du dataset utilisé

Pour réaliser ce projet on a utilisé une dataset composé de deux fichiers, comme illustré dans le diagramme ERD ci-dessus :



*Figure 1: Diagramme ERD*

Le premier fichier, "reviews", contient des retours d'information sur 3 millions d'utilisateurs portant sur 212 404 livres uniques. Ce dataset fait partie de l'ensemble de données des avis Amazon, incluant 142,8 millions d'avis couvrant la période de mai 1996 à juillet 2014.

Ce fichier comprend les attributs suivant :

*Tableau 1: Description des attributs du premier fichier*

Attribut	Description
Id	L'identifiant du livre
Title	Le titre du livre
Price	Prix du livre
User_Id	L'identifiant d'utilisateur qui a évalué le livre
profileName	Nom du profile de l'utilisateur

review/helpfulness	Note d'assistance de l'avis, par exemple 2/3
review/score	Evaluation varie de 0 à 5
review/time	Date de l'évaluation
review/summary	Résumé de l'évaluation
review/text	Le texte de l'évaluation

Le deuxième fichier, "Books Details", contient des informations détaillées sur les 212,404 livres uniques. Il a été construit en utilisant l'API Google Books pour obtenir des informations détaillées sur les livres notés dans le premier fichier. Les caractéristiques de ce fichier incluent :

*Tableau 2: Description des attributs du second fichier*

Attribut	Description
Title	Titre du livre
Describe	Description du livre
authors	Nom d'auteur du livre
image	url de la couverture du livre
previewLink	Lien pour accéder à ce livre sur Google Books
Publisher	Nom de l'éditeur
publishedDate	Date de publication
infoLink	lien d'information supplémentaire sur le livre sur Google Books
categories	Genres du livre
ratingCount	La note moyenne pour le livre



## **4. Conclusion**

Dans ce chapitre, on a abordé, dans un premier lieu, le contexte du projet ainsi la problématique. Et le choix du dataset, dans un second lieu.

# Chapitre 2: Prétraitement et stockage du données

## 1. Introduction

Ce chapitre se consacre à détailler la phase de prétraitement des données, ainsi que leur stockage dans une base de données orientée document, tout en fournissant une description de l'environnement de travail.

## 2. Environnement de travail

Pour mener à bien ce projet, il est crucial de sélectionner les outils et technologies appropriés. C'est pourquoi, dans cette section, nous exposons nos choix technologiques.

### 2.1. Apache Spark

Apache Spark est un framework open source de calcul distribué. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie.

C'est un moteur d'analyse unifié et ultra-rapide pour le traitement de données à grande échelle. Il permet d'effectuer des analyses de grande ampleur par le biais de machines de Clusters. Il est essentiellement dédié au Big Data et Machine Learning.

### 2.2. MongoDB

MongoDB est une base de données NoSQL orientée document, qui permet de stocker des données sous forme de documents JSON. Elle est très utilisée pour les applications web et mobiles, car elle permet de stocker des données de manière flexible et de les interroger facilement. Elle est également très performante et évolutive, ce qui la rend adaptée aux applications à forte charge.

### 3. Phase de prétraitement

La phase de prétraitement des données a été entreprise avec rigueur en utilisant Apache Spark pour effectuer le nettoyage des données. Le processus a débuté par l'établissement d'une connexion solide à notre ensemble de données massives. Par la suite, différentes étapes cruciales ont été suivies pour garantir la qualité et la cohérence des informations tel que :

- Analyser les données afin de mettre une stratégie adéquate pour les nettoyer.
- Extraire les données relatives aux utilisateurs et les mettre dans une autre fichier csv.
- Supprimer les colonnes inutiles.
- Renommer les colonnes par des noms significatifs.
- Eliminer les éléments nuls et les duplications.

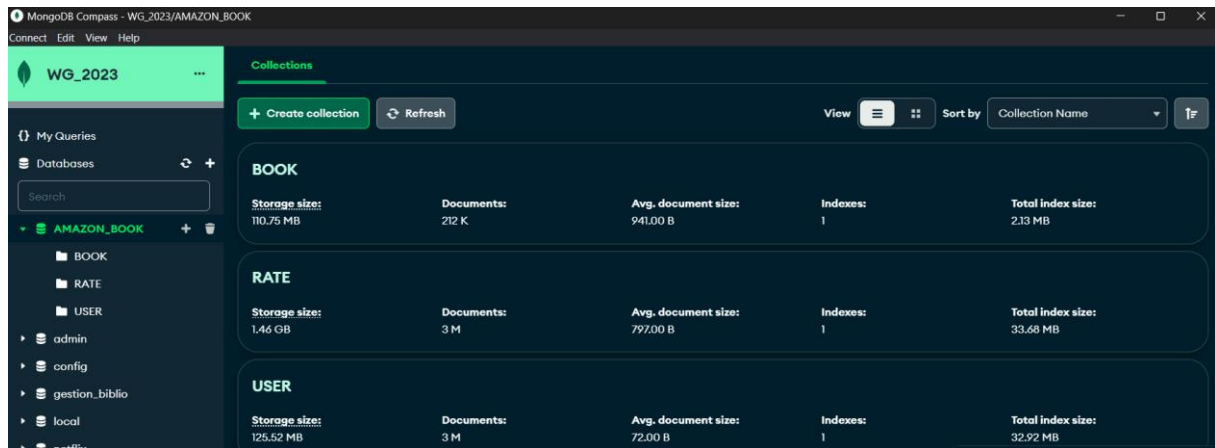
⇒ Ces étapes successives ont été orchestrées dans un pipeline Spark, permettant une exécution efficace et évolutive du prétraitement des données, jetant ainsi les bases d'une analyse et d'une modélisation ultérieures de qualité.

### 4. Phase de stockage

La phase de stockage des données a été exécutée de manière efficace en intégrant directement le jeu de données prétraité dans une base de données document MongoDB.

La décision stratégique d'opter pour MongoDB repose sur sa capacité exceptionnelle à gérer des volumes massifs de données de manière flexible et évolutive. Cette base de données NoSQL offre une structure de stockage orientée document qui s'adapte parfaitement à la nature variée de nos données, simplifiant ainsi la modélisation et l'accès aux informations. La robustesse de MongoDB dans la gestion des données massives se manifeste notamment dans sa capacité à évoluer horizontalement pour répondre à l'augmentation continue du volume de données.

En adoptant MongoDB, nous garantissons une gestion efficace, une évolutivité sans heurts et une rapidité d'accès aux données, éléments cruciaux pour notre projet de recommandation de livres basé sur des données massives et diversifiées.



The screenshot shows the MongoDB Compass interface for a database named 'WG\_2023'. The left sidebar lists the database 'AMAZON\_BOOK' and its collections: 'BOOK', 'RATE', and 'USER'. The main panel displays the 'Collections' tab with a table of collection statistics.

Collection	Storage size	Documents	Avg. document size	Indexes	Total index size
BOOK	110.75 MB	212 K	941.00 B	1	2.13 MB
RATE	1.46 GB	3 M	797.00 B	1	33.68 MB
USER	125.52 MB	3 M	72.00 B	1	32.92 MB

*Figure 2: La base de donnée*

## 5. Conclusion

Ce chapitre explore le prétraitement des données avec Apache Spark et leur stockage efficace dans MongoDB, soulignant la stratégie technologique pour gérer les données massives. Ces étapes fondamentales jettent les bases d'un système de recommandation de livres robuste.

# Chapitre 3 : Phase de modélisation

## 1. Introduction

Ce qui suit est dédiée à l'exploration approfondie de la phase de modélisation, mettant en lumière le processus de création de modèles de recommandation. Nous nous pencherons également sur leur implémentation pratique et leur intégration dans notre environnement de travail, tout en considérant le stockage dans une base de données orientée document.

## 2. Modélisation

Dans cette section, on va mettre en place notre system de recommandation tout en commençant par un prétraitement dédié à cette phase et tout mettant l'accent sur deux différents modèles de Machine Learning et la différence entre eux.

### 2.1. Modèle 1 : KnnWithMeans

Ce modèle est trainé sur un échantillon de 22,000 exemple (ainsi que l'ensemble de test est de 20% du taille de l'échantillon) : Nous avons travaillé avec un ensemble de données de petite taille en raison des limitations de performance matérielles. Il est toutefois possible de travailler sur l'ensemble complet de données.

Le principe de KNNWithMeans (k-nearest neighbors with means), repose sur l'idée de prédire les préférences d'un utilisateur en se basant sur les notations moyennes de ses voisins les plus proches. Les paramètres clés de l'algorithme comprennent:

- **k** (nombre de voisins): Il détermine le nombre de voisins à considérer lors de la prédiction. Un choix judicieux de **k** peut influencer la précision de la recommandation.
- **Min\_k** (nombre minimum de voisins requis): Il spécifie le nombre minimum de voisins nécessaires pour qu'une prédiction soit effectuée. Cela peut aider à éviter des prédictions basées sur un petit nombre de voisins.

- Sim\_options (options de similarité): Ces options définissent la métrique de similarité utilisée pour mesurer la proximité entre utilisateurs ou items. Des métriques telles que la similarité cosinus ou la similarité de Pearson peuvent être configurées.
- Verbose (verboosité): Ce paramètre contrôle le niveau de détails des messages de sortie pendant l'exécution de l'algorithme.

En utilisant ces paramètres, KNNWithMeans cherche à créer des recommandations personnalisées en exploitant les préférences moyennes des utilisateurs similaires, offrant ainsi une approche collaborative pour la recommandation de livres.

## 2.2. Modèle 2 : ALS

Le modèle ALS (Alternating Least Squares) est entraîné sur un échantillon de 1 million d'exemples, avec l'ensemble de test représentant 20% de la taille de l'échantillon. Cette taille d'échantillon a été choisie en raison des limitations de performance matérielles, bien qu'il soit envisageable de travailler avec l'ensemble complet de données. L'algorithme ALS fonctionne sur le principe de minimisation des erreurs quadratiques alternatives pour prédire les préférences des utilisateurs. Ses paramètres clés comprennent:

- Rank (rang): Détermine la dimension latente du modèle, influençant la complexité et la capacité de généralisation.
- MaxIter (nombre maximum d'itérations): Spécifie le nombre maximum d'itérations que l'algorithme effectue pendant l'entraînement.
- RegParam (paramètre de régularisation): Contrôle la régularisation du modèle pour éviter le surajustement.
- Nonnegative (contrainte de non-négativité): Si activé, impose des contraintes pour que les valeurs latentes restent non négatives.
- ImplicitPrefs (préférences implicites): Si activé, adapte le modèle pour les préférences implicites au lieu des notations explicites.

- ColdStartStrategy (stratégie pour les éléments inconnus): Spécifie la stratégie à adopter pour les éléments sans données d'entraînement.

En utilisant ces paramètres, le modèle ALS cherche à apprendre les caractéristiques latentes des utilisateurs et des items, fournissant ainsi des recommandations de livres personnalisées basées sur des données massives, avec une approche collaborative robuste.

En expliquant encore, Le modèle ALS fonctionne en minimisant alternativement les erreurs quadratiques entre les prédictions du modèle et les notations réelles des utilisateurs pour les items, et vice versa. Il s'agit d'une technique de factorisation matricielle qui apprend des représentations latentes pour les utilisateurs et les items.

En d'autres termes, le modèle ALS cherche à décomposer la matrice de notation utilisateur-item en deux matrices plus petites représentant les caractéristiques latentes des utilisateurs et des items. Cette approche permet de capturer les relations complexes entre les utilisateurs et les items, offrant ainsi la capacité de faire des recommandations personnalisées. Les paramètres tels que le rang (Rank) et le paramètre de régularisation (RegParam) influencent la dimension de ces caractéristiques latentes et la régularisation du modèle, respectivement, contribuant ainsi à la précision des prédictions.

### **Mais c'est quoi la représentation latente ?**

En apprentissage automatique, la "représentation latente" fait référence à une représentation intermédiaire apprise d'un ensemble de données. C'est une représentation mathématique qui capture les caractéristiques importantes et sous-jacentes des données de manière abstraite. Cette représentation est souvent de dimension réduite par rapport aux données d'origine et permet de condenser l'information tout en préservant les relations significatives entre les exemples.

Dans le contexte des modèles de recommandation, comme le modèle ALS, la représentation latente peut être associée aux caractéristiques cachées des utilisateurs et des items. Ces caractéristiques ne sont pas directement observables, mais le modèle les apprend afin de capturer les relations complexes entre les utilisateurs et les items, facilitant ainsi la génération de recommandations personnalisées.

### 3. Evaluation des modèles

L'évaluation des modèles de recommandation revêt une importance cruciale pour mesurer leur performance et leur capacité à générer des prédictions précises.

Dans notre étude, le modèle KNNWithMeans, formé sur un échantillon de 20 000 exemples, affiche un RMSE (Root Mean Squared Error) de 1.19. Ce résultat témoigne d'une performance relativement bonne du modèle, indiquant que les prédictions du modèle sont en moyenne éloignées d'environ 1.19 unité de la réalité.

D'autre part, le modèle ALS, entraîné sur un ensemble de données plus vaste de 1 million d'exemples, présente initialement un RMSE de 4. Cependant, grâce à un processus d'optimisation, le RMSE est considérablement amélioré, passant successivement à 3.65, puis à 2.96, jusqu'à atteindre une valeur de 1.5. Ces résultats démontrent l'efficacité de l'optimisation du modèle ALS, qui a permis de réduire significativement l'erreur de prédiction.

En conclusion, bien que le modèle KNNWithMeans affiche un RMSE inférieur, le modèle ALS, après optimisation, atteint un niveau de performance remarquable, se positionnant avec un RMSE final de 1.5, ce qui suggère une meilleure précision dans la prédiction des préférences des utilisateurs. Ainsi, le modèle ALS optimisé peut être considéré comme plus approprié pour notre application de recommandation de livres, offrant des prédictions plus précises et adaptées aux besoins des utilisateurs.

### 4. Conclusion

Ce chapitre de modélisation a évalué les performances de deux modèles, KNNWithMeans et ALS, pour le système de recommandation de livres. Le KNNWithMeans a montré un RMSE de 1.19 sur 20 000 exemples, tandis que le modèle ALS, après optimisation, a atteint un RMSE de 1.5 sur un ensemble de données plus vaste de 1 million d'exemples. En raison de sa précision élevée, le modèle ALS optimisé est privilégié pour les étapes ultérieures du projet.



# Chapitre 4 : Phase de visualisation des données

## 1. Introduction

Ce chapitre est consacré à explorer l'importance cruciale de la visualisation dans notre projet. La représentation graphique des données offre un moyen puissant de synthétiser des informations complexes, facilitant ainsi la transmission efficace des idées et des concepts.

## 2. Environnement de travail

### 2.1. MongoDB

En tant que base de données NoSQL, se distingue par sa flexibilité et sa capacité à stocker des données semi-structurées ou non structurées. Dans ce contexte, il sert de réservoir robuste pour les informations diverses générées par diverses sources.

### 2.2. Power BI

Power BI, une solution phare de business intelligence, se distingue par son ensemble complet de fonctionnalités qui vont au-delà de la simple visualisation et analyse de données. Avec son interface intuitive et ses outils puissants, Power BI permet une exploration approfondie des tendances, la création de rapports interactifs et la génération de tableaux de bord dynamiques. Cette plateforme offre une perspective holistique, favorisant une prise de décision éclairée au sein des organisations. Sa capacité à intégrer des données de sources diverses, à automatiser les mises à jour et à faciliter la collaboration entre les équipes en fait une plateforme incontournable, propulsant les entreprises vers une gestion agile et stratégique de leurs données.

### 2.3. Connection entre eux

La connexion entre MongoDB et Power BI devient un élément essentiel de cette infrastructure. Elle permet l'extraction dynamique des données stockées dans MongoDB

pour les intégrer de manière transparente dans Power BI. Cette liaison facilite la mise à jour en temps réel des visualisations, garantissant que les rapports et les tableaux de bord reflètent toujours les données les plus récentes.

L'ensemble du processus de l'extraction des données depuis MongoDB jusqu'à la création de tableaux de bord dynamiques sur Power BI, offre aux utilisateurs un accès direct et convivial à des parties précieuses.

Pour connecter au MongoDB et Power BI a la fois on doit ouvrir ces deux terminales :

```

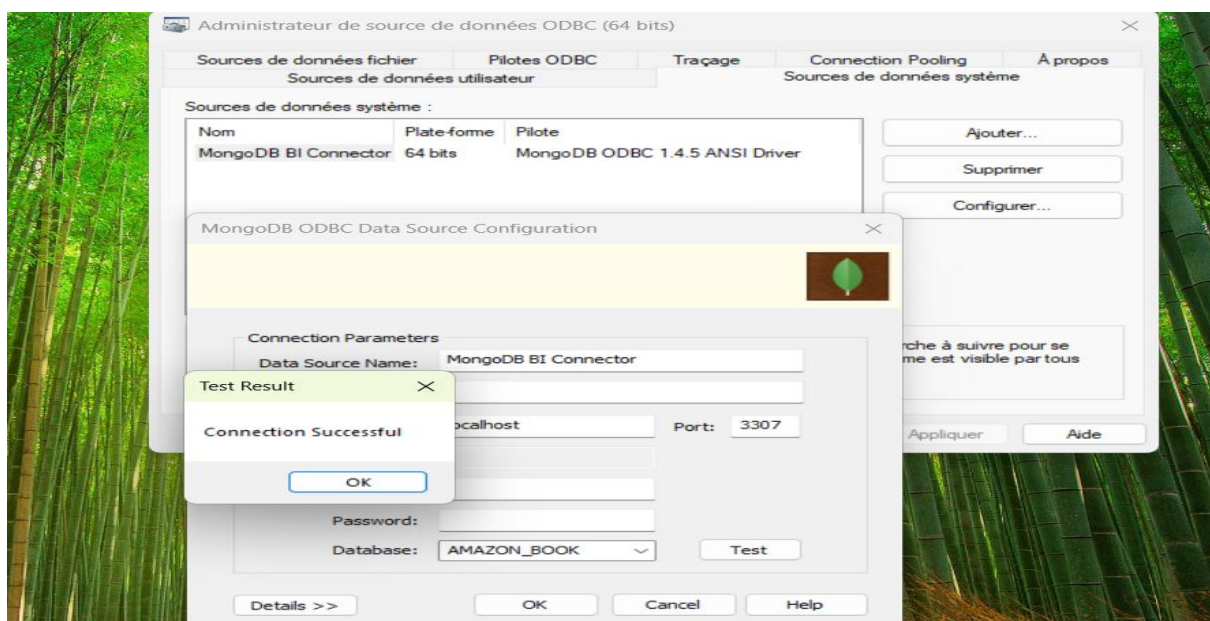
C:\Program Files\MongoDB\ x + -
2023-12-12T13:42:19.462+0100 I CONTROL [initandlisten] mongosql starting
: version=v2.14.11 pid=9752 host=LAPTOP-6V0LK1PN
2023-12-12T13:42:19.491+0100 I CONTROL [initandlisten] git version: 50701
14f09366f9d61a9de205596ef733f704d5b
2023-12-12T13:42:19.491+0100 I CONTROL [initandlisten] OpenSSL version Op
enSSL 1.0.2n-fips 7 Dec 2017 (built with OpenSSL 1.0.2s 28 May 2019)
2023-12-12T13:42:19.491+0100 I CONTROL [initandlisten] options: {}
2023-12-12T13:42:19.491+0100 I CONTROL [initandlisten] ** WARNING: Access
control is not enabled for mongosql.
2023-12-12T13:42:19.491+0100 I CONTROL [initandlisten]
2023-12-12T13:42:19.494+0100 I NETWORK [initandlisten] waiting for connec
tions at 127.0.0.1:3307
2023-12-12T13:42:20.143+0100 I SCHEMA [sampler] sampling MongoDB for sch
ema...
2023-12-12T13:42:24.690+0100 I SCHEMA [sampler] mapped schema for 3 name
spaces: "AMAZON_BOOK" (3): [{"book", "rate", "user"}]
2023-12-12T13:43:11.603+0100 I NETWORK [conn1] connection accepted from 1
27.0.0.1:53201 #1 (1 connection now open)
2023-12-12T13:43:11.643+0100 I NETWORK [conn1] end connection 127.0.0.1:5
3201 (0 connections now open)
2023-12-12T13:43:20.588+0100 I NETWORK [conn2] connection accepted from 1
27.0.0.1:53207 #2 (1 connection now open)
2023-12-12T13:43:20.632+0100 I NETWORK [conn2] end connection 127.0.0.1:5
3207 (0 connections now open)

mongosh mongodb://127.0.0.1 x + -
Please enter a MongoDB connection string (Default: mongodb://localhost/):
Current Mongosh Log ID: 657855465a3856583a5aa90d
Connecting to: mongodb://127.0.0.1:27017/?directConnection=true&ser
verSelectionTimeoutMS=2000&appName=mongosh+2.1.0
Using MongoDB: 7.0.4
Using Mongosh: 2.1.0
mongosh 2.1.1 is available for download: https://www.mongodb.com/try/downloa
d/shell
For mongosh info see: https://docs.mongodb.com/mongodb-shell/

-----
The server generated these startup warnings when booting
2023-12-11T08:59:04.149+01:00: Access control is not enabled for the data
base. Read and write access to data and configuration is unrestricted
-----
test> |
  
```

**Figure 3:** Terminale de connexion entre la BD et le Power BI

Puis, on établit la connexion :



**Figure 4:** connexion établit

### 3. Ggénération du rapport

À mesure que les organisations accumulent d'énormes volumes d'informations, la nécessité de donner un sens à ces données devient impérative. Dans ce paysage complexe, la visualisation de données émerge comme un moyen essentiel pour rendre compréhensibles des ensembles de données souvent vastes et complexes. Les graphiques, les tableaux de bord interactifs et d'autres formes de représentations visuelles offrent la possibilité de transcender les barrières linguistiques des données brutes, permettant aux utilisateurs de percevoir rapidement des tendances, des schémas et des relations.

Cette partie se penchera sur les enjeux de la visualisation de données, cherchant à définir des approches et des lignes directrices qui maximisent la clarté et l'efficacité de la communication visuelle tout en surmontant les défis inhérents à la complexité des données tout en utilisant Power BI pour générer un rapport qui contient des représentations graphiques des données pour faciliter la prise des décisions futur.

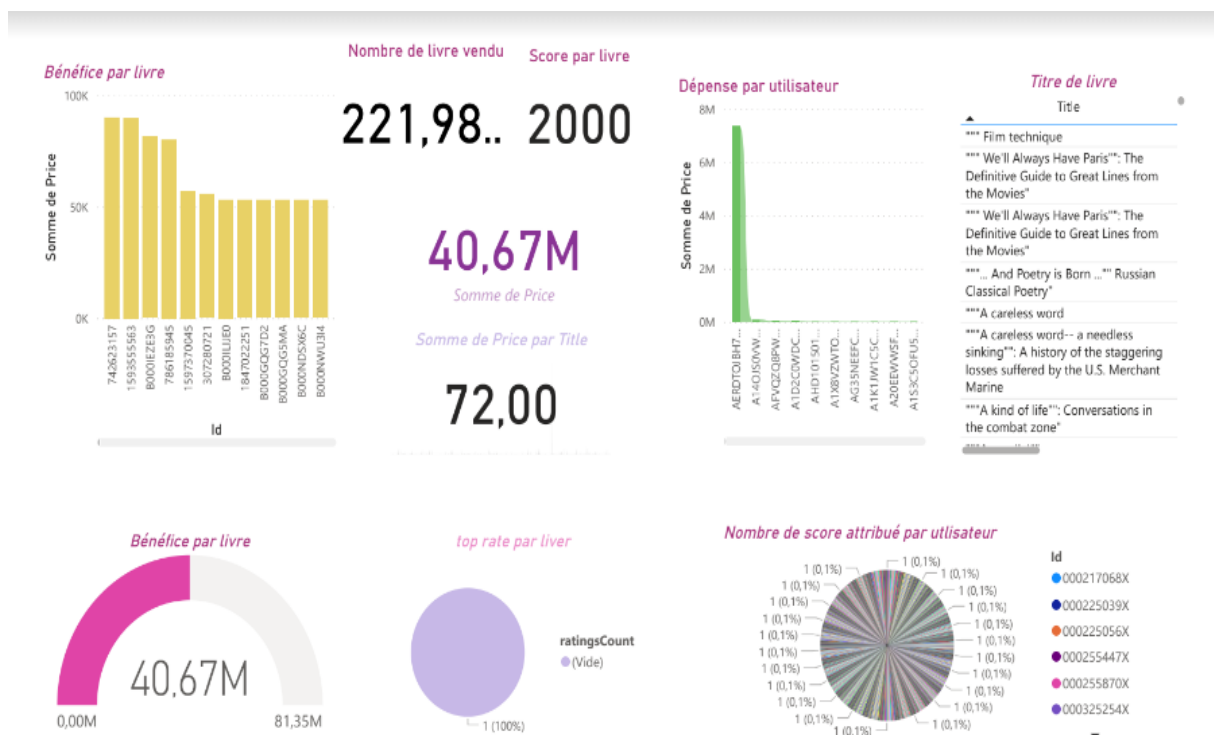


Figure 5: Rapport / Tableau de bord Power BI

## **4. Conclusion**

Ce chapitre de visualisation a permis d'exploiter les données stockées dans MongoDB en les connectant à Power BI. La création d'un tableau de bord interactif a offert une représentation visuelle riche et informative.