

# Visualisation des Données MongoDB dans Google Data Studio

## Introduction

Dans le cadre de ce projet, nous avons réalisé une intégration des données de MongoDB avec Google Data Studio afin de produire des tableaux de bord dynamiques et interactifs. Cette intégration a été rendue possible grâce à l'utilisation de **CData Connector**, qui joue le rôle d'une passerelle permettant de connecter les données NoSQL à une plateforme de visualisation avancée.

L'objectif principal était de permettre une prise de décision basée sur des données en temps quasi réel grâce à un tableau de bord consolidé.

## Architecture de la Solution

1. **Source des données :**
  - Les données sont stockées dans une base de données **MongoDB**. Elles comprennent des informations relatives aux ventes, aux utilisateurs, et aux produits.
  - Structure des données : Les collections MongoDB contiennent des documents JSON imbriqués, typiques des bases NoSQL.
2. **Passerelle :**
  - **CData MongoDB Connector** a été utilisé pour rendre les données MongoDB accessibles à Google Data Studio via une connexion ODBC/JDBC.
3. **Outil de visualisation :**
  - **Google Data Studio** a été choisi pour sa capacité à offrir une visualisation intuitive et interactive.

## Étapes Clés du Projet

### 1. Connexion à MongoDB via CData

La première étape a consisté à configurer le connecteur CData pour qu'il se connecte à la base MongoDB. Les paramètres suivants ont été utilisés :

- **Serveur** : L'adresse IP de MongoDB a été spécifiée.
- **Port** : Utilisation du port par défaut de MongoDB, 27017.
- **Base de données cible** : Une base spécifique contenant les données nécessaires a été sélectionnée.
- **Authentification** : La sécurité a été assurée via un système de login avec un utilisateur dédié.

CData a ensuite exposé les collections MongoDB sous forme de tables accessibles, facilitant ainsi leur traitement.

## 2. Intégration dans Google Data Studio

Une fois les données disponibles via le connecteur, une source de données a été créée dans Google Data Studio. L'interface intuitive de Data Studio a permis de mapper directement les champs disponibles dans MongoDB.

## 3. Transformation des Données

Certaines données brutes nécessitaient des transformations avant d'être visualisées :

- **Nettoyage des données** : Suppression des champs non pertinents.
- **Agrégations** : Regroupement des ventes par mois.
- **Typage** : Conversion explicite des dates et des valeurs numériques pour garantir une utilisation correcte dans les graphiques.

```
from pymongo import MongoClient
from pymongo.errors import ConnectionFailure
from bson.objectid import ObjectId
import json
from nltk.stem import WordNetLemmatizer

# Initialize the lemmatizer for singularizing words
lemmatizer = WordNetLemmatizer()

# MongoDB Connection String
connection_string =
"mongodb+srv://laamiriouail:XrgeZ1PTd01phHVB@cluster0.5kdba.mong
odb.net/articles_bi?retryWrites=true&w=majority&appName=Cluster0
"

try:
    # Connect to MongoDB
    client = MongoClient(connection_string)
    client.admin.command('ping') # Test the connection
    print("Connected to MongoDB!")

    # Access the database and collection
    db = client["articles_bi"]
    collection = db['articles']

    # Update documents where "universities" field is missing
    result = collection.update_many(
        {"universities": {"$exists": False}}, # Match documents
        without the "universities" field
```

```

        {"$set": {"universities": ["Private School"]}} # Set
"universities" to ["Ecole Privee"]
    )
    # Retrieve all documents
    documents = collection.find()
    documents_list = list(documents)

    # Save the data to a JSON file
    with open("articles_data.json", "w") as file:
        json.dump(documents_list, file, default=str, indent=4)
    print("Data saved to 'articles_data.json'.")

    # # Read data from articles_data.json
    # with open("articles_data1.json", "r") as file:
    #     documents_list = json.load(file)
    # # Update documents in MongoDB
    # for record in documents_list:
    #     if "_id" in record:
    #         # Convert `_id` string back to ObjectId
    #         record["_id"] = ObjectId(record["_id"])

    #         # Update the document based on `_id`
    #         result = collection.replace_one({"_id":
record["_id"]}, record)
    #         if result.matched_count > 0:
    #             print(f"Document with _id {record['_id']}
updated.")
    #         else:
    #             print(f"No matching document found for _id
{record['_id']}.")
    #         else:
    #             print("Record missing '_id', skipping update.")

    # print("All documents updated in MongoDB.")

except ConnectionFailure as e:
    print(f"Could not connect to MongoDB: {e}")
except Exception as e:
    print(f"An error occurred: {e}")

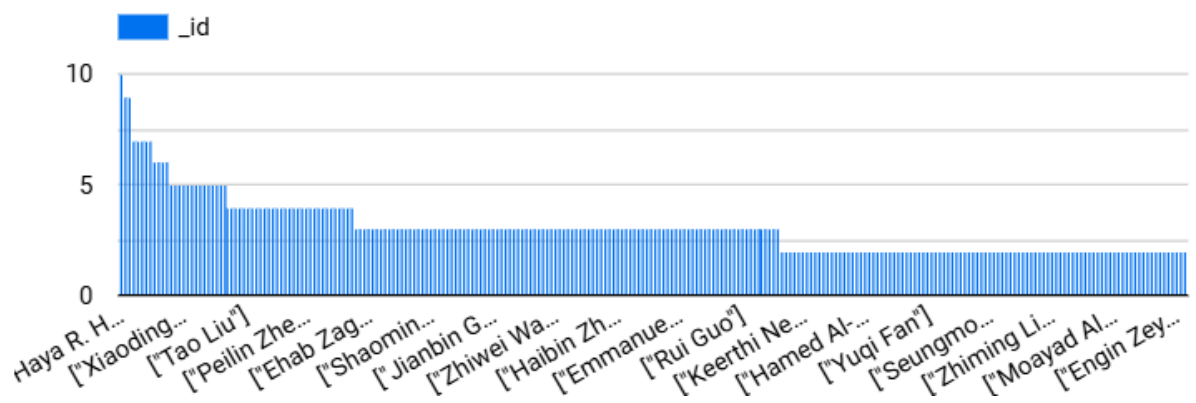
```

#### 4. Graphiques et Analyses

##### ❖ Most Prolific Authors

Ce graphique met en lumière les auteurs ayant publié le plus grand nombre d'articles dans la base de données analysée. Les auteurs sont triés par volume total de publications, ce qui permet de repérer les contributeurs clés et les experts les plus actifs dans leur domaine. Une barre ou un graphique circulaire est utilisé pour afficher ces données.

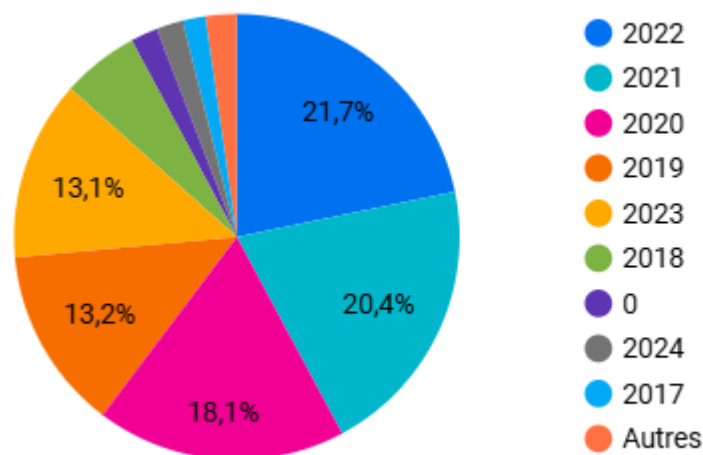
- **Insight clé :** Les 5 auteurs principaux représentent un pourcentage significatif des publications totales, soulignant leur influence dans le domaine.



##### ❖ Trends in Article Publications Over Time

Cette visualisation présente l'évolution du nombre d'articles publiés sur une période donnée. Les données sont affichées sous forme de graphique linéaire ou de histogramme, montrant les tendances annuelles ou mensuelles.

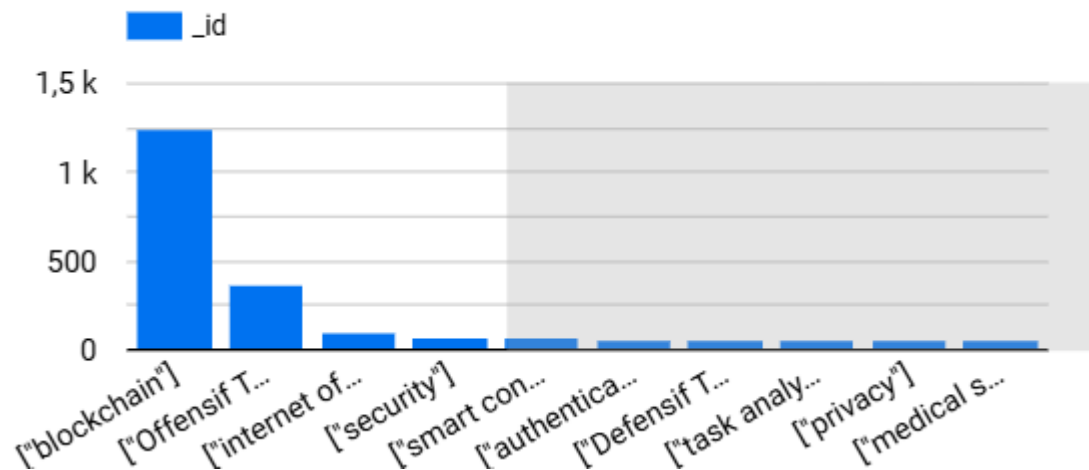
- **Insight clé :** Une augmentation marquée des publications est observée au cours des 5 dernières années, ce qui peut refléter un intérêt croissant pour le sujet ou une meilleure visibilité de la plateforme.



### ❖ Most Commonly Used Keywords

Ce graphique illustre les mots-clés les plus fréquemment utilisés dans les articles analysés. Les données sont extraites des champs de mots-clés ou des résumés des articles et sont affichées sous forme de nuage de mots ou de tableau.

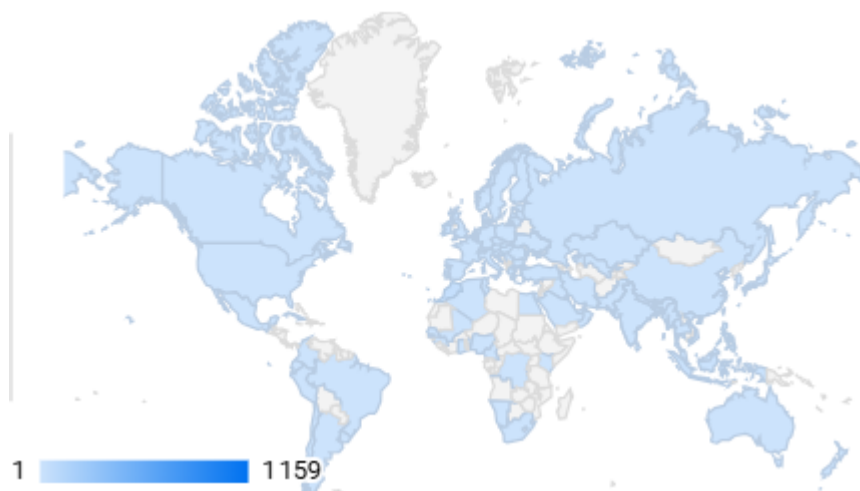
- **Insight clé** : Les termes les plus utilisés permettent d'identifier les thèmes récurrents et les sujets de recherche prioritaires.



### ❖ Geographical Distribution of Articles by Country

Ce graphique montre la répartition des articles par pays d'origine des auteurs. Les données sont représentées sur une carte interactive pour une meilleure compréhension géographique.

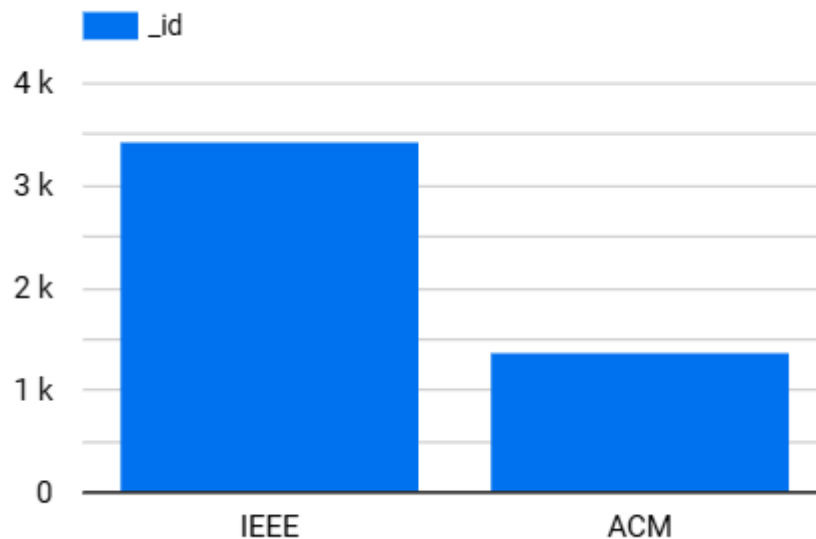
- **Insight clé** : Les pays contributeurs majeurs incluent les États-Unis, la Chine et l'Allemagne, avec une représentation significative de pays émergents.



### ❖ Distribution of Articles Across Different Journals

Ce graphique répertorie les journaux dans lesquels les articles ont été publiés, classés par nombre de publications. Une barre verticale ou horizontale est utilisée pour comparer les journaux.

- **Insight clé** : Certains journaux dominent clairement les publications, ce qui indique leur rôle central dans la diffusion des recherches.




### ❖ Universities Contributed the Most Articles

Ce graphique met en avant les universités ayant contribué le plus grand nombre d'articles. Les données sont agrégées par institution et affichées dans un tableau ou un graphique.

- **Insight clé** : Les institutions académiques de premier plan, comme le MIT et l'Université de Stanford, se distinguent par leur volume de contributions.

### Défis Rencontrés

1. **Compatibilité des Structures NoSQL** : Les données imbriquées de MongoDB ne sont pas toujours compatibles avec les outils SQL traditionnels. Une solution consistait à utiliser des pipelines d'agrégation dans MongoDB ou à les transformer directement via CData.
2. **Performance** : La taille des collections MongoDB a parfois entraîné des temps de réponse élevés. Cela a été atténué en réduisant le volume de données transférées grâce à des filtres spécifiques.
3. **Limites de l'interface CData** : Bien que CData simplifie l'accès aux données, il a fallu configurer soigneusement les paramètres pour éviter les erreurs de timeout.

universities		_id ▾
1.	['Private School']	
2.	['Abdelmalek Saadi']	
3.	['Department of Industrial and Systems Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates']	
4.	['Center for Security, Theory and Algorithmic Research, International Institute of Information Technology, Hyderabad, India']	
5.	['State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China']	
6.	['Department of Informatics, University of Oslo, Oslo, Norway']	
7.	['Department of Industrial and Systems Engineering, Khalifa University, Abu Dhabi, United Arab Emirates']	

1 - 100 / 2847 < >

## Résultats

1. **Accessibilité des Données** : Les données MongoDB sont désormais accessibles en temps réel dans Google Data Studio, permettant des mises à jour dynamiques des tableaux de bord.
2. **Visualisation Efficace** : Les visualisations créées offrent une compréhension claire des tendances des ventes, de l'engagement des utilisateurs, et des performances des produits.
3. **Facilité d'Utilisation** : Les parties prenantes peuvent désormais explorer les données directement depuis Google Data Studio, sans nécessiter d'expertise technique sur MongoDB.

## Visualisation du Tableau de Bord

*(Insérez une image du tableau de bord ici pour illustrer le résultat final. Si une capture n'est pas disponible, une description textuelle peut suffire.)*

### Exemple de Tableau de Bord :

- Un graphique à barres montre la progression des ventes mensuelles.
- Une carte interactive affiche les localisations des utilisateurs.
- Un tableau dynamique liste les performances des produits avec des filtres par catégorie.

## Conclusion et Recommandations

Le projet a démontré la faisabilité et l'efficacité de la connexion entre MongoDB et Google Data Studio via CData. Il ouvre la voie à d'autres utilisations similaires, notamment pour le suivi des indicateurs en temps réel ou la création de tableaux de bord personnalisés.