# Big Data Management - Création d’un Data Lake

## Introduction :

L’objectif de ce projet était de concevoir un « mini » lac de données (Data Lake en anglais) afin de mettre en pratique les connaissances vues en cours. Pour ce faire, nous avons utilisé Neo4J afin de modéliser les données sous la forme d’un graphe. Nous avons choisi le jeu de données : Blog Authorship Corpus disponible sur la plateforme Kaggle à l’adresse suivante : <https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus>. Il traite de blog écrit par des utilisateurs avant ou en 2004 comportant le texte de l’utilisateur et plusieurs métadonnées : Genre, Age, Sexe, topic, Signe astrologique, date. Pour agrémenter la quantité de métadonnées nous avons décidé d’ajouter un nom choisit par nos soins.

Le jeu de données étant beaucoup trop volumineux (611653 lignes), nous en avons gardé une dizaine que nous avons choisi unitairement afin d’en avoir de topics différents.

Cette sélection a permis de concevoir notre jeu de données (data.csv) que nous avons ensuite modélisé sous la forme d’un graphe dans Neo4j.

La prochaine étape a été de réalisé des traitements des textes sur Python afin d’explorer les données.

Une fois le Data Lake conçut, nous avons fait diverses expériences en Python sur nos données afin de distinguer des similitudes entre ces dernières.

L’ensemble des données et du code est accessible au repository GitHub à l’adresse suivante : [https://github.com/FeckNeck/data-lake-project](https://github.com/FeckNeck/data-lake-project%20).

## Données :

Comme décrit dans l’introduction, nous avons choisit des données de blog d’utilisateur que nous avons trouvé sur Kaggle. Nous extrait une dizaine de ligne car le fichier initial était trop volumineux ce qui nous a permis de concevoir le jeu de données suivant :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Id | Gender | Age | Topic | Sign | Date | Text |
| 3581210 | male | 33 | InvestmentBanking | Aquarius | 05/08/2004 | … |
| 3539003 | female | 14 | indUnk | Aries | 05/06/2004 | … |
| 3539003 | female | 14 | indUnk | Aries | 11/08/2004 | … |
| 4172416 | female | 25 | indUnk | Capricorn | 08/08/2004 | … |
| 4030905 | female | 17 | Student | Aries | 30/07/2004 | … |
| 3705830 | male | 25 | Non-Profit | Cancer | 23/06/2004 | … |
| 4120194 | female | 17 | Arts | Aries | 18/08/2004 | … |
| 913315 | male | 25 | Communications-Media | Aquarius | 29/06/2004 | … |
| 1877178 | male | 16 | Student | Taurus | 01/04/2004 | … |
| 1107146 | female | 16 | Student | Libra | 25/05/2003 | … |

Tableau - Dataset

Le texte étant pour certain blog top long, nous avons décidé de ne pas l’ajouter dans le tableau mais il peut etre visionné dans le fichier data.csv.

Ensuite, nous avons implementé dans Neo4j notre dataset en créant pour chaque ligne un noeud ayant pour labels les meme caracteristique que le tableau (Id, Gender etc…).

Cela nous a permis d’obtenir le graphe suivant :

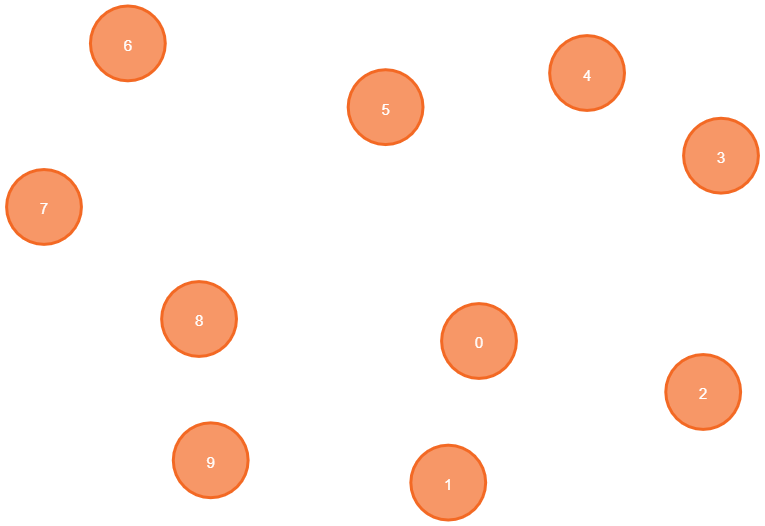


Figure - Graphe des blogs

## Métadonnées :

Titou

## Analyses :

Titou

## Conclusion :

Titou