Projet TDM
Master MIAGE – 2022-2023

L'objectif pédagogique de ce projet consiste à vous vous faire manipuler des moteurs NoSQL.

Objectif

Vous devez **installer/manipuler et analyser au minimum 2 moteurs NoSQL, basés sur le même modèle de données** (ex. 2 outils basés sur le modèle clé-valeur tels que *Redis* et *Riak*, ou 2 outils basés sur le modèle document tels que *CouchBase* ou *MongoDB* etc.).

La liste des moteurs triés par modèle est disponible sur : https://db-engines.com/en/ranking

Vous n'êtes pas obligé de travailler sur les outils présentés en cours et vous pouvez en étudier de nouveaux.

L'objectif de ce projet est de vous faire manipuler ces outils afin (1) d'expliquer comment on les utilise, (2) de les manipuler en utilisant un même jeu de données et les mêmes cas d'usage, et (3) de décrire leurs différences.

Ce projet est à réaliser, si vous le souhaitez, en binôme ou trinôme[1]. La partie analyse théorique en revanche devra être réalisée individuellement.

Choix du jeu de données

Vous pouvez : soit créer votre propre jeu de données, soit utiliser un jeu de données existant, tels que ceux cités ci-dessous, ou que vous aurez trouvé par vos propres moyens sur le Web. Le volume de données doit être si possible conséquent (voir partie cas d'usage)

Il est possible que vous ayez à adapter les données en fonction du moteur utilisé. Chaque jeu de données correspond à un cas d'usage particulier que vous saurez à spécifier.

- Données décrivant des tweets (pour *Riak*): https://github.com/dmitrizagidulin/riak-tableau
- Données portant sur des films *(*pour *AmazonDB* ou *MongoDB)* :
  - https://docs.aws.amazon.com/fr_fr/amazondynamodb/latest/developerguide/GettingStarted.NodeJs.02.html
  - http://b3d.bdpedia.fr/files/movies-mongochef.json
- Données décrivant des lignes de métro (pour *MongoDB*) :
  - http://b3d.bdpedia.fr/files/metro-lines.json
  - http://b3d.bdpedia.fr/files/metro-stops.json
- Plusieurs jeux de données JSON (particulièrement pour *MongoDB*, *CouchDB* ou *Elasticsearch*) : https://chewbii.com/json-datasets
- Jeux de données sur des gymnases et des sportifs (En BSON pour MongoDB) : https://drive.google.com/drive/folders/1RIJLBPSAmD7U0YLc4a1Rkq17j3NpRbWy
- Données gouvernementales (tapez JSON en mot-clé et un thème) :

[1] Les groupes de 3 sont interdits. Les trinômes doivent

- Générales : https://www.data.gouv.fr/en/datasets/
- Concernant l'éducation : https://data.education.gouv.fr/pages/accueil/
- Spécifiques à Paris : https://opendata.paris.fr/page/home/
- Dédiées à un département : https://opendata.hauts-de-seine.fr/

- Pour rechercher des jeux de données depuis *Google* : https://datasetsearch.research.google.com/

Cas d'usage

Vous devez étudier les cas d'usage sur votre jeu de données en considérant deux points de vus distincts :

- Point de vue Utilisateur standard : Vous devez pour cela définir, en langage courant, 4 types d'interrogations sur votre jeu de données. On estimera que celles-ci sont effectuées très fréquemment.
- Point de vue Analyste de données : Vous devez pour cela définir, en langage courant, 2 types d'interrogations complexes sur votre jeu de données (agrégation, transformation, calcul complexe).

Ces requêtes devront par la suite traduite dans le langage propre aux moteurs que vous avez choisis d'étudier.

Pour tester le temps d'exécution des requêtes, vous devrez (en dupliquant et en découpant votre jeu de données initial) les tester sur des jeux de données de taille variable (ex. des paquets de 10, 100, 1000, 5000 … éléments).

Analyse théorique (à réaliser individuellement – un moteur/personne)

Pour chaque moteur NoSQL vous devrez analyser et décrire les aspects concernant :

- L'architecture (maître/esclave, pair à pair …)
- La gestion de la réplication
- La gestion des transactions et de reprise sur panne
- La gestion de la cohérence, de la disponibilité et de la tolérance au partitionnement

Travail à réaliser

Pour ce projet vous devez rédiger un **rapport[2] (20 pages maximum)** expliquant :

- La motivation de vos choix des 2 moteurs à comparer.
- Les étapes à suivre pour pouvoir installer et/ou utiliser chaque moteur.
- La structure de votre jeu de donnée et la description des cas d'usages.
- Les étapes de création ou de chargement de votre jeu de données dans chaque moteur**.**
- L'expression de vos requêtes pour chaque moteur, le résultat obtenu et leurs temps d'exécution pour des jeux de données de différentes tailles.
- Une comparaison des 2 moteurs en termes d'installation/utilisation par rapport à votre jeu de données et de vos cas d'usage.
- L'analyse théorique de chaque moteur (**à réaliser de manière individuelle** – un moteur par personne) et un comparatif de ces analyses.
- Une conclusion indiquant les difficultés rencontrées, la répartition du travail entre les membres du groupe et le temps passé sur ce projet.

Modalités

Le projet est **à réaliser en binôme ou trinôme[3]**.

La **date de remise du projet** est le **04 décembre 2022**.

Un fichier .zip (dont le nom indiquera les noms de familles des membres du groupe) devra être **déposé sur *CEL* avant le 04/12/2022.** Ce fichier devra contenir :

- La version électronique du rapport.
- Le jeu de données si ce dernier est créé par vos soins. Si vous utilisez un jeu existant, l'adresse où récupérer ce jeu de données devra être précisée dans le document.

La présentation du projet est le 07 décembre 2022.

[2] Rappel : tout rapport doit comporter une introduction et une conclusion.
[3] Aucun groupe de plus de 3 membres n'est autorisé.