M2 MALIA – Année 2023-2024
Big data management
Projet Lac de données
J. Darmont (https://eric.univ-lyon2.fr/jdarmont/), 18/01/24

Objectif : Mettre en œuvre un mini-lac de données textuelles (travail en **binômes**)

1. Données

- Choisir un corpus de textes, par exemple sur la plateforme Kaggle ou encore dans la liste de corpus de Wikipedia. Si des données ou des métadonnées structurées ou semi-structurées y sont associées, les inclure dans le lac.
- Prévoir un mode de stockage pour les données (système de fichiers d'un ordinateur, HDFS…) et les stocker effectivement.

2. Métadonnées

- Définir un modèle de métadonnées simple (métadonnées intra et inter-objets ou goldMEDAL, par exemple) qui vous sera nécessaire pour interroger le lac de données textuelles.
- Prévoir un mode de stockage pour les métadonnées, par exemple OpenMetadata, Apache Atlas, Neo4J, MongoDB ou même un SGBD relationnel.
- Instancier les métadonnées dans les outils de stockage définis à l'étape précédente. Apache Tika peut aider pour les métadonnées intra-objet.
- Les métadonnées globales doivent au moins inclure un index inversé des termes des documents textuels. Il peut être généré par exemple à l'aide d'Elasticsearch ou de Solr.

3. Analyses

- Nettoyer/transformer les données si nécessaire, tout en conservant les données sources. Inclure les résultats et les transformations effectuées dans les métadonnées.
- Proposez des analyses relatives aux données textuelles du lac. Les outils de *Business Intelligence* peuvent être employés, ainsi que tous les outils utiles en *data science*.

4. Rendu

Rapport synthétique :

- Introduction/présentation des données
- Description des étapes 1 à 3
- Conclusion, problèmes rencontrés, perspectives
- Code en annexe si nécessaire

Rapport à rendre le 09/02/2024 sur Moodle : https://moodle.univ-lyon2.fr/course/view.php?id=13225