CS for Big Data
Christine Largouët
UP Informatique

Sujets de projets
Computer Science for Big Data

Préambule

Les sujets de projets proposés sont de différentes natures mais partagent tous les contraintes suivantes pour la restitution :

1. Les expérimentations doivent être réalisées à partir d'un jeu de données propre à chaque groupe. Le jeu de données choisi devra être adapté à la technologie Big Data choisie pour le projet (cette sélection fait partie du travail).
2. Des traitements spécifiques au Big Data doivent être appliqués comme des requêtes simples pour illustrer les principes mais pas uniquement. Sont également attendus des traitements de type : agrégations, map/reduce ou recherche de chemins selon le type de base ou autre selon les technologies employées.
3. Lorsqu'il s'agit de comparer différentes technologies, des analyses qualitatives (expressivité du langage de manipulation des données, interprétabilité des résultats, etc...) mais également quantitatives (lorsque c'est possible, temps de calcul ou autre) sont attendues.
4. Les traitements seront réalisés si cela est possible dans un programme Python. Si les problématiques traitées par les requêtes ainsi que les analyses s'y pretent, une visualisation graphique (matplotlib, seaborn) sera proposée.
5. Afin d'augmenter le bénéfice de ce projet, chaque groupe sera en charge d'évaluer le travail d'un autre groupe et de fournir une synthèse de l'évaluation (qui sera comptabilisée dans la notation du groupe évaluateur).

Travail Attendu :

On vous demande de présenter la technologie que vous aller tester par l'intermédiaire d'un jeu de données adapté et d'un ensemble de requêtes appropriées permettant de tester les avantages et les limites de cette solution.

Tous les documents textuels retournés en dehors des programmes seront en pdf (et non en word).

Le travail attendu est le suivant :

1. **Un rapport (en pdf) :** (etude et analyse)
   - Une présentation du SGDB choisi (objectifs et approche, type de données, type de langage de requêtes, etc) ainsi qu'un rapide manuel d'utilisation (installation, commandes de base, tout e qui vous parait utile)
   - Une présentation de votre jeu de données et de votre approche permettant de tester cette solution. Vous pourrez souligner votre raisonnement ou l'intuition qui vous amène à essayer telle requête, tel traitement en comparaison par exemple aux solutions vues dans le cours Big Data. 
   - Une partie plus technique donnera toutes les instructions (installation, précautions éventuelles, etc.) permettant de tester votre solution par vos camarades et par moi-même.
   - Une conclusion synthétisant votre travail ainsi que votre évaluation personnelle de la solution

2. **Le code :**
   - Un code en Python permettant de tester votre travail et un lien sur le jeu de données si celui-ci est volumineux. Le programme doit être correctement commenté, sans référence à vos propres répertoires afin de pouvoir être testé et analysé par vos camarades et moi-même/

3. **Une évaluation sur le sujet XXX (pdf) :**
   - Vous rédigerez un texte de quelques lignes pour chacune des questions suivantes :
     - Décrivez le travail proposé par le groupe XXX sur le sujet XXX
     - Quel est l'intérêt de la solution proposée et est-elle bien illustrée dans ce travail ?
     - Le jeu de données parait-il adapté à l'évaluation de la solution, commentez ?

- Les requêtes permettent-elles à votre avis de tester toutes les facettes de la solution ? Qu'est-ce qui aurait pu être envisagé de plus à votre avis ?
- Avez-vous pu lancer toutes les expérimentations (installation du SGBD et exécution du code) ? Si oui ou non commentez cet aspect pratique ?
- Le code proposé est-il clair et compréhensible ? Est-il efficace ? Correspond-il à vos standards de programmation ?
- L'analyse de la solution, l'évaluation de la méthode, la comparaison par rapport aux autres approches est-elle convaincante ? Commentez

Les sujets :

Sujet N°1 :

Etude et Analyse de la base de données NoSQL orientée colonne, **`Cassandra`**.

Sujet N°2 :

Etude et Analyse du système de gestion de base de données clé-valeur hautes performances **`Redis`**.

Sujet N°3 :

Etude et Analyse de la base de données proposée par AWS (Amazon Web Service) **`DynamoDB`** (25 Go de stockage gratuit, nécessité de créer un compte AWS)

https://aws.amazon.com/fr/free/?all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc

Sujet N°4 :

Etude et Analyse de la base de données orientée document **`CouchDB`**. Comparaison avec MongoDB.

Sujet N°5 :

Etude et Analyse de la base de données orientée document **`Couchbase`**. Comparaison avec MongoDB.

Sujet N°6:

Etude et Analyse de la base de données orientée graph **`OrientDB`**. Comparaison avec Neo4J.