Projet Traitement et Données Large Échelle

Vous pouvez réaliser ce projet seul ou en binôme.

- Comparaison de système : constituer une base de données avec un système NoSQL disponible. Évaluer les performances du système pour les opérations d'insertions et mises à jour.

- Étude : présenter, expliquer et approfondir un des aspects spécifiques du système NoSQL choisi.
  Cela peut être :
  - Une comparaison des performances des requêtes pour un système NoSQL avec un système relationnel (MySQL, postgreSQL, Oracle par exemple).
  - Comparer les performances avec ou sans utilisation d'index adaptés.
  - Évaluer l'impact en termes de performance de la réplication ou de la fragmentation.

Organisation du projet

Il vous sera demandé un notebook présentant le comparatif entre les vitesses d'exécutions des commandes CRUD pour le système relationnel et le système NoSQL.

Il faudra indiquer pour le 31 octobre si vous travaillez seul ou en binôme, ainsi que le système NoSQL choisi.

Le projet devra être rendu pour le 16 décembre 2024.

Choix du schéma

Pour effectuer vos comparaisons de performances, il faudra définir un schéma pour les données.
Vous pouvez choisir un contexte quelconque, partir d'une modélisation d'un problème réel, ou de ressources Open Data.

Choix du système NoSQL

Vous devez choisir un système NoSQL pour stocker vos données, et étudier un des aspects de ce système.

Vous pouvez consulter les liens listés dans les comparateurs de SGBD pour trouver des idées de système à étudier.

Quelques exemples de systèmes NoSQL : CouchDB, RethinkDB, CouchBase, Cassandra, MonetDB, HBase, BerkeleyDB, Voldemort, SolR, Ryak et tant d'autres.

Quelques exemples d'aspect à étudier particulièrement : langage de recherche, indexation interne, support de la concurrence d'accès, architecture, technique de distribution, réplication / fragmentation reprise sur panne, etc.

Benchmarking

Effectuer des comparaisons des performances pour un grand nombre d'insertion, de mises à jour et de requête entre le système NoSQL et le système relationnel choisi.

Pour le système NoSQL, essayer d'améliorer les performances de requêtes initialement lourdes choisies : création d'index, modification du stockage, modification des paramètres de durabilité et cohérence.

Tester l'impact sur les performances de la gestion de répliques.

Comparaison de SGBD

Vous pouvez consulter une liste de différents systèmes de gestion de bases de données, classée par type (relationnel, key-value store, bases de données documents, …) et par popularité à l'adresse suivante :

- https://db-engines.com/en/ranking

Open Data

L'Open Data correspond à la mise à disposition publique des données afin de favoriser leur diffusion et la construction d'applications les utilisant.

Les données peuvent être fournies sous différents formats (JSON, excel, accès BD). Ci-dessous, quelques sites d'exemple. Vous pouvez récupérer quelques documents, et les utiliser dans l'application test.

- http://data.enseignementsup-recherche.gouv.fr/
- https://data.strasbourg.eu/pages/accueil/
- https://www.data.gouv.fr/fr/

Produire un jeu de documents volumineux

Pour tester un système avec un nombre quelconque d'insertions, de mises à jour et de requêtes, il est possible d'utiliser des outils permettant de générer des jeux de données de taille paramétrable. Ci-dessous, quelques outils.

- http://generatedata.com/
- https://www.mockaroo.com/
- https://github.com/10gen-labs/ipsum

Critères de notations

- Notebook : 2 points
- Fonctionnement correct : 2 points
- Présentation du système NoSQL choisi : 2 points
- Avantages/inconvénients du système NoSQL choisi : 2 points
- Comparaison sur requêtes de sélections : 4 points
- Comparaison sur requêtes de mise à jour (CUD) : 3 points
- Graphiques (choix du diagrammes, incertitudes, etc.) : 1 points
- Performance avec / sans index : 2 points
- Prise en compte de la répartition des données : 2 points