Responsable UE : Tristan Allard.
Enseignants CM et TD : Tristan Allard, Tassadit Bouadi.
Enseignants TP : Khaled Arsalane, Louis Béziaud, Thomas Bouvier.
Contact : `prénom.nom@irisa.fr`

1 Objectifs des séances de TP

Les séances de TP de l'UE de bases de données ont pour objectif de réaliser une mise-en-oeuvre pratique des techniques vues en séances de CM et TD. Il est demandé aux étudiant·es de répondre en binôme à un cas d'usage concret de gestion de données. Plusieurs livrables jalonnent le module, code et rapports.

2 Cas d'usage

Vous venez de passer l'entretien de recrutement d'une entreprise spécialisée en analyse de données. La dernière étape du processus de recrutement est de démontrer vos compétences sur un cas d'usage concret. Vous avez un fichier CSV contenant plusieurs giga-octets de données, un ensemble de requêtes spécifiées en langage naturel, et des scénarios d'exécution. Votre objectif est de comparer les performances de deux SGBDs en faisant en sorte, sur chacun des SGBDs, que l'exécution des requêtes soit la plus efficace possible.

2.1 SGBDs considérés

Les deux SGBDs considérés sont les suivants :

1. PostgreSQL : Un SGBD *client/serveur* installé localement.
2. DuckDB : un SGBD *in-process* aussi installé localement.

2.2 Données

Thème Déplacements en taxis jaunes à New York en 2013

Téléchargement Ici : `https://archive.org/download/nycTaxiTripData2013/trip_data.7z`

Composition Une archive contenant 12 fichiers CSV. Vous commencerez par vous focaliser sur le premier fichier : trip_data_1.csv (2,46 Go décompressé). Vous pourrez augmenter progressivement le nombre de fichiers stockés.

2.3 Python Notebook

Vous travaillerez dans un notebook Python à partir duquel vous connecterez à PostgreSQL (ou vous utiliserez DuckDB) et vous effectuerez vos expérimentations. Vous pourrez aussi utiliser les bibliothèques graphiques disponibles en Python pour tracer vos graphes.

2.4 Requêtes

Restrictions :

R.0 : Les trajets effectués par le chauffeur dont le `medallion` est `0B57B9633A2FECD3D3B1944AFC7471CF`.

R.1 : Les trajets pour lesquels `passenger_count` est supérieur ou égal à 2.

R.2 : Les trajets pour lesquels `passenger_count` est égal à 1.

Agrégats :

A.0 : La distance totale cumulée effectuée par l'ensemble des trajets en taxi.

A.1 : Pour chaque medallion, la distance moyenne effectuée pour les trajets avec un unique passager.

A.2 : Pour chaque jour, le nombre de passagers moyen.

A.3 : Pour chaque chauffeur dont la distance moyenne parcourue est supérieur à la moyenne des distances, le nombre moyen de passagers par trajet.

3 Vos solutions

Chacune de vos solutions est le résultat d'un ensemble de choix de mise-en-oeuvre. Vos choix pourront porter sur les points suivants :

- Indexes variés,
- Organisations des données sur disque (hachées, triées),
- Matérialisation de vues,
- Procédures stockées,
- Syntaxe requête,
- *etc.* (à vous de proposer !)

4 Attendus

Il est attendu de vous une étude expérimentale des différentes étapes de votre raisonnement :

- **Des explications détaillées** sur vos solutions finales (syntaxe de vos requêtes et choix) mais aussi sur l'histoire qui vous a permis d'y arriver (vos étapes principales, les difficultés rencontrées, vos solutions).
- **Des graphes** qui montrent : (1) les mesures de performance de vos solutions à différentes étapes, (2) la façon dont vos solutions se comportent quand la taille des données augmente ou la quantité de `RAM` allouée au SGBD croît. La configuration de votre système (*e.g.,* `CPU`, support secondaire `SSD` ou `HD`, *etc*) et les paramètres de vos expérimentations (*e.g.,* `RAM` allouée, taille des données, *etc*) devront être détaillés.
- **Les plans d'exécution appliquées par votre SGBD** qui expliquent les mesures de performance obtenues. (En général tout ce qui peut expliquer les performances est bienvenu.)

Plus précisément vos expérimentations pourront être les suivantes :

Paramètres : Taille des données, **espace disponible en `RAM`**, nombre d'utilisateur·trices, *vos choix (cf Section 3).*

Mesures : Sur 10 exécutions par requête au total. Vous afficherez le temps d'exécution de la première exécution ainsi que le temps d'exécution moyen/min/max des 9 exécutions suivantes

Scénario mono-requête : Vous optimisez chaque requête indépendamment des autres (*i.e.,* la BD n'est optimisée que pour la requête en cours d'étude)

5 Exemple de calendrier

Pour illustration vous trouverez ci-dessous un exemple de calendrier des séances. La description est séquentielle mais en pratique le travail sur les différentes parties est entrelacé.

Partie 1 – environnement de travail (2 séances) :

- Téléchargement et prise en main des données (format `CSV`),
- Conception et implantation SQL du schéma conceptuel,
- Insertion d'un tout petit sous-ensemble de lignes dans la BD (*e.g.,* 10 lignes ?)
- Installation des SGBDs en local [1] et éventuellement d'une GUI (PostgreSQL, Pgadmin, DuckDB).

Partie 2 – Requêtes (2 séances et travail à la maison) :

- Écriture des requêtes en SQL.
- Vérifications de la correction de vos requêtes sur votre toute petite BD.
- Lecture et analyse des plans d'exécution.

Partie 3 – Plate-forme expérimentale (2 séances et travail à la maison) :

1. Postgresql est aussi installé à l'université.

- Définition des mesures qui vous intéressent (par ex, temps d'exécution, RAM utilisée), de la façon donc vous allez les faire, de l'information dont vous aurez besoin pour les traiter ensuite (par ex, id requête, timestamp, temps d'exécution), et de la manière dont vous allez les stocker (par ex, fichier CSV, json, en base de données).
- Développement de votre code permettant de lancer vos expérimentations, de faire vos mesures, et de les enregistrer.

Partie 4 – Visualisation de vos mesures (2 séances et travail à la maison) :

- Choix de votre environnement de visualisation (par ex, python, tableur).
- Conception de votre méthode pour tracer vos graphes.

Partie 5 – Comparaison et Optimisation (2 séances et travail à la maison) :

- Exécution de vos expérimentations.
- Comparaison des graphes des deux SGBD en compétition.
- Explications via l'analyse des plans d'exécution et hypothèse.
- Proposition d'amélioration des performances (et retour à l'exécution des expérimentations).

6 Modalités pratiques

Groupe : Binômes (sauf pour un groupe si le nombre de participant·es n'est pas un multiple de 2).

Matériel : Ordinateur personnel recommandé. Possibilité d'utiliser les postes accessibles à l'université.

Logiciels (a minima) : PostgreSQL, PGAdmin, DuckDB.

Langage : SQL et Python.

Livrables : Vous aurez à livrer les éléments suivants (envoi via Moodle) :

1. **Rapport mi-parcours - 03/11/2023 :** Rapport (5p max hors page de garde et table des matières, interligne simple, police de type Times New Roman taille 11) présentant l'état courant de votre travail, et discutant du travail restant et de l'organisation que vous adopterez pour le réaliser.
2. **Rapport terminal - 22/12/2023 :** Rapport (20p max hors page de garde et table des matières, interligne simple, police de type Times New Roman taille 11) présentant vos propositions finales (requêtes, stratégies d'amélioration des performances, avantages et inconvénients), les graphes demandés ainsi que la configuration et les paramètres à partir desquels ils ont été obtenus, les plans d'exécution correspondant, les analyses de graphes, et décrivant des pistes futures d'amélioration de vos résultats.
3. **Code - 22/12/2023 :** Code commenté (code de préparation des données, notebooks de pilotage des expérimentations, requêtes SQL incluant notamment la création de votre schéma) avec **fichier README** à la racine.

Calendrier : De fin septembre (première séance de TP) à fin novembre. 10 séances de TP encadrées au total, **travail personnel entre chaque séance indispensable**.