TP #01: Import des données IMDb et premières requêtes

Objectifs du TP

- Configurer l'environnement de base de données
- Importer les jeux de données IMDb
- Exécuter et analyser les premières requêtes
- Établir des mesures de performance de référence

Étapes détaillées

1. Préparation de l'environnement

- Installation/configuration du SGBD PostgreSQL (recommandé dans un environnement Docker avec pgAdmin)
    - Cloner le dépot GitHub : https://github.com/akaclasses/optimisation-sql
    - http://localhost:5050/
    - Login : admin@admin.com / adminpass
    - Créer un nouveau serveur:
        - Nom : OptimisationSQL
        - Connection / hostname : postgres
        - username : admin
        - password : adminpass
- Création d'une base de données dédiée (`imdb_clone` dans le Docker Compose fournit)

2. Téléchargement et préparation des données

- Téléchargement des fichiers TSV compressés depuis https://developer.imdb.com/non-commercial-datasets/
- Vérification de la structure et échantillonnage
- Déposer les fichiers dans le répertoire import du Docker Compose

⚠️ Attention à l’espace disque nécessaire ⚠️

L’import complet prend environ 35 Go

Auquel il faudra ajouter ensuite les index, etc…

3. Création du schéma

```
CREATE TABLE title_basics (
    tconst VARCHAR(12) PRIMARY KEY,
    title_type VARCHAR(20),
    primary_title VARCHAR(500),
    original_title VARCHAR(500),
    is_adult BOOLEAN,
    start_year INTEGER,
    end_year INTEGER,
    runtime_minutes INTEGER,
    genres VARCHAR(100)
);

-- Structures similaires pour les autres tables
```

4. Import des données

```
COPY title_basics FROM PROGRAM 'zcat /import/title.basics.tsv.gz'
WITH (FORMAT csv, DELIMITER E'\t', HEADER, NULL '\N', QUOTE E'\001');

-- Répéter pour les autres tables
```

5. Premières requêtes simples

```
-- Comptage simple
SELECT COUNT(*) FROM title_basics;

-- Distribution par type
SELECT title_type, COUNT(*)
FROM title_basics
GROUP BY title_type
ORDER BY COUNT(*) DESC;

-- Films les mieux notés
SELECT b.primary_title, r.average_rating, r.num_votes
FROM title_basics b
JOIN title_ratings r ON b.tconst = r.tconst
WHERE b.title_type = 'movie'
ORDER BY r.average_rating DESC
LIMIT 10;
```

6. Introduction aux plans d'exécution

```
EXPLAIN ANALYZE
SELECT b.primary_title, r.average_rating
FROM title_basics b
JOIN title_ratings r ON b.tconst = r.tconst
WHERE r.num_votes > 1000
ORDER BY r.average_rating DESC
LIMIT 10;
```