TP : Analyse des tendances musicales

Le but sera de rédiger un rapport qui pourra être utilisé comme structure pour votre prochain projet ou DP. Pensez à rajouter tous les éléments nécessaires lors de la rédaction pour une bonne compréhension.
Exemple : Lexique, Glossaire, Schémas, capture de code, etc.

📌 *Il n'est pas nécéssaire de mettre tout le code dans le rapport, mais de bien expliquer étape par étape. Exemple : utilisation d'une structure minimale.*

Objectif :

Ce TP a pour objectif d'analyser la relation entre les caractéristiques musicales (danceability et énergie) et la popularité (streams).

- **Mapper** :
    - Extraire le pourcentage de danceability (`danceability_%`), le pourcentage d'énergie (`energy_%`) et le nombre de streams (`streams`) pour chaque chanson.
- **Reducer** :
    - Agréger les données par combinaison de `danceability_%` et `energy_%` (clé composite).
    - Calculer la somme totale des streams pour chaque combinaison, ainsi que le nombre d'occurrences de chaque combinaison afin de calculer la moyenne des streams pour chaque combinaison de danceability et d'énergie.
    - Affichez le résultat pour le stocker dans HDFS, mais aussi, stockez les résultats dans HBase.
- **Visualisation** :
    - Créer un graphique avec Matplotlib pour visualiser les tendances des 10 meilleures combinaisons. (Un script python à part qui récupère les données depuis Hbase pour les mettre en forme).

Vous utiliserez **Hadoop MapReduce** pour traiter les données, **HappyBase** pour stocker les résultats dans **HBase**, et **Matplotlib** pour visualiser les résultats sous forme de graphique. Il est possible d'utiliser

Livrables attendus :

1. **Le code source** de votre job MapReduce.
2. **Le graphique généré** avec Matplotlib (au format PDF).
3. **Les données** stockées dans HBase (screenshot) et HDFS (part-00000).
4. **Un rapport détaillé** expliquant chaque étape du processus, les résultats obtenus, et une analyse des relations observées entre la danceability et les streams.

Étapes du TP :

1. Connexion à la machine virtuelle

- Indiquez comment vous vous connectez à la machine virtuelle avec l'IP, le port, le nom d'utilisateur et le mot de passe.
- Décrivez les étapes pour établir la connexion SSH.

2. Lancer les conteneurs et services Hadoop

- Expliquez les services nécessaires pour Hadoop (HDFS, YARN, Zookeeper, HBase).
- Indiquez les commandes nécessaires pour démarrer ces services sur la machine virtuelle.

3. Importer les données dans HDFS

- Décrivez la manière dont vous allez importer les données pour qu'elles puissent être traitées avec MapReduce.
- Précisez où stocker les fichiers.

4. Créer et exécuter un job MapReduce

- Expliquez la structure d'un job MapReduce (mapper, reducer).
- Donnez les commandes à exécuter pour soumettre le job sur Hadoop.

5. Visualiser les résultats

- Décrivez comment récupérer les résultats que se soit MapReduce ou Hbase.
- Indiquez comment trouver les données générées comme avec **Matplotlib**, **Pandas**...

6. Récupérer les résultats

- Expliquez comment récupérer les fichiers de sortie du job MapReduce depuis HDFS pour le mettre sur la partie linux de votre container hadoop-master.
- Donnez la procédure pour récupérer les données du container jusque son pc local.