Exercices MapReduce

Ces exercices vous permettront d'approfondir l'utilisation du processus MapReduce avec Hadoop, en utilisant des données réelles provenant de Spotify.

- Vous passerez aussi par l'utilisation de `Happybase`, une librairie Python pour interagir avec HBase, et de `Matplotlib` pour visualiser les résultats.
- Pour exporter les visuels, vous utiliserez la librairie `matplotlib.backends.backend_pdf`, pour exporter les graphiques en format PDF.
- pandas pourra être utilisé pour manipuler les données plus facilement et efficacement.

Données

Le Dataset Spotify_Most_Streamed_Songs.csv provient de kaggle. Un site d'entrainement de cours et de défis pour le Big Data et IA comportant des datasets divers et variés.

Plus d'information sur les données sur le site : https://www.kaggle.com/datasets/abdulszz/spotify-most-streamed-songs

Questions

Pour réaliser ces différentes questions, vous devrez utiliser le comportement MapReduce de Hadoop grâce à ces deux fichiers :

- le mapper, pour filtrer les données
- le reducer, pour aggréger et exporter les données

Pour simuler le comportement du MapReduce en local avec python, vous pouvez utiliser cette commande :

- Windows :

```
type fichier.txt | python mapper.py | sort | python reducer.py
```

- Linux/MacOS :

```
cat fichier.txt | python3 mapper.py | sort | python3 reducer.py
```

Exercice 1 : Nombre total de streams par année de sortie

Objectif : Calculer le nombre total de streams par année.

- **Mapper** : Lire le fichier CSV, extraire l'année de sortie (`released_year`) et le nombre de streams (`streams`).
- **Reducer** : Agréger le nombre total de streams par année.

Exercice 2 : Nombre moyen de playlists Spotify par année de sortie

Objectif : Calculer la moyenne du nombre de playlists Spotify contenant des chansons, regroupées par année de sortie.

- **Mapper** : Extraire l'année de sortie (`released_year`) et le nombre de playlists Spotify (`in_spotify_playlists`).
- **Reducer** : Calculer la moyenne du nombre de playlists par année.

Exercice 3 : Statistiques visuelles des streams par année

Objectif : Créer un graphique des streams par année, exporté en PDF.

- **Mapper** : Comme dans l'exercice 1, extraire l'année de sortie (`released_year`) et le nombre de streams (`streams`).
- **Reducer** : Agréger le nombre total de streams par année, créer un graphique matplotlib (pie) pour visualer les tendances et enfin, l'exporter en PDF.

Exercice 4 : Insertion dans HBase des streams par année

Objectif : Utiliser MapReduce pour insérer les résultats dans HBase.

- **Mapper** : Comme dans l'exercice 1, extraire l'année de sortie (`released_year`) et le nombre de streams (`streams`).
- **Reducer** : Agréger le nombre total de streams par année, puis insérer ces données agrégées dans une table HBase

Exercice 5 : Statistiques visuelles avec stockage HBase

Objectif : Combiner les concepts de MapReduce, HappyBase et Matplotlib.

- **Mapper** : Extraire le nom de l'artiste (`artist(s)_name`) et le nombre de streams (`streams`).
- **Reducer** : Agréger le nombre total de streams par artiste, puis insérer les résultats dans HBase.
  Récupérer les données et créer un graphique des streams par artiste, exporté en format PDF.