Projet Hadoop-MovieLens

Author : **Axel Delille**
Date : 17/17/2024

Pour répondre au projet, j'ai procédé en 3 étapes que je vais décrire ici.

1 - Détection du film favori de chaque utilisateur

Concrètement dans le ratings.csv je fais un map reduce qui va un WritableCustom pour avoir le movieid et la note permet dans le reduce de n'output que le user avec son movieId favori.

Cette partie est faite dans le job fr.etu.polytech.HighestRatedMoviePerUserId qui prends en entrée dans l'ordre :

- Le path sur hdfs de ratings.csv
- Le path de sortie sur hdfs que l'on nommera favoriteMoviePerUser.

2 - Mapping des titres de film au nombre d'utilisateur qui le préfère aux autres.

Ici on prends en entrée favoriteMoviePerUser et movies.csv. On a 2 mappers :

- Celui de favoriteMoviePerUser va retourner des tupes (movieId, "user:"+userId)
- Celui de movies.csv qui va retourner des tuples (moviesId, "title:"+movieTitle)

Dans le reducer on va donc avoir si les csv sont bien formaté pour chaque movieId quelque chose de la forme :

```
["title:Star Wars", "user:1", "user:2", ... ]
```

L'ordre des éléments n'est pas réellement important ici, le fait est que l'on a le titre ainsi que la liste des utilisateurs qui ont le film en favori donc on peut retourner à la fin le titre du film accompagné du nombre d'utilisateur qui l'ont préféré.

Cette partie est effectuée dans le job fr.etu.polytech.NumberOfLikePerMovie qui prends en entrée dans l'ordre :

- Le path sur hdfs de favoriteMoviePerUser
- Le path sur hdfs de movies.csv
- Le path de sortie sur hdfs que l'on nommera numberOfLikePerMovie

3 - Agrégation des données pour reproduire l'équivalent d'un group by sur le nombre de personne qui ont préféré chaque film.

Ici on prends juste en entrée numberOfLikePerMovie, le map va juste servir de passe plat où on va mettre le nombre de personne qui ont préféré le film en key et le titre du film en value.

Les key sont traité par ordre croissant par hadoop par défaut ce qui règle notre problème de faire un group by par ordre croissant.

On a alors le nombre de like et une liste de titre de film qui ont eu ce nombre de like.

Il suffit donc de retourner le nombre de like et la concaténation des titres de films séparé par un espace chacun.

Cette partie est effectuée dans le job fr.etu.polytech.movielens.GroupByNumberOfLikePerMovie qui prends en entrée dans l'ordre :

- Le path sur hdfs de numberOfLikePerMovie
- Le path de sortie sur hdfs du résultat

En complément des informations sur le projet donné ci-dessus, il contient à sa racine un README.md possédant un exemple d'utilisation du jar.