Epreuve Contrôle : Mini-Projet MapReduce

Instructions :

Ce mini-projet entre dans le cadre de l'épreuve finale de l'élément de module 'Big Data Labs'. C'est un projet à réaliser **maximum** en binôme et à remettre avant le **12 Décembre 2023**.

Sur le cours Google Classroom, merci de remettre un fichier Zip contenant un rapport avec le nom du binôme (ou monôme) avec les scripts associés.

Veuillez bien lire l'énoncé et les questions demandées avant de déposer une réponse finale.

Tous les fichiers de données sont fournis avec l'épreuve.

Enoncé :

PoliTV est une société internationale de streaming en ligne axée sur les films. Ses utilisateurs peuvent regarder les films disponibles via des téléviseurs intelligents, des PC ou des appareils mobiles. Les gestionnaires de PoliTV souhaitent calculer un ensemble de statistiques spécifiques basées sur les ensembles/fichiers de données d'entrée suivants.

- **Users.txt**

  Il s'agit d'un gros fichier texte contenant la liste des utilisateurs enregistrés de PoliTV. PoliTV compte des millions d'utilisateurs, c'est-à-dire que Users.txt compte des millions de lignes. Chaque ligne du fichier Users.txt est associée au profil d'un utilisateur et a le format suivant :

  Username, Gender, YearOfBirth, Country

  Ex : PaoloG76, Male, 1976, Italy

- **Movies.txt**

  Movies.txt est un gros fichier texte contenant le catalogue des films disponibles (plus de 100 000 films). Chaque ligne de Movies.txt est associée à un film et a le format suivant :

  MID, Title, Director, ReleaseDate

  Ex : MID124,Ghostbusters,Ivan Reitman,1984/05/01

- **WatchedMovies.txt**

  WatchedMovies.txt est un gros fichier texte. Chaque fois qu'un utilisateur regarde un film, une nouvelle ligne est ajoutée à la fin de WatchedMovies.txt. Les données collectées au cours des 10 dernières années sont actuellement stockées dans WatchedMovies.txt, c'est-à-dire que WatchedMovies.txt contient des milliards de lignes. Chaque ligne de WatchedMovies.txt a le format suivant :

  Username, MID, StartTimestamp, EndTimestamp

  Ex : PaoloG76,MID124,2010/06/01_14:18,2010/06/01_16:10

Travail à faire

Les responsables de PoliTV souhaitent effectuer des analyses sur les films regardés.

On doit développer une seule application pour traiter toutes les analyses qui les intéressent. L'application comporte cinq arguments : les trois fichiers d'entrée Users.txt, Movies.txt et WatchedMovies.txt et deux dossiers de sortie, "outPart1/" et "outPart2/" qui sont associés aux sorties des points 1 et 2 suivants, respectivement.

1. Hadoop MapReduce

On doit concevoir une application unique, basée sur Hadoop MapReduce, et écrire le code **Java** correspondant, pour répondre au point suivant :

Films regardés par un seul utilisateur au cours de l'année 2019.

L'application prend en compte uniquement les visualisations liées à l'année 2019 (c'est-à-dire les lignes de WatchedMovies avec StartTimestamp dans la plage du 1er janvier 2019 au 31 décembre 2019) et sélectionne les films qui ont été regardés par un seul utilisateur en 2019. On souhaite stocker les identifiants (MID) des films sélectionnés dans le dossier de sortie HDFS (un MID par ligne).

Note. **Si un film a été regardé plusieurs fois en 2019 mais toujours par le même utilisateur, ce film satisfait à la contrainte et doit être sélectionné.**

2. Spark et RDD

On doit concevoir une application unique, basée sur Spark, et écrire le code **Python** correspondant, pour aborder les points suivants :

Films qui ont été regardés fréquemment mais seulement pendant un an au cours des cinq dernières années.

En considérant uniquement les lignes de WatchedMovies.txt liées aux cinq dernières années (c'est-à-dire les lignes avec StartTimestamp dans la plage du 17 septembre 2015 au 16 septembre 2020), l'application sélectionne les films qui :

(i) n'ont été regardés qu'au cours d'une seule période de ces 5 années
(ii) et au moins 1 000 fois au cours de cette année.

Pour chacun des films sélectionnés, l'application stocke son identifiant (MID) et l'année au cours de laquelle il a été visionné au moins 1000 fois (une paire (MID, année) par ligne).

Note. **La valeur de StartTimestamp est utilisée pour décider en quelle année un utilisateur a regardé un film spécifique. Ne tenez pas compte de la valeur de EndTimestamp.**

Exemples

- Par exemple, supposons que le film MID15 ait été regardé 1 025 fois au cours de l'année 2016 et qu'il n'ait jamais été regardé au cours des quatre autres années. MID15 est sélectionné et la paire suivante est stockée dans la sortie : MID15,2016

- Par exemple, supposons que le film MID56 ait été regardé 1 056 fois en 2016 et 10 fois en 2018. MID56 ne doit pas être sélectionné car il a été regardé plus d'un an (en considérant les cinq dernières années).

Film le plus populaire depuis au moins deux ans.

En considérant toutes les lignes du fichier WatchedMovies.txt (c'est-à-dire toutes les années), l'application sélectionne les films qui ont été les plus populaires depuis au moins deux ans. La popularité annuelle d'un film au cours d'une année spécifique est donnée par le nombre d'utilisateurs distincts qui ont regardé ce film au cours de cette année spécifique. Un film est le plus populaire au cours d'une année spécifique s'il est associé à la plus grande popularité annuelle de cette année-là.

L'application stocke les identifiants (MID) des films sélectionnés (un MID par ligne).

Note. La valeur de StartTimestamp est utilisée pour décider en quelle année un utilisateur a regardé un film spécifique. Ne tenez pas compte de la valeur de EndTimestamp.

Exemple

On suppose qu'on est en 2013 et qu'on a 4 films et 5 années avec les visualisations suivantes :

	2010	2011	2012	2013	2014
MID1	**100**	**120**	5	**50**	20
MID2	10	3	15	10	5
MID3	15	10	50	30	**55**
MID4	50	10	**150**	**50**	10

On a donc :

- MID1 est le film le plus populaire des années 2010, 2011 et 2013.
- MID2 n'est jamais le film le plus populaire
- MID3 est le film le plus populaire de l'année et 2014
- MID4 est le film le plus populaire des années 2012 et 2013

Il s'ensuit que les films MID1 et MID4 sont sélectionnés et stockés dans le deuxième dossier de sortie de cette application car ils sont le film le plus populaire depuis au moins deux ans. Les films MID2 et MID3 sont ignorés car ils ne satisfont pas à la contrainte.

Bon courage !