==========

TP3
Le traitement Batch avec Hadoop
HDFS- MapReduce

==========

Exercice 1 : Filtrage

Utiliser le fichier nommé worldcitiespop.txt de taille 144MB.

https://github.com/CODAIT/redrock/blob/master/twitter-decahose/src/main/resources/Location/worldcitiespop.txt.gz

Ce fichier contient la liste des villes de la planète, leur géolocalisation, leur code région et leur population.

Le fichier "worldcitiespop.txt" est une base de données tabulaire qui recense des informations sur les villes du monde entier, offrant un aperçu détaillé de leur population et d'autres attributs clés.

Chaque ligne de ce fichier représente une entrée pour une ville spécifique et est structurée avec plusieurs champs tels que le nom de la ville, le pays dans lequel elle est située, la population, la latitude, la longitude, et d'autres données géographiques pertinentes.

Cette base de données fournit une ressource riche pour l'exploration des tendances démographiques à l'échelle mondiale et la réalisation d'analyses approfondies sur la répartition de la population dans différentes régions.

Le format tabulaire du fichier, associé à la diversité des informations qu'il contient, en fait un choix idéal pour des activités d'analyse de données, en particulier lorsqu'il est exploité à travers des techniques comme MapReduce pour extraire des insights significatifs sur les caractéristiques des villes à travers le monde.

1. Utilisez la commande hdfs tail pour regarder la structure de notre fichier de population.
2. Ecrivez un premier programme map reduce qui lit ce fichier et qui enlève toutes les lignes dont la population n'est pas connue. Pour lancer votre programme utilisez la commande :
3. yarn jar votrejar InputURI OutPutURI
4. Utilisez la commande hdfs cat pour compter le nombre de lignes du fichier généré par votre reducer.

   Information : Le résultat de chaque reducer est enregistré dans un fichier nommé part-r-XXXXX et situé dans le répertoire spécifié par setOutputPath.

   Attention : La première ligne du fichier est différente des autres.

Exercice 2 : Histogramme

Ecrivez un programme MapReduce qui calcul l'histogramme de fréquence des villes dont ont a la population. On utilisera une échelle logarithmique pour déterminer les classes d'équivalences des villes. Soit v1 et v2 deux villes v1 et v2 sont dans la même clasee d'équivalence si et seulement si int(log(pop(v1))) == int(log(pop(v_2))).

Le fichier généré par votre "reducer" devrait ressembler à cela :

10 45
100 585
1000 9429
10000 27065
100000 9820
1000000 979
10000000 57

Dans ce tableau on peut voir qu'il y a 45 villes avec entre 0 et 10 habitants et 27065 villes avec 10000 habitants. Visualisez votre résultat à l'aide d'un tableur.

Exercice 3 : Résumé

Modifiez votre programme précédent pour ajouter le calcul de la population moyenne, de la population minimale et de la population maximale de chaque classe d'équivalence.

Le fichier généré par votre "reducer" devrait ressembler à cela :

classe count avg max min
10 45 xx xx xx
100 585 xx xx xx
1000 9429 xx xx xx
10000 27065 xx xx xx
100000 9820 xx xx xx
1000000 979 xx xx xx
10000000 57 xx xx xx