TP 2 : Map-Reduce (Part 1)

idir.benouaret@epita.fr

Partie TD

Pas d'implémentation pour le moment, réfléchir aux exercices et écrire un pseudo-code.

Exercice I : Histogramme

Supposons que vous avez un fichier très volumineux, qui contient la liste de toutes les villes de la planète ainsi que leur population. Par exemple une ligne correspond à : `lyon,1416545`

Q1 – Écrivez un programme MapReduce qui calcule l'histogramme de fréquence des villes

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 classe d'équivalence si et seulement si int(log(pop(v1))) == int(log(pop(v2))).

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

```
10 34
100 85
1000 9429
10000 13065
100000 8830
1000000 979
10000000 57
```

Ce qui veut dire : il y a 34 villes avec entre 0 et 10 habitants, 9429 avec une population entre 100 et 1000, etc.

Exercice II : Moyenne

On considère un fichier CSV contenant des mesures de température. Chaque ligne du fichier correspond à `"year, month, temperature"`.

Q1 – Écrire un algorithme Map-Reduce capable de calculer la température moyenne pour chaque année

Partie TP

Pour cette partie d'abord réfléchir aux pseudo-codes. Puis implémenter avec la librairie de map-reduce MRJOB :

Exercice III : Découverte de la librairie MRJOB

`https://mrjob.readthedocs.io/`

Q1 – Installer mrjob

Q2 – lire, tester les programmes de bases de mrjob :
`https://mrjob.readthedocs.io/en/latest/guides/quickstart.html`

Exercice IV : WordCount

Télécharger les fichiers suivants et les mettre sur votre répertoire hdfs.
`https://gitlab.cri.epita.fr/idir.benouaret/bda_datasets/-/tree/main/Livre`

Q1 – Implémenter le programme Mapreduce qui permet de compter le nombre d'apparition de chaque mot dans le texte.

Q2 – Ecrire un programme qui compte le nombre exact de caractères, de mots et de ligne dans un fichier texte.

Q3 – Ecrire un programme qui retourne le mot plus plus long sur chaque ligne

Q4 – Écrire un programme qui retourne le mot le plus long du texte

Exercice V : Exercices sur le fichier d'arbres

Vous allez écrire plusieurs programmes *MapReduce* sur le fichier des arbres remarquables à Paris, qui se trouve ici :

- `https://gitlab.cri.epita.fr/idir.benouaret/bda_datasets/-/blob/main/arbres.csv`

Q1 – Tout d'abord, placer ce fichier sur un répertoire HDFS nommé data. i.e. le fichier doit se trouver sur hdfs :///user/root/data/arbres.csv

Pour chacun des programmes suivants que vous devez écrire, testez sur votre machine locale puis sur hadoop.

Q2 – Écrire un programme *mapreduce* qui affiche la liste de tous les arrondissement distincts contenant des arbres dans ce fichier. Il y en a forcément moins de 20 sauf s'il existe des erreurs de saisie.

Q3 – Écrire un programme qui calcule le nombre d'arbres existants pour chacun des genres. Par exemple, il y a 32 *Platanus*, 11 *Quercus*, etc.

Q4 – Écrire un programme qui calcule la hauteur du plus grand arbre de chaque genre. par exemple, le plus haut *Zelkova* fait 26 mètres

Q5 – Écrire un programme qui affiche l'arrondissement où se trouve le plus haut arbre

Q6 – Écrire un programme qui affiche l'arrondissement qui contient le plus d'arbres

Exercice VI : Exercice sur des données d'achat

L'objectif est de collecter des informations et de calculer des statistiques sur les résultats des ventes stockés dans un fichier volumineux. Le fichier `purchases.txt` se trouve à la racine de *hadoop-master* ou bien à télécharger ici (si vous travaillez en mode local) purchases.txt

le fichier est organisé comme suit :

- Date (format YYYY-MM-DD)
- Time (format hh :mm)
- Purchase city
- Purchase category (e.g., Book, Men's Clothing, DVDs...)
- Amount spent by the customer
- Payment method (e.g., Amex, Cash, MasterCard...)

les colonnes sont séparées par une tabulation

Q1 – Quelle est le nombre de commandes effectuées pour chacune des catégories d'achat ?

Q2 – Quel est le montant total dépensé dans chaque catégorie d'achat

Q3 – Combien d'argent est dépensé dans la ville de San Francisco pour chacune des méthodes de paiement

Q4 – Dans quelle ville la catégorie "Women's Clothing" a généré le plus d'argent en utilisant le mode de paiement "Cash"