

Atelier : Lancer des jobs Mapreduce

L'objectif : Comprendre le fonctionnement mapreduce

Réalisation : Démonstration et lancement d'un programme MapReduce qui compte le nombre d'occurrence de chaque mot dans un fichier.

Chapitre correspondant : Le framwork MapRedue/Yarn

1- Préparation des fichiers MapReduce :

a- Ouvrir le terminal en ligne de commande de votre VM, et récupérer le fichier de constitution.txt situé dans moodle

b- Mettre le fichier constitution.txt dans HDFS :

2- Lancer le job WordCount :

La commande ci-dessous lance le job MapReduce wordcount pour compter le nombre d'occurrences de chaque mot et écrire le résultat dans le répertoire wordcount_output

Hortonworks :

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar wordcount constitution.txt wordcount_output
```

Cloudera :

```
yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount constitution.txt wordcount_output
```

3- Visualiser le résultat :

a- Visualiser le contenu du répertoire wordcount_output

b- Visualiser le contenu du fichier par-r-00000

4- Lancer d'un programme MapReduce pour le fichier ville.txt:

a- Mettre le fichier ville.txt dans HDFS

b- Lancer le job ville et stocker le résultat dans ville_output:

Hortonworks

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar wordcount ville.txt ville_output
```

Cloudera :

```
yarn jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount  
constitution.txt ville_output
```

5- Visualiser le résultat :

- a- Visualiser le contenu du répertoire ville_output
- b- Visualiser le contenu du fichier

6- Questions Mapreduce :

- a- Pourquoi nous avons un seul fichier partiel part-r-00000 ?
- b- Que signifie le -r- dans le nom du fichier ?
- c- Pourquoi les mots sont triés dans un ordre alphabétique ?
- d- Quelle est la valeur du wordcount reducer
- e- Que représente la clé valeur pour le mapper ?

hadoop jar mr.jar MyLauncher

(alias vers la commande java sauf que ça rajoute l'intégralité du CLASSPATH d'hadoop dedans)