Atelier : Lancer des jobs Mapreduce

L'objectif : Comprendre le fonctionnement mapreduce

Emplacement du fichier : /formation/ateliers/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 aller lire le fichier de constitution.txt situé dans le répertoire /formation/ateliers/mapreduce/consitution.txt

```
cat /root/formation/ateliers/mapreduce/constitution.txt
```

b- Mettre le fichier constitution.txt dans HDFS :

```
Hadoop fs -put /root/formation/ateliers/mapreduce/constitution.txt
```

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

```
hadoop fs –ls wordcount_output
```

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

```
hadoop fs -cat wordcount_output/part-r-00000
```

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

a- Mettre le fichier ville.txt dans HDFS

```
hadoop fs -put /root/formation/mapreduce/ville.txt
```

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 réprtoire ville_output

```
hadoop fs -ls ville_output
```

b- Visualiser le contenu du fichier

```
hadoop fs -cat ville_output/part-r-00000
```

6- Questions Mapreduce :

a- Pourquoi nous avons un seul fichier partiel part-r-00000 ?

Parce que nous utilisant un seul reducer

b- Que signifie le –r- dans le nom du fichier ?

Le –r- fait référence au reducer

c- Pourquoi les mots sont triés dans un ordre alphabétique ?

La clé dans ce job mapreduce est les mots, ces derniers sont triés lors de la phase shuffle/sort

d- Quelle est la valeur du wordcount reducer

La valeur du reducer est la sum des occurrences pour le même mot dans le document

e- Que représente la clé valeur pour le mapper ?

La clé est chaque mot, la valeur est de 1

hadoop jar mr.jar MyLauncher

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