Travaux Pratiques Hadoop

Grégoire MASSOT - gregoire-massot.com

14 Novembre 2015. MàJ le 1 Avril 2016

Voici un compte-rendu des TPs Hadoop réalisés lors du cours Big Data en 3ème année à l'École des Mines, entre Septembre et Novembre 2015.
Ces TPs ont été réalisés sous Windows avec la machine virtuelle Cloudera CDH 5.5.

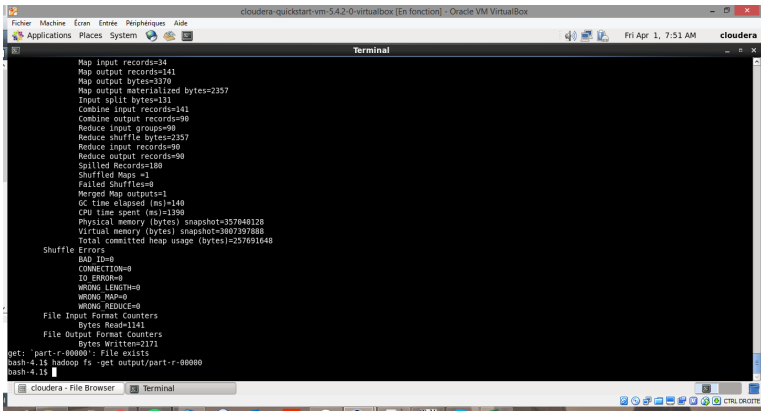

FIGURE 1 – Screenshot de la VM CDH5.5 de Cloudera

1 initialize.sh et launch.sh

Ce sont les scripts qui permettent respectivement d'automatiser les tâches d'initialisation de la machine virtuelle et d'exécution du programme. Plus de détails dans les commentaires.

2 tinygraph.txt

Le fichier d'entrée `tinygraph.txt` est un graphe de pages web. Chaque ligne correspond à une page web. Pour chaque ligne i :

- Le premier nombre est le numéro associé à la page web i
- Le second nombre est le PageRank initial de la page i. Il est égal à $\frac{1}{N}$ où N est le nombre de pages web contenues dans le graphe
- Les nombres suivants sont les numéros associés aux pages vers lesquels pointent les liens hypertextes de la page i

3 pageRank.java

C'est le programme qui va calculer les Page Rank des pages contenues dans `tinygraph.txt`.

- **Map** : Pour chaque ligne i on liste les liens sortants j et on produit les couples clé-valeur (page web j, $\frac{1}{N_i}$) avec N_i le nombre de liens sortants de la page i.
- **Reduce** : pour chaque paire clé-valeur, on sort somme les valeurs et on sort (page web j, somme valeurs associées à la clé j)

4 part-r-000000

C'est le fichier de sortie fourni par notre cluster Hadoop. Chaque ligne représente une page web. Pour chaque ligne i

— Le premier nombre correspond au numéro de la page web i

— Le second nombre est le PageRank de la page après calcul.