

Résumé semaine 1

Tout d'abord j'ai lu des documents définissant et expliquant le phénomène du Big Data. Cela m'a amené à étudier le patron MapReduce qui est considéré comme étant le plus efficace à régler ces problèmes de données massives. J'ai donc regardé son fonctionnement algorithmique et listé les différents frameworks disponibles. À partir de là je me suis concentré sur le framework Hadoop, le plus utilisé pour le Big Data. J'ai étudié son fonctionnement, sa composition et me suis renseigné sur les différentes améliorations possibles. J'ai ensuite repéré les distributions possibles pour installer Hadoop, et à partir de là essayé de le configurer sur ma machine.

J'ai essayé de faire une installation manuelle via Cloudera, au final Hadoop fonctionne, la commande est reconnue par le système, le problème est que je ne parviens pas à le configurer car les différents fichiers à configurer sont pour la plupart en lecture seule.

J'ai également installé sur VirtualBox, la sandbox de HortonWorks, sur laquelle j'arrive à faire fonctionner Hadoop à partir de l'interface web. C'est-à-dire utilisé l'interface web proposée par HortonWorks pour uploader des fichiers sur mon HDFS et utiliser les outils comme Hive, que j'ai essayé, et Pig, où j'ai des petits problèmes lors de mes essais. Cependant, il me faut essayer de faire fonctionner certains programmes Java sur la machine virtuelle.

J'ai également regardé pour d'autres possibilités, notamment avec Ambari de HortonWorks mais sans de réel résultat.

Durant cette semaine j'ai regardé plusieurs tutoriels sur le fonctionnement du patron MapReduce, et notamment par la mise en œuvre d'un programme Java permettant de compter les mots d'un fichier. Ceci étant un exemple classique d'utilisation. Il ne me reste plus qu'à le tester lorsque j'aurai réussi à configurer correctement Hadoop.

Par cet exemple et quelques autres j'ai pu appréhender certains aspects du langage utilisé par Hadoop. Ceci me permettra un peu plus tard de pouvoir réaliser des programmes plus complexes.

Il faut également que je vérifie avec quel format les fichiers téléchargés sur le site de l'observatoire doivent être édité pour pouvoir être récupérés sans problème par Hadoop.