Résumé semaine 1

Tout d'abord j'ai lu des des documents définissant et expliquant le phénomène du Big Data. Cela m'a amené à étudier le parton MapReduce qui est considéré comme étant le plus efficace à régler ces problème de données massives. J'ai donc regarder son fonctionnement algorithmique et listé les différents framework disponible. A partir de là je me suis concentré sur le framework Hadoop. Le framework le plus utilisé pour le Big Data. J'ai étudier son fonctionnement, sa composition et me suis renseigné sur les différentes améliorations possible. J'ai ensuite repérer les distribution possible pour installer Hadoop, et à partir de là essayer de le configurer sur ma machine.

J'ai essayé de faire une installation manuelle via cloudera, au final hadoop fonctionne, la commande est reconnu 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 en lecture

J'ai également installer sur virtualbox, la sandbox de HortonWorks, sur laquelle j'arrive à faire fonctionner hadoop mais surtout l'interface web. C'est-à-dire utilisé l'interface web proposé par HortonWorks pour upload des fichiers sur mon HDFS et utilisé les outils comme Hive, que j'ai essayé, et Pig, où j'ai des petits problème lors de mes essais. Cependant, il me faut essayer de faire fonctionner certains programmes Java sur la machine virtuel.

J'ai également regarder pour d'autres possibilité, 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 aspect 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 télécharger sur le site de l'observatoire doivent être édité pour pouvoir être récupérer sans problème par Hadoop.