Semaine 3

Durant cette 3ème semaine de stage, j'ai réalisé un certain nombre d'expérimentation pour tester le framework Spark. Ce framework proposer par la fondation Apache peut être vu comme une amélioration de Hadoop. Cependant, il ne faut pas le voir comme un remplacement d'Hadoop mais comme un projet s'intégrant à Hadoop. En effet, Spark a la possibilité notamment de se servir du HDFS pour réaliser ses jobs. J'ai donc rédiger un programme simple de compteur d'occurrence dans un fichier texte et l'ai essayé sur différents fichiers.

Le framework Spark propose l'utilisation de 3 langages : Java, Scala ou Python. Il propose notamment des shell en Scala et En Python pour pouvoir taper les commandes directement sur le terminal et ainsi ne pas avoir à créer un projet. De plus, Spark permet de créer un cluster de plusieurs nœuds assez facilement. J'ai donc réaliser un cluster à plusieurs nœuds en local sur ma machine. Pour cela, il suffit de configurer un terminal pour qu'il soit le maître et de configurer les autres pour qu'ils le reconnaissent pour leur maître et ainsi devenir des workers. Une fois le cluster configurer, j'ai à nouveau exécuté mon programme pour voir la différence. Je n'ai pas vu beaucoup de changement pour 2 raisons : tout d'abord les fichiers de tests sont de trop petites tailles, et ensuite bien que le cluster soit fait de plusieurs nœuds, il reste exécuté en local. On peut juste remarquer grâce à une interface disponible en localhost, le nombre de worker utilisé pour réaliser la tâche.

J'ai donc produit un fichier bien plus gros en essayant d'atteindre le Giga octets et j'ai pu remarquer que Spark à l'instar d'Hadoop produit en sortie un certains nombre de fichiers chacun correspondant à la sortie d'un reducer.

Durant cette semaine, j'ai cherché à trouver un cas concret d'application que l'on pourrait réaliser pour le CDS. Par l'intermédiaire de mon tuteur, j'ai pu discuté avec M. François Xavier Pineau qui avait déjà réalisé des travaux équivalents mais eux se basant sur un système de threads. Nous avons donc discutés ensemble pour pouvoir déterminer si les applications que lui avait réalisé pouvait être réalisé sur Hadoop. Je me suis alors rendu compte qu'il me manquait certaines compétences pour pouvoir déterminer si ces travaux étaient applicable à Hadoop.

J'ai donc à nouveau réalisé une série de recherche visant à approfondir les points qu'il avait mis en lumière. Une fois ceci fait, j'ai eu une nouvelle conversation avec lui où nous avons discuter plus en détails des applications possible. 2 ont été retenu :

- La première, mais aussi la plus intéressante, est de se servir d'Hadoop pour réaliser une comparaison de 2 ensembles de données. C'est-à-dire, de les comparer et de retourner uniquement les données communes aux deux ensembles.

- La seconde, aurait pour objectif de lire un ensemble de données et d'enlever toutes les données ne correspondant pas à certains critères.

Une fois ces objectifs fixés, j'ai commencé quelques recherches sur ce qui a déjà été fait.