Semaine 2

Début de semaine :
Expérimentation diverses avec Hadoop via la SandBox d'HortonWorks lancé sur une machine virtuel. J'ai regarder de la documentation sur le langage de script Pig, qui est sous projet réalisé par la fondation Apache pour compléter Hadoop, et qui est inclus dans la distribution.

J'ai donc étudier les différentes commandes possible avec ce langage, notamment les filtre, les jointures et les créations de relations. Puis je m'y suis essayer en tentant quelques script basé sur des exemples simple. Par exemple, charger un fichier texte et compter les occurrences des mots, puis en ajoutant des filtres pour limiter le nombre de résultat. Et bien d'autres, sans aller toutefois vers un cas concret car je n'avais pas trop d'idée et que l'objectif principale était de me faire la main.

J'ai ensuite voulu tester quelques scripts directement sur des fichiers de test récupéré sur la plate forme Vizier, j'ai donc généré 3 fichiers de tailles différentes, qui me serviront ici mais aussi pour plus tard. Une fois généré, j'ai dû supprimer les entêtes des fichiers car ils gênaient la génération de la table avec Hcatalog. J'ai donc pu réaliser un script qui travaillait sur des données concrètes. Cependant, je n'avais pas encore d'idée concrète sur le type de réalisation utiles à faire et je n'ai donc réaliser que quelques petits essai. Tout d'abord j'ai chargé la table des valeurs via Hcatalog, cependant, les colonnes ne comportaient pas de nom cohérents ce qui rendaient plus difficile le traitement. J'ai donc chargé le fichier puis l'ai rangé dans une relation où les noms de colonnes étaient plus intuitif. Ensuite j'ai réalisé une opération de filtrage classique permettant de ne récupérer que certaines valeurs.

C'est à peu près tout ce que j'ai réalisé comme travail autour de Pig.
J'ai essayé de réaliser un projet Hadoop classique qui lancerait l'exécution d'un programme Java simple visant à compter les occurrences des mots dans un fichier texte. Cependant, le programme une fois réaliser je me suis rendu compte qu'il manquait à mon projet l'ensemble des librairies utilisés par Hadoop. En cherchant un peu j'ai trouvé un projet de la fondation Apache, Maven qui permet de construire des architectures de projets et de créer le .jar en ajoutant automatiquement l'ensemble des librairies nécessaire. J'ai donc installé ce programme, cependant j'ai des problèmes de version avec certaines sources de Hadoop qui empêche l'exécution de maven. Je n'ai pas été plus loin dans cet essai.

Milieu et fin de semaine

Après une discussion avec mon tuteur de stage, nous avons mis en avant un autres framework de BigData qui serait 100fois plus puissante que Hadoop et plus récent. Il s'agit également d'un projet de la fondation Apache : Spark. Il s'agit d'un framework fonctionnant sur certaines base de Hadoop mais comportant des caractéristiques le rendant plus performants. Ce framework n'a pas pour objectif de remplacer Hadoop mais de le compléter. Il est donc possible de se servir du HDFS pour réaliser le cluster de Spark. Cependant, il faut que je voie comment ceci fonctionne.

J'ai donc réaliser une étude du framework Spark, ses performances, sa composition, son fonctionnement, et autre. J'ai téléchargé une release de Spark mais j'ai eu un petit problème avec mon espace de stockage que j'ai fin par réglé. J'ai donc installé le framework Sparksur ma machine. Spark étant un framework fonctionnant avec 3 langages, Python, Java et Scala ; j'ai donc décidé d'installer également Scala pour pouvoir éventuellement faire des essais avec.
En observant quelques tutoriels du langages en lui-même par quelques exemples simple, j'ai pu me

rencontre que la mise en œuvre d'un programme en Spark pourrait s'avérer plus simple qu'avec Hadoop, également son exécution. Cependant, il me faut le vérifier

Conclusion

Cette semaine j'ai donc :

- expérimenté un peu plus en détail Pig et son fonctionnement en me basant d'abord sur des exemples simple puis sur un ensemble de données récupéré sur la plate forme Vizier
- J'ai découvert le framework Spark qui pourrait s'avérer plus efficace que Hadoop, j'ai donc du faire des recherches et des installations.

Objectif :

- faire des expérimentations avec Spark : écrire un programme simple et l'exécuter
- me renseigner sur la mise en œuvre d'un cluster avec Spark, simple-node,multi-node en local et multi-node sur machines distantes.
- continuer de regarder ce qu'il est possible de faire avec Hadoop et Spark