Projet pour l'UE: Infrastructures pour le Cloud et Big Data

Responsables:

- Boris Teabe
- Daniel Hagimont

Contexte

Dans le cadre de votre cours sur les Infrastructures pour le Cloud et le Big Data, vous devez réaliser un projet. Ce projet vise à consolider les connaissances acquises durant le cours. Il aura également pour but d'approfondir les sujets abordés durant les séances de travaux pratiques.

Résumé

En une phrase, ce projet consistera à implémenter une plateforme de Big Data As-A-Sevice (BaDaaS). Vous serez responsable de l'implantation/gestion de votre l'infrastructure, l'implémentation d'une interface (simple) d'accès à votre service et enfin l'utilisation de votre service par le client.

Description détaillée

Comme vous avez pu le constater dans le résumé, votre projet comporte plusieurs aspects que je détaille dans les sections suivantes sous forme d'étape.

1) Infrastructure matérielle avec des machines virtuelles

Durant nos séances de TP, nous avons déployé Xen qui est un hyperviseur Opensource et nous avons créé et exécuté des machines virtuelles (VM). Dans cette prémière étape de votre projet, il sera question de mettre en place, un cluster de VM (minimum 4 VM) qui s'exécutera sur vos différentes machines (laptops). Vous pouvez utiliser un hyperviseur qui vous semble simple de prise en main, c-a-d Xen, KVM, VMWare ou bien HyperV. Libre à vous de choisir votre hyperviseur. La seule contrainte est que vos VM ne doivent pas s'exécuter sur une seule machine physique.

2) Déploiement de Spark sur votre cluster de VMs

Durant le cours, nous avons travaillé sur la plateforme Spark. Dans cette étape du projet, vous devez déployer Spark sur cotre cluster de VM avec HDFS comme système

de fichiers. Votre déploiement de Spark doit être en mode cluster, c-a-d avoir plusieurs datanodes.

3) Interface de votre BaDaaS

La prochaine étape sera l'implémentation d'un portail pour l'accès à votre service. Pas besoin d'avoir une gestion des comptes utilisateurs. Votre portail devra permettre de réaliser 3 opérations:

1. **Chargement de données**. L'utilisation devra pouvoir charger des données vers votre infrastructure Spark, plus précisément vers HDFS, à travers votre interface en spécifiant un chemin.
2. **Enregistrement d'un progamme Big Data**. Votre interface doit permettre d'uploader un programme (un jar si vous faite en java) qui pourra être utilisé par la suite pour lancer un job.
3. **Exécution d'une application Big Data.** Après que l'application soit uploader, ce dernier doit être utilisable. Ainsi, vous devrez disposer d'une interface permettant au client de choisir le programme qu'il souhaite utiliser, de spécifier le fichier d'entrée (un fichier chargé dans HDFS) et le repertoire de sortie dans HDFS. Noter que l'output sera fonction de l'algorithme big data implémenté. Il devra alors par la suite être possible de download le ouput dans un fichier (hors de hdfs).

4) Rédaction du rapport

La rédaction du rapport est également une phase de votre projet, une des plus importante car elle permet de présenter votre travail. Vous devez rédiger un rapport avec des sections correspondant à chacune des phases sus-citées. Pour chaque phase, il faudra décrire votre implantation et justifier vos choix.

Organisation

Le projet sera réalisé en groupe de 4 étudiants. Il n'y a pas de contraintes sur votre organisation interne et sur la répartition des tâches au sein de votre groupe. Il est important de mentionner dans le rapport la participation de chaque membre, et durant la remise des projets vous serez également ammené à la spécifier. C**onséquemment, les notes seront individuelles**.

Quelques contraintes

J'impose quelques contraintes parce que je peux.

1) Pour le langage de programmation, vous pouvez utiliser uniquement du **JAVA** ou **Python**

2) Et pour l'interface web de votre projet, libre à vous de choisir les langages de programmation qui vous conviennent.