Projet

Infrastructures pour le Big Data

Valdom

décembre 2024

Daniel Hagimont

Contexte

Dans le cadre de votre cours sur les Infrastructures pour le Big Data, vous devez réaliser un projet. Ce projet vise à consolider les connaissances acquises durant le cours. et les séances de travaux pratiques.

Résumé

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

Description

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

1) Infrastructure d'hébergement

Le but est ici d'implanter une infrastructure qui hébergera des machines virtuelles, machines virtuelles qui seront utilisées pour exécuter vos applications Big Data. Dans le cadre de ce projet, je vous propose d'utiliser la plate-forme AWS comme infrastructure d'hébergement.

Une partie administration de l'interface web doit permettre de gérer l'infrastructure, ajout/retrait de machines virtuelles et configuration de celles-ci.

Il vous faut implanter des scripts permettant de déployer et de-déployer votre infrastructure. On pourra notamment utiliser Boto3 qui permet d'accéder à AWS depuis Python.

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 votre cluster de VM avec HDFS comme système de fichiers. Votre déploiement de Spark doit être en mode cluster, c-a-d avoir plusieurs datanodes. Votre infrastructure Spark/HDFS devra évoluer en fonction de l'ajout/retrait de machines virtuelles.

3) Interface de votre BDaaS

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 4 opérations:

1. **Chargement de données**. L'utilisateur 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 uploadée, cette dernière 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. Notez que l'output sera fonction de l'algorithme big data implémenté.
4. **Récupération des résultats.** Après l'exécution, votre interface doit permettre de télécharger les résultats (à partir de HDFS).

4) Rédaction du rapport

La rédaction du rapport est également une phase de votre projet, une des plus importantes 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 3 ou 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.