

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 (longue), ce projet consistera à : (1) automatiser le déploiement de Kubernetes sur un cluster de VM sur AWS, (2) déployer un outil de monitoring d'application dans Kubernetes ([kube-opex-analytics](#)), (3) déployer Hadoop Spark sur votre cluster Kubernetes et exécuter l'application WordCount sur votre déploiement tout ceci en monitorant l'utilisation des ressources (CPU et mémoire).

Description détaillée

Comme vous avez pu le constater dans le résumé, votre projet consiste en plusieurs phases. Cette section donne plus de détails sur chacune de ces phases.

1) Automatisation du déploiement de Kubernetes sur un Cluster de VM sur AWS

Durant nos séances de TP, nous avons utilisé le service EC2 d'Amazon via le portail qu'il propose. L'objectif de cette phase du projet est d'utiliser ce même service mais à travers les API que propose AWS. Vous devez donc, écrire un programme qui permettra d'instancier des VM sur AWS, de se connecter à ces VM et d'y déployer Kubernetes, ceci sans l'intervention de l'homme.

2) Déploiement d'un outil de monitoring d'application dans Kubernetes (Kube-opex-analytics)

Cette phase du projet porte sur le monitoring d'application sur Kubernetes. Pour ce faire, vous allez utiliser l'outil **Kube-opex-analytics** [1], qui est un outil open source qui permet d'extraire l'utilisation CPU et Mémoire d'une application s'exécutant dans un cluster Kubernetes et de les présenter via des graphiques.

3) Déploiement de Hadoop Spark et exécution de WordCount

Durant le cours, nous avons travaillé sur la plateforme Spark. Dans cette phase du projet, vous devez déployer Spark sur votre cluster Kubernetes. Votre déploiement de Spark doit être en mode cluster, c-a-d avoir plusieurs datanodes. Il faudra par la suite tester votre déploiement en exécutant l'application WordCount qui vous a déjà été fournie durant les séances de TP. L'outil de monitoring sera utilisé pour afficher des statistiques sur l'utilisation CPU et mémoire de notre application WordCount.

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 amené à la spécifier. **Conséquemment, les notes seront individuelles.**

Quelques contraintes

J'impose quelques contraintes parce que je peux.

- 1) Le projet sera rendu via un/plusieurs fichier(s) docker Compose, c-a-d le jour de la démo vous viendrez tout simplement des fichiers docker-compose.
- 2) Pour le langage de programmation, vous pouvez utiliser uniquement du **JAVA** ou **Python**

Références

[1] <https://github.com/rchakode/kube-opex-analytics>

[2] <https://spark.apache.org/docs/latest/running-on-kubernetes.html>