Déploiement Rapide de Middleware pour le Big Data

Fabrice Huet
fabrice.huet@unice.fr

30 janvier 2017

Nombre d'étudiants souhaités : 3

Description du sujet

Pour pouvoir manipuler de grandes quantité de donnés (Big Data) il est nécessaire d'avoir recours à des outils spécialisés appelés *frameworks*. Ces outils fournissent souvent un modèle de programmation comme par exemple Map-Reduce, mais aussi tout un environnement d'exécution. Cet environnement permet d'exécuter des programme sur un ensemble de machines sans avoir à se soucier des problèmes de communication entre les machines ou d'éventuelles pannes.

Trois outils sont actuellement très utilisés à la fois dans l'industrie et la recherche : Hadoop [1], Spark [2] et Storm [3]. Chacun a un fonctionnement légèrement différent mais néanmoins proche dans les concepts.

Lors qu'un chercheur ou un industriel veut utiliser ces outils, il doit les déployer sur un ensemble de machine. Il s'agit d'une phase technique qui nécessite d'installer et de configurer un ensemble de machines. Quand on utilise une ressource de type Cloud, les machines ne sont disponibles que pendant leur utilisation et donc il est nécessaire de redéployer le framework avant chaque expérimentation.

Le but de ce TER est de mettre en place des outils pour déployer rapidement les 3 logiciels précédents le plus rapidement possible et sans aucune intervention de l'utilisateur. Les étudiants auront accès à un cluster de centaines de machines [4] afin de tester leur outil.

Mots clés : Docker, Puppet, Chef, Grid'5000, Cluster, Cloud, Big Data

Lieu

Sessions de travail régulières au laboratoire I3S (Sophia Antipolis)

Prérequis

Connaissance de Java, programmation shell, scripting et système.

1. http ://hadoop.apache.org/
2. https ://spark.apache.org/
3. http ://storm.apache.org/
4. http ://www.grid5000.fr

Informations complémentaires

Ce TER constitue un bon moyen de se former à des technologies qui ne sont pas normalement enseignées pendant les cours.