EISTI

Big Data et applications sur Spark

Cahier des charges - TIPE

CPI2 C2

Maxime LUNDQUIST
Mandry MBUNDU
David RIGAUX
Mehdi DALAA

26 novembre 2016

Table des matières

Introduction

Dans le cadre du TIPE, notre groupe allons travailler sur le monde du Big Data. Sachant que le Big Data représente un enjeux majeur dès aujourd'hui mais également pour l'avenir, il nous semblait important d'étudier ce domaine. En effet, lors de ce projet nous allons travailler avec *Spark*, qui est un framework open source de calcul distribué appliqué au Big Data. Le Big Data étant une notion relativement complexe par rapport à notre niveau, ce projet se construit petit à petit, et au cours du temps avec notre professeur référant, Monsieur Senoussi. C'est donc pour cela que nous allons tout d'abord découvrir les fonctionalitées de Spark en écrivant des algorithmes d'analyses en Python.

1 Découvrir les fonctionalitées de Spark

Spark est un outil pour le Big Data relativement puissant dû à ces nombreuses fonctionalitées. Pour les découvrir, nous allons commencer par étudier les algorithmes de base dans le domaine du Big Data. Spark nous permet de coder les algorithmes en Python ou en Scala et propose un shell intéractif, qui nous permet d'aborder ce framework plus facilement que d'autres.

2 Écrire des algorithmes d'analyse de données en Python sous Spark

Après avoir étudié ces algorithmes d'analyse de données, nous allons les coder en Python et les appliquer à des ensembles de données qui reste simple (le plus simple sont des coordonnées de points de dimension 2). Nous codons en Python car il est plus simple de coder dans ce langage de programmation par rapport au Scala. Un exemple d'algorithme que nous avons déjà commencé à coder mais que nous devons encore amélioré est *Kmeans*, ou en français *K-moyennes*. Cet algorithme permet de partitionner un ensemble de données en K clusters.

Conclusion

Réaliser un projet Big Data est complexe à notre niveau et cette complexité va augmenter étant donné que nous allons nous pencher sur l'étude d'algorithmes appliqué

à des données plus variées. Les fonctionnalités et les outils mis à notre disposition vont nous permettre d'avancer relativement vite sur l'application de nos algorithme. Il nous reste encore a bien étudiés ces outils afin d'être le plus productif possible dans l'avancée de notre projet.