Big Data - Data scientist

Visualisation et analyse des données dans le Cloud/Cluster.

Objectifs : Conception et développement d'une application web pour la visualisation de données Big-Data et le déploiement d'algorithmes sur le Cloud et/ou Cluster.

Vous serez en contact avec les partenaires du projet investissement d'avenir Big Data : **Arrow group**, Isthma , Digital & Ethics, Assureur (AXA). http://square-predict.net/

Connaissances requises : Etudiants motivés, développeurs souhaitant découvrir et acquérir des compétences très recherchées par le monde industriel.

Coordonnées du client : Mohammed Ghesmoune, Mustapha Lebbah, Hanane Azzag (mohammed.ghesmoune@lipn.univ-paris13.fr; mustapha.lebbah@univ-paris13.fr)

Sujet:

Le phénomène *big data* est considéré comme l'un des grands défis informatiques de la décennie 2010-2020 (http://fr.wikipedia.org/wiki/Big_data). Face à l'accroissement des volumes de données et au défi du Big Data, les outils de visualisation d'indicateurs voient leur utilité décupler. La bibliothèque D3.js (http://d3js.org/) est peut-être considérée comme le meilleur choix. Le modèle MapReduce (avec ses implémentations Hadoop ou Spark (https://spark.apache.org/)), inventé par Google, et qui permet un traitement distribué au sein d'un cluster de nœuds, connaît un vif succès auprès de grandes sociétés telles Amazon ou Facebook.

Dans le cadre d'un projet investissement d'avenir big-data de grande envergure, nous cherchons une équipe de travail qui aura pour principales missions :

- Le Responsive design est un enjeu majeur dans le développement des IHM moderne et industrielle de demain. Du développement web sur des langages et libraires innovants (HTML5/CSS3, Angular.js, Node.js) au développement d'application mobile cross plateforme Hybride (PhoneGap, iosc…, Suncha) l'équipe assurera la mise en œuvre d'une application web dynamique et interactive, accessible, via le net, à partir de PCs, Tablettes, et SmartPhones et compatible avec un déploiement sous tutelle du Cloud Computing.
- Le chargement de données massives sur le HDFS. Ces données peuvent être sous différents formats (CSV, JSON, RDF, etc.). L'hétérogénéité des sources et des données devra amener l'équipe à explorer de nouvelles approches de collecte, en streaming, notamment via l'utilisation de librairie comme Spring XD.
- La visualisation de ces données de manière interactive. Pour cela, nous recommandons l'utilisation de la bibliothèque D3.js. L'accès exploratoire et donc l'accès directe et par requête aux données sur le HDFS sera nécessaire. L'équipe sera amenée à utiliser des technologies classiques (HIVE, PIG, SPARQL) en passant aussi par des techniques plus innovantes et puissantes (Elastic Search…)
- Le déploiement et le lancement des algorithmes, écrits en Scala/Java/Python sous Spark ou Hadoop, sur le Cloud et/ou Cluster tel que Amazon EC2, Google Engine, Teralab. En premier lieu, il faudrait déployer l'application sur la plateforme Teralab (http://www.teralab-datascience.fr/). L'application doit permettre de lancer des algorithmes existants, mais aussi prévoir le lancement de nouveaux algorithmes qu'il faudrait récrire par la suite sous spark.
- L'équipe projet sera amenée à travailler sur toutes les phases de réalisation d'un logiciel (conception, mise en œuvre, test et déploiement) en adoptant que ce soit une méthodologie agile ou classique.
- Le choix des technologies de l'implémentation est laissé à l'équipe de développement, même si nous préconisons les nouvelles technologies telles que Java JEE ou encore .NET.