Conseiller, Accompagner, Concevoir, Diffuser en ligne.

Jouve IT Solutions libère vos contenus

Stage Bigdata BI

- **Titre** : **Systèmes d'analyse des liens dans un fonds documentaire**
- **Lieu** : Cesson-Sévigné
- **Durée :** 4 à 6 mois
- **Formation :** Bac + 5
- **Date :** à partir de mars 2015
- **Référence :** APU-2
- **Rémunération :** stage indemnisé

Sujet : L'objectif du projet est de faire évoluer la plateforme de valorisation des données actuelle en y ajoutant des nouvelles fonctionnalités. La plateforme permet d'extraire tous les liens d'un fonds contenant plus de 4 millions de documents juridiques et de visualiser des statistiques de répartitions de ces liens.

Les modules technologiques utilisés dans cette première version sont :

- Processus Map/Reduce pour l'extraction des données sur un cluster Hadoop.
- MongoDB pour stocker la base des liens et permettre le calcul des premières statistiques.
- Node.js pour la visualisation des statistiques.
- D3.js pour afficher les différents graphes de l'application.

Vous serez amené à réaliser les nouvelles fonctionnalités suivantes :

- Apport du « pseudo temps réel » dans les processus Map/Reduce en intégrant la surcouche Spark au cluster Hadoop existant.
- Intégration de l'application avec le Workflow de production du fonds documentaire de notre client pour un rafraichissement fréquent des données.
- Mise en place des statistiques plus avancées pour l'analyse des liens du fond documentaire en exploitant les modules MLib et GraphX fournis avec Spark.
- Intégration dans l'application de graphiques d'exploration du graphe de liens avec la bibliothèque graphique D3.js (ou autres, au choix du stagiaire).

Connaissance en programmation Java exigée. Un goût prononcé pour l'abstraction, la modélisation et une connaissance académique des **algorithmes** mis en œuvre dans **MapReduce**, les bases **NoSQL** et les **moteurs de recherche**.

Environnement :

Vous serez intégré dans l'équipe d'expertise **Search et Bigdata** (5 ingénieurs confirmés) et serez amené à contribuer aux projets Apache.

Cette équipe est experte dans l'intégration des moteurs de recherche propriétaires (Antidot, Exalead, Sinequa, IDOL) et open source (Lucene, Solr, ElasticSearch) , ainsi que dans le stockage et le traitement des données volumineuses (hadoop, Spark, Pig, MongoDB, Cassandra, Talend…).