



Conseiller, Accompagner, Concevoir, Diffuser  
en ligne.

**Jouve IT Solutions libère vos  
contenus**

# Stage Bigdata ETL

- **Titre** : Systèmes distribués d'extraction, de transformation et de chargement de données
- **Lieu** : Cesson-Sévigné
- **Formation** : Bac + 5
- **Référence** : APU-1
- **Rémunération** : stage indemnisé

**Durée** : 4 à 6 mois

**Date** : à partir de mars 2015

**Sujet** : L'objectif du projet est de qualifier puis de mettre en place un système distribué d'extraction, de transformation et de chargement de données volumineuses (jusqu'à 1 To de données XML).

Ce système sera utilisé pour préparer de très grandes quantités de données pour par exemple les indexer avec des moteurs de recherche "fulltext" tel que Apache Solr ou Elasticsearch.

Des technologies open-source telles que MapReduce, Hadoop, Spark, Solr, Elasticsearch seront mises en œuvre.

Dans ce cadre, vous serez amené à :

- Etudier les solutions existantes et fournir un état de l'art.
- Réaliser un prototype pour le traitement et la valorisation des données sur un cas client précis.
- Etudier et dimensionner les architectures matérielles et logicielles de nos plateformes de traitement des données.

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**. Mise à disposition d'un cluster de serveurs physiques sous Linux pour la réalisation et les tests.

---

### 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...).

---