FORMATION HADOOP

Développeur pour Hadoop (Apache)

Direction commerciale et marketing : 87, avenue du Maine 75014 Paris - Tél. : 01 72 50 01 26
Centre de services : 19 rue Edouard Vaillant – 37000 Tours - Tel : 02 47 68 48 50 - Fax : 02 47 68 48 59 - www.cyres.fr
SAS au capital de 300 000 Euros - R.C.S. Tours B 442 155 818 - Code NAF: 6201Z

Sommaire

I. Objectifs

Encadrée par un formateur qualifié, cette formation va vous amener à découvrir et comprendre le fonctionnement de l'écosystème Hadoop. Les thématiques traitées seront les suivantes :

- Le système de fichiers distribué HDFS
- Le traitement MapReduce et l'écriture de code
- Les bonnes pratiques de développement et d'implémentation des algorithmes courants
- L'optimisation des configurations et les techniques d'amélioration des performances
- Les jointures des données via MapReduce
- Les projets de l'écosystème Hadoop : Hive, Pig, Flume, Mahout et Sqoop
- Préparation à la certification Cloudera

II. Public concerné

Cette formation convient aux ingénieurs et développeurs qui ont déjà une expérience dans la programmation et qui souhaitent acquérir les connaissances nécessaires pour développer des applications MapReduce.

III. Pré-requis

- Connaissances minimales en système d'exploitation Linux.
- Expériences en développement informatique Java.

IV. Conditions Générales

Formation	Développeur pour Hadoop (Apache)
Référence	HADDEV1
Durée	4 jours (28 heures)
Tarif	A partir de 1 500 € H.T. / personne

V. Contenu de la formation

- **Introduction**

Objectifs :

Tour d'horizon de Hadoop, cette introduction revient sur les origines du projet et détaille les problématiques « Big Data » auxquelles les entreprises sont confrontées. A l'issue de ce module le stagiaire a une vision claire des tenants et des aboutissants du projet Hadoop.

Thèmes abordés :

- Enjeux et limites des systèmes actuels
- Quels besoins ?
- Approche « Big Data »

- **Hadoop : concept et bases**

Objectifs :

Ce module présente l'architecture interne de Hadoop notamment les deux composants principaux que sont HDFS et MapReduce. La fin de cette partie sera consacrée à l'écosystème Hadoop et aux projets qui gravitent autour. A la fin de ce module, le stagiaire possède les connaissances nécessaires pour comprendre et utiliser un environnement Hadoop.

Thèmes abordés :

- Projet Apache
- Système de fichiers distribués : HDFS
- Traitements distribués : MapReduce
- Présentation de l'écosystème Hadoop
- Exercices : « Premiers pas »

- **Développer un programme MapReduce (Java)**

Objectifs :

Point clé de la formation, le stagiaire va être amené à développer des programmes MapReduce. Le fonctionnement de l'algorithme et sa mise en œuvre seront détaillés tout au long de ce module. Le stagiaire possédera les compétences nécessaires pour écrire, compiler et lancer un programme MapReduce sur un cluster Hadoop.

Thèmes abordés :

- Processus MapReduce
- Développer des drivers, mappers et reducers en Java
- Développer en d'autres langages
- BestPractices : Tests unitaires
- Exercices : « Ecrire un programme MapReduce »

- **Détail de l'API Hadoop**

Objectifs :

Module faisant suite aux premiers développements, le stagiaire utilisera les fonctions avancées de l'API Hadoop pour être plus efficace et plus performant dans la création d'application. Ce module est centré sur la qualité et l'optimisation des traitements pour le code MapReduce. Le stagiaire sera en mesure de poser un diagnostic et comprendre les axes d'améliorations des développements MapReduce.

Thèmes abordés :

- Réduire les données intermédiaires : Combiner
- Améliorer le load-balancing : Partitionner
- Mise en cache et accès à HDFS
- Debugger efficacement votre code
- Exercices : « Optimisation des développements »

- **Typage des données**

Objectifs :

Tour d'horizon des différentes techniques liées aux données. Ce module porte sur 3 axes : le typage des données, la compression et la sérialisation de ces données. A la fin de ce module, le stagiaire peut créer ses propres types de données et peut choisir la méthode de compression et de sérialisation adéquate.

Thèmes abordés :

- Fonctionnement interne
- Créer ses propres formats
- Compression et sérialisation des données
- Exercices : « Utiliser les types de données »

- **Algorithmes et jointure des données**

Objectifs :

Découvrir les algorithmes qui sont le plus couramment utilisés pour traiter les données. Plusieurs techniques de jointure de données seront présentées. Au terme de ce module, le stagiaire a une vue globale des différentes techniques et est en mesure de choisir la plus adaptée à son besoin.

Thèmes abordés :

- Tri et recherche dans les données
- Indexer les données
- Algorithmes fréquents : TF-IDF, cooccurrence, etc.
- Jointure de datasets
- Exercices : « Implémenter TF-IDF »

- **Intégration SI**

Objectifs :

Ce module présente les différents aspects pour permettre l'intégration d'un environnement Hadoop au sein d'un système d'information existant. La première approche est centrée sur la cohabitation de SGBDR avec Hadoop, tandis que la seconde est centrée sur le flux de données « temps-réel ».

Thèmes abordés :

- Concepts et enjeux
- SGBDR vers HDFS : Sqoop
- Intégration de flux en temps réel : Flume

- **Ecosystème Hadoop**

Objectifs :

Tour d'horizon des différentes « briques » Hadoop permettant d'aller encore plus loin avec Hadoop. Trois points sont mis en avant dans ce module : l'interaction SQL avec Hadoop, l'utilisation d'un « langage de flux » dédié et un aperçu des techniques de datamining optimisées pour Hadoop.

Thèmes abordés :

- Approche SQL : Hive
- Approche flux de données : Pig
- Analyse avancée : Mahout
- Exercices : « Découverte de l'écosystème Hadoop »

- **Conclusion**