Développez vos talents
Certifiez vos compétences

Big Data - Développement

Hadoop - Développement (HDFS et MapReduce)

2 jours (14h00) | ★★★★½ 4,3/5 | HAD-DEV | Évaluation qualitative de fin de stage | Formation délivrée en présentiel ou distanciel [1]

Formations Informatique › Big Data › Big Data - Développement

À l'issue de ce stage vous serez capable de :

- Présenter les principes du Framework Hadoop
- Utiliser la technologie MapReduce pour paralléliser des calculs sur des volumes importants de données
- Identifier les commandes shell courantes pour HDFS.

Niveau requis

Avoir la connaissance d'un langage de programmation objet comme Java et du scripting.

Public concerné

Développeurs, BI, ETL, architectes et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop.

Cette formation :

- Est animée par un consultant-formateur dont les compétences techniques, professionnelles et pédagogiques ont été validées par des diplômes et/ou testées et approuvées par l'éditeur et/ou par M2i Formation
- Bénéficie d'un suivi de son exécution par une feuille de présence émargée par demi-journée par les stagiaires et le formateur.

(1) Modalité et moyens pédagogique :

Formation délivrée en présentiel ou distanciel * (e-learning, classe virtuelle, présentiel à distance). Le formateur alterne entre méthodes ** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation). La validation des acquis peut se faire via des études de cas, des quiz et/ou une certification.

Les moyens pédagogiques mis en oeuvre (variables suivant les formations) sont : ordinateurs Mac ou PC (sauf pour les cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel). Environnements de formation installés sur les postes de travail ou en ligne. Supports de cours et exercices.

* Nous consulter pour la faisabilité en distanciel. ** Ratio variable selon le cours suivi.

Programme

Introduction

- Big Data, introduction
- Les métiers du Big Data
- Big Data, architecture
- Les fonctionnalités du framework Hadoop
- Hadoop, l'écosystème
  - Hadoop Common
  - HDFS
  - YARN
  - Spark
  - MapReduce
  - L'ingestion de données : Kafka, Nifi

MapReduce

- Principe et objectifs du modèle de programmation MapReduce
- Fonctions "map" et "reduce"
- Couples (clés et valeurs)
- Implémentation par le framework Hadoop
- Etude de la collection d'exemples
- Rédaction d'un premier programme et exécution avec Hadoop

Programmation MapReduce

- Configuration des jobs
- Notion de configuration
- Les interfaces principales
  - Mapper
  - Reducer
- La chaîne de production
  - Entrées
  - Input splits
  - Mapper
  - Combiner
  - Shuffle / sort
  - Reducer
  - Sortie
  - Partitioner
  - OutputCollector
  - Codecs
  - Compresseurs
- Format des entrées et sorties d'un job MapReduce
  - InputFormat
  - OutputFormat
- Type personnalisé : création d'un Writable spécifique
- Utilisation
- Contraintes
- Répartition du job sur la ferme au travers de YARN

Streaming

- Définition du streaming MapReduce
- Création d'un job MapReduce dans Python
- Répartition sur la ferme
- Avantages et inconvénients

- Liaisons avec des systèmes externes
- Introduction au pont Hadoop
- Suivi d'un job en streaming

HDFS

- Concept de HDFS
- Architecture
- NameNode et DataNode
- Communications
- Gestionnaire et équilibreur de blocs
- Vérification de l'état / sécurité
- Interaction de ligne de commande avec HDFS
- Import/Export de données externes (fichiers, BDDR, CSV) vers HDFS
- Manipulation des fichiers HDFS
- Données hors HDFS (Hbase)

Hadoop, analyse de données

- Apache
  - Hive
  - Pig
  - Impala
- Différence entre Hive, Pig et Impala

Modalités d'évaluation des acquis

- En cours de formation, par des études de cas ou des travaux pratiques
- Et, en fin de formation, par un questionnaire d'auto-évaluation ou une certification (M2i ou éditeur)