

• **DURÉE**

3 jours / 21h

APPRÉCIATION

Évaluation qualitative de fin de stage

MODALITÉS

ET MOYENS

PÉDAGOGIQUES

Démonstrations, cas pratiques, synthèse et évaluation des acquis

► **Objectifs**

pédagogiques

- Comprendre le fonctionnement de MapReduce
- Comprendre le fonctionnement de Spark
- Pouvoir développer des applications distribuées dans un environnement Hadoop

► **Public concerné**

Chef de projets, Développeur, Ingénieur d'études, Architecte technique et toute personne souhaitant développer des applications distribuées dans un environnement Hadoop

► **Prérequis**

Notions en programmation Java et/ou Python et SQL

Big Data DÉVELOPPEMENT MAPREDUCE & SPARK SUR HADOOP

2 030 € HT

Inter-Entreprises
Intra-entreprises sur devis

Hadoop est un framework libre et Open Source destiné à faciliter la création d'applications distribuées pour le stockage et le traitement d'immenses volumes de données. Cette formation vous permettra d'acquérir les compétences nécessaires pour développer des solutions compatibles avec la plateforme Hadoop. Vous apprendrez à développer des applications MapReduce et Spark sous Hadoop.

Introduction

- Introduction au Big Data
- Introduction à Hadoop
- Introduction à Spark

La plateforme Hadoop

- Architecture et composants de la plateforme Hadoop
- HDFS: NameNode / DataNode
- Paradigme MapReduce et YARN
- Les technologies émergentes

MapReduce

- Détailler l'approche MapReduce
- Programmation MapReduce
- Configuration des jobs
- Comment interagir avec les jobs MapReduce
- Principales Interfaces utilisateurs: Mapper, Reducer, Partitioner, Counter
- Les entrées / sorties d'une application MapReduce
- Comment réaliser des tests unitaires d'applications MapReduce
- Comment utiliser des combinateurs MapReduce, des partitioners et des caches distribués
- Meilleures pratiques pour développer et déboguer des applications MapReduce

Apache Spark

- Concept et principe
- Pourquoi Spark ?
- Spark vs MapReduce
- Architecture et fonctionnement de Spark

Programmation avec Spark

- Les bases de Spark
- Les RDD
- La programmation parallèle avec Spark
- Partitionnement, jobs, stage et tasks
- Changer le niveau de parallélisation
- Soumettre une application Spark à un cluster
- Manipulation des données avec Spark : Spark SQL, DataFrames et DataSets
- Spark Streaming

Travaux pratiques

Exercices de mise en pratique de la programmation MapReduce et de la programmation Spark

