Sparks.

Programme de formation
Apache Spark

• Objectifs
Cette formation vous permettra d'explorer les applications de Spark, la solution Big Data d'Apache. Apprenez à exploiter la polyvalence et les capacités de ce framework pour concevoir des applications efficaces et optimisées de traitements de données pour satisfaire au mieux vos besoins dans de nombreux secteurs d'activités. Créez des applications sophistiquées pour analyser et travailler sur une grande variété de données avec des techniques de traitement par lots, de DataStreaming et de Machine Learning.

• Pré requis
Bases en langage Java, Scala ou Python, connaissances sur Apache Hadoop

• Durée
3 jours

• Public
Architectes-techniques, Developpeurs, DSI

•Plan de formation

Introduction à Apache Spark
Quelles solutions apportent Spark au Big Data ?
Principes de base du fonctionnement de Spark
Différences et complémentarités avec Apache Hadoop
Spécification de Spark Shell
Environnement et outils de Spark

Fonctionnement et utilisation des RDD (Resilient Distributed Datasets)
Gérer les opérations de RDD
RDD et MapReduce
Spark SQL

Combiner Spark au Système de Fichiers Distribués Hadoop (HDFS)
Intérêts de l'utilisation du HDFS dans Spark
Intégrer le HDFS dans l'architecture Spark
Utiliser le HDFS

Spark en cluster
Créer la structure en clusters
Hébergement et déploiement
Interface Web de Spark

Partitionnement et programmation parallèle
Localiser les données du HDFS
Partitionner les RDD
Programmer et exécuter les opérations parallèles
Mettre en cache le partitionnement des données
Gérer la persistance des données

Concevoir une application avec Spark
Présentation et configuration des propriétés de Spark
Prototypage d'opérations avec Spark Shell
Compilation et génération d'une application

Traiter les données en temps réel avec Spark Streaming
Fonctionnement et concepts de base
Notions de DStream
Intégration, transformation et opérations de sorties des DStreams
Gestion des performances

Machine Learning et implémentation d'algorithmes
Algorithmes itératifs et résolution de

problèmes
Machine Learning Library MLlib
Opérations sur les données graphiques

Optimisation de Spark
Gestion des variables partagées
Données broadcastées
Accumulateurs
Méthodes et outils d'optimisation des performances