

Big Data avec SPARK

Date et durée
Code formation : BDT01FR Durée : 5 jours Nombre d'heures : 35 heures
Description
<p>Le Big Data est considéré aujourd'hui comme l'un des plus grands défis informatiques de notre époque. Il représente un tournant pour les organisations au moins aussi important qu'Internet en son temps. Les plus enthousiastes y voient même une révolution industrielle comparable à la découverte de l'électricité au 19ème siècle ou de l'informatique (fin 20ème siècle).</p> <p>Le Big Data et les analytics sont utilisés dans presque tous les domaines et concernent les organisations de toutes tailles. C'est devenu au fil des années un enjeu économique et stratégique important pour les entreprises qui répond généralement à plusieurs objectifs comme l'amélioration de l'expérience client, l'optimisation des processus et de la performance opérationnelle, le renforcement ou diversification du business model.</p> <p>C'est ainsi que de plus en plus d'entreprises recherchent des personnes capables d'analyser et gérer les quantités de données générées par leurs activités sur les réseaux.</p> <p>Apache Spark est considéré aujourd'hui comme le framework le plus abouti et le plus utilisé pour les problématiques d'analyse des données à large échelle dans le monde. En somme, c'est une solution qui s'avère être le successeur de MapReduce, d'autant qu'il a l'avantage de fusionner une grande partie des outils nécessaires dans un cluster Hadoop.</p>
Objectifs
<p>A l'issue de la formation les participants seront capable de :</p> <ul style="list-style-type: none">• Maîtriser les concepts fondamentaux de Spark• Développer des applications avec Spark• Découvrir et comprendre les RDD• Explorer et manipuler des données à l'aide de Zeppelin• Exploiter des données avec Spark SQL• Comprendre le fonctionnement de Spark MLlib• Construire des modèles prédictifs avec Spark-ML• Mise en place d'un cluster Spark
Pré-requis
<p>Aucune connaissance sur Spark n'est requise. Une bonne connaissances du langage Java est requise.</p>
Public
<p>Cette formation s'adresse aux développeurs, Data scientists, architectes système et responsables techniques qui veulent déployer des solutions Spark dans leur entreprise.</p>

Cette formation s'adresse aux profils suivants

Développeur

Directeur des Systèmes d'Information (DSI)

Programme

Pourquoi Spark

- Introduction
- Les problématiques du BigData
- La révolution MapReduce
- MapReduce versus Spark
- Apache Hadoop et son écosystème
- HDFS et le Stockage de fichiers sur Hadoop : Namenode et le Datanode
- YARN et le processing des données sur un cluster Hadoop
- Dimensionner et configurer un cluster.
- SQOOP et l'import de données sur Hadoop

Comprendre et Développer sous Spark

- Les bases
- La programmation fonctionnelle avec Scala
- La programmation parallèle avec Scala
- Découvrir et comprendre les RDD
- Agréger les données avec les « paired RDD »
- Ecrire et exécuter des applications Spark
- Transformations et actions
- Configurer des applications Spark
- Exécuter des traitements dans un environnement distribué
- Cycle de vie d'un RDD
- Processing de données avec Spark
- DataFrame et Spark SQL
- Travailler avec Zeppelin
- Fonctionnalités avancées et amélioration des performances
- Brève initiation à Apache Flume et Apache Kafka

Spark pour la Data Science

- Introduction au Machine Learning
- Apprentissage supervisé et non supervisé
- Test et évaluation
- Les différentes classes d'algorithmes
- Présentation de SparkML et MLlib
- Implémentations des différents algorithmes dans MLlib
- Conclusion.

Travaux Pratiques

- Installation et configuration de Spark
- Premiers pas avec Spark
- Manipulation de différents Datasets à l'aide de RDD
- Manipulation de Datasets via des requêtes SQL
- Connexion avec une base externe via JDBC
- Mise en place d'un cluster Spark
- Utilisation de SparkML et MLlib

- Les commandes HDFS
- Utiliser le stockage HDFS
- Programmation parallèle sur Spark - Cacher et persister la donnée
- Les accumulateurs pour vérifier la qualité des données et l'utilisation des variables « broadcast »
- Partitionnement avancé et opérations, point de départ pour l'optimisation
- Manipuler les données avec Zeppelin
- SparkSQL avec utilisation d'UDF
- SparkSQL avec Hive
- SparkSQL et les requêtes.