

Formation Spark – développement des Applications pour le Big Data

Code : NTE15

Durée : 5 jours

Classe : Présentiel / à distance

Public

- Cette formation s'adresse aux développeurs, chefs de projet, architectes, ingénieurs data et passionnés du Big Data souhaitant développer des applications distribuées avec Spark.

Prérequis

- Bases en Java ou Python
- Connaissances en développement informatique
- Notions de bases de données
- Culture générale Big Data appréciée

Objectifs

- Comprendre la structure d'Apache Spark
- Utiliser Spark SQL pour la manipulation des données
- Développer des applications distribuées de Big Data
- Explorer Spark Streaming et Spark MLlib

Programme détaillé

1- Introduction à Hadoop et son écosystème

- Introduction générale à hadoop
- La place de mapreduce
- Le traitement de données avec Hadoop
- Les composants d'un cluster Hadoop
- Un système de fichiers distribué (HDFS)
- Traitement distribué sur un cluster Hadoop (mapreduce)
- Travailler avec Yarn
- En quoi Spark complète-t-il Hadoop ?

2- Architecture de Spark

- Un framework offrant de nombreux service mais pas de stockage (Hadoop, AWS S3, Cassandra, MongoDB, etc.)
- Rôle du coeur de Spark (moteur)
- RDD, la couche d'abstraction des données (Resilient Distributed Datasets)
- Accéder aux données avec Spark SQL
- Traiter les données en pseudo temps réel avec Spark Streaming
- Développer des applications distribuées de machine learning (Spark MLlib)
- Quels liens entre Spark et les langages de programmation (Java, Python, R, ...) ?
- Manipuler les graphes avec GraphX
- Limites de Spark

3- Les RDD, structures fondamentales de Spark

- Introduction aux RDD
- Les sources de données de RDD
- Créer et sauvegarder des RDD
- Les opérations sur les RDD
- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark
- Les RDD clé-valeur
- Map-Reduce : principe et usage dans Spark
- Autres opérations sur les RDD de paires
- Exécuter des requêtes SQL (Spark SQL)
- Interopérabilité avec les RDD



Formation Spark – développement des Applications pour le Big Data

Code : NTE15

Durée : 5 jours

Classe : Présentiel / à distance

4- Manipuler les données avec les Dataframe et Datasets

- Créer des DataFrames depuis diverses sources de données
- Les schémas des DataFrames
- Afficher le Dataframe en mode texte (take)
- Visualiser graphiquement le DataFrame (display)
- Sauvegarder des DataFrames
- Requêter des DataFrames avec des expressions sur les colonnes nommées
- Les requêtes de groupement et d'aggrégation
- Les jointures
- Les différences entre Datasets et DataFrames
- Créer des Datasets
- Charger et sauvegarder des Datasets
- Les opérations sur les Datasets
- Conversion entre RDD et DataFrames

5- Machine learning avec Spark

- Introduction au machine learning
- Les différentes classes d'algorithmes
- Présentation de SparkML et MLlib
- Implémentations des différents algorithmes dans MLlib

6- Analyser en temps réel avec Spark Streaming

- Comprendre l'architecture du streaming
- Présentation des Discretized Streams (DStreams)
- Les différents types de sources
- Manipulation de l'API (agrégations, watermarking...)
- Machine Learning en temps réel

7- Écriture d'une application compilée

- Écrire, configurer et lancer des applications spark
- Écrire une application Spark
- Compiler et lancer une application
- Le mode de déploiement d'une application
- L'interface utilisateur web des applications Spark
- Configurer les propriétés d'une application

