# Formation Spark - développement des Applications pour le Big Data

**Code** : NTE15 **Durée** : 5jours **Classe** : Présentiel/àdistance

## Public

- Ce cours Spark s'adresse à des chefs de projet, développeurs, data scientists ou architectes.

## Prérequis

- Pour suivre cette formation Spark, il est essentiel d'avoir des connaissances de base en développement dans les univers Java ou Python. Si vous connaissez un autre langage vous serez moins autonome pour réaliser les TP mais la formation gardera un sens au niveau des concepts et des librairies présentées.

## Objectifs

### Objectif opérationnel :

- Savoir utiliser Spark pour intégrer des données, les manipuler et utiliser les outils appropriés à chaque situation.

### Objectifs pédagogiques :

Plus concrètement, à l'issue de cette formation Spark , vous aurez acquis les connaissances et compétences nécessaires pour :

- Comprendre la philosophie de Spark et ses limites
- Utiliser Spark avec Hadoop
- Développer avec Spark streaming pour de l'analyse de flux en temps réel
- Développer des applications réparties avec Spark (parallélisme sur Cluster)
- Accéder à des données structurées dans vos applications (Spark SQL)
- Découvrir le machine learning avec Spark ML

## 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 services...
- ... 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 MLib)
- 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

# Formation Spark - développement des Applications pour le Big DataMicrosoft Azure AI

**Code** : NTE15 **Durée** : 5jours **Classe** : Présentiel/àdistance

- 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

**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