# Data Science
## SPARK POUR LA DATA SCIENCE

**2 030 € HT**
Inter-Entreprises
*Intra-entreprises sur devis*

La Data Science est aujourd'hui centrale dans les entreprises digitalisées. Le but de cette formation est de se familiariser avec Spark et de comprendre son utilisation pour traiter des problèmes de Data Science

● **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**
- Se familiariser avec Spark
- Apprendre à aborder les problématiques relatives à la Data Science avec Spark
- Comprendre le principe de fonctionnement de Spark et découvrir les principales librairies (Streaming, SQL, Machine Learning, GraphX).

**► Public concerné**
Développeurs, Data Analystes, Data Scientists, et toute personne souhaitant acquérir des connaissances dans le domaine de la Data Science et Spark.

**► Prérequis**
Une première expérience en programmation.
Avoir des connaissances en SQL, mathématiques et statistiques.

### Introduction à la Data Science
- Qu'est-ce que la Data Science ?
- Définition
- Domaines d'application de la Data Science
- Outils et algorithmes pour la Data Science
- De l'analyse statistique au Machine Learning
- Enjeux de la Data Science

### Introduction Spark
- Qu'est-ce que Spark ?
- Fonctionnement : RDD, DataFrames et DataSets
- Comment interagir avec Spark ?
- Programmer avec Spark : APIs Java, Python, Scala

### Manipulation des données
- Formats basiques (fichiers textes, JSON, CSV, SequencesFiles, fichiers compressés)
- Interagir avec des sources de données externes : connecteurs Hive, JDC, HBase, ElasticSearch…

### Spark Streaming
- Introduction à Spark Streaming
- La notion de « DStream »
- Principales sources de données
- Utilisation de l'API
- Manipulation des données

### Spark SQL
- Initiation à Spark SQL
- Création de DataFrames
- Manipulation des DataFrames (opérations basiques, agrégations et Groupby, Missing Data)
- Chargement et stockage de données (avec Hive, JSON, etc.)

### Spark ML avec MLlib
- Modélisation statistique et apprentissage
- Types de données (Vector, LabeledPoint, Model)
- Préparation des données
- Utilisation d'algorithmes de MLlib (k-means, régression logistique, arbre de discrimination, forêt aléatoire)
- Exemple de création d'un modèle d'évaluation avec Spark MLlib sur un jeu de données

### GraphX et GraphFrames
- Présentation de GraphX
- Principe de création des graphes
- API GraphX
- Présentation GraphFrames
- GraphX vs GraphFrames

### Travaux pratiques
Alternance d'apports théoriques, d'exercices pratiques et de mise en situation sous forme de travaux pratiques permettant de tester les différentes notions abordées.