

<b>Stage pratique de 3 jour(s)</b> <b>Réf : SPK</b>
<b>Participants</b> Développeurs, architectes.
<b>Pré-requis</b> Bonnes connaissances du langage Java.
<b>Prix 2020 : 2040€ HT</b>
<b>Dates des sessions</b> <b>AIX</b> 06 avr. 2020, 29 juin 2020 05 oct. 2020, 07 déc. 2020 <b>BORDEAUX</b> 30 mar. 2020, 27 juil. 2020 28 sep. 2020, 30 nov. 2020 <b>LILLE</b> 23 mar. 2020, 20 juil. 2020 21 sep. 2020, 23 nov. 2020 <b>LYON</b> 23 mar. 2020, 20 juil. 2020 12 oct. 2020, 23 nov. 2020 07 déc. 2020 <b>NANTES</b> 09 mar. 2020, 06 juil. 2020 07 sep. 2020, 23 nov. 2020 <b>PARIS</b> 27 jan. 2020, 23 mar. 2020 25 mai 2020, 20 juil. 2020 21 sep. 2020, 23 nov. 2020 <b>SOPHIA-ANTIPOLIS</b> 09 mar. 2020, 06 juil. 2020 07 sep. 2020, 23 nov. 2020 <b>STRASBOURG</b> 06 avr. 2020, 29 juin 2020 05 oct. 2020, 07 déc. 2020 <b>TOULOUSE</b> 30 mar. 2020, 27 juil. 2020 28 sep. 2020, 30 nov. 2020
<b>Modalités d'évaluation</b> L'évaluation des acquis se fait tout au long de la session au travers des multiples exercices à réaliser (50 à 70% du temps).
<b>Compétences du formateur</b> Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

## Spark, développer des applications pour le Big Data

*Vous développerez des applications en Java en vue de traiter en temps réel des données issues du Big Data. Vous collecterez, stockerez et traiterez avec Spark des données de formats hétérogènes afin de mettre en place des chaînes de traitement intégrées à votre Système d'Information.*

### OBJECTIFS PEDAGOGIQUES

Maîtriser les concepts fondamentaux de Spark  
 Développer des applications avec Spark Streaming  
 Faire de la programmation parallèle avec Spark sur un cluster  
 Exploiter des données avec Spark SQL  
 Avoir une première approche du Machine Learning

#### 1) Présentation d'Apache Spark

#### 2) Programmer avec les Resilient Distributed Dataset (RDD)

#### 3) Manipuler des données structurées avec Spark SQL

#### 4) Spark sur un cluster

#### 5) Analyser en temps réel avec Spark Streaming

#### 6) Manipuler des graphes avec GraphX

#### 7) Machine Learning avec Spark

### Travaux pratiques

*Mise en pratique des notions vues en cours à l'aide du langage Java.*

### 1) Présentation d'Apache Spark

- Historique du Framework.
- Les différentes versions de Spark (Scala, Python et Java).
- Comparaison avec l'environnement Apache Hadoop.
- Les différents modules de Spark.

#### Travaux pratiques

*Installation et configuration de Spark. Exécution d'un premier exemple avec le comptage de mots.*

### 2) Programmer avec les Resilient Distributed Dataset (RDD)

- Présentation des RDD.
- Créer, manipuler et réutiliser des RDD.
- Accumulateurs et variables broadcastées.
- Utiliser des partitions.

#### Travaux pratiques

*Manipulation de différents Datasets à l'aide de RDD et utilisation de l'API fournie par Spark.*

### 3) Manipuler des données structurées avec Spark SQL

- SQL, DataFrames et Datasets.
- Les différents types de sources de données.
- Interopérabilité avec les RDD.
- Performance de Spark SQL.
- JDBC/ODBC server et Spark SQL CLI.

#### Travaux pratiques

*Manipulation de Datasets via des requêtes SQL. Connexion avec une base externe via JDBC.*

### 4) Spark sur un cluster

- Les différents types d'architecture : Standalone, Apache Mesos ou Hadoop YARN.
- Configurer un cluster en mode Standalone.
- Packager une application avec ses dépendances.
- Déployer des applications avec Spark-submit.
- Dimensionner un cluster.

#### Travaux pratiques

*Mise en place d'un cluster Spark.*

### 5) Analyser en temps réel avec Spark Streaming

- Principe de fonctionnement.
- Présentation des Discretized Streams (DStreams).
- Les différents types de sources.
- Manipulation de l'API.
- Comparaison avec Apache Storm.

## Moyens pédagogiques et techniques

- Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
- A l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
- Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

### Travaux pratiques

*Consommation de logs avec Spark Streaming.*

## 6) Manipuler des graphes avec GraphX

- Présentation de GraphX.
- Les différentes opérations.
- Créer des graphes.
- Vertex and Edge RDD.
- Présentation de différents algorithmes.

### Travaux pratiques

*Manipulation de l'API GraphX à travers différents exemples.*

## 7) Machine Learning avec Spark

- Introduction au Machine Learning.
- Les différentes classes d'algorithmes.
- Présentation de SparkML et MLlib.
- Implémentations des différents algorithmes dans MLlib.

### Travaux pratiques

*Utilisation de SparkML et MLlib.*