

# Spark - Mise en oeuvre et programmation

3 j (21 heures)

Ref : SPAR

## Public

Chefs de projets, data scientists, développeurs

## Pré-requis

Avoir connaissance de Java ou Python, des bases Hadoop et des notions de calculs statistiques

## Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue

Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois

Un poste par stagiaire, vidéoprojecteur, support de cours fourni à chaque stagiaire

## Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur

Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires

Questionnaire d'évaluation de la satisfaction en fin de stage

Auto-évaluation des acquis de la formation par les stagiaires

Attestation de fin de formation

## Objectifs

Mettre en oeuvre Spark pour optimiser des calculs

Développer des applications avec Spark Streaming

Mettre en oeuvre un cluster Spark

## Programme détaillé

### INTRODUCTION

---

Présentation de Spark

Origine du projet

Apports

Principe de fonctionnement

Langages supportés

## PREMIERS PAS

---

Utilisation du Shell Spark avec Scala ou Python

Gestion du cache

## REGLES DE DEVELOPPEMENT

---

Mise en pratique en Java et Python

Notion de contexte Spark

Différentes méthodes de création des RDD

- Depuis un fichier texte, un stockage externe

Manipulations sur les RDD (Resilient Distributed Dataset)

- Fonctions

- Gestion de la persistance

## STREAMING

---

Objectifs

Principe de fonctionnement

Notion de StreamingContext

DStreams

Démonstrations

## CLUSTER

---

Différents cluster managers

- Spark en autonome

- Mesos

- YARN

- Amazon EC2

Architecture

- SparkContext

- Cluster manager

- Executor sur chaque nœud

Définitions

- Driver program

- Cluster manager

- Deploy mode

- Executor

- Task

- Job

Mise en oeuvre avec Spark et Amazon EC2

Soumission de jobs

Supervision depuis l'interface Web

## **INTEGRATION HADOOP**

---

Création et exploitation d'un cluster Spark / YARN

## **SUPPORT CASSANDRA**

---

Description rapide de l'architecture Cassandra

Mise en oeuvre depuis Spark

Exécution de travaux Spark s'appuyant sur une grappe Cassandra

---