

Spark Scala – Traitement des données

3 jours (21 heures)

Perfectionnement

Présentiel ou classe à distance

à partir de

2 355 € HT par personne (TVA en vigueur) **Repas inclus en présentiel**

Contenu mis à jour le 25/09/2026. Document téléchargé le 03/10/2026.

Objectifs

A l'issue de cette formation, vous serez capable de :

- Naviguer dans l'écosystème Spark en utilisant Scala comme langage de programmation.

Les compétences visées par cette formation font l'objet d'une évaluation.

À qui s'adresse cette formation ?

Statisticiens, consultants Big Data, data analysts, data scientists.

Prérequis

Connaître la programmation fonctionnelle et avoir des connaissances sur la gestion des bases de données.

Programme

1 Pendant votre formation

Spark Scala – Traitement des données

Jour 1

Introduction à Scala pour Apache Spark

- Présentation de Scala
- Pourquoi Scala avec Spark ?
- Scala dans les autres framework
- Introduction à Scala REPL
- Les opérations basiques sur Scala
- Les types de variables dans Scala
- Les structures de contrôles dans Scala
 - Les boucles
 - Les fonctions
 - Les procédures
- Les collections dans Scala (Array, ArrayBuffer, Map, Tuples, Lists...)

Introduction au Big Data et Apache Spark

- Introduction au Big Data
- Les challenges du Big Data
- Batch vs le temps réel dans le Big Data Analytics
- Analyse en Batch Hadoop
- Vue d'ensemble de l'écosystème
- Les options de l'analyse en temps réel
- Streaming Data Spark
- In-memory Data Spark
- Présentation de Spark
- Ecosystème Spark
- Les modes de Spark
- Installation de Spark
- Vue d'ensemble de Spark en cluster
- Spark Standalone cluster
- Spark Web UI

Jour 2

Les opérations communes sur Spark

- Utilisation de Spark Shell
- Création d'un contexte Spark
- Chargement d'un fichier en Shell
- Réalisation d'opérations basiques sur un fichier avec Spark Shell
- Présentation de l'environnement de développement SBT
- Créer un projet Spark avec SBT
- Exécuter un projet Spark avec SBT
- Le mode local

- Le mode Spark
- Le caching sur Spark
- Persistance distribuée

Introduction aux RDD et DataFrame

- Transformations dans le RDD
- Actions dans le RDD
- Chargement de données dans RDD
- Enregistrement des données à travers RDD
- Paire clé-valeur "RDD MapReduce" et les paires "RDD Operations"
- Intégration HDFS avec Spark et Hadoop
- Intégration YARN avec Spark et Hadoop
- Gestion des fichiers de séquences et les partitionner

Exemple de travaux pratiques (à titre indicatif)

- *Data PreProcessing avec Spark DataFrame*

Jour 3

Spark Streaming et MLlib

- Architecture de Spark Streaming
- Premier programme avec Spark Streaming
- Les transformations dans Spark Streaming
- La "fault tolerance" dans Spark Streaming
- Checkpointing
- Niveaux de parallélismes
- Machine Learning avec Spark
- Types de données
- Algorithmes et statistiques
- Classification et régression
- Clustering
- Filtrage collaboratif

GraphX, SparkSQL et amélioration des performances dans Spark

- Analyse de l'architecture de Hive et Spark SQL
- SQLContext dans Spark SQL
- Travailler avec les DataFrames
- Implémentation d'un exemple pour Spark SQL
- Intégration de Hive et Spark SQL
- Support pour JSON et les formats des "Parquet File"
- Implémentation de la Data Visualization avec Spark
- Chargement de données

- Les requêtes Hive à travers Spark
- Les techniques de tests dans Scala
- Les astuces d'amélioration de performance dans Spark
- Les variables partagées
- Diffusion des variables
- Partage de variables
- Accumulateurs

Financement

Plusieurs dispositifs de financement sont accessibles via les OPCO (Opérateurs de Compétences), organismes agréés par le ministère du Travail dont le rôle est d'accompagner, collecter et gérer les contributions des entreprises au titre du financement de la formation professionnelle.

Pour plus d'information, une équipe de gestionnaires SKOLAE Formation spécialisée vous accompagne dans le choix de vos formations et la gestion administrative.

Nous contacter pour en savoir plus →

Modalités pédagogiques

Formation délivrée en présentiel ou distanciel* (blended-learning, e-learning, classe virtuelle, présentiel à distance).

Le formateur alterne entre méthode** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation).

Variables suivant les formations, les moyens pédagogiques mis en oeuvre sont :

- Ordinateurs Mac ou PC (sauf pour certains cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel)
- Environnements de formation installés sur les postes de travail ou en ligne
- Supports de cours et exercices

En cas de formation intra sur site externe à M2I, le client s'assure et s'engage également à avoir toutes les ressources matérielles pédagogiques nécessaires (équipements informatiques...) au bon déroulement de l'action de formation visée conformément aux prérequis indiqués dans le programme de formation communiqué.

* nous consulter pour la faisabilité en distanciel

** ratio variable selon le cours suivi

Accessibilité aux personnes en situation de handicap

Détails de la session

Du 21 octobre 2026 au 23 octobre 2026 (3 jours / 21 heures)

Angers

Du 21 octobre 2026 au 23 octobre 2026

Angers

- 21 octobre 2026 ----- 9h00 – 16h00
- 22 octobre 2026 ----- 9h00 – 16h00
- 23 octobre 2026 ----- 9h00 – 16h00