

PySpark – 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 02/10/2026.

Objectifs

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

- Décrire le principe de fonctionnement de Spark
- Utiliser l'API PySpark pour interagir avec Spark en Python
- Mettre en oeuvre les méthodes de Machine Learning avec la librairie MLlib de Spark
- Traiter les flux de données avec Spark Streaming
- Manipuler les données avec Spark SQL.

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

À qui s'adresse cette formation ?

Développeurs, Data analysts, Data scientists, architectes Big Data et toute personne souhaitant acquérir des connaissances dans le domaine de la Data Science et sur Spark.

Prérequis

Avoir des notions de SQL et des connaissances de base en mathématiques et statistiques. Une première expérience en programmation Python est requise.

Programme

1 Pendant votre formation

PySpark – Traitement des données

Jour 1

Introduction à Hadoop

- L'ère du Big Data
- Architecture et composants de la plateforme Hadoop
- HDFS
- NameNode / DataNode / ResourceManager
- MapReduce et YARN

Introduction à Spark

- Qu'est-ce que Spark ?
- Spark vs MapReduce
- Fonctionnement
 - RDD
 - DataFrames
 - Data Sets
- Comment interagir avec Spark ?
- PySpark : programmer avec Spark en Python

Installation de Spark

- Sur une infrastructure distribuée
- En local
- En Cloud (présentation avec Amazon AWS et Microsoft Azure)

Jour 2

Spark pour la manipulation des données – PySpark

- Utilisation de SparkSQL et des DataFrames pour manipuler des données
- Charger des données depuis Hadoop, depuis des fichiers CSV, texte, JSON...
- Transformer des données (création de DataFrames, ajout de colonnes, filtres...)

Exemple de travaux pratiques (à titre indicatif)

- *Chargement et modifications de données avec Spark et PySpark*

L'utilisation de spark.ml pour le Machine Learning

- Apprentissage supervisé

- Forêts aléatoires avec Spark
- Mise en place d'un outil de recommandation
- Traitement de données textuelles
- Automatiser vos analyses avec des pipelines

Jour 3

Spark Streaming

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

Spark SQL

- Initialisation à Spark SQL
- Création de DataFrames
- Manipulation des DataFrames (opérations basiques, agrégations et groupBy, missing data)
- Chargement et stockage de données (avec Hive, JSON...)

GraphX et GraphFrames

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

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 4 novembre 2026 au 6 novembre 2026 (3 jours / 21 heures)

Toulouse

Du 4 novembre 2026 au 6 novembre 2026 **Toulouse**

- 4 novembre 2026 ----- 9h00 – 16h00
- 5 novembre 2026 ----- 9h00 – 16h00
- 6 novembre 2026 ----- 9h00 – 16h00