Apache Spark : maîtrisez le traitement de données distribuées et devenez expert Big Data

Lien : https://innov-systems.com/formation/apache-spark-maitrisez-le-traitement-de-donnees-distribuees-et-devenez-expert-big-data

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI15

CATÉGORIE
Big Data : développement, langages et NOSQL

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre l’architecture d’Apache Spark
- Maîtriser le traitement distribué des données
- Savoir utiliser RDD, DataFrame et Dataset
- Exploiter Spark SQL pour l’analyse de données
- Traiter des données en batch et en streaming
- Utiliser Spark MLlib pour le machine learning
- Optimiser les performances des jobs Spark
- Déployer des solutions Big Data scalables

POUR QUI ?

- Data engineers
- Data scientists
- Développeurs Big Data
- Architectes data
- Ingénieurs logiciels
- Analystes de données
- DevOps / MLOps engineers
- Toute personne travaillant sur le traitement de données massives

Innov Systems

Programme détaillé

1 / Introduction à Apache Spark

- Définition et rôle de Spark dans le Big Data
- Comparaison avec Hadoop MapReduce
- Cas d’usage industriels

2 / Architecture de Spark

- Driver, Executors et Cluster Manager
- Fonctionnement du traitement distribué
- Résilience et tolérance aux pannes

3 / RDD (Resilient Distributed Dataset)

- Concept de RDD
- Transformations et actions
- Lazy evaluation

4 / DataFrames et Dataset API

- Structure des DataFrames
- Optimisation des requêtes
- API Spark SQL

5 / Spark SQL

- Requêtes SQL sur données distribuées
- Intégration avec bases de données
- Optimisation des performances

Innov Systems

6 / Spark Streaming

- Traitement de données en temps réel
- Micro-batching
- Sources de streaming (Kafka, files, sockets)

7 / Machine Learning avec Spark MLlib

- Introduction au machine learning distribué
- Algorithmes de classification et régression
- Pipelines ML

8 / Optimisation des performances Spark

- Partitionnement des données
- Cache et persistance
- Optimisation des jobs Spark

9 / Intégration avec l’écosystème Big Data

- Hadoop, Hive, Kafka
- Cloud (Azure, AWS, GCP)
- Data lakes et data warehouses

10 / Bonnes pratiques et déploiement

- Gestion des ressources cluster
- Monitoring des jobs Spark
- Industrialisation des pipelines data

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

07 au 11 Déc. 2026	Présentiel - Casablanca
18 au 22 Jan. 2027	Présentiel - Casablanca
15 au 19 Mars 2027	Présentiel - Casablanca
10 au 14 Mai 2027	Présentiel - Casablanca
05 au 09 Juil. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems

Innov Systems — Tous droits réservés