Apache Spark & PySpark — Traitement distribué de données massives

Lien : https://innov-systems.com/formation/apache-spark-pyspark-traitement-distribue-de-donnees-massives

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI166

CATÉGORIE
Big Data - Développement

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 de données massives
- Utiliser PySpark pour le développement de pipelines data
- Manipuler DataFrames et RDD efficacement
- Optimiser les performances des jobs Spark
- Implémenter des traitements batch et streaming
- Intégrer Spark dans des architectures Big Data
- Développer des workflows de transformation de données
- Exploiter Spark SQL pour l’analyse de données
- Gérer des volumes de données à grande échelle
- Déployer des applications Spark en production
- Étudier des cas concrets de Big Data processing

POUR QUI ?

- Data engineers
- Data scientists
- Big Data engineers
- Machine learning engineers
- Backend developers
- DevOps engineers
- Architectes data
- SRE (Site Reliability Engineers)
- Analysts data
- Consultants Big Data
- Toute personne impliquée dans le traitement de données massives

Innov Systems

Programme détaillé

1 / Introduction à Apache Spark

- Présentation du framework de calcul distribué
- Positionnement dans l'écosystème Big Data
- Cas d'usage du traitement de données massives

2 / Architecture de Spark

- Driver, Executors et Cluster Manager
- Mode local vs cluster (YARN, Kubernetes, Standalone)
- Résilience via DAG et RDD lineage

3 / Fondamentaux du modèle de calcul Spark

- RDD (Resilient Distributed Dataset)
- DataFrames et Datasets
- Lazy evaluation et optimisation des plans

4 / PySpark : utilisation de Spark avec Python

- API PySpark et environnement de développement
- Manipulation de DataFrames
- Intégration avec l'écosystème Python (Pandas, NumPy)

5 / Traitement de données structurées et non structurées

- Nettoyage et transformation des données
- Parsing de fichiers massifs (JSON, CSV, Parquet)
- Gestion des données semi-structurées

Innov Systems

6 / Spark SQL et requêtage distribué

- Requêtes SQL sur dataframes
- Optimisation via Catalyst optimizer
- Jointures et agrégations à grande échelle

7 / Streaming de données avec Spark

- Spark Structured Streaming
- Traitement en temps réel vs batch
- Cas d’usage IoT et event streaming

8 / Performance et optimisation

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

9 / Intégration avec l’écosystème Big Data

- Hadoop, Hive et HDFS
- Cloud data platforms (AWS, Azure, GCP)
- Data lake et architectures modernes

10 / Cas pratiques et projets réels

- Pipeline de traitement de données massif
- Analyse de logs à grande échelle
- Déploiement de jobs Spark en production

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

08 au 12 Fév. 2027	Présentiel - Casablanca
05 au 09 Avr. 2027	Présentiel - Casablanca
31 Mai au 04 Juin 2027	Présentiel - Casablanca
26 au 30 Juil. 2027	Présentiel - Casablanca
20 au 24 Sep. 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 — Tous droits réservés