Spark Scala : traitement des données

Lien : https://innov-systems.com/formation/spark-scala-traitement-des-donnees

DURÉE
4 jours (28h)

RÉFÉRENCE
BSI21

CATÉGORIE
Big Data : développement, langages et NOSQL

OBJECTIFS DE LA FORMATION

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

- Développer des applications avec Apache Spark en utilisant Scala comme language de programmation
- Découvrir MLLib pour du machine learning sur Spark
- Optimiser une application Spark

POUR QUI ?

- Statisticiens
- Data analysts
- Data scientists

Innov Systems

Programme détaillé

1 / Introduction au Big Data

- Introduction au Big Data
- Les challenges du Big Data
- Batch vs le temps réel dans le Big Data Analytics
- Vue d'ensemble de l'écosystème
- Les options de l'analyse en temps réel
- Les solutions Spark apportées au Big Data

2 / Introduction à Spark

- Introduction au Big Data
- Historique du Framework
- Streaming Data - Spark
- In-memory Data - Spark
- Présentation de Spark
- Origine du projet
- Principes de base du fonctionnement de Spark
- Environnement et outils de Spark
- Comparaison avec l'environnement Apache Hadoop
- Les différents modules de Spark
- Langages supportés
- Ecosystème Spark
- Les modes de Spark
- Installation de Spark
- Vue d'ensemble de Spark en cluster
- Spark Standalone cluster
- Spark Web UI

Innov Systems

3 / 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 (boucles, fonctions, procédures)
- Les collections dans Scala (Array, ArrayBuffer, Map, Tuples, Lists...)

4 / 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

5 / 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"

Inov Systems

- Intégration HDFS avec Spark et Hadoop
- Intégration YARN avec Spark et Hadoop
- Gestion des fichiers de séquences et les partitionner

6 / Apache spark 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

7 / Manipuler des graphes avec GraphX

- 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

Inov Systems

- Les variables partagées
- Diffusion des variables
- Partage de variables
- Accumulateurs

8 / Optimisation de Spark

- Gestion des variables partagées
- Données broadcastées
- Accumulateurs
- Méthodes et outils d’optimisation des performances

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

02 au 05 Fév. 2027	Présentiel - Casablanca
30 Mars au 02 Avr. 2027	Présentiel - Casablanca
25 au 28 Mai 2027	Présentiel - Casablanca
20 au 23 Juil. 2027	Présentiel - Casablanca
21 au 24 Sep. 2027	Présentiel - Casablanca

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

Innov Systems

Réservation & Renseignements

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

Innov Systems — Tous droits réservés

Innov Systems