EduGroupe
Accompagner pour réussir

APACHE SPARK POUR LES DÉVELOPPEURS JAVA

Durée : 3 jours (21 heures)

CONNAISSANCES PREALABLES

- Bonne maîtrise du langage Java
- Connaissances en programmation orientée objet
- Connaissances de base en SQL
- Familiarité avec Maven ou Gradle
- Une première connaissance des architectures distribuées constitue un avantage.

PROFIL DES STAGIAIRES

- Développeurs Java
- Développeurs Java EE / Spring
- Développeurs Big Data
- Data Engineers débutants
- Architectes techniques
- Consultants techniques
- Équipes de développement souhaitant migrer vers des traitements Big Data
- Toute personne maîtrisant Java et souhaitant développer avec Apache Spark.

OBJECTIFS

À l'issue de cette formation, les participants seront capables de :

- Comprendre les principes du calcul distribué avec Apache Spark
- Développer des applications Spark en Java
- Manipuler les RDD, DataFrames et Datasets
- Utiliser Spark SQL dans des applications Java
- Développer des traitements Batch et Streaming
- Optimiser les performances d'applications Spark
- Déployer et superviser des applications Spark
- Adopter les bonnes pratiques de développement Big Data.

CERTIFICATION PREPAREE

Aucune

METHODES PEDAGOGIQUES

- Apports théoriques illustrés par des démonstrations
- Développement progressif d'applications Spark Java
- Travaux pratiques à chaque module
- Études de cas Big Data

- Cas fil rouge couvrant les trois journées
- Ateliers d'optimisation et de debugging

FORMATEUR

- Consultant expert Apache Spark, Java et Data Engineering, intervenant sur des projets Big Data distribués et des architectures analytiques à grande échelle.

METHODE D'EVALUATION DES ACQUIS

- Quiz de validation des connaissances
- Exercices de développement
- Travaux pratiques individuels
- Études de cas
- Projet fil rouge
- Évaluation continue par le formateur.

CONTENU DU COURS

Module 1 – Découvrir Apache Spark et les architectures Big Data (2h)

Objectifs

- Comprendre les limites des applications Java traditionnelles face aux volumes massifs
- Découvrir les concepts fondamentaux de Spark

Contenu

- Introduction au Big Data
- Hadoop et MapReduce
- Pourquoi Apache Spark ?
- Architecture Spark
- Driver et Executors
- Cluster Manager :
  - Standalone
  - YARN
  - Kubernetes
- Les modules Spark

Mise en pratique

- Analyse d'une architecture Spark
- Exécution d'une première application Spark
- Découverte du Spark UI

Module 2 – Mettre en place un environnement de développement Spark Java (2h)

Objectifs

- Configurer un environnement Spark Java complet
- Développer et exécuter ses premières applications

Contenu

- Installation de Spark
- Configuration Maven
- Gestion des dépendances
- Structure d'un projet Spark

- Création d'une SparkSession
- Exécution locale et distribuée

Mise en pratique

- Création d'un projet Spark Java
- Développement du premier programme Spark
- Tests d'exécution locale

Module 3 – Développer avec les RDD (3h)

Objectifs

- Comprendre les fondements de Spark Core
- Manipuler les collections distribuées

Contenu

- RDD (Resilient Distributed Dataset)
- Transformations
- Actions
- Évaluation paresseuse (Lazy Evaluation)
- Partitions
- Cache et persistance
- Gestion des erreurs

Mise en pratique

- Développement de traitements distribués
- Analyse du comportement des RDD
- Optimisation simple de traitements

Module 4 – Exploiter les DataFrames et Datasets en Java (4h)

Objectifs

- Utiliser les API modernes de Spark
- Manipuler efficacement les données structurées

Contenu

- DataFrames
- Datasets
- Encoders Java
- Lecture de données :
  - CSV
  - JSON
  - Parquet
- Sélection et filtrage
- Jointures
- Agrégations
- Gestion des schémas

Mise en pratique

- Construction de pipelines de traitement
- Manipulation de grands volumes de données
- Développement d'applications analytiques

Module 5 – Utiliser Spark SQL dans les applications Java (2h)

Objectifs

- Intégrer SQL dans les traitements Spark
- Réaliser des analyses avancées

Contenu

- Création de vues temporaires
- Requêtes Spark SQL
- Fonctions analytiques
- Agrégations avancées
- Fenêtres analytiques
- Intégration SQL et API Java

Mise en pratique

- Création de requêtes analytiques
- Comparaison SQL et API DataFrame
- Exploitation de données métiers

Module 6 – Développer des applications temps réel avec Structured Streaming (3h)

Objectifs

- Comprendre les traitements temps réel
- Développer des applications Streaming

Contenu

- Architecture Streaming
- Structured Streaming
- Sources Kafka
- Sources fichiers
- Gestion des événements
- Fenêtrage
- Checkpointing
- Tolérance aux pannes

Mise en pratique

- Développement d'une application Streaming Java
- Traitement d'événements Kafka
- Analyse de flux temps réel

Module 7 – Optimiser les performances des applications Spark (2h)

Objectifs

- Comprendre les mécanismes d'optimisation
- Améliorer les performances des traitements

Contenu

- Optimiseur Catalyst
- Tungsten
- Adaptive Query Execution
- Broadcast Join
- Partitionnement
- Shuffle
- Cache
- Gestion mémoire

Mise en pratique

- Analyse des plans d'exécution
- Optimisation d'une application Spark
- Mesure des gains de performance

Module 8 – Déployer et superviser des applications Spark (1h)

Objectifs

- Industrialiser les développements Spark

- Superviser les traitements

Contenu

- Packaging Maven
- spark-submit
- Déploiement sur cluster
- Logs et monitoring
- Spark History Server
- Gestion des incidents

Mise en pratique

- Déploiement d'une application
- Analyse des logs
- Diagnostic d'un incident d'exécution

Module 9 – Atelier fil rouge : développer une application Spark Java complète (2h)

Objectifs

- Mettre en œuvre l'ensemble des compétences acquises
- Réaliser une application Big Data de bout en bout

Contenu

- Ingestion de données
- Transformation distribuée
- Analyse SQL
- Optimisation
- Déploiement
- Monitoring

Mise en pratique

- Développement complet d'une application Spark Java
- Traitement d'un volume significatif de données
- Présentation des résultats
- Revue de code collective

Notre référent handicap se tient à votre disposition au 01.71.19.70.30 ou par mail à mailto:referent.handicap@edugroupe.com pour recueillir vos éventuels besoins d'aménagements, afin de vous offrir la meilleure expérience possible.