EduGroupe
Accompagner pour réussir

SPARK JAVA – DÉVELOPPER DES APPLICATIONS POUR LE BIG DATA

Durée : 3 jours (21 heures)

CONNAISSANCES PREALABLES

- Bonne maîtrise du langage Java
- Connaissances des concepts de programmation orientée objet
- Notions de SQL
- Connaissances générales des bases de données
- Une première expérience des systèmes distribués ou du Big Data constitue un avantage.

PROFIL DES STAGIAIRES

- Développeurs Java
- Data Engineers
- Développeurs Big Data
- Architectes techniques
- Consultants Data
- Ingénieurs logiciels
- Développeurs Hadoop souhaitant migrer vers Spark
- Toute personne amenée à développer des applications de traitement de données massives avec Apache Spark.

OBJECTIFS

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

- Comprendre l'architecture et le fonctionnement d'Apache Spark
- Développer des applications Big Data en Java avec Spark
- Manipuler les RDD, DataFrames et Datasets
- Mettre en œuvre des traitements distribués performants
- Exploiter Spark SQL pour l'analyse de données
- Développer des traitements Batch et Streaming
- Optimiser les performances d'une application Spark
- Déployer et superviser des applications Spark en environnement distribué.

CERTIFICATION PREPAREE

Aucune

METHODES PEDAGOGIQUES

- Alternance d'apports théoriques et de travaux pratiques
- Développement progressif d'applications Spark
- Exercices de programmation Java

- Études de cas Big Data
- Ateliers de performance et d'optimisation
- Cas fil rouge couvrant l'ensemble de la formation.

FORMATEUR

- Consultant expert Big Data, Apache Spark et développement Java, intervenant sur des projets Data Engineering, plateformes analytiques et architectures distribuées.

METHODE D'EVALUATION DES ACQUIS

- Quiz de validation des connaissances
- Exercices pratiques de développement
- Études de cas
- Réalisation d'applications Spark
- Évaluation continue par le formateur
- Validation des acquis en fin de formation.

CONTENU DU COURS

Module 1 – Comprendre Apache Spark et son écosystème (2h)

Objectifs

- Comprendre les principes du calcul distribué
- Identifier les composants de l'écosystème Spark

Contenu

- Présentation du Big Data moderne
- Limites du modèle MapReduce
- Architecture Apache Spark
- Concepts de cluster
- Driver et Executors
- Gestionnaire de ressources :
  - YARN
  - Kubernetes
  - Standalone
- Les composants Spark :
  - Spark Core
  - Spark SQL
  - Structured Streaming
  - MLlib
  - GraphX

Mise en pratique

- Exploration d'une architecture Spark
- Découverte de l'environnement de développement
- Exécution d'une première application Spark

Module 2 – Développer avec Spark Core et Java (4h)

Objectifs

- Développer des traitements distribués avec Spark

- Maîtriser les concepts fondamentaux de Spark Core

Contenu

- Création d'un projet Spark Java
- SparkSession
- SparkContext
- RDD (Resilient Distributed Datasets)
- Transformations :
  - map
  - flatMap
  - filter
  - distinct
- Actions :
  - collect
  - count
  - reduce
  - foreach
- Persistance et cache

Mise en pratique

- Développement d'applications Spark Core
- Manipulation de RDD
- Analyse de performances de traitements distribués

Module 3 – Manipuler les données avec DataFrames et Datasets (4h)

Objectifs

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

Contenu

- Introduction aux DataFrames
- Introduction aux Datasets
- Lecture de données :
  - CSV
  - JSON
  - Parquet
  - Avro
- Sélection et transformation des données
- Agrégations
- Jointures
- Gestion des schémas
- Encoders Java

Mise en pratique

- Création de pipelines DataFrames
- Transformation de jeux de données volumineux
- Réalisation d'analyses distribuées

Module 4 – Interroger les données avec Spark SQL (2h30)

Objectifs

- Utiliser SQL dans les traitements Big Data
- Réaliser des analyses avancées

Contenu

- Spark SQL

- Création de vues temporaires
- Requêtes SQL distribuées
- Fonctions intégrées
- Agrégations avancées
- Fenêtres analytiques
- Intégration SQL et Java

Mise en pratique

- Création de requêtes analytiques
- Exploitation de données métiers
- Analyse comparative SQL / API DataFrame

Module 5 – Développer des traitements temps réel avec Structured Streaming (2h30)

Objectifs

- Comprendre les architectures Streaming
- Développer des traitements temps réel

Contenu

- Principes du Streaming Data
- Architecture Structured Streaming
- Sources de données :
  - Kafka
  - Fichiers
  - Sockets
- Traitements temps réel
- Fenêtres temporelles
- Gestion des événements

Mise en pratique

- Développement d'une application Streaming
- Consommation de flux Kafka
- Analyse de données en temps réel

Module 6 – Optimisation et tuning des applications Spark (2h)

Objectifs

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

Contenu

- Optimiseur Catalyst
- Tungsten Engine
- Gestion mémoire
- Partitionnement
- Shuffle
- Cache et persistance
- Broadcast Joins
- Adaptive Query Execution (AQE)

Mise en pratique

- Analyse des plans d'exécution
- Optimisation d'applications Spark
- Comparaison des performances avant/après tuning

Module 7 – Déploiement et exploitation des applications Spark (1h30)

Objectifs

- Déployer des applications Spark en environnement distribué
- Superviser leur exécution

Contenu

- Packaging Maven
- Soumission d'applications avec spark-submit
- Déploiement sur YARN
- Déploiement sur Kubernetes
- Spark History Server
- Monitoring et logs
- Gestion des erreurs

Mise en pratique

- Déploiement d'une application Spark
- Analyse des journaux d'exécution
- Diagnostic des incidents courants

Module 8 – Atelier fil rouge : développer une application Big Data complète avec Spark Java (2h30)

Objectifs

- Mettre en œuvre les compétences acquises
- Développer une application Spark de bout en bout

Contenu

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

Mise en pratique

- Réalisation d'un projet complet :
  - Lecture de données volumineuses
  - Traitements distribués
  - Analyses métier
  - Production des résultats
- Présentation des solutions développées
- Débriefing collectif

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.