EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Développer des applications pour Apache Spark avec Python ou Scala

Code formation : EMAPS
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Spark est né en 2009 dans le laboratoire AMPLab de l'université de Berkeley. Ce framework offre un modèle de programmation plus simple que celui du MapReduce d'Hadoop et surtout plus rapide avec des temps d'exécution jusqu'à 100 fois plus courts. Avec Spark, les développeurs peuvent écrire simplement des applications distribuées complexes qui permettent de prendre des meilleures décisions plus rapidement et des actions en temps réel, appliquées à une grande variété de cas d'utilisations, d'architecture et de secteurs d'activités.

Cette formation s'adresse aux développeurs qui souhaitent créer et déployer des applications Big Data complètes et uniques en combinant batchs, le streaming et analyses interactives sur l'ensemble des données.

| Objectifs pédagogiques

- Identifier et définir les différents composants de l'écosystème Hadoop
- Appréhender le fonctionnement de Spark
- Développer des applications avec Apache Spark
- Optimiser une application Spark
- Utiliser Spark SQL et les dataframes
- Faire de l'analyse en temps réel avec Spark streaming
- Découvrir MLLib pour du machine learning sur Spark
- Explorer, manipuler et visualiser votre donnée avec Zeppelin

| Publics

- Développeur d'applications avec des contraintes temps réel
- Ingénieur d'études
- Architecte technique
- Chef de projet technique

| Pré-requis

- Connaissances de base en programmation ou en scripting (Python/Scala)
- Expérience basique en ligne de commande
- Aucune connaissance sur Hadoop n'est requise
- Connaissances en SQL et conception d'application temps réel utiles mais non obligatoire

| Méthode pédagogique

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation. Cette formation prépare à la certification éditeur Hortonworks.

| Programme détaillé

Jour 1

Comprendre HADOOP 2.X

- L'architecture de Hadoop 2.X
- The Hortonworks Data Platform (HDP)

Introduction à hadoop, hortonworks et au big data

- Cas d'usage pour Hadoop
- Qu'est-ce que Big Data ?
- HDP, Hortonworks et l'écosystème Hadoop
- Pourquoi utiliser Hortonworks ?

Introduction à apache spark

- Qu'est-ce que Spark et d'où vient-il ?
- Pourquoi utiliser Spark ?
- Spark vs MapReduce
- L'évolution rapide de Spark et l'engagement d'Hortonworks

Programmer avec apache spark

- Les composants de Spark
- Premiers pas avec Spark
- Les RDD
- Transformations et actions
- Spark Hello World (wordcount)
- Lazy evaluation
- Mise en pratique: "Assurer ses premiers pas avec Apache Spark"

Vue d'ensemble de HDFS et YARN

- Vue d'ensemble de HDFS
- Le Namenode et le Datanode
- Vue d'ensemble de YARN
- Composants cœur de YARN
- Mise en pratique: "Utiliser les commandes HDFS"

Programmation rdd avancée

- D’autres fonctions de RDD "cœur"
- Fonctions de RDD paires
- Utiliser la documentation de Spark
- Mise en pratique : "Utiliser le stockage HDFS"

Jour 2

Programmation parallèle avec SPARK

- Partitionnement, jobs, stage et tasks
- L’UI de Spark
- Changer le niveau de parrallélisation
- Mise en pratique : Programmation parallèle sur Spark

Cacher et persister la donnée

- Cache et persistance
- Mise en pratique : "cacher et persister la donnée"
- Exemple d’application itérative : PageRank
- Checkpointing
- Mise en pratique : "Checkpointing et RDD lineage"

Créer des applications Spark

- Créer une application à soumettre au cluster
- Soumettre une application au cluster
- Yarn client vs Yarn cluster
- Points importants de configuration
- Gérer/packager les dépendances
- Mise en pratique : "Créer une application Spark standalone"

Jour 3

Fonctionnalités avancées et amélioration des performances

- Accumulateurs
- Mise en pratique : "Utiliser les accumulateurs pour vérifier la qualité des données"
- Variables « broadcast »
- Mise en pratique : "Utiliser les variables broadcast"
- Partitionnement avancé et opérations
- Point de départ pour l’optimisation

Travailler vos données avec Zeppelin

- L'exploration de données en Spark avec Zeppelin
- Visualisation de données avec Zeppelin
- Faire du reporting avec Zeppelin

SPARK SQL

- Les concepts de Spark SQL
- Créer une Dataframe
- Sauvegarder une Dataframe
- Spark SQL et UDF

- Mise en pratique : "Spark SQL avec utilisation d'UDF"
- Mise en pratique : "Spark SQL avec Hive"

Jour 4

Spark Streaming

- L'architecture de Spark Streaming
- Vue d'ensemble de Spark Streaming
- Fiabilité des récepteurs et des sources
- Transformations et opérations de sorties
- Mise en pratique : "Wordcount en Spark Streaming"
- Configurer le checkpointing

SPARK MLLIB

- Vue d'ensemble de MLLib
- Apprentissage supervisé
- Apprentissage non supervisé