DOCAPOSTE Institute

Formation Introduction à Spark Scala pour le traitement des données

Présentation

Ce programme de vise à fournir aux participants une compréhension solide des bases de Spark Scala pour le traitement de données.
Il couvre progressivement des concepts allant des RDD et DataFrames aux opérations SQL et à l'optimisation des performances, en leur permettant de devenir autonomes dans la manipulation et l'analyse de données massives avec Spark.

Durée : 21,00 heures (3 jours)
Tarif INTRA : Nous consulter

Objectifs de la formation

- Comprendre les fondamentaux de Spark et de Scala
- Maîtriser les bases du traitement de données avec Spark
- Savoir manipuler des données structurées avec Spark DataFrame
- Utiliser Spark SQL pour l'interrogation de données
- Comprendre l'optimisation et les performances dans Spark

Prérequis

Cette formation Introduction à Spark Scala pour le traitement des données est conçue pour les débutants dans le domaine de Spark Scala, certaines connaissances de base en programmation et en bases de données sont généralement recommandées pour tirer pleinement parti du contenu du cours.

Public

Développeurs, analyste de données, Ingénieurs Big Data

Programme de la formation

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France

DOCAPOSTE Institute

Jour 1 : Introduction à Apache Spark et Scala

Introduction à Apache Spark

- Présentation de Spark et son rôle dans le traitement de données à grande échelle.
- Avantages de l'utilisation de Spark pour le traitement de données par rapport à d'autres technologies.

Introduction à Scala

- Présentation du langage de programmation Scala et de ses caractéristiques.
- Les principales différences entre Scala et d'autres langages de programmation courants.

Installation et configuration de l'environnement Spark

- Installation de Spark et de Scala sur l'environnement de développement.
- Configuration des variables d'environnement et des paramètres de Spark.

Les RDD (Resilient Distributed Datasets) en Spark

- Comprendre les RDD en tant que principal concept de stockage et de traitement des données dans Spark.
- Création et manipulation de RDD à partir de diverses sources de données.

Les transformations et les actions de base sur les RDD

- Comprendre les transformations et les actions et leur rôle dans les opérations de traitement de données.
- Utiliser les transformations (map, filter, reduceByKey, etc.) et les actions (count, collect, saveAsTextFile, etc.) pour analyser les données.

Manipulation de données avec Spark RDD

- Appliquer des opérations de traitement de données sur des RDD pour résoudre des problèmes pratiques.
- Exemples pratiques d'utilisation de Spark RDD pour effectuer des tâches de traitement de données.

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France

DOCAPOSTE Institute

Jour 2 : Manipulation de données avec DataFrames en Spark

Introduction aux DataFrames en Spark

- Comprendre les DataFrames en tant que structure de données tabulaire organisée et distribuée dans Spark.
- Avantages d'utilisation des DataFrames par rapport aux RDD.

Création et manipulation de DataFrames

- Charger des données dans un DataFrame à partir de différentes sources (fichiers CSV, JSON, bases de données, etc.).
- Appliquer des transformations sur les DataFrames pour filtrer, sélectionner et grouper les données.

Opérations de base sur les DataFrames

- Utiliser des fonctions d'agrégation telles que sum, avg, max, min, etc.
- Appliquer des opérations de jointure pour combiner des DataFrames.

Introduction aux Datasets en Spark

- Comprendre les Datasets en tant que nouvelle API de haut niveau pour manipuler des données structurées.
- Comparaison des Datasets avec les RDD et les DataFrames.

Traitement de données structurées avec les Datasets

- Utiliser des Datasets pour manipuler des données structurées et bénéficier de la typage statique.
- Appliquer des opérations de filtrage, de tri et de groupage sur les Datasets.

Les fonctions d'agrégation et les opérations avancées sur les DataFrames

- Utiliser des fonctions d'agrégation avancées pour effectuer des calculs complexes sur les données.
- Utiliser des fonctions de fenêtrage pour effectuer des opérations sur des fenêtres de données.

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France

Jour 3 : Utilisation de SQL et optimisation dans Spark

Introduction à Spark SQL

- Comprendre le rôle de Spark SQL dans le traitement de données.
- Utiliser SQL pour interroger des données dans Spark et tirer parti des capacités de Spark SQL.

Utilisation de SQL pour interroger des données dans Spark

- Écrire des requêtes SQL pour effectuer des analyses et des transformations de données.
- Comparaison des requêtes SQL avec les opérations DataFrame et RDD.

Introduction aux performances et à l'optimisation dans Spark

- Comprendre l'importance des performances dans le traitement de données à grande échelle.
- Identifier les goulots d'étranglement et les problèmes de performances courants.

Gestion des partitions pour améliorer les performances

- Comprendre le concept de partitionnement des données dans Spark.
- Optimiser les partitions pour améliorer l'efficacité des opérations de traitement de données.

Résolution de problèmes de traitement de données courants avec Spark Scala

- Exemples pratiques de résolution de problèmes courants liés au traitement de données avec Spark et Scala.
- Réalisation de projets pratiques pour appliquer les connaissances acquises au cours de la formation.

Organisation

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France

Formateur

Les formateurs de Docaposte Institute sont des experts de leur domaine, disposant d'une expérience terrain qu'ils enrichissent continuellement. Leurs connaissances techniques et pédagogiques sont rigoureusement validées en amont par nos référents internes.

Riches de leur expérience sur le sujet, ils sauront accompagner vos collaborateurs dans leur montée en compétence.

Moyens pédagogiques et techniques

- Apports des connaissances communes.
- Mises en situation sur le thème de la formation et des cas concrets.
- Méthodologie d'apprentissage attractive, interactive et participative.
- Equilibre théorie / pratique : 60 % / 40 %.
- Supports de cours fournis au format papier et/ou numérique.
- Ressources documentaires en ligne et références mises à disposition par le formateur.
- Pour les formations en présentiel dans les locaux mis à disposition, les apprenants sont accueillis dans une salle de cours équipée d'un réseau Wi-Fi, d'un tableau blanc ou paperboard. Un ordinateur avec les logiciels appropriés est mis à disposition (le cas échéant).

Dispositif de suivi de l'exécution et de l'évaluation des résultats de la formation

En amont de la formation

- Recueil des besoins des apprenants afin de disposer des informations essentielles au bon déroulé de la formation (profil, niveau, attentes particulières...).
- Auto-positionnement des apprenants afin de mesurer le niveau de départ.

Tout au long de la formation

- Évaluation continue des acquis avec des questions orales, des exercices, des QCM, des cas pratiques ou mises en situation...

A la fin de la formation

- Auto-positionnement des apprenants afin de mesurer l'acquisition des compétences.
- Evaluation par le formateur des compétences acquises par les apprenants.

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France

- Questionnaire de satisfaction à chaud afin de recueillir la satisfaction des apprenants à l'issue de la formation.
- Questionnaire de satisfaction à froid afin d'évaluer les apports ancrés de la formation et leurs mises en application au quotidien.

Accessibilité

Nos formations peuvent être adaptées à certaines conditions de handicap. Nous contacter pour toute information et demande spécifique.

Docaposte Institute
ACI 3A3-303 - 45-47 Boulevard Paul Vaillant Couturier, 94200 IVRY SUR SEINE
Tél : Variable non renseignée - Email : hello.institute@docaposte.fr
SIRET : 813 058 542 00032 - Code APE : 8559A - N TVA Intra : FR 58 813 058 542
Déclaration d'activité enregistrée sous le numéro 11941117994 auprès du préfet de région d'Ile-de-France