DOCAPOSTE Institute

Formation Spark, Développer des applications pour le Big Data

Présentation

Cette formation vous fournira une solide introduction technique à l'architecture Spark et au fonctionnement de Spark. Vous apprendrez les éléments de base de Spark, notamment les RDD et le moteur de calcul distribué, ainsi que les constructions de niveau supérieur, qui fournissent une interface plus simple et plus performante, notamment Spark SQL et DataFrames.

Vous verrez également des capacités plus avancées telles que l'utilisation de Spark Streaming pour traiter les données en continu, et aurez un aperçu du traitement graphique Spark (GraphX et GraphFrames) et du Machine Learning Spark (SparkML Pipelines).

Enfin, vous explorerez les éventuels problèmes de performance, le dépannage, les techniques de déploiement de grappes et les stratégies d'optimisation.

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

Objectifs de la formation

- Comprendre la nécessité de Spark dans le traitement des données
- Comprendre l'architecture Spark et la distribution des calculs aux nœuds de cluster.
- Se familiariser avec l'installation de base / la configuration / l'agencement de Spark
- Utiliser Spark pour des opérations interactives et ad hoc
- Utiliser Dataset/DataFrame/Spark SQL pour traiter efficacement les données structurées
- Comprendre les bases des RDD (Resilient Distributed Datasets), ainsi que le partitionnement, la circulation dans les pipelines et le calcul des données
- Comprendre la mise en cache des données de Spark et son utilisation
- Comprendre les implications et les optimisations des performances lors de l'utilisation de Spark

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

- Se familiariser avec le traitement graphique et l'apprentissage machine SparkML

Prérequis

- Connaissance de la programmation fonctionnelle avec les langages Java ou Python,
- Connaissances en gestion des bases de données,
- Notions de calculs statistiques.

Public

- Chefs de projet,
- Data Scientist,
- Développeurs,
- Architectes...

Programme de la formation

La montée en puissance de Scala

- Introduction à Scala, variables, types de données, flux de contrôle
- L'interpréteur Scala
- Collections et méthodes standard (par exemple map())
- Fonctions, méthodes, fonctions littérales
- Classe, objet, trait

Introduction à Spark

- Vue d'ensemble, motivations, systèmes Spark
- Ecosystème de Spark
- Spark vs. Hadoop
- Environnements typiques de déploiement et d'utilisation de Spark

Les RDD et l'architecture Spark

- Concepts de RDD, partitions, cycle de vie, évaluation paresseuse
- Travailler avec les RDD - Créer et transformer (carte, filtre, etc.)
- Mise en cache – Concepts, type de stockage, directives

DataSets/DataFrames et Spark SQL

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

- Introduction et utilisation
- Création et utilisation d'un ensemble de données
- Travailler avec JSON
- Utilisation du DataSet DSL
- Utiliser SQL avec Spark
- Formats de données
- Optimisations : Catalyst et Tungsten
- DataSets vs. DataFrames vs. RDD

Créer des applications Spark

- Aperçu, code de pilote simple, SparkConf
- Création et utilisation d'un contexte SparkContext/SparkSession
- Création et fonctionnement des applications
- Cycle de vie des applications
- Gestionnaires de clusters
- Logging et débogage

Spark Streaming

- Vue d'ensemble et principes de base de la diffusion en continu
- Streaming structuré
- DStreams (Discretized Steams),
- Architecture, Stateless, Stateful, et Windowed Transformations
- API de diffusion en continu (Spark Streaming)
- Programmation et transformations

Caractéristiques et optimisation des performances

- UI Spark
- Dépendances étroites vs. larges
- Réduire au minimum le traitement et le brassage des données
- Mise en cache - Concepts, type de stockage, lignes directrices
- Utilisation de la mise en cache
- Utilisation des variables de diffusion et des accumulateurs

Aperçu de Spark GraphX

- Introduction
- Construire des graphiques simples
- API GraphX
- Exemple de chemin le plus court

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

Aperçu de MLLib

- Introduction
- Vecteurs caractéristiques
- Regroupement / Groupement, K-Means
- Recommandations
- Classifications

Conclusion

Organisation

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.

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

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

- 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.
- 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.

NB : dans le cadre d'une Action collective, chaque stagiaire bénéficiaire sera contacté par un prestataire choisi par l'Opco Atlas afin d'évaluer « à chaud » la qualité de la formation suivie.

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