Big Data - Développement

Spark Java - Traitement des données

3 jours (21h00) | 4,6/5 | BIG-SPKJAV | Évaluation qualitative de fin de stage | Formation délivrée en présentiel ou distanciel

Contenu mis à jour le 13/10/2023. Document téléchargé le 29/05/2024.

Objectifs de formation

A l'issue de cette formation, vous serez capable de :

- Utiliser le style fonctionnel Java pour définir des tâches complexes de traitement des données
- Différencier les API RDD (Resilient Distributed Dataset) et DataFrame
- Utiliser une syntaxe de style SQL pour produire des rapports sur des ensembles de Big Data
- Mettre en place des algorithmes d'apprentissage automatique avec le Big Data et Spark ML (Machine Learning)
- Connecter Spark à Apache Kafka pour traiter des flux de Big Data
- Expliquer comment le streaming structuré peut être utilisé pour construire des pipelines avec Kafka.

Modalités, méthodes et moyens pédagogiques

Formation délivrée en présentiel ou distanciel* (blended-learning, e-learning, classe virtuelle, présentiel à distance).

Le formateur alterne entre méthode** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation).

Variables suivant les formations, les moyens pédagogiques mis en oeuvre sont :

- Ordinateurs Mac ou PC (sauf pour certains cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel)
- Environnements de formation installés sur les postes de travail ou en ligne
- Supports de cours et exercices

En cas de formation intra sur site externe à M2i, le client s'assure et s'engage également à avoir toutes les ressources matérielles pédagogiques nécessaires (équipements informatiques...) au bon déroulement de l'action de formation visée conformément aux prérequis indiqués dans le programme de formation communiqué.

** nous consulter pour la faisabilité en distanciel*

*** ratio variable selon le cours suivi*

Prérequis

Avoir une connaissance préalable de Java est supposée, mais tout ce qui dépasse les bases est expliqué. Une connaissance préalable de SQL sera utile pour une partie du cours, mais si vous ne l'avez jamais utilisé auparavant, ce sera une bonne première expérience.
Java 8 est requis pour le cours (Spark ne supporte pas actuellement Java 9+, cependant Java 8 est nécessaire pour la syntaxe lambda fonctionnelle).

Public concerné

Développeurs, statisticiens, consultants Big Data, Data Analysts, Data Scientists, architectes.

Cette formation :

- Est animée par un consultant-formateur dont les compétences techniques, professionnelles et pédagogiques ont été validées par des diplômes et/ou testées et approuvées par l'éditeur et/ou par M2i Formation
- Bénéficie d'un suivi de son exécution par une feuille de présence émargée par demi-journée par les stagiaires et le formateur.

Programme

Jour 1

Introduction

- Architecture de Spark et les RDD

Installation du Spark

Reduce sur les RDD

Mapping et outputting

- Les opérations de mapping
- Outputting des résultats sur la console
- Compter les éléments du Big Data
- "NotSerializableException" avec Spark

Tuples

- RDD des objets
- Tuples et RDD

Pair RDD

- Vue d'ensemble des Pair RDD
- Construire Pair RDD
- Coder le "reduceByKey"
- Utilisation de l'API Fluent
- Groupement par clé (BY KEY)

FlatMaps et filters

Lecture du disque

Classement des mots-clés en pratique

- Exigences pratiques
- Solution pratique (avec tri)

Tri et coalescence

- Coalesce dans Spark ?

Déploiement vers AWS EMR (Amazon Elastic MapReduce)

- Comment démarrer un cluster Spark pour EMR
- Emballage d'un Spark JAR pour EMR
- Exécuter un travail Spark sur EMR
- Comprendre la sortie de la progression du travail
- Calculer les coûts d'EMR et terminer le cluster

Jointures

- Internes
- Externes de gauche et optionnelles
- Externes à droite
- Complètes et cartésiennes

Exemples de travaux pratiques (à titre indicatif)

- *Big Data (grand exercice)*
  - *Présentation des exigences*
  - *Echauffement*
  - *Exigences de l'exercice principal*
  - *Marche à suivre*

La performance des RDD

- Transformations et actions
- Le DAG (Directed Acyclic Graph) et Spark UI
- Transformations étroites et larges
- Shuffles
- Gérer les BY KEY
- "map-side-reduces"
- Mise en cache et persistance

Jour 2

Spark SQL : introduction

- Utilisation pratique de Spark SQL

Datasets

- Les bases du Dataset
- Filtrage en utilisant les expressions, lambda et colonnes

SQL : syntaxe

- Utilisation d'une vue temporaire Spark pour SQL

Données en mémoire

Groupements et agrégations

Date Formatting

Multiple Groupings

Ordering

DataFrame API

- SQL vs DataFrame
- Groupement DataFrame

Pivot tables

- Coder Pivot table en Spark

Plus d'agrégations

- Comment utiliser la méthode "agg" en Spark

Exemples de travaux pratiques (à titre indicatif)

- *Comment utiliser lambda pour écrire un UDF (User Defined Functions) en Spark*
- *Utilisation de multiples paramètres d'entrées en Spark UDF*
- *Utilisation des UDF en Spark SQL*

Performance de Spark SQL

- Comprendre le Spark UI pour Spark SQL
- Performances de SQL et de DataFrame ?
- Mise à jour et réglage "spark.sql.shuffle.partitions"

HashAggregation

- Explication des plans d'exécution
- HashAggregation

Performance Spark SQL vs RDD

- Introduction de ML
- Apprentissage supervisé et non-supervisé
- Processus de construction d'un modèle

Régression linéaire

- Introduction
- Programmation des modèles de régression linéaire
- Assemblage des vecteurs des paramètres
- Fitting des modèles

Données d'apprentissage

- Training vs test et holdout Data
- Guide pratique
- Evaluation de la précision des modèles avec R2 et RMSE (Root Mean Square Error)

Paramètres d'ajustement des modèles

- Ajustement des paramètres des modèles de régression linéaire
- Training, test et holdout Data

Sélection des caractéristiques (features)

- Description des caractéristiques
- Corrélation des caractéristiques
- Identification et élimination des caractéristiques dupliquées
- Préparation des données

Données non numériques

- Utilisation "OneHotEncoding"
- Comprendre les Vectors

Pipelines

Cas d'étude

Régression logistique

- True vs false / negatives vs postives
- Implémentation de la régression logistique

Les arbres de décision

- Aperçu des arbres de décision
- Construction du modèle
- Interprétation d'un arbre de décision
- Random Forest

K-means clustering

Jour 3

Spark Streaming et streaming structuré avec Kafka

Introduction au streaming

- DStreams
- Commencer Streaming Job
- Transformations et agrégations streaming
- Spark UI pour les Streaming Jobs
- Traitement des lots

Streaming avec Apache Kafka

- Introduction et installation
- Utilisation du Kafka Event Simulator
- Intégration de Kafka avec Spark
- Utilisation de KafkaUtils pour accéder au DStream
- Ecrire une agrégation Kafka
- Ajouter une fenêtre et "slide interval"

Streaming structuré

- Aperçu du streaming structuré
- Les puits de données
- Les modes de sortie du streaming structuré
- Fenêtres et filigranes
- Batch pour le streaming structuré ?
- Kafka Structured Streaming Pipelines

Le contenu de ce programme peut faire l'objet d'adaptation selon les niveaux, prérequis et besoins des apprenants.

Modalités d'évaluation des acquis

- En cours de formation, par des études de cas ou des travaux pratiques
- Et, en fin de formation, par un questionnaire d'auto-évaluation

Les + de la formation

Le cours comprend :

- un module couvrant Spark ML, un ajout passionnant à Spark qui vous permet d'appliquer des modèles d'apprentissage automatique à vos Big Data ! Aucune expérience mathématiques n'est nécessaire !

- un module complet de 3 heures couvrant Spark Streaming, où vous aurez une expérience pratique de l'intégration de Spark avec Apache Kafka pour gérer les flux de données en temps réel. Nous utilisons à la fois les API DStream et streaming structuré.

Accessibilité de la formation

Le groupe M2i s'engage pour faciliter l'accessibilité de ses formations. Les détails de l'accueil des personnes en situation de handicap sont consultables sur la page Accueil et Handicap.

Modalités et délais d’accès à la formation

Les formations M2i sont disponibles selon les modalités proposées sur la page programme. Les inscriptions sont possibles jusqu'à 48 heures ouvrées avant le début de la formation. Dans le cas d'une formation financée par le CPF, ce délai est porté à 11 jours ouvrés.