Spark Java, développer des applications pour le Big Data

Cours Pratique de 3 jours - 21h
Réf : SPK - Prix 2024 : 2 280€ HT

Souvent présenté comme le successeur de Hadoop, SPARK simplifie la programmation des traitements BigData permettant l'utilisation de scala, Python ou Java . Cette formation apprendra aux programmeurs à traiter un flux de données en temps réel et à effectuer des traitements batch (du SQL jusqu'au Machine Learning).

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Maîtriser les concepts fondamentaux de Spark

Développer des applications avec Spark Streaming

Mettre en œuvre un cluster Spark

Exploiter des données avec Spark SQL

Avoir une première approche du Machine Learning

TRAVAUX PRATIQUES

Mise en pratique des notions vues en cours à l'aide du langage Java.

LE PROGRAMME

dernière mise à jour : 11/2022

1) Présentation d'Apache Spark

- Historique du Framework.
- Les différentes versions de Spark (Scala, Python et Java).
- Comparaison avec l'environnement Apache Hadoop.
- Les différents modules de Spark.

Travaux pratiques : Installation et configuration de Spark. Exécution d'un premier exemple avec le comptage de mots.

2) Programmer avec les Resilient Distributed Dataset (RDD)

- Présentation des RDD.
- Créer, manipuler et réutiliser des RDD.
- Accumulateurs et variables broadcastées.
- Utiliser des partitions.

Travaux pratiques : Manipulation de différents Datasets à l'aide de RDD et utilisation de l'API fournie par Spark.

3) Manipuler des données structurées avec Spark SQL

- SQL, DataFrames et Datasets.
- Les différents types de sources de données.
- Interopérabilité avec les RDD.
- Performance de Spark SQL.
- JDBC/ODBC server et Spark SQL CLI.

Travaux pratiques : Manipulation de Datasets via des requêtes SQL. Connexion avec une base externe via JDBC.

PARTICIPANTS

Chefs de projet, data scientists, développeurs, architectes.

PRÉREQUIS

Bonnes connaissances du langage Java. Connaissances en Big Data.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

4) Spark sur un cluster

- Les différents types d'architecture : Standalone, Apache Mesos ou Hadoop YARN.
- Configurer un cluster en mode Standalone.
- Packager une application avec ses dépendances.
- Déployer des applications avec Spark-submit.
- Dimensionner un cluster .

Travaux pratiques : Mise en place d'un cluster Spark.

5) Analyser en temps réel avec Spark Streaming

- Principe de fonctionnement.
- Présentation des Discretized Streams (DStreams).
- Les différents types de sources.
- Manipulation de l'API.
- Comparaison avec Apache Storm.

Travaux pratiques : Consommation de logs avec Spark Streaming.

6) Manipuler des graphes avec GraphX

- Présentation de GraphX.
- Les différentes opérations.
- Créer des graphes.
- Vertex and Edge RDD.
- Présentation de différents algorithmes.

Travaux pratiques : Manipulation de l'API GraphX à travers différents exemples.

7) Machine Learning avec Spark

- Introduction au Machine Learning.
- Les différentes classes d'algorithmes.
- Présentation de SparkML et MLlib.
- Implémentations des différents algorithmes dans MLlib.

Travaux pratiques : Utilisation de SparkML et MLlib.

LES DATES

CLASSE À DISTANCE	LILLE	PARIS
2024 : 12 juin, 18 sept., 18 déc.	2024 : 12 juin, 18 sept., 18 déc.	2024 : 05 juin, 11 sept., 18 déc.