Programme de formation

Spark - Traitement de données

3 jours - 21 heures	Code formation : ADHDEV0645

adhara France

adhara.fr

Objectifs

Comprendre les fondamentaux du développement d'applications Big Data en temps réel. Appliquer les systèmes de calculs distribués en temps réel. Traiter des grosses quantités de données en temps réel.

Participants

Développeurs informatiques, Chefs de projet, Data Scientists, Consultants en business intelligence, Responsables système d'informations.

Prérequis

Avoir connaissance langages orientés objet (Java, Python…).

Pédagogie

La pédagogie est basée sur le principe de la dynamique de groupe avec alternance d'apports théoriques, de phases de réflexion collectives et individuelles, d'exercices, d'études de cas et de mises en situations observées. Formation / Action participative et interactive : les participants sont acteurs de leur formation notamment lors des mises en situation car ils s'appuient sur leurs connaissances, les expériences et mettront en œuvre les nouveaux outils présentés au cours de la session.

Remarques

Certification

Profil de l'intervenant

Consultant-formateur expert. Suivi des compétences techniques et pédagogiques assuré par nos services.

Moyens techniques

Encadrement complet des stagiaires durant la formation. Espace d'accueil, configuration technique des salles et matériel pédagogique dédié pour les formations en centre. Remise d'une documentation pédagogique papier ou numérique à échéance de la formation.

Méthodes d'évaluation des acquis

Un contact téléphonique est systématiquement établi avec le stagiaire ou la personne chargée de son inscription afin de définir le positionnement. Si besoin, un questionnaire est adressé pour valider les prérequis en correspondance et obtenir toute précision nécessaire permettant l'adaptation de l'action. Durant la formation, des exercices individuels et collectifs sont proposés pour évaluer et valider les acquis du stagiaire. La feuille d'émargement signée par demi-journée ainsi que l'évaluation des acquis sont adressées avec la facture.

Programme

Introduction

Présentation Spark, origine du projet, apports, principe de fonctionnement
Langages supportés

Programme de formation

Spark - Traitement de données

3 jours - 21 heures	Code formation : ADHDEV0645

adhara France

adhara.fr

Premiers pas

Utilisation du shell Spark avec Scala ou Python
Gestion du cache

Règles de développement

Mise en pratique en Java et Python
Notion de contexte Spark
Différentes méthodes de création des RDD : depuis un fichier texte, un stockage externe
Manipulations sur les RDD (Resilient Distributed Dataset)
Fonctions, gestion de la persistence

Cluster

Différents cluster managers : Spark en autonome, avec Mesos, avec Yarn, avec Amazon EC2
Architecture : SparkContext,Cluster Manager, Executor sur chaque nœud
Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job
Mise en oeuvre avec Spark et Amazon EC2
Soumission de jobs, supervision depuis l'interface web

Intégration hadoop

Travaux pratiques avec YARN
Création et exploitation d'un cluster Spark/YARN

Support Cassandra

Description rapide de l'architecture Cassandra. Mise en oeuvre depuis Spark
Exécution de travaux Spark s'appuyant sur une grappe Cassandra

Spark SQL

Objectifs : traitement de données structurées
Optimisation des requêtes
Mise en oeuvre de Spark SQL
Comptabilité Hive

Streaming

Objectifs, principe de fonctionnement : stream processing
Source de données : HDFS, Flume, Kafka, ...
Notion de Streaming : Contexte, DStreams, démonstrations

MLib

Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistence, statistiques
Support de RDD
Mise en œuvre avec les DataFrames

GraphX

Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes