Spark : traitement de données

VOTRE FORMATION

DURÉE : 3 JOURS

21 heures

PROCHAINE SESSION :

Du 15 au 17 juin 2020

LIEU : En distanciel

PRIX : 2 220 €

net de taxes

LE PUBLIC

Chefs de projet, data scientists, développeurs.

LES OBJECTIFS

Comprendre le fonctionnement de Spark et son utilisation dans un environnement Hadoop.

Savoir intégrer Spark dans un environnement Hadoop, traiter des données Cassandra, HBase, Kafka, Flume, Sqoop, S3.

Cette formation permet de se présenter à l'examen "Certification Hadoop avec Spark pour développeur de Cloudera"

PRÉ-REQUIS

Connaissances des principes du BigData, et des architectures techniques mises en oeuvre

MODALITÉS

La formation est accessible à distance en classe virtuelle : accès à l'infrastructure de travaux pratiques, machines physiques, outils pédagogiques, échanges avec le formateur

De 4 à 12 participants

Financement éligible au FNE Formation pour tout salarié d'entreprise en activité partielle

VOTRE CONTACT :

Andrea FALLOURD
Conseillère en formation
06 74 51 44 97
afallourd@itescia.fr

ITESCIA - Campus de Pontoise
8 rue Pierre de Coubertin
95300 PONTOISE

www.itescia.fr

VOTRE PROGRAMME

Introduction

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

Premiers pas

Utilisation du shell Spark avec Scala ou Python. Modes de fonctionnement. Interprété, compilé.

Utilisation des outils de construction. Gestion des versions de bibliothèques.

Règles de développement

Mise en pratique en Java, Scala 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 persistance.

Cluster

Différents cluster managers : Spark en autonome, avec Mesos, avec Yarn, avec Amazon EC2

Architecture : SparkContext, Cluster Manager, Executor sur chaque noeud.

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

Traitements

Lecture/écriture de données : Texte, JSon, Parquet, HDFS, fichiers séquentiels.

Jointures. Filtrage de données, enrichissement. Calculs distribués de base. Introduction aux traitements de données avec map/reduce.

Travail sur les RDDs. Transformations et actions. Lazy execution. Impact du shuffle sur les performances.

RDD de base, key-pair RDDs.Variables partagées : accumulateurs et variables broadcast.

Intégration hadoop

Présentation de l'écosystème Hadoop de base : HDFS/Yarn.Travaux pratiques avec YARN
Création et exploitation d'un cluster Spark/YARN. Intégration de données sqoop, kafka, flume vers une architecture Hadoop.
Intégration de données AWS S3.

Support Cassandra

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

DataFrames

Spark et SQL

Objectifs : traitement de données structurées. L'API Dataset et DataFrames

Optimisation des requêtes.

Mise en oeuvre des Dataframes et DataSet.Comptabilité Hive

Travaux pratiques: extraction, modification de données dans une base distribuée.

Collections de données distribuées.

Exemples.

Streaming

Objectifs , principe de fonctionnement : stream processing.

Source de données : HDFS, Flume, Kafka, ...
Notion de Streaming. Contexte, DStreams, démonstrations.

Travaux pratiques : traitement de flux

DStreams en Scala.

Machine Learning

Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistance, statistiques.

Support de RDD. Mise en œuvre avec les DataFrames.

Spark GraphX

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

Travaux pratiques :exemples d'opérations sur les graphes.

Rejoignez nos réseaux sociaux