Pythagore F.D.

Apprendre à apprendre

CB037

Spark : traitement de données

Durée: 3 jours

1750 €	26 au 28 mars	16 au 18 octobre
	27 au 30 juin	17 au 19 décembre

Public:

Chefs de projet, data scientists, développeurs.

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. Ce stage permet de se présenter à l'examen "Certification Hadoop avec Spark pour développeur de Cloudera"

Connaissances préalables nécessaires:

Connaissance de Java ou Python, notions de calculs statistiques et des bases Hadoop ou avoir suivi le stage "Hadoop, l'écosystème".

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 persistence.

Pythagore F.D.

Apprendre à apprendre

CB037

Spark : traitement de données

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 oeuvre 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.

Pythagore F.D.

Apprendre à apprendre

CB037

Spark : traitement de données

Streaming : Objectifs , principe de fonctionnement : stream processing.
Source de données : HDFS, Flume, Kafka, ...
Notion de StreamingContexte, DStreams, démonstrations
Travaux pratiques : traitement de flux DStreams en Scala.

Machine Learning: Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistence, statistiques.
Support de RDD.
Mise en oeuvre 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.