Mise en œuvre des traitements Big Data avec Spark

Programme

Introduction

- Présentation de Spark, origine du projet
- Apports et principes 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

Traitement

- 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

Référence

THBI1211

Durée

3 jours / 21 heures

Prix HT / stagiaire

1500€

Objectifs pédagogiques

- Concevoir le fonctionnement de Spark et son utilisation dans un environnement Hadoop
- Intégrer Spark dans un environnement Hadoop
- Traiter des données Cassandra, HBase, Kafka, Flume, Sqoop et S3

Niveau requis

- Garantir avoir des connaissances sur Java ou Python
- Garantir avoir les bases Hadoop Notions de calculs statistiques

Public concerné

- Chefs de projet Data Scientists Développeurs

Formateur

Les formateurs intervenants pour Themanis sont qualifiés par notre Responsable Technique Olivier Astre pour les formations informatiques et bureautiques et par Didier Payen pour les formations management.

Moyens pédagogiques et techniques

Salles de formation (accessibles et adaptables aux besoins des personnes en situation de handicap) équipée d'un ordinateur de dernière génération par stagiaire, réseau haut débit et vidéo-projection UHD
Documents supports de formation projetés
Apports théoriques, étude de cas concrets et exercices
Mise à disposition en ligne de documents supports à la suite de la formation

Dispositif de suivi de l'éxécution de l'évaluation des résultats de la formation

Feuilles d'émargement (signature électronique privilégiée)
Evaluations formatives et des acquis sous forme de questions orales et/ou écrites (QCM) et/ou mises en situation
Questionnaires de satisfaction (enquête électronique privilégiée)

- 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

Streaming

- Objectifs , principe de fonctionnement : stream processing
- Source de données : HDFS, Flume, Kafka, …
- Notion de Streaming
- Contexte, DStreams, démonstrations
- 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 oeuvre avec les DataFrames

Spark Graphx

- Fourniture d’algorithmes, d’opérateurs simples pour des calculs statistiques sur les graphes
- Exemples d’opérations sur les graphes