Tél : +352 27 39 351
Fax : +352 27 39 52 68
Email : formation@oxiane.lu

Apache Spark

Spark : traitement de données

Apache Spark est un moteur de traitements distribués sur des gros volumes de données.

Souvent mis en opposition au modèle mapreduce implémenté dans Hadoop, il en est en fait une extension qui peut en diviser les temps d'exécution jusqu'à un facteur de 100 en maximisant le travail « in-memory».

Spark exploite les principes de programmation fonctionnelle afin d'optimiser l'empreinte mémoire nécessaire à son exécution. Conçu pour mettre en œuvre des traitements distribués, Spark peut s'appuyer sur plusieurs types de clusters, dont YARN le négociateur de ressources intégré à Hadoop.

Détails

- Code : DB-SPK
- Durée : 3 jours (21 heures)

Public
- Chefs de projets
- Data miner
- Data Scientist
- Developpeurs

Pré-requis
- Connaissance d'un langage de programmation

Objectifs
- Concevoir une application avec Spark
- Comprendre le principe de distribution des traitements
- Maîtriser les concepts fondamentaux des et des Resilient Distributed Dataset
- Utiliser les dataframes via Spark SQL
- Utiliser SparkUI afin d'analyser les jobs et tâches de Spark

Programme

Présentation de Spark
- Spark : un besoin de distribuer vos traitements
- Architecture de Spark runtime : driver, executor, master
- Positionner Spark vs Hadoop
- Les langages du framework : Java | Scala | Python | R

RDD : Resilient Distributed Dataset
- RDD : Le composant fondateur du fonctionnement de Spark
- Les partitions : la base de la distribution
- Transformations, actions et directed acyclic Graph
- Manipuler un RDD : Une API riche
- Le cas particulier des Pairs RDD

SparkSQL, Dataframes et Datasets
- Un modèle de programmation haut niveau
Initialisation d'un dataframe
- Manipulation : sélection, tri et fonctions d'agrégation.
- Dataset : une surcouche typée des dataframes
- Comprendre le plan d'exécution d'une requête
- Bonnes et mauvaises pratiques avec SparkSQL

Mise en cluster : Les infrastructures de déploiement
- Les composants d'une exécution Spark : Jobs, stages et tasks
- Un principe important : Data locality
- Distribution des données dans le cadre d'un cluster : les partitions
- Redistribution des données : le shuffle
- Bonnes pratiques et performance

Modalités
- Type d'action :Acquisition des connaissances
- Moyens de la formation :Formation présentielle – 1 poste par stagiaire – 1 vidéo projecteur – Support de cours fourni à chaque stagiaire
- Modalités pédagogiques :Exposés – Cas pratiques – Synthèse
- Validation :Exercices de validation – Attestation de stages

Tél : +352 27 39 351
Fax : +352 27 39 52 68
Email : formation@oxiane.lu