Python – Formation Spark

Programme

Présentation d'Apache Spark

- Historique du Framework. Sa place dans les technologies Big Data
- Les différentes versions de Spark (Scala, Python et Java)
- Les différents modules de Spark
- Travaux pratiques : Installation et configuration de Spark, Scala et les librairies Python sur un PC Windows

Resilient Distributed Dataset (RDD)

- Présentation des RDD
- Créer, manipuler et réutiliser des RDD
- Utiliser des partitions
- Travaux pratiques : Manipulation des RDD en python, Création de programmes avec pyspark

Spark sur un Cluster

- Les différents types d'architecture : Standalone, Apache Mesos ou Hadoop YARN
- Déployer des applications avec Spark-submit
- Dimensionner un cluster
- Travaux pratiques : Mise en place d'un cluster Spark dans le Cloud (Amazon aws et Databricks), Exécution d'un programme Python sur un cluster Spark (Amazon aws)

Manipuler des données structurées avec Spark SQL

- SQL, DataFrames vs Datasets
- Les différents types de sources de données
- JDBC/ODBC server et Spark SQL CLI
- Travaux pratiques : Manipulation de Datasets via des requêtes SQL. Connexion avec une base externe via JDBC

Spark Streaming : Analyser des flux en temps réel

- Principe de fonctionnement.Gestion du cache
- Présentation des Discretized Streams (DStreams)
- Les différents types de sources
- Flume vs Kafka vs Storm
- Travaux pratiques : Consommation de logs avec Spark Streaming. Analyse de flux avec Spark et Kafka

Spark et les technologies Big Data

- Utiliser Spark avec des bases de données NoSQL (MongoDb, Cassandra, Neo4j)
- Spark vs Hadoop
- Spark et les Notebooks (Jupyter, Zeppelin)

Machine Learning avec Spark

- Introduction au Machine Learning
- Les principaux algorithmes
- Présentation de SparkML et MLlib

Référence

THBI1112

Durée

3 jours / 21 heures

Prix HT / stagiaire

1875€

Objectifs pédagogiques

- Concevoir les concepts fondamentaux de Spark
- Développer des applications avec Spark en Python
- Utiliser des données avec Spark SQL
- Faire des algorithmes de Machine Learning

Niveau requis

- Garantir posséder des connaissances en Python

Public concerné

- Chef de projets, Développeur, Consultant

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)

- Implémentations des différents algorithmes dans MLlib
- Travaux pratiques : Utilisation de SparkML et MLlib

Optimisation de Spark

- Gestion des variables partagées
- Données broadcastées
- Accumulateurs