Apache Spark : traiter les données

Cette formation vous apporte les compétences nécessaires pour mener à bien des analyses de données en tirant parti de l'écosystème Spark : bases de Spark et Hadoop, dataframes et des schémas, transformer et agréger les données avec RDD, applications Spark, traitement distribué et persistance, Spark SQL et Spark Streaming. Python (PySpark) est utilisé dans ce cours. *Pré-requis :* Python, SQL.

Modalité : à distance

Formule au choix : Live Training+ *ou bien* Classe virtuelle Teams ou Zoom

Durée totale : 21 H (3 jours)

PLAN DETAILLE

Introduction à Hadoop : #1

Le monde Big Data
Hadoop : architecture et composants
Le système HDFS
MapReduce et YARN

Le framework Spark : #2

Spark : historique, principe
Spark comparé à MapReduce
Spark : SQL, Streaming, MLlib, GraphX
RDD, DataFrames et Data Sets
Spark : CLI ou stand alone
Programmation de Spark
Utiliser Spark en Python : PySpark

Installation : #3

Spark en local
Sur un environnement distribué
En Cloud : AWS, Azure

Comprendre et utiliser RDD: #4

Contextes, sessions.
RDD, qu'est-ce que c'est?
RDD : créer, manipuler, réutiliser
Principales fonctions/transformations
Algorithmes de type map/reduce
Utiliser des partitions.
Soumission de travaux.

Machine Learning avec Spark.ML : #6

Introduction au ML
Différentes classes d'algorithmes
Apprentissage supervisé
Forêts aléatoires avec Spark
Mise en place d'un outil de recommandation
Traitement de données textuelles
Créer des pipelines et automatiser

Manipuler les données, Spark SQL: #5

DataFrames et Data Sets
Créer des DataFrames, PySpark Pandas
Charger les données : Hadoop, CSV, JSON,..
Transformer avec les DataFrames
Le tockage de données
Interopérabilité avec les RDD
Spark SQL : prise en main
TP : mise en oeuvre.

Apache Spark : traiter les données

Cette formation vous apporte les compétences nécessaires pour mener à bien des analyses de données en tirant parti de l'écosystème Spark : bases de Spark et Hadoop, dataframes et des schémas, transformer et agréger les données avec RDD, applications Spark, traitement distribué et persistance, Spark SQL et Spark Streaming. Python (PySpark) est utilisé dans ce cours. *Pré-requis :* Python, SQL.

Modalité : à distance

Formule au choix : Live Training+ *ou bien*
Classe virtuelle Teams ou Zoom

Durée totale : 21 H (3 jours)

PLAN DETAILLE

Spark Streaming: #7

Introduction et architecture
Discretized Streams (DStreams)
Les sources de données
Utilisation de l'API
Manipulation des données
Machine learning en temps réel.

Spark et les graphes: #8

GraphX : présentation
Principe de création des graphes
API GraphX
Présentation de GraphFrames
GraphX vs GraphFrames