PySpark - Traitement des données

3 j (21 heures) Ref : PYT-TD

Public

Développeurs, Data analysts, Data scientists, architectes Big Data et toute personne souhaitant acquérir des connaissances dans le domaine de la Data Science et sur Spark

Pré-requis

Avoir des notions de SQL et des connaissances de base en mathématiques et statistiques
Une première expérience en programmation Python est requise

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois
Un poste par stagiaire, vidéoprojecteur, support de cours fourni à chaque stagiaire

Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires
Questionnaire d'évaluation de la satisfaction en fin de stage
Auto-évaluation des acquis de la formation par les stagiaires
Attestation de fin de formation

Objectifs

Comprendre le principe de fonctionnement de Spark
Utiliser l'API PySpark pour interagir avec Spark en Python
Utiliser les méthodes de Machine Learning avec la librairie MLlib de Spark
Traiter les flux de données avec Spark Streaming
Manipuler les données avec Spark SQL

Programme détaillé

INTRODUCTION A HADOOP

L'ère du Big Data
Architecture et composants de la plateforme Hadoop
HDFS

NameNode / DataNode / ResourceManager

MapReduce et YARN

INTRODUCTION A SPARK

Qu'est-ce que Spark ?

Spark vs MapReduce

Fonctionnement

RDD

DataFrames

Data Sets

Comment interagir avec Spark ?

PySpark : programmer avec Spark en Python

INSTALLATION DE SPARK

Sur une infrastructure distribuée

En local

SPARK POUR LA MANIPULATION DES DONNEES - PYSPARK

Utilisation de SparkSQL et des DataFrames pour manipuler des données

Charger des données depuis Hadoop, depuis des fichiers CSV, texte, JSON...

Transformer des données (création de DataFrames, ajout de colonnes, filtres...)

UTILISATION DE SPARK.ML POUR LE MACHINE LEARNING

Apprentissage supervisé

Forêts aléatoires avec Spark

Mise en place d'un outil de recommandation

Traitement de données textuelles

Automatiser vos analyses avec des pipelines

SPARK STREAMING

Introduction à Spark Streaming

La notion de "DStream"

Principales sources de données

Utilisation de l'API

Manipulation des données

SPARK SQL

Initialisation à Spark SQL

Création de DataFrames

Manipulation des DataFrames (opérations basiques, agrégations et groupBy, missing data)

Chargement et stockage de données (avec Hive, JSON...)

DEMONSTRATION GRAPHX ET GRAPHFRAMES

Présentation de GraphX

Principe de création des graphes

API GraphX

Présentation de GraphFrames

GraphX vs GraphFrames