STAT4DECISION
conseil et formation data science

FORMATION SPARK AVEC PYTHON

REF. : PYTHONSPARK101

DESCRIPTION

L'environnement Apache Spark est aujourd'hui central dans l'approche big data de la donnée. Cette formation spark avec python vous permet de maîtriser les principes de l'environnement Apache Spark et l'utilisation de la bibliothèque pyspark pour gérer des données, appliquer des algorithmes de machine learning ou accélérer vos processus.

Cette formation spark s'adresse à tous ceux qui veulent manipuler Apache Spark en utilisant le langage Python.

OBJECTIFS

- Comprendre l'environnement Apache Spark
- Savoir utiliser le package PySpark pour communiquer avec Spark
- Maîtriser l'utilisation de Spark SQL
- Maîtriser l'utilisation de Spark.ml

PROGRAMME DETAILLE

Rappels sur Python et la manipulation des données

Introduction à l'environnement Big Data et à Spark

- Pour qui ? Pour quoi faire ? Comment ?
- Comment installer Apache Spark
- PySpark un package Python pour gérer votre environnement Apache Spark
- Quelle infrastructure pour utiliser Spark en entreprise ?
- Les principes de l'environnement : RDD, DataFrame, DataSet…

Installation de Spark :

- Sur une infrastructure distribuée
- En local
- En cloud (exemples avec Amazon AWS et Microsoft Azure)

Spark pour la manipulation des données

- Utilisation de SparkSQL et des DataFrames pour manipuler des données
- Charger des données depuis Hadoop, depuis des fihiers csv…

DUREE

2 jours

PUBLIC ET PREREQUIS

Public ayant des bases en programmation. Une connaissance de Python est fortement conseillée.

MOYENS PEDAGOGIQUES

Alternance d'exposés et d'applications pratiques avec des exercices sur des données.

PLUS D'INFORMATIONS

https://www.stat4decision.com/fr/formations/formation-spark-avec-python/

Formation disponible en intra ou en inter-entreprises

- Transformer des données (création de DataFrames, ajout de colonnes, filtres…)
- Cas pratiques de chargement et de modifications de données avec Spark et PySpark

L'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

Introduction et utilisation de Spark Streaming avec PySpark