Hadoop : Analytics

Référence : **PYCB036** Durée : **2 jours** Certification : **Aucune**

CONNAISSANCES PREALABLES

• Connaissances des principes du BigData, d'un langage de programmation comme Java ou Scala ou Python.

PROFIL DES STAGIAIRES

• Chefs de projet, développeurs, data scientists, architectes souhaitant mettre en œuvre des solutions analytics avec hadoop.

OBJECTIFS

• Savoir mettre en œuvre les frameworks analytics dans un environnement hadoop.

CERTIFICATION PREPAREE

Aucune

METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

FORMATEUR

Consultant-Formateur expert Bigdata

METHODE D'EVALUATION DES ACQUIS

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation de fin de stage adressée avec la facture

CONTENU DU COURS

Introduction

- Définitions : Analytics
- Arbres de décision, de régression, régression automatique
- Apprentissage supervisé, apprentissage automatique
- Présentation du data munging

Hadoop et les outils d'analyse

- Rôle des différents composants : socle hadoop, yarn, hdfs
- Frameworks analytics : Mahout, Flink, Spark ML

Mahout

- Principe de fonctionnement
- Sources de données, format de stockage des données
- Génération de recommandations, traitement, filtrage
- Exemples de base : génération de recommandations, traitement, filtrage
- Présentation des algorithmes les plus courants
- Compatibilité avec Hadoop Yarn, Spark, H2O, Flink

Flink

- Origine du projet, fonctionnalités
- Traitement distribué de flux de données, en temps réel ou batch
- APIs disponibles
- Mise en oeuvre avec des programmes Java/Scala
- Analyse de graphe avec l'API Gelly

Spark MLib

- Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistence,statistiques
- Support de RDD
- Mise en oeuvre avec les DataFrames

GraphX

- Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes
- Travaux pratiques : exemples d'opérations sur les graphes