Programmation R et intégration bigData

Référence : **PYCB052** Durée : **3 jours** Certification : **Aucune**

CONNAISSANCES PREALABLES

• Notions de calculs statistiques.

PROFIL DES STAGIAIRES

• Chefs de projet, data scientists, statisticiens, développeurs souhaitant comprendre les apports de R pour l'analyse des données, et savoir l'intégrer à un environnement Hadoop.

OBJECTIFS

• Connaître les principales fonctions statistiques de R, et savoir utiliser des programmes R dans un environnement BigData, en s'appuyant sur le système distribué hdfs.

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

Présentation R

- Le projet R Programming
- Calculs statistiques et génération de graphiques
- Points forts de R Programming
- Besoins du BigData
- Positionnement R programming par rapport à Hadoop

Mise en oeuvre de R

- Travaux pratiques : installation et tests sur une plate-forme CentOS
- Utilisation de R en mode commande
- Commandes de base. Syntaxe
- Opérations de base. Expressions
- Manipulations de nombres, vecteurs, tableaux, matrices.listes, etc.

Tableaux et matrices

- Déclaration, dimensionnement, indexation
- Opérations de base : produit de tableaux, transposition, produits de matrices
- Matrices : équations linéaires, inversion, valeur propre, vecteur propre, déterminant, moindre carré, ...

Liste et DataFrames

- Définitions, cas d'utilisation
- Attachement, détachement. Chargement d'un dataframe. La fonction scan

Statistiques

- Distributions embarquées : uniforme, normale, poisson, exponentielle, ...
- Calculs statistiques. Modèles statistiques
- Affichage en graphes, histogrammes

Import/export

- Formats texte, csv, xml, binaire, largeur fixe, images (jpeg, png). Encodage. Filtrage

- Importation SQL. Importation depuis un socket réseau
- Travaux pratiques : importation de données géodésiques et export au format Json

Intégration Hadoop

- Association de la puissance du calcul distribué fourni par les outils hadoop
- Différents moyens d'intégration : sparkR, RHbase, RHDFS, RHadoop, rmr2 pour utiliser le système distribué hdfs depuis R, pour accéder à HBase depuis les programmes en R
- Transformation d'un dataframe R en un dataframe Spark
- Travaux pratiques avec Hadoop

Fonctions spécifiques

- Définition de nouvelles fonctions. Appels. Passage d'argument
- Construction d'une bibliothèque
- Diffusion, installation avec R CMD INSTALL

Evolutions

- Les acteurs : IBM avec BigInsights, Revolution R avec ScaleR