

Programmation R et intégration Big Data

Comprendre les apports de R pour l'analyse des données et savoir l'intégrer à un environnement Hadoop

Langage R : Programmation R pour Hadoop.

Détails

- Code : DB-LR
- Durée : 2 jours (14 heures)

Public

- Chefs de projets
- Data Scientist
- Développeurs

Pré-requis

- Connaissances de base en statistiques et en programmation

Objectifs

- Connaître les principales fonctions statistiques de R
- Utiliser des programmes R dans un environnement Hadoop en s'appuyant sur le système distribué hdfs et le stockage avec HBase
- Intégrer R à un environnement Hadoop

Programme

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
- Manipulations de nombres, vecteurs, tableaux, matrices, listes, ...

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 et de la richesse des outils d'analyse statistique de R
- 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

Évolutions

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

Modalités

- **Type d'action** :Acquisition des connaissances
- **Moyens de la formation** :Formation présentielle - 1 poste par stagiaire - 1 vidéo projecteur - Support de cours fourni à chaque stagiaire
- **Modalités pédagogiques** :Exposés - Cas pratiques - Synthèse
- **Validation** :Exercices de validation - Attestation de stages