Pythagore F.D.
Apprendre à apprendre

Développement applications BigData et la dataviz

CB500

Durée: 3 jours

1750 €

18 au 20 juin

1er au 2 octobre
10 au 12 décembre

Public:

Concepteurs, développeurs, chefs de projet, et toute personne souhaitant comprendre et pratiquer les méthodes, outils et langages de développement BigData

Objectifs:

Comprendre les concepts et l'apport du Big Data par rapport aux enjeux métiers. Comprendre et pratiquer les méthodes, outils et langages de développement BigData et de datavisualisation, faire le lien avec les équipes d'infrastructure pour concevoir et développer des applications sur un socle distribué.

Connaissances préalables nécessaires:

Il est demandé aux participants d'avoir une bonne culture générale sur les systèmes d'information (systèmes, réseaux, architectures cloud), de maîtriser un langage de développement objet.

Programme:

Définition et contexte spécifique des projets BigData :
Panorama technologique et enjeux socio-économiques
Propriété de la donnée, environnement juridique du traitement, sécurité
Impact des choix technologiques en matière d'analyse et de visualisation de données.
L'essentiel du BigData : calcul distribué, données non structurées.
Besoins fonctionnels et caractéristiques techniques des projets.
La valorisation des données.
Le positionnement respectif des technologies de cloud, BigData et noSQL, et les liens, implications.

L'environnement : Hadoop :
Les distributions Hadoop : hortonworks, MapR, Cloudera, ...
Les fonctionnalités du framework Hadoop
Le projet et les modules :
Hadoop Common, HDFS, YARN, Spark, MapReduce

Pythagore F.D.

Apprendre à apprendre

Développement applications BigData et la dataviz

CB500

Le développement : MapReduce	:	Principe et objectifs du modèle de programmation MapReduce. Fonctions map() et reduce(). Couples (clés, valeurs). Implémentation par le framework Hadoop. Etude de la collection d'exemples. Configuration des jobs, notion de configuration. Les interfaces principales : mapper, reducer, La chaîne de production : entrées, input splits, mapper, combiner, shuffle/sort, reducer, sortie. partitioner, outputcollector, codecs, compresseurs.. Format des entrées et sorties d'un job MapReduce : InputFormat et OutputFormat. Type personnalisés : création d'un writable spécifique. Utilisation. Contraintes. Travaux pratiques: rédaction d'un premier programme et exécution avec Hadoop.

Pythagore F.D.
Apprendre à apprendre

Développement applications BigData et la dataviz

CB500

Langages de programmation : Présentation des langages utilisés dans les applications BigData : Python, R, Scala,...
Zoom sur 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.
Exemples de mise en oeuvre :
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,etc ..
Intégration avec 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 :
RHive : fonctions R de calculs statistiques s'appuyant sur HiveQL
RHadoop : packages rmr2,
rhdfs pour utiliser le système distribué hdfs depuis R,
rhbase pour accéder à HBase depuis les programmes en R.
Travaux pratiques avec Hadoop :
Installation d'un cluster,
rmr2:traduction programmes R en mapreduce,
rhdfs:API d'accès R à des données stockéss sur HDFS
rhbase:API d'accès à des données stockées sur HBase

Le Deep Machine Learning : Le besoin : types de données,
exemples de démarches et d'analyse
Définition, les attentes par rapport au Machine Learning
Les valeurs d'observation, et les variables cibles.
Ingénierie des variables.
Les différentes méthodes : apprentissage supervisé, apprentissage automatique.
Classification des données,
Algorithmes : régression linéaire, k-voisins,classification naïve bayésienne, arbres de décision, etc ..
Mise en oeuvre : classifieurs. scoring

Pythagore F.D.

Apprendre à apprendre

Développement applications BigData et la dataviz

CB500

La Data Visualisation : Fonctionnalités des outils de dataviz :
analyses statistiques,
classifications, rapprochements,
production de recommandations,
représentations graphiques,
Présentation de quelques outils : Mahout, Giraph, Agile, spagobi
Présentation de Mahout.
Positionnement dans l'offre BigData et Machine Learning : Hadoop, Spark,..
Fonctionnalités.
Mode autonome ou mode distribué
Exemples d'algorithmes fournis avec Mahout.