Pythagore F.D.

Apprendre à apprendre

Techniques d'analyse et de visualisation BigData

CB400

Durée: 3 jours

1750 €

2 au 4 mai

10 au 12 septembre

26 au 28 novembre

Public:

Concepteurs, développeurs, data/business analysts, data scientists, et toute personne souhaitant comprendre et pratiquer les méthodes, outils et langages d'analyse et de visualisation

Objectifs:

Comprendre les concepts et l'apport du Big Data par rapport aux enjeux métiers.Comprendre et pratiquer les méthodes, outils et langages d'analyse et de visualisation dans un projet BigData.

Connaissances préalables nécessaires:

Il est demandé aux participants d'avoir une bonne culture générale sur les systèmes d'information, des connaissances sur les formats et outils classiques de stockage de données, des notions de base sur les méthodes et outils statistiques.

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.

Pythagore F.D.

Apprendre à apprendre

CB400

Techniques d'analyse et de visualisation BigData

Gérer la structure de données : Données structurées / non structurées
Origine des bases de données, les notions de transaction, les SGBD, la standardisation SQL.
Caractéristiques NoSQL :
adaptabilité, extensibilité,
structure de données proches des utilisateurs, développeurs
Les types de bases de données : clé/valeur, document, colonne, graphe.
Données structurées et non structurées, documents, images, fichiers XML, JSON, CSV, ...
Les différents modes et formats de stockage.
Stockage réparti : réplication, sharding, gossip protocol, hachage,
Systèmes de fichiers distribués : GFS, HDFS,
Quelques exemples de produits et leurs caractéristiques :
Cassandra, MongoDB, CouchDB, DynamoDB, Riak, Hadoop, HBase, BigTable, ...
Protocoles d'accès aux données, interfaces depuis les langages classiques.
Parallélisation des traitements : implémentation de MapReduce.
Cohérence des données et gestion des accès concurrents :
"eventual consistancy" et multi-version concurrency control.
Démonstrations avec Cassandra et MongoDB.

La collecte de données : Production de données classiques,
gestion de flux de données.
Le besoin d'indexation : définitions et techniques d'indexation.
Positionnement Elasticsearch et les produits complémentaires : Shield, Watcher, Marvel, Kibana,
Logstash, Beats
Principe : base technique Lucene et apports d'ElasticSearch.

Pythagore F.D.

Apprendre à apprendre

CB400

Techniques d'analyse et de visualisation BigData

L'entrepôt / le stockage de données :

Architectures de stockage.
Principe des entrepôts de données.
Le stockage réparti.
Définitions : réplication, sharding, protocole gossip, hachage,..
Exemple de Cassandra : notion de noeud, de grappe.
Exemple de MongoDB : structure des données :notions de documents, de collections
Le format BSON (Binary JSON), comparaison avec JSON
Fonctionnalités de MongoDB.
Interfaces disponibles.

Méthode d'analyse et de visualisation :

Les requêtes : différents outils selon les bases de données .
Comment écrire des requêtes? Les différentes approches.
Présentation du langage de requêtes CQL de Cassandra.
Exécution de scripts.
Présentation du shell MongoDB.
Les outils Pig et Hive : fonctionnalités,
le langage latin.
Rédaction d'exemples simples de scripts pig.
Les différentes méthodes d'analyse.
Algorithmes : régression linéaire, k-voisins,classification naïve bayésienne, arbres de décision, etc ..
Mise en oeuvre : classifieurs. scoring
Fonctionnalités des outils de dataviz :
analyses statistiques,
classifications, rapprochements,
production de recommandations,
représentations graphiques,
Présentation de quelques outils : Mahout, Giraph, Agile, spagobi
Travaux pratiques : mise en oeuvre de Mahout en mode distribué.
Exemples d'algorithmes fournis avec Mahout.