www.inow- formation.com
contact@inow- formation.com
01 77 49 46 50

Formation Big Data - Fondamentaux

Acquérir les connaissances pour exploiter les nouveaux outils dédiés au Big Data

Pré-requis :

Avoir des connaissances pratiques de la plateforme Microsoft Windows. Avoir des notions de programmation

Référence : BIG1

Durée : 3 jours

Niveau : Débutant

Tarif : 1590 € HT

Objectifs :

§ Sélectionner des entrepôts de Big Data adaptés pour gérer plusieurs ensembles de données
§ Traiter des ensembles de données volumineux avec Hadoop pour faciliter la prise de décisions techniques et métier
§ Interroger des ensembles de données volumineux en temps réel

Programme du cours :

Big Data : enjeux et opportunités

Evolution des données
Le Big Data et ses 5 grands défis : volume, variété, vélocité, véracité, validité.
Données massives : Web, réseaux sociaux, Open Data, capteurs, données scientifiques.
Ouverture des données publiques : le mouvement Open Data.
Interconnexion des données : le Linked Open Data.
Variété, distribution, mobilité des données sur Internet.
Vélocité et flux continus de données.

Les enjeux pour les entreprises
Véracité et validité des données provenant de sources variées pour la prise de décision.
Analyses complexes sur Big Data, Big Analytics.
Production d'informations en temps réel à partir de Big Data.
Croisement et visualisation de données publiques et privées.
Réactivité : traitement de flux de données en temps réel, Complex Event Processing (CEP).
Exemples de succès et d'échecs de projets Big Data.
Cloud et Big Data : le mariage parfait ?

Opportunités offertes par les progrès matériels
Le stockage : mémoires flash, disques HDD versus SSD, la nouvelle hiérarchie de mémoires.
Bientôt 1 teraoctet de RAM sur un chip : l'avènement du traitement de données in-memory ?
Processeurs multi-cœurs, la combinaison CPU/GPU, la nouvelle hiérarchie du calcul parallèle.
Le stockage disque en réseau NAS/SAN : impact sur les architectures de gestion de données?
Les architectures massivement parallèles (MPP) : speed-up, scale-up, scale-out, élasticité.
Cloud et microserveurs.

Architectures de bases de données

Concepts de base
Partage de données, définition et évolution de schéma, cohérence et protection des données.
Requêtes, transactions, vues, contraintes d'intégrité et triggers.
Optimisation et réglage, l'importance du placement et des index.
Le modèle ACID (Atomicité, Cohérence, Isolation, Durabilité) des transactions.
Transactions distribuées : le protocole 2PC, tolérance aux pannes et scalabilité.
Réplication de données : cohérence des copies, propagation des mises à jour.

Modèles de données
Le modèle relationnel : domaine de valeurs, relation, algèbre et calcul, le concept de valeur nulle.
SQL2 : les types de données, les niveaux d'isolation, la portabilité.
SQL3 : tables imbriquées, types complexes et extensions objet.
Nouveaux modèles : clé-valeur, tabulaire, document, graphe, stream.

L'analyse de données
Décisionnel et OLAP : le benchmark TPC-H, analyse multi-dimensionnelle.
Business Intelligence et data mining : extraction de connaissances à partir des données.

Architectures des SGBD
Parallélisme de données : inter-requête, inter-opération, intra-opération, pipeline.
Architectures MPP: SMP et NUMA, cluster shared-disk, cluster shared-nothing.

Architectures Big Data

Motivations
La fin de l'approche « taille unique » du relationnel.
Architecture 3-tiers dans le cloud.
Le théorème CAP (Consistency, Availability, Partition tolerance) : analyse et impact.

La pile logicielle big data
Les niveaux fonctionnels : stockage, organisation, traitement, intégration, outils d'analyse.
La gestion de clusters.
L'architecture Hadoop, comparaison avec l'architecture Lambda.
Comparaison avec les SGBD relationnels.

Techniques de base
Organisation des données : en ligne ou en colonne.
Placement des données : partitionnement et sharding, réplication, indexation.
Parallélisation des requêtes, équilibrage de charge.
Haute disponibilité : le failover, les points de sauvegarde pour requêtes lourdes.

Stockage de Big Data

Stockage d'objets

Stockage en fichiers distribués
Systèmes de fichiers distribués : Hadoop HDFS, Google File System, IBM GPFS, GlusterFS, Lustre.

Stockage clé-valeur
Systèmes clé-valeur : Amazon DynamoDB, Amazon SimpleDB, Apache Cassandra, Linkedin Voldemort.

Les principaux SGBD NoSQL

SGBD tabulaires
Modèle de données : table, orienté ligne/colonne, opérateurs ensemblistes.
Architecture : partitionnement et réplication de tables, stockage en fichiers.
Exemples : Google Bigtable sur GFS, Hadoop Hbase sur HDSF, Apache Accumulo.

SGBD orientés- documents

SGBD orientés- graphes
Modèle de données : graphe, RDF, opérateurs de parcours de graphes, langages de requêtes.
Architecture : partitionnement et réplication de graphes, stockage en fichiers, index.

Exemples : Neo4J, DEX/Sparksee, AllegroGraph, InfiniteGraph, IBM DB2 Sparql.

Intégration SQL/NoSQL
Le relâchement de la cohérence : problèmes pour les développeurs et les utilisateurs.
NoSQL versus relationnel. L'intégration SQL/NoSQL avec Google F1.