Big Data – De la stratégie à la mise en oeuvre

Description :

La formation a pour objet de brosser sans concession le tableau du Big Data.

Les participants repartiront de cette formation en ayant une vision claire de la stratégie et de l'éventuelle mise en œuvre d'un Big Data.

Objectifs

- Présentation du Big Data
- Stockage du Big Data
- Techniques et outils d'analyse et de traitement du Big Data
- Le traitement des Big Data
- L'accès des données en temps réel
- La protection des données
- Mise en œuvre d'une stratégie dédiée au Big Data

Publics

Architectes, Chefs de Projet, Directeur informatique, DSI

Durée

3 jours

Pré-requis

Un niveau technique minimal et de management des SI sont requis pour tirer pleinement parti de cette formation.

Programme de cette formation

Définition

- Les quatre dimensions du Big Data : volume, vélocité, variété, véracité
- Présentation de l'ensemble MapReduce, stockage et requêtes

Améliorer les résultats de l'entreprise grâce au Big Data

- Mesurer l'importance du Big Data au sein d'une entreprise
- Big Data et la performance de l'entreprise
- Clients, produits, processus, infrastructures : les nouveaux enjeux de performance des entreprises.
- L'analyse des données au service de la performance : comment identifier les nouveaux leviers de performance ?
- En quoi Big Data est un enjeu de performance pour les entreprises et les organisations ?
- Réussir à extraire des données utiles
- Intégrer le Big Data aux données traditionnelles

Sources de données publiques et privées

- Panorama des données publiques disponibles (Open Data) : données économiques, données individuelles, données produits.
- Comment exploiter les données des réseaux sociaux ?
- La perception des tendances de marché : l'analyse des sentiments du marché et des influences.
- Comment croiser données publiques et privées ?

Traitements des données non structurées

- Les types de données non structurées : message, document, semi-structuré.
- Les principes de l'analyse sémantique : sentiment, univers, corrélation.
- Les outils d'analyse sémantique et de recherche de corrélations.

Machine Learning pour les données

- Les principales méthodes descriptives et prédictives.
- Le Machine Learning : arbre de décision, règle d'association, Support Vector Machines.
- Les spécificités du Machine Learning pour le Big Data : haute dimension, occurrences élevées.
- Les meilleures pratiques pour le Machine Learning : cross-validation, grid search, semi-supervision.
- Domaines d'application privilégiés : marketing, finance, e-commerce.

Visualisation des données à valeur ajoutée

- Les limites des outils traditionnels d'analyse de données : Excel, BO, etc.
- Comment représenter efficacement des données analytiques ?
- Les outils et composants de visualisation des données Big Data.Analyser les

Analyser les caractéristiques de vos données

- Sélectionner les sources de données à analyser
- Supprimer les doublons
- Définir le rôle de NoSQL
- Partitionnement des données NoSQL : répartition sur les nœuds, redondance, élasticité.
- La granularité de cohérence des données : les différents niveaux de cohérence.

Présentation des entrepôts de Big Data

- Modèles de données : valeur clé, graphique, document, famille de colonnes
- Système de fichiers distribué Hadoop (HDFS)
- La technologie HDFS : principes et fonctionnement.
- Manipulation des données : Hive, Pig, Streaming, SQL/Hadoop

Choisir un entrepôt de Big Data

- Choisir un entrepôt de données en fonction des caractéristiques de vos données
- Injecter du code dans les données, mettre en œuvre des solutions de stockage des données multilingues
- Choisir un entrepôt de données capable de s'aligner avec les objectifs de l'entreprise

Archivage et sécurité

- L'évolution de l'archivage pour les Big Data.
- Comment maintenir la compatibilité dans le temps ?
- La réplication sur plusieurs datacenters et le filtrage intelligent.

Infrastructure de stockage

- Anatomie d'un datacenter de stockage Big Data : enjeux de coûts et de capacité.
- Quelles sont les contraintes hardware ? Système, stockage, réseau, place.
- Commodity hardware ou hardware spécialisé ? Les critères du choix.
- Technologies Big Data et Cloud Computing : comment utiliser les offres de Cloud publiques et privées ?

Introduction à NoSQL

- Le paradigme NoSql
- NoSql et la scalabilité
- « no SQL » vs « Not Only SQL »
- Les types de bases de données NoSql (stockage clé-valeur, graphe, Map-Reduce, ...)

Introduction à Hadoop

- Qu'est ce qu'Hadoop?
- L'écosystème Hadoop : Pig, Hive, HBase, Zookeeper...
- Comprendre MapReduce et HDFS (Hadoop Distributed File System)
- S'assurer de l'intégrité des données (checksum...)
- Gagner de la place : compression des données d'entrée/sortie dans Hadoop
- L'Openstack (Ceph)
- Le Complex Event Processing

Programmation parallèle

- Les limites de la programmation séquentielle traditionnelle.
- Les objectifs et les contraintes de la programmation parallèle.
- Les différents types de système de traitements parallèles : Grid, GPU et multi-core.

Architecture Grid Computing

- Architecture d'une grille de calculs.

- Comment accéder aux données dans une grille de calculs ?
- Les avantages et les limites du Grid Computing.

Architecture GPU et multi-core

- Comprendre les évolutions CPU et GPU : panorama des évolutions.
- Les architectures GPU et leurs limites.
- Les architectures multi-core et leurs limites.

Machine Learning en environnement parallèle

- Les contraintes de parallélisme des algorithmes de Machine Learning.
- Les librairies de Machine Learning : Mahout, Hama, Scikit Learn, R.

Bases de données distribuées, NoSQL et InMemory

- Quelles bases de données choisir : InMemory, Verticale, NoSQL ?
- Les bases de données Clé/Valeur : SimpleDB, Riak, Redis.
- Les bases de données Document : MongoDB, CouchDB.
- Les bases de données Colonne : HBase, Cassandra, XVelocity.
- Les bases de données Graphe : Neo4j, HyperGraphDB.

Le transactionnel Big Data

- Comment faire converger système opérationnel et Machine Learning ?
- Comment gérer l'apprentissage permanent ?
- Temps reel avec Hadoop : Stinger, Impala, Drill.
- Convergence Hadoop.
- MPP : MS Polybase, Asterdata.

Le streaming en grille : Storm

- Les techniques de streaming.
- Les outils du marché CEP pour Big Data : Storm.
- Concevoir des applications pour la prise de décision en temps réel avec Storm.

Le streaming pour le Big Data

- Comment collecter et contrôler les données au fil de l'eau, avec un grand débit ?
- Comment gérer les chargements en masse ?
- Les technologies low latency messaging : Kafka.
- Les caractéristiques du streaming dans un environnement Big Data.

Le streaming en grille : Storm

- Les techniques de streaming.
- Les outils du marché CEP pour Big Data : StreamInsight, Storm.
- Comment concevoir des applications pour la prise de décision en temps réel ?

Relations entre Cloud et Big Data

- Motivations des Clouds publics et privés.
- Les services XaaS.

- Les objectifs et avantages des architectures Clouds.
- les infrastructures.
- Les égalités et les différences entre Clouds et Big Data.
- Les Clouds de stockage.

Élaborer une stratégie dédiée au Big Data

- Définir les besoins en matière de Big Data
- Atteindre les objectifs grâce à la pertinence des données
- Évaluer les différents outils du marché dédiés au Big Data
- Répondre aux attentes du personnel de l'entreprise

Une méthode analytique innovante

- Identifier l'importance des traitements métier
- Cerner le problème
- Choisir les bons outils
- Obtenir des résultats exploitables

Analyse statistique du Big Data

- Exploiter la fonctionnalité RHadoop
- Générer des états statistiques avec RHadoop
- Utiliser la visualisation RHadoop
- Exploiter les résultats des analyses

L'efficacité des diverses technologies

- La conservation dans le temps face aux accroissements de volumétrie.
- La sauvegarde, en ligne ou locale ?
- L'archive traditionnelle et l'archive active.
- Les liens avec la gestion de hiérarchie de stockage : avenir des bandes magnétiques.
- La réplication multisites.
- La dégradation des supports de stockage

Classification, sécurité et confidentialité des données

- La structure comme critère de classification : non structurées, structurées, semi structurées.
- Selon le cycle de vie : temporaires, permanentes, archives actives.
- Difficultés supplémentaires liées aux augmentations de volumétries.
- Les solutions potentielles.

Mettre en œuvre une solution Big Data

- Bien choisir les fournisseurs et options d'hébergement
- Trouver le juste équilibre entre les coûts engendrés et la valeur apportée à l'entreprise
- Garder une longueur d'avance

Rôle de la DSI dans Big Data

- La DSI comme fournisseur de services IT ou fournisseur de données à valeur ajoutée ?
- Comment, concrètement, le DSI peut-il saisir l'opportunité Big Data ?
- Quels sont les nouveaux challenges pour la DSI ?
- Les acteurs métier et leurs rôles.

- La gouvernance et les indicateurs de pilotage du projet.

Création de valeur à partir des données

- Comment identifier les données à valeur ajoutée ?
- Comment tirer profit des données clients, produits ou de suivi ?
- L'anticipation, la sécurité et les recommandations.
- Analyses marketing et analyses d'impact.
- Distribution de la donnée.
- PRA. Recours aux Clouds publics
- Quel avenir pour les données ?
- Gouvernance du stockage dans l'entreprise ?
- Grille d’analyse de la valeur des données et des objectifs d’analyse.