Pythagore F.D.

Apprendre à apprendre

CB001

BigData Architecture et technologies

Durée: 2 jours

1160 €	14 au 15 février	5 et 6 septembre
	30 au 31 mai	14 au 15 novembre

Public:

Chefs de projets, architectes, développeurs, data-scientists, et toute personne souhaitant connaître les outils et solutions pour concevoir et mettre en oeuvre une architecture BigData.

Objectifs:

Comprendre les concepts essentiels du BigData, et les technologies implémentées. Savoir analyser les difficultés propres à un projet BigData, les freins, les apports, tant sur les aspects techniques que sur les points liés à la gestion du projet.

Connaissances préalables nécessaires:

Il est demandé aux participants d'avoir une bonne culture générale sur les systèmes d'information.

Programme:

Pythagore F.D.

Apprendre à apprendre

CB001

BigData Architecture et technologies

Introduction : 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.
Quelques éléments d'architecture.
Concepts clés : ETL, Extract Transform Load, CAP, 3V, 4V, données non structurées, prédictif, Machine Learning.
Quelques applications : Watson (IBM), Amazon Rekognition
L'écosystème du BigData : les acteurs, les produits, état de l'art.
Cycle de vie des projets BigData.
Emergence de nouveaux métiers : Datascientists, Data labs, Hadoop scientists, CDO, ...
Rôle de la DSI dans la démarche BigData
Gouvernance des données :
importance de la qualité des données, fiabilité, durée de validité, sécurité des données
Aspects législatifs : sur le stockage, la conservation de données, etc ...
sur les traitements, la commercialisation des données, des résultats

Pythagore F.D.

Apprendre à apprendre

CB001

BigData Architecture et technologies

Stockage distribué : Caractéristiques NoSQL
Les différents modes et formats de stockage.
Les types de bases de données : clé/valeur, document, colonne, graphe.
Besoin de distribution. Définition de la notion d'élasticité.
Principe du stockage réparti :
Définitions : réplication, sharding, gossip protocol, hachage,
Systèmes de fichiers distribués : GFS, HDFS, Ceph
Les bases de données : Cassandra, DynamoDB, Accumulo, HBase, MongoDB, CouchBase, Riak, BigTable, ..
Caractéristiques NoSQL :
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, ...
Qualité des données, gouvernance de données.

Indexation et recherche : Moteurs de recherche.Principe de fonctionnement.
Méthodes d'indexation. Mise en oeuvre avec elasticsearch.
Exemple de Lucene/solr.
Recherche dans les bases de volumes importants.
Exemples de produits et comparaison : Dremel, Drill, ElasticSearch, MapReduce,

Pythagore F.D.

Apprendre à apprendre

CB001

BigData Architecture et technologies

Calcul et restitution, intégration	: Différentes solutions : calculs en mode batch, ou en temps réel, sur des flux de données ou des données statiques. Les produits : langage de calculs statistiques, R Statistics Language, sas, RStudio; outils de visualisation : Tableau, QlikView Ponts entre les outils statistiques et les bases BigData Outils de calcul sur des volumes importants : storm en temps réel, hadoop en mode batch. Zoom sur Hadoop : complémentarité de HDFS et MapReduce. Restitution et analyse : logstash, kibana, elk, pentaho Présentation de pig pour la conception de tâches MapReduce sur une grappe Hadoop.