

## Big Data

Mettre en oeuvre un projet Big Data pour tirer le meilleur parti des données

Les "Big Data", aussi appelées mégadonnées, désignent des ensembles volumineux de données, venant de sources et de formats différents, et qui doivent être analysées en temps réel pour pouvoir fournir des éléments cruciaux de décisions ou encore offrir la meilleure expérience à vos utilisateurs.

Les outils conventionnels de gestion de bases de données et d'analyse ne permettent plus de suivre ces trois propriétés de volume, de variété et de vélocité. A l'issue de cette formation, vous connaîtrez les enjeux du Big Data, et vous saurez choisir les bon outils, autant en termes de bases NoSQL, d'algorithmes MapReduce, de stockage ou d'analyse pour réaliser vos projets Big Data.

### Détails

- Code : DB-BIG
- Durée : 2 jours ( 14 heures )

#### Public

- Chefs de projets
- Architectes

#### Pré-requis

- Savoir manipuler un tableur et connaître le fonctionnement d'un système de gestion de base de données

### Objectifs

- Comprendre les concepts du BigData et savoir quelles sont les technologies implémentées

### Programme

#### Introduction au Big Data

- Le besoin : volumes importants de données, traitements optimisés de flux de données au fil de l'eau, liés aux nouvelles technologies et aux nouveaux usages
- Domaines concernés : recherche scientifique, médical, e-commerce, sécurité, ...
- Développement des techniques sur différents aspects : stockage, indexation/recherche, calcul
- Définition ETL : Extract Transform Load
- Les acteurs

#### Stockage

- Caractéristiques NoSQL :
- Structure de données proches des utilisateurs, développeurs
- 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, HBase, BigTable, ..
- Les bases de données
- Quelques exemples de produits et leurs caractéristiques : cassandra, MongoDB, CouchDB, DynamoDB, Riak, Hadoop

#### Indexation et recherche

- Moteurs de recherche
- Principe de fonctionnement
- Méthodes d'indexation
- Exemple de Lucene, et mise en oeuvre avec solr
- Recherche dans les bases de volumes importants
- Exemples de produits et comparaison : dremel, drill, elasticsearch, MapReduce

#### 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
- Outils de calcul sur des volumes importants : storm en temps réel, hadoop en mode batch
- Zoom sur Hadoop : complémentarité de HDFS et MapReduce

#### Evolutions

- Les offres Saas BigData comme Google BigQuery
- Les limites. Les nouveautés annoncées

### Modalité

- Séminaire en présentiel
- Séminaire en distanciel
- Nombre de stagiaires minimum : 4
- Nombre de stagiaires maximum : 10

### Méthodes pédagogiques

- Exposés
- Echanges d'expérience

### Profils des intervenants

- Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

### Modalités d'évaluation

- Evaluation des acquis de la formation par le biais de cas pratiques et/ou mises en situation.
- Attestation de formation remise à chaque participant.

### Démarche qualité

- Questionnaire d'évaluation de satisfaction à chaud complété par chaque participant à l'issue de la formation.

### Moyens pédagogiques

- Salle équipée de PC (1 poste par stagiaire), vidéo-projecteur.
- Espace de pause.

Dernière mise à jour le 26/02/2020