



## Etat de l'art du BigData

CB000

Durée: 1 jour

830 €

31 mars  
30 juin

29 septembre  
17 novembre

### Public :

Directeurs SI, Responsables SI, Chefs de projets, Architectes, Consultants ou toute personne amenée à participer à un projet Big Data ...

### Objectifs :

A l'issue de la formation, le stagiaire sera capable d'intégrer efficacement dans une stratégie d'entreprise les enjeux organisationnels et techniques du Big Data, en disposant d'une vue d'ensemble de son écosystème.

### Connaissances préalables nécessaires :

Avoir des connaissances de base des architectures techniques et du fonctionnement d'un système de gestion de base de données (SGBD).

### Objectifs pédagogiques :

Découvrir les principaux concepts du Big Data  
Appréhender les avantages et les contraintes du Big Data  
Comprendre les enjeux économiques du Big Data  
Connaître l'écosystème du Big Data et appréhender les technologies associées  
Savoir anticiper son intégration dans les activités informatiques de l'entreprise  
Prendre en compte la sécurité et la confidentialité des données dans l'exploitation du Big Data

### Programme :

#### Découvrir les principaux concepts du Big Data

Besoins fonctionnels et caractéristiques techniques des projets.  
Concepts clés : ETL/ELT, Extract Transform Load, CAP, 3V, 4V, données non structurées, semi-structurées.  
Le positionnement respectif des technologies de cloud, BigData et noSQL, et les liens, implications.

Atelier : Amazon Rekognition, Polly, EMR.

#### Appréhender les avantages et les contraintes du Big Data

L'essentiel du BigData : calcul distribué, données non structurées.  
La valorisation des données. Prédicatif, Machine Learning.

Atelier : Mise en oeuvre d'un cas d'utilisation de choix : quand utiliser le BigData



# Phirio

---

## Comprendre les enjeux économiques du Big Data

Les moteurs de recherche. Méthodes d'indexation. Recherche dans les bases de volumes importants. Présentation de Elasticsearch et SolR.  
Les lacs de données : caractéristiques NoSQL.  
Le calcul distribué et les capacités d'analyse. Vers l'apprentissage automatique (Machine Learning) et l'intelligence artificielle.  
Les nouveaux métiers dans l'entreprise : DataScientists, DataEngineer, DataLabs, ...

Atelier : comparaison du mode de fonctionnement de quelques grands acteurs du BigData

---

## Connaître l'écosystème du Big Data et appréhender les technologies associées

L'écosystème du BigData : les acteurs, les produits, état de l'art. Cycle de vie des projets BigData.  
Systèmes de fichiers distribués : GFS, HDFS, Ceph. Les bases de données : Redis, Cassandra, DynamoDB, Accumulo, HBase, MongoDB, BigTable, Neo4j, ...  
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.  
Données structurées et non structurées, documents, images, fichiers XML, JSON, CSV, ...  
Stockage distribué : réplication, sharding, gossip, hachage,  
Principe du schemaless, schéma de stockage, clé de distribution, clé de hachage

Atelier : construction d'une architecture BigData à partir de briques de bases de l'écosystème pour répondre à un cas d'utilisation donné

---

## Savoir anticiper son intégration dans les activités informatiques de l'entreprise

Emergence de nouveaux métiers : Data-scientists, Data labs, Hadoop scientists, CDO, ...  
Intégration avec les outils statistiques présents et les outils BigData futurs. Outils de calcul sur des volumes importants : Kafka/Spark Structured Streaming/Storm en temps réel, Hadoop/Spark en mode batch.  
Intégration avec les produits de DataScience : langage de calculs statistiques, R Statistics Language, sas, RStudio; outils de visualisation : Tableau, QlikView  
Intégration avec les produits de restitution et d'analyse : Logstash, Kibana, ElasticSearch, Zeppelin

Atelier : démonstration d'un ETL SQL/Hadoop

---

## Prendre en compte la sécurité et la confidentialité des données dans l'exploitation du Big Data

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, ... sur les traitements, la commercialisation des données, des résultats

Atelier : mise en évidence des problèmes liés à la réplication inter-régions et concernant les aspects juridiques des données : droits d'exploitation, propriété intellectuelle, ...