Phirio

Les architectures et infrastructures pour le Bigdata

CB001

Durée: 2 jours

1 700 €

12 au 13 mars

26 au 27 mai
10 au 11 septembre
9 au 10 novembre

Public :

Chefs de projets, Architectes, Développeurs, Data Scientists ou toute personne souhaitant connaître les outils pour concevoir une architecture Big Data, ...

Objectifs :

A l'issue de la formation, le stagiaire sera capable d'initier la conception d'une architecture et d'une infrastructure Big Data, depuis la mise en place d'un datalake jusqu'à l'exploitation des données, en disposant d'une vue d'ensemble des différentes solutions dédiées au traitement des données de masse.

Connaissances préalables nécessaires :

avoir une bonne culture générale des systèmes d'information et plus particulièrement, avoir des connaissances de base des modèles relationnels, des statistiques et des langages de programmation.

Objectifs pédagogiques :

Comprendre les principaux concepts du Big Data ainsi que l'écosystème technologique d'un projet Big Data
Savoir analyser les difficultés propres à un projet Big Data
Déterminer la nature des données manipulées
Appréhender les éléments de sécurité, d'éthique et les enjeux juridiques
Exploiter les architectures Big Data
Mettre en place des socles techniques complets pour des projets Big Data.
Concevoir l'infrastructure d'un datalake : collecte, stockage, traitement des données.

Programme :

Comprendre les principaux concepts du Big Data ainsi que l'écosystème technologique d'un projet Big Data

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.
Concepts clés : ETL, Extract Transform Load, CAP, 3V, 4V, données non structurées, prédictif, Machine Learning.
L'écosystème du BigData : les acteurs, les produits, état de l'art. Cycle de vie des projets BigData.

Savoir analyser les difficultés propres à un projet 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
Emergence de nouveaux métiers : Data-scientists, Data labs, Hadoop scientists, CDO, ...
Intégration avec les outils statistiques présents et les outils BigData futurs.

Déterminer la nature des données manipulées

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, ...

Atelier : installation d'un lac de données Hadoop/HDFS

Atelier : démonstrations avec une base MongoDB et une base Cassandra sur des données de différents types.

Appréhender les éléments de sécurité, d'éthique et les enjeux juridiques

Les risques et points à sécuriser dans un système distribué.
Aspects législatifs et éthiques: 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, ...

Etude des failles de sécurité sur une infrastructure Hadoop.

Exploiter les architectures Big Data

Les objectifs de la supervision, les techniques disponibles. La supervision d'une ferme BigData.
Objets supervisés. Les services et ressources. Protocoles d'accès. Exporteurs distribués de données.
Définition des ressources à surveiller. Journaux et métriques.
Application aux fermes BigData : Hadoop, Cassandra, HBase, MongoDB
Besoin de base de données avec agents distribués, de stockage temporel (timeseriesDB)
Produits : Prometheus, Graphite, ElasticSearch.
Présentation, architectures.
Les sur-couches : Kibana, Grafana.

Atelier : mise en oeuvre de prometheus pour la supervision d'une ferme Cassandra sur une infrastructure distribuée multi-noeuds.

Mettre en place des socles techniques complets pour des projets Big Data.

Etude des différents composants d'une infrastructure BigData :
Datalake : collecte des différents types de données
Stockage distribué : réplication, sharding, gossip, hachage,
Principe du schemaless, schéma de stockage, clé de distribution, clé de hachage
Systèmes de fichiers distribués : GFS, HDFS, Ceph. Les bases de données : Redis, Cassandra, DynamoDB, Accumulo, HBase, MongoDB, BigTable, Neo4j, ...
Calcul et restitution :Apport des outils de calculs statistiques
Langages adaptés aux statistiques, liens avec les outils BigData.
Outils de calcul et visualisation :R, SAS, Spark, Tableau, QlikView, ...
Caractéristiques et points forts des différentes solutions.

Atelier : mise en oeuvre du sharding avec une base de données MongoDB sur une infrastructure distribuée