Lefebvre Dalloz
Compétences.

Big data : les architectures et infrastructures

Tél.: 01 83 10 10 10
src-formation@lefebvre-dalloz.fr
formation.lefebvre-dalloz.fr

Code	Durée	Tarif
701985	Sur mesure / 7 heures	Nous contacter

Repas inclus (en présentiel)

PUBLIC

Chefs de projets - Architectes - Développeurs - Data Scientists - Toute personne souhaitant connaître les outils pour concevoir une architecture Big Data

PRÉ-REQUIS

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.

NIVEAU D'EXPERTISE

Fondamentaux

LES POINTS FORTS

Docaposte Institute propose plusieurs dispositifs pédagogiques adaptés aux apprenants :

Formation en présentiel

En groupe (inter-entreprises ou intra-entreprise)

En individuel (monitorat)

En journée ou en cours du soir (sur demande spécifique)

Formation en distanciel

Distanciel synchrone

Distanciel asynchrone

MOYENS PÉDAGOGIQUES

- Apports des connaissances communes.
- Mises en situation sur le thème de la formation et des cas concrets.
- Méthodologie d'apprentissage attractive, interactive et participative.
- Equilibre théorie / pratique : 60 % / 40 %.
- Supports de cours fournis au format papier et/ou numérique.
- Ressources documentaires en ligne et références mises à disposition par le formateur.
- Pour les formations en présentiel

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

Programme de la formation

Définition et contexte spécifique des projets Big Data

- Origines du Big Data
- Les données au cœur des sujets : explosion des données, connexions Big Data et IoT (Internet des objets), données structurées, données semi-structurées, données non structurées et données structurées
- Les limites des architecturales actuelles
- Les définitions des systèmes Big Data
- Principes de fonctionnement
- Différentes offres de marché

Propriété des données, environnement de traitement légal et sécurité

- Sécurité éthique et questions juridiques
- Données personnelles
- Informations confidentielles, interdictions
- Réglementation des Données Numériques par la CNIL
- Accords Nationaux

Impact des choix technologiques liés à l'infrastructure et à l'architecture Big Data

dans les locaux mis à disposition, les apprenants sont accueillis dans une salle de cours équipée d'un réseau Wi-Fi, d'un tableau blanc ou paperboard. Un ordinateur avec les logiciels appropriés est mis à disposition (le cas échéant).

SATISFACTION ET EVALUATION

- En amont de la formation
  - Recueil des besoins des apprenants afin de disposer des informations essentielles au bon déroulé de la formation (profil, niveau, attentes particulières...).
  - Auto-positionnement des apprenants afin de mesurer le niveau de départ.
- Tout au long de la formation
  - Évaluation continue des acquis avec des questions orales, des exercices, des QCM, des cas pratiques ou mises en situation...
- A la fin de la formation
  - Auto-positionnement des apprenants afin de mesurer l'acquisition des compétences.
  - Evaluation par le formateur des compétences acquises par les apprenants.
  - Questionnaire de satisfaction à chaud afin de recueillir la satisfaction des apprenants à l'issue de la formation.
  - Questionnaire de satisfaction à froid afin d'évaluer les apports ancrés de la formation et leurs mises en application au quotidien.

ACCOMPAGNEMENT FORMATION À DISTANCE

En cas de nécessité, une assistance technique et pédagogique est joignable entre 8h30 et 18h (jours ouvrés):

- par téléphone : 01 83 10 10 10
- par mail : care-formation@lefebvre-dalloz.fr

Une réponse immédiate est apportée ; si besoin, le demandeur est mis en relation avec un expert dans un délai maximum de 48h.

- Architectures décisionnelles "traditionnelles" (Datastores, Data Warehouses, Data Marts, etc.)
- Philosophie des bases NoSQL : Column Family, orienté document, clé-valeur, diagramme
- Plusieurs acteurs (MongoDB, Cassandra, etc.)
- Big Table / Big Query
- Moteur de base de données (Exadata)
- Base de données vectorielle (Sybase IQ)
- Hadoop, système entièrement autonome ?
- Impacts économiques

Mise en œuvre et élaboration d'une stratégie dédiée au Big Data

- Besoins en sujet de Big Data
- Atteindre les impartiaux cabinet au bon droit des conjoncture
- Outils du marché dédiés au Big Data
- Répondre aux attentes d'un collaborateur

Architectures distribuées

- Problématiques et objectifs
- Des conjoncture cohérentes, disponibles et tolérantes aux pannes ?
- Les architectures lourdement parallèles
- L'ouverture aux traitements complexes (datamining, intention learning, etc.)
- Paradigmes de calculs distribués
- Les bases NoSQL et le calcul distribué
- Qualité des données (Dataquality)
- Liens entre infrastructure et qualité des données
- Pas de qualité, pas d'analyse
- Les 4 V
- Bases à chaud et à froid
- Les apports d'un outil de Dataquality
- Pourquoi utiliser un ETL ?
- Illustration via Talend Data Integration
- Analyser les données en les fusionnant avec les données internes
- Le Master Data Management (MDM)

Préparation et visage du cluster Hadoop

- Principes de fonctionnement de Hadoop Distributed File System (HDFS)
- Principes de fonctionnement de MapReduce
- Design « type » du cluster

Installation d'une plateforme Hadoop

- Type de déploiement
- Installation d'Hadoop
- Installation de divers composants (Hive, Pig, HBase, Flume...)
- Différences parmi les distributions Cloudera, Hortonworks et MapR

Gestion d'un cluster Hadoop

- Gestion des nœuds du cluster Hadoop
- Les TaskTracker, JobTracker dans MapReduce
- Gestion des services via les schedulers
- Gestion des logs

Gestion des données pour HDFS

- Import de conjoncture externes (fichiers, bases de conjoncture relationnelles) enthousiasme HDFS
- Manipulation des fichiers HDFS

Configuration avancée

- Gestion des autorisations et de la sécurité
- Reprise sur échec d'un name node (MRV1)
- Haute disponibilité d'un NameNode (MRV2/YARN)

Monitoring et optimisation

- Monitoring (Ambari, Ganglia...)
- Benchmarking/profiling d'un cluster
- Les outils Apache GridMix, Vaidya
- Taille des blocs
- Autres options de tuning (maniement de la compression, visage mémoire...)

A noter

En amont et en aval de la formation, le positionnement pédagogique sera effectué à l'aide d'un questionnaire d'auto-positionnement.