Mis à à jour : 19/08/2021

Formation Big Data - Les fondamentaux de l'analyse de données Acquérir une première expérience du Big Data

21 heures
BD540

Objectifs pédagogiques

- Comprendre le rôle stratégique de la gestion des données pour l'entreprise
- Identifier ce qu'est la donnée, et en quoi consiste le fait d'assurer la qualité de données
- Synthétiser le cycle de vie de la donnée
- Assurer l'alignement des usages métiers avec le cycle de vie de la donnée
- Découvrir les bonnes pratiques en matière de contrôle de qualité des données
- Assurer la mise en oeuvre de la gouvernance de la donnée
- Disposer d'un premier aperçu des possibilités de traitement proposé par MapR et Hadoop

Public(s)

- MOA, chef de projet, urbaniste fonctionnel, responsable de domaine, analystes, développeurs, data miners ...
- Futurs data scientists, data analysts et data stewards

Pré-requis

- Si aucune connaissance technique particulière n'est nécessaire, il est toutefois recommandé d'avoir suivi le module "Big Data - Enjeux et perspectives" (BD500) pour suivre cette formation dans des conditions optimales
- Une connaissance de SQL est un plus pour suivre cette formation

Modalités pédagogiques

Alternance théorie et pratique

Moyens et supports pédagogiques

Support(s) de formation par apprenant

Modalités d'évaluation et de suivi

un vidéocast "L'écosystème Hadoop"
deux vidéos-tutos "Installation d'un environnement Hadoop de base" et "Développement d'un premier MapReduce"
Cette formation ne fait pas l'objet d'un contrôle des acquis via une certification

Formateur

Programme

LES NOUVELLES FRONTIÈRES DU BIG DATA (INTRODUCTION)

- Immersion
- L'approche des 4 Vs
- Cas d'usages du Big Data
- Technologies
- Architecture
- Master-less vs Master-Slaves
- Stockage
- Machine Learning
- Data Scientist et Big Data
- Compétences
- La vision du Gartner
- Valeur ajoutée du Big Data en entrep

LA COLLECTE DES DONNÉES BIG DATA

- Typologie des sources
- Les données non structurées
- Typologie 3V des sources
- Les données ouvertes (Open Data)
- Caractéristiques intrinsèques des sources
- Nouveau paradigme de l'ETL à l'ELT
- Du "schema On Write" au "Schema on Read"
- Le concept du Data Lake
- La vision d'Hortonworks
- Les collecteurs Apache on Hadoop
- SQOOP versus NIFI
- Apache SQOOP - Présentation
- Apache NIFI - Présentation
- Les API de réseaux sociaux

- Lab : Ingestion de données dans un cluster avec Apache NIFI

LE CALCUL MASSIVEMENT PARALLÈLE

- Genèse et étapes clés
- Hadoop : Fonctions coeurs
- HDFS - Différenciation
- HDFS - Un système distribué
- HDFS - Gestion des blocs et réplication
- Exemples de commandes de base HDFS
- MapReduce : aspects fonctionnels et techniques
- Apache PIG et Apache HIVE
- Comparatif des 3 approches
- Les limitations de MapReduce
- L'émergence de systèmes spécialisés
- Le moteur d'exécution Apache TEZ
- La rupture Apache SPARK
- SPARK point clés principaux
- SPARK vs Hadoop Performance
- L'écosystème SPARK
- IMPALA - Moteur d'exécution scalable natif SQL
- Le moteur d'exécution Apache TEZ
- Hive in Memory : LLAP
- Big Deep Learning
- La rupture Hardware à venir
- Labs : Exemples de manipulations HDFS + HIVE et Benchmark moteurs d'exécutions HIVE

LES NOUVELLES FORMES DE STOCKAGE

- Enjeux
- Le “théorème” CAP
- Nouveaux standards : ACID => BASE
- Les bases de données NoSQL
- Panorama des solutions
- Positionnement CAP des éditeurs NoSQL
- Les bases de données Clé-Valeur
- Focus Redis
- Les Bases de données Document
- Focus mongoDB
- Les bases de données colonnes
- Focus Cassandra et HBase
- Les bases de données Graphes
- Tendances 1 : Le NewSql
- Tendances 2 : OLAP distribué
- Lab : Exemple d'utilisation d'une base NoSQL (HBASE)

LE BIG DATA ANALYTICS (PARTIE I - FONDAMENTAUX)

- Analyse de cas concrets
- Définition de l'apprentissage machine
- Exemples de tâches (T) du machine learning
- Que peuvent apprendre les machines ?
- Les différentes expériences (E)
- L'apprentissage
- Approche fonctionnelle de base
- Les variables prédictives
- Les variables à prédire
- Les fonctions hypothèses
- Pléthore d'algorithmes
- Choisir un algorithme d'apprentissage machine
- Sous et sur-apprentissage
- La descente de gradient
- Optimisation batch et stochastique
- Anatomie d'un modèle d'apprentissage automatique
- La chaîne de traitement standard
- Composantes clés et Big Data
- Trois familles d'outils machine learning
- Les librairies de machine learning standards et Deep Learning
- Les librairies Scalables Big Data
- Les plates-formes de Data Science
- Lab : Exemples de traitement Machine Learning avec Notebook

LE BIG DATA ANALYTICS (PARTIE II - L'ÉCOSYSTÈME SPARK)

- Les différents modes de travail avec Spark
- Les trois systèmes de gestion de cluster
- Modes d'écriture des commandes Spark
- Les quatre API Langage de Spark
- Le machine learning avec Spark

- Spark SQL - Le moteur d'exécution SQL
- La création d'une session Spark
- Spark Dataframes
- Spark ML
- L'API pipeline
- Travail sur les variables prédictives
- La classification et la régression
- Clustering et filtrage coopératif
- Lab : Exemple d'un traitement machine learning avec Spark

TRAITEMENT EN FLUX DU BIG DATA (STREAMING)

- Architectures types de traitement de Streams Big Data
- Apache NIFI - Description, composants et interface
- Apache KAFKA - Description, terminologies, les APIs
- Articulation NIFI et KAFKA (NIFI ON KAFKA)
- Apache STORM - Description, terminologies, langage (agnostique)
- Articulation KAFKA et STORM (KAFKA ON STORM)
- Apache SPARK Streaming et Structured Streaming
- Articulation KAFKA et SPARK
- Comparatif STORM / SPARK
- Deux cas concrets
- Lab : Réalisation d'un traitement Big Data en Streaming (Big Data streaming analytics

DÉPLOIEMENT D'UN PROJET BIG DATA

- Qu'est ce que le Cloud Computing
- Cinq caractéristiques essentielles
- Trois modèles de services
- Services Cloud et utilisateurs
- Mode SaaS
- Mode PaaS
- Mode IaaS
- Modèles de déploiement
- Tendances déploiement
- Cloud Privé Virtuel (VPC)
- Focus offre de Cloud Public
- Caractéristiques communes des différentes offres de Cloud Public
- Focus Amazon AWS
- Focus Google Cloud Platform
- Focus Microsoft Azure
- Classement indicatif des acteurs
- Points de vigilance
- Lab : Visite d'une plate-forme de Cloud
-

HADOOP ÉCOSYSTÈME ET DISTRIBUTIONS

- L'écosystème Hadoop
- Apache Hadoop - Fonctions coeurs
- HDFS - Système de gestion de fichiers distribué (rappel)
- Map Reduce : système de traitement distribué (rappel)
- L'infrastructure YARN
- YARN - Gestion d'une application
- Docker on YARN
- Les projets Apache principaux et associés
- Les architectures types Hadoop
- Les distributions Hadoop
- Qu'est ce qu'une distribution Hadoop
- Les acteurs aujourd'hui
- Focus Cloudera
- Cloudera Distribution including Apache Hadoop (CDH)
- Focus Hortonworks
- Hortonworks Platforms HDP et HDF
- Nouvelle plate-forme Cloudera
- Vision Cloudera
- Cloudera Data Platform
- Cloudera Data Flow
- Lab : Visite d'une distribution Hortonworks dans le Cloud

ARCHITECTURES DE TRAITEMENT BIG DATA

A - Traitement de données par lots (BATCH) : - le batch en Big Data - schéma de fonctionnement - usages types du batch processing - l'orchestrateur Apache OOZIE - les workflows OOZIE - les coordinateurs OOZIE (Coordinators) - limitations de OOZIE => FALCON - points de vigilance

B - Traitement de données en flux (Streaming) : - principes - fonctionnement - rappel : modèles types de traitement de Flux Big Data - points de vigilance

C - Modèles d'architecture de traitements de données Big Data : - objectifs - les composantes d'une architecture Big Data - deux modèles génériques : λ et K - architecture Lambda - les 3 couches de l'architecture Lambda - architecture Lambda : schéma de fonctionnement - solutions logicielles Lambda - exemple d'architecture logicielle Lambda - architecture Lambda : les + et les - - architecture Kappa - architecture Kappa : schéma de fonctionnement - solutions logicielles Kappa - architecture Kappa : les + et les -
L'heure du choix

- L'heure du choix
- Lab : Analyse architecturale de deux cas de figure

LA GOUVERNANCE DES DONNÉES BIG DATA

- Challenges Big Data pour la gouvernance des données
- L'écosystème des outils de gouvernance Big Data
- Les 3 piliers de la gouvernance Big Data
- Mise en perspective dans une architecture Big Data
- Management de la qualité des données Big Data
- Tests de validation de données dans Hadoop
- Les acteurs face à la qualité des données Big Data
- Management des métadonnées Big Data
- Focus Apache HCatalog
- Focus Apache ATLAS
- Management de la sécurité, de la conformité et la confidentialité Big Data
- Focus Apache RANGER
- Tendances sécurisation des SI
- Points de vigilance
- Lab : Réflexion collective ou individuelle sur des opportunités de projets Big Data dans l'organisation et définition des objectifs et des premiers jalons