Big Data : fondamentaux de l’analyse de données

Lien : https://innov-maroc.com/formation/big-data-fondamentaux-de-lanalyse-de-donnees

DURÉE
5 jours (35h)

RÉFÉRENCE
BSI98

CATÉGORIE
Big Data - Analyse des Données et Datavisualisation

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Appréhender le rôle stratégique de la gestion des données
- Assurer la qualité de données
- Appréhender L’importance du cycle de vie des données, des données de référence, et de la gouvernance
- Acquérir les bonnes pratiques en matière de contrôle de qualité des données
- Assurer la mise en oeuvre de la gouvernance de la donnée

POUR QUI ?

- MOA
- Chef de projet
- Urbaniste fonctionnel
- Responsable de domaine
- Analystes
- Développeurs
- Data miners
- Futurs data scientists
- Data analysts
- Data stewards

Programme détaillé

1 / Comprendre le Big Data

- Les origines du Big Data
- Les dimensions en V du Big Data
- Cas d'usages du Big Data
- Les technologies essentielles
- Architecture Big Data
- Master-less vs Master-Slaves
- Les architectures Big Data orientées stockage
- Spécificités du Machine Learning pour le Big Data et domaines d’application
- Nouveaux métiers (Data Scientist, Data Steward...)
- Compétences nouvelles à acquérir
- La vision du Gartner
- Valeur ajoutée du Big Data en entreprise

2 / La collecte des données Big Data

- Où et comment collecter des données ?
- Les sources de données, les API, les fournisseurs, les agrégateurs...
- Les principaux outils de collecte et de traitement de l’information (ETL)
- Les particularités de la collecte des données semi-structurées et non-structurées

3 / Le calcul massivement parallèle

- Genèse et étapes clés
- Hadoop : Fonctions coeurs
- Le système de fichiers Hadoop (HDFS)
- MapReduce : aspects fonctionnels et techniques
- Apache PIG et Apache HIVE
- Les limitations de MapReduce
- Le moteur d'exécution Apache TEZ
- L’apport d’Apache Spark
- Impala
- Le moteur d’exécution Apache TEZ
- Hive in Memory : LLAP
- Big Deep Learning
- La rupture Hardware à venir

4 / Le stockage des données

- Enjeux
- Le “théorème” CAP
- CAP vs ACID
- Bases de données NoSQL
- Positionnement CAP des éditeurs NoSQL
- Modèle de données (clé, valeur)
- Vue d'ensemble deRedis
- Les Bases de données Document
- Vue d'ensemble de mongoDB
- Les bases de données colonnes
- Vue d'ensemble de Cassandra et HBase
- Bases de données graph
- Le NewSql
- OLAP distribué

5 / Analyse de données : Fondamentaux

- Analyse de cas concrets
- Définition de l’apprentissage machine
- Exemples de tâches (T) du machine learning
- 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

6 / L’écosystème SPARK

- 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

7 / Traitement en flux du Big Data (streaming)

- Architectures types de traitement de Streams Big Data
- NIFI : présentation, composants et interface
- Kafka : présentation, terminologies, les APIs
- Articulation NIFI et Kafka
- Storm : présentation, terminologies, langage (agnostique)
- Articulation Kafka et Storm
- Spark Streaming et Structured Streaming
- Articulation Kafka et Spark
- Storm vs Spark

8 / Déploiement d’un projet Big Data

- Présentation du 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
- Vue d'ensemble de Amazon AWS
- Vue d'ensemble de Google Cloud Platform
- Vue d'ensemble de Microsoft Azure
- Classement indicatif des acteurs

9 / L'écosystème Hadoop

- Présentation des principaux modules de la distribution Apache Hadoop
- Présentation et comparaison des principales distributions commerciales (Cloudera, Hortonworks…)
- L’infrastructure matérielle et logicielle nécessaire au fonctionnement de Hadoop
- Serveur local ou cloud
- Les concepts de base de l’architecture Hadoop: Data Node, Name Node, Job Tracker, Task Tracker
- Présentation de HDFS (Système de gestion des fichiers de Hadoop)
- Présentation de MapReduce (Outil de traitement de Hadoop)
- Les commandes exécutées au travers de PIG
- Présentation de HIVE pour transformer du SQL en MapReduce

10 / 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
- Vue d'ensemble d'Apache HCatalog
- Vue d'ensemble d'Apache ATLAS

- Management de la sécurité, de la conformité et la confidentialité Big Data
- Vue d'ensemble d'Apache RANGER
- Sécurisation des SI : Tendances

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

03 au 07 Août 2026	Présentiel - Casablanca
28 Sep. au 02 Oct. 2026	Distanciel
23 au 27 Nov. 2026	Distanciel

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-maroc.com
Web : https//www.innov-maroc.com