



ORGANISME DE FORMATION AUX TECHNOLOGIES ET METIERS DE L'INFORMATIQUE

Formation Cursus Data Steward

N° ACTIVITÉ : 11 92 18558 92

TÉLÉPHONE : 01 85 77 07 07

E-MAIL : inscription@hubformation.com

Résumé

Une étude réalisée à Stanford a démontré que les participants à une réunion se souviennent pour 5% des statistiques qui leur sont présentées, et pour 63% de l'histoire qui leur est racontée. Pourtant beaucoup des présentations réalisées en entreprise se contentent de mettre bout à bout des graphiques similaires. Comment alors retenir l'attention du lecteur ou de l'auditeur ? D'abord, il faut connaître la grammaire graphique, autrement dit, savoir quelle représentation graphique utiliser pour chaque type de données. Il faut ensuite savoir construire une histoire en s'appuyant sur des outils et méthodes scientifiques reconnues. Et une fois cette étape achevée, reste à illustrer l'histoire créée à l'aide de graphiques parlants et de visualisations inspirantes.

Objectifs

- | Comprendre le rôle stratégique de la gestion des données pour l'entreprise ou l'organisation
- | 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
- | Comprendre les différences entre apprentissage automatique supervisé, non supervisé et méta-apprentissage
- | Savoir transformer un gros volume de données à priori hétérogènes en informations utiles
- | Maîtriser l'utilisation d'algorithmes d'auto-apprentissage adaptés à une solution d'analyse
- | exploiter de gros volumes de données textuelles
- | appliquer ces différentes techniques aux projets Big Data
- | concevoir un modèle de documents répondant aux attentes de l'entreprise ou de l'organisation, en fonction du sujet analysé
- | maîtriser une méthode simple et efficace de restitution de données
- | Connaître la grammaire graphique et savoir sélectionner le bon graphique pour représenter la bonne donnée
- | Savoir bâtir un schéma narratif qui captive votre auditoire et renforce la crédibilité de vos analyses
- | Apprendre à maîtriser les outils de Tableau Software pour restituer les résultats

Public

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

Prérequis

| Si aucune connaissance technique particulière n'est nécessaire, il est toutefois recommandé d'avoir suivi le module bd500 "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

Référence	MET062
Durée	8 jours (56h)
Tarif	6 800 €HT

SESSIONS PROGRAMMÉES

A DISTANCE (FRA)

- le 18 mai 2026
- le 20 juillet 2026
- le 7 septembre 2026

PARIS

- le 18 mai 2026
- le 20 juillet 2026
- le 7 septembre 2026

AIX-EN-PROVENCE

- le 20 juillet 2026
- le 7 septembre 2026
- le 21 décembre 2026

BORDEAUX

- le 18 mai 2026
- le 7 septembre 2026
- le 21 décembre 2026

LILLE

- le 20 juillet 2026
- le 7 septembre 2026
- le 21 décembre 2026

LYON

- le 18 mai 2026
- le 7 septembre 2026
- le 21 décembre 2026

ROUEN

- le 18 mai 2026
- le 7 septembre 2026
- le 5 octobre 2026

Programme de la formation

PARTIE 1 - Big Data - Les fondamentaux de l'analyse de données

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 & Big Data
- | Compétences
- | La vision du Gartner
- | Valeur ajoutée du Big Data en entreprise

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 coeur
- | 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 bibliothèques de machine Learning standards et Deep Learning
- | Les bibliothèques 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 & Structured Streaming
- | Articulation KAFKA et SPARK
- | Comparatif STORM / SPARK
- | Deux cas concrets

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 cœurs
- | 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 & 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 OOOZIE - les workflows OOOZIE - les coordinateurs OOOZIE (Coordinators) - limitations de OOOZIE => 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 : ? et ? - 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
- | 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

PARTIE 2 - Les bases de l'apprentissage Machine (Machine Learning)

L'apprentissage machine (Introduction)

- | Introduction
- | Champs de compétences
- | Focus Data Science (Data Mining)
- | Focus Machine Learning
- | Focus Big Data
- | Focus Deep Learning
- | Définition de l'apprentissage machine
- | Exemples de tâches du machine Learning
- | Que peuvent apprendre les machines
- | Les différents modes d'entraînement

Les fondamentaux de l'apprentissage machine

- | Un problème d'optimisation
- | Quête de la capacité optimale du modèle
- | Relation capacité et erreurs
- | Un apport philosophique
- | Cadre statistique
- | Anatomie d'un modèle d'apprentissage machine
- | Jeux de données d'entraînement :
- | Cadre statistique
- | Les variables prédictives
- | Chaîne de traitement des variables prédictives
- | Les variables à prédire
- | Fonctions hypothèses :
- | Principe : jeux de fonctions hypothèses
- | Contexte de sélection des fonctions hypothèses
- | Caractéristiques des fonctions hypothèses
- | Modèles probabilistes Fréquentistes et Bayésiens
- | Fonctions de coûts :
- | Les estimateurs
- | Principe du maximum de vraisemblance (MLE*)
- | MAP - Maximum A Posteriori
- | Le biais d'un estimateur
- | La variance d'un estimateur
- | Le compromis biais - variance
- | Les fonctions de coûts
- | La régularisation des paramètres
- | Algorithmes d'optimisations :
- | Les grandes classes d'algorithmes d'optimisation
- | La descente de gradient (1er ordre)
- | Descente de gradient (détails)
- | Les approches de Newton (2nd ordre)
- | Optimisation batch et stochastique
- | Pour aller plus loin
- | Lab : Mise en oeuvre de l'environnement de travail machine Learning

La classification

- | Introduction : - Choisir un algorithme de classification
- | La régression logistique :
- | Du Perceptron à la régression logistique

- | Hypothèses du modèle
- | Apprentissage des poids du modèle
- | Exemple d'implémentation : scikit-learn
- | Régression logistique
- | Fiche Synthèse
- | SVM :
- | Classification à marge maximum
- | La notion de marge souple (soft margin)
- | Les machines à noyau (kernel machines)
- | L'astuce du noyau (kernel trick)
- | Les fonctions noyaux - SVM - Maths - SVM - Fiche Synthèse
- | Arbres de décision :
- | Principe de base - Fonctionnement
- | Maximisation du Gain Informationnel
- | Mesure d'impureté d'un noeud
- | Exemple d'implémentation : scikit-learn
- | Arbres de décision - Fiche Synthèse
- | K plus proches voisins (kNN) :
- | L'apprentissage à base d'exemples
- | Principe de fonctionnement
- | Avantages et désavantages
- | kNN - Fiche synthèse
- | Lab : Expérimentation des algorithmes de classification sur cas concrets

Les pratiques

- | Prétraitement :
- | Gestion des données manquantes
- | Transformateurs et estimateurs
- | Le traitement des données catégorielles
- | Le partitionnement des jeux de données
- | Mise à l'échelle des données
- | Ingénierie des variables prédictives (Feature Engineering) :
- | Sélection des variables prédictives
- | Sélection induite par régularisation L1
- | Sélection séquentielle des variables
- | Déterminer l'importance des variables
- | Réduction dimensionnelle par Compression des données
- | L'extraction de variables prédictives
- | Analyse en composante principale (ACP)
- | Analyse linéaire discriminante (ADL) - l'ACP à noyau (KPCA)
- | Réglages des hyper-paramètres et évaluation des modèles :
- | Bonnes pratiques
- | La notion de Pipeline
- | La validation croisée (cross validation)
- | Courbes d'apprentissage
- | Courbes de validation
- | La recherche par grille (grid search)
- | Validation croisée imbriquée (grid searchcv)
- | Métriques de performance
- | Lab : Expérimentation des pratiques du machine learning sur cas concrets

L'apprentissage d'ensembles (ensemble learning)

- | Introduction
- | L'approche par vote
- | Une variante : l'empilement (stacking)
- | Le bagging
- | Les forêts aléatoires
- | Le boosting
- | La variante Adaboost
- | Gradient Boosting
- | Fiches synthèses
- | Lab : L'apprentissage d'ensemble sur un cas concret

La régression

- | Régression linéaire simple
- | Régression linéaire multi-variée
- | Relations entre les variables
- | Valeurs aberrantes (RANSAC)
- | Évaluation de la performance des modèles de régression
- | La régularisation des modèles de régression linéaire
- | Régression polynomiale
- | La régression avec les forêts aléatoires
- | Synthèse
- | Lab : La régression sur un cas concret

Le clustering

- | Introduction
- | Le regroupement d'objets par similarité avec les k-moyens (k-means)
- | k-means : algorithme
- | L'inertie d'un cluster
- | Variante k-means ++
- | Le clustering flou
- | Trouver le nombre optimal de clusters avec la méthode Elbow
- | Appréhender la qualité des clusters avec la méthode des silhouettes
- | Le clustering hiérarchique
- | Le clustering par mesure de densité DBSCAN
- | Autres approches du Clustering
- | Synthèse
- | Lab : Le clustering sur un cas concret

PARTIE 3 - Analyse, Data Visualisation et introduction au Data StoryTelling pour la restitution de données

Data Visualisation ou la découverte de la grammaire graphique

- | Passer simplement des chiffres aux graphiques
- | Jouer avec les 3 dimensions
- | Les concepts essentiels de la grammaire graphique : Quels sont les principaux types de graphique existants ?
- | Les graphiques proposés par Excel et tous les autres.
- | Comment choisir le bon graphique pour représenter la bonne donnée ?
- | Couleurs et formes : comment les choisir
- | Présentation détaillée de Tableau Software : menus, fenêtres, fonctions, vocabulaire
- | Passer de l'idée d'un graphique, à sa représentation physique, puis à sa formalisation dans un outil

Data Storytelling : Introduction

- | Qu'est-ce que le storytelling : exemples concrets en vidéo
- | Le data storytelling : appliquer les techniques de la mise en récit aux données d'entreprise
- | Présentation et analyse critique des méthodes de data storytelling de Apple, et de Hans Rosling
- | Storytelling des idées
- | Storytelling des données

Construire son histoire avec Tableau Software

- | Le Pitch
- | Le scénario
- | Le schéma narratif

Les outils

- | Panorama des fonctions de storytelling des outils de BI
- | Le module Data Storytelling de Tableau Software
- | Panorama des autres outils : outils de représentation graphiques, outils de développement

Méthode pédagogique

Chaque participant travaille sur un poste informatique qui lui est dédié. Un support de cours lui est remis soit en début soit en fin de cours. La théorie est complétée par des cas pratiques ou exercices corrigés et discutés avec le formateur. Le formateur projette une présentation pour animer la formation et reste disponible pour répondre à toutes les questions.

Méthode d'évaluation

Tout au long de la formation, les exercices et mises en situation permettent de valider et contrôler les acquis du stagiaire. En fin de formation, le stagiaire complète un QCM d'auto-évaluation.

Suivre cette formation à distance

Voici les prérequis techniques pour pouvoir suivre le cours à distance :

- | Un ordinateur avec webcam, micro, haut-parleur et un navigateur (de préférence Chrome ou Firefox). Un casque n'est pas nécessaire suivant l'environnement.
- | Une connexion Internet de type ADSL ou supérieure. Attention, une connexion Internet ne permettant pas, par exemple, de recevoir la télévision par Internet, ne sera pas suffisante, cela engendrera des déconnexions intempestives du stagiaire et dérangera toute la classe.
- | Privilégier une connexion filaire plutôt que le Wifi.
- | Avoir accès au poste depuis lequel vous suivrez le cours à distance au moins 2 jours avant la formation pour effectuer les tests de connexion préalables.
- | Votre numéro de téléphone portable (pour l'envoi du mot de passe d'accès aux supports de cours et pour une messagerie instantanée autre que celle intégrée à la classe virtuelle).
- | Selon la formation, une configuration spécifique de votre machine peut être attendue, merci de nous contacter.
- | Pour les formations incluant le passage d'une certification la dernière journée, un voucher vous est fourni pour passer l'examen en ligne.
- | Pour les formations logiciel (Adobe, Microsoft Office...), il est nécessaire d'avoir le logiciel installé sur votre machine, nous ne fournissons pas de licence ou de version test.
- | Horaires identiques au présentiel.