

Cursus Data Analyst

Cursus Métier

 Présentiel ou en classe à distance

 12 jours (84 h)

Prix inter : 8.200,00 € HT

Réf.: CM061

De nombreuses entreprises et organisations s'interrogent pour exploiter au mieux de leurs intérêts les immenses quantités de données dont elles disposent. Mais comment faire pour les faire "parler" ? En fonction du secteur d'activité de l'entreprise ou de l'organisation et des axes stratégiques fixés par la direction, le Data Analyst doit collecter et fouiller toutes les données à sa disposition pour finalement fournir des jeux de données de qualité qui permettront à l'aide d'outils spécifiques de Data Mining et d'apprentissage automatique (Machine Learning) de déterminer des modèles, schémas ou tendances qui au final aideront l'entité à améliorer ses performances. A l'issue de ce cursus de 12 jours, les participants auront acquis les connaissances et compétences techniques propres au métier du Data Analyst.

A qui s'adresse cette formation ?



Pour qui

- Toute personne souhaitant évoluer vers une fonction de Data Analyst



Prérequis

- La connaissance du langage SQL est nécessaire pour suivre ce cursus. Il est conseillé d'avoir suivi la formation "["Interroger des bases de données avec le langage SQL" \(LA300\)](#)" ou de disposer des connaissances équivalentes
- ["Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !"](#)

Programme

Avant la session

- Un quiz de consolidation des prérequis

En présentiel / A distance

1 - Big Data - Les fondamentaux de l'analyse de données (3j)

- Objectif : Disposer des connaissances et compétences nécessaires pour identifier et collecter des données et s'assurer de leur qualité et de leur alignement sur les besoins et usages métiers de l'entreprise ou organisation
- 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
- La collecte des données : typologie des sources, les données non structurées, typologie 3V des sources, les données ouvertes (Open Data), nouveau paradigme de l'ETL à l'ELT, le concept du Data Lake, les API de réseaux sociaux, ...
- Le calcul massivement parallèle : genèse et étapes clés, Hadoop, HDFS, MapReduce, Apache PIG et Apache HIVE, comparatif des 3 approches, limitations de MapReduce, moteur d'exécution Apache TEZ, la rupture Apache SPARK, Hive in Memory (LLAP), Big Deep Learning, ...

- Les nouvelles formes de stockage : enjeux, le "théorème" CAP, nouveaux standards : ACID => BASE, panorama des bases de données NoSQL, bases de données Clé-Valeur, bases de données Document, bases de données colonnes, bases de données Graphes, ...
- Le Big Data Analytics (fondamentaux) : analyse de cas concrets, que peuvent apprendre les machines ?, les différentes expériences (E), l'apprentissage, choisir un algorithme d'apprentissage machine, anatomie d'un modèle d'apprentissage automatique, les librairies de machine learning standards et Deep Learning, les plates-formes de Data Science
- Le Big Data Analytics (écosystème SPARK) : les différents modes de travail, les 3 systèmes de gestion de cluster, modes d'écriture des commandes Spark, machine learning avec Spark , travail sur les variables prédictives, la classification et la régression
- Traitement en flux du Big Data (streaming) : architectures types de traitement de Streams Big Data, Apache NIFI, Apache KAFKA, articulation NIFI et KAFKA, Apache STORM, articulation KAFKA et STORM, comparatif STORM/SPARK
- Déploiement d'un projet Big Data : Cloud Computing , 5 caractéristiques essentielles, 3 modèles de services, modes (SaaS , PaaS, IaaS), Cloud Privé virtuel (VPC), focus AWS, GCP et Azure
- Hadoop écosystème et distributions : écosystème, fonctions coeurs, HDFS, MapReduce, infrastructure YARN, distributions Hadoop, focus Cloudera, Focus Hortonworks,...
- Architecture de traitement Big Data : traitement de données par lots, traitement de données en flux, modèles d'architecture de traitement de données Big Data, l'heure du choix
- La gouvernance des données Big Data : outils de gouvernance Big Data, les 3 piliers, le management de la qualité des données, le management des métadonnées Big Data, management de la sécurité, de la conformité et la confidentialité

2 - Les bases de l'apprentissage Machine (Machine Learning) (3j)

- Objectif : Maîtriser le Data Mining et le Machine Learning pour explorer de très importants volumes de données et construire des modèles répondant aux problèmes très variés des entreprises du secteur privé et public
- L'apprentissage machine : champs de compétences, focus : Data Science, Machine Learning, Big Data, Deep Learning, définition de l'apprentissage machine, les différents modes d'entraînement
- Les fondamentaux de l'apprentissage machine : préambule, jeux de données d'entraînement, fonctions hypothèses, fonctions de coûts, algorithmes d'optimisations
- La classification : introduction, la régression logistique, SVM, arbres de décision, K plus proches voisins (kNN), synthèse
- Les pratiques : prétraitement, ingénierie des variables prédictives (feature engineering), réglages des hyper-paramètres et évaluation des modèles, synthèse
- L'apprentissage d'ensembles (ensemble learning) : introduction, l'approche par vote, une variante, le bagging, les forêts aléatoires, le boosting, la variante Adaboost, gradient boosting, fiches synthèses
- La régression : régression linéaire simple et multi-variee, relations entre les variables, valeurs aberrantes, évaluation de la performance des modèles de régression, régularisation des modèles de régression linéaire, régression polynomiale, régression avec les forêts aléatoires, synthèse
- Le clustering : introduction, regroupement d'objets par similarité avec les k-moyens, k-means, l'inertie d'un cluster, variante k-means++, clustering flou, clustering hiérarchique, clustering par mesure de densité DBSCAN, autres approches du clustering, synthèse

3 - Big Data : mise en oeuvre pratique d'une solution complète d'analyse des données (4j)

- Objectif : Savoir mettre en oeuvre une solution complète de Big Data en environnement Hadoop et disposer des compétences nécessaires au traitement et à l'analyse des données
- Introduction : objectifs, schématisation du projet, écosystème et stack technologique, résultats attendus
- Ingestion de données massives : description, caractéristiques clés des outils d'ingestion, focus Apache NIFI et KAFKA, ingestion de données en streaming NIFI sur KAFKA, réalisation d'un workflow NIFI d'ingestion de donnée streaming dans HDFS
- Traitement de données Big Data en batch : diagramme de fonctionnement, solutions logicielles associées, Big Data Batch scripting, Data Warehousing Big Data, Big Data analytics
- Traitement avancé Big Data : l'apprentissage machine, l'écosystème Spark, création d'un modèle de ML, d'un modèle de clusterisation de données, d'un modèle d'analyse prédictive supervisé, application d'un modèle ML
- Stockage de données distribuées : principes des bases de données distribuées, solutions (NoSQL, NewSQL), création, ingestion de données et interrogation d'une base de données distribuées
- Automatisation de chaîne de traitement Batch : l'orchestre Oozie, ordonnancement de scripts HIVE, combinaison avec des scripts SPARK
- Traitement de données massives en flux (streaming) : principe de fonctionnement, solutions logicielles, l'inscription de streams à un Hub Streaming, le traitement avancé de données en flux (machine learning)
- Mise en oeuvre dans une architecture Big Data : approches standards, réalisation d'une solution complète de traitement de données de type Lamda ou Kappa

4 - Big Data - Analyse, Data Visualization et introduction au Data StoryTelling pour la restitution de données (2j)

- Objectif : Être en mesure de concevoir des modèles de documents adaptés aux besoins métiers de l'entreprise ou organisation et savoir mettre en oeuvre différentes techniques de visualisation graphique, de mise en récit et de présentation permettant de valoriser les données
- Data Visualisation ou la découverte de la grammaire graphique : des chiffres aux graphiques, les 3 dimensions, présentation de Tableau Software, de l'idée d'un graphique à sa formalisation dans un outil
- Data Storytelling : présentation, exemples, techniques de la mise en récit des données, Storytelling des idées et des données
- Comment construire son histoire : Pitch, scénario, schéma narratif
- Les outils : fonctions de Storytelling des outils de BI, le module Data Storytelling de Tableau Software, autres outils

Après la session

- Un vidéocast "L'écosystème Hadoop"
- Deux vidéos-tutos "Installation d'un environnement Hadoop de base" et "Développement d'un premier MapReduce"



Les objectifs de la formation

- Disposer d'une vision claire du Big Data, de ses enjeux, de son écosystème et des principales technologies et solutions qui y sont associées
- Maîtriser le cycle de vie de la donnée et savoir garantir la qualité des données
- Être en mesure d'aligner les usages métiers avec le cycle de vie de la donnée
- Savoir transformer de gros volumes de données hétérogènes en informations utiles
- Comprendre comment utiliser des algorithmes d'auto-apprentissage adaptés à une solution d'analyse
- Disposer des compétences techniques nécessaires à la mise en oeuvre d'analyses Big Data
- Maîtriser la boite à outils technologique que constitue Hadoop et son écosystème et savoir comment utiliser chaque brique (PIG, HIVE, MapR...)
- Savoir concevoir des modèles de documents et des graphiques répondant aux attentes de l'organisation, en fonction du sujet analysé



Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.



Les points forts de la formation

- Chaque participant établit son propre planning de formation. En fonction de la date de début choisie parmi celles proposées ci-dessous, nos Conseillers Formation proposent différentes dates pour chacun des modules du cursus. Pour des raisons d'efficacité pédagogique, il est fortement recommandé de suivre les modules dans l'ordre présenté sur ce programme.
- L'alternance de formations et de périodes de mise en pratique en entreprise ou organisation favorise l'acquisition rapide et durable de nouveaux savoirs.
- Animé par un expert spécialiste du sujet traité, chacun des 4 modules aborde un aspect spécifique de la thématique de formation.
- A travers de nombreuses mises en situation, les participants mettront en pratique les aspects théoriques abordés au cours des différentes étapes du cursus.



Dates et villes 2026 - Référence CM061



Dernières places disponibles



Session garantie

Lyon

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 7 sept. au 5 mars
du 21 déc. au 18 juin

Rennes

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 5 oct. au 2 avr.
du 21 déc. au 18 juin

Paris

du 5 janv. au 3 juil.
du 16 mars au 11 sept.
du 18 mai au 13 nov.

du 20 juil. au 15 janv.
du 7 sept. au 5 mars
du 5 oct. au 2 avr.

du 21 déc. au 18 juin

Strasbourg

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 7 sept. au 5 mars
du 5 oct. au 2 avr.

A distance

du 5 janv. au 3 juil.
du 16 mars au 11 sept.
du 18 mai au 13 nov.

du 20 juil. au 15 janv.
du 7 sept. au 5 mars
du 5 oct. au 2 avr.

du 21 déc. au 18 juin

Nantes

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 5 oct. au 2 avr.
du 21 déc. au 18 juin

Rouen

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 7 sept. au 5 mars
du 5 oct. au 2 avr.

Toulouse

du 5 janv. au 3 juil.
du 16 mars au 11 sept.

du 20 juil. au 15 janv.
du 5 oct. au 2 avr.

Bordeaux

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 7 sept. au 5 mars
du 21 déc. au 18 juin

Sophia Antipolis

du 5 janv. au 3 juil.
du 18 mai au 13 nov.

du 7 sept. au 5 mars
du 5 oct. au 2 avr.

Marseille

du 16 mars au 11 sept.
du 20 juil. au 15 janv.

du 7 sept. au 5 mars
du 21 déc. au 18 juin

Lille

du 16 mars au 11 sept.

du 20 juil. au 15 janv.

du 7 sept. au 5 mars

du 21 déc. au 18 juin

Aix-en-Provence

du 16 mars au 11 sept.

du 20 juil. au 15 janv.

du 7 sept. au 5 mars

du 21 déc. au 18 juin