

Big Data – Mise en oeuvre pratique d'une solution complète d'analyse des données

INFORMATIONS GÉNÉRALES

Type de formation : Formation continue

Éligible au CPF : Non

Domaine : IA, Big Data et Bases de données

Action collective : Non

Filière : Big Data

Rubrique : Fondamentaux

Code de formation : BD006

€ Tarifs

Prix public : 2690 €

Tarif & financement :

Nous vous accompagnons pour trouver la meilleure solution de financement parmi les suivantes :

- **Le plan de développement des compétences** de votre entreprise : rapprochez-vous de votre service RH.
- **Le dispositif FNE-Formation.**
- **L'OPCO** (opérateurs de compétences) de votre entreprise.
- **Pôle Emploi** sous réserve de l'acceptation de votre dossier par votre conseiller Pôle Emploi.
- **CPF -MonCompteFormation**

[Contactez nous](#) pour plus d'information

PRÉSENTATION

Objectifs & compétences

- Disposer des compétences techniques nécessaires à la mise en oeuvre d'analyses Big Data
- Comprendre le cadre juridique du stockage et de l'analyse de données
- Savoir utiliser des outils de collecte opensource
- Être en mesure de choisir la bonne solution de stockage de données au regard des spécificités d'un projet (OLAP, NoSQL, graph)

Explorer la boîte à outils technologique que constitue Hadoop et son écosystème et savoir comment utiliser chaque brique (MapReduce, HIVE, SPARK,...)

Public visé

Chefs de projet Data Scientists, Data Analysts Développeurs Analystes et statisticien Toute personne en charge de la mise en oeuvre opérationnelle d'un projet Big Data en environnement Hadoop

Pré-requis

Il est recommandé d'avoir suivi le module «Big Data - Les fondamentaux de l'analyse des données» (BD007) pour suivre cette formation dans des conditions optimales Être familier des environnement techniques décisionnels traditionnels et connaître les principes de base d'algorithme est vivement recommandé Disposer d'une première approche pratique d'Hadoop est un plus pour suivre cette formation

📍 Lieux & Horaires

Campus : Ensemble des sites

Durée : 28 heures

Délai d'accès :

Jusqu'à 8 jours avant le début de la formation

Distanciel possible : Oui

PROGRAMME

LA COLLECTE DE DONNÉES

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)

Prise en main de Talend ETL et de Talend Data

Préparation (outils libres)

Les particularités de la collecte des données semi-structurées et non-structurées

LE STOCKAGE DES DONNÉES

Les différentes formes de stockage des données : rappel de l'architecture relationnelle de stockage des données transactionnelles (SGBD/R) et multidimensionnelles (OLAP)

Les nouvelles formes de stockage des données

- compréhension, positionnement et comparaison : Bases orientées clé-valeur, documents, colonnes, graphes Panorama des bases de données NoSQL

Prise en main d'une base de données orientée colonne (Hbase)

Particularités liées au stockage des données non-structurées

Comment transformer des données non structurées en données structurées

L'ÉCOSYSTÈME HADOOP

Présentation des principaux modules de la distribution

📅 Prochaines sessions

Cliquez sur la date choisie pour vous inscrire :

■ 28 / 10 / 2024

📍 : Ensemble des sites

✓ : Distanciel possible

🕒 : 28 heures

📅 : 4 jours

Apache Hadoop Présentation et comparaison des principales distributions commerciales (Cloudera, Hortonworks...) L'infrastructure matérielle et logicielle nécessaire au fonctionnement d'une distribution Hadoop en local ou dans le 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)

Prise en main et exercices pratiques dans HDFS

Présentation de MapReduce (Outil de traitement de Hadoop)

Les commandes exécutées au travers de PIG Utilisation de HIVE pour transformer du SQL en MapReduce

L'ANALYSE DE DONNÉES

Requêter les données

Analyser et comprendre la signification des données extraites

Particularités liées à l'analyse des données non structurées

Analyse statistique : notions de base

Analyse prédictive : comment transformer des données du passé en prévisions pour le futur

Calculer des tendances

Développer des programmes simples d'automatisation des analyses (en Python)

Machine Learning : les bases de l'apprentissage machine avec Spark Deep

Learning : notions de base de l'analyse future automatisée de données non structurées

MISE EN OEUVRE DE PROJETS BIG DATA

Automatisation de tâches avec Oozie

Mise en production de programmes de Machine Learning

L'utilisation des notebooks comme livrables

Traitement du temps réel

Gouvernance de données Big Data

MODALITÉS

Modalités

Modalités : en présentiel, distanciel ou mixte – Horaires de 9H à 12H30 et de 14H à 17H30 soit 7H – Intra et Inter entreprise

Pédagogie : essentiellement participative et ludique, centrée sur l'expérience, l'immersion et la mise en pratique. Alternance d'apports théoriques et d'outils pratiques.

Ressources techniques et pédagogiques : Support de formation au format PDF ou PPT Ordinateur, vidéoprojecteur, Tableau blanc, Visioconférence : Cisco Webex / Teams / Zoom

Pendant la formation : mises en situation, autodiagnostic, travail individuel ou en sous-groupe sur des cas réels

Méthode

Fin de formation : entretien individuel

Satisfaction des participants : questionnaire de satisfaction réalisé en fin de formation

Assiduité : certificat de réalisation (validation des acquis)