

Plan de cours N° : 1053

Durée : 2 jours (14h)

BIG DATA

La synthèse

PARTICIPANTS / PRE-REQUIS

Développeur, chef de projets proche du développement, ingénieur scientifique sachant lire du code.

Avoir une appétence pour les technologies informatiques.

OBJECTIFS PEDAGOGIQUES

Appréhender les concepts fondamentaux et les origines du Big Data. Découvrir les approches structurées, semi-structurées et non structurées, ainsi que les solutions Cloud. Apprendre à utiliser HDFS, MapReduce, Yarn et Spark. Comprendre leur rôle et leur utilisation dans la BI. Travailler avec MongoDB, Cassandra et Hive. Utiliser Pandas, NumPy, SciPy et Matplotlib pour le traitement et la visualisation des données. Implémenter des modèles de machine learning et d'intelligence artificielle avec Python et SKLearn. Utiliser MongoDB et Hadoop HDFS pour gérer les données IoT.

MOYENS PEDAGOGIQUES

Tour de table au début de chaque formation pour définir les objectifs de chaque participant,

Alternance entre apports théoriques (en moyenne 30%) et exercices pratiques (en moyenne 70%),

Utilisation de cas concrets issus de l'expérience professionnelle de nos formateurs,

Remise d'un support de cours,

Assistance post-formation d'une durée de 1 an sur le contenu de la formation via notre adresse mail dédiée formateurs@atp-formation.com

MOYENS PERMETTANT LE SUIVI DE L'EXECUTION ET DES RESULTATS

Positionnement préalable oral ou écrit,

Evaluation des acquis tout au long de la formation par des exercices de synthèse,

Attestation de stage remise à chaque apprenant, avec son niveau d'acquisition pour chaque objectif pédagogique,

Feuille de présence signée par demi-journée,

Questionnaire de satisfaction pour évaluer la qualité de l'enseignement,

En option: passage certification possible selon les thématiques.

MOYENS TECHNIQUES EN PRESENTIEL

Accueil des stagiaires dans une salle dédiée à la formation, équipée d'ordinateurs récents et performants, d'un vidéo projecteur et d'un tableau blanc.

MOYENS TECHNIQUES DES CLASSES A DISTANCE

Grâce à un logiciel comme Teams, suivez une formation en temps réel et entièrement à distance. Lors de la classe en ligne, les apprenants interagissent et communiquent entre eux et avec le formateur.

Nous vous conseillons très fortement l'utilisation de votre webcam et de disposer d'un double écran.

Pour toute question avant et pendant le parcours, une assistance technique et pédagogique est à disposition par téléphone au 04.76.41.14.20.

ORGANISATION

Les cours ont lieu de 9h00-12h30 13h30-17h00 (adaptable à la demande).

PROFIL FORMATEUR

Nous recrutons méticuleusement nos formateurs selon 3 critères: expertise, pédagogie et agilité.

ACCESSIBILITE

Les personnes atteintes de handicap souhaitant suivre nos formations sont invitées à nous contacter directement, afin d'étudier ensemble les possibilités d'organisation.

MISE A JOUR

15/07/2024

Siège social :
31 avenue du Granier
38240 MEYLAN

Agences :
170 rue de Chatagnon
38430 Moirans

Le Thélème
1501/1503 route des Dolines
06560 Valbonne

Plan de cours N° : 1053

Durée : 2 jours (14h)

BIG DATA

La synthèse

Introduction aux Data Sciences

- Qu'est-ce que la Data Science?
- Qu'est-ce qu'une base de données SQL?
- L'IoT
- Les origines du Big Data
- Les chiffres clés, le ROI
- Les langages Java et Python
- Qu'est-ce que l'informatique distribué ?

Les technologies de Big Data

- Les approches non structurées
 - Apache Hadoop HDFS
- Les approches semi structurées
 - MongoDb, Cassandra
- La répartition des bases et le modèle de distribution
- Les approches structurées : SQL
- La perte de l'ACIDité des bases
- Le modèle 3V : Volume, Vélocité, Variété
- Les différences entre les approches
- Les approches Cloud
 - Google, Amazon Web Service, Microsoft Azure

Apache Hadoop

- Le système HDFS
- L'approche semi-structurée avec Hive
- Apache Hadoop
- Le traitement des données réparties
 - Map – Reduce
- Le scheduler Yarn
- Spark : le futur
- Les distributions
 - Hortonworks, Cloudera, MapR
- L'approche 3V
- Création d'un datanode
- Création d'un namenode distribué
- Python HDFS

Introduction aux DataLake et DataLab

- Qu'est-ce qu'un DataLake ?
- Qu'est-ce qu'un DataLab ?
- Les DataLake pour la BI
 - Les DataMart et DataWarehouse
- Utiliser Hadoop comme DataLake

Les bases semi-structurées

- Le stockage NoSql
- Accès à MongoDB avec Python
- L'approche tabulaire avec Cassandra
- Les bases de Cassandra
- Accès à Cassandra avec CQL
- L'approche SQL avec Hive et HDFS
- Accès à Cassandra en Python
- L'approche semi-structurée avec JSON
- MongoDB
- L'IoT avec MongoDB
- Utiliser MongoDB avec Hadoop HDFS

Traitement et affichage des données

- Traitement des données avec Pandas
- Calculs sur les données avec Numpy et SciPy
- Le Machine Learning
- Les traitements avec Python, R et Scala
- La segmentation, la randomisation, le regroupement avec les modèles
- Implémenter de l'intelligence artificielle avec SKLearn
- Le Deep Learning avec les réseaux neuronaux
- Utilisation de la bibliothèque graphes MathPlotLib
- Affichage des données avec QlikView