BIG DATA ET DATAVISUALISATION

3 jours en présentiel (21 heures)

Objectifs pédagogiques

Cette formation Le développement d'applications Big Data et la Data Visualisation vous permettra de : Définir et identifier le contexte spécifique des projets Big Data, connaitre le panorama technologique et enjeux socio-économiques, mesurer l'impact des choix technologiques en matière de développement Big Data, appréhender l'environnement : Hadoop (distribution HortonWorks), maîtriser les techniques de développement : MapReduce, mettre en œuvre les langages de programmation : Python, connaitre Le Deep Machine Learning, sélectionner le mode pertinent de Data Visualisation, consolider ses connaissances à travers un cas d'usage.

Population visée

Cette formation Le développement d'applications Big Data et la Data Visualisation est destinée aux développeurs Big Data.

Pré-requis

Cette formation Le développement d'applications Big Data et la Data Visualisation nécessite une expérience dans le développement, si possible avec Java. Une compréhension des algorithmes est un plus.

Méthodes pédagogiques

1 poste et 1 support par stagiaire 8 à 10 stagiaires par salle Remise d'une documentation pédagogique papier ou numérique pendant le stage La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience

Formateur

Formateur consultant expert en BIG DATA

Modalités de validation des acquis

Auto-évaluation des acquis par le stagiaire via un questionnaire en ligne Attestation de fin de stage remise au stagiaire

Contenu

PANORAMA TECHNOLOGIQUE ET ENJEUX SOCIO-ECONOMIQUES

- Bâtir une vision Data Centric pour l'entreprise
- Etudier l'environnement concurrentiel de l'entreprise
- Comment créer de la valeur ou apporter de la valeur complémentaire aux données
- Comment utiliser les Big Data qui doivent être un levier technologique pour accompagner les enjeux métiers et non l'inverse
- Comprendre les acteurs du Big Data et leur positionnement
- Quelle vision à 3 ans
- Propriété de la donnée, environnement juridique du traitement, sécurité
- La nécessité de la gouvernance des données
- Qu'est-ce qu'un CDO ?

ASPECTS JURIDIQUES ET ETHIQUES : QUELLES DONNEES POUR QUELS USAGES ?

- Données objectives
- Données à caractère personnel
- Quelle gestion des données personnelles ? (donnée se rapportant à une personne physique, qui peut être identifiée quel que soit le moyen utilisé)
- Quels Impact sur la vie privée
- Surveillance et sanction de la CNIL
- Déclaration préalable
- Exemples
- Présentation du socle (la finalité du traitement) et de 4 conditions
- Finalité explicite et légitime
- Loyauté dans la mise en œuvre du traitement
- Données pertinentes
- Durée de conservation non excessive
- Sécurité

IMPACT DES CHOIX TECHNOLOGIQUES EN MATIERE DE DEVELOPPEMENT BIG DATA

- Les nouveaux frameworks Big Data
- Prendre en compte l'architecture de donnée distribuée
- Prendre en compte les traitements distribués
- L'importance de Java au sein des architectures Hadoop
- Le management des données

L'ENVIRONNEMENT : APACHE HADOOP

- Découvrir Hortonworks la distribution 100% Apache Hadoop
- Hortonworks et l'ODPi (Open Data Platform)
- Fondamentaux d'Hadoop
- L'intérêt d'Hadoop
- Vue globale d'Hadoop
- HDFS
- MapReduce
- YARN
- L'écosystème Hadoop

LE DEVELOPPEMENT : MAPREDUCE

- Introduction à PIG
- Fondamentaux de PIG
- Pourquoi utiliser Hive ?
- Comparer PIG aux ETL traditionnelles
- Cas d'utilisation de PIG
- Introduction à Hive
- Introduction à Impala et Hive
- Pourquoi utiliser Impala et Hive ?
- Comparer Hive aux Bases de données traditionnelles

BIG DATA ET DATAVISUALISATION

- Cas d'utilisation de Hive
- Modélisation et gestion des données avec Impala et Hive
- Aperçu sur le stockage de données
- Création de bases de données et de tableaux
- Remplir les données dans les tableaux
- HCatalog
- Mettre en mémoire-cache les Métadonnées Impala
- Les formats de données
- Sélectionner un format de fichier
- Support d'outils Hadoop pour les formats de fichier
- Schémas Avro
- Utiliser Avro avec Hive et Sqoop
- Evolution du Schéma Avro
- Compression
- Capturer les données avec Apache Flume
- Qu'est-ce qu'Apache Flume ?
- Architecture basique de Flume
- Les sources de Flume
- Flume Sinks
- Les réseaux de Flume
- La configuration de Flume
- Les bases de Spark
- Qu'est-ce qu'Apache Spark ?
- Utiliser « Spark Shell »
- RDDs (Resilient Distributed Datasets)
- La programmation fonctionnelle dans Spark
- Travailler avec des « RDD » dans Spark
- Ecrire et déployer des applications Spark
- La programmation parallèle avec Spark
- Aperçu de Shark (Spark SQL)

LANGAGES DE PROGRAMMATION : PYTHON, R, …

- Python
- Syntaxe basique
- Structures procédurales
- Bibliothèques essentielles
- La programmation orientée objet
- Le langage R
- Variables et types de bases (numeric, character, list, …)
- Tests
- Boucles
- Fonctions
- Fusion de données
- Traitement des valeurs manquantes
- Représentations graphiques des données
- Pie charts et graphiques à double échelle

LE DEEP MACHINE LEARNING

- Approche fréquentiste
- Apprentissage statistique
- Conditionnement des données et réduction de dimension
- Machines à vecteurs supports (SVM) et méthodes à noyaux
- Quantification Vectorielle
- Réseaux de neurones et deep learning
- Ensemble learning et arbres de décision
- Bandits

LA DATA VISUALISATION

- Connaître les modes de représentation des données
- Déterminer le graphe le plus pertinent selon le message à délivrer
- Concevoir et expérimenter des concepts
- Justifier ses analyses et choix graphiques
- Savoir sélectionner les outils de datavisualisation à positionner sur les plateformes Big Data

ETUDES DE CAS

- Mise en place d'une architecture Big Data orientée Data Lake chez Hermès et mise en place d'une solution de Datavizualisation pour la gestion de la console de Data Stewardship.