Formation
'Big Data et Machine Learning' - 2J

Derrière le terme Big Data se cache de nouvelles techniques permettant de stocker et traiter un volume et une variabilité des données qui étaient inaccessibles jusqu'alors. Quant à elles, les techniques de Machine Learning permettent de mettre au point des algorithmes qui apprennent à partir des données ou qui découvrent des corrélations insoupçonnées. Mais surtout, ces nouvelles techniques permettent de nouveaux services et usages.

Public concerné

Cette formation s'adresse à toute personne désireuse de comprendre ce qui se cache derrière le Hype Big Data et qui souhaite avoir une vue d'ensemble des cas d'utilisation métier et des technologies sous-jacentes.

Pré requis

Quelques bases mathématiques (calcul matriciel, probabilités basiques, dérivées partielles), quelques bases en algorithmique.

Durée – 2 jours

Coût – 1200 Euros HT par participant.

Pour une formation intra-entreprise : nous consulter.

Exemples et exercices

Le formateur étaye le cours en présentant des exemples sous Hadoop (HDFS, MapReduce, Pig, Ambari), Spark, MongoDB, MathLab, Orange Data Mining.
Les exercices réalisés par les stagiaires ne nécessitent pas de manipulation sur machines.

Support de formation

Le support de formation est en anglais. La reprographie du support de formation est assurée par inspearit.

Lieu

Dans nos locaux ou dans les vôtres en cas de formation intra-entreprise.

Organisation

Chaque session est limitée à 12 participants maximum.
N° d'agrément formation : 11755207775.

Formateurs

La formation est dispensée en français, par des consultants de inspearit.

Contact

inspearit
21 rue de la Banque
75002 Paris
Téléphone : +33 1 80 06 84 33
Fax : +33 1 80 06 84 34
E-mail : info.fr@inspearit.com
www.inspearit.fr

Programme

- **Qu'est-ce que le Big Data et le Big Data Analytics?**
  Suite au constat de l'évolution exponentielle du volume de données, de la multiplication des types et des sources de données, la formation présente ce qui se cache derrière le 'hype' Big Data : de la définition très générale s'appuyant sur les 4V, en passant par la présentation du nouveau rôle de Data Scientist jusqu'à la déclinaison opérationnelle en domaines technologiques.

- **Les cas d'utilisation métier**
  Présentation de cas d'utilisation des Big Data Analytics dans des domaines aussi divers que le marketing, la sécurité, la santé, la finance, l'assurance, l'aide à la prise de décision, l'optimisation des processus métier (industriel, agricole, services), l'optimisation des machines, la science, la politique…

- **L'architecture et les services Hadoop**
  Présentation de l'architecture du système Hadoop et des services associés (HDFS, Yarn, MapReduce, Pig, Spark, Hive, Hbase, Storm, Soir, Mahout, Kafka, Sqoop, Ambari, Oozie, Zookeeper…)

- **HDFS (Hadoop Distributed File System)**
  Présentation des caractéristiques et de l'architecture d'HDFS et des commandes courantes.

- **MapReduce**
  Présentation des concepts MapReduce qui seront illustrés par un exemple sous Python Mincemeat, puis sous Hadoop en Java et enfin sous Elastic MapReduce.
  Un exercice de multiplication de matrices par blocs en MapReduce permettra d'ancrer les concepts.

- **Pig Latin**
  Présentation du langage de haut niveau Pig Latin qui permet de s'affranchir d'une partie de la complexité de MapReduce.

- **Bases de données NoSQL**
  Présentation des différences entre les bases de données classiques et les bases NoSQL, énumération des bases NoSQL existantes, théorème de BGL.
  Présentation détaillée de la base orientée document Mongo DB

- **Introduction au Machine Learning**
  Le formateur présente de manière conceptuelle l'apprentissage supervisé et les processus associés, l'apprentissage non supervisé, la classification, la régression et le partitionnement de données (clustering) ainsi qu'un guide d'utilisation des différentes techniques en fonction du contexte.

  Deux exemples mis en œuvre par le formateur viendront illustrer les concepts :
  - une analyse de sentiment sur twitter en utilisant un classifieur bayésien naïf en python
  - une analyse génétique en utilisant l'outil Orange data mining avec un classifieur bayésien naïf et en analysant la performance avec une courbe de ROC.
  Suivra un exercice d'identification des 'features' sur un problème de télématique.

  Nous étudierons ensuite en détail les différentes techniques de Machine Learning et réaliserons des exercices correspondants :
  - La régression linéaire en utilisant l'algorithme du gradient et l'algorithme du gradient stochastique. Nous étudierons également comment identifier et régler les problèmes de biais et de variance : analyse de la courbe d'apprentissage, mécanisme de régularisation,
  - La régression logistique,
  - Les réseaux neuronaux,
  - Le Support Vector Machine (SVM),
  - Le partitionnement en k-moyennes (k-means),
  - L'analyse en composantes principales (PCA),
  - Les algorithmes de détection d'anomalies,
  - Les systèmes de recommandation.

- **Le Big Data analytics / Spark**
  Introduction au langage Scala et à Spark qui est une alternative au MapReduce d'Hadoop alliant performance et richesse de l'écosystème.
  Un exemple sous Spark et mllib viendra illustrer les concepts.

- **DataViz**
  Nous finirons cette formation par une introduction au DataViz.