PROGRAMME DE FORMATION SUR L'ANALYSE DES DONNEES BIG DATA

Description de la formation

Cette formation introduit les concepts clés du Big Data et couvre les outils essentiels pour collecter, traiter et analyser des volumes massifs de données. Les participants apprendront à utiliser des technologies telles que Hadoop, Spark et NoSQL pour gérer des données complexes. La formation inclut des méthodes de traitement en batch et en temps réel, ainsi que des techniques de visualisation des résultats. Des études de cas concrets permettront d'appliquer les compétences acquises à des problématiques réelles. Elle s'adresse à ceux qui souhaitent maîtriser l'analyse des données massives dans différents secteurs.

Module 1 : Introduction au Big Data

- Définition et caractéristiques du Big Data (les 5 V : Volume, Vélocité, Variété, Véracité, Valeur)
- Enjeux économiques, technologiques et éthiques
- Écosystème Big Data : bases de données, outils, métiers
- Architecture d'un système Big Data (data lake, data warehouse, etc.)

Module 2 : Collecte et stockage des données massives

- Méthode de collecte des données
  - Web scraping : extraction automatisée de données sur des sites web via des outils
    comme **Octoparse, Webscraper.io.**
  - Formulaires en ligne : collecte directe via inscriptions, sondages, questionnaires.
  - Cookies et trackers : suivi des comportements en ligne.
  - Collecte en temps réel via des plateformes infonuagiques (cloud), avec des outils
  - comme **Amazon Kinesis, Apache Kafka.**
  - Utilisation de proxies pour contourner les blocages lors de la collecte automatique

- **Méthode de stockage**
  - Bases de données NoSQL (MongoDB, Cassandra, etc.)
  - Systèmes distribués de stockage : Hadoop HDFS, Amazon S3
  - Ingestion des données : Apache Kafka, Flume, Sqoop

Module 3 : Traitement et manipulation des données

- Prétraitement des données
- Traitement batch avec Hadoop / MapReduce

- Traitement en temps réel avec Apache Spark, Flink
- Nettoyage, transformation et enrichissement des données
- Outils de manipulation des données : Spark SQL, PySpark, Pandas

Module 4 : Analyse et modélisation des données

- Statistiques descriptives et exploratoires sur grands volumes
- Analyse prédictive
- Classification, régression, clustering dans un environnement Big Data
- Détection d'anomalies et recommandations

Module 5 : Visualisation et interprétation des résultats

- Outils de data visualisation : Power BI, Tableau, Kibana, matplotlib
- Création de tableaux de bord interactifs
- Communication des résultats aux décideurs
- Cas pratiques de storytelling avec les données

Module 6 : Mise en œuvre d'un projet Big Data

- Étapes d'un projet Big Data (de la problématique à l'analyse)
- Choix technologiques selon les objectifs
- Bonnes pratiques de gestion de projet et de gouvernance des données
- Étude de cas complet (ex. : analyse des données clients, trafic, santé, etc.)