Développez vos talents
Certifiez vos compétences

Data Engineering

Extraction de données avec Python

3 jours (21 heures) | 4,6/5 | BIG-COLL | Évaluation qualitative de fin de stage | Formation délivrée en présentiel ou distanciel [1]

Formations Informatique › Big Data › Data Engineering

À l'issue de ce stage vous serez capable de :

- Réaliser du scraping de données
- Faire les actions d'ingestion nécessaires pour alimenter un Data Lake.

Niveau requis

Avoir des connaissances en algorithmiques.

Public concerné

Développeurs, intégrateurs, chefs de projets, consultants BI.

Cette formation :

- Est animée par un consultant-formateur dont les compétences techniques, professionnelles et pédagogiques ont été validées par des diplômes et/ou testées et approuvées par l'éditeur et/ou par M2i Formation
- Bénéficie d'un suivi de son exécution par une feuille de présence émargée par demi-journée par les stagiaires et le formateur.

(1) Modalité et moyens pédagogique :

Formation délivrée en présentiel ou distanciel * (e-learning, classe virtuelle, présentiel à distance). Le formateur alterne entre méthodes ** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation). La validation des acquis peut se faire via des études de cas, des quiz et/ou une certification.

Les moyens pédagogiques mis en oeuvre (variables suivant les formations) sont : ordinateurs Mac ou PC (sauf pour les cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel). Environnements de formation installés sur les postes de travail ou en ligne. Supports de cours et exercices.

* Nous consulter pour la faisabilité en distanciel. ** Ratio variable selon le cours suivi.

Programme

Les bases du langage Python

- Les caractéristiques du langage Python
- Pourquoi choisir Python pour l'analyse de données ?
- Types de bases
- Les instructions de bases
- Les procédures et fonctions

L'ingestion avec Python

- Utiliser la librairie Pandas pour manipuler les données
- Introduction du concept de DataFrame
- Les structures :
  - Interrogation
  - Indexation
- Traitement de "données manquantes"
- Fusion de DataFrames
- Manipulation des dates
- Application de mesures statistiques variées sur les DataFrames
- Bonne compréhension des problèmes d'échelle de mesure, de normalisation
- Création de métriques d'analyse

Scraping de données

- Qu'est-ce que le scraping ?
- Définition du scraping et de ses différents niveaux de difficulté sur plusieurs supports
  - Depuis le Web
  - Depuis du papier
  - Depuis des PDF
- Exemples de projets réalisés grâce au scraping
- L'architecture d'Internet
- Qu'est-ce qu'un "client" ? Qu'est-ce qu'un "serveur" ? Pourquoi est-ce important ?
- Comment HTTP et HTML impactent-t-ils nos scrapers ?
- Qu'est-ce qu'une balise HTML ? Un attribut ?
- Comment identifier certains éléments avec une "class" ou un "id" ?

Python comme solution ETL

- Les formats de données structurées : CSV, flux XML et JSON
- Lecture et écriture de fichiers
- Exploitation des données de fichiers de différentes sources
- Fonctions d'accès et de chargement de données en blocs de lignes
- Outils spécifiquement dédiés au scraping :
  - Beautiful Soup
  - CSS Select

Mise en oeuvre d'un scraper

- Un scraper simple (requêtes GET, pages séquencées)
- Identifier la stratégie à adopter pour naviguer sur le site
- Coder le scraper
- Un scraper complexe : envoyer des données à un site Internet pour obtenir des résultats plus complexes
- Qu'est ce qu'une requête POST et une requête GET ?
- Parcourir un site pour trouver les données
- Identifier la stratégie à adopter
- Coder le scraper