

Formation Data Engineering avec Python – ETL, Pandas et pipelines de données

Jour 1 – Préparation et analyse des données

- Introduction à NumPy et Pandas
- Lecture de fichiers CSV
- Lecture de fichiers Excel
- Manipulation de données JSON
- Introduction à l'utilisation de données provenant de bases SQL
- Sélection et filtrage des données
- Jointures
- Agrégations
- Gestion des valeurs manquantes
- Détection et traitement des doublons
- Gestion des valeurs aberrantes
- Transformation des variables
- Normalisation et standardisation
- Analyse exploratoire
- Visualisation
- Contrôle de la qualité des données

Travaux pratiques – Jour 1

Nettoyage et analyse d'un jeu de données contenant des erreurs, des valeurs manquantes et des données incohérentes.

Jour 2 – Développement de pipelines de données

- Principe d'un pipeline de données
- ETL : Extract, Transform, Load
- Organisation d'un projet Python
- Découpage du traitement en fonctions
- Automatisation des traitements
- Gestion des erreurs et exceptions
- Journalisation avec logging
- Validation automatique des données
- Tests simples du pipeline
- Formats de sortie : CSV et Parquet
- Préparation des données pour le Machine Learning
- Introduction à Git et à la reproductibilité

Mini-projet pratique

Données brutes → validation → nettoyage → transformation → analyse →
préparation ML → données finales → rapport d'erreurs.