Data management

ECTS : 6

Description du contenu de l'enseignement :

Le cours sera principalement axé sur les travaux pratiques (TP). Les étudiants travailleront sur un projet intégré où ils devront concevoir une pipeline complète d'ingestion de données, incluant la collecte, le traitement, et la visualisation des données. Chaque session abordera une étape clé du pipeline, suivie d'exercices pratiques pour consolider les connaissances.

Compétence à acquérir :

Ce cours permettra d'acquérir les compétences pour créer et gérer des pipelines de données complets, depuis la collecte jusqu'au traitement et à la visualisation, en utilisant des pratiques de développement orienté production :

- Récupération de données à partir de multiples sources
  - Extraction de données via des APIs Web ou via scrapping.
  - Peuplement et manipulation de bases de données SQL.
  - Manipulation de données stockées dans des buckets S3.
- Processing de données brutes
  - Utilisation avancée de *pandas* pour le chaînage de transformations.
  - Optimisation des calculs numériques avec *numpy*.
  - Introduction à l'analyse de données (modélisation et problématiques éthiques).
- Code orienté production
  - Utilisation de *VS Code* pour le développement.
  - Gestion de version avec *Git*.
  - Structuration et gestion de scripts pour automatiser les tâches.
  - Principes de la *Programmation Orientée Objet (OOP)* pour la modularité et la maintenance du code.
- Visualisation des données
  - Création de graphiques pour explorer et communiquer les résultats.
  - Utilisation de bibliothèques de visualisation comme *Matplotlib*, *Seaborn*, et autres.

Mode de contrôle des connaissances :

L'évaluation se fera sous la forme d'un devoir maison en groupe. Les étudiants devront livrer un projet final comprenant une pipeline fonctionnelle et documentée, qui met en œuvre les compétences acquises tout au long du cours. Le projet sera évalué sur la qualité du code, la robustesse de la solution, et la clarté de la documentation.