Syllabus du cours « Big Data »

Durée : 48h

Crédits : 6 ECTS

Objectifs pédagogiques

L'objectif principal de ce cours est de fournir aux étudiants une compréhension solide et complète de la technologie Apache Spark et son écosystème. À la fin du cours, les étudiants auront :
- une meilleure compréhension des enjeux liés au Big Data et du cyle de la donnée
- une meilleure connaissance théorique et pratique des technologies associées
- une maitrise de leur utilisation en environnement professionnel (use case réel et en mode DevOps)

Prérequis

Maitrise du language de programmation Python (débutant/intermédiaire)

Contenu

Thèmes :

- → Introduction au Big Data (2 heures)
  - Définition et enjeux du Big Data
  - Écosystème Hadoop

- → Fondamentaux HDFS et MapReduce (2 heures)
  - HDFS : architecture et fonctionnement
  - Paradigme MapReduce

- → Traitement & analyse de données avec Spark (10 heures)
  - Introduction à Apache Spark et comparaison avec MapReduce
  - Architecture de Spark et principales librairies (Spark SQL, Spark Streaming, MLlib, GraphX)
  - RDDs (Resilient Distributed Datasets)
  - Plan d'exécution & Opérations (transformations vs actions)
  - Applications Spark et gestion des ressources clusters (Standalone/YARN, Zookeeper)
  - Introduction à PySpark (API Python pour Spark)
  - Introduction à Dbt

- Exercices et cas pratiques

→ Streaming (4 heures)

- Introduction à Stream Data Processing
- Introduction à Flink & Kafka
- Comparaison Flink, Kafka, Spark streaming
- Hands-on Kafka & Spark Streaming

→ Data Visualization (2 heures)

- Introduction à Data Visualization (Presentation of PowerBI & Tableau)
- Hands-on PowerBI & Tableau

→ Stockage et gestion des données (6 heures)

- Base de données relationnelles
- No SQL & HDFS (Hadoop Distributed File System)
- Inputs/Outputs formats (Avro, Parquet, Csv, images, blob etc...)
- NoSQL vs SQL comparaisons
- Introduction à MongoDB, Cassandra, Neo4j, HBase

→ Cloud services (4 heures)

- Déploiement sur le cloud (AWS, Azure, Google cloud)
- Utilisation de (Azure) Databricks
- Intégration avec d'autres services cloud (Azure)

→ Architecture Big Data & DevOps (8 heures)

- Architectures Big Data: Lambda, Kappa, Medallion
- CI/CD pour les applications Big Data
- Orchestration (Airflow) & Data pipelines en entreprise

→ Gouvernance de données (4 heures)

- Data stewardship
- Sécurité et gouvernance des données

→ Projets pratiques (4 heures)

– Au choix: Analyse de données en temps réel / Machine Learning avec Spark MLlib / Traitement de graphes avec GraphX

Compétences

Ce cours permettra de développer des compétences en acquisition, traitement et stockage de données variées et à forte volumétrie. L'accent est mis afin qu'ils maitrisent les compétences attendues en entreprise. A savoir, travail en collaboration, développement en mode CI/CD (DevOps) et méta-apprentissage (indispensable pour maintenir des compétences à jour).

Approche pédagogique

Quelle est l'approche retenue pour ce cours ? Qu'attend-on des élèves pendant le cours ? Entre deux cours ?...

L'approche pédagogique est classique tout en gardant un certain pragmatisme. L'idée est d'équiper d'un savoir théorique indispensable pour comprendre les enjeux, besoins et l'univers technologique lié au big data en tant qu'ingénieur. Mais surtout de mettre l'accent sur la pratique, la pratique des outils, la confrontation aux principales problématiques ainsi que les modes d'organisations et fonctionnements au sein de l'entreprise. Le but n'est pas d'être exhaustif mais plutôt de les "armer" au mieux afin qu'il puisse rapidement s'intégrer au sein d'une équipe et relever les challenges lié à leur fonction.

A cette fin, il est attendu des élèves une écoute et participation active, un travail de recherche en dehors des cours et la réalisation des devois et exercices.

Déroulement par séance

Lors de la première séance, le plan de cours et les différentes sections seront présentés. Les thèmes pour les exposés (offrant la possibilité de se sensiblliser à des enjeux non couvert en cours) seront répartis entre les étudiants. L'enseignant présentera le cours, exercices et travaux pratiques. Certaines séances seront introduites/complétées par les exposés des élèves.

Modalités d'évaluation

Les étudiants seront évalués sur :
> participation et résolution des exercices/TPs (vu en cours)

> leur **présentation/restitution** d'exposé du projet pratique (comprenant)
- la qualité de leurs recherches,
- la qualité et l'originalité de leur présentation
- la qualité du rendu écrit (inclue les slides et tout autre document)
- l'évaluation du code rendu et déposé sur Github

> **Une évaluation finale** (théorie (QCM) + pratique (rendu de code)).

Bibliographie

- Designing Data-Intensive Applications (*Martin Kleppmann*)
- **"Spark: The Definitive Guide"** par *Bill Chambers, Matei Zaharia*

- **"Architecting the Cloud: Design Decisions for Cloud Computing Service**
- **Models (SaaS, PaaS, and laaS)"** par *Michael J.*
  Kavis
- **"Data Science for Business: What You Need to Know about Data Mining and**
- **Data-Analytic Thinking"** par Foster Provost, Tom
  Fawcett