3MD1060 - Infrastructure de traitement parallèle des Big Data

Responsables : **Pauline LAFITTE , Gianluca QUERCINI**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE PARIS - SACLAY**

Nombre d'heures d'études élèves (HEE) : **40**

Nombre d'heures présentielles d'enseignement (HPE) : **24**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Ce cours a pour objectif d'introduire les principales technologies pour faire face aux nombreux défis posés par les Big Data.

Big Data est un terme utilisé pour décrire une collection de données qui est énorme en volume et qui pourtant croît de façon exponentielle avec le temps. En bref, ces données sont si volumineuses et complexes qu'aucun des outils traditionnels de gestion des données n'est capable de les stocker ou de les traiter efficacement.

Ce cours présente les technologies existantes qui rendent possible le traitement efficace de grands volumes de données, à savoir Hadoop MapReduce et Apache Spark.

Période(s) du cours (n° de séquence ou hors séquence) :

SD9

Prérequis :

- Programmation Python
- Notions en bases de données relationnelles

Plan détaillé du cours (contenu) :

Programmation MapReduce

- Introduction à MapReduce: motivations et exemples.
- Mise en œuvre de MapReduce.

Hadoop

- Introduction à Hadoop et ses fonctionnalités.
- Hadoop Distributed File System (HDFS)

Introduction à Apache Spark

- Introduction à Spark et ses fonctionnalités (RDD, transformations, actions, dataframes, datasets).
- Architecture de Spark et modèle d'exécution.
- Mise en œuvre d'Apache Spark.

Composantes d'Apache Spark

- SparkSQL, Spark Streaming, Machine learning and Graph analysis.

Organisation de l'évaluation :

Examen écrit à la fin (3h) et un TP noté

Moyens :

Equipe pédagogique : Gianluca Quercini

Acquis d'apprentissage visés dans le cours :

- Comprendre les concepts à la base du Big Data.
- Utiliser des paradigmes de calcul distribué : MapReduce et Spark.
- Concevoir des algorithmes de calcul distribué sur les données.

Description des compétences acquises à l'issue du cours :

Exploit any type of data, structured or not, including massive data.

Bibliographie :

- Transparents fournis par l'enseignant.
- Singh, Chanchal, and Manish Kumar. Mastering Hadoop 3: Big data processing at scale to unlock unique business insights. Packt Publishing Ltd, 2019.
- Mehrotra, Shrey, and Akash Grade. Apache Spark Quick Start Guide: Quickly learn the art of writing efficient big data applications with Apache Spark. Packt Publishing Ltd, 2019.
- Karau, Holden, et al. Learning spark: lightning-fast big data analysis. " O'Reilly Media, Inc.", 2015.