3MD4130 - Modèles de calcul du Big Data

Responsables : **Stephane VIALLE**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE METZ**

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

Nombre d'heures présentielles d'enseignement (HPE) : **21**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Ce cours a pour objectif d'apprendre aux élèves à développer des applications performantes d'analyse de données en environnement Spark sur des plates-formes distribuées (clusters et Clouds). Des mécanismes de systèmes de fichiers distribués comme HDFS seront étudiés, ainsi que le modèle de programmation et l'algorithmique du "map-reduce étendu" de Spark au dessus des Spark "RDD", puis des modèles de programmation de plus haut niveau au dessus de Spark "Data Frames", et enfin des modèles de programmation sur Cloud. Des critères et métriques de passage à l'échelle seront également étudiés. Tout au long du cours des mises en oeuvres auront lieu sur des clusters et dans un Cloud, et les solutions développées seront évaluées par les performances obtenues sur les cas-tests, et par leur aptitude à passer à l'échelle.

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

SM11

Prérequis :

- Cours commun de 1A : Systèmes d'Information et Programmation (1CC1000)
- Cours commun de 1A : Algorithmique & Complexité (1CC2000)
- Cours de la mention SDI Metz : Ingénierie des données et du logiciel (3MD1510)

Plan détaillé du cours (contenu) :

- Emergence des technologies Big Data : motivations, besoins industriels, principaux acteurs.
- Pile logicielle d'Hadoop, architecture et fonctionnement de son système de fichier distribué (HDFS)
- Architecture et mécanisme de déploiement de calculs distribués de Spark
- Modèle de programmation par "RDD" et algorithmique du "map-reduce" étendu de Spark
- Modèle de programmation de Spark par "Data Frames" appliqué à l'analyse de graphes (GraphX)
- Architecture et environnement d'analyse de données sur Cloud
- Expérimentations et mesures de performances
- Critères et métriques de performances

Déroulement, organisation du cours :

Ce cours enchaîne 3 parties relatives à des modèles de calculs du "Big Data" : la première sur clusters de PC, la seconde dans les Cloud, et la troisième qui évalue des solutions de "passage à l'échelle".

Répartition globale : CM : 10h30, TD : 1h30, TP: 9h00 (total 21,00 HPE)

Plan du cours en 4 parties :

- Partie 1 : Architecture et développement en Spark RDD sur HDFS et clusters de PC.
  - CM : 4h30, TD : 1h30, TP : 3h00
- Partie 2 : Critères et métriques pour la performance et le passage à l'échelle.
  - CM : 1h30
- Partie 3 : Calcul et analyse de données large échelle sur Cloud.
  - CM : 3h00, TP : 3h00
- Partie 4 : Développement en Spark Data Frames sur HDFS et clusters de PC.
  - CM : 1h30, TP : 3h00

Organisation de l'évaluation :

Evaluation à partir des TP :

- **Les comptes rendus des TP seront notés** (le contenu et le nombre de pages des comptes rendus sont contraints, afin de forcer les étudiants à un effort de synthèse et de clarté)
- En cas d'absence non justifié à un TP la note de 0 sera appliquée, en cas d'absence justifiée le TP n'interviendra pas dans la note finale.
- L'examen de rattrapage sera un examen oral, qui constituera 100% de la note de rattrapage.

Moyens :

- Equipe enseignante : **Stéphane Vialle** et **Gianluca Quercini** (CentraleSupelec), **Wilfried Kirschenmann** (ANEO)
- Plateforme de développement et d'exécution :
  - **clusters de calcul du Data Center d'Enseignement (DCE)** du campus de Metz de CentraleSupélec
  - **accès à un Cloud professionnel**
- Environnements de développement :
  - **Spark + HDFS sur le DCE**
  - **autre environnement sur le Cloud**

Acquis d'apprentissage visés dans le cours :

A l'issue de ce cours, les élèves seront en mesure :

- **AA1** : de concevoir et d'implanter des algorithmes "map-reduce étendu" en Spark, efficaces sur des plates-formes distribuées, et passant à l'échelle.
- **AA2 :** d'analyser les capacités de "passage à l'échelle" d'une application
- **AA3** : d'utiliser un cluster ou un cloud pour réaliser des analyses de données distribuées à large échelle.
- **AA4** : de présenter synthétiquement une solution d'analyse de données conçue en "map-reduce"

Description des compétences acquises à l'issue du cours :

Design of algorithms and implementation of codes with scaling capabilities allow to:

- extend the acquisition of 2 skills:
  - **C6.3**: Conceive, design, implement and authenticate complex software
    - **Marker 3** : Parallel software development (considering large scale "Big Data" software)
    - According to the Learning Outcomes **AA1**, **AA2**
  - **C6.4:** Solve problems through mastery of computational thinking skills
    - Marker 2 : Exploitation of parallel architectures (considering large scale distributed architectures for "Big Data")
    - According to the Learning Outcomes AA3
- acquire a skill allowing to present clearly an original software approach
  - **C7.1:** Persuade at core value level; to be clear about objectives and expected results. To apply rigour when it comes to assumptions and structured undertakings, and in doing so structure and problematise the ideas themselves. Highlight the added value.
    - **Marker 1**: evaluated from a report or a presentation of a solution within the framework of "classic" lessons
    - According to the Learning Outcomes AA4

Bibliographie :

- Slides des enseignants

- BdD NoSQL :
  - Kristina Chorodorw. MongoDB. The Definitive Guide. 2nd edition. O'Reilly. 2013.
  - Rudi Bruchez. Les bases de données NoSQL et le Big Data. 2ème édition. Eyrolles. 2016.
- Hadoop & Map-Reduce :
  - Tom White. Hadoop. The definitive Guide. 3rd edition. O'Reilly. 2013.
  - Donald Miner and Adam Shook. MapReduce Design Patterns. O'Reilly. 2013.
- Spark :
  - M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, and I. Stoica. Resilient Distributed Datasets : A Fault-tolerant Abstraction for In-memory Cluster Computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, NSDI’12, 2012.
  - H. Karau, A. Konwinski, P.Wendell, and M.Zaharia. Learning Spark. O’Reilly, 1st edition, 2015.
  - H. Karau and R. Warren. High Performance Spark. O’Reilly, 1st edition, 2017.