2EL5040 - Big Data : collecte, stockage et analyse de données sur clusters et sur Cloud

Responsables : **Stephane VIALLE**

Langues d'enseignement : **FRANCAIS**

Type de cours : **Electif 2A**

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

Nombre d'heures d'études élèves (HEE) : **60**

Nombre d'heures présentielles d'enseignement (HPE) : **30**

Année académique : **2024-2025**

Catégorie d'électif : **Sciences fondamentales**

Niveau avancé : **oui**

Présentation, objectifs généraux du cours :

La baisse du coût des capteurs favorise leur utilisation fréquente dans tous les environnements (industriels, urbains, transports...), et provoque l'émergence de flux de données brutes. De même les données structurées accessibles sur le web ou dans des archives privées d'entreprises ne cessent d'augmenter. Des technologies "Big Data" ont vu le jour, et ont évolué rapidement, pour gérer, analyser et tirer parti ces sources de données.

- Ce cours présente les environnements du Big Data qui ont émergé pour stocker et interroger ces nouvelles masses de données : notamment des BdD NoSQL et des environnements distribués comme Hadoop et Spark. Ces environnements sont nés dans les industries innovantes du web, et ont amenés de nouveaux paradigmes de programmation comme Map-Reduce (implanté selon plusieurs variantes).
- Une partie importante du cours est consacré à la conception d'algorithmes de filtrage, d'enrichissement et d'analyse des données stockées dans les environnements du Big Data. La plupart de ces algorithmes sont basés sur le paradigme de programmation Map-Reduce et seront expérimentés lors de TP. Des métriques de performance et des critères de passage à l'échelle de systèmes distribués seront également présentés et utilisés en TP.
- La dernière partie du cours présente des algorithmes de Machine Learning, utilisés pour traiter et analyser des ensembles de données, et qui demandent parfois d'avoir recours à du calcul parallèle massif sur GPU.

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

SG8

Prérequis :

- Cours commun "Systèmes d'Information et Programmation" de la SG1 (1CC1000)
- Cours commun "Algorithmique & Complexité" de la ST2 (1CC2000)
- Cours commun "Statistique et Apprentissage" de la ST4 (1CC5000)

Plan détaillé du cours (contenu) :

- **Introduction et terminologie (1CM – 1h30):** Data Engineering vs Data Science, architectures distribuées matérielles et logicielle, analyse de données à haute performances, parallélisation SMPD vs Map-Reduce.
- **Environnement et technologie Hadoop (1CM – 1h30):** Système de fichiers distribué (HDFS), principe du Map-Reduce d'Hadoop, gestion de ressources version 1 avec limite d'échelle, et version 2 optimisée (YARN).
- **Environnement et technologie Spark (3CM – 4h30):** Architecture et mécanismes orientés performances de Spark, algorithmique Map-Reduce simple, algorithmique Map-Reduce pour l'analyse de graphes, bibliothèques Spark-SQL et de traitements de flux.
  - **TD1 & TD2 (3h00)**
  - **TP1 & TP2 (6h00)** sur clusters de PC
- **Métriques et limites de passage à l'échelle (1CM - 1h30):** métriques d'accélération et d'efficacité, critères de passage à l'échelle.
- **Exploration et préparation des données (1CM - 1h30):** problèmes classiques rencontrés avec les données, besoin d'exploration et de préparation des données
- **BdD NoSQL (2CM – 3h00):** Emergence des BdD NoSQL, technologies NoSQL, utilisation de MongoDB
  - **TP3 (3h00)**
- **Introduction aux technologies de Machine Learning (ML) (3CM : 4h30):** classification des algorithmes de ML, algorithmes de clustering, exemples de bibliothèques de ML en Python
  - **TP4 (3h00)**
- **Examen Ecrit (1h30)**

Déroulement, organisation du cours :

Les concepts vus en cours seront mis en oeuvre lors de TP sur des clusters Big Data du Data Center d'Enseignement de CentraleSupélec. Ces plateformes permettront de manipuler et d'interroger des environnements Spark et MongoDB distribués sur des clusters de PC, et hébergeant de gros volumes de données. Des serveurs de calcul permettront également la mise en oeuvre de bibliothèques de Machine Learning dans la dernière partie du cours. Des mesures de performances compléteront l'évaluation des solutions développées en TP.

Composition du cours : 12 CM (18h00), 2 TD (3h00), 4 TP (12h00) et 1 EE (1h30)

Séquencement possible du cours :

- 3CM, 1TD, 1TP, 1CM, 1TD, 1TP, 6CM, 1TP, 2CM, 1TP
- Examen écrit de 1h30

Organisation de l'évaluation :

Poids relatifs des différents examens :

- 40% : Rapports de TP. Une absence injustifiée à un TP entraînera une notre de 0/20 à ce TP. Une absence justifiée à un TP neutralisera la note de ce TP et augmentera le poids des autres TP.
- 60% : examen écrit final de 1h30, avec documents.

Examen de rattrapage : En cas de rattrapage, 100% de la note viendra d'un examen écrit de 1h30, dans la même modalité que l'examen écrit initial.

Moyens :

- 18h00 de cours de Data Engineering, incluant la présentation d'environnements standards et distribués de Big Data, et la conception de solutions performantes et passant à l'échelle.
- 3h00 de TD de dimensionnement d'architecture et d'algorithmique Map-Reduce.
- 12h00 de TP de mise en oeuvre de logiciels Big Data standard et Opensource (Hadoop HDFS, Spark, MongoDB, bibliothèques de Machine Learning), sur des serveurs et clusters de calculs à haute performants (ressources du Data Center d'Enseignement de CentraleSupelec).

Acquis d'apprentissage visés dans le cours :

A la fin de cet enseignement, l'élève sera capable de :

- **[Acquis d'Apprentissage 1* (AA1*)]** Spécifier, concevoir et présenter un système complexe et cohérent d'analyse de données large échelle (contribuant aux compétences **C2 C6 C7**) :
  - spécifier et dimensionner une architecture matérielle Big Data
  - choisir un environnement Big Data adapté (ex : Spark et certains de ses modules, ou un certain type de BdD NoSQL...)
  - concevoir un algorithme et une chaîne logicielle Map-Reduce, selon la variante du paradigme Map-Reduce qui est disponible (permettant de nettoyer, préparer, filtrer et interroger de gros volumes de données)
  - optimiser un algorithme Map-Reduce dans un objectif de performance et de passage à l'échelle
  - spécifier et dimensionner une architecture matérielle de Machine Learning (ex : CPU, cluster de CPU, GPU, cluster de GPU...)
  - présenter un résumé convaincant de l'architecture logicielle et matérielle mise au point
- **[Acquis d'Apprentissage 2* (AA2*)]** Evaluer et présenter les performances et la robustesse d'une architecture Big Data (contribuant aux compétences **C2 C6 C7**) :
  - définir une métrique et une procédure de test de passage à l'échelle adaptée au cas d'utilisation
  - identifier et présenter les goulots d'étranglement de l'architecture matérielle et logicielle (en cas d'augmentation du volume de données)
  - identifier et présenter les points faibles de l'architecture en cas de panne (single point of failure)
  - identifier et présenter les types d'imperfections de données perturbant la chaîne d'analyse

Description des compétences acquises à l'issue du cours :

- **C2:** Develop an in-depth skills in an engineering field and in a family of professions
- **C6:** Be operational, responsible, and innovative in the digital world
- **C7:** Know how to convince

Bibliographie :

Supports fournis aux étudiants :

- Slides et polycopié du cours.

Livres suggérés :

- Pirmin Lemberger, Marc Batty, Médéric Morel et Jean-Luc Raffaëlli. Big Data et Machine Learning. Dunod. 2015.
- Eric Biernat et Michel Lutz. Data Science : Fondamentaux et études de cas. Eyrolles. 2015.
- Bahaaldine Azarmi. Scalable Big Data Architecture. Apress. 2016.
- Kristina Chorodorw. MongoDB. The Definitive Guide. 2nd edition. O'Reilly. 2013.

- H. Karau, A. Konwinski, P. Wendell and M. Zaharia. Learning Spark. O'Reilly. 2015.
- Rudi Bruchez. Les bases de données NoSQL et le Big Data. 2ème édition. Eyrolles. 2016.
- Tom White. Hadoop. The definitive Guide. 3rd edition. O'Reilly. 2013.
- Donald Miner and Adam Shook. MapReduce Design Patterns. O'Reilly. 2013.
- Matthew Kirk. Thoughtful Machine Learning with Python. O'Reilly. 2017.