2EL5040 – Big Data: data gathering, storage and analysis on clusters and Clouds

Instructors: Stephane Vialle
Department: CAMPUS DE METZ
Language of instruction: FRANCAIS
Campus: CAMPUS DE METZ
Workload (HEE): 60
On-site hours (HPE): 35,00
Elective Category : Fundamental Sciences
Advanced level : Yes

Description

Decrease of sensor price make easier their usage in various environments (inside factories, cities, transports…) and generate many raw data flows. A similar increase can be observed with structured data available on the web or in private archives of companies. Some "Big Data" technologies have appeared and quickly evolved to manage and analyse these data sources.

This course presents the Big Data environments that have emerged to store and interrogate these new Big Data: in particular NoSQL BdD and distributed environments like Hadoop and Spark.These environments were born in the innovative web industries, and have brought new programming paradigms like Map-Reduce (implemented in several variants).

An important part of the course is devoted to the design of algorithms for filtering, enriching and analyzing data stored in Big Data environments. Most of these algorithms are based on the Map-Reduce programming paradigm and will be tested during labs. Performance metrics and criteria for scaling up distributed systems will also be presented and used in labs.

The last part of the course presents Machine Learning algorithms, used to process and analyze data sets, and which sometimes require the use of massive parallel computing on GPUs.

Quarter number

SG8

Prerequisites (in terms of CS courses)

- SG1 common course "*Systèmes d'Information et Programmation*" (1CC1000)
- ST2 common course "*Algorithmique & Complexité*" (1CC2000)
- ST4 common course "*Statistique et Apprentissage*" (1CC5000)

Syllabus

- Introduction and terminology (1CM - 1h30): Data Engineering vs Data Science, distributed hardware and software architectures, high performance data analysis, SMPD vs Map-Reduce parallelization.
- Hadoop environment and technology (1CM - 1h30): Distributed file system (HDFS), Hadoop Map-Reduce principle, resource manager version 1 with scale limit, and optimized version 2 (YARN).
- Spark environment and technology (3CM - 4h30): Spark performance-oriented architecture and mechanisms, simple Map-Reduce algorithm, Map-Reduce algorithm for graph analysis, Spark-SQL libraries and stream processing.
    - Tutorial courses 1 & 2 (3h00)
    - Labs 1 & 2 (6h00) on PC clusters
- Metrics and scaling limits (1CM - 1h30): acceleration and efficiency metrics, scaling criteria.
- Data exploration and preparation (1CM - 1h30): classic problems encountered with data, need for data exploration and preparation
- NoSQL data bases (2CM – 3h00): Emergence of NoSQL databases, NoSQL technologies, use of MongoDB
    - Lab 3 (3h00)
- Introduction to Machine Learning (ML) technologies (3CM : 4h30): classification of ML algorithms, clustering algorithms, examples of ML libraries in Python
    - Lab 4 (3h00)
- Written examination (1h30)

Class components (lecture, labs, etc.)

Theoretical issues introduced during the different lectures will be experimented during some labs on Big Data clusters of CentraleSupelec *Data Center for Education.* These experimental plateforms will allow to request Spark and MongoDB environments, distributed on PC clusters and managing large volumes of data. During the last part of the course, some computing servers will allow to efficiently run Machine Learning libraries. Some performance measurements will complete the evaluation of the different solutions developped during the labs.

Composition of the course: lectures 18h00 (12 x 1h30), tutorials 3h00 (2 x 1h30), labs 12h00 (4 x 3h00) and a final written exam (1h30)

Possible schedule of the course:

- 3 courses, 1 tutorial course, 1 lab, 1 course, 1 tutorial course, 1 lab, 6 courses, 1 lab, 2 courses, 1lab
- Written exam (1h30)

Grading

Relative weights of the different examinations:

- 40%: lab reports. Any unjustified absence at labworks will result in a score of 0. A justified absence at a labwork will neutralize the score of this labwork and will increase the weight of the others.
- 60%: final written exam of 1h30, with documents.

Remedial examination: If a remedial exam is necessary, 100% of the score will depend on a written exam of 1h30, with same modalities than the initial written exam.

Course support, bibliography

Documents supplied to the students:

- Slides et notebook of the teachers.

Suggested books:

- Pirmin Lemberger, Marc Batty, Médéric Morel et Jean-Luc Raffaëlli. *Big Data et Machine Learning*. Dunod. 2015 (in french).
- Eric Biernat et Michel Lutz. Data Science : *Fondamentaux et études de cas*. Eyrolles. 2015 (in french).
- 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.

Resources

- 18h00 of lectures about Data Engineering including: the introduction to standard and distributed Big Data environments, and the design of fast and scalable solutions.
- 3h00 of tutorials about architecture sizing and Map-Reduce algorithmics.

- 12h00 of labs about experimentation of standard and Opensource Big Data software (Hadoop HDFS, Spark, MongoDB, Machine Learning libraries), on high performance computing servers and clusters (resources of the *Data Center for Education* of CentraleSupelec).

Learning outcomes covered on the course

When finishing the course, the students will be able:

- **[Learning Outcomes 1* (AA1*)]** Specify, design and present a complex and consistent system for large scale data analysis (contributing to core skills **C2 C6**):
    - to specify and to set the size of a Big Data hardware architecture
    - to choose a Big Data environment adapted to the use case (ex: Spark and some of its libraries, or some kind of NoSQL dtabases...)
    - to design a Map-Reduce based software architecture and algorithm, function of the available Map-Reduce variant (in order to clean, to prepare, to filter and to request large data)
    - to optimize a Map-Reduce based algorithm to improve its performances and scalability
    - to specify and to set the size of a Machine Learning hardware architecture (ex: CPU, CPU cluster, GPU, GPU cluster...)
    - present a convincing summary of the software and hardware architecture developed

- **[Learning Outcomes 2* (AA2*)]** Evaluate and present performances and strength of a Big Data architecture (contributing to core skills C2 C6):
    - to define a metric and a scaling benchmark adapted to the use case
    - to identify the bottlenecks of the hardware and software architectures (when increasing the data volume)
    - to identify the single points of failure of the global architecture
    - to identify the kind of incorrect data disturbing the analysis

Description of the skills acquired at the end of the course

- **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