Big Data: Distributed Data Management and Scalable Analytics

Lecturers
Dimitrios SACHARIDIS (Coordinator) and Gianluca BONTEMPI

Course mnemonic
INFO-H515

ECTS credits
5 credits

Language(s) of instruction
English

Course period
Second term

Campuses
Solbosch and Plaine

Course content

The course is subdivided into 2 parts: Big Data Management and Big Data Analytics. The part on Big Data Analytics builds on concepts introduced in the part Big Data Management.

Part I: Big Data Management:

1. Introduction & Map/Reduce
2. Spark
3. Streaming: Spark Streaming & Storm; Big Data Architectures
4. Consistency and Availability
5. Distributed and parallel query evaluation
6. Stream Processing and Sublinear Algorithms

Part II: Big Data Analytics:

1. Introduction
2. Batch distributed machine learning
3. Sequential machine learning and streaming
4. Recommender systems and Collaborative filtering
5. Deep learning

Objectives (and/or specific learning outcomes)

This is an introductory course on big data management and analytics. Its objective is to introduce students to the fundamental notions, principles, and research results concerning modern, scalable, and fault-tolerant ways for managing and analyzing massive amounts of data using parallel and distributed systems. Armed with this knowledge, the student will be able to decipher, use, and compare the plethora of big-data technologies currently used in industry.

Learning outcomes

After successful completion of this course, the student:

1. Understands the characteristics of big data, and the challenges these represent
2. Knows the principal architectures of Big Data Management and Analytics Systems (BDMAS), is able to explain the purpose of each their components, and is able to recognize and explain the key properties, strengths and limitations of each type of BDMAS and their components.
3. Understands the key bottlenecks in managing and analyzing massive amounts of data and is familiar with modern algorithms for overcoming these bottlenecks using parallel and distributed computation.
4. Is able to actively use this algorithmic knowledge in the design and implementation of applications that solve common data management and analytics problems using different types of BDMAS.
5. Is able to build applications using specific instances of each type of BDMAS.
6. In addition, is able to use established software frameworks for reproducing/sharing her/his results, including virtualization software (Docker), version control systems (Git), and notebooks (Jupyter, Zeppelin)
7. Is able to implement an analytics pipeline (e.g. in Spark) able to process and learn predictive models from massive datasets

Pre-requisits and co-requisits

Required knowledge and skills

- Databases, SQL
- Supervised machine learning (classification, regression, feature selection)
- Basic notions of statistics and probability
- Programming in Python

Teaching method and learning activities

Combination of Ex-Cathedra Lectures, Exercise sessions, Computer labs, Self-study, and Project Work.

Contribution to the teaching profile

- Be capable of formulating and solving complex or open-ended technical and scientific problems by using abstraction, modeling, simulation, and multi-disciplinary analysis while satisfying the requirements of university-level research and responding to requirements, constraints, the set context and the technical, socio-economical ethical and environmental stakes—all with the purpose of obtaining concrete solutions.
- Have in-depth knowledge and understanding of a structured body of knowledge, both transversal and specialised. Be capable of autonomously and critically following current trends and advances in this body of knowledge.
- Define, plan, manage, and execute projects taking into account their objectives, the available resources and constraints; assuring the coherence and quality of the work and deliverables.

- Work efficiently with other professionals (in group, in partnership, or in competition), make decisions and develop leadership, in a variety of professional contexts, disciplines, and cultures.
- Communicate and share information in a structured manner: orally, graphically and written, in French and in one or more other languages. Communicate on scientific, technical and cultural aspects, adapting him/herself to the desired goal as well as the target audience.

References, bibliography and recommended reading

See recommended references in the UV page.

Course notes

Université virtuelle

Other information

Place(s) of teaching

Plaine and Solbosch

Contact(s)

Part II: Big Data Analytics: Pr. G. Bontempi (ULB)

Evaluation method(s)

written examination and Project

Evaluation method(s) (additional information)

Combination of written exam and project work.

Determination of the mark (including the weighting of partial marks)

Combination of written exam (10/20) and project work (10/20)

Main language(s) of evaluation

English

Other language(s) of evaluation, if applicable

French

Programmes

Programmes proposing this course at the Brussels School of Engineering

MA-IRCB | Master of science in Biomedical Engineering | finalité Professional/unit 2, MA-IRIF | Master of science in Computer Science and Engineering | finalité Professional/unit 2 and MS-BGDA | Specialized Master in data science, Big data | unit U

Programmes proposing this course at the Solvay Brussels School of Economics and Management

MS-BGDA | Specialized Master in data science, Big data | unit U

Programmes proposing this course at the faculty of Sciences

MA-BINF | Master in Bio-informatics and Modelling | finalité Research/unit 2, MA-INFO | Master in Computer science | finalité Professional/unit 1 and finalité Professional/unit 2 and MS-BGDA | Specialized Master in data science, Big data | unit U