INFO-H515: BIG DATA: DISTRIBUTED DATA MANAGEMENT AND SCALABLE ANALYTICS

Preliminaries

Dimitris Sacharidis, Gianluca Bontempi

2024–2025

LECTURE OUTLINE

General Course Information

GENERAL COURSE INFORMATION

WHAT IS THIS COURSE ABOUT ?

Objective:

Introduce 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.

Key questions:

- what is “big data”, what are the characteristics of such data ?
- what is a “compute cluster”?
- how do clusters store data ?
- how are they programmed ?
- what are notions of efficiency for distributed algorithms ?
- what is “big data analytics” ?
- how do you perform machine learning on big data ?
- ...

COMPETENCES TO DEVELOP

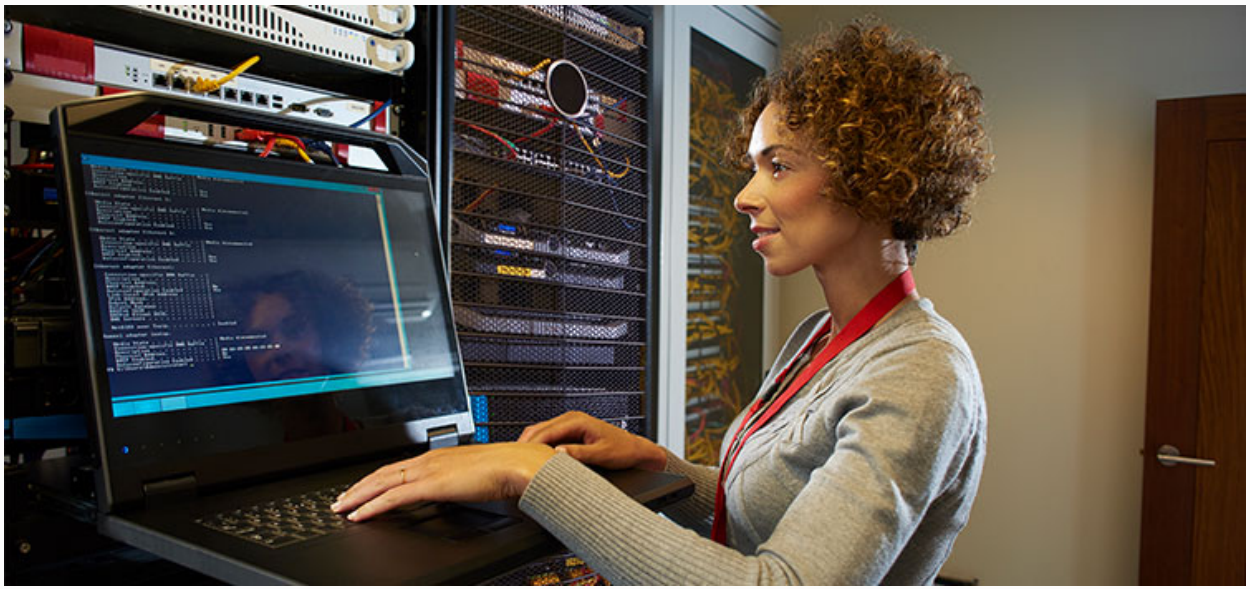
After successful completion of this course you should be able to:

1. understand the characteristics of big data, and the challenges these represent;
2. know the principal architectures of Big Data Management and Analytics Systems, be able to explain the purpose of each their components, and be able to recognize and explain the key properties, strengths and limitations of each type of system and their components;
3. understand the key bottlenecks in managing and analyzing massive amounts of data and be familiar with modern algorithms for overcoming these bottlenecks using parallel and distributed computation;
4. 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. build aplications using specific instances of each type of BDMAS.

WHAT THIS COURSE IS NOT

1. A course on how to build compute clusters

WHAT THIS COURSE IS NOT



1. A course on how to build compute clusters
2. A course on how to install big data frameworks

WHAT THIS COURSE IS NOT

1. A course on how to build compute clusters
2. A course on how to install big data frameworks
3. An exhaustive and detailed look into all possible big data frameworks that currently exist.

INFO-H515: 2 PARTS

Part 1 — Distributed Management

- Lectures by prof. D. Sacharidis (ULB)
- dimitris.sacharidis@ulb.be

Part 2 — Scalable Analytics

- Lectures by prof. G. Bontempi (ULB)
- gianluca.bontempi@ulb.be

PREREQUISITES

Required

- Good programming skills
- Introductory course on data management
- Introductory course on algorithms and data structures
- Introductory course on Machine Learning:
  - At ULB: INFO-F422 Statistical foundations of machine learning
  - At VUB: 1002080CNR Machine Learning, or 4004728DNR Techniques of Artificial Intelligence

INFO-H515 ORGANIZATION

The course is organized as a mixture of:

- Lectures
- Reading assignments
- Project work

INFO-H515 ORGANIZATION

Part 1: Distributed Management

Theory

Fridays 10h-12h

14 Feb – 21 Mar

ULB Solbosch

S.UB4.136, S.DC2.206

Exercises

Wednesdays 10h-12h

26 Feb, 5 Mar, 12 Mar, 19 Mar

ULB Solbosch TBA

Part 2: Scalable Analytics

Theory

Fridays 10h-12h

4 Apr – 23 May

ULB Plaine P.FORUM

Exercises

Wednesdays 10h-12h

TBA

Check UV and ULB Schedules for schedule and room updates

`https://www.ulb.be/en/schedules`

INFO-H515 SYLLABUS

The syllabus, available at the Virtual University consists of:

- Slides
- Associated reading assignments

Course material, exercises, reading assignments are all published on the ULB Virtual University

`https://uv.ulb.ac.be/course/view.php?id=124821`

Check regularly for updates!

INFO-H515 EVALUATION

- Project (60% of final score)
- Written exam (40% of final score)