

# Sistemi e Architetture per Big Data

## A.A. 2023/24

Valeria Cardellini, Matteo Nardelli

Laurea Magistrale in  
Ingegneria Informatica

### Teaching staff

---

- Valeria Cardellini
  - 4 CFU
  - Tel: 06 72597388, office: Ing. Informazione, room D1-17
  - Email: [cardellini@ing.uniroma2.it](mailto:cardellini@ing.uniroma2.it)
  - [www.ce.uniroma2.it/~valeria](http://www.ce.uniroma2.it/~valeria)
- Matteo Nardelli
  - 2 CFU
  - Email: [nardelli@ing.uniroma2.it](mailto:nardelli@ing.uniroma2.it)
  - [www.matteonardelli.it](http://www.matteonardelli.it)
- Email: use [SABD] in the subject line
- Office hours:
  - When: after lesson (in presence) or by appointment (either in presence or on Teams)

# General information

---

- Course web site  
[www.ce.uniroma2.it/courses/sabd2324](http://www.ce.uniroma2.it/courses/sabd2324)
- Virtual class on Teams
- Number of credits: 6 CFU
  - 60 hours of lessons (each lesson of 105 minutes)
- Class period: 2nd semester
  - From 4/3/2024 to 15/6/2024
- Class schedule
  - Monday 11:30-13:15, room C5
  - Thursday 11:30-13:15, room B8

📧 Please [register on Delphi](#) to join the course

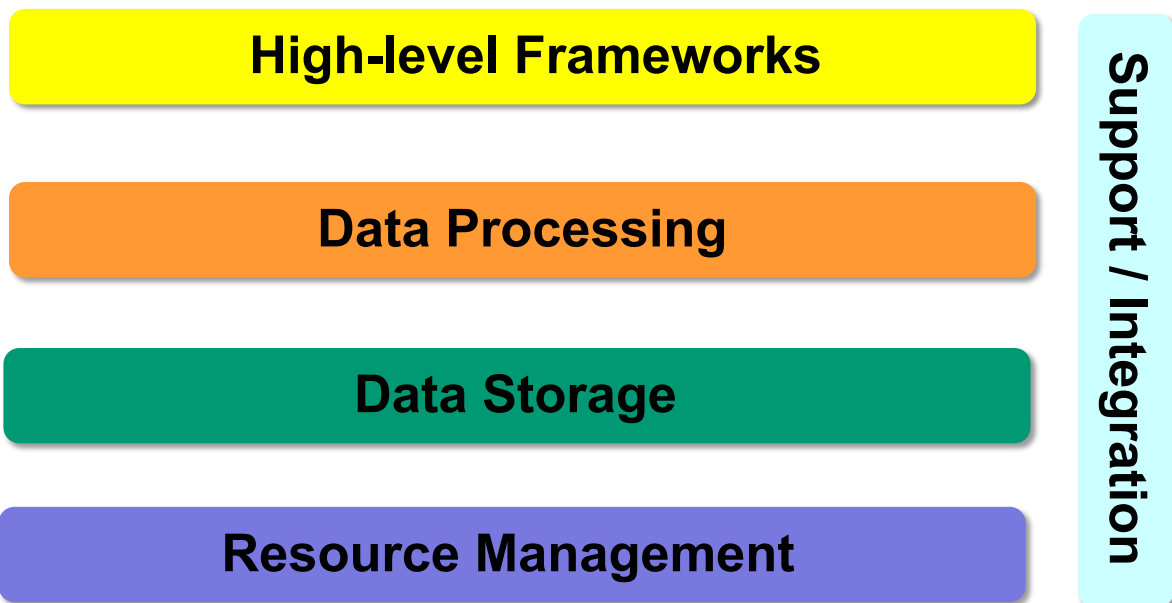
# Educational objectives

---

- Principles, paradigms, tools and technologies to design and manage distributed **systems** and **architectures** for **big data analytics** services and applications

# The Big Data stack we will consider

---



## Course program at-a-glance

---

- Frameworks for **resource management**
- Systems and frameworks for storing data either temporary or permanently, including distributed file systems and non-relational (NoSQL) databases for **data storage**
- Frameworks and tools for **collecting and ingesting data** from various sources into the big data analytics infrastructure
- **Processing** frameworks for *batch* and *real-time* analytics, including their architectural and programming aspects
- **High-level** frameworks and tools for **large scale** analytics, including *distributed ML*

## Course program in details

---

- Introduction to Big Data: issues and challenges
- Data storage: distributed file systems and NoSQL data stores
  - Case studies: HDFS, Cassandra, Dynamo, HBase, MongoDB, Neo4j
  - Hands-on: HDFS and NoSQL databases (Redis, MongoDB, HBase and Neo4j)
- Systems for batch processing
  - Case studies: Hadoop, Spark
  - Hands-on: Spark and Spark SQL
- Systems for data acquisition: pub/sub, message queues, collection systems
  - Case studies: Kafka, Nifi, Flume
  - Hands-on: Kafka

## Course program in details (2)

---

- Systems for stream processing
  - Case studies: Storm, Flink, Spark Streaming
  - Hands-on: Flink, Kafka Streams and Spark Streaming
- Frameworks for distributed machine learning and federated learning
  - Case study: Spark MLlib
- Frameworks for cluster resource management
  - Case study: Mesos
- Where data processing occurs?
  - In the Cloud
  - At the network edges

# Teaching material

---

- Your notes
- Lesson slides on web site and Teams
- Scientific papers, articles, etc. on web site
- Suggested textbooks:



- A. Bahga, V. Madiseti, [Cloud Computing Solutions Architect - A Hands-On Approach](#), 2019.



- M. Kleppman, [Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems](#), O'Reilly, 2017.

## Exam

---

a) 2 programming projects assigned during the course

- [Programming project #1](#): assigned at the end of April 2024, due at the end of May 2024
- [Programming project #2](#): assigned at the end of May 2024, due at the end of June 2024
- Possibly in groups of 2

b) [Final oral exam](#) on the course program

- When:
  - 2 dates in each exam period (June-July 2024, September 2024 and January-February 2025)

# Grading

---

- Programming project #1: 35%
- Programming project #2: 35%
- Final oral exam: 30%
  
- Participation during class will also be taken into account