TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Macroarea di Ingegneria
Dipartimento di Ingegneria Civile e Ingegneria Informatica

Sistemi e Architetture per Big Data

A.A. 2025/26

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
  - http://www.ce.uniroma2.it/~valeria
- Matteo Nardelli
  - 2 CFU
  - Email: nardelli@ing.uniroma2.it
  - https://www.matteonardelli.it/

- Email: use [SABD] in the subject line
- Office hours:
  - When: after lesson (in presence) or by appointment

General information

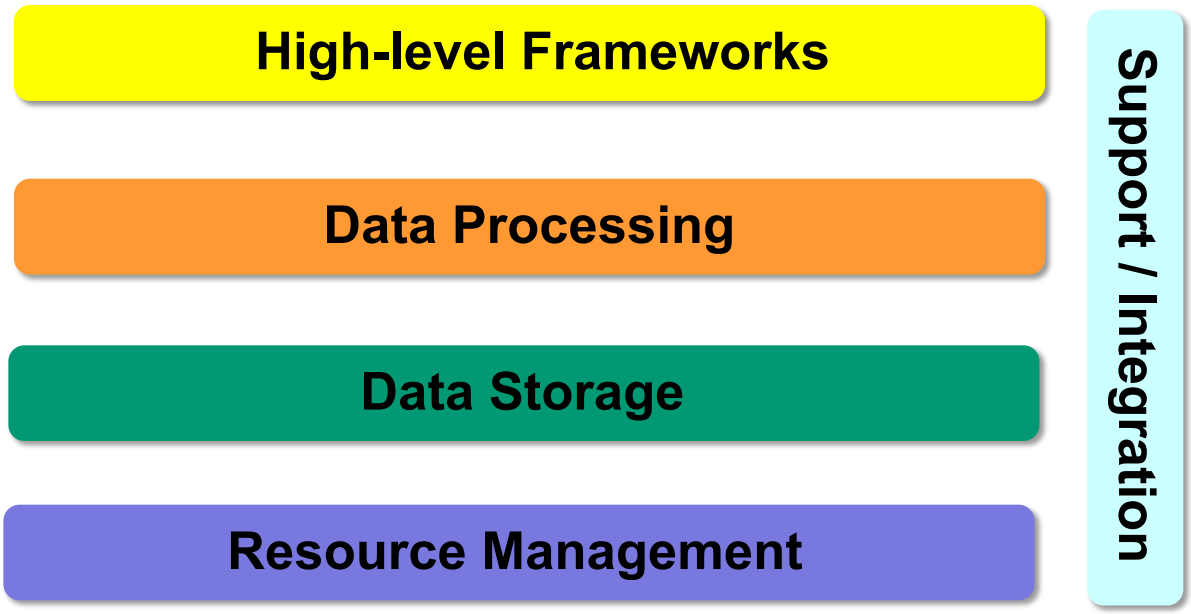
- Course web site
  http://www.ce.uniroma2.it/courses/sabd2526
- Virtual class on Teams
- Number of credits: 6 CFU
  - 60 hours of lectures (each lecture lasts 105 minutes)
- Class period: 2nd semester
  - From 2/3/2026 to 11/6/2026
- Class schedule
  - Monday 11:30-13:15, room C5
  - Thursday 11:30-13:15, room B8

☞ Please register on Delphi to join course

Educational objectives

- Principles, paradigms, tools, and technologies for designing and managing 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 **data storage**, both temporary and permanent, including distributed file systems and object stores, NoSQL data stores, and vector databases
- Frameworks and tools for **data collection and ingestion** from multiple sources into a 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 object stores, NoSQL data stores, and vector databases
  - Case studies: GFS, HDFS, Cassandra, Dynamo, DynamoDB, Bigtable, HBase, MongoDB, Neo4j
  - Hands-on: HDFS and NoSQL data stores (Redis, MongoDB, Hbase, Neo4j, InfluxDB)
- Systems and models for batch processing
  - Case study: Spark
  - Hands-on: Spark and Spark SQL
- Systems for data acquisition: pub/sub, message queues, and data collection systems
  - Case studies: Kafka (recall), Nifi, Airflow
  - Data formats and serialization: Parquet, Avro, ORC

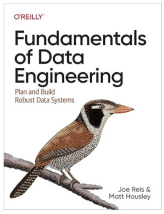
Course program in details

- Data pipeline architectures and ETL/ELT workflows
- Systems and models for stream processing
  - Case studies: Flink, Spark Streaming
  - Hands-on: Flink, Kafka Streams, Spark Streaming
- Frameworks for distributed machine learning and federated learning
  - Case study: Spark MLlib
- Frameworks for resource management
- Where does data processing occur?
  - In the Cloud
  - At the network edges

Teaching material

- Your lecture notes
- Lesson slides (web site and Teams)
- Scientific papers, videos, etc. (web site)
- Code and other material (web site Teams)
- Suggested textbooks:

M. Kleppman, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly, 2017. https://dataintensive.net



J. Reis, M. Housley, Fundamentals of Data Engineering: Plan and Build Robust Data Systems 1st Edition, O'Reilly, 2022. https://www.oreilly.com/library/view/fundamentals-of-data/9781098108298/

Exam

a) 2 programming projects assigned during the course
   - Programming project #1: assigned at the beginning of May 2026, due by the end of May 2026
   - Programming project #2: assigned at the beginning of June 2026, due at the end of June 2026
   - Possibly in groups of 2

b) Final oral exam on the course program
   - When:
     - 2 dates in each exam period (June-July 2026, September 2026 and January-February 2027)

Grading

- Programming project #1: 35%
- Programming project #2: 35%
- Final oral exam: 30%

- Class participation will also be taken into account