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. 2021/22

Valeria Cardellini, Matteo Nardelli

Laurea Magistrale in
Ingegneria Informatica

Teaching staff

- Valeria Cardellini
  - Tel: 06 72597388, office: Ing. Informazione, room D1-17
  - Email: cardellini@ing.uniroma2.it
  - http://www.ce.uniroma2.it/~valeria/
- Matteo Nardelli
  - Supplementary course “Hands-on storage systems and processing frameworks for Big Data”
  - Email: nardelli@ing.uniroma2.it
  - http://www.ce.uniroma2.it/~nardelli
- Email: use [SABD] in the subject line
- Office hours:
  - When: after lesson (in presence) or by appointment
  - Where: on Teams

General information

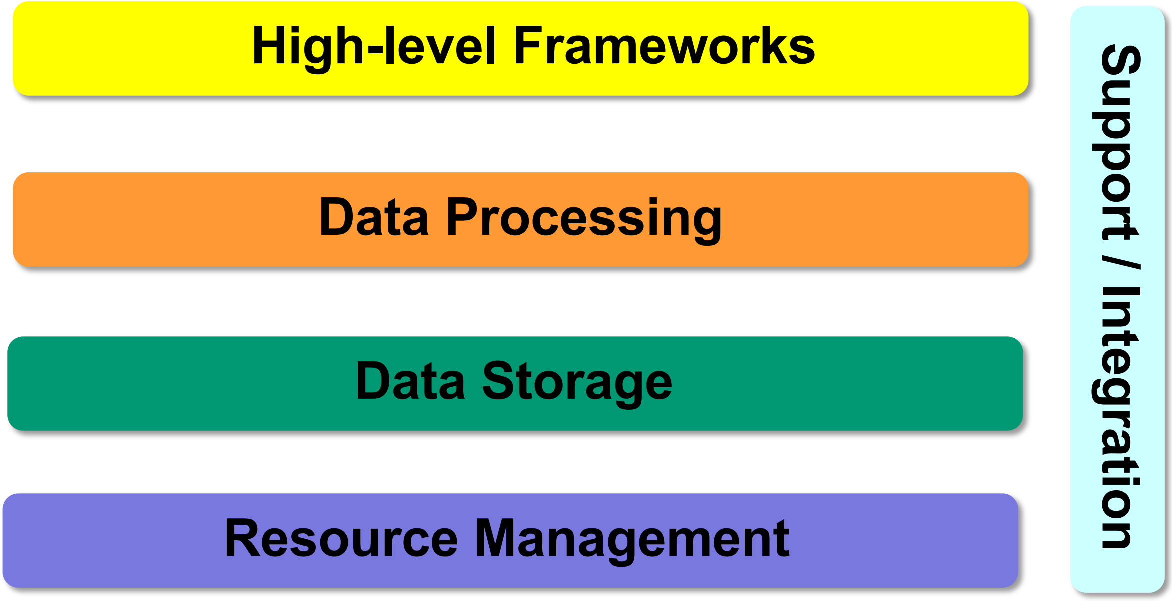
- Web site of the course
  http://www.ce.uniroma2.it/courses/sabd2122/
- Virtual class on Teams
- Number of credits: 6 CFU
  - 60 hours of lessons (each lesson of 105 minutes)
- Class period: 2nd semester
  - From 28/2/2022 to 9/6/2022
- Class schedule
  - Monday 11:30-13:15, room C5
  - Thursday 11:30-13:15, 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



High-level Frameworks

Data Processing

Data Storage

Resource Management

Support / Integration

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

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, Pig, Hive, Spark
  - Batch processing in the Cloud
  - Hands-on: Hadoop, Spark and Spark SQL
- Systems for data acquisition
  - Pub/sub, message queues, collection systems
  - Hands-on: Kafka

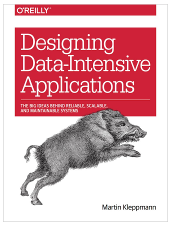
Course program in details (2)

- Systems for stream processing
  - Case studies: Storm, Flink, Heron, Samza, Spark Streaming
  - Stream processing in the Cloud
  - Hands-on: Kafka Streams and Spark Streaming
- Frameworks for distributed machine learning
- Frameworks for cluster resource management
  - Case studies: Mesos, YARN
- The new reference infrastructure: edge/fog computing

Teaching material

- Your notes
- Lesson slides on the course web site (after the lesson!)
- Scientific papers, articles, etc. on the course web site
- Suggested textbooks:

- A. Bahga, V. Madisetti, 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 2022, due at the end of May 2022
- Programming project #2: assigned at the end of May 2022, due at the end of June 2022
- Possibly in groups of 2

b) Final oral exam on the course program
- When:
  - 2 dates in each exam period (June-July 2022, September 2022 and January-February 2023)

Grading

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

- Participation during class will also be taken into account