Distributed architectures for big data processing and analytics

Teachers

- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Simone Papicchio
  - simone.papicchio@polito.it

Office hours

- Class-time (break, end of lesson)
- Or send and e-mail for an appointment

Weekly schedule

- Lectures (62 hours)
  - Monday 8:30-10:00
    - Classroom R3+ Virtual Classroom
  - Wednesday 10:00-13:00
    - Classroom 10A+ Virtual Classroom
- Lab activities (18 hours)
  - Team 1: Students from A to D – Tuesday from 11:30 to 13:00 (First lab activity – March 3, 2026) @ LABINF
  - Team 2: Students from E to M – Friday from 11:30 to 13:00 (First lab activity – March 6, 2026) @ LABINF
  - Team 3: Students from N to Z – Friday from 16:00 to 17:30 (First lab activity – March 6, 2026) @ LABINF

Lab activities

- LABINF account
  - https://www.labinf.polito.it/
  - You will receive an email from the LABINF's administrator with your credentials
  - Please make sure you have the account at LABINF before starting the lab activities
  - It is not the account you use to log into the PCs of the other labs at Politecnico di Torino
- **No lab activities during the first week**

Lab activities

- We will also provide you with a specific account on the BigData@Polito cluster
  - http://bigdata.polito.it/
- Detailed information will be provided before the first laboratory practice
  - You will receive an email with credentials and detailed information in the following days

Topics

- Lectures
  - Introduction to Big data
  - Big data pipeline and lambda architecture
  - Hadoop
    - Architecture
    - MapReduce programming paradigm
  - Spark
    - Architecture
    - Spark programs based on RDDs (Resilient Distributed Data sets)
    - Spark SQL and DataFrames

Topics

- Data mining and Machine learning libraries for Big Data
  - MLlib (Apache Spark's scalable machine learning library)
  - GraphX and GraphFrame (Apache Spark's API for graphs)
- Data streaming analytics
  - Spark Streaming
  - High level introduction to other frameworks (Apache Flink, Storm,Kafka, ..)

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

Prerequisites / prior knowledge

- Programming skills
  - **Java language (basic)**
  - **Python language**
- and basic knowledge of database concepts
  - Relational data model
  - SQL language

Material

- Web page
  - https://dbdmg.polito.it/dbdmg_web/2026/distributed-architectures-for-big-data-processing-and-analytics-2025-2026/
  - Slides, exercises, tools
- Video lectures/Virtual classrooms
  - The video lectures will be available on the Teaching portal
    - https://didattica.polito.it

Books and Readings

- Reference books:
  - Matei Zaharia, Bill Chambers. Spark: The Definitive Guide (Big Data Processing Made Simple). O'Reilly Media, 2018.
  - Advanced Analytics and Real-Time Data Processing in Apache Spark. Packt Publishing, 2018.
  - Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. Learning Spark (Lightning-Fast Big Data Analytics). O'Reilly, 2015.
  - Tom White. Hadoop, The Definitive Guide. (Third edition). O'Reilly Media, 2015.
  - Donald Miner, Adam Shook . "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems." O'Reilly, 2012

Exam rules

- Written exam
  - 2 programming exercises (27 points)
    - Design and develop programs based on the MapReduce programming paradigm and Spark APIs
  - 2 multiple-choice questions (4 points)
    - Topics
      - Technological characteristics and architecture of Hadoop and Spark
      - HDFS
      - MapReduce programming paradigm
      - Spark RDDs, transformations and actions
      - Spark SQL and DataFrames
      - Data mining and Machine learning libraries for Big data (Spark MLlib, GraphX/GraphFrame)
      - Data streaming analytics

Exam rules

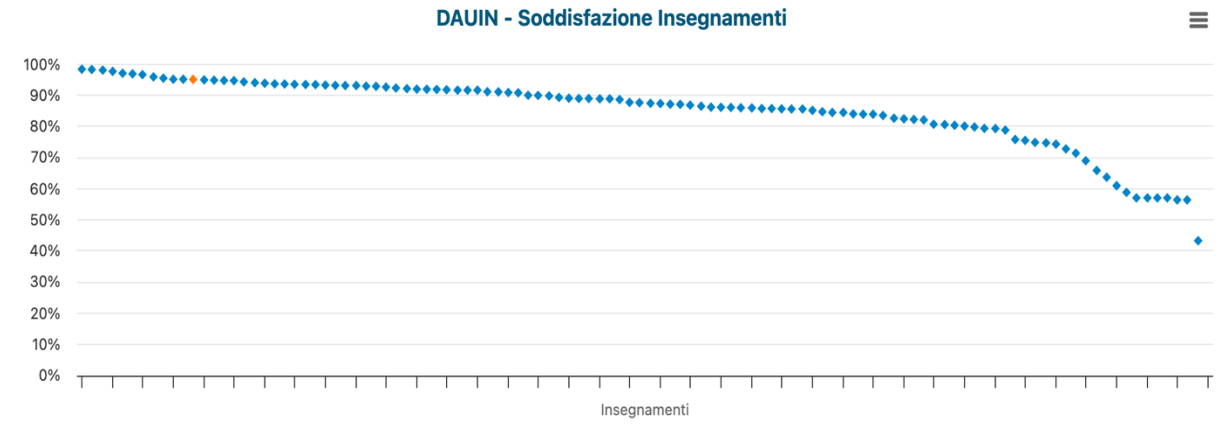
- On-site written exam on the **Exam/Moodle platform with Lockdown browser** – **You must bring your own PC**
  - 90 minutes
  - The exam is **open book**
    - Books, notes, handwritten notes, and any other paper materials are allowed
    - Electronic devices of any kind (PC, mobile phone, calculators, etc.) are not allowed, except the PC used for the exam itself
- Exam examples will be available on the web page of the course

A.Y. 2024/2025 - End-of-course Questionnaire

Elaborazione dati

Indice docente (indice medio delle risposte per Efficacia del/della docente)	3.69	**Percentuale di soddisfazione docente** (% delle risposte positive per Efficacia del/della docente)	96.40 %
Indice insegnamento (indice medio delle risposte Pt.2)	3.64	**Percentuale di soddisfazione insegnamento** (% risposte positive questionario Pt.2)	95.15 %
Tasso di compilazione (questionari compilati/studenti frequentanti)	0.74	**Tasso di risposta** (questionari compilati e schede bianche/studenti frequentanti)	0.82
Percentuale di compilazione (questionari compilati/studenti abilitati * 100)	73.75 %	**Percentuale di risposta** (questionari compilati e schede bianche/studenti abilitati * 100)	82.08 %

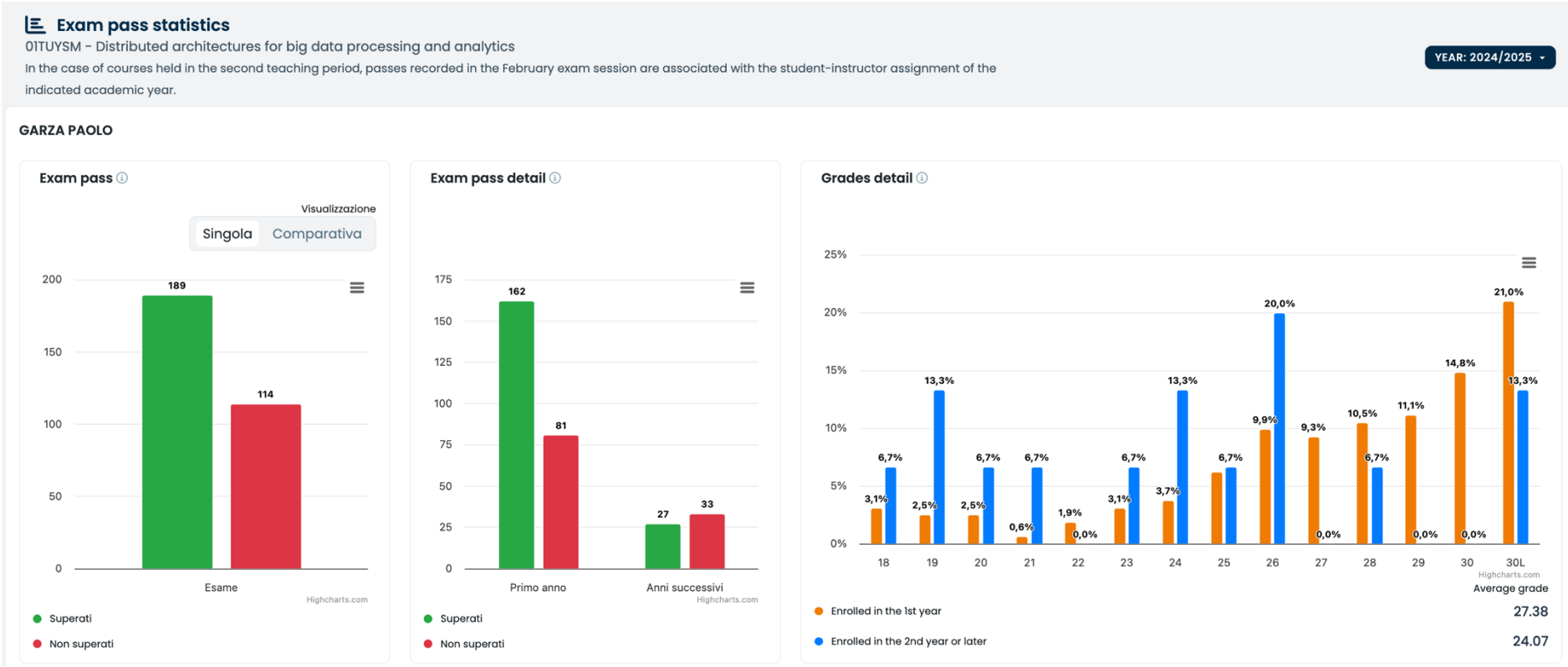
ⓘ I dati mostrati fanno riferimento a tutti gli insegnamenti afferenti ai Corsi di Studio incardinati in uno specifico dipartimento



- Relevant critiques*
  - The labs are crowded and just a few assistants.
  - The professor sometimes speaks too quickly during lectures, making it challenging to keep up with the material.
  - Some advanced topics (e.g., Spark Streaming and MLlib) receive less attention from students, partly because they are perceived as having lower weight in the exam.

**considered in planning this year's course*

Exam statistics(A.Y. 2024/25)



https://didattica.polito.it/pls/static/esami/statistiche/?p_cod_ins=01TUYSM&p_a_acc=2025