

Big Data Processing and Analytics

Teachers

- Paolo Garza
 - paolo.garza@polito.it
 - 011-090-7022
- Luca Colomba

Office hours

- Class-time (break, end of lesson)
- Or send an e-mail for an appointment
- Or Piazza for Q&A offline:

<http://piazza.com/polito.it/fall2022/o1dshov>

Weekly schedule

- Lectures (45 hours)
 - Tuesday - 10:00-11:30 - Room 3P
 - Wednesday - 16:00-19:00 - Room R3B
- Practices (15 hours)
 - Friday - 8:30-10:00 - LAIB5 - Team 1 (A-L)
 - Friday - 10:00-11:30 - LAIB5 - Team 2 (M-Z)
 - There are **no lab activities** during the **first week**

Practices

- We will provide you a specific account on the BigData@Polito cluster
 - <https://jupyter.polito.it>
 - <https://hue.polito.it>
- Detailed information will be provided next week
 - You will receive an email from the administrator of the cluster with username and password

Topics

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

Topics

- Data mining and Machine learning libraries for Big Data
 - **MLlib** (Apache Spark's scalable machine learning library)
- Streaming data analysis
 - **Spark Streaming**
- SQL databases for relational big data and NoSQL databases
 - Data models, Design, Querying

Topics

- Laboratory activities
 - Application development on Hadoop and Spark

Prerequisites / prior knowledge

- Object-oriented programming skills
 - **Java language (mandatory)**
- Basic knowledge of traditional database concepts (recommended)
 - Relational data model
 - SQL language

Material

- Web page
 - https://dbdmg.polito.it/dbdmg_web/index.php/2022/09/21/big-data-processing-and-analytics-2022-23/
 - Slides, exercises, lab activities, ..
- Video lectures/Virtual classrooms
 - 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 (max 27 points)
 - Design and develop Java programs based on the Hadoop MapReduce programming paradigm and/or Spark RDDs
 - 2 questions/theoretical exercises (max 4 points)
 - Topics
 - Technological characteristics and architecture of Hadoop and Spark
 - HDFS
 - MapReduce programming paradigm
 - Spark RDDs, transformations and actions
 - Spark SQL
 - Spark Streaming
 - Spark MLlib
 - NoSQL databases and data models for big data

Exam rules

- On-site written exam on the **Exam platform with Lockdown browser** – **You must bring your own PC**
 - 90 minutes
 - The exam is **open book**
 - Books, notes, and any other paper material 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