Big data: architectures and data analytics

Teachers

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

Office hours

- Class-time (break, end of lesson)
- Or send and e-mail for an appointment
- Or Piazza for Q&A offline: https://piazza.com/polito.it/fall2021/01qydov

Weekly schedule

- Lectures (45 hours)
  - Monday 16:00-17:30
    - Blended lecture – On-site (Room R1) + Online virtual classroom
  - Tuesday 10:00-13:00
    - Blended lecture – On-site (Room R1) + Online virtual classroom
- Practices (15 hours)
  - Monday 17:30-19:00 Team 1 (A-L)
    - Blended lab – On-site (LAIB1) + Online virtual classroom
  - Wednesday 14:30-16:00 Team 2 (M-Z)
    - On-site (LAIB1)
  - No lab activities during the first two weeks

Practices

- We will provide you a specific account on the BigData@Polito cluster
    - http://bigdata.polito.it/
- Detailed information will be provided next week
    - You will receive an email from the admin 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 (e.g., Hive) and NoSQL databases (e.g., HBASE)
  - Data models, Design, Querying

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

Prerequisites / prior knowledge

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

Material

- Web page
  - https://dbdmg.polito.it/dbdmg_web/index.php/2021/09/16/big-data-architectures-and-data-analytics-2021-2022
  - 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 (or Exams + Respondus for those who cannot be at Polito)
  - 2 hours
  - The exam is **closed book**
    - Books, notes, and any other paper material are not allowed.
    - Electronic devices of any kind (PC, laptop mobile phone, calculators, etc.) are not allowed.
- Past exams are available on the web page of the course