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 14+ Virtual Classroom
  - Wednesday 10:00-13:00
    - Classroom R4+ Virtual Classroom
- Lab activities (18 hours)
  - Tuesday 11:30-13:00 – Team #1 – A-L
    - LABINF
  - Friday 11:30-13:00 – Team #2 – M-Z
    - LABINF

Lab activities

- Please make sure you have a specific 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
  - You can register an account at LABINF every day from 2pm to 3pm (check the LABINF website for further details)
  - http://www.labinf.polito.it

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
  - We will send you an email with username and password

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
  - Apache Flink, Storm,Kafka, ..

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

Prerequisites / prior knowledge

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

Material

- Web page
  - https://dbdmg.polito.it/dbdmg_web/2024/distributed-architectures-for-big-data-processing-and-analytics-2023-2024/
  - 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 (max 27 points)
    - Design and develop programs based on the MapReduce programming paradigm and Spark APIs
  - 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 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 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