Big data for internet applications

Teachers

- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Luca Vassio
  - luca.vassio@polito.it
- Matteo Boffa
  - matteo.boffa@polito.it

Office hours

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

Weekly schedule

- Lectures (42 hours)
  - Wednesday 10:00-11:30
    - Sala C
  - Thursday 13:00-16:00
    - Room 17

Weekly schedule

- Practices (18 hours)
  - Thursday 13:00-16:00
    - Room 17
    - We will inform you in which dates this slot will be used for the practice/lab activities
    - There are **no lab activities** for the **first two weeks**
  - **You must bring your own PC for the practice/lab activities**

Practices

- We will provide you a specific account on the BigData@Polito cluster
  - http://bigdatalab.polito.it/
- Detailed information will be provided next week
  - You will receive an email with username and password. It will be sent by the administrator of the BigData@Polito cluster

Topics

- Lectures
  - Introduction to Big data
  - Hadoop
    - Infrastructure and basic components
  - Spark
    - Architecture
    - Spark programs based on RDDs (Resilient Distributed Data sets) and DataFrames

Topics

- Data mining and Machine learning libraries for Big Data
  - MLlib (Apache Spark's scalable machine learning library)
- Streaming data analysis
  - Spark Streaming
- Graph analysis
  - Spark GraphX
- Databases for big data
  - Data models, Design, Querying

Topics

- Laboratory activities
  - Development of Spark-based applications for analyzing data
  - Programming language: Python

Prerequisites/prior knowledge

- Basic object-oriented programming skills
  - We will use **Python**

Materials

- Teaching portal
  - News about the course
  - Slides, exercises, etc

Exam rules

- Written exam
  - 31 points
- Individual report
  - 31 points

Exam rules

- Final grade
  - Grade of the written exam*0.7 + Grade of the report*0.3
  - The exam is passed if
    - (i) Grade of the written exam >= 18 and
    - (ii) Grade of the individual report >=18

Exam rules

- On-site written exam on the Exam platform with Lockdown browser - **You must bring your own PC**
  - 2 hours
  - 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 rules

- Written exam
  - 2 programming exercises (max 27 points)
    - Design and develop of Python programs based on Spark
  - 2 questions/theoretical exercises (max 4 points)
    - Topics
      - Technological characteristics and architecture of Hadoop and Spark
      - Spark-based programming (RDDs, Datasets, transformations and actions)
      - Spark streaming, Mllib, GraphX
      - Databases for Big data and data models

Exam rules

- Individual reports on the practices assigned during the course and developed in laboratories
  - To be delivered **10 days before the written exam**
  - **One report for each lab**
  - The reports are **valid** for the **entire academic year**
  - **If the reports are sufficient, you cannot resubmit them**