Big data: architectures and data analytics

Teachers
- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Alessandro Farasin
- Francesco Ventura
- Andrea Pasini
- Marilisa Montemurro

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

Practices
- Please make sure you have a specific account at LABINF before starting the lab practices
  - 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

Practices (2)
- We will also provide you a specific account on the BigData@Polito cluster
  - http://bigdata.polito.it/
  - This account is different from the LABINF one
  - Detailed information will be provided next week
  - I will send you an email 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)
Topics

- Data mining and Machine learning libraries for Big Data
  - MLib (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

Prerequisites / prior knowledge

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

Books and Readings

- Reference books:

Materials

- Web page
  - News about the course
  - Slides, exercises, tools
- Video lectures
  - The course is video recorded and the video lectures are available on the Teaching portal
    - https://didattica.polito.it

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
      - NoSQL databases and data models
Exam rules

- Written exam
  - 2 hours
- Open book exam
  - Paper books and paper notes are allowed
  - Instead, no electronic devices (PC, laptop mobile phone, calculators, etc.) are allowed