Big data: architectures and data analytics

Teachers

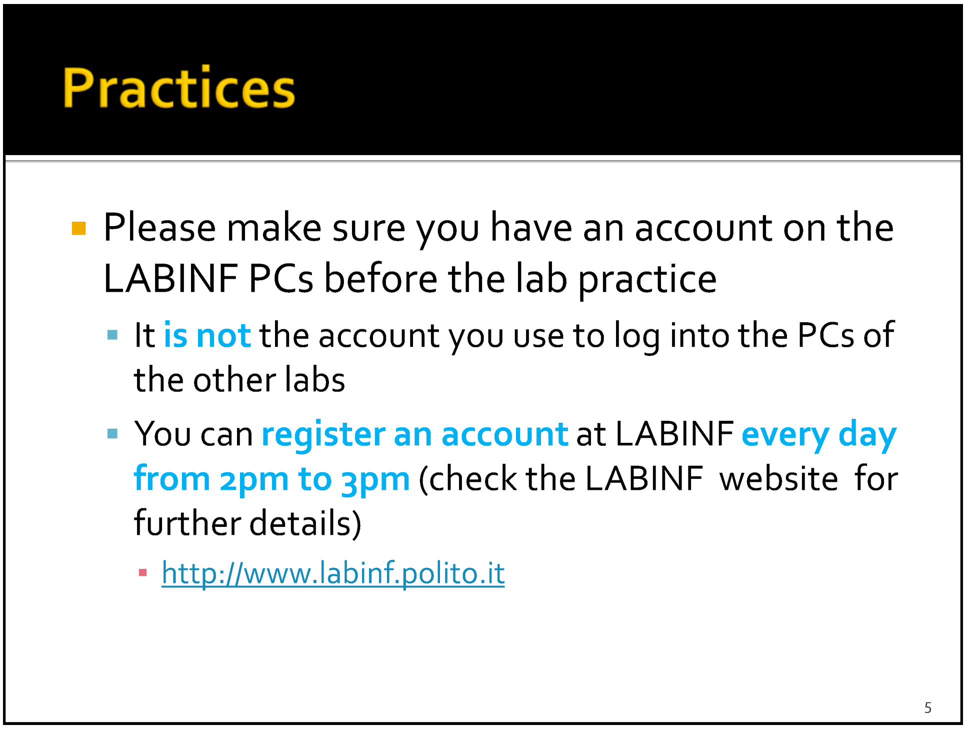
- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Luca Venturini
  - luca.venturini@polito.it
  - 011-090-7084

Office hours

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

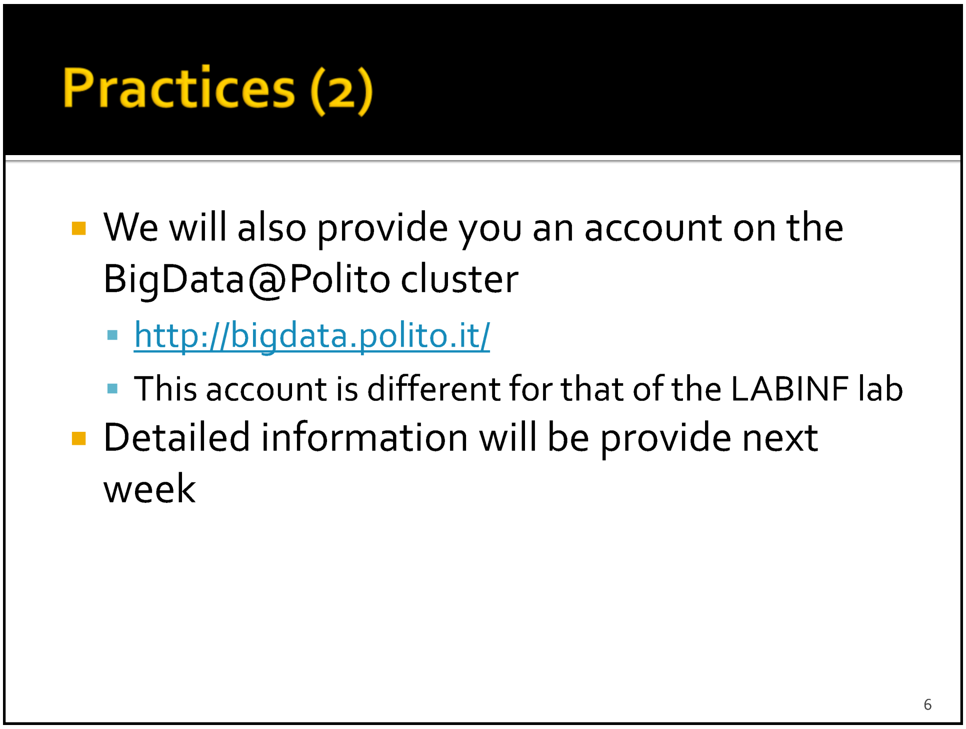
Weekly schedule

- Lectures
  - Thursday 13:00-16:00
    - Classroom 12
  - Friday 16:00-17:30
    - Classroom 12
- Practices
  - Tuesday 17:30-19:00 Group 1
  - Thursday 17:30-19:00 Group 2
    - LABINF
    - The first lab practice will be on Tuesday, March 21, 2017 at 17:30



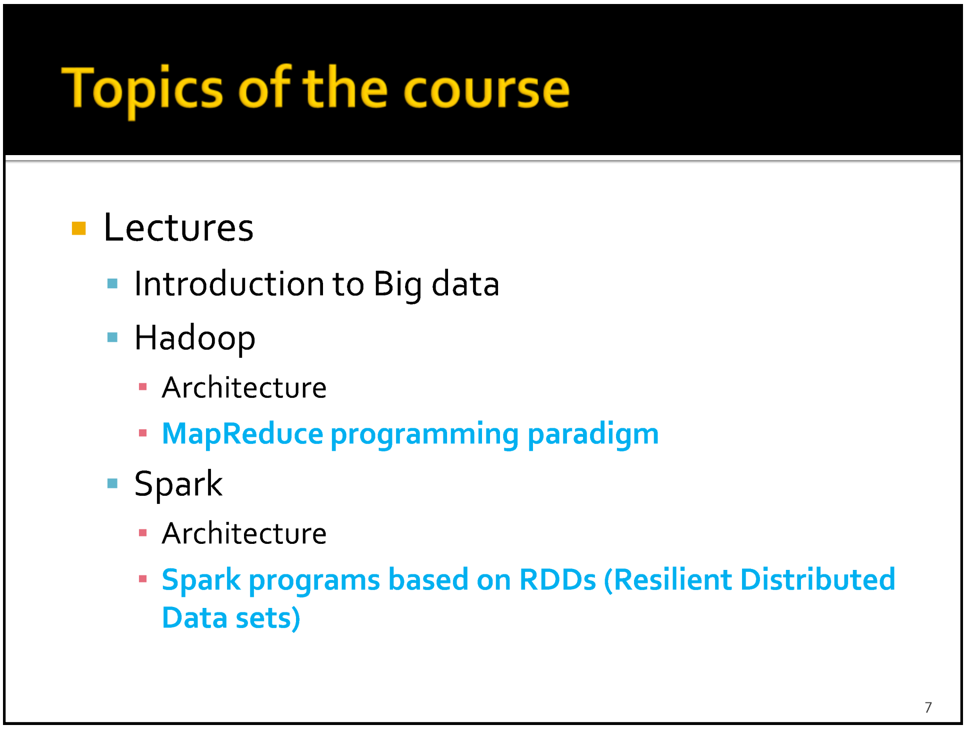
Practices

- Please make sure you have an account on the LABINF PCs before the lab practice
  - It **is not** the account you use to log into the PCs of the other labs
  - 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



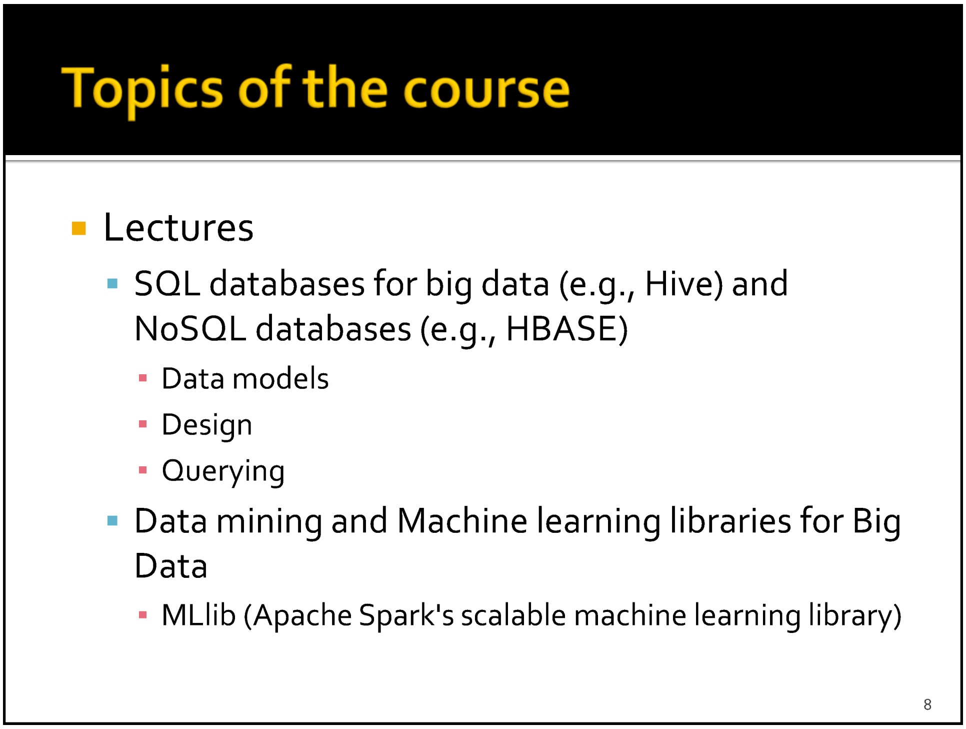
Practices (2)

- We will also provide you an account on the BigData@Polito cluster
  - http://bigdata.polito.it/
  - This account is different for that of the LABINF lab
- Detailed information will be provide next week



Topics of the course

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



Topics of the course

- Lectures
  - SQL databases for big data (e.g., Hive) and NoSQL databases (e.g., HBASE)
    - Data models
    - Design
    - Querying
  - Data mining and Machine learning libraries for Big Data
    - MLlib (Apache Spark's scalable machine learning library)

Topics of the course

- Laboratory activities
  - Developing of applications by means of Hadoop, Spark

Prerequisites/Assumed knowledge

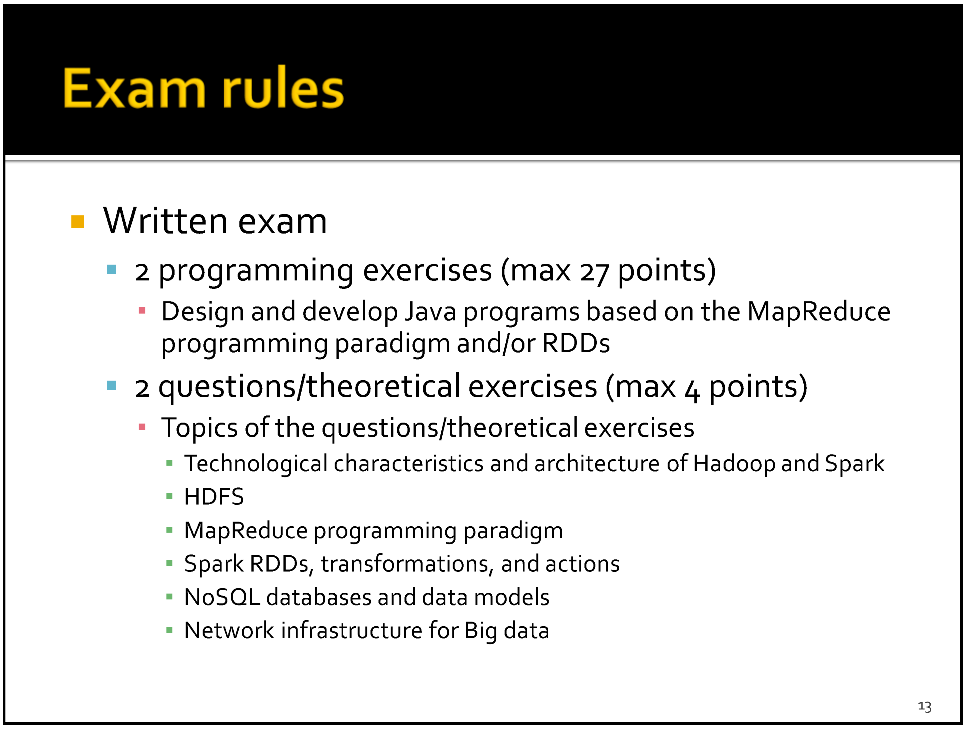
- Basic object-oriented programming skills
  - **Java language (mandatory)**
- and basic knowledge of traditional database concepts (recommended)
  - Relational data model
  - SQL language

Course materials

- Course web site
  - http://dbdmg.polito.it/wordpress/teaching/big-data-architectures-and-data-analytics
  - News about the course
  - Slides, exercises, tools
- Video lectures
  - Available on the Teaching portal
    - https://didattica.polito.it

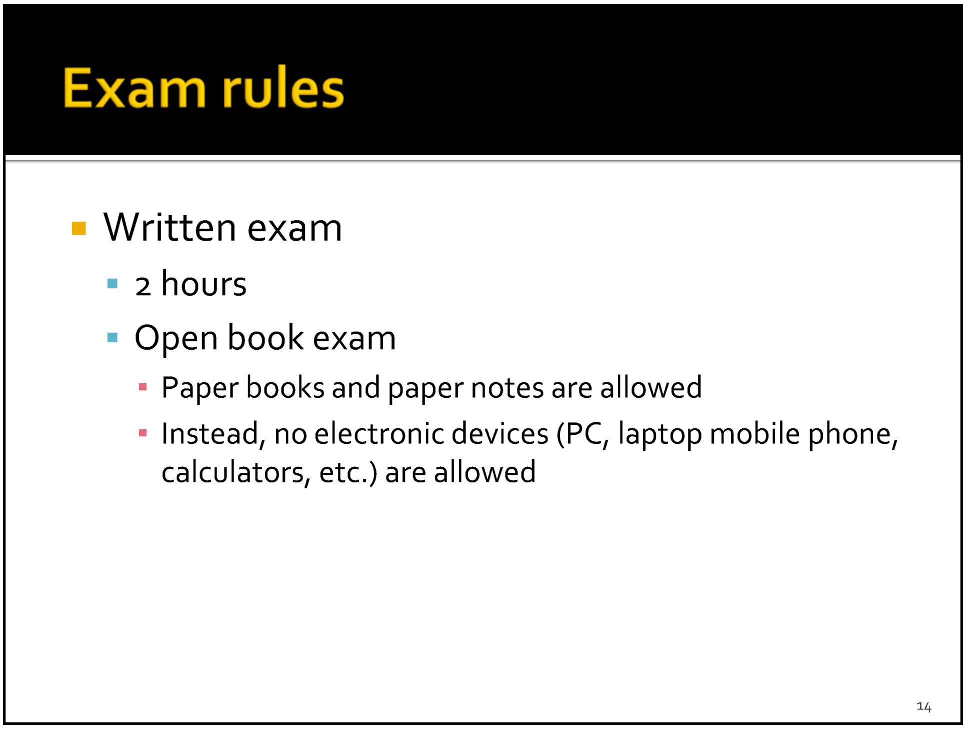
Books and Readings

- Reference books:
  - Tom White. "Hadoop, The Definitive Guide." (Third edition). O'Reilly, 2012.
  - Donald Miner, Adam Shook . "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems." O'Reilly, 2012
  - Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. "Learning Spark: Lightning-Fast Big Data Analytics." O'Reilly, 2015.
  - Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. "Advanced Analytics with Spark." O'Reilly, 2014.



Exam rules

- Written exam
  - 2 programming exercises (max 27 points)
    - Design and develop Java programs based on the MapReduce programming paradigm and/or RDDs
  - 2 questions/theoretical exercises (max 4 points)
    - Topics of the questions/theoretical exercises
      - Technological characteristics and architecture of Hadoop and Spark
      - HDFS
      - MapReduce programming paradigm
      - Spark RDDs, transformations, and actions
      - NoSQL databases and data models
      - Network infrastructure for Big data



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