

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

2

Office hours

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

3

Weekly schedule

- Lectures
 - Monday 13:00-14:30
 - Classroom 5I
 - Thursday 11:30-14:30
 - Classroom 5I
- Practices
 - Thursday 17:30-19:00
 - LABINF
 - The first lab practice will be on Thursday, March 17, 2016 at 17:30

4

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>

5

Practices (2)

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

6

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)**

7

Topics of the course

- Lectures
 - 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)
 - Cloud computing platforms and Network infrastructures for Big Data applications

8

Topics of the course

- Laboratory activities
 - Developing of applications by means of Hadoop, Spark, and NoSQL databases

9

Prerequisites/Assumed knowledge

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

10

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>

11

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.

12

Exam rules

- Written exam
 - 1 or 2 programming exercises (max 27 points)
 - Design and develop Java programs based on the MapReduce programming paradigm and/or RDDs
 - 2 multiple choice questions (max 4 points)
 - Topics of the multiple choice questions
 - 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

33

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

34