Big data: architectures and data analytics

Teachers

- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Daniele Apiletti
- Francesco Ventura
- Alessandro Farasin
- Andrea Pasini
- Eliana Pastor

2

Office hours

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

3

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

4

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 account

5

Topics

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

6

Topics

- Data mining and Machine learning libraries for Big Data
  - **MLlib** (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

7

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

8

Prerequisites / prior knowledge

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

9

Materials

- Web page
  - http://dbdmg.polito.it/wordpress/teaching/big-data-architectures-and-data-analytics-2017-2018
  - News about the course
  - Slides, exercises, tools
- Video lectures
  - The video lectures of the A.Y. 2016/2017 are available on the Teaching portal
    - https://didattica.polito.it

10

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.

11

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

12

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

13