Distributed architectures for big data processing and analytics

Teachers

- Paolo Garza
  - paolo.garza@polito.it
  - 011-090-7022
- Martino Trevisan

2

Office hours

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

3

Weekly schedule

- Lectures (~62 hours)
  - Tuesday 11:30-14:30
  - Wednesday 13:00-14:30
  - Friday 11:30-13:00
    - Classroom 3I
- Practices (18 hours)
  - Friday 13:00-14:30
    - No lab activities during the first week

4

Practices

- We will also provide you a specific account on the BigData@Polito cluster
  - http://bigdata.polito.it/
- Detailed information will be provided before the first laboratory practice
  - We will send you an email with username and password

6

Topics

- Lectures
  - Introduction to Big data
  - Big data pipelines and lambda architecture
  - Hadoop
    - Architecture
    - MapReduce programming paradigm
  - Spark
    - Architecture
    - Spark programs based on RDDs (Resilient Distributed Data sets)
    - Spark SQL and DataFrames

7

Topics

- Data mining and Machine learning libraries for Big Data
  - MLlib (Apache Spark's scalable machine learning library)
  - GraphX and GraphFrame (Apache Spark's API for graphs)
- Data streaming analytics
  - Spark Streaming
  - Apache Flink, Storm,Kafka, ..
- Relational and NoSQL databases for big data
  - Hive (relational)
  - HBase (open-source, distributed, versioned, non-relational, NoSQL database)

8

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

9

Prerequisites / prior knowledge

- Programming skills **(mandatory)**
  - **Java language**
  - **Python language**
- and basic knowledge of basic database concepts (recommended)
  - Relational data model
  - NoSQL data models
  - SQL language

10

Material

- Web page
  - https://dbdmg.polito.it/wordpress/teaching/distributed-architectures-for-big-data-processing-and-analytics-2019-2020
  - News about the course
  - Slides, exercises, tools
- Video lectures
  - The video lectures are available on the Teaching portal
    - https://didattica.polito.it

11

Books and Readings

- Reference books:
  - Matei Zaharia, Bill Chambers. Spark: The Definitive Guide (Big Data Processing Made Simple). O'Reilly Media, 2018.
  - Advanced Analytics and Real-Time Data Processing in Apache Spark. Packt Publishing, 2018.
  - Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. Learning Spark (Lightning-Fast Big Data Analytics). O'Reilly, 2015.
  - Tom White. Hadoop, The Definitive Guide. (Third edition). O'Reilly Media, 2015.
  - Donald Miner, Adam Shook . "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems." O'Reilly, 2012

12

Exam rules

- Written exam
  - 2 programming exercises (max 27 points)
    - Design and develop programs based on the MapReduce programming paradigm and Spark APIs
  - 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
      - Spark SQL and DataFrames
      - Data mining and Machine learning libraries for Big data (Spark MLlib, GraphX/GraphFrame)
      - Data streaming analytics (Spark Streaming, Flink, Storm, Kafka, ..)
      - Relational and NoSQL databases for big data (HIVE, Hbase)

13

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

14