COURSE: Architectures, systems and algorithms for big data computing

INSTRUCTORS: Marco Bianchi; Simone Angelini
EMAIL: mbianchi@fub.it
sangelini@fub.it
WEB PAGE: http://www.fub.it/ricercatori/marco-bianchi

COURSE DESCRIPTION
The main goal of the course is to introduce technologies, methodologies and algorithms to manage Big Data problems. More precisely, this course provides students with the theoretical knowledge and technical skills to collect, store, organize and analyze data using Big Data solutions based on the Hadoop Ecosystem.
This course assumes a basic knowledge of basic data structures (such as lists, hashes and graphs) and a prior knowledge of: the Java programming language, SQL and the Linux shell commands.

LEARNING OUTCOMES

- ✓ principles of MapReduce programming paradigm, reinforced by presentation of some algorithms implemented in MapReduce style;
- ✓ main features and architectural components of the following big data frameworks: Hadoop, Storm, Hive, Spark;
- ✓ how to analyse and select the proper technologies to face a Big Data problem;
- ✓ the "Lambda-Architecture" pattern;
- ✓ how to learn by-example using a pre-configured Big Data Platform.

METHODOLOGY
Theoretical lessons, discussion, question and answers, demonstrations, practical sessions (hands-on practice).

ASSESSMENT
Written exam; weighting: 100%

OUTLINE

- Introduction to Big Data, Map-Reduce and Hadoop:
  1. Introduction to big data problems and platforms.
  2. Map-Reduce and examples: word count, average temperature, image smoothing, page-rank.
  3. Hadoop, HDFS.

- Hadoop2 in practice:
  1. data logistics (data serialization, organizing and optimizing data in HDFS, moving data into and out of Hadoop),
  2. big data patterns (joining),
  3. data structures and algorithms at scale (e.g. Bloom Filters).

- Beyond MapReduce:
  1. SQL on Hadoop (Hive).
  2. Outline of Apache Storm and Apache Spark.
  3. Lambda architecture.

- Laboratory sessions focused on Hadoop, Hive and Spark.

TEXTBOOKS

Slides and references at free resources on the Web. Some of these are:

- ✓ MapReduce: Simplified Data Processing on Large Clusters
- ✓ The Google File System
- ✓ Hive – A Petabyte Scale Data Warehouse Using Hadoop
- ✓ Jure Leskovec, Anand Rajaraman, Jeff Ullman - *Mining of Massive Datasets* – free ebook: http://www.m-mds.org

ADDITIONAL SUGGESTED READING

- Alex Holmes - *Hadoop in Practice (second edition)* - 2015 – Manning
- Allen, Jankowsky, Pathirana – *Storm Applied – strategies for real-time event processing* – 2015 Manning
- Nathan Marz – *Big Data – principles and best practice of scalable real-time data system* – 2015 – Manning
- H. Karau et al. - Learning Spark: Lightning-Fast Big Data Analysis – 2015 – O'Reilly
- Grover, Malaska, Seidman & Shapira - *Hadoop Application Architectures Designing real-world big data application – 2015 – O'Reilly*
- S. Ryza et al. - Advanced Analytics with Spark: Patterns for Learning from Data at Scale - *2015 – O'Reilly*