Big Data:
Technologies for Big Data

Course description

Outline

- General information
- Course objectives
- Topics
- Textbooks
- Tools

General information

- Big Data
  - Technologies for Big Data (6 CFU)
    - Prof. Marco La Cascia
  - Analysis for Big Data (6 CFU)
    - Prof. Roberto Pirrone

General information

- Technologies for Big Data
- Marco La Cascia
  - marco.lacascia@unipa.it
  - Monday 15:00-17:00 (wztkv0u oppure Ed.6, III piano, stanza 7)
- Lecture time and place
  - Monday 10:00-13:00, F180 (Lecture)
  - Thursday 8:00-10:00, A320 (Practice)

~~marco.lacascia@community.unipa.it~~
~~marco.lacascia@you.unipa.it~~
~~MS Teams chat~~

General information

- Course homepage
  - portale.unipa.it
- Prerequisites
  - Basic knowledge of probability, statistics, and linear algebra
  - Basic Python programming
  - Relational DB, basic SQL

Course objectives

- The course is aimed at providing students with a knowledge of the software architectures for Big Data along with both the main algorithms for data analysis and preprocessing techniques with the aim of developing autonomously whole data analysis pipelines for real case studies
- The module allows acquiring 6 ECTS, and it is arranged in lessons and exercise sections.

Lessons

- Lessons start presenting at first the whole data analysis process. Next, preprocessing techniques, most widespread similarity measures in data analysis and frequent patterns analysis algorithms.
- Software architectures for Big Data wil be treated: databases noSQL will be presented along with the MapReduce algorithm, the Apache Hadoop ecosystem and the Apache Spark framework.

Exercise sections

- Exercises are mainly related to Python implementations of data processing and analysis.
- Configurations of the software environments that are used throughout the course.
- Exercises also include the discussion of numerical and theoretical problems proposed in the textbook.

Topics (1)

- Introduction to Data Mining
  - Data Mining Process, Basic Data Types, Major Building Blocks, Scalability Issues
- Data Preparation
  - Feature Extraction and Portability, Data Cleaning, Data Reduction and Transformation
- Similarity and Distances
  - Multidimensional Data, Text Similarity Measures, Temporal Similarity Measures, Graph Similarity Measures, Supervised Similarity Functions
- Association Pattern Mining
  - Frequent Pattern Mining Model, Association Rule Generation Framework, Frequent Itemset Mining Algorithms

Topics (2)

- Software architectures for Big Data
- Database noSQL, MongoDB
- MapReduce
- Apache Hadoop, HDFS
- Apache Spark and its libraries
- Graph databases and graph processing

Textbooks

- Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411 (https://link.springer.com/book/10.1007/978-3-319-14142-8)
- Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, Bill Chambers & Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218
- Introduzione a Python. Per l'informatica e la data science, 2021, Paul J. Deitel & Harvey M. Deitel, Pearson, ISBN 978-8891915924
- MongoDB official documentation
- Graph Algorithms Practical Examples in Apache Spark & Neo4j, 2019, Mark Needham & Amy E. Hodler, O'Really, ISBN 978-1492047681

Examination

- Computer test and oral examination where the result of the test will be discussed.
- Theoretical topics will be assessed through open questions, Python coding will be required to answer the practical questions.
- Written test will last for two hours.
- **No project work!**
  - **Interested students can ask for thesis.**