Data science and database technology

Introduction to the course

Transaction processing

- On Line Transaction Processing (OLTP)
  - Traditional DBMS usage
- Characterized by
  - snapshot of current data values
  - detailed data, relational representation
  - structured, repetitive operations
  - read/write access to few records
  - short transactions
  - isolation, reliability, and integrity are critical (ACID)
  - database size ≈ 100MB-GB

Analytical processing

- On Line Analytical Processing (OLAP)
  - Decision support applications
- Characterized by
  - "historical" data
  - consolidated, integrated data
  - ad hoc applications
  - read access to millions of records
  - complex queries
  - consistency before and after periodical loads
  - database size ≈ 100GB-TB

Course content

- First part (weeks 1-7)
  - Data warehouse design
  - OLAP analysis
  - Data science and data mining
- Second part (weeks 8-14)
  - DBMS server technology
  - Distributed databases
  - NoSQL databases (MongoDB, Elastic)

Course structure

- The course includes
  - lessons
  - classroom exercises
  - laboratories
- Laboratory sessions propose experimental activities on the most widespread commercial and open-source products
  - Students are partitioned in two groups
  - Lab sessions will start on the fourth week

Books

- Course books
  - Golfarelli, Rizzi, 'Data Warehouse Design: modern principles and methodologies', McGraw Hill, 2021
  - Tan, Steinbach, Kumar, *Introduction to data mining*, Pearson, 2006
  - Atzeni, Ceri, Fraternali, Paraboschi, Torlone, 'Basi di dati', 5 ed., McGraw Hill, 2018.
  - Dan Sullivan, NoSQL for Mere Mortals, Addison-Wesley Professional, 2015
  - Kristina Chodorow, Shannon Bradshaw, MongoDB: The Definitive Guide (Powerful and Scalable Data Storage), 3 ed. O'Reilly Media, 2018
  - Gormley, Tong, Elastic Search: The Definitive Guide, O'Reilly, 2015

Books

- Other books
  - Ramakrishnan, Gehrke, *Database Management Systems*, McGraw-Hill, 2004
  - Kimball e altri, *several books and white papers on data warehouse design methodologies and case studies,* Wiley
  - Han, Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann, 2006

Assessment and grading criteria

- Exam
  - Written test (mandatory)
    - A set of design exercises
    - A set of theory questions and exercises
    - Textbooks, notes, electronic devices of any kind are *not allowed* during the written part
  - Individual project on the main topics of the lectures (optional)
    - homework assigned during the course
    - homework must be delivered at predefined deadlines during the course
- Further details about exam structure and grading criteria are available on the didactic portal