UNIVERSITÀ degli STUDI di CATANIA

DEPARTMENT OF ECONOMICS AND BUSINESS
Master's Degree in Data Science for Management
Academic Year 2019/2020 - 1° Year

DATA BASE AND BIG DATA ANALYTICS

12 CFU - 1° and 2° Semester

Teaching Staff

CONCETTO SPAMPINATO - Module DATA BASE - ING-INF/05 - 6 CFU
Email: cspampin@dieei.unict.it
Office: DIEEI, Plesso 13, stanza 8, Cittadella Universitaria
Phone: 095/7382057
Office Hours: Su prenotazione via email
ORAZIO TOMARCHIO - Module Big Data Analytics - ING-INF/05 - 6 CFU
Email: orazio.tomarchio@unict.it
Office: DIEEI, Cittadella Universitaria, Viale Andrea Doria 6, Edificio 13
Phone: 095 7382357
Office Hours: Pubblicato sulla pagina del docente nel sito web del DIEEI. Il docente è disponibile anche a incontri di ricevimento in modalità telematica, previo appuntamento.

LEARNING OBJECTIVES

- **DATA BASE**
  This module covers the fundamental concepts of management database systems at scale as well as the analysis of existing benchmarks in different application scenarios. Topics include data models (relational); query languages (SQL); implementation techniques of database management systems even at large scale; noSQL databases, temporal, patial, Multimedia, and Deductive Databases. The module will also discuss available large scale multimedia datasets and how to query them as well as the state of the art techniques on how to create benchmarks for testing data analytics techniques. Principles on how to detect mistakes, biases, systematic errors, and other unexpected problems will be analyzed.

  The learning objectives are:

  - To understand and use the main technologies for database management;
  - To use SQL language for performing efficient queries in cases of large datasets;
  - To understand how to index and query multimedia datasets
  - To become aware of the existing benchmarks and their liminations for training and comparing data analysis techniques.

  Knowledge and understanding

  - To understand the main concepts of management database systems
  - To understand concepts and tools for generating and querying datasets at different scales

- To understand techniques for indexing and searching multimedia datasets
- To understand how potential biases in data collection may affect analytics methods

Applying knowledge and understanding

- To be able to effectively understand and use the main tools for creating and querying SQL and NoSQL datasets.
- To query and analysis multimedia at large scale
- To understand proper benchmarks and analysing achieved results also in terms of potential biases

- **Big Data Analytics**
  This module covers the fundamental concepts of management and design of a business intelligence system. Topics include data models for building a data warehouse; ETL (extract, transform and load) functionalities; OLAP analysis; basic data mining; reporting and interactive dashboards, evolution of BI architectures on large datasets. The module covers techniques and algorithms for data visualization and exploratory analysis based on principles and techniques from graphic design, perceptual psychology and cognitive science. It is targeted to using visualization in their data analytics work.

  The learning objectives are:

  a. to understand and use the main methodologies and techniques for data analysis
  b. to understand the main methodologies to design a data warehouse
  c. to understand the main methodologies to transform data into sources of knowledge through visual representation

  Knowledge and understanding

  - To understand the main concepts of management and design of database systems
  - To understand concepts and tools for generating and querying datasets at different scales
  - To understand the most important methodologies and techniques used by industries to analyse data in order to support the decision process
  - To understand the main methodologies to design a data warehouse
  - To understand the main methodologies to transform data into sources of knowledge through visual representation

  Applying knowledge and understanding

  - To be able to effectively understand and use the main tools for creating and querying SQL and NoSQL datasets.
  - To query and analysis data at large scale
  - To understand proper benchmarks and analysing achieved results also in terms of potential biases
  - To be able to apply methodologies and techniques to analyse data.
  - To be able to design a data warehouse.
  - To be able to build report and data analysis and organize them into interactive dashboards

COURSE STRUCTURE

- **DATA BASE**
  The main teaching methods are as follows:

  - Lectures to provide the basic theoretical and methodological knowledge for understanding how to manage data at scale;
  - Hands-on exercises to make students apply the learned methods, thus to improve their solving problem skills
  - Paper reading and student presentations in order to provide critical thinking skills
  - Seminars by renowned reaserach and industrial experts in the field.

- **Big Data Analytics**
  The main teaching methods are as follows:

  - Lectures, to provide theoretical and methodological knowledge of the subject;
  - Hands-on exercises, to provide “problem solving” skills and to apply design methodology;
  - Laboratories, to learn and test the usage of related tools

---

DETAILED COURSE CONTENT

- **DATA BASE**
  1) Models and Languages for Database Management (15 hours)

  - Fundamentals of Database Management Systems (DBMS)
  - Relational Model: basic concepts, integrity constraints and keys.
  - SQL language: data definition, data modification, queries, views, transactions.
  - NO-SQL database: MongoDB

  2) Querying and processing big data (10 hours)

  - Apache Spark SQL with Python
  - Dataset and Dataframes
  - Window functions
  - Caching and logging functions
  - The Spark UI
  - Examples of data analysis with Spark SQL

  3) Analyzing existing benchmarks (15 hours)

  - Comparative analysis of benchmarks for testing out data analysis and machine learning methods for several tasks from classification to regression to generation.
  - Categorization of common biases in benchmarks: selection bias, negative bias, cross-generalization bias
  - Identifying and correcting dataset-related biased results

- **Big Data Analytics**
  1. Introduction to Business Intelligence and Big Data Analytics (6 hours)

  - Goal and rationale of BI systems

- The value of knowledge - data driven decision making
- The structure and evolution of BI and Big Data analytics systems
- OLAP vs OLTP
- Data warehouse and Business intelligence
- Advanced tools and platforms for BI and analytics

2. Data models for data warehouse (12 hours)

- Conceptual modeling
- Dimensions and facts
- Multi-dimensional data model
- Conceptual, logical and physical design

3. BI Architecture (12 hours)

- ETL (extract, transform and load) functionalities
- OLAP analysis
- OLAP query
- Reporting
- Interactive Dashboard

4. Data Visualization (10 hours)

- Introduction to Visualization
- Data transformation into sources of knowledge through visual representation
- Charts and standard views: relevance, appropriateness and best practices
- Advanced and innovative tools for data visualization
- The evaluation of the quality of visualizations

TEXTBOOK INFORMATION

- **DATA BASE**
  1. R. Elmasri and S. Navathe, Fundamentals of Database Systems, 7th Edition, Pearson, 2016.
  2. Denny Lee, Tomasz Drabas, Learning Spark SQL, Packt Publishing, 2017
  3. Instructor’s notes
  4. Research papers (a list will be published on the page course)

- **Big Data Analytics**
  1. Ralph Kimball, Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition, Wiley, 2013
  2. Instructor’s notes (published on Studium)
