UNIVERSITÀ degli STUDI di CATANIA

DIPARTIMENTO DI ECONOMIA E IMPRESA
Corso di laurea magistrale in Data Science for Management
Anno accademico 2019/2020 - 1° anno

DATA BASE AND BIG DATA ANALYTICS

12 CFU - 1° e 2° semestre

Docenti titolari dell'insegnamento

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

OBIETTIVI FORMATIVI

- **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 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 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

MODALITÀ DI SVOLGIMENTO DELL'INSEGNAMENTO

- **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

PREREQUISITI RICHIESTI

- **DATA BASE**
  Basic programming skills.

- **BIG DATA ANALYTICS**
  - Basic knowledge of database systems
  - Basic knowledge of SQL

FREQUENZA LEZIONI

- **DATA BASE**
  Strongly recommended. Attending and actively participating in the classroom activities will contribute to the overall assessment of the final exam (see evaluation procedure section) .

- **BIG DATA ANALYTICS**
  Strongly recommended. Attending and actively participating in the classroom activities will contribute positively towards the overall assessment of the oral exam.

CONTENUTI DEL CORSO

- **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

TESTI DI RIFERIMENTO

- **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)

ALTRO MATERIALE DIDATTICO

- **BIG DATA ANALYTICS**
  Instructor's notes will be made available on the Studium web site

---

PROGRAMMAZIONE DEL CORSO

DATA BASE

	Argomenti	Riferimenti testi
1	Introduction to databases: Concepts and Architecture	Book 1 - Chapter 1 and 2
2	Relational Data Model	Book 1 - Chapter 5
3	Basic SQL: data definition, SQL query, update instruction set.	Book 1 - Chapter 6 + Notes
4	Advanced SQL: Complex Queries, Triggers, Views	Book 1 - Chapter 7 + Notes
5	Query processing and optimization	Book 1 - Chapter 18 and 19
6	NOSQL Databases and Big Data Storage Systems	Book 1 - Chapter 24 + Notes
7	Active, Temporal, Spatial, Multimedia, and Deductive Databases	Book 1 - Chapter 26
8	Getting started with Spark SQL for Data Processing	Book 2 - Chapter 1 and 2 + Notes
9	Spark SQL for Data Exploration	Book 2 - Chapter 3 + Notes
10	Spark SQL for Learning Applications	Book 2 - Chapter 6 and 10 + Notes
11	Multimedia benchmarks for bias identification and analysis	Research paper list on course course

---

VERIFICA DELL'APPRENDIMENTO

MODALITÀ DI VERIFICA DELL'APPRENDIMENTO

- **DATA BASE**

The final exam consists of a) a lab test aiming at assessing the capabilities in writing SQL and NoSQL queries using also SPARK SQL, b) a final report critically analyzing, in terms of possible biases, an existing dataset. The exam is evaluated according to the ability to write SQL queries, derive aggregated information from data and to discover correctly biases in data and motivate solutions for solving such biases.

The vote on the database module will account for 50% of the total grade for the entire course.

The module also foresees intermediate tests for students attending the course. These tests include: a lab test on SQL writing, a presentation, two written reports analyzing existing benchmarks (whose list will be given at the beginning of the course). The choice of the datasets to analyze will be done during classes in order to avoid overlap.

The grading policy for intermediate tests is:

- SQL test: 30%
- Paper presentation: 30%
- Reports: 30%
- Attendance and Discussion during classes: 10%

- **BIG DATA ANALYTICS**
  The final exam consists of a) a project work aiming at assessing the capabilities in developing a BI system including the analysis and the visualization of relevant information, b) an oral exam that will consist of the discussion of the project work.

  Assessment criteria include: depth of analysis, adequacy, quality and correctness of the proposed solutions to the project work, ability to justify and critically evaluate the adopted solutions, clarity.

  The vote on the Big Data Analytics module will account for 50% of the total grade for the entire course.

ESEMPI DI DOMANDE E/O ESERCIZI FREQUENTI

- **DATA BASE**
  Examples of questions and exercises will be available on the webpage course and on the Studium platform.

- **BIG DATA ANALYTICS**
  Examples of questions and exercises are available on the Studium platform