Testi del Syllabus

Resp. Did. MONTANARI ANGELO Matricola: 029635

Docenti
MONTANARI ANGELO, 3 CFU
DELLA MONICA DARIO, 3 CFU
CONTRATTO, 3 CFU

Anno offerta: 2018/2019

Insegnamento: 583SM - DATA MANAGEMENT FOR BIG DATA

Corso di studio: SM35 - DATA SCIENCE AND SCIENTIFIC COMPUTING

Anno regolamento: 2017

CFU: 9

Settore: INF/01

Tipo Attività: B - Caratterizzante

Anno corso: 1

Periodo: Secondo Semestre

Sede: TRIESTE

Lingua insegnamento: English

Contenuti (Dipl.Sup.)

1. Fundamentals of database systems.
The students will learn, and practice, how to design, develop, populate, and manipulate (query and update) a relational database (data models, integrity constraints, normal forms, query and update languages, transactions, indexes).

2. Advanced database models, languages, and systems.
The students will learn, and practice, advanced query processing techniques for relational databases as well as alternative data models and languages (XML databases). Moreover, they will be introduced to the basic elements of distributed and parallel database management systems that play a fundamental role in the management of big data.

3. Data analysis and big data.
The students will learn, and practice, the main techniques and tools for data analysis and big data management. A special attention will be given to data warehousing, data mining and methods and tools for big data. A number of key topics will be addressed,

ranging from the MapReduce paradigm to blockchain and its applications.

Testi di riferimento

- Fundamentals of Database Systems (7th Edition), Elmasri and Navathe, Pearson, 2016
- Readings in Database Systems (online, http://www.redbook.io)
- Principles of Distributed Database Systems (3rd Edition), Özsu and Valduriez, Springer, 2011
- Mining of Massive Datasets (online, http://mmds.org)
- scientific papers on advanced topics

Obiettivi formativi

The students must learn how to organize, manipulate, and analyze small and big data with a variety methods, techniques, and tools.

Knowledge and understanding: the students must acquire the necessary knowledge to model, import, tidy, transform, query, visualize, and analyze data as well as to communicate the results of the analysis. We take into consideration relational data as well as semistructured and unstructured data.

Applied knowledge and understanding: the students must learn languages and tools for the manipulation, analysis, and visualization of data, such as, for instance, PostgreSQL and BaseX for the management of relational and XML data, R and RStudio environment for data analysis and visualization, Processing for the visualization of data, and R Markdown language for the communication of the results of the analysis.

Making judgments: the students must be able to interpret the experimental results of the analysis
and draw effective conclusions relevant to the domain of discourse.

Communication skills: the students must be able to communicate effectively the results of the analysis. This includes both analyst-to-analyst communication and analyst-to-decision-maker communication.

Learning skills: the students must demonstrate that they have learned the ability to choose a sufficiently rich row data set, to analyze the data to extract meaningful information, to draw and to communicate conclusions.

Prerequisiti

Some basic knowledge about programming, algorithms and data structures, logic, and statistics are desirable.

Metodi didattici

The course will be multi-task (learn, make, use, read, dig, listen) and multi-teacher (the course is organized in three parts given by three different teachers; in addition, some advanced topics will be covered by invited experts in the field).

A task-tag legend is below:
Learn: professors teach, students listen (and hopefully learn).
Make: professors give assignments to students, that make them during classes. The solutions are discussed in one of the following class.
Use: Students use software: download, install, and run it for the first time. Professors give them a brief practical introduction to it.
Read: students read book's chapters and papers, typically at home. We discuss them together during one of the following class.
Listen: Some classes are given by invited speakers, experts in some specific fields.

Altre informazioni

Additional suggested books:
- PostgreSQL: Up and Running (3rd Edition), Regina Obe and Leo Hsu, O'Reilly Media, 2017
- An Introduction to XML and Web Technologies, Anders Møller and Michael I. Schwartzbach,
Addison-Wesley, 2006
- R Cookbook, Paul Teetor, O'Reilly Media. 2011.
- ggplot2: Elegant Graphics for Data Analysis, Hadley Wickham, Springer, 2010.
- Introduction to Automata Theory, Languages, And Computation, John E. Hopcroft, Rajeev Motwani,
and Jeffrey D. Ullman, Addison-Wesley, 2006.
- The Visual Display of Quantitative Information, Edward. R. Tufte, Graphics Press. 2001

Modalità di verifica dell'apprendimento

The exam consists either of a traditional oral examination or ofa project to be jointly done by a small group of students. Each group is asked to: (i) choose a sufficiently rich row data set, (ii) analyse the data to extract meaningful information using the methods and tools given during the course, (iii) draw and document conclusions, and (iv) communicate the outcomes and conclusions during an oral presentation

Programma esteso

Part 1 - 3 cfu (24 hours): Fundamentals of database systems

- Introduction to the DataBase Management Systems (DBM) - 2 hours

- Data models - 2 hours
+ Conceptual models (Entity-Relationship / ER Model)
+ Logical models (Relational Model)

- Design methodologies - 4 hours
+ Mapping ER schemas into relational ones
+ Functional dependencies and normalization

- Data definition, update, and query languages - 8 hours
+ Relation algebra and relational Calculus
+ SQL

- Transactions - 4 hours

- Indexes - 4 hours

Part 2 - 3 cfu (24 hours): Advanced database models, languages, and systems

- Query processing and optimization - 6 hours
+ Query processing
+ Algorithms for the join operation
+ Cost-based optimization and heuristics

- Semistructured Data and XML - 4 hours
+ Definition of semistructured data in XML
+ Querying XML data (XPath and XQuery)
+ XML and relational DBMS
+ Native XML databases

- Distributed and parallel database architectures - 12 hours
+ An introduction to parallel and distributed DBMS
+ Design of distributed databases (fragmentation and replication)
+ Distributed query processing
+ Optimization of distributed queries
+ Transaction processing in distributed databases: the two-phase commit (2PC) protocol
+ Parallel DBMS

- Cloud computing and DBMS - 2 hours

Part 3 - 3 cfu (24 hours): Data analysis and big data

- Data warehousing - 4 hours

- Data mining - 6 hours
+ Mining massive datasets
+ Data reduction (feature selection, instance selection, ..)
+ Text mining
+ Time series analysis

- Fundamentals of big data - 4 hours
+ Distinctive features, data science, and applications

+ Technologies for the management of big data
+ Internet of things and big data

- The MapReduce paradigm - 2 hours

- NewSQL systems - 2 hours
+ In-memory databases
+ Column-oriented databases

- NoSQL systems - 2 hours

- Blockchain and applications - 4 hours