UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Matematica e Informatica
ACADEMIC YEAR	2020/2021
MASTER'S DEGREE (MSC)	COMPUTER SCIENCE
INTEGRATED COURSE	DATA PROCESSING
CODE	19742
MODULES	Yes
NUMBER OF MODULES	2
SCIENTIFIC SECTOR(S)	INF/01
HEAD PROFESSOR(S)	GIANCARLO RAFFAELE Professore Ordinario Univ. di PALERMO
OTHER PROFESSOR(S)	GIANCARLO RAFFAELE Professore Ordinario Univ. di PALERMO ROMBO SIMONA ESTER Professore Ordinario Univ. di PALERMO
CREDITS	12
PROPAEDEUTICAL SUBJECTS	
MUTUALIZATION	
YEAR	1
TERM (SEMESTER)	2° semester
ATTENDANCE	Not mandatory
EVALUATION	Out of 30
TEACHER OFFICE HOURS	**GIANCARLO RAFFAELE** Monday 15:00 17:00 Stanza 106 Dipartimento di Matematica ed Informatica Thursday 15:00 17:00 Stanza 106 Dipartimento di Matematica ed Informatica **ROMBO SIMONA ESTER** Monday 9:30 13:30 In presenza: Via Archirafi 34, Piano II, Stanza 220 - Telematico: via Microsoft Teams o altro canale - In entrambi i casi, e' consigliabile prenotarsi tramite email alla docente

DOCENTE: Prof. RAFFAELE GIANCARLO

PREREQUISITES	Basic Knowledge in: algebra, geometry, algorithms and data structures, databases. Advanced Programming knowledge.
LEARNING OUTCOMES	Knowledge and ability to understand Acquisition of knowledge about the organization of business information systems, on the use and design of decision support systems, on problems related to big data management. Acquisition of advanced methods for the design and analysis of algorithms. Ability to use the specific language proper of this specialised area. Ability to apply knowledge and understanding Ability to analyze issues related to contexts characterized by large data sets and propose design solutions for the management and analysis of data in such contexts and decision support. Ability to develop software based on efficient algorithms for large datasets. Making judgments Ability to analyze and evaluate solutions for the management of large quantities of data. Ability to design decision support systems by analyzing the technical specifications provided. Being able to evaluate the implications and the results of algorithmic studies and of computational complexity associated to those topics. Enable communication Ability to describe designing solutions for complex information systems and decision support systems, and also to analyze their performance. Ability to cooperate for determining the appropriate design solutions in application contexts characterized by the presence of large amounts of data. Ability to explain algorithmic results even to a non-expert audience. To be able to illustrate the technological fall-out coming from algorithmic theory. Learning Abilities Ability to update by consulting advanced books and scientific publications related to the topics covered during the course. Ability to follow, using the knowledge acquired, both second master level and advanced courses and specialized seminars in the fields of Business Intelligence and Algorithm and Data Structures Design.
ASSESSMENT METHODS	Project, oral examination. Project (for the Big Data module): Design and implementation of a project based on big data technologies. It will be mandatory that the project is carried out by a team consisting of a minimum of two to a maximum of five students, in order to encourage them to work in team. Two revisions of the project will be necessary, during which the teacher will verify the individual contribution of each team member and assign a rating to the different phases of design. An overall score will be given to the project, weighted for each student depending on his/her individual contribution. A positive evaluation will depend on the originality of the proposed solution, the methodological rigor, the acquisition of technicall skills supplied by the course. The requirements to achieve the minimum score to access the second part of the examination for this module consists of the ability to realize a project that, while so basic, satisfy the assigned requirements and it is correct. Oral test: it will serve to check the knowledge gained during both modules, the autonomy in deepening even complex content and the ability to find individual solutions to the proposed problems. The oral test will begin, for both modules, with the presentation of a topic chosen from those proposed by the teacher during the course, on which the student will have produced a short essay. For Big Data Management, the student will produce a short paper and slide presentation. So, it will be checked for critical capacity and autonomy of the student's judgment through a sufficient number of questions. The assessment of the oral examination will be different for the two modules and it will complement the marks obtained by each student as a result of the project evaluation for Big Data Management. The final evaluation will be obtained by arithmetic average of the evaluations for the two modules. In particular: 18-21: Sufficient knowledge of all the parts of the program. 22-24: Discreet knowledge of all the parts of the program. 25-27: Good knowledge of all the parts of the program. 28-30: Excellent knowledge of all the parts of the program. Lode. It is awarded when, provided full score, the student has achieved in both modules an excellent ability to apply in authonimy the contents of the modules.
TEACHING METHODS	Frontal Teaching

MODULE
ALGORITHM SCIENCE AND ENGINEERING

Prof. RAFFAELE GIANCARLO

SUGGESTED BIBLIOGRAPHY

William J.Cook, William H. Cunningham, William R. Pulleyblank, Alexander SchrijverI. Combinatorial Optimization, Wiley 1997

Per le parti che rigurdano algoritmi di approssimazione- For the part regarding approximation algorithms

Robert Endre Tarjan. Data Structure and Network Algoritms, SIAM 1984
Per le parti che riguardano strutture dati in memoria interna-for the part regarding data structures in internal memory

Camil Demetrescu, Irene Finocchi,Giuseppe F. Italiano, Algoritmi e Strutture dati, McGraw Hill, 2005
Per le parti che riguardano l'analisi ammortizzata di algoritmi. For the part regarding amortised analysis of algorithms.

H. Cormen. C. Leiserson, R, Rivest, C. Stein Introduzione agli algoritmi e strutture dati, McGraw Hill, 2001
Per le parti che riguardano hashing e schemi di approssimazione-for the parts regarding hashing and approximation schemes

Materiale distribuito dal docente.
Tutto il resto del programma The remaining part of the syllabus.

AMBIT	50341-Discipline Informatiche
INDIVIDUAL STUDY (Hrs)	102
COURSE ACTIVITY (Hrs)	48

EDUCATIONAL OBJECTIVES OF THE MODULE

To expose the student to the advanced techniques for the design and analysis of computer algorithms. In particular, the entire spectrum of dynamic data structures is covered, with an in-depth study of the intrinsic computational complexity of difficult problems or problems that involve a large quantity of data.

SYLLABUS

Hrs	Frontal teaching
4	Preliminary Notions and Background Amortised Analysis of Algorithms: the method of credits; the potential method. Experimental analysis of Algorithms.
6	Internal Memory Advanced Data Structures: Balancing Red-Black Trees and analysis of the operations they support. Linking and Cutting of Trees.
6	Self-adjusting binary trees and analysis of the operations they support. Self-organizing Data Structure. Self-organizing List.
6	Space Succinct Dictionaries: Universal e Perfect Hashing.
5	Space Succinct Dictionaries: Bloom Filters.
6	Steaming Model. Motivation. The straming model of computation. Examples of streaming algorithms for mining large datasets.
6	Steaming Algorithms. Frequency Moments. Sketches.
4	Data Compression Schemes and their analysis. Static and adaptive data compression schemes. Compression Boosting and its engineering. Efficient data structures for data compression. Benchmarks for data compression performance analysis.
3	NP-Completeness Theory and Polynomial Approximations. Polynomial time approximation schemes. Non-approximability.
2	Approximation Algorithms and Heuristic Methods. TSP with triangle inequality.

MODULE
BID DATA MANAGEMENT

Prof.ssa SIMONA ESTER ROMBO

SUGGESTED BIBLIOGRAPHY

Parti su DATA WAREHOUSE: M. Golfarelli, S. Rizzi, “Data Warehouse – Teoria e pratica della progettazione”, Seconda Edizione, McGraw Hill, 2005.
Parti su FRAMEWORK, TECNOLOGIE, BIG DATA MINING: Jure Leskovec, Anand Rajaraman, Jeff Ullman. "Mining of Massive Datasets", Third Edition, 2020.
Parti su DATA MINING: J. Han, M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

AMBIT	50341-Discipline Informatiche
INDIVIDUAL STUDY (Hrs)	102
COURSE ACTIVITY (Hrs)	48

EDUCATIONAL OBJECTIVES OF THE MODULE

The main goal of Big Data Management is to provide students with knowledge about the organization of business information systems, the use and the design of decision support systems, and also on issues related to the management of large amounts of data. The course will begin with an overview of Business Intelligence and it will focus on the design of data warehouse and OLAP. The course will also address the design of non-relational database and the use of technologies such as Apache Hadoop and Spark. Finally,real application scenarios will be illustrated such as social networks, smart cities and biological data.

SYLLABUS

Hrs	Frontal teaching
2	INTRODUCTION TO BIG DATA Needs. Application contexts. An overview of the technologies.
3	BUSINESS INTELLIGENCE AND DATA WAREHOUSING Data acquisition. Processing of data in order to produce information. Archiving of raw data and information. Transmission of data and information. Presentation of data and information. Decision support systems. Architectures for Data Warehousing. The metadata. Quality of a Data Warehouse. The multidimensional model and OLAP. The main OLAP operations. Logical models of Data Warehouse.
4	DESIGN OF A DATA WAREHOUSE Design methodologies, process selection, choice of granularity, identifying and bringing the size, selection of measures, precalculations in the fact table, complete the dimension table, choice of the duration of the database, track "slowly changing dimension". Conceptual design. Logical design.
4	DATA CLEANING ETL. Data cleaning problems and techniques.
8	DESIGN OF DATA WAREHOUSE BY PENTAHO Study of the modules: Kettle, Data Integration, OLAP.
4	STRUCTURED STORAGE Non-relational database. NoSQL database types. Analysis of the advantages and disadvantages of not-relational databases. Implementation. Examples and exercises on NoSQL database.
6	MAP REDUCE Distributed File Systems. The MapReduce Paradigm. Design and implementation of algorithms in MapReduce (Arrays product, relational algebra, word counts, PageRank).
4	OTHER BIG DATA FRAMEWORKS AND TOOLS Resilient Distributed Datasets (RDD), TensorFlow, Apache Hadoop, Apache Spark.
6	BIG DATA MINING Basic notions. Market Basket Analysis. Classification and prediction. Decision trees. K-Nearest Neighbor. Clustering. Outlier detection. Data mining problems on large datasets, challenges and solutions. Artificial Intelligence and Big Data. Data Mining, Machine Learning and Statistics. Social Networks and Web Advertising. Recommendation Systems. Spark Mllib and applications to Collaborative Filtering.
4	BIG NETWORKS Social networks. Biological networks. Centrality measures. Classification results validation. Spark GraphX.
3	BIG SEQUENCES AND OTHER PROBLEMS. Analysis of NGS data and biological sequences. Indexing and compressing in MapReduce. Precision Medicine. Datasets e Data Frames. Data Lake. ELT vs ETL.