UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Ingegneria		
ACADEMIC YEAR	2023/2024		
MASTER'S DEGREE (MSC)	COMPUTER ENGINEERING		
INTEGRATED COURSE	BIG DATA - INTEGRATED COURSE		
CODE	22216		
MODULES	Yes		
NUMBER OF MODULES	2		
SCIENTIFIC SECTOR(S)	ING-INF/05		
HEAD PROFESSOR(S)	PIRRONE ROBERTO	Professore Ordinario	Univ. di PALERMO
OTHER PROFESSOR(S)	PIRRONE ROBERTO LA CASCIA MARCO	Professore Ordinario Professore Ordinario	Univ. di PALERMO Univ. di PALERMO
CREDITS	12		
PROPAEDEUTICAL SUBJECTS			
MUTUALIZATION			
YEAR	1		
TERM (SEMESTER)	Annual		
ATTENDANCE	Not mandatory		
EVALUATION	Out of 30		
TEACHER OFFICE HOURS	**LA CASCIA MARCO** Monday 15:00 17:00 Microsoft Teams Codice: wztkv0u **PIRRONE ROBERTO** Wednesday 11:30 13:00 Studio del docente, Edificio 6, terzo piano, stanza 3025		

DOCENTE: Prof. ROBERTO PIRRONE

PREREQUISITES	Base statistics
LEARNING OUTCOMES	Knowledge and understanding When having attended the course, students will own knowledge and methodologies to solve problems related to both the analysis of the most well-known data types and the use of software architectures for Big Data. Students will know adequately the differences between heterogeneous algorithm according to different data types; they will know the most suited preprocessing techniques, and how to define the most effective Big Data architecture for their analysis purposes. To reach this objective, the course is arranged in lessons. Such an objective will be verified through the theoretical questions in the written test related to each module, and the discussion of the results of each written test. Applying knowledge and understanding When having attended the course, students will own knowledge and methodologies solve problems related to the implementation of analysis pipelines for both classical datasets and Big Data. Students will know deeply the Python progrmming language along with the main library for visualizing and analyzing data like Numpy, SciPy, Scikit-learn, Matplotlib, Pandas, Tensorflow, and Keras. Moreover, students will know adequately both noSQL databases like Apache Cassandra and the Apache Hadoop ecosystem. Finally, they will know deeply the Apache Spark framework and the Python API for its library. To reach this objective, the course is arranged in exercises to develop pipelines for data analysis. Such an objective will be verified through the practical questions in the written test related to each module, and the discussion of the results of each written test. Making judgements Students will be able to compare the features of different IDEs and/or frameworks for Big Data analysis to find the solution for specific problems. They will be able to face unstructured problems at an operating level, and to take decisions in uncertain contexts. The methodologies learnt during the course will allow students to deepen new applicative problems in the field of Big Data and data analysis. To reach this objective, the course is arranged in two series of exercises, one for each module. Such an objective will be verified through the theoretical questions in the written test related to each module, and the discussion of the results of each written test. Communication Students will be able to talk about complex Big Data issues in highly specialized contexts, using the proper language. To reach this objective, the course is arranged in two series of exercises, one for each module. Such an objective will be verified through the the discussion of the results of each written test. Lifelong learning skills Students will be able to face autonomously whatever Big Data related issue. They will be able to deepen complex topics such as comparing the performances of different Big Data frameworks to devise their strengths and weaknesses. To reach this objective, the course is arranged in two series of exercises, one for each module. Such an objective will be verified through the the discussion of the results of each written test.
ASSESSMENT METHODS	The final exam consists of two separate written tests, one for each module. Each written test will be followed by an oral examination where the result of the corresponding test will be discussed. Every written test will last for two hours, and it is aimed at assessing both the degree of theoretical knowledge of the topic covered by each module and the competence attained in facing the topics covered by exercises. Theoretical topics will be assessed through open questions, while some coding will be required to answer the practical questions. Students with a minimun mark of 18/30 will be able to undergo the oral examination. Grades will be measured according to the following levels: -18/30 – 20/30: the student has an almost sufficient knowledge of the thoretical topics covered during the course; he is able to develop just some parts required to answer the practical questions. -21/30 – 23/30: the student has a discrete knowledge of the thoretical topics covered during the course; he is able to develop roughly all the components

	required to answer the practical questions. -24/30 – 26/30: the student has a good knowledge of the thoretical topics covered during the course; he is able to develop completely all the components required to answer the practical questions. -27/30 – 30/30: the student has a full knowledge of the thoretical topics covered during the course; he is able to provide a complete and correct implementation of all the components required to answer the practical questions. -30 cum laude: the student has extremely good knowledge of the thoretical topics covered during the course; he is able to to provide very good implementation of all the components required to answer the practical questions. Moreover, the student exhibits originality and autonomous deepening of the topics covered by the course. Finally, also her/his implementation is original.
TEACHING METHODS	Lessons, exercises, and workgroups for developing Big Data analysis pipelines.

MODULE
BIG DATA TECHNOLOGIES

Prof. MARCO LA CASCIA

SUGGESTED BIBLIOGRAPHY

Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, disponibile gratuitamente in forma elettronica per gli studenti dell'Ateneo.
Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00.
Introduzione a Python. Per l'informatica e la data science, 2021, di Paul Deitel & Harvey Deitel, Pearson, ISBN 978-8891915924, prezzo orientativo € 45,00.
Note fornite dal docente/Lecture notes

AMBIT	50369-Ingegneria informatica
INDIVIDUAL STUDY (Hrs)	96
COURSE ACTIVITY (Hrs)	54

EDUCATIONAL OBJECTIVES OF THE MODULE

The course is aimed at providing students with a deep knowledge of the software architectures for Big Data along with both the main algorithms for data analysis and preprocessing techniques with the aim of developing autonomously whole data analysis pipelines for real case studies.
The module allows acquiring 6 ECTS, and it is arranged in lessons and exercise sections.
Lessons start presenting at first the whole data analysis process. Next, preprocessing techniques are faced like dimensionality reduction and missing data management, while introducing some of the most widespread similarity measures in data analysis and frequent patterns analysis algorithms. Then software architectures for Big Data wil be treated: databases noSQL will be presented along with the MapReduce algorithm, the Apache Hadoop ecosystem and the Apache Spark framework.
Exercises cover the Python language and the related modules (numy, pandas, matplotlib, sklearn), the configurations of the software environments that are used throughout the course, and the implementation of some topics covered in class.

SYLLABUS

Hrs	Frontal teaching
2	Introduction. Data analysis workflow: data gathering, preprocessing, applying analysis techniques, knowledge extraction.
3	Data preparation: data types, data cleaning, missing data, sampling.
3	Dimensionality reduction: Principal Component Analysis, Singular Value Decomposition, Wavelet transform, Muti Dimensional Scaling, Graph embedding.
4	Similarities and distances for different data types: quantitative data, categorical data, text data, temporal sequences, graphs.
4	Mining frequent patterns: Apriori algorithm, correlation statistics.
4	Software architectures for Big Data: database noSQL, MongoDB. Data lake.
8	Software architectures for Big Data: l'algoritmo MapReduce, Apache Hadoop, HDFS
8	Software architectures for Big Data: Apache Spark and its libraries.

Hrs	Practice
9	Python and numpy, pandas, matplotlib, sklearn modules review
3	MongoDB
3	Apache Hadoop, HDFS
3	Analyzing data in Spark SQL

MODULE
ANALYSIS FOR BIG DATA

Prof. ROBERTO PIRRONE

SUGGESTED BIBLIOGRAPHY

Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, prezzo orientativo € 70,00

Deep Learning, (2016), di Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press, ISBN 978-0262035613, prezzo orientativo €65,00

Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, (2017) Sebastian Raschka, Vahidm Mirjalili, Packt Publishing, ISBN 978-1787125933, prezzo orientativo € 35,00

Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, prezzo orientativo € 45,00.

Repository per slide ed esercitazioni:
https://github.com/fredffsixty/Big-Data

Siti web con manuali di riferimento per le esercitazioni ed i testi:
https://link.springer.com/book/10.1007%2F978-3-319-14142-8
http://www.deeplearningbook.org/
https://github.com/PacktPublishing/Python-Machine-Learning-Second-Edition
https://github.com/databricks/Spark-The-Definitive-Guide

AMBIT	50369-Ingegneria informatica
INDIVIDUAL STUDY (Hrs)	96
COURSE ACTIVITY (Hrs)	54

EDUCATIONAL OBJECTIVES OF THE MODULE

The module provides students with a deep knowledge of the main algorithms for data analysis both in Big Data perspective and in a classic Machine Learning setup, with the aim of building autonomously a complete data analysis pipeline for real problems.

The module allows to acquire 6 ECTS, and it consists of lessons and exercises.

Lessons start with an introduction to Probability, and Information Theories along with statistical estimate and sampling concepts. Then the part of the course devoted to machine learning is faced, and in particular clustering, classifiers, neural networks, and deep learning wil be studied. Finally, some application scenarios are presented: medical image analysis, natural language processing, and web data analysis.

Exercises are related to using the Python libraries sci-kit learn, Spark ML and Tensorflow for explaining the topics faced in the lessons using some examples already developed.

SYLLABUS

Hrs	Frontal teaching
3	Introduction to Probability, and Information Theories along with statistical estimate and sampling concepts.
2	Introduction to Machine Learning: supervised and unsupervised learning, reinforcement learning, model capacity, parameters and hyperparameters, error types, training.
5	Clustering: k-means algorithm and similar approaches, hierarchical clustering, density-based and grid-based clustering, high dimensional data, clustering evaluation, outliers.
5	Classification basics: feature selection, decision trees, rule-based classifiers, Naïve Bayes, logistic regression, Support Vector Machines, Nearest Neighbor, classifiers evaluation.
2	Classification advanced: Multi-class and rare class learning, regression on numeric data, semi-supervised learning, ensemble methods.
8	deep learning: structure of a neural network, hidden and output units, loss functions, computational graph, stachastic gradient descent, optimization and regularization, CNN, Autoencoders, LSTM, GAN, Graph Neural Networks, fine tuning and transfer learning.
3	Medical image analysis: segmentation of CT/MR volumes using CNN 3D
3	Natural Language Processing: text classification using Word2Vec.
5	Web data analysis: PageRank algorithm, recommender systems, web usage analysis, social network analysis.

Hrs	Practice
3	Statistical estimate of a Gaussian distribution varying the sampling type.

3	Clustering with sci-kit learn
3	Classification using sci-kit learn
3	Implementing a Spark ML pipeline for clustering
3	Implementing a Spark ML pipeline for classification
3	Tensorflow: implementation of very simple DNNs.