Modello 3 – Scheda di trasparenza proposta

SUBJECT	BIG DATA TOOLS (MODULE OF BIG DATA C.I. TEACHING)

PREREQUISITES	Base Statistics
LEARNING OUTCOMES	Conoscenza e capacità di comprensione 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, and the discussion of its results.
	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. Moreover, students will know adequately both noSQL databases like 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 includes a series of exercises to develop pipelines for data analysis. Such an objective will be verified through the practical questions in the written test, and the discussion of its results.
	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 includes a series of exercises. Such an objective will be verified through the theoretical questions in the written test, and the discussion of its results.
	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 includes a series of exercises.

	Such an objective will be verified through the the discussion of the results of the 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 includes a series of exercises. Such an objective will be verified through the the discussion of the results of the written test.
ASSESSMENT METHODS	The final exam consists of a written test, which will be followed by an oral examination where the result of the written test will be discussed. The written test will last for two hours, and it is aimed at assessing both the degree of theoretical knowledge of the topic covered 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 theoretical topics covered during the course; he/she 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 theoretical topics covered during the course; he/she 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/she 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/she 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 thoretica topics covered during the course; he/she 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. For students with disabilities and neurodiversity, the compensatory tools and dispensatory measures identified will be guaranteed, from CeNDis - Centro di Ateneo per la disabilita e la neurodiversita, based on specific needs and in implementation of current legislation.
EDUCATIONAL OBJECTIVES	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 couse 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 paradigm, the Apache Hadoop ecosystem and the Apache Spark framework. Exercises cover the Python language and the related modules (numpy, pandas, matplotlib, sklearn), the configurations of the software environments that are used throughout the course, and the implementation of some topics covered in class.
TEACHING METHODS	Lectures; Theoretical exercises; Group tutorials for developing data analytics pipelines with big data technologies.
SUGGESTED BIBLIOGRAPHY	Data Mining: The Textbook, 2015, Charu C. Aggarwal, Springer-Verlag New York, ISBN 978-3319141411, available free of charge in electronic form for students of the University. Spark: The Definitive Guide: Big Data Processing Made Simple, 2018, di Bill Chambers e Matei Zaharia, Oreilly & Associates Inc, ISBN 978-1491912218, € 45,00. Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and The Cloud, 2019, di Paul Deitel & Harvey Deitel, Pearson, ISBN 978-0135404676, € 80,00. Lecture notes

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.
4	Software architectures for Big Data: l'algoritmo MapReduce, Apache Hadoop, HDFS
4	Software architectures for Big Data: Apache Spark and its libraries.

Modello 3 – Scheda di trasparenza proposta

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