SYLLABUS
Academic year 2024-2025

Dean,
Prof. dr. eng. Vasile-Ion Manta

1. Program data

1.1 Higher education institution	"Gheorghe Asachi" Technical University of Iași
1.2 Faculty	Automatic Control and Computer Engineering
1.3 Department	Computers
1.4 Field of studies	Computers and Information Technology
1.5 The cycle of studies[1]	Master
1.6 Study program	Artificial Intelligence

2. Subject data

2.1 Name of the subject / Code				**Big Data Techniques** *(Tehnici Big Data)* / **AI.113**			
2.2 Course coordinator				Lect. dr. eng. Alexandru Archip / Lect. dr. eng. Marius Gavrilescu			
2.3 Application instructor				Lect. dr. eng. Alexandru Archip / Lect. dr. eng. Marius Gavrilescu			
2.4 Year of study[2]	1	2.5 Semester[3]	2	2.6 Type of assessment[4]	exam	2.7 Type of subject[5]	DS

3. Estimated total time of daily activities (hours per semester)

3.1 Number of hours per week	4	3.2 lectures	2	3.3a sem.		3.3b laboratory	2	3.3c project	
3.4 Total hours in curriculum[6]	56	3.5 lectures	28	3.6a sem.		3.6b laboratory	28	3.6c project	

Distribution of the time fund[7]	No. hours
Study by textbook, course support, bibliography and notes	35
Additional documentation in the library, on specialist electronic platforms and in the field	35
Preparation of seminars/labs/projects, assignments, reports and portfolios	20
Tutorial[8]	
Examinations[9]	4
Other activities:	

3.7 Total hours of individual study[10]	94
3.8 Total hours per semester[11]	150
3.9 Number of credits	6

4. Prerequisites (where applicable)

4.1 curriculum[12]	
4.2 competences	

5. Conditions (where applicable)

5.1 conducting the lectures[13]	● Blackboard, video projector
5.2 conducting the seminar / laboratory / project[14]	● Laboratory room with computers and Internet access ● IntelliJ/PyCharm or similar IDE (academic license) for Java/Python ● Access to a Hadoop/Spark cluster (optional).

[1] *Bachelor / Master*
[2] *1-4 for Bachelor's, 1-2 for Master's*
[3] *1-8 for Bachelors, 1-3 for Masters*
[4] *Exam, colloquium or VP A/R – from the curriculum*
[5] *DF - fundamental subject, DID - subject in the field, DS - specialized subject or DC - complementary subject - from the education plan*
[6] *It is equal to 14 weeksx number of hours from point 3.1 (similar for 3.5, 3.6abc)*
[7] *The lines below refer to the individual study; the total is completed at point 3.7.*
[8] *Between 7 and 14 hours*
[9] *Between 2 and 6 hours*
[10] *The sum of the values on the previous lines, which refer to the individual study.*
[11] *The sum of the number of hours of direct teaching activity (3.4) and the number of hours of individual study (3.7); must be equal to the number of credits allocated to the subject (point 3.9)x 24 hours per credit.*
[12] *Mention the subjects that must be passed previously or equivalent*
[13] *Blackboard, video projector, flipchart, specific teaching materials, etc.*
[14] *Computing technique, software packages, experimental stands, etc.*

6. Specific competences accumulated[15]

Number of credits assigned to the subject[16]:		**6**	Distribution of credits per competences[17]
Professional competences	CP1	Knowledge of advanced concepts of computer science and information technology and the ability to work with these concepts.	1.1
	CP2	Scientific and practical research in the field of artificial intelligence.	1.1
	CP3	Design and development of artificial intelligence systems.	1.1
	CP4	Problem solving using artificial intelligence methods and techniques.	1
	CP5	Utilization of artificial intelligence tools and technologies.	1
	CP6		
	CPS1		
	CPS2		
Transversal competences	CT1	Legislation compliant application of the intellectual property rights and of the principles, norms and values of the professional ethics code within their own strategies for rigorous, effective and responsible work.	0.1
	CT2	Application of communication techniques and effective group work; developing empathic interpersonal communication skills and assuming leadership roles/functions in a multi-specialized team.	0.3
	CT3	Creating opportunities for continuous training and the effective utilization of learning resources and techniques for personal development.	0.3
	CTS		

7. Objectives of the subject (resulting from the grid of specific competences accumulated)

7.1 General objective of the subject	The course is designed to provide students with the skills and knowledge needed to develop applications and techniques for the processing of large-scale data sets. The course covers a wide array of topics, including MapReduce techniques, data acquisition and retrieval, algorithm design for big data processing, pattern identification, statistical data processing and data analytics. Additional related areas of interest covered by the course are data compression, clustering and classification methods, stream processing, and big data visualization.
7.2 Specific objectives	The goal of this course is to familiarize the student with Big Data techniques and particularities of analyzing large volumes of data. Specific study topics include: - batch (MapReduce) and stream processing techniques; - information retrieval within large volumes of data; - algorithm design approaches for large volumes of data; - pattern analysis and association rule mining; - statistical data analysis concepts; - clustering and classification in Big Data; - data compression techniques.

8. Contents

8.1 Course[18]	Teaching methods[19]	Remarks
1. Introduction to Big Data (2h) Key concepts; formal definitions; examples of applicability and necessity; cloud computing in Big Data; batch processing basics. ***2. MapReduce Fundamentals (2h)*** Definitions; fundamental concepts; data flow within MapReduce applications; fundamental stages: map, reduce; designing a MapReduce application; Hadoop and Spark frameworks. ***3. Search Engines I (2h)*** Indexing techniques: definitions and fundamental concepts; indexing and specific data structures; processing collections of text documents; index	Lectures via Powerpoint presentations, explanations and open discussions with students	

[15] *Competencies from the G1 and G1bis Grids of the study program, adapted to the specifics of the subject, for which credits are allocated (www.rncis.ro or the faculty website)*

[16] *From the education plan*

[17] *The credits allocated to the subject are distributed on professional and transversal competences according to the specifics of the subject*

[18] *Chapter and paragraph headings*

[19] *Exposition, lecture, blackboard presentation of the studied issue, use of video projector, discussions with students (for each chapter, if applicable)*

compression techniques; applying MapReduce in indexing. ***4. Search Engines II (2h)*** Techniques for finding information in massive data; definitions and fundamental concepts; information retrieval models: boolean, vector, n-gram techniques; AI-based retrieval techniques; applying MapReduce for search problems. **5. *Graph Analysis (2h)*** Methods and techniques specific to MapReduce; definitions and fundamental concepts; specific data structures: arrays and adjacency lists; BFS traversal; minimum paths; MapReduce approaches; the PageRank algorithm. ***6. Frequent Patterns and Association Rules (2h)*** Definitions and fundamental concepts; specific metrics; classic sequential algorithms: Apriori, FP-Growth; parallelization methods and challenges; MapReduce approaches. ***7. Big Data Applications in Research and for Real-world Problems (2h)*** Discussion on the topics studied so far; recap of key concepts and methods; research ideas and directions - smart IDPS solutions. ***8. Dimensionality in Big Data (2h)*** Challenges of large, high-dimensional data sets; variation and correlation; data projection and reconstruction; dimensionality reduction of large data sets; case study: PCA + Eigenfaces. ***9. Clustering of Large Data (2h)*** Similarity metrics; popular clustering methods applied for large data sets (K-Means, hierarchical clustering, DBSCAN); clustering evaluation metrics; effects of data dimensionality and size on clustering quality. ***10. Supervised Learning from Large Data (2h)*** Problem spaces, feature extraction and prioritization; linear classifiers, K-Nearest Neighbors applied to big data; data partitioning strategies, analysis of decision regions; performance considerations and challenges. ***11. Statistics for Big Data (2h)*** Recap of fundamental concepts: random variables, probability distributions, empirical and parametric distributions; distribution fitting; evaluating goodness-of-fit; Monte Carlo approaches in Big Data modeling; uncertainty and sensitivity analysis; challenges when working with large data sets. ***12. Data Compression (2h)*** Lossless and lossy compression; Run-Length Encoding, Huffman, LZ compression; discrete cosine transform, MPEG and JPEG compression. Evaluation and performance of compression methods; data distortion of lossy compression methods. ***13. Stream Processing (2h)*** Challenges of very large data streams; advantages and disadvantages over batch processing; stream sources; data normalization and standardization; stream processing models; windowing and buffering techniques; sampling methods for data streams. ***14. Visualization of Big Data (2h)*** Fundamentals; visual data representations; information visualization models; visual encodings; parallel coordinates; radar charts; heatmaps; Euler diagrams; radial sets; NodeTrix and network-oriented visualization; word clouds, word trees and topic streams.		

Course references:

[1] Sivarajah U, Kamal M, Irani Z and Weerakkody V 2016 Critical analysis of Big Data challenges and analytical methods J. Bus. Res. 70 263–86.
[2] K.U J and M.David J 2014 Issues, Challenges and Solutions : Big Data Mining CS IT-CSCP 4 131–40.

[3] Yaqoob I, Hashem I A T, Gani A, Mokhtar S, Ahmed E, Anuar N B and Vasilakos A V 2016 Big data: from beginning to future Int. J. Inf. Manage. 36 1231–47
[4] Gandomi A and Haider M 2015 Beyond the hype: Big data concepts, methods, and analytics Int. J. Inf. Manage. 35 137–44
[5] Hariri R H, Fredericks E M and Bowers K M 2019 Uncertainty in big data analytics: survey, opportunities, and challenges J. Big Data 6 44
[6] Martin A 2011 a Framework for Business Intelligence Application Using Ontological Classification Int. J. Eng. Sci. Technol. 3 1213–21
[7] Ma C, Zhang H H and Wang X 2014 Machine learning for Big Data analytics in plants Trends Plant Sci. 19 798–808
[8] Tsai C-W, Lai C-F, Chao H-C and Vasilakos A V 2015 Big data analytics: a survey J. Big Data 2 21
[9] Galetsi P, Katsaliaki K and Kumar S 2020 Big data analytics in health sector: Theoretical framework, techniques and prospects Int. J. Inf. Manage. 50 206–16
[10] Khan N, Yaqoob I, Hashem I, Inayat Z, Kamaleldin W, Alam M, Shiraz M and Gani A 2014 Big data: survey, technologies, opportunities, and challenges Sci. World J. 2014
[11] Saha P, Mittal M, Gupta S and Sharawi M 2017 Big Data trends and analytics: A survey Int. J. Comput. Appl. 180 9–20
[12] Shah T, Rabhi F and Ray P 2015 IInvestigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health condi Cluster Comput. 18 351–67
[13] Schroeck M, Shockley R, Smart J, Romero Morales D and Tufano P Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report IBM Inst. Bus. Value Said Bus. Sch. Univ. Oxford
[14] Khan N, Alsaqer M, Shah H, Badsha G, Abbasi A and Salehian S 2018 The 10 Vs, issues and challenges of big data Proceedings of the 2018 International Conference on Big Data and Education (New York, NY, USA: Association for Computing Machinery) pp 52–6
[15] Nair L R, Shetty S D and Shetty S D 2018 Applying spark based machine learning model on streaming big data for health status prediction Comput. Electr. Eng. 65 393–9
[16] García S, Ramírez-Gallego S, Luengo J, Benítez J M and Herrera F 2016 Big data preprocessing: methods and prospects Big Data Anal. 1
[17] Hashem I A T, Yaqoob I, Anuar N B, Mokhtar S, Gani A and Ullah Khan S 2015 The rise of "big data" on cloud computing: Review and open research issues Inf. Syst. 47 98–115
[18] Dritsas E, Livieris I E, Giotopoulos K and Theodorakopoulos L 2018 An Apache Spark implementation for graph-based hashtag sentiment classification on Twitter Proceedings of the 22nd Pan-Hellenic Conference on Informatics pp 255–60
[19] Oo, Myat Cho Mon and Thein T 2019 An efficient predictive analytics system for high dimensional big data J. King Saud Univ. - Comput. Inf. Sci.
[20] Murdoch T B and Detsky A S 2013 The inevitable application of big data to health care J. Am. Med. Assoc. 309 1351–2
[21] Sin K and Muthu L 2015 "Application of big data in education data mining and learning analytics – a literature review " ICTACT J. Soft Comput. 05 1035–49
[22] Wolfert S, Ge L, Verdouw C and Bogaardt M-J 2017 Big Data in smart farming – A review Agric. Syst. 153 69–80.
[23] Jiawei Han and Micheline Kamber, Data Mining - Concepts and Techniques, Second Edition, The Morgan Kaufmann Series in Data Management Systems, Ed. Morgan Kaufmann, 2006
[24] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press (http://nlp.stanford.edu/IR-book/), 2009
[25] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified data processing on large clusters. In OSDI'04: Sixth Symposium on Operating System Design and Implementation, pages 137–150, San Francisco, CA, 2004.
[26] Jing Zhang, Gongqing Wu, Xuegang Hu, Shiying Li and Shuilong Hao, A Parallel Clustering Algorithm with MPI – MKmeans, JOURNAL OF COMPUTERS, VOL. 8, NO. 1, JANUARY 2013
[27] Martin Ester, Hans-Peter Kriegl, Jorg Sandr, Xiaowei XU, A Density-Based Algorithm for Discovering Cluster in Large Spatial Databases with Noise, KDD-96 Proceedings, 1996
[28] G. Onofrei, A. Archip, Achieving better recommendations with overclassification: Practical considerations, Proceedings of the 20th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, România, p. 410-416 (2016)
[29] A. Archip, M. Craus, The Fast Itemset Miner: A detailed analysis of the candidate generation stages, 15th International Conference on System Theory, Control, and Computing (ICSTCC), Sinaia, Romania, p. 1 - 5, IEEE (2011)
[30] M.I. Astratiei, A. Archip, A Case Study on Improving the Performance of Text Classifiers, 14th International Conference on System Theory, Control, and Computing (ICSTCC), Sinaia, Romania, p. 37 - 42, IEEE (2010)
[31] A. Archip, V. Manta, G. Danilet, Parallel K-Means Revisited: A Hypercube Approach, 14th International Conference on System Theory, Control, and Computing (ICSTCC), Sinaia, Romania, p. 43 - 48, IEEE (2010)
[32] Dell Zhang, Cosmin Stamate, Cloud Computing – lecture notes (MSc. studies), Birkbeck University of London

8.2a Seminar	Teaching methods[20]	Remarks

[20] *Discussions, debates, presentation and/or analysis of papers, solving exercises and problems*

8.2b Laboratory	Teaching methods[21]	Remarks
1. Introduction to MapReduce (1) (2h) Simple examples meant to present the main stages of MapReduce and the main differences from classic *map* and *reduce* primitives. 2. ***Introduction to MapReduce (2) (2h)*** Hadoop and Spark fundamentals. Case study: matrix multiplication. ***3. Text indexing techniques (2h)*** Information retrieval within large text collections. Inverted indexing and postings. Index types: boolean, quantitative and positional indexes. Index compression techniques. ***4. Text based information retrieval (2h)*** Boolean and vector-based search techniques. ***5. Graph analysis techniques (2h)*** BFS graph traversal. BFS extensions for Dijkstra and PageRank ***6. Frequent patterns and association rule mining (2h)*** Apriori based approaches for batch processing. ***7. Open discussions on study topics. Research use cases and applicability of studied concepts (2h)*** ***8. Dimensionality reduction (2h)*** Study of variance/covariance. Principal component analysis (PCA). Data projection and reconstruction using PCA. ***9. Clustering in Big Data (2h)*** K-Means clustering algorithm. Impact of dimensionality reduction of clustering results. Cluster quality assessment. ***10. KNN decision regions (2h)*** KNN – unweighted vs. weighted approaches. KNN on large data sets: determination and visualization of decision regions. Study of the impact of hyperparameters on decision region shape. ***11. Statistical data analysis (2h)*** Exercises using uniform- and normally-distributed data. Parametric distribution fitting of large numerical data sets. Simple applications using Monte Carlo techniques. ***12. Data compression (2h)*** RLE/LZ compression/decompression of strings. Huffman compression/decompression for binary values. DCT compression. Evaluation of compression techniques. ***13. Stream processing (2h)*** Application of various stream processing techniques using simple mathematical/statistical methods on numeric data streams. Impact of various buffer sizes. Techniques for sampling from streams.	General and individual explanations, individual computer work.	
8.2c Project	Teaching methods[22]	Remarks

[21] *Practical demonstration, exercise, experiment*

[22] *Case study, demonstration, exercise, error analysis, etc.*

Applications (laboratory / project) references:

See "Course references"

9. Corroboration of the contents of the subject with the expectations of representatives of the epistemic community, professional associations and representative employers in the field related to the program[23]

The included study topics aim to familiarize the students with the field of Big Data and its impact on Artificial Intelligence/Machine Learning (AI/ML) techniques. The significance of the course's theoretical and practical concepts is emphasized by its various applications in Web data analysis, Internet-of-Things and Cybersecurity. Furthermore, the amount of data required to build high quality AI/ML training models enforces the necessity of Big Data analytical techniques. Similar courses can be found within the curricula of university master's studies, both nationally and internationally.

10. Evaluation

Type of activity	10.1 Evaluation criteria	10.2 Evaluation methods		10.3 Weight in the final grade
10.4a Exam	Acquired theoretical and practical knowledge (quantity, correctness, accuracy)	Periodic tests[24]:		50% (minimum 5)
		Homework:		
		Other activities[25]:		
		Final evaluation:	100%	
10.4b Seminar				
10.4c Laboratory	Knowledge of related techniques, ability to use dedicated frameworks, evaluation and interpretation of results	● Practical demonstrations ● Oral answers ● Written questionnaires		50% (minimum 5)
10.4d Project				
10.5 Minimum performance standard[26]: grade 5 in the exam and applications (the average between laboratory and project)				

Date of completion,
5 December 2023

Signature of course coordinator,
Lect. dr. eng. Alexandru Archip

Lect. dr. eng. Marius Gavrilescu

Signature of application instructor,
Lect. dr. eng. Alexandru Archip

Lect. dr. eng. Marius Gavrilescu

Date of approval in the department,
7 December 2023

Director of department,
Assoc. prof. dr. eng. Andrei Stan

[23] *The connection with other subjects, the usefulness of the subject on the labor market*

[24] *The number of tests and the weeks in which they will be held will be specified.*

[25] *Scientific circles, professional competitions, etc.*

[26] *The minimum performance standard from the competences grid of the study program is customized to the specifics of the subject, if applicable.*