SYLLABUS

1. Information regarding the programme

1.1 Higher education institution	Babeş-Bolyai University of Cluj-Napoca
1.2 Faculty	Faculty of Mathematics and Computer Science
1.3 Department	Department of Computer Science
1.4 Field of study	Computer Science
1.5 Study cycle	Bachelor of Science
1.6 Study programme / Qualification	Artificial Intelligence

2. Information regarding the discipline

2.1 Name of the discipline (en) (ro)			Introduction to Big Data				
2.2 Course coordinator			Lect. Dr. Ioana-Georgiana Ciuciu				
2.3 Seminar coordinator			Lect. Dr. Ioana-Georgiana Ciuciu				
2.4. Year of study	2	2.5 Semester	4	2.6. Type of evaluation	C	2.7 Type of discipline	Compulsory
2.8 Code of the discipline		MLE5203					

3. Total estimated time (hours/semester of didactic activities)

3.1 Hours per week	4	Of which: 3.2 course	2	3.3 seminar/laboratory	1/1
3.4 Total hours in the curriculum	56	Of which: 3.5 course	28	3.6 seminar/laboratory	14/14
Time allotment:					hours
Learning using manual, course support, bibliography, course notes					24
Additional documentation (in libraries, on electronic platforms, field documentation)					24
Preparation for seminars/labs, homework, papers, portfolios and essays					32
Tutorship					6
Evaluations					8
Other activities:					-

3.7 Total individual study hours	94
3.8 Total hours per semester	150
3.9 Number of ECTS credits	6

4. Prerequisites (if necessary)

4.1. curriculum	
4.2. competencies	• Basic knowledge of data analytics, preferably

	• Basic knowledge of data visualization, preferably • Programming skills

5. Conditions (if necessary)

5.1. for the course	• Room with video projector
5.2. for the seminar /lab activities	• Room with computers as needed; • Big Data software installed • High level programming language environment

6. Specific competencies acquired

Professional competencies	**Knowledge and understanding** • Understanding the specificities of Big Data • Knowing the main Big Data sources and the main methods to store and process these data • Understanding the role of Big Data analysis in various domains **Explanation and interpretation** • Explaining decisions using complex models based on Big Data • Interpreting the results of the Big Data analysis **Instrumental - applicative** • Using non-traditional databases for the storage and processing of large volumes of data • Advanced querrying over distributed information resources • Evaluation, testing and validation with real-world data **Attitude** • Manifesting an open attitude towards the contributions of Big Data and the underlying technologies in a multitude of domains
Transversal competencies	• Methods and algorithms for data processing and analysis applied to Big Data • Multidisciplinary competencies spanning various application sectors (e.g., life sciences and bioinformatics, telco, media, finance, security, health, energy, etc.) • Data Science competencies, combining data analyst and data engineer- specific competencies (e.g., competencies from the fields of mathematics, statistics, information science, computer science, databases, machine learning, data mining, visualization, etc.)

7. Objectives of the discipline (outcome of the acquired competencies)

7.1 General objective of the discipline	• Handling (extremely) large amounts of digital data in various formats (text, video, financial, medical, etc.)
7.2 Specific objective of the discipline	• Enable the use of novel algorithms, software infrastructures and methodologies for the purpose of processing (store, retrieve, analyze) large amounts of data • Provide decision support over large volumes of data • Enable the creation of applications and services for various business domains based on the results of big data analysis.

8. Content

8.1 Course	Teaching methods	Remarks
1. Introduction to Data Science and Big Data – part I	Exposure, description, explanation, examples, case studies	Data Science main concepts, the Data Science Process, examples, case studies
2. Introduction to Data Science and Big Data – part II	Exposure, description, explanation, examples, case studies	Data types, data sources, data availability, main Big Data concepts, Big Data emerging technologies, case study examples
3. Industrial Standards for Data Mining Projects	Exposure, description, explanation, examples, case studies	Methodology for Data Science projects (CRISP-DM)
4. Big Data Architecture - part I	Exposure, description, explanation, examples, case studies	Traditional database systems versus Big Data systems
5. Big Data architecture – part II	Exposure, description, explanation, examples, case studies	The Lambda Architecture - a model for building a Big Data system, case studies and examples
6. Batch processing (Batch Layer) – part I	Exposure, description, explanation, examples, case studies	Big Data storage, data model for Big Data, batch computing, the Hadoop Ecosystem, Batch processing, technologies
7. Batch processing (Batch Layer) - part II	Exposure, description, explanation, examples, case studies	
8. Speed processing (Speed Layer) - part I	Exposure, description, explanation, examples, case studies	Computing and storing of real time views, real time updates, tools
9. Speed processing (Speed Layer) - part II	Exposure, description, explanation, examples, case studies	
10. Data Ingestion	Exposure, description, explanation, examples, case studies	Definitions and design considerations, batch ingestion, real time ingestion, tools

11. NoSQL Solutions for Big Data	Exposure, description, explanation, examples, case studies	NoSQL databases, NoSQL Data Models Tutorial provided
12. Data Visualization	Exposure, description, explanation, examples, case studies	Scientific data visualization principles, examples, technologies
13. Big Data Case Studies	Exposure, description, explanation, examples, case studies	Presentation of Big Data real-world case studies
14. Big Data Research Essays Presentation	Exposure, description, explanation, examples, case studies	Student essay presentation

Bibliography

Marz, N., & Warren, J. (2015). *Big Data. Principles and Best Practices of scalable real-time systems*. Manning Publications

Cielen, D., Meysman, A.D.B., & Ali, M. (2016). *Introducing Data Science. Big Data, machine learning, and more, using Python tools*. Manning Publications

Grus, J. (2019). *Data Science from Scratch: First Principles with Python*. O'Reilly Media, Inc.

Damji, J.S., Wenig, B., Das, T., & Lee, D. (2020). *Learning Spark*. O'Reilly Media, Inc.

Sadalage, P., Fowler, M. (2013). *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Pearson Education, Inc.

Agneeswaran, V. (2014). *Big Data Analytics Beyond Hadoop*. Pearson Education

White, T. (2009). *Hadoop: The Definitive Guide*. O'Reilly

McCallum, Q. E. (2012). *Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work*. O'Reilly

8.2 Seminar / laboratory	Teaching methods	Remarks
Seminar	Solving Big Data processing related problems	The seminar takes place every two weeks and takes two hours
1. Introduction in Data Science and Big Data	• Knowledge synthesis • Conceptual clarification • Group activities • Practical activities	Illustration of the main concepts of Data Science Examples and exercises related to the Data Science process and applied to various domains Approaching the challenges related to data in various domains using Big Data

		Exercises aimed at practicing the work with various data types
2. Big Data Architecture	• Group activities • Guided discovery • Practical activities	Modeling a Big Data system based on the Lambda Architecture
3. Batch storage and processing of data	• Group activities • Practical activities	Realizing batch storage and processing operations The seminar is organised as a tutorial
4. Real-time storage and processing of data	• Group activities • Practical activities	Realizing real-time storage and processing operations The seminar is organised as a tutorial
5. Data ingestion	• Group activities • Practical activities	Exemplifying data ingestion The seminar is organised as a tutorial
6. NoSQL databases	• Group activities • Practical activities	Working with NoSQL databases The seminar is organised as a tutorial
7. Data visualization	• Group activities • Practical activities	Basic data visualization operations The seminar is organised as a tutorial
Laboratory		
The laboratory will be organized as a semester project with groups of about 2-3 students (depending on the requirements and the equipment needed) Team work will be autonomous (focus on creativity and critical thinking) Technical tutorials will be provided to support student work around the most important aspects of Big Data storage and processing (e.g., Hadoop shell, PySpark, Data Ingestion with Apache Sqoop, NoSQL, etc.)	• Research-informed Learning • Tutorial-based • Problem-solving approach • Team work • Big Data solutions for concrete problems and case studies	The laboratory takes place every two weeks and takes two hours Groups will be monitored via a project wiki managed with the course/lab the responsible

Bibliography

1. http://mahout.apache.org/
2. http://www.tutorialspoint.com/mahout/mahout_introduction.htm
3. http://spark.apache.org/documentation.html
4. http://shark.cs.berkeley.edu/
5. http://spark.apache.org/
6. http://nosql-database.org/
7. https://www.mongodb.com/nosql-explained

9. Corroborating the content of the discipline with the expectations of the epistemic community, professional associations and representative employers within the field of the program

- Synergies with various local and EU initiatives: local industry, national institutions, European Data Science Academy (EDSA, https://edsa-project.eu/),EU projects such as Big Data for Next Generation Energy (BD4NRG, https://www.bd4nrg.eu/), LETHE (https://cordis.europa.eu/project/id/101017405), FARE (https://cordis.europa.eu/project/id/853566), the Human Brain Project (https://www.humanbrainproject.eu/en/), SoBigData (http://project.sobigdata.eu/), etc.
- Collaboration with the IT industry: invited lectures with real-life use cases, semester project topics, equipment (e.g., smart sensors).

10. Evaluation

Type of activity	10.1 Evaluation criteria	10.2 Evaluation methods	10.3 Share in the grade (%)
10.4 Course	- to be familiar with the main concepts of the domain -to be able to model a problem from a specific application field relying on emergent Big Data technologies - to be able to apply these principles in real-life use cases	Evaluation of a research essay	50%
10.5 Seminar/lab activities	- to be able to propose viable creative solutions to real-life big data challenges from various application domains - to be able to consume (query, analyze)Big Data in order to derive information relevant to use cases from various application domains - to demonstrate critical thinking - to successfully perform	Semester project	50%

	individual and team-based tasks		

10.6 Minimum performance standards

- A minimum grade of 5 (on a scale from 1 to 10) is necessary for the final grade (average between the research essay and the semester project)
- The lab attendance is compulsory at a rate of 90%, according to the decision of the Computer Science Department Council (http://www.cs.ubbcluj.ro/wp-content/uploads/Hotarare-CDI-15.03.2017.pdf)

Date	Signature of course coordinator	Signature of seminar coordinator
...25 April 2023.....	Lect. Dr. Ioana-Georgiana Ciuciu	Lect. Dr. Ioana-Georgiana Ciuciu

Date of approval

.....

Signature of the head of department

Prof. Dr. Laura Diosan