SYLLABUS

Big Data Processing and Applications

University year 2025-2026

1. Information regarding the programme

1.1. Higher education institution	Babeş-Bolyai University of Cluj-Napoca
1.2. Faculty	Faculty of Biology and Geology
1.3. Department	Department of Molecular Biology and Biotechnology
1.4. Field of study	Biology
1.5. Study cycle	Master
1.6. Study programme/Qualification	Bioinformatics
1.7. Form of education	Full time

2. Information regarding the discipline

2.1. Name of the discipline	**Big Data Processing and Applications**					Discipline code	**MME8158**
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	3	2.6. Type of evaluation	E	2.7. Discipline regime	Compulsory

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

3.1. Hours per week	**3**	of which: 3.2 course	**2**	3.3 seminar/laboratory/project	**2**
3.4. Total hours in the curriculum	56	of which: 3.5 course	28	3.6 seminar/laboratory/project	**28**

Time allotment for individual study (ID) and self-study activities (SA)	**hours**
Learning using manual, course support, bibliography, course notes (SA)	42
Additional documentation (in libraries, on electronic platforms, field documentation)	42
Preparation for seminars/labs, homework, papers, portfolios and essays	41
Tutorship	10
Evaluations	4
Other activities: bi-directional communication with the course responsible	5
3.7. Total individual study hours	**144**
3.8. Total hours per semester	**200**
3.9. Number of ECTS credits	**8**

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

6.1. Specific competencies acquired [1]

[1] One can choose either competences or learning outcomes, or both. If only one option is chosen, the row related to the other option will be deleted, and the kept one will be numbered 6.

Professional/essential competencies	• Capability of developing of high performance programs based on parallel and distributed programming • Efficient modeling and solving real-life problems;
Transversal competencies	• Team work capabilities; able to fulfill different roles; • Professional communication skills; concise and precise description, both oral and written, of professional results, negociation abilities;

6.2. Learning outcomes

Knowledge	• The student knows how to handle (extremely) large amounts of digital data in various formats (text, video, financial, medical, etc.) • The student knows the key concepts of parallel cluster architectures • The student acquires the fundamental knowledge that allows parallelizing and solving large and complex problems on scalable systems
Skills	• The student is able to use novel algorithms, software infrastructures and methodologies for the purpose of processing (store, retrieve, analyze) large amounts of data • The student is able to develop applications and services for various business domains based on the results of big data analysis
Responsibility and autonomy:	• The student manages a workflow and interacts inside a team, makes decisions and manages unforeseen situations, develops creative ideas and innovative techniques • The student knows and follows ethical and deontological norms and rules in scientific research • The student develops the ability to translate academic knowledge into a professional, economic, social and ethical context. • The student uses efficient strategies, methods and techniques for lifelong education, in order to self educate and self develop his/her personal and professional skills

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 (retrieve, store, 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	Exposure, description, explanation, examples, case studies	Data Science main concepts, the Data Science Process, challenges, data availability, data types, tools
2. Industrial Standards for Data Mining Projects	Exposure, description, explanation, examples, case studies	Methodology for Data Science projects (CRISP-DM)
3. Big Data Architecture	Exposure, description, explanation, examples, case studies	Traditional database systems versus Big Data systems, the Lambda Architecture, a model for building a Big Data system, case studies and examples
4. Batch Layer	Exposure, description, explanation, examples, case studies	Big Data storage, data model for Big Data, batch computing, the Hadoop Ecosystem
5. Serving Layer	Exposure, description, explanation, examples, case studies	Requirements, performance metrics, the normalization/denormalization problem, tools
6. Spark for data processing - part I	Exposure, description, explanation, examples, case studies	Batch data processing using Apache Spark. Examples
7. Speed Layer - part I	Exposure, description, explanation, examples, case studies	Computing and storing of real time views, real time updates, tools
8. Spark for data processing - part II	Exposure, description, explanation, examples, case studies	Real-time data processing using Apache Spark. Examples
9. Data Ingestion	Exposure, description, explanation, examples, case studies	Definitions and design considerations, batch ingestion, real time ingestion, tools
10. NoSQL Solutions for Big Data	Exposure, description, explanation, examples, case studies	NoSQL databases, NoSQL Data Models Tutorial provided
11. Ethical Challenges Related to Big Data	Exposure, description, explanation, examples, case studies	Challenges in developing and using big data applications including (i) security and privacy of data; (ii) algorithmic bias and fairness; (iii) transparency and; and (iv) social and ethical implications
12. Big Data Case Studies	Exposure, description, explanation, examples, case studies	Presentation of Big Data (industrial) case studies
13. Big Data Research Essays Presentation	Exposure, description, explanation, examples, case studies	Student essay presentation
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.

Zečević, P., Bonaći, M. (2017). *Spark in Action*, Manning Publications

Perrin, J.G. (2020). *Spark in Action*, 2nd Ed., Manning Publications

Zelenin, A., Kropp, A. (2025). *Apache Kafka in Action*, Manning Publications

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

Banker, K., Bakkum, P., Verch, S., Garrett, D. and Hawkins, T. (2016). *MongoDB in Action*, Second Edition, Manning Publications

Borucki, A. (2024). *MongoDB in Action*, 3rd Ed., Manning Publications

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

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

Holmes, A. (2015). *Hadoop in Practice*, 2nd ed., Manning Publications

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

Grigorev, A. (2021). *Machine Learning Bookcamp*, Manning Publications

Rioux, J. (2022). *Data Analysis with Python and PySpark*, Manning Publications

Khalil, M. (2025). *Effective Data Analysis*, Manning Publications

8.2 Seminar / laboratory	Teaching methods	Remarks
Semester project organized 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	Groups will be monitored via a project team (in MS Teams) managed with the course/seminar responsible The seminar takes place every two weeks and takes two hours

Bibliography (same as for the course)

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 European Federation of Data Driven Innovation Hubs (EUHubs4Data, https://euhubs4data.eu/), Incubator of Trusted B2B Data Sharing ecosystems of collaborating SMEs linked to Digital Innovation Hubs (i4Trust, https://i4trust.org/), REACH EuRopEAn incubator for trusted and secure data value Chains (https://www.reach-incubator.eu/), 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).
- Collaboration with other study programs (e.g., Bioinformatics Master from the Faculty of Biology) around the semester project or with students and professors from other faculties and universities (e.g, collaborative projects, invited courses, etc.)

10. Evaluation

Activity type	10.1 Evaluation criteria	10.2 Evaluation methods	10.3 Percentage of final 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	Written exam/ Evaluation of a research essay	50%
10.5 Seminar/laboratory	- 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 individual and team-based tasks	Semester project	50%
10.6 Minimum standard of performance			

- A minimum grade of 5 (on a scale from 1 to 10) is necessary for the written exam, the practical work and the research essay
- 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)

11. Labels ODD (Sustainable Development Goals)[2]

Not applicable.

Date: 15 April 2025	Signature of course coordinator Assist.Prof. PhD. Ioana CIUCIU	Signature of seminar coordinator Assist.Prof. PhD. Ioana CIUCIU
Date of approval: ...		Signature of the head of department Assoc.prof.phd. Adrian STERCA

[2] Keep only the labels that, according to the *Procedure for applying ODD labels in the academic process*, suit the discipline and delete the others, including the general one for *Sustainable Development* – if not applicable. If no label describes the discipline, delete them all and write „*Not applicable*.".