COURSE DESCRIPTION

1. Information about the programme

1.1 Institution of higher education	**Alexandru Ioan Cuza University of Iasi**
1.2 Faculty	**Faculty of Economics and Business Administration**
1.3 Department	**Department of Accounting, Information Systems and Statistics**
1.4 Field of study	**Business Informatics**
1.5 Level	**Master**
1.6 Study programme/ Qualification	**Software Development and Business Information Systems**

2. Information about the course

2.1 Course name	**Software Tools for Big Data and Machine Learning**						
2.2 Course coordinator	**Prof. Marin Fotache, Ph.D.**						
2.3 Seminar coordinators	**Marius-Iulian Cluci**						
2.4 Year of study	I	2.5 Semester	II	2.6 Type of assessment	P	2.7 Discipline status	**C**

* *C – Compulsory / E - Elective*

3. Total estimated time (hours alloted to didactic activity per semester)

3.1 Total number of hours per week	4	of which: 3.2 lecture	2	3.3 seminar/lab	2
3.4 Total number of hours in the curriculum	56	of which: 3.5 lecture	28	3.6 seminar/lab	28

Time distribution	hours
Study of the handbook, coursebook, bibliography and notes	30
Additional research in the library, online and on the field	14
Preparation of seminars/labs, homeworks and projects	30
Tutorials	14
Assessment	6
Other activities.....	

3.7 Total number of self-study hours	94
3.9 Total number of hours per semester	150
3. 10 Number of credits	6

4. Prerequisites (if applicable)

4.1 curriculum-based	• Database Logic in Business Applications
4.2 competence-based	• Not applicable

5. Conditions (if applicable)

5.1. for lectures	• Online scenario: Moodle + Microsoft Teams • On-site scenatio: Lecture rooms shall be provided with video projector ; Moodle platform for some tests • Big Data ang Machine Learning cluster (MongoDB, Neo4j), Hadoop, Apache Spark
5.2. for seminars/labs	• Big Data ang Machine Learning cluster (MongoDB, Neo4j), Hadoop, Apache Spark • R/RStudio installed on the personal computers

6. Assimilated specific competences

Professional competences	• C2.1 Mastering theoretical and technological knowledge and tools concerning business data modeling, query, processing, administration and analysis, including Big Data (1 credit) • C2.2 Selection and refinement of the methods and techniques for data modeling, persistence, query and analysis, according to the nature of problems and available resources (0.5 credits) • C2.3 Assess the degree of information integrity and validity for organizational data; find the appropriate tools for administration and analysis of business data (0.5 credits) • C2.4 Design the most appropriate solutions for gathering, storage, processing, administration and analysis of business data according to the organizational resources and constraints (0.5 credits) • C2.5 Develop projects and case-studies concerning modeling, implementation (database logic), administration and analysis of data for real-world applications (0.5 credits) • C4.1 Gaining detailed knowledge on all aspects of methodological and technological regarding the representation and persistence of data formats, the protocols and means of communication and integration of applications and services within distributed business information systems (0.5 credits) • C4.3 Identification of information sources, application modules and available services, both inside and outside the business system; estimate the solutions of their integration in order to match the current and future information needs of the organization (0.5 credits) • C4.4 Define the most appropriate solutions for data and modules integration, in order to meet the organizational requirements towards information integrity and security (0.5 credits) • C4.5 Write the specifications and deploy the modules regarding data, applications and services integration (0.5 credits)
Transversal competences	• CT1 – The ability to communicate and collaborate in teams of different professionals (0.5 credits) • CT3 – Continuous improvement of specific skills and knowledge towards approaching information systems, development of new software technologies and management of information systems (0.5 credits)

7. Discipline objectives (provided by the assimilated specific competences grid)

7.1 The general objective of the discipline	• To provide the core knowledge, methodologies and tools in order to model, store, process and analyze huge volumes of business data in various formats
7.2 Specific objectives	• Handle of various formats and technologies for data storage • Acquiring basic knowledge and skills for deploying Big Data and Machine Learning solutions • Application of high-level framework for data processing, analysis and machine learning in R (mainly the tidyverse and tidymodels ecosystems) • Getting the necessary knowledge and skills for data storage, query and processing in document databases (MongoDB) • Getting the necessary knowledge and skills for data storage, query and processing in graph databases (Neo4j) • Deploying Big Data solutions with the Hadoop ecosystem (incl. Hive) • Usage of big data processing frameworks (Spark) for Big Data processing (SparkSQL) and Machine Learninh (MLib)

	• R and Big Data (incl. the sparklyR package)

8. Content

8. 1 Subject	**Teaching methods**	**Observations**
Big Data and Machine Learning landscape in terms of data persistence, processing and analysis	PPT presentation, conversation	1 hour Marz & Warren (2014)
The **tidyverse** ecosystem for data processsing in R	PPT presentation, script writing and execution, discussion. Case studies.	4 hours Wickham & Grolemund (2022) Baumer et al. (2021) Ismay & Kim (2022) Fotache (2023a)
Exploratory Data Analysis with (mainly) the tidyverse	PPT presentation, script writing and execution, discussion. Case studies.	2 hours Wilke (2019) Staniak & Biecek, (2019) Wickham, 2022 Fotache (2023a)
High-level Machine Learning frameworks in R. The **tidymodels** ecosystem.	PPT presentation, script writing and execution, discussion. Case studies.	4 hours Fotache (2023a) Silge (2022) Kuhn & Silge (2022)
Distributed Architectures for SQL Servers. Case study: Citus (PostgreSQL)	PPT presentation, script writing and execution, discussion.	1 hour Cubukcu (2021) Fotache (2023b)
JSON Data Management in SQL Servers. PostgreSQL implementation	PPT presentation, script writing and execution, discussion.	2 hours Fotache (2023b)
Distributed Architectures for MongoDB	PPT presentation, script writing and execution, discussion.	1 hour Fotache (2022b)
Graph databases. Data model in Neo4j. Queries with Cypher language. Relational DB to Graph DB migration. Case study Distributed Architectures for Neo4j	PPT presentation, script writing and execution, discussion. Case studies.	4 hours Robinson et al. (2015) Panzarino (2014) Fotache (2021)
The ecosystem of Hadoop: HDFS, Map-Reduce; data processing with Hive	PPT presentation, explanation, conversation, questioning.	4 hours Du (2015) Hrubaru (2019)
Unified Big Data Processing Frameworks. Apache Spark. SparkSQL and MLLib.	PPT presentation, explanation, conversation, questioning.	5 hours Spark (2022) Javier & Kuo (2019)
Main References Banker K., Bakkum P., Verch S., Garrett D., Hawkins T. (2015). MongoDB in action (2nd edition), Manning Publications		

Baumer, B.S., Kaplan, D.T., Horton, N.J. (2021). Modern Data Science with R (2nd ed.). CRC Press (Taylor & Francis Group, LLC). disponibilă gratuit la adresa https://mdsr-book.github.io/mdsr2e/
Capriolo E., Wampler D., Rutherglen J. (2012). Programming Hive, O'Reilly
Copeland R. (2013). MongoDB Applied Design Patterns, O'Reilly
Cubukcu, U., Erdogan, O., Pathak, S., Sannakkayala, S., Slot, M. (2021). Citus: Distributed PostgreSQL for Data-Intensive Applications. In Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21). Association for Computing Machinery, New York, NY, USA, 2490–2502. https://doi.org/10.1145/3448016.345755 (Open Access)
Du D. (2015). Apache Hive Essentials, Packt Publishing
Fotache M. (2023a), Data Processing, Analysis and Science with R. Available on GitHub: (https://github.com/marinfotache/Data-Processing-Analysis-Science-with-R)
Fotache M. (2023b), Polyglot Persistence and Big Data, Available on GitHub: (https://github.com/marinfotache/Polyglot-Persistence-and-Big-Data)
Hrubaru, I. (2016). Polyglot persistence for business applications, Ph.D. Thesis, Univ. Al.I.Cuza Iași, 2016
Ismay, C. and Kim, A.Y. (2022). An Introduction to Statistical and Data Sciences via R. A ModernDive into R and the Tidyverse. Modern Drive. Disponibilă gratuit la adresa https://moderndive.com/index.html
Javier L. , Kuo K., Ruiz E. (2019). Mastering Spark with R: The Complete Guide to Large-Scale Analysis and Modeling, O'Reilly, freely available at: https://therinspark.com
Karau H., Warren R. (2017). High Performance Spark, O'Reilly Media, ISBN: 9781491943205
Kuhn, M and Silge, J. (2022). Tidy Modeling with R. O'Reilly. Disponibilă gratuit la adresa https://www.tmwr.org
Luranschi, J, Kuo, K., Ruiz, E. (2019). Mastering Spark with R: The Complete Guide to Large-Scale Analysis and Modeling. ISBN : 978-1492046370; freely available at: https://therinspark.com
Marz N., Warren J. (2014). Big Data. Principles and best practices of scalable realtime data systems, Manning Publications
MongoDB (2023). MongoDB Tutorials, https://docs.mongodb.com/manual/tutorial/
Panzarino O. (2014). Learning Cypher, Packt Publishing
Robinson I., Webber J., Eifren E. (2015). Graph Databases, O'Reilly, freely available at https://neo4j.com/graph-databases-book/
Silge, J., Julia Silge Keynote Wednesday (Introd to tidymodels) - https://www.youtube.com/watch?v=KFatBWeEecs
Sintonen, M. (2019). Scaling Our Analytical Processing Service: Sharding a PostgreSQL Database with Citus, https://www.smartly.io/blog/scaling-our-analytical-processing-service-sharding-a-postgresql-database-with-citus
Spark (2023). Apache Spark documentation, https://spark.apache.org/docs/latest/
Staniak, M., Biecek, P. (2019). The landscape of R packages for automated exploratory data analysis. arXiv preprint arXiv:1904.02101. https://arxiv.org/pdf/1904.02101.pdf
Webber, J., van Bruggen, R. (2020). Graph Databases For Dummies. Freely available at: https://neo4j.com/graph-databases-for-dummies/
White T. (2015). Hadoop. The Definitive Guide (4th Edition), O'Reilly
Wilke, C.O. (2019). Fundamentals of Data Visualization. O'Reilly. Freely available at: https://clauswilke.com/dataviz/
Wickham, H. (2022). ggplot2: elegant graphics for data analysis (3rd ed.), Springer. Freely available at: https://ggplot2-book.org/index.html
Wickham, H., Grolemund, G. (2022). R for Data Science, 2nd ed., O'Reilly, disponibilă gratuit la adresa https://r4ds.hadley.nz

8. 2 Seminar/lab	**Teaching methods**	**Observations**
Perform data processing tasks with the tidyverse ecosystem	Writing R code, discusssion. Case study	4 hours
Individual assessment 1 – tidyverse	Moodle	2 hours
Perform Exploratory Data Analysis with the tidyverse ecosystem	Writing R code, discusssion. Case study	2 hours
Build and tune Machine Learning models with the tidymodels ecosystem	Writing R code, discusssion. Case study	4 hours

Presentation of mini-project 1 (team): Data Processing and Machine Learning models using tidyverse and tidymodels	Prezentation of team solution	1 hour
Basic features on Open Stack Projects on Raa-IS (private cloud) platform. A distributed setup for MongoDB	Discussion, Scripts and code execution	2 hours
Practice CRUD operations in Neo4j	Discussion, scripts execution	2 hours
Practice database queries in Cypher (Neo4j)	Discussion, scripts execution	3 hours
Presentation of mini-project 2 (team): Relational DB -> Graph DB (Neo4j) – database design and queries		1 hour
Setup and deploy Hadoop clusters. CRUD and queries with Hive	Discussion, Scripts and code execution	3 hours
Big data processing and Machine Learning with Spark.	Discussion, Scripts and code execution	3 hours
Presentation: mini-project 3 (team): Hadoop + Spark	Discussion, Scripts and code execution, feedback, assessment	1 hour
Main references: Fotache M. (2023a), Data Processing, Analysis and Science with R. Available on GitHub: (https://github.com/marinfotache/Data-Processing-Analysis-Science-with-R) Fotache M. (2023b), Polyglot Persistence and Big Data, Available on GitHub: (https://github.com/marinfotache/Polyglot-Persistence-and-Big-Data)		

9. Corroboration of the course content with the expectations of epistemic community representatives, professional associations as well as of representative employers in the programme related field.

- The content of this discipline has been decided upon by taking into account both the curricula of some prestigious Western Universities and the demands of the economic environment provided by potential employers, either in the public or in the private IT companies.

10. Assessment

Type of activity	10.1 Assessment criteria	10.2 Assessment methods	10.3 Share of final grade
Individual assessment 1 – tidyverse	Finding appropriate solutions for a given set of requirements	Test on Moodle	20%
Mini-project 1: EDA and ML with tidyverse and tidymodels (team)	Solution quality	Presentation, code execution, discussion of each team’s solution	20%
Mini-project 2: Relational DB to Graph DB (team)	Quality of schema translation, query viability and finesse	Presentation, code execution, discussion of each team’s solution	20%
Mini-project 3: Hadoop/Hive + Spark (team)	Solution relevance and finesse	Presentation, code execution, discussion of each team’s solution	30%
Solving the ad-hoc problems/tests during lectures and labs	Solutions quality (& sometimes originality)	Assessment by course lecturer or lab coordinator	10%

10.6 Minimum performance standard

- Design and implement scripts and code modules for CRUD operations and queries in NoSQL data stores (MongoDB and Neo4j) and Hadoop/in-memory systems (Spark) and building predictive models with high-level machine learning frameworks in R.

Date of completion	Lecture Coordinator	Lab Coordinator
September 20, 2022	Prof. Marin Fotache	Marius-Iulian Cluci

Date of approval within the department

Head of Department
Prof. Florin Dumitriu, Ph.D.