DOCTORAL (PHD) STUDIES

COURSE UNIT DESCRIPTION

Course unit title	Scientific areas	Faculty	Institute, department
Big Data Analytics	Informatics engineering (T 007)	Faculty of Mathematics and Informatics	Institute of Data Science and Digital technologies

Study method	Number of credits	Study method	Number of credits
Lectures	1 (autumn)	Consultations	1
Individual works	4	Seminars	1

Summary

Course objective. The aim is to enable the doctoral student to acquire essential skills in analyzing Big Data and develop practical abilities to apply methods and technologies for Big Data storage and processing, and to critically evaluate modern scalable machine learning paradigms and distributed architectures as well as methods of research.

Introductory part. Definition and evolution of Big Data. The main sources of Big Data. Types of Big Data: structured and unstructured. Big Data storage technologies and techniques. Data Lakehouses, Data Fabric, and Data Mesh architectures. Big Data mining: practical applications.

Big Data analytics:
Knowledge discovery in databases and the data mining process.
Methods of large-scale preprocessing: data profiling, cleaning, feature engineering, and data sampling.
Big Data analysis tasks.
Big Data mining and scalable machine learning:

- Frequent pattern analysis and association rules.
- Forecasting, classification, and anomaly detection at scale; model evaluation under concept drift.
- Streaming data analysis.
- Clustering methods for large-scale data.
- Dimension reduction and representation learning (including embeddings).
- Machine and deep learning for Big Data processing: neural networks (CNNs, RNNs, Transformers), transfer learning, distributed training.
- Large Language Models (LLMs) and Generative AI. Retrieval-Augmented Generation (RAG) and vector databases.
- Text mining and Natural Language Processing (NLP): TF-IDF, transformer-based models, topic modeling, sentiment analysis.
- Machine Learning Model Operationalization management (MLOps) practices for scalable systems.
- Big Data visualization challenges.

Big Data technologies: MapReduce (concepts), Apache Hadoop and Spark, Apache Kafka, Spark MLlib Library for Machine Learning, Big Data analytics using the R package, PySpark and Python.
Cloud computing solutions for Big Data analytics: Microsoft Azure Machine Learning, Amazon Web Services (AWS), Google Cloud (BigQuery, Dataproc), Databricks (Lakehouse, Delta Lake), Containerization and orchestration basics (Docker/Kubernetes) for reproducible analytics.

Practical aspects of Big Data processing. Practical applications in addressing the challenges arising from the business, finance, medicine, cybersecurity, and other domains.

Practical task: to solve a real-world data analysis problem using Big Data and machine learning or deep learning, including data collection, preprocessing, and visualization.
Main literature
Kleppmann, M. (2017). Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. O'Reilly Media, Inc.
Chambers, B., & Zaharia, M. (2018). Spark: The definitive guide: Big data processing made simple. " O'Reilly Media, Inc.".
Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PloS one*, *15*(2), e0228987.
Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Kaluri, R., Rajput, D.S., Srivastava, G. and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, pp.54776-54788.
Bhatia, P. (2019). Data Mining and Data Warehousing Principles and Practical Techniques. Cambridge University Press. https://doi.org/10.1017/9781108635592
Jurafsky, D. and Martin, J.H. 2026. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 6, 2026. https://web.stanford.edu/~jurafsky/slp3.
Pietsch, W. (2021), Big Data, Cambridge University Press. https://doi.org/10.1017/9781108588676
Mining Massive Data Sets, SOE-YCS0007, Stanford School of Engineering https://online.stanford.edu/courses/soe-ycs0007-mining-massive-data-sets
Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. Mining of massive data sets. Cambridge university press, 2020. https://doi.org/10.1017/9781108684163, http://www.mmds.org/
Spark and Python for Big Data with PySpark, Udemy https://www.udemy.com/course/spark-and-python-for-big-data-with-pyspark/

Lecturer(s) (name, surname)	Science degree	Main publications
Viktor Medvedev	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Viktor+Medvedev
Jolita Bernatavičienė	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Jolita+Bernataviči enė
Olga Kurasova	dr.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova