University of Gdańsk

Subject card

Subject name and code	Data bases & big data, PG_00117811						
Field of study	Chemistry						
Date of commencement of studies	October 2024		Academic year of realisation of subject	2024/2025			
Education level	postgraduate studies		Subject group	Optional subject group			
Mode of study	full-time studies		Mode of delivery	at the university			
Year of study	1		Language of instruction	English			
Semester of study	2		ECTS credits	2.0			
Learning profile	academic		Assessment form				
Conducting unit	Pracownia Chemoinformatyki Środowiska -> Katedra Chemii i Radiochemii Środowiska -> Faculty of Chemistry						
Name and surname of lecturer (lecturers)	Subject supervisor		mgr inż. Michał Kałapus				
	Teachers						
Lesson types	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	0.0	0.0	30.0	0.0	0.0	30
	E-learning hours included: 0.0						
Learning activity and number of study hours	Learning activity	Participation in didactic classes included in study plan		Participation in consultation hours	Self-study		SUM
	Number of study hours	30		5.0	15.0		50
Subject objectives	The goal of the course is to introduce students to the basics of Big Data processing using MapReduce and PySpark technologies. Students will learn to design, implement and optimize data processing algorithms in these technologies, as well as use cloud resources (AWS, Google Cloud) to run them.						

Learning outcomes

Course outcome	Subject outcome	Method of verification
[CHEMMU2_K06] Undertakes research tasks consciously and responsibly, understanding the social aspects of the practical application of the acquired knowledge and skills and the responsibility related to it.	The student is able to identify and analyze the social, ethical, and legal aspects associated with the use of large data sets, including chemical data, particularly in the context of privacy, data security, and the potential social impacts of data analysis. The student is able to consider these aspects in the planning and implementation of research projects using computational and information technology methods.	[SK5] implementation of a problem task [SK8] observation of student's independent or team work
[CHEMMU2_U01] Plans and implements chemical experiments of medium complexity.	The student is able to design and conduct analysis of large data sets from experiments, including chemical experiments, with an in-depth level of complexity. The student is able to use computational and computer methods to process, analyze and interpret data (MapReduce and PySpark), taking into account the specifics of the experiment and potential sources of error.	[SU1] oral statement/conversation/ discussion
[CHEMMU2_W02] Has extended and in-depth knowledge in the field of basic chemistry.	The student is able to apply in-depth knowledge of the basic branches of chemistry (organic, inorganic, physical, analytical chemistry) to the analysis of large sets of chemical data. The student is able to identify chemical compounds, predict their properties and reactivity, and evaluate the influence of various factors on the course of chemical reactions.	[SW5] implementation of a problem task
[CHEMMU2_W08] Demonstrates knowledge of theoretical computational and IT methods used to solve problems in chemistry.	The student is able to apply in-depth theoretical knowledge of computational and computer methods (e.g., machine learning, statistics, network analysis) to the analysis of large chemical data sets. The student is able to select appropriate methods for a specific research problem, conduct the analysis, interpret the results and assess their reliability.	[SW1] oral statement/ conversation/discussion [SW5] implementation of a problem task
[CHEMMU2_U09] Has deepened ability to prepare various forms of oral presentations on chemistry in Polish and English.	The student is able to prepare and deliver various forms of oral presentations (paper, poster, seminar, scientific communication) in Polish or English on the analysis of large sets of chemical data, adapting the style and level of detail to the audience. The student is able to use a variety of data visualization tools and techniques to convey information in an attractive and understandable manner.	[SU1] oral statement/conversation/ discussion [SU8] observation of student's independent or team work
[CHEMMU2_U08] Prepares and presents oral presentations in various fields of chemistry in Polish and English, using acquired knowledge and skills as well as basic sources of scientific information.	The student is able to prepare and deliver an oral presentation in Polish or English on the analysis of large data sets, using appropriate data processing tools and techniques (MapReduce and PySpark). The student is able to critically evaluate the results of the analysis, draw conclusions and present them in a way that can be understood by audiences with different levels of expertise.	[SU1] oral statement/conversation/ discussion [SU2] presentation/project/paper/ report [SU5] implementation of a problem task

<table>
<tr><td>Subject contents</td><td>concept of large databases and BigData, basics of big datasets engineering, big data hardware infrastructure (local and cloud), MapReduce
algorithm, introduction to Python programming language, data analysis in Python, machine learning (supervised and unsupervised methods),
introduction to Apache Spark and Hadoop
setting up working environment (Python, Spark) and Big Data datasets, Spark basics, Resilient distributed datasets, RDDs examples and exercises,
introduction to SparkSQL, SQL commands exercises, Spark MLlib (linear regression and decision trees with Spark ML)</td></tr>
<tr><td>Prerequisites and co-requisites</td><td></td></tr>
<tr><td>Assessment methods and criteria</td><td>
<table>
<tr><th>Subject passing criteria</th><th>Passing threshold</th><th>Percentage of the final grade</th></tr>
<tr><td>Observation of the student's behavior during classes and during consultations.</td><td>50.0%</td><td>30.0%</td></tr>
<tr><td>Project</td><td>50.0%</td><td>70.0%</td></tr>
</table>
</td></tr>
<tr><td>Recommended reading</td><td>
<table>
<tr><td>Basic literature</td><td>Apache Spark and PySpark documentation - https://spark.apache.org/docs/
Python documentation - https://docs.python.org/3/
M. Bowles - Machine Learning with Spark and Python®: Essential Techniques for Predictive Analytics</td></tr>
<tr><td>Supplementary literature</td><td>The Google File System
Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google</td></tr>
<tr><td>eResources addresses</td><td>Adresy na platformie eNauczanie:</td></tr>
</table>
</td></tr>
<tr><td>Example issues/ example questions/ tasks being completed</td><td>main concepts of Big Data; understand the structure and properties
of databases; understand the hardware requirements and differences in the
infrastructure for big data; MapReduce algorithm and its mapper and reduction functions; understand basics of the Python programming language (types,
data structures, functions, libraries), know essential methods and libraries used in
data analysis and machine learning in Python; know Apache Spark and Hadoop
engines and its modules</td></tr>
<tr><td>Work placement</td><td>Not applicable</td></tr>
</table>

Document generated electronically. Does not require a seal or signature.