GDAŃSK UNIVERSITY OF TECHNOLOGY

Subject card

Subject name and code	Data engineering tools, PG_00062742		
Field of study	Technologies for Industry 5.0		
Date of commencement of studies	October 2024	Academic year of realisation of subject	2026/2027
Education level	first-cycle studies	Subject group	Obligatory subject group in the field of study Subject group related to scientific research in the field of study
Mode of study	Full-time studies	Mode of delivery	at the university
Year of study	3	Language of instruction	Polish
Semester of study	6	ECTS credits	4.0
Learning profile	general academic profile	Assessment form	assessment
Conducting unit	Division of Theoretical Physics and Quantum Informaton -> Institute of Physics and Applied Computer Science -> Faculty of Applied Physics and Mathematics		
Name and surname of lecturer (lecturers)	Subject supervisor	dr inż. Patryk Jasik	
	Teachers		

Lesson types and methods of instruction

Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
Number of study hours	0.0	0.0	30.0	15.0	0.0	45

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	45	5.0	50.0	100

Subject objectives

Gaining knowledge and skills related to the practical use of data engineering tools.

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K6_U06] performs analysis, exploration and cleaning of data sets, can use statistical models and machine learning models, integrate various analytical, management and data storage tools	The student performs analysis, exploration, and cleaning of data sets using data engineering tools, can utilize statistical models and machine learning models, and can integrate various tools for analytics, data management, and storage.	[SU1] Assessment of task fulfilment
[K6_W06] demonstrates knowledge related to data analysis and engineering, machine learning, knows the principles of integrating data with management systems to analyze complex engineering and technological problems	The student demonstrates knowledge of data engineering tools to analyze complex engineering and technological problems.	[SW3] Assessment of knowledge contained in written work and projects
[K6_K03] effectively, clearly and unambiguously conveys information, describes activities and communicates their results and opinions of a specialist engineer using appropriate communication methods and tools	The student effectively, clearly, and unambiguously conveys information about data engineering tools, describes actions, and communicates their results and the specialist engineer's opinions using appropriate communication methods and tools.	[SK2] Assessment of progress of work [SK5] Assessment of ability to solve problems that arise in practice

Subject contents	Introduction to Data Engineering (2 hours) 1. **Basics of Data Engineering**: Introduction to concepts and technologies. • Definition of data engineering and its role in data processing. • Overview of data engineering process stages: acquisition, cleaning, processing, analysis, and presentation of data. • Discussion of popular tools and technologies used in data engineering. Python and Databases (6 hours) 1. **Python for Data Engineering**: • Basic Python libraries for data engineering. • Practical exercises on data manipulation using selected Python packages. 2. **SQL: Relational Databases**: • SQL review: syntax basics, queries. • Database management: e.g., MariaDB, PostgreSQL. • Practical exercises on creating and managing SQL databases. • Connecting to databases from Python code and executing queries. 3. **NoSQL: Non-relational Databases**: • Introduction to NoSQL: principles, types (document, key-value, column, graph). • Examples of popular NoSQL databases: MongoDB, Cassandra. • Practical exercises on using MongoDB in data engineering. • Connecting to MongoDB from Python code and performing data operations. Apache Spark (6 hours) 1. **Apache Spark: Processing Large Data Sets**: • Introduction to Apache Spark: architecture and components. • Spark SQL and DataFrames: processing data in tabular format. • Spark Streaming: processing streaming data. • Practical exercises on using Spark for data analysis. Docker and Kubernetes (6 hours) 1. **Docker: Application Containerization**: • Introduction to Docker: concepts, architecture. • Creating and managing Docker containers. • Practical exercises on application containerization. 2. **Kubernetes: Container Orchestration**: • Introduction to Kubernetes: architecture, basic concepts. • Creating and managing Kubernetes clusters. • Practical exercises on using Kubernetes for container management. Cloud Solutions (5 hours) 1. **Cloud Data Platforms: Introduction to Cloud Solutions**: • Overview of major cloud providers: AWS, Google Cloud Platform, Microsoft Azure. • Review of cloud services: Amazon S3, Google BigQuery, Azure Data Lake. • Practical exercises on using cloud data platforms. 2. **Infrastructure as a Service (IaaS) and Platform as a Service (PaaS)**: • Introduction to IaaS and PaaS: differences, advantages, and disadvantages. • Creating and managing cloud computing environments. • Practical exercises on deploying and managing applications in the cloud. Apache Airflow (5 hours) 1. **Apache Airflow: Workflow Orchestration**: • Introduction to Apache Airflow: architecture and basic concepts. • Defining DAG (Directed Acyclic Graph): structure and principles of operation. • Schedules and operators: creating and managing tasks. • Practical exercises on configuring and deploying workflows in Airflow. • Monitoring and managing tasks in Apache Airflow. Team Project (max 2 people) (15 hours) • **Defining the Problem and Collecting Data**. • **Designing and Implementing Data Engineering Solutions**. • **Presenting Results and Discussing Challenges and Solutions**.
Prerequisites and co-requisites	Knowledge of Python and SQL.
Assessment methods and criteria	

Subject passing criteria	Passing threshold	Percentage of the final grade
Team Project (max 2 students)	60.0%	100.0%

Recommended reading	Basic literature	Sreeram Nudurupati, "Essential PySpark for Scalable Data Analytics. A beginner's guide to harnessing the power and ease of PySpark 3", Packt PublishingCuantum Technologies LLC, "Data Analysis Foundations with Python. Master Data Analysis with Python: From Basics to Advanced Techniques", Packt PublishingCuantum Technologies LLC, "Python and SQL Bible. From Beginner to World Expert: Unleash the true potential of data analysis and manipulation", Packt Publishing
	Supplementary literature	Pulkit Chadha, Data Engineering with Databricks Cookbook. Build effective data and AI solutions using Apache Spark, Databricks, and Delta Lake, Packt Publishing
	eResources addresses	Adresy na platformie eNauczanie:
Example issues/ example questions/ tasks being completed	Guidelines for Creating the Project Report: 1. Report Title 2. Introduction - Motivation, Goals 3. Data Description - Data structure, variable description, origin 4. Setting Up the Environment for Data Processing 5. Description of Data Preparation Process - Step-by-step 6. Data Analysis and/or Modeling - Assumptions, brief description of methods and chosen analysis and/or modeling methodology 7. Data Presentation Using the Created Application 8. Results, Conclusions, and Discussion The report, along with all code, should be placed in a chosen repository (e.g., GitLab, GitHub).	
Work placement	Not applicable	

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