COURSE DESCRIPTION CARD - SYLLABUS

Course name
Big Data Technologies [S1DSwB1>TBD]

Course

Field of study Data Science in Business	Year/Semester 3/6
Area of study (specialization) –	Profile of study general academic
Level of study first-cycle	Course offered in Polish
Form of study full-time	Requirements compulsory

Number of hours

Lecture 30	Laboratory classes 0	Other 0
Tutorials 30	Projects/seminars 0	

Number of credit points

5,00

Coordinators

dr hab. Grzegorz Pawłowski
grzegorz.pawlowski@put.poznan.pl

Lecturers

Prerequisites

Basic knowledge of programming and database management.

Course objective

Providing basic knowledge and skills in the organization, management and processing of very large, variable and diverse Big Data sets.

Course-related learning outcomes

Knowledge:
Characterizes methods for analyzing and processing large datasets, including exploration techniques, machine learning, and data visualization in the Big Data environment [DSB1_W01].
Describes the architecture of distributed systems and NoSQL databases used in Big Data technology [DSB1_W02].
Explains the concept of data processing in the MapReduce model and the use of Apache Spark for Big Data analysis [DSB1_W03].
Analyzes key aspects of security, privacy management, and legal regulations regarding large-scale data [DSB1_W05].

Skills:
Selects appropriate methods and tools for Big Data analysis, considering their efficiency and scalability [DSB1_U02].
Designs and conducts analytical experiments in a Big Data environment, utilizing distributed computing systems [DSB1_U03].
Creates data analysis and processing pipelines using Spark SQL and PySpark, optimizing the management of large data resources [DSB1_U08].
Applies machine learning techniques to analyze large datasets, implementing predictive and exploratory models [DSB1_U09].
Effectively collaborates in interdisciplinary analytical teams, integrating knowledge from cloud technologies and distributed systems [DSB1_U14].

Social competences:
Takes business initiatives related to the implementation of Big Data technologies, utilizing innovative data analysis methods [DSB1_K04].

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Summative grade for the lecture is based on the percentage result from the test. Questions and tasks checking understanding of the topics. Passing threshold - 50%.
Formative laboratory assessment consists of grades that the student receives for completing individual tasks during classes. The summary grade from the laboratory is given as the average of these grades. The assessment takes into account the correctness and completeness of the results achieved.

Programme content

Lecture: Definition, importance, applications and challenges of Big Data. Introduction to distributed file systems and NoSQL databases. MapReduce programming model, Sparc architecture and components, data retrieval and streaming. Data warehousing and the use of data mining and machine learning in data analysis and visualization. The idea of cloud computing and characteristics of the largest platforms: MS Azure, Amazon Web Service (AWS) and Google Cloud Platform (GCP) and their services. Data management, security and responsible use of data.
Lab: Familiarizing students with the Apache Hadoop platform, in order to operate on distributed file systems. Introduction to batch processing engines, optimization and data decomposition using the example of MapReduce. Relational data processing using Spark SQL. Using PySparc to manage large data
resources.

Course topics

1. Fundamentals of Big Data
2. Big Data resource characteristics
3. Methods for processing large data sets
4. Big Data analysis and visualization
5. Cloud computing platforms and their services
6. Selected Big Data Management Systems
7. Secure management of private data

Teaching methods

Lectures: informative lecture, multimedia presentation, problem-based lecture.
Laboratories: laboratory method, case study method, workshop method.

Bibliography

Basic:
Marz N., Warren J., Big Data. Principles and best practices of scalable realtime data systems, Manning Pubications Co., 2015.
White T., Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, O'Reilly Media, 2015.

Zaharia M., Chambers B., Spark: The Definitive Guide, O'Reilly Media / Helion, 2018.
Chalkiopoulos A., Programming MapReduce with Scalding. A practical guide to designing, testing, and implementing complex MapReduce applications in Scala, Helion, 2014.

Additional:
Garcia-Molina H., Database Systems The Complete Book, Pearson India, 2013.
Ryza S., Lasersson U., Owen S., Wills J., Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly Media, 2015.

Breakdown of average student's workload

	Hours	ECTS
Total workload	125	5,00
Classes requiring direct contact with the teacher	62	2,50
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	63	2,50