COURSE DESCRIPTION CARD - SYLLABUS

Course name
Big Data Processing

Course

Field of study
Computing

Year/Semester
4/7

Area of study (specialization)

Profile of study
general academic

Level of study
First-cycle studies

Course offered in
Polish

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture
30

Laboratory classes
30

Number of credit points

4

Lecturers

Responsible for the course/lecturer:
dr inż. Krzysztof Jankiewicz
email: Krzysztof.Jankiewicz@put.poznan.pl
phone: 61 6652960
faculty: Faculty of Computing and Telecommunications
address: ul. Piotrowo 2, 60-965 Poznań

Responsible for the course/lecturer:

Prerequisites

Knowledge of relational database systems. Knowledge of the SQL language. Basic knowledge of object oriented programming languages such as Java.

Course objective

1. Provide students with basic knowledge in the field of organization, management and Big Data processing.
2. Developing students' ability to solve problems related to the organization, management and processing of Big Data.

Course-related learning outcomes

Knowledge

Has knowledge of significant development directions and the most important achievements made in Big

Data processing. (K1st_W5)
Has systematized and theoretically founded general knowledge in the field of processing large volumes of data as well as detailed knowledge of selected issues related to this area of computer science. (K1st_W4)
She/He knows the basic techniques, methods and tools used in the processing of Big Data, mainly of engineering. (K1st_W7)

Skills

Is able to formulate and solve Big Data processing tasks, use appropriately selected methods, including analytical, simulation or experimental methods. (K1st_U4)
She/He can properly use Big Data processing techniques, applicable at various stages of the implementation of IT projects. (K1st_U2)
She/he Is able to obtain information from various sources, including literature and databases, both in Polish and in English, integrate it properly, interpret and critically evaluate it, draw conclusions, and exhaustively justify her/his opinions. (K1st_U1)
She/He is able - according to the given specification - to design and implement a Big Data processing project, selecting appropriate methods, techniques and programming tools. (K1st_U10)
She/He is able to plan and implement the process of his own permanent learning and knows the possibilities of further education (2nd and 3rd degree studies, courses and lectures available on the Internet). (K1st_U19)
Has the ability to formulate Big Data processing algorithms and implement them using at least one of the popular programming tools. (K1st_U11)

Social competences

She/He understands that knowledge and skills related to Big Data processing become obsolete very quickly (K1st_K1)
Is aware of the importance of knowledge in solving engineering problems in the field of Big Data processing, knows examples and understands the causes of malfunctioning information systems that have led to serious financial and social losses, or to a serious loss of health and even life. (K1st_K2)

Methods for verifying learning outcomes and assessment criteria

Learning outcomespresented above are verified as follows:

The learning outcomes presented above are verified as follows:
Formative assessment:
a) in the field of lectures:
- on the basis of answers to questions about the material discussed in the lectures.
b) in the field of laboratories / exercises:
- on the basis of an assessment of the current progress in the implementation of tasks.
Summative assessment:
a) in the field of lectures, verification of the assumed learning outcomes is carried out by:
- assessment of the knowledge and skills demonstrated in the exam with different characteristics and complexity of problems to be solved (simple tasks related to basic knowledge, more difficult tasks requiring calculations or simulation of algorithms, problem tasks of high complexity); the total number

of questions in the exam is about 10; all questions are scored similarly, a total of 100 points can be obtained; passing the exam is from 50 points; the final grade is a weighted average of the written and laboratory exam.
- examination of the exam results,
b) in the field of laboratories, verification of the assumed learning outcomes is carried out by:
- evaluation of the implementation of tasks related to given laboratory classes; during each laboratory class, the student receives a list of tasks to be performed (consisting of non-marked tasks, scored tasks and homework) for which 30% of points can be obtained, in addition, the student carries out two projects in the middle and at the end of the semester, for which he can receive 30% respectively and 40% points; completion of the laboratory is from 50% of the points obtained during the whole semester; it is possible to get additional points for activity during classes.

Programme content

Wykład:
The lecture program covers the following topics:
- Introduction to Big Data systems, motivations, definitions, problems in the Big Data world, types of tool processing. Big Data systems architectures (Lambda, Kappa). NoSQL database models, BASE, CAP theorem.
- Hadoop platform, distributed file systems on the example of HDFS, task scheduling systems in Big Data systems on the example of YARN, data batch processing engines on the example of MapReduce, MapReduce processing optimization techniques, decomposition of complex problems into MapReduce action sequences, Hadoop Streaming
- Higher level programming tools on the example of Pig and Hive systems, architecture, processing optimization techniques, Pig Latin, Hive SQL. Physical data organization, ORC file format, Bloom filter.
- Introduction to Scala functional programming
- Modern Big Data processing engines on the example of the Spark platform, architecture, techniques of unstructured data processing using RDD, RDD support for key-value pairs, optimization of RDD processing.
- Relational data processing using Spark SQL, DataFrame and Dataset data types, data processing in Spark SQL, processing optimization mechanisms.Laboratoria:
Laboratory classes are conducted in the form of fifteen two-hour exercises, held in the laboratory. Exercises are carried out individually, with the exception of some tasks that can be carried out in teams of two. The laboratory program covers the following topics:
- Familiarization with the environments used in laboratories
- Hadoop - introduction, MapReduce
- HDFS, YARN
- High-level batch processing - Pig
- High-level batch processing - Hive
- Introduction to Scala language
- Spark platform - introduction
- Spark - RDD - basics
- Spark - RDD - key-value

- Spark - RDD - performance
- Spark - DataFrames
- Spark - Datasets

Teaching methods

1. lecture: multimedia presentation illustrated with examples given on the board, discussion and problem analysis.
2. laboratory exercises: problem solving, discussion, team work.

Bibliography

Basic

1. N. Marz, J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Pubications Co., 2015.
2. T. White, Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale, O'Reilly Media; 4th edition (April 14, 2015)
3. Matei Zaharia, Bill Chambers, Spark: The Definitive Guide, O'Reilly Media, 2018
4. M. Odersky, L. Spoon, B. Venners, Programming in Scala, 3rd edition, Artima Inc, 2016.
5. Mining of Massive Datasets, A. Rajaraman, J. D. Ullman, Cambridge University Press, 2012 (http://infolab.stanford.edu/~ullman/mmds.html)
6. Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom, Database Systems: The Complete Book, Pearson; 2nd edition (June 5, 2008)

Additional

1. S. Ryza, U. Lasersson, S. Owen, J. Wills, Spark. Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly Media; 2nd edition (June 12, 2017)
2. C. Horstmann, Scala for the Impatient, Addison-Wesley, 2016.
3. Ch. Lam, Hadoop in Action, Manning Publications Co., 2011.
4. R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, John Wiley & Sons 2002
5. P. Raghavan, H. Schütze, Introduction to Information Retrieval, Ch. D. Manning, Cambridge University Press 2008, (http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html)

Breakdown of average student's workload

	Hours	ECTS
Total workload	101	4
Classes requiring direct contact with the teacher	61	2.4
Participation in laboratory classes / exercises[1]	30	1.2
Completing (as part of own work) tasks from laboratory exercises	5	0.2
Homework: 5 x 1 hour	5	0.2
Participation in consultations related to the implementation of the education process	1	0.0
Preparation for classes with obligatory scored tasks	10	0.4
Participation in lectures	30	1.2
Getting familiar with the indicated literature and teaching materials (10 pages of scientific text = 1 hour), 100 pages	10	0.4
Preparation for the exam	10	0.4

[1] delete or add other activities as appropriate