COURSE DESCRIPTION CARD - SYLLABUS

Course name
Processing of Massive Datasets - BigData [N1Inf1>BIGD]

Course

Field of study
Computing

Year/Semester
4/8

Area of study (specialization)
–

Profile of study
general academic

Level of study
first-cycle

Course offered in
Polish

Form of study
part-time

Requirements
compulsory

Number of hours

Lecture
16

Laboratory classes
16

Other
0

Tutorials
0

Projects/seminars
8

Number of credit points

4,00

Coordinators

dr inż. Krzysztof Jankiewicz
krzysztof.jankiewicz@put.poznan.pl

Lecturers

Prerequisites

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

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 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 knowledge and skills demonstrated in the lecture final exam with various characteristics and complexity of problems to be solved (simple tasks concerning basic knowledge, more difficult tasks requiring calculations or simulation of algorithms, problem tasks of high complexity); the total number of questions is approximately 10; all questions are scored similarly, you can get a total of 100 points; passing the lecture test is from 50 points; the final grade is a weighted average from the written lecture exam and laboratory results.
- discussion of the results of the lecture test,
b) in the field of laboratories, verification of the assumed learning outcomes is carried out by assessing the implementation of tasks related to current laboratory classes; during each laboratory class, the student receives a list of tasks to be performed; it is possible to obtain additional points for activity during classes; passing the laboratory requires obtaining at least 50% of the points collected throughout the semester, the final grade results from the points obtained in the laboratory and the project,
c) in terms of the project, the student carries out two projects in the middle and at the end of the semester; passing the project requires obtaining 50% of the possible points

Programme content

The program includes: an introduction to Big Data systems, details about distributed file systems on the example of HDFS, task scheduling systems on the example of YARN, batch data processing engines on the example of MapReduce, the Hadoop platform, higher-level programming tools on the example of the Hive platform, modern Big Data processing engines on the example of the Spark platform.

Course topics

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 Hive platform, architecture, processing, optimization techniques, 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.

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 - Hive
- Introduction to Scala language
- Spark platform - introduction
- Spark - RDD
- Spark - SQL (DataFrames)
- Spark - Datasets/Pandas API on Spark

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.
3. project: independent work, consultations with the teacher, implementation of Big Data data processing systems

Bibliography

Basic
1. N. Marz, J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Publications 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	100	4,00
Classes requiring direct contact with the teacher	40	2,00
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	60	2,00