COURSE DESCRIPTION CARD - SYLLABUS

Course name

Processing of Massive Datasets

Course

Field of study	Year/semester
Computing	1/1
Area of study (specialization)	Profile of study
Artificial Intelligence	general academic
Level of study	Course offered in
Second-cycle studies	Polish
Form of study	Requirements
full-time	compulsory

Number of hours

Lecture	Laboratory classes	Other (e.g. online)
30	30	
Tutorials	Projects/seminars	

Number of credit points

4

Lecturers

Responsible for the course/lecturer:	Responsible for the course/lecturer:
Anna Kobusińska, PhD, DSc	Krzysztof Jankiewicz, PhD
email: Anna.Kobusinska@cs.put.poznan.pl	email: Krzysztof.Jankiewicz@cs.put.poznan.pl
tel. 61 665-2964	tel: 61 6652960
Institute of Computing Science, Faculty of Computing and Telecommunications	Institute of Computing Science, Faculty of Computing and Telecommunications
Piotrowo 2, 60-965 Poznan	Piotrowo 2, 60-965 Poznan

Prerequisites

Learning objectives of the first cycle studies defined in the resolution of the PUT Academic Senate that are verified in the admission process to the second cycle studies. The learning objectives are available at the website of the faculty www.cat.put.poznan.pl. In particular, students starting the course Processing of Massive Datasets should have basic knowledge of operating systems, distributed processing, computer networks, relational database systems as well as SQL and object-oriented programming languages.

Students should also be capable of continuous learning and knowledge acquisition from selected sources, understand the need to expand their competencies, as well as express the readiness for collaborating as part of a team.

Course objective

The objective for this course is to give the students basic knowledge in the field of processing of massive datasets, in particular the presentation of theoretical and practical aspects of the design, of large scale systems that process such massive datasets, and the challenges related to their development and management. Developing students' skills in solving problems of processing of massive datasets in large-scale distributed environments.

Course-related learning outcomes

Knowledge

1. has advanced detailed knowledge regarding selected IT issues such as architecture and classification of systems that process massive datasets, programming tools used in massive datasets processing environments [K2st_W3]

1. has knowledge about development trends and the most important cutting edge achievements in computer science and other selected and related scientific disciplines in the field of processing of massive datasets [K2st_W4]

2. has advanced and detailed knowledge of the processes occurring in the life cycle of hardware or software information systems [K2st_W5]

3. knows advanced methods, techniques and tools used to solve complex engineering tasks and conduct research in a selected area of computer science in the field of processing of massive datasets [K2st_W6]

Skills

1. is able to obtain information from literature, databases and other sources (both in Polish and English), integrate them, interpret and critically evaluate them, draw conclusions and formulate and fully justify opinions [K2st_U1]

4. is able to plan and carry out experiments, including computer measurements and simulations, interpret the obtained results and draw conclusions and formulate and verify hypotheses related to complex engineering problems and simple research problems related to processing of massive datasets [K2st_U3]

5. can use analytical, simulation and experimental methods to formulate and solve engineering problems and simple research problems related to processing of massive datasets [K2st_U4]

6. is able to assess the suitability and the possibility of using new achievements (methods and tools) and new IT products in the field of processing of massive datasets [K2st_U6]

7. is able - using among others conceptually new methods - to solve complex IT tasks related to processing of massive datasets, including atypical tasks and tasks containing a research component [K2st_U10]

Social competencies

1. understands that in the field of IT the knowledge and skills related to processing of massive datasets quickly become obsolete [K2st_K1]

2. understands the importance of using the latest knowledge in the field of processing of massive datasets in solving research and practical problems [K2st_K2]

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Formative assessment:

a) in relation to lectures - on the basis of answers to questions related to the course material discussed during the lectures.

b) in relation to laboratories - on the basis of an assessment of the current progress in the implementation of tasks.

Summative assessment:

a) in relation to lectures - verification of the assumed learning outcomes is carried out in two tests that consist of questions of varied characteristics and complexity (simple basic knowledge tasks, more difficult tasks requiring calculations, problem tasks of high complexity). Students must obtain at least 50% of the total points available at each test. The final grade is based on the results of both tests.

b) in relation to laboratories - verification of the assumed learning outcomes is carried out by assessing the implementation of tasks related to given laboratory classes; during each laboratory class, students receive a list of tasks to be performed; moreover, students carry out two projects in the middle and at the end of the semester. Students must obtain at least 50% of the possible points in the first and second half of the semester; it is possible to get additional points for activity during laboratory classes; the final grade results from the points collected throughout the semester.

Programme content

Lectures cover the following topics:

1. Presentation of the challenges related to the processing of massive datasets: sources of massive datasets, definitions of massive datasets, various aspects of processing massive datasets

2. Introduction to large-scale distributed systems that process massive datasets: classifications of massive datasets processing systems, Big Data systems architecture (Lambda, Kappa).

3. Introduction to NoSQL databases: classification (key value, column-oriented, document-oriented, column-oriented, graph-oriented models); construction of NoSQL systems (data partitioning, load

balancing, replication, data versioning, membership management, failure handling) based on Google BigTable, Cassandra, Neo4j; CAP, PACELC theorems

4. Resource management in Big Data systems based on Mesos and YARN resource management systems (architecture, resource allocation algorithms)

5. Storage of massive data - distributed file systems based on Google File System (architecture, applied algorithms)

6. Approaches to processing massive graph data – based on Pregel system

7. Concurrent processing of massive data – based on Apache Spark platform (architecture), processing techniques using Resilient Distributed Datasets (RDD)

8. Relational data processing using Spark SQL, DataFrame and Dataset data types, data processing in Spark SQL, processing optimization mechanisms

9. Modern techniques of streaming data processing – based on streaming platforms: Apache Flink, Apache Kafka, Apache Spark Streaming

During laboratories the following topics are covered:

1. Introduction to the environments used during laboratories - installation, configuration, programming interface, data types, basic operations available in a given system.

2. Practical use of systems that process massive datasets:
   - implementation of applications in the Cassandra environment (introduction to Cassandra, CQL, Java Api database)
   - implementation of jobs in the Apache Flink environment
   - implementation of jobs in the Apache Kafka environment
   - implementation of jobs in the Apache Spark environment (introduction to Apache Spark platform, Spark SQL, Spark Structured Streaming)

Teaching methods

1. Lectures: multimedia presentation illustrated with examples given on the blackboard.

2. Laboratory classes: multimedia presentation illustrated with examples given on the blackboard and demonstration, discussion, workshops, practical exercises, team work.

Bibliography

Basic

1. J.Berman, Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information, Morgan-Kaufman, 2013

2. N. Marz, J. Warren, Big Data. Principles and best practices of scalable realtime data systems, Manning Pubications Co., 2015

3. M. Zaharia, B. Chambers, Spark: The Definitive Guide, O'Reilly Media, 2018

4. A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2012 (podręcznik dostępny w wersji elektronicznej: http://infolab.stanford.edu/~ullman/mmds.html)

5. P. Sadalage, M. Flower, NoSQL distilled, Addison-Wesley, 2013

Additional

1. S. Ryza, U. Lasersson, S. Owen, J. Wills, Spark. Zaawansowana analiza danych, Helion, 2015

2. J. S. Damji et al., Learning Spark - Lightning-Fast Data Analytics, O'Relly Media, 2020

3. I. Robinson, J. Webber, E. Eifrem, Graph Databases: New Opportunities for Connected Data, O'Reilly Media, Inc., 2015

4. A. Kobusińska, C. Leung, C.-H. Hsu, S. Raghavendra , V. Chang, Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing, Future Generation Computer Systems, 87, 2018

Breakdown of average student's workload

	Hours	ECTS
Total workload	100	4,0
Classes requiring direct contact with the teacher	60	2,0
Student's own work (literature studies, preparation for laboratory classes/tutorials, preparation for tests/exams, project preparation)[1]	40	2,0

[1] niepotrzebne skreślić lub dopisać inne czynności