COURSE DESCRIPTION CARD - SYLLABUS

Course name
Big data and distributed processing [S1SI1E>BIGD]

Course

Field of study
Artificial Intelligence

Year/Semester
3/6

Area of study (specialization)
–

Profile of study
general academic

Level of study
first-cycle

Course offered in
English

Form of study
full-time

Requirements
elective

Number of hours

Lecture
30

Laboratory classes
30

Other
0

Tutorials
0

Projects/seminars
0

Number of credit points

5,00

Coordinators

dr hab. inż. Anna Kobusińska prof. PP
anna.kobusinska@put.poznan.pl

mgr inż. Adam Godziński
adam.godzinski@put.poznan.pl

Lecturers

Prerequisites

Students starting the course should have basic knowledge of operating systems, 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 big data and distributed processing. In particular the presentation Big Data organization, and theoretical and practical aspects of the design of distributed systems that process such Big Data, as well as the challenges related to their development and management. Developing students" skills to solve problems related to the organization, management and processing of Big Data in distributed environments.

Course-related learning outcomes

Knowledge:
1. Students have a well-grounded knowledge in the domain of distributed systems, distributed processing, and classification and management of Big Data
2. Student know and understand the basic paradigmes, techniques, methods, algorithms, and tools used for solving distributed computing problems and Big Data processing problems, including synchronisation of time in distributed processing; various approaches to data and service replication, as well as the concept of replica consistency; implications of individual node and network communications failures; effects of large scale on the provision of fundamental services and the tradeoffs arising from scale and Big Data; range of distributed algorithms, such as broadcast and consensus; NoSql approaches to data processing and management
3. Students have the 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 Big Data distributed processing

Skills:
1. Students understand that knowledge and skills quickly become outdated in computer science and, in particular, Big Data distributed processing and perceives the need for constant additional training and raising one"s qualifications
2. Students can analyse computational and communication complexity of distributed algorithms
3. Students can use proper methods (analytical, simulation, experimental) for solving of specific distributed computing problems
4. Students can efficiently plan and carry out experiments, including computer measurements and simulations, interpret the obtained results and draw conclusions based on the experimental outcomes in the context of distributed processing and Big Data processing and management problems
5. Students can design and implement a distributed algorithm, choosing proper language for the task and using proper techniques, methods and tools

Social competences:
1. Students understand that in the field of IT the knowledge and skills related to processing of massive datasets quickly become obsolete
2. Students understand the importance of using the latest knowledge in the field of distributed and Big Data processing in solving research and practical problems

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

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) Lectures: verification of the assumed learning outcomes is carried out during the exam which has the form of multi-choice test and tasks of varied characteristics and complexity (simple basic knowledge tasks, more difficult tasks requiring calculations, problem tasks of high complexity) concerning the subjects presented during all lectures. Each task is evaluated individually, being allocated a certain number of points. The points are summed up and a standard scale is used to derive the final marks: <50% - 2.0, [50%, 60%) - 3.0, [60%, 70%) - 3.5, [70%, 80%) - 4.0, [80%, 90%) - 4.5, and [90%,100%] - 5.0.
b) Laboratory classes: 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 three projects. Students must obtain at least 50% of the possible points from the projects. 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

The course covers the following topics: introduction to distributed systems; avantages and challenges of distributed systems; time, in distributed systems; challenges related to the processing of Big Data;

introduction to NoSQL databases; CAP and PACELC theorems; replication problem; quorum algorithms; processing of massive data based on Apache Spark platform and MapReduce

Course topics

The following topics are discussed during the course:
- Introduction to distributed systems; avantages and challenges of distributed systems; unbounded delay and partial failure; network protocols; transparency; client-server systems; remote procedure call (RPC).
- System models and faults. Synchronous, partially synchronous, and asynchronous network models; crash-stop, crash-recovery, and Byzantine faults; failures, faults, and fault tolerance; two generals problem.
- Time, clocks, and ordering of events. Physical clocks; leap seconds; UTC; clock synchronisation and drift; Network Time Protocol (NTP). Causality; happens-before relation. Logical time; Lamport clocks; vector clocks.
- Broadcast (FIFO, causal, total order); gossip protocols.
- Challenges related to the processing of Big Data: sources of Big Data, definitions and characteristics of Big Data, various aspects of processing Big Data Classifications of Big Data distributed processing systems,
- Big Data systems architectures(Lambda, Kappa).
- 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, Dynamo, Cassandra;
- CAP and PACELC theorems
- Replication. Quorums; idempotence; replica consistency;state machine replication; leader-based replication; inearizability; eventual consistency; consensus and total order broadcast. FLP result; leader election; the Paxos and Raft consensus algorithms.
- Concurrent processing of massive data based on Apache Spark platform (architecture), processing techniques using Resilient Distributed Datasets (RDD);relational data processing using Spark SQL, DataFrame and Dataset data types, data processing in Spark SQL, processing optimization mechanisms

Teaching methods

1. Lectures: multimedia presentation, illustrated with examples given on the blackboard.
2. Laboratory classes: a multimedia presentation illustrated with examples given on the blackboard and project.

Bibliography

Basic
1. Modern Operating Systems (free PDF available online) by Andrew S Tanenbaum, Herbert Bos
2. Kleppmann, M. (2017). Designing data-intensive applications. O'Reilly.
3. Tanenbaum, A.S. and van Steen, M. (2017). Distributed systems, 3rd edition. available online.
4. Cachin, C., Guerraoui, R. and Rodrigues, L. (2011) Introduction to Reliable and Secure Distributed Programming. Springer (2nd edition).
5. NoSQL distilled, P. Sadalage, M. Flower, Addison-Wesley, 2013
6. M. Zaharia, B. Chambers, Spark: The Definitive Guide, O"Reilly Media, 2018
Additional
1. Spark in Action, Bonaći M., Zečević P., Manning, 2015
2. A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2012 (online: http://infolab.stanford.edu/~ullman/mmds.html)
3. J. S. Damji et al., Learning Spark - Lightning-Fast Data Analytics, O"Relly Media, 2020
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	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