COURSE DESCRIPTION CARD - SYLLABUS

Course name
Big Data and Cloud Computing [S2Inf1-TPD>BIGD]

Course

Field of study
Computing

Year/Semester
1/1

Area of study (specialization)
Data Processing Technologies

Profile of study
general academic

Level of study
second-cycle

Course offered in
Polish

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture
30

Laboratory classes
0

Other
0

Tutorials
0

Projects/seminars
0

Number of credit points

2,00

Coordinators

dr inż. Tomasz Koszlajda
tomasz.koszlajda@put.poznan.pl

Lecturers

Prerequisites

The learning outcomes from the 1st cycle studies defined in the Resolution of the Senate of PUT, especially the effects K_W1-2, K_W4, K_W6-15, verified in the recruitment process for the 2nd cycle studies - these effects are presented on the faculty"s website. The learning outcomes of first cycle studies defined in the Resolution of the Senate of PUT, in particular effects K_U1-2, K_U4, K_U7-8, K_U14-20, K_U22-23, K_U26, verified in the process of recruitment for second cycle studies - these effects are presented on the faculty"s website. The learning outcomes from the 1st cycle studies defined in the Resolution of the Senate of PUT, especially the effects K_K1-9, verified in the process of recruitment for the 2nd cycle studies - these effects are presented on the faculty"s website. Moreover, in the scope of social competences, a student must present such attitudes as honesty, responsibility, perseverance, cognitive curiosity, creativity, personal culture, respect for other people.

Course objective

1.To provide students with basic knowledge of the new fields of application of database systems and new models of database systems, in terms of cloud computing and in particular the processing of huge data sets - Big Data. 2.To develop in students the ability to solve problems of analysis, design and implementation of applications of new generations of databases.

Course-related learning outcomes

Knowledge:
has advanced and in-depth knowledge of broadly understood information systems, theoretical foundations of their construction, and methods, tools and programming environments used in their implementation (k2st_w1)
has knowledge concerning problems of efficiency, fault tolerance and coherence of processing in distributed, replicated and parallel computing platforms, based on theoretical foundations: queue theory, coherence models of replicated data processing (k2st_w2)
has advanced detailed knowledge of selected topics in computer science (k2st_w3)
has knowledge of development trends and most significant new achievements of computer science and other, selected, related disciplines (k2st_w4)
has knowledge of advanced methods, techniques and tools applied in solving complex engineering tasks and carrying out research works in the selected area of computer science (k2st_w6)

Skills:
is able to acquire information from literature, databases and other sources (in native language and english), in the scope of alternative solutions to problems presented in classes; (k2st_u1)
is able to use analytical, simulation and experimental methods to formulate and solve engineering tasks and simple research problems; (k2st_u4)
is able to integrate knowledge from different areas of computer science, e.g. database systems or operating systems (k2st_u5)
is able to assess the usefulness and possibility of using new developments and new it products, e.g. in selecting an appropriate nosql class system; (k2st_u6)
is able to critically analyze existing technical solutions and propose their improvements, e.g. in load balancing; (k2st_u8)
is able to assess the suitability of methods and tools for solving an engineering task, involving, for example, appropriate solutions for big data analysis; (k2st_u9)
is able to solve complex it tasks, e.g. requiring multiple data analysis interactions; (k2st_u10)

Social competences:
understands that in computer science, knowledge and skills become obsolete very quickly (k2st_k1)
understands the importance of using the latest knowledge of computer science in solving research and practical problems (k2st_k2)

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:
- participation in lectures, activity during lectures: looking for answers to questions posed by the lecturer, being critical of lecturers" translations, being interested in extending lectures, finding errors in lecture materials.
Summative assessment:
- evaluation of the knowledge and skills demonstrated on a problem-based written exam (the student may use a limited set of teaching materials); a score of at least 50% is required to obtain a grade of 3.0. The final grade also takes into account the evaluation of activity during lectures.

Programme content

The lecture program covers the following topics:
1. distributed database technology: fragmentation, partitioning and data sharding.
2. Cloud performance - load balancing in cloud computing; Managing concurrent execution of large computing tasks.
3. Performance correctness of databases with data replication.
4 Parallel databases. Architectures of parallel databases.
5. BigData technology. Map-Reduce processing model and architecture. In-memory database technology.
6. New generation of NoSQL class databases.

Course topics

The lecture program covers the following topics:
1. Rationale for database cloud technology. Data processing service - DaaS. Big Data processing. Distributed database technology: fragmentation, partitioning and sharding of data, basics of data fragmentation - Consistent Hashing.
2. Cloud performance - load balancing in cloud computing; basic concepts of queue theory, Kendall"s notation; Little"s law; Kingman"s formula; cloud load balancing protocols; job scheduling protocols. Impact of variability in job size and job submission frequency on the quality of load balancing and job scheduling; G/G/N queueing systems. Management of virtual machines - Distributed Resource Scheduler algorithm. Management of concurrent execution of large computing tasks. Fair resource allocation algorithms: Max-min fairness and Dominant Resource Fairness.
3 Correctness of databases with data replication. Consistency of replicated databases: Brewer"s theorem, iPACeLC classification; consistency models for replicated databases; methods for maintaining Primary Copy, MultiMaster Copies and Quorum replicas. Algorithms for maintaining replicas; logical clocks, version vectors, Paxos protocol and RAFT algorithm.
4 Parallel databases. Architectures of parallel databases. Data partitioning methods. Algorithms for parallel database processing.
5 Big Data technology. Map-Reduce processing model and architecture: HDFS , YARN and ZooKeeper. Spark platform: data structures and functionality. In-memory database technology; algorithms and data structures: red-black tree, AVL-tree, T-tree, linear hashing.
6. New generation of NoSQL class databases. New logical models: key-value, column families, document and graph data model. CRUD processing paradigm. Performance of NoSQL family database systems. Sharding and replication in NoSQL systems.
7. NewSQL combination of relational data model with sharding and data replication technologies.

Teaching methods

Lecture: multimedia presentation, illustrated with examples given on the blackboard.

Bibliography

Basic
1. Big data: efektywna analiza danych, Mayer-Schonberger, MT Biznes 2017
2. Big data: najlepsze praktyki budowy skalowalnych systemów obsługi danych w czasie rzeczywistym, N. Marz, J. Warren, Helion 2016
3. Cloud Computing: Theory and Practice, D. Marinescu, Morgan Kaufmann 2013
4. Principles of Distributed Database Systems, M. Özsu, P. Valduriez, Springer 2011
5. Spark. Zaawansowana analiza danych, S.Ryza, U.Laserson, S.Owen, J.Wills, Helion 2016
6. Hadoop. Kompletny przewodnik. Analiza i przechowywanie danych, T. White, Hekion 2016
7. Performance Modeling and Design of Computer Systems, M. Harchol-Balter, Cambridge University 2013
Additional

Breakdown of average student's workload

	Hours	ECTS
Total workload	50	2,00
Classes requiring direct contact with the teacher	30	1,50
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	20	0,50