COURSE DESCRIPTION CARD - SYLLABUS

Course name
Processing of Data Streams in Big Data Systems

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 studies

Course offered in
Polish

Form of study
full-time

Requirements
elective

Number of hours

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

Number of credit points

5

Lecturers

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

Responsible for the course/lecturer:

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 Data Streams in Big Data Systems should have basic knowledge of operating systems, distributed processing, computer networks, relational database systems as well as SQL and object-oriented programming languages, as well as massive data processing systems (Big Data) using batch data processing.

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 stream data processing and processing of massive data streams, in particular the presentation of theoretical and practical aspects of the design, of large scale systems that process such massive data streams, and the challenges related to their development and management. Developing students' skills in solving problems of processing of massive data streams 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 data streams, programming tools used in massive data streams 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 data streams [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 data streams [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 data streams [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 data streams [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 data streams [K2st_U6]

7. is able - using among others conceptually new methods - to solve complex IT tasks related to processing of massive data streams, including a typical 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 data streams quickly become obsolete [K2st_K1]

2. understands the importance of using the latest knowledge in the field of processing of massive data streams 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 one test 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 the test. The final grade is based on the results of the test and the grade of laboratories.

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; 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 data streams and massive data streams: data stream sources, data stream definitions, data stream processing aspects.
2. Introduction to data stream processing systems, basic concepts, system architecture, levels of programming interfaces on the example of centralized platforms: Esper and Oracle.

3. Introduction to massive data stream processing systems, data stream acquisition, message queuing systems, publisher/subcriber solutions on the example of Apache Kafka platform.
4. The first generation of massive data stream processing systems, concept, concepts, limitations, on the example of the Spark Streaming library.
5. Next generations of massive data stream processing systems, support for event timestamps, unordered data, delayed data, state processing, triggers, on the example of Kafka Streams and Spark Structured Streaming libraries.
6. Processing of massive data streams using high-level programming interfaces, Table API, Complex Event Processing on the example of the Apache Flink platform.
7. Platforms for building declarative data flow systems with support for massive data streams, architecture, concepts on the example of Apache NiFi

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 data streams and massive data streams:
   - implementation of applications in the Esper and Oracle platforms
   - implementation of jobs in the Apache Kafka environment
   - implementation of jobs in the Apache Spark environment
   - implementation of jobs in the Apache Flink environment
   - implementation of dataflows in the Apache NiFi platform

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

2. Tyler Akidau, Slava Chernyak, Reuven Lax, Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing, O'Reilly, 2018

3. Fabian Hueske, Vasiliki Kalavri, Stream Processing with Apache Flink. Fundamentals, Implementation, and Operation of Streaming Applications, O'Reilly Media, 2019

4. Gwen Shapira, Todd Palino, Rajini Sivaram, Krit Petty, Kafka - The Definitive Guide: Real-time data and stream processing at scale, O'Reilly Media; Wydanie II, 2022

5. 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)

6. 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. 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

4. Documentation of systems / platforms used during the course available on-line

Breakdown of average student's workload

	Hours	ECTS
Total workload	125	5,0
Classes requiring direct contact with the teacher	60	2,5
Student's own work (literature studies, preparation for laboratory classes/tutorials, preparation for tests/exams, project preparation)[1]	65	2,5

[1] delete as appropriate or add other activities