

Big Data Infrastructure

Course Workload		Assessment form (examination/ graded test/ ungraded test)
ECTS	Hours	
6	216	Exam

Students will study the principles of construction and the basics of organizing the development of modern software solutions for processing big data, typical errors that occur when working with big data, and the signs of their manifestation, methods for eliminating typical errors that arise when working with big data. an integration solution, the main mechanisms and algorithms for analyzing big data and extracting knowledge from them, principles and technologies for the functioning of the chosen integration platform, the capabilities of modern and promising means of integrating systems and other applications and services, principles for processing, storing and protecting data.

Course structure:

1. Evolution of big data processing systems

1.1. The main stages of the development of big data processing systems, the main types of systems and their purpose

2. Distributed File System HDFS

2.1. The purpose of the HDFS distributed file system, the basic principles of the HDFS device, the procedure for data replication and ensuring fault tolerance

3. MapReduce technology

3.1. Apache Zookeeper architecture and principles, consensus algorithms, PAXOS algorithm

4. Providing fault tolerance using Apache ZooKeeper

4.1. An introduction to self-organizing cards; methods for constructing self-organizing maps; evaluating and configuring self-organizing maps

5. YARN and Mesos Resource Managers

5.1. The purpose and tasks of the infrastructure manager, YARN architecture and principles, Mesos architecture and principles, centralized and two-tier planning approaches

6. Batch processing big data

6.1. Principles of organizing data batch processing, architecture and principles of the Apache Spark device, data processing using Spark

7. Streaming Big Data

7.1. Principles of organizing streaming data processing, architecture and principles of the Apache Kafka and Apache Flink device, processing streaming data using Spark Streaming and Apache Flink

8. Interactive Big Data Processing

8.1. Principles of organizing interactive data processing, Lambda and Kappa architectures, interactive data processing using Spark SQL

9. Graph data processing

9.1. Representation of graph data for batch processing, processing graph data with Spark GraphX