ODESSA I.I. MECHNYKOV NATIONAL UNIVERSITY
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATION TECHNOLOGIES
DEPARTMENT OF MATHEMATICAL SUPPORT OF COMPUTER SYSTEMS

Syllabus of the course "Analysis and visualization of huge data sets (Big Data)"

Amount	the total number of: credits – 4; hours – 120; content modules - 3
Semester	spring
Days, Time, Place	according to the class schedule
Teacher(s)	Tatiana Petrushyna, Ph.D. (physics and mathematics), Associate Professor of the Department of Mathematical Support of Computer Systems
Contact phone number	(0674860042)
E-mail	Tatyana.Petrushina@gmail.com
Workplace	department of mathematical support of computer systems
Consultations	face-to-face consultations: Monday from 13.00-14.00 online consultations: ZOOM (link is generated at the beginning of classes)

COMMUNICATION

Communication with students will be carried out by e-mail, in the classroom or via ZOOM.

COURSE ABSTRACT

Subject of the study of the course is the technical and methodological means of organization, storage and analysis of large data (Big Data).

Course Prerequisites

The course material is based on the previously acquired knowledge, practical skills and skills of the students on topics and areas related to algorithms, data structures, relational databases, the SQL language, discrete mathematics and probability theory. The corresponding courses are taught within the educational program of the first (bachelor) level of higher education in specialty 126 "Information systems and technologies".

Course Post-requisites

This course complements the discipline "Intelligent data analysis and machine learning methods" in the field of data analysis and processing and is the basis for mastering the following disciplines of the educational and professional master's training program in the specialty 126 "Information systems and technologies": "Professional research practice", " Completion of master's qualification work".

Purpose of the course is the formation of system knowledge regarding the organization of computational processes of extraction and interpretation of hidden knowledge using various types of application program packages (tensor toolbox, Hadoop, Map reduce, etc.); the ability to develop interface software for combining the source of big data (Big Data) with the appropriate software environment; the ability to implement a computing process based on cloud services and technologies, parallel and distributed computing; acquisition of knowledge, practical skills and abilities to use modern methods, algorithms for processing highly dimensional, large-volume, rapidly changing multi-format data to extract hidden knowledge needed to support decision-making.

Course content

Considered:

- *Introduction to Big Data analytics. Basic concepts and definitions. Sources of Big Data. Big Data formation and processing technologies. Application of Big Data: in economy, business, health care, industry, examples of use in scientific fields. Concepts of Data Mining, Datafication.*
- *Data management, Big Data life cycle. Phases of the data life cycle, conditionality of the cycle, principles of formation. Data management, 7 key considerations in creating a hybrid integration strategy.*

- *Problems of representation and modeling of knowledge and their connection with the problem of Big Data. Computer knowledge - database + rule base. Conceptual and empirical models of knowledge. Data Mining - research and discovery of hidden knowledge by "machine" in raw data.*
- *"Artificial intelligence" as a synthesis of "data mining", "knowledge discovery" and "hidden knowledge". Methods of extracting new knowledge from fact bases. Matrix decompositions and their role in data analysis.*
- *Multidimensional arrays, tensor models. Traditional DataScience - Big Data - common features and differences. The modern trend is from 2D matrices to 3D tensors. The expediency and necessity of machine learning when processing the results of Big Data processing.*
- *Base of the main conceptual apparatus for Big Data analysis: methods of numerical analysis based on tensor representations of data; the latest methods and models of big data processing based on tensor networks and tensor decompositions. Matrices, matrix expansions.*
- *Tools and software for working with Big Data. Basic principles of working with data; tools; examples of solving practical problems.*
- *Machine learning. Principles of working with Big Data, the MapReduce paradigm, Hadoop and corporate systems, Hadoop and DBMS.*

EXPECTED RESULTS

As a result of studying the course, the student must

know: basic concepts and terminology of the organization, storage and analysis of big data, technologies and areas of application of Big Data tools and data mining methods (Data Mining), understand its role and place in decision-making automation systems, basic methods, models, software and technical tools, which allow processing of many dimensional, fast-changing and high-volume arrays of data;

be able: to apply the acquired knowledge in professional activities during the development, adjustment and operation of information systems using artificial intelligence and under non-standard conditions of conducting research.

Competencies that the student receives as a result of studying the course:

- *the ability to develop mathematical, informational, and computer models of objects and processes related to informatization.*
- *the ability to utilize modern data analysis technologies for optimizing processes in information systems.*
- *the ability to conduct information analysis and create multi-dimensional models of subject areas.*

Learning outcomes: upon completion of the course, the student will have skills

- *develop and use data repositories, and perform data analysis to support decision-making.*
- *solve digital transformation tasks in new or unfamiliar environments based on specialized conceptual knowledge, including modern scientific achievements in the field of information technology, research, and knowledge integration from various fields.*
- *develop and support autonomous distributed intelligent systems for automated information search and analysis.*
- *create optimized pipelines for data preparation for subsequent storage and processing.*

FORMS AND METHODS OF TEACHING

The course will be taught in the form of lectures (18 hours) and laboratory classes (18 hours), organization of students' independent work (84 hours).

The basic training of students is carried out in lectures and laboratory classes.

During the teaching of the course, the following teaching methods are used: verbal (lecture, explanation); face-to-face (Power Point presentation); practical (laboratory works); work with literary sources (independent work of students).