MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
ODESSA I.I. Mechnikov NATIONAL UNIVERSITY
Department of mathematical support of computer systems

"APPROVED"
Vice-rector for scientific and pedagogical work
_____
_____ 20___

WORKING PROGRAM OF EDUCATIONAL COURSE

OK10 "Analysis and visualization of huge data sets (Big Data)"
(course name)

Level of higher education Second (master's)

Field of knowledge 12 – Information technologies

Specialty 126 – Information systems and technologies
(code and name of specialty)

Educational and professional program Information systems and technologies
(EPP/ESP name)

ONU - 2022

The working program of the educational course "Analysis and visualization of huge data sets (Big Data)". – Odesa: ONU, 2022. – 9 p.

Developer:
Petrushyna T.I., Ph.D. (Ph.-M.), Associate Professor of the Department of MSCS

The working program was approved at the meeting of the Department of Mathematical Support of Computer Systems

Protocol No. 1 from "25" 08 2022 year

Head of the department _____ (Eugene MALAKHOV)
(signature) (First Name Surname)

Agreed with the guarantor of the EPP "Information systems and technologies"

_____ (Eugene MALAKHOV)
(signature) (First Name Surname)

Approved by the educational and methodological commission (EMC) for IT specialties of the FMPhIT

Protocol No. 1 from "31" 08 2022 year

Head of EMC _____ (Alla RACHYNSKA)
(signature) (First Name Surname)

Reviewed and approved at the meeting of the department _____

Protocol No. 1 from "29" 08 2022 year

Head of Department _____ (_____)
(signature) (First Name Surname)

Reviewed and approved at the meeting of the department _____

Protocol No. ___ from "____" _____ 20__ year

Head of Department _____ (_____)
(signature) (First Name Surname)

1. Course description

<table>
<tr><td rowspan="2">Name of indicators</td><td rowspan="2">Field of knowledge, direction of training, educational and qualification level</td><td colspan="2">Course characteristics</td></tr>
<tr><td>full-time education</td><td>external form of education</td></tr>
<tr><td rowspan="15">The total number of:

credits - 4

hours - 120

content modules - 3</td><td rowspan="15">Branch of knowledge
12 - Information technologies

Specialty
126 – Information systems and technologies

Level of higher education:
Second (master's)</td><td colspan="2">Mandatory</td></tr>
<tr><td colspan="2">Year of training:</td></tr>
<tr><td>1st</td><td></td></tr>
<tr><td colspan="2">Semester</td></tr>
<tr><td>2nd</td><td></td></tr>
<tr><td colspan="2">Lectures</td></tr>
<tr><td>18 hours</td><td>8 hours</td></tr>
<tr><td colspan="2">Practical, seminar</td></tr>
<tr><td>hours</td><td>hours</td></tr>
<tr><td colspan="2">Laboratory</td></tr>
<tr><td>18 hours</td><td>6 hours</td></tr>
<tr><td colspan="2">Independent work</td></tr>
<tr><td>84 hours</td><td>106 hours</td></tr>
<tr><td colspan="2">Individual tasks: hours</td></tr>
<tr><td colspan="2">Form of final control: exam</td></tr>
</table>

* - in the presence

2. The purpose and tasks of the educational course

The purpose teaching of the course is the formation of students' ideas about technical and methodological means of organization, storage and analysis of Big Data.

Task: acquisition of competences based on the assimilation of the main theoretical provisions regarding the existing technical and methodological means of analyzing super-large data, which ensure the storage and management of the amount of data in hundreds of terabytes or petabytes, which conventional relational databases do not allow to use effectively; mastering effective methods of organizing unstructured information (text, images, video, etc.), means and methods of working with it, generation of analytical reports, implementation of predictive models.

The process of studying the course is aimed at forming elements of the following **competencies**:

1) general:

—

2) professional:

SC04. The ability to develop mathematical, informational, and computer models of objects and processes related to informatization.

SC05. The ability to utilize modern data analysis technologies for optimizing processes in information systems.

SCM07. The ability to conduct information analysis and create multi-dimensional models of subject areas.

Program learning outcomes:

LO09. Develop and use data repositories, and perform data analysis to support decision-making.

LO11. 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.

LOM07. Develop and support autonomous distributed intelligent systems for automated information search and analysis.

LOM08. Create optimized pipelines for data preparation for subsequent storage and processing.

Expected learning outcomes. As a result of studying the academic course, the student should

***know**: basic concepts, methods and tools in the field of Big Data; basics of machine learning, visualization and storage of big data; the basics of working with data warehouses and NoSQL DBMS*

***be able**: apply approaches of analytical processing of big data when solving problems related to management in complex technical systems; master the methods of using software tools that support Big Data technologies to solve practical problems in the subject area; translate subject area problems to big data processing technology.*

3. Course content

Content module 1. General concepts and definitions of Big Data.

Topic 1. Modern approaches to the processing and storage of extremely large data.

Topic 2. Data processing, visualization, primary statistical analysis.

Topic 3. Forms of data presentation, types and types of data.

Content module 2. Software in the field of big data analysis.

Topic 1. Analytical platforms: classification and application features.

Topic 2. Visual modeling languages. The R language

Topic 3. NoSQL DBMS. Hadoop.

Topic 4. MapReduce. Full text search. ETL process for processing reports.

Content module 3.KDD and Data Mining technologies.

Topic 1. Data preparation for analysis. Methods of extracting knowledge.

Topic 2. Finding associative rules. Application of classification and regression. Statistical methods.

4. Course structure

Names of content modules and topics	Number of hours									
	Full-time					Correspondence form				
	That's all	including				That's all	including			
		l	p	lab	Wed		l	p	lab	Wed
1	2	3	4	5	6	7	8	9	10	11
Content module 1. General concepts and definitions of Big Data.										
Topic 1.	5	1			4	6				6
Topic 2.	5	1			4	10	1		1	8
Topic 3.	5	1			4	10				8
Content module 2. Software in the field of big data analysis										
Topic 1.	10	2			8	13.5	1		0.5	12
Topic 2.	14	2		2	10	13			1	12
Topic 3.	15	3		2	10	15.5	1		0.5	14
Topic 4.	19	3		4	12	18	1		1	16
Content module 3. KDD and Data Mining technologies										
Topic 1.	23	3		4	16	16	1		1	14
Topic 2.	24	2		6	16	18	1		1	16
Total hours	**120**	**18**		**18**	**84**	**120**	**8**		**6**	**106**

5. Topics of seminar classes

Seminar classes are not provided

6. Topics of practical classes

Practical classes are not provided

7. Topics of laboratory classes

No s/p	Topic name	Number hours	
		Full-time	Correspondence form
1	Preparation, processing, analysis and visualization of results using MS SQL Analytical Services.	4	2
2	Cluster analysis in the R environment	4	2
3	Regression analysis in the R environment	6	2
4	Bayesian classification in the R environment	4	
	Total hours	**18**	**6**

8. Independent work

No s/p	Title of the topic / types of tasks	Number hours	
		Full-time	Correspondence form
1	Analysis of existing Big Data software tools.[1]	10	12
2	The specifics of applying the KDD technology to big data. [1]	24	30
3	The application of neural networks to the analysis of big data.[1]	24	30
4	Organization of data processing in HDFS.[1]	26	34
	Total hours	**84**	**106**

Independent work includes:
[1] – preparation for lectures and laboratory classes;

8.1. Individual educational and research task (course project or calculation and graphic work) is not provided

9. Teaching methods

Lectures using multimedia presentation material.

10. Control methods

Current control methods: assessment of laboratory work performance.

Final control: Exam. During the final control, the student must answer the theoretical questions from the list given in point 11.

10.1. Evaluation criteria for the final modular control:

1. The answer should be complete and short. It should not contain material that does not relate to the essence of the question.

2. Clearly formulate statements, skillfully apply the necessary formulas and knowledge of the main issues of the program.
3. Answers with false statements are evaluated based on the closeness of the answer to the correct one.
4. Omissions in the justification of statements are taken into account and this leads to a decrease in the number of points.
5. Small flaws, inaccuracies in the presentation of the material, reduce the number of points.
6. Ignorance and misunderstanding of the main idea of a theoretical question or problem leads to the withdrawal of up to 90% of points.
7. If there is no answer to the question, zero points are assigned.

11. Questions for the final control

1. Define the essence of the concept of "big data".
2. Describe the techniques of big data analysis.
3. Describe the process of big data analysis.
4. What are the features of big data storage.
5. Give the characteristics of Big Data in the world market.
6. Define the concept of Data Mining.
7. Define the concept of KDD
8. Identify the differences between parametric, nonparametric, and nominal methods.
9. What are the challenges of maintaining big data security?
10. What is cognitive data analysis.
11. What data models do you know?
12. Basic descriptive statistics.
13. Peculiarities of applying correlation-regression analysis of big data.
14. The essence of cluster analysis in application to big data.
15. Finding associative rules in big data.
16. Classification of data using a neural network.
17. Classification using decision trees.
18. Software tools for big data analysis and their shortcomings.
19. The main capabilities of storing big data in the R programming language.
20. Basic features of the R language for big data analysis.

12. Distribution of points received by students

Current testing and independent work									Exam	Sum
Content module 1			Content module 2				Content module 3			
T1	T2	T3	T1	T2	T3	T4	T1	T2		
8	8	10	8	8	10	10	8	10	20	100

T1, T2 ... - topics of content modules

Evaluation scale: national and ECTS

Total points	ECTS assessment	National scale	
90 — 100	A - "excellent"	5 "excellent"	"test"
85 - 89	B - "very good"	4 "good"	
75 - 84	C - "good"		
70 - 74	D - "satisfactory"	3 "satisfactory"	
60 - 69	E - "permissible"		
35 — 59	F - "unsatisfactory with the possibility of reassembly"	2 "unsatisfactory"	"uncounta
0 — 34	FX – "unsatisfactory with mandatory repeat course"		

13. Methodical support

Synopsis of lectures in electronic format; a complex of educational and methodological support of the course; regulations; presentation materials.

14. Recommended Books

14.1. Basic literature

1. Weigend A. Big Data. All technology in one book. - Internet publishing house "Exmo", 2018. - 384 p.
2. Devy S. Fundamentals of Data Science and Big Data. Python and data science. // S. Devy, M. Arno, A. Mohamed — St. Petersburg: Peter, 2017. — 336 e.: ill.
3. Richart V., Coelho P.L. Construction of machine learning systems in Python. - M.: DMK Press, 2016, - 302 p.: ill.
4. S. Khaikin. Neural networks: full course, 2nd edition. : Trans. with English-M. : Publishing House "Williams", 2006.-1104p.
5. Intellectual data analysis: Textbook / O.I. Chernyak, P.V. Zakharchenko/ K.: Znannia, 2014. - 599 p.
6. Akimenko V.V., Zagorodniy Yu.V. Designing of the CSPR based on fuzzy logic. Educational and methodological manual. - K.: KNU Publishing House, 2007. - 94c.
7. White T. Hadoop: A detailed guide. — St. Petersburg: Peter, 2013. — 672 p.: ill.

14.2. Auxiliary literature

1. Mayer-Schönberger, V. Big data. A revolution that will change the way we live, work and think / Viktor Mayer-Schönberger, Kenneth Cukier; trans. with English Inny Haydyuk. — M.: Mann, Ivanov and Ferber, 2014. — 240 p.
2. Chubukova I. A. Data Mining: textbook. — M.: Internet University of Information Technologies: BYNOM: Laboratory of Knowledge, 2006. — 382 p.
3. Chuck L. Hadoop in action. - M.: DMK Press, 2012. - 424 p.: ill.

4. Shipunov A.B., Baldin E.M., Volkova P.A., Korobeynikov A.I., Nazarova S.A., Petrov S.V., Sufiyanov V.G. Visual statistics. Let's use R! - M.: DMK Press, 2012. - 298p.

15. Electronic information resources

1. Big data. [Electronic resource]. - Access mode:https://uk.wikipedia.org/wiki/Big_data
2. Big Data for Development: From Information- to Knowledge Societies [Electronic resource]. – Access mode: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145
3. Big Data from A to Z. [Electronic resource]. – Access mode: https://habrahabr.ru/company/dca/blog/267361/
4. Big Data and blockchain. [Electronic resource]. - Access mode:
5. https://forklog.com/big-data-i-blokchejn-proryv-v-oblasti-analiza-dannyh/
6. Machine learning online coursehttps://www.coursera.org/course/ml
7. Big Data Overview online coursehttps://education.emc.com/academicalliance/elearning/Big_Data_Overview/index.htm
8. Online course R programminghttps://www.coursera.org/course/rprog
9. Introduction to Data Science online coursehttps://www.coursera.org/course/datasci
10. Online course "Introduction to Big Data Analytics" http://bit.ly/IntuitBDA.
11. Textbook on statistical traininghttp://statweb.stanford.edu/~tibs/ElemStatLearn/