МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

ЗАТВЕРДЖЕНО
на засіданні кафедри
інформаційних систем
Протокол № 1 від 27.08.2024 р.

ПОГОДЖЕНО
Проректор з навчально-методичної
роботи

___________________ Каріна НЕМАШКАЛО

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ *
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ СЕМЕНА КУЗНЕЦЯ *
№02071211

АНАЛІТИКА ВЕЛИКИХ ДАНИХ

робоча програма навчальної дисципліни (РПНД)

Галузь знань	**12 "Інформаційні технології"**
Спеціальність	**126 "Інформаційні системи та технології"**
Освітній рівень	**другий (магістерський)**
Освітня програма	**"Інформаційні системи та технології"**

Статус дисципліни	**вибіркова**
Мова викладання, навчання та оцінювання	**англійська**

Розробник:
к.е.н., доцент ___________________ Сергій ЗНАХУР

Завідувач кафедри
інформаційних систем ___________________ Дмитро БОНДАРЕНКО

Гарант програми підписано КЕП Олександр КОЛГАТІН

Харків
2024

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF ECONOMICS

APPROVED

at the meeting of the informational systems department
Protocol № 1 of 27.08.2024

AGREED

Vice-rector for educational and methodological work

________ Karina NEMASHKALO

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ ІМЕНІ СЕМЕНА КУЗНЕЦЯ
№02071211

BIG DATA

Program of the course

Field of knowledge **12 "Information technologies"**
Specialty **126 "Information systems and technologies"**
Study cycle **second (master's)**
Study programme **"Information Systems and Technologies**

Course status **elective**
Language **english**

Developers:
PhD in Economics, Associate Professor ________________ Serhii ZNAKHUR

Head of the Department information systems ________________ Dmytro BONDARENKO

Head of Study Programme digital signature Oleksandr KOLGATIN

Kharkiv
2024

INTRODUCTION

The course "Big Data" is included in the group of educational and professional courses of master's training in specialty 126 "Information systems and technologies". The educational course "Management IS and data storage" is a basic educational course and is studied in accordance with the curriculum for training specialists of the second (master's) level of specialty 126 "Information systems and technologies".

The purpose of studying the course "Big Data" is to provide in-depth knowledge and practical skills in working with big data, building and using distributed systems for building pipelines for processing big data, forming a system of theoretical knowledge and acquiring practical skills and skills in applying big data and distributed technologies bases

The task of the educational course "Big Data" is to form a system of theoretical knowledge and practical skills necessary for working with big data, as well as to master modern technologies and methods for their processing. The main tasks of the course are:

1. Providing knowledge about the construction and use of distributed systems for processing large volumes of data.
2. Formation of skills to create and configure data processing pipelines using Pandas, Dask, Spark technologies.
3. Providing practical skills in the implementation of parallel and distributed computing processes for data analysis.
4. Familiarization with modern Data Mining methods and algorithms for revealing hidden patterns and knowledge in big data.

The training is aimed at training specialists who can effectively work with big data.

The result of the educational course is mastering the following competencies: the ability to ensure the processes of obtaining and interpreting hidden knowledge using Pandas, Dask, Spark; the ability to implement computing processes based on the use of parallel and distributed computing, the ability to develop data analysis pipelines; acquisition of knowledge, practical skills and the ability to use modern methods and Data Mining algorithms.

The object of the educational course is the process of data analysis.

The subject of the educational course is the main approaches to data processing and methods of developing and implementing pipelines for data analysis.

In the process of training, students acquire the necessary knowledge during lectures and performing laboratory work. Independent work of students is also of great importance in the process of studying and consolidating knowledge. All types of classes are developed in accordance with the transfer system of the organization of the educational process. The material taught in this course is used by master's students to write a thesis and study subjects in the master's degree.

The learning outcomes and competences that the course forms are defined in table 1.

Table 1

Learning outcomes and competencies formed by the course

Learning Outcomes	**Competencies**
LO07	GK01,GK04, SK01,SK02,SK03
LO09	SK04, SK05

LO07. Making a grounded choice of project solutions and design a service-oriented information architecture of the enterprise (institution, organisation, etc.).

LO09. Developing and use data warehouses, to perform data analysis for supporting decision-making.

GC01. Ability to abstract thinking, analysis and synthesis.

GC04. Ability to develop and manage projects.

SC01. Ability to develop and apply IST necessary for solving strategic and current tasks.

SC02. Ability to formulate requirements for life cycle stages of service-oriented information systems.

SC03. Ability to design information systems taking into account the specifics of their purpose, incomplete/insufficient information and conflicting requirements.

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

SC05. Ability to use modern data analysis technologies to optimize processes in information systems.

COURSE CONTENT

Content module 1. Basics of big data analysis

Topic 1. Introduction to the course

Introduction to the course. The purpose and tasks of the course, its place in the educational process. The structure of the course, recommendations for its study. Basic concepts of big data analysis (Big Data). Professional requirements. Data Science specializations. Stages and typical steps of data analysis. Data analysis protocol ("Cross-Industry Standard Protocol for Data Mining" CRISP-DM). Pandas, Dask, Spark tools and technologies. Data processing based on Pandas. Features of data processing in DASK. Features of data processing in Spark.

Topic 2. Data analysis based on Dask

Defining Big Data. 3V: Volume, Velocity, Variety. Variety of data types and data sources. Structured and unstructured data. General description of the Dask library. Advantages and features of Dask. Parallelization by the Dask method. An example of statistical analysis of a large data set on Dask.

Topic 3 . Apache Spark Basics

The development history of Spark. Big Data framework. Advantages and features of Apache Spark. MapReduce and Spark. Development of MapReduce-Tez. Basics of DAG. Pros and cons of Lazy Evaluation. Apache Spark modules. Spark processes. Launch Spark. An example of statistical analysis of a large data set on Apache Spark.

Content module 2. Construction of solutions in Apache Spark

Topic 4. Implementation of SQL queries in Apache Spark

Apache Spark Platform. RDD (Resilient Distributed Datasets). Spark SQL module. Integrating SQL with Spark. DataFrame API. Creating DataFrames or Tables. Syntax of basic SQL queries. Executing SQL queries in Apache Spark. Examples of SQL queries in Apache Spark.

Topic 5. Machine learning with Apache Spark

General description of ML methods. Classification task. Classification metrics. Logistic regression. Trees and ensemble of trees. Boosting. Stacking. The structure of Apache Spark. MLlib machine learning library. Functionality of MLlib. Principles of using MLlib. An example of solving a classification problem based on ML Apache Spark.

Topic 6. Using Spark ML and NLP algorithms to build a data analysis pipeline

Basics of NLP (Natural language processing). NLP tasks. The structure of the NLP pipeline. Description of each stage of the pipeline. Overview of frameworks for NLP tasks on a cluster. Implementing NLP in Spark. Basics of Deep Learning on the example of the BERT architecture. Implementing Deep Learning in Spark NLP.

The list of laboratory studies in the course is given in table 2.

Table 2

The list of laboratory studies

Name of the topic and/or task	Content
Topic 1. Lab 1	Basics of data analysis using the DASK library
Topic 2. Lab 2	Using basic PySpark functions
Topic 3. Lab 3	Analytical processing based on PySpark SQL&RDD
Topic 4. Laboratory work 4	Using ML methods to solve data analysis problems in Spark ML (regression)
Topic 5. Lab 5	Using ML methods to solve data analysis problems in Spark ML (classification)
Topic 6. Laboratory work 6	Solving NLP problems

The list of self-studies in the course is given in table 3.

Table 3

List of self-studies

Name of the topic and/or task	**Content**
Topic 1: Task 1.	Research on Big Data trends
Topic 2. Task 2.	Analysis of modern Big Data tools
Topic 3. Task 3.	A comparison of Spark and MapReduce
Topic 4. Task 4.	Spark Streaming Research

Topic 5. Task 5.	Study of MLlib methods and algorithms
Topic 6. Task 6.	Exploring transformers and LLM in Spark NLP

The number of hours of lectures, practical (seminar) studies and hours of self-study is given in the technological card of the course.

TEACHING METHODS

In the process of teaching the course, in order to acquire certain learning outcomes, to activate the educational process, it is envisaged to use such teaching methods as:
Problem lecture (Topics 1, 2, 3, 5, 6), mini-lecture and discussion (Topics 4).
Visual (demonstration (Topics 1-6)).
Individual laboratory work (Topics 1 - 6)

FORMS AND METHODS OF ASSESSMENT

The University uses a 100-point cumulative system for assessing the learning outcomes of students.
Current control is carried out during lectures, practical, laboratory and seminar classes and is aimed at checking the level of readiness of the student to perform a specific job and is evaluated by the amount of points scored:
– for courses with a form of semester control as an exam: maximum amount is 60 points; minimum amount required is 35 points.
The final control includes current control and assessment of the student .
Semester control is carried out in the form of a semester exam or grading.
The final grade in the course is determined:
– for courses with a form of exam, the final grade is the amount of all points received during the current control and the exam grade.
During the teaching of the course, the following control measures are used:
Current control: defense of laboratory work (50 points), written tests (10 points).
Semester control: Grading including Exam (40 points).
More detailed information on the assessment system is provided in technological card of the course.

An example of an exam card and assessment criteria.

Semyon Kuznets Kharkiv National University of Economics
The second (master's) level of higher education

Specialty "126 "Information systems and technologies"
Educational and professional program "Information systems and technologies".
Semester II
Educational course "Big Data"

EXAMINATION TICKET No. 1

Task 1 (diagnostic, 10 points) .
1)Give the difference between Pandas based ETL and DASK based ETL.
2) Provide ETL examples.

Task 2 (heuristic, 30 points).

1. Generate 40 text files of 10000000 characters.
2. Analyze the time of calculating the number of characters in all files using Dask.
3. Analyze the time of calculating the number of characters in all files using Pandas.
4. Compare and analyze the results of the two approaches.

Protocol No. 1 of August 27, 2024 was approved at the meeting of the Department of Information Systems.

Examiner, Doctor of Economics, Assoc. Serhii ZNAKHUR

Chief Department of Ph.D., Assoc. Dmytro BONDARENKO

Evaluation criteria

The examination ticket includes two heuristic tasks. Google Cloud and PowerBI cloud services are used in the process of performing exam tasks . The maximum number is 40 points; the minimum that is counted is 25 points. At the same time, for completely correctly completed tasks, the student receives:
Task 1 – 10 points;
Task 2 – 30 points .
The final scores for the exam consist of the sum of points for the completion of all tasks, rounded to a whole number according to the rules of mathematics.
Task 1 (diagnostic) is estimated at ***10 points*** as follows:
5 points – implementation of the first task;

5 points – implementation of the second task;

Task 2 (heuristic) is valued at ***30 points*** as follows:
8 points – implementation of the first task;
8 points – implementation of the second task;
8 points – implementation of the third task;
6 points – implementation of the fourth task;

RECOMMENDED LITERATURE

Main

1. Foreman, John W. Data Smart: Using Data Science to Transform Information into Insight / John W. Foreman; DB books24x7. - John Wiley & Sons, 2014. - 432 pages. - ISBN 978-1-118-03496-5.: http://common.books24x7.com/toc.aspx?bookid=58144.
2. Devi S. Fundamentals of Data Science and BigData. Python and data science.//S. Devy, M. Arno, A. Mohamed — St. Petersburg: Peter, 2017. — 336 e.: ill.
3. Intellectual data analysis: Textbook / O. I. Chernyak, P. V. Zakharchenko / K.: Znannia, 2014. – 599 p.
4. Information systems and technologies: monograph / by general ed. V. S. Ponomarenko. - Kh.: FOP Brovin O.V., 2019. - 212 p. http://repository.hneu.edu.ua/handle/123456789/21743

Additional

1. Marchenko O.O., Rossada T.V. Current problems of Data Mining: a study guide for students of the Faculty of Computer Science and Cybernetics. - Kyiv. - 2017. - 150
2. Zavadsky I.O. Basics of databases: [Study. manual] / I.O. Zavadsky -K.: Publisher I.O. Zavadskyi, 2011. – 192 p.

Information resources

1.IoT Fundamentals: Big Data & Analytics // Electronic resource. Access mode: https://www.netacad.com/courses/iot/big-data-analytics
2. Apache Spark // Electronic resource. Access mode: https://spark.apache.org/
3. Local Spark cluster // Electronic resource. Access mode: https://r-analytics.blogspot.com/2020/02/spark-r-connect.html
4. Azzalini A., Bruno S. Data Analysis and Data Mining. An Introduction. New York: Oxford University Press, 2012. 289 p. Access mode: http://ebooks.znu.edu.ua/files/Bibliobooks/Kudin/0036206.pdf .

5. Gisele LP, Alex AF Automating the Design of Data Mining Algorithms: an Evolutionary Computation Approach. Heidelberg: Springer-Verlag Berlin Heidelberg, 2010. 197 p. Access mode: http://ebooks.znu.edu.ua/files/Bibliobooks/Kudin/0036216.pdf .
6. Databases and information systems. Training course - Access mode: http://simulation.kiev.ua/dbis/lection06.html
7. The site of personal educational systems of S. Kuznets State University. course "Big Data Analytics". – Access mode: https://pns.hneu.edu.ua/course/view.php?id=9759