National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"

Computer Systems
Software Department

BIG DATA TECHNOLOGIES

Syllabus

Requisites of the Course

Cycle of Higher Education	***First cycle of higher education (Bachelor's degree)***
Field of Study	*12 Information Technologies*
Speciality	*121 Software engineering*
Education Program	*Software Engineering of Multimedia and Information Retrieval Systems*
Type of Course	*Elective*
Mode of Studies	*full-time*
Year of studies, semester	*3 year (5 semester)*
ECTS workload	*4 credits (ECTS). Time allotment - 120 hours, including 54 hours of classroom work, and 66 hours of self-study.*
Testing and assessment	*Credit test*
Course Schedule	*Classes by the timetable http://rozklad.kpi.ua/*
Language of Instruction	*English*
Course Instructors	*Lecturer: PhD, Associate Professor, Liubov Oleshchenko, oleshchenkoliubov@gmail.com* *Teacher of computer workshop: PhD, Associate Professor, Liubov Oleshchenko, oleshchenkoliubov@gmail.com*
Access to the course	*Google classroom: Access is given to registered students.*

Outline of the Course

1. Course description, goals, objectives, and learning outcomes

The study of the discipline "Big Data Technologies" allows students to form the competencies necessary to solve practical problems of professional and scientific activities related to the preparation and analysis of big data.

*The **purpose** of studying the discipline "Big Data Technologies" is to form students' ability to use software methods and data processing tools for big data analysis and management decisions in various fields of science and business.*

*The **subject** of the discipline "Big Data Technologies" is the general principles and approaches to big data processing, capabilities and technologies of Python and R programming languages for processing, analysis and visualization of big data.*

*The study of the discipline "Fundamentals of computer systems and networks" contributes to the formation of the following **professional competence (PC)** in students according to the educational program:*

***PC8** Ability to apply fundamental and interdisciplinary knowledge to successfully solve software engineering problems.*

*The study of the discipline "Fundamentals of computer systems and networks" contributes to the formation of the following **program learning outcomes (PLO)** for students according to the educational program:*

PLO01 *To analyze, purposefully search and select the necessary information and reference resources and knowledge to solve professional problems, taking into account modern advances in science and technology.*

PLO07 *To know and to apply in practice the fundamental concepts, paradigms and basic principles of the functioning of language, instrumental and computational tools of software engineering.*

PLO18 *To know and be able to apply information technology of processing, storage and transmission of data.*

2. Prerequisites and post-requisites of the course (the place of the course in the scheme of studies in accordance with curriculum)

Successful study of the discipline "Big Data Technologies" is preceded by the study of disciplines "Programming" and "Databases" of the curriculum for bachelors in the specialty 121 Software Engineering.

The theoretical knowledge and practical skills obtained during the mastering of the discipline "Big Data Technologies" are necessary for studying the discipline "Network Software Design and Development" of the curriculum for master's degree in specialty 121 Software Engineering.

To successfully master the discipline requires a basic level of English not less than A2.

3. Content of the course

Discipline "Big Data Technologies" involves the study of the following topics:

Topic 1. Sources and types of big data. Preparation and data analysis in Python.

Test 1 (Topic 1.)

Computer Workshops 1-3.

Topic 2. Programming language R capabilities. Architectural models of Big Data.

Test 2 (Topic 2.)

Credit test.

4. Coursebooks and teaching resources

Basic references:

1. Big Data and Big Data Analytics: Concepts, Types and Technologies / https://www.researchgate.net/publication/328783489_Big_Data_and_Big_Data_Analytics_Concepts_Types_and_Technologies/related

2. Understanding Big Data. Analytics for Enterprise Class. Hadoop and Streaming Data / https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/I111025E.pdf?__cf_chl_jschl_tk__=pmd_224e99ec9955bf0d1a8e4ab864b26016568b9271-1628851175-0-gqNtZGzNAfijcnBszQdi

3. Programming with Databases – Python / https://swcarpentry.github.io/sql-novice-survey/10-prog/index.html

4. Pandas / https://www.w3schools.com/python/pandas/pandas_intro.asp

5. Plotly Python Open Source Graphing Library / https://plot.ly/python/

6. Getting Started with Plotly in Python / https://plot.ly/python/getting-started/

7. Microsoft R Application Network / https://mran.microsoft.com/documents/what-is-rDecision Tree / https://www.geeksforgeeks.org/decision-tree/

8. Apache Hadoop / http://hadoop.apache.org/

9. HDFS / https://www.ibm.com/analytics/hadoop/hdfs

10. About the Cassandra File System (CFS) – deprecated / https://docs.datastax.com/en/dse/5.1/dse-dev/datastax_enterprise/ analytics/ cfsAbout.html

Additional references:

1. Extract, transform, and load (ETL) / https://docs.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl

Educational content

5. Methodology

№	Type of training session	Description of the lesson
Topic 1. Sources and types of big data. Preparation and data analysis in Python.		
1	Lecture 1. Sources of big data. Internet of Things. Definition of Big Data.	History of machine learning. History of machine learning development. Predictive analysis and tasks of machine learning. Stages of scientific research. Errors in predictive analysis. Evaluation of the results of machine learning models. Types of machine learning.
2	Lecture 2. Development of software for the analysis of websites that provide open data using Python Pandas. Open data, their formats and processing means.	Possibilities of data analysis tools. The role of Python in data analysis. Traditional big data analytics and next generation analytics. Data analysis life cycle. Open data, their formats and processing means. Web scraping. Extract, convert and download data.
3	Computer Workshop 1. Analysis and visualization of data in Python.	Objective: to demonstrate knowledge of the data analysis lifecycle using a given data set and specified tools, import Python packages needed for data set analysis, use Python and Jupyter tools to prepare this data for analysis, analyze it, build graphs.
4	Lecture 3. Formatting time and date data, reading and writing files in Python.	Format time and date data in Python. Read and write files in Python. Interaction with external applications.
5	Lecture 4. Python and SQLite programming. Purpose of the csvsql utility.	Basic SQL operations. Python work with SQLite. Purpose of the csvsql utility. The execute () method.
6	Lecture 5. Procedure for importing data from files into Pandas. Import data from the Internet using Pandas. Correlation analysis tools in Pandas.	Statistical approaches to big data analytics. Using Pandas. Import data from files. Import data from the Internet. Descriptive statistics in Pandas. Correlation analysis tools in Pandas.
7	Computer Workshop 2. Correlation Analysis in Python.	Objective: Demonstrate skills in performing correlation data analysis using a given data set and specified Python tools to calculate correlation. To configure the dataset, determine if the variables in that dataset are correlated, use Python to calculate the correlation between the two sets of variables, and visualize the results of the research.
8	Lecture 6. Processing of missing data. Convert data types and manipulate date frames.	Processing of missing data. Data type conversion. Manipulating date frames in Python.
9	Lecture 7. Regression analysis of data in Python.	Regression analysis. Types of regression analysis. Application of regression analysis for data analysis.

10	Computer Workshop 3. Building a linear regression in Python.	Objective: to get acquainted with the concepts of linear regression and work with data for forecasting in Python, analyze the proposed sales data and build a linear regression to forecast annual net sales based on the number of stores in the area.
11	Lecture 8. Errors in data analysis and forecasting analytics. Estimation of regression errors by means of Python.	Errors in data analysis and forecasting. Estimation of regression errors by means of Python. Purpose of the scikit-learn library.
12	Lecture 9. Data classification algorithms. Application and problems of classifications.	Classification problems. Classification algorithms. Visualization of classifications. Application and validation of classifications. Decision tree classifier model.
13	Lecture 10. Pyplot module. Plotly tool. Types of data visualization. Visualization of anomalies. Using the Folium and Leaflet.js libraries to build maps.	Pyplot module. Plotly tool. Types of data visualization. Visualization of anomalies. Using the Folium and Leaflet.js libraries to build maps.
Topic 2. Programming language R capabilities. Architectural models of Big Data.		
14	Lecture 11. Data analysis in R. Factors, lists, frames and actions on them.	History of language development R. Possibilities of language R. Objects, packages, functions. Vectors, matrices and operations on them in R. Factors, lists, frames.
15	Lecture 12. Export, import and data processing in R.	Export and import data into R. Use R to analyze time series. Data processing in R.
16	Lecture 13. Basic tools for data analysis and visualization in R.	The plot () function and its parameters. Management of general parameters - arguments of graphic functions. Types of graphs in R.
17	Lecture 14. Big Data architectural models. Virtualization technologies. Hypervisors. Container technology of program code execution on the server. SaaS, PaaS and IaaS.	Architectural models of Big Data engineering. Virtualization technologies. Layers of abstraction. Hypervisors. Container technology of program code execution on the server. Data engineering.
18	Lecture 15. Hadoop Big Data technologies. Distributed MapReduce processing. HDFS.	Scalability with big data. Data storage and processing in distributed file systems. Distributed databases. Distributed Hadoop file system (HDFS).
19	Lecture 16. Distributed Kafka streaming platform. Advantages of Cassandra.	Data reception problem. Kafka streaming platform is distributed. Advantages of Cassandra.
20	Lecture 17. Apache Spark platform. Lambda and Kappa big data processing architectures. Test 2 (Topic 2)	The problem of computational function. Spark technology. Comparison of Spark and MapReduce. Spark and sparklyr for working with big data in R. Lambda - architecture. Advantages and disadvantages of Lambda -architecture. Kappa - architecture. Advantages and disadvantages of Kappa-architecture. Final Test.
21	Lecture 18. Final lecture.	Credit Test.

6. Self-study

The discipline "Big Data Technologies" is based on independent preparation for classroom classes on theoretical and practical topics.

№	*The name of the topic that is submitted for independent study*	*Hours of study*	*References*
1	*Preparing for Computer Workshops 1-3.*	*22*	*[4-6]*
2	*Preparing for Test 1 (Topic 1).*	*12*	*[1-6]*
3	*Preparing for Test 2 (Topic 2).*	*12*	*[7-10]*
4	*Preparing for Credit Test.*	*20*	*[1-10]*

Policy and Assessment

7. Course policy

- *Attendance at lectures is mandatory.*
- *Attendance at computer workshops can be sporadic and if you need to defend a computer workshop.*
- *Rules of conduct in the classroom: activity, respect for those present, turning off the phones.*
- *Adherence to the policy of academic integrity.*
- *Rules for protecting the work of the computer workshop: the work should be done according to the option of the student, which is determined by his number in the group list.*
- *The rules for assigning penalty points are as follows.*

8. Monitoring and grading policy

During the semester, students complete 3 computer workshops. Maximum number of points for each computer workshop: 20 points.

Points are awarded for:

- quality of computer workshop: 0-8 points;

- answer during the defense of the computer workshop: 0-8 points;

- timely submission of work to the defense: 0-4 points.

Maximum number of points for performing and defending computer workshops:

20 points × 3 comp. work. = 60 points.

The task for the final test consists of 8 questions - 5 theoretical and 3 practical. The answer to each theoretical question is evaluated by 5 points, the answer to the practical question is evaluated by 5 points.

Criteria for evaluating each theoretical question of the final test:

5 points - the answer is correct, complete, well-argued;

3-4 points - there are minor errors in the answer;

1-2 points - there are significant errors in the answer;

0 points - no answer or the answer is incorrect.

Criteria for evaluating the practical question of the final test:

5 points - the answer is correct, complete, well-argued;

3-4 points - there are minor errors in the answer;

1-2 points - there are significant errors in the answer;

0 points - no answer or the answer is incorrect.

Maximum number of points for the final test:

5 points × 5 theoretical questions + 5 points × 3 practical questions = 40 points.

The rating scale for the discipline is equal to:

R = Rs = 60 points + 40 points = 1000 points.

Calendar control: conducted twice a semester as a monitoring of the current state of compliance with the requirements of the syllabus.

At the first attestation (8th week) the student receives "credited" if his current rating is not less than 50% of the maximum number of points that the student can receive before the first attestation.

At the second attestation (14th week) the student receives "credited" if his current rating is not less than 50% of the maximum number of points that the student can receive before the second attestation.

Semester control: Credit Test.

Conditions of admission to semester control:

With a semester rating (Rs) of at least 60 points and enrollment in all computer workshops, the student receives the final test"automatically" according to the table (Table of correspondence of rating points to grades on the university scale).

Prerequisite for admission to the final test is the implementation and defense of a computer workshop.

The final performance score or the results of Final test the Fail/ Pass are adopted by university grading system as follows:

Score	*Grade*
100-95	*Excellent*
94-85	*Very good*
84-75	*Good*
74-65	*Satisfactory*
64-60	*Sufficient*
Below 60	*Fail*
Course requirements are not met	*Not Graded*

9. Additional information about the course

Enrollment in Certificates for Distance or Online Courses: Certificates for Online Courses "Programming Essentials in Python" and "IoT Fundamentals: Big Data & Analytics" are credited as modular tests 1 and 2, respectively, received in the Network Academy. The presence of certificates for similar courses in big data processing technologies, writing articles or participation in conferences / projects on relevant topics is also assessed as an additional 5 points.

Syllabus of the course

Is designed by teacher PhD, Associate Professor, Liubov Oleshchenko

Adopted by Computer Systems Software Department (protocol № 8, 25.01.23)

Approved by the Faculty Board of Methodology (protocol № 6, 27.01.2023)