Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Computer Systems Software
Department

BIG DATA AND ANALYTICS

Syllabus

Details of the academic discipline

Level of higher education	***Second (master's)***
Branch of knowledge	*12 Information technologies*
Specialty	*121 Software engineering*
Educational program	*Software engineering of multimedia and information-search systems*
Discipline status	*Selective*
Form of education	*full-time*
Year of training, semester	*2nd year of training, 3rd semester*
Scope of the discipline	*Lectures: 32 hours, laboratory work: 14 hours, independent work: 74 hours.*
Semester control/ control measures	*Final test, modular test , calendar control*
Lessons schedule	*According to the schedule of the current academic year (http://roz.kpi.ua/)*
Language of teaching	*English*
Information about the head of the course / teachers	Lecturer: *Ph.D., associate professor, Liubov Oleshchenko, oleshchenkoliubov@gmail.com* Laboratory work: *Ph.D., associate professor, Liubov Oleshchenko,* oleshchenkoliubov@gmail.com
Placement of the course	Google classroom: Access is given to registered students.

Program of educational discipline

1. Description of the educational discipline, its purpose, subject of study and learning outcomes

The study of the discipline "Big data and analytics" allows students to develop the competencies necessary for solving practical problems of professional and scientific activities related to the analysis of big data.

***The purpose** of studying the discipline "Big data and analytics" is to form students' abilities to apply software methods of big data analytics to solve professional and scientific problems.*

***The subject of** the discipline "Big data and analytics" is software methods and technologies of big data analytics. After mastering the discipline "Big data and analytics", **the learning outcomes** are:*

knowledge:

- *big data storage and support technologies;*
- *the Hadoop stack for big data and the Hadoop Distributed File System (HDFS);*
- *components of the Hadoop stack, including the YARN resource and task management system, HDFS, and the MapReduce programming model;*
- *methods of parallel programming using Spark, built-in Spark libraries;*
- *data placement strategies, using Zookeeper capabilities;*
- *Spark Streaming and Sliding Window Analytics, Spark GraphX & Graph Analytics;*
- *machine learning algorithms using Map Reduce for Big Data Analytics.*

skills:

- *use big data analytics technologies using the capabilities of the Python programming languages.*

experience:

- *apply acquired knowledge for further scientific and professional activities.*

The study of the discipline "Big data and analytics" contributes to the formation of higher education students who study under the educational program "Software Engineering of Multimedia and Information Retrieval Systems" competencies necessary for solving practical problems of professional activity related to big data analytics:

- *Ability to abstract thinking, analysis and synthesis.*
- *Ability to conduct research at an appropriate level.*
- *Ability to develop and implement scientific and/or applied projects in the field of software engineering.*

The study of the discipline "Big data and analytics" contributes to the formation of higher education students studying under the educational program "Software Engineering of Multimedia and Information Retrieval Systems" competencies necessary for solving practical problems of professional activities related to big data analytics:

GC01 *Ability to abstract thinking, analysis and synthesis.*

GC03 *Ability to conduct research at the appropriate level.*

PC02 *Ability to develop and implement scientific and / or applied projects in the field of software engineering.*

2. Pre-requisites and post-requisites of the discipline

(the place of the course in the scheme of studies in accordance with curriculum)

The successful study of the discipline "Big data and analytics" is preceded by the study of the regulatory disciplines "Programming and "Databases" of the curriculum for bachelors in the specialty 121"Software engineering".

The discipline "Big data and analytics" ensures the implementation of course projects and master's theses in the specialty 121 "Software engineering".

3. Content of the academic discipline

The discipline "Big data and analytics" involves the study of the following topics:
Topic 1. Architectural models of big data.
Topic 2. Software methods of big data analytics.
Modular test.
Final test.

4. Educational materials and resources

Basic literature:

1. Oleshchenko, L. M. Big data and analytics [Electronic resource] : textbook for higher education students studying in the specialty F2 «Software Engineering» / L. M. Oleshchenko ; Igor Sikorsky Kyiv Polytechnic Institute. – Electronic text data (1 file: 9,65 MB). – Kyiv: Prosvita Ltd, 2026. – 317 p. ISBN 978-617-7010-42-4. https://ela.kpi.ua/handle/123456789/81869

2. Big Data & Analytics. https://www.tutorialspoint.com/big_data_tutorials.htm

Additional literature:

1. *MapReduce. Tutorial. Electronic resource. Access mode: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html*
2. *Python programming language. Electronic resource. Access mode: https://www.python.org/*
3. *Running ZooKeeper in Production. Electronic resource. Access mode:*

https://docs.confluent.io/platform/current/zookeeper/deployment.html

4. *Running ZooKeeper, A Distributed System Coordinator. Electronic resource. Access mode: https://kubernetes.io/docs/tutorials/stateful-application/zookeeper/*
5. *Hadoop YARN Architecture. Electronic resource. Access mode: https://www.geeksforgeeks.org/hadoop-yarn-architecture/*
6. *Understanding YARN architecture and features. Electronic resource. Access mode: https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.0.0/data-operating-system/content/apache_yarn.html*
7. *Machine Learning Algorithm K-means using Map Reduce for Big Data Analytics. Electronic resource. Access mode: http://www.nitttrc.edu.in/nptel/courses/video/106104189/lec25.pdf*
8. *Big Data Predictive Analytics. Electronic resource. Access mode: https://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics/*
9. *Page Rank Algorithm and Implementation. Electronic resource. Access mode: https://www.geeksforgeeks.org/page-rank-algorithm-implementation/*
10. *PageRank Algorithm in Big Data. Electronic resource. Access mode: http://www.nitttrc.edu.in/nptel/courses/video/106104189/lec31.pdf*
11. *GraphX Programming Guide. Electronic resource. Access mode: https://spark.apache.org/docs/latest/graphx-programming-guide.html*

The materials are freely available on the Internet.

Educational content

5. Methodology

№	Type of training session	Description of the training session	Hours
		Topic 1. Architectural models of big data.	
1	*Lecture 1. Big data and its role in the modern world. Problems of storage and analytics of big data.*	*Growth of data in today's world. Growth of Internet of Things devices. Definition of big data and its classification. Structured and unstructured data. Cloud computing. Fog computing.*	2
2	*Lecture 2. Big Data Infrastructure. Distributed Data and Analytics. Quantum Computing and Its Applications.*	*Big data infrastructure. Distributed data and its analytics. Quantum computing and its applications for intelligent processing of EEG data using hybrid quantum-classical machine learning models.*	2
3	*Lecture 3. The life cycle of big data analytics.*	*Big data analytics life cycle. Software methods for extracting, transforming, and loading big data.*	2
4	*Laboratory work 1. Big data analytics and machine learning.*	*Lesson 1. Data analysis and preparation. Loading the dataset. Primary analysis: checking data types, detecting missing values, processing categorical variables, scaling numeric features.*	2
		Lesson 2. Building a machine learning model. Creating and training a machine learning model. Evaluating the model: accuracy, recall, precision, F1-measure.	2
		Lesson 3. Building an intelligent system. model integration. Data visualization.	2

5	Web scraping, web crawling, indexing, and web automation technologies.	Technologies and software methods for web scraping, web crawling, indexing and web automation . Machine-readable data and APIs. HTML parsers. DOM analysis. Semantic annotation recognition. Web page analyzers using computer vision.	2
6	Laboratory work 2. Analysis of financial tweets users for stocks traded on NYSE, NASDAQ and SNP.	Lesson 1. Preprocessing and analysis of tweets. Uploading the file. Data structure analysis: text cleaning (removing stop words, symbols, hashtags, mentions), normalization, tokenization, lemmatization. Studying the distribution of tweets by time, languages, and main topics.	2
		Lesson 2. Building a financial sentiment classifier. Designating or using existing labels for sentiment analysis (positive / negative / neutral sentiment). Text vectorization (TF-IDF or Embeddings). Model training (Logistic Regression, SVM, Naive Bayes or LSTM). Model evaluation by metrics (accuracy, F1-score). Analysis of results and tracking sentiment. Analysis of model results. Visualization of sentiment by companies (graphs, diagrams, comparisons). Building an interface and script for monitoring current sentiment regarding individual companies or cryptocurrencies.	2
7	Lecture 5. Virtualization technologies for big data analytics. Container technology for executing program code on the server.	The role of virtualization for big data analytics. Virtualization technologies. Layers of abstraction. Hypervisors. Container technology for executing program code on the server. Big data engineering. SaaS, PaaS and IaaS.	2
8	Lecture 6. Hadoop technologies for big data. MapReduce programming model and framework. HDFS capabilities.	Hadoop technologies for big data. Scalability with big data. MapReduce programming model and framework. Data storage and processing in distributed file systems. Distributed databases. HDFS distributed file system.	2
9	Lecture 7. Kafka distributed streaming platform. Cassandra DBMS.	Data ingestion problem. Kafka distributed streaming platform. Advantages of Cassandra distributed DBMS.	2
10	Lecture 8. Apache Spark platform and its capabilities.	The problem of computational function. Spark technology. Comparison of Spark and MapReduce. Capabilities of Spark MLlib, Spark Streaming GraphX modules.	2
11	Lecture 9. Apache Flink capabilities for streaming and batch data processing.	Apache Flink Architecture . Stateful Functions. Flink ML Capabilities	2
Topic 2. Software methods for big data analytics.			
12	Lecture 10. Distributed configuration and service management system for distributed applications ZooKeeper.	Running ZooKeeper in production. Physical hardware. Memory. CPU and GPU. Configuration parameters.	2

13	Laboratory 3. Distributed big data computing using Spark cluster.	Lesson 1. Installing and configuring Spark. Installing Apache Spark on a local machine (or in a Google Colab environment with PySpark). Checking its functionality. Connecting to a Spark session and configuring a Spark cluster.	2
		Lesson 2. Loading and processing big data. Import a large dataset in CSV format. Read data using Spark DataFrame API. Primary processing: cleaning, filtering, aggregations. Perform analytical operations in distributed mode: grouping, counting, sorting, calculating statistics. Plot and save results. Analyze the advantages of distributed processing compared to traditional processing.	2
14	Lecture 11. Using Hadoop YARN for Big Data Analytics.	Hadoop YARN Architecture. Map Reduce in Hadoop. Using Hadoop YARN for Big Data Analytics.	2
15	Lecture 12. Software methods for clustering big data.	Software methods for clustering big data . Using the K-means and Map Reduce algorithms.	2
16	Lecture 13. Predictive analytics of big data.	Predictive analytics of big data, basic methods and technologies. Regression analysis of big data.	2
17	Lecture 14. Using Big Cloud Providers Data .	Big Cloud Providers Overview Data . AWS Big Data Platform.	2
18	Lecture 15. Using the PageRank algorithm for big data analytics.	Page Rank Algorithm and its Software Implementation for Big Data Analytics . Property Graph. Example of a Property Graph. Graph Operators. Aggregated Operator List. Property Operators. Structural Operators. Join Operators . Neighborhood Aggregation. Aggregate Messages (aggregateMessages). Map Reduce Triplets Transition Guide (Legacy). Neighbor Gathering. Caching and Decaching. Pregel API. Graph Constructors.	2
19	Lecture 16. Conclusion lesson.	Repetition of studied material. Modular test work.	2

6. Self-study

The discipline "Big data and analytics" is based on independent preparation for classroom classes on theoretical topics.

№	Title of the topic to be studied independently	Number of hours	Literature
1	Preparation for lecture 1	3	1, pp. 9–39
2	Preparation for lecture 2	4	1, pp. 39–46; 113–137
3	Preparation for lecture 3	3	1, pp. 50–60
4	Preparation for laboratory work 1	4	1, pp. 198–214; 241–263
5	Preparation for lecture 4	4	1, pp. 61–84
6	Preparation for laboratory work 2	4	1, pp. 198–240

7	*Preparation for lecture 5*	*4*	*1, pp. 89–112*
8	*Preparation for lecture 6*	*4*	*1, pp. 138–154*
9	*Preparation for lecture 7*	*3*	*1, pp. 159–162*
10	*Preparation for lecture 8*	*4*	*1, pp. 163–171*
11	*Preparation for lecture 9*	*4*	*1, pp. 172–190*
12	*Preparation for lecture 10*	*3*	*3 (add.), 4 (add.)*
13	*Preparation for laboratory work 3*	*4*	*1, pp. 163–171*
14	*Preparation for lecture 11*	*4*	*1, pp. 141–145*
15	*Preparation for lecture 12*	*4*	*1, pp. 224–234*
16	*Preparation for lecture 13*	*4*	*1, pp. 203–214*
17	*Preparing for lecture 14*	*4*	*1, pp. 297–309, 5 (add.), 6 (add.)*
18	*Preparing for lecture 15*	*4*	*7 (add.)*
19	*Preparation for the modular test*	*6*	*1, pp. 9–309, 8 (add.), 9 (add.), 10 (add.), 11 (add.)*

Policy and Assessment

7. Course policy

Forms of organizing the educational process, types of training sessions and assessment of learning outcomes are regulated by the Regulations on the Organization of the Educational Process at the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".

***Class attendance policy**. Attendance at lecture classes is mandatory. Attendance at laboratory classes may be occasional and required for the defense of laboratory work. The presence or absence of a student at a class is not assessed by awarding or deducting points. If a student cannot attend classes, he or she is still responsible for studying the theoretical material and completing practical assignments.*

***Policy on ethical norms in the classroom**: discipline; compliance with subordination; honesty; responsibility; respect for those present, turning off phones.*

***Policy on assessing learning outcomes**. The policy on assessing learning outcomes is regulated by the Regulations on the system of assessing learning outcomes at Igor Sikorsky Kyiv Polytechnic Institute. According to the Regulations, each grade is given in accordance with the criteria developed by the teacher and announced to students in advance. If a student fails to complete all four laboratory tests, he/she will not be allowed to take the test. Failure to pass the current control measure (modular test) without good reason is assessed as 0 points.*

The policy and principles of academic integrity are regulated by the norms set of the Code of Honor of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" (https://kpi.ua/files/honorcode_en.pdf); Regulations on the Organization of the Educational Process, Regulations on the System for Preventing Academic Plagiarism, Regulations on the Commission on Ethics and Academic Integrity. Plagiarism and other forms of violation of the principles of academic integrity are unacceptable. The student must complete all laboratory practical tasks independently using open sources of information and acquired knowledge and skills.

Plagiarism and other forms of violation of the principles of academic integrity are unacceptable. All work for current and semester tests must be completed independently by the student using open sources of information and the acquired knowledge and skills.

All works that violate the principles of academic integrity (the program code does not match the assignment option, the identity of the program code among different works, etc.) are evaluated at 0 points. To gain access to the test, the student must independently complete the laboratory work (without changing the current rating). In the case of semester control work that violates the principles of academic integrity, the semester control report is marked "Eliminate".

Policy on appealing the results of assessment of control measures. According to the "Regulations on resolving conflict situations at Igor Sikorsky Kyiv Polytechnic Institute" (https://osvita.kpi.ua/sites/default/files/downloads/regulations_resolving_conflict_situations_2020.pdf) , students have the right to appeal the results of control measures with arguments, explaining which criterion they disagree with according to the assessment. A student can raise any issue related to the procedure of control measures and expect that it will be considered in accordance with predetermined procedures.

Policy on assigning incentive points. According to the "Regulations on the system of assessing learning outcomes at Igor Sikorsky Kyiv Polytechnic Institute", incentive points are not included in the main RSO scale, and their sum cannot exceed 10 points.

Incentive points are awarded for a creative approach in performing laboratory work (the maximum number of points for all work is 10 points), as well as for participation in scientific projects and conferences related to the topic of this course.

8. Monitoring and grading policy

*During the semester, students perform **3 laboratory works.** The maximum number of points for each laboratory work: 20 points.*

Points are awarded for:
- quality of performance of the laboratory work: 0-8 points;
- answer during the defense of the laboratory work: 0-8 points;
- timely submission of work for defense: 0-4 points.
Performance evaluation criteria:
7-8 points – the work is done qualitatively, in full;
5-6 points - the work is done qualitatively, in full, but has shortcomings;
3-4 points - the work is done qualitatively, but not in full, has flaws;
1-2 points – the work is not done well, not in full, has flaws;
0 points – the work is incomplete or contains significant errors.
Answer evaluation criteria:
7-8 points – the answer is complete, well-argued;
5-6 points – the answer is incomplete, but well argued;
3-4 points – there are minor errors in the answer;
1-2 points – there are significant errors in the answer;
0 points - there is no answer or the answer is incorrect.
Criteria for evaluating the timeliness of work submission for defense:
4 points – the work is presented for defense no later than the specified deadline;
0-3 points – the work is submitted for defense later than the specified deadline.

The maximum number of points for performing and defending computer practicals:
20 points × 3 comp. practice = 60 points.

*The assignment for **the modular test** consists of 5 questions - 3 theoretical and 2 practical.*

The answer to each theoretical/practical question is evaluated by 8 points.
Evaluation criteria for each theoretical/practical question of the modular test:
7-8 points – the answer is correct, complete, well-argued;
5-6 points – the answer is correct, but incomplete or poorly argued;
3-4 points – there are minor errors in the answer;
1-2 points – there are significant errors in the answer;
0 points - there is no answer or the answer is incorrect.

The maximum number of points for a modular test:

8 points × 3 theoretical questions + 8 points × 2 practical questions = 40 points.

The rating scale for the discipline is equal to:

$R = R_{laboratory\ work} + R_{modular\ test} = R_S = 60$ points + 40 points = 100 points.

Calendar control: is carried out twice a semester as a monitoring of the current state of fulfillment of the syllabus requirements.

At the first attestation, the student receives "passed" if his current rating is at least 50% of the maximum number of points that the student can receive before the first certification (20 points).

At the second attestation, the student receives "passed" if his current rating is at least 50% of the maximum number of points that the student can receive before the second certification (30 points).

Semester control: ***Final test****.*

Conditions for admission to semester control:

with a semester rating of not less than 60% (60 points) and the enrollment of all the works of the computer workshop.

Completion and defense of a computer workshop is a necessary condition for admission to the final test.

Table of correspondence of rating points to grades on the university scale:

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

According to the Regulations on the recognition of learning outcomes acquired in non-formal/informal education at Igor Sikorsky Kyiv Polytechnic Institute https://osvita.kpi.ua/node/179, during the study of the discipline, it is possible to enroll in an online course or other element of non-formal education in the discipline.

The availability of a certificate of completion of a similar course in big data analytics is evaluated as 20 points (if the course topic corresponds to the subject of the laboratory practicum). Writing research articles or participating in conferences/projects in the relevant field of this discipline can also be evaluated as an additional 5 points.

For example, a laboratory practicum may be automatically credited with 20 points upon submission in Classroom of a Coursera certificate corresponding to the practicum's subject. Certificates of completion of the following or similar courses may be used to credit Computer Workshop 1 (20 points):

Introduction to Big Data with Spark and Hadoop (IBM) *— a course with practical lab exercises on Hadoop, Spark, HDFS, Hive, MapReduce, SparkSQL in a Jupyter/Docker environment.*

https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop

Machine Learning With Big Data *— overview of machine learning methods for big data using Spark, with practical tasks for scaling ML models.*

https://www.coursera.org/learn/big-data-machine-learning

Google Data Analytics Professional Certificate (Google) *— a series of 8 courses covering data analysis to practical projects.*

https://www.coursera.org/professional-certificates/google-data-analytics

Certificates of completion of the following or similar courses may be used to credit Computer Workshop 2 (20 points):

NLP: Twitter Sentiment Analysis *— a practical project implementing a Naive Bayes–based classifier for sentiment analysis on large datasets of tweets, including financial messages.*

https://www.coursera.org/projects/twitter-sentiment-analysis

Certificates of completion of the following or similar courses may be used to credit Computer Workshop 3 (20 points):

Introduction to Big Data with Spark and Hadoop *— covering both Hadoop and distributed processing with Spark.*

https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop

Scalable Machine Learning on Big Data using Apache Spark *— course on using Spark for scalable ML tasks on clusters, SparkSQL, SparkML Pipelines.*

https://www.coursera.org/learn/machine-learning-big-data-apache-spark

Big Data Analysis with Scala and Spark *— distributed computing with Scala and Spark, including the MapReduce paradigm.*

https://www.coursera.org/learn/scala-spark-big-data

Big Data Specialization (UC San Diego) *— specialization covering Hadoop, Spark, MapReduce, and Hive.*

https://www.coursera.org/specializations/big-data

etc.

The student must inform the instructor about a completed or planned course and clarify the results and prospects for crediting the acquired learning outcomes obtained through non-formal or informal education.

Syllabus of the course

Is designed by teacher PhD, Associate Professor, Liubov Oleshchenko

Adopted by Computer Systems Software Department (protocol № 12, 11.02.2026)

Approved by the Faculty Board of Methodology (protocol № 7, 23.02.2026)