Syllabus
Course Program

Big data infrastructure and management

Specialty
113 Applied mathematics

Institute
Educational and Scientific Institute of Computer Science and Information Technology

Educational program
Intelligent Data Analysis

Department
Computer Mathematics and Data Analysis

Level of education
Bachelor's level

Course type
Special (professional), Selective

Semester
7

Language of instruction
Ukrainian

Lecturers and course developers

Костюк Ольга Василівна

Olha.Kostiuk@khpi.edu.ua
Candidate of Technical Sciences, Associate Professor of the Department of Computer Mathematics and Data Analysis of KhPI National Technical University

Author of scientific and educational and methodical works. Leading lecturer in the disciplines: "Decision theory", "Fuzzy models and methods", "Introduction to the specialty and engineering activity"

General information

Summary

The discipline is aimed at mastering the theoretical foundations of the main principles, approaches and directions of big data technologies and infrastructure. Approaches and definitions are considered, an overview of the big data system is provided, and the topic of big data management systems and their practical application is revealed.

Course objectives and goals

The goal of studying the discipline is to acquire the necessary competencies for applied application of the basics of managing large-scale databases, in particular: the ability to discuss and solve problems that arise in connection with big data; the ability to compare and contrast different software and infrastructure architectures that can be applied to manage big data and choose the appropriate architecture for a given problem; the ability to compare different approaches to the implementation of big data systems; the ability to design and build a large, scalable database and use it to answer complex queries; the ability to identify and eliminate performance bottlenecks in the architecture and implementation of the data environment.

Format of classes

Lectures, laboratory classes, self-study, consultations. The final control is in the form of an exam.

Competencies

GC 1. Ability to learn and master modern knowledge.
GC 2. Ability to apply knowledge in practical situations.
GC 7. Ability to search, process and analyze information from various sources.
GC 8. Knowledge and understanding of the subject area and understanding of professional activities.
GC 10. Skills in the use of information and communication technologies.
SC 5. Ability to develop algorithms and data structures, software tools and program documentation.
SC 6. Ability to design databases, information systems and resources.
SC 8. Ability to operate and maintain software of automated and information systems for various purposes.
SC 21. Ability to develop and operate software tools for processing large amounts of data based on information technologies of distributed and cloud computing.

Learning outcomes

LO 1. Demonstrate knowledge and understanding of basic concepts, principles, theories of applied mathematics and use them in practice.
LO 7. Be able to conduct practical research and find a solution of incorrect tasks.
LO 11. Be able to apply modern programming technologies and software development, software implementation numerical and symbolic algorithms.
LO 13. To use specialized software programs in practical work products and software systems for computer mathematics.
LO 14. Demonstrate the ability to self-learn and continue professional development.
L0 25. Be able to apply modern information technologies and software for processing large amounts of data based on distributed and cloud services.

Student workload

The total volume of the course is 120 hours (4 ECTS credits): lectures – 30 hours, laboratory classes – 30 hours, self-study – 60 hours.

Course prerequisites

"Theory and design of databases", "Algorithmic languages (optional)", "Algorithmization and programming", "Discrete structures and data structures"

Features of the course, teaching and learning methods, and technologies

A feature of teaching is the use of elements of project work.

Program of the course

Topics of the lectures

Topic 1. Basics of Big Data
- basic concepts;
- technologies;
- features;
- the history and importance of big data.

Topic 2. Reference architecture of Big Data and examples of use
- overview of the main components;
- design principles.

Topic 3. Concept of Open Data
- datasets for data analysis and machine learning.

Topic 4. Basics of Data Warehousing
- Data Warehouse;

- layers.

Topic 5. Features of Data Warehousing
- denormalization;
- canonical uniform.

Topic 6. Basics of SQL, relational databases
Topic 7. Types and overview of NoSQL databases
Topic 8. Hadoop Ecosystem for Big Data
Topic 9. Components of the Hadoop ecosystem
- HDFS;
- HBase;
- MapReduce;
- YARN.

Topic 10. Components of the Hadoop ecosystem for Big Data processing
- Hadoop Data Warehouse Hive,
- data flow processing with Pig.

Topic 11. Basics of Apache Spark
- processing with spark;
- Spark SQL.

Topic 12. Components of Apache Spark:
- data bricks;
- delta;
- data frames.

Topic 13. Using Apache Spark mllib for machine learning
Topic 14. Basics of Kafka
Topic 15. Real-time data processing (stream processing) with Spark

Topics of the workshops

Workshops are not provided for in the curriculum.

Topics of the laboratory classes

Laboratory work 1. Building a database scheme (ER diagram)
Laboratory work 2. An example of Web scraping using Python
Laboratory work 3. An example of Web scraping using R-language
Laboratory work 4. Data import from open sources
Laboratory work 5. Creating an interactive Jupyter notebook using Google Colab and Binder
Laboratory work 6. Working with SQL: basic commands
Laboratory work 7. Working with NoSQL databases
Laboratory work 8. Execution of a simple task with MapReduce
Laboratory work 9 Setting up a single-node Hadoop cluster using Python
Laboratory work 10. Creation of interactive visualizations of big data
Laboratory work 11. Introduction to Apache Spark
Laboratory work 12. Big Data Project Management Using Jira: Planning and Tracking Project Progress
Laboratory work 13. Big Data Security: Developing and testing security measures to protect big data.
Laboratory work 14. Stream data processing with Kafka: Working with real data streams
Laboratory work 15. Development and analysis of solutions for real business problems based on big data

Self-study

Big Data infrastructure and its components.
Big Data in industry and the concept of Industry 4.0.
Application of MapReduce algorithms and distributed data storage technologies.
Learning the Java MapReduce library.
Programming for Spark.
Learning the PySpark library.
Working with Kafka.

Non-formal education

Within the framework of non-formal education according to the relevant Regulation (z0328-22), the educational component or its separate topics can be taken into account in case of independent completion of professional courses/training, obtaining civic education, online education, professional internship, etc.
In particular, individual topics of this component may be taken into account upon successful completion of the following courses:
- Topic 6. Basics of SQL, relational databases
https://www.coursera.org/specializations/cloudera-big-data-analysis-sql
- Topic 7. Types and overview of NoSQL databases
https://www.coursera.org/learn/introduction-to-nosql-databases
- Topic 8. Hadoop Ecosystem for Big Data
https://www.coursera.org/learn/hadoop
- Topic 13. Using Apache Spark mllib for machine learning
https://www.coursera.org/learn/machine-learning-big-data-apache-spark
- Topic 14. Basics of Kafka
https://www.coursera.org/projects/googlecloud-creating-a-streaming-data-pipeline-with-apache-kafka-totbh

Course materials and recommended reading

Basic literature

1. Wiktorski Tomasz. Data-intensive Systems: Principles and Fundamentals using Hadoop and Spark. Springer, 2019. – 105 p.
2. Zgurovsky M.Z., Zaychenko Y.P. Big Data: Conceptual Analysis and Applications. Springer, 2020. – 298 p.
3. Akerkar R. Models of Computation for Big Data Cham: Springer International Publishing, 2018. – 110 p.
4. Raheem N. Big Data: A Tutorial-Based Approach. Taylor & Francis Group LLC, CRC Press, 2019. – 203 p.

Additional literature

1. Davy Cielen, Arno D. B. Meysman, and Mohamed Ali. Introducing Data Science. Big data, machine learning, and more, using Python tools https://www.manning.com/books/introducing-data-science
2. Технології Big Data. Лабораторний практикум. Навч. посібник для здобувачів ступеня магістр за спеціальністю 123 «Комп'ютерні системи та мережі» / Таран В. К.: КПІ ім. Ігоря Сікорського, 2022. – 26 с.
https://comsys.kpi.ua/metodichni-vkazannya-po-disciplinam
3. Peter Ghavami. Big Data Management. Data Governance Principles for Big Data Analytics
https://doi.org/10.1515/9783110664065

Assessment and grading

Criteria for assessment of student performance, and the final score structure

Description of the final score structure, course requirements, and necessary steps to earn points, especially paying attention to self-study and individual assignments.

Grading scale

Total points	National	ECTS
90–100	Excellent	A
82–89	Good	B
75–81	Good	C
64–74	Satisfactory	D
60–63	Satisfactory	E
35–59	Unsatisfactory (requires additional learning)	FX
1–34	Unsatisfactory (requires repetition of the course)	F

Norms of academic integrity and course policy

The student must adhere to the Code of Ethics of Academic Relations and Integrity of NTU «KhPI»: to demonstrate discipline, good manners, kindness, honesty, and responsibility. Conflict situations should be openly discussed in academic groups with a lecturer, and if it is impossible to resolve the conflict, they should be brought to the attention of the Institute's management.
Regulatory and legal documents related to the implementation of the principles of academic integrity at NTU «KhPI» are available on the website: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Approval

Approved by	Date, signature 29.08.2024		Head of the Department Olena AKHIIEZER
	Date, signature 29.08.2024		Guarantor of the Educational Program Olena AKHIIEZER