Syllabus
Course Program

BigData technologies

Specialty
121 – Software Engineering
122 – Computer Science

Institute
Institute of Computer Science and Information Technology

Educational program
Software Engineering
Computer Science and Intelligent Systems

Department
Software Engineering and Management Intelligent Technologies (321)

Level of education
Bachelor's level

Course type
Special (professional), Elective

Semester
5

Language of instruction
English, Ukrainian

Lecturers and course developers

Volodymyr Petrovych Burdaev

volodymyr.burdaev@khpi.edu.ua
Ph.D., S.Sc., Associate Professor of the Department of Software Engineering and Intelligent Management Technologies

Google Scholar:https://scholar.google.com/citations?user=&user=RX9JedIAAAAJ
ORCID: https://orcid.org/0000-0001-9848 -9059
Scopus: https://www.scopus.com/authid/detail.uri?authorId=6507982230, https://www.scopus.com/authid/detail.uri?authorId=57224197566.
More about the lecturer on the department's website

General information

Summary

The task of the discipline is for students to acquire the required level of knowledge in the use of big data technology: storage, processing, analysis, visualization and application. Students will study the basics of modern computing platforms for big data, as well as remote service functions based on the organization and application of cloud computing technologies

Course objectives and goals

Formation of students' system of theoretical knowledge and acquisition of practical abilities and skills regarding the use of Big Data technologies, basic models of providing cloud computing services, development of web applications for conducting scientific research in a cloud environment

Format of classes

Lectures, laboratory classes, self-study, consultations. Final control in the form of credit.

Competencies

121 - Software Engineering
К05. Ability to learn and master modern knowledge.
К06. Ability to search, process and analyze information from various sources.

К19. Knowledge of data information models, ability to create software for storing, extracting and processing data.
К24. Ability to carry out the system integration process, apply change management standards and procedures to maintain the integrity, overall functionality and reliability of the software.
К25. Ability to reasonably choose and master the tools for software development and maintenance.
122 - Computer Science and Intelligent Systems
GC1. Ability to think abstractly, analyze and synthesize.
GC2. Ability to apply knowledge in practical situations.
GC3. Knowledge and understanding of the subject area and understanding of professional activities.
GC6. Ability to learn and master modern knowledge.
GC9. Ability to work in a team.
PC8. Ability to design and develop software using various programming paradigms: generalized, object-oriented, functional, logical, with appropriate models, methods and algorithms of computation, data structures and control mechanisms.
PC9. Ability to implement a multi-level computing model based on client-server architecture, including databases, knowledge and data warehouses, to perform distributed processing of large data sets on clusters of standard servers to meet the computing needs of users, including cloud services.

Learning outcomes

121 - Software Engineering
PLO01. Analyze, purposefully search and select information and reference resources and knowledge necessary for solving professional problems, taking into account modern achievements of science and technology.
PLO07. To know and apply in practice the fundamental concepts, paradigms and basic principles of functioning of language, tools and computing tools of software engineering.
PLO13. Know and apply methods of developing algorithms, designing software and data structures and knowledge.
PLO14. Apply in practice instrumental software tools for domain analysis, design, testing, visualization, measurement and software documentation.
PLO15. Motivated to choose programming languages and development technologies to solve the tasks of creating and maintaining software.
PLO18. To know and be able to apply information technologies for data processing, storage and transmission.
122 - Computer Science and Intelligent Systems
PLO9. Develop software models of subject environments, choose a programming paradigm from the standpoint of convenience and quality of application for the implementation of methods and algorithms for solving problems in the field of computer science.
PLO10. To use tools for developing client-server applications, design conceptual, logical and physical models of databases, develop and optimize queries to them, create distributed databases, data warehouses and showcases, knowledge bases, including cloud services, using web programming languages.

Student workload

The total volume of the course is 150 hours (5 ECTS credits): lectures – 32 hours, laboratory classes – 32 hours, self-study – 86 hours.

Course prerequisites

Object-oriented programming
Basics of web development

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

Teaching and learning methods:
interactive lectures with presentations, discussions, laboratory classes, teamwork, case method, student feedback, problem-based learning.
Forms of assessment:

written individual assignments for laboratory work (CAS), assessment of knowledge in laboratory classes (CAS), express surveys (CAS), online tests (CAS), final/semester control in the form of a semester exam, according to the schedule of the educational process (FAS).

Program of the course

Topics of the lectures

Topic 1. Sources of Big Data. Definition of Big Data
The Internet of Things and the growth of data. Defining Big Data. Big data examples in the real world. Open data. Data privacy. Structured and unstructured data. Cloud computing. Big data infrastructure. Distributed data and its processing.
Topic 2. Ontology for Big Data
Human brain and ontology. Ontology properties. Advantages of ontologies. Components of ontologies. The role of ontology in Big Data. Alignment of ontologies. Objectives of ontology in Big Data. Ontology problems in Big Data. RDF is a universal data format. Using OWL, the Web Ontology Language.
Topic 3. Study of Big Data
Supervised and unsupervised machine learning. The Spark programming model. Spark MLlib library. Pipeline. Regression analysis. Linear regression. The method of least squares. Data clustering. K-means algorithm. Content-based recommendation systems.
Topic 4. Virtualization technologies
Basic components of cloud computing. Concept of virtualization of computer systems and networks. Overview of network virtualization systems, computer resources, applications and data storage. Definition of application and operating system level virtualization. Server virtualization. Concept of virtualization of operating systems. Familiarization with the concepts and technologies of converting a server solution to a virtual machine, migration of virtual machines and "live migration" data.
Topic 5. Stacks of cloud platforms
Classification of cloud computing systems. Definition of systems: IaaS, PaaS, SaaS. IaaS is infrastructure as a service. PaaS is a platform as a service. SaaS - software as a service. The concept of a business model for providing software for rent. An overview of the main cloud computing providers.
Topic 6. Microsoft Azure cloud platform
Features of the platform. Historical information about the implementation of the platform. The main components of the platform. Technologies supported by the Microsoft Azure cloud. Application examples.
Topic 7. Open data, their formats and processing tools
The capabilities of data analysis tools using Python Randas. The role of Python in data analysis. Traditional big data analytics and next generation analytics. Life cycle of data analysis. Open data, their formats and means of processing Web scraping. Data extraction, conversion and loading.
Topic 8. Formatting time and date data, reading and writing files in Python
Interaction with external applications. Formatting time and date data in Python. Reading and writing files in Python. Interaction with external applications.
Topic 9. Python programming and basic SQL operations
Python work with SQL. Python and three types of relational databases: SQLite, MySQL, and PostgreSQL. Using Jupyter Notebook for data processing. Efficient navigation in the notebook. Features of Python for practical data science. Dictionaries. Classes. Visualizations in Jupyter Notebook.
Topic 10. The procedure for importing data from files in Pandas
Import data from the Internet. Tools for correlation analysis in Pandas. Statistical approaches to Big Data analytics. Using Pandas. Import data from files. Import data from the Internet. Descriptive statistics in Pandas. Tools for correlation analysis in Pandas.
Topic 11. Converting data types and manipulating dataframes in Python
Processing of missing data. Conversion of data types. Manipulation of data frames. Methods and types of machine learning analysis. Regression analysis. Types of regression analysis. Application of regression analysis.
Topic 12. Errors in data analysis and predictive analytics
Estimating regression errors using Python tools. Purpose of the scikit-learn library. Errors in data analysis and predictive analytics. Estimating regression errors using Python tools. Purpose of the scikit-learn library.
Topic 13. Data classification algorithms

Applications and problems of classifications. Decision tree classifier model. Problems of classification. Classification algorithms. Visualization of classifications. Application and validation of classifications.

Topic 14. Types of data visualization

Visualization of anomalies. Pyplot module. The Plotly tool. Visualization of anomalies.

Topic 15. The Apache Spark platform

A computational function problem. Spark technology. A comparison of Spark and MapReduce. Resilient Distributed Dataset (RDD). DataFrame. DataSet. Machine learning library - Mllib. ML Pipelines.

Topic 16. Big data and artificial intelligence systems

Pyramid of results. What the human brain does best. Touch input. Storage. Computing power. Low power consumption. What the electronic brain does best. Complete processing. Big data. Evolution from dumb to smart machines. Intelligence. Types of intelligence. Classification of intelligence tasks. Big data platforms. Batch processing. Real-time processing. Intelligent programs with big data. Areas of artificial intelligence.

Topics of the workshops

Workshops are not provided within the discipline.

Topics of the laboratory classes

Topic 1: Deploy ASP.NET web applications to Azure App Service using Visual Studio
Topic 2. Version control systems (Git)
Topic 3. Database deployment in Microsoft Azure. Creation of informational model and queries
Topic 4. SQL Server DB migration to Microsoft Azure using the Data Migration Assistant
Topic 5. Big data processing with Apache Spark
Topic 6. Big data processing with Apache Spark (DataFrame API and Spark SQL)
Topic 7. Building predictive models in Microsoft Azure ML Studio (regression model)
Topic 8. Building predictive models in Microsoft Azure ML Studio (classification model, clustering model)

Self-study

Individual assignments are not provided in the curriculum.
Students are recommended with additional materials (videos, articles) for self-study and processing.

Course materials and recommended reading

Key literature

1. J. Perrin, Spark in action (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala), Manning, 2020, 576 p.
2. J. Damji, B. Wenig, T. Das, D. Lee, Learning spark (2nd ed.) O'Reilly Media Inc., 2020, 672 p.
3. S. Nudurupati, Essential PySpark for scalable data analytics: A beginner's guide to harnessing the power and ease of PySpark 3 Packt Publishing, 2021, 308 p.
4. K. Ramcharan, K. Sundar, S. Alla, Applied data science using PySpark: Learn the end-to-end predictive model-building cycle Apress, 2020, 138 p.
5. M. Lathkar, Python Data Persistence: With SQL and NoSQL Databases, BPB Publications, 2019, 316 p.

Additional literature

1. B. Murray, Big Data for Beginners: Book 1 - An Introduction to the Data Collection, Storage, Data Cleaning and Preprocessing, Independently published, 2023, 57 p.
2. Saswat Sarangi, Pankaj Sharma, Big Data: A Beginner's Introduction, Routledge India, 2019, 138 p.
3. B. Murray. Big Data for Beginners: Book 2 - An Introduction to the Data Analysis, Visualization, Integration, Interoperability, Governance and Ethics, Independently published, 2023, 60 p.
4. Thomas H. Davenport, Nitin Mittal, All-in On AI: How Smart Companies Win Big with Artificial Intelligence, Harvard Business Review Press , 2023, 224 p.
5. A. J. Gutman, J. Goldmeier, Becoming a Data Head: How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning, Wiley, 2021, 272 p.

Assessment and grading

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

100% of the final grade consists of the results of the assessment in the form of credit (40%) and current assessment (60%).
40% credit
60% current assessment:
Test №1 (10%)
Test №2 (10%)
Laboratory works (40%)
Laboratory work №1 (5%)
Laboratory work №2 (5%)
Laboratory work №3 (5%)
Laboratory work №4 (5%)
Laboratory work №5 (5%)
Laboratory work №6 (5%)
Laboratory work №7 (5%)
Laboratory work №8 (5%)

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 08.06.2023 Head of the department
Ihor HAMAIUN

08.06.2023 Guarantor of the educational program
Andrii KOPP
Uliya LITVINOVA