Syllabus
Course Program

METHODS AND TECHNOLOGIES OF WORKING WITH BIG DATA

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
Master's level

Course type
Special (professional), Selective

Semester
1

Language of instruction
English,

Lecturers and course developers

First name and surname

leonid.liubchyk@khpi.edu.ua
DSc, Professor of CMAD department.

The number of scientific and educational publications is more than 200.
Leading Topicr in the disciplines: "Control theory", "Incorrect data processing problems", "Predictive analysis".
More about the lecturer on the department's website

First name and surname

klym.yamkovyi@cs.khpi.edu.ua
PhD, Assistant Professor of CMAD department.

The number of scientific and educational publications is more than 10.
More about the lecturer on the department's website

General information

Summary

Functional capabilities of the main infrastructure components and Big Data tools are studied. Emphasis is placed on understanding and using the Apache Hadoop ecosystem as the primary platform for Big Data, its core functional components MapReduce, Spark, HBase, Hive, Pig, and the supported Pig Latin and Hive programming languages. The course also provides information on security issues and compliance with industry data management requirements, including issues related to the EU General Data Protection Regulation (GDPR).

Course objectives and goals

The purpose of studying the discipline is to study the main concepts of Big Data and related technologies, to acquire knowledge and skills in the selection and evaluation of Big Data infrastructure services from the main cloud service providers (Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform

(GCP) and others), to solve the main tasks of managing and analyzing enterprise data, acquiring skills for choosing and deploying a Hadoop or Spark cluster on one of the cloud platforms (Azure HDInsight, Amazon EMR). or others), programming tasks using one of the scripting languages HiveQL, Pig Latin.

Format of classes

Lectures, laboratories, self-study, consultations. The final control is an exam

Competencies

SC 10. Ability to conduct mathematical and computer modeling, data analysis and processing, computational experiments, solving formalized problems using specialized software tools.
SC 19. Ability to apply mathematical methods and algorithms of machine learning, soft computing and computational intelligence to analyze uncertain data, forecast and make decisions.

Learning outcomes

LO 11. To be able to apply modern technologies of programming and software development, software implementation of numerical and symbolic algorithms.
LO 23. Be able to apply existing and develop new algorithms and software tools for statistical and intellectual analysis of uncertain data.
LO 25. Be able to apply modern information technologies and software for processing large data sets based on distributed and cloud services.

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

Bachelor's level in the specialty "Applied Mathematics".

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

Labs are performed on a real cloud platform and Hadoop cluster (AWS or Azure).

Program of the course

Topics of the lectures

Topic 1. Basics of cloud technologies.
Topic 2. Models of cloud services, cloud resources, functioning of cloud services.
Topic 3. Reference architecture of Big Data and examples of use.
Topic 4. Basics of the Big Data.
Topic 5. Cloud platforms for Big Data. Review and comparison.
Topic 6. Cloud relational databases AWS RDS, AWS Aurora.
Topic 7. Basic concept of orchestration and AirFlow.
Topic 8. AirFlow components.
Topic 9. Modern large-scale databases AWS Aurora, Azure CosmosDB, Google Spanner.
Topic 10. Basics of Secondary Index: SOLR.
Topic 11. Basics of Elasticsearch/OpenSearch.
Topic 12. Basics of full-text search.
Topic 13. Basics of vector search.
Topic 14. Enterprise Big Data architecture and Big Data management.
Topic 15. Security problems of Big Data, data protection. Access control and identity management.
Topic 16. Review of the course and discussion of the achieved results.

Topics of the workshops

None.

Topics of the laboratory classes

Topic 1. The main providers of cloud services AWS, Microsoft Azure, Google Cloud Platform.
Topic 2. Working with the Amazon Web Services (AWS) cloud.
Topic 3. Deployment and access to EC2, S3, VM instances.
Topic 4. SSH client configuration and VM access.
Topic 5. Installation and configuration of Airflow.
Topic 6. Creating an ETL pipeline in Airflow.
Topic 7. Performing a simple task with MapReduce.
Topic 8. Installing and configuring a stand-alone Hadoop cluster for personal use.
Topic 9. Working with the Hadoop cluster.
Topic 10. Getting to know the Hue interface, downloading data and files.
Topic 11. Creating an ETL pipeline in Hue.
Topic 12. Working with SQL, basic commands.
Topic 13. Introduction to AWS RDS cloud services for creating your own relational database.
Topic 14. Working with AWS RDS Maria DB.
Topic 15. Working with AWS RDS PostgreSQL.
Topic 16. Working with the Databrick online educational cluster.
Topic 17. Working with Spark SQL and DataFrame API.
Topic 18. Introduction to Amazon OpenSearch Service.
Topic 19. Working with full-text search in OpenSearch.
Topic 20. Work with vector search in OpenSearch.
Topic 21. Project management regarding Big Data.
Topic 22. Data life cycle management and DataOps.
Topic 23. Definition of Big Data corporate infrastructure and data processing services.
Topic 24. Selection of infrastructure services and components.

Self-study

Virtual hybrid/ dynamic cloud data center
Cloud outsourcing of the IT infrastructure of the enterprise.
Big Data infrastructure and its components.
Big Data in industry and the concept of Industry 4.0.
Familiarity with cloud platforms: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP)
Application of MapReduce algorithms and distributed data storage technologies.
Learning the Java MapReduce library.
AWS Elastic Map Reduce (EMR) Overview.
AWS cloud services for working with data; cloud data storage and database services.
Popular platforms for Spark, DataBricks.
Programming for Spark.
Data management maturity model and aspects of data quality assurance.
Data Management Plan (DMP). Principles of data efficiency FAIR (Findable – Accessible – Interoperable – Reusable).
Data governance in industry, ensuring data trust and data sovereignty.
Cloud compliance standards and assessment of cloud service provider services.

Course materials and recommended reading

1. Balusamy, Balamurugan, Seifedine Kadry, and Amir H. Gandomi. Big Data: Concepts, Technology, and Architecture. John Wiley & Sons, 2021.
https://media.wiley.com/product_data/excerpt/21/11197018/1119701821-11.pdf
2. M. Collier, R. Shahan. Microsoft Azure Essentials: Fundamentals of Azure, Second Edition. Microsoft Press, 2016. – 246 p. https://download.microsoft.com/download/6/6/2/662DD05E-BAD7-46EF-9431-135F9BAE6332/9781509302963_Microsoft%20Azure%20Essentials%20Fundamentals%20of%20Azure%202nd%20ed%20pdf.pdf
3. IoT Fundamentals: Big Data & Analytics. https://www.netacad.com/courses/iot/big-data-analytics

4. Apache Hadoop. http://hadoop.apache.org/
5. MapReduce Tutorial. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
6. Apache Spark. https://spark.apache.org/

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.

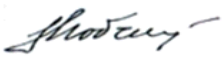
Grading scale

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 Topicr, 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 31.08.2023		Head of the department Olena AKHIIEZER
	Date, signature 31.08.2023		Guarantor of the educational program Leonid LYUBCHYK