

**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ**

ЗАТВЕРДЖЕНО

на засіданні кафедри
інформаційних систем
Протокол № 1 від 27.08.2024 р.

ПОГОДЖЕНО

Проректор з навчально-методичної роботи

Каріна НЕМАШКАЛО



**ВИСОКОПРОДУКТИВНІ СИСТЕМИ ОБРОБКИ
ТА АНАЛІЗУ ВЕЛИКИХ ДАНИХ**

робоча програма навчальної дисципліни (РПНД)

Галузь знань	12 "Інформаційні технології"
Спеціальність	122 "Комп'ютерні науки"
Освітній рівень	другий (магістерський)
Освітня програма	"Комп'ютерні науки"

Статус дисципліни
Мова викладання, навчання та оцінювання

**обов'язкова
англійська**

Розробник:
д.т.н., професор

підписано КЕП

Сергій МІНУХІН

Завідувач кафедри
інформаційних систем

Дмитро БОНДАРЕНКО

Гарант програми

підписано КЕП

Сергій МІНУХІН

Харків
2024

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF ECONOMICS

APPROVED

at the meeting of the department
information systems
Protocol No. 1 dated August 27, 2024

AGREED

Vice-rector for educational and methodical work

 Karina NEMASHKALO

**HIGH PERFORMANCE SYSTEMS OF PROCESSING
AND ANALYSIS OF BIG DATA**

Program of the course

Field of knowledge	12 "Information technologies"
Specialty	122 "Computer Sciences"
Study cycle	second (master's)
Study programme	"Computer Sciences"


Discipline status	mandatory
Language of teaching, learning and assessment	English

Developer:
Doctor of Science, Professor

digital signature

Serhii MINUKHIN

Head of the department
information systems



Dmytro BONDARENKO

Guarantor of the program

digital signature

Serhii MINUKHIN

Kharkiv

2024

INTRODUCTION

The conditions for the growth of data volumes and the increase in the dependence of the quality of the business processes of commercial activities of enterprises on the flows and intensity of data lead to the need to create distributed information systems that must ensure a sufficient level of operational efficiency of processing such data. The development of distributed and parallel computing technologies, as well as the availability of high-performance systems that are developing and are publicly available for commercial and research organizations to support their activities, now allow us to process large volumes of data efficiently enough. Software systems and technologies of scalable computing systems with the capabilities of distributed processing of extremely large data sets have a significant influence on this development.

The problem of scalability of computer systems with increasing volumes and intensity of data should be solved by modern means of distributed environments in combination with technologies of distributed and parallel computing, distributed file systems and data warehouses, which in general will ensure effective data processing.

The course "High Performance Systems of Processing and Analysis of Big Data" is studied by students of the specialty 122 "Computer Sciences" of the Educational and professional program "Computer Sciences" of all forms of education in the first year of study during the first semester. The study of the course involves acquiring theoretical knowledge and mastering practical skills related to the use of technologies for performing time-consuming tasks based on high-performance data processing and storage systems. The study of the course is aimed at forming among students of higher education a general understanding of the place and essence of big data in the modern activity of enterprises and institutions, the features of the use and processing of big data in comparison with classical standards and technologies of distributed and parallel processing, in particular, technologies of data storage and processing in real life time or in batch mode, mastering architectural solutions and components of modern high-performance computing ecosystems.

The purpose of teaching the course "High Performance Systems of Processing and Analysis of Big Data" is to provide higher education students with a system of theoretical knowledge and acquire practical skills to understand the essence of problems that arise when using big data, modern approaches and tools for their processing and analysis.

The tasks of the course are:

- acquisition of competences in working with big data, their analysis in order to make effective management decisions;
- acquisition of competences regarding the choice of framework architecture, selection, installation and configuration of software for work in software environments at the level of a distributed system and a local resource;

- acquisition of practical skills in the use of big data processing software systems using modern integrated software systems and technologies (frameworks);

- deployment and configuration basic software (frameworks) for launching, performing tasks and analyzing the obtained results using technologies and tools of distributed systems and modern paradigms of parallel programming.

The object of the course is the processes of processing and analyzing big data of various nature to improve the quality of management of enterprises and institutions.

The subject of the course is the methods, models and technologies of processing, storing and analyzing big data of various nature.

The results of training and competence, which are formed by the educational course, are presented in the table. 1.

Table 1

Learning outcomes and competences formed by the course

Learning outcomes	Competencies that must be mastered by a student of higher education
LO1	GC05, GC07, SC04, SC07
LO2	SC05, SC06, SC08, SC09, SC12
LO4	GC05, SC02, SC07, SC08, SC12
LO5	SC09
LO6	GC02, GC03, SC05, SC08, SC09, SC12
LO7	GC05, GC06, GC07, SC01, SC04, SC05
LO8	GC01, GC03, GC05, GC07, SC02, SC04, SC06, SC12
LO9	GC01, GC02, GC03, GC05, GC07, SC04, SC05, SC07, SC08, SC12
LO10	SC02, SC12
LO11	GC01, GC03, GC05, SC06, SC07, SC08, SC12
LO12	SC10
LO15	SC04
LO16	SC03, SC07, SC08
LO18	GC05, SC04
LO19	SC07, SC08, SC12
LO 20	GC01, GC02, GC03, GC05, GC07, SC01, SC02, SC03, SC04, SC05, SC06, SC07, SC11, SC12

where, LO1. Have specialized conceptual knowledge that includes modern scientific achievements in the field of computer science and is the basis for original thinking and conducting research, critical understanding of problems in the field of computer science and at the border of fields of knowledge.

LO2. Have specialized computer science problem-solving skills necessary for conducting research and/or carrying out innovative activities in order to develop new knowledge and procedures.

LO4. Manage work processes in the field of information technology, which are complex, unpredictable and require new strategic approaches.

- LO5. Evaluate the results of teams and collectives in the field of information technologies and ensure the effectiveness of their activities.
- LO6. Develop a conceptual model of an information or computer system.
- LO7. Develop and apply mathematical methods for the analysis of information models.
- LO8. Develop mathematical models and data analysis methods (including large ones).
- LO9. Develop algorithmic and software for data analysis (including large data).
- LO10. To design architectural solutions of information and computer systems for various purposes.
- LO11. Create new algorithms for solving problems in the field of computer science, evaluate their effectiveness and limitations on their application.
- LO12. Design and support databases and knowledge
- LO15. Identify the needs of potential customers regarding the automation of information processing.
- LO16. Conduct research in the field of computer science.
- LO18. Collect, formalize, systematize and analyze the needs and requirements for the information or computer system being developed, operated or supported.
- LO19. To analyze the current state and global trends in the development of computer sciences and information technologies.
- LO 20. Develop algorithms and software components of computer information systems for high-performance big data processing systems (including distributed and parallel computing) and cloud platform services.
- GC01. Ability to abstract thinking, analysis and synthesis.
- GC02. Ability to apply knowledge in practical situations.
- GC03. Ability to communicate in the national language both orally and in writing.
- GC05. Ability to learn and master modern knowledge.
- GC06. The ability to be critical and self-critical.
- GC07. Ability to generate new ideas (creativity).
- SC01. Awareness of the theoretical foundations of computer science.
- SC02. The ability to formalize the subject area of a certain project in the form of an appropriate information model.
- SC03. Ability to use mathematical methods to analyze formalized models of the subject area.
- SC04. The ability to collect and analyze data (including large data) to ensure the quality of project decision-making.
- SC05. Ability to develop, describe, analyze and optimize architectural solutions of information and computer systems for various purposes.
- SC06. Ability to apply existing and develop new algorithms for solving problems in the field of computer science.
- SC07. Ability to develop software according to formulated requirements, taking into account available resources and constraints.
- SC08. The ability to develop and implement software development projects, including in unpredictable conditions, with unclear requirements and the need to apply new strategic approaches, use software tools to organize teamwork on the project.
- SC09. Ability to develop and administer databases and knowledge bases.
- SC10. The ability to evaluate and ensure the quality of IT projects, information and computer systems of various purposes, to apply international standards for assessing the quality of software of information and computer systems, models for assessing the maturity of information and computer systems development processes.
- SC11. Ability to initiate, plan and implement the development processes of information and computer systems and software, including its development, analysis, testing, system integration, implementation and support.

SC12. Ability to develop, apply and integrate data processing and analysis technologies in high-performance systems and cloud platforms to ensure efficient use of computing resources of computer systems.

COURSE PROGRAM

Content of the academic course

Content module 1. Basic concepts, essence and features of big data. Principles of organizing the construction of systems for working with big data

Topic 1. Concepts, characteristics of big data and their processing systems

1.1. Concept, definition and characteristics of big data. The current state of the use of big data in global practice. The 5 Vs of Big Data: Volume, Velocity, Variety, Validity, and Value. Directions and application of big data in the activities of the world's leading companies.

1.2. Application of the latest information technologies in information systems for processing big data.

Topic 2. Modern big data processing systems. Composition of components and their purpose

2.1. Big data framework structure: Big Data Strategy, Big Data Architecture, Big Data Algorithms, Big Data Processes, Big Data Functions, Artificial Intelligence.

2.2. Architecture.

2.2. Data Storage component.

2.3. Calculation component.

2.4. Calculation modes: in-memory, parallel programming, distributed data storage.

2.5. Analysis component.

2.6. Classification characteristics of big data:

domain areas: Health and Medical Care, Social Networks and Internet, Government and Public Sector, Natural Resource Management, Economic and Business Sector;

data format: structured (relational databases); unstructured data (video, text, geographic location information, etc.); semi-structured (JSON, XML);

data processing modes: batch processing, stream processing, real-time processing.

Topic 3. Apache Hadoop: a framework for processing big data. Basic components for building Hadoop: Google's MapReduce, Google File System

3.1. Organization of distributed information processing in the Apache Hadoop framework. Composition of components for distributed processing, storage and parallel computing. Purpose of MapReduce and distributed data storage system

3.2. Google's MapReduce is a basic programming model for processing large data sets in a massively parallel way:

3.3. Google File System (GFS) is a basic file system for processing batch workloads of large volumes of data to ensure: fault tolerance, efficient processing of large files; optimization of processes of reading, writing and adding data.

Topic 4. Architecture of Apache Hadoop

4.1. Apache HDFS - Hadoop distributed file system: Master/Slave architecture. Purpose and functions of NameNode (Master node), Secondary NameNode and DataNodes (Slave nodes). Principles of organizing the interaction between the master and the working nodes of the cluster when processing tasks. Block organization of data storage on working nodes (Slave nodes). Writing and reading data to the HDFS system.

4.2. Apache Hadoop Map/Reduce.

Map/Reduce architecture. The content of the stages of task processing according to the Map/Reduce parallel programming paradigm: iteration on input data; calculating key/value pairs for each piece of input data; grouping of all intermediate values by key; iteration on resulting groups; reduction of each group. Composition and assignment of stages (phases) of parallel programming in Map/Reduce: map phase: splitting, mapping, partition, combined; reduce phase: read, sort, reduce. The composition and format of the input and output files of the tasks being processed. An example of using Map/Reduce for a test data set of Wordcount type.

Content Module 2. Apache Spark: A Universal Platform for Big Data Processing and Analytics

Topic 5. Architecture of Apache Spark.

5.1. Composition and purpose of components and tools. Apache Spark Deployment Systems.

5.2. Architecture and programming interfaces.

5.3. Spark Context, Spark Driver, Cluster Manager, Data Node, JobTracker, TaskTracker functions.

5.4. The composition and purpose of the components of the Spark ecosystem: Spark Core; Spark Streaming, Spark SQL, GraphX, MLlib.

Topic 6. Apache Spark deployment modes.

6.1. Client, Cluster modes.

6.2. Local mode.

6.3. Standalone Scheduler mode.

6.4. YARN mode.

6.5. Mesos mode.

Topic 7. Scheduling tasks in Apache Spark.

7.1. Apache Spark Driver: DAGScheduler, TaskScheduler, BlockManager.

7.2. DAGScheduler implementation for RDD operation.

Topic 8. Working with databases and data stores in SparkSQL. RDD, Dataframe and Dataset.

8.1. RDD: concept, purpose, and usage principles for structured data and SQL-like operations.

8.2. Concepts and functions of Dataframe. Features of implementation and use.

8.3. Dataset: basic functions and purposes for handling relational databases. The content is a combination of RDD and Dataframe functions.

Topic 9. Deployment and configuration of Apache Spark and Apache Hadoop frameworks in distributed and virtual environments.

9.1. Deployment and configuration of Apache Spark and Apache Hadoop frameworks in a distributed cluster. Selection and justification of deployment modes.

9.2. Deployment and configuration of Apache Spark and Apache Hadoop frameworks in a virtual environment (VirtualBox). Selection and justification of deployment modes.

The list of laboratory classes by course is given in the table. 2.

Table 2

List of laboratory classes

Topic name	Content
Topic 1-4, 5, 6, 9.Laboratory work No. 1	Installing and deploying Apache Spark using VAGRANT software
Topic 1-4, 5, 6, 9.Laboratory work No. 2	Installing an Apache Spark cluster offline
Topic 1-4, 6, 9.Laboratory work no3	Installing and configuring an Apache Spark YARN cluster

The list of self-studies by course is given in the table. 3.

Table 3

List of self-studies

Topic name	Content
Topic 1 - 9	Studying lecture material
Topic 1 - 9	Preparation for laboratory classes
Topic 1 - 9	Preparation for the exam

The number of hours of lectures, laboratory classes and hours of independent work is given in the work plan (technological map) for the academic course.

TEACHING METHODS

In the process of teaching an course, in order to acquire certain learning outcomes, to activate the educational process, it is envisaged to use such teaching methods as:

Verbal lectures (Topic 1-9), problematic lecture (Topic 7, 9), provocation lecture (Topic 8).

In person (demonstration (Topic 1-9)).

Laboratory work (Topic 1 – 4, 5, 6, 9).

FORMS AND METHODS OF ASSESSMENT

The university uses a 100-point accumulative system for evaluating the learning outcomes of higher education applicants.

Current control is carried out during lectures and laboratory classes and is aimed at checking the level of readiness of a higher education applicant to perform specific work and is evaluated by the sum of points scored:

– for courses with a form of semester control as an exam: maximum amount is 60 points; minimum amount required is 35 points.

Final control includes semester control and certification of the student of higher education.

Semester control is conducted in the form of a semester exam (exam). The semester exam (exam) is taken during the exam session.

The maximum number of points that a student of higher education can receive during the examination (examination) is 40 points. The minimum amount for which the exam is considered passed is 25 points.

The final grade in the course is determined:

– **for courses with a form of exam, the final grade is the amount of all points received during the current control and the exam grade.**

During the teaching of the academic discipline, the following control measures are used:

Current control: defense of laboratory work (50 points), written control work (10 points).

Semester control: exam (40 points).

More detailed information about the evaluation system is given in the work plan (technological map) for the academic course.

An example of an examination ticket and evaluation criteria for an academic course.

An example of an examination ticket

Semyon Kuznets Kharkiv National University of Economics

Second (master's) level

Specialty "Computer Sciences"

Educational and professional program "Computer sciences".

Semester 1

Course "High-performance systems of processing and analysis of big data"

EXAMINATION TICKET No. 1

Task1 (diagnostic, 10 points).

Describe and characterize the architecture of a typical Hadoop cluster. List the functions of the wizard.

Task 2(stereotypical, 10 points).

Describe the HDFS distributed file system in Hadoop. The purpose of the secondary NameNode node, its differences from the NameNode node. Give a rationale for its use in a cluster.

Task 3 (heuristic, 12 points). List the main deployment modes of Apache Spark. Provide the characteristics and features of the Standalone mode.

Task 4(stereotypical, 8 points).List the options and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

Protocol No. 1 of August 27, 2024

was approved at the meeting of the Department of Information Systems.

Examiner, Doctor of Technical Sciences, prof. Serhii MINUKHIN

Chief Dmytro BONDARENKO, Ph.D. of the department

Evaluation criteria

The final marks for the exam consist of the sum of the marks for the completion of all tasks, rounded to a whole number according to the rules of mathematics.

The algorithm for solving each task includes separate stages that differ in complexity, time-consumingness, and importance for solving the task. Therefore, individual tasks and stages of their solution are evaluated separately from each other in this way.

Task 1.

This task is evaluated on a 10-point scale.

A score of 10 points is given if the acquirer provides a complete composition of the components of the Apache Hadoop cluster architecture in accordance with the defined functions. The functions of the master and its place in the cluster architecture, the main differences from the functions of the working node are given and detailed.

A score of 9 points is given if the acquirer provides a complete composition of the components of the Apache Hadoop cluster architecture in accordance with the defined functions. However, the answer contains certain inaccuracies in defining the differences between the functions of the wizard and the worker nodes.

A score of 8 points is given if the acquirer does not fully provide the components of the Apache Hadoop cluster architecture in accordance with the defined functions. A comparison of the functions of the cluster nodes was made, but there are inaccuracies regarding the definition of the functionality of the wizard.

A score of 7 points is given if the acquirer does not fully provide the components of the Apache Hadoop cluster architecture in accordance with the defined functions. A comparison of the functions of the cluster nodes has been made, but there are errors regarding the presentation of their functionality characteristics.

A score of 6 points is given if the acquirer does not provide the composition and functions of the cluster components in full and with errors. The functions of the cluster management wizard are not fully described and explained.

A score of 5 points is given if the acquirer does not provide the composition and functions of the cluster components in full and with errors. The functions of the wizard or working nodes are not fully specified and justified.

A score of 4 points is given if the acquirer does not provide the composition and functions of the cluster components in full and with errors. The functions of the wizard and the working nodes of the cluster are not fully explained and substantiated.

A score of 3 points is given if the composition of the components of the cluster is given by the acquirer with errors. There are a significant number of errors and inaccuracies in the description of the functions of the wizard.

A score of 2 points is given if the acquirer incorrectly specified the composition of the components of the cluster architecture. There are a significant number of errors in the description of the functions of the wizard and working nodes.

A score of 1 point is given if the acquirer incorrectly specified the composition of the components of the cluster architecture. There is an incorrect description of the functions of the wizard.

A score of 0 points is given for failure to complete the task in general.

Task 2.

This task is evaluated on a 10-point scale.

A score of 10 points is given if the applicant has given the functions in all of the HDFS distributed file system, the purpose of the secondary NameNode node, its differences from the NameNode node. The need for its use from the point of view of fault tolerance and stability of cluster operation is presented and substantiated.

A score of 9 points is given if the acquirer has given the functions in full of the HDFS distributed file system, the purpose of the secondary NameNode node, its differences from the NameNode node. The necessity of its use is not fully substantiated from the point of view of fault tolerance and stability of cluster operation.

An estimate of 8 points is given if the acquirer has given the functions in full HDFS distributed file system, in particular, the purpose of the secondary NameNode, but its differences from the NameNode are not fully presented. In general, the necessity of its use is substantiated from the point of view of fault tolerance and stability of cluster operation.

A score of 7 points is given if the applicant does not provide the full range of functions of the HDFS distributed file system, in particular, regarding the purpose of the secondary NameNode node, its differences from the NameNode node are not fully presented. In general, the necessity of its use is substantiated from the point of view of fault tolerance and stability of cluster operation.

A score of 6 points is given if the applicant does not provide the full range of functions distributed file system HDFS, in particular, there are inaccuracies regarding the explanation of the purpose of the secondary NameNode node, its differences from the NameNode node are not fully presented. In general, the necessity of its use is substantiated from the point of view of fault tolerance and stability of cluster operation.

A score of 5 points is given if the applicant does not provide the full range of functions of the HDFS distributed file system, in particular, there are errors regarding the explanation of the purpose of the secondary NameNode node, there are inaccuracies regarding its definition of the difference from the NameNode node. There is no clear justification for the need to use it from the point of view of fault tolerance and stability of cluster operation.

A score of 4 points is given if the acquirer has given the functions with inaccuracies of the HDFS distributed file system, there are errors regarding the designation of the secondary NameNode, and inaccuracies in explaining the differences from the NameNode. There is no clear justification and explanation of the need for its use from the point of view of fault tolerance and stability of cluster operation.

A score of 3 points is given if the acquirer gives functions with errorsHDFS distributed file system, there are errors in assigning a secondary NameNode and errors in defining and explaining differences from a NameNode. There is no justification and explanation of the need for use of the secondary NameNodefrom the point of view of fault tolerance and stability of cluster operation.

An assessment of 2 points is given if the acquirer provides functions with certain errors of the HDFS distributed file system, there are errors regarding the designation of the secondary NameNode, errors regarding the definition and explanation of the differences from the NameNode. There is no justification and explanation of the need for use of the secondary NameNode from the point of view of fault tolerance and stability of cluster operation.

A score of 1 point is given if the applicant does not specify the functions of the distributed file system HDFS, the destination of the secondary node NameNode. There is no justification for the need for use of the secondary NameNode for cluster operation.

A score of 0 points is given for failure to complete the task in general.

Task 3.

This task is evaluated on a 12-point scale.

A score of 12 points is given if the acquirer provides a complete and explained list, features and general principles of each of the deployment modes of the Apache Spark cluster. A comprehensive description and features of cluster deployment in Standalone mode, the applied method of task queue processing, are provided.

An estimate of 11 points is given if the acquirer provides a complete and explained list, features and general principles of each of the deployment modes of the Apache Spark cluster. Sufficient characteristics and some features of the cluster deployment mode in Standalone mode, the applied method of processing the task queue are presented.

A score of 10 points is given if the acquirer does not list, features and general principles of each of the modes of deployment of the Apache Spark cluster in full and with appropriate explanations. An incomplete description and unclear features of the cluster deployment mode in Standalone mode, the applied task queue processing method, are provided.

A score of 9 points is given if the acquirer does not list, features and general principles of each Apache Spark cluster deployment mode in full and with appropriate explanations. An incomplete description and not all features of the cluster deployment mode in Standalone mode, the applied task queue processing method, are provided.

A score of 8 points is assigned if the acquirer does not provide a complete list, features and general principles of each of the deployment modes of the Apache Spark cluster without reasonable explanations. An incomplete description is given and the features of the cluster deployment mode in Standalone mode are not given in full. The algorithm of the used method of processing the task queue in the cluster is not described.

A score of 7 points is given if the acquirer does not provide a complete list, features and basic principles of each of the deployment modes of the Apache Spark cluster without reasonable explanations. An incomplete description is given and the features of the cluster deployment mode in Standalone mode are not given in full. The algorithm of the used method of processing the task queue in the cluster is not given.

A score of 6 points is given if the acquirer does not provide a complete list, features and basic principles of each of the deployment modes of the Apache Spark cluster without reasonable explanations. An incomplete description is given and the features of the cluster deployment mode in Standalone mode are not given in full. There is no description of the used method of processing the task queue in the cluster.

A score of 5 points is given if the acquirer lists, with some inaccuracies and without explanations, the features and basic principles of each of the deployment modes of the Apache Spark cluster. An incomplete description of the features of cluster deployment in Standalone mode is provided. There is no description of the essence of the task queue processing method in the cluster.

A score of 4 points is given if the acquirer lists with certain inaccuracies and without explanations, some features and some principles of each mode of deployment of the Apache Spark cluster. An incomplete description of the features of cluster deployment in Standalone mode with

certain errors and inaccuracies is provided. There is no description of the task queue processing algorithm in the cluster.

A score of 3 points is given if the acquirer lists with errors and without explanations, some features and implementation principles of some deployment modes of the Apache Spark cluster. An incomplete description of the features of cluster deployment in Standalone mode is provided with some errors and inaccuracies. There is no description of the content of the task queue processing algorithm in the cluster.

A score of 2 points is given if the acquirer provides a list with errors and without explanations, some features, the principles of implementation of Apache Spark cluster deployment modes are not given. Features of cluster deployment in Standalone mode with certain errors and inaccuracies are provided. There is no algorithm for processing the task queue in the cluster.

A score of 1 point is given if the acquirer provides a list with inaccuracies and without explanations, some features, the principles of deployment modes of the Apache Spark cluster are not given. The description of cluster deployment in Standalone mode with significant errors and inaccuracies is provided. There is no content of the task queue processing algorithm in the cluster.

A score of 0 points is given for failure to complete the task in general.

Task 4.

This task is evaluated on an 8-point scale.

An assessment of 8 points is given if the acquirer has given in full and with justified explanations parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

A score of 7 points is given if the acquirer fully, but without substantiation, their appointment is given parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

An assessment of 6 points is given if the applicant does not fully and the appropriate justification is given parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

A score of 5 points is given if the winner of inaccuracies in the explanations include the composition parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

Score 4 points are given if the applicant provides significant inaccuracies and explanations parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

An assessment of 3 points is given if the acquirer has some problems errors and lack of explanations are given parameters and their description of the contents of the Vagrantfile when installing and configuring Apache Spark (Standalone mode).

An assessment of 2 points is given if the applicant with significant errors lacks explanations regarding the composition parameters and their Vagrantfile characteristics when installing and configuring Apache Spark (Standalone mode).

A score of 1 point is given if the applicant incorrectly specified the composition parameters, their characteristics with fundamental errors contents of Vagrantfile when installing and configuring Apache Spark (Standalone mode).

A score of 0 points is given for failure to complete the task in general.

RECOMMENDED LITERATURE

Main

1. Zgurovsky M. Z., Zaychenko Y. P. Big data: conceptual analysis and applications. – Springer International Publishing, 2020. <https://link.springer.com/book/10.1007/978-3-030-14298-8/>.
2. The Big Data-Driven Digital Economy: Artificial and Computational Intelligence. 78-3-030-73057-4 (eBook) <https://doi.org/10.1007/978-3-030-73057-4>.
3. Aroraa G., Lele C., Jindal M. Data Analytics: Principles, Tools, and Practices: A Complete Guide for Advanced Data Analytics Using the Latest Trends, Tools, and Technologies (English Edition). – BPB Publications, 2022.
4. Spark: The Definitive Guide: Big Data Processing Made Simple https://books.google.de/books?hl=ru&lr=&id=oitLDwAAQBAJ&oi=fnd&pg=PP1&dq=Apache+spark+guide&ots=1BtsSveVbd&sig=WkEJjdpEcZp7bKoNoqLgL_5eVAg&redir_esc=y#v=onepage&q=Apache%20spark%20guide&f=false.
5. Beginning Apache Spark 2: With Resilient Distributed Datasets, Spark SQL . structured streaming and Spark machine learning library https://books.google.de/books?hl=ru&lr=&id=wzppDwAAQBAJ&oi=fnd&pg=PR3&dq=Apache+spark+guide&ots=H8sY9Hz7xA&sig=odbg6Dz_D5okP1b_EIMOj51R_20&redir_esc=y#v=onepage&q=Apache%20spark%20guide&f=false.
6. Singh C., Kumar M. Mastering Hadoop 3: Big data processing at scale to unlock unique business insights. – Packt Publishing Ltd, 2019.
7. Mendelevitch O., Stella C., Eadline D. Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale. – Addison-Wesley Professional, 2016.
8. Turkington G., Deshpande T., Karanth S. Hadoop: Data Processing and Modelling. – Packt Publishing Ltd, 2016.
9. Коцовський В. М. Теорія паралельних обчислень: навчальний посібник. / В. М. Коцовський. – Ужгород: ПП «АУТДОР-Шарк», 2021. – 188 с. http://195.230.140.114/jspui/bitstream/123456789/10630/1/Par_rozp_obch_2021.pdf.
10. Інформатика в сфері комунікацій [Електронний ресурс] : навч.-практ. посіб. : у 3-х ч. Ч. 3 : Використання web-технологій у сфері комунікацій / С. Г. Удовенко, В. А. Затхей, О. В. Гороховатський [та ін.] ; за заг. ред. С. Г. Удовенка; Харківський національний економічний університет ім. С. Кузнеця. - Електрон. текстові дан. (10.5 МБ). - Харків : ХНЕУ ім. С. Кузнеця, 2020. - 154 с. : іл. - Загол. з титул. екрану. - Бібліогр.: с. 153. <http://www.repository.hneu.edu.ua/handle/123456789/24506>.

Additional

11. Кислова О. Великі дані в контексті дослідження проблем сучасного суспільства / О. Кислова // Вісник Харківського національного університету імені В. Н. Каразіна, 2019 р. Серія «Соціологічні дослідження сучасного

суспільства: методологія, теорія, методи». 2019. № 42. С. 59–68. URL: <https://periodicals.karazin.ua/ssms/article/view/14869>.

12. Мінухін С., Коптілов Н. Метод збільшення продуктивності Apache Spark на основі сегментування даних і налаштувань конфігураційних параметрів // Сучасний стан наукових досліджень та технологій в промисловості. – 2024. – №. 1 (27). – С. 128–139. <https://doi.org/10.30837/ITSSI.2024.27.128>.

13. Hoger K. Omar, Alaa Khalil Jumaa. Distributed big data analysis using Spark parallel data processing // Bulletin of Electrical Engineering and Informatics. Vol. 11, No. 3, 2022, pp. 1505~1515. DOI:10.11591/eei.v11i3.3187.

14. Сучасні інформаційні технології та системи [Електронний ресурс] : монографія / Н. Г. Аксак, Л. Е. Гризун, С. В. Мінухін [та ін.] ; за заг. ред. Пономаренка В. С. – Харків : ХНЕУ ім. С. Кузнеця, 2022. – 270 с. <http://www.repository.hneu.edu.ua/handle/123456789/29233>.

15. Dong Z. Research of big data information mining and analysis: Technology based on Hadoop technology //2022 International Conference on Big Data, Information and Computer Network (BDICN). IEEE, 2022. – С. 173–176. DOI:10.1109/BDICN55575.2022.00041.

16. Dai H. et al. Research and implementation of big data preprocessing system based on Hadoop //2016 IEEE International Conference on Big Data Analysis (ICBDA). – IEEE, 2016. – С. 1–5. DOI:10.1109/ICBDA.2016.7509802.

17. Bhadani A. K., Jothimani D. Big data: challenges, opportunities, and realities // Effective big data management and opportunities for implementation. – 2016. – С. 1–24. <https://arxiv.org/abs/1705.04928/>.

18. Klemm M., Cownie J. High Performance Parallel Runtimes: Design and Implementation. – Walter de Gruyter GmbH & Co KG, 2021. <https://github.com/parallel-runtimes/lomp>.

19. A Tang S. et al. A survey on spark ecosystem: Big data processing infrastructure, machine learning, and applications //IEEE Transactions on Knowledge and Data Engineering. – 2020. – Т. 34. – №. 1. – С. 71–91.

Information resources

20. Apache Spark™ - Unified Engine for large-scale data analytics.

<https://spark.apache.org/>.

21. Apache Hadoop <https://hadoop.apache.org/>.

22. HDFS Architecture <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html>.

23. MapReduce Tutorial - Apache Hadoop 3.4.0.

<https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>.