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	**обов'язкова**
Мова викладання, навчання та оцінювання	**українська**

Розробник:
д.т.н., професор підписано КЕП Сергій МІНУХІН

Завідувач кафедри
інформаційних систем _____ Дмитро БОНДАРЕНКО

Гарант програми підписано КЕП Сергій МІНУХІН

Харків
2024

ВСТУП

Умови зростання обсягів даних й збільшення залежності якості бізнес-процесів комерційної діяльності підприємств від потоків та інтенсивності даних призводять до необхідності створення розподілених інформаційних систем, які мають забезпечити достатній рівень оперативності оброблення таких даних. Розвиток технологій розподілених та паралельних обчислень, а також наявність можливостей надпродуктивних систем, що розвиваються, та є загальнодоступними для комерційних та науково-дослідних організацій для забезпечення їх діяльності, натепер дозволяють достатньо ефективно обробляти дані великих об'ємів. Значний вплив на цей розвиток мають програмні системи та технології масштабованих обчислювальних систем з можливостями розподіленої обробки надвеликих масивів даних.

Проблема масштабованості комп'ютерних систем при збільшенні об'ємів та інтенсивності даних має розв'язуватися сучасними засобами розподілених середовищ у поєднанні технологіями розподілених та паралельних обчислень, розподілених файлових систем та сховищ даних, які загалом забезпечать ефективну обробку даних.

Навчальна дисципліна "Високопродуктивні системи обробки та аналізу великих даних" вивчається здобувачами спеціальності 122 "Комп'ютерні науки" ОПП "Комп'ютерні науки" усіх форм навчання на першому році навчання протягом першого семестру. Вивчення навчальної дисципліни передбачає набуття теоретичних знань та опанування практичними навичками, пов'язаними з використанням технологій виконання трудомістких завдань на основі надпродуктивних систем оброблення та зберігання даних. Вивчення навчальної дисципліни спрямовано на формування у здобувачів вищої освіти загального розуміння місця та сутності великих даних у сучасній діяльності підприємств та установ, особливостей використання та оброблення великих даних у зрівнянні з класичними стандартами та технологіями розподіленої та паралельної обробки, зокрема технологіями зберігання та оброблення даних у реальному часі або в пакетному режимі, оволодіння архітектурними рішеннями та компонентами сучасних екосистем надпродуктивних обчислень.

Метою викладання навчальної дисципліни "Високопродуктивні системи обробки та аналізу великих даних" є надання здобувачам вищої освіти системи теоретичних знань і придбання практичних навичок розуміння сутності проблем, які виникають при використанні великих даних, сучасних підходів та інструментів їх оброблення та аналізу.

Завданнями навчальної дисципліни є:

- набуття компетентностей щодо роботи з великими даними, їх аналізу задля прийняття ефективних управлінських рішень;

- набуття компетентностей щодо вибору архітектури фрейворку, вибір, встановлення та налаштування програмного забезпечення для роботи у програмних середовищах на рівні розподіленої системи та локальному ресурсі;

- набуття практичних навичок з використання програмних систем обробки великих даних з застосуванням сучасних інтегрованих програмних систем та технологій (фреймворків),

- розгортання та налаштування базового програмного забезпечення (фреймворків) для запуску, виконання завдань та аналізу отриманих результатів з використанням технологій і засобів розподілених систем та сучасних парадигм паралельного програмування.

Об'єктом навчальної дисципліни є процеси оброблення та аналізу великих даних різної природи для підвищення якості управління підприємствами та установами.

Предметом навчальної дисципліни є методи, моделі та технології оброблення, зберігання та аналізу великих даних різної природи.

Результати навчання та компетентності, які формує навчальна дисципліна, подано в табл. 1.

Таблиця 1

Результати навчання та компетентності, які формує навчальна дисципліна

Результати навчання	**Компетентності, якими повинен оволодіти здобувач вищої освіти**
РН1	ЗК05, ЗК07, СК04, СК07
РН2	СК05, СК06, СК08, СК09, СК12
РН4	ЗК05, СК02, СК07, СК08, СК12
РН5	СК09
РН6	ЗК02, ЗК03, СК05, СК08, СК09, СК12
РН7	ЗК05, ЗК06, ЗК07, СК01, СК04, СК05
РН8	ЗК01, ЗК03, ЗК05, ЗК07, СК02, СК04, СК06, СК12
РН9	ЗК01, ЗК02, ЗК03, ЗК05, ЗК07, СК04, СК05, СК07, СК08, СК12
РН10	СК02, СК12
РН11	ЗК01, ЗК03, ЗК05, СК06, СК07, СК08, СК12
РН12	СК10
РН15	СК04
РН16	СК03, СК07, СК08
РН18	ЗК05, СК04
РН19	СК07, СК08, СК12
РН20	ЗК01, ЗК02, ЗК03, ЗК05, ЗК07, СК01, СК02, СК03, СК04, СК05, СК06, СК07, СК11, СК12

де, РН1. Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп’ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп’ютерних наук та на межі галузей знань.

РН2. Мати спеціалізовані уміння/навички розв’язання проблем комп’ютерних наук, необхідні для проведення досліджень та/або провадження інноваційної діяльності з метою розвитку нових знань та процедур.

РН4. Управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів.

РН5. Оцінювати результати діяльності команд та колективів у сфері інформаційних технологій забезпечувати ефективність їх діяльності.

РН6. Розробляти концептуальну модель інформаційної або комп’ютерної системи.

РН7. Розробляти та застосовувати математичні методи для аналізу інформаційних моделей.

РН8. Розробляти математичні моделі та методи аналізу даних (включно з великими).

РН9. Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).

РН10. Проєктувати архітектурні рішення інформаційних та комп’ютерних систем різного призначення.

РН11. Створювати нові алгоритми розв’язування задач у сфері комп’ютерних наук, оцінювати їх ефективність та обмеження на їх застосування.

РН12. Проєктувати та супроводжувати бази даних та знань

РН15. Виявляти потреби потенційних замовників щодо автоматизації обробки інформації.

РН16. Виконувати дослідження у сфері комп’ютерних наук.

РН18. Збирати, формалізувати, систематизувати і аналізувати потреби та вимоги до інформаційної або комп’ютерної системи, що розробляється, експлуатується чи супроводжується.

РН19. Аналізувати сучасний стан і світові тенденції розвитку комп’ютерних наук та інформаційних технологій.

РН20. Розробляти алгоритми та компоненти програмного забезпечення комп’ютерних інформаційних систем для надпродуктивних систем оброблення великих даних (включно з розподіленими та паралельними обчисленнями) та сервісів хмарних платформ.

ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК02. Здатність застосовувати знання у практичних ситуаціях.

ЗК03. Здатність спілкуватися державною мовою як усно, так і письмово.

ЗК05. Здатність вчитися й оволодівати сучасними знаннями.

ЗК06. Здатність бути критичним і самокритичним.

ЗК07. Здатність генерувати нові ідеї (креативність).

СК01. Усвідомлення теоретичних засад комп’ютерних наук.

СК02. Здатність формалізувати предметну область певного проєкту у вигляді відповідної інформаційної моделі.

СК03. Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області.

СК04. Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проєктних рішень.

СК05. Здатність розробляти, описувати, аналізувати та оптимізувати архітектурні рішення інформаційних та комп’ютерних систем різного призначення.

СК06. Здатність застосовувати існуючі і розробляти нові алгоритми розв’язування задач у галузі комп’ютерних наук.

СК07. Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень.

СК08. Здатність розробляти і реалізовувати проєкти зі створення програмного забезпечення, у тому числі в непередбачуваних умовах, за нечітких вимог та необхідності

застосовувати нові стратегічні підходи, використовувати програмні інструменти для організації командної роботи над проєктом.

СК09. Здатність розробляти та адмініструвати бази даних та знань.

СК10. Здатність оцінювати та забезпечувати якість ІТ-проєктів, інформаційних та комп'ютерних систем різного призначення, застосовувати міжнародні стандарти оцінки якості програмного забезпечення інформаційних та комп'ютерних систем, моделі оцінки зрілості процесів розробки інформаційних та комп'ютерних систем.

СК11. Здатність ініціювати, планувати та реалізовувати процеси розробки інформаційних та комп'ютерних систем та програмного забезпечення, включно з його розробкою, аналізом, тестуванням, системною інтеграцією, впровадженням і супроводом.

СК12. Здатність розробляти, застосовувати та інтегрувати технології оброблення та аналізу даних в надпродуктивних системах та хмарних платформах для забезпечення ефективного використання обчислювальних ресурсів комп'ютерних систем.

ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Зміст навчальної дисципліни

Змістовий модуль 1. Основні поняття, сутність та особливості великих даних. Принципи організації побудови систем для роботи з великими даними

Тема 1. Поняття, характеристики великих даних та системи їх оброблення

1.1. Поняття, визначення та характеристики великих даних. Сучасний стан використання великих даних у світовій практиці. Характеристики великих даних 5 V: обсяг, швидкість, різноманітність, достовірність і цінність. Напрями та застосування великих даних у діяльності провідних світових компаній.

1.2. Застосування новітніх інформаційних технологій в інформаційних системах оброблення великих даних.

Тема 2. Сучасні системи оброблення великих даних. Склад компонентів та їх призначення

2.1. Структура фреймворку великих даних: Big Data Strategy, Big Data Architecture, Big Data Algorithms, Big Data Processes, Big Data Functions, Artificial Intelligence.

2.2. Архітектура.

2.2. Компонента Зберігання даних.

2.3. Компонента Обчислення.

2.4. Режими обчислень: in-memory, паралельне програмування, розподілене зберігання (сховище) даних.

2.5. Компонента Аналіз.

2.6. Класифікаційні ознаки великих даних:

доменні області: Здоров'я та медичне обслуговування, Соціальні мережі та Інтернет, Уряд та державний сектор, Управління природними ресурсами, Економічний та бізнес-сектор;

формат даних: структуровані (реляційні бази даних); неструктуровані дані (відео, текст, інформація про географічне розташування тощо); напівструктуровані (JSON, XML);

режими оброблення даних: пакетна обробка, потокова обробка, обробка у реальному часі.

Тема 3. Apache Hadoop: фреймворк для оброблення великих даних. Базові складові для побудови Hadoop: Google's MapReduce, Google File System

3.1. Організація розподіленої обробки інформацій в фреймворку Apache Hadoop. Склад компонентів для розподіленої обробки, зберігання та паралельних обчислень. Призначення MapReduce та розподіленої системи зберігання даних

3.2. Google's MapReduce – базова модель програмування для обробки великих наборів даних у масово паралельний спосіб:

3.3. Google File System (GFS) – базова файлова система для обробки пакетних робочих навантажень даних великих об'ємів задля забезпечення: відмовостійкості, ефективного оброблення великих файлів; оптимізації процесів читання, запису та додавання даних.

Тема 4. Архітектура Apache Hadoop

4.1. Apache HDFS - розподілена файлова система Hadoop: архітектура Master/Slave. Призначення та функції вузлів NameNode (Master node), Secondary NameNode та DataNodes (Slave nodes). Принципи організації взаємодії між майстром та робочими вузлами кластера при обробленні завдань. Блочна організація зберігання даних на робочих вузлах (Slave nodes). Запис та читання даних в систему HDFS.

4.2. Apache Hadoop Map/Reduce.

Архітектура Map/Reduce. Зміст етапів оброблення завдань за парадигмою паралельного програмування Map/Reduce: ітерація по вхідних даних; обчислення пар «ключ/значення» для кожної частини вхідних даних; групування всіх проміжних значень за ключем; ітерація по результуючих групах; редукція кожної групи. Склад та призначення етапів (фаз) реалізації паралельного програмування в Map/Reduce: **фаза map**: splitting, mapping, partition, combined; **фаза reduce**: read, sort, reduce. Склад і формат вхідних та вихідних файлів завдань, що обробляються. Приклад застосування Map/Reduce для тестового набору даних типу Wordcount.

Змістовий модуль 2. Apache Spark: універсальна платформа для обробки та аналітики великих даних

Тема 5. Архітектура Apache Spark.

5.1. Склад та призначення компонентів та інструментів. Системи розгортання Apache Spark.

5.2. Архітектура та інтерфейси програмування.

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

5.4. Склад та призначення компонентів екосистеми Spark: Spark Core; Spark Streaming, Spark SQL, GraphX, MLlib.

Тема 6. Режими розгортання Apache Spark.

6.1. Режими Client, Cluster.

6.2. Режим Local.

6.3. Режим Standalone Scheduler.

6.4. Режим YARN.

6.5. Режим Mesos.

Тема 7. Планування виконання завдань в Apache Spark.

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

7.2. Реалізація DAGScheduler для роботи RDD.

Тема 8. Робота з базами та сховищами даних в SparkSQL. RDD, Dataframe і Dataset.

8.1. RDD: поняття, призначення та принципи використання для структурованих даних і SQL-подібних операцій.

8.2. Поняття та функції Dataframe. Особливості реалізації та використання.

8.3. Dataset: основні функції та призначення для оброблення реляційних баз даних. Вміст поєднання функцій RDD та Dataframe.

Тема 9. Розгортання та налаштування фреймворків Apache Spark та Apache Hadoop в розподіленому та віртуальному середовищах.

9.1. Розгортання та налаштування фреймворків Apache Spark та Apache Hadoop в розподіленому кластері. Вибір та обґрунтування режимів розгортання.

9.2. Розгортання та налаштування фреймворків Apache Spark та Apache Hadoop в віртуальному середовищі (VirtualBox). Вибір та обґрунтування режимів розгортання.

Перелік лабораторних занять за навчальною дисципліною наведено в табл. 2.

Таблиця 2

Перелік лабораторних занять

Назва теми	Зміст
Тема 1-4, 5, 6, 9.Лабораторна робота № 1	Установка та розгортання Apache Spark за допомогою ПЗ VAGRANT
Тема 1-4, 5, 6, 9. Лабораторна робота № 2	Установка кластера Apache Spark в автономному режимі
Тема 1-4, 6, 9. Лабораторна робота № 3	Установка та налаштування Apache Spark YARN кластера

Перелік самостійної роботи за навчальною дисципліною наведено в табл. 3.

Таблиця 3

Перелік самостійної роботи

Назва теми	Зміст
Тема 1 - 9	Вивчення лекційного матеріалу
Тема 1 - 9	Підготовка до лабораторних занять
Тема 1 - 9	Підготовка до екзамену

Кількість годин лекційних, лабораторних занять та годин самостійної роботи наведено в робочому плані (технологічній карті) з навчальної дисципліни.

МЕТОДИ НАВЧАННЯ

У процесі викладання навчальної дисципліни для набуття визначених результатів навчання, активізації освітнього процесу передбачено застосування таких методів навчання, як:

Словесні лекції (Тема 1-9), проблемна лекція (Тема 7, 9), лекція-провокація (Тема 8).

Наочні (демонстрація (Тема 1-9)).

Лабораторна робота (Тема 1 – 4, 5, 6, 9).

ФОРМИ ТА МЕТОДИ ОЦІНЮВАННЯ

Університет використовує 100-бальну накопичувальну систему оцінювання результатів навчання здобувачів вищої освіти.

Поточний контроль здійснюється під час проведення лекційних, лабораторних занять і має на меті перевірку рівня підготовленості здобувача вищої освіти до виконання конкретної роботи і оцінюється сумою набраних балів:

− для дисциплін з формою семестрового контролю екзамен (іспит): максимальна сума – 60 балів;

− мінімальна сума, що дозволяє здобувачу вищої освіти складати екзамен (іспит) – 35 балів.

Підсумковий контроль включає семестровий контроль та атестацію здобувача вищої освіти.

Семестровий контроль проводиться у формі семестрового екзамену (іспиту). Складання семестрового екзамену (іспиту) здійснюється під час екзаменаційної сесії.

Максимальна сума балів, яку може отримати здобувач вищої освіти під час екзамену (іспиту) – 40 балів. Мінімальна сума, за якою екзамен (іспит) вважається складеним – 25 балів.

Підсумкова оцінка за навчальною дисципліною визначається сумуванням балів за поточний та підсумковий контроль.

Під час викладання навчальної дисципліни використовуються наступні контрольні заходи:

Поточний контроль: захист лабораторних робіт (50 балів), письмова

контрольна робота (10 балів).

Семестровий контроль: екзамен (40 балів).

Більш детальну інформацію щодо системи оцінювання наведено в робочому плані (технологічній карті) з навчальної дисципліни.

Приклад екзаменаційного білета та критерії оцінювання для навчальної дисципліни.

Приклад екзаменаційного білета

Харківський національний економічний університет імені Семена Кузнеця
Другий (магістерський) рівень
Спеціальність «Комп'ютерні науки»
Освітньо-професійна програма «Комп'ютерні науки».
Семестр 1
Навчальна дисципліна «Високопродуктивні системи обробки та аналізу великих даних"

ЕКЗАМЕНАЦІЙНИЙ БІЛЕТ № 1

Завдання 1 (діагностичне, 10 балів).

Опишіть та надайте характеристику архітектурі типового кластера Hadoop. Перелічите функції майстра.

Завдання 2 (стереотипне,10 балів).

Дайте характеристику розподіленій файловій системі HDFS в Hadoop. Призначення вузла вторинний NameNode, його відмінності від вузла NameNode. Наведіть обґрунтування його використання в кластері.

Завдання 3 (евристичне, 12 балів). Наведіть основні режими розгортання Apache Spark. Надайте характеристику та особливості режиму Standalone.

Завдання 4 (стереотипне, 8 балів). Наведіть параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Затверджено на засіданні кафедри інформаційних систем протокол № 1 від «27» серпня 2024 р.

Екзаменатор д.т.н., проф. Сергій МІНУХІН

Зав. кафедрою к.т.н., Дмитро БОНДАРЕНКО

Критерії оцінювання

Підсумкові бали за екзамен складаються із суми балів за виконання всіх завдань, що округлені до цілого числа за правилами математики.

Алгоритм вирішення кожного завдання включає окремі етапи, які відрізняються за складністю, трудомісткістю та значенням для розв'язання завдання. Тому окремі завдання та етапи їх розв'язання оцінюються відокремлене один від одного таким чином.

Завдання 1.

Дане завдання оцінюється за 10-бальною шкалою.

Оцінка 10 балів ставиться, якщо здобувачем в повному обсязі наведено склад компонентів архітектури кластера Apache Hadoop відповідно до визначених функцій.

Наведено та деталізовано функції майстра та його місце в архітектурі кластера, основні відмінності від функцій робочого вузла.

Оцінка 9 балів ставиться, якщо здобувачем в повному обсязі наведено склад компонентів архітектури кластера Apache Hadoop відповідно до визначених функцій. Проте у відповіді є певні неточності у визначенні відмінностей функцій майстра та робочих вузлів.

Оцінка 8 балів ставиться, якщо здобувачем не в повному обсязі наведено компоненти архітектури кластера Apache Hadoop відповідно до визначених функцій. Здійснене порівняння функцій вузлів кластера, але є неточності щодо визначення функціональності майстра.

Оцінка 7 балів ставиться, якщо здобувачем не в повному обсязі наведено компоненти архітектури кластера Apache Hadoop відповідно до визначених функцій. Здійснене порівняння функцій вузлів кластера, але є помилки щодо подання характеристик їх функціональності.

Оцінка 6 балів ставиться, якщо здобувачем не в повному обсязі та з помилками наведено склад та функції компонентів кластера. Не в повному обсязі наведено та пояснена функції майстра для управління кластером.

Оцінка 5 балів ставиться, якщо здобувачем не в повному обсязі та з помилками наведено склад та функції компонентів кластера. Не в повному обсязі наведено та обґрунтовано функції майстра або робочих вузлів.

Оцінка 4 бали ставиться, якщо здобувачем не в повному обсязі та з помилками наведено склад та функції компонентів кластера. Не в повному обсязі наведено та обґрунтовано функції майстра та робочих вузлів кластера.

Оцінка 3 бали ставиться, якщо здобувачем з помилками наведено склад компонентів кластера. Наявна значна кількість помилок та неточностей при описі функцій майстра.

Оцінка 2 бали ставиться, якщо здобувачем невірно наведено склад компонентів архітектури кластера. Наявна значна кількість помилок при описі функцій майстра та робочих вузлів.

Оцінка 1 бал ставиться, якщо здобувачем невірно наведено склад компонентів архітектури кластера. Наявний невірний опис функцій майстра.

Оцінка 0 балів ставиться за невиконання завдання загалом.

Завдання 2.

Дане завдання оцінюється за 10-бальною шкалою.

Оцінка 10 балів ставиться, якщо здобувачем в повному обсязі наведено функції розподіленої файлової системі HDFS, призначення вторинного вузла NameNode, його відмінності від вузла NameNode. Наведено та обґрунтовано необхідність його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 9 балів ставиться, якщо здобувачем в повному обсязі наведено функції розподіленої файлової системі HDFS, призначення вторинного вузла NameNode, його відмінності від вузла NameNode. Не в повній мірі обґрунтовано необхідність його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 8 балів ставиться, якщо здобувачем в повному обсязі наведено функції розподіленої файлової системі HDFS, зокрема, призначення вторинного вузла NameNode, але не в повній мірі представлені його відмінності від вузла NameNode. Загалом обґрунтовано необхідність його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 7 балів ставиться, якщо здобувачем не в повному обсязі наведено функції розподіленої файлової системі HDFS, зокрема, щодо призначення вторинного вузла NameNode, не в повній мірі представлені його відмінності від вузла NameNode. Загалом обґрунтовано необхідність його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 6 балів ставиться, якщо здобувачем не в повному обсязі наведено функції розподіленої файлової системі HDFS, зокрема, присутні неточності щодо пояснення призначення вторинного вузла NameNode, не в повній мірі представлені його відмінності від

вузла NameNode. Загалом обґрунтовано необхідність його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 5 балів ставиться, якщо здобувачем не в повному обсязі наведено функції розподіленої файлової системі HDFS, зокрема, присутні помилки щодо пояснень призначення вторинного вузла NameNode, мають місце неточності щодо його визначення відмінності від вузла NameNode. Відсутнє чітке обґрунтування необхідності його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 4 бали ставиться, якщо здобувачем з неточнотями наведено функції розподіленої файлової системі HDFS, присутні помилки щодо призначення вторинного вузла NameNode, і неточності щодо пояснення відмінностей від вузла NameNode. Відсутнє чітке обґрунтування та пояснення необхідності його використання з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 3 бали ставиться, якщо здобувачем з помилками наведено функції розподіленої файлової системі HDFS, присутні помилки щодо призначення вторинного вузла NameNode та помилки щодо визначення та пояснення відмінностей від вузла NameNode. Немає обґрунтування та пояснення необхідності використання вторинного вузла NameNode з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 2 бали ставиться, якщо здобувачем з певними помилками наведено функції розподіленої файлової системі HDFS, є помилки щодо призначення вторинного вузла NameNode, помилки щодо визначення та пояснення відмінностей від вузла NameNode. Немає обґрунтування та пояснення необхідності використання вторинного вузла NameNode з точки зору відмовостійкості та стабільності роботи кластера.

Оцінка 1 бал ставиться, якщо здобувачем не наведено функції розподіленої файлової системі HDFS, призначення вторинного вузла NameNode. Немає обґрунтування необхідності використання вторинного вузла NameNode для роботи кластера.

Оцінка 0 балів ставиться за невиконання завдання загалом.

Завдання 3.

Дане завдання оцінюється за 12-бальною шкалою.

Оцінка 12 балів ставиться, якщо здобувачем в повному обсязі та з поясненнями наведено перелік, особливості та загальні принципи кожного з режимів розгортання кластера Apache Spark. Надана всебічна характеристика та особливості розгортання кластера у режимі Standalone, застосованого методу оброблення черги завдань.

Оцінка 11 балів ставиться, якщо здобувачем в повному обсязі та з поясненнями наведено перелік, особливості та загальні принципи кожного з режимів розгортання кластера Apache Spark. Подана достатня характеристика та деякі особливості режиму розгортання кластера у режимі Standalone, застосованого методу оброблення черги завдань.

Оцінка 10 балів ставиться, якщо здобувачем не в повному обсязі та з відповідними поясненнями наведено перелік, особливості та загальні принципи кожного з режимів розгортання кластера Apache Spark. Подана неповна характеристика та нечіткі особливості режиму розгортання кластера у режимі Standalone, застосованого методу оброблення черги завдань.

Оцінка 9 балів ставиться, якщо здобувачем не в повному обсязі та з відповідними поясненнями наведено перелік, особливості та загальні принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика та не всі особливості режиму розгортання кластера у режимі Standalone, застосованого методу оброблення черги завдань.

Оцінка 8 балів ставиться, якщо здобувачем не в повному обсязі та без обґрунтованих пояснень наведено перелік, особливості та загальні принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика та наведено не в повному обсязі

особливості режиму розгортання кластера у режимі Standalone. Не описано алгоритм використованого методу оброблення черги завдань у кластері.

Оцінка 7 балів ставиться, якщо здобувачем не в повному обсязі та без обґрунтованих пояснень наведено перелік, особливості та основні принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика та наведено не в повному обсязі особливості режиму розгортання кластера у режимі Standalone. Не наведено алгоритм використаного методу оброблення черги завдань у кластері.

Оцінка 6 балів ставиться, якщо здобувачем не в повному обсязі та без обґрунтованих пояснень наведено перелік, особливості та основні принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика та наведено не в повному обсязі особливості режиму розгортання кластера у режимі Standalone. Не наведено опис використаного методу оброблення черги завдань у кластері.

Оцінка 5 балів ставиться, якщо здобувачем з деякими неточностями та без пояснень наведено перелік, особливості та основні принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика особливостей розгортання кластера у режимі Standalone. Немає опису сутності методу оброблення черги завдань у кластері.

Оцінка 4 бали ставиться, якщо здобувачем з певними неточностями та без пояснень наведено перелік, деякі особливості та деякі принципи кожного з режимів розгортання кластера Apache Spark. Надана неповна характеристика особливостей розгортання кластера у режимі Standalone з певними помилками та неточностями. Немає опису алгоритму оброблення черги завдань у кластері.

Оцінка 3 бали ставиться, якщо здобувачем з помилками та без пояснень наведено перелік, деякі особливості та принципи реалізації деяких режимів розгортання кластера Apache Spark. Надана неповна характеристика особливостей розгортання кластера у режимі Standalone з деякими помилками та неточностями. Немає опису вмісту алгоритму оброблення черги завдань у кластері.

Оцінка 2 бали ставиться, якщо здобувачем з помилками та без пояснень наведено перелік, деякі особливості, не наведено принципи реалізації режимів розгортання кластера Apache Spark. Надана характеристика особливостей розгортання кластера у режимі Standalone з певними помилками та неточностями. Немає алгоритму оброблення черги завдань у кластері.

Оцінка 1 бал ставиться, якщо здобувачем з неточностями та без пояснень наведено перелік, деякі особливості, не наведено принципи режимів розгортання кластера Apache Spark. Надана характеристика розгортання кластера у режимі Standalone з суттєвими помилками та неточностями. Немає вмісту алгоритму оброблення черги завдань у кластері.

Оцінка 0 балів ставиться за невиконання завдання загалом.

Завдання 4.

Дане завдання оцінюється за 8-бальною шкалою.

Оцінка 8 балів ставиться, якщо здобувачем в повному обсязі та з обґрунтованими поясненнями наведено параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 7 балів ставиться, якщо здобувачем в повному обсязі але без обґрунтування їх призначення наведено параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 6 балів ставиться, якщо здобувачем не в повному обсязі та відповідного обґрунтування наведено параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 5 балів ставиться, якщо здобувачем з неточностями у поясненнях наведено склад параметрів та їх характеристика вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 4 бали ставиться, якщо здобувачем з значними неточностями та поясненнями наведено параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 3 бали ставиться, якщо здобувачем з деякими помилками та відсутністю пояснень наведено параметри та їх характеристику вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 2 бали ставиться, якщо здобувачем з суттєвими помилками відсутністю пояснень щодо складу параметрів та їх характеристик файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 1 бал ставиться, якщо здобувачем невірно наведений склад параметрів, їх характеристику з принциповими помилками вмісту файлу Vagrantfile при установці та налаштуванні Apache Spark (режим Standalone).

Оцінка 0 балів ставиться за невиконання завдання загалом.

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Основна

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.

Додаткова

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.

Інформаційні ресурси

20. Сайт персональної навчальної системи з навчальної дисципліни «Високопродуктивні системи обробки та аналізу великих даних https://pns.hneu.edu.ua/course/view.php?id=5476