Київський столичний університет імені Бориса Грінченка
Факультет інформаційних технологій та математики
Комп'ютерних наук і математики

«ЗАТВЕРДЖУЮ»
Проректор з науково-педагогічної
та навчальної роботи
_____ О.Б.Жильцов
«___» _____ 2024 р.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

АНАЛІЗ ТА ОБРОБКА ВЕЛИКИХ ДАНИХ

для студентів

спеціальності	122 Комп'ютерні науки
освітнього рівня	другого (магістерського)
освітньої програми	122.00.02 Інформаційно-аналітичні системи

КИЇВСЬКИЙ СТОЛИЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ БОРИСА ГРІНЧЕНКА
Код ЄДРПОУ 45307986
Програма № 3336/24
Начальник відділу моніторингу якості освіти
(підпис) (прізвище, ініціали)
«___» _____ 2024 р.

Київ – 2024

Розробник:
Гладун Анатолій Ясонович, кандидат технічних наук, доцент комп'ютерних наук факультету Інформаційних технологій і математики Київського столичного університету імені Бориса Грінченка

Викладач:
Гладун Анатолій Ясонович, кандидат технічних наук, доцент комп'ютерних наук факультету Інформаційних технологій і математики Київського столичного університету імені Бориса Грінченка

Робочу програму розглянуто і затверджено на засіданні кафедри комп'ютерних наук
Протокол від _7 лютого 2024 р. № 1
Завідувач кафедри _____ Ірина МАШКІНА
(підпис)

Робочу програму погоджено з гарантом освітньої програми (керівником освітньої програми 122.00.02 Інформаційно-аналітичні системи
___.___. 2024 р.
Керівник освітньої програми _____ Олександр БУШМА
(підпис)

Робочу програму перевірено
___.___. 2024 р.
Заступник декана _____ Євген ІВАНІЧЕНКО
(підпис)

Пролонговано:

на 20__/20__ н.р. _____ (_____), «____»____ 20___ р., протокол № ___
(підпис) (ПІБ)

на 20__/20__ н.р. _____ (_____), «____»____ 20___ р., протокол № ___
(підпис) (ПІБ)

на 20__/20__ н.р. _____ (_____), «____»____ 20___ р., протокол № ___
(підпис) (ПІБ)

на 20__/20__ н.р. _____ (_____), «____»____ 20___ р., протокол № ___
(підпис) (ПІБ)р.

1. Опис навчальної дисципліни

Найменування показників	Характеристика дисципліни за формами навчання
	денна
Вид дисципліни	обов'язкова
Мова викладання, навчання та оцінювання	українська
Загальний обсяг кредитів / годин	6 / 180
Курс	1(5)
Семестр	2
Кількість змістових модулів з розподілом:	6
Обсяг кредитів	6
Обсяг годин, в тому числі:	180
Аудиторні	48
Модульний контроль	12
Семестровий контроль	
Самостійна робота	90
Форма семестрового контролю	
екзамен	30

2. Мета та завдання навчальної дисципліни

Мета – формування у магістрів необхідного обсягу теоретичних знань і практичних умінь та навичок, що дозволяють обробляти, аналізувати та видобувати корисну інформацію із швидко змінюваних надвеликих масивів складних неструктурованих даних.

Вивчення дисципліни передбачає ознайомлення з рядом програмних продуктів, які лежать в основі платформи Hadoop, що признана еталоном у середовищі аналітики великих даних.

Вивчення цієї дисципліни підготує випускника до виконання професійних задач: постановка задачі аналізу даних; попереднє оброблення даних; візуалізація даних; розробка, реалізація і застосування методів інтелектуального аналізу даних до надвеликих масивів даних; подання результатів роботи.

Завдання:

- надати студентам теоретичних знань з базовиих основ концепції великих даних, характеристик великих даних, архітектур для обробки та аналізу великих даних;
- надати студентам знання щодо особливостей архітектурних рішень при створенні та розгортанні систем обробки надвеликих масивів даних;
- надати студентам знання щодо вибору релевантної технології зберігання й обробки великих даних, використання сучасних високопродуктивних систем зберігання великих даних;
- сформувати вміння розробки етапів життєвого циклу великих даних та вибору типу аналітики при обробці великих даних, яка відповідає поставленій задачі обробки;
- навчити способам формалізації алгоритмів та ч вибору релевантних моделей при обробці великих даних (поєднання статистичних та кібернетичних методів обробки);
- надати студентам знання щодо характеристик внутрішніх і зовнішніх типів великих даних, застосування метаданих для оцінювання зовнішніх блоків великих даних (достовірність, актуальність, цінність тощо);
- навчити застосовувати кібернетичні методи машинного навчання в алгоритмах аналізу надвеликих даних;
- надати студентам знання щодо розробки нових математичних методів, методів проектування й аналізу алгоритмів, програм для створення систем прийняття рішень;

- сформувати вміння застосовувати методи аналізу великих даних, візуального аналізу даних та сучасних інструментальних засобів аналізу великих даних;

Загальні компетентності

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.
ЗК07 Здатність генерувати нові ідеї (креативність).

Спеціальні фахові компетентності

СК01 Усвідомлення теоретичних засад комп'ютерних наук
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
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. наявних ресурсів та обмежень.

3. Результати навчання за дисципліною

У результаті вивчення навчальної дисципліни студент повинен

знати:

- причини виникнення нового напрямку «великі дані» та проблем і можливостей, пов'язаних з появою великих даних;
- можливості технологій аналізу надвеликих масивів даних для вирішення проблем підприємств, організацій чи бізнесу, а також можливостей застосування наукових методів, у т.ч. методів інтелектуального аналізу даних, до великих даних;
- формалізацію алгоритмів у парадигмі Map Reduce та методів оцінки якості моделей, алгоритмів, методів експериментальної перевірки гіпотез, методів обґрунтування гіпотез;
- існуючі в сучасному світі джерела й типи даних, тенденції великих даних, основні елементи процесу аналізу надвеликих масивів даних, основні підходи до обробки надвеликих масивів даних;
- застосування кібернетичних методів машинного навчання в алгоритмах аналізу надвеликих даних та програмно-інструментальних засобів, орієнтованих на застосування цих методів;
- основні технології і інструменти роботи з великими даними: Hadoop, HDFS, MapReduce, YARN, Storm, Apache Spark;
- основні характеристики кластерної архітектури Hadoop для великих даних і її переваги;
- прикладні застосування великих даних (в електроенергетиці, банківській справі, наукових дослідженнях тощо).

вміти:

- виконувати дослідження процесів створення, нагромадження та обробки інформації, включаючи аналіз і створення моделей даних і знань, мов їхнього опису та маніпулювання,
- виконувати розробку нових математичних методів, розробляти методи проектування й аналізу алгоритмів, програм для створення систем прийняття рішень;
- ставити завдання в області великих даних та виявлення конкретних можливостей використання великих даних для вирішення завдань власної компанії, розуміти ролі всіх учасників процесу роботи з великими даними;
- володіти методами аналізу великих даних, візуального аналізу даних та навичками вибору сучасних інструментальних засобів аналізу великих даних для створення інфраструктури сучасного підприємства;
- застосовувати методи та методики оброблення «сирих» даних великого обсягу і перетворення їх в суттєву інформацію для можливості подальшого видобування нетривіальних прихованих знань.

Знання та розуміння

ПРз-1 наукознавчого понятійного апарату, методології, методів, форм наукових досліджень, вимог та правил наукових публікацій, етичних аспектів наукових досліджень;

Застосування знань та розумінь

ПРу-1 формулювати та вирішувати дослідницьке завдання, збирати, оброблювати та систематизувати інформацію для його вирішення, формулювати висновки, публікувати результати наукових досліджень;

ПРу-2 здатність до формулювання логічних міркувань і висловлень, заснованих на інтерпретації даних, та приймати рішення на підставі неповних даних

ПРу-5 обирати і застосовувати відповідні типові аналітичні, розрахункові та експериментальні методи при розв'язання професійних задач, інтерпретувати результати

4. Структура навчальної дисципліни

Назви змістових модулів і тем	Кількість годин					
	денна форма					
	Усього	у тому числі				
		л.	лр.	м.к.	інд.	с.р.
Змістовий модуль 1. Введення у великі дані						
Тема 1. Введення у великі дані. Концептуальні положення, типи та характеристики великих даних.	10	2	-	-	-	5
Тема 2. Багатовимірний аналіз даних: сутність, методи, специфіка застосування.	14	2	2	2	-	5
Разом за змістовим модулем 1	**24**	**4**	**2**	**2**	**-**	**10**
Змістовий модуль 2. Архітектура систем обробки великих даних						
Тема 3. Огляд архітектур для оброблення надвеликих масивів даних.	14	2	2	-	-	5
Тема 4. Основні технології і інструменти роботи з великими даними.	16	2	2	2	-	5
Разом за змістовим модулем 2	**30**	**4**	**4**	**2**	**-**	**10**
Змістовий модуль 3. Життєвий цикл великих даних						
Тема 5. Технології збору та зберігання великих даних.	14	1	2	-	-	5
Тема 6. Організація оброблення надвеликих масивів даних.	16	1	4	2	-	5

Разом за змістовим модулем 3	**30**	**2**	**6**	**2**	**-**	**10**
Змістовий модуль 4. Методи аналізу великих даних						
Тема 7. Сучасні методи оброблення та аналізу надвеликих масивів даних в Web.	16	2	2	-	-	10
Тема 8. Машинне навчання в аналітичному обробленні надвеликих масивів даних.	20	2	4	2	-	10
Разом за змістовим модулем 4	**36**	**4**	**6**	**2**	**-**	**20**
Змістовий модуль 5. Програмування алгоритмів обробки великих даних						
Тема 9. Мови Phyton і R, стек бібліотек аналізу великих даних. Готові рішення аналізу великих даних Weka та інші.	14	2	2	-	-	10
Тема 10. MapReduce - парадигма розподілених обчислень для великих даних.	16	-	4	2	-	10
Разом за змістовим модулем 5	**30**	**2**	**6**	**2**	**-**	**20**
Змістовий модуль 6. Прикладні аспекти застосування результатів аналізу великих даних						
Тема 11. Технології зберігання великих даних. Великі дані та Інтернет речей (Internet of Things).	14	2	2	-	-	10
Тема 12. Використання результатів оброблення надвеликих даних для Business Intelligence.	16	-	4	2	-	10
Разом за змістовим модулем 6	**30**	**2**	**6**	**2**	**-**	**20**
Усього годин	**180**	**18**	**30**	**12**	**-**	**90**

5. Програма навчальної дисципліни

Змістовий модуль 1. Введення у великі дані.

Тема 1. Введення у великі дані. Концептуальні положення, типи та характеристики великих даних.

Визначення та термінологія великих даних. Великі дані, як сучасний феномен та їх роль в техніці, науці, економіці та суспільному житті. Застосування великих даних і базова модель. Структуровані, неструктуровані та напівструктуровані типи великих даних. Характеристики великих даних: об'єм, швидкість, різноманітність, достовірність, корисність. Метадані. Огляд можливих джерел великих даних (дані соціальних мереж; персональні дані; дані моніторингових систем; сенсорні дані; дані транзакцій; адміністративні дані). Методики збирання великих даних. Роль великих даних для національної економіки.

Тема 2. Багатовимірний аналіз даних: сутність, методи, специфіка застосування.

Використання аналітичної обробки в реальному часі OLAP для надвеликих масивів даних. Концепція аналітичної обробки OLAP (Online Analytical Processing). OLTP (Online Transaction Processing). Типи OLAP. Технічні реалізації OLAP. Технологія OLAP, вимоги і правила для реалізації. Сховища даних. Тест FASMl (Fast of Analysis Shared Multidimensional Information). Операції над багатомірною моделлю даних OLAP. Архітектура OLAP-систем. MOLAP. ROLAP. HOLAP. Недоліки традиційних баз даних для зберігання надвеликих масивів даних. NoSQL-бази даних.

Змістовий модуль 2. Архітектура систем обробки великих даних

Тема 3. Огляд архітектур для оброблення надвеликих масивів даних.

Кластери. Файлові системи FS і розподілені файлові системи DFS. NoSQL. Реплікація масивів даних. Еталонна архітектура для великих даних (HPE Big Data Reference Architecture, BDRA), лямбда-архітектура для аналітики великих даних.

Технології нереляційних СУБД у розподілених сховищах даних. Особливості систем MongoDB, CouchDB та Redis. Особливості розробки інформаційних систем на базі NoSQL-рішень. Особливості та технологічні рішення на прикладі СУБД MongoDB CouchDB та Redis. Порівняння та оцінювання сучасних рішень на базі концепції NoSQL. Технології реплікації. Масштабування NoSQL-рішень на основі сегментування даних. Особливості розгортання та підтримки рішень на базі розподіленої бази даних, яка застосовує нереляційну модель даних.

Тема 4. Основні технології і інструменти роботи з великими даними.

Типи сховищ даних: технологія NoSQL. Технологія MapReduce - особливості розподіленого паралельного оброблення великих масивів даних з використанням великого числа комп'ютерів (кластерів). Переваги та недоліки застосування моделі Map Reduce. Технологія Apache Hadoop для організації розподіленого оброблення великих об'ємів даних: утиліти, бібліотеки та фреймворк для розробки та виконання програм. HDFS - файлова система для зберігання файлів надвеликих розмірів. YARN - модуль, що забезпечує керування ресурсами кластерів та планування завдань. Storm - система для розподіленої обробки великих даних в режимі реального часу, Apache Spark - для розроблення високопродуктивних розподілених систем для вирішення завдань оброблення великих даних і машинного навчання.

Змістовий модуль 3. Життєвий цикл великих даних.

Тема 5. Технології збору та зберігання великих даних.

Огляд технологій зберігання великих даних. Бази даних. Системи керування базами даних. Моделі даних. Підготовка вихідних даних для аналізу: первинна обробка й візуалізація наявних даних.

Процес збору великих даних і проблема якості. Концепція загальної помилки для великих даних. Розширення концепції для великих даних. Труднощі збору великих даних та шляхи їх вирішення. Адміністративні та етичні проблеми: право власності на дані; контроль даних; керування збором даних; конфіденційність і повторна ідентифікація: відсутність однозначного

визначення «виправданих заходів». Технічна проблема: навички, необхідні для інтегрування великих обсягів даних.

Тема 6. Організація оброблення надвеликих масивів даних.

Визначення проблеми (постановка задачі). Збір та підготовка даних. Оцінка даних. Об'єднання й очищення даних. Відбір даних. Перетворення. Побудова моделі. Оцінка й інтерпретація. Зовнішня перевірка. Використання моделі. Спостереження за моделлю. Пониження розмірності даних. Кібернетичні методи аналізу даних (нейронні мережі, карти Кохонена, генетичні алгоритми). Виявлення закономірностей у багатомірному потоку даних за допомогою нейромереж. Приклади використання нейромереж для вирішення прикладних задач аналізу великих даних. Типи задач прийняття рішень, формальні означення дерева рішень і системи прийняття рішень. Семантична задача розпізнавання об'єктів віртуального (цифрового) світу.

<u>Змістовий модуль 4.</u> Методи аналізу великих даних.

Тема 7. Сучасні методи оброблення та аналізу надвеликих масивів даних в Web.

Web Mining - видобування знань з Web. Проблеми аналізу інформації з Web. Етапи Web Mining. Web Mining та інші Інтернет-технології. Категорії Web Mining. Методи добування Web-контенту. Добування Web-контенту в процесі інформаційного пошуку. Добування Web-контенту для формування баз даних. Добування Web-структур. Подання Web-структур. Оцінка важливості Web-структур. Пошук Web-документів з урахуванням гіперпосилань, кластерізація Web-структур. Дослідження використання Web-ресурсів. Дослідницька інформація, етап препроцесінгу, етап добування шаблонів, етап аналізу шаблонів й їхнє застосування. Технологія Opinion Mining та її використання в сучасних інформаційних системах прийняття рішень. Семантичний пошук в Web Mining.

Тема 8. Машинне навчання в аналітичному обробленні надвеликих масивів даних.

Визначення термінів. Роль машинного навчання в обробленні великих даних. Типи задач машинного навчання: навчання з учителем, навчання без учителя та навчання з підкріпленням. Взаємозв'язок машинного навчання з іншими областями аналітики великих даних. Навчання дерев рішень. Навчання асоціативних правил. Штучні нейронні мережі. Глибинне навчання. Індуктивне логічне програмування. Метод опорних векторів. Кластеризація. Баєсові мережі. Навчання з підкріпленням. Навчання представлень. Навчання подібностей та мір. Навчання розріджених словників. Генетичні алгоритми. Прикладні застосування машинного навчання: рекомендаційні системи; інформаційний пошук; маркетинг; Інтернет-реклама; комп'ютерний зір; машинна діагностика; оптимізація та метаевристика; обробка природної мови; мультилінгвістичний переклад; аналіз ринку цінних паперів; біоінформатика; банківська система та страхування. Програмне забезпечення для реалізації алгоритмів машинного навчання для великих даних.

<u>Змістовий модуль 5.</u> Програмування алгоритмів обробки великих даних.

Тема 9. Мови Phyton і R, стек бібліотек аналізу великих даних. Готові рішення аналізу великих даних Weka та інші.

Мови Phyton і R суперники у сфері роботи з даними. Переваги та недоліки цих мов. Типи і структури даних в Phyton. Модулі бібліотеки та пакети програмно-інструментального середовища. Особливості мови програмування для статистики R. Пакети та бібліотеки R. Графічні Редактори скриптів та IDE. Графічні інтерфейси (GUI) для роботи з R. Текстові редактори та середовища розробки (IDE) з частковою підтримкою R. Взаємодія R з іншими мовами програмування. Підтримка R пропієтарними програмними продуктами. Weka (Waikato Environment for Knowledge Analysis) - вільне програмне забезпечення для аналізу даних та машинного навчання. Засоби візуалізації та алгоритми для аналізу даних у Weka. Аналіз великих даних: підготовка даних (preprocessing), відбір ознак (feature selection), кластеризація, класифікація, регресійний аналіз та візуалізація результатів на основі Weka.

Тема 10. MapReduce - парадигма розподілених обчислень для великих даних

Масштабованість. Відмовостійкість. Універсальність MapReduce. Низькорівневий характер MapReduce. Логічне виконання є тісно пов'язаним з фізичним. Конвеєрні схеми для розгляду пакетних обчислень на більш високому рівні. Принципи побудови конвейерних схем обчислень великих даних. Лямбда-архітектура обчислень надвеликих масивів даних. Опис рівнів обробки даних в Лямбда-архітектурі. Візуалізація результатів та ефективність знань отриманих при обробці надвеликих масивів даних.

<u>Змістовий модуль 6.</u> Прикладні аспекти застосування результатів аналізу великих даних.

Тема 11. Технології зберігання великих даних. Великі дані та Інтернет речей (Internet of Things).

Вимоги до зберігання надвеликих масивів даних. Вибір рішення для зберігання даних на рівні пакетної обробки. Застосування сховища пар «ключ-значення».Розподілені файлові системи. Принцип дії розподілених файлових систем. Накопичення надвеликих масиві даних в технології Інтернету речей (Internet of Things). Історія виникнення Інтернету речей. Базові принципи IoT. Архітектура IoT. Зрілість концепції IoT і складових її технологій. Взаємодія IoT з перспективними інфокомунікаційними технологіями. Радіочастотна ідентифікація RFID. Загальні відомості про радіочастотну ідентифікацію RFID. Мітки RFID. Пристрої, що зчитують RFID. Стандартизація технології RFID. Бездротові сенсорні мережі WSN. Основні поняття й принципи сенсорних мереж. Базова архітектура сенсорної мережі. Вузли бездротової сенсорної мережі. Способи передачі даних у БСМ. Протоколи й технології передачі даних у БСМ. Генерація та засоби оброблення надвеликих масивів даних у Інтернеті речей. Web of Things, як надбудова над IoT для забезпечення функцій зберігання та обробки великих масивів даних.

Тема 12. Використання результатів оброблення надвеликих даних для Business Intelligence.

Введення в технологію Business Intelligence. Зв’язки Business Intelligence з методами аналізу великих даних. Класифікація продуктів Business Intelligence. Архітектура Business Intelligence. Використання метаданих в BI.Тенденції розвитку Business Intelligence. Зв`язок Web-сервісів з BI. Business Intelligence 2.0. Основні елементи платформи Business Intelligence. Тенденції розвитку BI. Інтеграція технологій Semantic Web з системами Business Intelligence.

6. Контроль навчальних досягнень

6.1 . Система оцінювання навчальних досягнень студентів

Вид діяльності студента	Максимальна к-сть балів за одиницю	Модуль 1		Модуль 2		Модуль 3		Модуль 4		Модуль 5		Модуль 6	
		кількість одиниць	максимальна кількість балів	кількість одиниць	максимальна кількість балів	кількість одиниць	максимальна кількість балів	кількість одиниць	максимальна кількість балів	кількість одиниць	максимальна кількість балів	кількість одиниць	максимальна кількість балів
Відвідування лекцій	1	2	2	2	2	1	1	2	2	1	1	1	1
Відвідування лабораторних занять	1	1	1	2	2	3	3	3	3	3	3	3	3
Лабораторна робота (в тому числі допуск, виконання, захист)	10	1	10	2	20	3	30	3	30	3	30	3	30
Виконання завдань для самостійної роботи	5	1	5	1	5	1	5	1	5	1	5	1	5
Виконання модульної роботи	25	1	25	1	25	1	25	1	25	1	25	1	25
Разом		-	43	-	54	-	64	-	65		64		64
Максимальна кількість балів: 354													
Розрахунок коефіцієнта: =60/354=0,17													
Екзамен 40 балів													

7.1Завдання для самостійної роботи та критерії її оцінювання

Самостійна робота виконується протягом опрацювання відповідного змістового модуля на лекційних та практичних заняттях і здається на перевірку викладачу у вигляді **авторського** (2-3 сторінки друкованого тексту) реферативного дослідження на вказану в таблиці тему.

Кількість балів за самостійну роботу залежить від дотримання таких вимог:

- своєчасність і самостійність виконання завдань;
- якість виконання завдань (повнота викладення теми, наявність прикладів і джерел, на які спирався студент при опрацюванні теми тощо);
- творчий підхід у виконанні завдань.

№ з/п	Назва теми	Кількість годин	Бали
Змістовий модуль 1.		**16**	**2**
1	Скласти еволюційну схему понять із області великих даних та проаналізувати поняття «аналіз даних» та «аналітика даних» із області великих даних.	8	1
2	Виявити особливості у двох технологіях багатовимірного аналізу даних: OLTP обробка транзакцій у реальному часі та OLAP аналітична обробка у реальному часі. Описати процедуру обробки великих даних ETL: добування перетворення і завантаження.	8	1
Змістовий модуль 2.		**20**	**3**
3	1.Дослідити відомі сучасні архітектури побудови систем для оброблення великих даних. 2. Описати лямбда-архітектуру для великих даних та її особливості.	10	2

№ з/п	Назва теми	Кількість годин	Бали
Змістовий модуль 1.		**16**	**2**
	3.Описати поняття кластерів та інших компонент системи для обробки великих даних.		
4	Сховища та вітрини даних: особливості застосування.	10	1
Змістовий модуль 3.		**20**	**3**
5	Особливості роботи бази даних NoSQL та бази даних NewSQL. Аналіз пристроїв для збереження даних у внутрішній оперативній пам'яті.	10	2
6	Описати кількісний та якісний аналіз великих даних. Дати класифікацію задач з обробки великих даних в технології інтелектуального аналізу даних (Data Mining).	10	1
Змістовий модуль 4.		**24**	**2**
7	1. Аналіз Web–середовища та джерел інформації, призначених для обробки. 2. Виникнення технології Web-Mining і зростання її значення у сучасному світі.	12	1
8	Аналіз алгоритмів «машинного навчання» та зростання його ролі в сучасних системах обробки великих даних.	12	1
Змістовий модуль 5.		**20**	**3**
9	Аналіз режимів рооти аналітичної платформи обробки великих даних Tableau	10	2
10	1. Сформулювати призначення Hadoop для великих даних. 2. Аналіз обробки робочих завдань для великих даних: пакетна обробка та транзакцій на обробка. 3. Аналіз пакетної обробки за допомогою процесора MapReduce. 4. Аналіз задач Map і Reduce.	10	1
	Змістовий модуль 6.	**20**	**2**
11	1.Аналіз сучасних сенсорних мереж та особливостей накопичення, зберігання та передачі інформації. 2.Основін концепції технології Internet of Things. 3. Основін концепції технології Web of Things	10	1
12	Застосування технології Семантичного вебу у Business Intelligence.	10	1
	Усього	**120**	**15**

8.Форми проведення модульного контролю та критерії оцінювання.

Модульний контроль проводиться у формі комп'ютерного тесту. Тести містять 25 питань різного типу, вага кожного питання – 1 бал.

***Форми** проведення семестрового контролю та критерії оцінювання.*

Семестровий контроль проводиться у формі екзамену. Підсумкова оцінка рівня досягнення результатів навчання є сумою всіх оцінок за змістові модулі.

Контроль успішності студентів з урахуванням поточного і підсумкового оцінювання здійснюється відповідно до навчально-методичної карти дисципліни (п. 10), де зазначено види контролю і кількість балів за видами. Систему рейтингових балів для різних видів контролю та порядок їх переведення у національну (4-бальну) та європейську (ECTS) шкалу подано нижче у таблицях.

Шкала відповідності оцінок

Рейтингова оцінка	Оцінка за стобальною шкалою	Значення оцінки
A	**90 – 100** балів	**Відмінно** – відмінний рівень знань (умінь) в межах обов'язкового матеріалу з можливими незначними недоліками
B	**82-89** балів	**Дуже добре** – достатньо високий рівень знань (умінь) в межах обов'язкового матеріалу без суттєвих (грубих) помилок
C	**75-81** балів	**Добре** – в цілому добрий рівень знань (умінь) з незначною кількістю помилок
D	**69-74** балів	**Задовільно –** посередній рівень знань (умінь) із значною кількістю недоліків, достатній для подальшого навчання або професійної діяльності
E	**60-68** балів	**Достатньо** – мінімально можливий допустимий рівень знань (умінь)
FX	**35-59** балів	**Незадовільно з можливістю повторного складання** – незадовільний рівень знань, з можливістю повторного перескладання за умови належного самостійного доопрацювання
F	**1-34** балів	**Незадовільно з обов'язковим повторним вивченням курсу** – досить низький рівень знань (умінь), що вимагає повторного вивчення дисципліни

10. Навчально-методична карта дисципліни

Разом: 180 год., із них: лекції – 18 год., практичні заняття – 30 год., модульний контроль – 12 год., самостійна робота – 90 год.

Модулі (назви, бали)	Змістовий модуль 1 та 2 (69 балів			Змістовий модуль 3 та 4 (75 балів)		Змістовий модуль 5 та 6 (67 балів)	
Лекції (теми, бали)	Введення у великі дані. Концептуальні положення, типи та характеристики великих даних. (1 бал)			Технології збору та зберігання великих даних. (1 бал)	Машинне навчання в аналітичному обробленні надвеликих масивів даних. (1 бал)	MapReduce - парадигма розподілених обчислень для великих даних (1 бал)	Використання результатів оброблення надвеликих даних для Business Intelligence. (1 бал)
Практичні заняття (теми, бали)	Аналітична обробка в реальному часі OLAP для надвеликих масивів даних. (11 балів))	Кореляція. Регресійний аналіз. Завдання в області великих даних, розв'язувані методом регресійного аналізу. (11 балів)	Постановка задачі класифікації. Постановка задачі кластеризації. Задача побудови асоціативних правил. Аналітична платформа Deductor. (11 балів)	Розвязання задачі асоціації та регресії для супермаркету на основі введення первинних даних за допомогою аналітичної платформи Deductor. (11 балів)	Розвязання задачі "Чи видавати кредит клієнту?" на основі карт Коохонена. (11 балів)	Обробка інформації у Web. Web-аналітика. Goooogle-аналітика. Засоби SEO у Яндекс і Google. Аналітично-пошукова робота. (11 балів)	Обробка інформації у Web. Opinion Mining – аналіз текстової інформації для прийняття рішень. (11 балів)
Самостійна робота	Самостійна робота (5 балів)			Самостійна робота (5 балів)		Самостійна робота (5 балів)	
Поточний контроль (вид, бали)	Модульна контрольна робота 1 та 2 (по 25 балів)			Модульна контрольна робота 3 та 4 (по 25 балів)		Модульна контрольна робота 5 та 6 (по 25 балів)	
Підсумковий контроль	Залік (40 балів)						

Теми, що виносяться на семестровий контроль

1. Визначення та термінологія великих даних.
2. Роль великих даних для національної економіки.
3. Великі дані, як сучасний феномен та їх роль в техніці, науці, економіці та суспільному житті.
4. Застосування великих даних і базова модель.
5. Структуровані, неструктуровані та напівструктуровані типи великих даних.
6. Метадані.
7. Характеристики великих даних: об’єм, швидкість, різноманітність, достовірність, корисність.
8. Огляд джерел великих даних.
9. Походження даних
10. Процес здобуття даних
11. Ідентифікація даних
12. Збір і фільтрація даних
13. Витяг даних
14. Перевірка й очищення даних
15. Агрегація та подання даних
16. Візуалізація даних
17. Використання результатів аналізу
18. Поняття аналізу даних для великих даних
19. Поняття аналітики даних для великих даних
20. Поняття багатовимірних даних
21. Аналітична обробка в реальному масштабі часу OLAP для надвеликих масивів даних.
22. Концепція аналітичної обробки OLAP на основі багатовимірного куба.
23. Види архітектур OLAP-систем.
24. Технологія OLTP – обробка транзакцій в реальному масштабі часу.
25. Використання сховища даних для надвеликих масивів даних.
26. Вітрини даних для надвеликих масивів даних.
27. Недоліки традиційних баз даних для зберігання надвеликих масивів даних. NoSQL-бази даних.
28. Життєвий цикл аналітики великих даних
29. Процедура ETL: витяг, перетворення й завантаження великих даних.
30. Поняття кластера в сфері обробки надвеликих масивів даних.
31. Файлові системи FS і розподілені файлові системи DFS для великих даних.
32. Система керування базою даних NoSQL. Реплікація масивів даних.
33. Еталонна архітектура для великих даних (Big Data Reference Architecture)
34. Поняття лямбда-архітектури для аналітики великих даних.
35. Особливості розробки інформаційних систем на базі NoSQL-рішень.
36. Технології реплікації.
37. Особливості роботи розподіленої бази даних, яка застосовує нереляційну модель даних.
38. Шардинг для організації сховищ великих даних.
39. Реплікація для організації обробки великих даних.
40. Режим реплікації "головний-підлеглий
41. Взаємозвязок шардинга і реплікації
42. Теорема CAP (теорема Брюера) для обробки великих даних
43. ACID - набір властивостей для надійної роботи транзакцій бази даних (атомарність, узгодженість, ізольованість, довговічність)
44. BASE - принцип проектування баз даних на основі теореми CAP
45. Основні поняття обробки великих даних
46. Паралельна обробка великих даних

47. Розподілена обробка великих даних
48. Особливості Hadoop - програмної платформи для організації розподіленої обробки великих даних
49. Поняття обробка робочих завдань в системах аналітики великих даних
50. Пакетна обробка великих даних
51. Транзакційна обробка великих даних
52. Обробка великих даних в пакетному режимі
53. Пакетна обробка за допомогою MapReduce
54. Переваги та недоліки застосування моделі Map Reduce.
55. Технологія Apache Hadoop для організації розподіленого оброблення великих об’ємів даних.
56. HDFS - файлова система для зберігання файлів надвеликих розмірів.
57. YARN - модуль, що забезпечує керування ресурсами кластерів та планування завдань.
58. Storm - система для розподіленої обробки великих даних в режимі реального часу,
59. Apache Spark - для розроблення високопродуктивних розподілених систем.
60. Задачі Map і Reduce в обробці великих даних
61. Задача Map
62. Об'єднання даних
63. Розбивка даних
64. Перетасування й сортування даних
65. Задача Reduce
66. Інтерпретація алгоритмів MapReduce
67. Обробка великих даних в режимі реального часу
68. Обсяг, погодженість, швидкість (ОПШ) великих даних
69. Обробка потоку подій
70. Обробка складних подій
71. Огляд технологій зберігання великих даних.
72. Бази даних та моделі даних.
73. Підготовка вихідних даних для аналізу: первинна обробка й візуалізація наявних даних.
74. Процес збору великих даних і проблема якості.
75. Концепція загальної помилки для великих даних.
76. Розширення концепції для великих даних.
77. Труднощі збору великих даних та шляхи їх вирішення.
78. Адміністративні та етичні проблеми: право власності на дані; контроль даних; керування збором даних; конфіденційність і повторна ідентифікація.
79. Web Mining - видобування знань з Web.
80. Проблеми аналізу інформації з Web.
81. Етапи Web Mining.
82. Категорії Web Mining.
83. Методи добування Web-контенту.
84. Добування Web-контенту в процесі інформаційного пошуку.
85. Добування Web-контенту для формування баз даних.
86. Добування Web-структур та подання Web-структур.
87. Оцінка важливості Web-структур.
88. Пошук Web-документів з урахуванням гіперпосилань, кластерізація Web-структур.
89. Дослідження використання Web-ресурсів.
90. Дослідницька інформація, етап препроцесінгу, етап добування шаблонів, етап аналізу шаблонів й їхнє застосування.
91. Технологія Opinion Mining та її використання в сучасних інформаційних системах прийняття рішень.
92. Визначення термінів машинного навчання.
93. Роль машинного навчання в обробленні великих даних.

94. Типи задач машинного навчання: навчання з учителем, навчання без учителя та навчання з підкріпленням.
95. Взаємозв’язок машинного навчання з іншими областями аналітики великих даних.
96. Навчання дерев рішень.
97. Навчання асоціативних правил.
98. Штучні нейронні мережі.
99. Мови Phyton і R у сфері роботи з даними.
100. Вибір рішення для зберігання даних на рівні пакетної обробки.
101. Internet of Things, як джерело накопичення надвеликих масивів даних.
102. Базові принципи IoT. Архітектура IoT.
103. Web of Things, як надбудова над IoT для забезпечення функцій зберігання та обробки великих масивів даних.

14. Рекомендована література

Основна

1. Гладун А.Я., Рогушина Ю.В. Data Mining: пошук знань в даних (Підручник). – К. «Універсаріум». 2016.- 468с.
2. Гладун А.Я., Рогушина Ю.В. Семантичні технології: принципи та практики (монографія) – К. «Універсаріум». 2016.- 365с.
3. Гладун А.Я., Рогушина Ю.В., Осадчий В.В., Прийма С.М. Онтологічний аналіз у Web (монографія).-Мелітополь, МДПУ ім. Б.Хмельницького. 2015.- 287с. ISBN: 978-617-7346-27-1.
4. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-876138.
5. Patil D.J. Building Data Science Teams. O'Reilly. 2011. ISBN: 978-1-449-31623-5 (http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf)
6. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-876138.
7. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data // EMC Education Services. 2015. — 432p. — ISBN: 978-1-118-876138
8. Frontiers in Massive Data Analysis, National Research Council, 2013 - http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis
9. J. Hopcroft, R. Kannan. Foundations of Data Science. 2013. — 412 p. - (https://www.dropbox.com/s/j2s5dn5w5g7ics5/Data%20Science%20Foundations%20book-dec-30-2013.pdf)
10. Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future (англ.). EWeek (15 August 2011).

Допоміжна

1. Shneiderman. The big picture for big data: Visualization. Science, 343:730, February 2014.
2. Keim D. Qu H., Ma K.-L. Big-Data Visualization // IEEE Computer Graphics and Applications. July/August 2013. Pp. 50-51.
3. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations // Proceedings of the IEEE Conference on Visual Languages, September 3-6, 1996. Pp. 336-343.
4. Maletic J.I., Marcus A., Collard M.L. A task oriented view of software visualization // International Workshop on Visualizing Software for Understanding and Analysis. 2002. Pp. 32–40. 338
5. Baker, M. P., Wickens, C. D. Human Factors in Virtual Environments for the Visual Analysis of Scientific Data // Technical Report, NCSA. 1995.
6. *Martin Hilbert.* Big Data for Development: From Information- to Knowledge Societies", - 2013. - SSRN Scholarly Paper No. ID 2205145). Rochester, NY: Social Science Research Network;
7. *Hortonworks.* 7 Key Drivers for the Big Data Market. - 2012
8. Big Data analytics: Future architectures, Skills and roadmaps for the CIO - 2011. - IDC/SAS
9. Big Data: The Next Frontier for Innovation, Competition, and Productivity - 2011. - McKinsey Global Institute.
10. *Hasso Plattner, Alexander Zeier.* In-Memory Data Management: Technology and Applications.
11. *Gaurav Vaish.* Getting Started with NoSQL - 2013. - Packt Publishing. ISBN- 13: 978-1849694988
12. *Jim Webber, Emil Eifrem.* Graph Databases by Ian Robinson. - 2013 - O'Reilly Media. ISBN-13: 978-1449356262
13. *DJPatil.* Building Data Science Teams. O'Reilly. 2011. ISBN: 978-1-44931623-5 (http://cdn.oreilly.com/radar/2011/09/Building-Data-Science-Teams.pdf)

Ресурси Інтернет:

1. Acquia, Examples of Big Data Projects. URL: http://www.acquia.com/ examples-big-data-projects
2. Manyika James, Chui Michael, Brad Brown, Bughin Jacques, Dobbs Richard, Roxburgh Charles, Hung Byers Angela. Report of McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity. URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_fro ntier_for_innovation