ДЕРЖАВНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

СИСТЕМА УПРАВЛІННЯ ЯКІСТЮ
Система забезпечення якості освітньої діяльності та якості вищої освіти
сертифікована на відповідність ДСТУ ISO 9001:2015 / ISO 9001:2015

Кафедра комп'ютерних наук та інформаційних систем

ЗАТВЕРДЖЕНО
вченою радою ФІТ
(пост. п.13 від «28» 05 2024 р.)
Декан

_____ Олександр ХАРЧЕНКО

КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA) /
COMPUTER TECHNOLOGIES OF BIG DATA PROCESSING

РОБОЧА ПРОГРАМА /
COURSE OUTLINE

освітній ступінь	**магістр**	/	bachelor
галузь знань	**12 Інформаційні технології**	/	Information Technology
спеціальність	**122 Комп'ютерні науки**	/	122 Computer Science
освітня програма	**Комп'ютерні науки**	/	Computer Science

Київ 2024

Автори: Томашевська Т.В., кандидат технічних наук, доцент

Робочу програму розглянуто і затверджено на засіданні кафедри комп'ютерних наук та інформаційних систем 14 травня 2024р., протокол № 39.

Рецензенти: Т.О.Філімонова, канд. фіз.-мат. наук, доц., доцент кафедри комп'ютерних наук та інформаційних систем
Н.О.Гордійко, кан. техн. наук, доц., доцент кафедри прикладної фізики Фізико-технічного інституту Національного технічного університету України «КПІ імені Ігоря Сікорського»

КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA) /
COMPUTER TECHNOLOGIES OF BIG DATA PROCESSING

РОБОЧА ПРОГРАМА /
COURSE OUTLINE

освітній ступінь	**магістр**	/	bachelor
галузь знань	**12 Інформаційні технології**	/	Information Technology
спеціальність	**122 Комп'ютерні науки**	/	122 Computer Science
освітня програма	**Комп'ютерні науки**	/	Computer Science

1. СТРУКТУРА ДИСЦИПЛІНИ ТА РОЗПОДІЛ ГОДИН ЗА ТЕМАМИ (ТЕМАТИЧНИЙ ПЛАН)

Назва теми	Кількість годин				
	Усього годин / кредитів	з них			
		лекції	практичні (семінарські) заняття/МК	самостійна робота студентів	Форми контролю
1	2	3	5	6	7
РОЗДІЛ 1. Вступ до аналітики великих даних	38	8	4	26	
Тема 1.1. Глобальні групи даних. Категорії даних	16	4	2	10	О, ППР, ПСР
Тема 1.2. Інструменти по роботі з великими даними	22	4	2	16	О, ППР, ПСР
РОЗДІЛ 2. Робота з PySpark	142	22	26	98	
Тема 2.1. Введення в PySpark та архітектура Apache Spark.	20	2	2	16	О, ППР, ПСР
Тема 2.2. Робота з PySpark.	14	2	4	8	О, ППР, ПСР
Тема 2.3. Абстракції даних в PySpark	18	4	4	10	О, ППР, ПСР
Тема 2.4. Машинне навчання з PySpark MLlib:	16	4	4	8	О, ППР, ПСР
Тема 2.5. Обробка потокових даних за допомогою PySpark Structured Streaming.	18	2	4	12	О, ППР, ПСР
Тема 2.6. Інтеграція PySpark з базами даних.	20	2	4	14	О, ППР, ПСР
Тема 2.7. Обробка графових даних із GraphFrames	14	2	2	10	О, ППР, ПСР
Тема 2.8. Розгортання та масштабування PySpark додатків.	22	4	2	16	О, ППР, ПСР
Разом	180	30	30	124	
Підсумковий контроль – екзамен					

Умовні позначення: ПСР – перевірка самостійної роботи; МК – модульний контроль; ППР – перевірка практичної роботи; О – опитування

2. ТЕМАТИКА ТА ЗМІСТ ЛЕКЦІЙНИХ, ПРАКТИЧНИХ (СЕМІНАРСЬКИХ), ЛАБОРАТОРНИХ ЗАНЯТЬ, САМОСТІЙНОЇ РОБОТИ СТУДЕНТІВ

Результати навчання	Навчальна діяльність*	Робочий час студента, год
1	2	3
РОЗДІЛ 1. ВСТУП ДО АНАЛІТИКИ ВЕЛИКИХ ДАНИХ		
Знати: Теоретичний матеріал, основні поняття з теми. **Вміти**: Аналізувати групи даних, вміти їх розрізняти	**Тема 1.1. Глобальні групи даних. Категорії даних** **Лекція №1. Великі дані. Категорії даних.** *План лекції*: 1. Поняття про глобальні групи даних. Структуризація та класифікація груп даних. 2. Ідентифікація машинних даних, потокових даних, озер даних. 3. Категорії структурованих, неструктурованих та напівструктурованих даних 4. Піраміда Data Science. Процес CRISP-DM. **Список рекомендованих джерел:** Основний: 1[14-42], 2[10-48] Додатковий: 5[7-29] Інтернет-ресурси: 11	16 4
	<u>Самостійна робота студента</u> Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. П'ять "V" великих даних. Об’єм. Швидкість. Різнорідність. Достовірність. Цінність. 2. Еволюція даних. 3. Ієрархія DIKW. Піраміда DIKW. 4. Розуміння бізнес-цілей та початкове вивчення даних. 5. Підготовка даних. Моделювання. 6. Розподіл часу між основними завданнями обробки даних. **Список рекомендованих джерел:** Основний: 1[14-42], 2[10-48] Додатковий: 5[7-29] Інтернет-ресурси: 11	10
	<u>Лабораторне заняття №1</u> *Основні проблеми створення та використання великих даних* 1. *Основні джерела великих даних.* 2. *Відкриті джерела великих даних для дослідників та аналітиків. World Bank Open Data data.worldbank.org. www.imf.org/en/Data Data. xVIEW. MIMIC-III. Berkeley DeepDrive BDD 100k. CREMA-D.* 3. *Організація використання великих даних*	2

Знати: Принципи побудови систем обробки великих даних **Вміти**: Обирати та встановлювати Інструменти роботи з великими даними	**Тема 1.2. Інструменти по роботі з великими даними** **Лекція №1. Архітектура екосистем по роботі з великими даними** *План лекції*: 1. Інструменти для маніпуляції з великими даними 2. Інструменти для візуалізації великих даних 3. Моделі даних. 4. Алгоритм MapReduce. Архітектура та фази використання. **Список рекомендованих джерел:** Основний: 1[с.42-69], 4[с.52-84] Додатковий: 6[с.11-29] Інтернет-ресурси: 11, 12	22 4
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Розподілені сховища. Прозорий доступ до даних. 2. Приклади розподілених та хмарних сховищ даних. Хмарні послуги, які стосуються великих даних. 3. Стратегія вибору інструментів для роботи з великими даними. **Список рекомендованих джерел:** Основний: 1[с.42-69], 4[с.52-84] Додатковий: 6[с.11-29] **,** 10 [1-24] Інтернет-ресурси: 11, 12	16
	Лабораторне заняття №2 *Організація роботи з PySpark* 1. *Ознайомлення з алгоритмами MapReduce* 2. *Найпростіші приклади використання MapReduce* 3. *Робота з алгоритмами MapReduce.*	2
РОЗДІЛ 2. РОБОТА З PYSPARK		
Знати: Принципи та концепції Spark. **Вміти**: Розгортати PySpark, налаштовувати віртуальне середовище	**Тема 2.1. Введення в PySpark та архітектура Apache Spark** **Лекція 1. Apache Spark та фреймворк PySpark** *План лекції:* 1. Архітектура Spark. 2. Архітектура PySpark. 3. Огляд основних модулів та компонентів Spark 4. Приклади використання Spark. **Список рекомендованих джерел:** Основний: 1[с.123-158], 3с.[15-54] Додатковий: 6[с.24-89] , 10 [1-24] Інтернет-ресурси: 13, 14	20 2
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку	16

	інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Проекти, засновані на використанні Spark 2. Scala як основа розробки програм з використанням Spark 3. Порівняння MapReduce і Spark 4. Причини та історія появи фреймворка PySpark **Список рекомендованих джерел:** Основний: 1[с.123-158], 3с.[15-54] Додатковий: 6[с.24-89] , 10 [1-24] Інтернет-ресурси: 13, 14	
	<u>Лабораторне заняття №3</u> *Ознайомлення з фреймворком PySpark* 1. *Встановлення PySpark.* 2. *Налагодження віртуального середовища.* 3. *Налаштування PySpark.*	2
Знати: Основні поняття Spark. **Вміти**: Створювати програми з використанням фреймворку PySpark	**Тема 2.2. Робота з PySpark** **Лекція 1. Робота з PySpark** *План лекції:* 1. Робота з PySpark. 2. Виконання програм PySpark. 3. Об'єкти PySpark 4. Побудови програми з використанням PySpark **5.** Скорочена нотація для перетворень **Список рекомендованих джерел:** Основний: 1[с.123-158], 3[с.15-54] Додатковий: 6[с.24-89] , 10 [1-24] Інтернет-ресурси: 13, 14	14 2
	<u>Самостійна робота студента</u> Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Найпростіші приклади реалізації 2. Інтерактивна оболонка 2. Виконання програм (основні параметри) **Список рекомендованих джерел:** Основний: 1[с.123-158], 3[с.15-54] Додатковий: 6[с.24-89] , 10 [1-24] Інтернет-ресурси: 13, 14	8
	<u>Лабораторне заняття №4</u> *Основні принципи роботи з PySpark.* 1. *Робота з PySpark. Лямбда-вирази* 2. *Робота з функцією ReduceByKey().* 3. *Зчитування та обробка даних.* 4. *Збереження даних*	4

Знати: Принципи роботи з даними в PySpark **Вміти**: Створювати набори даних, проводити їх обробку.	**Тема 2.3. Абстракції даних в PySpark** **Лекція №1. Виведення даних у вікно документа** *План лекції:* 1. RDD та DataFrame. 2. Основні операції, доступні RDD. 3. Робота з DataFrame. 4. Злиття даних з різних джерел. **Список рекомендованих джерел** Основний: 1[с.199-430], 2[с.21-65] Додатковий: 6[с.235-348] , 10 [1-24] Інтернет-ресурси: 13, 14	18 4
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Причини використання RDD 2. Використання різних форматів даних 3. Оптимізація запитів Spark SQL **Список рекомендованих джерел** Основний: 1[с.199-430], 2[с.21-65] Додатковий: 6[с.235-348] , 10 [1-24] Інтернет-ресурси: 13, 14	10
	Лабораторне заняття №4 *Розробка програм з використанням абстракцій PySpark.* 1. *Операції з RDD.* 2. *Parquet файл та створення DataFrame* 3. *Агрегування колонок та стовпців.* 4. *Маніпуляція з даними.*	4
Знати: Організацію машинного навчання в PySpark **Вміти**: Розробляти алгоритми машинного навчання з використанням функцій PySpark	**Тема 2.4. Машинне навчання з PySpark MLlib** **Лекція 1. Процес організації машинного навчання в PySpark** *План лекції* 1. Можливості бібліотеки MLLib. 2. Моделі регресії 3. Ансамблеві архітектури 4. Виявлення патернів **Список рекомендованих джерел** Основний: 1[с.434-463], 3[с.98-156] Додатковий: 7[с.86-148] , 10 [1-24] Інтернет-ресурси: 13, 14	16 4
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку	8

	інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Рекомендаційні системи з PySpark 2. Робота з текстами 3. Загальний конвеєр обробки великих даних з бібліотекою MLLib **Список рекомендованих джерел** Основний: 1[с.434-463], 3[с.98-156] Додатковий: 7[с.86-148] , 10 [1-24] Інтернет-ресурси: 13, 14	
	<u>Лабораторне заняття №6</u> *Розробка програм для реалізації задач машинного навчання* 1. *Логістична та лінійна регресія* 2. *Кластерізація* 3. *Знаходження патернів.*	4
Знати: Потокові дані, принципи роботи з ними **Вміти**: Створювати програми обробки потокових даних	**Тема 2.5. Обробка потокових даних за допомогою PySpark Structured Streaming.** **Лекція 1. Обробка потокових даних за допомогою PySpark Structured Streaming** *План лекції:* 1. Потокові дані. 2. Принципи роботи з потоковими даними 3. Основні функції бібліотеки PySpark Structured Streaming 4. Організація роботи з PySpark Structured Streaming **Список рекомендованих джерел** Основний: 3[с.167-196] Додатковий: 5[с.20-83] , 10 [1-24] Інтернет-ресурси: 13, 14	18 2
	<u>Самостійна робота студента</u> Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Джерела потокових даних 2. Основні алгоритми роботи з потоковими даними. 3. Інструменти роботи з потоковими даними **Список рекомендованих джерел** Основний: 3[с.167-196] Додатковий: 5[с.20-83] , 10 [1-24] Інтернет-ресурси: 11, 12	12
	<u>Лабораторне заняття №7</u> *Розробка програм для роботи з потоковими даними* 1. *Знайомство з бібліотекою PySpark Structured Streaming* 2. *Розробка алгоритму розрахунку вибіркових значень потоку даних*	4

	3. *Створення програм з обробки потокових даних.*	
Знати: Технологію взаємодії PySpark з різними БД. **Вміти:** Організовувати роботу з БД, виконувати обробку великих даних в БД	**Тема 2.6. Інтеграція PySpark з базами даних** **Лекція №1. Організація взаємодії PySpark з БД** *План лекції:* 1. Організація роботи з різними БД в PySpark. 2. Інтеграція з NoSQL базами даних. 3. Взаємодія з NoSQL сховищами даних. 4. Забезпечення безпеки під час роботи із зовнішніми сховищами даних **Список рекомендованих джерел:** Основний: 1[с.434-463], 3[с.98-156] Додатковий: 5[с.126-178] , 10 [1-24] Інтернет-ресурси: 13, 14	20 2
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Бази даних NoSQL (Not Only SQL). Завдання, які вирішує NoSQL. Бази даних NewSQL. 2. Порівняння баз даних SQL (реляційних) та NoSQL (не реляційних). 3. Класифікація баз даних nosql та newsql. 4. Основні типи даних NoSQL. 5. Key-value сховища. 6. Сімейство стовпцевих баз даних. 7. Переваги баз даних NoSQL. Недоліки бази даних NoSQL. **Список рекомендованих джерел:** Основний: 1[с.434-463], 3[с.98-156] Додатковий: 5[с.126-178, 10 [1-24]] Інтернет-ресурси: 13, 14	14
	Лабораторне заняття №8 *Розробка програмного додатку по роботі з БД* 1. *Методи роботи з БД* 2. *Обробка транзакцій при взаємодії з базою даних* 3. *Робота з індексами* 4. *Створення проекту по роботі з БД*	4
Знати: Графові моделі великих даних. **Вміти:** Створити додатки з використанням графових	**Тема 2.7. Обробка графових даних із GraphFrames** **Лекція 1. Обробка графових даних із GraphFrames** *План лекції*: 1. Термінологія графових баз даних. 2. Графові БД, складність та розмір даних. 3. Приклади використання графових баз даних. 4. Використання PySpark при роботі з графами.	14 2

структур	**Список рекомендованих джерел:** Основний: 3[с.390-410], 4[с.172-190] Додатковий: 7[86-148] , 10 [1-24] Інтернет-ресурси: 13, 14	
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Графові СУБД. 2. Приклади графових СУБД. 3. Архітектура графових СУБД. 4. Переваги та недоліки графових СУЬД **Список рекомендованих джерел:** Основний: 3[с.390-410], 4[с.172-190] Додатковий: 7[86-148] , 10 [1-24] Інтернет-ресурси: 13, 14	10
	Лабораторне заняття № 9 *Використання графових структур в PySpark.* 1. *Створення та маніпуляція графових структур.* 2. *Застосування алгоритмів на графах.* 3. *Візуалізація графових структур.* 4. *Розробка програм з використанням графових структур.*	2
Знати: основні поняття організації роботи PySpark додатків. **Вміти:** Здійснювати розгортання додатків PySpark.	**Тема 2.8. Розгортання та масштабування PySpark додатків** **Лекція №1. Розгортання та масштабування PySpark додатків.** *План лекції:* 1. Організація продуктивності PySpark. 2. Профілювання та моніторинг 3. Огляд середовищ розгортання та їх особливостей. 4. Методи горизонтального та вертикального масштабування. **Список рекомендованих джерел:** Основний: 3[с.52-90] Додатковий: 8[с.90-160], 9[с.54-129] , 10 [1-24] Інтернет-ресурси: 12, 13, 14	16 4
	Самостійна робота студента Самостійна робота передбачає вивчення окремих питань дисципліни на основі опрацювання літератури та пошуку інформаційних джерел у середовищі Інтернет. Пропонуються такі питання для самостійного опрацювання: 1. Рефакторинг коду для більш ефективного виконання операцій. 2. Використання поділу даних. 3. Кешування **Список рекомендованих джерел:** Основний: 3[с.52-90]	16

	Додатковий: 8[с.90-160], 9[с.54-129], 10 [1-24] Інтернет-ресурси: 12, 13, 14	
	Лабораторне заняття № 11 *Розгортання додатків PySpark* 1. *Оптимізація продуктивності в PySpark)* 2. *Інструменти для аналізу продуктивності програми. Оптимізація під кластер.* 3. *Робота з налаштуваннями Spark для оптимального використання ресурсів кластера*. 4. *Моніторинг та налагодження у продакшені.* 5. *Інструменти для відстеження та вирішення проблем у робочому середовищі*	2
	Разом	180
Підсумковий контроль – екзамен		

* +20% інтерактиву – зазначені курсивом

3. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

1. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основний:*

1. *Талах М.В. Технології обробки Big Data. Навчальний посібник. Чернівці: Чернівецький нац.ун-т, 2024. 454 с*.
2. Ryza S. Advanced Analytics with PySpark, Patterns for Learning from Data at Scale Using Python and Spark, 2022. 233 p.
3. Rioux J. Data Analysis with Python and PySpark. Manning, 2022. 456 p.
4. Ghavami P. Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics. CreateSpace Independent Publishing Platform, 2016. 204 p

Додатковий

5. Akerkar R. Models of Computation for Big Data Cham: Springer International Publishing, 2018.
6. Chambers B., Zaharia M. The Definitive Guide: Big Data Processing Made Simple. 2018. 603 p.
7. Singh P. Machine Learning with PySpark. 2nd Ed. Apress. 2022. 220p.
8. *Big Data processing methods, models and information technologies: Monograph / edited by Oleg I. Pursky. – Shioda GmbH, Steyr, Austria, 2019. – 234 p.*
9. *Pursky O.I. Identifying customer segments in e-trade using system analysis and clustering methods / O.I.Pursky //Monograph – Agenda Publishing House, Coventry, United Kingdom, 2018. – 140p.*
10. *Томашевська Т.В. Комп'ютерні технології обробки великих даних (Big Data): Методичні рекомендації до лабораторних занять / Т.В. Томашевська. – Київ: Державний торговельно-економічний ун-т, 2024. – 24 с.*

Інтернет-ресурси

11. KDNuggets: Data Mining Community Top Resource for Analytics, Data Mining, and Data Science Software, Companies, Data, Jobs, Education, News, and more. URL: http://www.kdnuggets.com
12. The Data Mine. URL: http://www.the-data-mine.com
13. Уроки PySpark Tutorial for Beginners. URL: https://zecourse.com/course/youtubecomchanneluckw4jcwtegrdhisyiiko4tqabout/uroki-pyspark-tutorial-for-beginners
14. PySpark. User Guide. URL: https://spark.apache.org/docs/latest/api/python/user_guide/index.html

*Курсивом позначені видання, що присутні у бібліотеці ДТЕУ