ДЕРЖАВНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
Факультет інформаційних технологій
Кафедра комп'ютерних наук та інформаційних систем

СИЛАБУС (SYLLABUS)
Дисципліна «Комп'ютерні технології обробки великих даних(Big Data) / Computer technologies of big data processing (Big Data)»

ІНФОРМАЦІЯ ПРО ВИКЛАДАЧА

Викладач	Томашевська Тетяна Володимирівна
Науковий ступінь	Кандидат технічних наук
Вчене звання	Доцент
Посада	Доцент кафедри комп'ютерних наук та інформаційних систем
Адреса кафедри	м.Київ, вул. Кіото 19, каб. Б-507, Б-526
E-mail	compdep@knute.edu.ua
Консультації	Відповідно до графіку індивідуальних консультацій на сайті кафедри

ПОЛІТИКА АКАДЕМІЧНОЇ ДОБРОЧЕСНОСТІ

https://knute.edu.ua/file/MzEyMQ==/c12a9f74e87d9154696ca0f761da2e5c.pdf

Дотримання академічної доброчесності передбачає:

- самостійне виконання навчальних завдань, завдань поточного та підсумкового контролю результатів навчання (для осіб з особливими освітніми потребами ця вимога застосовується з урахуванням їхніх індивідуальних потреб і можливостей);
- посилання на джерела інформації у разі використання не авторських ідей, розробок, тверджень, відомостей і т.п.;
- дотримання норм законодавства про авторське право і суміжні права;
- надання достовірної інформації про результати власної наукової діяльності, використані методики досліджень і джерела інформації.

Порушенням академічної доброчесності вважається:

- академічний плагіат – оприлюднення (частково або повністю) наукових (творчих) результатів, отриманих іншими особами, як результатів власного дослідження (творчості) та/або відтворення опублікованих текстів (оприлюднених творів мистецтва) інших авторів без зазначення авторства;
- самоплагіат – оприлюднення (частково або повністю) власних раніше опублікованих наукових результатів як нових наукових результатів;
- фабрикація – вигадування даних чи фактів, що використовуються в наукових дослідженнях;
- фальсифікація – свідома зміна чи модифікація вже наявних даних, що стосуються наукових досліджень.

За порушення академічної доброчесності здобувачі освіти можуть бути притягнені до академічної відповідальності:

- повторне проходження оцінювання (модульний контроль, іспит, залік тощо);
- повторне проходження відповідного освітнього компонента освітньо-професійної програми;
- відрахування з Університету;
- позбавлення наданих університетом пільг;
- відмова у присудженні відповідного ступеня вищої освіти;

ПОЛІТИКА ЩОДО ВІДВІДУВАННЯ ЗАНЯТЬ

- відвідування занять є обов'язковим;
- за об'єктивних причин (наприклад, хвороба, міжнародне стажування та ін.) навчання може відбуватись в он-лайн формі за погодженням із викладачем дисципліни.

ПОЛОЖЕННЯ ПРО АПЕЛЯЦІЮ РЕЗУЛЬТАТІВ ЕКЗАМЕНІВ У ДТЕУ

https://knute.edu.ua/file/MjkwNQ==/cf2f392763bdbe0447eed3c254854ec5.pdf

ВРЕГУЛЮВАННЯ КОНФЛІКТНИХ СИТУАЦІЙ

Учасники освітнього процесу повинні дотримуватися принципів гідності, взаємоповаги, толерантності, доброчесності. Адміністрація ДТЕУ забезпечує попередження, запобігання, своєчасне виявлення та врегулювання конфліктних ситуацій, пов'язаних із цькуванням, дискримінацією, сексуальними домаганнями (див. Положення про врегулювання конфліктних ситуацій ДТЕУ (https://knute.edu.ua/file/MjkwMjQ=/b91ca19cb0c629d8b9938ba46ccc41f5.pdf).

УЗГОДЖЕННЯ ЗМІСТУ ДИСЦИПЛІНИ З ЦІЛЯМИ СТАЛОГО РОЗВИТКУ (ЦСР)

Дисципліна вивчає, як методи роботи з великими даними співвідноситься з цілями сталого розвитку. Студенти під час вивчення дисципліни досліджують, як застосування підходів до обробки великих даних може сприяти або перешкоджати системній роботі щодо досягнення конкретних цілей сталого розвитку, розбудові сталої, справедливої економіки. Наступні Цілі сталого розвитку мають особливе відношення до тем, що розглядаються навчальною дисципліною:

1 **ЦСР 9. Створення стійкої інфраструктури, сприяння всеохоплюючій і сталій індустріалізації та новаціям.** Упродовж вивчення дисципліни студенти з'ясують, як використання великих даних буде сприяти допомагати підприємствам впроваджувати інноваційні підходи для підвищення продуктивності і створення більш стійких процесів в індустрії, розвитку цифрової інфраструктури.

2 **ЦСР 12. Забезпечення переходу до раціональних моделей споживання і виробництва.** Досліджуючи методи роботи з великими даними студенти усвідомлять, як за їх допомогою розробляти раціональні моделі споживання та здійснювати науково обґрунтований аналіз та прогнозування ефективності використання ресурсів

ОПИС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Назва дисципліни / тип дисципліни	Комп'ютерні технології обробки великих даних (Big Data) / обов'язкова
Навчальний рік	2024-2025
Факультет	Факультет інформаційних технологій
Курс	1
Семестр	1
Освітній ступінь	Магістр
Галузь знань	12 «Інформаційні технології»
Спеціальність	122 «Комп'ютерні науки»
Загальна характеристика	Кількість годин –180 Кількість кредитів – 6 **Види занять:** лекції, лабораторні, самостійна робота. **Співвідношення аудиторних годин і годин самостійної роботи** - 60/120 **Мова викладання** – українська **Форма викладання** – очна
Підсумковий контроль	Екзамен

Програмне забезпечення	Python 2.7, 3.*, Spark 3.*, PySpark, MongoDB
Обладнання	Проектор, комп'ютерна техніка із встановленим програмним забезпеченням та доступом до мережі Інтернет.
Необхідні попередні дисципліни	Дисципліна «Технології розподілених систем та паралельних обчислень»; дисципліна «Машинне навчання»
Методика вивчення	Методика вивчення дисципліни полягає у набутті студентами знань теоретичного і практично-прикладного характеру під час лекцій, лабораторних занять, самостійної роботи та вивчення першоджерел і навчально-методичної літератури.
Мета і завдання	**Метою** вивчення дисципліни «Комп'ютерні технології обробки великих даних (Big Data)» є формування у студентів знань щодо принципів роботи з великими даними, навичок володіння технологіями обробки великих даних та розробки додатків для роботи з великими даними на базі мови програмування Python. Завданням вивчення дисципліни є теоретична та практична підготовка студентів з таких питань: методи зберігання та обробки великих даних з використанням сучасних фреймворків; методи аналізу великих даних та технології використання нереляційних баз даних; технології паралельної та розподіленої обробки великих даних в пакетному та реальному режимах; алгоритми фільтрації, варіації та зберігання великих даних із застосовуванням специфікацій, стандартів, правил і рекомендацій; методи та засоби візуалізації великих даних.
Місце дисципліни в освітньо-професійній програмі	
Загальні компетентності	ЗК 2 Здатність застосовувати знання у практичних ситуаціях ЗК 5 Здатність вчитися й оволодівати сучасними знаннями.
Фахові компетентності (результати навчання)	СК 04 Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проектних рішень. СК 06 Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі комп'ютерних наук. СК 07 Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень. СК 09 Здатність розробляти та адмініструвати бази даних та знань.
Програмні результати навчання	РН8 Розробляти математичні моделі та методи аналізу даних (включно з великим). РН9 Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими). РН12 Проектувати та супроводжувати бази даних та знань.

ТЕМАТИКА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

РОЗДІЛ 1. Вступ до аналітики великих даних

Тема 1.1. Глобальні групи даних. Категорії даних.

Поняття про глобальні групи даних – Shallow data, Deep Data, Micro-data, Nano-data, їх структуризацію та класифікацію. Ідентифікація машинних даних, потокових даних, озер даних. Категорії структурованих, неструктурованих та напівструктурованих даних.

Тема 1.2. Інструменти по роботі з великими даними

Інструменти маніпуляції з великими даними: Hadoop, Apache Spark, Hive, Apache Kafra. Інструменти для візуалізації великих даних. Інструменти для зберігання великих даних. Бази даних для роботи з великими даними. Моделі даних No-SQL. Типи No-SQL БД. Алгоритм MapReduce. Архітектура MapReduce. Фази MapReduce.

РОЗДІЛ 2. Робота з PySpark

Тема 2.1. Введення в PySpark та архітектура Apache Spark.

Екосистема Spark. Ознайомлення з основними концепціями та інструментами PySpark. Модулі та компоненти Spark: Огляд основних модулів, таких як Spark SQL, Spark Streaming та MLlib. Приклади використання. Встановлення та налаштування PySpark. Конфігурація Spark. Інтеграція з Jupyter та IDE. Оптимізація ресурсів. Методи ефективного використання обчислювальних ресурсів. Обробка помилок та моніторинг.

Тема 2.2. Робота з PySpark.

Робота з PySpark. Скорочена нотація перетворень. Виконання програм PySpark. Визначення основної функції-драйвера. Об'єкт SparkSession. Зчитування даних та перетворення їх. Застосування перетворень Spark. Збереження результатів

Тема 2.3. Абстракції даних в PySpark.

Операції, доступні для RDD (Resilient Distributed Datasets). Кешування: оптимізація продуктивності з використанням кешування. Робота з файловою системою. Взаємодія із розподіленим файловим сховищем. Оптимізація та розподіл даних. Методи оптимізації процесу обробки даних. DataFrame та Dataset: Основні операції з даними у форматі DataFrame та Dataset. Об'єднання даних: Злиття та об'єднання даних із різних джерел. Операції з колонками: Маніпуляції з даними на рівні колонок. Обробка пропущених даних. Оптимізація запитів Spark SQL.

Тема 2.4. Машинне навчання з PySpark MLlib

Основні алгоритми машинного навчання: Огляд доступних алгоритмів у MLlib. Навчання та оцінка моделей: Процес навчання моделей та оцінка їх ефективності. Гіперпараметри та крос-валідація. Оптимізація моделей з використанням крос-валідації. Інтеграція з DataFrame. Робота з даними у форматі DataFrame у контексті машинного навчання. Оптимізація продуктивності ML-задач: Методи оптимізації продуктивності під час роботи з великими обсягами даних.

Тема 2.5. Обробка потокових даних за допомогою PySpark Structured Streaming

Введення у потокову обробку: Основи роботи з потоковими даними. Структурований стрімінг. Використання Structured Streaming для обробки даних. Вікна часу та агрегації: Робота з вікнами часу для агрегації даних. Стан та надійність: Забезпечення надійності обробки потокових даних. Інтеграція із зовнішніми джерелами: Взаємодія з різними джерелами потокових даних.

Тема 2.6. Інтеграція PySpark з базами даних

Методи взаємодії з різними базами даних. Оптимізація SQL-запитів. Робота з індексами. Обробка транзакцій при взаємодії з базою даних. Інтеграція з NoSQL базами даних: Взаємодія з NoSQL сховищами даних. Забезпечення безпеки під час роботи із зовнішніми сховищами даних.

Тема 2.7. Обробка графових даних із GraphFrames

Введення у графові структури даних: Розуміння базових понять графів. Створення та маніпуляції графами. Алгоритми обробки графів. Застосування різних алгоритмів до графових структур. Візуалізація графів. Інструменти та методи для візуалізації графових даних.

Тема 2.8. Розгортання та масштабування PySpark додатків

Оптимізація продуктивності в PySpark. Кешування та Broadcast. Використання кешування та передачі даних на рівні вузлів. Оптимізація коду на рівні операцій: Рефакторинг коду для більш ефективного виконання операцій. Використання поділу даних. Профілювання та моніторинг. Інструменти для аналізу продуктивності програми. Оптимізація під кластер. Робота з налаштуваннями Spark для оптимального використання ресурсів кластера. Огляд середовищ розгортання та їх особливостей. Масштабування додатків. Методи горизонтального та вертикального масштабування. Налаштування параметрів для великих обсягів даних. Ефективне використання обчислювальних ресурсів. Моніторинг та налагодження у продакшені. Інструменти для відстеження та вирішення проблем у робочому середовищі.

Перелік навчальних робіт з дисципліни «Комп'ютерні технології обробки даних (Big Data)»

Види робіт	**К-сть балів**
Лабораторне заняття №1. Тема: "Основні проблеми створення та використання великих даних"	8
Лабораторне заняття №2. Тема: "Знайомство з PySpark"	8
Лабораторне заняття №3. Тема: "Розробка програм з використанням абстракцій PySpark"	8
Лабораторне заняття №4. Тема: "Розробка програм для реалізації задач машинного навчання"	8
Лабораторне заняття №5. Тема: "Розробка програм для роботи з потоковими даними"	8
Лабораторне заняття №6. Тема: "Розробка програмного додатку по роботі з БД"	8
Лабораторне заняття №7. Тема: "Використання графових структур в PySpark"	8
Лабораторне заняття №8. Тема: "Розгортання додатків PySpark"	8
Модульний контроль	16
Разом: Аудиторна робота	**80**
Самостійна робота (СР)	**20**
Всього:	**100**

КОНТРОЛЬ ТА КРИТЕРІЇ ОЦІНЮВАННЯ ЗНАНЬ СТУДЕНТІВ

При вивченні дисципліни використовуються наступні форми контролю знань студентів: поточний; модульний; підсумковий.

Поточний контроль передбачає перевірку теоретичних питань, самостійної роботи, практичних робіт та усне опитування по кожній практичній роботі. По даному виду контролю оцінювання знань здійснюється у відповідності до бального розподілу наведеного в попередній таблиці.

Модульний контроль передбачає виконання модульної контрольної роботи. Всі завдання оцінюються в 16 балів. Перше завдання (теоретичне) – 6 балів, друге завдання (практичне) – 5 балів, третє завдання (практичне) – 5 балів.

Формою підсумкового контролю є екзамен. Екзаменаційна оцінка (100 балів) є результатом виконання одного теоретичного питання (40 балів) та двох практичних завдань (2х30=60 балів).

Результуюча оцінка з дисципліни визначається як середня від балів набраних протягом семестру та отриманих на іспиті.

СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

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.