



СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

МЕТОДИ І ЗАСОБИ ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ

ID 6537

Шифр, назва спеціальності та освітній рівень	122 Комп'ютерні науки (бакалавр)	Назва освітньої програми	Комп'ютерні науки (2024) Computer science (2024)
Тип програми	Освітньо-професійна	Мова викладання	Українська
Факультет	Факультет комп'ютерно-інформаційних систем і програмної інженерії (ФІС)	Кафедра	Каф. комп'ютерних наук (КН)

Викладач/викладачі

Готович Володимир Анатолійович, канд. техн. наук, доцент, доцент кафедри комп'ютерних наук, [профіль на порталі "Науковці ТНТУ"](#)

Загальна інформація про дисципліну

Мета курсу	Мета дисципліни «Методи і засоби опрацювання великих даних» полягає у формуванні компетентностей щодо методів та засобів опрацювання великих даних, згідно освітньої програми, шляхом вивчення теоретичних основ, методів, алгоритмів і архітектур, та набуття навичок практичного застосування популярних на сьогоднішній день інструментальних засобів для опрацювання великих масивів даних.
Формат курсу	Змішаний – курс, що передбачає проведення лекцій, практичних робіт, лабораторних та консультацій для кращого розуміння викладеного матеріалу і має супровід в електронному навчальному курсі системи A-Tutor, має структуру, контент, завдання і систему оцінювання.
Компетентності ОП	ЗК1. Здатність до абстрактного мислення, аналізу та синтезу. ЗК2. Здатність застосовувати знання у практичних ситуаціях. ЗК7. Здатність до пошуку, оброблення та аналізу інформації з різних джерел. СК9. Здатність реалізувати багаторівневу обчислювальну модель на основі архітектури клієнт-сервер, включаючи бази даних, знань і сховища даних, виконувати розподілену обробку великих наборів даних на кластерах стандартних серверів для забезпечення обчислювальних потреб користувачів, у тому числі на хмарних сервісах. СК11. Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв’язування прикладних задач. СК17. Здатність проектувати, розробляти та/чи супроводжувати програмно-алгоритмічні засоби для отримання, зберігання, обробки, аналізу та візуалізації різнотипових наборів та колекцій великих даних; впроваджувати рішення, які можуть ефективно масштабуватися при зростанні обсягів даних.
Програмні результати навчання з ОП	ПР17. Використовувати знання архітектури великих даних, структурних моделей для забезпечення ефективності та швидкодії обробки та візуалізації великих за обсягом даних. ПР18. Застосовувати знання інформаційних технологій обробки, аналізу та візуалізації великих даних на практиці, розробляти стратегії для підтримки та покращення якості великих даних протягом їхнього життєвого циклу. ПР19. Використовувати спеціалізовані інструменти та технології для обробки та візуалізації великих обсягів даних, такі як Apache Hadoop, Spark, або інші фреймворки для роботи з Big Data.
Обсяг курсу	Очна (денна) форма здобуття освіти: Кількість кредитів ECTS — 4; лекції — 28 год.; лабораторні заняття — 28 год.; самостійна робота — 64 год.; Заочна форма здобуття освіти:

	Кількість кредитів ECTS — 4; лекції — 6 год.; лабораторні заняття — 6 год.; самостійна робота — 108 год.;
Ознаки курсу	Рік навчання — 4; семестр — 8; Обов’язкова (для здобувачів інших ОП може бути вибірковою) дисципліна; кількість модулів — 2;
Форма контролю	Поточний контроль: захист лабораторних робіт та модульне тестування Підсумковий контроль: екзамен
Компетентності та дисципліни, що є передумовою для вивчення	Сховища великих даних, інтелектуальний аналіз даних.
Матеріально-технічне та/або інформаційне забезпечення	Комп’ютер під управлінням ОС Windows версії не нижче 10, та з об’ємом оперативної пам’яті 16 Гбайт, прикладне ПЗ Oracle VirtualBox. Середовище Google Colab (доступне з допомогою веб-браузера).

СТРУКТУРА КУРСУ

Лекційний курс	Годин	
	ОФЗО	ЗФЗО
Тема 1. Поняття та характеристики великих даних. Поняття великих даних. Термінологія. Характеристики великих даних. Різні типи даних.	2	0,5
Тема 2. Перехід до великих даних та питання планування. Проблеми та перспективи переходу бізнесу до технологій великих даних. Організаційні передумови. Збір (отримання) даних. Конфіденційність. Безпека великих даних. Джерела походження даних. Обмеження підтримки в реальному часі. Особливі проблеми продуктивності. Особливі вимоги до керівництва. Методологія управління потоками даних та програмними рішеннями для великих даних. Використання хмарних технологій.	2	0,5
Тема 3. Apache Hadoop. Загальний огляд. Реплікація даних. Сфери застосування Hadoop. Переваги, які надає Hadoop. Виклики та труднощі при роботі з Hadoop.	1	0,5
Тема 4. Життєвий цикл великих даних. Традиційний життєвий цикл Data Mining. Методологія CRISP-DM. Методологія SEMMA. Життєвий цикл аналітики великих даних. Етапи життєвого циклу. Етап 1. Оцінка бізнес-ситуації. Етап 2. Ідентифікація даних. Етап 3. Збір та фільтрація даних. Етап 4. Видобування даних. Етап 5. Перевірка та очищення даних. Етап 6. Агрегування та подання (представлення) даних. Етап 7. Аналіз даних. Етап 8. Візуалізація даних. Етап 9. Використання результатів аналізу.	4	0,5
Тема 5. Модель розподілених обчислень MapReduce. Пакетна обробка даних за допомогою MapReduce. Завдання Map та Reduce. Простий приклад MapReduce. Інтерпретація алгоритмів MapReduce.	1	0,5
Тема 6. Корпоративні технології та Business Intelligence для великих даних. Загальна характеристика корпоративних технологій. Обробка транзакцій у реальному часі (OLTP). Аналітична обробка даних в реальному часі (OLAP). Вилучення, перетворення та завантаження (ETL) даних. Сховища даних. Вітрини даних. Традиційний BI. Використання хмарних технологій. BI для великих даних. Візуалізація даних. Традиційна візуалізація. Візуалізація даних при обробці великих даних.	3	0,5
Тема 7. Apache Hive. Загальна характеристика. Архітектура та функціонування Apache Hive. Компоненти Apache Hive. Функціонування Apache Hive. Обробка запитів. Відмінності між Apache Hive і традиційними СУБД. Переваги та недоліки застосування Apache Hive.	1	0,5

Тема 8. Принципи зберігання великих даних. Кластери. Файлові системи та розподілені файлові системи. БД типу NoSQL. Шардинг. Реплікація. Поеднання шардингу та реплікації. Теорема CAP. Принципи ACID. Принципи BASE.	4	0,5
Тема 9. Apache Pig. Загальна характеристика. Архітектура Apache Pig. Модель даних Pig Latin.	1	
Тема 10. Принципи обробки великих даних. Паралельна обробка даних. Розподілена обробка даних. Hadoop. Обробка робочих завдань. Обробка із застосування кластера. Обробка в пакетному режимі. Обробка в режимі реального часу. Принцип SCV.	2	0,5
Тема 11. Apache Spark. Загальна характеристика. Архітектура Apache Spark та принципи її функціонування. Архітектура. Режими виконання. Resilient Distributed Datasets. Directed Acyclic Graph (DAG). DataFrames та Datasets. Spark Core. Spark APIs. Переваги Apache Spark.	1	0,5
Тема 12. Засоби і технології зберігання великих даних. Загальна характеристика засобів і технологій зберігання великих даних. Дискові системи зберігання. Розподілені файлові системи. Системи управління реляційними базами даних. Бази даних NoSQL. Системи зберігання даних в оперативній пам'яті.	2	0,5
Тема 13. Основні методи аналізу великих даних. Загальна характеристика методів аналізу великих даних. Кількісний аналіз. Якісний аналіз. Data Mining. Статистичний аналіз. Машинне навчання. Семантичний аналіз. Візуальний аналіз. Метод вибору категорії засобів аналітичного опрацювання «Великих даних. Метод вибору платформи аналітичного опрацювання міських колекцій даних. Інформаційна технологія «Великі дані» в системах «розумних міст». Інформаційна технологія «Аналітика великих даних» в системах «розумних міст».	4	0,5
РАЗОМ:	28	6
Лабораторний практикум (теми)	Годин ОФЗО ОБОВ'ЯЗКОВО	ЗФЗО ЗА ДОПОМОГОЮ
Тема 1. Ознайомлення з робочим середовищем та робота з файловою системою кластера.	4	1
Тема 2. Виконання простого завдання MapReduce.	4	1
Тема 3. Робота з Apache Hive.	6	1
Тема 4. Робота з Apache Pig.	4	1

Тема 5. Робота з Apache Spark.

6

1

Тема 6. Exploratory Data Analysis з використанням засобів мови Python.

4

1

РАЗОМ:

28

6

Курсова робота/проект

Мета виконання курсового проекту	Метою виконання курсового проекту з дисципліни «Методи і засоби опрацювання великих даних» є систематизація, закріплення та поглиблення теоретичних знань, їхнє застосування для вирішення конкретного практичного завдання відповідно до вимог формування компетентностей згідно освітньої програми.
Завдання курсового проекту	1) аналіз проблеми та/чи методів і алгоритмів для її вирішення; 2) критичний аналіз існуючих методів і засобів розв’язання аналогічної проблеми; 3) обґрунтування вибору методів та/чи засобів розв’язання проблеми; 4) виконання поставленого індивідуального завдання та оформлення звітної документації.
Структура курсового проекту	Титульний лист; завдання на курсовий проект; анотація; зміст; перелік умовних позначень; вступ; основна частина; висновки; список використаних джерел; додатки.
Обсяг курсового проекту	Рекомендований обсяг - 30-35 сторінок (не включаючи додатків).
Етапи виконання	Вибір та затвердження теми курсового проекту; критичний аналіз нормативно-правової бази, спеціальної літератури з проблем, що розглядаються, пошук додаткових джерел інформації; складання плану курсового проекту; узагальнення та аналіз накопиченого матеріалу, обробка даних, обґрунтування пропозицій; написання тексту і оформлення курсового проекту; захист курсового проекту згідно з встановленим графіком.
Оцінювання курсового проекту	Зміст курсового проекту – 75 балів, захист курсового проекту – 25 балів.
Форма контролю	Захист курсового проекту передбачає: - стислу доповідь (5 хв.), в якій необхідно відокремити мету, об’єкт, предмет дослідження та коротко висвітлити зміст одержаних результатів дослідження. Зробити акцент на висновках та рекомендаціях. Бажано, щоб доповідь під час захисту супроводжувалась презентацією результатів; - співбесіду і відповіді на запитання керівника та членів комісії. Курсовий проект та її захист оцінюється відповідно до вимог кредитно-модульної системи.
Технічне й програмне забезпечення	Технічні засоби для демонстрування результатів виконання курсового проекту (ноутбук, проектор).

ІНШІ ВИДИ РОБІТ

Теми, короткий зміст

Самостійна робота студента включає опрацювання теоретичного матеріалу за допомогою рекомендованих літературних та Інтернет джерел (а також шляхом самостійного пошуку інформації) по наступних темах:

Тема 1. Поняття та характеристики великих даних. Дата аналітичне мислення. Покоління архітектур великих даних.

Тема 2. Перехід до великих даних та питання планування. Задачі бізнесу і рішення Data Science.

Тема 3. Apache Hadoop. Консольні команди hdfs.

Тема 4. Розподілені файлові системи HDFS, S3 та GCS.

Тема 5. ETL- та ELT-процеси. Міграція та інтеграція даних. Формати даних Parquet та Avro.

Тема 6. Apache Hive. Мова HiveQL.

Тема 7. Принципи зберігання великих даних. Сучасні розподілені сховища даних та озера даних.

Тема 8. Мова Apache Pig Latin.

Тема 9. Принципи обробки великих даних. Пакетна та потокова обробка даних.

Тема 10. Apache Spark. Способи організації даних RDD, Data Frame та Dataset.

Тема 11. Apache Airflow.

Тема 12. Сучасні хмарні рішення для великих даних. Найпопулярніші сервіси на базі AWS, Google Cloud та Microsoft Azure.

Тема 13. Найпопулярніші бібліотеки мови Python для обробки великих даних. Dask, Pandas, NumPy, SQLAlchemy.

Тема 14. Основні методи аналізу великих даних. Рішення на базі машинного навчання та технологій штучного інтелекту.

Крім того, до самостійної роботи також належать підготовка до захисту лабораторних робіт і підготовка до змістових модулів та екзамену.

Інформаційні джерела для вивчення курсу

Рекомендована література

Основна:

1. Thomas Erl et al. Big Data Fundamentals : Concepts, Drivers & Techniques // Thomas Erl, Wajid Khattak, Paul Buhler. – Pearson, 2016. – 240 p. ISBN: 0134291077
2. Foster Provost, Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. – O'Reilly Media, 2013. – 413 p. ISBN: 9781449361327
3. Davy Cielen, Arno Meysman, Mohamed Ali. Introducing Data Science: Big Data, Machine Learning, and more, using Python tools. – Manning Publications Co, 2016. – 300 p. ISBN: 9781633430037

Допоміжна:

1. Vince Reynolds. Big Data For Beginners. – CreateSpace Independent Publishing Platform, 2016. – 150 p. ISBN: 1530412048
2. Nathan Marz, James Warren. Big Data. Principles and best practices of scalable realtime data systems. – Manning Publications Co, 2015. – 308 p. ISBN: 9781617290343
3. Brian Steele, John Chandler, Swarna Reddy. Algorithms for Data Science. – Springer, 2016. – 430 p. ISBN: 978-3-319-45795-6
4. Дуда О.М. Інформаційні технології супроводу процесів в міських ресурсних та соціокомунікаційних мережах : дис. ... канд. техн. наук : 05.13.06 / О.М. Дуда - Тернопіль, 2019. - 233 с. <https://elartu.tntu.edu.ua/handle/lib/31118>
5. Duda, O., Kunanets, N., Matsiuk, O., Pasichnyk, V., Rzhеuskyi, A. (2021). Aggregation, Storing, Multidimensional Representation and Processing of COVID-19 Data. In: Shakhovska, N., Medykovsky, M.O. (eds) Advances in Intelligent Systems and Computing V. CSIT 2020. Advances in Intelligent Systems and Computing, vol 1293. Springer, Cham. https://doi.org/10.1007/978-3-030-63270-0_60. ISSN: 2194-5357. EISSN: 2194-5365.
6. O. Duda, V. Pasichnyk, N. Kunanets, R. Antonii and O. Matsiuk, Multidimensional Representation of COVID-19 Data Using OLAP Information Technology, 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, 2020, pp. 277-280, doi: 10.1109/CSIT49958.2020.9321889. ISSN 27663655.
7. Duda, O., Matsiuk, O., Kunanets, N., Pasichnyk, V., Rzhеuskyi, A., & Bilak, Y. (2021). Formation of hypercubes based on data obtained from systems of IoT devices of urban resource networks. International Journal of Sensors, Wireless Communications and Control, 11(5), 498-504. DOI 10.2174/2210327910999201210145151. ISSN 22103279.
8. Duda, O., Kunanets, N., Martsenko, S., Nykytyuk, V., & Pasichnyk, V. (2021). Information technology platform for the selection and analytical processing of information on COVID-19. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 231-238. <https://doi.org/10.1109/CSIT52700.2021.9648839>. ISSN 27663655.
9. Duda, O., Kunanets, N., Martsenko, S., Nykytyuk, V., & Pasichnyk, V. (2021). COVID-19 data collections and analytical processing. 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), 252-257. doi: 10.1109/CSIT52700.2021.9648658. ISSN: 2766-3639. EISSN: 2766-3655.
10. Duda, O. O., Kunanets, N. O., Matsiuk, O. V., Pasichnyk, V. V., Veretennikova, N. M., Fedonuyk, A. M., & Yunchyk, V. M. (2020). Selection of effective methods of big data analytical processing in information systems of smart cities. CEUR Workshop Proceedings, 2631, 68-78. (2nd International Workshop on Modern Machine Learning Technologies and Data Science, MoMLeT+DS 2020, Lviv-Shatsk, 2020, Code 161592). ISSN 1613-0073.
11. Bodnarchuk, I., Duda, O., Kharchenko, A., Kunanets, N., Matsiuk, O., & Pasichnyk, V. (2020). Choice method of analytical platform for smart cities. CEUR Workshop Proceedings, 2732, 116-127. (16th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops, ICTERI 2020, Kharkiv, 2020, Code 164721). ISSN 1613-0073.

Інформаційні ресурси:

1. Електронний навчальний курс “Методи і засоби опрацювання великих даних” в системі дистанційного навчання ТНТУ імені І. Пулюя. [Електронний ресурс]. – Режим доступу: <https://dl.tntu.edu.ua/bounce.php?course=6537>.
2. Hadoop Tutorial for Beginners. [Електронний ресурс]. – Режим доступу: https://www.youtube.com/watch?v=JK2MdJAWEGc&list=PLlgLmuG_KgbasW0lpInSAIxYd2vqAEPit&index=1
3. Glossary. [Електронний ресурс]. – Режим доступу: <https://www.databricks.com/glossary>
4. Cloud Computing Concepts Hub. [Електронний ресурс]. – Режим доступу: <https://aws.amazon.com/what-is>
5. MapReduce Tutorial For Beginners | Simplilearn. [Електронний ресурс]. – Режим доступу: <https://www.youtube.com/watch?v=rll6EnW95R8>
6. Big Data Analytics Tutorial. [Електронний ресурс]. – Режим доступу: https://www.tutorialspoint.com/big_data_analytics/index.htm
7. Hadoop - Quick Guide. [Електронний ресурс]. – Режим доступу: https://www.tutorialspoint.com/hadoop/hadoop_quick_guide.htm
8. The Ultimate Hands-On Hadoop: Tame your Big Data! [Електронний ресурс]. – Режим доступу: <https://ua.udemy.com/course/the-ultimate-hands-on-hadoop-tame-your-big-data>

Політики курсу

Політика контролю	Використовуються такі засоби оцінювання та методи демонстрування результатів навчання: поточне опитування; тестування; виконання індивідуальних завдань та презентацій; оцінювання результатів виконаних самостійних робіт; бесіди та обговорення проблемних питань; дискусії; індивідуальні консультації; екзамен. Можливий ректорський контроль.
Політика щодо консультування	Консультації при вивченні дисципліни проводяться згідно затвердженого на кафедрі . Консультування передбачено як очно ,так і з використанням ресурсів електронного навчального курсу у середовищі електронного навчання університету.
Політика щодо перескладання	Студент має право на повторне складання модульного контролю з метою підвищення рейтингу протягом тижня після складання модульного контролю за графіком. Перескладання екзамену відбувається в терміни, визначені графіком освітнього процесу. Здобувач ВО має право на зарахування результатів навчання здобутих у неформальній чи інформальній освіті.
Політика щодо академічної доброчесності	При складанні усіх видів контролю у середовищі електронного навчання завжди активується система розпізнавання особи, що складає контроль. Усі практичні роботи у ЕНК перевіряються вбудованою системою Антиплагіат. При складанні усіх форм контролю забороняється списування, у тому числі з використанням сучасних інформаційних технологій.
Політика щодо відвідування	Відвідування занять є обов’язковим компонентом освітнього процесу. За наявності поважних причин (наприклад, хвороба, особливі потреби, відрядження, сімейні обставини, участь у програмах академічної мобільності тощо) навчання може здійснюватися за індивідуальним графіком, погодженим з деканом факультету.

СИСТЕМА ОЦІНЮВАННЯ

Розподіл балів, які отримують студенти за курс

Модуль 1			Модуль 2			Підсумковий контроль		Разом з дисципліни
Аудиторна та самостійна робота			Аудиторна та самостійна робота			Теоретичний курс	Практичне завдання	100
Теоретичний курс (тестування)	Лабораторна робота		Теоретичний курс (тестування)	Лабораторна робота				
16	21		17	21		15	10	
№ лекції	Види робіт	К-ть балів	№ лекції	Види робіт	К-ть балів			
Тема 1			Тема 8					
Тема 2	Лабораторна робота №1	7	Тема 9	Лабораторна робота №4	7			
Тема 3			Тема 10					
Тема 4			Тема 11	Лабораторна робота №5	7			
Тема 5	Лабораторна робота №2	7	Тема 12					
Тема 6			Тема 13	Лабораторна робота №6	7			
Тема 7	Лабораторна робота №3	7						

Розподіл балів, які отримують студенти за виконання та захист КП							
Модуль 1		Модуль 2		Підсумковий контроль		Разом за КП	
Виконання розділу 1		Виконання розділу 2		Захист КП		100	
30		45		25			
Види робіт	К-ть балів	Види робіт	К-ть балів				
Етап 1.1	10	Етап 2.1	15				
Етап 1.2	10	Етап 2.2	15				
Етап 1.3	10	Етап 2.3	15				
Розподіл оцінок							
Сума балів за навчальну діяльність		Шкала ECTS		Оцінка за національною шкалою			
90-100		A		Відмінно			
82-89		B		Добре			
75-81		C		Добре			
67-74		D		Задовільно			
60-66		E		Задовільно			
35-59		FX		Незадовільно з можливістю повторного складання			
1-34		F		Незадовільно з обов’язковим повторним вивченням дисципліни			

Затверджено рішенням кафедри КН, протокол №1 від «26» серпня 2024 року.

ПОГОДЖЕНО

Гарант освітньої програми канд. техн. наук, доцент кафедри КН

Леся ДМИТРОЦА

Затверджено рішенням кафедри КН, протокол №1 від «26» серпня 2024 року.

ПОГОДЖЕНО

Гарант освітньої програми канд. техн. наук, доцент кафедри КН

Леся ДМИТРОЦА

Затверджено рішенням кафедри КН, протокол №1 від «26» серпня 2024 року.

ПОГОДЖЕНО

Гарант освітньої програми канд. техн. наук, доцент кафедри КН

Леся ДМИТРОЦА

Затверджено рішенням кафедри КН, протокол №1 від «26» серпня 2024 року.

ПОГОДЖЕНО

Гарант освітньої програми канд. техн. наук, доцент кафедри КН

Леся ДМИТРОЦА