Інститут телекомунікацій і глобального інформаційного простору НАН України

Відділ природних ресурсів

Сучасні методи, моделі та технології оброблення даних в інформаційно-комунікаційних системах

Робоча програма навчальної дисципліни (силабус)

Реквізити навчальної дисципліни	
Рівень вищої освіти	***Третій (освітньо-науковий)***
Галузь знань	*Галузь F Інформаційні технології*
Спеціальність	*F3 – Комп'ютерні науки*
Освітньо-наукова програма	*Інформаційні технології*
Статус дисципліни	*Обов'язкова*
Форма навчання	*очна(денна)*
Рік підготовки, семестр	*2 курс осінній семестр*
Обсяг дисципліни	*120 годин / 4 кредити ЄКТС* *(лекції – 28 год., практичні заняття – 12 год., СРС – 80 год.)*
Семестровий контроль/ контрольні заходи	*іспит*
Розклад занять	*2 год лекційних та 1 год практичних занять на тиждень*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лекції та комп'ютерні практикуми заняття проводить: *член-кореспондент, д.т.н., професор, Трофимчук Олександр Миколайович, trofymchuk@nas.gov.ua*
Розміщення курсу	https://classroom.google.com/c/ODI2MTE3NDM2Mzg1?cjc=gmcgh36a

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Вивчення даної дисципліни базується на знаннях, уміннях і компетентностях, набутих під час опанування дисциплін обов'язкового компоненту циклів загальної та професійної підготовки, зокрема «Стан та перспективи розроблення інформаційних технологій та комп'ютерних систем», а також результатах науково-дослідної діяльності здобувача. Вивчення дисципліни «Сучасні методи, моделі та технології оброблення даних в інформаційно-комунікаційних системах» формує теоретичну та методологічну основу для подальшого опанування вибіркових освітніх компонентів професійної підготовки, зокрема дисциплін, пов'язаних із застосуванням методів штучного інтелекту, машинного навчання, інтелектуального аналізу даних, оброблення великих та неструктурованих даних, а також для виконання власного наукового дослідження.

Мета: *формування у здобувачів ступеня доктора філософії здатності до самостійного проведення наукових досліджень у сфері розроблення та застосування сучасних методів, моделей і технологій збирання, оброблення, аналізу, представлення та зберігання даних в*

інформаційно-комунікаційних системах, з використанням математичного та комп'ютерного моделювання, сучасних інформаційних технологій і методів аналізу даних для розв'язання складних науково-прикладних задач у галузі комп'ютерних наук.

***Предмет вивчення** – сучасні математичні та комп'ютерні моделі, методи, алгоритми і технології збирання, оброблення, аналізу, представлення, передавання та зберігання даних в інформаційно-комунікаційних системах, зокрема методи інтеграції, структурування, перетворення та інтелектуального аналізу даних, оцінювання їх якості й достовірності, виявлення закономірностей та отримання нових знань для розв'язання складних науково-прикладних задач у галузі комп'ютерних наук.*

Програмні результати навчання:

Загальні компетентності

ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.
ЗК02. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
ЗК04. Здатність розв'язувати комплексні проблеми комп'ютерних наук на основі системного наукового світогляду та загального культурного кругозору із дотриманням принципів професійної етики та академічної доброчесності.
ЗК06. Здатність до набуття концептуальних та методологічних знань в галузі чи на межі галузей знань або професійної діяльності, використання спеціалізованих умінь/навичок і методів, необхідних для розв'язання значущих проблем у сфері професійної діяльності, науки та/або інновацій, розширення та переоцінки вже існуючих знань і професійної практики, безперервного саморозвитку та самовдосконалення.

Спеціальні (фахові) компетентності

СК02. Здатність застосовувати сучасні методології, методи та інструменти експериментальних і теоретичних досліджень у сфері комп'ютерних наук, сучасні цифрові технології, бази даних та інші електронні ресурси у науковій та освітній діяльності.
СК03. Здатність виявляти, ставити та розв'язувати дослідницькі науково-прикладні завдання та/або проблеми в сфері комп'ютерних наук, оцінювати та забезпечувати якість виконуваних досліджень.
СК06. Здатність аналізувати та оцінювати сучасний стан і тенденції розвитку комп'ютерних наук та інформаційних технологій.
СК08. Здатність дотримуватися академічної та професійної доброчесності, морально-етичних правил поведінки, етики досліджень в академічному середовищі.

Програмні результати навчання

ПРН05. Планувати і виконувати експериментальні та/або теоретичні дослідження з комп'ютерних наук та дотичних міждисциплінарних напрямів з використанням сучасних інструментів, критично аналізувати результати власних досліджень і результати інших дослідників у контексті усього комплексу сучасних знань щодо досліджуваної проблеми.
ПРН06. Застосовувати сучасні інструменти і технології пошуку, оброблення та аналізу інформації, зокрема, статистичні методи аналізу даних великого обсягу та/або складної структури, спеціалізовані бази даних та інформаційні системи.

ПРН07. Розробляти та реалізовувати наукові та/або інноваційні інженерні проєкти, які дають можливість переосмислити наявне та створити нове цілісне знання та/або професійну практику і розв'язувати значущі наукові та технологічні проблеми комп'ютерних наук з дотриманням норм академічної етики і врахуванням соціальних, економічних, екологічних та правових аспектів.

ПРН08. Визначати актуальні наукові та практичні проблеми у сфері комп'ютерних наук, глибоко розуміти загальні принципи та методи комп'ютерних наук, а також методологію наукових досліджень, застосовувати їх у власних дослідженнях у сфері комп'ютерних наук та у викладацькій практиці.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Передумовами вивчення дисципліни є знання та компетентності, сформовані під час опанування обов'язкових освітніх компонентів ОНП, насамперед дисципліни «Стан та перспективи розроблення інформаційних технологій та комп'ютерних систем». Необхідними є базові знання з комп'ютерних наук, інформаційних технологій, методів і алгоритмів розв'язання теоретичних та прикладних задач, математичного і комп'ютерного моделювання, а також розуміння принципів функціонування інформаційно-комунікаційних систем і технологій оброблення даних.

Дисципліна «Сучасні методи, моделі та технології оброблення даних в інформаційно-комунікаційних системах» є обов'язковим компонентом Н.07 та вивчається у третьому семестрі після дисципліни Н.06 «Стан та перспективи розроблення інформаційних технологій та комп'ютерних систем». Результати навчання за дисципліною можуть використовуватися під час опанування вибіркових освітніх компонентів, зокрема дисциплін, пов'язаних із аналізом даних, машинним навчанням, штучним інтелектом, інтелектуальними інформаційними системами, обробленням великих і неструктурованих даних та текстовою аналітикою, а також під час виконання наукового дослідження і підготовки дисертації доктора філософії.

Результати навчання за дисципліною забезпечують формування методологічної та інструментальної основи для подальшого застосування сучасних методів, моделей і технологій оброблення та аналізу даних у науковій діяльності, зокрема під час виконання дослідницьких завдань, розроблення та дослідження математичних і комп'ютерних моделей, проведення експериментальних досліджень та підготовки наукових публікацій.

3. Зміст навчальної дисципліни

Тема 1. *Сучасні інформаційно-комунікаційні системи та дані.*

Тема 2. *Моделі та способи представлення даних.*

Тема 3. *Методи збирання та первинного оброблення даних.*

Тема 4. *Якість, достовірність та невизначеність даних.*

Тема 5. *Системи управління базами даних та технології зберігання даних.*

Тема 6. *Інтеграція, трансформація та консолідація даних.*

Тема 7. *Розподілене та паралельне оброблення даних.*

Тема 8. *Методи статистичного та розвідувального аналізу даних.*

Тема 9. *Методи оброблення даних великого обсягу та складної структури.*

Тема 10. *Оброблення потокових та динамічних даних.*

Тема 11. *Методи аналізу багатовимірних та складних даних.*

Тема 12. *Візуалізація та графічне представлення даних.*

Тема 13. *Інтелектуалізація процесів оброблення та аналізу даних.*

Тема 14. *Технології оброблення та аналізу даних у наукових дослідженнях.*

4. Навчальні матеріали та ресурси

Базова література

1. Dordal P. L. An Introduction to Computer Networks. 2nd ed. Chicago : Loyola University Chicago, 2022. – URL: https://intronetworks.cs.luc.edu/current2/html/

2. Curry E., Auer S., Berre A. J., Metzger A., Perez M. S., Zillner S. (eds.). Technologies and Applications for Big Data Value. Cham: Springer. – 2022. – XXIV, 544 p. – DOI: 10.1007/978-3-030-78307-5. – URL: https://link.springer.com/book/10.1007/978-3-030-78307-5

3. Bast H., Korzen C., Meyer U., Sanders P., et al. (eds.). Algorithms for Big Data: DFG Priority Program 1736. Cham: Springer. – 2022. – DOI: 10.1007/978-3-031-21534-6. – URL: https://link.springer.com/book/10.1007/978-3-031-21534-6

4. Curry E., Scerri S., Tuikka T. (eds.). Data Spaces: Design, Deployment and Future Directions. Cham: Springer. – 2022. – 357 p. – DOI: 10.1007/978-3-030-98636-0. – URL: https://link.springer.com/book/10.1007/978-3-030-98636-0

5. Peterson L. L., Davie B. S. Computer Networks: A Systems Approach. 5th ed. Burlington : Morgan Kaufmann, 2011. 928 p. DOI: 10.1016/C2009-0-63439-9. – URL: https://open.umn.edu/opentextbooks/textbooks/computer-networks-a-systems-approach

6. Brito P., Dias J. G., Lausen B., Montanari A., Nugent R. (eds.). Classification and Data Science in the Digital Age. Cham: Springer, – 2023. – DOI: 10.1007/978-3-031-09034-9. – URL: https://link.springer.com/book/10.1007/978-3-031-09034-9

7. Zhu H., Li T., Zhao B. Statistical Learning Methods for Neuroimaging Data Analysis with Applications // Annual Review of Biomedical Data Science. – 2023. – Vol. 6. – P. 73–104. – DOI: 10.1146/annurev-biodatasci-020722-100353. – URL: https://www.annualreviews.org/content/journals/10.1146/annurev-biodatasci-020722-100353

8. Khan W., Kumar T., Zhang C., Raj K., Roy A. M., Luo B. SQL and NoSQL Database Software Architecture Performance Analysis and Assessments - A Systematic Literature Review // Big Data and Cognitive Computing. – 2023. – Vol. 7, No. 2. – Art. 97. – DOI: 10.3390/bdcc7020097. – URL: https://www.mdpi.com/2504-2289/7/2/97

9. Flemisch B., Hermann S., Herschel M., et al. Research Data Management in Simulation Science: Infrastructure, Tools, and Applications // Datenbank-Spektrum. – 2024. – Vol. 24. – P. 97–105. – DOI: 10.1007/s13222-024-00475-4. – URL: https://link.springer.com/article/10.1007/s13222-024-00475-4

10. Bossaller J., Million A. J. The research data life cycle, legacy data, and dilemmas in research data management // Journal of the Association for Information Science and Technology. – 2023. – Vol. 74, No. 6. – P. 701–706. – DOI: 10.1002/asi.24645. – URL: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.24645

11. Patel H., Guttula S., Gupta N., Hans S., Mittal R. S., et al. A Data-centric AI Framework for Automating Exploratory Data Analysis and Data Quality Tasks // ACM Journal of Data and Information Quality. – 2023. – Vol. 15, No. 4. – Art. 44. – P. 1–26. – DOI: 10.1145/3603709. – URL: https://dl.acm.org/doi/10.1145/3603709

12. Taipalus T. Database management system performance comparisons: A systematic literature review // Journal of Systems and Software. – 2024. – Vol. 208. – Art. 111872. – DOI: 10.1016/j.jss.2023.111872. – URL: https://www.sciencedirect.com/science/article/pii/S0164121223002674?via%3Dihub

13. Khan W., Kumar T., Zhang C., Raj K., Roy A. M., Luo B. SQL and NoSQL Database Software Architecture Performance Analysis and Assessments - A Systematic Literature Review // Big Data and

Cognitive Computing. – 2023. – Vol. 7, No. 2. – Art. 97. – DOI: 10.3390/bdcc7020097. – URL: https://www.mdpi.com/2504-2289/7/2/97

14. Fragkoulis M., Carbone P., Kalavri V., Katsifodimos A. A survey on the evolution of stream processing systems // The VLDB Journal. – 2024. – Vol. 33, No. 2. – P. 507–541. – DOI: 10.1007/s00778-023-00819-8. – URL: https://link.springer.com/article/10.1007/s00778-023-00819-8

15. Jia W., Sun M., Lian J., Hou S. Feature dimensionality reduction: a review // Complex & Intelligent Systems. – 2022. – Vol. 8. – P. 2663–2693. – DOI: 10.1007/s40747-021-00637-x. – URL: https://link.springer.com/article/10.1007/s40747-021-00637-x

16. Ulmer A., Angelini M., Fekete J.-D., Kohlhammer J., May T. A Survey on Progressive Visualization // IEEE Transactions on Visualization and Computer Graphics. – 2024. – Vol. 30, No. 9. – P. 6447–6467. – DOI: 10.1109/TVCG.2023.3346641. – URL: https://ieeexplore.ieee.org/document/10373169

17. Zhou S., Xu H., Zheng Z., Chen J., et al. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions // ACM Computing Surveys. – 2025. – Vol. 57, No. 3. – Art. 69. – P. 1–38. – DOI: 10.1145/3689036. – URL: https://dl.acm.org/doi/10.1145/3689036

18. Mitchell S. N., Lahiff A., Cummings N., et al. FAIR data pipeline: provenance-driven data management for traceable scientific workflows // Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. – 2022. – Vol. 380, No. 2233. – Art. 20210300. – DOI: 10.1098/rsta.2021.0300. – URL: https://royalsocietypublishing.org/rsta/article/380/2233/20210300/112235/FAIR-data-pipeline-provenance-driven-data

19. Wilkinson S. R., Aloqalaa M., Belhajjame K., et al. Applying the FAIR Principles to computational workflows // Scientific Data. – 2025. – Vol. 12. – Art. 328. – DOI: 10.1038/s41597-025-04451-9. – URL: https://www.nature.com/articles/s41597-025-04451-9

20. Dovgiy S., Kopiika O., Kozlov O. Architectures for the Information Systems, Network Resources, and Network Services / Ceur Workshop Proceedings. – 2021. – p. 293–301. – https://ceur-ws.org/Vol-3187/short9.pdf

21. Trofymchuk O., Anpilova Y., Hordiienko O., Karpenko O. Method for Detecting Natural and Anthropogenic Changes That Filled with Water in Landscapes Using Radar Satellite Imagery // Book Chapter in Systems, Decision and Control in Energy. – 2023. – p. 799-813. – DOI: 10.1007/978-3-031-35088-7_46

22. Trofymchuk O., Vishnyakov V., Sheviakina N., Klymenko V., Zahorodnia S. Geoinformation terrain analysis for optimizing military equipment routes / International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEMConference Paper. – 2024. – ISSN: 1314-2704. – DOI: 10.5593/sgem2024v/3.2/s12.281 – https://epslibrary.at/sgem_jresearch_publication_view.php?page=view&editid1=10073

23. Довгий С.О., Копійка О.В. Зміна бізнес-моделі управління ІТ на підприємстві природокористування у зв'язку з розвитком сервісно-орієнтованих інформаційних технологій // Екологічна безпека та природокористування. – 2021. – № 1 (37). – ISSN: 2411-4049. – https://es-journal.in.ua/article/view/228374

24. Довгий С.О., Копійка О.В., Козлов О.С., Литвиненко А.О. Інформаційні технології ІТ-сервісів архітектури додатків // Екологічна безпека та природокористування. – 2023. – 3 (47). – ISSN: 2411-4049. – https://www.researchgate.net/publication/376372858_Information_technologies_of_application_architectures_IT_services

Додаткова література

25. Kuznietsova N.V., Trofymchuk O.M., Bidyuk P.I., Terentiev O.M., Levenchuk L.B. Bayesian modelling of risks of various origin // KPI Science News, no. 4, pp. 7–18, 2021. – doi: 10.20535/kpisn.2021.4.251684. – http://scinews.kpi.ua/article/view/251684

26. Трофимчук О.М., Бідюк П.І., Терентьєв О.М., Просянкіна-Жарова Т.І. Математичне моделювання, інтелектуальний аналіз даних та штучний інтелект для підтримки прийняття рішень з повоєнного відновлення // Екологічна безпека та природокористування. – 2025. – вип. 3 (55). – с. 33-49. – ISSN: 2411-4049. – https://doi.org/10.32347/2411-4049.2025.3.33-49

27. Dovgyi S., Zoziuk M., Koroliouk D. Adaptive Deep Learning System for Investigating General Data // Cybernetics and Systems Analysis. – 2023. – 59(5). – p. 704–712. – https://link.springer.com/article/10.1007/s10559-023-00606-x
2025. – 2 (9). – с. 10-19. – http://mmtt.khpi.edu.ua/article/view/341038/329077

28. Довгий С.О., Буланчук Г.Г., Буланчук О.М, Листопадова В.В. Точність розрахунків поля швидкостей від системи вихрових відрізків при використанні графічних процесорів // Вісник Національного технічного університету «ХПІ». Серія: Математичне моделювання в техніці та технологіях. – 2023. – Випуск 1. – с. 103-109.

29. Терентьєв О.М., Дуда В.О. Метод відновлення пропусків у даних на основі комбінованої моделі експоненційного згладжування // Екологічна безпека та природокористування. – 2025. – вип. 1 (53). – с. 125-131. – ISSN: 2411-4049. – https://doi.org/10.32347/2411-4049.2025.1.125-131 - ISSN: 2411-4049. – https://es-journal.in.ua/article/view/328615

Нормативно-правова література:

1. Розпорядження Кабінету Міністрів України «Про схвалення Концепції Державної цільової науково-технічної програми з використання технологій штучного інтелекту в пріоритетних галузях економіки на період до 2026 року» від 13 квітня 2024 р. № 320-р // База даних «Законодавство України» / Верховна Рада України. – Режим доступу: https://zakon.rada.gov.ua/go/320-2024-%D1%80 (дата звернення: 19.01.2026).

2. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (Artificial Intelligence Act) // Official Journal of the European Union. – L 1689, 12.7.2024. – URL: http://data.europa.eu/eli/reg/2024/1689/oj (дата звернення: 19.01.2026).

Інформаційні ресурси:

1. Світовий центр даних КПІ - http://wdc.kpi.ua/uk (дата звернення: 06.06.2025).

2. Аналітичний онлайн-ресурс KDnuggets про Data Science, Machine Learning, AI та аналітики великих даних – https://www.kdnuggets.com/ (дата звернення: 06.06.2025).

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

Для лекційних занять використовуються пояснювально-ілюстративний метод та метод проблемного виконання, для проведення практичних занять використовується дослідницький метод навчання: викладач ставить перед аспірантами проблему, і ті вирішують її самостійно або під керівництвом викладача.

За дистанційної форми навчання заняття проводять за допомогою платформи для проведення онлайн-зустрічей Zoom

Назви тем	**Кількість навчальних годин**				**Форми контролю**
	Усього годин (кредитів)	**Лекції**	**Комп'ютерний практикум**	**Самостійна робота студентів**	
	120 год	*28 год*	*12 год*	*80 год*	
Тема 1. Сучасні інформаційно-комунікаційні системи та дані.	8	2		6	Участь у дискусії, тестування

Тема 2. Моделі та способи представлення даних.	10	2	2	6	Участь у дискусії, тестування
Тема 3. Методи збирання та первинного оброблення даних.	8	2		6	Участь у дискусії, тестування
Тема 4. Якість, достовірність та невизначеність даних.	10	2	2	6	Участь у дискусії, тестування
Тема 5. Системи управління базами даних та технології зберігання даних.	8	2		6	Участь у дискусії, тестування
Тема 6. Інтеграція, трансформація та консолідація даних.	10	2	2	6	Участь у дискусії, тестування
Тема 7. Розподілене та паралельне оброблення даних.	8	2		6	Участь у дискусії, тестування
Тема 8. Методи статистичного та розвідувального аналізу даних.	10	2	2	6	Участь у дискусії, тестування
Тема 9. Методи оброблення даних великого обсягу та складної структури.	8	2		6	Участь у дискусії, тестування
Тема 10. Оброблення потокових та динамічних даних.	10	2	2	6	Участь у дискусії, тестування
Тема 11. Методи аналізу багатовимірних та складних даних.	8	2		6	Участь у дискусії, тестування
Тема 12. Візуалізація та графічне представлення даних.	8	2		6	Участь у дискусії, тестування
Тема 13. Інтелектуалізація процесів оброблення та аналізу даних.	8	2		6	Участь у дискусії, тестування
Тема 14. Технології оброблення та аналізу даних у наукових дослідженнях.	6	2	2	2	Участь у дискусії, тестування
Загалом	**120**	**28**	**12**	**80**	

6. Самостійна робота аспіранта

Розподіл годин самостійної роботи для аспірантів:

1.Підготовка до іспиту – 2 год.

2.Опрацювання питань, які не розглядаються на аудиторних заняттях – 78 год., з них на виконання індивідуальних дослідницьких завдань – 12 год.

Усього 80 год.

Передбачено три індивідуально-дослідницькі завдання у межах самостійної роботи, що спрямовані на формування і розвиток практичних умінь підготовки публікацій, доповідей та проєктів з актуальних проблем

Індивідуально-дослідницьке завдання 1.

Тема: дослідження методів підготовки та оцінювання якості даних в інформаційно-комунікаційних системах.

Мета: дослідження методів збирання, представлення, первинного оброблення та оцінювання якості даних, отриманих із джерел інформаційно-комунікаційних систем, а також визначення впливу пропущених, аномальних, суперечливих та невизначених даних на результати їх подальшого аналізу.

Послідовність виконання завдання:

1. Визначити предметну область та сформулювати дослідницьку задачу.
2. Обрати та охарактеризувати набір даних і джерела його отримання.
3. Визначити структуру, типи, формати та особливості представлення даних.
4. Провести аналіз повноти, достовірності, узгодженості та інших характеристик якості даних.
5. Виявити пропущені, аномальні, помилкові та суперечливі значення.
6. Обґрунтувати методи первинного оброблення, очищення та трансформації даних.
7. Реалізувати обрані методи засобами програмного забезпечення.
8. Провести порівняльний аналіз характеристик даних до та після оброблення.
9. Дослідити вплив якості підготовлених даних на результати подальшого статистичного або інтелектуального аналізу.
10. Проаналізувати отримані результати та сформулювати науково обґрунтовані висновки.

Це індивідуальне дослідницьке завдання пов’язане із темами 1-4.

Індивідуально-дослідницьке завдання 2.

Тема: дослідження методів інтеграції та масштабованого оброблення даних в інформаційно-комунікаційних системах.

Мета: lослідження та експериментальне порівняння методів зберігання, інтеграції, трансформації, консолідації та розподіленого оброблення різнорідних даних з оцінюванням їх обчислювальної ефективності та масштабованості.

Послідовність виконання завдання:

1. Визначити предметну область та сформулювати дослідницьку задачу.

2. Визначити джерела та охарактеризувати різнорідні набори даних.

3. Обґрунтувати модель та технологію зберігання даних.

4. Розробити схему інтеграції, трансформації та консолідації даних.

5. Визначити технологічний pipeline оброблення даних.

6. Реалізувати запропонований підхід засобами відповідного програмного забезпечення.

7. Дослідити можливості централізованого, паралельного або розподіленого оброблення даних.

8. Провести експерименти на наборах даних різного обсягу та/або складності.

9. Оцінити часові, обчислювальні та інші характеристики ефективності оброблення.

10. Порівняти результати застосування різних підходів і визначити умови їх ефективного використання.

11. Сформулювати висновки та рекомендації щодо вибору технології оброблення даних.

Це індивідуальне дослідницьке завдання пов’язане із темами 5-9.

Індивідуально-дослідницьке завдання 3.

Тема: Дослідження методів аналізу потокових, багатовимірних та складних даних у наукових дослідженнях.

Мета: розроблення та експериментальне дослідження методики оброблення й аналізу потокових, багатовимірних або складних даних із використанням статистичних, інтелектуальних та візуальних методів для розв’язання науково-дослідницької задачі.

Послідовність виконання завдання:

1. Сформулювати наукову проблему, дослідницьке питання та, за необхідності, дослідницьку гіпотезу.

2. Визначити джерело та охарактеризувати досліджувані дані.

3. Визначити особливості їх динамічності, багатовимірності або складної структури.

4. Розробити послідовність оброблення та аналізу даних.

5. Обґрунтувати вибір статистичних, інтелектуальних та/або алгоритмічних методів.

6. Реалізувати обрані методи та провести обчислювальний експеримент.

7. Дослідити методи зниження розмірності, кластеризації, класифікації, прогнозування або виявлення аномалій відповідно до поставленої задачі.

8. Виконати графічну та візуальну інтерпретацію отриманих результатів.

9. Провести порівняння альтернативних методів за визначеними критеріями.

10. Оцінити достовірність, стійкість та відтворюваність отриманих результатів.

11. Інтерпретувати результати з позицій поставленої наукової задачі.

12. Сформулювати наукові висновки та визначити перспективи подальших досліджень.

Це індивідуальне дослідницьке завдання пов’язане із темами 10-14.

Результати опанування відповідного матеріалу перевіряються під час усного опитування і підсумкового (іспит) контролю знань.

Самостійна робота аспірантів передбачає підготовку та поглиблене вивчення тем навчальної дисципліни за напрямками:

1. Дослідження сучасних моделей представлення, зберігання та управління даними в інформаційно-комунікаційних системах, зокрема реляційних, нереляційних, розподілених та інших моделей даних.

2. Дослідження методів збирання, первинного оброблення та оцінювання якості даних, включаючи очищення, нормалізацію, оброблення пропущених і аномальних значень, а також оцінювання достовірності та невизначеності даних.

3. Дослідження методів інтеграції, трансформації, консолідації та масштабованого оброблення даних у розподілених і хмарних інформаційно-комунікаційних системах, зокрема для великих та складноструктурованих наборів даних.

4. Дослідження методів статистичного, інтелектуального та багатовимірного аналізу даних, включаючи виявлення закономірностей, кластеризацію, класифікацію, зниження розмірності, прогнозування та виявлення аномалій.

5. Дослідження сучасних технологій оброблення потокових і складних даних та візуалізації результатів їх аналізу з використанням програмних засобів для проведення обчислювальних експериментів, оцінювання результатів і забезпечення відтворюваності наукових досліджень.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Вимоги, яких має дотримуватися аспірант в рамках даної дисципліни:

- правила відвідування занять: відвідування лекцій та практичних занять, а також відсутність на них, не оцінюється. Однак, аспірантам рекомендується відвідувати заняття, оскільки на них викладається теоретичний матеріал та розвиваються навички, необхідні для виконання семестрового індивідуального завдання та проводяться контрольні заходи (тести) з поточної оцінки самостійної роботи аспірантів з засвоєння поточного матеріалу. Останні є складовою частиною поточного рейтингу і проводяться тільки у день проведення відповідних лекцій та практичних занять. Система оцінювання орієнтована на отримання балів за своєчасність виконання аспірантами практичних та контрольних робіт, а також виконання завдань, які здатні розвинути практичні уміння та навички;
- правила поведінки на заняттях: аспірант повинен брати участь у розв’язку задач, готувати короткі доповіді;
- захист комп’ютерних практикумів – захист відбувається у визначені терміні під час аудиторних занять;
- політика щодо академічної доброчесності – політика та принципи академічної доброчесності визначені у Етичному кодексі вченого Інституту телекомунікацій та глобального інформаційного простору НАН України.

8. Види контролю та рейтингова система оцінювання результатів навчання (РСО)

Семестровий контроль - іспит. Рейтингова система оцінювання результатів навчання передбачає оцінювання заходів поточного контролю з дисципліни впродовж семестру. Рейтингова оцінка здобувача складається з балів, отриманих здобувачем за результатами заходів поточного контролю. Рейтингова оцінка доводиться до здобувачів на передостанньому занятті з дисципліни в семестрі. Здобувачі, які виконали всі умови допуску до іспиту та мають рейтингову оцінку 60 і більше балів, отримують відповідну до набраного рейтингу оцінку без додаткових випробувань

Зі здобувачами, які виконали всі умови допуску до іспиту та мають рейтингову оцінку менше 60 балів, а також з тими здобувачами, хто бажає підвищити свою рейтингову оцінку, на останньому

за розкладом занятті з дисципліни в семестрі викладач проводить семестровий контроль у вигляді контрольної роботи.

Для посилення зацікавленості здобувачів у якісному виконанні індивідуальних семестрових завдань, передбачених індивідуальним навчальним планом здобувача, рейтингову оцінку, у разі виконання контрольної роботи, можна визначати як суму балів за контрольну роботу та балів за індивідуальне семестрове завдання. У цьому випадку розмір шкали оцінювання контрольної роботи зменшується на максимальне значення балів, передбачених за виконання відповідного індивідуального семестрового завдання.

Після виконання контрольної роботи, якщо оцінка за контрольну роботу більша ніж за рейтингом, здобувач отримує оцінку за результатами контрольної роботи. Якщо оцінка за контрольну роботу менша ніж за рейтингом, то здобувач отримує більшу з оцінок, що отримані за результатами контрольної роботи або за рейтингом.

До відомості семестрового контролю викладач заносить рейтингові бали, отримані здобувачем у семестрі або за результатами виконання контрольної роботи, та оцінку відповідно до цих балів

Критерії нарахування балів:

1. Комп'ютерні практикуми оцінюються виходячи з максимальної кількості балів - 10 балів кожне:

- «відмінно» –95 відсотків максимального балу;
- «добре» –75-95 відсотків максимального балу;
- «задовільно» –60-75 відсотків максимального балу;
- «достатньо» – 50 відсотків – робота виконана, але не захищена.

Умови допуску до підсумкового контролю є зарахування усіх практичних робіт Рейтинг аспіранта з дисципліни складається з балів, що він отримує за:

– виконання ПРАКТИЧНИХ (лабораторних) робіт;

– виконання самостійної роботи.

За період вивчення дисципліни аспірант може набрати 100 балів. Їх розподіл між видами робіт наведено в таблиці 1.

Таблиця 1. Розподіл між видами робіт

Бали за виконання	Номер комп'ютерного практикуму або індивідуально дослідницького завдання						Разом
	1	2	3	4	5	6	
Комп'ютерний практикум	10	10	10	10	10	10	100
Індивідуально дослідницьке завдання	20	10	10				

2. Контрольна робота оцінюється за такими критеріями:

- «відмінно » – повна відповідь (не менше 90% потрібної інформації), надані відповідні обґрунтування та особистий погляд;
- «добре» – достатньо повна відповідь (не менше 75% потрібної інформації), що виконана згідно з вимогами до рівня «умінь», або незначні неточності);
- «задовільно» – неповна відповідь (не менше 60% потрібної інформації. що виконана згідно з вимогами до «стереотипного» рівня та деякі помилки);
- «незадовільно» – незадовільна відповідь–0 балів.

Контрольна робота передбачається у вигляді тесту, критерії оцінювання тесту:

Кількість правильних відповідей	*Відсоток правильних відповідей*	*Оцінка за національною шкалою*	*Оцінка за шкалою ECTS*
48-50	*95-100*	*Відмінно*	*A*
41-47	*82-94*	*Дуже добре*	*B*
37-40	*75-81*	*Добре*	*C*
34-36	*69-74*	*Задовільно*	*D*
30-33	*60-68*	*Достатньо*	*E*
5-29	*10-59*	*Не задовільно*	*FX*

Відповідність рейтингових балів оцінкам за шкалою Інституту та *шкалою ECTS*

Рейтингова оцінка	*Оцінка за національною шкалою*	*Оцінка за шкалою ECTS*
90-100	*Відмінно*	*A*
82-89	*Дуже добре*	*B*
75-81	*Добре*	*C*
69-74	*Задовільно*	*D*
60-68	*Достатньо*	*E*
45-59	*Не задовільно*	*FX*
Невиконання умов допуску до семестрового контролю	*Не допущено*	

9. Додаткова інформація з дисципліни (освітнього компонента)

- *перелік питань, які виносяться на семестровий контроль у додатку*
- є можливість зарахування сертифікатів проходження дистанційних чи онлайн курсів Coursera, EdX за відповідною тематикою – зараховується додатково до 10 балів до загального рейтингу аспіранта, якщо аспірант набрав не менше 75 балів за період вивчення курсу та отримав відповідний сертифікат.

Робочу програму навчальної дисципліни (силабус):

Склав: д.т.н., професор, член-кореспондент НАН України Трофимчук Олександр Миколайович

Затверджено:

Вченою радою Інституту телекомунікацій
і глобального інформаційного простору
НАН України Протокол №10 від 13.08.2026

Введено в дію:

Наказом директора
Наказ від 14.08.2026 № 12

Додаток 1. Теми і зміст лекцій

Тема 1. Сучасні інформаційно-комунікаційні системи та дані.

Поняття даних та інформації. Життєвий цикл даних в інформаційно-комунікаційних системах. Джерела та типи даних. Структуровані, напівструктуровані та неструктуровані дані. Архітектура сучасних систем оброблення даних. Основні етапи: збирання, передавання, зберігання, оброблення, аналіз та представлення даних.

Тема 2. Моделі та способи представлення даних.

Моделі даних та їх класифікація. Реляційна, об'єктна, документна, графова та інші моделі. Структури даних і метадані. Представлення складних та багатовимірних даних. Вибір моделі даних відповідно до характеру наукової та прикладної задачі.

Тема 3. Методи збирання та первинного оброблення даних.

Джерела отримання даних. Методи автоматизованого та експериментального збирання даних. Інтеграція даних із різнорідних джерел. Первинне перетворення даних. Фільтрація, нормалізація, кодування та агрегування. Формування наборів даних для подальшого аналізу.

Тема 4. Якість, достовірність та невизначеність даних.

Основні характеристики якості даних. Повнота, точність, узгодженість, актуальність та достовірність. Пропущені, суперечливі та аномальні значення. Методи виявлення та оброблення помилок. Джерела невизначеності в даних та їх врахування під час наукових досліджень.

Тема 5. Системи управління базами даних та технології зберігання даних.

Принципи організації баз даних. Реляційні бази даних та SQL. Проєктування структури даних. Індексація та оптимізація доступу. Сучасні нереляційні системи зберігання даних. Вибір технології зберігання залежно від структури та обсягу даних.

Тема 6. Інтеграція, трансформація та консолідація даних.

Проблеми інтеграції різнорідних даних. ETL та ELT-підходи. Очищення та трансформація даних. Узгодження схем і форматів. Консолідація розподіленої інформації. Інтероперабельність інформаційних систем. Формування інтегрованого інформаційного середовища.

Тема 7. Розподілене та паралельне оброблення даних.

Архітектури розподіленого оброблення. Розподіл даних і обчислень. Паралельне оброблення. Масштабованість та відмовостійкість. Розподілені сховища. Оброблення даних у мережевих, мобільних та хмарних середовищах.

Тема 8. Методи статистичного та розвідувального аналізу даних.

Основи статистичного аналізу даних. Описова статистика. Дослідження розподілів та взаємозв'язків між змінними. Кореляційний та регресійний аналіз. Виявлення закономірностей та аномалій. Exploratory Data Analysis. Статистичне обґрунтування результатів оброблення даних.

Тема 9. Методи оброблення даних великого обсягу та складної структури.

Поняття Big Data та його характеристики. Проблеми масштабованості. Методи оброблення великих масивів даних. Пакетне та потокове оброблення. Масштабовані обчислення. Особливості роботи з високовимірними та складними даними.

Тема 10. Оброблення потокових та динамічних даних.

Особливості потокових даних. Дані реального та близького до реального часу. Потокове оброблення та віконні методи. Фільтрація та агрегування потоків. Виявлення змін і аномалій у потоках. Оброблення даних від сенсорних та мережевих джерел. Застосування до дослідження динамічних процесів.

Тема 11. Методи аналізу багатовимірних та складних даних.

Багатовимірні дані та проблема розмірності. Методи зниження розмірності. Виявлення прихованої структури даних. Кластеризація та групування. Аналіз залежностей між ознаками. Вибір інформативних характеристик. Підготовка даних для математичного та комп'ютерного моделювання.

Тема 12. Візуалізація та графічне представлення даних.

Принципи візуального аналізу даних. Табличне, графічне та інтерактивне представлення. Візуалізація багатовимірних даних. Вибір способу візуального представлення залежно від типу даних і дослідницької задачі. Інтерактивна аналітика. Візуалізація результатів наукового дослідження.

Тема 13. Інтелектуалізація процесів оброблення та аналізу даних.

Поняття інтелектуального оброблення даних. Автоматизоване виявлення закономірностей. Методи класифікації, прогнозування та виявлення аномалій. Основи машинного навчання як інструменту аналізу даних. Поєднання статистичних, алгоритмічних та інтелектуальних методів.

Тема 14. Технології оброблення та аналізу даних у наукових дослідженнях.

Побудова повного конвеєра оброблення даних у науковому дослідженні. Формулювання дослідницької задачі та вибір методів. Підготовка, оброблення та аналіз даних. Побудова математичних і комп'ютерних моделей. Проведення обчислювальних експериментів. Валідація та інтерпретація результатів. Відтворюваність досліджень. Представлення та документування результатів.

Додаток 2. Теми і зміст комп'ютерних практикумів

Комп'ютерний практикум 1.
Збір даних, первинна обробка та оцінювання якості даних.

Зміст заняття.
Практичне дослідження повного циклу підготовки даних до подальшого аналізу. Формування або отримання набору даних із відкритих джерел, аналіз його структури та типів даних, первинне профілювання, виявлення пропущених, некоректних, дубльованих та аномальних значень. Застосування методів очищення, нормалізації, стандартизації, кодування та трансформації даних. Визначення критеріїв оцінювання якості даних та кількісне порівняння характеристик набору до і після оброблення. Візуальний та статистичний аналіз результатів підготовки даних із використанням програмних засобів.

Основні завдання практикуму:
1. Обрати предметну область та визначити джерело отримання даних.
2. Отримати та завантажити реальний набір даних із відкритого джерела.
3. Визначити структуру набору даних, типи змінних, обсяг та характеристики даних.
4. Провести первинне профілювання та дослідницький аналіз набору даних.
5. Виявити пропущені, дубльовані, некоректні, суперечливі та аномальні значення.
6. Визначити основні проблеми якості даних та їх можливі причини.
7. Обґрунтувати методи очищення та первинної обробки даних.
8. Виконати оброблення пропущених та аномальних значень.
9. Застосувати необхідні методи нормалізації, стандартизації, кодування та трансформації даних.
10. Сформувати кількісні показники якості даних до та після оброблення.
11. Порівняти отримані результати та оцінити вплив первинного оброблення на подальшу аналітичну придатність даних.
12. Візуалізувати результати та сформулювати висновки щодо ефективності застосованих методів.

Очікувані результати.
Після виконання практикуму здобувач повинен:
1. уміти отримувати та структурувати дані з відкритих джерел;
2. визначати тип, структуру та основні характеристики наборів даних;
3. проводити профілювання та первинний аналіз даних;
4. виявляти основні проблеми якості даних;
5. застосовувати методи очищення, нормалізації, стандартизації та трансформації даних;
6. обґрунтовано обирати методи оброблення пропущених та аномальних значень;
7. використовувати кількісні критерії для оцінювання якості даних;
8. проводити порівняльний аналіз даних до та після preprocessing;
9. програмно реалізовувати процедури первинного оброблення даних;
10. інтерпретувати результати обчислювального експерименту та формулювати науково обґрунтовані висновки.

Літературні джерела:
1. Starbuck C. Data Preparation // The Fundamentals of People Analytics: With Applications in R. Cham: Springer. – 2023. – P. 79–95. – DOI: 10.1007/978-3-031-28674-2_6. – URL: https://link.springer.com/chapter/10.1007/978-3-031-28674-2_6
2. Maharana K., Mondal S., Nemade B. A review: Data pre-processing and data augmentation techniques // Global Transitions Proceedings. – 2022. – Vol. 3, No. 1. – P. 91–99. – DOI: 10.1016/j.gltp.2022.04.020. – URL: https://www.sciencedirect.com/science/article/pii/S2666285X22000565?via%3Dihub

3. Zhu J., Zhao X., Sun Y., Song S., Yuan X. Relational Data Cleaning Meets Artificial Intelligence: A Survey // Data Science and Engineering. – 2025. – Vol. 10, No. 2. – P. 147–174. – DOI: 10.1007/s41019-024-00266-7. – URL: https://link.springer.com/article/10.1007/s41019-024-00266-7

4. Li N., Qi Y., Li C., Zhao Z. Active Learning for Data Quality Control: A Survey // ACM Journal of Data and Information Quality. – 2024. – Vol. 16, No. 2. – Art. 11. – P. 1–45. – DOI: 10.1145/3663369. – URL: https://dl.acm.org/doi/10.1145/3663369

Комп'ютерний практикум 2.

Моделювання та організація зберігання даних у системах управління базами даних

Зміст заняття:

Дослідження принципів концептуального, логічного та фізичного моделювання даних. Побудова ER-моделі предметної області та її перетворення у реляційну схему. Визначення сутностей, атрибутів, зв'язків, ключів та обмежень цілісності. Нормалізація структури бази даних. Реалізація створеної моделі у системі управління базами даних із використанням SQL. Дослідження альтернативних підходів до організації зберігання даних у NoSQL-системах. Порівняння реляційної та нереляційної моделей за структурою даних, способом доступу та характеристиками виконання запитів. Дослідження впливу структури схеми, індексації та способу організації даних на продуктивність СУБД.

Основні завдання практикуму:

1. Визначити предметну область та сформулювати вимоги до інформаційної системи.
2. Визначити основні сутності, атрибути та зв'язки між ними.
3. Побудувати концептуальну модель даних у вигляді ER-діаграми.
4. Перетворити концептуальну модель у логічну реляційну модель.
5. Визначити первинні та зовнішні ключі, обмеження цілісності та типи даних.
6. Виконати нормалізацію розробленої структури бази даних.
7. Реалізувати базу даних у вибраній СУБД.
8. Створити таблиці, індекси, обмеження цілісності та необхідні SQL-запити.
9. Сформувати тестовий набір даних та виконати операції додавання, модифікації, пошуку й видалення даних.
10. Дослідити вплив індексації та структури схеми на час виконання запитів.
11. За необхідності реалізувати альтернативне представлення даних у NoSQL-СУБД.
12. Порівняти реляційний та нереляційний підходи до організації зберігання даних.
13. Провести експериментальне оцінювання продуктивності розроблених структур.
14. Проаналізувати отримані результати та сформулювати рекомендації щодо вибору моделі даних і СУБД.

Очікувані результати. Після виконання практикуму здобувач повинен:

1. уміти формалізувати вимоги до зберігання даних інформаційної системи;
2. розробляти концептуальні та логічні моделі даних;
3. будувати ER-моделі та перетворювати їх у реляційні схеми;
4. визначати ключі, зв'язки та обмеження цілісності;
5. застосовувати принципи нормалізації баз даних;
6. створювати та налаштовувати структури баз даних у СУБД;
7. використовувати SQL для створення, модифікації та опрацювання даних;
8. застосовувати індексацію та інші засоби оптимізації доступу до даних;
9. розуміти особливості реляційних і NoSQL-моделей;
10. оцінювати вплив структури бази даних на продуктивність запитів;
11. проводити експериментальне порівняння альтернативних способів організації зберігання даних;
12. обґрунтовувати вибір моделі даних і СУБД відповідно до характеристик прикладної задачі.

Літературні джерела:

1. Huawei Technologies Co., Ltd. Database Principles and Technologies – Based on Huawei GaussDB. Singapore: Springer, 2022. XIII, 314 p. – DOI: 10.1007/978-981-19-3032-4. – URL: https://link.springer.com/book/10.1007/978-981-19-3032-4

2. Mazurova O., Syvolovskyi I., Syvolovska O. NOSQL database logic design methods for MongoDB and Neo4j // Innovative Technologies and Scientific Solutions for Industries. – 2022. No. 2 (20). – P. 52–63. – DOI: 10.30837/ITSSI.2022.20.052. – URL: https://journals.uran.ua/itssi/article/view/262054

3. Fernández Candel C. J., Sevilla Ruiz D., García Molina J. J. A unified metamodel for NoSQL and relational databases // Information Systems. – 2022. – Vol. 104. – Art. 101898. – DOI: 10.1016/j.is.2021.101898. – URL: https://www.sciencedirect.com/science/article/pii/S0306437921001149?via%3Dihub

4. Ha M., Shichkina Y. Translating a Distributed Relational Database to a Document Database // Data Science and Engineering. – 2022. – Vol. 7. – P. 136–155. – DOI: 10.1007/s41019-022-00181-9. – URL: https://link.springer.com/article/10.1007/s41019-022-00181-9

Комп'ютерний практикум 3.

Інтеграція та трансформація різнорідних даних.

Зміст заняття:

Дослідження методів інтеграції даних із різнорідних джерел, що відрізняються структурою, форматом, схемою, семантикою та способом організації. Формування єдиного представлення даних на основі декількох джерел. Дослідження методів ETL/ELT, зіставлення схем, очищення та трансформації даних, гармонізації форматів і семантики. Реалізація pipeline інтеграції структурованих, напівструктурованих та, за можливості, неструктурованих даних. Дослідження підходів до централізованої інтеграції, data federation та віртуалізації даних. Оцінювання повноти, узгодженості та якості результатів інтеграції.

Основні завдання практикуму:

1. Визначити предметну область та сформулювати задачу інтеграції даних.
2. Обрати щонайменше два різнорідних джерела даних.
3. Дослідити структуру, формати, схеми та семантику даних у вибраних джерелах.
4. Визначити відмінності між джерелами, зокрема у назвах, типах, форматах та одиницях вимірювання атрибутів.
5. Розробити схему зіставлення структур та відповідностей між джерелами даних.
6. Обґрунтувати метод інтеграції та послідовність операцій ETL/ELT.
7. Виконати очищення, перетворення, нормалізацію та гармонізацію даних.
8. Реалізувати pipeline інтеграції засобами відповідного програмного забезпечення.
9. Сформувати консолідований набір даних або єдине представлення інтегрованих даних.
10. Перевірити повноту, узгодженість, відсутність дублювання та коректність інтегрованого набору.
11. Порівняти характеристики вихідних та інтегрованого наборів даних.
12. Дослідити вплив різних способів трансформації на результати інтеграції.
13. Візуалізувати результати та оцінити якість виконаної інтеграції.
14. Сформулювати висновки щодо ефективності застосованого підходу та можливостей його використання для інших джерел даних.

Очікувані результати. Після виконання практикуму здобувач повинен:

1. розуміти основні проблеми інтеграції різнорідних даних;
2. уміти аналізувати структуру, формати та семантику різних джерел даних;
3. визначати структурні та семантичні невідповідності між джерелами;
4. розробляти схеми зіставлення та узгодження даних;
5. застосовувати методи ETL/ELT;
6. виконувати очищення, трансформацію та гармонізацію даних;
7. інтегрувати структуровані та напівструктуровані дані;
8. реалізовувати відтворювані pipelines інтеграції;

9. оцінювати повноту, узгодженість та якість результатів інтеграції;
10. розуміти принципи data federation та віртуалізації даних;
11. порівнювати альтернативні підходи до інтеграції;
12. проводити експериментальне дослідження та обґрунтовувати вибір технології інтеграції.

Літературні джерела:

1. Putrama I. M., Martinek P. Heterogeneous data integration: Challenges and opportunities // Data in Brief. – 2024. – Vol. 56. – Art. 110853. – DOI: 10.1016/j.dib.2024.110853. – URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11402636

2. Masmoudi M., Ben Abdallah Ben Lamine S., Karray M. H., Archimede B., Baazaoui Zghal H. Semantic Data Integration and Querying: A Survey and Challenges // ACM Computing Surveys. – 2024. Vol. 56, No. 8. – Art. 209. – P. 1–35. – DOI: 10.1145/3653317. – URL: https://dl.acm.org/doi/10.1145/3653317

3. Cheng C., Messerschmidt L., Bravo I., Waldbauer M., Bhavikatti R., Schenk C., Grujic V., Model T., Kubinec R., Barceló J. A General Primer for Data Harmonization // Scientific Data. – 2024. – Vol. 11. Art. 152. – DOI: 10.1038/s41597-024-02956-3. – URL: https://www.nature.com/articles/s41597-024-02956-3

Комп'ютерний практикум 4.

Розподілене та масштабоване оброблення великих масивів даних

Зміст заняття:

Дослідження принципів розподіленого та паралельного оброблення великих масивів даних. Аналіз архітектури розподілених систем зберігання та оброблення даних. Дослідження технологій Apache Hadoop та Apache Spark, принципів розподілу даних між обчислювальними вузлами, планування та виконання паралельних задач. Реалізація операцій фільтрації, агрегації, сортування та з'єднання великих наборів даних у розподіленому середовищі. Дослідження впливу обсягу даних, кількості обчислювальних вузлів та параметрів конфігурації на продуктивність системи. Експериментальне оцінювання часу виконання, масштабованості та використання обчислювальних ресурсів.

Основні завдання практикуму:

1. Визначити предметну область та сформулювати задачу оброблення великого масиву даних.
2. Обрати та підготувати набір даних достатнього обсягу для проведення експерименту.
3. Проаналізувати архітектуру та принципи функціонування вибраної технології розподіленого оброблення.
4, Організувати зберігання набору даних у розподіленій файловій системі або відповідному розподіленому сховищі.
5. Реалізувати операції розподіленого оброблення даних.
6. Дослідити механізми розподілу даних і паралельного виконання обчислювальних задач.
7. Реалізувати та виконати операції фільтрації, агрегації, сортування та/або з'єднання великих наборів даних.
8. Провести експерименти з різним обсягом вхідних даних.
9. Дослідити залежність часу виконання від кількості обчислювальних ресурсів або вузлів.
10. Оцінити масштабованість розробленого рішення.
11. Дослідити вплив параметрів конфігурації системи на продуктивність оброблення.
12. Проаналізувати використання процесорних, оперативних, дискових та мережевих ресурсів.
13. За можливості порівняти два підходи або технології розподіленого оброблення даних.
14. Узагальнити результати експериментів та сформулювати рекомендації щодо вибору технології й параметрів її налаштування.

Очікувані результати. Після виконання практикуму здобувач повинен:

1. розуміти принципи організації розподіленого та паралельного оброблення даних;
2. знати основні компоненти та принципи функціонування сучасних платформ Big Data;
3. уміти організовувати розподілене зберігання великих масивів даних;
4. застосовувати технології Apache Hadoop та/або Apache Spark для розподіленого оброблення;
5. реалізовувати паралельні операції над великими наборами даних;
6. розуміти механізми розподілу даних, виконання задач та обміну даними між обчислювальними вузлами;
7. оцінювати часову та ресурсну складність розподілених обчислень;
8. досліджувати горизонтальну та вертикальну масштабованість;
9. аналізувати вплив конфігурації обчислювального середовища на продуктивність;
10. проводити експериментальне порівняння технологій та алгоритмів розподіленого оброблення;
11. інтерпретувати результати обчислювальних експериментів;
12. обґрунтовувати вибір технологій та параметрів розподіленого оброблення для конкретної задачі.

Літературні джерела:

1. Curry E., Auer S., Berre A. J., Metzger A., Perez M. S., Zillner S., eds. Technologies and Applications for Big Data Value. – Cham: Springer, 2022. – XXIV. – 544 p. – DOI: 10.1007/978-3-030-78307-5. – URL: https://link.springer.com/book/10.1007/978-3-030-78307-5

2. Bast H., Korzen C., Meyer U., Penschuck M., eds. Algorithms for Big Data: DFG Priority Program 1736. – Cham: Springer, 2022. – XIV. – 285 p. – DOI: 10.1007/978-3-031-21534-6. – URL: https://link.springer.com/book/10.1007/978-3-031-21534-6

3. Wang X., Carey M. J., Tsotras V. J. Subscribing to big data at scale // Distributed and Parallel Databases. – 2022. – Vol. 40, No. 2–3. – P. 475–520. – DOI: 10.1007/s10619-022-07406-w. – URL: https://link.springer.com/article/10.1007/s10619-022-07406-w

4. Zhang D., Dai Z.-Y., Sun X.-P., Wu X.-T., Li H., Tang L., He J.-H. A distributed data processing scheme based on Hadoop for synchrotron radiation experiments // Journal of Synchrotron Radiation. – 2024. – Vol. 31, Pt. 3. – P. 635–645. – DOI: 10.1107/S1600577524002637. – URL: https://journals.iucr.org/s/issues/2024/03/00/wz5039/index.html

Комп’ютерний практикум 5.

Статистичний, багатовимірний та інтелектуальний аналіз даних

Зміст заняття: дослідження методів статистичного та інтелектуального аналізу даних із використанням програмних засобів. Проведення описового та дослідницького аналізу даних, визначення статистичних характеристик і взаємозв’язків між змінними. Дослідження структури багатовимірних даних та застосування методів зниження розмірності, зокрема аналізу головних компонент. Застосування методів кластеризації та класифікації для виявлення прихованих закономірностей і структур у даних. Побудова та оцінювання моделей машинного навчання для задач класифікації та прогнозування. Проведення порівняльного обчислювального експерименту, оцінювання якості отриманих моделей та інтерпретація результатів.

Основні завдання практикуму:

1. Обрати предметну область та сформулювати дослідницьку задачу аналізу даних.
2. Обрати та підготувати набір даних, що містить достатню кількість змінних і спостережень.
3. Провести описовий та дослідницький статистичний аналіз даних.
4. Визначити статистичні характеристики змінних та дослідити взаємозв’язки між ними.
5. Проаналізувати кореляційну структуру багатовимірного набору даних.
6. Обґрунтувати та застосувати метод зниження розмірності, зокрема метод головних компонент (PCA) або альтернативний метод.
7. Дослідити структуру даних із використанням методів кластерного аналізу.

8. Визначити оптимальну або обґрунтовану кількість кластерів та оцінити якість кластеризації.

9. Побудувати одну або декілька моделей класифікації та/або прогнозування.

10. Провести навчання моделей та оцінити їх якість за відповідними статистичними метриками.

11. Порівняти результати декількох методів аналізу.

12. Дослідити вплив попереднього перетворення даних та/або зниження розмірності на результати аналізу.

13. Виконати графічну візуалізацію структури даних та результатів кластеризації, класифікації або прогнозування.

14. Провести інтерпретацію отриманих результатів та сформулювати науково обґрунтовані висновки.

Очікувані результати. Після виконання практикуму здобувач повинен:

1. уміти проводити описовий та дослідницький статистичний аналіз даних;
2. визначати та інтерпретувати статистичні характеристики багатовимірних даних;
3. досліджувати кореляційні та інші статистичні взаємозв'язки між змінними;
4. застосовувати методи зниження розмірності багатовимірних даних;
5. використовувати методи кластерного аналізу для виявлення прихованої структури даних;
6. застосовувати методи класифікації та прогнозування;
7. будувати та оцінювати моделі машинного навчання;
8. обґрунтовано вибирати методи аналізу відповідно до властивостей даних та поставленої задачі;
9. використовувати відповідні метрики для порівняння якості моделей;
10. проводити обчислювальні експерименти та статистично інтерпретувати їх результати;
11. візуалізувати результати багатовимірного та інтелектуального аналізу;
12. порівнювати альтернативні методи та формулювати науково обґрунтовані висновки.

Літературні джерела:

1. Mathai A. M., Provost S. B., Haubold H. J. Multivariate Statistical Analysis in the Real and Complex Domains. – Cham: Springer, 2022. – XXVII, 921 p. – DOI: 10.1007/978-3-030-95864-0. – URL: https://link.springer.com/book/10.1007/978-3-030-95864-0

2. Montesinos López O. A., Montesinos López A., Crossa J. Multivariate Statistical Machine Learning Methods for Genomic Prediction. – Cham: Springer, 2022. – XXIV, 691 p. – DOI: 10.1007/978-3-030-89010-0. – URL: https://link.springer.com/book/10.1007/978-3-030-89010-0

3. Brazdil P., van Rijn J. N., Soares C., Vanschoren J. Metalearning: Applications to Automated Machine Learning and Data Mining. 2nd ed. – Cham: Springer, 2022. – XII, 346 p. – DOI: 10.1007/978-3-030-67024-5. – URL: https://link.springer.com/book/10.1007/978-3-030-67024-5

4. Brito P., Dias J. G., Lausen B., Montanari A., Nugent R., eds. Classification and Data Science in the Digital Age. – Cham: Springer, 2023. – XIV, 416 p. – DOI: 10.1007/978-3-031-09034-9. – URL: https://link.springer.com/book/10.1007/978-3-031-09034-9

Комп'ютерний практикум 6.

Комплексне оброблення, візуалізація та дослідницький аналіз даних.

Зміст заняття: комплексне дослідження повного циклу оброблення та аналізу даних для розв'язання науково-дослідницької задачі. Формування та реалізація відтворюваного data science pipeline, що включає отримання, підготовку, трансформацію та аналіз даних. Застосування статистичних, багатовимірних та інтелектуальних методів дослідження даних. Побудова інформативних графічних представлень та інтерактивних візуалізацій для виявлення закономірностей, взаємозв'язків, тенденцій та аномалій. Інтерпретація результатів комплексного аналізу, оцінювання їх достовірності та відтворюваності. Підготовка результатів обчислювального експерименту у формі, придатній для використання в науковому дослідженні.

Основні завдання практикуму:

1. Сформулювати науково-дослідницьку задачу та визначити дослідницькі питання.
2. Обрати та обґрунтувати набір даних, необхідний для проведення дослідження.
3. Сформувати послідовність операцій комплексного оброблення даних.
4. Виконати підготовку, очищення, трансформацію та необхідне перетворення даних.
5. Провести описовий та дослідницький статистичний аналіз.
6. Дослідити багатовимірну структуру даних та взаємозв'язки між досліджуваними ознаками.
7. Застосувати методи зниження розмірності, кластеризації, класифікації, прогнозування або виявлення аномалій відповідно до поставленої задачі.
8. Розробити систему візуалізації результатів оброблення та аналізу даних.
9. Застосувати відповідні графічні та інтерактивні способи представлення багатовимірних, часових та інших складних даних.
10. Порівняти результати декількох методів оброблення та аналізу.
11. Дослідити вплив параметрів обраних методів на отримані результати.
12. Оцінити достовірність, стійкість та відтворюваність результатів обчислювального експерименту.
13. Оформити результати дослідження у вигляді таблиць, графіків, візуалізацій та аналітичних висновків.
14. Сформулювати науково обґрунтовані висновки та визначити можливості використання отриманих результатів у подальшому науковому дослідженні.

Очікувані результати. Після виконання практикуму здобувач повинен:
1. уміти організовувати повний цикл оброблення та аналізу даних;
2. розробляти відтворювані pipelines дослідження даних;
3. застосовувати статистичні, багатовимірні та інтелектуальні методи аналізу;
4. визначати закономірності, взаємозв'язки, тенденції та аномалії у даних;
5. обґрунтовано вибирати методи аналізу відповідно до структури та властивостей даних;
6. застосовувати методи зниження розмірності та візуального дослідження багатовимірних даних;
7. створювати інформативні статичні та інтерактивні візуалізації;
8. використовувати візуалізацію як інструмент дослідницького аналізу, а не лише як засіб графічного представлення результатів;
9. порівнювати результати застосування різних методів;
10. оцінювати достовірність, стійкість та відтворюваність отриманих результатів;
11. представляти результати обчислювального експерименту у формі, придатній для наукової публікації або дисертаційного дослідження;
12. формулювати наукові висновки на основі результатів комплексного аналізу даних.

Літературні джерела:
1. Aigner W., Miksch S., Schumann H., Tominski C. Visualization of Time-Oriented Data. 2nd ed. – London: Springer, 2023. – XIX, 441 p. – DOI: 10.1007/978-1-4471-7527-8. – URL: https://link.springer.com/book/10.1007/978-1-4471-7527-8
2. Barth L. S., Fahimi H., Joharinad P., Jost J., Keck J. Data Visualization with Category Theory and Geometry: With a Critical Analysis and Refinement of UMAP. – Cham: Springer, 2025. – XIII, 272 p. – DOI: 10.1007/978-3-031-97973-6. – URL: https://link.springer.com/book/10.1007/978-3-031-97973-6
3. Saqr M., López-Pernas S., eds. Learning Analytics Methods and Tutorials: A Practical Guide Using R. – Cham: Springer, 2024. – XXXIV, 736 p. – DOI: 10.1007/978-3-031-54464-4. – URL: https://link.springer.com/book/10.1007/978-3-031-54464-4
4. Wagner A. S., Waite L. K., Wierzba M. et al. FAIRly big: A framework for computationally reproducible processing of large-scale data // Scientific Data. – 2022. – Vol. 9. – Art. 80. – 17 p. – DOI: 10.1038/s41597-022-01163-2. – URL: https://www.nature.com/articles/s41597-022-01163-2

Додаток 3. Перелік питань до іспиту з дисципліни

1. Що таке інформаційно-комунікаційна система та які основні функції вона виконує щодо даних?
2. Які основні типи даних використовуються в сучасних інформаційно-комунікаційних системах?
3. У чому полягають основні відмінності структурованих, напівструктурованих та неструктурованих даних?
4. Які характеристики сучасних даних визначають вимоги до технологій їх оброблення?
5. Як архітектура інформаційно-комунікаційної системи впливає на процеси збирання, зберігання та оброблення даних?
6. Які основні етапи життєвого циклу даних в інформаційно-комунікаційній системі?
7. Які проблеми виникають під час оброблення даних великого обсягу, високої швидкості надходження та складної структури?
8. Що таке модель даних і яке її призначення в інформаційно-комунікаційній системі?
9. Які основні типи моделей даних використовуються в сучасних інформаційних системах?
10. У чому полягають особливості реляційної моделі даних?
11. Які переваги та обмеження має документна модель представлення даних?
12. У чому полягають особливості графової моделі даних?
13. У яких випадках доцільно використовувати реляційну, документну, ключ-значення або графову модель?
14. Що таке концептуальна, логічна та фізична моделі даних?
15. Які проблеми виникають під час перетворення даних між різними моделями представлення?
16. Які основні методи та джерела збирання даних використовуються в інформаційно-комунікаційних системах?
17. Які фактори необхідно враховувати під час вибору джерел даних для наукового дослідження?
18. Що таке профілювання даних та яке його призначення?
19. Які основні операції первинного оброблення даних?
20. Які методи використовуються для оброблення пропущених значень?
21. Які підходи застосовуються для виявлення та оброблення аномальних значень?
22. У чому полягають відмінності між очищенням, нормалізацією, стандартизацією та трансформацією даних?
23. Як вибір методу первинного оброблення може впливати на результати подальшого аналізу?
24. Що розуміють під якістю даних?
25. Які основні характеристики якості даних необхідно оцінювати?
26. Як оцінюються повнота, узгодженість та достовірність даних?
27. Які причини виникнення помилок у даних?
28. Які методи використовуються для виявлення помилок у даних?
29. У чому полягає проблема невизначеності даних?
30. Як невизначеність та низька якість вихідних даних можуть впливати на результати статистичного та інтелектуального аналізу?
31. Як побудувати систему кількісного оцінювання якості даних до та після їх очищення?
32. Що таке система управління базами даних та які основні функції вона виконує?
33. Які основні етапи проєктування бази даних?
34. Що таке ER-модель та які її основні компоненти?
35. Яке призначення первинних та зовнішніх ключів?
36. Що таке нормалізація бази даних і для чого вона застосовується?
37. Які переваги та недоліки надмірної нормалізації?
38. Яке призначення індексів у системах управління базами даних?
39. Як структура бази даних та індексація впливають на продуктивність виконання запитів?
40. У чому полягають основні відмінності реляційних та NoSQL-СУБД?

41. Що таке інтеграція даних і які основні проблеми виникають під час інтеграції різнорідних джерел?

42. У чому полягають структурні та семантичні відмінності між джерелами даних?

43. Що таке ETL та які основні етапи цього процесу?

44. У чому полягає відмінність між ETL та ELT?

45. Що таке зіставлення схем та для чого воно використовується?

46. Що розуміють під семантичною інтеграцією даних?

47. Що таке гармонізація даних та які її основні завдання?

48. У чому полягає консолідація даних і як оцінити якість її результатів?

49. Що таке розподілене оброблення даних і чим воно відрізняється від централізованого?

50. Які основні переваги паралельного оброблення великих масивів даних?

51. Які архітектурні принципи лежать в основі розподілених систем оброблення даних?

52. Які функції виконує розподілена файлова система в системах Big Data?

53. Які основні принципи функціонування Apache Hadoop?

54. Які основні принципи функціонування Apache Spark?

55. У чому полягають відмінності між Hadoop MapReduce та Spark?

56. Як кількість обчислювальних вузлів впливає на продуктивність розподіленого оброблення?

57. Що таке горизонтальна масштабованість і як її експериментально оцінити?

58. Які основні завдання статистичного аналізу даних?

59. Які статистичні характеристики використовуються для опису наборів даних?

60. У чому полягає дослідницький аналіз даних?

61. Як досліджувати статистичні взаємозв’язки між змінними?

62. Що таке кореляційний аналіз та які його основні обмеження?

63. Які основні методи кластерного аналізу використовуються для виявлення структури даних?

64. Які критерії можна використовувати для оцінювання якості кластеризації?

65. У чому полягає задача класифікації та які основні підходи до її розв’язання?

66. Як здійснюється порівняння якості різних моделей статистичного та інтелектуального аналізу?

67. Що розуміють під Big Data та які характеристики визначають необхідність спеціальних технологій їх оброблення?

68. Які основні проблеми виникають під час оброблення великих масивів даних?

69. Які технології використовуються для масштабованого зберігання та оброблення Big Data?

70. У чому полягають особливості оброблення складноструктурованих даних?

71. Як обсяг даних впливає на вибір алгоритму та технології їх оброблення?

72. Які критерії використовуються для оцінювання масштабованості системи оброблення даних?

73. Як оцінити залежність продуктивності системи від обсягу даних та доступних обчислювальних ресурсів?

74. У чому полягають основні відмінності потокових та статичних даних?

75. Які особливості необхідно враховувати під час оброблення потокових даних у реальному або близькому до реального часу?

76. Що таке потокове оброблення даних та які основні операції можуть виконуватися над потоком?

77. Які методи використовуються для виявлення аномалій у потокових даних?

78. Як забезпечується масштабованість систем потокового оброблення?

79. Що таке багатовимірні дані та які проблеми виникають під час їх аналізу?

80. У чому полягає проблема «прокляття розмірності»?

81. Яке призначення методів зниження розмірності даних?

82. У чому полягає метод головних компонент (PCA) та які задачі він дає змогу розв’язувати?

83. Які переваги та обмеження має застосування PCA для аналізу багатовимірних даних?

84. Які сучасні методи нелінійного зниження розмірності можуть застосовуватися для дослідження складних даних?

85. Як вибір способу представлення багатовимірних даних впливає на результати їх аналізу?

86. Які основні завдання візуалізації даних у науковому дослідженні?

87. Як обрати тип графічного представлення відповідно до структури та типу даних?

88. Які особливості візуалізації багатовимірних даних?

89. Які методи використовуються для візуалізації часових та потокових даних?

90. У чому полягає відмінність між звичайною графічною візуалізацією та visual analytics?

91. Які переваги надають інтерактивні засоби візуалізації під час дослідницького аналізу?

92. Що розуміють під інтелектуалізацією процесів оброблення та аналізу даних?

93. Які методи машинного навчання можуть використовуватися для автоматизації аналізу даних?

94. Як здійснюється автоматизований вибір алгоритмів та гіперпараметрів моделей?

95. Які переваги та ризики автоматизації процесів аналізу даних?

96. Як оцінити інтерпретованість та надійність результатів інтелектуального аналізу?

97. Які етапи повинен містити відтворюваний pipeline оброблення та аналізу даних у науковому дослідженні?

98. Які принципи необхідно дотримуватися для забезпечення відтворюваності обчислювального експерименту?

99. Як обґрунтувати вибір методів оброблення та аналізу даних у дисертаційному дослідженні?

100. Як побудувати комплексну методику оброблення, аналізу та візуалізації даних для розв’язання конкретної наукової задачі?