Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Кафедра інформатики
та програмної інженерії

КУРСОВА РОБОТА З ОБРОБЛЕННЯ НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***другий (магістерський)***
Галузь знань	*Інформаційні технології*
Спеціальність	*121 Інженерія програмного забезпечення*
Освітня програма	*Інженерія програмного забезпечення інформаційних систем*
Статус дисципліни	*основна*
Форма навчання	*очна(денна)/дистанційна/змішана*
Рік підготовки, семестр	*1-й курс, осінній семестр*
Обсяг дисципліни	*1 кредити, 30 годин*
Семестровий контроль/ контрольні заходи	*Залік (захист курсової роботи)*
Розклад занять	*перший семестр*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	доцент кафедри ІПІ Олійник Ю.О., oliyura@gmail.com асистент кафедри ІПІ Зарічковий О.А. alexkirnas1998@gmail.com
Розміщення курсу	https://ecampus.kpi.ua/ https://classroom.google.com/c/NzExNzA5MzMxODg0?cjc=nc5wtj7

Програма навчальної дисципліни

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

Мета вивчення дисципліни – набуття ключових фахових компетентностей, теоретичних знань і практичних навичок аналізу, аргументування, прийняття рішень при розв'язанні задач та практичних проблем оброблення та аналізу надвеликих масивів даних.

Предметом вивчення дисципліни є технології, моделі, архітектура розподілених сховищ даних; процеси обробки та аналізу надвеликих масивів даних.

Завдання вивчення дисципліни: – оволодіння основними поняттями обробки та аналізу надвеликих масивів даних; – ознайомлення з новітніми підходами створення розподілених сховищ даних; – набуття практичних навичок обробки та аналізу надвеликих масивів даних для вирішення задач підтримки прийняття рішень.

Навчальна дисципліна покликана допомогти студенту отримати:

- вивчення сучасних концепцій та підходів до оброблення надвеликих масивів даних та створення сховищ даних;
- уміння вільно орієнтуватися на сучасному світі розподілених сховищ даних; проектувати та створювати розподілені сховища даних, застосовувати сучасні методи та технології обробки та аналізу надвеликих масивів даних**.**
- здатність використовувати можливості сучасних засобів та технологій обробки потоків даних.

КОМПЕТЕНТНОСТІ

Спеціальні (фахові, предметні) компетентності

- ФК11 - Здатність до аналізу, проектування та розробки нових та використання існуючих систем зберігання та обробки надвеликих масивів даних

ПРОГРАМНІ РЕЗУЛЬТАТИ НАВЧАННЯ

- ПРН20 – Розробляти, реалізувати та застосовувати різні методи інтелектуального аналізу даних до Big Data, формулювати алгоритми обробки в парадигмі Map Reduce, обирати відповідну технологію зберігання і оброблення надвеликих даних, використовувати сучасні високонавантажені системи зберігання та оброблення великих даних.

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

Пререквізити:

- Бази даних.
- Аналіз даних.
- Проєктування інформаційних систем.
- Теорія алгоритмів.

Знання, одержані студентами при вивченні дисципліни, використовуються у магістерських дослідженнях.

3. Зміст навчальної дисципліни

Основні типові етапи виконання курсової роботи:

- постановка задачі;
- розробка процесу обробки даних;
- розробка ETL процесів;
- розробка структури БД;
- опис методів обробки даних;
- дослідження ефективності методів обробки даних
- оформлення пояснювальної записки;
- захист курсової роботи.

Етапи виконання можуть відрізнятись в залежності від теми.

Тема курсової роботи повинна корелюватись з задачами магістерської дисертації. Задачі повинні вирішені за допомогою технологій BigData.

Приклади тем:

- Сегментація вподобань користувачів на основі технології Apache Spark.
- Класифікація наукових текстів за кодом УДК.
- Розпізнавання ієрогліфів японської мови в потоці даних.

4. Навчальні матеріали та ресурси

Базова

1. Harness the Power of Big Data. Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles. ISBN: 978-0-07180818-7
2. Big Data Beyond the Hype. A Guide to Conversations for Today's Data Center. Paul Zikopoulos, Dirk deRoos, Christopher Bienko, Rick Buglio, Marc Andrews. ISBN: 978-0-07-184466-6
3. Damji, Jules S., et al. Learning Spark. O'Reilly Media, 2020
4. Needham, Mark, and Amy E. Hodler. Graph algorithms: practical examples in Apache Spark and Neo4j. O'Reilly Media, 2019.
5. Chuck Lam. Hadoop in Action. - WILEY, 2011. - 336с: ил. ISBN 978-8177228137.
6. Edward Capriolo, Dean Wampler, and Jason Rutherglen. Programming Hive. ISBN: 978-1-449-31933-5
7. Ralph Kimball, Margy Ross, Warren Thornthwaite, Joy Mundy, Bob Becker. The Data Warehouse Lifecycle Toolkit, 2nd Edition. Wiley Publishing, Inc, 2008.672с.:ил. ISBN: ISBN 978-0-470-14977-5.
8. Ralph Kimball, Margy Ross, Warren Thornthwaite (Contributions by), Joy Mundy (Contributions by), Bob Becker (Contributions by). Relentlessly Practical Tools for Data Warehousing and Business Intelligence. Wiley Publishing, Inc, 2010. 744 .:ил. ISBN: 978-0-470-56310-6.
9. Пасічник В.В., Шаховська Н.Б. Сховища даних: Навчальний посібник. - Львів: "Магнолія 2006", 2008.-496 С. ISBN 978-966-2025-18-7.
10. Daniel Linstedt, Michael Olschimke Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufmann Waltham, MA 02451, USA. ISBN: 978-0-12-802510-9. P.661.

Допоміжна література

1. Thomas C. Hammergren and Alan R. Simon Data Warehousing For Dummies. Wiley Publishing, Inc. Hoboken, NJ, USA. ISBN: 978-0-470-40747-9. P.388
2. Vaisman, A., Zimányi, E. Data warehouse systems. Data-Centric Systems and Applications. 2014.
3. Додаткові матеріали до курсу «Оброблення надвеликих масивів даних». Електронний ресурс: https://drive.google.com/drive/u/1/folders/159vqraxrOrrZej2CqaOJzxe0wAuZyWcr

Інформаційні ресурси

- https://ecampus.kpi.ua/

Для викладання дисципліни необхідні наступні ресурси:

- В лекційній аудиторії має бути комп'ютер з доступом до мережі Інтернет, а також проектор;
- В локальній мережі або в хмарному середовищі мають бути встановлені :

  Apache Hadoop, Apache Hive, Apache Spark версії 3 та вище, які розповсюджуються по безкоштовній ліцензії.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

№ п/п	Освітні компоненти (навчальні дисципліни, курсові проекти (роботи), практики, кваліфікаційна робота)	Назва кафедри	К-ть здобувачів, які вибрали дисципліну		Обсяг дисциплін		Аудиторні години								Самостійна робота студентів	Контрольні заходи та їх розподіл за семестрами								Розподіл аудиторних годин на тиждень за курсами семестрами							
								В тому числі																1 курс						
								Лекції			Практичні (комп.практ)		Лабораторні											ІП-11мн(5+0)						
																								1 семестр				2 семестр		
																								18 тижнів				18 тижнів		
																								у тому числі				у тому чис		
			Б	К	Кредитів ЄСТС	Годин	Всього	за НП	з урахуван. інд. занять	за НП	з урахуван. інд. занять	за НП	з урахуван. інд. занять	Індивідуальні заняття		Екзамени	Заліки	Модульн.(темат), контр.роботи	Курсові проекти	Курсові роботи	РГР,РР,ГР	ДКР	Реферати	Всього	Лекцій	Практичні	Лабораторні	Всього	Лекцій	Практичні
1	2	3			4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
6	Оброблення надвеликих масивів даних	Інформатики та програмної інженерії			5	150	72	36				36			78	1		1					4	2		2			

МЕТОДИ НАВЧАННЯ:

Лекційні заняття проходять з використанням мультимедійних технологій та наступних методів:

- Пояснювально-ілюстративного методу Послідовна та логічно ув'язана подача матеріалу надає уявлення та знання у його логічної цілісності.

- Метод проблемного викладу надає уяву та методи отримання нових знань та фактів з використанням вже відомих фактів та тверджень.

- Інтерактивний метод під час лекційних занять використовується для встановлення діалогу з аудиторією.

Комп'ютерні практикуми проходять з використанням наступних методів:

1) репродуктивного методу, завдяки якому студенти закріплюють вивчений теоретичний матеріал та навчаються використовувати його в конкретних задачах

2) проблемного методу, при застосуванні якого студенти залучаються до обговорення та вирішення задач, пов'язаних з новітніми інформаційними технологіями аналітичної обробки інформації

Самостійна робота з можливістю особистих консультацій з викладачем.

Тематичний план проведення лекційних занять

Лекція 1. Задачі оброблення надвеликих масивів даних	2

Лекція 2. Архітектура та принципи обробки даних в Apache Hadoop. Концепція MapReduce.	2
Лекція 3. Архітектура та принципи обробки даних в Apache Hive	2
Лекція 4-5. Архітектура та принципи обробки даних в Apache Spark	4
Лекція 6-7. Обробка графів в Apache Spark	4
Лекція 8. Класифікація даних в Apache Spark	2
Лекція 9. Кластеризація даних в Apache Spark	2
Лекції 10-11. Створення рекомендаційних систем на основі Apache Spark	4
Лекції 12-13. Обробка текстової інформації	4
Лекція 14-15. Обробка потоків даних	4
Лекція 16-18. Кращі практики створення рішень з обробки надвеликих масивів даних	6

6. Самостійна робота студента

Самостійна робота

Матеріали для самостійної роботи розміщені на гугл диску викладача.

Тематика самостійної роботи
Тема 1. Характеристики технологій обробки надвеликих масивів даних
Тема 2. Виконання операцій join в концепції MapReduce
Тема 3. Типи моделей сховищ даних
Тема 4. ETL процеси в технологіях обробки надвеликих масивів даних
Тема 5. Бібліотеки ML, MLib в Apache Spark
Тема 6. Графові алгоритми в Neo4j
Тема 7. Методи визначення аномалій для надвеликих масивів даних
Тема 8. Фреймворк TEZ
Тема 9. Технології управління ресурсами в задачах обробки надвеликих масивів даних
Тема 10. Кращі практики рішень з обробки надвеликих масивів даних

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Форми організації освітнього процесу, види навчальних занять і оцінювання результатів навчання регламентуються Положенням про організацію освітнього процесу в Національному технічному університеті України «Київському політехнічному інституті імені Ігоря Сікорського».

Політика виставлення оцінок: кожна оцінка виставляється відповідно до розроблених викладачем та заздалегідь оголошених студентам критеріїв, а також мотивується в індивідуальному порядку на вимогу студента.

Політика академічної поведінки та доброчесності: конфліктні ситуації мають відкрито обговорюватись в академічних групах з викладачем, необхідно бути взаємно толерантним, поважати думку іншого. Плагіат та інші форми нечесної роботи неприпустимі. Всі індивідуальні завдання студент має виконати самостійно із використанням рекомендованої літератури й отриманих знань та навичок. Цитування

в письмових роботах допускається тільки із відповідним посиланням на авторський текст. Недопустимі підказки і списування у ході захисту практикумів, на контрольних роботах, на іспиті.

Норми академічної етики: дисциплінованість; дотримання субординації; чесність; відповідальність; робота в аудиторії з відключеними мобільними телефонами. Повага один до одного дає можливість ефективніше досягати поставлених командних результатів. При виконанні практикумів студент може користуватися ноутбуками. Проте під час лекційних занять та обговорення завдань практикумів не слід використовувати ноутбуки, смартфони, планшети чи комп'ютери. Це відволікає викладача і студентів групи та перешкоджає навчальному процесу. Якщо ви використовуєте свій ноутбук чи телефон для аудіо- чи відеозапису, необхідно заздалегідь отримати дозвіл викладача.

Дотримання академічної доброчесності студентів й викладачів регламентується кодекс честі Національного технічного університету України «Київський політехнічний інститут», положення про організацію освітнього процесу в КПІ ім. Ігоря Сікорського. За порушення принципів академічної доброчесності, зокрема плагіат практикумів, студент втрачає всі бали за даний практикум.

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

Поточний контроль: перевірка дотримання графіку виконання календарного плану курсової роботи.

Календарний контроль: проводиться двічі за семестр як моніторинг виконання календарного плану курсової роботи

Семестровий контроль: залік.

Керівник здійснює контроль за ходом виконання студентом курсової роботи, надає йому необхідну консультативну допомогу. Протягом семестру студент демонструє викладачу поточні результати роботи над проектом. Після перевірки роботи викладач призначає день, час і місце захисту.
Напередодні захисту студенту необхідно повторити теоретичний матеріал, що стосується роботи, та переглянути безпосередньо її зміст. Захист КР проводиться у формі співбесіди зі з'ясуванням усіх питань, що виникли у керівника під час перевірки роботи.

За результатами захисту, у відповідності до критеріїв оцінювання викладачі виставляють студенту оцінку.

На оцінку за КР впливають:

- якість розробленого програмного забезпечення;
- якість розробленої програмної документації;
- компетентність та загальна ерудиція студента при відповідях на запитання під час захисту;
- ступінь виконання графіку підготовки курсової роботи.

Якщо студент подав на захист не самостійно виконану роботу, про що свідчить його некомпетентність у рішеннях та матеріалах роботи, КР до захисту не допускається, що супроводжується записом "не допущений" у екзаменаційній відомості. Такий самий запис робиться у випадку, якщо КР не завершена на час захисту. В цих випадках запис "не допущений" еквівалентний отриманню оцінки "незадовільно".

Рейтингова оцінка з курсової роботи має дві складові: виконання курсової роботи та її захист. Перша (стартова) складова характеризує роботу студента з курсового проектування та її результат - якість пояснювальної записки та розробленого програмного забезпечення; друга складова характеризує якість захисту студентом курсової роботи.

Розмір шкали першої складової (*r1*) дорівнює 70 балів, а другої складової *(r2)*- 30 балів.

Система рейтингових балів.

Стартова складова виконання курсової роботи (r1):

- ступінь розкриття теоретичних аспектів теми та коректність використання понятійного апарату – до 5 балів;
- повнота та коректність предметної області та задачі обробки даних – до 10 балів;
- якість написання та оформлення програмного коду та моделі бази даних та ETL процесів – до 20 балів;
- складність обраних методів обробки – до 20 балів;
- якість оформлення пояснювальної записки з урахуванням виконання вимог нормативних документів – до 15 балів;
- не своєчасність виконання основних етапів графіку підготовки курсової роботи – -1 бал за кожен день запізнення графіку.

Складова захисту курсової роботи (r2):

- ступінь володіння теоретичним матеріалом – до 5 балів;
- ступінь володіння кодом програми в цілому – до 15 балів;
- вміння внести зміни у програмний код – до 10 балів.

Узагальнені критерії оцінювання захисту курсової роботи та деталізовані бальні шкали наведено в наступній таблиці

Параметри оцінювання	*Діапазон балів*	*Критерії оцінювання за бальною шкалою*
Ступінь володіння теоретичним матеріалом	*0-5*	*0 – студент на дав відповіді на теоретичні питання*
		1-4 – відповідь, на одне чи два теоретичних питання
		5 - відповідь, на усі теоретичні питання
Ступінь володіння кодом програми в цілому	*0-15*	*0 – студент не володіє кодом програми*
		1-3 – студент володіє лише власним кодом
		4-14 – студент частково володіє кодом своїх колег по курсовій роботі
		15 – студент повністю володіє кодом своїх колег по курсовій роботі
Вміння внести зміни у програмний код	*0-10*	*0 – студент не може виконати додаткові завдання, пов'язані із внесенням змін у програмний код*
		1-9 – студент частково може виконати додаткові завдання, пов'язані із внесенням змін у програмний код
		10 – студент грамотно виконує додаткові завдання, пов'язані із внесенням змін у програмний код

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

R = r1+ r2	Оцінка ECTS	Національна оцінка
95… 100	A	відмінно
85 … 94	B	добре

75 … 84	C	
65 … 74	D	задовільно
60 … 64	E	
Менше 60	Fx	незадовільно
Курсовий проект не допущено до захисту	F	не допущено

9. Додаткова інформація з дисципліни (освітнього компонента)

Перелік питань, які виносяться на семестровий контроль розміщений в системі «Електронний кампус КПІ» або на платформі Classroom компанії Google.

Робочу програму навчальної дисципліни (силабус):

Складено доцент кафедри ІПІ, Олійник Ю.О.

Ухвалено кафедрою ІПІ (протокол №16 від 29.05.2024 р.)

Погоджено Методичною комісією факультету (протокол № 10 від 21.06.2024 р.)