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

Кафедра інформатики
та програмної інженерії

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	*заочна*
Рік підготовки, семестр	*1-й курс, осінній семестр*
Обсяг дисципліни	*5 кредити, 150 годин (10 годин – Лекції, 10 годин - Лабораторні роботи, 130 годин – СРС)*
Семестровий контроль/ контрольні заходи	*іспит*
Розклад занять	*перший семестр*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: доцент кафедри ІПІ Олійник Ю.О., oliyura@gmail.com Комп'ютерний практикум: доцент кафедри ІПІ Олійник Ю.О. oliyura@gmail.com
Розміщення курсу	https://ecampus.kpi.ua/

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 дипломному проєктуванні.

3. Зміст навчальної дисципліни

Тема 1. Обробка даних та побудова розподілених сховищ в Hadoop та Hive
Тема 2. Архітектура Apache Spark
Тема 3. Графові алгоритми в Apache Spark
Тема 4. Інтелектуальний аналіз даних в Apache Spark
Тема 5. Обробка потоків даних
Тема 6. Обробка текстових 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 дисц., курс. пр.(роб.), практ., кваліф. роб.)	Кафедра	К-ть здобув.		Обсяг дисциплін		Аудиторні години								СРС	Контрольні заходи								Розподіл аудиторних годин на тиждень за курсами і семестрами							
								Лекції		Практ. (комп. прк)		Лабор												1 курс							
																								ІП-з41мп (Б: 0, К: 10);							
																								1 семестр				2 семестр			
																								17 тижнів				28 тижнів			
																									у т.ч.				у т.ч.		
			Бюджет	Контракт	Кред ECTS	Години	Всього	за НП	з урах. інд занять	за НП	з урах. інд занять	за НП	з урах. інд занять	Інд. зан.		Екзамени	Заліки	МКР	Курсові роботи	Курсові проекти	РГР,РР,ГР	ДКР	Реф.	Всього	Лекц	Практ	Лаб	Всього	Лекц	Практ	Лаб
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	30	31	32
	Оброблення надвеликих масивів даних	ІПІ			5.0	150	20	10	-	-	-	10	-	0	130	1		1						20	10		10				

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	0.5
Лекція 2. Архітектура та принципи обробки даних в Apache Hadoop. Концепція MapReduce.	0.5
Лекція 3. Архітектура та принципи обробки даних в Apache Hive	0.5
Лекція 4-5. Архітектура та принципи обробки даних в Apache Spark	0.5
Лекція 6-7. Обробка графів в Apache Spark	1
Лекція 8. Класифікація даних в Apache Spark	1
Лекція 9. Кластеризація даних в Apache Spark	1
Лекції 10-11. Створення рекомендаційних систем на основі Apache Spark	1
Лекції 12-13. Обробка текстової інформації	1
Лекція 14-15. Обробка потоків даних	1

Лекція 16-18. Кращі практики створення рішень з обробки надвеликих масивів даних	2

Тематичний план проведення лабораторних робіт та комп'ютерних практикумів

№	Лабораторна робота / комп'ютерний практикум	Тема	Аудиторних години	СРС, годин
1	ЛР 1	Розподілена обробка даних в Apache Hadoop та Apache Hive	2	6
2	ЛР 2	Обробка текстової інформації	2	6
3	КР 1	Використання графових алгоритмів в Apache Spark	2	4
4	КР 2	Класифікація та кластеризація даних в Apache Spark	2	2
5	КР 3	Обробка потоків даних в Apache Spark	2	2
	В підсумку		**10**	**20**

6. Самостійна робота студента

Самостійна робота

Тема 1. Характеристики технологій обробки надвеликих масивів даних	10
Тема 2. Виконання операцій join в концепції MapReduce	10
Тема 3. Типи моделей сховищ даних	6
Тема 4. ETL процеси в технологіях обробки надвеликих масивів даних	6
Тема 5. Бібліотеки ML, MLib в Apache Spark	20
Тема 6. Графові алгоритми в Neo4j	12
Тема 7. Методи визначення аномалій для надвеликих масивів даних	6
Тема 8. Фреймворк TEZ	6
Тема 9. Технології управління ресурсами в задачах обробки надвеликих масивів даних	6
Тема 10. Кращі практики рішень з обробки надвеликих масивів даних	6
Виконання лабораторних робіт на комп'ютерних практикумів	20
Підготовка до модульних контрольних робіт	6
Підготовка до екзаменаційної роботи по всьому матеріалу модуля.	16
ВСЬОГО	**130**

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; у випадку не виконання студентом усіх передбачених навчальним планом видів занять до екзамену він не допускається.

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)

Рейтинг студента з дисципліни складається з балів, що він отримує за:

1. виконання та захист 3 практикумів та 2х лабораторних робіт;
2. виконання модульної контрольної роботи (МКР);

Лабораторна робота (ЛР): до 15 балів

Комп'ютерні практикуми здається очно/в дистанційному форматі та оцінюються так:

– «відмінно», повна відповідь на питання під час захисту (не менш ніж 90% потрібної інформації) та оформлений належним чином електронний протокол до практикуму – 15/13 балів;

– «добре», достатньо повна відповідь на питання під час захисту (не менш ніж 75% потрібної інформації) та оформлений належним чином електронний протокол до практикуму – 12/11 бали;

– «задовільно», неповна відповідь на питання під час захисту (не менш ніж 60% потрібної інформації), незначні помилки та оформлений належним чином електронний протокол до практикуму – 10/8 бал;

– «незадовільно», незадовільна відповідь та/або не оформлений належним чином електронний протокол до практикуму – 0/7 балів.

Комп'ютерний практикум (КП): до 10 балів

Комп'ютерні практикуми здається очно/в дистанційному форматі та оцінюються так:

– «відмінно», повна відповідь на питання під час захисту (не менш ніж 90% потрібної інформації) та

оформлений належним чином електронний протокол до практикуму – 10/9 балів;

– «добре», достатньо повна відповідь на питання під час захисту (не менш ніж 75% потрібної інформації) та оформлений належним чином електронний протокол до практикуму – 8/7 бали;

– «задовільно», неповна відповідь на питання під час захисту (не менш ніж 60% потрібної інформації), незначні помилки та оформлений належним чином електронний протокол до практикуму – 5/6 бал;

– «незадовільно», незадовільна відповідь та/або не оформлений належним чином електронний протокол до практикуму – 0/4 балів.

Заохочувальні бали

– за виконання творчих робіт з кредитного модуля (наприклад, участь у факультетських та інститутських олімпіадах з навчальних дисциплін, участь у конкурсах робіт, підготовка оглядів наукових праць тощо); за активну роботу на лекції (питання, доповнення, зауваження за темою лекції, коли лектор пропонує студентам задати свої питання) 1-2 бали, але в сумі не більше 10;

– презентації по СРС – від 1 до 5 балів.

– додаткові факультативні лабораторні роботи/комп'ютерні практикуми – від 1 до 10 балів.

Модульна контрольна робота

Ваговий бал МКР – 30 балів. *Критерії оцінювання кожної частини МКР:*

– "*відмінно*", повна відповідь (не менше 90% потрібної інформації) – 25–30 балів;

– "*добре*", достатньо повна відповідь (не менше 75% потрібної інформації), або повна відповідь з незначними помилками – 17-24 балів;

– "*задовільно*", неповна відповідь (не менше 60% потрібної інформації) та незначні помилки – 9-16 балів;

– "*незадовільно*", незадовільна відповідь (взагалі неправильна відповідь) – 0-8 балів.

РОЗПОДІЛ БАЛІВ, ЯКІ ОТРИМУЮТЬ СТУДЕНТИ З ДИСЦИПЛІНИ

Види контролю	*бали*
ЛР «Розподілена обробка даних в Apache Hadoop та Apache Hive»	*15*
ЛР «Обробка текстової інформації»	*15*
КП «Використання графових алгоритмів в Apache Spark»	*10*
КП «Класифікація та кластеризація даних в Apache Spark»	*10*
КП «Обробка потоків даних в Apache Spark»	*10*
МКР	*30*
Заохочувальні бали / Самостійна робота	*10*

*R=2*15+3*10+30+10=100*

Семестровий контроль: іспит

1. Студенти, які набрали протягом семестру кількість балів $R_D \geq 60$, мають можливість:

- отримати іспит з кредитного модуля «автоматом» відповідно набраного рейтингу;
- виконувати екзаменаційну контрольну роботу з метою підвищення оцінки.

Якщо оцінка за екзаменаційну контрольну роботу більша ніж «автоматом» за рейтингом, студент отримує оцінку за результатами екзаменаційної контрольної роботи. Інакше – застосовується варіант жорсткої РСО:

якщо студент хоче отримати вищу оцінку, його попередні бали анулюються і він пише екзаменаційну роботу, яка оцінюється в 100 балів.

Екзаменаційна контрольна робота складається з 2 теоретичних питань з оцінюванням по 25 балів кожне та 2 практичних завдань по 25 балів кожне.

2. Студенти, які наприкінці семестру мають рейтинг $40 \le R_D < 60$ виконують екзаменаційну контрольну роботу. При цьому рейтингова оцінка з кредитного модуля складається з балів за екзаменаційну контрольну роботу і ця рейтингова оцінка є остаточною.

3. Студенти, які наприкінці семестру мають рейтинг $R_D < 40$ до іспиту не допускаються і повинні виконувати додаткову роботу для підвищення свого рейтингу.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	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.)