Міністерство освіти і науки України
Національний аерокосмічний університет ім. М. Є. Жуковського
«Харківський авіаційний інститут»

Кафедра комп'ютерних систем, мереж і кібербезпеки (№ 503)
(назва кафедри)

ЗАТВЕРДЖУЮ

Голова НМК
_____ Д.М. Крицький
(підпис) (ініціали та прізвище)

« 31 » серпня 2022 р.

РОБОЧА ПРОГРАМА ОБОВ'ЯЗКОВОЇ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Технології обробки великих даних
(назва навчальної дисципліни)

Галузь знань: 12 «Інформаційні технології»
(шифр і найменування галузі знань)

Спеціальність: 123 «Комп'ютерна інженерія»
(код та найменування спеціальності)

Освітні програми: «Комп'ютерні системи та мережі», «Системне програмування», «Програмовні мобільні системи та Інтернет речей»
(найменування освітньої програми)

Форма навчання: денна

Рівень вищої освіти: другий (магістерський)

Харків 2022 рік

Розробник: Фесенко Герман Вікторович, професор, д.т.н., професор _____
(прізвище та ініціали, посада, науковий ступінь та вчене звання) (підпис)

Робочу програму розглянуто на засіданні кафедри комп'ютерних систем, мереж і кібербезпеки
(назва кафедри)

Протокол № 1 від « 30 » серпня 2022 року

Завідувач кафедри д.т.н., професор _____ В.С. Харченко
(науковий ступінь та вчене звання) (підпис) (ініціали та прізвище)

1. Опис навчальної дисципліни

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, спеціальність, спеціалізація, рівень вищої освіти</td><td>Характеристика навчальної дисципліни</td></tr>
<tr><td>Денна форма навчання</td></tr>
<tr><td>Кількість кредитів – 4,5</td><td>Галузь знань:
12 «Інформаційні технології»</td><td>Обов’язкова</td></tr>
<tr><td>Модулів – 1</td><td rowspan="4">Спеціальність:
123 «Комп’ютерна інженерія»
Освітні програми:
«Комп’ютерні системи та мережі», «Системне програмування», «Програмовні мобільні системи та Інтернет речей»</td><td rowspan="2">Навчальний рік
2022/2023</td></tr>
<tr><td>Змістовних модулів – 2</td></tr>
<tr><td>Індивідуальне науково-дослідне завдання: немає</td><td rowspan="2">Семестр</td></tr>
<tr><td>Загальна кількість годин – денна – 481)/135</td><td>1-й</td></tr>
<tr><td rowspan="12">Тижневих годин для денної форми навчання:
аудиторних – 3
самостійної роботи здобувача – 5,4</td><td rowspan="12">Рівень вищої освіти:
другий (магістерський)</td><td>Лекції1)</td></tr>
<tr><td>32 години</td></tr>
<tr><td>Практичні1)</td></tr>
<tr><td>0 годин</td></tr>
<tr><td>Лабораторні1)</td></tr>
<tr><td>16 годин</td></tr>
<tr><td>Самостійна робота</td></tr>
<tr><td>87 годин</td></tr>
<tr><td>Вид контролю</td></tr>
<tr><td>Іспит</td></tr>
</table>

Співвідношення кількості годин аудиторних занять до самостійної роботи становить *48/87*.

[1)] Аудиторне навантаження може бути зменшене або збільшене на одну годину в залежності від розкладу занять.

2. Мета та завдання навчальної дисципліни

Мета: формування знань та умінь про застосування технологій розподіленої обробки структурованих та неструктурованих наборів великих даних з використанням сучасних методів та інструментів.

Завдання:

- формування знань і навичок щодо організації зберігання великих даних;
- формування знань і навичок щодо організації розподіленої обробки великих даних;
- отримання знань щодо побудови і функціонування еталонної архітектури великих даних.

Компетентності, які набуваються: Дисципліна має допомогти сформувати у здобувачів такі загальні та спеціальні компетентності:

- ЗК2. Здатність до абстрактного мислення, аналізу і синтезу.
- ЗК4. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
- ЗК6. Здатність виявляти, ставити та вирішувати проблеми.
- СК7. Здатність досліджувати, розробляти та обирати технології створення великих і надвеликих систем.
- СК11. Здатність обирати ефективні методи розв'язування складних задач комп'ютерної інженерії, критично оцінювати отримані результати та аргументувати прийняті рішення.

Очікувані результати навчання. В результаті вивчення дисципліни здобувачі мають досягти такі результати навчання:

- ПРН2. Знаходити необхідні дані, аналізувати та оцінювати їх.
- ПРН10. Здійснювати пошук інформації в різних джерелах для розв'язання задач комп'ютерної інженерії, аналізувати та оцінювати цю інформацію.

Пререквізити: дисципліна є обов'язковим компонентом освітньої програми і базується на знаннях, отриманих під час вивчення дисциплін у циклі загальної і професійної підготовки, передбачених навчальним планом спеціальності.

Кореквізіти: Матеріал, засвоєний під час вивчення цієї дисципліни, є базою для дисципліни «Комп'ютерні системи штучного інтелекту».

3. Зміст навчальної дисципліни

Модуль 1

Змістовний модуль 1. Технології зберігання великих даних, обробка великих даних з використанням екосистеми Hadoop

Тема 1. Розуміння великих даних

Термінологія та стандарти у галузі великих даних. Характеристики великих даних. Організаційні передумови, придбання, конфіденційність, безпека даних та походження даних. Особливі вимоги до управління, методологія, хмарні середовища. Життєвий цикл аналітики великих даних. Обробка транзакцій (OLTP) та аналітична обробка (OLAP) в реальному часі, добування, перетворення і завантаження (ETL). Сховища та вітрини даних.

Тема 2. Концепції та технології зберігання великих даних

Кластери, файлові системи та розподілені файлові системи, база даних NoSQL. Шардінг і реплікація. Поєднання шардінга і реплікації. Теорема CAP, принцип проектування бази даних ACID та принцип BASE. Дискові системи зберігання. Системи зберігання в оперативній пам'яті.

Тема 3. Екосистема Hadoop: MapReduce, HDFS, YARN
Основні відомості про Hadoop та його екосистему. MapReduce. Розподілена файлова система Hadoop. Менеджер ресурсів YARN

Тема 4. Екосистема Hadoop: Kafka
Обмін повідомленнями на кшталт «публікація/підписка». Створення та налаштування конфігурації. Конфігурування апаратних та програмних компонентів. Екосистема даних і Kafka. Сценарії використання. Принципи роботи виробників та споживачів Kafka. Фіксація та зміщення. Особливості внутрішньої побудови Kafka.

Змістовний модуль 2. Технології потокової обробки великих даних та еталонна архітектура

Тема 5. Потокова обробка з використанням Kafka Streams
Що таке потокова обробка. Швидкість, узгодженість, обсяг (SCV). Основні поняття потокової обробки. Патерни проєктування потокової обробки. Kafka Streams: огляд архітектури. Локальність даних. Відновлення після збою та відмовостійкість. Типи вікон Kafka Streams. Приклад потокової обробки. Обробка транзакції придбання товару.

Тема 6. Обробка великих даних у режимі часу, наближеному до реального
Рецепт застосунку для обробки великих даних у режимі часу, наближеного до реального (NRT-застосунку). NRT-застосунки на основі Storm та Spark. Лямбда-архітектура для аналітики в режимі реального часу. NRT-застосунок і часові обмеження. Будівельні блоки NRT-застосунку. Функційний та системний погляд на NRT-застосунок. Потокова обробка з використанням Flink та Spark streaming.

Тема 7. Обробка великих даних з використанням Spark
Загальні відомості про Spark. Чотири складові Spark. Spark у процесі обробки даних / інженерії даних. Spark у наукових дослідженнях у галузі обробки даних. Що можна робити за допомогою Spark. Фрейм даних. Ментальна модель Spark. Взаємодія зі Spark. Стислий огляд компонент та взаємодій між ними. Порівняння Hadoop і Spark.

Тема 8. Еталонна архітектура великих даних
Загальні відомості про еталонну архітектуру великих даних. Концепція еталонної архітектури великих даних. Представлення користувача. Наскрізні аспекти. Функційна архітектура. Функційні компоненти.

4. Структура навчальної дисципліни

Назва змістовного модуля і тем	Кількість годин				
	Усього	У тому числі			
		л	п	лаб.	с. р.
1	2	3	4	5	6
Модуль 1					
Змістовний модуль 1. Технології зберігання великих даних, обробка великих даних з використанням екосистеми Hadoop.					
Тема 1. Розуміння великих даних.	19	4		4	11
Тема 2. Концепції та технології зберігання великих даних.	15	4			11
Тема 3. Екосистема Hadoop: MapReduce, HDFS, YARN.	18	4		3	11
Тема 4. Екосистема Hadoop: Kafka.	14	4			10
Модульний контроль.	1			1	
Разом за змістовним модулем 1	67	16		8	43

Змістовний модуль 2. Технології потокової обробки великих даних та еталонна архітектура.					
Тема 5. Технології обробки великих даних з використанням корпоративних сховищ.	19	4		4	11
Тема 6. Особливості обробки великих даних з використанням хмарного середовища.	15	4			11
Тема 7. Технології обробки великих даних з використанням хмарних сховищ загального призначення.	18	4		3	11
Тема 8. Технології обробки великих даних з використанням реляційних та нереляційних баз даних хмарного середовища.	15	4			11
Модульний контроль.	1			1	
Разом за змістовним модулем 2	68	16		8	44
Усього годин	**135**	**32**		**16**	**87**

5. Теми семінарських занять

№ з/п	Назва теми	Кількість годин
1	*Не передбачено*	
	Разом	

6. Теми практичних занять

№ з/п	Назва теми	Кількість годин
1	*Не передбачено*	
	Разом	

7. Теми лабораторних занять

№ з/п	Назва теми	Кількість годин
1	Робота зі вбудованими функціями модуля Spark SQL	4
2	Робота зі Spark Machine Learning Library: Transformers and Estimators	4
3	Робота зі Spark Machine Learning Library: Supervised Learning (навчання з учителем)	4
4	Робота зі Spark Machine Learning Library: Recommendation Engines	4
	Разом	**16**

8. Самостійна робота

№ з/п	Назва теми	Кількість годин
1	Відпрацювання матеріалів лекційних занять за темою 1	11
2	Відпрацювання матеріалів лекційних занять за темою 2	11
3	Відпрацювання матеріалів лекційних занять за темою 3	11
4	Відпрацювання матеріалів лекційних занять за темою 4	10
5	Відпрацювання матеріалів лекційних занять за темою 5	11
6	Відпрацювання матеріалів лекційних занять за темою 6	11
7	Відпрацювання матеріалів лекційних занять за темою 7	11
8	Відпрацювання матеріалів лекційних занять за темою 8	11
	Разом	**87**

9. Індивідуальні завдання

№ з/п	Назва теми	Кількість годин
1	*Не передбачено*	
	Разом	

10. Методи навчання

Проведення аудиторних лекцій, лабораторних робіт, консультацій, а також самостійна робота здобувачів з використанням відповідних матеріалів (п.14, 15).

11. Методи контролю

Проведення поточного контролю, електронного тестування, підсумковий контроль у вигляді іспиту.

12. Критерії оцінювання та розподіл балів, які отримують здобувачі

Складові навчальної роботи	Бали за одне заняття (завдання)	Кількість занять (завдань)	Сумарна кількість балів
Змістовний модуль 1			
Робота на лекціях	0…2	8	0…16
Виконання і захист лабораторних робіт	0...7	2	0...14
Модульний контроль	0…20	1	0…20
Змістовний модуль 2			
Робота на лекціях	0…2	8	0…16
Виконання і захист лабораторних робіт	0...7	2	0...14
Модульний контроль	0…20	1	0…20
Усього за семестр			**0…100**

Білет для іспиту складається з двох теоретичних питань (30 балів за кожне питання) та практичного завдання (40 балів).

Під час складання семестрового іспиту здобувач має можливість отримати максимум 100 балів.

Критерії оцінювання роботи здобувача протягом семестру

Задовільно (60-74). Показати мінімум знань та умінь. Захистити не менше 75% від усіх завдань лабораторних занять. Знати основні концепції обробки та зберігання великих даних. Уміти працювати зі вбудованими функціями модуля Spark SQL та зі Spark Machine Learning Library з використанням Transformers and Estimators.

Добре (75-89). Твердо знати мінімум, захистити не менше 90% завдань лабораторних занять. Знати ключові принципи організації розподіленої та потокової оброки великих даних, а також принципи побудови еталонної архітектури великих даних. Уміти працювати зі вбудованими функціями модуля Spark SQL та зі Spark Machine Learning Library з використанням Transformers, Estimators та Supervised Learning.

Відмінно (90-100). Здати всі контрольні точки з оцінкою «відмінно». Досконально знати всі теми та уміти їх застосовувати. Вміти ефективно застосовувати функції модуля Spark SQL та основні функції Spark Machine Learning Library.

Шкала оцінювання: бальна і традиційна

Сума балів	Оцінка за традиційною шкалою	
	Іспит, диференційований залік	Залік
90 – 100	Відмінно	Зараховано
75 – 89	Добре	
60 – 74	Задовільно	
0 – 59	Незадовільно	Не зараховано

13. Методичне забезпечення

Навчально-методичний комплекс дисципліни розміщений у системі дистанційного навчання «Ментор».

1. Сторінка дисципліни у системі дистанційного навчання «Ментор» [Ел. ресурс]. URL: https://mentor.khai.edu/course/view.php?id=3702

14. Рекомендована література

Базова

1. Ерл Т., Хаттак В., Булер П. *Основи Big Data: концепції, алгоритми й технології* : пер. з англ. Дніпро: Баланс Бізнес Букс, 2018. 320 с.
2. Ravi V., Cherukuri A. *Handbook of Big Data Analytics*. Vol. 1. *Methodology*. UK, London : The Institution of Engineering and Technology, 2021. 390 p.
3. Ravi V., Cherukuri A. *Handbook of Big Data* Analytics. Vol. 2. *Applications in ICT, security and business analytics*. UK, London : The Institution of Engineering and Technology, 2021. 420 p.
4. Perrin J.-G. *Spark in action*. USA, NY, Shelter Island : Manning Publications Co., 2020, 576 p.

Допоміжна

1. Akhtar S. *Big Data Architect's Handbook*. Birmingham – Mumbai : Packt Publ., 2018. 477 p.
2. Chellappan S., Ganesan D. *Practical Apache Spark*. USA, New York : Apress, 2018. 288 p.

3. Shapira C., Palino T., Sivaram R., Petty K. *Kafka: The Definitive Guide*. USA, CA, Sebastopol : O'Reilly Media, 2022. 486 p.

15. Інформаційні ресурси

1. The 9 Best Free Online Big Data And Data Science Courses [Ел. ресурс]. URL: https://bernardmarr.com/the-9-best-free-online-big-data-and-data-science-courses.
2. Big Data And Data Science Courses [Ел. ресурс]. URL: https://www.edx.org/learn/big-data
3. Mastering Big Data Analytics [Ел. ресурс]. URL: https://www.mygreatlearning.com/academy/learn-for-free/courses/mastering-big-data-analytics