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

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

ЗАТВЕРДЖУЮ

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

«31» серпня 2021 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)

Освітня програма: Системне програмування
(найменування освітньої програми)

Форма навчання: денна

Рівень вищої освіти: другий (магістерський)

Харків 2021 рік

Розробник: Фесенко Г. В., доцент, д.т.н., доцент _____
(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» 08 2021 р.

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>Найменування показників</td><td>Галузь знань, спеціальність, освітня програма, рівень вищої освіти</td><td>Характеристика навчальної дисципліни
(денна форма навчання)</td></tr>
<tr><td>Кількість кредитів – 4,5</td><td rowspan="14">Галузь знань
12 «Інформаційні технології»
(шифр та найменування)

Спеціальність
123 «Комп’ютерна інженерія»
(код та найменування)

Освітня програма
Системне програмування
(найменування)

Рівень вищої освіти:
другий (магістерський)</td><td>Обов’язкова</td></tr>
<tr><td>Кількість модулів – 1</td><td>Навчальний рік</td></tr>
<tr><td>Кількість змістовних модулів – 2</td><td>2021/ 2022</td></tr>
<tr><td>Індивідуальне завдання - немає
(назва)</td><td>Семестр</td></tr>
<tr><td rowspan="2">Загальна кількість годин – 48/135</td><td>1-й</td></tr>
<tr><td>Лекції *</td></tr>
<tr><td rowspan="6">Кількість тижневих годин для денної форми навчання: аудиторних – 3, самостійної роботи студента – 5,4</td><td>32 години</td></tr>
<tr><td>Практичні, семінарські*</td></tr>
<tr><td>0 годин</td></tr>
<tr><td>Лабораторні*</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>залік</td></tr>
</table>

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

*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: комплексне застосування методів та інструментів розподіленої обробки великих даних.

Компетентності, які набуваються:

– здатність до абстрактного мислення, аналізу і синтезу;

– здатність до пошуку, оброблення та аналізу інформації з різних джерел;

– здатність виявляти, ставити та вирішувати проблеми;

– здатність досліджувати, розробляти та обирати технології створення великих і надвеликих систем;

– здатність обирати ефективні методи розв'язування складних задач комп'ютерної інженерії, критично оцінювати отримані результати та аргументувати прийняті рішення.

Очікувані результати навчання:

– знаходити необхідні дані, аналізувати та оцінювати їх;

– здійснювати пошук інформації в різних джерелах для розв'язання задач комп'ютерної інженерії, аналізувати та оцінювати цю інформацію.

Пререквізити – «Бази даних», «Паралельні та розподілені обчислення», «Інженерія програмного забезпечення», «Моделі та структури даних».

Кореквізити – «Комп'ютерні системи штучного інтелекту», «Дипломне проєктування».

3. Зміст навчальної дисципліни

Модуль 1

Змістовний модуль 1. Технології потокової обробки, перенесення великих даних та управління ресурсами кластерів.

Тема 1. Технології потокової розподіленої обробки великих даних.

Фреймворки потокової розподіленої обробки великих даних Apache Kafka Streams, Spark Streaming, Flink, Storm і Samza: архітектура, принципи роботи, переваги, недоліки, галузі застосування. Відмінності між пакетною обробкою і обробкою в реальному часі. П'ять загальних характеристик та десять відмінностей фреймворків Apache Kafka Streams, Spark Streaming, Flink, Storm і Samza.

Тема 2. Технології управління ресурсами кластерів під час роботи з великими даними.

Apache ZooKeeper: архітектура «Клієнт-Сервер», ієрархічний простір імен, підтримувані операції, система розподілених блокувань, продуктівність, надійність, переваги і недоліки. YARN: основні поняття, ключові компоненти, архітектура, протоколи взаємодії, взаємодія ResourceManager і ApplicationManager, переваги і недоліки. Apache Mesos: визначення, історія, принципи роботи, виконувані функції, архітектура, Mesos-masters і Mesos-slaves, фреймворки Aurora, Marathon, Chronos, Jenkins.

Тема 3. Технології перенесення великих даних з різних джерел до аналітичних застосунків та сховищ.

Основні цілі і завдання ETL-процесу. Узагальнена структура ETL-процесу. Особливості та способи витягування даних в ETL-процесі. Очищення даних в ETL-процесі. Два рівня очищення великих даних. Критерії оцінювання якості великих даних. Особливості та способи перетворення даних в ETL-процесі. Перетворення структури та агрегування великих даних. Переведення значень. Створення нових даних. Вибір місця для перетворення даних. Особливості та організація завантаження даних в ETL-процесі. Неповне завантаження даних. Багатопотокова організація процесу завантаження даних. Пост-завантажувальні операції. Завантаження даних з локальних джерел. Особливості безпосереднього завантаження даних з найбільш поширених типів джерел.

Тема 4. Технології обробки великих даних з використання розподіленої файлової системи Hadoop.

Основні поняття про розподілену файлову систему Hadoop (Hadoop Distributed File System (HDFS)). Застосування HDFS. Архітектура HDFS: вузол імен даних, вторинний вузол імен, сервер даних, клієнт. Особливі характеристики HDFS. Файлові операції HDFS: запис, реплікація, читання, видалення. Схема взаємодії вузла імен даних із сервером HDFS. Взаємодія клієнта і кластера. Основні поняття про Apache Hive. Архітектура Apache Hive: Користувальницький інтерфейс, сховище метаданих, процесор Hive SQL, механізм виконання. Основні поняття про Impala. Архітектура Impala: системна служба, служба імен, служба координації метаданих. Головні принципи виконання SQL-запитів.

Модульний контроль.

Змістовний модуль 2. Технології обробки великих даних з використанням корпоративних та хмарних сховищ.

Тема 5. Технології обробки великих даних з використанням корпоративних сховищ.

Огляд архітектур даних. Мультимодальна архітектура даних. Огляд методологій, принципів і концепцій різних типів корпоративних сховищ даних (КСД). Підходи до побудови КСД Кімбала та Інмона. Лямда- і каппа-архітектура. Методологія Data Vault. Метод Anchor. Принцип Data Mesh. Фабрика даних. Принципи побудови КСД. Класичне та віртуальне КСД. Озера та вітрини даних. Ключові відмінності КСД та озер даних.

Тема 6. Особливості обробки великих даних з використанням хмарного середовища.

Еластичні обчислювальні ресурси. Способи управління хмарною інфраструктурою. Безпека хмарного середовища. Розміщення та пакетна обробка великих даних у хмарному середовищі. Хмарні сервіси, пов'язані з великими даними. Рух повідомлень та способи доставки великих даних у хмарне сховище. Обробка інформації і розміщення результатів. Монолітна та багатошарова архітектура. Архітектура мікросервісів. Загальна структура системи, що масштабується і побудована на основі кластера. Сервіси Amazon Web Services (AWS) EMR і Azure HDInsight.

Тема 7. Технології обробки великих даних з використанням хмарних сховищ загального призначення.

Загальні відомості про хмарні сховища загального призначення. Формати зберігання даних. Особливості зберігання великих даних у хмарному сховищі Microsoft Azure Storage. Види сервісів Azure Storage Account. Особливості зберігання великих даних у хмарному сховищі AWS. Amazon Simple Storage Service. Особливості зберігання дисків віртуальних машин.

Тема 8. Технології обробки великих даних з використанням реляційних та нереляційних баз даних хмарного середовища.

Azure SQL: загальні відомості; особливості розміщення; ціновий рівень та продуктивність; концептуальний устрій; конфігурація; управління і моніторинг; сервіс автоматичної оптимізації продуктивності; вбудований редактор запитів. Робота декількох баз даних Azure SQL на одному сервері. Elastic Database Pool. AWS Relational Database Service: загальні відомості; екземпляри. AWS CloudWatch. Архітектура сервісу Aurora. Бази даних типу «ключ - значення» в середовищах Azure і AWS. Графові бази даних в середовищах Azure і AWS. DocumentDB. Нереляційні бази даних типу сімейства стовпців. Azure Table Storage і Azure CosmosDB.

Модульний контроль.

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 потокової обробки, перенесення великих даних та управління ресурсами кластерів.					
Тема 1. Технології потокової розподіленої обробки великих даних.	19	4		4	11
Тема 2. Технології управління ресурсами кластерів під час роботи з великими даними.	15	4			11
Тема 3. Технології перенесення великих даних з різних джерел до аналітичних застосунків та сховищ.	18	4		3	11

Тема 4. Технології обробки великих даних з використання розподіленої файлової системи Hadoop.	14	4			10
Модульний контроль.	1			1	
Разом за змістовним модулем 1	67	16		8	43
Змістовний модуль 2. Технології обробки великих даних з використанням корпоративних та хмарних сховищ.					
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

№ п/п	Назва теми	Кількість годин
1	*Не передбачено*	
2		
	Разом	

6. Теми практичних занять

№ п/п	Назва теми	Кількість годин
1	*Не передбачено*	
2		
	Разом	

7. Теми лабораторних занять

№ п/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

№ п/п	Назва теми	Кількість годин
1	Особливості роботи з великими даним в Spark MLib: основні відомості про використання машинного навчання для обробки великих даних	11
2	Особливості роботи з великими даним в Spark MLib: ML Pipelines**,** Feature Extraction	11
3	Особливості роботи з великими даним в Spark MLib:Transformation, and Selection**,** Evaluation Metrics	11
4	Особливості роботи з великими даним в Spark MLib: навчання з учителем (Supervised Learning)	10
5	Особливості роботи з великими даним в Spark MLib: навчання без учителя (Unsupervised Learning)	11
6	Особливості роботи з великими даним в Spark MLib: Recommendation Engines	11
7	Особливості роботи з великими даним в Spark MLib: Graph Analysis	11
8	Особливості роботи з великими даним в Spark MLib: Deep Learning	11
	Разом	**87**

9. Індивідуальні завдання

Не передбачено

10. Методи навчання

Проведення аудиторних лекцій, лабораторних занять, консультацій, а також самостійна робота студентів за відповідними матеріалами.

11. Методи контролю

Проведення поточного контролю, письмового модульного контролю, підсумковий контроль у вигляді іспиту.

12. Критерії оцінювання та розподіл балів, які отримують студенти

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**

Критерії оцінювання роботи здобувача протягом семестру

Задовільно (60-74). Мати мінімум знань та умінь. Відпрацювати та захистити всі лабораторні роботи та домашні завдання. Вміти самостійно давати характеристику використовуваній технології потокової розподіленої обробки великих даних та технології управління ресурсами кластерів під час роботи з великими даними, працювати зі вбудованими функціями модуля Spark SQL та зі Spark Machine Learning Library з використанням Transformers and Estimators.

Добре (75 - 89). Твердо знать мінімум знань, виконати усі завдання. Показати вміння виконувати та захищати всі лабораторні роботи в обумовлений викладачем строк з обґрунтуванням рішень та заходів, які запропоновано у роботах. Вміти пояснювати вибір технологій потокової обробки, перенесення великих даних та управління ресурсами кластерів, правильно обирати архітектуру корпоративного або хмарного сховища. Вміти ефективно застосовувати функції модуля Spark SQL та основні функції Spark Machine Learning Library.

Відмінно (90 - 100). Повно знати основний та додатковий матеріал. Знати усі теми. Орієнтуватися у підручниках та посібниках. Досконально знати усі основні технології та фреймворки, які використовуються для потокової обробки, перенесення великих даних та управління ресурсами кластерів, а також технології обробки великих даних з використанням корпоративних та хмарних сховищ. Вміти удосконалювати архітектуру корпоративного або хмарного сховища в залежності від зміни завдань, для яких воно використовується. Безпомилково виконувати та захищати всі лабораторні роботи в обумовлений викладачем строк з докладним обґрунтуванням рішень та заходів, які запропоновано у роботах.

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

Дистанційний курс в системі дистанційного навчання Ментор, розташований за адресою: https://mentor.khai.edu/course/view.php?id=3702.

14. Рекомендована література

Базова

1. ISO/IEC TR 20547-1:2020. Information technology – Big data reference architecture – Part 1: Framework and application process.
2. ISO/IEC TR 20547-2:2020. Information technology – Big data reference architecture – Part 2: Use cases and derived requirements.
3. ISO/IEC DIS 20547-3:2020. Information technology – Big data reference architecture – Part 3: Reference architecture.
4. ISO/IEC CD 20547-4:2020. Information technology – Big data reference architecture – Part 4: Security and privacy.
5. ISO/IEC TR 20547-5:2020. Information technology – Big data reference architecture – Part 5: Standards roadmap.
6. Ravi V., Cherukuri A. *Handbook of Big Data Analytics*. Vol. 1. *Methodology*. UK, London : The Institution of Engineering and Technology, 2021. 390 p.
7. 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.
8. Ерл Т., Хаттак В., Булер П. *Основи Big Data: концепції, алгоритми й технології* : пер. з англ. Дніпро: Баланс Бізнес Букс, 2018. 320 с.

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.

15. Інформаційні ресурси

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