Затверджено Вченою Радою ОНУ
імені І.І. Мечникова
від "20"грудня 20 16 р. № 4

Одеський національний університет імені І.І. Мечникова
Кафедра математичного забезпечення комп'ютерних систем

"ЗАТВЕРДЖУЮ"
Проректор з науково-педагогічної роботи
_____
"____" _____ 20___ р.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

ВБ4 Методи Big Data
(назва навчальної дисципліни)

Рівень вищої освіти _____ третій (доктор філософії)

Спеціальність _____ 122 – Комп'ютерні науки
(код і назва спеціальності (тей)

Факультет _____ математики, фізики та інформаційних технологій
(назва інституту, факультету)

2020 – 2021

Робоча програма складена на основі навчальної програми з дисципліни «Методи Big Data».

Розробник:
Петрушина Т.І., к.ф.-м.н., доцент кафедри МЗКС

Робоча програма затверджена на засіданні кафедри математичного забезпечення комп'ютерних систем

Протокол № 1 від. "28" 08 2020 р.

Завідувач кафедри _____ (підпис) (Є.В. Малахов) (прізвище та ініціали)

Схвалено навчально-методичною комісією (НМК) з ІТ факультету МФІТ

Протокол № 1 від. "31" 08 2020 р.

Голова НМК _____ (підпис) (О.В. Савастру) (прізвище та ініціали)

Переглянуто та затверджено на засіданні кафедри _____

Протокол № 1 від. "30" 08 2021 р.

Завідувач кафедри _____ (підпис) (Є.В. Малахов) (прізвище та ініціали)

Переглянуто та затверджено на засіданні кафедри _____

Протокол № ___ від. "____" _____ 20__ р.

Завідувач кафедри _____ (підпис) (_____) (прізвище та ініціали)

1. Опис навчальної дисципліни

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, напрям підготовки, освітньо-кваліфікаційний рівень</td><td colspan="2">Характеристика навчальної дисципліни</td></tr>
<tr><td>денна форма навчання</td><td>заочна форма навчання</td></tr>
<tr><td rowspan="13">Загальна кількість:

кредитів – 3

годин – 90

залікових модулів – 1

змістовних модулів – 3

ІНДЗ* – _____
(вид завдання)</td><td>Галузь знань
<u>12 – Інформаційні технології</u>
(шифр і назва)</td><td colspan="2">Нормативна</td></tr>
<tr><td rowspan="5">Спеціальність
<u>122 – Комп’ютерні науки</u>
(шифр і назва)</td><td colspan="2">Рік підготовки:</td></tr>
<tr><td>2</td><td></td></tr>
<tr><td colspan="2">Семестр</td></tr>
<tr><td>2</td><td></td></tr>
<tr><td colspan="2">Лекції</td></tr>
<tr><td rowspan="7">Рівень вищої освіти:
<u>доктор філософії</u></td><td>16 год.</td><td></td></tr>
<tr><td colspan="2">Практичні, семінарські</td></tr>
<tr><td>14 год.</td><td></td></tr>
<tr><td colspan="2">Лабораторні</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Самостійна робота</td></tr>
<tr><td>60 год.</td><td></td></tr>
<tr><td colspan="2">Індивідуальні завдання:</td></tr>
<tr><td colspan="2">Вид контролю: залік</td></tr>
</table>

* – за наявності

1. Мета дисципліни

Метою викладання курсу є формування у аспірантів представлення про технічні та методологічні засоби організації, зберігання та аналізу надвеликих даних (Big Data).

Завдання: набуття компетентностей на основі засвоєння основних теоретичних положень щодо існуючих технічних та методологічних засобів аналізу надвеликих даних, що забезпечують зберігання і управління обсягом даних в сотні терабайт або петабайт, які звичайні реляційні бази даних не дозволяють ефективно використовувати; освоєння ефективних методів організації неструктурованою інформації (текст, зображення, відео та ін), засобів і способів роботи з нею, генерацію аналітичних звітів, впровадження прогностичних моделей.

Процес вивчення дисципліни спрямований на формування елементів наступних **компетентностей**:

а) загальних:

ЗК 01. Здатність до абстрактного мислення, аналізу та синтезу.

б) фахових:

СК 02. Здатність застосовувати сучасні інформаційні технології, бази даних та інші електронні ресурси, спеціалізоване програмне забезпечення у науковій та навчальній діяльності.

СК 06. Здатність будувати математичні, інформаційні, структурні, онтологічні та інші моделі для опису і подальшого вивчення процесів чи об'єктів предметних областей.

СК 07. Здатність розуміти, виявляти сутність та розв'язувати проблеми перетворення, аналізу та представлення інформації.

СК 10. Здатність створювати та удосконалювати методи аналізу і обробки інформації, а також розробляти на їх основі відповідні інформаційні технології.

Програмні результати навчання:

ПРН 08. Глибоке розуміння загальних принципів та методів комп'ютерних наук, а також методології наукових досліджень, та вміння застосувати їх у власних дослідженнях у сфері комп'ютерних наук та у викладацькій практиці.

ПРН 14. Глибоке розуміння загальних принципів побудови математичних, структурних, онтологічних моделей і метамоделей предметних областей та вміння проводити аналіз таких моделей.

ПРН 15. Глибоке розуміння методів аналізу, обробки і перетворення даних, принципів вдосконалення таких методів і створення нових, вміння розробляти на їх основі нові інформаційні технології та відповідні інтелектуальні та спеціалізовані системи.

Очікувані результати навчання. У результаті вивчення навчальної дисципліни аспірант повинен

***знати**: основні поняття, методи та засоби в області Big Data; основи машинного навчання, візуалізації та зберігання великих даних; основи роботи з хранилищами даних і NoSQL СУБД*

***вміти**: застосовувати підходи аналітичної обробки великих даних при вирішенні задач, пов'язаних з управлінням у складних технічних системах; освоювати методики використання програмних засобів, що підтримують технології Big Data для вирішення практичних задач в предметній області; перевести проблеми предметної області на технологію обробки великих даних.*

2. Зміст навчальної дисципліни

Змістовний модуль 1. Загальні поняття та визначення Big Data.
Тема 1. Сучасні підходи до обробки та зберігання надвеликих даних.
Тема 2. Передобробка даних, візуалізація, первинний статистичний аналіз.
Тема 3. Форми представлення даних, типи і види даних.

Змістовний модуль 2. Програмне забезпечення в області аналізу великих даних.
Тема 1. Аналітичні платформи: класифікація та особливості застосування.
Тема 2. Мови візуального моделювання. Мова R
Тема 3. NoSQL СУБД. Hadoop.
Тема 4. MapReduce. Повнотекстовий пошук. ETL процес по обробці звітів.

Змістовний модуль 3. Технології KDD і Data Mining.
Тема 1. Підготовка даних до аналізу. Методика вилучення знань.
Тема 2. Пошук асоціативних правил. Застосування класифікації і регресії. Статистичні методи.

3. Структура навчальної дисципліни

<table>
<tr><th rowspan="4">Назви змістових модулів і тем</th><th colspan="6">Кількість годин</th></tr>
<tr><th colspan="6">Денна форма</th></tr>
<tr><th rowspan="2">Форма контролю</th><th rowspan="2">Усього</th><th colspan="4">у тому числі</th></tr>
<tr><th>л</th><th>п</th><th>лаб</th><th>ср</th></tr>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr>
<tr><td colspan="7">Змістовний модуль 1. Загальні поняття та визначення Big Data.</td></tr>
<tr><td>Тема 1.</td><td>КО</td><td>9</td><td>1</td><td rowspan="3">4</td><td></td><td rowspan="2">4</td></tr>
<tr><td>Тема 2.</td><td>КО</td><td>21</td><td>2</td><td></td></tr>
<tr><td>Тема 3.</td><td>КО</td><td>22</td><td>2</td><td></td><td>8</td></tr>
<tr><td colspan="7">Змістовний модуль 2. Програмне забезпечення в області аналізу великих даних</td></tr>
<tr><td>Тема 1.</td><td>КО</td><td>22</td><td>1</td><td>1</td><td></td><td>8</td></tr>
<tr><td>Тема 2.</td><td>КО</td><td>22</td><td>2</td><td rowspan="3">5</td><td></td><td>8</td></tr>
<tr><td>Тема 3.</td><td>КО</td><td>24</td><td>2</td><td></td><td>8</td></tr>
<tr><td>Тема 4.</td><td>КО</td><td>24</td><td>2</td><td></td><td>8</td></tr>
<tr><td colspan="7">Змістовний модуль 3. Технології KDD і Data Mining</td></tr>
<tr><td>Тема 1.</td><td>КО</td><td>24</td><td>2</td><td rowspan="2">4</td><td></td><td>8</td></tr>
<tr><td>Тема 2.</td><td>КО</td><td>24</td><td>2</td><td></td><td>8</td></tr>
<tr><td>Всього годин</td><td></td><td>90</td><td>16</td><td>14</td><td></td><td>60</td></tr>
</table>

Форма контролю: **КО** –контрольне опитування (поточне)

4. Теми семінарських/практичних занять

№ з/п	Назва теми	Кількість годин
1	Підготовка, обробка, аналіз та візуалізація результатів з допомогою MS SQL Analytical Services.	4
2	Кластерний аналіз у середовищі R	4
3	Регресійний аналіз в середовищі R	4
4	Байєсова класифікація в середовищі R	2
	Разом	**14**

5. Теми лабораторних занять

Немає

6. Самостійна робота

Основні види самостійної роботи з Модулю1:

– робота над індивідуальним науково-дослідним завданням (10 год);

– підготовка до практичних занять; опрацювання лекційного матеріалу та рекомендованої літератури; самоперевірка аспирантом власних знань за запитаннями до самоконтроля (розподіл годин за темами модулю наведено в таблиці).

№ з/п	Назва теми	Кількість годин
1	Аналіз існуючих програмних інструментів Big Data.	10
2	Специфіка застосування технології KDD до великих даних.	15
3	Застосування нейронних мереж до аналізу великих даних.	15
4	Організація обробки даних в HDFS.	20
	Разом	**60**

До самостійної роботи відноситься:

[1] – підготовка до лекцій та лабораторних занять;

[2] – виконання індивідуальних завдань.

7. Індивідуальне навчально-дослідне завдання

Окремого завдання немає. Практична робота виконується на основі дисертаційного дослідження

8. Курсовий проект

Немає

9. Методи навчання

Лекції з використанням мультимедійного презентаційного матеріалу.

10. Методи контролю

Навчальна дисципліна вивчається протягом одного семестру та складається з трьох змістових модулів (ЗМ). По завершенні курсу аспіранти складають залік. Залік виставляється за результатами роботи здобувача вищої освіти впродовж усього семестру.

Здобувач вищої освіти одержує підсумкову оцінку, якщо за результатами поточного контролю він набрав 60 і більше балів.

Якщо за результатами поточного контролю здобувач вищої освіти набрав менше 60 балів, він повинен виконати залікове завдання і з урахуванням його результатів одержати відповідну кількість залікових балів із дисципліни.

Якщо за результатами поточного контролю здобувач вищої освіти набрав 60 і більше балів, проте хоче поліпшити свій підсумковий результат, він також може виконати залікове завдання і з урахуванням його результатів підвищити свою підсумкову оцінку з дисципліни.

Під час викладання матеріалу лекцій лектор може здійснювати контрольні опитування здобувачів. Максимальна оцінка, яку аспірант може отримати за ці опитування, складає 10 балів, які є складовими комплексного підсумкового контролю.

Оцінювання під час поточного та підсумкового контролю здійснюється за 100-бальною шкалою шляхом:

– усне опитування на практичних заняттях;
– оцінювання якості виконання індивідуальної роботи (у вигляді проєкту / реферату).

Підсумкове оцінювання здійснюється шляхом врахування усіх отриманих здобувачем вищої освіти оцінок. Підсумкова оцінка визначається як зважена сума балів за результатами поточного контролю. За три змістових модулі аспірант може одержати максимум 100 балів.

10.1. Критерії оцінювання на підсумковому модульному контролі:

1. Відповідь повинна бути повною і короткою. Вона не повинна мати в собі матеріал, що не відноситься до сутті питання.
2. Чітко формулювати твердження, вправно застосовувати необхідні формули і знання основних питань програми.
3. Відповіді, що мають помилкові твердження оцінюються виходячи з близькості відповіді до правильної.
4. Пропуски в обґрунтуванні тверджень враховуються і це призводить до зменшення кількості балів.

5. Малі недоліки, неточності при викладенні матеріалу, зменшують кількість балів.
6. Незнання і нерозуміння основної ідеї теоретичного питання або задачі призводить до зняття до 90 % балів.
7. Якщо відповідь на питання відсутня то виставляється нуль балів.

10.2. Питання до підсумкового контролю

1. Визначте сутність поняття «великі дані».

2. Охарактеризуйте методики аналізу великих даних.

3. Опишіть процес аналізу великих даних.

4. Які особливості зберігання великих даних.

5. Дайте характеристи Big Data на світовому ринку.

6. Визначте поняття Data Mining.

7. Визначте поняття KDD

8. Визначте відмінності між параметричними, непараметричними і номінальними методами.

9. Які завдання підтримки безпеки великих даних.

10. У чому полягає когнітивний аналіз даних.

11. Які моделі даних ви знаєте?

12. Основні описові статистики.

13. Особливості застосування кореляційно-регресивного аналізу великих даних.

14. Сутність кластерного аналізу в застосуванні до великих даними.

15. Пошук асоціативних правил в великих даних.

16. Класифікація даних за допомогою нейромережі.

17. Класифікація за допомогою дерев рішень.

18. Програмні засоби аналізу великих даних і їх недоліки.

19. Основні можливості зберігання великих даних на мові програмування R.

20. Основні можливості мови R для аналізу великих даних.

11. Розподіл балів, які отримують аспіранти

Поточне тестування та самостійна робота			Сума
Змістовий модуль 1	Змістовий модуль 2		

							Змістовий модуль 3		
Т1	Т2	Т3	Т1	Т2	Т3	Т4	Т1	Т2	
8	10	12	10	10	10	10	15	15	100

Т1, Т2 ... – теми змістових модулів

Шкала оцінювання: національна та ECTS

Загальна сума балів	Оцінка ECTS	Національна шкала	
90 — 100	A – «відмінно»	5 «відмінно»	«залік»
85 — 89	B – «дуже добре»	4 «добре»	
75 — 84	C – «добре»		
70 — 74	D – «задовільно»	3 «задовільно»	
60 — 69	E – «допустимо»		
35 — 59	F – «незадовільно з можливістю повторного складання»	2 «незадовільно»	«незалік»
0 — 34	FX – «незадовільно з обов’язковим повторним курсом»		

12. Методичне забезпечення

Конспект лекцій в електронному форматі; комплекс навчально-методичного забезпечення дисципліни; нормативні документи; презентаційні матеріали.

13. Рекомендована література

13.1. Основна література

1. Вайгенд А. Big Data. Вся технология в одной книге. – Интернет издательство «Эксмо», 2018. – 384 с.
2. Дэви С. Основы Data Science и Big Data. Python и наука о данных.//С.Дэви, М.Арно, А.Мохамед — СПб.: Питер, 2017. — 336 е.: ил.
3. Ричарт В., Коэльо П.Л. Построение систем машинного обучения на языке Python. – М.: ДМК Пресс, 2016, - 302 с: ил.
4. С. Хайкин. Нейронные сети: полный курс, 2-е издание. :Пер. с англ.-М. : Издательский дом «Вильямс», 2006.-1104с.
5. Інтелектуальний аналіз даних: Підручник / Черняк О.І., Захарченко П.В./ К.: Знання, 2014р. - 599 с.

6. Акіменко В.В., Загородній Ю.В. Проектування СППР на основі нечіткої логіки. Навчально- методичний посібник. – К.: Вид-во КНУ, 2007. – 94с.
7. Уайт Т. Hadoop: Подробное руководство. — СПб.: Питер, 2013. — 672 с.:ил.

13.2. Допоміжна література

1. Майер-Шенбергер, В.Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. — М. : Манн, Иванов и Фербер, 2014. — 240 с.
2. Чубукова И. А. Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с
3. Чак Л. Hadoop в действии. - М.: ДМК Пресс, 2012. - 424с: ил.
4. Шипунов А.Б., Балдин Е.М., Волкова П.А., Коробейников А.И., Назарова С.А., Петров С.В., Суфиянов В.Г. Наглядная статистика. Используем R! - М.: ДМК Пресс, 2012. - 298с.
5. Маккинли, У. Python и анализ данных / У. Маккинли. - М.: ДМК, 2015. - 482 с.

7.1. Електронні інформаційні ресурси

1. Великі дані. [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/Великі_дані
2. Big Data for Development: From Information- to Knowledge Societies [Електронний ресурс]. – Режим доступу: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145
3. Big Data от А до Я. [Електронний ресурс]. – Режим доступу: https://habrahabr.ru/company/dca/blog/267361/
4. Big Data и блокчейн. [Електронний ресурс]. – Режим доступу:
5. https://forklog.com/big-data-i-blokchejn-proryv-v-oblasti-analiza-dannyh/
6. Онлайн курс Machine learning https://www.coursera.org/course/ml
7. Онлайн курс Big Data Overview https://education.emc.com/academicalliance/elearning/Big_Data_Overview/index.htm
8. Онлайн курс R programming https://www.coursera.org/course/rprog
9. Онлайн курс Introduction to Data Science https://www.coursera.org/course/datasci
10. Онлайн курс «Введение в аналитику больших массивов данных» http://bit.ly/IntuitBDA.
11. Учебник по статистическому обучению http://statweb.stanford.edu/~tibs/ElemStatLearn/