ДЕРЖАВНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

СИСТЕМА УПРАВЛІННЯ ЯКІСТЮ
Система забезпечення якості освітньої діяльності та якості вищої освіти
сертифікована на відповідність ДСТУ ISO 9001:2015 / ISO 9001:2015

Кафедра комп'ютерних наук та інформаційних систем

ЗАТВЕРДЖЕНО
вченою радою ДТЕУ
(прот. п.__ від «23» 04 2024 р.)
Ректор
_____ Анатолій Мазаракі

КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA) /
COMPUTER TECHNOLOGIES OF BIG DATA PROCESSING
ПРОГРАМА /
COURSE SUMMARY

Київ 2024

Автори: Т.В. Томашевська, канд. тех. наук, доц.

Програму розглянуто і затверджено на засіданні кафедри комп'ютерних наук та інформаційних систем «26» березня 2024р., протокол № 31

Рецензенти: Т.О.Філімонова, канд. фіз.-мат. наук, доц., доцент кафедри комп'ютерних наук та інформаційних систем
Н.О.Гордійко, кан. техн. наук, доц., доцент кафедри прикладної фізики Фізико-технічного інституту Національного технічного університету України «КПІ імені Ігоря Сікорського»

КОМП'ЮТЕРНІ ТЕХНОЛОГІЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ (BIG DATA) /
COMPUTER TECHNOLOGIES OF BIG DATA PROCESSING

ПРОГРАМА /
COURSE SUMMARY

ВСТУП

Програма дисципліни «Комп’ютерні технології обробки великих даних (Big Data)» призначена для здобувачів другого (магістерського) рівня вищої освіти галузі знань 12 "Інформаційні технології" спеціальності 122 "Комп’ютерні науки".

Програму підготовлено з урахуванням вимог Стандарту вищої освіти України із зазначеної спеціальності та відповідної освітньо-професійньої програми підготовки магістрів.

Розроблена програма складається з таких розділів:

1. Мета, завдання та предмет дисципліни.
2. Передумови вивчення дисципліни як вибіркової компоненти освітньої програми.
3. Результати вивчення дисципліни.
4. Зміст дисципліни.
5. Список рекомендованих джерел.

1. МЕТА, ЗАВДАННЯ ТА ПРЕДМЕТ ДИСЦИПЛІНИ

Метою вивчення дисципліни "Комп’ютерних технологій обробки великих даних (Big Data)" є формування у студентів навичок оволодіння технологіями обробки великих даних, опанування інструментами для розробки програмного забезпечення для розв’язання задач з використанням великих даних.

Завданням вивчення дисципліни є теоретична та практична підготовка студентів з таких питань:

- методи зберігання та обробки великих даних з використанням сучасних фреймворків;
- методи аналізу великих даних та технології використання нереляційних баз даних;
- технології паралельної та розподіленої обробки великих даних в пакетному та реальному режимах;
- алгоритми фільтрації, варіації та зберігання великих даних із застосовуванням специфікацій, стандартів, правил і рекомендацій;
- методи та засоби візуалізації великих даних.

Предметом дисципліни є знання у сфері технологій обробки великих даних з використанням інструментів роботи

2. ПЕРЕДУМОВИ ВИВЧЕННЯ ДИСЦИПЛІНИ ЯК ВИБІРКОВОЇ КОМПОНЕНТИ ОСВІТНЬОЇ ПРОГРАМИ

Знання:

- предмету і сутності комп'ютерних технологій обробки даних;
- основ дискретного аналізу, математичної статистики;
- основ методології математичного моделювання економічних процесів;
- механізмів застосування теоретичних методів і моделей у відображенні економічних процесів.

Вміння:

- базові знання з програмування мовою Python;
- навички роботи з СУБД реляційного типу та знання мови запитів SQL;
- розуміти математичні моделі структурованих даних.

3. РЕЗУЛЬТАТИ ВИВЧЕННЯ ДИСЦИПЛІНИ

Дисципліна дисципліни "Комп'ютерних технологій обробки великих даних (Big Data)", як обов'язкова компонента освітньої-професійної програми, забезпечує оволодіння студентами загальними та фаховими компетентностями і досягнення ними програмних результатів навчання за відповідними освітньо-професійними програмами:

"Комп'ютерні науки" (ОС магістра)

Номер в освітній програмі	**Зміст компетентності**	**Номер теми, що розкриває зміст компетентності**
Загальні компетентності за освітньо-професійною програмою		
ЗК 02	Здатність застосовувати знання у практичних ситуаціях	1.1, 2.3, 2.8
ЗК 05	Здатність вчитися й оволодівати сучасними знаннями.	1.1, 1.2, 2.1-2.8
Фахові компетентності за освітньо-професійною програмою		
СК 04	Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проектних рішень.	1.1, 1.2, 2.1, 2.3, 2.4, 2.5, 2.6
СК 06	Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі	2.4, 2.7

	комп’ютерних наук.	
СК 07	Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень.	1.2, 2.8
СК 09	Здатність розробляти та адмініструвати бази даних та знань.	1.2, 2.3, 2.6
Програмні результати навчання за освітньо-професійною програмою		
РН8	Розробляти математичні моделі та методи аналізу даних (включно з великим).	2.4, 2.5
РН9	Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими).	1.2, 2.1-2.8
РН12	Проектувати та супроводжувати бази даних та знань.	1.2, 2.3, 2.6

4. ЗМІСТ ДИСЦИПЛІНИ

РОЗДІЛ 1. Вступ до аналітики великих даних

Тема 1.1. Глобальні групи даних. Категорії даних.

Поняття про глобальні групи даних – Shallow data, Deep Data, Micro-data, Nano-data, їх структуризацію та класифікацію. Ідентифікація машинних даних, потокових даних, озер даних. Категорії структурованих, неструктурованих та напівструктурованих даних.

Список рекомендованих джерел:
Основний: 1[14-42], 2[10-48]
Додатковий: 5[7-29]
Інтернет-ресурси: 11

Тема 1.2. Інструменти по роботі з великими даними

Інструменти маніпуляції з великими даними: Hadoop, Apache Spark, Hive, Apache Kafra. Інструменти для візуалізації великих даних. Інструменти для зберігання великих даних. Бази даних для роботи з великими даними. Моделі даних No-SQL. Типи No-SQL БД. Алгоритм MapReduce. Архітектура MapReduce. Фази MapReduce.

Список рекомендованих джерел:
Основний: 1[42-69], 4[52-84]
Додатковий: 6[11-29]. 8 [90-160] , 10 [1-24]
Інтернет-ресурси: 11, 12

РОЗДІЛ 2. Робота з PySpark

Тема 2.1. Введення в PySpark та архітектура Apache Spark.

Екосистема Spark. Ознайомлення з основними концепціями та інструментами PySpark. Модулі та компоненти Spark: Огляд основних модулів, таких як Spark SQL, Spark Streaming та MLlib. Приклади використання. Встановлення та налаштування PySpark. Конфігурація Spark. Інтеграція з Jupyter та IDE. Оптимізація ресурсів. Методи ефективного використання обчислювальних ресурсів. Обробка помилок та моніторинг.

Список рекомендованих джерел:
Основний: 1[123-158], 3[15-54]
Додатковий: 6[24-89] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.2. Робота з PySpark.

Робота з PySpark. Скорочена нотація перетворень. Виконання програм PySpark. Визначення основної функції-драйвера. Об’єкт SparkSession. Зчитування даних та перетворення їх. Застосування перетворень Spark. Збереження результатів

Список рекомендованих джерел:
Основний: 1[123-158], 3[15-54]
Додатковий: 6[24-89] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.3. Абстракції даних в PySpark.

Операції, доступні для RDD **(Resilient Distributed Datasets)**. Кешування: оптимізація продуктивності з використанням кешування. Робота з файловою системою. Взаємодія із розподіленим файловим сховищем. Оптимізація та розподіл даних. Методи оптимізації процесу обробки даних. DataFrame та Dataset: Основні операції з даними у форматі DataFrame та Dataset. Об’єднання даних: Злиття та об’єднання даних із різних джерел. Операції з колонками: Маніпуляції з даними на рівні колонок. Обробка пропущених даних. Оптимізація запитів Spark SQL.

Список рекомендованих джерел:
Основний: 1[199-430], 2[21-65]
Додатковий: 6[235-348] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.4. Машинне навчання з PySpark MLlib:

Основні алгоритми машинного навчання: Огляд доступних алгоритмів у MLlib. Навчання та оцінка моделей: Процес навчання моделей та оцінка їх

ефективності. Гіперпараметри та крос-валідація. Оптимізація моделей з використанням крос-валідації. Інтеграція з DataFrame. Робота з даними у форматі DataFrame у контексті машинного навчання. Оптимізація продуктивності ML-задач: Методи оптимізації продуктивності під час роботи з великими обсягами даних.

Список рекомендованих джерел:
Основний: 1[434-463], 3[98-156]
Додатковий: 7[86-148] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.5. Обробка потокових даних за допомогою PySpark Structured Streaming:

Введення у потокову обробку: Основи роботи з потоковими даними. Структурований стрімінг. Використання Structured Streaming для обробки даних. Вікна часу та агрегації: Робота з вікнами часу для агрегації даних. Стан та надійність: Забезпечення надійності обробки потокових даних. Інтеграція із зовнішніми джерелами: Взаємодія з різними джерелами потокових даних.

Список рекомендованих джерел:
Основний: 3[167-196]
Додатковий: 5[200-220] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.6. Інтеграція PySpark з базами даних:

Методи взаємодії з різними базами даних. Оптимізація SQL-запитів. Робота з індексами. Обробка транзакцій при взаємодії з базою даних. Інтеграція з NoSQL базами даних: Взаємодія з NoSQL сховищами даних. Забезпечення безпеки під час роботи із зовнішніми сховищами даних.

Список рекомендованих джерел:
Основний: 1[434-463], 3[98-156]
Додатковий: 5[126-178] , 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.7. Обробка графових даних із GraphFrames:

Введення у графові структури даних: Розуміння базових понять графів. Створення та маніпуляції графами. Алгоритми обробки графів. Застосування різних алгоритмів до графових структур. Візуалізація графів. Інструменти та методи для візуалізації графових даних.

Список рекомендованих джерел:
Основний: 3[390-410], 4[172-190]
Додатковий: 7[86-148], 9 [54-129], 10 [1-24]
Інтернет-ресурси: 13, 14

Тема 2.8. Розгортання та масштабування PySpark додатків:

Оптимізація продуктивності в PySpark. Кешування та Broadcast. Використання кешування та передачі даних на рівні вузлів. Оптимізація коду на рівні операцій: Рефакторинг коду для більш ефективного виконання операцій. Використання поділу даних. Профілювання та моніторинг. Інструменти для аналізу продуктивності програми. Оптимізація під кластер. Робота з налаштуваннями Spark для оптимального використання ресурсів кластера. Огляд середовищ розгортання та їх особливостей. Масштабування додатків. Методи горизонтального та вертикального масштабування. Налаштування параметрів для великих обсягів даних. Ефективне використання обчислювальних ресурсів. Моніторинг та налагодження у продакшені. Інструменти для відстеження та вирішення проблем у робочому середовищі.

Список рекомендованих джерел:

Основний: 3[52-90]
Додатковий: 5[-148], 10 [1-24]
Інтернет-ресурси: 12, 13, 14

5. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основний:*

1. Талах М.В. Технології обробки Big Data. Навчальний посібник. Чернівці: Чернівецький нац.ун-т, 2024. 454 с.

2. Ryza S. Advanced Analytics with PySpark, Patterns for Learning from Data at Scale Using Python and Spark, 2022. 233 p.

3. Rioux J. Data Analysis with Python and PySpark. Manning, 2022. 456 p.

4. Ghavami P. Big Data Governance: Modern Data Management Principles for Hadoop, NoSQL & Big Data Analytics. CreateSpace Independent Publishing Platform, 2016. 204 p

Додатковий

5. Akerkar R. Models of Computation for Big Data Cham: Springer International Publishing, 2018.

6. Chambers B., Zaharia M. The Definitive Guide: Big Data Processing Made Simple. 2018. 603 p.

7. Singh P. Machine Learning with PySpark. 2nd Ed. Apress. 2022. 220p.

8. Big Data processing methods, models and information technologies: Monograph / edited by Oleg I. Pursky. – Shioda GmbH, Steyr, Austria, 2019. – 234 p.

9. Pursky O.I. Identifying customer segments in e-trade using system analysis and clustering methods / O.I.Pursky //Monograph – Agenda Publishing House, Coventry, United Kingdom, 2018. – 140p.

10. Томашевська Т.В. Комп'ютерні технології обробки великих даних (Big Data): Методичні рекомендації до лабораторних занять / Т.В. Томашевська. – Київ: Державний торговельно-економічний ун-т, 2024. – 24 с.

Інтернет ресурси

11. KDNuggets: Data Mining Community Top Resource for Analytics, Data Mining, and Data Science Software, Companies, Data, Jobs, Education, News, and more. URL: http://www.kdnuggets.com

12. The Data Mine. URL: http://www.the-data-mine.com

13. Уроки PySpark Tutorial for Beginners. URL: https://zecourse.com/course/youtubecomchanneluckw4jcwtegrdhisyiiko4tqabout/uroki-pyspark-tutorial-for-beginners

14. PySpark. User Guide. URL: https://spark.apache.org/docs/latest/api/python/user_guide/index.html

*Курсивом позначені видання, що присутні у бібліотеці ДТЕУ