МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

ЗАТВЕРДЖЕНО
на засіданні кафедри
інформаційних систем.
Протокол № 1 від 27.08.2024 р.

ПОГОДЖЕНО
Проректор з навчально-методичної роботи

_____ Каріна НЕМАШКАЛО

АНАЛІТИКА ВЕЛИКИХ ДАНИХ

робоча програма навчальної дисципліни (РПНД)

Галузь знань	**12 "Інформаційні технології"**
Спеціальність	**126 "Інформаційні системи та технології"**
Освітній рівень	**другий (магістерський)**
Освітня програма	**"Інформаційні системи та технології"**

Статус дисципліни	**вибіркова**
Мова викладання, навчання та оцінювання	**українська**

Розробник:
к.е.н., доцент _____ Сергій ЗНАХУР

Завідувач кафедри
інформаційних систем _____ Дмитро БОНДАРЕНКО

Гарант програми підписано КЕП Олександр КОЛГАТІН

Харків
2024

ВСТУП

Дисципліну “Аналітика великих даних” віднесено до групи освітньо-професійних дисциплін підготовки магістрів зі спеціальності 126 "Інформаційні системи та технології". Навчальна дисципліна "Управлінські ІС та сховища даних" є базовою навчальною дисципліною та вивчається згідно з навчальним планом підготовки фахівців другого (магістерського) рівня спеціальності 126 "Інформаційні системи та технології".

Метою вивчення дисципліни “Аналітика великих даних” є надання поглиблених знань та практичних навичок щодо роботи з великими даними, побудови й використання розподілених систем для побудови конвеєрів опрацювання великих даних, формування системи теоретичних знань і набуття практичних умінь та навичок щодо застосування технологій великих даних та розподілених баз.

Завдання навчальної дисципліни "Аналітика великих даних" полягає у формуванні системи теоретичних знань і практичних навичок, необхідних для роботи з великими даними, а також у засвоєнні сучасних технологій і методів для їх опрацювання. Основними завданнями курсу є:

1. Надання знань про побудову та використання розподілених систем для опрацювання великих обсягів даних.
2. Формування вмінь створювати та налаштовувати конвеєри обробки даних із використанням технологій Pandas, Dask, Spark.
3. Забезпечення практичних навичок у реалізації паралельних і розподілених обчислювальних процесів для аналізу даних.
4. Ознайомлення з сучасними методами та алгоритмами Data Mining для виявлення прихованих закономірностей і знань у великих даних.

дання спрямовані на підготовку фахівців, здатних ефективно працювати з великими даними.

Результатом навчальної дисципліни є оволодіння наступними компетентностями: здатність забезпечити процеси отримання та інтерпретації прихованих знань з використанням Pandas, Dask, Spark; здатність реалізовувати обчислювальні процеси на основі використання паралельних і розподілених обчислень, здатність розробляти конвеєри аналізу даних; набуття знань, практичних навичок та умінь використання сучасних методів, алгоритмів Data Mining.

Об'єктом навчальної дисципліни є процеси аналізу даних.

Предметом навчальної дисципліни є основні підходи щодо обробки даних та методи розробки та реалізації конвеєрів щодо аналізу даних.

У процесі навчання здобувачі отримують необхідні знання під час лекційних занять та виконання лабораторних робіт. Також велике значення в процесі

вивчення та закріплення знань має самостійна робота здобувачів. Усі види занять розроблені відповідно до трансферної системи організації навчального процесу. Матеріал, який викладається у цій дисципліні, використовується магістрами для написання дипломної роботи та при вивченні дисциплін у магістратурі.
Результати навчання та компетентності, які формує навчальна дисципліна визначено в табл. 1.

Таблиця 1

Результати навчання та компетентності, які формує навчальна дисципліна

Результати навчання	**Компетентності, якими повинен оволодіти здобувач вищої освіти**
РН07	ЗК01, ЗК04, СК01, СК02, СК03
РН09	СК04, СК05

РН07. Здійснювати обґрунтований вибір проектних рішень та проектувати сервіс-орієнтовану інформаційну архітектуру підприємства (установи, організації тощо).
РН09. Розробляти і використовувати сховища даних, здійснювати аналіз даних для підтримки прийняття рішень.
ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.
ЗК04. Здатність розробляти проекти та управляти ними.
СК01. Здатність розробляти та застосувати ІСТ, необхідні для розв’язання стратегічних і поточних задач.
СК02. Здатність формулювати вимоги до етапів життєвого циклу сервіс-орієнтованих інформаційних систем.
СК03. Здатність проектувати інформаційні системи з урахуванням особливостей їх призначення, неповної/недостатньої інформації та суперечливих вимог.
СК04. Здатність розробляти математичні, інформаційні та комп’ютерні моделі об’єктів і процесів інформатизації
СК05. Здатність використовувати сучасні технології аналізу даних для оптимізації процесів в інформаційних системах.

ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Зміст навчальної дисципліни

Змістовий модуль 1. Основи аналізу великих даних

Тема 1. Введення в дисципліну

Вступ до дисципліни. Мета та завдання дисципліни, її місце у навчальному

процесі. Структура дисципліни, рекомендації щодо її вивчення. Основні поняття аналізу великих даних (Big Data). Професійні вимоги. Спеціалізації Data Science. Етапи та типові кроки аналізу даних. Протокол аналізу даних ("Cross-Industry Standard Protocol for Data Mining" CRISP-DM). Інструменти та технології Pandas, Dask, Spark. Обробка даних на основі Pandas. Особливості обробки даних в DASK. Особливості обробки даних в Spark.

Тема 2. Аналіз даних на основі Dask

Визначення великих даних. 3V:Volume, Velocity, Variety. Різноманітність типів даних і джерел даних. Структуровані та неструктуровані дані. Загальний опис бібліотеки Dask. Переваги та особливості Dask. Паралелізація методом Dask. Приклад статистичного аналізу великого набору даних на Dask.

Тема 3. Основи Apache Spark

Історія розвитку Spark. Big Data фреймворк. Переваги та особливості Apache Spark. MapReduce і Spark. Розвиток MapReduce-Tez. Основи DAG. Плюси і мінуси Lazy Evaluation. Модулі Apache Spark. Процеси Spark. Запуск Spark. Приклад статистичного аналізу великого набору даних на Apache Spark.

Змістовий модуль 2. Побудова рішень у Apache Spark

Тема 4. Реалізації SQL запитів у Apache Spark

Платформа Apache Spark. RDD (Resilient Distributed Datasets). Модуль Spark SQL. Інтеграція SQL з Spark. DataFrame API. Створення DataFrame або таблиць. Синтаксис основних SQL-запитів. Виконання SQL-запитів у Apache Spark. Приклади запитів SQL у Apache Spark.

Тема 5. Машинне навчання з Apache Spark

Загальний опис методів ML. Задача класифікації. Метрики класифікації. Логістична регресія. Дерева та ансамбль дерев. Boosting. Stacking. Структура Apache Spark. Бібліотека машинного навчання MLlib. Функціональність MLlib. Принципи використання MLlib. Приклад рішення задачі класифікації на основі ML Apache Spark.

Тема 6. Використання алгоритмів Spark ML та NLP для побудови конвеєру аналіза даних

Основи NLP (Natural language processing). NLP завдання. Структура NLP конвеєра. Опис кожного етапу конвеєра. Огляд фреймворків для задач NLP на кластері. Реалізація NLP у Spark. Основи Deep Learning на прикладі архітектури BERT. Реалізація Deep Learning у Spark NLP.

Перелік лабораторних занять за навчальною дисципліною наведено в табл. 2.

Таблиця 2

Перелік лабораторних занять

Назва теми	Зміст
Тема 1-2 Лабораторна робота 1	Основи аналізу даних за допомогою бібліотеки DASK
Тема 3 Лабораторна робота 2	Використання базових функцій PySpark
Тема 4 Лабораторна робота 3	Аналітична обробка на основі PySpark SQL&RDD
Тема 5. Лабораторна робота 4	Використання методів ML для вирішення задач аналізу даних у Spark ML (регресія)
Тема 5. Лабораторна робота 5	Використання методів ML для вирішення задач аналізу даних у Spark ML (класифікація)
Тема 6. Лабораторна робота 6	Рішення задач NLP

Перелік самостійної роботи за навчальною дисципліною наведено в табл. 3.

Таблиця 3

Перелік самостійної роботи

Назва теми	Зміст
Тема 1. Завдання 1.	Дослідження тенденцій Big Data
Тема 2. Завдання 2.	Аналіз сучасних інструментів Big Data
Тема 3. Завдання 3.	Порівняння Spark та MapReduce
Тема 4. Завдання 4.	Дослідження Spark Streaming

Тема 5. Завдання 5.	Дослідження методів та алгоритмів MLlib
Тема 6. Завдання 6.	Дослідження трансформерів та LLM у Spark NLP

Кількість годин лекційних, лабораторних занять та годин самостійної роботи наведено в робочому плані (технологічній карті) з навчальної дисципліни.

МЕТОДИ НАВЧАННЯ

У процесі викладання навчальної дисципліни для набуття визначених результатів навчання, активізації освітнього процесу передбачено застосування таких методів навчання, як:
Проблемна лекція (Тема 1 ,2, 3, 5, 6), міні-лекція та дискусія (Тема 4).
Наочні (демонстрація (Тема 1-6)).
Лабораторна робота індивідуальна (Тема 1 – 6)

ФОРМИ ТА МЕТОДИ ОЦІНЮВАННЯ

Університет використовує 100 бальну накопичувальну систему оцінювання результатів навчання здобувачів вищої освіти.
Поточний контроль здійснюється під час проведення лекційних, практичних, лабораторних та семінарських занять і має на меті перевірку рівня підготовленості здобувача вищої освіти до виконання конкретної роботи і оцінюється сумою набраних балів для дисциплін з формою семестрового контролю екзамен (іспит): максимальна сума – 60 балів; мінімальна сума, що дозволяє здобувачу вищої освіти складати екзамен (іспит) – 35 балів.
Підсумковий контроль включає семестровий контроль та атестацію здобувача вищої освіти.
Семестровий контроль проводиться у формах семестрового екзамену (іспиту), диференційованого заліку або заліку. Складання семестрового екзамену (іспиту) здійснюється під час екзаменаційної сесії.
Максимальна сума балів, яку може отримати здобувач вищої освіти під час екзамену (іспиту) – 40 балів. Мінімальна сума, за якою екзамен (іспит) вважається складеним – 25 балів.
Підсумкова оцінка за навчальною дисципліною визначається як
сума балів за поточний та підсумковий контроль.
Під час викладання навчальної дисципліни використовуються наступні контрольні заходи:
Поточний контроль: захист лабораторних робіт (48 бали), письмові контрольні

роботи (12 балів).
Семестровий контроль: Екзамен (40 балів).
Більш детальну інформацію щодо системи оцінювання наведено в робочому плані (технологічній карті) з навчальної дисципліни.

Приклад екзаменаційного білета та критерії оцінювання для навчальної дисципліни.

Приклад екзаменаційного білета

Харківський національний економічний університет імені Семена Кузнеця
Другий (магістерський) рівень вищої освіти
Спеціальність 126 «Інформаційні системи та технології»
Освітньо-професійна програма «Інформаційні системи та технології».
Семестр II
Навчальна дисципліна «Управлінські ІС та сховища даних»

ЕКЗАМЕНАЦІЙНИЙ БІЛЕТ № 1

Завдання 1 (діагностичне, 10 балів).

1)Наведіть різницю між ETL на основі Pandas та ETL на основі DASK.
2) Надайте приклади ETL.

Завдання 2 (евристичне, 30 балів).
1) Згенеруйте 40 текстових файлів розміром 10000000 символів.
2) Зробіть аналіз часу розрахунку кількості символів у всіх файлах засобами Dask.
3) Зробіть аналіз часу розрахунку кількості символів у всіх файлах засобами Pandas.
4) Зробіть зіставлення та аналіз результатів двох підходів.

Затверджено на засіданні кафедри інформаційних систем протокол № 1 від «27» серпня 2024 р.

Екзаменатор к.е.н., доц. Сергій ЗНАХУР

Зав. кафедрою к.т.н., доц. Дмитро БОНДАРЕНКО

Критерії оцінювання

Екзаменаційний білет включає два евристичних завдання. В процесі виконання екзаменаційних завдань використовуються хмарний сервіс Google Cloud та PowerBI. Максимальна кількість – 40 балів; мінімальна, що зараховується – 25 балів. При цьому за повністю правильно виконані завдання студент отримує:

Завдання 1 – 10 балів;

Завдання 2 – 30 балів.

Підсумкові бали за іспитом складаються із суми балів за виконання всіх завдань, що округлені до цілого числа за правилами математики.

Завдання 1 (діагностичне) оцінюється у ***10 балів*** наступним чином:

5 бали – реалізація першого завдання;

5 балів – реалізація другого завдання;

Завдання 2 (евристичне) оцінюється у ***30 балів*** наступним чином:

8 балів – реалізація першого завдання;

8 балів – реалізація другого завдання;

8 балів – реалізація третього завдання;

6 балів – реалізація четвертого завдання;

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Основна

1.Foreman, John W. Data Smart: Using Data Science to Transform Information into Insight / John W. Foreman; БД books24x7. – John Wiley & Sons, 2014. – 432 pages. – ISBN 978-1-118-03496-5.: http://common.books24x7.com/toc.aspx?bookid=58144.

2.Дэви С. Основы Data Science и BigData. Python и наука о данных.//С.Дэви, М.Арно, А.Мохамед — СПб.: Питер, 2017. — 336 е.: ил.

3.Інтелектуальний аналіз даних: Підручник / Черняк О.І., Захарченко П.В./ К.: Знання, 2014. – 599 с.

4. Інформаційні системи та технології: монографія / за заг. ред . В. С. Пономаренка. - Х. : ФОП Бровін О.В., 2019. - 212 с .
http://repository.hneu.edu.ua/handle/123456789/21743

Додаткова

1. Марченко О. О., Россада Т.В. Актуальні проблеми Data Mining: навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. – Київ. – 2017. – 150

2. Завадський І.О. Основи баз даних: [Навч. посіб.] / І.О. Завадський. –К.: Видавець І.О. Завадський, 2011. – 192 с.

Інформаційні ресурси

1.IoT Fundamentals: Big Data & Analytics // Електронний ресурс. Режим доступу: https://www.netacad.com/courses/iot/big-data-analytics
2. Apache Spark // Електронний ресурс. Режим доступу: https://spark.apache.org/
3. Локальний Spark-кластер // Електронний ресурс. Режим доступу: https://r-analytics.blogspot.com/2020/02/spark-r-connect.html
4. Azzalini A., Bruno S. Data Analysis and Data Mining. An Introduction. New York : Oxford University Press, 2012. 289 p. Режим доступу: http://ebooks.znu.edu.ua/files/Bibliobooks/Kudin/0036206.pdf.
5. Gisele L.P., Alex A.F. Automating the Design of Data Mining Algorithms: an Evolutionary Computation Approach. Heidelberg : Springer-Verlag Berlin Heidelberg, 2010. 197 p. Режим доступу: http://ebooks.znu.edu.ua/files/Bibliobooks/Kudin/0036216.pdf.
6. Бази даних та інформаційні системи. Навчальний курс – Режим доступу: http://simulation.kiev.ua/dbis/lection06.html
7. Сайт персональних навчальних систем ХНЕУ ім..С.Кузнеця. Дисципліна «Аналітика великих даних». – Режим доступу: https://pns.hneu.edu.ua/course/view.php?id=9759