Силабус освітнього компонента
Програма навчальної дисципліни

Інженерія великих даних

Шифр та назва спеціальності
122 – Комп'ютерні науки

Інститут
ННІ комп'ютерних наук та інформаційних технологій

Освітня програма
Комп'ютерні науки. Штучний інтелект та управління проєктами

Кафедра
Системний аналіз та інформаційно-аналітичні технології (322)

Рівень освіти
Бакалавр

Тип дисципліни
Профільна, Обов'язкова

Семестр
6

Мова викладання
Українська

Викладачі, розробники

Любченко Наталія Юріївна

Nataliia.Liubchenko@khpi.edu.ua
Кандидат технічних наук, доцент, доцент кафедри системного аналізу та інформаційно-аналітичних технологій

Досвід роботи – 20 років. Автор понад 60 наукових та навчально-методичних праць. Провідний лектор з дисциплін: «Основи програмування», «Інженерія великих даних», «Семантичний WEB», «Об`єктно-орієнтоване програмування», «Основи розподілених та паралельних обчислень».
Детальніше про викладача на сайті кафедри

Загальна інформація

Анотація

Предметом дисципліни є теоретичні та практичні основи оброблення великих даних. Розглядаються загальні методи оброблення великих даних, можливості різних мов програмування для аналізу та візуалізації даних Дисципліна забезпечує теоретичну і практичну підготовку в області паралельних та розподілених обчислень, оволодіння концепціями сучасного програмування в рамках парадигм паралельного та розподіленого програмування.

Мета та цілі дисципліни

Формування у студентів сучасного рівня інформаційної та програмістської культури, оволодіння основними принципами інженерії програмного забезпечення, набуття ними практичних навичок самостійної розробки програмного забезпечення і використання сучасних інформаційних технологій для розв'язання практичних задач, а також надання знань та навичок щодо застосування технологій, концепцій зберігання, обробки та методів аналізу великих даних під час організації корпоративних сховищ даних та розробки програмного забезпечення бізнес-аналітики.

Формат занять

Лекції, лабораторні роботи, самостійна робота, консультації. Підсумковий контроль – залік.
.

Компетентності

ЗК7. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
СК9. Здатність реалізувати багаторівневу обчислювальну модель на основі архітектури клієнт-сервер, включаючи бази даних, знань і сховища даних, виконувати розподілену обробку великих наборів даних на кластерах стандартних серверів для забезпечення обчислювальних потреб користувачів, у тому числі на хмарних сервісах.
СК11. Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач.
СК16. Здатність реалізовувати високопродуктивні обчислення на основі хмарних сервісів і технологій, паралельних і розподілених обчислень при розробці й експлуатації розподілених систем паралельної обробки інформації.

Результати навчання

РН12. Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використанням програмних інструментів підтримки багатовимірного аналізу даних на основі технологій DataMining, TextMining, WebMining.
РН16. Виконувати паралельні та розподілені обчислення, застосовувати чисельні методи та алгоритми для паралельних структур, мови паралельного програмування при розробці та експлуатації паралельного та розподіленого програмного забезпечення.
РН17. Розробляти системи штучного інтелекту на основі використання моделей, методів та засобів інженерії даних та знань

Обсяг дисципліни

Загальний обсяг дисципліни 120 год. (4 кредитів ECTS): лекції – 32 год., лабораторні роботи – 16 год., самостійна робота – 72 год.

Передумови вивчення дисципліни (пререквізити)

Для успішного проходження курсу необхідно мати знання та практичні навички з наступних дисциплін: «Об'єктно орієнтоване програмування», «Бази даних в інформаційних системах», «Технології та засоби машинного навчання».

Особливості дисципліни, методи та технології навчання

На лекційних заняттях викладання матеріалу здійснюється в усній формі із записом основних положень лекції у конспект. Для демонстрації презентацій застосовується медіа проектор та комп'ютер.
На лабораторних заняттях студенти виконують індивідуальні завдання на комп'ютерах у середовищі IntelliJ IDEA, Eclips для роботи с проектами Scala, за допомогою інструментів Oracle JDK, Scala Build Tool (SBT), Scala IDE, Apache Spark.

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Розуміння великих даних.
Набори та аналіз даних. Дескриптивна, діагностична, прогностична та прескриптивна аналітика з використанням великих даних. Бізнес аналітика. Ключові показники ефективності. Характеристики великих даних.

Тема 2. Перехід до великих даних та питання планування.
Організаційні передумови. Набуття даних. Конфіденційність. Безпека. Походження. Обмеження підтримки у реальному часі.
Тема 3. Корпоративні технології і бізнес-аналітика для великих даних.
Обробка транзакцій в режимі реального часу (OLTP). Аналітична обробка у реальному часі (OLAP). Вилучення, перетворення та завантаження (ETL). Сховища даних.
Тема 4. Концепції зберігання великих даних.
Кластери. Файлові системи та розподілені файлові системи. NoSQL. Шардінг. Реплікація. Режим реплікації «ведучий-ведений», режим реплікації «одноранговий».
Тема 5. Концепції обробки великих даних.
Паралельна обробка даних. Розподілена обробка даних. Hadoop. Пакетна обробка за допомогою MapReduce.
Тема 6. Технології зберігання великих даних.
Дискові пристрої зберігання (розподілені файлові системи, системи управління базами даних (СУБД), бази даних NoSQL.
Тема 7. Основні методи аналізу великих даних.
Кількісний аналіз. Якісний аналіз. Data Mining. Статистичний аналіз (A/B тестування, кореляція, регресія). Машинне навчання (класифікація, кластеризація, виявлення викидів, фільтрація). Семантичний аналіз (обробка природної мови, обробка тексту, аналіз емоціонального забарвлення висловлювань). Візуальний аналіз (кольорові карти, часові ряди, мережеві графи).
Тема 8. Введення в розподілені та паралельні обчислення.
Поняття про паралельні та розподілені обчислення. Послідовні обчислення. Паралельні обчислення. Засоби для здійснення паралельних обчислень.
Тема 9. Введення в розподілені розрахунки.
Поняття розподілених обчислень та розподіленої системи. Цілі побудови розподілених систем. Вимоги до розподілених систем: прозорість, відкритість, масштабованість. Складність розробки розподілених систем.
Тема 10. Розподілений аналіз даних і кластерні обчислення. Розподілені обчислення в задачах машинного навчання.
Розгляд задачі розподілення обчислень великих даних. Методики аналізу великих даних. Дослідження принципу та вимог розподілення обчислень.
Тема 11. Обробка робочих завдань (пакетна обробка).
Транзакційна обробка. Кластер. Обробка у пакетному режимі. Інтерпретація алгоритмів MapReduce. Обробка в режимі реального часу.
Тема 12. Мета та задачі паралельної обробки даних. Принципи розробки паралельних методів.
Проблеми використання паралелізму. «Послідовність» існуючих алгоритмів і програмного забезпечення. Складність розробки паралельних алгоритмів.
Тема 13. Моделі обчислень та методи аналізу ефективності.
Загальні принципи побудови паралельних алгоритмів і програм. Показники ефекту розпаралелення (прискорення, продуктивність, ефективність). Способи оцінки показників.
Тема 14. Паралельні алгоритми розв'язку задач.
Матрично-векторне множення, множення матриць, розв'язок систем лінійних рівнянь. Системи лінійних рівнянь. Паралельні алгоритми.

Теми практичних занять

Практичні роботи в рамках дисципліни не передбачені.

Теми лабораторних робіт

Лабораторна робота №1.
Об'єктно-орієнтоване програмування в Scala. Функціональне програмування в Scala.
Лабораторна робота №2.
Програмування операцій зі стійкими розподіленими наборами даних (Resilient Distributed Datasets, RDD).
Лабораторна робота №3.
RDD. Агрегування в Spark.

Лабораторна робота №4.
Створення та робота з DataFrames з використанням Spark.
Лабораторна робота №5.
Робота з графами з використанням Spark GraphX.
Лабораторна робота №6.
Реалізація алгоритмів машинного навчання з використанням Spark MLlib і пакета spark.ml..
Лабораторна робота №7.
Введення в аналіз даних за допомогою Scala і Spark.
Лабораторна робота №8.
Модель розподілених обчислень MapReduce.
Рішення задач MapReduce в Spark.

Самостійна робота

Опрацьовування лекційного матеріалу.
Підготовка до лабораторних занять та модульних контролів.
Самостійне вивчення тем та питань, які не викладаються на лекційних заняттях.

Література та навчальні матеріали

Основна література

1. Олещенко Л.М. Технології оброблення великих даних: навч. посіб. / Л.М. Олещенко. –Київ: КПІ ім. Ігоря Сікорського, 2021. – 227 с.
https://ela.kpi.ua/bitstream/123456789/42206/1/%D0%9AonspLekts_Tekhnolohii-obroblennia-velykykh-danykh_%D0%9Eleshchenko.pdf
2. Таран В. І. Технології Big Data: навч. посіб. / В. І. Таран, Ю. Г. Гордієнко, С. Г. Стіренко. – Київ КПІ ім. Ігоря Сікорського, 2022. – 56 с.
https://ela.kpi.ua/bitstream/123456789/50014/1/BigData_praktykum.pdf
3. Jules S. Damji. Learning Spark. Lightning-Fast Data Analytics. Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee / Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. - 2020. – 399 p. https://pages.databricks.com/rs/094-YMS-629/images/LearningSpark2.0.pdf
4. Spark by Robert D. Schneider and Jeff Karmiol / Published by John Wiley & Sons, Inc. 111 River St. Hoboken, NJ 07030-5774. – 2019. — 80 p. https://www.ibm.com/downloads/cas/WEB4XBOR
5. Dean Wampler Programming Scala / Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. – 2021. – 556 p. https://books-library.net/files/books-library.net-10211445Tf3F5.pdf

Додаткова література

1. Любченко Н. Ю. Основи мови програмування SCALA [Електронний ресурс] : навч.-метод. посібник / Н. Ю. Любченко, А. О. Подорожняк, О. П. Черних ; Нац. техн. ун-т "Харків. політехн. ін-т". – Електрон. текст. дані. – Харків, 2023. – 148 с. https://repository.kpi.kharkov.ua/items/3f5ee53b-258b-472a-ba35-3fdc412bfbed
2. Big Data [Електронний ресурс]. Режим доступу: https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye
3. Проект Apache Hadoop [Електронний ресурс]. – Режим доступу: https://hadoop.apache.org/
4. Unified engine for large-scale data analytics [Електронний ресурс]. – Режим доступу: https://spark.apache.org/
5. Big Data (Біг Дата / Великі дані) – що це таке, суть, визначення, як працює та навіщо потрібно [Електронний ресурс]. Режим доступу: https://termin.in.ua/big-data-velyki-dani/

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням:
модульні тести - 20 балів; лабораторні роботи - 60 балів; залік - 20 балів

Шкала оцінювання

Сума балів	Національна оцінка	ECTS
90–100	Відмінно	A
82–89	Добре	B
75–81	Добре	C
64–74	Задовільно	D
60–63	Задовільно	E
35–59	Незадовільно (потрібне додаткове вивчення)	FX
1–34	Незадовільно (потрібне повторне вивчення)	F

Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту. Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Погодження

Силабус погоджено	29.08.2023	Завідувач кафедри Юрій ДОРОФЄЄВ
	29.08.2023	Гарант ОП Марина ГРИНЧЕНКО