Силабус освітнього компонента
Програма навчальної дисципліни

Технології Big Data

Шифр та назва спеціальності
122 – Комп’ютерні науки

Інститут
ННІ Комп'ютерного моделювання, прикладної фізики та математики

Освітня програма
Комп’ютерні науки. Моделювання, проектування та комп’ютерна графіка

Кафедра
Комп'ютерне моделювання процесів та систем (162)

Рівень освіти
Бакалавр

Тип дисципліни
Профільна підготовка, Вибіркова

Семестр
8

Мова викладання
Українська

Викладачі, розробники

Багмут Іван Олександрович
(відповідальний лектор)
ivan.bagmut@khpi.edu.ua
кандидат технічних наук, доцент

Спеціаліст у галузі науки про дані, машинного навчання та штучного інтелекту. Автор понад 40 наукових статей і матеріалів доповідей, співавтор методичних посібників.

Детальніше про викладача на сайті кафедри

Овсяніков Владислав Валерійович
(асистент з лабораторних робіт)
vladyslav.ovsianikov@khpi.edu.ua
Аспірант, Senior Big Data Engineer в компанії EPAM

Детальніше про викладача на сайті кафедри

Загальна інформація

Анотація

Дисципліна “Технології Big Data” спрямована на вивчення підходів, методів і механізмів функціонування та використання інфраструктури для розподілених обчислень на базі кластеру Hadoop та парадигми MapReduce. Необхідність в використанні нових підходів обумовлена тим, що сучасні підходи до вирішення складних завдань, які потребують обробки надзвичайно великого обсягу даних, потребують використання великої кількості обчислювальних ресурсів. Вивчення даної дисципліни майбутніми фахівцями дозволить їм набути важливих компетенцій в плані розвитку існуючих і використанню нових підходів для організації розподілених обчислень.

Мета та цілі дисципліни

Підготовка фахівців, здатних розв'язувати комплексні проблеми у сфері обробки "великих даних" та використовувати сучасні засоби для організації обчислень в розподілених системах.

Формат занять

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

Компетентності

Здатність використовувати сучасні інфраструктури для розподілених обчислень.
Здатність створювати та використовувати програмне забезпечення для розподілених обчислень.
Прогнозувати вплив і ефект застосовуваних методів, технічних засобів і технологій Big Data.
ЗК2: Здатність застосовувати знання у практичних ситуаціях.
ЗК3: Знання та розуміння предметної області та розуміння професійної діяльності.
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.
СК19: Здатність застосовувати сучасні математичні концепції та алгоритмічні стратегії у сфері штучного інтелекту та машинного навчання для розробки новітніх моделей та систем, які здатні ефективно аналізувати, інтерпретувати, обробляти та використовувати складні дані, орієнтуючись на розширення та вдосконалення існуючих методів та технологій штучного інтелекту

Результати навчання

Мати передові концептуальні та методологічні знання у сфері обробки та аналізу великих обсягів даних.
Мати методологічні знання в плані застосування сучасних підходів та засобів для організації обчислень у розподілених обчислювальних системах.
Розробляти програмне забезпечення для обробки великих даних у розподілених обчислювальних системах.
ПР10: Використовувати інструментальні засоби розробки клієнт-серверних застосувань, проектувати концептуальні, логічні та фізичні моделі баз даних, розробляти та оптимізувати запити до них, створювати розподілені бази даних, сховища та вітрини даних, бази знань, у тому числі на хмарних сервісах, із застосуванням мов веб-програмування.
ПР12: 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.
ПР16: 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.
ПР20: Застосовувати вдосконалені математичні та алгоритмічні знання в області штучного інтелекту для створення інноваційних моделей та систем, які спроможні комплексно аналізувати та інтерпретувати складні та багатовимірні дані, відкриваючи нові можливості для поліпшення та оптимізації інтелектуальних технологій

Обсяг дисципліни

Загальний обсяг дисципліни 120 год. (4 кредити ECTS): лекції – 20 год., практичні заняття – 10 год., самостійна робота – 90 год.

Передумови вивчення дисципліни (пререквізити)

Для успішного освоєння курсу необхідні знання набути в дисциплінах "Алгоритмізація та програмування", "Архітектура обчислювальних систем", "Дискретна математика", "Організація баз даних", "Комп'ютерні мережі та розподілені обчислення", "Інтелектуальний аналіз даних".

Особливості дисципліни, методи та технології навчання

Основною мовою програмування, що використовується в рамках курсу, є мова Python. Доступ до програми хмарних обчислень AWS Educate, що надає безкоштовні можливості для використання студентами.
Лекції ведуться з активним використанням мультимедійних ресурсів та інтерактивних методів навчання, що включає аналіз прикладів, кейсів та реальних проектів. Матеріал подається через демонстраційний підхід з акцентом на систематизацію ключових концепцій, а також залучення студентів до дискусій та критичного аналізу.
Навчальні матеріали доступні студентам на Microsoft OneDrive.

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Основні поняття обробки даних. Методи та засоби збирання та зберігання даних.
Структура дисципліни та РСО. Поняття даних, основні завдання обробки даних, особливості обробки даних. Класифікація та загальний огляд етапів та методів обробки даних. Основні етапи процесу роботи з великими даними: постановка задачі, визначення даних, фільтрація, видобування, валідація та підготовка, аналіз, візуалізація. Види систем для роботи з великими даними. Джерела великих даних, питання приватності та безпеки. Засоби збереження великих даних. Хмарні сховища, розподілені сховища. Складові фреймворку для роботи з великими даними.

Тема 2. Фреймворк Hadoop.
Основні особливості Hadoop, переваги та недоліки. Задачі, які вирішує Hadoop. Складові фреймворку Hadoop. Структура HDFS(Hadoop Distributed File System). Файлова система Hadoop, операції читання та збереження файлів. Map Reduce, його функції, основні операції. Завдання та планування завдань в Map Reduce. Менеджер ресурсів YARN. Інструменти для написання запитів Pig та Hive. Інструменти для реалізації алгоритмів машинного навчання Mahout. Інструмент для роботи з графами Giraph.

Тема 3. Методи машинного навчання для обробки великих даних.
Основні види машинного навчання. Етапи аналізу даних методами машинного навчання. Масштабування ознак. Вибір моделі та способу її навчання. Методи оцінки моделей. Види методів класифікації, їх особливості, переваги та недоліки. Сфери застосування методів класифікації. Лінійна, нелінійна та логістична регресії. Критерії вибору та перевірки регресійної моделі. Додаткові регресійні моделі. Класифікація алгоритмів кластерного аналізу. Ієрархічні алгоритми. Метод k-середніх. Нечіткі методи кластерного аналізу. Концепції та категорії глибинного навчання. Основні моделі глибинного навчання. Паралельна оптимізація для глибинного навчання.

Тема 4. Системи для аналізу великих даних в реальному часі. Методи та засоби візуалізації великих даних.
Основні концепції, характеристики та платформи для обробки великих даних в реальному часі. Поняття події, потоку подій, потокової обробки. Основні платформи для потокової обробки даних, їх особливості. Фреймворк Spark. Особливості платформ Storm та Kafka, їх інтеграція з Hadoop. Основні задачі, пов’язані з аналітикою даних в соціальних мережах. Методи визначення мови. Інтелектуальний аналіз тексту. Визначення трендових тем. Побудова рекомендаційних систем. Виявлення аномалій. Типи ресурсів. Основні методи та платформи для управління ресурсами. Управління ресурсами з одного та з багатьох джерел в хмарі. Методи візуального аналізу даних.

Графіки числових рядів. Представлення мережі у вигляді графа. Обробка графів. Аналіз просторових даних для задачі таргетування. Недоліки пакетної обробки даних в Hadoop. Метод повтору. Організація завдань в методі повтору. Алгоритми пакетної обробки. Аналіз продуктивності алгоритмів. Приватність даних з соціальних мереж. Політики диференційованої приватності. Безпека великих даних. Шифрування великих даних.

Теми практичних занять

Не передбачено навчальним планом

Теми лабораторних робіт

Тема 1. Hadoop та MapReduce.
Встановлення та налаштування Hadoop. Виконання базових завдань в MapReduce.

Тема 2. Обробка даних.
Попередня обробка даних. Обробка запитів. Статистична обробка даних.

Тема 3. Алгоритми в Mahout
Алгоритми класифікації в Mahout. Алгоритми регресії в Mahout. Алгоритми кластеризації в Mahout. Глибинне навчання для великих даних.

Тема 4. Обробка великих даних в реальному часі. Візуалізація великих даних.
Обробка великих даних в реальному часі. Обробка даних в фреймворку Spark. Візуалізація великих даних.

Самостійна робота

Самостійне опрацювання окремих тем дисципліни призначено для закріплення знань, умінь та навичок, отриманих студентами в ході освоєння лекційного матеріалу курсу на лабораторних заняттях. Вивчення додаткових тем курсу, що не викладається в рамках лекційного курсу, але необхідних для майбутнього Big Data розробника.

Література та навчальні матеріали

1. Zgurovsky M.Z., Zaychenko Y.P. Big Data: Conceptual Analysis and Applications. Springer, 2020. – 298 p.
2. Олійник А. О. Інтелектуальний аналіз даних : навчальний посібник. — Запоріжжя : ЗНТУ, 2012. — 278 с.
3. Rajkumar Buyya. Big Data. Principles and Paradigms. — Elsevier, 2016. – 496p.
4. Cloud computing / N. B. Ruparelia. – Cambridge; London: The MIT Press, 2016. – 260 c. – (The MIT Press essential knowledge series)
5. NoSQL: database for storage and retrieval of data in cloud / Ed. by G. C. Deka. – Boca Raton [etc.]: CRC Press: Taylor & Francis Group, 2017. – 455 c

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

Опрацювання питань, що винесені на самостійну роботу, оцінюється лектором на заліку наприкінці навчального семестру.

Бали нараховуються за наступним співвідношенням:
• лабораторні роботи: 70% семестрової оцінки
• залік: 30% семестрової оцінки

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	28.08.2023	Завідувач кафедри Дмитро БРЕСЛАВСЬКИЙ
	28.08.2023	Гарант ОП Оксана ТАТАРІНОВА