Силабус освітнього компонента
Програма навчальної дисципліни

Інформаційні технології аналізу великих гетерогенних даних

Шифр та назва спеціальності
113 – Прикладна математика

Інститут
ННІ Комп'ютерних наук та інформаційних технологій

Освітня програма
Інтелектуальний аналіз даних

Кафедра
Комп'ютерна математика і аналіз даних

Рівень освіти
Магістр

Тип дисципліни
Профільна, Вибіркова

Семестр
2

Мова викладання
Українська

Викладачі, розробники

Любчик Леонід Михайлович
Leonid.Liubchyk@khpi.edu.ua
Доктор технічних наук, професор, академік Академії наук вищої школи України, Лауреат Державної премії України, професор кафедри комп'ютерної математики і аналізу даних НТУ «ХПІ». Досвід роботи з 1981 року. Кількість наукових та навчальних публікацій понад 200. Провідний лектор з дисциплін: «Випадкові процеси і стохастичні системи», «Теорія керування», «Некоректні задачі обробки даних», «Прогнозний аналіз» Наукові напрямки: керування та прийняття рішень в умовах невизначеності, машинне навчання.
https://scholar.google.com/citations?user=Jn_UBfEAAAAJ&hl=en
https://www.scopus.com/authid/detail.uri?authorId=24723278200
https://orcid.org/0000-0003-0237-8915
https://ua.linkedin.com/in/leonid-lyubchyk-8a60071b5
Детальніше про викладача на сайті кафедри

Ямковий Клим Сергійович
klym.yamkovyi@cs.khpi.edu.ua
PhD, асистент
Кількість наукових та навчальних публікацій понад 10.

https://scholar.google.com/citations?user=P7suw-4AAAAJ&hl=en
https://www.scopus.com/authid/detail.uri?authorId=57204824595

Ковальов Олексій Миколайович
Oleksii.M.Kovalov@cs.khpi.edu.ua
Аспірант
https://orcid.org/my-orcid?orcid=0009-0007-1484-4334

Загальна інформація

Анотація

Курс зосереджується на вивченні ключових аспектів інфраструктури та інструментів для роботи з Великими Гетерогенними Даними. У рамках курсу особлива увага приділяється Apache Hadoop, провідній платформі для обробки Великих Даних, включаючи її важливі компоненти, такі як MapReduce, Spark, HBase, Hive і Pig, а також мови програмування Pig Latin і Hive, які використовуються для обробки даних.
Також курс охоплює важливі аспекти безпеки та відповідності даних галузевим стандартам, з особливим акцентом на вимоги Загального Регламенту Захисту Даних ЄС (GDPR). Це забезпечує студентам глибоке розуміння не тільки технічних засад обробки Великих Даних, але й важливості їх безпечного та відповідального використання у відповідності до законодавчих вимог.

Мета та цілі дисципліни

Мета курсу полягає в засвоєнні ключових концепцій і технологій у сфері великих даних. Цілі курсу включають:
Ознайомлення з основними поняттями Великих Даних і сучасними технологічними рішеннями в цій області. Розвиток здатності оцінювати та вибирати сервіси інфраструктури Великих Даних від провідних провайдерів хмарних послуг, таких як Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP) та інших. Знайомство с засобами завантаження та обробки гетерогенних даних та ETL. Набуття практичних навичок управління та аналізу даних для бізнес-задач, включаючи вибір та налаштування кластерів Hadoop або Spark на одній з хмарних платформ (наприклад, Azure HDInsight або Amazon EMR). Розвиток навичок програмування для обробки даних, використовуючи скриптові мови програмування, такі як HiveQL та Pig Latin.

Формат занять

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

Компетентності

ЗК 3. Здатність до безперервного навчання, придбання нових знань і умінь, у тому числі в галузі, відмінній від професійної.
ЗК 4. Здатність виявляти, ставити та вирішувати проблеми у професійній діяльності.
ЗК 5. Здатність генерувати нові ідеї (креативність) і нестандартні підходи до їхньої реалізації, гнучке адаптування до реальних професійних ситуацій, проявляти творчий підхід, ініціативу.
ЗК 6. Здатність критично оцінювати й переосмислювати накопичений досвід (власний і чужий), аналізувати свою професійну і соціальну діяльність.
ЗК 7. Здатність працювати з інформацією: знаходити і використовувати інформацію з різних джерел, потрібну для розв'язання професійних завдань.
ЗК 11. Здатність до соціальної і професійної взаємодії та співпраці у колективі, командної роботи.
СК 2. Здатність обирати, розробляти та досліджувати математичний аналітичний або чисельний метод розв'язання практичних задач, що забезпечує потрібні точність і надійність результату.
СК 5. Здатність до проведення математичного і комп'ютерного моделювання та обчислювального експерименту, збору, візуалізації, аналізу та обробки отриманих даних, розв'язання формалізованих задач за допомогою спеціалізованих програмних засобів.

СК 7. Здатність до пошуку, вивчення та аналізу науково-технічної інформації, вітчизняного і закордонного досвіду, пов'язаного із застосуванням математичних методів для дослідження процесів та систем.
СК 13. Здатність до розробки та експлуатації спеціалізованих програмних засобів обробки великих масивів даних на основі інформаційних технологій розподілених і хмарних обчислень.
СК 14. Здатність до використання сучасних інформаційних технологій інтелектуального аналізу даних, прогнозування, прийняття рішень, інформаційного пошуку і видобування знань.

Результати навчання

РН 1. Демонструвати знання і розуміння основних концепцій, принципів, теорій фундаментальної та прикладної математики і використовувати їх на практиці.
РН 2. Уміти формалізувати задачі, сформульовані мовою певної предметної галузі й обирати раціональний метод вирішення; розв'язувати задачі аналітичними або чисельними методами, оцінювати точність і достовірність отриманих результатів та виконувати їхню інтерпретацію.
РН 4. Уміти поєднувати методи математичного і комп'ютерного моделювання з неформальними процедурами експертного аналізу для пошуку оптимальних рішень.
РН 7. Уміти застосовувати сучасні технології програмування та розроблення програмного забезпечення, програмної реалізації чисельних і символьних алгоритмів.
РН 8. Уміти застосовувати у практичній роботі спеціалізовані програмні продукти і програмні системи комп'ютерної математики, аналізу великих даних тощо.
РН 13. Знати і розуміти методи розв'язання математичних задач інтелектуального інформаційного пошуку та видобування знань.

Обсяг дисципліни

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

Передумови вивчення дисципліни (пререквізити)

Для успішного проходження курсу необхідно мати знання та практичні навички з дисциплін профільного пакету ВП*.1, а саме, ВП1.1 – «Методи та технології роботи з великими даними», або ВП2.1 – «Аналіз і синтез природньомовної інформації». А також, необхідно мати знання та практичні навички з дисциплін вільного вибору (ДВВ) профільної підготовки, які викладались в першому семестрі.

Особливості дисципліни, методи та технології навчання

Лабораторні роботи виконуються на реальній хмарній платформі та кластері Hadoop (AWS або Azure).

Програма навчальної дисципліни

Теми лекційних занять

Лекція 1.
Основи хмарних технологій. Історія розвитку.
Лекція 2.
Моделі хмарних сервісів, хмарні ресурси, функціонування хмарних служб
Лекція 3.
Еталонна архітектура Великих Даних та приклади використання.
Лекція 4.
Хмарні платформи для великих даних. Огляд та порівняння.
Лекція 5.
Екосистема Hadoop для Великих Даних. HDFS, HBase, MapReduce, YARN.
Лекція 6.
Компоненти екосистеми Hadoop для обробки Великих Даних: Hadoop Data Warehouse Hive, data flow processing with Pig

Лекція 7.
Основи SQL, реляційні бази даних.
Лекція 8.
Типи та огляд баз даних NoSQL.
Лекція 9.
Сучасні великомасштабні бази даних AWS Aurora, Azure CosmosDB, Google Spanner.
Лекція 10.
Потоки даних та потоковий аналіз даних. Архітектура та компоненти Spark.
Лекція 11.
Платформи для Spark, DataBricks.
Лекція 12.
Концепція Відкритих Даних (Open Data), набори даних для аналізу даних та машинного навчання.
Лекція 13.
Гетерогенні дані. Інструменти для завантаження та обробки
Лекція 14.
Гетерогенні дані. Очищення, стандартизація, інтеграція.
Лекція 15.
Архітектура великих даних підприємства та управління великими даними.
Лекція 16.
Проблеми безпеки Великих Даних, захист даних. Контроль доступу та управління ідентичністю.

Теми практичних занять

Лабораторне заняття 1.
Основні провайдери хмарних послуг AWS, Microsoft Azure, Google Cloud Platform
Лабораторне заняття 2.
Робота з хмарою Amazon Web Services(AWS).
Лабораторне заняття 3.
Розгортання та доступ до екземплярів EC2, S3, VM
Лабораторне заняття 4.
Конфігурація клієнта SSH і доступ до VM.
Лабораторне заняття 5.
Виконання простого завдання з MapReduce.
Лабораторне заняття 6.
Установка та налаштування автономного кластера Hadoop для особистого користування.
Лабораторне заняття 7.
Ознайомлення з інтерфейсом Hue, завантаження даних і файлів.
Лабораторне заняття 8.
Робота з SQL, базові команди
Лабораторне заняття 9.
Ознайомлення з хмарними сервісами AWS RDS для створення власної реляційної бази даних.
Лабораторне заняття 10.
Робота з учбовим онлайн кластером Databrick.
Лабораторне заняття 11.
Приклад Web scrapping, робота з простим скриптом python.
Лабораторне заняття 12.
Інструменти для завантаження гетерогенних даних. ETL.
Лабораторне заняття 13.
Управління проєктами щодо Великих Даних.
Лабораторне заняття 14.
Управління життєвим циклом даних та DataOps.
Лабораторне заняття 15.
Визначення корпоративної інфраструктури Великих Даних та сервісів обробки даних.
Лабораторне заняття 16.
Вибір інфраструктурних послуг та компонентів.

Теми лабораторних робіт

N/A

Самостійна робота

Віртуальний гібридний/ динамічний хмарний центр обробки даних
Хмарний аутсорсинг IT-інфраструктури підприємства.
Інфраструктура Великих Даних та її компоненти.
Великі Дані в промисловості та концепція Industry 4.0.
Знайомство з хмарними платформами: AWS, Microsoft Azure, Google Cloud Platform (GCP)
Застосування алгоритмів MapReduce та технологій розподілених сховищ даних.
Вивчення бібліотеки Java MapReduce.
Огляд AWS Elastic Map Reduce (EMR).
Хмарні сервіси AWS для роботи з даними; хмарне зберігання даних та сервіси баз даних.
Популярні платформи для Spark, DataBricks.
Програмування для Spark.
Модель зрілості управління даними та аспекти забезпечення якості даних.
План управління даними (DMP). Принципи ефективності даних FAIR (Findable – Accessible – Interoperable – Reusable).
Інструменти ETL
Управління даними в промисловості, забезпечення довіри до даних та суверенітету даних.
Хмарні стандарти відповідності та оцінка послуг провайдера хмарних послуг.

Література та навчальні матеріали

1. Л.М. Олещенко. Технології оброблення великих даних. Навчальний посібник. 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. B. Schoenborn. Big Data Analytics Infrastructure For Dummies, IBM Limited Edition Published by John Wiley & Sons, Inc. 2014. https://www.ibm.com/downloads/cas/AYWDRYLW
3. M. Collier, R. Shahan. Microsoft Azure Essentials: Fundamentals of Azure, Second Edition. Microsoft Press, 2016. – 246 p. https://download.microsoft.com/download/6/6/2/662DD05E-BAD7-46EF-9431-135F9BAE6332/9781509302963_Microsoft%20Azure%20Essentials%20Fundamentals%20of%20Azure%202nd%20ed%20pdf.pdf
4. IoT Fundamentals: Big Data & Analytics // Електр. ресурс. Режим доступу: https://www.netacad.com/courses/iot/big-data-analytics
5. Apache Hadoop // Електронний ресурс. Режим доступу: http://hadoop.apache.org/
6. MapReduce Tutorial // Електронний ресурс. Режим доступу: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
7. Apache Spark // Електронний ресурс. Режим доступу: https://spark.apache.org/
8. Lidong Wang, Heterogeneous Data and Big Data Analytics. Automatic Control and Information Sciences. 2017, 3(1), 8-15. DOI: 10.12691/acis-3-1-3. https://pubs.sciepub.com/acis/3/1/3/

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

Студент зобов'язаний відвідувати всі заняття згідно розкладу. Без особистої присутності студента підсумковий контроль не проводиться. Бали студента з дисципліни нараховуються за наступним співвідношенням:

- контрольні роботи: 40% семестрової оцінки;
- самостійна робота: 20% семестрової оцінки;
- іспит: 40% 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	Дата погодження, підпис 31.08.2023 р.		Завідувач кафедри Олена АХІЄЗЕР
	Дата погодження, підпис 31.08.2023 р.		Гарант ОП Олексій ГАЛУЗА