Силабус освітнього компонента
Програма навчальної дисципліни

Технології обробки великих обсягів даних

Шифр та назва спеціальності
124 – Системний аналіз

Інститут
ННІ Комп'ютерних наук та інформаційних технологій

Освітня програма
Системний аналіз і управління

Кафедра
Системного аналізу та інформаційно-аналітичних технологій

Рівень освіти
Магістр

Тип дисципліни
Спеціальна (фахова), вибіркова

Семестр
2

Мова викладання
Українська

Викладачі, розробники

Колбасін Вячеслав Олександрович
viacheslav.kolbasin@khpi.edu.ua
Кандидат технічних наук, доцент, доцент кафедри системного аналізу та інформаційно-аналітичних технологій НТУ "ХПІ"

Досвід роботи – 20 років. Автор понад 40 наукових та навчально-методичних праць. Провідний лектор з дисциплін: «Програмування та підтримка веб-застосувань», «Платформи корпоративних інформаційних систем», «Обробка великих обсягів даних у корпоративних системах», «Технології обробки великих обсягів даних». Має професійні сертифікації: AWS Certified Solutions Architect – Associate, AWS Certified Machine Learning – Specialty, Oracle Certified Associate Java SE7, Oracle Certified Professional Java SE7
Детальніше про викладача на сайті кафедри

Загальна інформація

Анотація

Дисципліна спрямована на опанування студентами платформ та фреймворків, що використовуються при обробці великих обсягів даних . Розглянуто фреймворки Apache Hadoop, Apache Hive, Apache Spark, основи NoSQL баз даних та архітектури, що застосовуються при обробці великих обсягів даних, типові задачі інтелектуальної обробки великих обсягів даних. В ході лабораторних робіт та розрахункової роботи студентами виконується аналіз великих обсягів даних польотів цивільної авіації.

Мета та цілі дисципліни

Мета викладання дисципліни полягає в формування у студентів теоретичних знань і практичних навичок застосувань технологій обробки великих обсягів даних для вирішення прикладних та наукових задач за допомогою класичних технологій BigData, таких як Apache Hadoop, Apache Hive, Apache Spark, NoSQL баз даних та засобів машинного навчання.

Формат занять

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

Компетентності

ЗК1. Здатність до абстрактного мислення, аналізу та синтезу.
ЗК3. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
ЗК4. Здатність спілкуватися з представниками інших професійних груп різного рівня (з експертами з інших галузей знань/видів економічної діяльності).
СК2. Здатність проєктувати архітектуру інформаційних систем
СК3. Здатність розробляти системи підтримки прийняття рішень та рекомендаційні системи.
СК6. Здатність застосовувати теорію і методи Data Science для здійснення інтелектуального аналізу даних з метою виявлення нових властивостей та генерації нових знань про складні системи.
СК11. Вміння використовувати моделі та методи Data Mining для розв'язання задач інтелектуального аналізу даних.

Результати навчання

РН1. Спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері системного аналізу та інформаційних технологій і є основою для оригінального мислення та проведення досліджень.
РН6. Застосовувати методи машинного навчання та інтелектуального аналізу даних, математичний апарат нечіткої логіки, теорії ігор та розподіленого штучного інтелекту для розв'язання складних задач системного аналізу.
РН7. Розробляти інтелектуальні системи в умовах слабо структурованих даних різної природи.
РН8. Здійснювати ідентифікацію та оцінювання параметрів математичних моделей об'єктів керування.

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 вивчення дисципліни необхідно мати знання та практичні навички з дисципліни "Обробка даних засобами Python".

Особливості дисципліни, методи та технології навчання

Лекції проводяться з використанням мультимедійних технологій. Для виконання лабораторних робіт може використовуватись хмарне середовище або віртуальна машина, образ якої є в навчальних матеріалах.
Навчальні матеріали доступні студентам через OneDrive кафедри.

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Вступ.
Задача обробки великих обсягів даних. Технології обробки великих обсягів даних. Горизонтальне та вертикальне масштабування.
Тема 2. Розподілене зберігання та обробка великих обсягів даних.
Файлова система HDFS. Консольні операції та API для роботи з файлами в HDFS.
Тема 3. Підхід до паралельної обробки даних та фреймворк Map-Reduce.

Алгоритми паралельної обробки даних. Структура фреймворку Map-Reduce. Реалізація простих задач обробки даних.

Тема 4. Практичне використання Map-Reduce підходу.

Оптимізація Map-Reduce задач. Рядкові та колонкові формати файлів. Управління сортуванням та передачею даних між вузлами.

Тема 5. Управління кластером.

Система управління кластером YARN. Запуск задач та управління їх виконанням за допомогою YARN. Інші фреймворки управління кластером.

Тема 6. Фреймворк Apache Hive.

Принципи побудови та архітектура Apache Hive. Виконання SQL запитів за допомогою Hive. Обмеження та особливості SQL в Hive. Направлені ациклічні графи (DAG) та фреймворк Tez.

Тема 7. Фреймворк Apache Spark.

Принципи побудови та архітектура Apache Spark. Зберігання даних у RDD. Конвеєр обробки даних у Spark. Створення простих задач у Spark.

Тема 8. Робота за даними у Spark.

RDD, DataFrame, DataSet. Використання Java та Python у Spark. Модель пам'яті Python застосування. Оптимізація Spark застосувань.

Тема 9. Spark SQL.

Виконання запитів Spark SQL. Особливості використання Spark SQL. Засоби оптимізації Spark SQL запитів.

Тема 10. Обробка потокових даних.

Архітектура та засоби обробки потокових даних. Система обміну повідомленнями Apache Kafka. Робота з потоковими даними в Spark Streaming.

Тема 11. Бази даних для роботи з великими обсягами даних.

Концепція та різновиди NoSQL баз даних. Орієнтована на колонки БД Apache Cassandra. Документо-орієнтована БД MongoDB. New SQL бази даних.

Тема 12. Організація та моделювання даних.

Рівні моделювання даних. Багатовимірні моделі (multi dimensional data models). Обробка даних, що змінюються (Slowly Changing Dimensions, SCD).

Тема 13. Архітектура систем обробки даних.

Типовий процес обробки. Data warehouse та datamart. Архітектура Data Lake. Лямбда-архітектура.

Тема 14. Інтелектуальна обробка даних.

Засоби інтелектуальної обробки даних в Apache Spark. Технології машинного навчання у Spark.

Тема 15. Обробка текстових даних.

Моделі текстових даних – модель мішка слів та векторні моделі. Тематичні моделі тексту. Аналіз текстів за допомогою тематичних моделей.

Тема 16. Рекомендаційні системи.

Рекомендаційні системи. Колаборативна фільтрація. Засоби Spark для реалізації колаборативної фільтрації.

Теми практичних занять

Практичні заняття в рамках дисципліни не передбачені

Теми лабораторних робіт

Тема 1. Встановлення та ознайомлення з віртуальною машиною для роботи з BigData технологіями. Програмний доступ до файлів в HDFS.

Тема 2. Пошук слів найбільшої довжини за допомогою Map-Reduce. Дослідження впливу формату файлу на швидкість обробки даних та обсяг зайнятої дискової пам'яті.

Тема 3. Обчислення статистичних показників по датасету за допомогою Apache Hive.

Тема 4. Пошук найдовшого слова за допомогою Apache Spark.

Тема 5. Обчислення статистичних показників по датасету з використанням Apache Spark SQL.

Тема 6. Обчислення статистики по продажах мережі роздрібної торгівлі у режимі реального часу з використанням Apache Spark Streaming.

Тема 7. Побудова моделі даних для обліку польотів цивільної авіації та обліку вильотів пілотів. Використання архітектури Data Lake.

Тема 8. Побудова простої рекомендаційної системи.

Самостійна робота

Дисципліна передбачає виконання індивідуальної курсової роботи щодо аналізу статистичних даних по індивідуально заданому зрізу датасету польотів цивільної авіації США. Результат роботи оформлюється у письмовий звіт.
Студентам рекомендуються додаткові матеріали для самостійного ознайомлення та вивчення.

Література та навчальні матеріали

Основна література

1. Олещенко Л.М. Технології оброблення великих даних. Конспект лекцій. Київ: КПІ ім. Ігоря Сікорського, 2021. 227 с. URL: https://ela.kpi.ua/bitstreams/d05e72c9-26d1-41ad-bc6f-97f88d8e0938/download
2. Warren J., Marz N. Big Data. Manning, 2015. 328 p. URL: https://www.manning.com/books/big-data
3. Holmes A. Hadoop in Practice, 2nd Edition. Manning, 2014. 512 p. URL: https://www.manning.com/books/hadoop-in-practice-second-edition
4. Perrin J. Spark in Action, 2nd Edition. Manning, 2020. 576 p. URL: https://www.manning.com/books/spark-in-action-second-edition
5. Scott D., Gamov V., Klein D. Kafka in Action, 2022. 272 p. URL: https://www.manning.com/books/kafka-in-action
6. Wilson B. Machine Learning Engineering in Action. Manning, 2022. 576 p. URL: https://www.manning.com/books/machine-learning-engineering-in-action
7. Falk K. Practical Recommender Systems, Manning. 2019. 432 p. URL: https://www.manning.com/books/practical-recommender-systems

Додаткова література

1. Chitturi P. Apache Spark for Data Science Cookbook. Packt publishing, 2016. 392 p. URL: https://www.packtpub.com/en-us/product/apache-spark-for-data-science-cookbook-9781785880100?srsltid=AfmBOooILFZsGdnwL_KL9fxDtz258XrXGdiKxT-4CR1XbF78-io0Pqhf
2. Shapira G., Palino T., Sivaram R. Kafka: The Definitive Guide, 2nd edition. O`Reilly Media, 2021. 485p. URL: https://www.oreilly.com/library/view/kafka-the-definitive/9781491936153/
3. Banik R. Hands-On Recommendation Systems with Python. Packt publishing, 2018. 146 p. URL: https://www.oreilly.com/library/view/hands-on-recommendation-systems/9781788993753/
4. Apache Hadoop documentation. URL: http://hadoop.apache.org/docs/current/.
5. Apache Hive Language Manual. URL: https://cwiki.apache.org/confluence/display/Hive/LanguageManual
6. Apache Spark documentation. URL: https://spark.apache.org/docs/latest/.
7. Kafka Documentation, URL: https://kafka.apache.org/documentation/

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

Для оцінки роботи студентів протягом семестру підсумкова оцінка розраховується як середньо-зважена сума оцінок за контрольні заходи (максимальна сума –200 балів):
а) виконання контрольної роботи № 1: максимальна оцінка – 35 балів, вага оцінки – 17.5% кредитів дисципліни);
б) виконання контрольної роботи № 2: максимальна оцінка – 35 балів, вага оцінки – 17.5% кредитів дисципліни);
в) виконання лабораторних робіт: максимальна оцінка – 80 балів, вага оцінки – 40% кредитів дисципліни);
г) виконання розрахункового завдання: максимальна оцінка – 50 балів, вага оцінки – 25% кредитів дисципліни).

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	25.08.2024	Завідувач кафедри Юрій ДОРОФЄЄВ
	25.08.2024	Гарант ОП Валерій СЕВЕРИН