СИЛАБУС
НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

«Обробка великих обсягів даних у корпоративних системах»

Рівень освіти	**Магістр**	**Тип дисципліни**	**Вибіркова. Професійна**
Шифр та назва спеціальності	**122 – Комп'ютерні науки**	**Інститут**	**ННІ КНІТ** **Навчально науковий інститут комп'ютерних наук та інформаційних технологій**
Назва освітньо-професійної програми	**Комп'ютерні науки**	**Кафедра**	**Системного аналізу та інформаційно-аналітичних технологій**

ВИКЛАДАЧ

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.

Персональна сторінка - *https://web.kpi.kharkov.ua/say/uk/uaabout/uaprofs/kolbasinvo/*

ЗАГАЛЬНА ІНФОРМАЦІЯ ПРО ДИСЦИПЛІНУ

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 курсової роботи студентами виконується аналіз великих обсягів даних польотів цивільної авіації.
Мета та цілі	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	Лекції, лабораторні роботи, консультації, курсова робота, самостійна робота. Підсумковий контроль – екзамен.

Результати навчання	Студент повинен: Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп'ютерних наук і є основою для оригінального мислення та проведення досліджень, критичне осмислення проблем у сфері комп'ютерних наук та на межі галузей знань. Мати спеціалізовані уміння/навички розв'язання проблем комп'ютерних наук, необхідні для проведення досліджень та/або провадження інноваційної діяльності з метою розвитку нових знань та процедур. Зрозуміло і недвозначно доносити власні знання, висновки та аргументацію у сфері комп'ютерних наук до фахівців і нефахівців, зокрема до осіб, які навчаються. Розробляти концептуальну модель інформаційної або комп'ютерної системи. Розробляти та застосовувати математичні методи для аналізу інформаційних моделей. Розробляти математичні моделі та методи аналізу даних (включно з великим). Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими). Проектувати архітектурні рішення інформаційних та комп'ютерних систем різного призначення Створювати нові алгоритми розв'язування задач у сфері комп'ютерних наук, оцінювати їх ефективність та обмеження на їх застосування
Обсяг	Загальний обсяг дисципліни 150 год.: лекції – 32 год., лабораторні роботи – 32 год., самостійна робота – 86 год.
Пререквізити	Інтелектуальний аналіз даних. Експертні системи та бази знань.
Вимоги викладача	Студент зобов'язаний відвідувати всі заняття згідно розкладу, не спізнюватися. Дотримуватися етики поведінки. Працювати з навчальною та додатковою літературою. Пропущені лабораторні та практичні заняття відпрацьовуються самостійно. Без особистої присутності студента підсумковий контроль не проводиться.

СТРУКТУРА ДИСЦИПЛІНИ

Лекція 1	Задача обробки великих обсягів даних. Технології обробки великих обсягів даних. Горизонтальне та вертикальне масштабування.	**Лабораторна робота 1**	Встановлення та ознайомлення з віртуальною машиною для роботи з BigData технологіями.	Самостійна робота	Інші віртуальні машини для роботи з BigData стеком та відповідні засоби хмарних провайдерів.
Лекція 2	Розподілене зберігання та обробка великих обсягів даних. Файлова система HDFS. Консольні операції та API для роботи з файлами в HDFS.	**Лабораторна робота 2**	Програмний доступ до файлів в HDFS.		Ознайомлення з документацією по консольних операціях та Java API для роботи з HDFS.
Лекція 3	Алгоритми паралельної обробки даних. Структура фреймворку Map-Reduce. Реалізація простих задач обробки даних.	**Лабораторна робота 3**	Пошук слів найбільшої довжини за допомогою Map-Reduce.		Класи та інтерфейси для реалізації Map-Reduce задач.
Лекція 4	Оптимізація Map-Reduce задач. Рядкові та колонкові формати файлів. Управління сортуванням та передачею даних між вузлами.	**Лабораторна робота 4**	Дослідження впливу формату файлу на швидкість обробки даних та обсяг зайнятої дискової пам'яті		Структура та особливості форматів файлів ORC, Parquet, Avro.

Лекція 5	Система управління кластером YARN. Запуск задач управління їх виконанням за допомогою YARN. Інші фреймворки управління кластером.	**Лабораторна робота 5**	Робота з журналами задач в YARN та управління виконанням задач на кластері.		Засоби моніторингу Yarn кластеру.
Лекція 6	Принципи побудови та архітектура Apache Hive. Виконання SQL запитів за допомогою Hive. Обмеження та особливості SQL. Фреймворк Tez.	**Лабораторна робота 6**	Обчислення статистичних показників по польотах цивільної авіації в США		Виконання частини курсової роботи за допомогою Hive.
Лекція 7	Принципи побудови та архітектура Apache Spark. Зберігання даних у RDD. Конвеєр обробки даних у Spark. Створення простих задач у Spark.	**Лабораторна робота 7**	Пошук найдовшого слова за допомогою Spark.		Методи обробки Spark RDD для мов Java, Scala, Python.
Лекція 8	Робота за даними у Spark: RDD, DataFrame, DataSet. Модель пам'яті PySpark застосування. Оптимізація Spark застосувань.	**Лабораторна робота 8**	Написання застосування для пошуку даних по авіаперельотам.		Засоби Python та Java для роботи з Apache Spark.
Лекція 9	Spark SQL. Виконання запитів та особливості використання Spark SQL. Засоби оптимізації Spark SQL запитів.	**Лабораторна робота 9**	Обчислення статистичних показників по польотах цивільної авіації в США та порівняння з використанням Spark SQL запитів.		Виконання частини курсової роботи за допомогою Spark SQL.
Лекція 10	Архітектура та засоби обробки потокових даних. Система обміну повідомленнями Apache Kafka. Робота з потоковими даними в Spark Streaming.	**Лабораторна робота 10**	Обчислення статистики по продажах мережі роздрібної торгівлі у режимі реального часу.		Можливості Kafka Connect та Kafka Streams.
Лекція 11	Бази даних для роботи з великими обсягами даних. Концепція та різновиди NoSQL баз даних. БД Apache Cassandra. БД MongoDB. New SQL бази даних.	**Лабораторна робота 11**	Збереження результатів потокової обробки даних у NoSQL базу даних.		Підходи до гарантування цілісності даних у NoSQL системах.
Лекція 12	Моделювання даних. Багатовимірні моделі (multi dimensional data models). Обробка даних, що змінюються (SCD).	**Лабораторна робота 12**	Побудова моделі даних для обліку польотів цивільної авіації та обліку вильотів пілотів.		Підготовка звіту з виконання курсової роботи.
Лекція 13	Архітектура систем обробки даних. Типовий процес обробки. Data warehouse та datamart. Архітектура Data Lake. Лямбда-архітектура.	**Лабораторна робота 13**	Ознайомлення з прикладом застосування, що реалізує архітектуру Data Lake.		Різновидності архітектур обробки великих обсягів даних.

Лекція 14	Інтелектуальна обробка даних. Засоби інтелектуальної обробки даних та технології машинного навчання в Apache Spark.	**Лабораторна робота 14**	Побудова регресійної моделі ціни товару.		Методи та класи машинного навчання в Apache Spark.
Лекція 15	Обробка текстових даних. Моделі текстових даних – модель мішка слів та векторні моделі. Аналіз текстів за допомогою тематичних моделей.	**Лабораторна робота 15**	Побудова тематичних моделей відгуків на товар та аналіз відгуків на основі цих моделей		Засоби тематичного моделювання текстів в Apache Spark.
Лекція 16	Рекомендаційні системи. Колаборативна фільтрація. Засоби Sprak для реалізації колаборативної фільтрації.	**Лабораторна робота 16**	Побудова простої рекомендаційної системи		Завершення звіту з виконання курсової роботи та її захист.

ЛІТЕРАТУРА ТА НАВЧАЛЬНІ МАТЕРІАЛИ

Основна

1. Олещенко Л.М. Технології оброблення великих даних. Конспект лекцій. Київ: КПІ ім. Ігоря Сікорського, 2021. 227 с.
2. Warren J., Marz N. Big Data. Manning, 2015. 328 p.
3. White T. Hadoop: The Definitive guide. O'Reilly Media, 2015. 756 p.
4. Du D. Apache Hive Essentials, 2nd edition. Packt publishing, 2018. 210 p.
5. Frampton M. Mastering Apache Spark. Packt publishing, 2015. 318 p.
6. Damji J., Wenig B., Das T. Learning Spark, 2nd edition. O`Reilly Media, 2020. 397 p.
7. Lui A. Apache Spark Machine Learning Blueprints. Packt publishing, 2016. 252 p.
8. Estrada R. Apache Kafka Quick Start Guide. Packt publishing, 2018. 186 p.
9. Ploetz A., Kandhare D., Kadambi S. Seven NoSQL Databases in a Week. Packt publishing, 2018. 308 p.
10. Pasupuleti P., Purra B. Data Lake Development with Big Data. Packt publishing, 2015. 252 p.
11. Kasliwal N. Natural Language Processing with Python Quick Start Guide. Packt publishing, 2018. 182 p.
12. Banik R. Hands-On Recommendation Systems with Python. Packt publishing, 2018. 146 p.

Додаткова

13. Bansal H., Chauhan S., Mehrotra S. Apache Hive Cookbook. Packt publishing, 2016. 268 p.
14. Luu H. Beginning Apache Spark 3. APress, 2021. 445 p.
15. Chitturi P. Apache Spark for Data Science Cookbook. Packt publishing, 2016. 392 p.
16. Shapira G., Palino T., Sivaram R. Kafka: The Definitive Guide, 2nd edition. O`Reilly Media, 2021. 485p.
17. Raj P., Deka G. A Deep Dive into NoSQL Databases: The Use Cases and Applications. Academic Press, 2018. 400 p.
18. Falk K. Practical Recommender Systems. Manning, 2019. 432 p.
19. Apache Hadoop documentation. Режим доступу: http://hadoop.apache.org/docs/current/.
20. Apache Spark documentation. Режим доступу: https://spark.apache.org/docs/latest/.

ПЕРЕЛІК ЗАПИТАНЬ ДЛЯ ПІДГОТОВКИ ДО ЕКЗАМЕНУ

Принципи обробки великих обсягів даних. Розподілена файлова система HDFS. Технологія Hadoop. Базові можливості. Підходи до оптимізація швидкодії. Формати зберігання даних ORC, Parquet, Avro. Технологія Hive - головні принципи побудови. Особливості трансляції Hive запитів до Tez робіт. Оптимізація Join запитів у Hive. Головні принципи побудови технології Apache Spark. Поняття RDD та основні типи операцій над ними. Широкі та вузькі операції в Spark. Процес виконання Spark коду. Розподіл коду за вузлами кластера. Технологія Spark SQL. Головні принципи побудови та оптимізація запитів. Обробка потокових даних за допомогою Spark Streaming. NoSql бази даних - головні принципи побудови та класифікація. Архітектура Data Lake. Підходи та засоби для інтелектуальної обробки текстових даних. Машинне навчання з використанням Spark ML. Рекомендаційні системи.

ПЕРЕЛІК ОБЛАДНАННЯ

Мультимедійний комп'ютерний клас; Windows 10 Education (Academic Open License); локально встановлена безкоштовна версія Java SDK, VirtualBox, середовище розробки IntelliJ IDEA або Eclipse, веб-браузер.

СИСТЕМА ОЦІНЮВАННЯ

Розподіл балів для оцінювання успішності студента

Сума балів за всі види навчальної діяльності	Оцінка ECTS	Оцінка за національною шкалою
90-100	A	Відмінно
82-89	B	Добре
74-81	C	
64-73	D	Задовільно
60-63	E	
35-59	FX	незадовільно з можливістю повторного складання
0-34	F	незадовільно з обов'язковим повторним вивченням дисципліни

Нарахування 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).

НОРМИ АКАДЕМІЧНОЇ ЕТИКИ

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 нерозв'язності конфлікту доводитися до співробітників деканату.

Силабус за змістом повністю відповідає робочій програмі навчальної дисципліни