



Побудова масштабованих систем обробки даних у реальному часі

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	<i>Перший (бакалаврський)</i>
Галузь знань	<i>12 Інформаційні технології</i>
Спеціальність	<i>121 Інженерія програмного забезпечення</i>
Освітня програма	<i>Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці</i>
Статус дисципліни	<i>Нормативна</i>
Форма навчання	<i>Заочна</i>
Рік підготовки, семестр	<i>4 курс, весняний семестр</i>
Обсяг дисципліни	<i>4,5 кредитів ЄКТС/135 год (10 год. лекції, 8 год. практичні заняття, 117 год. самостійна робота)</i>
Семестровий контроль/ контрольні заходи	<i>Екзамен, модульна контрольна робота, календарний контроль</i>
Розклад занять	<i>Згідно розкладу на весняний семестр поточного навчального року (rozklad.kpi.ua)</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	<i>Лектор: старший викладач, Колумбет В.П., kvplinux@gmail.com, тел. 093-405-35-39 Практичні: старший викладач, Колумбет В.П., kvplinux@gmail.com, тел. 093-405-35-39</i>
Розміщення курсу	<i>Кампус</i>

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Вивчення дисципліни «Побудова масштабованих систем обробки даних у реальному часі» дозволяє сформуванню у здобувачів освіти компетенції, необхідні для розв'язання практичних задач професійної та наукової діяльності, пов'язаної з підготовкою, аналізом великих даних у масштабованих системах обробки даних.

***Метою** вивчення дисципліни «Побудова масштабованих систем обробки даних у реальному часі» є формування у студентів навичку застосовувати програмні методи та засоби для побудови масштабованих систем, обробки, аналізу великих даних та прийняття управлінських рішень в різних галузях науки та бізнесу.*

***Предметом** дисципліни «Побудова масштабованих систем обробки даних у реальному часі» є загальні принципи та підходи до побудови масштабованих систем, оброблення великих даних, можливості мов програмування Python та R для побудови систем, оброблення, аналізу та візуалізації великих даних.*

Вивчення дисципліни «Побудова масштабованих систем обробки даних у реальному часі» сприяє формуванню у студентів фахової компетентності (ФК) за освітньою програмою:

ФК08 - Здатність застосовувати фундаментальні і міждисциплінарні знання для успішного розв'язання завдань інженерії програмного забезпечення.

Вивчення дисципліни «Побудова масштабованих систем обробки даних у реальному часі» сприяє формуванню у студентів наступних програмних результатів навчання (ПРН) за освітньою програмою:

ПРН01 - Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідникові ресурси і знання з урахуванням сучасних досягнень науки і техніки.

ПРН07 - Знати і застосовувати на практиці фундаментальні концепції, парадигми і основні принципи функціонування мовних, інструментальних і обчислювальних засобів інженерії програмного забезпечення.

ПРН18 - Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Успішному вивченню дисципліни «Побудова масштабованих систем обробки даних у реальному часі» знання отримані при вивченні дисциплін «Програмування» та «Бази даних» навчального плану підготовки бакалаврів за спеціальністю 121 Інженерія програмного забезпечення.

Отримані при засвоєнні дисципліни «Побудова масштабованих систем обробки даних у реальному часі» теоретичні знання та практичні уміння сприяють успішному виконанню курсових проєктів, стартапів та дипломної роботи бакалаврів.

3. Зміст навчальної дисципліни

Дисципліна «Побудова масштабованих систем обробки даних у реальному часі» передбачає вивчення таких тем:

Тема 1. Джерела та типи великих даних. Підготовка та аналіз даних з використанням мови Python.

Тема 2. Архітектурні моделі BigData. Можливості мови R.

Модульна контрольна робота.

Екзамен.

4. Навчальні матеріали та ресурси

Основна література:

1. BIG DATA: Інноваційні можливості підвищення прибутковості агробізнесу / [Електронний ресурс] – Режим доступу: <http://www.agrobusiness.com.ua/ideii-i-trendy/8383-bigdata-innovatsiini-mozhlyvistipidvyschennia-prybutkovostiagrobiznesu.html>
2. IoT Analytics Platform / Електронний ресурс. Режим доступу: <https://blog.codecentric.de/en/2016/07/iot-analytics-platform/>
3. IoT Fundamentals: Big Data & Analytics / Електронний ресурс. Режим доступу: <https://www.netacad.com/courses/iot/big-data-analytics>
4. Akka // Електронний ресурс. Режим доступу: <https://akka.io/>

5. *A brief introduction to two data processing architectures — Lambda and Kappa for Big Data* / Електронний ресурс. Режим доступу: <https://towardsdatascience.com/a-briefintroduction-to-two-data-processing-architectures-lambda-and-kappa-for-big-data4f35c28005bb>
6. *What Is Apache Hadoop?* / [Електронний ресурс] – Режим доступу: <http://hadoop.apache.org/>
7. *Apache Kafka. A distributed streaming platform* / [Електронний ресурс] – Режим доступу: <https://kafka.apache.org/>
8. *Apache Cassandra* / [Електронний ресурс] – Режим доступу: <http://cassandra.apache.org/>
9. *Apache Spark. A fast and general engine for large-scale data processing* / [Електронний ресурс] – Режим доступу: <https://spark.apache.org>
10. *IoT-Analyse-Plattform: Floating Bus Data* / Електронний ресурс. Режим доступу: https://www.youtube.com/watch?v=VYxc-3ZRRL4&ab_channel=codecentricAG.

Додаткова література:

1. *MapReduce. Tutorial.* Електронний ресурс. Режим доступу: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
2. *Pandas.* Електронний ресурс. Режим доступу: https://www.w3schools.com/python/pandas/pandas_intro.asp
3. *Pandas documentation: веб-сайт.* URL: <https://pandas.pydata.org/docs/>
4. *Parcel documentation: веб-сайт.* URL: <https://parceljs.org/docs/>
5. *Python documentation: веб-сайт.* URL: <https://docs.python.org/3/>
6. Slatkin B. *Effective Python: 90 Specific Ways to Write Better Python.* Addison-Wesley Professional, 2019. 480 p.
7. *Programming with Databases – Python.* Електронний ресурс. Режим доступу: <https://swcarpentry.github.io/sql-novice-survey/10-prog/index.html>
8. *Мова програмування R.* Електронний ресурс. Режим доступу: <https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r-programmirovanie>

Матеріали знаходяться у вільному доступі в Інтернеті.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

№ з/п	Тип навчального заняття	Опис навчального заняття
<i>Тема 1. Джерела та типи великих даних. Підготовка та аналіз даних з використанням мови Python.</i>		
1	<i>Лекція 1. Джерела великих даних. Інтернет Речей. Визначення BigData. Python Pandas. Відкриті дані, їх формати та засоби обробки.</i>	<i>Розвиток Інтернету Речей та збільшення обсягу інформації. Використання платформи Kaggle та DrivenData для роботи з великими обсягами даних. Аналіз визначення великих даних. Наведення прикладів використання великих обсягів даних у реальних сценаріях. Розгляд відкритих та приватних даних, структурованих та неструктурованих. Вивчення хмарних та туманних обчислень. Розгляд рухомих та нерухомих даних. Огляд інфраструктури для обробки великих обсягів даних. Аналіз розподілених даних та їх обробка, відкритих даних, їх різноманітних форматів та інструментів для їх обробки. Вивчення технік web-скрапінгу та процесів вилучення, трансформації і завантаження даних.</i>

2	Комп'ютерний практикум 1. Аналіз та візуалізація даних у Python.	Мета: продемонструвати володіння життєвим циклом аналізу даних через використання наданого набору інформації та відзначених інструментів. Здійснити импорт пакетів Python, необхідних для обробки даного набору, та використати інструменти Python та Jupyter для підготовки даних до аналізу. Провести аналіз та побудувати графіки на основі отриманих результатів.
3	Лекція 2. Програмування Python та SQLite. Процедура імпорту даних із файлів у Pandas та засоби для кореляційного аналізу в Pandas. Оброблення даних, регресійний аналіз даних в Python.	Робота Python з SQLite. Основні операції SQL. Призначення утиліти csvsql. Метод "execute". Здійснення імпорту даних з різних джерел, таких як файли та Інтернет, у середовище Pandas. Використання Pandas для проведення описової статистики та інструментів для кореляційного аналізу даних. Оброблення відсутніх даних. Маніпулювання дата фреймами у Python. Дослідження регресійного аналізу та його різновидів. Використання регресійного аналізу у здійсненні аналізу інформації.
4	Комп'ютерний практикум 2. Побудова лінійної регресії в Python.	Мета: освоїти концепції лінійної регресії та роботи з даними для прогнозування в середовищі Python. Аналіз запропонованих даних щодо обсягів продажів, та побудова лінійної регресії для прогнозування річного чистого обсягу продажів на основі кількості магазинів у даному районі.
Тема 2. Архітектурні моделі BigData. Можливості мови R.		
5	Лекція 3. Аналіз даних в R. Фактори, списки, фрейми та дії над ними. Експорт, импорт та оброблення даних в R. Основні інструменти аналізу та візуалізації даних в R.	Еволюція мови програмування R. Функціональні можливості R. Робота з об'єктами, пакетами та функціями. Операції з векторами, матрицями у мові R. Робота із факторами, списками, фреймами та їх обробка. Використання R для аналізу часових рядів. Експорт та импорт даних в R. Оброблення даних в R. Функція "plot" і її параметри. Типи графіків в R. Управління загальними параметрами - аргументами графічних функцій.
6	Комп'ютерний практикум 3. Аналіз та візуалізація даних в R.	Мета: вивчити функціональність мови програмування R для проведення аналізу та візуалізації даних. Використати бібліотеку dplyr у R для оптимізації та трансформації даних, а також використати бібліотеку ggplot2 для візуалізації отриманих результатів.
7	Лекція 4. Архітектурні моделі Big Data. Технології віртуалізації. Гіпервізори. Контейнерні технології виконання. SaaS, PaaS і IaaS. Технології Hadoop Big Data. Розподілена обробка MapReduce. HDFS.	Моделі архітектури в області Big Data Engineering. Віртуалізаційні технології та рівні абстракції. Гіпервізори. Використання контейнерної технології для виконання програмного коду на сервері. Процес інжинірингу даних. Масштабування за рахунок використання великих обсягів даних. Методи зберігання та оброблення інформації у розподілених файлових системах. Використання

		<i>розподілених баз даних. Використання розподіленої файлової системи Hadoop (HDFS).</i>
8	<i>Лекція 5. Платформа Apache Spark. Lambda та Карра архітектури оброблення великих даних.</i>	<i>Проблема обчислювальної функції та її вирішення за допомогою технології Spark. Порівняння Spark та MapReduce. Використання Spark та sparklyr для обробки великих обсягів даних в мові програмування R. Архітектура Lambda та її переваги та недоліки. Також, архітектура Карра та аналіз її переваг і недоліків.</i>
9	<i>Комп'ютерний практикум 4. Розподілені обчислення даних з використанням Spark кластера та мови R.</i>	<i>Мета: встановити Spark на власному комп'ютері та провести розподілені обчислення для конкретного набору даних, використовуючи Spark-кластер та мову програмування R.</i>
10	<i>Підсумкове заняття. Екзамен.</i>	<i>Екзаменаційна контрольна робота.</i>

6. Самостійна робота студента

Дисципліна «Побудова масштабованих систем обробки даних у реальному часі» ґрунтується на самостійній підготовці до аудиторних занять на теоретичні та практичні теми.

<i>№ з/п</i>	<i>Назва теми, що виноситься на самостійне опрацювання</i>	<i>Кількість годин</i>
1	<i>Підготовка до лекції 1</i>	8
2	<i>Підготовка до комп'ютерного практикуму 1</i>	8
3	<i>Підготовка до лекції 2</i>	8
4	<i>Підготовка до комп'ютерного практикуму 2</i>	8
5	<i>Підготовка до лекції 3</i>	8
6	<i>Підготовка до комп'ютерного практикуму 3</i>	8
7	<i>Підготовка до лекції 4</i>	8
8	<i>Підготовка до лекції 5</i>	8
9	<i>Підготовка до комп'ютерного практикуму 4</i>	8
10	<i>Підготовка до модульної контрольної роботи</i>	45

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Відвідування лекційних занять є обов'язковим.

- Відвідування занять комп'ютерного практикуму може бути епізодичним та за потреби захисту робіт комп'ютерного практикуму.
- Правила поведінки на заняттях: активність, повага до присутніх, відключення телефонів.
- Дотримання політики академічної доброчесності.

8. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

Протягом семестру студенти виконують 4 комп'ютерних практикумів. Максимальна кількість балів за кожний комп'ютерний практикум: 15 балів.

Бали нараховуються за:

- якість виконання комп'ютерного практикуму: 0-13 балів;
- відповідь під час захисту комп'ютерного практикуму: 0-4 балів; - своєчасне представлення роботи до захисту: 0-2 балів.

Критерії оцінювання якості виконання:

- 10-11 балів – робота виконана якісно, в повному обсязі;
- 1-9 балів – робота виконана якісно, в повному обсязі, але має недоліки; 0 балів – робота виконана не в повному обсязі, або містить суттєві помилки.

Критерії оцінювання відповіді:

- 3-4 балів – відповідь повна, добре аргументована;
- 1-2 бал – у відповіді є суттєві помилки;
- 0 балів – немає відповіді або відповідь невірна.

Критерії оцінювання своєчасності представлення роботи до захисту:

- 1-2 балів – робота представлена до захисту не пізніше вказаного терміну; 0 балів – робота представлена до захисту пізніше вказаного терміну.

Максимальна кількість балів за виконання та захист комп'ютерних практикумів: 15 балів × 4 комп. практ. = 60 балів.

Завдання на **модульну контрольну роботу** складається з 8 питань – 5 теоретичних та 3 практичного. Відповідь на кожне теоретичну та практичну запитання оцінюється 5 балами. Критерії оцінювання кожного теоретичного/практичного запитання модульної контрольної роботи:

- 5 балів – відповідь вірна, повна, добре аргументована;
- 3-4 балів – відповідь вірна, але неповна або погано аргументована;
- 2 балів – у відповіді є незначні помилки;
- 1 бал – у відповіді є суттєві помилки;
- 0 балів – немає відповіді або відповідь невірна.

Максимальна кількість балів за модульну контрольну роботу:

5 балів × 5 теоретичні запитання + 5 балів × 3 практичні запитання = 40 балів.

Рейтингова шкала з дисципліни дорівнює:

$R = R_c = 60 \text{ балів} + 40 \text{ балів} = 100 \text{ балів}$.

За описом: $R = R_{\text{комп.практ}} + R_{\text{МКР}} = 60 + 40 \text{ балів} = 100 \text{ балів}$

Календарний контроль: провадиться двічі на семестр як моніторинг поточного стану виконання вимог силабусу.

На першій атестації (8-й тиждень) студент отримує «зараховано», якщо його поточний рейтинг не менше 50 % від максимальної кількості балів, яку може отримати студент до першої атестації (20 балів).

На другій атестації (14-й тиждень) студент отримує «зараховано», якщо його поточний рейтинг не менше 50 % від максимальної кількості балів, яку може отримати студент до другої атестації (30 балів).

Семестровий контроль: **екзамен.**

Умови допуску до семестрового контролю:

При семестровому рейтингу (r_c) не менше 60 % (60 балів) та зарахуванні усіх робіт комп'ютерного практикуму.

Необхідною умовою допуску до екзамену є виконання і захист комп'ютерного практикуму.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	Оцінка
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

1. Додаткова інформація з дисципліни (освітнього компонента)

Сертифікати проходження онлайн курсів «IoT Fundamentals: Big Data & Analytics» та «Programming Essentials in Python» дозволяють зарахувати модульну контрольну роботу згідно отриманої загальної оцінки в Мережевій академії (Cisco Networking Academy), написання статей або участь у конференціях та проєктах за відповідною тематикою також оцінюється як додаткові 5 балів.

Складено старшим викладачем ІПЗЕ, Колумбетом Вадимом Петровичем

Ухвалено кафедрою ІПЗЕ (протокол № 28 від 15.05.2023 р.)

Погоджено Методичною комісією факультету (протокол № 9 від 26.05.2023 р.)