Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Кафедра інженерії програмного забезпечення в енергетиці (ІПЗЕ)

Побудова масштабованих систем обробки даних у реальному часі

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни	
Рівень вищої освіти	***Перший (бакалаврський)***
Галузь знань	*12 Інформаційні технології*
Спеціальність	*121 Інженерія програмного забезпечення*
Освітня програма	*Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці*
Статус дисципліни	*Нормативна*
Форма навчання	*Очна (денна)*
Рік підготовки, семестр	*4 курс, весняний семестр*
Обсяг дисципліни	*4,5 кредитів ЄКТС/135 год (36 год. лекції, 18 год. практичні заняття, 81 год. самостійна робота)*
Семестровий контроль/ контрольні заходи	*Екзамен, модульна контрольна робота, календарний контроль*
Розклад занять	*Згідно розкладу на весняний семестр поточного навчального року (rozklad.kpi.ua)*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	*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*
Розміщення курсу	*Кампус*

Програма навчальної дисципліни

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

Вивчення дисципліни «Побудова масштабованих систем обробки даних у реальному часі» дозволяє сформувати у здобувачів освіти компетенції, необхідні для розв'язання практичних задач професійної та наукової діяльності, пов'язаної з підготовкою, аналізом великих даних у масштабованих систем 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 вивчення 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-mozhlyvostipidvyschennia-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*
Тема 1. Джерела та типи великих даних. Підготовка та аналіз даних з використанням мови Python.		
1	*Лекція 1. Джерела великих даних. Інтернет Речей. Визначення BigData.*	*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.*

2	*Комп'ютерний практикум 1. Аналіз джерел відкритих даних. Завантаження та даних збереження в форматі csv.*	*Мета: аналіз джерел відкритих даних через Open Government Partnership та платформи, що надають відкриті дані, для можливостей збереження та візуалізації інформації. Використання web-сайтів www.knoema.com та www.gapminder.org для вивчення власності персональних даних у випадках їх відсутності на локальному рівні та обмежень електронних таблиць під час завантаження інформації.*
3	*Лекція 2. Розроблення програмного забезпечення для аналізу web-сайтів, які надають відкриті дані за допомогою Python Pandas. Відкриті дані, їх формати та засоби обробки.*	*Можливості інструментів для обробки даних. Роль мови програмування Python у проведенні аналізу інформації. Порівняння традиційного підходу до обробки великих обсягів даних і нового покоління аналітики. Загальний цикл життя процесу аналізу даних. Огляд відкритих даних, їх різноманітних форматів та інструментів для їх обробки. Вивчення технік web-скрапінгу та процесів вилучення, трансформації і завантаження даних.*
4	*Комп'ютерний практикум 2. Аналіз та візуалізація даних у Python.*	*Мета: продемонструвати володіння життєвим циклом аналізу даних через використання наданого набору інформації та відзначених інструментів. Здійснити імпорт 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.*
5	*Лекція 3. Форматування даних про час та дату, читання та запис файлів в Python.*	*Опрацювання даних часу та дати у мові програмування Python. Робота з читанням та записом файлів у середовищі Python. Взаємодія з зовнішніми програмами в рамках Python.*
6	*Лекція 4. Програмування Python та SQLite. Призначення утиліти csvsql.*	*Робота Python з SQLite. Основні операції SQL. Призначення утиліти csvsql. Метод execute().*
7	*Лекція 5. Процедура імпорту даних із файлів у Pandas. Імпорт даних з Інтернету за допомогою Pandas. Засоби для кореляційного аналізу в Pandas.*	*Використання статистичних методів у великоаналітичних завданнях та їх реалізація за допомогою бібліотеки Pandas у Python. 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.*
8	*Комп'ютерний практикум 3. Кореляційний аналіз у Python.*	*Мета: демонстрація вмінь проведення кореляційного аналізу даних з використанням вказаного набору та інструментів Python для розрахунку ступеня кореляції. Необхідно налаштувати набір даних, визначити, чи існує кореляція між змінними в даному наборі, використовуючи Python для обчислення кореляції між двома наборами змінних, та створити візуальне представлення результатів аналізу.*

9	*Лекція 6. Оброблення відсутніх даних. Перетворення типів даних та маніпулювання дата фреймами.*	*Перетворення типів даних. Оброблення відсутніх даних. Маніпулювання дата фреймами у Python.*
10	*Лекція 7. Регресійний аналіз даних в Python.*	*Дослідження регресійного аналізу та його різновидів. Використання регресійного аналізу у здійсненні аналізу інформації.*
11	*Комп'ютерний практикум 4. 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.*
12	*Лекція 8. Помилки в аналізі даних та прогнозній аналітиці. Оцінка помилок регресії засобами Python.*	*Недоліки у виконанні аналізу даних та прогнозній аналітиці. Визначення та оцінка помилок у регресійному аналізі за допомогою інструментів Python. Роль бібліотеки scikit-learn у цьому контексті.*
13	*Лекція 9. Алгоритми класифікації даних. Застосування класифікацій.*	*Труднощі у сфері класифікації. Методи класифікації. Графічне зображення результатів класифікації. Використання та перевірка правильності класифікацій. Огляд моделі класифікатора на основі дерева рішень.*
14	*Лекція 10. Бібліотека Pyplot. Засіб Plotly. Типи візуалізації даних. Використання бібліотек Folium та Leaflet.js для побудови карт. Візуалізація аномалій.*	*Бібліотека Pyplot. Засіб Plotly. Різноманітність методів візуалізації даних. Графічне відображення аномалій. Використання Folium та Leaflet.js для створення картографічних візуалізацій.*
15	*Комп'ютерний практикум 5.*	*Мета: освоїти бібліотеку Pyplot та Засіб Plotly, моделей класифікатора на основі дерева рішень.*
Тема 2. Архітектурні моделі BigData. Можливості мови R.		
16	*Лекція 11. Аналіз даних в R. Фактори, списки, фрейми та дії над ними.*	*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.*
17	*Комп'ютерний практикум 6.*	*Мета: освоїти функціональні можливості R, роботу з об'єктами, пакетами та функціями.*
18	*Лекція 12. Експорт, імпорт та оброблення даних в R.*	*Використання R для аналізу часових рядів. Експорт та імпорт даних в R. Оброблення даних в R.*
19	*Лекція 13. Основні інструменти аналізу та візуалізації даних в R.*	*Функція "plot" і її параметри. Типи графіків в R. Управління загальними параметрами - аргументами графічних функцій.*

20	*Комп'ютерний практикум 7. 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.*
21	*Лекція 14. Архітектурні моделі Big Data. Технології віртуалізації. Гіпервізори. Контейнерна технологія виконання програмного коду на сервері. SaaS, PaaS і IaaS.*	*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.*
22	*Лекція 15. Технології Hadoop Big Data. Розподілена обробка MapReduce. HDFS.*	*Масштабування за рахунок використання великих обсягів даних. Методи зберігання та оброблення інформації у розподілених файлових системах. Використання розподілених баз даних. Використання розподіленої файлової системи Hadoop (HDFS).*
23	*Комп'ютерний практикум 8.*	*Мета: освоїти технології Hadoop, використання контейнерних технологій.*
24	*Лекція 16. Розподілена потокова платформа Kafka. Переваги Cassandra.*	*Проблема прийому даних. Переваги Cassandra . Розподілена потокова платформа Kafka.*
25	*Лекція 17. Платформа Apache Spark. Lambda та Kappa архітектури оброблення великих даних.*	*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 Kappa та аналіз її переваг і недоліків.*
26	*Комп'ютерний практикум 9. Розподілені обчислення даних з використанням Spark кластера та мови R.*	*Мета: встановити Spark на власному комп'ютері та провести розподілені обчислення для конкретного набору даних, використовуючи Spark-кластер та мову програмування R.*
27	*Лекція 18. Підсумкове лекційне заняття. Екзамен.*	*Повторення вивченого матеріалу. Екзаменаційна контрольна робота.*

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.

№ з/п	*Назва теми, що виноситься на самостійне опрацювання*	*Кількість годин*
1	*Підготовка до лекції 1*	*2*
2	*Підготовка до комп'ютерного практикуму 1*	*2*

3	*Підготовка до лекції 2*	*2*
4	*Підготовка до комп'ютерного практикуму 2*	*2*
5	*Підготовка до лекції 3*	*2*
6	*Підготовка до лекції 4*	*2*
7	*Підготовка до лекції 5*	*2*
8	*Підготовка до комп'ютерного практикуму 3*	*2*
9	*Підготовка до лекції 6*	*2*
10	*Підготовка до лекції 7*	*2*
11	*Підготовка до комп'ютерного практикуму 4*	*2*
12	*Підготовка до лекції 8*	*2*
13	*Підготовка до лекції 9*	*2*
14	*Підготовка до лекції 10*	*2*
15	*Підготовка до комп'ютерного практикуму 5*	
16	*Підготовка до лекції 11*	*2*
17	*Підготовка до лекції 12*	*2*
18	*Підготовка до комп'ютерного практикуму 6*	
19	*Підготовка до лекції 13*	*2*
20	*Підготовка до комп'ютерного практикуму 7*	*2*
21	*Підготовка до лекції 14*	*2*
22	*Підготовка до лекції 15*	*2*
23	*Підготовка до лекції 16*	*2*
25	*Підготовка до комп'ютерного практикуму 8*	
26	*Підготовка до лекції 17*	*2*
27	*Підготовка до комп'ютерного практикуму 9*	*2*
28	*Підготовка до модульної контрольної роботи*	*27*

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 (РСО)

Протягом семестру студенти виконують 9 комп'ютерних практикумів. Максимальна кількість балів за кожний комп'ютерний практикум: 7 балів.

Бали нараховуються за:
- якість виконання комп'ютерного практикуму: 0-4 бали;
- відповідь під час захисту комп'ютерного практикуму: 0-4 бали; - своєчасне представлення роботи до захисту: 0-2 бали.

Критерії оцінювання якості виконання:
3-4 бали – робота виконана якісно, в повному обсязі;
1-2 бали – 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 бали – 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 бали – 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: 7 балів × 9 комп. практ. = 63 бали.

*Завдання на **модульну контрольну роботу** складається з 7 питань – 6 теоретичних та 1 практичного. Відповідь на кожне теоретичне запитання оцінюється 5 балами, практичне - 7. Критерії оцінювання кожного теоретичного/практичного запитання модульної контрольної роботи:*
5(7) балів – відповідь вірна, повна, добре аргументована;
3-4(5-6) балів – відповідь вірна, але неповна або погано аргументована;
2(3-4) бали – у відповіді є незначні помилки;
1(1-2) 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 + 7 балів × 1 практичні запитання = 37 балів.

Рейтингова шкала з дисципліни дорівнює:
$R = R_C$ = 63 балів + 37 балів = 100 балів.
За описом: $R = R_{комп.практ} + R_{МКР.}$ = 63+37 балів = 100 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 інженерії програмного забезпечення в енергетиці НН ІАТЕ (протокол № 34 від 10.05.2024 р.)*

Погоджено *Методичною комісією НН ІАТЕ (протокол № 9 від 31.05.2024 р.)*