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

Кафедра програмного забезпечення комп'ютерних систем

# ВЕЛИКІ ДАНІ ТА АНАЛІТИКА

## Робоча програма навчальної дисципліни (Силабус)

| Реквізити навчальної дисципліни | |
|---|---|
| **Рівень вищої освіти** | ***Другий (магістерський)*** |
| **Галузь знань** | *12 Інформаційні технології* |
| **Спеціальність** | *121 Інженерія програмного забезпечення* |
| **Освітня програма** | *Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем* |
| **Статус дисципліни** | *Вибіркова* |
| **Форма навчання** | *Очна (денна)* |
| **Рік підготовки, семестр** | *2 рік підготовки, 3 семестр* |
| **Обсяг дисципліни** | *Лекції: 36 год., комп'ютерний практикум: 18 год., самостійна робота: 66 год.* |
| **Семестровий контроль/ контрольні заходи** | *Залік, модульна контрольна робота, календарний контроль* |
| **Розклад занять** | *Згідно розкладу поточного навчального року (rozklad.kpi.ua)* |
| **Мова викладання** | *Українська* |
| **Інформація про керівника курсу / викладачів** | Лектор: *к.т.н., доцент, Олещенко Любов Михайлівна, oleshchenkoliubov@gmail.com*<br>Комп'ютерний практикум: *к.т.н., доцент, Олещенко Любов Михайлівна, oleshchenkoliubov@gmail.com* |
| **Розміщення курсу** | Google classroom: *https://classroom.google.com/u/1/c/NTQ1NjUxMjY2NzEy* |

## Програма навчальної дисципліни

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

*Вивчення дисципліни «Великі дані та аналітика» дозволяє сформувати у здобувачів освіти компетенції, необхідні для розв'язання практичних задач професійної та наукової діяльності, пов'язаної з аналізом великих даних.*

***Метою** вивчення дисципліни «Великі дані та аналітика» є формування у студентів здатностей застосовувати програмні методи аналітики великих даних для вирішення професійних та наукових задач.*

***Предметом** дисципліни «Великі дані та аналітика» є програмні методи та технології аналітики великих даних. Після засвоєння дисципліни «Великі дані та аналітика» **результатами навчання** є:*

***знання:***

- *технологій зберігання та підтримки великих даних;*
- *стеку Hadoop для великих даних та розподіленої файлової системи Hadoop (HDFS);*
- *компонентів стеку Hadoop, включаючи систему керування ресурсами та завданнями YARN, HDFS і модель програмування MapReduce;*
- *методів паралельного програмування за допомогою Spark, вбудовані бібліотеки Spark;*
- *стратегії розміщення даних, використання можливостей Zookeeper;*
- *Spark Streaming і Sliding Window Analytics, Spark GraphX & Graph Analytics.*
- *алгоритмів машинного навчання з використанням Map Reduce для Big Data Analytics.*

***уміння:***

- *використовувати технології аналітики великих даних за допомогою можливостей мови програмування Python.*

***досвід:***

- *застосовувати набуті знання для подальшої наукової та професійної діяльності.*

*Вивчення дисципліни «Великі дані та аналітика» сприяє формуванню у здобувачів вищої освіти, які навчаються за освітньою програмою «Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем» компетентностей, необхідних для розв'язання практичних задач професійної діяльності, пов'язаної з аналітикою великих даних:*

***ЗК01*** *Здатність до абстрактного мислення, аналізу та синтезу.*

***ЗК03*** *Здатність проводити дослідження на відповідному рівні.*

***ФК02*** *Здатність розробляти і реалізовувати наукові та/або прикладні проєкти у сфері інженерії програмного забезпечення.*

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

*Успішному вивченню дисципліни «Великі дані та аналітика» передує вивчення нормативних дисциплін «Програмування» та «Бази даних», вибіркової дисципліни «Технології оброблення великих даних» навчального плану підготовки бакалаврів за спеціальністю 121 «Інженерія програмного забезпечення».*

*Дисципліна «Великі дані та аналітика» забезпечує виконання курсових проєктів та магістерських дисертацій за спеціальністю 121 «Інженерія програмного забезпечення».*

## 3. Зміст навчальної дисципліни

*Дисципліна «Великі дані та аналітика» передбачає вивчення таких тем:*

*Тема 1. Архітектурні моделі великих даних.*

*Тема 2. Програмні методи аналітики великих даних.*

*Залік.*

## 4. Навчальні матеріали та ресурси

### *Базова література:*


1. *Олещенко Л. М. Технології оброблення великих даних: конспект лекцій з дисципліни «Технології оброблення великих даних»: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» / Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – 2021. – 225 с.*
2. *Олещенко Л. М. Технології оброблення великих даних: комп'ютерний практикум з дисципліни «Технології оброблення великих даних»: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» / Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – 2021. – 85 с.*


### *Додаткова література:*


1. *MapReduce. Tutorial. Електронний ресурс. Режим доступу: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html*
2. *Мова програмування R. Електронний ресурс. Режим доступу: https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r-programmirovanie*
3. *Running ZooKeeper in Production. Електронний ресурс. Режим доступу: https://docs.confluent.io/platform/current/zookeeper/deployment.html*
4. *Running ZooKeeper, A Distributed System Coordinator. Електронний ресурс. Режим доступу: https://kubernetes.io/docs/tutorials/stateful-application/zookeeper/*
5. *Hadoop YARN Architecture. Електронний ресурс. Режим доступу: https://www.geeksforgeeks.org/hadoop-yarn-architecture/*
6. *Understanding YARN architecture and features. Електронний ресурс. Режим доступу:*

https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.0.0/data-operating-system/content/apache_yarn.html
7. Machine Learning Algorithm K-means using Map Reduce for Big Data Analytics. Електронний ресурс. Режим доступу: http://www.nitttrc.edu.in/nptel/courses/video/106104189/lec25.pdf
8. Big Data Predictive Analytics. Електронний ресурс. Режим доступу: https://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics/
9. Page Rank Algorithm and Implementation. Електронний ресурс. Режим доступу: https://www.geeksforgeeks.org/page-rank-algorithm-implementation/
10. PageRank Algorithm in Big Data. Електронний ресурс. Режим доступу: http://www.nitttrc.edu.in/nptel/courses/video/106104189/lec31.pdf
11. GraphX Programming Guide. Електронний ресурс. Режим доступу: https://spark.apache.org/docs/latest/graphx-programming-guide.html

*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. Архітектурні моделі великих даних.* | |
| 1 | *Лекція 1. Великі дані та їх роль у сучасному світі. Проблеми зберігання та аналітики великих даних.* | *Зростання даних у сучасному світі. Зростання пристроїв Інтернету речей. Визначення великих даних та їх класифікація. Структуровані та неструктуровані дані. Хмарні обчислення. Туманні обчислення.* |
| 2 | *Лекція 2. Інфраструктура великих даних. Розподілені дані та їх аналітика.* | *Інфраструктура великих даних. Розподілені дані та їх аналітика.* |
| 3 | *Лекція 3. Життєвий цикл аналітики великих даних.* | *Життєвий цикл аналітики великих даних. Програмні методи витягування, перетворення та завантаження великих даних.* |
| 4 | *Комп'ютерний практикум 1. Аналітика великих даних та машинне навчання.* | *Завдання: проаналізувати набір даних про банківські реквізити клієнтів та багато іншої інформації, пов'язаної з кредитуванням, яку протягом багатьох років збирала велика фінансова компанія (файл test.csv, що містить 50000 записів). Потрібно побудувати інтелектуальну систему для розподілу людей за кредитними балами, враховуючи кредитну інформацію особи та модель машинного навчання, яка зможе класифікувати кредитну оцінку.* |
| 5 | *Лекція 4. Технології вебскрапінгу, вебкраулінгу, індексації та вебавтоматизації.* | *Технології та програмні методи вебскрапінгу, вебкраулінгу, індексації та вебавтоматизації. Машинозчитувані дані та API. HTML-аналізатори. DOM-аналіз. Розпізнавання семантичних анотацій. Аналізатори вебсторінки з використанням комп'ютерного зору.* |

| | | |
|---|---|---|
| *6* | *Комп'ютерний практикум 2. Аналіз фінансових твітів користувачів щодо акцій, якими торгують NYSE, NASDAQ і SNP.* | *Завдання: Файл stockerbot-export.csv містить 28 тисяч твітів про публічні компанії (і кілька криптовалют), які позначені компанією, про яку вони твітують, і її символом, а також іншими полями. Потрібно розробити класифікатор фінансових настроїв, який зможе відстежувати почуття у Twitter (і всієї громадськості) щодо будь-якої публічної компанії (і криптовалюти).* |
| *7* | *Лекція 5. Технології віртуалізації для аналітики великих даних.* | *Роль віртуалізації для аналітики великих даних. Технології віртуалізації. Шари абстракції. Гіпервізори.* |
| *8* | *Лекція 6. Контейнерна технологія виконання програмного коду на сервері.* | *Контейнерна технологія виконання програмного коду на сервері. Інжиніринг великих даних. SaaS, PaaS і IaaS.* |
| *9* | *Лекція 7. Технології Hadoop для великих даних.* | *Технології Hadoop для великих даних. Масштабованість за допомогою великих даних.* |
| *10* | *Лекція 8. Програмна модель та програмний каркас MapReduce. Можливості HDFS.* | *Програмна модель та програмний каркас MapReduce. Зберігання та оброблення даних в розподілених файлових системах. Розподілені бази даних. Розподілена файлова система Hadoop (HDFS).* |
| *11* | *Лекція 9. Розподілена потокова платформа Kafka. Переваги розподіленої СУБД Cassandra.* | *Проблема прийому даних. Розподілена потокова платформа Kafka. Переваги розподіленої СУБД Cassandra.* |
| *12* | *Лекція 10. Платформа Apache Spark та її можливості.* | *Проблема обчислювальної функції. Технологія Spark. Порівняння Spark та MapReduce. Можливості модулів Spark MLlib, Spark Streaming GraphX.* |
| *13* | *Лекція 11. Можливості Apache Flink для потокової та пакетної обробки даних.* | *Архітектура Apache Flink. Функції з відстеженням стану. Можливості Flink ML.* |
| *Тема 2. Програмні методи аналітики великих даних.* | | |
| *14* | *Лекція 12. Розподілена система керування конфігурацією та службами для розподілених програм ZooKeeper.* | *Запуск ZooKeeper у виробництві. Обладнання. Пам'ять. CPU та GPU. Параметри конфігурації.* |
| *15* | *Комп'ютерний практикум 3. Розподілені обчислення великих даних з використанням Spark-кластера.* | *Завдання: Встановити Spark на локальній машині, виконати розподілену аналітику для набору великих даних з використанням Spark-кластера.* |
| *16* | *Лекція 13. Використання Hadoop YARN для аналітики великих даних.* | *Архітектура Hadoop YARN. Map Reduce у Hadoop. Використання Hadoop YARN для аналітики великих даних.* |
| *17* | *Лекція 14. Програмні методи кластеризації великих даних.* | *Програмні методи кластеризації великих даних. Використання алгоритму K-means та Map Reduce.* |
| *18* | *Лекція 15. Предиктивна аналітика великих даних.* | *Предиктивна аналітика великих даних, основні методи та технології. Регресійний аналіз великих даних.* |
| *19* | *Лекція 16. Використання хмарних провайдерів Big Data.* | *Огляд хмарних провайдерів Big Data. Платформа AWS Big Data.* |

| | | |
|---|---|---|
| *20* | *Лекція 17. Використання алгоритму PageRank для аналітики великих даних.* | *Алгоритм Page Rank і його програмна реалізація для аналітики великих даних. Графік властивостей. Приклад діаграми властивостей. Оператори графів. Зведений список операторів. Оператори власності. Структурні оператори. Оператори приєднання. Агрегація сусідства. Сукупні повідомлення (aggregateMessages). Map Reduce Triplets Transition Guide (Legacy). Інформація про ступінь комп'ютера. Збирання сусідів. Кешування та декешування. Pregel API. Конструктори графів.* |
| *21* | *Лекція 18.Підсумкове заняття.* | *Модульна контрольна робота.* |

## 6. Самостійна робота студента

*Дисципліна «Великі дані та аналітика» ґрунтується на самостійній підготовці до аудиторних занять на теоретичні теми.*

| *№ з/п* | *Назва теми, що виноситься на самостійне опрацювання* | *Кількість годин* | *Література* |
|---|---|---|---|
| *1* | *Підготовка до лекції 1* | *2* | *1, стор. 8-19.* |
| *2* | *Підготовка до лекції 2* | *2* | *1, стор. 20-21, 1(дод.).* |
| *3* | *Підготовка до лекції 3* | *2* | *1, стор. 20-26.* |
| *4* | *Підготовка до комп'ютерного практикуму 1* | *4* | *2, стор.17-23.* |
| *5* | *Підготовка до лекції 4* | *2* | *1, стор.28-40.* |
| *6* | *Підготовка до комп'ютерного практикуму 2* | *4* | *2, стор.24-38.* |
| *7* | *Підготовка до лекції 5* | *2* | *1, стор.169-173.* |
| *8* | *Підготовка до лекції 6* | *2* | *1, стор.176-182.* |
| *9* | *Підготовка до лекції 7* | *2* | *1, стор.183-185.* |
| *10* | *Підготовка до лекції 8* | *2* | *1, стор. 185-189.* |
| *11* | *Підготовка до лекції 9* | *2* | *1, стор.190-201.* |
| *12* | *Підготовка до лекції 10* | *2* | *1, стор.201-206.* |
| *13* | *Підготовка до лекції 11* | *2* | *1, стор.219-226.* |
| *14* | *Підготовка до лекції 12* | *2* | *1, стор.119-168, 2 (дод.).* |
| *15* | *Підготовка до комп'ютерного практикуму 3* | *4* | *2, стор.70-85.* |
| *16* | *Підготовка до лекції 13* | *2* | *3 (дод.), 4 (дод.).* |
| *17* | *Підготовка до лекції 14* | *2* | *5 (дод.), 6 (дод.).* |
| *18* | *Підготовка до лекції 15* | *2* | *7 (дод.).* |
| *19* | *Підготовка до лекції 16* | *2* | *8 (дод.).* |
| *20* | *Підготовка до лекції 17* | *2* | *9 (дод.), 10 (дод.), 11 (дод.).* |
| *21* | *Підготовка до модульної контрольної роботи* | *20* | *1, стор.28-40, 119-226.* |

## 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 **3 комп'ютерні практикуми**. Максимальна кількість балів за кожний комп'ютерний практикум: 20 балів.*

*Бали нараховуються за:*
*- якість виконання комп'ютерного практикуму: 0-8 бали;*
*- відповідь під час захисту комп'ютерного практикуму: 0-8 бали;*
*- своєчасне представлення роботи до захисту: 0-4 бали.*
*Критерії оцінювання якості виконання:*
*7-8 балів – робота виконана якісно, в повному обсязі;*
*5-6 балів – робота виконана якісно, в повному обсязі, але має недоліки;*
*3-4 бали – робота виконана якісно, але не в повному обсязі, має недоліки;*
*1-2 бали – робота виконана не якісно, не в повному обсязі, має недоліки;*
*0 балів – робота виконана не в повному обсязі, або містить суттєві помилки.*
*Критерії оцінювання відповіді:*
*7-8 балів – відповідь повна, добре аргументована;*
*5-6 балів – відповідь неповна, проте добре аргументована;*
*3-4 бали – у відповіді є незначні помилки;*
*1-2 бали – у відповіді є суттєві помилки;*
*0 балів – немає відповіді або відповідь невірна.*
*Критерії оцінювання своєчасності представлення роботи до захисту:*
*4 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:*
*20 балів × 3 комп. практ. = 60 балів.*

*Завдання на **модульну контрольну роботу** складається з 5 питань – 3 теоретичних та 2 практичних. Відповідь на кожне теоретичне/практичне запитання оцінюється 8 балами.*
*Критерії оцінювання кожного теоретичного/практичного запитання модульної контрольної роботи:*
*7-8 балів – відповідь вірна, повна, добре аргументована;*
*5-6 балів – відповідь вірна, але неповна або погано аргументована;*
*3-4 бали – у відповіді є незначні помилки;*
*1-2 бали – у відповіді є суттєві помилки;*
*0 балів – немає відповіді або відповідь невірна.*
*Максимальна кількість балів за модульну контрольну роботу:*
*8 балів × 3 теоретичні запитання + 8 балів × 2 практичні запитання = 40 балів.*

*Рейтингова шкала з дисципліни дорівнює:*
*$R = R_C$ = 60 балів + 40 балів = 100 балів.*
*За описом: $R = R_{комп.практ} + R_{МКР.}$ = 60+40 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:* ***залік****.*

*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 заліку 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 |

## 9. Додаткова інформація з дисципліни (освітнього компонента)

*Наявність сертифікату проходження аналогічного курсу з аналітики великих даних оцінюється як 20 балів, написання статей або участь у конференціях/ проєктах за відповідною тематикою також оцінюється як додаткові 5 балів.*

**Складено** к.т.н., доц. Олещенко Л.М.

**Ухвалено** кафедрою ПЗКС (протокол №8 від 25.01.23)

**Погоджено** Методичною комісією факультету прикладної математики (протокол № 6 від 27.01.2023)