**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**
**Львівський національний університет імені Івана Франка**
**Факультет електроніки та комп'ютерних технологій**
**Кафедра системного проектування**

**Затверджено**
На засіданні
кафедри системного проектування
факультету електроніки та комп'ютерних технологій
Львівського національного університету імені Івана Франка
(протокол № 1 від 30.08 2022 р.)

Завідувач кафедри:

_____ Роман ШУВАР

# Силабус з навчальної дисципліни "Методи та технології обробки великих та надвеликих даних", що викладається в межах ОПП " Високопродуктивний комп'ютинг " першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 121 – Інженерія програмного забезпечення

**Львів 2022 р.**

| | |
|---|---|
| **Назва дисципліни** | Методи та технології обробки великих та надвеликих даних |
| **Адреса викладання дисципліни** | м. Львів, вул. Драгоманова, 50 |
| **Факультет та кафедра, за якою закріплена дисципліна** | Факультет електроніки та комп'ютерних технологій, кафедра системного проектування |
| **Галузь знань, шифр та назва спеціальності** | 12 Інформаційні технології<br>121 Інженерія програмного забезпечення (ВПК) |
| **Викладачі дисципліни** | Стахіра Роман Йосипович, доцент<br>Ляшкевич Василь Яремович, доцент |
| **Контактна інформація** | roman.stakhira@lnu.edu.ua<br>vasyl.lyashkevych@lnu.edu.ua |
| **Консультації з питань навчання по дисципліні відбуваються** | Консультації в день проведення лекційних занять (за попередньою домовленістю). Також можливі он-лайн консультації через MS Teams або систему електронного навчання Moodle. Для погодження часу онлайн консультацій слід писати на електронну пошту викладача. |
| **Сторінка дисципліни** | https://moodle.elct.lnu.edu.ua/course/view.php?id=318 |
| **Інформація про дисципліну** | Дисципліна "Методи та технології обробки великих та надвеликих даних" є дисципліною з циклу професійної та практичної підготовки спеціальності 121 Інженерія програмного забезпечення для освітньої програми «Високопродуктивний комп'ютинг», яка викладається в 6 семестрі в обсязі 4,5 кредитів (за Європейською Кредитно-Трансферною Системою ECTS). |
| **Коротка анотація дисципліни** | Навчальну дисципліну розроблено таким чином, щоб оволодіти базовими поняттями великих даних, організацією та використанням розподілених технологій для опрацювання великих даних, управлінні ресурсами віддалених розподілених систем, баз та сховищ даних, використанням технологій розподілених обчислень для побудови конвеєрів великих даних, вирішення питання аналітики на основі великих даних для вирішення задач в галузі науки про дані та систем штучного інтелекту. |
| **Мета та цілі дисципліни** | **Метою** вивчення дисципліни "Методи та технології обробки великих та надвеликих даних" є надання поглиблених знань та практичних навичок щодо роботи з великими даними, побудови й використання розподілених систем для побудови конвеєрів опрацювання великих даних, формування системи теоретичних знань і набуття практичних умінь та навичок щодо застосування, налагодження й адміністрування систем на базі технологій розподілених сховищ даних та проектування відповідних надійних та економічно |

| | |
|---|---|
| | привабливих систем для збереження великих об'ємів даних. **Цілями** дисципліни є засвоєння методів створення розподілених систем та технології їх проектування для вирішення задач наук про дані на основі великих даних, наповнення даними та підтримання в робочому стані, вивчення методів і засобів обробки великих даних. |
| **Література для вивчення дисципліни** | Основна література:<br>1. Павленко Л. А. Корпоративні інформаційні системи: Навчальний посібник./ Л. А. Павленко - Харків: ВД "ІНЖЕК", 2005. – 260<br>2. Катренко А.В., Системний аналіз об'єктів та процесів комп'ютеризації : Навчальний посібник./А.В. Катренко - Львів: "Новий світ-2000".-2003.-424с.<br>3. Michael Armbrust. Makeing Apache Spark better with Delta Lake: Databricks, 2020. - 399 p.<br>4. Gerardus Blokdyk. Databricks A complete Guide, 2021. - 205 p. - [Режим доступу]: https://www.everand.com/book/487839900/Databricks-A-Complete-Guide-2021-Edition<br>5. Tom White. Hadoop: The definitive Guide: O`Reilly, 2015. - 805 p.<br>6. Документація Apache Hadoop [Електронний ресурс] // Apache Hadoop. – 2021. – Режим доступу до ресурсу: https://hadoop.apache.org/docs/stable/.<br>7. Документація Apache Spark [Електронний ресурс] // Apache Spark. – 2019. – Режим доступу до ресурсу: https://spark.apache.org/docs/latest/.<br>8. Документація HBase [Електронний ресурс] // HBase. – 2019. – Режим доступу до ресурсу: https://hbase.apache.org/book.html.<br>9. RabbitMq [Електронний ресурс] // RabbitMq. – 2020. – Режим доступу до ресурсу: https://www.rabbitmq.com/documentation.html.<br>10. Ifeyinwa A. A. Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications / A. A. Ifeyinwa, H. N. Friday. – Nigeria: Abakaliki, 2019. – 30 c.<br>11. Donald Miner, Adam Sbook. MapReduce Design Patterns: O`Reilly, 2013. - 251 p. - Режим доступу: http://vargas-solar.com/bigdata-fest/wp-content/uploads/sites/33/2014/11/MapReduce-Design-Patterns-V413HAV.pdf<br>12. Michael Crump, Chris Pietschmann, Vahe Minasyan. The Developer's Guide to Azure. Microsoft Press, A division of Microsoft Corporation One Microsoft Way, Redmond, Washington 98052-6399.<br>13. Kai Hwang, Min Chen. Big-Data Analytics for Cloud, IoT and Cognitive Computing: Willey, 2017. - 428 p.<br>14. Designing Distributed System. - [Режим доступу]: https://azure.microsoft.com/mediahandler/files/resourcefiles/designing-distributed-systems/Designing_Distributed_Systems.pdf<br>15. Kristina Chodorow. Scaling MongoDB: O'Reilly, 2011. - 58 p.<br>16. Query-By-Example (QBE). - [Електронний ресурс]. - Режим доступу: https://link.springer.com/chapter/10.1007/978-3-642-58763-4_10<br>17. Google file system. - [Електронний ресурс]. - Режим доступу: https://docplayer.net/10419940-The-google-file-system.html<br>18. Google. Cloud Bigtable. - [Електронний ресурс]. - Режим доступу: https://cloud.google.com/bigtable<br>19. Rik Van Bruggen. Learning Neo4j: Packt Publishing, 2014. - 222 p.<br>20. MySQL Cluster Manager 8.0.31 User Manual. - [Електронний ресурс]. - Режим доступу: https://downloads.mysql.com/docs/mysql-cluster-manager-1.4-en.a4.pdf<br>21. Alex Holmes. Hadoop in Practice: Manning Publications, 2012. - 537 p. - Режим доступу: https://ia600201.us.archive.org/7/items/HadoopInPractice/Hadoop%20in%20Practice.pdf<br>22. Neha Narkhede. Kafka: The Definitive Guide: O`Reilly, 2017. - 322 p. - [Електронний ресурс]. - Режим доступу: https://book.huihoo.com/pdf/confluent-kafka-definitive-guide-complete.pdf<br>23. Bas Harenslak, Julian de Ruiter. Data Pipelines with Apache Airflow: Manning Publications, 2021. - 482 p. - [Електронний ресурс]. - Режим доступу: https://biconsult.ru/files/Data_warehouse/Bas_P_Harenslak%2C_Julian_Rutger_de_Ruiter_Data_Pipelines_with_Apache.pdf<br>24. Apache HBase Team. Apache HBase ™ Reference Guide. - [Електронний ресурс]. - Режим доступу: https://hbase.apache.org/apache_hbase_reference_guide.pdf<br>25. Informatica PowerCenter Designer Guide 10.4.0: Informatica, 2019. - 286. - [Електронний ресурс]. - Режим доступу: https://docs.informatica.com/content/dam/source/GUID-B/GUID-B54ED1F4-60B8-4F11-8E22-48C4BECE109A/27/en/PC_1040_DesignerGuide_en.pdf<br>26. Joshua N.Milligan. Learning Tableau 2019. Third Edition: Packt Publications, 2019. - 808 p. - [Електронний ресурс]. - Режим доступу: http://projanco.com/Library/Learning%20Tableau%202019%20Tools%20for%20Business%20Intelligence,%20data%20prep,%20and%20visual%20analytics.pdf<br>27. Nagios. - [Електронний ресурс]. - Режим доступу: https://www.tutorialspoint.com/nagios/nagios_tutorial.pdf<br>28. Icinga2open source monitoring. - [Електронний ресурс]. - Режим доступу: https://docplayer.net/11107242-Icinga2-open-source-monitoring.html |
| **Обсяг курсу** | Кількість кретитів ЄКТС: 4,5 (135 год), з них: 64 годин аудиторних занять (лекції: 32 год, лабораторні: 32 год.) та 71 год. самостійної роботи. |

| | |
|---|---|
| **Очікувані результати навчання** | Після вивчення даного курсу здобувачі набудуть таких Загальних(ЗК)/Фахових(ФК) компетентностей та Програмних результатів навчання (ПРН):<br><br>ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.<br>ЗК02. Здатність застосовувати знання у практичних ситуаціях.<br>ЗК04. Здатність спілкуватися іноземною мовою як усно, так і письмово.<br>ЗК05. Здатність вчитися і оволодівати сучасними знаннями.<br>ФК13. Здатність ідентифікувати, класифікувати та формулювати вимоги до програмного забезпечення.<br>ФК14. Здатність брати участь у проектуванні програмного забезпечення, включаючи проведення моделювання (формальний опис) його структури, поведінки та процесів функціонування.<br>ФК15. Здатність розробляти архітектури, модулі та компоненти програмних систем.<br>ФК22. Здатність накопичувати, обробляти та систематизувати професійні знання щодо створення і супроводження програмного забезпечення та визнання важливості навчання протягом всього життя.<br>ФК24. Здатність здійснювати процес інтеграції системи, застосовувати стандарти і процедури управління змінами для підтримки цілісності, загальної функціональності і надійності програмного забезпечення.<br>ФК25. Здатність обґрунтовано обирати та освоювати інструментарій з розробки та супроводження програмного забезпечення.<br>ФК27. Здатність розробляти високопродуктивні програмні комплекси для вирішення завдань наук про дані, систем штучного інтелекту, вбудованих та інших інноваційних систем.<br>ФК28. Володіння методами розроблення систем підвищеної продуктивності, серверними та розподіленими технологіями, інструментальними засобами проектування та розробки веб-застосувань і нових технологій.<br>ФК29. Здатність здійснювати розробку програмного забезпечення використовуючи різні методології та засоби програмування з метою забезпечення їх високої надійності та продуктивності в роботі.<br>ПРН01. Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідникові ресурси і знання з урахуванням сучасних досягнень науки і техніки.<br>ПРН09. Знати та вміти використовувати методи та засоби збору, формулювання та аналізу вимог до програмного забезпечення.<br>ПРН10. Проводити передпроектне обстеження предметної області, системний аналіз об’єкта проектування.<br>ПРН13. Знати і застосовувати методи розробки алгоритмів, конструювання програмного забезпечення та структур даних і знань.<br>ПРН14. Застосовувати на практиці інструментальні програмні |

| | |
|---|---|
| | засоби доменного аналізу, проектування, тестування, візуалізації, вимірювань та документування програмного забезпечення.<br>ПРН18. Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних.<br>ПРН21. Знати, аналізувати, вибирати, кваліфіковано застосовувати засоби забезпечення інформаційної безпеки (в тому числі кібербезпеки) і цілісності даних відповідно до розв'язуваних прикладних завдань та створюваних програмних систем.<br>ПРН26. Знати засоби інтеграції, розгортання та підтримки спеціалізованих програмних компонентів, розроблених на основі інноваційних технологій для вирішення завдань високопродуктивних обчислень.<br>ПРН27. Знати основи інженерії даних і конструювання конвеєрі даних та вміти обирати оптимальні алгоритми і технології розробки інноваційних рішень, зокрема для вирішення задач наук про дані та вбудованих систем. |
| **Ключові слова** | Розподілені системи, Великі дані, Конвеєри обробки великих даних, розподілені бази даних, кластери даних, сховища даних, озера даних, Data Warehouse, Data Mart, Data Mesh, RabbitMQ. |
| **Формат курсу** | Очний. Проведення лекцій, практичних робіт та консультації для кращого розуміння тем. |
| **Теми** | Див. СХЕМА КУРСУ |
| **Підсумковий контроль, форма** | Залік в кінці семестру |
| **Пререквізити** | Для вивчення курсу студенти потребують базових знань з дисциплін: "Методи та технології аналізу даних" та "Бази даних". |
| **Навчальні методи та техніки, які будуть використовуватися під час викладання курсу** | Презентація, лекції, лабораторні роботи, обговорення, дискусія. |
| **Необхідне обладнання** | Мультимедійне обладнання, комп'ютерний клас, програми та сервіси MS Teams, Moodle, Databricks, Hadoop, Apache Spark, PySpark, Airflow, Microsoft Azure, Azure Cosmos, Apache Kafka, Python |
| **Критерії оцінювання (окремо для кожного виду навчальної діяльності)** | Оцінювання проводиться упродовж семестру за 100-бальною шкалою, де враховано бали за два контрольні заміри по 35 балів за кожний модуль та 30 балів за складання заліку.<br>Бали нараховуються за видами робіт з співвідношенням:<br>• контрольні заміри (2 модулі): 70% семестрової оцінки; максимальна кількість балів: 70, а саме:<br>- лабораторні роботи: 60% оцінки контрольного заміру; максимальна кількість балів: 42 (12 лабораторних робіт).<br>- теоретичний матеріал: 40% оцінки контрольного заміру; максимальна кількість балів: 28 (2 тести по 14 балів кожний). |

| | |
|---|---|
| | • залік: 30% семестрової оцінки, максимально 30 балів.<br>Оцінки за лабораторні заняття розподіляються наступним чином: виконання лабораторних завдань – 60 %, відповіді на запитання викладача по темі заняття – 40 %.<br>Бали за лабораторними роботами розподіляються так:<br>- лаб. роботи 1 - 10 оцінюються в 3 бали, де від 0 до 2 балів за виконання та від 0 до 1 бали за тестування чи опитування.<br>- лаб. роботи 11-12 оцінюються в 4 бали, де від 0 до 3 балів за виконання та від 0 до 1 бали за тестування чи опитування.<br>У випадку тестування, вага питання має 0.2 бала (5 питань для однієї лабораторної роботи). Критерії виконання та оцінювання більш детально розписані у завданнях для лабораториних робіт.<br><br>**Оцінювання залікових питань:**<br>10 балів - розглянута тема відтворюється в повному обсязі, правильно, обґрунтовано, логічно, які містять аналіз і систематизацію, аргументовані висновки. Засвідчено глибоке володіння матеріалом. Наведені приклади коду повністю робочі та відповідають темі. Можуть бути присутні несуттєві помилки та невідповідності;<br>8 балів - відтворюється значна частина розглянутої теми. Виявлено знання і розуміння основних положень навчальної дисципліни, проте присутні неточності та/або невідповідності основній темі. Наведені приклади коду частково робочі, проте в загальному відповідають темі;<br>5 балів - відстежується загальне розуміння розглянутої теми. Виявлені множинні неточності та невідповідності, пояснення наведеного коду відстуні, код функціонує із значними неточностями (або відсутні приклади запуску коду на виконання взагалі);<br>3 бали – студент погано розуміє розглянуту тему. Виявлені суттєві неточності та невідповідності. Наведені приклади коду з суттєвими недоліками, або не відповідають темі;<br>Менше 3 балів – студент взагалі не розуміє розглянуту тему. Тему не розкрито, кількість викладеного матеріалу не відповідає загальним нормам обраного виду роботи. Наведений код не робочий, або відсутній як такий.<br>**Академічна доброчесність**: Очікується, що лабораторні та контрольні роботи студентів будуть їх оригінальними дослідженнями чи міркуваннями. Відсутність посилань на використані джерела, фабрикування джерел, списування, втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в роботі студента є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману.<br>**Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції і лабораторні заняття курсу. Студенти мають інформувати викладача про неможливість відвідати заняття. |

| | |
|---|---|
| | Студенти зобов'язані дотримуватися усіх термінів визначених для виконання усіх видів робіт, передбачених курсом.<br>**Література**. Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих.<br>**Політика виставлення балів**. Враховуються бали набрані на поточному тестуванні, самостійній роботі та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання. Жодні форми порушення академічної доброчесності не толеруються. |
| **Питання до контрольних робіт** | Перелік питань та завдань для проведення підсумкової оцінки знань певних тем до контрольних робіт:<br>1. Вступ до великих даних (Великі дані. Атрибути великих даних. Загальне використання. Принцип роботи. Проблеми Великих даних. Спільні характеристики. Джерела великих даних. Загальні поняття для накопичення та опрацювання великих даних.)<br>2. Парадигма MapReduce та доступ до даних (*Історія MapReduce. Розгляд системи MapReduce. Розгляд платформи Apache Pig. Apache Hive. Hive QL. Архітектра MRv1. Планувальник YARN.*)<br>3. Мова запитів GraphQL (*Що таке є мова запитів GraphQL. Переваги GraphQL. GraphQL в порівнянні REST. Схеми GraphQL. Схема визначення. Вирішувачі (Resovers). Запит даних. Повернення даних. Apollo GraphQL.*)<br>4. Шаблони у MapReduce (*Поняття шаблону. Питання проектування шаблонів MapReduce. Шаблон фільтрування. Фільтр Bloom. Шаблони об'єднання. Мета шаблони.*)<br>5. Шаблони сумаризації у MapReduce (*Шаблон сумаризації. Числова сумаризація. Сумаризація з інвертованих індексом. Сумаризація обрахунків.*)<br>6. Шаблони організації даних у MapReduce (Шаблон організації даних. Структурований шаблон. Ієрархічний шаблон. Шаблон роздріблення та зв'язування.)<br>7. Концепція розподілених систем (*Приклади типових розподілених систем. Компоненти розподілених систем. Інтернет та інтранет. Використання пристроїв у розподілених ситемах. Обмін ресурсів в WWW. Веб-сервери та веб-браузери. HTTP/HTTPS протоколи. Програмні та апаратні сервісні рівні. Клуєнт-серверні моделі. Веб-прохі сервери.*)<br>8. Веб-застосунки (*Веб-застосунки. Мобільні агенти. Комп'ютерні мережі. Синхронні та асинхронні розподілені системи. Події в системі. Порядок подій в реальному часі. Канали зв'язку. Питання захисту.*) |

| | |
|---|---|
| | 9. Поняття нереляційних баз даних (*Поняття NoSQL. Модель даних. Типи моделей даних. Логічна модель. Створення фізичної моделі даних. Фізична модель даних для реляційних баз даних. Інструменти для моделювання даних.*)<br>10. MongoDB (*Гнучке моделювання даних за допомогою MongoDB Atlas. Приклад моделювання бази даних. Запити і агрегування. Детальніше про мову запитів в MongoDB.* )<br>11. MongoDB SQL (*Розуміння зіставлення MongoDB та SQL. MongoDB SQL: термінологія. MongoDB SQL: Виконувані файли бази даних. MongoDB SQL: команди. Приклад CRUD функцій в MongoDB. Побудова запитів в MongoDB.*)<br>12. Графові бази даних (*Що таке є мова запитів GraphQL. Переваги GraphQL. GraphQL в порівнянні REST. Схеми GraphQL. Схема визначення. Вирішувачі (Resovers). Запит даних. Повернення даних. Apollo GraphQL. Приклад з NodeJS та Express. Приклад запиту для React client.*)<br>13. Резидентна система управління базами даних Redis (*Розподілене сховище пар ключ-значення. Конфігурації Redis. Типи даних Redis. Команди з використанням ключа. Redis server. Redis on python. Backup. Безпека Redis. Redis benchmark. Транзакції Redis. Redis pipelining. Redis на docker.*)<br>14. Графова база даних Neo4j (*Graph Databases. Приклади графових баз даних. Порівняння графових та реляційних баз даних. Графова база даних Neo4j. Neo4j браузер. Схема Neo4j. Структура зашифрованого запиту. Написання зашифрованих запитів.*)<br>15. Робота з Neo4j (*Схема Neo4j. Структура зашифрованого запиту. Написання зашифрованих запитів. Створення та запит до вершин. Встановлення взаємозв'язку між вершинами. Використання транзакційного зашифрованого HTTP end-point. Використовувані протоколи.*)<br>16. Графові моделі Orient db (*Огляд Orient db. Мульти-модель систем керування базами даних (СКДБ). Моделі документів. Графові моделі. Orient db синтаксис. Типи транзакцій. ETL (Extract, Transform, Load). Можливості Orient db. Використання Orient db.*)<br>17. База даних Hbase (*Кластерна архітектура HBase. Відмінності архітектури HBase від інших розподілених файлових систем. Особливості побудови архітектури HBase. Hbase на Python. Операції HBase. Функції операцій HBase. Огляд переваг операцій HBase над іншими файловими системами.*)<br>18. Особливості розроблення веб-додатків і веб-сервісів із застосуванням технологій розподілених сховищ даних (Поняття веб-додатку та веб-сервісу. Розподілені бази даних. Сховища даних. Розподілені технології. Особливості розподілених сховищ даних)<br>19. Концепція Hadoop (*Знайомство з Hadoop. Історія Hadoop. Компоненти Hadoop. Вузли та демони Hadoop.* |

| | |
|---|---|
| | *Архітектура Hadoop. Hadoop характеристики. oogle File System (GFS). HDFS. Концепція та архітектура HDFS. Відмінності архітектури HDFS від інших розподілених файлових систем.* )<br>20. Особливості HDFS (*Особливості побудви архітектури HDFS. Операції HDFS. Функції операцій HDFS. Огляд переваг операцій HDFS над іншими файловими системами. Типи запитів. Мовна підтримка. MapReduce. Властивості MapReduce. Трекер задач. Hive. HiveQL. Hadoop Fea.*)<br>21. Різновид реляційних мов запиту (*Реляційні мови запитів. Запит як приклад (Query-by-Example). Опис запитів мовою QBE. Вибірка даних з умовою. Базові оператори мови SQL та особливості їх запису.* )<br>22. Формування SQL запитів (*Формування запитів мовою SQL. Вибірка рядків конструкцією WHERE. Сортування результатів (конструкція ORDER BY). Вкладені запити (підзапити). Запити для кількох зв'язків. Умовний ящик. Microsoft Access. Datalog.*)<br>23. Знайомство з Apache Spark (Виклики та рішення. Що таке Apache Spark? Модель Spark. Потужний стек – гнучка розробка. Компоненти Apache Spark)<br>24. Spark SQL (*Spark SQL. Інтерфейс програмування. Модель даних. Операції DataFrame. Запити рідних наборів даних. Функції, визначені користувачем. Оптимізація та виконання плану. Логічний план. Фізичний план. Фізичний план з предикатом Pushdown і Column Pruning. Генерація коду. Розширення. Spark MLlib конвеєр.*)<br>25. Поняття хмарної платформи Snowflake (*Поняття платформи даних. Переваги хмарних платформ даних. Використання платформ даних для бізнесу. Традиційні архітектури. Сучасна архітектура даних з Snowflake. Порівняльна характеристика Snowflake в порівнянні з іншими платформами.*)<br>26. Поняття стеку ELK (*ELK стек. ElasticSearch. Logstash. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. Різниця між ELK і EFK.*)<br>27. Поняття стеку EFK (*EFK стек. ElasticSearch. Fluentd. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. Різниця між ELK і EFK.*)<br>28. Хмарні сервіси з Snowflake (*Архітектура Snowflake. Хмарні сервіси. Збільшення існуючих озер даних. Низька затримка. Транзакція перетворення, що масштабується. Безпечний доступ до даних. Інтеграція з Snowflake. Snowflake з Tableau. Робота з Snowflake.*)<br>29. Оркестрування потоків операцій в Airflow (*Поняття Apache Airflow. Основні відомості та призначення. Поняття прямих ациклічних графів (DAG). Робочий процес. Airflow веб-сервер. Запуск DAG. Оператори Airflow. Налаштування Airflow операторів. Airflow сенсори. Проектне навчання. Використання Python для програмування DAG. Панель керування Airflow.*) |

| | |
|---|---|
| | 30. Організація конвеєрів великих даних на RabbitMQ (*Поняття RabbitMQ. Опис. Взаємозв'язок. Використання. Протокол AMQ. Масштабування RabbitMQ. Інтеграція з базами даних.*)<br>31. Організація конвеєрів великих даних на Kafka (*Поняття технології Apache Kafka. Поняття брокера. Поняття споживача. Взаємодія сервісів через меседж брокери. Порівняння Kafka і RabbitMQ.*)<br>32. Робота з Tableau (*Поняття Tableau. Візуалізація даних. Тип діаграми та потоки інформаційної панелі. Попередня уважна обробка. Заголовок і підказка. Наступні кроки та додаткові ресурси. Вибір правильного графіку. Порівняння методів графічного представлення інформації. Розподіли. Взаємозв'язок. Панель інструментів. Шарування. Тестування. Попередня уважна обробка даних. Кольорові гами. Вибір стилю та кольору. Заголовок. Контекст. Tooltip.* )<br>33. Маштабування даних (*Перша зустріч науки з великими даними. Поняття машинного навчання. Дані і знання. Великі масштаби даних. Надвеликі розміри моделей. Класичні алгоритми машинного навчання. Питання масштабування. Стратегія паралелізму. Використання MapReduce.* ) |
| **Опитування** | Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу. |

# СХЕМА КУРСУ

| Тиж. | Тема, план, короткі тези | Форма діяльності (заняття) | Література. Ресурси в Інтернеті | Завдання, год | Термін виконання, тиж. |
|---|---|---|---|---|---|
| 1 | **Вступ до великих даних.** Великі дані. Атрибути великих даних. Загальне використання. Принцип роботи. Проблеми Великих даних. Спільні характеристики. Джерела великих даних. Загальні поняття для накопичення та опрацювання великих даних. | лекція | 1 - 5 | 2 | кінець поточного тижня |
| | **Лаб 1.** Реєстрація та знайомство з Databricks | лаб. робота | 3-4 | 2 | кінець поточного тижня |
| | Природа та особливості великих даних | сам. робота | 1-2 | 4,4375 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| 2 | **Поняття розподілених систем.** Приклади типових розподілених систем. Компоненти розподілених систем. Інтернет та інтранет. Використання пристроїв у розподілених ситемах. Обмін ресурсів в WWW. Веб-сервери та веб-браузери. HTTP/HTTPS протоколи. Програмні та апаратні сервісні рівні. Клуєнт-серверні моделі. Веб-прохі сервери. Веб-застосунки. Мобільні агенти. Комп'ютерні мережі. Синхронні та асинхронні розподілені системи. Події в системі. Порядок подій в реальному часі. Канали зв'язку. Питання захисту. | лекція | 1-9 | 2 | кінець поточного тижня |
| | **Лаб 2.** Поняття керування великими даними на Databriks | лаб. робота | 3-4, 6-7 | 2 | кінець поточного тижня |
| | Проблеми побудови розподілених систем для роботи з великими даними | сам. робота | 8-9 | 4,4375 | кінець поточного тижня |
| 3 | **Концепція та компоненти Hadoop.** Знайомство з Hadoop. Історія Hadoop. Компоненти Hadoop. Вузли та демони Hadoop. Архітектура Hadoop. Hadoop характеристики. oogle File System (GFS). HDFS. Концепція та архітектура HDFS. Відмінності архітектури HDFS від інших розподілених файлових систем. Особливості побудви архітектури HDFS. Операції HDFS. Функції операцій HDFS. Огляд переваг операцій HDFS над іншими файловими системами. Типи запитів. Мовна підтримка. MapReduce. Властивості MapReduce. Трекер задач. Hive. HiveQL. Hadoop Fea. | лекція | 1-2, 8-16 | 2 | кінець поточного тижня |
| | **Лаб 3.** Організація конвеєру на Databriks для роботи із великими даними | лаб. робота | 6-7 | 2 | кінець поточного тижня |
| | Обливості проектування кластерів для опрацювання великих даних | сам. робота | 1-2, 9 | 4,4375 | кінець поточного тижня |
| 4 | **Парадигма MapReduce та доступ до даних.** | лекція | 1-2, 10 | 2 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| | Історія MapReduce. Розгляд системи MapReduce. Розгляд платформи Apache Pig. Apache Hive. Hive QL. Архітектра MRv1. Планувальник YARN. Що таке є мова запитів GraphQL. Переваги GraphQL. GraphQL в порівнянні REST. Схеми GraphQL. Схема визначення. Вирішувачі (Resovers). Запит даних. Повернення даних. Apollo GraphQL. | | | | |
| | **Лаб 4.** Знайомство з Pyspark | лаб. робота | 3, 5 | 2 | кінець поточного тижня |
| | Методи синхронізації процесів та даних в розподілених програмах | сам. робота | 1-2, 8-9 | 4,4375 | кінець поточного тижня |
| 5 | **Шаблони MapReduce.** Поняття шаблону. Питання проектування шаблонів MapReduce. Шаблон сумаризації. Числова сумаризація. Сумаризація з інвертованих індексом. Сумаризація обрахунків. Шаблон фільтрування. Фільтр Bloom. Шаблон організації даних. Структурований шаблон. Ієрархічний шаблон. Шаблон роздріблення та зв'язування. Шаблони об'єднання. Мета шаблони. | лекція | 11 | 2 | кінець поточного тижня |
| | **Лаб 5.** Побудова конвеєру машинного навчання на Spark | лаб. робота | 3, 5 | 2 | кінець поточного тижня |
| | Особливості побудови архітектури системи на основі технології MapReduce | сам. робота | 12-13 | 4,4375 | кінець поточного тижня |
| 6 | **Робота з MongoDB.** Модель даних. Типи моделей даних. Логічна модель. Створення фізичної моделі даних. Фізична модель даних для реляційних баз даних. Інструменти для моделювання даних. Гнучке моделювання даних за допомогою MongoDB Atlas. Приклад моделювання бази даних. Запити і агрегування. Детальніше про мову запитів в MongoDB. Розуміння зіставлення MongoDB та SQL. MongoDB SQL: термінологія. MongoDB SQL: Виконувані файли бази даних. MongoDB SQL: команди. | лекція | 6, 10-15 | 2 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| | Приклад CRUD функцій в MongoDB. Побудова запитів в MongoDB. | | | | |
| | **Лаб 6.** Основи структурованої потокової подачі даних | лаб. робота | 8-9, 14 | 2 | кінець поточного тижня |
| | Розподілені файлові сховища даних на основі MongoDB | сам. робота | 4-6, 8-10, 15 | 4,4375 | кінець поточного тижня |
| 7 | **Резидентна система управління базами даних Redis.** Розподілене сховище пар ключ-значення. Конфігурації Redis. Типи даних Redis. Команди з використанням ключа. Redis server. Redis on python. Backup. Безпека Redis. Redis benchmark. Транзакції Redis. Redis pipelining. Redis на docker. | лекція | 7-8 | 2 | кінець поточного тижня |
| | **Лаб 7.** Структурована потокова подача даних з Apache Spark | лаб. робота | 1-2, 8-9, 17 | 2 | кінець поточного тижня |
| | Поняття кешування в системах з великими даними | сам. робота | 11 | 4,4375 | кінець поточного тижня |
| 8 | **Робота з Neo4j.** Graph Databases. Приклади графових баз даних. Порівняння графових та реляційних баз даних. Графова база даних Neo4j. Neo4j браузер. Схема Neo4j. Структура зашифрованого запиту. Написання зашифрованих запитів. Створення та запит до вершин. Встановлення взаємозв'язку між вершинами. Використання транзакційного зашифрованого HTTP end-point. Використовувані протоколи. | лекція | 8, 9, 14 | 2 | кінець поточного тижня |
| | **Лаб 8.** Поняття сучасних форматів для зберігання великих даних | лаб. робота | 3, 9-10, 13 | 2 | кінець поточного тижня |
| | Особливості мови запитів SQL для великих даних | сам. робота | 15 | 4,4375 | кінець поточного тижня |
| 9 | **Бази даних Hbase та Orient db.** Особливості розроблення веб-додатків і веб-сервісів із | лекція | 3, 5, 8 | 2 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| | застосуванням технологій розподілених сховищ даних. Кластерна архітектура HBase. Відмінності архітектури HBase від інших розподілених файлових систем. Особливості побудови архітектури HBase. Hbase на Python. Операції HBase. Функції операцій HBase. Огляд переваг операцій HBase над іншими файловими системами. Огляд Orient db. Мульти-модель систем керування базами даних (СКДБ). Моделі документів. Графові моделі. Orient db синтаксис. Типи транзакцій. ETL (Extract, Transform, Load). Можливості Orient db. Використання Orient db. | | | | |
| | **Лаб 9.** Робота із дельта-форматом даних | лаб. робота | 3, 9-10, 13 | 2 | кінець поточного тижня |
| | Методи управління паралельним доступом до сховищ з великими даними | сам. робота | 9-10, 13 | 4,4375 | кінець поточного тижня |
| 10 | **Стеки великих даних.** ELK стек. ElasticSearch. Logstash. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. EFK стек. ElasticSearch. Fluentd. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. Різниця між ELK і EFK. | лекція | 7-12 | 2 | кінець поточного тижня |
| | **Лаб 10.** Побудова Delta Live Tables (DLT) | лаб. робота | 3 | 2 | кінець поточного тижня |
| | Поняття IoT систем | сам. робота | 9-11, 13-15 | 4,4375 | кінець поточного тижня |
| 11 | **Робота з Spark SQL.** Виклики та рішення. Що таке Apache Spark? Модель Spark. Потужний стек – гнучка розробка. Spark SQL. Інтерфейс програмування. Модель даних. Операції DataFrame. Запити рідних наборів даних. Функції, визначені користувачем. Оптимізація та виконання плану. Логічний план. Фізичний план. Фізичний план з предикатом Pushdown і Column | лекція | 9, 11, 15 | 2 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| | Pruning. Генерація коду. Розширення. Spark MLlib конвеєр. | | | | |
| | **Лаб 11.** Проектування SQL Warehouse на Azure/AWS/GCP | лаб. робота | 3-4, 6, 15, 18 | 6 | кінець 13-го тижня |
| | Резервування та відновлення у розподілених базах даних | сам. робота | 1-2, 9 | 4,4375 | кінець поточного тижня |
| 12 | **Хмарна платформа Snowflake.** Поняття платформи даних. Переваги хмарних платформ даних. Використання платформ даних для бізнесу. Традиційні архітектури. Сучасна архітектура даних з Snowflake. Архітектура Snowflake. Хмарні сервіси. Збільшення існуючих озер даних. Низька затримка. Транзакція перетворення, що масштабується. Безпечний доступ до даних. Інтеграція з Snowflake. Snowflake з Tableau. Робота з Snowflake. | лекція | 6-7 | 2 | кінець поточного тижня |
| | Особливості розподілених файлових систем Snowflake | сам. робота | 9-10 | 4,4375 | кінець поточного тижня |
| 13 | **Оркестрування потоків операцій в Airflow.** Поняття Apache Airflow. Основні відомості та призначення. Поняття прямих ациклічних графів (DAG). Робочий процес. Airflow веб-сервер. Запуск DAG. Оператори Airflow. Налаштування Airflow операторів. Airflow сенсори. Проектне навчання. Використання Python для програмування DAG. Панель керування Airflow. | лекція | 14, 18 | 2 | кінець поточного тижня |
| | Використання Airflow у конвеєрах обробки великих даних | сам. робота | 19 | 4,4375 | кінець поточного тижня |
| 14 | **Організація конвеєрів великих даних.** Поняття RabbitMQ. Опис. Взаємозв'язок. Використання. Протокол AMQ. Масштабування RabbitMQ. Інтеграція з базами даних. Поняття технології Apache Kafka. Взаємодія сервісів через меседж | лекція | 9, 20-22 | 2 | кінець поточного тижня |

| | | | | | |
|---|---|---|---|---|---|
| | брокери. Порівняння Kafka і RabbitMQ. | | | | |
| | **Лаб 12.** Побудова SQL Warehouse | лаб. робота | 3-4, 6, 15, 18 | 6 | кінець 16-го тижня |
| | Особливості роботи з RabbitMQ | сам. робота | 19 | 4,4375 | кінець поточного тижня |
| 15 | **Знайомство із Tableau.** Вступ. Візуалізація даних. Тип діаграми та потоки інформаційної панелі. Попередня уважна обробка. Заголовок і підказка. Наступні кроки та додаткові ресурси. Вибір правильного графіку. Порівняня методів графічного представлення інформації. Розподіли. Взаємозв'язок. Панель інструментів. Шарування. Тестування. Попередня уважна обробка даних. | лекція | 13, 21 | 2 | кінець поточного тижня |
| | Операції та функції операцій Google BigTable | сам. робота | 13 | 4,4375 | кінець поточного тижня |
| 16 | **Машинне навчання на великих даних.** Перша зустріч науки з великими даними. Поняття машинного навчання. Дані і знання. Великі масштаби даних. Надвеликі розміри моделей. Класичні алгоритми машинного навчання. Питання масштабування. Стратегія паралелізму. Використання MapReduce. Традиційна обробка даних. | лекція | 1-4, 8-9, 16-17 | 2 | кінець поточного тижня |
| | Моніторинг стану розподілених обчислювальних систем з великими даними | сам. робота | 22-23 | 4,4375 | кінець поточного тижня |