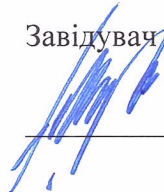


МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет електроніки та комп'ютерних технологій
Кафедра системного проектування

Затверджено

На засіданні
кафедри системного проектування
факультету електроніки та комп'ютерних
технологій
Львівського національного університету
імені Івана Франка
(протокол № 1 від 28.08.2023 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 2023 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 121 Інженерія програмного забезпечення (ВПК)
Викладачі дисципліни	Ляшкевич Василь Яремович, канд. техн. наук, доцент Стахіра Роман Йосипович, канд. фіз.-мат. наук, доцент
Контактна інформація	vasyl.lyashkevych@lnu.edu.ua , https://electronics.lnu.edu.ua/employee/liashkevych-v-ya-roman.stakhira@lnu.edu.ua https://electronics.lnu.edu.ua/employee/stakhira-roman-josypovych/
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 на основі аналізу актуальних проблем при роботі із великими даними у різних сферах їх застосування, що регулярно проводиться на відповідних конференціях чи семінарах. Серед найголовніших питань, що розглядаються в дисципліні, є питання базових понять великих даних, організації та використання розподілених технологій для опрацювання великих даних, управлінні ресурсами віддалених розподілених систем та сховищ великих та надвеликих даних, побудові конвеєрів для опрацювання великих даних, вирішенні питань аналітики та інших 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

	<p>теоретичних знань і набуття практичних умінь та навичок щодо застосування, налагодження й адміністрування систем на базі технологій великих даних та розподілених баз та сховищ даних, проектуванні надійних та економічно доцільних систем для збереження великих об'ємів даних.</p> <p>Цілями дисципліни є отримання практичних навичок та засвоєння методів опрацювання великих даних, створення кластерів даних та технології їх проектування для вирішення задач наук про дані на основі великих даних, наповнення даними та підтримання в робочому стані систем з даними.</p>
<p>Література для вивчення дисципліни</p>	<p>Основна література:</p> <ol style="list-style-type: none"> 1. Rajkumar Buyya, Rodrigo N. Calheiros, Amir Vahid Dastjerdi. Big Data: Principles and Paradigms: Elsevier, 2016. - 496 p. - Режим доступу: https://dphoto.lecturer.pens.ac.id/lecture_notes/internet_of_things/Big%20Data%20Principles%20and%20Paradigms.pdf 2. Катренко А.В., Системний аналіз об'єктів та процесів комп'ютеризації : Навчальний посібник./А.В. Катренко - Львів: "Новий світ-2000".-2003.-424с. 3. Michael Armbrust. Making Apache Spark better with Delta Lake: Databricks, 2020. - 399 p. 4. Gerardus Blokdyk. Databricks A complete Guide, 2021. - 205 p. - [Режим доступу]: https://www.everand.com/book/487839900/Databricks-A-Complete-Guide-2021-Edition 5. Tom White. Hadoop: The definitive Guide: O'Reilly, 2015. - 805 p. 6. Документація Apache Hadoop [Електронний ресурс] // Apache Hadoop. – 2021. – Режим доступу до ресурсу: https://hadoop.apache.org/docs/stable/. 7. Документація Apache Spark [Електронний ресурс] // Apache Spark. – 2019. – Режим доступу до ресурсу: https://spark.apache.org/docs/latest/. 8. Документація HBase [Електронний ресурс] // HBase. – 2019. – Режим доступу до ресурсу: https://hbase.apache.org/book.html. 9. RabbitMQ [Електронний ресурс] // RabbitMQ. – 2020. – Режим доступу до ресурсу: https://www.rabbitmq.com/documentation.html. 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 с. 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 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. 13. Kai Hwang, Min Chen. Big-Data Analytics for Cloud, IoT and Cognitive Computing: Willey, 2017. - 428 p. 14. Designing Distributed System. - [Режим доступу]: https://azure.microsoft.com/mediaservices/files/resourcefiles/designing-distributed-systems/Designing-Distributed-Systems.pdf 15. Kristina Chodorow. Scaling MongoDB: O'Reilly, 2011. - 58 p. 16. Joyce Kay Avila. Snowflake The Definitive Guide: O'Reilly, 2022. - 468 p. - [Електронний ресурс]. - Режим доступу: https://www.snowflake.com/wp-content/uploads/2022/03/Snowflake-The-Definitive-Guide-1.pdf 17. Google. Cloud Bigtable. - [Електронний ресурс]. - Режим доступу: https://cloud.google.com/bigtable 18. Rik Van Bruggen. Learning Neo4j: Packt Publishing, 2014. - 222 p. 19. 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 20. Alex Holmes. Hadoop in Practice: Manning Publications, 2012. - 537 p. - Режим доступу: https://ia600201.us.archive.org/7/items/HadoopInPractice/Hadoop%20in%20Practice.pdf 21. 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 22. Bas Harenslak, Julian de Rooter. 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_Rooter_Data_Pipelines_with_Apache.pdf 23. Apache HBase Team. Apache HBase™ Reference Guide. - [Електронний ресурс]. - Режим доступу: https://hbase.apache.org/apache_hbase_reference_guide.pdf 24. 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
<p>Обсяг курсу</p>	<p>Кількість кредитів ЄКТС: 4,5 (135 год), з них: 64 годин аудиторних занять (лекції: 32 год, лабораторні: 32 год.) та 71 год. самостійної роботи.</p>

Очікувані результати навчання

Після вивчення даного курсу здобувачі набувають таких Загальних та Фахових компетентностей та Програмних результатів навчання:

K01. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК02. Здатність застосовувати знання у практичних ситуаціях.

ЗК04. Здатність спілкуватися іноземною мовою як усно, так і письмово.

ЗК05. Здатність вчитися і оволодівати сучасними знаннями.

ФК13. Здатність ідентифікувати, класифікувати та формулювати вимоги до програмного забезпечення.

ФК14. Здатність брати участь у проектуванні програмного забезпечення, включаючи проведення моделювання (формальний опис) його структури, поведінки та процесів функціонування.

ФК15. Здатність розробляти архітектури, модулі та компоненти програмних систем.

ФК22. Здатність накопичувати, обробляти та систематизувати професійні знання щодо створення і супроводження програмного забезпечення та визнання важливості навчання протягом всього життя.

ФК24. Здатність здійснювати процес інтеграції системи, застосовувати стандарти і процедури управління змінами для підтримки цілісності, загальної функціональності і надійності програмного забезпечення.

ФК25. Здатність обґрунтовано обирати та освоювати інструментарій з розробки та супроводження програмного забезпечення.

ФК27. Здатність розробляти високопродуктивні програмні комплекси для вирішення задач наук про дані, систем штучного інтелекту, вбудованих та інших інноваційних систем.

ФК28. Володіння методами розроблення і впровадження систем підвищеної продуктивності, серверних, мікросервісних, хмаркових, розподілених та інших новітніх технологій.

ФК29. Здатність здійснювати розробку програмних продуктів використовуючи різні методології і технології із забезпеченням їх інноваційності та високої продуктивності.

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

ПРН09. Знати та вміти використовувати методи та засоби збору, формулювання та аналізу вимог до програмного забезпечення.

ПРН10. Проводити передпроектне обстеження предметної області, системний аналіз об'єкта проектування.

ПРН13. Знати і застосовувати методи розробки алгоритмів, конструювання програмного забезпечення та структур даних і знань.

ПРН14. Застосовувати на практиці інструментальні програмні засоби доменного аналізу, проектування, тестування, візуалізації, вимірювань та документування програмного

	<p>забезпечення.</p> <p>ПРН18. Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних.</p> <p>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.</p> <p>ПРН26. Знати засоби інтеграції, розгортання та підтримки спеціалізованих програмних компонентів, розроблених на основі інноваційних технологій для вирішення завдань високопродуктивних технологій.</p> <p>ПРН27. Знати основи інженерії й аналізу даних та вміти вибрати оптимальні алгоритми і технології для розробки інноваційних рішень при розв'язанні задач наук про дані, вбудованих систем та систем штучного інтелекту.</p>
Ключові слова	Великі дані, Конвеєри обробки великих даних, розподілені бази даних, кластери даних, сховища даних, озера даних, технологія MapReduce, Hive, Pig, Airflow, Apache Spark, 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)	<p>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.</p> <p>Бали нараховуються за видами робіт з співвідношенням:</p> <ul style="list-style-type: none"> • контрольні заміри (2 модулі): 70% семестрової оцінки; максимальна кількість балів: 70, а саме: <ul style="list-style-type: none"> - лабораторні роботи: 60% оцінки контрольного заміру; максимальна кількість балів: 42 (16 лабораторних робіт по 3 бали за кожну, де: від 0 до 2-х балів за виконання та від 0 до 1 балу за оцінювання чи тестування. Для тестування використовуються 5 вибраних випадкових питань з бази питань по 0.2 бали за питання.).

- теоретичний матеріал: 40% оцінки контрольного заміру; максимальна кількість балів: 28 (2 тести по 14 балів кожний, де вага питання складає 0.5 бала за правильну відповідь).

- залік: 30% семестрової оцінки, максимально 30 балів.

На заліку використовуються білети з 3 питань або тестування. При тестування використовуються 60 питань з вагою 0.5 бала за питання. На одне питання відводиться 45 секунд.

Оцінювання залікових питань:

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;

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;

5 балів - відстежується загальне розуміння розглянутої теми. Виявлені множинні неточності та невідповідності, пояснення наведеного коду відсутні, код функціонує із значними неточностями (або відсутні приклади запуску коду на виконання взагалі);

3 бали – студент погано розуміє розглянуту тему. Виявлені суттєві неточності та невідповідності. Наведені приклади коду з суттєвими недоліками, або не відповідають темі;

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.

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 незарахування викладачем, незалежно від масштабів плагіату чи обману.

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.

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.

Політика виставлення балів. Враховуються бали набрані на поточному тестуванні, самостійній роботі та бали підсумкового тестування. При цьому обов'язково

	<p>враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнь 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.</p> <p>Критерії оцінювання результатів неформальної освіти: Нарахування балів відбувається за написання студентом тез доповідей на конференціях, наукових статей, участь у діяльності наукових гуртків, участь у наукових семінарах та круглих столах, конкурсах, участь у заходах неформальної освіти за отримання сертифікатів про проходження навчання на різних освітніх платформах (Coursera, Prometheus тощо), курсах на провідних ІТ компаніях за тематикою навчальної дисципліни. Кількість балів визначається відсотком покриття результатів відповідної активності до вимог результатів навчання з навчальної дисципліни.</p>
<p>Питання до контрольних робіт</p>	<p>Перелік питань та завдань для проведення підсумкової оцінки знань певних тем до контрольних робіт:</p> <ol style="list-style-type: none"> 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.) 2. Парадигма MapReduce та доступ до даних (<i>Історія MapReduce. Розгляд системи MapReduce. Розгляд платформи Apache Pig. Apache Hive. Hive QL. Архітектура MRv1. Планувальник YARN.</i>) 3. Мова запитів GraphQL (<i>Що таке є мова запитів GraphQL. Переваги GraphQL. GraphQL в порівнянні REST. Схеми GraphQL. Схеми визначення. Вирішувачі (Resolvers). Запит даних. Повернення даних. Apollo GraphQL.</i>) 4. Шаблони у MapReduce (<i>Поняття шаблону. Питання проектування шаблонів MapReduce. Шаблон фільтрування. Фільтр Bloom. Шаблони об'єднання. Мета шаблони.</i>) 5. Шаблони сумаризації у MapReduce (<i>Шаблон сумаризації. Числова сумаризація. Сумаризація з інвертованих індексом. Сумаризація обрахунків.</i>) 6. Шаблони організації даних у MapReduce (<i>Шаблон організації даних. Структурований шаблон. Ієрархічний шаблон. Шаблон розділення та зв'язування.</i>) 7. Концепція розподілених систем (<i>Приклади типових розподілених систем. Компоненти розподілених систем. Інтернет та інтранет. Використання пристроїв у розподілених ситемах.</i>) 8. Протоколи обміну даними в розподілених системах (<i>Обмін ресурсів в WWW. Веб-сервери та веб-браузери. HTTP/HTTPS протоколи. Програмні та апаратні сервісні рівні. Клуєнт-серверні моделі. Веб-прохі сервери.</i>) 9. Веб-застосунки (<i>Веб-застосунки. Мобільні агенти. Комп'ютерні мережі. Синхронні та асинхронні</i>

розподілені системи. Події в системі. Порядок подій в реальному часі. Канали зв'язку. Питання захисту.)

10. MongoDB (Гнучке моделювання даних за допомогою MongoDB Atlas. Приклад моделювання бази даних. Запити і агрегування. Детальніше про мову запитів в MongoDB.)
11. MongoDB SQL (Розуміння зіставлення MongoDB та SQL. MongoDB SQL: термінологія. MongoDB SQL: Виконувани файли бази даних. MongoDB SQL: команди. Приклад CRUD функцій в MongoDB. Побудова запитів в MongoDB.)
12. 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.)
13. 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.)
14. 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.)
15. 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.)
16. 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)
17. 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.)
18. Особливості HDFS (Особливості побудви архітектури HDFS. Операції HDFS. Функції операцій HDFS. Огляд переваг операцій HDFS над іншими файловими системами. Типи запитів. Мовна підтримка. Hadoop Fea.)
19. 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.)

	<p>20. Знайомство з Apache Spark (Виклики та рішення. Що таке Apache Spark? Модель Spark. Потужний стек – гнучка розробка. Компоненти Apache Spark)</p> <p>21. Spark SQL (<i>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.</i>)</p> <p>22. 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.)</p> <p>23. Поняття стеку ELK (<i>ELK стек. Elasticsearch. Logstash. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. Різниця між ELK і EFK.</i>)</p> <p>24. Поняття стеку EFK (<i>EFK стек. Elasticsearch. Fluentd. Kibana. Опис. Взаємозв'язок. Використання. Принцип роботи. Різниця між ELK і EFK.</i>)</p> <p>25. Хмарні сервіси з Snowflake (<i>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.</i>)</p> <p>26. 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.)</p> <p>27. Організація конвеєрів великих даних на RabbitMQ (Поняття RabbitMQ. Опис. Взаємозв'язок. Використання. Протокол AMQ. Масштабування RabbitMQ. Інтеграція з базами даних.)</p> <p>28. Організація конвеєрів великих даних на Kafka (Поняття технології Apache Kafka. Поняття брокера. Поняття споживача. Взаємодія сервісів через меседж брокери. Порівняння Kafka і RabbitMQ.)</p> <p>29. 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.)</p>
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 - 10	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, 10	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.	лекція	1-9	2	кінець поточного тижня
	Лаб 2. Поняття керування великими даними на Databriks	лаб. робота	3-4	2	кінець поточного тижня
	Проблеми побудови розподілених систем для роботи з великими даними	сам. робота	5, 8-10	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.	лекція	1-2, 8-16	2	кінець поточного тижня

	Особливості побудви архітектури HDFS. Операції HDFS. Функції операцій HDFS. Огляд переваг операцій HDFS над іншими файловими системами. Типи запитів.				
	Лаб 3. Організація конвеєру на Databricks для роботи із великими даними	лаб. робота	3-4, 6-7	2	кінець поточного тижня
	Особливості проектування кластерів для опрацювання великих даних	сам. робота	1-2, 9	4,4375	кінець поточного тижня
4	Парадигма MapReduce та доступ до даних. Історія MapReduce. Розгляд системи MapReduce. Розгляд платформи Apache Pig. Apache Hive. Hive QL. Архітектура MRv1. Планувальник YARN. Мовна підтримка. MapReduce. Властивості MapReduce. Трекер задач. Hive. HiveQL.	лекція	1-2, 10	2	кінець поточного тижня
	Лаб 4. Знайомство з PySpark	лаб. робота	3, 7	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 розділення та зв'язування. Шаблони об'єднання. Мета шаблони.	лекція	1, 11	2	кінець поточного тижня
	Лаб 5. Побудова конвеєру машинного навчання на Spark	лаб. робота	3, 7	2	кінець поточного тижня
	Особливості побудови архітектури системи на основі технології MapReduce	сам. робота	11	4,4375	кінець поточного тижня

6	<p>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. Приклад CRUD функцій в MongoDB. Побудова запитів в MongoDB.</p>	лекція	6, 10-15	2	кінець поточного тижня
	<p>Лаб 6. Основи структурованої потокової подачі даних</p>	лаб. робота	8-9, 14	2	кінець поточного тижня
	<p>Розподілені файлові сховища даних на основі MongoDB</p>	сам. робота	8-10, 15-17	4,4375	кінець поточного тижня
7	<p>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.</p>	лекція	7-8	2	кінець поточного тижня
	<p>Лаб 7. Структурована потокова подача даних з Apache Spark</p>	лаб. робота	3, 7	2	кінець поточного тижня
	<p>Поняття кешування в системах з великими даними</p>	сам. робота	11	4,4375	кінець поточного тижня
8	<p>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</p>	лекція	8, 9, 18	2	кінець поточного тижня

	транзакційного зашифрованого HTTP end-point. Використовувані протоколи.				
	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	20-25	4,4375	кінець поточного тижня
9	Бази даних HBase. Особливості розроблення веб-додатків і веб-сервісів із 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.	лекція	8, 24	2	кінець поточного тижня
	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	1, 6-8	2	кінець поточного тижня
	Лаб 10. Побудова Delta Live Tables (DLT)	лаб. робота	3	2	кінець поточного тижня
	Поняття IoT систем	сам. робота	13	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.	лекція	3, 7	2	кінець поточного тижня

	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.				
	Лаб 11. Проектування конвеєру даних для міграції даних з NoSQL у SQL	лаб. робота	3-4, 6, 15, 18	6	кінець поточного 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	16	2	кінець поточного тижня
	Лаб 12. Програмування конвеєру для міграції даних з NoSQL у SQL	лаб. робота	3-4, 6, 15, 18	6	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	22	2	кінець поточного тижня

	Лаб 13. Проектування конвеєру даних для міграції даних з SQL у NoSQL сховище даних	лаб. робота	3-4, 6, 15, 18	6	кінець поточного тижня
	Використання Airflow у конвеєрах обробки великих даних	сам. робота	18	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 брокери. Порівняння Kafka і RabbitMQ.	лекція	1, 9, 20-22	2	кінець поточного тижня
	Лаб 14. Програмування конвеєру для міграції даних з SQL у NoSQL	лаб. робота	1-4, 6, 15, 18	6	кінець поточного 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. Порівняння методів графічного представлення інформації. Розподіли. Взаємозв'язок. Панель інструментів. Шарування. Тестування. Попередня уважна обробка даних.	лекція	17, 24	2	кінець поточного тижня
	Лаб 15. Проектування Data Warehouse для великих даних	лаб. робота	1-4, 6, 15, 18	6	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	лекція	1-4, 8-9, 16-17	2	кінець поточного тижня

	машинного навчання. Питання масштабування. Стратегія паралелізму. Використання MapReduce. Традиційна обробка даних.				
	Лаб 16. Проектування Data Lake для великих даних	лаб. робота	1, 3-4, 6, 15, 18	6	кінець поточного тижня
	Побудова сховищ даних для великих даних	сам. робота	1, 14, 19-23	4,4375	кінець поточного тижня