МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Факультет прикладної математики та інформатики
Кафедра обчислювальної математики

Затверджено
на засіданні
кафедри обчислювальної математики
факультету прикладної математики та інформатики
Львівського національного університету імені Івана Франка
(протокол № _1_ від _29 серпня_ 2023 р.)

Завідувач кафедри

Роман ХАПКО

Силабус з навчальної дисципліни «Організація та обробка великих даних», що викладається в межах ОПП Прикладна математика першого (бакалаврського) рівня вищої освіти для здобувачів зі спеціальності 113 Прикладна математика

Львів 2023 р.

Назва дисципліни	Організація та обробка великих даних
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1
Факультет та кафедра, за якою закріплена дисципліна	Факультет прикладної математики та інформатики Кафедра обчислювальної математики
Галузь знань, шифр та назва спеціальності	11 Математика та статистика 113 Прикладна математика
Викладачі дисципліни	Ільницька Ольга Володимирівна, кандидат фізико-математичних наук, асистент кафедри обчислювальної математики
Контактна інформація викладачів	Olha.ilnytska@lnu.edu.ua; https://ami.lnu.edu.ua/employee/o-v-ilnytska; Головний корпус ЛНУ ім. І. Франка, каб. 262, 361. м. Львів, вул. Університетська, 1
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/лабораторних занять (за попередньою домовленістю).
Сторінка курсу	
Інформація про дисципліну	Дисципліна «Організація та обробка великих даних» є вибірковою дисципліною з спеціальності 113 Прикладна математика для освітньої програми «Прикладна математика», яка викладається в 6-му семестрі (5 кредитів ECTS).
Коротка анотація дисципліни	Курс розроблено таким чином, щоб ознайомити студентів з підходами до роботи з великими даними: основ вилучення, трансформації та завантаження великих даних, створення ефективних запитів, роботи у хмарному середовищі, основи аналізу та візуалізації великих даних.
Мета та цілі дисципліни	Метою вивчення вибіркової дисципліни «Організація та обробка великих даних» є освоєння студентами основ вилучення, трансформації та завантаження великих даних, створення ефективних запитів, роботи у хмарному середовищі, основи аналізу та візуалізації великих даних.
Література для вивчення дисципліни	**Основна література** 1. Фостер Провост, Том Фоусет. Data Science для бізнесу. Як збирати, аналізувати і використовувати дані/ Ф. Провост , Т. Фоусет // Наш Формат. –2019 2. https://cloud.google.com/docs 3. Unwin, A. Why is Data Visualization Important? What is Important in Data Visualization? / A. Unwin // Harvard Data Science Review – 2020, 2(1). https://doi.org/10.1162/99608f92.8ae4d525 4. https://pandas.pydata.org/docs/ 5. https://numpy.org/doc/

<table>
<tr><td></td><td>Допоміжна література:
1. https://www.ibm.com/topics/etl
2. https://aws.amazon.com/what-is/etl/</td></tr>
<tr><td>Обсяг курсу</td><td>Загальний обсяг:
150 годин (аудиторних занять: 64 год., з них 32 год. лекцій та 32 год. лабораторних робіт; самостійної роботи: 86 год).</td></tr>
<tr><td>Очікувані результати навчання</td><td>Після завершення цього курсу студент буде :
Знати основи роботи з великими даними:
– вилучення, трансформації та завантаження;
– створення ефективних запитів;
– основні метрики даних різних галузей
– основні типи візуалізації.
Вміти:
– застосовувати вказані вище методи для побудови ETL процесів;
– працювати у хмарних середовищах;
– реалізовувати (програмно) алгоритми вивчених методів.</td></tr>
<tr><td>Ключові слова</td><td>Великі дані; вилучення, трансформація та завантаження великих даних; робота у хмарному середовищі; візуалізація даних; основні метрики різних галузей.</td></tr>
<tr><td>Формат курсу</td><td>Очний
Проведення лекцій, лабораторних занять і консультацій.</td></tr>
<tr><td>Теми</td><td>Подано нижче у таблиці Схема курсу «Організація та обробка великих даних».</td></tr>
<tr><td>Підсумковий контроль, форма</td><td>Залік.</td></tr>
<tr><td>Пререквізити</td><td>Для вивчення курсу студенти потребують базових знань з
- алгебри;
- програмування.</td></tr>
<tr><td>Навчальні методи та техніки, які будуть використовуватися під час викладання курсу</td><td>Презентації, лекції (лекції-бесіди, лекції-розповіді).
Індивідуальні завдання.</td></tr>
<tr><td>Необхідне обладнання</td><td>Комп’ютер із програмним забезпеченням python, Visual Studio або PyCharm, доступ до Internet мережі.</td></tr>
<tr><td>Критерії оцінювання (окремо для кожного виду навчальної діяльності)</td><td>Оцінювання проводиться за 100-бальною шкалою.
<table>
<tr><th colspan="2" rowspan="2">Оцінка за шкалою ECTS</th><th rowspan="2">Оцінка в балах</th><th colspan="3">Оцінка за національною шкалою</th></tr>
<tr><th colspan="2">Екзамен, диференційований залік</th><th>залік</th></tr>
<tr><td>A</td><td>Відмінно</td><td>100 - 90</td><td>Відмінно</td><td>5</td><td rowspan="2"></td></tr>
<tr><td>B</td><td>Дуже добре</td><td>81- 89</td><td>Добре</td><td>4</td></tr>
</table></td></tr>
</table>

C	Добре	71 -80			зараховано
D	Задовільно	61 - 70	Задовільно	3	
E	Достатньо	51- 60			
FX (F)	Незадовільно	0 - 50	Незадовільно	2	не зараховано

Впродовж семестру студент може отримати 100 балів. З них:
- **за роботу на лабораторних заняттях:** максимальна кількість – 70 балів (6 програм (індивідуальні завдання) по 5 балів та 4 програми (індивідуальні завдання) по 10 балів); для кожного завдання встановлено терміни здачі. Роботи, які здаються із порушенням термінів без поважних причин, оцінюються на нижчу оцінку (кожен блок тем на 2бали менше).
- **колоквіуми**: максимальна кількість – 30 балів (3 колоквіуми по 10 тестових завдань по 1б.).

Підсумкова максимальна кількість балів 100.

Критерії оцінювання індивідуальних завдань:

10 (5) балів	студент повністю виконав умови завдання, алгоритм реалізовано правильно, відповідає на всі запитання, пов'язані з тематикою завдання, проводить чіткий аналіз та порівняння отриманих результатів, пропонує інші підходи до вирішення поставленого завдання;
8-9 (4) балів	студент повністю виконав умови завдання, на деякі запитання, алгоритм реалізовано правильно, пов'язані з тематикою завдання, відповідає з незначними неточностями, проводить аналіз отриманих результатів з незначними неточностями;
6-7 (3) балів	студент виконав завдання з незначними помилками, але самостійно їх виправляє, якщо на них вкаже викладач, на деякі запитання, пов'язані з тематикою завдання, відповідає з неточностями, проводить аналіз отриманих результатів з неточностями;
4-5 (2) бали	студент виконав завдання частково, алгоритм реалізовано з помилками, які частково може виправити, якщо на них вкаже викладач, на запитання відповідає з помилками, проводить аналіз отриманих результатів з помилками;
2-3 бали	студент виконав завдання частково, алгоритм реалізовано з помилками, які самостійно не може виправити, переважно не відповідає на запитання;
1 бал	студент виконав завдання частково з грубими помилками, які самостійно не може виправити, демонструє незнання матеріалу;
0 балів	студент не виконав завдання.

	Критерії оцінювання тестових завдань (колоквіум): **1 бал:** відповідь на завдання правильна; **0 балів:** відповідь на завдання неправильна. **Академічна доброчесність**: Очікується, що 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 письмовій роботі студента є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману. **Відвідання занять** є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та лабораторні зайняття курсу. Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх індивідуальних завдань, 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 за індивідуальні завдання та за колоквіуми. При цьому обов'язково враховуються присутність на заняттях та активність студента під час лабораторного 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 і т. ін. Жодні форми порушення академічної доброчесності не толеруються.
Питання до колоквіумів.	1. Поняття великих даних: типи та класифікація. 2. Поняття ETL та ELT. 3. Способи вилучення даних. 4. Типи перетворення даних. 5. Типи завантаження даних. 6. Вичитка та збереження даних за допомогою пакету pandas. 7. Вичитка та збереження даних за допомогою пакету numpy. 8. CGP та GCP API: робота з BigQuery, способи створення таблиць, модифікація та видалення таблиць. 9. Віртуальні машини Linux: налаштування прав користувача, робота з файлами. 10. GCP SDK: передача файлів між машинами. 11. Агреговані метрики. 12. Типи та способи візуалізації даних.
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

Схема курсу «Організація та обробка великих даних»

Тиждень	Тема, план, короткі тези	Форма діяльності (заняття)	Література. Ресурси в інтернеті	Завдання, год.	Термін виконання
1	**Тема 1.** Поняття великих даних та їх класифікація. Концепції великих даних	лекція (2 год.)	[1] (додаткова [1,2])	Опрацювання лекційного матеріалу (3 год.)	1 тиждень
2-3	**Тема 2.** Вилучення, трансформація та завантаження даних. Вичитка даних різних форматів за допомогою python, numpy array, pandas dataframes, зберігання у різних форматах	лекція (4 год.)	[1,4,5]	Опрацювання лекційного матеріалу (6 год.)	1 тиждень
4-5	**Тема 3.** Google Cloud Platform: налаштування акаунту, огляд платформи та її можливості. Створення таблиць та написання запитів BigQuery	лекція (4 год.)	[2]	Опрацювання лекційного матеріалу (6 год.)	1 тиждень
6	**Тема 4.** GCP API: робота з новими таблицями	лекція (2 год.)	[2]	Опрацювання лекційного матеріалу (3 год.)	1 тиждень
7	**Тема 5.** GCP API: робота з існуючими таблицями.	лекція (2 год.)	[2]	Опрацювання лекційного матеріалу (3 год.)	1 тиждень
8	**Тема 6.** GCP API: потоковий запис у таблицю.	лекція (2 год.)	[2]	Опрацювання лекційного матеріалу (3 год.)	1 тиждень
9	**Тема 7.** Віртуальні машини Linux та GCP SDK: налаштування VM, робота з файлами	лекція (2 год.)	[2]	Опрацювання лекційного матеріалу (3 год.)	1 тиждень
10-11	**Тема 8.** Віртуальні машини Linux та GCP SDK: налаштування автоматичного запуску завдань та емейл сповіщення	лекція (4 год.)	[2]	Опрацювання лекційного матеріалу (6 год.)	1 тиждень
12-13	**Тема 9.** Основні метрики різних галузь: фінанси, маркетинг та мобільні додатки (CPI, RPI, ROAS, Retention, Conversion та інші)	лекція (4 год.)	[1]	Опрацювання лекційного матеріалу (6 год.)	1 тиждень
14	**Тема 10.** Основні метрики різних галузь: медицина	лекція	[1]	Опрацювання	1 тиждень

	(Індекс виживання, Disease-Free Survival, та інші)	(2 год.)		лекційного матеріалу (3 год.)	
15-16	**Тема 11.** Основи візуалізації даних з використанням хмарних технологій GCP Looker	лекція (4 год.)	[1, 2,3]	Опрацювання лекційного матеріалу (6 год.)	1 тиждень

1.	**Тема 1.** Вичитка даних різних форматів за допомогою python, numpy array, pandas dataframes, зберігання у різних форматах. (*Індивідуальне завдання №1. Вичитати дані форматів csv, txt, json за допомогою пакетів python numpy як narray, pandas як dstsframe. Зберегти дані у форматах csv, txt, json)*	лабораторне (2 год.)	[1,4,5]	Виконання індивідуального завдання №1 (3 год.)	під час заняття 2 тижні
2.	**Тема 2.** Google Cloud Platform: налаштування акаунту, написання запитів BigQuery *(Індивідуальне завдання №2. Написати SQL запит до тестових таблиць використовуючи платформу GCP)*	лабораторне (2 год.)	[2]	Виконання індивідуального завдання №2 (3 год.)	під час заняття 1 тиждень
3.	Здача індивідуальних завдань №1, 2	лабораторне (2 год.)			під час заняття
4.	**Тема 3.** GCP API: робота з новими таблицями. *(Індивідуальне завдання №3. Створити нову таблицю у середовищі BigQuery використовуючи GCP API)*	лабораторне (2 год.)	[2]	Виконання індивідуального завдання №3 (4 год.)	під час заняття 2 тижні
5.	**Тема 4.** GCP API: робота з існуючими таблицями. *(Індивідуальне завдання №4. Створити запит до існуючої таблиці BigQuery, змінити дані існуючої таблиці, видалити таблицю використовуючи GCP API.)*	лабораторне (2 год.)	[2]	Виконання індивідуального завдання №4 (4 год.)	під час заняття 1 тиждень
6.	Здача індивідуальних завдань №3,4	лабораторне (2 год.)			під час заняття
7.	**Тема 5.** GCP API: потоковий запис у таблицю.	лабораторне (2 год.)	[2]	Виконання індивідуального	під час заняття

	(Індивідуальне завдання №5. Реалізувати різницевий метод для крайових задач)			завдання №5 (4 год.)	1 тиждень
8.	Здача індивідуального завдання №5	лабораторне (2 год.)			під час заняття
9.	**Тема 6.** Віртуальні машини Linux та GCP SDK: налаштування VM, робота з файлами **Колоквіум 1** (*Індивідуальне завдання №6. Налаштувати Linux VM використовуючи пратформу GCP, перенести файл з локальної робочої станції на VM та назад використовуючи GSP SDK*)	лабораторне (2 год.)	[2]	Виконання індивідуального завдання №6 (4 год.)	під час заняття 1 тиждень
10.	Здача індивідуального завдання №6	лабораторне (2 год.)			під час заняття
11.	**Тема 7.** Віртуальні машини Linux та GCP SDK: налаштування автоматичного запуску завдань та емейл сповіщення. **Колоквіум 2** (*Індивідуальне завдання №7. Налаштувати автоматичний запуску завдання та емейл сповіщення про статус його виконання*)	лабораторне (2 год.)	[2]	Виконання індивідуального завдання №7 (4 год.)	під час заняття 2 тижні
12.	**Тема 8.** Основні метрики різних галузь: фінанси, маркетинг та мобільні додатки (CPI, RPI, ROAS, Retention, Conversionб та інші). (*Індивідуальне завдання №8. Використовуючи GCP API, створити запит до таблиці, який агрегує основні метрики та зберегти результат у нову таблицю*)	лабораторне (2 год.)	[1]	Виконання індивідуального завдання №8 (4 год.)	під час заняття 1 тиждень
13.	Здача індивідуальних завдань № 7, 8	лабораторне (2 год.)			під час заняття
14.	**Тема 9.** Основні метрики різних галузь: медицина (Індекс виживання, Disease-Free Survival, та інші). (*Індивідуальне завдання №9. Використовуючи GCP API, створити запит до таблиці, який агрегує основні метрики*	лабораторне (2 год.)	[1]	Виконання індивідуального завдання №9 (4 год.)	під час заняття 2 тижні

	та зберегти результат у нову таблицю)				
15.	**Тема 10.** Основи візуалізації даних з використанням хмарних технологій GCP Looker. (*Індивідуальне завдання №10. Побудувати візуалізацію результатів до лабораторних №8 та №9*)	лабораторне (2 год.)	[1, 2,3]	Виконання індивідуального завдання №10 (4 год.)	під час заняття 1 тиждень
16.	**Колоквіум 3** Здача індивідуальних завдань № 9, 10	лабораторне (2 год.)			під час заняття