

Шифр Спеціальності і: F3	Назва спеціальності, освітньої програми: Комп'ютерні науки, «Комп'ютерні науки»	Освітній рівень: Магістр
-----------------------------------	--	-----------------------------

«Затверджую»

Завідувачка кафедри
Тетяна ГОНЧАРЕНКО

Розробник силабуса
Ольга СОЛОВЕЙ



СИЛАБУС

БК «Хмарні сервіси та технології обробки великих даних»

(назва, шифр освітньої компоненти (дисципліни))

1) Статус освітньої компоненти: вибіркова

2) Контактні дані викладача: (зазначається посада, вчений ступінь, ПІБ викладача, корпоративна адреса електронної пошти, телефон, посилання на сторінку викладача на сайті КНУБА)

к.т.н., доцент Соловей Ольга Леонідівна, solovey.ol@knuba.edu.ua, +380 (67) 169 60 14
<https://www.knuba.edu.ua/elementor-161775/>

3) Пререквізити (дисципліни-попередники, які необхідно вивчити, щоб слухати цей курс):
Дисципліни, в яких вивчаються: основи програмування, алгоритми та структури даних, основи баз даних

4) Коротка анотація дисципліни

Дисципліна "Хмарні сервіси та технології обробки великих даних" присвячена вивченню сучасних технологій управління та обробки великих даних у хмарному середовищі з використанням платформи Databricks. Програма курсу охоплює два змістовні модулі, які послідовно розкривають як базові, так і поглиблені аспекти роботи з великими даними, включаючи архітектуру платформи, інструменти для інтеграції, зберігання, трансформації та автоматизації даних.

У першому змістовому модулі розглядаються основи управління даними, типи хмарних сховищ, можливості Unity Catalog, техніки завантаження даних (пакетне, потокове), роль SQL Warehouses у виконанні запитів, а також архітектура Databricks та інструменти моніторингу даних і виконання завдань.

Другий модуль зосереджений на техніках ефективного обробки даних за допомогою Apache Spark, Delta Lake і концепції Lakehouse (Дім Даних). Особливу увагу приділено побудові Medallion-архітектури, автоматизації робочих навантажень за допомогою Lakeflow Jobs, а також інтеграції SQL та Python у щоденну аналітичну практику.

Метою курсу є надання студентам практичних знань і навичок для проектування, реалізації та оптимізації рішень на базі сучасних хмарних платформ для роботи з великими даними.

5) Структура курсу: Лекції, лабораторні роботи, самостійна робота, РГР, залік

Загальна кількість кредитів ECTS	4,5
Сума годин:	135
Вид індивідуального завдання	РГР
Форма контролю	залік

6) Зміст курсу:

Лекція 1. Вступ до хмарних технологій та обробки великих даних: Платформи, типи сховищ та виклики.

Тема 1. Визначення великих даних (Big Data). Основні характеристики великих даних.

Тема 2. Проблеми та виклики обробки великих даних.

Тема 3. Огляд хмарних сховищ даних. Хмарні платформи для управління великими даними.

Тема 4. Рішення для управління даними в Databricks.

Тема 5. Роль Unity Catalog у керуванні наборами даних та метаданими.

Тема 6. Компоненти Unity Catalog для управління даними - каталог, схема, таблиці, представлення.

Лекція 2. Завантаження даних у Databricks

Тема 1. Визначення процесів завантаження даних у Databricks - пакетне та потокове завантаження. Вибір оптимальної техніки завантаження даних.

Тема 2. Використання SQL для програмного завантаження даних у Databricks.

Тема 3. Використання Databricks SQL для запитів, трансформацій та маніпулювання наборами даних.

Лекція 3. SQL Warehouses в Databricks та їх роль в виконанні запитів.

Тема 1. Створення таблиць, матеріалізованого подання (materialized view), динамічного подання (dynamic view).

Тема 2. Характеристики продуктивності SQL запитів - Photon Engine.

Тема 3. Виконання поширених SQL-трансформацій (агрегації, з'єднання, фільтрація) за допомогою Об'єднаного SQL-редактора.

Лекція 4. Структура та основні компоненти архітектури Databricks.

Тема 1. Компоненти та їх взаємодія в архітектурі Databricks.

Тема 2. Clusters: масштабування та управління ресурсами. Notebooks: інтерактивна робота з даними та кодування в середовищі Databricks.

Тема 3. Jobs: автоматизація та оркестрація завдань. Delta Lake: основи роботи з Delta таблицями, переваги для зберігання та обробки даних.

Тема 4. Використання сторонніх баз даних, файлових систем та інструментів аналітики з Databricks.

Лекція 5. Панелі моніторингу в архітектурі DataBricks

Тема 1. Визначення та роль панелей моніторингу у DataBricks. Особливості створення.

Тема 2. Приклад та переваги використання панелей моніторингу у DataBricks.

Тема 3. Визначення та основні функції Genie у DataBricks — інтелектуальний аналіз даних.

Змістовий модуль 2. Техніки та підходи для ефективної обробки даних хмарної платформи Databricks

Лекція 6. Apache Spark у Databricks: Основи, Інтеграція та Оптимізація для Обробки Великих Даних

Тема 1. Основні компоненти Apache Spark: Driver, Executor, Cluster Manager.

Тема 2. Основи роботи в Databricks з Apache Spark.

Тема 3. Використання Delta Lake в Spark для роботи з даними: транзакційна підтримка та управління даними.

Лекція 7. Delta таблиці та архітектура Medallion у Databricks

Тема 1. Основні концепції Delta таблиць: транзакційна підтримка, підтримка ACID властивостей, змінення історії даних.

Тема 2. Порівняння Delta таблиць із звичайними таблицями в Spark: швидкість, надійність, можливості для зберігання та управління даними.

Тема 3. Рівні архітектури Medallion - переваги архітектури Medallion для підтримки багаторівневої обробки та зберігання даних.

Тема 4. Техніки додавання метаданих під час завантаження даних у таблиці.

Лекція 8. Техніки та підходи для ефективної обробки даних - Дім-Даних (LakeHouse)

Тема 1. Основні принципи побудови Дому Даних (LakeHouse): інтеграція структурованих та неструктурованих даних. Переваги використання Lake House для зберігання та обробки великих даних.

Тема 2. Використання CREATE TABLE AS (CTAS) для завантаження даних.

Тема 3. Пояснення техніки завантаження за допомогою команди COPY INTO.

Тема 4. Використання MERGE INTO для обробки змін та конфліктів при завантаженні нових даних.

Тема 5. Автоматизація процесу завантаження даних на основі методу Auto Loader.

Лекція 9. Організація та автоматизація робочих навантажень за допомогою процесів Lakeflow Jobs у Databricks

Тема 1. Основні функції та можливості Lakeflow Jobs у платформі Databrick - створення та керування робочими процесами. Функції: планування завдань, моніторинг статусу, інтеграція з іншими інструментами.

Тема 2. Механізми автоматичного запуску завдань і налаштування розкладу виконання.

Тема 3. Принципи роботи з конвейерами завдань у Lakeflow Jobs. Основні інструменти для побудови контейнерів завдань.

Тема 4. Інтеграція Python та SQL у робочі процеси Lakeflow Jobs.

Теми лабораторних робіт

Лабораторна робота № 1. Налаштування і використання Unity Catalog для управління метаданими в Databricks.

Лабораторна робота № 2. Виконання процесів завантаження даних у Databricks (паketне та потокове завантаження).

Лабораторна робота № 3. Створення динамічних та матеріалізованих переглядів в SQL Warehouses. Оцінка продуктивності SQL запитів за допомогою Photon Engine.

Лабораторна робота № 4. Робота з компонентами Databricks: налаштування кластерів, створення та виконання ноутбуків.

Лабораторна робота № 5. Створення панелей моніторингу у DataBricks та використання DataBricks Genie для автоматизації аналізу даних

Лабораторна робота № 6. Робота з Delta Lake для збереження та обробки даних.

Лабораторна робота № 7. Порівняння Delta таблиць з традиційними таблицями в Spark.

Лабораторна робота № 8. Завантаження даних у Databricks Lake House за допомогою CTAS, COPY INTO, MERGE INTO та Auto Loader.

Індивідуальне завдання

Створення конвейеру завдань у Lakeflow Jobs

7) Посилання на сторінку електронного навчально-методичного комплексу

дисципліни: <https://org2.knuba.edu.ua/course/view.php?id=5324>