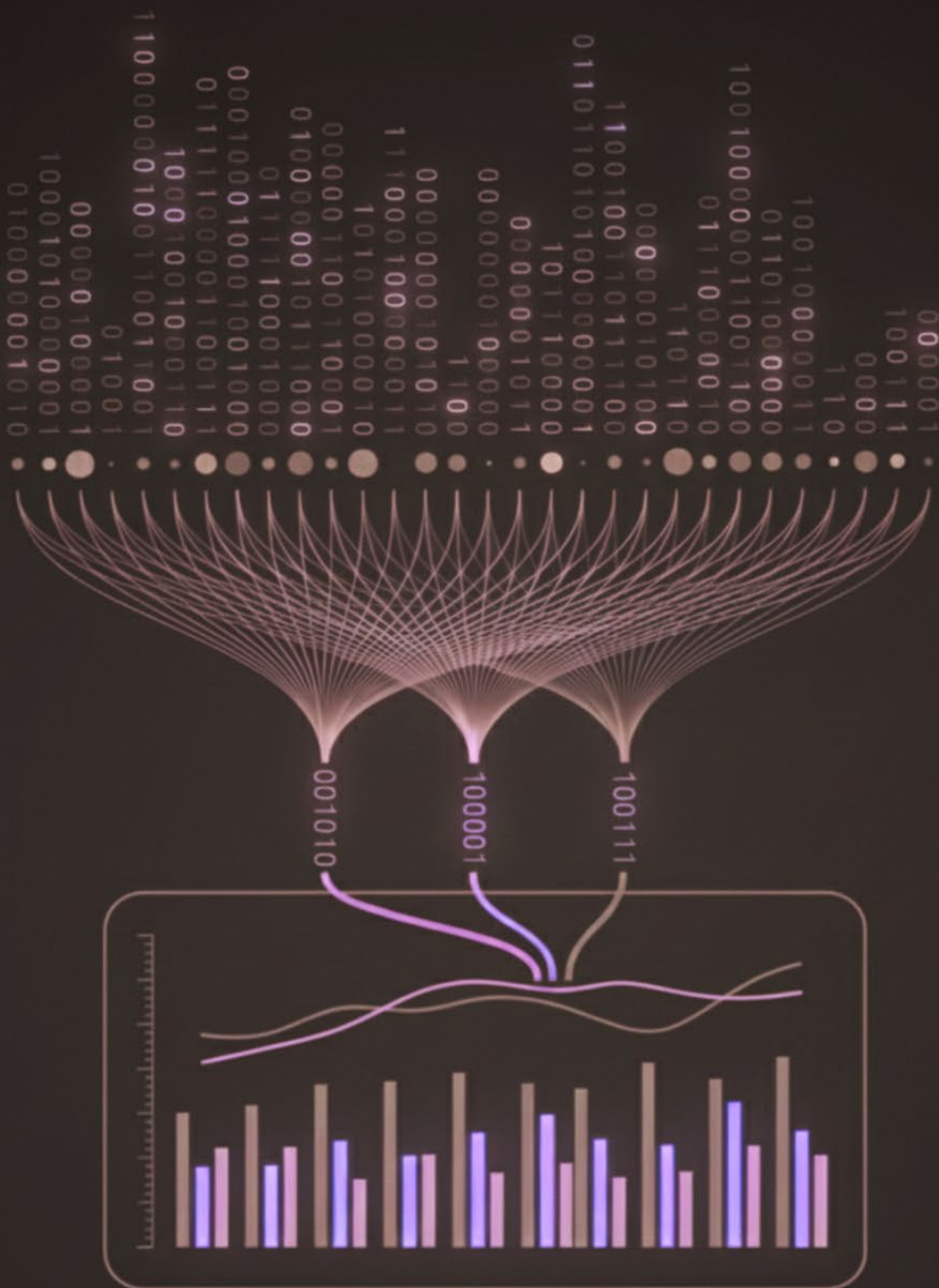




ОСНОВИ ІНЖИНІРИНГУ ДАНИХ

Ушенко Ю.О. Ватаманіца Е.В. Івашко В.В.

Ю. О. Ушенко, Е. В. Ватаманіца, В. В. Івашко

ОСНОВИ ІНЖИНІРИНГУ ДАНИХ

Навчально-методичний посібник

(видання електронне)

ЧЕРНІВЦІ 2025

УДК 004.6, 004.8

Т-56

Рекомендовано до друку вченою радою Навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького національного університету імені Юрія Федьковича (протокол № 17 від 18 грудня 2025)

Рецензенти:

Шпирка В.І., провідний розробник програмного забезпечення в компанії «ТОВ ІНТЕЛІАС»

Валь О.Д., кандидат фізико-математичних наук, доцент кафедри програмного забезпечення комп'ютерних систем Чернівецького національного університету імені Юрія Федьковича

Ушенко Ю.О., Ватаманіца Е.В., Івашко В.В.

Т-56 Основи інжинірингу даних / Ушенко Ю.О. , Ватаманіца Е.В., Івашко В.В. –

Чернівці: Чернівецький національний університет, 2025. – 220с.

Це видання є комплексним посібником для студентів, які прагнуть опанувати повний життєвий цикл роботи з інформацією в сучасних технологічних екосистемах. Видання детально розглядає фундаментальні принципи моделювання та архітектури даних, навчаючи читача будувати масштабовані та гнучкі системи. Особливу увагу приділено основам Kafka як центральному елементу для обробки подій у реальному часі, а також проектуванню ефективних потоків і конвеєрів даних (pipelines), що забезпечують безперебійне переміщення та трансформацію даних.

Окрім технічних аспектів переміщення даних, автори заглиблюються у стратегії вибору та експлуатації сучасних сховищ, від класичних реляційних баз до хмарних озер даних. Видання призначене для студентів спеціальності F3 “Комп’ютерні науки” та всіх, хто бажає ознайомитися з інжинірингом даних.

УДК 004.6, 004.8

©Чернівецький національний університет , 2025

Зміст

1 Вступ у Інжиніринг даних	7
1.1 Що таке інжиніринг даних?.....	7
1.1.1 Визначення інженерії даних	7
1.1.2 Вступ до життєвого цикл інженерії даних	8
1.1.3 Еволюція інженера даних.....	9
1.1.4 Інженерія даних та наука про дані	14
1.2 Навички і діяльність для інжинірингу даних.....	15
1.2.1 Набір навичок інженера даних	15
1.2.2 Зрілість даних та інженер даних.....	17
1.2.3 Ділові та технічні обов'язки інженера з даних	21
1.2.4 Континуум ролей інженера з даних: від А до Б.....	24
1.3 Життєвий цикл інжинірингу даних	26
1.3.1 Що таке життєвий цикл інженерії даних?	26
1.3.2 Генерація: Вихідні системи	28
1.3.3 Зберігання	31
1.3.4 Завантаження даних.....	34
1.3.5 Трансформація	38
1.3.6 Подача	39
1.3.7 Машинне навчання	42
1.3.8 Зворотній ETL	44
1.3.9 Основні підводні течії в життєвому циклі інженерії даних.....	45
1.3.10 Безпека	45
1.3.11 Управління даними.....	46
1.3.12 DataOps	56
1.3.13 Архітектура даних	61
1.3.14 Оркестрація.....	62
1.3.15 Інженерія програмного забезпечення	64
2 Моделювання і архітектура даних.....	68
2.1 Підхід до моделювання даних.....	68
2.2 Рівні моделювання даних	69
2.3 Методи моделювання даних.....	71
2.3.1 Нормалізація.....	71
2.3.2 Методи моделювання серійних аналітичних даних	76
2.3.3 Моделювання потокових даних.....	92
2.4 Якість даних та забезпечення цілісності	94
2.4.1 Що таке якість даних	94
2.4.2 Показники якості даних.....	95
2.4.2.1 Правильні назви заголовків і стовпців.	96
2.4.2.2 Правильне форматування файлів.....	96

2.4.2.3 Правильні типи даних	97
2.4.2.4 Діапазони значень і цілісність значень.	98
2.5 Політики управління даними	98
2.5.1 Що таке управління даними?.....	98
2.5.2 Які переваги управління даними для бізнесу?	99
2.5.3 Ключові елементи управління даними	100
2.5.3.1 Каталогізація даних.....	100
2.5.3.2 Якість даних.....	100
2.5.3.3 Класифікація даних	101
2.5.3.4 Безпека даних.....	101
2.5.3.5 Аудит прав на доступ до даних.....	102
2.5.3.6 Походження даних	102
2.5.3.7 Виявлення даних	103
2.5.3.8 Обмін даними та співпраця	103
3 Основи Kafka	105
3.1 Повідомлення та пакети.....	105
3.1.1 Схеми	106
3.2 Теми та розділи(Topics and partitions)	107
3.3 Виробники і споживачі	108
3.4 Брокери і кластери.....	109
3.4.1 Кілька кластерів	111
3.5 Переваги Apache Kafka	112
3.6 Екосистема даних	114
3.6.1 Варіанти використання.....	116
3.7 Походження Кафки	118
3.7.1 Народження Кафки.....	119
3.8 Написання повідомлень до Кафки.....	120
3.8.1 Створення виробника Kafka.....	123
3.9 Надсилання повідомлення до Kafka	125
3.9.1 Синхронне надсилання повідомлення	126
3.9.2 Надсилання повідомлення асинхронно	128
3.10 Налаштування виробників.....	129
3.10.1 Час доставки повідомлень.....	131
3.10.2 Тип стиснення	135
4 Потік і конвеєр даних.....	138
4.1 Основи переміщення даних.....	139
4.1.1 Передача даних від виробників до споживачів.....	139
4.1.2 Немає єдиного стандарту для передачі даних.....	139
4.1.3 Програмування на основі потоків(Flow-Based Programming)	140
4.1.4 Еволюція підходів до переміщення даних.....	142
4.1.5 Потік даних та виклики пов'язані з ним	142

4.1.6 Що таке конвеєр?	144
4.2 Архітектура конвеєра даних	145
4.2.1 Чому архітектура конвеєра даних є важливою?	145
4.2.1.1 З архітектури конвеєра даних та їх еволюція	146
4.2.2 Поширені рішення для конвеєрів даних у сучасному стеку даних	147
4.2.2.1 Зберігання та обробка даних	148
4.2.2.2 Введення даних	149
4.2.2.3 Оркестрування даних	150
4.2.2.4 Трансформація даних	150
4.2.2.5 Бізнес-аналітика та аналітика	151
4.2.3 7 найкращих практик архітектури конвеєра даних	151
4.3 Виклики масштабування та оптимізація продуктивності	152
4.3.1 Розуміння оптимізації конвеєрів даних	153
4.3.1.1 Виявлення вузьких місць у вашому каналі передачі даних	153
4.3.1.2 Розуміння важливості оптимізації конвеєра даних	154
4.3.1.3 Як виміряти продуктивність конвеєра даних	154
4.3.2 Кращі практики оптимізації конвеєра даних	155
4.3.2.1 Оптимізація обробки та зберігання даних	155
4.3.2.2 Використовуйте інструменти оркестрування даних для ефективних конвеєрів даних	155
4.3.2.3 Впровадьте перевірку якості даних та моніторинг	156
4.3.3 Інструменти та методи оптимізації конвеєра даних	156
4.3.3.1 Огляд популярних інструментів для оптимізації конвеєра даних	156
4.3.3.2 Методи оптимізації обробки та зберігання даних	157
4.3.3.3 Використання інструментів оркестрування даних для управління конвеєрами даних	157
5 Сховища	158
5.1 Системи зберігання даних	158
5.1.1 Зберігання файлів	160
5.1.2 Блок-сховище	163
5.1.3 Зберігання об'єктів	168
5.1.4 Кеш-пам'ять і системи зберігання даних на основі пам'яті	175
5.1.5 Розподілена файлова система Hadoop	176
5.1.6 Потокowe сховище	177
5.1.7 Індеси, розбиття на розділи та кластеризація	178
5.2 Абстракції зберігання інженерних даних	180
5.2.1 Сховище даних	181
5.2.2 Озеро даних	182
5.2.3 Data Lakehouse	182
5.2.4 Платформи даних	183
5.2.5 Архітектура потокового сховища для пакетного зберігання даних	184

5.3 Тренди у сфері зберігання даних	184
5.3.1 Каталог даних.....	184
5.3.2 Обмін даними.....	186
5.3.3 Схема.....	186
5.3.4 Відокремлення обчислень від зберігання.....	187
5.3.5 Життєвий цикл зберігання даних і збереження даних	192
5.3.6 Сховище для одного орендаря чи для багатьох	196
5.4 Безпека даних та відповідність регуляторним вимогам	199
5.4.1 Конфіденційність	199
6 DataOps та автоматизація процесів.....	200
6.1 Паралелі між DevOps і DataOps	200
6.1.1 Спільні принципи.....	201
6.1.2 Зміни для даних.....	202
6.2 Важливість DataOps	202
6.2.1 Швидкість отримання результату	203
6.2.2 Ефективність співпраці	203
6.2.3 Якість і надійність даних	204
6.3 Виклики, на які відповідає DataOps.....	205
6.3.1 Масштабованість.....	205
6.3.2 Різноманітність даних	206
6.3.3 Обробка в режимі реального часу	206
6.3.4 Безпека даних та відповідність вимогам	207
6.4 Кращі практики впровадження DataOps	207
6.4.1 Встановлення цілей	208
6.4.2 Створення правильної команди:.....	208
6.4.3 Вибір правильних інструментів.....	209
6.4.4 Просування культури DataOps	209
6.5 Знайомство з Apache Airflow.....	210
6.5.1 Конвеєри даних у вигляді графів.....	210
6.5.2 Виконання конвеєра даних у вигляді графа	210
6.5.3 Пайплайн графів проти послідовних скриптів.....	212
6.5.4 Гнучке визначення конвеєрів кодом(Python)	212
6.5.5 Планування та виконання конвеєрів	213
6.5.6 моніторинг та обробка збоїв	214
6.5.7 Коли використовувати Airflow	215
7 Практичні завдання	216
7.1. Проектування та підготовка (Тема 1 і 2).....	216
7.2. Реалізація Streaming шару з Kafka (Тема 3).....	216
7.3. Побудова конвеєра даних (ETL) (Тема 4).....	217
7.4. Організація сховища даних (Тема 5)	217
7.5. Оркестрація та автоматизація (Тема 6)	217

Навчальне видання
(електронне)

Ушенко Юрій Олександрович, Ватаманіца Едгар Вадимович, Івашко Віктор Вікторович

ОСНОВИ ІНЖИНІРИНГУ ДАНИХ

Літературний редактор: О. В. Лупул

Папір офсетний. Формат 60x84/16.

Умов. друк. арк.. 13,02. Обл.- вид. арк. 11,2. Тираж – 50.

Видавець та виготівник: Чернівецький національний університет
Свідоцтво суб'єкта видавничої справи ДК №1841 від 10.06.2004 р.
58000, м. Чернівці, вул. І. Франка, 20, оф.18, тел. (0372) 55-05-85