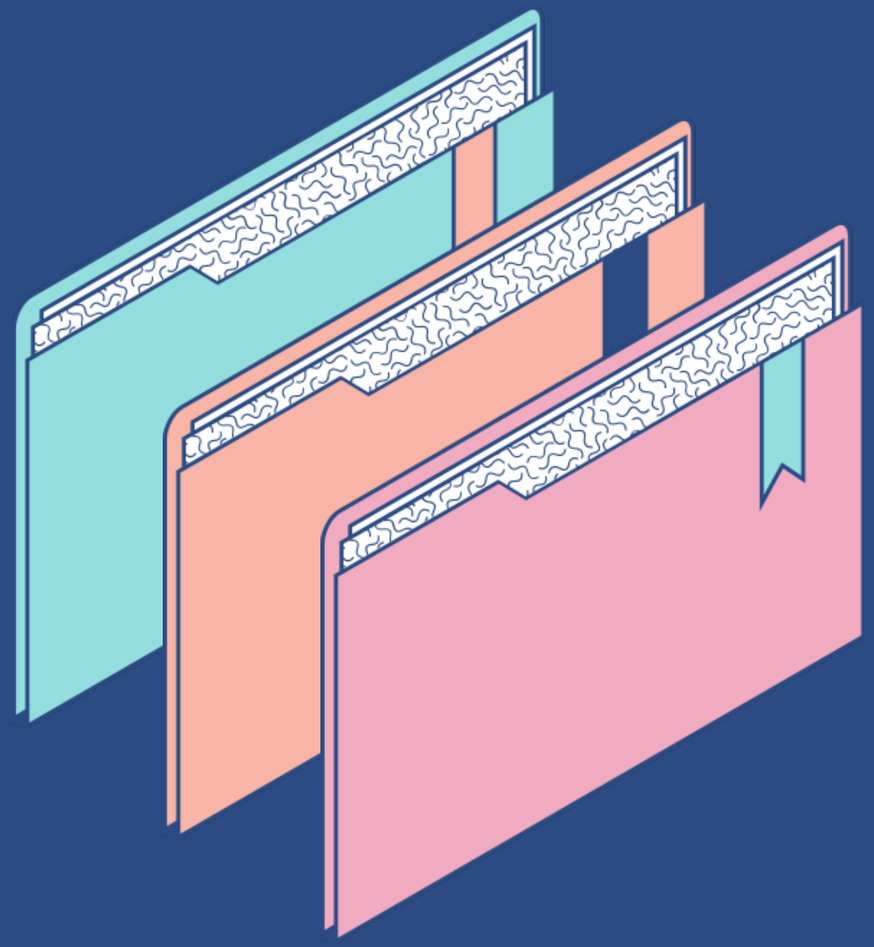
ПРЕЗЕНТАЦІЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕННЯ У РАМКАХ ДИПЛОМНОЇ РОБОТИ НА ТЕМУ:

Оптимізація побудови векторних і пошукових індексів для корпусів знань за визначеними тематиками (розподілені системи)

Підготував Білокінь Єгор, ІПЗ-4



Зміст

ВСТУП

АКТУАЛЬНІСТЬ

Великий об’єм інформації, яку необхідно структурувати та індексувати, зростання попиту користувачів для пошуку інформації формує завдання побудови та оптимізації пошукової системи для компаній та розробників.

На 65% зросла кількість пошукових запитів у Google

230 000 Тб нових даних з’являється щохвилини

ВСТУП

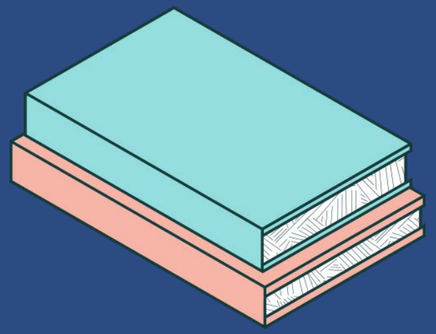
Об’єкт та предмет дослідження

- Оптимізація побудови пошукових індексів.
- Розподілені системи для побудови пошукових та векторних індексів.

Мета дослідження

Досягти максимальної ефективності побудови індексів за допомогою розподілених систем.

Загальні відомості

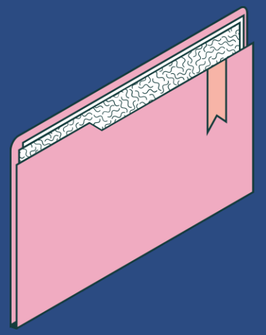


КОРПУС ЗНАНЬ

Колекція даних, що об’єднана за темою або категорією даних, що містяться у цій колекції.

СИСТЕМА ІНФОРМАЦІЙНОГО ПОШУКУ

Працює із індексами, що будуються для кожної одиниці інформаційних даних та здійснює пошук за допомогою різноманітних методів, що використовують побудований індекс для пошуку найкращого результату.



ПОШУКОВИЙ ІНДЕКС

Структура даних, що містить інформацію про об’єкт індексування (документ).

Проблематика

РЕСУРСИ

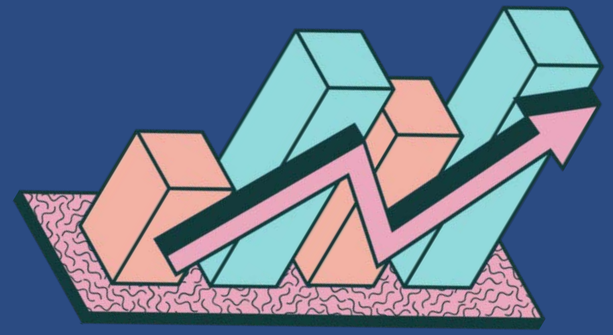
СІП вимагає значних ресурсів апаратного забезпечення: оперативної пам’яті, місткості накопичувача та іншого, витрати яких значно зростають із збільшенням інформації, що міститься у колекції.

МАСШТАБУВАННЯ

Дані постійно оновлюються та додаються нові

ШВИДКІСТЬ ТА ЯКІСТЬ ПОШУКУ

Оптимізація швидкодії побудови індексів може вплинути на якість пошуку в обраному корпусі

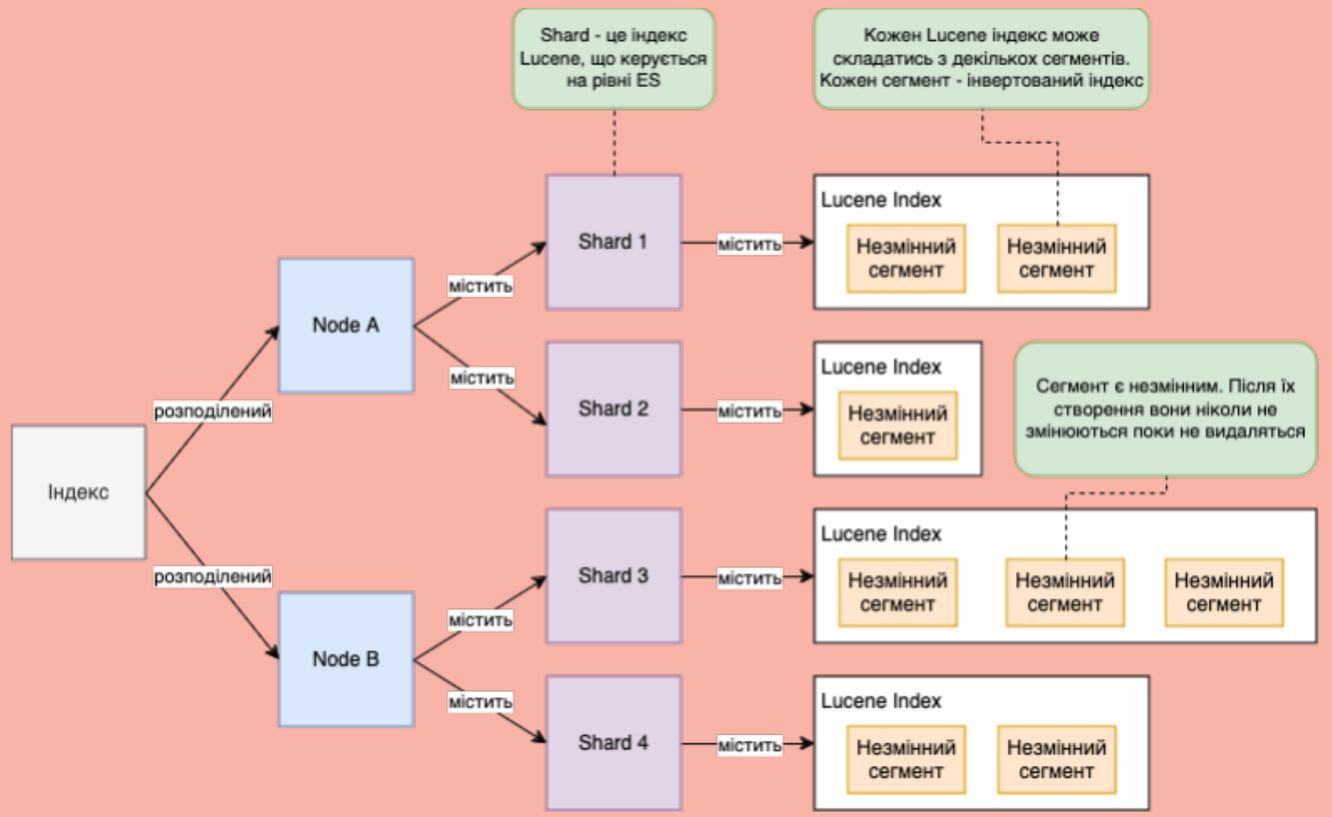


ElasticSearch

ІНТЕГРОВАНИЙ З LUCENE ДВИГУН ДЛЯ ПОВНОТЕКСТОВОГО ПОШУКУ

- Lucene виконує пошукову роль, тоді як ES забезпечує масштабованість, доступність, REST API та структури даних.
- Редагування або видалення документу, який присутній у сегменті Lucene, замість зміни попереднього створюється новий сегмент.

АРХІТЕКТУРА



ES розбиває кожен індекс Lucene на доступні вузли. Shard може бути основним shard або shard-копією. Кожен shard є індексом Lucene та може мати кілька сегментів, кожен з яких є інвертованим індексом

Огляд розподілених систем для побудови індексів

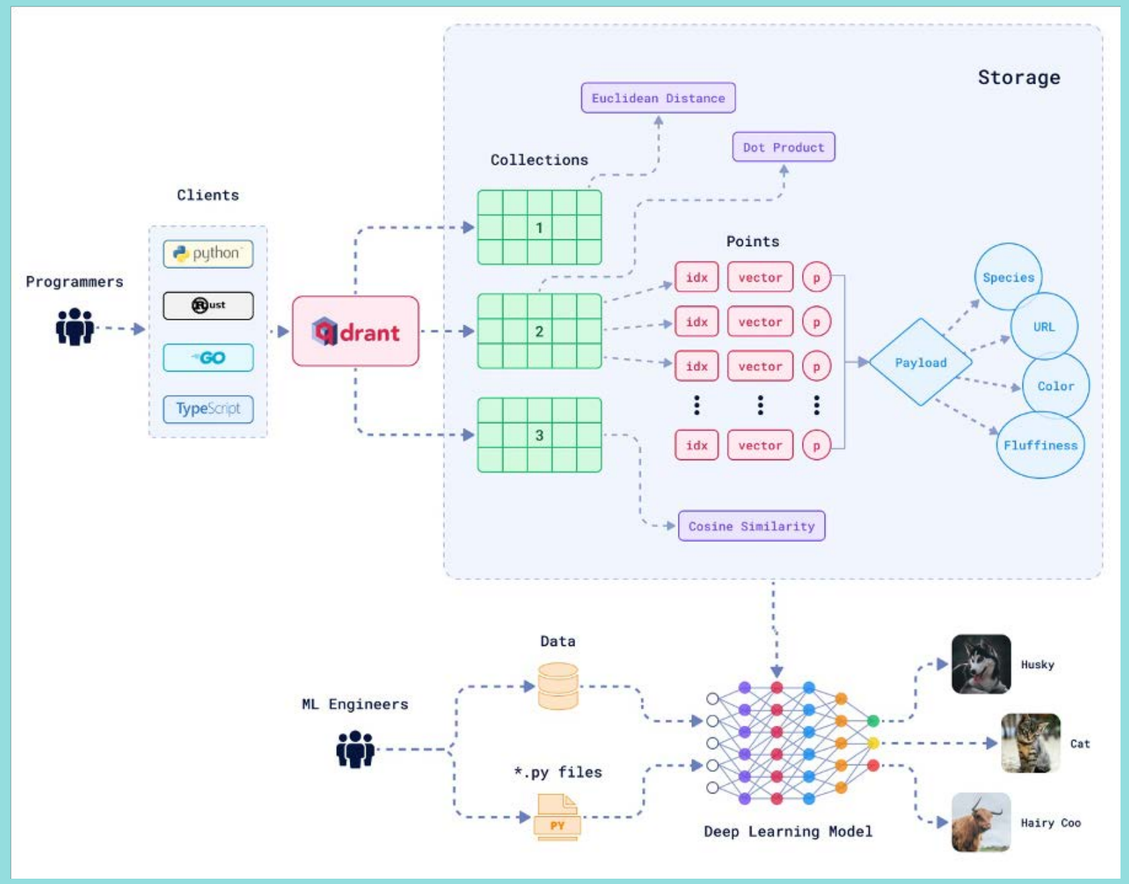
Qdrant

ВЕКТОРНА БД З ПОШУКОВОЮ СИСТЕМОЮ СХОЖОСТІ ВЕКТОРІВ

- Векторні БД оптимізовано для ефективного зберігання та запитів до високовимірних векторів.
- Точки - головна сутність, з якою працює Qdrant. Точки складаються з вектора, ідентифікатора та корисного навантаження

Огляд розподілених систем для побудови індексів

АРХІТЕКТУРА



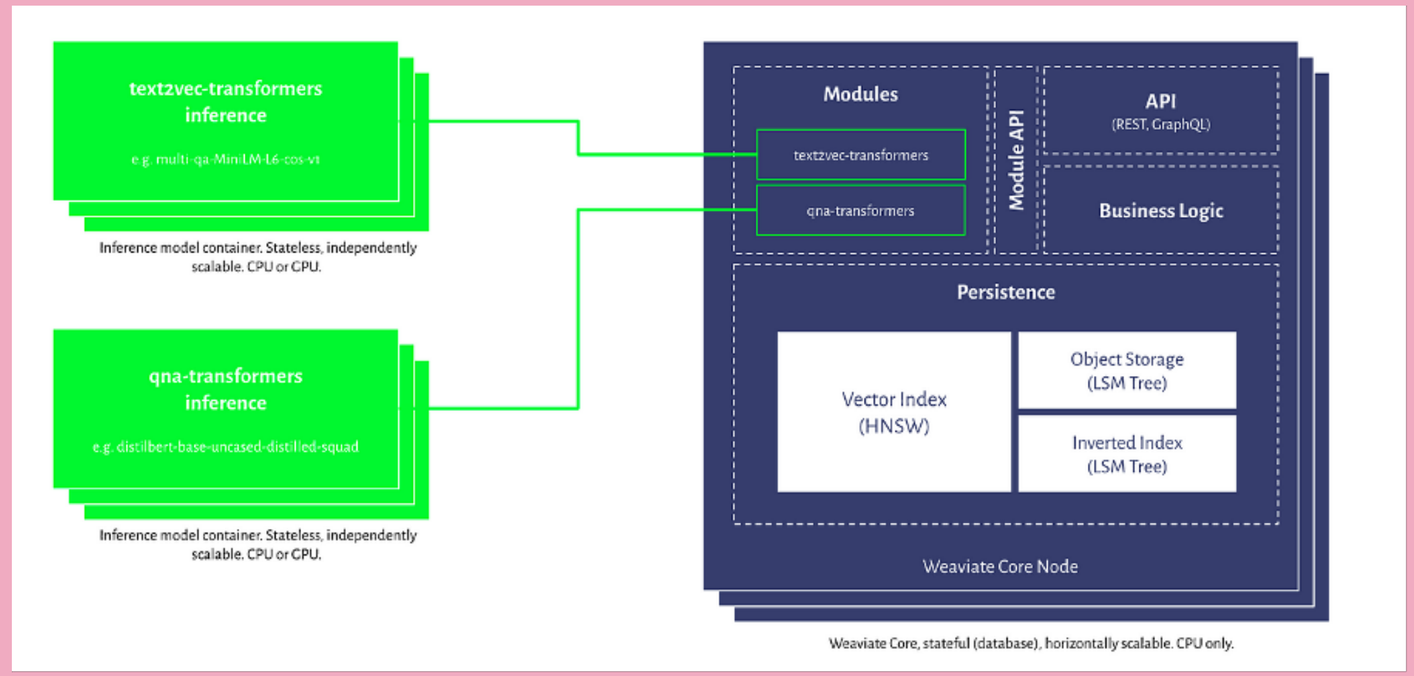
Представлення даних у векторних БД називають векторами, що є стисненою версією даних. Забезпечують швидкий пошук подібності та семантичний пошук, одночасно дозволяючи користувачам знаходити вектори, які є найближчими до даного вектора запиту на основі деякої метрики відстані.

Weaviate

ВЕКТОРНА БД, ЯКА ЗБЕРІГАЄ ЯК ОБ’ЄКТИ, ТАК І ВЕКТОРИ

- Weaviate зберігає об’єкти даних у колекціях на основі класів. Об'єкти даних представлені у вигляді JSON-документів.
- Weaviate забезпечує модульну структуру. Ключові функції винесені та обробляються додатковими модулями.
- Індексування у Weaviate підтримується за допомогою двох типів індексів.

АРХІТЕКТУРА



Сховище складається із shard’s, певних контейнерів компонентів. Кожен shard містить об’єкт, його вектор та інвертований індекс. Кожна операція запису негайно зберігається, а також стійка до збоїв програми та системи. За запитом на векторний пошук Weaviate повертає весь об’єкт. Об'єкти та їх вектори можна оновлювати або видаляти за бажанням; навіть під час читання з БД.

Огляд розподілених систем для побудови індексів

Очищення тексту для побудови індексу

1 КРОК

Аналіз даних

дослідження вхідних даних та їх характеристик

2 КРОК

Нормалізація

перетворення вхідних даних у формат за вимогами

3 КРОК

Токенізація

процес розподілу вхідних даних на менші одиниці

4 КРОК

Стоп-слова

видалення неінформативних слів

5 КРОК

Стеммінг

відсікання суфіксів слів

Кодування даних у вектор

Sentence-BERT

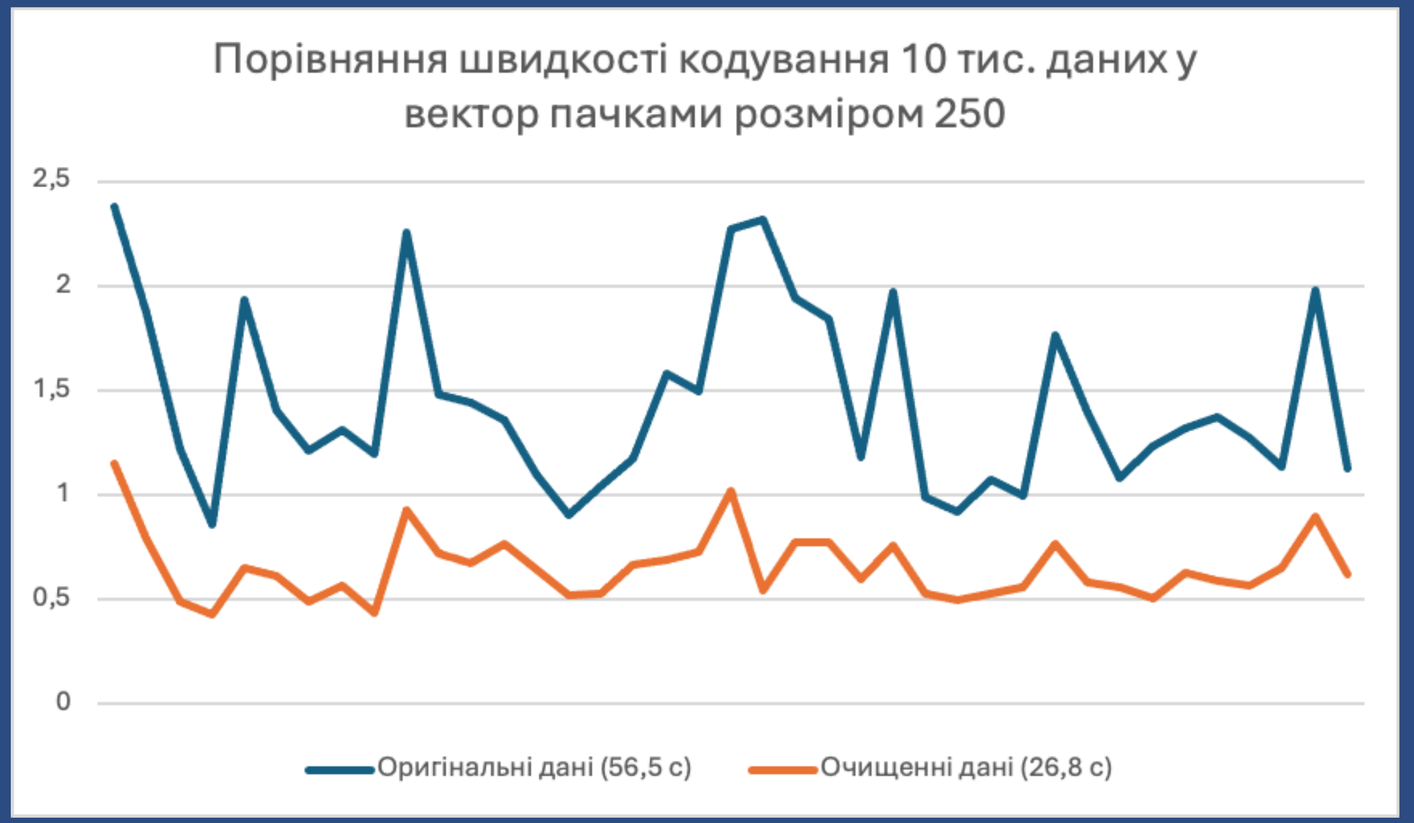
Використовує треновану модель, що значно пришвидшує процес перетворення даних у вектор.

LLM для кодування

Вирішено зупинитись на моделі «multi-qa-MiniLM-L6-cos-v1», що призначена для семантичного пошуку та має найкращі показники співвідношення швидкості з втратами точності пошуку.

Оптимізація шляхом паралельності

SentenceTransformers multiprocess pool



Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
paraphrase-MiniLM-L3-v2	62.29	39.19	50.74	19000	61 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1	64.33	51.83	58.08	14200	80 MB

Оптимізація розподілених систем

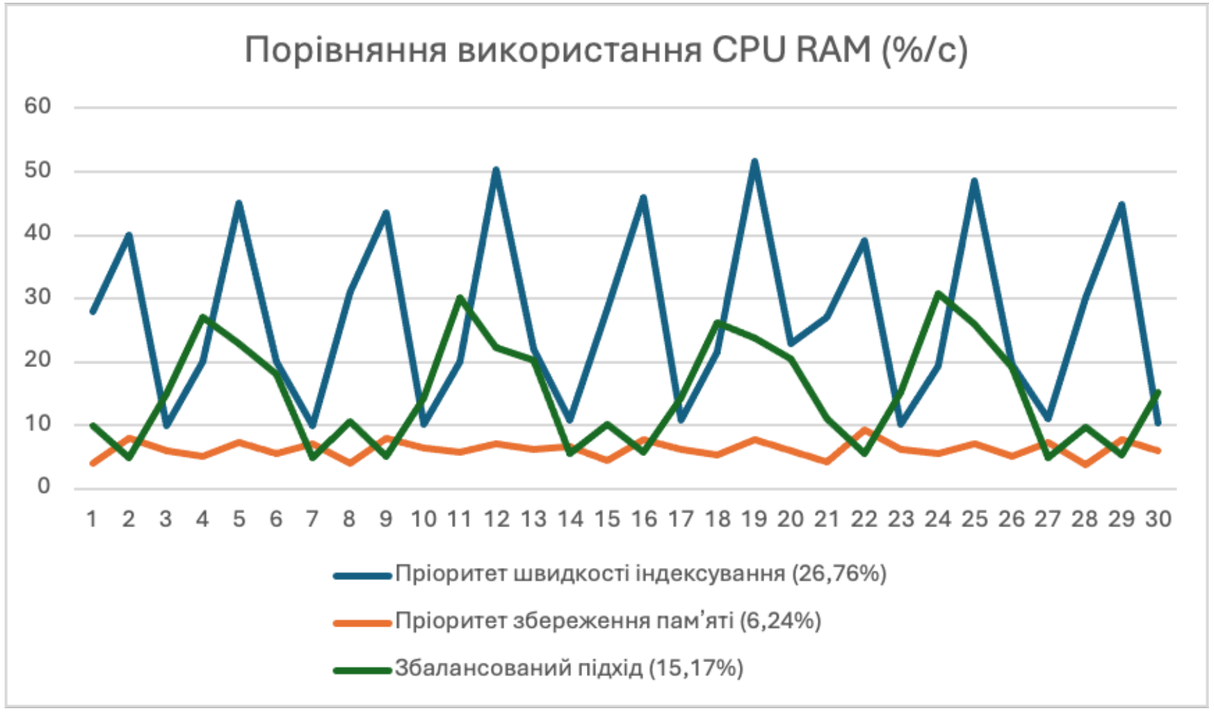
КОНФІГУРАЦІЇ КЛІЄНТУ, РОЗГОРТАННЯ ТА МЕТОДИ

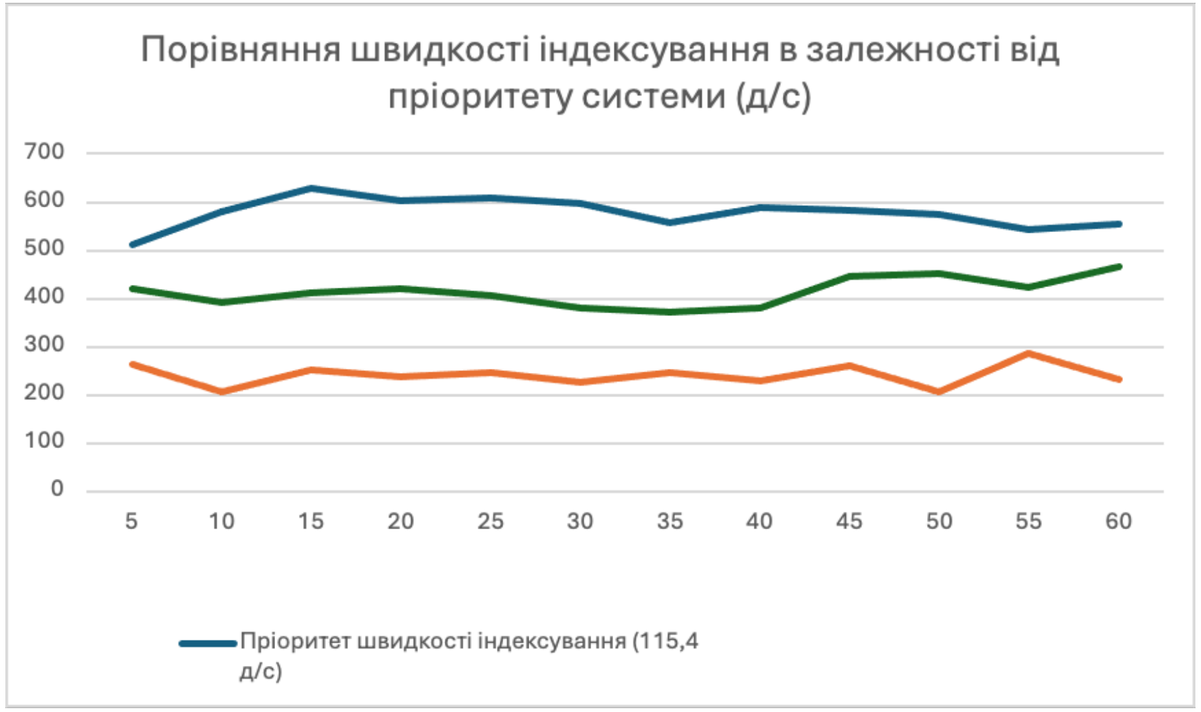
ШЛЯХИ ОПТИМІЗАЦІЇ	ELASTICSEARCH	QDRANT	WEAVIATE
Масові операції	Bulk API	upsert, upload_points, upload_collection	dynamic batch upload
Vector Quantization	-	Scalar, Binary, Product	Product, Binary
Конфігурація	репліки, індекс за полем, dynamic template	оптимізатори	тип індексу, репліки
Розгортання	виділення пам’яті	memmap, сегменти	memmap, асинхронне індексування

ElasticSearch

Інтегрований з Lucene двигун для повнотекстового пошуку

Метрики продуктивності

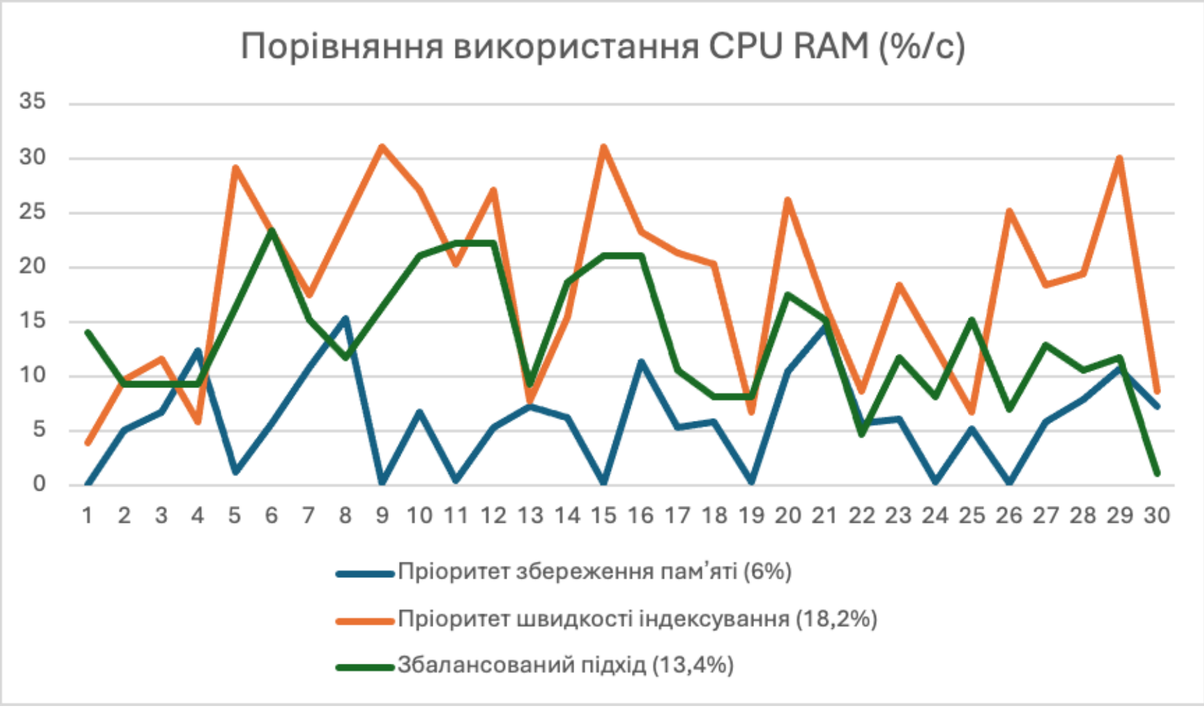


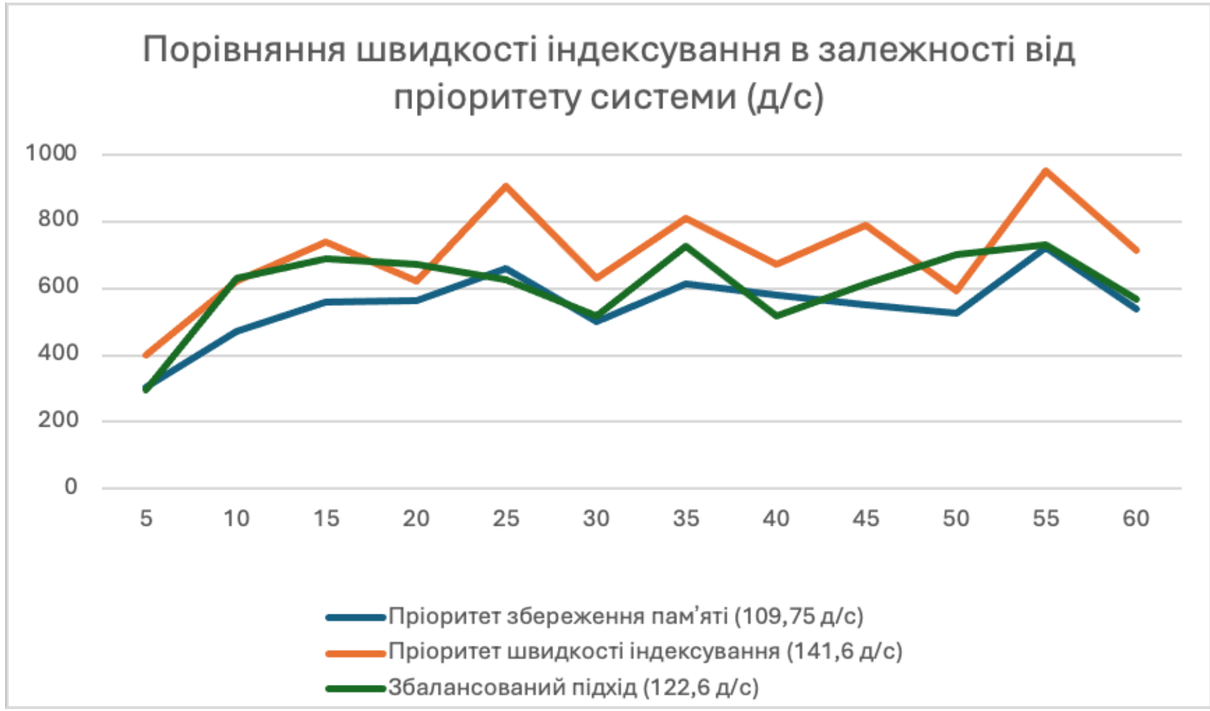


Qdrant

ВЕКТОРНА БД З ПОШУКОВОЮ СИСТЕМОЮ СХОЖОСТІ ВЕКТОРІВ

Метрики продуктивності

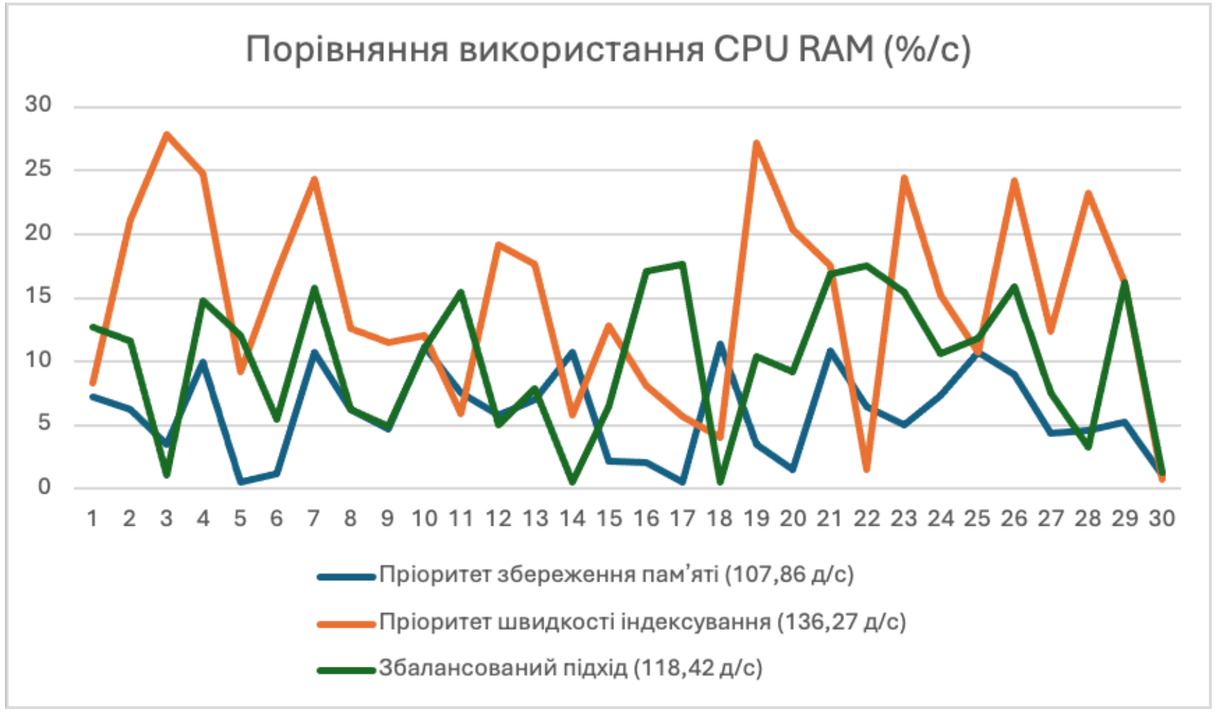


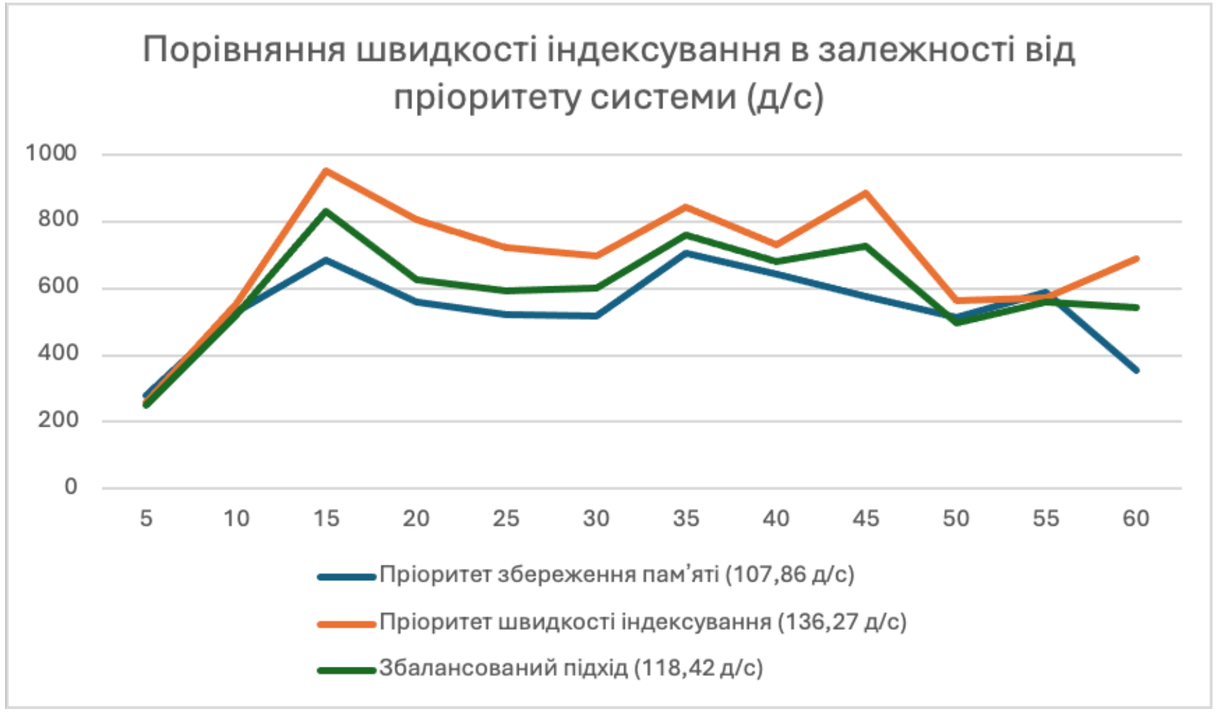


Weaviate

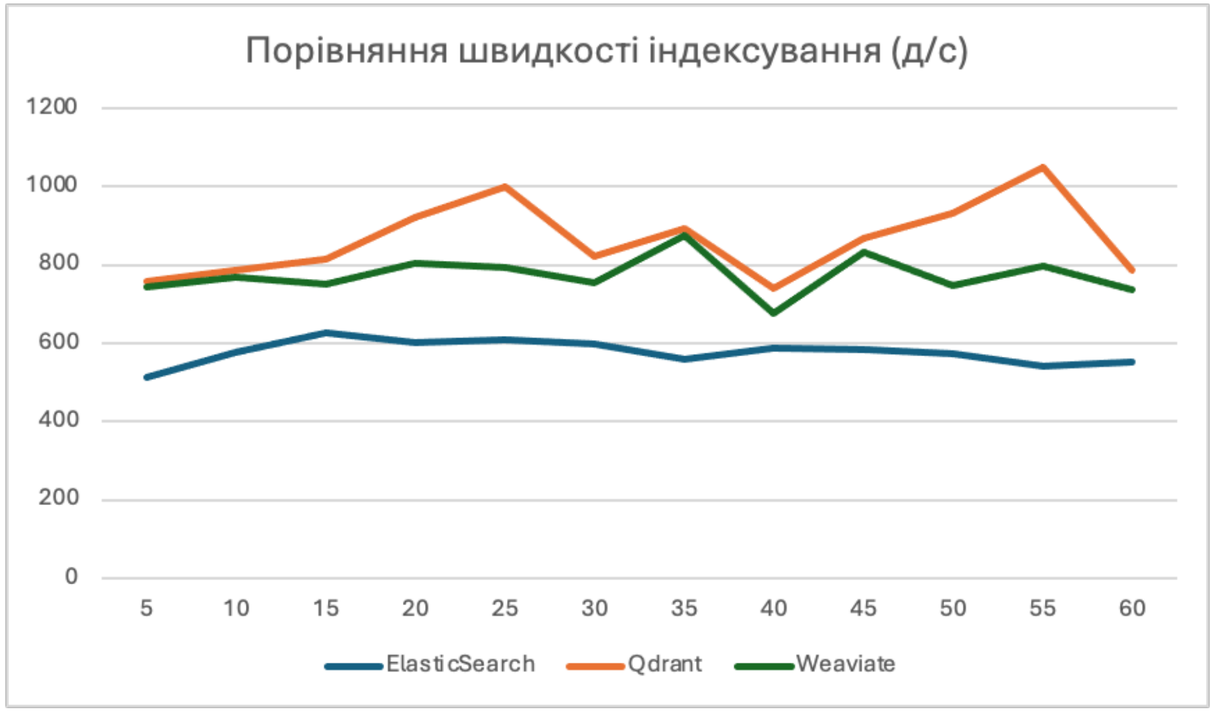
ВЕКТОРНА БД, ЯКА ЗБЕРІГАЄ ЯК ОБ’ЄКТИ, ТАК І ВЕКТОРИ

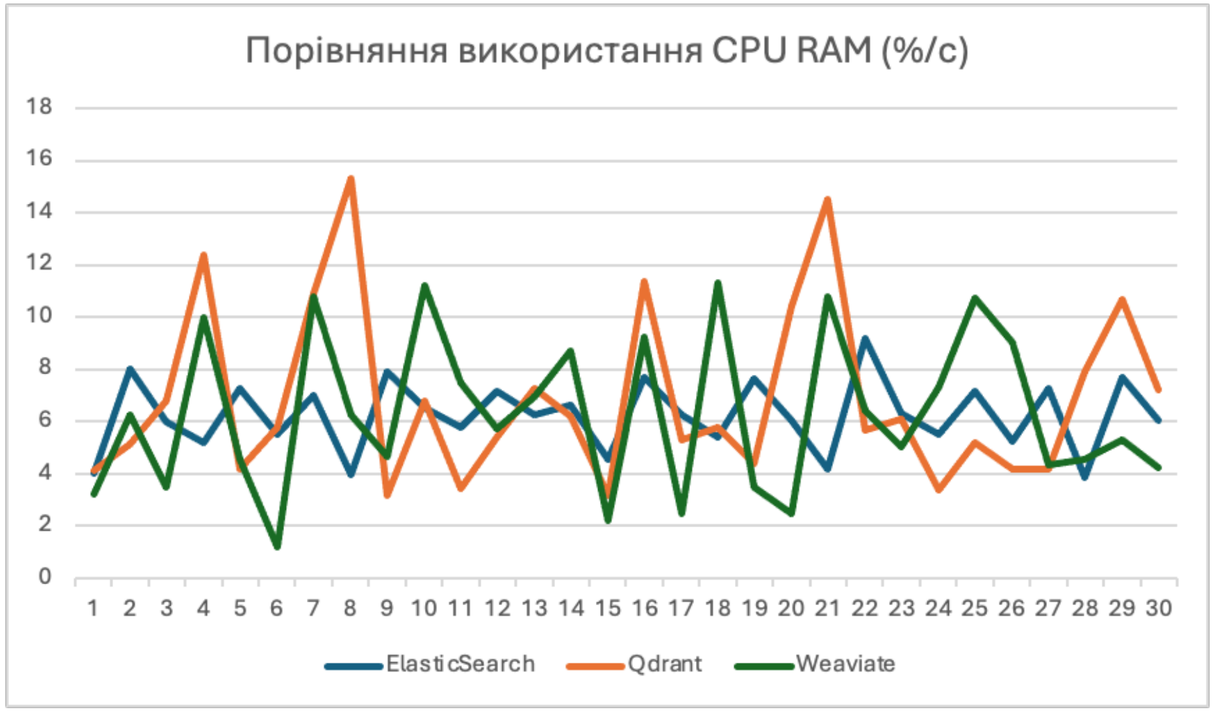
Метрики продуктивності





Аналіз результатів





ВИСНОВКИ

Presenting live not your thing? No worries! Record your Canva Presentation your audience can watch at their own pace.

Don't forget to delete or hide this page before presenting.

Попередня обробка даних за допомогою потужностей AWS
Створено алгоритм очищення даних, що включає в себе декілька етапів обробки
Sentence-BERT перетворення даних у векторне представлення
Розподілені системи ElasticSearch, Qdrant та Weaviate
Методи оптимізації систем
Аналіз результатів дослідження