Elasticsearch як ядро пошукової системи

Виконав студент прикладної математики - 3

Федусов С.В.

Науковий керівник

Доктор технічних наук, доцент Глибовець А.М.

Вступ

- Сучасні пошукові системи мають наступні базові функції: швидкий повнотекстовий пошук, фільтрацію, сортування і ранжування документів, нечіткий пошук, швидке збереження та індексування структурованих даних.
- Необхідність отримувати результати пошукових запитів за мілісекунди, маючи десятки або сотні гігабайт інформації вимагає використання ефективних алгоритмів та структур даних. Окрім цього, сама система повинна гарантувати цілісність та відмовостійкість. Надає необхідний функціонал та задовольняє наведені вимоги пошуковий двигун Elasticsearch.

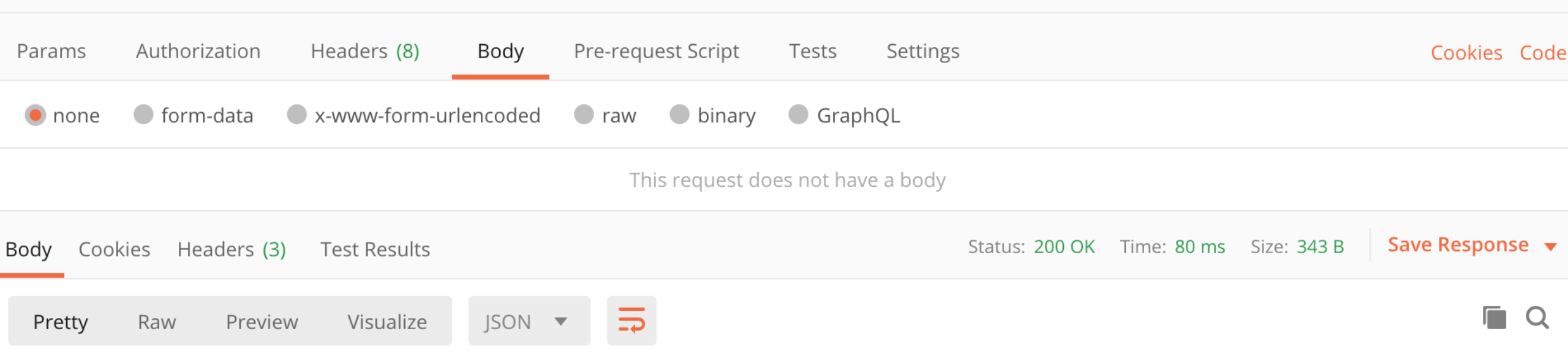
Постановка задачі

- Ознайомитися з базовим функціоналом пошукового двигуна Elasticsearch
- Дослідити алгоритми та структури даних, що використовуються для індексації
- Розглянути архітектуру пошукового двигуна
- Застосувати Elasticsearch для індексації та аналізу тестових даних

Тестові дані

- Для дослідження можливостей індексації, пошуку та аналізу даних була використана глобальна статистика поширення вірусу COVID-19

Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20
	Afghanistan	33.0	65.0	0	0	0	0	0
	Albania	41.1533	20.1683	0	0	0	0	0
	Algeria	28.0339	1.6596	0	0	0	0	0
	Andorra	42.5063	1.5218	0	0	0	0	0
	Angola	-11.2027	17.8739	0	0	0	0	0
	Antigua and Barbuda	17.0608	-61.7964	0	0	0	0	0
	Argentina	-38.4161	-63.6167	0	0	0	0	0
	Armenia	40.0691	45.0382	0	0	0	0	0
Australian Capital Territory	Australia	-35.4735	149.0124	0	0	0	0	0
New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3
Northern Territory	Australia	-12.4634	130.8456	0	0	0	0	0
Queensland	Australia	-28.0167	153.4	0	0	0	0	0
South Australia	Australia	-34.9285	138.6007	0	0	0	0	0
Tasmania	Australia	-41.4545	145.9707	0	0	0	0	0
Victoria	Australia	-37.8136	144.9631	0	0	0	0	1
Western Australia	Australia	-31.9505	115.8605	0	0	0	0	0
	Austria	47.5162	14.5501	0	0	0	0	0
	Azerbaijan	40.1431	47.5769	0	0	0	0	0
	Bahamas	25.0343	-77.3963	0	0	0	0	0
	Bahrain	26.0275	50.55	0	0	0	0	0
	Bangladesh	23.685	90.3563	0	0	0	0	0
	Barbados	13.1939	-59.5432	0	0	0	0	0



Робота з Elasticsearch

Взаємодія з кластером Elasticsearch відбувається за допомогою http запитів.

POST http://localhost:9200/covid_19/_search Send Save

Params Authorization Headers (10) Body Pre-request Script Tests Settings Cookies Code

none form-data x-www-form-urlencoded raw binary GraphQL JSON Beautify

```
{
    "query": {
        "match_all": {}
    },
    "from": 0,
    "size": 10
}
```

POST http://localhost:9200/covid_19/_doc/ Send Save

Params Authorization Headers (10) Body Pre-request Script Tests Settings Cookies Code

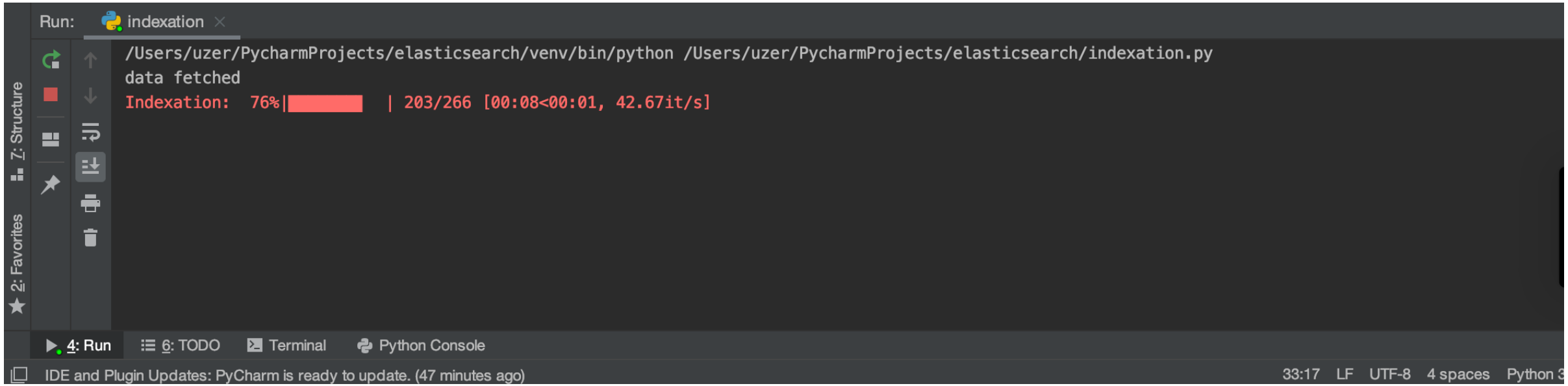
none form-data x-www-form-urlencoded raw binary GraphQL JSON Beautify


```
{
    "country": "Ukraine",
    "location": {
        "lat": 48.3794,
        "lon": 31.1656
    },
    "daily": [
        {
            "date": "2020-01-22",
            "cases": 0
        },
        {
            "date": "2020-01-23",
            "cases": 0
        },
        {
            "date": "2020-01-24",
            "cases": 0
        },
```

Демо-застосунок для індексації

Для індексації тестових даних було використано Python, Pandas та Elasticsearch Python API.

Для моніторингу процесу індексації було використано Python бібліотеку tqdm, що відображає індикатор виконання індексації, а також базову інформацію: час виконання, швидкість.

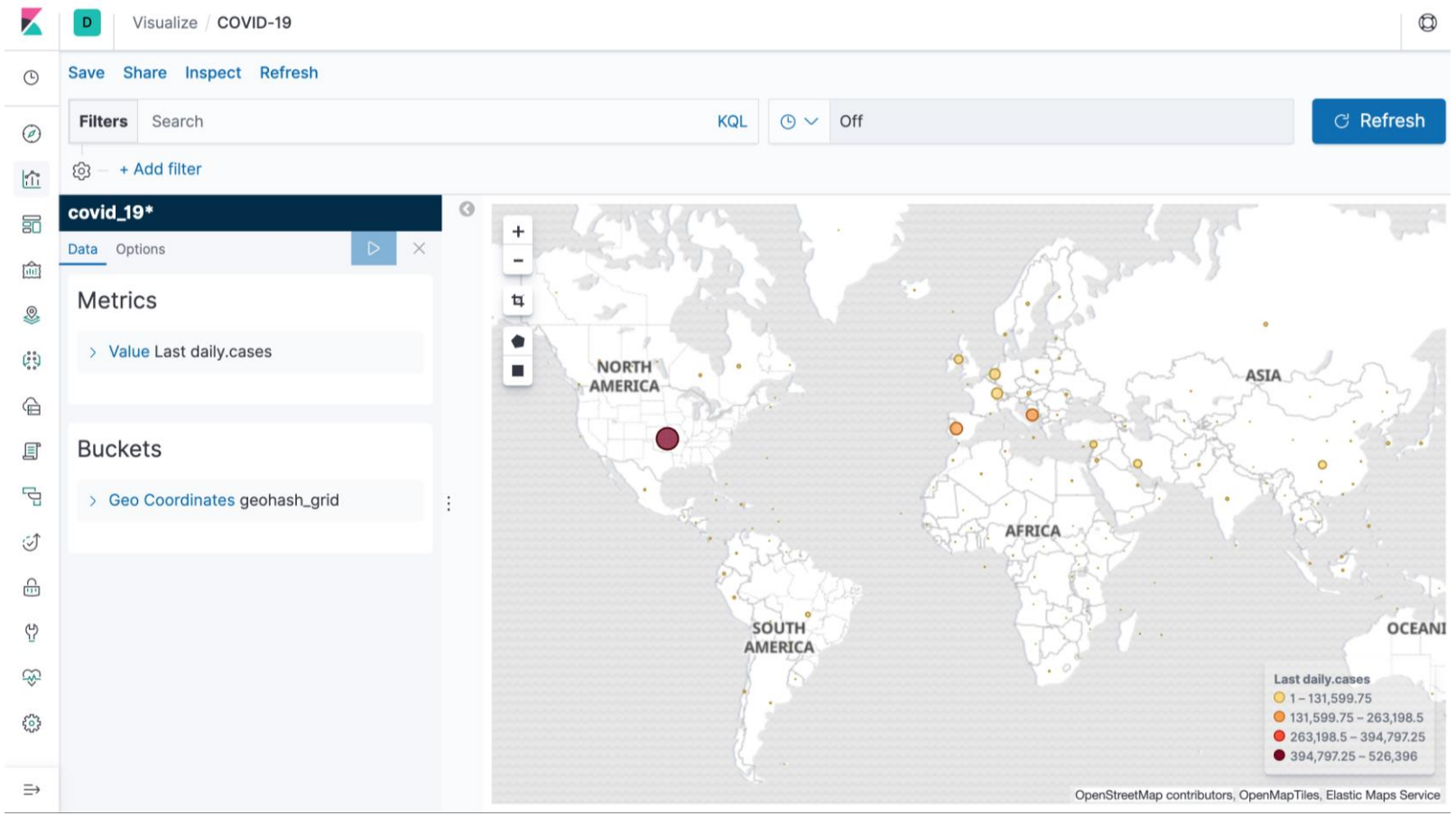


Розподілена архітектура

Індекс було створено з налаштуванням п'яти шард і двох реплік. Таким чином, у кластері з двох машин кожна з них містить усі шарди індексу.

Metrics | Shards | Aliases | Mappings | Administration

Shard	State	# Docs	Size	Primary	Node
0	STARTED	52	107.6kb	true	es01
0	STARTED	52	107.6kb	false	es02
1	STARTED	36	81.2kb	false	es01
1	STARTED	36	81.3kb	true	es02
2	STARTED	55	115.2kb	true	es01
2	STARTED	55	110.4kb	false	es02



Використовуючи дані розташування країн за допомогою Kibana було побудовано карту поширення вірусу. Кожна країна має відмітку у вигляді кола, що іллюструє кількість випадків захворювання

Висновки

- Схема індексу дозволяє описати будь-яку структуру, а мова запитів задовольняє будь-яку інформаційну потребу.
- Завдяки розподіленій архітектурі нівелюються фізичні обмеження однієї обчислювальної машини та забезпечується більша надійність та відмовостійкість системи.
- Elasticsearch є гарним рішенням для побудови пошукових систем

Дякую за увагу