



**МІНІСТЕРСТВО ОСВІТИ, НАУКИ, МОЛОДІ ТА СПОРТУ УКРАЇНИ  
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
«КІЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»  
ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ**

**Лабораторна робота №3**

**Аналіз даних**

**Підготував:**

студент 4 курсу  
групи ФІ-84

Коломієць Андрій Юрійович

**E-mail:** andrew.kolomiets.work@gmail.com

**Київ – 2021**

## Лабораторна робота №3

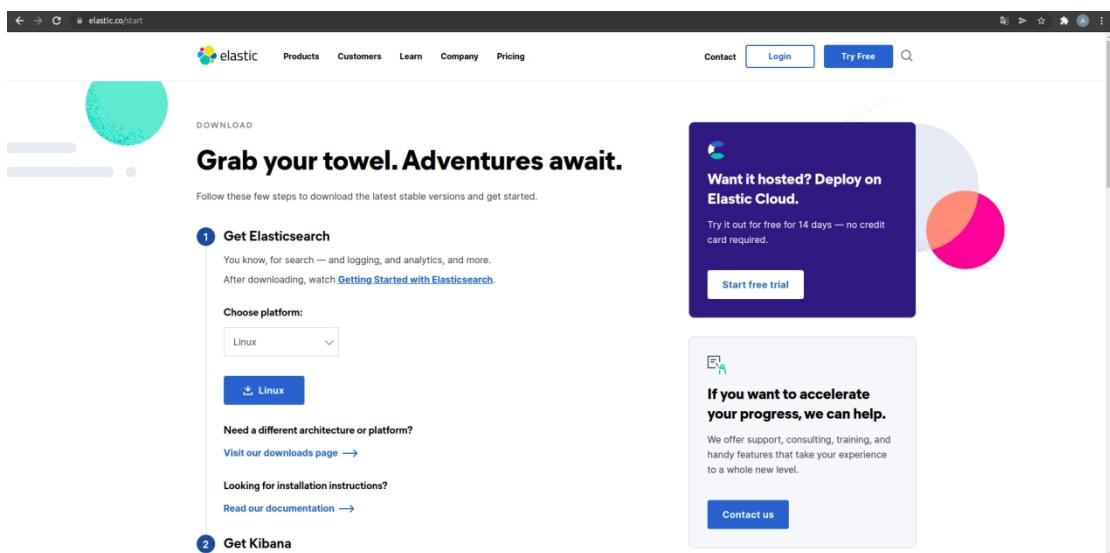
### Аналіз даних

#### Завдання на самостійну роботу

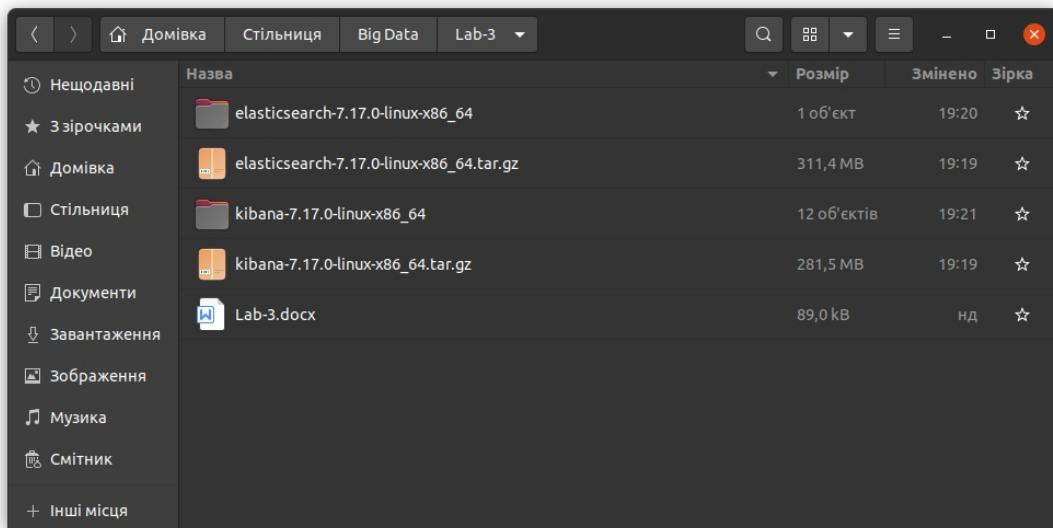
1. Написати програму формування пакетного файлу для завантаження в **Elasticsearch** на основі сформованого на попередньому занятті файлу в форматі **JSON**.
2. Встановити на комп'ютері утиліту **curl** (<http://curl.se>) і вивчити склад і значення її основних параметрів.
3. Ознайомитися зі складом і змістом **CRUD-операцій** в **Elasticsearch**.

#### Виконання завдання

Переходимо на сайт скачування продукту та завантажуємо **Elasticsearch** та **Kibana**:



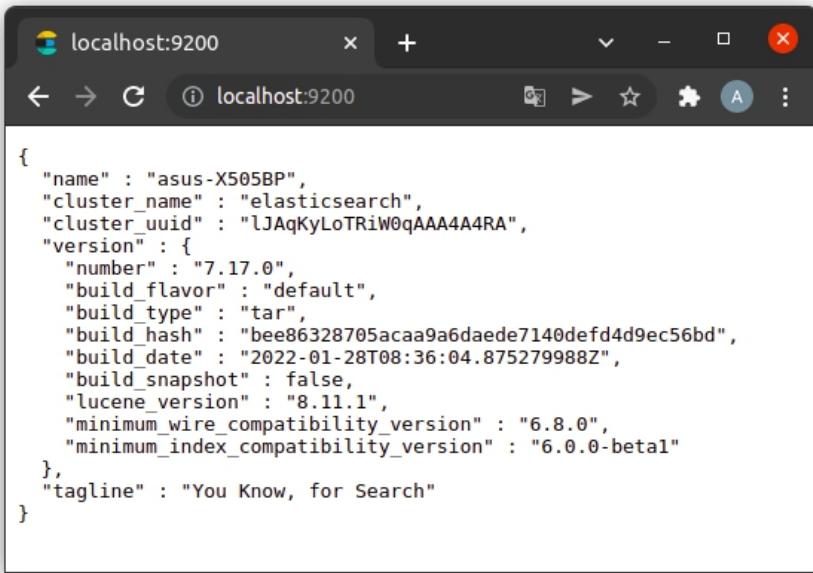
В результаті буде завантажено два архіви, впевнимося в цьому.



Запускаємо відповідно до інструкцій на сайті два вебінтерфейси:

Перевіряємо коректність запуску програми:

-використання браузера:



A screenshot of a web browser window titled "localhost:9200". The address bar also shows "localhost:9200". The page content is a JSON object representing the Elasticsearch version. The JSON is as follows:

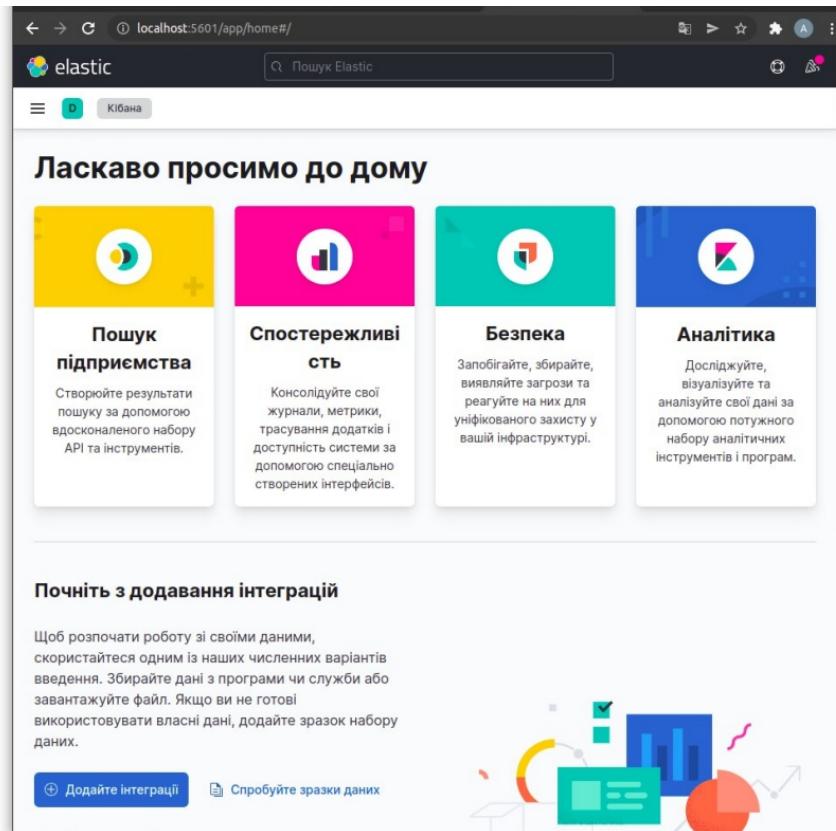
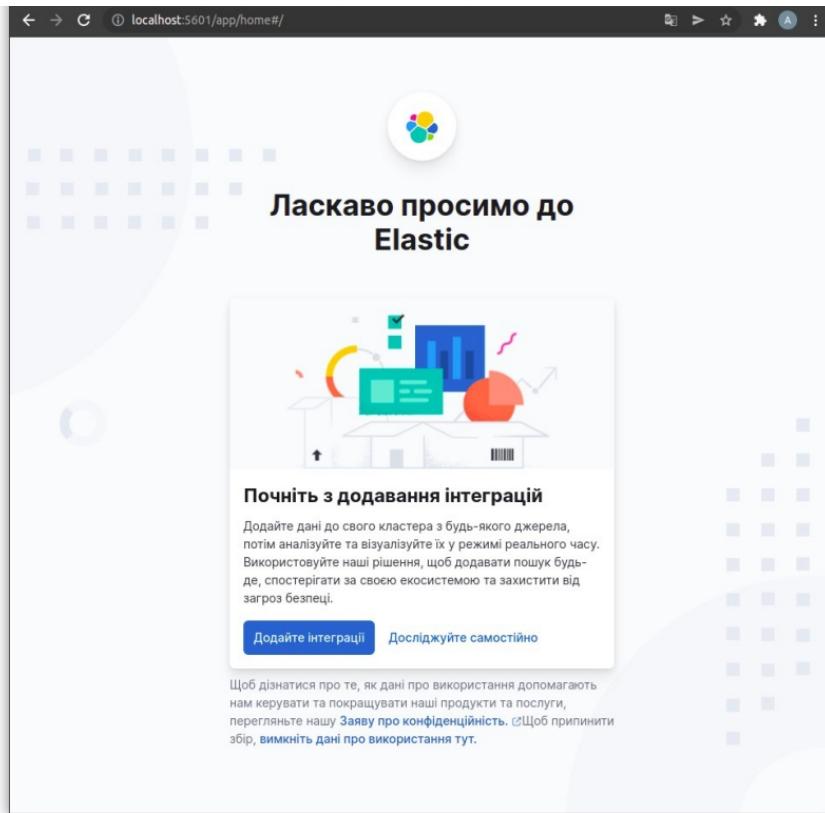
```
{  
  "name" : "asus-X505BP",  
  "cluster_name" : "elasticsearch",  
  "cluster_uuid" : "lJAqKyLoTRiW0qAAA4A4RA",  
  "version" : {  
    "number" : "7.17.0",  
    "build_flavor" : "default",  
    "build_type" : "tar",  
    "build_hash" : "bee86328705acaa9a6daede7140defd4d9ec56bd",  
    "build_date" : "2022-01-28T08:36:04.875279988Z",  
    "build_snapshot" : false,  
    "lucene_version" : "8.11.1",  
    "minimum_wire_compatibility_version" : "6.8.0",  
    "minimum_index_compatibility_version" : "6.0.0-beta1"  
  },  
  "tagline" : "You Know, for Search"  
}
```

-використання терміналу:

```
curl -X GET "localhost:9200/?pretty"
```

```
{  
  "name" : "asus-X505BP",  
  "cluster_name" : "elasticsearch",  
  "cluster_uuid" : "lJAqKyLoTRiW0qAAA4A4RA",  
  "version" : {  
    "number" : "7.17.0",  
    "build_flavor" : "default",  
    "build_type" : "tar",  
    "build_hash" : "bee86328705acaa9a6daede7140defd4d9ec56bd",  
    "build_date" : "2022-01-28T08:36:04.875279988Z",  
    "build_snapshot" : false,  
    "lucene_version" : "8.11.1",  
    "minimum_wire_compatibility_version" : "6.8.0",  
    "minimum_index_compatibility_version" : "6.0.0-beta1"  
  },  
  "tagline" : "You Know, for Search"  
}
```

Запустимо **Kibana**:



Можна проексперементувати та завантажити зразок JSON та поглянути на результати:

## More ways to add data

In addition to adding [integrations](#), you can try our sample data or upload your own data.

[Sample data](#) [Upload file](#)

**result.json**

### File contents

First 998 lines

```
1 {  
2   "title": "Covid-19: Major IT companies to extend WFH till March 2021",  
3   "textBody": "&lt;a href=\"https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoeshow/79640742.cms\"&gt;&lt;img  
4     width='100' height='75' border='0' hspace='10' align='left' src=\"https://img.etimg.com/photo/79640742.cms\" /&gt;&lt;/a&gt;",  
5   "source": "ETAG launched on Play Store worldwide",  
6   "PubDate": "2022-02-08T12:11:00Z",  
7   "URL": "https://economictimes.indiatimes.com/tech/software/covid-19-major-it-companies-to-extend-wfh-till-march-2021/videoeshow/79640742.cms"  
8 }  
9 {  
10  "title": "Persistent to acquire US-based Capit Software for $6.34 million",  
11  "textBody": "&lt;a href=\"https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-capit-software-for-6-34-million/articleshow/78681554.cms\"&gt;&lt;  
12    img width='100' height='75' border='0' hspace='10' align='left' src=\"https://img.etimg.com/photo/78681554.cms\" /&gt;&lt;/a&gt;Capit, founded in 2014, specializes in  
13    enterprise integration with expertise in MuleSoft, Red Hat and TIBCO. It also delivers enterprise modernization, with advanced proficiency in key partner platforms,  
14    frameworks and industry data models.",  
15  "source": "ETAG Launched on Play Store worldwide",  
16  "PubDate": "2022-02-08T12:11:00Z",  
17  "URL": "https://economictimes.indiatimes.com/tech/software/persistent-to-acquire-us-based-capit-software-for-6-34-million/articleshow/78681554.cms"
```

### Summary

**Number of lines analyzed**

998

**Format**

semi\_structured\_text

**Grok pattern**

```
%{QUOTEDSTRING:field}:"%{TIMESTAMP_ISO8601:timestamp}","%{QUOTEDSTRING:field2}:"%{URI:uri}":"%{QUOTEDSTRING:field3}:%{QUOTEDSTRING:field4},%{QUOTEDSTRING:field5}:%{QUOTEDSTRING:field6}.%{QUOTEDSTRING:field7}.%{QUOTEDSTRING:field8}.*%{QUOTEDSTRING:field8}.*"
```

**Time field**

timestamp

**Time format**

ISO8601

[Override settings](#)

[Analysis explanation](#)

### File stats

All fields 11 of 11 total

Number fields 0 of 0 total

Field name 11 ▾

Field type 3 ▾

Type	Name	Documents (%)	Distinct values	Distributions
>	field	142 (100%)	1	1 category
>	field2	142 (100%)	1	1 category
>	field3	142 (100%)	1	1 category
>	field4	142 (100%)	142	top 10 of 142 categories
>	field5	142 (100%)	1	1 category
>	field6	142 (100%)	97	top 10 of 97 categories
>	field7	142 (100%)	5	5 categories
>	field8	142 (100%)	3	3 categories
>	message	142 (100%)	142	top 10 of 142 categories
>	timestamp	142 (100%)	1	1 category
>	uri	142 (100%)	142	top 10 of 142 categories

Rows per page: 25 ▾

< 1 >

Маємо котрийсь аналіз але поки невідомо, спробуємо використати термінал.

## Виконання з використанням терміналу Linux

### Створюємо скрипт

**Зauważення:** команди не вводилися вручну! Використовуючи **Visual Studio Code** можна замінити кожне входження певного символу на команду відповідну. В данній лабораторній роботі для створення скрипту було здійснено заміну симовола:

```
{
```

на

```
curl -X POST http://localhost:9200/test/_doc/ -H Content-Type:application/json -d '  
{
```

та

```
,
```

на

```
'}
```

Додамо нове розширення до файлу `#!/bin/bash` та зробимо файл виконуваним `$chmod a+x script.sh`.

Здійснимо таку процедуру відповідно кожне входження легко заміняється:

```
#!/bin/bash

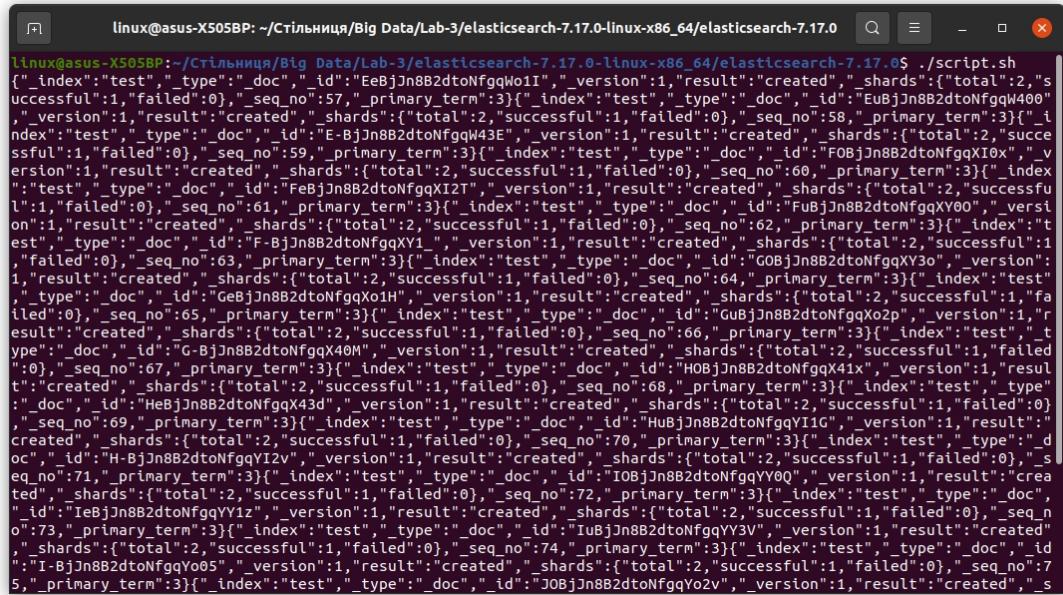
curl -XPOST 'http://localhost:9200/test/_doc/' -H 'Content-Type: application/json' -d '
{
  "title": "НБУ не втримав долар у межах 29 гривень",
  "textBody": "Торги на міжбанківському валютному ринку станом на 12:40 відбувалися на рівні 28 гривень 28–32 копійки за долар",
  "source": "НБУ не втримав долар у межах 29 гривень",
  "PubDate": "2022-02-23T13:06:00Z",
  "URL": " https://www.radiosvoboda.org/a/news-hryvnya-dolar-kurs-nbu/31718337.html "
}

curl -XPOST 'http://localhost:9200/test/_doc/' -H 'Content-Type: application/json' -d '
{
  "title": "Засідання РНБО лишається відкритим, ніхто Київ не покидає – Данілов",
  "textBody": "У разі необхідності РНБО збиратиметься у дуже короткий проміжок часу",
  "source": "НБУ не втримав долар у межах 29 гривень",
  "PubDate": "2022-02-23T13:06:00Z",
  "URL": " https://www.radiosvoboda.org/a/news-rnbo-zasidannia-danilov/31718332.html "
}

curl -XPOST 'http://localhost:9200/test/_doc/' -H 'Content-Type: application/json' -d '
{
  "title": "Визнання Росією «ДНР». Які це несе загрози для прифронтового Маріуполя",
  "textBody": "Маріуполі сподіваються, що рішення Путіна спричинить швидші та жорсткіші санкції проти нього",
  "source": "НБУ не втримав долар у межах 29 гривень",
  "PubDate": "2022-02-23T13:06:00Z",
  "URL": " https://www.radiosvoboda.org/a/novyny-pryazovya-rosiya-dnr-mariupol/31718220.html "
}

curl -XPOST 'http://localhost:9200/test/_doc/' -H 'Content-Type: application/json' -d '
{
  "title": "В більшості областей України запровадять надзвичайний стан – секретар РНБО",
  "textBody": "Заходи можуть передбачати посилення охорони громадського порядку, обмеження певного руху транспорту, перевірку документів у громадян",
  "source": "НБУ не втримав долар у межах 29 гривень",
  "PubDate": "2022-02-23T13:06:00Z",
  "URL": " https://www.radiosvoboda.org/a/news-nadzvychainyi-stan-rnbo/31718296.html "
}
```

Використовуємо `./script.sh` для запуску файлу та отримуємо:



```
linux@asus-X505BP:~/Стільниця/Big Data/Lab-3/elasticsearch-7.17.0-linux-x86_64/elasticsearch-7.17.0$ ./script.sh
{"_index": "test", "_type": "doc", "_id": "EebjJn8B2dtoNfgqW01", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 57, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "EebjJn8B2dtoNfgqW01", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 58, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "E-BjJn8B2dtoNfgqW43E", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 59, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "F0BjJn8B2dtoNfgqXIoX", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 60, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "FeBjJn8B2dtoNfgqX12T", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 61, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "FuBjJn8B2dtoNfgqXY00", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 62, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "F-BjJn8B2dtoNfgqXY1", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 63, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "G0BjJn8B2dtoNfgqXY3o", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 64, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "G-BjJn8B2dtoNfgqXoI", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 65, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "GuBjJn8B2dtoNfgqXo2p", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 66, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "H-BjJn8B2dtoNfgqX40M", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 67, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "H-BjJn8B2dtoNfgqYI2v", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 68, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "HeBjJn8B2dtoNfgqX43d", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 69, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "HuBjJn8B2dtoNfgqYIIG", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 70, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "I-BjJn8B2dtoNfgqYo5", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 71, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "I0BjJn8B2dtoNfgqYY00", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 72, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "IeBjJn8B2dtoNfgqYY1z", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 73, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "IuBjJn8B2dtoNfgqYY3V", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 74, "primary_term": 3} {"_index": "test", "_type": "doc", "_id": "I-BjJn8B2dtoNfgqYo5", "_version": 1, "result": "created", "shards": {"total": 2, "successful": 1, "failed": 0}, "seq_no": 75, "primary_term": 3}
```

## Перевірка результатів завантаження даних

```
linux@asus-X505BP:~/Стільниця/Big Data/Lab-3/elasticsearch-7.17.0-linux-x86_64/elasticsearch-7.17.0$ curl -XGET 'http://localhost:9200/test/_doc/_count'
```

```
{"count": 815, "shards": {"total": 1, "successful": 1, "skipped": 0, "failed": 0}}
```

## Перевірка коректності створення індексу

```
linux@asus-X505BP:~/Стільниця/Big Data/Lab-3/elasticsearch-7.17.0-linux-x86_64/elasticsearch-7.17.0$ cat check_index.sh
#!/bin/bash

curl -XGET 'http://localhost:9200/test/_doc/_search?pretty=true' -H 'Content-Type: application/json' -d '
{
  "query" : {
    "match" : { "textBody": "сервіс" }
  }
}'

linux@asus-X505BP:~/Стільниця/Big Data/Lab-3/elasticsearch-7.17.0-linux-x86_64/elasticsearch-7.17.0$ ./check_index.sh
{
  "took" : 3,
  "timed_out" : false,
  "_shards" : {
    "total" : 1,
    "successful" : 1,
    "skipped" : 0,
    "failed" : 0
  },
  "hits" : {
    "total" : {
      "value" : 0,
      "relation" : "eq"
    },
    "max_score" : null,
    "hits" : [ ]
  }
}
```

## CRUD

*CRUD – англ. Create, Read, Update, Delete – 4 основні функції керування даними «створення, читання, оновлення і вилучення». У системі Elasticsearch операції CRUD націлені на документи*

*Для операцій CRUD реалізовано такі API:*

- Index API;
- Get API;
- Update API;
- Delete API.