Napredne arhitekture informacionih sistema

Elasticsearch

Izvođači nastave:
dr Marko Vještica
Elena Akik
Sanja Radić

DAIS
Data Science and Information Systems Group

Sadržaj

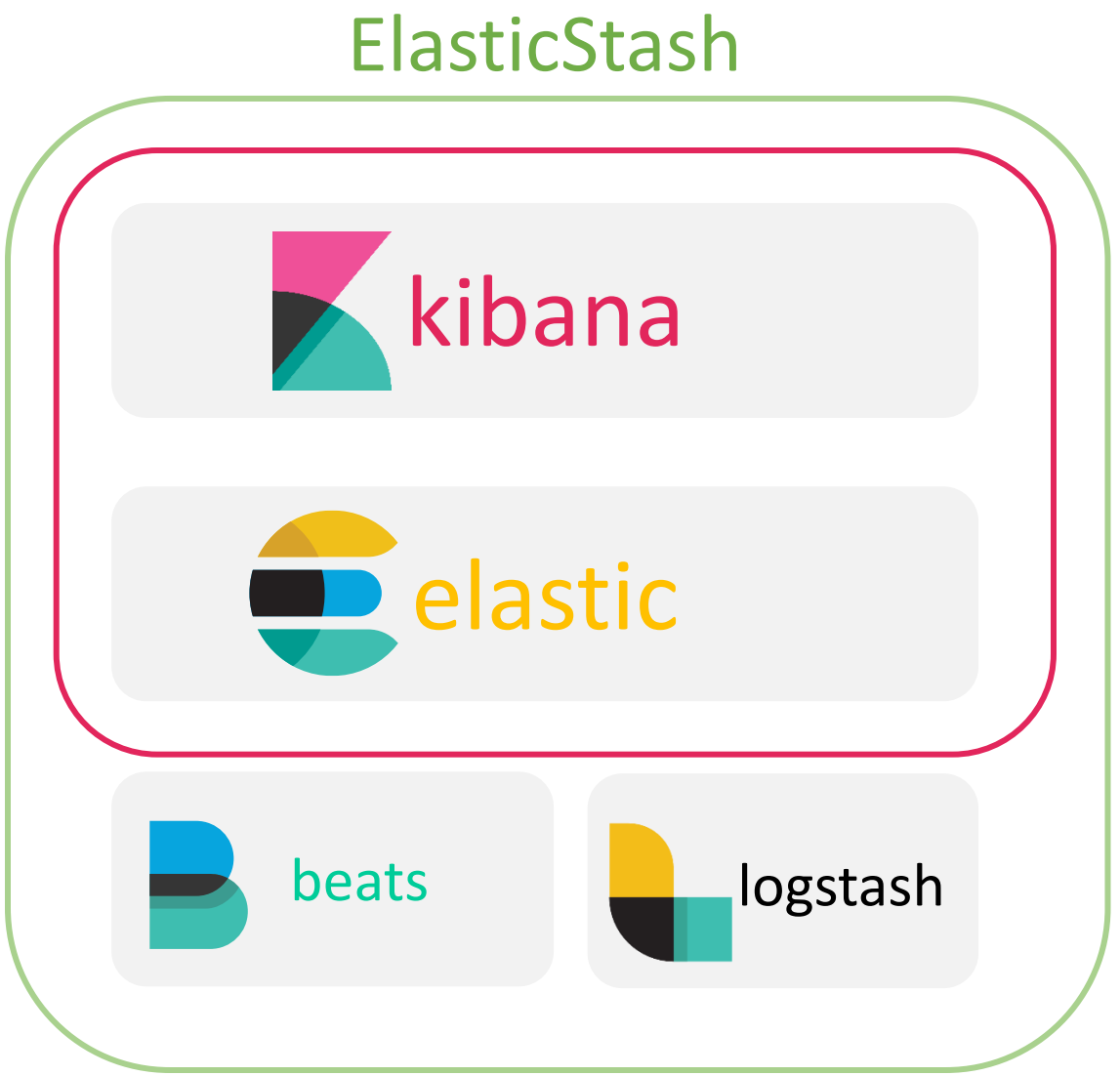
Softverska podrška

- **Lokalno** - *Elasticsearch with Kibana (Elasticsearch visualization tool)*
  - *Link: https://dev.to/elastic/downloading-elasticsearch-and-kibana-macos-linux-and-windows-1mmo*
- **Preko *Docker*-a** – pomoću pokretanja *docker-compose* fajla u kome se nalaze *Kibana* i *Elasticsearch*
  - *Kibani* je moguće pristupiti na adresi: *http://localhost:5601*
  - *Elasticserch*-u je moguće direktno pristupiti na adresi: *http://localhost:9200*

Sadržaj

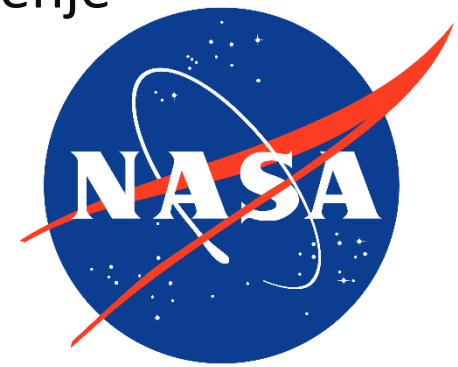
Šta je *ElasticStash*?

- ***ElasticStash*** je skup alata i tehnologija koji se koriste za prikupljanje, obradu, skladištenje i analizu velike količine podataka u realnom vremenu
- ***Elasticsearch*** je platforma za skladištenje, pretraživanje i analizu podataka u realnom vremenu
- ***Kibana*** je alat za vizualizaciju koji komunicira sa *Elasticsearch*-om, omogućavajući korisnicima da kreiraju upite, grafikone i tablice



Primena *Elasticsearch*-a

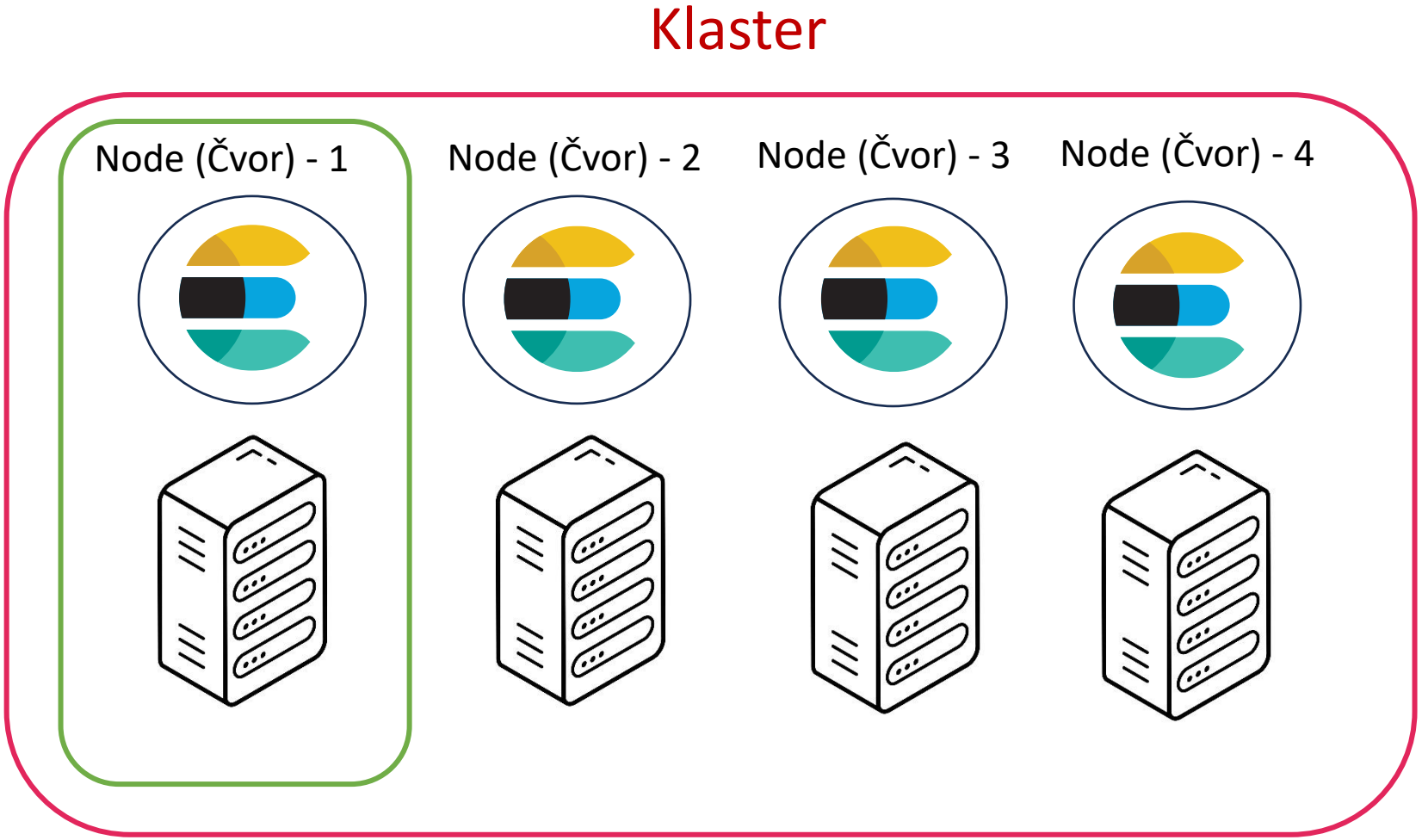
- **Evidencija događaja** – Prikupljanje i analiza zapisa događaja radi praćenja performansi i otkrivanja problema
- **Merenje performansi** – Prikupljanje i analiza metrika performansi sistema radi praćenja stanja i identifikacije potencijalnih problema
- **Analiza sigurnosti** – Analiza podataka radi otkrivanja i reagovanja na sigurnosne pretnje
- **Biznis analitika** – Analiza podataka radi donošenja informisanih poslovnih odluka i identifikacije prilika za unapređenje



Arhitektura *Elasticsearch*-a

Čvor je jedna instanca *Elasticsearch*-a i ima svoj jedinstveni naziv

Prilikom pokretanja *docker-compose* fajla, kreira se jedan čvor (instancu), a njen klaster se kreira automatski



Čuvanje podataka

- U okviru *Elasticsearch*-a podaci se čuvaju kao **dokumenti** (objekti JSON)
  - Svaki dokument ima svoja **polja** unutar kojih se čuvaju podaci
- Dokumenti koji su slični po šemi i semantici grupišu se u **indekse**, čime je moguće identifikovati određene podatke pristupom odgovarajućeg indeksa
- Ukoliko u okviru indeksa postoje raznorodni dokumenti, moguće je kreirati **tipove**, odnosno podgrupu dokumenata unutar indeksa

Indeks namirnice

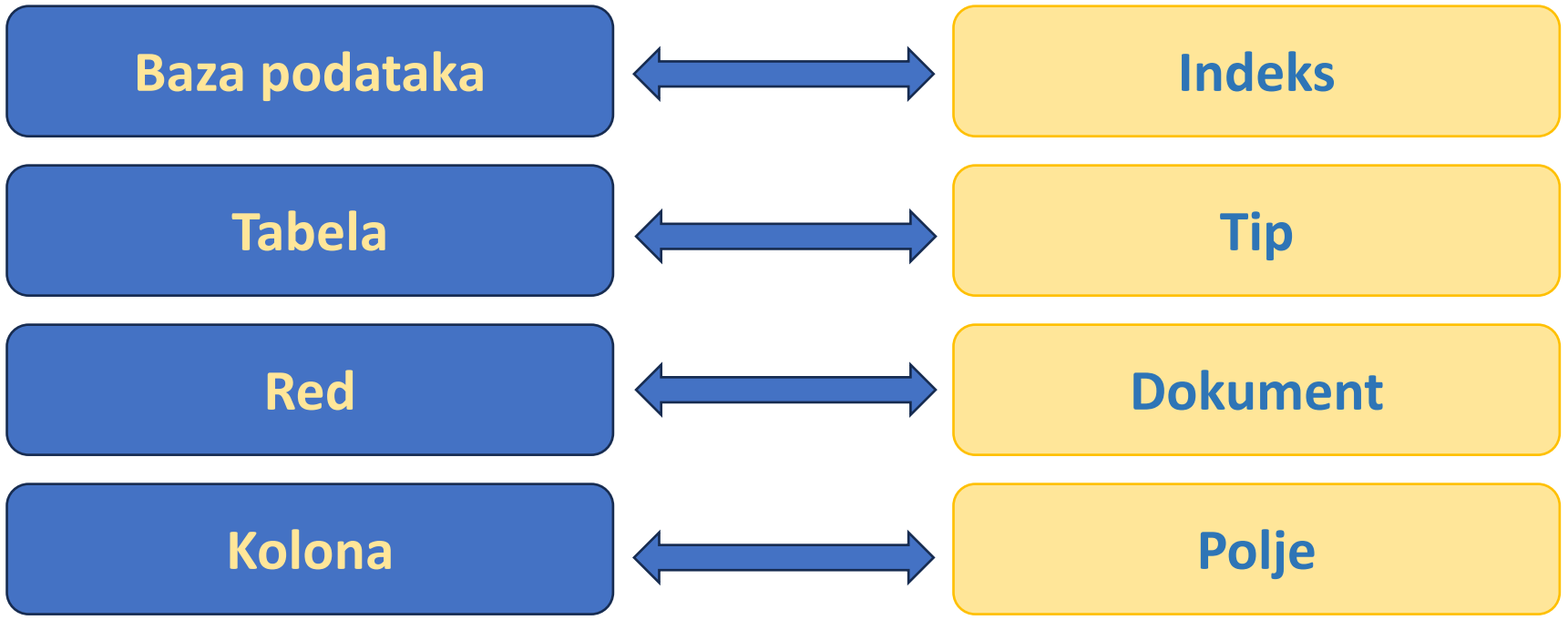
```
{
  naziv: "Šargarepa (1 kg)"
  kategorija: "povrće"
}
{
  naziv: "Kruška (1 kg)"
  kategorija: "voće"
}
```

Indeks piće

```
{
  naziv: "Sok od breskve (0,75 l)"
  kategorija: "negazirano piće“
}
{
  naziv: "Coca Cola (1,5 l)"
  kategorija: "gazirano piće"
}
```

Poređenje relacione baze i *Elasticsearch*-a

- Organizacija **relacionih** baza
- Organizacija ***Elasticsearch*-a**

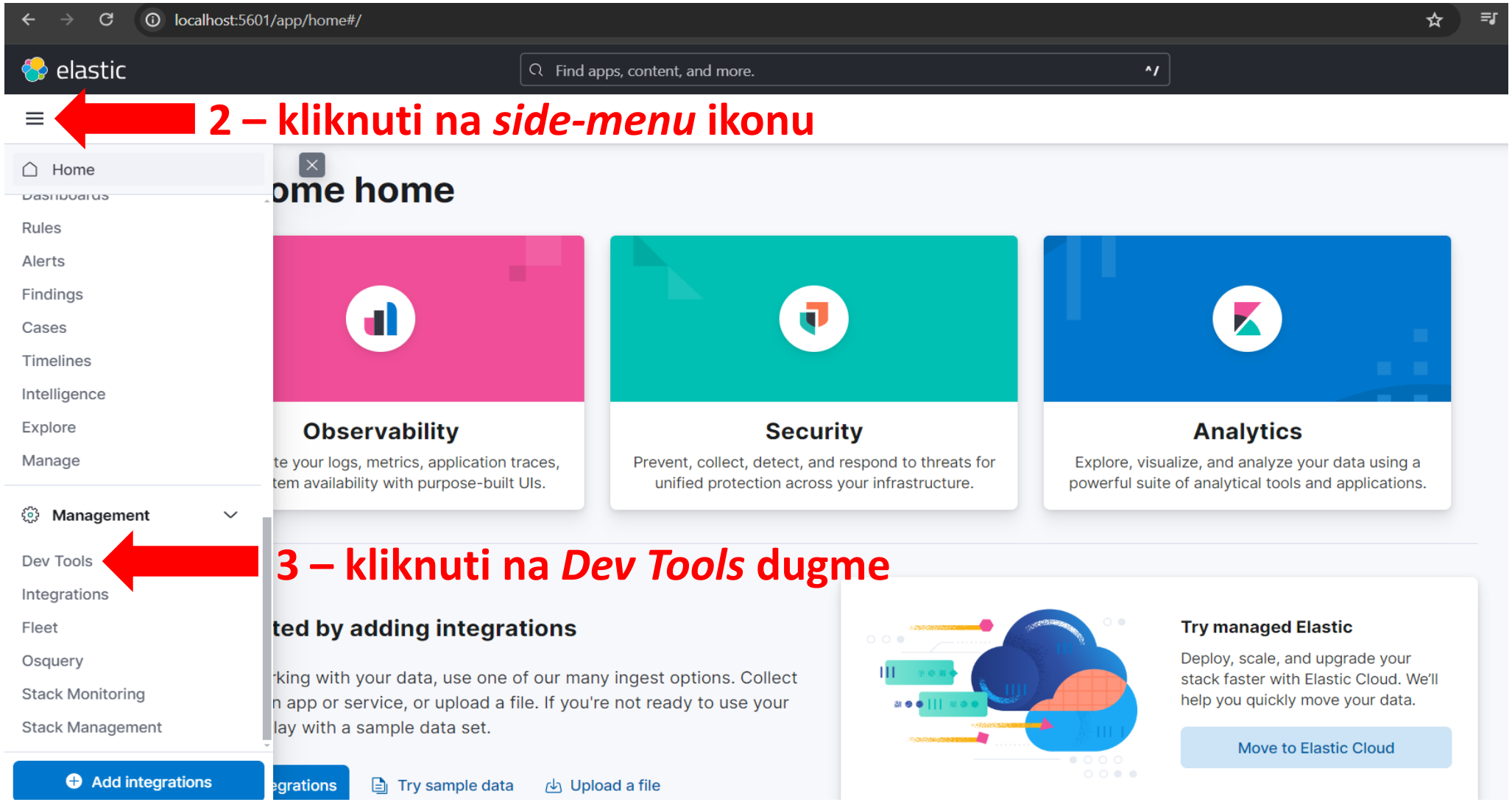


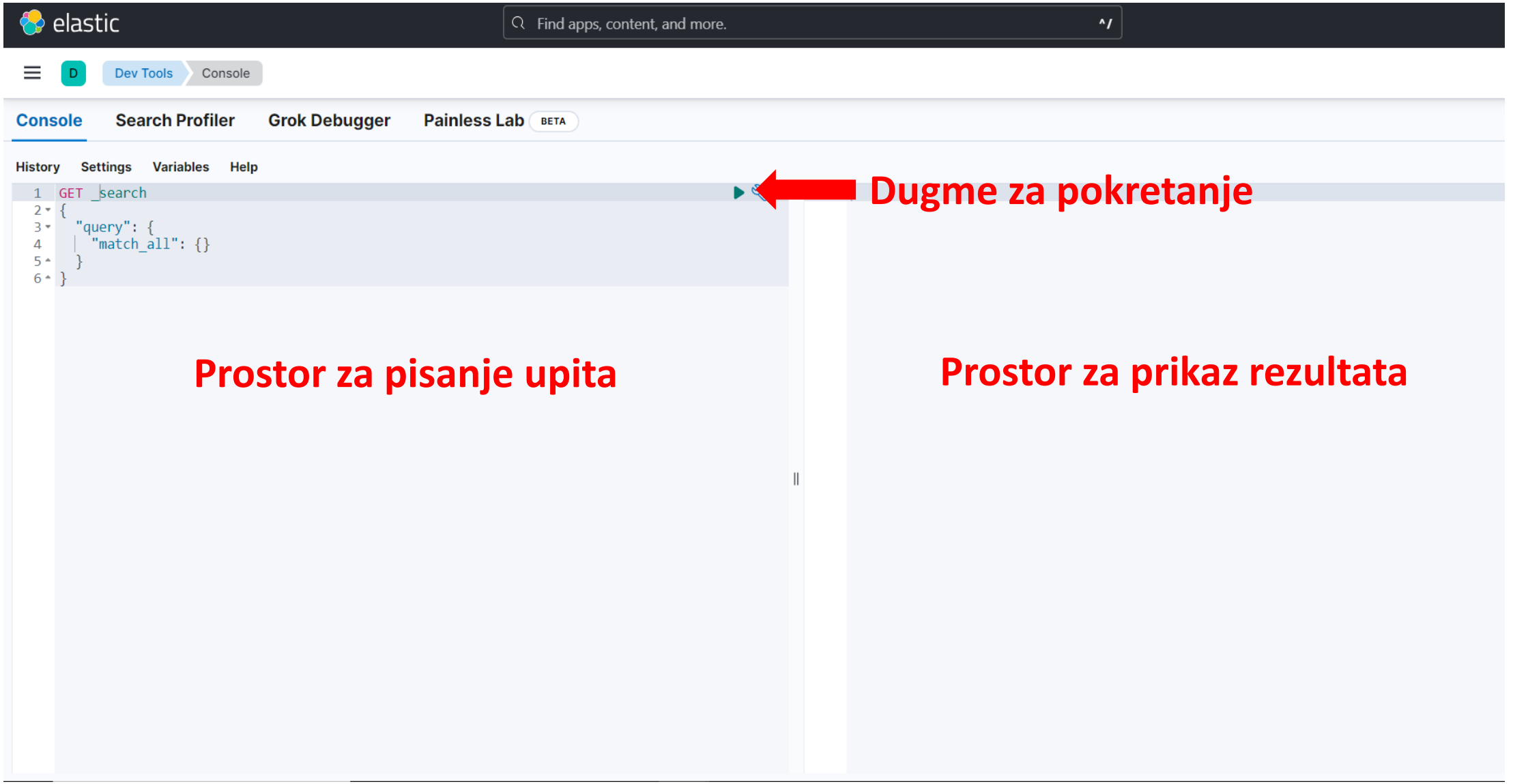
Poređenje relacione baze i *Elasticearch*-a

Relaciona baza podataka	*Elasticsearch*
Strogo definisana šema	Fleksibilna šema, omogućava nestrukturirane podatke
SQL bazirani upiti	JSON bazirani upiti, podržava kompleksne pretrage
Ograničena horizontalna skalabilnost	Horizontalna skalabilnost dodavanjem čvorova
Transakcioni sistemi, obezbeđenje ACID	Analitika, logovanje, pretraga teksta, monitoring
Skladište podataka	Pretraživač (engl. *search engine*) i skladište podataka

Sadržaj

1 – pokrenuti *Kibanu* u pretraživaču: *localhost:5601*




elastic
Find apps, content, and more.
^/
D
Dev Tools
Console
Console
Search Profiler
Grok Debugger
Painless Lab
BETA
History
Settings
Variables
Help
1 GET _search
2 {
3 "query": {
4 "match_all": {}
5 }
6 }
Dugme za pokretanje
Prostor za pisanje upita
Prostor za prikaz rezultata

Informacije o klasterima čvorovima

- Format pisanja upita u *Kibani* kako bi se dobile informacije o klasteru i čvorovima:

```
1 GET _API/parameter
```

- Provera pokrenutosti *Elasticsearch* klastera (engl. *cluster*)

```
1 GET _cluster/health
```

- Ispis informacija o aktivnim čvorima (engl. *nodes*) unutar klastera

```
1 GET _nodes/stats
```

Sadržaj

Kreiranje indeksa i dokumenata

- Kreiranje indeksa čiji je naziv „proizvodi“

```
PUT proizvodi
```

- Format za kreiranje dokumenata u indeksu

```
POST <naziv_indeksa>/_doc
{
  "<polje>": "<vrednost>"
}
```

```
PUT naziv_indeksa/_doc/id_dokumenta
{
  "<polje>": "<vrednost>"
}
```

Napomena: Kada se koristi *POST* metoda, identifikator dokumenta se automatiski generiše, dok kada se koristi *PUT* metoda, kreiraće se dokument sa navedenim identifikatorom ukoliko ne postoji

Kreiranje indeksa i dokumenata – zadatak

- Kreirati novi dokument u indeksu „voće“. Novi dokument treba da sadrži informacije o voću: naziv, kategorija i cena.
  - **Naziv**: Jabuka (1 kg)
  - **Kategorija**: Prehrana
  - **Cena**: 95

  Napomena: Uraditi zadatak i sa *POST* i sa *PUT* metodom

```
PUT voce
```

```
POST voce/_doc
{
  "naziv": "Jabuka (1 kg)",
  "kategorija": "Prehrana",
  "cena": "95"
}
```

```
PUT voce/_doc/1
{
  "naziv": "Jabuka (1 kg)",
  "kategorija": "Prehrana",
  "cena": "95"
}
```

Dobavljane dokumenata

- Format dobavljanja dokumenata

```
1  GET <naziv_indeksa>/_doc/<id_dokumenta>
```

- Dobaviti dokument proizvoda čiji je id = 1

```
1  GET voce/_doc/1
```

Brisanje dokumenata

- Format brisanja dokumenata

```
1  DELETE <naziv_indeksa>/_doc/<id_dokumenta>
```

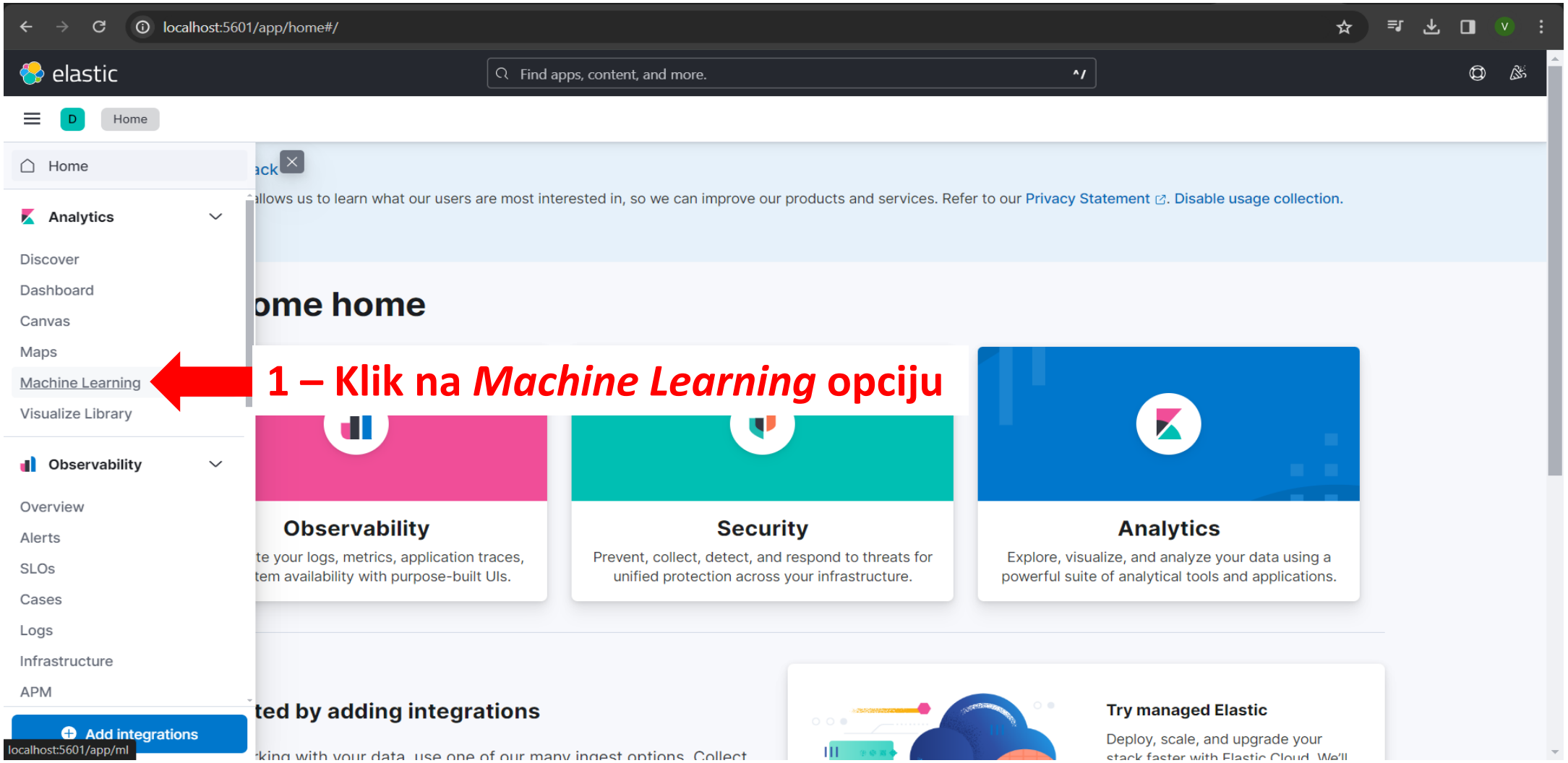
- Obrisati dokument proizvoda čiji je id = 1

```
1  DELETE voce/_doc/1
```

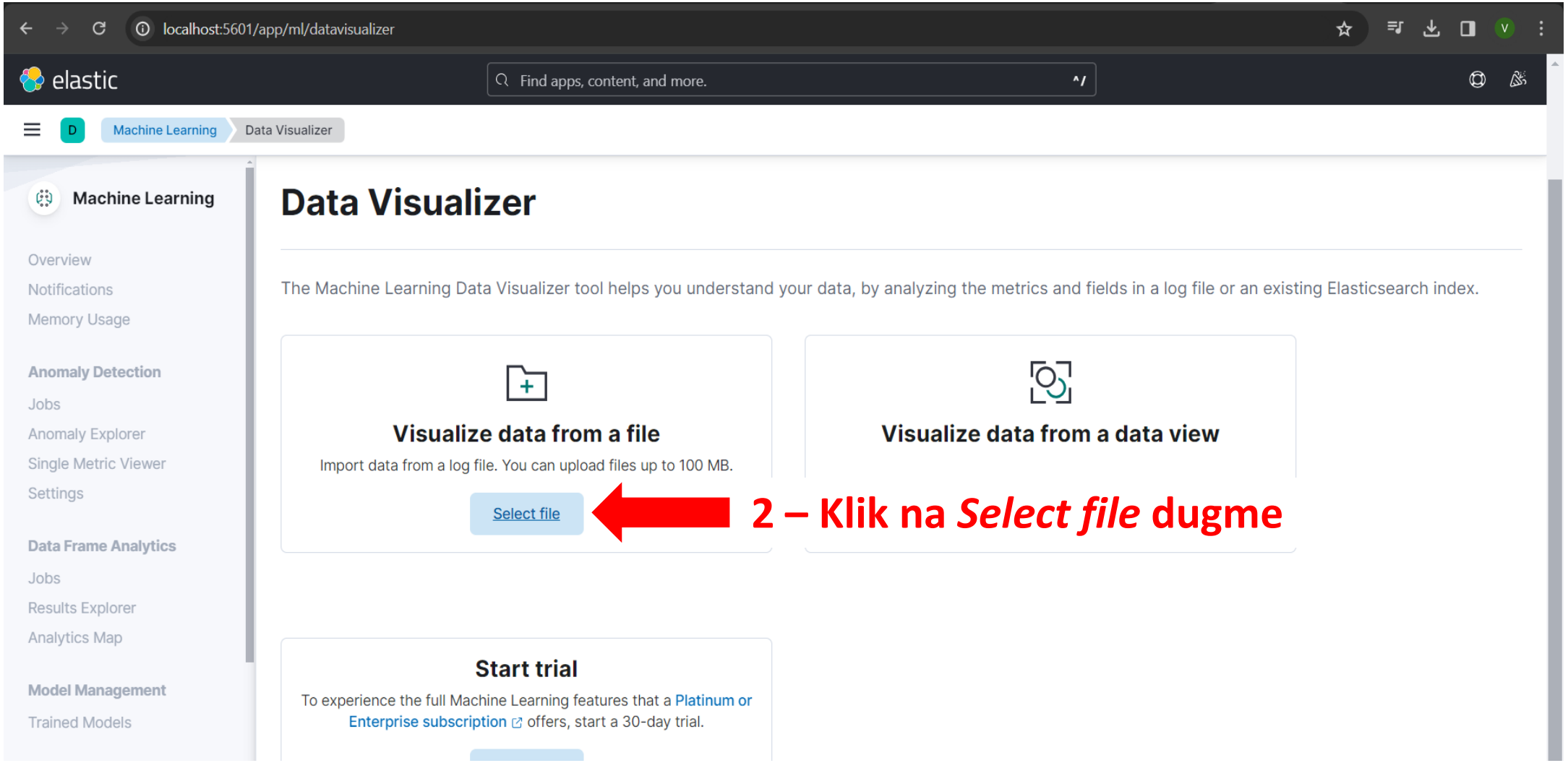
Sadržaj

- Softverska podrška
- Uvod u *Elasticsearch*
- Pregled alata za vizualizaciju – *Kibana*
- Operacije CRUD pomoću *Elasticsearch*-a
- Pretraživanje podataka pomoću upitnog jezika *Elasticsearch*
- Korisni linkovi

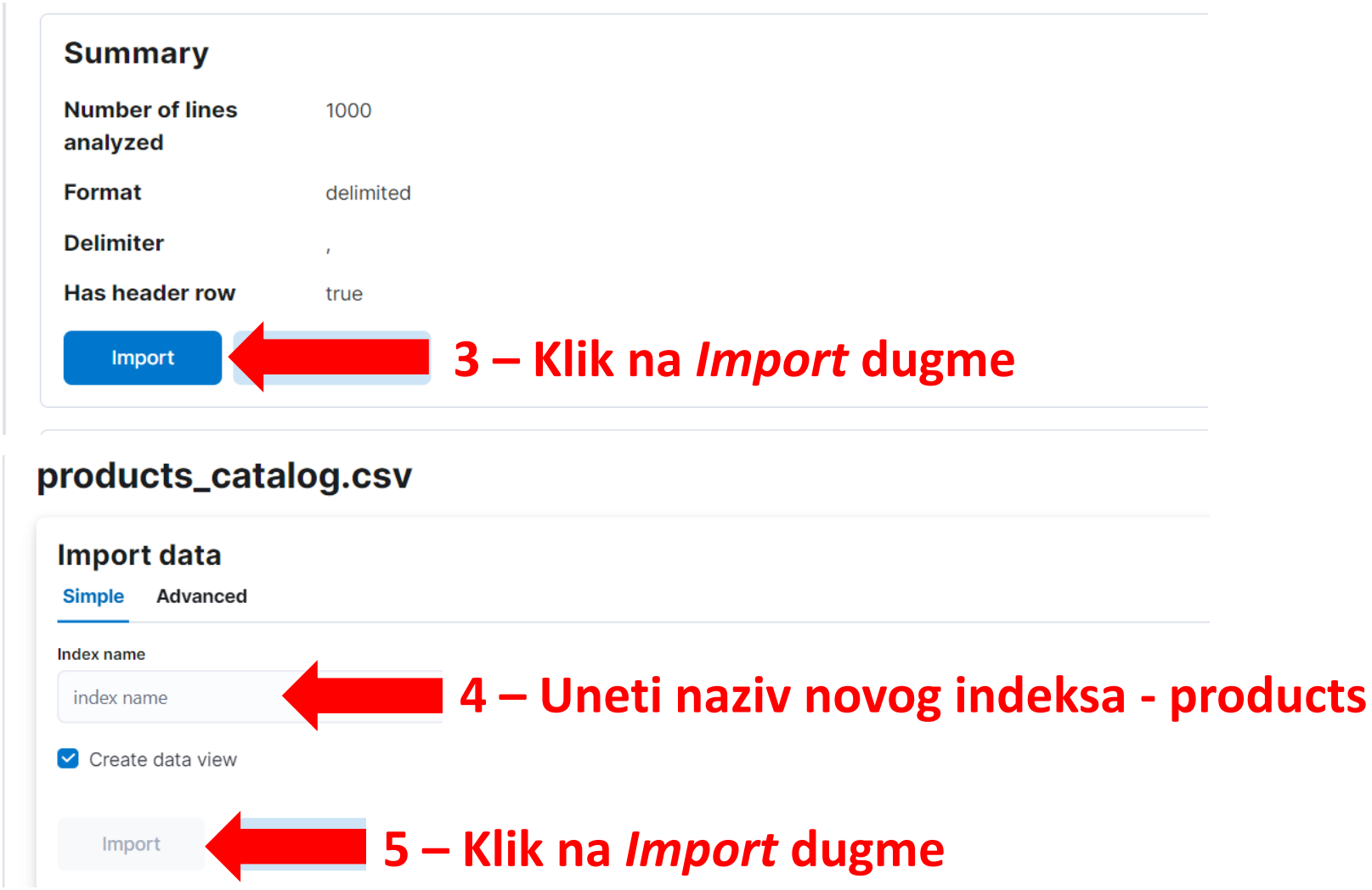
Dodavanje skupa podataka (*dataset*) u *Kibana* okruženje



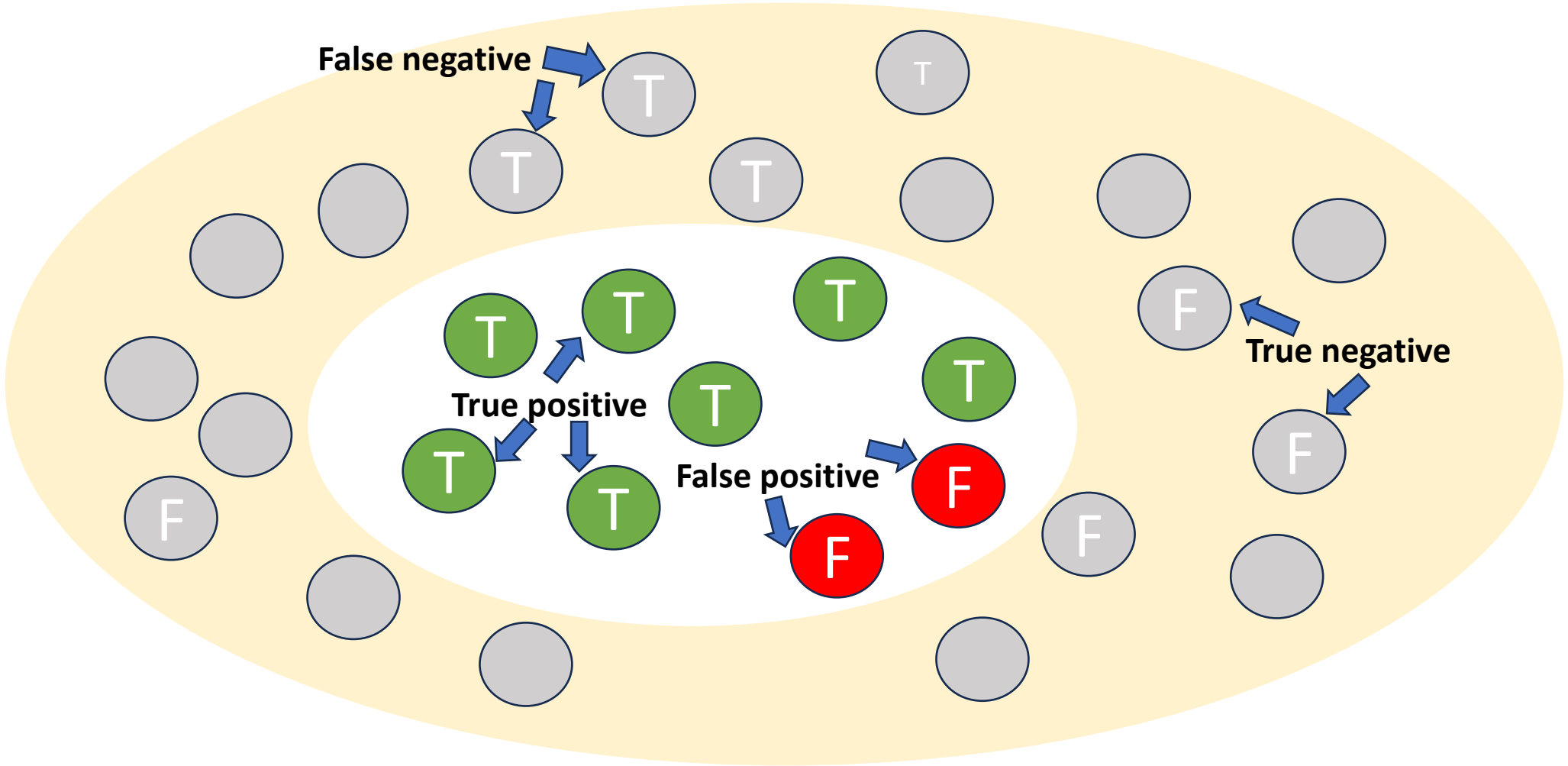
Dodavanje skupa podataka (*dataset*) u *Kibana* okruženje



Dodavanje skupa podataka (*dataset*) u *Kibana* okruženje



Tačnost i odziv



Tačnost i odziv

$$\textbf{Tačnost} = \frac{True\ positives}{True\ positives + False\ positives}$$

Proporcija stvarno pozitivnih predikcija u odnosu na ukupan broj pozitivnih predikcija

$$\textbf{Odziv} = \frac{True\ positives}{True\ positives + False\ negatives}$$

Proporcija stvarno pozitivnih predikcija u odnosu na ukupan broj traženih elemenata

Osnovni načini pretrage dokumenata

- **Upiti**
  - **Klasični upiti** – Pretraga dokumenata prema određenim kriterijumima
  - ***Full-Text* upiti** – Pretraga po celom tekstu dokumenata
  - ***Fuzzy* upiti** – Pronalaženje sličnih dokumenata (npr. po zvuku ili pravopisu)
  - **Geografski upiti** – Pretraga dokumenata po geografskoj lokaciji
- **Agregacije**
  - **Metričke agregacije** – Izračunavanje statističkih metrika na osnovu vrednosti polja
  - **Baket agregacije** – Grupisanje dokumenata u bakete na osnovu zadatog kriterijuma
  - ***Pipeline* agregacije** – Obrađivanje rezultata drugih agregacija za generisanje novih rezultata

Pretraga dokumenata – upiti

- Format kreiranja dokumenata u indeksu

```
GET <naziv_indeksa>/_search
```

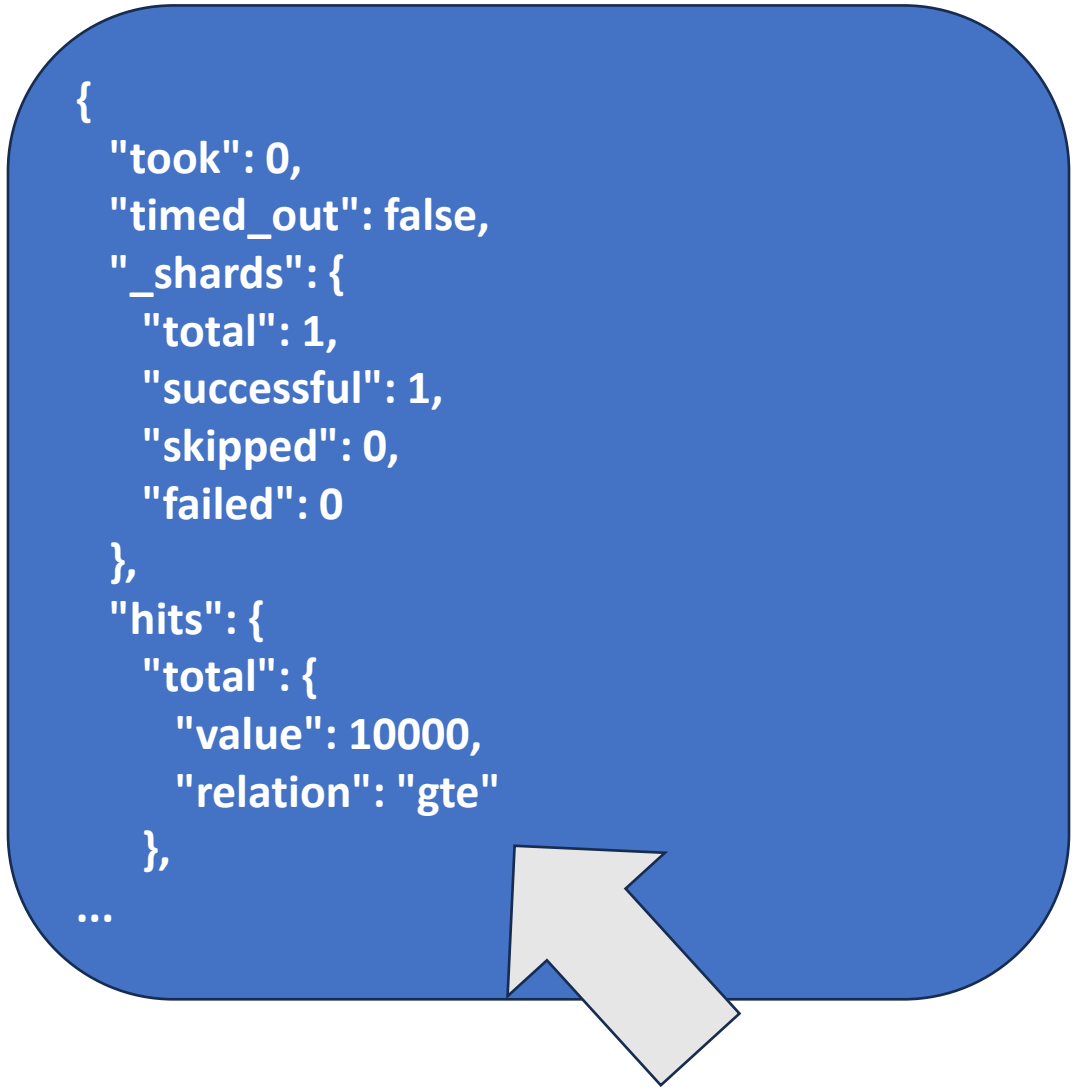
- Dobaviti sve dokumente iz indeksa *products*

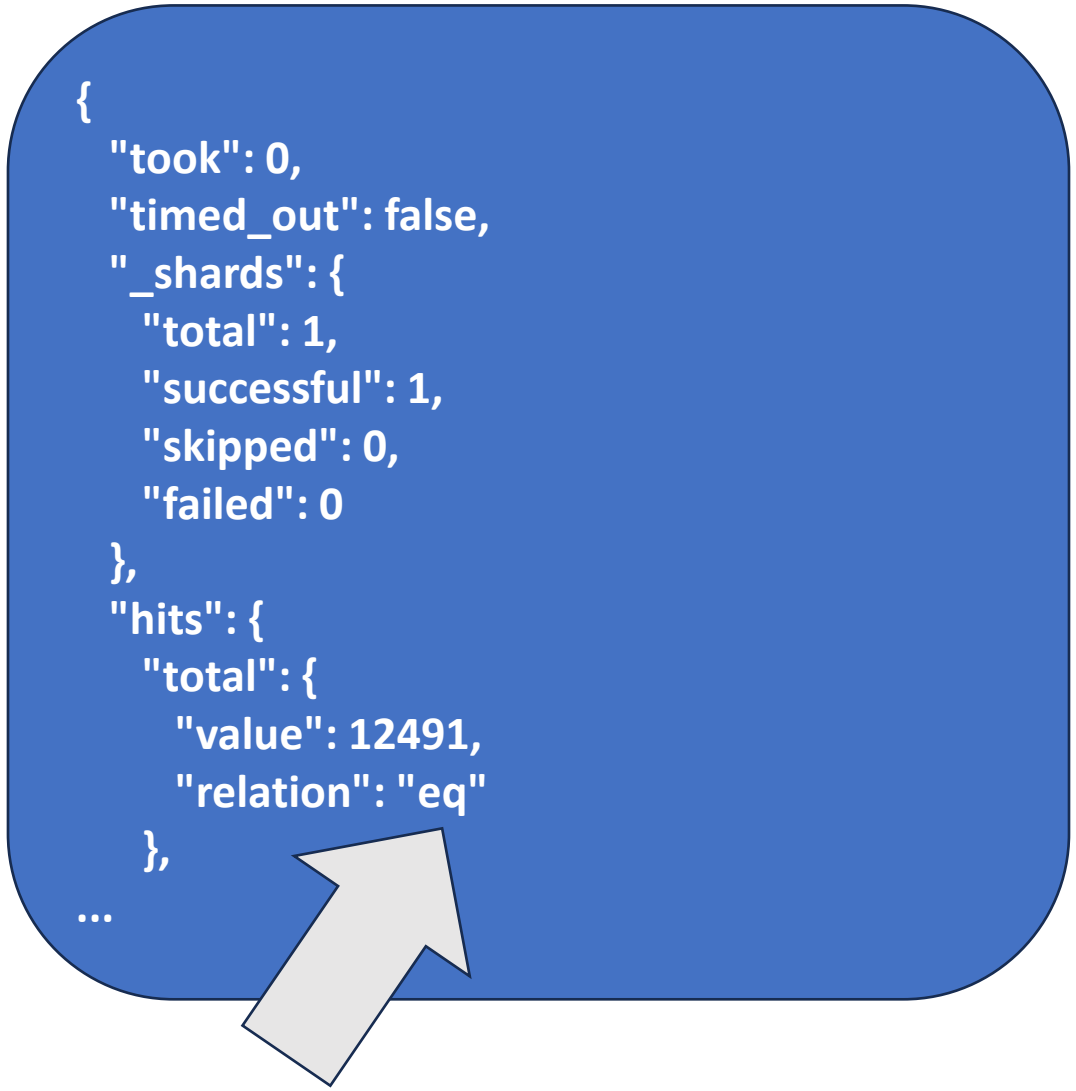
```
GET products/_search
```

- Da li su dobijeni svi dokumenti?

```
GET products/_search
{
  "track_total_hits": true
}
```

Napomena: Pomoću opcije *track_total_hits,* moguće je naglasiti da je potrebno dobaviti sve dokumente. U suprotnom, biće dobavljen optimalan broj dokumenata u skladu sa resursima (do 10000 dokumenata).





Razlika u izlaznim rezultatima

Pretraga dokumenata – upiti

- Format pretrage dokumenata u indeksu po opsegu vrednosti

```
GET <naziv_indeksa>/_search
{
  "query": {
    "range": {
      "<naziv_polja>": {
        "gte": "<donja_granica>",
        "lte": "<gornja_granica>"
      }
    }
  }
}
```

Pretraga dokumenata – upiti – zadatak

- Prikazati sve dokumente koji su nastali u avgustu 2022. godine (polje *DateCreated*)

```
GET products/_search
{
  "query": {
    "range": {
      "DateCreated": {
        "gte": "8/1/2022",
        "lte": "8/31/2022"
      }
    }
  }
}
```

Pretraga dokumenata – upiti

- Format pretrage dokumenata u indeksu na osnovu sadržaja tekstualnog polja

```
GET <naziv_indeksa>/_search
{
  "query": {
    "match": {
      "<naziv_polja>": {
        "query": "<ključne_reči_na_osnovu_kojih_se_vrši_pretraga>"
      }
    }
  }
}
```

Pretraga dokumenata – upiti – zadatak

- Prikazati sve dokumente koji sadrže u svom opisu (polje *Description*) sledeći niz ključnih reči *„Grey textured handheld bag“*

```
GET products/_search
{
  "query": {
    "match": {
      "Description": {
        "query": "Grey textured handheld bag"
      }
    }
  }
}
```

Koliko proizvoda je dobijeno kao rezultat upita? Da li je prednost pružena preciznosti ili odzivu?

Pretraga dokumenata – upiti

- Prikazati sve dokumente koji sadrže u svom opisu (polje *Description*) sledeći niz ključnih reči *„Grey textured handheld bag“*

```
GET products/_search
{
  "query": {
    "match": {
      "Description": {
        "query": "Grey textured handheld bag",
        "operator": "and"
      }
    }
  }
}
```

Napomena: Ukoliko je dodat operator *and*, biće prikazani samo dokumenti koji sadrže sve navedene ključne reči

Pretraga dokumenata – upiti

- Prikazati sve dokumente koji sadrže u svom opisu (polje *Description*) sledeći niz ključnih reči *„Grey textured handheld bag“*

```
GET products/_search
{
  "query": {
    "match": {
      "Description": {
        "query": "Grey textured handheld bag",
        "minimum_should_match": 3
      }
    }
  }
}
```

Napomena: Ukoliko je dodato polje *minimum_should_match*, biće prikazani samo dokumenti koji sadrže minimalno *N* navedenih ključnih reči (u predstavljenom primeru minimalno tri ključne reči)

Pretraga dokumenata – agregacije

- **Agregacija po pojmovima (engl. *Term Aggregation*):**
  - Grupiše dokumente po jedinstvenim vrednostima određenog polja
  - Korisno za analizu distribucije podataka po kategorijama ili ključnim rečima
- **Agregacija po histogramu datuma (engl. *Date Histogram Aggregation*):**
  - Grupiše dokumente u vremenske intervale (npr. po danima, mesecima ili godinama) na osnovu datuma
  - Omogućava analizu trendova ili sezonskih varijacija u podacima
- **Agregacija suma (engl. *Sum Aggregation*):**
  - Računa sumu vrednosti numeričkog polja
  - Korisno za izračunavanje ukupnih količina, prihoda ili drugih numeričkih metrika
- **Agregacija prosek (engl. *Average Aggregation*):**
  - Računa prosečnu vrednost numeričkog polja
  - Omogućava analizu prosečnih vrednosti ili performansi
- **Agregacija minimum/maksimum (engl. *Min/Max Aggregation*):**
  - Pronalazi minimalnu ili maksimalnu vrednost numeričkog polja
  - Korisno za identifikovanje najmanjih ili najvećih vrednosti u skupu podataka

Pretraga dokumenata – agregacije

- Opšti format agregiranja dokumenata – agregrianje po pojmovima

```
GET naziv_indeksa/_search
{
  "aggs": {
    "<naziv_agregacije>": {
      "<tip_agregacije>": {
        "field": "<naziv_polja_za_agregaciju>",
        ["size": "<broj_jednistvenih_vrednosti_koje_je_potrebno_vratiti>" –
        koristi se za agregaciju po pojmovima]
        ["interval": "<dan_ili_mesec_ili_godina>" – koristi se za agregaciju
        po histogramu datuma]
      }
    }
  }
}
```

Pretraga dokumenata – agregacije

- Format agregiranja dokumenata po pojmovima

```
GET naziv_indeksa/_search
{
  "aggs": {
    "<naziv_agregacije>": {
      "terms": {
        "field": "<naziv_polja_za_agregaciju>",
        "size": "<broj_jednistvenih_vrednosti_koje_je_potrebno_vratiti>"
      }
    }
  }
}
```

Pretraga dokumenata – agregacije – zadatak

- Prebrojati broj proizvoda koji se nude u prodavnici po brendu (polje *ProductBrand*) i ograničiti broj jedinstvenih vrednosti na 10

```
GET products/_search
{
  "aggs": {
    "by_brend": {
      "terms": {
        "field": "ProductBrand",
        "size": 10
      }
    }
  }
}
```

Sadržaj

Korisni linkovi

- **ElasticSearch – zvanična dokumentacija**
  - *https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html*
- **YouTube tutorijal o osnovama ElasticSearch-a**
  - *https://youtube.com/playlist?list=PL_mJOmq4zsHZYAyK606y7wjQtC0aoE6Es&si=sniZIgAmU-6lE76w*

Napredne arhitekture informacionih sistema

Elasticsearch

Pitanja?

Izvođači nastave:
dr Marko Vještica
Elena Akik
Sanja Radić

DAIS
Data Science and Information Systems Group