WYBRANE METODY PRZEPROWADZANIA I DETEKCJI WEB SCRAPINGU

Michał Tułowiecki
dr inż. Mariusz Sepczuk

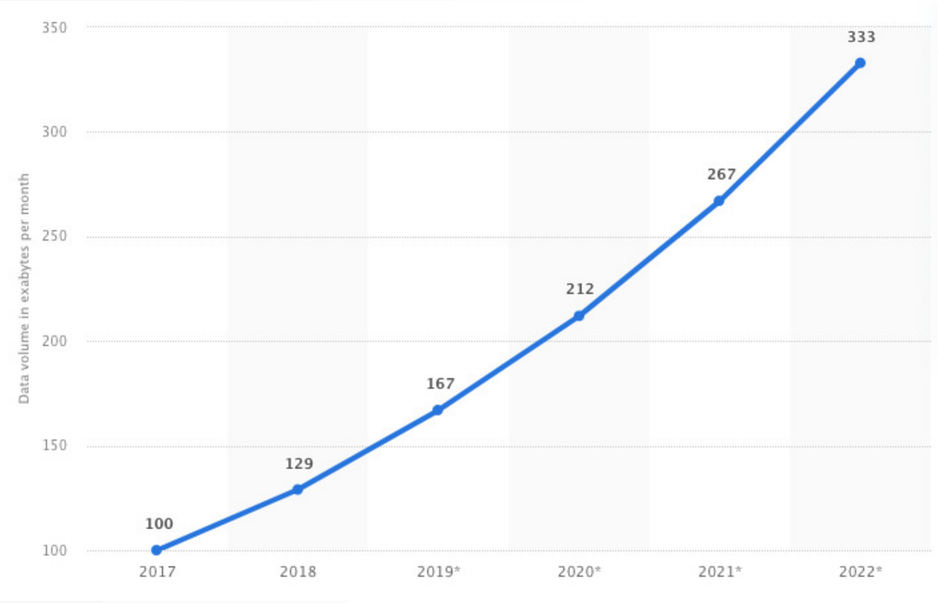
AGENDA

WEB SCRAPING

Web Scraping to **technika pozyskiwania informacji** z zasobów WWW. Proces ten najczęściej wykorzystuje automatyzację za pomocą specjalnego oprogramowania.

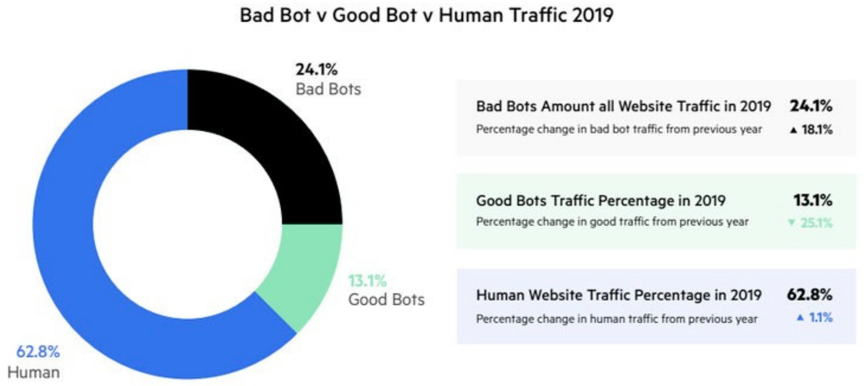
MOTYWACJA

Ograniczone zasoby ludzkie pozwalają na przetworzenie jedynie niewielkiego ułamka danych umieszczonych w Internecie.



Wolumen danych w globalnym konsumenckim ruchu IP w latach 2017–2022 (w eksabajtach na miesiąc), Tiago Bianchi

Technika Web Scrapingu wciąż zyskuje na popularności stając się nieodzownym narzędziem w pracy z danymi.



Procentowy rozkład ruchu generowany w internecie przez: złe boty, dobre boty i ludzi w roku 2019, lunio.ai

01

ANALIZA I PRAKTYCZNE WYKORZYSTANIE WEB SCRAPINGU

CEL PRACY

02

OPRACOWANIE I WDROŻENIE METOD BLOKUJĄCYCH WEB SCRAPING

REALIZACJA

SKLEP INTERNETOWY

- Witryna internetowa
- Zaplecze technologiczne

SCRAPER

- Podejście typu black box
- API Scraping
- Cel: pobranie szczegółowych danych o wszystkich produktach

REALIZACJA

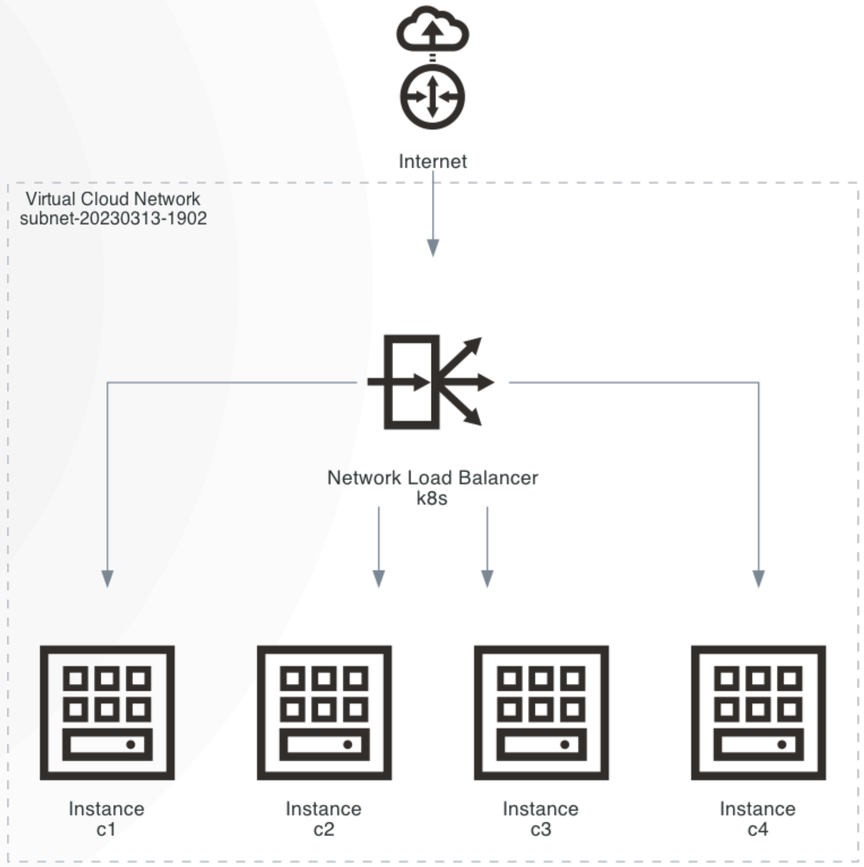
SKLEP INTERNETOWY

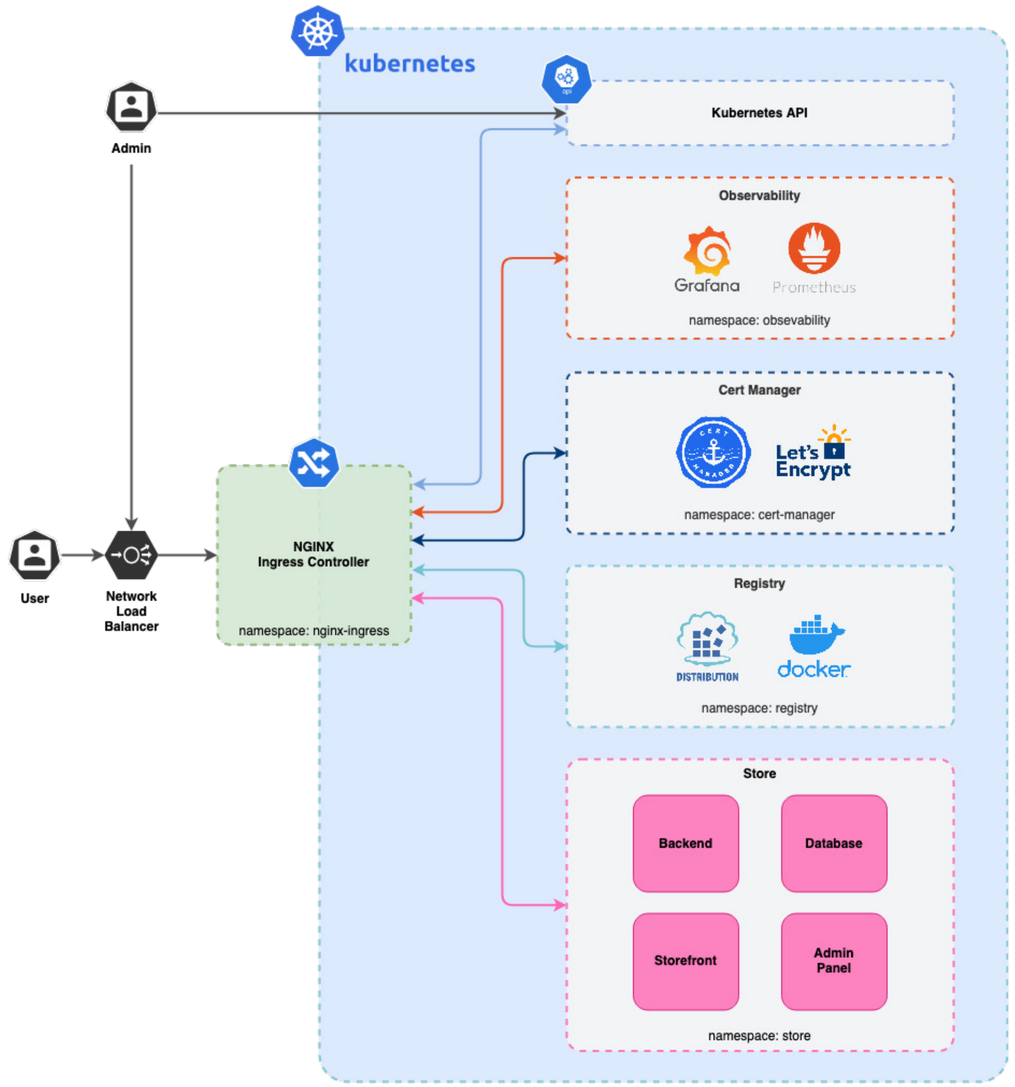
- Witryna internetowa
- Zaplecze technologiczne
- Metody detekcji web scrapingu:
  - Rate Limiting
  - Bot Blocker Proxy
  - Browser Fingerprinting

SCRAPER

- Podejście typu black box
- API Scraping
- Cel: pobranie szczegółowych danych o wszystkich produktach

INFRASTRUKTURA





RATE LIMITING

```yaml
apiVersion: k8s.nginx.org/v1
kind: VirtualServer
metadata:
  name: backend-vs
  namespace: store
spec:
  ingressClassName: public
  host: api.tulski.com
  upstreams:
    - name: backend
      service: backend
      port: 9000
  routes:
    - path: /store
      policies:
        - name: store-api-rate-limit-policy
      action:
        pass: backend
    ...
```

```yaml
apiVersion: k8s.nginx.org/v1
kind: Policy
metadata:
  name: store-api-rate-limit-policy
  namespace: store
spec:
  rateLimit:
    rate: 30r/m
    key: ${binary_remote_addr}
    burst: 5
    delay: 10
    zoneSize: 10M
    rejectCode: 429
```

BOT BLOCKER PROXY

robots.txt I Bad Referrers I IP Addresses and Ranges I User-Agents

BROWSER FINGERPRINTING

Bot analysis result:


```
{
  "id": "n-FGkHkLXvYFb8EmIbH4b",
  "created_at": "2024-01-07T19:05:18.207Z",
  "bot": {
    "result": "bad_bot"
  }
}
```


Fingerprint:


```
{
  "plugins": [
    "PDF Viewer::Portable Document Format::internal-pdf-v
    "Chrome PDF Viewer::Portable Document Format::interna
    "Chromium PDF Viewer::Portable Document Format::inter
    "Microsoft Edge PDF Viewer::Portable Document Format:
    "WebKit built-in PDF::Portable Document Format::inter
  ],
  "mimeTypes": [
    "Portable Document Format~~application/pdf~~pdf",
    "Portable Document Format~~text/pdf~~pdf"
  ],
  "userAgent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_
  "platform": "MacIntel",
  "languages": [
    "en-GB",
    "en-US",
    "en"
  ],
  "screen": {
```

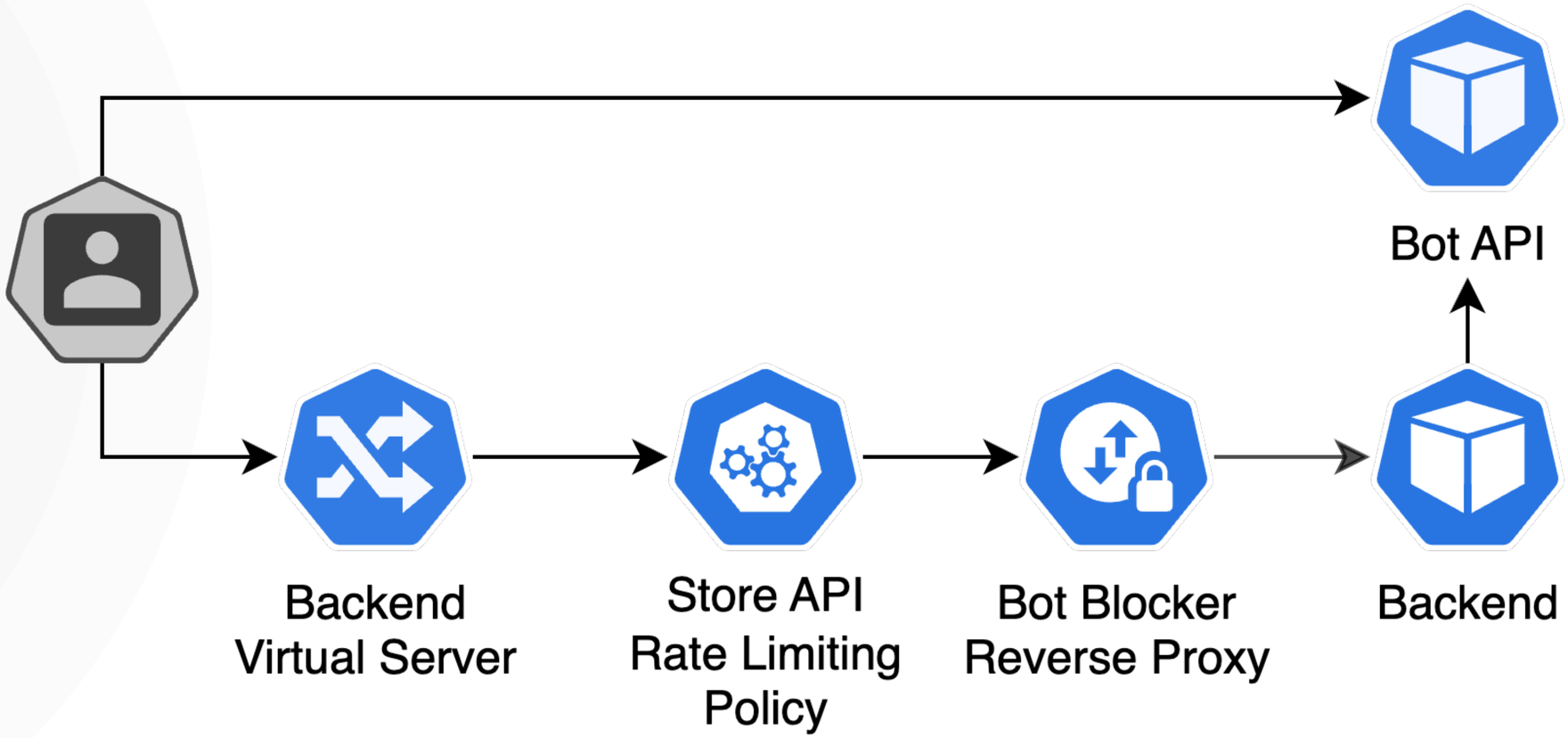

Bot analysis result:


```
{
  "id": "Y7353MmNy3R8OGwCt4Dgm",
  "created_at": "2024-01-07T19:08:17.331Z",
  "bot": {
    "result": "not_detected"
  }
}
```


Fingerprint:


```
{
  "plugins": [
    "PDF Viewer::Portable Document Format::internal-pdf-v
    "Chrome PDF Viewer::Portable Document Format::interna
    "Chromium PDF Viewer::Portable Document Format::inter
    "Microsoft Edge PDF Viewer::Portable Document Format:
    "WebKit built-in PDF::Portable Document Format::inter
  ],
  "mimeTypes": [
    "Portable Document Format~~application/pdf~~pdf",
    "Portable Document Format~~text/pdf~~pdf"
  ],
  "userAgent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10
  "platform": "MacIntel",
  "languages": [
    "en-US",
    "en",
    "pl"
  ],
  "screen": {
```

SCHEMAT WDROŻONYCH METOD



TESTY

Przeprowadzono serię testów manualnych mających na celu weryfikację poprawności działania scrapera i sklepu internetowego z uwzględnieniem wdrożonych metod detekcji web scrapingu.

1. ***Bez zabezpieczeń***
2. ***tylko Bot Blocker Proxy***
3. ***tylko rate limiting***
4. ***tylko Bot API***
5. ***włączone wszystkie metody detekcji***

PODSUMOWANIE

- Web scraping jawi się jako **atrakcyjne i efektywne narzędzie** do zbierania danych.
- Scraper okazał się relatywnie **łatwy i szybki w implementacji**.
- Wdrożone metody detekcji web scrapingu **skutecznie zablokowały** scraper.
  - Rate Limiting
  - Bot Blocker Proxy
  - Browser Fingerprinting
- Projekt posiada **potencjał do dalszego rozwoju** w zakresie detekcji web scrapingu.