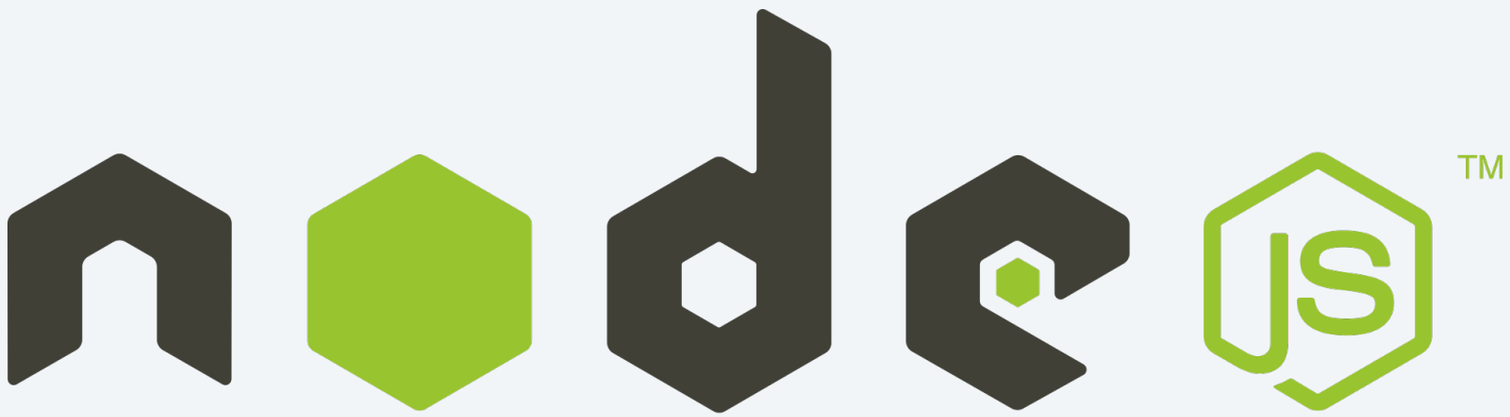
Web Scrapping

with



ABOUT ME

Hengki Sihombing

Building Karejo.com - Organizer JakartaJS

twitter @hengkiardo

github @aredo

hengki@karejo.com

Schedule

- What is Web Scrapping
- Why we do that
- How we do that in Node.js
- Our Target
- Demo Code

What is Web Scrapping

Web scraping is a computer software technique of extracting information from websites.

https://en.wikipedia.org/wiki/Web_scraping

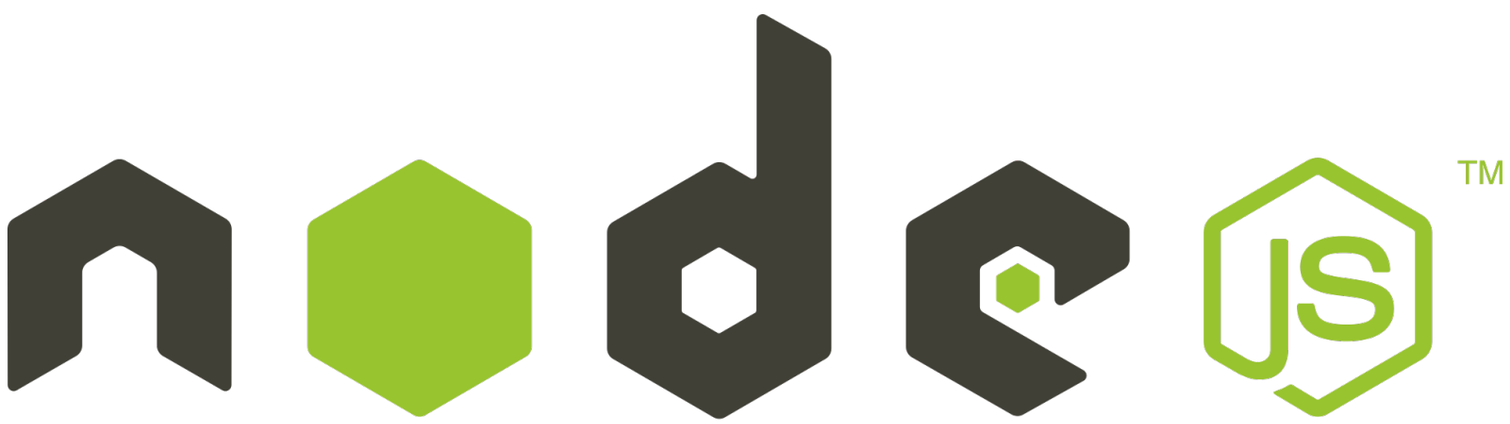
Why we do that

- To get more accessible data
- Improve our hacking skill
- Have fun and build some nice project

Some product doing Web scrapping

- Wego, Skyscanner
- Flipboard, Instapaper, Nuzzel
- Telunjuk.com
- Karejo.com
- etc..

How we do with



Our Dependencies

Request
Cheerio
Async
Mongoose
Express
Agenda

```
{
  "name": "jakartajs-des-2015",
  "version": "1.0.0",
  "main": "app.js",
  "scripts": {
    "start": "nodemon app.js"
  },
  "repository": {
    "type": "git",
    "url": "git@github.com:aredo/jakartajs-des-2015.git"
  },
  "author": "Hengki Sihombing <hengki.sihombing@gmail.com>",
  "license": "(ISC OR GPL-3.0)",
  "dependencies": {
    "agenda": "^0.6.28",
    "agenda-ui": "0.0.7",
    "async": "^1.4.2",
    "cheerio": "^0.19.0",
    "express": "^4.13.3",
    "moment": "2.10.6",
    "mongoose": "^4.1.11",
    "mongoose-timestamp": "^0.4.0",
    "request": "^2.65.0",
    "lodash": "^3.10.1"
  }
}

```

Step by step Scapping

```
[
    "Requesting Pages",
    "Parsing HTML",
    "Traversing Dom",
    "Collecting Data",
    "Save to MongoDB"
]
```

Request

```
var Request    = require('request')

var options = {
  url: url,
  headers: {
    'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_10_1
      AppleWebKit/537.36 (KHTML, like Gecko) Chrome/44.0.2403.125
      Safari/537.36',
  }
};

Request(options, function (err, response, body) {

  if (err) {
    return throw new Error()
  }

  callback(err, body)
})
```

Cheerio

```
var cheerio = require('cheerio')

var html = `<!DOCTYPE html>
<html lang="en" class="no-js">
  <head>
  </head>
  <body>
    <h2 class="title">Hello world</h2>
  </body>
</html>`

$ = cheerio.load(html)

var title = $('h2').text()

console.log(title)
```

Our Target

SEMUA KATEGORI

Cari produk, kategori, atau merk

Cari

1

VOUCHER RP 50.000
DAPATKAN DISINI

SISTEM OPERASI

- Android (6130)
- Android 2.2 Froyo (5)
- Android 2.3 (4)
- Android 2.3 Gingerbread (3)
- Android 4.0 Ic (2)
- Android 4.0 Ice Cream (3)
- Android 4.1 Jelly Bean (25)
- Android 4.2 Jelly Bean (13)

more...

HARGA

6000 RP – 4646464

JUMLAH PORT USB

- 1 (11)
- 2 (10)
- 3 (2)

TIPE KABEL

- Aux (117)
- DVI (24)
- HDMI (273)
- Lainnya (113)
- Lighting (301)
- Micro USB (3035)
- USB (2932)

Handphone & Tablet

840427 Results

ATUR BERDASARKAN: Terpopuler

Grid List

Mi 4i - 16GB - Putih

~~RP 2.999.000~~ - 13%

RP 2.599.000

(122 reviews)

Seller lainnya dari RP 2.622.000

Lenovo A7000 Special Edition - 16 GB - ...

~~RP 2.399.000~~ - 4%

RP 2.299.000

DECLEN100 **RP 2.199.000**

(196 reviews)

Alcatel Flash 2 - 16 GB - Volcanic Gre ...

~~RP 2.499.000~~ - 20%

RP 1.999.000

(167 reviews)

Infinix Note 2 X600 4G LTE - 16 GB - A ...

~~RP 1.999.000~~ - 5%

RP 1.899.000

(21 reviews)

Samsung Galaxy J5 Dual SIM - 8 GB - Putih

~~RP 2.899.000~~ - 9%

RP 2.649.990

Lenovo A6000 - 4G LTE - 8 GB - Hitam

~~RP 1.699.000~~ - 19%

RP 1.375.000

Infinix Android One X510 Hot 2 - 16GB ...

~~RP 1.499.000~~ - 13%

RP 1.299.000

Delcell Power Bank Note Polymer Batter ...

~~RP 229.000~~ - 31%

RP 159.000

DemoCode

Agenda

```
var config = {
  mongodb: 'mongodb://127.0.0.1/jakartajs-des-2015',
}

var agendaOptions = {
    address: config.mongodb,
    collection: "cron"
  }

var agenda = new Agenda({ db: agendaOptions })

agenda.define('lazada.co.id - handphone tablet', function (job, done) {

require('./handphone-tablet.js')(done, job)

})

agenda.every('10 minutes', 'lazada.co.id - handphone tablet');

agenda.start()
```

Please!!

Let’s Join Us

KAREJO

Software Engineer

Front-End Developer

UI/UX Designer

let me know by email :

hengki@karejo.com

Thank you

https://jakartajs-join.herokuapp.com

meetup.com/JakartaJS