BIG DATA : Crawling

Ahmad Rio Adriansyah S.Si. M.Si.

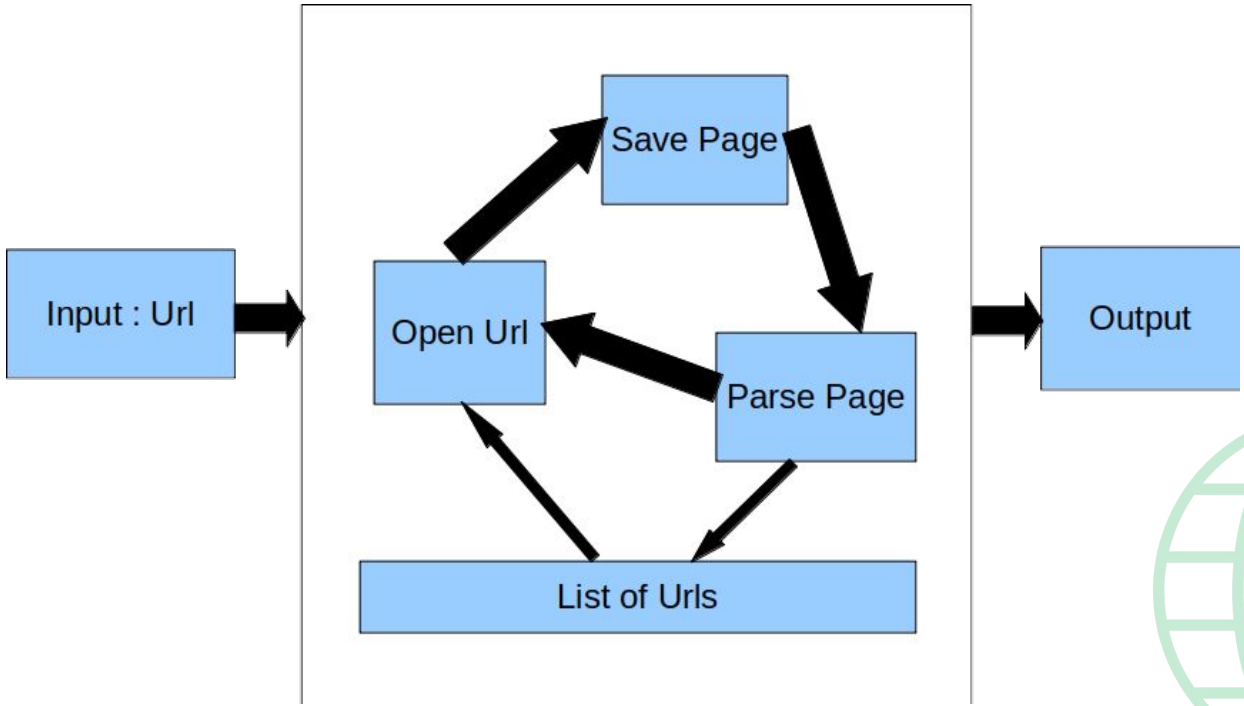
Getting Data

- Pengumpulan data primer mandiri
- Open Data dalam bentuk csv/sql/json/lainnya
- Web
  - Tidak semua web menyediakan data dalam bentuk yang siap pakai
  - Tetapi, selama data tersebut bisa diakses dari web, bisa kita ambil dan manfaatkan

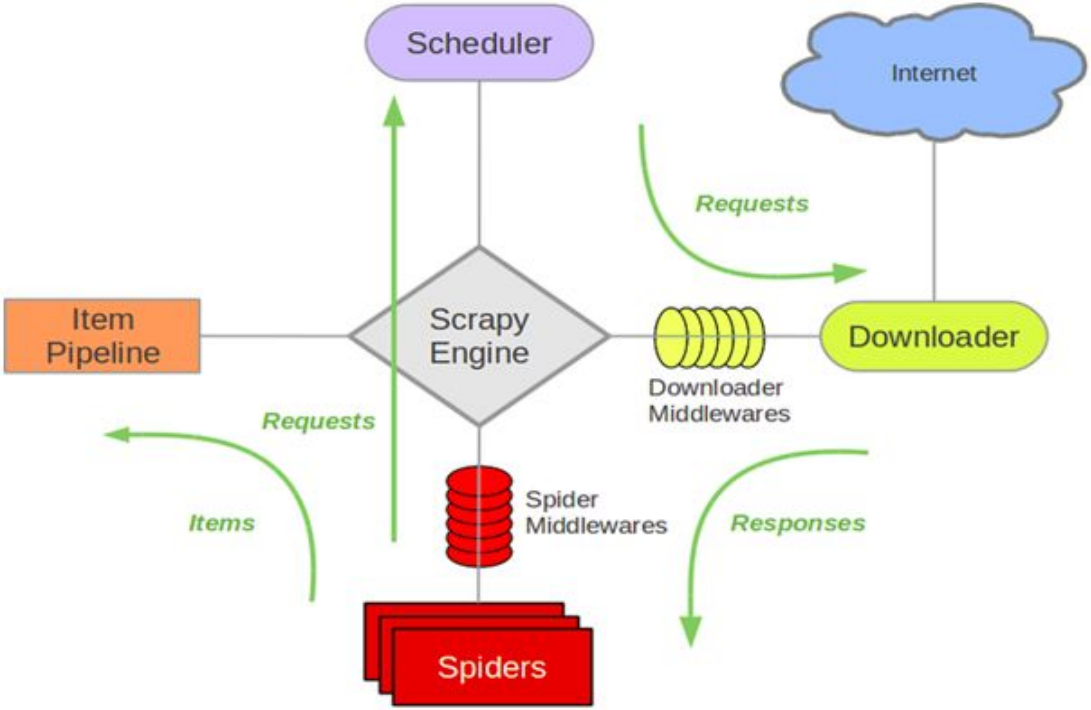
Scraping

- Teknik yang digunakan untuk mengambil data (teks, gambar, video, atau file) dari website secara otomatis
- Bisa menggunakan :
  - Add-on pada Browser (Web Scraping, Scraper, Data Scraping, Scrapbook, dll)
  - Script sederhana
  - Framework/Software untuk scraping (Scrapy, Grab, Yakuza, FMiner)

Alur



Alur Scrapy


Scheduler
Internet
Requests
Item
Pipeline
Scrapy
Engine
Downloader
Downloader
Middlewares
Requests
Items
Spider
Middlewares
Responses
Spiders

Scraping dengan Python

- Library yang dibutuhkan
  - urllib / urllib2 / urllib3 / requests
  - beautifulsoup (https://www.crummy.com/software/BeautifulSoup/)
  - Others (tergantung kebutuhan)

Akses URL

```
import requests
from bs4 import BeautifulSoup as BS
import pandas as pd
import time
```

```
url = 'http://detik.com'
s = requests.session()
r = s.get(url)
```

```
soup = BS(r.text)
```

Parsing

- Inspeksi elemen konten yang ingin diambil
- Daftarkan halaman yang akan dicrawling selanjutnya

Beautiful Soup

- Library untuk membantu parsing halaman html

```
print(soup.prettify())
```

```
soup.find_all('a')
```

```
soup.find('div',{'class':'menu'})
```

Looping dan simpan hasil

- Hasil bisa disimpan dalam bentuk array atau dictionary python
- Dapat disimpan dalam file baik berupa
  - File biasa (dengan fungsi print ke file connection)
  - File csv (dengan bantuan modul pandas)
  - File pickle (dengan bantuan modul pickle)

  *penyimpanan dengan pickel mudah tapi berbahaya

Demo

Ada website yang ingin dicoba ambil hasilnya?

Memasukkan data ke MongoDB

Lakukan perintah berikut untuk menyimpan file hasil crawling berformat json,csv, atau tsv ke dalam basis data MongoDB menggunakan command **mongoimport** :

```
$ mongoimport -d <nama basis data> -c <nama collection>
--type <json|csv|tsv> --file <nama file> --headerline
```

*gunakan --headerline untuk memanggil selain file json