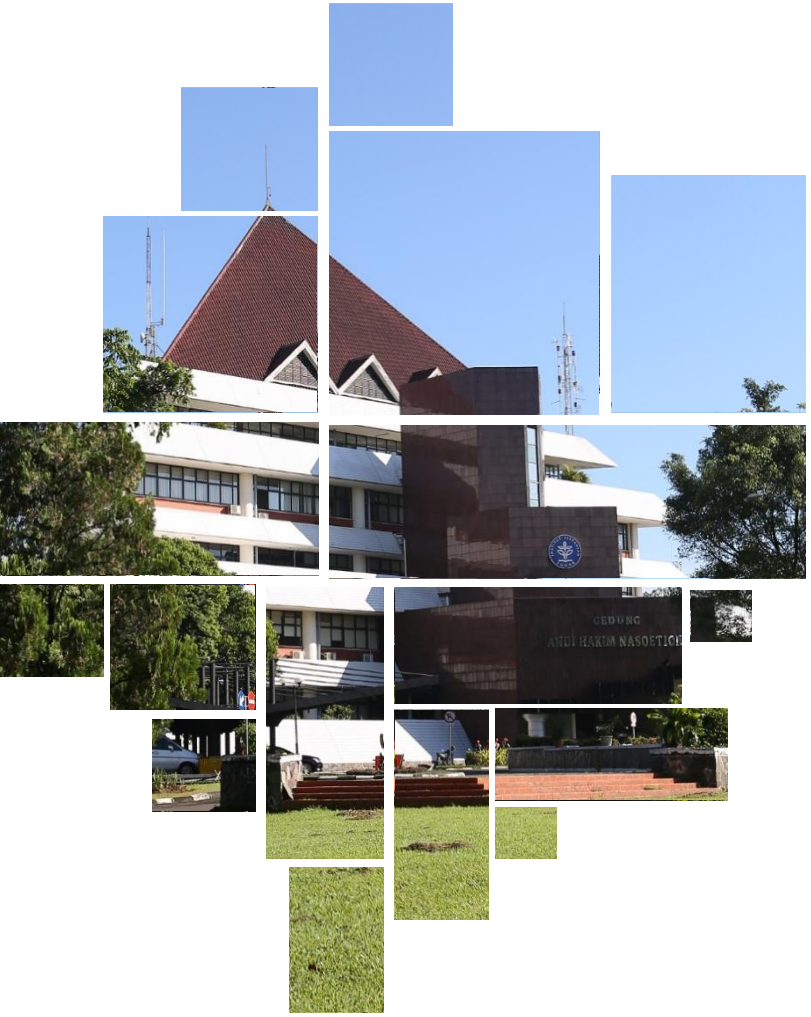
IPB University
Bogor Indonesia

Department of Statistics
Faculty of Mathematics Natural Sciences

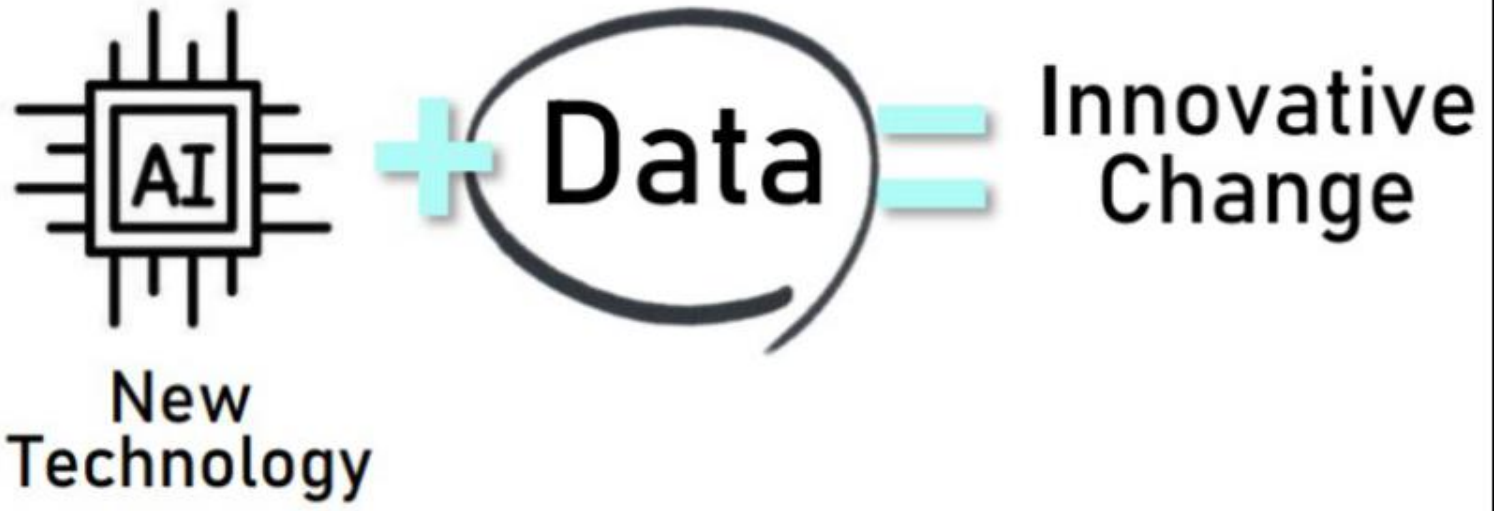


Digital Data Collection

Kuliah 11 | STA221 Metode Pengumpulan Data

Senin, 8 November 2021

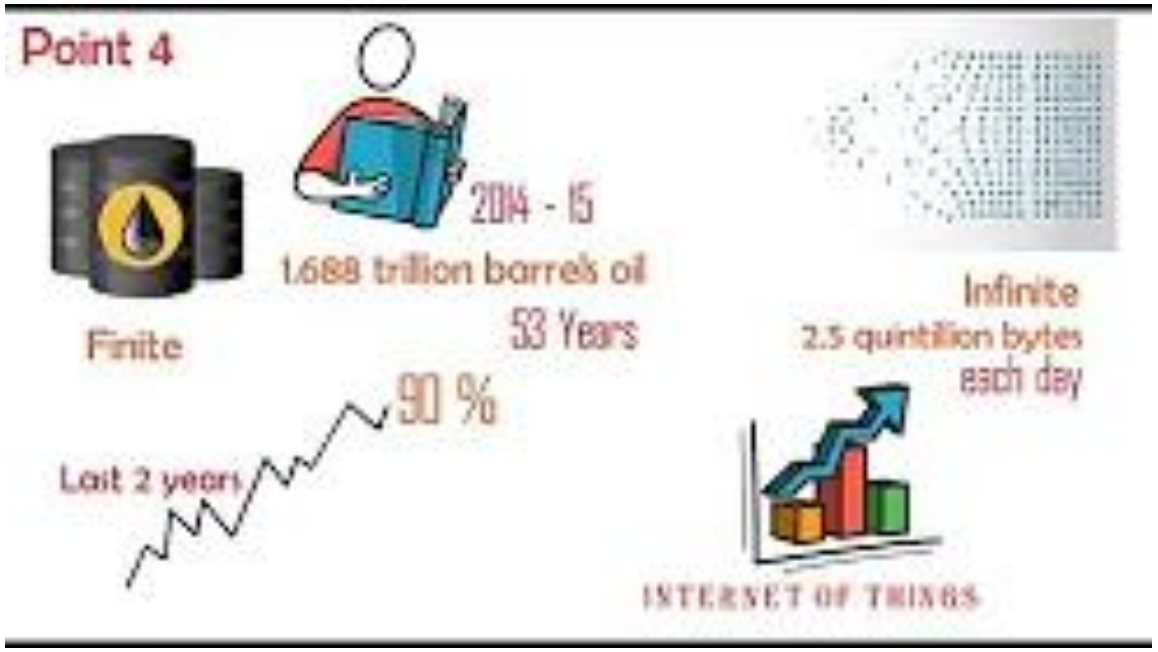
- *"The world's most valuable resource is no longer oil, but data."* **(Economist, May 6th 2017)**



2020 This Is What Happens In An Internet Minute

Created By:
@LoriLewis
@OfficiallyChadd

Simak video-video berikut



Why data is the new oil?
https://youtu.be/QGv02fDcdg4

Apa itu Big Data?
https://youtu.be/aC2CmTTZTVU

Simak video-video berikut

What is sport and exercise science?
https://youtu.be/9Hqdrx1HPfA

How Big Data Could Transform
The Health Care Industry
https://youtu.be/_mXrZEIpNMw

“

Data adalah jenis kekayaan baru bangsa kita. Kini data lebih berharga dari minyak. Oleh karena itu, kedaulatan data harus diwujudkan. Hak warga negara atas data pribadi harus dilindungi. Regulasinya harus segera disiapkan, tidak boleh ada kompromi.

Presiden RI, Joko Widodo dalam Pidato Kenegaraan 16 Agustus 2019

DULU sumber data berasal dari:

- SENSUS
- SURVEI
- REGISTRASI

SEKARANG sumber data juga berasal dari:

- BIG DATA
  - *Data administratif*
  - *Data digital komersial atau transaksional*
  - *Perangkat pelacakan GPS*
  - *Data perilaku*
  - *Data opini*

Sumber: Presentasi Direktorat Analisis dan Pengembangan Statistik – BPS, 22 Sept 2021

DATA SOURCES

- *Mobile phone data*
- *Financial transactions*
- *Online search and access logs*
- *Citizen card*
- *Postal data*

Exhaust data

- *Satellite and UAV imagery*
- *Sensors in cities, transport and homes*
- *Sensors in nature, agriculture and water*
- *Wearable technology*
- *Biometric data*
- *Internet of Things (IoT)*

Sensing data

What People Do

- *Social media data*
- *Web scraping*
- *Participatory sensing / crowdsourcing*
- *Health records*
- *Radio content*

Digital Content

What People Say

Sumber: Presentasi Direktorat Analisis dan Pengembangan Statistik – BPS, 22 Sept 2021

Big Data Initiatives and Developments in BPS

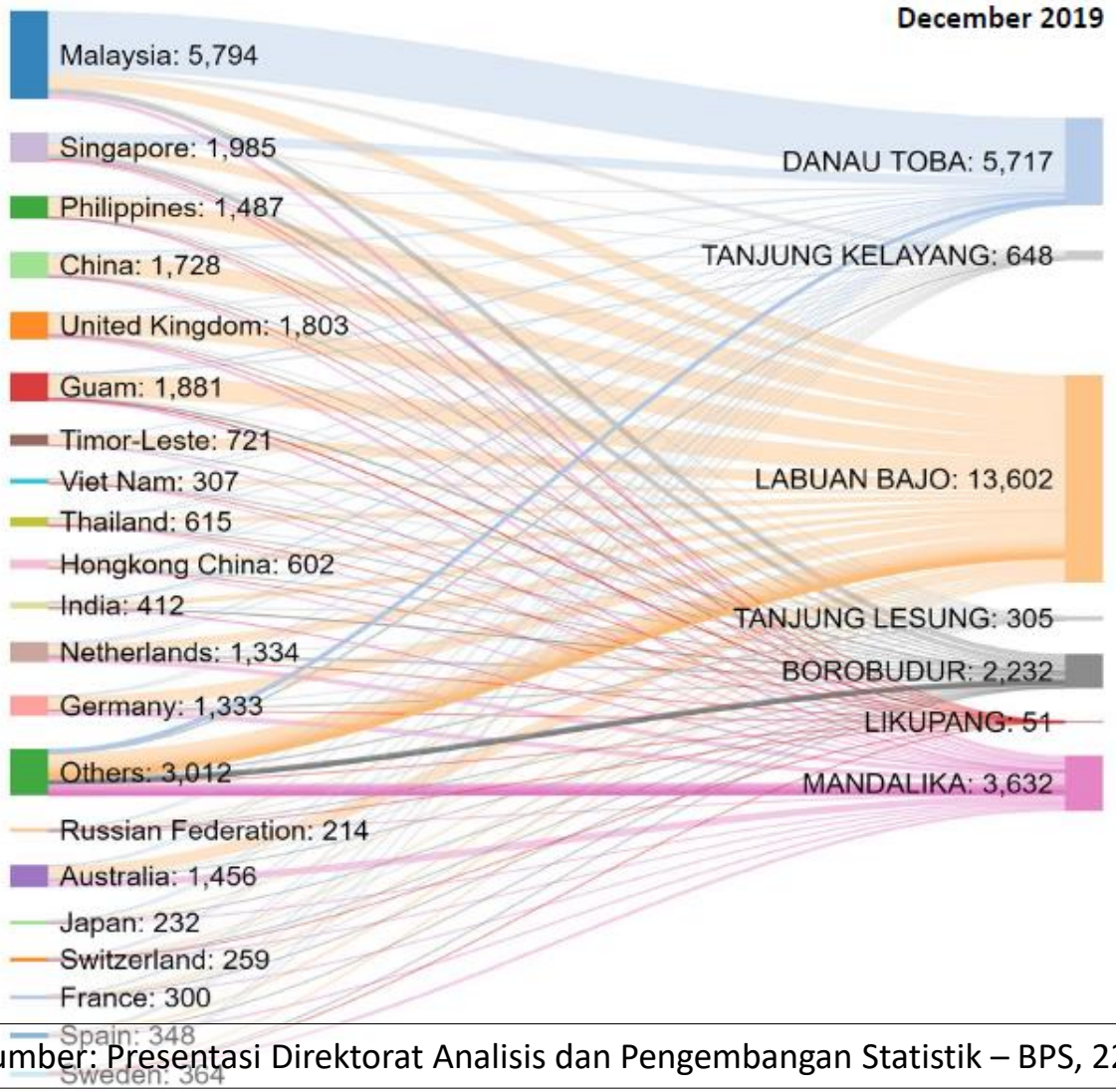
1. *Web-crawling*
   - a. *Marketplace* ▶ *E-commerce Data*
   - b. *Flight Tracker, bus booking site* ▶ Transportation analytics
   - c. Job Vacancy Site ▶ Labor analysis
   - d. Online booking site and review ▶ Room occupancy rate, Number of tourists, etc
   - d. Air Quality, weather reporting site ▶ Enviromental and disaster statistics
   - d. Online news and social media ▶ Current fenomena, citizen sensing
2. *Google and Facebook mobility index* ▶ People mobility
3. Satelite Imagery ▶ Sample Frame Area, Poverty mapping
4. *Mobile Phone Data* ▶ Number of international and domestic visitors, Metropolitan Statistical Area

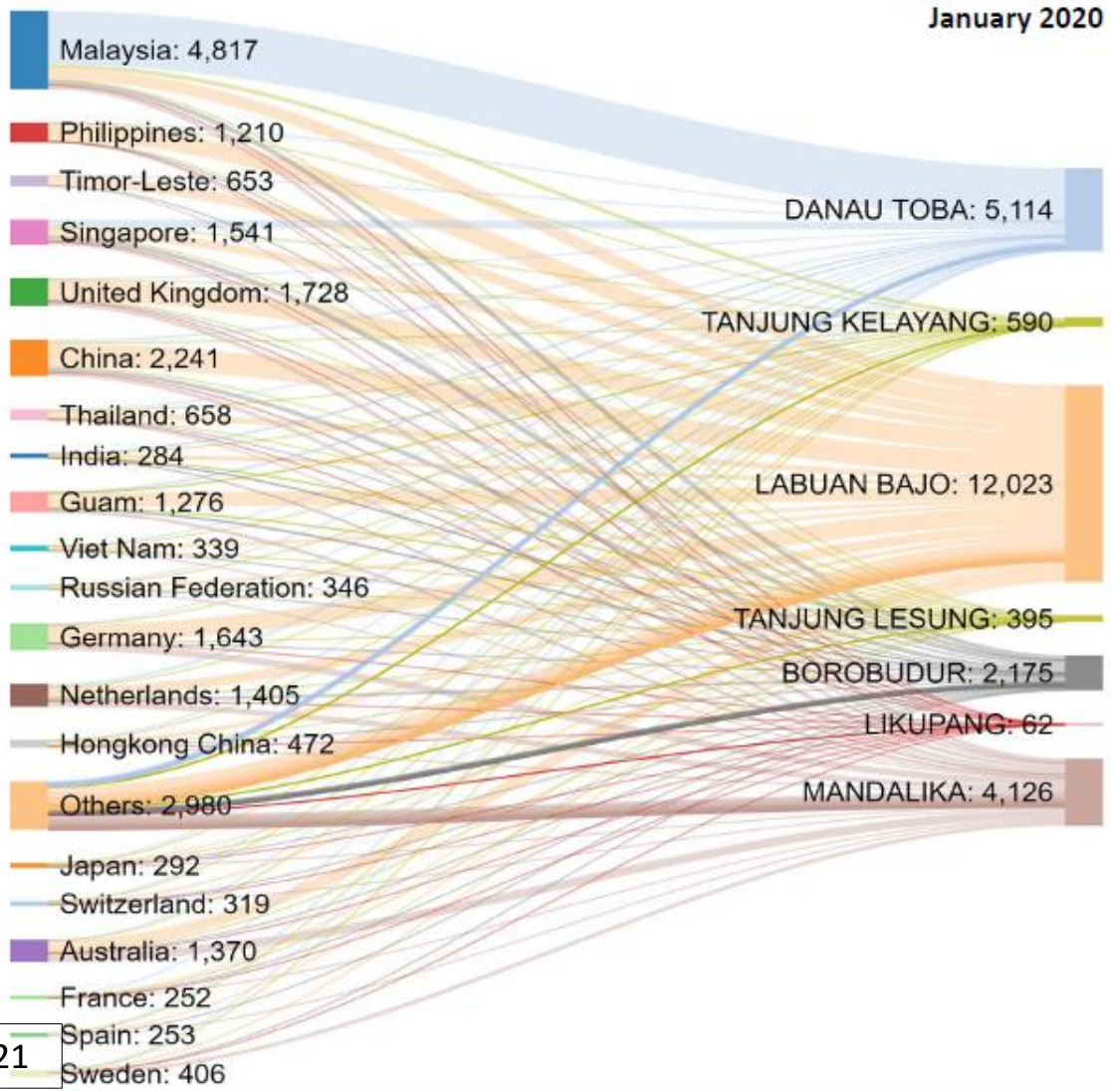
Challenges

- Data Access and Acquisition
- Data Source Quality
- Statistical Methodology
- New skills profile and tech
- Data Privacy and protection
- Regulation on National Statistical System
- Interoperability

Sumber: Presentasi Direktorat Analisis dan Pengembangan Statistik – BPS, 22 Sept 2021

MPD: Foreign Tourists: Monthly Visitors





Sumber: Presentasi Direktorat Analisis dan Pengembangan Statistik – BPS, 22 Sept 2021

Web Crawler

- Web crawler, dikenal juga dengan web spider atau web robot, adalah program yang bekerja dengan metode tertentu dan secara otomatis mengumpulkan semua informasi yang ada dalam suatu website.
- Web crawler akan mengunjungi setiap alamat website yang diberikan kepadanya, kemudian mengorek, mengambil, dan menyimpan semua informasi yang terdapat di dalam website tersebut.
- Setiap kali web crawler mengunjungi sebuah website, maka dia juga akan mendata semua link/URL yang ada di website tersebut untuk kemudian dikunjungi lagi satu persatu.
- Dikutip dari Totally Tech, *web crawling* adalah proses di mana *search engine* menemukan konten yang di-*update* di sebuah situs atau halaman baru, perubahan situs, atau *link* yang mati.
- Menurut Moz, *web crawling* adalah proses di mana mesin pencari mengirimkan tim robot (*crawler* atau *spider*) untuk menemukan konten-konten baru dan konten yang telah di-*update.*

Web Crawler

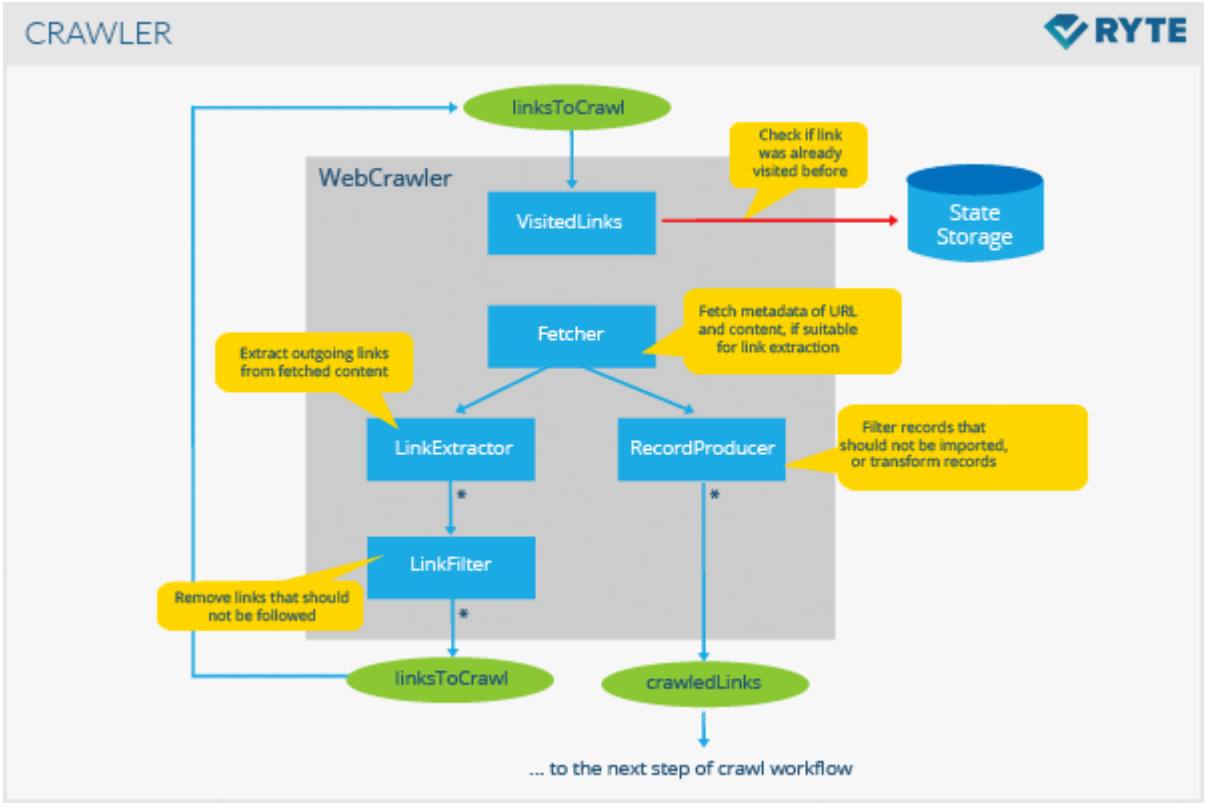
- Setiap *search engine* mempunyai web crawler sendiri, sehingga pencarian dengan keyword yang sama di *search engine* yang berbeda akan menghasilkan hasil yang berbeda
- Contoh web crawler
  - Googlebot dari google
  - Bingbot dari Bing
  - Slurp Bot dari Yahoo
  - DuckDuckBot dari DuckDuckGO
  - Baiduspider dari Baidu (mesin pencari dari China)
  - Yandex Bot dari Yandex (mesin pencari dari Rusia)
  - Sogou Spider dari Sogou (mesin pencari dari China)
  - Exabot dari Exalead
  - Alexa Crawler dari Amazon

Web Crawler

IPB University
Bogor Indonesia

Cara Kerja Web Crawler

1. Proses crawling dimulai dari daftar link halaman yang sudah dikenal oleh crawler dari sitemap suatu website.
2. Crawler mengunjungi masing-masing link pada suatu website untuk memperoleh berbagai macam URL baru.
3. URL tersebut kemudian dikumpulkan dan dimasukkan dalam indeks *search engine* yang disimpan dalam suatu database. Setiap kali ada perubahan pada website, maka indeks akan terupdate secara otomatis.
4. Proses 1-3 akan berlangsung secara terus menerus hingga tanpa batas



Web Crawler

Fungsi Web Crawler

Pada prinsipnya, *web crawler* berfungsi untuk merayapi dan mengindeks seluruh halaman atau konten yang ada di internet. Fungsi lainnya dari web crawler:

1. Melihat data perbandingan harga

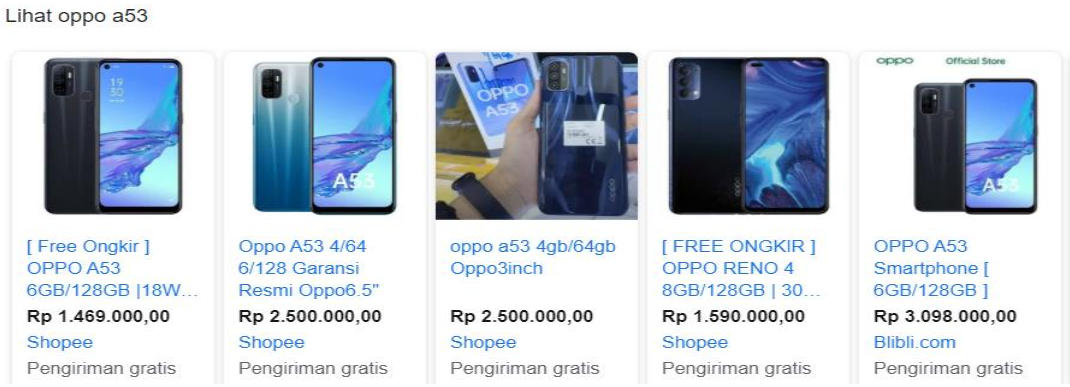
Web crawler bisa membandingkan harga dari suatu produk di internet. Jadi jika dilakukan pencarian suatu produk di internet, harga dari produk tersebut akan langsung muncul tanpa perlu masuk ke website penjualnya.

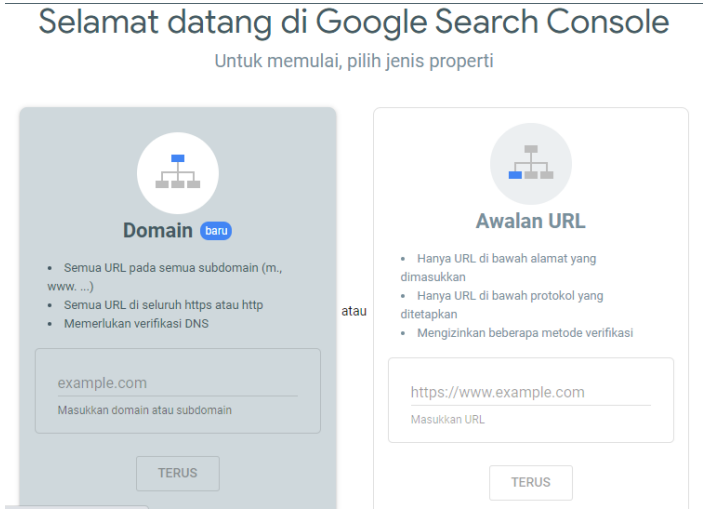
2. Menunjang *Web Analysis Tool*

Google Search Console adalah salah satu *web analysis tool* milik Google. *Tools* ini bisa membantu untuk menganalisis suatu web untuk mengetahui *page view*, *backlink*, *internal link*, dll. Tools ini menggunakan web crawler dalam mengumpulkan data-data tersebut.

3. Menunjang data mining

Web crawler berguna untuk mengumpulkan set data dari sumber terbuka di internet, seperti data alamat email dan nomor telepon sejumlah perusahaan yang terbuka untuk umum.





Web Scraping

- Web scraping adalah proses pengambilan data dari suatu website
- Web scraper adalah program (perangkat lunak) yang secara otomatis dapat mengekstrak data-data yang diperlukan dari satu atau lebih laman web (*web pages*) yang dijadikan target, kemudian menyimpan data-data tersebut kedalam suatu sistem database untuk dimanipulasi sesuai kebutuhan.

Terdapat 2 metode web scraping :

1. Manual : metode dengan menyalin data dengan cara copy-paste dari sebuah website
2. Otomatis : metode yang menggunakan koding, aplikasi, atau extension browser

Cara Kerja Web Scraper

1. Masukkan satu atau lebih URL untuk dilakukan scraping
2. Scraper akan memuat seluruh kode HTML untuk halaman tersebut. Web Scraper yang lebih canggih akan memberikan lebih banyak data termasuk elemen CSS dan Javascript.
3. Setelah itu, scraper akan mulai mengekstrak data pada halaman atau data tertentu yang dipilih pengguna untuk dijalankan. Jadi pengguna harus memilih data spesifik yang ingin diperoleh dari suatu halaman.

   Contoh : data produk dan harga dari suatu halaman *e-commerce*.
4. Web Scraper akan mengekstraksi semua data yang sudah dikumpulkan ke dalam format yang lebih mudah dipahami oleh pengguna, seperti format csv atau excel.

Web Scraping

IPB University
Bogor Indonesia

Teknik-teknik Web Scraping

1. Menyalin data secara manual
2. Menggunakan regular expression
3. Parsing HTML
4. Menganalisa DOM
5. Menggunakan Xpath
6. Menggunakan Google Sheet

Web Scraping

1. Menyalin Data secara Manual

- Merupakan teknik web scraping yang paling sederhana.
- Teknik ini memakan waktu lama karena informasi yang diinginkan harus diambil dan disimpan satu persatu secara manual.
- Merupakan metode yang paling efektif dari segi pencarian data, karena letak informasi yang ingin disalin dari suatu website sudah diketahui.
- Dianjurkan untuk digunakan jika jumlah website yang ingin disaring terbatas.

2. Menggunakan Regular Expression

- Regular expression adalah baris kode yang digunakan dalam algoritma pencarian untuk menemukan tipe data tertentu dari sebuah file. Dalam web scraping, file yang dimaksud adalah file-file penunjang sebuah website.
- Keuntungan menggunakan regular expression untuk web scraping adalah konsistensi syntaxnya dalam berbagai bahasa pemrograman.
- Keuntungan lainnya, regular expression dapat digunakan untuk mencari data berdasarkan jenisnya, seperti nama produk, harga, dan alamat email

Web Scraping

IPB University
Bogor Indonesia

3. Parsing HTML

- Parsing HTML adalah metode yang dilakukan dengan mengirimkan HTTP request kepada server yang menyimpan data website yang datanya ingin diekstrak.
- Keuntungan : Web scraping bisa dilakukan pada website statis dan website dinamis; memungkinkan untuk menyalin data dalam jumlah yang besar dalam waktu singkat.
- Kelemahan : Parsing HTML dapat dicegah dengan proteksi website; Jika teknik ini terlalu sering digunakan pada suatu situs/website, maka pengguna bisa diblokir dari situs tersebut.

4. Menganalisa DOM (Document Object Model)

- DOM atau *document object model* adalah representasi struktur sebuah halaman website yang ditulis dengan HTML
- Ketika melakukan scraping dengan parsing HTML, DOM dari halaman yang akan diekstrak akan dimuat terlebih dahulu. DOM tersebut juga membawa data yang ada pada HTML.
- Analisa DOM bisa dijadikan alternatif untuk melakukan web scraping terhadap website dinamis jika parsing HTML tidak membuahkan hasil

Web Scraping

IPB University
Bogor Indonesia

5. Menggunakan Xpath

- XPath adalah bahasa query yang digunakan untuk memilih node dari struktur file XML dan HTML
- Xpath digunakan sebagai salah satu teknik web scraping karena XPath bisa digunakan untuk mencari data pada elemen teks dalam file XML dan HTML

6. Menggunakan Google Sheet

- untuk melakukan web scraping dengan google sheet, diperlukan browser yang memiliki fitur inspect element.
- untuk melakukan web scraping, copy expression XPath dari elemen halaman website yang datanya ingin diekstrak ke dalam command IMPORTXML yang ada di Google Sheet

Tugas

- Eksplorasi bagaimana cara mendapatkan data dari beberapa sumber berikut, beserta :
  - Google Analytics
  - Google Trends
  - Google Traffic
  - Facebook Mobility
- Lakukan secara berkelompok, hasilnya dipresentasikan pekan depan

Tugas

- 16 kelompok dibagi merata untuk membahas keempat platform.
- Selain platform wajib per kelompok, diperkenankan jika ingin membahas selain keempat platform tersebut, akan ada apresiasi tambahan bagi kelompok yang melakuannya.
- 4 kelompok (perwakilan setiap platform) yang dipilih secara acak mempresentasi hasil penelusurannya dalam waktu 10-15 menit.
- Selain dipresentasikan, hasil penelusuran juga dikumpulkan dalam bentuk softcopy via newlms. Tidak ada ketentuan format hasil penelusuran yang dikumpulkan, silakan menggunakan kreativitas masing-masing kelompok.
- Pengumpulan cukup dilakukan oleh salah satu anggota kelompok paling lambat hari Rabu, 17 November 2021