Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

PERTEMUAN KE: 1

PENGANTAR PENGUMPULAN DATA (CRAWLING)

KULIAH TEORI

OLEH: Nurjoko

Capaian Pembelajaran

1. Memahami konsep dasar dan definisi pengumpulan data (crawling)
2. Mampu menjelaskan tujuan utama dari proses crawling dalam konteks pengolahan data.
3. Memahami konsep dasar dari data crawling sebagai proses pengumpulan informasi dari berbagai sumber data online.
4. Mampu menjelaskan peran dan pentingnya data crawling dalam pengolahan data
5. Menjelaskan manfaat penggunaan data crawling dalam konteks analisis data dan pengambilan keputusan.

Pengenalan *Data Crawling* (1)

- Big Data adalah terminologi pada komputer yang merujuk kepada pengolahan data yang memiliki ukuran yang besar.
- Salah satu sumber yang dapat menyebabkan ukuran data yang besar adalah *data crawling*.
- *Data crawling* adalah teknik untuk mencari informasi dari berbagai sumber data. *Data crawling* akan melacak informasi ke setiap level sumber data yang dapat diakses.

Pengenalan *Data Crawling* (2)

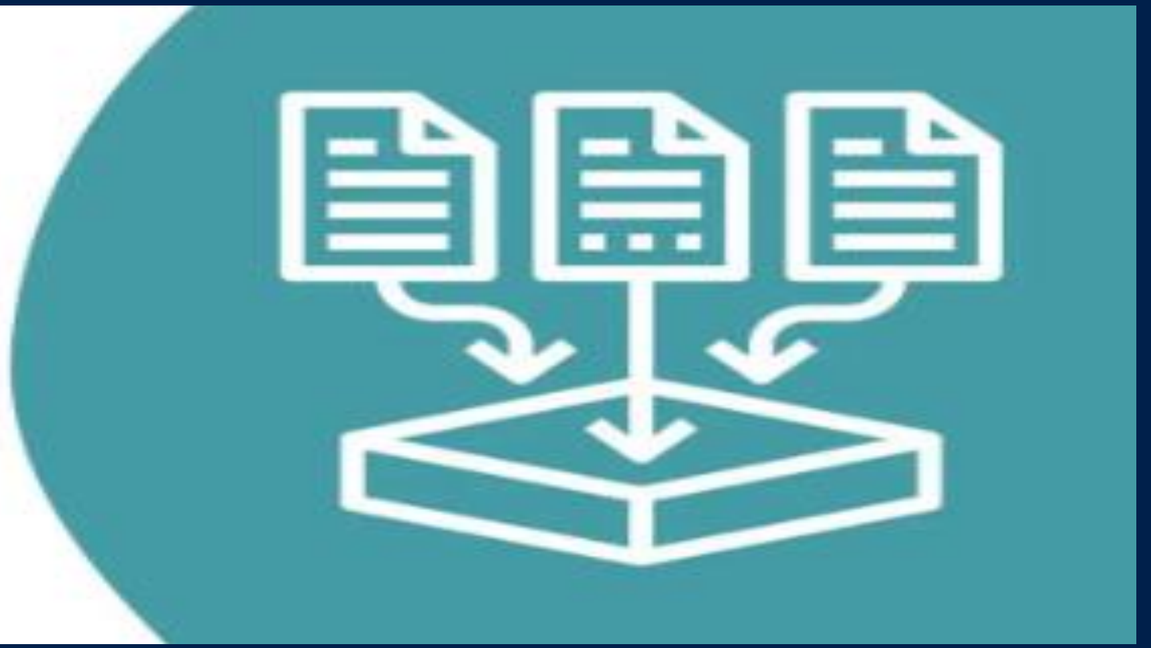
- Sumber data crawling yang paling umum adalah "website".
  - *Data crawling* yang bersumber dari website biasa disebut *web crawling*.
- *Data crawling* sangat berguna bagi perusahaan dalam mencari berbagai informasi yang tersebar di berbagai website.
- Terutama untuk perusahaan *search engine* yang menggunakan *web crawling* untuk membuat index web dari sebuah mesin pencarian di internet.

Apa Itu Crawling Data?

Pengertian, Proses dan Caranya

Definisi Crawling Data

Merupakan sebuah proses yang dilakukan oleh sebuah program komputer untuk mengumpulkan data dari berbagai sumber



Data yang sudah dikumpulkan nantinya bisa dijadikan sebagai alat untuk pengembangan sebuah sistem atau hanya sebagai data penelitian. Proses crawling dimulai dengan crawler yang mengunjungi sebuah web tertentu dan kemudian mengikuti tautan yang ada di halaman tersebut. Data yang ditemukan ini akan disimpan dalam basis data yang biasanya dibutuhkan mesin pencari, sehingga bisa menampilkan hasil pencarian yang lebih relevan.

Dampak Melakukan Crawling Data

Dampak Melakukan Crawling Data

- Meningkatkan Kualitas Produk dan Layanan
- Mengembangkan Daya Saing
- Meningkatkan Efisiensi Operasional

1. Meningkatkan Kualitas Produk dan Layanan

Crawling data adalah langkah mendapatkan data yang lebih lengkap mengenai keinginan dari konsumen, sehingga perusahaan lebih memahaminya. Jadi perusahaan akan mampu meningkatkan kualitas produk dan juga layanan kepada konsumen dengan dasar data tersebut. Selain itu perusahaan juga bisa secara cepat merespons umpan balik dari konsumen dan memproduksi produk terbaik berdasarkan konsumen.

2. Mengembangkan Daya Saing

Perusahaan bisa memiliki daya saing yang lebih besar apabila melakukan crawling data ini dan mendapatkan data yang lebih luas. Perusahaan jadi bisa memahami lebih baik mengenai pesaing dan juga bisa beradaptasi dengan berbagai perubahan yang terjadi di lingkup bisnis. Data tersebut jelas memberi kemampuan bagi perusahaan untuk bisa bersaing dengan lebih baik dengan kompetitor.

3. Meningkatkan Efisiensi Operasional

Crawling data mampu memberikan kesempatan perusahaan mampu meningkatkan efisiensi operasional yang menjadi kunci kesuksesan. Mulai dari manajemen logistik, memantau kinerja staf dan lain sebagainya bisa dilakukan dengan baik dan akhirnya operasional lebih efisien. Efisiensi ini akan berhubungan dengan efisiensi biaya dan juga produktivitas yang bisa semakin meningkat.

Proses Kerja Crawling Data

1. Memulai pada titik awal yakni web atau kumpulan URL yang sudah ditentukan sebelumnya dalam perencanaan crawling data. Crawling data adalah penjelajahan, dan mengunjungi web yang ditentukan itu menjadi titik awal.

2. Setelah itu crawler akan mengunjungi tautan yang ada di halaman tersebut dan hal tersebut dilakukan secara terus menerus. Langkah ini akan membuat crawler akan mengunjungi banyak situs untuk mengumpulkan data yang dibutuhkan.

3. Pada saat menjelajah crawler akan mengambil data yang diperlukan dan kemudian melakukan indeks dalam kumpulan data.

4. Crawler kemudian mengunjungi website yang sudah diindeks sebelumnya secara berkala untuk memastikan data yang diambil tetap yang terbaru.

5. Dalam prosesnya crawler tetap mengikuti etika dalam pengambilan data, caranya dengan mematuhi file robots.txt yang biasanya terdapat pada host website. Jadi crawler tahu mana halaman yang boleh dan tidak boleh diindeks.

Cara Melakukan Crawling Data

Bila ingin melakukan crawling data, maka perlu tahu dulu bagaimana cara crawling data yang baik dan benar. Berikut ini cara untuk melakukan crawling data.

1. Tentukan dulu sumber data yang nantinya akan menjadi target crawling

2. Manfaatkan software crawling untuk bisa mengumpulkan informasi dari sumber data yang ditentukan.

3. Konfigurasi software crawling data adalah langkah penting dalam crawling data untuk menentukan berapa jumlah halaman dan data yang diambil.

Contoh Penerapan Crawling Data

e-Commerce

e-Commerce melakukan penerapan crawling data untuk beberapa tujuan, berikut beberapa di antaranya.

- Menganalisis ulasan untuk memahami konsumen dengan jauh lebih baik.
- Memantau harga pesaing secara real time untuk menyesuaikan harga supaya tetap kompetitif.
- Memantau persediaan supaya tidak berlebihan namun juga tidak kurang.

Perbedaan dari Crawling Data dengan Web Scraping

Perbedaan dari Crawling Data dengan Web Scraping

Web scraping dengan crawling data memang memiliki karakteristik yang sama, akan tetapi sebenarnya berbeda. Berikut perbedaan crawling data dan web scraping.

1. Tujuan

Tujuan crawling data itu lebih luas yakni untuk menjelajahi semua web, memahami struktur dan mengumpulkan data yang relevan. Sementara web scraping hanya mengambil data tertentu.

2. Ruang Lingkup

Ruang lingkup crawling data lebih luas karena tidak hanya satu situs dan satu jenis data saja, sementara web scraping hanya web tertentu dan data tertentu saja.

3. Otomatisasi

Crawling data sudah pasti bisa berjalan otomatis, sementara web scraping membutuhkan program khusus untuk bisa otomatis.

Crawling data adalah proses penting dalam analisis data termasuk bagi perusahaan. Kalian bisa mempelajari crawling data di Coding Studio dengan mudah dan murah.

Data Scraping vs Data Crawling

Data Crawling	Data Scraping
Dilakukan dalam skala besar	Dapat dilakukan dalam skala yang tidak terlalu besar
Hanya membutuhkan *crawl agent*	Membutuhkan *crawl agent* dan *parser*
Melibatkan deduplikasi data	Tidak selalu melibatkan deduplikasi data
Merayapi data pada target tertentu, kemudian mengindeksnya	Hanya mengambil data yang dipilih, kemudian mengunduhnya

Data Crawling vs Web Crawling

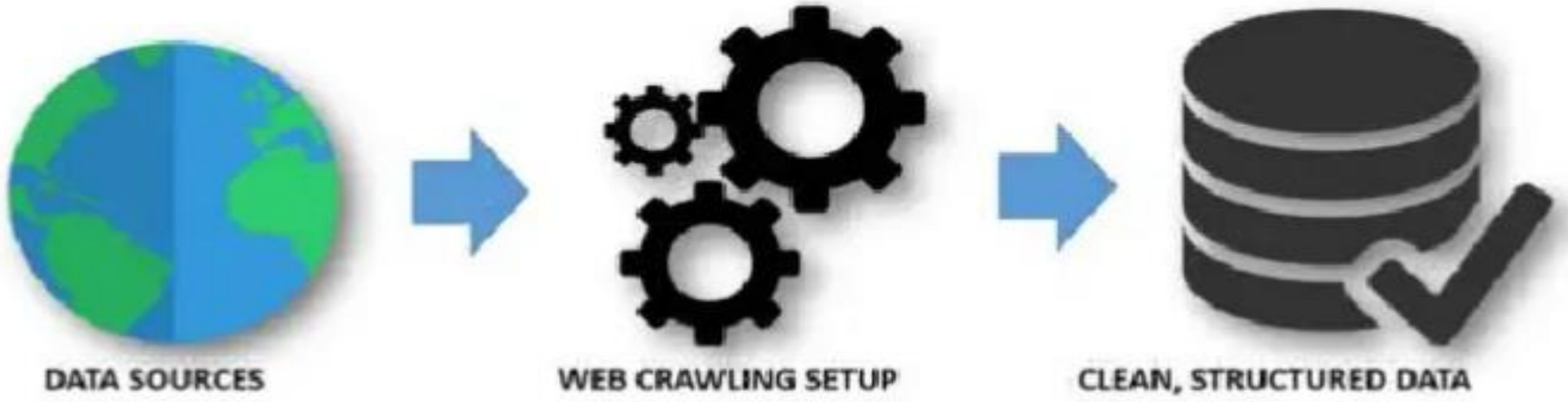
Data Crawling	Web Crawling
Memungkinkan pengambilan data dari seluruh sumber seperti *database*, *file*, atau API	Pengambilan data berfokus pada situs-situs web yang ada di internet
Bertujuan mengambil data untuk dianalisis demi kebutuhan pengembangan atau penelitian	Bertujuan mengambil data dari sebuah situs untuk memperbaharui mesin pencari

Konsep Dasar *Data Crawling* (1)

- *Data crawling* dilakukan menggunakan sebuah aplikasi *crawler* dengan konfigurasi tertentu.
- Aplikasi *crawler* adalah sebuah program yang dibuat untuk menjangkau semua sumber informasi dan melakukan aksi yang sudah ditentukan seperti memeriksa "kesegaran" dari sebuah informasi atau mengambil data dari sebuah sumber informasi.

Konsep Dasar *Data Crawling* (2)

- Langkah-langkah dasar dari sebuah proses *data crawling* adalah seperti terlihat pada gambar berikut:



Konsep Dasar *Data Crawling* (3)

- Langkah-langkah *Data Crawling*:
  - Menentukan sumber-sumber informasi
    - Langkah pertama adalah menentukan atau membuat daftar sumber-sumber informasi. Misalnya untuk *web crawling*, membuat daftar-daftar URL website yang akan diambil informasinya. Daftar website harus kredibel dan hindari website yang tidak mengijinkan *automated crawling* pada konfigurasi robots.txt atau di halaman TOS.
  - Mengkonfigurasi aplikasi *crawler*
    - Langkah kedua membutuhkan kemampuan teknis khususnya kemampuan pemrograman untuk memahami struktur data pada sumber informasi dan mengidentifikasi poin-poin yang bisa diambil dari sumber informasi tersebut sesuai dengan tugas yang sudah ditentukan.

Konsep Dasar *Data Crawling* (4)

- Melakukan *cleansing* dan menghilangkan duplikasi data
  - Data awal hasil *crawler* umumnya penuh dengan data-data anomali dan mengandung duplikasi informasi. Kondisi ini dapat mempengaruhi akurasi dari proses dan analisa data. Karena itu, langkah ini menjadi penting untuk membersihkan data dari data-data anomali serta data yang terduplikasi.
- Restrukturisasi data
  - Data yang didapat dari hasil *cleansing* dan penghilangan duplikasi, perlu diubah struktur-nya ke dalam skema yang dipahami oleh komputer. Dengan data yang terstruktur, akan mempermudah pemrosesan dan analisa lebih lanjut.

Learning Objective n

Fill in .....

CONCLUSION

Fill in .....

REFERENCES

[1] digitalent.kominfo.go.id

[2] Proses Kerja Crawling Data (https://codingstudio.id/blog/crawling-data-adalah/#)

[3] https://cmlabs.co/id-id/seo-guidelines/data-crawling#apa-itu-data-crawling

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”

[12-09-2023] | [KODE MK: SSD23406 | MATAKULIAH: CRAWLING AND DATA PREPROCESSING | SKS: 2/2 | VERSI: 01