Grow With EDM

Data Fundamentals & Web Scraping

Concepts, Tools & Application

GWE #1

Apa Itu Data?

GWE #1

Enterprise Data Management Laboratory

Apa itu Data dan Informasi?

- Data: Fakta atau angka mentah yang belum diproses atau dianalisis. Contoh: 100, Jakarta, 25 Januari 2025.
- Informasi: Data yang telah diproses, dianalisis, dan diorganisir sehingga memiliki makna dan dapat digunakan untuk pengambilan keputusan. Contoh: Penjualan 100 produk di Jakarta pada 25 Januari 2025 menunjukkan peningkatan dibandingkan bulan sebelumnya.

GWE #1

Enterprise Data Management Laboratory

Kategori Data

Data Terstruktur: Data yang diorganisasi dalam format tertentu, memudahkan pencarian, analisis, dan pemrosesan (misalnya, dalam basis data relasional).

Data Semi Terstruktur: Data dengan struktur parsial atau tidak sepenuhnya terorganisasi, seperti JSON atau XML, yang masih memerlukan pengolahan tambahan.

Data Tidak Terstruktur: Data tanpa format tertentu, seperti teks, gambar, dan video, membutuhkan pemrosesan tambahan untuk dianalisis.

GWE #1

Enterprise Data Management Laboratory

Tipe Data

Data Numerikal

Continuous Data (Data Kontinu): Data yang bisa memiliki nilai apa pun dalam rentang tertentu, termasuk desimal. Contoh: tinggi badan, berat badan, suhu.

Discrete Data (Data Diskrit): Data yang hanya dapat memiliki nilai-nilai tertentu dan terbatas. Biasanya data ini berupa bilangan bulat. Contoh: jumlah anak, jumlah mobil.

Data Kategorikal

Nominal Data: Data yang tidak memiliki urutan atau hierarki antar kategori. Contoh: warna, jenis kelamin, jenis makanan.

Ordinal Data: Data yang memiliki urutan atau tingkatan, tetapi perbedaan antar nilai tidak selalu terukur. Contoh: level pendidikan (SD, SMP, SMA), skala kepuasan (sangat tidak puas, tidak puas, puas, sangat puas).

GWE #1

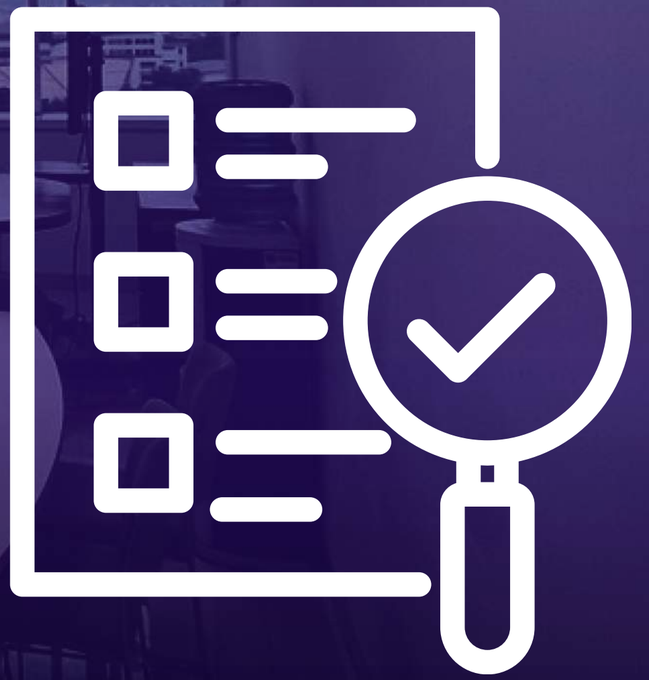
Mengapa data itu penting?

GWE #1

Enterprise Data Management Laboratory

Pentingnya Data

- Data membantu dalam membuat keputusan yang lebih baik.
- Data membantu memecahkan masalah dengan menemukan alasan di balik kinerja yang buruk.
- Data membantu seseorang mengevaluasi kinerja.
- Data membantu seseorang meningkatkan proses.
- Data membantu seseorang memahami konsumen dan pasar.

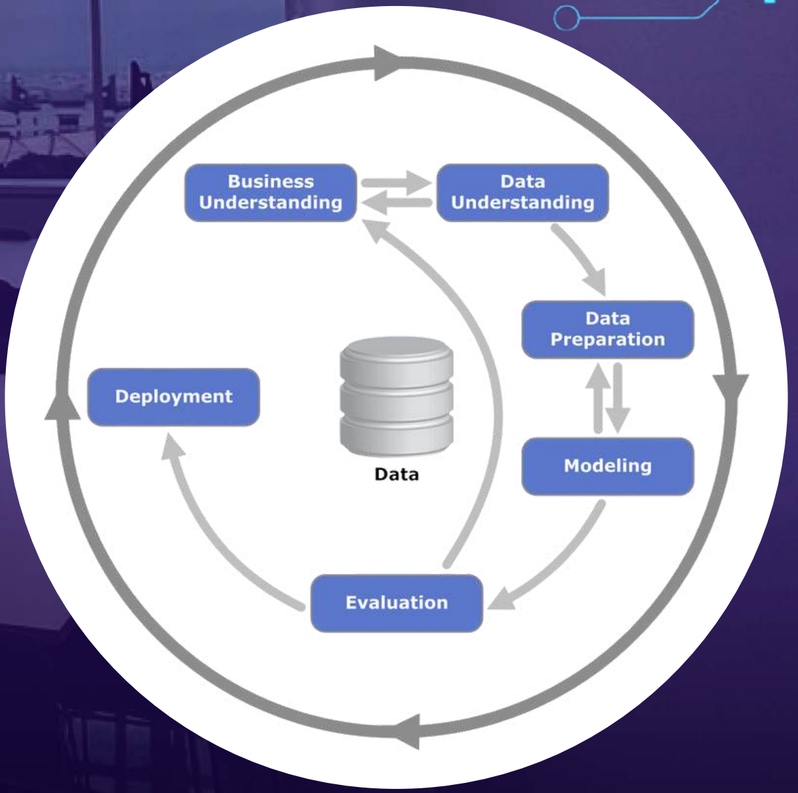


GWE #1

Enterprise Data Management Laboratory

Data Life Cycle

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



GWE #1

Enterprise Data Management Laboratory

Tools/Library yang biasa digunakan

GWE #1

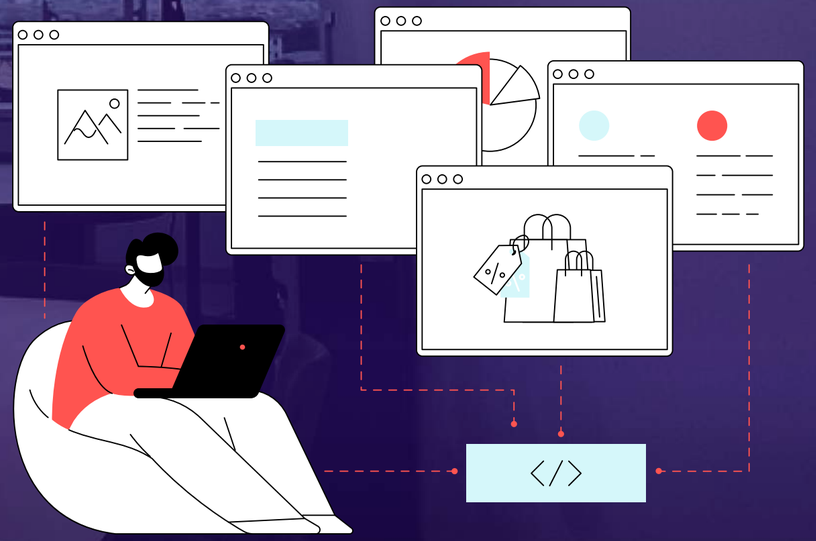
Apa Itu Web Scraping?

GWE #1

Enterprise Data Management Laboratory

Pengertian Web Scraping

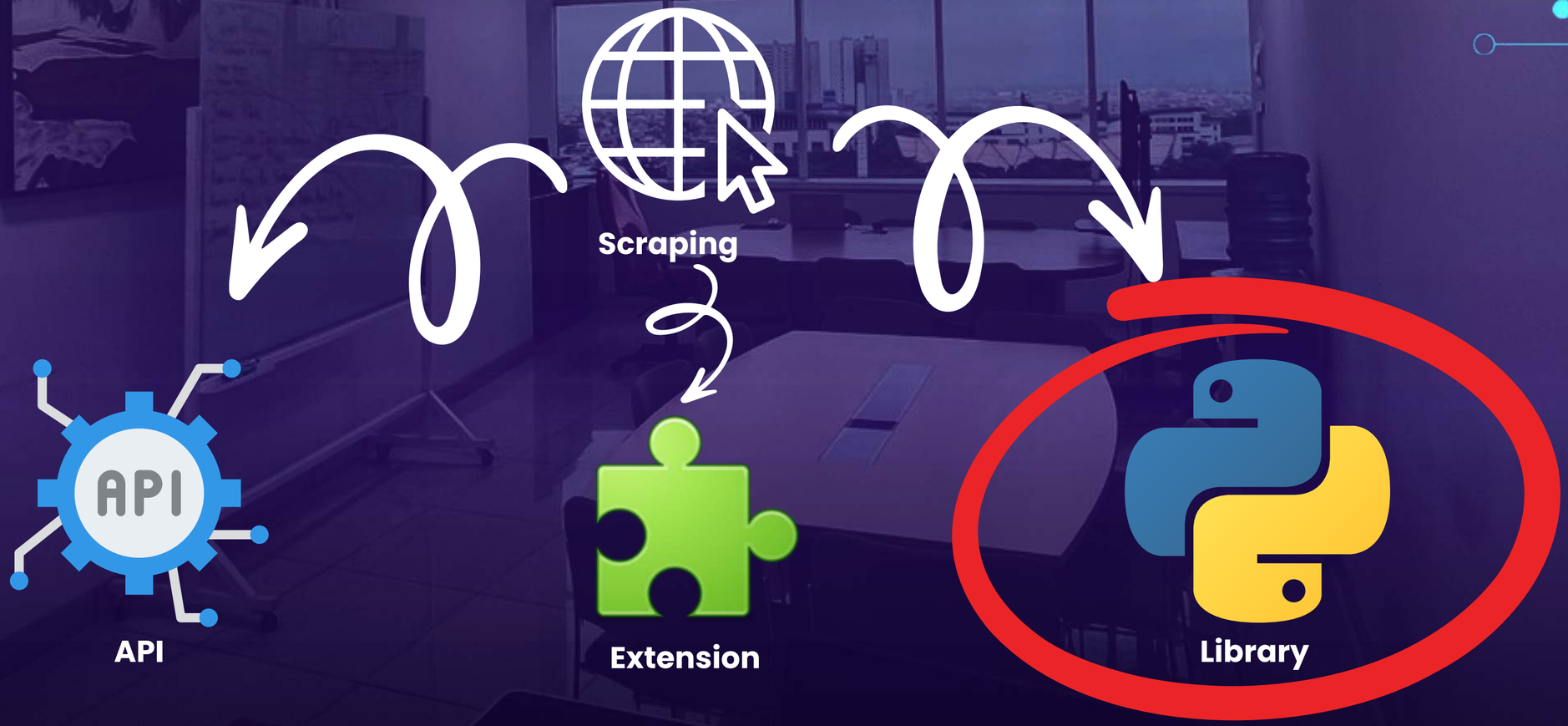
Web Scraping adalah proses buat kita mengoleksi data dari sumber-sumber data buat kita simpan dan olah lebih lanjut lagi. Nah biasanya karena data sifatnya kredensial kita sebagai user awam akan memanfaatkan data2 dari internet / website nih maka sering dinamakan sebagai Web Scraping.



GWE #1

Enterprise Data Management Laboratory

Cara Web Scraping

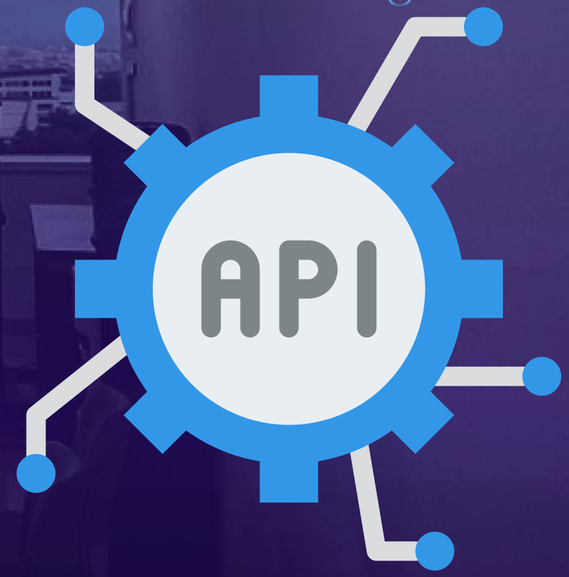


GWE #1

Enterprise Data Management Laboratory

Cara Web Scraping #1

API (Application Programming Interface) adalah lapisan yang memungkinkan sistem saling berkomunikasi. Dalam komunikasi, unsur pentingnya adalah pertukaran data. Beberapa situs web menyediakan API untuk memungkinkan pertukaran data sesuai dengan protokol yang telah ditentukan.



Dimana data akan didapatkan melalui method GET dan akan kita miliki dalam berbagai format seperti JSON, XML, Protobuf dan sebagainya.

pros: mudah, tidak bermasalah jika tampilan web berganti
cons: harus tersedia dulu API nya, biasanya berbayar / based on kerjasama

*bisa dipelajari lebih lanjut di mata kuliah EAI(Semester 5).
Atau memalui Youtube Pak Faqih

GWE #1

Enterprise Data Management Laboratory

Cara Web Scraping #2

Extention pada browser seperti MS. Edge, Chrome, dll

Dapat menggunakan extention **webscraping.io** jika kalian ingin mencoba

pros: mudah, plug and play
cons: pasti ada keterbatasannya, example cuma bisa format .csv / per page

Extension

GWE #1

Enterprise Data Management Laboratory

Cara Web Scraping #3

*Ini yang akan kita pelajari

Taukan kalo py itu salah satu bahasa yang powerful. Karena tinggal import import kita bisa segalanya

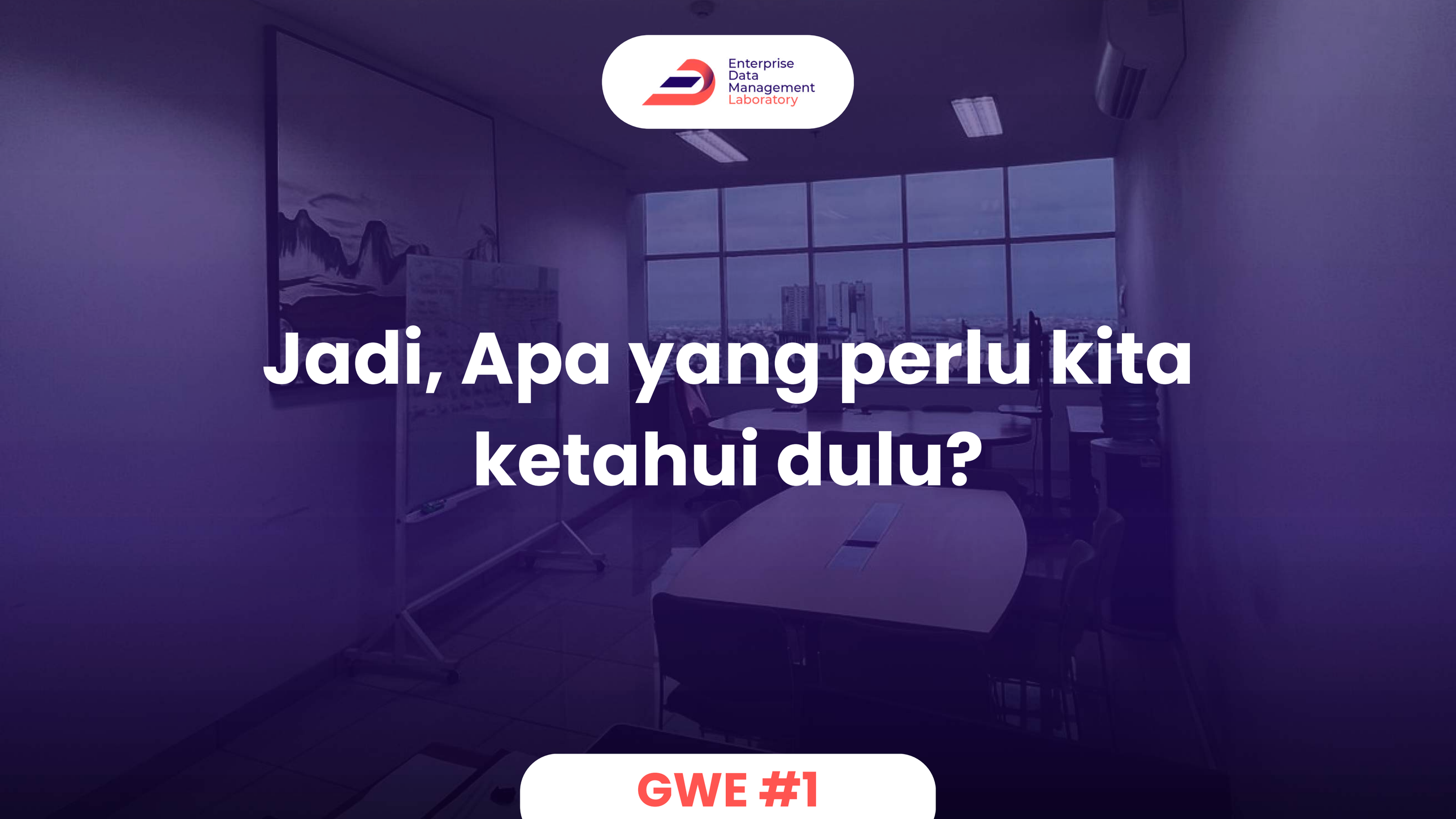
Akan menggunakan library BeautifulSoup & requests

Dengan cara parsing HTML (tampilan web) yang ada di website. Jadi kita bisa ambil data yang terlihat di web

pros: sangat fleksibel, mostly bisa web apapun
cons: complex, tampilan berubah bisa saja cara berubah

Library

GWE #1



Jadi, Apa yang perlu kita ketahui dulu?

GWE #1

Enterprise Data Management Laboratory

Yang perlu kita ketahui dulu #1

Python Library

Kumpulan kode kode yang udah di compile dan dapat digunakan kembali untuk tugas-tugas tujuannya.

Requests

untuk mendapatkan kode HTML (berinteraksi dengan web) pada pada program Python. Bisa juga handle authentication lhï

BeautifulSoup

Library untuk parsing struktur data terutama di HTML dan XML. Sehingga bisa mengetahui elemen2 yang ada di website.

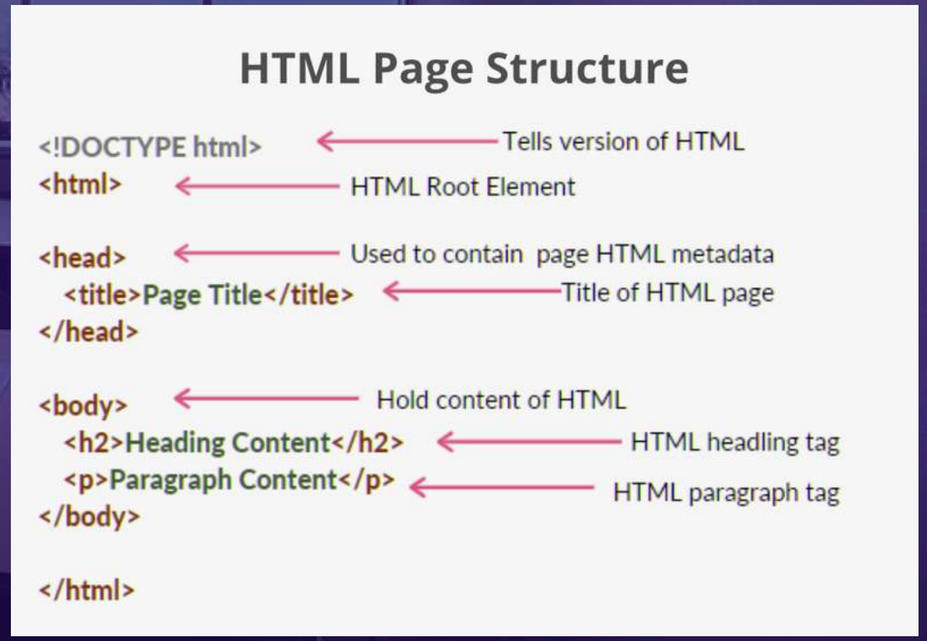
GWE #1

Enterprise Data Management Laboratory

Yang perlu kita ketahui dulu #2

Website Structure

Kita akan mengambil data berdasarkan struktur HTML nya, maka penting untuk tau ada apa saja sih elemen yang digunakan untuk menyimpan data di web. Serta elemen seperti class, href dan sebagainya.
Dengan BeautifulSoup kita akan mengambil isi data berdasarkan nama elemennya



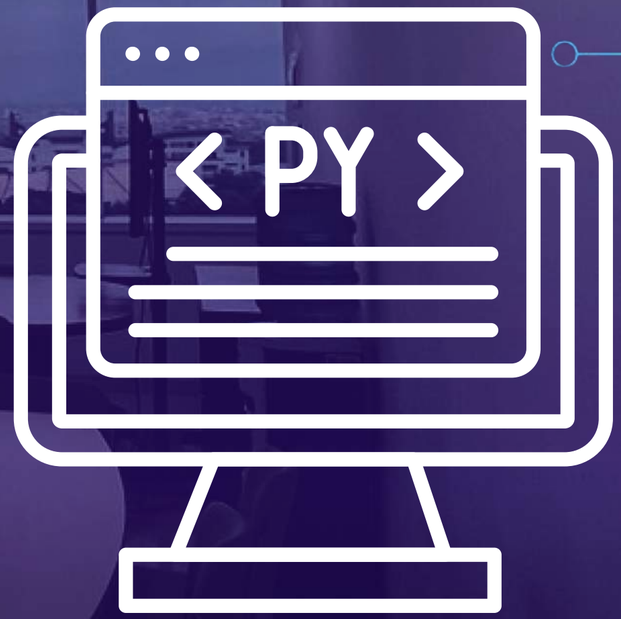
GWE #1

Enterprise Data Management Laboratory

Yang perlu kita ketahui dulu #3

Python Syntax

Percayalah cheat sheet nanti akan banyak kepakai. Terutama kita akan pakai untuk looping, format data, simpan data, data wrangling dan sebagainya.



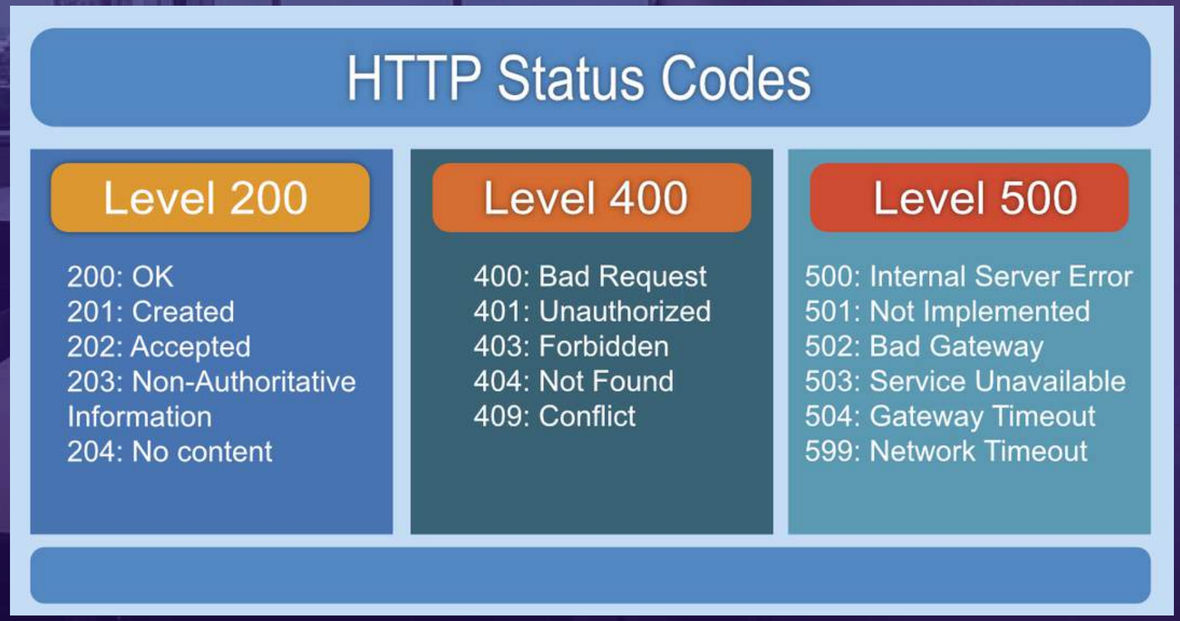
GWE #1

Enterprise Data Management Laboratory

Yang perlu kita ketahui dulu #4

HTTP Response Code

Kita dapat mengetahui apakah sebuah web dapat diakses atau tidak, serta apakah kita berhasil terhubung dengan servernya, melalui response code yang diterima. Response code ini memberikan informasi penting, seperti status keberhasilan, kesalahan, atau izin akses yang berguna untuk memastikan proses komunikasi dengan web berjalan dengan baik.



GWE #1

Enterprise Data Management Laboratory

Yang perlu kita ketahui dulu #5

Alur

Request

1. Menyambungkan ke web tujuan dengan method GIT
2. Mempelajari struktur html website
3. Mengkonfigurasi agar dapat mengambil data dari website

Parse

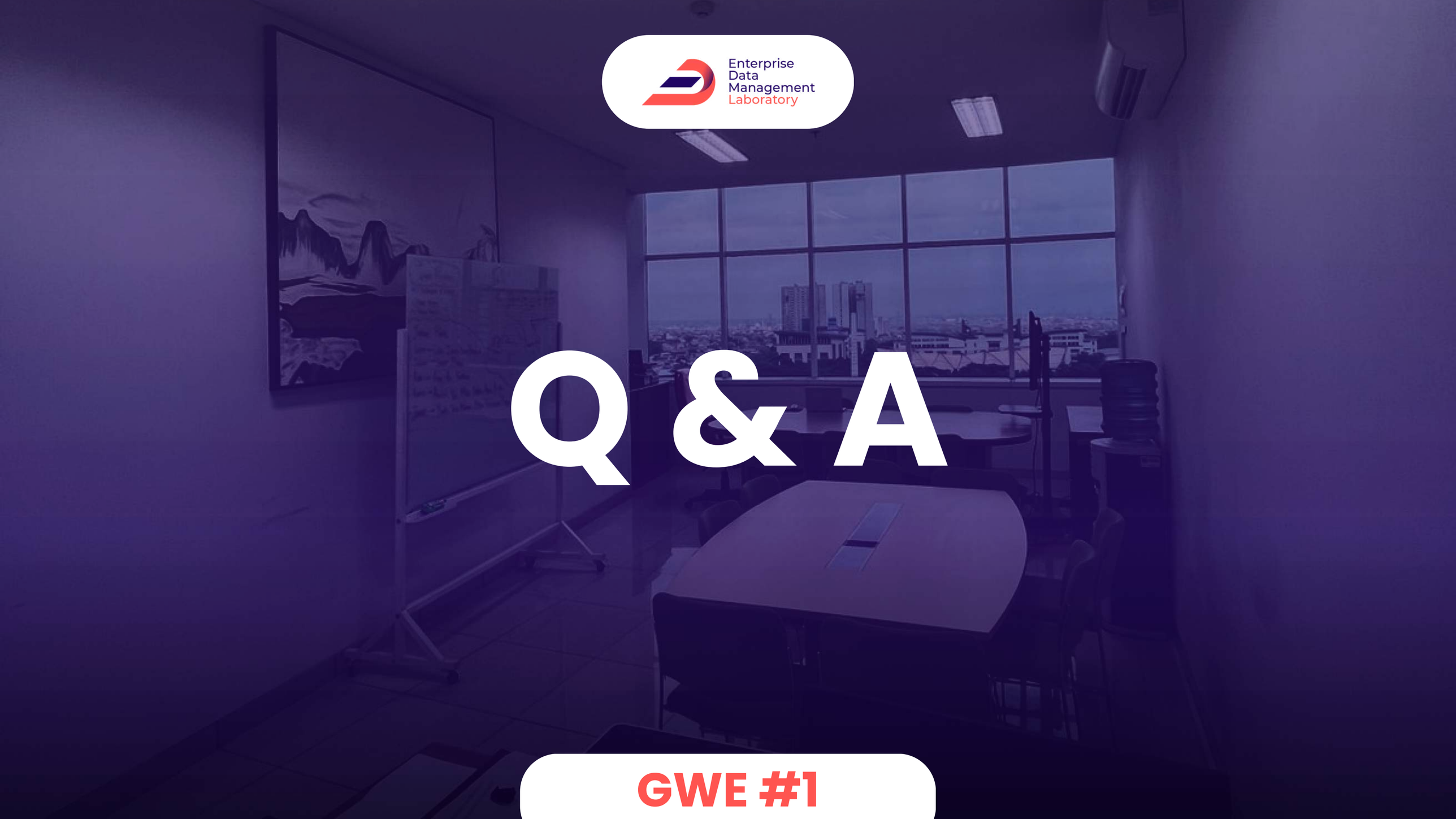
1. Inspect elemen HTML websitH
2. Memilih elemen yang akan kita ambil dan perlukan untuk tujuan data
3. Scrape text

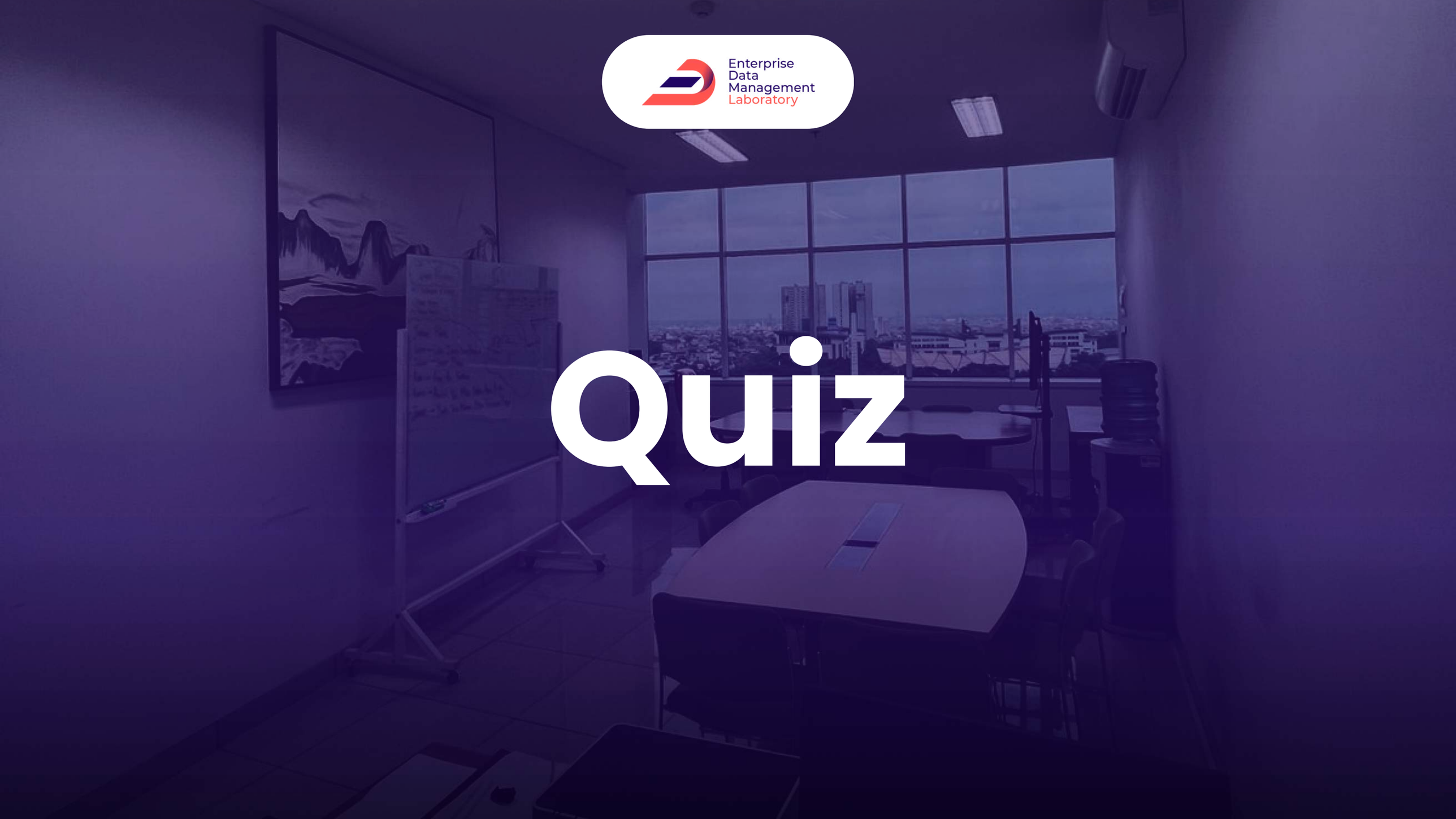
Store

1. Simpan data yang telah berhasil di scrape ke dalam file / database
2. Melakukan otomasi pengambilan data

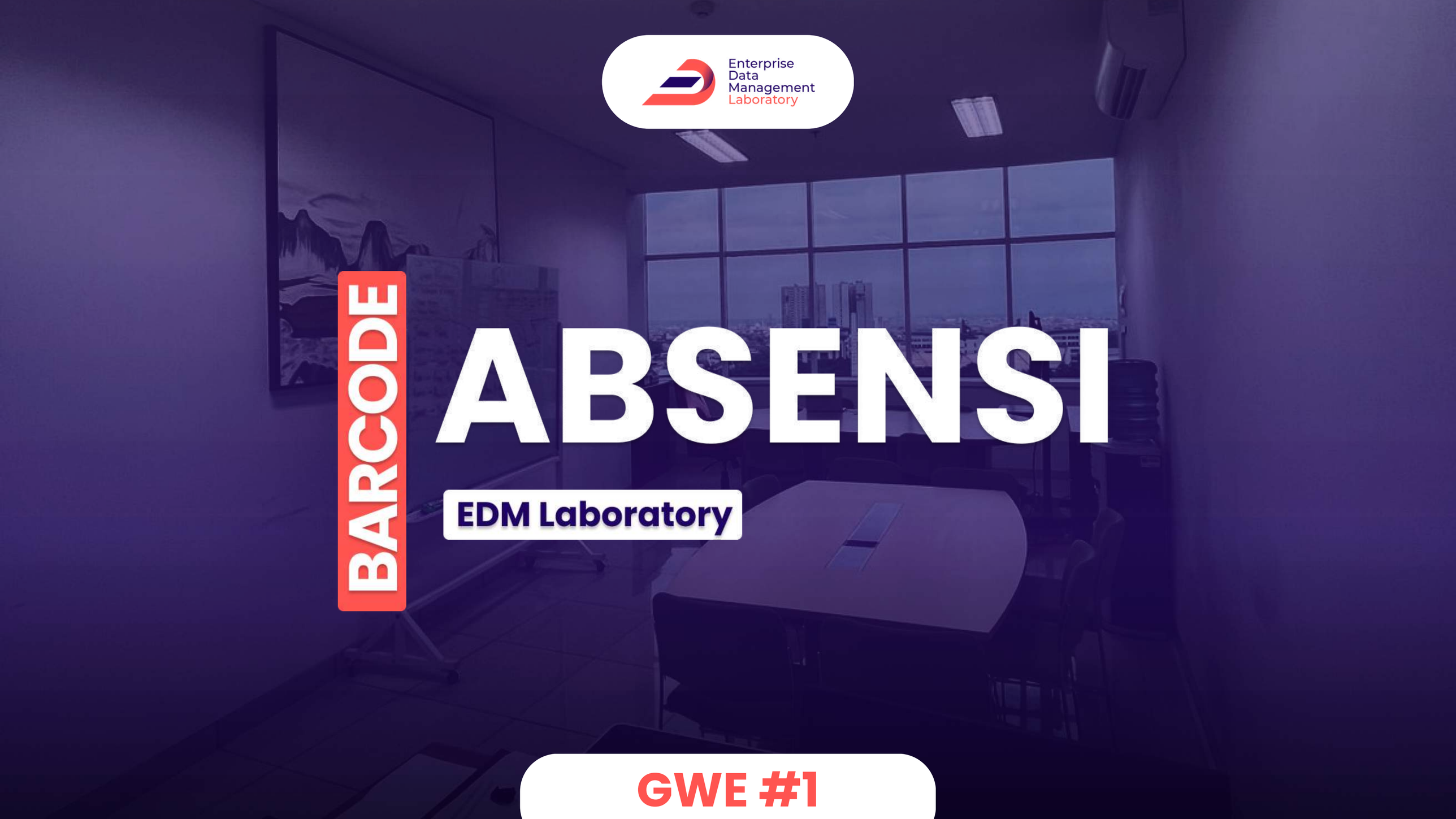
GWE #1


Enterprise
Data
Management
Laboratory
Let’s try


Enterprise
Data
Management
Laboratory
Q & A
GWE #1



Quiz


Enterprise
Data
Management
Laboratory
BARCODE
ABSENSI
EDM Laboratory
GWE #1

Enterprise Data Management Laboratory

Presensi GWE #1

https://tel-u.ac.id/presensigwe

GWE #1

Enterprise Data Management Laboratory

Terima Kasih

GWE #2

@lifeatedmlab @186mcgwc