

Khoá học:
PYTHON WEB SCRAPING
THU THẬP “MỌI” DỮ LIỆU TỪ
INTERNET



LEARN ANYTHING



VNSTOCK

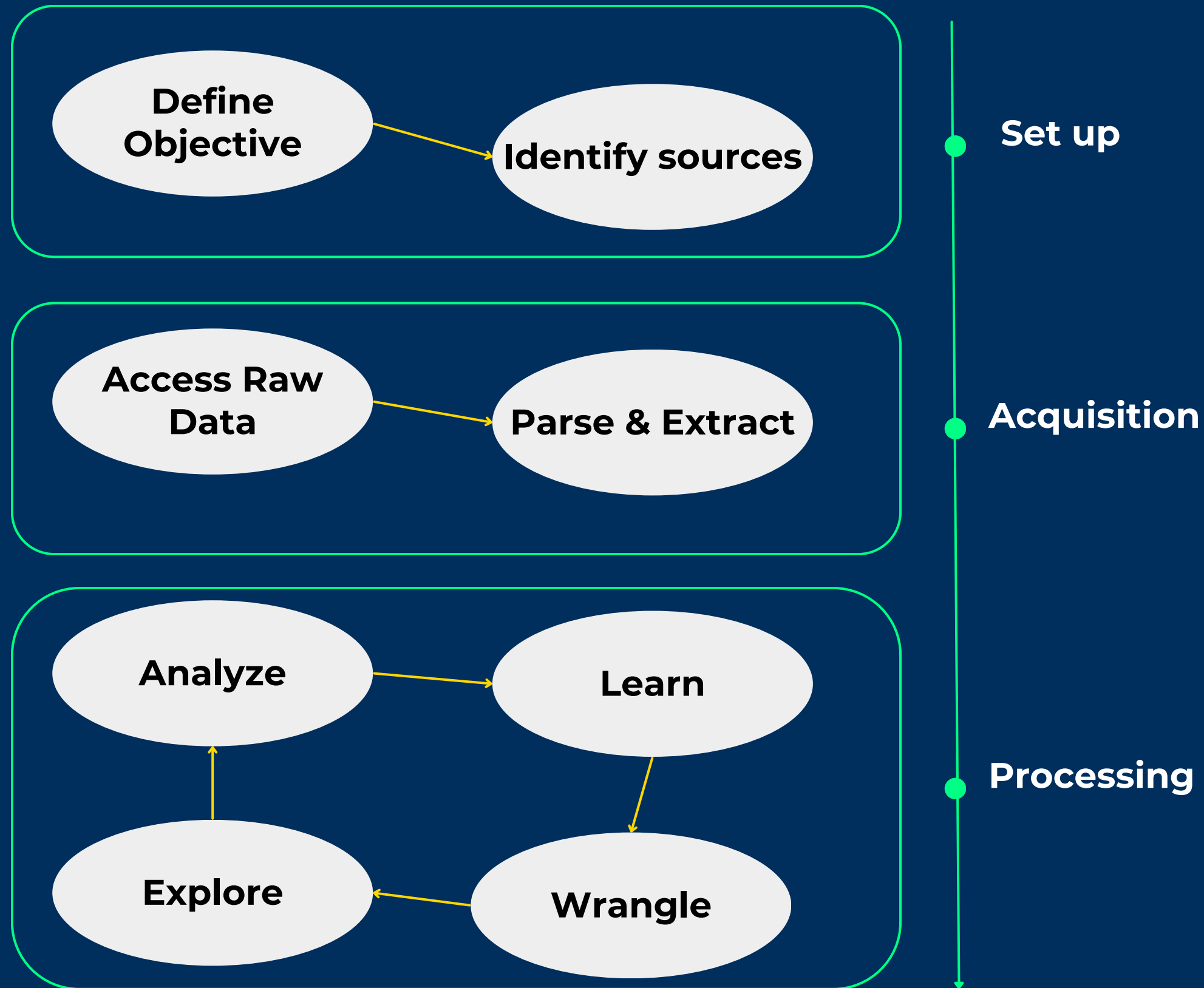


TỔNG QUAN

Giới thiệu kiến thức tổng quát về Web Scraping

WEB SCRAPING LÀ GÌ?

WEB SCRAPING LÀ GÌ?



- Là quá trình thu thập dữ liệu từ các trang web. Dữ liệu sau đó được lưu trữ, xử lý, phân tích và giúp cung cấp thông tin trong việc ra quyết định cụ thể.
- Chia thành 3 công đoạn chính: Thiết lập, Thu thập, Xử lý
- Python cung cấp bộ công cụ web scraping mạnh mẽ.

WEB SCRAPING CÓ HỢP PHÁP?

- Thu thập dữ liệu công khai là hợp pháp, ngoại trừ
 - Phục vụ cho mục đích xấu (đánh cắp sở hữu trí tuệ)
 - Gây ảnh hưởng xấu đến hoạt động (DDos - tấn công từ chối dịch vụ)
 - Thu thập/mua bán thông tin nhận dạng cá nhân mà không được sự cho phép
 - GDPR ở châu Âu
 - CCPA ở California (Hoa Kỳ)
 - Nghị định 13/2023/NĐ-CP bảo vệ dữ liệu cá nhân tại Việt Nam từ 1/7/2023
- Một số sự kiện nổi bật liên quan Web scraping
 - OpenAI sử dụng dữ liệu công khai để đào tạo mô hình ngôn ngữ lớn AI
 - Bê bối dữ liệu Facebook-Cambridge Analytica năm 2014 (tác động bầu cử Mỹ)
 - LinkedIn vs hiQ Labs năm 2019

PHẠM TRÙ ĐẠO ĐỨC CỦA WEB SCRAPING

- **Đừng phá hoại**
 - Ý thức rõ ảnh hưởng của tác vụ thu thập dữ liệu lên chủ thể và rủi ro pháp lý
- **Tôn trọng bản quyền**
 - Ý thức, cân nhắc việc thu thập các tài nguyên riêng tư được bảo vệ bản quyền (không phát tán hoặc lạm dụng)
- **“Better be safe than sorry”**
 - Hiểu rõ và bảo vệ chính mình tránh các rủi ro không đáng có.
- **“Be nice”**: xin phép & chia sẻ nếu có thể
 - Cân nhắc việc xin phép chủ sở hữu về chia sẻ thông tin khi thực hiện nghiên cứu khoa học.

WEB SCRAPING ĐỂ LÀM GÌ?

LÝ DO CẦN WEB SCRAPING

1. Tạo doanh thu

- Dữ liệu là sản phẩm kinh doanh (dịch vụ dữ liệu)
- Insights là sản phẩm kinh doanh (dịch vụ phân tích, nghiên cứu thị trường)
- Dữ liệu là đầu vào cho sản phẩm (AI)
- Dữ liệu phục vụ đầu tư

2. Giảm chi phí

- Tự xây dựng sản phẩm thay vì mua (build vs buy)

3. Tăng năng suất & hiệu quả

- Thực hiện công việc với số lượng lớn hàng loạt

4. Ra quyết định tự tin

- Tăng tỉ lệ thành công khi ra quyết định dựa trên dữ liệu

5. Tính cá nhân hóa cao

- Cá nhân hóa sản phẩm tùy thích

WEB SCRAPING THẾ NÀO?

ĐỐI TƯỢNG MỤC TIÊU CỦA WEB SCRAPING

- Văn bản
- Đối tượng
 - Bảng
 - File
 - Sách điện tử
- Đa phương tiện
 - Hình ảnh
 - Âm thanh
 - Video
- Tác vụ
 - Gửi yêu cầu (request GET)
 - Thay đổi thông tin (request POST)
- API chính thức
- Web công khai
- Web riêng tư

CÔNG CỤ WEB SCRAPING TRONG PYTHON

Trình duyệt web



Thư viện Python



Code/test generator



Công cụ AI



IDE / Code Editor



Công cụ phụ trợ khác

