Crawling Data

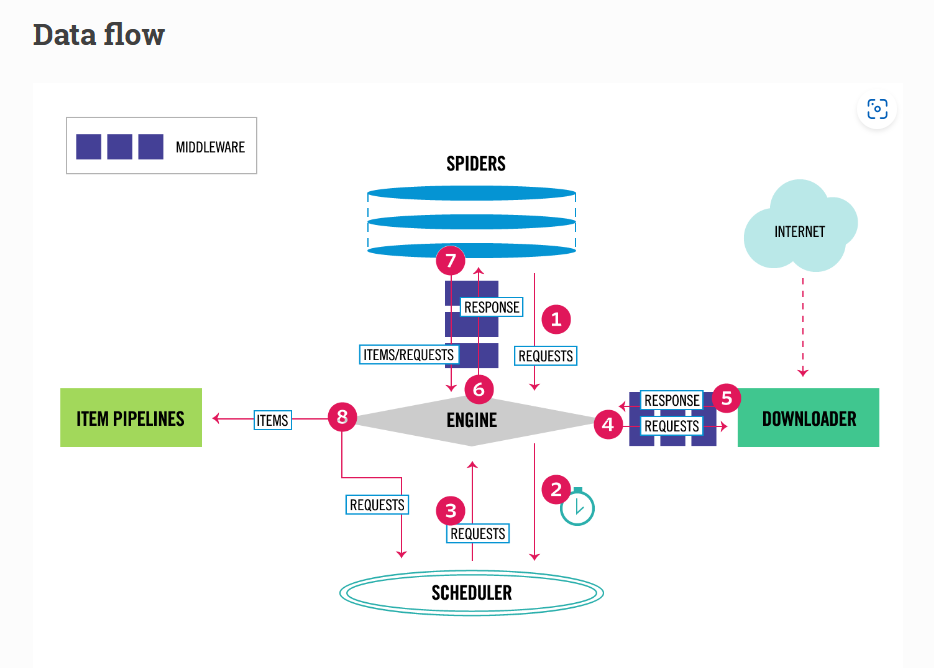
1. Đinh nghĩa, tổng quan về scrapy

Sử dụng scrapy để crawling data, với scrapy là một thư viện của python

**Cấu trúc dự án Scrapy**:

Scrapy project được tổ chức thành các thành phần chính bao gồm

* **Spiders:** Định nghĩa cách mà Scrapy thu thập dữ liệu từ các trang web cụ thể.
* **Items:** Định nghĩa cấu trúc của dữ liệu mà bạn muốn thu thập từ các trang web.
* **Pipelines:** Xử lý và lưu trữ dữ liệu sau khi nó được thu thập bởi spider.
* **Middlewares** Xử lý các yêu cầu và phản hồi trước hoặc sau khi chúng được gửi hoặc nhận bởi spider.



1. **Tạo Spider**: Bạn sẽ tạo ra một spider để định nghĩa cách thu thập dữ liệu từ một trang web cụ thể. Spider sẽ chứa các hướng dẫn cho Scrapy về cách tìm kiếm và trích xuất dữ liệu từ các trang web, bằng cách sử dụng CSS selectors hoặc XPath expressions.
2. **Định nghĩa Items**: Bạn sẽ định nghĩa các items để xác định cấu trúc của dữ liệu mà bạn muốn thu thập từ các trang web. Mỗi item sẽ bao gồm các trường mô tả thông tin cụ thể mà bạn muốn thu thập, chẳng hạn như tiêu đề, URL, nội dung, vv.
3. **Viết Pipelines**: Pipelines sẽ được sử dụng để xử lý và lưu trữ dữ liệu sau khi nó được thu thập bởi spider. Bạn có thể viết pipelines để thực hiện các tác vụ như lọc dữ liệu, kiểm tra tính hợp lệ hoặc lưu trữ dữ liệu vào cơ sở dữ liệu.
4. **Chạy và cấu hình Scrapy**: Bạn sẽ cấu hình và chạy Scrapy để bắt đầu quá trình thu thập dữ liệu. Bạn có thể cấu hình các tham số như tên miền, số lượng luồng, tốc độ crawl, và các thiết lập khác trong tệp cấu hình Scrapy.
5. **Xử lý dữ liệu thu thập**: Sau khi quá trình crawling hoàn tất, dữ liệu thu thập sẽ được xử lý bởi các pipelines và lưu trữ theo cách mà bạn đã định nghĩa.