BÁO CÁO

BÀI 7: CRAWLER DỮ LIỆU GIÁ SẢN PHẨM TỪ NHIỀU NGUỒN

1. **Giới thiệu:**

Microservice là một phương pháp thiết kế phần mềm, trong đó một ứng dụng được chia thành nhiều phần nhỏ, mỗi phần có thể được phát triển, triển khai và duy trì độc lập. Trong báo cáo này, chúng ta sẽ tập trung vào việc sử dụng microservice để crawl dữ liệu giá sản phẩm từ nhiều nguồn như Amazon và eBay.

1. **Khái niệm:**
2. Ngôn ngữ lập trình:

**JavaScript** là một ngôn ngữ lập trình thông dụng được sử dụng rộng rãi trên cả client-side và server-side, giúp xây dựng các ứng dụng web động và tương tác.

1. Backend:

* **Node.js** là môi trường thời gian chạy (runtime environment) JavaScript đa nền tảng và mã nguồn mở. NodeJS cho phép các lập trình viên tạo cả ứng dụng front-end và back-end bằng JavaScript.
* **Expressjs** là một framework được xây dựng trên nền tảng của Nodejs. Nó cung cấp các tính năng mạnh mẽ để phát triển web hoặc mobile. Expressjs hỗ trợ các method HTTP và midleware tạo ra API vô cùng mạnh mẽ và dễ sử dụng.

1. Frontend:

**React.js** là một thư viện JavaScript dùng để xây dựng giao diện người dùng, đặc biệt là các ứng dụng đơn trang (SPA). React.js cho phép tạo ra các component tái sử dụng và quản lý trạng thái ứng dụng hiệu quả.

1. Cơ sở dữ liệu:

**MongoDB** là một cơ sở dữ liệu NoSQL, sử dụng cấu trúc lưu trữ dạng JSON, cho phép linh hoạt trong việc lưu trữ và truy xuất dữ liệu.

1. Thư viện Crawler:

**Cheerio** là một thư viện JavaScript giúp xử lý và phân tích các tài liệu HTML bằng cách cung cấp API tương tự như jQuery. Cheerio giúp dễ dàng lấy dữ liệu từ các trang web và xử lý chúng theo yêu cầu.

1. Quản lý mã nguồn:

**Git**: Là một hệ thống quản lý phiên bản phân tán, cho phép theo dõi các thay đổi trong mã nguồn và cộng tác hiệu quả giữa các nhà phát triển.

**GitHub**: Là một nền tảng lưu trữ mã nguồn dựa trên Git, cung cấp các công cụ và tính năng giúp quản lý dự án và cộng tác dễ dàng.

1. **Kế hoạch thực hiện:**

Ngày 1 - 2: Lập kế hoạch, tìm hiểu các thư viện crawler và chọn thư viện phù hợp (Cheerio).

Ngày 3 – 5: Khởi tạo môi trường, tìm hiểu tài liệu về thư viện cheerio.

Ngày 6 – 9: Tạo và kết nối cơ sở dữ liệu (sử dụng mongoose để kết nối mongodb), viết chức năng crawler giá sản phẩm, kiểm thử chức năng

Ngày 10 – 12: Xây dựng giao diện nhập URL sản phẩm để thực hiện chức năng crawler dữ liệu giá.

Ngày 13 – 14: Hoàn thiện và viết báo cáo.

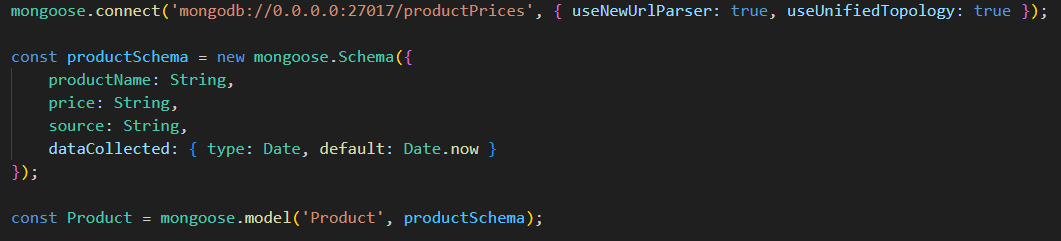
1. **Hiện thực microservice:**
2. Thiết lập môi trường:

Cài đặt môi trường và các thư viện cần thiết:

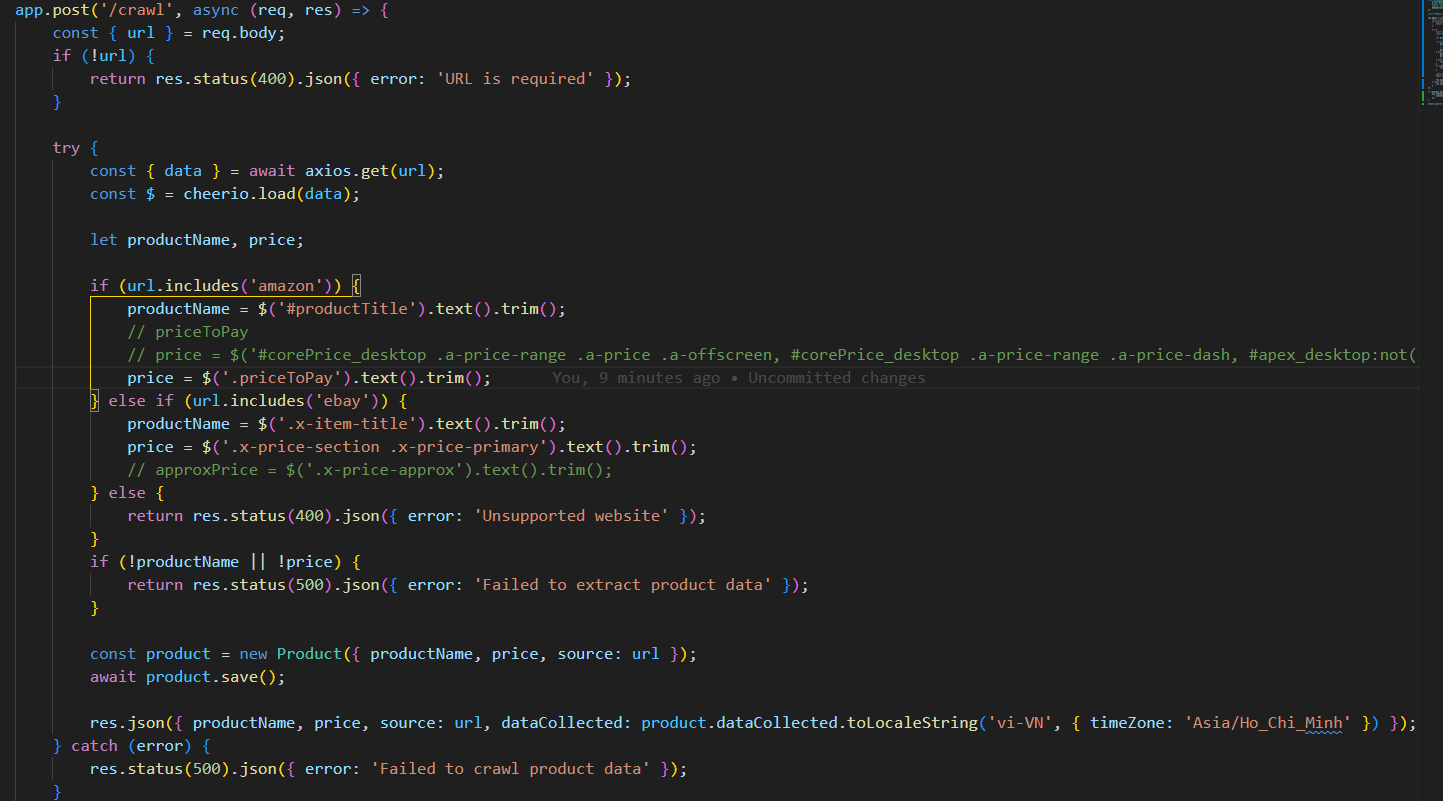
* Tạo server backend bằng Node.js và cài đặt các thư viện Express.js, cheerio, axios, mongoose.
* Tạo giao diện frontend bằng Reactjs.
* Tạo testcase bằng thư viện supertest.

1. Tạo và kết nối cơ sở dữ liệu:

* Kết nối cơ sở dữ liệu bằng mongoose và tạo mô hình cơ sở dữ liệu Product.



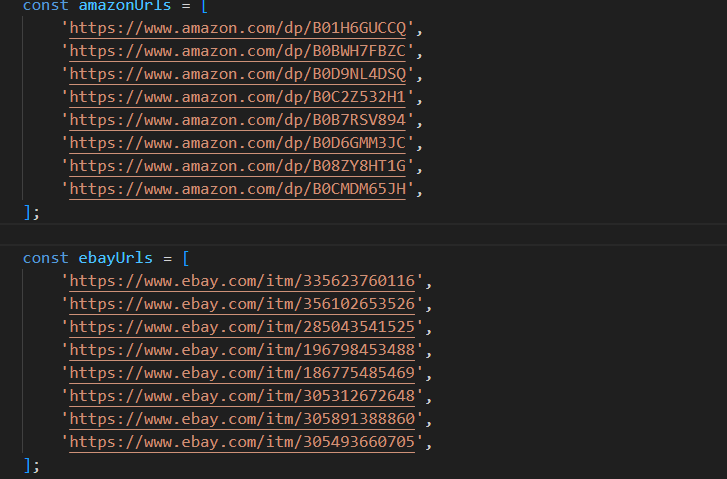
1. Viết chức năng crawler:



* Định nghĩa route POST cho /crawl
* Lấy URL sản phẩm từ body.
* Gửi yêu cầu HTTP GET để lấy dữ liệu HTML từ URL.
* Tải HTML vào Cheerio để xử lý và trích xuất thông tin.
* Trích xuất tên sản phẩm (#productTitle và .x-item-title) và giá (.priceToPay và .x-price-primary) từ HTML.
* Trả về JSON chứa thông tin sản phẩm.
* Khó khăn gặp phải: Xác định selector của giá sản phẩm.

1. Kiểm thử và hoàn thiện:

* Sử dụng thư viện supertest để kiểm thử một vài sản phẩm thuộc Amazon và Ebay.





* Khó khăn gặp phải: Tìm hiểu thêm thư viện supertest

1. **Kết luận:**

Việc sử dụng microservice để crawl dữ liệu từ Amazon và eBay là một phương pháp hiệu quả để thu thập và phân tích dữ liệu từ các trang web lớn. Bằng cách sử dụng các microservice, chúng ta có thể tối ưu hóa quá trình phát triển và quản lý dữ liệu một cách hiệu quả hơn.