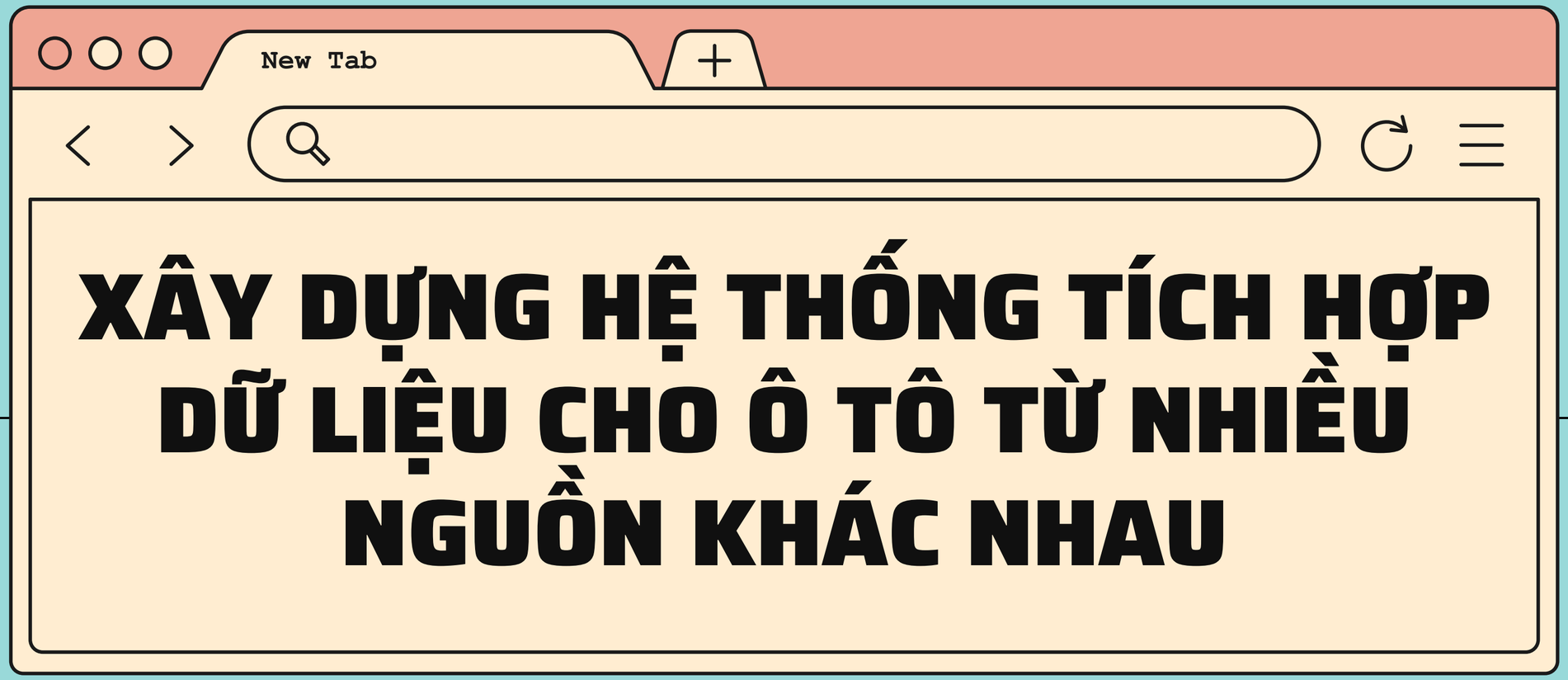
Học phần Tích Hợp Dữ Liệu



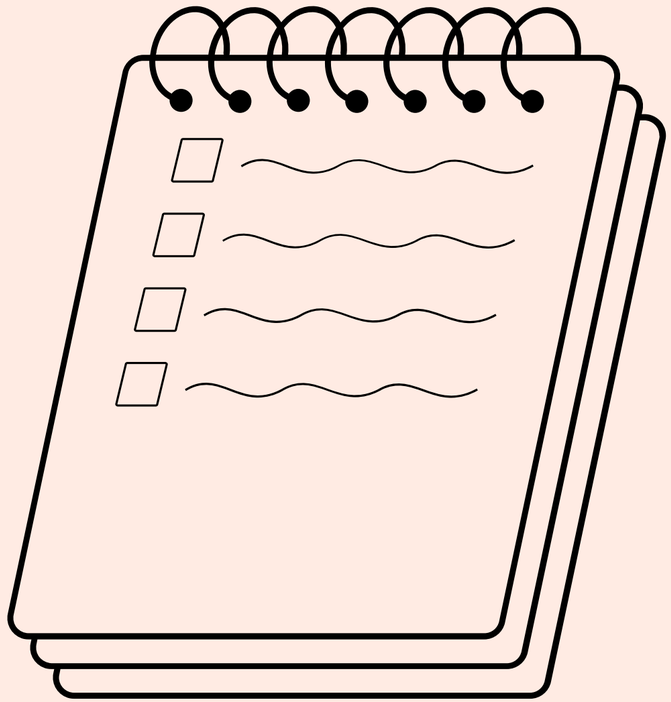
XÂY DỰNG HỆ THỐNG TÍCH HỢP DỮ LIỆU CHO Ô TÔ TỪ NHIỀU NGUỒN KHÁC NHAU

Giảng viên hướng dẫn: TS. Đỗ Bá Lâm

Nhóm 11
Phạm Minh Khôi - 20183566
Đỗ Lương Kiên - 20183568
Nguyễn Hữu Kiệt - 20183571
Nguyễn Đình Lâm - 20183574

Nội dung trình bày

1	Mô tả bài toán
2	Hướng giải quyết bài toán
3	Schema Matching
4	Data Matching
5	Demo chương trình



PHẦN 1: MÔ TẢ BÀI TOÁN

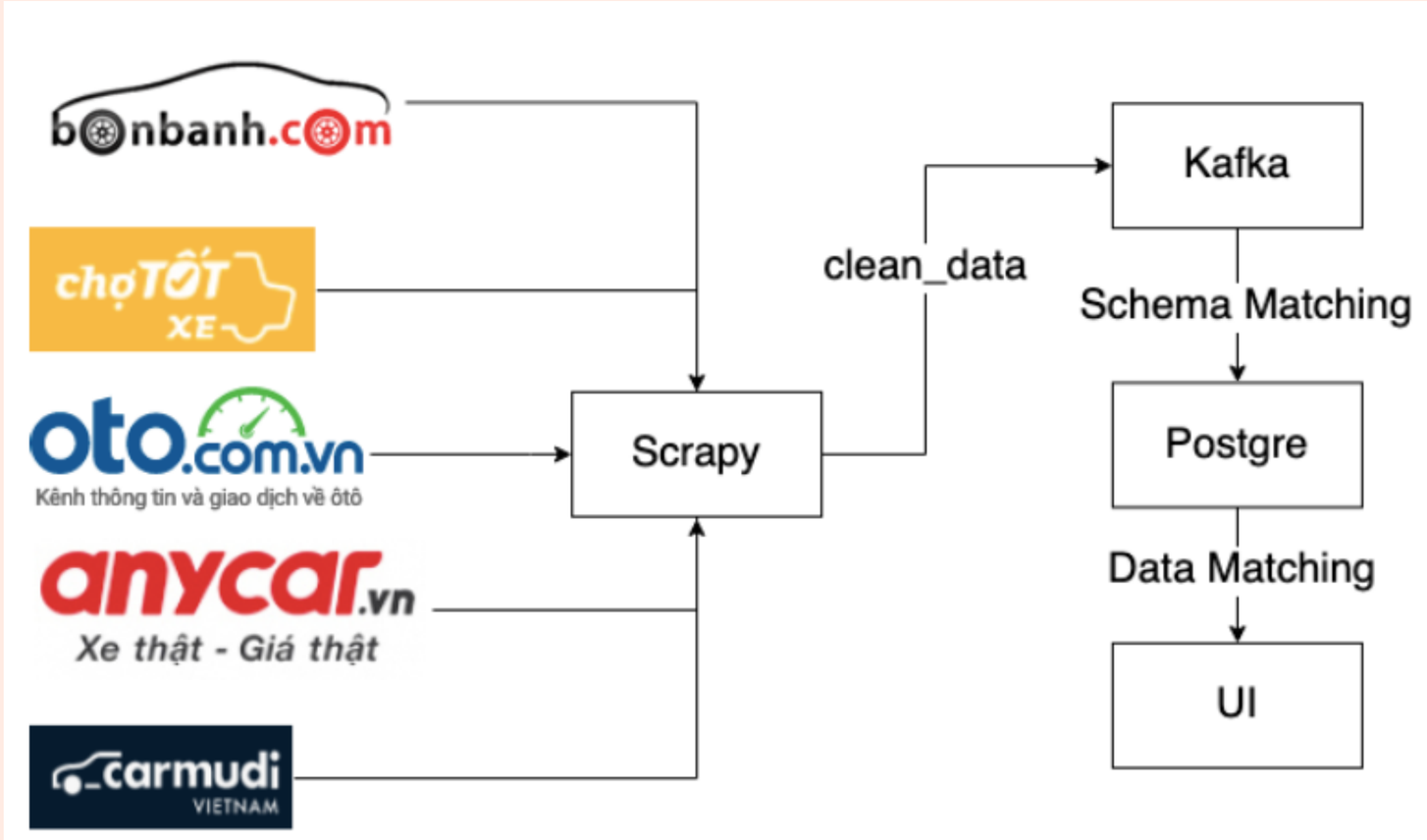
- Sau thời gian khủng hoảng bởi đại dịch covid, nhu cầu mua sắm và đi lại của mọi người đều tăng cao.

- Ô tô dần trở thành một phương tiện không thể thiếu của mọi gia đình vì những ưu điểm của nó.

- Để phục vụ cho nhu cầu mua sắm, tìm kiếm thông tin 1 cách nhanh - gọn, nhóm đã quyết định xây dựng hệ thống tích hợp dữ liệu ô tô từ nhiều nguồn khác nhau, cùng với giao diện thân thiện, dễ sử dụng, giúp cho việc tìm kiếm thông tin của người dùng dễ dàng hơn bao giờ hết.

PHẦN 2: HƯỚNG GIẢI QUYẾT BÀI TOÁN

1	Crawl dữ liệu từ 5 nguồn khác nhau bằng Scrapy
2	Làm sạch dữ liệu crawl được và đưa vào hàng đợi Kafka
3	Tích hợp dữ liệu vào 1 nguồn đích và lưu trữ vào database (Postgre)
4	Thực hiện data matchching
5	Trả dữ liệu về cho phía Client



PHẦN 3: SCHEMA MATCHING

- Sử dụng thuật toán COMA của thư viện Valentine

  - COMA sử dụng cả tên trường và tên thuộc tính để đối sánh lược đồ.
  - Ánh xạ lần lượt từng cặp thuộc tính của 2 lược đồ, từ đó tính ra độ tương đồng

--> Xây dựng lược đồ chung, ánh xạ các lược đồ với lược đồ chung để tổng hợp

PHẦN 3: SCHEMA MATCHING

Thuộc tính	Mô tả
id	Id sản phẩm khi đưa vào lược đồ chung
domain	trang web nguồn sản phẩm
url	nguồn sản phẩm
crawled_date	thời gian crawl
anh_xe	url của ảnh xe chính của bài đăng bán
ten	tên sản phẩm
gia_ban	giá sản phẩm
nam_san_xuat	năm sản xuất
xuat_xu	nguồn gốc sản phẩm (trong nước/ nhập k
tinh_trang	tình trạng sản phẩm (cũ/ mới)

kieu_dang	kiểu dáng
so_km_da_di	quãng đường đã đi
mau_ngoai_that	màu xe bên ngoài
mau_noi_that	màu xe bên trong
so_cua	số cửa
so_cho_ngoi	số chỗ ngồi
nhien_lieu	nhiên liệu
hop_so	hộp số
dan_dong	dẫn động
dung_tich_xi_lanh	dung tích xi lanh
thong_tin_mo_ta	thông tin mô tả

PHẦN 4: DATA MATCHING

- Thực hiện phân cụm dựa trên độ đo khoảng cách.

- Các vấn đề cần giải quyết
  - Lựa chọn độ đo khoảng cách nào?
  - Chọn số lượng cụm là bao nhiêu?
  - Giải thuật dừng khi nào? (gặp điều kiện dừng, sau 1 số lượng vòng lặp...)
  - Hiệu năng
  - Nhiễu
  - ...

PHẦN 4: DATA MATCHING

- Lựa chọn độ đo khoảng cách:
  - Dựa trên thực nghiệm
  - Dựa trên mục đích của việc tích hợp đó là đem các xe của cùng một dòng xe gom vào một nhóm, đồng thời loại hạn chế query các bản ghi trùng lặp
  - Dựa trên đặc trưng của dữ liệu, vì tên của các xe giữa các trang sẽ có sự tương đồng về cách viết. Ví dụ: Xe Toyota Vios 2.4G 2019, Toyota Vios 2018,...

--> Sử dụng độ đo khoảng cách Levenshtein làm độ đo cho giải thuật phân cụm

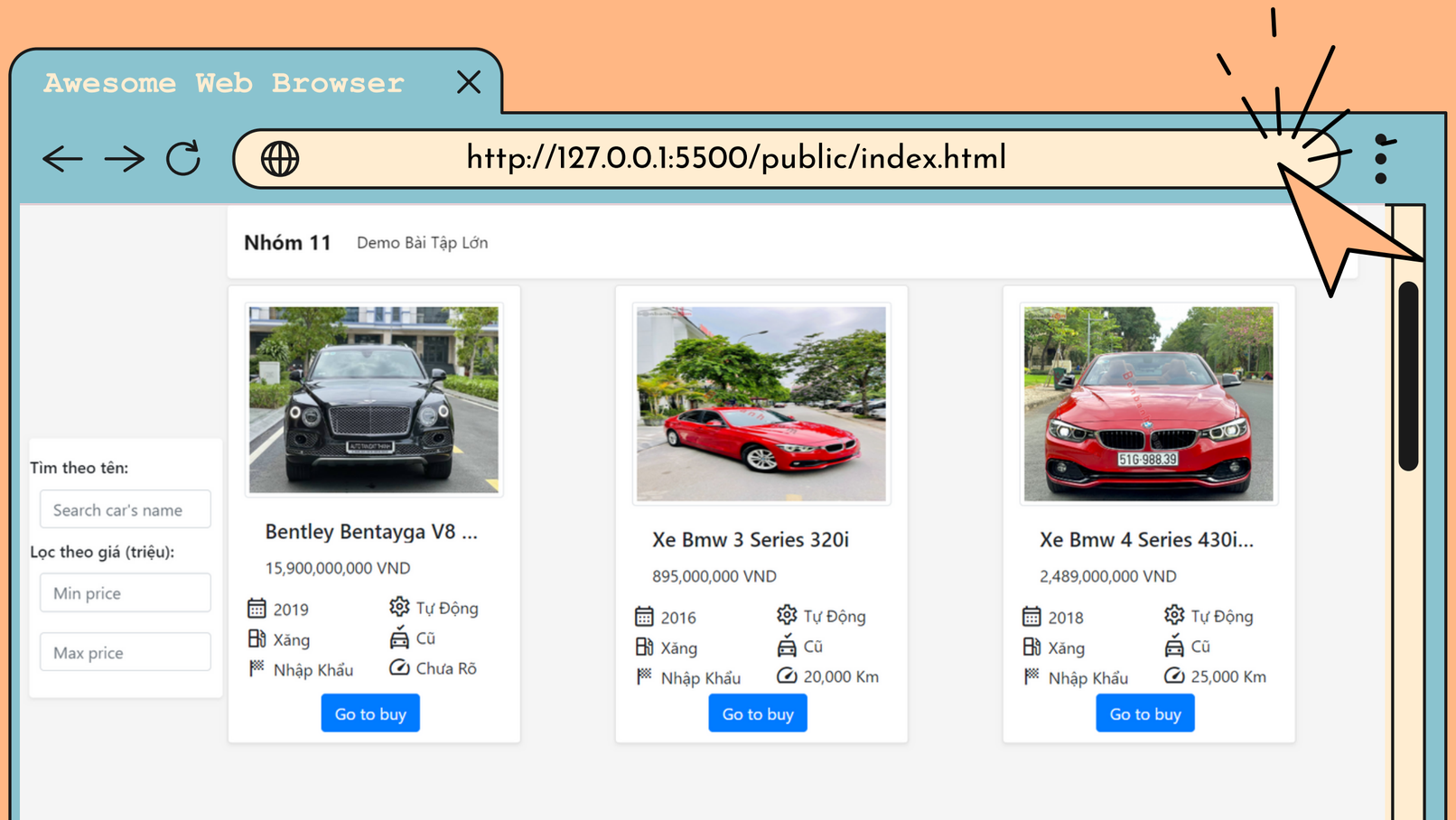
PHẦN 4: DATA MATCHING

- Lựa chọn số lượng cụm để bắt đầu:
  - Dựa trên dữ liệu hiện tại: duyệt lần lượt qua các xe để làm sample, tính khoảng cách đến các xe còn lại, nếu khoảng cách nhỏ hơn 1 ngưỡng thì tạm gom làm 1 cụm, đồng thời những xe đó được loại ra tập xe ban đầu.

  - Với mỗi cụm, tính lại điểm trung bình, rồi lấy các điểm trung bình đó như điểm khởi tạo ban đầu của giải thuật phân cụm.

- Hiệu năng
  - Khi dữ liệu phình to ra, việc update regular data warehouse sẽ rất tốn kém
  - Cải thiện bằng cách đánh chỉ mục hoặc dùng lại các cụm đã có

PHẦN 5: DEMO CHƯƠNG TRÌNH



PHÂN CÔNG CÔNG VIỆC

Tên	Công việc	Nhóm trưởng
Phạm Minh Khôi	Crawl data, Kafka, Báo cáo	
Đỗ Lương Kiên	Crawl data, Frontend, Slide, Thuyết trình	
Nguyễn Hữu Kiệt	Crawl data, Data matching, backend	
Nguyễn Đình Lâm	Crawl data, schema matching, hệ thống	x

Link github: https://github.com/nguyendinhlam88/Data-Integration

TỔNG HỢP SỐ LƯỢT COMMIT

```
PS D:\Documents\20212\Tich hop du lieu\btl\Data-Integration> git shortlog -s
     1  Do Luong Kien
     7  Kien Do
    27  LamYaYa88
     6  Nguyen Huu Kiet
    15  ghuioio
    10  lam.nguyen3
     4  nguyendinhlam88
```

Phạm Minh Khôi	ghuioio	15
Đỗ Lương Kiên	doflamingo0	8
Nguyễn Hữu Kiệt	KietNguyen10112000	6
Nguyễn Đình Lâm	nguyendinhlam88	51


New Tab
THANK YOU FOR LISTENING