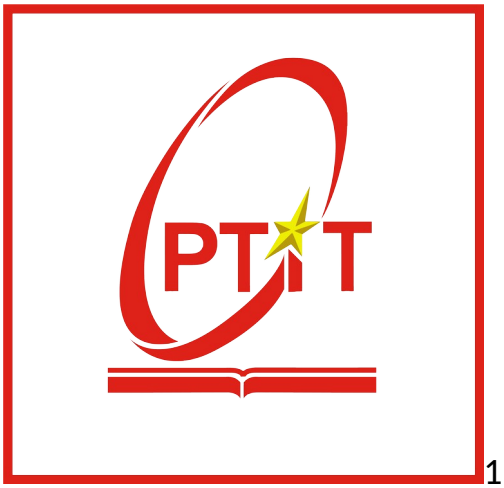
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



1

Báo cáo hàng tuần

Môn học: Thực tập cơ sở

Giảng viên: Kim Ngọc Bách

Họ và tên: Nguyễn Hữu Phúc

Mã SV: B22DCAT224

Lớp: E22CQCN04-B

BÁO CÁO TUẦN 3

I. Phân tích dữ liệu và chọn lọc dữ liệu cần thiết

1. Mô tả dữ liệu

Dữ liệu mẫu trên MongoDB


```
_id: ObjectId('5ed8cb2bc671fc36b74653ad')
time_stamp : 1591266092
ip : "37.170.17.183"
user_agent : "Mozilla/5.0 (iPhone; CPU iPhone OS 13_4_1 like Mac OS X) AppleWebKit/6…"
resolution : "375x667"
user_id_db : "502567"
device_id : "beb2cacb-20af-4f05-9c03-c98e54a1b71a"
api_version : "1.0"
store_id : "12"
local_time : "2020-06-04 12:21:27"
show_recommendation : "false"
current_url : "https://www.glamira.fr/glamira-pendant-viktor.html?alloy=yellow-375"
referrer_url : "https://www.glamira.fr/men-s-necklaces/"
email_address : "pereira.vivien@yahoo.fr"
recommendation : false
utm_source : false
utm_medium : false
collection : "view_product_detail"
product_id : "110474"
▸ option : Array (2)
```


1.1. Thông tin hệ thống và thời gian

- **_id.$oid**: Mã định danh duy nhất của bản ghi trong MongoDB.
- **time_stamp**: Dấu thời gian Unix (giây) khi sự kiện xảy ra.
- **local_time**: Thời gian sự kiện theo múi giờ địa phương của người dùng.
- **api_version**: Phiên bản API của hệ thống.

1.2. Thông tin thiết bị và trình duyệt

- **ip**: Địa chỉ IP của người dùng, có thể sử dụng để định vị địa lý.
- **user_agent**: Chuỗi nhận diện trình duyệt và hệ điều hành.
- **resolution**: Độ phân giải màn hình thiết bị của người dùng.
- **device_id**: Mã định danh duy nhất của thiết bị.

1.3. Thông tin người dùng

- **user_id_db**: ID người dùng trong cơ sở dữ liệu (nếu có).
- **email_address**: Địa chỉ email của người dùng (có thể trống).

1.4. Hành vi và trang web

- **store_id**: ID của cửa hàng mà người dùng đang truy cập.
- **current_url**: URL hiện tại mà người dùng đang xem.

- **referrer_url**: URL trước đó của người dùng, có thể là trang tìm kiếm hoặc trang sản phẩm khác.
- **collection**: Hành động của người dùng, như:
    - **view_product_detail**: Xem chi tiết sản phẩm.
    - **checkout**: Thanh toán.
    - **view_listing_page**: Xem danh sách sản phẩm.
    - **view_home_page**: Xem trang chủ.
    - Còn nhiều nữa
- **show_recommendation**: Cho biết hệ thống có hiển thị gợi ý sản phẩm hay không.

1.5. Thông tin sản phẩm và mua sắm

- **product_id**: ID sản phẩm mà người dùng quan tâm hoặc mua.
- **order_id**: ID đơn hàng (nếu có, trong quá trình thanh toán).
- **cart_products**: Danh sách sản phẩm trong giỏ hàng, bao gồm:
    - **product_id**: ID sản phẩm.
    - **amount**: Số lượng sản phẩm.
    - **option**: Danh sách các tùy chọn sản phẩm mà người dùng đã chọn, chẳng hạn như:
        - **alloy**: Loại hợp kim.
        - **diamond**: Loại kim cương.
        - **quality**: Chất lượng sản phẩm.

1.6. Thông tin chiến dịch marketing

- **utm_source**: Nguồn chiến dịch quảng cáo (ví dụ: Google, Facebook, hoặc `false` nếu không có).
- **utm_medium**: Phương thức quảng cáo (ví dụ: CPC, banner, hoặc `false` nếu không có).

2. Những dữ liệu cần lấy ra để phân tích ETL

2.1. Dữ liệu định danh (ID & Thiết bị)

- **_id.$oid** – ID duy nhất của bản ghi (giữ để theo dõi dữ liệu).
- **device_id** – Mã thiết bị (giúp xác định người dùng ẩn danh).
- **ip** – Địa chỉ IP (có thể dùng để phân tích theo vị trí).

2.2. Dữ liệu về hành vi người dùng trên trang web

- **current_url** – Trang web người dùng đang xem.
- **referrer_url** – Nguồn gốc trước khi truy cập trang hiện tại.
- **collection** – Loại hành động người dùng thực hiện (xem sản phẩm, vào trang chủ, checkout, v.v.).

2.3. Dữ liệu sản phẩm & mua hàng

- **product_id** – ID sản phẩm được xem hoặc thêm vào giỏ hàng.

2.4. Dữ liệu về thời gian & thiết bị

- **time_stamp** – Dấu thời gian Unix (giúp phân tích theo ngày, giờ).

II. Triển khai Kafka với Docker

1. Lý do triển khai Kafka với Docker

- **Dễ dàng triển khai**: Docker giúp cài đặt Kafka nhanh chóng mà không cần cấu hình thủ công từng thành phần.
- Tính di động cao: Chạy Kafka trên nhiều môi trường khác nhau (local, server, cloud) chỉ với một file **docker-compose.yml.**
- **Quản lý tài nguyên tốt**: Docker cô lập Kafka trong container, tránh xung đột với các ứng dụng khác.
- **Dễ dàng mở rộng**: Có thể nhanh chóng thêm nhiều broker Kafka để tạo cluster.
- **Tự động khởi động lại**: Nếu Kafka bị dừng do lỗi hoặc cập nhật hệ thống, Docker có thể tự động khởi động lại.

2. Folder kafka-docker đã được em đẩy lên github