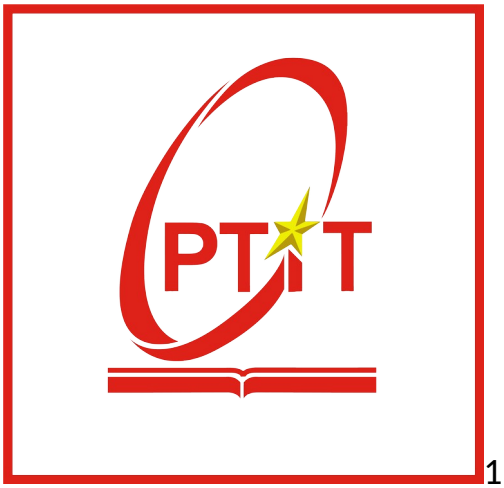
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



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 1

Tổng quan Dự án: Glamira UserFlow Insights

1. Thông tin Dự án

- **Tên dự án**: Glamira UserFlow Insights
- **Mục tiêu**: Xây dựng một pipeline dữ liệu phân tán để xử lý 32 GB dữ liệu log người dùng từ trang web Glamira, lưu trữ dữ liệu đã xử lý vào MongoDB, và áp dụng machine learning để phân tích hành vi người dùng, dự đoán khả năng mua sắm, hoặc hỗ trợ đề xuất sản phẩm cá nhân hóa.
- **Thời gian bắt đầu**: Tháng 3 năm 2025
- **Thành viên tham gia**: Nguyễn Hữu Phúc

2. Mô tả Dữ liệu

- **Nguồn dữ liệu**: Dữ liệu log người dùng từ trang web Glamira, một nền tảng thương mại điện tử chuyên về trang sức và phụ kiện cá nhân hóa.
- **Dung lượng**: 32 GB, bao gồm nhiều tệp log (BSON/JSON).
- **Đặc điểm**: Dữ liệu lớn, không đồng nhất (có thể chứa giá trị thiếu, định dạng không chuẩn).

3. Kiến trúc Pipeline Dữ liệu

Dự án sử dụng một pipeline dữ liệu phân tán với các giai đoạn chính:

3.1. Giai đoạn Thu thập (Ingestion) - Apache Kafka

- **Mô tả**: Đọc 32 GB dữ liệu log và đẩy vào Kafka để xử lý luồng dữ liệu.
- **Công cụ**: Apache Kafka.
- **Quy trình**:
    - Dữ liệu log (tệp BSON/JSON) được đọc tuần tự và gửi vào topic Kafka glamira-user-logs.
    - Kafka đảm bảo xử lý dữ liệu lớn hiệu quả, chia thành nhiều partition để tối ưu hóa tốc độ.
- **Kết quả**: Dữ liệu thô được truyền liên tục qua Kafka, sẵn sàng cho giai đoạn ETL.

3.2. Giai đoạn ETL (Extract, Transform, Load)

- **Mô tả**: Trích xuất dữ liệu từ Kafka, biến đổi (làm sạch, trích xuất đặc trưng), và lưu trữ vào MongoDB.
- **Công cụ**: Apache Spark (PySpark) để xử lý dữ liệu lớn, Python (pandas, kafka-python).
- **Quy trình**:

    - **Extract**: Đọc dữ liệu từ topic glamira-user-logs bằng Spark Streaming.
    - **Transform**:
        - Làm sạch: Xử lý giá trị thiếu, chuẩn hóa timestamp, loại bỏ dữ liệu trùng lặp.
        - Trích xuất đặc trưng: Tính toán số lần nhấp chuột, thời gian phiên (session duration), số sản phẩm xem/thêm vào giỏ.
    - **Load**: Lưu dữ liệu đã xử lý vào MongoDB dưới dạng tài liệu (document) trong collection user_logs.
- **Kết quả**: Dữ liệu sạch, có cấu trúc, được lưu trữ trong MongoDB, sẵn sàng cho phân tích.

3.3. Giai đoạn Lưu trữ - MongoDB

- **Mô tả**: MongoDB được chọn làm cơ sở dữ liệu NoSQL để lưu trữ dữ liệu log đã xử lý, hỗ trợ tốt cho dữ liệu không đồng nhất.
- **Ưu điểm**: MongoDB cho phép truy vấn nhanh, mở rộng dễ dàng với 32 GB dữ liệu, và phù hợp với dữ liệu dạng JSON.

3.4. Giai đoạn Machine Learning

- **Mô tả**: Sử dụng dữ liệu từ MongoDB để huấn luyện mô hình machine learning, dự đoán hành vi người dùng.
- **Công cụ**: Python (scikit-learn, PyTorch), PyMongo (kết nối MongoDB).
- **Quy trình**:
    - Trích xuất dữ liệu từ MongoDB.
    - Xây dựng mô hình phân loại
    - Đánh giá mô hình bằng các chỉ số như accuracy, precision, recall.
- **Kết quả**: Mô hình dự đoán chính xác hành vi người dùng, hỗ trợ đề xuất sản phẩm hoặc tối ưu hóa chiến lược bán hàng.

4. Công cụ và Công nghệ

- **Apache Kafka**: Quản lý luồng dữ liệu thời gian thực.
- **Apache Spark (PySpark)**: Xử lý dữ liệu lớn trong giai đoạn ETL.
- **MongoDB**: Lưu trữ dữ liệu đã xử lý.
- **Python**: Điều phối pipeline và xây dựng mô hình.
- **Scikit-learn/PyTorch**: Xây dựng và huấn luyện mô hình machine learning.
- **Visual Studio Code**: Môi trường phát triển
- **Hệ điều hành:** Linux

5. Kết quả Mong đợi

- **Pipeline dữ liệu**: Một hệ thống xử lý luồng dữ liệu 32 GB hiệu quả, từ Kafka đến MongoDB.
- **Mô hình machine learning**: Mô hình dự đoán khả năng mua sắm với độ chính xác cao (mục tiêu: >80% accuracy).

- **Ứng dụng**: Phân tích hành vi người dùng, đề xuất sản phẩm, tối ưu hóa chiến lược thương mại điện tử cho Glamira.

6. Rủi ro và Giải pháp

- **Rủi ro 1**: Xử lý 32 GB dữ liệu có thể gây tắc nghẽn.
    - **Giải pháp**: Sử dụng Spark để xử lý phân tán, chia dữ liệu thành nhiều partition trong Kafka.
- **Rủi ro 2**: Dữ liệu không đồng nhất (giá trị thiếu, định dạng không chuẩn).
    - **Giải pháp**: Áp dụng quy trình ETL mạnh mẽ với Spark để làm sạch và chuẩn hóa.
- **Rủi ro 3**: Mô hình machine learning không đạt hiệu suất mong đợi.
    - **Giải pháp**: Thử nghiệm nhiều mô hình (Random Forest, Neural Networks), tối ưu hóa đặc trưng.

8. Kết luận

Dự án **Glamira UserFlow Insights** tận dụng các công nghệ hiện đại như Kafka, Spark, và MongoDB để xây dựng một pipeline dữ liệu mạnh mẽ, kết hợp với machine learning để khai thác giá trị từ 32 GB dữ liệu log người dùng. Dự án không chỉ nâng cao khả năng xử lý dữ liệu lớn mà còn cung cấp công cụ phân tích và dự đoán hữu ích cho Glamira, giúp tối ưu hóa trải nghiệm khách hàng và chiến lược kinh doanh.