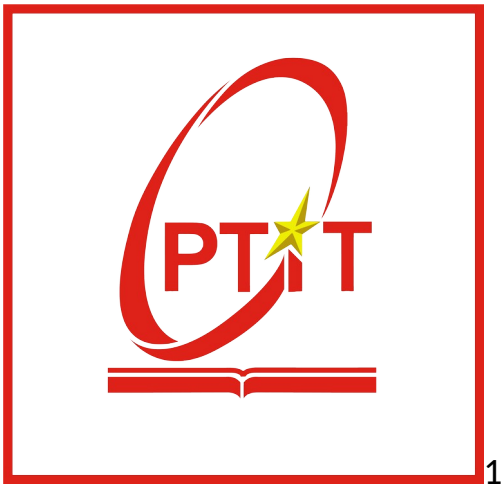
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



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

ĐỀ CƯƠNG DỰ ÁN: GLAMIRA USERFLOW INSIGHTS

1. Thông tin chung về dự án

- **Tên dự án**: Glamira UserFlow Insights
- **Mục tiêu**:
    - Xử lý và phân tích 32 GB dữ liệu log người dùng từ trang web Glamira.
    - Xây dựng các mô hình machine learning để dự đoán hành vi khách hàng dựa trên dữ liệu phân tích.
- **Phạm vi**:
    - Thu thập, lưu trữ, xử lý dữ liệu log người dùng.
    - Thiết kế pipeline dữ liệu từ nguồn thô đến mô hình ML.
    - Tập trung vào hành vi người dùng (user behavior) trên trang web Glamira.
- **Đối tượng hướng đến**:
    - Đội ngũ phát triển sản phẩm, marketing của Glamira.
    - Các bên liên quan cần hiểu rõ hành vi khách hàng để tối ưu hóa trải nghiệm người dùng và chiến lược kinh doanh.

---

2. Quy trình Pipeline chi tiết

2.1. Thu thập dữ liệu (Data)

- **Mô tả**: Thu thập 32 GB dữ liệu log người dùng từ hệ thống của Glamira (web server logs, tracking events, v.v.).
- **Kết quả đầu ra**: Dữ liệu thô được tập hợp và sẵn sàng để đưa vào hệ thống lưu trữ.

2.2. Lưu trữ dữ liệu ban đầu (MongoDB)

- **Mô tả**: Chuyển dữ liệu thô vào MongoDB để lưu trữ dưới dạng NoSQL, phù hợp với dữ liệu log không cấu trúc hoặc bán cấu trúc.
- **Công cụ**:
    - MongoDB
- **Quy trình**:
    - Tạo schema cơ bản
    - Phân vùng dữ liệu theo thời gian hoặc loại sự kiện để tối ưu truy vấn.
- **Kết quả đầu ra**: Dữ liệu được lưu trữ an toàn trong MongoDB, sẵn sàng để stream.

2.3. Stream dữ liệu (Kafka)

- **Mô tả**: Sử dụng Apache Kafka để truyền dữ liệu từ MongoDB sang hệ thống xử lý tiếp theo theo thời gian thực hoặc batch.
- **Công cụ**:
    - Kafka Producer (đẩy dữ liệu từ MongoDB).
    - Kafka Consumer (nhận dữ liệu cho bước ETL).
- **Quy trình**:
    - Thiết lập topic trong Kafka

    - Đảm bảo tính toàn vẹn dữ liệu trong quá trình stream (fault tolerance, retry mechanism).
- **Kết quả đầu ra**: Dữ liệu được stream ổn định, sẵn sàng cho bước xử lý ETL.

2.4. Xử lý dữ liệu (ETL)

- **Mô tả**: Thực hiện quá trình Extract, Transform, Load để làm sạch và chuyển đổi dữ liệu từ Kafka sang định dạng phù hợp cho SQL Server.
- **Công cụ**:
    - Apache Spark hoặc Python (pandas, pyspark) cho xử lý dữ liệu lớn.
    - Công cụ ETL như Apache NiFi hoặc Talend (tùy chọn).
- **Quy trình**:
    - **Extract**: Lấy dữ liệu từ Kafka.
    - **Transform**:
        - Làm sạch dữ liệu (loại bỏ trùng lặp, xử lý giá trị thiếu).
        - Chuẩn hóa dữ liệu (ví dụ: chuyển timestamp sang định dạng thống nhất).
        - Tạo các feature cần thiết cho ML (ví dụ: thời gian trên trang, số lần nhấp chuột).
    - **Load**: Chuẩn bị dữ liệu để đẩy vào SQL Server.
- **Kết quả đầu ra**: Dữ liệu sạch, có cấu trúc, sẵn sàng lưu trữ trong SQL Server.

2.5. Lưu trữ dữ liệu đã xử lý (SQL Server)

- **Mô tả**: Lưu trữ dữ liệu đã qua xử lý trong SQL Server để phục vụ truy vấn và phân tích.
- **Công cụ**:
    - Microsoft SQL Server
- **Quy trình**:
    - Thiết kế schema quan hệ .
    - Import dữ liệu từ bước ETL vào SQL Server.
    - Tối ưu hóa truy vấn
- **Kết quả đầu ra**: Cơ sở dữ liệu quan hệ chứa dữ liệu đã xử lý, sẵn sàng cho bước ML.

2.6. Xây dựng mô hình Machine Learning (Machine Learning)

- **Mô tả**: Sử dụng dữ liệu từ SQL Server để huấn luyện và triển khai các mô hình ML dự đoán hành vi khách hàng.
- **Công cụ**:
    - Python (scikit-learn, TensorFlow, PyTorch).
- **Quy trình**:
    - **Chuẩn bị dữ liệu**: Trích xuất dữ liệu từ SQL Server, chia tập train/test.
    - **Chọn mô hình**:
        - Phân loại (classification) để dự đoán hành vi (mua hàng, rời bỏ giỏ hàng).
        - Hồi quy (regression) để dự đoán giá trị (thời gian trên trang, doanh thu tiềm năng).
    - **Huấn luyện**: Tối ưu hóa mô hình với hyperparameter tuning.

    - **Đánh giá**: Sử dụng các chỉ số như accuracy, precision, recall, hoặc RMSE.
    - **Triển khai**: Đưa mô hình vào production (API hoặc batch prediction).
- **Kết quả đầu ra**: Mô hình ML hoạt động, dự đoán chính xác hành vi khách hàng.

---

3. Thời gian dự kiến (Timeline)

- **Thu thập dữ liệu**: 1 tuần.
- **Lưu trữ vào MongoDB**: 1 tuần.
- **Thiết lập Kafka**: 1 tuần.
- **Xử lý ETL**: 2 tuần.
- **Lưu trữ vào SQL Server**: 1 tuần.
- **Xây dựng và triển khai ML**: 3-4 tuần.
- **Tổng thời gian**: 9-10 tuần (tùy thuộc vào nguồn lực và độ phức tạp).

---

4. Nguồn lực

- **Công nghệ**:
    - MongoDB, Kafka, SQL Server, Spark, Python, Docker

---

5. Kết quả mong đợi

- Một pipeline dữ liệu hoàn chỉnh từ nguồn thô đến mô hình ML.
- Báo cáo phân tích hành vi khách hàng dựa trên dữ liệu log.
- Mô hình ML có khả năng dự đoán hành vi với độ chính xác cao (target >80% tùy mô hình).