BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO TIẾN ĐỘ DỰ ÁN – HOÀN THÀNH 90%

XÂY DỰNG 1 HỆ THỐNG DATAPIPLE LINE ĐỂ PHÂN TÍCH DỮ LIỆU TỪ GLAMIRA

Giảng viên hướng dẫn : Kim Ngọc Bách

Họ và tên : Ngô Vũ Minh Quý

Mã sinh viên : B22DCVT427

Lớp : E22CQCN02-B

1. Tiến độ thực hiện

Sau gần 5 tuần làm việc nghiêm túc và phối hợp hiệu quả giữa các thành viên, dự án hiện đã **hoàn thành khoảng 90% khối lượng công việc theo kế hoạch đề ra**. Cụ thể:

1.1. Phần đã hoàn thành

- **Hạ tầng dữ liệu**:
    - Thiết lập thành công **máy ảo (VM)** và cài đặt **MongoDB**.
    - Tạo và kết nối **MinIO** để thay thế cho GCS.
    - Cài đặt và cấu hình **PostgreSQL** thay thế cho BigQuery.
- **Thu thập và xử lý dữ liệu**:
    - Import dữ liệu từ file gốc vào MongoDB.
    - Thực hiện **xử lý địa chỉ IP** và tạo collection dữ liệu vị trí người dùng.
    - Trích xuất thông tin sản phẩm và crawl tên sản phẩm từ URL.
    - Lưu toàn bộ dữ liệu đã xử lý ra định dạng CSV/JSON để lưu trữ.
- **Xây dựng pipeline**:
    - Hoàn thiện các **script Python** để:
        - Tự động xuất dữ liệu từ MongoDB lên MinIO.
        - Kích hoạt tải dữ liệu sang PostgreSQL (ETL).
    - Có thể chạy thử pipeline đầu–cuối (end-to-end) với tập dữ liệu mẫu.
- **Mô hình hóa dữ liệu**:
    - Đã tạo các bảng **dimension** và **fact** chính cho phân tích.

- Viết các truy vấn chuẩn hóa và kiểm thử sơ bộ chất lượng dữ liệu.
- Tiến hành profiling và mô tả dữ liệu (data dictionary).

1.2. Phần chưa hoàn thành (10%)

- **Trực quan hóa dữ liệu**:
  - Chưa hoàn tất kết nối PostgreSQL với công cụ như Superset/Metabase.
  - Một số dashboard (như doanh thu theo khu vực, phân tích sản phẩm) vẫn đang ở giai đoạn thiết kế và thử nghiệm.
- **Tài liệu hoá và báo cáo cuối kỳ**:
  - Còn thiếu bản tài liệu hướng dẫn người dùng (user guide).
  - Phần tài liệu mô hình hóa (schema diagrams, ERD) đang được hoàn thiện.

---

3. Kế hoạch hoàn thiện (dự kiến trong 2–3 ngày)

- Hoàn thiện trực quan hóa dữ liệu qua Superset hoặc Metabase.
- Kiểm tra lại toàn bộ pipeline với dữ liệu thật.
- Tối ưu script và thực hiện logging đầy đủ.
- Hoàn thành báo cáo, slide thuyết trình và tài liệu hướng dẫn sử dụng.

---

2. Đánh giá chung

Nhìn chung, nhóm đã thực hiện đúng lộ trình và hoàn thành phần lớn các hạng mục quan trọng, bao gồm cả hạ tầng, pipeline, xử lý dữ liệu và mô hình

hóa. Dự án có tính khả thi cao và có thể mở rộng được nếu tích hợp thêm các nguồn dữ liệu khác hoặc thêm các dashboard phân tích nâng cao.

Với tiến độ hiện tại, nhóm **hoàn toàn có thể hoàn thành 100% dự án trong vài ngày tới**, đồng thời đảm bảo chất lượng và khả năng trình bày sản phẩm cuối kỳ tốt nhất.