BỘ THÔNG TIN VÀ TRUYỀN THÔNG

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO THỰC TẬP CƠ SỞ

ĐỀ TÀI
Triển khai Big Data

Giảng viên hướng dẫn : Kim Ngọc Bách

Họ và tên : Ngô Vũ Minh Quý

Mã sinh viên : B22DCVT427

Lớp : E22CQCN02-B

I. Giới thiệu đề tài.

Thu thập, Xử lý & Trực quan hóa Dữ liệu

Tổng quan dự án

Dự án này là một pipeline dữ liệu từ đầu đến cuối, bao gồm thu thập, xử lý, lưu trữ, biến đổi và trực quan hóa dữ liệu. Mục tiêu là xây dựng một hạ tầng dữ liệu tự động, có khả năng mở rộng trên Google Cloud Platform (GCP) và MongoDB, đảm bảo dữ liệu có chất lượng cao, dễ truy cập và cung cấp thông tin giá trị thông qua các dashboard phân tích.

Giai đoạn của dự án

1. Thu thập & Lưu trữ Dữ liệu (Tuần 1-2)

Mục tiêu: Thiết lập hạ tầng cơ bản cho quá trình nhập và lưu trữ dữ liệu.

Công việc chính:

- Thiết lập hạ tầng Cloud:
  - Tạo tài khoản GCP và khởi tạo dự án.
  - Hiểu các khái niệm cơ bản của Cloud.
  - Tải và lưu trữ dataset Glamira.
- Cấu hình Google Cloud Storage (GCS):
  - Tạo bucket trên GCS và cài đặt các lớp lưu trữ.
  - Thiết lập cơ chế xác thực.
  - Tải dữ liệu thô lên GCS.
- Cài đặt Máy Ảo (VM) & MongoDB:
  - Tạo máy ảo trên GCP.
  - Cài đặt và thiết lập MongoDB.
  - Kiểm tra kết nối cơ sở dữ liệu.
- Tải dữ liệu & Khám phá ban đầu:
  - Nhập dữ liệu thô vào MongoDB.
  - Chạy các truy vấn cơ bản để tìm hiểu cấu trúc dữ liệu.
  - Ghi chép lại lược đồ và các mối quan hệ dữ liệu.
- Xử lý Định vị IP:
  - Cài đặt thư viện ip2location-python.
  - Viết script Python để trích xuất địa chỉ IP và lấy dữ liệu vị trí.

    - Lưu kết quả vào một collection mới trong MongoDB.
- Thu thập & Trích xuất Tên Sản phẩm:
    - Lọc dữ liệu sự kiện tương tác với sản phẩm.
    - Trích xuất product_id và current_url.
    - Crawl tên sản phẩm và lưu vào file CSV.
- Kiểm tra chất lượng dữ liệu & Tài liệu hóa:
    - Thực hiện profiling dữ liệu (kiểm tra null, giá trị duy nhất, tính nhất quán).
    - Ghi chép quá trình thiết lập và phát hiện dữ liệu.

2. Pipeline & Lưu trữ Dữ liệu (Tuần 3-4)

Mục tiêu: Tự động hóa quy trình di chuyển dữ liệu từ MongoDB đến GCP Storage và BigQuery.

Công việc chính:

- Xuất dữ liệu từ MongoDB:
    - Viết script Python để trích xuất dữ liệu từ MongoDB.
    - Chuyển đổi dữ liệu sang định dạng phù hợp (CSV, JSON, Parquet).
    - Tải dữ liệu lên GCS.
    - Triển khai logging và xử lý lỗi.
- Tích hợp với BigQuery:
    - Tạo dataset trong BigQuery và định nghĩa schema.
    - Tải dữ liệu từ GCS vào BigQuery.
    - Tự động hóa việc tải dữ liệu bằng Cloud Functions.
- Kiểm tra & Giám sát:
    - Cấu hình cảnh báo khi pipeline gặp lỗi.
    - Kiểm thử pipeline từ đầu đến cuối.
    - Ghi chép tài liệu về pipeline và các quy trình vận hành.

3. Biến đổi & Trực quan hóa Dữ liệu (Tuần 5-6)

Mục tiêu: Xây dựng mô hình dữ liệu và trực quan hóa dữ liệu để phân tích kinh doanh.

Công việc chính:

- Thiết lập dbt & Mô hình dữ liệu:
    - Cài đặt và cấu hình dbt.

    - Kết nối dbt với BigQuery và xây dựng project.
    - Tạo bảng dimensions (ví dụ: chi tiết sản phẩm, khách hàng).
    - Xây dựng bảng fact cho giao dịch mua hàng.
    - Viết các transformation SQL theo best practices.
    - Thêm các bài kiểm tra dữ liệu trong dbt.
- Xây dựng Dashboard trên Looker:
    - Kết nối Looker với BigQuery.
    - Tạo explores và dashboards bao gồm:
        - Phân tích doanh thu.
        - Phân bổ địa lý.
        - Xu hướng theo thời gian.
        - Hiệu suất sản phẩm.
- Tài liệu hóa & Tối ưu hóa:
    - Tối ưu hóa các truy vấn và biến đổi dữ liệu.
    - Cung cấp tài liệu chi tiết và hướng dẫn sử dụng.
    - Công cụ sử dụng
- Google Cloud Platform (GCP): Cloud Storage, Compute Engine, BigQuery, Cloud Functions.
- MongoDB: Lưu trữ dữ liệu gốc và đã xử lý.
- Python: Xử lý dữ liệu, ETL pipeline.
- ip2location-python: Xử lý vị trí địa lý từ địa chỉ IP.
- dbt (Data Build Tool): Chuyển đổi và kiểm thử dữ liệu trong BigQuery.
- Looker: Trực quan hóa dữ liệu và phân tích kinh doanh.
    - Kết quả mong đợi
- Hạ tầng dữ liệu:
    - GCP với các dịch vụ lưu trữ, tính toán và cơ sở dữ liệu đã cấu hình.
- Pipeline tự động:
    - Script Python để trích xuất, xử lý và nhập dữ liệu.
    - Lưu trữ dữ liệu trên GCS và BigQuery.
    - Cloud Functions tự động hóa quá trình ingestion.
- Mô hình dữ liệu & Phân tích:

- Các mô hình dữ liệu trong dbt.
- Báo cáo chất lượng dữ liệu.
- Dashboard Looker với các phân tích kinh doanh.

- Tài liệu & Báo cáo:
  - Hướng dẫn chi tiết về thiết lập và sử dụng pipeline.
  - Tóm tắt các phát hiện và insight quan trọng.
  - Trình bày workflow toàn bộ dự án.

Kết luận

Dự án này xây dựng nền tảng vững chắc cho quy trình xử lý dữ liệu hiện đại, từ thu thập đến phân tích. Việc tận dụng công nghệ cloud và best practices trong ETL, mô hình hóa dữ liệu và phân tích giúp đảm bảo hiệu suất, độ tin cậy và khả năng mở rộng của hệ thống.