BỘ GIÁO DỤC & ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

KHOA CÔNG NGHỆ THÔNG TIN

ꕥ&ꕥ

Bài tập : Hoạt động activity 2

Môn học: Xử lý dữ liệu trực tuyến

Bộ môn: Hệ thống thông tin

Mã lớp: CQ2022/1

GVHD: Nguyễn Trần Minh Thư

Tên nhóm: Nhóm 4

Thành viên:

Lê Đức Cường	MSSV:
Nguyễn Dương Trường Sinh	MSSV:
Phạm Trần Trung Hậu	MSSV:
Bùi Đoàn Thuý Vy	MSSV: 22120448

Thành phố Hồ Chí Minh, ngày 25 tháng 03 năm 2025

AG2_DL

I. Mô phỏng đầy đủ hoạt động và các khái niệm liên quan đến nội dung bài học Data Engineering LifeCycle

1. Sources of Data

- Là nơi dữ liệu được sinh ra:
  - Ứng dụng Web/mobile
  - Flat file
  - Database
  - Web Scraping
  - API, Web Services
  - Iot Systems
  - Data Lake, Data warehouse
- Dữ liệu có thể là: thời gian thực (stream), định kỳ(batch), theo sự kiện (event- driven)

2. Ingestion

- Đưa raw data từ nguồn vào data pipeline để xử lý
- Đây là quá trình kết hợp dữ liệu từ nhiều nguồn khác nhau(disparate sources) thành một hệ thống trung tâm duy nhất, để người dùng có thể truy vấn và phân tích.
- Các bước trong quy trình Ingestion:
  - Collated
  - Processed
  - Clean
  - Integrated
  - Users
- Có hai hình thức chính:
  - Batch ingestion: thu thập định kỳ, phù hợp với hệ thống truyền thống
  - Stream ingestion: thu thập thời gian thực, phù hợp với dữ liệu cập nhật liên tục
- Công cụ: Kafka, Apache NiFi, Flume, Airbyte, AWS Kinesis

3. Transformation

- Là biến raw data thành dữ liệu có ích -> phục vụ cho phân tích, mô hình hóa và ra quyết định
- Làm sạch và xử lý dữ liệu thành dạng phù hợp:
  - Loại bỏ trùng lặp, null
  - Format lại dữ liệu (định dạng ngày giờ, chuẩn hóa đơn vị đo)
  - Join nhiều nguồn, tính toán KPIs
  - Áp dụng các logic nghiệp vụ (business rules)
- Các hoạt động chính:
  - Queries:
  - Modeling: chuyển đổi dữ liệu thô thành cấu trúc phù hợp với mục tiêu phân tích

- Transformation: là quá trình biến đổi raw data -> dữ liệu chuẩn, có thể sử dụng trong các bước tiếp theo(downstream use) , gồm các hoạt động như Manipulated (xử lý), Enhanced (bổ sung giá trị), Saved (lưu trữ),

4. Storage

- Nơi dữ liệu được lưu trữ tạm thời hoặc lâu dài cho các bước xử lý tiếp theo
- Kiến trúc phân tầng trong Storage
  - Raw Ingredients (Nguyên liệu thô)
  - Storage Systems (Hệ thống lưu trữ)
  - Storage Abstractions (Trừu tượng lưu trữ)
- Đi từ Storage Abstractions đến Raw Ingredients thể hiện sự trừu tượng hóa và tích hợp cao dần
- Bắt đầu từ phần cứng (low-level) ➜ hệ thống quản lý ➜ khái niệm hóa giúp phân tích, xử lý, khai thác dữ liệu.

5. Serving

- Là giai đoạn đưa dữ liệu đã được xử lý đến tay người dùng cuối để phân tích, ra quyết định hoặc tích hợp vào hệ thống khác.
- Analytics trong Serving data: quá trình phân tích dữ liệu để tìm ra insights (thông tin quan trọng) và mô hình (patterns).
- Các loại Analytics :
  - Business Intelligence (BI)
  - Operational Analytics
  - Embedded Analytics
- Machine Learning in Serving Data: Là quá trình triển khai các mô hình học máy vào môi trường thực tế để:
  - Dự đoán theo thời gian thực dựa trên dữ liệu đầu vào (real-time inference)
  - Phục vụ kết quả mô hình cho các hệ thống hoặc người dùng
  - Theo dõi nguồn gốc dữ liệu và lịch sử mô hình (data lineage & model tracking) để đảm bảo tính chính xác và kiểm soát
- Reverse ETL: quá trình đẩy dữ liệu từ kho dữ liệu (data warehouse) trở lại các hệ thống hoạt động (như CRM, công cụ marketing, ứng dụng...), nhằm phục vụ cho các hoạt động phân tích, dự đoán và ra quyết định.

II. Phát triển bài nghiên cứu Grab Food ở cấp độ sâu hơn

a. Mục tiêu:

Cải thiện hệ thống đề xuất món ăn trên GrabFood. Cụ thể, *phân tích hành vi của người dùng để dự đoán món ăn tiếp theo* mà họ có khả năng đặt và tìm hiểu lý *do tại sao người dùng từ bỏ giỏ hàng khi chưa hoàn tất đơn hàng*.

b. Phát triển

1. Generation – Dữ liệu được sinh ra

- Nguồn phát sinh dữ liệu thô từ hệ thống thực tế
    - Người dùng thực hiện thao tác: tìm kiếm món, click vào món, thêm giỏ hàng, xóa khỏi giỏ, đặt hàng.
    - Nhà hàng cập nhật thực đơn, giảm giá, chương trình khuyến mãi.
    - Hệ thống thu thập thêm thông tin vị trí, thời tiết, giờ trong ngày...

2. Data Ingestion – Thu thập dữ liệu

- Dữ liệu cần thu thập:
    - **Hành vi người dùng**: click, xem món, thêm vào giỏ, mua hàng, thời gian tương tác, từ khóa tìm kiếm, thời gian sử dụng ứng dụng, thời gian xem sản phẩm,..
    - **Thông tin đơn hàng**: món đã đặt, thời gian đặt, vị trí người dùng, tổng giá trị đơn.
    - **Các hoạt động đối với giỏ hàng**: Thêm, xóa và thay đổi số lượng các món ăn hoặc các đơn hàng được thêm món nhưng không hoàn tất.
    - **Thông tin giỏ hàng bị bỏ**: thời gian rời đi, bước rời khỏi, món trong giỏ.
    - **Dữ liệu bổ sung**: thời tiết, sự kiện đặc biệt (VD: lễ, bóng đá), chương trình khuyến mãi.
    - **Thông tin thanh toán**: Phương thức thanh toán, việc sử dụng các ưu đãi.
    - **Feedback**: đánh giá món, khảo sát sau khi hủy đơn.
- Công cụ:
    - **Streaming ingestion**: dùng Apache Kafka, Kinesis hoặc Pub/Sub để thu thập sự kiện theo thời gian thực.
    - **Batch ingestion**: thu thập logs hoặc đơn hàng mỗi giờ/ngày qua job ETL.
    - Tích hợp từ các **nguồn bên ngoài** như API thời tiết hoặc Google Maps để người dùng có thể theo dõi hành trình đơn hàng theo thời gian thực hoặc thanh toán (các ứng dụng như Momo, Zalopay, Smartbanking).

3. Storage – Lưu trữ dữ liệu

- Tổ chức lưu trữ dữ liệu theo dạng có cấu trúc & bán cấu trúc:
    - Dữ liệu bán cấu trúc (event logs, clickstream) → lưu trong Data Lake (AWS S3, Google Cloud Storage).
    - Dữ liệu quan hệ (đơn hàng, người dùng, nhà hàng) → lưu trong Data Warehouse (BigQuery, Snowflake, Redshift).

  - Dữ liệu real-time cache cho gợi ý nhanh → Redis, Memcached.
- Tạo các lớp bộ nhớ đệm cho các dữ liệu thường xuyên được truy cập như các món ăn đã đặt của người dùng, các tìm kiếm trong lịch sử.
- Thiết lập các phân vùng (partition) dựa theo từng khu vực hoặc theo thời gian.
- Đảm bảo các chính sách về bảo mật dữ liệu người dùng.

4. Transformation

- Xử lý dữ liệu thô thành dữ liệu hữu ích
- Làm sạch dữ liệu: loại bỏ null, format lại timestamp
- Tính toán các feature như:
  - Tần suất đặt món
  - Thời gian người dùng rời khỏi giỏ hàng
  - Tổng tiền trung bình trong giỏ hàng
- Chuẩn hóa tên các món ăn để việc phân loại và gợi ý món ăn được dễ dàng hơn (ví dụ "cơm tấm" hay "cơm sườn" được xem như cùng một món)
- Tổng hợp các hành vi người dùng theo từng phiên hoạt động của họ (từ khi mở ứng dụng đến lúc giao hàng).
- Tính toán các thông số về việc bỏ giỏ hàng hoặc thời gian trung bình hoàn thành đơn hàng để đo lường độ hiệu quả của hệ thống gợi ý và cải thiện đề xuất.
- Công cụ : Apache Spark, dbt, Airflow.

5. Serving

- Làm cho dữ liệu được sử dụng cho hệ thống downstream (phía sử dụng)
- **Analytics:**
  - Phân nhóm người dùng theo hành vi (clustering: người thường đặt 1 món, người hay xem nhưng không mua...).
  - Phân tích dữ liệu theo từng phiên hoạt động của người dùng.
  - Trực quan hoá dữ liệu theo dạng phễu để tìm các điểm gây cản trở trong quá trình hoàn tất đơn hàng.
  - Phân tích thời gian để xác định khi nào người dùng hay bỏ giỏ hàng nhất (giai đoạn chọn món, thanh toán hay vận chuyển).
  - Xây dựng tính năng đầu vào cho hệ thống gợi ý:
    - Món ăn phổ biến
    - Món ăn gần đây nhất
    - Giờ ăn quen thuộc
    - Nhà hàng thường đặt
- **Mô hình máy học:**
  - Dùng Collaborative filtering models để gợi ý các món ăn theo từng thời điểm trong ngày hoặc các món ăn được đặt cùng nhau, dự đoán thời gian người dùng đặt hàng lại và hiển thị thông báo ứng dụng phù hợp.
  - Sử dụng Clustering algorithms để gom nhóm những người dùng có hành vi sử dụng giống nhau.

- Áp dụng các Classification models để dự đoán các khả năng người dùng bỏ giỏ hàng.
- Sử dụng xử lý ngôn ngữ tự nhiên (NLP) để phân tích các từ khoá mà người dùng tìm kiếm để gợi ý món ăn phù hợp.

- **Reverse ETL**
  - Liên tục cập nhật lịch sử đặt hàng để phân tích và gợi ý món cho những lần đặt hàng kế tiếp.
  - Đồng bộ hoá dữ liệu người dùng theo từng nhóm để cải thiện hệ thống gợi ý món.
  - Cung cấp các gợi ý cá nhân hoá thông qua thông báo ứng dụng hoặc kênh thông tin như email cá nhân.