***Bài hoạt động:*** *(Mã bài tập: AG2\_DL)* sinh viên hoàn thiện bài tập hoạt động Case study, mô phỏng đầy đủ hoạt động và các khái niệm liên quan nội dung bài học Data Engineering Lifecycle mức tổng quát. Lưu ý: phát triển tiếp bài nghiên cứu đang thực hiện của mỗi nhóm ở cấp độ sâu hơn về thông tin đã học.

**Data Engineering Lifecycle** (Vòng đời kỹ thuật dữ liệu) bao gồm toàn bộ quá trình chuyển đổi dữ liệu thô thành sản phẩm cuối hữu ích. Quá trình này bao gồm nhiều giai đoạn, mỗi giai đoạn có vai trò và trách nhiệm cụ thể. Vòng đời này đảm bảo rằng dữ liệu được xử lý hiệu quả và hiệu suất, từ khi được tạo ra ban đầu đến khi được sử dụng cuối cùng. Data Engineering Lifecycle được chia thành 5 giai đoạn chính:

*1. Generation - Tạo ra dữ liệu:* Thu thập dữ liệu từ nhiều nguồn khác nhau.

*2. Ingestion - Tiếp nhận dữ liệu:* Mang dữ liệu vào một hệ thống tập trung.

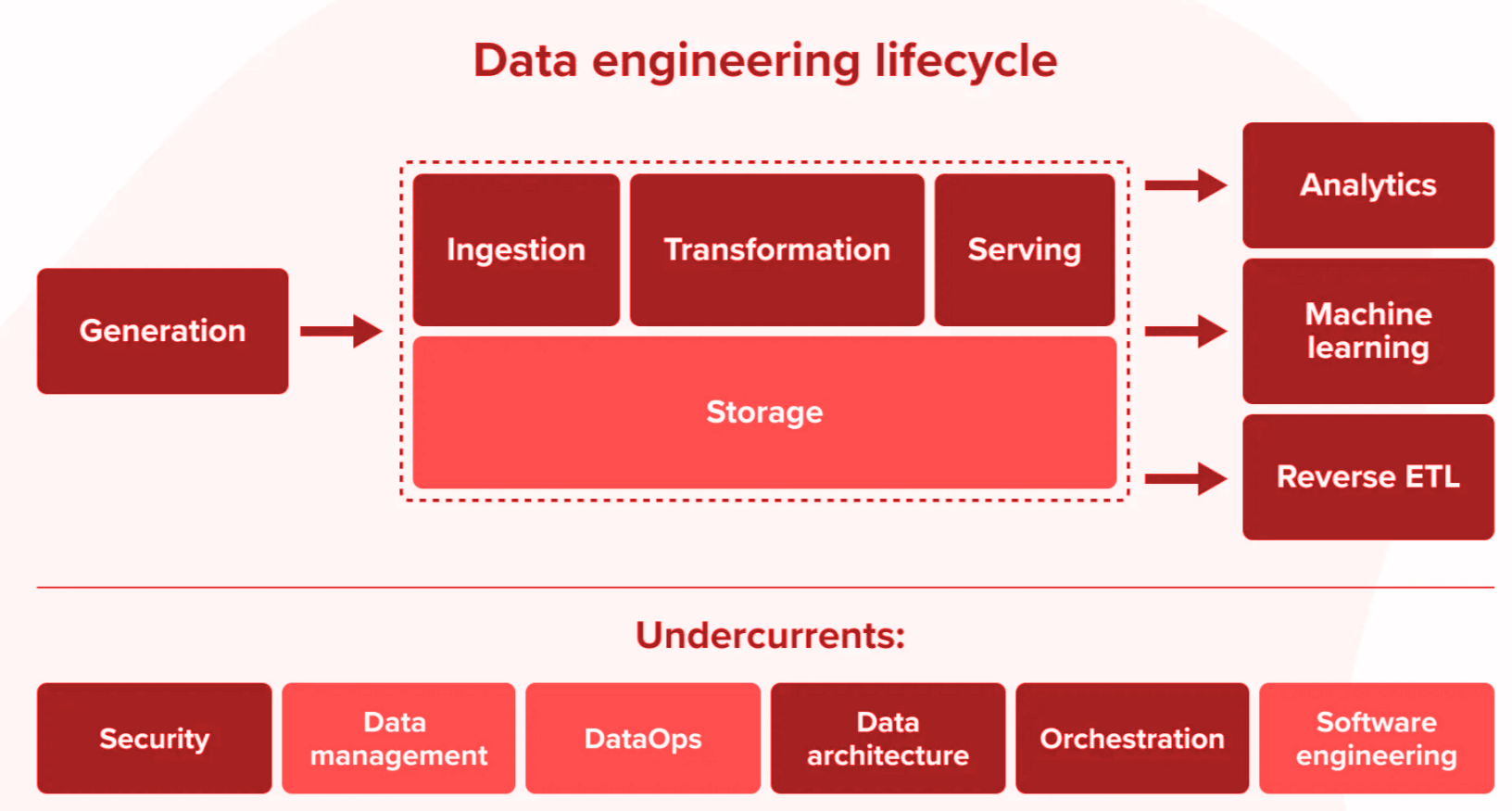
*3. Storage - Lưu trữ dữ liệu:* Lưu trữ dữ liệu an toàn để xử lý và phân tích trong tương lai.

*4. Transformation - Chuyển đổi dữ liệu:* Chuyển đổi dữ liệu sang định dạng phù hợp cho phân tích.

*5. Serving - Phân phối dữ liệu:* Cung cấp dữ liệu cho người dùng cuối cho mục đích ra quyết định và vận hành.

Quy trình chung có thể được hiểu như sau:

Dữ liệu đầu → Biến đổi thành dữ liệu có giá trị → Sử dụng dữ liệu cho các hành động ứng dụng, phân tích



**Case study:** *Phân tích hành vi người dùng Grabfood cho việc dự đoán và gợi ý món ăn tiếp theo mà người dùng có thể đặt. Đồng thời phân tích các nguyên do khiến người dùng từ bỏ giỏ hàng.*

**Các thành phần chính của chu trình:**

### *1. Generation*

Dữ liệu người dùng cần thu thập thông qua các thông tin sau đây:

- Tương tác của người dùng với ứng dụng: Từ khoá tìm kiếm, thời gian sử dụng ứng dụng, thời gian xem sản phẩm,...

- Lịch sử đơn hàng: Các món ăn đã đặt, giá tiền đơn hàng,...

- Các hoạt động đối với giỏ hàng: Thêm, xoá và thay đổi số lượng các món ăn hoặc các đơn hàng được thêm món nhưng không hoàn tất.

- Thông tin giao hàng: Địa điểm giao hàng, thời gian thường đặt đơn.

- Thông tin thanh toán: Phương thức thanh toán, việc sử dụng các ưu đãi.

- Đánh giá: Các phản hồi và đánh giá của người dùng đối với đơn hàng và người giao hàng.

### *2. Ingestion*

Ở giai đoạn này, hệ thống của *Grabfood* sẽ thực hiện những tác vụ sau để thu nhận được dữ liệu yêu cầu ở phần Generation:

- Cấu hình luồng dữ liệu thời gian thực phù hợp với việc hiển thị dữ liệu đề xuất các quán ăn, món ăn đối với từng người dùng ngay thời điểm họ sử dụng.

- Ghi nhận hành vi của người dùng thông qua các API calls và event logs.

- Sử dụng tracking pixels hoặc session recording để lập bản đồ hành trình người dùng từ lúc bắt đầu đến lúc họ thanh toán đơn hàng.

- Sử dụng database listeners để theo dõi những thay đổi giỏ hàng như việc thêm, xoá, sửa thông tin các món ăn trong giỏ hàng.

- Tích hợp API với bên thứ ba về vấn đề thanh toán (các ứng dụng như Momo, Zalopay, Smartbanking) và vận chuyển (Google Map) để người dùng có thể theo dõi hành trình đơn hàng theo thời gian thực.

### *3. Storage*

Các thiết lập về lưu trữ được cài đặt phục vụ cho các loại dữ liệu trước và sau khi xử lý như sau:

- Dữ liệu thô ban đầu sẽ được lưu trữ vào data lake.

- Đối với dữ liệu cấu trúc sau khi xử lý được lưu trữ trong data warehouse, điều này giúp cho việc phân tích và dự đoán dễ dàng và hiệu quả hơn.

- Thiết lập cơ sở dữ liệu NoSQL nhằm phục vụ cho việc lưu trữ được các loại như liệu phi cấu trúc như các tương tác của người dùng với ứng dụng.

- 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*

Dữ liệu cần được xử lý nhiều bước để thuận tiện cho việc phân tích và sử dụng các mô hình ML:

- Chuẩn hoá tên các món ăn để việc phân loại và gợi ý món ăn được dễ dàng hơn (ví dụ như “cơm tấm” hay “cơm sườn” được xem như là cùng một món).

- Trích xuất riêng các đặc điểm sử dụng của người dùng như thời gian người dùng ở một trang, độ cuộn sâu cho việc phân tích và gợi ý món ăn.

- Chuyển đổi các đoạn thời gian thành các mốc thời gian ý nghĩa (giờ ăn trưa, ăn tối hay ăn khuya) để phân tích hành vi dựa trên thời gian.

- Tổng hợp các hành vi người dùng theo từng phiên hoạt động của họ (từ khi mở ứng dụng đến lúc giao hàng).

- Xây dựng các vector đặc trưng (sở thích, thời gian) để làm đầu vào cho hệ thống gợi ý.

- 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.

### *5. Serving*

Các dữ liệu sau khi được xử lý và lưu trữ sẽ được thể hiện như sau:

- API cung cấp khả năng truy xuất dữ liệu gợi ý trong thời gian thực.

- Các dữ liệu không cần xử lý ngay lập tức được xử lý theo lô (batch) để phân tích xu hướng lâu dài hoặc theo từng tuần, từng tháng.

- Các khối dữ liệu được tổng hợp thành dashboards phục vụ cho BI, đồng thời giúp cho các quản lý theo dõi được dữ liệu nhanh chónh và dễ dàng.

- Sử dụng các event-based triggers như việc gửi thông báo về các ưu đãi giảm giá hoặc nhắc nhở nhằm khuyến khích người dùng sử dụng ứng dụng.

- Các báo cáo định kỳ (hàng ngày, hàng tuần hoặc hàng tháng) được tạo để phục vụ cập nhật tình hình kinh doanh và các chiến lược.

- Dùng Feature stores để lưu trữ các đặc trưng được chuẩn bị sẵn dùng làm cho các mô hình ML.

**Các mục đích sử dụng cho dữ liệu đầu ra:**

### *1. Analytics - Phân tích dữ liệu*

- Phân tích hành vi người dùng theo từng nhóm thông qua các thói quen mua sắm của các nhóm khác nhau.

- 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.

- Đo lường tỷ lệ chuyển đổi giữa các danh mục thực phẩm khác nhau để đưa ra gợi ý phù hợp và giảm tỷ lệ bỏ 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).

### *2. Machine Learning - Áp dụng các 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.

### *3. Reverse ETL (Extract, Transform, Load) - Tích hợp ngược*

- 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.