# *Sử dụng Data Warehouse và Data Lake vào case study:*

## **Giai đoạn 1**: Xây dựng kiến trúc dữ liệu Lakehouse

### - Thiết lập Data Lake

+ Bronze Layer: Thu thập dữ liệu thô từ tất cả nguồn

Lưu trữ đầy đủ logs ứng dụng, clickstream, giao dịch.

Tích hợp Delta Lake cho ACID transactions.

Định dạng: Parquet, Avro, JSON, text.

+ Silver Layer: Dữ liệu chuẩn hóa và làm sạch

Loại bỏ dữ liệu trùng lặp.

Chuẩn hóa trường dữ liệu.

Làm giàu dữ liệu với các thuộc tính phụ.

+ Gold Layer: Dữ liệu phân tích sẵn sàng

Tính toán các aggregates phổ biến.

Tạo các bảng feature cho Machine Learning.

Tạo các views cho phân tích nhanh.

### - Thiết lập Data Warehouse:

+ Dimensional Model:

Xây dựng star schema cho phân tích đa chiều.

Tạo các dimension tables: Customer, Restaurant, Food, Time, Location.

Tạo các fact tables: Orders, Cancellations, Interactions.

+ Real-time Data Warehouse:

Sử dụng Snowflake/BigQuery để hỗ trợ truy vấn thời gian thực.

Thiết lập materialized views cho truy vấn thường xuyên.

Cấu hình auto-scaling cho đáp ứng nhu cầu cao điểm.

## **Giai đoạn 2**: Phát triển pipeline dữ liệu

+ Batch Processing:

ETL hàng ngày từ Bronze đến Silver và Gold layers.

Tự động kiểm tra chất lượng dữ liệu.

Orchestration với Apache Airflow.

+ Stream Processing:

Kafka cho thu thập dữ liệu thời gian thực.

Spark Streaming/Flink cho xử lý luồng.

Cập nhật trực tiếp vào Silver layer và Data Warehouse.

+ Feature Store:

Lưu trữ và quản lý tính năng cho mô hình ML.

Đồng bộ hóa tính năng online và offline.

Hỗ trợ phục vụ tính năng độ trễ cực thấp.

## **Giai đoạn 3**: Phát triển mô hình phân tích

### - Mô hình dự đoán món ăn tiếp theo

+ Mô hình collaborative filtering: Dựa trên dữ liệu người dùng tương đồng từ Data Warehouse

+ Mô hình content-based: Phân tích đặc điểm món ăn và sở thích từ Data Lake

+ Mô hình kết hợp (hybrid): Tích hợp dữ liệu từ cả hai nguồn

+ Mô hình học sâu: Huấn luyện trên dữ liệu lớn từ Data Lake

### - Mô hình dự đoán hủy đơn

+ Mô hình phân loại: Huấn luyện trên dữ liệu có cấu trúc từ Data Warehouse

+ Mô hình xác suất: Tích hợp dữ liệu từ nhiều nguồn thông qua Data Lake

+ Mô hình thời gian thực: Sử dụng Feature Store để suy luận nhanh

+ Mô hình can thiệp chủ động: Tích hợp với hệ thống quyết định thời gian thực

## **Giai đoạn 4**: MLOps và triển khai

### + Tích hợp MLOps

Model Registry: Quản lý phiên bản mô hình

CI/CD pipeline: Tự động hóa việc huấn luyện và triển khai

Model Monitoring: Theo dõi hiệu suất và drift detection

Experiment Tracking: So sánh và lưu trữ các thử nghiệm

### + Microservices Architecture

API Gateway: Điểm truy cập thống nhất

Recommendation Service: Tính toán đề xuất thời gian thực

Cancellation Prediction Service: Dự báo hủy đơn

Feature Service: Kết nối với Feature Store

### + Thử nghiệm A/B

Sử dụng dữ liệu từ Data Warehouse để đánh giá kết quả

Phân tích chi tiết trải nghiệm người dùng từ Data Lake

Tối ưu hóa dựa trên các chỉ số: CTR, tỷ lệ chuyển đổi, tỷ lệ hủy đơn

## **Giai đoạn 5**: Phân tích nâng cao và trực quan hóa

### + Phân tích dữ liệu nâng cao

Customer 360: Tạo cái nhìn toàn diện về khách hàng từ dữ liệu đa nguồn

Time Series Analysis: Phân tích xu hướng dài hạn từ Data Warehouse

Network Analysis: Phân tích mối quan hệ giữa người dùng và món ăn

### + Nền tảng trực quan hóa

Business Intelligence: Dashboards từ Data Warehouse

Customer Analytics: Trực quan hóa từ phân tích Data Lake

Real-time Monitoring: Theo dõi hoạt động hệ thống và hiệu suất

## **Giai đoạn 6**: Mở rộng và phát triển

### + Mở rộng tích hợp

Cross-service Recommendation: Liên kết với các dịch vụ Grab khác

Partner Analytics: Cung cấp insights cho đối tác nhà hàng

Edge Computing: Đưa một phần xử lý đến thiết bị người dùng

### + Chiến lược dữ liệu dài hạn

Data Governance: Xây dựng chính sách quản trị dữ liệu

Data Retention: Chiến lược lưu trữ dữ liệu dài hạn

Privacy and Compliance: Đảm bảo tuân thủ các quy định về dữ liệu

## *1. Mô hình phân tích dự đoán món ăn tiếp theo với kiến trúc Lakehouse*

### - Phân tích dữ liệu từ nhiều nguồn:

+ Dữ liệu từ Data Lake:

Dữ liệu thô về lịch sử tương tác (clickstream)

Hình ảnh món ăn được người dùng tương tác nhiều

Dữ liệu văn bản từ đánh giá và bình luận

Logs ứng dụng đầy đủ

+ Dữ liệu từ Data Warehouse:

Lịch sử giao dịch đã chuẩn hóa

Tỷ lệ hoàn thành đơn theo thời gian

Phân khúc khách hàng

Metrics kinh doanh về nhà hàng và món ăn

### - Feature Engineering và Model Training:

+ Tạo Feature từ Data Lake:

Phân tích cảm xúc từ đánh giá bằng NLP

Trích xuất đặc trưng hình ảnh từ món ăn được click

Phát hiện mẫu hành vi người dùng phức tạp

Features theo thời gian thực từ stream processing

+ Tạo Feature từ Data Warehouse:

Tính toán các chỉ số tần suất, thời gian giữa các đơn hàng

Phân tích thời điểm mua hàng theo ngày trong tuần, thời điểm

Tổng hợp chi tiêu và sở thích theo nhóm sản phẩm

Tạo các chỉ số RFM (Recency, Frequency, Monetary)

### *2. Mô hình dự đoán hủy đơn hàng*

+ Phân tích offline (Data Warehouse):

Xác định các yếu tố dự báo hủy đơn mạnh nhất

Phân loại nguyên nhân hủy đơn từ dữ liệu lịch sử

Xây dựng các modeled tables cho phân tích đa chiều

+ Phân tích thời gian thực (Real-time Layer):

Sử dụng Feature Store để truy xuất tính năng nhanh chóng

Áp dụng mô hình dự đoán tức thời khi người dùng xem món

Can thiệp chủ động trước khi người dùng hủy đơn