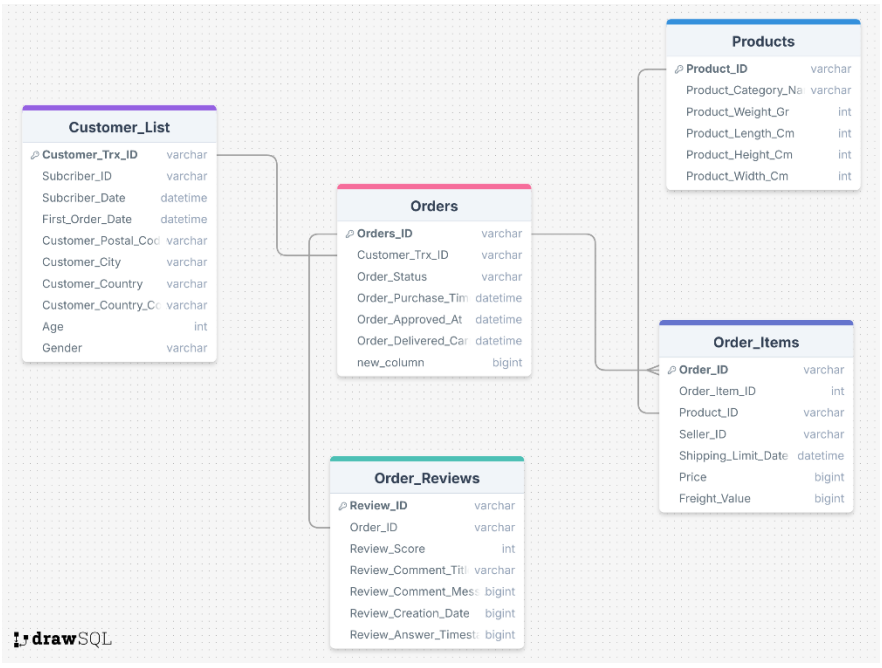
Lab 3 - Phân Tích Dữ Liệu Thương Mại sử dụng Spark DataFrame

Fecom Inc. là công ty thương mại điện tử có trụ sở tại Berlin, Đức. Từ năm 2022 đến 2024, công ty đã ghi nhận 99.441 đơn hàng từ 102.727 khách hàng duy nhất và theo dõi giao dịch của 3.095 người bán. Bộ dữ liệu chứa thông tin về:

- **Đơn hàng (Orders):** Thông tin về trạng thái đơn hàng, thời gian mua, duyệt, giao hàng...
- **Khách hàng (Customer_List):** Thông tin về ngày đăng ký, ngày đặt hàng đầu tiên, địa chỉ, độ tuổi, giới tính...
- **Chi tiết đơn hàng (Order_Items):** Danh sách sản phẩm, giá, phí vận chuyển, ngày giao hàng dự kiến...
- **Sản phẩm (Products):** Thông tin về danh mục, kích thước, trọng lượng sản phẩm...
- **Đánh giá đơn hàng (Order_Reviews):** Điểm đánh giá, tiêu đề và nội dung bình luận, thời gian đánh giá...

Dữ liệu này đến từ 338 thành phố tại 28 quốc gia, với 32.951 sản phẩm thuộc 72 danh mục khác nhau. Mục tiêu của bài thực hành là sử dụng Spark DataFrame để thực hiện các phân tích bán hàng và tiếp thị.



Hãy sử dụng Spark DataFrame thực hiện các yêu cầu bên dưới:

1. Hãy đọc dữ liệu từ các file csv, sử dụng tự suy ra kiểu dữ liệu cho mỗi cột.
2. Thống kê tổng số đơn hàng, số lượng khách hàng và người bán.
3. Phân tích số lượng đơn hàng theo quốc gia, sắp xếp theo thứ tự giảm dần.
4. Phân tích số lượng đơn hàng nhóm theo năm, tháng đặt hàng (Hiển thị theo năm tăng dần, tháng giảm dần)
5. Thống kê điểm đánh giá trung bình, số lượng đánh giá theo từng mức (ví dụ: 1 đến 5).
   Lưu ý: Cần xử lý các giá trị ngoại lệ và NULL trong cột Review_Score

Chọn 1 trong các câu sau để làm:

6. Tính doanh thu (giá sản phẩm + phí vận chuyển) trong năm 2024 và nhóm theo danh mục sản phẩm
7. Xác định sản phẩm có số lượng bán ra cao nhất và tính điểm đánh giá trung bình cho từng sản phẩm
8. Tính toán hiệu số giữa ngày giao hàng thực tế (Order_Delivered_Carrier_Date) và ngày giao hàng dự kiến (ví dụ: Shipping_Limit_Date từ bảng Order_Items) để đánh giá hiệu suất giao hàng.
9. Nhóm khách hàng dựa trên số lượng đơn hàng, giá trị trung bình của đơn hàng và tần suất mua sắm.
10. Xếp hạng các seller dựa trên tổng doanh thu và số lượng đơn hàng bán được.