

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN



Môn học: DATA WAREHOUSE

ĐỀ TÀI:

**XÂY DỰNG KHO DỮ LIỆU PHÂN TÍCH XU HƯỚNG KHÁCH
HÀNG CỦA ZOMATO APP**

GVHD: Gv.Nguyễn Văn Thành

Nhóm thực hiện đồ án: Nhóm 14

Họ Và Tên	MSSV
Đàm Trọng Hải Dương	20142481
Trần Sĩ Nguyên	21133059
Phan Cao Bằng	21133006

TP Hồ Chí Minh, tháng 05 năm 2024

LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin được gửi lời cảm ơn đến Thầy Nguyễn Văn Thành - Giảng viên phụ trách môn Kho Dữ Liệu – trường đại học Sư Phạm Kỹ Thuật Thành Phố Hồ Chí Minh.

Trong quá trình nhóm tụi em thực hiện làm đồ án đã nhận được nhiều sự giúp đỡ từ Thầy. Thầy đã cung cấp đầy đủ kiến thức, chỉ bảo và đóng góp những ý kiến quý báu giúp tụi em có thể hoàn thành được đồ án của mình một cách tốt nhất.

Sau một quá trình dài học tập và tìm hiểu thì nhóm chúng em đã thực hiện đồ án “Xây dựng và khai thác kho dữ liệu về thông tin quản lý nhà hàng của ứng dụng Zomato”. Trong quá trình thực hiện đồ án, dựa trên kiến thức được Thầy cung cấp qua các buổi học lý thuyết cũng như thực hành trên lớp, kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất. Tuy nhiên, đồ án còn chưa được hoàn thiện và có nhiều sai sót.

Nhóm rất mong nhận được sự góp ý từ Thầy nhằm rút ra những kinh nghiệm quý báu và hoàn thiện vốn kiến thức để nhóm có thể hoàn thành những đồ án khác trong tương lai.

NHẬN XÉT TỪ GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

PHÂN CÔNG CÔNG VIỆC

Công việc	Đàm Trọng Hải Dương	Trần Sĩ Nguyên	Phan Cao Bằng	Mức độ hoàn thành
Chọn dataset	✓	✓	✓	100%
Nghiên cứu tập dữ liệu	✓	✓	✓	100%
Chọn dữ liệu phù hợp	✓	✓	✓	100%
Mô tả bài toán	✓	✓	✓	100%
Tiền xử lý dữ liệu	✓	✓	✓	100%
Thiết kế data warehouse	✓	✓	✓	100%
ETL	✓	✓	✓	100%
Tạo OLAP cube	✓	✓	✓	100%
SSAS	✓	✓	✓	100%
Viết báo cáo	✓	✓	✓	100%
Trực quan lên Power BI	✓	✓	✓	100%

Mục Lục

.....	2
.....	2
I. TỔNG QUAN VỀ TẬP DỮ LIỆU	6
1. Lý do hình thành dự án:.....	6
1.1. Đặt vấn đề:	6
1.2. Giải pháp:.....	6
1.3. Mục tiêu và ý nghĩa của đề án:	6
2. Giới thiệu tổng quan về tập dữ liệu:	6
2.1. Nguồn dữ liệu sử dụng:.....	6
2.2. Mô tả chi tiết dữ liệu:	6
2.2.1. Tập dữ liệu:	7
2.2.2. Mô tả chi tiết các thuộc tính trong tập dữ liệu:	7
2.3. Công cụ sử dụng trong đề án:	10
II. Thiết kế xây dựng cơ sở dữ liệu tác nghiệp (OLAP)	10
1. Tiền xử lý dữ liệu:	10
1.1. Chọn các thuộc tính cần thiết cho bản Fact chính:.....	10
1.2. Tạo bảng Dim:	12
2. Thiết kế DataWarehouse:.....	14
2.1. Thiết kế StarSchema cho Fact chính:.....	14
III. Truyền dữ liệu vào các Dim và Fact (SQL)	16
3. Thực hiện các bước chuẩn bị trên SQL server:	16
3.1. Tạo các dim và fact:	16
3.2. Truyền dữ liệu từ csv vào database, và các fact:.....	18
3.2.1. Truyền dữ liệu vào các Dim	18
3.2.2. Truyền dữ liệu vào 2 Fact:	19
IV. Tích hợp dữ liệu vào kho (SSIS)	21
1. Control Flow:.....	21
1.1. Control Flow của các bảng Dim:	21

1.3.	Control Flow của 2 bảng Fact:	26
2.	Data Flow:	27
2.1.	Data Flow của Fact Sales:	27
2.2.	Data Flow của FactRestaurantPerformance:	28
V.	Thiết kế Cube (SSAS):	29
1.	Data SourceView:	29
2.	Cube:	29
VI.	Trả lời cho các câu hỏi đã đặt ra:	30
VII.	Trực quan hóa (Power BI)	33
	Tài liệu tham khảo:	37

I. TỔNG QUAN VỀ TẬP DỮ LIỆU

1. Lý do hình thành dự án:

1.1. Đặt vấn đề:

Trong thời đại kỹ thuật số, dữ liệu đóng vai trò cực kỳ quan trọng trong việc định hình các chiến lược kinh doanh và đưa ra quyết định chính xác. Ứng dụng Zomato, một nền tảng nổi tiếng trong lĩnh vực dịch vụ ăn uống và đánh giá nhà hàng, thu thập một lượng lớn dữ liệu từ người dùng và các nhà hàng trên toàn thế giới. Việc xây dựng một kho dữ liệu từ tập dữ liệu của Zomato có thể cung cấp những hiểu biết sâu sắc về xu hướng ẩm thực, hành vi người dùng, và hiệu quả kinh doanh của các nhà hàng.

1.2. Giải pháp:

Tích hợp và tổ chức dữ liệu: Thu thập, làm sạch và tổ chức lại dữ liệu từ Zomato để tạo thành một kho dữ liệu nhất quán và dễ truy xuất.

Phân tích và trực quan hóa dữ liệu: Sử dụng các công cụ phân tích và trực quan hóa để khám phá các mẫu dữ liệu và xu hướng quan trọng.

Hỗ trợ ra quyết định: Cung cấp các thông tin hữu ích cho các bên liên quan như nhà hàng, người dùng và các nhà quản lý của Zomato để cải thiện dịch vụ và trải nghiệm khách hàng.

Để có được những thông tin hữu ích, ta sẽ đặt ra một số câu hỏi:

- Top 10 nhà hàng có lợi nhuận cao nhất.
- Top 10 nhà hàng bị lỗ vốn.
- Những nhà hàng nào có số lượng đánh giá và xếp hạng cao nhất?

1.3. Mục tiêu và ý nghĩa của đề án:

Tạo ra một kho dữ liệu tích hợp và nhất quán đồng thời cung cấp các công cụ phân tích và trực quan hóa dữ liệu. Đề mạng đến những thông tin hữu ích, hỗ trợ doanh nghiệp ra quyết định về các định hướng trong tương lai. Đồng thời đảm bảo tính an toàn và bảo mật cho dữ liệu.

2. Giới thiệu tổng quan về tập dữ liệu:

2.1. Nguồn dữ liệu sử dụng:

- Nguồn dữ liệu được thu thập từ Kaggle và Github:

<https://www.kaggle.com/datasets/anas123siddiqui/zomato-database>

https://github.com/kayazay/zomato_analytics

- Bao gồm 5 tập dữ liệu tương ứng với 4 tập dữ liệu (menu, restaurant, users, orders) ở Kaggle và 1 tập dữ liệu ở Github (dim_veg_food)

2.2. Mô tả chi tiết dữ liệu:

Cơ sở dữ liệu ứng dụng giao đồ ăn Zomato app là một tập hợp toàn diện các bảng lưu trữ tất cả thông tin quan trọng liên quan đến ứng dụng giao đồ ăn. Nó bao gồm thông tin về đơn đặt hàng của người dùng, các món ăn có sẵn trên ứng dụng, thực đơn của các nhà hàng khác nhau, chính các nhà hàng và người dùng đã đăng ký trên ứng dụng. Các bảng được liên kết với nhau và lưu trữ thông tin cụ thể, cho phép truy xuất dữ liệu hiệu quả.

2.2.1. Tập dữ liệu:

Tập dữ liệu bao gồm 5 bảng khác nhau:

Users: gồm 100001(dòng) * 11(cột) với mỗi dòng là thông tin của user sử dụng app.

Menu: gồm 1037783(dòng) * 6(cột) với mỗi dòng là thông tin về việc buôn bán của nhà hàng.

Restaurant: gồm 148542(dòng) * 11(cột) với mỗi dòng là thông tin nhà hàng.

Orders: gồm 148542(dòng) * 7 (cột) với mỗi dòng là thông tin về số lượng order.

Ingrediant: gồm 367030(dòng) * 5(cột) với mỗi dòng là thông tin về tiêu thụ nguyên liệu.

2.2.2. Mô tả chi tiết các thuộc tính trong tập dữ liệu:

Tập dữ liệu	Tên thuộc tính	Mô tả
Restaurant.csv	id	Mã định danh duy nhất của nhà hàng.
	name	Tên của nhà hàng.
	city	Thành phố nơi nhà hàng đặt tại.
	rating	Điểm đánh giá của nhà hàng..
	Rating_count	Số lượng đánh giá mà nhà hàng đã nhận được
	cost	Giá trung bình của một bữa ăn tại nhà hàng.
	cuisine	Loại hình ẩm thực mà nhà hàng phục vụ.
	Lic_no	Số giấy phép kinh doanh của nhà hàng.

	link	Liên kết đến nhà hàng trên một trang web hoặc ứng dụng.
	adress	Địa chỉ vị trí của nhà hàng.
	menu	Đường dẫn đến tệp chứa menu của nhà hàng.
User.csv	user_id	Định danh duy nhất cho mỗi người dùng.
	name	Tên của người dùng.
	email	Địa chỉ email của người dùng.
	password	Mật khẩu liên kết với tài khoản của người dùng.
	Age	Tuổi của người dùng.
	Gender	Giới tính của người dùng
	Marital Status	Tình trạng hôn nhân của người dùng.
	Occupation	Nghề nghiệp hoặc chức vụ của người dùng.
	Monthly Income	Thu nhập hàng tháng của người dùng.
	Educational Qualifications	Trình độ học vấn cao nhất của người dùng
	Family size	Số thành viên trong gia đình của người dùng
Menu.csv	total_spent	Tổng số tiền đã chi tiêu bởi người dùng
	menu_id	Mã số duy nhất định danh cho mỗi mục menu.
	r_id	Mã số định danh của nhà hàng tương ứng.
	f_id	Mã số định danh của món ăn trong menu.

	cuisine	Loại hình ẩm thực của món ăn.
	price	Giá của món ăn.
	total_sold	Tổng số lượng món ăn đã bán.
Orders.csv	order_id	Mã số định danh duy nhất cho mỗi đơn đặt hàng.
	order_date	Ngày đặt hàng.
	sales_qty	Số lượng sản phẩm được bán trong đơn hàng.
	sales_amount	Tổng số tiền từ đơn hàng.
	currency	Đơn vị tiền tệ sử dụng trong giao dịch.
	user_id	Mã số định danh của người dùng đặt hàng.
	r_id	Mã số định danh của nhà hàng từ đó đặt hàng.
Dim_food_veg.csv	ID	Mã số định danh duy nhất cho mỗi nguyên liệu
	Item	Tên của mặt hàng.
	VEG	Trạng thái thực phẩm, có thể là "Vegetarian" (Chay) hoặc "Non-vegetarian" (Không chay).
	Item Cnt	Số lượng mặt hàng.
	total_used	Tổng số lượng mặt hàng đã sử dụng.

2.3. Công cụ sử dụng trong đồ án:

Visual Studio: Tích hợp các công nghệ

- Microsoft DataTools IntegrationServices (SSIS)
- Microsoft.DataTools.AnalysisServices (SSAS)

SQL Server 2019

Ngôn ngữ truy vấn:SQL

II. Thiết kế xây dựng cơ sở dữ liệu tác nghiệp (OLAP)

1. Tiền xử lý dữ liệu:

1.1. Chọn các thuộc tính cần thiết cho bản Fact chính:

Fact Sales

Tập dữ liệu	Tên thuộc tính	Mô tả
orders.csv	Order_id	Mã số định danh duy nhất cho mỗi đơn hàng.
	Order_date	Ngày đặt hàng.
	User_id	Mã số định danh của người dùng đặt hàng.
	R_id	Mã số định danh của nhà hàng từ đó đặt hàng.
	Sales_qty	Số lượng sản phẩm được bán trong đơn hàng.
	Sales_amont	Tổng số tiền từ đơn hàng.
	Currency	Đơn vị tiền tệ sử dụng trong giao dịch.
Menu.csv	Menu_id	Mã số định danh duy nhất cho mỗi mục menu.
	F_id	Mã số định danh của món ăn trong menu.
Restaurant.csv	cost	Giá của món ăn trong nhà hàng.
	rating	Điểm đánh giá của nhà hàng.

Tập dữ liệu	Tên thuộc tính	Mô tả
orders.csv	Order_id	Mã số định danh duy nhất cho mỗi đơn hàng.
	User_id	Mã số định danh của người dùng đặt hàng.
	Sales_qty	Số lượng sản phẩm được bán trong đơn hàng.
	Sales_amont	Tổng số tiền từ đơn hàng.
Restaurant.csv	cost	Giá của món ăn trong nhà hàng.
	rating	Điểm đánh giá của nhà hàng.
	id	Mã số định danh duy nhất cho mỗi nhà hàng.
	name	Tên của nhà hàng.
	city	Thành phố nơi nhà hàng đặt tại.

Fact RestaurantPerformance

Tập dữ liệu	Tên thuộc tính	Mô tả
orders.csv	Order_id	Mã số định danh duy nhất cho mỗi đơn hàng.
	Sales_qty	Số lượng sản phẩm được bán trong đơn hàng.
	Sales_amont	Tổng số tiền từ đơn hàng.
	Currency	Đơn vị tiền tệ sử dụng trong giao dịch.
Menu.csv	Price	Giá của món ăn.

Restaurant.csv	id	Mã số định danh duy nhất cho mỗi nhà hàng.
	name	Tên của nhà hàng.
	city	Thành phố nơi nhà hàng đặt tại.

1.2. Tạo bảng Dim:

Thực hiện xử lý các dữ liệu không khớp với data bằng cách quan sát và thực hiện xử lý ở excel.

Các bảng Dim thu được sau khi xử lý và làm sạch:

1.2.1. DimMenu:

	A	B	C	D	E	F	G
1		r_id	f_id	cuisine	price	total_sold	
2	0	567335	1	Beverages	40	28	
3	1	567335	2	Beverages	40	34	
4	2	158203	3	Beverages	65	64	
5	3	158203	4	Beverages	65	32	
6	4	158203	5	Beverages	65	19	
7	5	158203	6	Beverages	65	47	

1.2.2. DimUser:

1	user_id	name	email	password	Age	Gender	Marital Status	Occupation	Monthly Income	Educational Qualific	Family size	total_spent
2	1	Claire Ferguson	fordanthony@exam	NKz0fWDh15	20	Female	Single	Student	No Income	Post Graduate	4	0
3	2	Jennifer Young	ann96@example.cor	=+i5Q91jtl5	24	Female	Single	Student	Below Rs.10000	Graduate	3	0
4	3	Jermaine Roberson	uwalker@example.o	eO4GqGusF{	22	Male	Single	Student	Below Rs.10000	Post Graduate	3	0
5	4	Rachel Carpenter	kimberlypatterson@	ex:8J#E5RM1o	22	Female	Single	Student	No Income	Graduate	6	0
6	5	Shawn Parker	daniellebennett@ex	:8J#E5RM1o	22	Male	Single	Student	Below Rs.10000	Post Graduate	4	0
7	6	Timothy Clark	brettsantana@exam	qOJONAOY54	27	Female	Married	Employee	More than 50000	Post Graduate	2	0
8	7	Alexander Lucas	susan58@example.c	(^+21Yv3Uv	22	Male	Single	Student	No Income	Graduate	3	0
9	8	Christopher Curry	brookesmith@exam	^+5FP5zm(L	24	Female	Single	Student	No Income	Post Graduate	3	0
10	9	Daniel Mercado	imyers@example.co	eSDJ2tR0l#	23	Female	Single	Student	No Income	Post Graduate	2	0

1.2.3. DimRestaurant:

1	id	name	city	rating	rating_cou	cost	cuisine	lic_no	link	address	menu
2	567335	AB FOODS	Abohar		Too Few R	200	Beverages	2.21E+13	https://ww	AB FOODS Menu/567335.json	
3	531342	Janta Sweet	Abohar	4.4	50+ rating:	200	Sweets;Ba	1.21E+13	https://ww	Janta Sweet Menu/531342.json	
4	158203	theka coff	Abohar	3.8	100+ rating:	100	Beverages	2.21E+13	https://ww	theka coff Menu/158203.json	
5	187912	Singh Hut	Abohar	3.7	20+ rating:	250	Fast Food;	2.21E+13	https://ww	Singh Hut Menu/187912.json	
6	543530	GRILL MAS	Abohar		Too Few R	250	Italian-Am	1.21E+13	https://ww	GRILL MAS Menu/543530.json	
7	158204	Sam Uncle	Abohar	3.6	20+ rating:	200	Continent	2.21E+13	https://ww	Sam Uncle Menu/158204.json	
8	156588	shere punj	Abohar	4	100+ rating:	150	North Indi	2.21E+13	https://ww	shere punj Menu/156588.json	
9	244866	Shri Balaji	Abohar		Too Few R	100	North Indi	2.21E+13	https://ww	Shri Balaji Menu/244866.json	
10	156602	Hinglaj Ka	Abohar	4.2	20+ rating:	100	Snacks;Ch	2.21E+13	https://ww	Hinglaj Ka Menu/156602.json	

1.2.4. DimOrder:

1		order_date	sales_qty	sales_amo	currency	user_id	r_id
2	0	10/10/2017	100	41241	INR	49226	567335
3	1	5/8/2018	3	-1	INR	77359	531342
4	2	4/6/2018	1	875	INR	5321	158203
5	3	4/11/2018	1	583	INR	21343	187912
6	4	6/18/2018	6	7176	INR	75378	543530
7	5	11/20/2017	59	500	USD	34323	158204
8	6	11/22/2017	36	250	USD	33246	156588
9	7	11/23/2017	39	21412	INR	87420	244866
10	8	11/27/2017	35	19213	INR	31017	156602

1.2.5. DimIngredient:

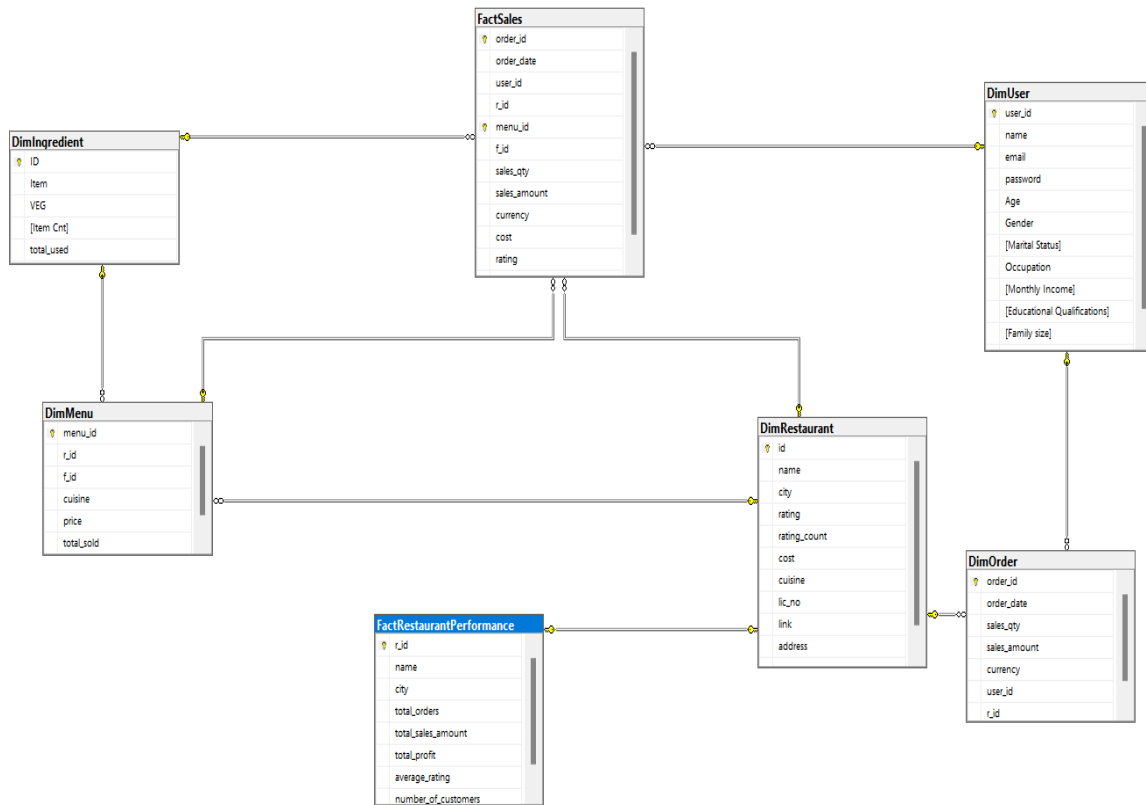
	A	B	C	D	E	F
1		Item	VEG	Item Cnt	total_used	
2	1	Egg	Non-veget	11K	0	
3	2	Rice	Non-veget	15K	0	
4	3	Noodles	Vegetarian	6K	0	
5	4	Paneer	Vegetarian	28K	0	
6	5	Chilli	Vegetarian	9K	0	
7	6	Noodles	Vegetarian	6K	0	
8	7	Dal	Vegetarian	5K	0	
9	8	Masala	Vegetarian	16K	0	

2. Thiết kế DataWarehouse:

2.1. Thiết kế StarSchema cho Fact chính:

Tiến hành tạo star schema dùng SQL. Star schema gồm các bảng DimIngredient, DimOrder, DimUser, DimMenu, DimRestaurant, FactRestaurantPerformance và FactSales. Ta sẽ thực hiện các bước sau:

Đây là lược đồ **StarSchema**:



III. Truyền dữ liệu vào các Dim và Fact (SQL)

3. Thực hiện các bước chuẩn bị trên SQL server:

3.1. Tạo các dim và fact:

Tạo DimIngredient

```
-- Create table DimIngredient
CREATE TABLE DimIngredient (
    [ID] INT NOT NULL PRIMARY KEY,
    [Item] VARCHAR(255) NULL,
    [VEG] VARCHAR(20) NULL,
    [Item Cnt] VARCHAR(255) NULL,
    [total_used] INT DEFAULT 0 NULL
) ON [PRIMARY];
```

Tạo DimOrder

```
-- Create table DimOrder
CREATE TABLE DimOrder (
    [order_id] INT NOT NULL PRIMARY KEY,
    [order_date] DATE NULL,
    [sales_qty] INT NULL,
    [sales_amount] float NULL,
    [currency] VARCHAR(3) NULL,
    [user_id] INT NOT NULL,
    [r_id] INT NOT NULL,
    FOREIGN KEY (user_id) REFERENCES DimUser(user_id),
    FOREIGN KEY (r_id) REFERENCES DimRestaurant(id)
) ON [PRIMARY];
```

Tạo DimUser

```
-- Create table DimUser
CREATE TABLE DimUser (
    [user_id] INT NOT NULL PRIMARY KEY,
    [name] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [email] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [password] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [Age] INT NULL DEFAULT 0,
    [Gender] VARCHAR(10) DEFAULT 'Unknown' NULL,
    [Marital Status] VARCHAR(20) DEFAULT 'Unknown' NULL,
    [Occupation] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [Monthly Income] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [Educational Qualifications] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [Family size] INT NULL DEFAULT 0,
    [total_spent] DECIMAL(10,2) DEFAULT 0 NULL
) ON [PRIMARY];
```


Tạo DimMenu

```
-- Create table DimMenu
CREATE TABLE DimMenu (
    [menu_id] int NOT NULL PRIMARY KEY DEFAULT 'N/A',
    [r_id] INT NOT NULL DEFAULT 0,
    [f_id] INT NOT NULL DEFAULT 0,
    [cuisine] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [price] DECIMAL(10,2) DEFAULT 0 NULL,
    [total_sold] INT DEFAULT 0 NULL,
    FOREIGN KEY (r_id) REFERENCES DimRestaurant(id),
    FOREIGN KEY (f_id) REFERENCES DimIngredient(ID)
) ON [PRIMARY];
```

Tạo DimRestaurant

```
-- Create table DimRestaurant
CREATE TABLE DimRestaurant (
    [id] INT NOT NULL PRIMARY KEY DEFAULT 0,
    [name] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [city] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [rating] float DEFAULT 0 NULL,
    [rating_count] VARCHAR(255) DEFAULT 0 NULL,
    [cost] float DEFAULT 0 NULL,
    [cuisine] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [lic_no] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [link] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [address] VARCHAR(255) DEFAULT 'Unknown' NULL,
    [menu] VARCHAR(255) DEFAULT 'Unknown' NULL,
) ON [PRIMARY];
```

Tạo FactSales

```
-- Create table FactSales
CREATE TABLE FactSales (
    [order_id] INT NOT NULL,
    [order_date] DATE NOT NULL,
    [user_id] INT NOT NULL,
    [r_id] INT NOT NULL,
    [menu_id] int NOT NULL,
    [f_id] INT NOT NULL,
    [sales_qty] INT NOT NULL,
    [sales_amount] float NOT NULL,
    [currency] VARCHAR(3) NOT NULL,
    [cost] float NOT NULL,
    [rating] float NOT NULL,
    [profit] float NOT NULL,
    CONSTRAINT PK_FactSales PRIMARY KEY NONCLUSTERED (order_id, menu_id),
    FOREIGN KEY (user_id) REFERENCES DimUser(user_id),
    FOREIGN KEY (r_id) REFERENCES DimRestaurant(id),
    FOREIGN KEY (menu_id) REFERENCES DimMenu(menu_id),
    FOREIGN KEY (f_id) REFERENCES DimIngredient(Id)
) ON [PRIMARY];
```

Tạo FactRestaurantPerformance

```
-- Create table FactRestaurantPerformance
CREATE TABLE FactRestaurantPerformance (
    [r_id] INT NOT NULL PRIMARY KEY,
    [name] VARCHAR(255) NULL,
    [city] VARCHAR(255) NULL,
    [total_orders] INT NULL,
    [total_sales_amount] FLOAT NULL,
    [total_profit] FLOAT NULL,
    [average_rating] FLOAT NULL,
    [number_of_customers] INT NULL,
    FOREIGN KEY (r_id) REFERENCES DimRestaurant(id)
) ON [PRIMARY];
```

Update Các bảng bằng hàm tính toán:

```
-- Tổng chi tiêu của người dùng (DimUser)
UPDATE DimUser
SET total_spent = (
    SELECT SUM(fs.sales_amount)
    FROM FactSales fs
    WHERE fs.user_id = DimUser.user_id
);

select * from FactSales
-- Tổng số lượng món ăn đã bán (DimMenu)
UPDATE DimMenu
SET total_sold = (
    SELECT SUM(fs.sales_qty)
    FROM FactSales fs
    WHERE fs.menu_id = DimMenu.menu_id
);

-- Tổng số lượng nguyên liệu đã sử dụng (DimIngredient)
UPDATE DimIngredient
SET total_used = (
    SELECT SUM(fs.sales_qty)
    FROM FactSales fs
    JOIN DimMenu dm ON fs.menu_id = dm.menu_id
    WHERE dm.f_id = DimIngredient.ID
);
```

3.2. Truyền dữ liệu từ csv vào database, và các fact:

3.2.1. Truyền dữ liệu vào các Dim

```
BULK INSERT DimRestaurant
FROM 'C:\Users\ASUS\Documents\Zalo Received Files\restaurant.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2 -- Use this if your CSV has a header row
);
```

```

BULK INSERT DimMenu
FROM 'C:\Users\ASUS\Documents\Zalo Received Files\menu.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2 -- Use this if your CSV has a header row
);

```

```

BULK INSERT DimUser
FROM 'C:\Users\ASUS\Documents\Zalo Received Files\users.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2 -- Use this if your CSV has a header row
);

```

```

BULK INSERT DimIngredient
FROM 'C:\Users\ASUS\Documents\Zalo Received Files\DIM_FOODS_VEG.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2 -- Use this if your CSV has a header row
);

```

```

BULK INSERT DimOrder
FROM 'C:\Users\ASUS\Documents\Zalo Received Files\orders.csv'
WITH (
    FIELDTERMINATOR = ',',
    ROWTERMINATOR = '\n',
    FIRSTROW = 2 -- Use this if your CSV has a header row
);

```



3.2.2. Truyền dữ liệu vào 2 Fact:

```

-- Insert data into FactSales from the dimension tables
INSERT INTO FactSales (
    order_id,
    order_date,
    user_id,
    r_id,
    menu_id,
    f_id,
    sales_qty,
    sales_amount,
    currency,

```

```

        cost,
        rating,
        profit
    )
SELECT
    do.order_id,
    do.order_date,
    do.user_id,
    do.r_id,
    dm.menu_id,
    dm.f_id,
    do.sales_qty,
    do.sales_amount,
    do.currency,
    dr.cost,
    dr.rating,
    (do.sales_amount - dr.cost) AS profit -- Example profit calculation
FROM
    DimOrder do
JOIN
    DimMenu dm ON dm.r_id = do.r_id -- Example join condition
JOIN
    DimRestaurant dr ON dr.id = do.r_id
JOIN
    DimIngredient di ON di.ID = dm.f_id
JOIN
    DimUser du ON du.user_id = do.user_id;

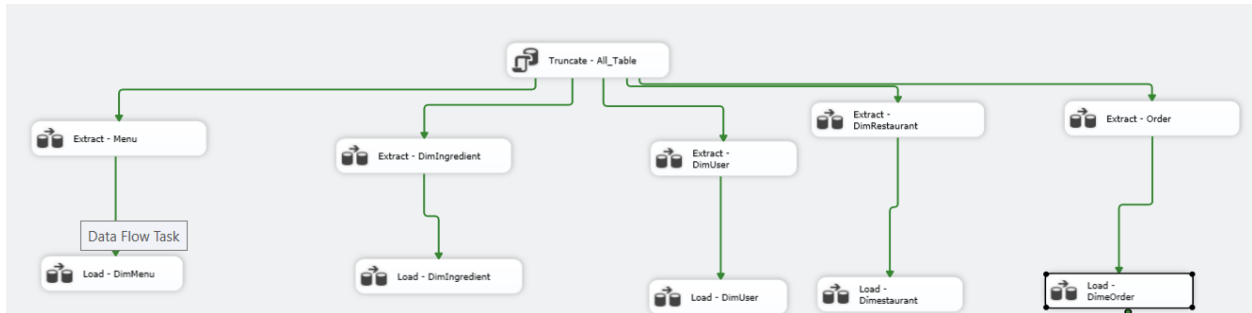
-- Insert data into FactRestaurantPerformance from the dimension tables
INSERT INTO FactRestaurantPerformance (r_id, name, city, total_orders,
total_sales_amount, total_profit, average_rating, number_of_customers)
SELECT
    r.id,
    r.name,
    r.city,
    COUNT(o.order_id) AS total_orders,
    SUM(o.sales_amount) AS total_sales_amount,
    SUM(o.sales_qty * m.price) AS total_profit,
    AVG(r.rating) AS average_rating,
    COUNT(DISTINCT o.user_id) AS number_of_customers
FROM
    DimRestaurant r
    LEFT JOIN DimOrder o ON r.id = o.r_id
    LEFT JOIN DimMenu m ON o.r_id = m.r_id
GROUP BY
    r.id,
    r.name,
    r.city;

```

IV. Tích hợp dữ liệu vào kho (SSIS)

1. Control Flow:

1.1. Control Flow của các bảng Dim:



1.2. Data Flow của các bảng Dim:

1.2.1. Data Flow DimIngredient:

Extract DimIngredient:

- Thực hiện tạo stgIngredient sau đó mapping giống hình bên dưới:

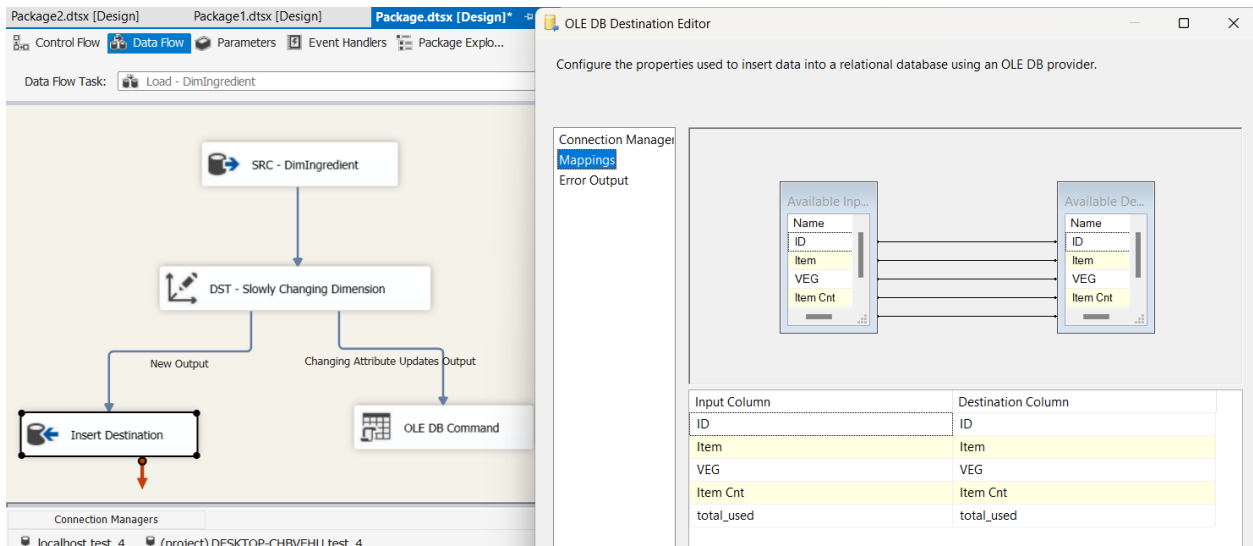
The image shows the SSIS Data Flow Task and the OLE DB Destination Editor for the 'DimIngredient' table.

Data Flow Task: The task is named 'Extract - DimIngredient'. It shows a source 'SRC - ExternalSource_DimIngredient' connected to a destination 'DST - DimIngredient'.

OLE DB Destination Editor: The editor is configured for the 'DST - DimIngredient' destination. It shows the mapping of input columns to destination columns.

Input Column	Destination Column
ID	ID
Item	Item
VEG	VEG
Item Cnt	Item Cnt
total_used	total_used

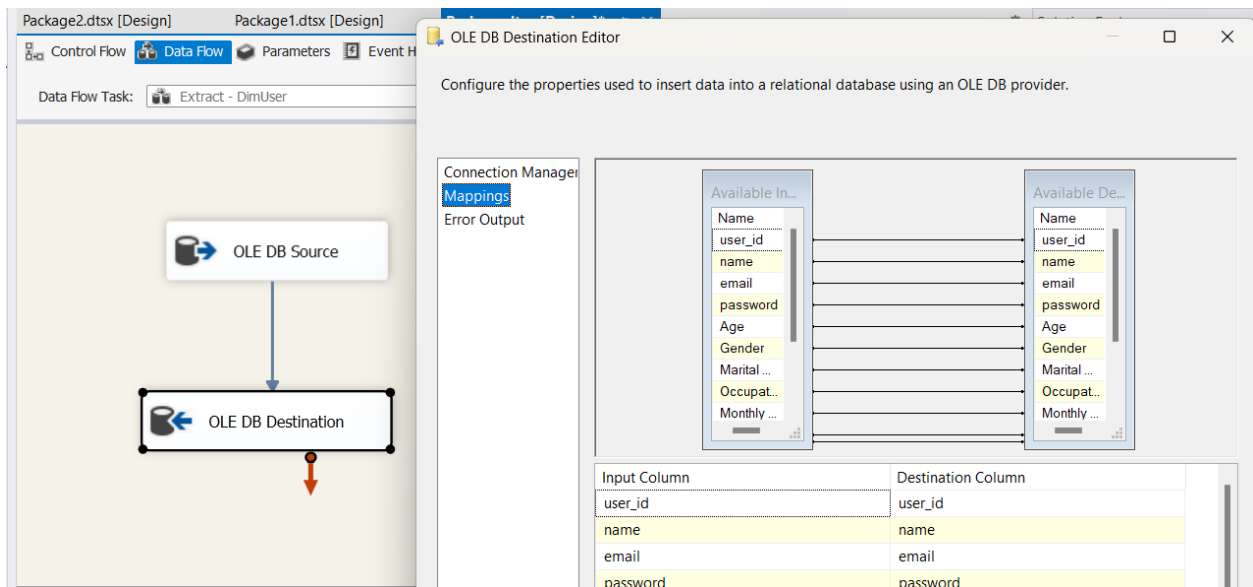
Load DimIngredient:



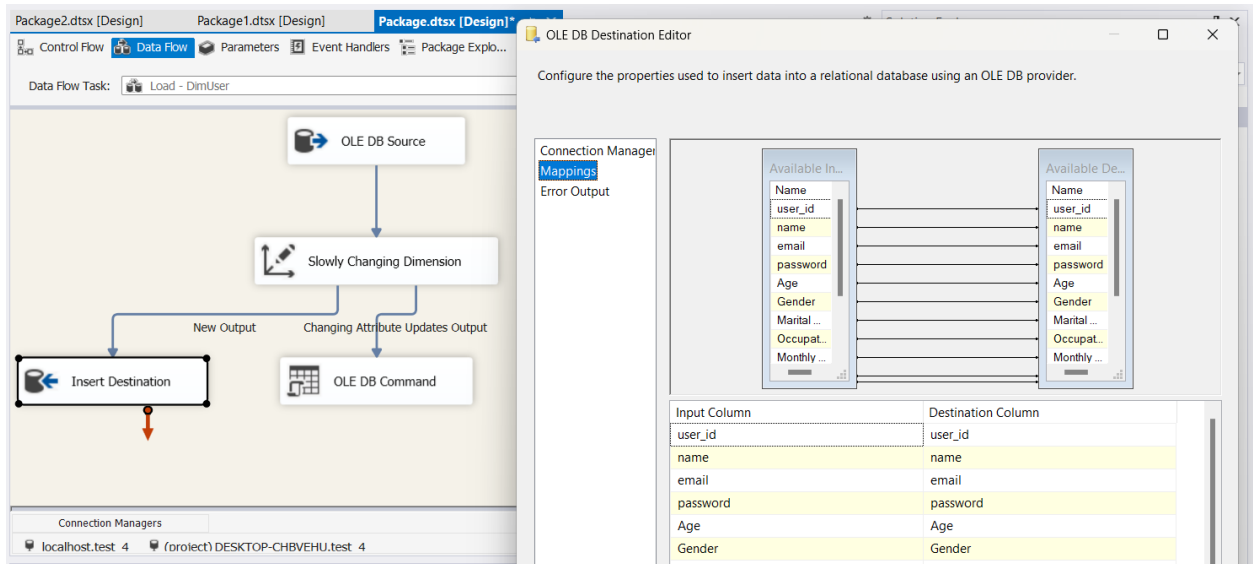
1.2.2. Data Flow DimUser:

Extract DimUser:

- Thực hiện tạo stgUser sau đó mapping giống hình bên dưới:



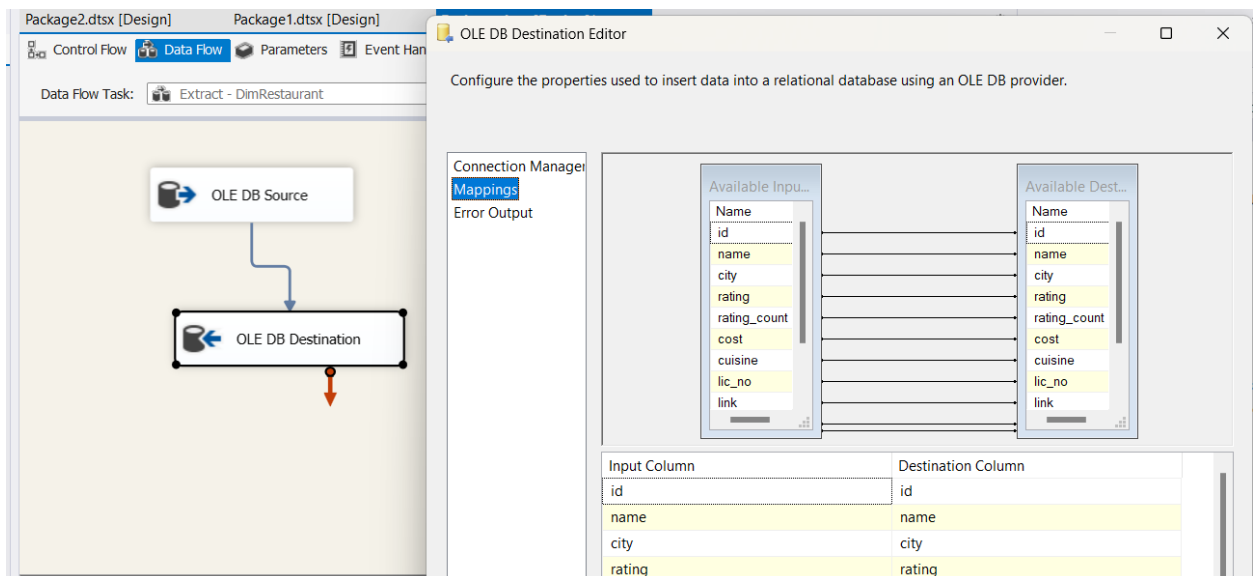
Load DimUser:



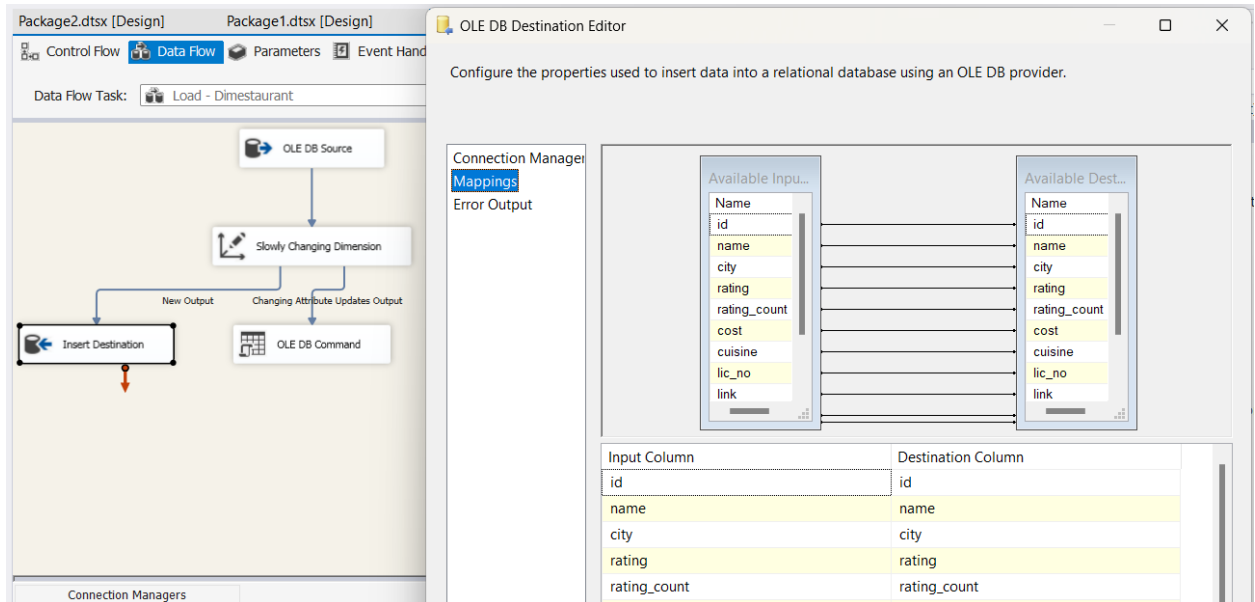
1.2.3. Data Flow DimRestaurant:

Extract DimRestaurant:

- Thực hiện tạo stgRestaurant sau đó mapping giống hình bên dưới:



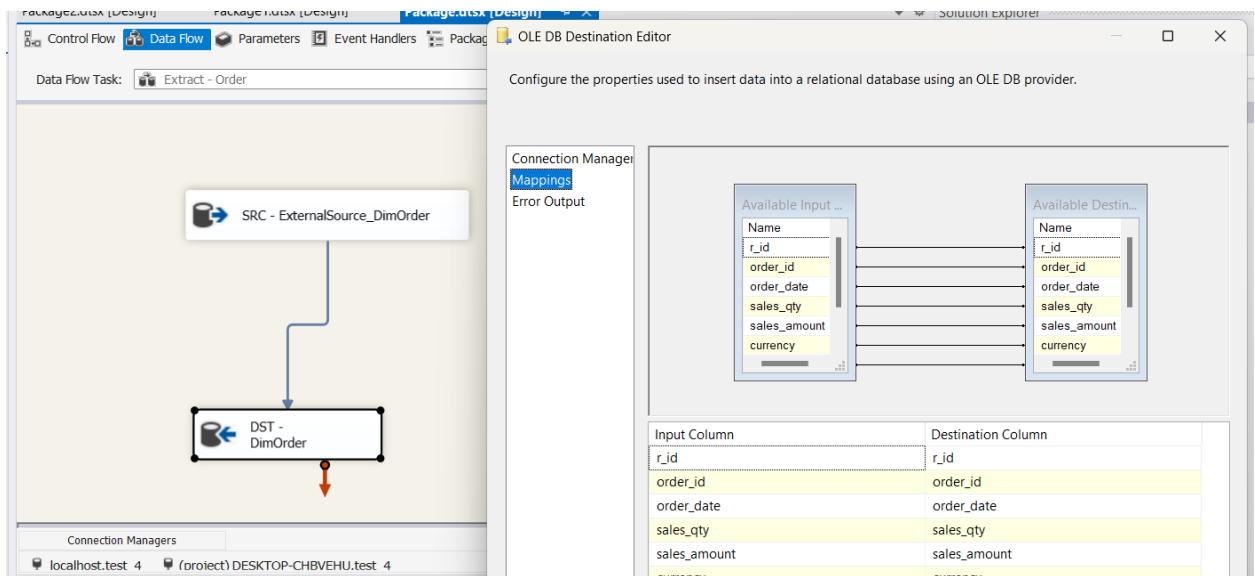
Load DimRestaurant:



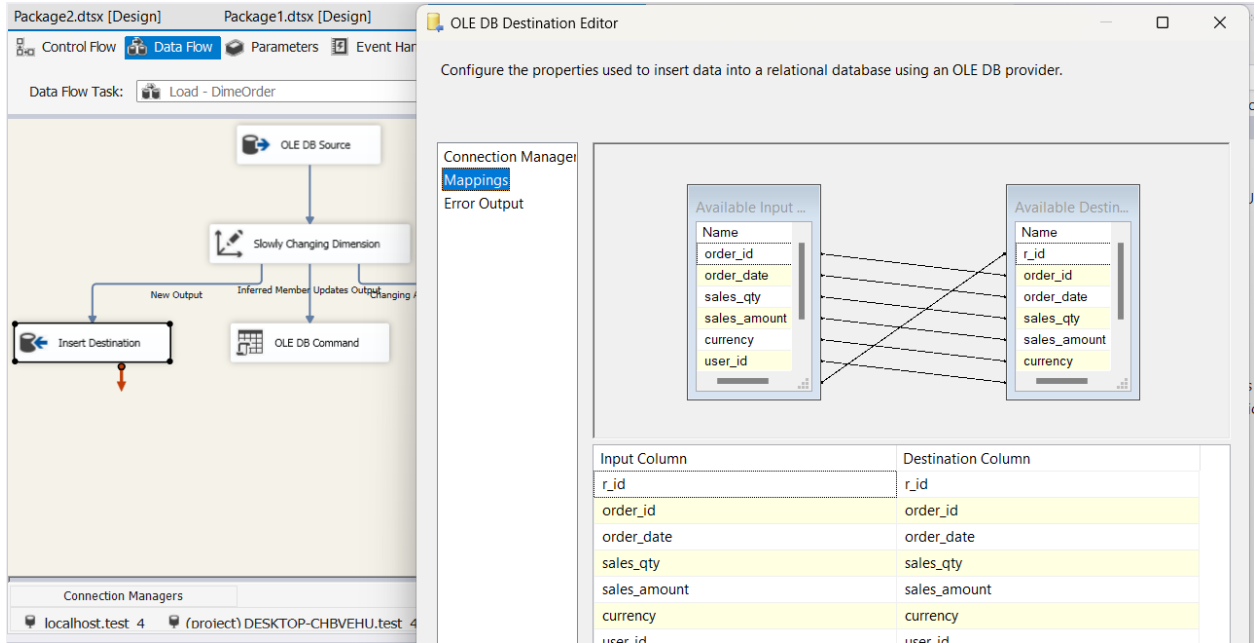
1.2.4. Data Flow DimOrder:

Extract DimOrder:

- Thực hiện tạo stgOrder sau đó mapping giống hình bên dưới:



Load DimOrder:



The image shows the SSIS Package Design window and the OLE DB Destination Editor for the 'Load - DimeOrder' task.

Package Design: The Data Flow Task 'Load - DimeOrder' contains an 'OLE DB Source' connected to a 'Slowly Changing Dimension' component. The 'Slowly Changing Dimension' component has three outputs: 'New Output', 'Inferred Member', and 'Updates Output'. The 'New Output' is connected to an 'Insert Destination' component. The 'Inferred Member' and 'Updates Output' are connected to an 'OLE DB Command' component.

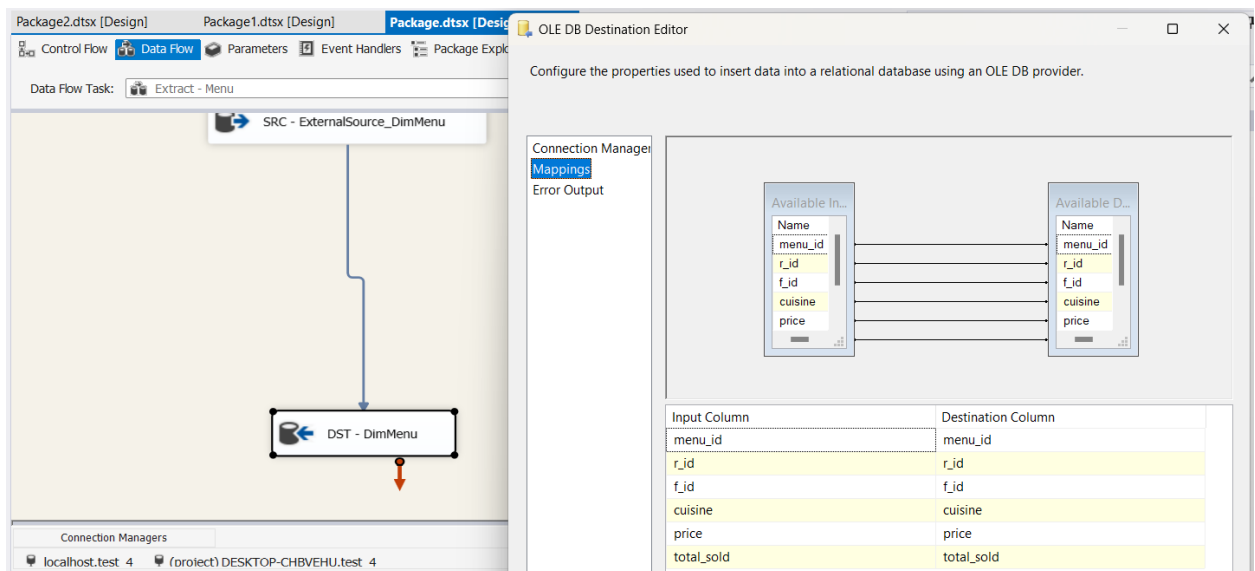
OLE DB Destination Editor: The 'Mappings' tab shows the column mappings between the 'Available Input Columns' and the 'Available Destination Columns'.

Input Column	Destination Column
r_id	r_id
order_id	order_id
order_date	order_date
sales_qty	sales_qty
sales_amount	sales_amount
currency	currency
user_id	user_id

1.2.5. Data Flow DimMenu:

Extract DimMenu:

- Thực hiện tạo stgMenu sau đó mapping giống hình bên dưới:



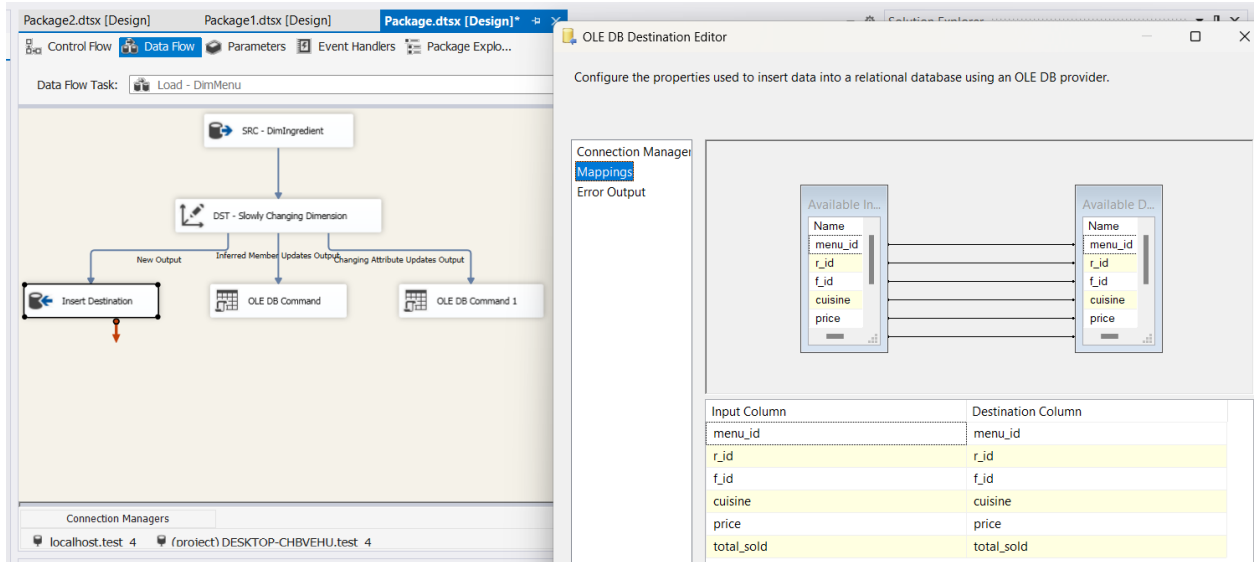
The image shows the SSIS Package Design window and the OLE DB Destination Editor for the 'Extract - Menu' task.

Package Design: The Data Flow Task 'Extract - Menu' contains an 'SRC - ExternalSource_DimMenu' component connected to a 'DST - DimMenu' component.

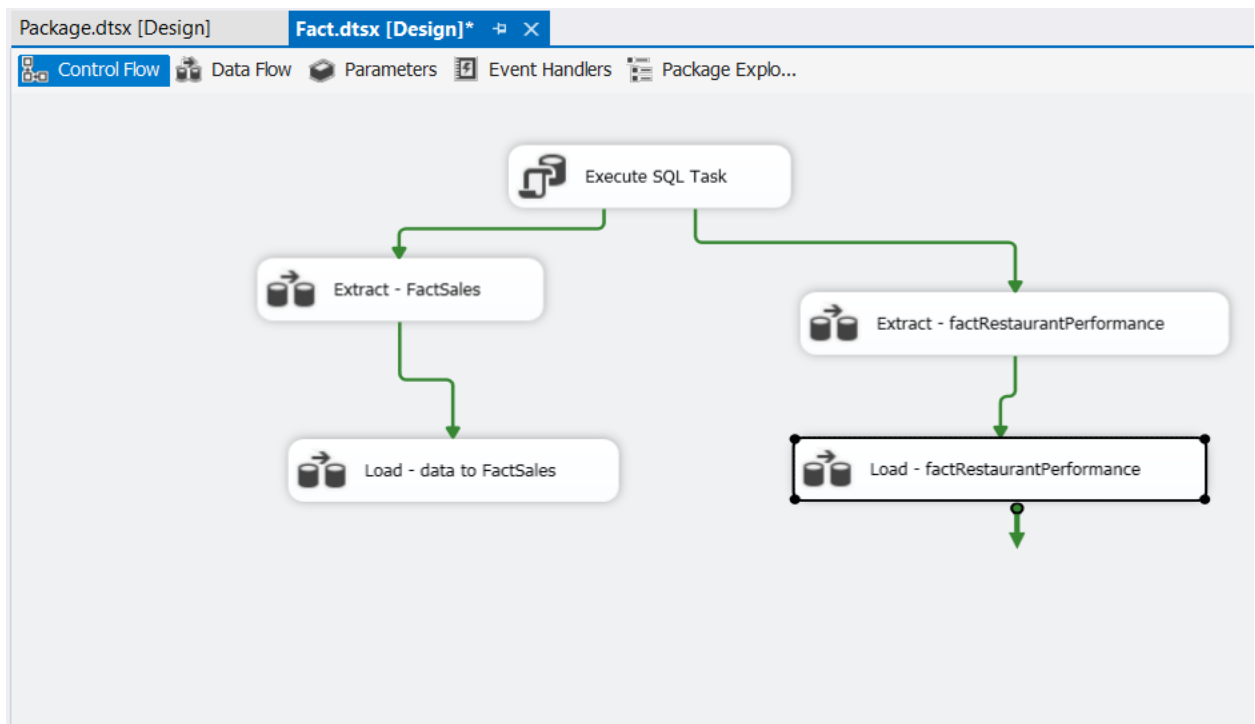
OLE DB Destination Editor: The 'Mappings' tab shows the column mappings between the 'Available Input Columns' and the 'Available Destination Columns'.

Input Column	Destination Column
menu_id	menu_id
r_id	r_id
f_id	f_id
cuisine	cuisine
price	price
total_sold	total_sold

Load DimMenu:



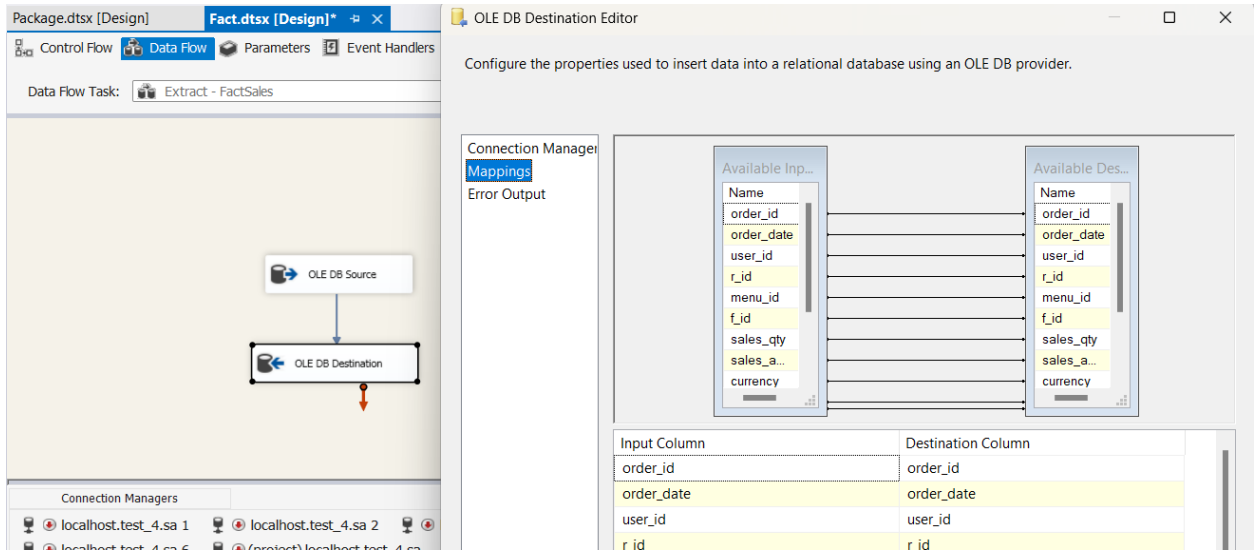
1.3. Control Flow của 2 bảng Fact:



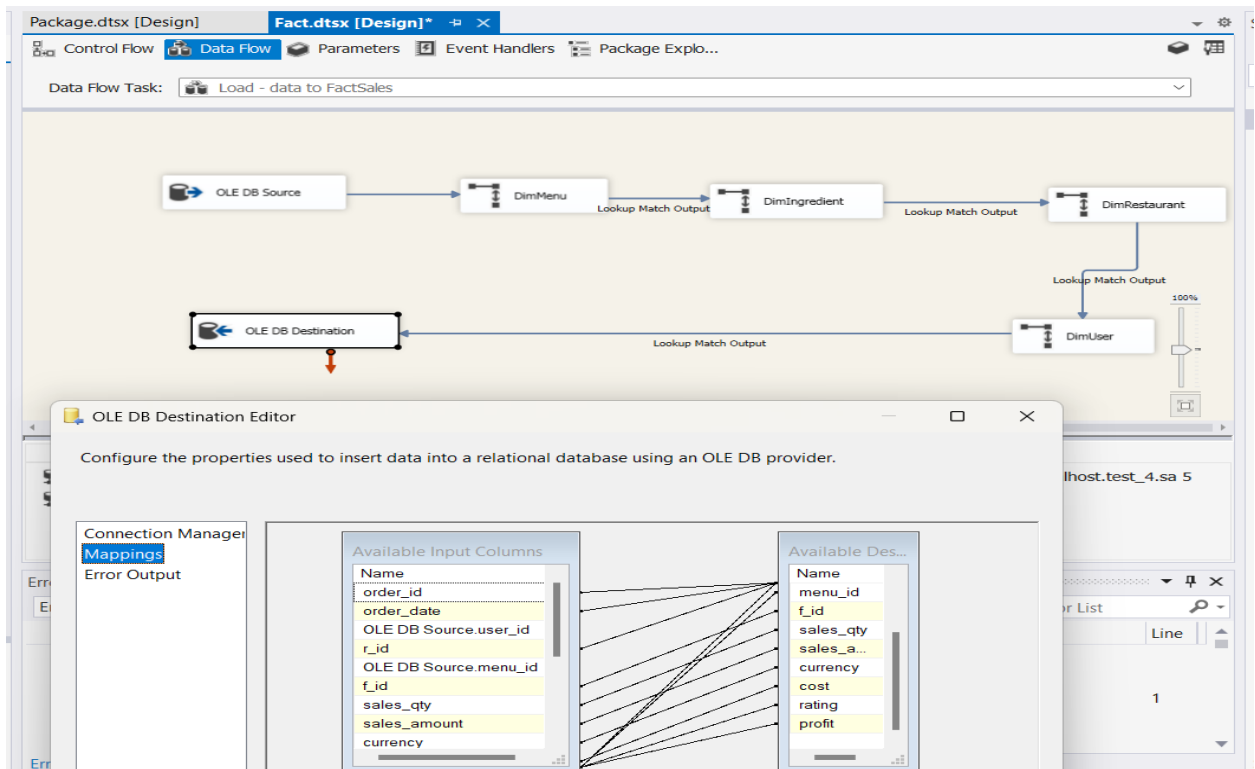
2. Data Flow:

2.1. Data Flow của Fact Sales:

Extract FactSales:

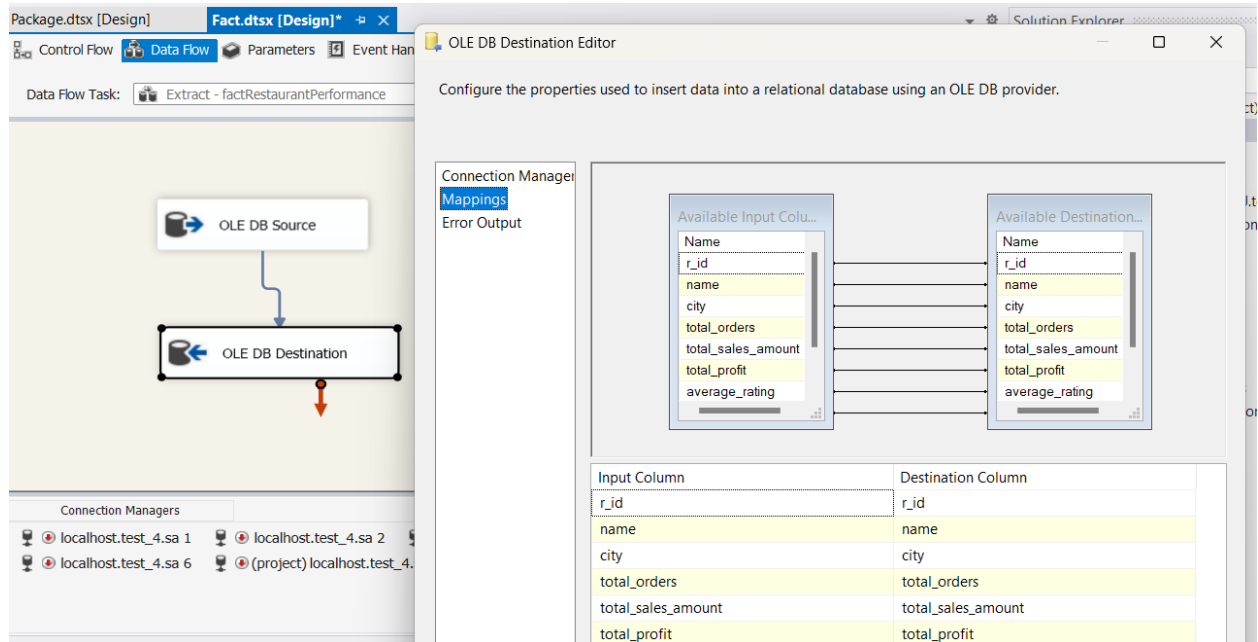


Load FactSales:

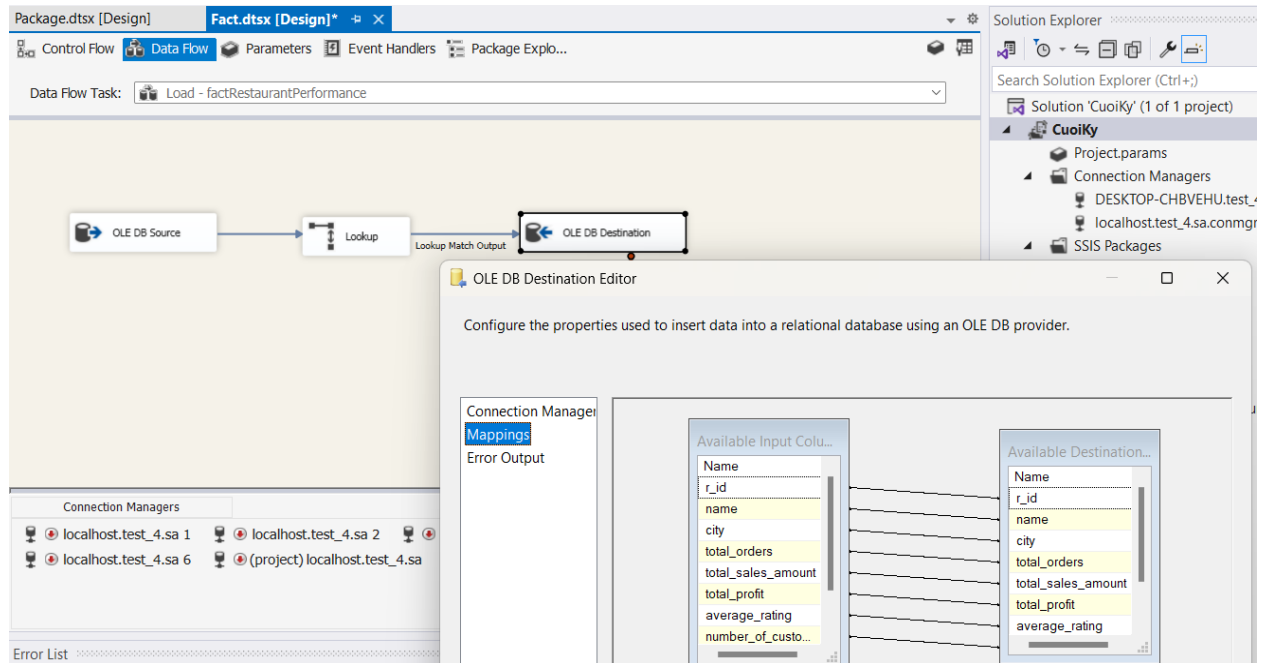


2.2. Data Flow của FactRestaurantPerformance:

Extrakt FactRestaurantPerformance:



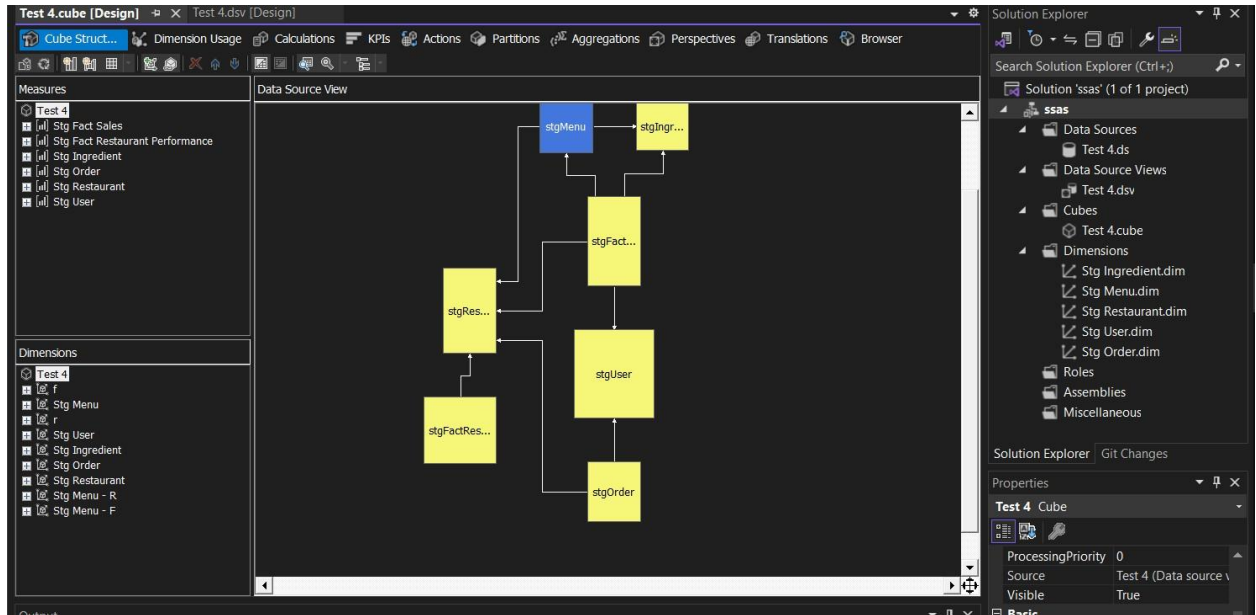
Load FactRestaurantPerformance:



V. Thiết kế Cube (SSAS)

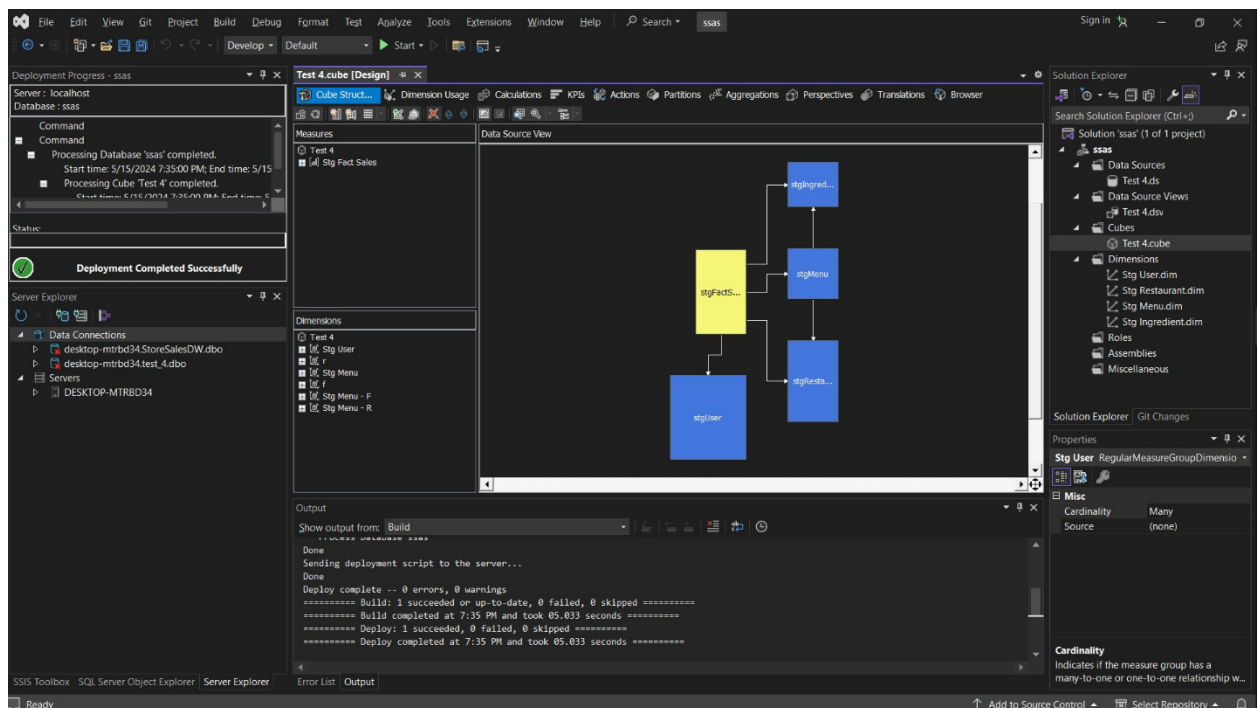
1. Data SourceView:

- Ở phần Data Source View chọn New Data Source View, chọn các bảng cần thiết và finish



2. Cube:

Ở phần Cube chọn New Cube, chọn Data Source View đã tạo , chọn các bảng độ đo, bảng fact, bảng dim cần thiết sau đó finish



VI. Trả lời cho các câu hỏi đã đặt ra

1. Lợi nhuận 10 nhà hàng có số lượt đánh giá cao nhất:

	A	B	C	D	E
1	ResId	Rating	Profit		
2	73719	953.4	1322275		
3	112399	899.1	2817342		
4	133352	1208	129860		
5	137701	1102	323930		
5	165532	925	121750		
7	183663	915.9	248784		
3	226299	997.2	105260		
9	251746	980.4	-83592		
0	332479	1315.8	78948		
1	531342	968	-44220		
2	Grand Total	10264.8	5020337		
3					
4					
5					

Dữ liệu cho thấy không có mối tương quan rõ ràng giữa Rating và Profit của các nhà hàng trên Zomato. Điều này cho thấy, việc nhà hàng được đánh giá cao trên Zomato chưa chắc đã đảm bảo lợi nhuận cao.

Ảnh hưởng của các yếu tố khác: Có nhiều yếu tố có thể tác động đến Profit của nhà hàng trên Zomato, chẳng hạn như: loại hình ẩm thực, giá cả, vị trí, chương trình khuyến mãi, chất lượng dịch vụ,... Việc một nhà hàng có Profit âm có thể do nhiều nguyên nhân, không chỉ đơn thuần là Rating thấp.

Hạn chế của dữ liệu: Dữ liệu chỉ phản ánh một phần hoạt động kinh doanh của các nhà hàng trên Zomato. Zomato chỉ là một trong số nhiều nền tảng đặt món trực tuyến, nên dữ liệu có thể không phản ánh toàn diện hiệu quả kinh doanh của nhà hàng.

Để có cái nhìn sâu sắc hơn về hoạt động của các nhà hàng trên Zomato, cần thu thập thêm dữ liệu về:

Tiêu chí lựa chọn 10 nhà hàng: Xác định rõ tiêu chí (top rating, top doanh thu,...) để đánh giá chính xác hơn ý nghĩa của dữ liệu.

Thông tin chi tiết về từng nhà hàng: Loại hình ẩm thực, mức giá, vị trí, số lượng đánh giá,...

Dữ liệu về hoạt động trên Zomato: Lượt xem, lượt đặt món, bình luận của khách hàng,...

Dữ liệu từ các nguồn khác: So sánh dữ liệu từ Zomato với các nền tảng đặt món khác, cũng như doanh thu thực tế của nhà hàng.

Bằng cách phân tích dữ liệu một cách đa chiều và toàn diện, chúng ta có thể hiểu rõ hơn về các yếu tố quyết định thành công của nhà hàng trên Zomato và đưa ra những chiến lược kinh doanh hiệu quả.

2. Top 10 khách hàng có chi tiêu cao nhất:

	A	B	C	D
1	User Id	Cost		
2	12121	115750		
3	13474	120000		
4	20513	174200		
5	38410	125191		
6	44604	114300		
7	54316	129600		
8	70364	138400		
9	75686	124800		
10	82668	117600		
11	99923	129000		
12	Grand Total	1288841		
13				

Phân bố chi tiêu: Có thể thấy sự chênh lệch về chi tiêu giữa các khách hàng trong Top 10. Khách hàng có User Id 20513 chi tiêu cao nhất (174200) trong khi User Id 44604 có chi tiêu thấp nhất (114300).

Hướng phát triển: Dữ liệu chỉ thể hiện chi tiêu của 10 khách hàng hàng đầu,ta có thể đưa ra những đãi ngộ, giảm giá và thăng hạng cho các khách hàng chi tiêu cao để có thể thu hút thêm nhiều tiềm lực từ họ, bên cạnh đó phân tích thêm về những nhà hàng, món ăn, yếu tố thu hút họ để phát triển và triển khai rộng trên nhiều khu vực hơn.

Để có thể phân tích sâu hơn, cần bổ sung:

Thông tin chi tiết về từng khách hàng: Ví dụ như lịch sử đặt món, loại món ăn ưa thích, tần suất sử dụng app,...

3. Top các nhà hàng bị lỗ vốn:

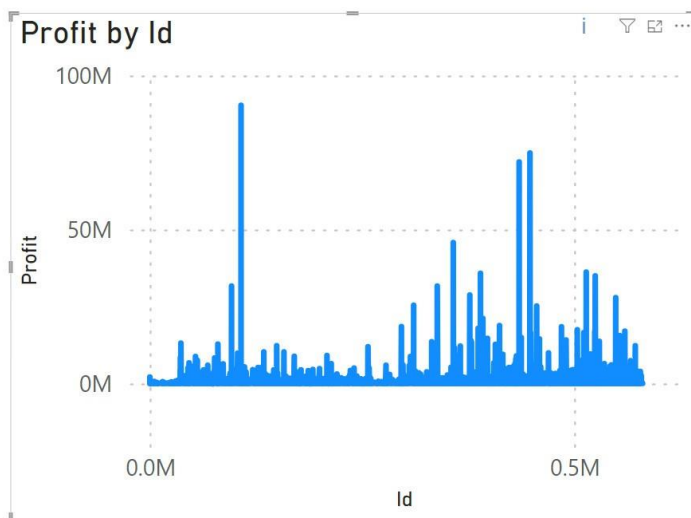
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Row Labels	Rating	Profit														
2	223	70.4	-4464														
3	227	15.6	-144														
4	232	55.9	-455														
5	237	323.4	-53438														
6	287	64.6	-3451														
7	297	266.5	-51480														
8	312	215.6	-15239														
9	425	304	-26828														
0	977	405.9	-71874														
1	1308	25.8	-1302														
2	2698	22	-3420														
3	3038	8.6	-1348														
4	3209	35.1	-882														
5	3306	0	-2730														
6	3454	176.3	-6437														
7	4979	208	-9984														
8	5350	300	-22350														
9	5770	453.6	-22248														
10	6842	58.8	-966														
11	7633	96	-2352														
12	7753	16.8	-1160														
13	10576	499.2	-5120														
14	14195	8	-148														
15	14415	49.2	-2256														
16	14837	718.2	-342														
17	14837	718.2	-342														

Mặc dù không có mối liên hệ giữ rating và ptofit do có những nhà hàng rating cao nhưng vẫn lỗ. Thì cũng mang lại những điều kiện để phân tích chẳng hạn có thể khảo sát xem các rating đó đa phần là do ai đánh giá, tại sao nhà hàng lỗ lại có lượng đánh giá cao, các nhà hàng có đánh giá ít thì khả năng tiếp cận của họ đến người tiêu dung đã đủ rộng chưa vì biết đâu họ buôn bán có tâm nhưng lại ít khách biết đến và dẫn đến lỗ.

Hướng phát triển là xem xét thêm các yếu tố đã nêu trên như khả năng tiếp cận, quảng cáo, vị trí, ngày thành lập (vì biết đâu do tuổi đời còn nhỏ nên chưa được sự tin tưởng của khách hàng hoặc còn nhiều sai sót) .

VII. Trực quan hóa (Power BI)

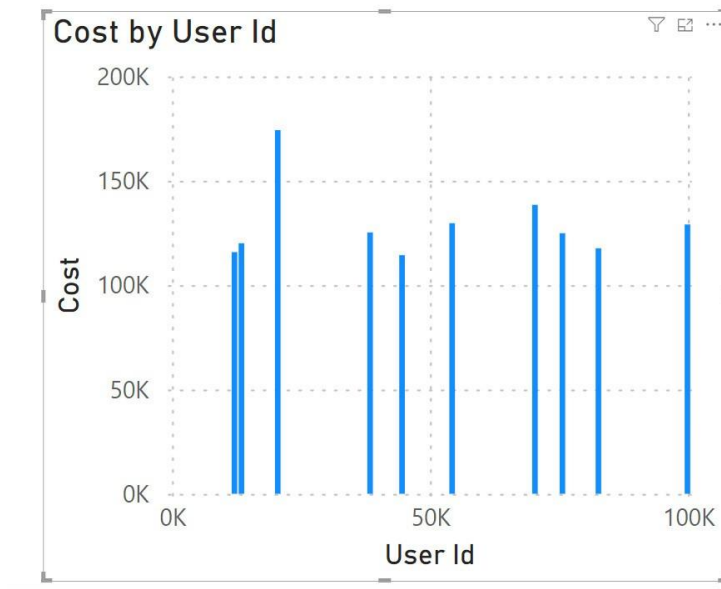
1. Lợi nhuận của tất cả nhà hàng:



Nắm bắt được tệp khách hàng có chỉ tiêu cao để tối ưu hóa dịch vụ, khuyến mãi và đãi ngộ với họ, tìm ra yếu tố tác động khi họ ủng hộ nhiều vào dịch vụ nhà hàng, các yếu tố đó có chủ quan hay khách quan nếu khách quan có thể tận dụng để phát triển.

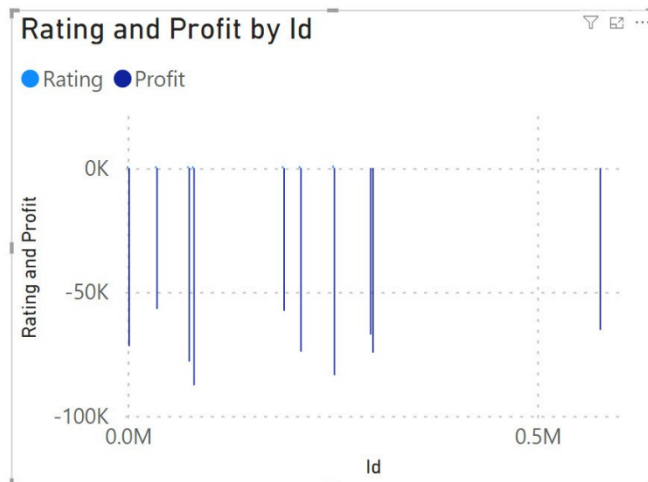
Nắm bắt được tệp khách hàng bình dân và giá rẻ để đưa ra nhiều đãi ngộ thu hút khách, ưu tiên các nhà hàng bình dân giá rẻ được xuất hiện quảng cáo đối với tệp khách này.

2. Top 10 khách hàng chi tiêu nhiều nhất:



Cũng như đã nói ở trên thì chúng ta sẽ chắc lọc ra để có phương án tối ưu giữ chân khách hàng và tạo tiềm lực cho nhà hàng.

3. Top 10 nhà hàng bị lỗ vốn:



Trực quan hơn về dữ liệu để có thể dễ dàng tìm ra hướng giải quyết phù hợp, có thể là chấn chỉnh lại dịch vụ, các món ăn, hay là nắm bắt mức độ yêu thích món ăn để cắt giảm hay tập trung đầu tư phát triển cho phù hợp với lợi thế.

VIII. TỔNG KẾT

1. Tổng Kết:

Dự án của bạn đã thành công trong việc:

Xây dựng kho dữ liệu: Bạn đã thu thập, làm sạch và tích hợp dữ liệu từ nhiều nguồn khác nhau của Zomato thành một kho dữ liệu nhất quán trên SQL Server, tạo nền tảng cho việc phân tích.

Thiết kế mô hình dữ liệu: Bạn sử dụng Star Schema để tổ chức dữ liệu, giúp cho việc truy vấn và phân tích hiệu quả hơn.

Trực quan hóa dữ liệu: Sử dụng Power BI, bạn đã tạo ra các biểu đồ trực quan, dễ hiểu để trình bày thông tin về lợi nhuận, chi tiêu của khách hàng và các nhà hàng bị lỗ.

Tuy nhiên, dự án vẫn còn một số hạn chế:

Thiếu tiêu chí rõ ràng: Việc xác định "Top" các nhà hàng còn mơ hồ, chưa có tiêu chí cụ thể.

Thiếu ngữ cảnh: Cần bổ sung thông tin về khoảng thời gian, đơn vị tiền tệ và quy mô dữ liệu để kết quả phân tích chính xác hơn.

Phân tích còn đơn giản: Bạn mới chỉ tập trung vào việc mô tả dữ liệu, chưa đi sâu vào phân tích các yếu tố ảnh hưởng và đưa ra insight giá trị.

2. Hướng phát triển:

Bổ sung thông tin: Cập nhật thông tin về tiêu chí "Top", đơn vị tiền tệ, khoảng thời gian, quy mô dữ liệu.

Phân tích chuyên sâu:

Phân tích yếu tố ảnh hưởng: Sử dụng các kỹ thuật phân tích thống kê, khai phá dữ liệu để tìm ra các yếu tố ảnh hưởng đến lợi nhuận nhà hàng, chi tiêu khách hàng (ví dụ: loại hình ẩm thực, giá cả, vị trí, đánh giá,...)

Phân khúc khách hàng: Phân loại khách hàng theo các tiêu chí khác nhau (như mức chi tiêu, sở thích ẩm thực, tần suất sử dụng app) để có chiến lược tiếp cận phù hợp.

Dự đoán: Xây dựng mô hình dự đoán doanh thu, lợi nhuận của nhà hàng hoặc hành vi của khách hàng dựa trên dữ liệu lịch sử.

Mở rộng hệ thống:

Kết nối với nguồn dữ liệu thời gian thực: Cập nhật dữ liệu Zomato liên tục để nắm bắt xu hướng thay đổi.

Xây dựng dashboard: Tạo dashboard trực quan, động để theo dõi các chỉ số kinh doanh quan trọng.

Phát triển ứng dụng: Phát triển ứng dụng di động hoặc website để cung cấp thông tin và phân tích cho nhà hàng và người dùng.

3. Kết luận:

Dự án phân tích dữ liệu Zomato của bạn là một khởi đầu tốt. Bằng cách bổ sung thông tin, phân tích chuyên sâu và mở rộng hệ thống, bạn có thể tạo ra một hệ thống phân tích dữ liệu giá trị, hỗ trợ ra quyết định hiệu quả cho Zomato, nhà hàng và người dùng.

Tài liệu tham khảo:

[1] Slide bài giảng và các file pdf bài tập của gv.Trần Văn Thành, giảng viên môn Kho Dữ Liệu trường đại học Sư phạm Kỹ thuật TP HCM

[2] Tự học Power BI cho người mới bắt đầu (47 phút)

https://www.youtube.com/watch?v=F7JRKUim-0&ab_channel=G%C3%A0Excel