A picture containing picture frame, rectangle, screenshot, frame

Description automatically generated**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH**

Ảnh có chứa văn bản, mẫu họa

Mô tả được tạo tự động**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 gởi 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 hang 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](#_Toc167433793)

[................................................................................................................................... 2](#_Toc167433794)

[**I.** **TỔNG QUAN VỀ TẬP DỮ LIỆU** 6](#_Toc167433795)

[**1.** **Lý do hình thành dự án:** 6](#_Toc167433796)

[**1.1.** **Đặt vấn đề:** 6](#_Toc167433797)

[**1.2.** **Giải pháp:** 6](#_Toc167433798)

[**1.3.** **Mục tiêu và ý nghĩa của đồ án:** 6](#_Toc167433799)

[**2.** **Giới thiệu tổng quan về tập dữ liệu:** 6](#_Toc167433800)

[**2.1.** **Nguồn dữ liệu sử dụng:** 6](#_Toc167433801)

[**2.2.** **Mô tả chi tiết dữ liệu:** 6](#_Toc167433802)

[**2.2.1.** **Tập dữ liệu:** 7](#_Toc167433803)

[**2.2.2.** **Mô tả chi tiết các thuộc tính trong tập dữ liệu:** 7](#_Toc167433804)

[**2.3.** **Công cụ sử dụng trong đồ án:** 10](#_Toc167433805)

[**II.** **Thiết kế xây dựng cơ sở dữ liệu tác nghiệp (OLAP)** 10](#_Toc167433806)

[**1.** **Tiền xử lý dữ liệu:** 10](#_Toc167433807)

[**1.1.** **Chọn các thuộc tính cần thiết cho bản Fact chính:** 10](#_Toc167433808)

[**1.2.** **Tạo bảng Dim:** 12](#_Toc167433809)

[**2.** **Thiết kế DataWarehouse:** 14](#_Toc167433810)

[**2.1.** **Thiết kế StarSchema cho Fact chính:** 14](#_Toc167433811)

[**III.** **Truyền dữ liệu vào các Dim và Fact (SQL)** 16](#_Toc167433812)

[**3.** **Thực hiện các bước chuẩn bị trên SQL server:** 16](#_Toc167433813)

[**3.1.** **Tạo các dim và fact:** 16](#_Toc167433814)

[**3.2.** **Truyền dữ liệu từ csv vào database, và các fact:** 18](#_Toc167433815)

[**3.2.1.** **Truyền dữ liệu vào các Dim** 18](#_Toc167433816)

[**3.2.2.** **Truyền dữ liệu vào 2 Fact:** 19](#_Toc167433817)

[**IV.** **Tích hợp dữ liệu vào kho (SSIS)** 21](#_Toc167433818)

[**1.** **Control Flow:** 21](#_Toc167433819)

[**1.1.** **Control Flow của các bảng Dim:** 21](#_Toc167433820)

[**1.3.** **Control Flow của 2 bảng Fact:** 26](#_Toc167433821)

[**2.** **Data Flow:** 27](#_Toc167433822)

[**2.1.** **Data Flow của Fact Sales:** 27](#_Toc167433823)

[**2.2.** **Data Flow của FactRestaurantPerformence:** 28](#_Toc167433824)

[**V.** **Thiết kế Cube (SSAS)** 29](#_Toc167433825)

[**1.** **Data SourceView:** 29](#_Toc167433826)

[**2.** **Cube:** 29](#_Toc167433827)

[**VI.** **Trả lời cho các câu hỏi đã đặt ra** 30](#_Toc167433828)

[**VII.** **Trực quan hóa (Power BI)** 33](#_Toc167433829)

[**Tài liệu tham khảo:** 37](#_Toc167433830)

# **TỔNG QUAN VỀ TẬP DỮ LIỆU**

## **Lý do hình thành dự án:**

### **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. **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 hang.

**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. **Mục tiêu và ý nghĩa của đồ 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. Để 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.

1. **Giới thiệu tổng quan về tập dữ liệu:**
   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)
  1. **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.

* + 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 hang.

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

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.

* + 1. **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. |
| **cusine** | 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 |
| **total\_spent** | Tổng số tiền đã chi tiêu bởi người dùng |
| **Menu.csv** | **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. |

* 1. **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

1. **Thiết kế xây dựng cơ sở dữ liệu tác nghiệp (OLAP)**
2. **Tiền xử lý dữ liệu:**
   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**

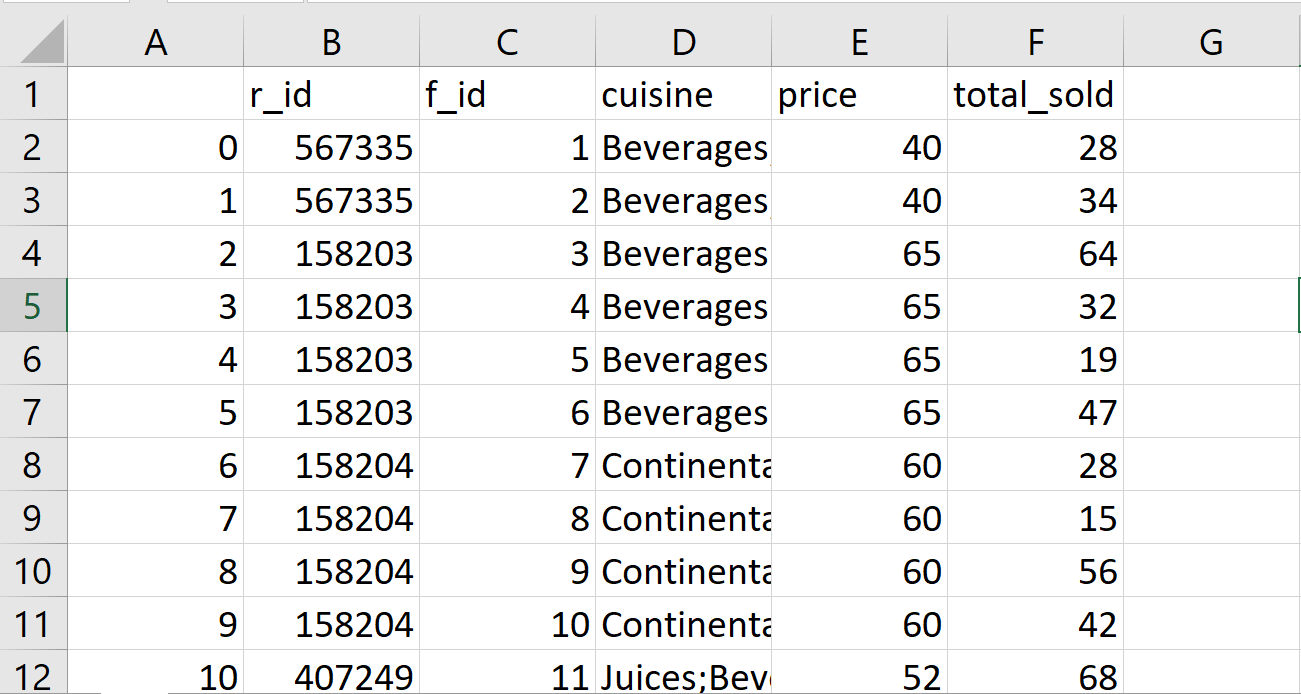
|  |  |  |
| --- | --- | --- |
| **Tập dữ liệu** | **Tên thuộc tính** | **Mô tả** |
| **orders.csv** | **Order\_id** | Mã số định danh duy nhất cho mỗi đơn hàng. |
| **Sales\_qty** | Số lượng sản phẩm được bán trong đơn hàng. |
| **Sales\_amont** | Tổng số tiền từ đơn hàng. |
| **Currency** | Đơn vị 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. **Tạo bảng Dim:**

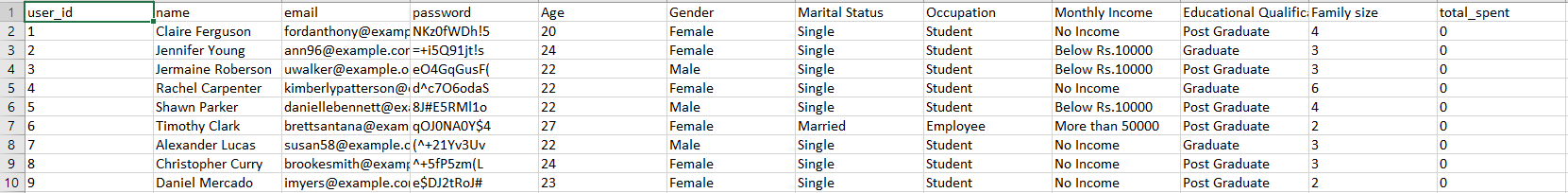
Thực hiện xử lí các dữ liệu không khớp với data bằng cách quan sát và thực hiện xử lí ở excel.

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

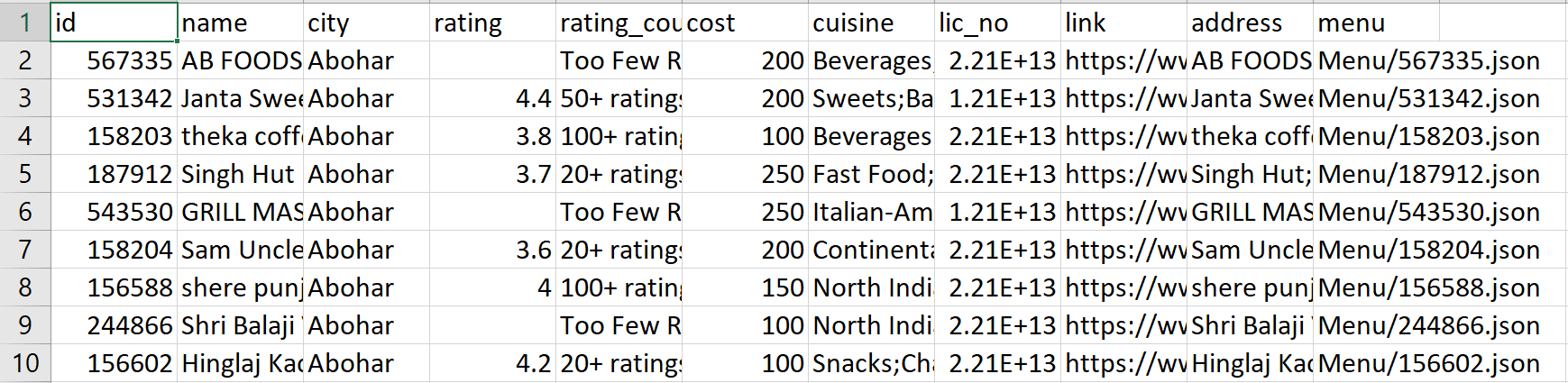
* + 1. DimMenu:



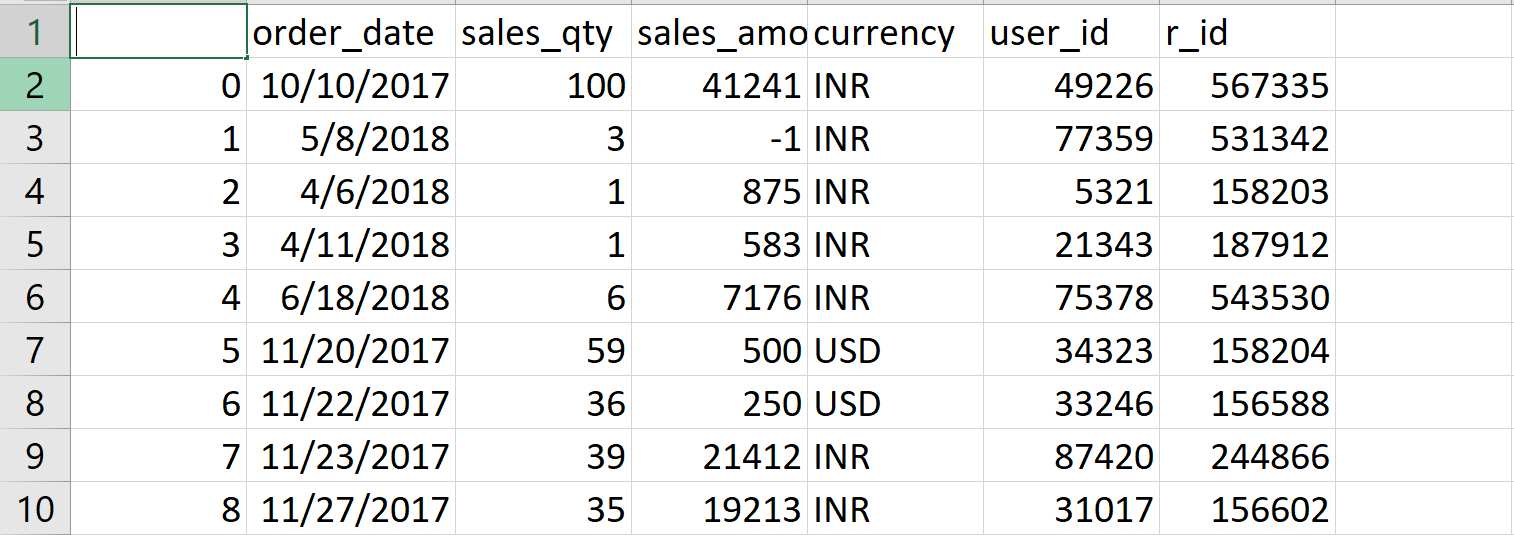
* + 1. DimUser:



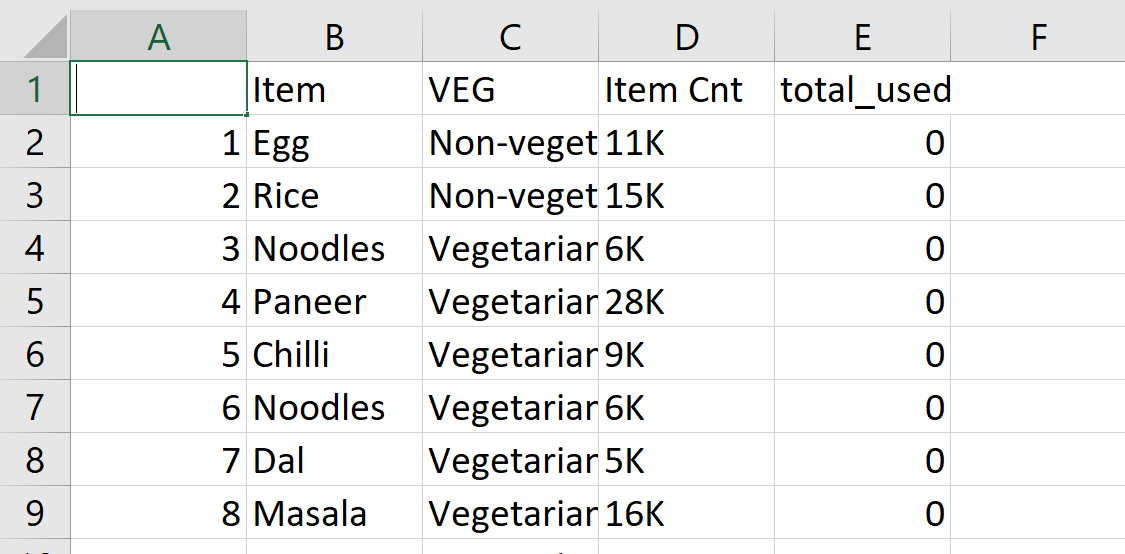
* + 1. DimRestaurant:



* + 1. DimOrder:



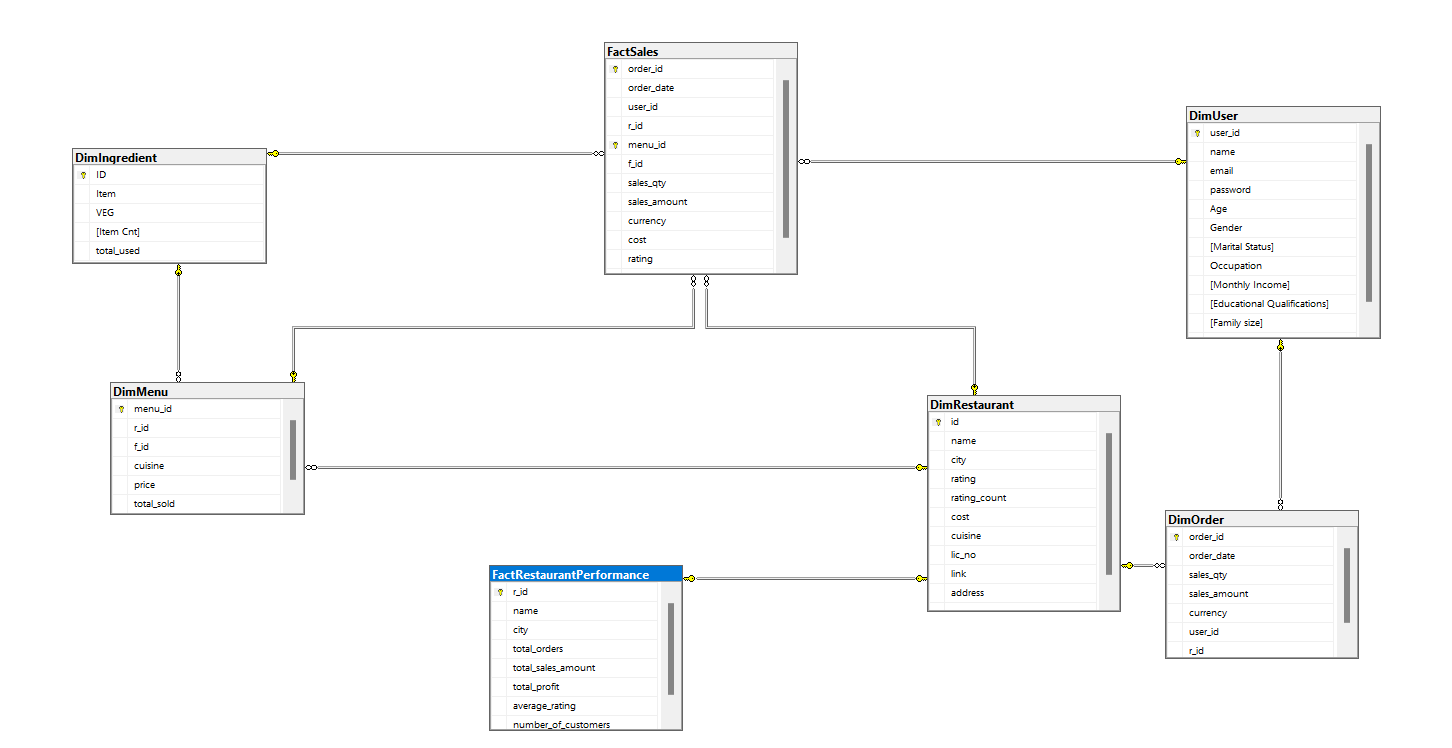
* + 1. DimIngredient:



1. **Thiết kế DataWarehouse:**
   1. **Thiết kế StarSchema cho Fact chính:**

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

Đây là lược đồ **StarSchema**:



1. **Truyền dữ liệu vào các Dim và Fact (SQL)**
2. **Thực hiện các bước chuẩn bị trên SQL server:**
   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 DimMenud

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

);

* 1. **Truyền dữ liệu từ csv vào database, và các fact:**
     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'

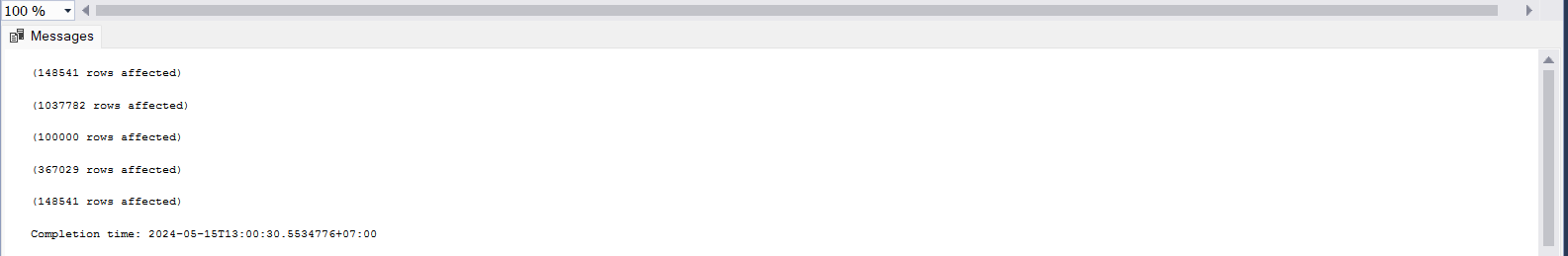
WITH (

FIELDTERMINATOR = ',',

ROWTERMINATOR = '\n',

FIRSTROW = 2 -- Use this if your CSV has a header row

);



* + 1. **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,

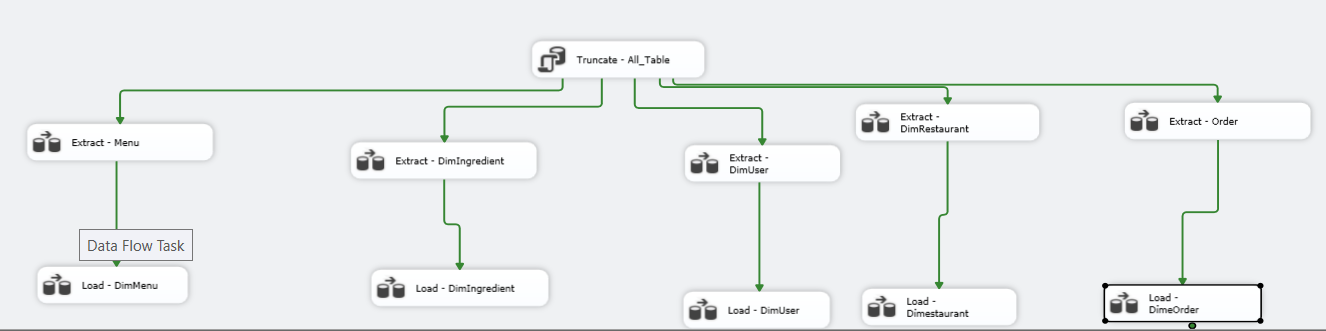
r.name,

r.city;

1. **Tích hợp dữ liệu vào kho (SSIS)**

## **Control Flow:**

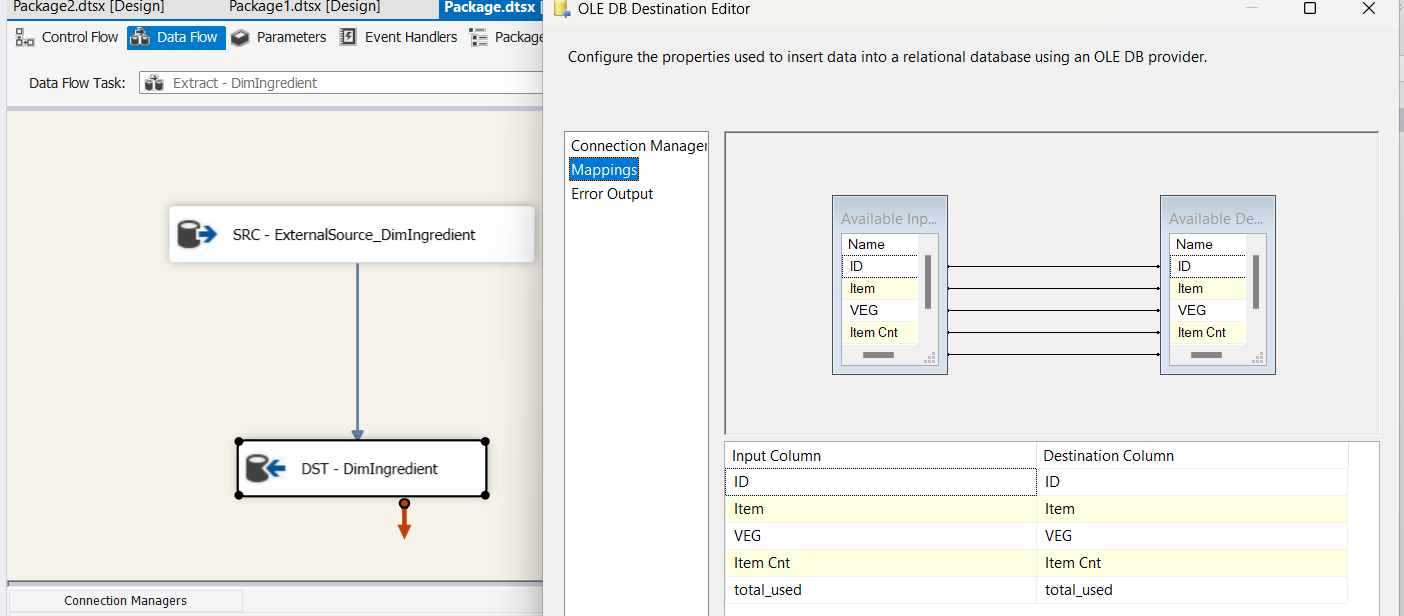
* 1. **Control Flow của các bảng Dim:**

****

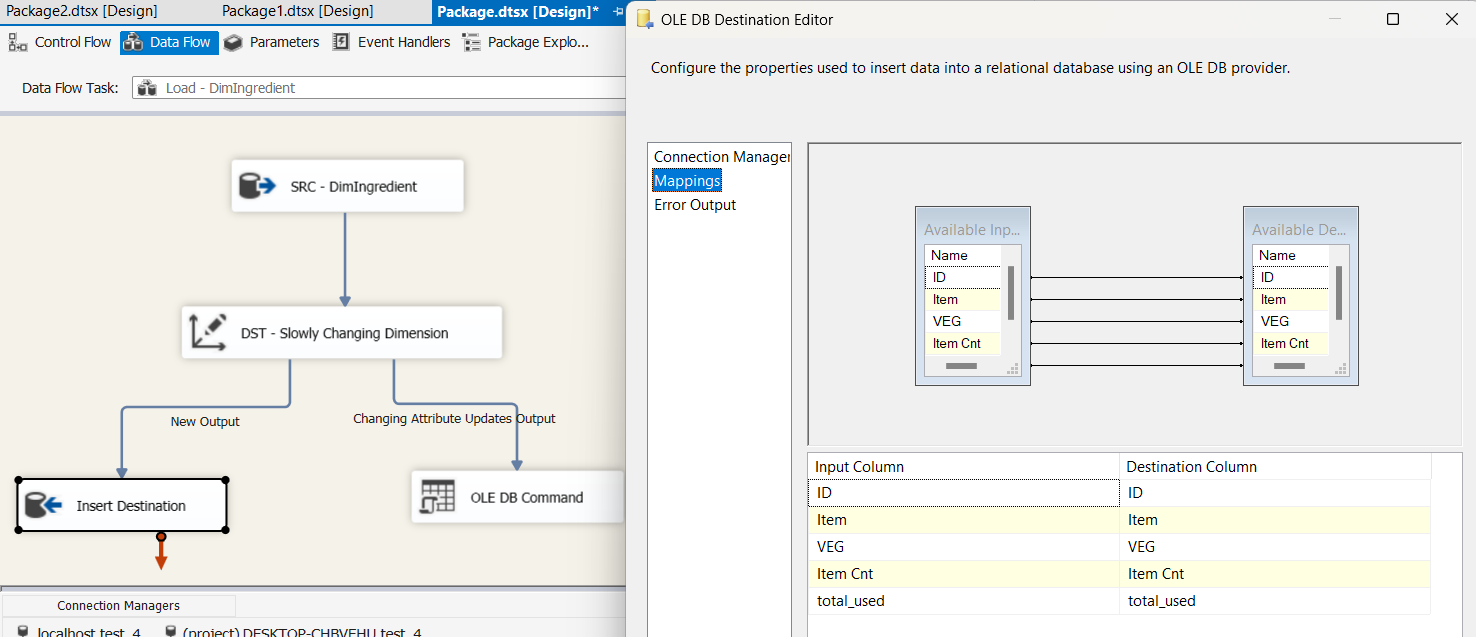
* 1. Data Flow của các bảng Dim:
     1. Data Flow DimIngredient:

Extract DimIngredient:

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



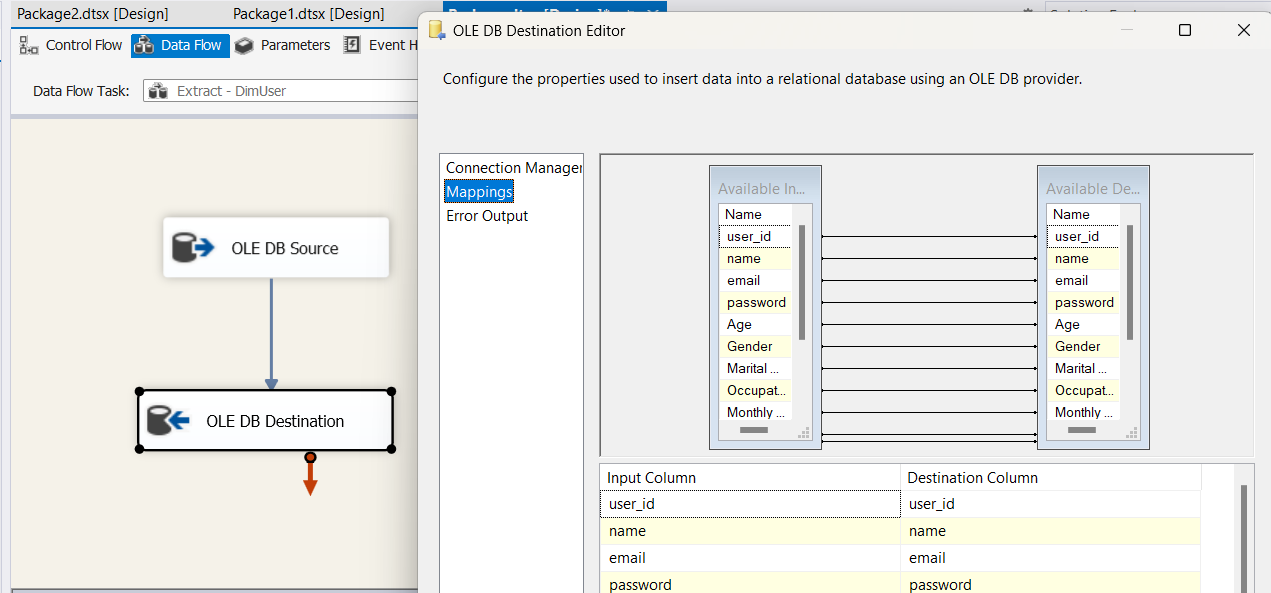
Load DimIngredient:



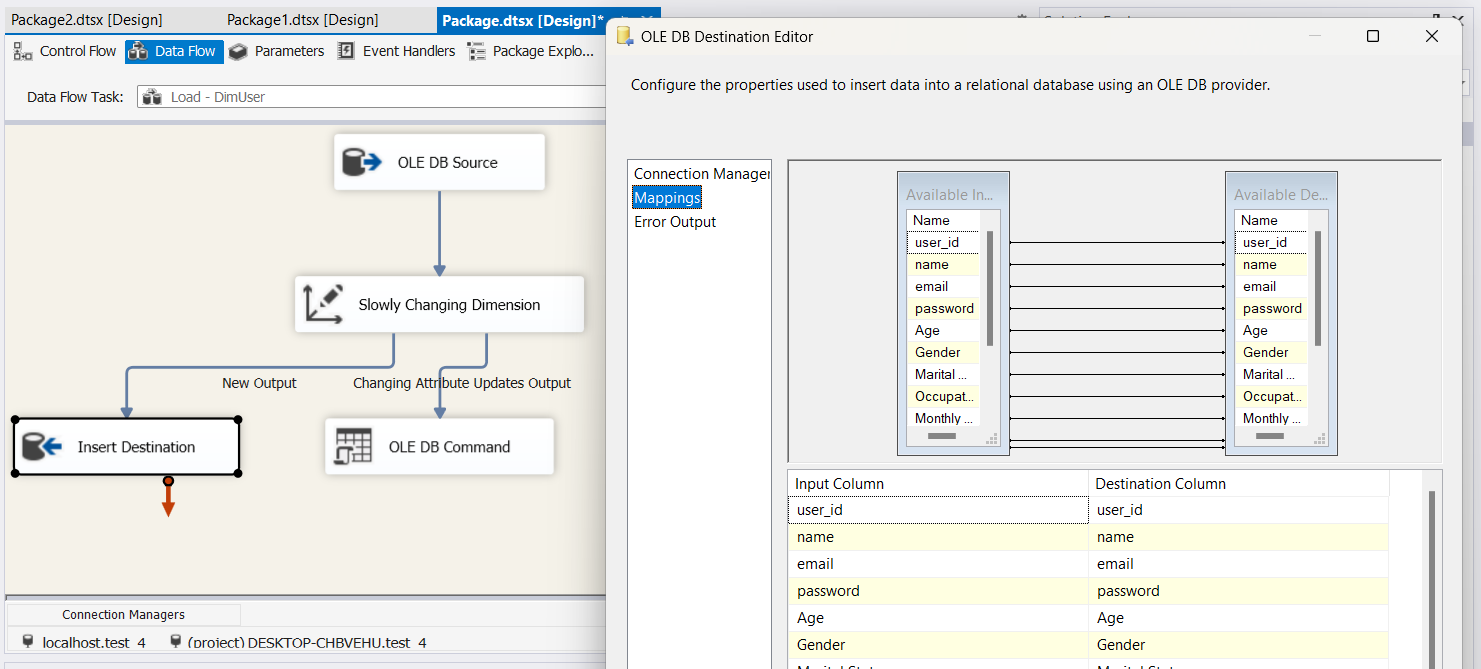
* + 1. Data Flow DimUser:

Extract DimUser:

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



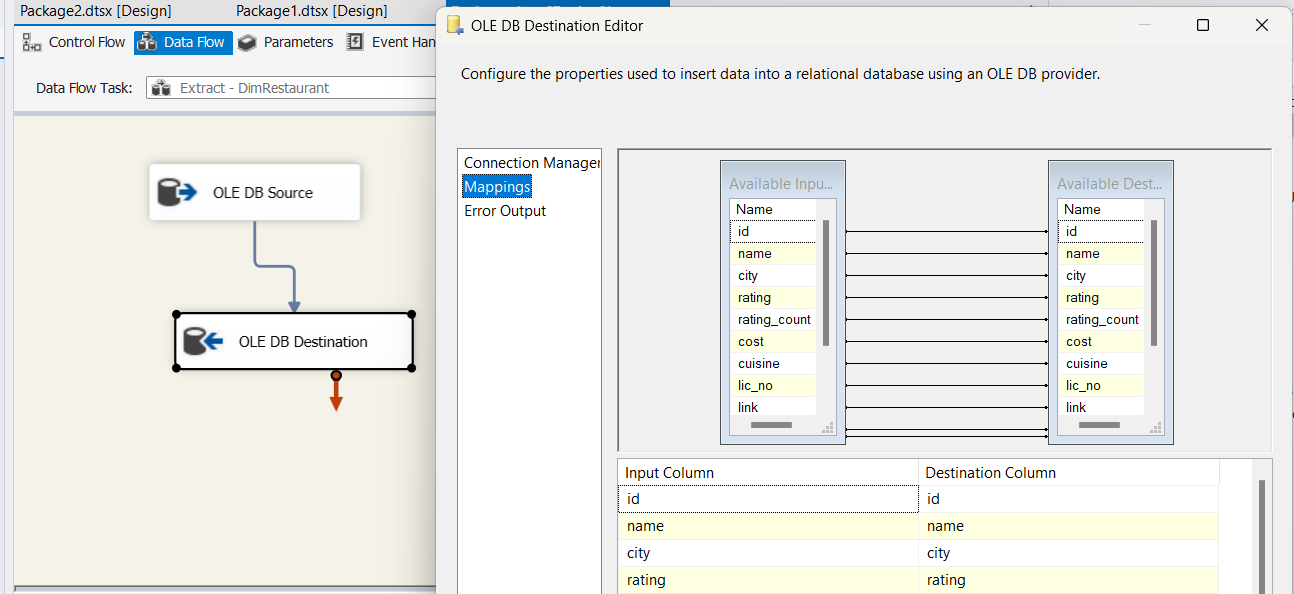
Load DimUser:



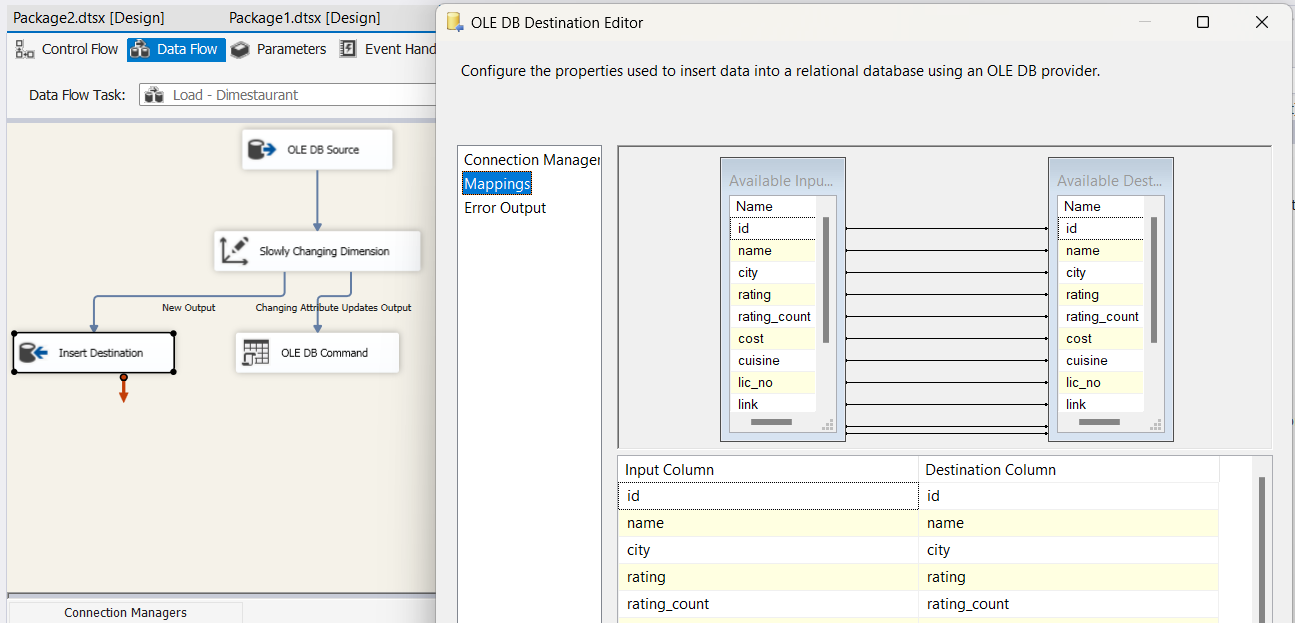
* + 1. Data Flow DimRestaurant:

Extract DimRestaurant:

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



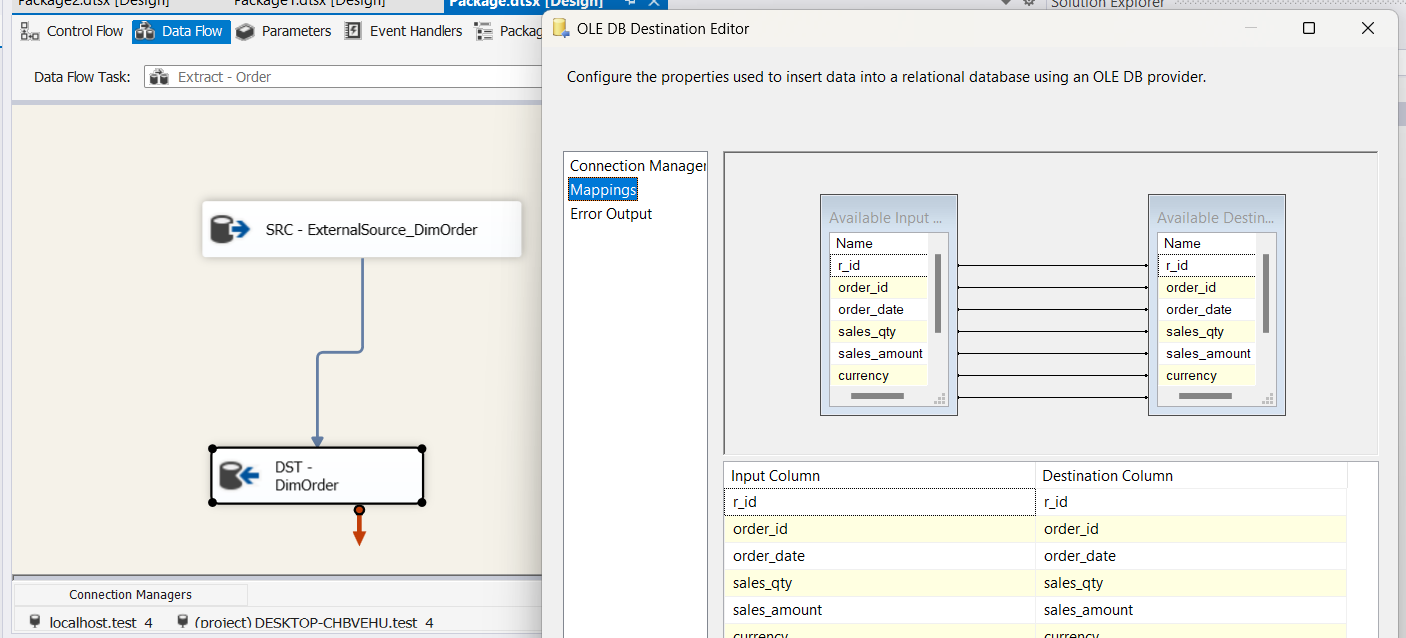
Load DimRestaurant:



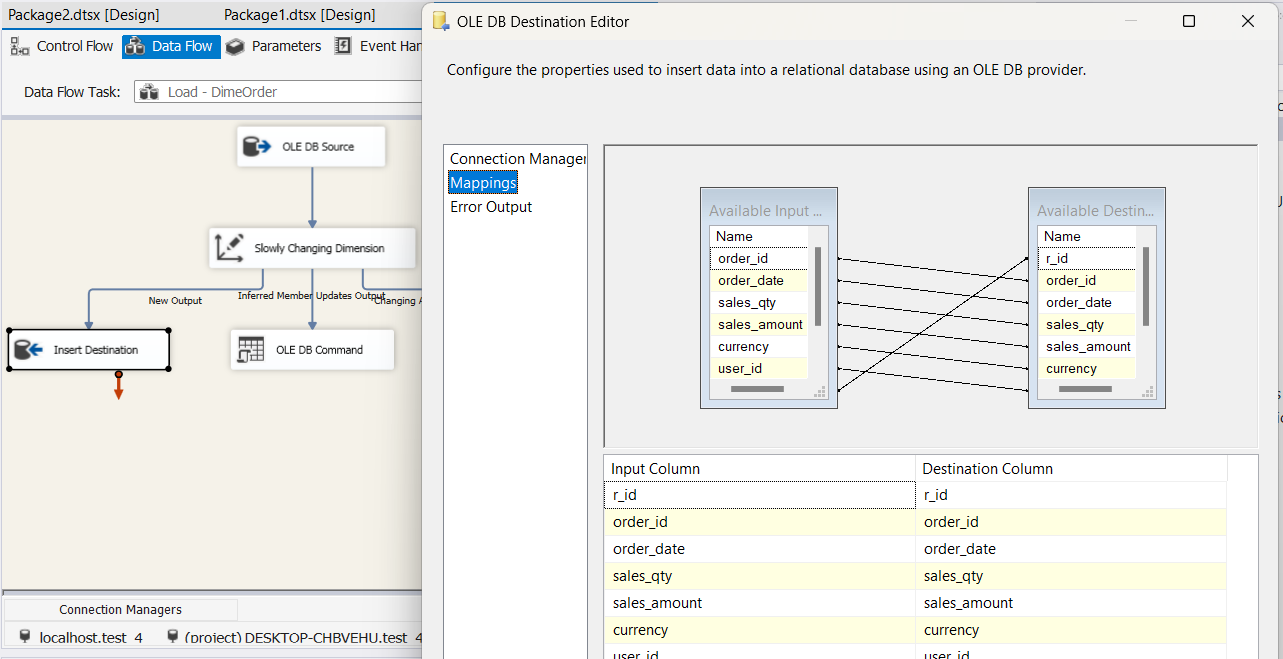
* + 1. Data Flow DimOrder:

Extract DimOrrder:

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



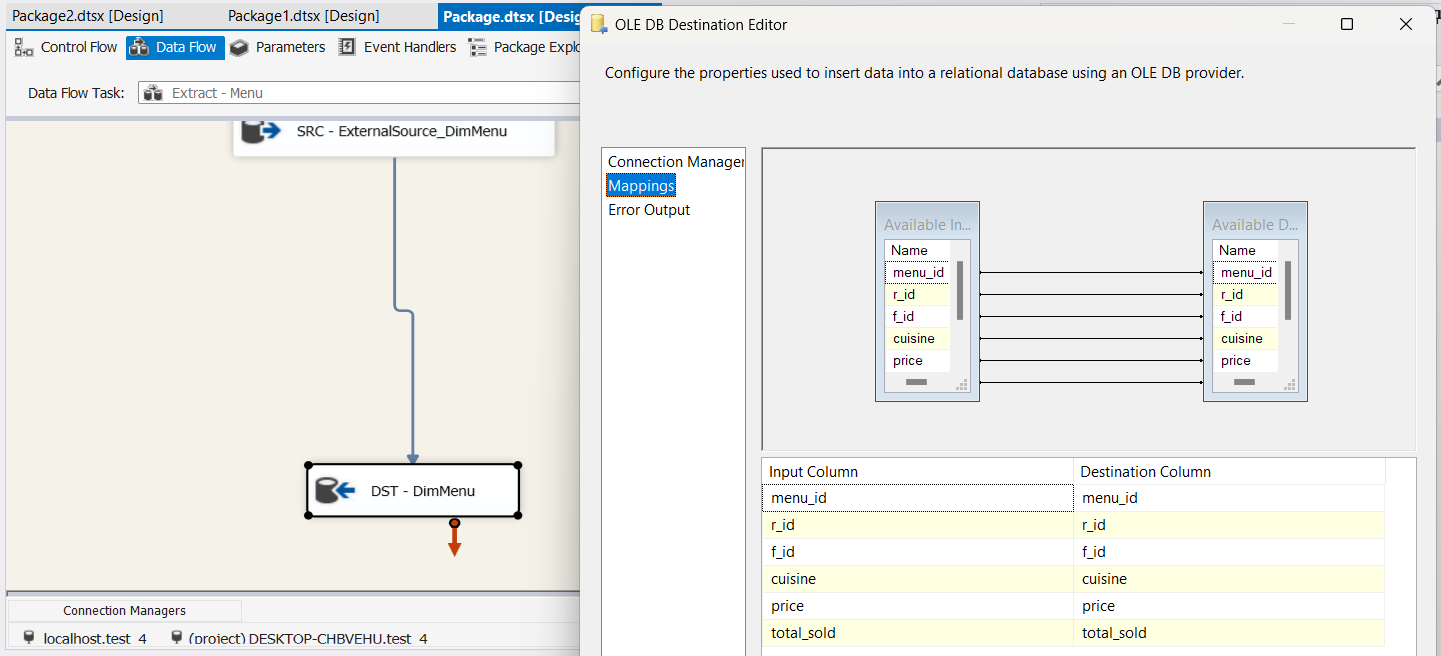
Load DimOrder:



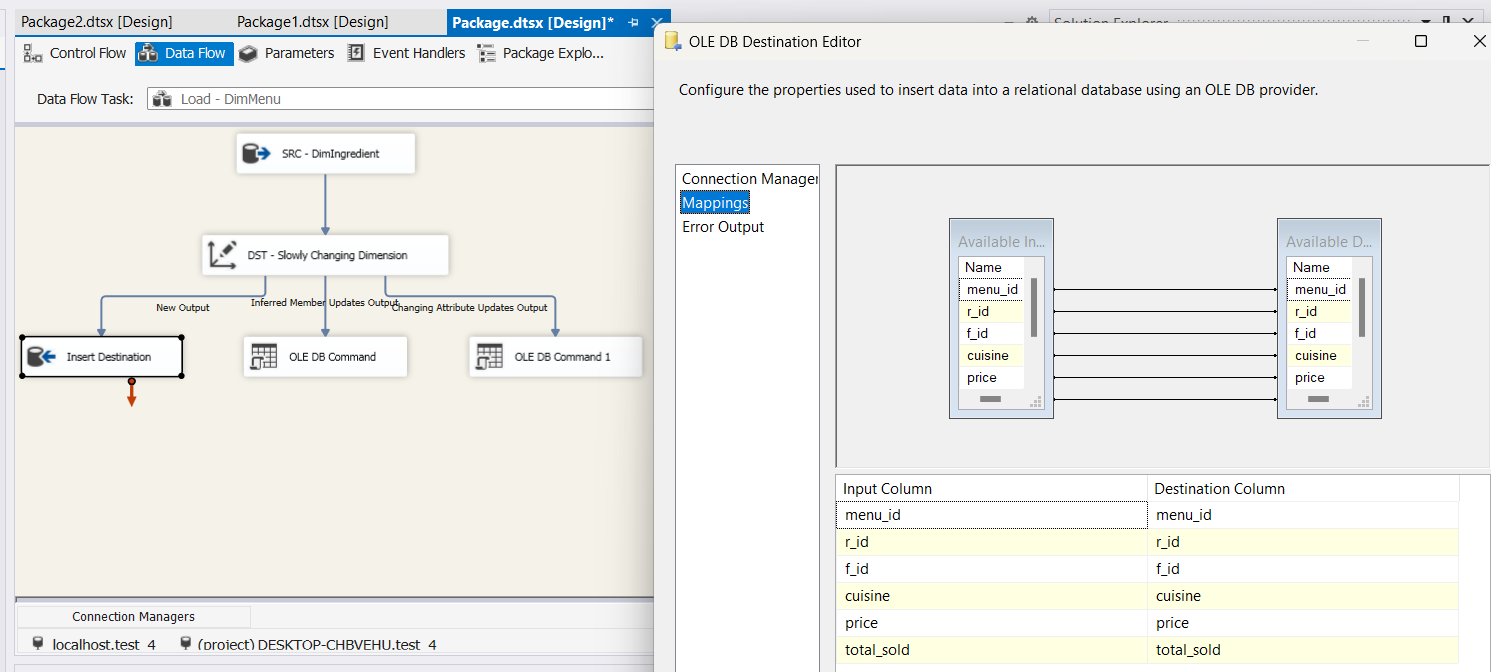
* + 1. Data Flow DimMenu:

Extract DimMenu:

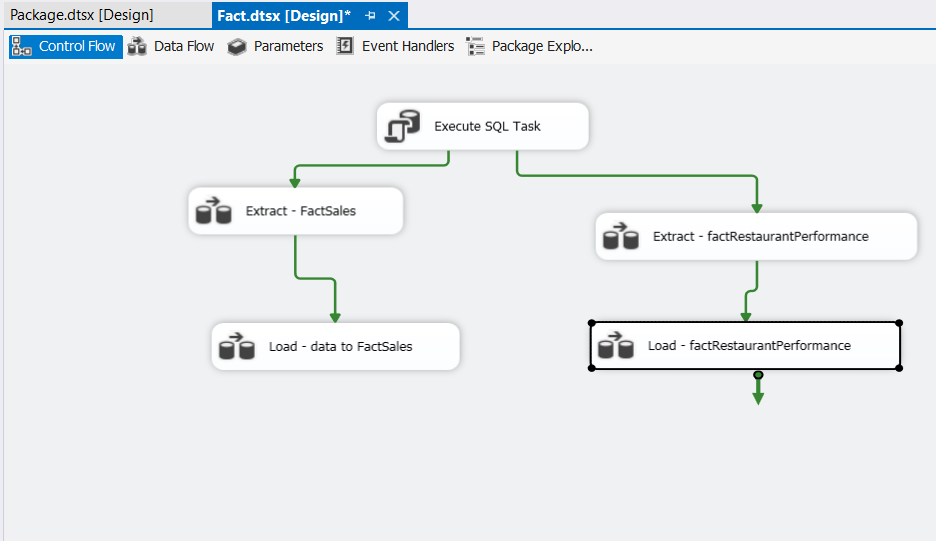
* Thực hiện tạo stgMenu sau đó mapping giống hình bên dưới:



Load DimMenu:

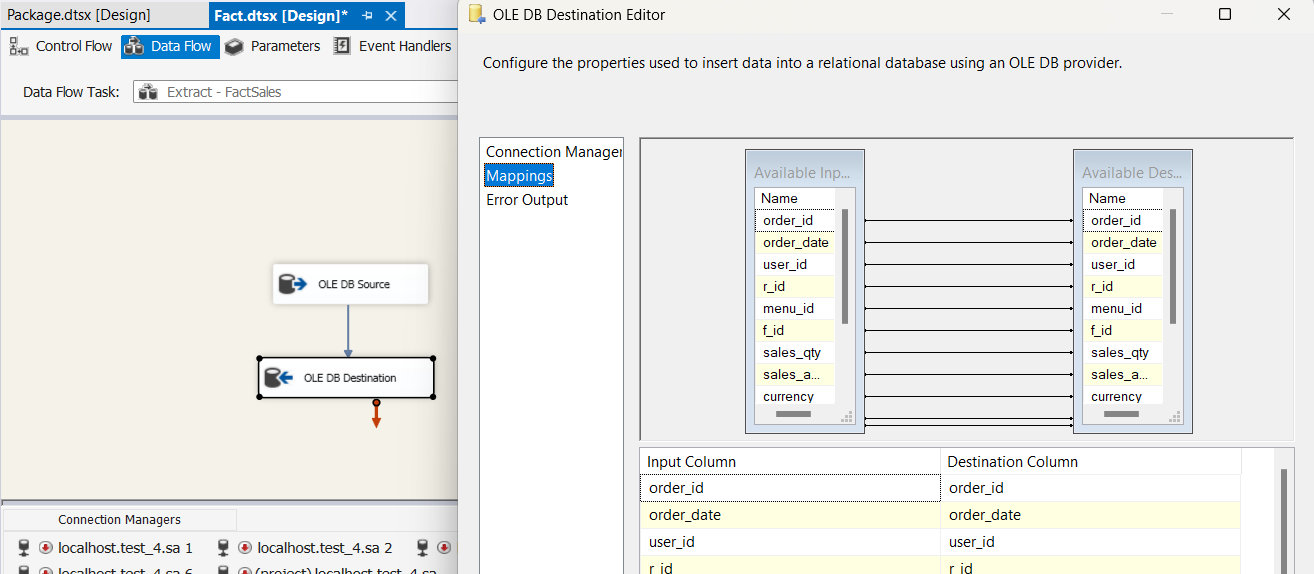


* 1. **Control Flow của 2 bảng Fact:**

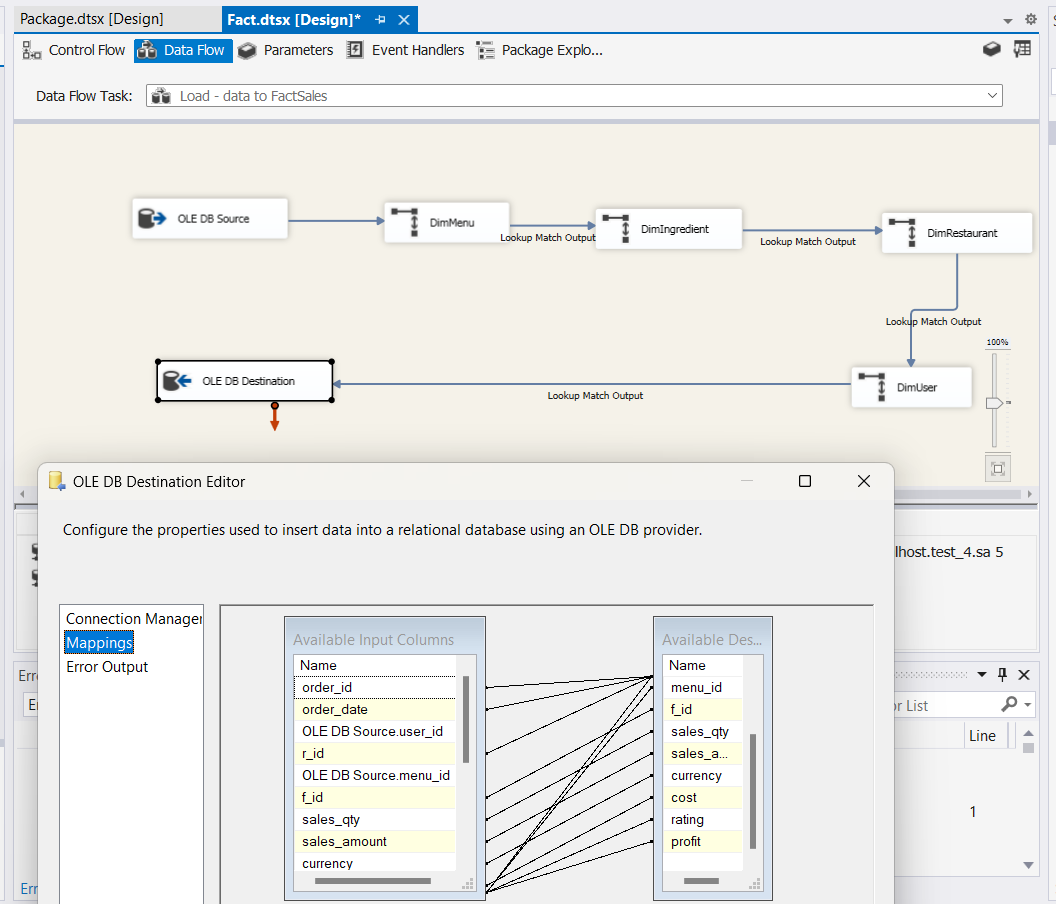
****

1. **Data Flow:**
   1. **Data Flow của Fact Sales:**

Extract FactSales:

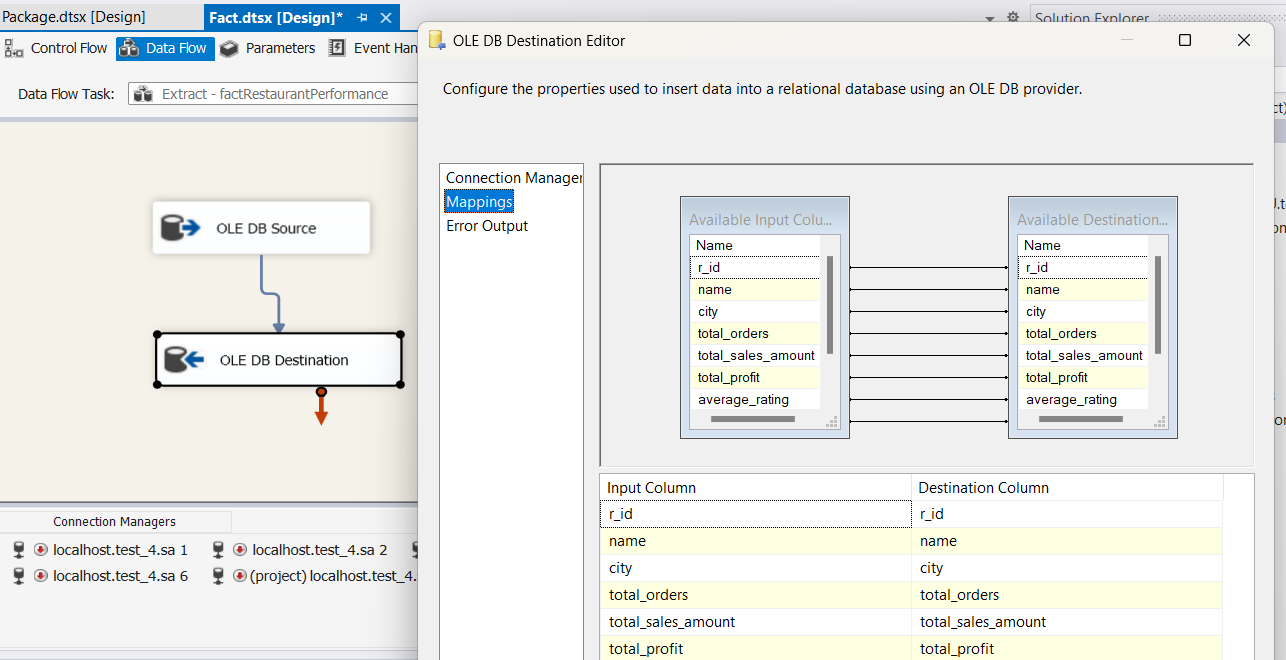


Load FactSales:

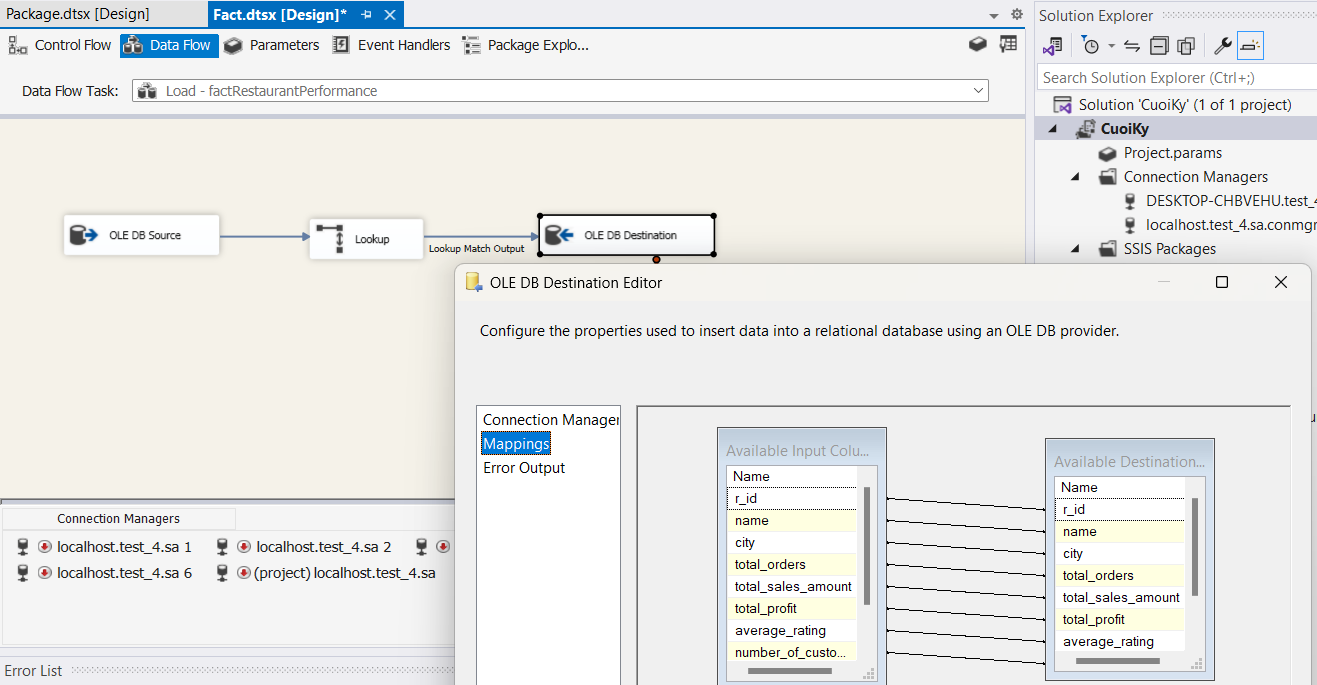
****

* 1. **Data Flow của FactRestaurantPerformence:**

Extraxt FactRestaurantPerformence:

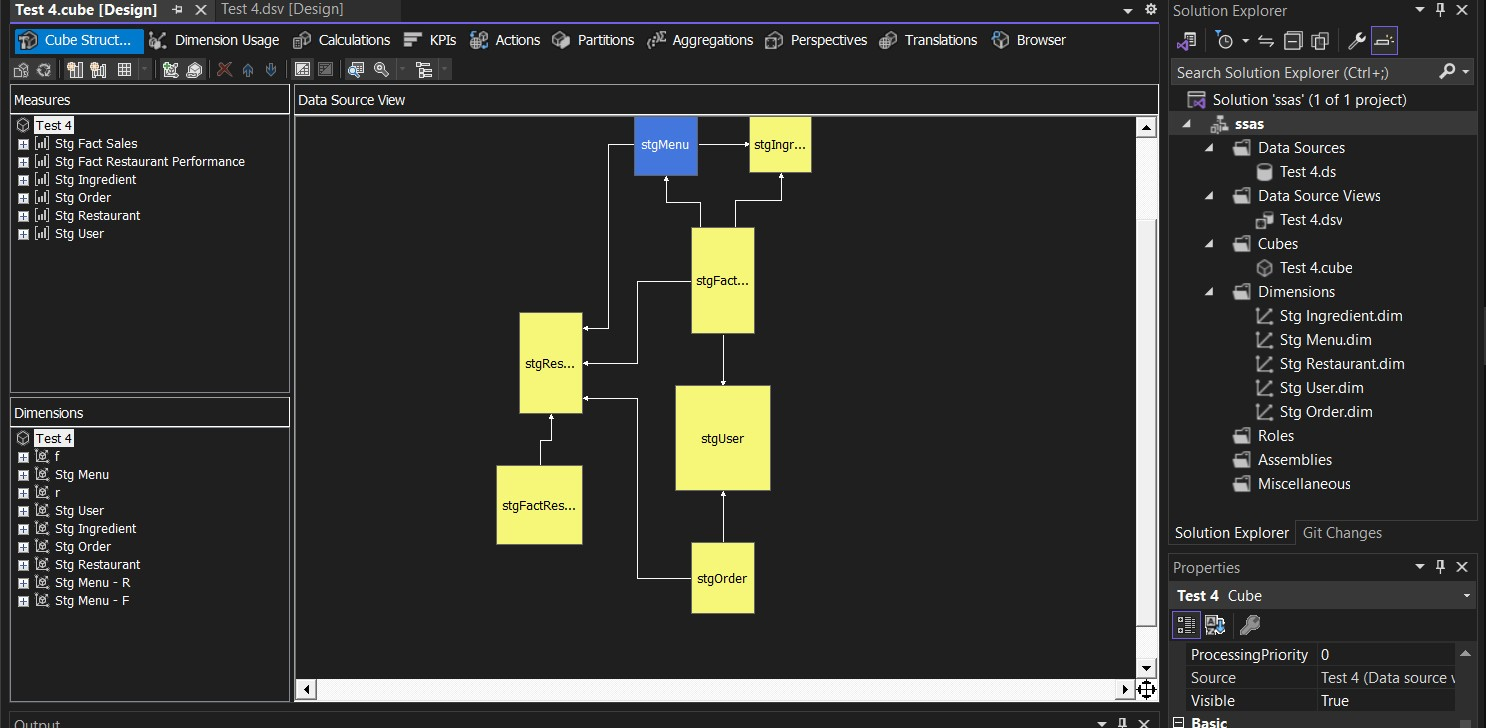


Load FactRestaurantPerformence:

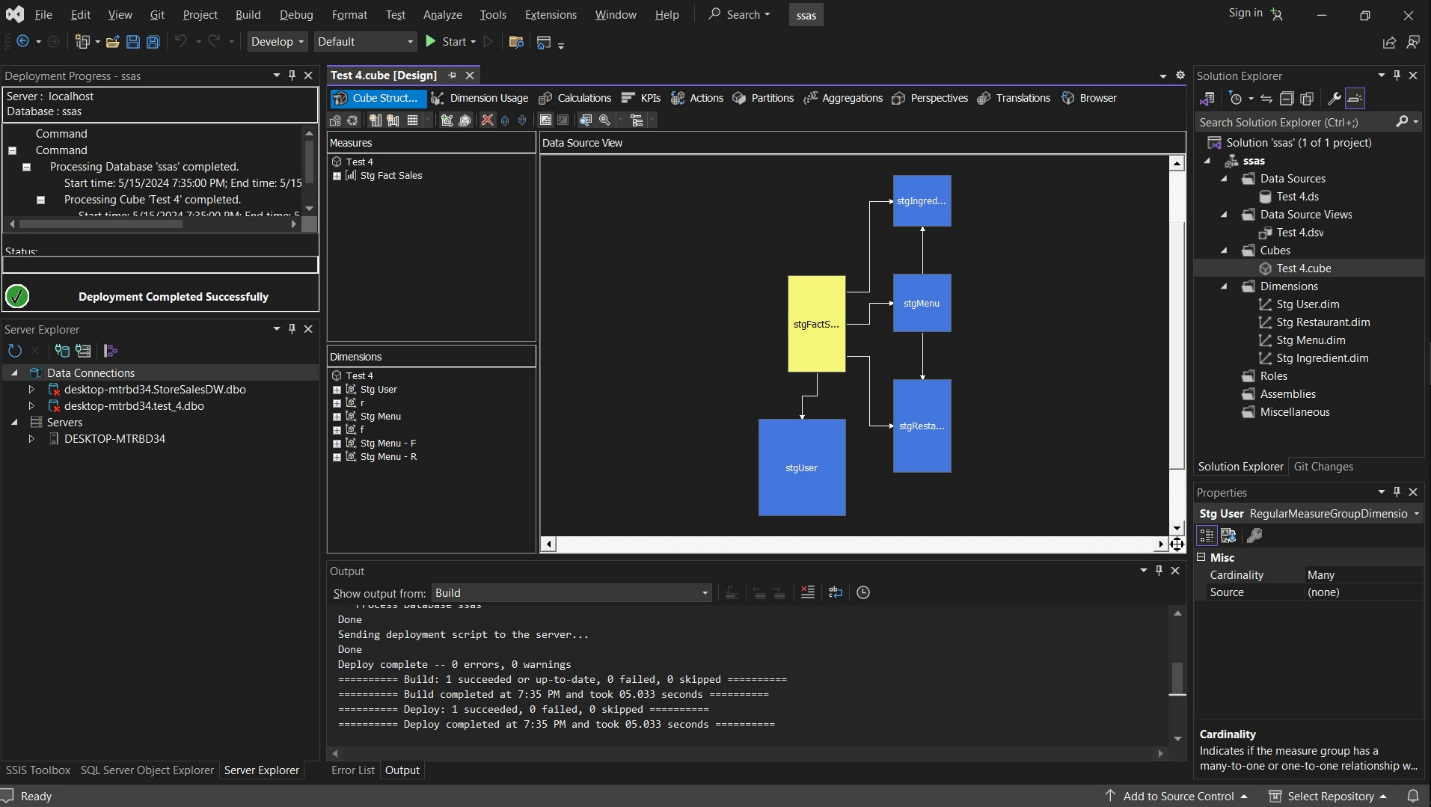
****

1. **Thiết kế Cube (SSAS)**
2. **Data SourceView:**

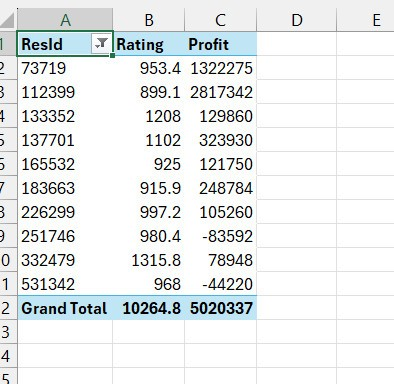
- Ở phần Data Source View chọn New Data Source View, chọn các bảng cần thiết và finish

****

1. **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

1. **Trả lời cho các câu hỏi đã đặt ra**
2. **Lợi nhuận 10 nhà hang có số lượt đánh giá cao nhất:**



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.

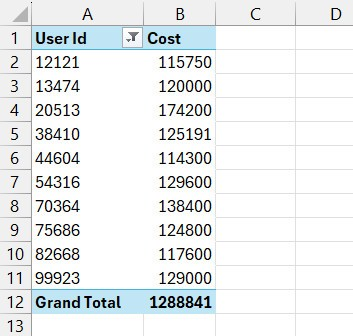
Thông tin chi tiết về từng nhà hàng: Loại hình ẩm thực, mức giá, vị trí, số lượng đánh giá,...

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.

1. **Top 10 khách hang có chi tiêu cao nhất:**



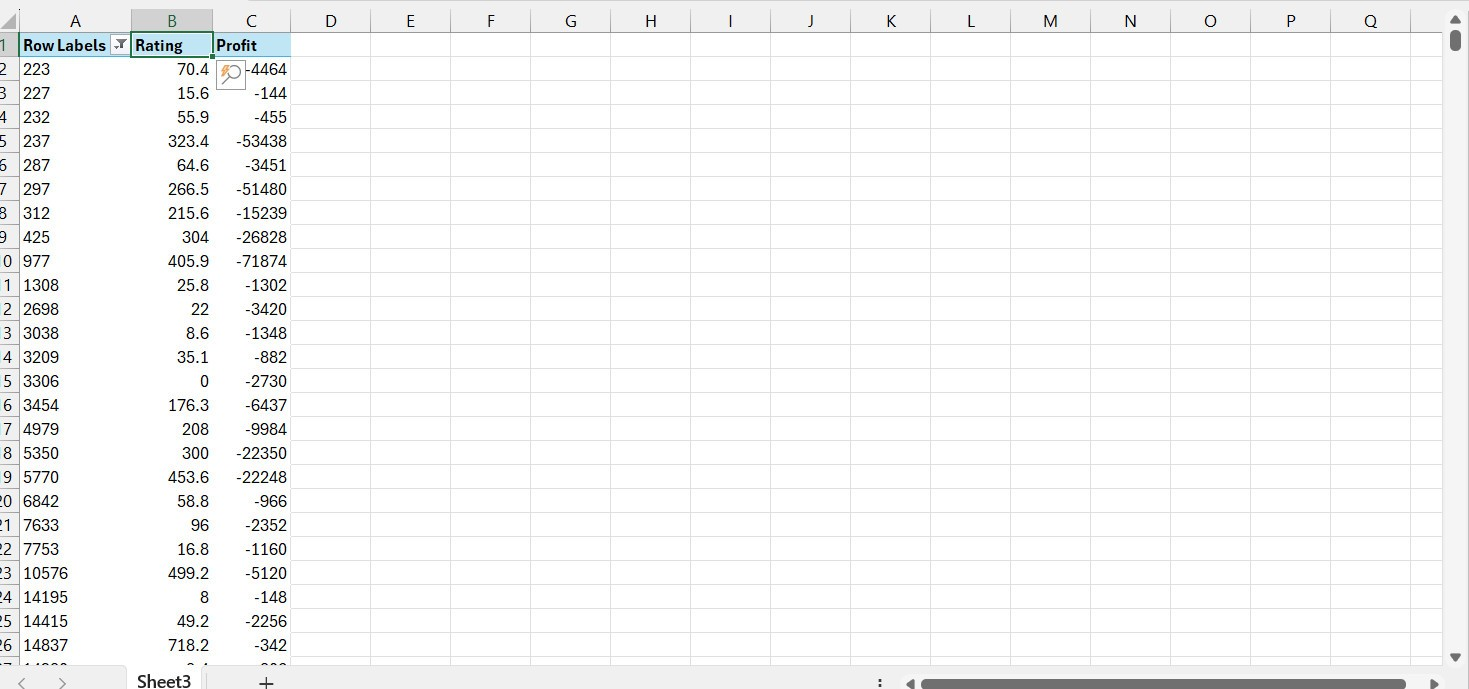
Phân bố chi tiêu: Có thể thấy sự chênh lệch về chi tiêu giữa các khách hàng trong Top 10. Khách hàng có User Id 20513 chi tiêu cao nhất (174200) trong khi User Id 44604 có chi tiêu thấp nhất (114300).

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 them nhiều tiềm lực từ họ, bên cạch đó phân tích them 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,...

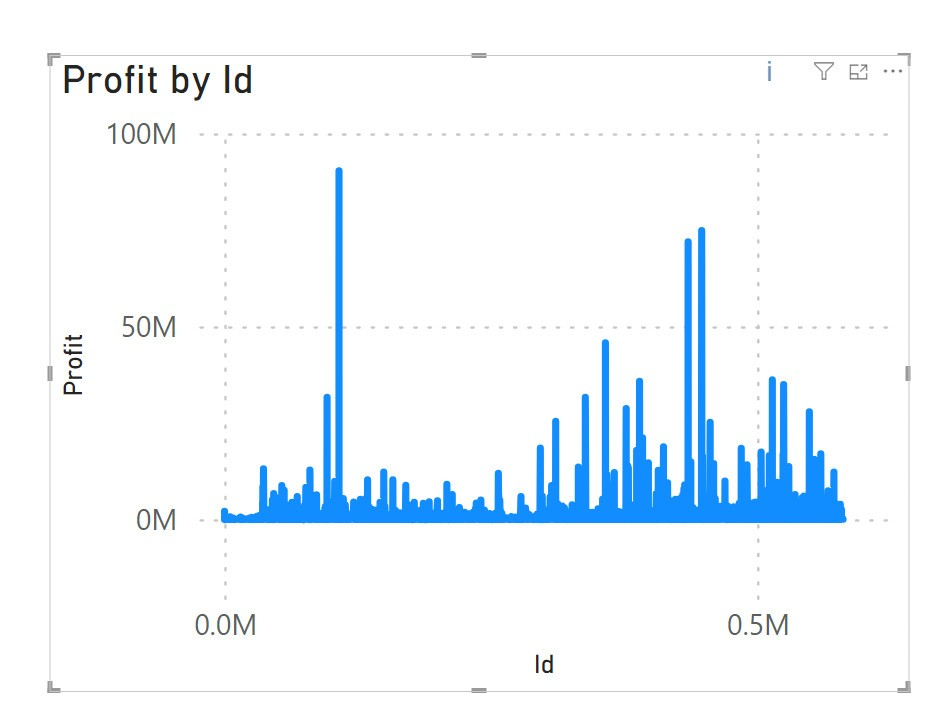
1. **Top các nhà hang bị lỗ vốn:**



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 hang lỗ lại có lượng đánh gía 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 them 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 xót) .

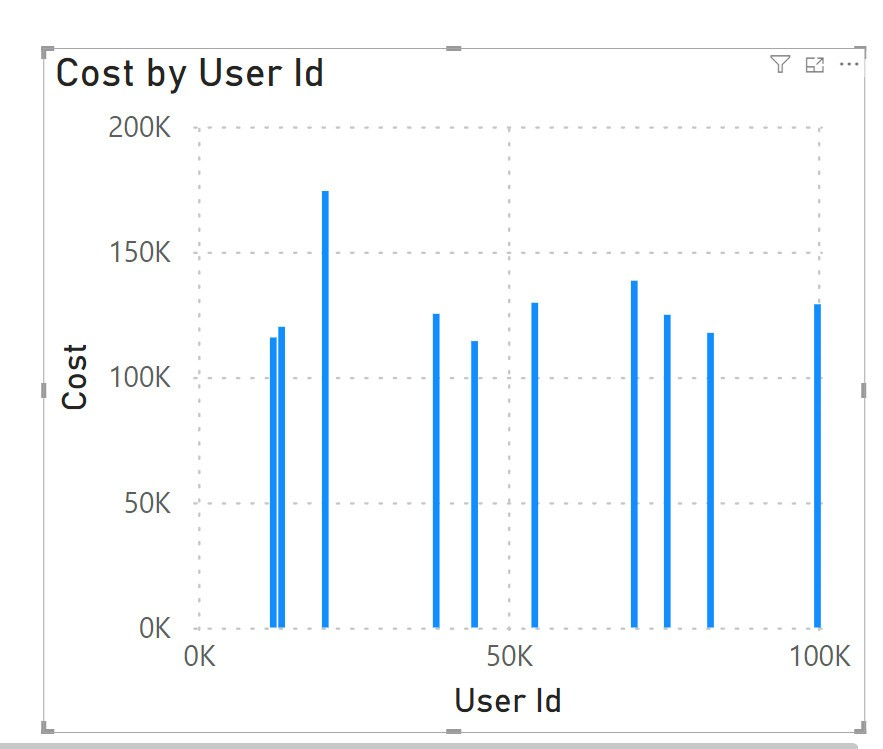
1. **Trực quan hóa (Power BI)**
2. **Lợi nhuận của tất cả nhà hàng:**



Nắm bắt được tệp khách hàng có chi 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.

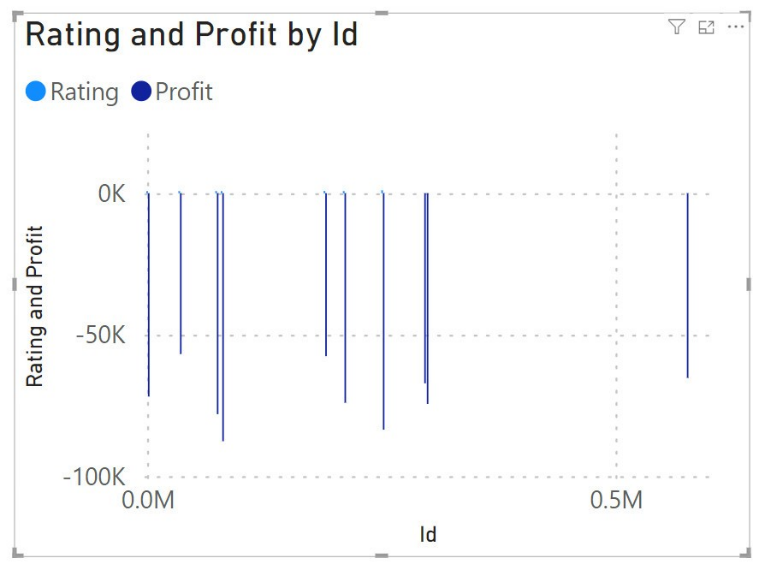
Nắm bắt được tệp khách hàng bình dân và giá rẻ để đưa ra nhiều đãi ngộ thu hút khách, ưu tiên các nhà hàng bình dân giá rẻ được xuất hiện quảng cáo đối với tệp khách này.

1. **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 lục cho nhà hàng.

1. **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.

1. **TỔNG KẾT**
2. **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.

1. **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.

1. **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 TPHCM

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

<https://www.youtube.com/watch?v=F7JRKUIim-0&ab_channel=G%C3%A0Excel>