|  |  |
| --- | --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**  **THÀNH PHỐ HỒ CHÍ MINH** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**KHOA CÔNG NGHỆ THÔNG TIN**

A logo of hands holding a book and a candle

Description automatically generated

**ĐỒ ÁN MÔN HỌC**

**MÔN:** KHO DỮ LIỆU

**MÃ MÔN HỌC:** DAWH430784

**XÂY DỰNG KHO DỮ LIỆU TỪ HỆ THỐNG THƯƠNG MẠI ĐIỆN TỬ OLIST TẠI BRAZIL**

**Nhóm sinh viên thực hiện:** Nhóm 5

**Học kỳ:** II

**Năm học:** 2024 – 2025

**GVHD: Nguyễn Văn Thành**

*Thành phố Hồ Chí Minh, tháng 5 năm 2025*

**PHÂN CÔNG NHIỆM VỤ**

**Nhóm:** 05

**Lớp:** DAWH430784

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Nhiệm vụ** | **Trần Khánh Dương 22133011** | **Nguyễn Đinh Hồng Phúc 22133041** | **Lê Văn Sang 22133047** | **Dương Thành Sơn 22133048**  **(Nhóm trưởng)** |
| Tìm kiếm dữ liệu |  |  |  | X |
| Xử lý dữ liệu đầu vào | X | X |  | X |
| Xác định  business process |  | X | X |  |
| Xác định Dim và Fact | ProductSellerPeformance (3) | Sales&RevenueAnalysis (1) OrderFulfillment&DeliveryPerformance (2) | Sales&RevenueAnalysis (1) OrderFulfillment&DeliveryPerformance (2) | ProductSellerPeformance (3) |
| Staging và load và các dim, fact bằng T-SQL | X | X | X | X |
| Staging và load và các dim, fact bằng SSIS |  | X | X | X |
| Nhập dữ liệu vào SSAS, tạo Data Cube |  |  |  | X |
| Đặt câu hỏi | Đặt câu hỏi cho trực quan | Đặt câu hỏi cho phân tích | Đặt câu hỏi cho phân tích |  |
| Trả lời bằng Pivot Table |  |  |  | X |
| Trực quan hóa dữ liệu bằng PowerBI | X |  |  |  |

**MỤC LỤC**

[**1. Giới thiệu chung** 1](#_heading=h.4nx8g23c14s5)

[**2. Xử lý dữ liệu đầu vào** 2](#_heading=h.l6bevtjh01v0)

[**2.1. Chuyển dữ liệu nguồn thành file CSV** 2](#_heading=h.fyece36atuqo)

[**2.2. Load dữ liệu gốc vào cơ sở dữ liệu trong SQL Server** 3](#_heading=h.alvbc4fazk6w)

[**3. Xây dựng kho dữ liệu** 5](#_heading=h.i8r0j1h0tgo9)

[**3.1. Star shema** 6](#_heading=h.b5rgdt74dmsy)

[**3.2. Cấu trúc triển khai các Business Process** 9](#_heading=h.h3odod22brs6)

[**4. Tải dữ liệu (ETL)** 15](#_heading=h.e2ml6ixb7hg0)

[**4.1. Load dữ liệu sử dụng T-SQL** 16](#_heading=h.7pjp0qlm91lp)

[**4.1.1. Load dữ liệu vào các Dimension** 16](#_heading=h.x0jlhkfrabl0)

[**4.1.2. Load dữ liệu vào các Fact** 26](#_heading=h.5erzi14ezp45)

[**4.2. Load bằng công cụ SSIS** 29](#_heading=h.z4gzvcymszod)

[**4.2.1. Kết Nối** 29](#_heading=h.onvzjwcr057l)

[**4.2.2. Dim Date Time** 30](#_heading=h.dv5fdikrrabi)

[**4.2.3. Business Process: Order Fulfillment & Delivery Performance** 31](#_heading=h.r19wplj9ajsv)

[**4.2.4. Business Process : Sales & Revenue Analysis** 34](#_heading=h.ea6lssr2ohbi)

[**4.2.5. Business Process: Product Seller Peformance** 37](#_heading=h.wrx4ncl5hvur)

[**5. Phân tích Dữ liệu bằng SSAS** 38](#_heading=h.qtyocflghz3i)

[**5.1. Tạo Data Sources và Data Sources View kết nối với OlistDW** 38](#_heading=h.wz6t687atstz)

[**5.2. Tạo Cubes** 39](#_heading=h.gv3evd338w2b)

[**5.2.1. Tạo Cube FactProductSellerPerformance** 39](#_heading=h.3r33joq19uc)

[**5.2.2. Tạo Cube FactSales** 40](#_heading=h.89feua6aoxvn)

[**5.2.3. Tạo Cube FactOrderFullfillment** 41](#_heading=h.fdrl9fit0p4w)

[**5.3. Thiết lập Measures và Dimensions** 42](#_heading=h.9rymo13ytt2)

[**5.4. Triển khai và Xử lý Cube** 43](#_heading=h.791guoxc95q)

[**6. Trả lời câu hỏi bằng Pivot Table** 45](#_heading=h.5r09a5fm98kd)

[**6.1. Kết nối Excel với Cube SSAS** 45](#_heading=h.1aeuwu1si32a)

[**6.2. Trả lời các câu hỏi phân tích** 47](#_heading=h.la11cl6tdb1c)

[**6.2.1. Business Process : Order Fulfillment & Delivery Performance** 47](#_heading=h.pptpkhf2e27s)

[**6.2.2. Business Process : Sales & Revenue Analysis** 49](#_heading=h.o8hepa9fja76)

[**6.2.3. Business Process: Product Seller Peformance** 53](#_heading=h.3v8rv6bctvh3)

[**7. Trực quan hóa bằng Power BI** 57](#_heading=h.x72288qkjhv6)

[**7.1. Thực hiện phân tích dữ liệu cho Business process: Sales & Revenue Analysis** 57](#_heading=h.evcl05ndek5x)

[**7.1.1. Câu hỏi: Cho biết số lượng đơn đặt hàng, doanh thu qua các năm 2017, 2018** 58](#_heading=h.yi9od871zu8w)

[**7.1.2. Câu hỏi: Cho biết sự tăng giảm doanh thu qua các tháng (2017, 2018) dựa theo thị trường/khu vực** 59](#_heading=h.t7727vpt7kn3)

[**7.2. Thực hiện phân tích dữ liệu cho Business process: Order Fulfillment** 60](#_heading=h.fpl31hv5v993)

[**7.2.1. Câu hỏi: Cho biết thời gian giao hàng trung bình qua các tháng của năm 2017 và 2018** 60](#_heading=h.sm58trupaay2)

[**7.2.2. Câu hỏi: Cho biết tỉ lệ đơn giao đúng hạn và trễ hạn qua các tháng của năm 2017 và 2018** 61](#_heading=h.48vm8fmsscnb)

[**7.3. Thực hiện phân tích dữ liệu cho Business process: Product & Seller Performance** 62](#_heading=h.c0a5h1bpw55a)

[**7.3.1. Câu hỏi: Cho biết 5 danh mục sản phẩm bán chạy và đem theo doanh thu nhiều nhất trong năm 2018** 63](#_heading=h.1qsxqb8zeg0u)

[**7.3.2. Câu hỏi: Cho biết 5 danh mục sản phẩm nhận phản hồi tệ nhất trong năm 2018** 64](#_heading=h.u3pqwkhnir3p)

[**7.3.3. Câu hỏi: Cho biết 3 Seller bán được nhiều đơn hàng và đem lại doanh thu nhiều nhất trong năm 2018** 65](#_heading=h.t2b2mqg6mbxl)

[**8. Kết luận** 66](#_heading=h.qtldl6fz9oyf)

[**8.1. Kết quả đạt được** 66](#_heading=h.je1wc7gve2h)

[**8.2. Những hạn chế** 66](#_heading=h.xc1fsgovye50)

[**9. Tài liệu tham khảo** 66](#_heading=h.w4mnmzey1f5)

# **1. Giới thiệu chung**

* Lý do chọn đề tài

Trong thời đại công nghệ số hiện nay, thương mại điện tử (e-commerce) đang phát triển mạnh mẽ và trở thành xu hướng tất yếu trong hoạt động kinh doanh toàn cầu. Việc khai thác và phân tích dữ liệu từ các nền tảng thương mại điện tử giúp các doanh nghiệp hiểu rõ hơn về hành vi tiêu dùng, xu hướng mua sắm, hiệu suất bán hàng, cũng như đánh giá chất lượng dịch vụ. Từ đó, họ có thể đưa ra các chiến lược kinh doanh hiệu quả, tối ưu hóa trải nghiệm khách hàng và tăng khả năng cạnh tranh.

Xuất phát từ nhu cầu thực tiễn đó, nhóm em chọn đề tài **"Xây dựng kho dữ liệu từ hệ thống thương mại điện tử Olist tại Brazil"**. Đây là một bài toán thực tế, giàu tiềm năng phân tích, với dữ liệu phong phú và đa chiều, bao gồm thông tin về khách hàng, đơn hàng, sản phẩm, nhà bán hàng, phương thức vận chuyển, đánh giá dịch vụ, thanh toán,...

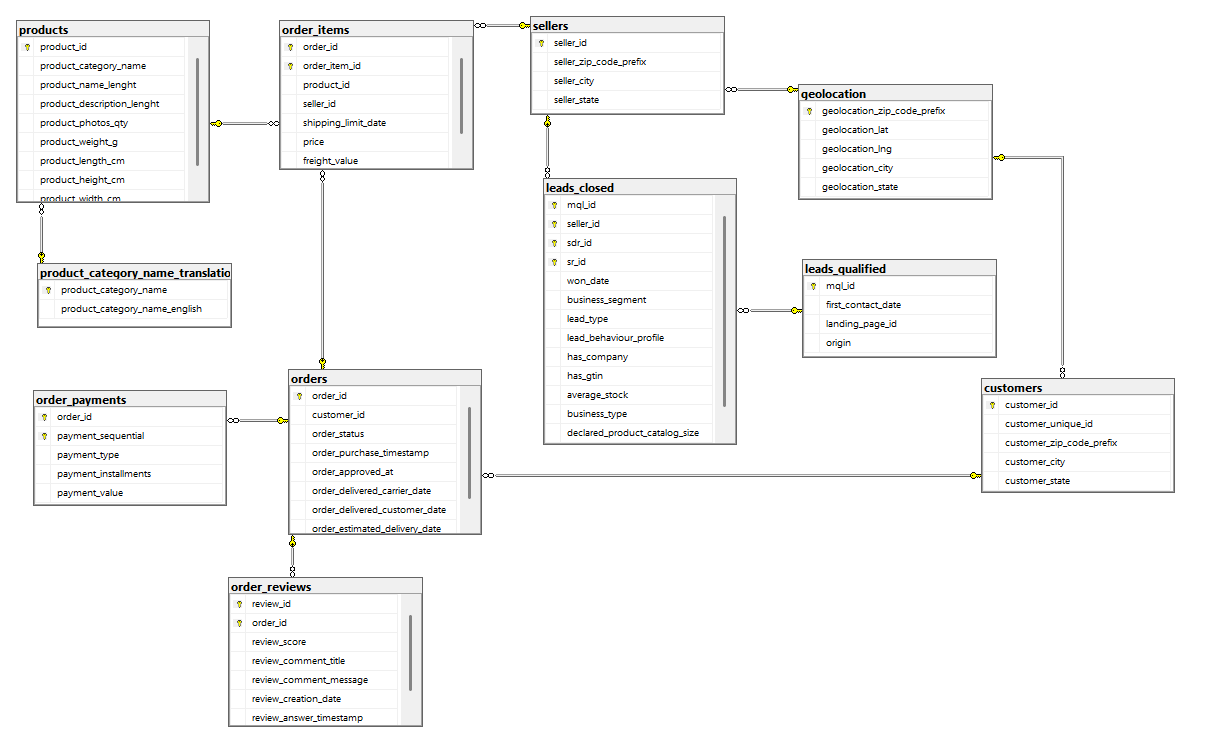
* Nguồn dữ liệu

Tập dữ liệu được sử dụng trong đồ án được lấy từ Kaggle - E-commerce dataset by Olist (SQLite), bao gồm hơn 100.000 đơn hàng được thực hiện từ năm 2016 đến 2018. Đây là một nguồn dữ liệu thực tế, đã được xử lý sơ bộ, phù hợp với mục tiêu xây dựng kho dữ liệu và khai thác thông tin nhằm hỗ trợ ra quyết định trong lĩnh vực thương mại điện tử.

Đường dẫn kaggle: https://www.kaggle.com/datasets/terencicp/e-commerce-dataset-by-olist-as-an-sqlite-database

* Tập dữ liệu gồm các bảng và thông tin từng bảng

|  |  |  |
| --- | --- | --- |
| Tên bảng | Số dòng | Số cột |
| customers | 99441 | 5 |
| geolocation | 1000163 | 5 |
| leads\_closed | 842 | 14 |
| leads\_qualified | 8000 | 4 |
| order\_items | 112650 | 7 |
| order\_payments | 103886 | 5 |
| order\_reviews | 99224 | 7 |
| orders | 99441 | 8 |
| product\_category\_name\_translation | 71 | 2 |
| products | 32951 | 9 |
| sellers | 3095 | 4 |



# **2. Xử lý dữ liệu đầu vào**

## **2.1. Chuyển dữ liệu nguồn thành file CSV**

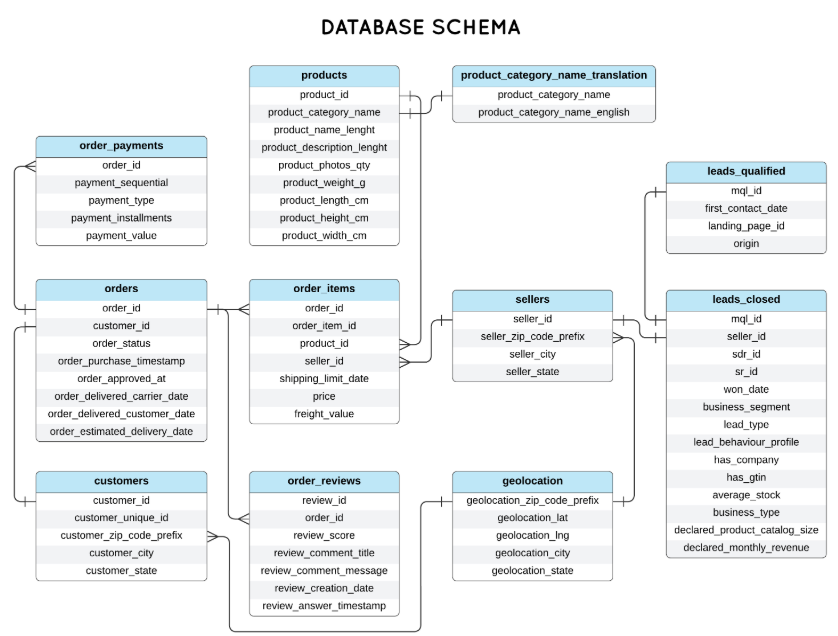
* Dữ liệu: Là 1 file dạng olist.sqlite
* Công cụ thực hiện (Python, SQLite) để xuất file CSV từ các bảng trong tập olist.sqlite



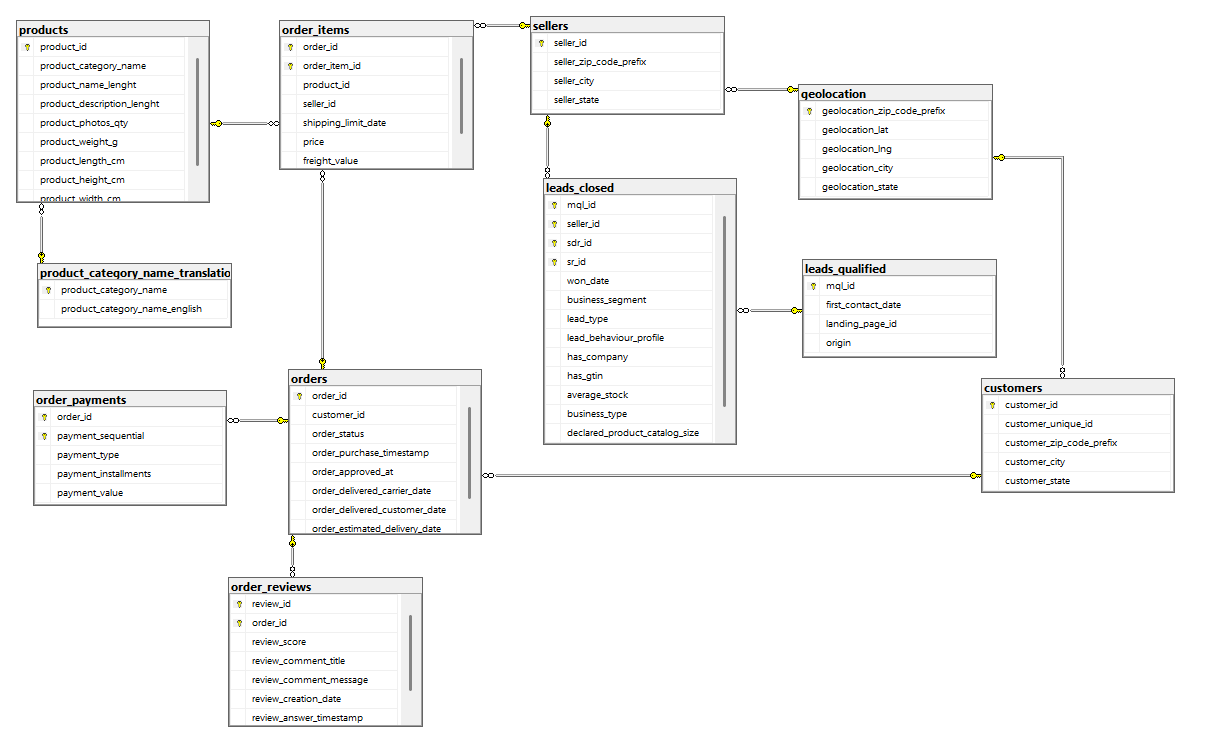
## **2.2. Load dữ liệu gốc vào cơ sở dữ liệu trong SQL Server**

**2.2.1. Tạo cơ sở dữ liệu từ trong SQL Server dựa trên dữ liệu nguồn**

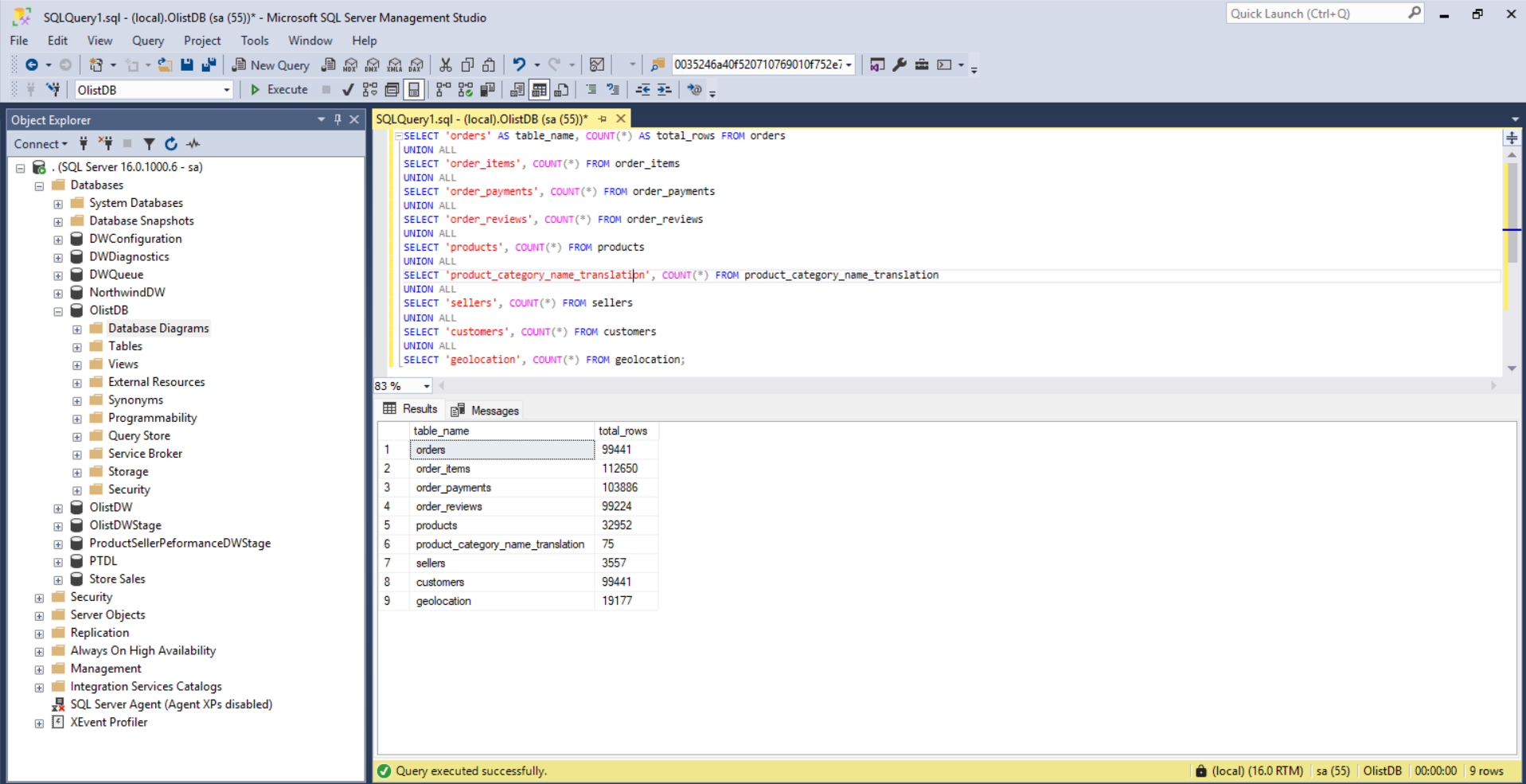
* Lược đồ SQL Lite của dữ liệu nguồn



* Lược đồ trong SQL Server được tạo lại



* Thực hiện load dữ liệu từ CSV vào database và xem kết quả



# **3. Xây dựng kho dữ liệu**

Olist – một nền tảng thương mại điện tử tổng hợp tại Brazil – cung cấp hệ sinh thái bán hàng cho các nhà cung cấp thông qua việc kết nối họ với nhiều marketplace khác nhau. Do đó, lượng dữ liệu phát sinh từ các giao dịch, đánh giá, thanh toán, vận chuyển… là rất lớn và phong phú, đòi hỏi một hệ thống lưu trữ và phân tích hiệu quả.

Tuy nhiên, dữ liệu gốc của Olist hiện được tổ chức dưới dạng các bảng transactional rời rạc như orders, order\_items, customers, sellers, order\_reviews, order\_payments,... Dạng dữ liệu này tuy phù hợp với hệ thống vận hành (OLTP), nhưng lại gây khó khăn cho các phân tích tổng hợp (OLAP) do thiếu tính tổ chức và khó truy vấn hiệu suất cao.

Chính vì vậy, việc xây dựng kho dữ liệu (Data Warehouse) cho Olist là cần thiết. Kho dữ liệu giúp tổ chức dữ liệu theo mô hình Star Schema, trong đó các bảng fact trung tâm kết nối với các bảng dimension mô tả, tạo điều kiện cho việc phân tích dễ dàng, trực quan và hiệu quả hơn.

Dựa trên dữ liệu Olist đã được cung cấp, đồ án này lựa chọn ba business process đặc trưng đại diện cho ba khía cạnh quan trọng trong hoạt động thương mại điện tử để xây dựng mô hình kho dữ liệu:

* **Order Fulfillment & Delivery Performance**: phân tích hiệu suất giao hàng để đánh giá độ hài lòng khách hàng và hiệu quả vận hành.
* **Sales & Revenue Analysis**: đánh giá doanh thu theo sản phẩm, thời gian, seller, và chi phí logistics.
* **Product & Seller Performance**: theo dõi chất lượng sản phẩm và hiệu suất bán hàng để hỗ trợ chiến lược kinh doanh.

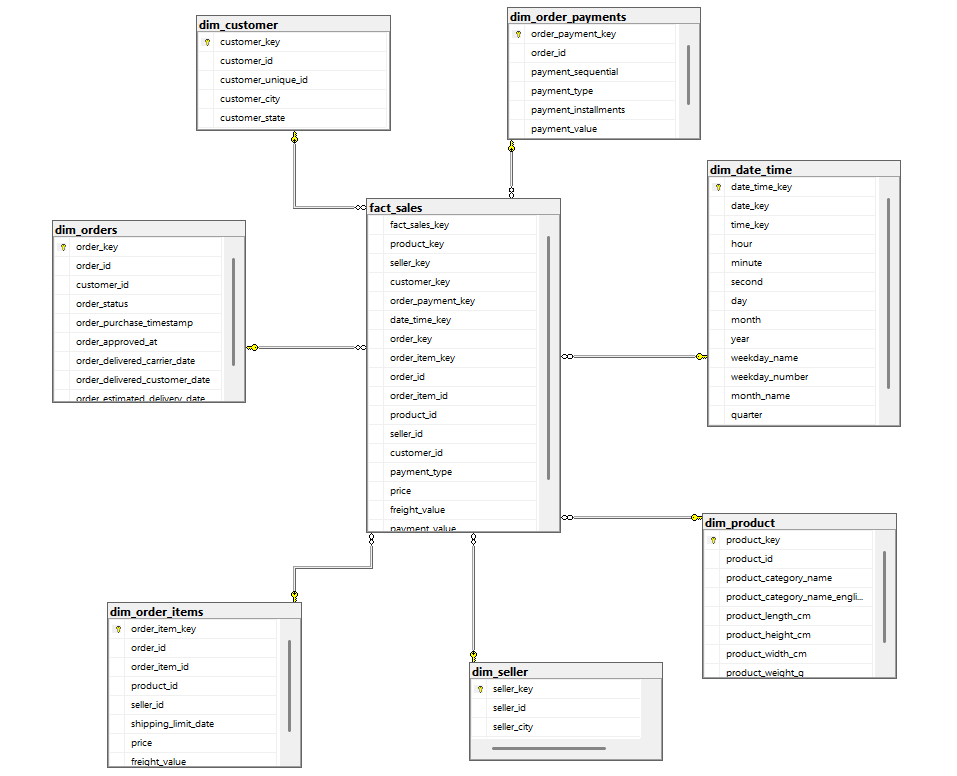
Mỗi business process đều được thiết kế tương ứng với một bảng fact chứa dữ liệu định lượng, cùng các bảng dimension chứa dữ liệu mô tả như khách hàng, sản phẩm, nhà bán hàng, thời gian, phương thức thanh toán,... Từ đó, hệ thống có thể trả lời được nhiều câu hỏi nghiệp vụ cụ thể như:

* Thời gian giao hàng trung bình là bao lâu?
* Doanh thu theo loại sản phẩm và theo thời gian ra sao? Seller nào tạo doanh thu cao nhất?
* Sản phẩm nào bị đánh giá thấp nhất? Seller nào thường xuyên bị khách hàng chê trách?

Việc triển khai kho dữ liệu giúp đơn giản hóa truy vấn, hỗ trợ trực quan hóa dữ liệu và tạo nền tảng vững chắc cho các hệ thống phân tích BI (Business Intelligence), từ đó giúp doanh nghiệp đưa ra quyết định nhanh chóng và chính xác hơn.

## **3.1. Star shema**

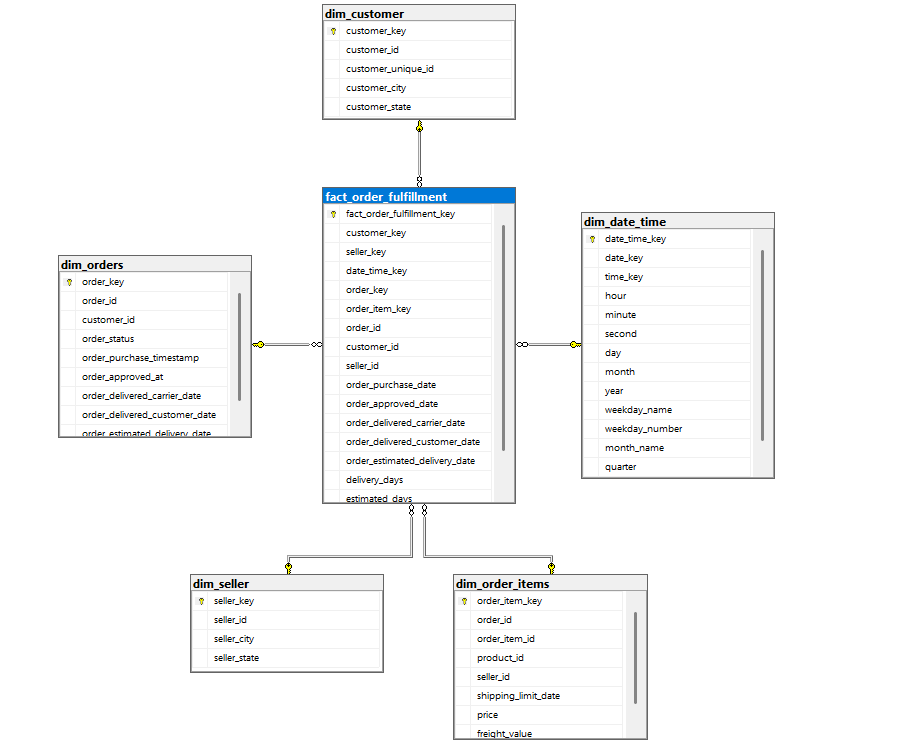
* Thiết kế cho Fact Sales



* Thiết kế cho Fact Product Seller Performance



* Thiết kế cho Fact Product Order Fullfillment

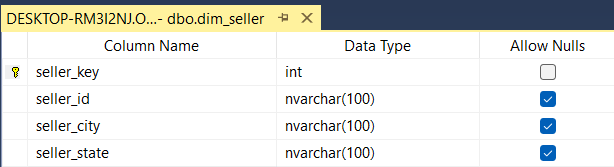


## **3.2. Cấu trúc triển khai các Business Process**

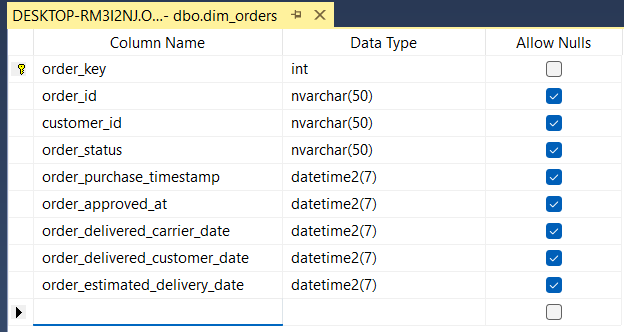
**3.2.1. Dim Customer**



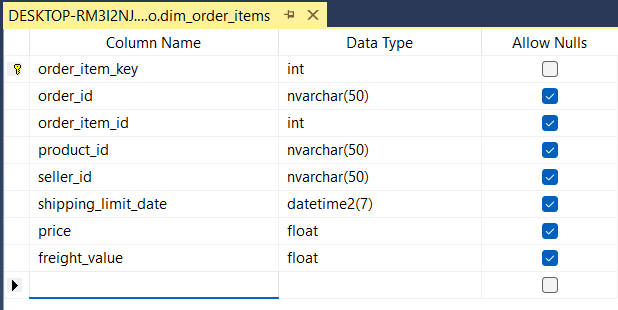
**3.2.2. Dim Seller**



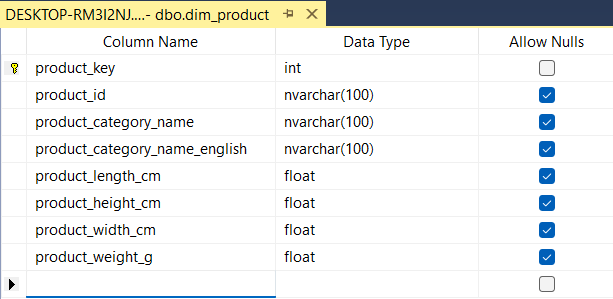
**3.2.3. Dim Orders**



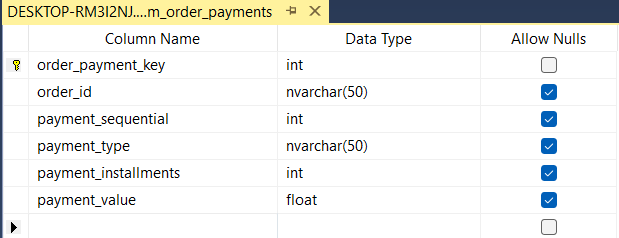
**3.2.4. Dim Order Items**



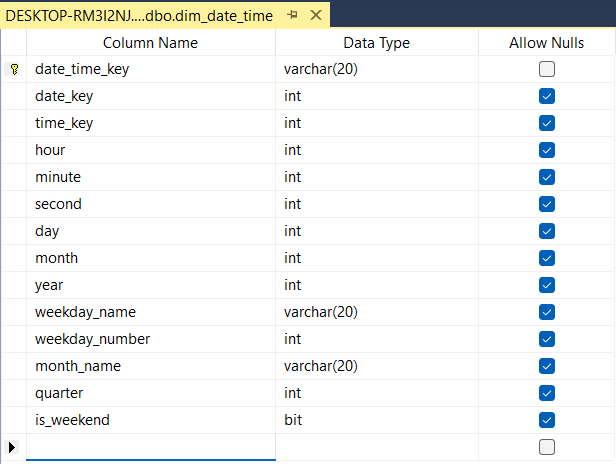
**3.2.5. Dim Product**



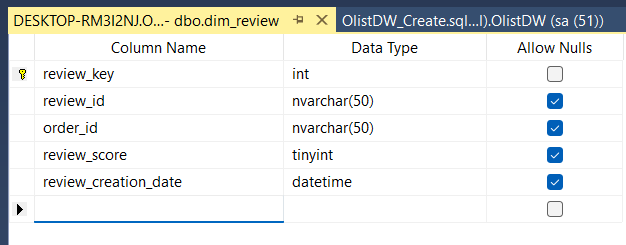
**3.2.6. Dim Order Payments**

****

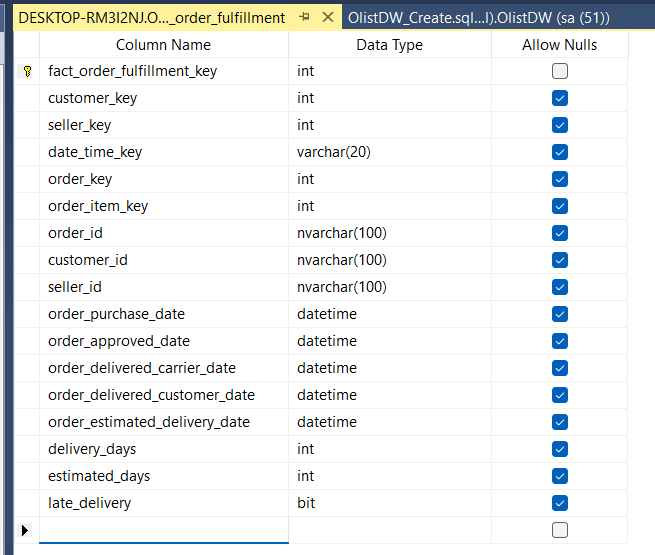
**3.2.7. Dim Date Time**



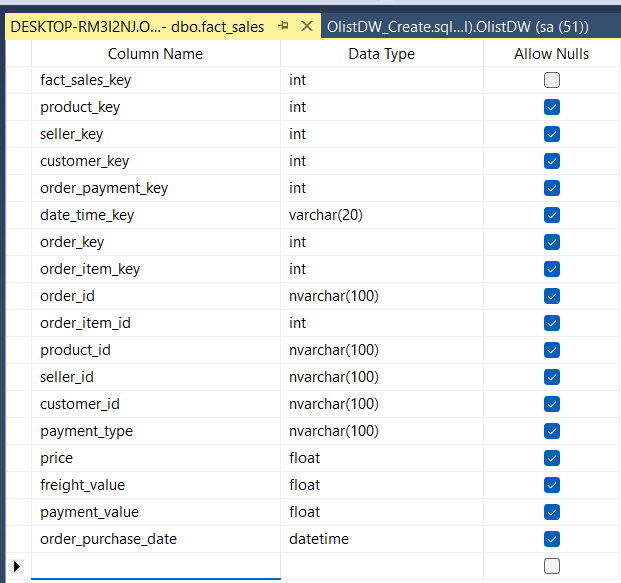
**3.2.8. Dim Review**



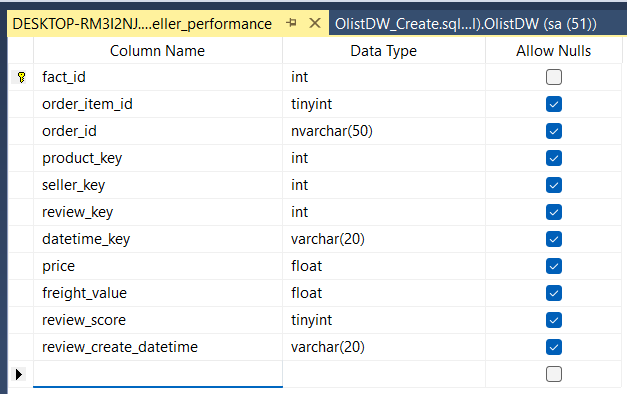
**3.2.9. Fact Order Fulfillment**



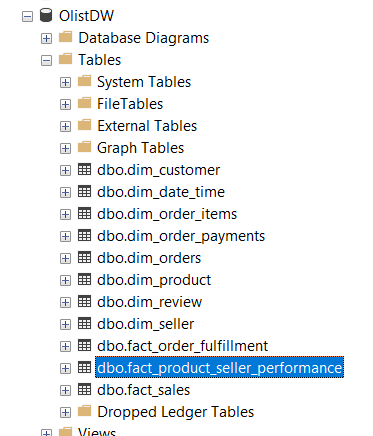
**3.2.10. Fact Sales**

****

**3.2.11 Fact Product Seller Performance**



* Cấu trúc tổng hợp



# **4. Tải dữ liệu (ETL)**

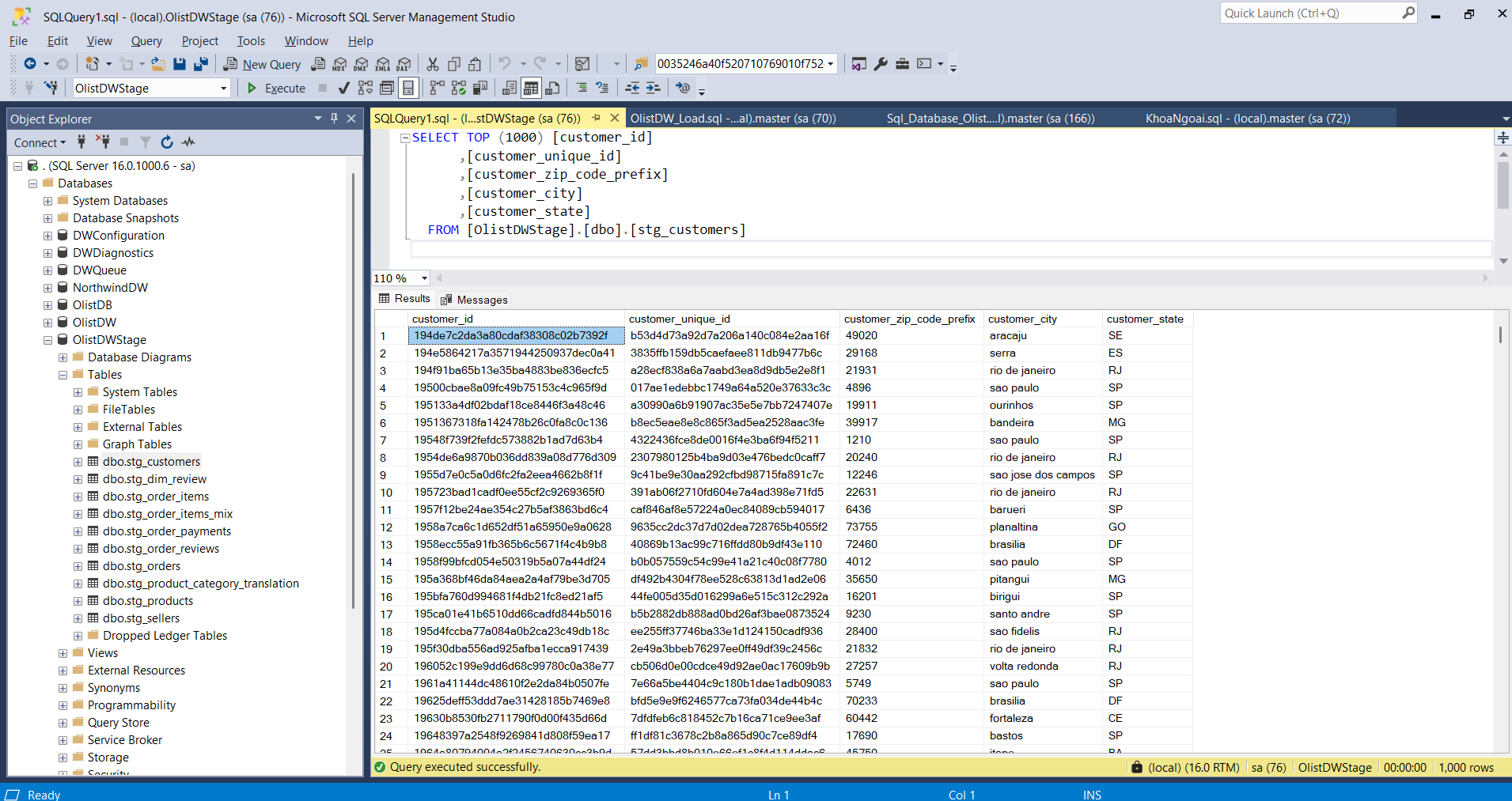
## **4.1. Load dữ liệu sử dụng T-SQL**

### **4.1.1. Load dữ liệu vào các Dimension**

**4.1.1.1. Dim Customer**

**Tạo Dim Customer Stage**

SELECT \* INTO stg\_customers FROM OlistDB.dbo.customers;

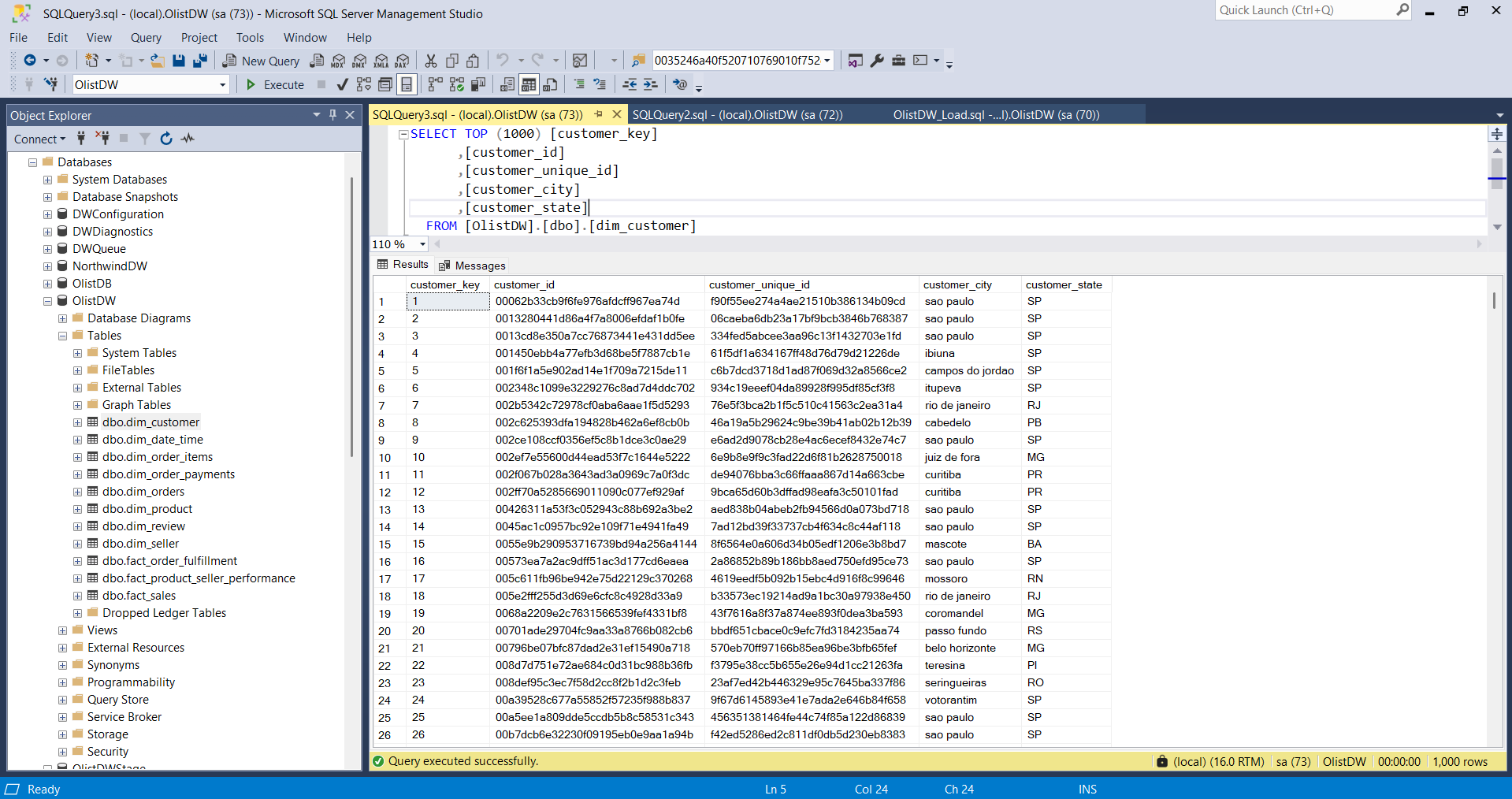


**Load vào dim\_customer**

INSERT INTO dbo.dim\_customer (customer\_id, customer\_unique\_id, customer\_city, customer\_state)

SELECT DISTINCT customer\_id, customer\_unique\_id, customer\_city, customer\_state

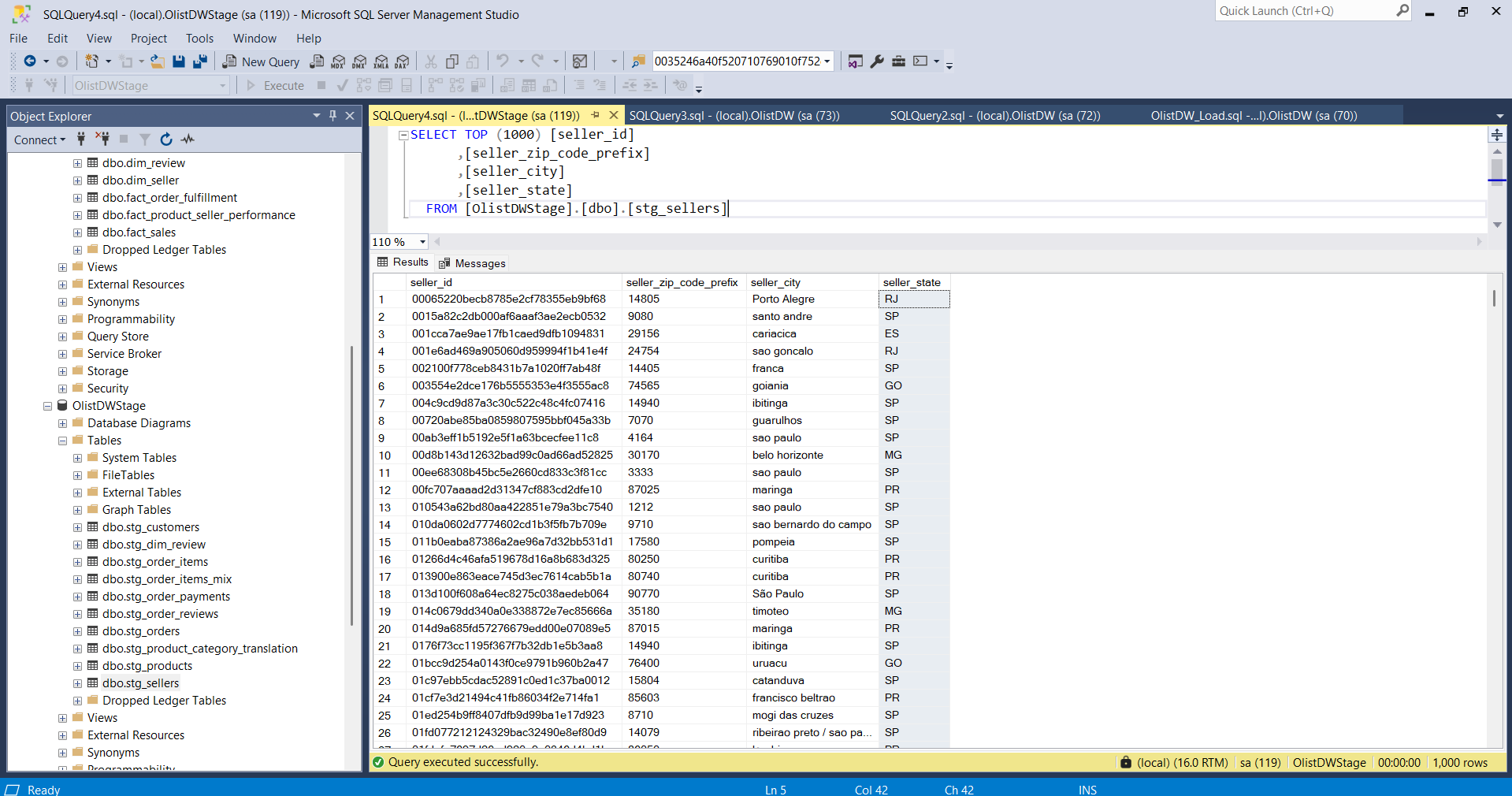
FROM OlistDWStage.dbo.stg\_customers;



**4.1.1.2. Dim Seller**

**Tạo Dim Seller Stage**

SELECT \* INTO stg\_sellers FROM OlistDB.dbo.sellers;

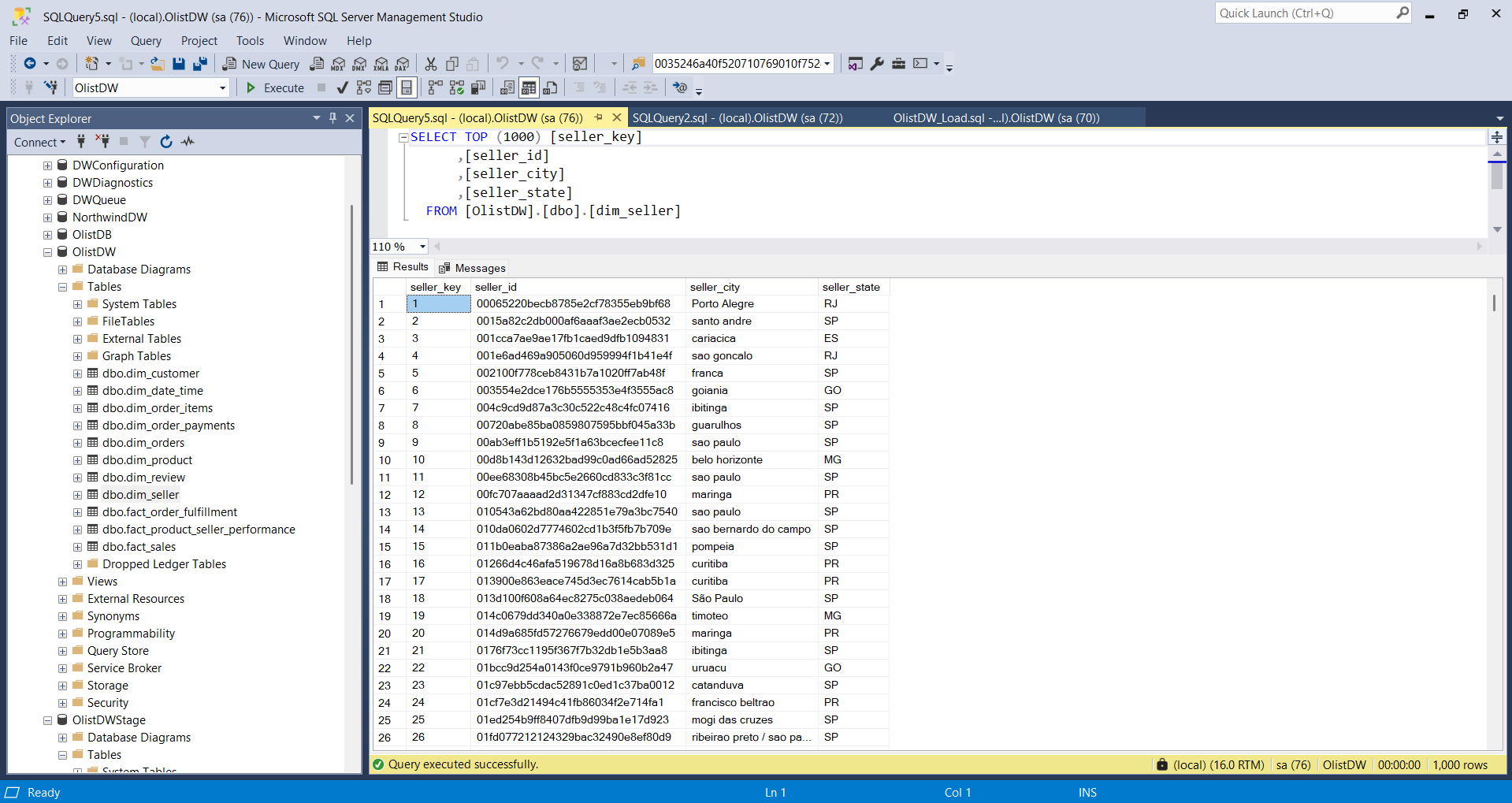
****

**Load vào dim\_seller**

INSERT INTO dbo.dim\_seller (seller\_id, seller\_city, seller\_state)

SELECT DISTINCT seller\_id, seller\_city, seller\_state

FROM OlistDWStage.dbo.stg\_sellers;

****

**4.1.1.3. Dim Date**

**Load dim\_date\_time**

BULK INSERT dim\_date\_time

FROM 'C:\Users\Asus\Desktop\Data\_Engineer\Kho\_du\_lieu\project\olist\_csv\_export\DimDateTime.csv'

WITH (

FORMAT = 'CSV',

FIRSTROW = 2,

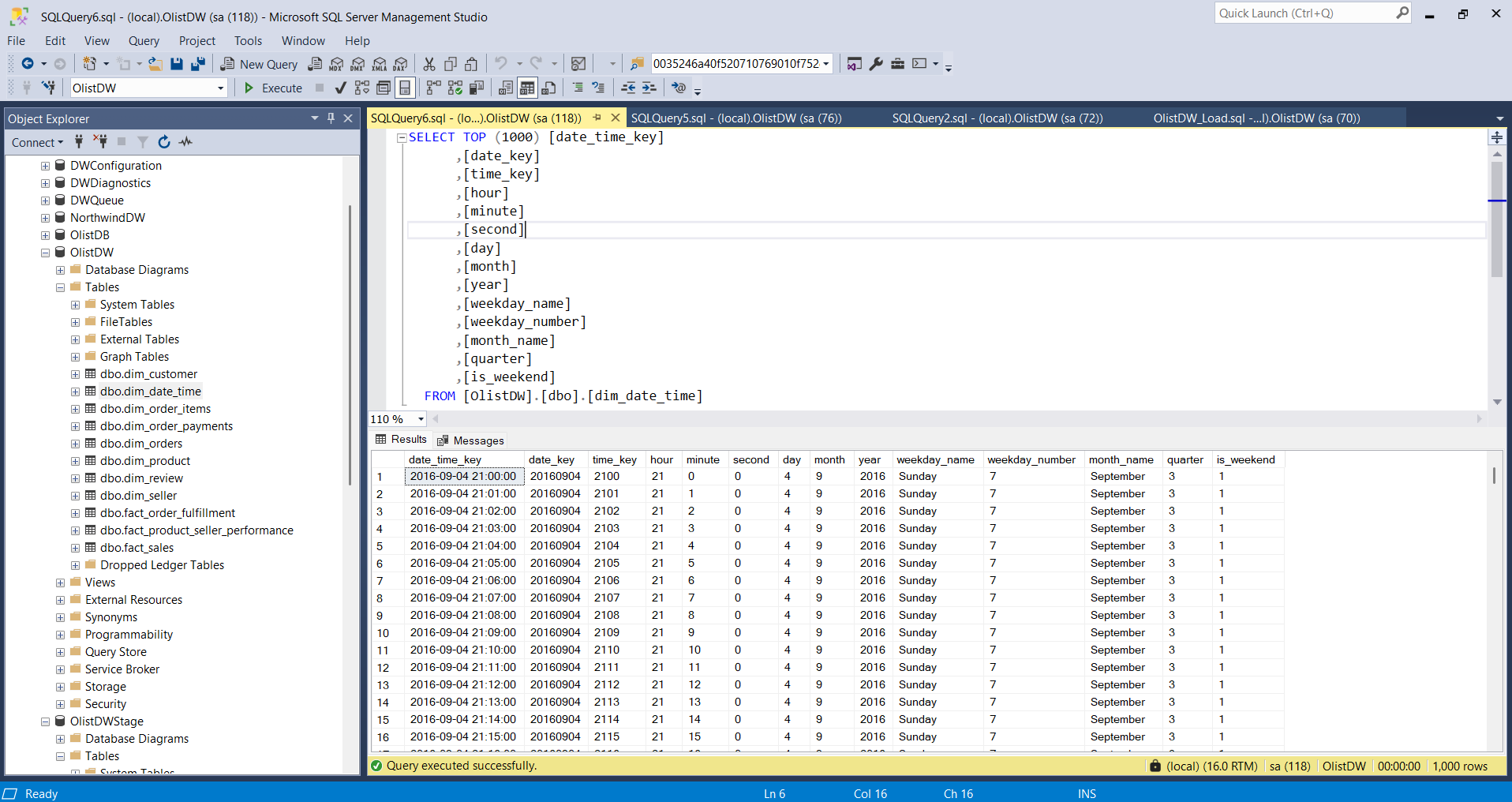
FIELDTERMINATOR = ',',

ROWTERMINATOR = '\n',

TABLOCK,

ERRORFILE = 'C:\Users\Asus\Desktop\DimDateTime\_ErrorRows.log'

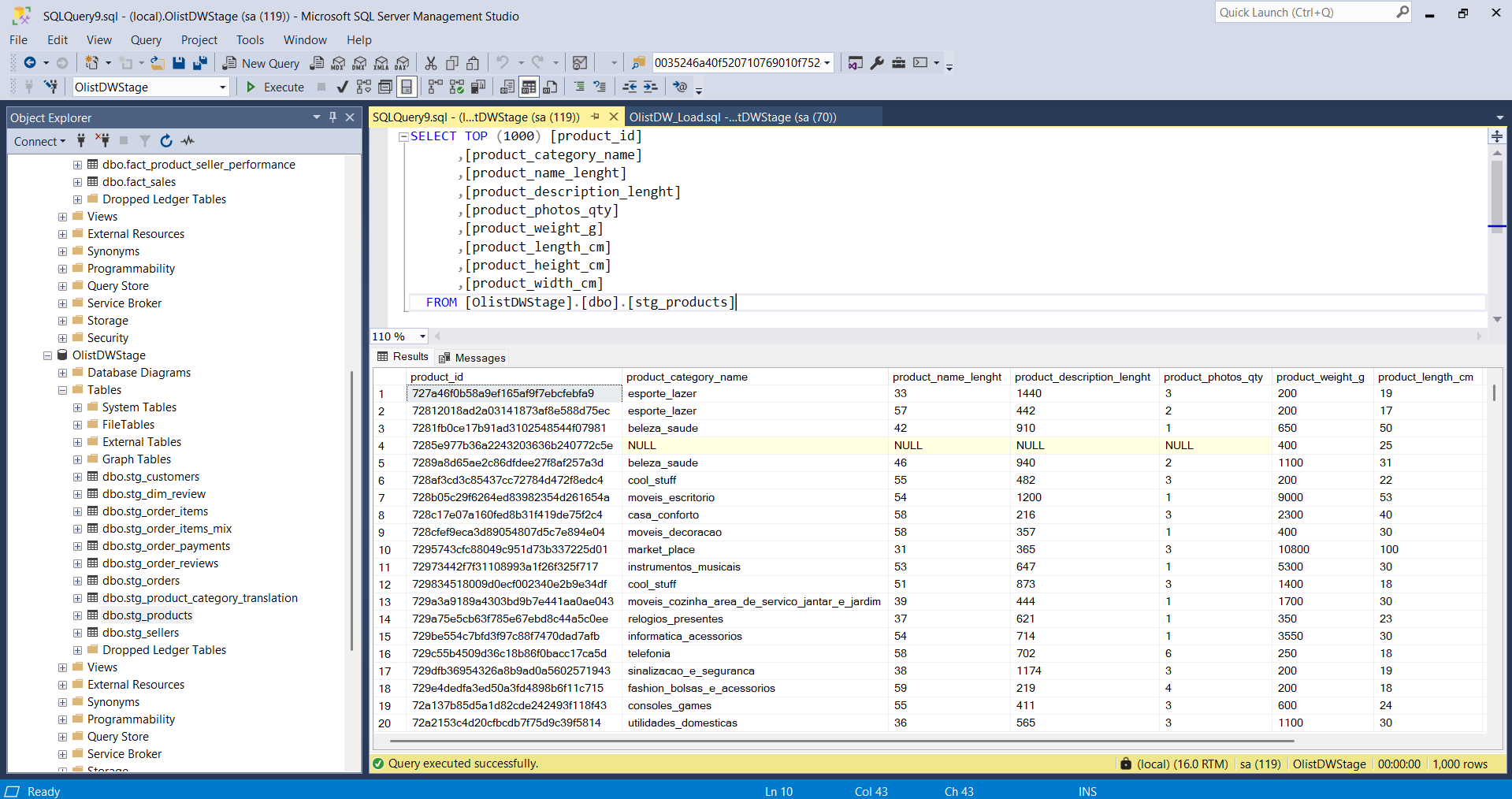
);



**4.1.1.4. Dim Product**

**Tạo Dim Product Stage**

SELECT \* INTO stg\_products FROM OlistDB.dbo.products;



**Load vào dim\_product**

INSERT INTO dbo.dim\_product (

product\_id, product\_category\_name, product\_category\_name\_english,

product\_length\_cm, product\_height\_cm, product\_width\_cm, product\_weight\_g

)

SELECT

p.product\_id,

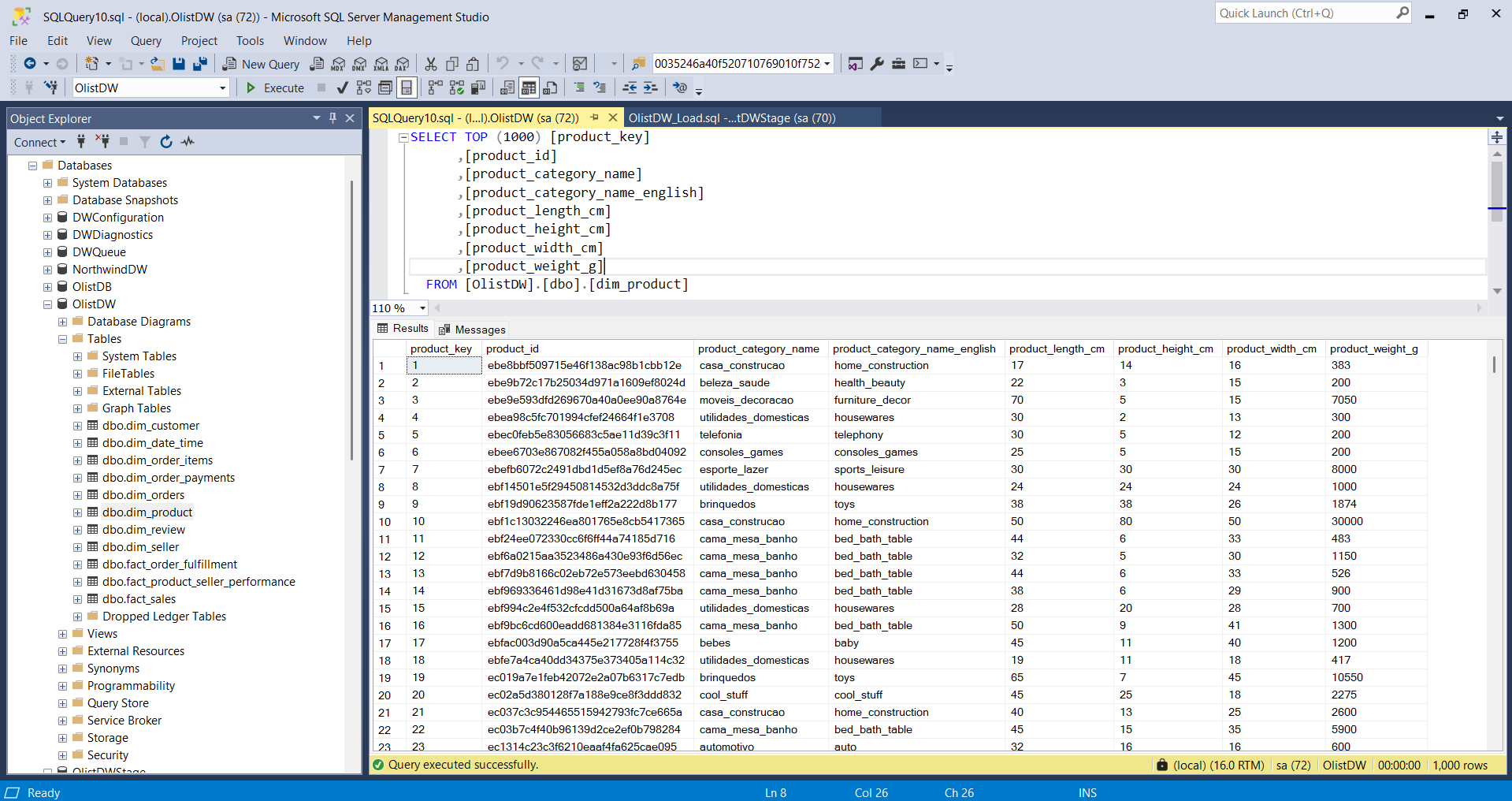
p.product\_category\_name,

t.product\_category\_name\_english,

p.product\_length\_cm, p.product\_height\_cm, p.product\_width\_cm, p.product\_weight\_g

FROM OlistDWStage.dbo.stg\_products p

JOIN OlistDWStage.dbo.stg\_product\_category\_translation t ON p.product\_category\_name = t.product\_category\_name;



**4.1.1.5. Dim Review**

**Tạo Dim Review Stage**

SELECT

review\_id,

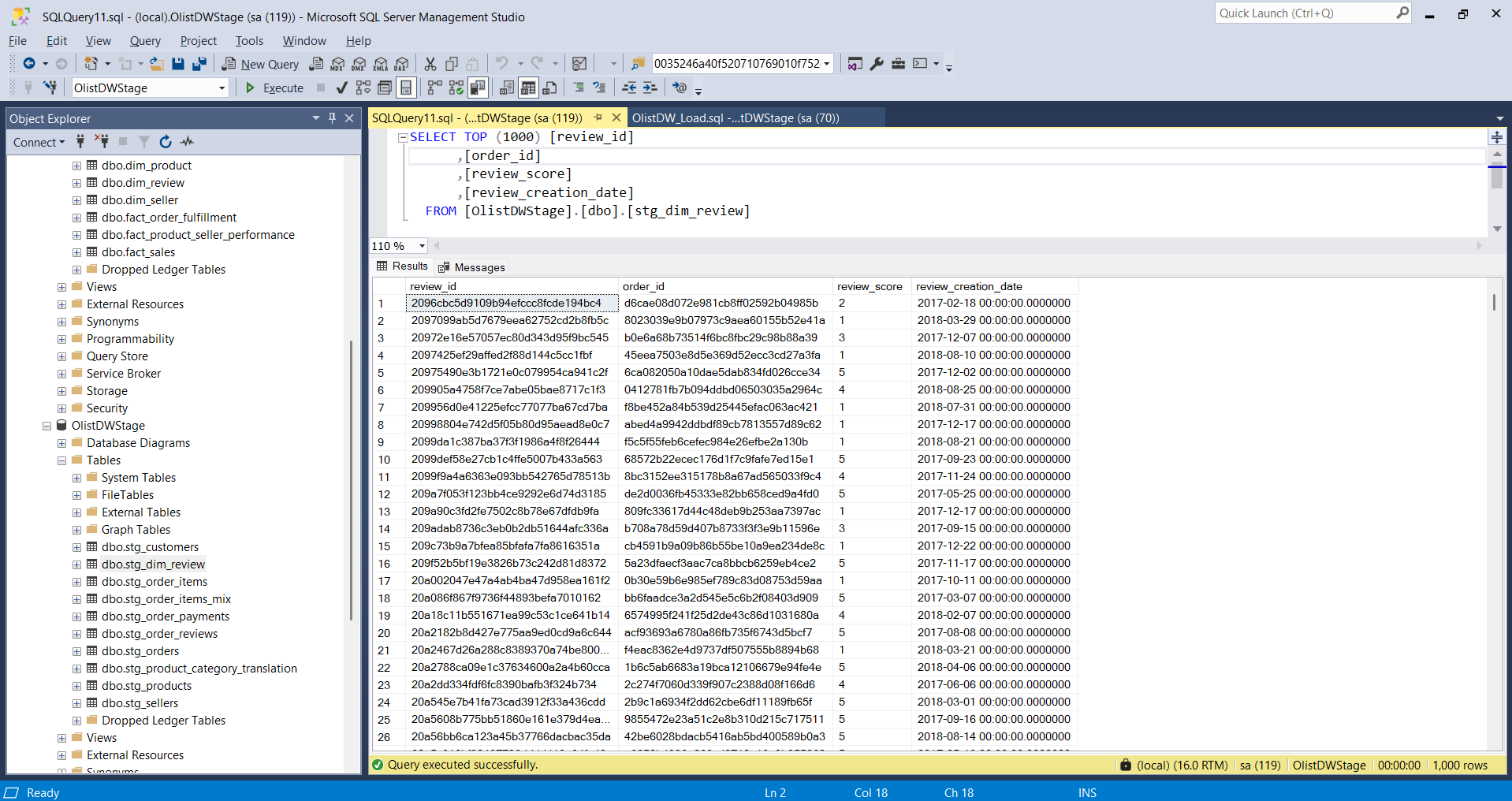
order\_id,

review\_score,

review\_creation\_date

INTO dbo.stg\_dim\_review

FROM OlistDB.dbo.order\_reviews;

****

**Load dim\_review**

INSERT INTO dim\_review (

review\_id,

order\_id,

review\_score,

review\_creation\_date

)

SELECT

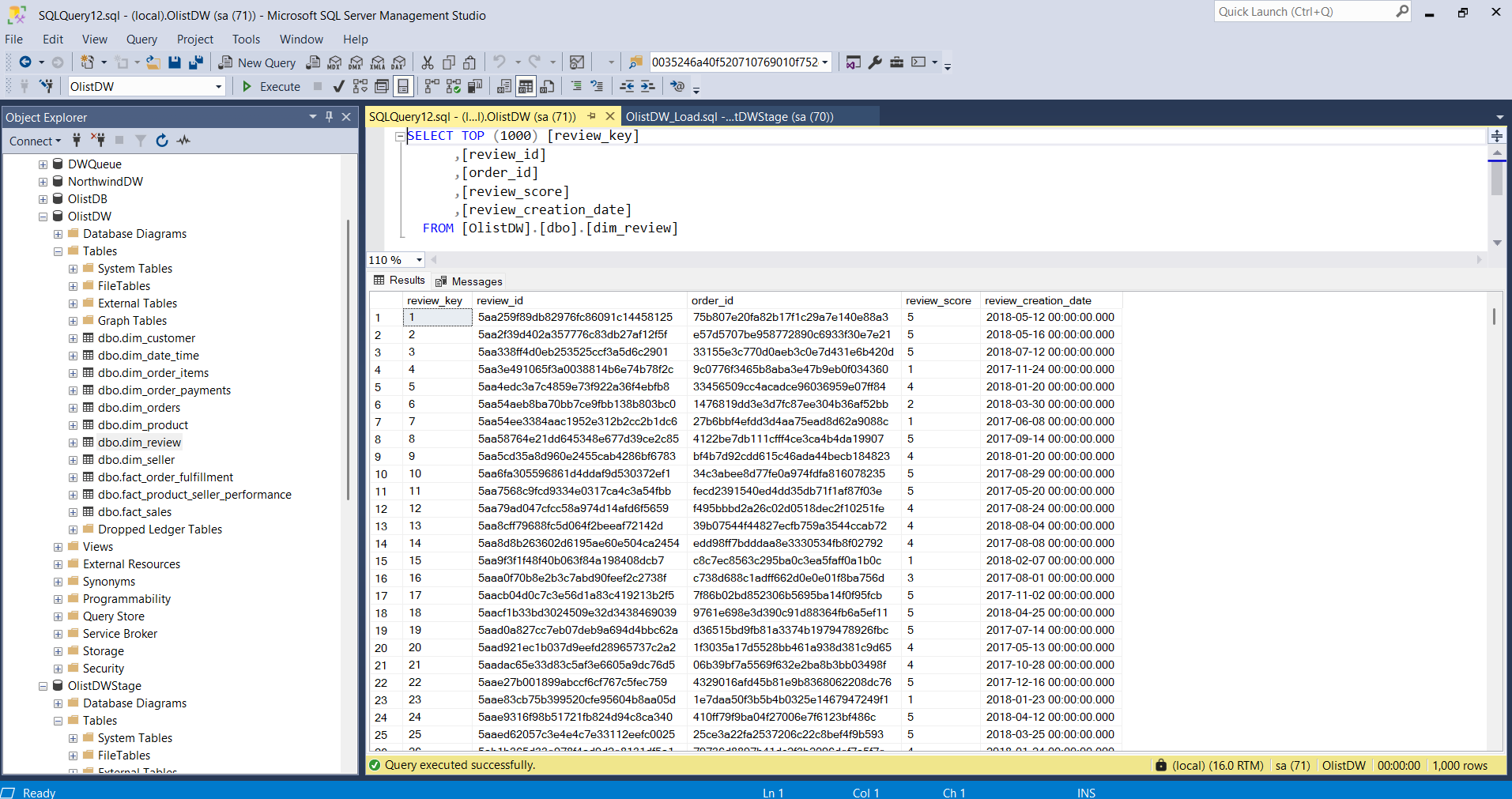
review\_id,

order\_id,

review\_score,

review\_creation\_date

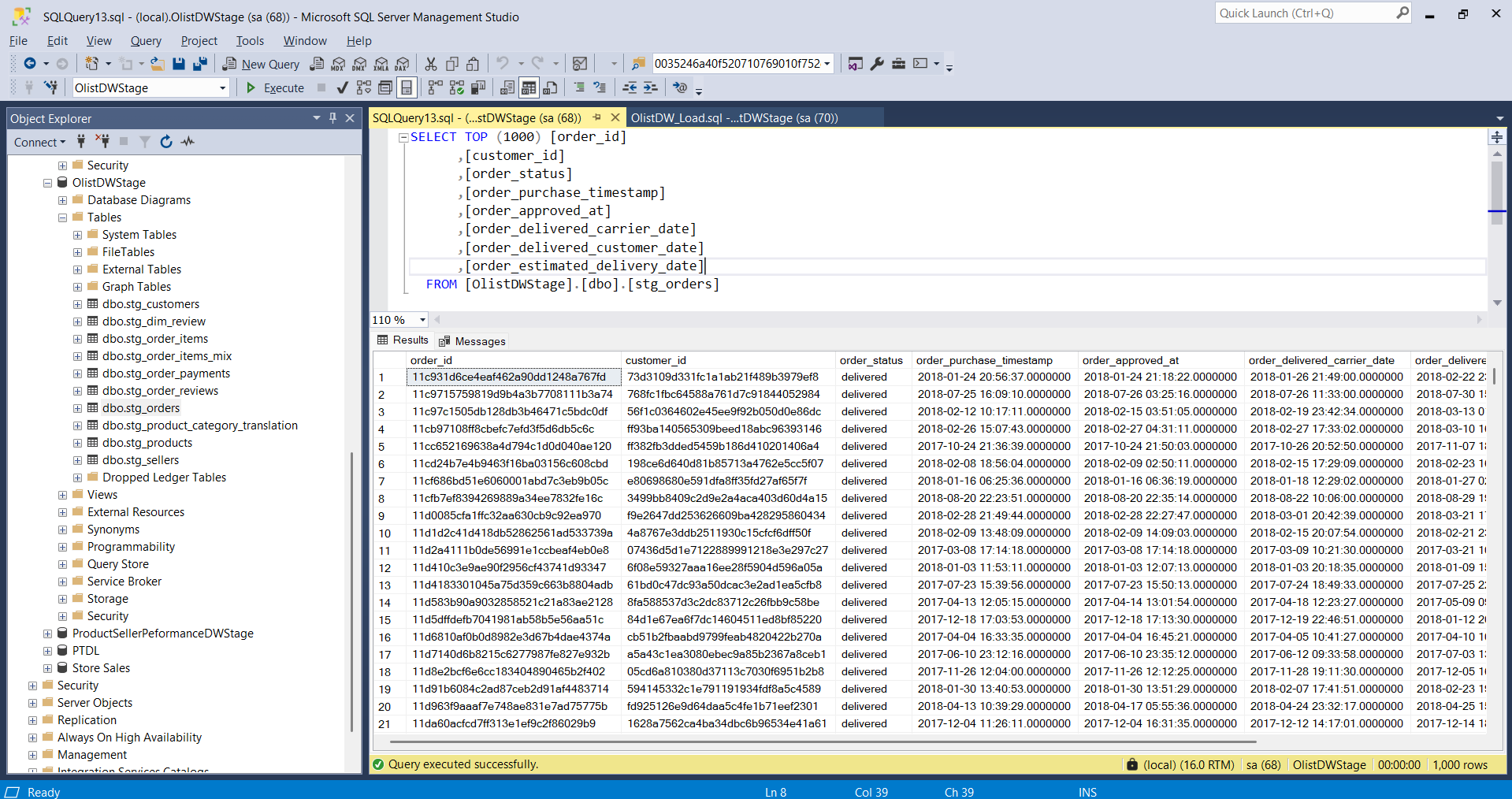
FROM OlistDWStage.dbo.stg\_dim\_review;

****

**4.1.1.6. Dim Orders**

**Tạo Dim Orders Stage**

SELECT \* INTO stg\_orders FROM OlistDB.dbo.orders;

****

**Load vào dim\_orders**

INSERT INTO dbo.dim\_orders (

order\_id, customer\_id, order\_status,

order\_purchase\_timestamp, order\_approved\_at,

order\_delivered\_carrier\_date, order\_delivered\_customer\_date,

order\_estimated\_delivery\_date

)

SELECT

order\_id,

customer\_id,

order\_status,

order\_purchase\_timestamp,

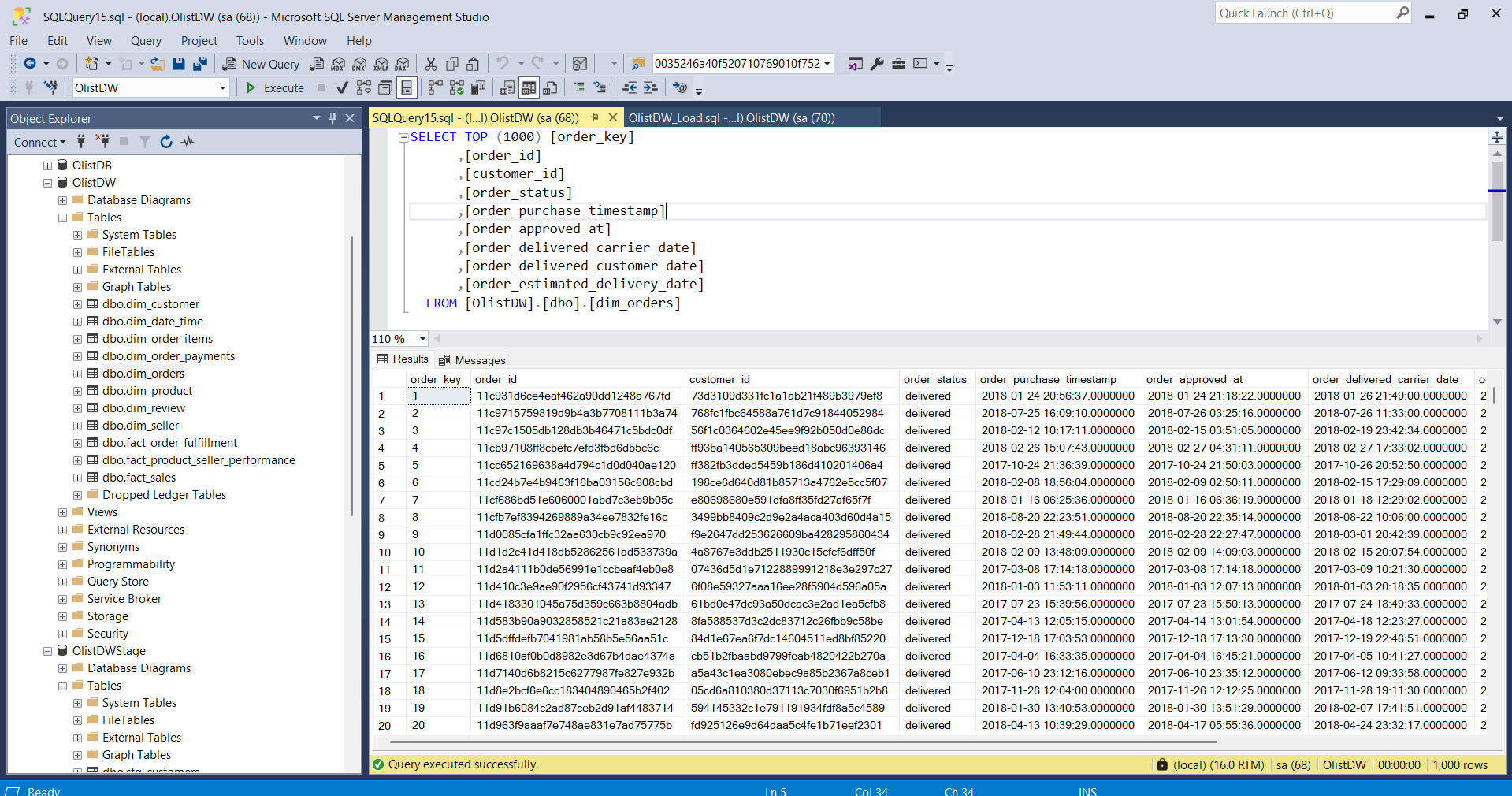
order\_approved\_at,

order\_delivered\_carrier\_date,

order\_delivered\_customer\_date,

order\_estimated\_delivery\_date

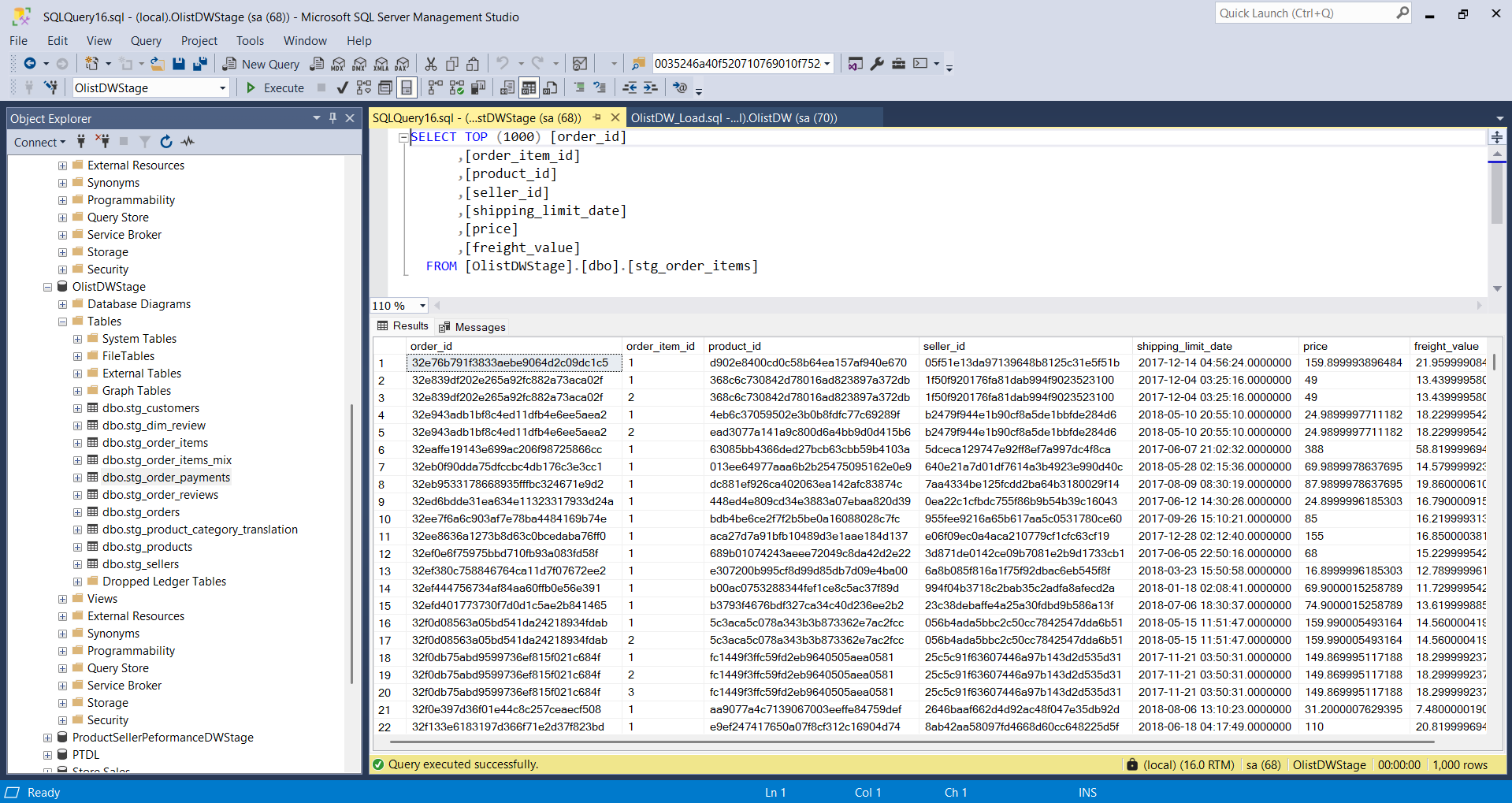
FROM OlistDWStage.dbo.stg\_orders;



**4.1.1.7. Dim Orders Items**

**Tạo Dim Orders Items Stage**

SELECT \* INTO stg\_order\_items FROM OlistDB.dbo.order\_items;



**Load vào dim\_order\_items**

INSERT INTO dbo.dim\_order\_items (

order\_id, order\_item\_id, product\_id, seller\_id,

shipping\_limit\_date, price, freight\_value

)

SELECT

order\_id,

order\_item\_id,

product\_id,

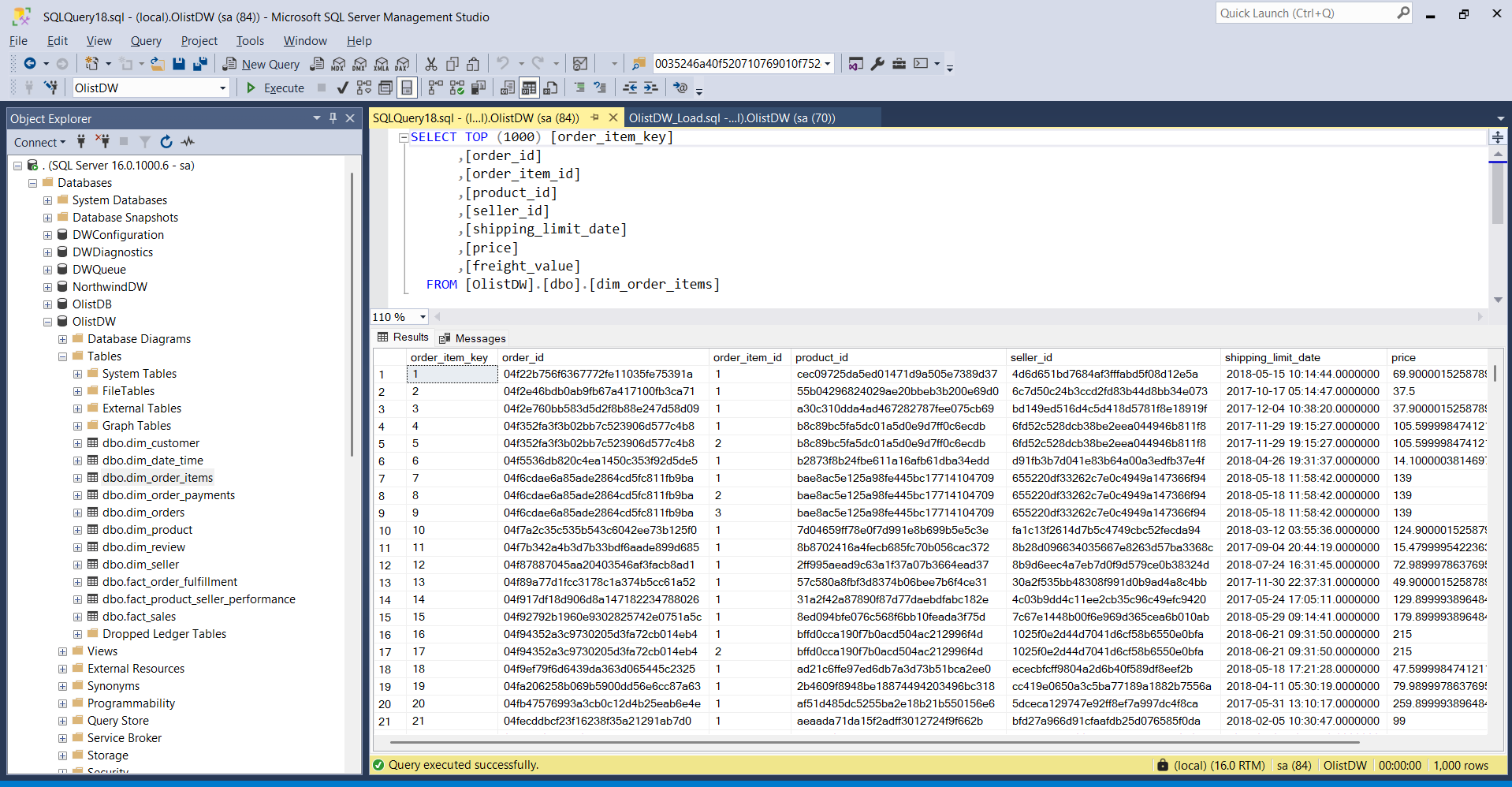
seller\_id,

shipping\_limit\_date,

price,

freight\_value

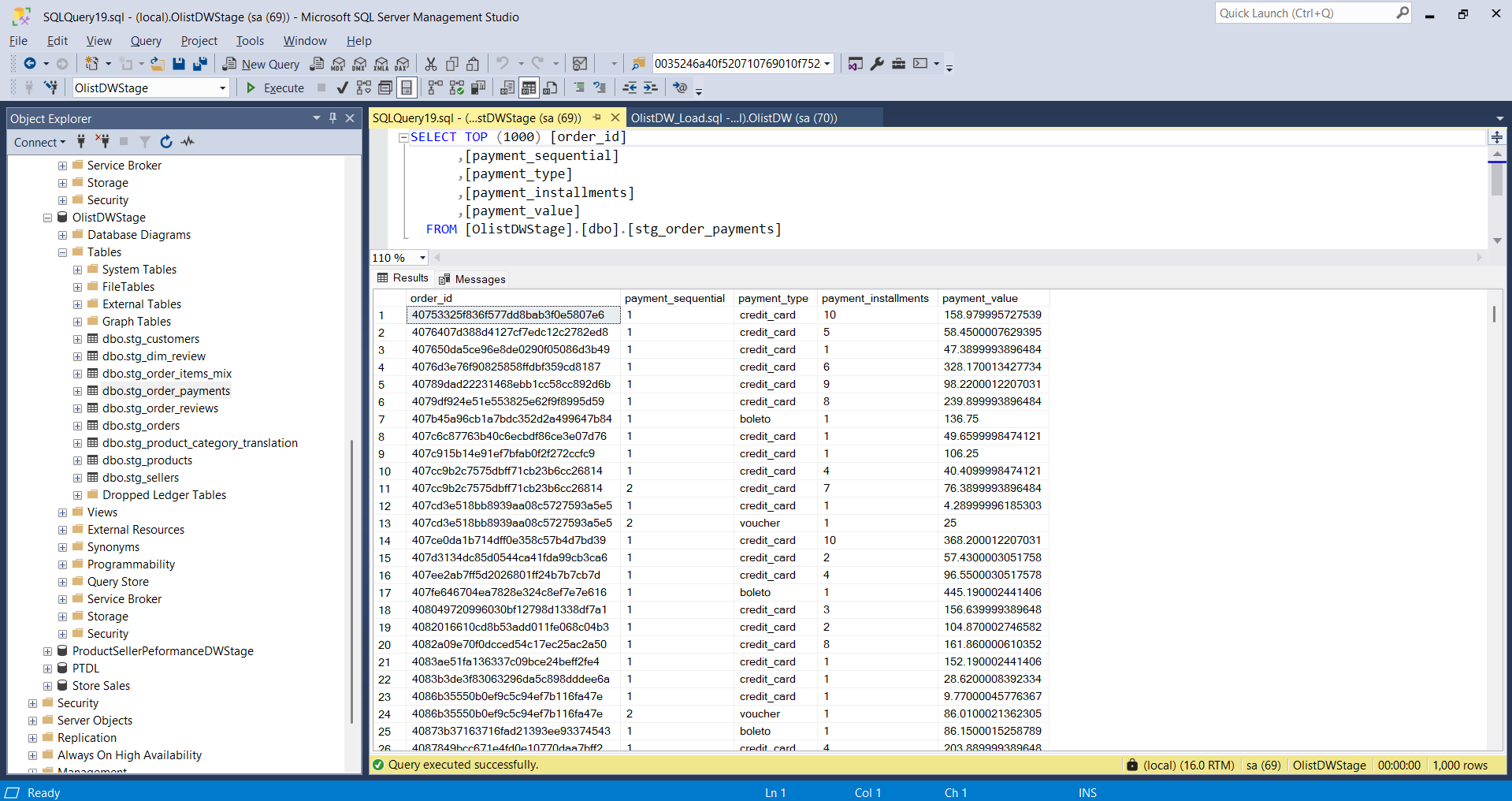
FROM OlistDWStage.dbo.stg\_order\_items;



**4.1.1.8. Dim Orders Payment**

**Tạo Orders Payment Stage**

SELECT \* INTO stg\_order\_payments FROM OlistDB.dbo.order\_payments;



**Load vào dim\_order\_payments**

INSERT INTO dim\_order\_payments (

order\_id,

payment\_sequential,

payment\_type,

payment\_installments,

payment\_value

)

SELECT

order\_id,

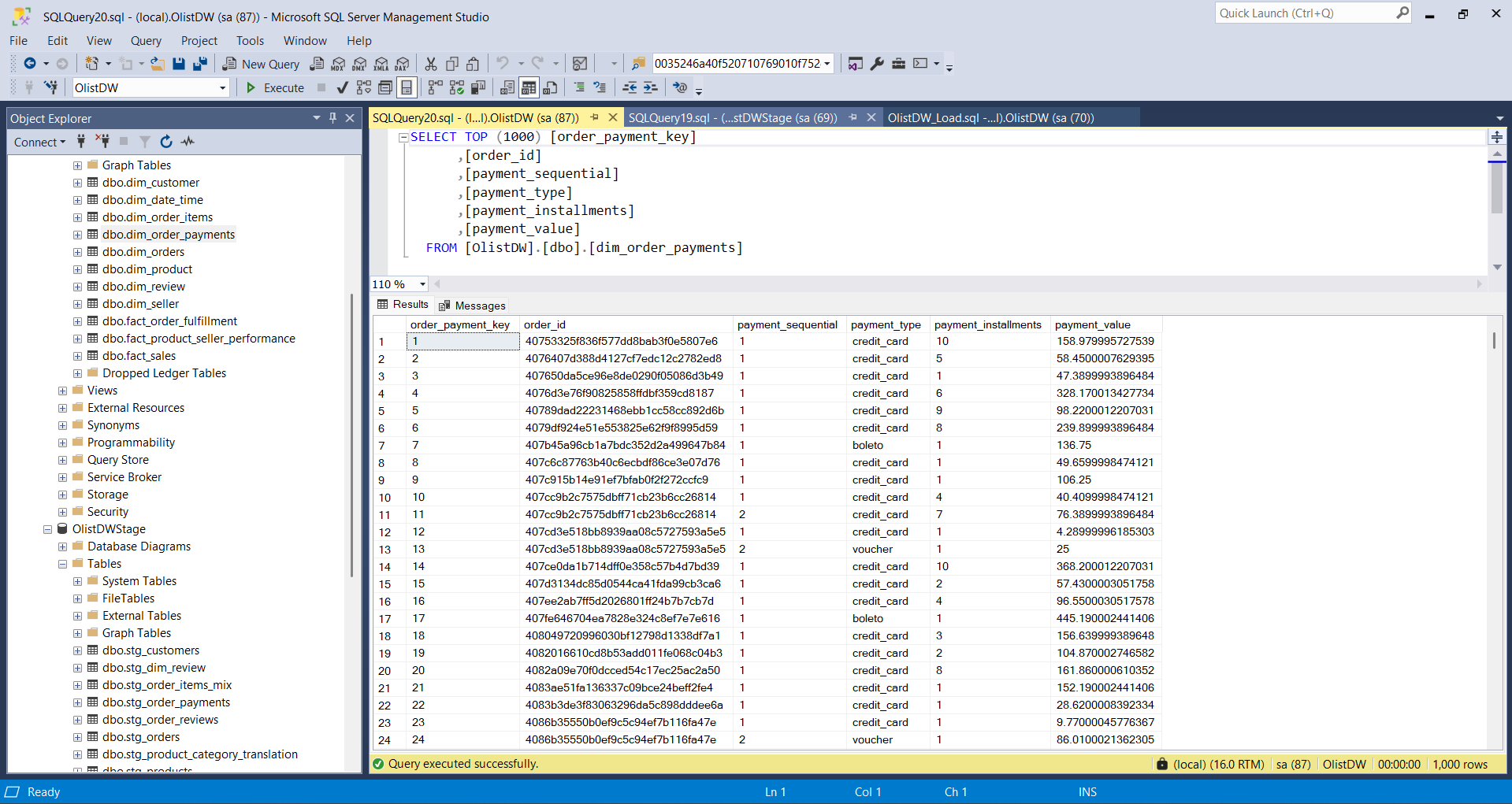
payment\_sequential,

payment\_type,

payment\_installments,

payment\_value

FROM OlistDWStage.dbo.stg\_order\_payments;



### **4.1.2. Load dữ liệu vào các Fact**

**4.1.2.1. Fact Order Fulfillment**

INSERT INTO dbo.fact\_order\_fulfillment (

customer\_key, seller\_key, date\_time\_key,

order\_key, order\_item\_key, order\_id,

customer\_id, seller\_id,

order\_purchase\_date, order\_approved\_date,

order\_delivered\_carrier\_date, order\_delivered\_customer\_date,

order\_estimated\_delivery\_date,

delivery\_days, estimated\_days, late\_delivery

)

SELECT

dc.customer\_key, ds.seller\_key, d.date\_time\_key,

o.order\_key, oi.order\_item\_key, o.order\_id,

o.customer\_id, oi.seller\_id,

o.order\_purchase\_timestamp, o.order\_approved\_at,

o.order\_delivered\_carrier\_date, o.order\_delivered\_customer\_date,

o.order\_estimated\_delivery\_date,

DATEDIFF(DAY, o.order\_purchase\_timestamp, o.order\_delivered\_customer\_date),

DATEDIFF(DAY, o.order\_purchase\_timestamp, o.order\_estimated\_delivery\_date),

CASE

WHEN o.order\_delivered\_customer\_date > o.order\_estimated\_delivery\_date THEN 1

ELSE 0

END

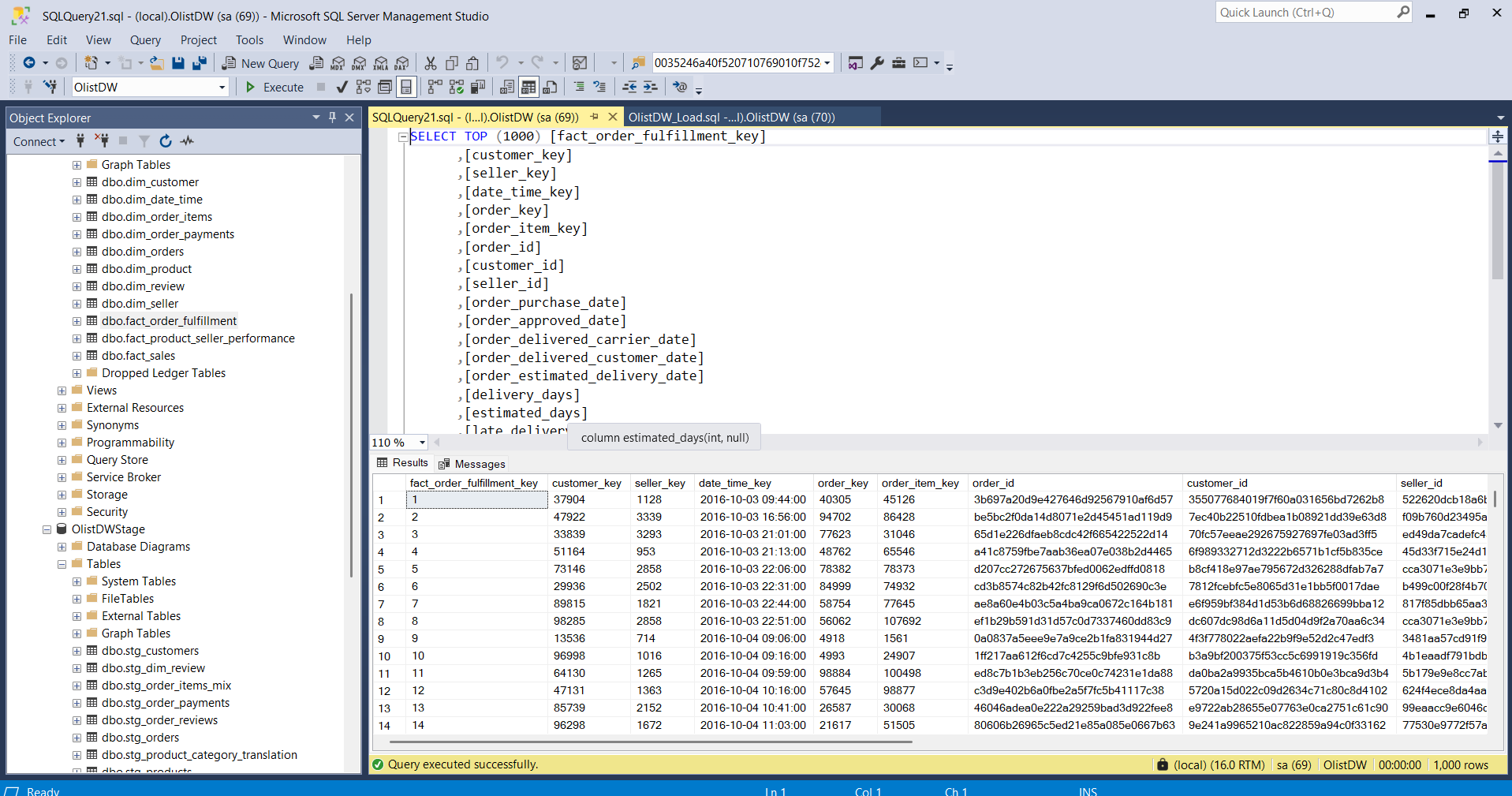
FROM OlistDW.dbo.dim\_orders o

JOIN OlistDW.dbo.dim\_order\_items oi ON o.order\_id = oi.order\_id

JOIN OlistDW.dbo.dim\_customer dc ON dc.customer\_id = o.customer\_id

JOIN OlistDW.dbo.dim\_seller ds ON ds.seller\_id = oi.seller\_id

JOIN OlistDW.dbo.dim\_date\_time d ON d.date\_time\_key = FORMAT(o.order\_purchase\_timestamp, 'yyyy-MM-dd HH:mm') + ':00';



**4.1.2.2. Fact Sales**

INSERT INTO dbo.fact\_sales (

product\_key, seller\_key, customer\_key,

order\_payment\_key, date\_time\_key,

order\_key, order\_item\_key,

order\_id, order\_item\_id, product\_id, seller\_id, customer\_id,

payment\_type, price, freight\_value, payment\_value, order\_purchase\_date

)

SELECT

product\_key, seller\_key, customer\_key,

p.order\_payment\_key, d.date\_time\_key,

o.order\_key, oi.order\_item\_key,

oi.order\_id, oi.order\_item\_id, oi.product\_id, oi.seller\_id, o.customer\_id,

p.payment\_type, oi.price, oi.freight\_value, p.payment\_value,

o.order\_purchase\_timestamp

FROM OlistDW.dbo.dim\_order\_items oi

JOIN OlistDW.dbo.dim\_orders o ON oi.order\_id = o.order\_id

LEFT JOIN (

SELECT

order\_payment\_key, order\_id,

MAX(payment\_type) AS payment\_type,

SUM(payment\_value) AS payment\_value

FROM OlistDW.dbo.dim\_order\_payments

GROUP BY order\_id, order\_payment\_key

) p ON oi.order\_id = p.order\_id

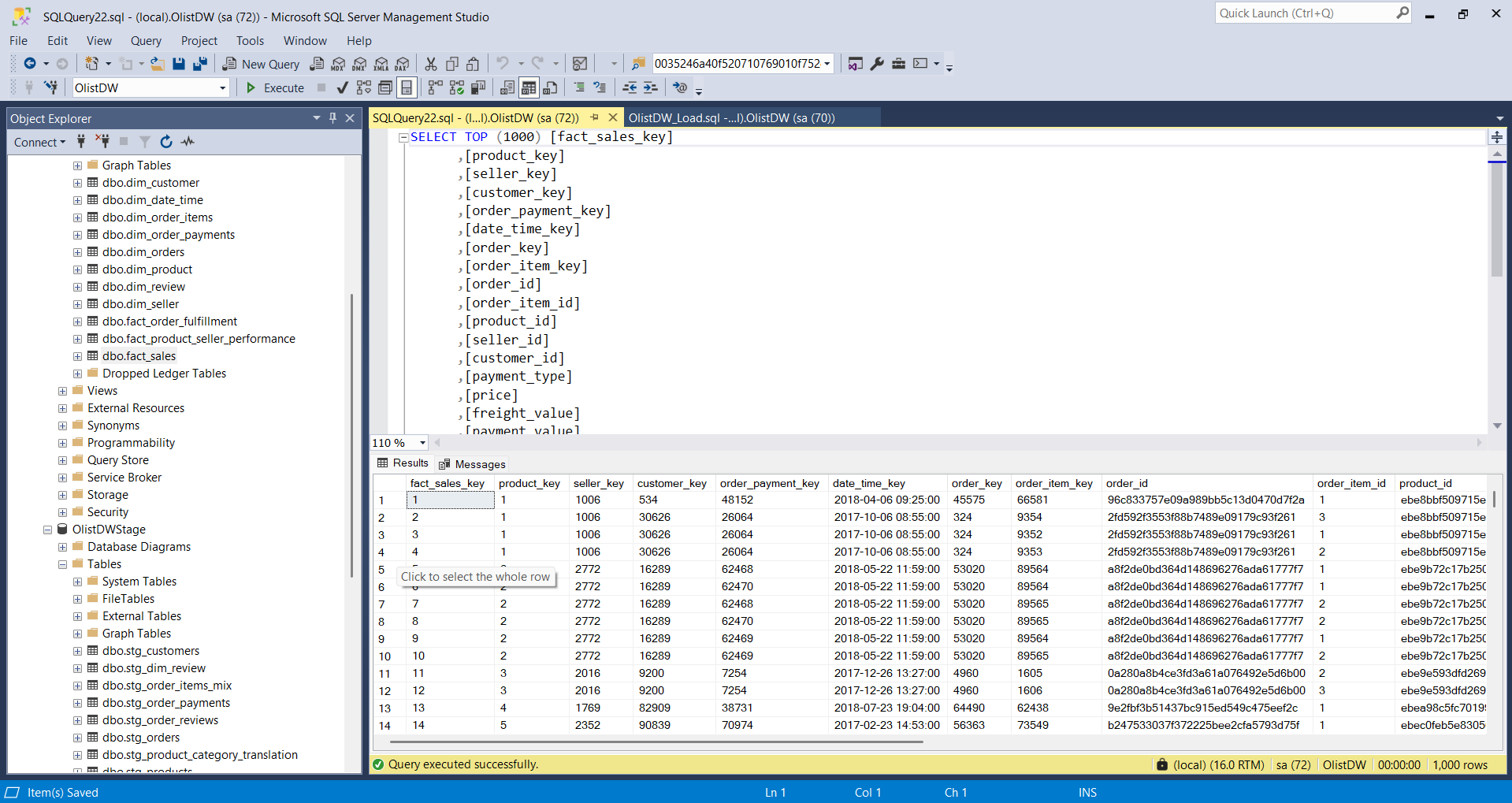
JOIN OlistDW.dbo.dim\_product dp ON dp.product\_id = oi.product\_id

JOIN OlistDW.dbo.dim\_seller ds ON ds.seller\_id = oi.seller\_id

JOIN OlistDW.dbo.dim\_customer dc ON dc.customer\_id = o.customer\_id

JOIN OlistDW.dbo.dim\_date\_time d

ON d.date\_time\_key = FORMAT(o.order\_purchase\_timestamp, 'yyyy-MM-dd HH:mm') + ':00';

****

**4.1.2.3. Fact Product Seller & Performance**

insert into OlistDW.dbo.fact\_product\_seller\_performance

(order\_item\_id, order\_id, price, freight\_value, review\_score, product\_key, seller\_key, review\_key, datetime\_key, review\_create\_datetime)

select

o.order\_item\_id,

o.order\_id,

o.price,

o.freight\_value,

r.review\_score,

p.product\_key,

s.seller\_key,

r.review\_key,

FORMAT(o.order\_purchase\_timestamp, 'yyyy-MM-dd HH:mm') + ':00' AS datetime\_key,

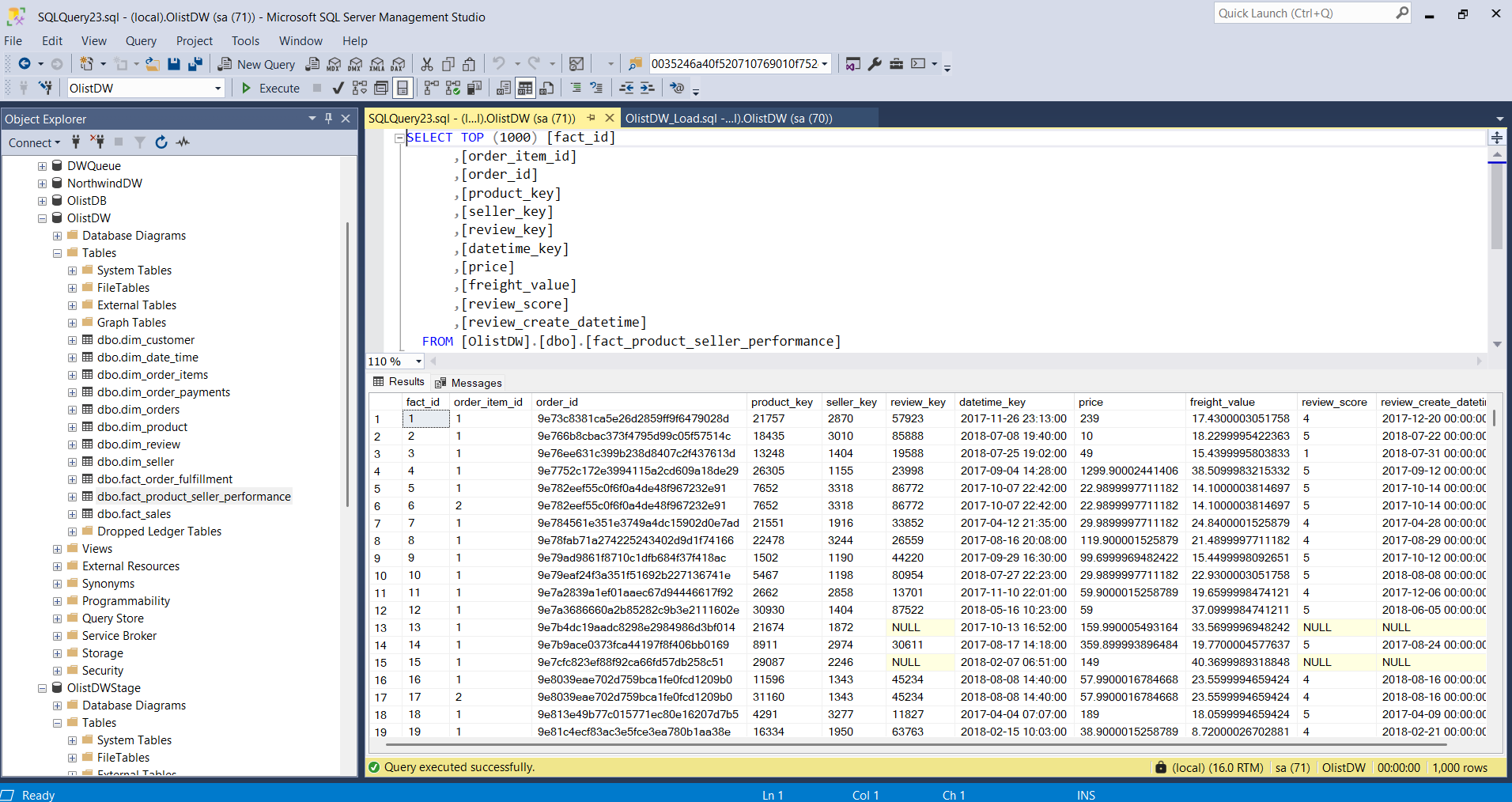
FORMAT(r.review\_creation\_date, 'yyyy-MM-dd HH:mm') + ':00' AS review\_create\_datetime

from OlistDWStage.dbo.stg\_order\_items\_mix o

join OlistDW.dbo.dim\_product p on o.product\_id = p.product\_id

join OlistDW.dbo.dim\_seller s on o.seller\_id = s.seller\_id

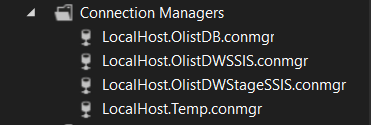
join OlistDW.dbo.dim\_review r on o.order\_id = r.order\_id;



## **4.2. Load bằng công cụ SSIS**

### **4.2.1. Kết Nối**

* Tiến hành thiết lập các kết nối:



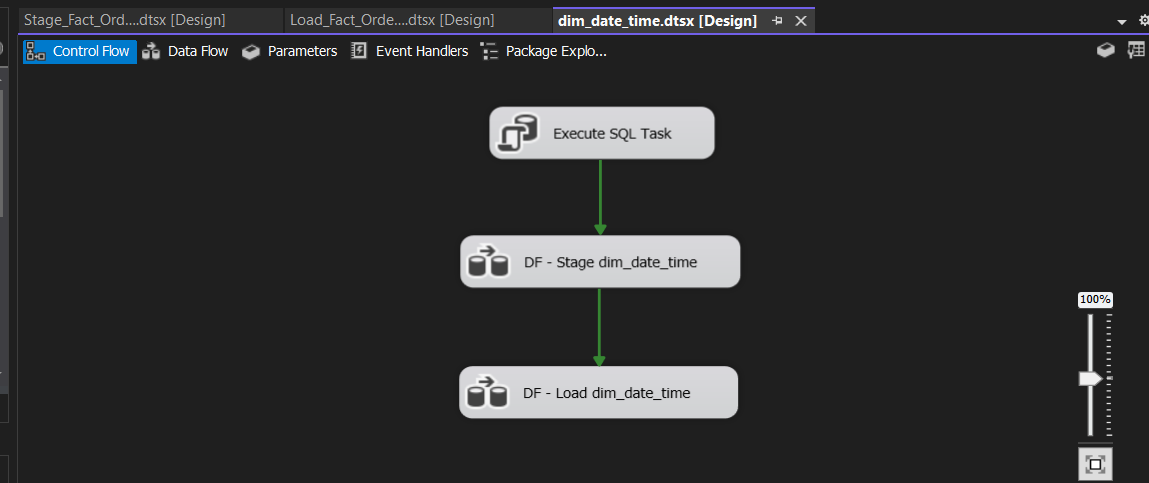
Với OlistDB chứa dữ liệu gốc, OlistDWStageSSIS dùng làm database cho staging, Temp chứa dữ liệu cho dim\_date\_time, và OlistDWSSIS là data warehouse sau khi load cuối cùng.

### **4.2.2. Dim Date Time**

Tạo package:



Control Flow:

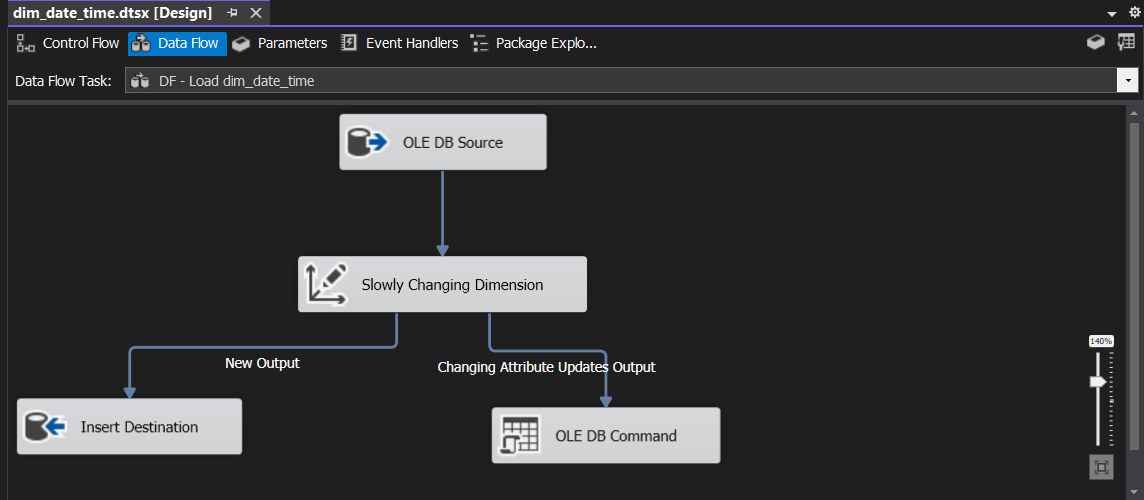


Các Data Flow:

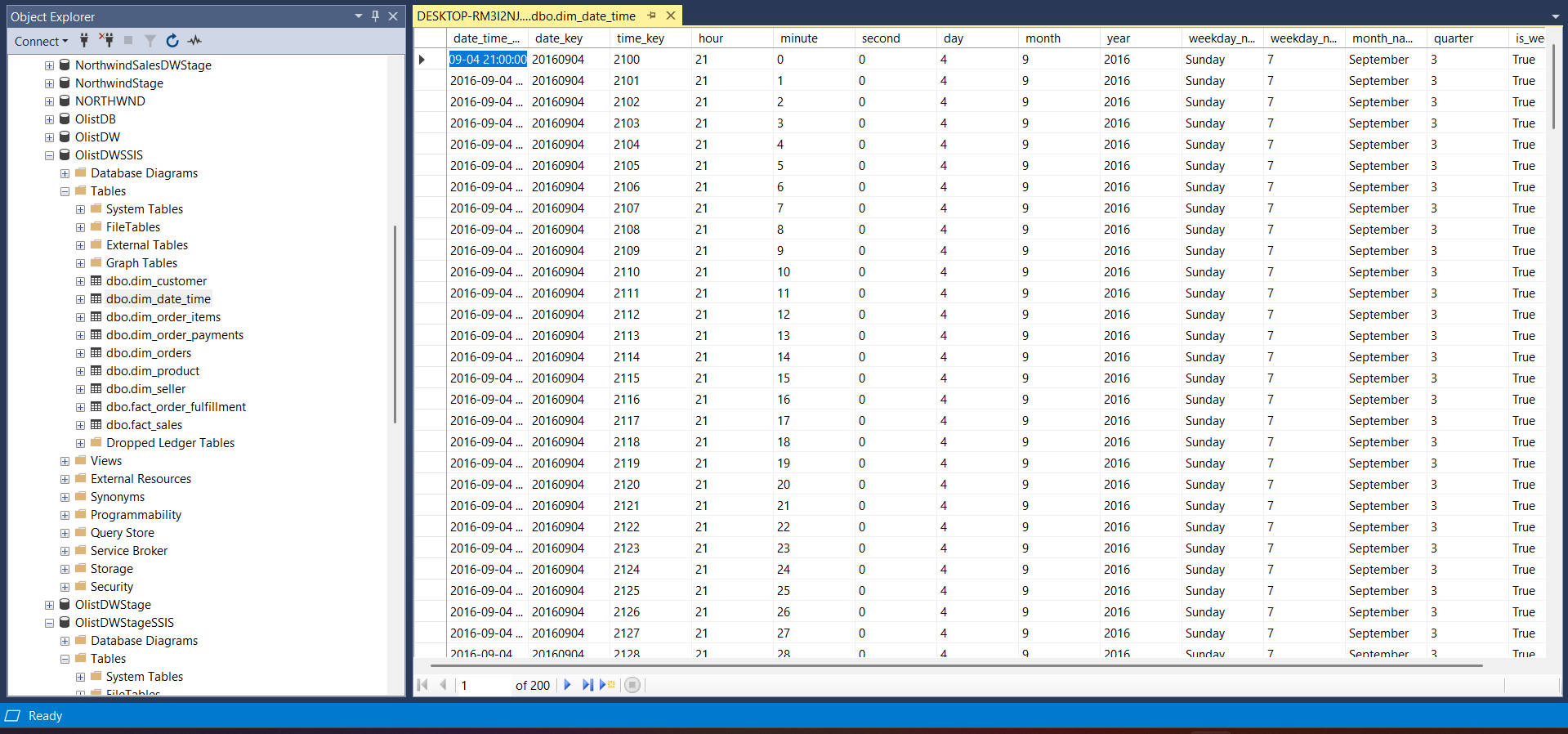
* DF - Stage dim\_date\_time



* DF - Load dim\_date\_time



Kết quả cuối cùng:



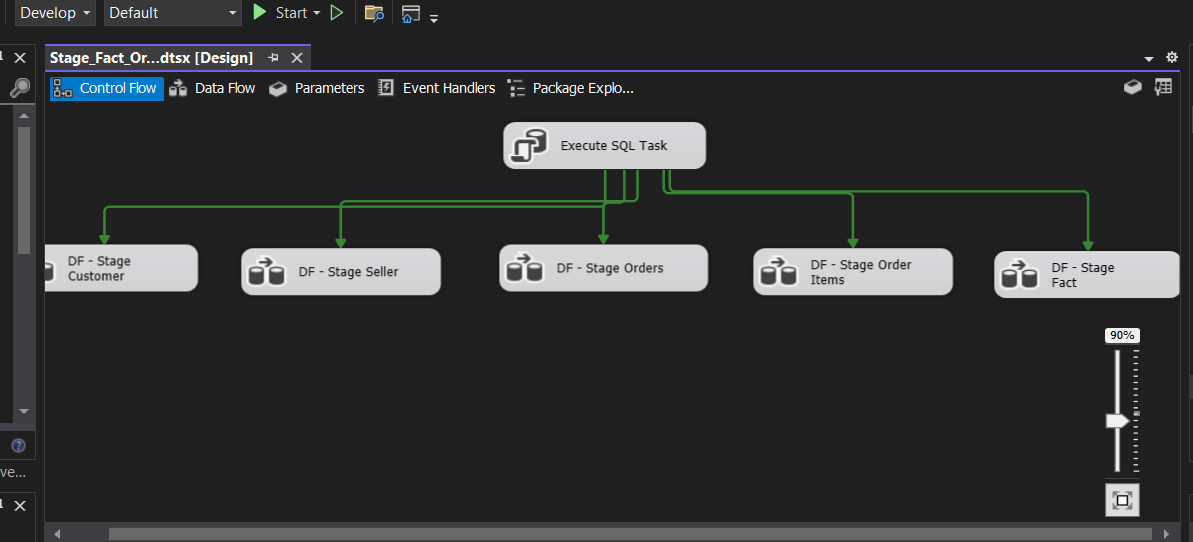
### **4.2.3. Business Process: Order Fulfillment & Delivery Performance**

**4.2.3.1. Tạo các Data Flow Stage cần thiết cho Order Fulfillment & Delivery Performance**

* Sử dụng lại dim\_date\_time đã staging và load trước đó.
* Tạo package:



* Tạo các Control Flow Staging cần thiết:



* Kết quả thực thi:

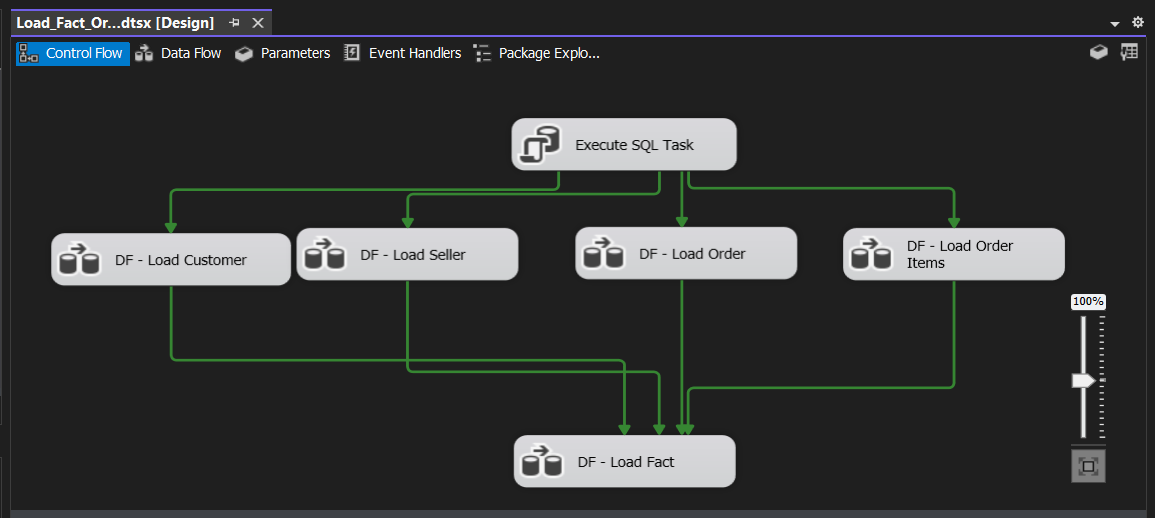


**4.2.3.2. Tạo các Data Flow load cần thiết cho Order Fulfillment & Delivery Performance**

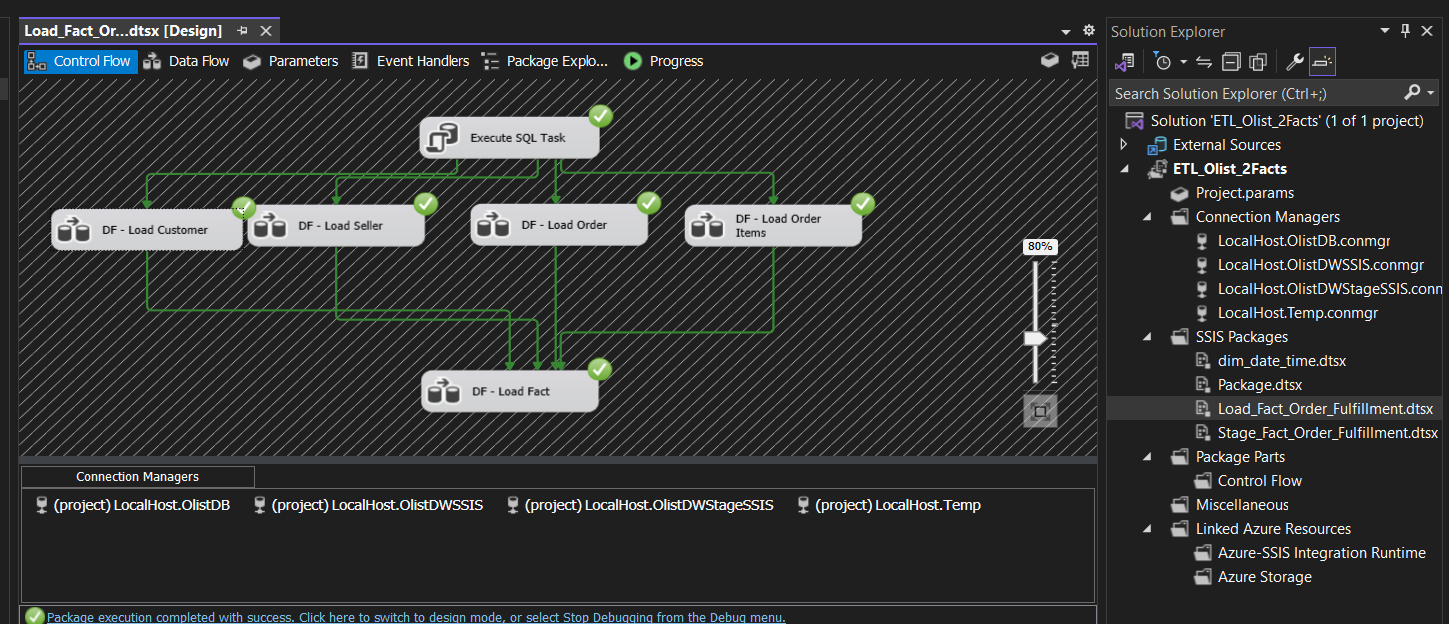
* Tạo package:



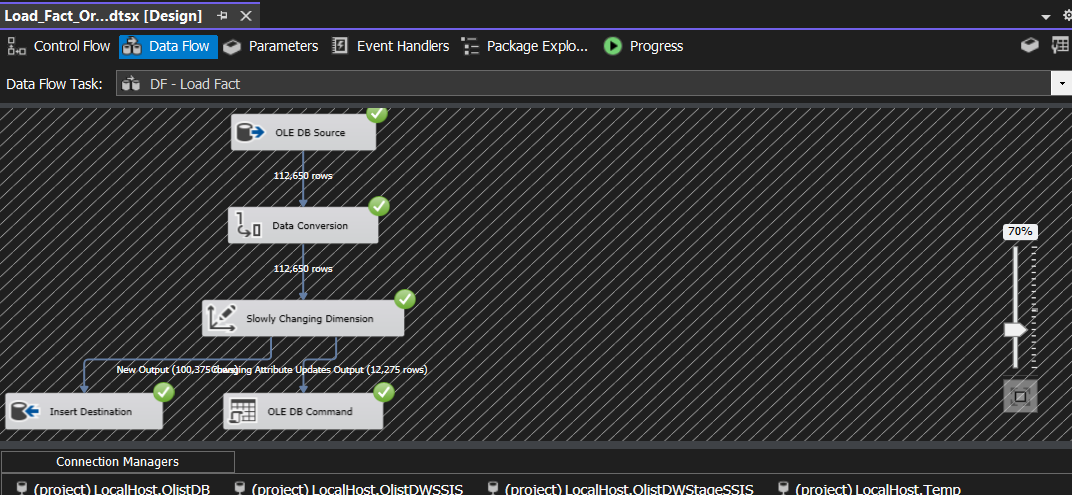
* Control Flow:

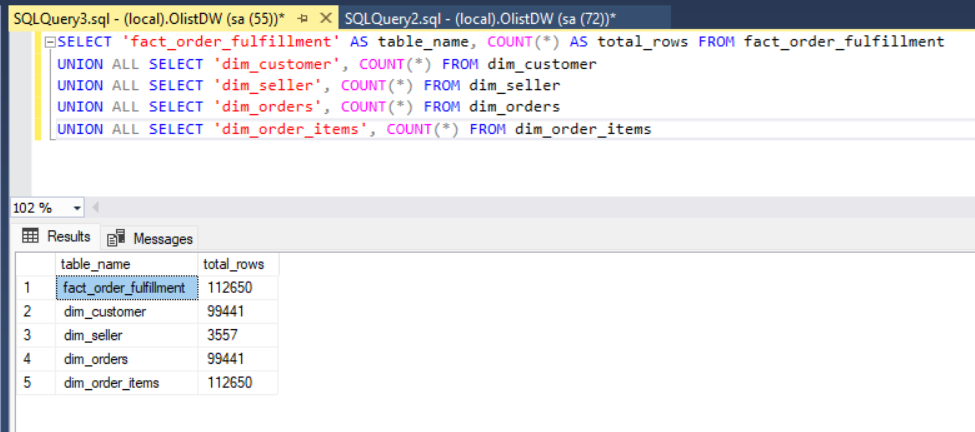


* Kết quả thực thi:



* Kết quả load cho Fact Order Fulfillment & Delivery Performance





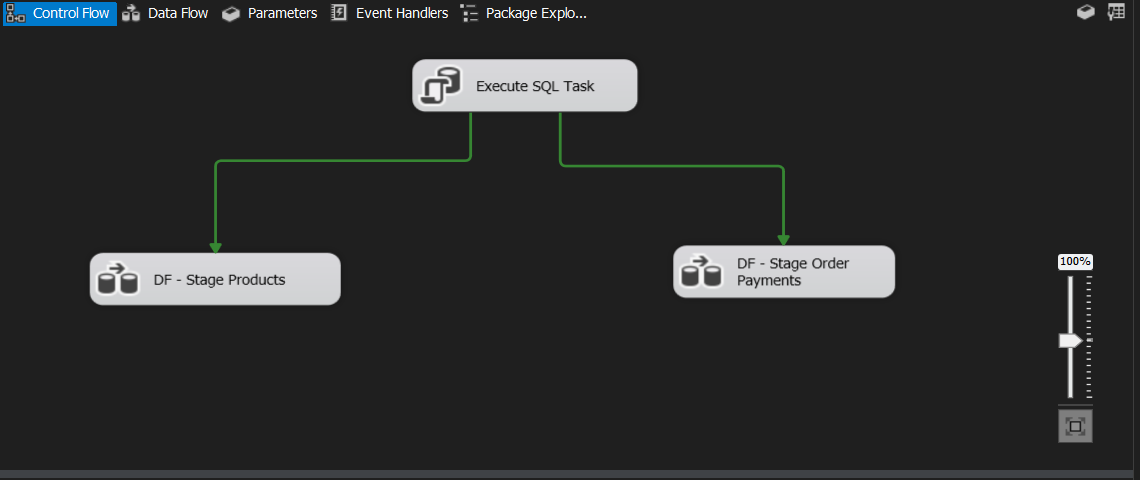
### **4.2.4. Business Process : Sales & Revenue Analysis**

**4.2.4.1. Tạo các Data Flow Stage cần thiết cho Sales & Revenue Analysis**

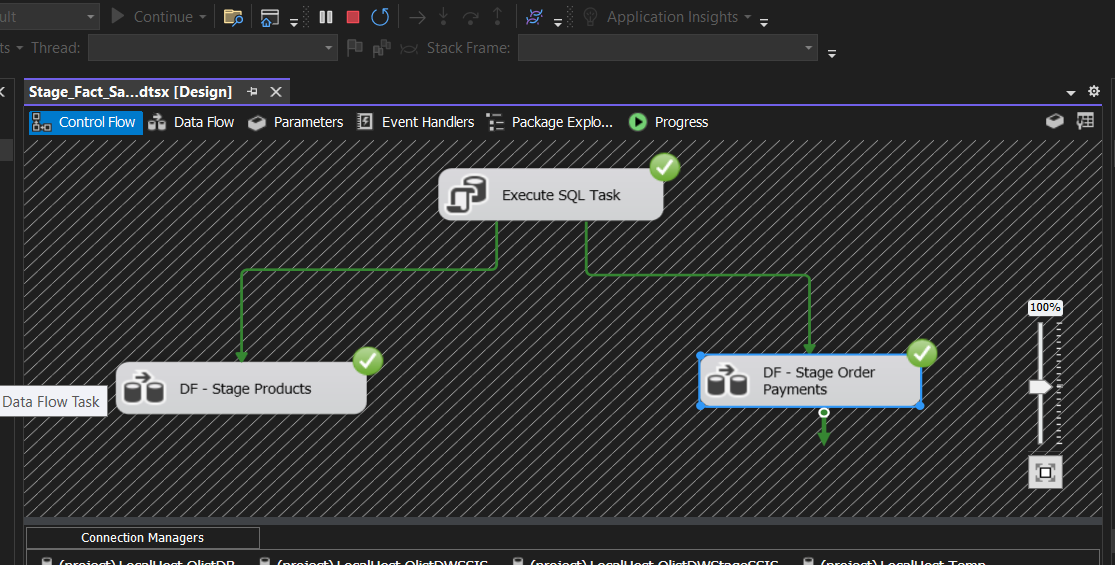
* Sử dụng lại các dim như customer, seller, orders, order items, date
* Tạo package:



* Control Flow:



* Kết quả thực thi:

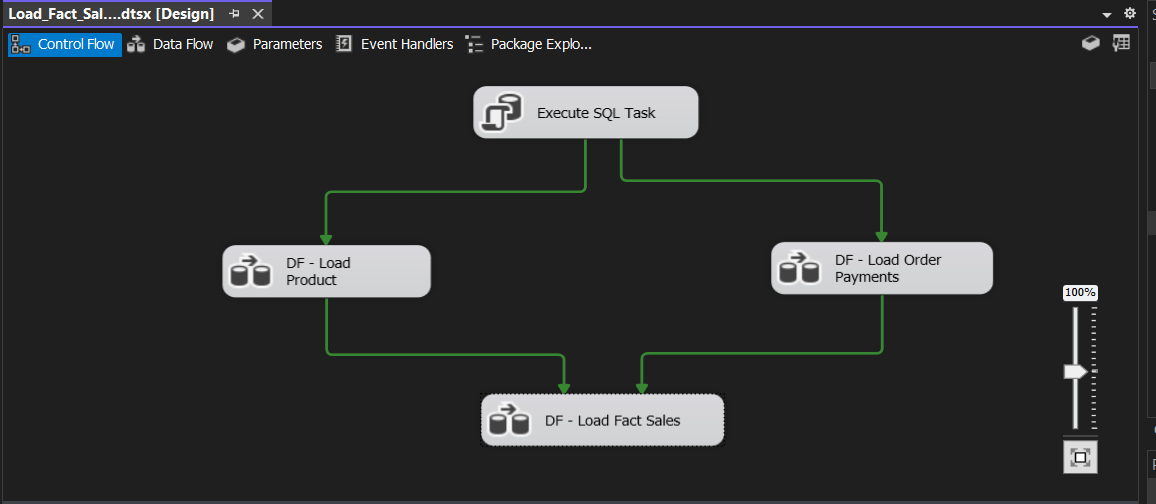


**4.2.4.2. Tạo các Data Flow load cần thiết cho Sales & Revenue Analysis**

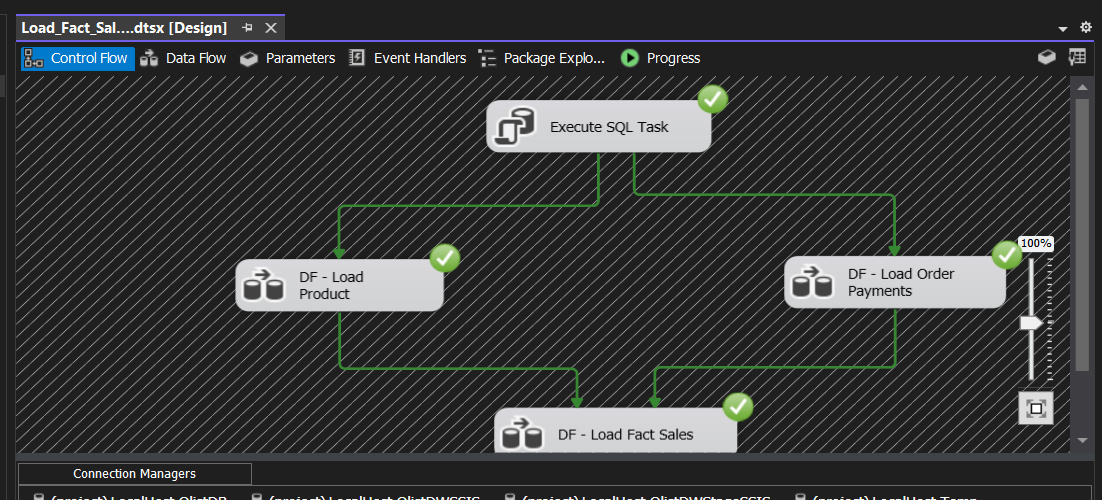
* Tạo package:



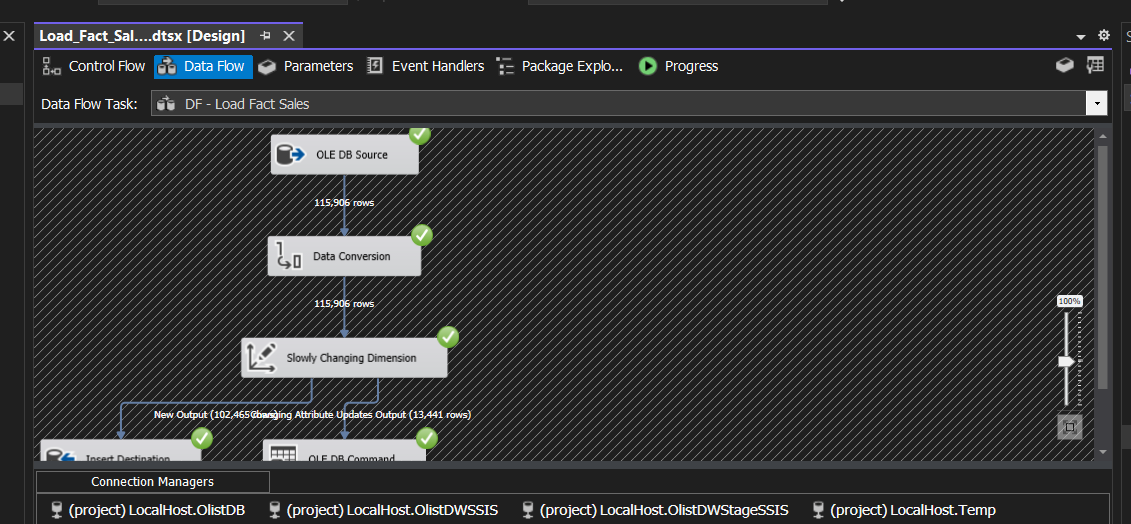
* Control Flow:

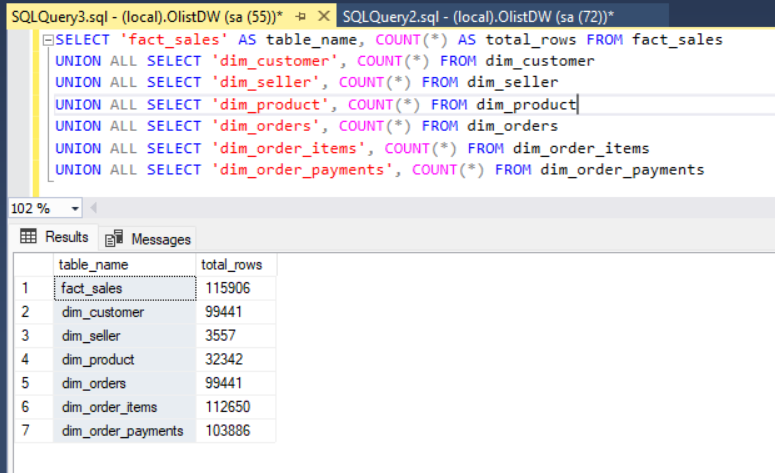


* Kết quả thực thi:



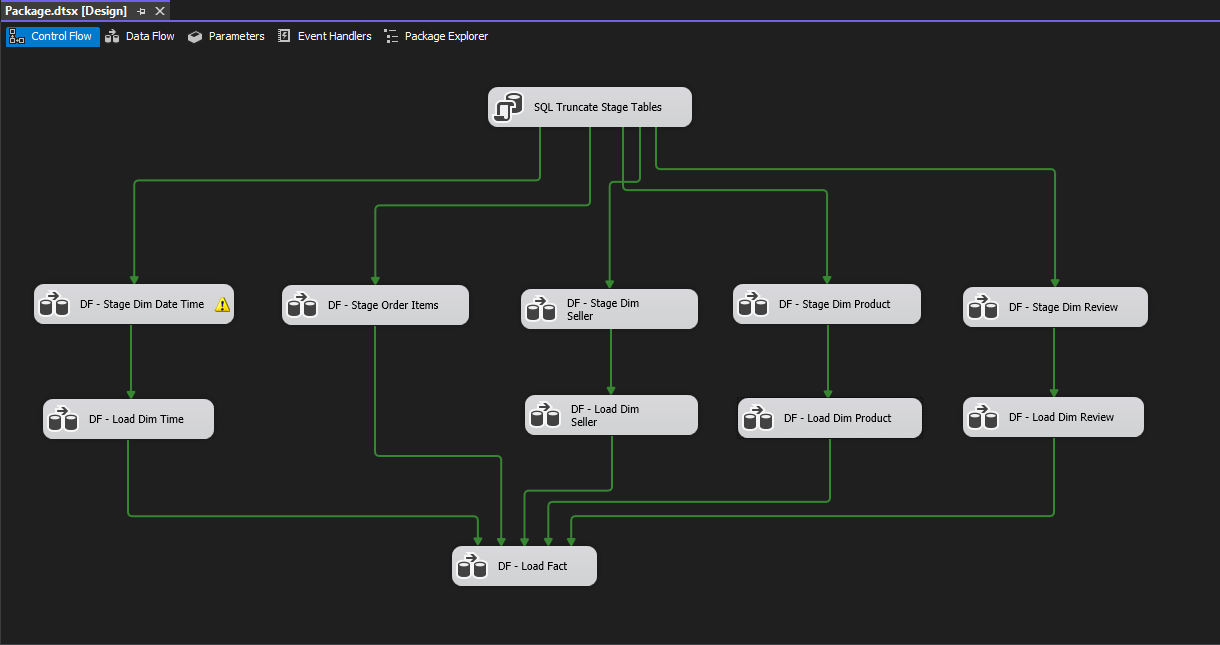
* Kết quả load cho Fact Sales & Revenue Analysis



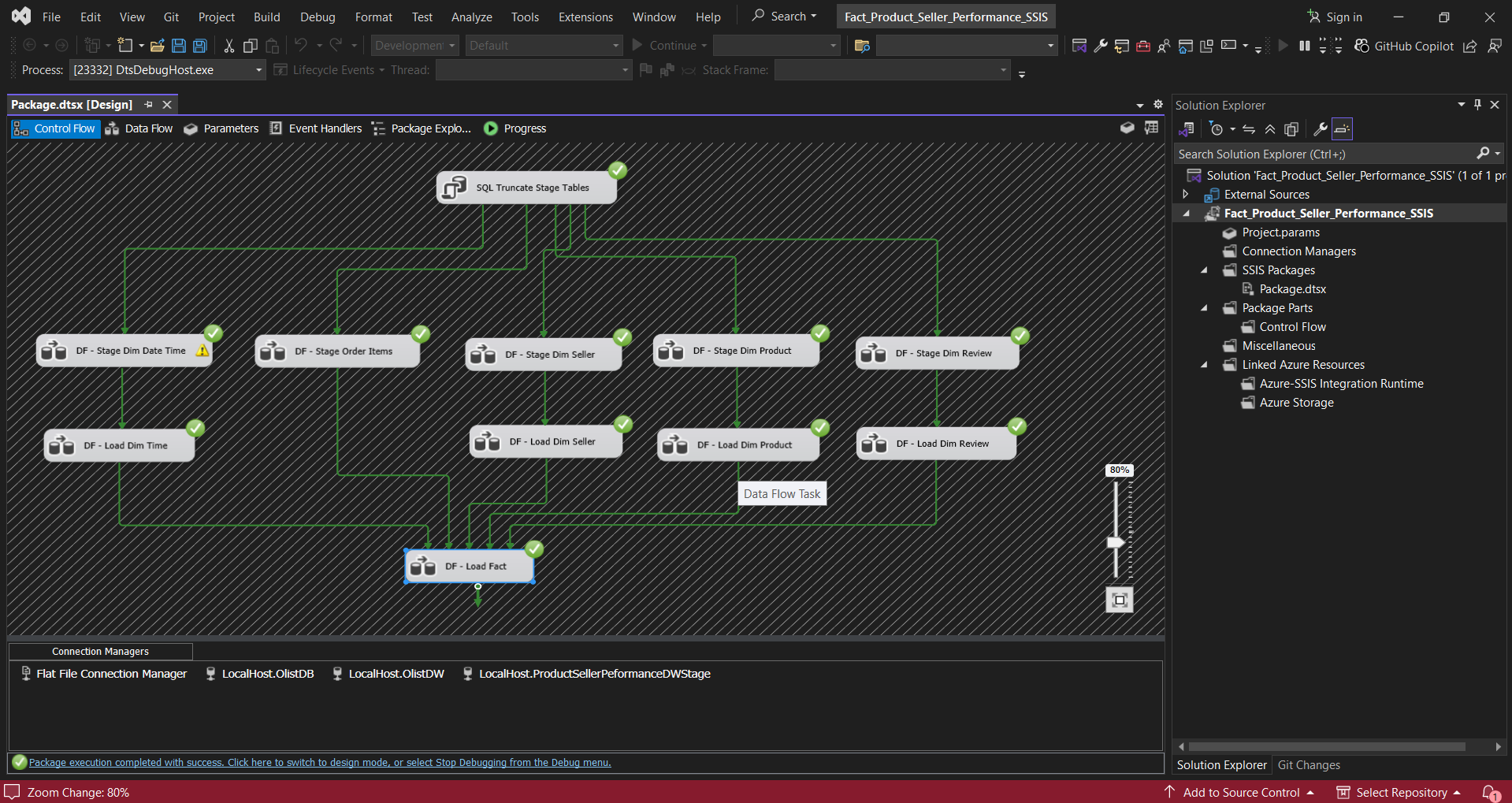


### **4.2.5. Business Process: Product Seller Peformance**

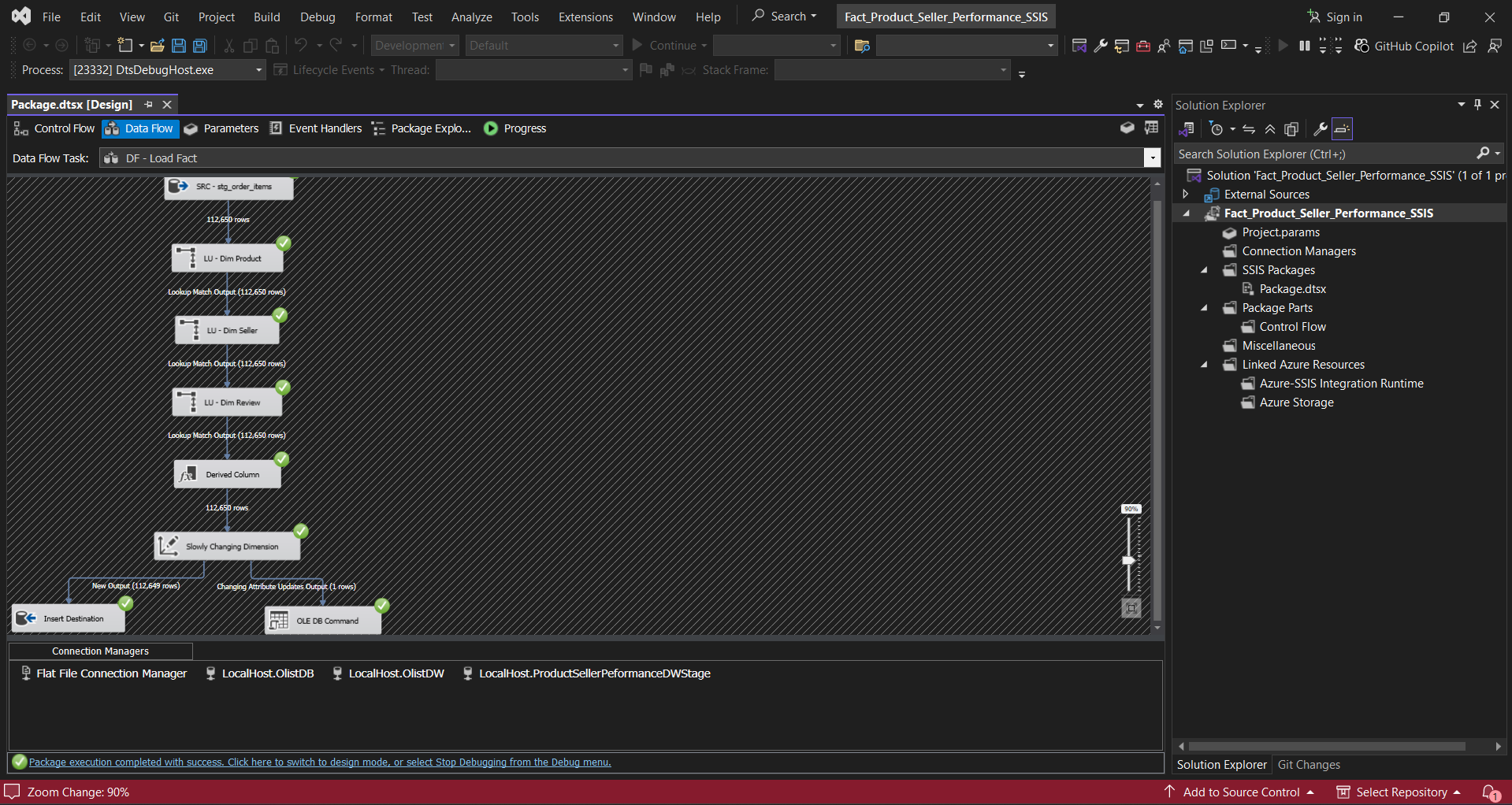
* Tạo Data Flow Stage và Load cần thiết cho Product Seller Peformance

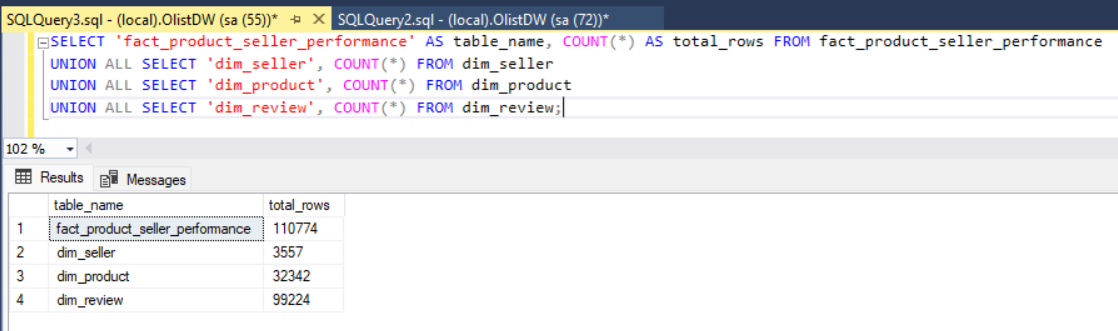


* Kết quả thực thi:



* Kết quả load cho Fact Product Seller Peformance

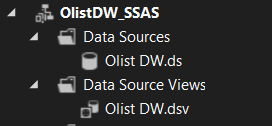




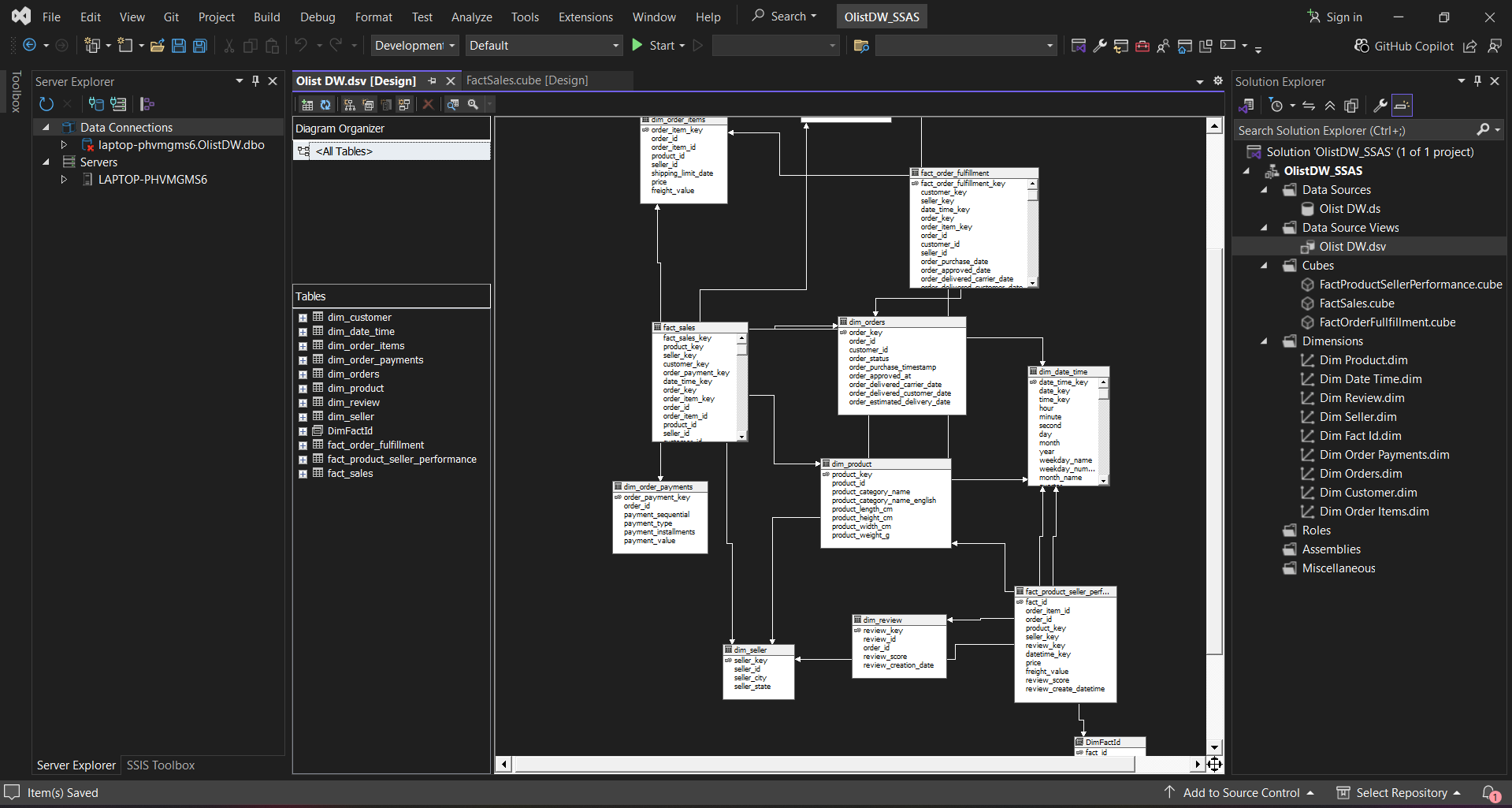
# **5. Phân tích Dữ liệu bằng SSAS**

## **5.1. Tạo Data Sources và Data Sources View kết nối với OlistDW**

* Tạo kết nối



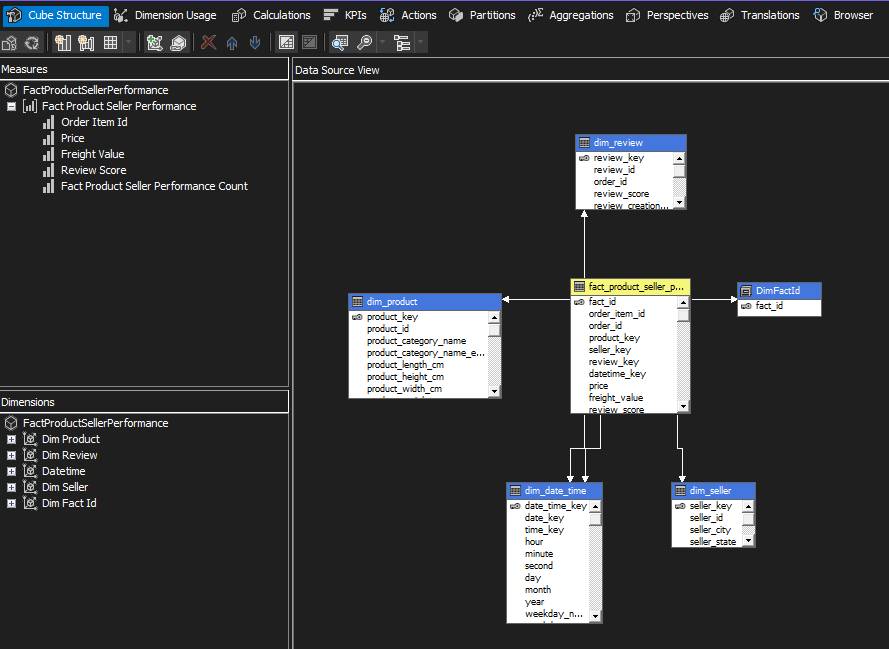
* Kết nối với OlistDW



## **5.2. Tạo Cubes**

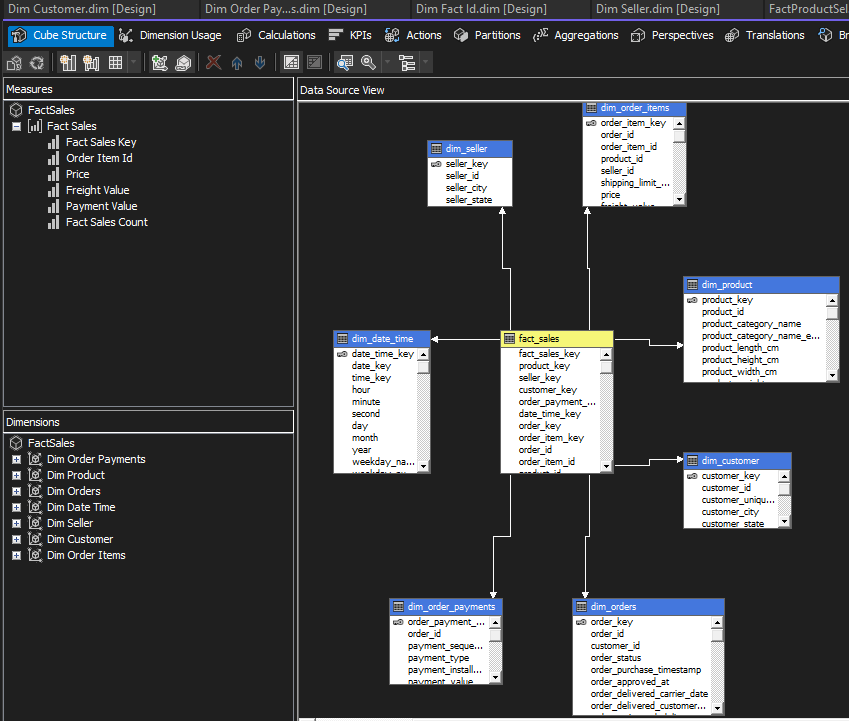
### **5.2.1. Tạo Cube FactProductSellerPerformance**

* Chọn các Dimensions và Fact làm Measures đúng cho phân tích



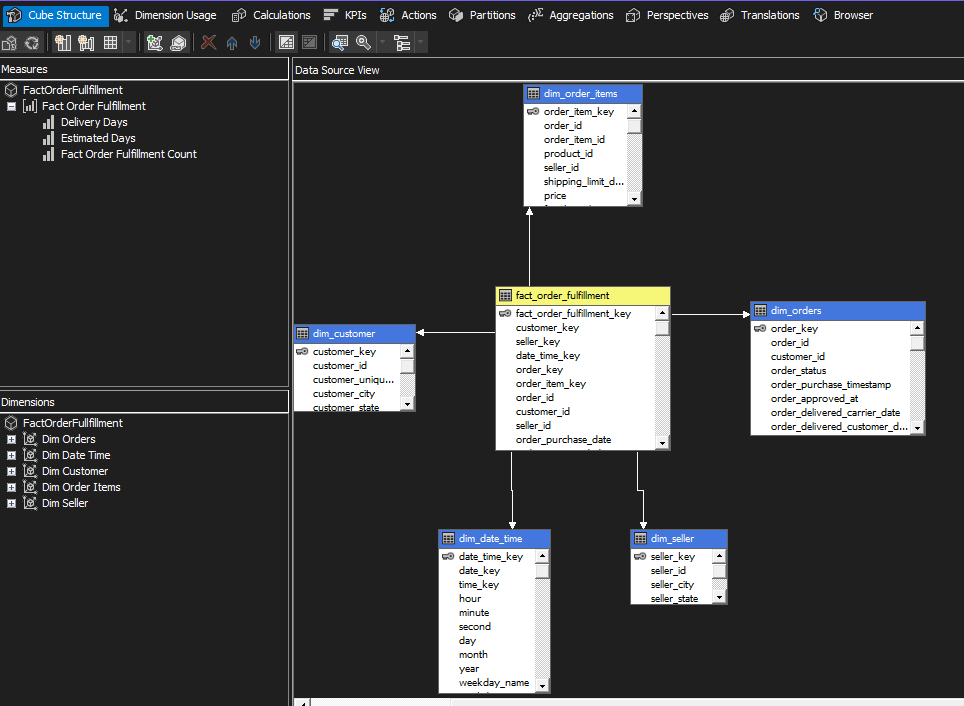
### **5.2.2. Tạo Cube FactSales**

* Chọn các Dimensions và Fact làm Measures đúng cho phân tích



### **5.2.3. Tạo Cube FactOrderFullfillment**

* Chọn các Dimensions và Fact làm Measures đúng cho phân tích

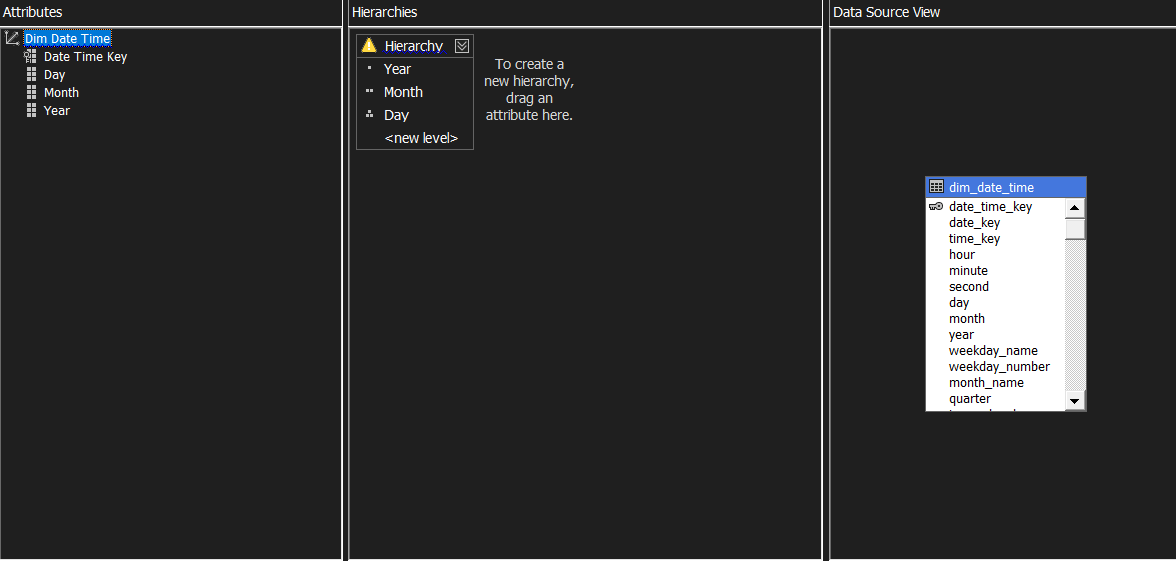


## **5.3. Thiết lập Measures và Dimensions**

* Thêm thuộc tính hiển thị cho Dim Product



* Thêm thuộc tính và phân cấp cho Dim Date Time

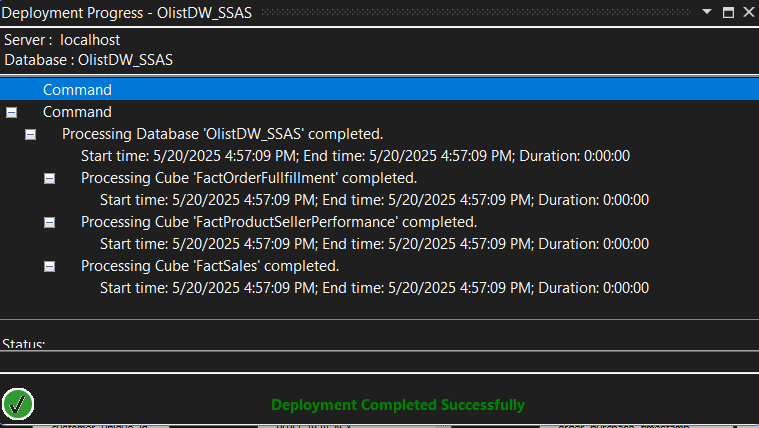


* Thêm thuộc tính hiển thị cho Dim Customer

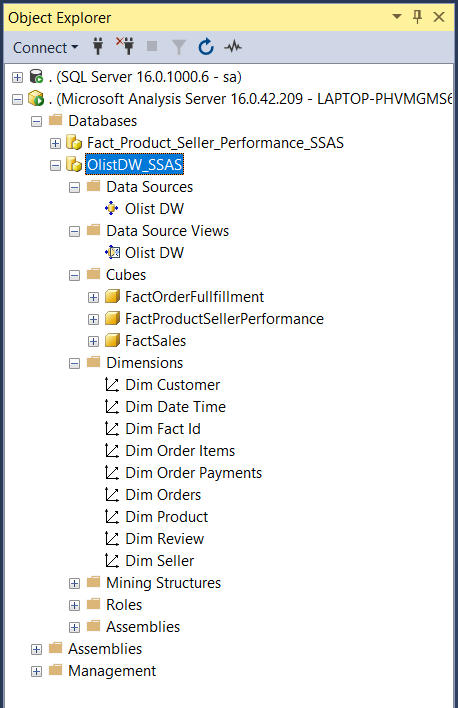


## **5.4. Triển khai và Xử lý Cube**

* Thực hiện Deloy Cube lên Analysis Services.



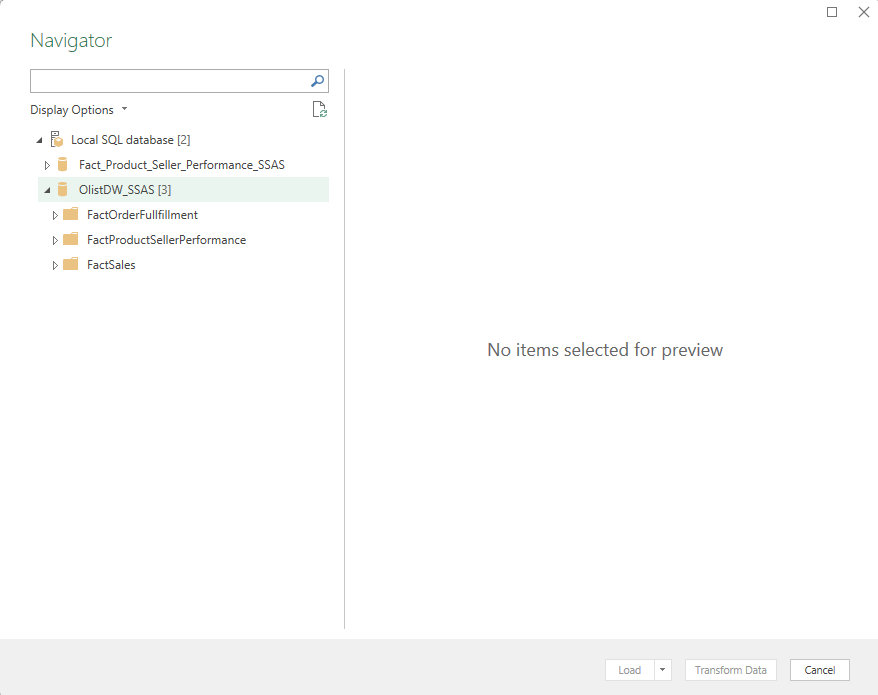
* Kiểm tra kết quả



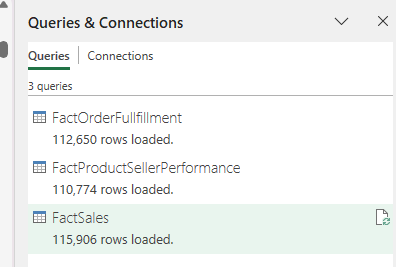
# **6. Trả lời câu hỏi bằng Pivot Table**

## **6.1. Kết nối Excel với Cube SSAS**

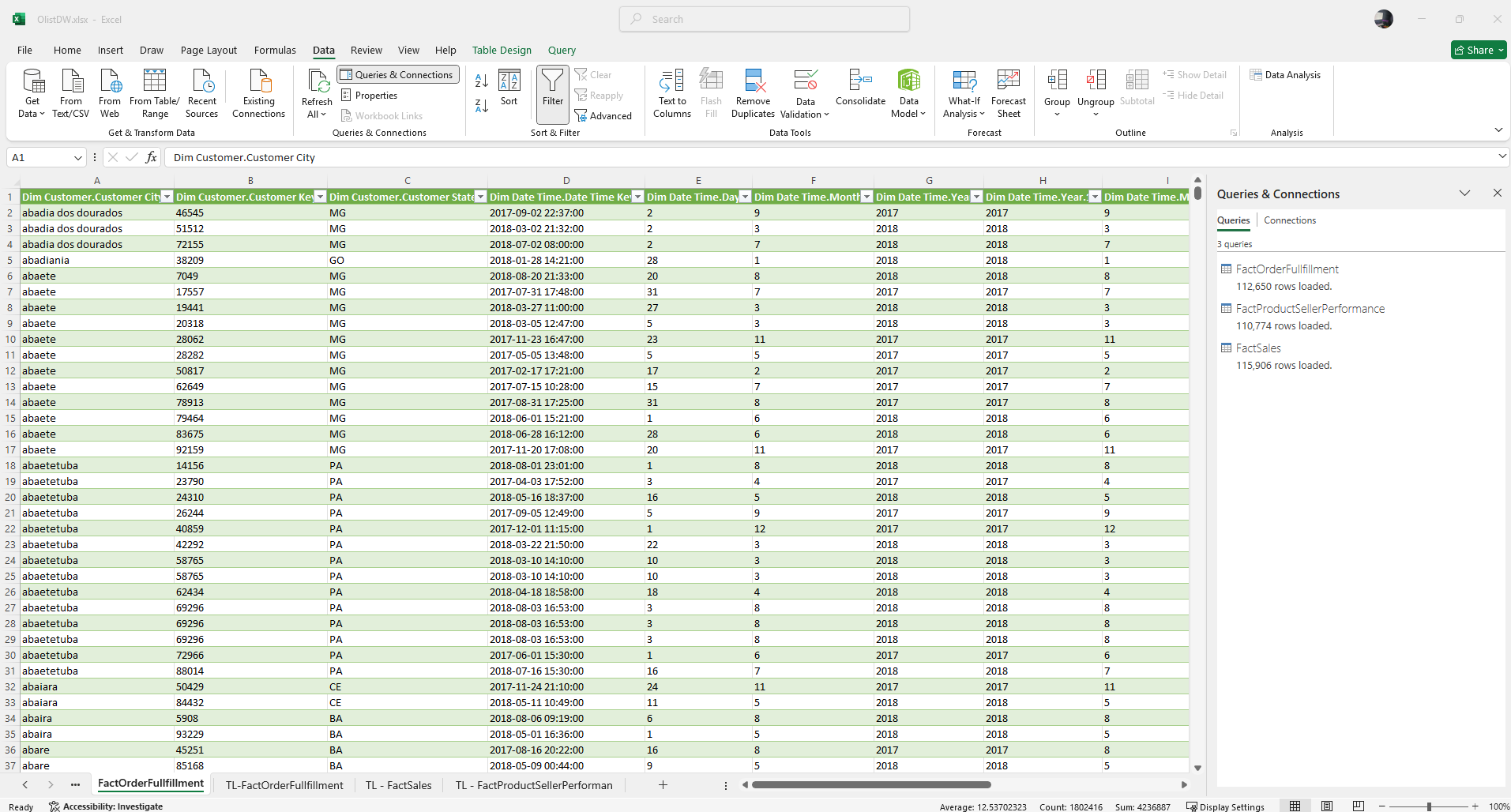
* Kết nối với Analysis Services sẽ có OlistDW\_SSAS



* Chọn các Fact để load



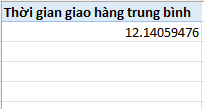
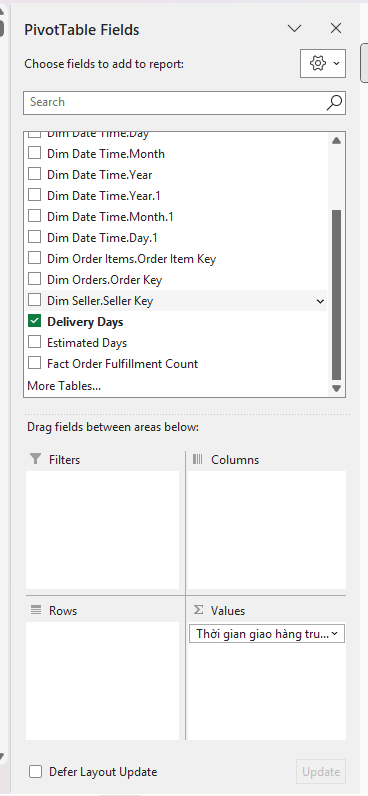
* Kết quả load thành công



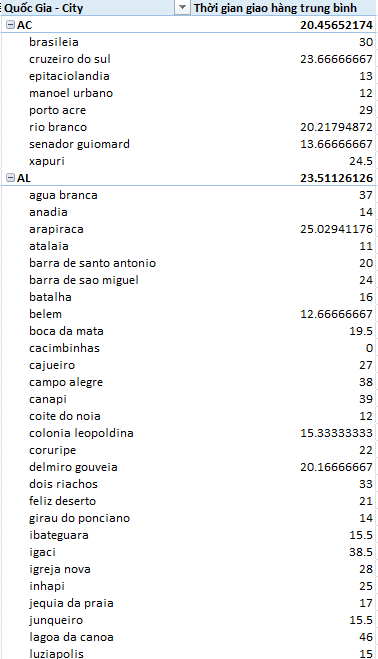
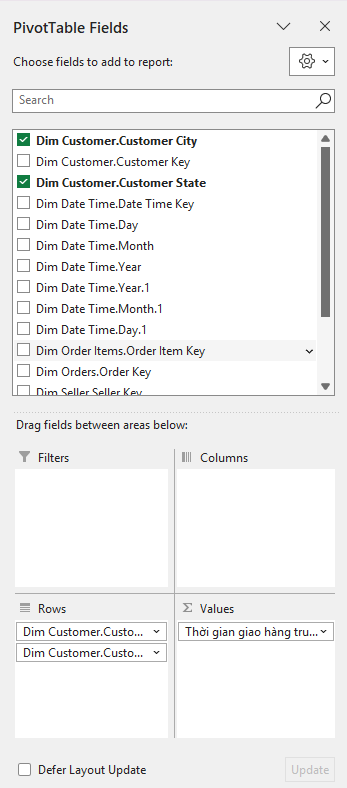
## **6.2. Trả lời các câu hỏi phân tích**

### **6.2.1. Business Process : Order Fulfillment & Delivery Performance**

**Câu hỏi 1: Thời gian giao hàng trung bình là bao lâu?**

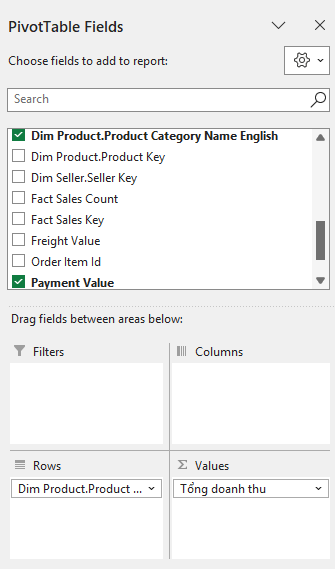
****

**Câu hỏi 2: Có sự khác biệt về thời gian giao hàng giữa các bang/khu vực không?**

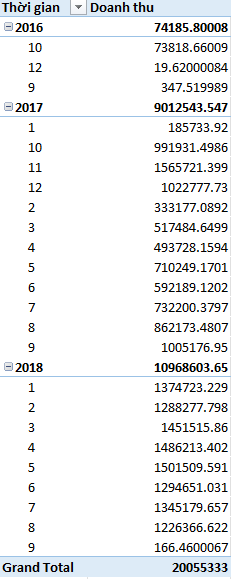
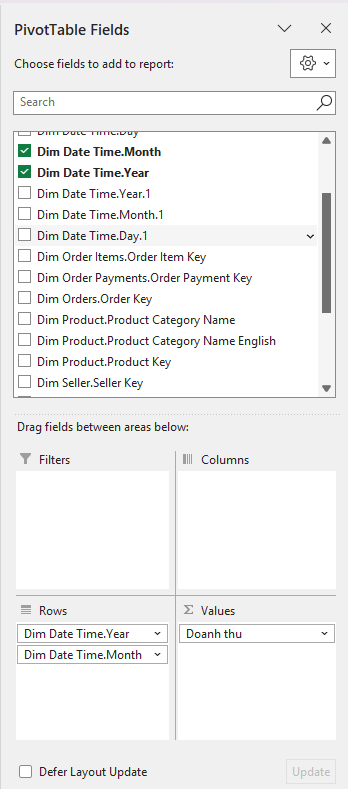
****

### **6.2.2. Business Process : Sales & Revenue Analysis**

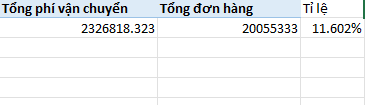
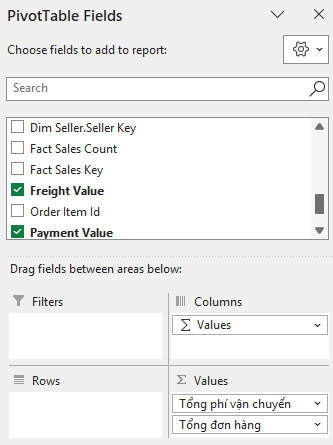
**Câu hỏi 3: Doanh thu theo loại sản phẩm?**

****

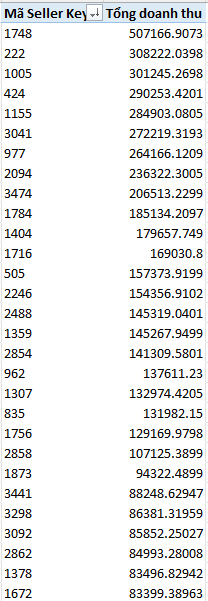
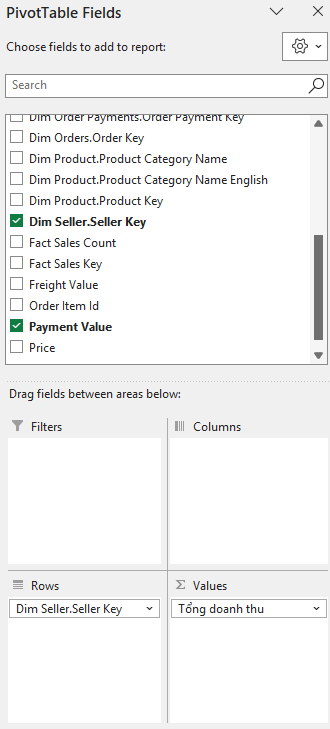
**Câu hỏi 4: Doanh thu theo thời gian?**

****

**Câu hỏi 5: Chi phí vận chuyển chiếm bao nhiêu % giá trị đơn hàng?**

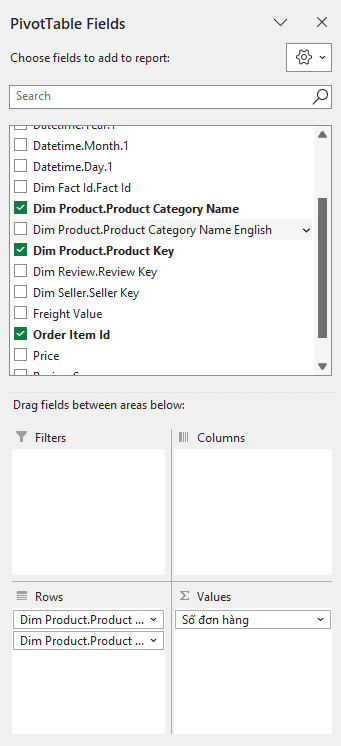
****

**Câu hỏi 6: Seller nào có doanh thu cao nhất?**

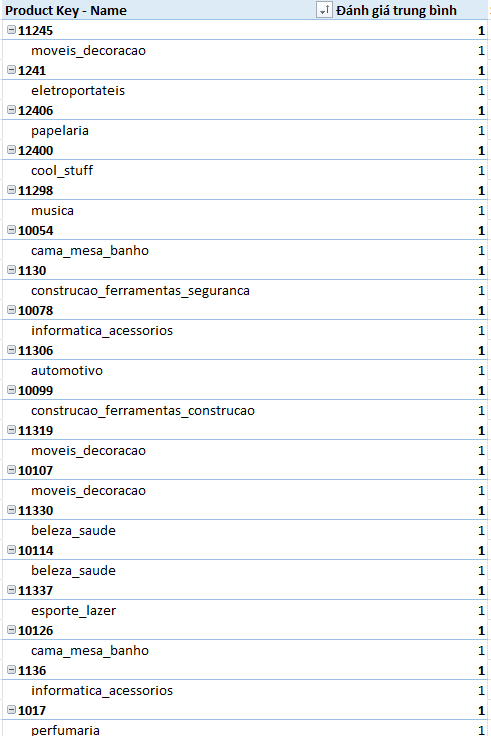
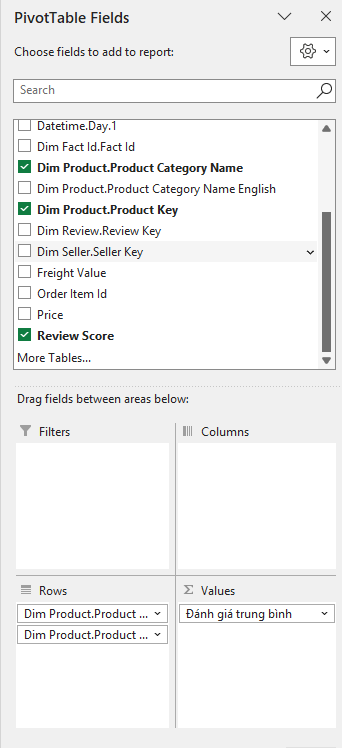
****

### **6.2.3. Business Process: Product Seller Peformance**

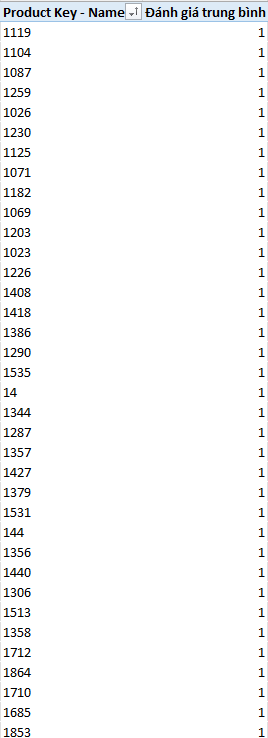
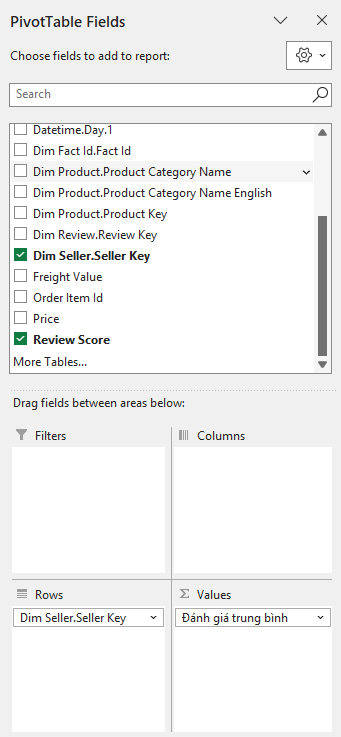
**Câu hỏi 7: Sản phẩm nào bán chạy nhất?**

****

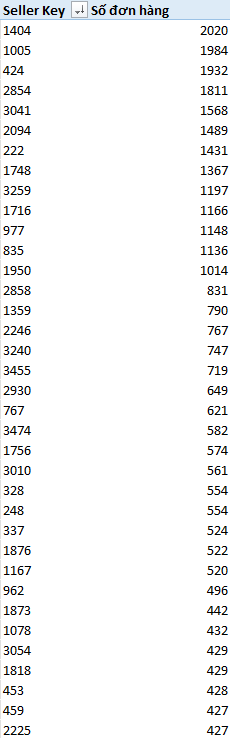
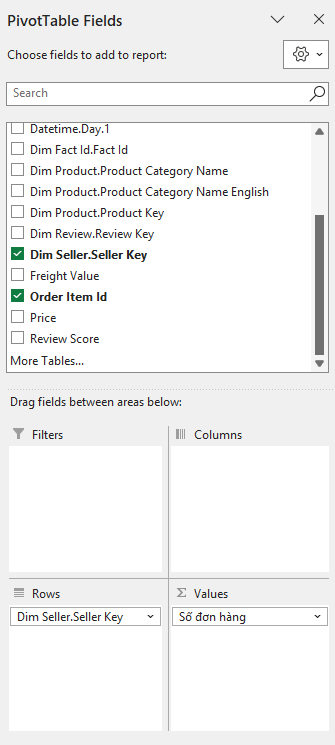
**Câu hỏi 8: Sản phẩm nào bị đánh giá thấp?**

****

**Câu hỏi 9: Seller nào có đánh giá thấp nhất?**

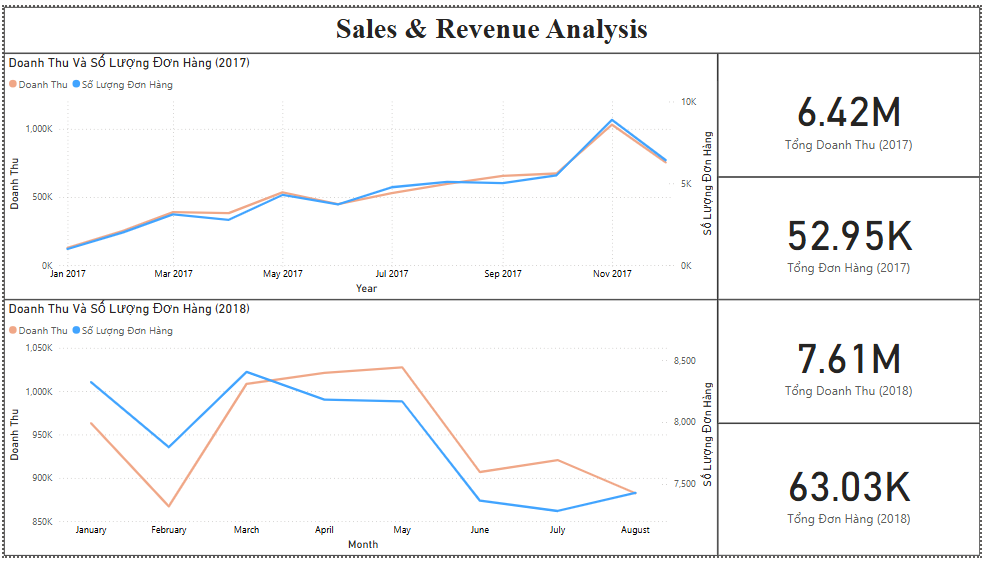
****

**Câu hỏi 10: Seller nào có nhiều đơn hàng nhất?**

****

# **7. Trực quan hóa bằng Power BI**

## **7.1. Thực hiện phân tích dữ liệu cho Business process: Sales & Revenue Analysis**



### **7.1.1. Câu hỏi: Cho biết số lượng đơn đặt hàng, doanh thu qua các năm 2017, 2018**

Từ DashBoard, Ta có thể xem được lợi nhuận và số đơn đặt hàng năm 2017 lần lượt là: 6,42 triệu đô doanh thu và 52,950 đơn đặt hàng và năm 2018 là: 7,61 triệu đô doanh thu và 63,030 đơn đặt hàng

A white rectangular object with a black line

AI-generated content may be incorrect.

A white rectangular object with a black line

AI-generated content may be incorrect.

### **7.1.2. Câu hỏi: Cho biết sự tăng giảm doanh thu qua các tháng (2017, 2018) dựa theo thị trường/khu vực**

A graph with a line and a line

AI-generated content may be incorrect.

Biểu đồ 2017 cho thấy xu hướng tăng trưởng ổn định của cả doanh thu và số lượng đơn hàng trong năm, đặc biệt là tăng mạnh vào tháng 11, cho thấy cao điểm bán hàng cuối năm.

A graph of a graph showing the number of months and months

AI-generated content may be incorrect.

Biểu đồ 2018 lại thể hiện sự biến động rõ rệt giữa các tháng. Mặc dù doanh thu và số đơn hàng có xu hướng tăng cùng nhau trong quý I, nhưng từ tháng 6 trở đi đều sụt giảm, phản ánh sự chững lại trong hoạt động kinh doanh.

## **7.2. Thực hiện phân tích dữ liệu cho Business process: Order Fulfillment**

A screenshot of a graph

AI-generated content may be incorrect.

### **7.2.1. Câu hỏi: Cho biết thời gian giao hàng trung bình qua các tháng của năm 2017 và 2018**

A graph of purple bars

AI-generated content may be incorrect.

Biểu đồ 2017: Thời gian giao hàng trung bình duy trì ở mức tương đối ổn định (khoảng 15–17 ngày), với xu hướng giảm nhẹ từ tháng 4 đến tháng 8, cho thấy sự cải thiện dần trong nửa cuối năm.

A graph of purple bars

AI-generated content may be incorrect.

Biểu đồ 2018: Đầu năm có thời gian giao hàng cao hơn, đặc biệt tháng 2 vượt 20 ngày – thể hiện hiệu suất thấp. Tuy nhiên, từ tháng 4 trở đi, thời gian giao giảm rõ rệt, cho thấy hiệu quả giao hàng được cải thiện rõ rệt qua các tháng.

### **7.2.2. Câu hỏi: Cho biết tỉ lệ đơn giao đúng hạn và trễ hạn qua các tháng của năm 2017 và 2018**

A blue rectangular bars with white text

AI-generated content may be incorrect.

Biểu đồ 2017: Tỷ lệ đơn giao đúng hạn duy trì rất cao qua các tháng, gần như tuyệt đối. Điều này cho thấy hiệu quả vận hành và cam kết giao hàng được đảm bảo ổn định trong suốt năm.

A screenshot of a graph

AI-generated content may be incorrect.

Biểu đồ 2018: Đầu năm (đặc biệt là tháng 2 và 3) ghi nhận tỷ lệ giao trễ cao hơn đáng kể so với các tháng còn lại. Tuy nhiên, từ tháng 4 trở đi, tỷ lệ giao đúng hạn được cải thiện rõ rệt và duy trì ở mức gần 100%, tương đương với năm 2017

## **7.3. Thực hiện phân tích dữ liệu cho Business process: Product & Seller Performance**

A close-up of a graph

AI-generated content may be incorrect.

### **7.3.1. Câu hỏi: Cho biết 5 danh mục sản phẩm bán chạy và đem theo doanh thu nhiều nhất trong năm 2018**

#### A graph of purple rectangular objects AI-generated content may be incorrect.

A graph of a bar

AI-generated content may be incorrect.

Hai biểu đồ trên cho thấy Top 5 danh mục sản phẩm bán chạy nhất và mang lại doanh thu cao nhất trong năm 2018. Có thể nhận thấy:

Danh mục "bed\_bath\_table" dẫn đầu cả về số lượng đơn hàng (6203 đơn) và doanh thu (770K), cho thấy đây là mặt hàng có sức mua lớn và giá trị cao. Các danh mục còn lại như sports\_leisure, furniture\_decor, health\_beauty, và housewares cũng nằm trong Top 5 ở cả hai biểu đồ, chứng tỏ chúng có sức tiêu thụ ổn định và đóng góp đáng kể vào doanh thu. Mặc dù thứ tự có sự thay đổi nhẹ giữa hai biểu đồ, nhưng nhìn chung, các danh mục có số lượng đơn hàng cao cũng đồng thời đem lại doanh thu lớn, cho thấy hiệu quả kinh doanh đồng đều.

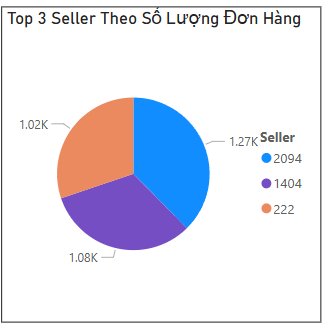
### **7.3.2. Câu hỏi: Cho biết 5 danh mục sản phẩm nhận phản hồi tệ nhất trong năm 2018**

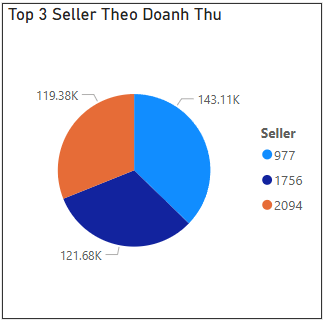
#### A graph of a bar chart AI-generated content may be incorrect.

Biểu đồ thể hiện điểm đánh giá trung bình của khách hàng theo từng danh mục sản phẩm. Có thể nhận thấy:

Danh mục "industry\_commerce" có điểm đánh giá cao nhất với 3.62, cho thấy mức độ hài lòng của khách hàng khá tốt đối với nhóm sản phẩm này. Các danh mục arts\_and\_crafts, music, và furniture\_matters có điểm đánh giá trung bình ở mức khá, dao động từ 3.08 đến 3.51. Danh mục "PC Gamer" có điểm đánh giá thấp nhất, chỉ đạt 1.83, phản ánh có thể tồn tại các vấn đề về chất lượng, dịch vụ hoặc kỳ vọng không được đáp ứng đúng với nhóm sản phẩm này. Từ biểu đồ này, doanh nghiệp có thể xác định được các danh mục cần cải thiện chất lượng hoặc nâng cao trải nghiệm khách hàng, đặc biệt là với nhóm "PC Gamer".

### **7.3.3. Câu hỏi: Cho biết 3 Seller bán được nhiều đơn hàng và đem lại doanh thu nhiều nhất trong năm 2018**





Hai biểu đồ tròn trên thể hiện Top 3 người bán hàng (Seller) theo số lượng đơn hàng và doanh thu:

Seller 2094 dẫn đầu về số lượng đơn hàng (1.27K) và cũng đứng Top 1 về doanh thu (143.11K), cho thấy hiệu quả bán hàng cao cả về số lượng lẫn giá trị. Seller 1404 có số lượng đơn hàng lớn (1.08K) nhưng không xuất hiện trong Top 3 doanh thu, cho thấy giá trị trung bình mỗi đơn hàng có thể thấp. Ngược lại, Seller 977 không nằm trong Top 3 về số đơn hàng nhưng đạt doanh thu cao thứ 2, cho thấy các đơn hàng của người bán này có giá trị cao. Seller 222 xuất hiện trong cả hai biểu đồ, cho thấy sự ổn định về mặt hiệu suất kinh doanh.

Từ đây có thể đánh giá được hiệu quả của từng Seller không chỉ dựa vào số lượng đơn hàng mà còn cần xem xét giá trị đơn hàng để đưa ra chiến lược hỗ trợ phù hợp.

# **8. Kết luận**

## **8.1. Kết quả đạt được**

Dự án “Xây dựng kho dữ liệu từ hệ thống thương mại điện tử OList tại Brazil” đã mang lại cho nhóm chúng em nhiều trải nghiệm quý báu, đồng thời giúp củng cố và nâng cao kiến thức chuyên môn về lĩnh vực kho dữ liệu một cách toàn diện hơn. Thông qua quá trình thực hiện, chúng em đã có cơ hội áp dụng những khái niệm lý thuyết đã học vào thực tế, từ việc xác định các quy trình nghiệp vụ, thiết kế mô hình dữ liệu đa chiều đến triển khai quy trình ETL bằng công cụ SSIS, xây dựng các khối phân tích với SSAS, và trực quan hóa dữ liệu thông qua Power BI Desktop.

Không chỉ dừng lại ở việc vận dụng kỹ thuật, dự án còn giúp nhóm rèn luyện tư duy phân tích, khả năng tiếp cận và xử lý dữ liệu đầu vào một cách logic và có hệ thống. Chúng em học được cách đặt ra các câu hỏi định hướng để khai thác dữ liệu hiệu quả hơn, cũng như cách làm việc nhóm, thảo luận, và đưa ra các quyết định kỹ thuật trong quá trình xây dựng kho dữ liệu. Ngoài ra, chúng em cũng tích lũy thêm kinh nghiệm thực tế trong việc xử lý các vấn đề phát sinh khi thao tác với dữ liệu thô, từ đó nâng cao khả năng thích ứng và giải quyết vấn đề trong môi trường làm việc thực tiễn.

## **8.2. Những hạn chế**

Trong quá trình thực hiện xây dựng kho dữ liệu, nhóm chúng em không tránh khỏi gặp một số những khó khăn vì còn hạn chế về mặt kiến thức. Tập dữ liệu còn nhiều dữ kiện chưa được khai thác hết. Những thực hiện của chúng em chưa đạt đến mức nâng cao trong tổng thể quá trình xây dựng một kho dữ liệu hoàn chỉnh.

# **9. Tài liệu tham khảo**

[1]. Tài liệu các file PDF hướng dẫn về Data Integration với SQL Server, SSIS, SSAS của GVHD Ths. Nguyễn Văn Thành.

[2] Inmon, W. H., “Building the Data Warehouse,” Wiley, 2005.

[3] Kimball, R., Ross, M., “The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling,” Wiley, 2013.

[4] Golfarelli, M., Rizzi, S., “Data Warehouse Design: Modern Principles and Methodologies,” McGraw-Hill, 2009.

[5] IBM, “IBM Data Warehousing and Business Intelligence,” IBM Redbooks, 2010.

[6] Microsoft, “Data Warehousing Guide,” Microsoft Learn, 2023. [Online]. Available: [https://learn.microsoft.com/en-us/sql/relational-databases](https://learn.microsoft.com/en-us/sql/relational-databases%20%20%20%20%20%20%20%20%20)

[7] Oracle, “Data Warehousing Concepts,” Oracle Documentation, 2022. [Online]. Available: <https://docs.oracle.com/en/database/oracle/oracle-database/19/dwhsg/index.html>

[8] Google Cloud, “Introduction to Data Warehousing,” Google Cloud Documentation, 2023. [Online]. Available: https://cloud.google.com/learn/what-is-a-data-warehouse