



**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TPHCM**

**Khoa: Công nghệ thông tin**

**Môn: HTTT phục vụ trí tuệ kinh doanh**

---

# **ĐỒ ÁN MÔN HỌC**

## **XÂY DỰNG VÀ KHAI THÁC KHO**

## **DỮ LIỆU**

---

- Nhóm thực hiện**

Nhóm CQ-BI-10

- Giảng viên hướng dẫn**

Hồ Thị Hoàng Vy

Nguyễn Thị Như Anh

## BẢNG THÔNG TIN CHI TIẾT THÀNH VIÊN

<b>Mã nhóm:</b> CQ-BI-10			
<b>STT</b>	<b>MSSV</b>	<b>Họ và tên</b>	<b>Email</b>
1	1712769	Trịnh Đức Thanh	<a href="mailto:1712769@student.hcmus.edu.vn">1712769@student.hcmus.edu.vn</a>
2	1712828	Huỳnh Thanh Khải Trân	<a href="mailto:1712828@student.hcmus.edu.vn">1712828@student.hcmus.edu.vn</a>
3	1712899	Dương Khánh Vi	<a href="mailto:1712899@student.hcmus.edu.vn">1712899@student.hcmus.edu.vn</a>
4	1712926	Lương Tường Vy	<a href="mailto:1712926@student.hcmus.edu.vn">1712926@student.hcmus.edu.vn</a>

## BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ CÔNG VIỆC

Công việc thực hiện	Người thực hiện	Mức độ hoàn thành
Mô tả chi tiết các bảng thuộc kho dữ liệu.	1712769 – Trịnh Đức Thanh	100%
Phân tích nghiệp vụ và thiết kế Datastore.		
Thực hiện report.		
Mô tả ý nghĩa các thuộc tính nguồn của dữ liệu Census Block.	1712828 – Huỳnh Thanh Khải Trân	100%
Phân tích nghiệp vụ và thiết kế Datastore.		
Thực hiện convert data.		
Thực hiện Cube, OLAP, MDX.	1712899 – Dương Khánh Vi	100%
Phân tích nghiệp vụ và thiết kế Datastore.		
Mô tả các khóa thuộc kho dữ liệu.		
Mô tả ý nghĩa các thuộc tính nguồn của dữ liệu Yellow Taxi Trip.	1712926 – Lương Tương Vy	100%
Phân tích nghiệp vụ và thiết kế Datastore.		
Thực hiện ETL, KPI.		

# MỤC LỤC

A-	YÊU CẦU.....	1
B-	KẾT QUẢ .....	2
I.	Project description .....	2
II.	Key Deliverables .....	2
III.	Critical Success Factors.....	2
IV.	Risks and Concerns .....	3
V.	Business Requirements.....	3
1.	Tiền tổng mỗi chuyến đi.....	3
2.	Thống kê doanh thu năm theo quận.....	4
3.	Phân tích dữ liệu chuyến đi .....	7
4.	Thống kê các loại dữ liệu theo tháng, quý, năm.....	9
VI.	Source data description .....	12
1.	Dữ liệu của nguồn Yellow Taxi năm 2014 .....	12
2.	Dữ liệu của nguồn Yellow Taxi năm 2015 .....	13
3.	Dữ liệu của nguồn Yellow Taxi năm 2016 .....	15
4.	Dữ liệu của nguồn Census Block .....	16
VII.	BI Dimensional Logical Model Design .....	17
a)	Stage .....	17
b)	NDS .....	21
3.	DDS.....	27
4.	Thiết kế Data Flow .....	32
VIII.	ETL.....	33
1.	Source to Stage.....	33
2.	Stage to NDS .....	39
3.	NDS to DDS.....	48
IX.	Khai thác dữ liệu .....	51
1.	OLAP .....	51
2.	Cube .....	58



3.	Truy vấn bằng MDX.....	69
4.	Mining .....	85
5.	KPI .....	85
6.	Report.....	88

# YÊU CẦU ĐỒ ÁN

<b>Loại bài tập</b>	<input checked="" type="checkbox"/> Lý thuyết <input type="checkbox"/> Thực hành <input checked="" type="checkbox"/> Đồ án <input type="checkbox"/> Bài tập
<b>Ngày bắt đầu</b>	<b>28/09/2020</b>
<b>Ngày kết thúc</b>	<b>06/01/2021</b>

## A- YÊU CẦU

Mô tả ý nghĩa các thuộc tính của 2 nguồn dữ liệu trên

Thiết kế kho dữ liệu (KDL) và tổng hợp, nạp dữ liệu các nguồn vào KDL

- Sử dụng dịch vụ Google API Reverse Geocoding (hoặc tương tự) để lấy địa chỉ (street, district, city, state) từ tọa độ đón và trả khách
- Sử dụng dịch vụ hoặc các thư viện có sẵn để lấy Census Block ID từ tọa độ đón và trả khách

Thiết kế và xây dựng Cube

Gợi ý:

- Chuyển đổi dữ liệu ngày tháng sao cho có thể tạo được Date dimension hierarchy Year-Quarter-Month-Date trong Cube
- Xác định và thiết kế các phân cấp chiều còn lại

Khai thác dữ liệu:

OLAP:

- Phân tích chuyển di theo Geography (street, district, city, state, boroughs)..., theo Census Block, theo thời gian, theo loại thanh toán...

Report:

- Dùng regional map để biểu diễn trực quan (bằng màu sắc) sự phân bố số lượng đón xe (pickups) và số lượng trả khách (drop-offs) ở các vùng (district, city) theo thời gian tháng, quý, năm?
- Dùng regional map để biểu diễn trực quan (bằng màu sắc) sự phân bố số lượng đón xe (pickups) và số lượng trả khách (drop-offs) ở các Census Block theo thời gian tháng, quý, năm?
- Phân tích thời điểm nào trong ngày là giờ cao điểm và thấp điểm của các chuyến taxi? Vẽ đồ thị phân bố số lượng đón taxi theo giờ trong ngày trong tất cả các năm.
- Thống kê doanh thu năm theo quận (district) đón khách.

- Dưới góc độ là người khai thác dữ liệu, SV tự đề xuất và thực hiện các nhu cầu phân tích khác.

## Prediction

Gợi ý:

- Dự đoán vào một thời điểm cụ thể trong năm theo từng district thì lượng khách đón taxi sẽ như thế nào
- Để phân chia thành phố new york thành khu vực để có thể thực hiện dự đoán theo khu vực • sử dụng thuật toán K-mean...
- Xây dựng một mô hình dự đoán cho số tiền boa theo tỷ lệ phần trăm của tổng giá vé
- ....
- Sinh viên tự đề xuất 1 yêu cầu phân tích, lựa chọn mô hình phù hợp.

## B- KẾT QUẢ

### I. Project description

Mục tiêu của đồ án là có thể giúp cho sinh viên có thể phân tích và khai thác dữ liệu dựa trên các nguồn dữ liệu đã được cung cấp thông qua bộ công cụ SSDT của Microsoft (Gồm SISS, SASS, SRSS). Đối tượng dữ liệu của đồ án là Yellow Taxi từ năm 2014 đến 2018 và tình trạng dân cư của bang New York (Census Block).

Với những đối tượng này, sinh viên tiến hành phân tích dữ liệu qua các công việc: xây dựng kho dữ liệu, kiến trúc Data flow, lọc dữ liệu (Data cleaning), ETL dữ liệu từ nguồn đến Kho dữ liệu; Khai thác Kho dữ liệu qua các công việc: report, OLAP, mining, lập lịch định kỳ thực hiện ETL. Qua việc phân tích và khai thác, chúng ta sẽ tổng kết, báo cáo, biểu diễn trực quan số lượng, địa điểm chuyển đi cùng với doanh số và dự đoán tình hình kinh doanh. Từ đó giúp họ có cái nhìn trực quan, rõ ràng hơn để thực hiện các quyết định trong kinh doanh sắp tới

### II. Key Deliverables

Đồ án có thể tạo ra KDL đáp ứng nhu cầu lưu trữ, phân tích, khai thác, trực quan dữ liệu và dự đoán tình hình kinh doanh cho người dùng. Giúp họ có cái nhìn rộng, nhiều chiều hơn về tình hình của tổ chức để có quyết định hợp lý.

Kết quả của đồ án là KDL, những report trên nhiều tiêu chí, biểu đồ trực quan và các dự đoán doanh thu, tình hình kinh doanh của doanh nghiệp.

### III. Critical Success Factors

- Nguồn dữ liệu: Nguồn dữ liệu cung cấp phải chính xác, đồng bộ, chính xác. Nếu chưa đáp ứng đủ các tiêu chí này thì cần phải thực hiện lọc, biến đổi để cho các thuộc tính trong nguồn có cùng chung tính chất.

- KDL: Kho dữ liệu phải đáp ứng được nhu cầu lưu trữ ở quá khứ và hiện tại, có khả năng đáp ứng các nhu cầu phân tích của người dùng.
- Yêu cầu người dùng: Cần phải nắm rõ được yêu cầu của người dùng để phân tích và khai thác dữ liệu ở khía cạnh mà họ cần.

#### **IV. Risks and Concerns**

- Nghiệp vụ liên quan: Cần có các thông tin chính xác về nghiệp vụ của đối tượng liên quan để có cái nhìn đúng đắn, tạo ra các measure chính xác cho dữ liệu
- Các công cụ chuyển đổi: Các công cụ chuyển đổi vị trí hiện tại vẫn còn giới hạn về số lượng dữ liệu. Để có thể chuyển đổi lượng lớn dữ liệu thì cần phải nghiên cứu và áp dụng nhiều công cụ vào cùng lúc, điều này đôi lúc gây trở ngại cho người phát triển đồ án.
- Các chính sách bên ngoài: Các chính sách liên quan như thuế, phí đường bộ, các chính sách của chính phủ,... có liên quan đến dữ liệu có thể thay đổi theo thời gian. Nếu dự án thực hiện trong thời gian dài thì đòi hỏi phải có sự hiểu biết, cập nhật kịp thời để vận hành, sửa chữa.

#### **V. Business Requirements**

<b>Identifier</b>	<b>Name</b>	<b>Description</b>	<b>Priority</b>
1	Tiền tổng mỗi chuyến đi	Nghiệp vụ về quy định tính tiền tổng mỗi chuyến đi, bao gồm tiền mỗi chuyến, tiền thuế, tiền phí cầu đường,... các loại tiền phát sinh	1
2	Thống kê doanh thu năm theo quận	Nghiệp vụ về quy định thống kê doanh thu năm theo quận	1
3	Phân tích dữ liệu chuyến đi	Nghiệp vụ để phân tích chuyến đi theo địa chỉ cụ thể (gồm đường, quận, thành phố, bang), theo thời gian,...	1
4	Thống kê các loại dữ liệu theo tháng, năm.	Nghiệp vụ quy định về các loại mốc thời gian cho việc thống kê dữ liệu chuyến đi.	1

##### **1. Tiền tổng mỗi chuyến đi**

###### **1.1. Requirement description**

Nghiệp vụ về quy định công thức tính tiền tổng mỗi chuyến đi, bao gồm tiền mỗi chuyến, tiền thuế, tiền phí đường bộ,... các loại tiền phát sinh.

###### **1.2. Data sources**

Column	Data Type	Description
fare_amount	Float	Số tiền gốc của chuyến đi tính theo đồng hồ. (Đơn vị khoảng cách là met)
extra	Float	Phí phụ thu Có 2 mức giá cho phí phụ thu: + 0.5\$ cho giờ cao điểm + 1\$ cho qua đêm
mta_tax	Float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
tip_amount	Float	Tiền boa
tolls_amount	Float	Tiền đi qua trạm thu phí (phí cầu đường)
improvement_surcharge	Float	Phí phát sinh cải thiện chuyến đi. Phụ phí quy định là 0.30\$ tại điểm thả cờ. Phí này được tính bắt đầu từ năm 2015
total_amount	Float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện)

### 1.3. Business and data transformations

- Nghệp vụ chính:* Cách tính tổng tiền của xe taxi sau mỗi chuyến đi.
- Data transformation:* Không có

### 1.4. Business metrics or Key Performance Indicators (KPIs)

- Total\_amount:* tiền tổng mỗi chuyến đi
- Quy định tính tiền của mỗi chuyến đi:*

$$\text{total\_amount} = \text{tolls\_amount} + \text{Extra} + \text{MTA\_tax}$$

### 1.5. Business processes

B1: Tài xế bấm hoàn thành chuyến đi

B2: Đồng hồ bấm giờ hiển thị số tiền

B3: Khách hàng trả tiền theo phương thức thanh toán mà mình chọn

B4: Thanh toán

### 1.6. List business groups involved and describe type of involvement

Các đối tượng liên quan đến yêu cầu: Doanh nghiệp Yellow Taxi và Chính phủ.

- Hệ thống của doanh nghiệp sẽ ước lượng số tiền của mỗi chuyến dựa trên khoảng cách mà mỗi chuyến đã đi + thuế + phí cầu đường.
- Chính phủ quy định các loại phí như thuế, phí cầu đường. Các loại phí này có thể thay đổi tùy theo chính sách mỗi năm

## 2. Thống kê doanh thu năm theo quận

### 2.1. Requirement description

Nghiệp vụ về quy định việc thống kê doanh thu năm theo khu vực quận. Nghiệp vụ chỉ rõ thời điểm bắt đầu và kết thúc việc thống kê doanh thu.

## 2.2. Data sources

- Dữ liệu NYC Yellow Taxi

<b>Column</b>	<b>Data Type</b>	<b>Description</b>
tpep_pickup_datetime	Datetime	Thời gian đồng hồ bắt đầu hoạt động
tpep_dropoff_datetime	Datetime	Thời gian đồng hồ ngừng hoạt động
passenger_count	Int	Số lượng hành khách trên chuyến đi
trip_distance	Float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
PULocationID	Int	Khu vực mà đồng hồ tính tiền của taxi bắt đầu hoạt động
DOLocationID	Int	Khu vực mà đồng hồ tính tiền của taxi ngừng hoạt động
RatecodeID	Int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
store_and_fwd_flag	Bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
Pickup_latitude	Decimal	Vĩ độ điểm đón khách
Pickup_longitude	Decimal	Kinh độ điểm đón khách
Dropoff_longitude	Decimal	Kinh độ điểm trả khách
Dropoff_latttude	Decimal	Vĩ độ điểm trả khách
payment_type	Int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn

fare_amount	Float	Số tiền gốc của chuyến đi tính theo đồng hồ
extra	Float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
mta_tax	Float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
tip_amount	Float	Tiền boa
tolls_amount	Float	Tiền đi qua trạm thu phí (phí cầu đường)
improvement_surcharge	Float	Phí phát sinh cải thiện chuyến đi.
total_amount	Float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa)

- Dữ liệu nguồn Census Block

Column	Data Type	Description
the_geom	Multipolygon	Tọa độ địa lý để xác định hình thể của quận. Tọa độ gồm nhiều điểm để nối thành hình dạng của 1 khu vực.
CTLabel	String	Mã định danh cho đường điều tra dân số của 1 quận
BoroName	String	Boro là viết tắt của Borough Boundary. Tên các khu vực trong New York (Ở đây ta lấy đơn vị là quận)
BoroCode	String	Mã code để định vị các quận. Code gồm số từ 1 đến 5 tương ứng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
BoroCT2010	String	Chỉ số điều tra dân số của tất cả các khu dân cư trong quận. Dựa trên công thức $Boro * 1000000 + CT2010$
NTACode	String	Mã code đánh dấu cho các khu dân cư của quận trong bang New York

### 2.3. Business and data transformations

*Nghệp vụ chính:* Thống kê doanh thu các chuyến đi trong năm của một quận. Doanh thu được tính kể từ 12:00 am ngày 1/1 năm cần thống kê đến 12:00 am ngày 1/1 năm tiếp theo trong phạm vi một quận.

*Data transformation:*

- Pickup\_latitude, Pickup\_longitude, Dropoff\_longitude, Dropoff\_latitude: Đây là các thuộc tính tọa độ kinh độ và vĩ độ của 1 khu vực. Cần phải dùng công

cụ Google API để đổi các tọa độ này thành một vị trí cụ thể, có địa chỉ gồm: đường, quận, thành phố, bang.

- Payment\_type: có sự khác biệt kiểu dữ liệu giữa các năm 2014 và 2015, 2016. Năm 2014 là kiểu string, năm 2015 và 2016 là int.

## 2.4. Business metrics or Key Performance Indicators (KPIs)

Total\_amount

## 2.5. Business processes

Không có.

## 2.6. List business groups involved and describe type of involvement

Các đối tượng liên quan đến yêu cầu: Doanh nghiệp Yellow Taxi và Chính phủ

## 3. Phân tích dữ liệu chuyến đi

### 3.1. Requirement description

Nghiệp vụ để phân tích chuyến đi theo địa chỉ cụ thể (gồm đường, quận, thành phố, bang), theo Census Block, theo thời gian, theo loại thanh toán...

- Phân tích theo địa chỉ cụ thể: Tại 1 địa chỉ cụ thể (gồm đường, quận, thành phố, bang) có bao nhiêu chuyến xe, giờ lên xe, giờ xuống xe, quãng đường chuyến đi, số lượng hành khách, đơn giá. Ở địa chỉ nào thì có số lượng chuyến đi nhiều nhất, ít nhất.
- Phân tích theo Census Block: Tại 1 vùng dân cư của quận có bao nhiêu chuyến xe, giờ lên xe, giờ xuống xe, quãng đường chuyến đi, số lượng hành khách, đơn giá. Vùng có chuyến đi nhiều nhất, ít nhất
- Phân tích theo thời gian: Tại 1 thời điểm cụ thể (giờ, ngày, tháng, năm) có bao nhiêu chuyến xe, giờ lên xe, giờ xuống xe, quãng đường chuyến đi, số lượng hành khách, đơn giá. Phân tích từng chuyến xe ở từng thời điểm cụ thể. Tại thời điểm nào là cao điểm, thấp điểm, doanh thu trung bình mỗi chuyến đi theo thời gian.
- Phân tích theo loại thanh toán: Phân tích chuyến đi với các loại thanh toán hiện có. Chuyến đi thực hiện loại thanh toán nào, loại thanh toán nào được áp dụng nhiều nhất, ít nhất.

### 3.2. Data sources

- Dữ liệu NYC Yellow Taxi

Column	Data Type	Description
tpep_pickup_datetime	Datetime	Thời gian đồng hồ bắt đầu hoạt động
tpep_dropoff_datetime	Datetime	Thời gian đồng hồ ngừng hoạt động
passenger_count	Int	Số lượng hành khách trên chuyến đi

trip_distance	Float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
PULocationID	Int	Khu vực mà đồng hồ tính tiền của taxi bắt đầu hoạt động
DOLocationID	Int	Khu vực mà đồng hồ tính tiền của taxi ngừng hoạt động
RatecodeID	Int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
store_and_fwd_flag	Bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y = có lưu lại và gửi cho server N = không lưu lại và gửi cho server
Pickup_latitude	Decimal	Vĩ độ điểm đón khách
Pickup_longitude	Decimal	Kinh độ điểm đón khách
Dropoff_longitude	Decimal	Kinh độ điểm trả khách
Dropoff_lattitude	Decimal	Vĩ độ điểm trả khách
payment_type	Int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
fare_amount	Float	Số tiền gốc của chuyến đi tính theo đồng hồ
extra	Float	Phí phụ thu Có 2 mức giá cho phí phụ thu: + 0.5\$ cho giờ cao điểm + 1\$ cho qua đêm
mta_tax	Float	Thuế

		Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
tip_amount	Float	Tiền boa
tolls_amount	Float	Tiền đi qua trạm thu phí (phí cầu đường)
improvement_surcharge	Float	Phí phát sinh cải thiện chuyến đi.
total_amount	Float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa)

- Dữ liệu Census Block

Column	Data Type	Description
the_geom	Multipolygon	Tọa độ địa lý để xác định hình thể của quận. Tọa độ gồm nhiều điểm để nối thành hình dạng của 1 khu vực.
CTLabel	String	Mã định danh cho đường điều tra dân số của 1 quận
BoroName	String	Boro là viết tắt của Borough Boundary. Tên các khu vực trong New York (Ở đây ta lấy đơn vị là quận)
BoroCode	String	Mã code để định vị các quận. Code gồm số từ 1 đến 5 tương trưng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
NTACode	String	Mã code đánh dấu cho các khu dân cư của quận trong bang New York

### 3.3. Business and data transformations

*Nghệp vụ chính:* Phân tích dữ liệu chuyến đi theo nhiều tiêu chí, bao gồm: theo địa chỉ cụ thể, theo Census Block, theo thời gian, theo loại thanh toán.

*Data transformation:*

- Pickup\_latitude, Pickup\_longitude, Dropoff\_longitude, Dropoff\_latitude: Đây là các thuộc tính tọa độ kinh độ và vĩ độ của 1 khu vực. Cần phải dùng công cụ Google API để đổi các tọa độ này thành một vị trí cụ thể, có địa chỉ gồm: đường, quận, thành phố, bang.

### 3.4. Business metrics or Key Performance Indicators (KPIs)

Không có.

### 3.5. Business processes

Không có.

### 3.6. List business groups involved and describe type of involvement

Các đối tượng liên quan đến yêu cầu: Doanh nghiệp Yellow Taxi

## 4. Thống kê các loại dữ liệu theo tháng, quý, năm

#### **4.1. Requirement description**

Nghệp vụ quy định về thời gian cho việc thống kê, gồm các loại thời gian: tháng, quý, năm

- Theo tháng: báo cáo vào lúc 12:00 am ngày 1 hàng tháng
- Theo quý: Một năm chia thành 4 quý:
  - Quý 1: từ 12:00 am ngày 1/1 đến 12:00 am ngày 1/4
  - Quý 2: từ 12:00 am ngày 1/4 đến 12:00 am ngày 1/7
  - Quý 3: từ 12:00 am ngày 1/7 đến 12:00 am ngày 1/10
  - Quý 4: từ 12:00 am ngày 1/10 đến 12:00 am ngày 1/1
- Theo năm: Báo cáo được tính kể từ 12:00 pm ngày 1/1 năm cần thống kê đến 12:00 pm ngày 1/1 năm tiếp theo.

#### **4.2. Data sources**

- Dữ liệu NYC Yellow Taxi

<b>Column</b>	<b>Data Type</b>	<b>Description</b>
VendorID	Int	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMD = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
tpep_pickup_datetime	Datetime	Thời gian đồng hồ bắt đầu hoạt động
tpep_dropoff_datetime	Datetime	Thời gian đồng hồ ngừng hoạt động
passenger_count	Int	Số lượng hành khách trên chuyến đi
trip_distance	Float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
PULocationID	Int	Khu vực mà đồng hồ tính tiền của taxi bắt đầu hoạt động
DOLocationID	Int	Khu vực mà đồng hồ tính tiền của taxi ngừng hoạt động
RatecodeID	Int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
store_and_fwd_flag	Bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không.

		Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
Pickup_lattude	Decimal	Vĩ độ điểm đón khách
Pickup_longtitude	Decimal	Kinh độ điểm đón khách
Dropoff_longtitude	Decimal	Kinh độ điểm trả khách
Dropoff_lattude	Decimal	Vĩ độ điểm trả khách
payment_type	Int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
fare_amount	Float	Số tiền gốc của chuyến đi tính theo đồng hồ
extra	Float	Phí phụ thu Có 2 mức giá cho phí phụ thu: + 0.5\$ cho giờ cao điểm + 1\$ cho qua đêm
mta_tax	Float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
tip_amount	Float	Tiền boa
tolls_amount	Float	Tiền đi qua trạm thu phí (phí cầu đường)
improvement_surcharge	Float	Phí phát sinh cải thiện chuyến đi.
total_amount	Float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa)

- Dữ liệu Census Block

Column	Data Type	Description
the_geom	Multipolygon	Tọa độ địa lý để xác định hình thể của quận. Tọa độ gồm nhiều điểm để nối thành hình dạng của 1 khu vực.
CTLabel	String	Mã định danh cho đường điều tra dân số của 1 quận
BoroName	String	Boro là viết tắt của Borough Boundary. Tên các khu vực trong New York (Ở đây ta lấy đơn vị là quận)
BoroCode	String	Mã code để định vị các quận.

		Code gồm số từ 1 đến 5 tượng trưng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
NTACode	String	Mã code đánh dấu cho các khu dân cư của quận trong bang New York

#### 4.3. Business and data transformations

*Nghiệp vụ chính:* Nghiệp vụ quy định về khoảng thời gian cho việc thống kê.

*Data transformation:*

- Các loại ngày tháng của pickup\_datetime, dropoff\_datetime, tpep\_pickup\_datetime, tpep\_dropoff\_datetime: kiểu dữ liệu giống nhau nhưng hình thức của dữ liệu không giống nhau, năm 2014 là ngày/tháng/năm, giờ (24 giờ), năm 2015 và 2016 là năm, tên tháng, ngày, giờ (12 giờ) có pm và am.

#### 4.4. Business metrics or Key Performance Indicators (KPIs)

Không có.

#### 4.5. Business processes

Không có.

#### 4.6. List business groups involved and describe type of involvement

Các đối tượng liên quan đến yêu cầu: Doanh nghiệp Yellow Taxi

### VI. Source data description

#### 1. Dữ liệu của nguồn Yellow Taxi năm 2014

STT	Column	Data Type	Description
1	vendor_id	string	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
2	pickup_datetime	datetime (dd/mm/yy time)	Thời gian đón khách
3	dropoff_datetime	datetime (dd/mm/yy time)	Thời gian trả khách
4	passenger_count	int	Số lượng hành khách trên chuyến đi
5	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
6	pickup_longitude	decimal	Kinh độ điểm đón khách
7	pickup_latitude	decimal	Vĩ độ điểm đón khách
8	store_and_fwd_flag	bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không.

			Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
9	dropoff_longitude	decimal	Kinh độ điểm trả khách
10	dropoff_latitude	decimal	Vĩ độ điểm trả khách
11	payment_type	string	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
12	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
13	mta_tax	float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
14	tip_amount	float	Tiền boa
15	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
16	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện)
17	imp_surcharge	float	Phí phát sinh cải thiện chuyến đi.
18	extra	float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
19	rate_code	int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm

## 2. Dữ liệu của nguồn Yellow Taxi năm 2015

STT	Column	Data Type	Description
1	vendor_id	int	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System

2	pickup_datetime	datetime (year month day time)	Thời gian đón khách
3	dropoff_datetime	datetime (year month day time)	Thời gian trả khách
4	passenger_count	int	Số lượng hành khách trên chuyến đi
5	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
6	pickup_longitude	decimal	Kinh độ điểm đón khách
7	pickup_latitude	decimal	Vĩ độ điểm đón khách
8	store_and_fwd_flag	bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
9	dropoff_longitude	decimal	Kinh độ điểm trả khách
10	dropoff_latitude	decimal	Vĩ độ điểm trả khách
11	payment_type	int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
12	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
13	mta_tax	float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
14	tip_amount	float	Tiền bo
15	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
16	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền bo và phí phát sinh cải thiện)
17	imp_surcharge	float	Phí phát sinh cải thiện chuyến đi.
18	extra	float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
19	rate_code	NULL	

20	RateCodeID	int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
----	------------	-----	--

### 3. Dữ liệu của nguồn Yellow Taxi năm 2016

STT	Column	Data Type	Description
1	VendorID	int	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
2	tpep_pickup_datetime	(datetime year month day time)	Thời gian đón khách
3	tpep_dropoff_datetime	(datetime year month day time)	Thời gian trả khách
4	passenger_count	int	Số lượng hành khách trên chuyến đi
5	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
6	store_and_fwd_flag	bool	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
7	payment_type	int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
8	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
9	mta_tax	float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
10	tip_amount	float	Tiền boa

11	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
12	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền bo và phí phát sinh cải thiện)
13	improvement_surcharge	float	Phí phát sinh cải thiện chuyến đi.
14	extra	float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
15	RatecodeID	int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền bo, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
16	PULocationID	int	ID định vị vị trí đón khách
17	DOLocationID	int	ID định vị vị trí trả khách

#### 4. Dữ liệu của nguồn Census Block

STT	Column	Data Type	Description
1	the_geom	Multipolygon	Tọa độ địa lý để xác định hình thể của quận. Tọa độ gồm nhiều điểm để nối thành hình dạng của 1 khu vực.
2	CTLabel	String	Mã định danh cho đường điều tra dân số của 1 quận
3	BoroName	String	Boro là viết tắt của Borough Boundary. Tên các khu vực trong New York (Ở đây ta lấy đơn vị là quận)
4	BoroCode	String	Mã code để định vị các quận. Code gồm số từ 1 đến 5 tương ứng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
5	CT2010	Int	Chỉ số điều tra dân số của một khu dân cư trong quận.
6	BoroCT2010	String	Chỉ số điều tra dân số của tất cả các khu dân cư trong quận. Dựa trên công thức $Boro * 1000000 + CT2010$
7	CDEligibil	String	Điều kiện nhận tài trợ phát triển cộng đồng của khu dân cư. Gồm 2 loại: Đạt và Không đạt (Ineligible và Eligible)
8	NTACode	String	Mã code đánh dấu cho các khu dân cư của quận trong bang New York
9	NTAName	String	Viết tắt của Neighborhood Tabulation Areas. Là các khu dân cư tập trung đông dân số của quận trong bang New York
10	PUMA	String	Khu vực sử dụng dữ liệu công cộng

11	Shape_Leng	Double	Chiều dài của khu vực
12	Shape_Area	Double	Diện tích của khu vực

## VII. BI Dimensional Logical Model Design

### a) Stage

Table	Column	Data Type	Description
YellowTaxi 2014	vendor_id	Varchar(10)	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
	pickup_datetime	datetime	Thời gian đón khách
	dropoff_datetime	datetime	Thời gian trả khách
	passenger_count	int	Số lượng hành khách trên chuyến đi
	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
	pickup_longitude	decimal	Kinh độ điểm đón khách
	pickup_latitude	decimal	Vĩ độ điểm đón khách
	store_and_fwd_flag	bit	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
	dropoff_longitude	decimal	Kinh độ điểm trả khách
	dropoff_latitude	decimal	Vĩ độ điểm trả khách
	payment_type	Varchar(10)	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
	mta_tax	float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
	tip_amount	float	Tiền bo
	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền bo và phí phát sinh cải thiện)
	imp_surcharge	float	Phí phát sinh cải thiện chuyến đi.
	extra	float	Phí phụ thu

YellowTaxi 2015			Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
	rate_code	Varchar(10)	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
	vendor_id	int	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
	pickup_datetime	datetime	Thời gian đón khách
	dropoff_datetime	datetime	Thời gian trả khách
	passenger_count	int	Số lượng hành khách trên chuyến đi
	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
	pickup_longitude	decimal	Kinh độ điểm đón khách
	pickup_latitude	decimal	Vĩ độ điểm đón khách
	store_and_fwd_flag	bit	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
	dropoff_longitude	decimal	Kinh độ điểm trả khách
	dropoff_latitude	decimal	Vĩ độ điểm trả khách
	payment_type	Int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyến đi bị hoãn
	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
	mta_tax	float	Thuế

			Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
	tip_amount	float	Tiền boa
	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện)
	imp_surcharge	float	Phí phát sinh cải thiện chuyến đi.
	extra	float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
	rate_code	NULL	
	RateCodeID	int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
Yellow Taxi 2016	VendorID	int	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ. CMT = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
	tpep_pickup_datetime	datetime	Thời gian đón khách
	tpep_dropoff_datetime	datetime	Thời gian trả khách
	passenger_count	int	Số lượng hành khách trên chuyến đi
	trip_distance	float	Khoảng cách chuyến đi dựa trên đồng hồ taxi
	store_and_fwd_flag	bit	Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không. Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
	payment_type	int	Mã biểu thị phương thức thanh toán cho mỗi chuyến đi. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí

			4 = Tranh chấp 5 = Không xác định 6 = Chuyển đi bị hoãn
	fare_amount	float	Số tiền gốc của chuyến đi tính theo đồng hồ
	mta_tax	float	Thuế Mức giá mặc định là 0.5\$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại
	tip_amount	float	Tiền boa
	tolls_amount	float	Tiền đi qua trạm thu phí (phí cầu đường)
	total_amount	float	Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện)
	improvement_surcharge	float	Phí phát sinh cải thiện chuyến đi.
	extra	float	Phí phụ thu Có 2 mức giá cho phí phụ thu: 0.5\$ cho giờ cao điểm 1\$ cho qua đêm
CensusBlock	RatecodeID	int	Mã được gom nhóm theo giá trị trung bình của số tiền gốc, số tiền boa, tổng số tiền và khoảng cách chuyến đi (hay mã giá cuối cùng có hiệu lực vào cuối chuyến đi). 1 = Tỷ lệ chuẩn 2 = JFK 3 = Newark 4 = Nassau hoặc Westchester 5 = Số tiền thương lượng 6 = Đi theo nhóm
	PULocationID	int	ID định vị vị trí đón khách
	DOLocationID	int	ID định vị vị trí trả khách
	the_geom	Varchar(1000)	Tọa độ địa lý để xác định hình thể của quận. Tọa độ gồm nhiều điểm để nối thành hình dạng của 1 khu vực.
	CTLabel	Varchar(10)	Mã định danh cho đường điều tra dân số của 1 quận
	BoroName	Varchar(50)	Boro là viết tắt của Borough Boundary. Tên các khu vực trong New York (Ở đây ta lấy đơn vị là quận)
	BoroCode	Varchar(10)	Mã code để định vị các quận. Code gồm số từ 1 đến 5 tương trưng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
	CT2010	Varchar(10)	Chỉ số điều tra dân số của một khu dân cư trong quận.

	BoroCT2010	Varchar(10)	Chỉ số điều tra dân số của tất cả các khu dân cư trong quận. Dựa trên công thức Boro*1000000+CT2010
	CDEligibil	Varchar(10)	Điều kiện nhận tài trợ phát triển cộng đồng của khu dân cư. Gồm 2 loại: Đạt và Không đạt (Ineligible và Eligible)
	NTACode	Varchar(10)	Mã code đánh dấu cho các khu dân cư của quận trong bang New York
	NTAName	Varchar(50)	Viết tắt của Neighborhood Tabulation Areas. Là các khu dân cư tập trung đông dân số của quận trong bang New York
	PUMA	Varchar(10)	Khu vực sử dụng dữ liệu công cộng
	Shape_Leng	Double	Chiều dài của khu vực
	Shape_Area	Double	Diện tích của khu vực

## b) NDS

### 2.1. COLUMNS

Table	Column	Key Type	Data Type	Is Nullable	Default Values	Description
BILL	Bill_ID	PK	Int Identity(1,1)	N		Mã hóa đơn của một chuyến đi
	Vendor_ID	FK	Int	N		Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ.
	PU_ID	FK	Int	N		Khu vực mà đồng hồ tính tiền của taxi bắt đầu hoạt động
	DO_ID	FK	Int	N		Khu vực mà đồng hồ tính tiền của taxi ngừng hoạt động
	Payment_type	FK	Int	N		Mã phương thức thanh toán.
	RatecodeID	FK	Int	N		Mã giá cuối cùng có hiệu lực vào cuối chuyến đi.
	Extra		Float	N		Phí phụ thu.
	Mta_tax		Float	N		Thuế.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
	TrangThai	FK	Int	N		Mã trạng thái
	store_and_fwd_flag		Varchar(1)	N		Cờ hiệu cho biết chuyến đi có được lưu lại và gửi cho server vendor hay không.

					Y= có lưu lại và gửi cho server N= không lưu lại và gửi cho server
	pass_count	Int	N		Số lượng hành khách trên chuyến đi
	trip_distance	Float	N		Khoảng cách chuyến đi dựa trên đồng hồ taxi
	fare_amount	Float	N		Số tiền gốc của chuyến đi tính theo đồng hồ
	tip_amount	Float	N		Tiền boa
	tolls_amount	Float	Y		Tiền đi qua trạm thu phí (phí cầu đường)
	Impr_surcharge	Float	N		Phí phát sinh cải thiện chuyến đi.
	total_amount	Float	N		Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện)
	NgayCapNhat	Datetime	Y		Ngày cập nhật dữ liệu.
NGUỒN DỮ LIỆU	ID	PK Int Identity(1,1)	N		Mã tăng tự động cho bảng Nguồn dữ liệu.
	NguonDL	Varchar(50)	N		Tên nguồn dữ liệu.
	NgayCapNhat	Datetime	Y		Ngày cập nhật nguồn dữ liệu.
TRẠNG THÁI	ID	PK Int Identity(1,1)	N		Mã tăng tự động cho bảng Trạng thái.
	TrangThai	Varchar(50)	N		Tên trạng thái.
VENDOR	ID	PK Int Identity(1,1)	N		Mã tăng tự động cho bảng Vendor.
	VendorID	Varchar(10)	N	CMD / VTS / DDS	Mã cho biết nhà cung cấp TPEP đã cung cấp các bảng lưu trữ.
	VendorName	Varchar(50)	N		Tên Vendor tương ứng với VendorID. CMD = Creative Mobile Technologies; VTS = VeriFone Inc; DDS = Digital Dispatch System
	NgayCapNhat	Datetime	N		Ngày cập nhật dữ liệu bảng Vendor.
	NguonDL	FK Int	N		Mã nguồn dữ liệu.
	TrangThai	FK Int	N		Mã trạng thái.

<b>PAYMENT</b>	ID	PK	Int Identity(1,1)	N		Mã tăng tự động cho bảng Payment.
	Payment_type		Int	N	Từ 1 đến 6	Mã phương thức thanh toán.
	PaymentName		Varchar(50)	N		Phương thức thanh toán. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyển đi bị hoãn
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu bảng Payment.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
	TrangThai	FK	Int	N		Mã trạng thái.
<b>PICK_UP</b>	PU_ID	PK	Int Identity(1,1)	N		Mã tăng tự động cho bảng PUlocation.
	PU_Datetime		Datetime	N		Thời điểm mà đồng hồ tính tiền của taxi bắt đầu hoạt động.
	PU_latitude		Varchar(100)	N		Vĩ độ điểm đón khách.
	PU_longitude		Varchar(100)	N		Kinh độ điểm đón khách.
	PU_Number		Varchar(100)	Y		Số địa chỉ đón khách
	PU_Street		Varchar(100)	Y		Đường đón khách
	PU_CensusBlock		Varchar(100)	Y		Census block đón khách
	PU_City		Varchar(100)	Y		Thành phố đón khách
	PU_State		Varchar(100)	Y		Bang đón khách
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu bảng Pick_up.
	PU_NTA	FK	Int	N		Mã khu dân cư thuộc quận.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
<b>DROP-OFF</b>	TrangThai	FK	Int	N		Mã trạng thái.
	ID	PK	Int Identity(1,1)	N		Mã tăng tự động cho bảng DOLocation.
	DO_Datetime		Datetime	N		Thời điểm mà đồng hồ tính tiền của taxi ngừng hoạt động.
	DO_longitude		Varchar(100)			Kinh độ điểm trả khách.
	DO_latitude		Varchar(100)	N		Vĩ độ điểm trả khách.

	DO_Number		Varchar(10)	Y		Số địa chỉ trả khách
	DO_Street		Varchar(100)	Y		Đường trả khách
	DO_City		Varchar(100)	Y		Thành phố trả khách
	DO_CensusBlock		Varchar(100)	Y		Census block trả khách
	DO_State		Varchar(100)	Y		Bang trả khách
	DO_NTA	FK	Int	N		Mã khu dân cư thuộc quận.
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu bảng Drop_off.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
	TrangThai	FK	Int	N		Mã trạng thái.
NTA	ID	PK	Int Identity(1,1)	N		Mã tăng tự động cho bảng NTA.
	NTACode		Varchar(100)	N		Mã code đánh dấu cho các khu dân cư của quận trong bang New York
	NTAName		Varchar(100)	N		Viết tắt của Neighborhood Tabulation Areas. Là các khu dân cư tập trung đông dân số của quận trong bang New York
	PUMA		Varchar(100)	N		Khu vực sử dụng dữ liệu công cộng
	BoroID	FK	Int	N		Mã quận.
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu bảng NTA.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
	TrangThai	FK	Int	N		Mã trạng thái.
BORO	ID	PK	Int Identity(1,1)	N		Mã tăng tự động cho bảng Boro.
	BoroCode		Varchar(10)	N	Từ 1 đến 5	Mã định vị quận. Code gồm số từ 1 đến 5 tương ứng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
	BoroName		Varchar(50)	N		Tên các quận trong New York
	CTLable		Varchar(100)	N		Mã định danh cho đường điều tra dân số của 1 quận

	CT2010		Varchar(100)	N		Chỉ số điều tra dân số của một khu dân cư trong quận.
	CDEligibil		Varchar(100)	N		Điều kiện nhận tài trợ phát triển cộng đồng của khu dân cư. Gồm 2 loại: Đạt và Không đạt (Ineligible và Eligible)
	Shape_leng		Varchar(100)	N		Chiều dài của khu vực
	Shape_area		Varchar(100)	N		Diện tích của khu vực
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu bảng Boro.
	NguonDL	FK	Int	N		Mã nguồn dữ liệu.
	TrangThai	FK	Int	N		Mã trạng thái.

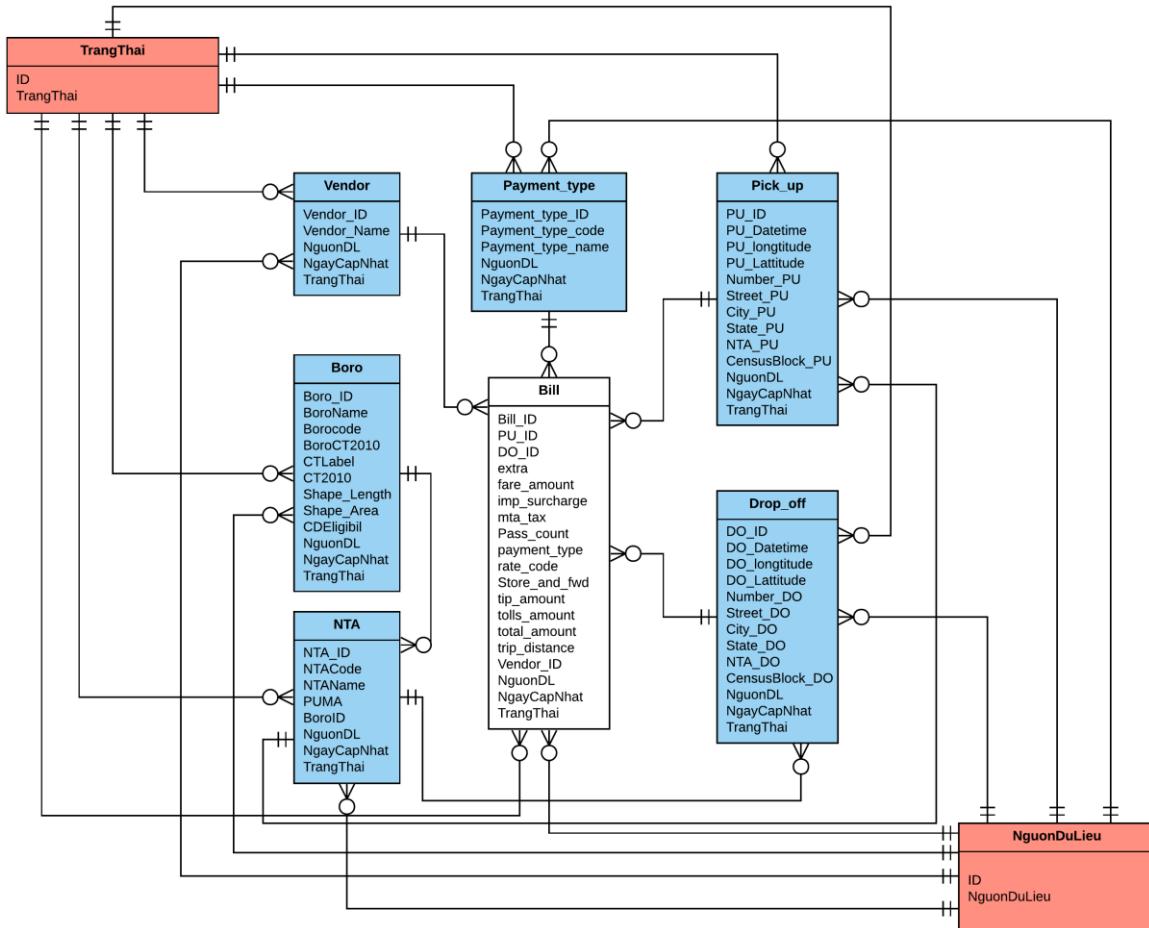
## 2.2. CONSTRAINTS

Constraint Name	Constraint Type	Table	Column	Description
PK_BILL	Primary key	BILL	Bill_ID	Khoá chính của bảng BILL
FK_BILL_VENDOR	Foreign key	BILL	VendorID	Khoá ngoại tham chiếu từ BILL đến VENDOR
FK_BILL_PULOCATION	Foreign key	BILL	PUID	Khoá ngoại tham chiếu từ BILL đến PULOCATION
FK_BILL_DOLOACTION	Foreign key	BILL	DOID	Khoá ngoại tham chiếu từ BILL đến DOLOACTION
FK_BILL_PAYMENT	Foreign key	BILL	Payment_type	Khoá ngoại tham chiếu từ BILL đến PAYMENT
FK_BILL_NGUONDL	Foreign key	BILL	NguonDL	Khoá ngoại tham chiếu từ BILL đến NGUONDL
FK_BILL_TRANGTHAI	Foreign key	BILL	TrangThai	Khoá ngoại tham chiếu từ BILL đến TRANGTHAI
PK_NGUONDL	Primary key	NGUONDL	ID	Khoá chính của bảng NGUONDL
PK_TRANGTHAI	Primary key	TRANGTHAI	ID	Khoá chính của bảng TRANGTHAI
PK_VENDOR	Primary key	VENDOR	ID	Khoá chính của bảng VENDOR
FK_VENDOR_NGUONDL	Foreign key	VENDOR	NguonDL	Khoá ngoại tham chiếu từ VENDOR đến NGUONDL
FK_VENDOR_TRANGTHAI	Foreign key	VENDOR	TrangThai	Khoá ngoại tham chiếu từ VENDOR đến TRANGTHAI

PK_PAYMENT	Primary key	PAYMENT	ID	Khoá chính của bảng PAYMENT
FK_PAYMENT_NGUONDL	Foreign key	PAYMENT	NguonDL	Khoá ngoại tham chiếu từ PAYMENT đến NGUONDL
FK_PAYMENT_TRANGTHAI	Foreign key	PAYMENT	TrangThai	Khoá ngoại tham chiếu từ PAYMENT đến TRANGTHAI
PK_PICKUP	Primary key	PICK_UP	ID	Khoá chính của bảng PICK_UP
FK_PICKUP_NGUONDL	Foreign key	PICK_UP	NguonDL	Khoá ngoại tham chiếu từ PICK_UP đến NGUONDL
FK_PICKUP_TRANGTHAI	Foreign key	PICK_UP	TrangThai	Khoá ngoại tham chiếu từ PICK_UP đến TRANGTHAI
PK_DROPOFF	Primary key	DROP_OFF	ID	Khoá chính của bảng DROP_OFF
FK_DROPOFF_NGUONDL	Foreign key	DROP_OFF	NguonDL	Khoá ngoại tham chiếu từ DROP_OFF đến NGUONDL
FK_DROPOFF_TRANGTHAI	Foreign key	DROP_OFF	TrangThai	Khoá ngoại tham chiếu từ DROP_OFF đến TRANGTHAI
PK_NTA	Primary key	NTA	ID	Khoá chính của bảng NTA
FK_NTA_NGUONDL	Foreign key	NTA	NguonDL	Khoá ngoại tham chiếu từ NTA đến NGUONDL
FK_NTA_TRANGTHAI	Foreign key	NTA	TrangThai	Khoá ngoại tham chiếu từ NTA đến TRANGTHAI
PK_BORO	Primary key	BORO	ID	Khoá chính của bảng BORO
FK_BORO_NGUONDL	Foreign key	BORO	NguonDL	Khoá ngoại tham chiếu từ BORO đến NGUONDL
FK_BORO_TRANGTHAI	Foreign key	BORO	TrangThai	Khoá ngoại tham chiếu từ BORO đến TRANGTHAI

### 2.3. DATA STORE LOGICAL MODEL

# NDS



## 3. DDS

### 3.1. COLUMNS

Table	Column	Key Type	Data Type	Is Nulla ble	Default Values	Description
BILL	Bill_ID	PK	Int	N		Mã hóa đơn của một chuyến đi
	PU_ID	FK	Int	N		Khu vực mà đồng hồ tính tiền của taxi bắt đầu hoạt động
	DO_ID	FK	Int	N		Khu vực mà đồng hồ tính tiền của taxi ngừng hoạt động
	Payment_type	FK	Int	N		Phương thức thanh toán.
	PU_Co	FK	Int	N		Mã địa chỉ của địa điểm bắt đầu chuyến đi.

	PU_Hour	FK	Int	N		Mã giờ trong ngày bắt đầu chuyến đi.
	PU_Month	FK	Int	N		Mã tháng bắt đầu chuyến đi.
	PU_Year	FK	Int	N		Mã năm bắt đầu chuyến đi.
	DO_Co	FK	Int	N		Mã địa chỉ của địa điểm kết thúc chuyến đi.
	DO_Hour	FK	Int	N		Mã giờ trong ngày kết thúc chuyến đi.
	DO_Month	FK	Int	N		Mã tháng kết thúc chuyến đi.
	DO_Year	FK	Int	N		Mã năm kết thúc chuyến đi.
	Extra		Float	N		Phí phụ thu.
	Mta_tax		Float	N		Thuế.
	Imp_Sur		Float	N		Phí phát sinh cải thiện chuyến đi.
	fare_amount		Float	N		Số tiền gốc của chuyến đi tính theo đồng hồ
	tip_amount		Float	N		Tiền bo
	tolls_amount		Float	Y		Tiền đi qua trạm thu phí (phí cầu đường)
	Pass_count		Int	N		Số lượng hành khách trên chuyến đi.
	TripDistance		Float	N		Khoảng cách của chuyến đi.
	total_amount		Float	N		Tổng chi phí hành khách phải trả (không bao gồm tiền bo)
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu.
	NguonDL		Int	N		Mã nguồn dữ liệu.
	TrangThai		Int	N		Mã trạng thái.
DIM_CO ORDINA TES	Co_ID	PK	Int	N		Mã địa chỉ.
	Longitude		Varchar(10 0)	N		Kinh độ địa chỉ.
	Latitude		Varchar(10 0)	N		Vĩ độ địa chỉ.
	StreetID	FK	Int	N		Mã đường.
	BorolD	FK	Int	N		Mã quận.
	CensusBlock		Varchar(10 0)	N		Mã census block.

	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu.
	NguonDL		Int	N		Mã nguồn dữ liệu.
	TrangThai		Int	N		Mã trạng thái.
DIM_PAYMENT	PaymentID	PK	Int	N		Mã phương thức thanh toán.
	Payment_type		Int	N	Từ 1 đến 6	Phương thức thanh toán. Với quy định: 1 = Thẻ tín dụng 2 = Tiền mặt 3 = Không tính phí 4 = Tranh chấp 5 = Không xác định 6 = Chuyển đi bị hoãn
	PaymentName		Varchr(100)	N		Tên của phương thức thanh toán.
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu.
	NguonDL		Int	N		Mã nguồn dữ liệu.
	TrangThai		Int	N		Mã trạng thái.
	BoroID	PK	Int	N		Mã quận.
DIM_BORO	BoroCode		Varchar(100)	N	Từ 1 đến 5	Mã định vị quận. Code gồm số từ 1 đến 5 tương ứng cho 5 quận của bang New York, với: 1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, 5: Staten Island
	BoroName		Varchar(100)	N		Tên các quận trong New York
	NTACode		Varchar(10)	N		Mã đánh dấu các khu dân cư của quận trong bang New York
	NTAName		Varchar(50)	N		Tên các khu dân cư của quận trong bang New York
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu.
	NguonDL		Int	N		Mã nguồn dữ liệu.
	TrangThai		Int	N		Mã trạng thái.
DIM_STREET	Street_ID	PK	Int	N		Mã đường.
	Street		Varchar(100)	N		Tên đường.
	Co_ID	FK	Int	N		Mã địa chỉ.
	Number		Varchar(100)	N		Số nhà.

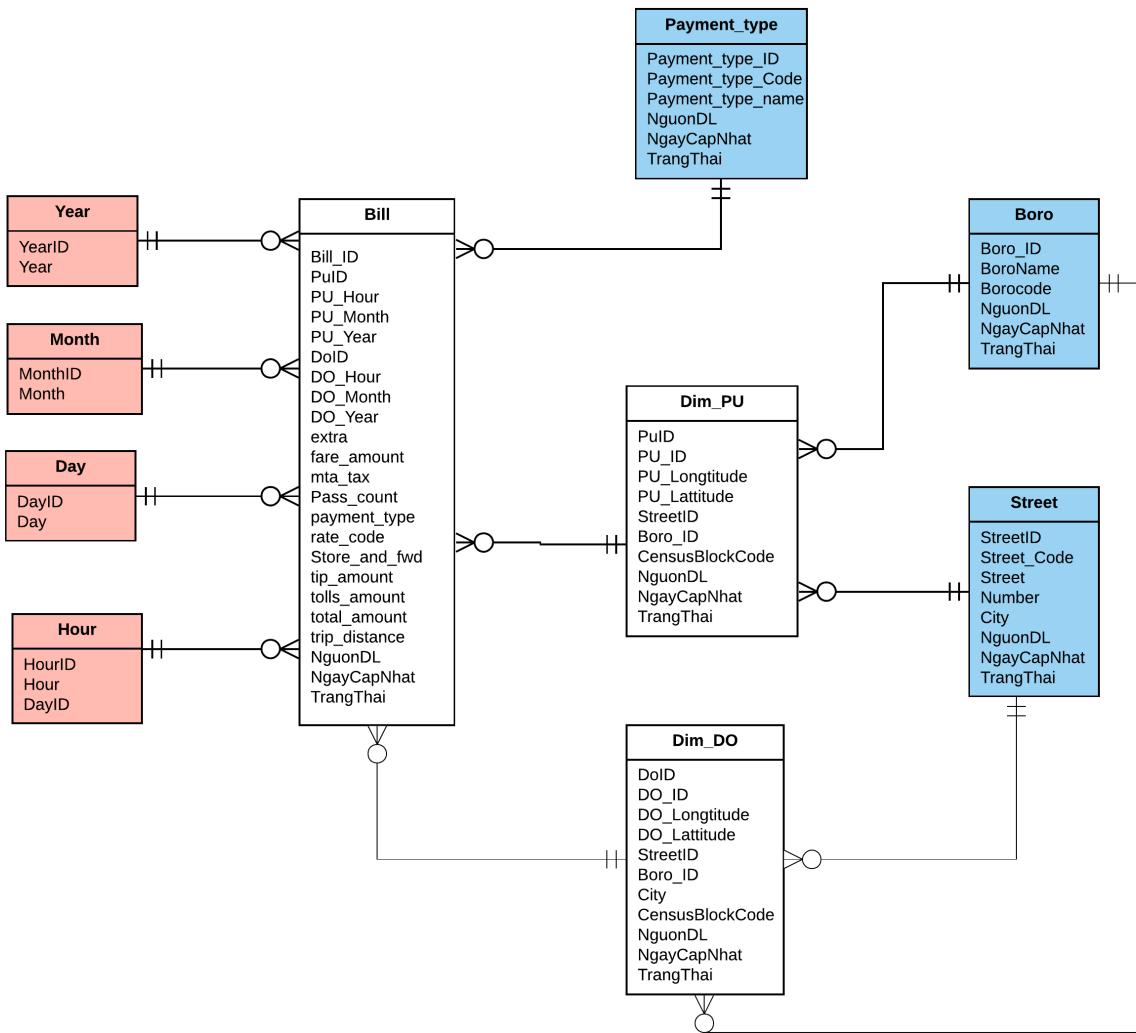
	NgayCapNhat		Datetime	N		Ngày cập nhật dữ liệu.
	NguonDL		Int	N		Mã nguồn dữ liệu.
	TrangThai		Int	N		Mã trạng thái.
DIM_HO UR	HourID	PK	Int	N		Mã giờ
	HHour		Int	N		Giờ.
DIM_DA Y	DayID	PK	Int	N		Mã ngày
	DDay		Int	N		Ngày
DIM_M ONTH	MonthID	PK	Int	N		Mã tháng
	Monthh		Int	N		Tháng
DIM_YE AR	YearID	PK	Int	N		Mã năm
	Yearr		Int	N		Năm

### 3.2. CONSTRAINT

Constraint Name	Constraint Type	Table	Column	Description
PK_BILL	Primary key	BILL	Bill_ID	Khoá chính của bảng BILL
FK_BILL_PAYMENT	Foreign key	BILL	Payment_ty pe	Khoá ngoại tham chiếu từ BILL đến DIM_PAYMENT
FK_BILL_PUCORD INATES	Foreign key	BILL	PU_Co	Khoá ngoại tham chiếu từ BILL đến DIM_COORDINATES
FK_BILL_PUHOUR	Foreign key	BILL	PU_Hour	Khoá ngoại tham chiếu từ BILL đến DIM_HOUR
FK_BILL_PUMONT H	Foreign key	BILL	PU_Month	Khoá ngoại tham chiếu từ BILL đến DIM_MONTH
FK_BILL_PUYEAR	Foreign key	BILL	PU_Year	Khoá ngoại tham chiếu từ BILL đến DIM_YEAR
FK_BILL_DOCOOR DINATES	Foreign key	BILL	DO_Co	Khoá ngoại tham chiếu từ BILL đến DIM_COORDINATES
FK_BILL_DOHOUR	Foreign key	BILL	DO_Hour	Khoá ngoại tham chiếu từ BILL đến DIM_HOUR
FK_BILL_DOMONT H	Foreign key	BILL	DO_Month	Khoá ngoại tham chiếu từ BILL đến DIM_MONTH
FK_BILL_DOYEAR	Foreign key	BILL	DO_Year	Khoá ngoại tham chiếu từ BILL đến DIM_YEAR
PK_BORO	Primary key	DIM_BORO	BoroID	Khoá chính của bảng DIM_BORO
PK_STREET	Primary key	DIM_STREET	StreetID	Khoá chính của bảng DIM_STREET
PK_PAYMENT	Primary key	DIM_PAYME NT	PaymentID	Khoá chính của bảng DIM_PAYMENT
PK_HOUR	Primary key	DIM_HOUR	HourID	Khoá chính của bảng DIM_HOUR

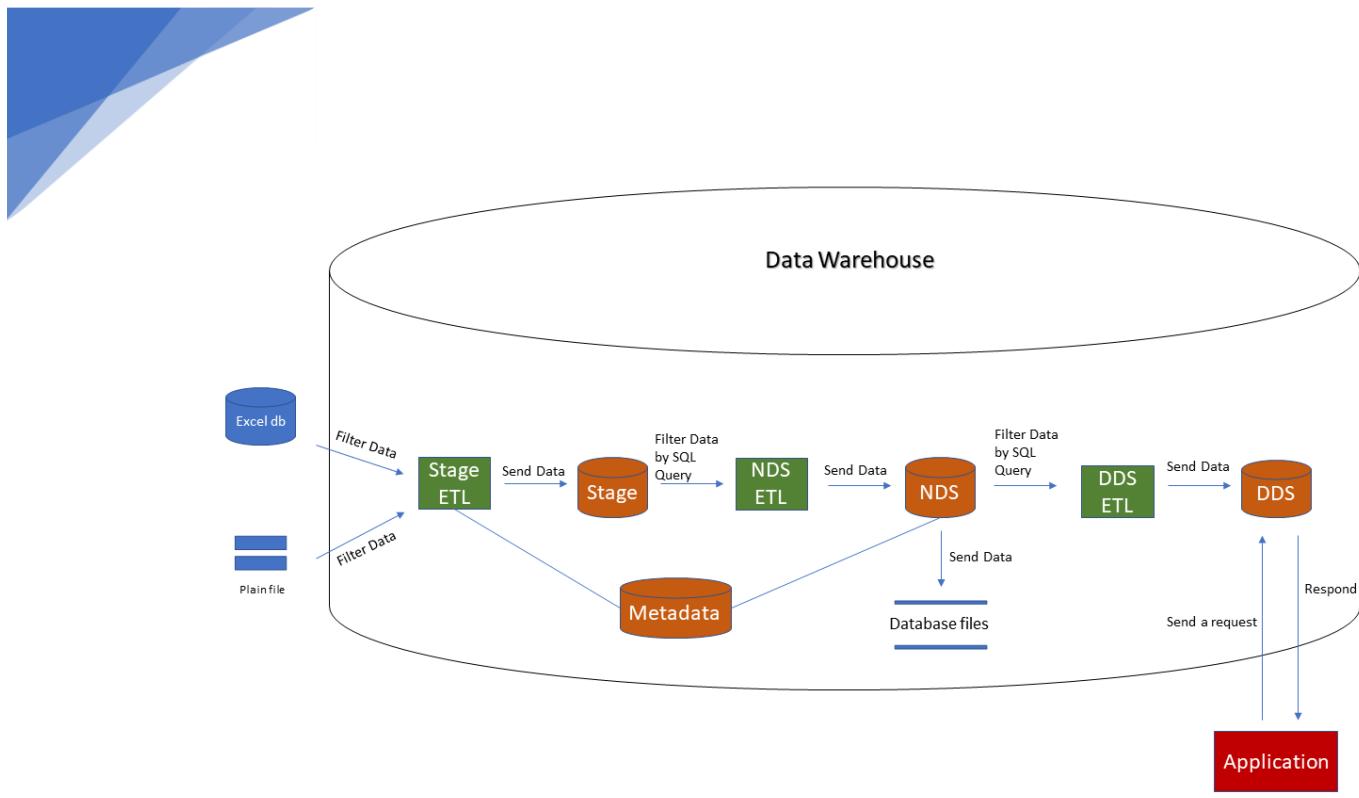
FK_HOUR_DAY	Foreign key	DIM_HOUR	DayID	Khoá ngoại tham chiếu từ DIM_HOUR đến DIM_DAY
PK_DAY	Primary key	DIM_DAY	DayID	Khoá chính của bảng DIM_DAY
PK_MONTH	Primary key	DIM_MONTH	MonthID	Khoá chính của bảng DIM_MONTH
PK_YEAR	Primary key	DIM_DATETIME_YEAR	YearID	Khoá chính của bảng DIM_YEAR
PK_CO	Primary key	DIM_COORDINATES	Co_ID	Khoá chính của bảng DIM_COORDINATES
FK_CO_BORO	Foreign key	DIM_COORDINATES	BoroID	Khoá ngoại tham chiếu từ DIM_COORDINATES đến DIM_BORO
FK_CO_STREET	Foreign key	DIM_COORDINATES	StreetID	Khoá ngoại tham chiếu từ DIM_COORDINATES đến DIM_STREET

### 3.3. DATASTORE LOGICAL MODEL



#### 4. Thiết kế Data Flow

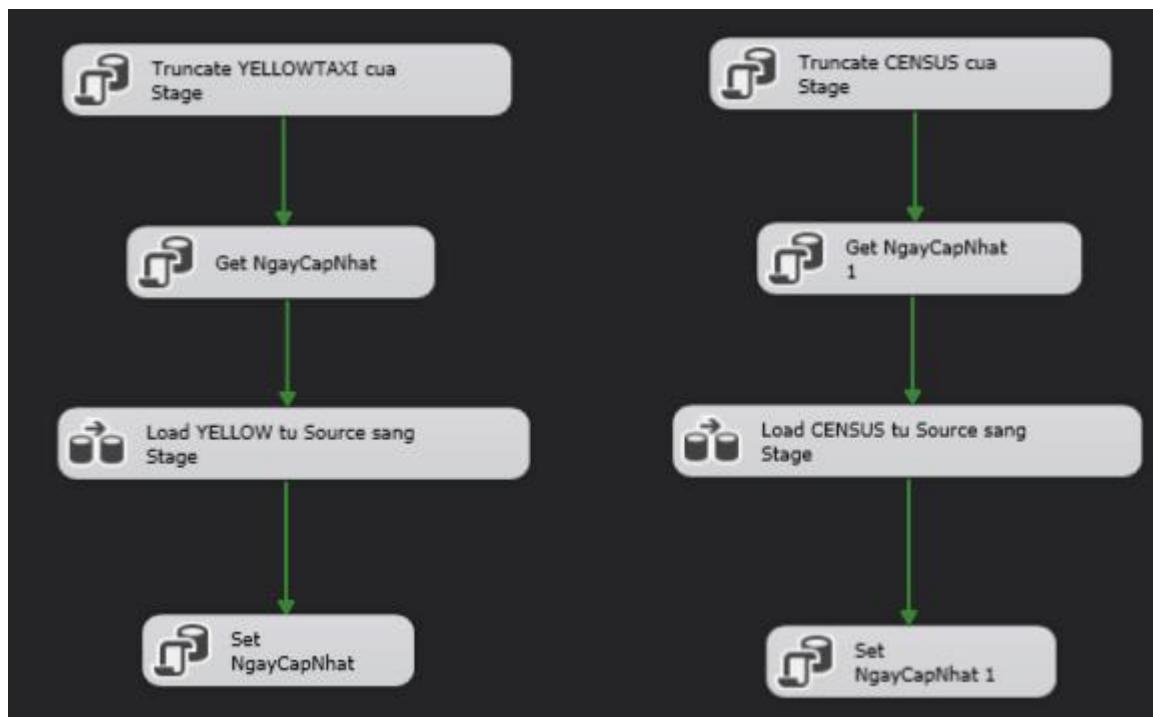
❖ Data flow architecture theo NDS và DDS



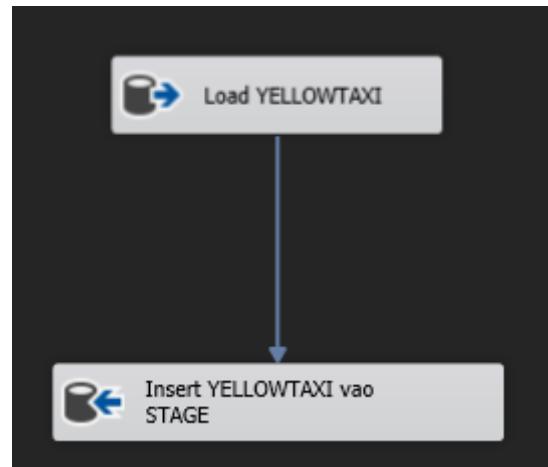
## VIII. ETL

### 1. Source to Stage

Ta thiết kế dòng Data flow như hình dưới đây để nạp các dữ liệu từ nguồn Yellow Taxi và Census Block vào Stage:



Ở phần *Load YELLOW tu Source sang Stage* ta có một luồng như hình dưới:



Ta tạo database BI\_YellowTaxi và thực hiện thêm các cột NgayTao, NgayCapNhat:

```

CREATE DATABASE BI_YellowTaxi
GO

USE BI_YellowTaxi
GO

--Them cot cho YellowTaxi
ALTER TABLE YL2015
    ADD [NgayTao] datetime NULL,
        [NgayCapNhat] datetime NULL,
        [TrangThai] int NULL;
GO

UPDATE YL2015
SET [TrangThai] = 1;
GO

UPDATE YL2015
SET [NgayTao] = '20201213 22:12:11.853',
    [NgayCapNhat] = '20201213 22:12:11.853';

--Them cot cho CensusBlock
ALTER TABLE CensusBlock
    ADD [NgayTao] datetime NULL,
        [NgayCapNhat] datetime NULL,
        [TrangThai] int NULL;
GO

```

Sau đó ta thực hiện kết nối với database và nhập câu lệnh SQL như dưới đây:

OLE DB connection manager:

LAPTOP-O59F8M98.BI\_YellowTaxi

New...

Data access mode:

SQL command

SQL command text:

```
SELECT *
FROM YL2015
WHERE (NgayTao < GETDATE()) AND (NgayTao > ?) OR
(NgayCapNhat < GETDATE()) AND (NgayCapNhat > ?)
```

Parameters...

Build Query...

Và ta thực hiện nạp dữ liệu vào bảng YELLOWTAXI của Stage

OLE DB connection manager:

LAPTOP-O59F8M98.BI\_STAGE

New...

Data access mode:

Table or view - fast load

Name of the table or the view:

[dbo].[YELLOWTAXI]

New...

Ta tạo một Truncate và thao tác như hình dưới:

<b>General</b>	
Name	<b>Truncate YELLOWTAXI cua Stage</b>
Description	<b>Execute SQL Task</b>
<b>Options</b>	
TimeOut	<b>0</b>
CodePage	<b>1252</b>
TypeConversionMode	<b>Allowed</b>
<b>Result Set</b>	
ResultSet	<b>None</b>
<b>SQL Statement</b>	
ConnectionType	<b>OLE DB</b>
Connection	<b>LAPTOP-O59F8M98.BI_STAGE</b>
SQLSourceType	<b>Direct input</b>
SQLStatement	<b>truncate table YELLOWTAXI</b>
IsQueryStoredProcedure	<b>False</b>
BypassPrepare	<b>True</b>
<b>Name</b>	

Ta tạo GET NgayCapNhat và thao tác như hình dưới:

<b>General</b>	
Name	<b>Get NgayCapNhat</b>
Description	<b>Execute SQL Task</b>
<b>Options</b>	
TimeOut	<b>0</b>
CodePage	<b>1252</b>
TypeConversionMode	<b>Allowed</b>
<b>Result Set</b>	
ResultSet	<b>Single row</b>
<b>SQL Statement</b>	
ConnectionType	<b>OLE DB</b>
Connection	<b>LAPTOP-O59F8M98.BI_METADATA</b>
SQLSourceType	<b>Direct input</b>
SQLStatement	<b>SELECT NgayCapNhat FROM Data_Flow</b>
IsQueryStoredProcedure	<b>False</b>
BypassPrepare	<b>True</b>
<b>Name</b>	

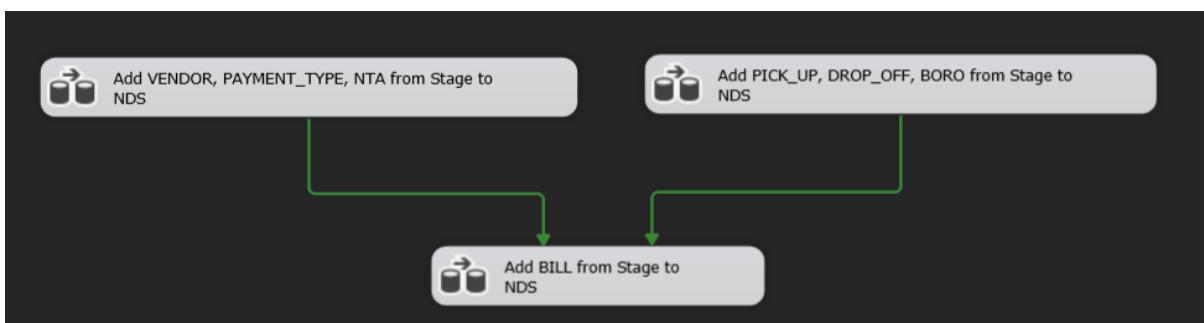
Ta tạo SET NgayCapNhat và thao tác như hình dưới:

<b>General</b>	
Name	<b>Set NgayCapNhat</b>
Description	<b>Execute SQL Task</b>
<b>Options</b>	
TimeOut	<b>0</b>
CodePage	<b>1252</b>
TypeConversionMode	<b>Allowed</b>
<b>Result Set</b>	
ResultSet	<b>None</b>
<b>SQL Statement</b>	
ConnectionType	<b>OLE DB</b>
Connection	<b>LAPTOP-O59F8M98.BI_METADATA</b>
SQLSourceType	<b>Direct input</b>
SQLStatement	<b>update Data_Flow set NgayCapNhat = g</b>
IsQueryStoredProcedure	<b>False</b>
BypassPrepare	<b>True</b>

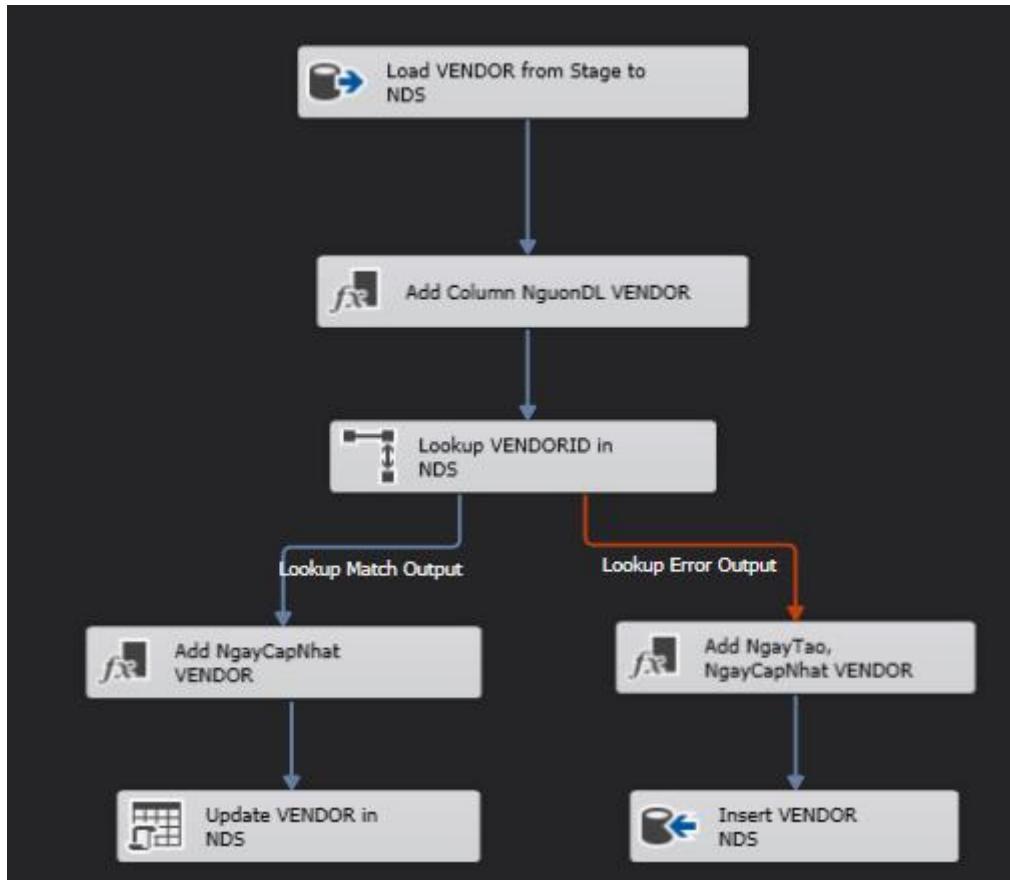
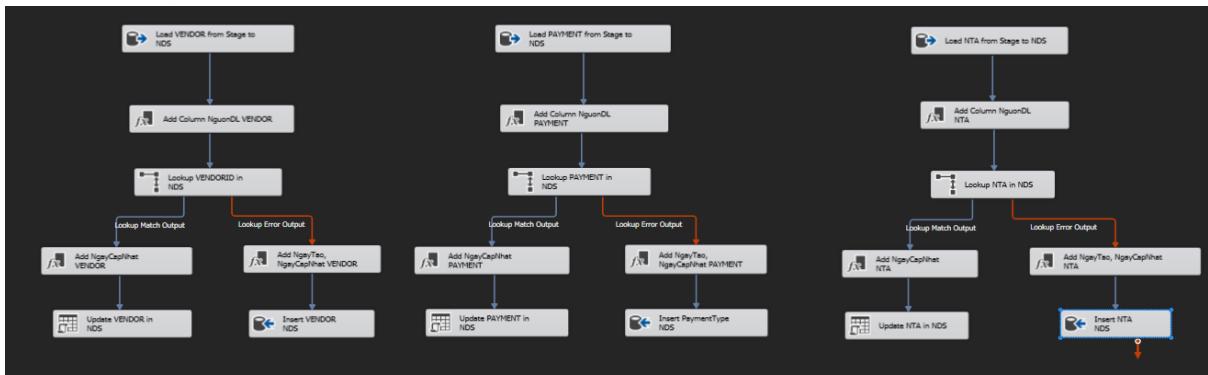
Ta làm tương tự trên với nạp dữ liệu từ nguồn Census Block và Stage

## 2. Stage to NDS

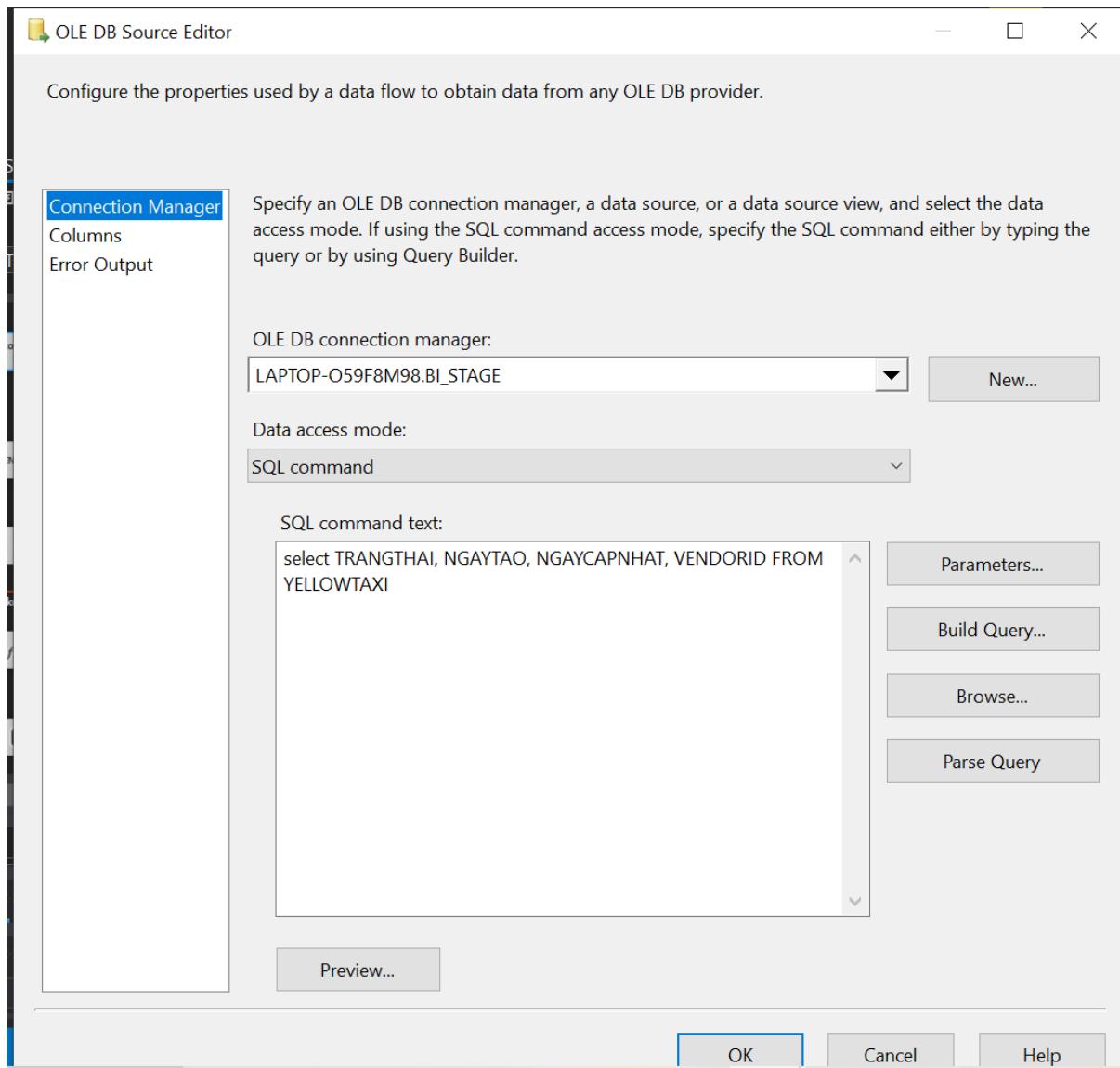
Ta tạo một Data Flow như hình dưới đây:



Ở phần *Add VENNDOR, PAYMENT, NTA from Stage to NDS* ta tạo các luồng như sau:



Đầu tiên, ta tải dữ liệu của bảng Vendor từ Stage:



Thêm cột NguonDL cho bảng Vendor:

Description				
Derived Column Name	Derived Column	Expression	Data Type	Length
NguonDL	<add as new column>	1	four-byte signed integer	

Tiếp theo, ta lookup xem VendorID đã tồn tại trong NDS chưa

Nếu đã tồn tại, ta thêm cột NgayCapNhat:

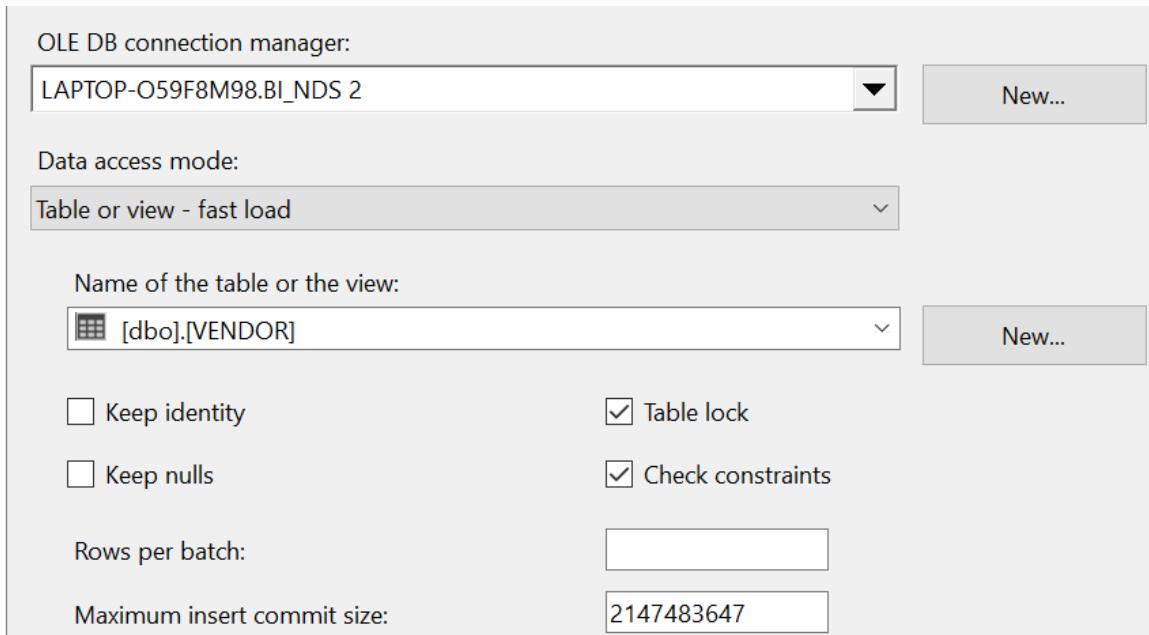
Và update lại dữ liệu bảng Vendor với VendorID tương ứng:

Input Column	Destination Column
NgayCapNhat_NEW	Param_0
TRANGTHAI	Param_1
VENDORID	Param_2
NguonDL	Param_3

Nếu chưa tồn tại, ta thêm cột NgayCapNhat và cột NgayTao:

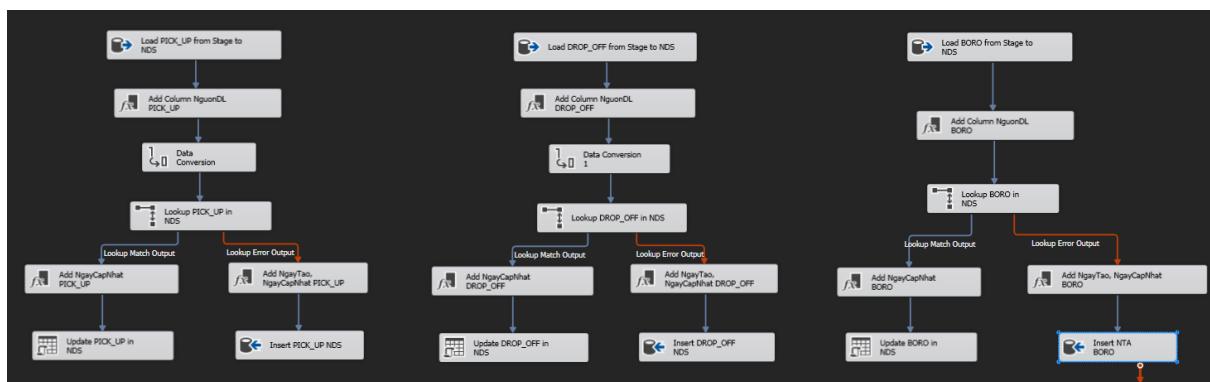
Derived Column Name	Derived Column	Expression	Data Type	Length
NgayCapNhat_NEW	<add as new column>	GETDATE()	database timestamp	[...]
NgayTao_NEW	<add as new column>	GETDATE()	database timestamp	[...]

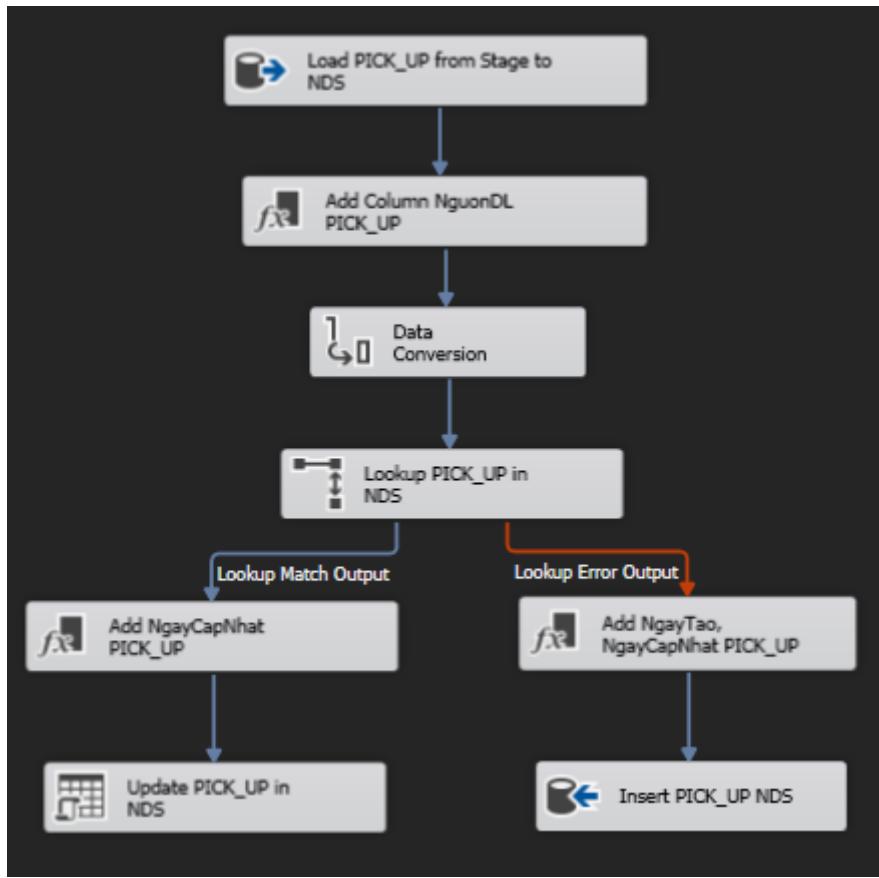
Sau đó, ta nạp dữ liệu mới vào bảng Vendor của NDS:



Ta thực hiện tương tự ở các luồng khác.

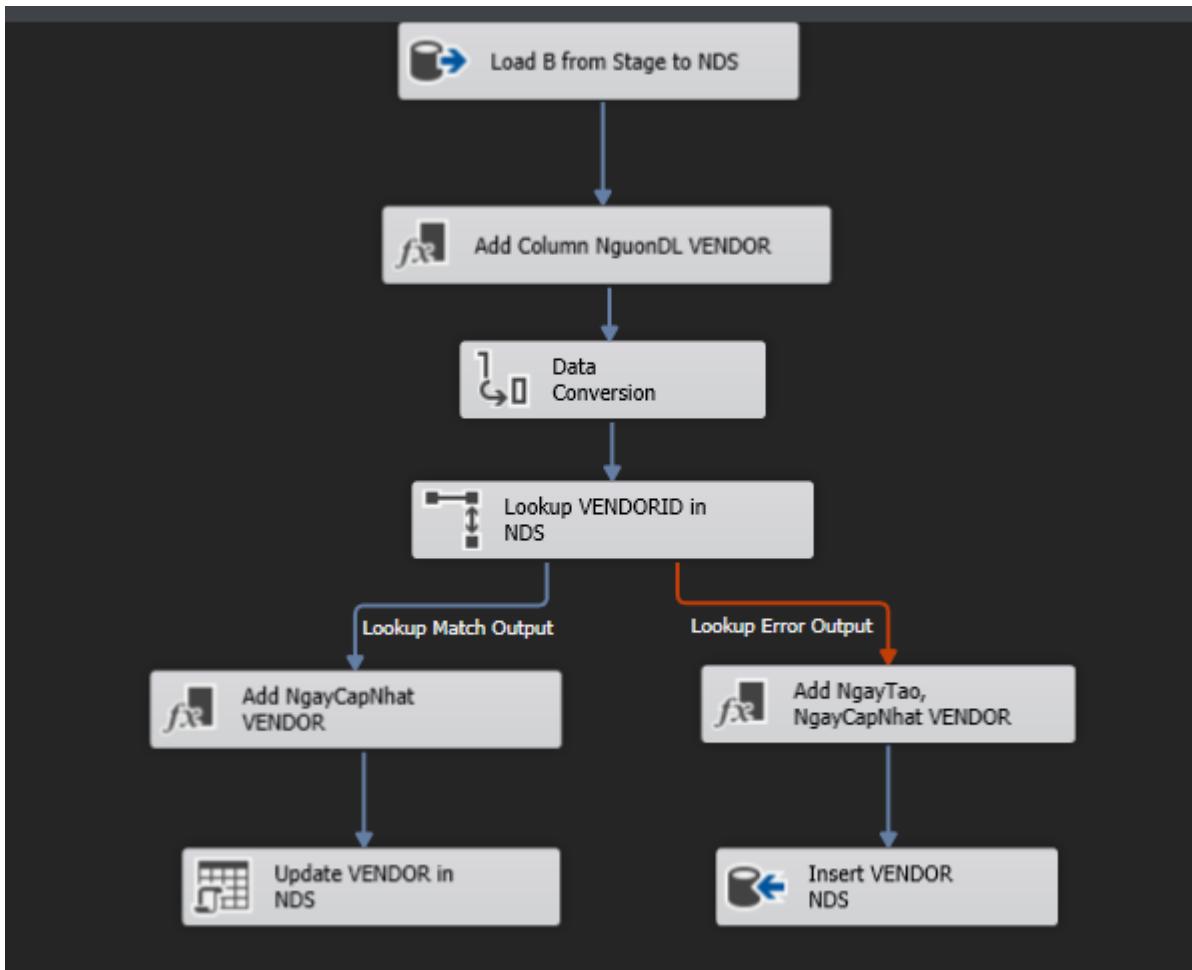
Ở phần *Add PICK\_UP, DROP\_OFF, BORO from Stage to NDS* ta tạo các luồng như sau:





Ở đây, ta cũng thực hiện tương tự *Add VENDOR, PAYMENT, NTA from Stage to NDS*

Ở phần *Add BILL from Stage to NDS* ta tạo luồng như sau:



Trước tiên, ta tải dữ liệu bảng Bill từ Stage:

OLE DB connection manager:

LAPTOP-O59F8M98.BI\_STAGE

New...

Data access mode:

SQL command

SQL command text:

```
select TRANGTHAI, NGAYTAO, NGAYCAPNHAT, PUDatetime,
VendorID,
PassCount,
TripDistance,
RateCode,
StoreAndFwd,
PaymentType,
FareAmount,
Extra,
MtaTax,
TipAmount,
TollsAmount,
TotalAmount,
PULongitude,
```

Parameters...

Build Query...

Browse...

Parse Query

Tiếp theo, ta thêm cột NguonDL cho bảng Bill:

Derived Column Name	Derived Column	Expression	Data Type	Length
NguonDL	<add as new column>	1	four-byte signed integer	

Sau đó, ta convert kiểu dữ liệu của tọa độ sang varchar(100):

Input Column	Output Alias	Data Type	Length	Precision	Scale	Code Page
PULongitude	Copy of PULongitude	string [DT_STR]	100			1252 (ANSI)
PULatitude	Copy of PULatitude	string [DT_STR]	100			1252 (ANSI)

Tiếp theo, ta lookup các tọa độ và thời gian đón xe:

OLE DB connection manager:

LAPTOP-O59F8M98.BI\_NDS

New...

Use a table or a view:

New...

New...

Use results of an SQL query:

```
select PULongitude, PULatitude, PUDatetime, ID from
PICK_UP where NguonDL = 1
```

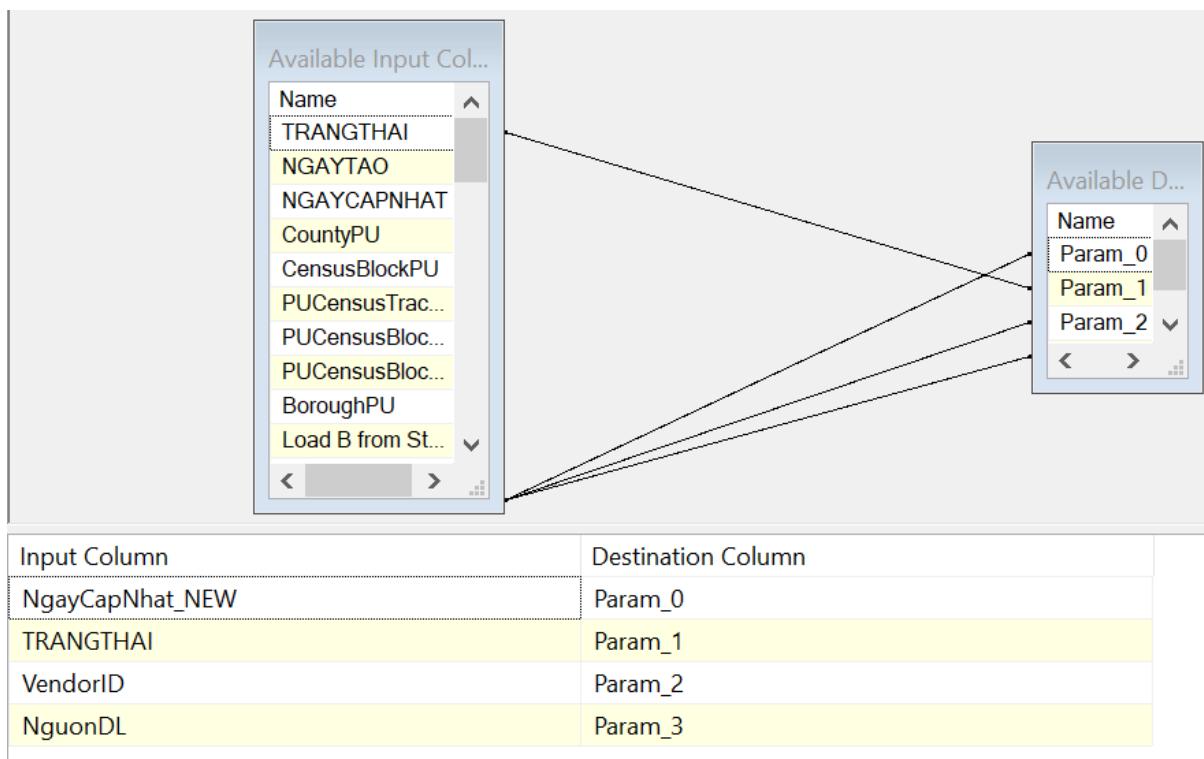
Build Query...

Browse...

Nếu đã tồn tại, ta thêm cột NgayCapNhat:

Derived Column Name	Derived Column	Expression	Data Type	Length
NgayCapNhat_NEW	<add as new column>	GETDATE()	database timestamp [...]	

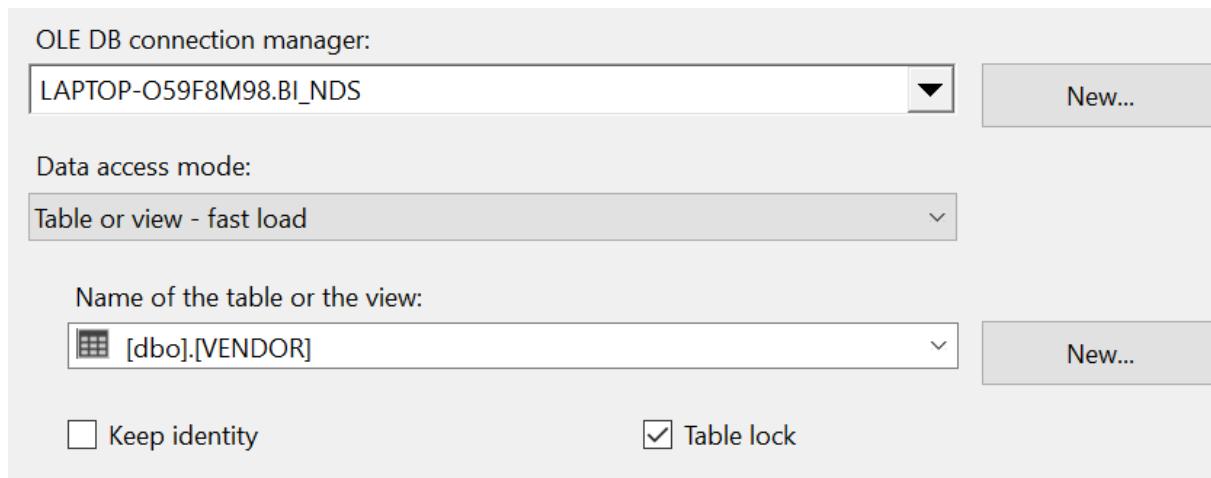
Và update lại dữ liệu tương ứng:



Nếu chưa tồn tại, ta thêm cột NgayCapNhat và NgayTao:

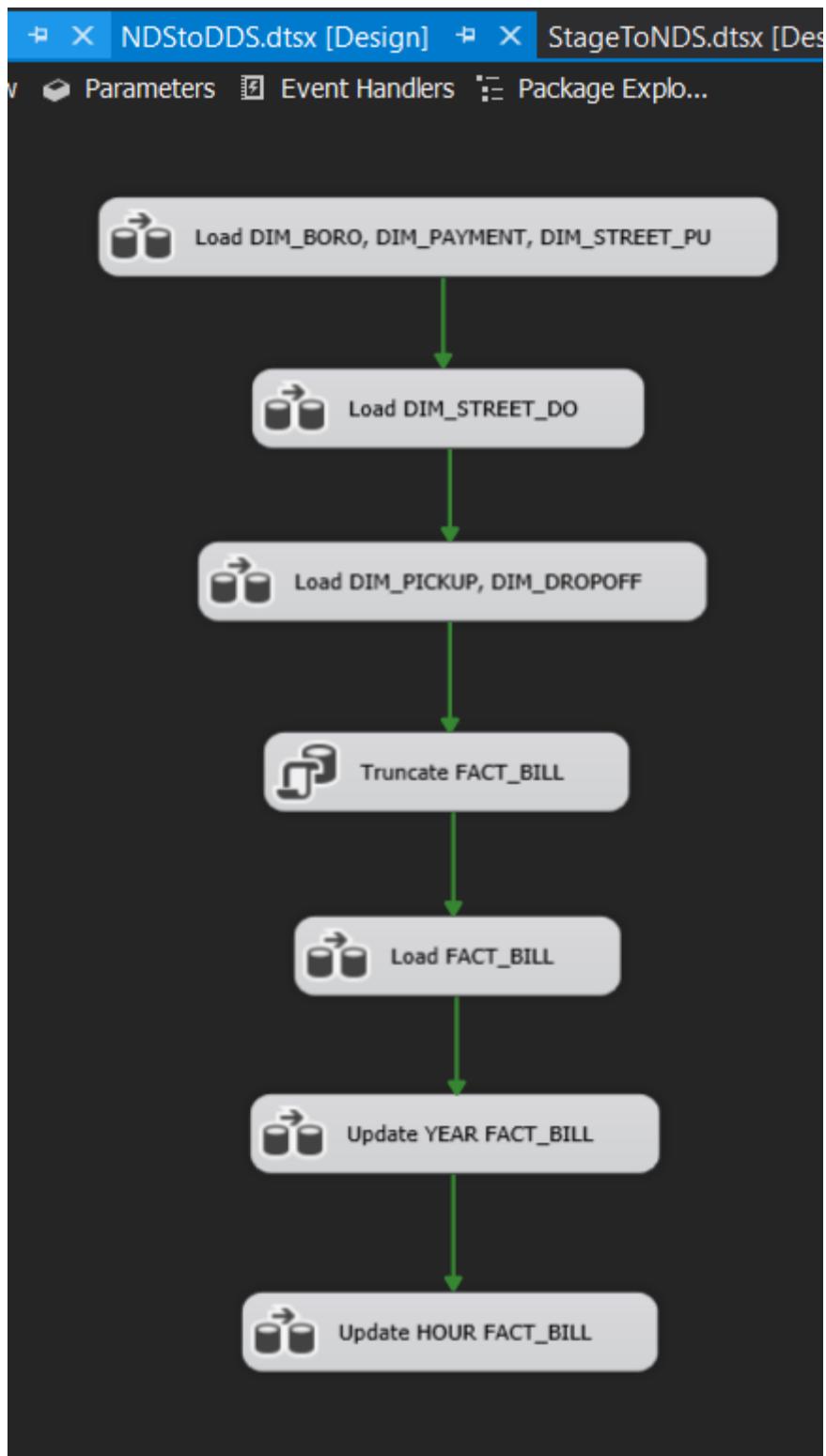
Derived Column Name	Derived Column	Expression	Data Type
NgayCapNhat_NEW	<add as new column>	GETDATE()	database timestamp [...]
NgayTao_NEW	<add as new column>	GETDATE()	database timestamp [...]

Và thực hiện nạp dữ liệu mới từ Stage vào bảng Bill của NDS

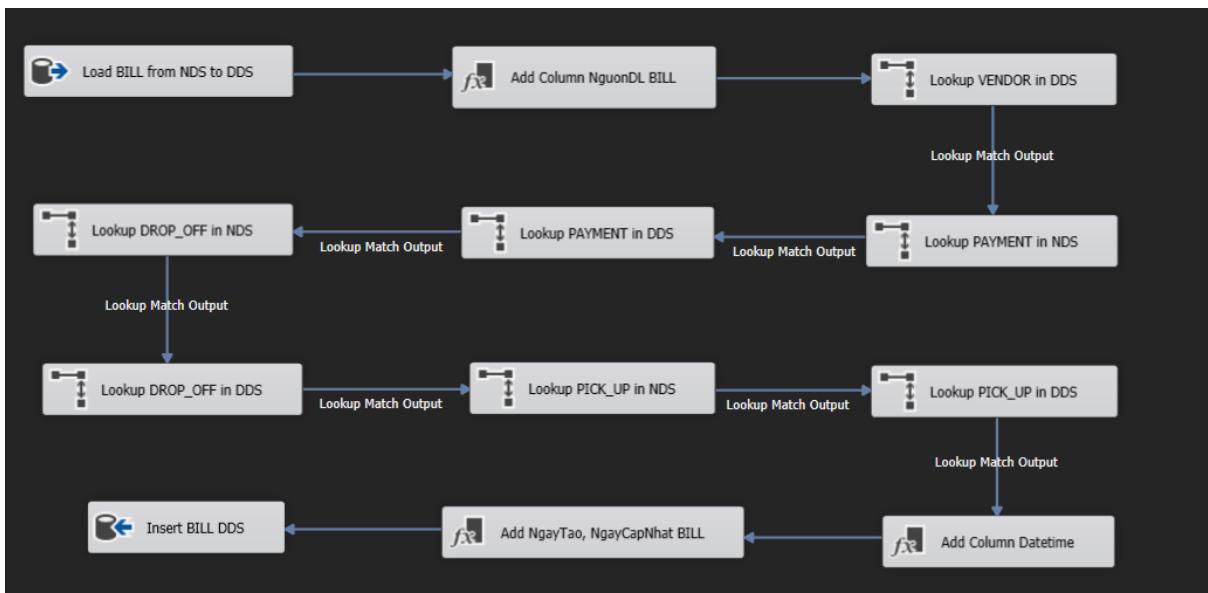


### 3. NDS to DDS

Ta tạo database BI\_DDS và lần lượt nạp dữ liệu vào các bảng Dim.



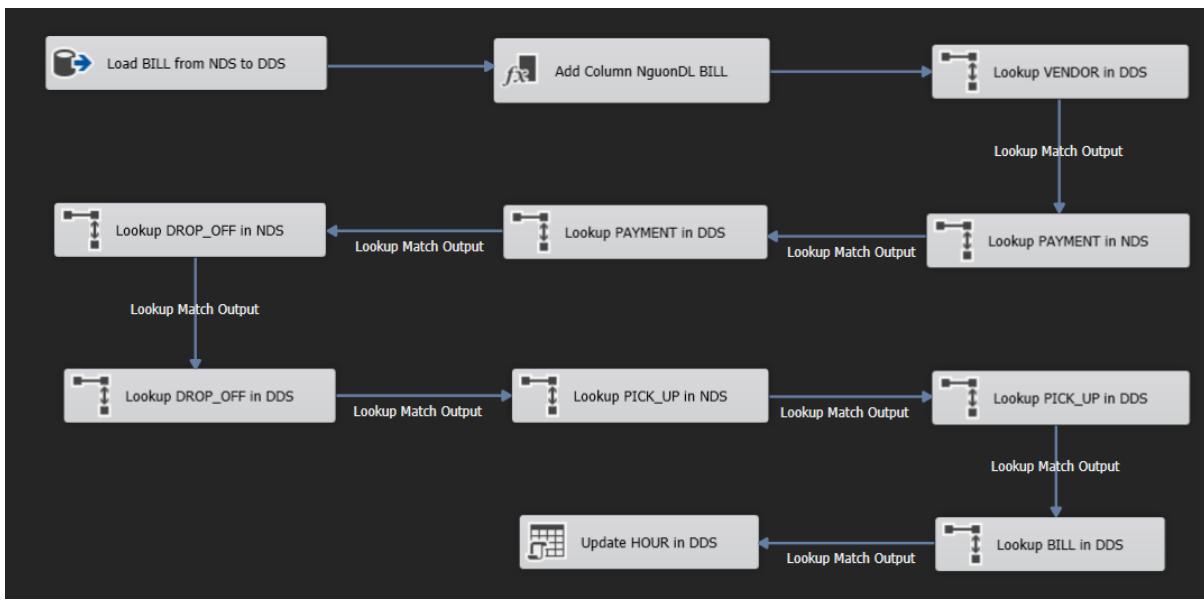
Sau khi đã hoàn tất nạp vào các bảng Dim, ta tiếp tục tạo data flow như hình dưới để nạp dữ liệu vào bảng Fact là bảng Bill.



Update nǎm:



Update giờ:



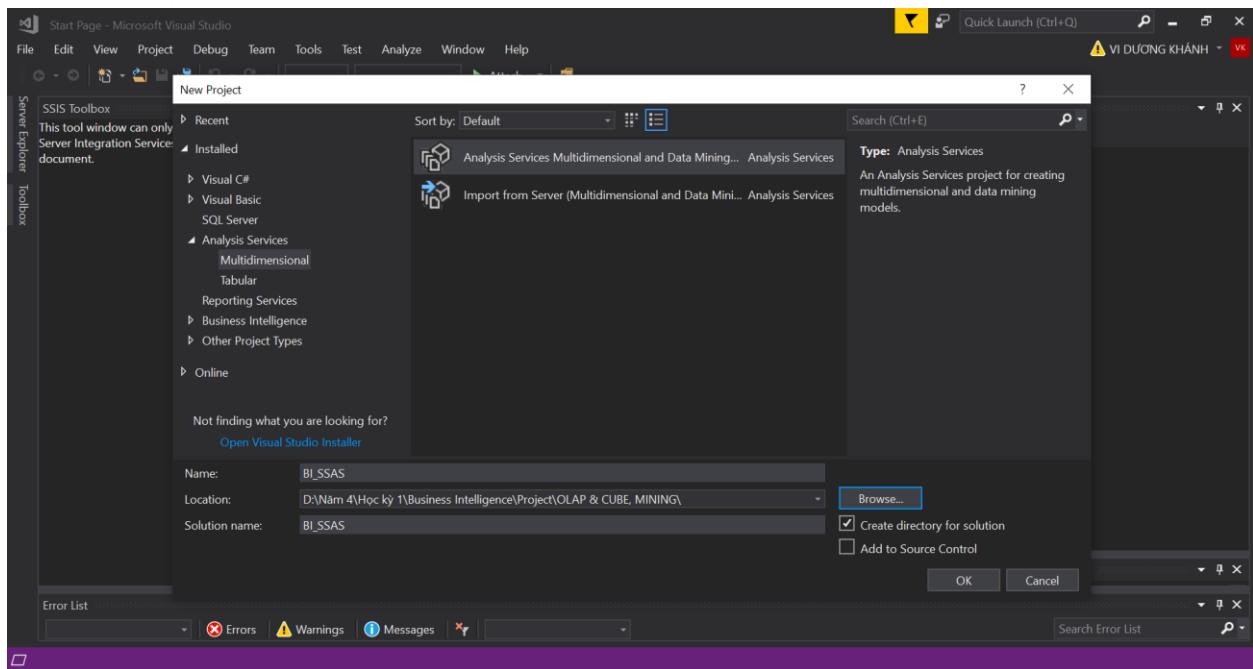
Ta lấy dữ liệu từ bảng Bill ở NDS. Sau đó, lần lượt lookup các thuộc tính khóa ngoại của bảng, nếu tất cả đều đã tồn tại trong các bảng Dim, ta thực hiện insert dữ liệu vào bảng Fact\_Bill. Sau khi chạy, ta kiểm tra lại ở SQL Server thì thấy đã có dữ liệu.

ID	PUID	DOID	passenger_count	trip_distance	rate_code	payment_type	fare_amount	surcharge	mta_tax	tip_amount	tolls_amount	total_amount	PU_Hour	PU_Day	PU_Month
1	1	1	1	0.8	1	2	5	0.5	0.5	0	0	6	1	9	1
2	2	2	2	4.8	1	2	15.5	0.5	0.5	0	0	16.5	1	9	1
3	3	3	3	1	1	4	3	0.5	0.5	0	0	4	1	9	1
4	4	4	4	2.1	1	2	9.5	0.5	0.5	0	0	10.5	1	9	1
5	5	5	5	1.8	1	2	8	0.5	0.5	0	0	9	1	9	1
6	6	6	6	0.9	1	2	5.5	0.5	0.5	0	0	6.5	1	9	1
7	7	7	7	2.3	1	2	8.5	0.5	0.5	0	0	9.5	1	9	1
8	8	8	8	18.1	2	2	52	0	0.5	0	5.33	57.83	1	9	1
9	9	9	9	11.8	1	2	34	0.5	0.5	0	0	35	1	9	1
10	10	10	10	1.8	1	2	7	0.5	0.5	0	0	8	5	9	1
11	11	11	11	0	1	2	2.5	0.5	0.5	0	0	3.5	5	9	1
12	12	12	12	1.1	1	2	6	0.5	0.5	0	0	7	5	9	1
13	13	13	13	2	1	2	8.5	0.5	0.5	0	0	9.5	6	9	1
14	14	14	14	0.9	1	2	5	0	0.5	0	0	5.5	6	9	1
15	15	15	15	4.1	1	4	14.5	0	0.5	0	0	15	6	9	1
16	16	16	16	0.9	1	2	6	0	0.5	0	0	6.5	6	9	1
17	17	17	17	2.7	1	2	9	0	0.5	0	0	9.5	6	9	1
18	18	18	18	1	1	2	10	0	0.5	0	0	10.5	6	11	1
19	19	19	19	4.7	1	2	16.5	0	0.5	0	0	17	6	11	1
20	20	20	20	1	0.9	1	2	5	0	0.5	0	5.5	6	11	1
21	21	21	21	1.8	1	2	8	0	0.5	0	0	8.5	13	11	1
22	22	22	22	0.3	1	2	4	0	0.5	0	0	4.5	13	11	1
23	23	23	23	0.5	1	2	4	0	0.5	0	0	4.5	13	11	1
24	24	24	24	1.3	1	2	8	0	0.5	0	0	8.5	13	11	1
25	25	25	25	0.6	1	2	5	0	0.5	0	0	5.5	13	11	1

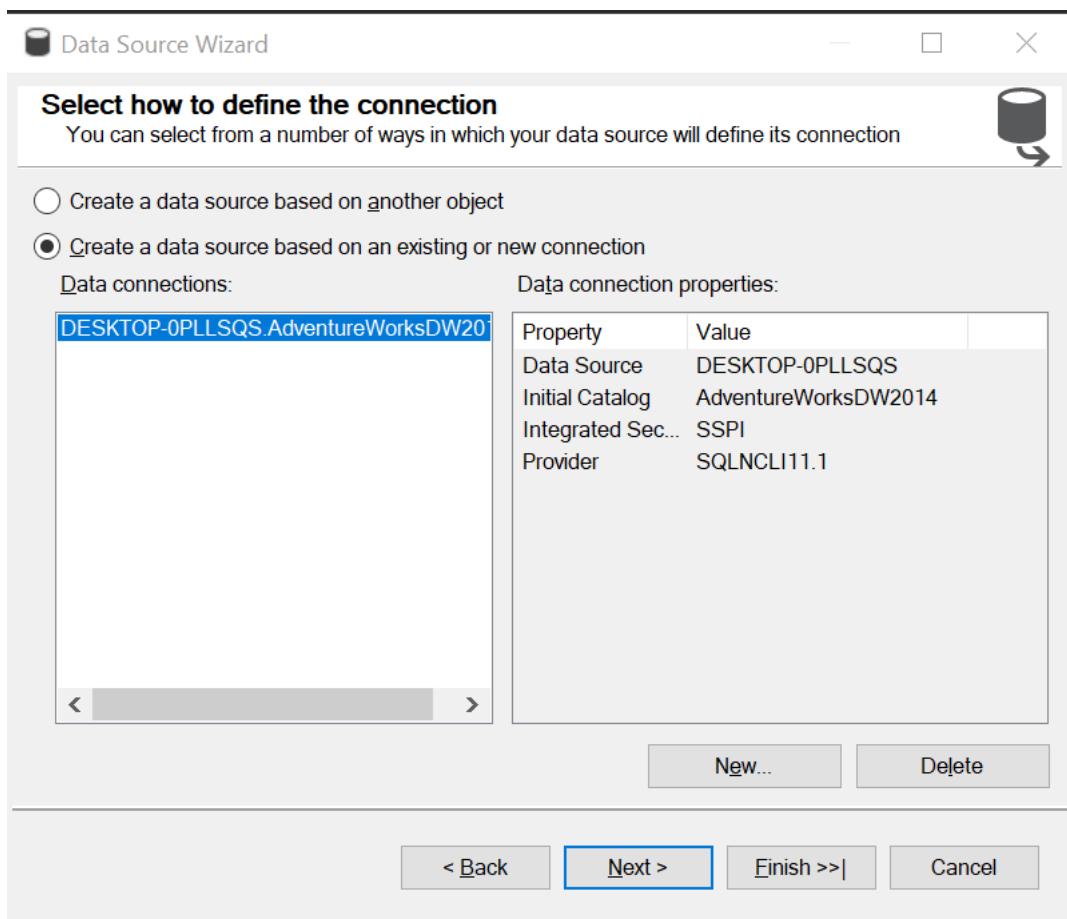
## IX. Khai thác dữ liệu

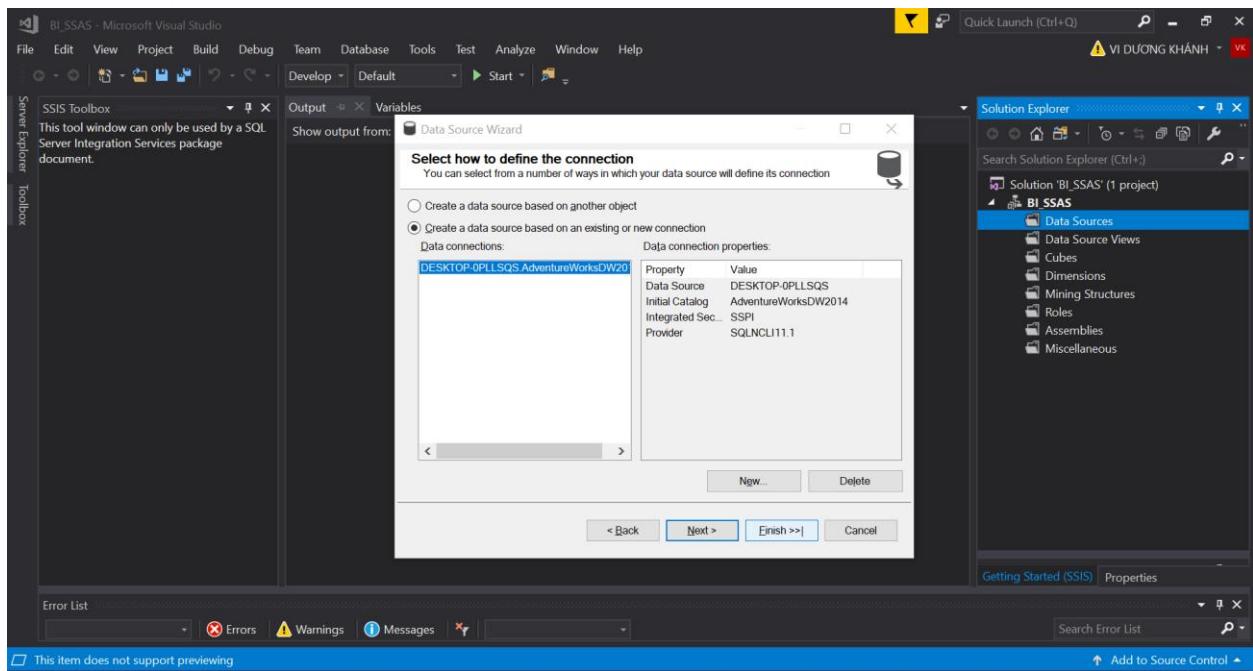
### 1. OLAP

Bước 1: Tạo project SSAS

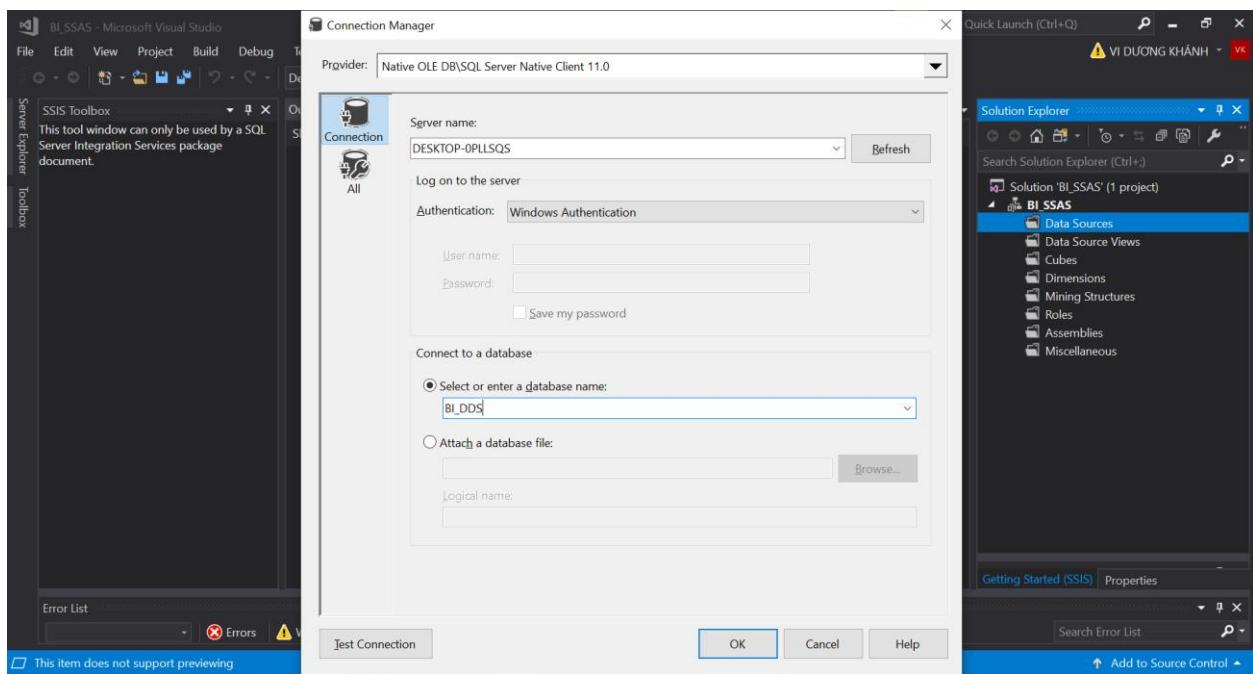


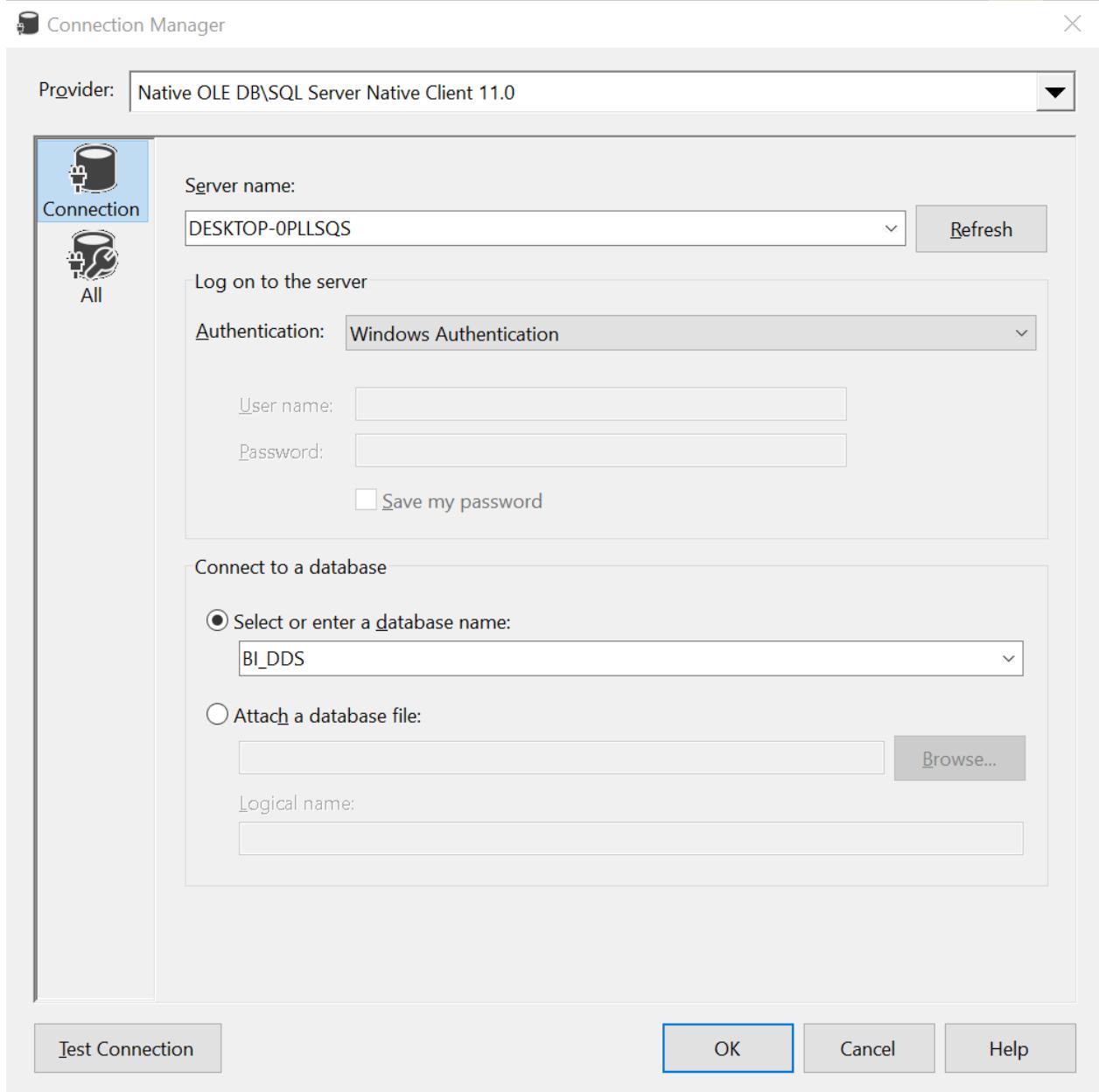
## Bước 2: Tạo Data Source



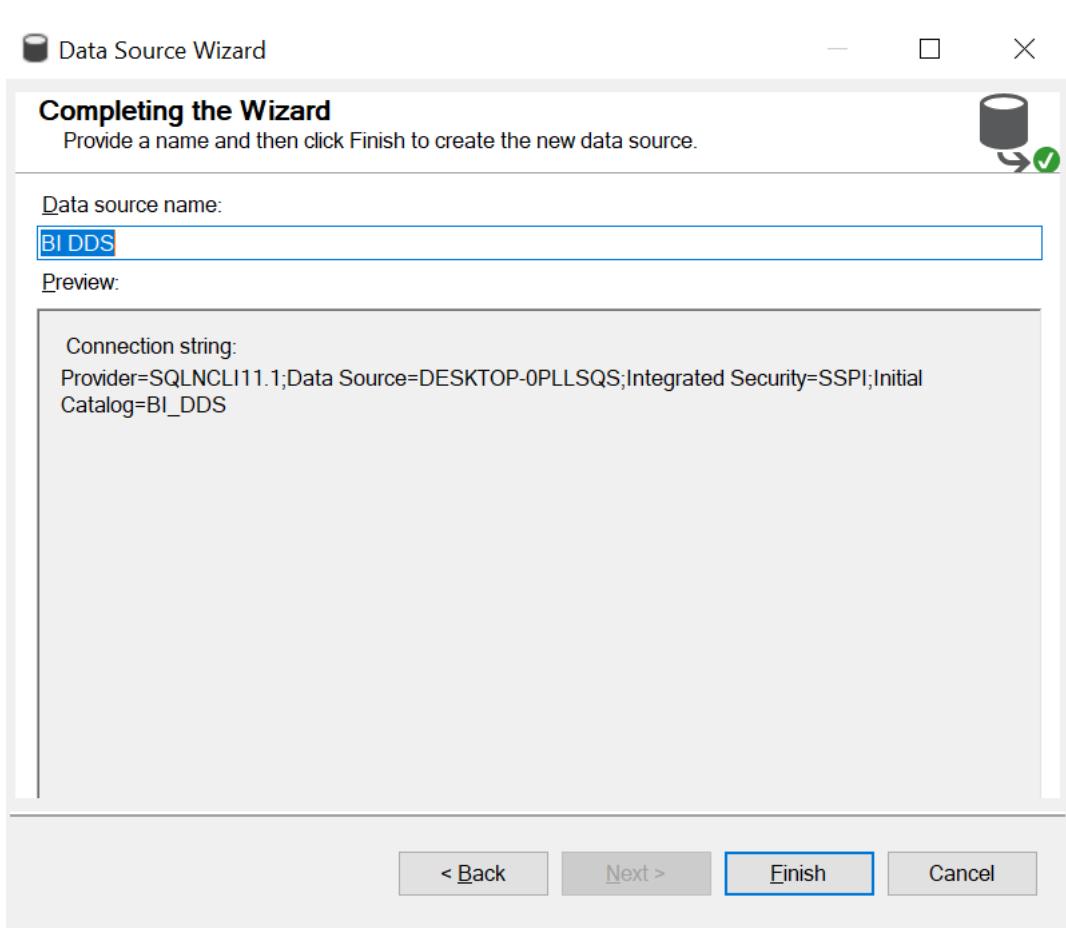
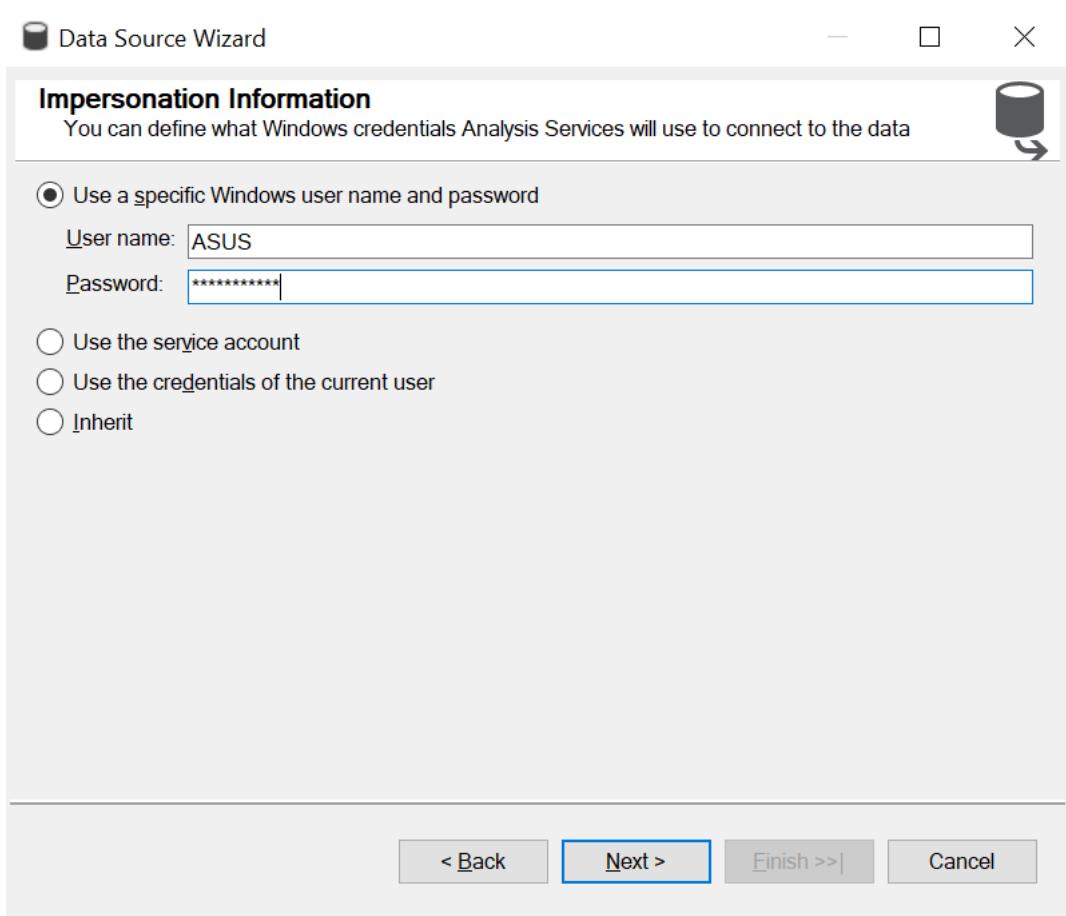


Tạo kết nối Database:

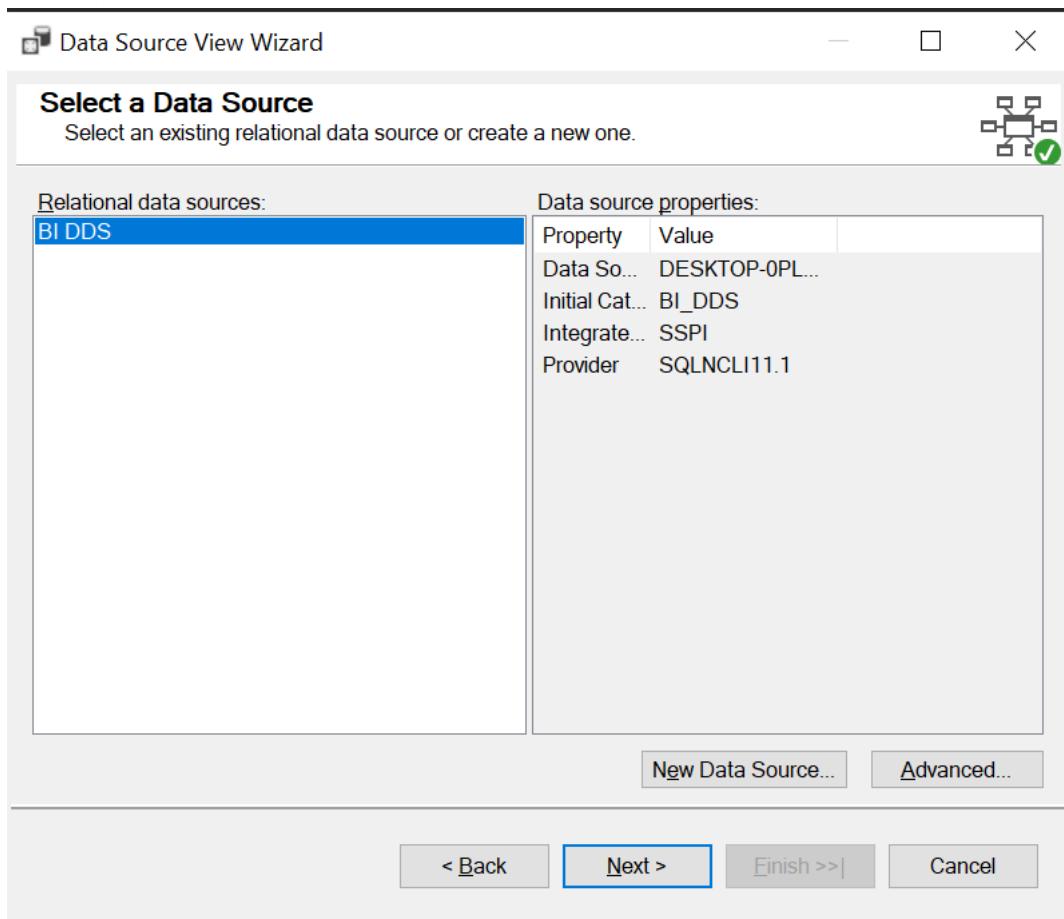


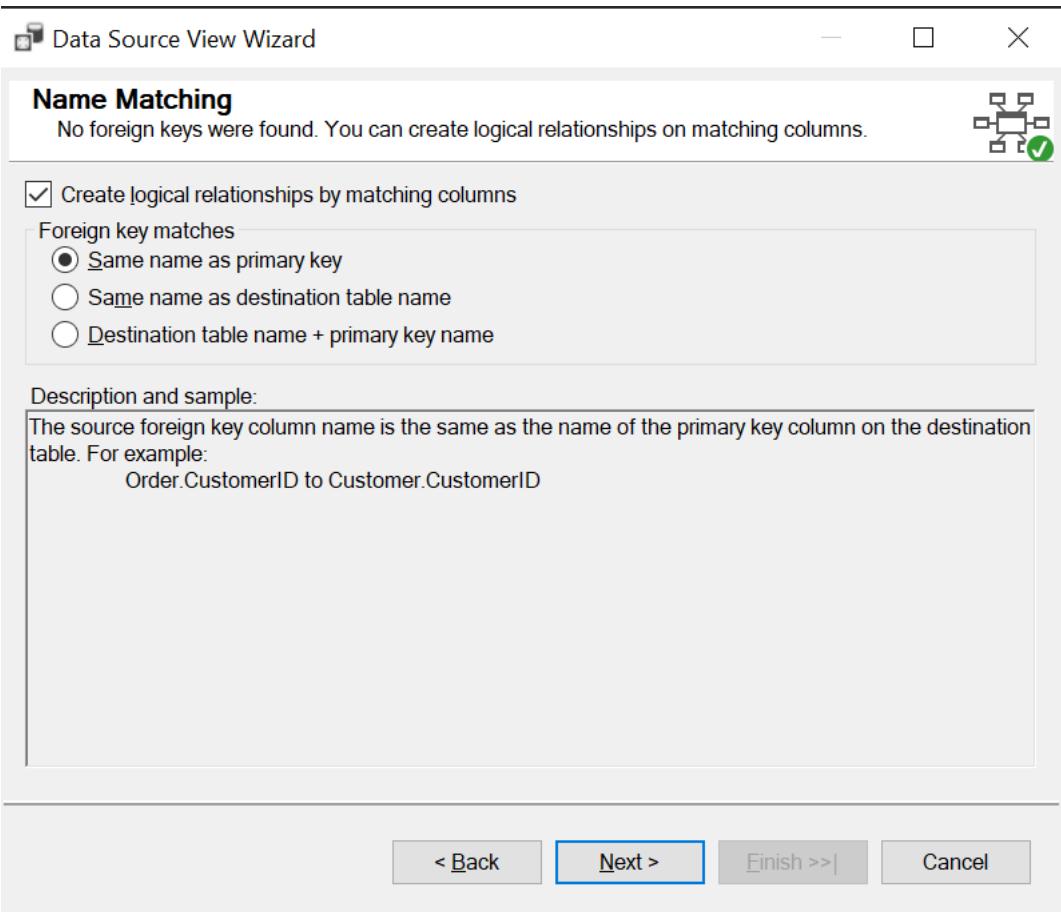


Sau khi tạo kết nối Database, ta sẽ chọn sử dụng các loại tài khoản để thực hiện process. Ở đây em chọn tài khoản của Window trên máy tính.

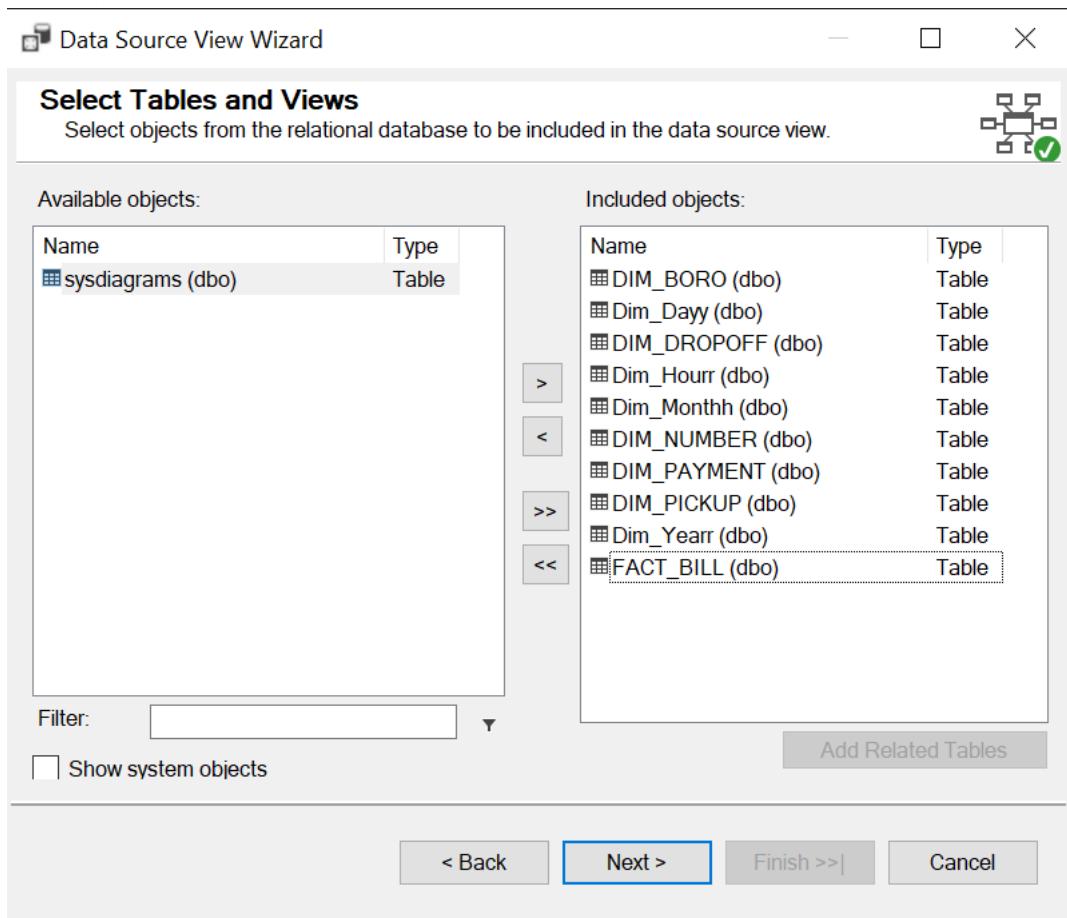


### Bước 3: Tạo Data Source View

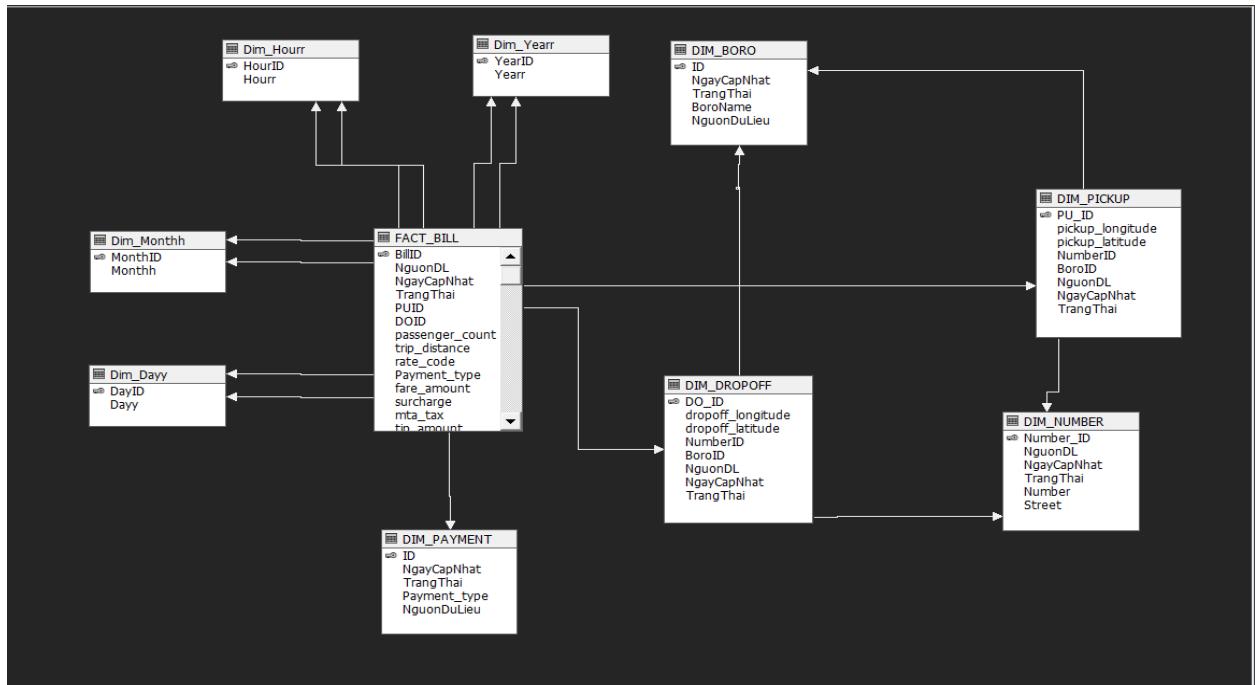




Chọn các bảng cần có trong Data Source Views.

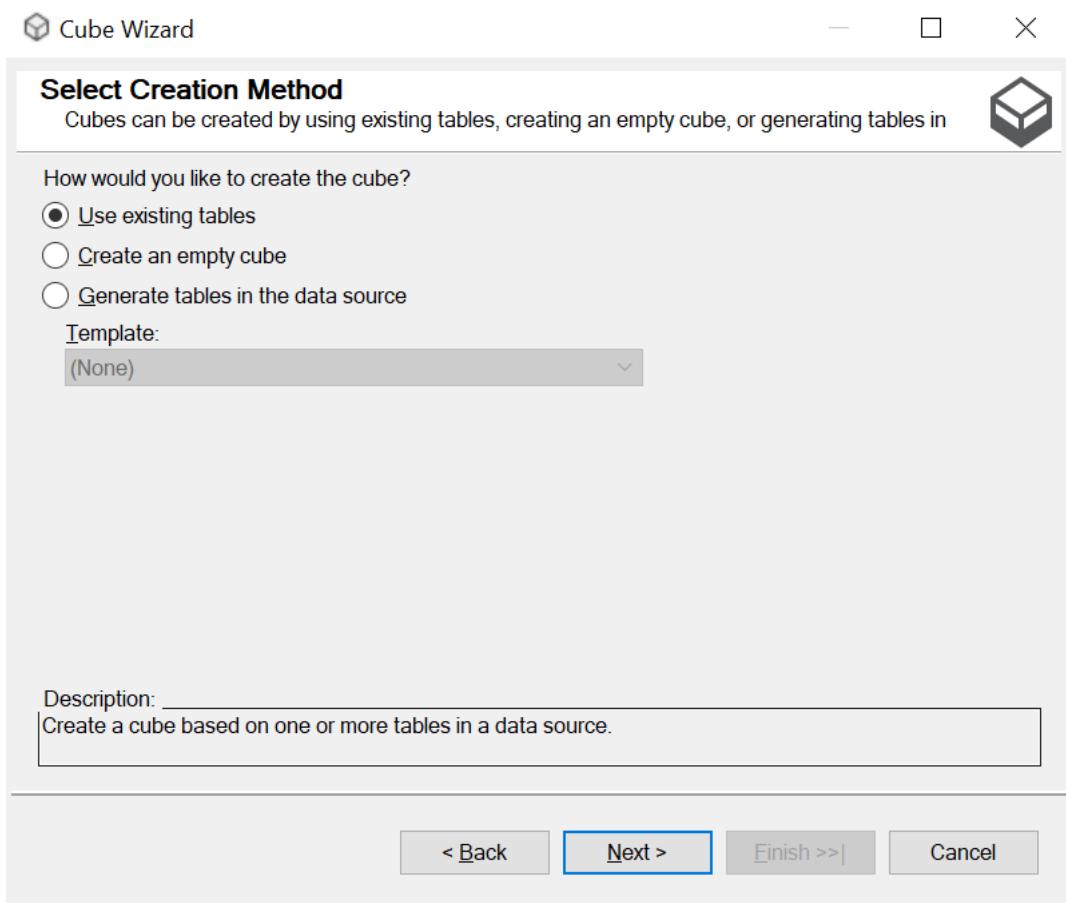
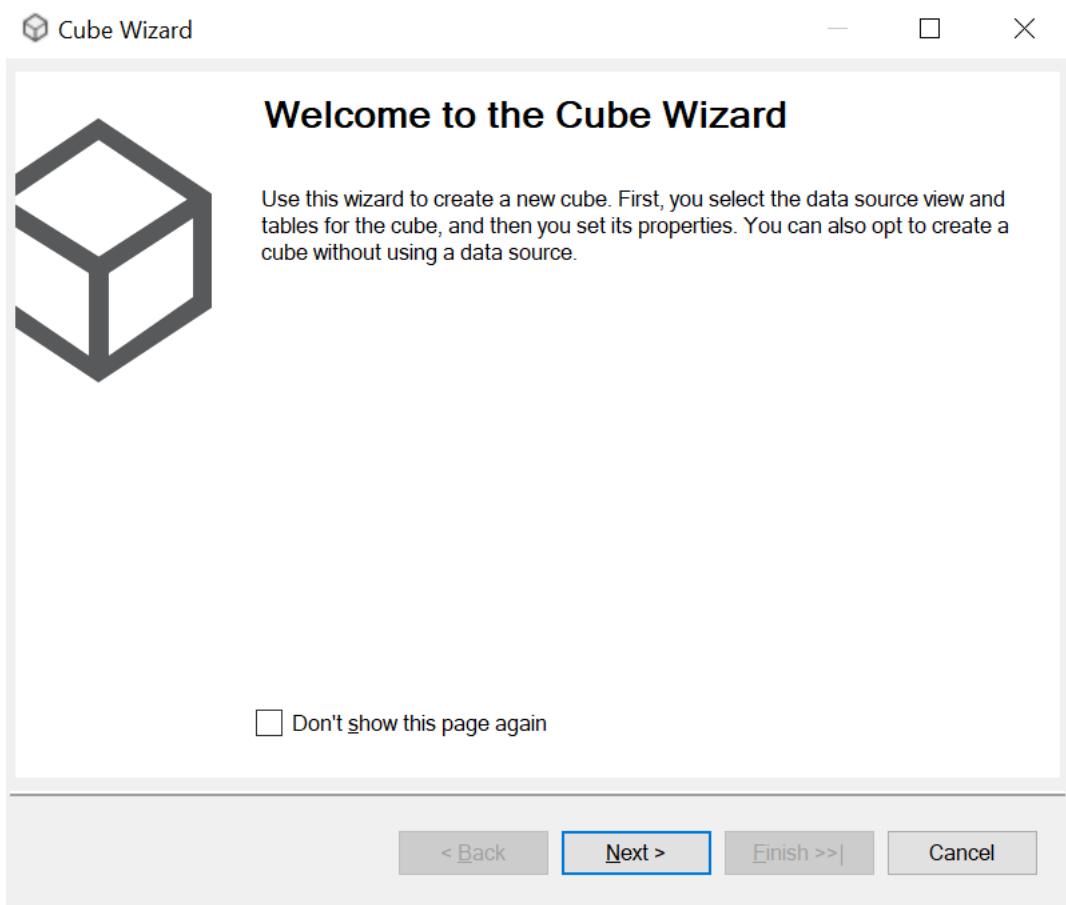


Quá trình sau khi hoàn thành:

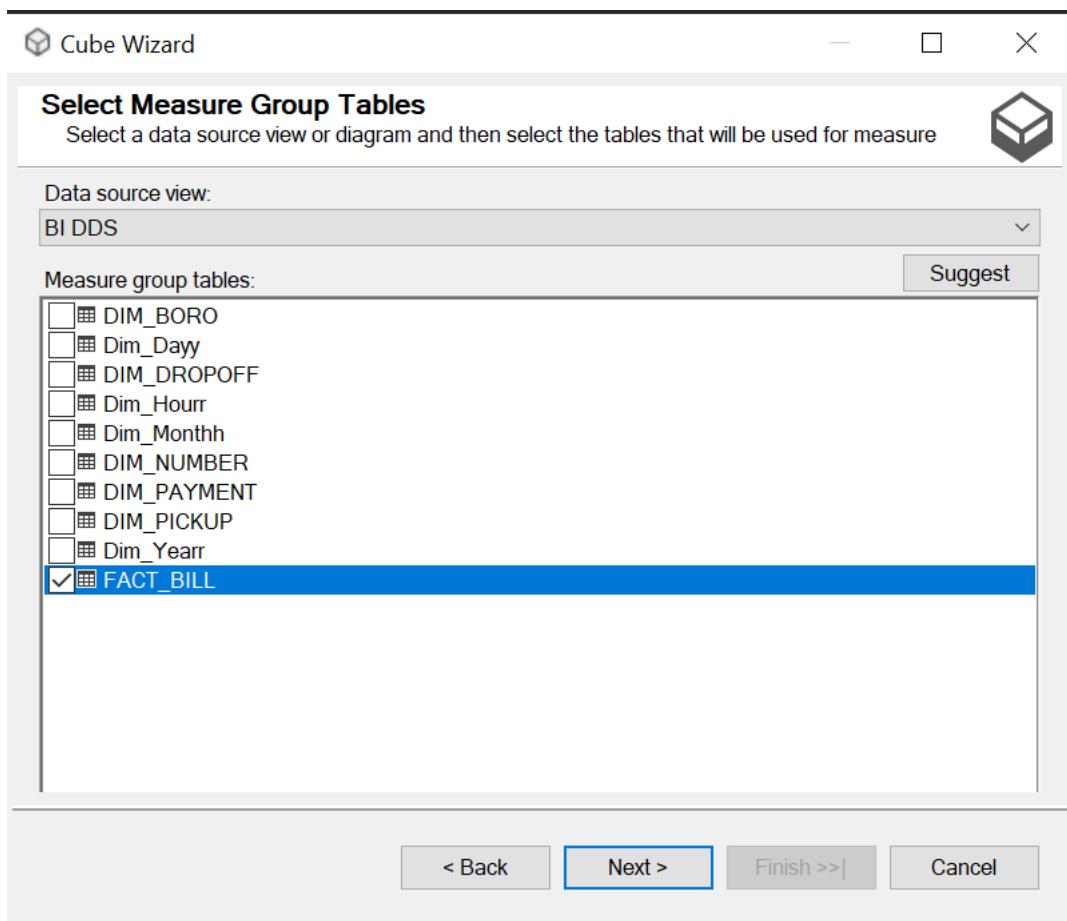


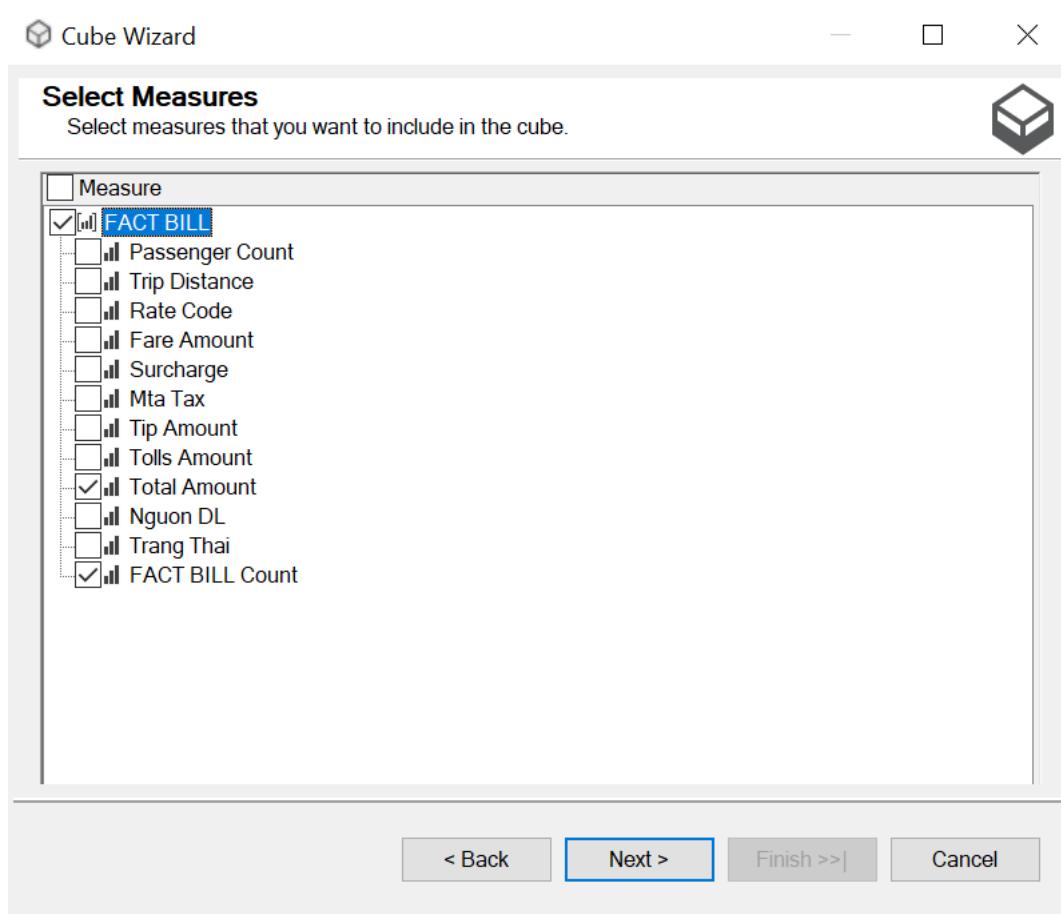
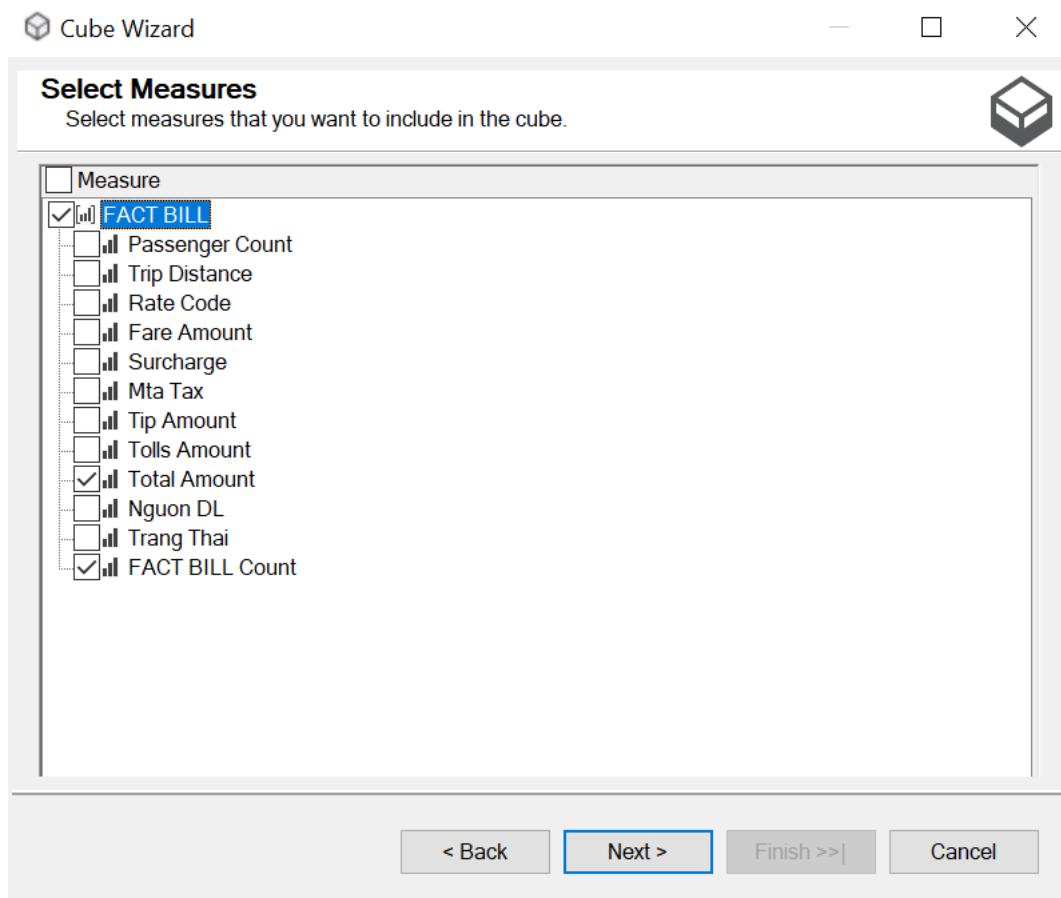
## 2. Cube

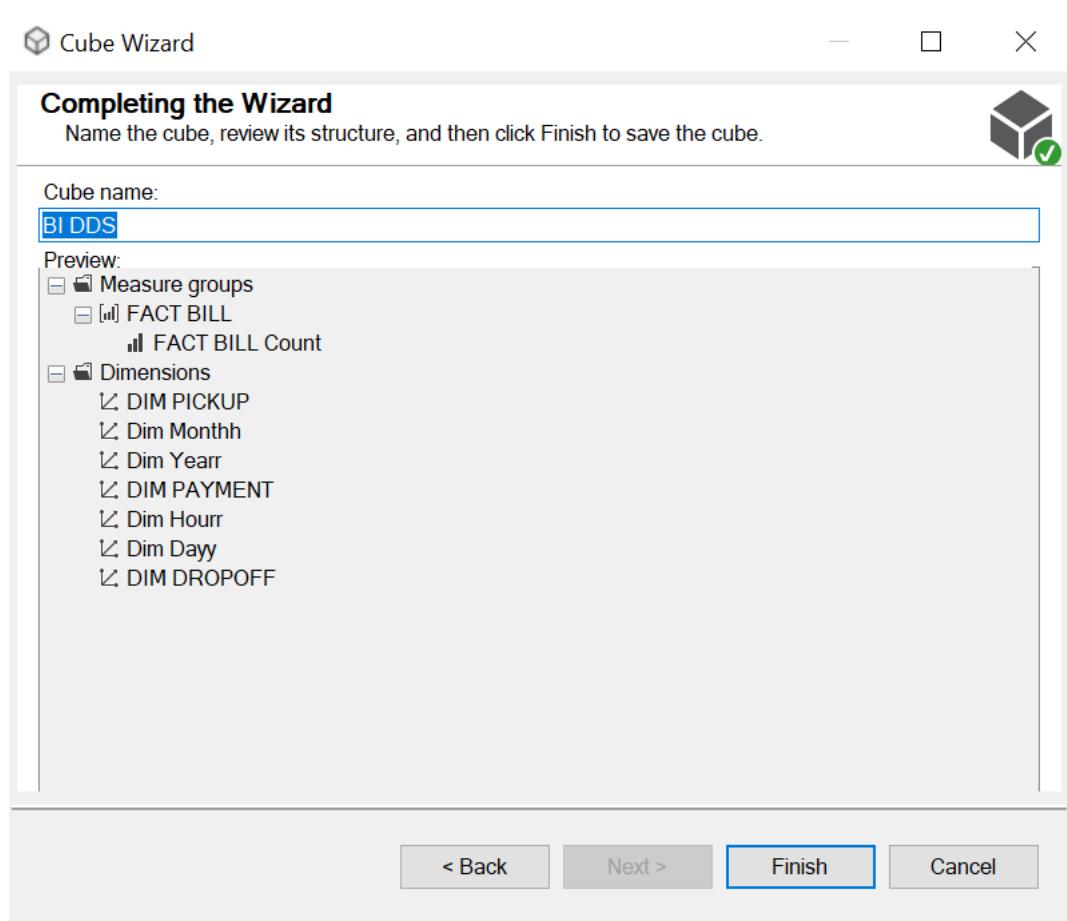
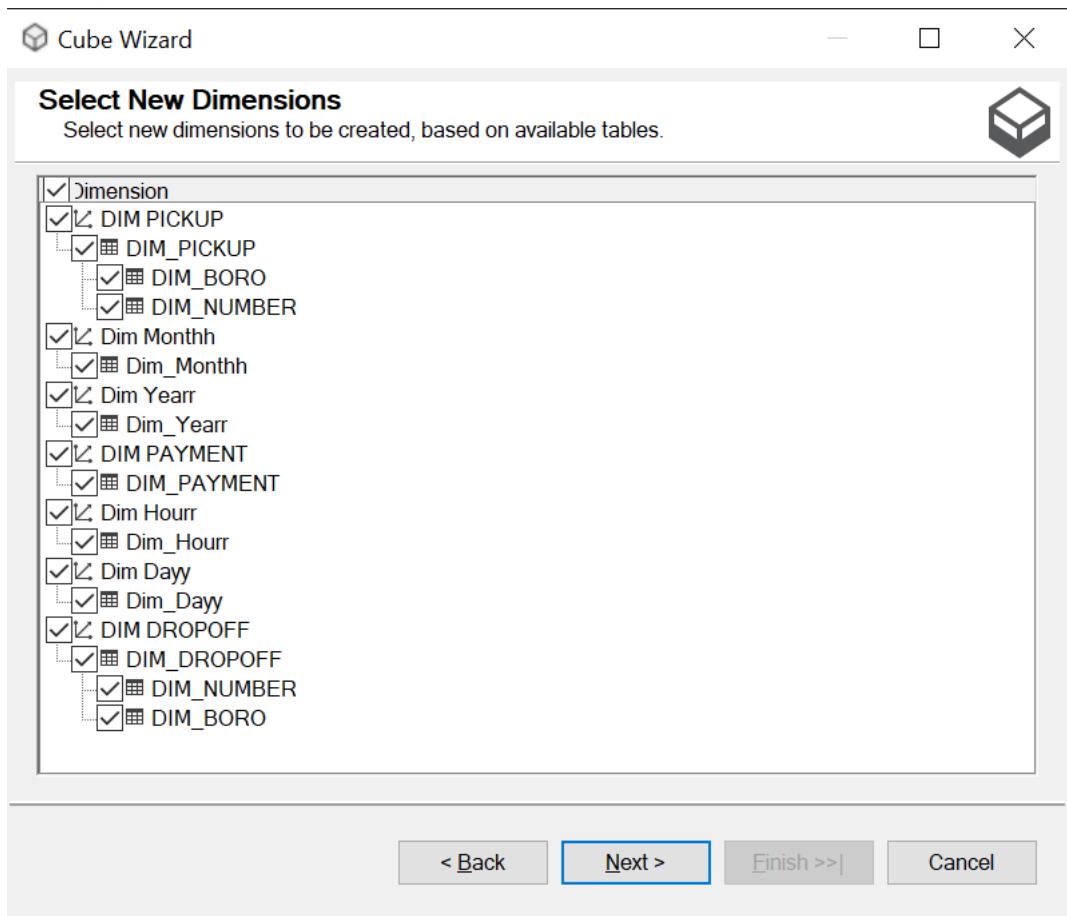
Tiếp nối phần OLAP ở trên, ta sẽ tiến hành tạo Cube:



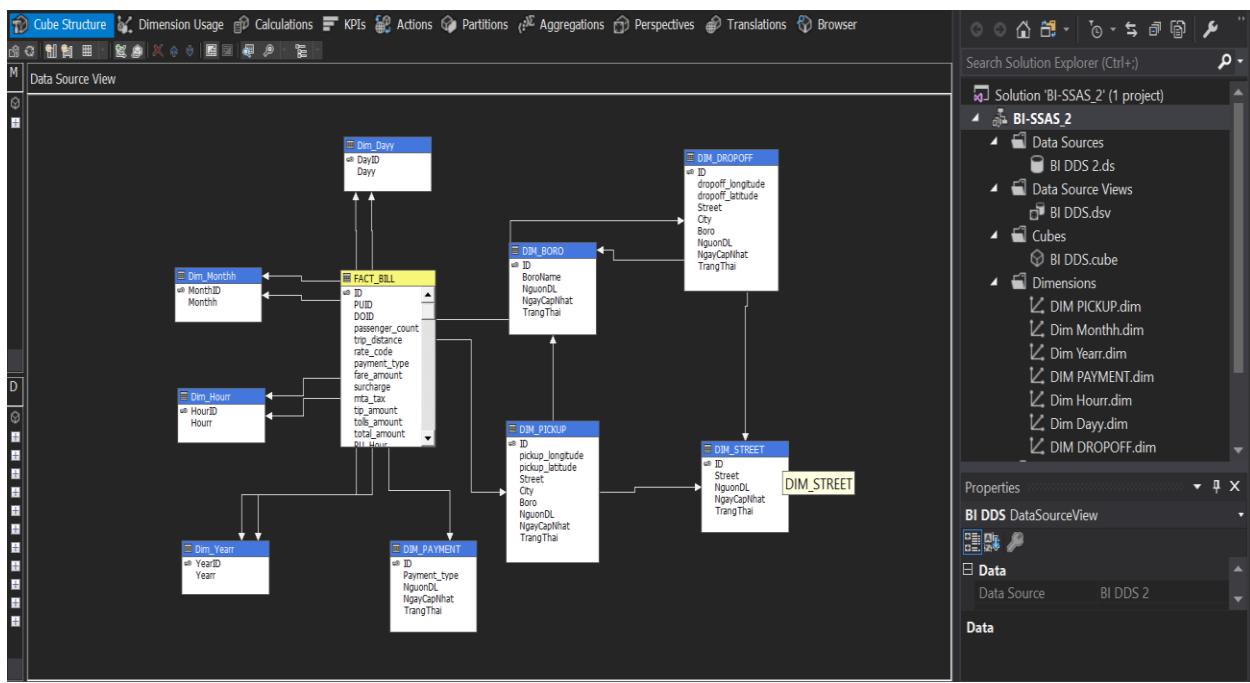
Chọn 1 bảng làm Fact, measure và dimension như hình bên dưới:





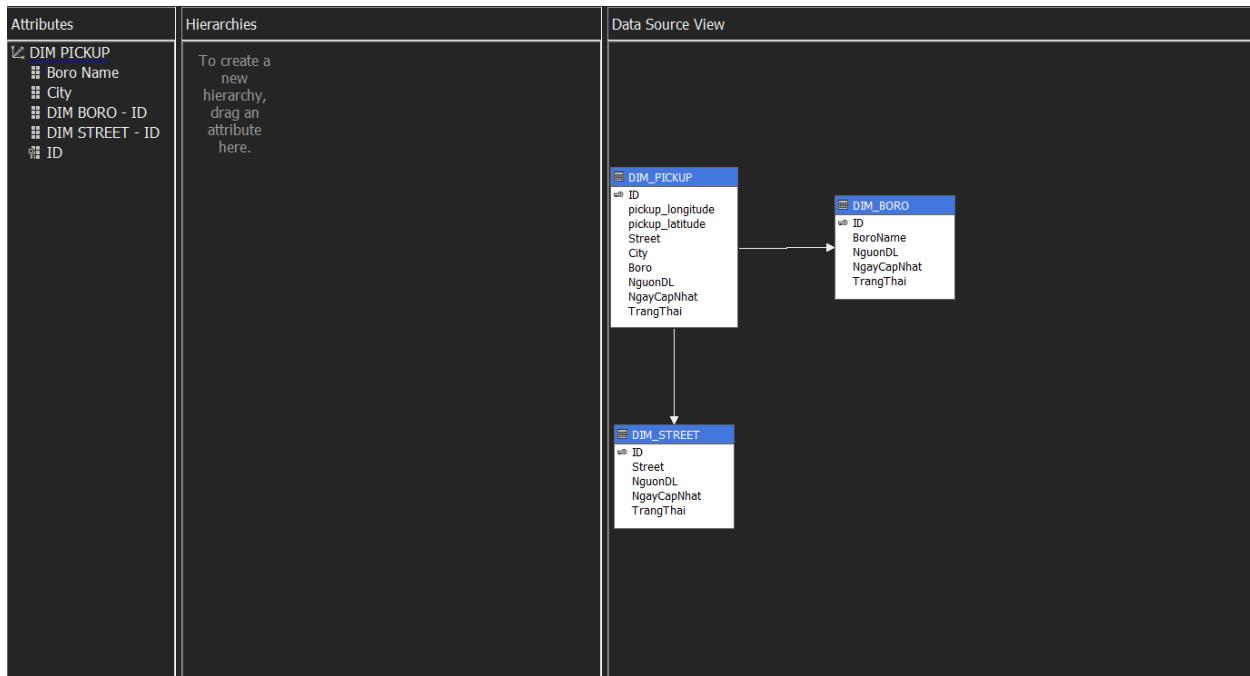


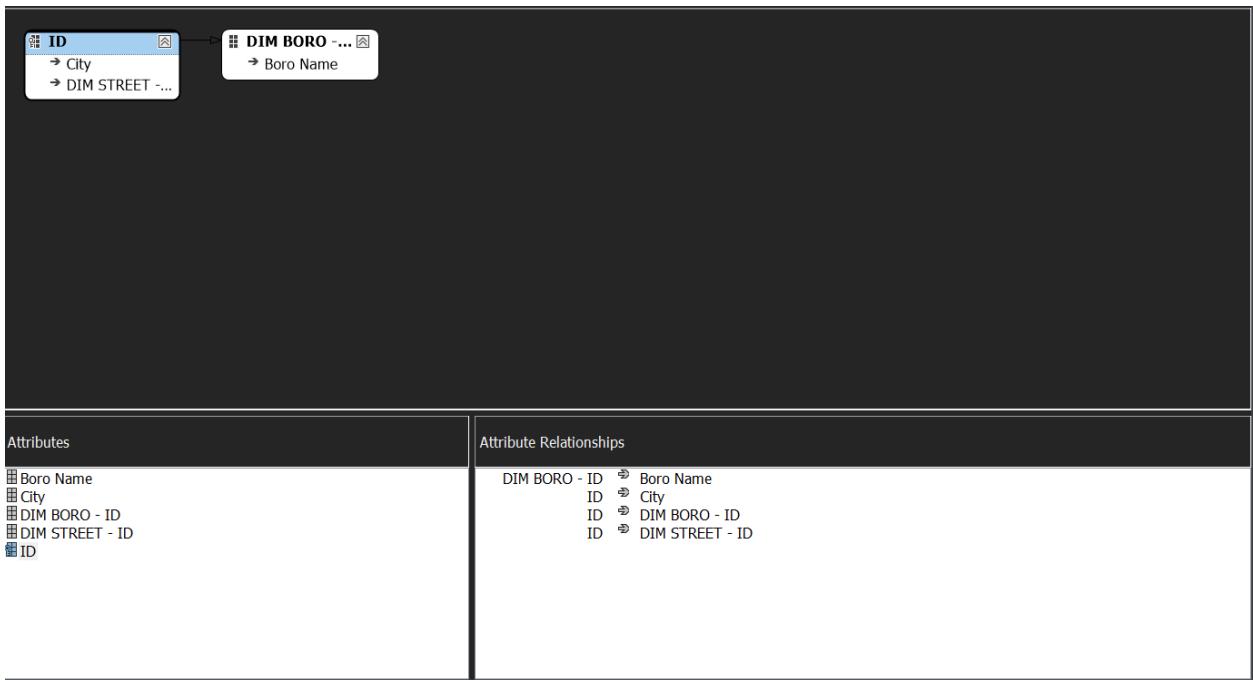
Kết quả sau khi tạo Cube:



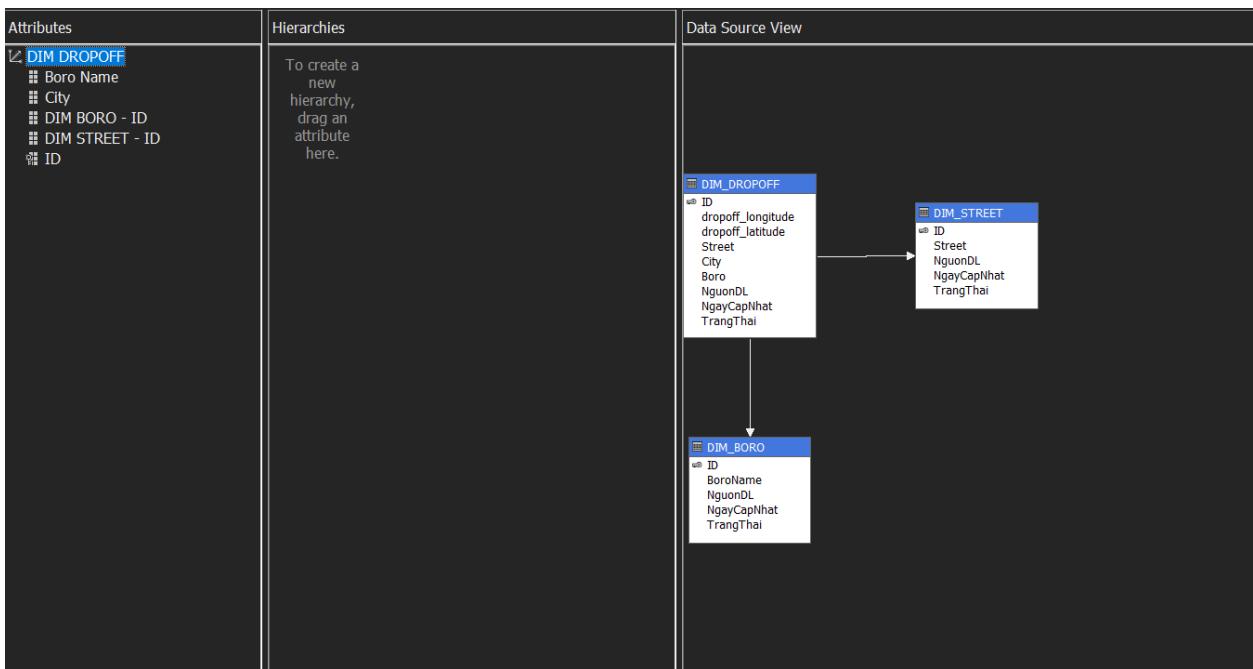
Bước 5: chỉnh sửa các dimemsion

#### a. Dim\_Pickup

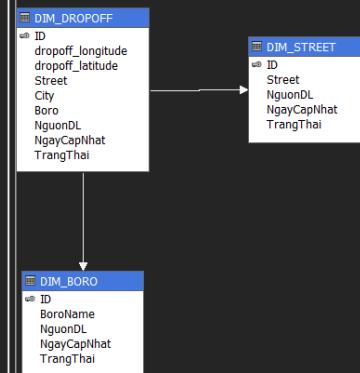


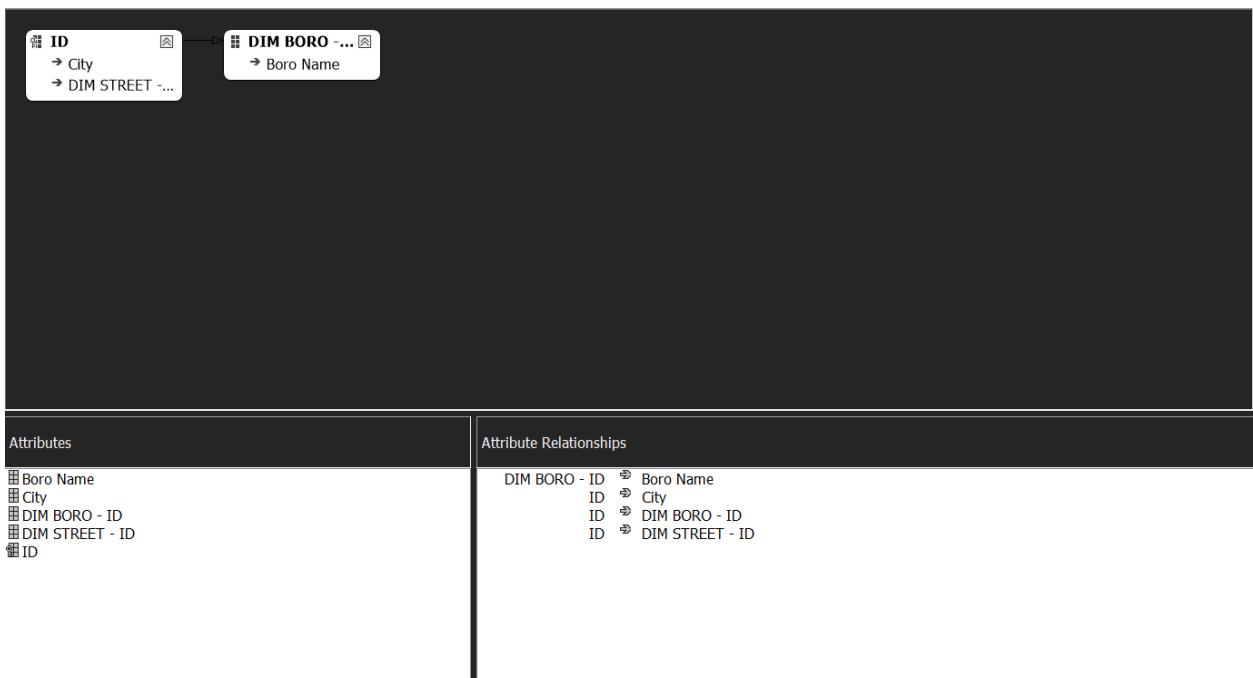


## b. Dim\_Dropoff

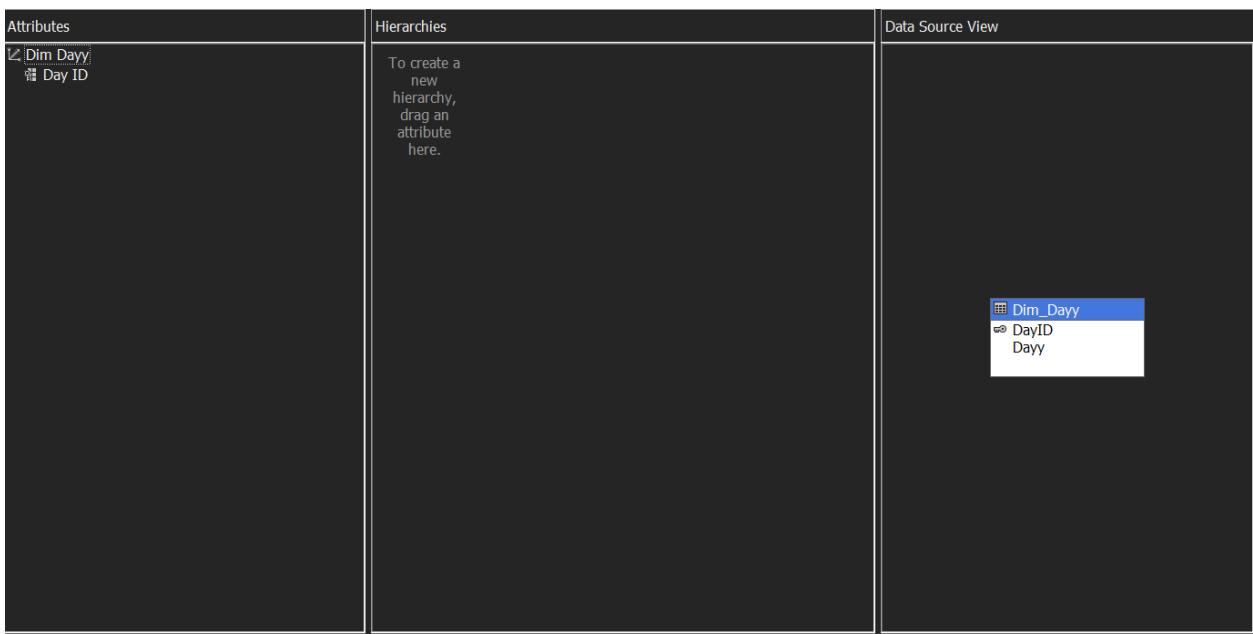


### Data Source View

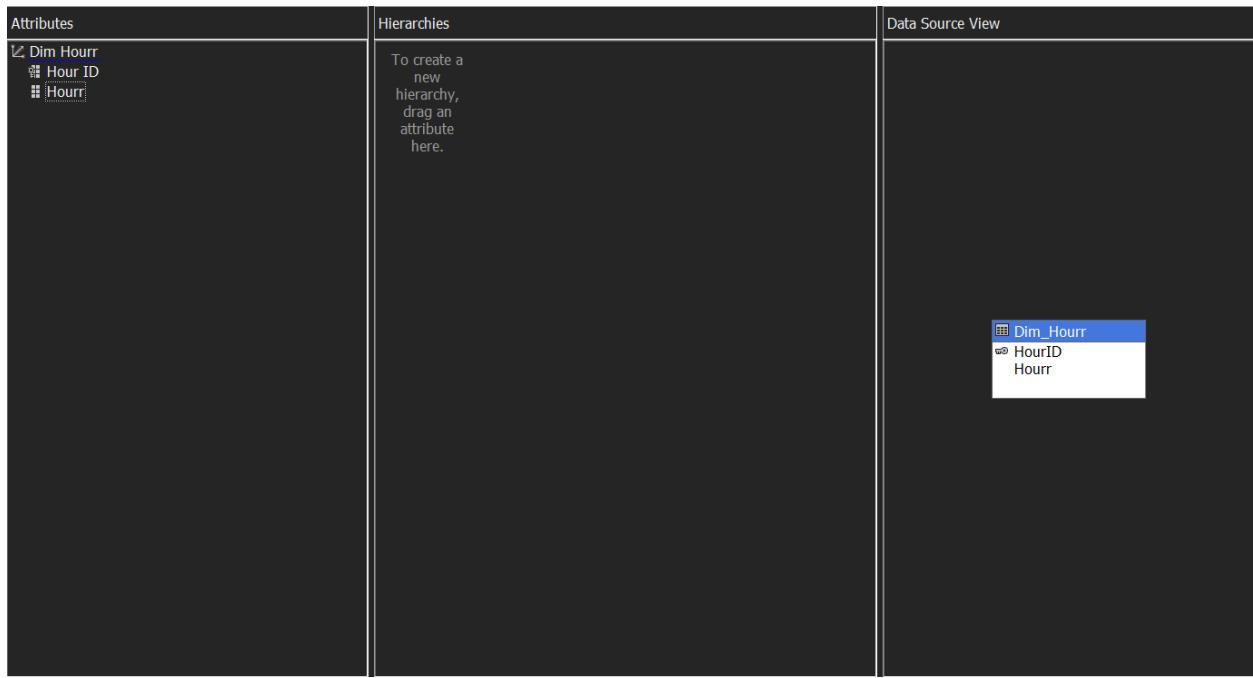




### c. Dim\_Dayy

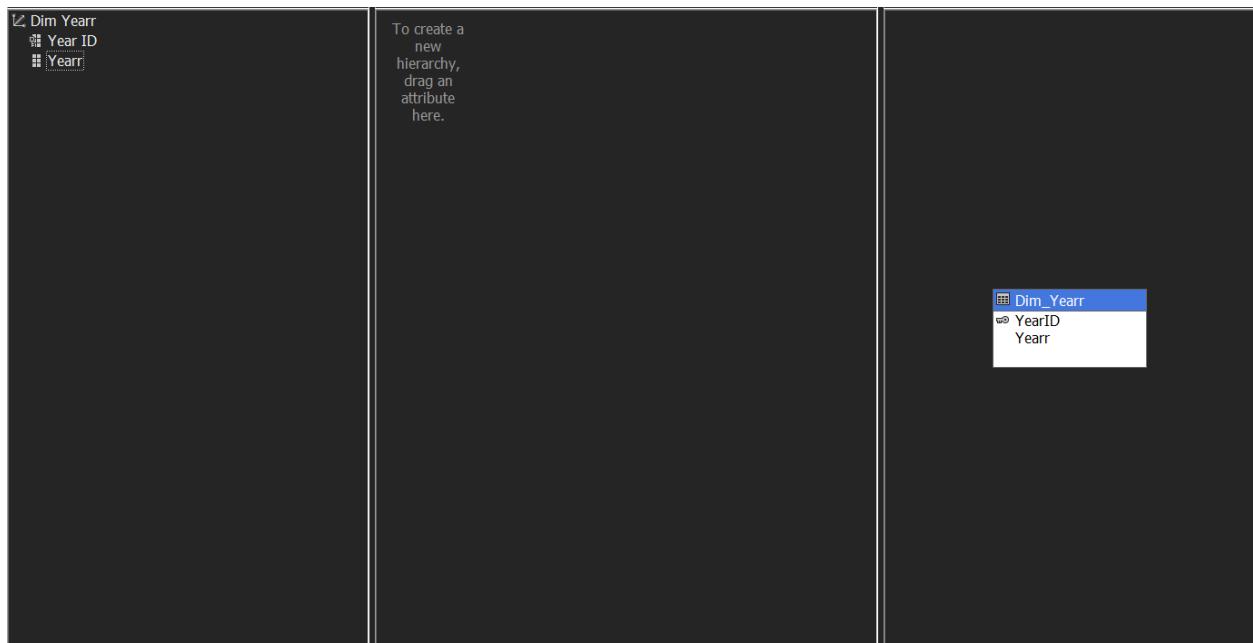


### d. Dim\_Hourr



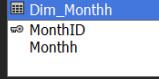
Attributes	Hierarchies	Data Source View			
<ul style="list-style-type: none"> <li>Dim Hourr           <ul style="list-style-type: none"> <li>Hour ID</li> <li>Hourr</li> </ul> </li> </ul>	To create a new hierarchy, drag an attribute here.	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">Dim_Hourr</td> </tr> <tr> <td style="padding: 2px;">HourID</td> </tr> <tr> <td style="padding: 2px;">Hourr</td> </tr> </table>	Dim_Hourr	HourID	Hourr
Dim_Hourr					
HourID					
Hourr					

#### e. *Dim\_Yearr*



Attributes	Hierarchies	Data Source View			
<ul style="list-style-type: none"> <li>Dim Yearr           <ul style="list-style-type: none"> <li>Year ID</li> <li>Yearr</li> </ul> </li> </ul>	To create a new hierarchy, drag an attribute here.	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 2px;">Dim_Yearr</td> </tr> <tr> <td style="padding: 2px;">YearID</td> </tr> <tr> <td style="padding: 2px;">Yearr</td> </tr> </table>	Dim_Yearr	YearID	Yearr
Dim_Yearr					
YearID					
Yearr					

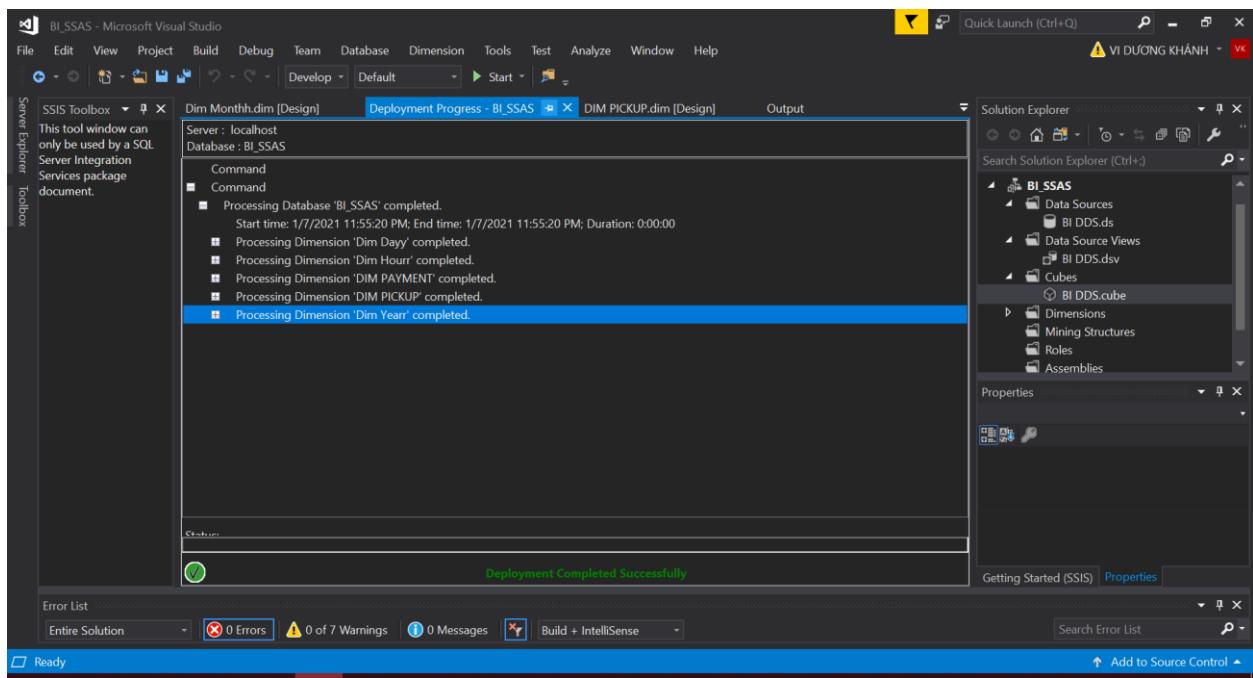
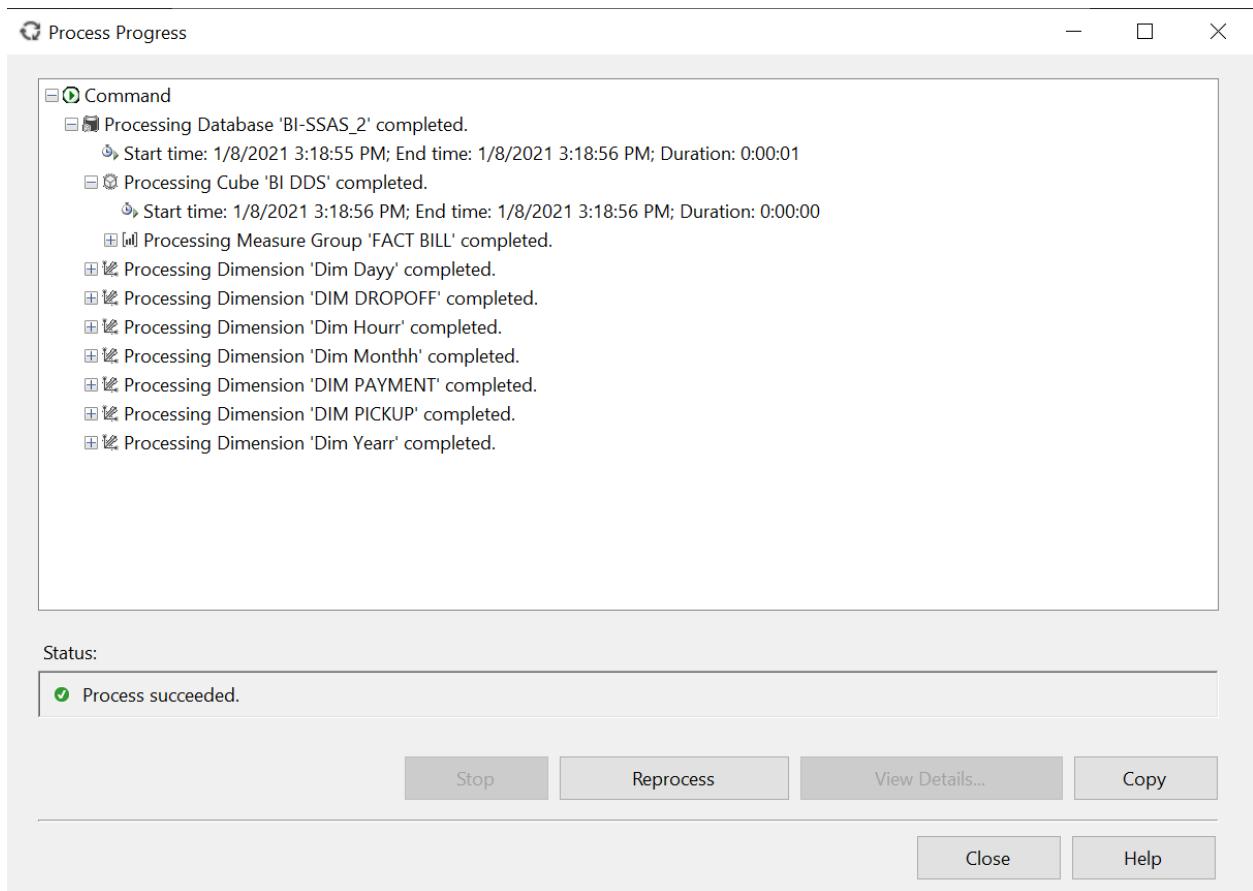
#### f. *Dim\_Monthh*

Attributes	Hierarchies	Data Source View
<input checked="" type="checkbox"/> <b>Dim Monthh</b> <input type="checkbox"/> Month ID	To create a new hierarchy, drag an attribute here.	

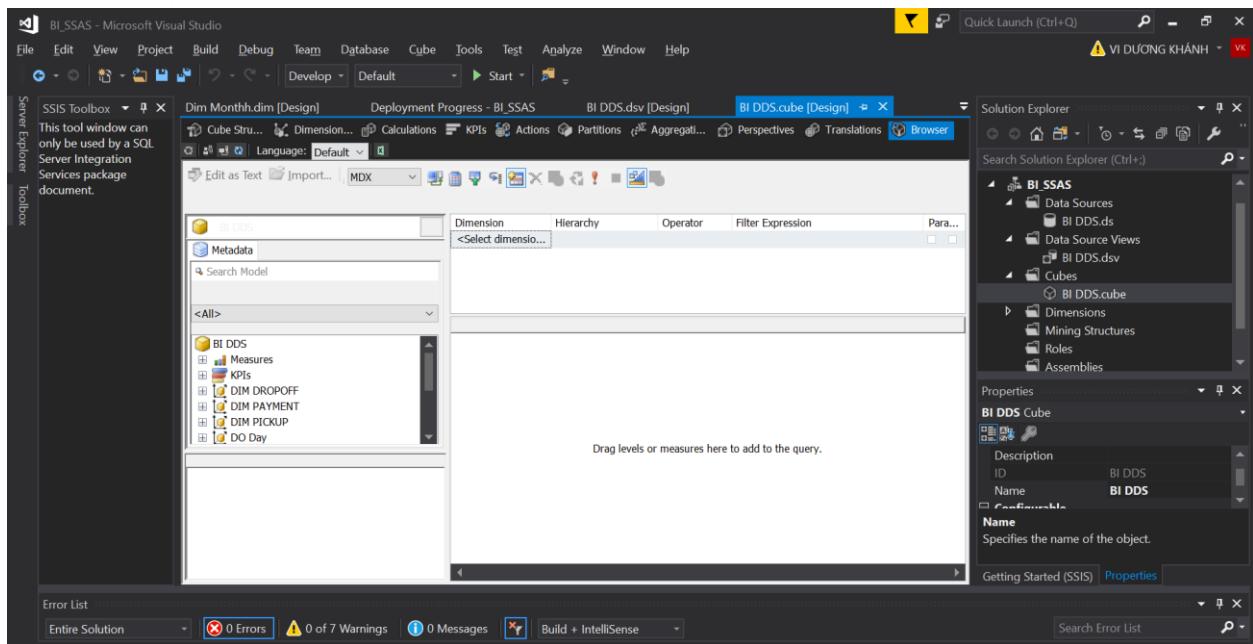
### ***g. Dim\_Payment***

Attributes	Hierarchies	Data Source View
<input checked="" type="checkbox"/> <b>DIM PAYMENT</b> <input type="checkbox"/> ID <input type="checkbox"/> Payment Type	To create a new hierarchy, drag an attribute here.	

Sau khi chỉnh sửa các dimension, tiến hành process:



- Kết quả sau khi process:



Thử thực hiện truy vấn trên cube và ta có kết quả:

DIM BORO - ID	Boro Name	Yearr	FACT BILL Count
1	Manhattan	2014	5225
1	Manhattan	2015	3163
2	Brox	2014	9
2	Brox	2015	3
3	Brooklyn	2014	204
3	Brooklyn	2015	181
4	Queens	2014	209
4	Queens	2015	160

### 3. Truy vấn bằng MDX

#### 3.1. Tính doanh thu tổng

- Doanh thu tổng

Total Amount
492740.879999931

- Doanh thu theo từng năm

```

5 | SELECT [DO Year].[Year].&[2014] ON COLUMNS FROM [BI DDS] WHERE [Measures].[Total Amount]
100 % <
Messages Results
2014
331915.429999999

```

```

7 | SELECT [DO Year].[Year].&[2015] ON COLUMNS FROM [BI DDS] WHERE [Measures].[Total Amount]
100 % <
Messages Results
2015
160825.449999997

```

### 3.2. Tính doanh thu của một giờ

- Tính doanh thu của một giờ cụ thể trong ngày

```

15 | -- TINH DOANH THU CUA 1 GIO CU THE
16 | SELECT [PU Hour].[Hour].&[01:00] ON COLUMNS ,
17 | [DO Day].[Day ID].[Day ID] ON ROWS
18 | FROM [BI DDS]
19 | WHERE [Measures].[Total Amount]
100 % <
Messages Results
01:00
1 (null)
2 35.6
3 243.3
4 75.5
5 (null)
6 (null)
7 1098.43
8 337.33
9 1584.83
10 (null)
11 355.83
12 416.5
13 (null)
14 3745.6
15 2860.01
16 6

```

	01:00
17	5076.03
18	340.33
19	192.76
20	713.53
21	(null)
22	(null)
23	(null)
24	3870.04
25	588.27
26	4253.86
27	4309.57
28	854.81
29	4257.89
30	(null)
31	29.5
Unknown	(null)

- Tính doanh thu của các giờ trong ngày

```

18 -- TINH DOANH THU CUA CAC GIO TRONG NGAY
19 SELECT [DO Hour].[Hourr].[Hourr] ON COLUMNS,
20 [DO Day].[Day ID].[Day ID] ON ROWS
21 FROM [BI DDS]
22 WHERE [Measures].[Total Amount]
23 -----3. DOANH THU THEO THANG-----

```

Messages Results

	01:00	02:00	03:00	04:00	05:00	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00	
1	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	1746.72	9077.650000000002	69.75	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	
2	35.6	55	88.4	104.18	110.8	220.02	75.92	167.79	170.6	177.37	252.78	140.1	83.5	185.38	261.51	2	2	2	2	2	2	2	2	
3	243.3	48.4	108.6	41.5	69.45	228.6	384.09	496.55	708.18	574.68	452.51	380.78	406.94	512.73	384.92	224.53	331.37	4	4	4	4	4	4	4
4	75.5	129.12	51.35	24.6	(null)	32	108	147.29	147.2	163.52	171.83	225.94	115.6	118.3	55.17	134.17	132.9	2	2	2	2	2	2	2
5	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	
6	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	
7	943.68	1131.2	1044.71	511.8	346.16	469.72	376.2	430.1	684.28	541.23	401.67	706.22	691.26	526.27	715.82	443.51	976.64	1	1	1	1	1	1	1
8	231.5	230.33	105	37	614.5	885.49	945.16	810.32	913.66	846.5	817.65	974.16	1038.83	731.83	570	189.5	21	21	21	21	21	21	21	
9	1121.5	1318.83	869.88	614.33	496.83	495.16	386.16	256.5	443.33	618.16	714.66	926.32	667	950.16	516.83	632.83	364.33	6	6	6	6	6	6	6
10	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	
11	222	394.33	350.5	317.9	123.5	321.79	543.18	810.88	1050.42	872.5	1261.39	1176.1	1372.15	1367.12	1869.27	1620.1	1671.04	24	24	24	24	24	24	24
12	287	299.5	329.5	419.83	109.33	226.83	73	241.83	173.83	224.5	251.5	254.5	312	191.5	225.33	193.5	275.01	68	68	68	68	68	68	68
13	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	(null)	240.25	(null)	1469.04	485.9	1460.99	1134.05	145.02	48	135.5	135.5	135.5	135.5	135.5	135.5	135.5
14	2427.23	3369.16	1449.98	877.12	180.67	686.25	557.25	1125.36	1738.08	1244.79	745.2	1029.3	1688.35	1269.11	1192.08	985.11	362.48	4	4	4	4	4	4	4
15	2277.7	1785.61	1209.2	599.21	774.69	786.17	815.82	505.63	556.96	523.48	389.02	354.48	339.62	192.27	371.33	255.65	215.6	1	1	1	1	1	1	1
16	6	(null)	(null)	(null)	(null)	(null)	(null)	(null)	18.5	20.5	(null)	(null)	(null)	3	(null)	7	(null)	(null)	(null)	(null)	(null)	(null)	(null)	
17	3996.25	2595.77	679.24	375.15	502.15	316.49	428.88	653.48	430.4	418.95	271.78	166.1	233.79	231.13	110.4	131.03	159.2	1	1	1	1	1	1	1
18	160.1	336.4	176.1	137.14	77.2	36.95	139.13	10	58.8	24.6	40.4	86.15	34	24.35	58	38.37	(null)	(null)	(null)	(null)	(null)	(null)	(null)	

### 3.3. Tính doanh thu theo tháng trong năm

```

25 -----3. DOANH THU THEO THANG-----
26 SELECT {[DO Year].[Yeaarr].[Yeaarr]} ON COLUMNS,
27 {[PU Month].[Month ID].[Month ID]} ON ROWS
28 FROM [BI DDS]
29 WHERE [Measures].[Total Amount]
30

```

100 % <

	2014	2015	Unknown
1	14314.21	24597.5399999996	(null)
2	27048.08	10894.12	(null)
3	29031.8	29416.3299999995	(null)
4	44388.98	57524.1900000018	(null)
5	(null)	38393.27	(null)
6	51017.33	(null)	(null)
7	23186.61	(null)	(null)
8	68398.5700000001	(null)	(null)
9	(null)	(null)	(null)
10	23830.09	(null)	(null)
11	(null)	(null)	(null)
12	50699.76	(null)	(null)
Unknown	(null)	(null)	(null)

### 3.4. Doanh thu theo boro

- Doanh thu theo boro theo năm

```

30 -----4. DOANH THU CUA CAC BORO THEO TUNG NAM-----
31 SELECT [DO Year].[Yeaarr].[Yeaarr] ON COLUMNS,
32 [DIM PICKUP].[Boro Name].[Boro Name] ON ROWS
33 FROM [BI DDS]
34 WHERE [Measures].[Total Amount]

```

100 % <

	2014	2015	Unknown
Brooklyn	13626.29	7468.2800000004	(null)
Brox	303.28	123.85	(null)
Manhattan	269295.29	138649.250000006	(null)
Queens	23497.13	14108.26	(null)
Staten Island	(null)	(null)	(null)
Unknown	25193.44	475.81	(null)

- Doanh thu theo boro theo tháng

```

36 -----5. DOANH THU CUA CAC BORO THEO THANG-----
37 SELECT [DO Month].[Month ID].[Month ID] ON ROWS,
38 [DIM PICKUP].[Boro Name].[Boro Name] ON COLUMNS
39 FROM [BI DDS]
40 WHERE [Measures].[Total Amount]
41

```

100 % <

	Brooklyn	Brox	Manhattan	Queens	Staten Island	Unknown
1	1750.71	(null)	33743.0799999995	3344.41	(null)	73.55
2	1418.13	17	34437.6199999999	2069.45	(null)	(null)
3	2671.18	34.1	50404.9400000018	4959.4500000001	(null)	378.46
4	4924.5300000002	139.25	88937.4300000052	7563.6600000002	(null)	319
5	1633.97	23.6	33348.0899999994	3393.11	(null)	23.8
6	1977.29	76.3	44481.35	4176.16	(null)	306.23
7	(null)	(null)	(null)	(null)	(null)	23186.61
8	3009.77	62.78	58118.16	6329.26	(null)	878.6
9	(null)	(null)	(null)	(null)	(null)	(null)
10	804.16	6.5	21076.77	1924.66	(null)	18
11	(null)	(null)	(null)	(null)	(null)	(null)
12	2904.83	67.6	43397.0999999999	3845.23	(null)	485
Unknown	(null)	(null)	(null)	(null)	(null)	(null)

### 3.5. Phân tích chuyến đi theo giờ để biết đâu là giờ cao điểm

```

42 -----6. PHAN TICH GIO CAO DIEM-----
43 -- THONG KE SO LUONG CHUYEN DI TRONG 1 GIO TRONG NAM
44 SELECT [DO Hour].[Hourr].[Hourr] ON ROWS,
45 [DO Year].[Yearrr].[Yearrr] ON COLUMNS
46 FROM [BI DDS]
47 WHERE [Measures].[FACT BILL Count]
48

```

100 % <

	2014	2015	Unknown
01:00	1737	(null)	(null)
02:00	1524	(null)	(null)
03:00	836	(null)	(null)
04:00	463	(null)	(null)
05:00	312	(null)	(null)
06:00	428	(null)	(null)
07:00	539	(null)	(null)
08:00	609	(null)	(null)
09:00	733	(null)	(null)
10:00	653	(null)	(null)
11:00	675	222	(null)
12:00	634	659	(null)
13:00	768	1	(null)
14:00	686	(null)	(null)
15:00	564	(null)	(null)
16:00	468	(null)	(null)
17:00	441	(null)	(null)
18:00	1071	(null)	(null)
19:00	723	(null)	(null)
20:00	670	(null)	(null)
21:00	652	(null)	(null)
22:00	615	(null)	(null)
23:00	577	(null)	(null)
24:00	5930	9348	(null)
Unknown	(null)	(null)	(null)

### 3.6. Phân tích chuyến đi theo địa chỉ cụ thể

- Thống kê số lượng điểm đón trong năm theo thành phố

```

49  -- THONG KE SO LUONG DIEM DON TRONG NAM THEO THANH PHO
50  SELECT [DO Year].[Year].[] ON COLUMNS,
51  [DIM PICKUP].[City].[City] ON ROWS
52  FROM [BI DDS]
53  WHERE [Measures].[FACT BILL Count]

```

100 % < Messages Results

	2014	2015	Unknown
Amityville	1	(null)	(null)
Astoria	142	36	(null)
Bronx	20	9	(null)
Brooklyn	874	479	(null)
Corona	5	(null)	(null)
Cresskill	(null)	1	(null)
East Elmhurst	41	33	(null)
East Orange	1	(null)	(null)
East Rutherford	1	(null)	(null)
Elmhurst	6	4	(null)
Far Rockaway	1	(null)	(null)
Farmingdale	1	(null)	(null)
Flushing	36	2	(null)
Forest Hills	7	2	(null)
Hoboken	2	(null)	(null)
Jackson Heights	6	1	(null)
Jamaica	303	216	(null)
Jersey City	5	1	(null)
Kenilworth	1	(null)	(null)
Kew Gardens	4	(null)	(null)

	2014	2015	Unknown
Kew Gardens	4	(null)	(null)
Little Ferry	(null)	1	(null)
Long Island City	102	39	(null)
Maspeth	3	(null)	(null)
Middle Village	1	1	(null)
Montclair	1	(null)	(null)
Nanuet	1	(null)	(null)
New Providence	(null)	1	(null)
New York	20561	9350	(null)
Newark	2	(null)	(null)
North Bergen	1	(null)	(null)
Ozone Park	1	(null)	(null)
Passaic	1	(null)	(null)
Rego Park	5	1	(null)
Richmond Hill	1	(null)	(null)
Ridgewood	1	1	(null)
Seaford	1	(null)	(null)
South Ozone Park	3	(null)	(null)
South Richmond Hill	2	(null)	(null)
Stamford	(null)	1	(null)

Stamford	(null)	1	(null)
Sunnyside	26	12	(null)
Tarrytown	1	(null)	(null)
The Bronx	96	30	(null)
Union City	1	(null)	(null)
Weehawken	3	(null)	(null)
Whitestone	(null)	1	(null)
Woodside	37	8	(null)
Unknown	(null)	(null)	(null)

- Thống kê điểm đón của thành phố theo tháng

```

55 -- THONG KE SO LUONG DIEM DON TRONG THANG THEO THANH PHO
56 SELECT [DO Month].[Month ID].[Month ID] ON COLUMNS,
57 [DIM PICKUP].[City].[City] ON ROWS
58 FROM [BI DDS]
59 WHERE [Measures].[FACT BILL Count]

```

100 %

	1	2	3	4	5	6	7	8	9	10	11	12	Unknown
Amityville	(null)	1	(null)	(null)	(null)	(null)	(null)						
Astoria	6	22	11	43	7	27	10	27	(null)	16	(null)	9	(null)
Bronx	(null)	1	3	10	2	4	2	2	(null)	1	(null)	4	(null)
Brooklyn	110	94	159	327	112	123	31	187	(null)	45	(null)	165	(null)
Corona	(null)	1	(null)	1	(null)	1	1	1	(null)	(null)	(null)	(null)	(null)
Cresskill	(null)	(null)	(null)	(null)	1	(null)							
East Elmhurst	8	6	14	16	5	9	5	4	(null)	6	(null)	1	(null)
East Orange	(null)	(null)	(null)	(null)	(null)	1	(null)						
East Rutherford	(null)	1	(null)	(null)	(null)	(null)	(null)						
Elmhurst	2	2	(null)	3	(null)	2	(null)	(null)	(null)	(null)	(null)	1	(null)
Far Rockaway	(null)	1	(null)	(null)	(null)	(null)	(null)						
Farmingdale	(null)	1	(null)	(null)	(null)	(null)	(null)						
Flushing	4	4	5	9	(null)	7	2	3	(null)	3	(null)	1	(null)
Forest Hills	2	1	3	2	(null)	1	(null)						
Hoboken	(null)	1	(null)	(null)	(null)	1	(null)						
Jackson Heights	1	(null)	1	2	(null)	(null)	(null)	(null)	(null)	2	(null)	1	(null)
Jamaica	48	18	70	96	59	43	43	79	(null)	21	(null)	42	(null)
Jersey City	1	(null)	(null)	2	(null)	1	(null)	1	(null)	(null)	(null)	1	(null)
Kenilworth	(null)	1	(null)										

	1	2	3	4	5	6	7	8	9	10	11	12	Unknown
Kew Gardens	(null)	(null)	1	(null)	(null)	1	(null)	1	(null)	(null)	(null)	1	(null)
Little Ferry	(null)	(null)	1	(null)									
Long Island City	10	12	8	25	13	17	9	19	(null)	7	(null)	21	(null)
Maspeth	(null)	(null)	1	(null)	2	(null)							
Middle Village	(null)	1	(null)	1	(null)								
Montclair	(null)	(null)	(null)	1	(null)								
Nanuet	(null)	1	(null)										
New Providence	(null)	(null)	1	(null)									
New York	2478	2703	3491	6059	2165	3195	1882	3564	(null)	1722	(null)	2652	(null)
Newark	(null)	2	(null)	(null)	(null)	(null)	(null)						
North Bergen	(null)	(null)	(null)	1	(null)								
Ozone Park	(null)	1	(null)										
Passaic	(null)	(null)	(null)	(null)	(null)	1	(null)						
Rego Park	(null)	(null)	(null)	1	(null)	4	(null)	(null)	(null)	1	(null)	(null)	(null)
Richmond Hill	(null)	(null)	(null)	(null)	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)
Ridgewood	(null)	(null)	(null)	1	(null)	(null)	1	(null)	(null)	(null)	(null)	(null)	(null)
Seaford	(null)	1	(null)										
South Ozone Park	(null)	(null)	(null)	1	(null)	(null)	(null)	2	(null)	(null)	(null)	(null)	(null)
South Richmond Hill	(null)	(null)	(null)	(null)	(null)	1	(null)	1	(null)	(null)	(null)	(null)	(null)
Unknown	(null)	(null)	1	(null)									

Stamford	(null)	(null)	1	(null)									
Sunnyside	3	4	3	10	1	7	2	3	(null)	1	(null)	4	(null)
Tarrytown	(null)	1	(null)	(null)	(null)	(null)	(null)						
The Bronx	13	10	16	19	5	19	9	14	(null)	8	(null)	13	(null)
Union City	(null)	1	(null)	(null)	(null)								
Weehawken	(null)	(null)	(null)	(null)	(null)	1	(null)	2	(null)	(null)	(null)	(null)	(null)
Whitestone	(null)	(null)	1	(null)									
Woodside	2	4	5	9	2	8	3	3	(null)	3	(null)	6	(null)
Unknown	(null)												

- Thống kê số lượng điểm đi trong năm theo thành phố

```

66  -- THONG KE SO LUONG DIEM DI TRONG NAM THEO THANH PHO
67  SELECT [DO Year].[Year].<Year> ON COLUMNS,
68  [DIM DROPOFF].[City].[City] ON ROWS
69  FROM [BI DDS]
70  WHERE [Measures].[FACT BILL Count]
71

```

100 % <

	2014	2015	Unknown
Amityville	(null)	(null)	(null)
Arverne	(null)	1	(null)
Astoria	392	198	(null)
Bayside	7	1	(null)
Bellerose Terrace	(null)	2	(null)
Breezy Point	1	(null)	(null)
Bronx	148	101	(null)
Bronxville	(null)	(null)	(null)
Brooklyn	2101	1346	(null)
College Point	3	2	(null)
Corona	19	16	(null)
East Atlantic Beach	1	(null)	(null)
East Brunswick	(null)	(null)	(null)
East Elmhurst	73	23	(null)
East Orange	(null)	(null)	(null)
East Rutherford	(null)	(null)	(null)
Edgewater	(null)	(null)	(null)

- Thống kê số lượng điểm đi trong tháng theo thành phố

```

72  -- THONG KE SO LUONG DIEM DI TRONG THANG THEO THANH PHO
73  SELECT [DO Month].[Month ID].[Month ID] ON COLUMNS,
74  [DIM DROPOFF].[City].[City] ON ROWS
75  FROM [BI DDS]
76  WHERE [Measures].[FACT BILL Count]

```

100 %

	1	2	3	4	5	6	7	8	9	10	11	12	Unknown
Amityville	(null)												
Arverne	(null)	(null)	1	(null)									
Astoria	43	44	65	149	51	49	17	94	(null)	26	(null)	52	(null)
Bayside	(null)	(null)	1	1	1	(null)	(null)	1	(null)	(null)	(null)	4	(null)
Bellerose Terrace	(null)	(null)	(null)	1	1	(null)							
Breezy Point	(null)	1	(null)										
Bronx	20	19	32	69	23	14	7	36	(null)	8	(null)	21	(null)
Bronxville	(null)												
Brooklyn	252	191	425	895	352	225	79	520	(null)	83	(null)	425	(null)
College Point	(null)	(null)	(null)	2	(null)	1	(null)	1	(null)	(null)	(null)	1	(null)
Corona	6	4	1	8	4	5	(null)	3	(null)	2	(null)	2	(null)
East Atlantic Beach	(null)	(null)	1	(null)									
East Brunswick	(null)												
East Elmhurst	7	11	8	10	1	13	7	18	(null)	9	(null)	12	(null)
East Orange	(null)												
East Rutherford	(null)												
Edgewater	(null)												
Elmhurst	5	4	3	15	2	6	3	10	(null)	3	(null)	4	(null)

- Thống kê số lượng điểm đón trong năm theo đường

```

78 -- THONG KE SO LUONG DIEM DON TRONG NAM THEO TEN DUONG
79 SELECT [DO Year].[Yeaarr].[Yeaarr] ON COLUMNS,
80 [DIM PICKUP].[DIM STREET - ID].[DIM STREET - ID] ON ROWS
81 FROM [BI DDS]
82

```

100 %

	2014	2015	Unknown
1	1183.23	607	(null)
2	516.75	470.58	(null)
3	5559.14	2229.07	(null)
4	12989.31	6192.98000000003	(null)
5	2105.71	941.869999999999	(null)
6	8470.79	4646.93000000002	(null)
7	588.71	314.63	(null)
8	8760.29	4743.36000000001	(null)
9	7916.65	3532.15000000001	(null)
10	981.05	648.72	(null)
11	188.35	70	(null)
12	19.4	37.88	(null)
13	3702.6	1969.47	(null)
14	236.5	112.05	(null)
15	19.5	(null)	(null)
16	875.8	418.14	(null)

- Thống kê số lượng điểm đi theo tên đường

```

-- THONG KE SO LUONG DIEM DI TRONG NAM THEO TEN DUONG
84 SELECT [DO Year].[Yeaarr].[Yeaarr] ON COLUMNS,
85 [DIM DROPOFF].[DIM STREET - ID].[DIM STREET - ID] ON ROWS
86 FROM [BI DDS]
87 WHERE [Measures].[FACT BILL Count]
88
89

```

100 %

	2014	2015	Unknown
1	92	64	(null)
2	23	24	(null)
3	411	114	(null)
4	533	253	(null)
5	107	47	(null)
6	395	200	(null)
7	28	21	(null)
8	335	147	(null)
9	401	166	(null)
10	72	17	(null)
11	42	19	(null)
12	3	6	(null)
13	156	36	(null)
14	43	15	(null)
15	(null)	1	(null)
16	62	38	(null)
17	28	20	(null)
18	43	18	(null)

### 3.7. Phân tích chuyển đi theo boro cụ thể

- Thống kê số lượng điểm đón của khu vực theo năm

```
91  -- THONG KE SO LUONG DIEM DON CUA KHU VUC THEO NAM
92  SELECT [DO Year].[Year]. [Year] ON COLUMNS,
93  [DIM PICKUP].[Boro Name].[Boro Name] ON ROWS
94  FROM [BI DDS]
95  WHERE [Measures].[FACT BILL Count]
```

100 % < >

	2014	2015	Unknown
Brooklyn	843	479	(null)
Brox	18	9	(null)
Manhattan	18767	9380	(null)
Queens	655	357	(null)
Staten Island	(null)	(null)	(null)
Unknown	2025	5	(null)

- Thống kê số lượng điểm đón của khu vực theo tháng

```
96
97  -- THONG KE SO LUONG DIEM DON CUA KHU VUC THEO THANG
98  SELECT [DO Month].[Month ID].[Month ID] ON COLUMNS,
99  [DIM PICKUP].[Boro Name].[Boro Name] ON ROWS
100 FROM [BI DDS]
101 WHERE [Measures].[FACT BILL Count]
102
```

100 % < >

	1	2	3	4	5	6	7	8	9	10	11	12	Unknown
Brooklyn	110	94	159	327	112	123	(null)	187	(null)	45	(null)	165	(null)
Brox	(null)	1	3	10	2	4	(null)	2	(null)	1	(null)	4	(null)
Manhattan	2491	2713	3507	6078	2170	3214	(null)	3579	(null)	1730	(null)	2665	(null)
Queens	86	75	123	220	87	128	(null)	143	(null)	60	(null)	90	(null)
Staten Island	(null)												
Unknown	1	(null)	3	4	1	4	2001	10	(null)	1	(null)	5	(null)

- Thống kê số lượng điểm đón của khu vực theo ngày

```

103  -- THONG KE SO LUONG DIEM DON CUA KHU VUC THEO NGAY
104  SELECT [DO Day].[Day ID].[Day ID] ON ROWS,
105  [DIM PICKUP].[Boro Name].[Boro Name] ON COLUMNS
106  FROM [BI DDS]
107  WHERE [Measures].[FACT BILL Count]

```

100 %

Messages Results

	Brooklyn	Brox	Manhattan	Queens	Staten Island	Unknown
1	78	2	1717	52	(null)	(null)
2	52	1	997	25	(null)	2
3	32	2	928	21	(null)	2
4	18	(null)	340	7	(null)	(null)
5	5	(null)	67	5	(null)	(null)
6	4	(null)	97	10	(null)	(null)
7	51	1	1130	39	(null)	447
8	41	1	701	28	(null)	945
9	53	1	821	33	(null)	607
10	4	(null)	60	1	(null)	4
11	39	(null)	2115	65	(null)	2
12	33	1	879	30	(null)	1
13	48	(null)	1203	53	(null)	1
14	141	6	2540	90	(null)	2
15	44	(null)	960	46	(null)	2
16	(null)	(null)	7	1	(null)	(null)
17	62	2	1275	39	(null)	3
18	(null)	(null)	102	4	(null)	(null)
19	25	1	570	24	(null)	(null)
20	51	1	1105	62	(null)	1
21	70	1	1630	43	(null)	(null)
22	1	(null)	5	(null)	(null)	(null)
23	(null)	(null)	49	1	(null)	(null)
24	53	1	833	17	(null)	1
25	15	(null)	852	43	(null)	4
26	57	(null)	928	45	(null)	2
27	87	3	1457	98	(null)	1
28	62	(null)	1039	35	(null)	2
29	82	3	1792	49	(null)	(null)
30	64	(null)	1036	26	(null)	(null)
31	50	(null)	912	20	(null)	1
Unknown	(null)	(null)	(null)	(null)	(null)	(null)

- Thống kê số lượng điểm trả của khu vực theo năm

```

108
109 -- THONG KE SO LUONG DIEM TRA CUA KHU VUC THEO NAM
110 SELECT [DO Year].[Year].[Year] ON COLUMNS,
111 [DIM DROPOFF].[Boro Name].[Boro Name] ON ROWS
112 FROM [BI DDS]
113 WHERE [Measures].[FACT BILL Count]
114

```

100 %

Messages Results

	2014	2015	Unknown
Brooklyn	1921	1346	(null)
Brox	127	101	(null)
Manhattan	18988	8108	(null)
Queens	1187	637	(null)
Staten Island	6	3	(null)
Unknown	79	35	(null)

- Thống kê số lượng điểm trả của khu vực theo tháng

```

115 -- THONG KE SO LUONG DIEM TRA CUA KHU VUC THEO THANG
116 SELECT [DO Month].[Month ID].[Month ID] ON COLUMNS,
117 [DIM DROPOFF].[Boro Name].[Boro Name] ON ROWS
118 FROM [BI DDS]
119 WHERE [Measures].[FACT BILL Count]
120

```

100 %

Messages Results

	1	2	3	4	5	6	7	8	9	10	11	12	Unknown
Brooklyn	252	191	293	895	352	225	31	520	(null)	83	(null)	425	(null)
Brox	20	19	18	69	23	14	2	34	(null)	8	(null)	21	(null)
Manhattan	2272	2537	3310	5238	1823	3060	1891	3048	(null)	1648	(null)	2269	(null)
Queens	141	132	166	411	163	159	77	285	(null)	89	(null)	201	(null)
Staten Island	(null)	2	(null)	(null)	3	2	(null)	(null)	(null)	(null)	(null)	2	(null)
Unknown	3	2	8	26	8	13	(null)	34	(null)	9	(null)	11	(null)

- Thống kê số lượng điểm trả của khu vực theo ngày

```

121 -- THONG KE SO LUONG DIEM TRA CUA KHU VUC THEO NGAY
122 SELECT [DO Day].[Day ID].[Day ID] ON ROWS,
123 [DIM DROPOFF].[Boro Name].[Boro Name] ON COLUMNS
124 FROM [BI DDS]
125 WHERE [Measures].[FACT BILL Count]

```

100 %

Messages Results

	Brooklyn	Brox	Manhattan	Queens	Staten Island	Unknown
1	187	17	1549	90	(null)	6
2	96	7	932	39	(null)	3
3	80	3	844	53	(null)	5
4	23	2	321	19	(null)	(null)
5	14	1	55	7	(null)	(null)
6	8	2	89	12	(null)	(null)
7	87	11	1479	86	1	4
8	72	5	1582	56	(null)	1
9	102	9	1296	107	1	(null)
10	15	(null)	53	1	(null)	(null)
11	122	8	2004	77	2	8
12	57	4	826	53	(null)	4
13	133	6	1083	78	(null)	5
14	286	19	2331	135	(null)	8
15	112	7	847	79	1	6
16	(null)	1	7	(null)	(null)	(null)
17	195	14	1073	93	1	5
18	10	1	86	9	(null)	(null)
19	81	3	489	45	1	1
20	163	15	954	83	(null)	5
21	151	7	1499	83	2	2
22	1	(null)	5	(null)	(null)	(null)
23	1	(null)	48	1	(null)	(null)
24	117	13	711	58	(null)	6
25	42	8	786	64	(null)	14
26	161	5	769	91	(null)	6
27	257	19	1237	126	(null)	7
28	176	16	854	84	(null)	8
29	208	9	1630	76	(null)	3
30	183	11	860	68	(null)	4
31	127	5	797	51	(null)	3
Unknown	(null)	(null)	(null)	(null)	(null)	(null)

- Thống kê số lượng điểm trả của khu vực theo giờ

```

127 -- THONG KE SO LUONG DIEM TRA CUA KHU VUC THEO GIO
128 SELECT [DO Year].[Year]. [Year] ON COLUMNS,
129 [DIM DROPOFF].[Boro Name].[Boro Name] ON ROWS
130 FROM [BI DDS]
131 WHERE [Measures].[FACT BILL Count]

```

100 %

	2014	2015	Unknown
Brooklyn	1921	1346	(null)
Brox	127	101	(null)
Manhattan	18988	8108	(null)
Queens	1187	637	(null)
Staten Island	6	3	(null)
Unknown	79	35	(null)

### 3.8. Phân tích chuyển đổi theo loại thanh toán

- Thống kê số lượng loại thanh toán trong năm

```

134 -- THONG KE SO LUONG LOAI THANH TOAN TRONG NAM
135 SELECT [DO Year].[Year]. [Year] ON COLUMNS,
136 [DIM PAYMENT].[Payment Type].[Payment Type] ON ROWS
137 FROM [BI DDS]
138 WHERE [Measures].[FACT BILL Count]
139

```

100 %

	2014	2015	Unknown
Cash	7617	3508	(null)
Credit card	14560	6677	(null)
Dispute	84	12	(null)
No charge	20	33	(null)
Unknown	27	(null)	(null)
Voided trip	(null)	(null)	(null)
Unknown	(null)	(null)	(null)

- Thống kê doanh thu của loại thanh toán trong năm

```

140 -- THONG KE DOANH THU LOAI THANH TOAN TRONG NAM
141 SELECT [DO Year].[Year]. [Year] ON COLUMNS,
142 [DIM PAYMENT].[Payment Type].[Payment Type] ON ROWS
143 FROM [BI DDS]
144 WHERE [Measures].[Total Amount]
145

```

100 %

	2014	2015	Unknown
Cash	88914.1500000002	46188.0000000015	(null)
Credit card	241448.05	113942.0900000004	(null)
Dispute	973.49	245.37	(null)
No charge	252	449.99	(null)
Unknown	327.74	(null)	(null)
Voided trip	(null)	(null)	(null)
Unknown	(null)	(null)	(null)

- Thống kê số lượng các loại thanh toán trong 2 năm

```

146 -- THONG KE SO LUONG CAC LOAI THANH TOAN TRONG CA 2 NAM
147 SELECT [DIM PAYMENT].[Payment Type].[Payment Type] ON COLUMNS
148 FROM [BI DDS]
149 WHERE [Measures].[FACT BILL Count]
150
100 %

```

Cash	Credit card	Dispute	No charge	Unknown	Voided trip	Unknown
11125	21237	96	53	27	(null)	(null)

- Thống kê số lượng các loại thanh toán trong 2 năm của khu vực

```

145 -- THONG KE SO LUONG CAC LOAI THANH TOAN TRONG 2 NAM CUA 1 KHU VUC
146 SELECT [DIM PAYMENT].[Payment Type].[Payment Type] ON COLUMNS,
147 [DIM PICKUP].[Boro Name].[Boro Name] ON ROWS
148 FROM [BI DDS]
149 WHERE [Measures].[FACT BILL Count]
150
100 %

```

	Cash	Credit card	Dispute	No charge	Unknown	Voided trip	Unknown
Brooklyn	385	929	2	4	2	(null)	(null)
Brox	12	15	(null)	(null)	(null)	(null)	(null)
Manhattan	8388	19640	57	37	25	(null)	(null)
Queens	369	627	12	4	(null)	(null)	(null)
Staten Island	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Unknown	1971	26	25	8	(null)	(null)	(null)

- Thống kê doanh thu các loại thanh toán trong 2 năm của 1 khu vực

```

150
157 -- THONG KE DOANH THU CAC LOAI THANH TOAN TRONG 2 NAM CUA 1 KHU VUC
158 SELECT [DIM PAYMENT].[Payment Type].[Payment Type] ON COLUMNS,
159 [DIM PICKUP].[Boro Name].[Boro Name] ON ROWS
160 FROM [BI DDS]
161 WHERE [Measures].[Total Amount]
162
100 %

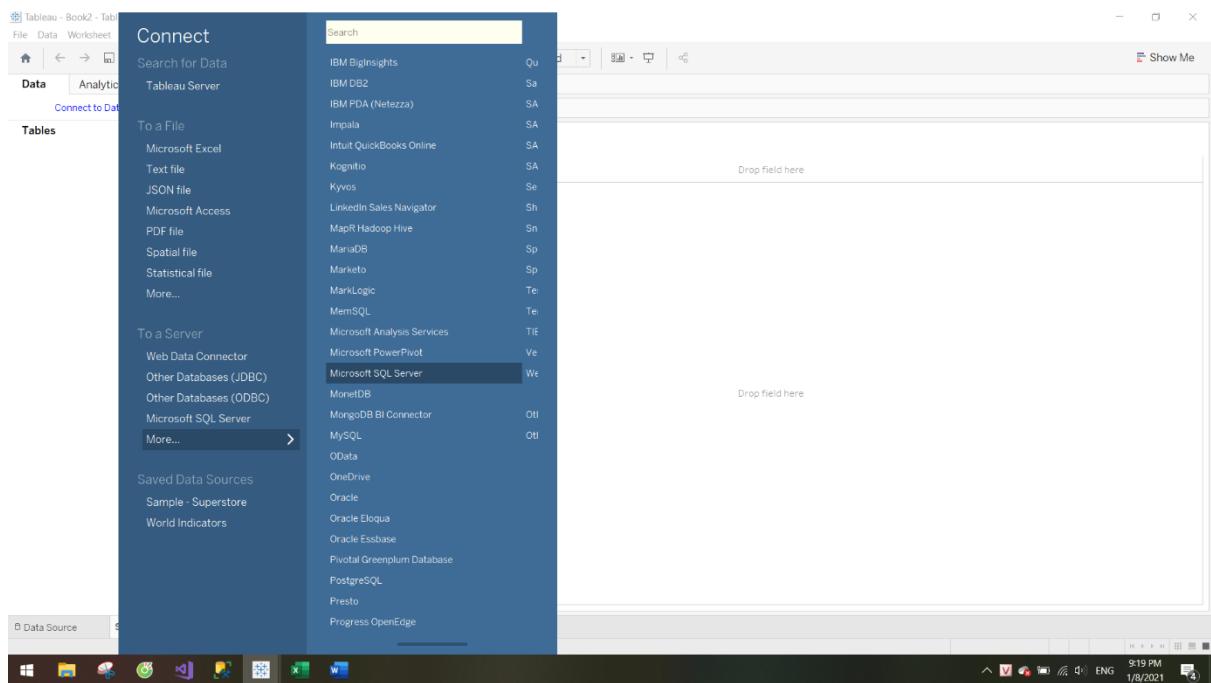
```

	Cash	Credit card	Dispute	No charge	Unknown	Voided trip	Unknown
Brooklyn	4991.46000000003	15987.2699999999	47	41.2	27.64	(null)	(null)
Brox	153.4	273.73	(null)	(null)	(null)	(null)	(null)
Manhattan	96833.4100000091	309579.049999974	743.59	488.39	300.1	(null)	(null)
Queens	10166.97	27122.7499999999	241.77	73.9	(null)	(null)	(null)
Staten Island	(null)	(null)	(null)	(null)	(null)	(null)	(null)
Unknown	22956.91	2427.34	186.5	98.5	(null)	(null)	(null)

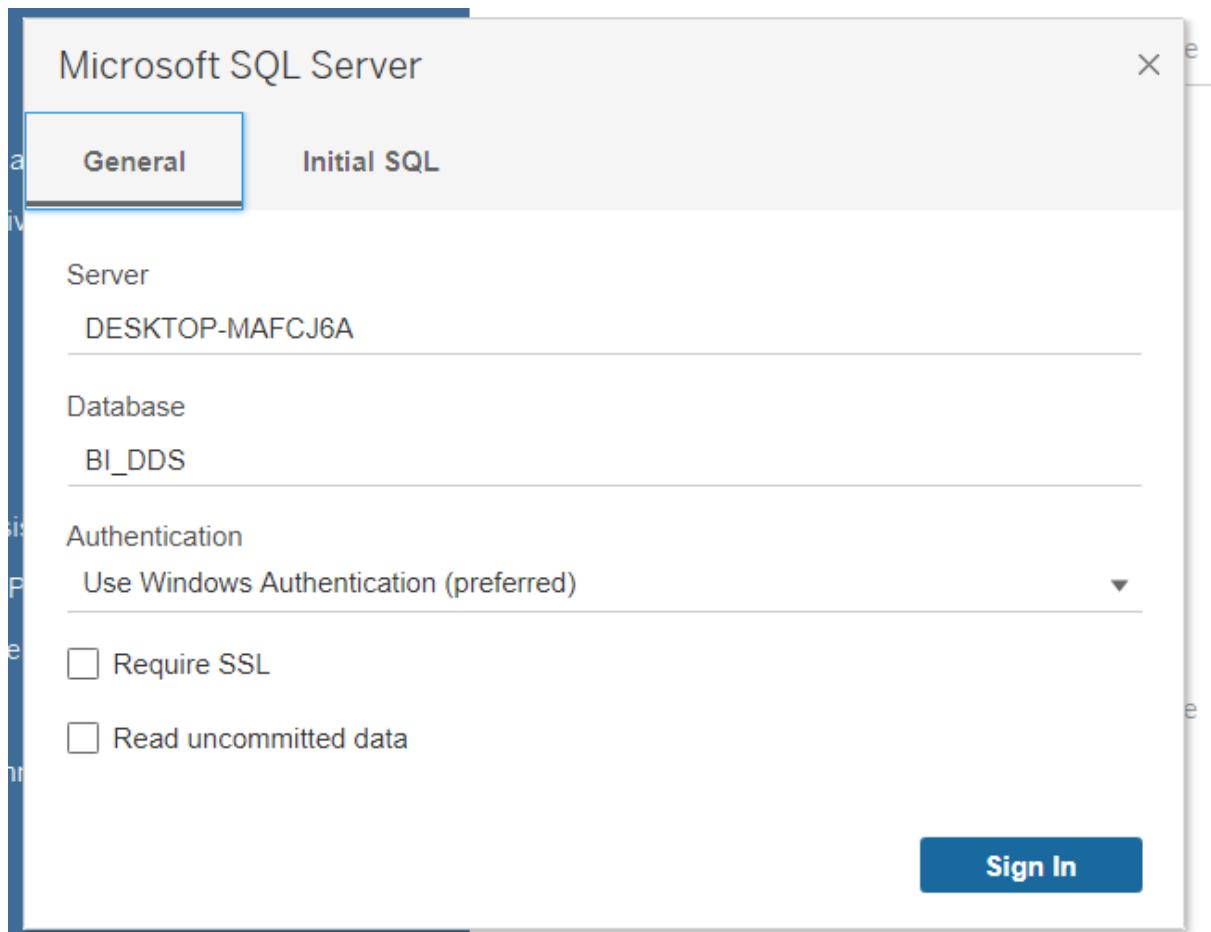
## 4. Mining

## 5. KPI

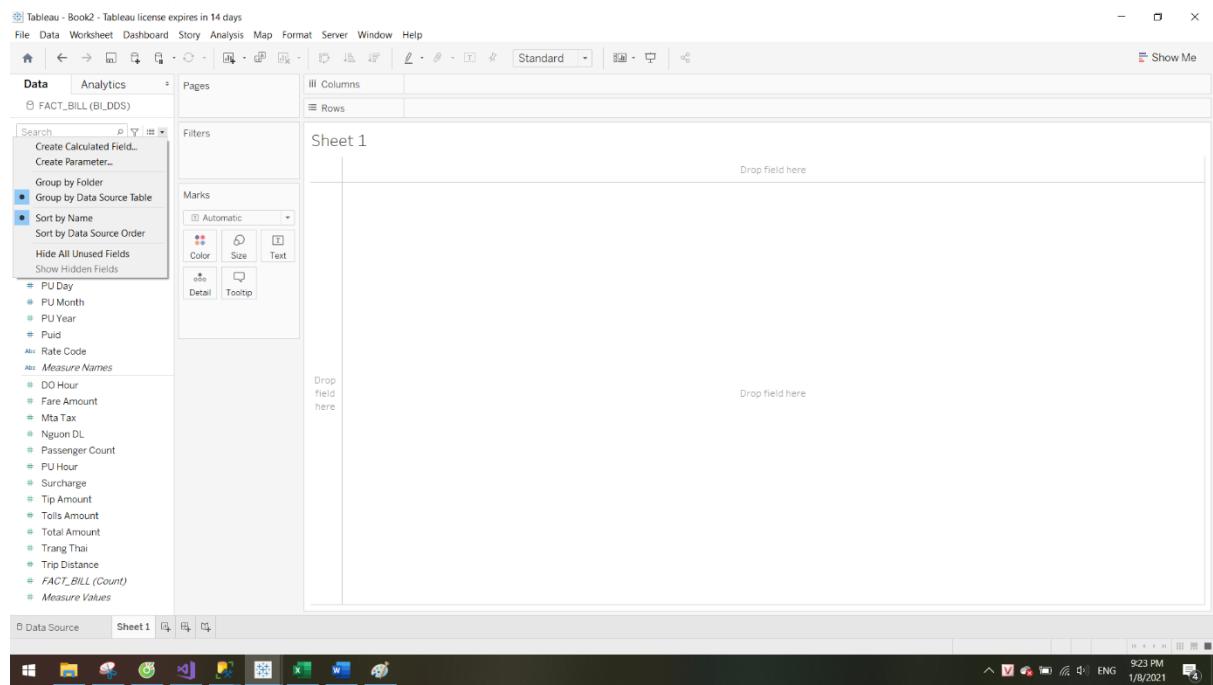
## Bước 1: Connect đến SQL Server



## Bước 2: Kết nối đến database

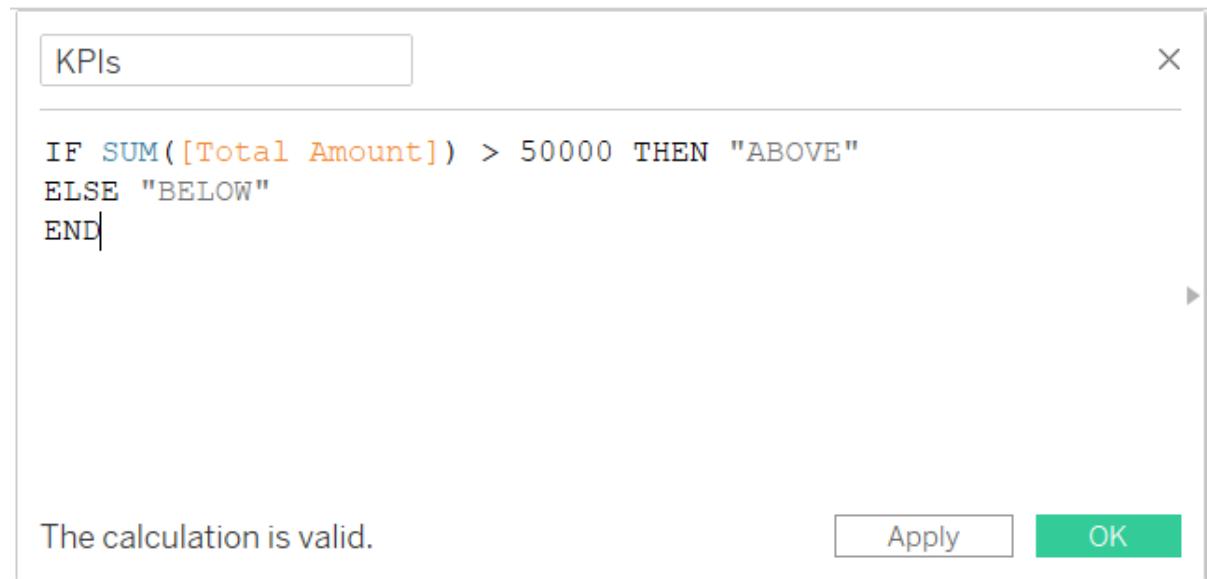


### Bước 3: Thêm Fact\_Bill vào sheet



The screenshot shows the Tableau Data Source interface. On the left, a sidebar lists various fields and measures from the FACT\_BILL table, such as PU Day, PU Month, PU Year, Puid, Rate Code, Measure Names, Measure Values, DO Hour, Fare Amount, Mta Tax, Nguon DL, Passenger Count, PU Hour, Surcharge, Tip Amount, Tolls Amount, Total Amount, Trang Thai, Trip Distance, FACT\_BILL (Count), and FACT\_BILL (Sum). A radio button next to 'Sort by Name' is selected. The main workspace is titled 'Sheet 1' and contains two empty 'Drop field here' areas.

### Bước 4: Tạo KPI và ràng buộc giá trị cho KPI



The screenshot shows the Tableau KPI creation dialog. The title bar says 'KPIs'. The main area contains the following DAX code:

```
IF SUM([Total Amount]) > 50000 THEN "ABOVE"  
ELSE "BELOW"  
END
```

Below the code, a message says 'The calculation is valid.' There are 'Apply' and 'OK' buttons at the bottom right.

Và đây là kết quả của tổng doanh thu theo tháng, với màu xanh là đạt KPI và đỏ là chưa đạt.

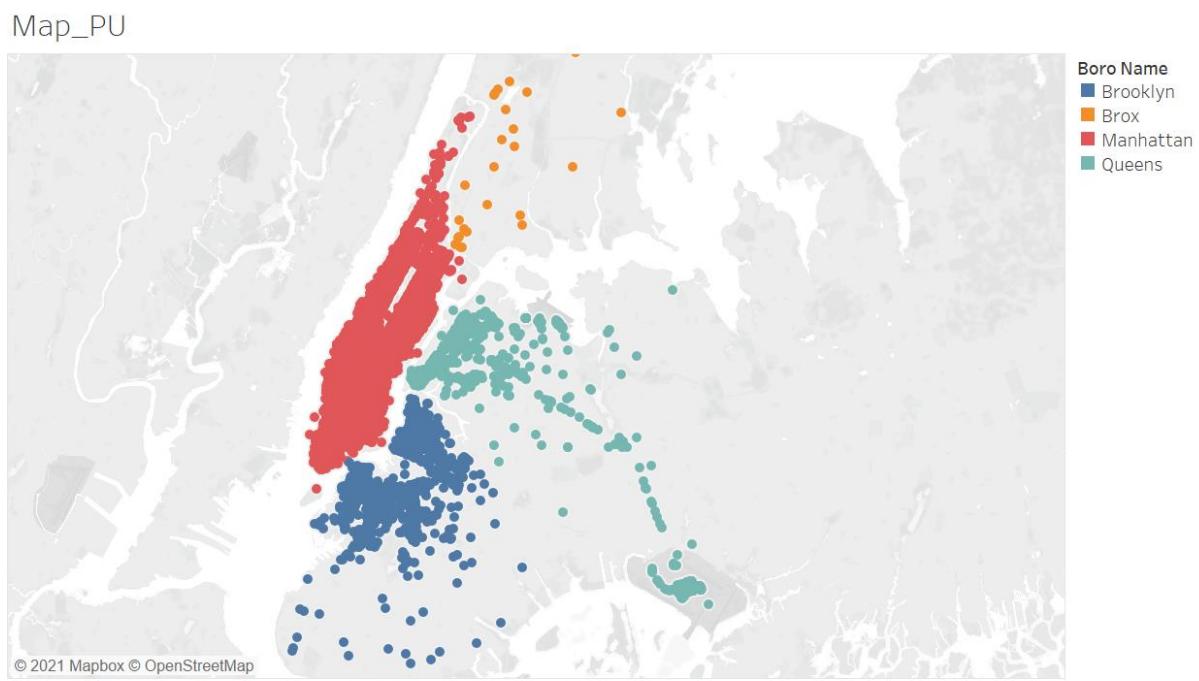


## 6. Report

Ta tiến hành tạo các biểu đồ báo cáo thông qua Tableau với dữ liệu từ DDS

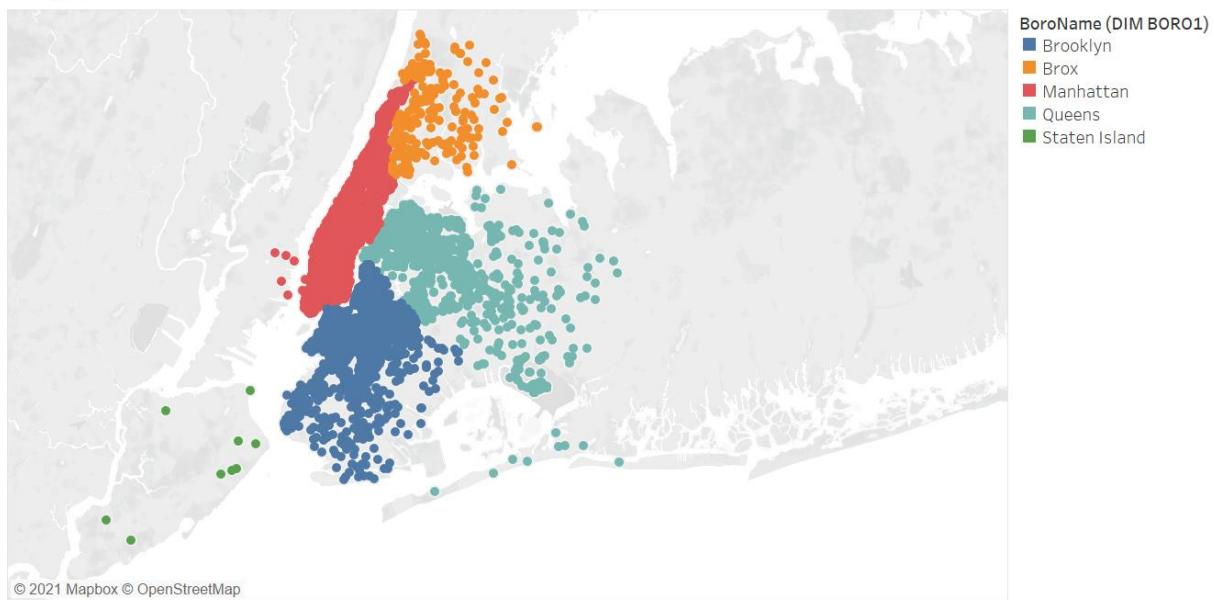
### 6.1. Phân tích các điểm đón và trả khách bằng regional map theo boro

- Các điểm đón khách



- Các điểm trả khách

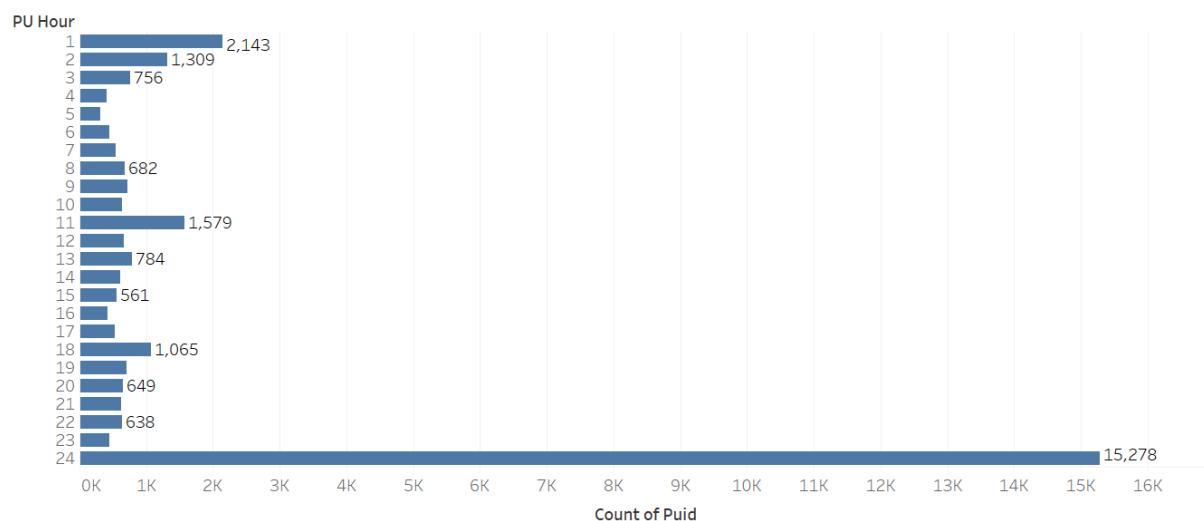
Map\_DO



## 6.2. Phân tích giờ, ngày cao điểm đón khách

- Giờ cao điểm

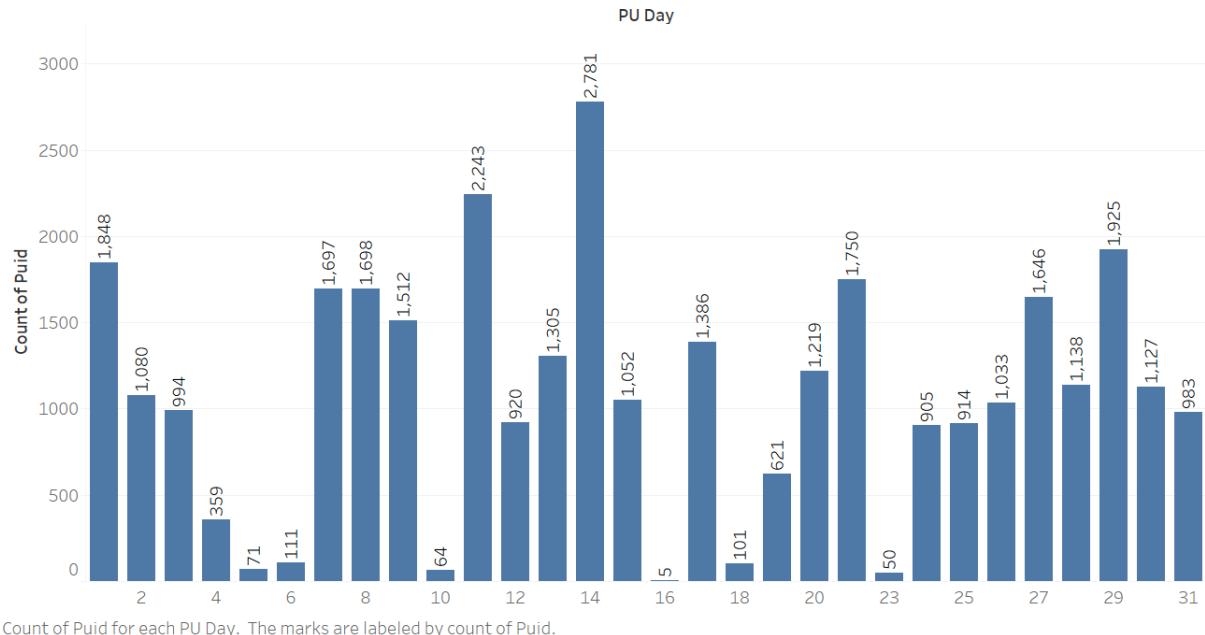
Gio\_cao\_diem



Count of Puid for each PU Hour. The marks are labeled by count of Puid.

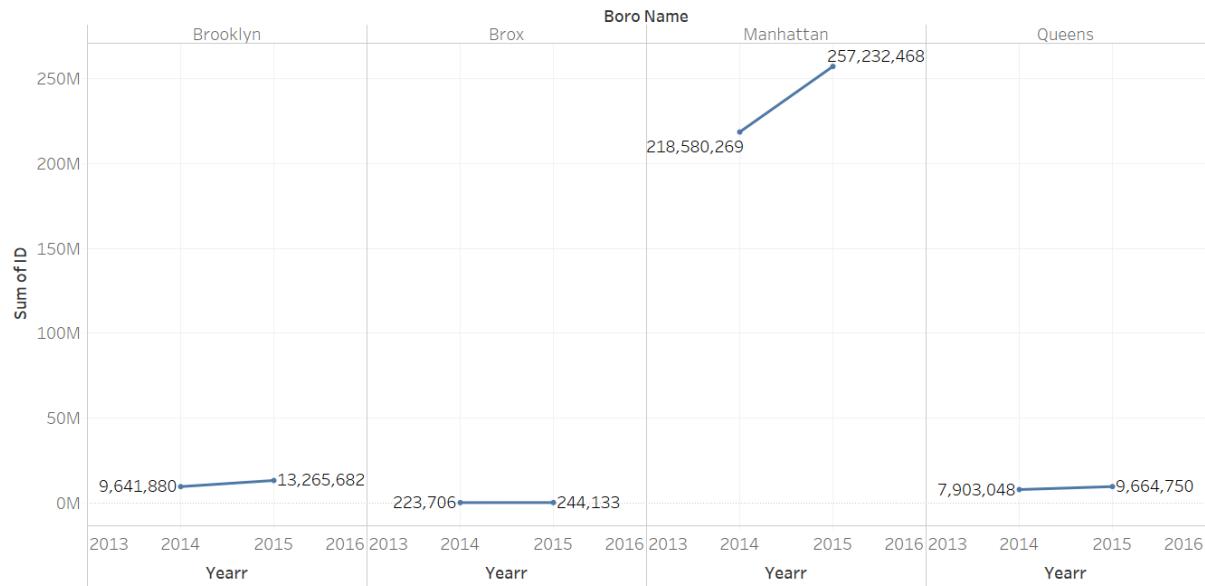
- Ngày cao điểm

## Ngay\_cao\_diem



## 6.3. Phân tích số lượng đón xe theo boro theo năm

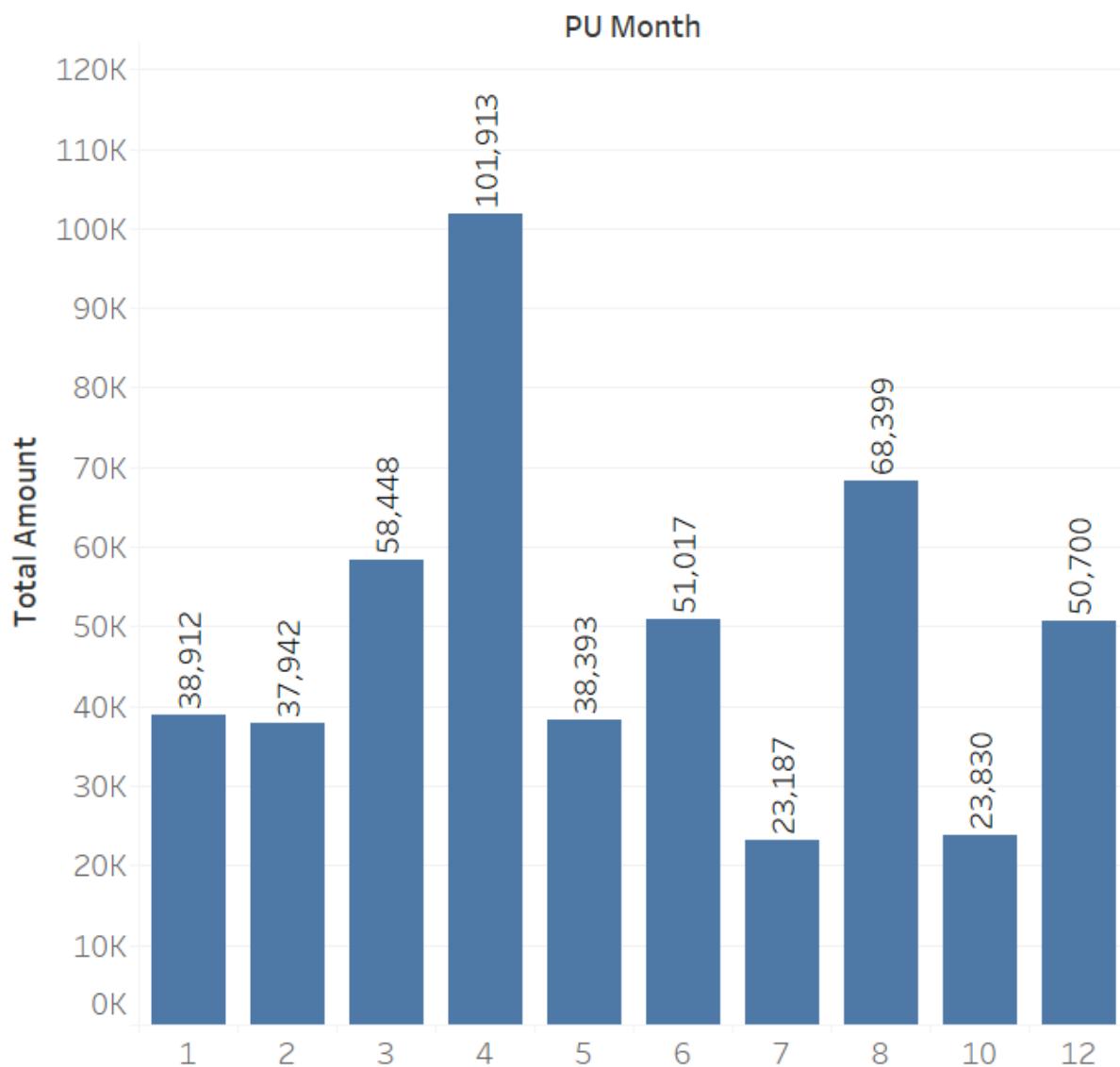
### Tong\_chuyen\_di\_theo\_boro\_theo\_nam



## 6.4. Phân tích tổng doanh thu theo tháng, năm

- Theo tháng

## Doanh\_thu\_theo\_thang

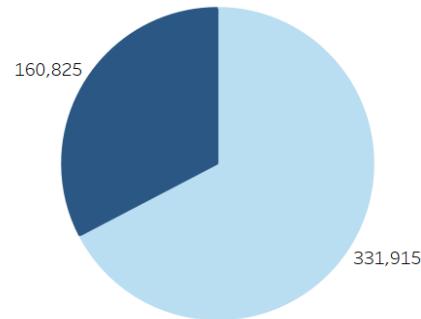


Sum of Total Amount for each PU Month. The marks are labeled by sum of Total Amount.

- Theo năm

## Doanh\_thu\_theo\_nam

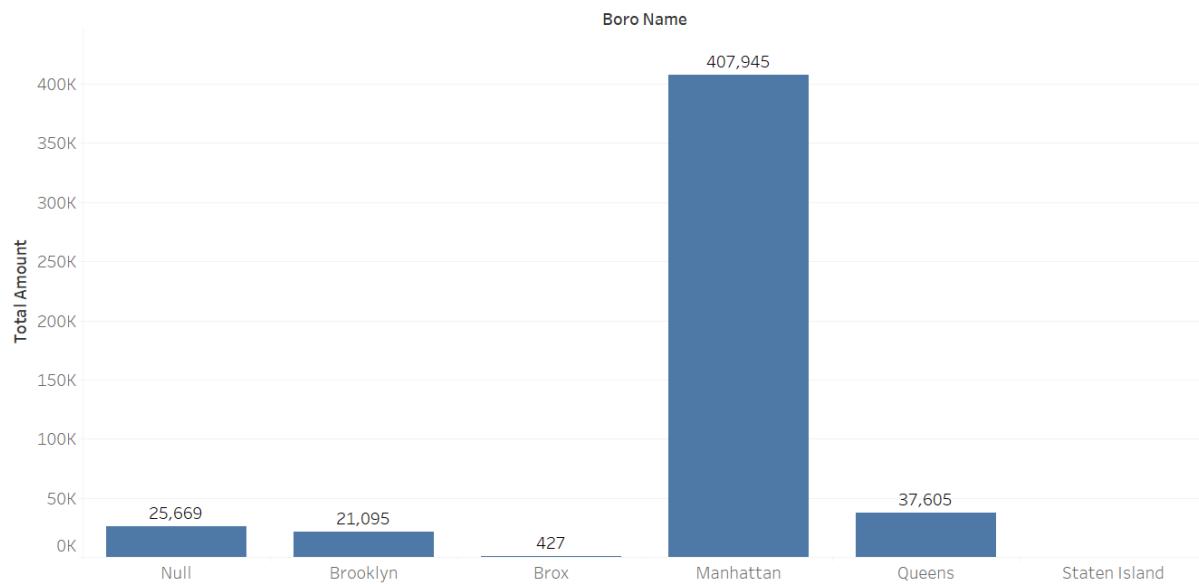
Year  
2014 2015



Sum of Total Amount. Color shows details about Year. The marks are labeled by sum of Total Amount.

## 6.5. Phân tích tổng doanh thu theo boro

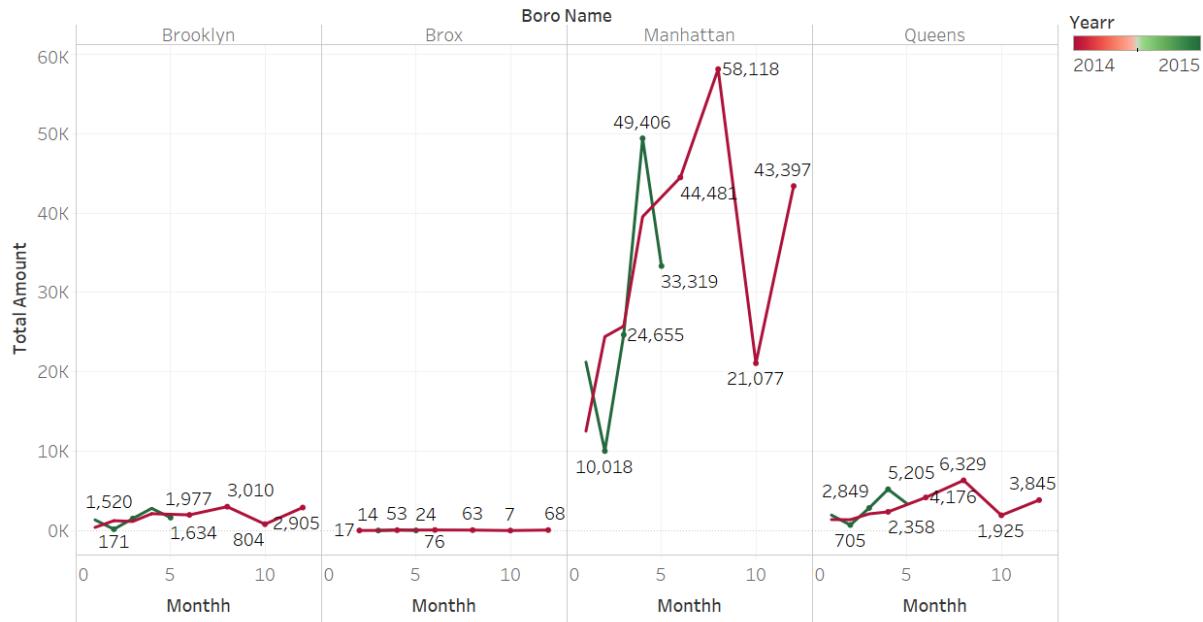
### Doanh\_thu\_theo\_boro



Sum of Total Amount for each Boro Name. The marks are labeled by sum of Total Amount.

- Theo tháng

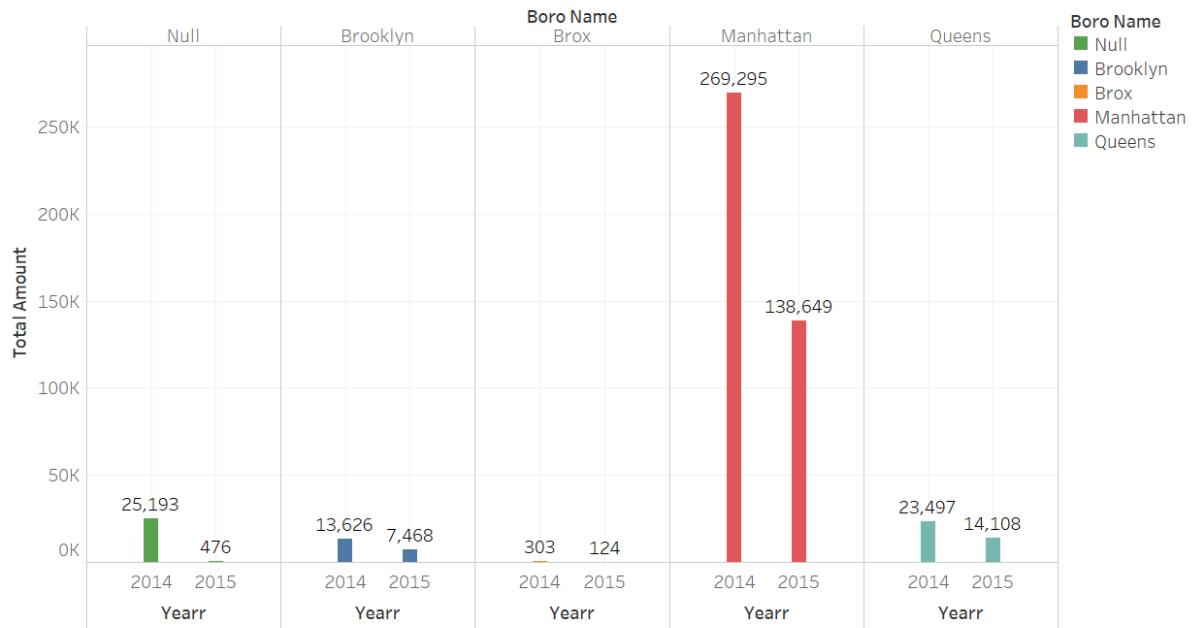
## Doanh\_thu\_theo\_boro\_theo\_thang



The trend of sum of Total Amount for Monthh broken down by Boro Name. Color shows details about Yearr. The marks are labeled by sum of Total Amount. The view is filtered on Boro Name and Monthh. The Boro Name filter keeps Brooklyn, Brox, Manhattan, Queens and Staten Island. The Monthh filter keeps non-Null values only.

- Theo năm

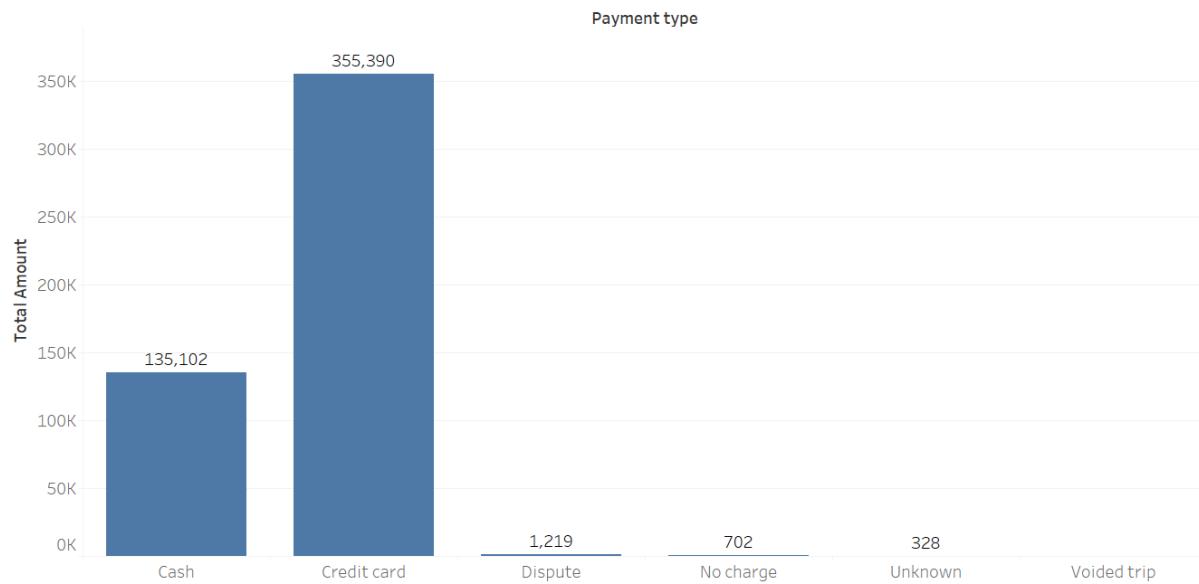
## Doanh\_thu\_theo\_boro\_theo\_nam



The plot of sum of Total Amount for Yearr broken down by Boro Name. Color shows details about Boro Name. The marks are labeled by sum of Total Amount. The view is filtered on Boro Name, which keeps 6 of 6 members.

## 6.6. Phân tích tổng doanh thu các loại thanh toán

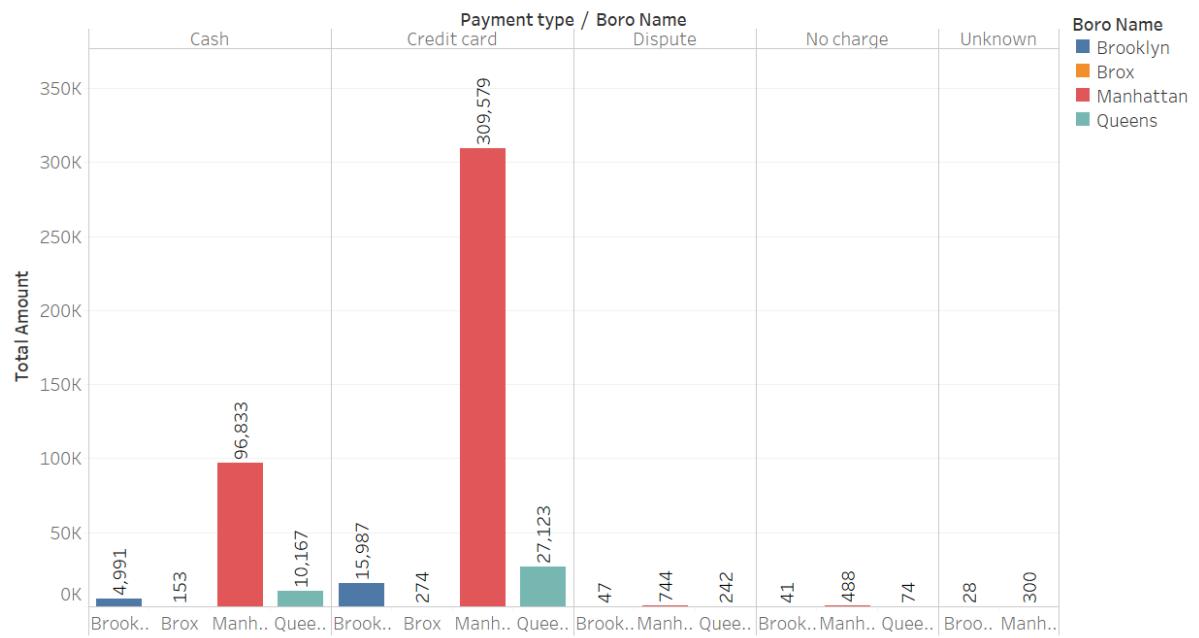
## Doanh\_thu\_theo\_payment



Sum of Total Amount for each Payment type. The marks are labeled by sum of Total Amount.

- Theo boro

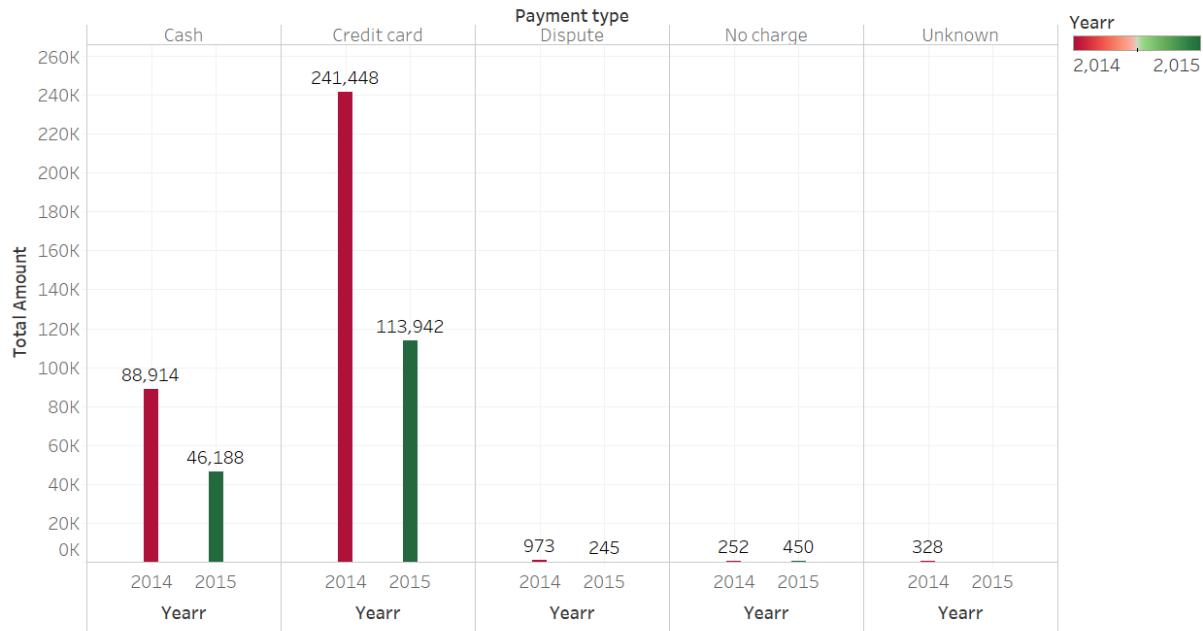
## Doanh\_thu\_theo\_payment\_theo\_boro



Sum of Total Amount for each Boro Name broken down by Payment type. Color shows details about Boro Name. The marks are labeled by sum of Total Amount. The view is filtered on Boro Name, which keeps Brooklyn, Bronx, Manhattan, Queens and Staten Island.

- Theo năm

## Doanh\_thu\_theo\_payment\_theo\_nam

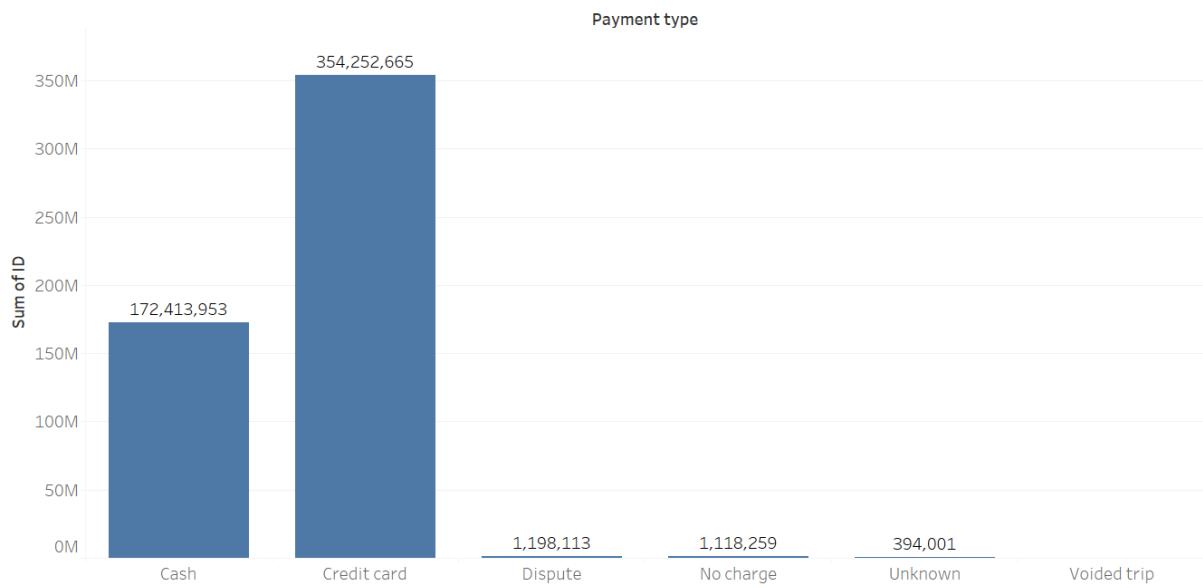


The plot of sum of Total Amount for Yearr broken down by Payment type. Color shows sum of Yearr. The marks are labeled by sum of Total Amount.

## 6.7. Phân tích tổng chuyển đi

- Theo loại thanh toán

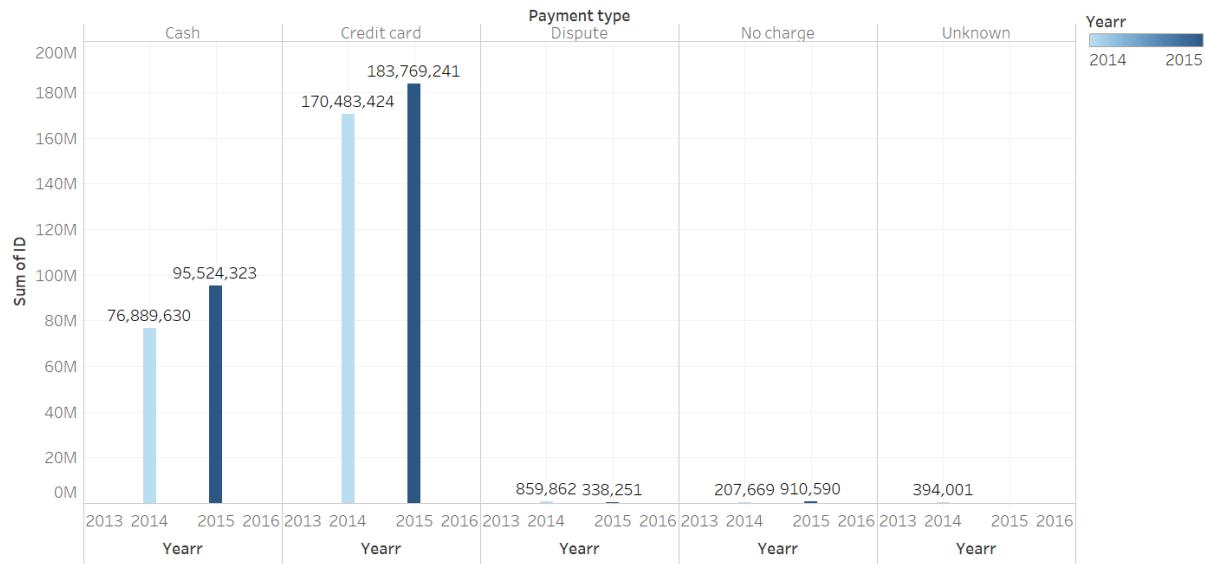
### Tong\_chuyen\_di\_theo\_payment



Sum of ID for each Payment type. The marks are labeled by sum of ID.

- Theo năm

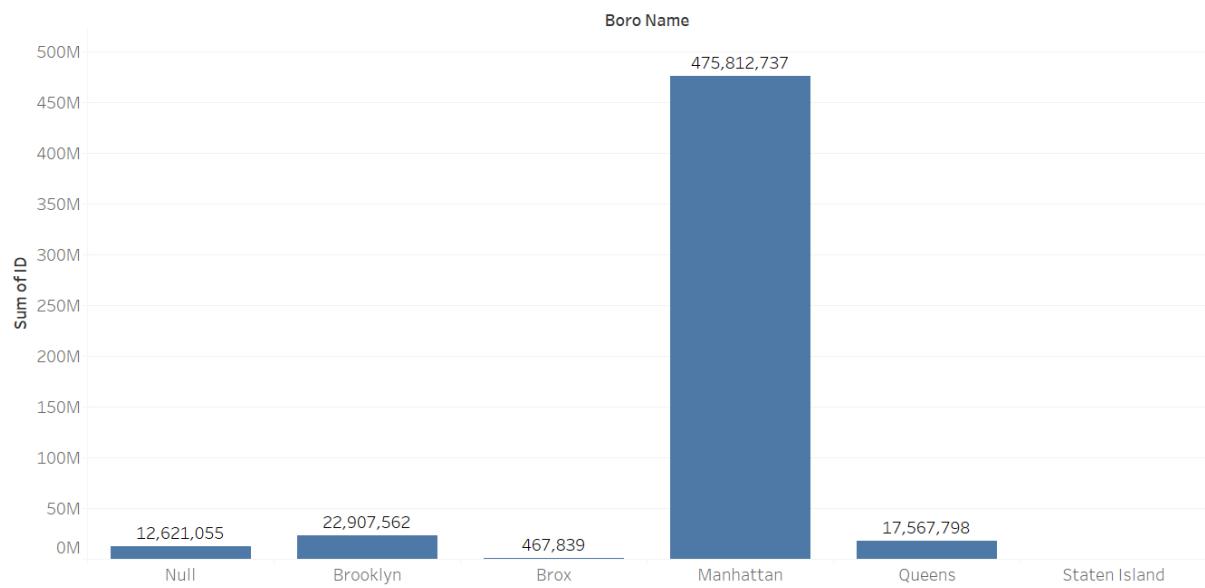
## Tong\_chuyen\_di\_theo\_payment\_theo\_nam



The plot of sum of ID for Yearrr broken down by Payment type. Color shows details about Yearrr. The marks are labeled by sum of ID.

- Theo boro

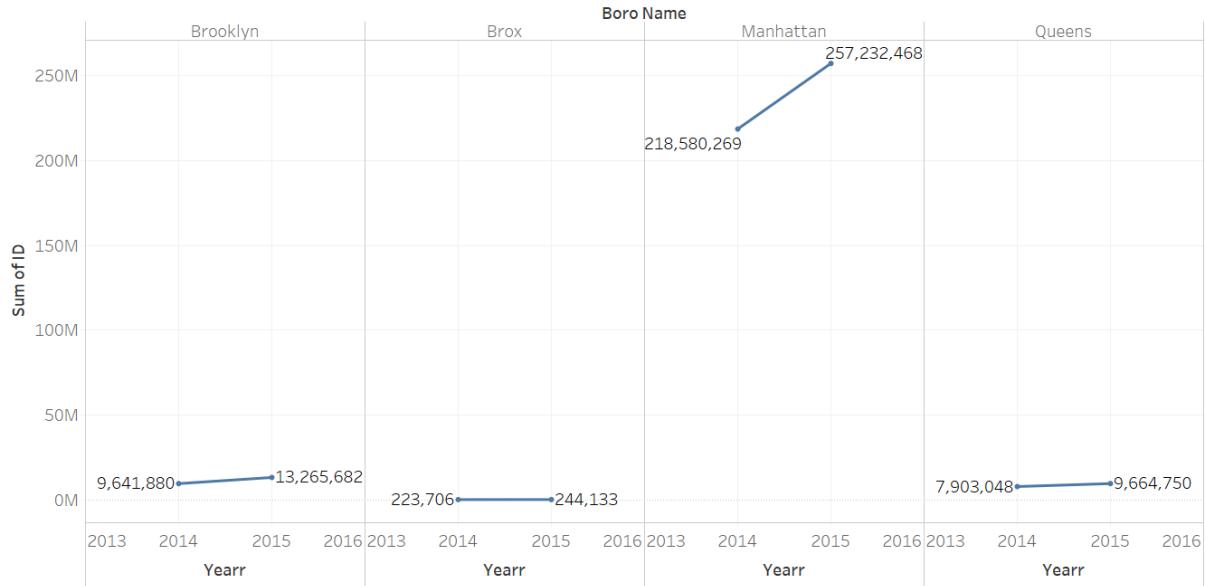
## Tong\_chuyen\_di\_theo\_boro



Sum of ID for each Boro Name. The marks are labeled by sum of ID.

- Theo năm

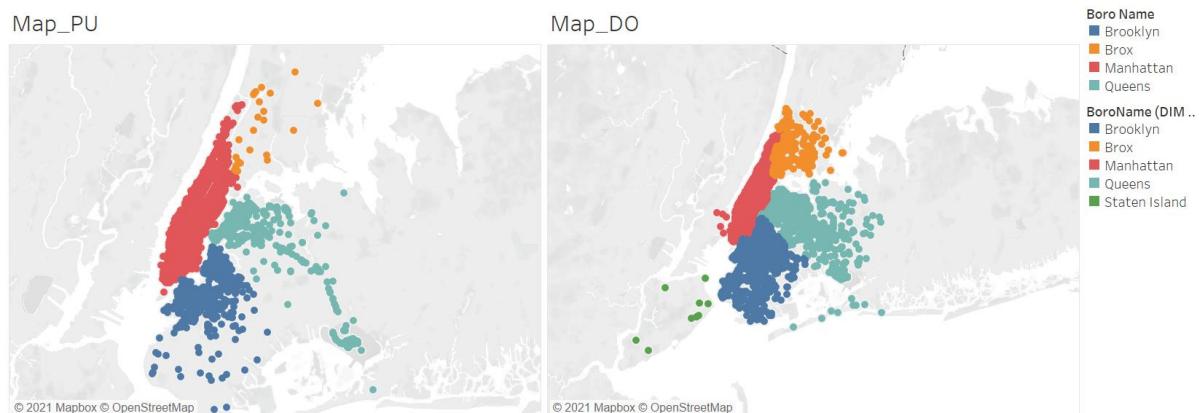
Tong\_chuyen\_di\_theo\_boro\_theo\_nam



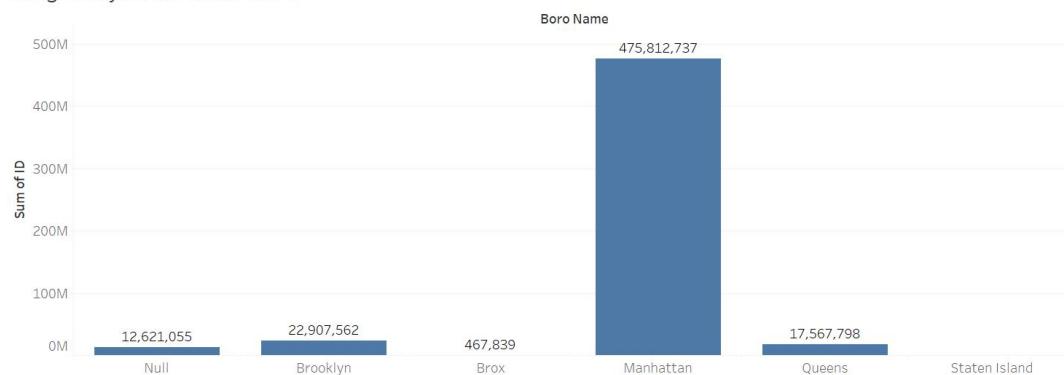
The trend of sum of ID for Yearr broken down by Boro Name. The marks are labeled by sum of ID. The view is filtered on Boro Name, which keeps Brooklyn, Bronx, Manhattan, Queens and Staten Island.

## 6.8. Dashborad report

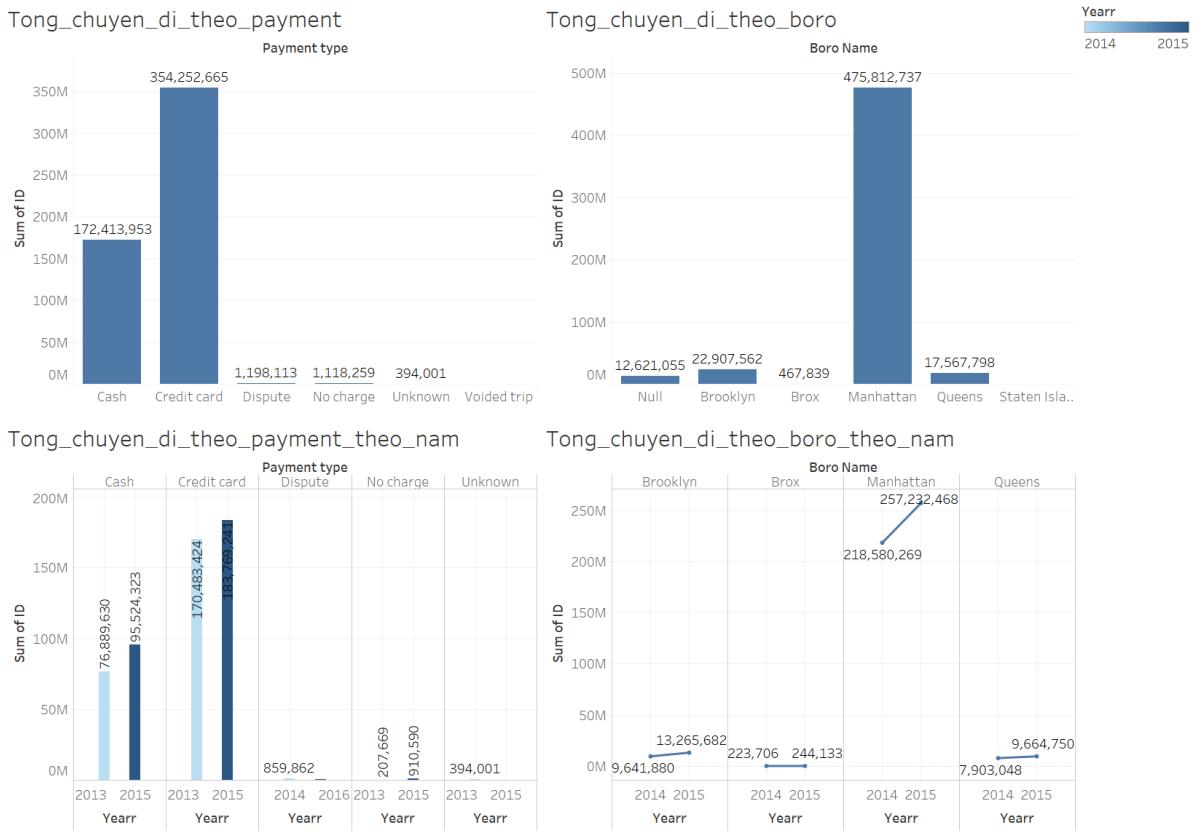
- Chuyển đi theo boro



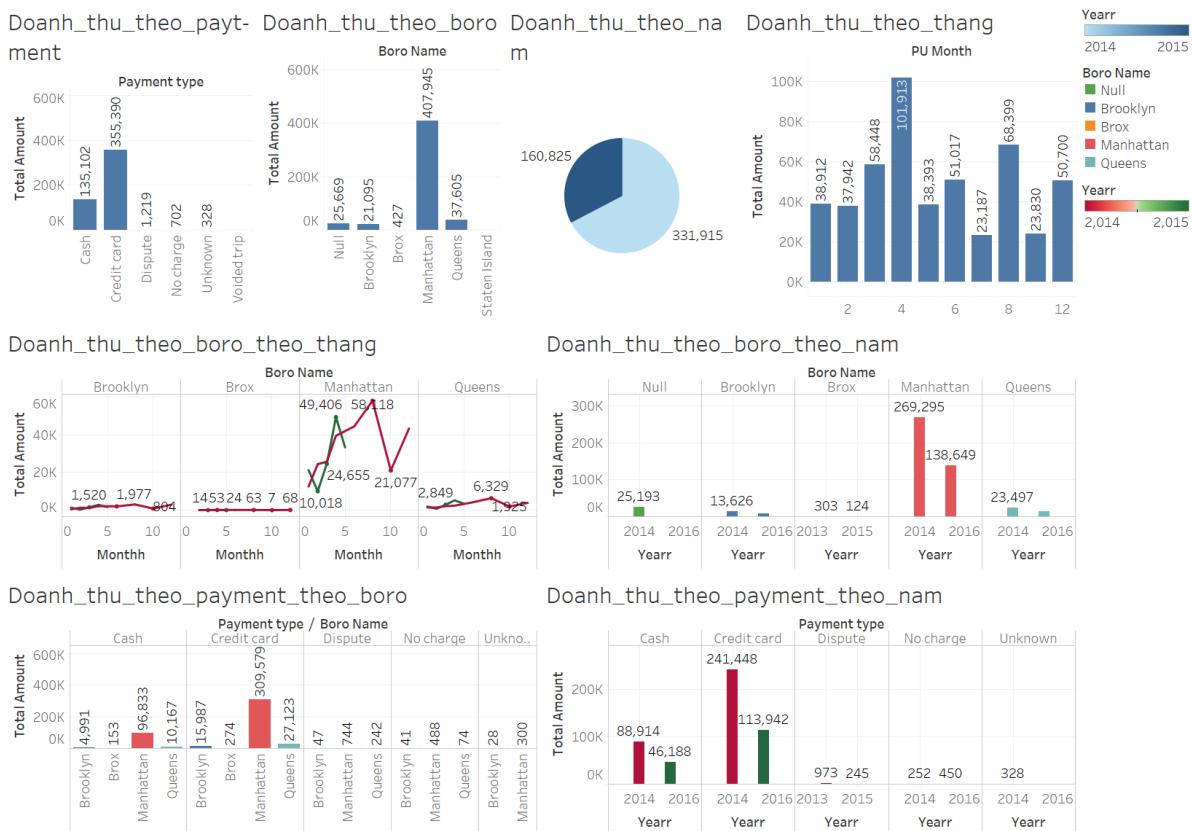
Tong\_chuyen\_di\_theo\_boro



- Phân tích tổng chuyển đi



- Phân tích doanh thu



---HẾT---