**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**

|  |  |  |  |
| --- | --- | --- | --- |
| **Mã nhóm:** CQ-BI-10 | | | |
| **STT** | **MSSV** | **Họ và tên** | **Email** |
| 1 | 1712769 | Trịnh Đức Thanh | [1712769@student.hcmus.edu.vn](mailto:1712769@student.hcmus.edu.vn) |
| 2 | 1712828 | Huỳnh Thanh Khải Trân | [1712828@student.hcmus.edu.vn](mailto:1712828@student.hcmus.edu.vn) |
| 3 | 1712899 | Dương Khánh Vi | [1712899@student.hcmus.edu.vn](mailto:1712899@student.hcmus.edu.vn) |
| 4 | 1712926 | Lương Tường Vy | [1712926@student.hcmus.edu.vn](mailto:1712926@student.hcmus.edu.vn) |

**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](#_Toc61347736)

[B- KẾT QUẢ 2](#_Toc61347737)

[I. Project description 2](#_Toc61347738)

[II. Key Deliverables 2](#_Toc61347739)

[III. Critical Success Factors 2](#_Toc61347740)

[IV. Risks and Concerns 3](#_Toc61347741)

[V. Business Requirements 3](#_Toc61347742)

[1. Tiền tổng mỗi chuyến đi 3](#_Toc61347743)

[2. Thống kê doanh thu năm theo quận 4](#_Toc61347744)

[3. Phân tích dữ liệu chuyến đi 7](#_Toc61347745)

[4. Thống kê các loại dữ liệu theo tháng, quý, năm 9](#_Toc61347746)

[VI. Source data description 12](#_Toc61347747)

[1. Dữ liệu của nguồn Yellow Taxi năm 2014 12](#_Toc61347748)

[2. Dữ liệu của nguồn Yellow Taxi năm 2015 13](#_Toc61347749)

[3. Dữ liệu của nguồn Yellow Taxi năm 2016 15](#_Toc61347750)

[4. Dữ liệu của nguồn Census Block 16](#_Toc61347751)

[VII. BI Dimensional Logical Model Design 17](#_Toc61347752)

[a) Stage 17](#_Toc61347753)

[b) NDS 21](#_Toc61347754)

[3. DDS 27](#_Toc61347755)

[4. Thiết kế Data Flow 32](#_Toc61347756)

[VIII. ETL 33](#_Toc61347757)

[1. Source to Stage 33](#_Toc61347758)

[2. Stage to NDS 39](#_Toc61347759)

[3. NDS to DDS 48](#_Toc61347760)

[IX. Khai thác dữ liệu 51](#_Toc61347761)

[1. OLAP 51](#_Toc61347762)

[2. Cube 58](#_Toc61347763)

[3. Truy vấn bằng MDX 69](#_Toc61347764)

[4. Mining 85](#_Toc61347765)

[5. KPI 85](#_Toc61347766)

[6. Report 88](#_Toc61347767)

**YÊU CẦU ĐỒ ÁN**

|  |  |
| --- | --- |
| **Loại bài tập** | **🗹 Lý thuyết Thực hành 🗹 Đồ án Bài tập** |
| **Ngày bắt đầu** | **28/09/2020** |
| **Ngày kết thúc** | **06/01/2021** |

1. **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 hierachy Year-Quarter-Month-Date trong Cube
* Xác định và thiết kế các phân cấp chiều còn lại

Khai thác dữ liệu:

OLAP:

* Phân tích chuyến đi 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.

1. **KẾT QUẢ**
2. **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 (Cencus 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 chuyến 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

1. **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.

1. **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.

1. **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.

1. **Business Requirements**

|  |  |  |  |
| --- | --- | --- | --- |
| **Identifier** | **Name** | **Description** | **Priority** |
| 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. **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. **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. **Business and data transformations**
* *Nghiệp vụ chính:* Cách tính tổng tiền của xe taxi sau mỗi chuyến đi.
* *Data transformation:* Không có
  1. **Business metrics or Key Performance Indicators (KPIs)**
* *Total\_amount:* tiền tổng mỗi chuyến đi
* *Quy định tính tiền của mỗi chuyến đi:*

total\_amount = tolls\_amount + Extra + MTA\_tax

* 1. **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. **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

1. **Thống kê doanh thu năm theo quận**
   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.

* 1. **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\_lattude | Decimal | Vĩ độ điểm đón khách |
| Pickup\_longtitude | Decimal | Kinh độ điểm đón khách |
| Dropoff\_longtitude | 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 | Multipoligon | 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 |
| 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 |

* 1. **Business and data transformations**

*Nghiệp 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\_longtitude, Dropoff\_longtitude, 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.
  1. **Business metrics or Key Performance Indicators (KPIs)**

Total\_amount

* 1. **Business processes**

Không có.

* 1. **List business groups involved and describe type of involvement**

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

1. **Phân tích dữ liệu chuyến đi**
   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.
  1. **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\_longtitude | Decimal | Kinh độ điểm đón khách |
| Dropoff\_longtitude | 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 Census Block

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| the\_geom | Multipoligon | 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 |

* 1. **Business and data transformations**

*Nghiệp 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\_longtitude, Dropoff\_longtitude, 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.
  1. **Business metrics or Key Performance Indicators (KPIs)**

Không có.

* 1. **Business processes**

Không có.

* 1. **List business groups involved and describe type of involvement**

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

1. **Thống kê các loại dữ liệu theo tháng, quý, năm**
   1. **Requirement description**

Nghiệp 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.
  1. **Data sources**
* Dữ liệu NYC Yellow Taxi

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| 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\_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 Census Block

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| the\_geom | Multipoligon | 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 |

* 1. **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.
  1. **Business metrics or Key Performance Indicators (KPIs)**

Không có.

* 1. **Business processes**

Không có.

* 1. **List business groups involved and describe type of involvement**

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

1. **Source data description**
2. **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 |

1. **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$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại |
| 14 | tip\_amount | float | Tiền boa |
| 15 | tolls\_amount | float | Tiền đi qua trạm thu phí (phí cầu đường) |
| 16 | total\_amount | float | Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện) |
| 17 | imp\_surcharge | float | Phí phát sinh cải thiện chuyến đi. |
| 18 | extra | float | Phí phụ thu  Có 2 mức giá cho phí phụ thu:  0.5$ cho giờ cao điểm  1$ cho qua đêm |
| 19 | rate\_code | 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 |

1. **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$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại |
| 10 | tip\_amount | float | Tiền boa |
| 11 | tolls\_amount | float | Tiền đi qua trạm thu phí (phí cầu đường) |
| 12 | total\_amount | float | Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện) |
| 13 | improvement\_surcharge | float | Phí phát sinh cải thiện chuyến đi. |
| 14 | extra | float | Phí phụ thu  Có 2 mức giá cho phí phụ thu:  0.5$ 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 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 |
| 16 | PULocationID | int | ID định vị vị trí đón khách |
| 17 | DOLocationID | int | ID định vị vị trí trả khách |

1. **Dữ liệu của nguồn Census Block**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Column** | **Data Type** | **Description** |
| 1 | the\_geom | Multipoligon | 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 tượng trưng 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 |

1. **BI Dimensional Logical Model Design**
2. **Stage**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Column** | **Data Type** | **Description** |
| **YellowTaxi2014** | 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$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại |
| tip\_amount | float | Tiền boa |
| tolls\_amount | float | Tiền đi qua trạm thu phí (phí cầu đường) |
| total\_amount | float | Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện) |
| imp\_surcharge | float | Phí phát sinh cải thiện chuyến đi. |
| extra | float | Phí phụ thu  Có 2 mức giá cho phí phụ thu:  0.5$ cho giờ cao điểm  1$ cho qua đêm |
| rate\_code | 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 |
| **YellowTaxi2015** | 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 |
| **YellowTaxi2016** | 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 = 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$, nhưng có thể tự động thay đổi theo tỉ giá hiện tại |
| tip\_amount | float | Tiền boa |
| tolls\_amount | float | Tiền đi qua trạm thu phí (phí cầu đường) |
| total\_amount | float | Tổng chi phí hành khách phải trả (không bao gồm tiền boa và phí phát sinh cải thiện) |
| improvement\_surcharge | float | Phí phát sinh cải thiện chuyến đi. |
| extra | float | Phí phụ thu  Có 2 mức giá cho phí phụ thu:  0.5$ cho giờ cao điểm  1$ cho qua đêm |
| 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 |
| **CensusBlock** | 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 tượng 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 |

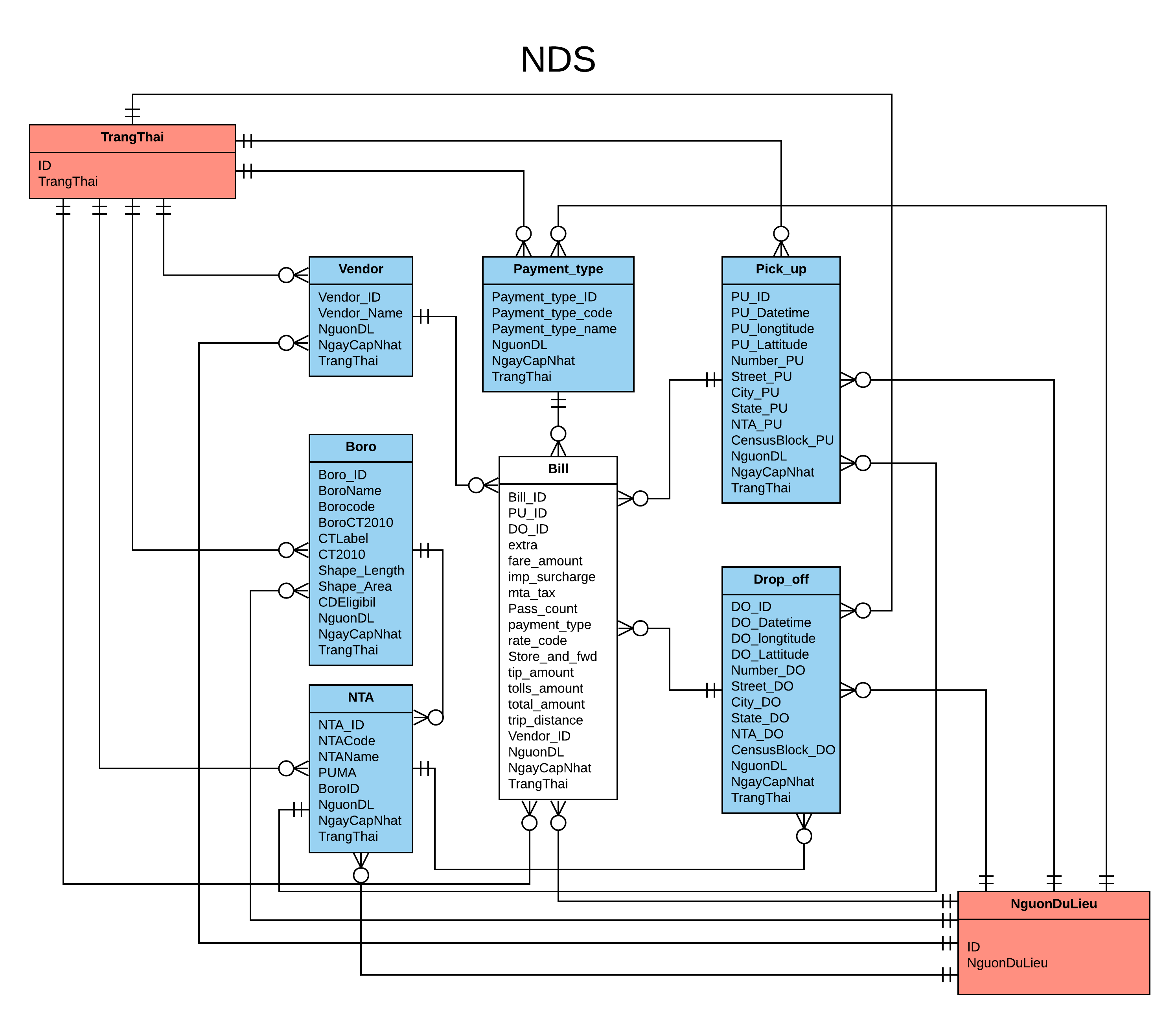
1. **NDS**
   1. **COLUMNS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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. |
| **NGUONDL** | 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. |
| **TRANGTHAI** | 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 Indentity(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. |
| **PAYMENT** | ID | PK | Int Indentity(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 = Chuyến 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. |
| **PICK\_UP** | PU\_ID | PK | Int Indentity(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\_lattitude |  | Varchar(100) | N |  | Vĩ độ điểm đón khách. |
| PU\_longtitude |  | 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. |
| TrangThai | FK | Int | N |  | Mã trạng thái. |
| **DROP\_OFF** | ID | PK | Int Indentity(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\_longtitude |  | 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 Indentity(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 Indentity(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 tượng trưng 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. |

* 1. **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 |

* 1. **DATA STORE LOGICAL MODEL**



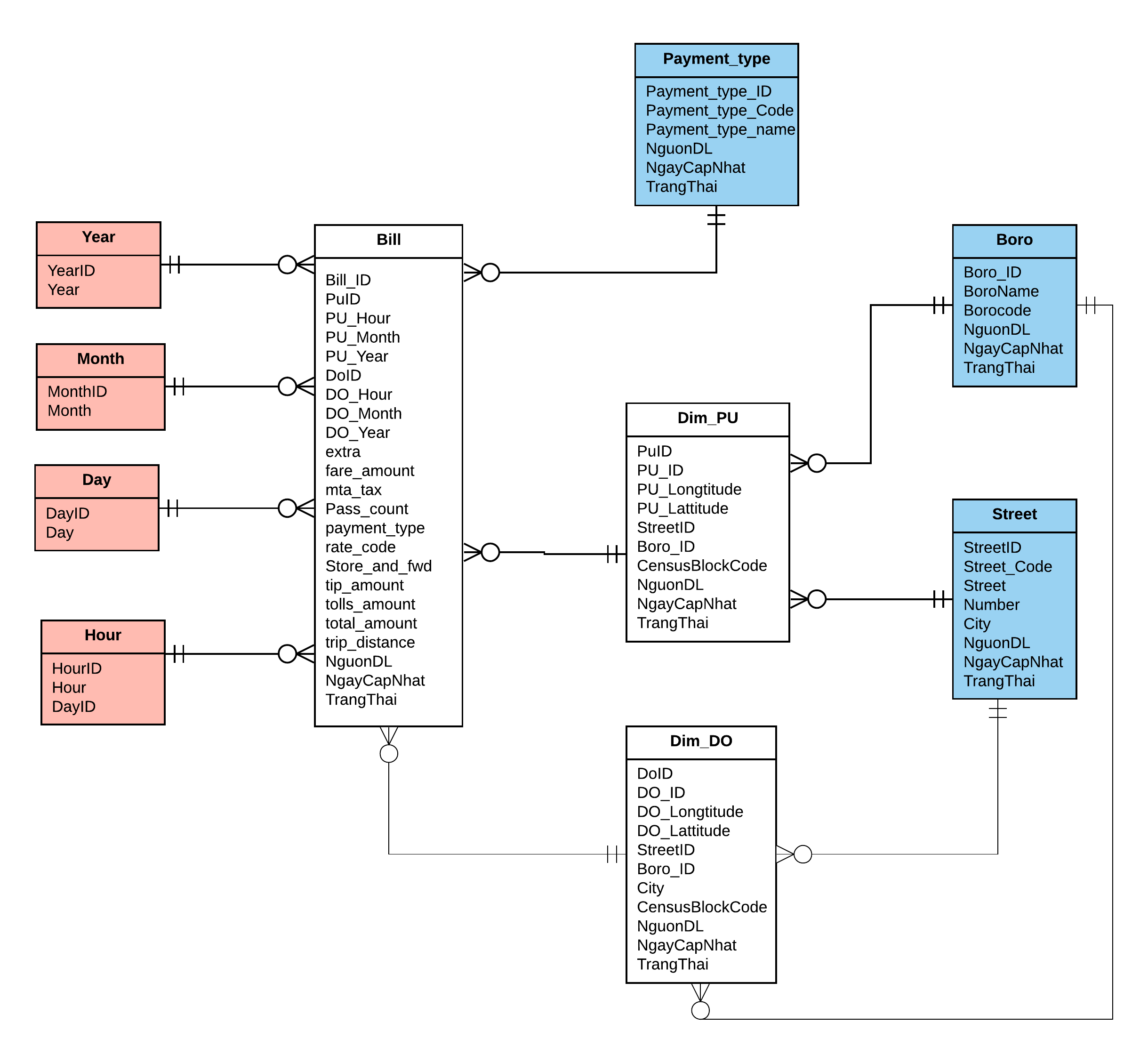
1. **DDS**
   1. **COLUMNS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table** | **Column** | **Key Type** | **Data Type** | **Is Nullable** | **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 boa |
| 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 boa) |
| NgayCapNhat |  | Datetime | N |  | Ngày cập nhật dữ liệu. |
| NguonDL |  | Int | N |  | Mã nguồn dữ liệu. |
| TrangThai |  | Int | N |  | Mã trạng thái. |
| **DIM\_COORDINATES** | Co\_ID | PK | Int | N |  | Mã địa chỉ. |
| Longtitude |  | Varchar(100) | N |  | Kinh độ địa chỉ. |
| Lattude |  | Varchar(100) | N |  | Vĩ độ địa chỉ. |
| StreetID | FK | Int | N |  | Mã đường. |
| BoroID | FK | Int | N |  | Mã quận. |
| CensusBlock |  | Varchar(100) | 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 = Chuyến 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. |
| **DIM\_BORO** | BoroID | PK | Int | N |  | Mã quận. |
| BoroCode |  | Varchar(100) | N | Từ 1 đến 5 | Mã định vị 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 |
| 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\_HOUR** | HourID | PK | Int | N |  | Mã giờ |
| HHour |  | Int | N |  | Giờ. |
| **DIM\_DAY** | DayID | PK | Int | N |  | Mã ngày |
| DDay |  | Int | N |  | Ngày |
| **DIM\_MONTH** | MonthID | PK | Int | N |  | Mã tháng |
| Monthh |  | Int | N |  | Tháng |
| **DIM\_YEAR** | YearID | PK | Int | N |  | Mã năm |
| Yearr |  | Int | N |  | Năm |

* 1. **CONSTRAINT**

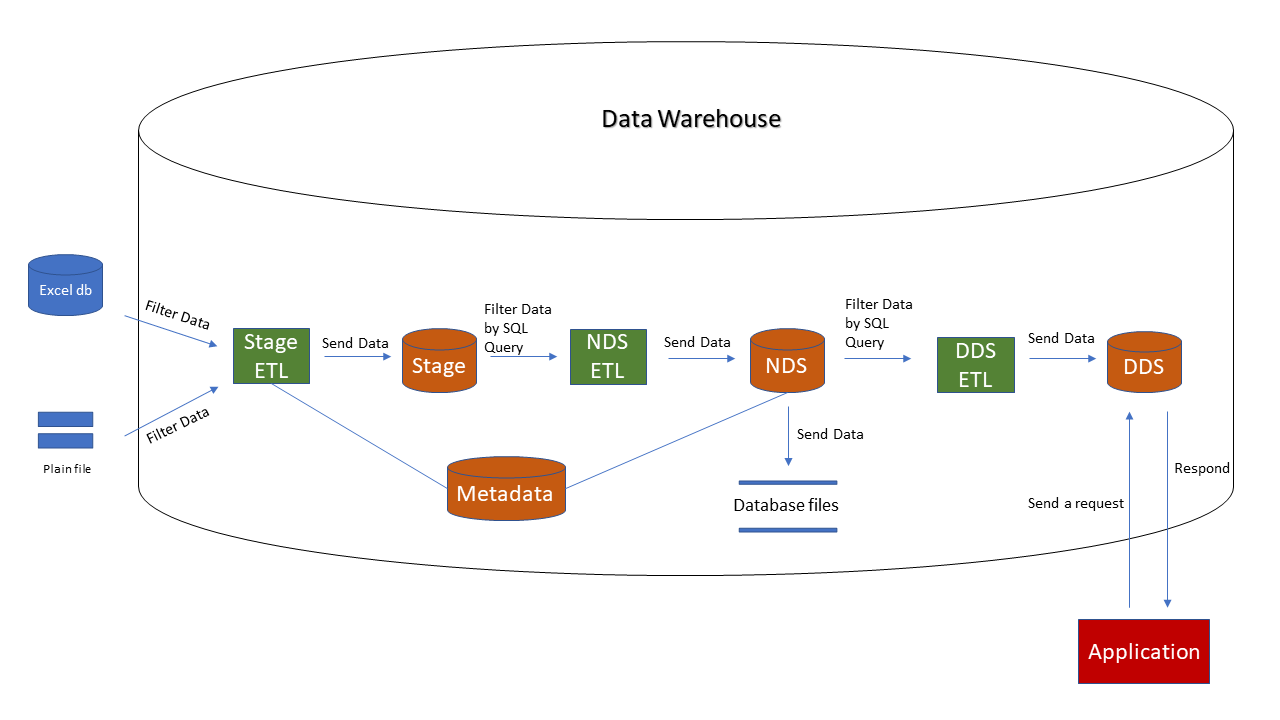
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Constraint Name** | **Constraint Type** | **Table** | **Column** | **Description** |
| PK\_BILL | Primary key | BILL | Bill\_ID | Khoá chính của bảng BILL |
| FK\_BILL\_PAYMENT | Foreign key | BILL | Payment\_type | Khoá ngoại tham chiếu từ BILL đến DIM\_PAYMENT |
| FK\_BILL\_PUCOORDINATES | 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\_PUMONTH | 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\_DOCOORDINATES | 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\_DOMONTH | 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\_PAYMENT | 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 |

* 1. **DATASTORE LOGICAL MODEL**



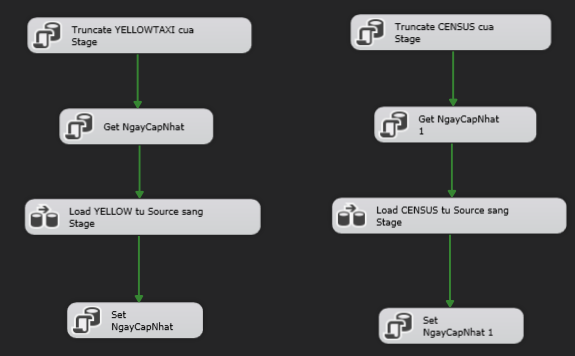
1. **Thiết kế Data Flow**

* **Data flow architecture theo NDS và DDS**

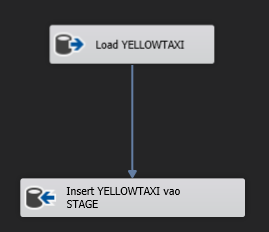
****

1. **ETL**
2. **Source to Stage**

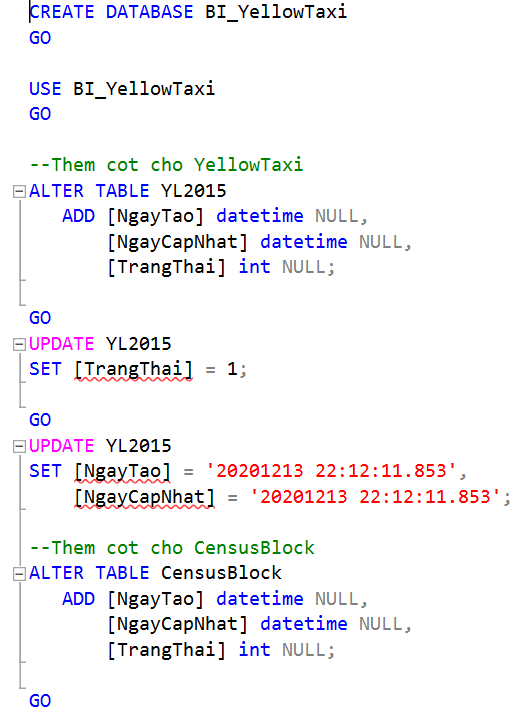
Ta thiết kế dòng Data flow như hình dưới đây để nạp các dữ liệu từ nguồn Yellow Taxi và Census Block vào Stage:



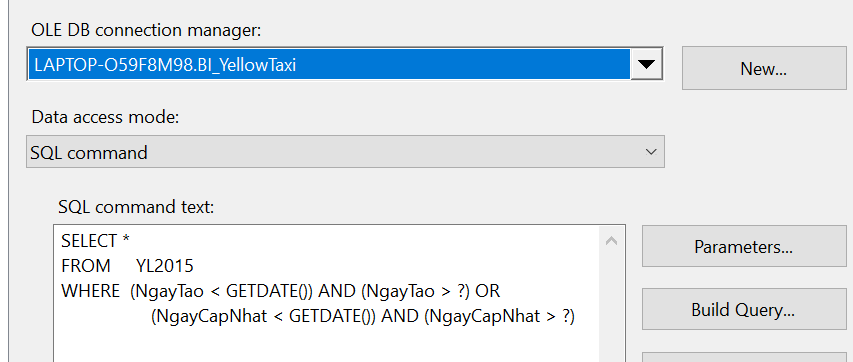
Ở phần *Load YELLOW tu Source sang Stage* ta có một luồng như hình dưới:



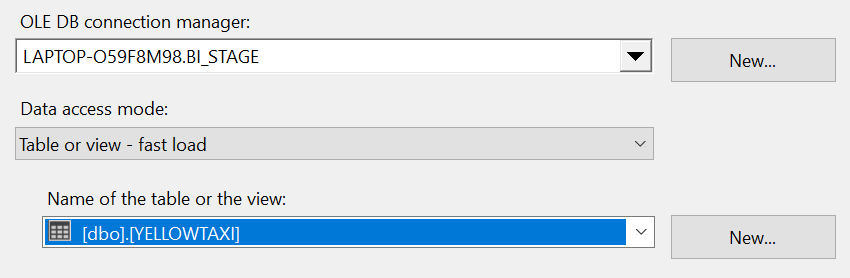
Ta tạo database BI\_YellowTaxi và thực hiện thêm các cột NgayTao, NgayCapNhat:



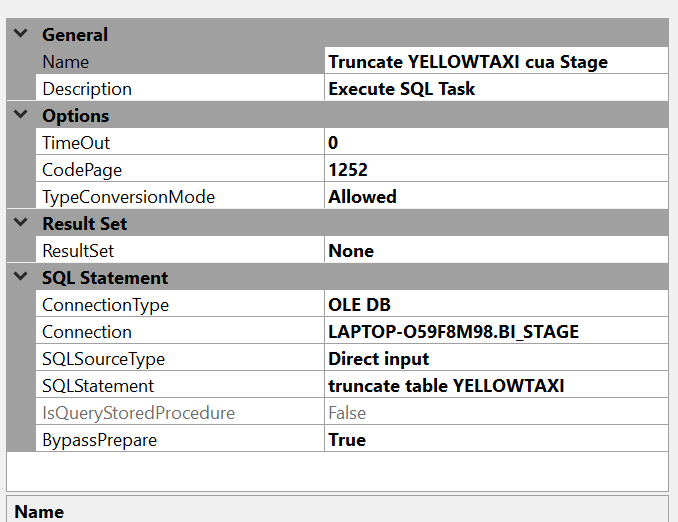
Sau đó ta thực hiện kết nối với database và nhập câu lệnh SQL như dưới đây:



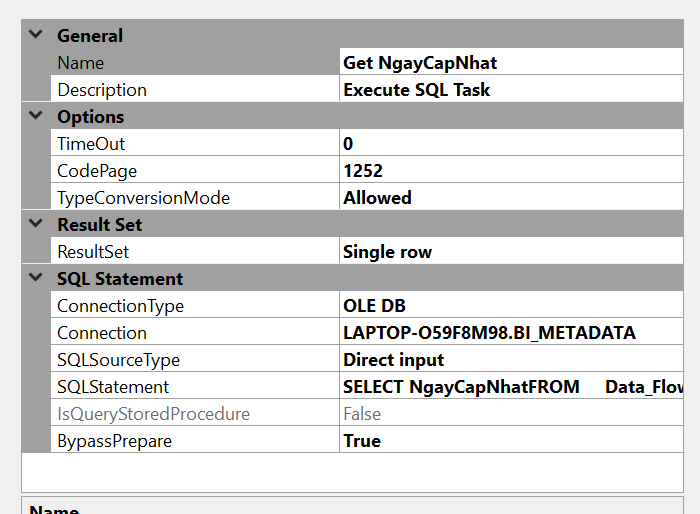
Và ta thực hiện nạp dữ liệu vào bảng YELLOWTAXI của Stage



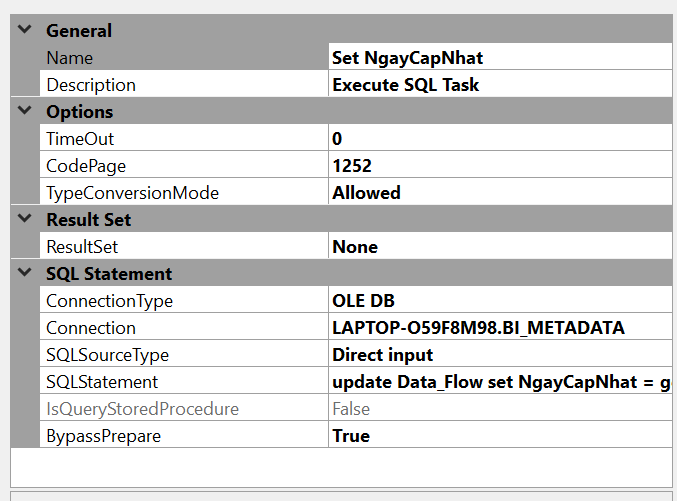
Ta tạo một Truncate và thao tác như hình dưới:



Ta tạo GET NgayCapNhat và thao tác như hình dưới:



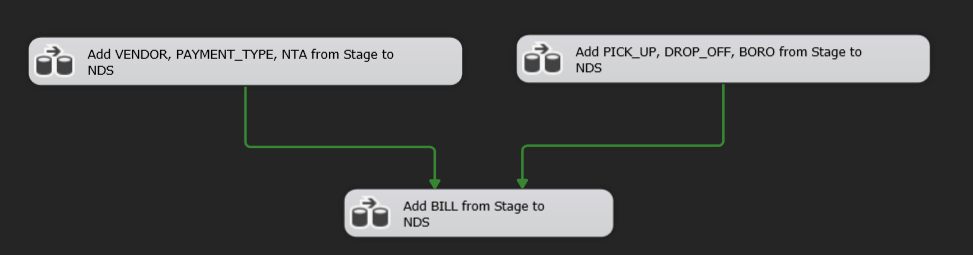
Ta tạo SET NgayCapNhat và thao tác như hình dưới:



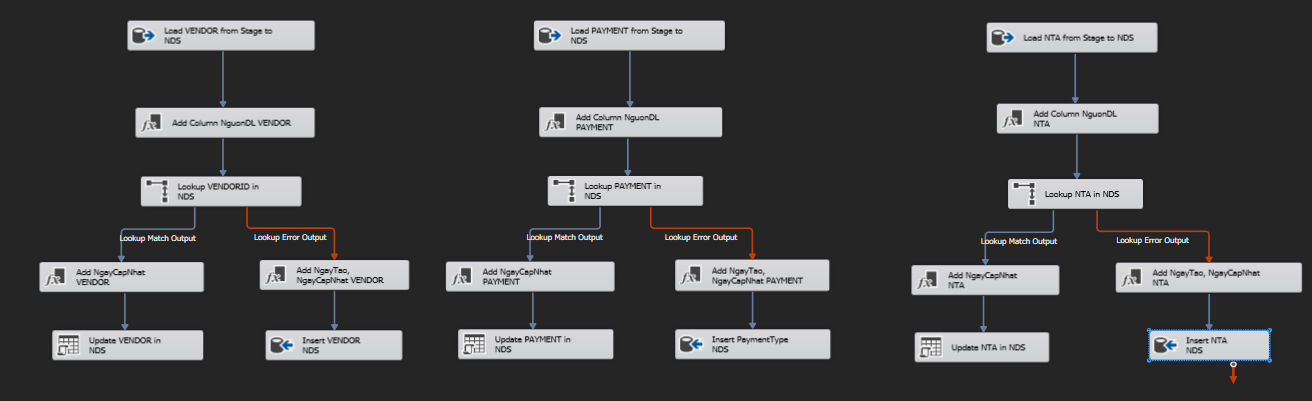
Ta làm tương tự trên với nạp dữ liệu từ nguồn Census Block và Stage

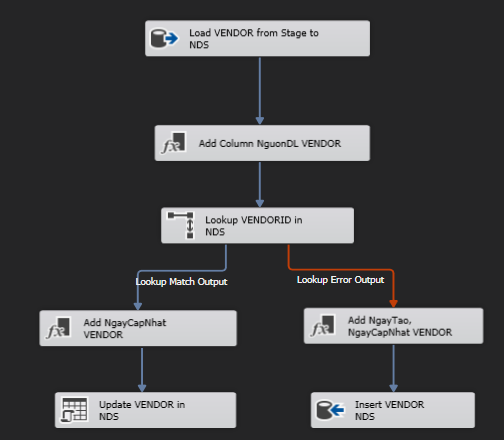
1. **Stage to NDS**

Ta tạo một Data Flow như hình dưới đây:

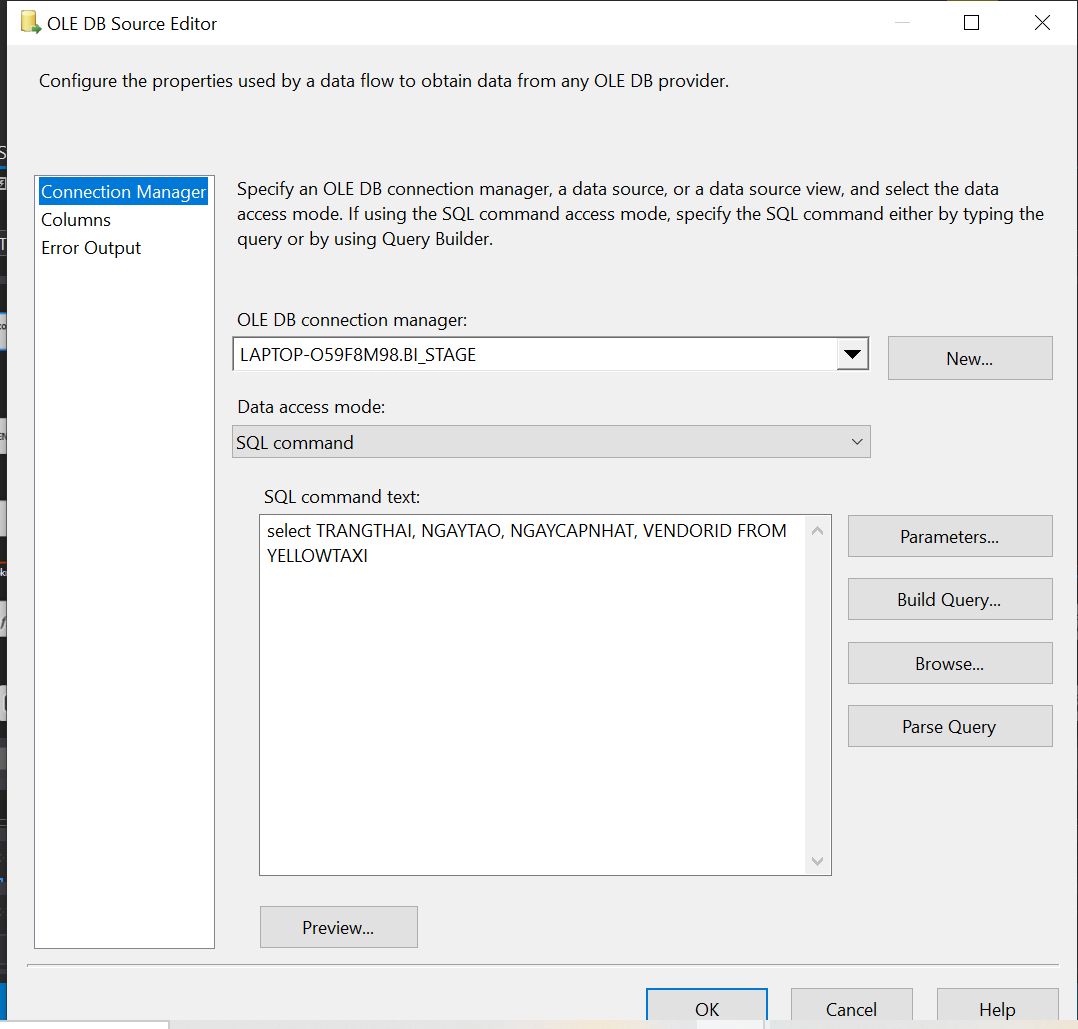


Ở phần *Add VENNDOR, PAYMENT, NTA from Stage to NDS* ta tạo các luồng như sau:

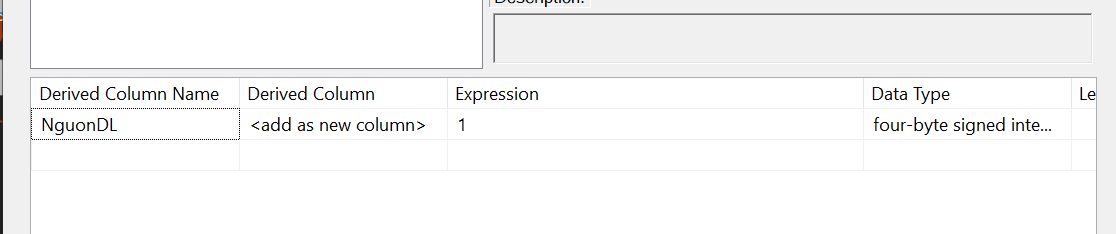




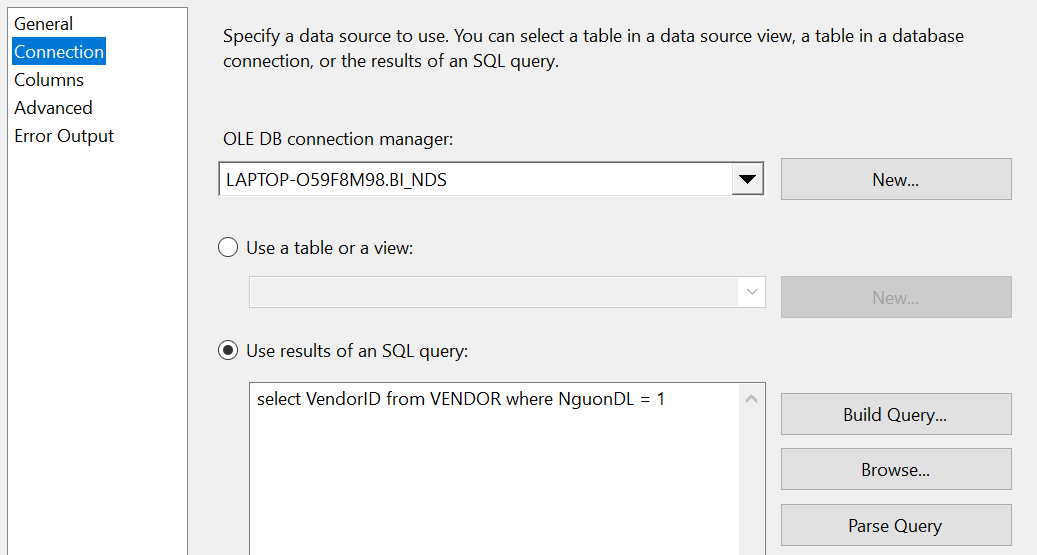
Đầu tiên, ta tải dữ liệu của bảng Vendor từ Stage:



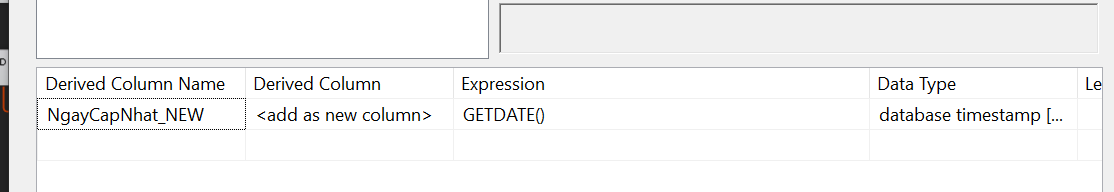
Thêm cột NguonDL cho bảng Vendor:



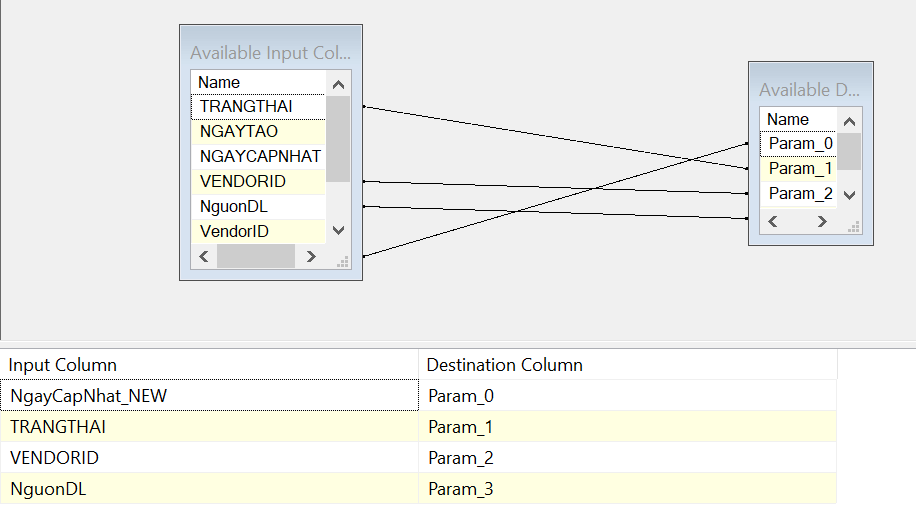
Tiếp theo, ta lookup xem VendorID đã tồn tại trong NDS chưa



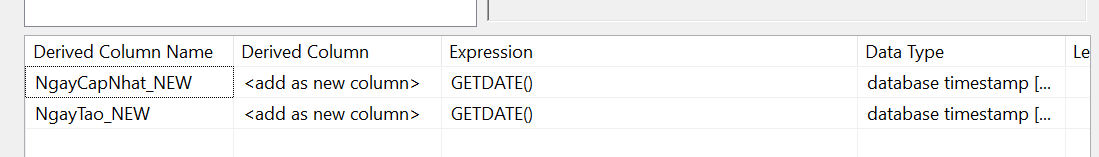
Nếu đã tồn tại, ta thêm cột NgayCapNhat:



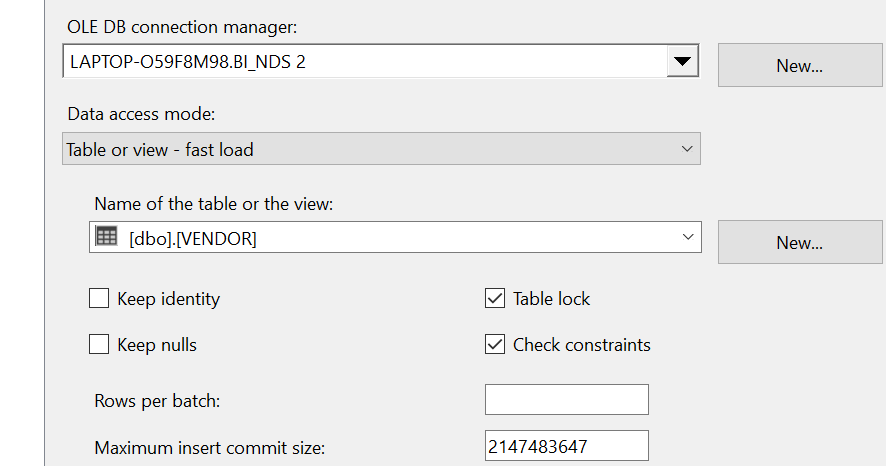
Và update lại dữ liệu bảng Vendor với VendorID tương ứng:



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

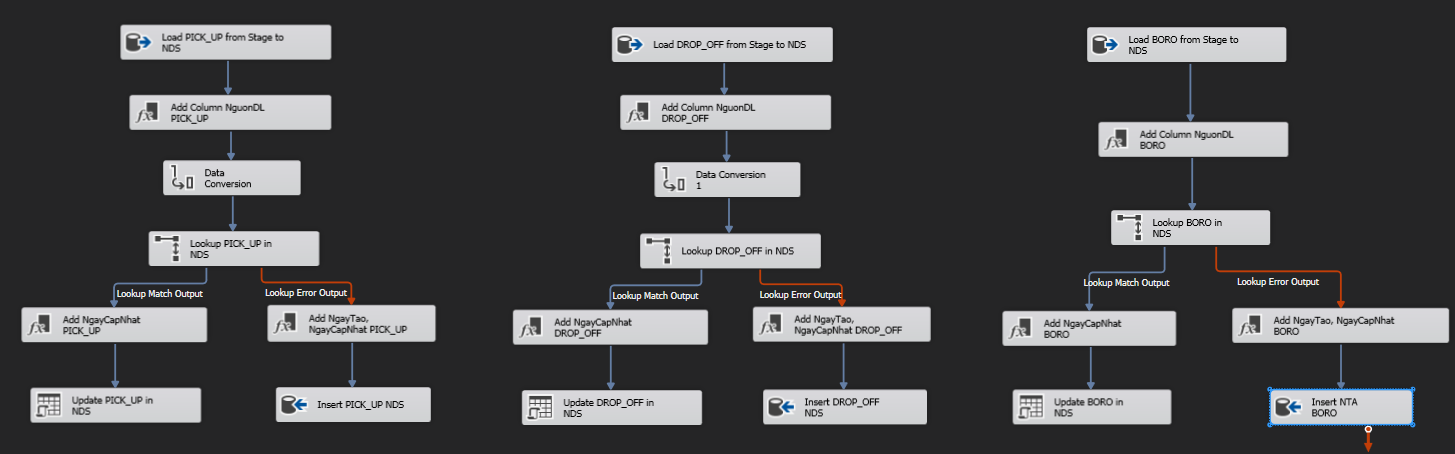


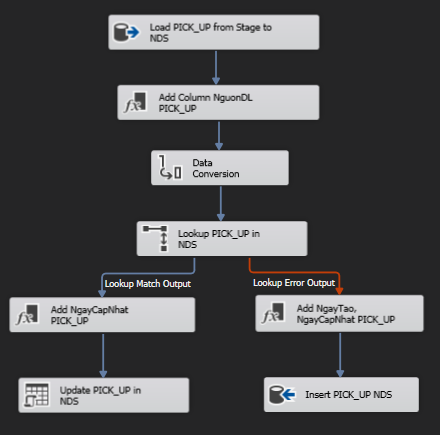
Sau đó, ta nạp dữ liệu mới vào bảng Vendor của NDS:



Ta thực hiện tương tự ở các luồng khác.

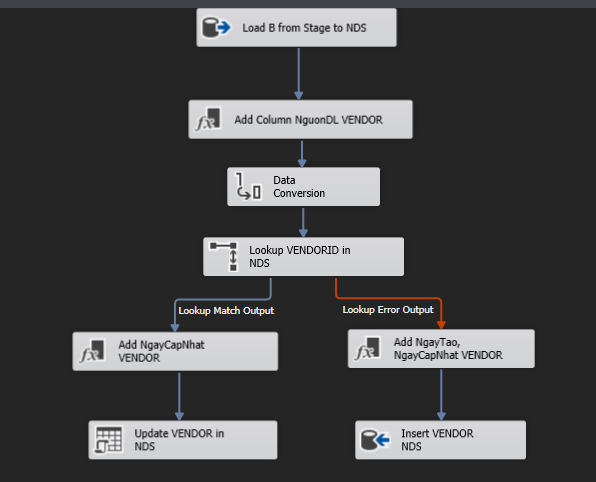
Ở phần *Add PICK\_UP, DROP\_OFF, BORO from Stage to NDS* ta tạo các luồng như sau:



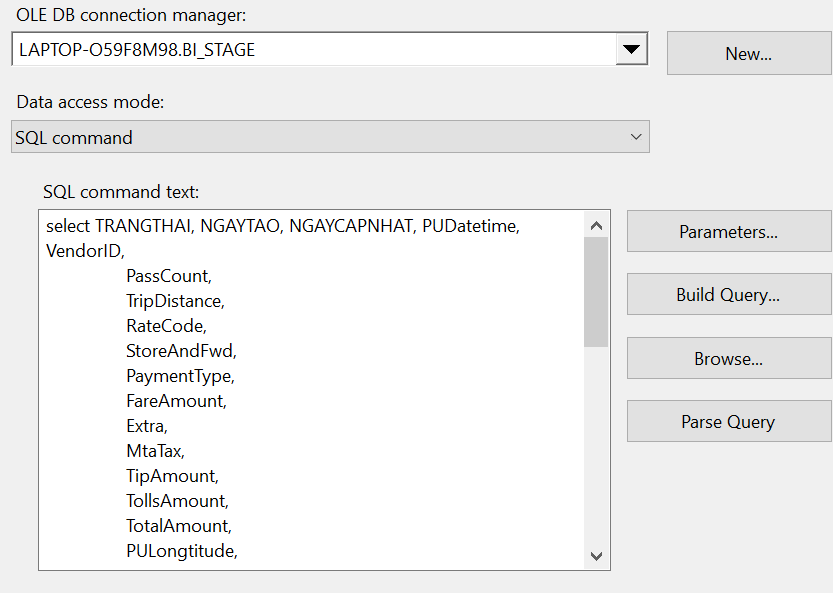


Ở đây, ta cũng thực hiện tương tự*Add VENDOR, PAYMENT, NTA from Stage to NDS*

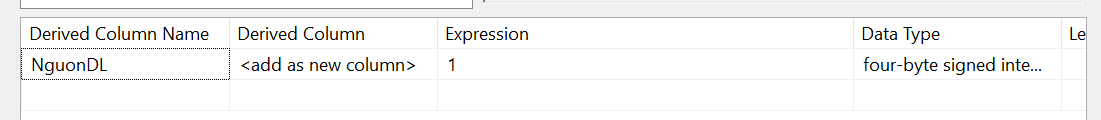
Ở phần *Add BILL from Stage to NDS* ta tạo luồng như sau:



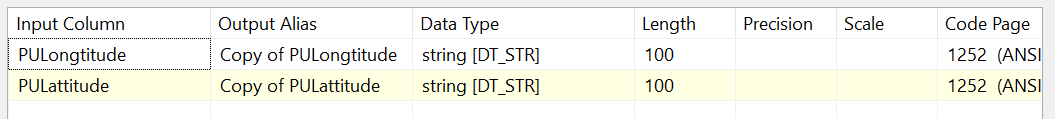
Trước tiên, ta tải dữ liệu bảng Bill từ Stage:



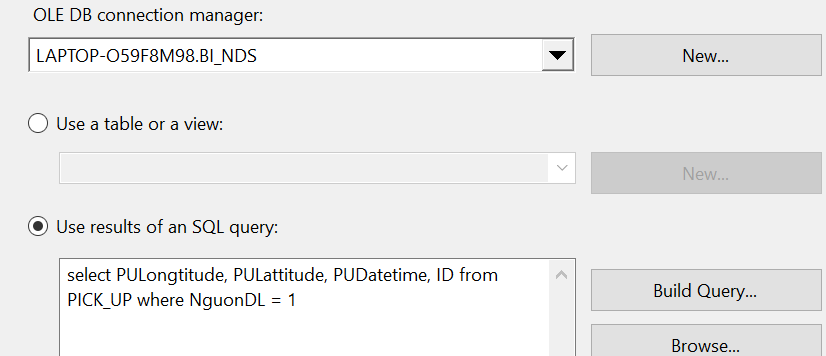
Tiếp theo, ta thêm cột NguonDL cho bảng Bill:



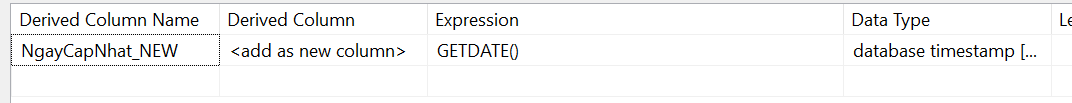
Sau đó, ta convert kiểu dữ liệu của tọa độ sang varchar(100):



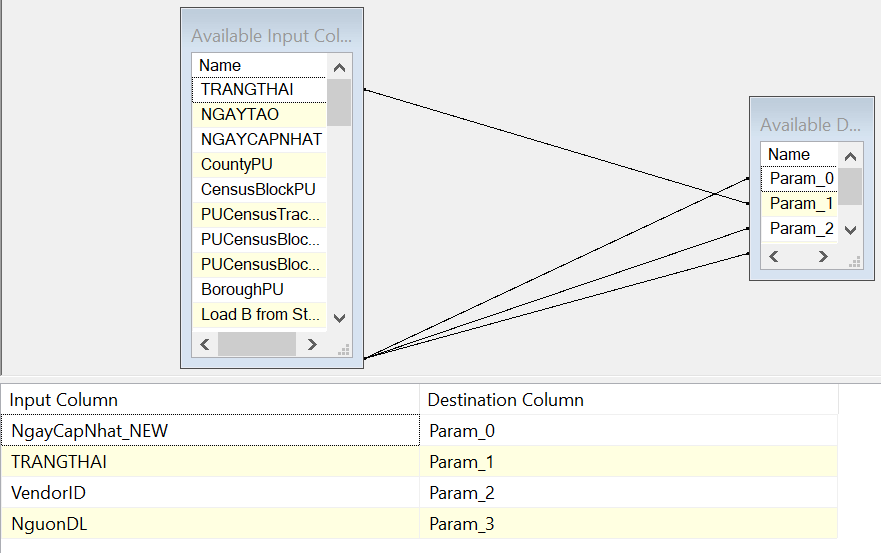
Tiếp theo, ta lookup các tọa độ và thời gian đón xe:



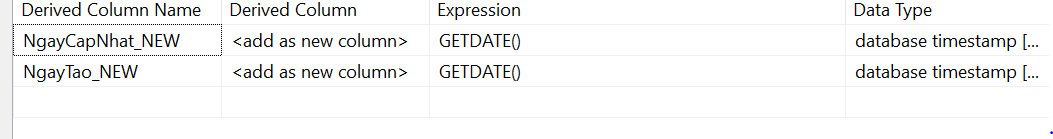
Nếu đã tồn tại, ta thêm cột NgayCapNhat:



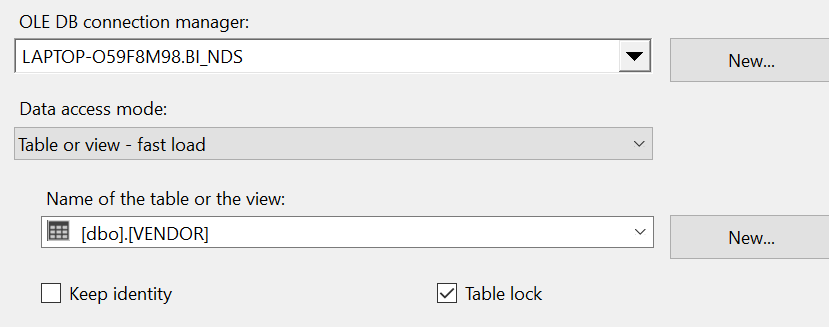
Và update lại dữ liệu tương ứng:



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

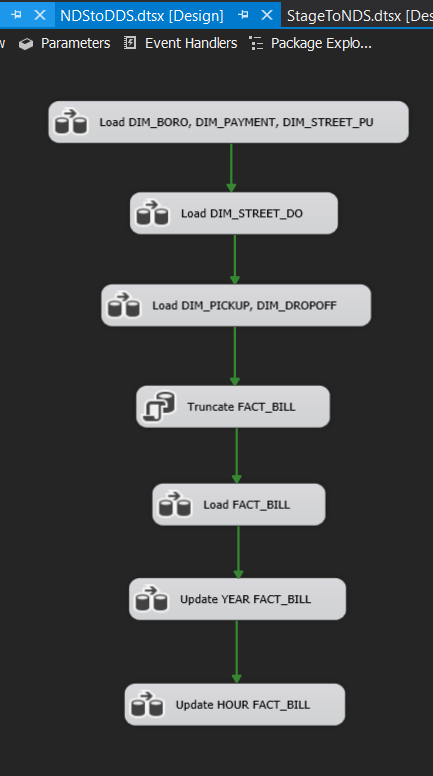


Và thực hiện nạp dữ liệu mới từ Stage vào bảng Bill của NDS

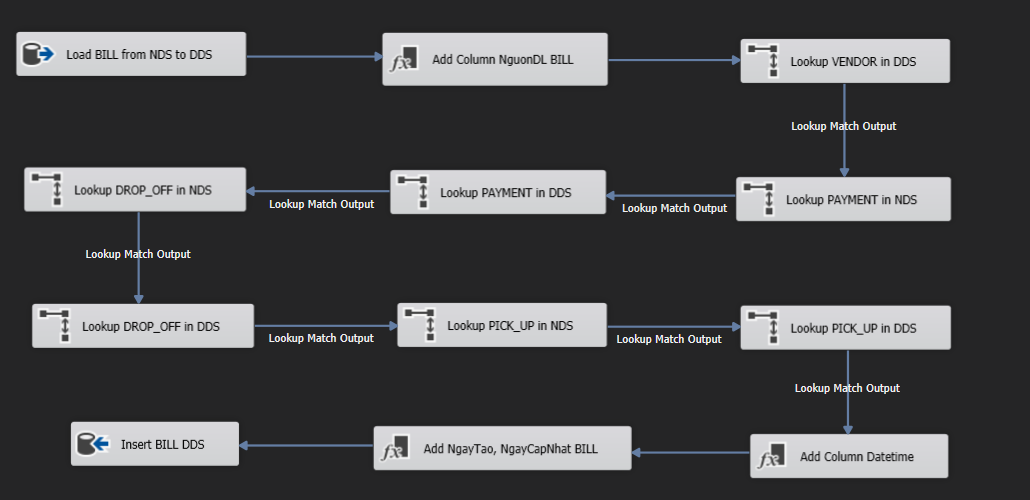


1. **NDS to DDS**

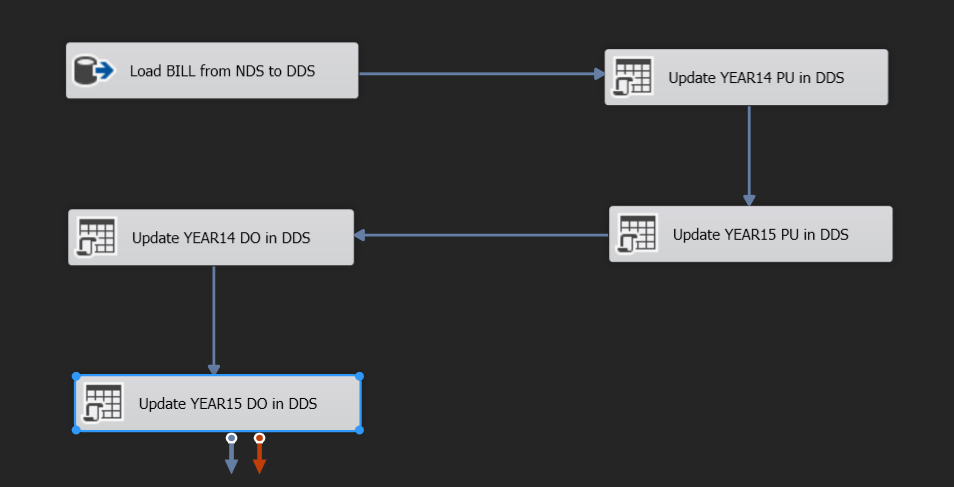
Ta tạo database BI\_DDS và lần lượt nạp dữ liệu vào các bảng Dim.



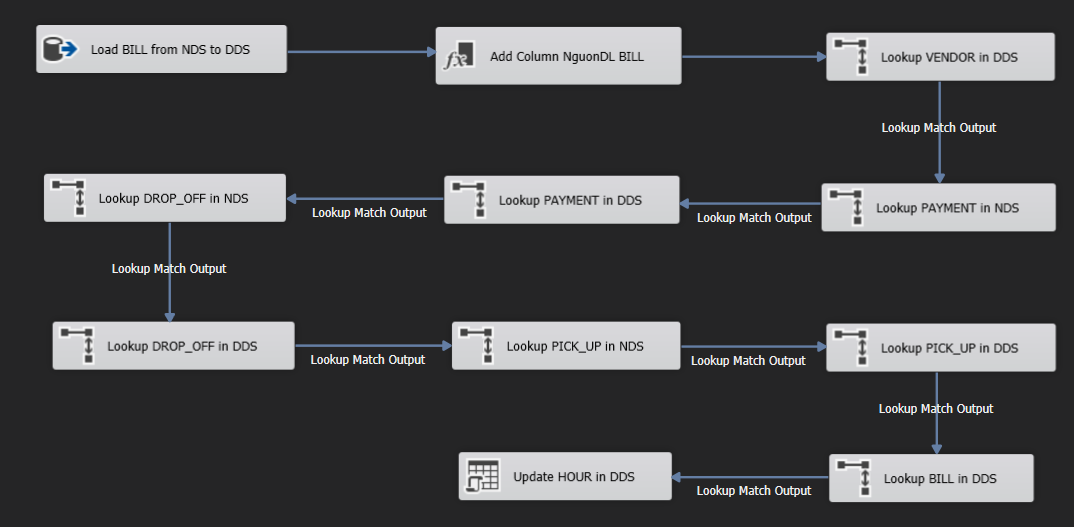
Sau khi đã hoàn tất nạp vào các bảng Dim, ta tiếp tục tạo data flow như hình dưới để nạp dữ liệu vào bảng Fact là bảng Bill.



Update năm:



Update giờ:

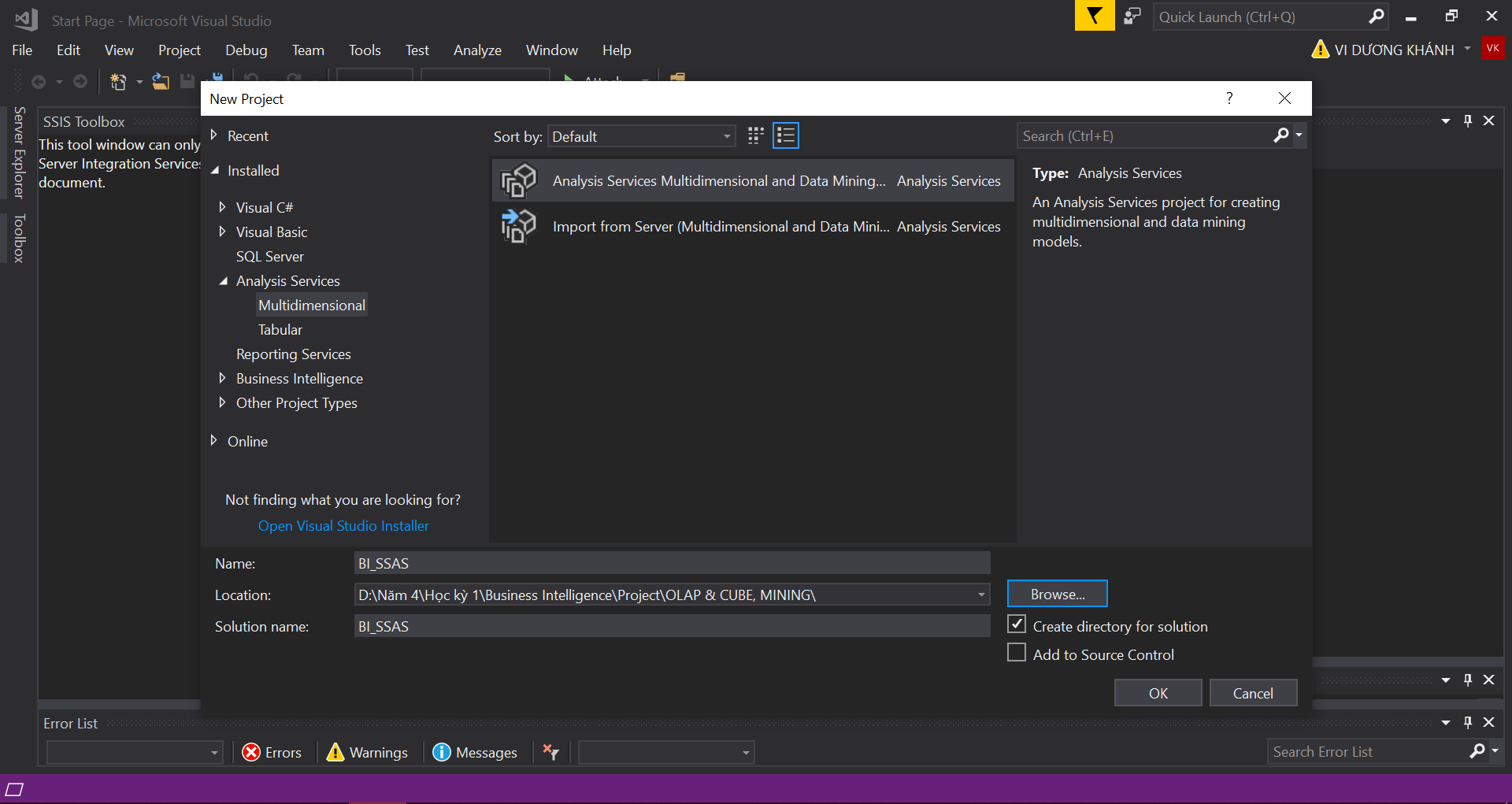


Ta lấy dữ liệu từ bảng Bill ở NDS. Sau đó, lần lượt lookup các thuộc tính khóa ngoại của bảng, nếu tất cả đều đã tồn tại trong các bảng Dim, ta thực hiện insert dữ liệu vào bảng Fact\_Bill. Sau khi chạy, ta kiểm tra lại ở SQL Server thì thấy đã có dữ liệu.

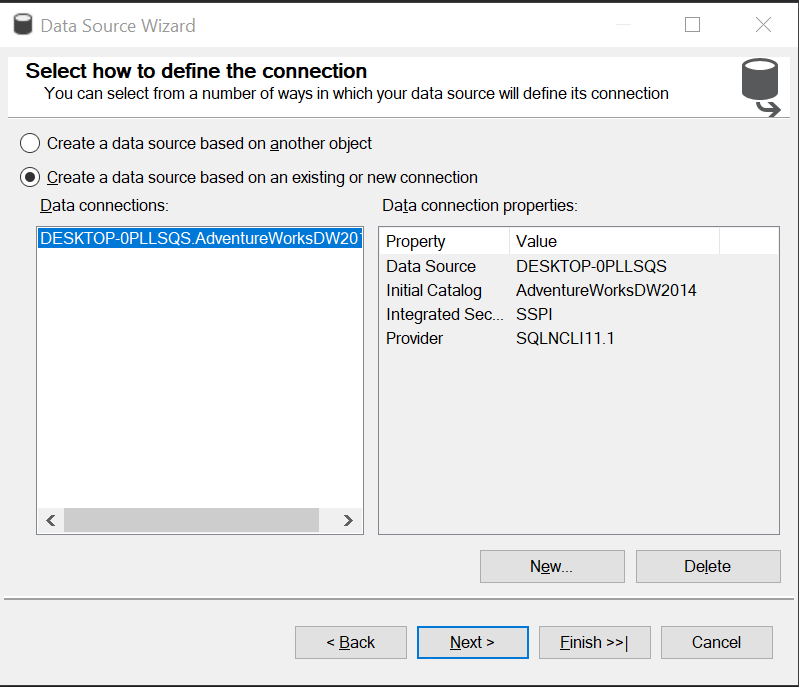


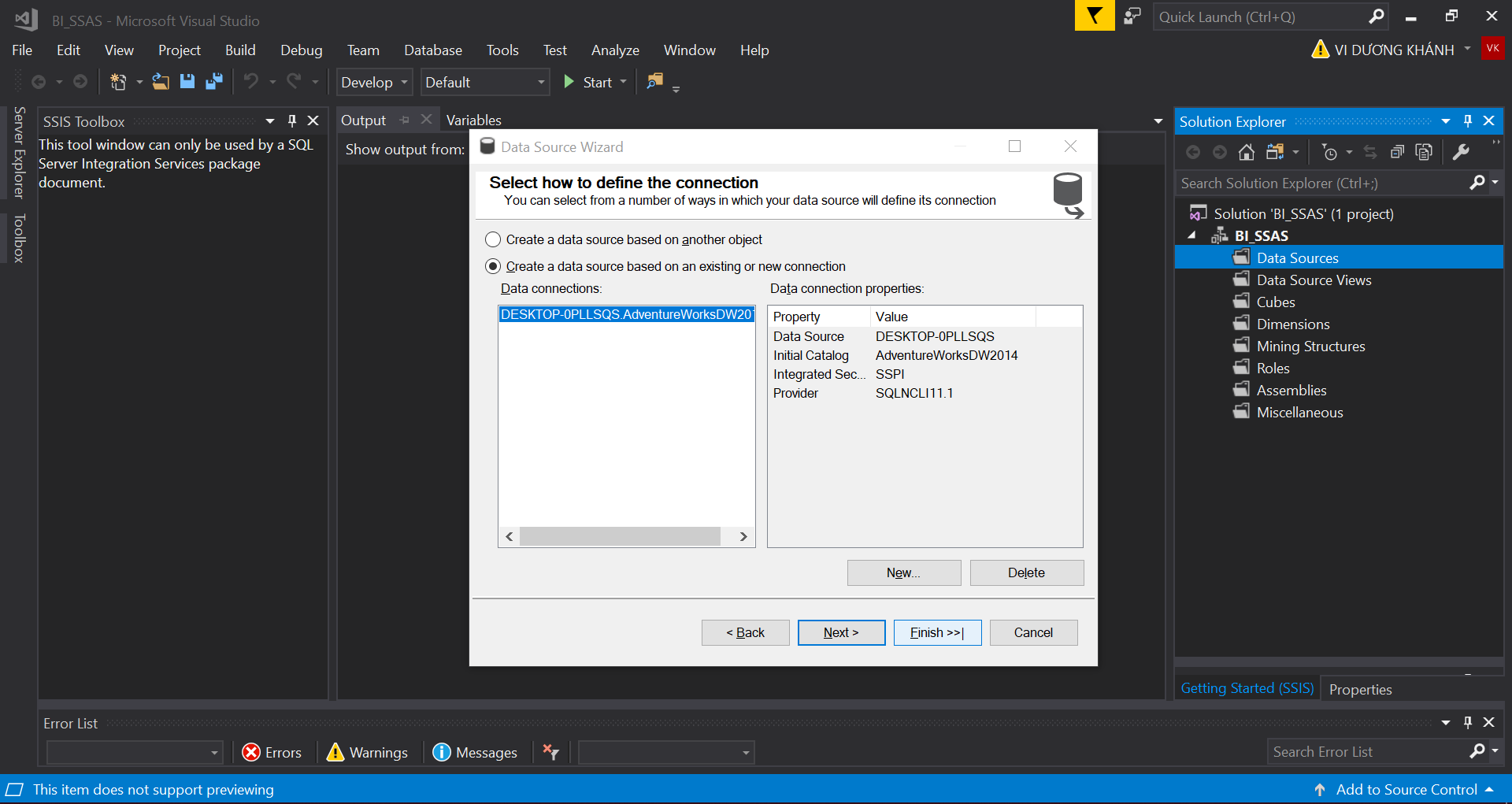
1. **Khai thác dữ liệu**
2. **OLAP**

Bước 1: Tạo project SSAS

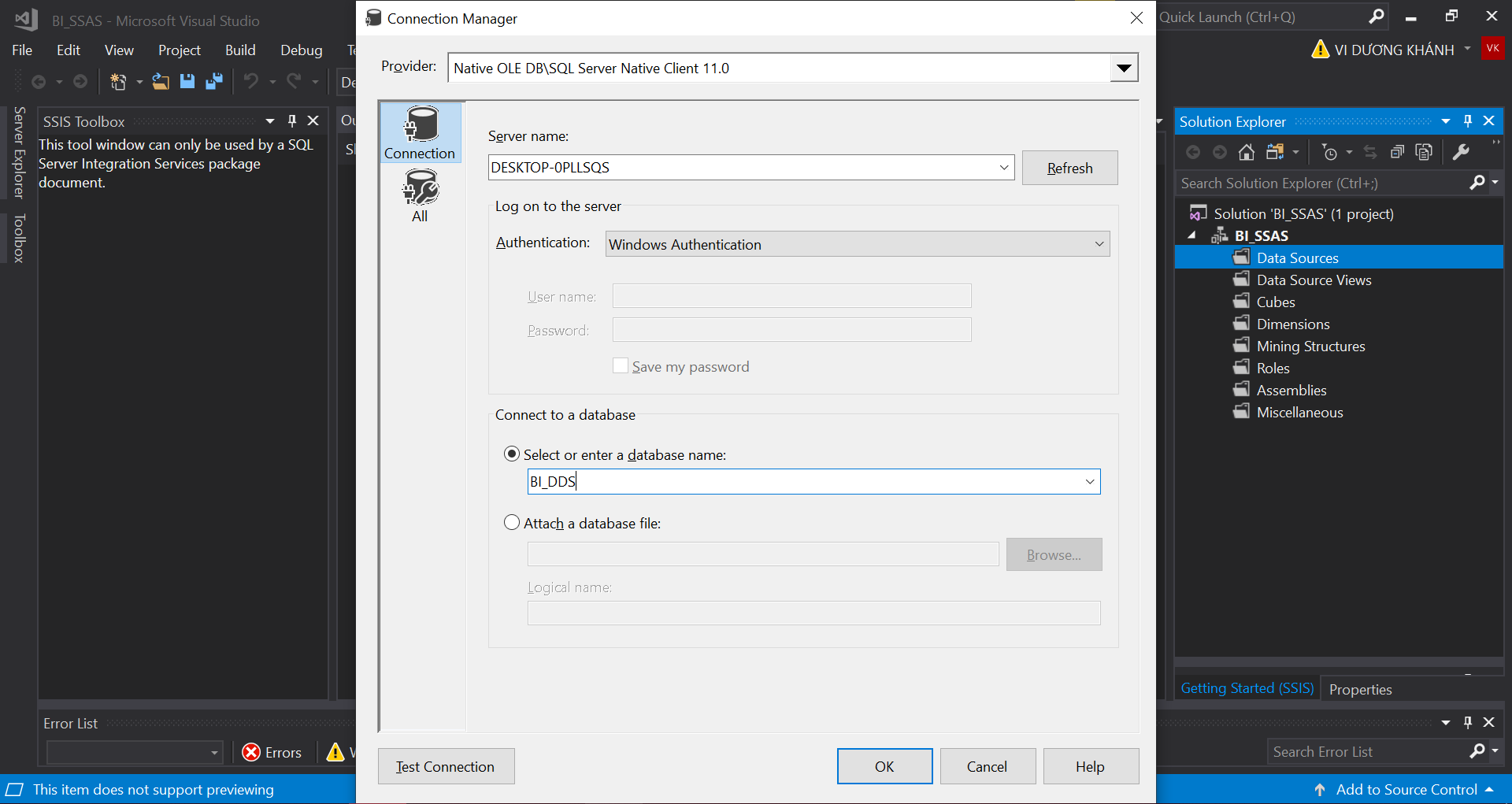


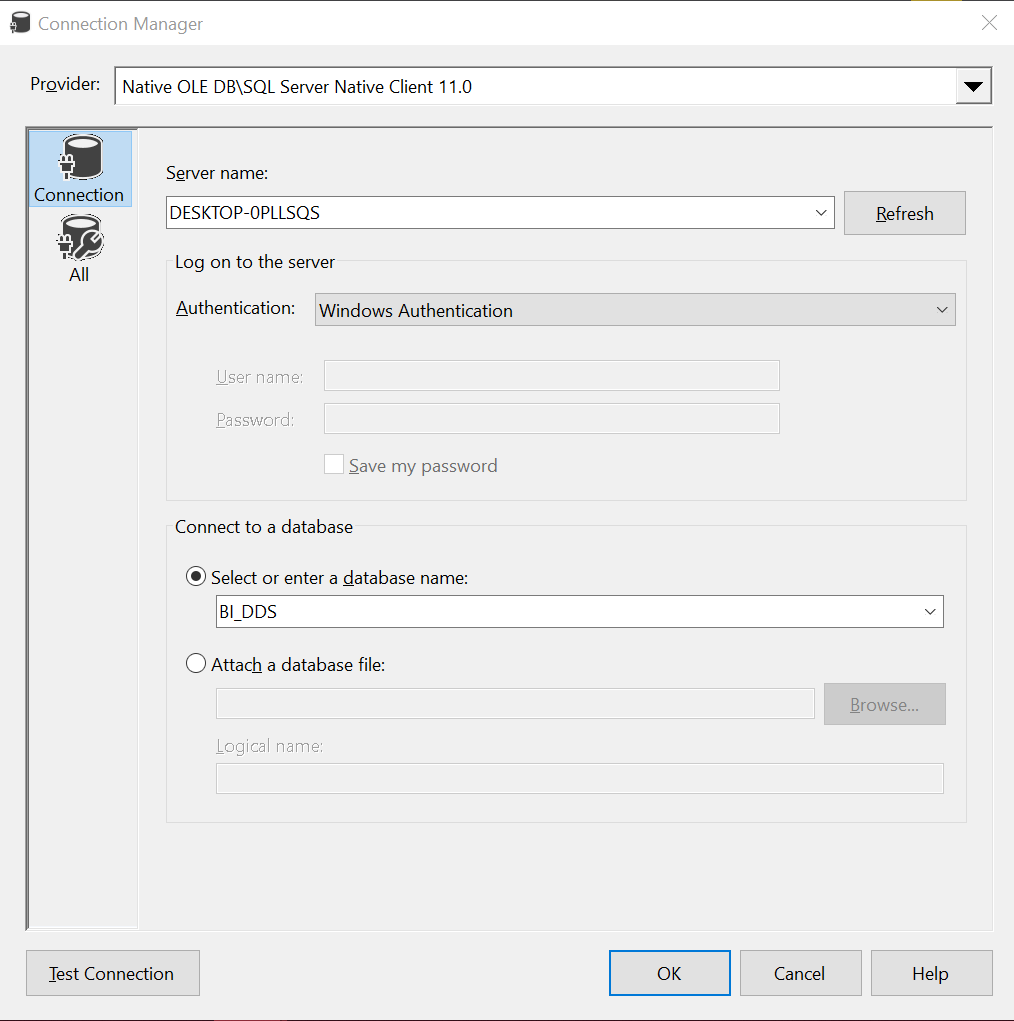
Bước 2: Tạo Data Source



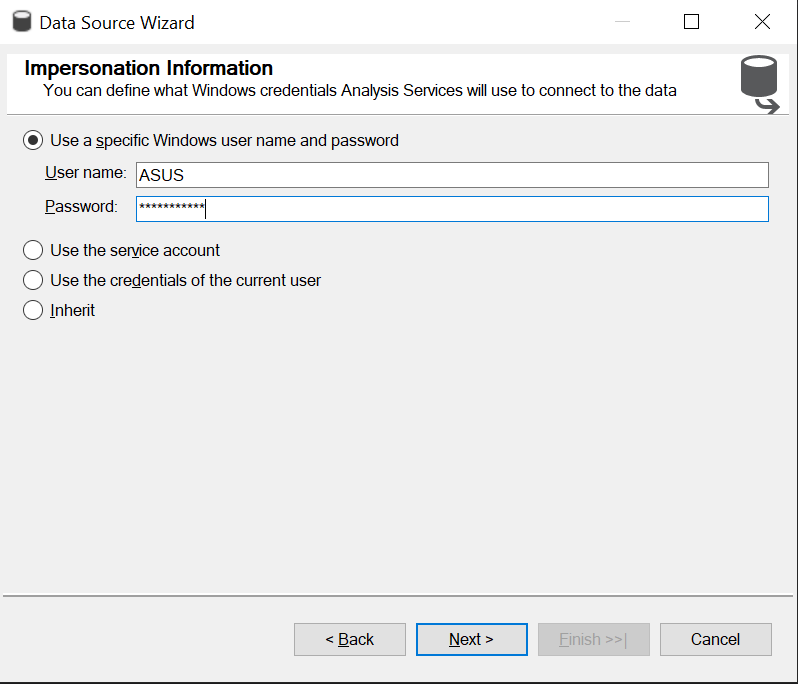


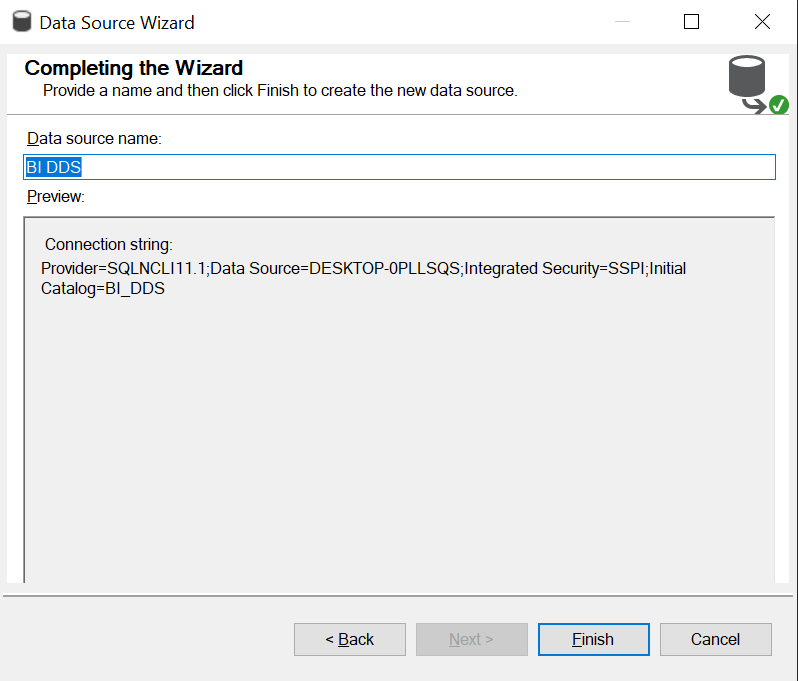
Tạo kết nối Database:





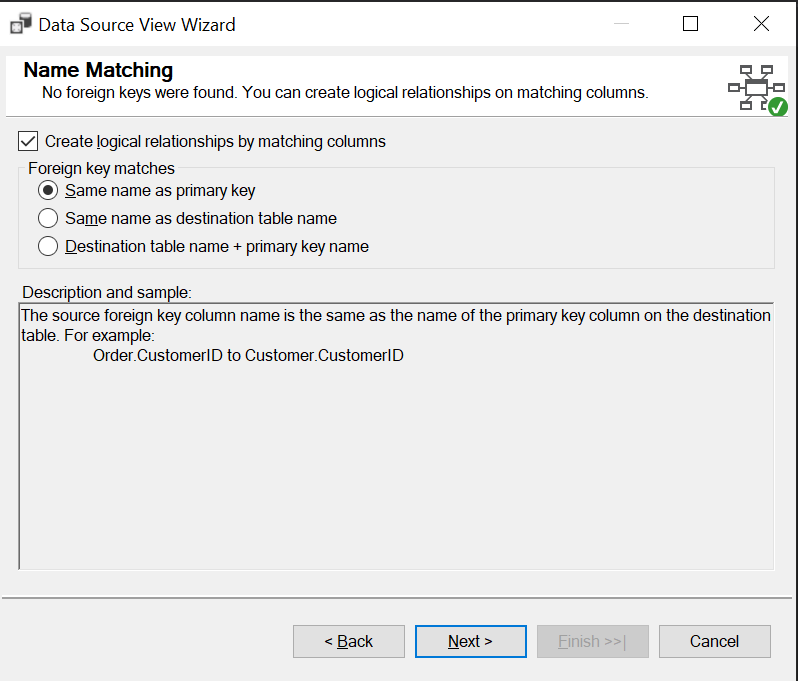
Sau khi tạo kết nối Database, ta sẽ chọn sử dụng các loại tài khoản để thực hiện process. Ở đây em chọn tài khoản của Window trên máy tính.



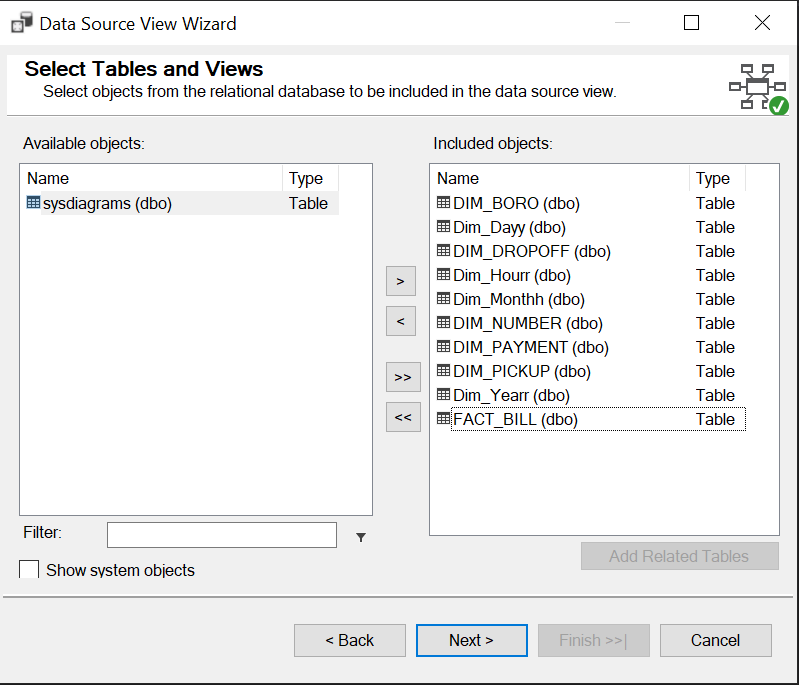


Bước 3: Tạo Data Source View

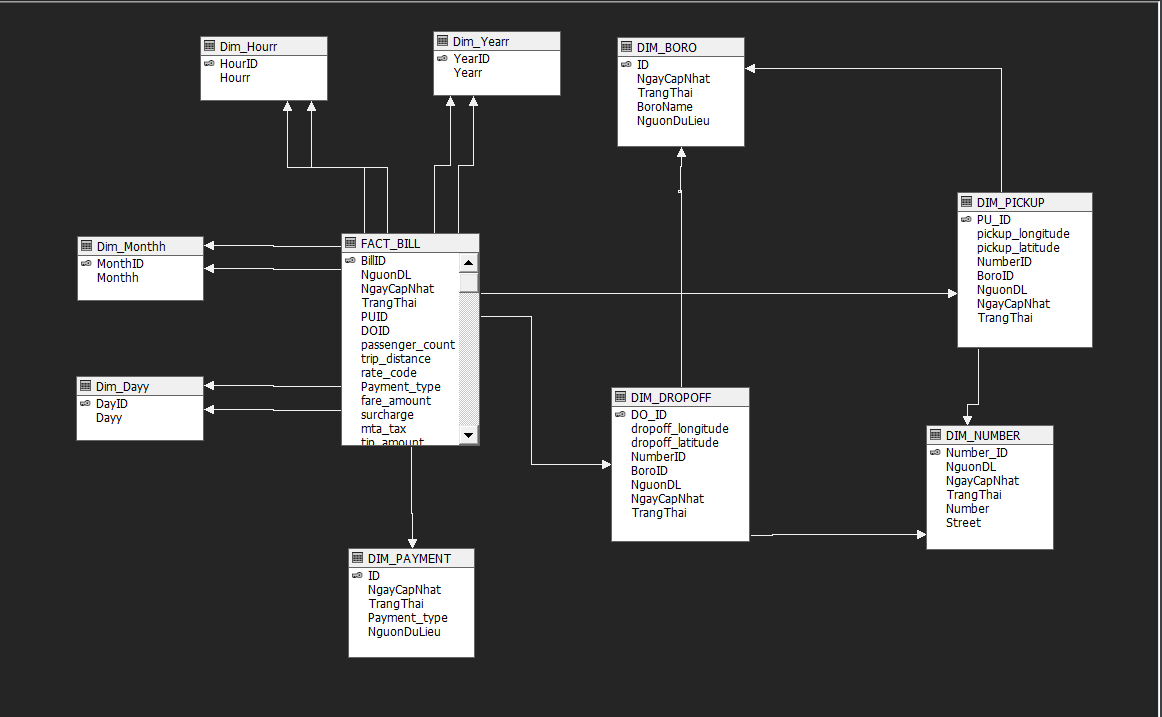




Chọn các bảng cần có trong Data Source Views.



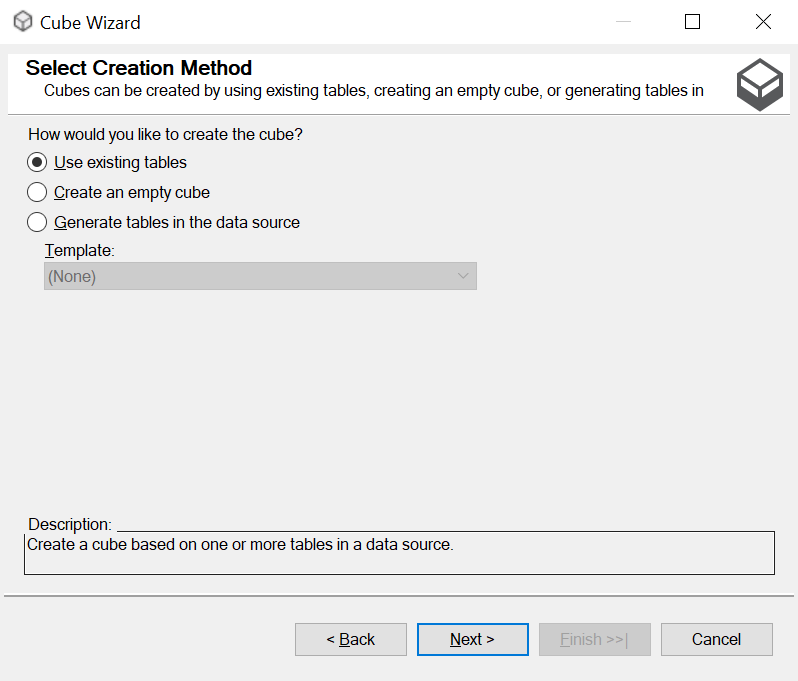
Quá trình sau khi hoàn thành:



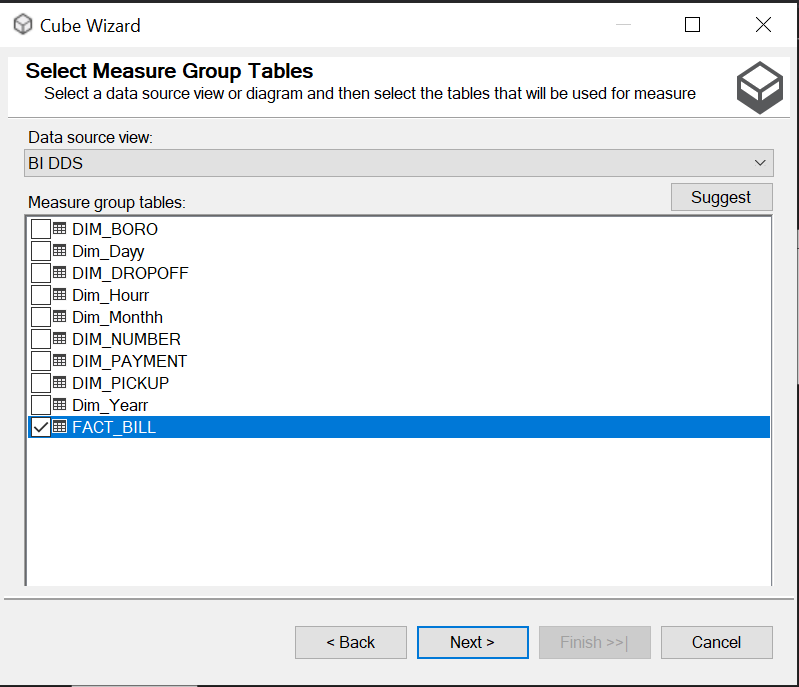
1. **Cube**

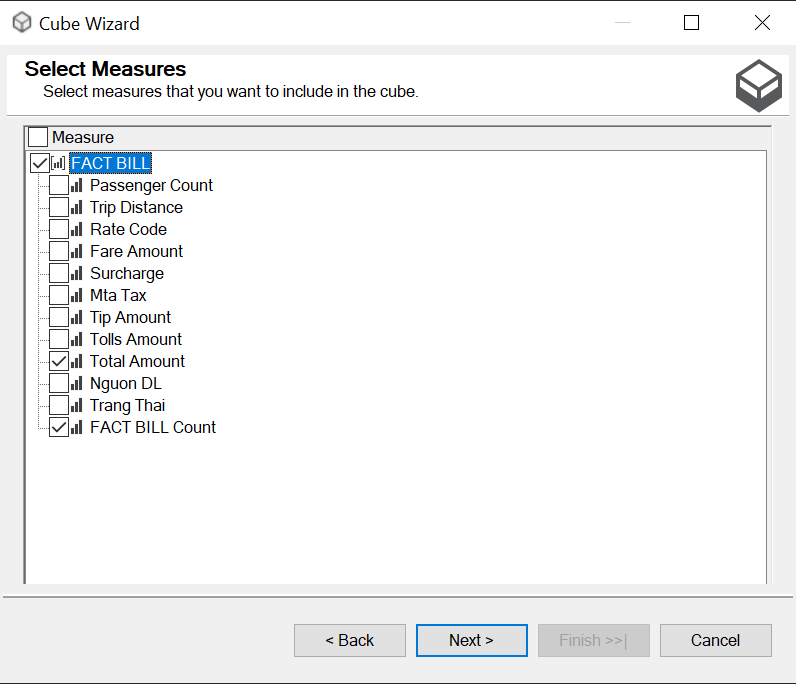
Tiếp nối phần OLAP ở trên, ta sẽ tiến hành tạo Cube:

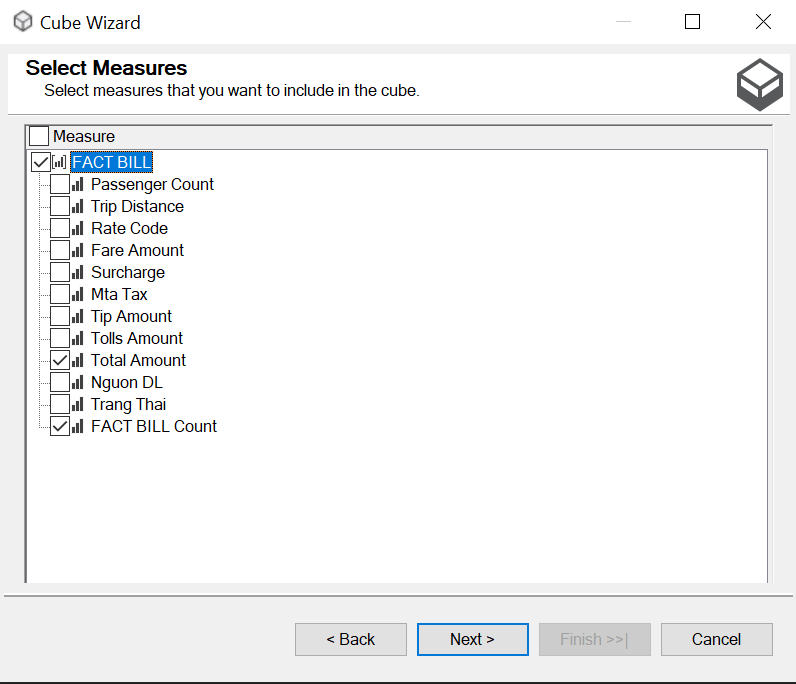


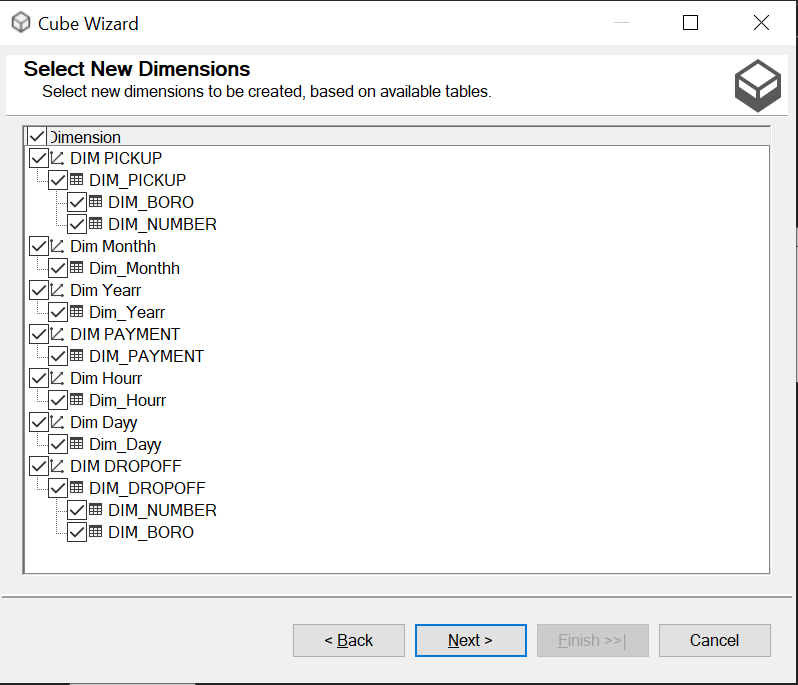


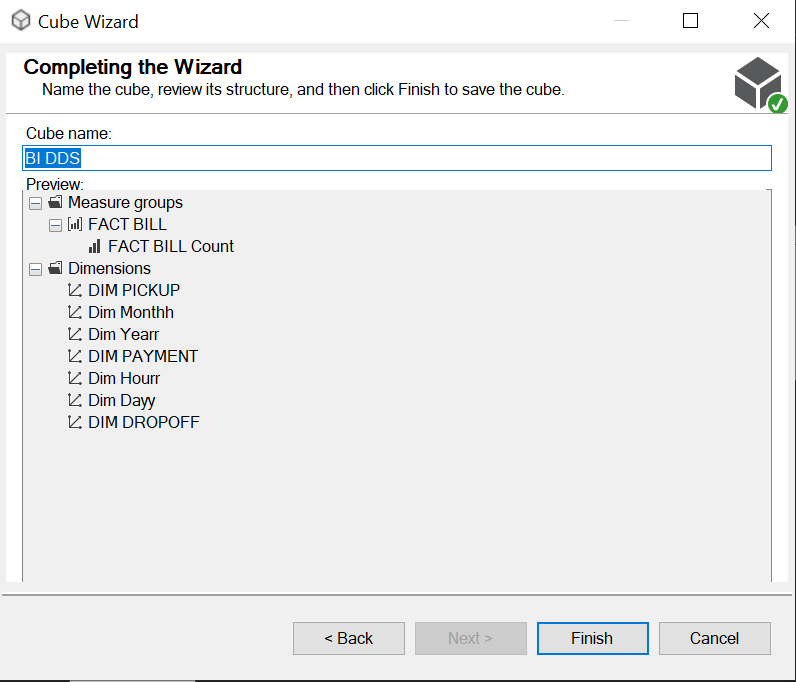
Chọn 1 bảng làm Fact, measure và dimension như hình bên dưới:



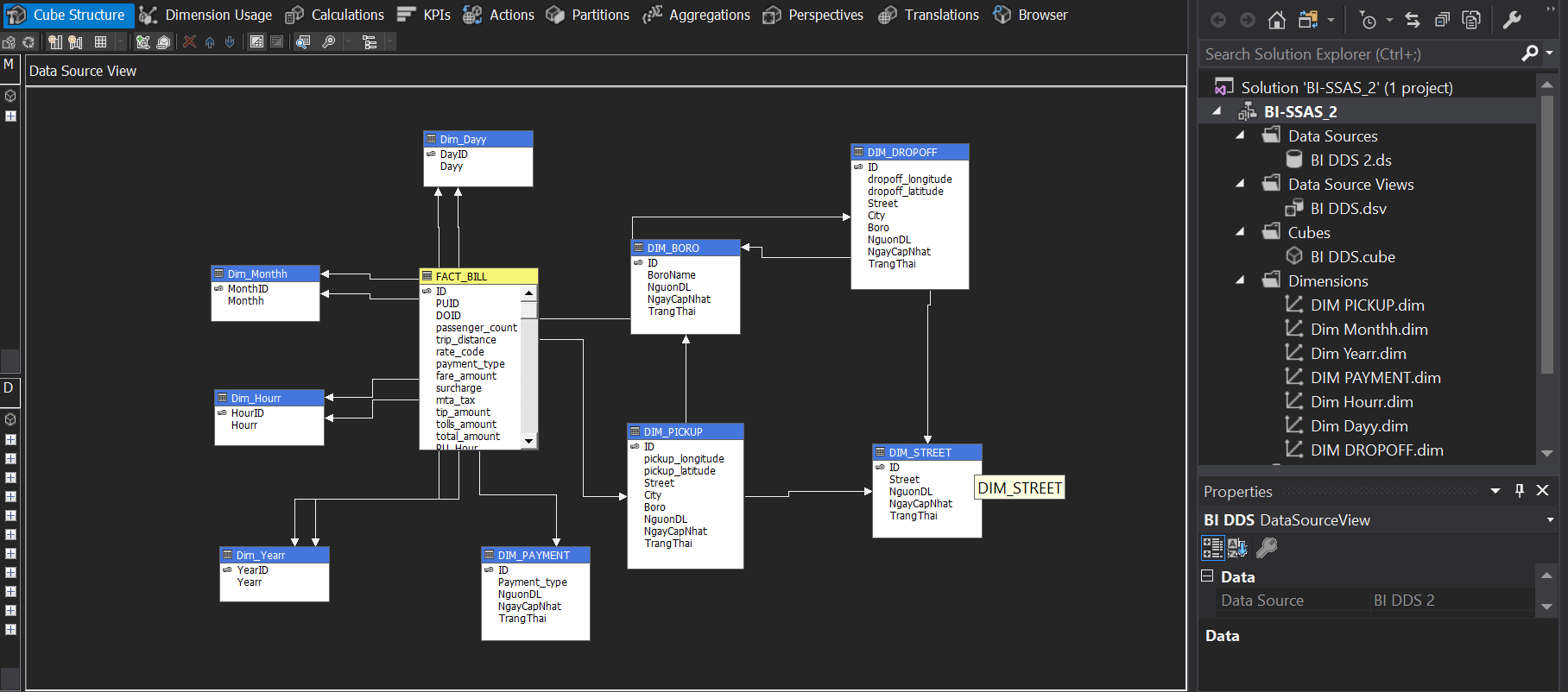






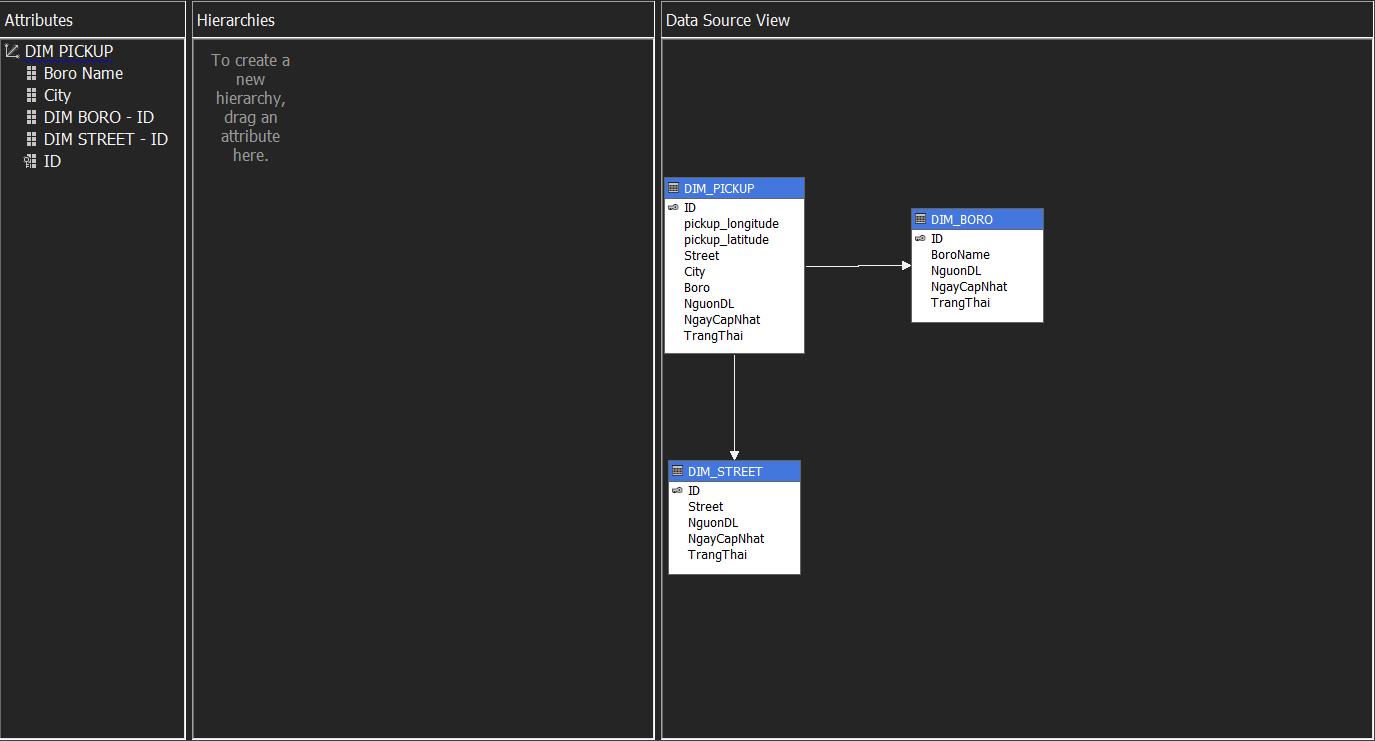


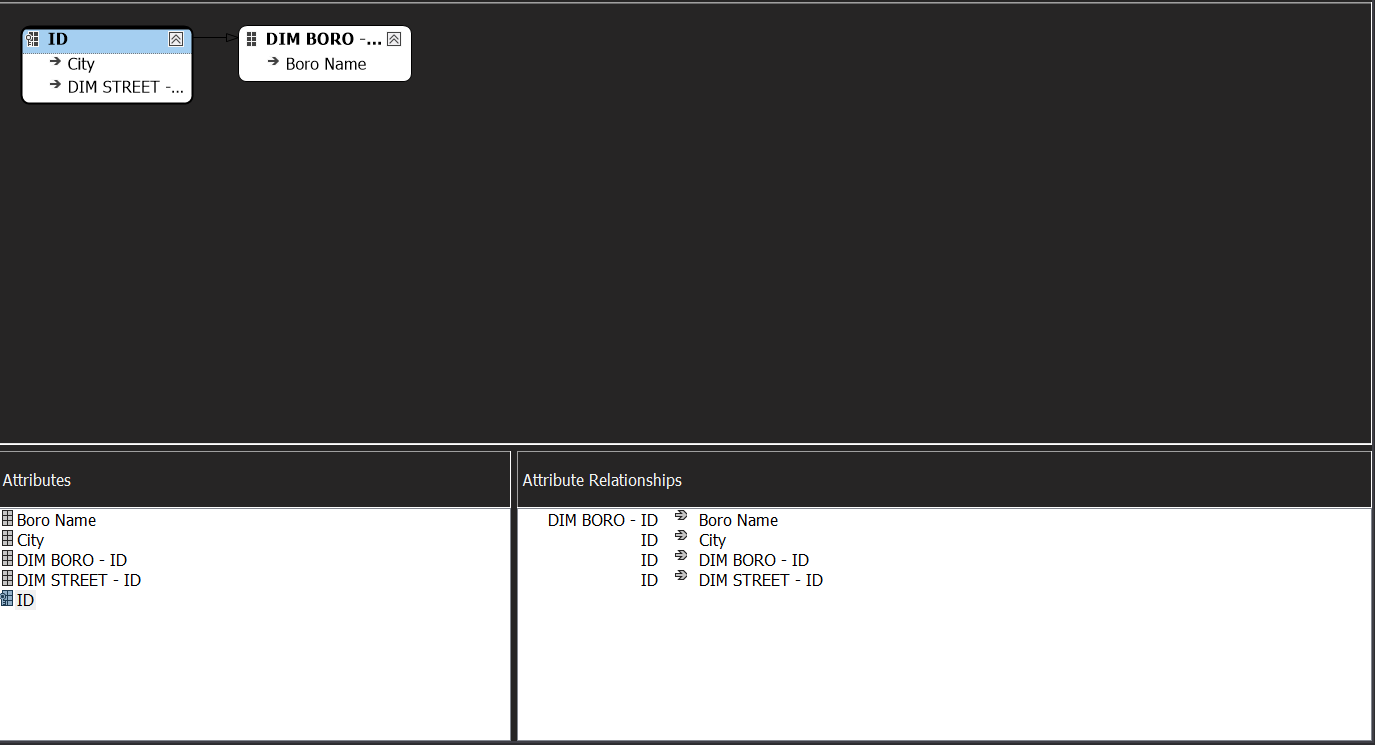
Kết quả sau khi tạo Cube:



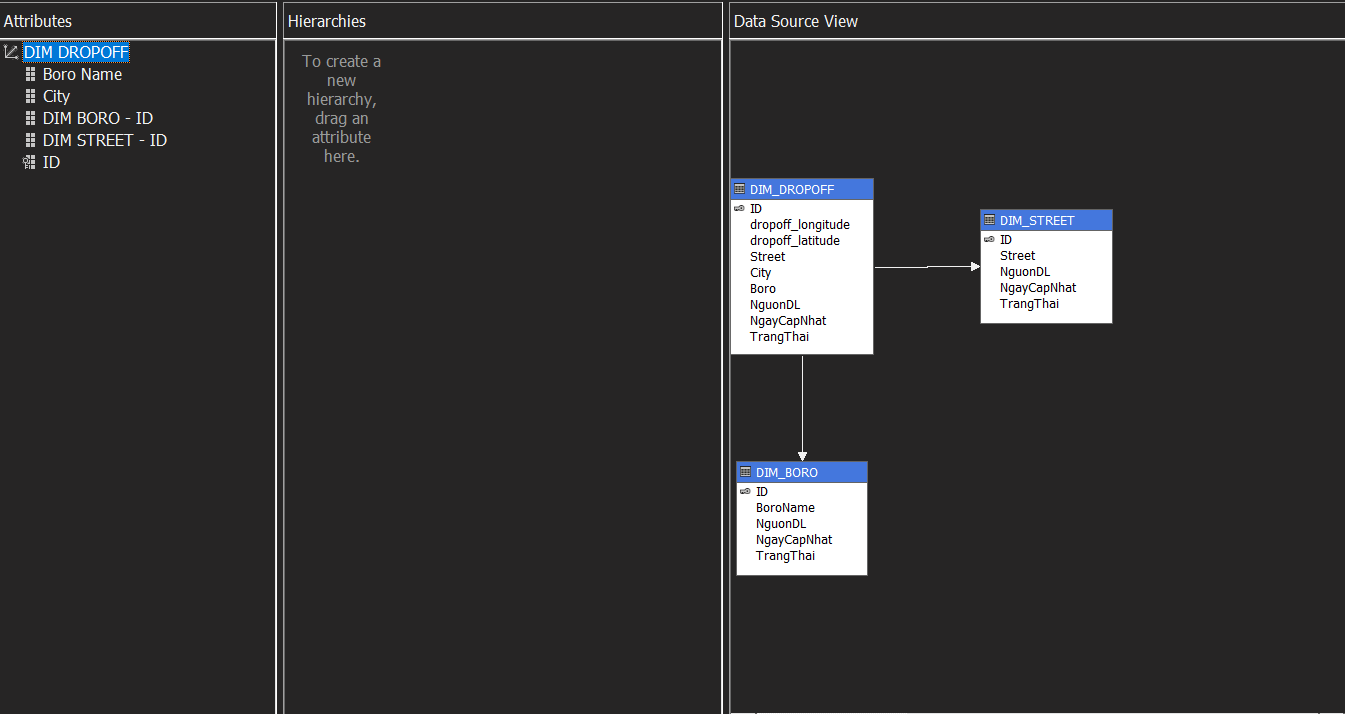
Bước 5: chỉnh sửa các dimemsion

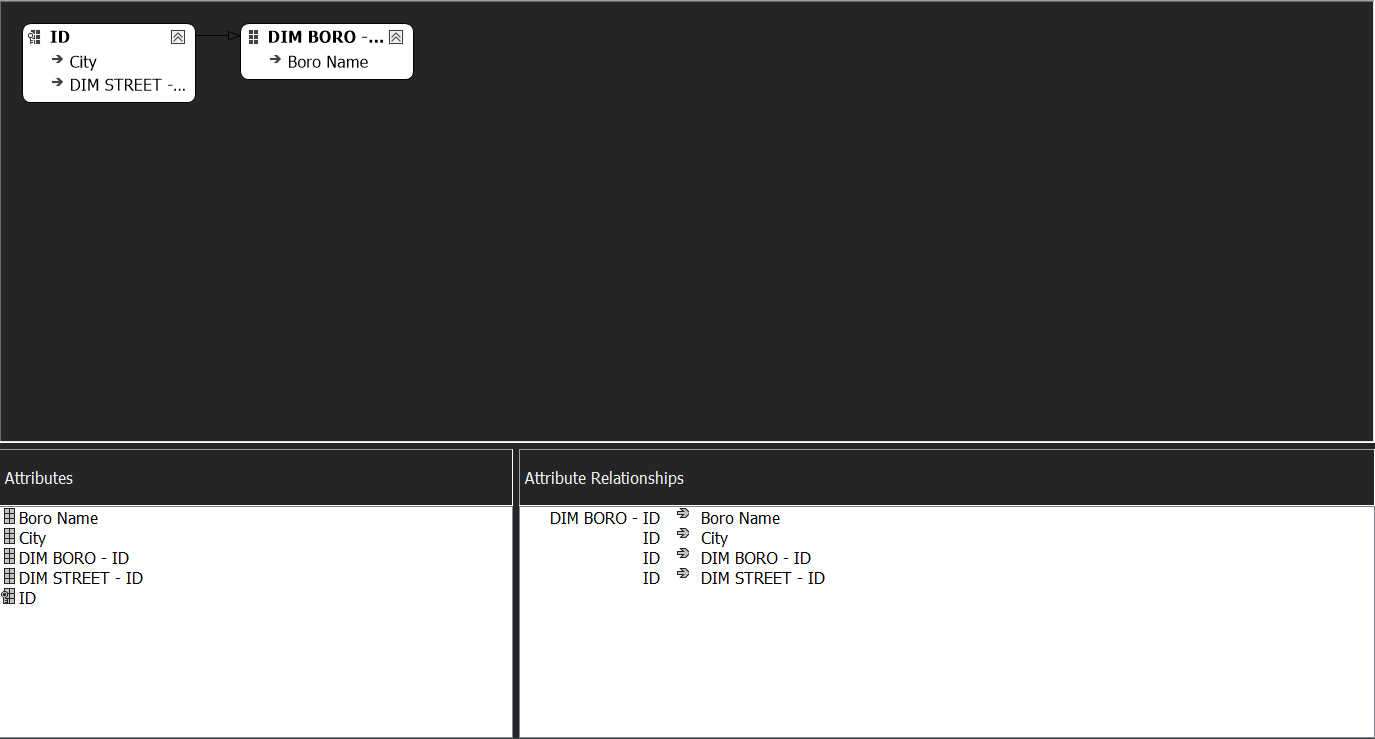
1. ***Dim\_Pickup***



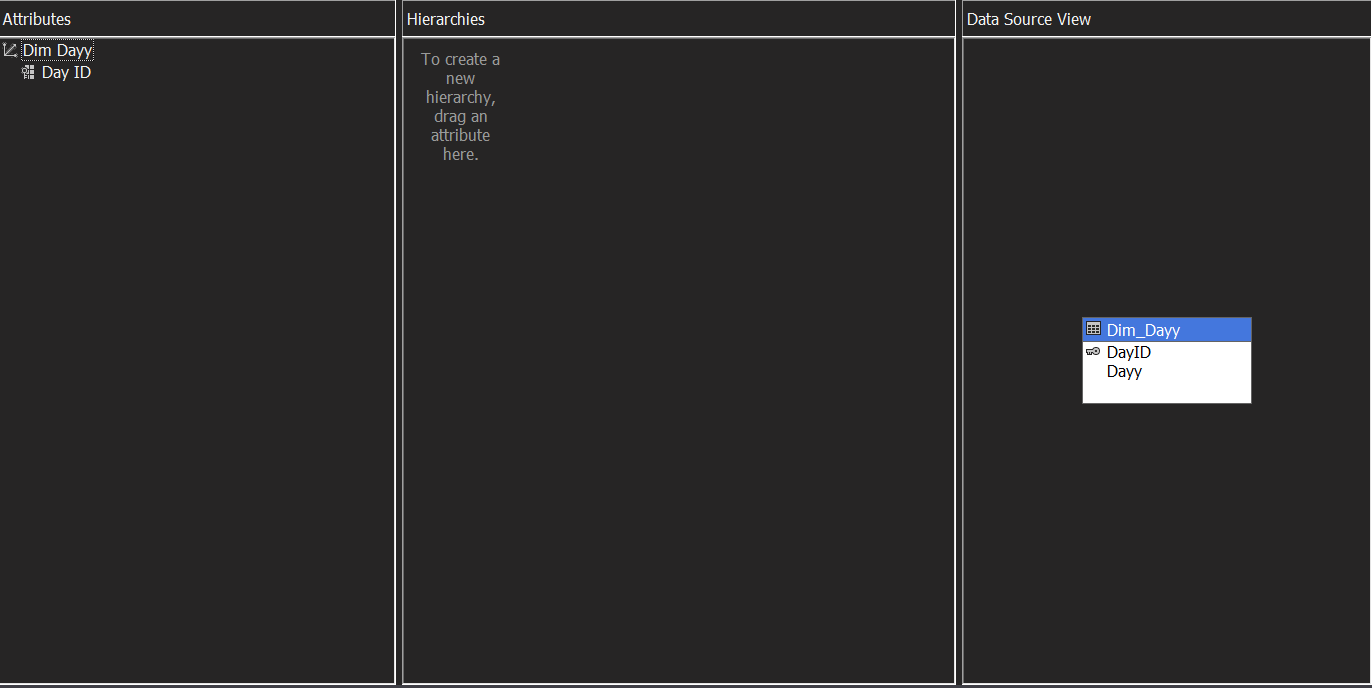


1. ***Dim\_Dropoff***

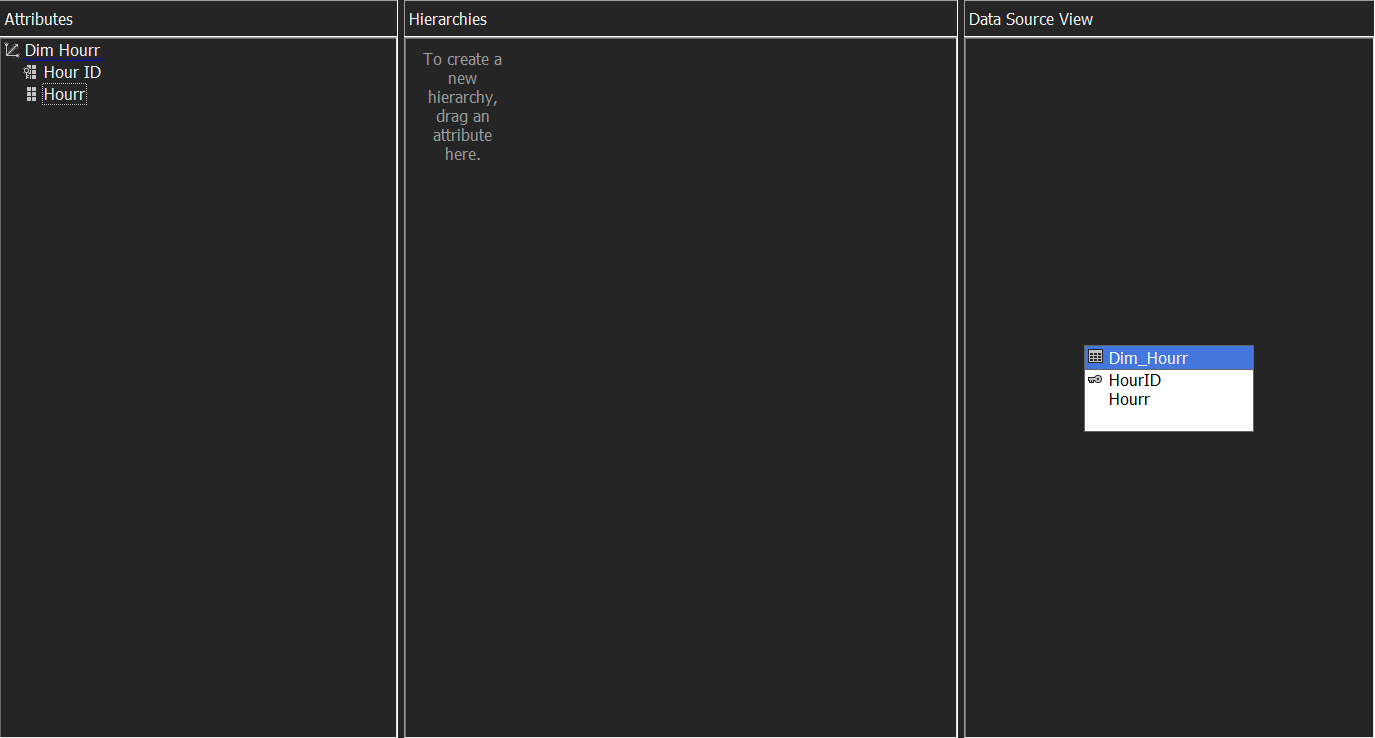




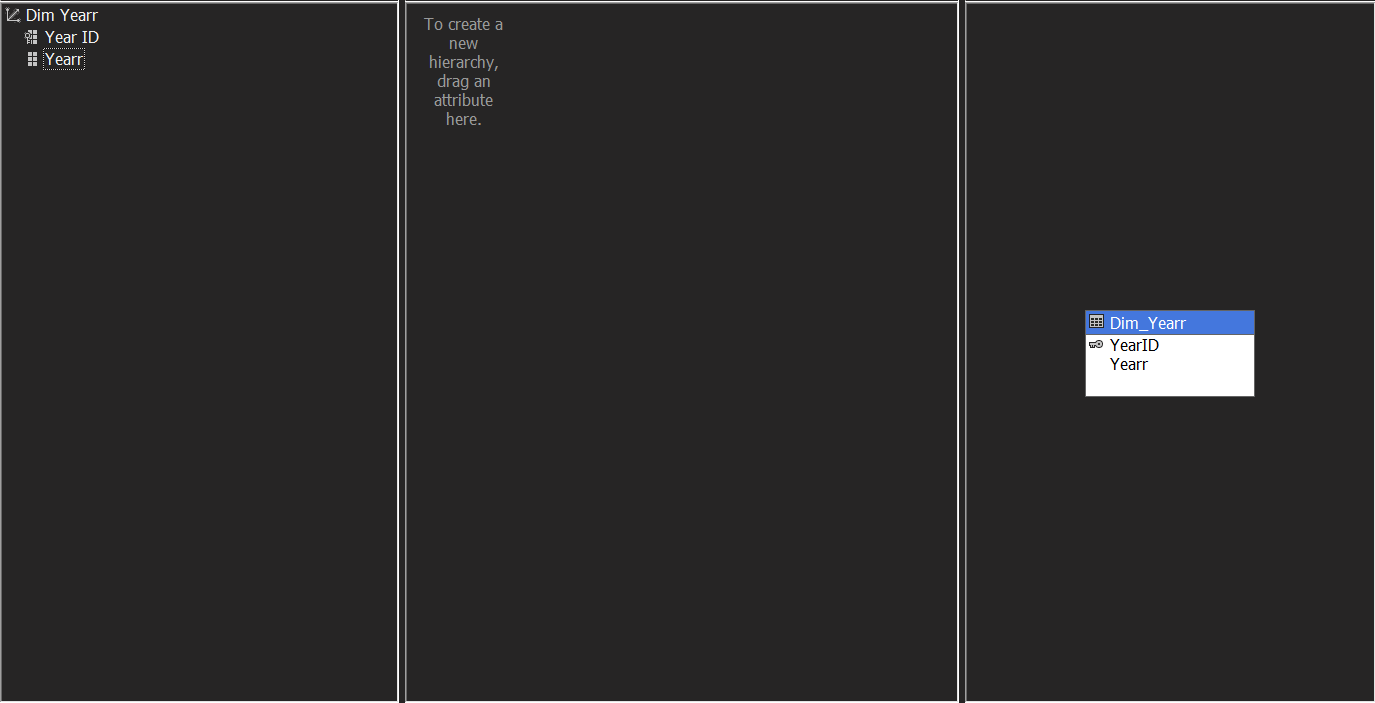
1. ***Dim\_Dayy***



1. ***Dim\_Hourr***



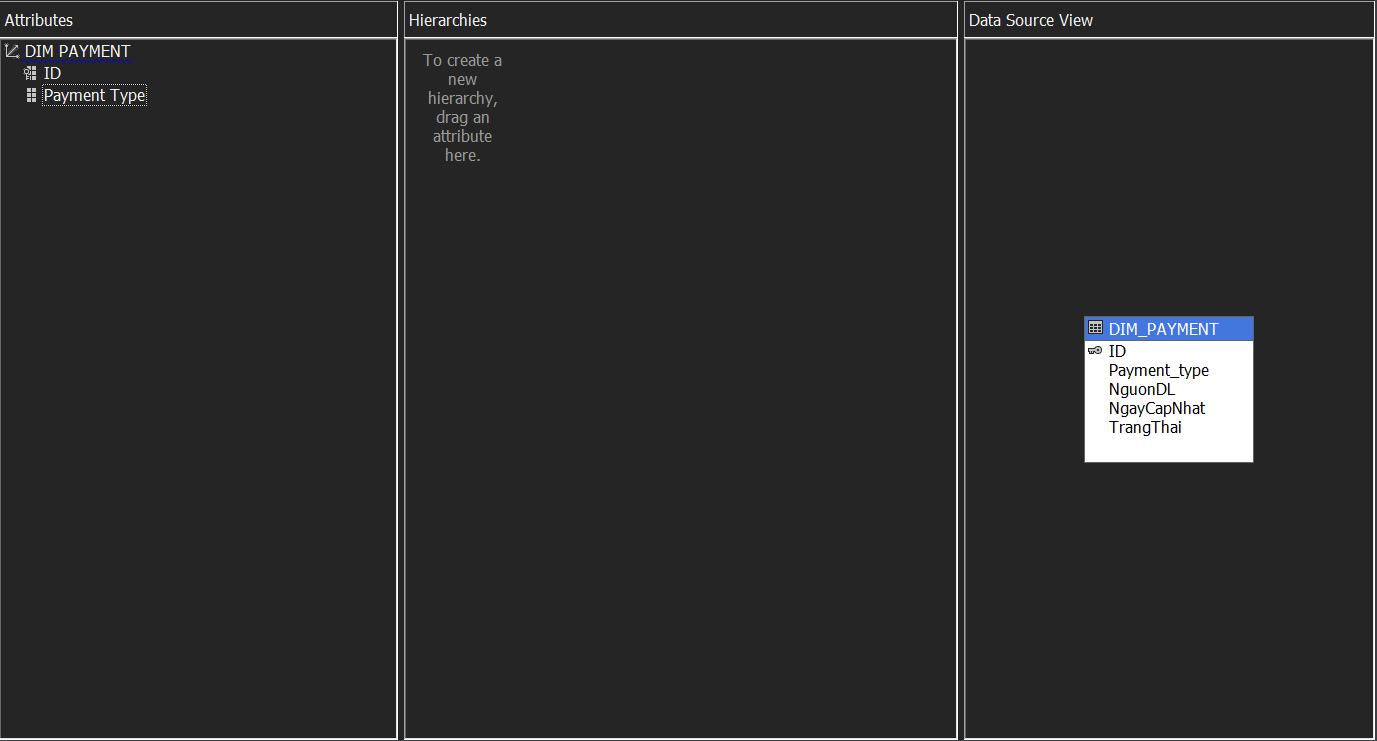
1. ***Dim\_Yearr***



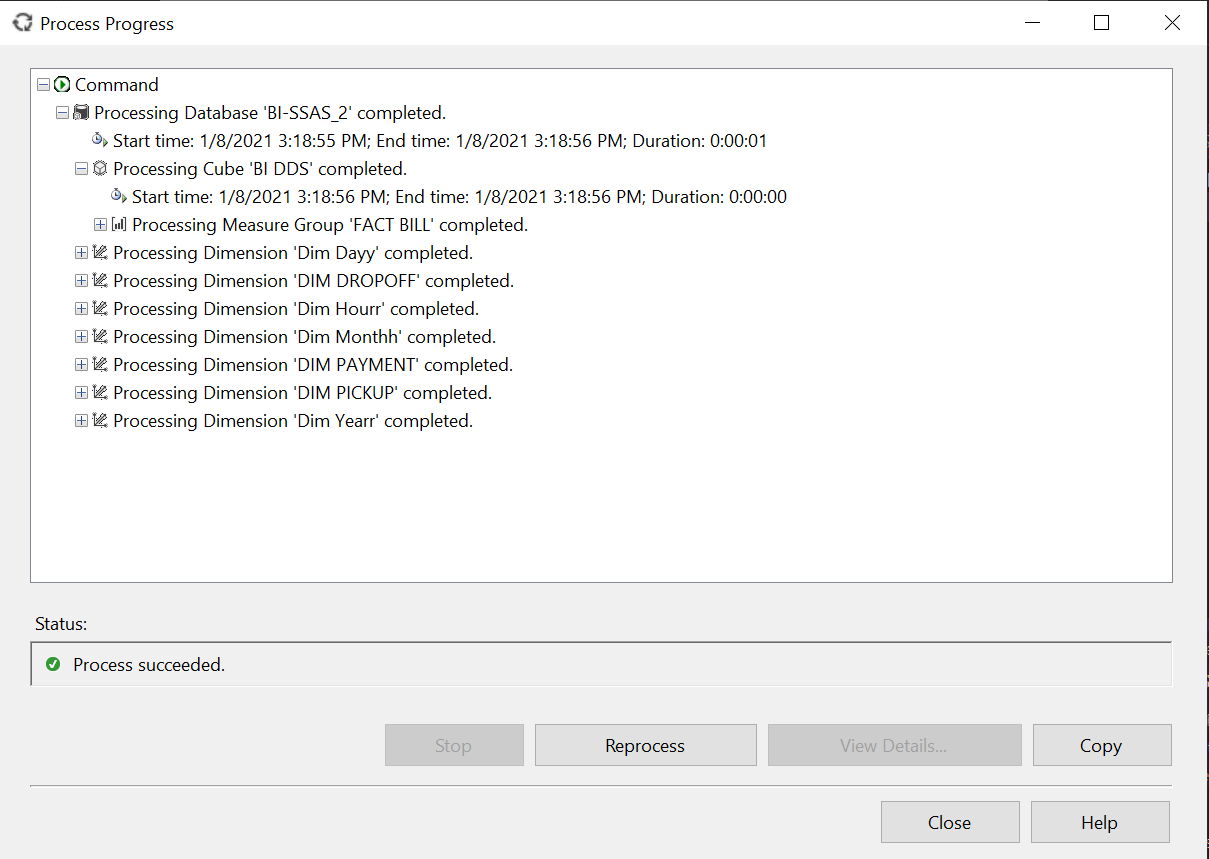
1. ***Dim\_Monthh***

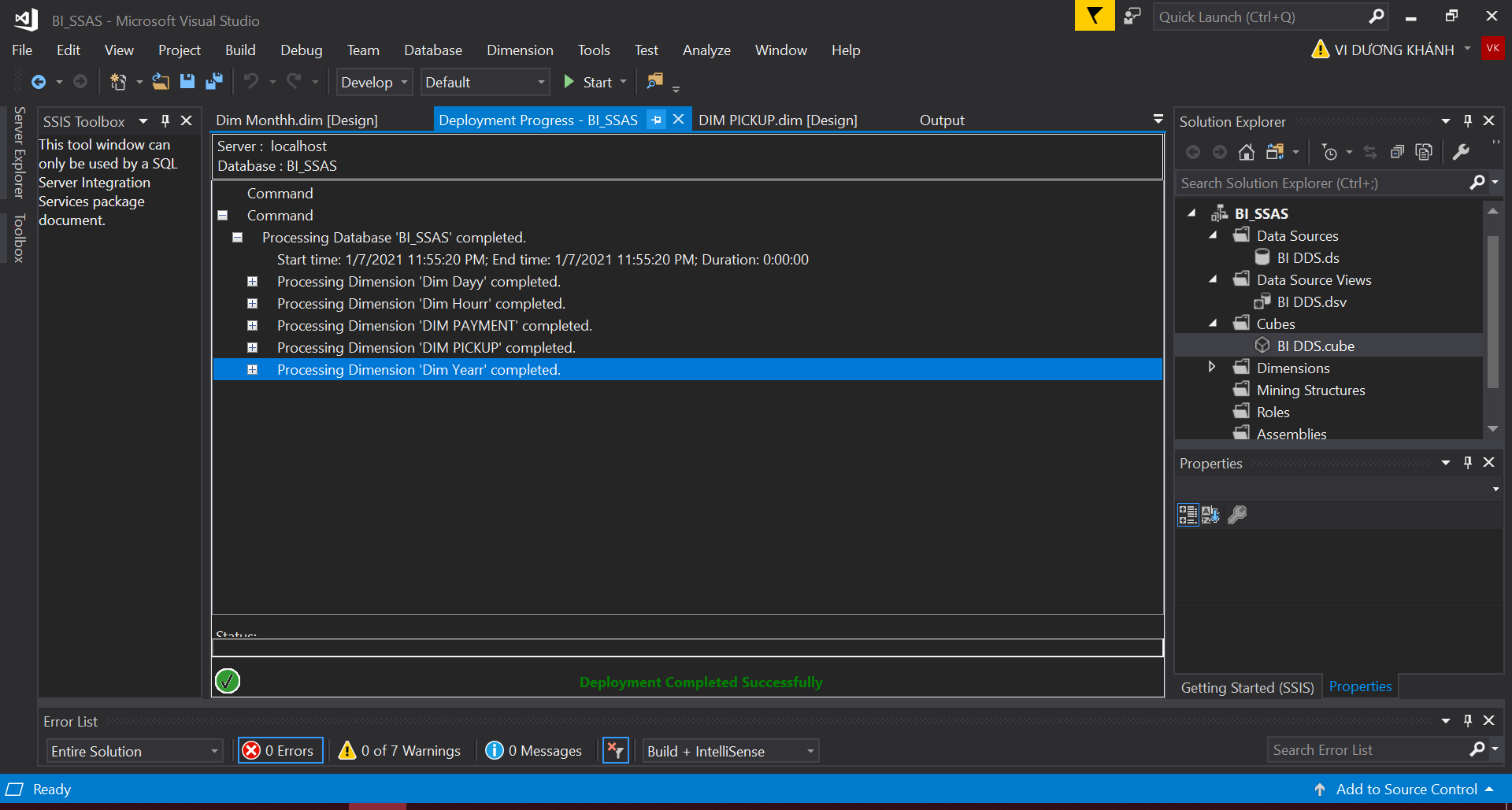


1. ***Dim\_Payment***

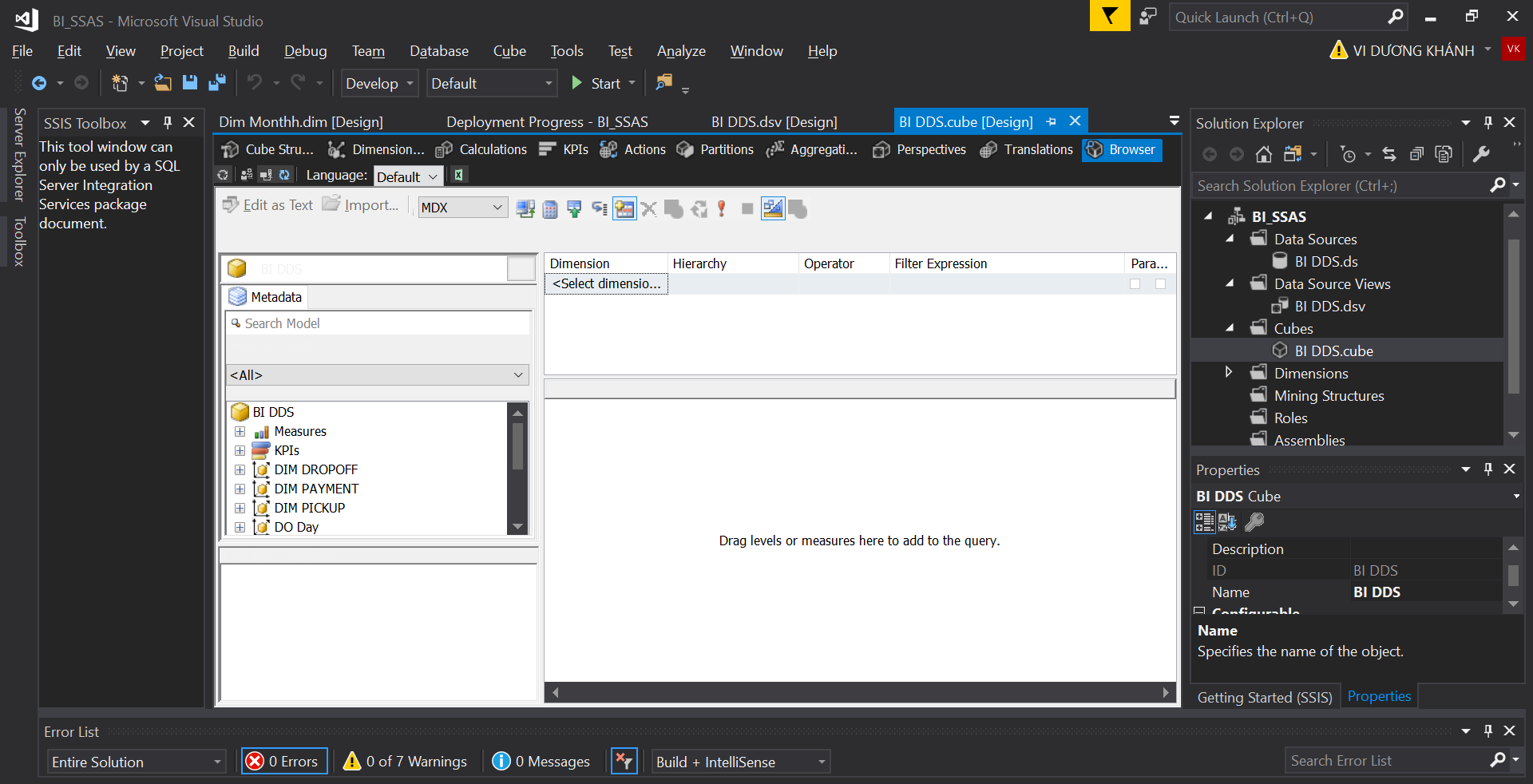


Sau khi chỉnh sửa các dimension, tiến hành process:

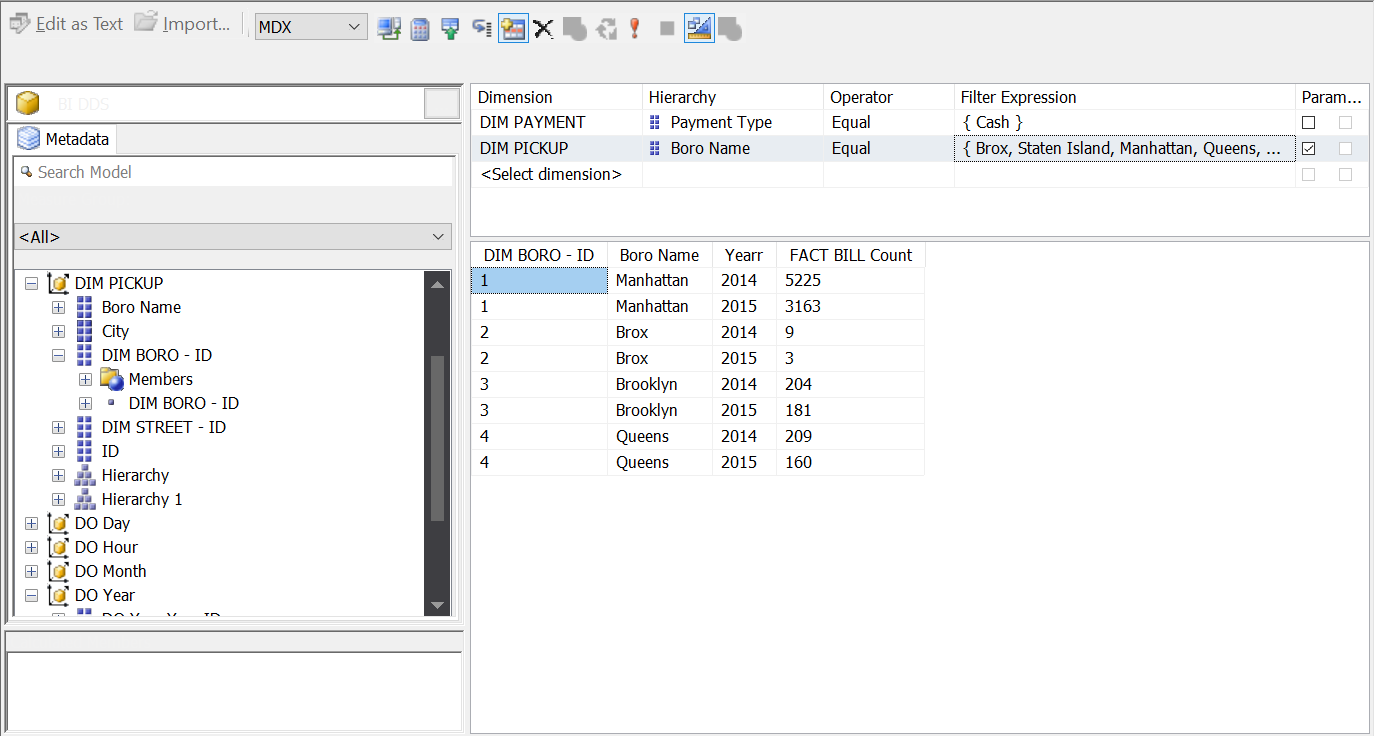




* Kết quả sau khi process:

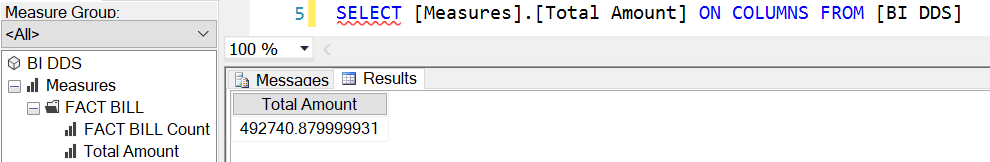


Thử thực hiện truy vấn trên cube và ta có kết quả:

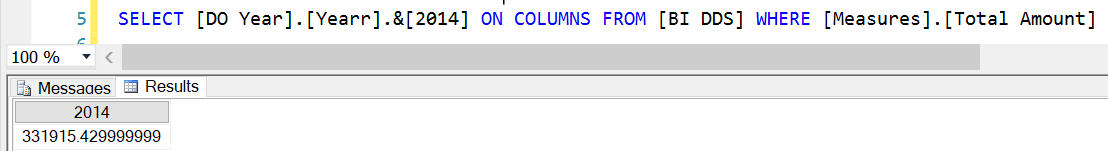


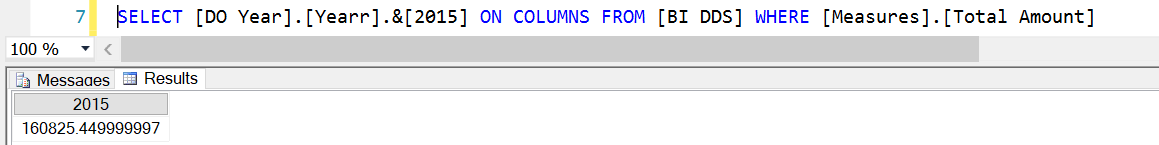
1. **Truy vấn bằng MDX**
   1. **Tính doanh thu tổng**

* Doanh thu tổng

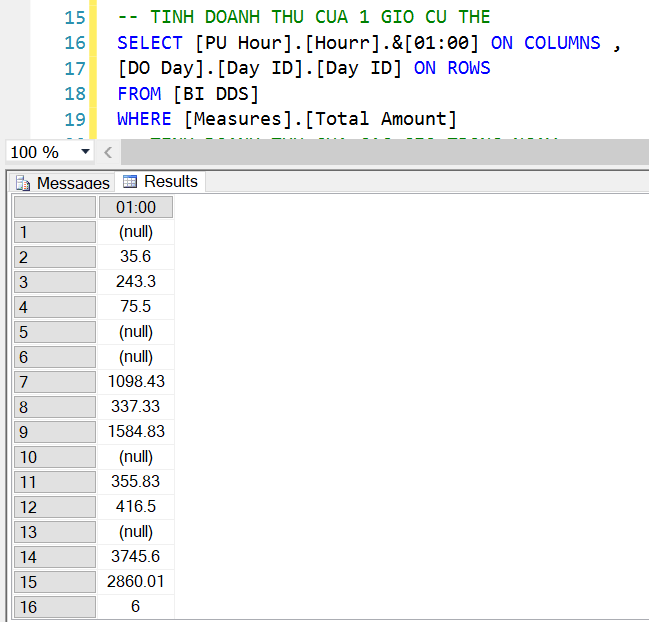


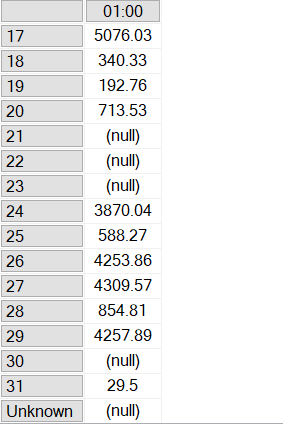
* Doanh thu theo từng năm





* 1. **Tính doanh thu của một giờ**
* Tính doanh thu của một giờ cụ thể trong ngày

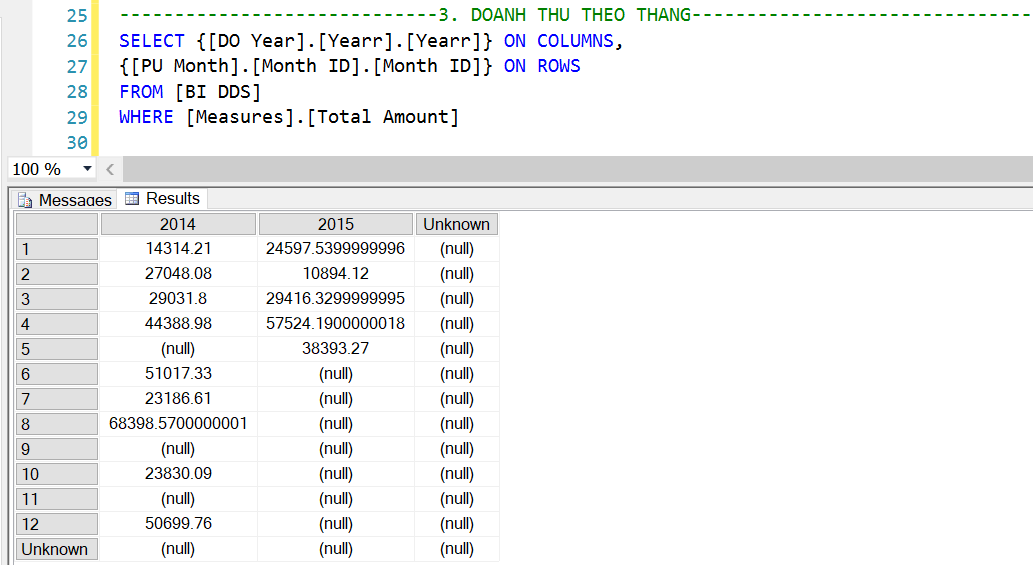




* Tính doanh thu của các giờ trong ngày



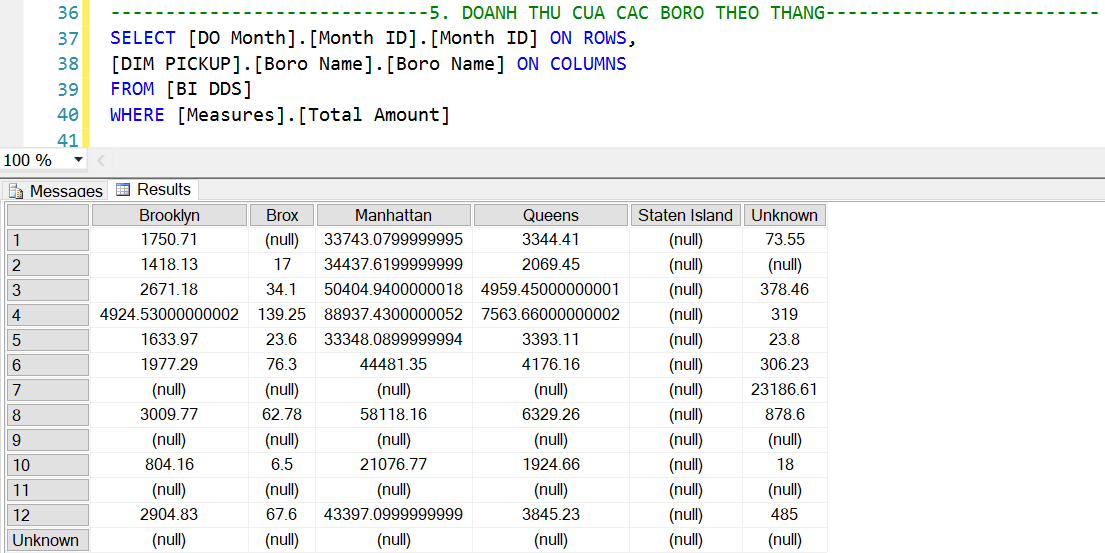
* 1. **Tính doanh thu theo tháng trong năm**



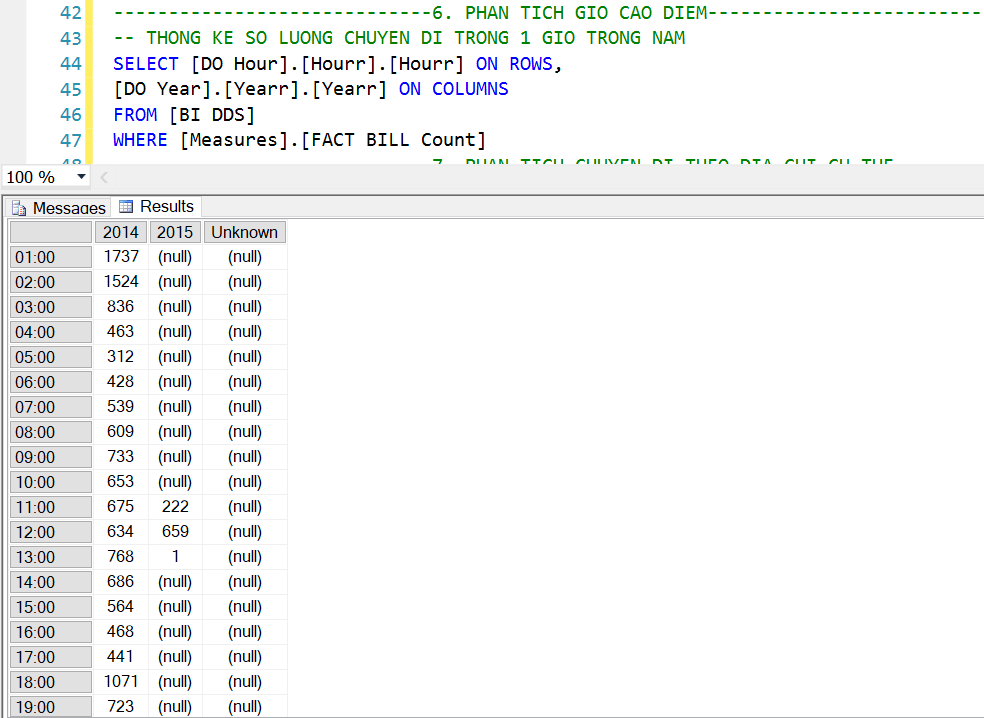
* 1. **Doanh thu theo boro**
* Doanh thu theo boro theo năm

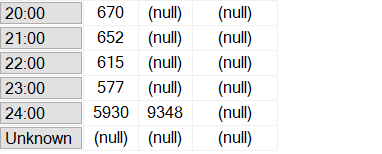


* Doanh thu theo boro theoo tháng

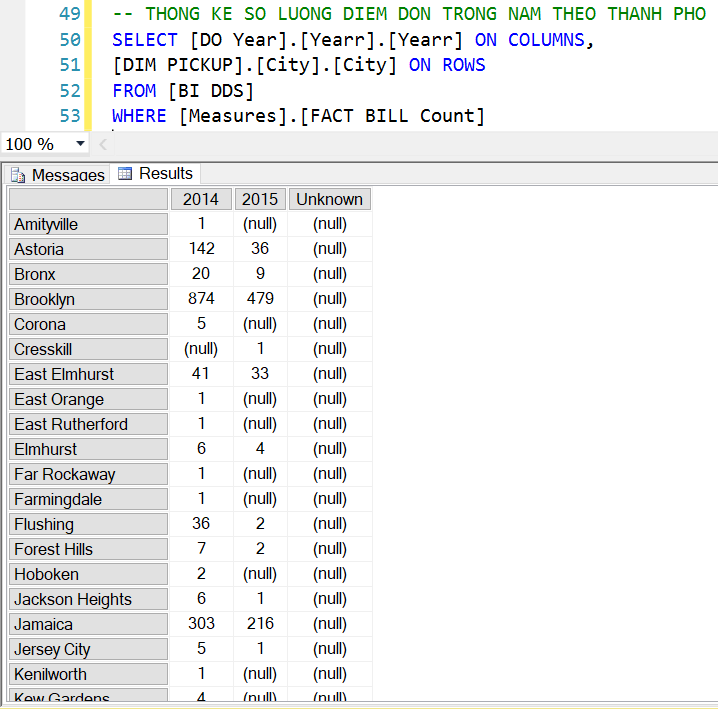


* 1. **Phân tích chuyến đi theo giờ để biết đâu là giờ cao điểm**

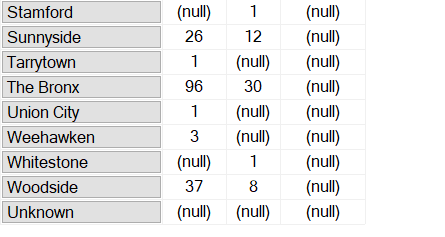




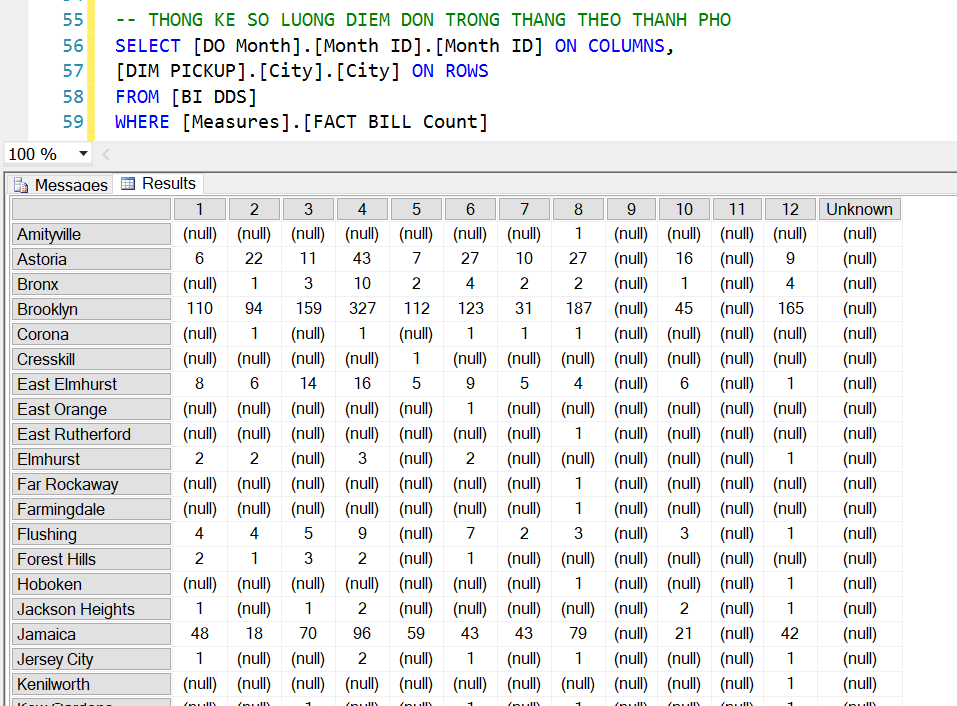
* 1. **Phân tích chuyến đi theo địa chỉ cụ thể**
* Thống kê số lượng điểm đón trong năm theo thành phố

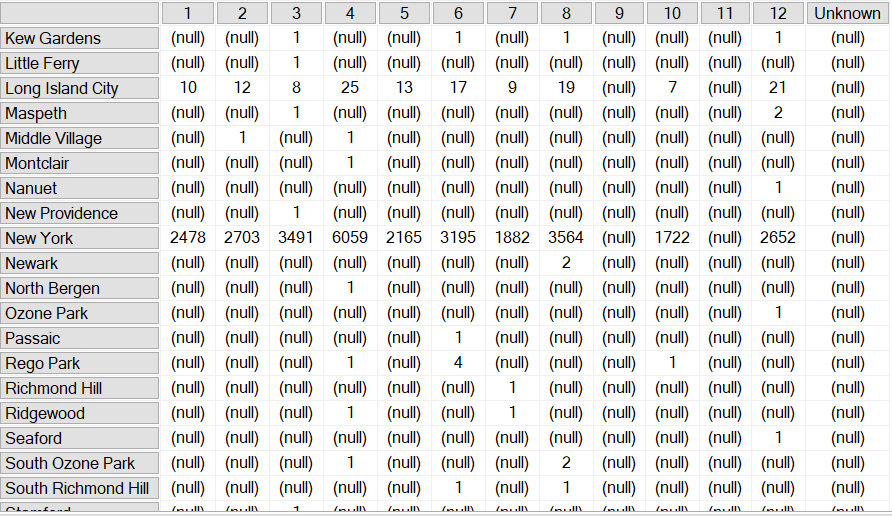


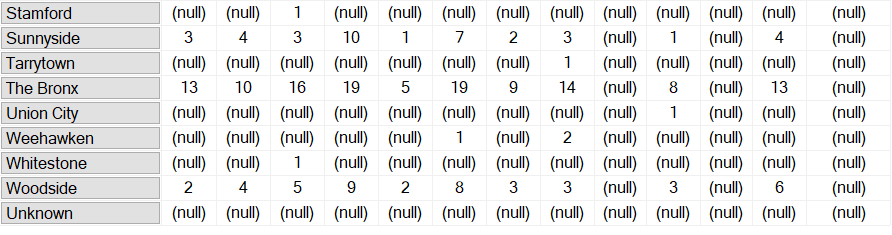




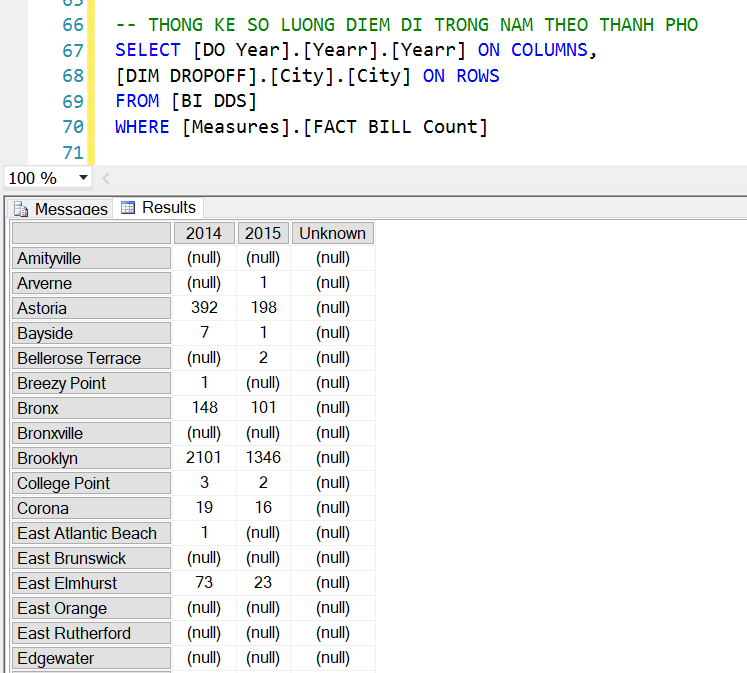
* Thống kê điểm đón của thành phố theo tháng



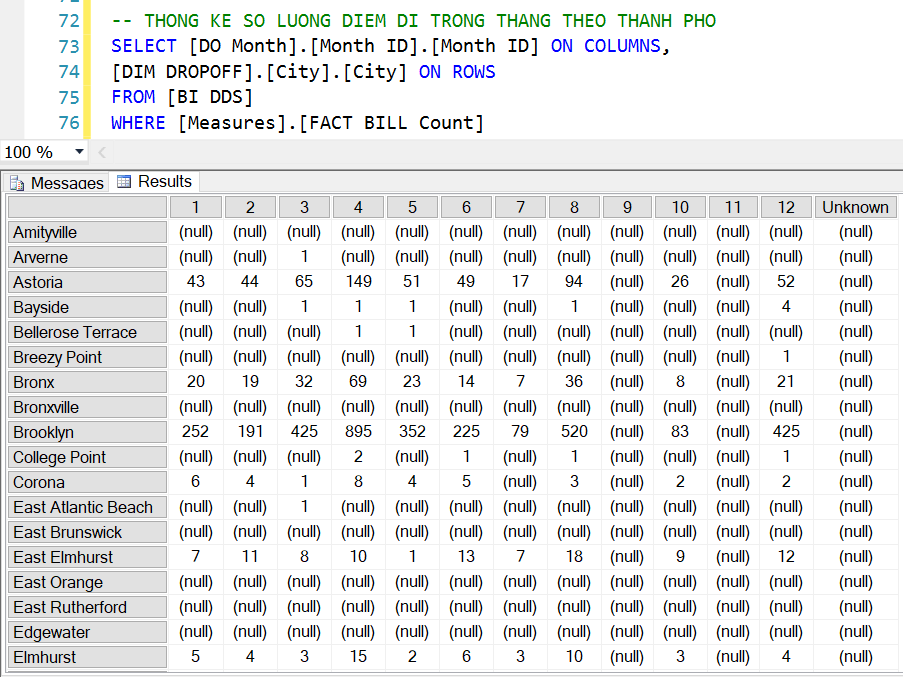




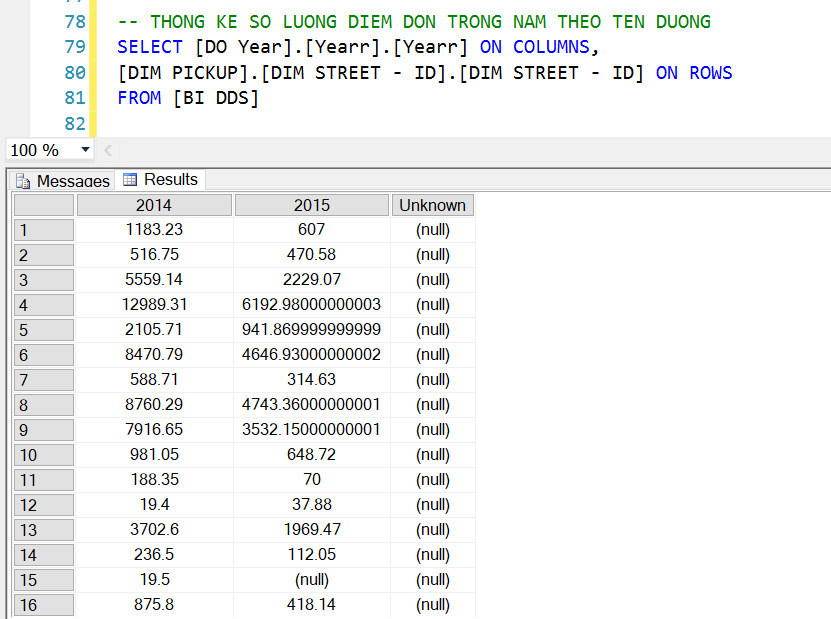
* Thống kê số lượng điểm đi trong năm theo thành phố



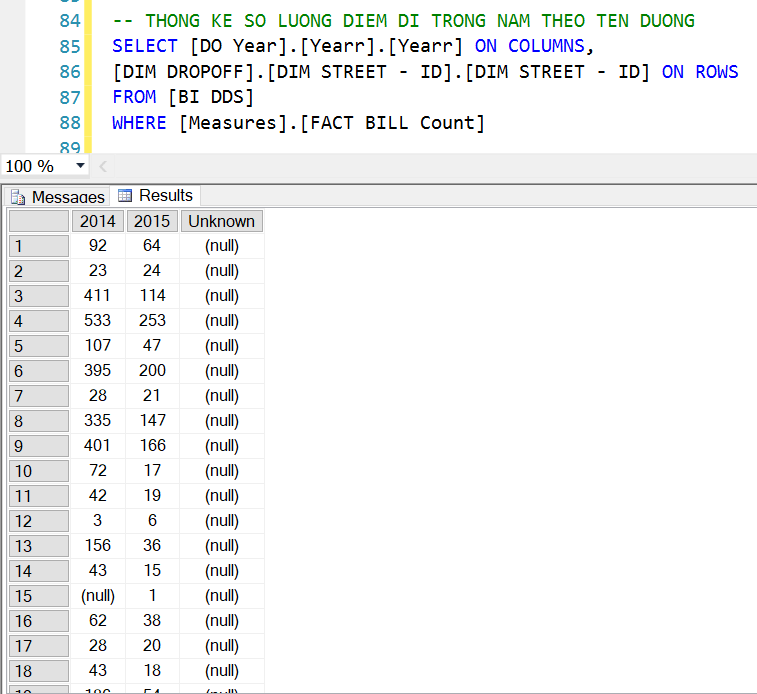
* Thống kê số lượng điểm đi trong tháng theo thành phố



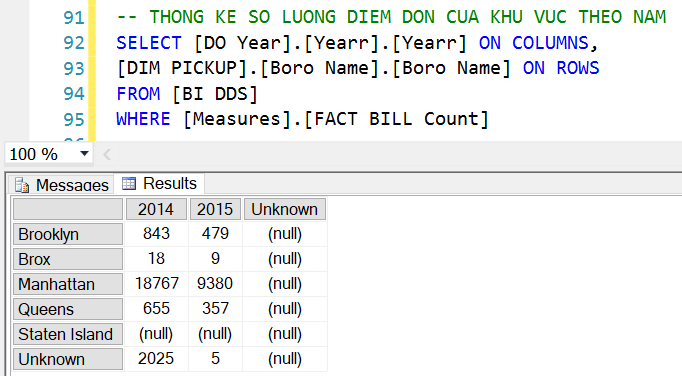
* Thống kê số lượng điểm đón trong năm theo đường



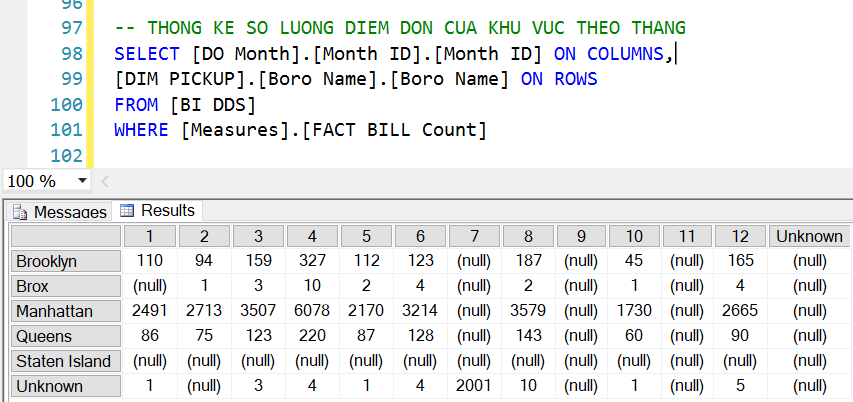
* Thống kê số lượng điểm đi theo tên đường



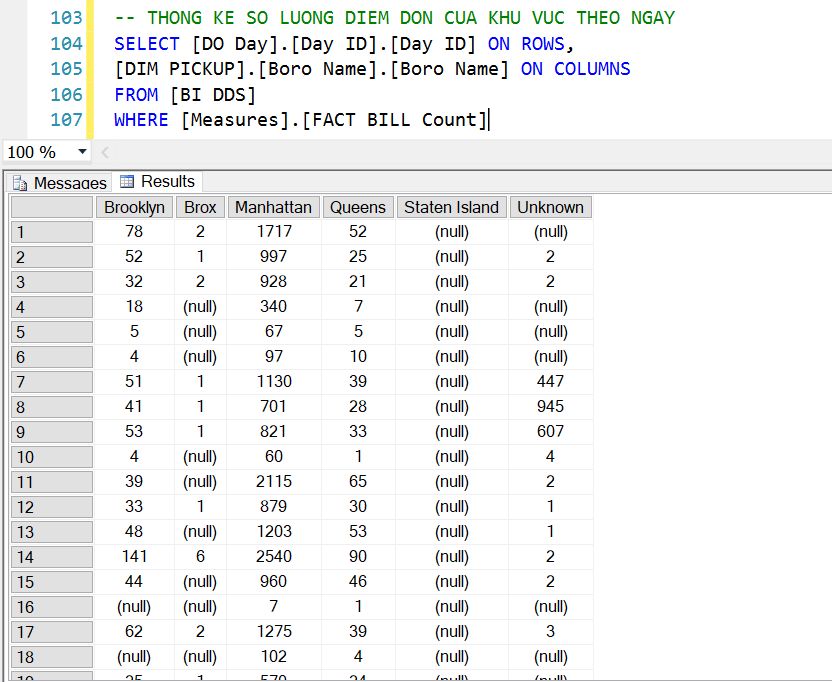
* 1. **Phân tích chuyến đi theo boro cụ thể**
* Thống kê số lượng điển đón của khu vực theo năm

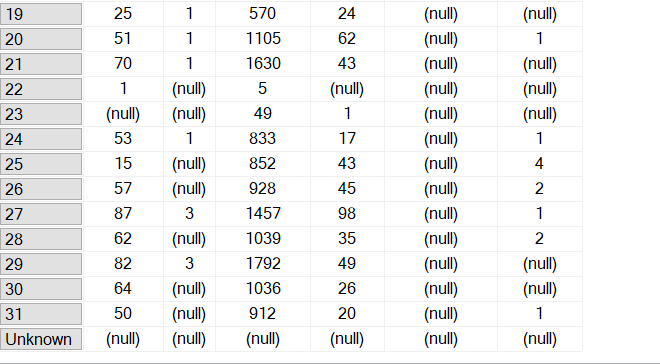


* Thống kê số lượng điểm đón của khu vực theo tháng

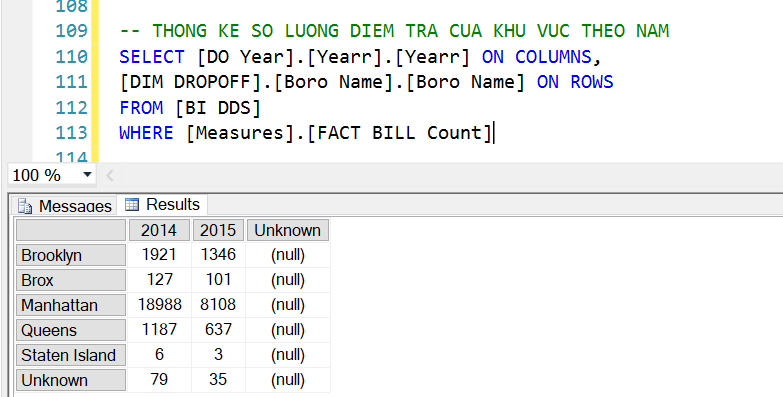


* Thống kê số lượng điểm đón của khu vực theo ngày

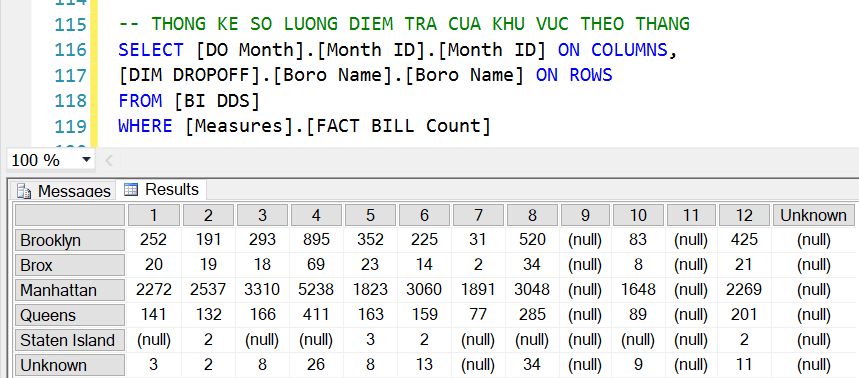




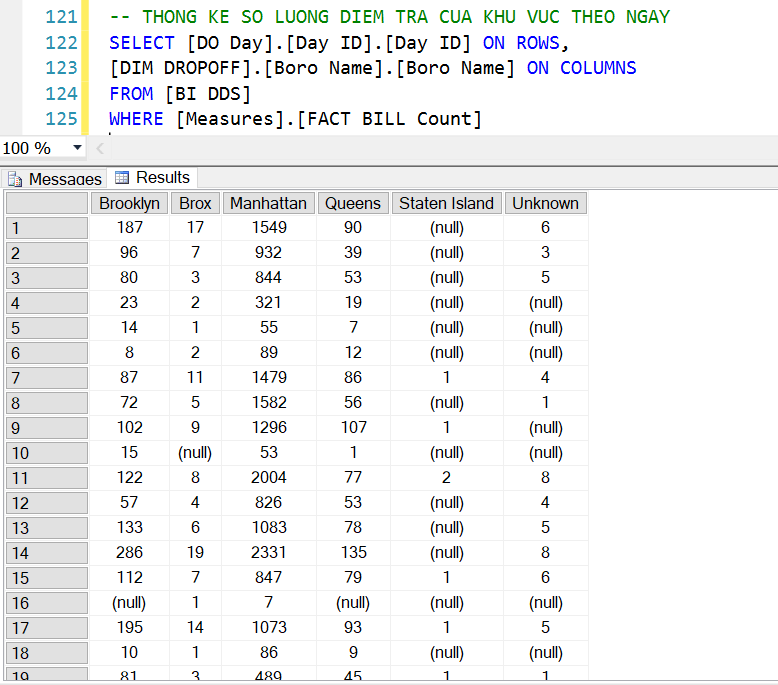
* Thống kê số lượng điểm trả của khu vực theo năm

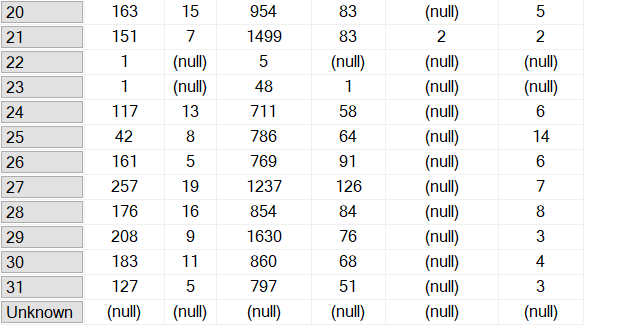


* Thống kê số lượng điểm trả cuả khu vực theo tháng

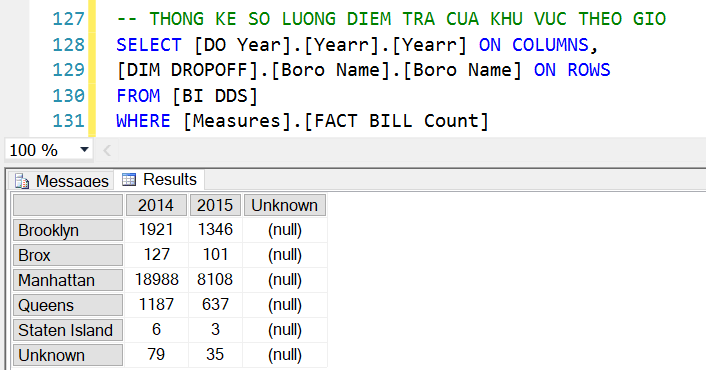


* Thống kê số lượng điểm trả cuả khu vực theo ngày

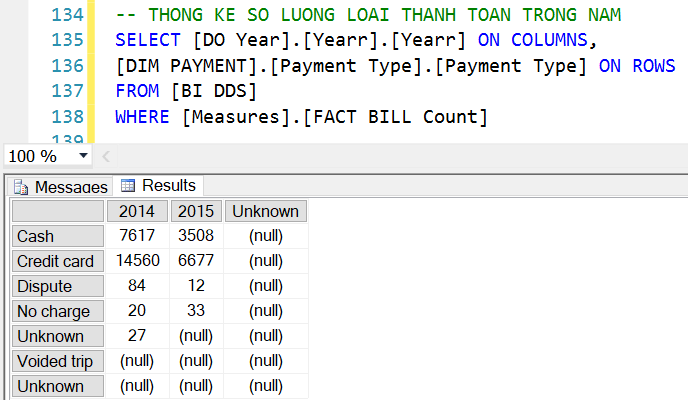




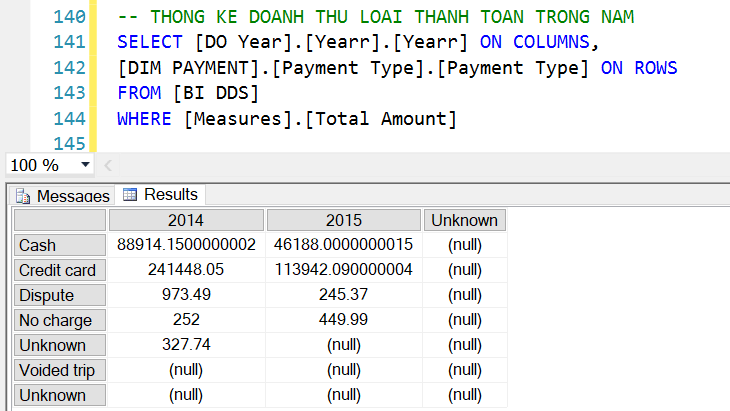
* Thống kê số lượng điểm trả cuả khu vực theo giờ



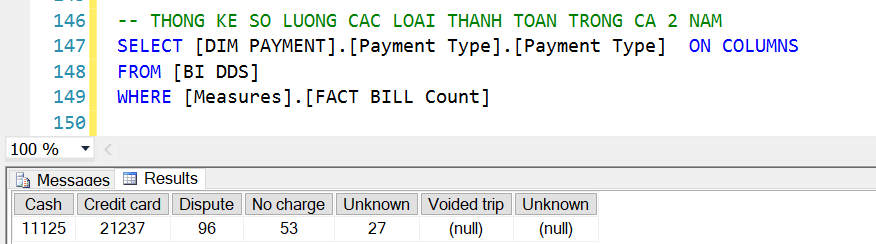
* 1. **Phân tích chuyến đi theo loại thanh toán**
* Thống kê số lượng loại thanh toán trong năm



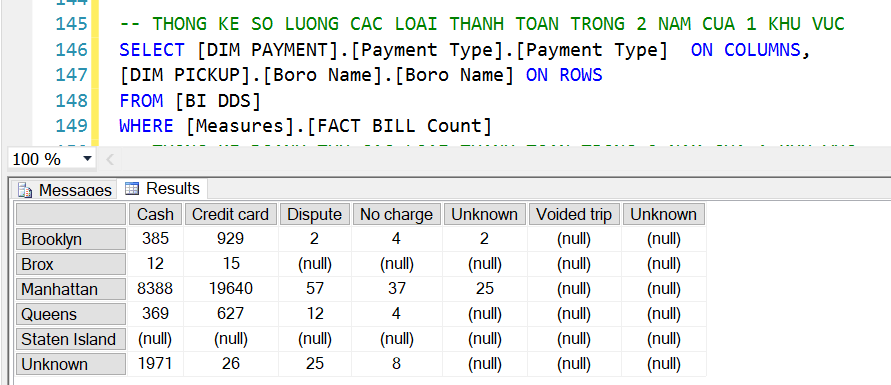
* Thống kê doanh thu của loại thanh toán trong năm



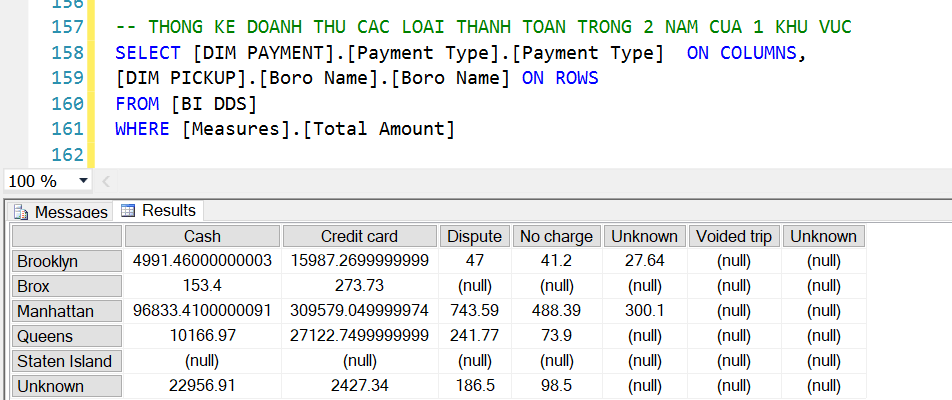
* Thống kê số lượng các loại thanh toán trong 2 năm



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

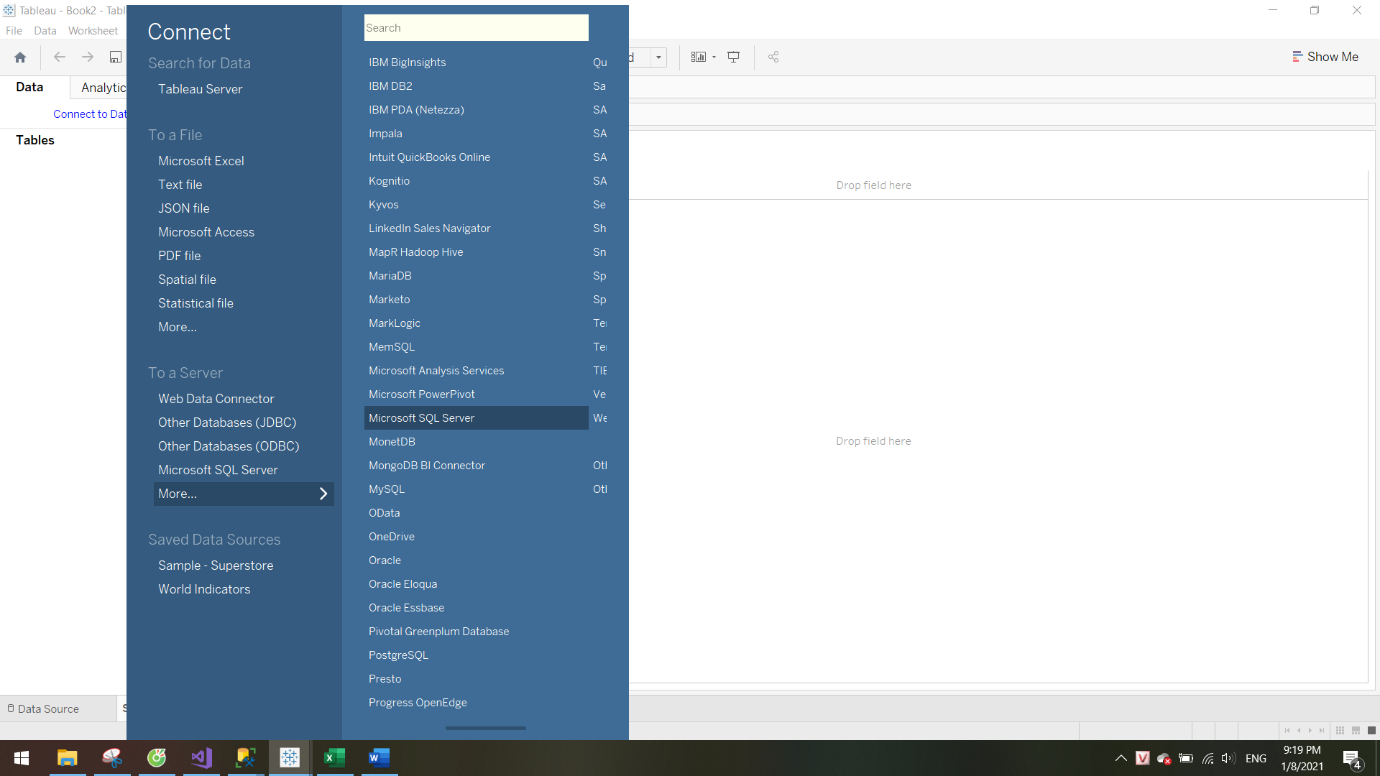


* Thống kê doanh thu các loại thanh toán trong 2 năm của 1 khu vực

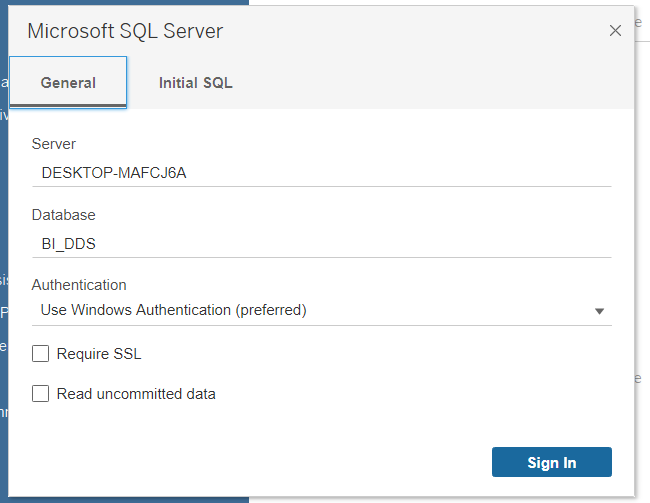


1. **Mining**
2. **KPI**

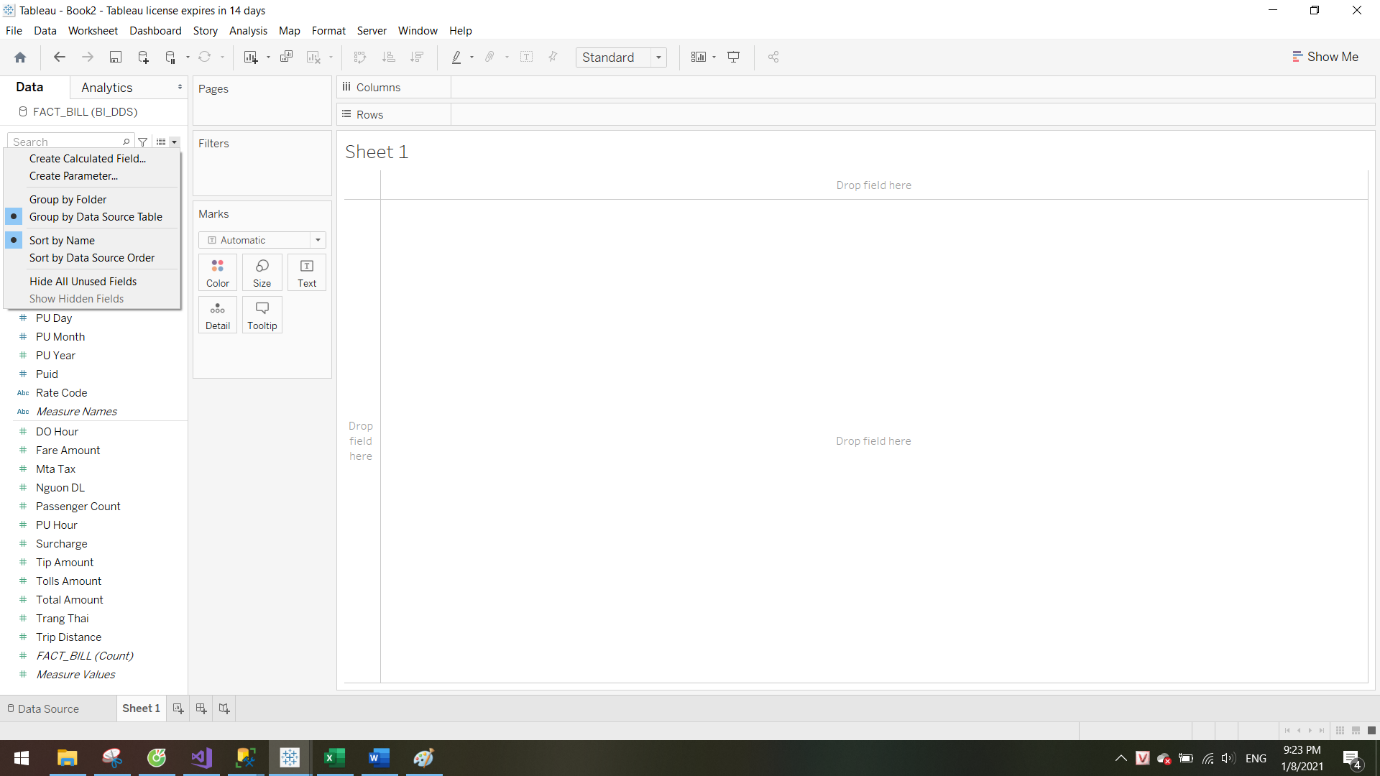
Bước 1: Connect đến SQL Server



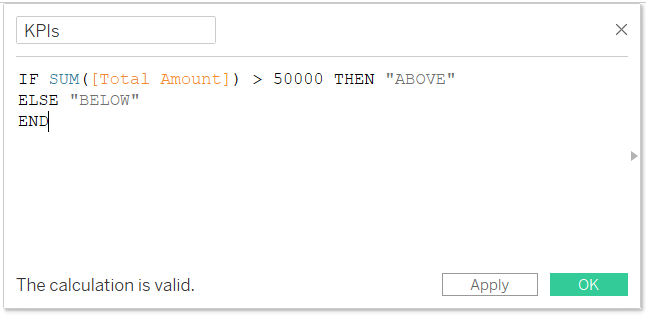
Bước 2: Kết nối đến database



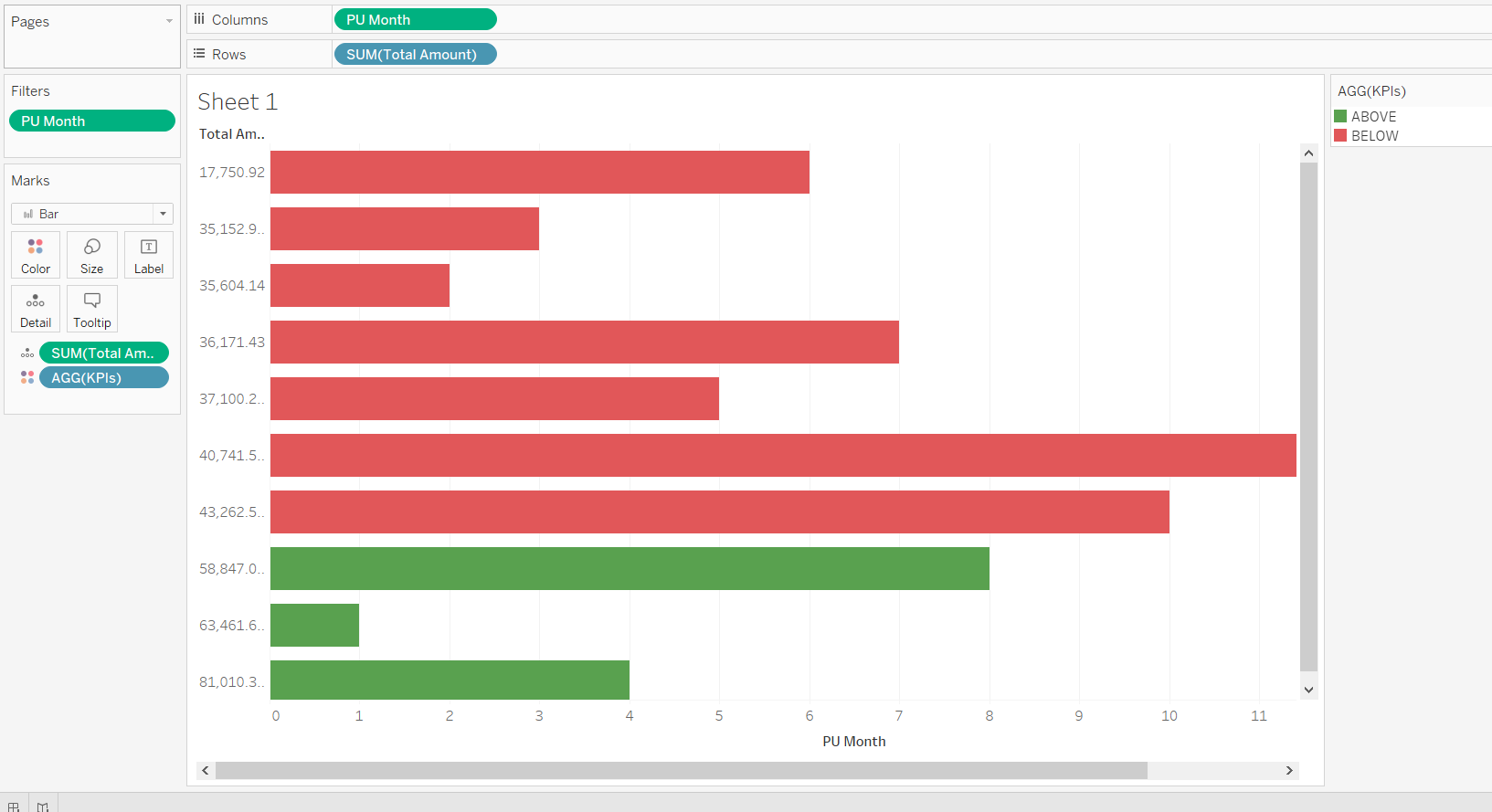
Bước 3: Thêm Fact\_Bill vào sheet



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



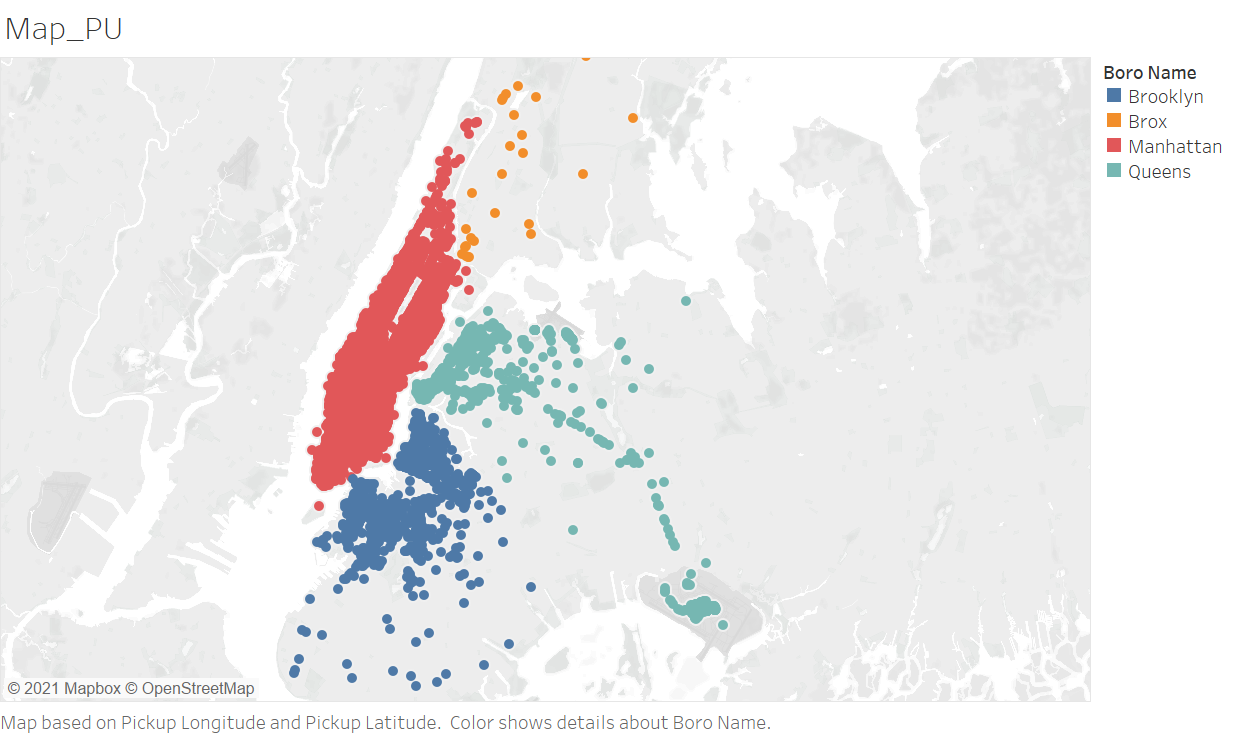
Và đây là kết quả của tổng doanh thu theo tháng, với màu xanh là đạt KPI và đỏ là chưa đạt.



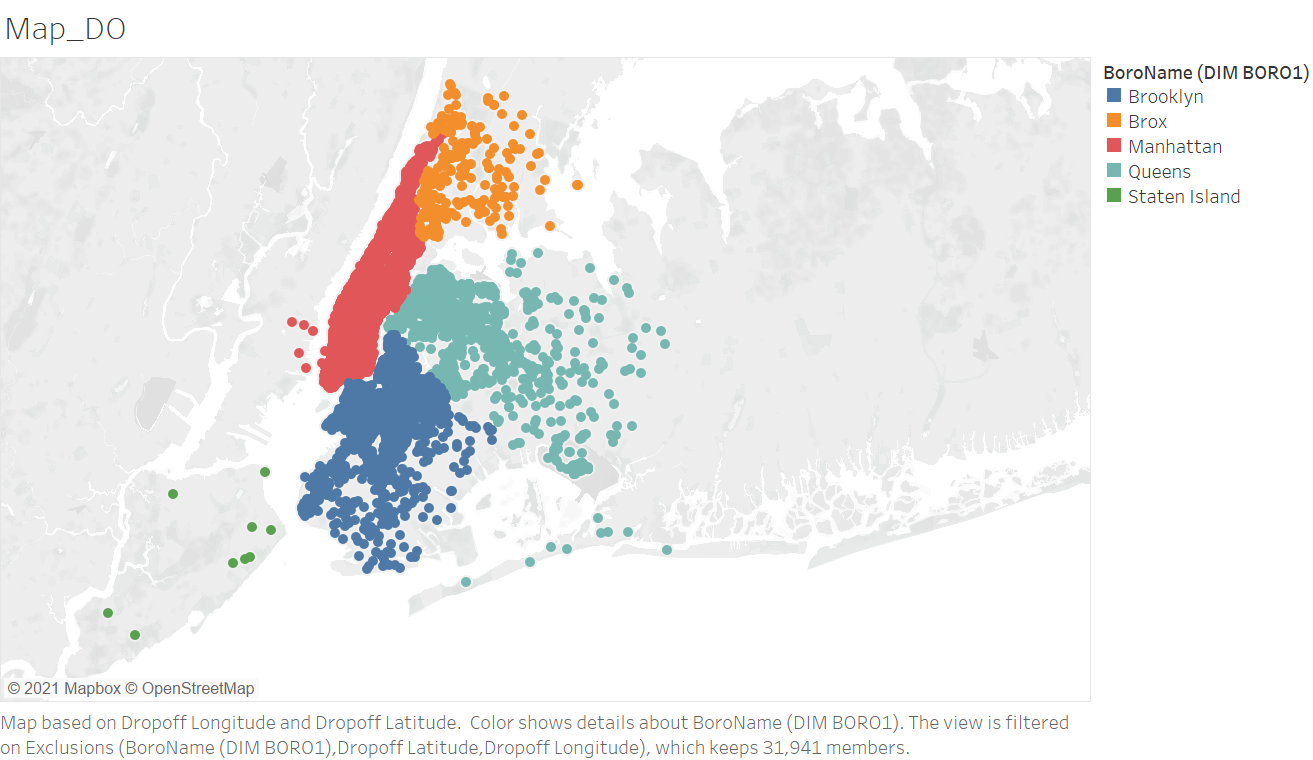
1. **Report**

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

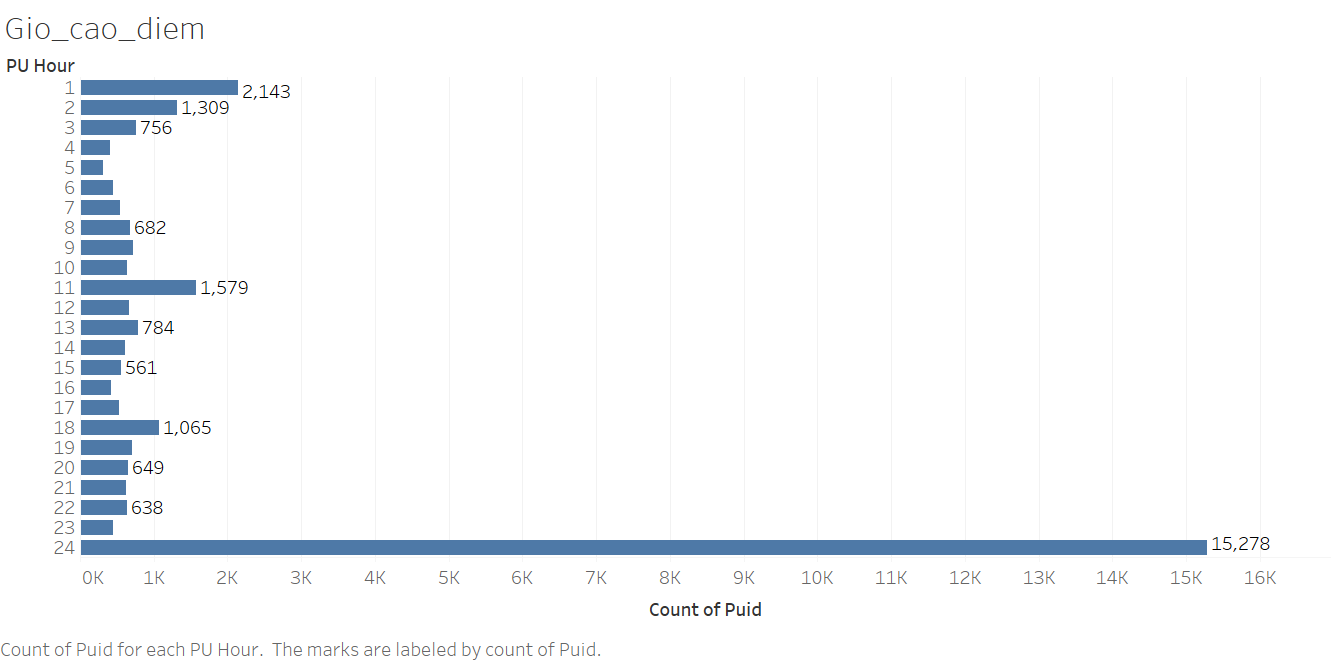
* 1. **Phân tích các điểm đón và trả khách bằng regional map theo boro**
* Các điểm đón khách



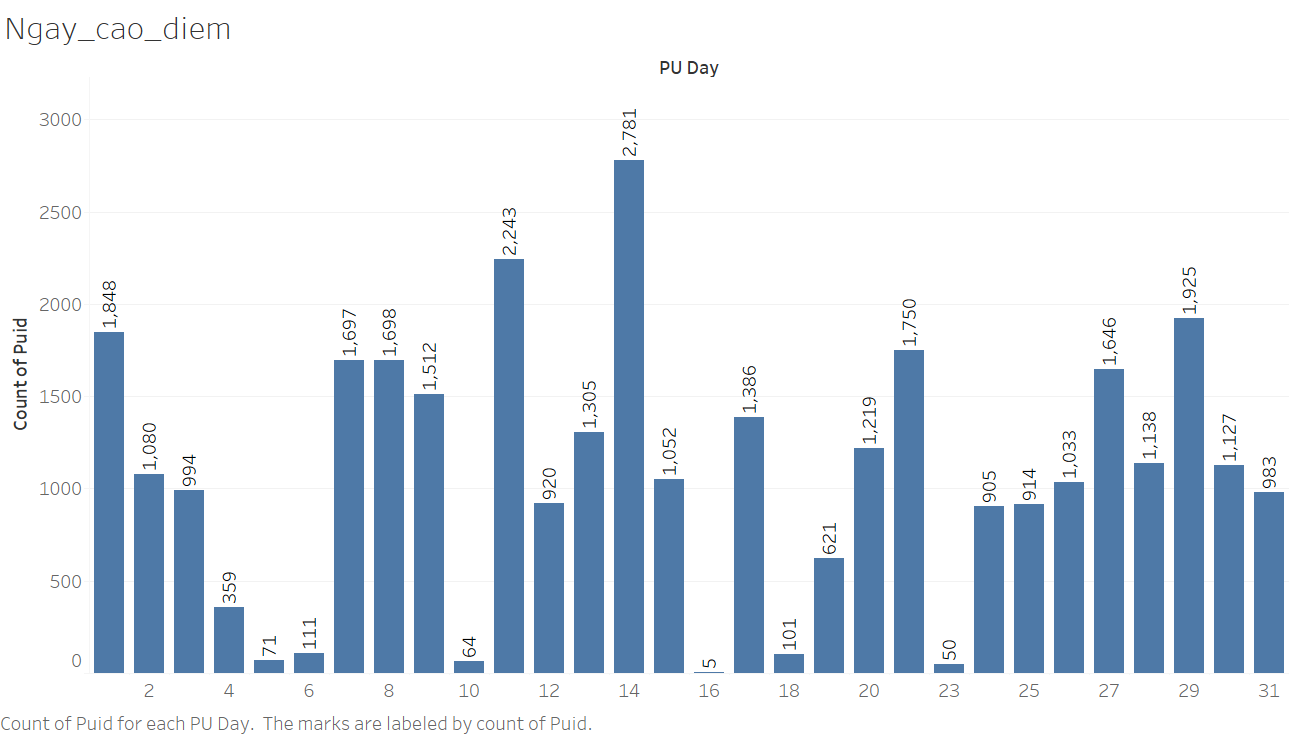
* Các điểm trả khách



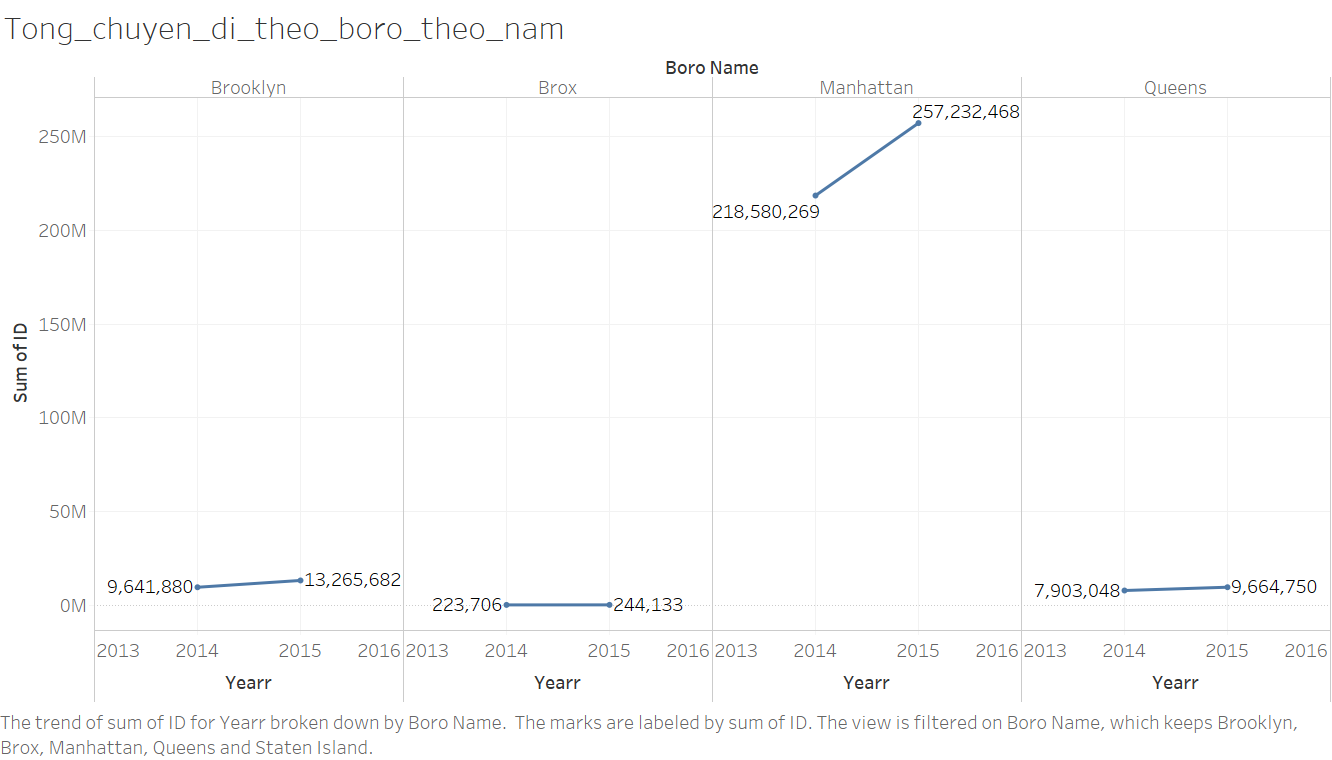
* 1. **Phân tích giờ, ngày cao điểm đón khách**
* Giờ cao điểm



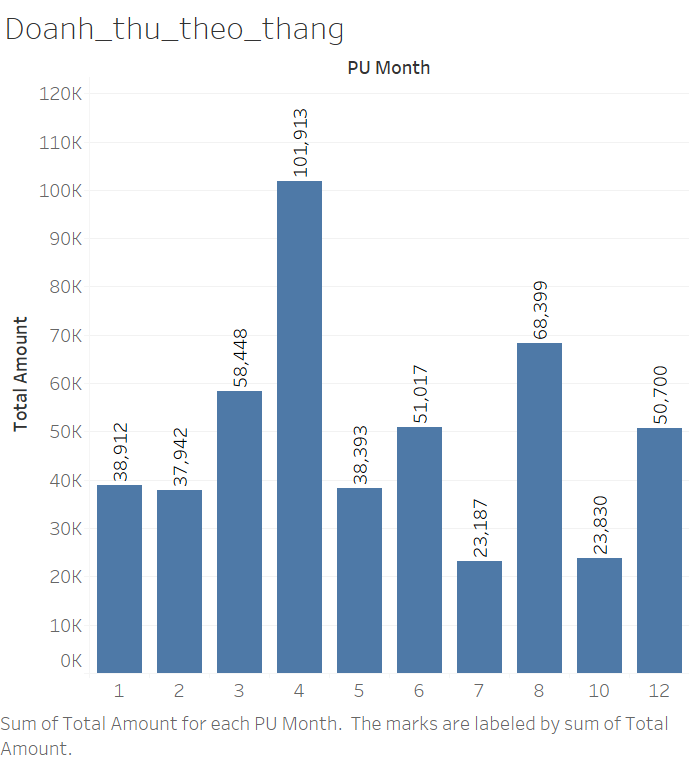
* Ngày cao điểm



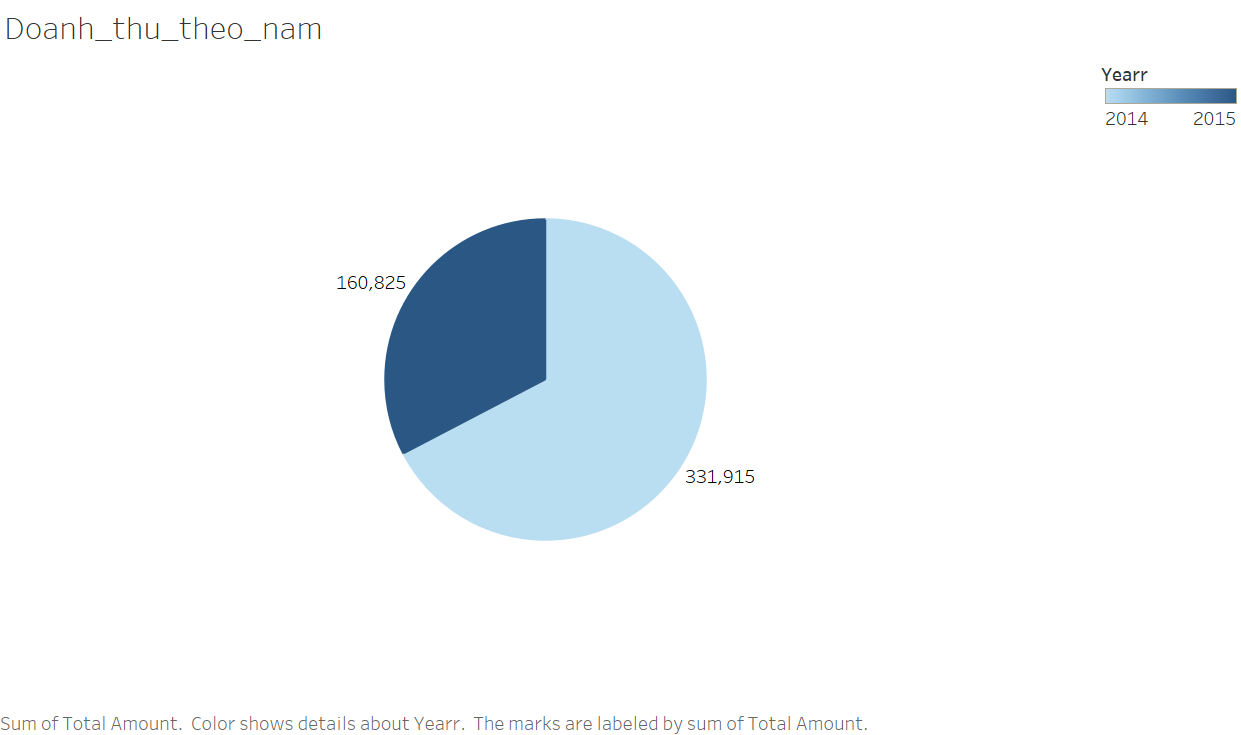
* 1. **Phân tích số lượng đón xe theo boro theo năm**



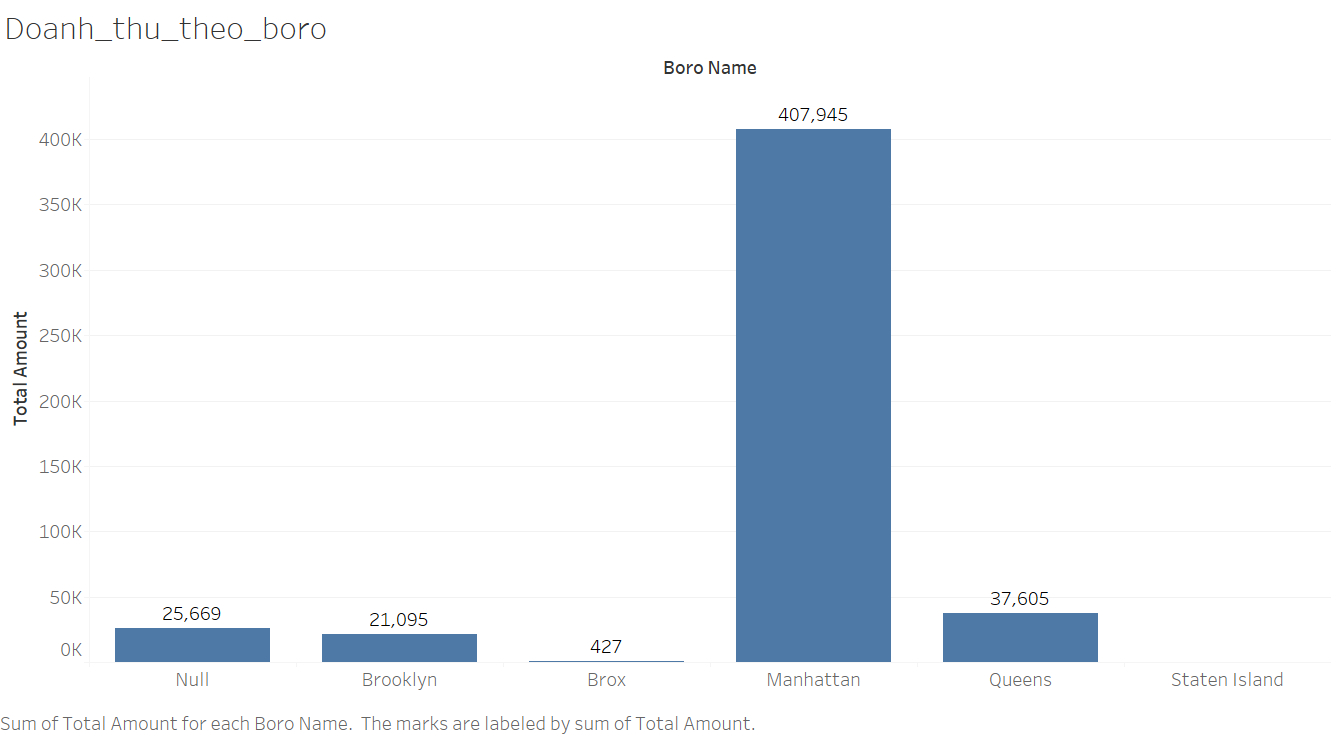
* 1. **Phân tích tổng doanh thu theo tháng, năm**
* Theo tháng



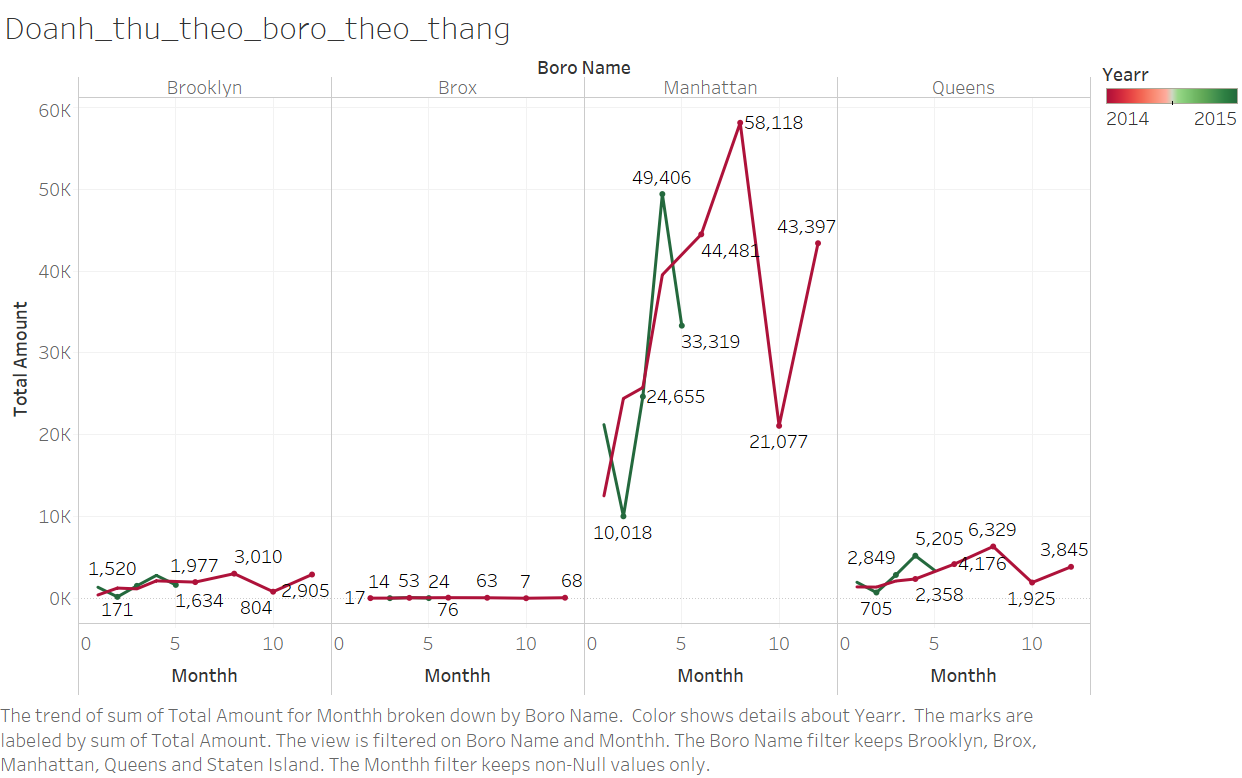
* Theo năm



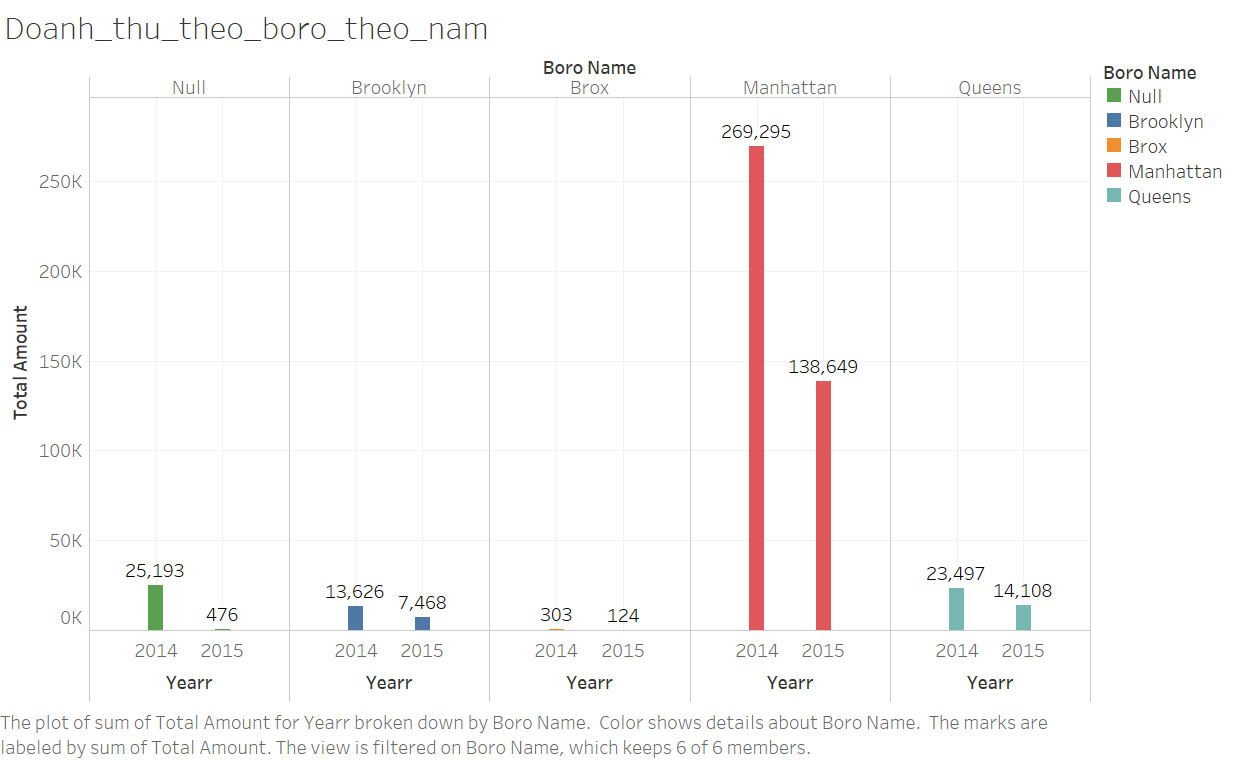
* 1. **Phân tích tổng doanh thu theo boro**



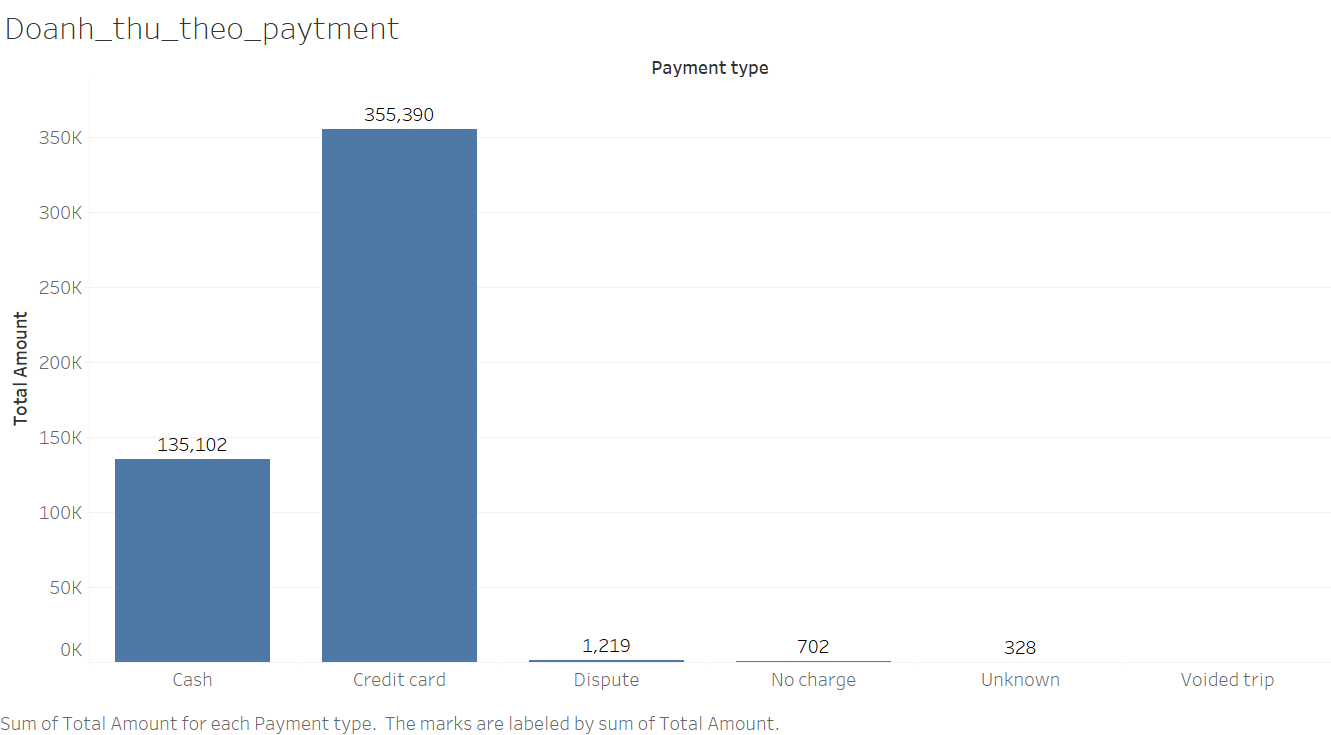
* Theo tháng



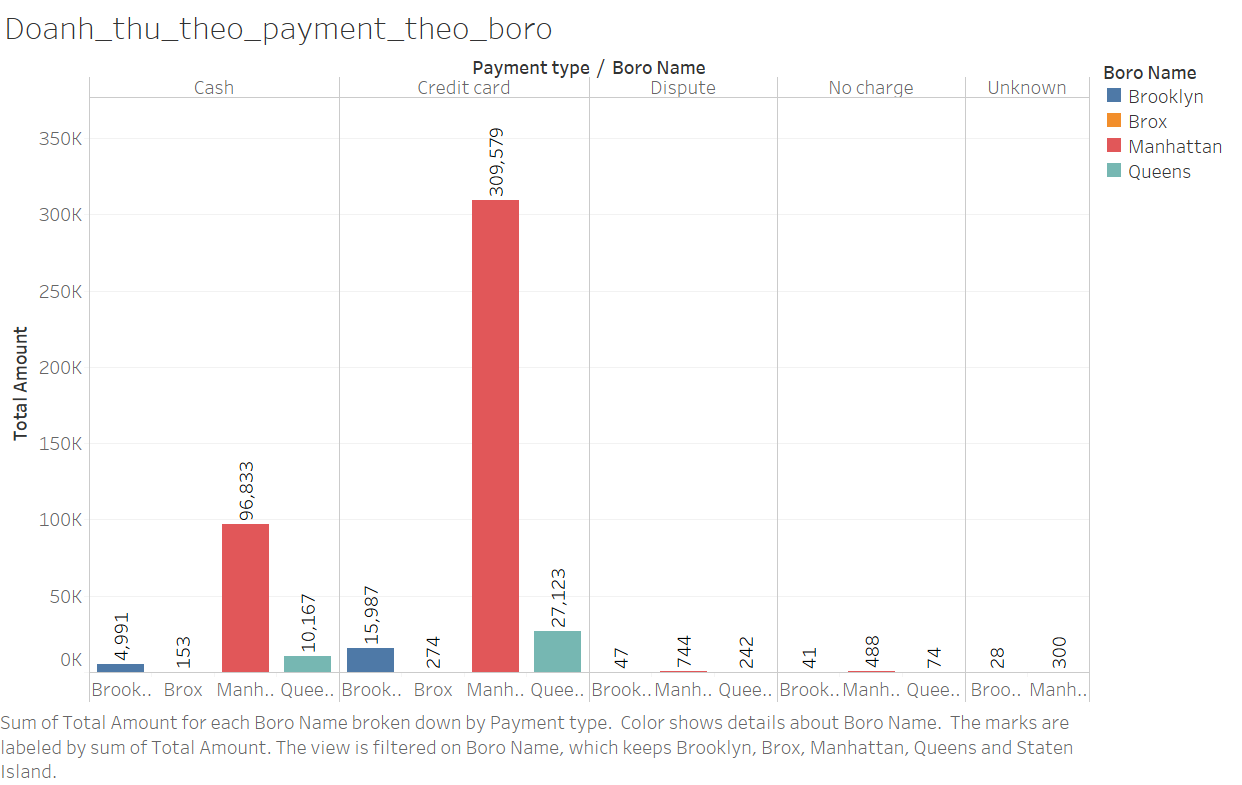
* Theo năm



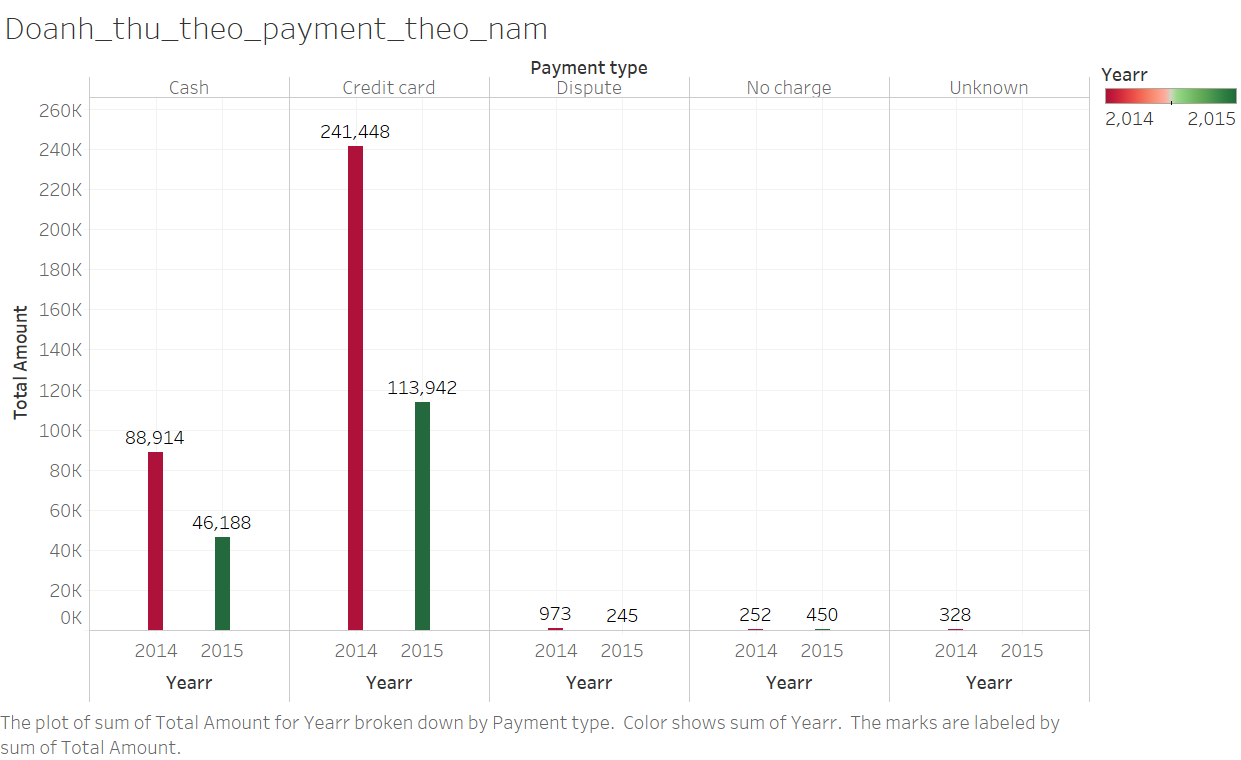
* 1. **Phân tích tổng doanh thu các loại thanh toán**



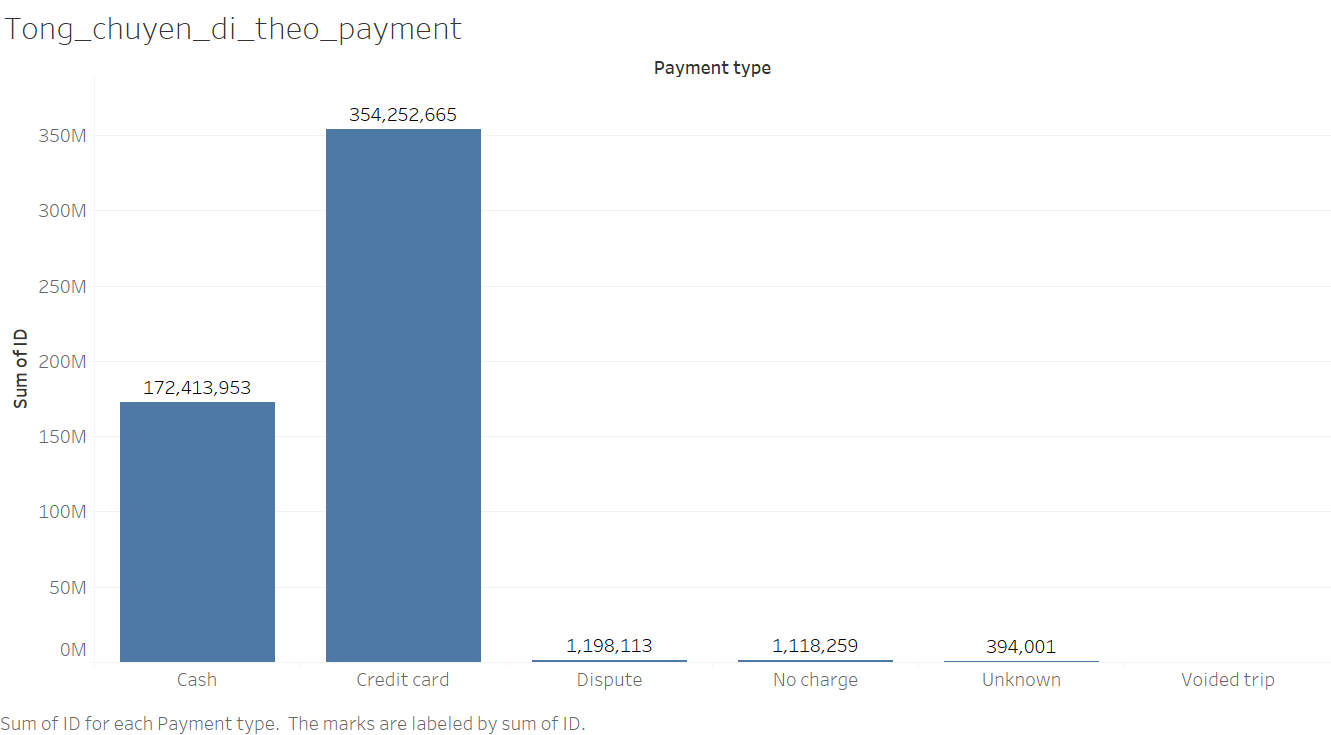
* Theo boro



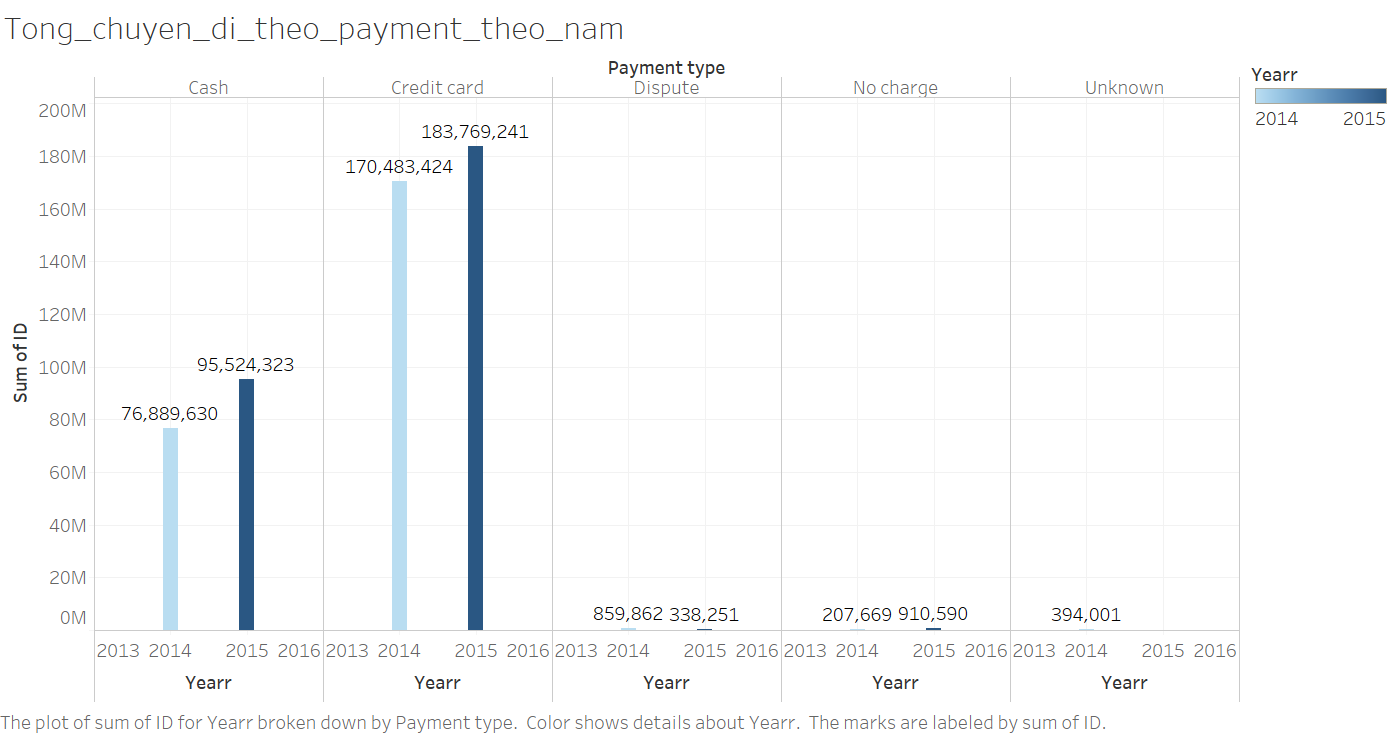
* Theo năm



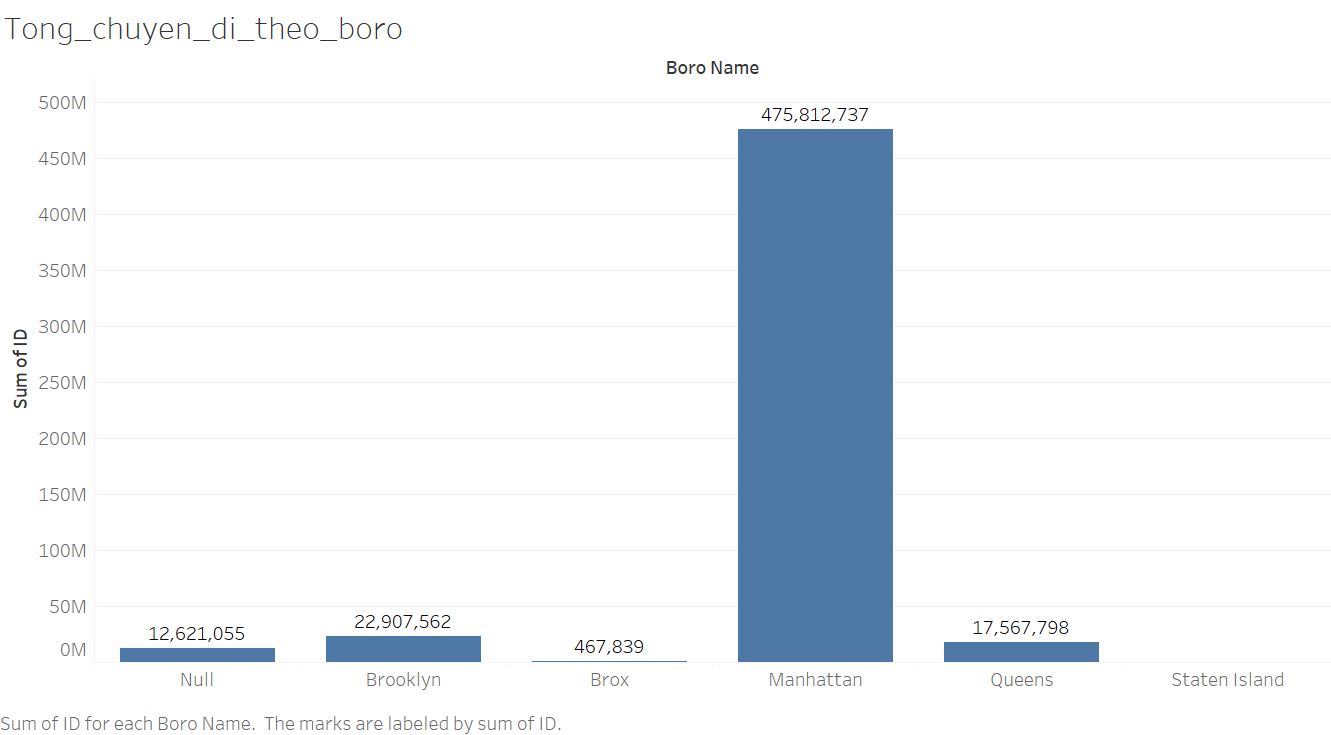
* 1. **Phân tích tổng chuyến đi**
* Theo loại thanh toán



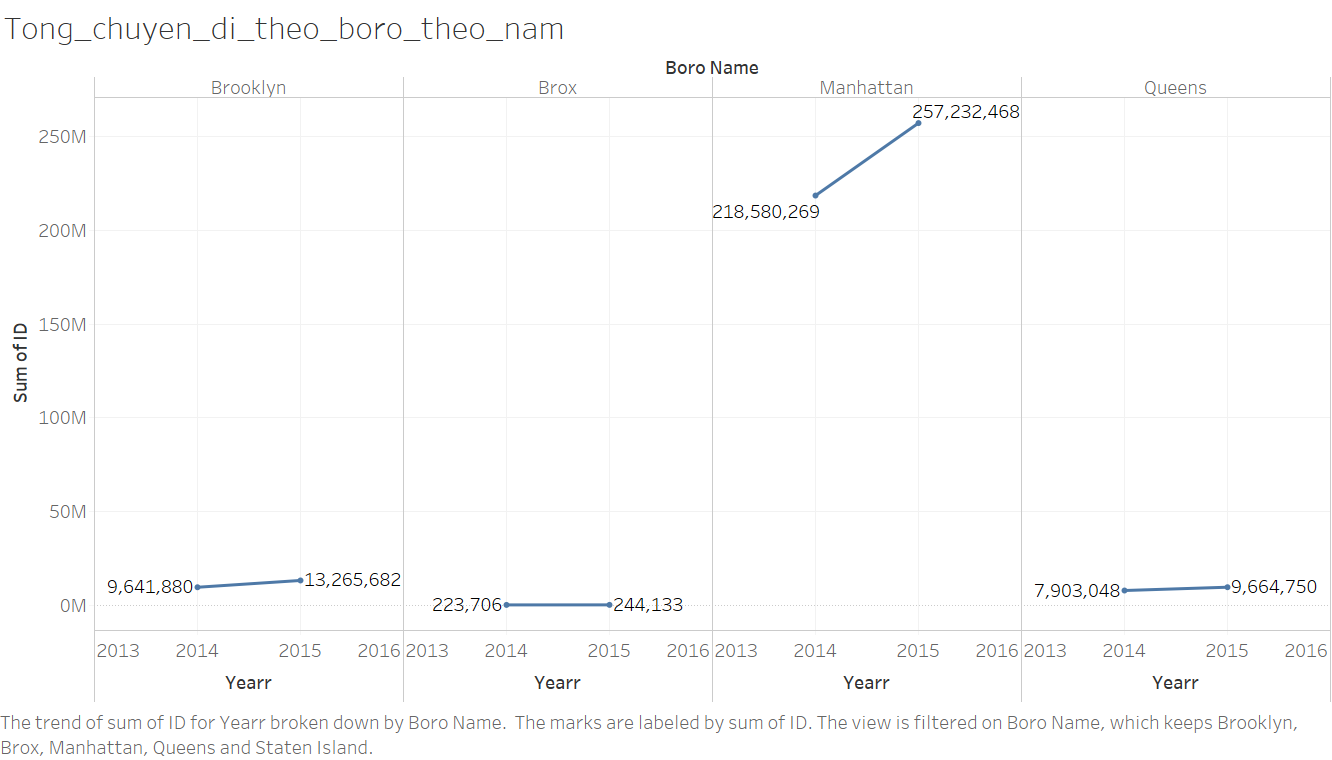
* Theo năm



* Theo boro



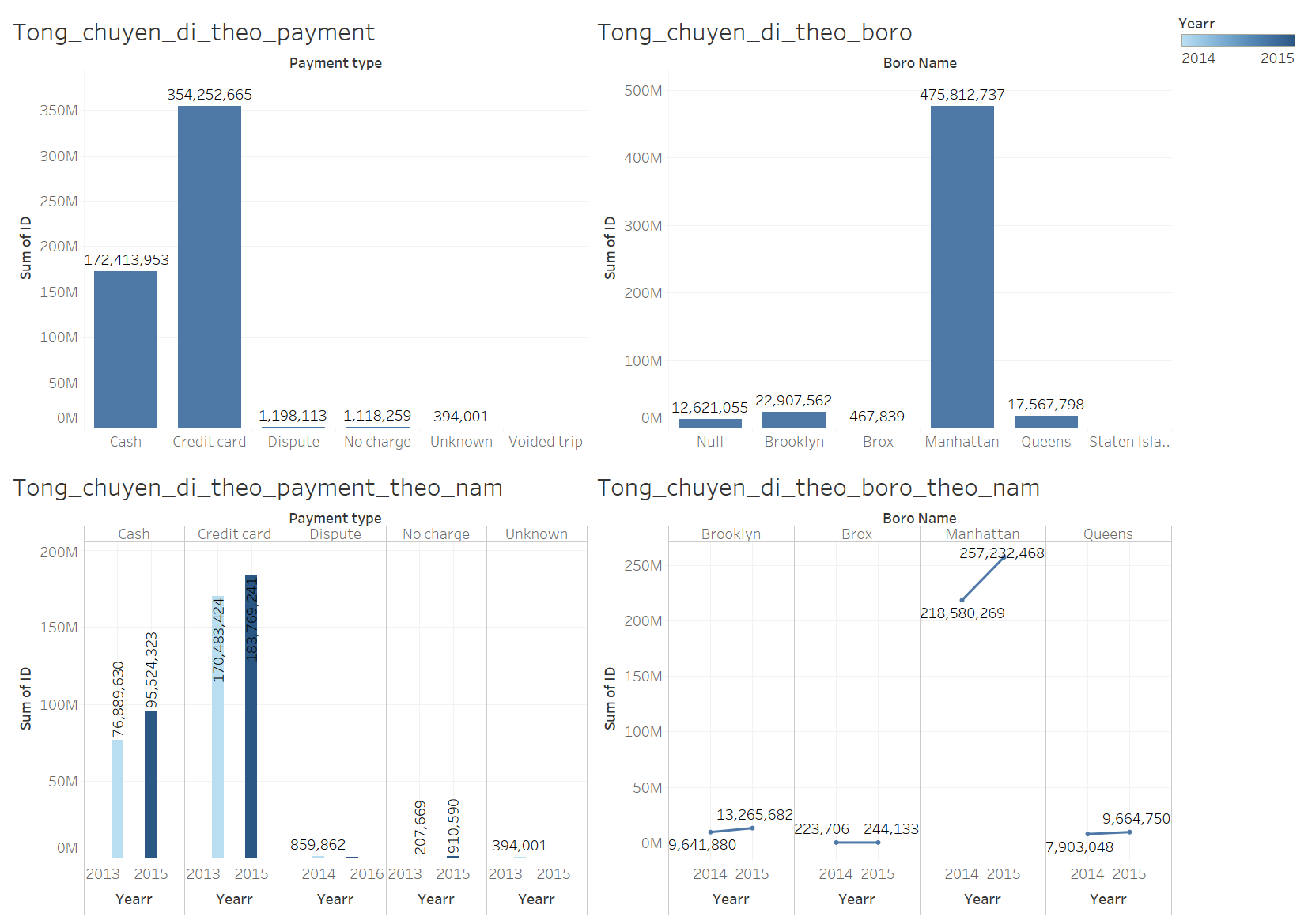
* Theo năm



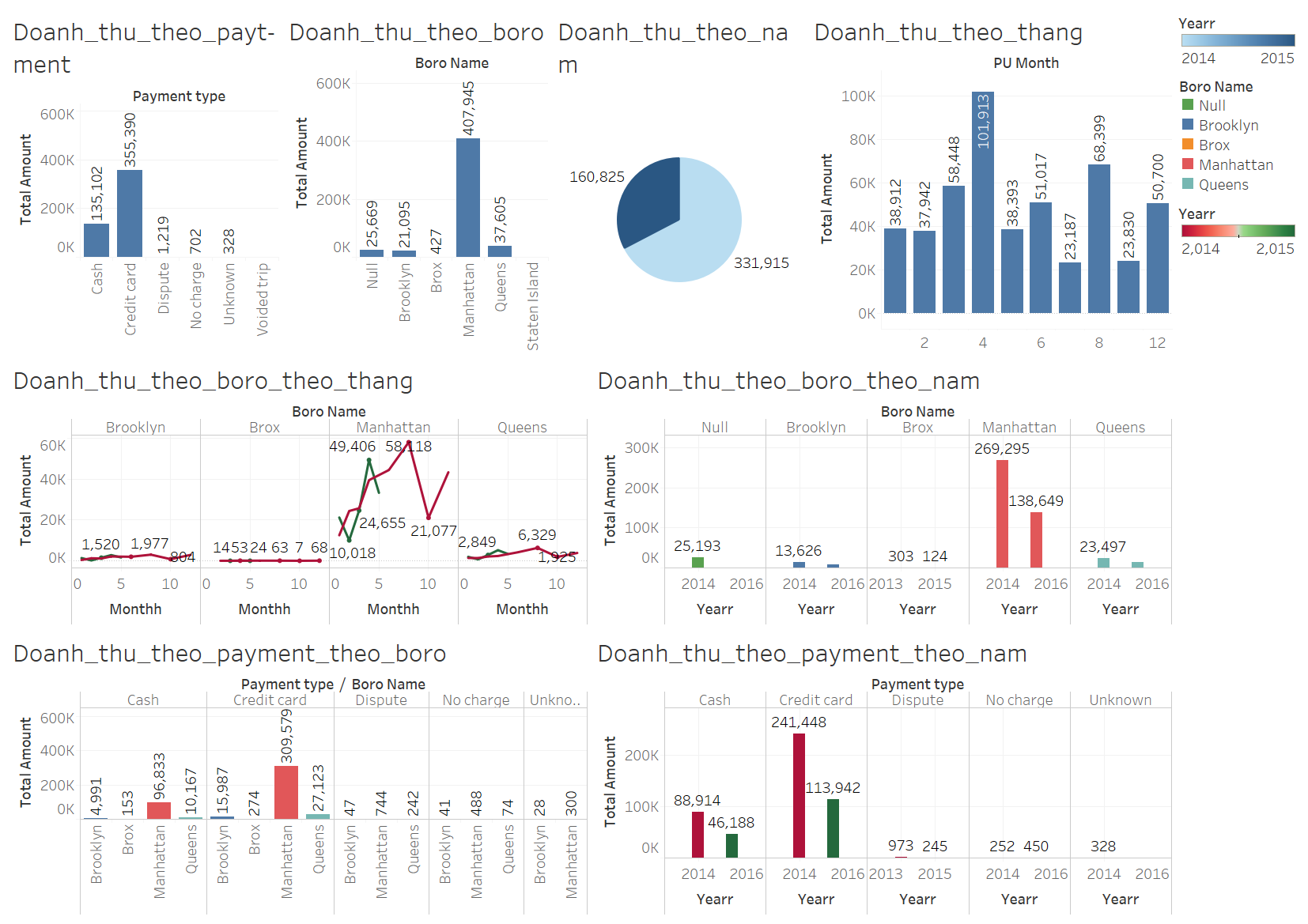
* 1. **Dashborad report**
* Chuyến đi theo boro



* Phân tích tổng chuyến đi



* Phân tích doanh thu

**---HẾT---**