**Đại học Quốc Gia Hà Nội**

**Đại học Công Nghệ**



**BÁO CÁO MÔN KHO DỮ LIỆU**

**Chủ đề: Viễn thông**

**(Telecommunications)**

**Giảng viên hướng dẫn: Nguyễn Hà Nam**

Thành viên nhóm:

1. Lô Quốc Khánh
2. Bàn Văn Mạnh
3. Đào Đức Hòa
4. Hoàng Tiến Mạnh

**Công cụ hỗ trợ:** SQL server 2012 developer edition, SQL server manager studio, Visual studio, Excel, Power BI

**Bản báo cáo gồm:** 1 bản dịch, 1 bản word, 1 bản powerpoint.

MỤC LỤC

[Những công việc cần làm 4](#_Toc514097652)

[I. Tổng quan về Kho dữ liệu và Viễn thông 5](#_Toc514097653)

[1. Kho dữ liệu 5](#_Toc514097654)

[a. Khái niệm 5](#_Toc514097655)

[b. Mục đích của kho dữ liệu 5](#_Toc514097656)

[c. Đặc tính của kho dữ liệu 6](#_Toc514097657)

[d. Ngôn ngữ cho kho dữ liệu 6](#_Toc514097658)

[2. Viễn thông (Telecommunications) 7](#_Toc514097659)

[II. Xây dựng hệ thống quản lí và Phân tích dữ liệu Viễn Thông 7](#_Toc514097660)

[1. Yêu cầu: 7](#_Toc514097661)

[a, bài toán 7](#_Toc514097662)

[b, một số câu hỏi để làm rõ yêu cầu bài toán 7](#_Toc514097663)

[2. Thiết kế cơ sở dữ liệu 8](#_Toc514097664)

[a, Custommer: Thể hiện thông tin khách hàng 8](#_Toc514097665)

[b, Billfact: Hóa đơn khách hàng 8](#_Toc514097666)

[c, SalesOrg: Kênh bán hàng 9](#_Toc514097667)

[d, Time: Thời gian 9](#_Toc514097668)

[e, Rate plan: Gói cước 10](#_Toc514097669)

[f, Service Line: Đường dây dịch vụ 10](#_Toc514097670)

[3. Dữ Liệu 11](#_Toc514097671)

[a, Customer 11](#_Toc514097672)

[b, BillFact 12](#_Toc514097673)

[c, Time 13](#_Toc514097674)

[d, SalesOrg 14](#_Toc514097675)

[e, RatePlant 14](#_Toc514097676)

[f, serviceLine 15](#_Toc514097677)

[4. Lược đồ hình sao 16](#_Toc514097678)

[5. Các bước tiến hành 16](#_Toc514097679)

[a. Import dữ liệu: 16](#_Toc514097680)

[b. Phân tích dữ liệu 18](#_Toc514097681)

[c. Sử dụng công cụ power BI để vẽ biểu đồ dữ liệu, nhìn dữ liệu một cách trực quan hơn. 30](#_Toc514097682)

Danh mục hình vẽ:

*Hình 1: mô hình liên kết – lược đồ sao của hệ thống*

*Hình 2. Đổ dữ liệu từ file excel chuẩn bị sẵn vào database*

*Hình 3. Database chứa data được import*

*Hình 4. Tạo nguồn dữ liệu*

*Hình 5. Tạo view cho Data source*

*Hình 6. Ta được view sau khi tạo xong*

*Hình 7. chọn bảng Billfact làm measure group*

*Hình 8. Chọn các thuộc tính từ bảng measure group*

*Hình 9. Chọn các demansion*

*Hình 10. Thông báo hoàn thành cube*

*Hình 11. Process và deploy thành công*

*Hình 12. Tổng số tin nhắn và cuộc gọi theo Quốc gia trong 3 năm*

*Hình 13. . Lượng data sử dụng theo quốc gia trong 3 năm*

*Hình 14. Tin nhắn, data sử dụng, số cuộc gọi theo nhà mạng năm 2013*

*Hình 15: Tin nhắn, data sử dụng, số cuộc gọi theo nhà mạng năm 2014*

*Hình 16.1 Top 5 nước sử dụng data nhiều nhất năm 2013*

*Hình 16.2 Top 5 nước sử dụng data nhiều nhất năm 2014*

*Hình 16.3 Top 5 nước sử dụng data nhiều nhất năm 2015*

*Hình 17.1 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2013*

*Hình 17.2 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2014*

*Hình 17.3 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2015*

*Hình 18. Biểu đồ: lượng data sử dụng của các quốc gia trong 3 năm từ 2013 -2015*

*Hình 19. Biểu đồ: số cuộc gọi sử dụng của các quốc gia trong 3 năm từ 2013 -2015*

*Hình 20: phí data của các nhà mạng từ 2013 -2015*

*Hình 21: Biểu đồ thể hiện sự phân bố của các nhà mạng theo quốc gia*

# Những công việc cần làm

* Cài đặt một hệ quản trị cơ sở dữ liệu tùy chọn (SQL Server)
* Đọc tài liệu, xác định ứng dụng cần làm
* Xây dựng cơ sở dữ liệu tương ứng với ứng dụng đó
* Nhập dữ liệu mẫu tương ứng
* Convert dữ liệu từ cơ sở dữ liệu sang kho dữ liệu
* Xây dựng một số câu hỏi và dùng kho dữ liệu để trả lời những câu hỏi ấy

# Tổng quan về Kho dữ liệu và Viễn thông

## Kho dữ liệu

### Khái niệm

Kho dữ liệu (Data Warehouse – DW) là tuyển tập các cơ sở dữ liệu  tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định.

Theo John Ladley, Công nghệ kho dữ liệu (Data Warehouse Technology) là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau.

Kho dữ liệu thường rất lớn tới hàng trăm GB hay thậm chí hàng Terabyte. Kho dữ liệu được xây dựng để tiện lợi cho việc truy cập theo nhiều nguồn, nhiều kiểu dữ liệu khác nhau sao cho có thể kết hợp được cả những ứng dụng của các công nghệ hiện đại và kế thừa được từ những hệ thống đã có sẵn từ trước.

### Mục đích của kho dữ liệu

Mục tiêu chính của kho dữ liệu là nhằm đáp ứng các tiêu chuẩn cơ bản sau:

* Phải có khả năng đáp ứng mọi yêu cầu về thông tin của người sử dụng.
* Hỗ trợ để các nhân viên của tổ chức thực hiện tốt, hiệu quả công việc của mình, như có những quyết định hợp lý, nhanh và bán được nhiều hàng hơn, năng suất cao hơn, thu được lợi nhuận cao hơn, v.v.
* Giúp cho tổ chức, xác định, quản lý và điều hành các dự án, các nghiệp vụ một cách hiệu quả và chính xác.
* Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau

Muốn đạt được những yêu cầu trên thì DW phải:

* Nâng cao chất lượng dữ liệu bằng các phương pháp làm sạch và tinh lọc dữ liệu theo những hướng chủ đề nhất định
* Tổng hợp và kết nối dữ liệu
* Đồng bộ hoá các nguồn dữ liệu với DW
* Phân định và đồng nhất các hệ quản trị cơ sở dữ liệu tác nghiệp như là các công cụ chuẩn để phục vụ cho DW.
* Quản lí siêu dữ liệu
* Cung cấp thông tin được tích hợp, tóm tắt hoặc được liên kết, tổ chức theo các chủ đề
* Dùng  trong các hệ thống hỗ trợ quyết định (Decision suport system - DSS), các hệ thống thông tin tác nghiệp hoặc hỗ trợ cho các truy vấn đặc biệt.

### Đặc tính của kho dữ liệu

Những đặc điểm cơ bản của Kho dữ liệu (DW) là một tập hợp dữ liệu có tính chất sau:

* ***Tính tích hợp (Integration)***
* ***Dữ liệu gắn thời gian và có tính lịch sử***
* ***Dữ liệu có tính ổn định (nonvolatility) và không biến động***
* ***Dữ liệu tổng hợp***

### Ngôn ngữ cho kho dữ liệu

Ngôn ngữ [xử lý phân tích trực tuyến](https://vi.wikipedia.org/wiki/X%E1%BB%AD_l%C3%BD_ph%C3%A2n_t%C3%ADch_tr%E1%BB%B1c_tuy%E1%BA%BFn) (OLAP - [On-Line Analytical Prosessing](https://vi.wikipedia.org/w/index.php?title=On-Line_Analytical_Prosessing&action=edit&redlink=1)), rất phù hợp với kho dữ liệu, ngôn ngữ này tương tự với [ngôn ngữ truy vấn](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_truy_v%E1%BA%A5n) SQL và tập trung vào các câu lệnh sau:

* + Thu nhỏ (roll-up): ví dụ: nhóm dữ liệu theo năm thay vì theo quý.
  + Mở rộng (drill-down): ví dụ: mở rộng dữ liệu, nhìn theo tháng thay vì theo quý.
  + Cắt lát (slice): nhìn theo từng lớp một. Ví dụ: từ danh mục bán hàng của Q1, Q2, Q3, Q4 chỉ xem của Q1.
  + Thu nhỏ (dice): bỏ bớt một phần của dữ liệu (tương ứng thêm điều kiện vào câu lệnh WHERE trong SQL).

## Viễn thông (Telecommunications)

Trong nền kinh tế thị trường hiện nay, thông tin là yếu tố sống còn đối với bất kỳ một doanh nghiệp nào. Việc nắm bắt thông tin giúp cho các doanh nghiệp hoạch định chiến lược kinh doanh cho mình một cách chính xác. Sự ra đời của công nghệ kho dữ liệu (Data Warehouse - DWH) trong những năm gần đây đã đáp ứng nhu cầu quản lý, lưu trữ một khối lượng dữ liệu lớn và có khả năng khai thác dữ liệu đa chiều và theo chiều sâu nhằm hỗ trợ việc ra quyết định của các nhà quản lý.

Ngoài việc thu thập lưu trữ các thông tin từ các hệ thống như Tính cước và chăm sóc khách hàng nói chung và các hệ thống khác nói riêng hệ thống DWH còn cung cấp các thông tin hữu ích giúp cho các nhà phát triển cập nhật các thay đổi hệ thống một cách nhanh chóng. Hơn thế nữa, hệ thống còn cung cấp cho nhà quản lý những thông tin quan trọng chính xác và nhanh chóng giúp họ có tầm nhìn chiến lược và hỗ trợ nhà quản lý ra những quyết định kịp thời và có lợi nhất cho doanh nghiệp trong điều kiện cạnh tranh của ngành viễn thông đang diễn ra rất gay gắt.

# II. Xây dựng hệ thống quản lí và Phân tích dữ liệu Viễn Thông

## Yêu cầu:

### a, bài toán

Xây dựng một kho dữ liệu để tổng hợp và phân tích quá trình sử dụng viễn thông của khác hàng

### b, một số câu hỏi để làm rõ yêu cầu bài toán

* Quốc gia nào là thị trường lớn, quốc gia nào có khả năng phát triển trong những năm tiếp theo?
* Các nhà mạng có cần phải đưa ra các gói khuyễn mại để tăng nhu cầu của khách hàng không?
* Sự cạnh tranh của các nhà mạng tại Trung Quốc ra sao?
* Xu thế sử dụng dịch vụ của khách hàng như thế nào?
* Đưa ra Top 5 Quốc gia sử dụng MB nhiều nhất?
* Đưa ra Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất?

## Thiết kế cơ sở dữ liệu

### a, Custommer: Thể hiện thông tin khách hàng

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Kiểu Dữ liệu** | **Mô tả** |
| Customer\_id | Int | ID khách hàng |
| Customer\_name | String | Tên Khách Hàng |
| Customer\_address | String | Địa chỉ khách hàng |
| Customer\_zip | String | Mã giao dịch khách hàng |
| Customer\_city | String | Thành phố khách hàng sinh sống |
| Orig\_authorization\_credis\_code | String | Mã tín dụng |
| Customer\_State | String | Quốc gia |

### b, Billfact: Hóa đơn khách hàng

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Kiểu dữ liệu** | **Mô tả** |
| ID | String | Mã hóa đơn |
| Bill Date Key | String | Mã thời gian |
| Customer\_ID | String | Mã khách hàng |
| Service Line Key | String | Mã dịch vụ |
| Sales Organization Key | String | Mã Kênh bán hàng |
| Rate Plan code | String | Mã gói cước |
| Bill Number | String | Số hiệu hoá đơn |
| Call Count | Int | Cuộc gọi |
| Total Minute Count | Int | Phút gọi |
| Night-Weekend Minute Count | Int | Phút gọi đêm cuối tuần |
| Roam Minute Count | Int | Số phút gọi ngoại tuyến |
| Message Count | Int | Tin nhắn |
| Data MB Used | Int | Dữ liệu sử dụng |
| Month Service Charge | Float | Phí dịch vụ tháng |
| Message Charge | Float | Phí tin nhắn |
| Data Charge | Float | Phí dữ liệu |
| Roaming Charge | Float | Phí ngoại tuyến |
| Taxes | Float | Thuế |
| Regulatory Charges | Float | Phí điều chỉnh |

### c, SalesOrg: Kênh bán hàng ( Nhà Mạng)

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Kiểu dữ liệu** | **Mô tả** |
| Sales Organization Key | String | Mã tổ chức |
| Sales Org Number | String | Số hiệu kênh bán hàng |
| Sales Organization Name | string | Tên Tổ chức bán hàng  ( Nhà Mạng) |
| Sales Channel ID | String | Mã kênh bán hàng |
| Sales Channel Name | String | Tên kênh bán hàng |

### d, Time: Thời gian

|  |  |  |
| --- | --- | --- |
| **Trường dữ liệu** | **Kiểu dữ liệu** | **Mô tả** |
| Bill Date Key | String | Mã thời gian |
| Year | Int | Năm |
| Month | Int | Tháng |
| Day | int | Ngày |

### e, Rate plan: Gói cước

|  |  |  |
| --- | --- | --- |
| **Trường** | **Kiểu dữ liệu** | **Mô tả** |
| Rate plan code | int | Mã gói cước |
| Rate Plan Name | String | Tên gói cước |
| Rate Plan Abbreviation | String | Tên viết tắt gói cước |
| Rate Plan Type Code | String | Loại mã gói cước |
| Rate Plan Type Description | String | Miêu tả gói cước |
| Plan Minutes Allowed | int | Số phút cho phép |
| Plan Messages Allowed | int | Số tin nhắn cho phép |
| Plan Data MB Allowed | int | Số dữ liệu cho phép |
| Night-Weekend Minute Ind | int | Số phút gọi đêm cuối tuần |

### f, Service Line: Đường dây dịch vụ

|  |  |  |
| --- | --- | --- |
| **Trường** | **Kiểu dữ liệu** | **Mô tả** |
| Service Line Key | int | Mã dịch vụ |
| Service Line Number | int | Số dịch vụ |
| Service Line Area Code | int | Mã vùng dịch vụ |
| Service Line Activation Date | datetime | Ngày kích hoạt dịch vụ |

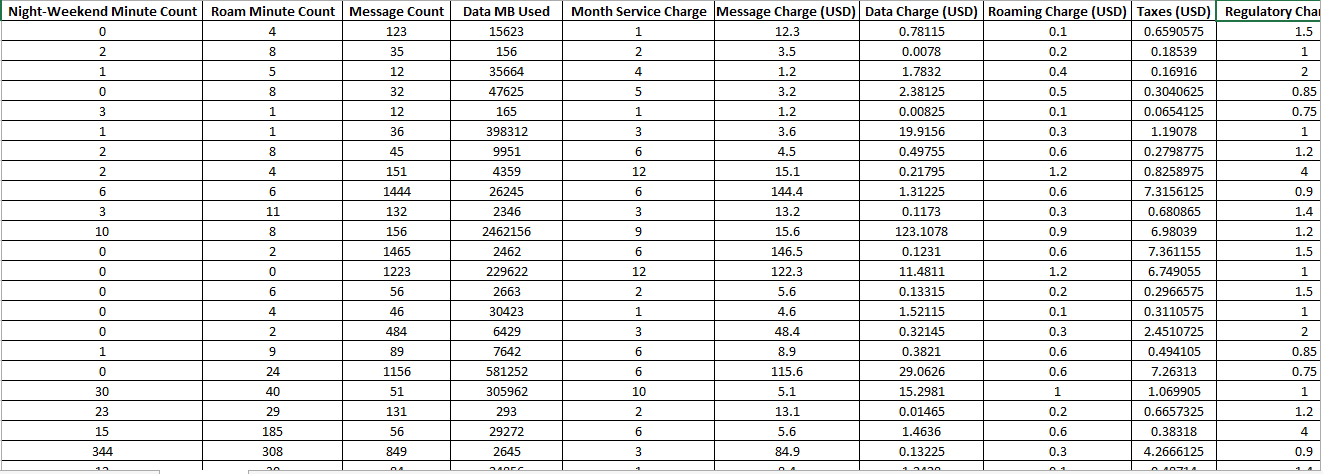
## Dữ Liệu

### a, Customer



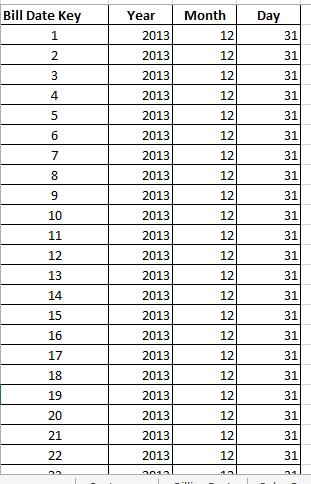
Dữ liệu customer

### b, BillFact



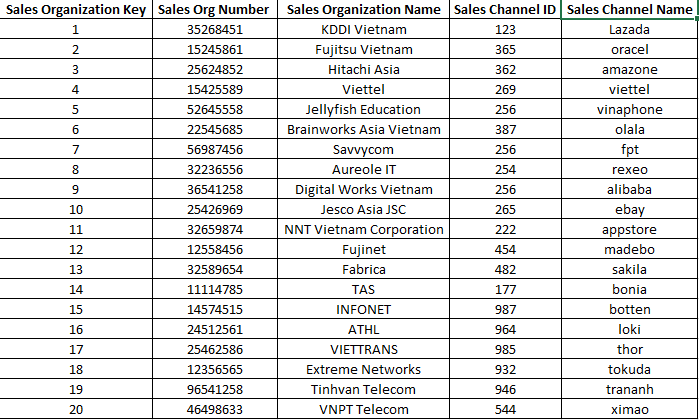
Dữ liệu BillFact

### c, Time



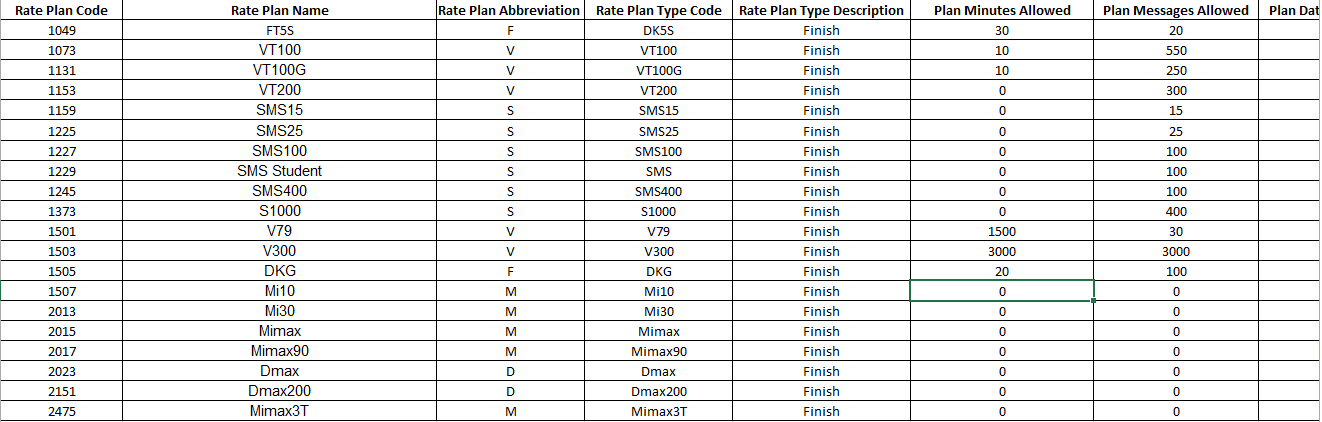
Time

### d, SalesOrg



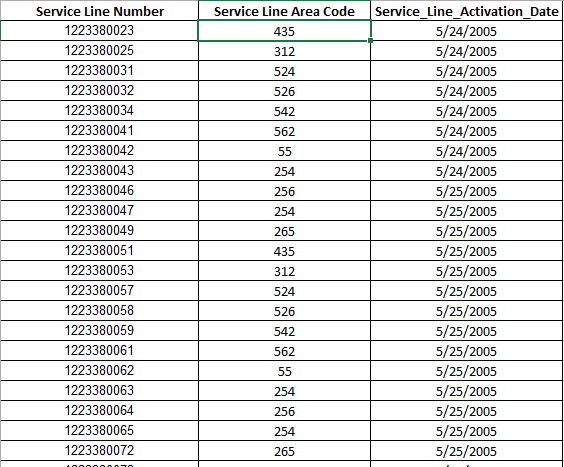
Salesorg

### e, RatePlant



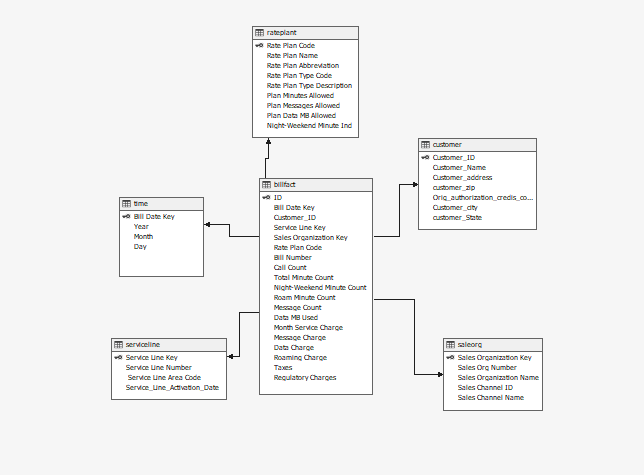
Dữ liệu rateplan

### f, serviceLine



ServicesLine

## Lược đồ hình sao

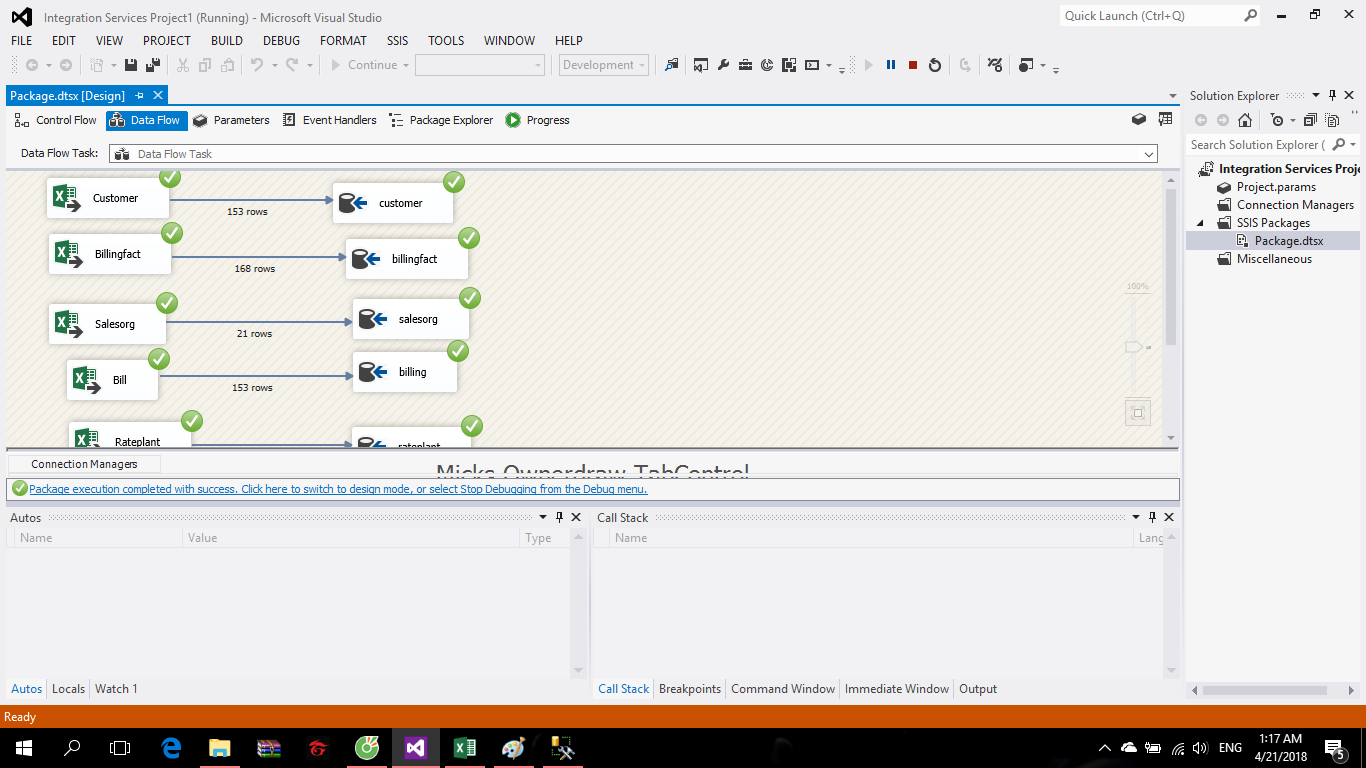


Hình 1: mô hình liên kết – lược đồ sao của hệ thống

## Các bước tiến hành

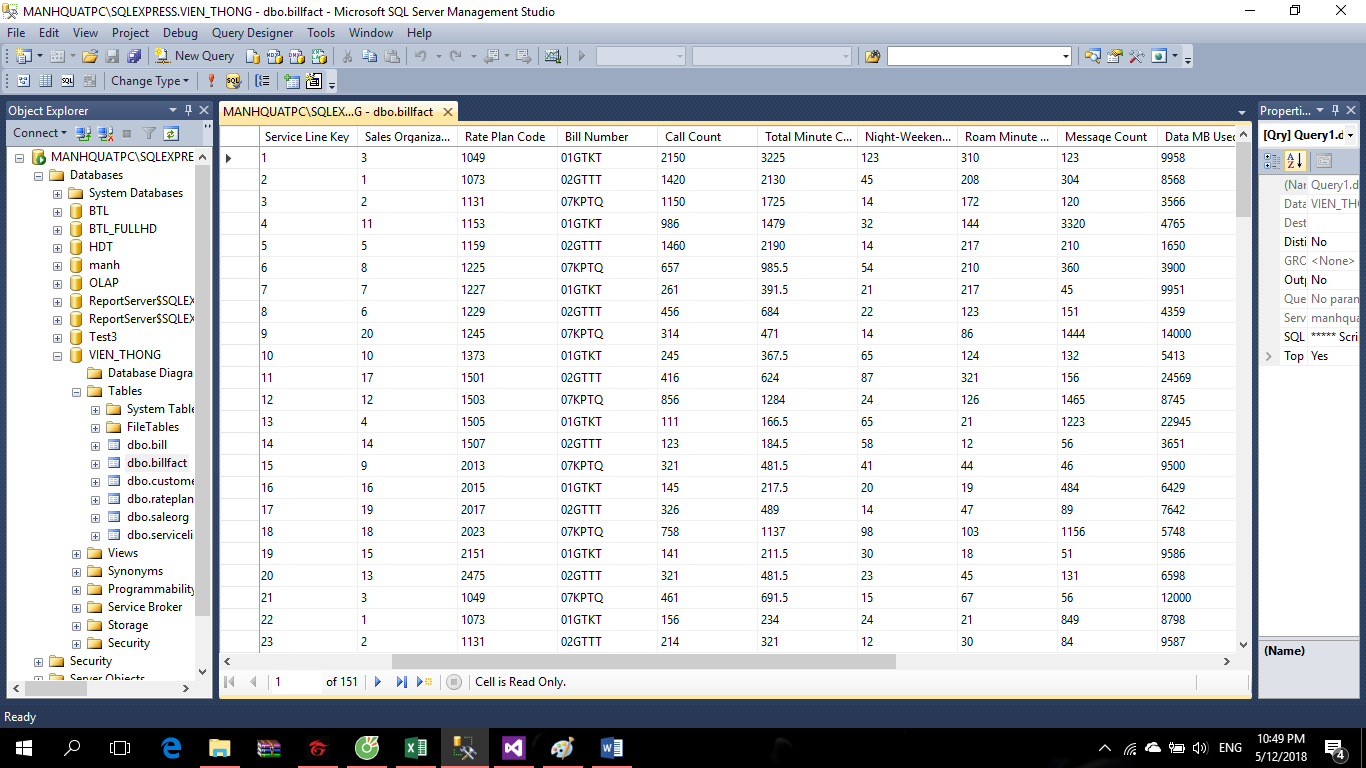
### Import dữ liệu:

Sử dụng SQL Server Integration Service (ssis) để import dữ liệu từ excell vào Database, cụ thể như sau:



Hình 2. Đổ dữ liệu từ file excel chuẩn bị sẵn vào database

Sau khi đổ dữ liệu thành công, database sẽ chứa dữ liệu

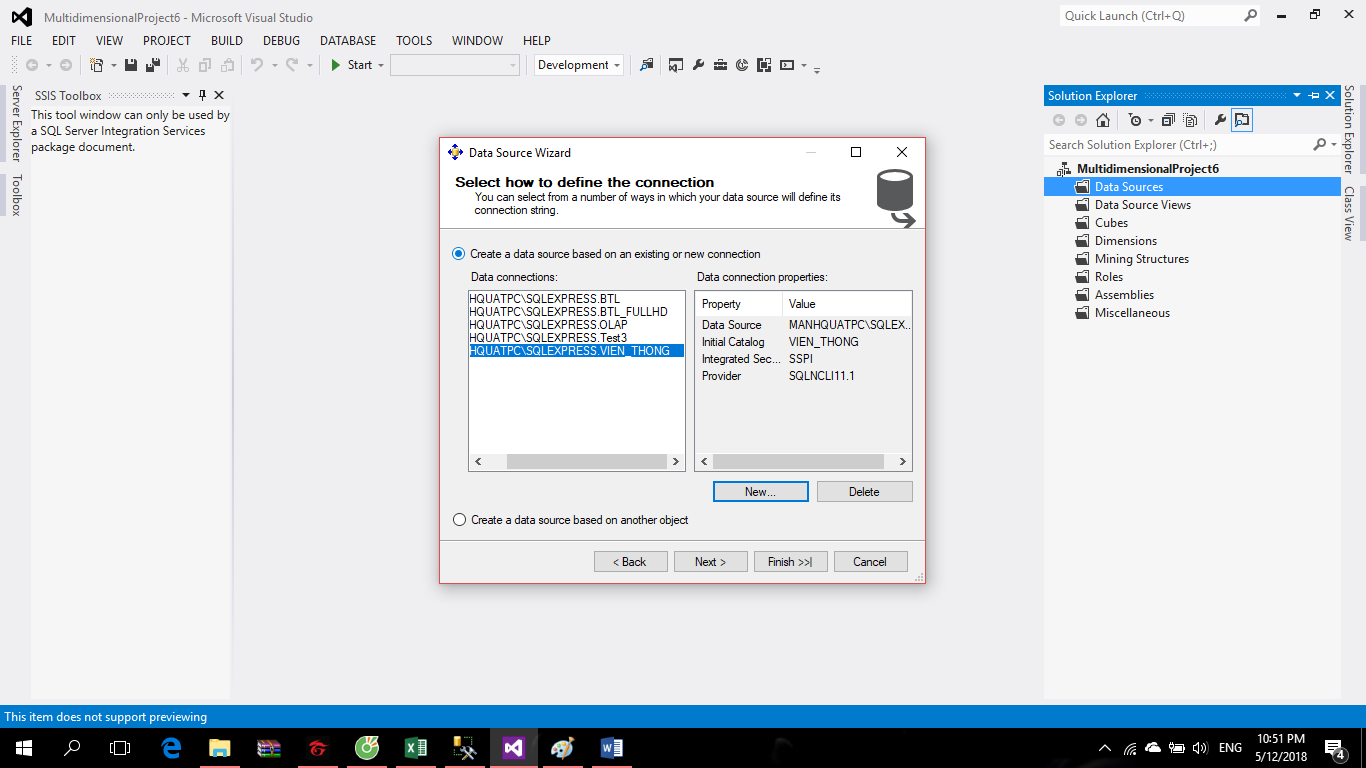


Hình 3. Database chứa data được import

### Phân tích dữ liệu

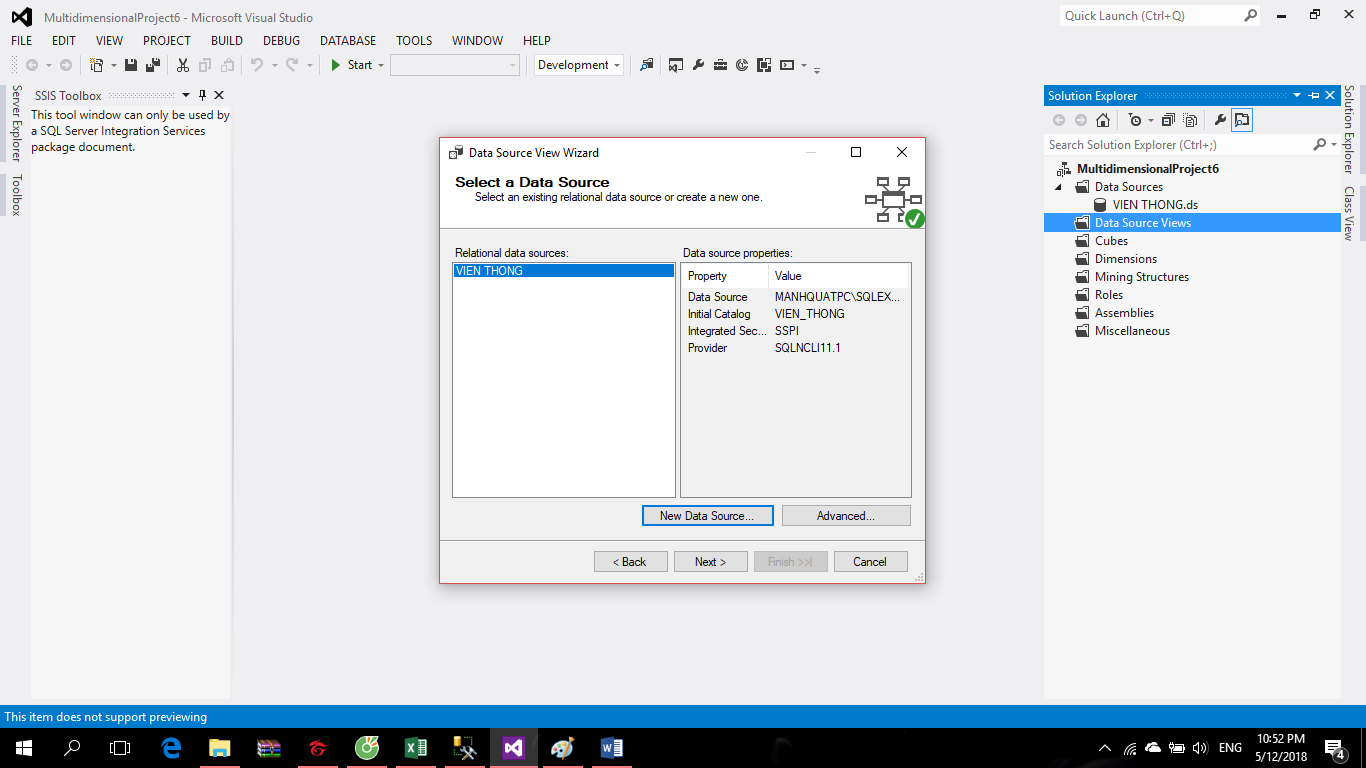
Sử dụng SQL Server Analysis Service ( SSAS ) để phân tích dữ liệu, cụ thể như sau:

* Tạo new project Analysis services Multidementional And Data Mining -> rồi chọn nguồn cho project. Nguồn project chính là database mà ta tạo ở bước 1.



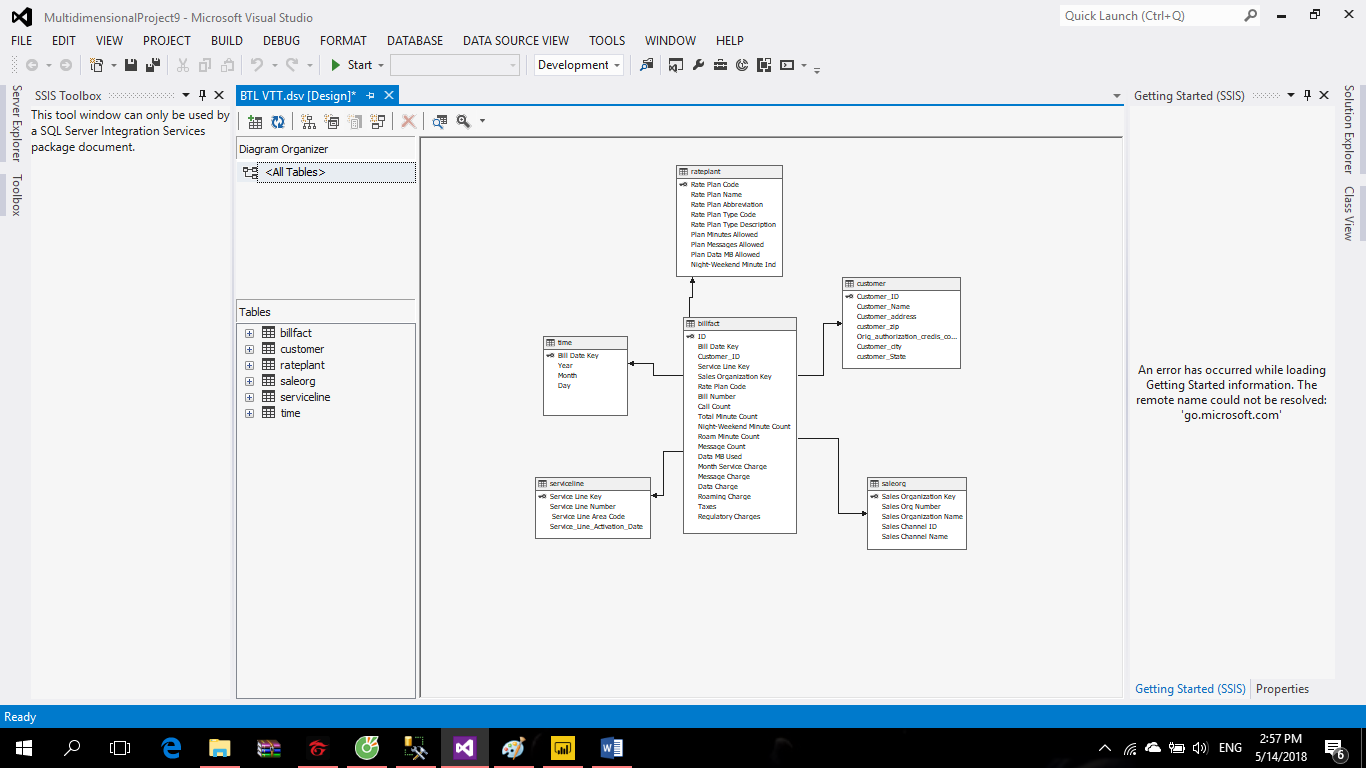
Hình 4 . Tạo nguồn dữ liệu

Tiếp theo, ta tạo view cho data source



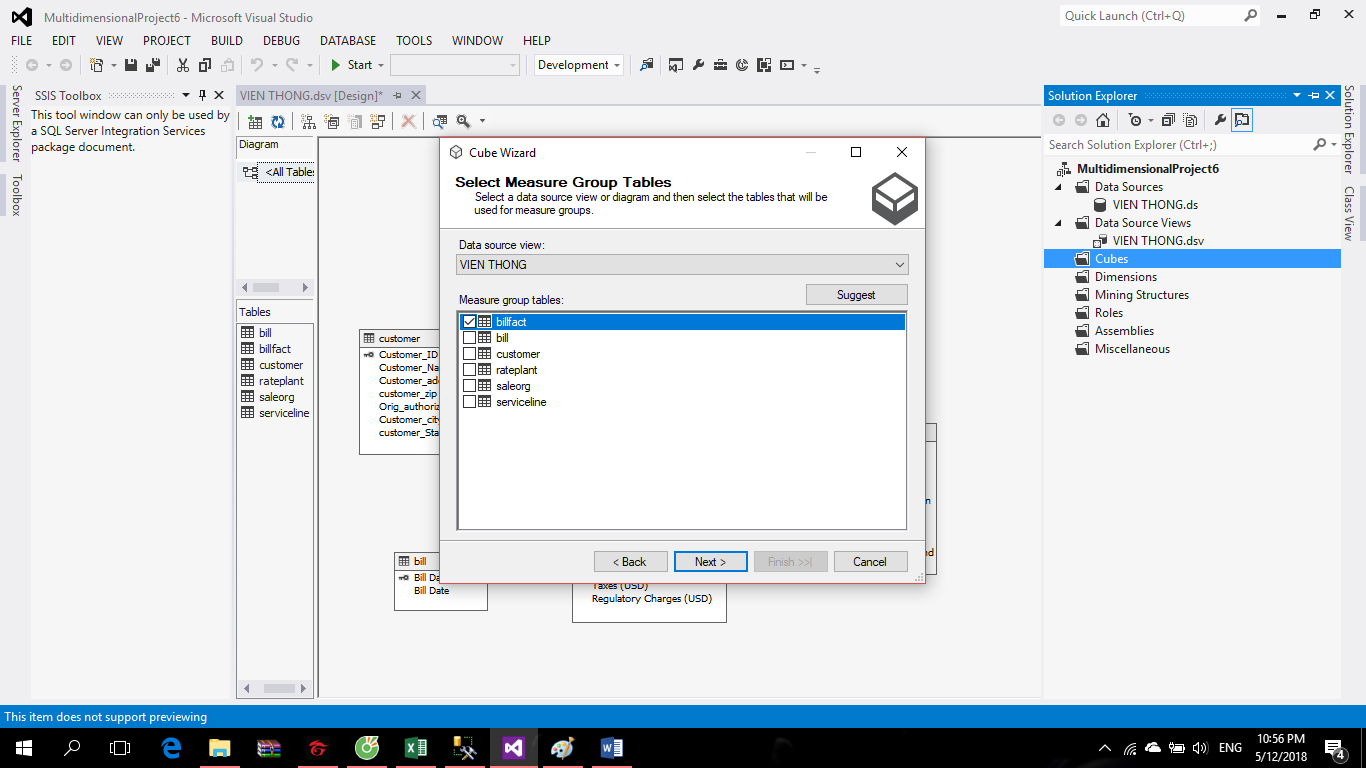
Hình 5. Tạo view cho Data source

và kết quả sau khi tạo xong view:



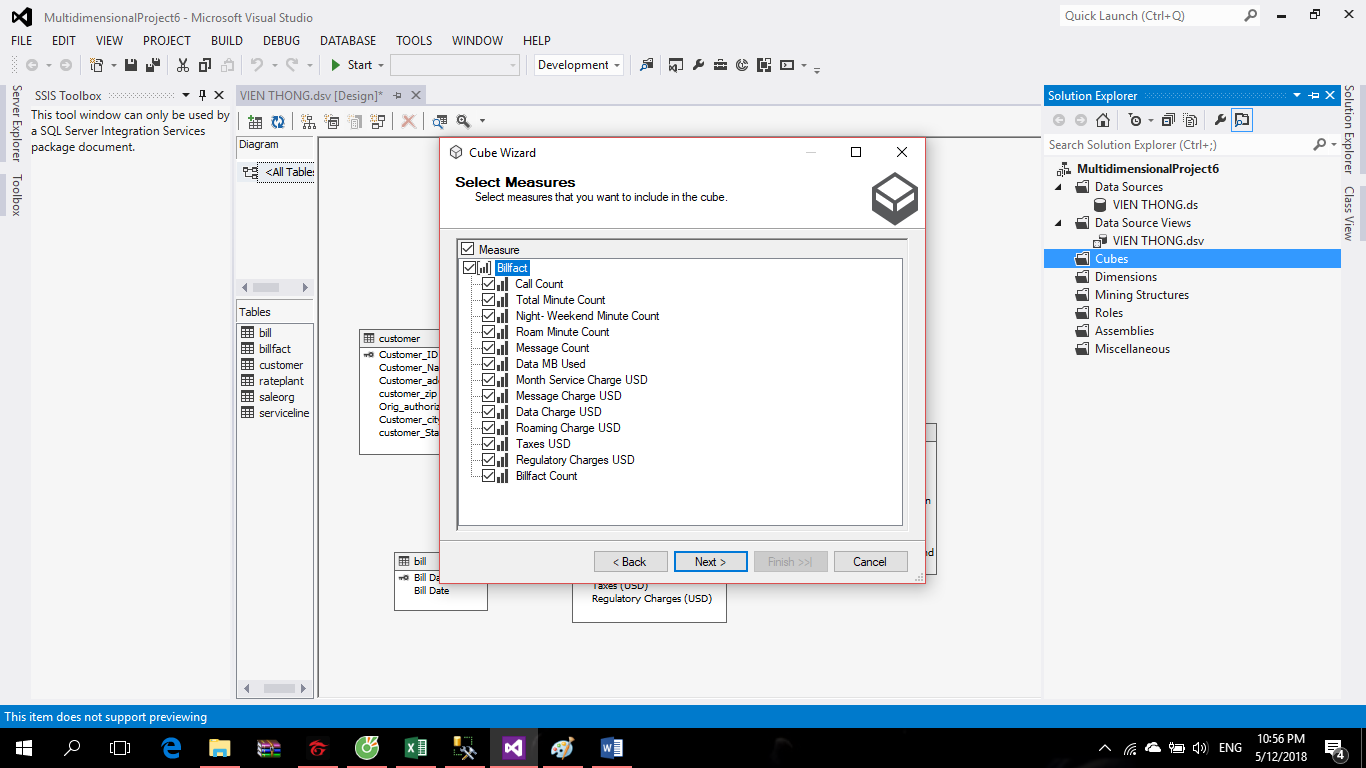
Hình 6. Ta được view sau khi tạo xong

Sau khi tạo xong view, ta bắt đầu tạo cube bằng việc chọn bảng billfact làm measugroup.



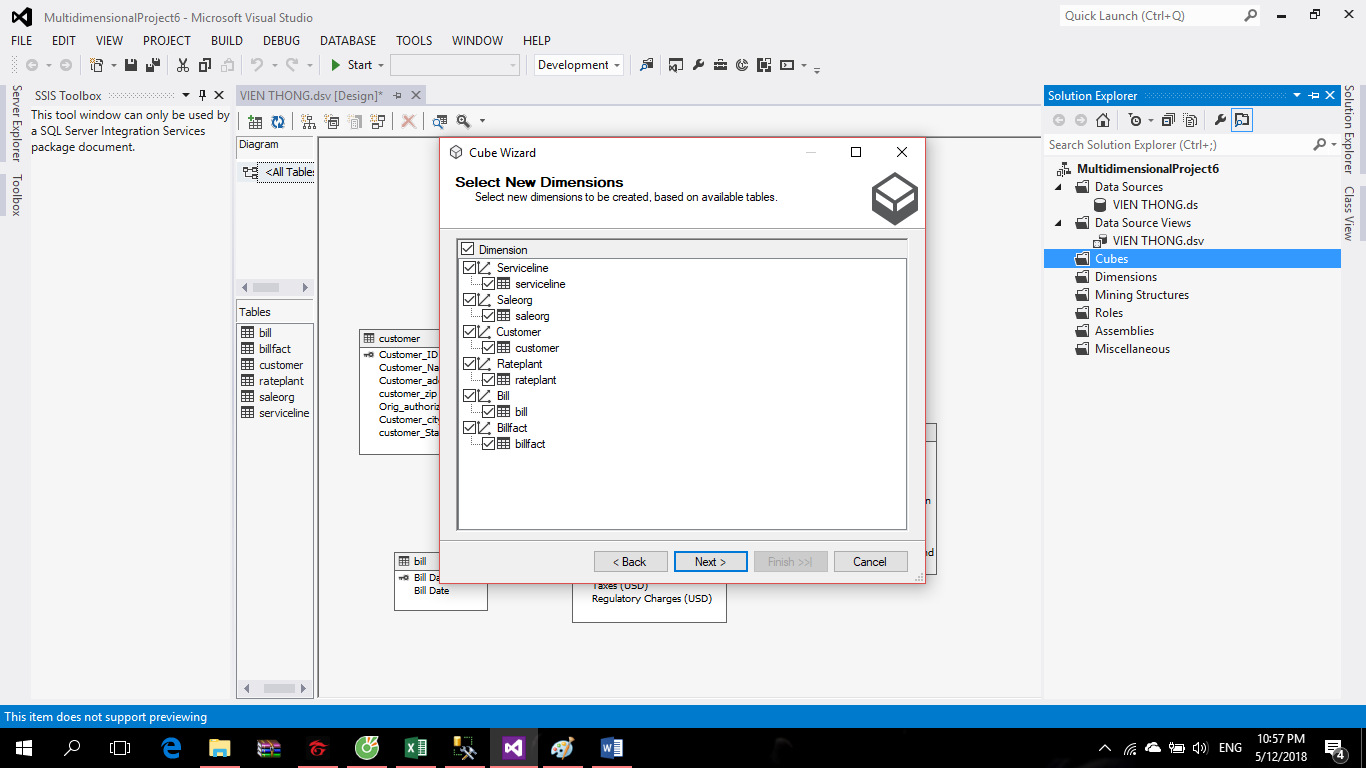
Hình 7. chọn bảng Billfact làm measure group

Chọn các thuộc tính trong bảng measure group:



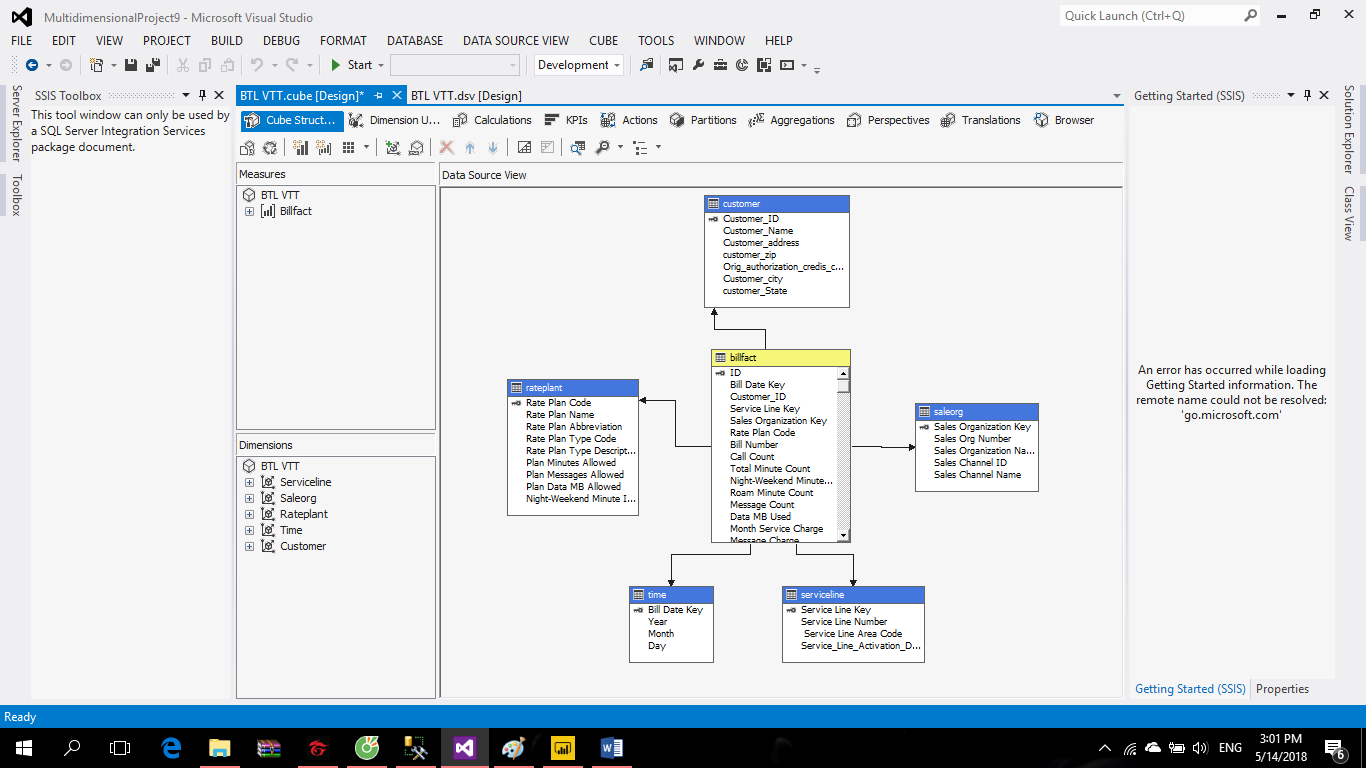
Hình 8. Chọn các thuộc tính từ bảng measure group

Chọn các demention:



Hình 9. Chọn các demansion

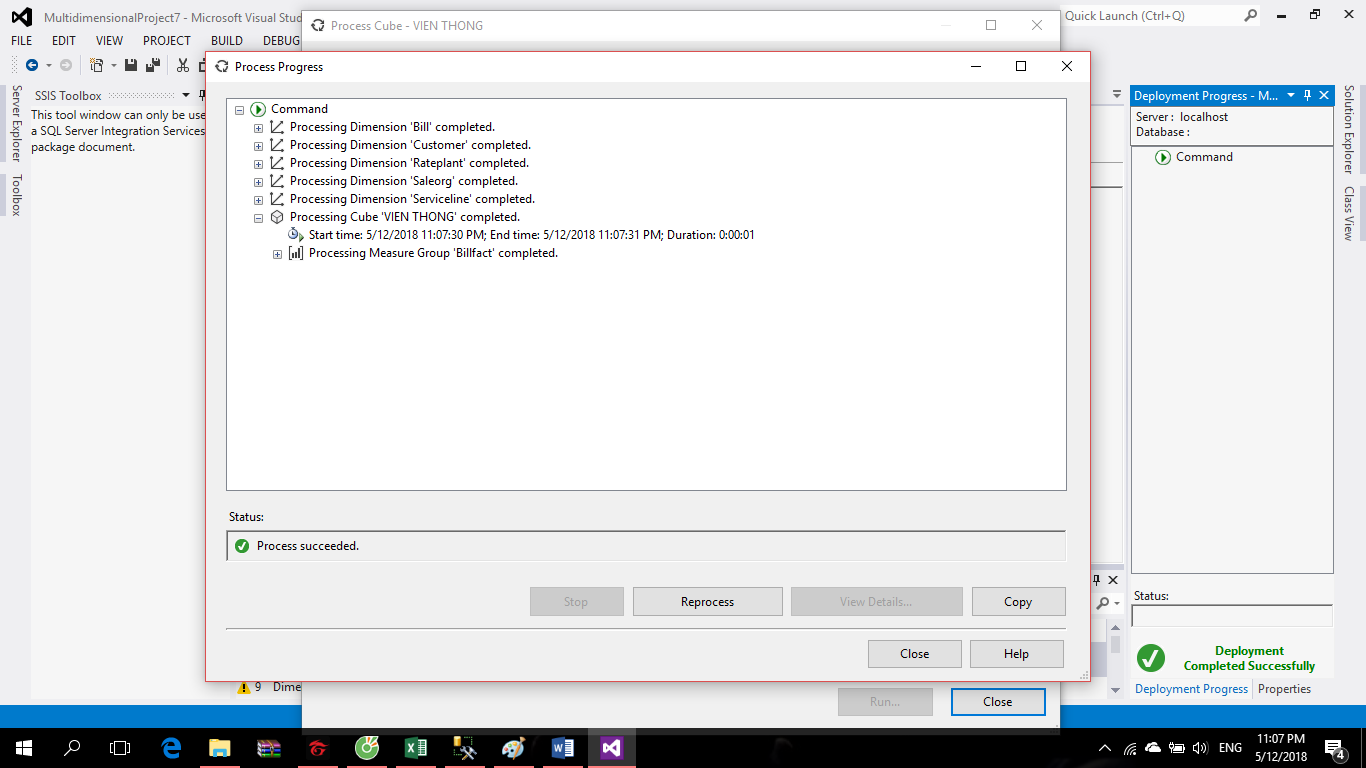
Kết quả:



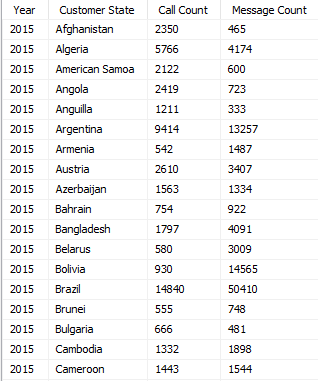
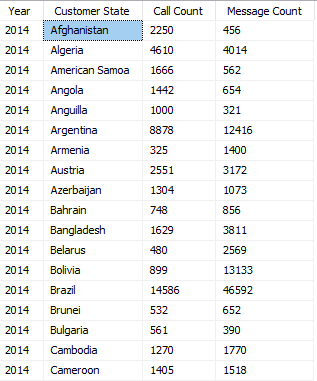
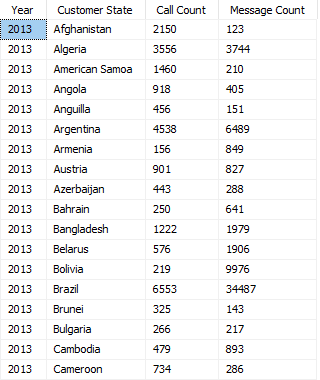
Hình 10. Thông báo hoàn thành cube

Như vậy đã tạo xong cube.

Tiếp theo, ta process và deploy cube vừa tạo

Hình 11. Process và deploy thành công

Như vậy là đã xong, phần tiếp theo là trích xuất và phân tích dữ liệu để giúp phần trả lời câu hỏi bài toán.



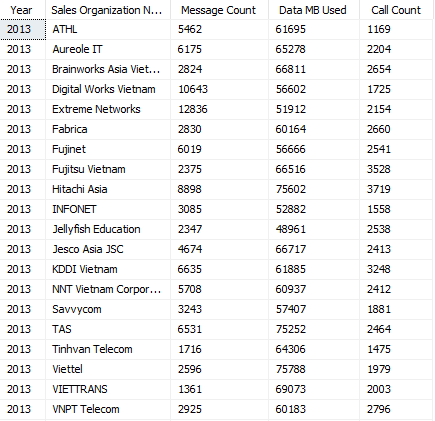
2013 2014 2015

Hình 12. Tổng số tin nhắn và cuộc gọi theo Quốc gia trong 3 năm

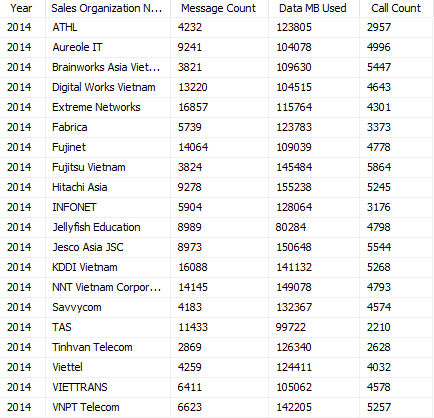


2013 2014 2015

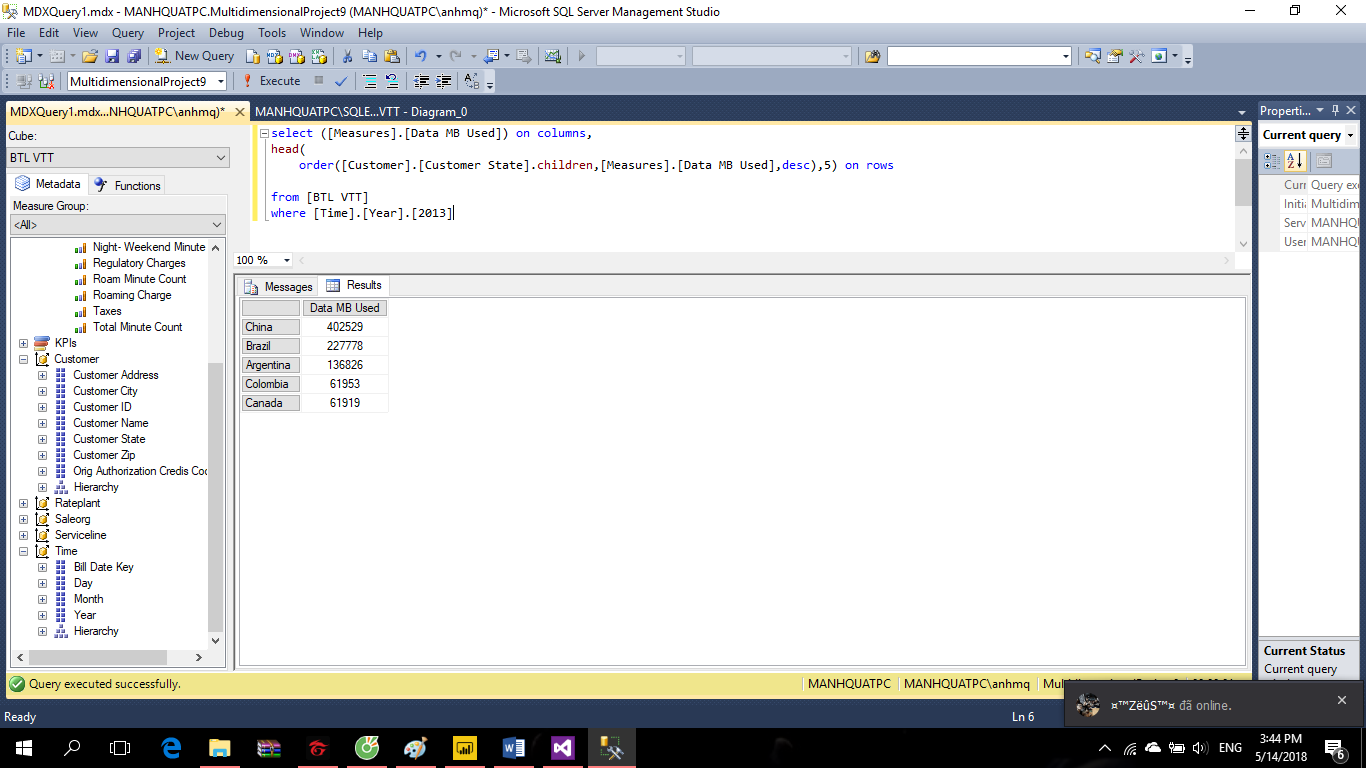
Hình 13. Lượng data sử dụng theo quốc gia trong 3 năm



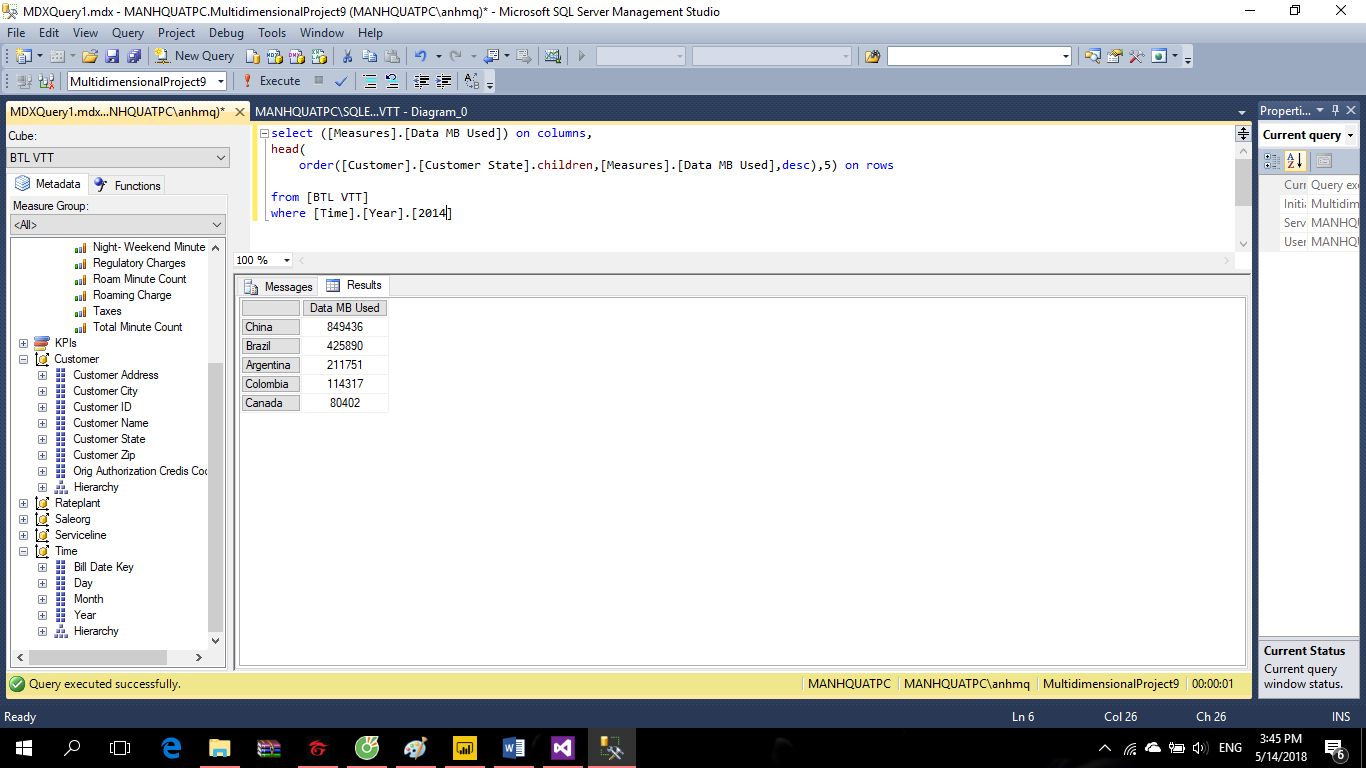
Hình 14. Tin nhắn, data sử dụng, số cuộc gọi theo nhà mạng năm 2013



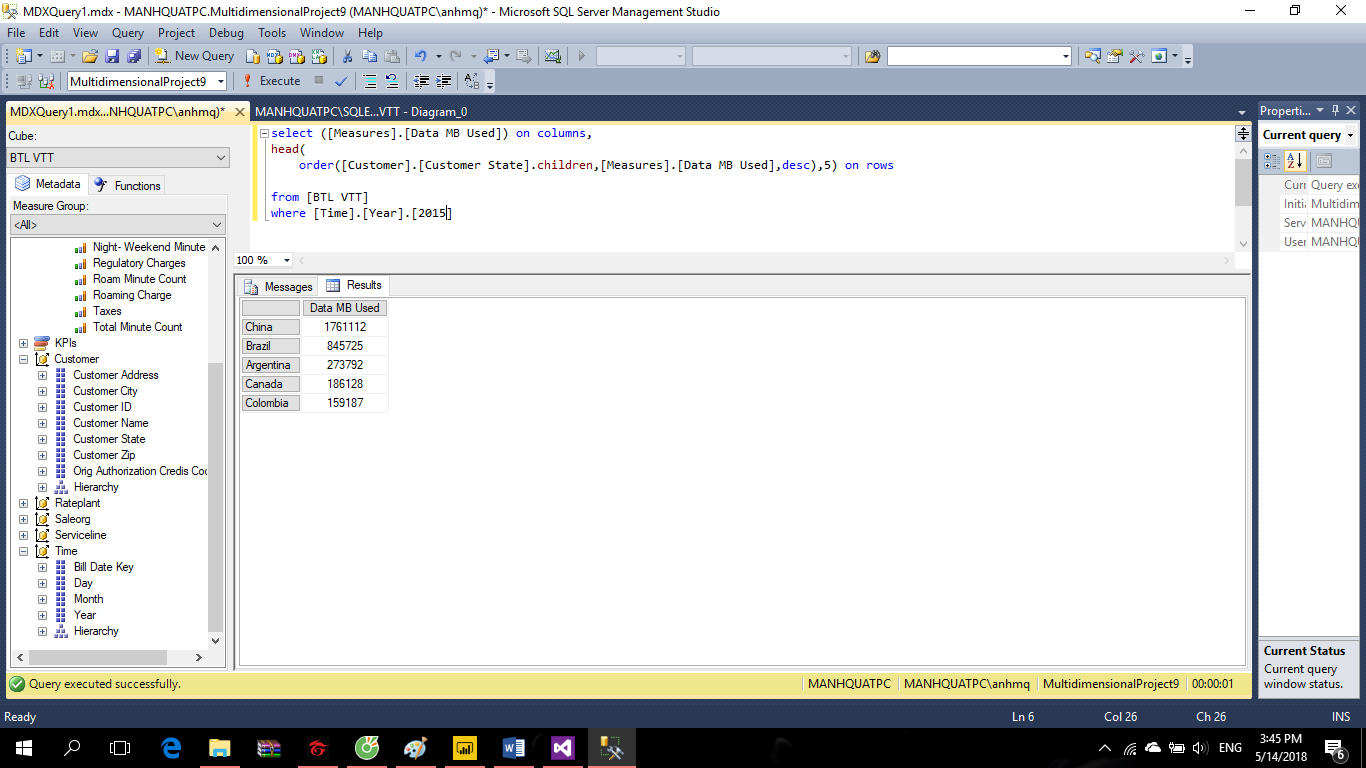
Hình 15. Tin nhắn, data sử dụng, số cuộc gọi theo nhà mạng năm 2014



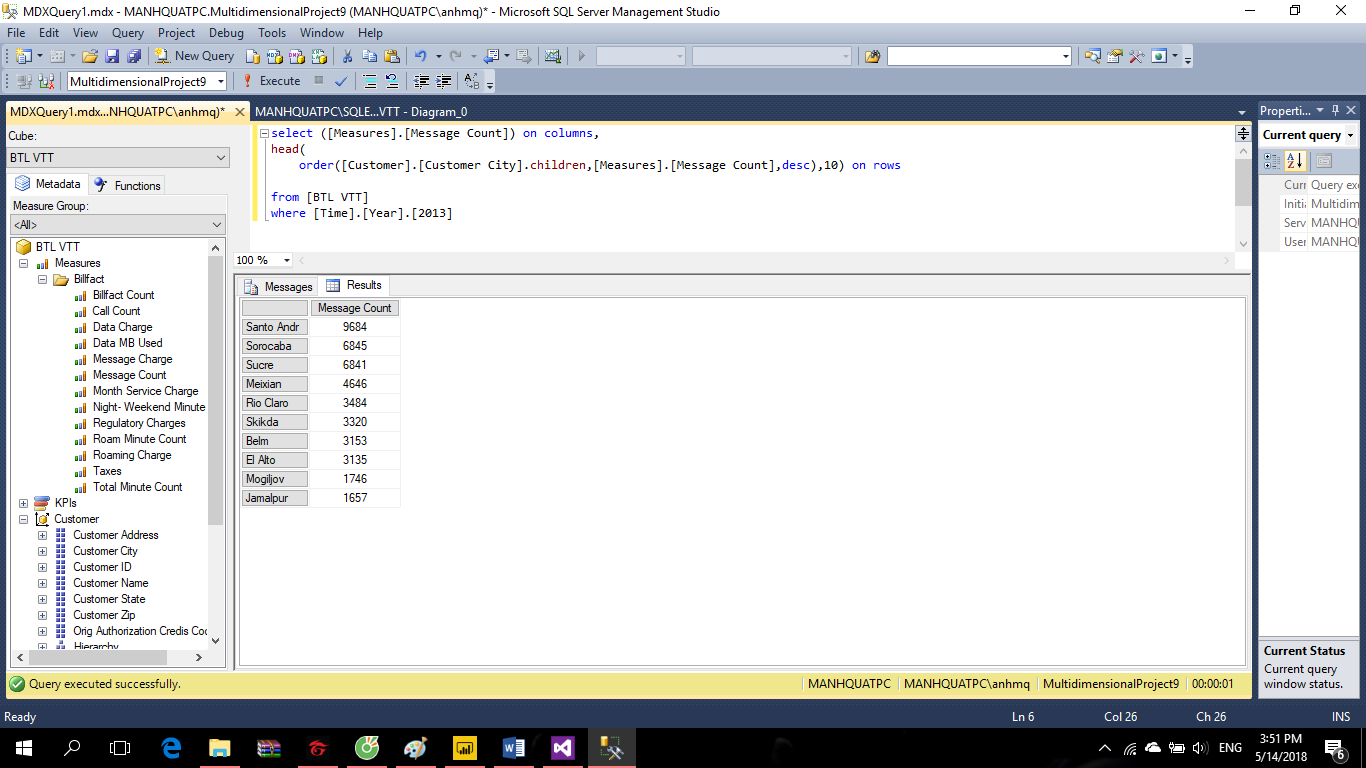
Hình 16.1 Top 5 nước sử dụng data nhiều nhất năm 2013



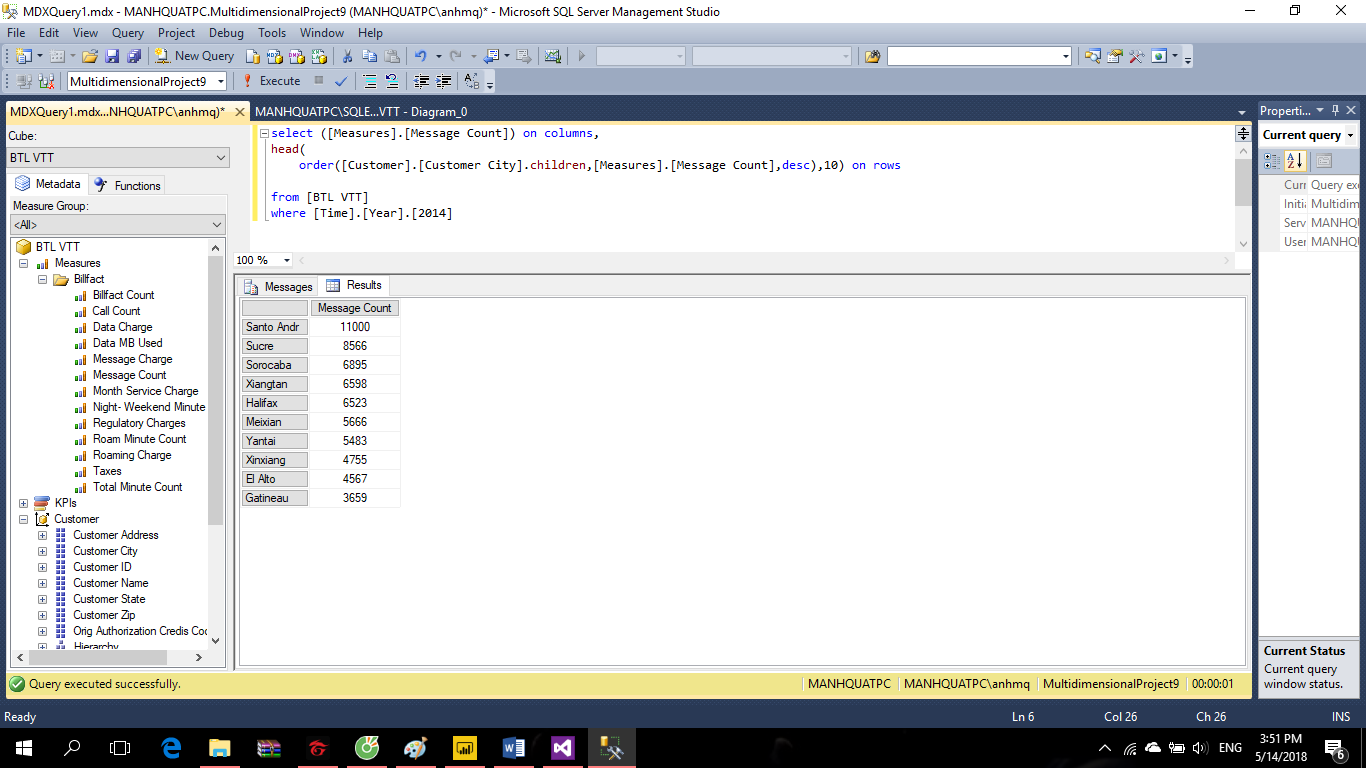
Hình 16.2 Top 5 nước sử dụng data nhiều nhất năm 2014



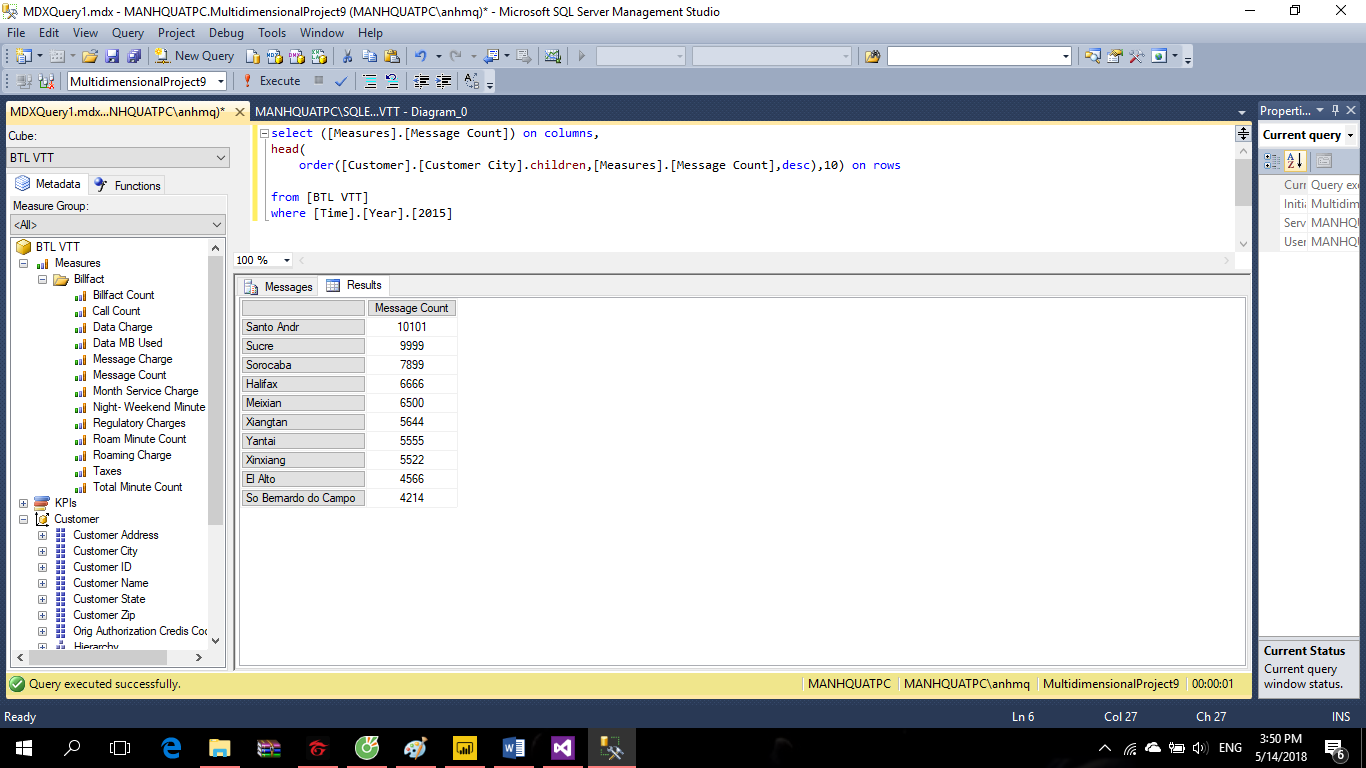
Hình 16.3 Top 5 nước sử dụng data nhiều nhất năm 2015



Hình 17.1 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2013

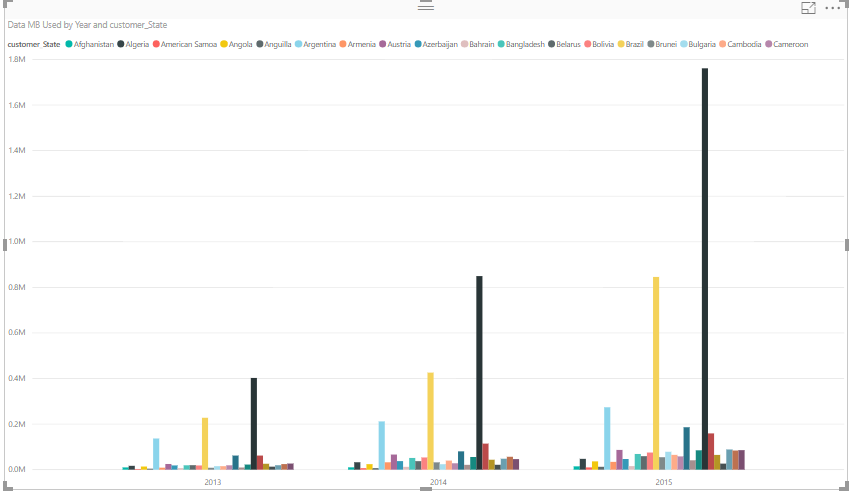


Hình 17.2 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2014



Hình 17.3 Top 10 thành phố sử dụng dịch vụ tin nhắn nhiều nhất 2015

### Sử dụng công cụ power BI để vẽ biểu đồ dữ liệu, nhìn dữ liệu một cách trực quan hơn.

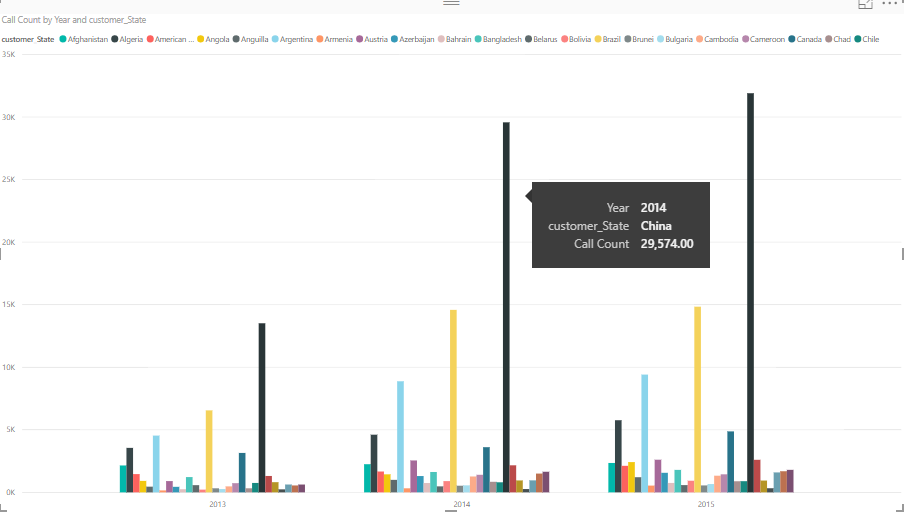


Hình 18. Biểu đồ: lượng data sử dụng của các quốc gia trong 3 năm từ 2013 -2015

Từ biểu đồ lượng data sử dụng của các quốc gia trong 3 năm từ 2013 – 2015 ta thấy:

* Dung lượng data sử dụng của các quốc gia đều tăng qua các năm và năm 2015 tăng khá mạnh so với năm 2014
* 3 quốc gia sử dụng dung lượng data nhiều nhất là: China, Brazil và Argentina
* Năm 2013, Canada và Colombia sử dụng khá ít dung lượng data nhưng đến năm 2015, 2 nước này tăng đáng kể lượng sử dụng data
* Trả lời câu hỏi: Quốc gia nào là thị trường lớn, quốc gia nào có khả năng phát triển trong những năm tiếp theo?

**→ Nhìn vào biểu đồ thấy được quốc gia nào có tốc độ tăng trưởng nhanh nhất, giúp các công ty trả lời được câu hỏi nên đầu tư vào thị trường nào để tăng lợi nhuận của nhà mạng, thị trường nào có tiềm năng phát triển**



Hình 19. Biểu đồ thể hiện số cuộc gọi theo quốc gia trong 3 năm 2013 - 2015

Từ biểu đồ thể hiện số cuộc gọi của các quốc gia trong 3 năm từ 2013 – 2015 ta thấy:

* Nhu cầu gọi điện rất lớn từ các thuê bao
* Năm 2014 so với năm 2013: Tốc độ tăng trưởng các cuộc gọi khá nhanh. Nhưng năm 2015 so với 2014: tốc độ tăng trưởng đã chậm lại, điển hình là China. Từ 2013 – 2014, số cuộc gọi tăng gấp đôi, nhưng 2014 – 2015, số cuộc gọi đã tăng rất chậm.
* Trả lời câu hỏi: Các nhà mạng có cần phải đưa ra các gói khuyễn mại để tăng nhu cầu của gọi khách hàng không?

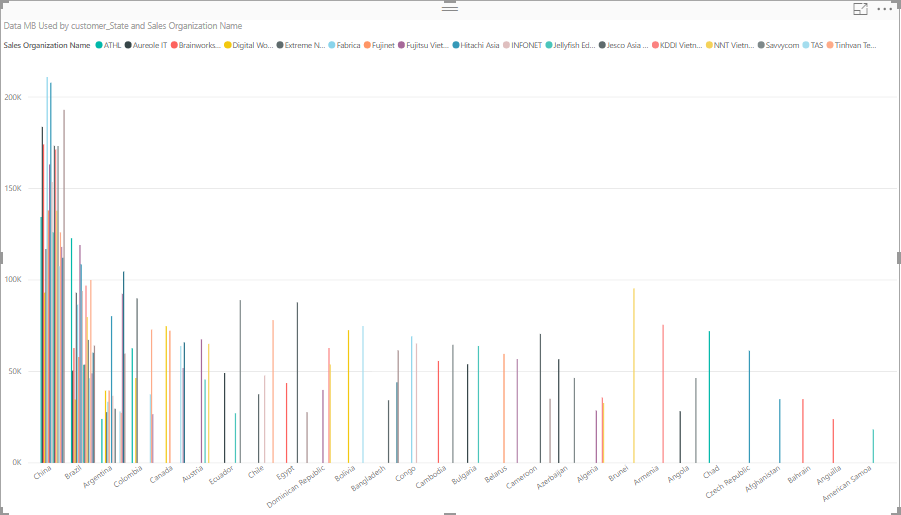
**→ Như vậy các nhà mạng cần phải có những giải pháp để tăng cuộc gọi từ các khách hàng như đưa ra các khuyến mãi, đưa ra các gói cước ưu đãi.**

****

**Hình 20.** Phí data của các nhà mạng từ 2013 – 2015

Nhìn vào biểu đồ doanh thu phí data của các nhà mạng từ 2013 – 2015 thấy rằng: Từ khoảng thời gian 2014 – 2015, doanh thu phí data tăng gần như gấp đôi so với thời gian trước đó. Và nhìn vào biểu đồ cũng thấy rõ nhà mạng nào đang phát triển nhanh nhất. Doanh thu từ phí data của các nhà mạng đều tăng, lượng data sử dụng của khách hàng ngày càng lớn, đây là một xu thế tất yếu hiện nay. Các nhà mạng nên đẩy mạnh phát triển vào mảng dung lượng data, doanh thu từ mảng này là doanh thu chính của các nhà mạng trong tương lai.

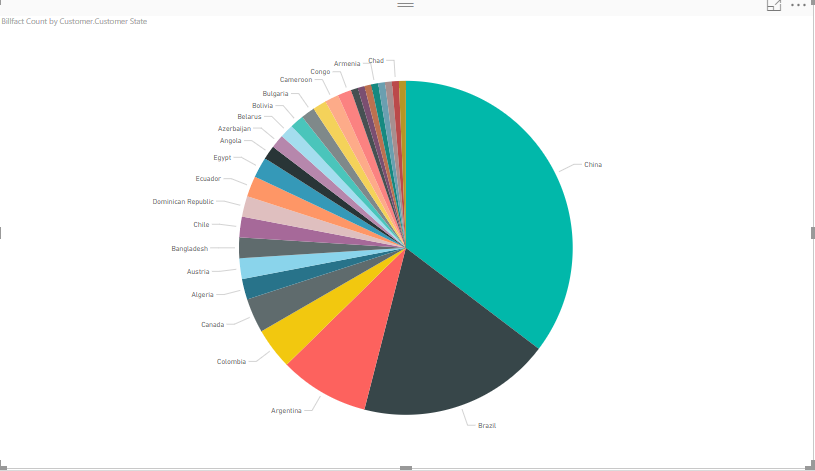
* Từ biểu đồ này và 2 biểu đồ trước sẽ giúp trả lời câu hỏi: Xu thế sử dụng dịch vụ của khách hàng như thế nào?



Hình 21: Biểu đồ thể hiện sự phân bố của các nhà mạng theo quốc gia

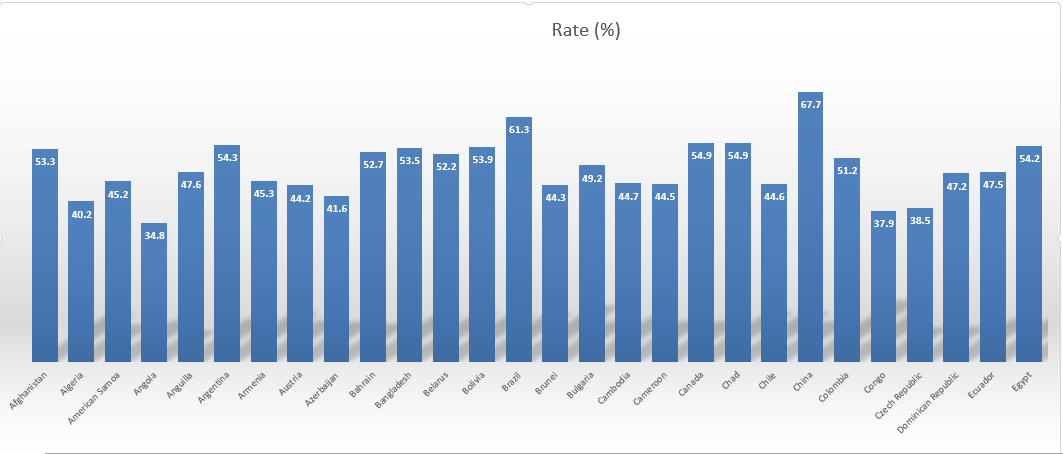
Từ biểu đồ thể hiện sự phân bố của các nhà mạng theo quốc gia ta thấy:

* Còn rất nhiều thị phần béo bở cho các nhà mạng, 1 số quốc gia hiện chỉ có 1, 2 nhà mạng. Nhu cầu phát triển còn rất lớn
* Các quốc gia như: China, Brazil, Argentina có sự cạnh tranh rất lớn của các nhà mạng.



Hình 22: Biểu đồ thể hiện số hoá đơn trên các quốc gia.

Như chúng ta thấy, Trung quốc và Brazil chiếm hơn 55% số hoá đơn viễn thông trên toàn thế giới. điều này cho thấy đây là 2 quốc gia này có lượng người dùng rất là lớn, là một thị trường tiềm năng cho các nhà phát triển viễn thông.



Hình 23: Biểu đồ tỉ lệ dân số sử dụng viễn thông của các quốc gia.

Nhìn vào biểu đồ ta thấy, Trung quốc có tỉ lệ dân số dùng viễn thông cao vượt trội so với các quốc gia khác ( 67.7 % dân số), tiếp theo là Brazil ( với 61.3 % dân số).

Có thể thấy, Trung Quốc là quốc gia có lượng người dùng rất lớn với số dân đông nhất thế giới và tỉ lệ người dùng viễn thông cao nhất thế giới.

* Trả lời câu hỏi: Sự cạnh tranh của các nhà mạng tại Trung Quốc ra sao?

**→ Các nhà mạng có thể chuyển hướng đầu tư cho các năm tiếp theo hoặc mở rộng thị phần của mình.**