**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**

**THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO KHO DỮ LIỆU**

**ĐỀ TÀI:**

**XÂY DỰNG KHO DỮ LIỆU**

**TRÊN TẬP DỮ LIỆU COVID-19**

**GV hướng dẫn: Thầy Quách Đình Hoàng**

**Nhóm sinh viên thực hiện:**

|  |  |  |
| --- | --- | --- |
| ***Họ và tên*** | ***MSSV*** |  |
| **Nguyễn Tiến Duy** | **17133008** |  |
| **Bồ Xuân Hoàng** | **17133022** |  |
| **Lê Kha** | **17133030** |  |
| **Nguyễn Hiền Nhung** | **17133044** |  |
| **Phan Tấn Thịnh** | **17133060** |  |
| **Hoàng Thị Cẩm Tú** | **17133071** |  |

**Hồ Chí Minh, 2020**

**Mục lục**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | |  |
|  | | |  |
|  |  | |  |
|  |  | |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  | |  |
|  |  | |  |
|  |  | |  |

**Lời cảm ơn**

Trong thời gian làm đồ án, nhóm em đã nhận được nhiều sự giúp đỡ, đóng góp ý kiến và chỉ bảo nhiệt tình của thầy cô, bạn bè.

Nhóm em xin gửi lời cảm ơn chân thành đến Th.s Quách Đình Hoàng, trường Đại Học Sư Phạm Kỹ Thuật Tp.HCM, người đã tận tình hướng dẫn, chỉ bảo chúng em trong suốt quá trình học tập và hoàn thiện đề tài.

Nhóm em cũng xin chân thành cảm ơn các thầy cô giáo trong trường Đại Học Sư Phạm Kỹ Thuật Tp.HCM nói chung, các thầy cô trong khoa Công Nghệ Thông Tin nói riêng đã dạy dỗ cho chúng em kiến thức về các môn đại cương cũng như các môn chuyên ngành, giúp chúng em có được cơ sở lý thuyết vững vàng và tạo điều kiện giúp đỡ em trong suốt quá trình học tập.

**Lời mở đầu**

Dịch Covid-19 bùng nổ trên toàn cầu vào đầu năm 2020. Và ở Hàn Quốc đã lây nhiễm cho hơn 10000 người. Vì vậy, đề tài của nhóm em đề xuất một thiết kế kho dữ liệu cho một hệ thống thông tin về các bệnh nhân nhiễm Covid trên lãnh thổ Hàn Quốc, nó có vai trò giúp đưa ra quyết định phòng tránh, cách ly, dự đoán được tình hình dịch bệnh một cách chính xác nhất.

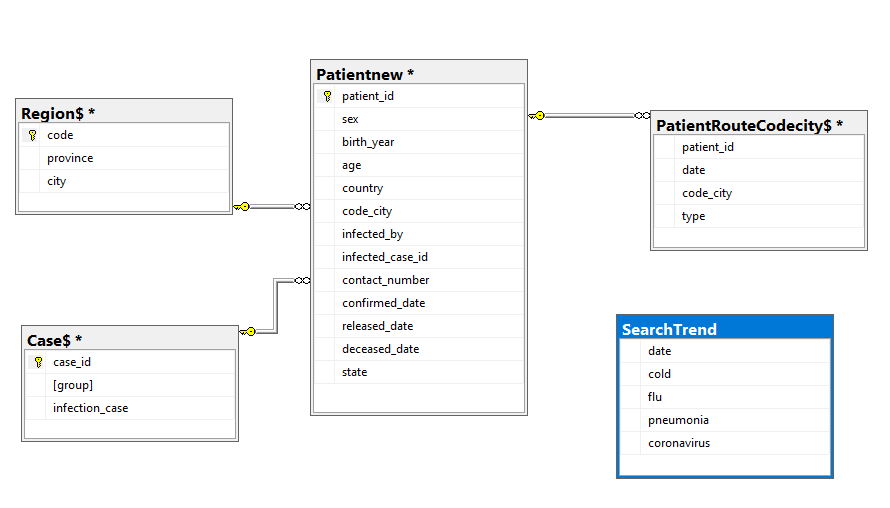
**Chương 1: Giới thiệu cơ sỡ dự liệu**

**1.1** **Giới thiệu về cơ sở dữ liệu**

Đây là dữ liệu về thông tin các bệnh nhân nhiễm COVID19 tại Hàn Quốc dựa trên các tài liệu báo cáo của KCDC ( Trung tâm kiểm soát và phòng ngừa dịch bệnh Hàn Quốc) và chính quyền địa phương. Dữ liệu được ghi nhận từ ngày 20/1/2020 tới ngày 1/6/2020.

Link dataset: <https://www.kaggle.com/kimjihoo/coronavirusdataset>

Database schema:



Patientnew: thể hiện thông tin của những người nhiễm Covid như sex, birth\_year, age,.., infected\_by là bị lây nhiễm bởi người nào, infected\_case\_id là lây từ nguồn dịch nào,contact\_number là số người đã tiếp xúc qua. Khóa chính là id của người nhiễm. Khóa ngoại là code\_city tham chiếu đến bảng Region. Infected\_case\_id tham chiếu đến bảng Case$.

PatientRouteCity$: trình bày những nơi mà người nhiễm Covid đã đi qua. Khóa ngoại là patient\_id tham chiếu đến bảng Patientnew.

Region$: thể hiện các tỉnh, thành phố của Hàn Quốc và mã số của mỗi vùng.

Case$: thể hiện các nguồn lây nhiễm dịch và mã id của mỗi nguồn.

**1.2** **Lý do cần có một data warehouse**

Datawarehouse giúp tổ chức thống kê thông tin về dịch tễ học của người nhiễm, các ổ dịch chính, các địa điểm người nhiễm bệnh lui tới nhiều nhất, province nào có nhiều người nhiễm, sự quan tâm dịch bệnh của cộng đồng để đưa ra quyết định phòng tránh, ngăn ngừa lây nhiễm bệnh và hướng giải quyết phù hợp.

**Chương 2: Hiện thực kho dữ liệu**

Đễ hiện thực kho dữ liệu việc đầu tiên cần làm là xác định các *Business Process* mà kho dữ liệu phục vụ. Từ đó thiết kế các *Data Mark* phù hợp và tích hợp dữ liệu đã có vào các *Data Mark* vừa được thiết kế để lưu trữ, tổng hợp, phân tích và có thể lập thành các báo cáo theo ý muốn.

**2.1** **Chọn các Business Process**

Tập dữ liệu *Covid\_19* là tập dữ liệu về dịch bênh covid-19 tại Korea. Một số thông tin có thể khai thác từ bộ dữ liệu như: thông tin người bị nhiễm bệnh, địa điểm nào có mức độ lây nhiễm cao nhất? Bệnh nhân đã đi qua những địa điểm nào nhiều nhất? Nguyên nhân bị nhiễm bệnh nhiều nhất là từ đâu?,… Chúng tôi đã lựa chọn ra một số câu hỏi để hiện thực kho dữ liệu trên. Những câu hỏi như sau:

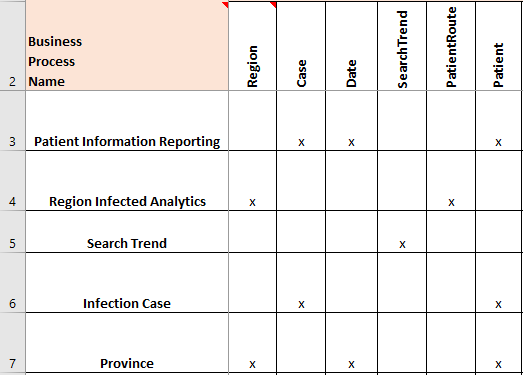
* Patient Information Reporting: theo dõi được thông tin người bệnh như độ tuổi, giới tính, ngày nhiễm bệnh, nơi sống, ngày khỏi bệnh, số người lây nhiễm, thời gian hồi phục.
* Province: theo dõi mức độ nhiễm bệnh như số người nhiễm bệnh, số người chết, số người hồi phục theo thành phố, địa điểm để tiến hành cách ly khu vực hay xóa bỏ cách ly. Mức độ nhiễm bệnh nên được cập nhập hằng ngày để phân tích.
* Infection Case: phân tích các trường hợp nhiễm bệnh để biết được nguồn bệnh từ đâu để phòng tránh, kiểm soát được nguồn nhiễm bệnh.
* Region Infected Analytics : phân tích lộ trình di chuyển của bệnh nhân từ đó biết được địa điểm có tỉ lệ lây nhiễm bệnh cao,
* Search Trend: theo dõi mức độ quan tâm các dịch bệnh theo từ khóa tìm kiếm mỗi năm. Mực độ tìm kiếm nên được cập nhập hằng ngày để

**2.2** **Thiết kế Star Schema**

Công việc tiếp theo sau khi xác định các *Business Process* là xây dựng mô hình kho dữ liệu cho phù hợp, ở đây chọn xây dựng kho bằng mô hình sao *Star Schema*. Trong mô hình *Star Schema* mỗi *Business process* sẽ ứng với một bảng trung tâm, gọi là bảng *Fact*. Bảng trung tâm sẽ kết nối với những bảng thông tin mở rộng khác để làm rõ thông tin trong nó, những bảng xung quanh nó được gọi là các bảng *Dimension*. Các bảng *Fact* có thể chia sẽ các Dimension với nhau, hay có thể nói một *Dimension* có thể được dùng bởi nhiều bảng *Fact* [3].

**2.2.1** **Bus Matrix**

Để xác định *Dimension* nào được sử dụng bởi bảng *Fact* nào, công cụ thường được dùng là *Bus Matrix*. *Bus Matrix* tương ứng cho các *Business process* được chọn bên trên được thể hiện như hình **2.1**.



Hình 2.1: Bus Matrix cho Business process trong báo cáo

**2.2.2** **Các Dimension Table**

Các bảng *Dimension* phục vụ trong báo cáo được định nghĩa như sau:

DimRegion: Chứa thông tin về khu vực bao gồm code, province, city.

DimCase: Chứa thông tin về các trường hợp nhiễm bệnh bao gồm case\_id, [group], infection\_case.

DimDate: Chứa thông tin chi tiết về thời gian bao gồm DateKey, Date, DateOfWeek, DayName, DayOfMonth, DayOfYear, WeekOfYear, MonthName, MonthOfYear, Quarter, QuaterName, Year, IsAWeekday.

DimSearchTrend: Thông tin về mức độ tìm kiếm từ cộng đồng bao gồm DateKey, Date, year, Cold, Flu, pneumonia, Coronavirus.

DimPatientRoute: Chứa thông tin về tuyến đường thành phố bao gồm Patient\_Rkey, patient\_id, date, code\_city, type.

DimPatient: Thông tin của người bệnh bao gồm patient\_id, sex, birth\_year, age, country, code\_city, infected\_by, infected\_case\_id, contact\_number, confirmed\_date, released\_date, deceased\_date, state.

**2.2.3** **Các bảng Fact**

Đối với các *Business process* đã được xác định Patient Information Reporting, Region Infected Analytics, Search Trend, Infection Case, Province các bảng *Fact* phục vụ cho các *Business process* được định nghĩa như sau:

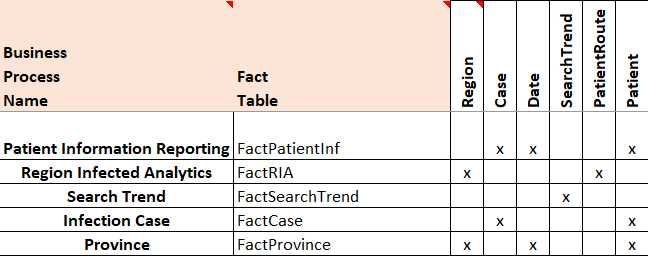
FactPatientInf là bảng Fact cho Business process Patient Information Reporting

FactRIA là bảng Fact cho Business process Region Infected Analytics

FactSearchTrend là bảng Fact cho Business process Search Trend

FactCase là bảng Fact cho Business process Infection Case

FactProvince là bảng Fact cho Business process Province



Hình 2.2: Các bảng Fact phục vụ cho các Business process

**2.3 Tích hợp dữ liệu (ELT – Extract Transform Load)**

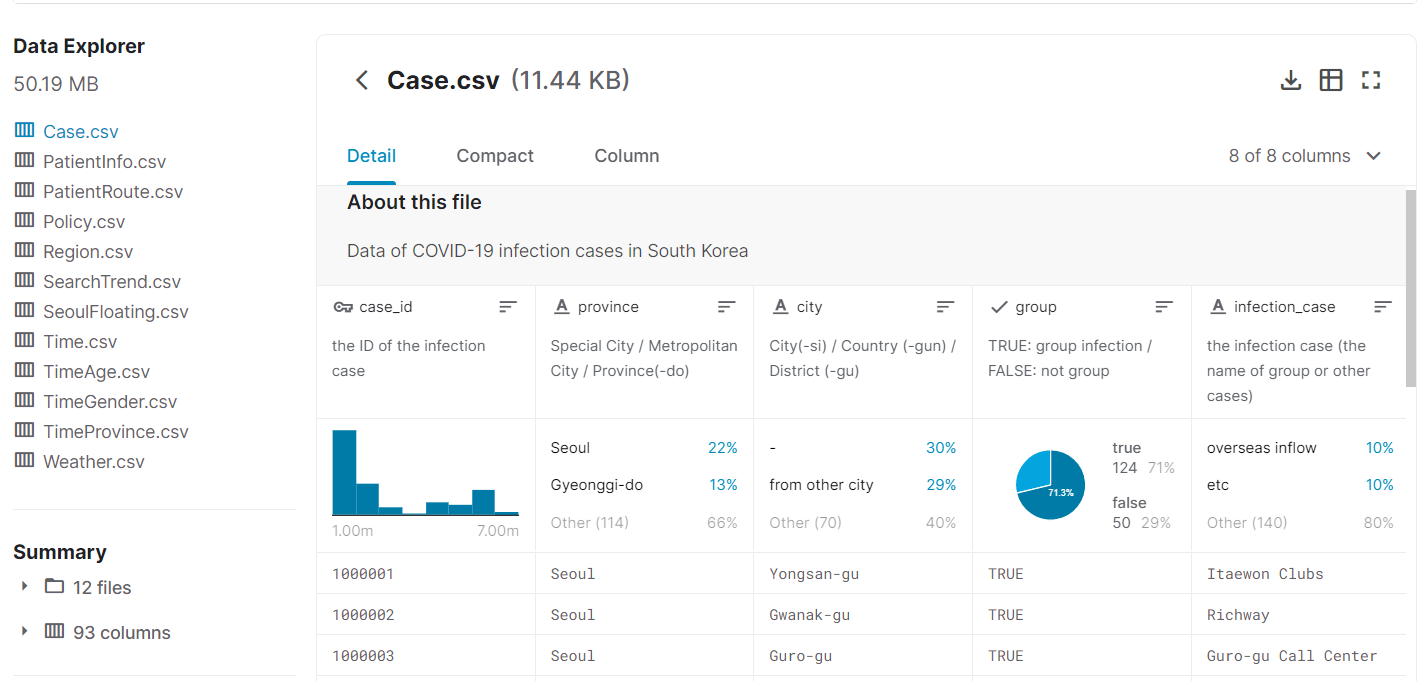
Trước khi bắt đầu tích hợp, ta cần chuẩn bị.

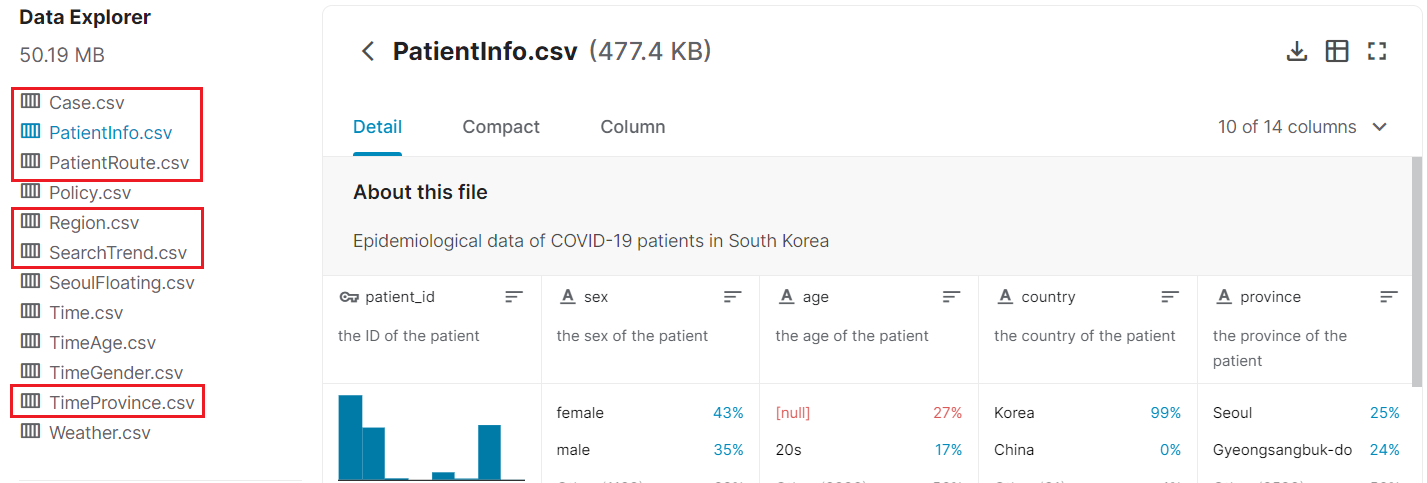
1) SQL Server Data tools: dùng để tạo CSDL cho Analysis Services.

2) SQL Server Analysis Services: dùng để triển khai Analysis Services.

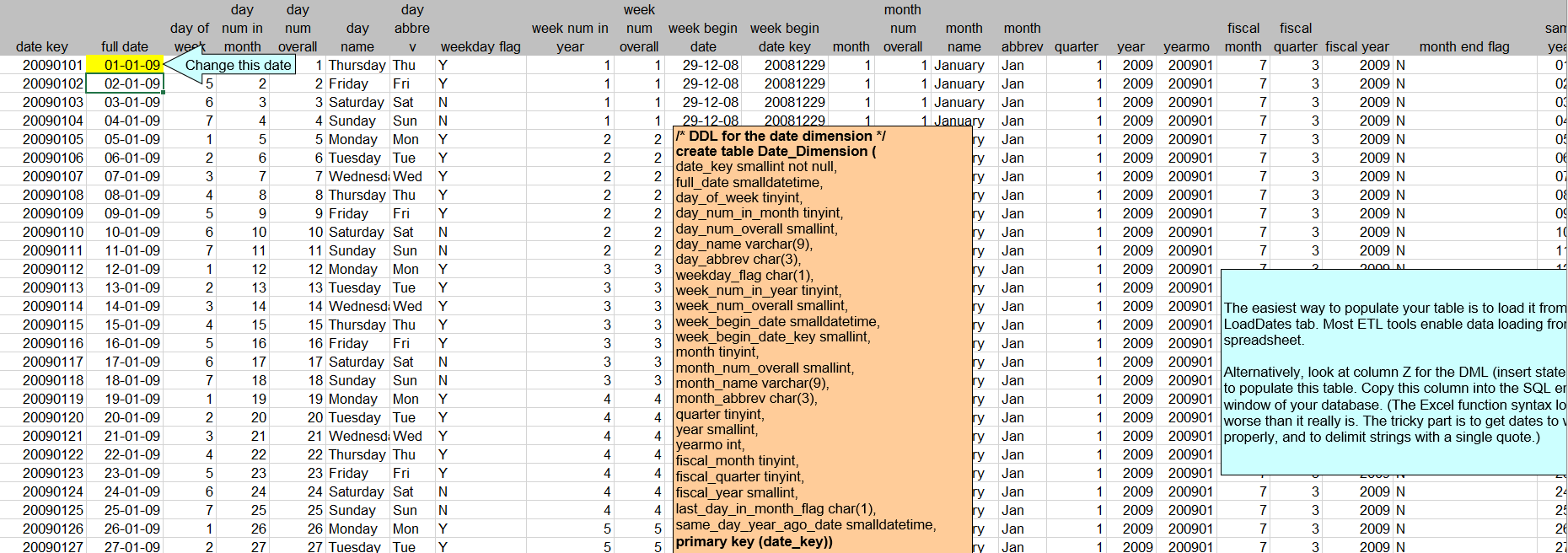
**a) Extracts dữ liệu**

Cơ sở dữ liệu ban đầu của database covid\_19 gồm 12 bảng.



Với việc có sẵn nguồn dữ liệu dataset của cơ sở dữ liệu Covid\_19, ta bắt đầu chiết xuất dữ liệu cần thiết để phù hợp cho bài làm. Dưới dây là dữ liệu đã được triết xuất từ dữ liệu nguồn bên trên bao gồm 6 bảng.

Thời gian trong dữ liệu nguồn là thời gian mà đã được thống kê trên từng thuộc tính (theo vùng, theo tuổi, theo bệnh nhân,….). Do đó ta cân thêm một bảng, bảng này chứa thông tin về thời gian, bảng có tên là DimDate.. Thay đổi ngày bắt đầu từ 1-1-2009 phù hợp với CSDL gốc.

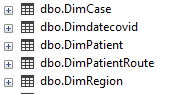


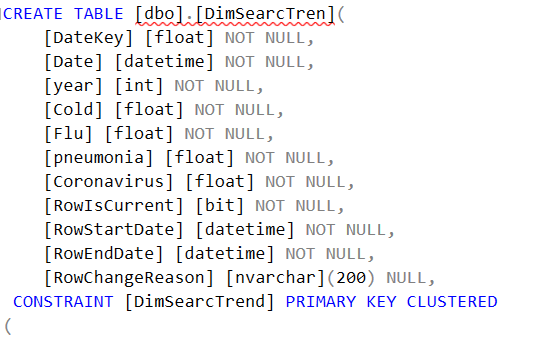
Sau khi có được cơ sở dữ liệu cần thiết, ta sẽ import vào Sql Server và thực hiện các bước tiếp theo.

**b) Tranforms dữ liệu**

Ta sẽ tiến hành tạo star schema dùng SQL. Star schema gồm các bảng: ***DimPatient, DimCase, DimPatientRoute, DimRegion, Dimdatecovid19, DimSearchTrend, FactCase, FactPatientInf, FactProvince, FactRIA, FactSearTrend.*** Ta sẽ thực hiện các bước sau:

* Tạo database trong SQL Server, đặt tên là **Covid\_19DW**. Database này chứa các bảng dimension và factable.
* **Dimensional**





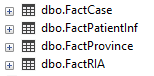
* **Facttable**

*FactPatientInf* : Thống kê thông tin từng bệnh nhân.

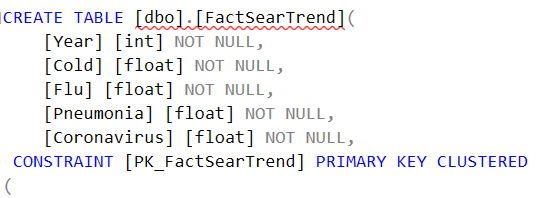
*FactProvince*: Thống kê những vùng có bị ảnh hưởng bởi covid19.

*FactCase*: Thống kê những ổ dịch chính

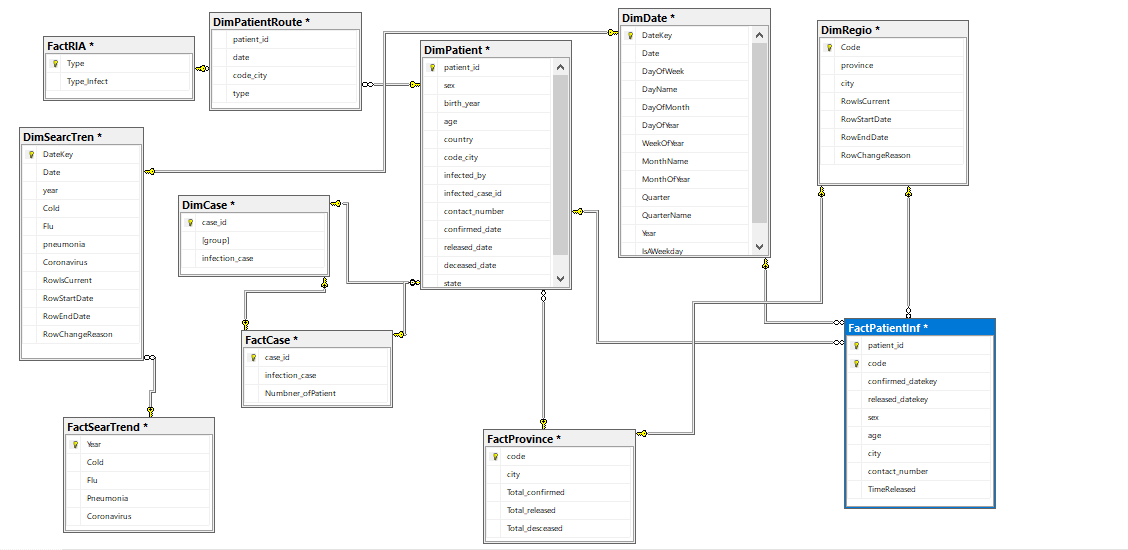
*FactRIA*:Thống kê địa điểm ảnh hưởng bởi dịch covid\_19.



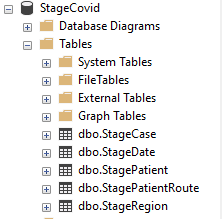
*FactSearchTren*:Thống kê mức độ quan tâm đến dịch bệnh của công đồng.



Sau khi tạo ta được star schema hoàn chỉnh như sau:



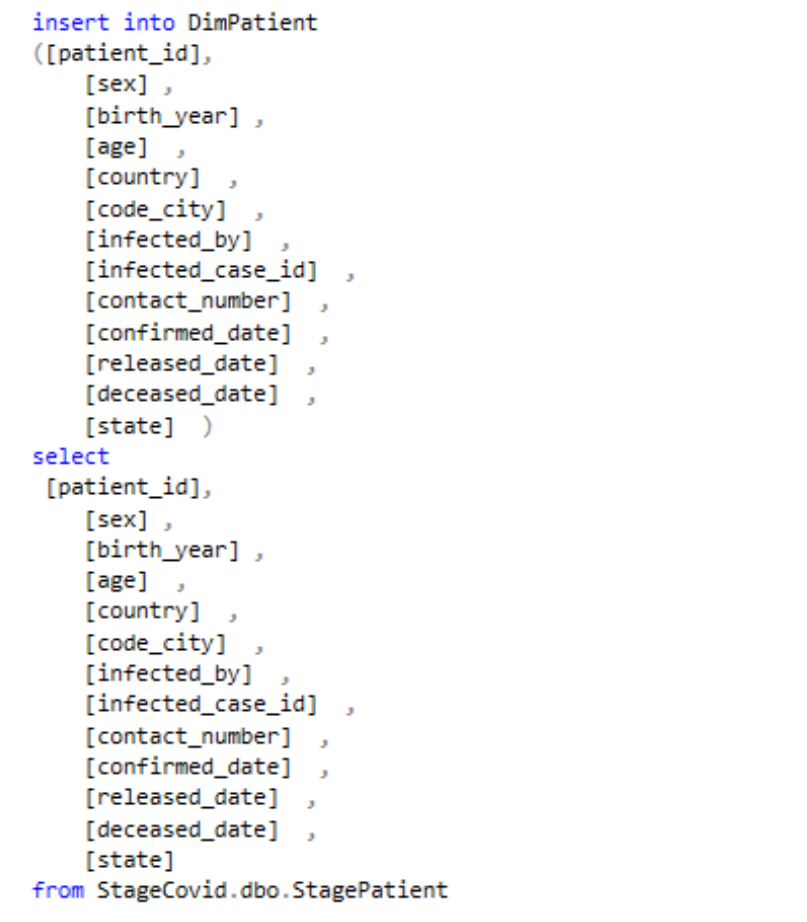
Trước khi load dữ liệu vào star schema, ta sẽ thực hiện bước staging dữ liệu nguồn. Tạo một CSDL staging, có thể đặt tên là **CovidStage**.



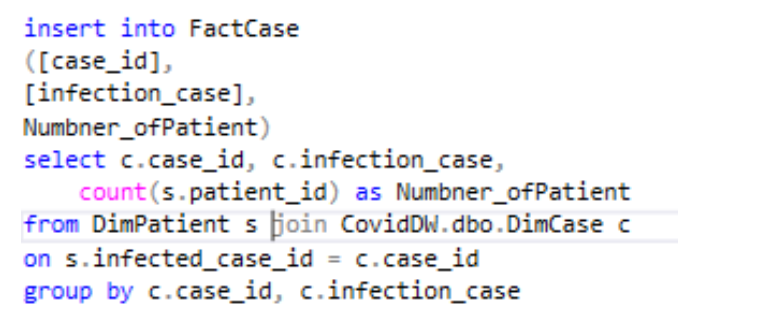
**c) Load dữ liệu**

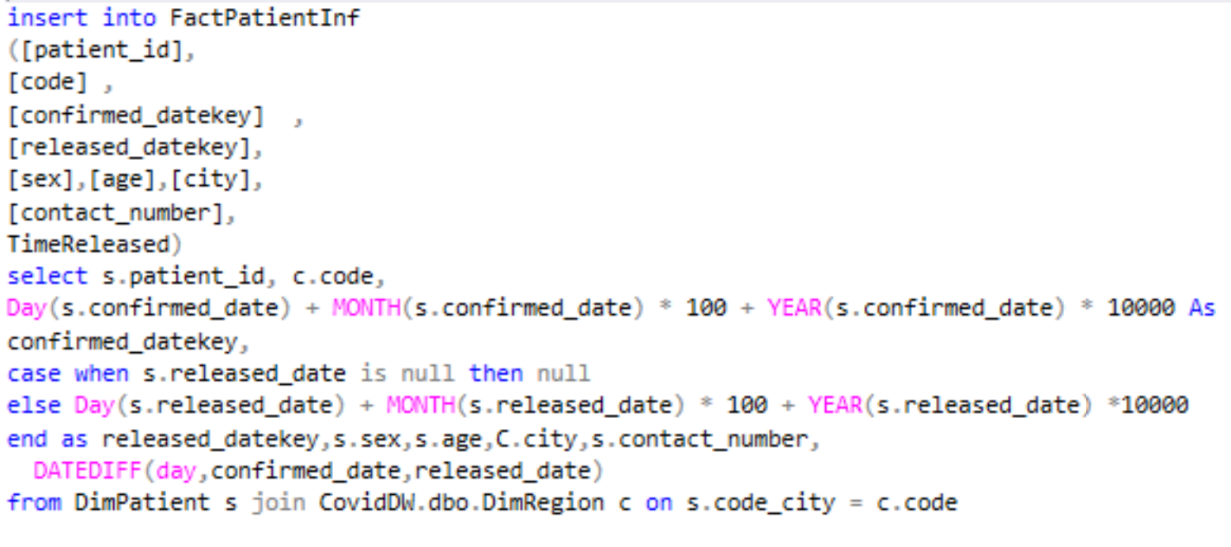
Sau khi có những bảng tạm, ta bắt đầu load dữ liệu từ stage table vào data warehouse.

Load bảng **DimPatient**

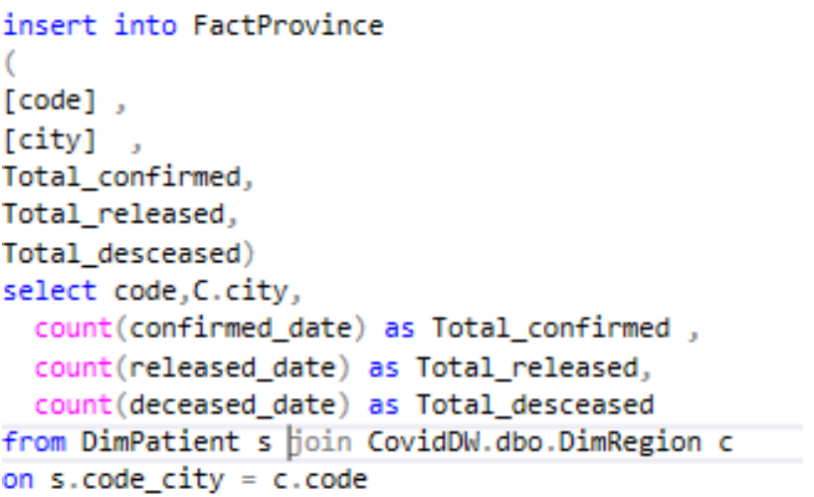


Tương tự ta load các bảng **DimPatientRoute, DimRegion, DimCase, Dimdatecovid**

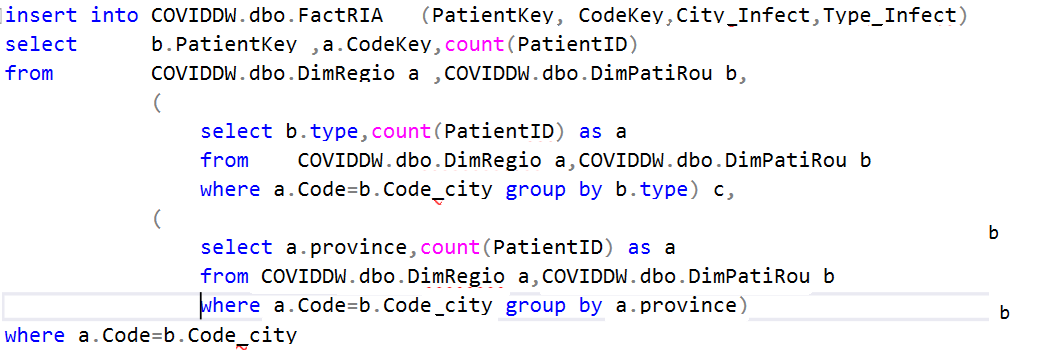
Load bảng **FactCase**

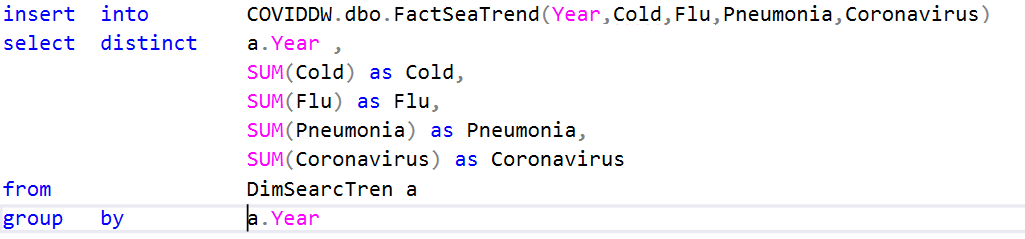
Load bảng **FactPatientInf**

Load bảng **FactProvince**



Load bảng **FactRIA**



Load bảng **FactSeaTrend**

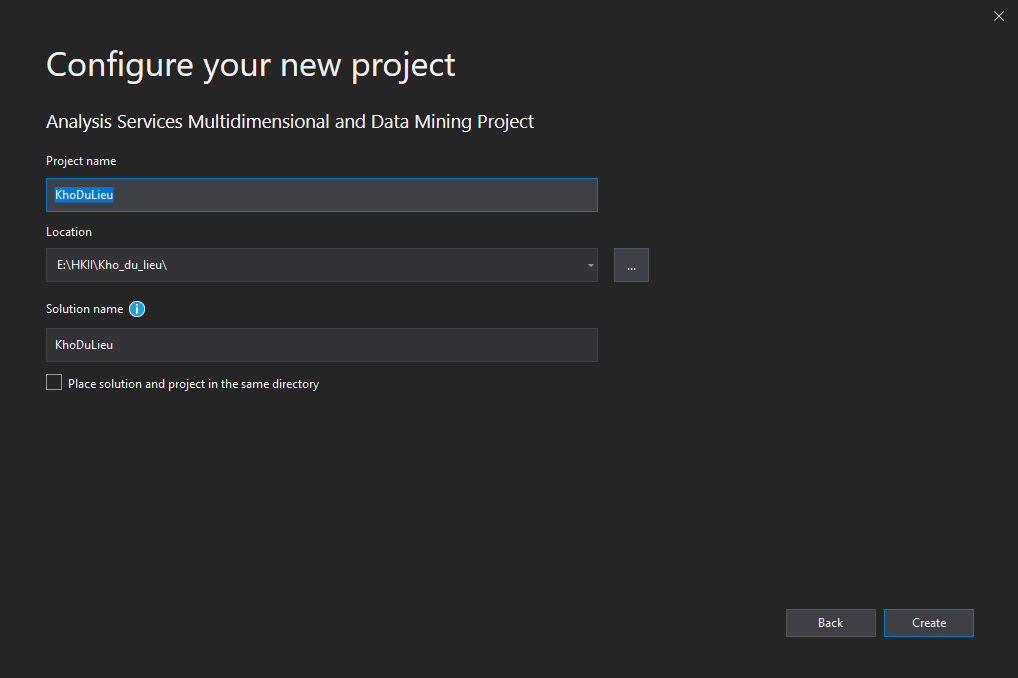
**2.4** **Hiện thực hóa Cube**

Sau khi dữ liệu được tích hợp vào kho dữ liệu, chúng ta sẽ tiến hành hiện thực hóacube phục vụ cho quá trình tổng hợp, phân tích dữ liệu. Việc hiện thực hóa cubeđược thực hiện theo trình tự sau:

* + 1. **Tạo một Analysis Services Project**

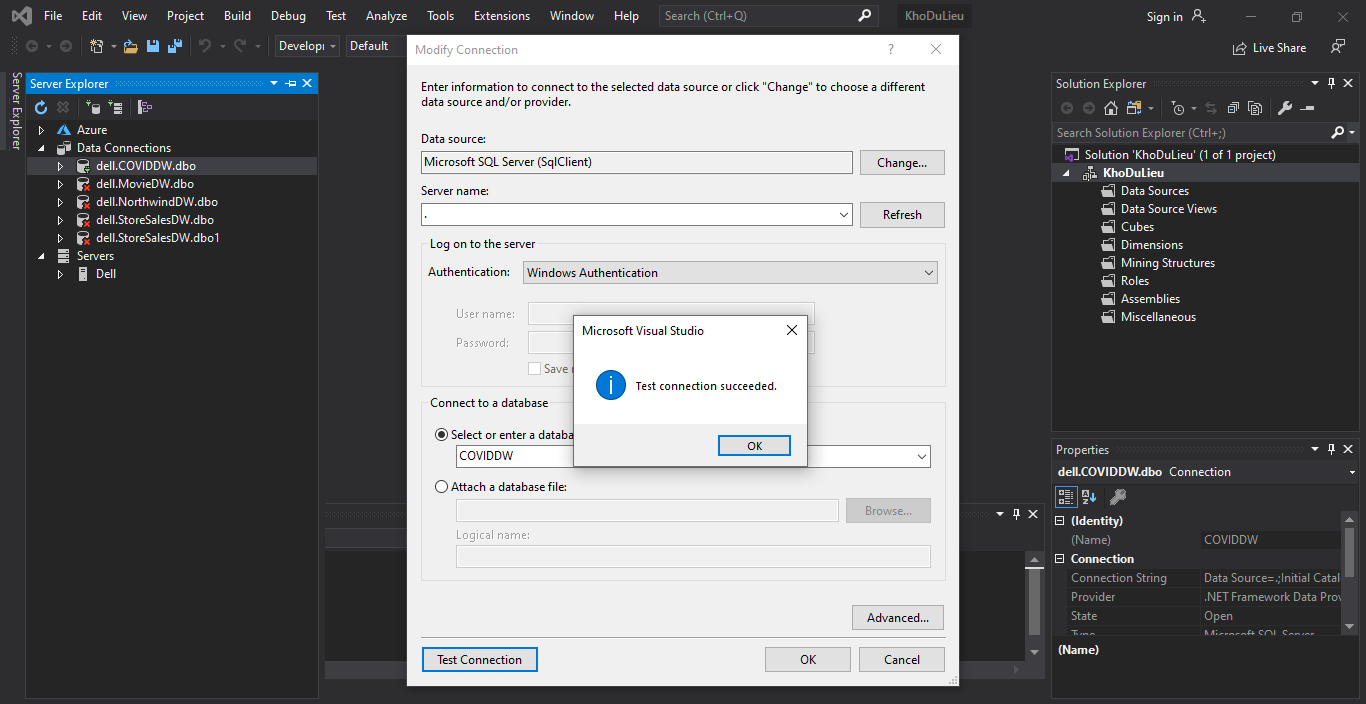
Sử dụng Visual Studio (có extension Microsoft Analysis Service)

Mở Visual Studio -> File -> New -> Project -> Analysis Services Multidimensional Project, đặt tên cho project và click OK.



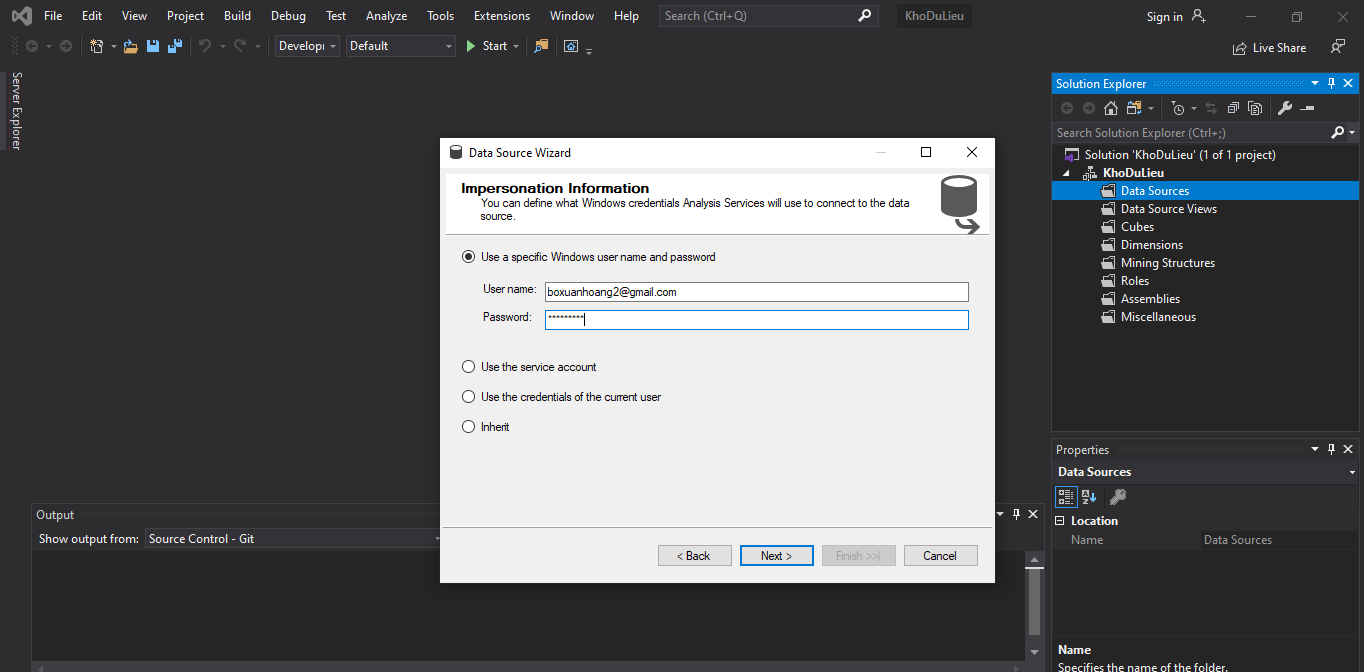
* + 1. **Kết nối đến Analysis Services Server**

Right click vào “Data Connections” ở mục Server Explorer chọn “Add Connection …” sẽ xuất hiện dialog “Choose Data Source”, chọn “Microsoft SQL Server” và click “Continue”

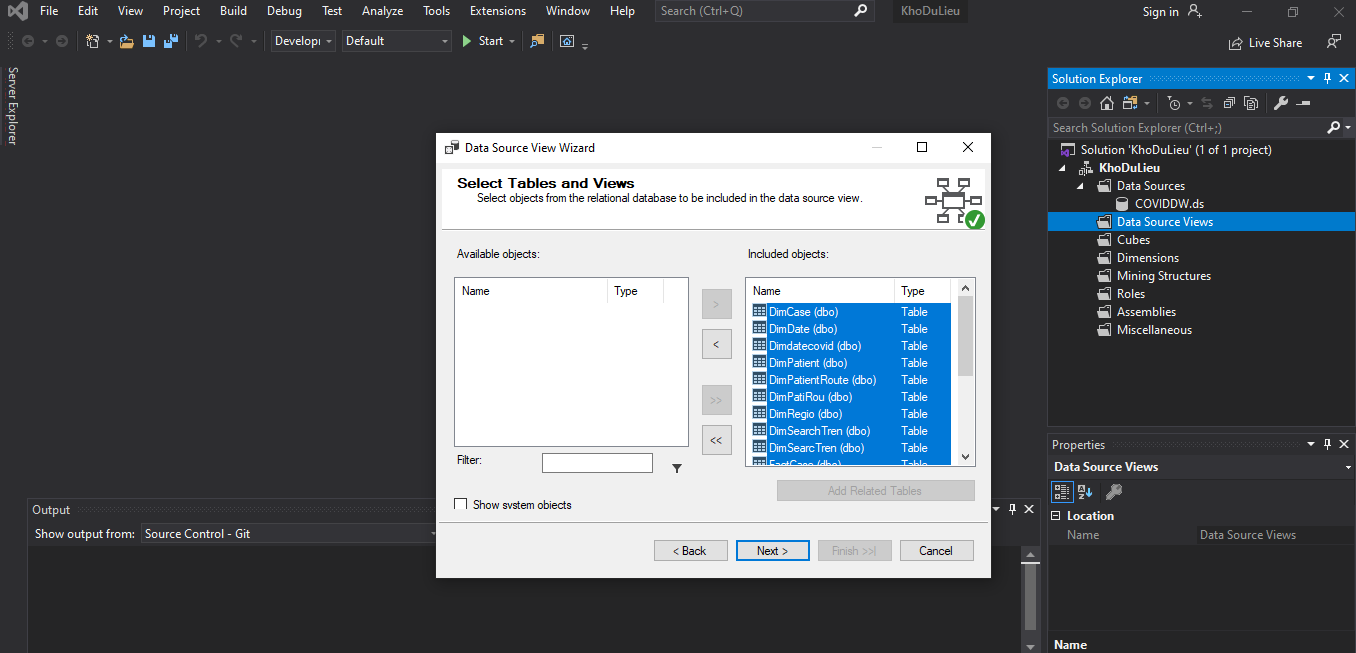


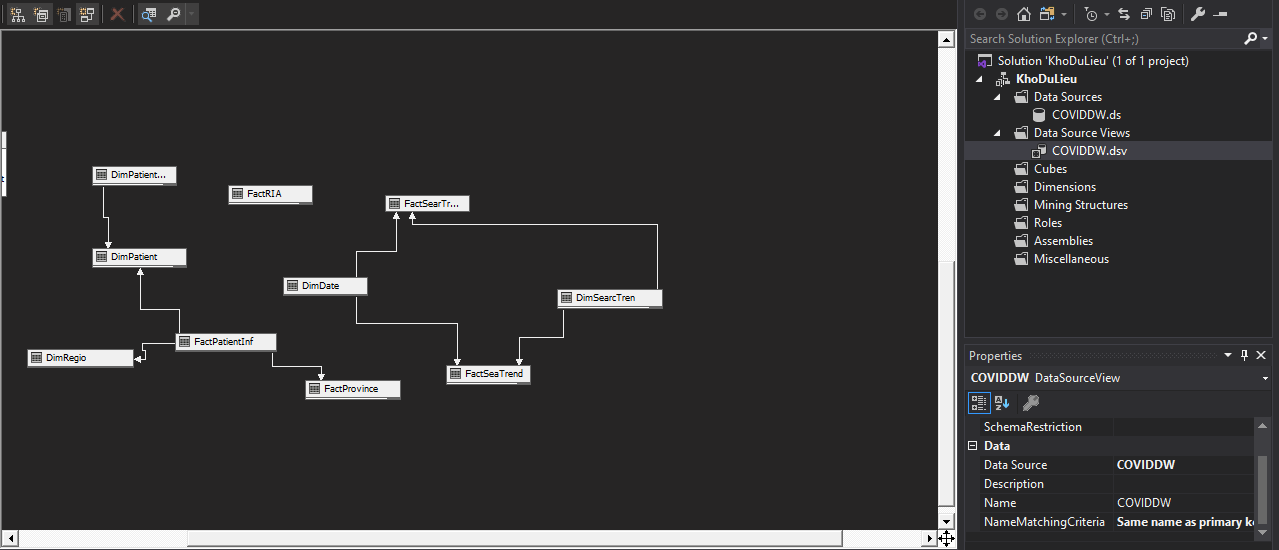
* + 1. **Tạo một Data Cube**

1. **Tạo một data source cho project**

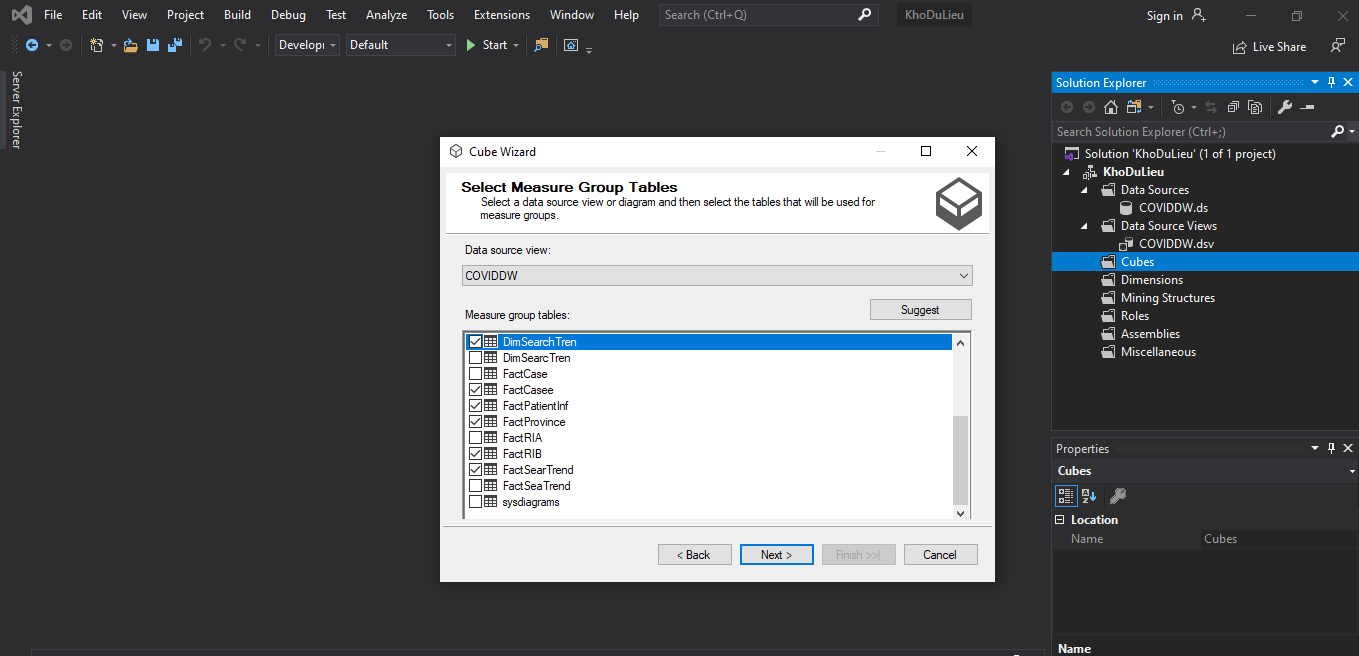


1. **Tạo data source view**

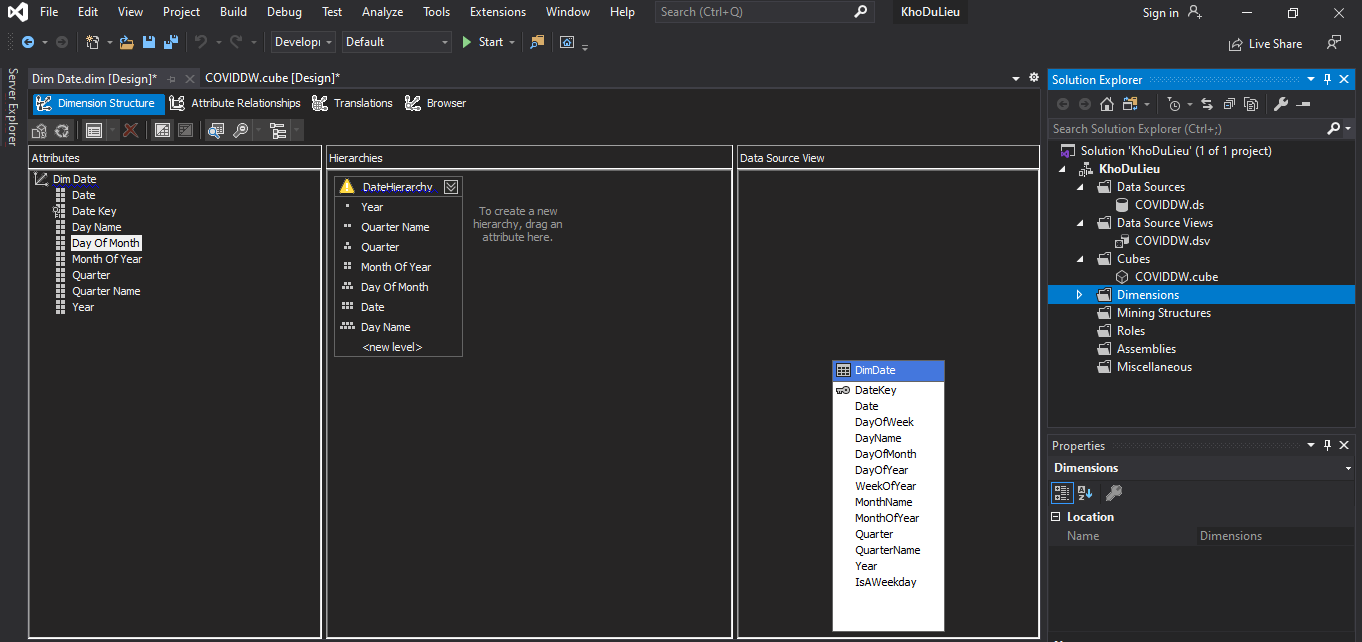




1. **Tạo data cube**

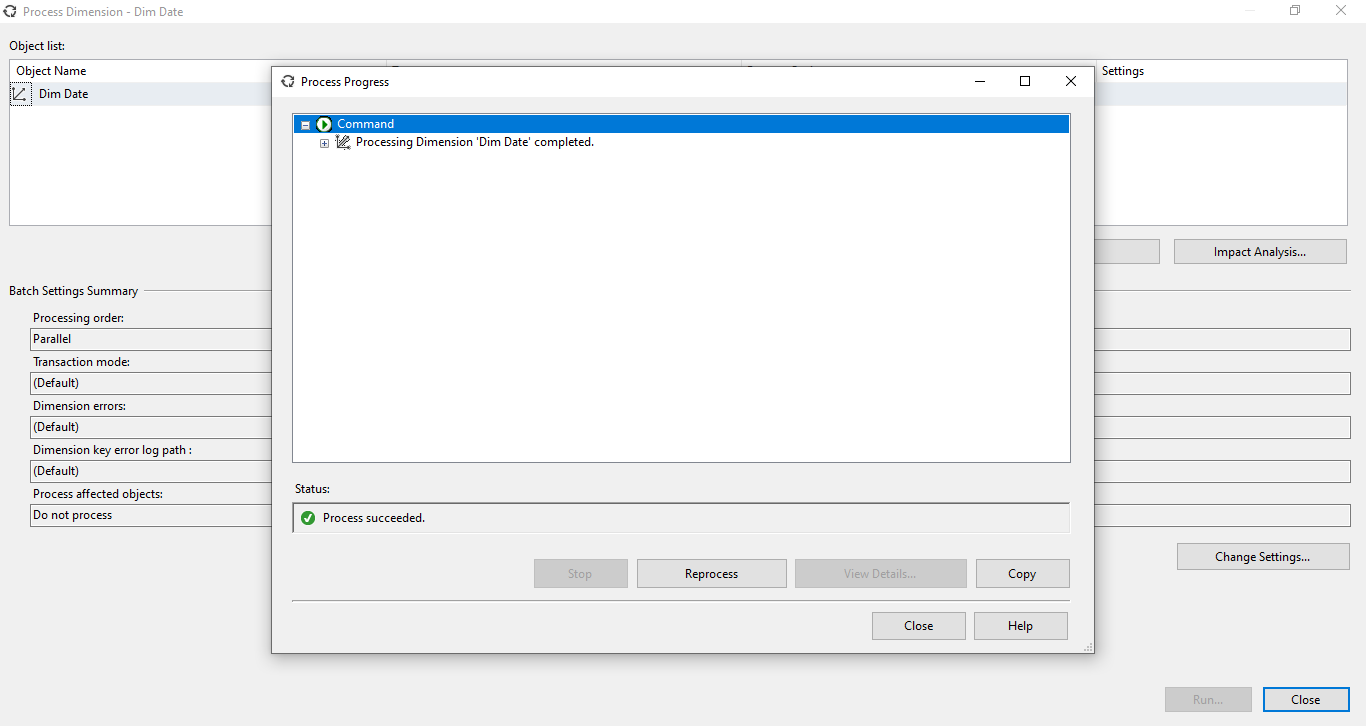


1. **Cấu hình các dimension**

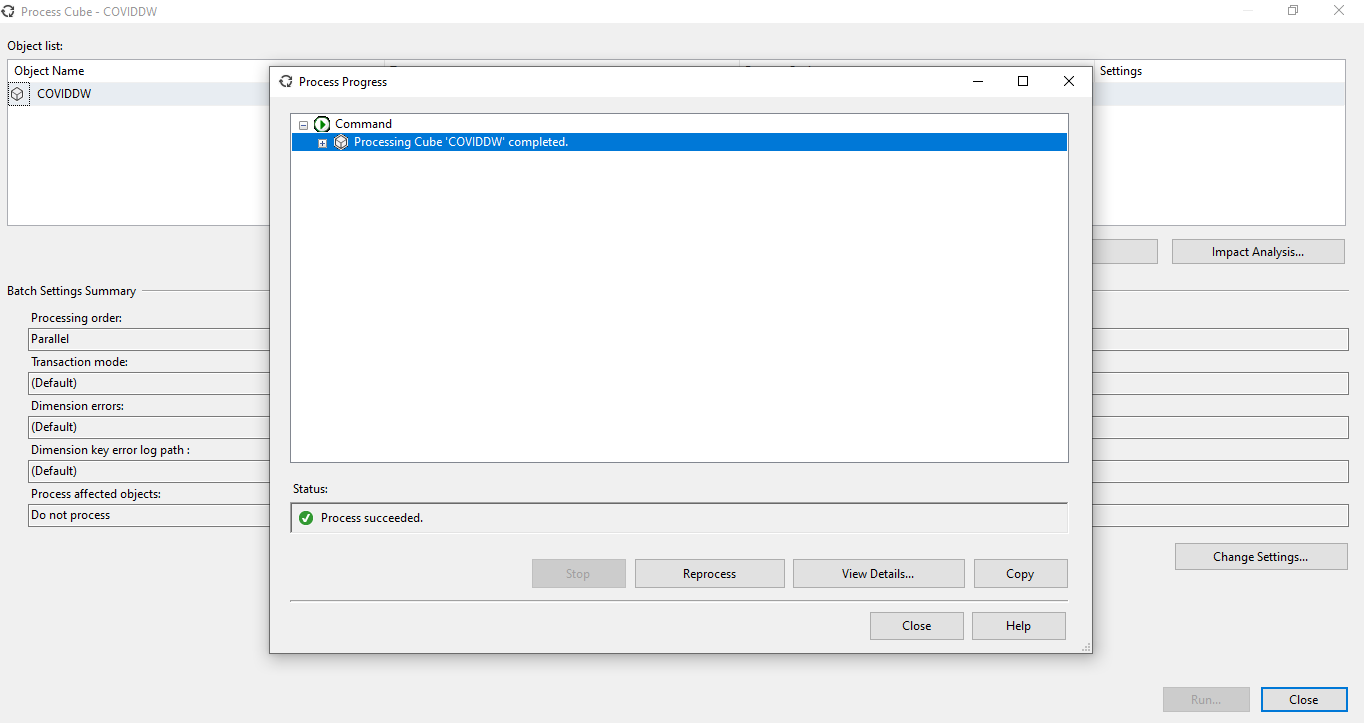


Click Process icon ( ) ở toolbar click Yes khi được hỏi để build và deploy project. Nếu build thành công, dialog sau sẽ xuất hiện.

Click Run để thực thi. Nếu thành công, Process Progress dialog sau sẽ xuất hiện.



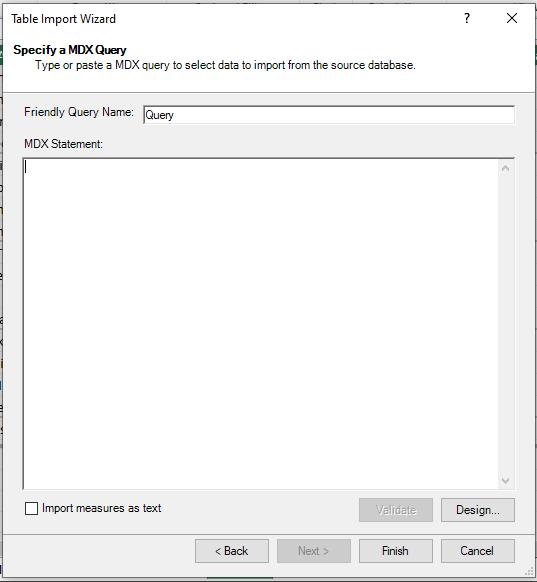
Sau khi đã cấu hình hết các dimension và process tất cả thì ta sẽ process cube và thực hiện các truy vấn phù hợp , nếu thành công thì sẽ có thông báo như sau :

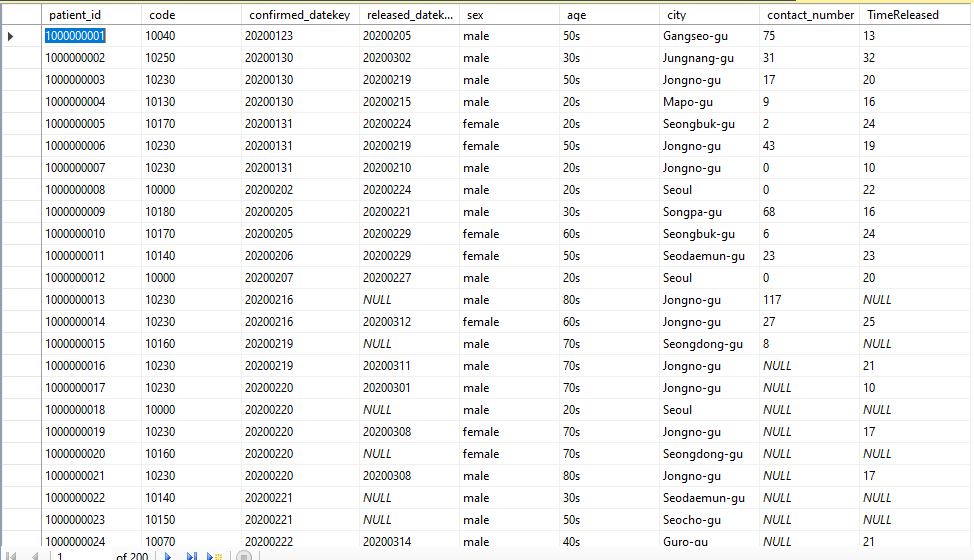


**Chương 3: Truy xuất và báo cáo**

* 1. **Xuất báo cáo**

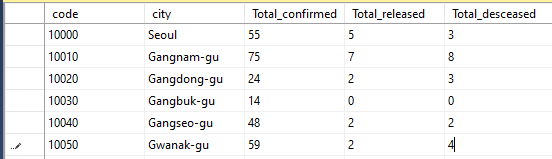
Khi đã có *cube* ta có thể tổng hợp, phân tích dữ liệu để cho ra các báo cáo hỗ trợ cho việc ra quyết định. Việc xuất báo cáo sẽ được thực hiện bảng *Power Pivot* của *Microsoft Excel*. Việc này được thực hiện theo các bước sau:

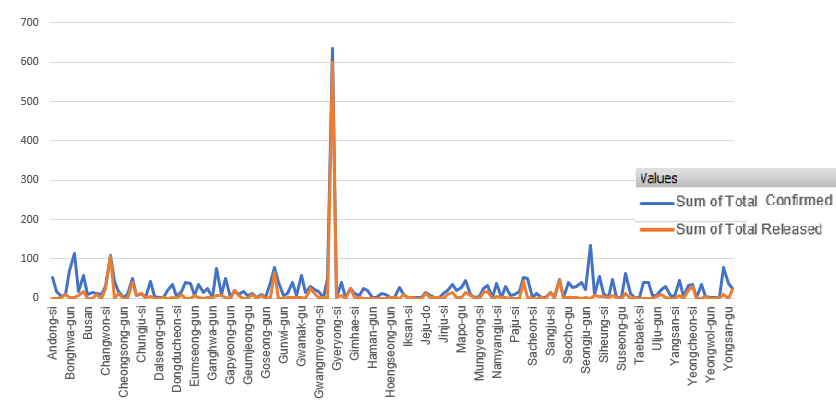
1. Kết nối với Analysis Service (cube): *Power Pivot> Management>Get Exter-nal Data > From Database > From Analysis Services or PowerPivot* Sau đóchọn và kết nối với cube vừa tạo.
2. Thiết kế câu truy vấn: Chọn Design và kéo thả các thuộc tính cần thiết vào phần bảng tính toán. Sau khi thiết kế theo ý muốn, chọn *OK > Finish* để hoàn thành câu truy vấn.
3. Thiết kế báo cáo: Sau khi thiết kế xong câu truy vấn ta có thể thiết kế báo cáo dự trên những dự liệu đã được tổng hợp trong câu truy vấn đó. Có thể tạo những biểu đồ dạng cột, hoặc dạng đường,… giúp tăng phân trực quan cho báo cáo.
   1. **Báo cáo**
      1. **Patient Information Reporting**



Từ bảng thống kê ta có thể biết được các thông của người bệnh như nơi ở, ngày nhiễm bệnh, ngày khỏi bệnh, giới tính, độ tuổi, sô người lây nhiễm, thời gian hồi phục. Từ đó ta có thể kiểm soát các bệnh nhân một cách dễ dàng và dựa vào thời gian hồi phục có thể tìm ra phương pháp chữa trị hợp lí.

* + 1. **Province**



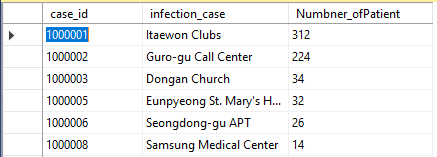


Với kết quả truy vấn này, chúng ra có thể theo dõi được số ca nhiễm bệnh, số người chết, số người khỏi bệnh của mỗi thành phố như thế nào. Từ đó có thể đưa ra những chính sách cách ly với những thành phố có tỉ lệ người nhiễm và chết cao, hoặc bỏ cách ly với những thành phố đã có tỉ lệ người khỏi bệnh cao.

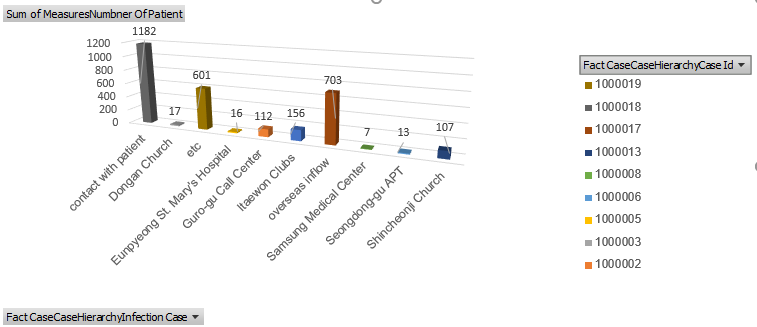
Ví dụ: ta thấy thành phố Bonghwa-gun có số người nhiễm bệnh trên 100 người nhưng số người hết bệnh thì chỉ là vài người, cho thấy đây là thành phố có tỉ lệ nhiễm bệnh cao cần được cách ly. Hoặc thành phố có người nhiễm cao là Gyerygong-si có số nhiễm cao nhất gần 650 người, nhưng cũng có tỉ lệ người khỏi bệnh gần 600 người tính tới tháng 5 năm 2020, cho thấy thành phố có phương pháp chữa trị bệnh tốt và không cần thiết cách ly nữa.

* + 1. **Infection Case**

Mục đích của bảng FactCase được tạo ra để trả lời cho câu hỏi đâu là những địa điểm có lượng bệnh nhân mắc covid\_19 đi qua nhiều nhất. Ta thực hiện truy vấn để tìm ra những nơi mà bệnh nhận đã mắc bệnh hay qua lại.



Kết quả sau truy vấn ta được như sau: Có thể thấy rõ được là số lượng người mắc bệnh đi qua từng vùng dịch. Ta thấy rằng vùng Iteawon Clubs có số lượng người mắc covid\_19 đi qua đây khá cao với 312 người từng đến đây.Số lượng này thay đổi theo từn khu vực, từng, ổ dịch: Guro-gu Call Center (224), contact with patient (1183), Dongan Church (34), overseas inflow (703),…

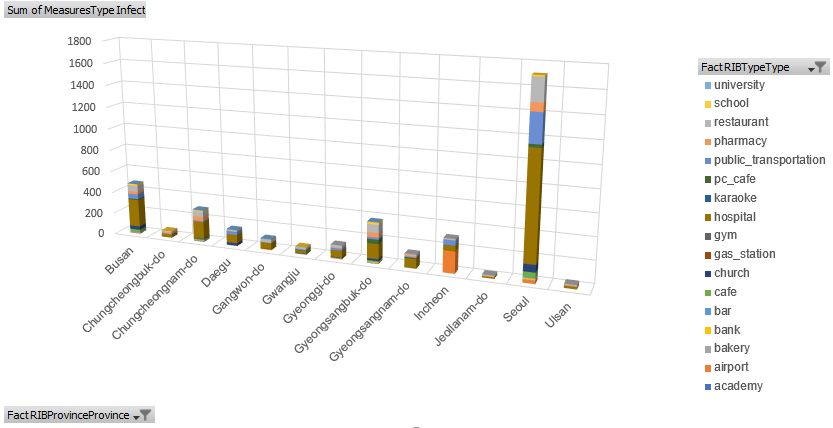


Dựa vào biểu đồ ta thấy được rằng: Những người mắc do tiếp xúc trực tiếp với bệnh nhân là có số lượng cao nhất với 1182 người, tiếp đến là những người từ nước ngoài trở về nước với 703 người, có 601 trường hợp là không xác đinh được nhiễm bệnh từ đâu.Còn lại những vùng dịch như: Iteawon Clubs, Shincheonji Church, Guro-gu Call Center,… có mức tương đương nhau.

Từ biểu đồ này thì nhà nước, chính phủ những cơ quan có quyền hạn sẽ biết được vùng dịch nào đang nằm trong điểm nóng, đang ở mức nguy hiểm nào để đưa ra được chính sách cách li, phong tỏa …hợp lệ.

* + 1. **Region Infected Analytics**

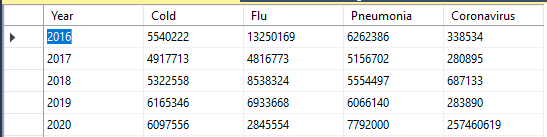
Mục đích của bảng FactCase được tạo ra để trả lời cho câu hỏi đâu là những địa điểm có lượng bệnh nhân mắc covid\_19 đi qua nhiều nhất. Ta thực hiện truy vấn để tìm ra những nơi mà bệnh nhận đã mắc bệnh hay qua lại



* + 1. **Search Trend**

Mục đích của bảng FactSearchTrend là để trả lời cho câu hỏi mức độ quan tâm tới dịch bệnh của cộng đồng hiện này.Ta sẽ thực hiện truy vấn trên bảng FactSearchTrend để thống kê ra được đâu là bệnh mà được người dân quan tâm đến nhất.

Dưới đây là kết quả sau khi ta thực hiện truy vấn:



Như kết quả sau khi thực hiện truy vấn, ta thấy:

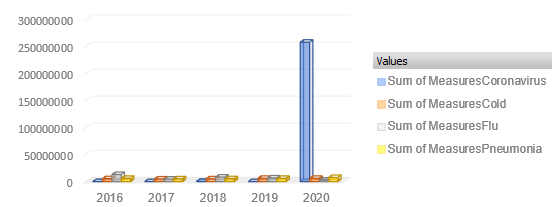
Trong năm 2016 thì từ khóa “Flu” là từ khóa được tìm kiếm nhiều nhất đối với cộng đồng người Korea với 13.250.169 lượt kiếm, còn từ khóa được tìm kiếm ít nhất đó chính là “coronavirus” chỉ với 338.534 lượt tìm kiếm, điều này hoàn toàn đúng vì dịch covid chỉ xuất hiện trong cuối năm 2019 nên lượt quan tâm của cộng đồng không đáng kể.Chứng tỏ rằng, ở thời điểm 2016 thì dịch bệnh Flu là dịch bệnh mà gây ảnh hưởng lớn nhất đối với đất nước Korea khi đó.

Sang năm 2017, nhìn chung thì các lượt tìm kiếm từ khóa “Flu”, “Cold”, “Pneumonia”, “Coronavirus” đều giảm xuống.Với lượt truy cập nhiều nhất là 5.156.702 với từ khóa “Pneumonia”.Lí do để giải thích cho lượt truy cập từ khóa dịch bệnh giảm có thể là do tình hình các dịch bệnh đều giảm, việc phòng chống ngăn chặn đều được thực hiện tốt nên tình hình các dịch bệnh không phải là cao điểm nữa, nên tâm lí người dân từ đó cũng thoải mái.Từ đó dấn đến việc quan tâm tới dịch không còn quá quan trọng nên lượt truy cập giảm xuống.

Sang đến năm 2018, thì lượt truy cập các từ khóa đều tăng và lượt được truy cập nhiều nhất đó là “Flu” với 8.538.324, chứng tỏ dịch bệnh này có thể đã bùng phát tại nước này.Việc tìm kiếm từ khóa để hiểu hơn về dịch bệnh cũng như cách phòng chống của người dân tăng mạnh.

Đến năm 2019, năm mà dịch covid\_19 bùng phát, tuy nhiên dịch bệnh bùng phát vào những ngày cuối năm 2019 và bắt đầu tại đất nước Korea là vào đầu năm 2020 nên từ khóa “Coronavirus” chưa thực sự được chú ý tới.Lượt tìm kiếm nhiều nhất vẫn là “Flu” với 6.33.668, có thể thấy được là dịch “Flu” xảy ra thường xuyên bùng phát tại Korea và mức độ quan tâm đến dịch bệnh này cũng được chú trọng đến.

Tiếp đến năm 2020, năm dịch covid\_19 bùng phát mạnh tại Korea.Ta có thể thế rõ được lượt truy cập từ khóa “coronavirus” với 257.460.619, gấp rất nhiều lần đối với những năm khác, đây có thể được xem là dịch bệnh lớn nhất và cũng khiến nhiều người chết nhất tại Korea, việc phòng chánh, tìm kiếm cách bảo vệ bản thân là điều nên bắt buộc với mỗi người, không chỉ vậy truyền thông Korea cũng đẩy mạnh tuyên truyền dịch bệnh, nên việc từ khóa “coronavirus” tăng đột biến là điều dễ hiểu.



Như biểu đồ trên cho ta thấy, màu xanh thể hiện cho dịch bệnh Corinavirus, màu cam thể hiện dịch bệnh Cold, màu trắng thể hiện cho dịch Flu và màu vàng thể hiện cho dịch Pneumonia. Trước năm 2020 thì mức độ tìm kiếm đều ở mức ngang tương đương nhau.Tuy nhiên, sang đến năm 2020, dịch Corinavirus tăng vọt, hơn rất nhiều so với những dịch bênh khác, những năm khác, mặc dù 2020 mới chỉ trôi qua được phân nửa nhưng số lượng tìm kiếm từ khóa virus corona đã gấp 200 bốn năm trước cộng lại.Chứng tỏ mức độ nguy hiểm và sự quan tâm đến dịch bệnh tăng cao , khi số lượng tìm kiếm gấp gần 5 lần dân số Hàn Quốc. Thống kê này sẽ có nhiều tác dụng ví dụ như khi biết được người dân thường xuyên tìm kiếm về virus corona thì chính phủ có thể đưa ra các thông báo , quyết sách , hỏa tốc trên mạng để người dân có thể nhanh chóng cập nhật được. Cũng như giúp các công ty có thể quảng cáo sản phẩm hỗ trợ dịch bệnh hiệu quả hơn .

KẾT LUẬN

1. **Những gì đạt được**

Sau quá trình thực hiện việc tạo kho dữ liệu cho cơ sở dữ liệu Covid -19 nhóm chúng tôi đã thực hiện các công việc như tìm kiếm cơ sở dữ liệu và chọn lọc dữ liệu về Covid-19 , xây dựng được các business process và bus matrix tương ứng, xây dựng các bảng fact dimension, thực hiện tích hợp dữ liệu, thiết kế data cube bằng SSAS, sử dụng Excel và SQL để tạo truy vấn và báo cáo biểu đồ. Kết quả cuối cùng việc đưa ra các business process tương ứng để giúp trả lời những câu hỏi đã đề ra.

1. **Những gì chưa đạt được**

Nhóm gặp khó khăn nhiều trong việc xử lý dữ liệu bởi vì các dữ liệu trong cơ sỡ dữ liệu Covid-19 gốc còn khá mới chưa được cập nhập đầy đủ các thông tin cần thiết để nhóm có thể thông kê và tận dụng một cách hiệu quả. Do thời gian cặp nhập theo từng ngày nên khi truy xuất dữ liệu quá lớn và chưa trực quan khi chuyển qua theo tuần hoặc tháng nên nhóm phải thống kê theo địa điểm như thành phố hoặc tỉnh để có phân tích dễ dàng hơn.

Trong tương lai nhóm sẽ cố gắng nghiên cứu và thu thập thêm nhiều dữ liệu hơn nữa để cải thiện chất lượng kho dữ liệu để truy vấn được rõ ràng và chính xác hơn.

**TÀI LIỆU THAM KHẢO**

<https://www.kaggle.com/kimjihoo/ds4c-what-is-this-dataset-detailed-description?fbclid=IwAR31UcGI57dqeCdKDP8PZssU3DswMwA-NPhFGgnEAK6ex0zidxRgLO7EZCA>

<http://classes.ischool.syr.edu/ist722fudge/old/units/09/Lab3-ETL-Development-with-SSIS.pdf?fbclid=IwAR2oa3Baplc95nCSq3kRZqBC_SCDMZ_3LMlvWMVdl7yjQ4JnZ86ZV9f6zBk>