**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu Covid 19 ở Hàn Quốc tham khảo sử dụng trong việc hoạch định chiến lược kiểm soát dịch Covid tại các Centers for Disease Control and Prevention tại các cấp**

Lớp: **181330A, 181330B, 181330C**

**NHÓM : 6**

THÀNH VIÊN: 02.Trần Hoàng An Bình [02/06]

**Ngày nộp: 15/06/2021**

**TP.HCM, ngày 15 Tháng 06 năm 2021**

**NHẬN XÉT CỦA GIẢNG VIÊN ĐỐI VỚI NHÓM:**

**………………………………………………………………..**

**………………………………………………………………..**

**……………………………………………………………….**

**NHÓM : 6**

**ĐIỂM ĐÁNH GIÁ**

1. Trần Hoàng An Bình [2] **:………….(…………………)………….**

**Ngày 15/06/2021**

**Giảng viên Ký tên**

THÀNH VIÊN:

2.Trần Hoàng An Bình [38/6]: Phân hệ phân tích thời gian khỏi bệnh trung bình của bệnh nhân theo thành phố (DimRegion, DimPatient, DimDATE,FactThoiGianTrungBinh)

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**ĐỀ TÀI ĐỒ ÁN HỌC PHẦN**

TÊN HỌC PHẦN: **Kho dữ liệu (DWH)**

MÃ SỐ LỚP HP: **DAWH430784\_06**

Tên đề tài Nhóm: **Phát triển mô hình kho dữ liệu Covid 19 ở Hàn Quốc tham khảo sử dụng trong việc hoạch định chiến lược kiểm soát dịch Covid tại các Centers for Disease Control and Prevention tại các cấp**

**Giảng viên giảng dạy: VÕ XUÂN THỂ \_ Giảng viên chính**

Lớp: **181330A, 181330B, 181330C (Sáng thứ 4)**

**Tên sản phẩm đề tài: 02N6.02.TranHoangAnBinh.DAHP.DWH.rar** (Microsoft SQL Server 2019)

**DB Nguồn: Covid19\_Original**

**DB DWH: Covid19\_final**

**Tên máy chủ DB SQL: DESKTOP-2984JF9\BI**

**Tên đăng nhập DB SQL:**

**Mật khẩu (password) đăng nhập SQL:**

**ProjectName BIDS: 02TranHoangAnBinhTGTB\_DWH (Microsoft Visual Studio 2019)**

**NHÓM : 06**

THÀNH VIÊN: 2.Tần Hoàng An Bình [02];

**TP.HCM, ngày 15 Tháng 06 năm 2021**

# LỜI CẢM ƠN

Để hoàn thành báo cáo môn học này trước tiên em xin gửi đến các quý thầy giáo, cô giáo khoa Công nghệ Thông tin trường Đại Học Sư Phạm Kỹ Thuật TP.HCM lời cảm ơn chân thành và sâu sắc nhất. Đặc biệt, em xin gửi đến thầy Võ Xuân Thể người đã tận tình hướng dẫn, giúp đỡ chúng em hoàn thành bài báo cáo này. Thầy đã cung cấp rất đầy đủ những thông tin, nguồn tài liệu chính xác để chúng em có thể tìm hiểu và vận dụng tốt nhất những gì đã học.

Vì hạn chế trong thời gian học lẫn thời gian làm báo cáo nên không thể tránh được những thiếu sót. Em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các quý thầy cô để em có điều kiện bổ sung, nâng cao về kiến thức, ý thức về nghiệp vụ của mình, để có thêm kinh nghiệm cho việc ra đời sau này.

# DANH MỤC CÁC TỪ VIẾT TẮT

|  |  |
| --- | --- |
| CSDL hoặc DB | Cơ sở dữ liệu: DataBase |
| LĐQH | Lược đồ quan hệ |
| PTH | Phụ thuộc hàm = Ràng buộc tham chiếu: Reference Constraint |
| PTTK | Phân tích và thiết kế |
| RBTV | Ràng buộc toàn vẹn [dữ liệu] |
|  |  |
| BIDS | Công cụ hỗ trợ tổ chức, quản lý và khai thác kho dữ liệu (Hệ thống tổ chức dữ liệu thông minh cho các doanh nghiệp): Business Intelligence Development Studio của Microsoft |
| DWH | Kho dữ liệu (Nhà kho dữ liệu): DataWareHouse |
| ERD | Sơ đồ thực thể kết hợp: Entity Relationship Diagram |
| MS | Công ty Microsoft |
| NF | Dạng chuẩn của CSDL: Normal Form |
| SQL | Ngôn ngữ vấn tin có cấu trúc: Structured Query Language |

# DANH MỤC CÁC THUẬT NGỮ

|  |  |
| --- | --- |
| Giao diện người dùng  (User-Interface) | Là hệ thống các màn hình giao tiếp cho phép người sử dụng tương tác với các thành phần phần mềm, điều khiển phần mềm hoạt động theo yêu cầu của người dùng - tương ứng các chức năng hiện có của phần mềm. |
| Người dùng (User):  Tài khoản (Account) | Là một quyền làm việc trên hệ thống phần mềm được cấp phát cho một cá nhân thông qua tên tài khoản (username) và mật khẩu (password). |
|  |  |
| BigData | Dữ liệu lớn: là một tập hợp dữ liệu rất lớn và phức tạp, không thể xử lý dữ liệu bằng các phương pháp truyền thống. |

# MỤC LỤC

[LỜI CẢM ƠN 4](#_Toc73666776)

[DANH MỤC CÁC TỪ VIẾT TẮT 5](#_Toc73666777)

[DANH MỤC CÁC THUẬT NGỮ 6](#_Toc73666778)

[MỤC LỤC 7](#_Toc73666779)

[Chương 1. GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN 10](#_Toc73666780)

[1.1. Tổng quan về ĐỒ ÁN HỌC PHẦN 10](#_Toc73666781)

[1.2. Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN 10](#_Toc73666782)

[1.3. Mục tiêu của ĐỒ ÁN HỌC PHẦN 10](#_Toc73666783)

[1.4. Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN 10](#_Toc73666784)

[1.5. Sản phẩm của ĐỒ ÁN HỌC PHẦN 10](#_Toc73666785)

[1.6. Bố cục của báo cáo 10](#_Toc73666786)

[Chương 2. CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN 11](#_Toc73666787)

[2.1. Tổng quan về kho dữ liệu (DWH) 11](#_Toc73666788)

[2.1.1. Các khái niệm cơ bản về Kho dữ liệu 11](#_Toc73666789)

[2.1.2. Kiến trúc cơ bản của kho dữ liệu: 11](#_Toc73666790)

[2.1.3. Mô hình CSDL đa chiều (Dimensional Modeling) 11](#_Toc73666791)

[2.2. Phân tích và thiết kế Kho dữ liệu 11](#_Toc73666792)

[2.2.1. Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling) 11](#_Toc73666793)

[2.2.2. Các dạng (phương pháp) tích hợp dữ liệu 11](#_Toc73666794)

[2.2.3. Các bước cơ bản trong tiến trình tích hợp dữ liệu 11](#_Toc73666795)

[2.2.4. Các công cụ tích hợp dữ liệu: 11](#_Toc73666796)

[2.2.5. Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows) 11](#_Toc73666797)

[2.2.6. Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến) 11](#_Toc73666798)

[2.3. Vấn tin khai thác sử dụng Kho dữ liệu 11](#_Toc73666799)

[2.3.1. Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu 11](#_Toc73666800)

[2.3.2. Vấn tin với các hàm phân tích của SQL trên kho dữ liệu 11](#_Toc73666801)

[2.3.3. Vấn tin dùng Materialized View để phân tích trên kho dữ liệu 11](#_Toc73666802)

[2.3.4. Hệ thống OLAP 11](#_Toc73666803)

[Chương 3. TỔNG QUAN VỀ KHO DỮ LIỆU …<theo tên đề tài>…… TẠI ……<theo tên đề tài>………… 12](#_Toc73666804)

[3.1. Ý tưởng hình thành DWH …<theo tên đề tài>…… tại ……<theo tên đề tài>………… 12](#_Toc73666805)

[3.2. Giới thiệu chung về Kho dữ liệu 12](#_Toc73666806)

[3.2.1. Mô tả tổng quan về DWH 12](#_Toc73666807)

[3.2.2. Các dữ liệu nguồn hình thành DWH 12](#_Toc73666808)

[3.3. Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH 12](#_Toc73666809)

[3.4. Các thành phần phân hệ Kho dữ liệu phân công các cá nhân phụ trách 12](#_Toc73666810)

[Chương 4. PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC PHÂN HỆ KHO DỮ LIỆU …<theo tên đề tài>…… TẠI ……<theo tên đề tài>………… 13](#_Toc73666811)

[4.1. <STT>\_ Ho Tên 1\_ <thành phần kho dữ liệu đươc phân công> 13](#_Toc73666812)

[4.1.1. Mô tả tổng quan về phân hệ DWH…<tên phân hệ DWH>… 13](#_Toc73666813)

[4.1.2. Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu 13](#_Toc73666814)

[4.1.3. Xác định ý tưởng hình thành phân hệ của Kho dữ liệu 13](#_Toc73666815)

[4.1.4. Mô tả hệ thống các DB gốc liên quan phân hệ DWH 13](#_Toc73666816)

[4.1.5. Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ 13](#_Toc73666817)

[4.1.6. Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ 14](#_Toc73666818)

[4.1.7. Thực hiện các thủ tục tích hợp các DB gốc vào phân hệ DWH 15](#_Toc73666819)

[4.1.8. Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH 15](#_Toc73666820)

[4.1.9. Triển khai phân hệ DWH lên BIDS 15](#_Toc73666821)

[4.1.10. Khai thác sử dụng phân hệ DWH dùng các vân tin SQL 15](#_Toc73666822)

[4.2. <STT>\_ Ho Tên 2\_ <thành phần kho dữ liệu đươc phân công> 15](#_Toc73666823)

[Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN 16](#_Toc73666824)

[5.1. Kết luận 16](#_Toc73666825)

[5.1.1. Những kết quả đạt được 16](#_Toc73666826)

[5.1.2. Hạn chế 16](#_Toc73666827)

[5.2. Hướng phát triển 16](#_Toc73666828)

[5.2.1. Hướng khắc phục các hạn chế 16](#_Toc73666829)

[5.2.2. Hướng mở rộng ĐỒ ÁN HỌC PHẦN 16](#_Toc73666830)

[DANH MỤC TÀI LIỆU THAM KHẢO 17](#_Toc73666831)

[CÁC PHỤ LỤC 18](#_Toc73666832)

[ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN 19](#_Toc73666833)

# GIỚI THIỆU VỀ ĐỒ ÁN HỌC PHẦN

## Tổng quan về ĐỒ ÁN HỌC PHẦN

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.

## Nội dung chuyên môn chính của ĐỒ ÁN HỌC PHẦN

Trích xuất dữ liệu

Làm sạch dữ liệu

Chuyển đổi dữ liệu sang dạng kho

Phân tích, sắp xếp, dữ liệu theo cấu trúc và đưa ra quyết định

## Mục tiêu của ĐỒ ÁN HỌC PHẦN

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, địa phương 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

* Phân tích và đưa ra hướng giải quyết

## Công cụ và nền tảng kỹ thuật thực hiện ĐỒ ÁN HỌC PHẦN

Các công cụ chính sử dụng trong đồ án này là

+ Microsoft SQL Server 2019

+ SQL Server Business Intelligence Development Studio (BIDS Visual Studio 2019)

## Sản phẩm của ĐỒ ÁN HỌC PHẦN

Sản phẩm thu được là dữ liệu DWH đã được phân tích giúp trả lời những câu hỏi đề ra chẳng hạn: Phần lớn số người nhiễm covid ở Hàn Quốc thuộc tỉnh nào, những người có bệnh nền tim và tiểu đường thì ai có nguy cơ tử vong cao hơn, và với tốc độ lây nhiễm này các bệnh viện có đáp ứng đủ giường bệnh cho bệnh nhân không ?

## Bố cục của báo cáo

Báo cáo gồm những nội dung như sau:

Chương 1: Giới thiệu đồ án của học phần DWH

Chương 2: Các cơ sở lý thuyết của HP phục việc thực hiện đề tài HP

Chương 3: Giới thiệu tổng quan về Kho dữ liệu (DWH) Kho dữ liệu Covid 19 tại Hàn Quốc

Chương 4: Phân tích, thiết kế và tổ chức khai thác sử dụng từng phân hệ của Kho dữ liệu

Chương 5: Tổng kết các kết quản đạt được và còn hạn chế của đồ án, đồng thời đề xuất hướng khắc phục hạn chế và phát triển Đồ án

# CƠ SỞ LÝ THUYẾT CỦA HỌC PHẦN

## Tổng quan về kho dữ liệu (DWH)

### Các khái niệm cơ bản về Kho dữ liệu

#### Kho dữ liệu (DWH)?

Data Warehouse tạm dịch là kho dữ liệu thường được viết tắt là DW hay DWH.

- Về cơ bản có thể hiểu DW là một tập hợp các dữ liệu, thông tin có chung một chủ đề, được tổng hợp từ nhiều nguồn khác nhau trong nhiều mốc thời gian và không chỉnh sửa. Được dùng cho việc hỗ trợ ra quyết định, phân tích dữ liệu và lập báo cáo trong công tác quản lý.

- Hiện nay khái niệm kho dữ liệu được phát triển rộng hơn, nó mô tả tập hợp các công nghệ, phương pháp, kỹ thuật có thể kết hợp với nhau thực hiện các chức năng tích hợp, lưu trữ, xử lý và phân tích dữ liệu để cung cấp thông tin cho người sử dụng. Một kho dữ liệu thường có dung lượng lên đến hàng trăm GB thậm chí tính bằng đơn vị TB.

#### Nguyên lý hình thành DWH

- Subject Oriented (Hướng chủ đề): DW được xây dựng nhằm đáp ứng yêu cầu phân tích ở các cấp độ khác nhau của tiến trình ra quyết định, không tập trung vào xử lý giao dịch/hoạt động hàng ngày.

> Cung cấp một khung nhìn đơn giản, súc tích về các chủ đề cần thiết như khách hàng, sản phẩm, bán hàng,.. loại bỏ các dữ liệu không hữu ích trong tiến trình ra quyết định.

- Integrated (Tính toàn vẹn): Tích hợp dữ liệu từ nhiều nguồn khác nhau (bất kể sai khác về trường/ý nghĩa/định dạng dữ liệu) vào một định dạng thống nhất.

- Nonvolatile (Tính bất biến): Dữ liệu phải thống nhất theo thời gian (Hạn chế tối đa sửa đổi/xoá dữ liệu) >> phân tích sự thay đổi theo thời gian.

- Time-Varying (Giá trị lịch sử): Cung cấp dữ liệu tại các thời điểm khác nhau của môt thông tin và thời điểm thay đổi

> Kết hợp dữ liệu sự kiện/giao dịch (fact) và dữ liệu tham chiếu (reference) chứa giá trị đại diện thời gian của sự kiện.

> Thích ứng những thay đổi trong tương lai, báo cáo chính xác lịch sử thay đổi.

#### Các đặc điểm cơ bản của DWH

#### Phân biệt [SS] DWH với CSDL truyền thống

### Kiến trúc cơ bản của kho dữ liệu:

#### Tầng đáy: nạp dữ liệu

#### Tầng giữa: OLAP (OLAP server): xử lý DWH

#### Tầng trên cùng: ứng dụng = khai thác sử dụng DWH

### Mô hình CSDL đa chiều (Dimensional Modeling)

## Phân tích và thiết kế Kho dữ liệu

### Phân tích và thiết kế kho dữ liệu theo hướng mô hình hóa nhiều chiều (Dimensional Modeling)

### Các dạng (phương pháp) tích hợp dữ liệu

### Các bước cơ bản trong tiến trình tích hợp dữ liệu

### Các công cụ tích hợp dữ liệu:

### Lập các luồng công việc tích hợp dữ liệu (Data Integration Workflows)

### Hệ thống OLTP (On-Line Transaction Processing = Xử lý giao dịch trực tuyến)

## Vấn tin khai thác sử dụng Kho dữ liệu

### Vấn tin với các phép toán tổng hợp của SQL trên kho dữ liệu

### Vấn tin với các hàm phân tích của SQL trên kho dữ liệu

### Vấn tin dùng Materialized View để phân tích trên kho dữ liệu

### Hệ thống OLAP

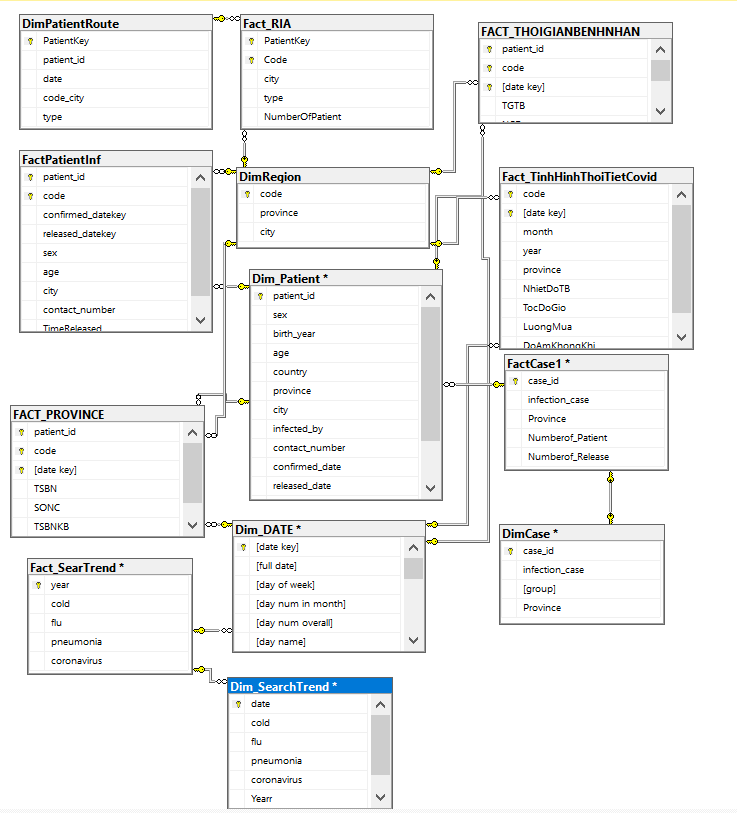
# TỔNG QUAN VỀ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU COVID 19 Ở HÀN QUỐC THAM KHẢO SỬ DỤNG TRONG VIỆC HOẠCH ĐỊNH CHIẾN LƯỢC KIỂM SOÁT DỊCH COVID TẠI CÁC CENTERS FOR DISEASE CONTROL AND PREVENTION TẠI CÁC CẤP

## Ý tưởng hình thành DWH Phát triển mô hình kho dữ liệu Covid 19 ở Hàn Quốc tham khảo sử dụng trong việc hoạch định chiến lược kiểm soát dịch Covid tại các Centers for Disease Control and Prevention tại các cấp

* Đầu tiên chúng em tham khảo các tập dữ liệu từ kaggle sau đó chúng em chọn dữ liệu covid19 để hình thành dwh. Sau khi đã chọn tập dữ liệu chúng em tiến hành download các file csv về import vào SQL server để hình thành database gốc sau khi đã có database gốc chúng em bắt đầu xây dựng 1 dataware house chung cho cả nhóm sau đó tiến hành phân chia công việc cho từng cá nhân (mỗi người 1 fact)

## Giới thiệu chung về Kho dữ liệu

### Mô tả tổng quan về DWH



Gồm 7 fact, 6 dim

Fact RIA: Thống kê tỉ lệ người nhiễm bệnh đã đi qua khu vực đó

FactPatientInf: Thống kê số lượng tiếp xúc và thời gian trung bình khỏi bệnh trong 1 khu vực

FACT\_THOIGIANBENHNHAN: thống kê thời gian trung bình khỏi bệnh của bệnh nhân

Fact\_TinhHinhThoiTietCovid: Thống kê về thời tiết có ảnh hưởng đến khả năng nhiễm bệnh

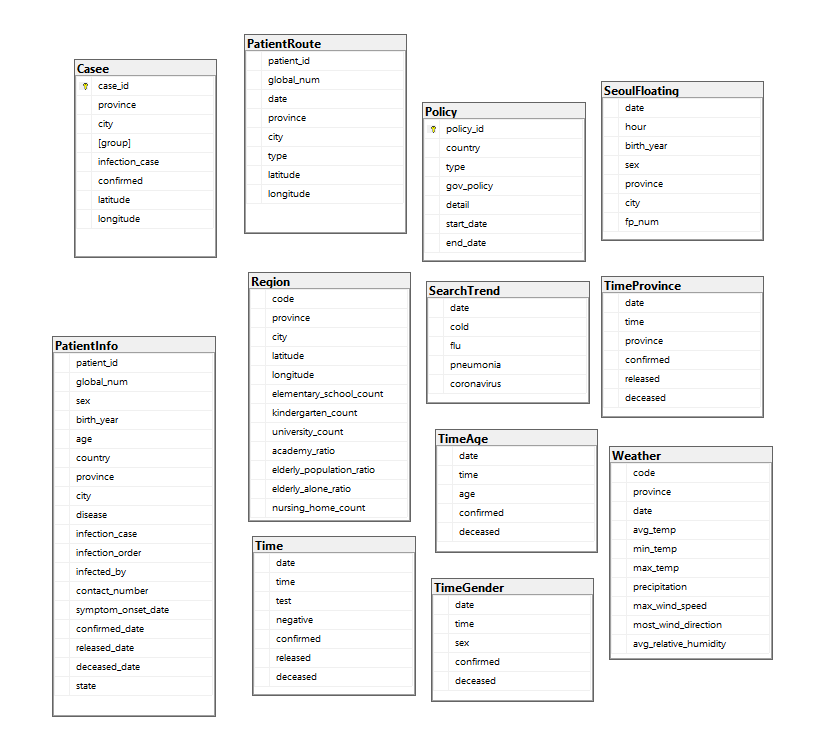
FACT\_PROVINCE: Thống kê số ca mắc số ca tử vong số ca khỏi bệnh theo province

Fact\_Case: Thống kê số lượng bệnh nhân và số bệnh nhân khỏi bệnh trong một ổ dịch

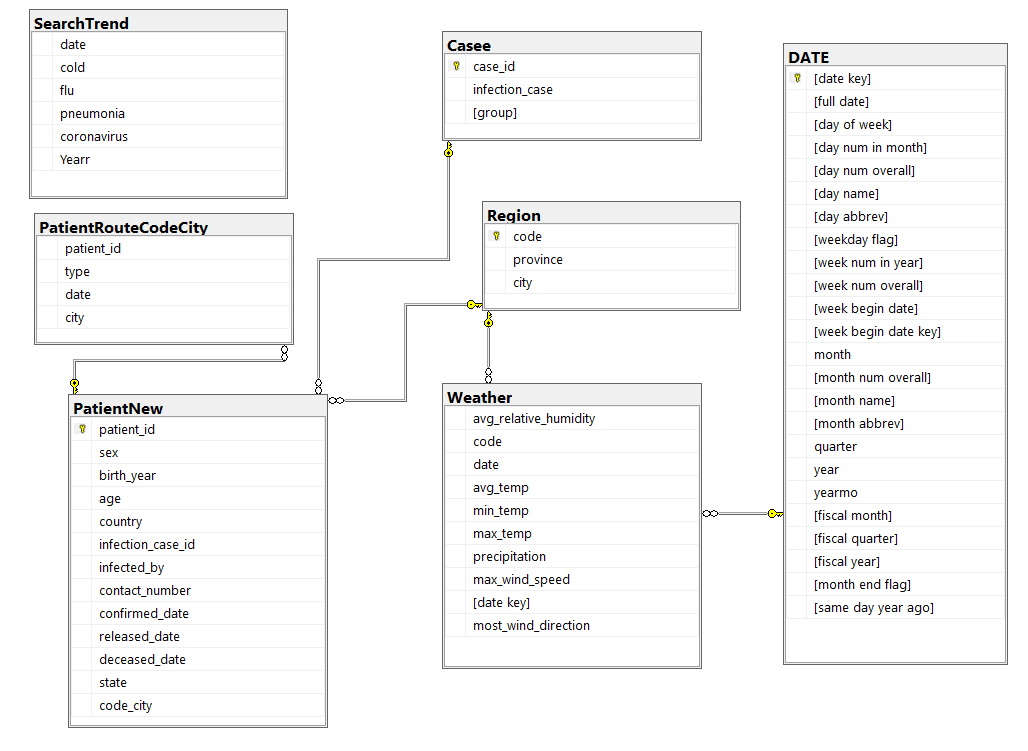
Fact\_searTrend: Thống kê lượt tìm kiếm tên từng loại bệnh theo từng năm và tháng

### Các dữ liệu nguồn hình thành DWH

* Dữ liệu từ file csv trên kaggle



* Database gốc:



## Xác định nhu cần tổ chức và phân tích dữ liệu của toàn bộ DWH

Nhu cầu của tổ chức:

+ Thống kê thời gian trung bình khỏi bệnh của bệnh nhân

+ Thống kê số lượng tiếp xúc và thời gian trung bình khỏi bệnh của bệnh nhân trong 1 khu vực

+ Thống kê thời gian trung bình khỏi bệnh của bệnh nhân

+ Thống kê thời tiết có ảnh hưởng đến khả năng nhiễm bệnh không

+ Thống kê tình hình bệnh nhân

+ Thống kê số lượng bệnh nhân và số bệnh nhân khỏi bệnh trong một ổ dịch

+ Thống kê lượt tìm kiếm tên từng loại bệnh theo từng năm và tháng (mức độ quan tâm của người bệnh đến dịch bệnh)

## Các thành phần phân hệ Kho dữ liệu phân công các cá nhân phụ trách

<02>\_ Trần Hoàng An Bình \_ Phân tích thống kê thời gian trung bình khỏi bệnh của bệnh nhân

# PHÂN TÍCH, THIẾT KẾ VÀ KHAI THÁC SỬ DỤNG CÁC PHÂN HỆ KHO DỮ LIỆU PHÁT TRIỂN MÔ HÌNH KHO DỮ LIỆU COVID 19 Ở HÀN QUỐC THAM KHẢO SỬ DỤNG TRONG VIỆC HOẠCH ĐỊNH CHIẾN LƯỢC KIỂM SOÁT DỊCH COVID TẠI CÁC CENTERS FOR DISEASE CONTROL AND PREVENTION TẠI CÁC CẤP

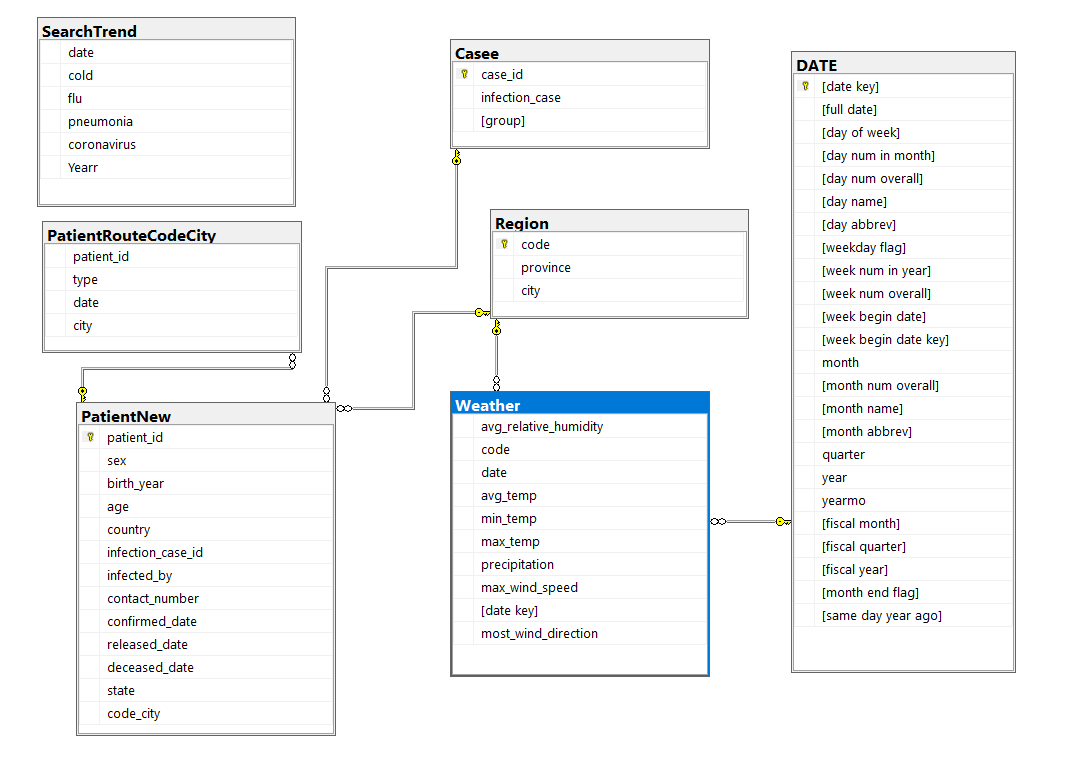
## 02\_ Trần Hoàng An Bình\_ Phân hệ phân tích thời gian khỏi bệnh của bệnh nhân theo vùng

### Mô tả tổng quan về phân hệ DWH phân tích thống kê số lượng bệnh nhân và người khỏi bệnh trong một ổ dịch

Trong phân hệ cá nhân này em sẽ phân tích thống kê số thời gian khỏi bệnh của một bệnh nhân để đánh giá sức để kháng và chất lượng phục vụ y tế theo từng vùng. Từ đó có kế hoạch cải thiện chất lượng y tế và sức khỏe của mọi người.

Phân hệ gồm Dim\_Region, DimPatient và DimDATE, FACTTHOIGIANBENHNHAN

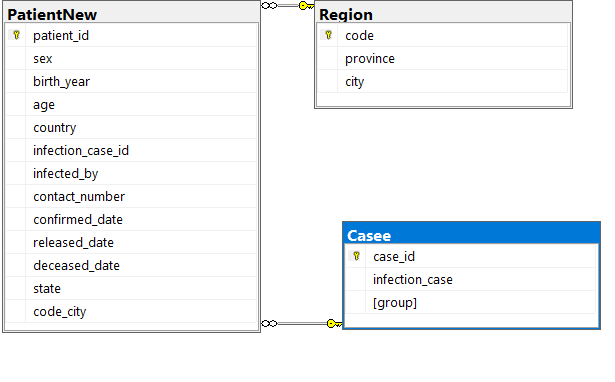
### Xác định nguồn dữ liệu tích hợp phân hệ Kho dữ liệu



### Xác định ý tưởng hình thành phân hệ của Kho dữ liệu

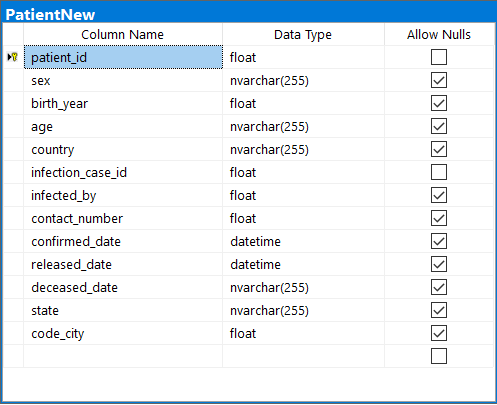
Dựa vào các bảng PatientNew là thông tin người bệnh và Region em bắt đầu hình thành ý tưởng DWH về phân tích thống thời gian trung bình khỏi bệnh của bệnh nhân, số bệnh nhân và số người khỏi bệnh theo từng vùng để đưa ra đánh giá về công tác y tế ở vùng đó và đưa ra các biện pháp phòng dịch hợp lý. Em sẽ hình thành các dim và các fact dựa trên 2 bảng này từ đó sẽ có những view thống kê sơ bộ

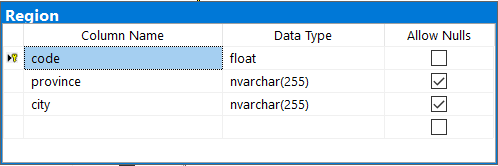
### Mô tả hệ thống các DB gốc liên quan phân hệ DWH

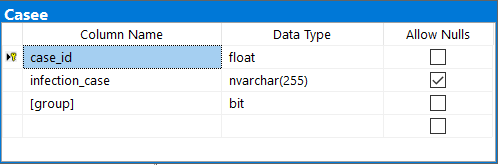


Mô tả DB:

* PatientNew(patient\_id, sex, birth\_year, age, country, infection\_case\_id, infection\_by, contact\_number, confirmed\_date, released\_date, deceased\_date, state, code\_city): thông tin bệnh nhân
* Region(code, province, city): vùng
* DATE(date key): ngày







### Phân phân tích các DB gốc xác định yêu cầu phân tích DWH của phân hệ

Dim:

DimRegion: Dim về các vùng trong nước gồm mã thành phố tên phành phố và tên tỉnh

DimPatient: Dim về thông tin của bệnh nhân như id bệnh nhân ngày mắc ngảy khỏi nơi ở,…

Fact:

FactCase: Thống kê số thời gian rung bình khỏi bệnh của bện nhân theo từng Vùng

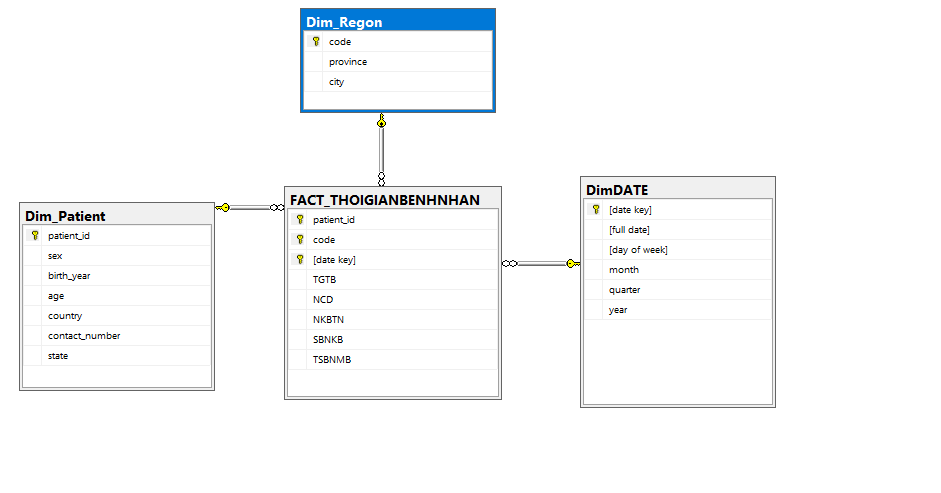
TGTB: Là thời gian khỏi bệnh Trung bình của bệnh nhân (AVG)

NCD:Là thời gian đầu tiên có người mắc bênh( MIN)

SBNKB: Tổng số bệnh nhân đã khỏi bệnh(count )

TSBNMB:Tổng số bệnh nhân đã khỏi và chưa khỏi (count)

### Thiết kế DB mới tổ chức phân tích Kho dữ liệu cho phân hệ

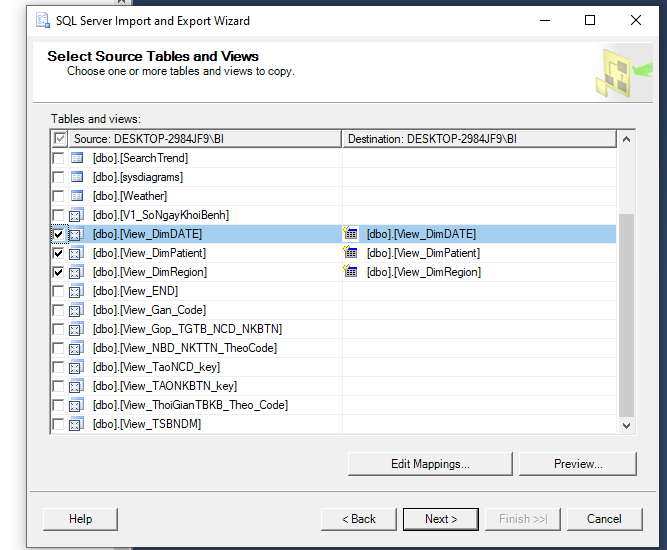


### Thực hiện các thủ tục tích hợp các DB gốc vào phân hệ DWH

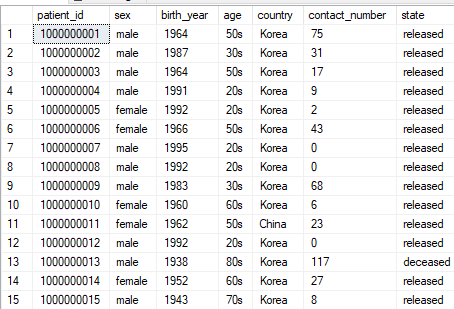
+ Copy DataBase cần thiết (để lập bản sao phục vụ phân tích) từ DB gốc sang DB của DHW phân tích. Chú ý: Để Copy DB phải kích hoạch SQL Server Agent (Local System)

+ Sử dụng import data import các dữ liệu từ db gốc vào dwh

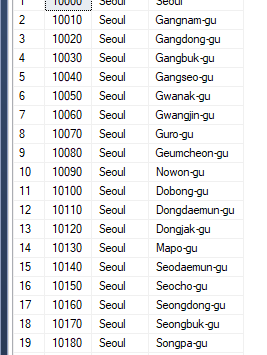
* Nạp dữ liệu vào bảng DimPatient, DimDate, DimRegion:



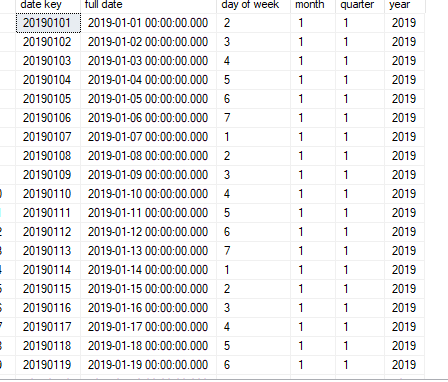
Dữ liệu sau khi them vào DimPatient



Dữ liệu sau khi them vào DimDate



Dữ liệu vào Dimdate



Lậo

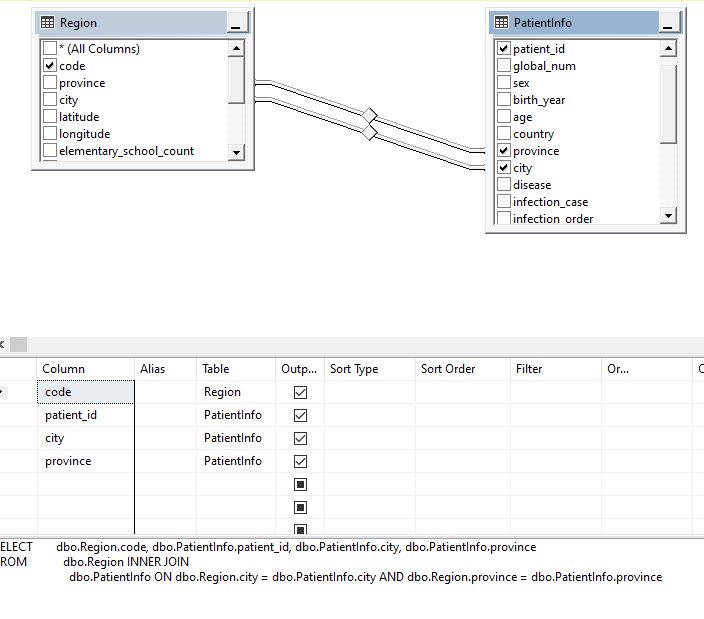
### Thiết lập các Views tính toán cần thiết để nạp dữ liệu từ DB gốc vào các Factors của phân hệ DWH

**V1\_Code\_TinhThanh** :Tạo Vùng Miền theo Code

SELECT dbo.Region.code, dbo.PatientInfo.patient\_id, dbo.PatientInfo.city, dbo.PatientInfo.province

FROM dbo.Region INNER JOIN

dbo.PatientInfo ON dbo.Region.city = dbo.PatientInfo.city AND dbo.Region.province = dbo.PatientInfo.province



**V1\_SoNgayKhoiBenh** :Tính số ngày khỏi bệnh theo vùng miền, tỉnh

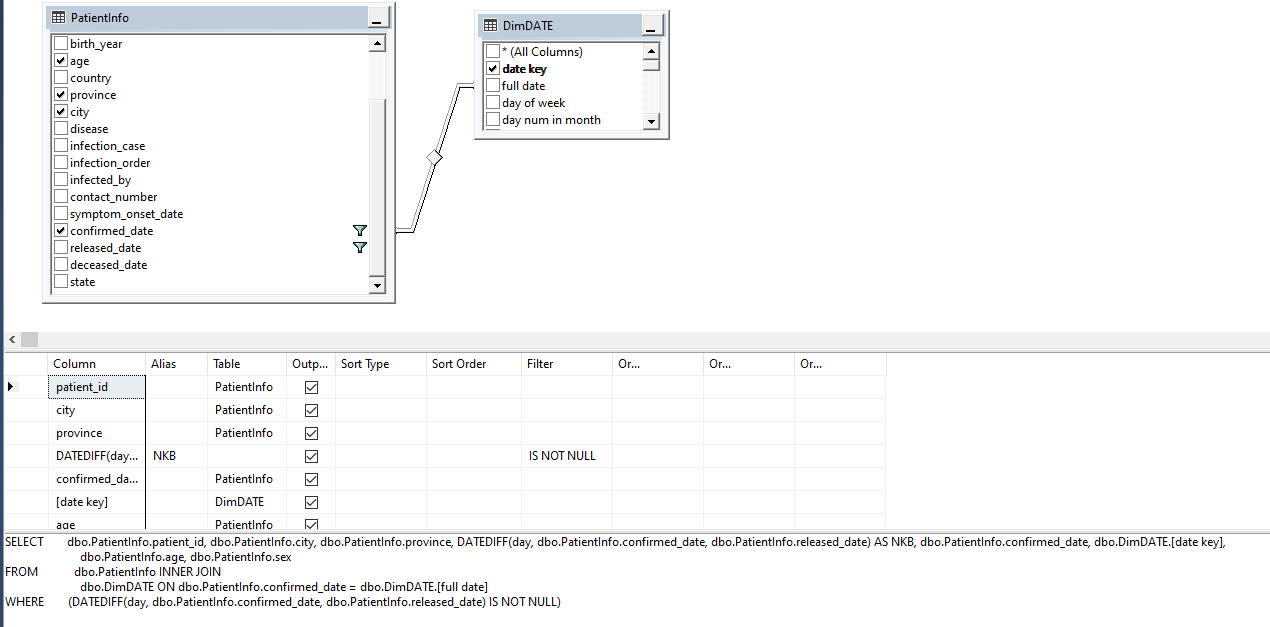
SELECT dbo.PatientInfo.patient\_id, dbo.PatientInfo.city, dbo.PatientInfo.province, DATEDIFF(day, dbo.PatientInfo.confirmed\_date, dbo.PatientInfo.released\_date) AS NKB, dbo.PatientInfo.confirmed\_date, dbo.DimDATE.[date key],

dbo.PatientInfo.age, dbo.PatientInfo.sex

FROM dbo.PatientInfo INNER JOIN

dbo.DimDATE ON dbo.PatientInfo.confirmed\_date = dbo.DimDATE.[full date]

WHERE (DATEDIFF(day, dbo.PatientInfo.confirmed\_date, dbo.PatientInfo.released\_date) IS NOT NULL)



**View\_END** : View thời gian trung bình khỏi bệnh, số bệnh nhân, số bệnh nhân khỏi bệnh ngày bắt đầu dịch, ngày gần nhất có bệnh nhân khỏi bệnh theo mỗi bệnh nhân.

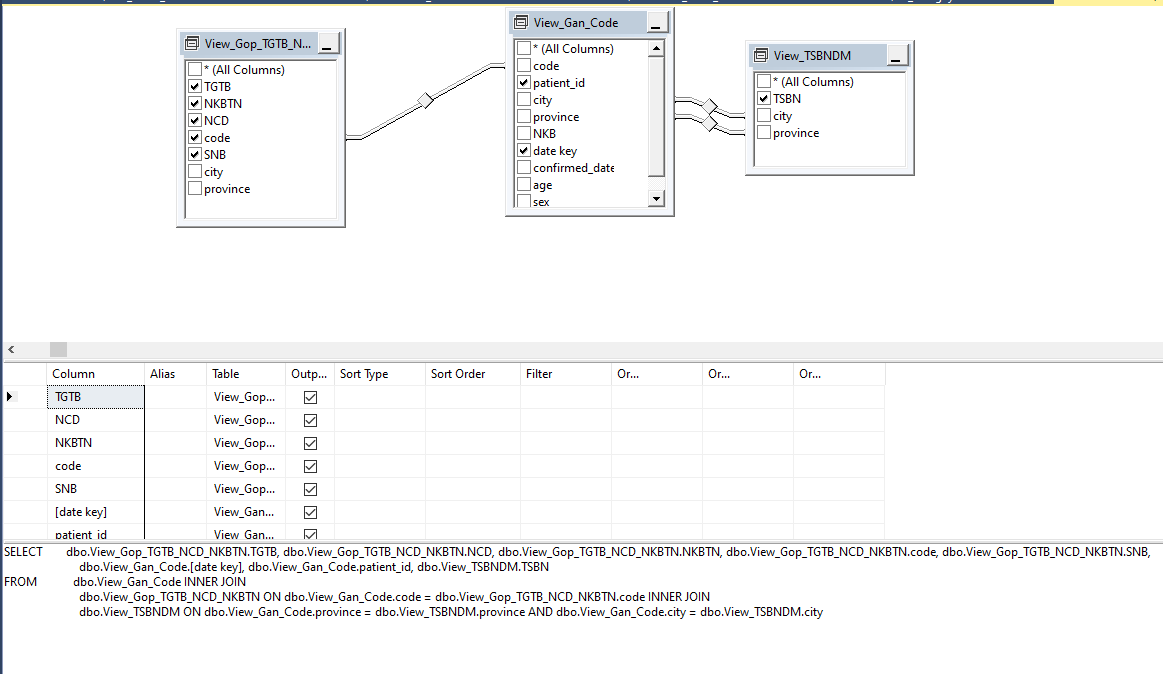
SELECT dbo.View\_Gop\_TGTB\_NCD\_NKBTN.TGTB, dbo.View\_Gop\_TGTB\_NCD\_NKBTN.NCD, dbo.View\_Gop\_TGTB\_NCD\_NKBTN.NKBTN, dbo.View\_Gop\_TGTB\_NCD\_NKBTN.code, dbo.View\_Gop\_TGTB\_NCD\_NKBTN.SNB,

dbo.View\_Gan\_Code.[date key], dbo.View\_Gan\_Code.patient\_id, dbo.View\_TSBNDM.TSBN

FROM dbo.View\_Gan\_Code INNER JOIN

dbo.View\_Gop\_TGTB\_NCD\_NKBTN ON dbo.View\_Gan\_Code.code = dbo.View\_Gop\_TGTB\_NCD\_NKBTN.code INNER JOIN

dbo.View\_TSBNDM ON dbo.View\_Gan\_Code.province = dbo.View\_TSBNDM.province AND dbo.View\_Gan\_Code.city = dbo.View\_TSBNDM.city



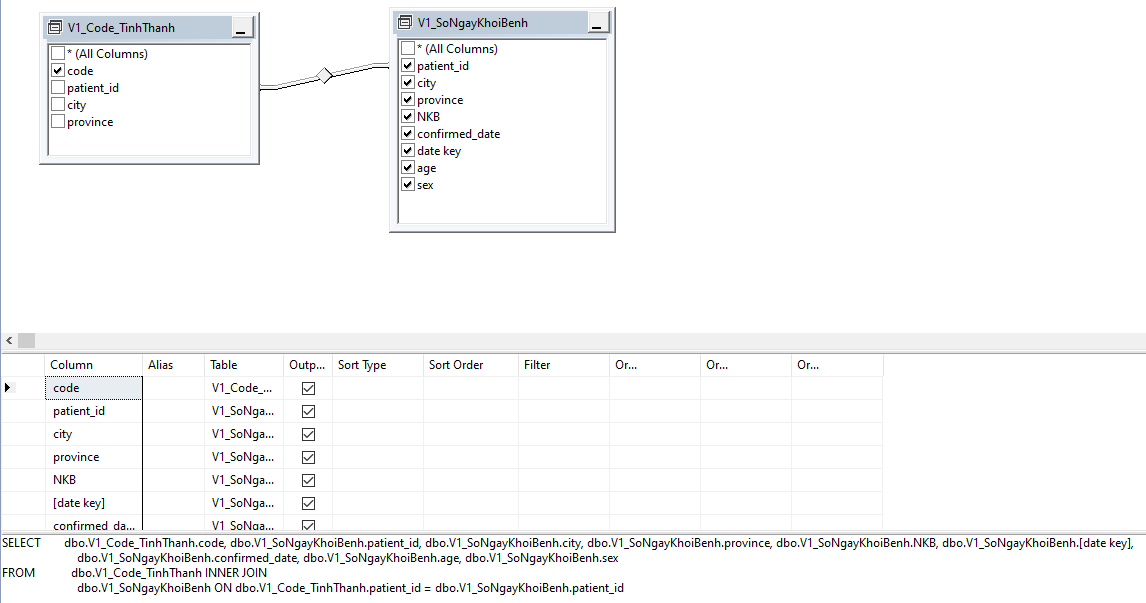
**View\_Gan\_Code** :Gắn mã Vùng miền vào số ngày vùng miền đã tính

SELECT dbo.V1\_Code\_TinhThanh.code, dbo.V1\_SoNgayKhoiBenh.patient\_id, dbo.V1\_SoNgayKhoiBenh.city, dbo.V1\_SoNgayKhoiBenh.province, dbo.V1\_SoNgayKhoiBenh.NKB, dbo.V1\_SoNgayKhoiBenh.[date key],

dbo.V1\_SoNgayKhoiBenh.confirmed\_date, dbo.V1\_SoNgayKhoiBenh.age, dbo.V1\_SoNgayKhoiBenh.sex

FROM dbo.V1\_Code\_TinhThanh INNER JOIN

dbo.V1\_SoNgayKhoiBenh ON dbo.V1\_Code\_TinhThanh.patient\_id = dbo.V1\_SoNgayKhoiBenh.patient\_id



**View\_Gop\_TGTB\_NCD\_NKBTN** : Gộp ngày có dịch và ngày bệnh khỏi bệnh gần nhất vào thời gian trung bình

SELECT dbo.View\_ThoiGianTBKB\_Theo\_Code.TGTB, dbo.View\_TAONKBTN\_key.NKBTN, dbo.View\_TaoNCD\_key.NCD, dbo.View\_ThoiGianTBKB\_Theo\_Code.code, dbo.View\_ThoiGianTBKB\_Theo\_Code.SNB,

dbo.View\_ThoiGianTBKB\_Theo\_Code.city, dbo.View\_ThoiGianTBKB\_Theo\_Code.province

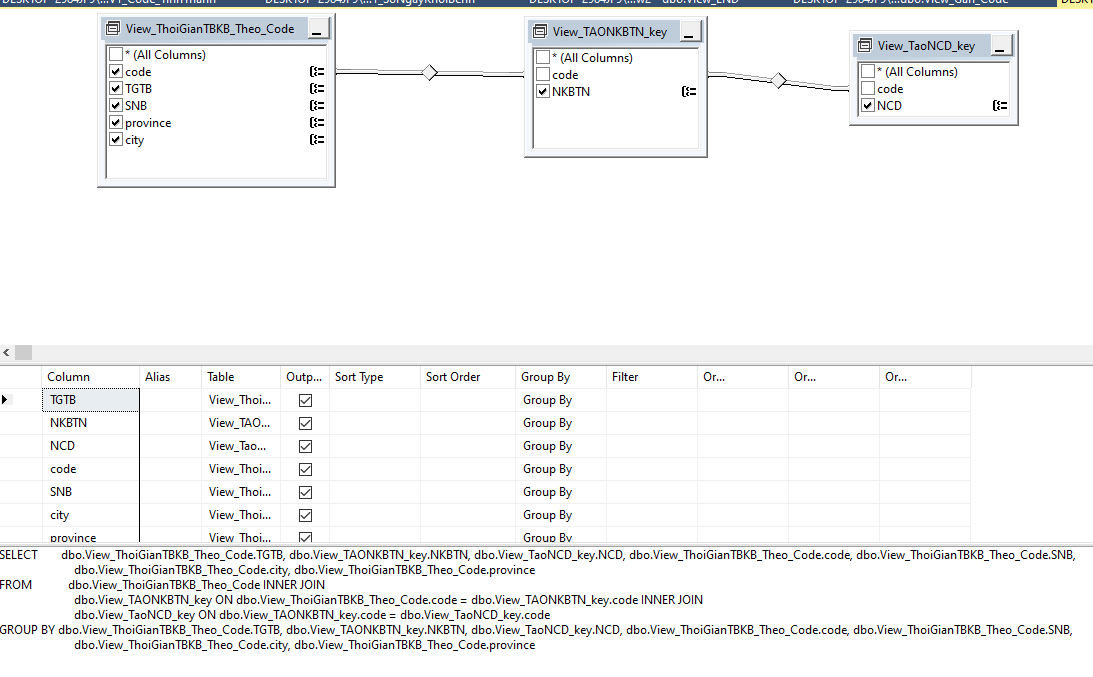
FROM dbo.View\_ThoiGianTBKB\_Theo\_Code INNER JOIN

dbo.View\_TAONKBTN\_key ON dbo.View\_ThoiGianTBKB\_Theo\_Code.code = dbo.View\_TAONKBTN\_key.code INNER JOIN

dbo.View\_TaoNCD\_key ON dbo.View\_TAONKBTN\_key.code = dbo.View\_TaoNCD\_key.code

GROUP BY dbo.View\_ThoiGianTBKB\_Theo\_Code.TGTB, dbo.View\_TAONKBTN\_key.NKBTN, dbo.View\_TaoNCD\_key.NCD, dbo.View\_ThoiGianTBKB\_Theo\_Code.code, dbo.View\_ThoiGianTBKB\_Theo\_Code.SNB,

dbo.View\_ThoiGianTBKB\_Theo\_Code.city, dbo.View\_ThoiGianTBKB\_Theo\_Code.province



**View\_NBD\_NKTTN\_TheoCode** :Tính ngày có dịch và ngày khỏi bệnh gần nhất

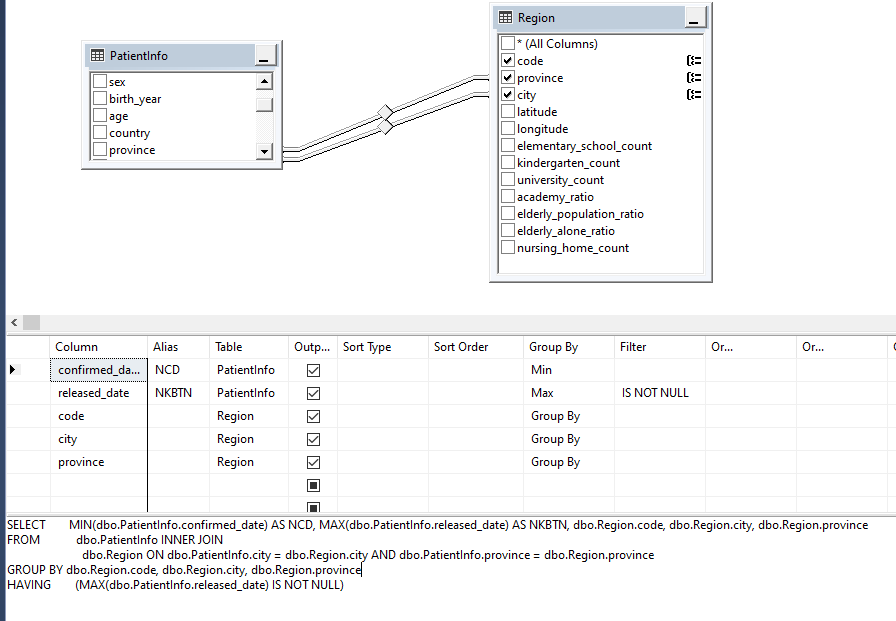
SELECT MIN(dbo.PatientInfo.confirmed\_date) AS NCD, MAX(dbo.PatientInfo.released\_date) AS NKBTN, dbo.Region.code, dbo.Region.city, dbo.Region.province

FROM dbo.PatientInfo INNER JOIN

dbo.Region ON dbo.PatientInfo.city = dbo.Region.city AND dbo.PatientInfo.province = dbo.Region.province

GROUP BY dbo.Region.code, dbo.Region.city, dbo.Region.province

HAVING (MAX(dbo.PatientInfo.released\_date) IS NOT NULL)

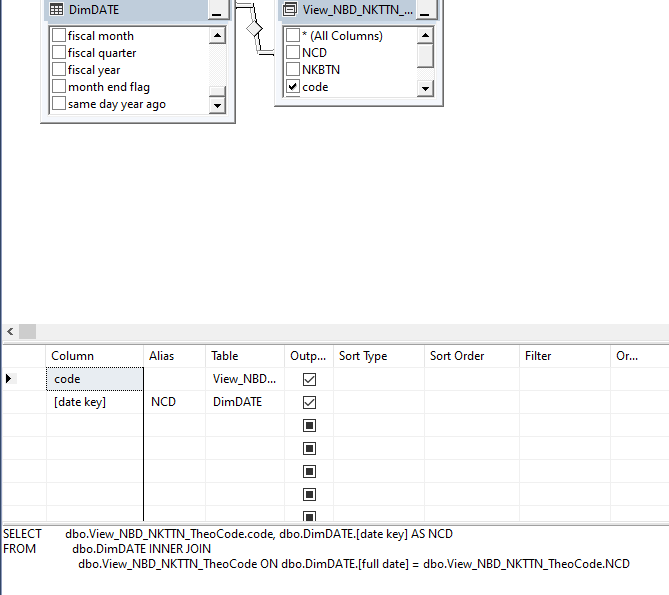


**View\_TaoNCD\_key**: Tạo key date cho ngày có dịch

SELECT dbo.View\_NBD\_NKTTN\_TheoCode.code, dbo.DimDATE.[date key] AS NCD

FROM dbo.DimDATE INNER JOIN

dbo.View\_NBD\_NKTTN\_TheoCode ON dbo.DimDATE.[full date] = dbo.View\_NBD\_NKTTN\_TheoCode.NCD

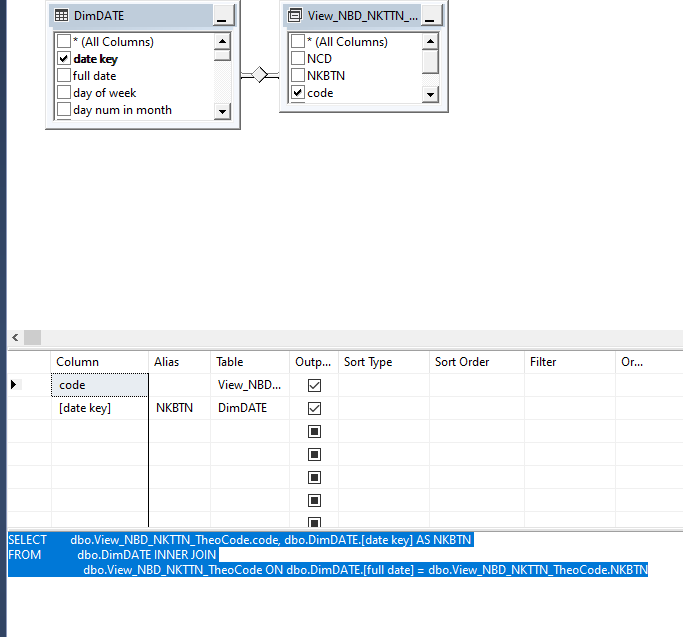


**View\_TAONKBTN\_key** : Tạo key date cho ngày khỏi bệnh gần nhất

SELECT dbo.View\_NBD\_NKTTN\_TheoCode.code, dbo.DimDATE.[date key] AS NKBTN

FROM dbo.DimDATE INNER JOIN

dbo.View\_NBD\_NKTTN\_TheoCode ON dbo.DimDATE.[full date] = dbo.View\_NBD\_NKTTN\_TheoCode.NKBTN

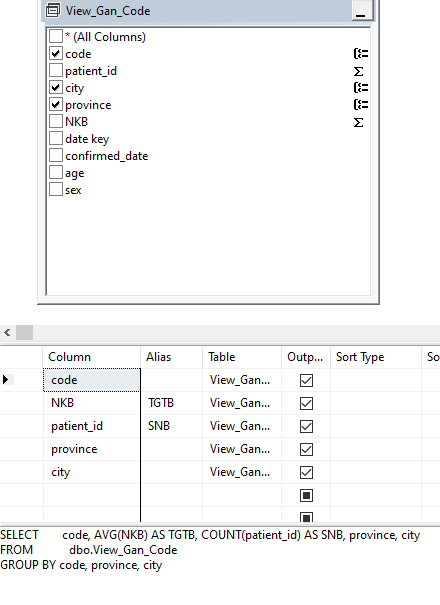


**View\_ThoiGianTBKB\_Theo\_Code**: Tính thời gian trung bình và tổng số bệnh nhân khỏi bệnh

SELECT code, AVG(NKB) AS TGTB, COUNT(patient\_id) AS SNB, province, city

FROM dbo.View\_Gan\_Code

GROUP BY code, province, city

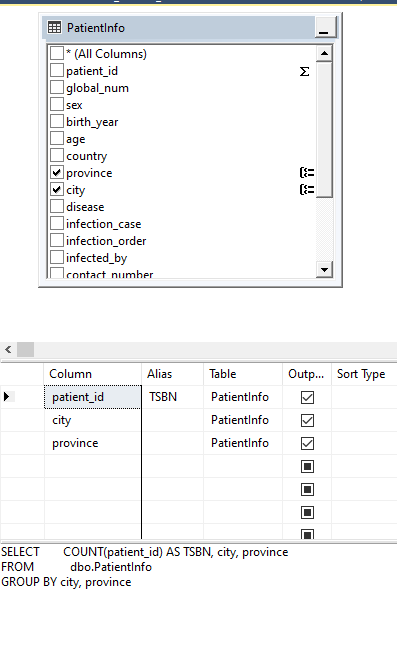


**View\_TSBNDM**:Tính tổng số bệnh mắc bệnh

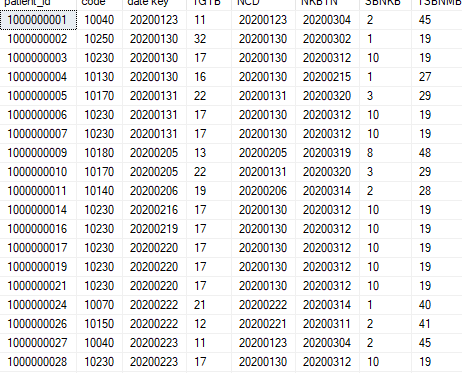
SELECT COUNT(patient\_id) AS TSBN, city, province

FROM dbo.PatientInfo

GROUP BY city, province

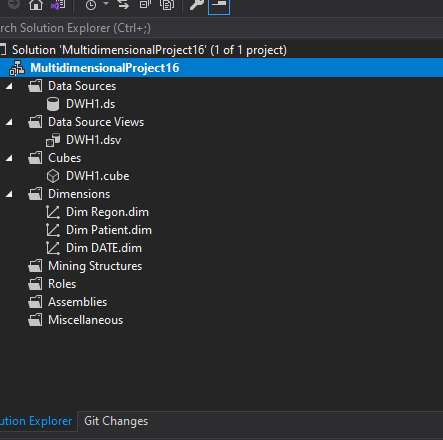


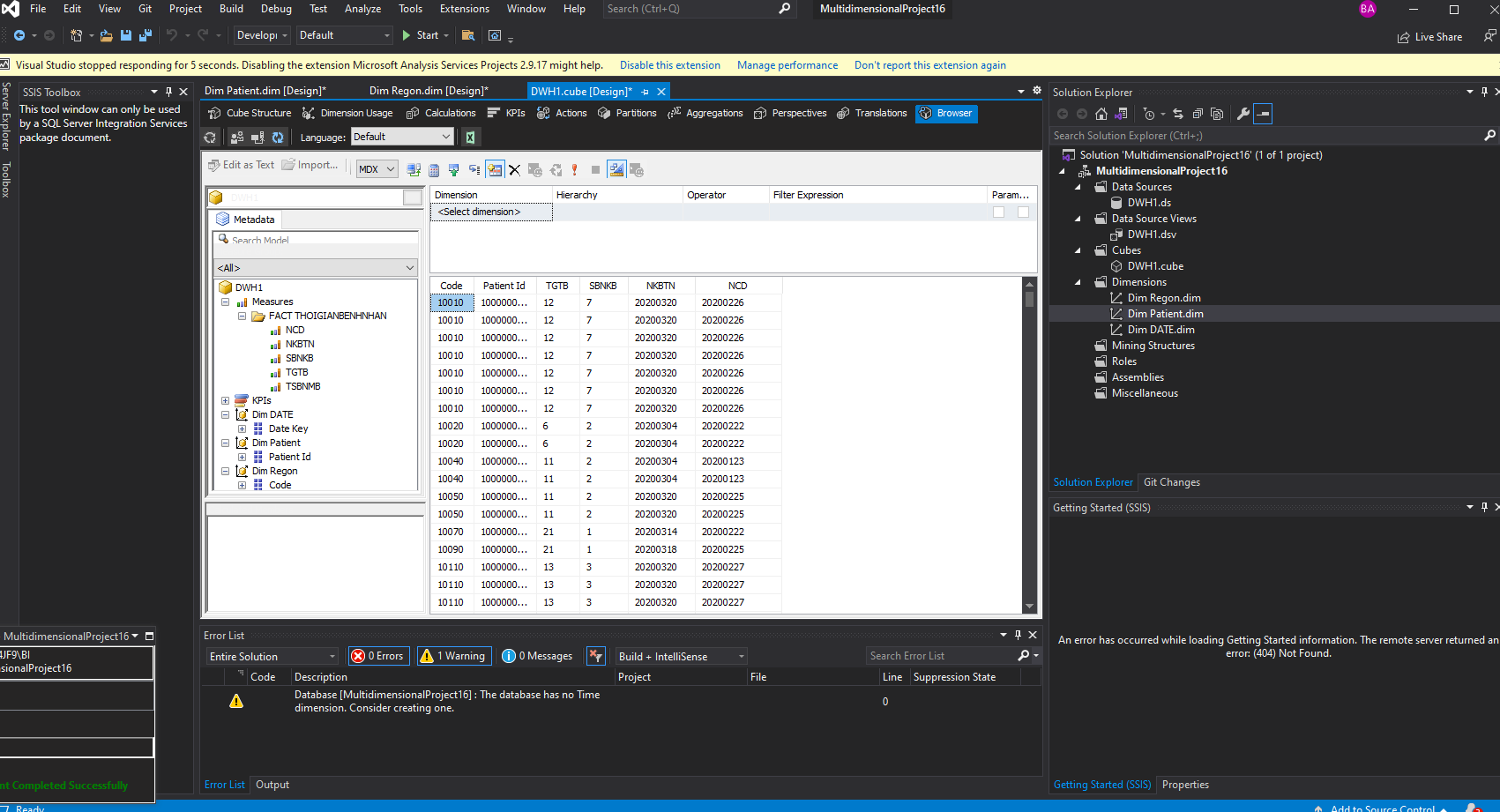
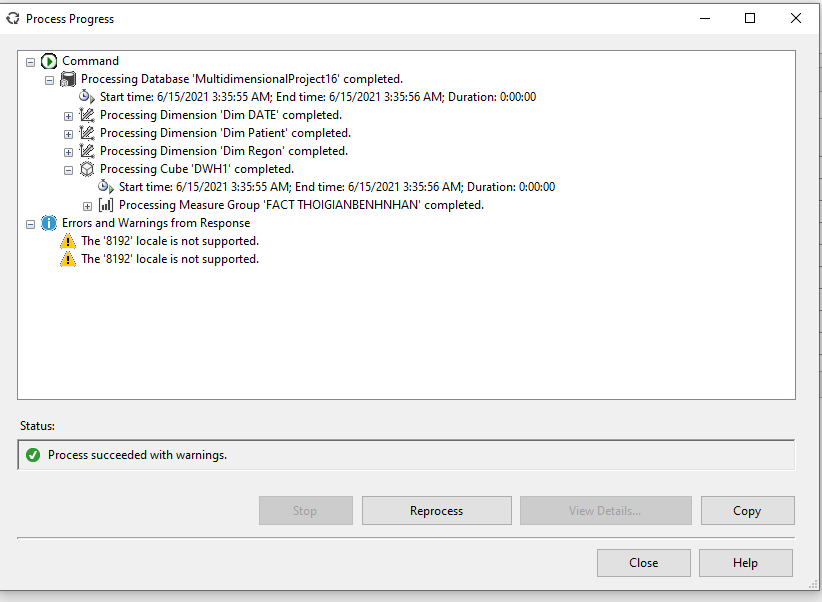
Nạp vào DWH



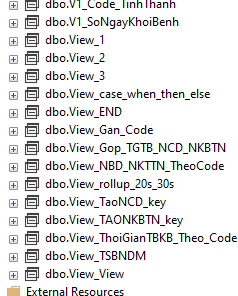
### Triển khai phân hệ DWH lên BIDS

Tham khảo Buổi 4, 5

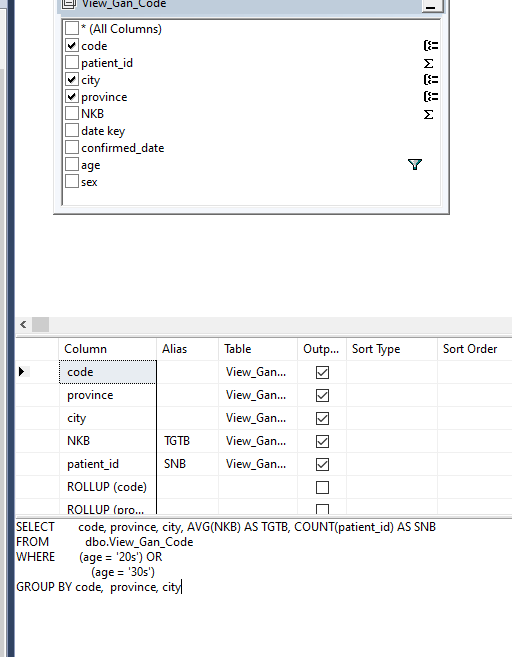




### Khai thác sử dụng phân hệ DWH dùng các vân tin SQL



**+ Câu hỏi : Từ dữ liệu gốc tạo view tính thời gian trung bình và số bệnh nhân khỏi bệnh có độ tuổi 20 -40**



SELECT code, province, city, AVG(NKB) AS TGTB, COUNT(patient\_id) AS SNB

FROM dbo.View\_Gan\_Code

WHERE (age = '20s') OR

(age = '30s')

GROUP BY code, province, city

Bênh nhân là có độ tuổi 20 – 40 dùng roll up

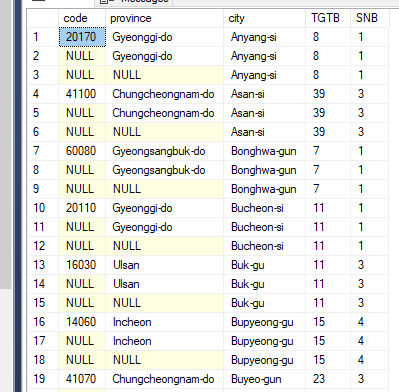
SELECT code, province, city, AVG(NKB) AS TGTB, COUNT(patient\_id) AS SNB

FROM dbo.View\_Gan\_Code

WHERE (age = '20s') OR

(age = '30s')

GROUP BY ROLLUP (code), ROLLUP (province), ROLLUP (city)



**Dùng CUBE**

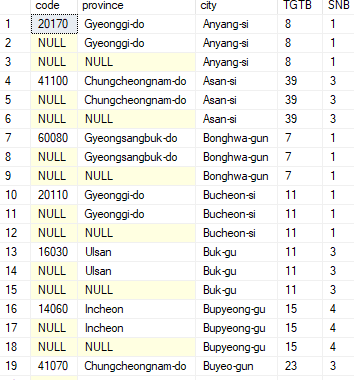
SELECT code, province, city, AVG(NKB) AS TGTB, COUNT(patient\_id) AS SNB

FROM dbo.View\_Gan\_Code

WHERE (age = '20s') OR

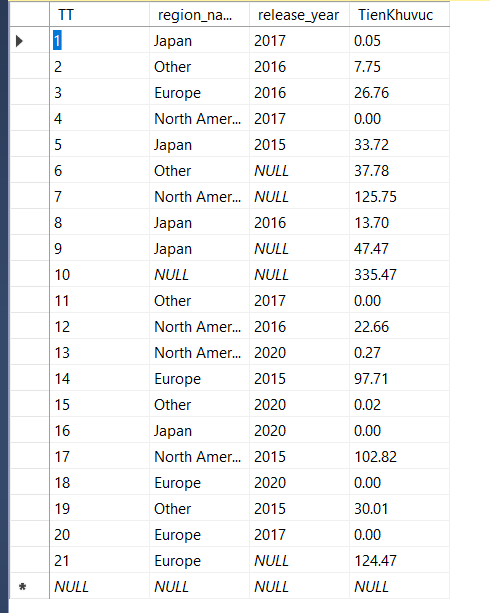
(age = '30s')

GROUP BY cube (code), cube (province), cube (city)

****

**Thêm cột STT vào view trên**

SELECT row\_number() OVER (ORDER BY newid()) AS TT, \*

FROM dbo.View\_rollup\_20s\_30s 

Dùng CASE WHEN… THEN …ELSE…END => thêm “Bênh nhân” vào các dòng thống kê

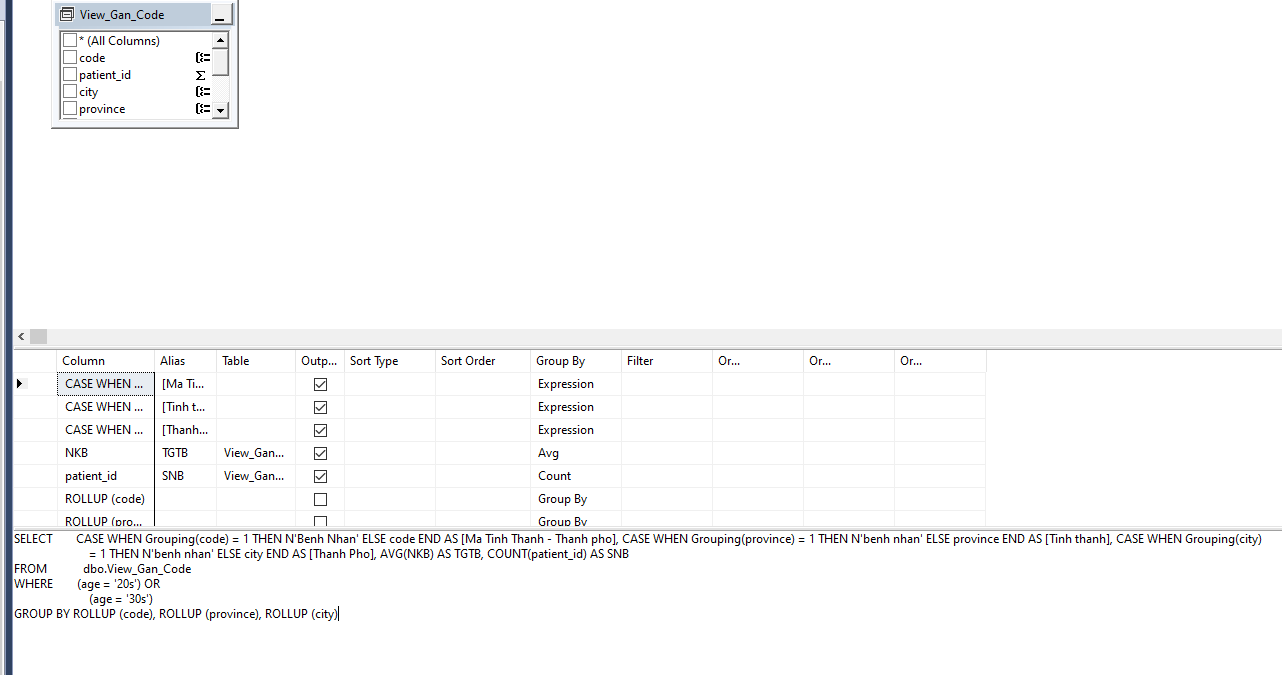
SELECT CASE WHEN Grouping(code) = 1 THEN N'Benh Nhan' ELSE code END AS [Ma Tinh Thanh - Thanh pho], CASE WHEN Grouping(province) = 1 THEN N'benh nhan' ELSE province END AS [Tinh thanh], CASE WHEN Grouping(city)

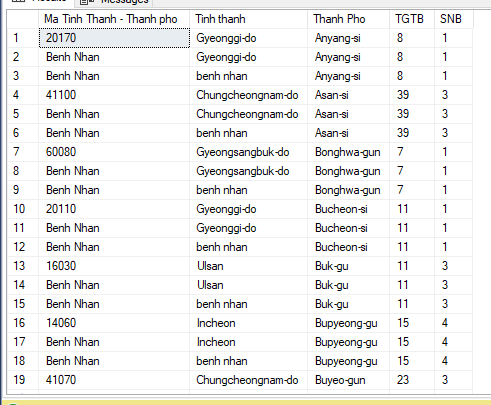
= 1 THEN N'benh nhan' ELSE city END AS [Thanh Pho], AVG(NKB) AS TGTB, COUNT(patient\_id) AS SNB

FROM dbo.View\_Gan\_Code

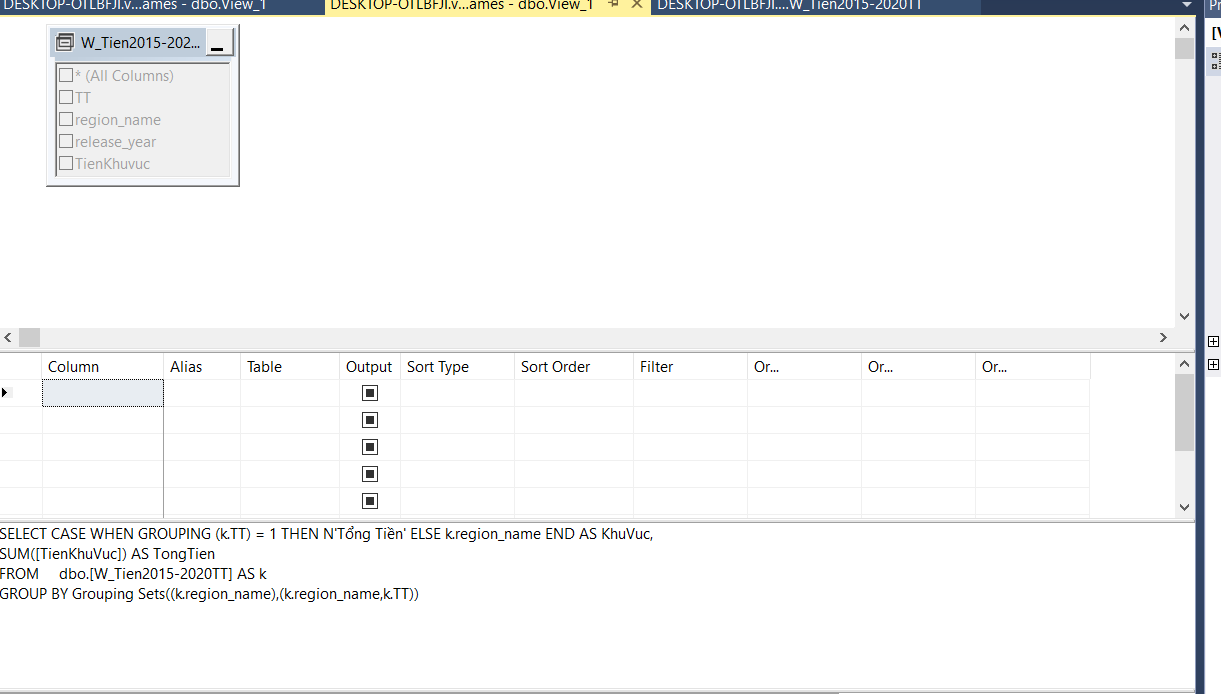
WHERE (age = '20s') OR

(age = '30s')

GROUP BY ROLLUP (code), ROLLUP (province), ROLLUP (city) 



Tính tiền thuốc theo đợt khám = thống kê thông qua gôm nhóm dữ liệu nguồn (Grouping sets)

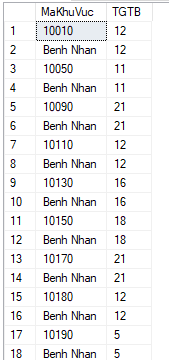


SELECT CASE WHEN GROUPING (TT) = 1 THEN N'Benh Nhan' ELSE code END AS MaKhuVuc,

SUM(TGTB) AS TGTB

FROM dbo.View\_2

GROUP BY Grouping Sets((code),(code,TT))



# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỒ ÁN HỌC PHẦN

## Kết luận

### Những kết quả đạt được

- Thống kê được số người mắc covid ở từng tỉnh, từng thành phố

- Thống kê được số người khỏi bệnh ở từng tỉnh, từng thành phố

- Thống kê số ca tử vong ở từng tỉnh, từng thành phố

- Hiểu rõ và sử dụng thành thạo các vấn tin SQL qua các view đã tạo

- Triển khai phân tích phân hệ DWH trên BIDS

- Kết quả thu được của đề tài giúp đưa ra các biện pháp tốt nhất để phòng dịch covid

### Hạn chế

- Chưa thống kê được các bệnh nhân đang chữa trị ở từng tỉnh, từng thành phố

- Chưa trực quan hóa dữ liệu ra biểu đồ để có thể dễ hình dung và đưa ra những dự đoán chính xác hơn.

## Hướng phát triển

### Hướng khắc phục các hạn chế

- Để khắc phục hạn chế thống kê bệnh nhân đang chữa trị, em sẽ lấy tổng bệnh nhân trừ cho tổng của các bệnh nhân đã khỏi bệnh và các bệnh nhân đã tử vong.

- Để khắc phục vấn đề trực quan hóa dữ liệu em sẽ sử dụng các tools để vẽ biểu đồ và quan sát

### Hướng mở rộng ĐỒ ÁN HỌC PHẦN

+ Dùng tool nâng cao là SSIS để làm sạch dữ liệu và đổ dữ liệu vào DWH

+ Triển khai dữ liệu này lên web đề người dùng có thể xem và tương tác với dữ liệu

+ Dùng nhiều tool khác nhau để vẽ biểu đồ

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Võ Xuân Thể (2021), *Bài giảng học phần Kho dữ liệu*, TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM (lưu hành nội bộ).

# CÁC PHỤ LỤC

(nếu có)

# ĐÁNH GIÁ KẾT QUẢ HỌC TẬP HỌC PHẦN

ĐỒ ÁN HP THAY CHO BÀI THI KẾT THÚC HỌC PHẦN

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hình thức KT** | **Nội dung** | **Chuẩn đầu ra đánh giá** | **Trình độ năng lực** | **Phương pháp đánh giá** | **Công cụ đánh giá** | **Tỉ lệ (%)** |
| **BÀI TẬP LẬP TRÌNH** | | | | | |  |
| BL#1 | Bài tập lập trình theo từng Buổi thực hành = tính bình quân  Kiểm tra viết | G1.1 | 3 | Bài tập  Kiểm tra | Bài tập  Kiểm tra | 50 |
| G2.1 | 3 |
| G3.1  G3.2 | 4  3 |
| G4.1 | 5 |
| **ĐỒ ÁN HỌC PHẦN (Project)** | | | | | |  |
| ĐA#1 | Phát triển một hệ thống KHO DỮ LIỆU đơn giản:  + Đề tài Nhóm có phân công phân hệ cụ thể cho từng thành viên thuộc nhóm  + Xây dựng và thực nghiệm Hệ thống Kho dữ liệu theo nhóm và phân hệ cá nhân  => báo cáo vào ngày thi. | G1.1 | 3 | Báo cáo, Demo và thuyết trình | Rubric | 50 |
| G2.1 | 3 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |
| G2.1 | 4 |
| G3.1 | 4 |
| G3.2  G4.1 | 3  5 |