**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

bdôca

A picture containing text

Description automatically generated

**BÁO CÁO ĐỒ ÁN**

**KHO DỮ LIỆU VÀ OLAP**

**Đề tài: Xây dựng và phân tích**

**kho dữ liệu ung thư ở Brazil**

**GVHD: ThS. Nguyễn Thị Kim Phụng**

**GVHD: Thầy Lê Võ Đình Kha**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | | |
| STT | Họ tên | MSSV |
| 1 | Nguyễn Minh Hiếu | 20521326 |
| 2 | Vũ Tiến Linh | 19521760 |

**TP. HỒ CHÍ MINH , NĂM 2023**

Mục lục

[CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI 7](#_Toc139321866)

[1.1. Lý do chọn đề tài 7](#_Toc139321867)

[1.2. Giới thiệu đề tài 8](#_Toc139321868)

[1.3. Mô tả dữ liệu gốc 9](#_Toc139321869)

[1.4. Sơ đồ hình sao 12](#_Toc139321870)

[1.5. Mô tả các bạn dữ liệu 12](#_Toc139321871)

[CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO 16](#_Toc139321872)

[2.1. Tạo kho dữ liệu bằng SQL Server 16](#_Toc139321873)

[2.2. Tạo Project trong Visual Studio 17](#_Toc139321874)

[2.3. Tạo bảng Dim và Fact 18](#_Toc139321875)

[2.3.1. Tạo bảng DimPatient 18](#_Toc139321876)

[2.3.2. DimTime 29](#_Toc139321877)

[2.3.3. Tạo bảng DimDisease 32](#_Toc139321878)

[2.3.4. Tạo bảng DimMorphology 35](#_Toc139321879)

[2.3.5. Tạo bảng DimTopography 38](#_Toc139321880)

[2.3.6. Tạo bảng FactMedical 41](#_Toc139321881)

[CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN – QUÁ TRÌNH SSAS 45](#_Toc139321882)

[3.1. Cấu hình project SSAS 45](#_Toc139321883)

[3.1.1. Tạo project 45](#_Toc139321884)

[3.1.2. Tạo Data Sources 46](#_Toc139321885)

[3.1.3. Tạo Data Sources View 47](#_Toc139321886)

[3.1.4. Xây dựng các Cube và các Measure 49](#_Toc139321887)

[3.1.5. Chỉnh sửa Dimensions 50](#_Toc139321888)

[3.1.6. Process Cube 51](#_Toc139321889)

[3.2. Quá trình phân tích dữ liệu bằng ngôn ngữ MDX 52](#_Toc139321890)

[3.2.1. Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177 52](#_Toc139321891)

[3.2.2. Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012 52](#_Toc139321892)

[3.2.3. Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo 53](#_Toc139321893)

[3.2.4. Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm 53](#_Toc139321894)

[3.2.5. Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó 54](#_Toc139321895)

[3.2.6. Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000 55](#_Toc139321896)

[3.2.7. Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown 56](#_Toc139321897)

[3.2.8. Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013 56](#_Toc139321898)

[3.2.9. Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần 57](#_Toc139321899)

[3.2.10. Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất 57](#_Toc139321900)

[3.2.11. Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000 58](#_Toc139321901)

[3.2.12. Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất 58](#_Toc139321902)

[3.3. Quá trình phân tích dữ liệu trên bằng công cụ SSAS trên các khối CUBE 59](#_Toc139321903)

[3.3.1. Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177 59](#_Toc139321904)

[3.3.2. Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012 59](#_Toc139321905)

[3.3.3. Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo 60](#_Toc139321906)

[3.3.4. Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm 60](#_Toc139321907)

[3.3.5. Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó 60](#_Toc139321908)

[3.3.6. Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000 61](#_Toc139321909)

[3.3.7. Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown 61](#_Toc139321910)

[3.3.8. Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013 61](#_Toc139321911)

[3.3.9. Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần 62](#_Toc139321912)

[3.3.10. Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất 62](#_Toc139321913)

[3.3.11. Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000 62](#_Toc139321914)

[3.3.12. Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất 63](#_Toc139321915)

[3.4. Quá trình phân tích dữ liệu trên bằng công cụ Privot Excel 63](#_Toc139321916)

[3.4.1. Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177 63](#_Toc139321917)

[3.4.2. Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012 64](#_Toc139321918)

[3.4.3. Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo 64](#_Toc139321919)

[3.4.4. Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm 65](#_Toc139321920)

[3.4.5. Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó 66](#_Toc139321921)

[3.4.6. Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000 67](#_Toc139321922)

[3.4.7. Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown 68](#_Toc139321923)

[3.4.8. Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013 68](#_Toc139321924)

[3.4.9. Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần 69](#_Toc139321925)

[3.4.10. Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất 69](#_Toc139321926)

[3.4.11. Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000 70](#_Toc139321927)

[3.4.12. Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất 71](#_Toc139321928)

[3.5. Quá trình phân tích dữ liệu trên bằng công cụ Power BI 72](#_Toc139321929)

[3.5.1. Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177 72](#_Toc139321930)

[3.5.2. Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012 72](#_Toc139321931)

[3.5.3. Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo 73](#_Toc139321932)

[3.5.4. Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm 73](#_Toc139321933)

[3.5.5. Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó 73](#_Toc139321934)

[3.5.6. Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000 74](#_Toc139321935)

[3.5.7. Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown 74](#_Toc139321936)

[3.5.8. Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013 75](#_Toc139321937)

[3.5.9. Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần 75](#_Toc139321938)

[3.5.10. Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất 75](#_Toc139321939)

[3.5.11. Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000 76](#_Toc139321940)

[3.5.12. Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất 77](#_Toc139321941)

[CHƯƠNG 4. QUÁ TRÌNH KHAI THÁC DỮ LIỆU - DATA MINING 77](#_Toc139321942)

[4.1. Tiền xử lý dữ liệu 77](#_Toc139321943)

[4.2. RandomForest 80](#_Toc139321944)

[4.3. Naive Bayes 81](#_Toc139321945)

[4.4. Nhận xét 83](#_Toc139321946)

**DANH SÁCH HÌNH ẢNH**

[Hình 1. Sơ đồ hình sao của kho dữ liệu ung thư 12](#_Toc139321817)

[Hình 2. Đặt tên và tùy biến kích thước kho dữ liệu 16](#_Toc139321818)

[Hình 3. Các tùy chọn cần thiết 17](#_Toc139321819)

[Hình 5. Đặt tên cho Project SSIS 18](#_Toc139321820)

[Hình 4. Giao diện chính của SSIS 18](https://uithcm-my.sharepoint.com/personal/20521326_ms_uit_edu_vn/Documents/HK%202%20Năm%203/Nhom40_20521326_Nguyễn%20Minh%20Hiếu_19521760_Vũ%20Tiến%20Linh/Báo%20cáo/Báo%20cáo.docx#_Toc139321821)

[Hình 6. Kéo Data Flow Task vào màn hình làm việc và đổi tên thành Pateint 19](#_Toc139321822)

[Hình 7. Kéo Flat File Source vào môi trường làm việc của Data Flow 19](#_Toc139321823)

[Hình 8. Chọn đường dẫn đến file csv 20](#_Toc139321824)

[Hình 9. Chọn thuộc tính cho DimPatient 21](#_Toc139321825)

[Hình 10. Chuyển đổi các giá trị 21](#_Toc139321826)

[Hình 11. Chọn Transformation 22](#_Toc139321827)

[Hình 12. Chọn input colum 23](#_Toc139321828)

[Hình 13. Đoạn code để bỏ các dòng khống đúng định dạng ngày 23](#_Toc139321829)

[Hình 14. Copy sang cột mới và đổi kiểu dữ liệu 24](#_Toc139321830)

[Hình 15. Đặt điều kiện cho Case 1 24](#_Toc139321831)

[Hình 16. Sắp sếp cột Copy of Patient Code tăng dần và bỏ các giá trị trùng lặp với Copy of Patient Code 25](#_Toc139321832)

[Hình 17. Kết nối với kho dữ liệu Cancer\_DW 26](#_Toc139321833)

[Hình 18. Tạo bảng DimPatient vào kho dữ liệu 26](#_Toc139321834)

[Hình 19. Quá trình Mappings dữ liệu bảng DimPatient 27](#_Toc139321835)

[Hình 20. Quá trình đổ dữ liệu vào DimPatient 27](#_Toc139321836)

[Hình 21. Đổ dữ liệu thành công 27](#_Toc139321837)

[Hình 22. Kiểm tra dữ liệu đã đổ thành công 28](#_Toc139321838)

[Hình 23. Kết nối Execute SQL Task với Sequence Container 28](#_Toc139321839)

[Hình 24. Viết câu lệnh loại bỏ trùng sau mỗi lần đổ dữ liệu của bảng DimPatient 29](#_Toc139321840)

[Hình 25. Chọn các thuộc tính cho bảng DimTime 30](#_Toc139321841)

[Hình 26. Đoạn code để gộp dữ liệu vào cột Date 30](#_Toc139321842)

[Hình 27. Sắp xếp ngày theo Date và xóa trùng lặp 31](#_Toc139321843)

[Hình 28. Thêm 2 cột mới Month và Year 31](#_Toc139321844)

[Hình 29. Tạo bảng DimTime 32](#_Toc139321845)

[Hình 30. Case 1 lọc giá trị null của IllnessCode 33](#_Toc139321846)

[Hình 31. Derived Column chuyển giá trị NA thành NULL 34](#_Toc139321847)

[Hình 32. Sắp xếp theo Illness Code 34](#_Toc139321848)

[Hình 33. Tạo bảo DimDisease 35](#_Toc139321849)

[Hình 34. Chọn các thuộc tính cho bảng DimMorphology 36](#_Toc139321850)

[Hình 35. Case 1 lọc các giá trị null 37](#_Toc139321851)

[Hình 36. Sort theo Code of Morphology và xóa trùng lặp 37](#_Toc139321852)

[Hình 37. Tạo bảng DimMorphology 38](#_Toc139321853)

[Hình 38. Chọn các thuộc tính cho bảng DimTopography 38](#_Toc139321854)

[Hình 39. Case 1 Lọc các giá trị null 39](#_Toc139321855)

[Hình 40. Sort theo Topography và xóa trùng lặp 40](#_Toc139321856)

[Hình 41. Tạo bảng DimTopography vào kho dữ liệu 40](#_Toc139321857)

[Hình 42. Chọn các thuộc tính cho bảng FactMedical 41](#_Toc139321858)

[Hình 43. Thay giá trị "NA" thành NULL 41](#_Toc139321859)

[Hình 44. Đoạn code để đổi các giá trị ngày đúng định dạng và copy sang các cột khác 42](#_Toc139321860)

[Hình 45. Copy thuộc tính Patient Code sang cột mới và chuyển sang kiểu dữ liệu int 42](#_Toc139321861)

[Hình 46. So sánh giữa Copy of Patient Code của bảng FactMedical với PatientCode của DimPatient 43](#_Toc139321862)

[Hình 47. Tạo bảng FactMedical 44](#_Toc139321863)

[Hình 48. Thêm khóa ngoại 44](#_Toc139321864)

[Hình 49. Xóa khóa ngoại 45](#_Toc139321865)

**DANH SÁCH CÁC BẢNG**

[Bảng 1. Mô tả các thuộc tính của dataset 11](#_Toc139321808)

[Bảng 2. Mô tả về bảng DimPatient 13](#_Toc139321809)

[Bảng 3. Mô tả bảng DimDisease 14](#_Toc139321810)

[Bảng 4. Mô tả bảng DimTime 14](#_Toc139321811)

[Bảng 5. Mô tả bảng DimMorphology 14](#_Toc139321812)

[Bảng 6. Mô tả bảng DimTopography 15](#_Toc139321813)

[Bảng 7 Mô tả bảng FactMedical 16](#_Toc139321814)

[Bảng 8. Thống kê mã bệnh ở người trẻ 83](#_Toc139321815)

[Bảng 9. Thống kê mã bệnh 83](#_Toc139321816)

# TỔNG QUAN VỀ ĐỀ TÀI

## Lý do chọn đề tài

Ung thư là một căn bệnh phức tạp đòi hỏi một cách tiếp cận đa ngành để điều trị. Một trong những thành phần quan trọng của việc chăm sóc ung thư hiệu quả là có quyền truy cập vào dữ liệu toàn diện và chính xác. Công nghệ thông tin trong chăm sóc sức khỏe vẫn là một chủ đề thời sự, đã kích thích sự phát triển trong việc nhập y lệnh của bác sĩ, hệ thống hỗ trợ quyết định và chia sẻ hồ sơ bệnh nhân. Bất chấp mọi nỗ lực, nhiều tổ chức chăm sóc sức khỏe vẫn có hệ thống thông tin độc lập không liên lạc với nhau. Quan trọng hơn, các hệ thống thông tin lâm sàng (CIS) như hồ sơ bệnh nhân điện tử, thường được thiết kế để hỗ trợ chăm sóc trực tiếp cho từng bệnh nhân, nhưng không phù hợp để phân tích ở cấp độ tổng hợp, ví dụ như trên các nhóm bệnh nhân mắc cùng một bệnh.

Do đó, việc xây dựng kho dữ liệu chăm sóc sức khỏe dành riêng cho các bệnh ung thư có thể được sử dụng bởi các nhà quản lý điều hành, bác sĩ, y sĩ và các chuyên gia y tế khác để hỗ trợ quá trình chăm sóc sức khỏe sẽ mang lại một số lợi ích như sau:

* Tích hợp dữ liệu: Kho dữ liệu giúp tích hợp dữ liệu từ nhiều nguồn khác nhau, như hồ sơ bệnh nhân, kết quả xét nghiệm, và thông tin điều trị, tạo ra một nền tảng dữ liệu thống nhất, giúp các nhà nghiên cứu, bác sĩ, và các bên liên quan dễ dàng truy cập và phân tích dữ liệu.
* Phân tích dữ liệu nhanh chóng: OLAP cho phép thực hiện các truy vấn phức tạp và phân tích dữ liệu nhanh chóng bằng cách sử dụng các kỹ thuật như tổng hợp dữ liệu và tiền xử lý. Điều này giúp giảm thời gian phân tích và đưa ra các quyết định chính xác hơn.
* Phân tích xu hướng và mô hình: Kho dữ liệu và OLAP giúp phát hiện các xu hướng, mô hình, và liên kết trong dữ liệu, giúp các nhà nghiên cứu hiểu rõ hơn về các yếu tố nguy cơ, dự đoán kết quả, và đề xuất các phương pháp điều trị phù hợp.
* Hỗ trợ quyết định: Kho dữ liệu và OLAP hỗ trợ các bác sĩ và nhà quản lý trong việc đưa ra các quyết định về chăm sóc bệnh nhân, phân bổ nguồn lực, và nghiên cứu ung thư.
* Giám sát và đánh giá: Kho dữ liệu và OLAP giúp giám sát hiệu quả của các phương pháp điều trị, đánh giá chất lượng dịch vụ y tế, và đo lường hiệu quả của các chương trình phòng chống ung thư.
* Nâng cao chất lượng dữ liệu: Việc xây dựng kho dữ liệu yêu cầu rà soát và làm sạch dữ liệu, giúp nâng cao chất lượng dữ liệu và đảm bảo tính nhất quán, chuẩn xác của thông tin.
* Hỗ trợ nghiên cứu và hợp tác: Kho dữ liệu và OLAP giúp chia sẻ dữ liệu ung thư giữa các tổ chức, nghiên cứu viên, và bác sĩ trên toàn cầu, thúc đẩy sự hợp tác trong nghiên cứu và phát triển các phương pháp điều trị mới.

Tóm lại, việc xây dựng kho dữ liệu và sử dụng OLAP trong dữ liệu ung thư giúp tăng cường khả năng phân tích, quản lý, và chia sẻ thông tin, từ đó hỗ trợ nghiên cứu, chăm sóc bệnh nhân, và phòng chống ung thư một cách hiệu quả hơn.

## Giới thiệu đề tài

* Tên Dataset: Cancer Data Brazil
* Link Dataset: [Cancer Data Brazil | Kaggle](https://www.kaggle.com/datasets/joaopedromedeiros/cancer-data-brazil?select=municipio.csv&fbclid=IwAR0U2u0auDQEizukn7MWRcampUdB95SwIH5NfC058Cqfbc3fbBHgVH9i-lk)
* Mô Tả: Dữ liệu này là tập hợp thông tin về bệnh ung thư ở các bệnh viện tại Brazil từ năm 2000 đến năm 2019. Dữ liệu này bao gồm các thông tin về bệnh nhân, bao gồm tuổi, giới tính và vị trí địa lý của bệnh viện, các thông tin về loại ung thư, phương pháp chẩn đoán và điều trị, và các chỉ số sức khỏe của bệnh nhân. Tập dữ liệu này hoàn toàn do Viện Ung thư Quốc gia INCA tạo ra và dữ liệu này phù hợp với luật pháp khu vực và có thể được yêu cầu bởi bất kỳ ai.
* Kích thước:
  + Với caner\_data\_eng có 38 thuộc tính với 1.048.574 dòng
  + Với municipio có 4 thuộc tính với 168.818 dòng

## Mô tả dữ liệu gốc

Bảng caner\_data\_eng

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Ý nghĩa | Kiểu dữ liệu |
| 1 | Patient.Code | Mã số bệnh nhân | Int |
| 2 | RCBP.Name | Tên của bệnh viện hoặc cơ sở y tế nơi bệnh nhân được điều trị | Varchar |
| 3 | Gender | Giới tính | Varchar |
| 4 | Date.of.Birth | Ngày sinh | Date |
| 5 | Age | Tuổi | Int |
| 6 | Raca.Color | Chủng tộc hoặc màu da của bệnh nhân. | Varchar |
| 7 | Nationality | Quốc tịch của bệnh nhân. | Varchar |
| 8 | Naturality.State | Tên tiểu bang hoặc tỉnh nơi bệnh nhân sinh ra | Varchar |
| 9 | Naturality | Quốc gia nơi bệnh nhân sinh ra | Varchar |
| 10 | Degree.of.Education | Trình độ học vấn của bệnh nhân | Varchar |
| 11 | State.Civil | Tình trạng hôn nhân của bệnh nhân | Varchar |
| 12 | Code.Profession | Mã số của nghề nghiệp của bệnh nhân | Int |
| 13 | Name.Occupation | Tên nghề nghiệp của bệnh nhân | Varchar |
| 14 | Status.Address | Tình trạng địa chỉ của bệnh nhân | Varchar |
| 15 | City.Address | Thành phố hoặc khu vực nơi bệnh nhân đang sống | Varchar |
| 16 | Description.of.Topography | Mô tả về vị trí của căn bệnh | Varchar |
| 17 | Topography.Code | Mã số của vị trí bệnh | Varchar |
| 18 | Morphology.Description | Mô tả về hình thái của căn bệnh | Varchar |
| 19 | Code.of.Morphology | Mã số về hình thái bệnh | Int |
| 20 | Description.of.Disease | Mô tả về căn bệnh | Varchar |
| 21 | Illness.Code | Mã số của căn bệnh | Varchar |
| 22 | Child.Illness.Description | Mô tả về căn bệnh ở trẻ em | Varchar |
| 23 | Child.Illness.Code | Mã số về căn bệnh ở trẻ em | Varchar |
| 24 | Youth.Adult.Illness.Description | Mô tả về căn bệnh ở người trưởng thành | Varchar |
| 25 | Code.of.Disease.Adult.Young | Mã số về căn bệnh ở người trưởng thành | Varchar |
| 26 | Indicator.of.Rare.Case | Chỉ số về trường hợp hiếm gặp | Binary |
| 27 | Diagnostic.means | Phương pháp chẩn đoán của căn bệnh | Varchar |
| 28 | Extension | Phạm vi của căn bệnh | Varchar |
| 29 | Laterality | Vị trí của căn bệnh | Varchar |
| 30 | Statement | Tình trạng của căn bệnh | Varchar |
| 31 | TNM | Mã số về tình trạng ung thư (tumor, nodes, metastasis) | Int |
| 32 | Status.Vital | Tình trạng sống còn của bệnh nhân | Varchar |
| 33 | Type.of.Death | Loại tử vong của bệnh nhân | Varchar |
| 34 | Date.of.Death | Ngày tử vong của bệnh nhân | Date |
| 35 | Date.of.Last.Contact | Ngày liên lạc cuối cùng với bệnh nhân | Date |
| 36 | Date.of.Diagnostic | Ngày chẩn đoán căn bệnh | Date |
| 37 | Distant.metastasis | Tình trạng bệnh lý xa của căn bệnh | Varchar |
| 38 | year | Năm khi dữ liệu được ghi lại | Int |

Bảng 1. Mô tả các thuộc tính của dataset

## Sơ đồ hình sao

A screenshot of a computer

Description automatically generated

Hình 1. Sơ đồ hình sao của kho dữ liệu ung thư

## Mô tả các bạn dữ liệu

DimPatient: lưu trữ thông tin cá nhân của bệnh nhân.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Patient.Code | Int | Not | Mã số bệnh nhân |
|  | RCBP.Name | Varchar | Allow | Tên của bệnh viện hoặc cơ sở y tế nơi bệnh nhân được điều trị |
|  | Gender | Varchar | Allow | Giới tính |
|  | Date.of.Birth | Date |  | Ngày sinh |
|  | Age | Int |  | Tuổi |
|  | Status.Address | Varchar |  | Tình trạng địa chỉ của bệnh nhân |
|  | Nationality | Varchar |  | Quốc tịnh bệnh nhân |
|  | Degree.of.Education | Varchar |  | Trình độ học vấn của bệnh nhân |
|  | State.Civil | Varchar |  | Tình trạng hôn nhân của bệnh nhân |
|  | Raca.Color |  |  | Chủng tộc hoặc màu da của bệnh nhân |
|  | Naturality.State | Varchar |  | Tên tiểu bang hoặc tỉnh nơi bệnh nhân sinh ra |
|  | Naturality | Varchar |  | Quốc gia nơi bệnh nhân sinh ra |
| Khóa ngoại | Code.Profession | Int |  | Mã số của nghề nghiệp của bệnh nhân |

Bảng 2. Mô tả về bảng DimPatient

DimDisease: lưu trữ thông tin của loại bệnh ung thư

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Illness.Code | Varchar | Not | Mã số của căn bệnh |
|  | Description.of.Disease | Varchar |  | Mô tả về căn bệnh |
|  | Chill.Illness.Code | Varchar |  | Mã số về căn bệnh ở trẻ em |
|  | Child.Illness.Description | Varchar |  | Mô tả về căn bệnh ở trẻ em |
|  | Code.of.Disease.Adult.Young | Varchar |  | Mã số về căn bệnh ở người trưởng thành |
|  | Young.Adult.Illness.Description | Varchar |  | Mô tả về căn bệnh ở người trưởng thành |

Bảng 3. Mô tả bảng DimDisease

DimTime: lưu trữ các thông tin về ngày tháng.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Date | date | Not | Ngày khi dữ liệu được ghi lại |
|  | Month | Int |  | Tháng khi dữ liệu được ghi lại |
|  | Year | Int |  | Năm khi dữ liệu được ghi lại |

Bảng 4. Mô tả bảng DimTime

DimMorphology: lưu trữ thông tin về hình thái của loại bệnh.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Code.of.Morphology | Int |  | Mã số về hình thái bệnh |
|  | Morphology.Description | Varchar |  | Mô tả về hình thái của căn bệnh |

Bảng 5. Mô tả bảng DimMorphology

DimTopography: lưu trữ thông tin về vị trí của bệnh.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Topography.Code | Int |  | Mã số về vị trí bệnh |
|  | Description.of.Topography | Varchar |  | Mô tả về vị trí của căn bệnh |

Bảng 6. Mô tả bảng DimTopography

FactMedical:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Tên cột | Kiểu dữ liệu | Null | Mô tả |
| Khóa chính | Id | Int | Not |  |
|  | Patient.Code | Int | Not | Mã số bệnh nhân |
|  | Topography.Code | Varchar | Not | Mã số của vị trí bệnh nhân |
|  | Code.of.Morphology | Int | Not | Mã số về hình thái bệnh |
|  | Illness.Code | Varchar | Not | Mã số của căn bệnh |
|  | Diagnostic.Means | Varchar | Allow | Phương pháp chẩn đoán của căn bệnh |
|  | Extension | Varchar | Allow | Phạn vi của căn bệnh |
|  | Laterality | Varchar | Allow | Vị trí của căn bệnh |
|  | Statement | Varchar | Allow | Tình trạng của căn bệnh |
|  | TNM | Int | Not | Mã số về tình trạng ung thư(tumor,nodes,metastasis) |
|  | Status.Vital | Varchar | Allow | Tình trạng sống còn của bệnh nhân |
|  | Type.of.Death | Varchar | Allow | Loại tử vong của bệnh nhân |
|  | Date.of.Death | Date | Allow | Ngày tử vong của bệnh nhân |
|  | Date.of.Last.Contract | Date | Allow | Ngày liên lạc cuối cùng với bệnh nhân |
|  | Date.of.Diagnostic | Date | Allow | Ngày chuẩn đoán căn bệnh |
|  | Distant.Metastasis | Varchar | Allow | Tình trạng di căn căn bệnh |
|  | Indicator.of.Race.Case | Binary | Allow | Chỉ số về trường hợp hiến gặp |

Bảng 7 Mô tả bảng FactMedical

# TÍCH HỢP DỮ LIỆU VÀO KHO

## Tạo kho dữ liệu bằng SQL Server

Đặt tên kho dữ liệu và tùy biến kích thước kho dữ liệu. Rồi nhấn **OK.**

A screenshot of a computer

Description automatically generated

Hình 2. Đặt tên và tùy biến kích thước kho dữ liệu

A screenshot of a computer

Description automatically generated with medium confidence

Hình 3. Các tùy chọn cần thiết

## Tạo Project trong Visual Studio

Chọn công cụ **Integration Services Project** chọn đường dẫn để lưu project, đặt tên và nhấn OK. Giao diện quá trình SSIS của công cụ BI hiện ra.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generatedHình 5. Đặt tên cho Project SSIS

Hình 4. Giao diện chính của SSIS

## Tạo bảng Dim và Fact

### Tạo bảng DimPatient

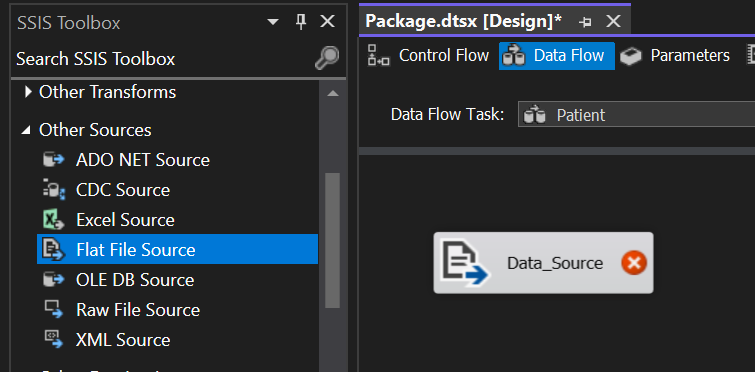
Tại Control Flow chọn Data Flow Task, đổi tên thành Patient, sau đó chuột phải chọn edit.

A screenshot of a computer

Description automatically generated

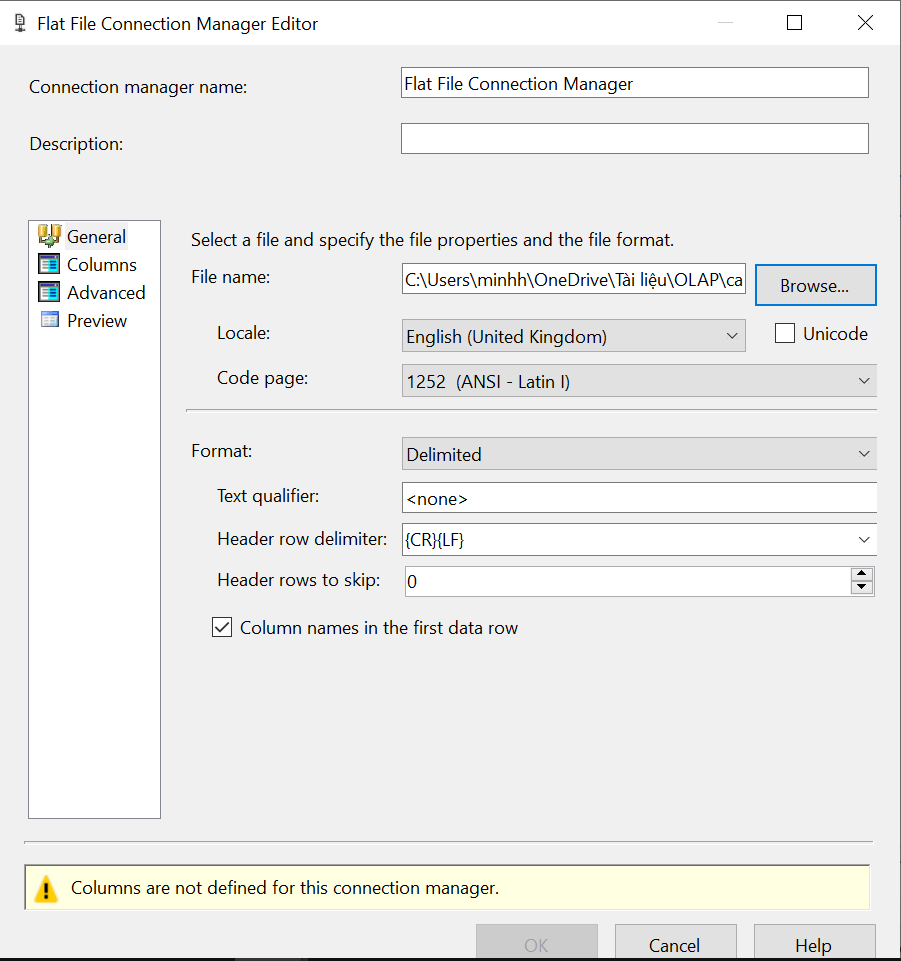
Hình 6. Kéo Data Flow Task vào màn hình làm việc và đổi tên thành Pateint

Ở màn hình làm việc qua tab **Data Flow.** Tìm kiếm chức năng Flat File Source kéo vào màn hình làm việc. Rồi đổi tên thành **Data\_Source**



Hình 7. Kéo Flat File Source vào môi trường làm việc của Data Flow

Nhấn đúp chuột 2 lần vào Flat File Source vừa kéo sau đó nhấn New, tiếp tục dẫn đến đường dẫn file CSV đã xử lý rồi nhấn OK.



Hình 8. Chọn đường dẫn đến file csv

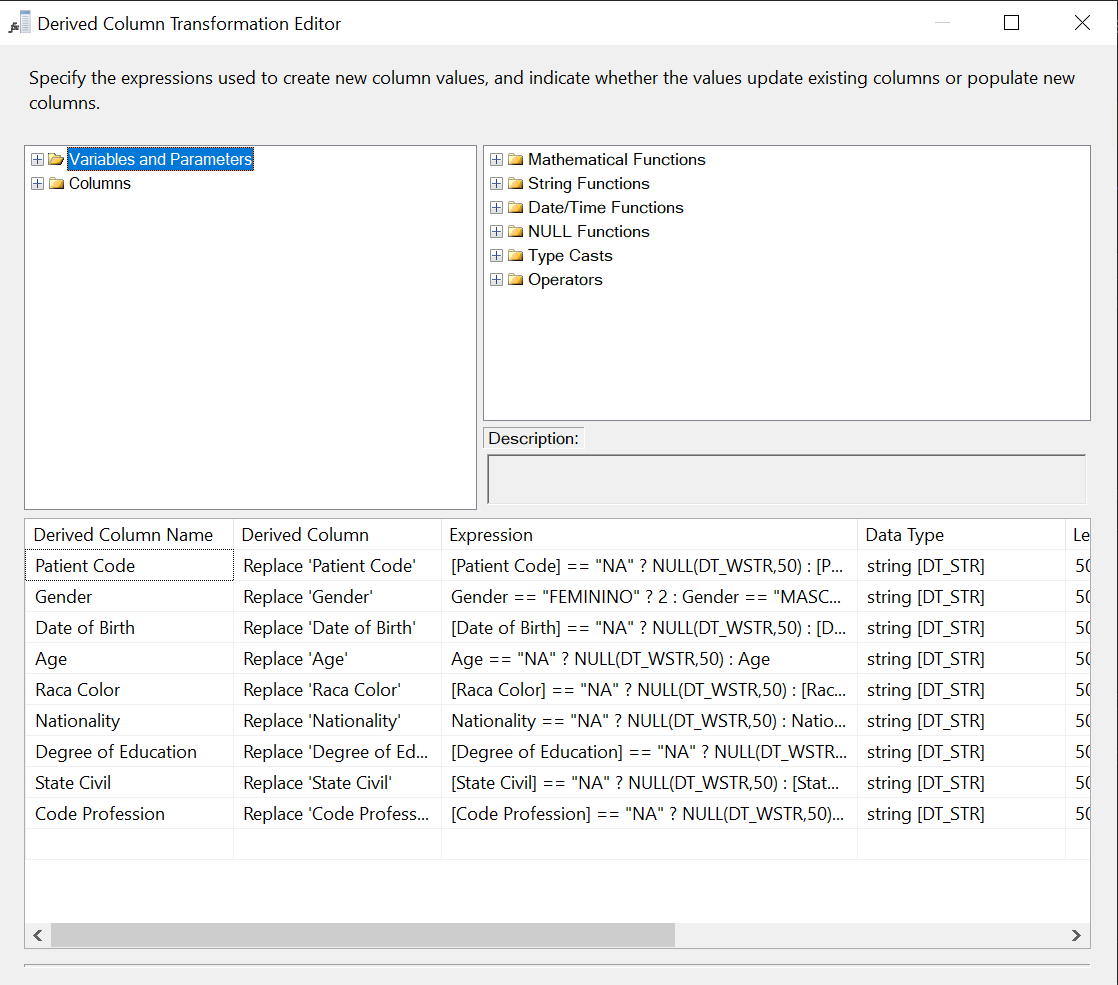
Chọn các thuộc tính cho bảng DimPatient

A screenshot of a computer

Description automatically generated

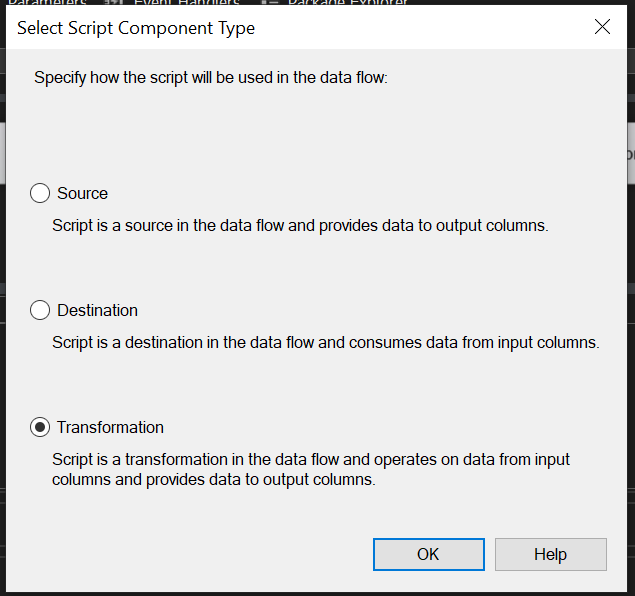
Hình 9. Chọn thuộc tính cho DimPatient

Sử dụng công cụ Derived Column để chuyển các giá trị NA thành NULL và chuyển các ký tự Gender thành số.



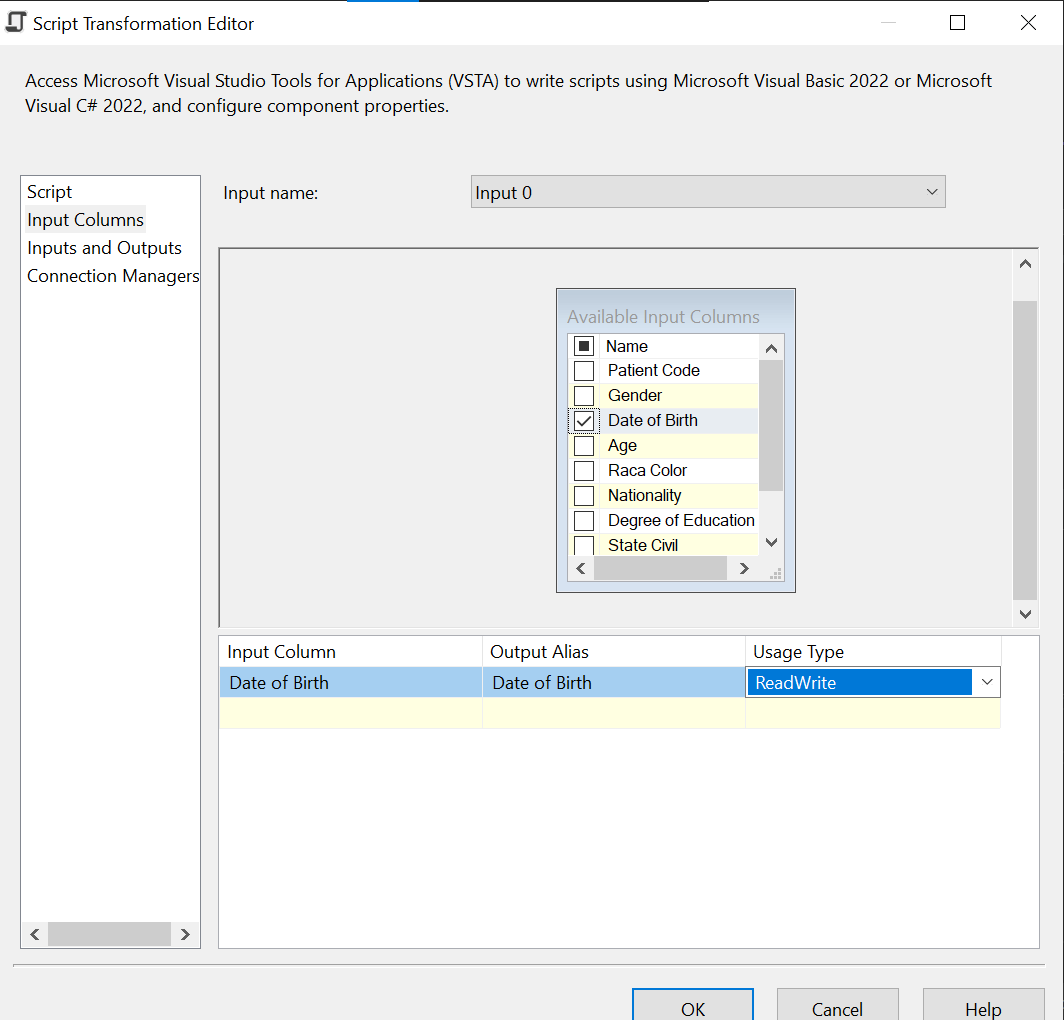
Hình 10. Chuyển đổi các giá trị

Sử dụng công cụ Script Component để loại bỏ các giá trị không thuộc định dạng ngày tháng năm của thuộc tính Date of Birth



Hình 11. Chọn Transformation

Chọn cột input là Date of Birth và chọn Usage Type là ReadWirte



Hình 12. Chọn input colum

Chọn Edit Script và thực hiện viết code

A screenshot of a computer program

Description automatically generated with medium confidence

Hình 13. Đoạn code để bỏ các dòng khống đúng định dạng ngày

Dùng công cụ Data Conversion để đổi kiểu dữ liệu thuộc tính và copy sang cột mới.

A screenshot of a computer

Description automatically generated

Hình 14. Copy sang cột mới và đổi kiểu dữ liệu

Sử dụng Conditional Split để loại bỏ các giá trị Null, Age bằng 0

A screenshot of a computer

Description automatically generated

Hình 15. Đặt điều kiện cho Case 1

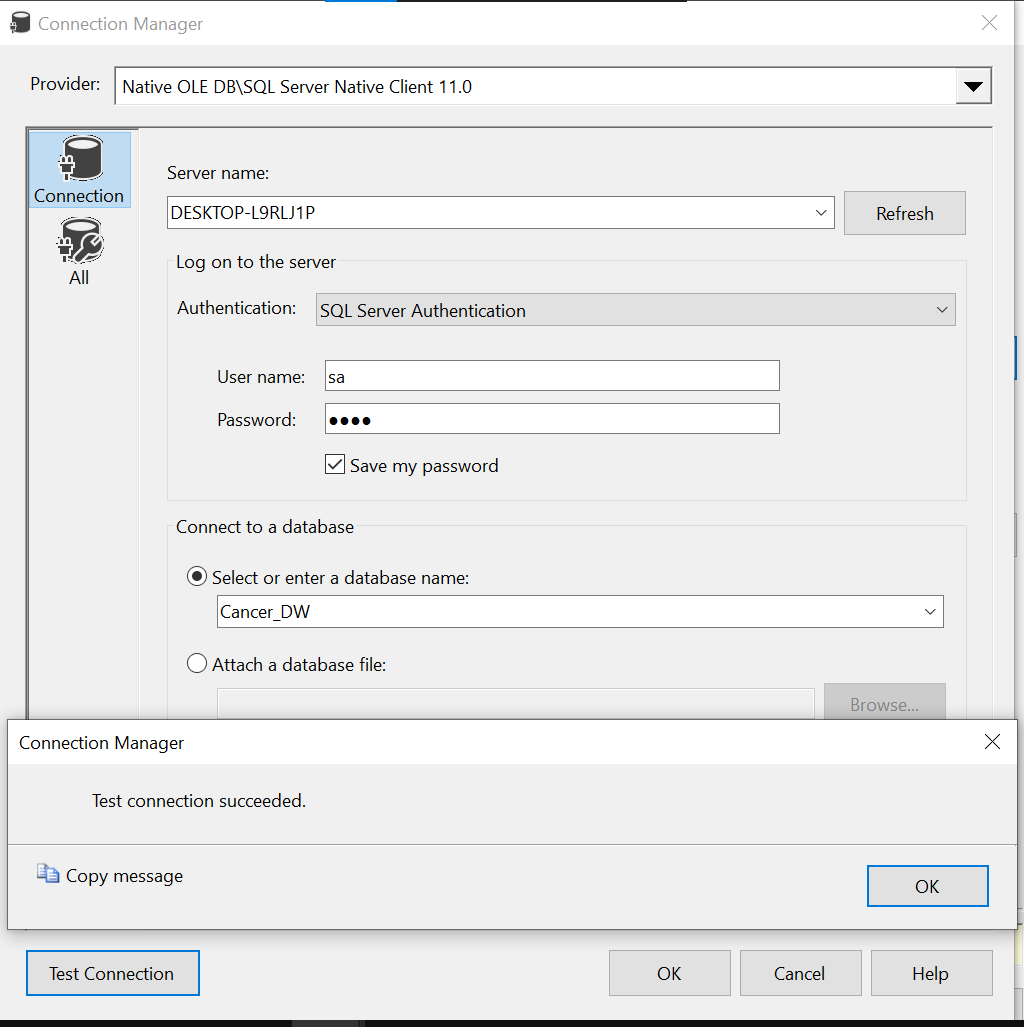
Sort dùng để sắp xếp các cột **Copy of Patient Code** và bỏ các giá trị trùng lặp.

A screenshot of a computer

Description automatically generated

Hình 16. Sắp sếp cột Copy of Patient Code tăng dần và bỏ các giá trị trùng lặp với Copy of Patient Code

Kéo công cụ OLE DB Destination sau đó đổi tên thành DimPatient. Nhấp chuột vào Sort kéo mũi tên màu xanh qua DimPatient. Nhấp chuột trái **DimPatient**. Chọn **New** tiếp tục sau đó chọn **New** để tạo một liên kết mới. Chọn **Service name** trong phần connect to database name chọn kho dữ liệu (Cancer\_DW) ban đầu đã tạo như hình minh họa.



Hình 17. Kết nối với kho dữ liệu Cancer\_DW

Tạo bảng DimPatient với [Patient Code] làm khóa chính

A screenshot of a computer program

Description automatically generated with medium confidence

Hình 18. Tạo bảng DimPatient vào kho dữ liệu

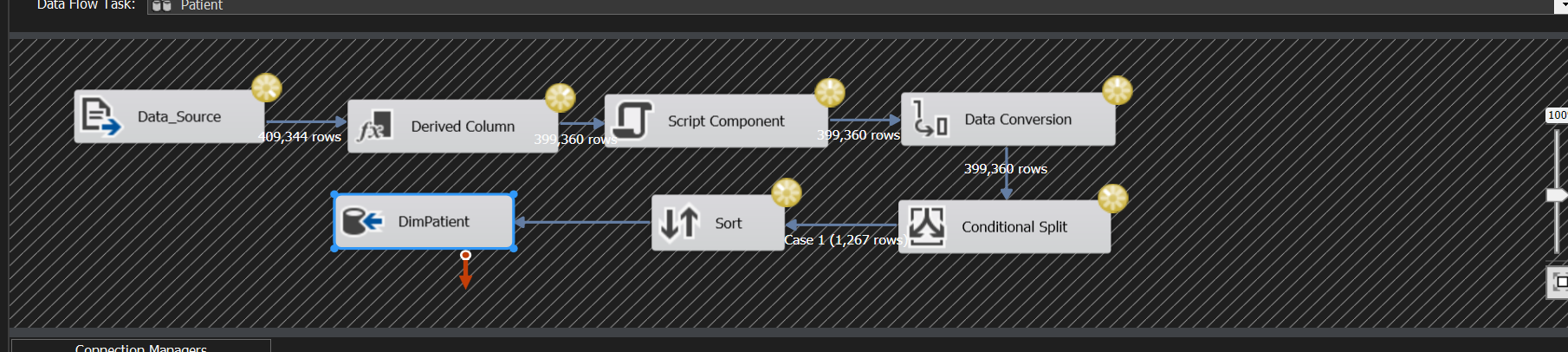
Qua phần **Mappings** nhấn **OK** nhấn **Start** để bắt đầu quá trình đổ dữ liệu. Khi đỗ dữ liệu thành công ta sẽ có kết quả như hình.

A screenshot of a computer

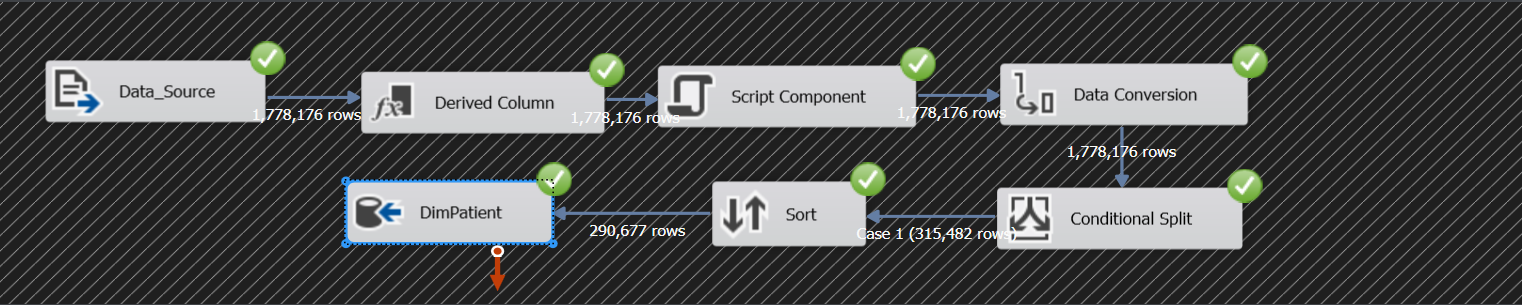
Description automatically generated

Hình 19. Quá trình Mappings dữ liệu bảng DimPatient

Nhấn Start để thực hiện quá trình đổ dữ liệu vào DimPAtient trong kho dữ liệu

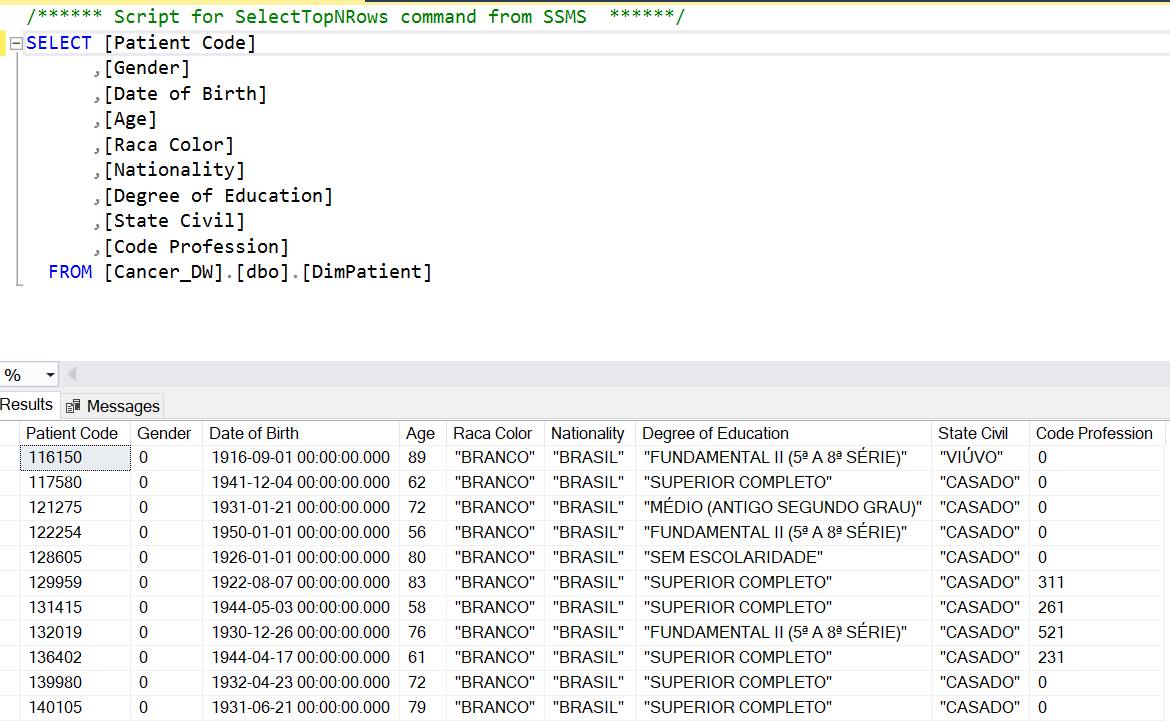


Hình 20. Quá trình đổ dữ liệu vào DimPatient



Hình 21. Đổ dữ liệu thành công

SQL Server truy vấn xem dữ liệu đã được đổ vào chưa. Kết quả dữ liệu đã đổ như hình minh họa



Hình 22. Kiểm tra dữ liệu đã đổ thành công

Quay lại tab **Control Flow** kéo **Execute SQL Task** vào màn hình làm việc sau đó nối mũi tên của **Execute SQL Task** sang **Patient** như hình minh họa.

A screenshot of a computer

Description automatically generated

Hình 23. Kết nối Execute SQL Task với Sequence Container

Nhấn chuột trái vào Execute SQL Task. Viết câu lệnh SQLStatement để ngăn ngừa việc trùng khi đổ dữ liệu vào những lần sau.

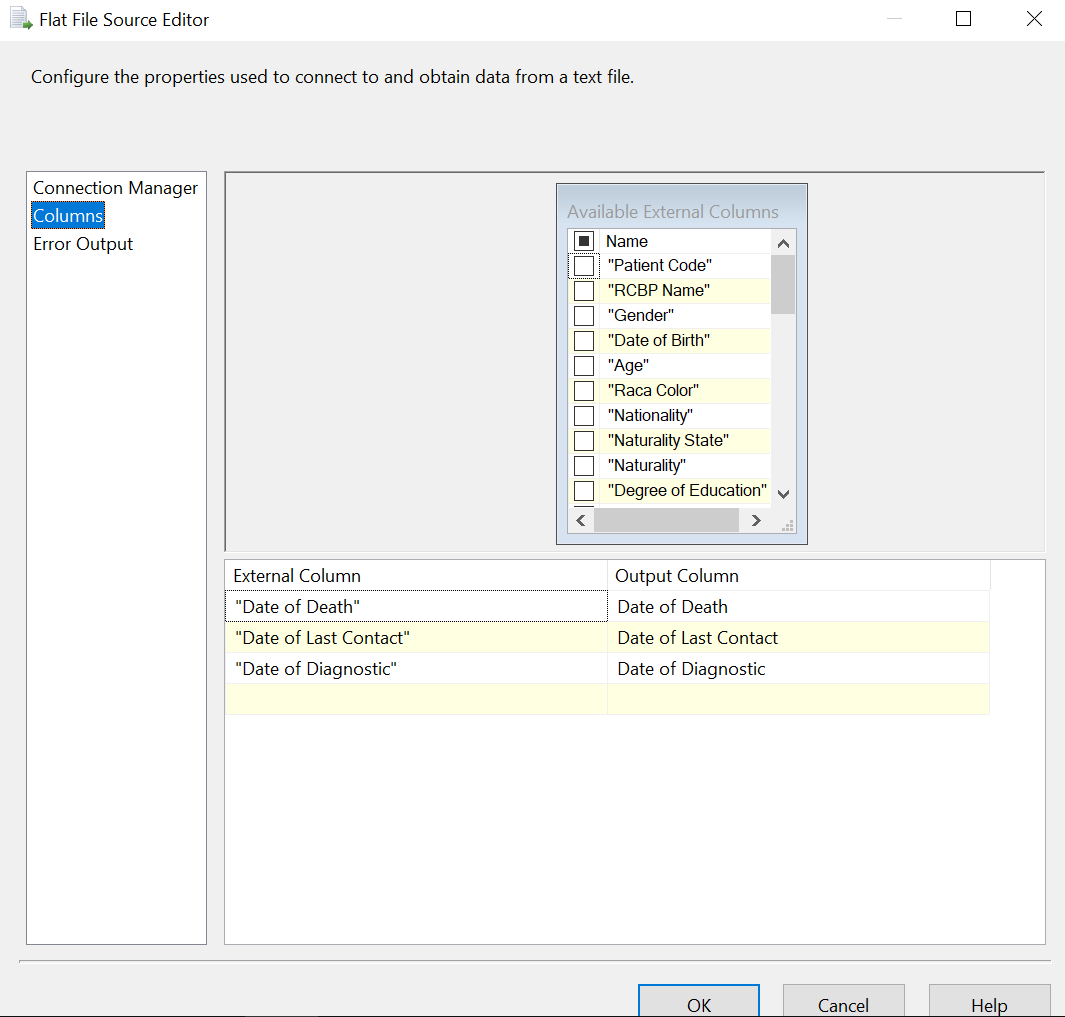
A screenshot of a computer

Description automatically generated

Hình 24. Viết câu lệnh loại bỏ trùng sau mỗi lần đổ dữ liệu của bảng DimPatient

### DimTime

Tương tự như như Patient, chúng ta cũng kéo Flat File Source vào và thêm đường dẫn vào file CSV.



Hình 25. Chọn các thuộc tính cho bảng DimTime

Sử dụng Script Component để gộp dữ liệu 3 cột vào 1 cột Date

A screenshot of a computer program

Description automatically generated with medium confidence

Hình 26. Đoạn code để gộp dữ liệu vào cột Date

Sử dụng Sort để sắp ngày tăng dần và xóa dữ liệu trùng lặp

A screenshot of a computer

Description automatically generated

Hình 27. Sắp xếp ngày theo Date và xóa trùng lặp

Derived Column dùng để thêm 2 cột mới là month và year từ Date.

A screenshot of a computer

Description automatically generated

Hình 28. Thêm 2 cột mới Month và Year

Dùng OLE DB Distinetion để tạo bảng DimTime.

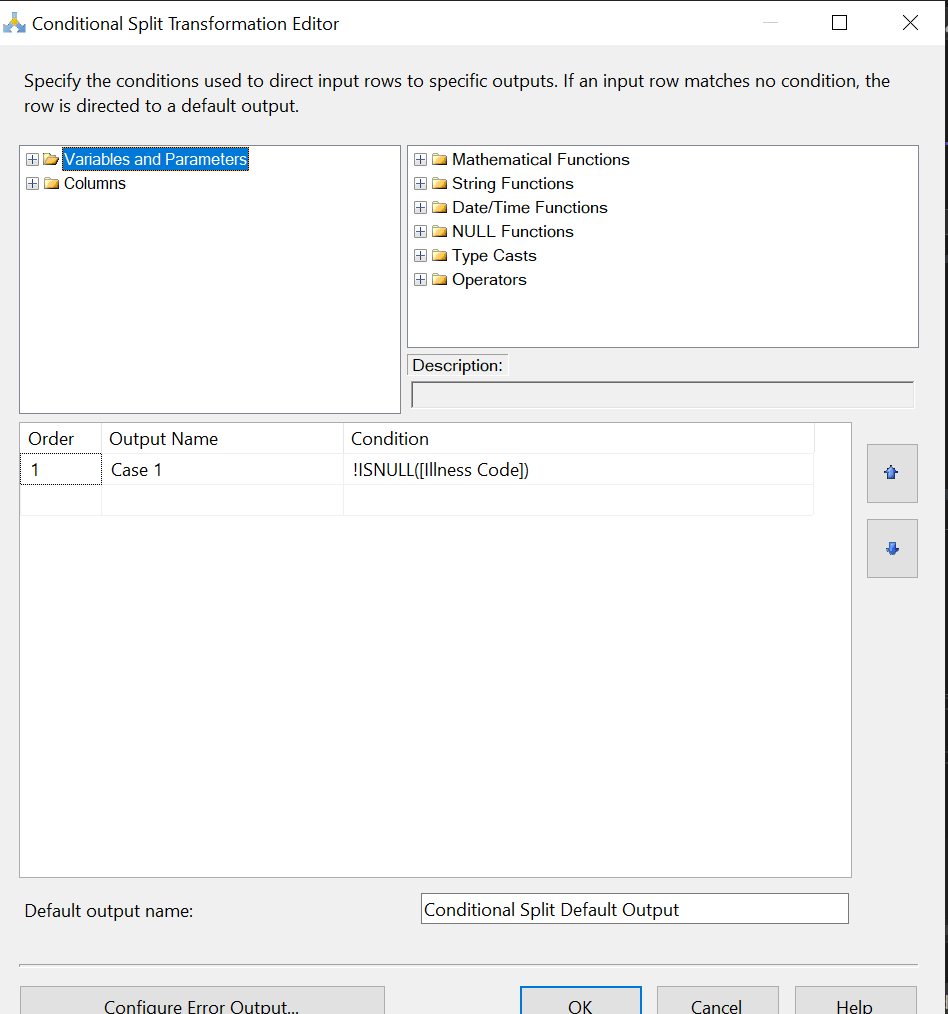
A screenshot of a computer

Description automatically generated

Hình 29. Tạo bảng DimTime

### Tạo bảng DimDisease

Tương tự như trên thêm Flat File Source và chọn đường dẫn đến CSV. Dùng Conditional Split để bỏ các giá trị Conditional Split để lọc các giá trị NULL.



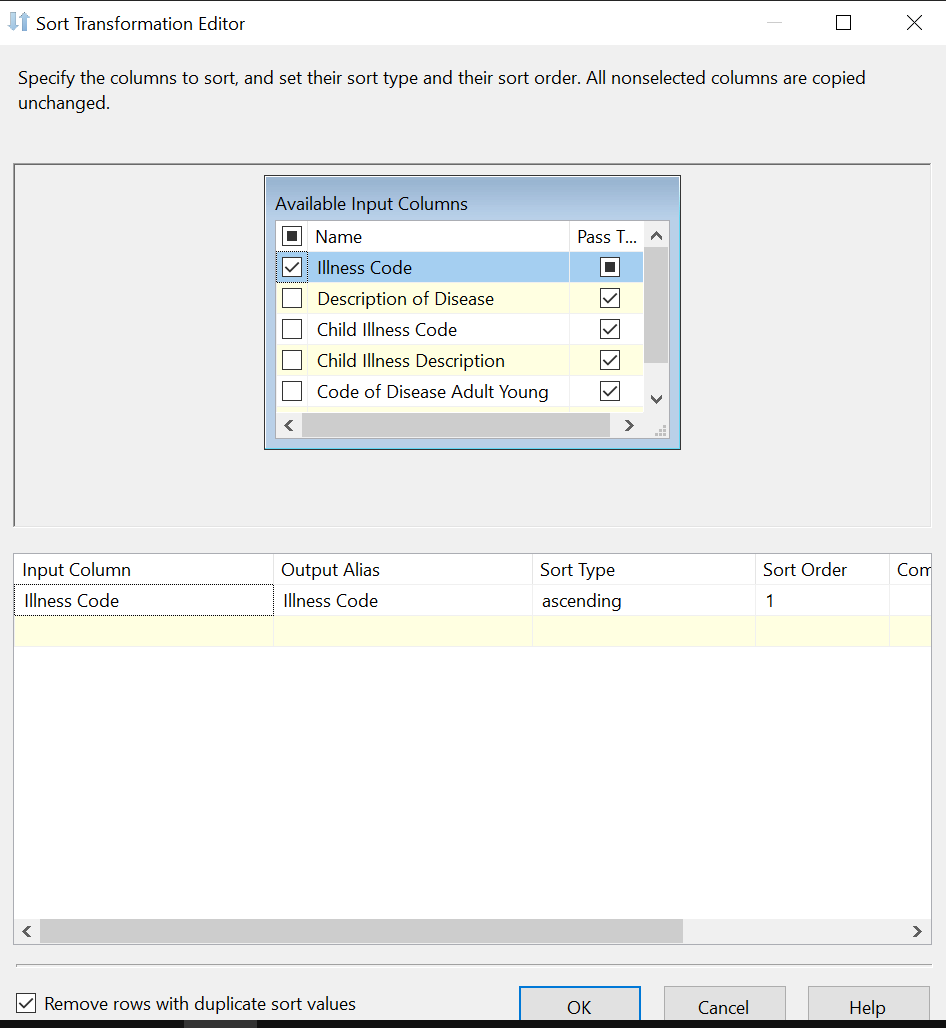
Hình 30. Case 1 lọc giá trị null của IllnessCode

Derived Column chuyển các giá trị NA thành NULL

A screenshot of a computer

Description automatically generated

Hình 31. Derived Column chuyển giá trị NA thành NULL



Hình 32. Sắp xếp theo Illness Code

A screenshot of a computer

Description automatically generated

Hình 33. Tạo bảo DimDisease

### Tạo bảng DimMorphology

Chọn các thuộc tính cho bảng DimMorphology

A screenshot of a computer

Description automatically generated

Hình 34. Chọn các thuộc tính cho bảng DimMorphology

Conditional Split lọc các dữ liệu null của thuộc tính.

A screenshot of a computer

Description automatically generated

Hình 35. Case 1 lọc các giá trị null

A screenshot of a computer

Description automatically generated

Hình 36. Sort theo Code of Morphology và xóa trùng lặp

A screenshot of a computer

Description automatically generated

Hình 37. Tạo bảng DimMorphology

### Tạo bảng DimTopography

A screenshot of a computer

Description automatically generated

Hình 38. Chọn các thuộc tính cho bảng DimTopography

Conditional Split lọc các giá trị null.

A screenshot of a computer

Description automatically generated

Hình 39. Case 1 Lọc các giá trị null

A screenshot of a computer

Description automatically generated

Hình 40. Sort theo Topography và xóa trùng lặp

A screenshot of a computer

Description automatically generated

Hình 41. Tạo bảng DimTopography vào kho dữ liệu

### Tạo bảng FactMedical

A screenshot of a computer

Description automatically generated

Hình 42. Chọn các thuộc tính cho bảng FactMedical

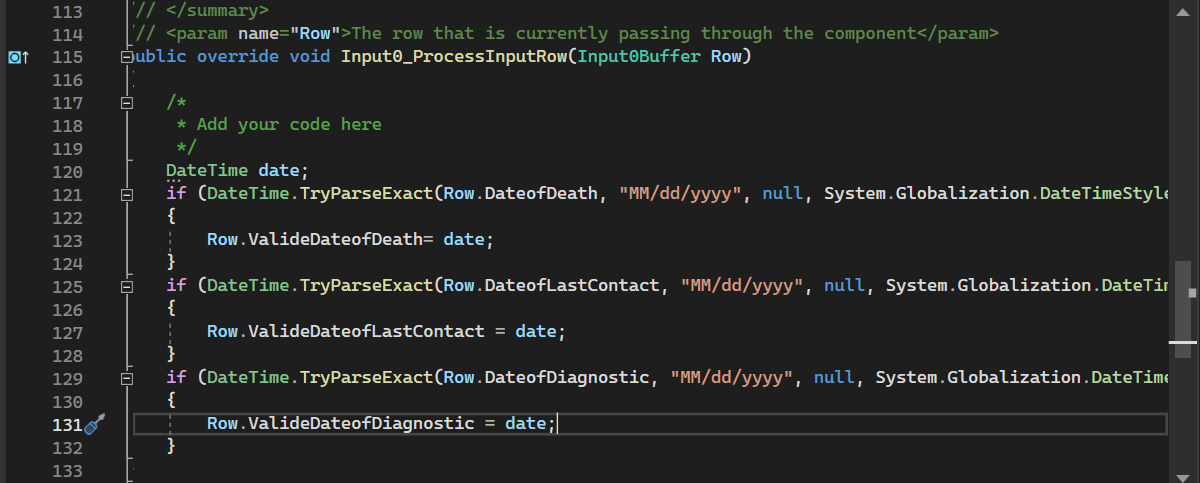
Sử dụng Derived Column để đổi các giá trị NA thành NULL.

A screenshot of a computer

Description automatically generated

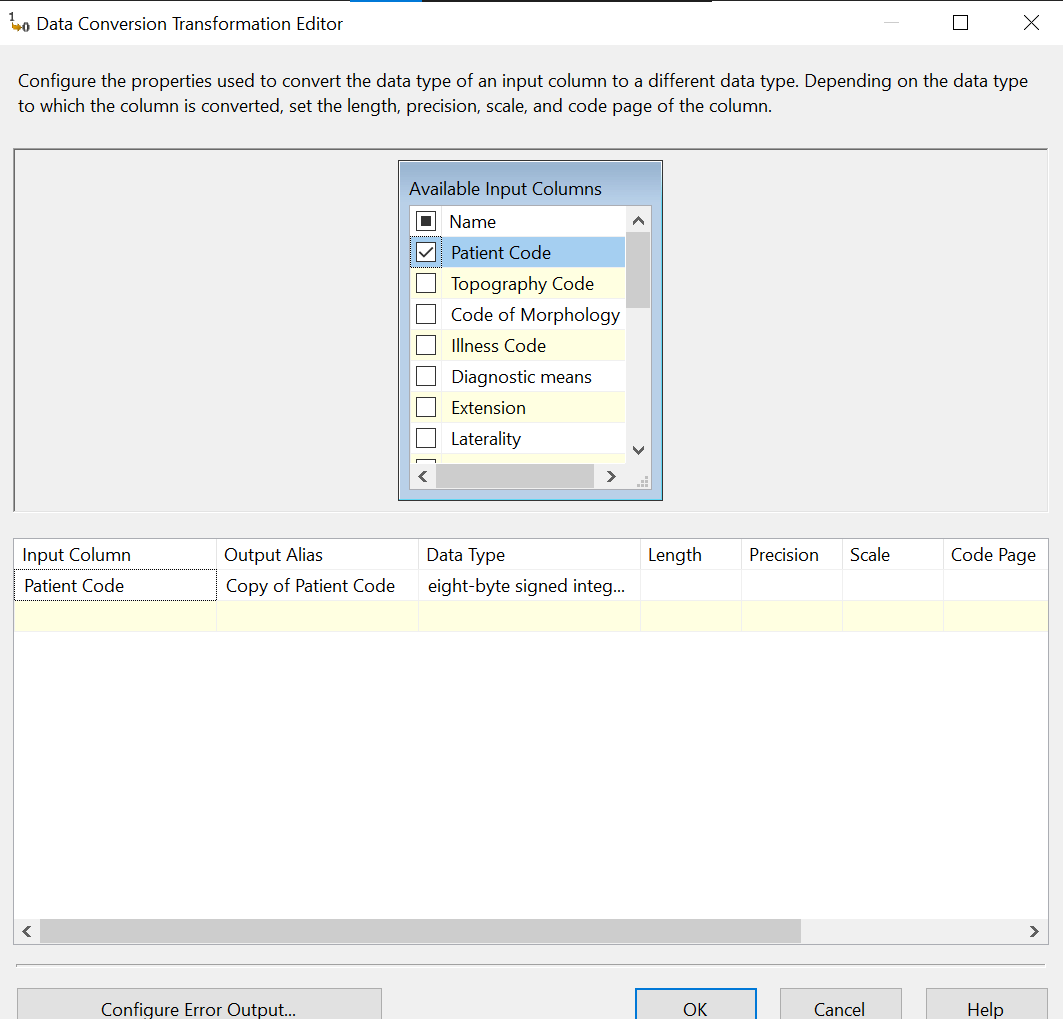
Hình 43. Thay giá trị "NA" thành NULL

Sử dụng Script Component để chuyển các giá trị Date of Death, Date of Last Contact, Date of Diagnostic thành đúng định dạng ngày.



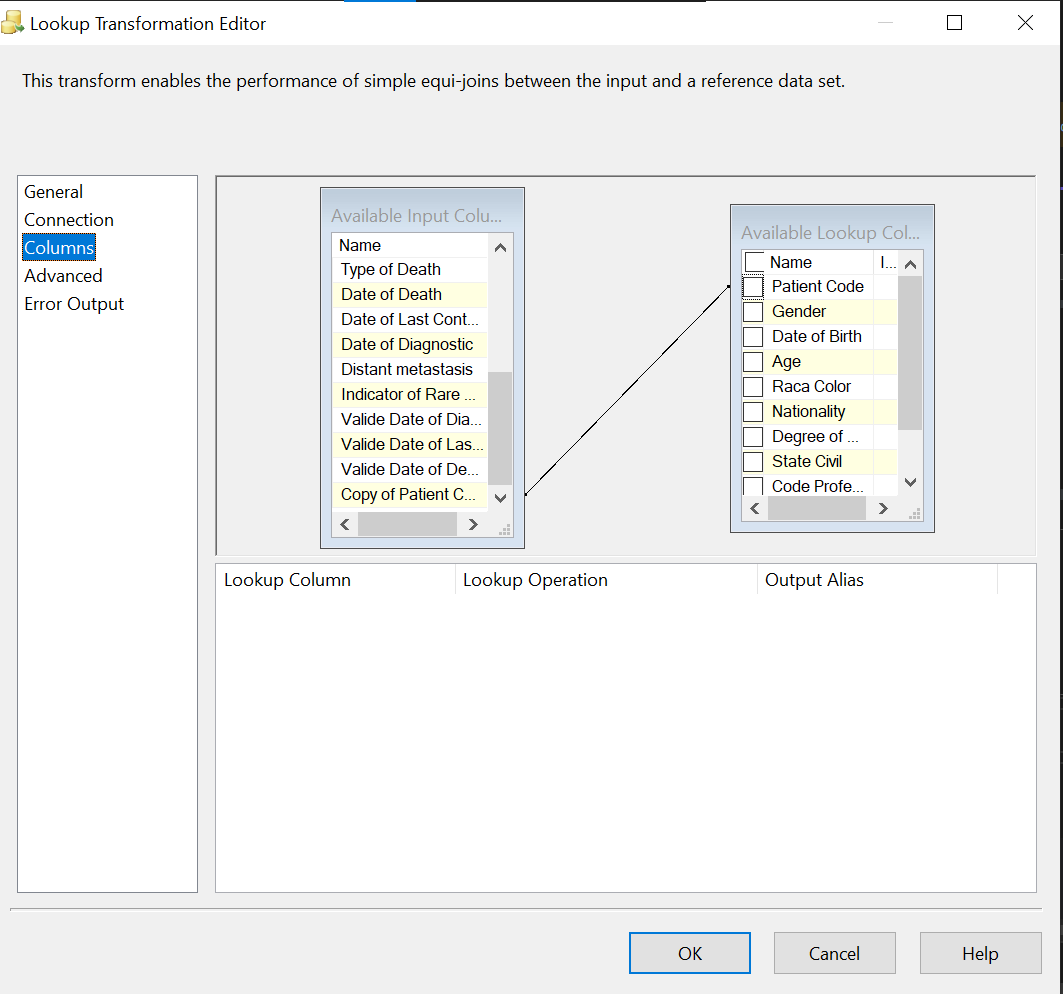
Hình 44. Đoạn code để đổi các giá trị ngày đúng định dạng và copy sang các cột khác

Copy dữ liệu của Patient Code và đổi kiểu dữ liệu sang int vào cột mới của bảng Fact



Hình 45. Copy thuộc tính Patient Code sang cột mới và chuyển sang kiểu dữ liệu int

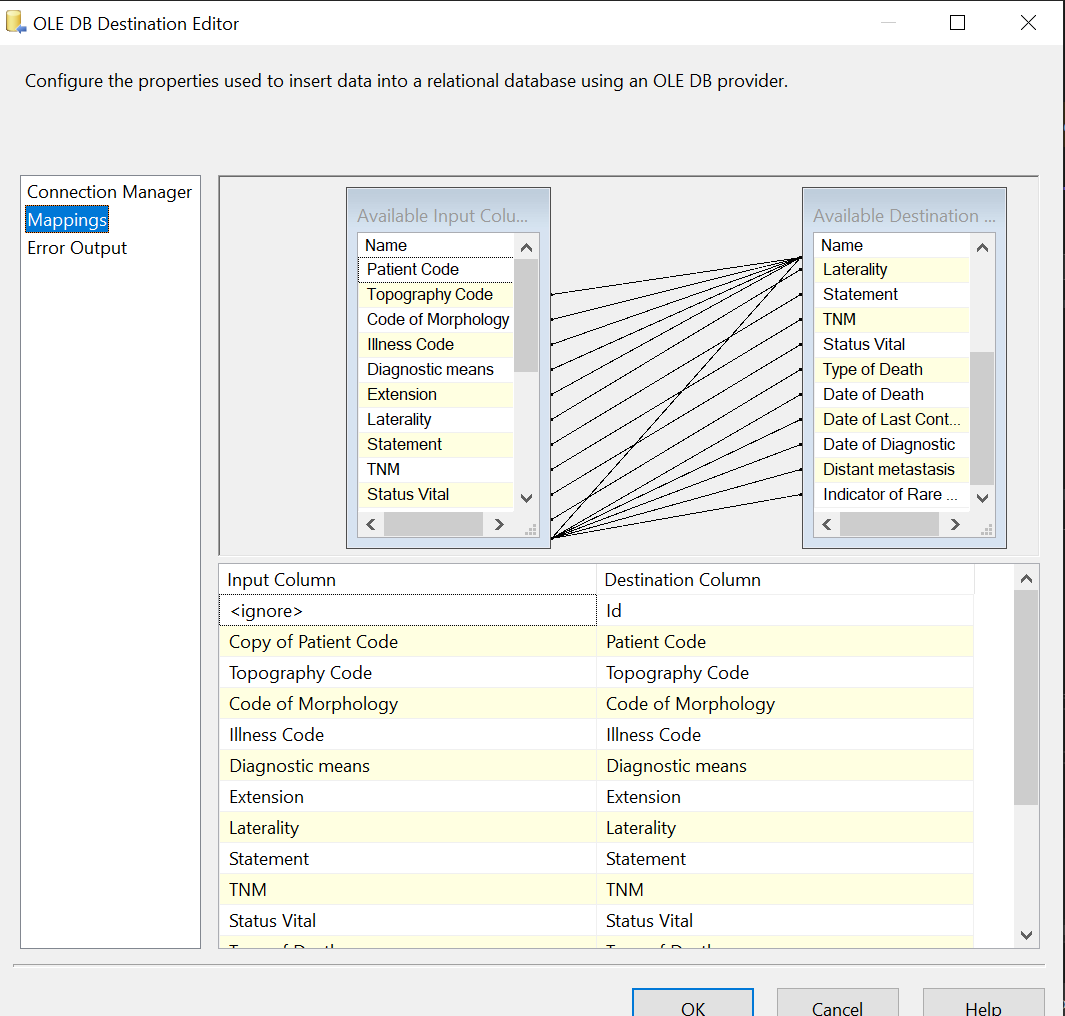
Sử dụng LookUp để lọc các giá trị không trùng với các bảng Dim



Hình 46. So sánh giữa Copy of Patient Code của bảng FactMedical với PatientCode của DimPatient

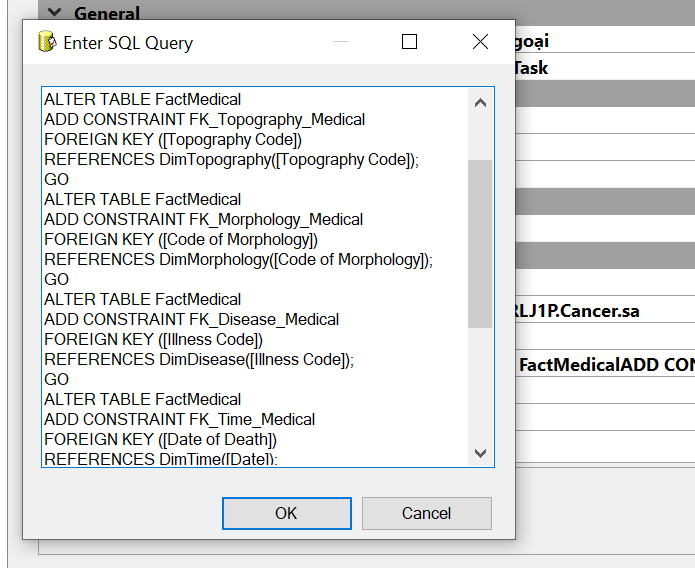
Tương tự so sánh với bảng DimDisease, DimTopography và Dimorphology

Tạo bảng FactMedical với id là khóa chính tăng tự động.



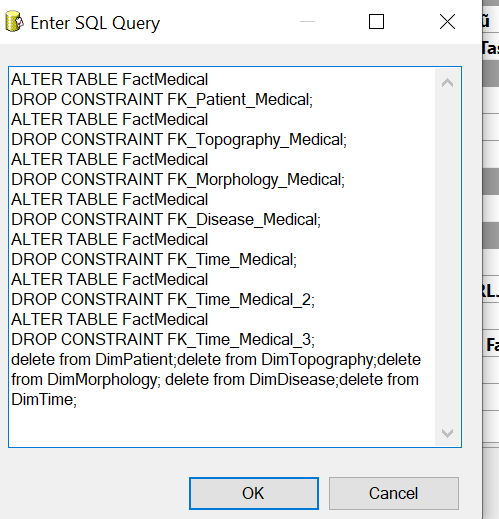
Hình 47. Tạo bảng FactMedical

Thêm Execute SQL Task để thêm các khóa ngoại cho các bảng



Hình 48. Thêm khóa ngoại

Thêm câu lệnh xóa khóa ngoại trước khi xóa bảng Dim



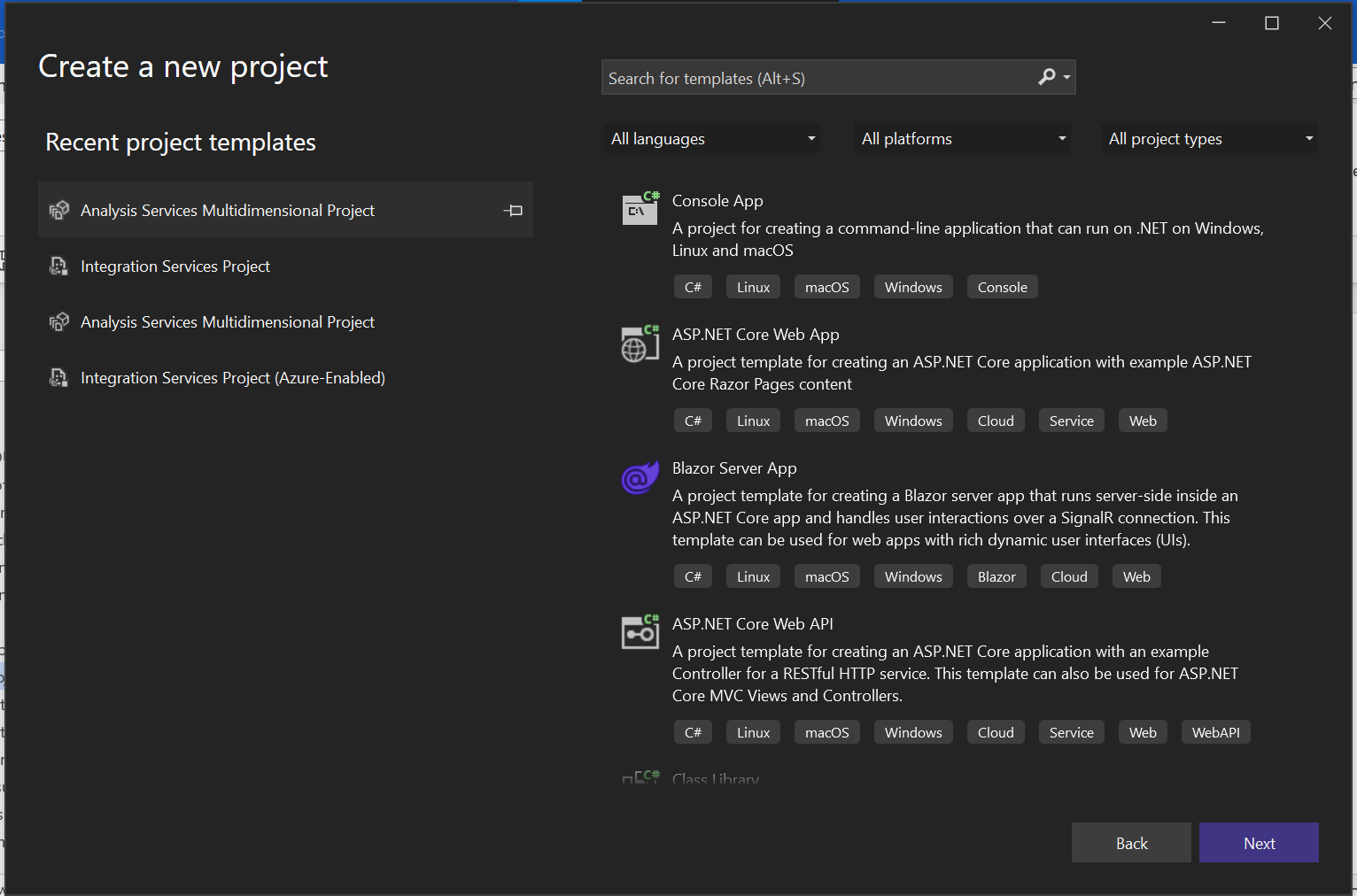
Hình 49. Xóa khóa ngoại

# PHÂN TÍCH DỮ LIỆU TRỰC TUYẾN – QUÁ TRÌNH SSAS

## Cấu hình project SSAS

### Tạo project

Tạo một project Analysis Services



Sau đó đặt tên project và chọn nơi lưu project

### Tạo Data Sources

* Chuột phải chọn New Data Source, sau đó chọn Next

A screenshot of a computer

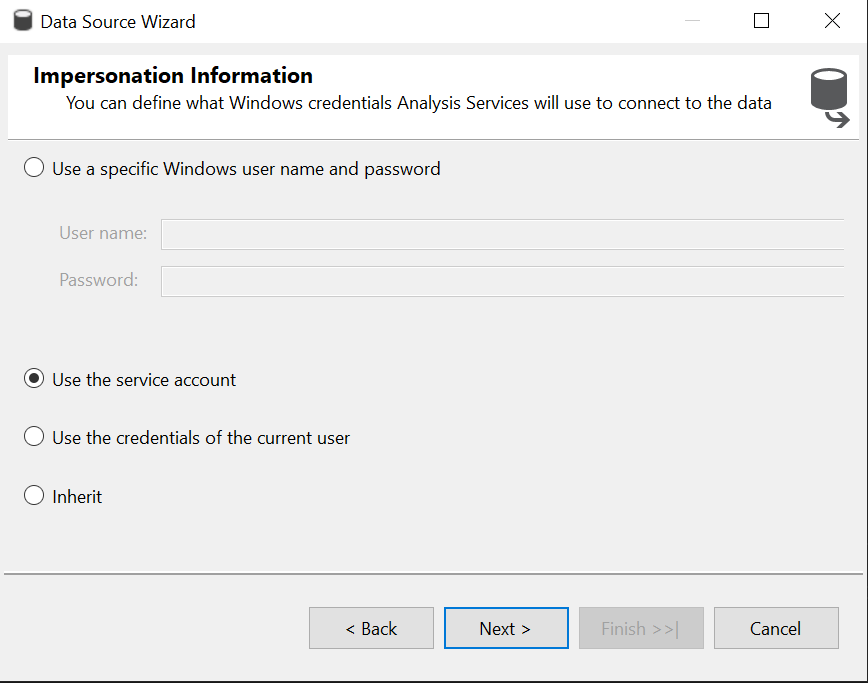
Description automatically generated with medium confidence

* Chọn kết nối đã tạo từ quá trình SSIS rồi nhấn Next

A screenshot of a computer

Description automatically generated

* Chọn Use service account



* Sau đó chọn finish

A screenshot of a computer

Description automatically generated

### Tạo Data Sources View

* Chuột phải vào Data Sources Views và chọn New Data Sources Views, sau đó nhấn Next

A screenshot of a computer

Description automatically generated with medium confidence

* Chọn Data Source Cancer mới tạo ở trên rồi nhấn Next

A screenshot of a computer

Description automatically generated with medium confidence

* Chọn bảng FactMedical và chọn mũi tên trên cùng

A screenshot of a computer

Description automatically generated with medium confidence

* Sau đó chọn Add Related Tables để chuyển những bảng khác có liên kết với FactMedical qua

A screenshot of a computer

Description automatically generated with medium confidence

* Chọn Finish để kết thúc quá trình tạo Data Source View

A screenshot of a computer

Description automatically generated

### Xây dựng các Cube và các Measure

* Chuột phải và Cube và chọn New Cube, sau đó chọn Next

A screenshot of a computer

Description automatically generated

* Chọn Use existing tables, sau đó chon Next

A screenshot of a computer

Description automatically generated with medium confidence

* Chọn Fact để phân chia các Measure

A screenshot of a computer

Description automatically generated with low confidence

* Chọn Finish để hoàn thành quá trình tạo Cube

A screenshot of a computer

Description automatically generated with medium confidence

### Chỉnh sửa Dimensions

* Bảng DimTime
* Kéo những thuộc tính cần phân cấp qua cửa sổ hierachies

A screenshot of a computer

Description automatically generated

Các bảng Dim khác cũng tương tự

### Process Cube

* Click chuột phải vào Cancer.cube -> Chọn Process xuất hiện thông báo ta chọn Yes, được như hình sau:
* Tiếp tụt chọn Run, kết quả chạy như hình sau là đã thành công phần tạo Cube

A screenshot of a computer

Description automatically generated with medium confidence

* Process Cube thành công

A screenshot of a computer

Description automatically generated with medium confidence

## Quá trình phân tích dữ liệu bằng ngôn ngữ MDX

### Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177

Câu lệnh MDX:

SELECT NON EMPTY [Measures].[Fact Medical Count] ON COLUMNS,

NON EMPTY([Dim Time].[Year].CHILDREN)

ON ROWS FROM [Cancer DW]

WHERE [Dim Disease].[Illness Code].&[177];

Kết quả câu truy vấn:

A screenshot of a computer

Description automatically generated with low confidence

### Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012

Select {[Measures].[Fact Medical Count]} on Columns ,

{TopCount({[Dim Disease].[Illness Code].CHILDREN\*[Dim Disease].[Description Of Disease].Children\*[Dim Disease].[Child Illness Description].Children\*[Dim Disease].[Code Of Disease Adult Young].Children},1,[Measures].[Fact Medical Count])} on Rows

From [Cancer DW]

Where [Dim Time].[Hierarchy].[Year].&[2012].&[1].&[3];

A picture containing text, screenshot, line, font

Description automatically generated

### Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo

Select {[Measures].[Fact Medical Count]} on Columns ,

{Filter

({[Dim Patient].[Patient Code].Children\*[Dim Patient].[Gender].CHILDREN\*[Dim Patient].[Age].Children\*[Dim Patient].[Raca Color].Children},[Measures].[Fact Medical Count] > 10)} on Rows

From [Cancer DW]

Where [Dim Patient].[City Address].&[SÄ‚O PAULO];

A screenshot of a number

Description automatically generated with low confidence

### Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm

WITH MEMBER [Measures].[Percentage 18-40] AS

SUM(

FILTER(

[Dim Patient].[Patient Code].MEMBERS,

[Dim Patient].[Age].CurrentMember.MemberValue > 18 AND [Dim Patient].[Age].CurrentMember.MemberValue < 40)

,[Measures].[Fact Medical Count])

/ [Measures].[Fact Medical Count], Format\_String = 'Percent'

SELECT [Measures].[Percentage 18-40] ON COLUMNS,

[Dim Time].[Year].MEMBERS on rows

FROM [Cancer DW];

A screenshot of a computer

Description automatically generated with low confidence

### Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó

Select NON EMPTY{[Dim Patient].[Raca Color].children

\*[Measures].[Fact Medical Count]} on Columns,

NON EMPTY{

DrillDownLevel(

DrillDownLevel(

DrillDownLevel(

[Dim Time].[Hierarchy])

)

)

} on Rows

From [Cancer DW];

A screenshot of a computer

Description automatically generated with low confidence

### Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000

Select {[Measures].[Fact Medical Count]} on Columns,

{Generate(

[Dim Patient].[City Address].Children,

[Dim Patient].[City Address].CurrentMember \*

Filter(

{[Dim Disease].[Child Illness Code].Children

\*[Dim Disease].[Child Illness Description].Children},

[Measures].[Fact Medical Count]>1000

)

)

} on Rows

From [Cancer DW];

A picture containing text, screenshot, number, font

Description automatically generated

### Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown

Select NON EMPTY{[Dim Time].[Year].Members} on Columns,

{Except(

{[Dim Patient].[Degree Of Education].Children},

{[Dim Patient].[Degree Of Education].[All].UNKNOWNMEMBER,

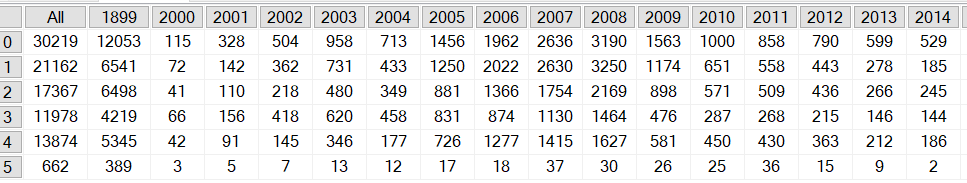
[Dim Patient].[Degree Of Education].&[6]}

)

} on Rows

From [Cancer DW]

Where {[Measures].[Fact Medical Count]};



### Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013

Select [Measures].[Fact Medical Count] on Columns,

{TopCount(

{[Dim Disease].[Illness Code].Children},

3,

[Measures].[Fact Medical Count]

)

} on Rows

From [Cancer DW]

Where {[Dim Time].[Year].&[2013]};

A screenshot of a computer

Description automatically generated with medium confidence

### Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần

SELECT NON EMPTY {[Measures].[Fact Medical Count]} ON COLUMNS,

NON EMPTY ORDER ([Dim Disease].[Illness Code].CHILDREN,[Measures].[Fact Medical Count] , desc) ON ROWS

FROM [Cancer DW]

WHERE ([Dim Topography].[Topography Code].&[132]);

A screenshot of a computer

Description automatically generated with medium confidence

### Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất

Select {[Measures].[Fact Medical Count]} on Columns,

{Union(

TopCount(NonEmpty({[Dim Patient].[City Address].children}),1,[Measures].[Fact Medical Count]),

BottomCount(NonEmpty({ [Dim Patient].[City Address].children}),1,[Measures].[Fact Medical Count])) } on rows

from [Cancer DW];

A screenshot of a computer

Description automatically generated with low confidence

### Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000

Select {[Measures].[Fact Medical Count]} on Columns,

{Intersect(

{Filter([Dim Disease].[Illness Code].children,[Measures].[Fact Medical Count] >= 1000)}

,{Filter([Dim Disease].[Illness Code].children,[Measures].[Fact Medical Count] <= 10000)}

)} on Rows

From [Cancer DW];

A screenshot of a computer

Description automatically generated with medium confidence

### Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất

Select {[Measures].[Fact Medical Count]} on Columns,

{Generate(

[Dim Patient].[City Address].Children,

TopCount(

{[Dim Patient].[City Address].CurrentMember

\* [Dim Disease].[Illness Code].Children}

,3

,[Measures].[Fact Medical Count]

)

)} on Rows

From [cancer DW];

A screenshot of a data

Description automatically generated with low confidence

## Quá trình phân tích dữ liệu trên bằng công cụ SSAS trên các khối CUBE

### Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177

A screenshot of a computer

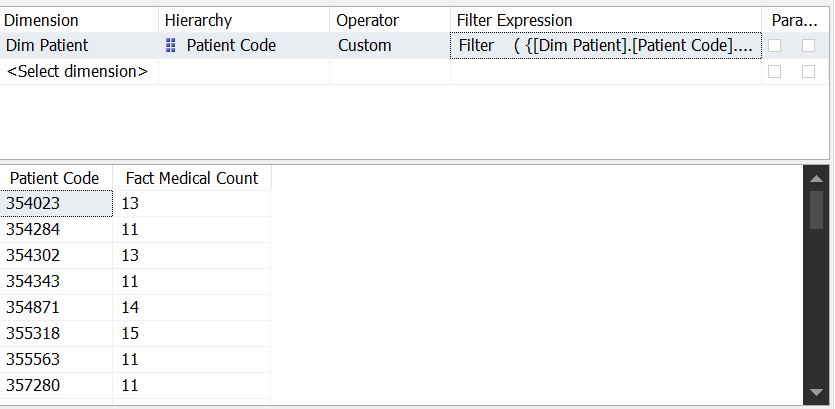
Description automatically generated with medium confidence

### Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012

A screenshot of a computer

Description automatically generated with medium confidence

### Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo



### Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm

A screenshot of a computer

Description automatically generated

### Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó

A screenshot of a table

Description automatically generated with low confidence

### Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000

A screenshot of a computer

Description automatically generated with medium confidence

### Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown

A screenshot of a computer

Description automatically generated with medium confidence

### Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013

A screenshot of a computer

Description automatically generated with medium confidence

### Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần

A screenshot of a computer

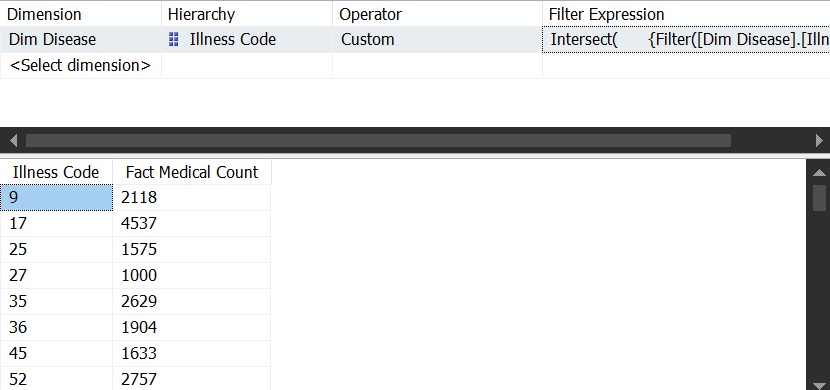
Description automatically generated with medium confidence

### Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất

A screenshot of a computer

Description automatically generated with low confidence

### Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000



### Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất



## Quá trình phân tích dữ liệu trên bằng công cụ Privot Excel

### Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177

A screenshot of a computer

Description automatically generated with medium confidence

### Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012

|  |  |
| --- | --- |
| Hierarchy | 3 |
|  |  |
| **Row Labels** | **Fact Medical Count** |
| 292 | 145 |
| **Grand Total** | **145** |

### Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo

A screenshot of a computer

Description automatically generated with medium confidence

### Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm

A picture containing text, screenshot, font

Description automatically generated

### Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó

A screenshot of a computer

Description automatically generated with medium confidence

### Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000

A picture containing text, screenshot, font, number

Description automatically generated

### Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown

A screenshot of a computer

Description automatically generated with medium confidence

### Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013

A screenshot of a computer

Description automatically generated with medium confidence

### Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần

A screenshot of a computer

Description automatically generated with medium confidence

### Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất

A screenshot of a computer

Description automatically generated with medium confidence

### Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000

A picture containing text, screenshot, font

Description automatically generated

### Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất

A screenshot of a computer

Description automatically generated

## Quá trình phân tích dữ liệu trên bằng công cụ Power BI

### Đếm số bệnh nhân ung thư chết qua từng năm với mã bệnh là 177

A screen shot of a graph

Description automatically generated with low confidence

### Cho biết Mã bệnh, mô tả bệnh, mô tả bệnh ở trẻ em, mô tả bệnh ở người lớn có nhiều người chết vì bệnh đó nhất trong tháng 3 năm 2012

A screenshot of a computer

Description automatically generated with low confidence

### Thống kê các bệnh nhân có số lượng điều trị lớn hơn 10 lần tại São Paulo

A picture containing text, screenshot, font, number

Description automatically generated

### Tính tỷ lệ phần trăm bệnh nhân có độ tuổi từ 18 đến 40 đã mất theo năm

A picture containing text, screenshot, line, plot

Description automatically generated

### Theo từng tháng, quý, năm chủng tộc và số lượng người mắc trong thời gian đó

A screenshot of a computer

Description automatically generated with medium confidence

### Cho biết mã bệnh ở trẻ em, mô tả bệnh của nó và tổng số lượng bệnh của nó theo từng thành phố lớn hơn 1000

A screenshot of a computer screen

Description automatically generated with low confidence

### Với từng trình độ học vấn liệt kê số lượng người bị ung thư qua từng năm trừ những người không hoàn thành bậc học và các giá trị unknown

A screenshot of a graph

Description automatically generated with low confidence

### Top 3 căn bệnh ung thư nhiều người bị nhất trong năm 2013

A screenshot of a computer

Description automatically generated with medium confidence

### Thông kê các bệnh ở tuyến yên theo thứ tự giảm dần

A screenshot of a graph

Description automatically generated with low confidence\

### Lấy ra 2 thành phố có số người bị ung thư cao nhất và thấp nhất

A blue square on a white background

Description automatically generated with medium confidence

A screenshot of a computer

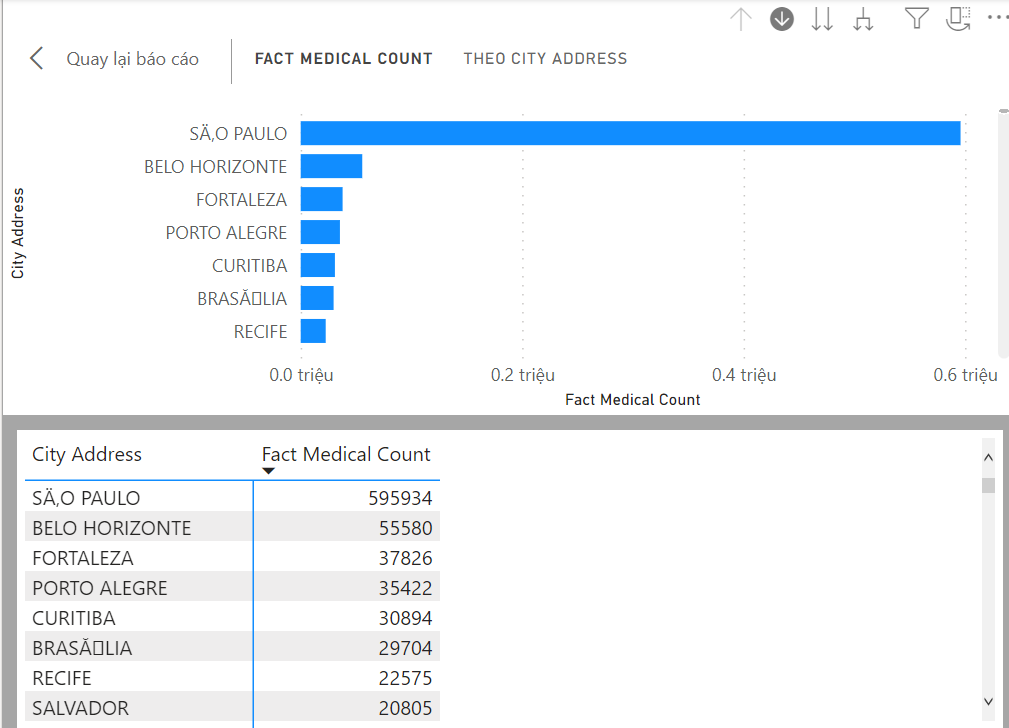
Description automatically generated

### Thông kê số bệnh nhân theo căn bệnh từ 1000 đến 10000

A screenshot of a graph

Description automatically generated with medium confidence

### Mỗi thành phố đưa ra 3 loại bệnh mà nhiều người mắc nhất



# QUÁ TRÌNH KHAI THÁC DỮ LIỆU - DATA MINING

## Tiền xử lý dữ liệu

* Import những thư viện cần thiết cho quá trình mining và Đọc dữ liệu từ file

A screenshot of a computer screen

Description automatically generated with medium confidence

* Xóa những thuộc tính có quá nhiều giá trị null và các thuộc tính có kiểu dữ liệu là object

A screen shot of a computer

Description automatically generated with medium confidence

* Loại bỏ các dữ liệu ở thuộc tính Type.of.Death bằng 1 vì nó là các giá trị Nan

A picture containing text, screenshot, clock, font

Description automatically generated

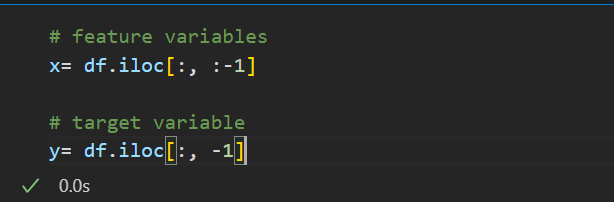
* Thống kê biến phụ thuộc

A picture containing text, screenshot, font

Description automatically generated

Ta thấy bộ dữ liệu đang mất cân bằng về 2

* Tách các biến độc lập và các biến phụ thuộc



* Kiểm tra thuộc tính có ảnh hưởng cao đến biến phụ thuộc

A screen shot of a computer

Description automatically generated with medium confidence

* Loại bỏ các thuộc tính có độ ảnh hưởng thấp

A picture containing screenshot, text, font

Description automatically generated

* Kiểm tra dộ tương quan

A screen shot of a computer code

Description automatically generated with low confidence

A picture containing screenshot, text, square, rectangle

Description automatically generated

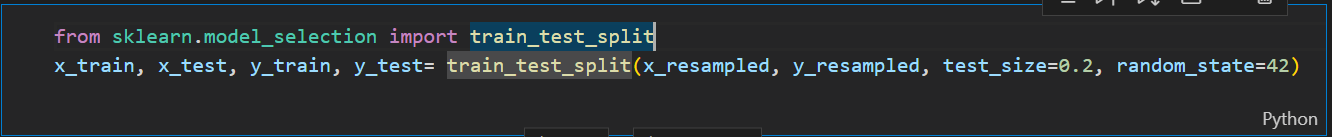
Ta thấy không có cặp dữ liệu nào tương quan quá cao

* Xử lý mất cân bằng dữ liệu

A picture containing text, screenshot, software, display

Description automatically generated

* Chia các tập train-test theo tỷ lệ 8-2



## RandomForest

* Khởi tạo mô hình với 100 cây quyết định

A screen shot of a computer

Description automatically generated with low confidence

* Huấn luyện mô hình trên tập huấn luyện và in ra kết quả đánh giá mô hình trên cả tập huấn luyện và tập kiểm tra.

A screen shot of a computer program

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with low confidence

Kết quả trên tập kiểm tra:

* Độ chính xác (accuracy) của mô hình trên tập kiểm tra là 81.93%.
* Báo cáo phân loại (classification report) cho thấy độ chính xác của lớp 0 là 73.97%, độ chính xác của lớp 2 là 98.02%.

Tổng quan, mô hình phân loại Random Forest có độ chính xác khá cao trên cả tập huấn luyện và tập kiểm tra, nhưng độ chính xác trên tập kiểm tra thấp hơn một chút so với tập huấn luyện, có thể là do overfitting.

* Độ ảnh hưởng của các biến độc lập

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, line, diagram

Description automatically generated

Ta thấy thuộc tính Child.Illness.Code có độ ảnh hưởng cao nhất với thuộc tính phụ thuộc

* Ma trận nhầm lẫn

**A picture containing text, screenshot, diagram, colorfulness

Description automatically generated**

Ma trận nhầm lẫn (confusion matrix) cho thấy có 155088 mẫu thuộc lớp 0 được phân loại đúng, 54607 mẫu thuộc lớp 0 bị phân loại sai, 101871 mẫu thuộc lớp 2 được phân loại đúng và 2059 mẫu thuộc lớp 2 bị phân loại sai.

## Naive Bayes

* Xây dựng mô hình trên tập dữ liệu train và thực hiện dự đoán

A screen shot of a computer program

Description automatically generated with medium confidence

Dựa trên kết quả đánh giá, mô hình phân loại Naive Bayes Gaussian của chúng em có độ chính xác (accuracy) là 79.34% và F1-score là 79.95%. Điều này cho thấy mô hình của bạn có khả năng phân loại tốt trên tập dữ liệu kiểm tra.

* Tính toán ma trận nhầm lẫn và biểu diễn ra đồ thị

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

Sau khi kiểm tra trên tập dữ liệu Test chúng em thấy:

TP (true positive) – mẫu mang nhãn 0 (Chết vì bệnh ung thư) được phân lớp đúng vào lớp 0 (Chết vì bệnh ung thư) là 151106

FN (false negative) – mẫu mang nhãn 0 bị phân lớp sai vào lớp 2 (Chết không phải vì bệnh ung thư) là 6041

FP (false positive) – mẫu mang nhãn 2 bị phân lớp sai vào lớp 0 là 58738

TN (true negative) – mẫu mang nhãn 2 được phân lớp đúng vào lớp 2 là 97740.

## Nhận xét

Xét 2 thuộc tính có dộ ảnh hưởng cao nhất

A screenshot of a graph

Description automatically generated with medium confidence

Bảng 8. Thống kê mã bệnh ở người lớn

Với mã bệnh ở người lớn là 58 (khối u không xác định rõ nguồn gốc) thì số bệnh bênh chết vì ung thư là cao nhất và cũng là mã bệnh ở người lớn có số lượng người tử vong cao nhất

A picture containing text, screenshot, diagram, plot

Description automatically generated

Bảng 9. Thống kê mã bệnh

Với mã bệnh ở trẻ em là 49 (các khối u ác tính khác không được chỉ định rõ ràng) thì số bệnh bênh chết vì ung thư là cao nhất và cũng là mã bệnh ở trẻ em có số lượng người tử vong cao nhất.

A picture containing diagram, text, line, plan

Description automatically generated

Những người có khối u không xác định rõ nguồn gốc và các khối u ác tính khác không được chỉ định rõ ràng thì sẽ có nguy cơ tử vong rất cao.

# TÀI LIỆU THAM KHẢO

1. SQL Server 2012 Tutorials: Analysis Services - Multidimentional ModelingFile

2. <https://www.distilled.net/google-data-studio-tutorial/>

3. Slide bài giảng và hướng dẫn thực hành do giảng viên cung cấp

4. Tài liệu Seminar “Quá trình SSIS”, tác giả: Dương Quốc Cường

5. Tài liệu Seminar “Quá trình SSAS”, tác giả: Dương Quốc Cường

6. Tài liệu Seminar “Quá trình data mining”, tác giả: Đỗ Hữu Lượng

7. SQL Server 2018 Tutorials: Analysis Services - Multidimentional ModelingFile

8. Tài liệu thiết kế kho dữ liệu <http://repository.vnu.edu.vn/bitstream/VNU_123/6559/1/V_L0_01959.pdf>