**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

# BÁO CÁO ĐỒ ÁN THỰC HÀNH MÔN HỌC KHO DỮ LIỆU VÀ OLAP

**ĐỀ TÀI:**

## XÂY DỰNG KHO DỮ LIỆU VÀ OLAP CHĂM SÓC SỨC KHỎE

Nhóm sinh viên thực hiện:

1.Nguyễn Thị Thanh Thúy MSSV: 22521457

2. Hoàng Xuân Thủy MSSV: 22521456

**TP. HỒ CHÍ MINH, NĂM 2024**

## NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....

*……, ngày……..tháng……năm 2024*

**Người nhận xét**

*(Ký tên và ghi rõ họ tên)*

# LỜI CẢM ƠN

*Kính thưa quý thầy cô,*

*Trước hết, thay mặt toàn thể thành viên nhóm, em xin gửi lời cảm ơn chân thành và sâu sắc đến quý thầy cô trường Đại học Công nghệ Thông tin đã tạo điều kiện thuận lợi để chúng em hoàn thành tốt đồ án. Đặc biệt, em xin được bày tỏ lòng biết ơn sâu sắc đến cô Nguyễn Thị Kim Phụng và chị Ngô Thùy Yến Nhi, người đã tận tình hướng dẫn và hỗ trợ nhóm trong suốt quá trình thực hiện. Nhờ sự tận tâm, nhiệt huyết và những kiến thức chuyên môn sâu rộng của cô cùng chị, chúng em đã không ngừng học hỏi và hoàn thiện bản thân.*

*Bên cạnh đó, em cũng xin gửi lời cảm ơn đến tất cả các thành viên trong nhóm. Mỗi thành viên đều đã đóng góp một phần không nhỏ vào thành công của dự án này. Sự hợp tác, chia sẻ và tinh thần trách nhiệm của các bạn đã tạo nên một tập thể đoàn kết và hiệu quả.*

*Quá trình thực hiện đồ án "Xây dựng kho dữ liệu và Olap Chăm sóc sức khỏe" là một trải nghiệm vô cùng quý báu. Chúng em đã được rèn luyện nhiều kỹ năng mềm và cứng, từ việc thu thập dữ liệu, phân tích, thiết kế đến trình bày. Tuy nhiên, chúng em cũng nhận thức được rằng sản phẩm còn nhiều hạn chế. Chúng em rất mong nhận được những góp ý quý báu từ quý thầy cô để hoàn thiện hơn nữa.*

*Cuối cùng, chúng em xin chúc cô và chị luôn mạnh khỏe, hạnh phúc và gặt hái nhiều thành công trong sự nghiệp trồng người."*

# MỤC LỤC

LỜI CẢM ƠN ..... 3
MỤC LỤC ..... 4
DANH MỤC HÌNH ẢNH ..... 9
DANH MỤC BẢNG ..... 17
CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU ..... 18
1. LÝ DO CHỌN ĐỀ TÀI ..... 18
2. PHÁT BIỂU VỀ DỮ LIỆU ..... 18
2.1. Mô tả dữ liệu ..... 18
2.2. Thuộc tính của kho dữ liệu ..... 19
2.3. Kho dữ liệu đã xử lý ..... 20
3. XÂY DỰNG KHO DỮ LIỆU ..... 22
3.1. Sơ đồ bông tuyết minh họa ..... 22
3.2. Dim_Location ..... 23
3.3. Dim_Date ..... 23
3.4. Dim_ Doctor ..... 23
3.5. Dim_Insurance _Provider ..... 23
3.6. Dim_Patient ..... 24
3.7. Dim_Medication ..... 24
3.8. Dim_Test_Result ..... 25
3.9. Fact ..... 25
4. Nội dung 15 câu truy vấn ..... 26
CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS) ..... 27
1. CHUẨN BỊ CÔNG CỤ VÀ DATAWAREHOUSE ..... 27
2. TẠO PROJECT SSIS TRONG VISUAL STUDIO 2022 ..... 32
3. TẠO BẢNG DIM VÀ BẢNG FACT ..... 33
3.1. DIM_LOCATION ..... 33
3.2. DIM_DOCTOR ..... 37
3.3. DIM_TEST_RESULT ..... 40
3.4. DIM_ INSURANCE_PROVIDER ..... 43
3.5. DIM_MEDICATION ..... 46
3.6. DIM_ PATIENT ..... 49

3.7. DIM_DATE ..... 52
3.8. FACT ..... 57
CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO( SSAS) ..... 79
1. TẠO PROJECT SSAS TRONG VISUAL STUDIO ..... 79
2. XÁC ĐỊNH DỮ LIỆU NGUỒN( DEFINE DATE SOURCE) ..... 80
3. XÁC ĐỊNH KHUNG NHÌN DỮ LIỆU NGUỒN( DEFINE DATE SOURCE VIEW) 84
4. XÁC ĐỊNH CÁC CHIỀU ( DEFINE A DIMENSION) ..... 88
5. XÂY DỰNG CÁC KHỐI VÀ CÁC MEASURE ..... 111
6. ĐỔI TÊN VÀ THUỘC TÍNH CÁC MEASURES ..... 116
7. PHÂN CẤP TRONG BẢNG CHIỀU ..... 116
7.1. Bảng Dim_Date ..... 116
8. CHẠY PROJECT SSAS ..... 123
9. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE ..... 125
9.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm ..... 125
9.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu. ..... 126
9.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm. 127
9.4. Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm .... 129
9.5. Đưa ra top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023. 131
9.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc top 5 ..... 133
9.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024 135
9.8. Liệt kê các bệnh nhân(mã bệnh nhân, tên bệnh nhân, giới tính, nhóm máu) có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024. .. 137
9.9. Trong mỗi loại bệnh, in ra bác sĩ có số lượng bênh nhân lớn nhất trong năm 2022 139
9.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.142
9.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể144
9.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $100000, sau đó sắp xếp theo tổng chi phí giảm dần. 147
9.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.. 149

9.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất / top 3 số lần nhập viện nhiều nhất. ..... 151

9.15. Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất. 153

10. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL154

10.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm. ..... 154

10.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu. ..... 156

10.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm. 159

10.4. Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm. ... 162

10.5. Đưa ra Top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023. 164

10.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc Top 5. ..... 167

10.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024. 170

10.8. Liệt kê các bệnh nhân có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024. ..... 173

10.9. Trong mỗi loại bệnh, in ra bác sĩ có số lượng bênh nhân lớn nhất trong năm 2022. 175

10.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.179

10.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể.181

10.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần. 183

10.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.. 186

10.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất hoặc top 3 số lần nhập viện nhiều nhất. ..... 189

10.15. Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất. 191

11. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX ..... 193

11.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm ..... 193

11.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu. ..... 194

11.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm. 195

11.4. Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm .... 196

11.5. Đưa ra Top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023. 197

11.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc Top 5. ..... 198

11.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024. 198

11.8. Liệt kê các bệnh nhân có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024. ..... 199

11.9. Trong mỗi loại bệnh, in ra bác sĩ có số lượng bệnh nhân lớn nhất trong năm 2022. 200

11.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.201

11.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể.202

11.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần. 203

11.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.. 204

11.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất hoặc top 3 số lần nhập viện nhiều nhất. ..... 205

11.15. Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất. 207

CHƯƠNG 4. REPORT VỚI POWER BI ..... 208

1. QUÁ TRÌNH LẬP BÁO BIỂU BẰNG CÔNG CỤ POWER BI..... 208

1.1. REPORT 1 ..... 208

1.2. REPORT 2 ..... 212

1.3. REPORT 3 ..... 216

1.4. REPORT 4 ..... 221

1.5. REPORT 5 ..... 224

1.6. REPORT 6 ..... 228

1.7. REPORT 7 ..... 232

1.8. REPORT 8 ..... 236

1.9. REPORT 9 ..... 240

1.10. REPORT 10 ..... 244

1.11. REPORT 11 ..... 248

1.12. REPORT 12 ..... 252

1.13. REPORT 13 ..... 256

1.14. REPORT 14 ..... 260

1.15. REPORT 15 ..... 264

CHƯƠNG 5. QUÁ TRÌNH DATAMINING ..... 269
1. Tổng quan quá trình ..... 269
2. Giới thiệu bộ dữ liệu ..... 269
3. Công cụ khai thác dữ liệu ..... 269
4. Tiền xử lý dữ liệu ..... 269
5. Khai thác dữ liệu ..... 270
5.1. Thuật toán K-Means ..... 270
5.2. Thuật toán Decision Tree và Random Forest ..... 274
6. Phân tích kết quả ..... 281
6.1. Phân tích kết quả thuật toán K-Means ..... 281
6.2. Phân tích kết quả thuật toán Decision Tree và Random Forest ..... 284
7. Đánh giá thuật toán ..... 285
7.1. Thuật toán K- Means ..... 285
7.2. Thuật toán Decision Tree ..... 286
7.3. Thuật toán Random Forest ..... 287
DANH MỤC TÀI LIỆU THAM KHẢO ..... 288

# DANH MỤC HÌNH ẢNH

Hình1. Thuộc tính của kho dữ liệu ..... 19
Hình 2 . Kho dữ liệu đã xử lý ..... 21
Hình 3. Sơ đồ bông tuyết ..... 22
Hình 4 . SQL Server Data Tools ..... 27
Hình 5 . Server ..... 27
Hình 6 . Tạo và cấu hình kho dữ liệu ..... 28
Hình 7. Tạo kho dữ liệu mới ..... 28
Hình 8.Tạo kho dữ liệu mới ..... 29
Hình 9. Tạo DW ..... 30
Hình 10. Tạo DW ..... 30
Hình 11 . Tùy chỉnh dữ liệu ..... 31
Hình 12. Configure Data Collection ..... 31
Hình 13. Hoàn thành ..... 32
Hình 14. Create Project ..... 32
Hình 15. Nhập tên cho Project ..... 33
Hình 16. Sort Dim_Location ..... 33
Hình 17 . Remove row with duplicate sort values ..... 34
Hình 18 . OLE DB Destination – Dim_Loaction ..... 34
Hình 19. Connection Manager ..... 35
Hình 20. Tạo bảng Dim_Location ..... 36
Hình 21. Mapping thuộc tính ..... 37
Hình 22. Sort Dim_Doctor ..... 37
Hình 23. Sort theo Doctor ..... 38
Hình 24. OLE DB DESTINATION -DIM_DOCTOR ..... 38
Hình 25. Tạo bảng DIM_DOCTOR ..... 39
Hình 26. Mapping thuộc tính ..... 40
Hình 27 . Kéo chức năng Sort ..... 40
Hình 28. Sort theo Test_Results ..... 41
Hình 29. OLE DB Destination- DIM_TEST_RESULT ..... 41
Hình 30. Tạo bảng DIM_TEST_RESULT ..... 42
Hình 31. Mapping thuộc tính ..... 43
Hình 32. Kéo chức năng Sort ..... 43
Hình 33. Sort theo Insurance_Provider ..... 44
Hình 34. OLE DB DESTINATION – DIM_INSURANCE_PROVIDER ..... 44
Hình 35. Tạo bảng DIM_INSURANCE_PROVIDER ..... 45
Hình 36. Mapping thuộc tính ..... 46
Hình 37 . Kéo chức năng Sort ..... 46
Hình 38 . Sort theo Medication ..... 47
Hình 39. OLE DB DESTINATION – DIM_MEDICATION ..... 47
Hình 40. Tạo bảng DIM_MEDICATION ..... 48
Hình 41. Mapping thuộc tính ..... 49
Hình 42. Kéo chức năng Sort ..... 49
Hình 43. Sort theo Name, Gender, BloodType, Medical Condition ..... 50
Hình 44. OLE DB DESTINATION – DIM_PATIENT ..... 50
Hình 45. Tạo bảng DIM_PATIENT ..... 51
Hình 46. Mapping thuộc tính ..... 52
Hình 47 . Xử lý dữ liệu Admission_Date ..... 53
Hình 48 . Xử lý dữ liệu Discharge Date ..... 54
Hình 49. Union All ..... 55
Hình 50. Sort theo day , month, year, quarter ..... 56

Hình 51. OLE DB DESTINATION – DIM_DATE ..... 56
Hình 52 . Tạo bảng DIM,_DATE ..... 57
Hình 53. Kéo chức năng Data Flow Task ..... 57
Hình 54. Flat File Source ..... 58
Hình 55. Xử lý dữ liệu ..... 58
Hình 56. Lookup ..... 59
Hình 57 . Edit Mapping ..... 60
Hình 58. Đổi tên ở Output Alias ..... 61
Hình 59. Lookup ..... 62
Hình 60. Edit Mapping ..... 63
Hình 61. Đổi tên ô Output Alias ..... 64
Hình 62. Lookup ..... 65
Hình 63 . Edit Mapping ..... 66
Hình 64. Lookup DIM_DOCTOR ..... 67
Hình 65. Edit Mapping ..... 68
Hình 66. Lookup DIM_TEST_RESULT ..... 69
Hình 67. Edit Mapping ..... 70
Hình 68. Lookup DIM_INSURANCE_PROVIDER. ..... 71
Hình 69. Edit Mapping ..... 72
Hình 70 . Lookup DIM_MEDICATION. ..... 73
Hình 71. Edit Mapping ..... 74
Hình 72. Lookup DIM_PATIENT. ..... 75
Hình 73. Edit Mapping ..... 76
Hình 74. Tạo bảng FACT ..... 77
Hình 75 . Đổ dữ liệu ..... 78
Hình 76. Tạo Excute SQL Task ..... 79
Hình 77 . Tạo Khóa ngoại ..... 79
Hình 78 . Tạo Project SSAS ..... 80
Hình 79 . Lưu project ..... 80
Hình 80. New Data Source ..... 81
Hình 81. New Data Source ..... 81
Hình 82. Kết nối database ..... 82
Hình 83 . Use the service account. ..... 83
Hình 84 . Finish ..... 84
Hình 85. New Data Source View ..... 85
Hình 86. Chọn DATASET ..... 86
Hình 87 . Chọn các bảng FACT và DIM ..... 87
Hình 88. FINISH ..... 88
Hình 89. New Dimension ..... 89
Hình 90 . Tạo Dimesion ..... 90
Hình 91. DIM_DATE ..... 91
Hình 92. Chọn thuộc tính cho DIM_DATE ..... 92
Hình 93 . Completing ..... 93
Hình 94. DIM_DOCTOR ..... 94
Hình 95. Chọn thuộc tính cho DIM_DOCTOR ..... 95
Hình 96. Hoàn thành tạo DIM_DOCTOR ..... 96
Hình 97. DIM_INSURANCE_PROVIDER ..... 97
Hình 98. Chọn thuộc tính cho DIM_INSURANCE_PROVIDER ..... 98
Hình 99 . Hoàn thành tạo DIM_INSURANCE_PROVIDER ..... 99
Hình 100. DIM_LOCATION ..... 100
Hình 101. Chọn thuộc tính cho DIM_LOCATION ..... 101
Hình 102 . Hoàn thành tạo DIM_LOCATION ..... 102

Hình 103. DIM_MEDICATION ..... 103
Hình 104. Chọn thuộc tính cho DIM_MEDICATION ..... 104
Hình 105. Hoàn thành tạo DIM_MEDICATION ..... 105
Hình 106 . DIM_PATIENT ..... 106
Hình 107. Chọn thuộc tính cho DIM_PATIENT ..... 107
Hình 108 . Hoàn thành tạo DIM_PATIENT ..... 108
Hình 109 . DIM_ TEST_RESULT ..... 109
Hình 110 . Chọn thuộc tính cho DIM_TEST_RESULT ..... 110
Hình 111 . Hoàn thành tạo DIM_TEST_RESULT ..... 111
Hình 112 . New Cube ..... 112
Hình 113 . Chọn bảng FACT ..... 113
Hình 114 . Chọn các độ đo ..... 114
Hình 115. Chọn các bảng DIM ..... 115
Hình 116 . Kết thúc quá trình tạo Cube ..... 116
Hình 117 . Phân cấp bảng DIM_DATE ..... 117
Hình 118. Phân cấp bảng DIM_DATE ..... 117
Hình 119. Đổi Name Column của Day ..... 118
Hình 120 . Phân cấp bảng DIM_DATE ..... 119
Hình 121. Đổi Name Column của Month ..... 120
Hình 122. Phân cấp bảng DIM_DATE ..... 121
Hình 123 . Đổi Name Column của Quarter. ..... 122
Hình 124. Đổi Name Column của Year. ..... 123
Hình 125. Attribute Relationship ..... 123
Hình 126. Deploy Project ..... 124
Hình 127 . Kết quả quá trình Deploy ..... 124
Hình 128. Câu truy vấn 1 ..... 125
Hình 129. Kết quả câu truy vấn 1 ..... 126
Hình 130. Câu truy vấn 2 ..... 126
Hình 131 . Filter Expression ( Abnormal ; Normal ) ..... 127
Hình 132. Kết quả câu truy vấn 2 ..... 127
Hình 133. Câu truy vấn 3 ..... 128
Hình 134 . Kết quả câu truy vấn 3 ..... 128
Hình 135 .Tạo calculated Member cho  Câu truy vấn 4 ..... 129
Hình 136 . Kéo thả các thuộc tính ..... 130
Hình 137 . Filter Expression chọn Cancer. ..... 130
Hình 138 . Design Mode sang Script Mode ..... 131
Hình 139. Kết quả câu truy vấn 4 ..... 131
Hình 140. Tạo Named Set mới cho câu truy vấn 5 ..... 132
Hình 141. Kéo thả các thuộc tính ..... 132
Hình 142. Filter Expression chọn 2023 và bộ lọc named set Cau5 ..... 133
Hình 143. Kết quả câu truy vấn 5 ..... 133
Hình 144. Tạo Named Set mới cho câu truy vấn 6 ..... 134
Hình 145. Kéo thả các thuộc tính ..... 134
Hình 146 .Filter Expression chọn 0+ ; Named Set Cau 6 ..... 135
Hình 147. Kết quả câu truy vấn 6 ..... 135
Hình 148. Tạo Named Set mới cho câu truy vấn 7 ..... 136
Hình 149. Kéo thả các thuộc tính ..... 136
Hình 150. Filter Expression chọn quý 2 của năm 2024 và chọn bộ lọc named set Cau 7 ..... 137
Hình 151 . Kết quả câu truy vấn 7 ..... 137
Hình 152 . Tạo Named Set mới cho câu truy vấn 8 ..... 138
Hình 153. Kéo thả các thuộc tính ..... 138
Hình 154 . Filter Expression chọn năm 2024 và chọn bộ lọc named set Cau 8 ..... 138

Hình 155 . Kết quả câu truy vấn 8 ..... 139
Hình 156 .Tại browser của công cụ BI Kéo thả các thuộc tính cần thiết ..... 140
Hình 157. Chuyển từ Design Mode sang Script Mode ..... 140
Hình 158 . Xóa bỏ bớt dòng code dư ..... 141
Hình 159. Dán code và Execute Query ..... 141
Hình 160. Kết quả câu truy vấn 9 ..... 142
Hình 161. Qua browser của công cụ BI kéo thả những thuộc tính cần thiết ..... 143
Hình 162. Chuyển từ Design Mode sang Script Mode ..... 143
Hình 163. Xóa bỏ bớt dòng code dư ..... 143
Hình 164. Dán code và Execute Query ..... 144
Hình 165 . Kết quả câu truy vấn 10 ..... 144
Hình 166 . Qua browser của công cụ BI kéo thả những thuộc tính cần thiết ..... 145
Hình 167. Chuyển từ Design Mode sang Script Mode ..... 145
Hình 168. Chuyển từ Design Mode sang Script Mode ..... 146
Hình 169 . Dán code và Execute Query ..... 146
Hình 170. Kết quả câu truy vấn 11 ..... 146
Hình 171. Tạo một Named Set mới Câu 12 ..... 147
Hình 172. Kéo thả các thuộc tính cần thiết ..... 147
Hình 173. Filter Expression chọn Diabetes và chọn bộ lọc named set Cau 12 ..... 148
Hình 174 175. ..... 148
Hình 175 . Chuyển từ Design Mode sang Script Mode ..... 148
Hình 176 . Thêm Named Set Cau 12 vào ..... 148
Hình 177 . Kết quả câu truy vấn 12 ..... 149
Hình 178. Tạo độ đo tính toán ..... 149
Hình 179. Kéo thả các thuộc tính cần thiết ..... 150
Hình 180 . Lọc theo năm 2023 ..... 150
Hình 181. Kết quả câu truy vấn 13 ..... 150
Hình 182. Tạo một Named Set mới ..... 151
Hình 183. Kéo thả các thuộc tính cần thiết ..... 152
Hình 184. Lọc theo Named Set Cau 14. ..... 152
Hình 185. Kết quả câu truy vấn 14 ..... 152
Hình 186. Tạo một Named Set mới ..... 153
Hình 187. Kéo thả các thuộc tính cần thiết ..... 153
Hình 188. Lọc theo Named Set Cau 15 ..... 154
Hình 189. Kết quả câu truy vấn 15 ..... 154
Hình 190. Chọn biểu tượng Excel ..... 154
Hình 191 . Chọn biểu tượng Excel ..... 154
Hình 192 . Chọn các thuộc tính cần thiết ..... 155
Hình 193 . Kết quả câu truy vấn 1 dạng Matrix ..... 155
Hình 194. Kết quả câu truy vấn 1 dạng Chart ..... 156
Hình 195 . Chọn các thuộc tính cần thiết ..... 157
Hình 196. Filter ( chọn Abnormal và Normal) ..... 158
Hình 197. Kết quả câu truy vấn 2 dạng Matrix ..... 159
Hình 198. Kết quả câu truy vấn 2 dạng Chart ..... 159
Hình 199. Chọn các thuộc tính cần thiết ..... 160
Hình 200. Kết quả câu truy vấn 3 dạng Matrix ..... 161
Hình 201. Kết quả câu truy vấn 3 dạng Chart ..... 161
Hình 202. Chọn các thuộc tính cần thiết ..... 162
Hình 203. Lọc bệnh theo Cancer ..... 163
Hình 204. Kết quả câu truy vấn 4 dạng Table ..... 163
Hình 205. Kết quả câu truy vấn 4 dạng Chart ..... 164
Hình 206. Chọn các thuộc tính cần thiết ..... 165

Hình 207. Lọc theo 2023 ..... 166
Hình 208. Kết quả câu truy vấn 5 dạng Table. ..... 166
Hình 209. Kết quả câu truy vấn 5 dạng Chart ..... 167
Hình 210. Chọn các thuộc tính cần thiết ..... 168
Hình 211. Lọc theo nhóm máu 0+ ..... 169
Hình 212. Kết quả câu truy vấn 6 dạng Table ..... 169
Hình 213. Kết quả câu truy vấn 6 dạng Chart ..... 170
Hình 214. Chọn các thuộc tính cần thiết ..... 171
Hình 215. Lọc theo quý 2 năm 2024 ..... 172
Hình 216. Kết quả câu truy vấn 7 dạng Table ..... 172
Hình 217. Kết quả câu truy vấn 7 dạng Chart ..... 173
Hình 218. Chọn các thuộc tính cần thiết ..... 174
Hình 219 . Lọc theo 2024 ..... 174
Hình 220. Kết quả câu truy vấn 8 dạng Table ..... 175
Hình 221. Chọn các thuộc tính cần thiết ..... 176
Hình 222. Lọc theo năm 2022 ..... 177
Hình 223. Kết quả câu truy vấn 9 dạng Table ..... 178
Hình 224. Kết quả câu truy vấn 9 dạng Chart ..... 178
Hình 225. Chọn các thuộc tính cần thiết ..... 179
Hình 226. Lọc theo năm 2022 ..... 180
Hình 227 . Kết quả câu truy vấn 10 dạng Table ..... 181
Hình 228. Kết quả câu truy vấn 10 dạng Chart ..... 181
Hình 229. Chọn các thuộc tính cần thiết ..... 182
Hình 230. Kết quả câu truy vấn 11 dạng Table ..... 183
Hình 231. Kết quả câu truy vấn 11 dạng Chart ..... 183
Hình 232. Chọn các thuộc tính cần thiết ..... 184
Hình 233. Lọc theo căn bệnh Diabetes ..... 185
Hình 234. Kết quả câu truy vấn 12 dạng Table ..... 185
Hình 235 . Kết quả câu truy vấn 12 dạng Chart ..... 186
Hình 236. Chọn các thuộc tính cần thiết ..... 187
Hình 237. Lọc theo năm 2023 ..... 188
Hình 238. Kết quả câu truy vấn 13 dạng Table ..... 189
Hình 239. Kết quả câu truy vấn 13 dạng Chart ..... 189
Hình 240. Chọn các thuộc tính cần thiết ..... 190
Hình 241. Kết quả câu truy vấn 14 dạng Table ..... 191
Hình 242 . Kết quả câu truy vấn 14 dạng Chart ..... 191
Hình 243 . Chọn các thuộc tính cần thiết ..... 192
Hình 244. Kết quả câu truy vấn 15 dạng Table ..... 192
Hình 245 . Kết quả câu truy vấn 15 dạng Chart ..... 193
Hình 246 . Câu truy vấn ngôn ngữ MDX 1 ..... 193
Hình 247 . Kết quả câu truy vấn MDX 1 ..... 194
Hình 248. Câu truy vấn ngôn ngữ MDX 2 ..... 194
Hình 249. Kết quả câu truy vấn MDX 2 ..... 195
Hình 250. Câu truy vấn ngôn ngữ MDX 3 ..... 195
Hình 251 . Kết quả câu truy vấn MDX 3 ..... 196
Hình 252. Câu truy vấn ngôn ngữ MDX 4 ..... 196
Hình 253. Kết quả câu truy vấn MDX 4 ..... 197
Hình 254. Câu truy vấn ngôn ngữ MDX 5 ..... 197
Hình 255 . Kết quả câu truy vấn MDX 5 ..... 197
Hình 256 . Câu truy vấn ngôn ngữ MDX 6 ..... 198
Hình 257. Kết quả câu truy vấn MDX 6 ..... 198
Hình 258. Câu truy vấn ngôn ngữ MDX 7 ..... 199

Hình 259 . Kết quả câu truy vấn MDX 7 ..... 199
Hình 260. Câu truy vấn ngôn ngữ MDX 8 ..... 200
Hình 261. Kết quả câu truy vấn MDX 8 ..... 200
Hình 262. Câu truy vấn ngôn ngữ MDX 9 ..... 201
Hình 263. Kết quả câu truy vấn MDX 9 ..... 201
Hình 264. Câu truy vấn ngôn ngữ MDX 10 ..... 202
Hình 265. Kết quả câu truy vấn MDX 10 ..... 202
Hình 266. Câu truy vấn ngôn ngữ MDX 11 ..... 203
Hình 267. Kết quả câu truy vấn MDX 11 ..... 203
Hình 268. Câu truy vấn ngôn ngữ MDX 12 ..... 204
Hình 269. Kết quả câu truy vấn MDX 12 ..... 204
Hình 270. Câu truy vấn ngôn ngữ MDX 13 ..... 205
Hình 271. Kết quả câu truy vấn MDX 13 ..... 205
Hình 272 . Câu truy vấn ngôn ngữ MDX 14 ..... 206
Hình 273. Kết quả câu truy vấn MDX 14 ..... 206
Hình 274. Câu truy vấn ngôn ngữ MDX 15 ..... 207
Hình 275 . Kết quả câu truy vấn MDX 15 ..... 207
Hình 276 . Get Data Report 1 ..... 208
Hình 277. Import câu lệnh MDX 1 ..... 209
Hình 278 . Transform Data ..... 210
Hình 279. Thay đổi tên cho các thuộc tính ..... 210
Hình 280. Close & Apply ..... 211
Hình 281 . Visualization matrix của Report 1 ..... 211
Hình 282 . Visualization Clustered Column Chart của Report 1 ..... 212
Hình 283 . Get Data Report 2 ..... 213
Hình 284. Import câu lệnh MDX 2 ..... 214
Hình 285 . Transform Data ..... 214
Hình 286 . Thay đổi tên cho các thuộc tính ..... 215
Hình 287 . Close & Apply ..... 215
Hình 288 . Visualization Clustered Column Chart của Report 2 ..... 216
Hình 289 . Get Data Report 3 ..... 217
Hình 290. Import câu lệnh MDX 3 ..... 218
Hình 291 . Transform Data ..... 219
Hình 292 . Thay đổi tên cho các thuộc tính ..... 219
Hình 293 . Close & Apply ..... 220
Hình 294 . Visualization Line Chart của Report 3 ..... 220
Hình 295 . Get Data Report 4 ..... 221
Hình 296 . Import câu lệnh MDX 4 ..... 222
Hình 297 . Transform Data ..... 222
Hình 298. Thay đổi tên cho các thuộc tính ..... 223
Hình 299 . Close & Apply ..... 223
Hình 300 . Visualization Clustered Bar Chart của Report 4 ..... 224
Hình 301. Get Data Report 5 ..... 225
Hình 302. Import câu lệnh MDX 5 ..... 226
Hình 303 . Transform Data ..... 226
Hình 304. Thay đổi tên cho các thuộc tính ..... 227
Hình 305. Close & Apply ..... 227
Hình 306. Visualization Pie Chart của Report 5 ..... 228
Hình 307 . Get Data Report 6 ..... 229
Hình 308 . Import câu lệnh MDX 6 ..... 230
Hình 309. Transform Data ..... 230
Hình 310. Thay đổi tên cho các thuộc tính ..... 231

Hình 311 . Close & Apply ..... 231
Hình 312 . Visualization matrix của Report 6 ..... 231
Hình 313 . Visualization Stacked Column Chart của Report 6 ..... 232
Hình 314 . Get Data Report 7 ..... 233
Hình 315 . Import câu lệnh MDX ..... 234
Hình 316 . Transform Data ..... 234
Hình 317. Thay đổi tên cho các thuộc tính ..... 235
Hình 318 . Close & Apply ..... 235
Hình 319 . Visualization Table của Report 7 ..... 235
Hình 320. Visualization Card  của Report 7 ..... 236
Hình 321 . Get Data Report 8 ..... 237
Hình 322 . Import câu lệnh MDX 8 ..... 238
Hình 323. Transform Data ..... 239
Hình 324 . Thay đổi tên cho các thuộc tính ..... 239
Hình 325 . Close & Apply ..... 240
Hình 326 . Visualization matrix của Report 8 ..... 240
Hình 327 . Get Data Report 9 ..... 241
Hình 328 . Import câu lệnh MDX 9 ..... 242
Hình 329 . Transform Data ..... 242
Hình 330 . Thay đổi tên cho các thuộc tính ..... 243
Hình 331 . Close & Apply ..... 243
Hình 332 . Visualization Stacked Bar Chart của Report 9 ..... 244
Hình 333 . Get Data Report 10 ..... 245
Hình 334 . Import câu lệnh MDX 10 ..... 246
Hình 335 . Transform Data ..... 246
Hình 336 . Thay đổi tên cho các thuộc tính ..... 247
Hình 337 . Close & Apply ..... 247
Hình 338 . Visualization Clustered Column Chart của Report 10 ..... 248
Hình 339 . Get Data Report 11 ..... 249
Hình 340 . Import câu lệnh MDX 11 ..... 250
Hình 341 . Transform Data ..... 250
Hình 342 . Thay đổi tên cho các thuộc tính ..... 251
Hình 343 . Close & Apply ..... 251
Hình 344 . Visualization Clustered Column Chart của Report 11 ..... 252
Hình 345 . Get Data Report 12 ..... 253
Hình 346 . Import câu lệnh MDX 12 ..... 254
Hình 347 . Transform Data ..... 254
Hình 348 . Thay đổi tên cho các thuộc tính ..... 255
Hình 349. Close & Apply ..... 255
Hình 350 . Visualization Stacked Bar Chart của Report 12 ..... 256
Hình 351. Get Data Report 13 ..... 257
Hình 352 . Import câu lệnh MDX 13 ..... 258
Hình 353 . Transform Data ..... 258
Hình 354 . Thay đổi tên cho các thuộc tính ..... 259
Hình 355 . Close & Apply ..... 259
Hình 356 . Visualization Pie Chart của Report 13 ..... 260
Hình 357 . Get Data Report 14 ..... 261
Hình 358 . Import câu lệnh MDX 14 ..... 262
Hình 359. Transform Data ..... 262
Hình 360 . Thay đổi tên cho các thuộc tính ..... 263
Hình 361 . Close & Apply ..... 263
Hình 362 . Visualization Clustered Column Chart của Report 14 ..... 264

Hình 363 . Get Data Report 15 ..... 265
Hình 364 . Import câu lệnh MDX 15 ..... 266
Hình 365 . Transform Data ..... 266
Hình 366 . Thay đổi tên cho các thuộc tính ..... 267
Hình 367 . Close & Apply ..... 267
Hình 368 . Visualization Stacked Column Chart của Report 15 ..... 268
Hình 369. Tiền xử lý dữ liệu ..... 270
Hình 370. Bước 1_Thuật toán K- Means ..... 270
Hình 371. Tiền xử lý dữ liệu ..... 271
Hình 372. Xác định số cụm ..... 271
Hình 373. Vẽ biểu đồ Elbow ..... 271
Hình 374. Biểu đồ Elbow ..... 272
Hình 375. K-Means Clustering ..... 272
Hình 376. Vẽ sơ đồ gom cụm ..... 273
Hình 377. Biểu đồ gom cụm ..... 274
Hình 378. Xuất file Excel ..... 274
Hình 379. Import thư viện ..... 275
Hình 380. Xử lý dữ liệu (1) ..... 275
Hình 381. Xử lý dữ liệu (2) ..... 275
Hình 382. Huấn luyện mô hình ..... 276
Hình 383. Kết quả huấn luyện mô hình ..... 276
Hình 384. Đánh giá mô hình ..... 276
Hình 385. Kết quả đánh giá mô hình ..... 277
Hình 386. Phân tích Feature Importance ..... 277
Hình 387. Biểu đồ tầm quan trọng với thuộc tính ..... 278
Hình 388. Trực quan hóa ma trận nhầm lẫn ..... 278
Hình 389. Ma trận nhầm lẫn Decision Tree ..... 279
Hình 390. In ra cây quyết định ..... 279
Hình 391. Decision Tree Visualization ..... 280
Hình 392. Ma trận nhầm lẫn Random Forest ..... 280
Hình 393. Cluster 0 ..... 281
Hình 394. Cluster 1 ..... 282
Hình 395. Cluster 2 ..... 283
Hình 396. Cluster 3 ..... 284

# DANH MỤC BẢNG

Bảng 1. Thuộc tính kho dữ liệu ..... 20
Bảng 2. Kho dữ liệu đã xử lý ..... 22
Bảng 3. DIM_LOCATION ..... 23
Bảng 4. DIM_DATE ..... 23
Bảng 5. DIM_DOCTOR ..... 23
Bảng 6. DIM_INSURANCE_PROVIDER ..... 24
Bảng 7. DIM_PATIENT ..... 24
Bảng 8. DIM_MEDICATION ..... 25
Bảng 9. DIM_TEST_RESULT ..... 25
Bảng 10. FACT ..... 26

# CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ DỮ LIỆU

## 1. LÝ DO CHỌN ĐỀ TÀI

Trong kỷ nguyên dữ liệu lớn, ngành chăm sóc sức khỏe đang trải qua một cuộc cách mạng. Dữ liệu sức khỏe, từ hồ sơ bệnh án điện tử đến dữ liệu về thiết bị y tế đeo được, ngày càng trở nên phong phú và đa dạng. Để khai thác hiệu quả lượng dữ liệu khổng lồ này, các công cụ phân tích dữ liệu đa chiều (OLAP) đóng vai trò vô cùng quan trọng. Việc áp dụng OLAP vào lĩnh vực chăm sóc sức khỏe không chỉ giúp các nhà nghiên cứu, bác sĩ và nhà quản lý có cái nhìn toàn diện về tình hình sức khỏe của cộng đồng mà còn hỗ trợ trong việc đưa ra các quyết định chính xác về phòng ngừa, chẩn đoán và điều trị bệnh.

## 2. PHÁT BIỂU VỀ DỮ LIỆU

### 2.1. Mô tả dữ liệu

- Kho dữ liệu chăm sóc sức khỏe  được thiết kế để mô phỏng dữ liệu chăm sóc sức khỏe trong thế giới thực, cho phép người dùng thực hành, phát triển và thể hiện kỹ năng xử lý và phân tích dữ liệu của họ trong bối cảnh ngành chăm sóc sức khỏe.
- Thông qua kho dữ liệu người dùng có thể biết được các thông tin về ngày nhập viện, ngày xuất viện của một bệnh nhân ; thông tin bảo hiểm, bệnh lý của một bệnh nhân; đơn thuốc; bác sĩ; viện phí;…
- Kho dữ liệu được xây dựng với hướng chủ đề chăm sóc sức khỏe (Healthcare)
- Kho dữ liệu gồm 55.500 dòng và 15 thuộc tính
- Link dữ liệu gốc : Healthcare Dataset (kaggle.com)

## 2.2. Thuộc tính của kho dữ liệu

| Name<br>Name | Age<br>Age | Gender<br>Gender | Blood Type<br>Blood Type | Medical Condition<br>Medical Condition | Date of<br>Date of A |
|---|---|---|---|---|---|
| 49992<br>unique values | 13 — 89 | Male 50%<br>Female 50% | A- 13%<br>A+ 13%<br>Other (41575) 75% | Arthritis 17%<br>Diabetes 17%<br>Other (36888) 66% | 2019-05- |
| Bobby JacksOn | 30 | Male | B- | Cancer | 2024-01- |
| LesLie TErRy | 62 | Male | A+ | Obesity | 2019-08- |
| DaNnY sMitH | 76 | Female | A- | Obesity | 2022-09- |
| andrEw waTtS | 28 | Female | O+ | Diabetes | 2020-11- |
| adrIENNE bEll | 43 | Female | AB+ | Cancer | 2022-09- |
| EMILY JOHNSOn | 36 | Male | A+ | Asthma | 2023-12- |
| edwArD EDWaRDs | 21 | Female | AB- | Diabetes | 2020-11- |
| CHrisTInA MARtinez | 20 | Female | A+ | Cancer | 2021-12- |
| JASmINe aGuIlaR | 82 | Male | AB+ | Asthma | 2020-07- |
| ChRISTopher BerG | 58 | Female | AB- | Cancer | 2021-05- |
| mIchElLe daniELs | 72 | Male | O+ | Cancer | 2020-04- |
| aaRon MARtiNeZ | 38 | Female | A- | Hypertension | 2023-08- |
| connOR HANsEn | 75 | Female | A+ | Diabetes | 2019-12- |

*Hình1. Thuộc tính của kho dữ liệu*

| STT | Tên thuộc tính | Kiểu dữ liệu | Ý nghĩa của thuộc tính |
|---|---|---|---|
| 1 | Patient_name | Text | Tên bệnh nhân |
| 2 | Age | Numberial | Tuổi của bệnh nhân |
| 3 | Gender | Text | Giới tính:<br>1. Nam<br>2. Nữ |
| 4 | Blood_type | Text | Nhóm máu |
| 5 | Medical_condition | Text | Tình trạng bệnh lý |
| 6 | Admission_date | Date | Ngày nhập viện<br>(Định dạng: dd/mm/yyyy) |
| 7 | Discharge_date | Date | Ngày xuất viện<br>(Định dạng: dd/mm/yyyy) |
| 8 | Doctor_name | Text | Tên bác sĩ |

| | | | |
|---|---|---|---|
| 9 | Hospital_name | Text | Tên bệnh viện |
| 10 | Room_number | Numberial | Số phòng |
| 11 | Insurance_provider | Text | Tên nhà cung cấp bảo hiểm:<br>1. Aetna<br>2. Blue Cross<br>3. Cigna<br>4. UnitedHealthcare<br>5. Medicare |
| 12 | Medication | Text | Tên thuốc:<br>• Aspirin<br>• Ibuprofen<br>• Penicillin<br>• Paracetamol<br>• Lipitor |
| 13 | Admission_type | Text | Loại nhập viện |
| 14 | Test_result | Text | Kết quả xét nghiệm:<br>• Normal<br>• Abnormal<br>• Inconclusive |
| 15 | Billing_amount | Numberial | Viện phí |

*Bảng 1. Thuộc tính kho dữ liệu*

### 2.3. Kho dữ liệu đã xử lý

Sau khi lọc dữ liệu ta được 55500 dòng và 13 thuộc tính để sử dụng cho việc phân tích đề tài.

100 %

Results | Messages

| | fact_id | BillingAmount | admission_date_id | discharge_date_id | date_id | location_id | test_result_id | doctor_id | medication_id | insurance_id | patient_id |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 18856.2813059782 | 57 | 58 | 1 | 43335 | 3 | 26613 | 4 | 2 | 5153 |
| 2 | 2 | 33643.3272865779 | 4 | 4 | 2 | 25285 | 2 | 33649 | 2 | 4 | 29834 |
| 3 | 3 | 27955.0960788425 | 41 | 42 | 3 | 11539 | 3 | 37829 | 1 | 1 | 11696 |
| 4 | 4 | 37909.7824098753 | 19 | 20 | 4 | 20123 | 1 | 22512 | 2 | 4 | 2702 |
| 5 | 5 | 14238.3178139376 | 41 | 42 | 3 | 48097 | 1 | 21260 | 5 | 1 | 397 |
| 6 | 6 | 48145.1109510419 | 56 | 56 | 6 | 34469 | 3 | 37122 | 2 | 5 | 15409 |
| 7 | 7 | 19580.8723448609 | 19 | 19 | 7 | 17999 | 2 | 21959 | 4 | 4 | 14742 |
| 8 | 8 | 45820.4627215946 | 32 | 33 | 8 | 37868 | 2 | 36660 | 4 | 3 | 8697 |
| 9 | 9 | 50119.2227915485 | 15 | 15 | 9 | 43847 | 1 | 9261 | 1 | 3 | 19801 |
| 10 | 10 | 19784.6310622107 | 25 | 26 | 10 | 35029 | 2 | 14680 | 4 | 5 | 9025 |
| 11 | 11 | 12576.7956090502 | 12 | 12 | 11 | 41100 | 3 | 18734 | 4 | 4 | 35210 |
| 12 | 12 | 7999.58687960419 | 52 | 53 | 12 | 29950 | 2 | 11559 | 3 | 4 | 83 |
| 13 | 13 | 43282.2833577043 | 8 | 8 | 13 | 37925 | 1 | 22138 | 5 | 3 | 10103 |
| 14 | 14 | 33207.7066337296 | 13 | 14 | 14 | 39210 | 3 | 37437 | 3 | 5 | 40319 |
| 15 | 15 | 40701.5992273088 | 30 | 30 | 15 | 33320 | 3 | 32894 | 4 | 5 | 6675 |
| 16 | 16 | 12263.3574250214 | 45 | 45 | 16 | 11006 | 2 | 25489 | 1 | 2 | 36606 |
| 17 | 17 | 24499.8479037366 | 14 | 15 | 17 | 10347 | 3 | 13098 | 4 | 5 | 17425 |
| 18 | 18 | 17440.4654441247 | 11 | 12 | 18 | 45682 | 1 | 20615 | 4 | 3 | 36410 |
| 19 | 19 | 18843.0230178342 | 23 | 23 | 19 | 21076 | 1 | 20638 | 1 | 2 | 30952 |
| 20 | 20 | 23762.2035790596 | 43 | 43 | 20 | 18941 | 3 | 10637 | 5 | 3 | 11420 |
| 21 | 21 | 42.5145885533243 | 50 | 51 | 21 | 24438 | 3 | 22926 | 1 | 2 | 47142 |
| 22 | 22 | 17695.9116223438 | 9 | 10 | 22 | 48564 | 3 | 14411 | 3 | 1 | 9037 |
| 23 | 23 | 5998.10290819591 | 37 | 38 | 23 | 10919 | 2 | 39124 | 3 | 1 | 26561 |
| 24 | 24 | 25250.0524282161 | 28 | 28 | 24 | 4970 | 2 | 11420 | 3 | 1 | 14234 |
| 25 | 25 | 33211.295419012 | 13 | 14 | 14 | 48892 | 2 | 35879 | 4 | 4 | 38390 |

*Hình 2 . Kho dữ liệu đã xử lý*

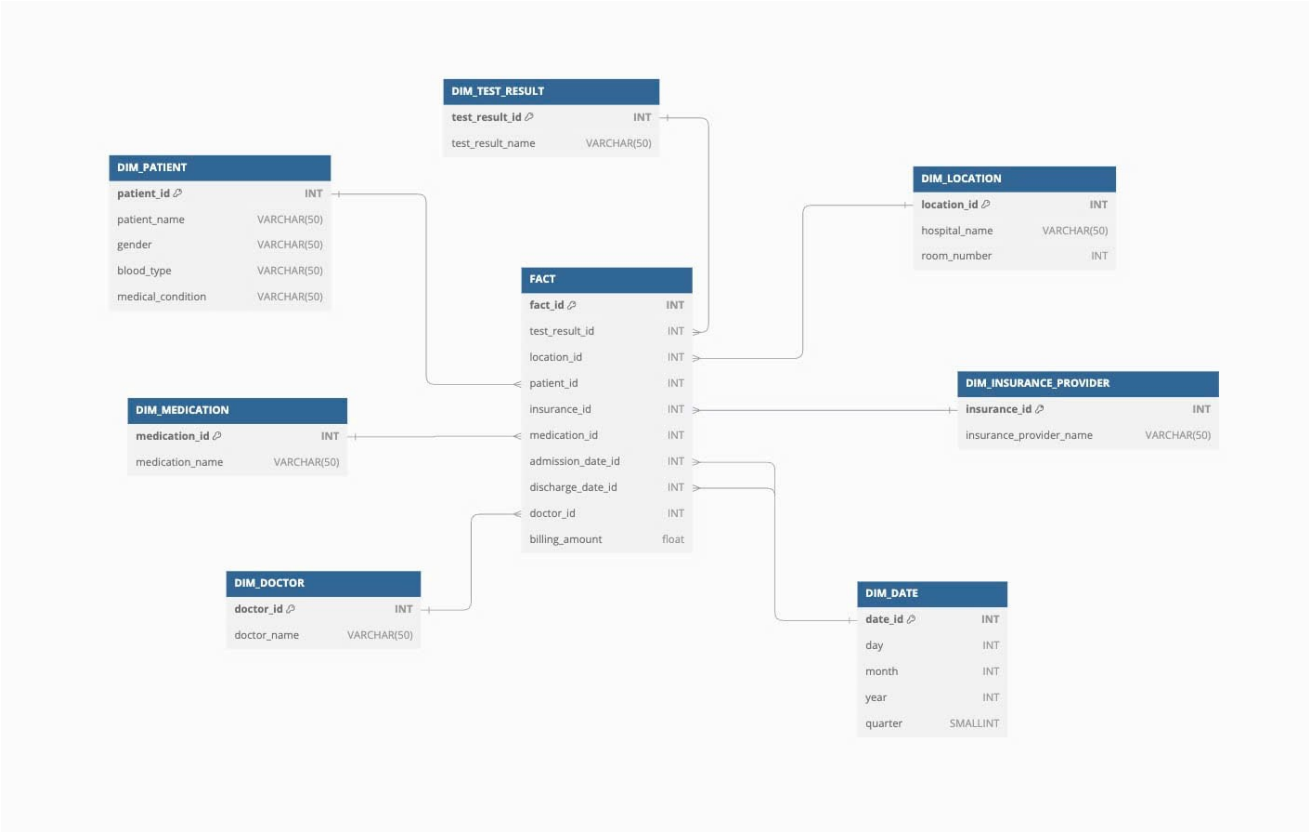
| STT | Tên thuộc tính | Kiểu dữ liệu | Ý nghĩa của thuộc tính |
|---|---|---|---|
| 1 | Patient_name | Text | Tên bệnh nhân |
| 2 | Gender | Text | Giới tính:<br>3. Nam<br>4. Nữ |
| 3 | Blood_type | Text | Nhóm máu |
| 4 | Medical_condition | Text | Tình trạng bệnh lý |
| 5 | Admission_date | Date | Ngày nhập viện<br>(Định dạng: dd/mm/yyyy) |
| 6 | Discharge_date | Date | Ngày xuất viện<br>(Định dạng: dd/mm/yyyy) |
| 7 | Doctor_name | Text | Tên bác sĩ |
| 8 | Hospital_name | Text | Tên bệnh viện |
| 9 | Room_number | Numberial | Số phòng |
| 10 | Insurance_provider | Text | Tên nhà cung cấp bảo hiểm:<br>6. Aetna<br>7. Blue Cross<br>8. Cigna |

| | | | |
|---|---|---|---|
| | | | 9. UnitedHealthcare<br>10. Medicare |
| 11 | Medication | Text | Tên thuốc:<br>• Aspirin<br>• Ibuprofen<br>• Penicillin<br>• Paracetamol<br>• Lipitor |
| 12 | Test_result | Text | Kết quả xét nghiệm:<br>• Normal<br>• Abnormal<br>• Inconclusive |
| 13 | Billing_amount | Numberial | Viện phí |

*Bảng 2. Kho dữ liệu đã xử lý*

## 3. XÂY DỰNG KHO DỮ LIỆU

### 3.1. Sơ đồ bông tuyết minh họa




*Hình 3. Sơ đồ bông tuyết*

### 3.2. Dim_Location

Gồm có 49921 dòng dữ liệu, 3 thuộc tính, mỗi dòng dữ liệu chứa thông tin phòng, bệnh viện của bệnh nhân.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | location_id | | int | Mã địa điểm |
| | hospital_name | | varchar | Tên bệnh viện |
| | room_number | | Int | Số phòng |

*Bảng 3. DIM_LOCATION*

### 3.3. Dim_Date

Gồm có 55500 dòng dữ liệu, 3 thuộc tính, mỗi dòng dữ liệu chứa thông tin mã nhập viện, xuất viện.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | date_id | | int | Mã ngày |
| | day | | int | Ngày |
| | month | | int | Tháng |
| | year | | int | Năm |
| | quarter | | Small Int | Qúy |

*Bảng 4. DIM_DATE*

### 3.4. Dim_ Doctor

Gồm có 40341 dòng dữ liệu, 2 thuộc tính, mỗi dòng dữ liệu chứa thông tin bác sĩ của bệnh viện

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | doctor_id | | int | Mã bác sĩ |
| | doctor_name | | varchar | Tên bác sĩ |

*Bảng 5. DIM_DOCTOR*

### 3.5. Dim_Insurance _Provider

Gồm có 5 dòng dữ liệu, 2 thuộc tính, mỗi dòng dữ liệu chứa thông tin của nhà cung cấp bảo hiểm.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | insurance_id | | int | Mã nhà cung cấp bảo hiểm |
| | insurance_provider_name | | varchar | Tên nhà cung cấp bảo hiểm:<br>• Aetna<br>• Blue Cross<br>• Cigna<br>• UnitedHealthcare<br>• Medicare |

*Bảng 6. DIM_INSURANCE_PROVIDER*

### 3.6. Dim_Patient

Gồm có 50000 dòng dữ liệu, 5 thuộc tính, mỗi dòng dữ liệu chứa thông tin của bệnh nhân.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | patient_id | | int | Mã bệnh nhân |
| | patient_name | | varchar | Tên bệnh nhân |
| | gender | | varchar | Giới tính:<br>1. Nam<br>2. Nữ |
| | blood_type | | varchar | Nhóm máu |
| | medical_condition | | varchar | Tình trạng bệnh lý |

*Bảng 7. DIM_PATIENT*

### 3.7. Dim_Medication

Gồm có 5 dòng dữ liệu, 2 thuộc tính, mỗi dòng dữ liệu chứa thông tin thuốc của bệnh nhân.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | medication_id | | int | Mã thuốc |

| | | | | |
|---|---|---|---|---|
| | medication_name | | varchar | Tên thuốc:<br>• Aspirin<br>• Ibuprofen<br>• Penicillin<br>• Paracetamol<br>• Lipitor |

*Bảng 8. DIM_MEDICATION*

### 3.8. Dim_Test_Result

Gồm có 3 dòng dữ liệu, 2 thuộc tính, mỗi dòng dữ liệu chứa thông tin kết quả xét nghiệm.

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | test_result_id | | int | Mã xét nghiệm |
| | test_result_name | | varchar | Kết quả xét nghiệm:<br>• Normal<br>• Abnormal<br>• Inconclusive |

*Bảng 9. DIM_TEST_RESULT*

### 3.9. Fact

Gồm có 55500 dòng dữ liệu, 9 thuộc tính, là bảng sự kiện của kho dữ liệu

| Khoá chính | Tên thuộc tính | NULL | Kiểu dữ liệu | Mô tả thuộc tính |
|---|---|---|---|---|
| 🔑 | fact_id | | int | Mã của bảng sự kiện |
| 🗝 | location_id | | int | Mã địa điểm |
| 🗝 | date_id | | int | Mã ngày |
| 🗝 | test_result_id | | int | Mã xét nghiệm |
| 🗝 | patient_id | | int | Mã bệnh nhân |
| 🗝 | insurance_id | | int | Mã nhà cung cấp bảo hiểm |

| | | | | |
|---|---|---|---|---|
| 🔑 | medication_id | | int | Mã thuốc |
| 🔑 | doctor_id | | int | Mã bác sĩ |
| | billing_amount | | double | Viện phí |

*Bảng 10. FACT*

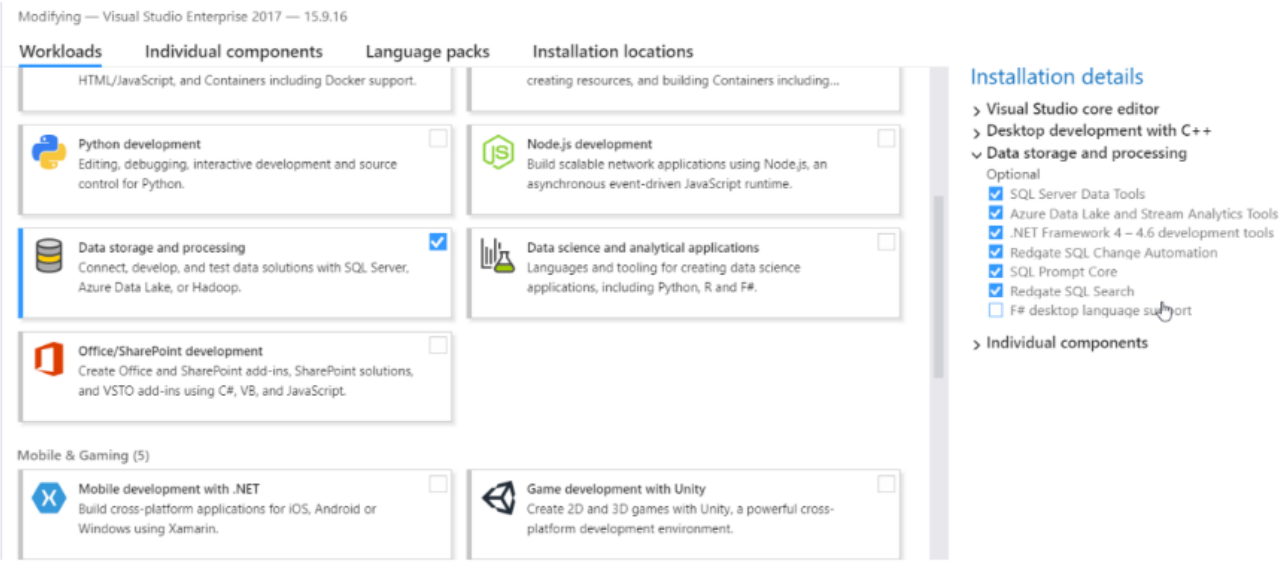
## 4. Nội dung 15 câu truy vấn

a. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm
b. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu.
c. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm.
d. Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm
e. Đưa ra top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023.
f. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc top 5
g. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024
h. Liệt kê các bệnh nhân(mã bệnh nhân, tên bệnh nhân, giới tính, nhóm máu) có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024.
i. Trong mỗi loại bệnh, in ra bác sĩ  có số lượng bênh nhân lớn nhất trong năm 2022
j. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.
k. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể
l. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần.
m. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.
n. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất / top 3 số lần nhập viện nhiều nhất.
o. Cho biết bệnh viện có doanh thu  cao nhất và bệnh viện có doanh thu thấp nhất.

# CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

## 1. CHUẨN BỊ CÔNG CỤ VÀ DATAWAREHOUSE

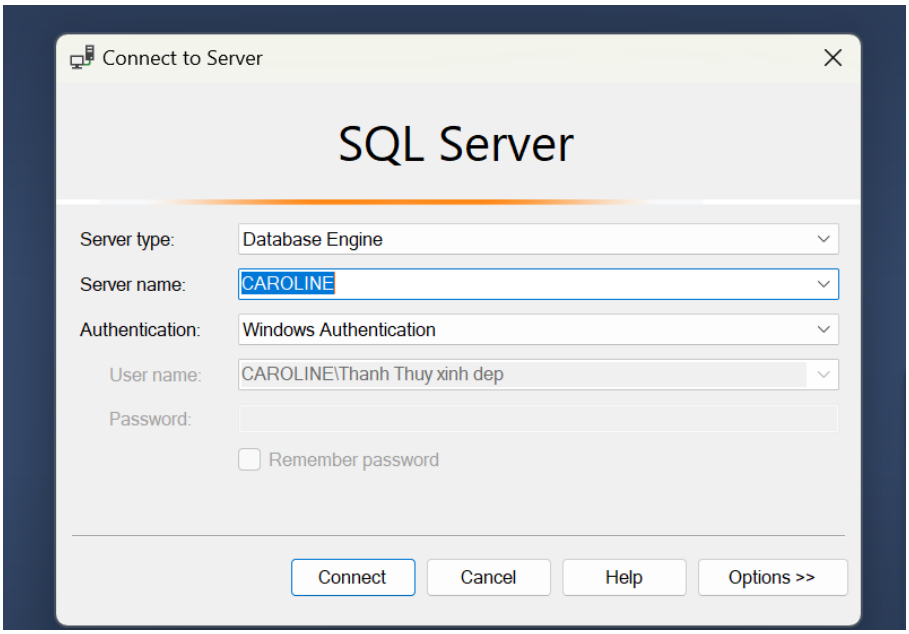
- Download công cụ **SQL Server Data Tools**




*Hình 4 . SQL Server Data Tools*

Chọn **Data storage and processing** và chọn những mục như hình minh họa.

- Tải các công cụ SSDT cho phiên bản **Visual Studio Enterprise 2022.**
- Các bước tạo cơ sở dữ liệu và thiết lập kho dữ liệu :
  - **Bước 1 :** Mở **Microsoft SQL Server Management Studio** và kết nối với **Server**

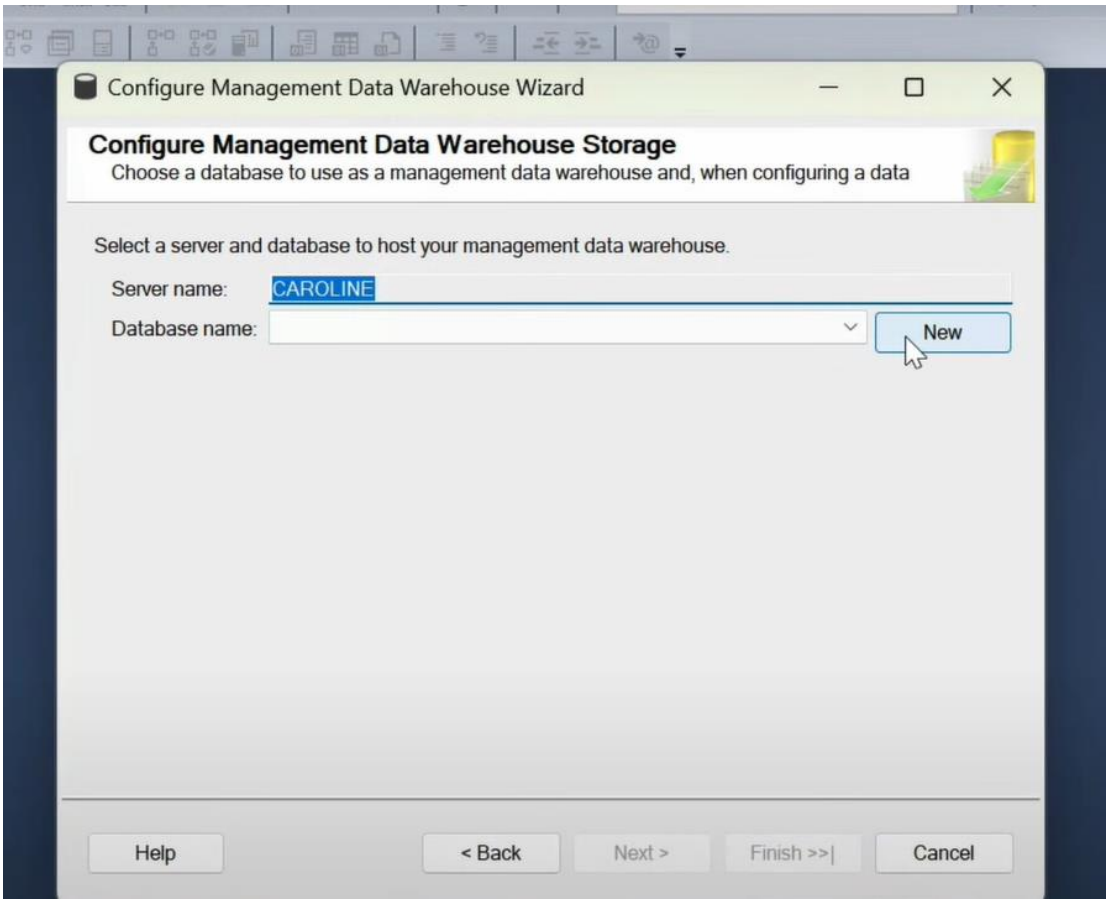



*Hình 5 . Server*

- **Bước 2** : Tạo và cấu hình kho dữ liệu

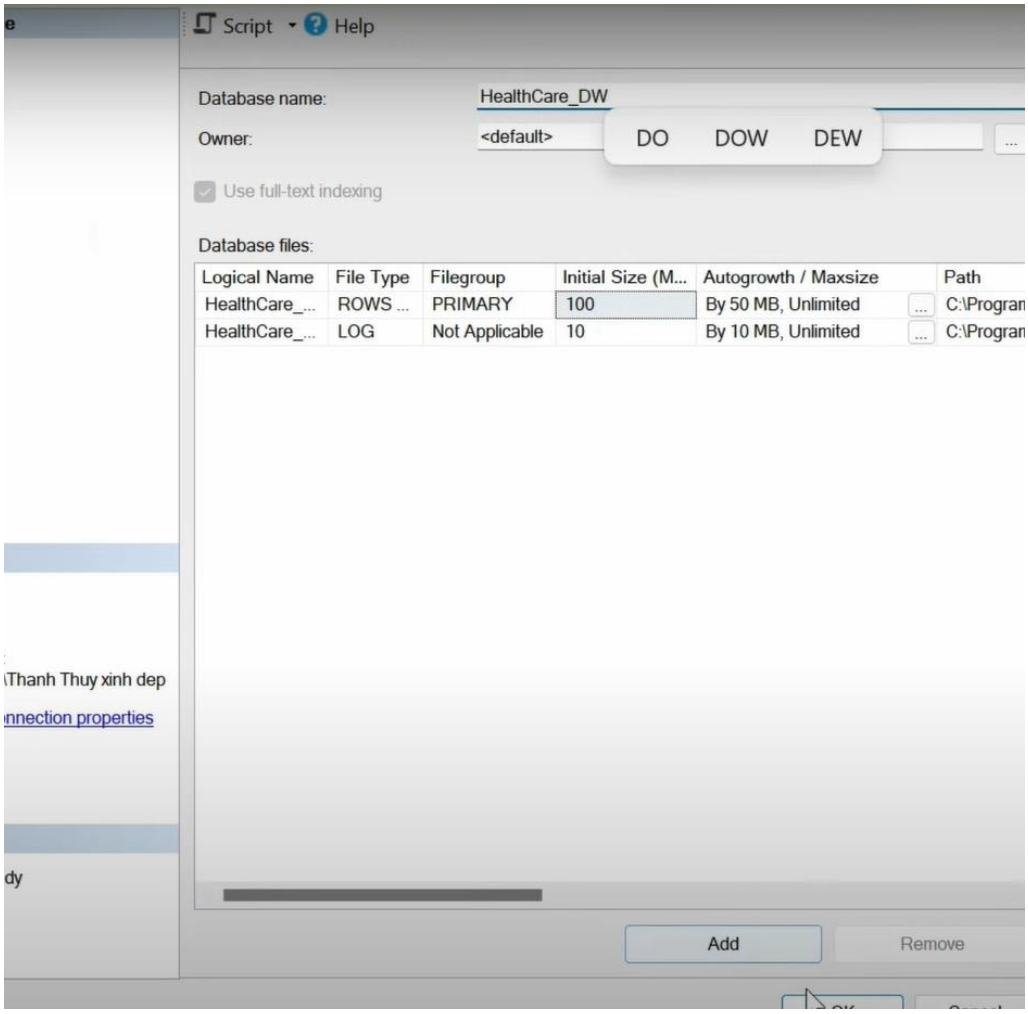
*Hình 6 . Tạo và cấu hình kho dữ liệu*

- **Bước 3 :** Nhấn **Next** để tiếp tục và sau đó nhấn **New** để tạo kho dữ liệu mới



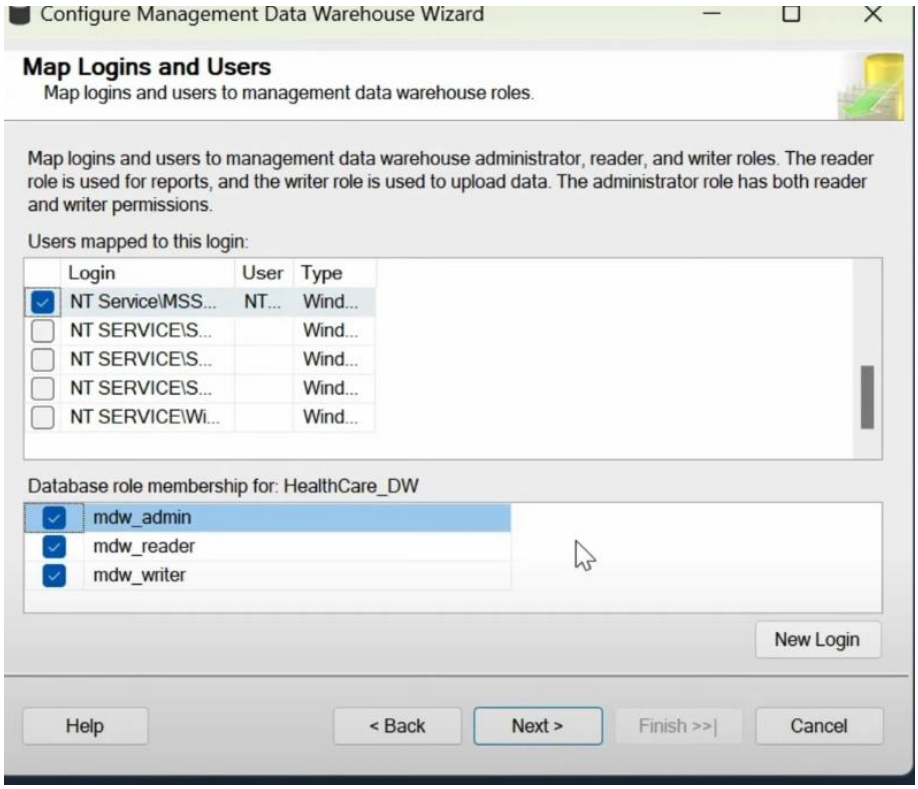

*Hình 7. Tạo kho dữ liệu mới*

- **Bước 4 :** Đặt tên kho dữ liệu và tùy biến kích thước kho dữ liệu. Rồi nhấn **Add.**

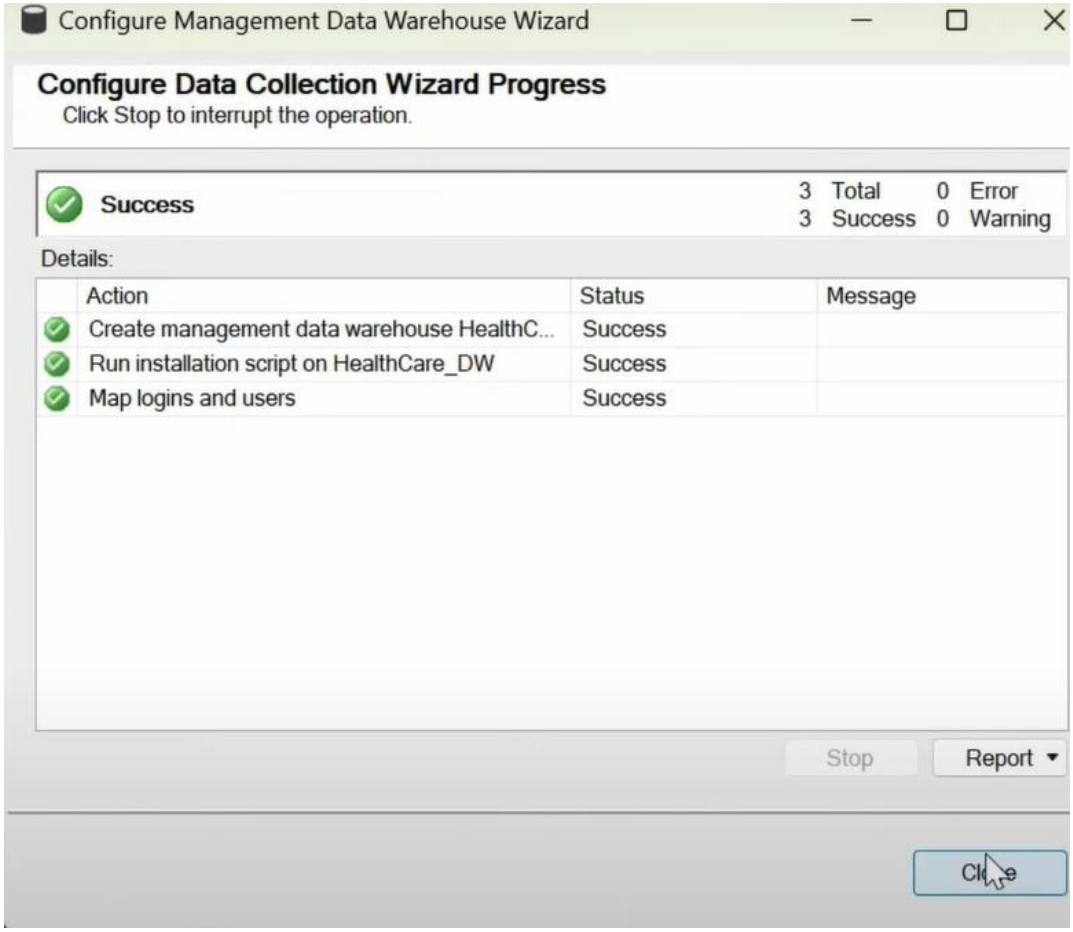



*Hình 8.Tạo kho dữ liệu mới*

- **Bước 5 :** Tiếp tục nhấn **Next**. Tiếp theo click vào như hình rồi nhấn **Next** sau đó nhấn **Finish** để kết thúc quá trình tạo data warehouse.



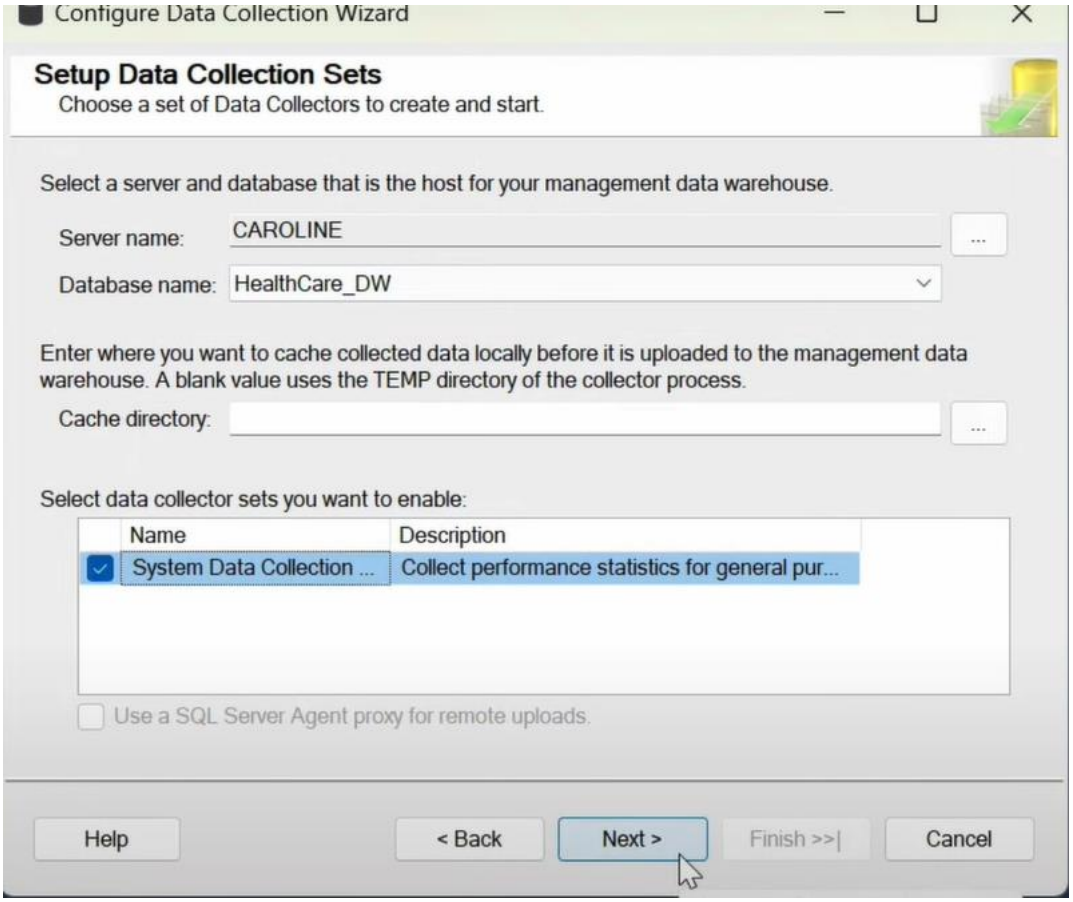

*Hình 9. Tạo DW*




*Hình 10. Tạo DW*

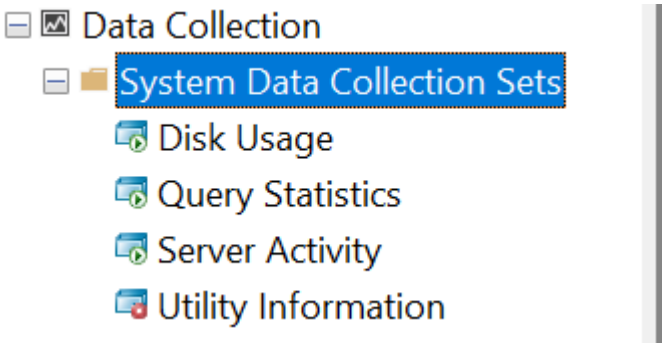
- **Bước 6**: Tùy chỉnh dữ liệu thu thập. Tiếp tục nhấn **Next**, sau đó **Finish** để hoàn thành rồi nhấn **Close.**

*Hình 11 . Tùy chỉnh dữ liệu*




*Hình 12. Configure Data Collection*

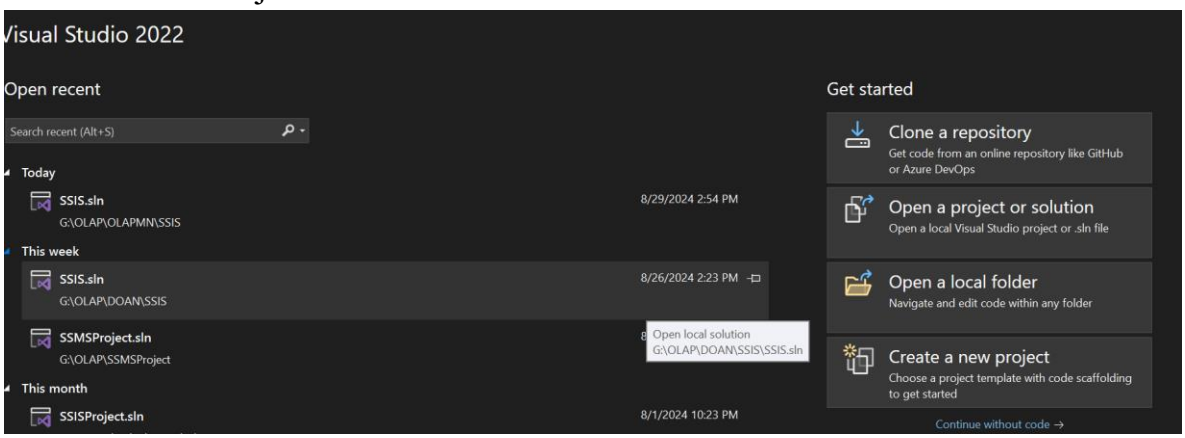
- **Bước 7 :** Kết quả chạy hoàn thành.



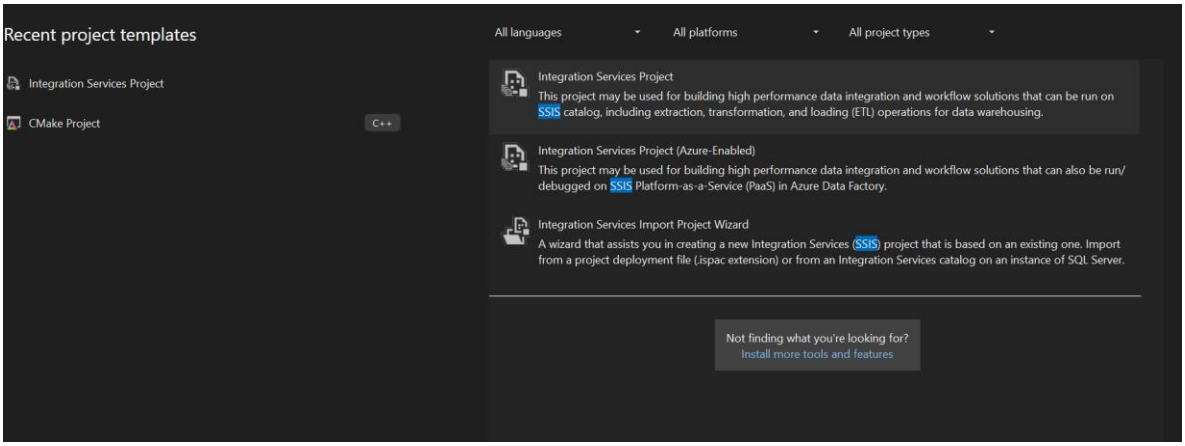

*Hình 13. Hoàn thành*

## 2. TẠO PROJECT SSIS TRONG VISUAL STUDIO 2022

- Chọn Create a new project → Hộp thoại sẽ hiển thị như hình → Chọn "Intergration Services Project"

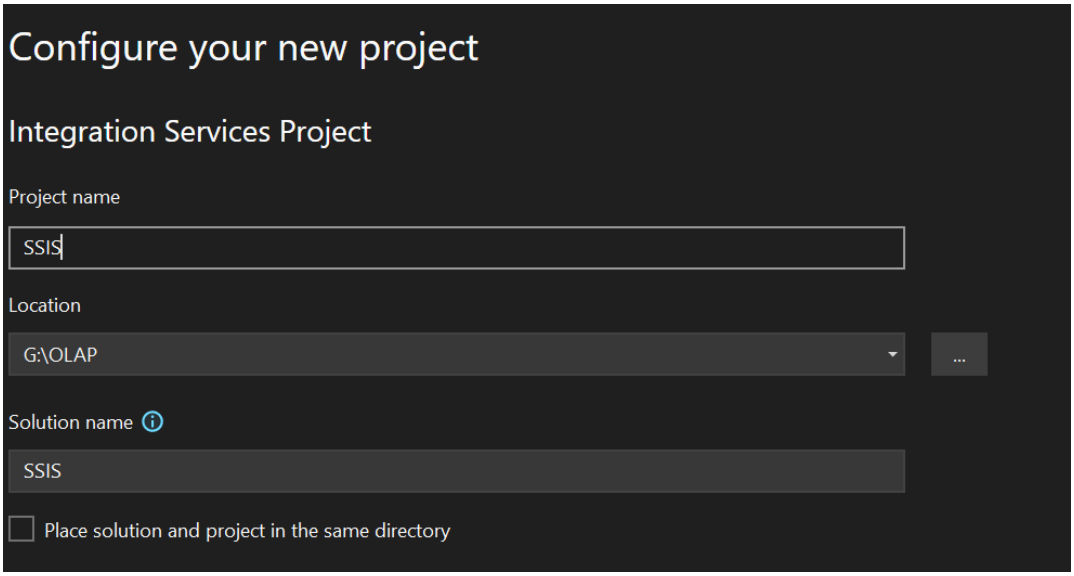



*Hình 14. Create Project*




*Hình 15. Create Project*

- Nhập tên cho Project

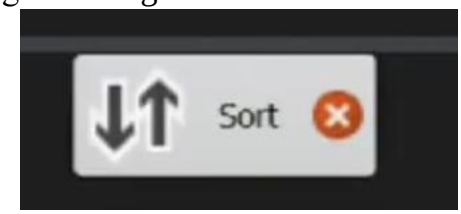



*Hình 15. Nhập tên cho Project*

## 3. TẠO BẢNG DIM VÀ BẢNG FACT

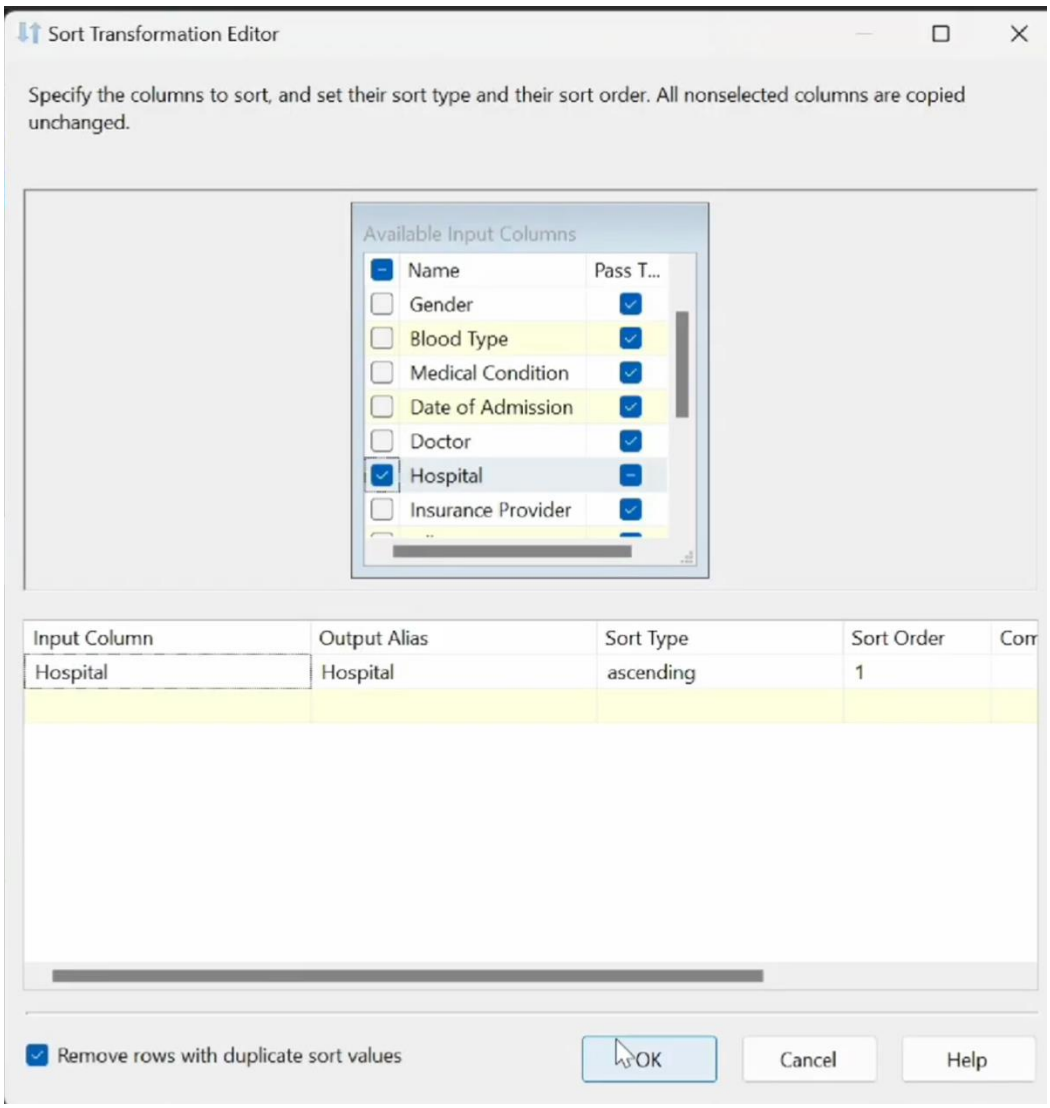
### 3.1. DIM_LOCATION

- Bước 1 : Kéo chức năng Sort sang.



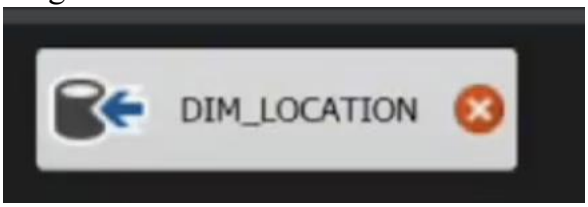

*Hình 16. Sort Dim_Location*

- Bước 2 : Sort theo Hospital, tích vào Remove rows with duplicate sort values và nhấn OK.



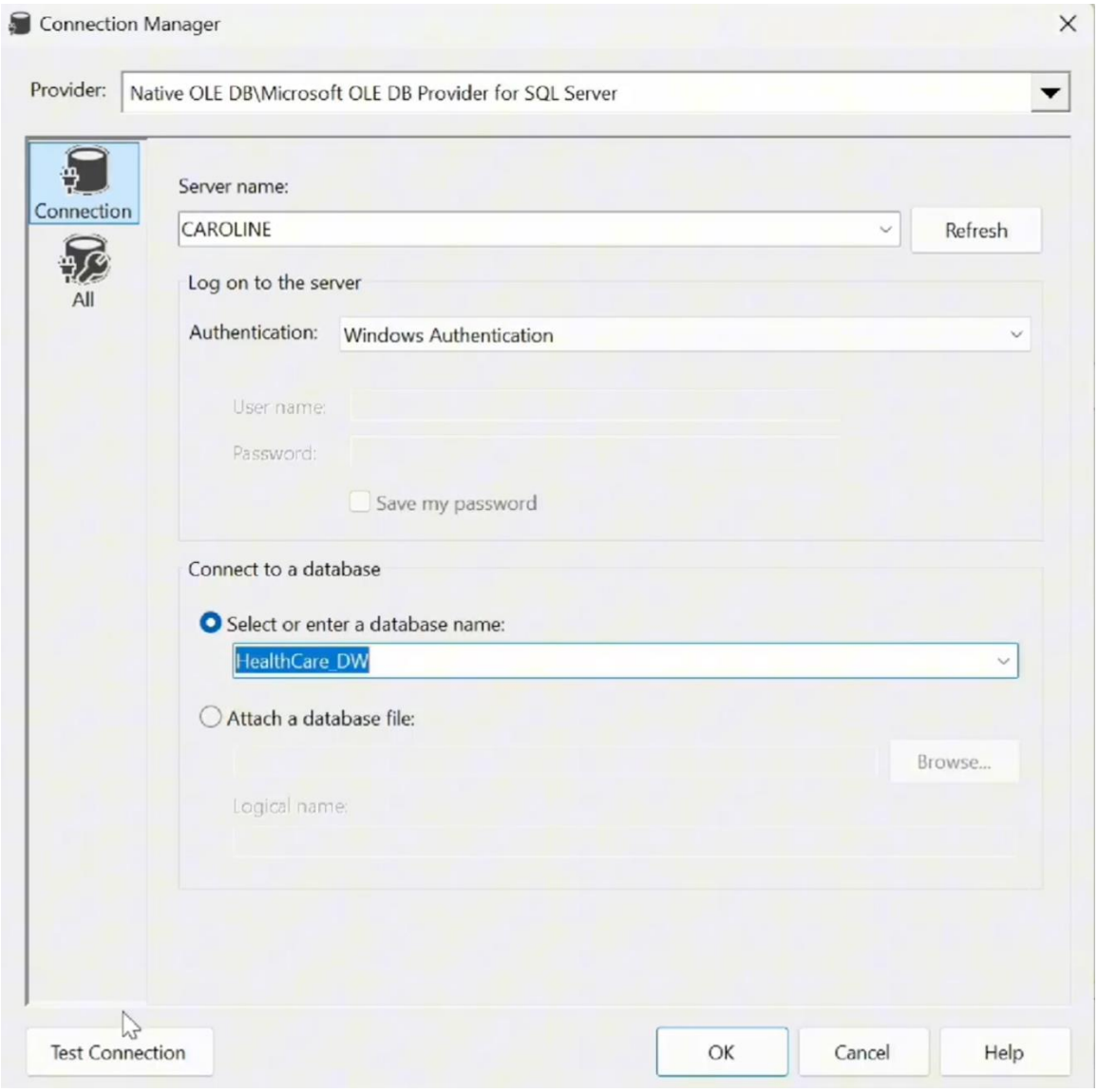

*Hình 17 . Remove row with duplicate sort values*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_LOCATION



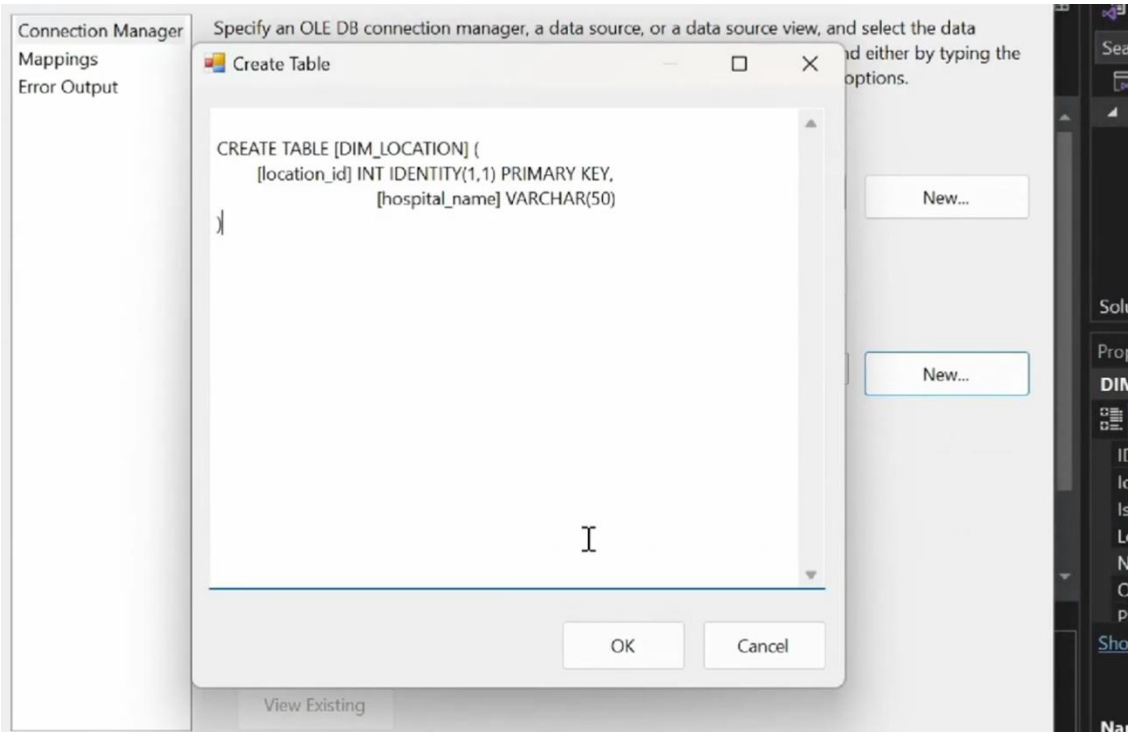

*Hình 18 . OLE DB Destination – Dim_Loaction*

- Bước 4: Tạo một Connection Manager mới với Server name là tên máy chủ và chọn database đã tạo.



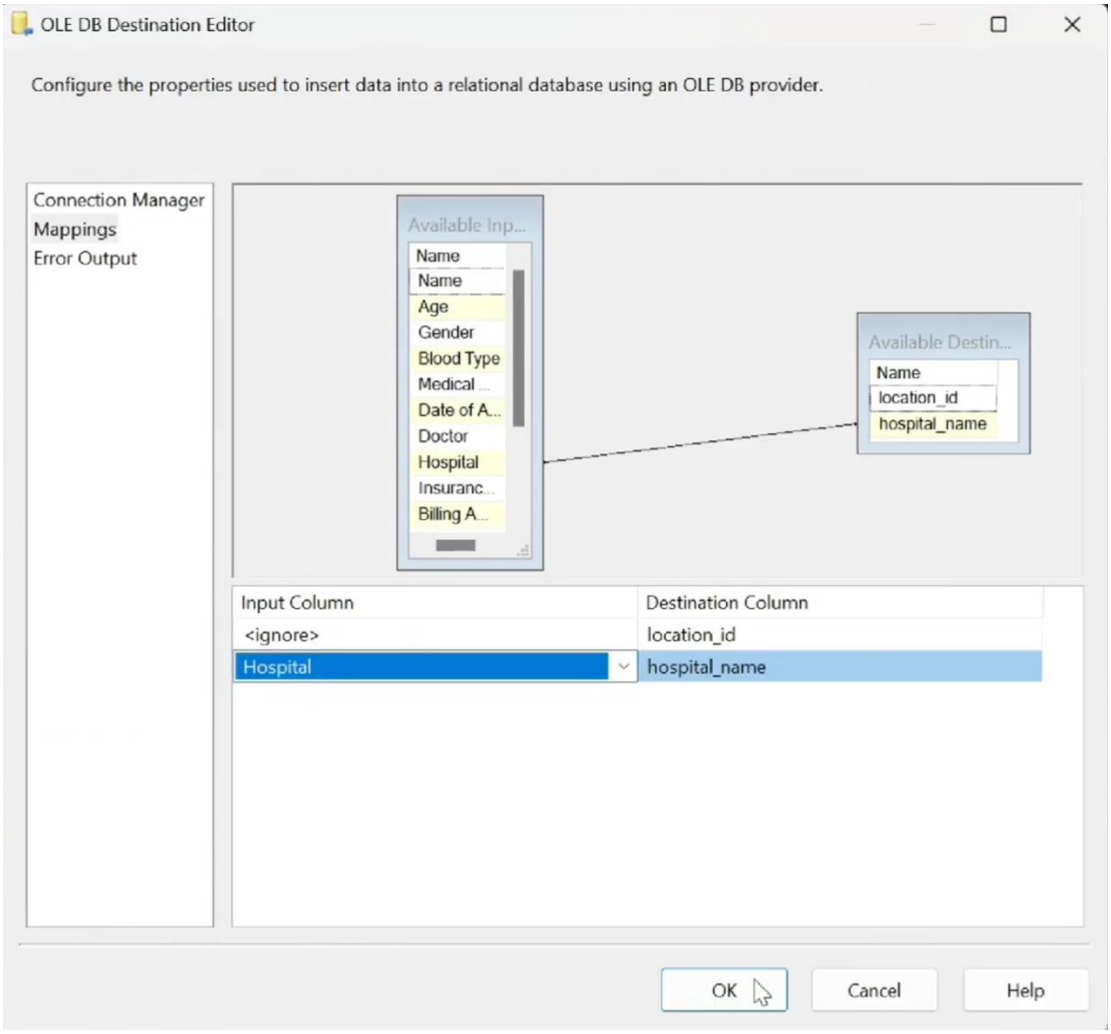

*Hình 19. Connection Manager*

- Bước 5: Tạo bảng DIM_LOCATION




*Hình 20. Tạo bảng Dim_Location*

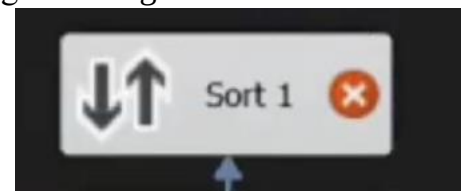
- Bước 6: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.




*Hình 21. Mapping thuộc tính*

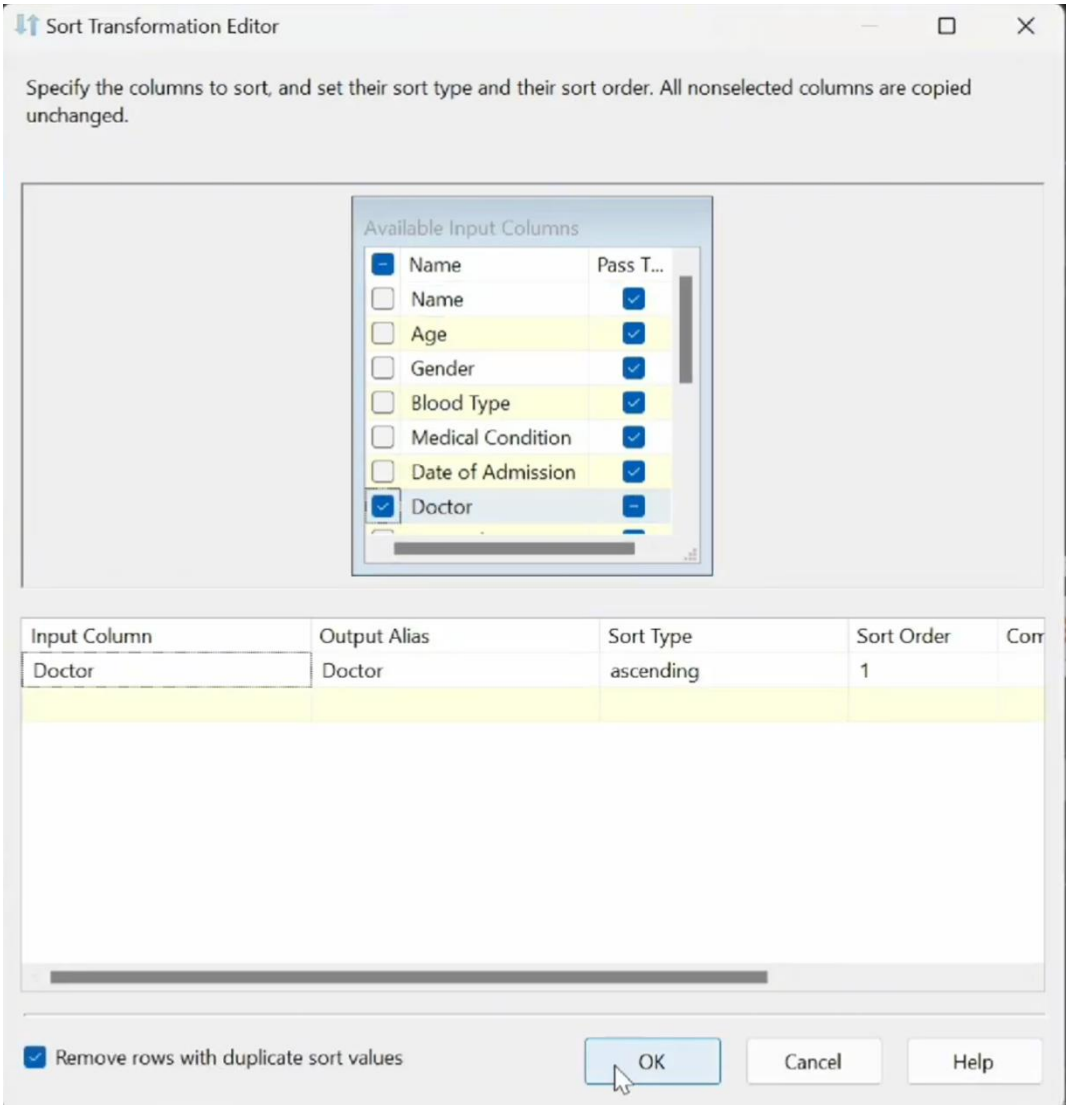
## 3.2. DIM_DOCTOR

- Bước 1 : Kéo chức năng Sort sang.



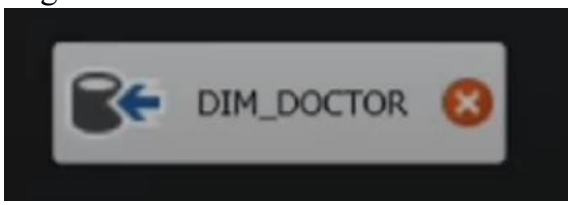

*Hình 22. Sort Dim_Doctor*

- Bước 2 : Sort theo Doctor, tích vào Remove rows with duplicate sort values và nhấn OK.



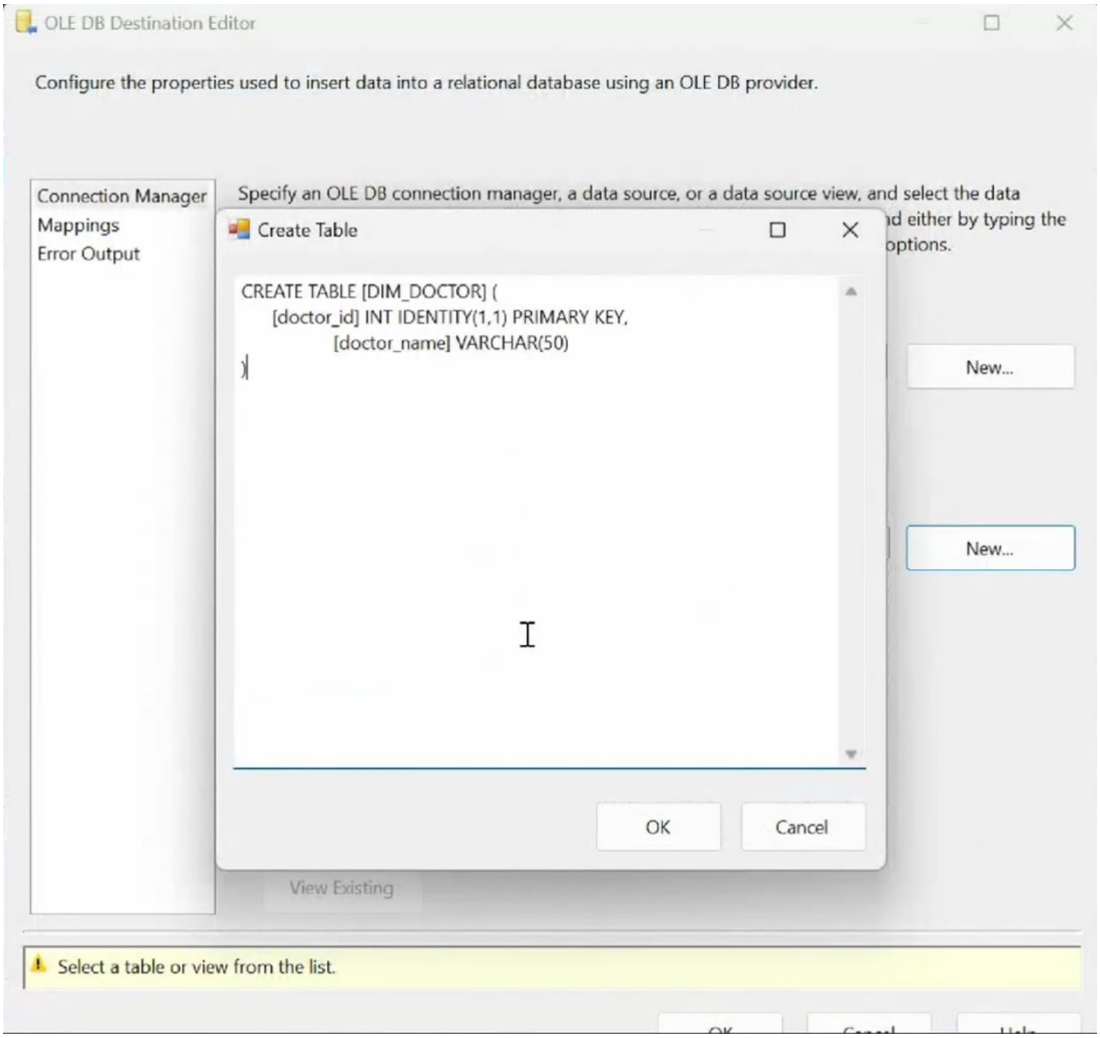

*Hình 23. Sort theo Doctor*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_DOCTOR



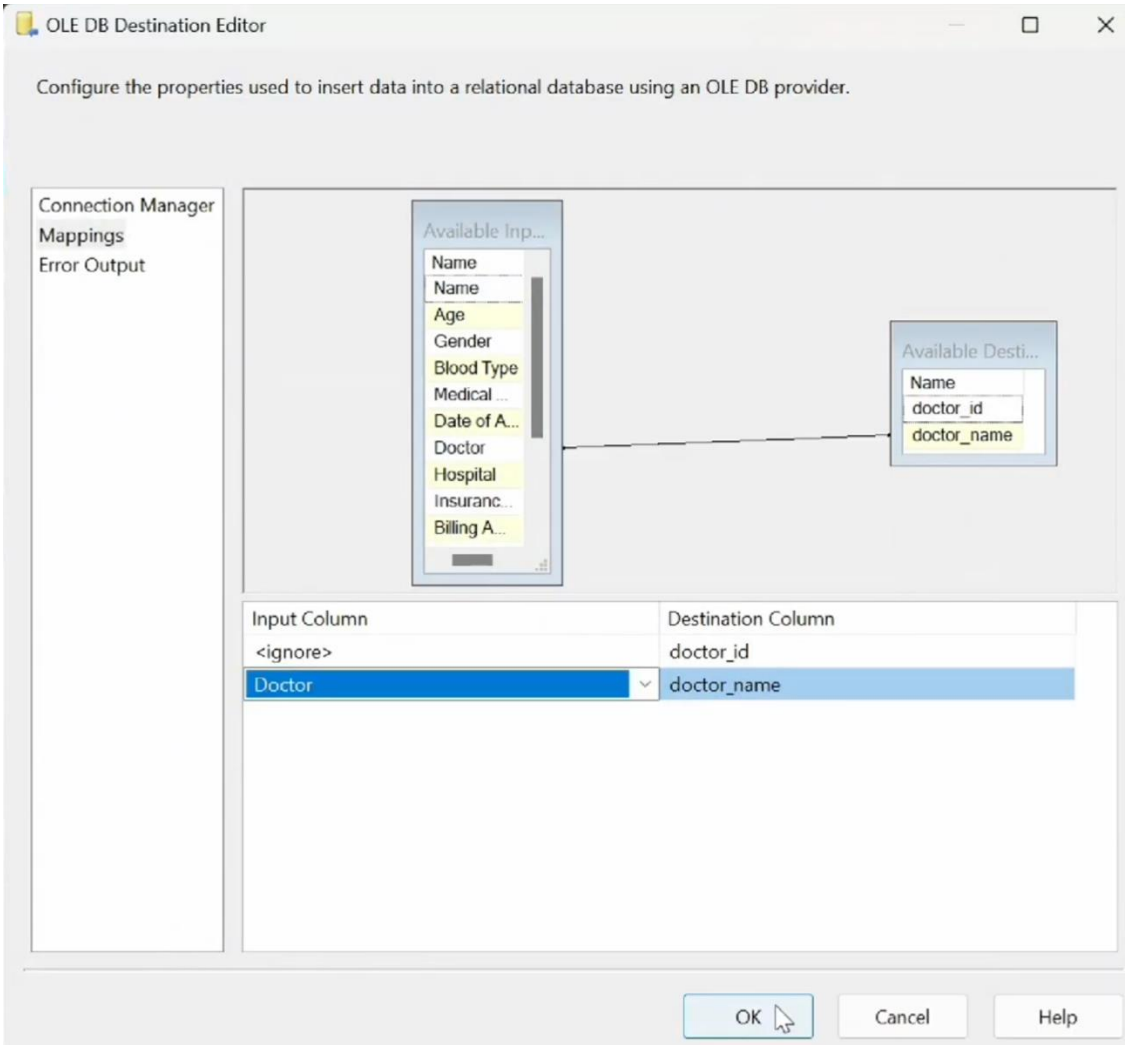

*Hình 24. OLE DB DESTINATION -DIM_DOCTOR*

- Bước 4: Tạo bảng DIM_DOCTOR




*Hình 25. Tạo bảng DIM_DOCTOR*

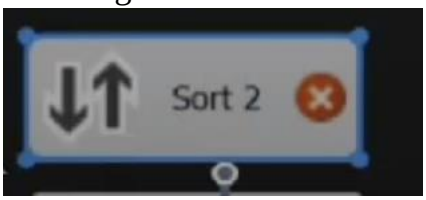
- Bước 5: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.




*Hình 26. Mapping thuộc tính*

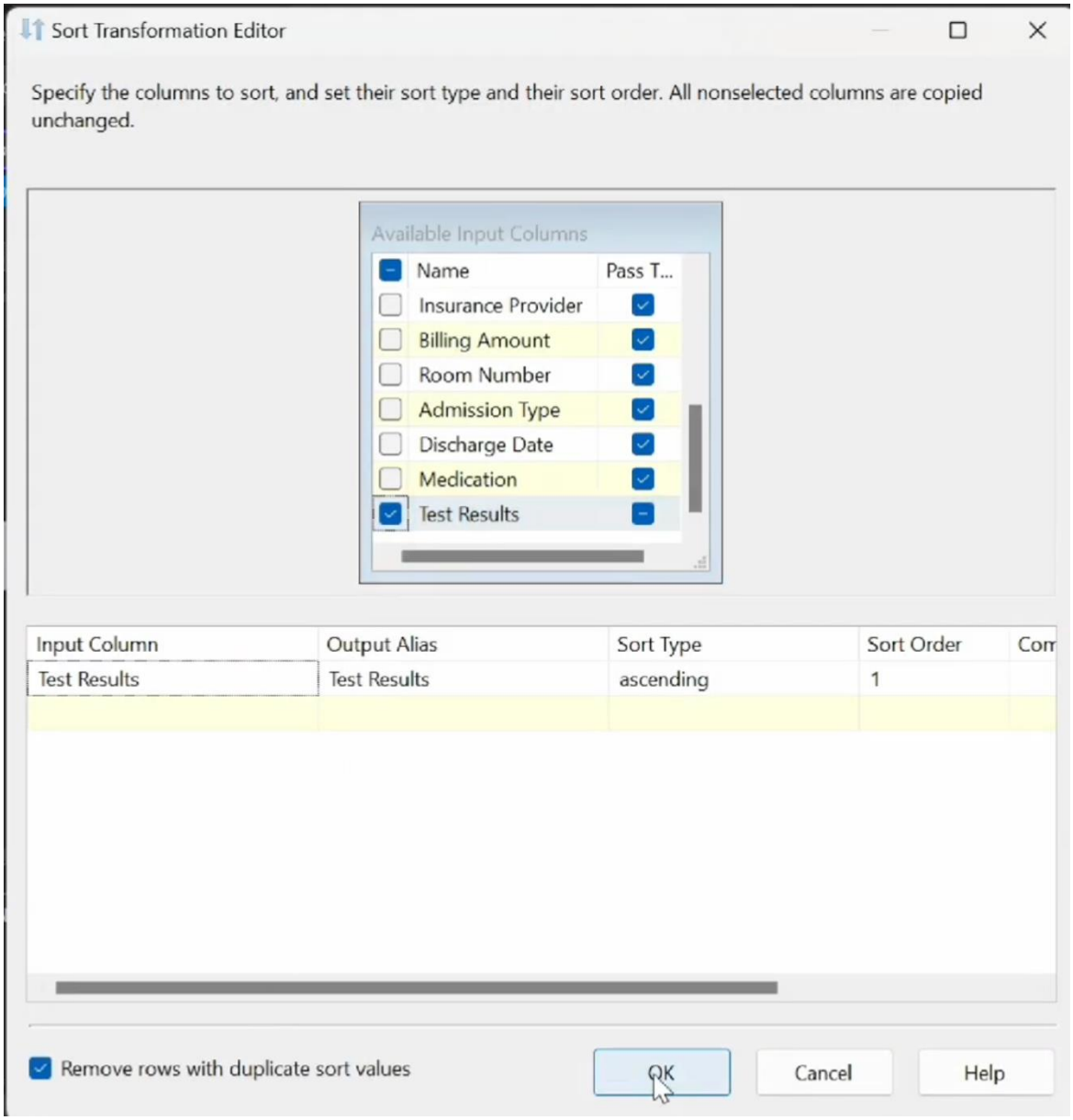
### 3.3. DIM_TEST_RESULT

- Bước 1 : Kéo chức năng Sort sang.



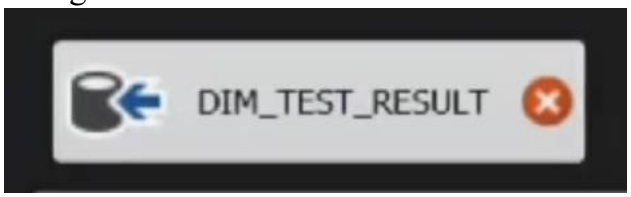

*Hình 27 . Kéo chức năng Sort*

- Bước 2 : Sort theo Test Result, tích vào Remove rows with duplicate sort values và nhấn OK.



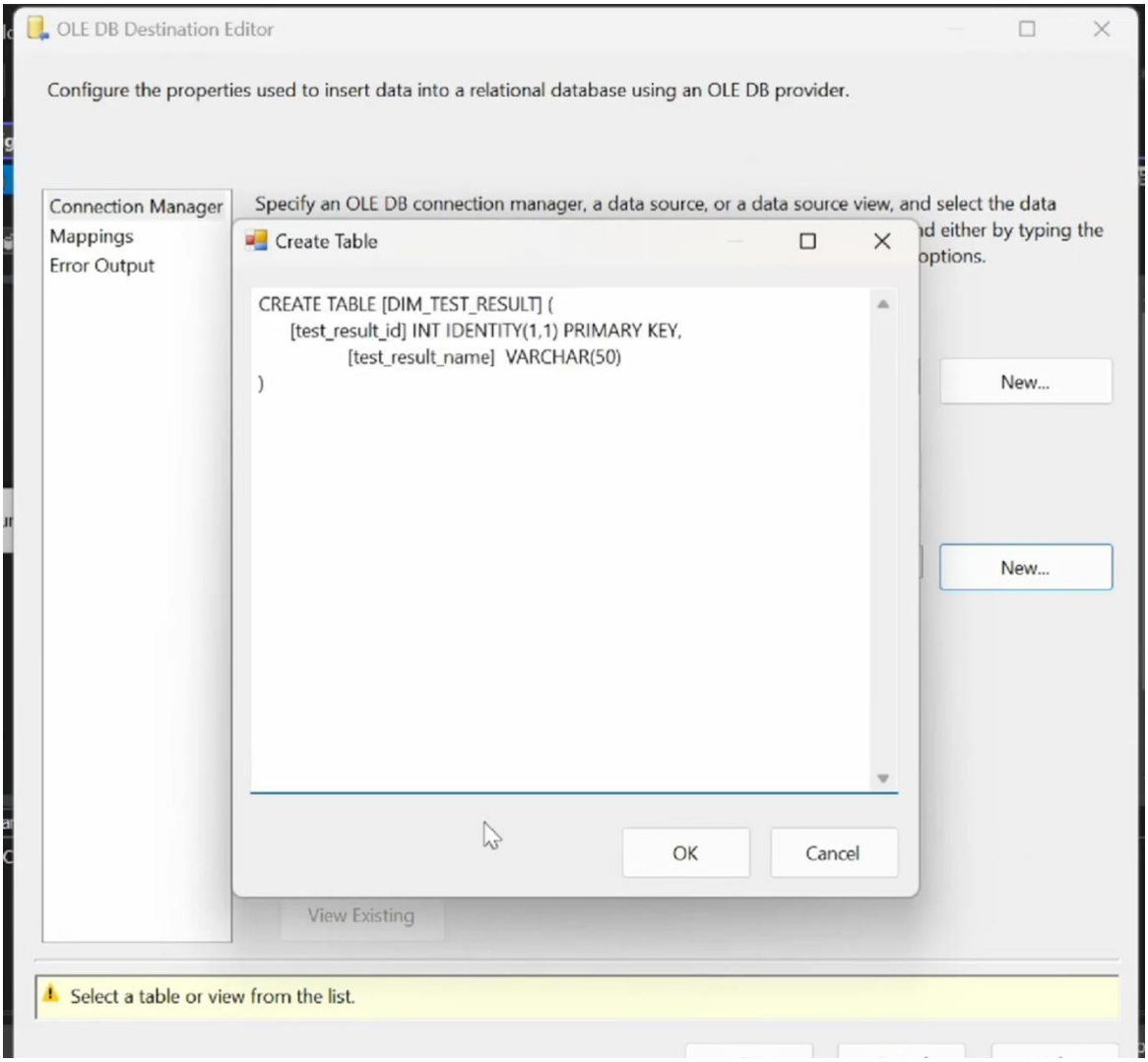

*Hình 28. Sort theo Test_Results*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_TEST_RESULT



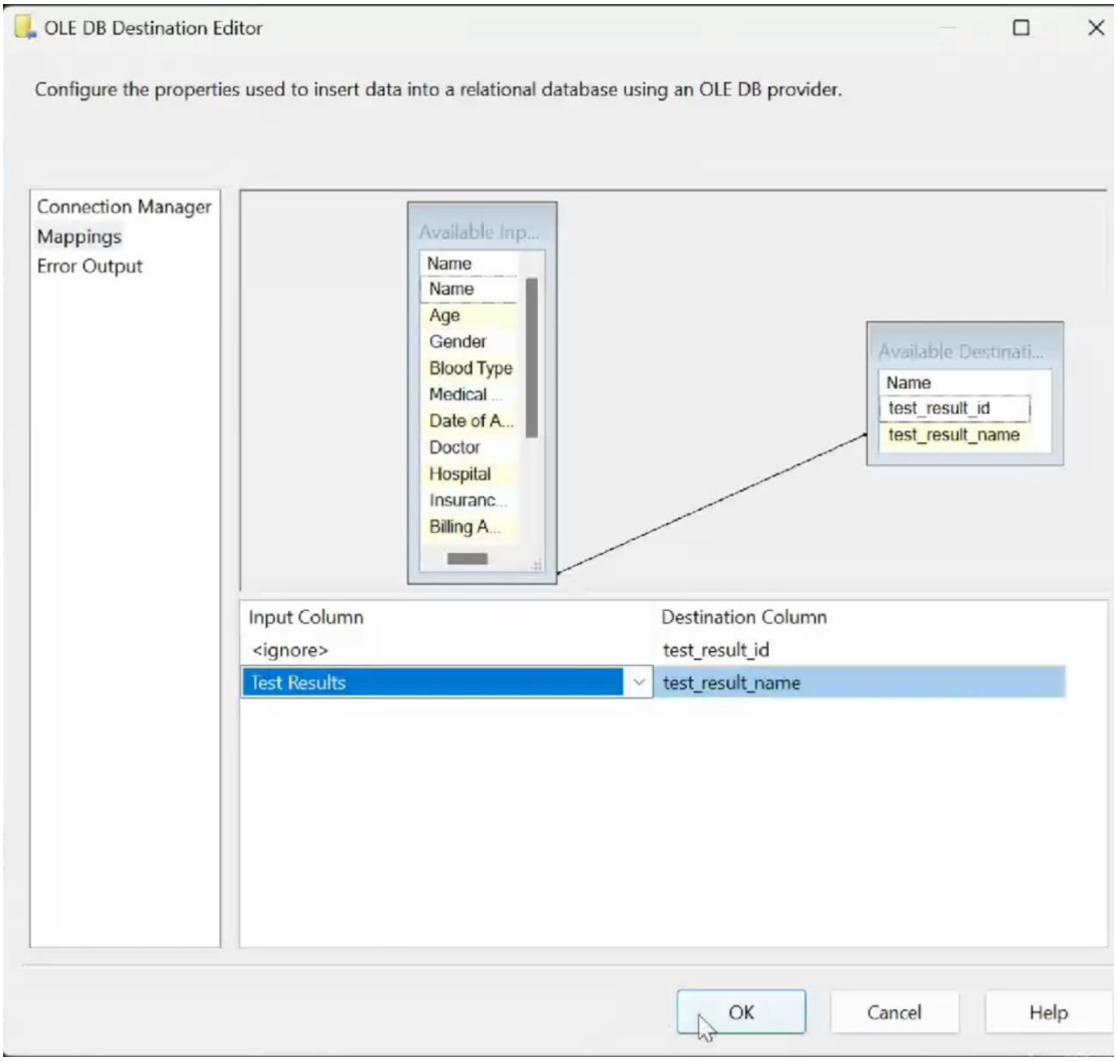

*Hình 29. OLE DB Destination- DIM_TEST_RESULT*

- Bước 4: Tạo bảng DIM_TEST_RESULT




*Hình 30. Tạo bảng DIM_TEST_RESULT*

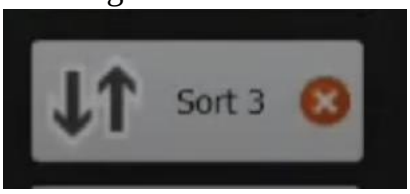
- Bước 5: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.




*Hình 31. Mapping thuộc tính*

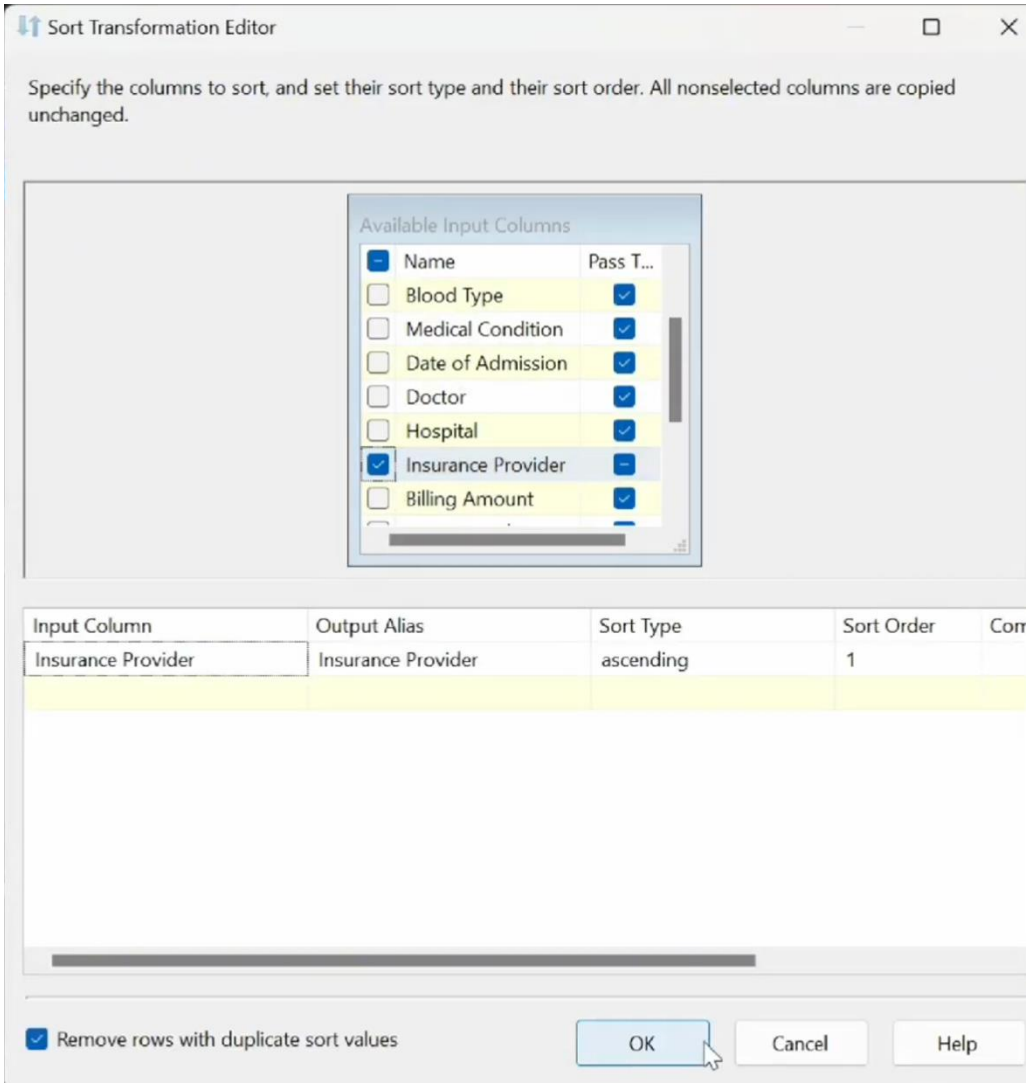
## 3.4. DIM_ INSURANCE_PROVIDER

- Bước 1 : Kéo chức năng Sort sang.



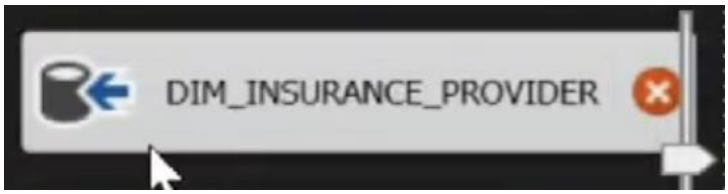

*Hình 32. Kéo chức năng Sort*

- Bước 2 : Sort theo Insurance Provider, tích vào Remove rows with duplicate sort values và nhấn OK.



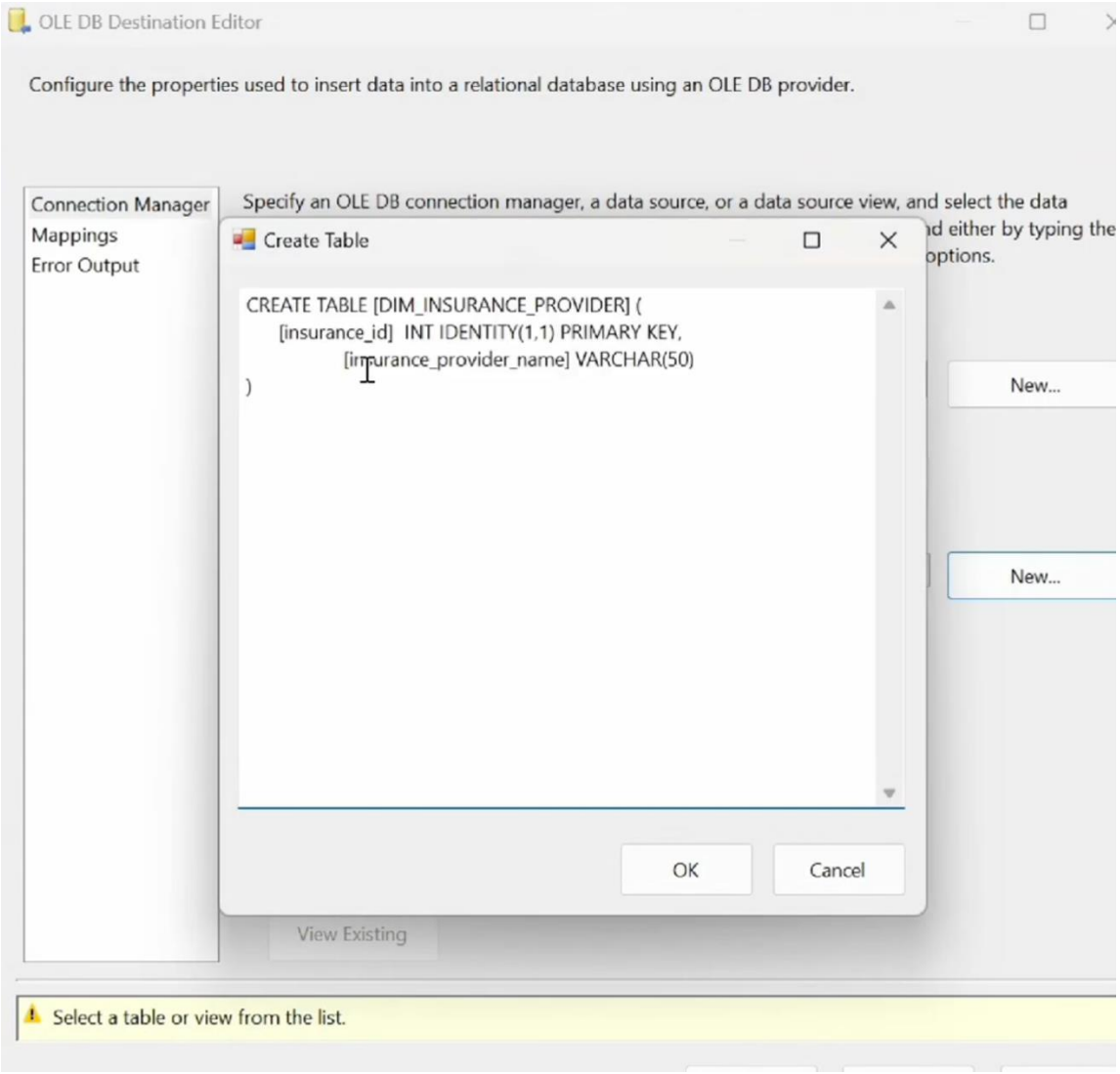

*Hình 33. Sort theo Insurance_Provider*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_INSURANCE_PROVIDER



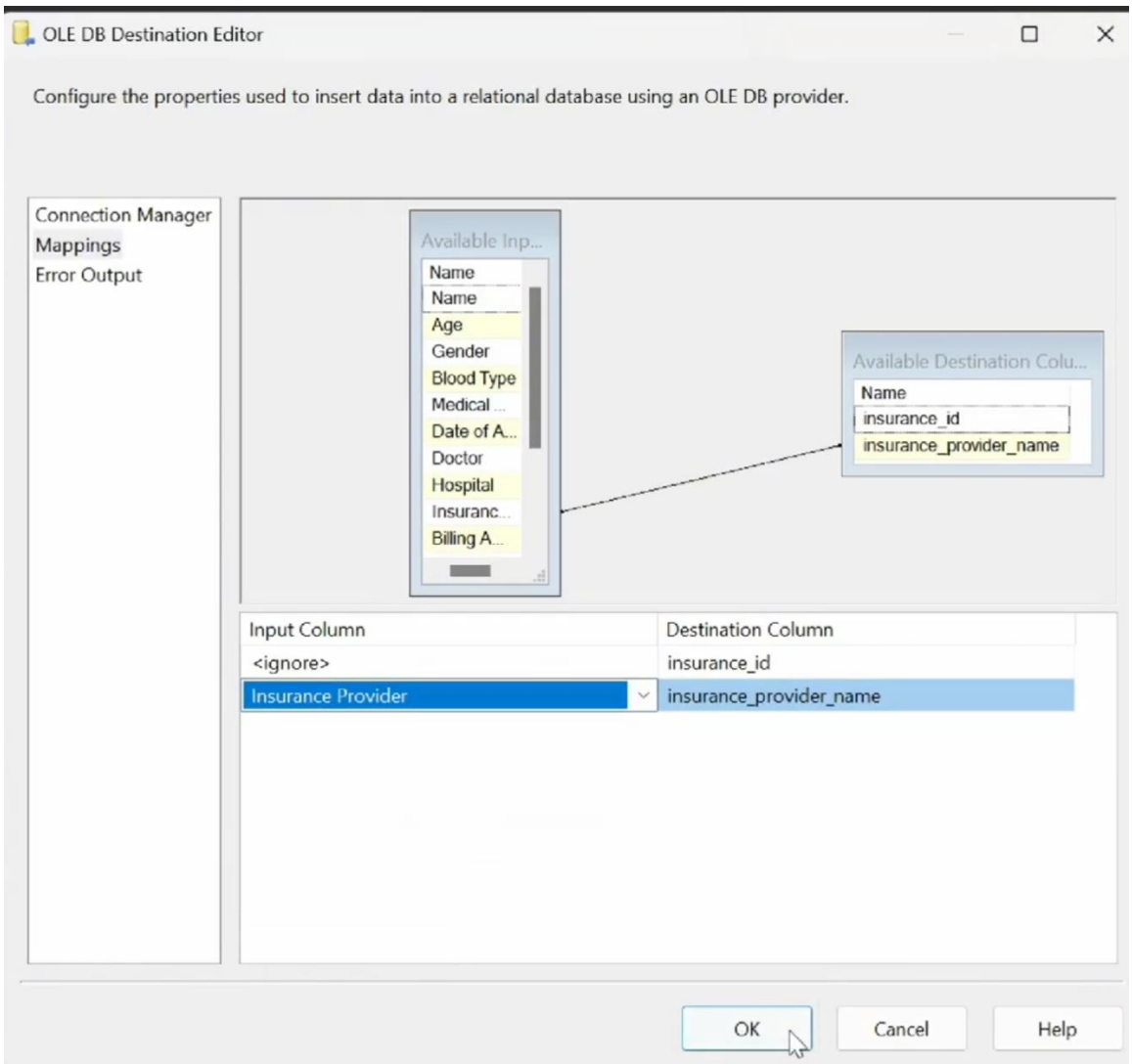

*Hình 34. OLE DB DESTINATION – DIM_INSURANCE_PROVIDER*

- Bước 4: Tạo bảng DIM_ INSURANCE_PROVIDER




*Hình 35. Tạo bảng DIM_INSURANCE_PROVIDER*

- Bước 5: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.



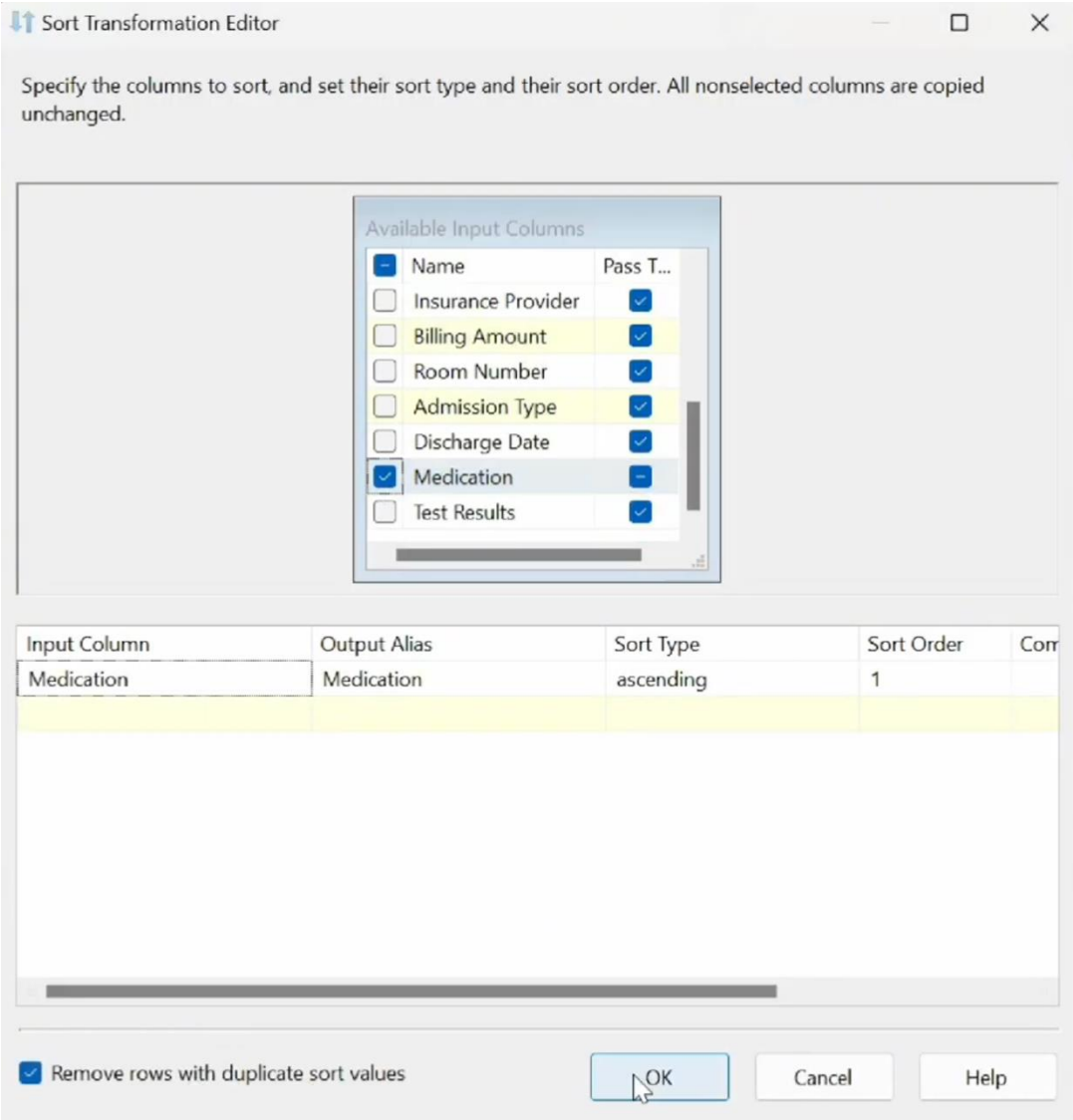

*Hình 36. Mapping thuộc tính*

## 3.5. DIM_MEDICATION

- Bước 1 : Kéo chức năng Sort sang.

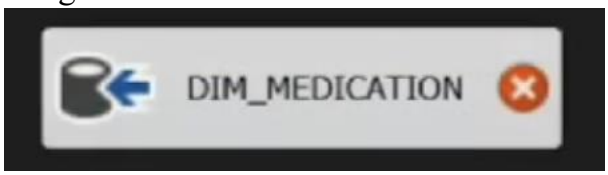
*Hình 37 . Kéo chức năng Sort*

- Bước 2 : Sort theo Medication, tích vào Remove rows with duplicate sort values và nhấn OK.



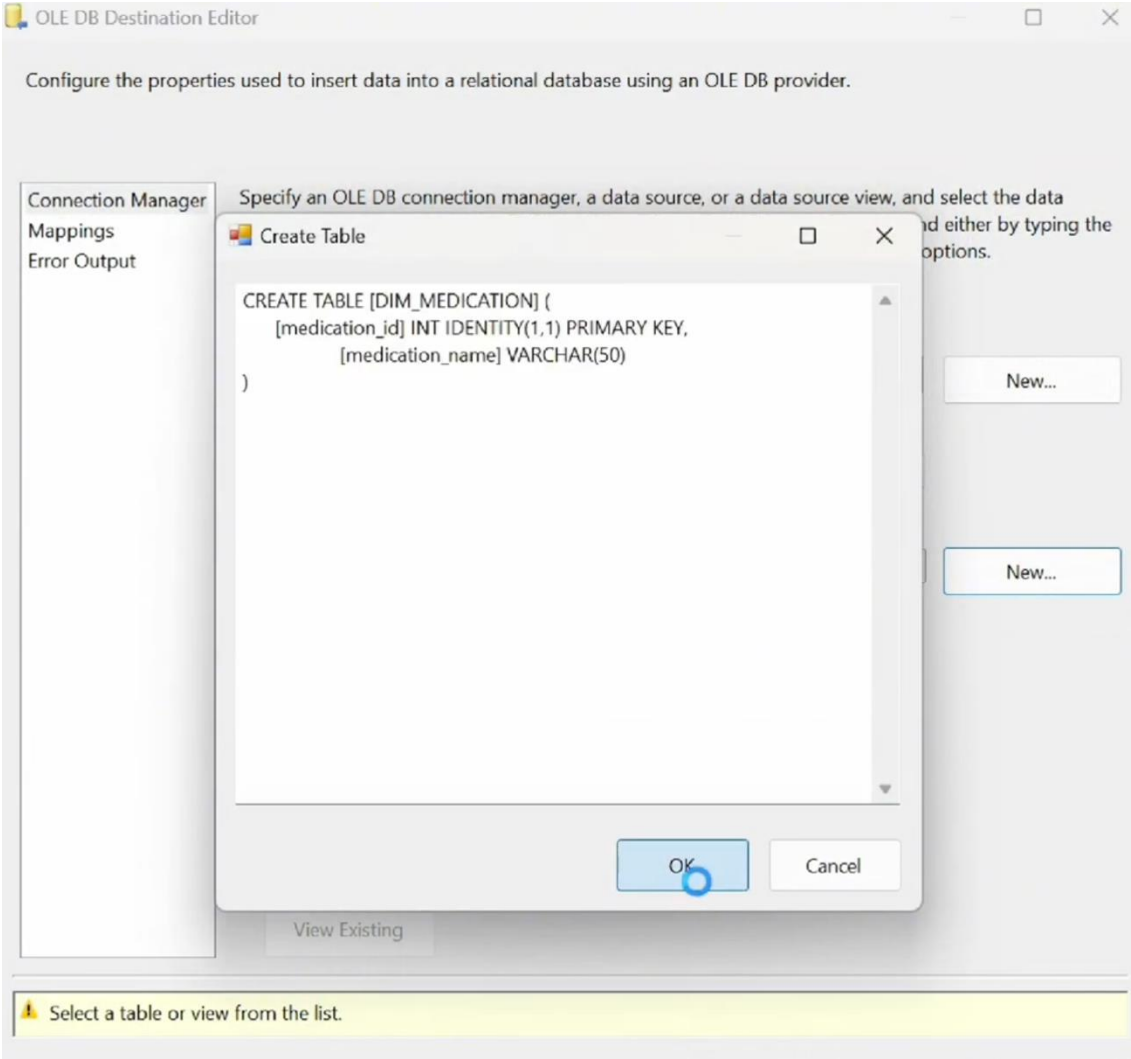

*Hình 38 . Sort theo Medication*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_MEDICATION



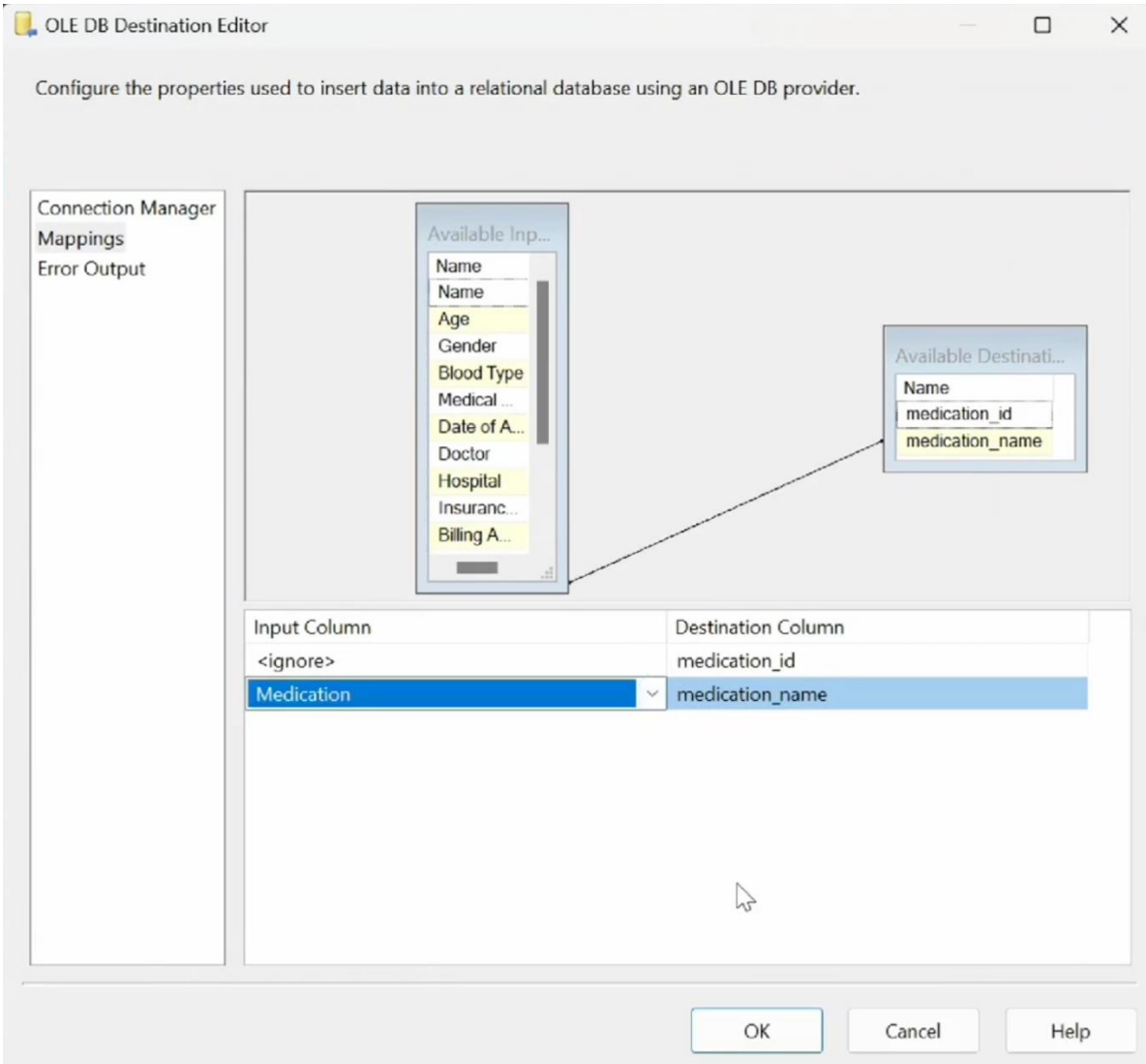

*Hình 39. OLE DB DESTINATION – DIM_MEDICATION*

- Bước 4: Tạo bảng DIM_ MEDICATION




*Hình 40. Tạo bảng DIM_MEDICATION*

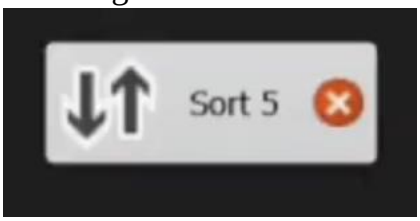
- Bước 5: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.




*Hình 41. Mapping thuộc tính*

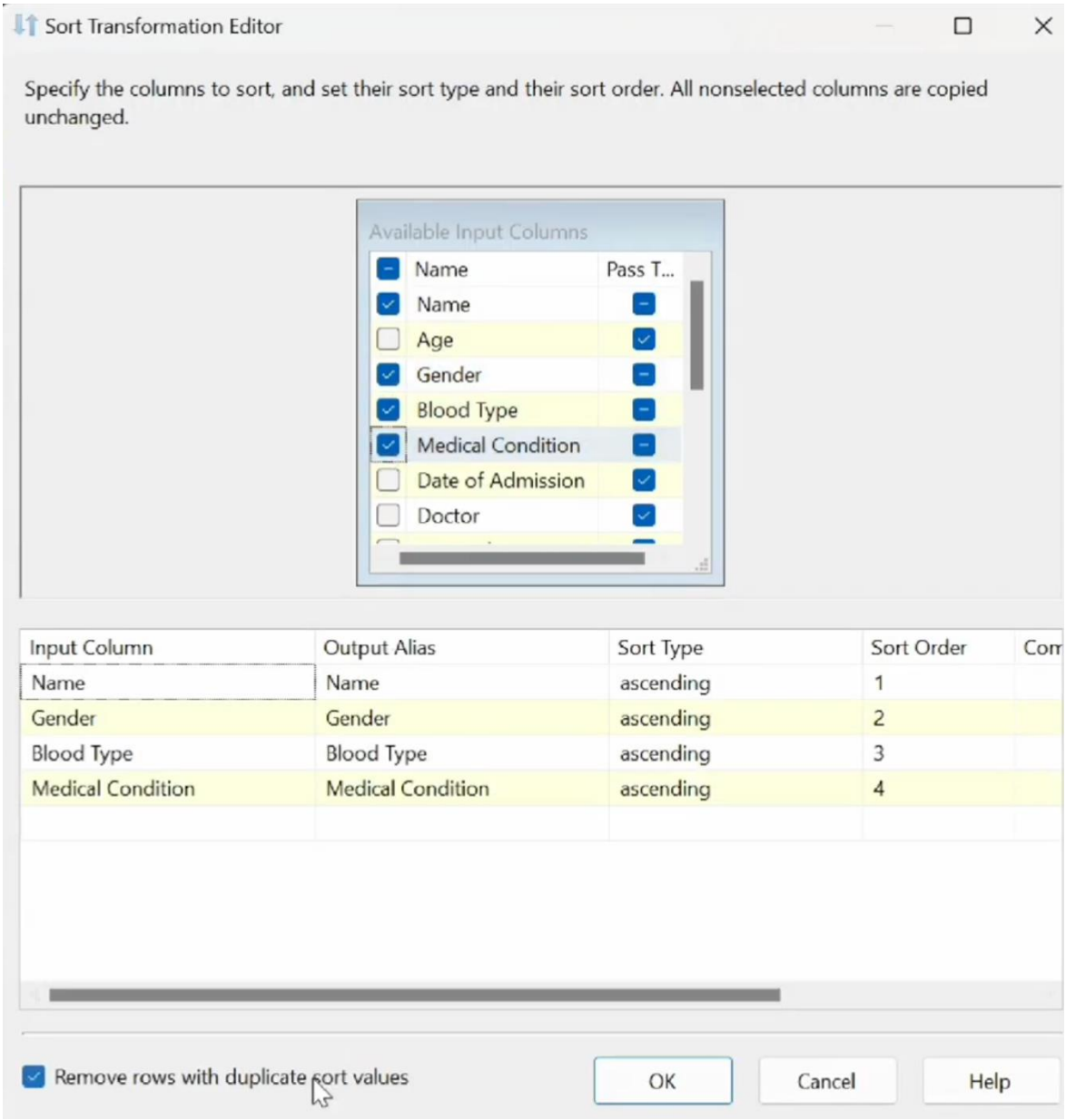
### 3.6. DIM_ PATIENT

- Bước 1 : Kéo chức năng Sort sang.



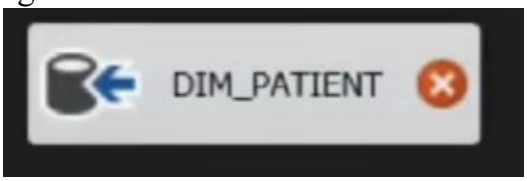

*Hình 42. Kéo chức năng Sort*

- Bước 2 : Sort theo Name, Gender, Blood Type, Medical Condition, tích vào Remove rows with duplicate sort values và nhấn OK.



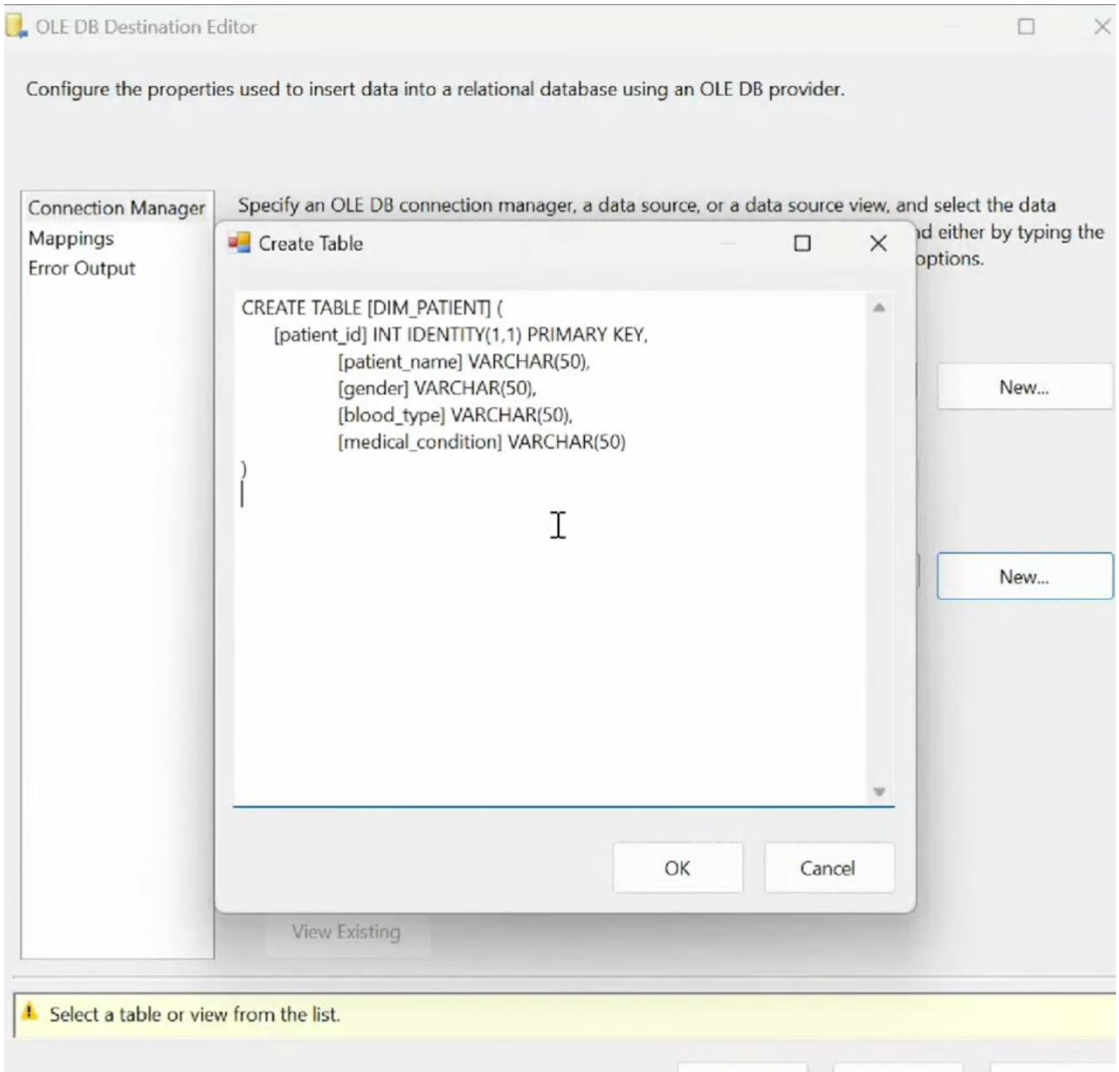

*Hình 43. Sort theo Name, Gender, BloodType, Medical Condition*

- Bước 3: Kéo chức năng OLE DB Destination và đổi tên thành DIM_PATIENT



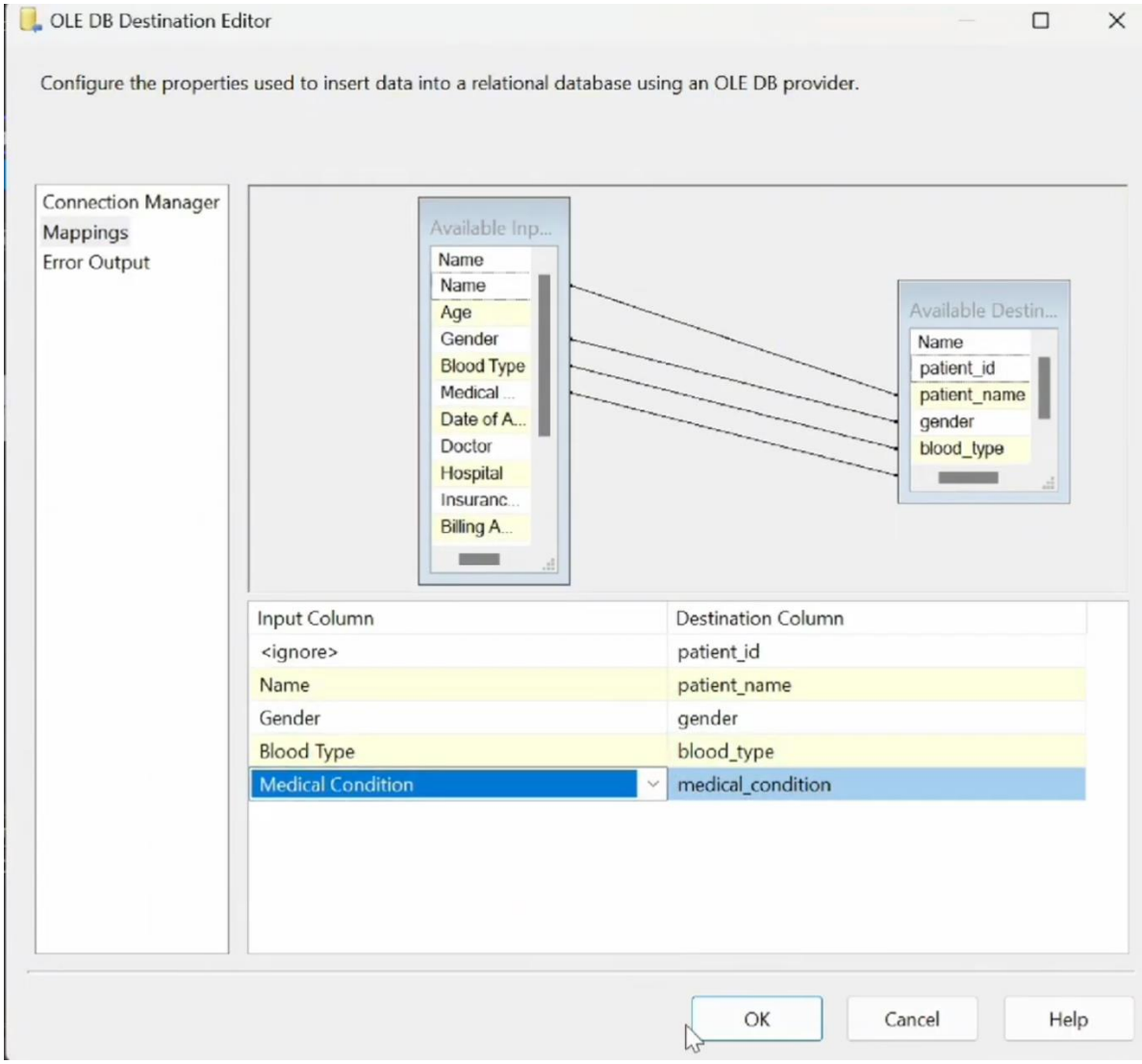

*Hình 44. OLE DB DESTINATION – DIM_PATIENT*

- Bước 4: Tạo bảng DIM_ PATIENT




*Hình 45. Tạo bảng DIM_PATIENT*

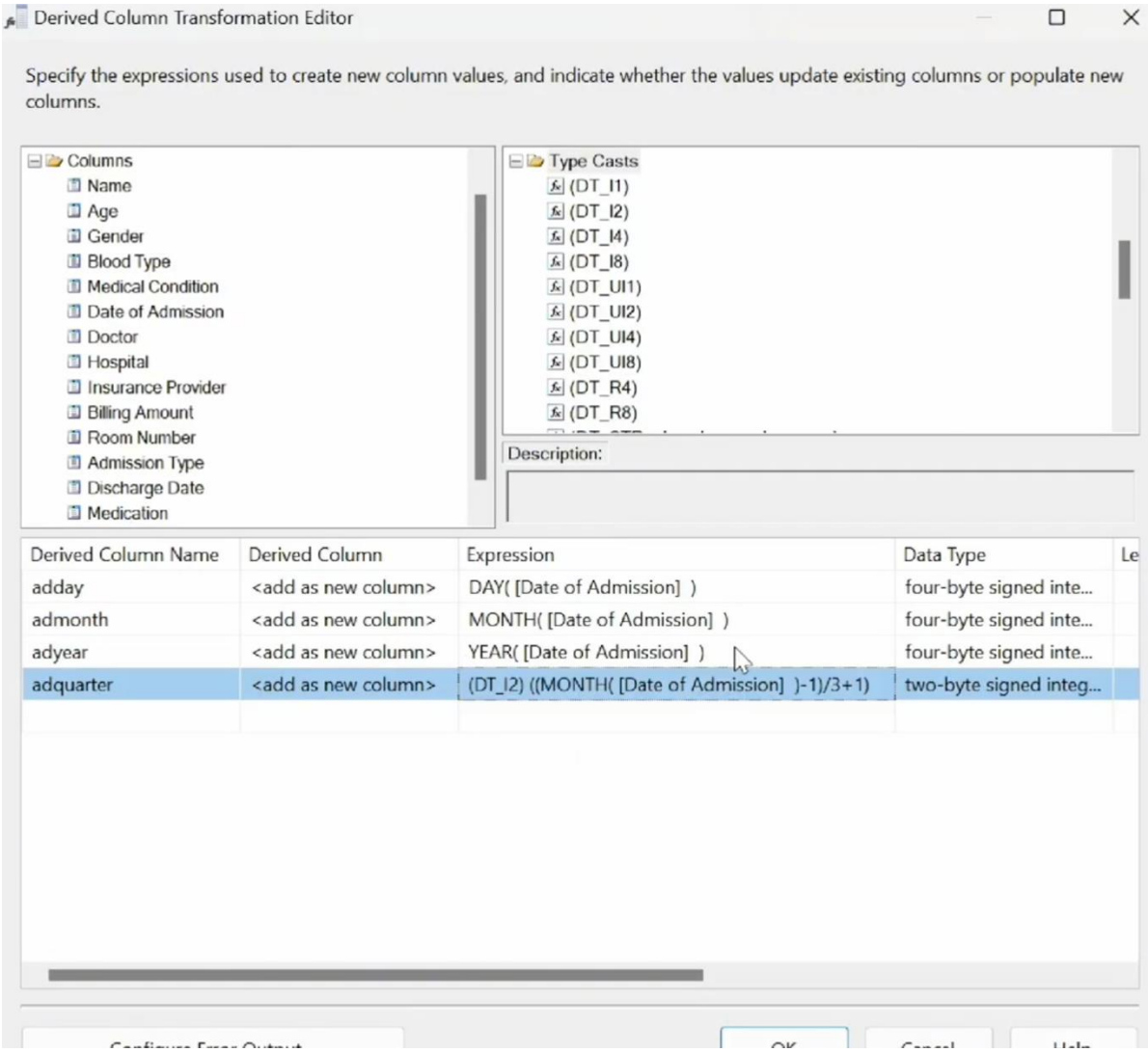
- Bước 5: Mapping lại thuộc tính của dữ liệu để phù hợp với thuộc tính đã tạo trong table.




*Hình 46. Mapping thuộc tính*

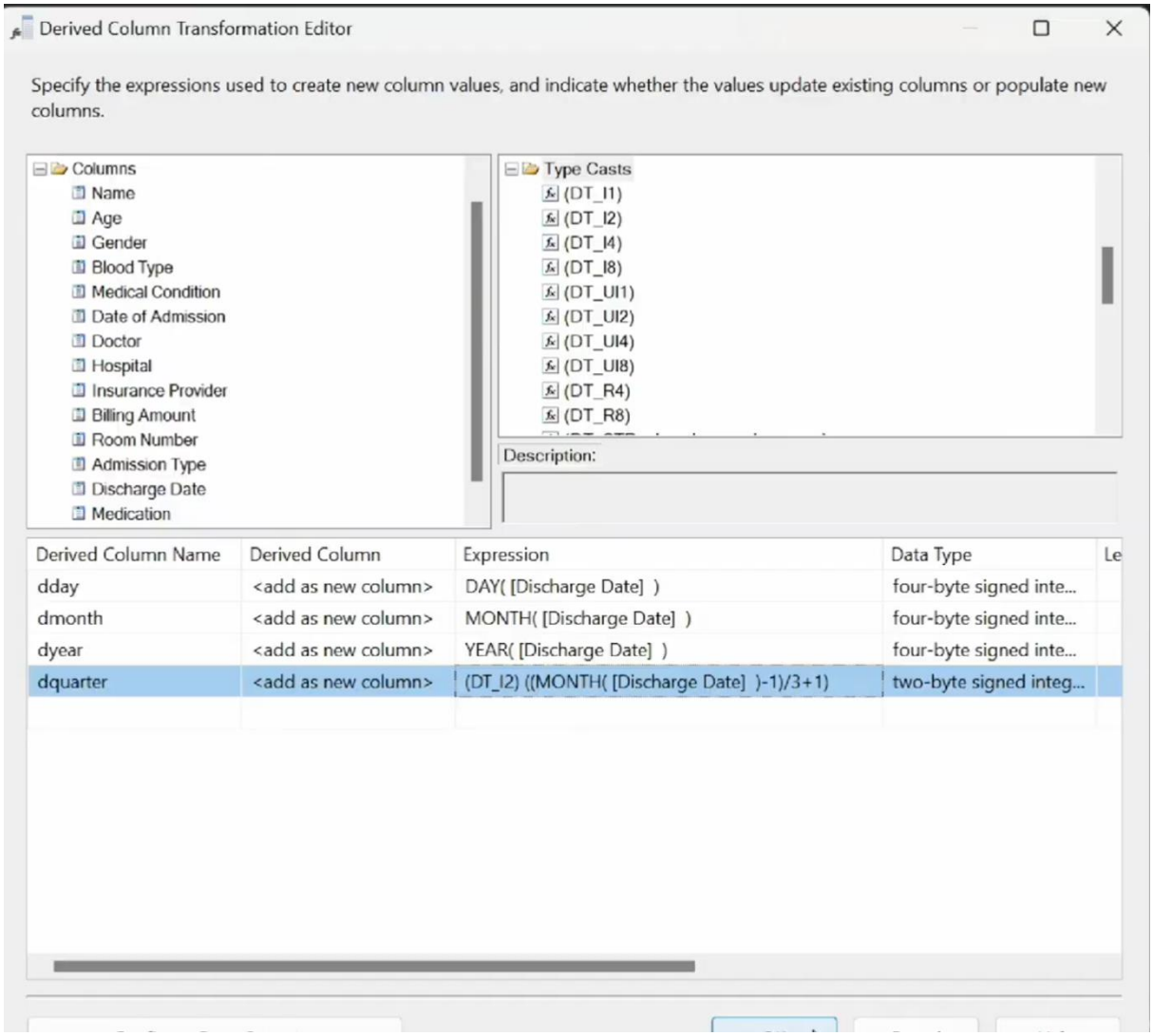
## 3.7. DIM_DATE

- Bước 1 : Tạo Derived Column và xử lí dữ liệu ngày, tháng, quý, năm của Admission Date



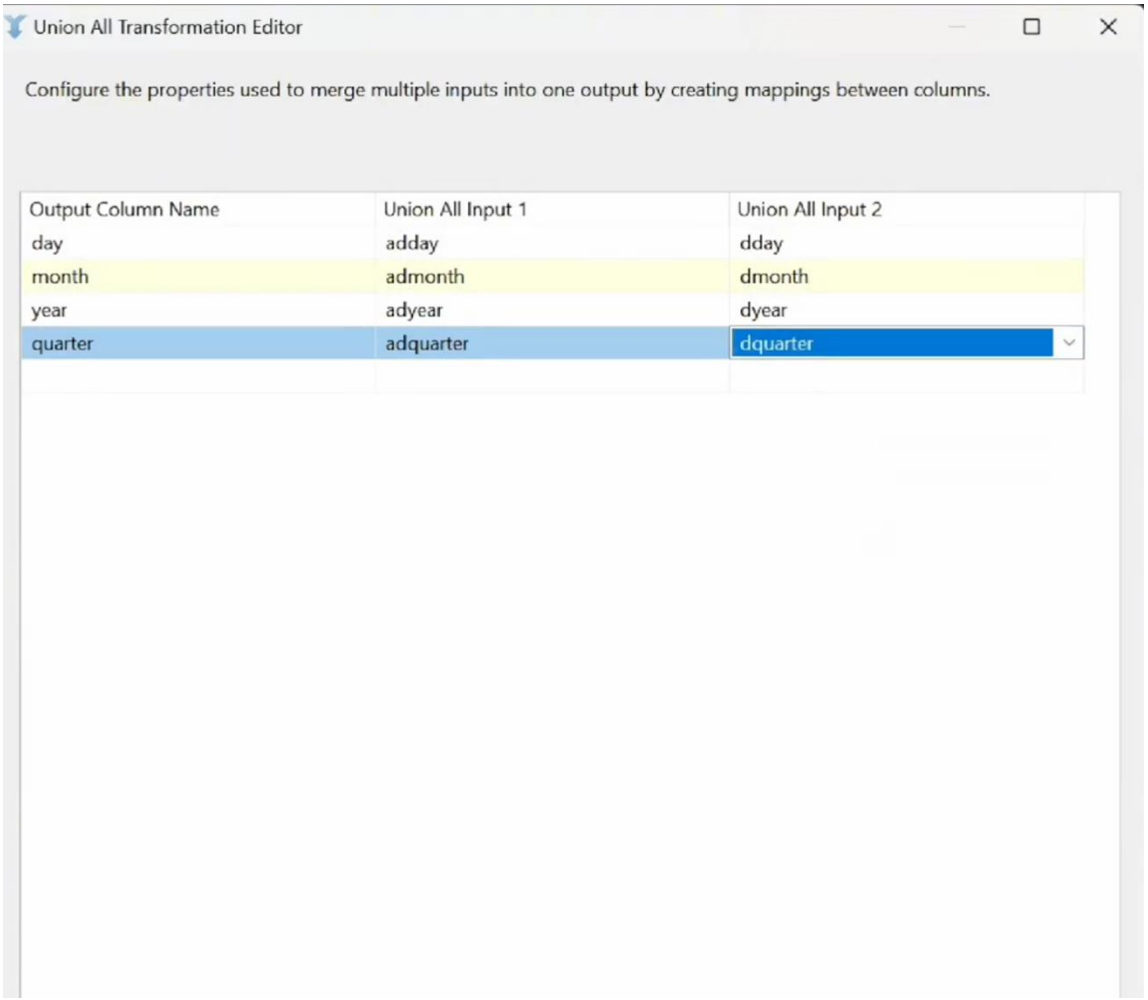

*Hình 47 . Xử lý dữ liệu Admission_Date*

- Bước 2 : Tạo Derived Column và xử lí dữ liệu ngày, tháng, quý, năm của Discharge Date



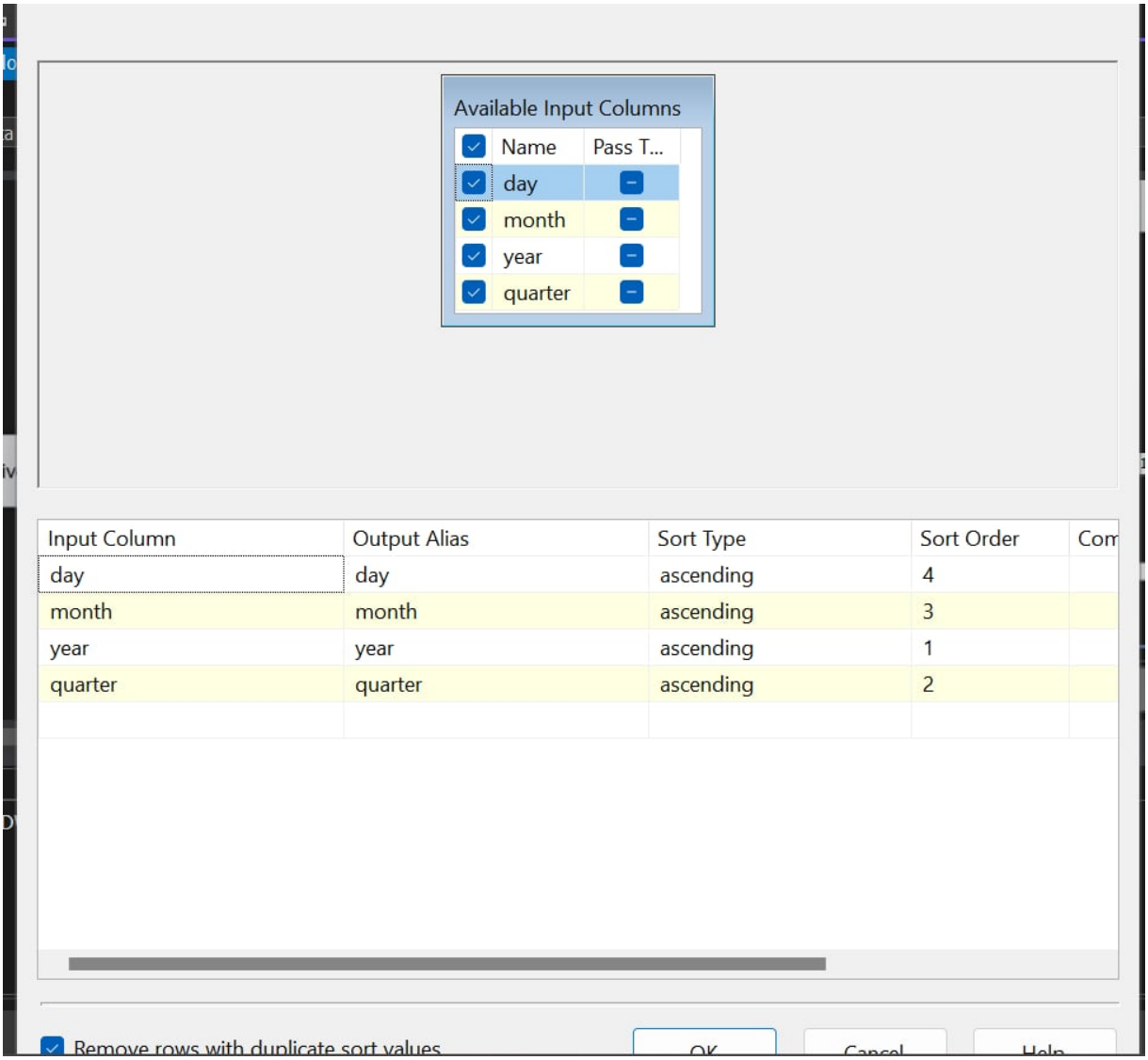

*Hình 48 . Xử lý dữ liệu Discharge Date*

- Bước 3 : Kéo Union All sang và gộp dữ liệu của 4 thuộc tính từ admission date và discharge date thành 4 thuộc tính chung là day, month, quarter, year.



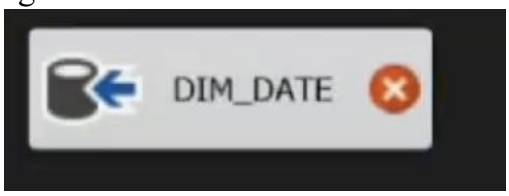

*Hình 49. Union All*

- Bước 4 : Sort theo Year, Quarter, Month, Day tích vào Remove rows with duplicate sort values và nhấn OK.



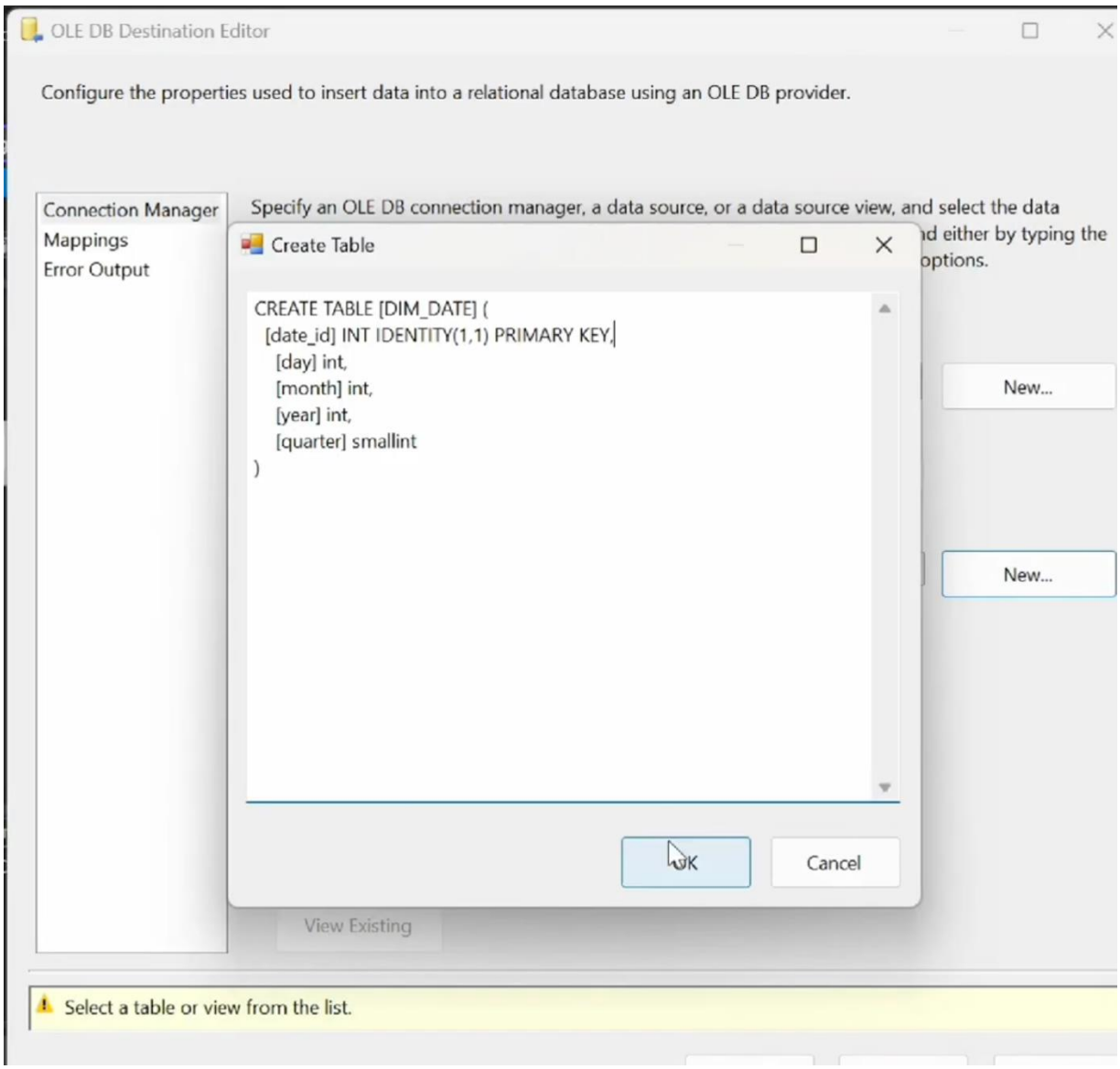

*Hình 50. Sort theo day , month, year, quarter*

- Bước 5 : Kéo chức năng OLE DB Destination và đổi tên thành DIM_DATE.




*Hình 51. OLE DB DESTINATION – DIM_DATE*

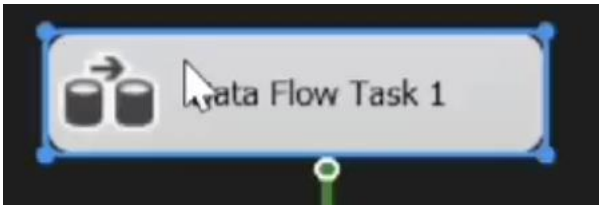
- Bước 6: Tạo bảng DIM_ DATE




*Hình 52 . Tạo bảng DIM,_DATE*

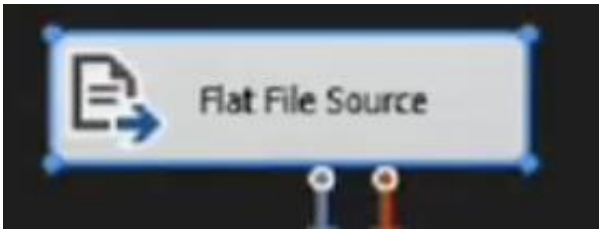
## 3.8. FACT

- Bước 1 : Kéo chức năng Data Flow Task từ thanh công cụ SSIS ToolBox sang màn hình làm việc.



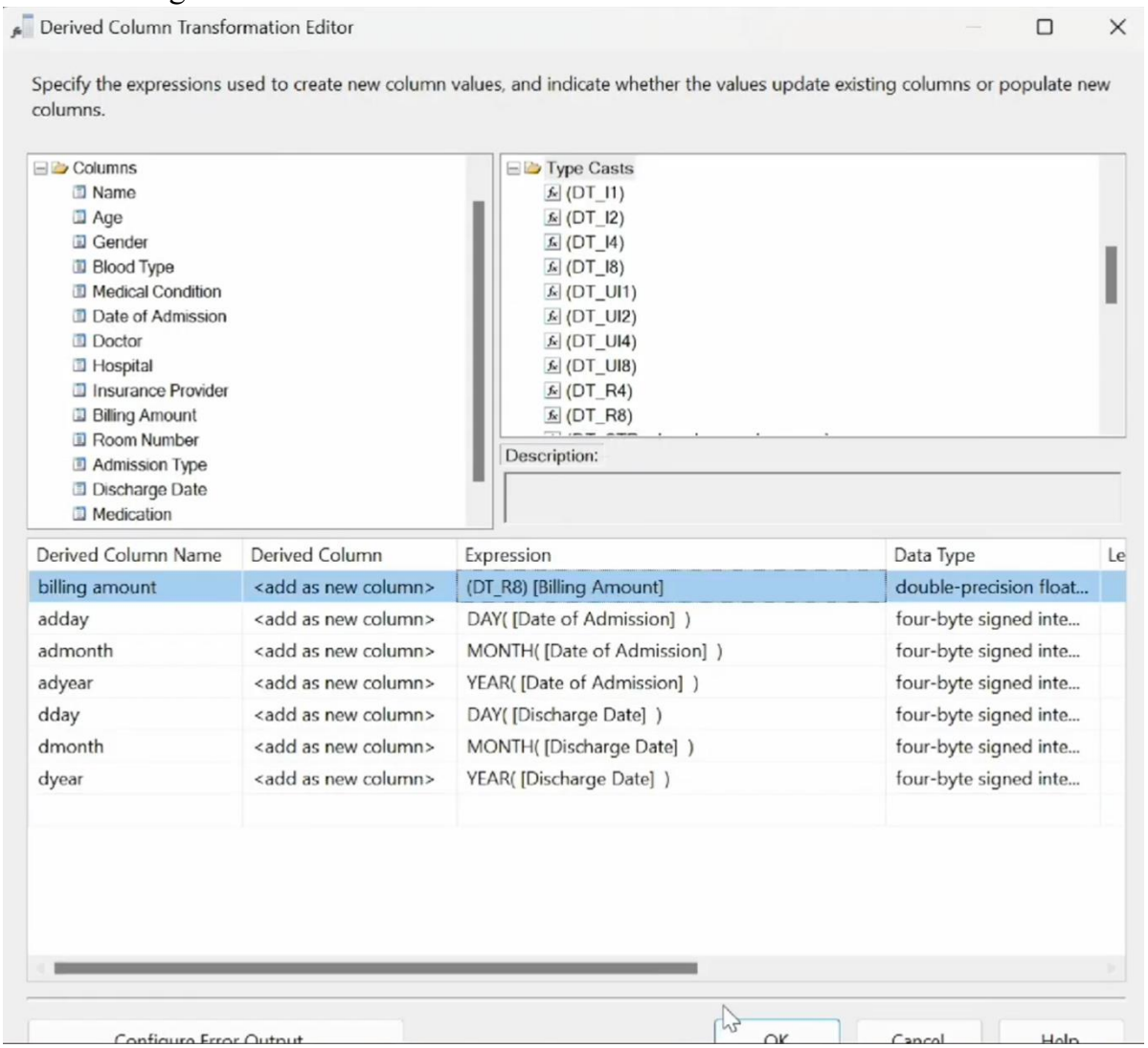

*Hình 53. Kéo chức năng Data Flow Task*

- Bước 2 : Click chuột vào ô Data Flow Task, màn hình chuyển sang cửa sổ Data Flow. Kéo thả chức năng **Flat File Source** từ cột trái sang và chọn dataset của mình



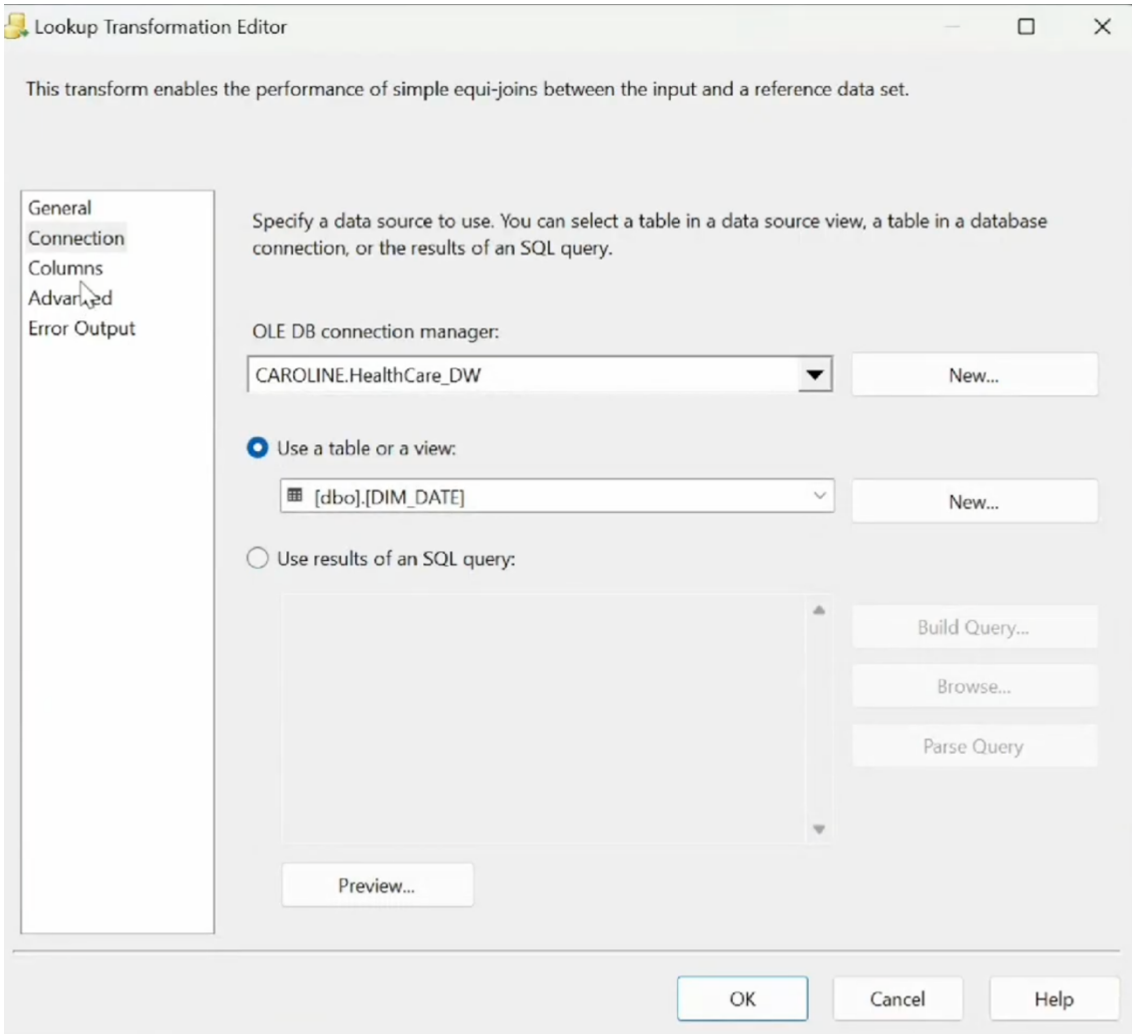

*Hình 54. Flat File Source*

- Bước 3 : Kéo chức năng **Devired Column** từ SSIS ToolBox sang màn hình làm việc, xử lí dữ liệu ngày tháng năm của **admission date** và **discharge date**, đổi kiểu dữ liệu của Billing Amount.



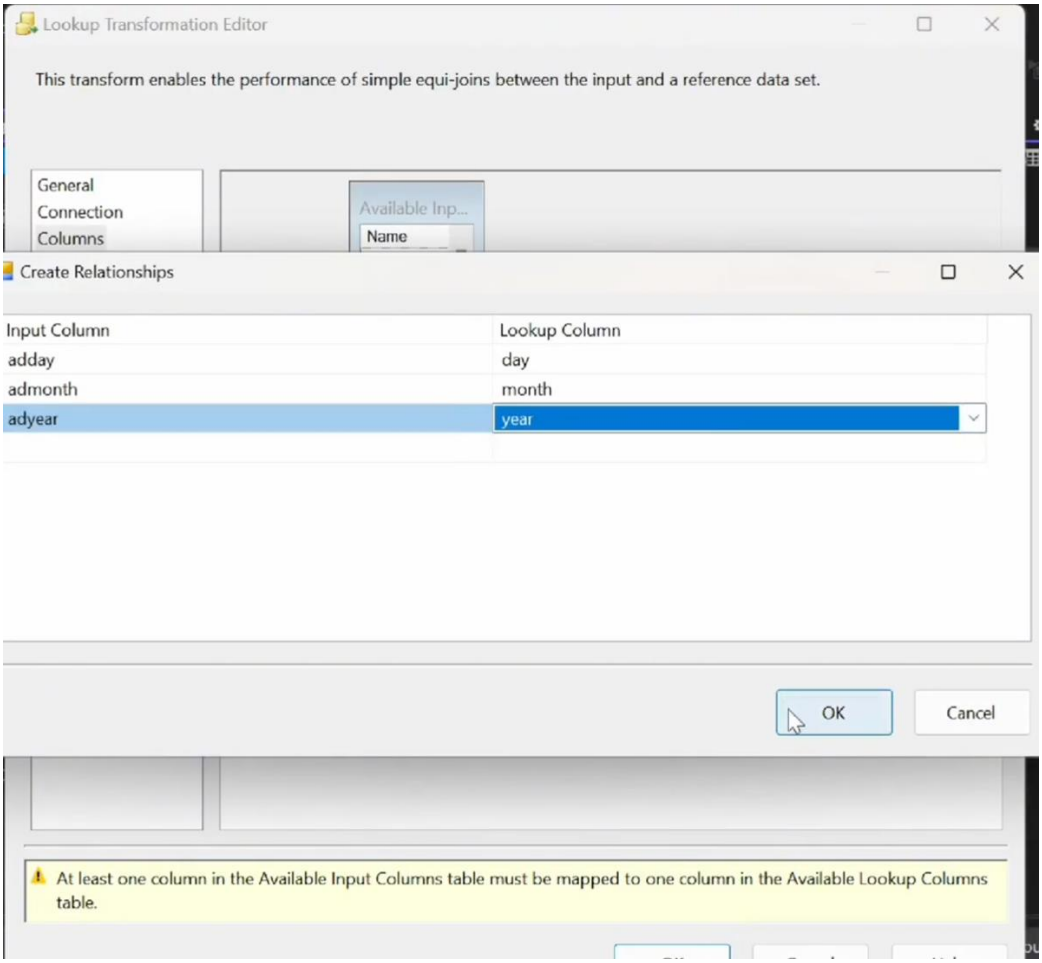

*Hình 55. Xử lý dữ liệu*

- Bước 5 : Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_DATE.



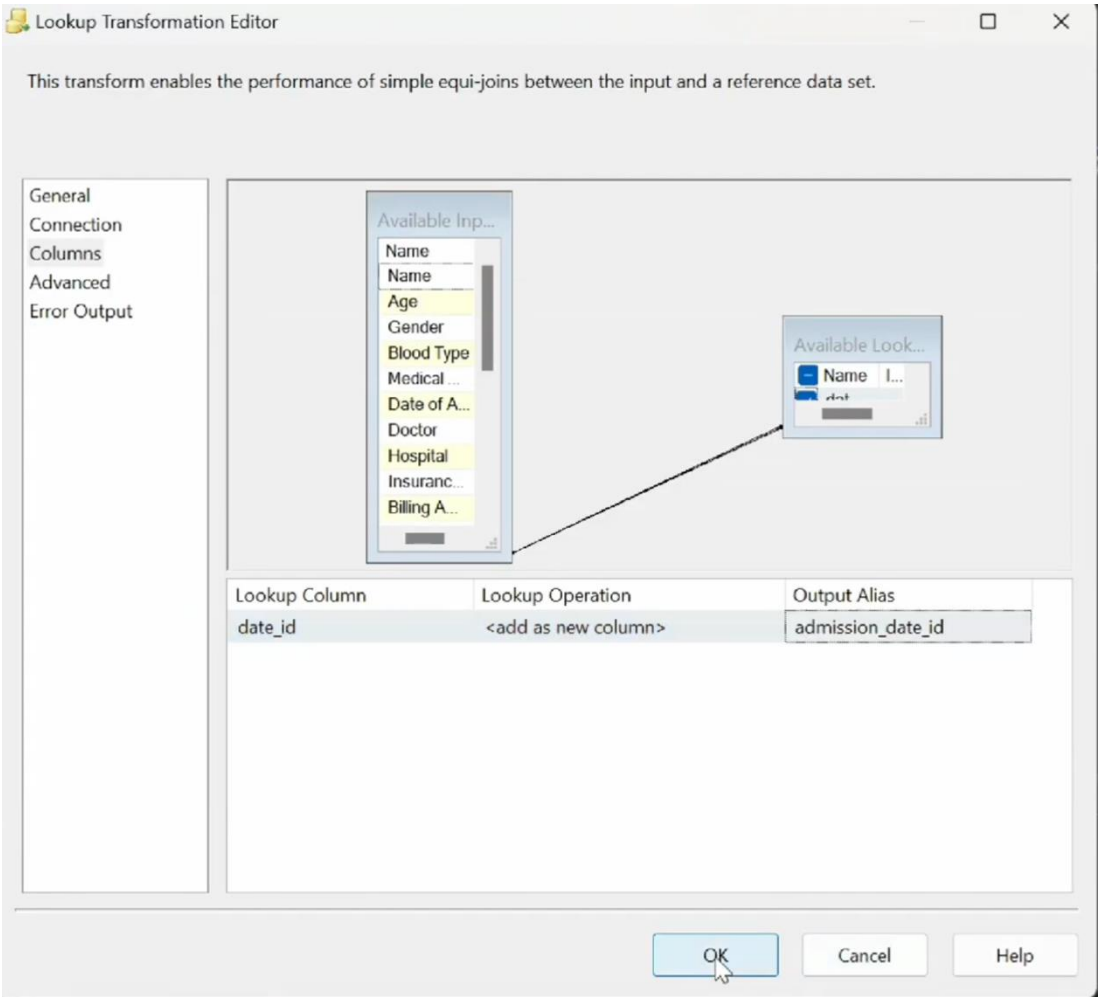

*Hình 56. Lookup*

- Bước 6: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của admission date.



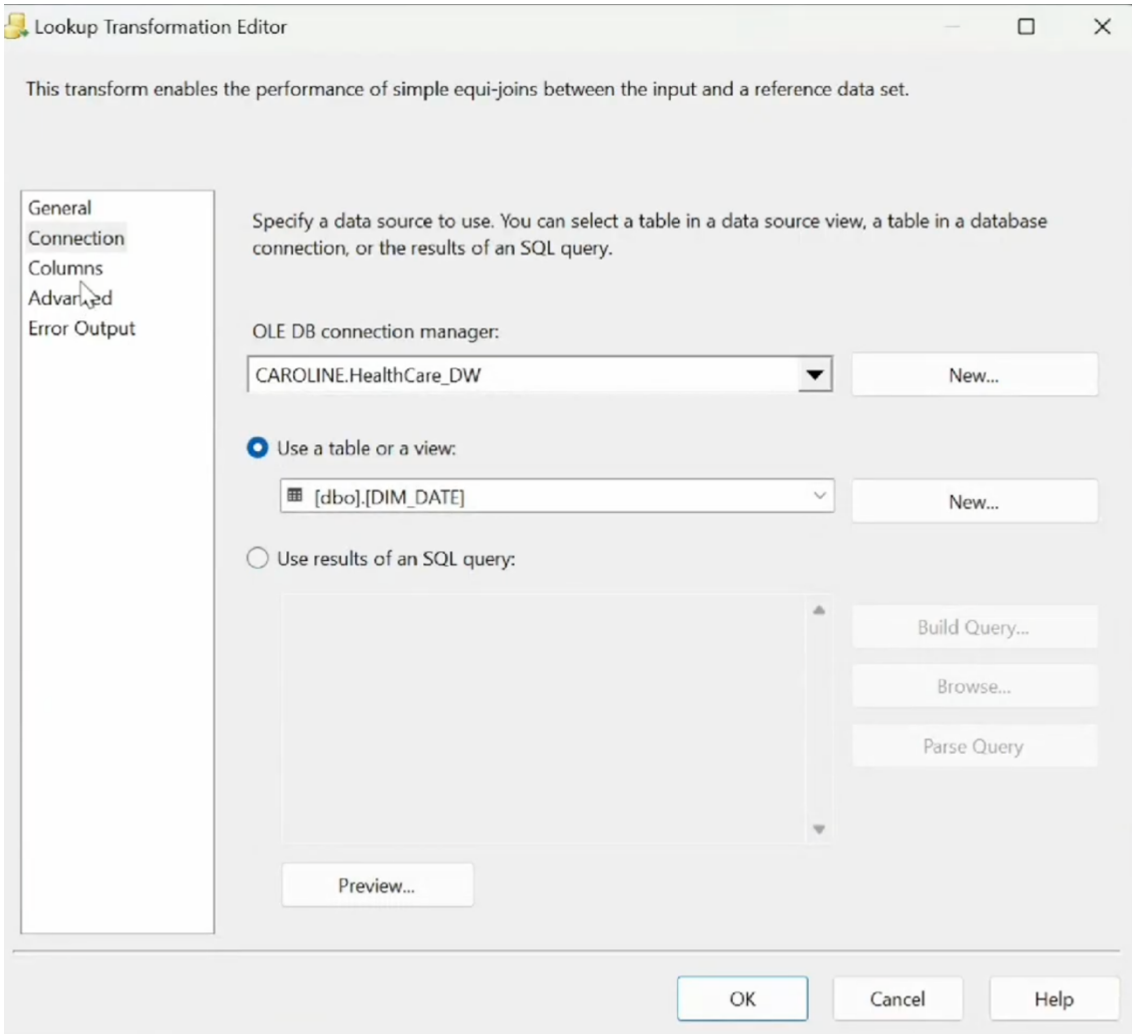

*Hình 57 . Edit Mapping*

- Bước 7 : Đổi tên ở ô **OutputAlias** từ date_id thành **admission_date_id.**



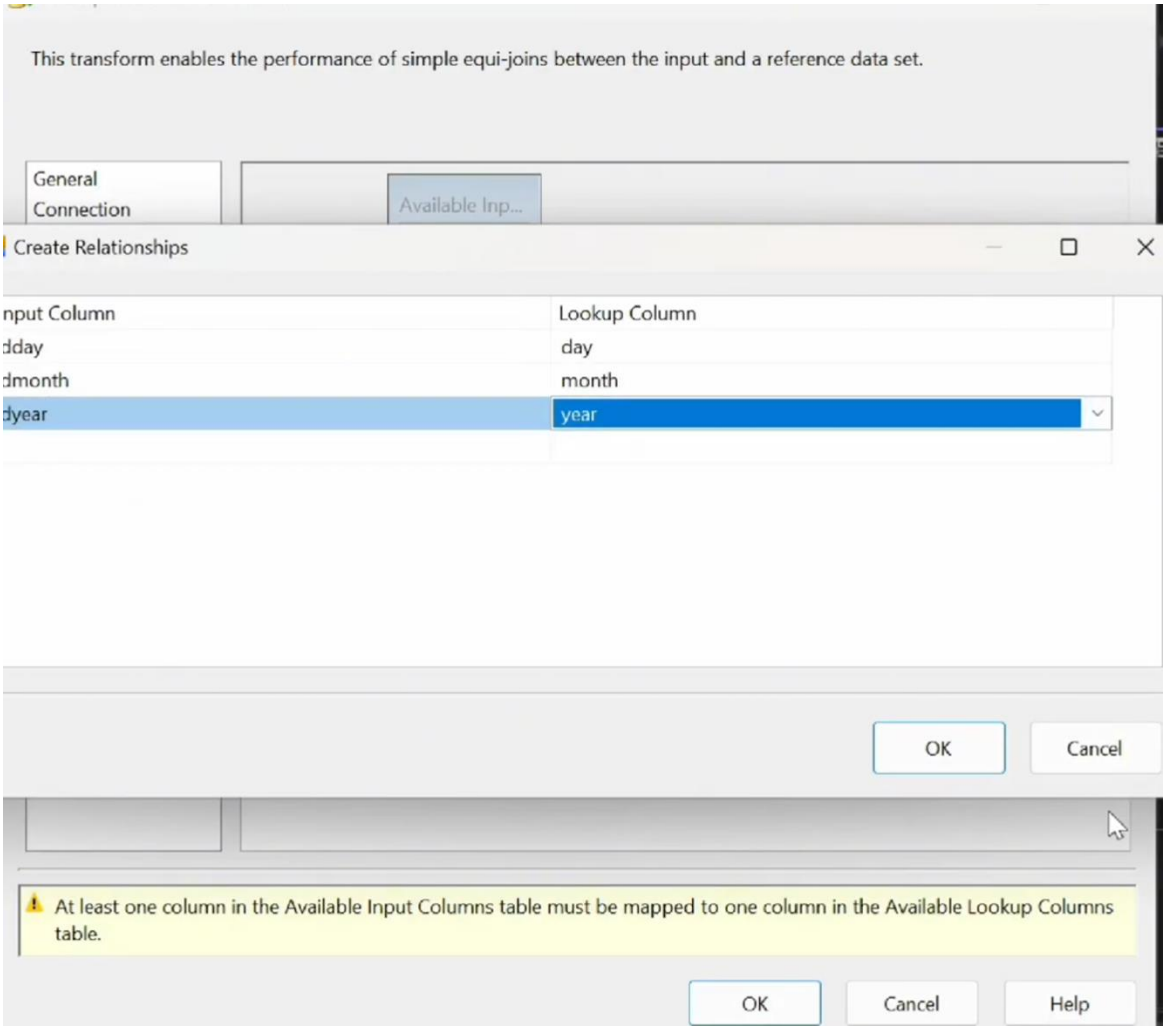

*Hình 58. Đổi tên ở Output Alias*

- Bước 8: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_DATE.



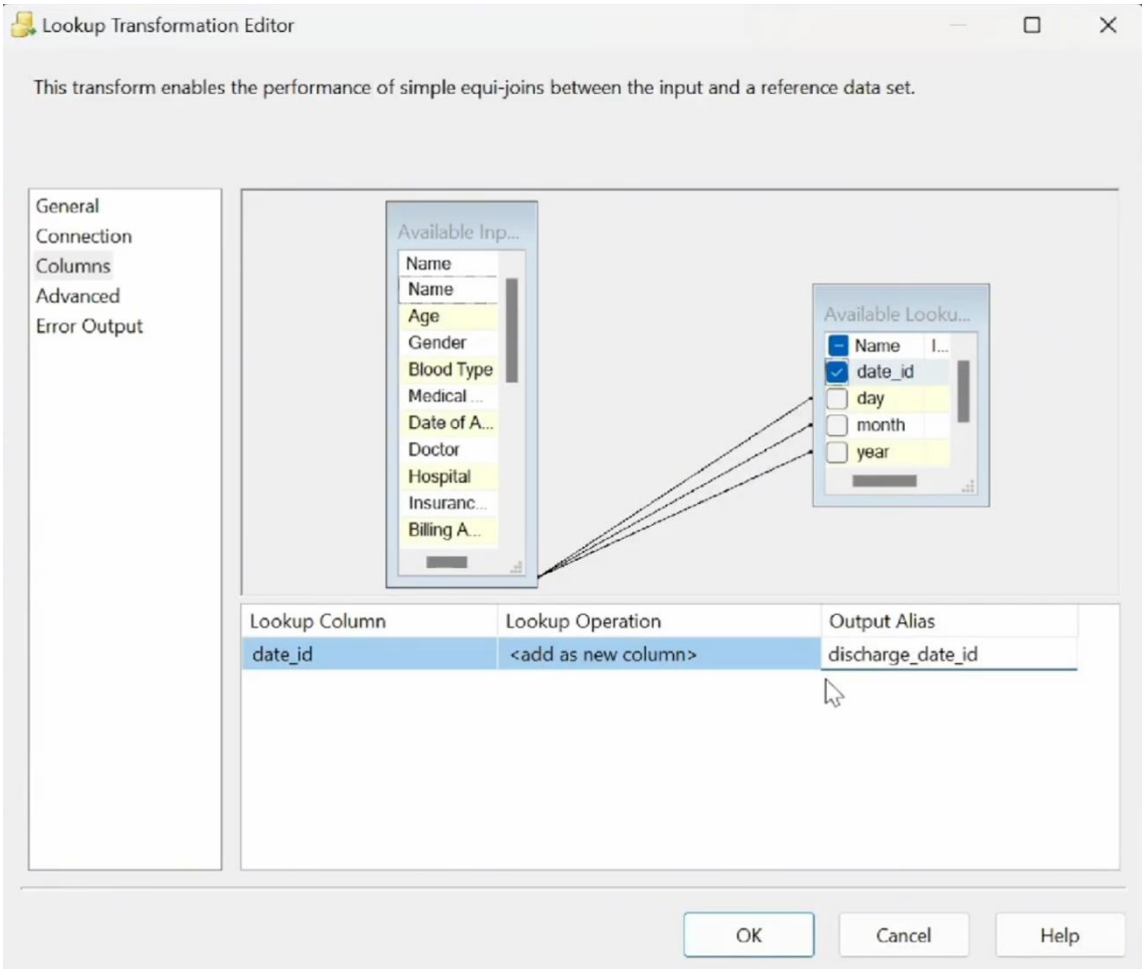

*Hình 59. Lookup*

- Bước 9: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của discharge date.



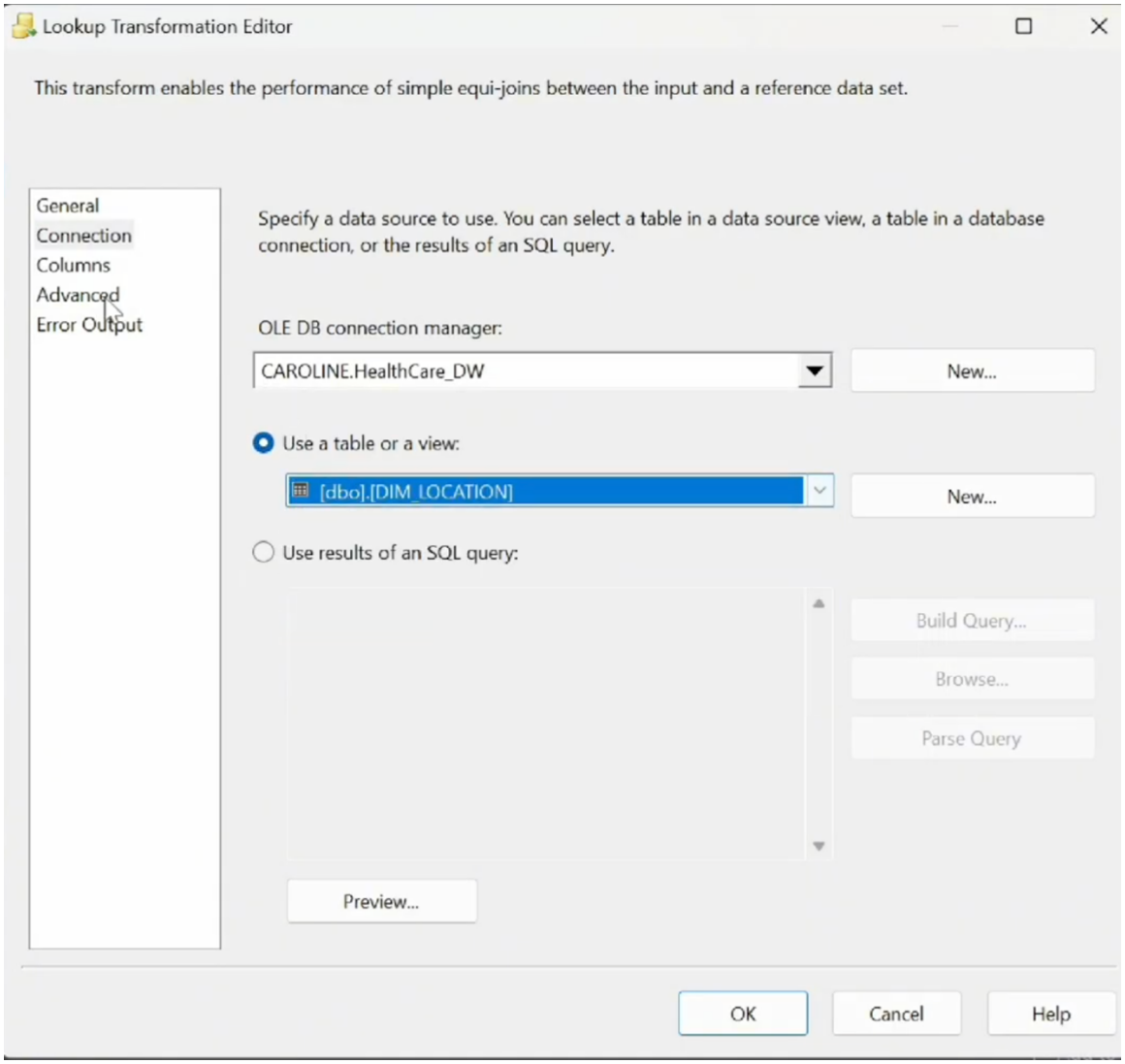

*Hình 60. Edit Mapping*

- Bước 10: Đổi tên ở ô **OutputAlias** từ date_id thành **discharge_date_id.**



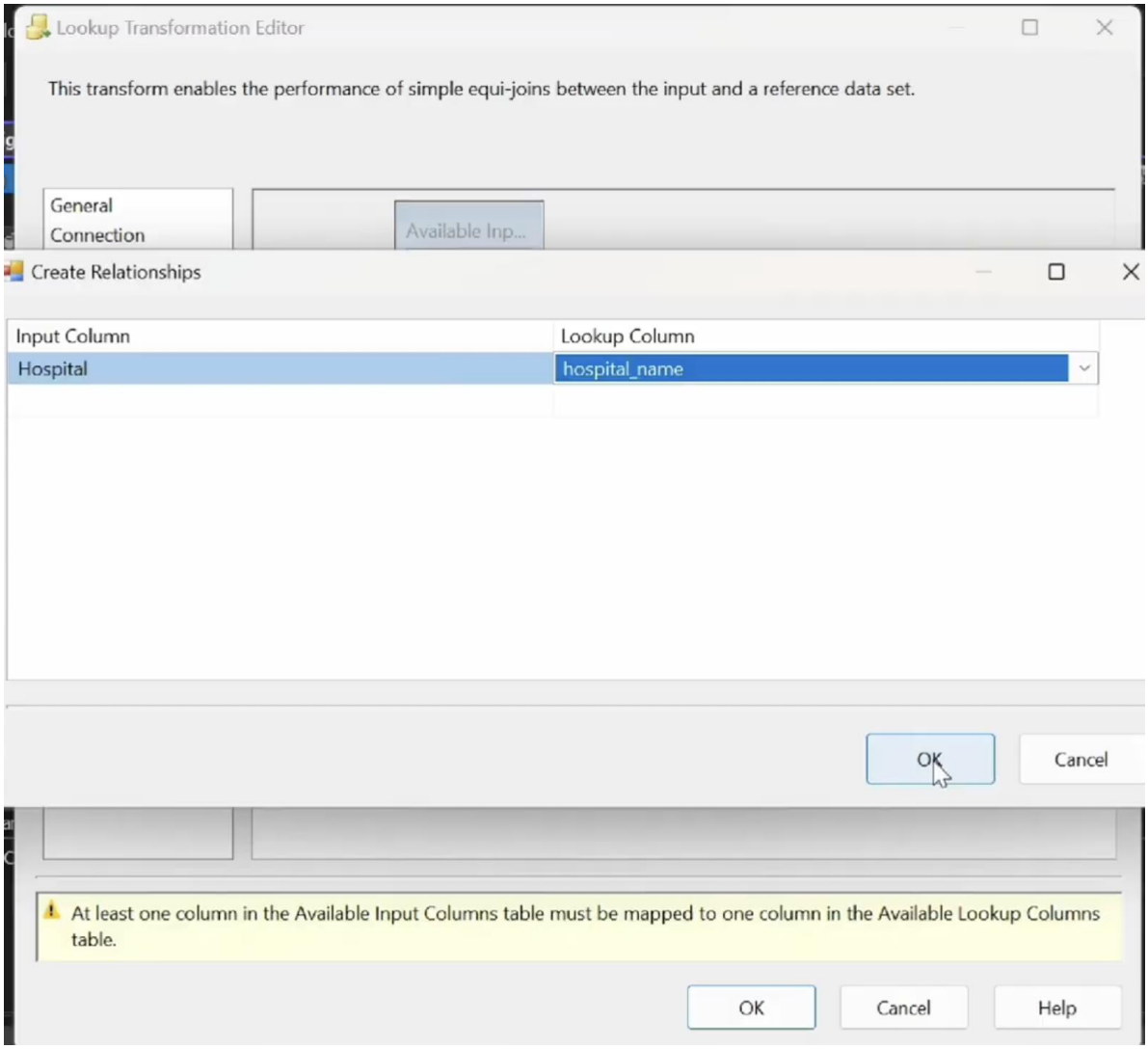

*Hình 61. Đổi tên ô Output Alias*

- Bước 11: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_LOCATION.



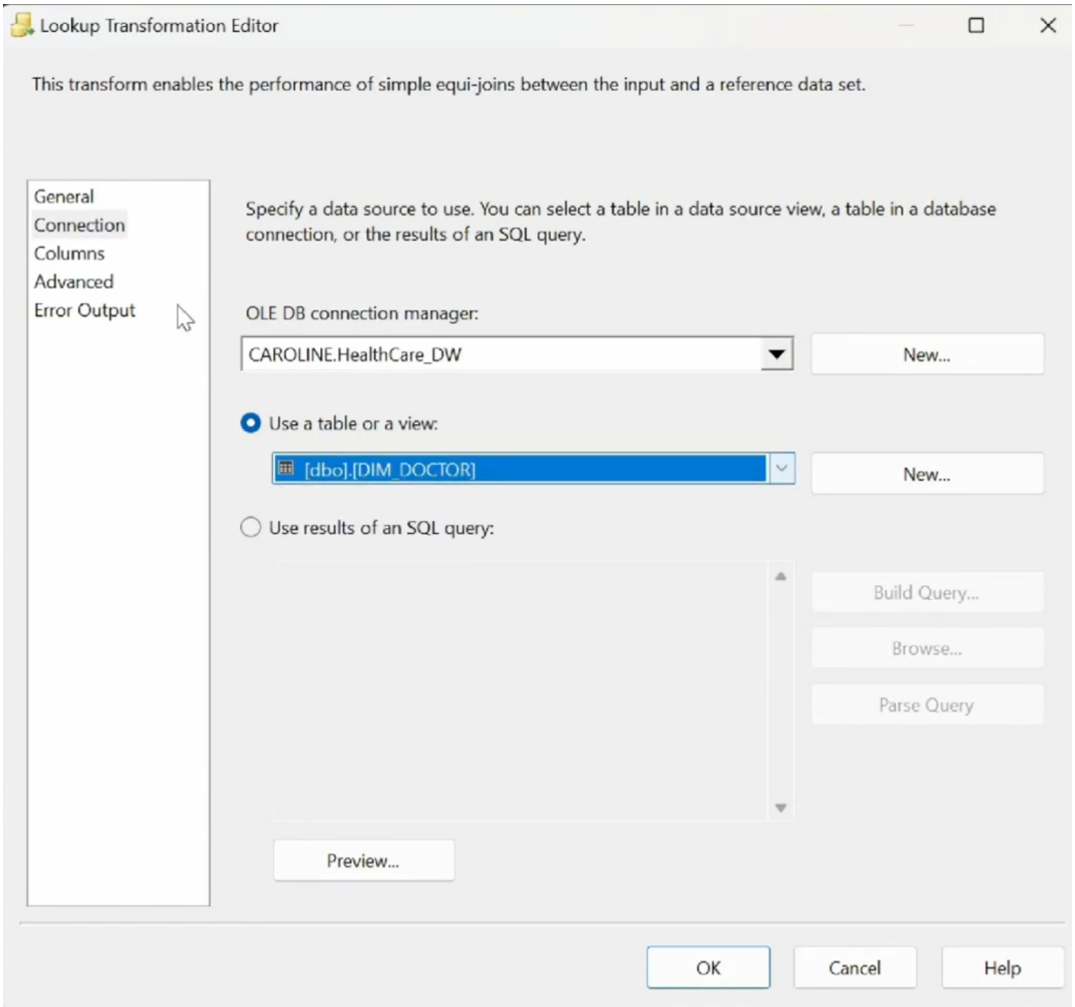

*Hình 62. Lookup*

- Bước 12: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của Location



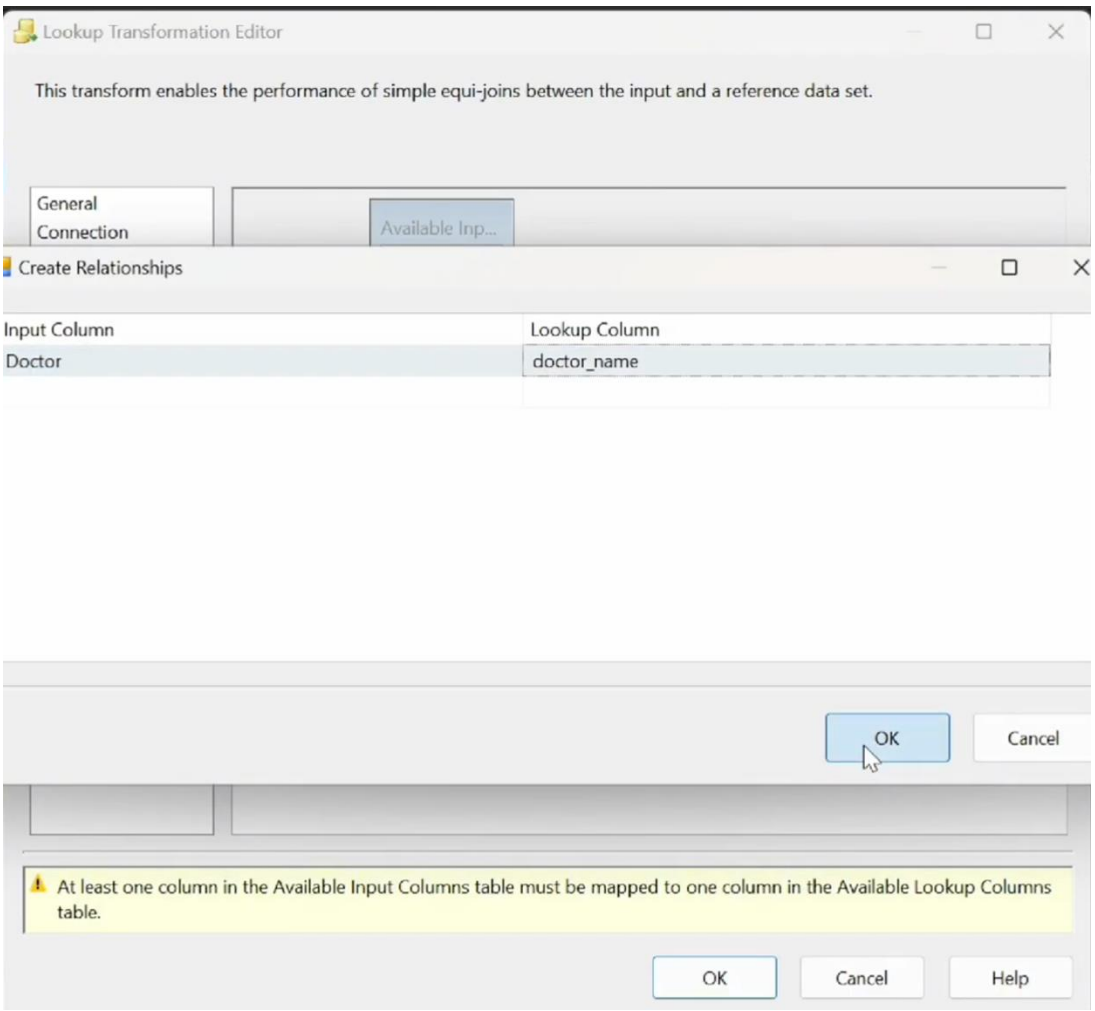

*Hình 63 . Edit Mapping*

- Bước 13: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_DOCTOR.



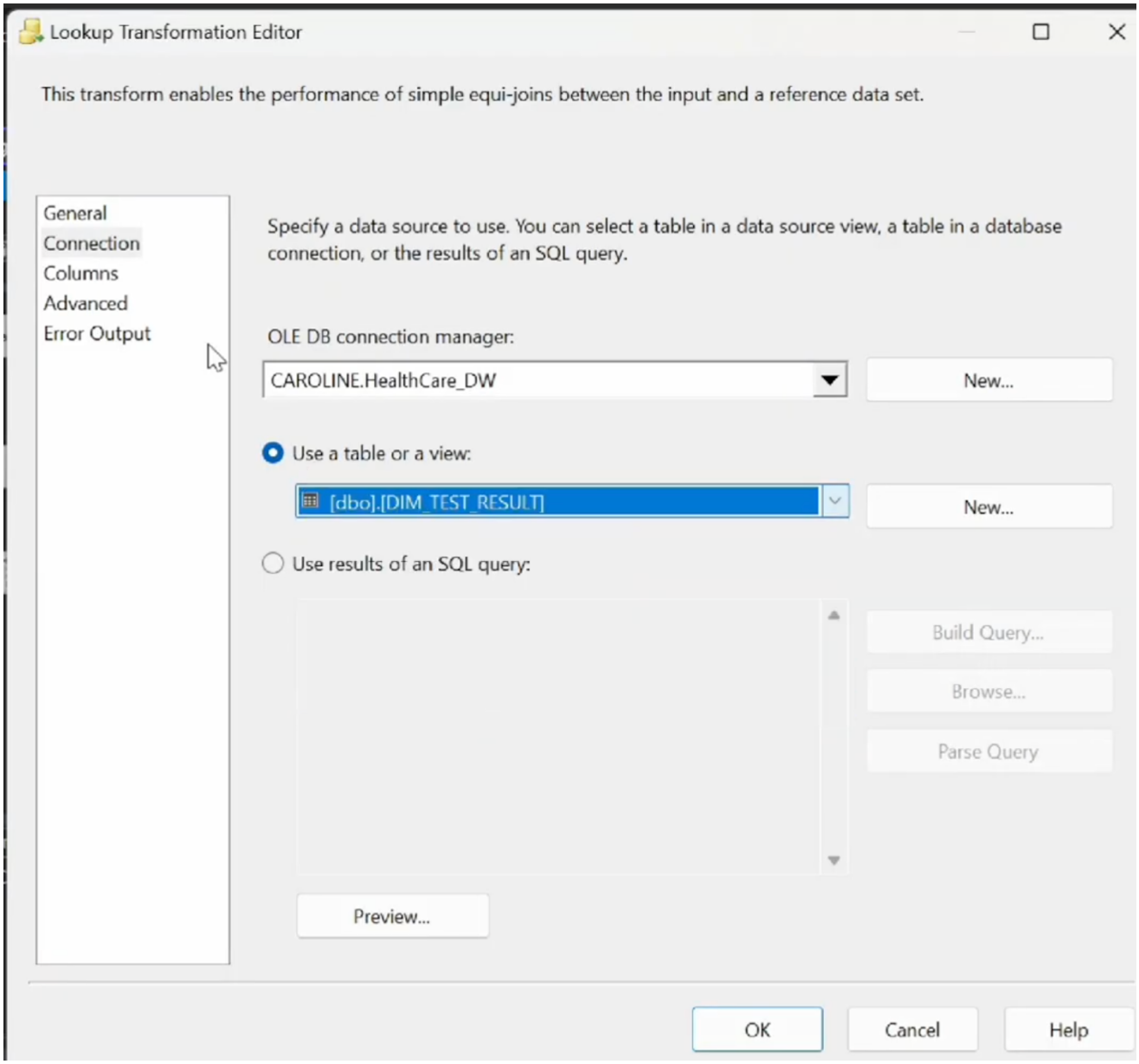

*Hình 64. Lookup DIM_DOCTOR*

- Bước 14: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của Doctor.



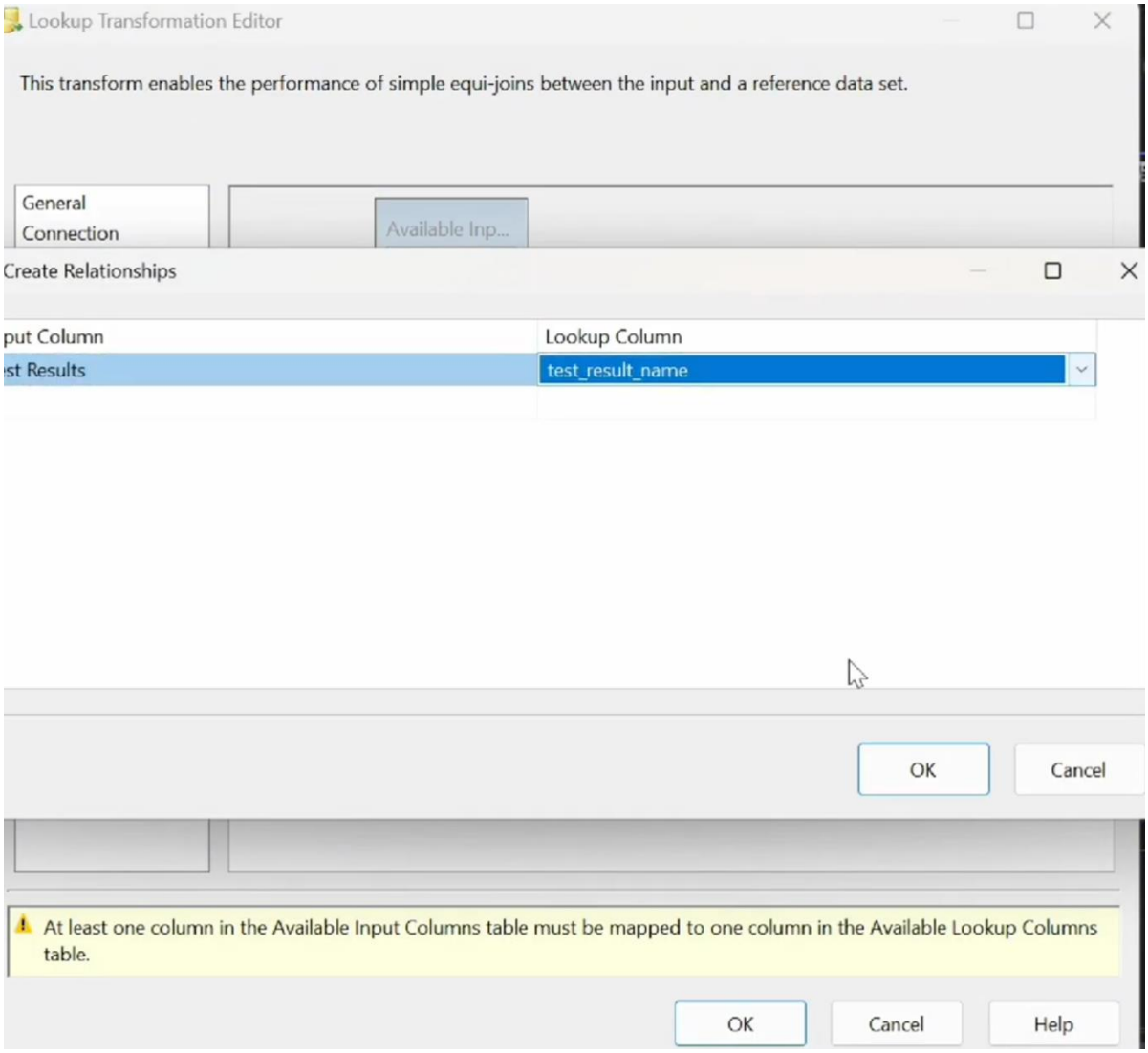

*Hình 65. Edit Mapping*

- Bước 15: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_TEST_RESULT.



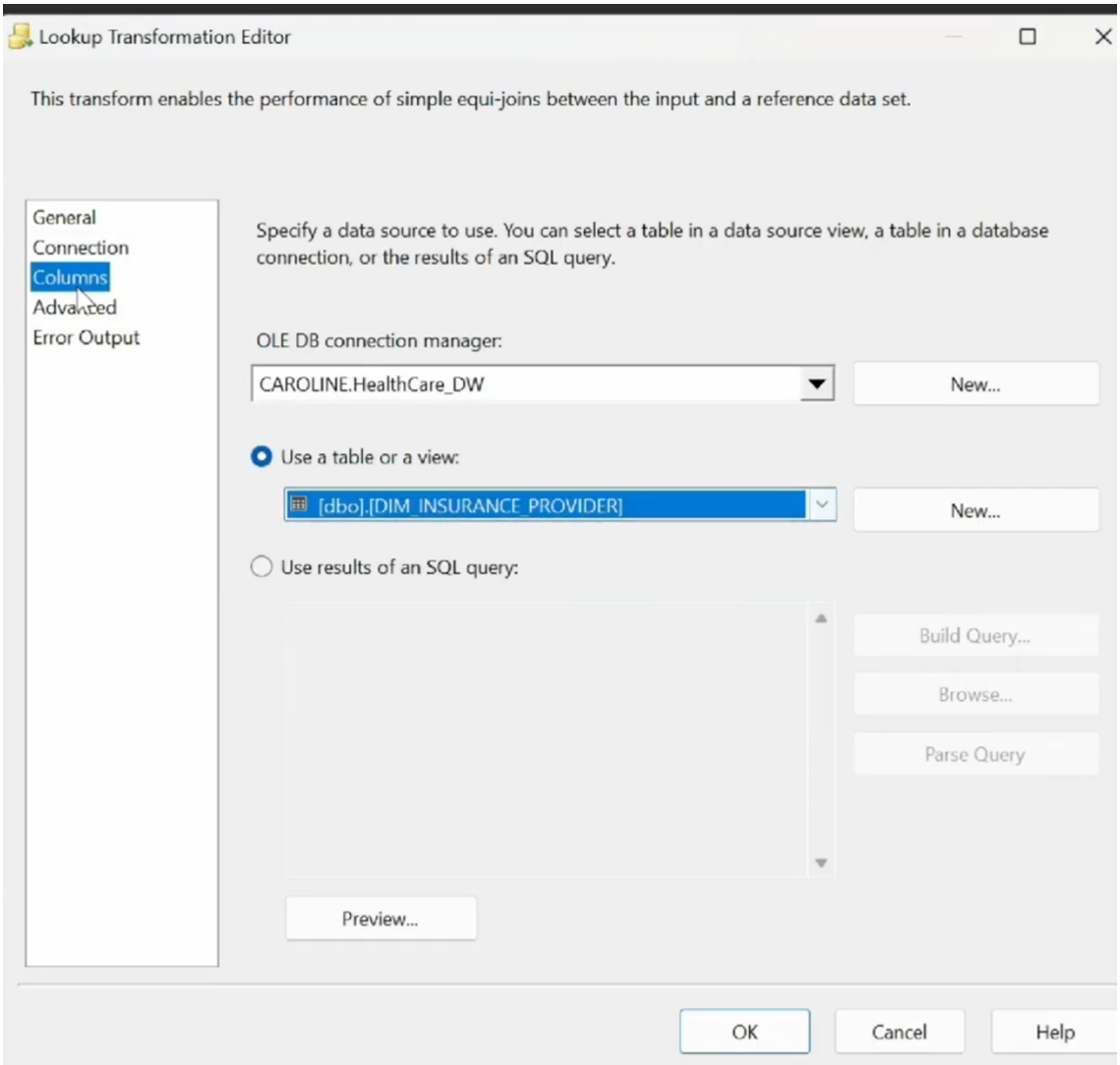

*Hình 66. Lookup DIM_TEST_RESULT*

- Bước 16: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của test result.



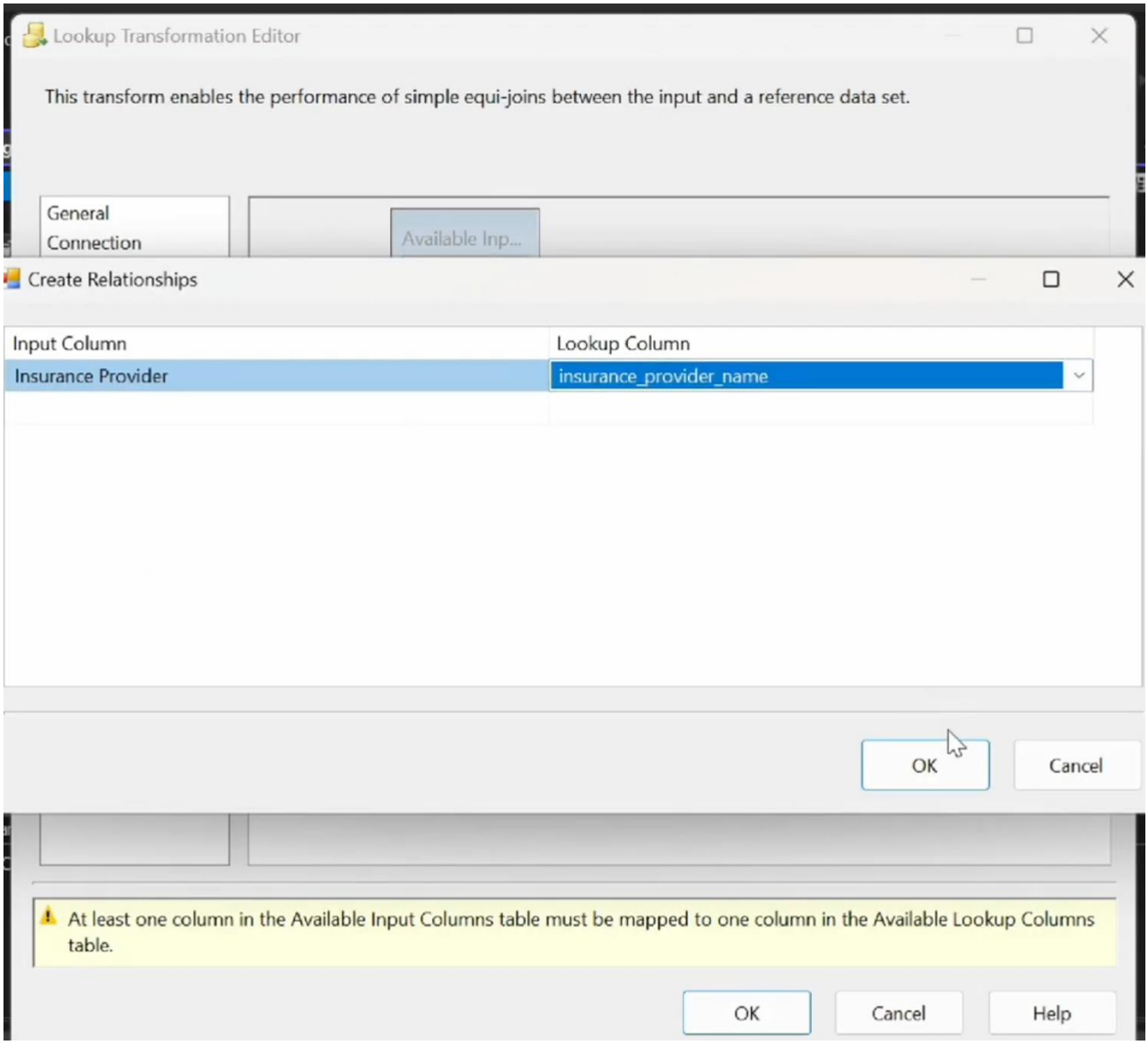

*Hình 67. Edit Mapping*

- Bước 17: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_INSURANCE_PROVIDER.



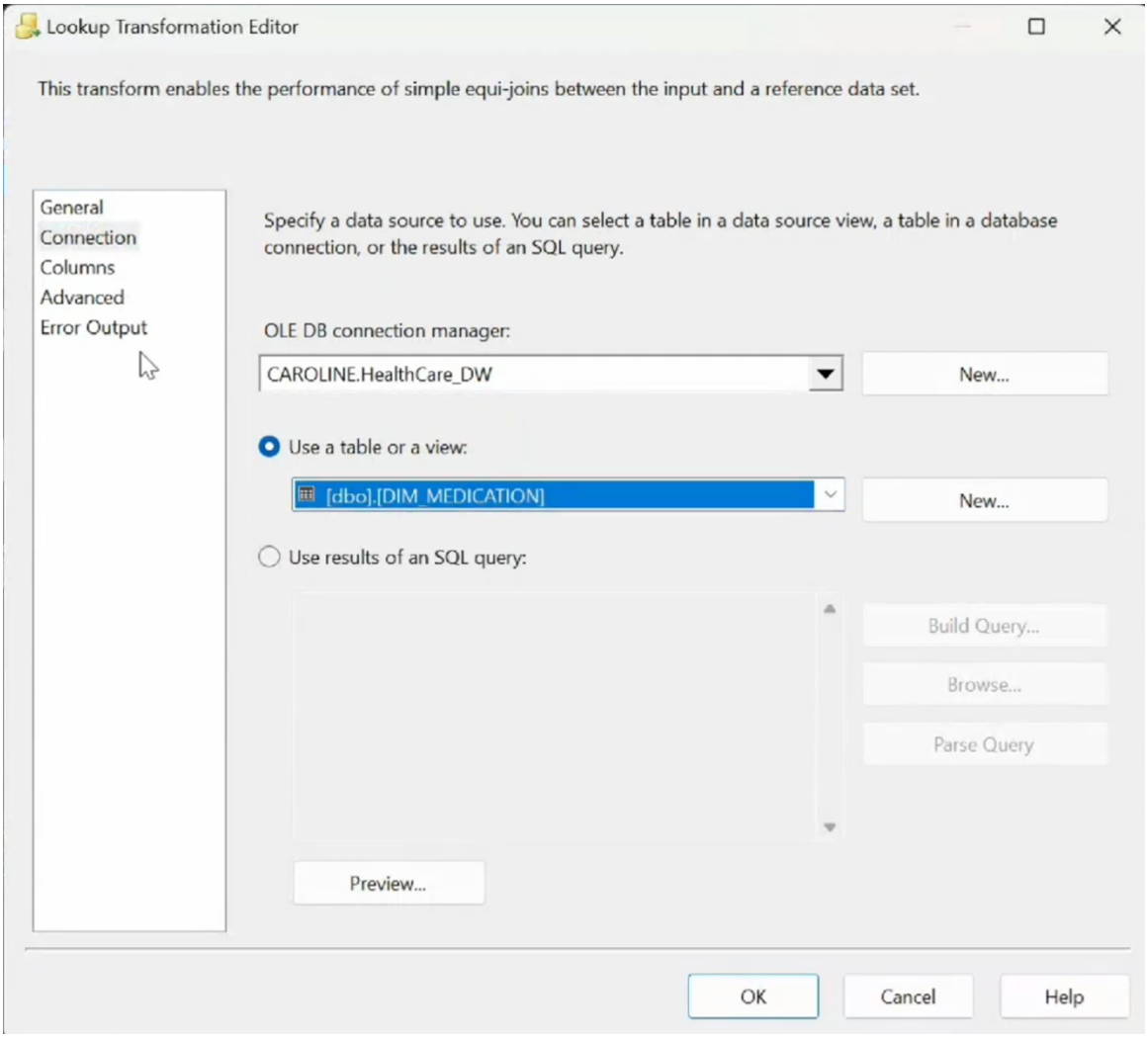

*Hình 68. Lookup DIM_INSURANCE_PROVIDER.*

- Bước 18: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của insurance provider.



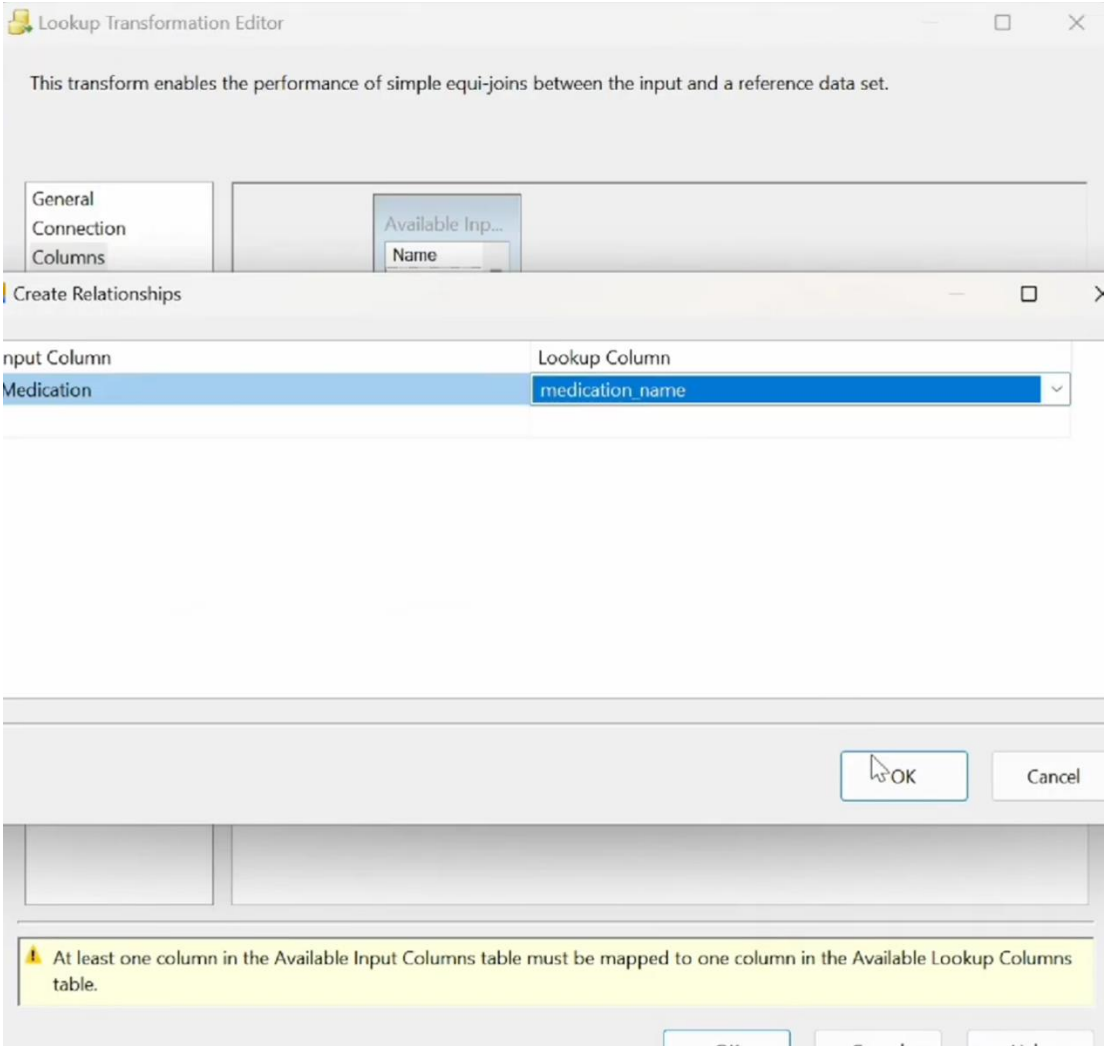

*Hình 69. Edit Mapping*

- Bước 19: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_MEDICATION.



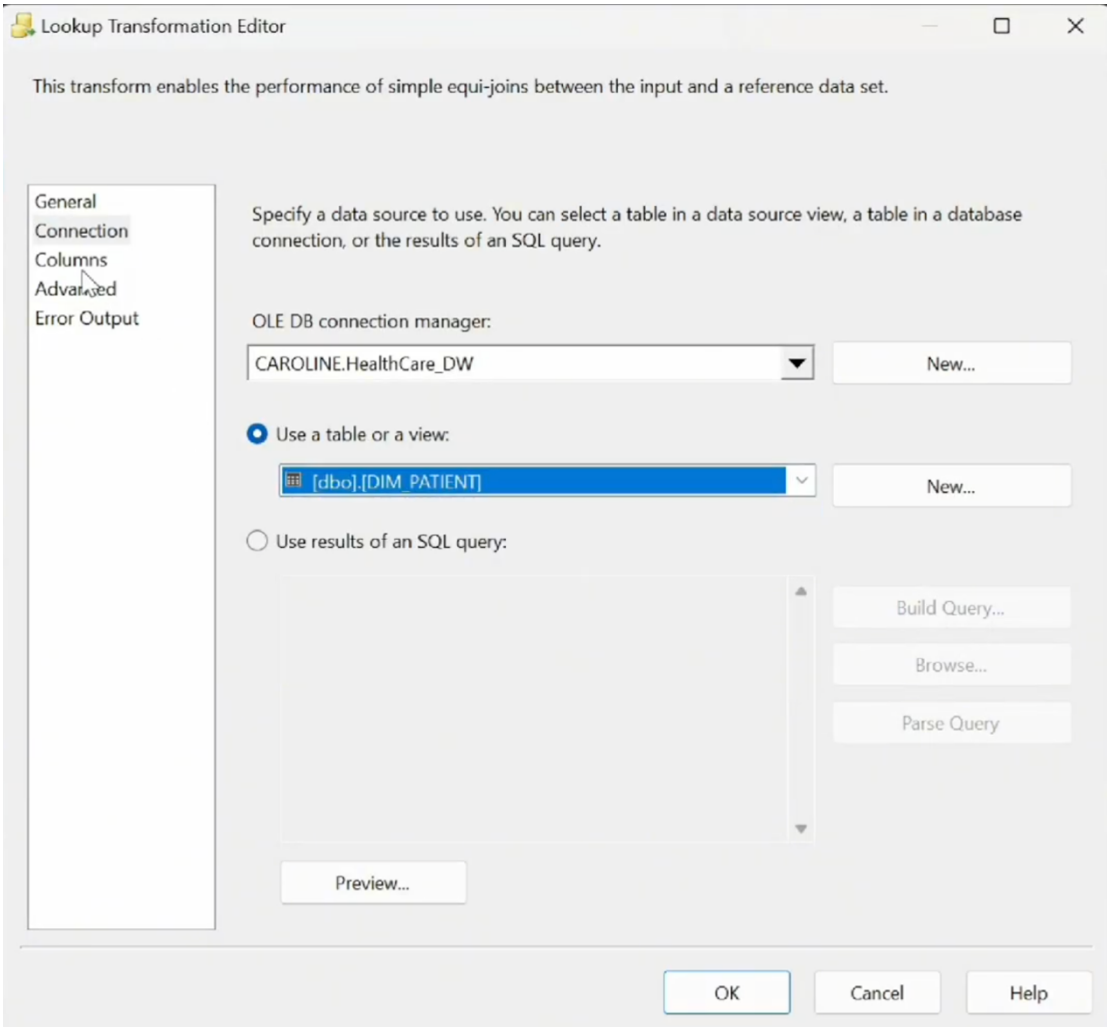

*Hình 70 . Lookup DIM_MEDICATION.*

- Bước 20: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của medication.



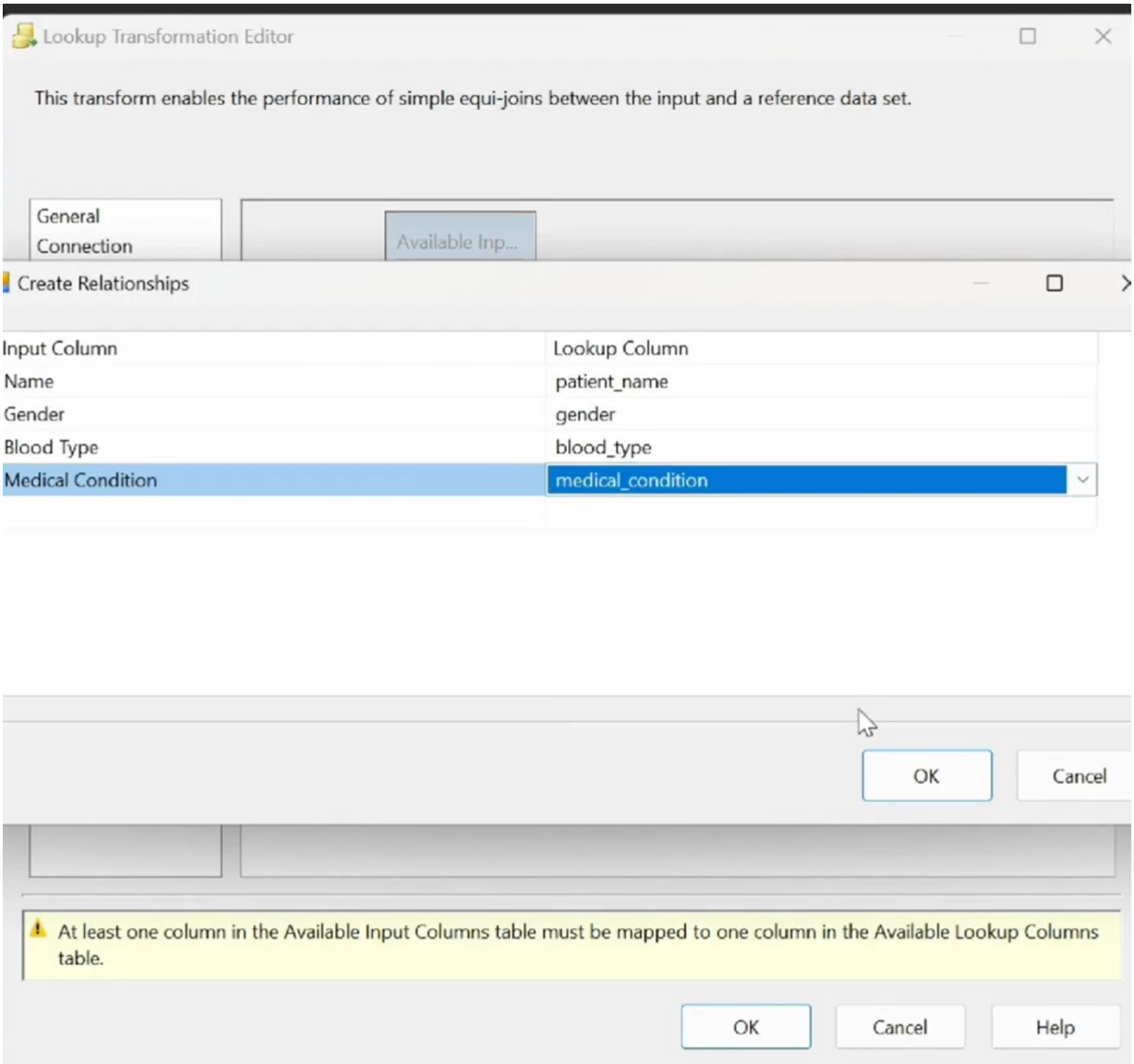

*Hình 71. Edit Mapping*

- Bước 21: Chọn công cụ **Lookup** từ SSIS ToolBox, chọn DIM_PATIENT.



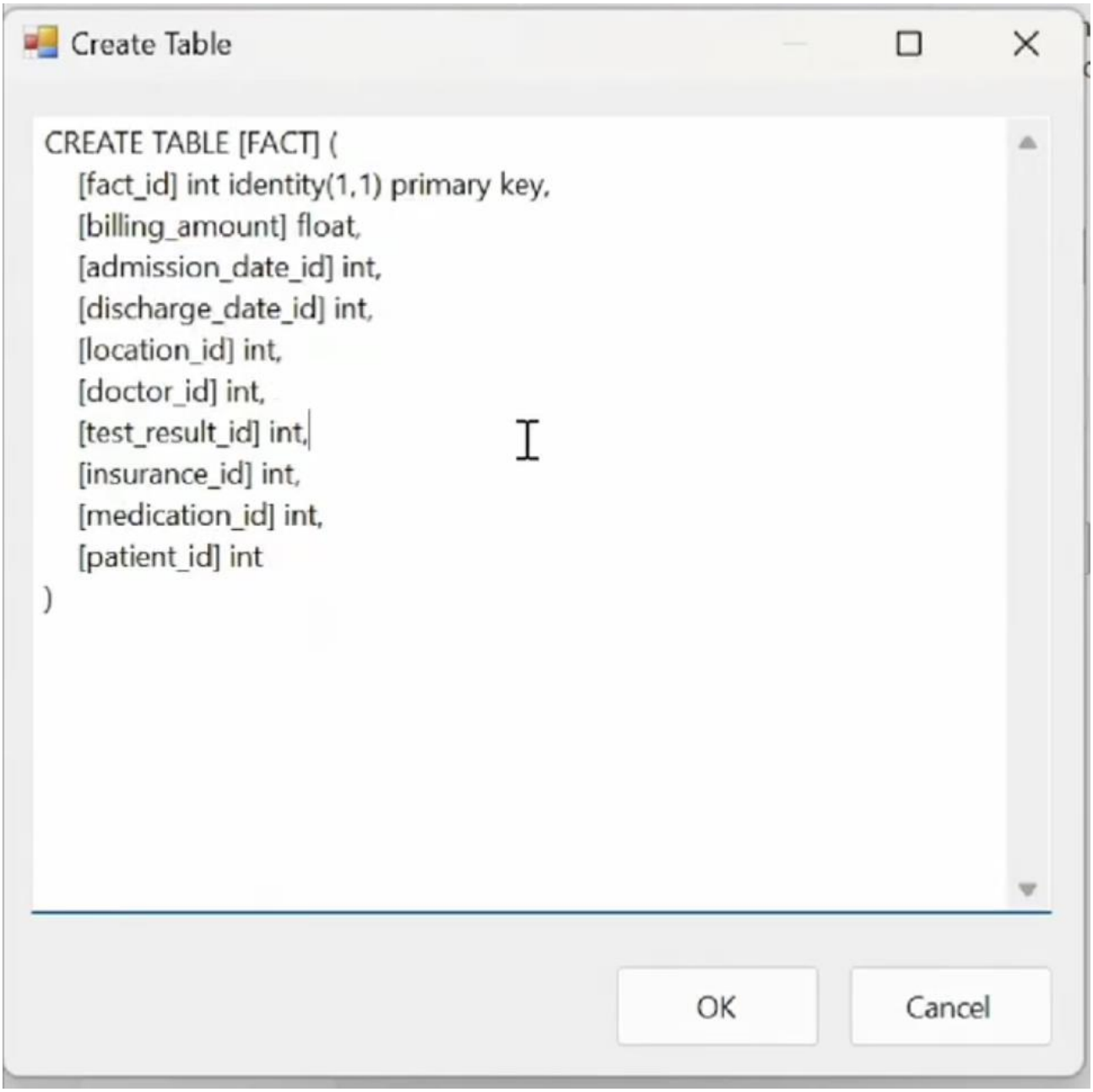

*Hình 72. Lookup DIM_PATIENT.*

- Bước 22: Chọn mục Column, chuột phải vào table 1 và chọn Edit Mapping, sau đó mapping các thuộc tính để lấy id của patient.




*Hình 73. Edit Mapping*

- Bước 23: Chọn công cụ OLE DB  Destination. Đổi tên thành FACT. Sau đó tạo bảng FACT.




*Hình 74. Tạo bảng FACT*

- Bước 25: Nhấn Start để bắt đầu quá trình đổ dữ liệu . Khi đỗ dữ liệu thành công ta sẽ có kết quả như hình. Vào SQL Server truy vấn xem dữ liệu đã được đổ vào chưa. Kết quả dữ liệu đã đổ như hình minh họa.

Results | Messages

| | fact_id | billing_amount | admission_date_id | discharge_date_id | location_id | doctor_id | test_result_id | insurance_id | medication_id | patient_id |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 18856.2813059782 | 1730 | 1732 | 34470 | 26613 | 3 | 2 | 4 | 5153 |
| 2 | 2 | 33643.3272865779 | 105 | 111 | 20830 | 33649 | 2 | 4 | 2 | 29834 |
| 3 | 3 | 27955.0960788425 | 1234 | 1249 | 10301 | 37829 | 3 | 1 | 1 | 11696 |
| 4 | 4 | 37909.7824098753 | 561 | 591 | 17148 | 22512 | 1 | 4 | 2 | 2702 |
| 5 | 5 | 14238.3178139376 | 1231 | 1251 | 38429 | 21260 | 1 | 1 | 5 | 397 |
| 6 | 6 | 48145.1109510419 | 1688 | 1692 | 27544 | 37122 | 3 | 5 | 2 | 15409 |
| 7 | 7 | 19580.8723448609 | 546 | 558 | 15572 | 21959 | 2 | 4 | 4 | 14742 |
| 8 | 8 | 45820.4627215946 | 966 | 976 | 29925 | 36660 | 2 | 3 | 4 | 8697 |
| 9 | 9 | 50119.2227915485 | 421 | 434 | 34813 | 9261 | 1 | 3 | 1 | 19801 |
| 10 | 10 | 19784.6310622107 | 747 | 777 | 28045 | 14680 | 2 | 5 | 4 | 9025 |
| 11 | 11 | 12576.7956090502 | 348 | 351 | 32723 | 18734 | 3 | 4 | 4 | 35210 |
| 12 | 12 | 7999.58687960419 | 1559 | 1582 | 23632 | 11559 | 2 | 4 | 3 | 83 |
| 13 | 13 | 43282.2833577043 | 219 | 235 | 29978 | 22138 | 1 | 3 | 5 | 10103 |
| 14 | 14 | 33207.7066337296 | 381 | 409 | 31108 | 37437 | 3 | 5 | 3 | 40319 |
| 15 | 15 | 40701.5992273088 | 885 | 890 | 26523 | 32894 | 3 | 5 | 4 | 6675 |
| 16 | 16 | 12263.3574250214 | 1335 | 1345 | 9842 | 25489 | 2 | 2 | 1 | 36606 |
| 17 | 17 | 24499.8479037366 | 413 | 434 | 9266 | 13098 | 3 | 5 | 4 | 17425 |
| 18 | 18 | 17440.4854441247 | 306 | 331 | 26321 | 20615 | 1 | 3 | 4 | 36410 |

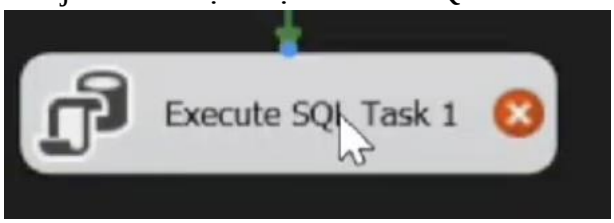
Query executed successfully. | CAROLINE (16.0 RTM) | CAROLINE\Thanh Thuy xi... | HealthCare_DW | 00:00:00 | 55,500 rows

*Hình 75 . Đổ dữ liệu*

# CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO( SSAS)

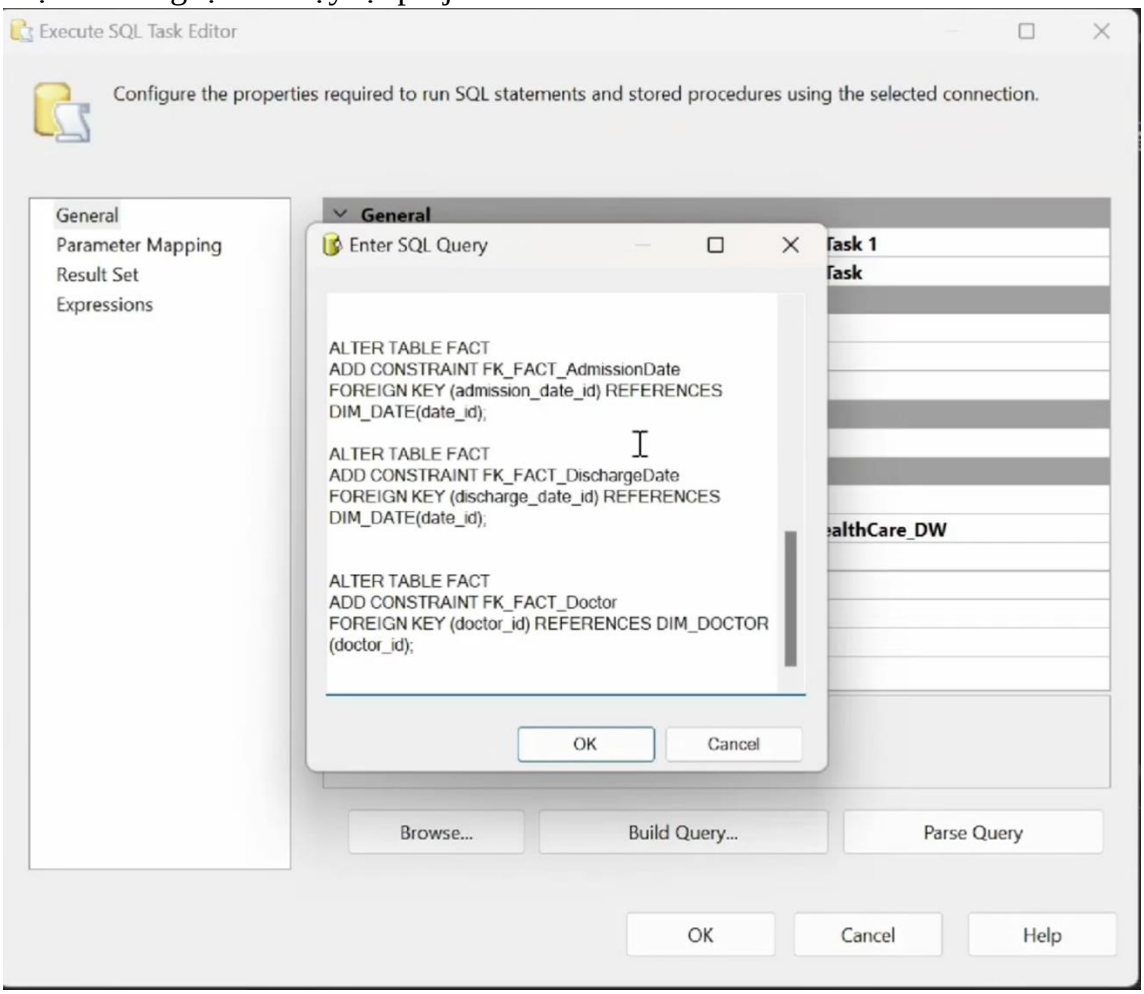
## 1. TẠO PROJECT SSAS TRONG VISUAL STUDIO

- Trước tiên tạo Mở Project SSIS tạo một Excute SQL Task mới:



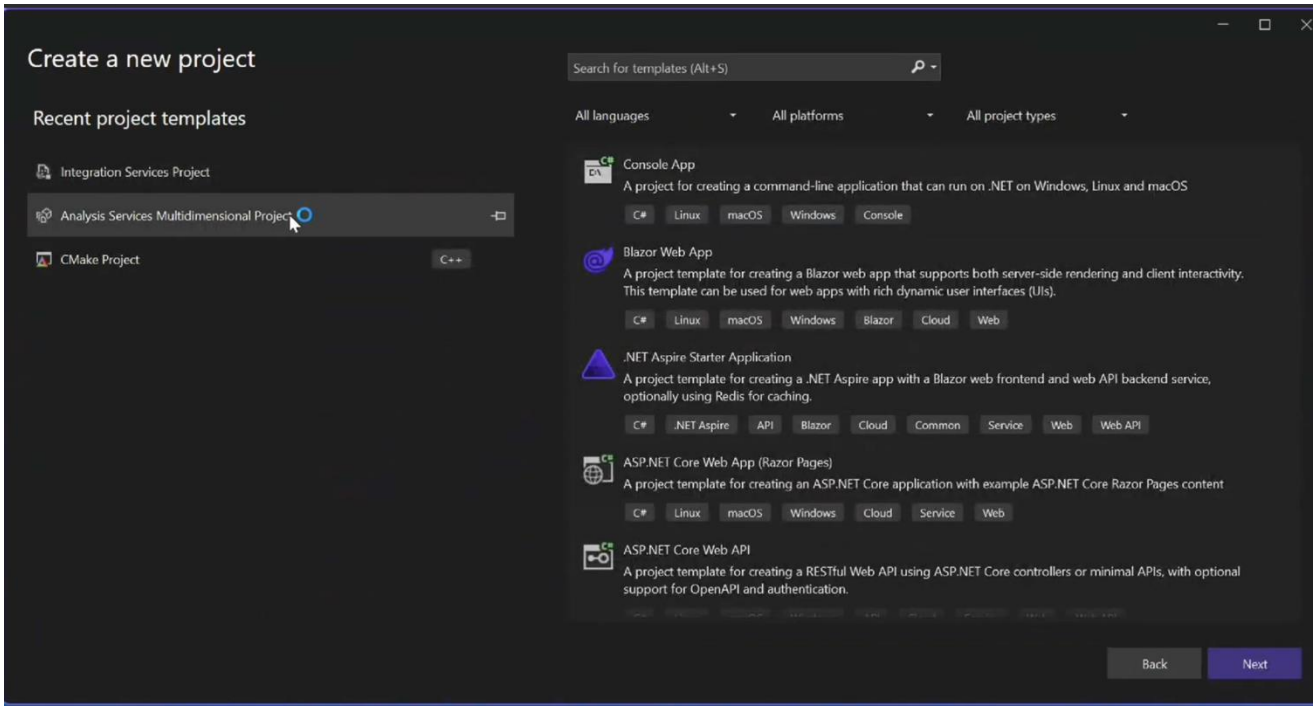

*Hình 76. Tạo Excute SQL Task*

- Tạo khoá ngoại và chạy lại project SSIS



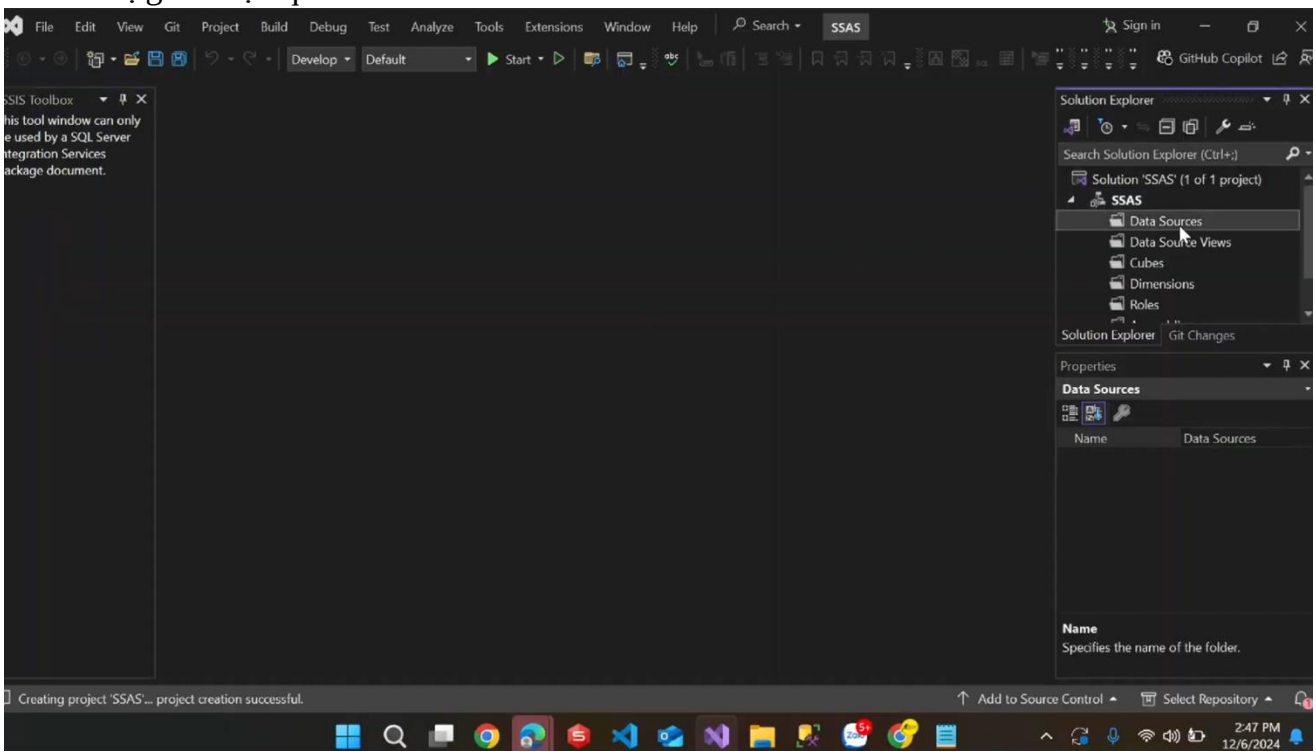

*Hình 77 . Tạo Khóa ngoại*

- **Bước 1**: Tạo project với template Analysis Services Multidimensional Project




*Hình 78 . Tạo Project SSAS*

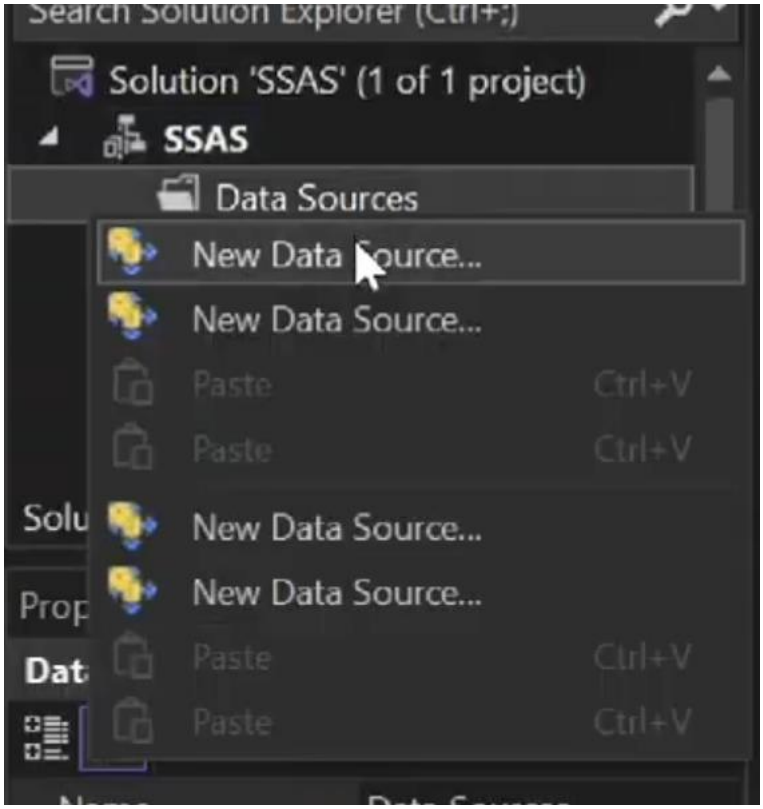
- **Bước 2**: Sau đó chọn đường dẫn để lưu project và đặt tên. Nhấn OK, màn hình sẽ hiển thị giao diện quá trình SSAS.




*Hình 79 . Lưu project*

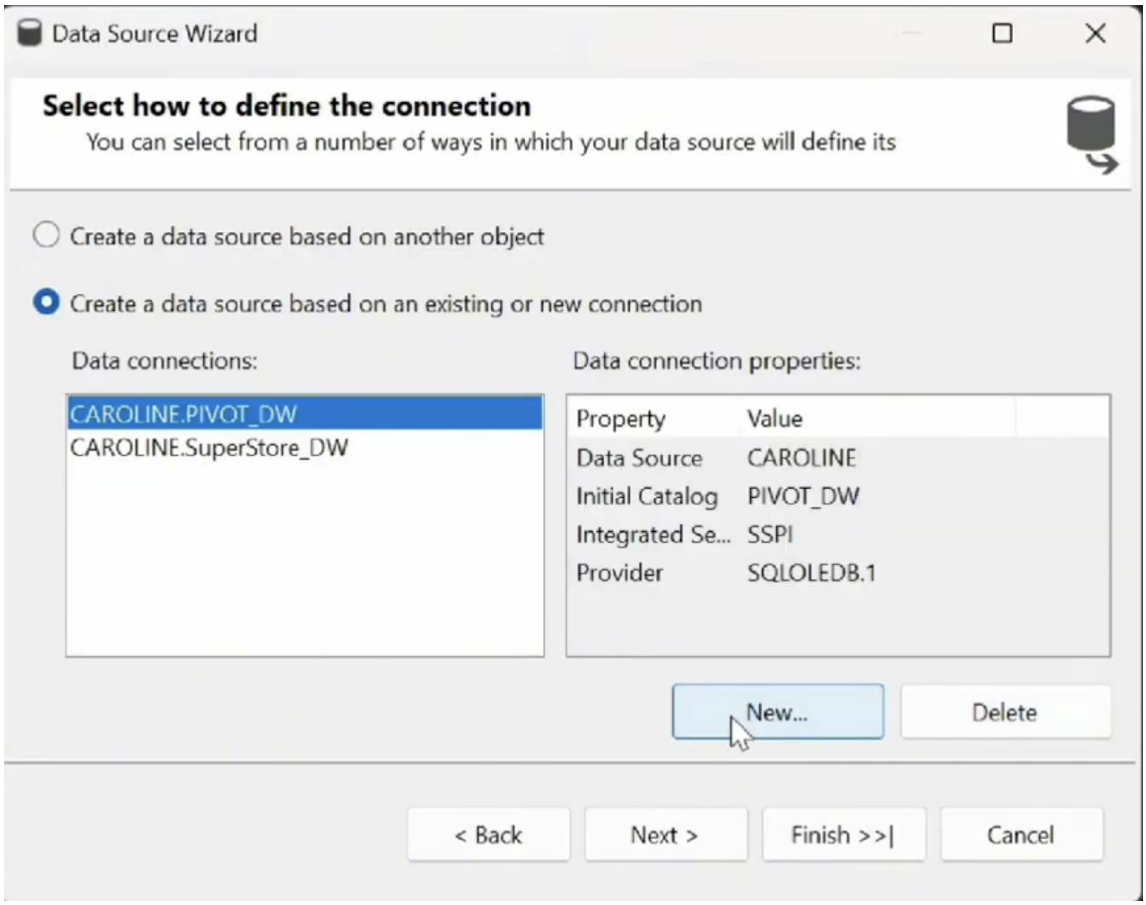
## 2. XÁC ĐỊNH DỮ LIỆU NGUỒN( DEFINE DATE SOURCE)

- **Bước 1:** Bên gốc phải phần Solution Explorer nhấn chuột phải vào Data Source.



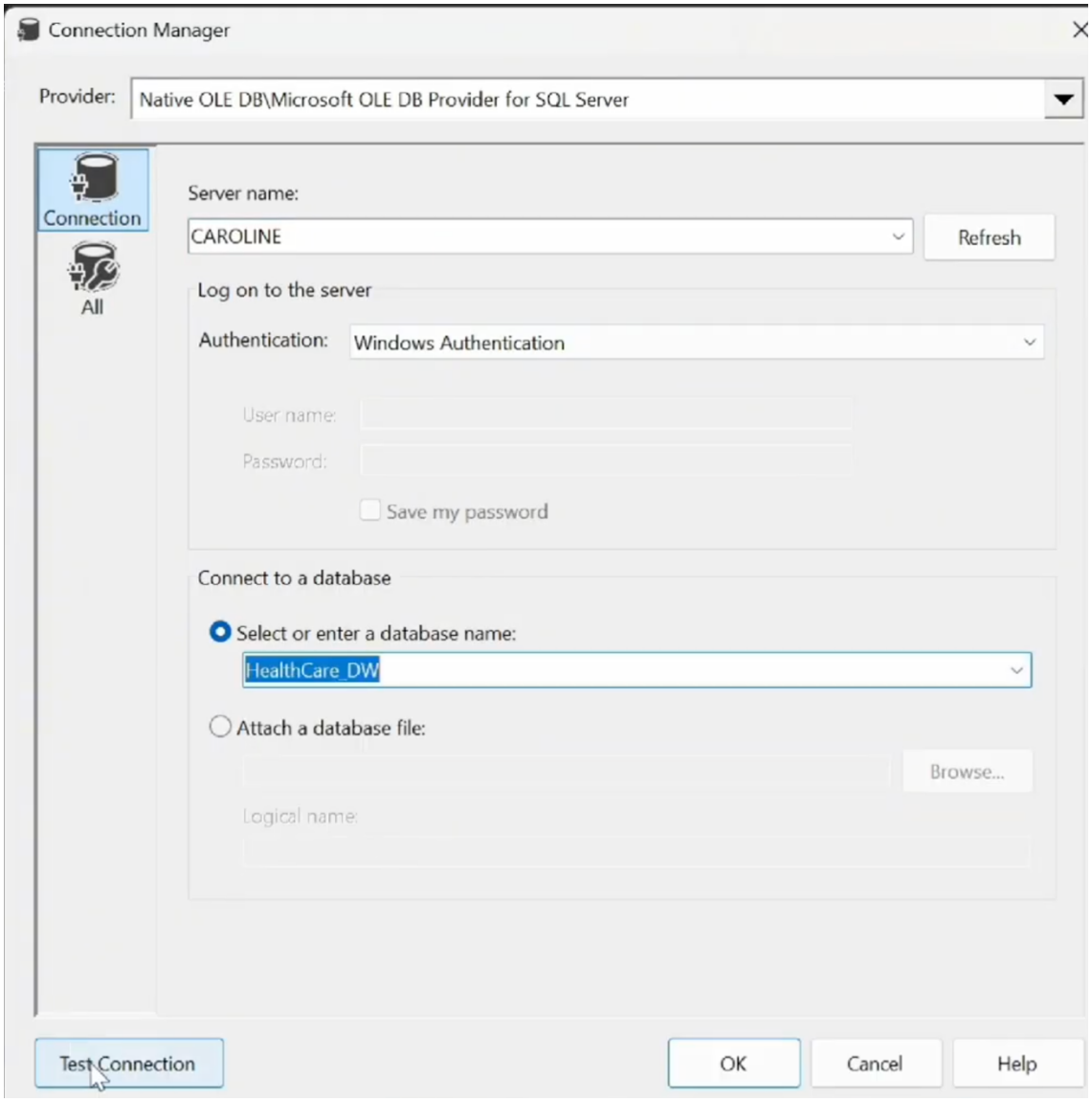

*Hình 80. New Data Source*

- **Bước 2:** Chọn New.



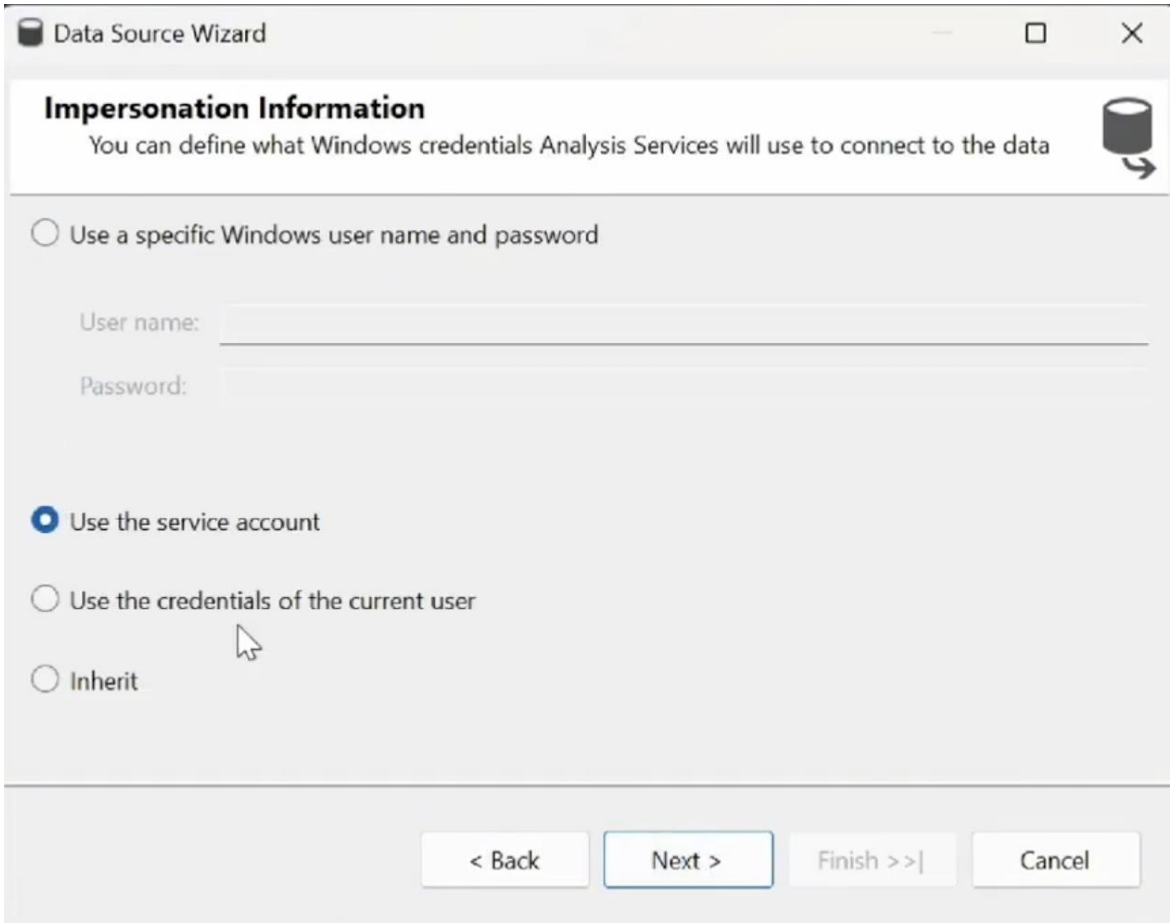

*Hình 81. New Data Source*

- **Bước 3:** Chọn server name và kết nối database.



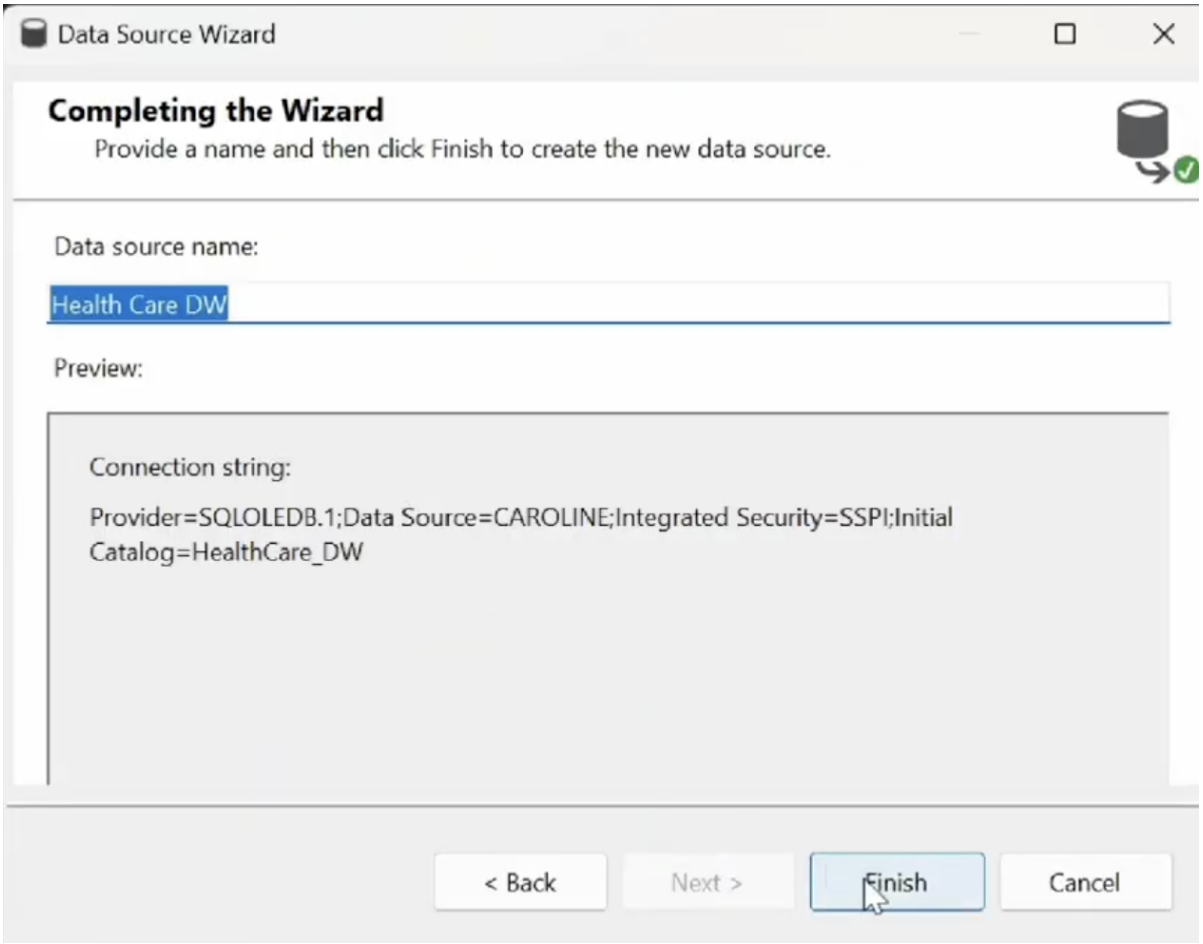

*Hình 82. Kết nối database*

- **Bước 4:** Bấm Next, chọn Use the service account.




*Hình 83 . Use the service account.*

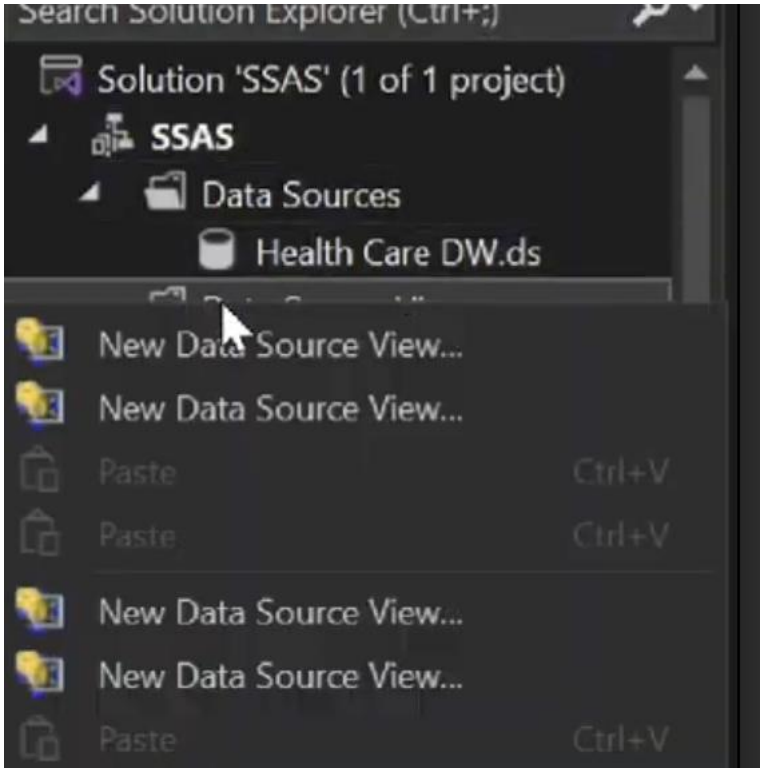
- **Bước 5:** Bấm Next, chọn finish.




*Hình 84 . Finish*

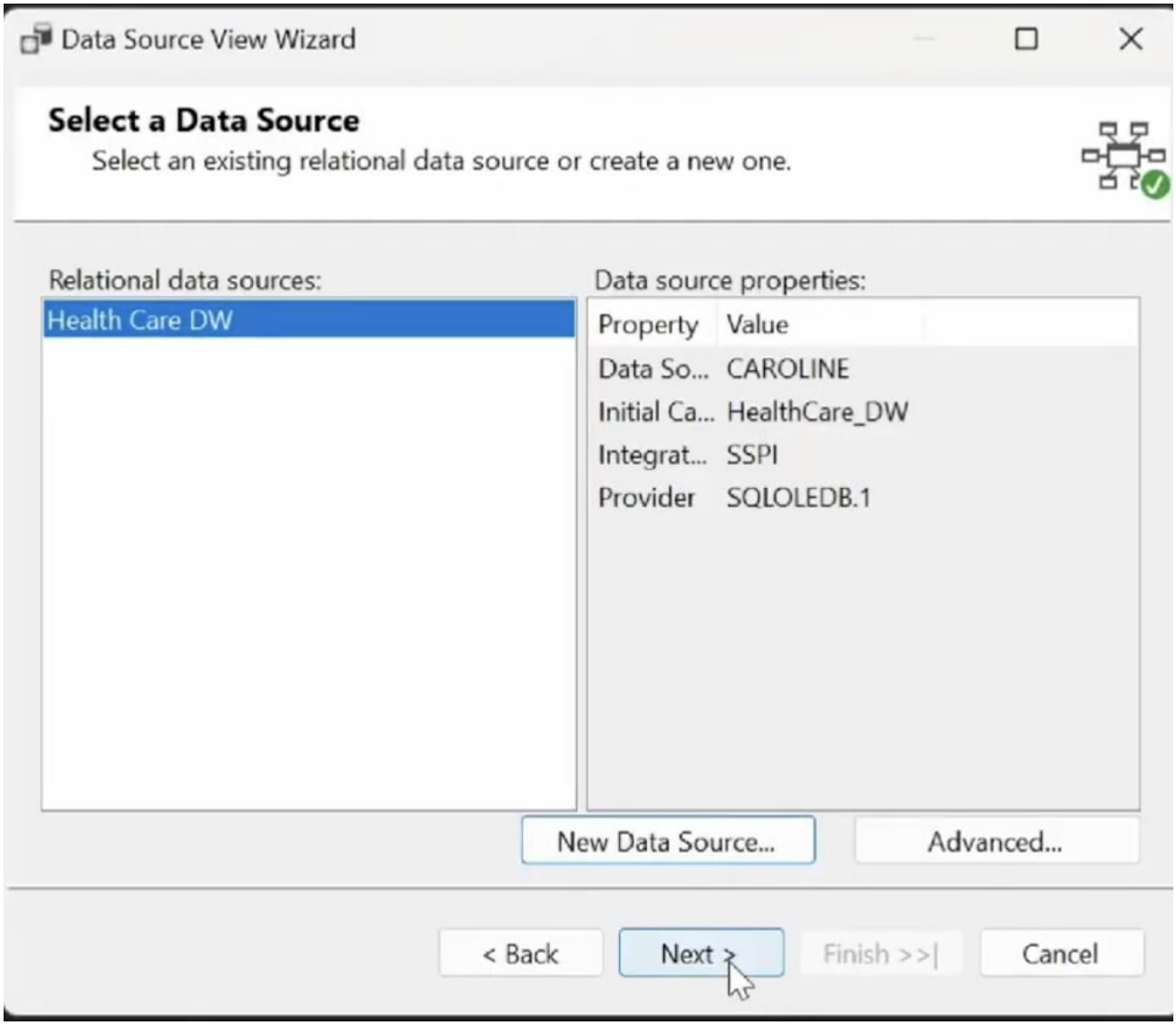
## 3. XÁC ĐỊNH KHUNG NHÌN DỮ LIỆU NGUỒN( DEFINE DATE SOURCE VIEW)

- **Bước 1:** Bên góc phải phần Solution Explorer nhấn chuột phải vào Data Source View



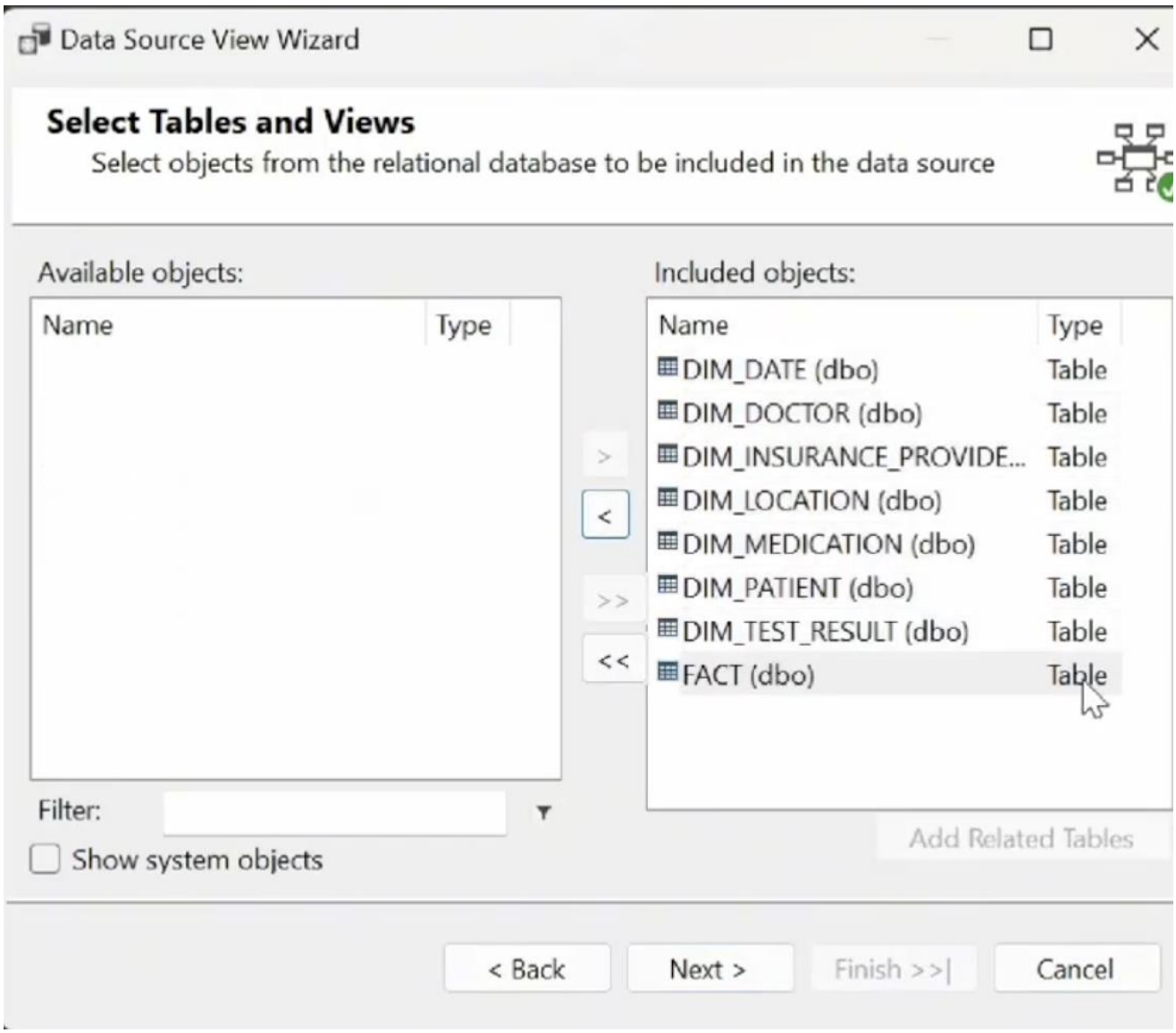

*Hình 85. New Data Source View*

- **Bước 2:** Chọn dataset của project và bấm next.



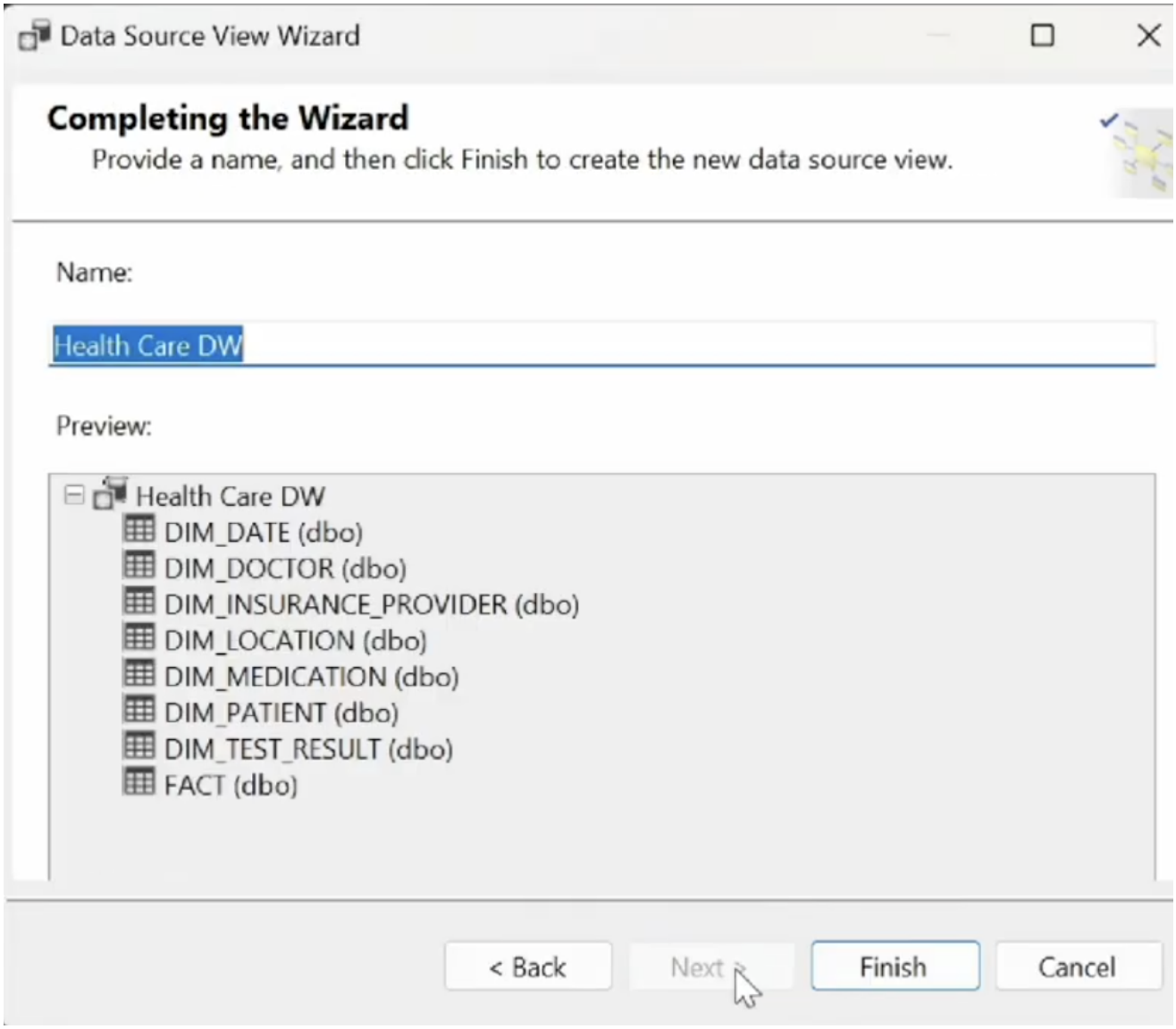

*Hình 86. Chọn DATASET*

- **Bước 3:** Chọn bảng FACT và DIM cho quá trình phân tích và bấm Next




*Hình 87 . Chọn các bảng FACT và DIM*

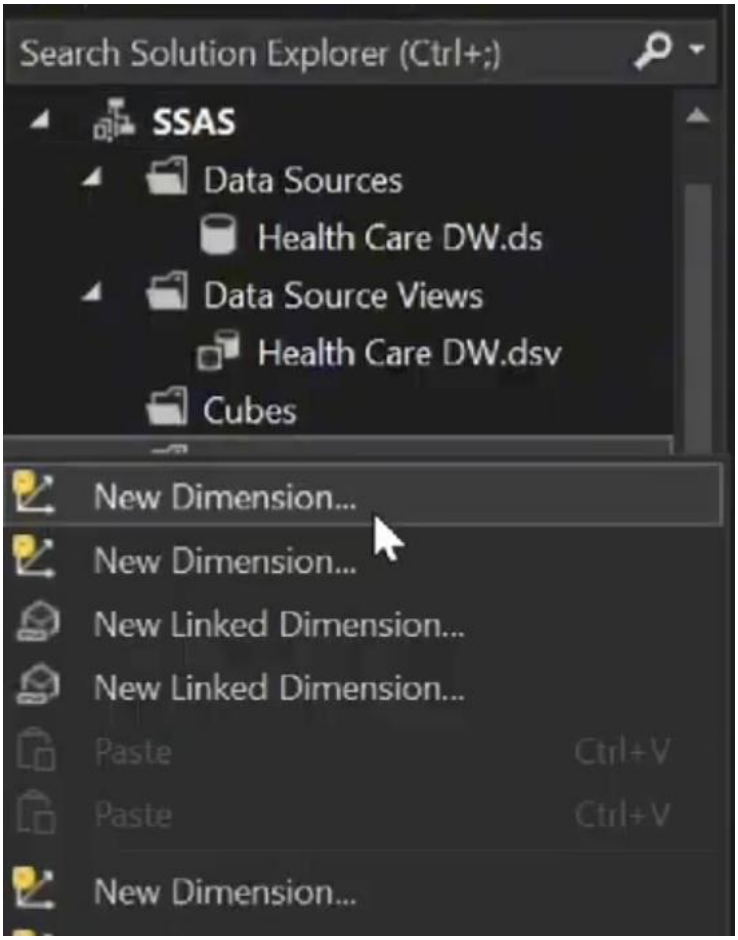
- **Bước 4:** Nhấn Finish để kết thúc quá trình.




*Hình 88. FINISH*

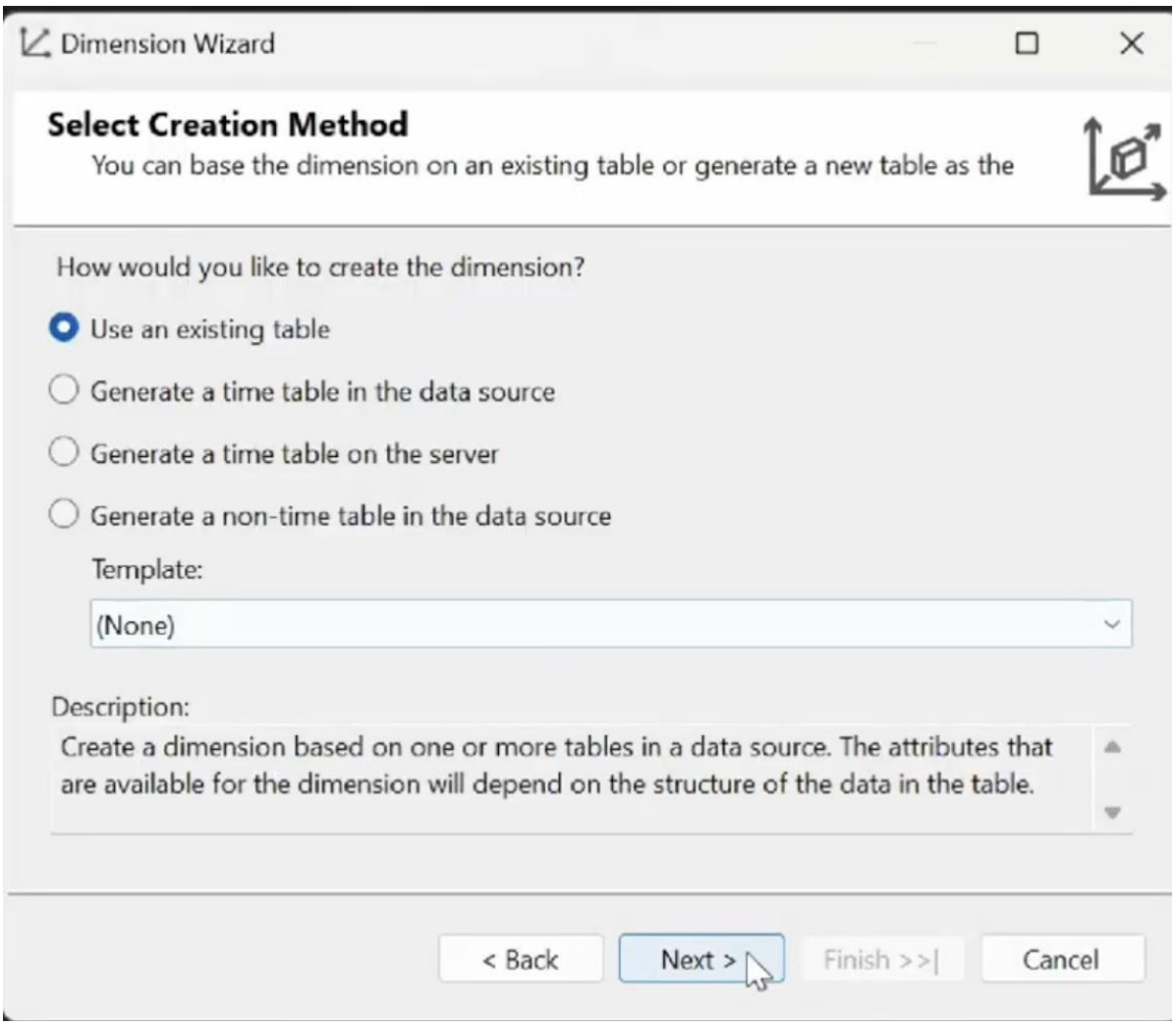
## 4. XÁC ĐỊNH CÁC CHIỀU ( DEFINE A DIMENSION)

- **Bước 1:** Bên gốc phải phần Solution Explorer nhấn chuột phải vào Dimension và chọn New Dimension.



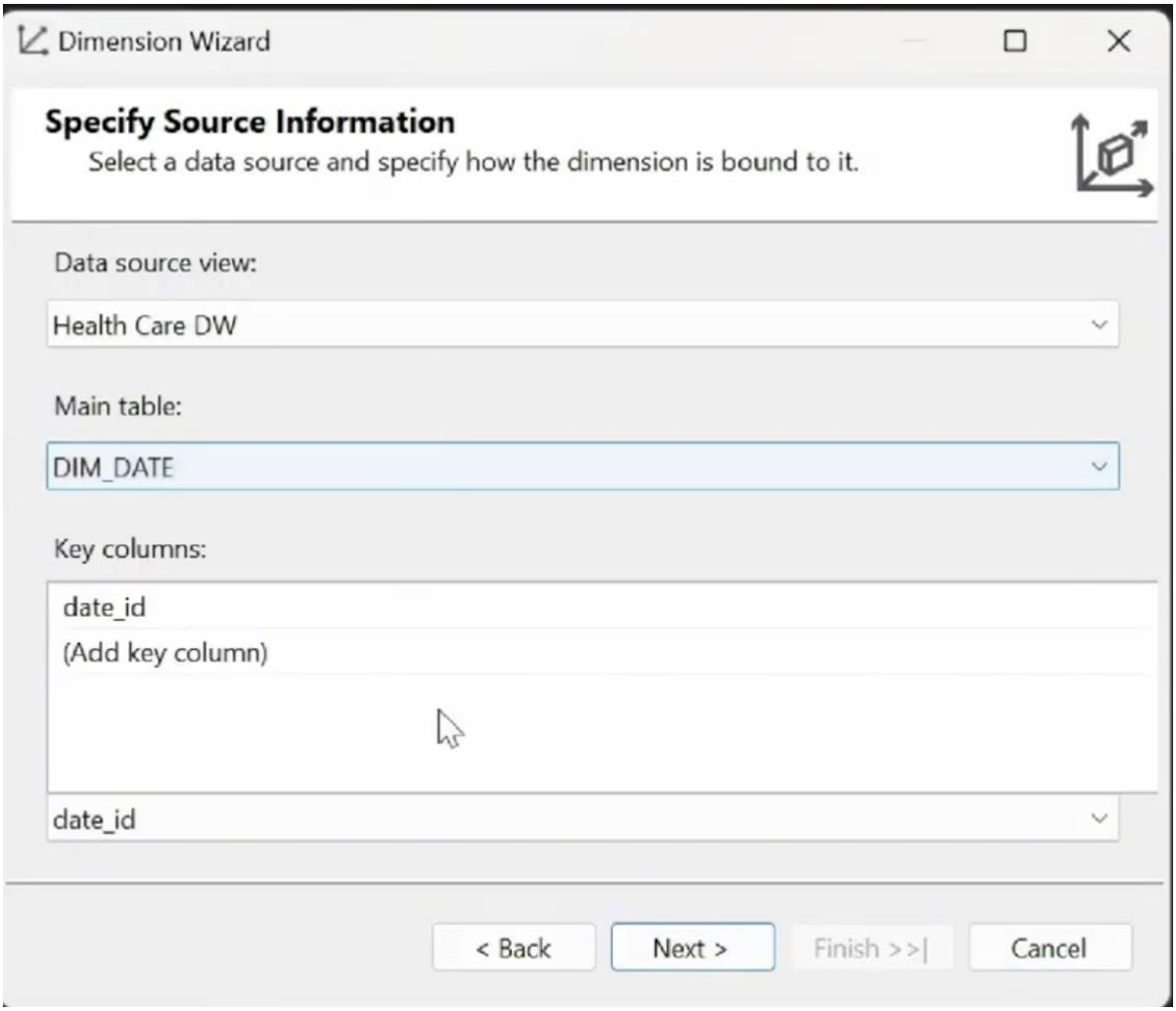

*Hình 89. New Dimension*

- **Bước 2:** Nhấn Next.

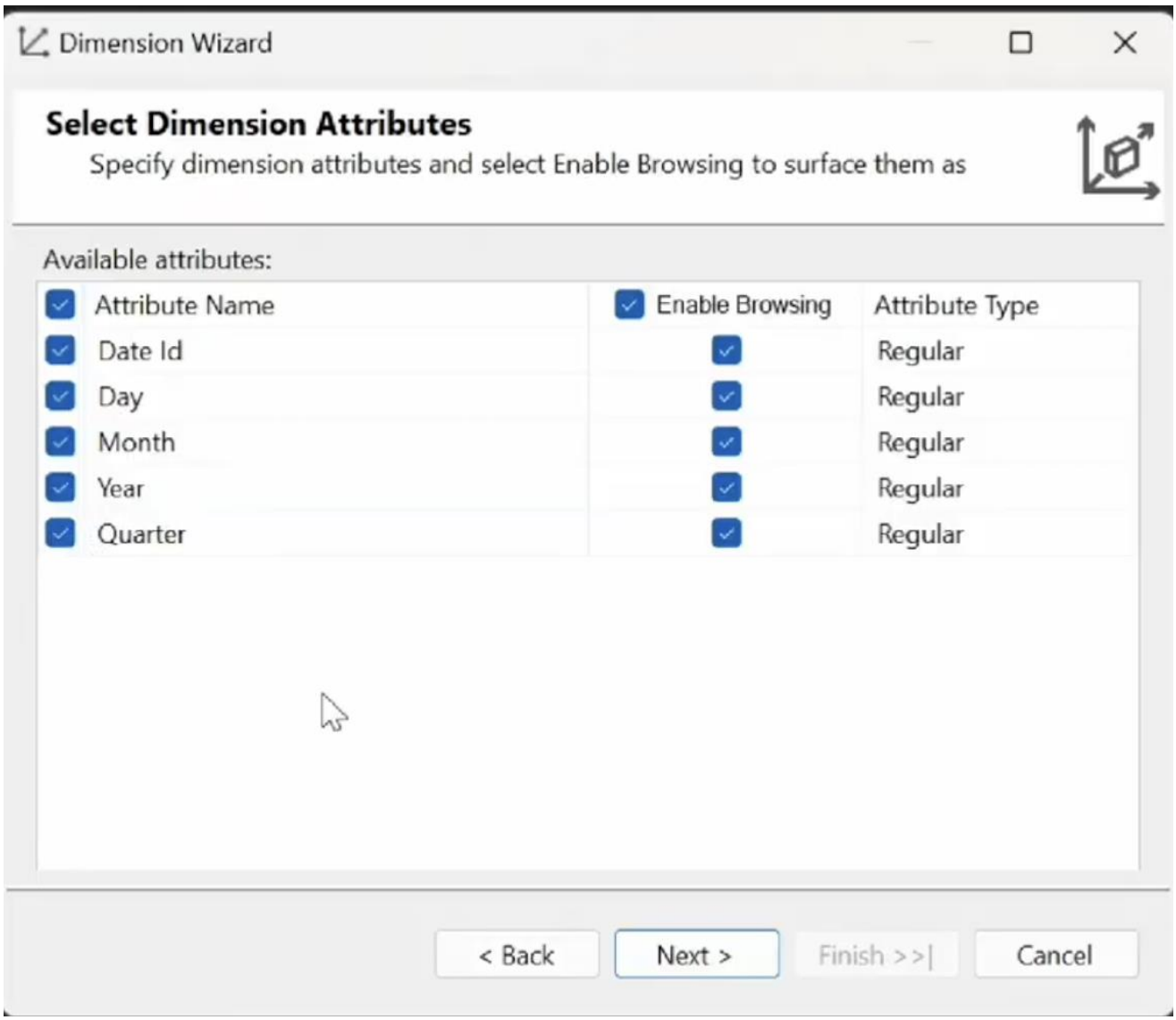



*Hình 90 . Tạo Dimesion*

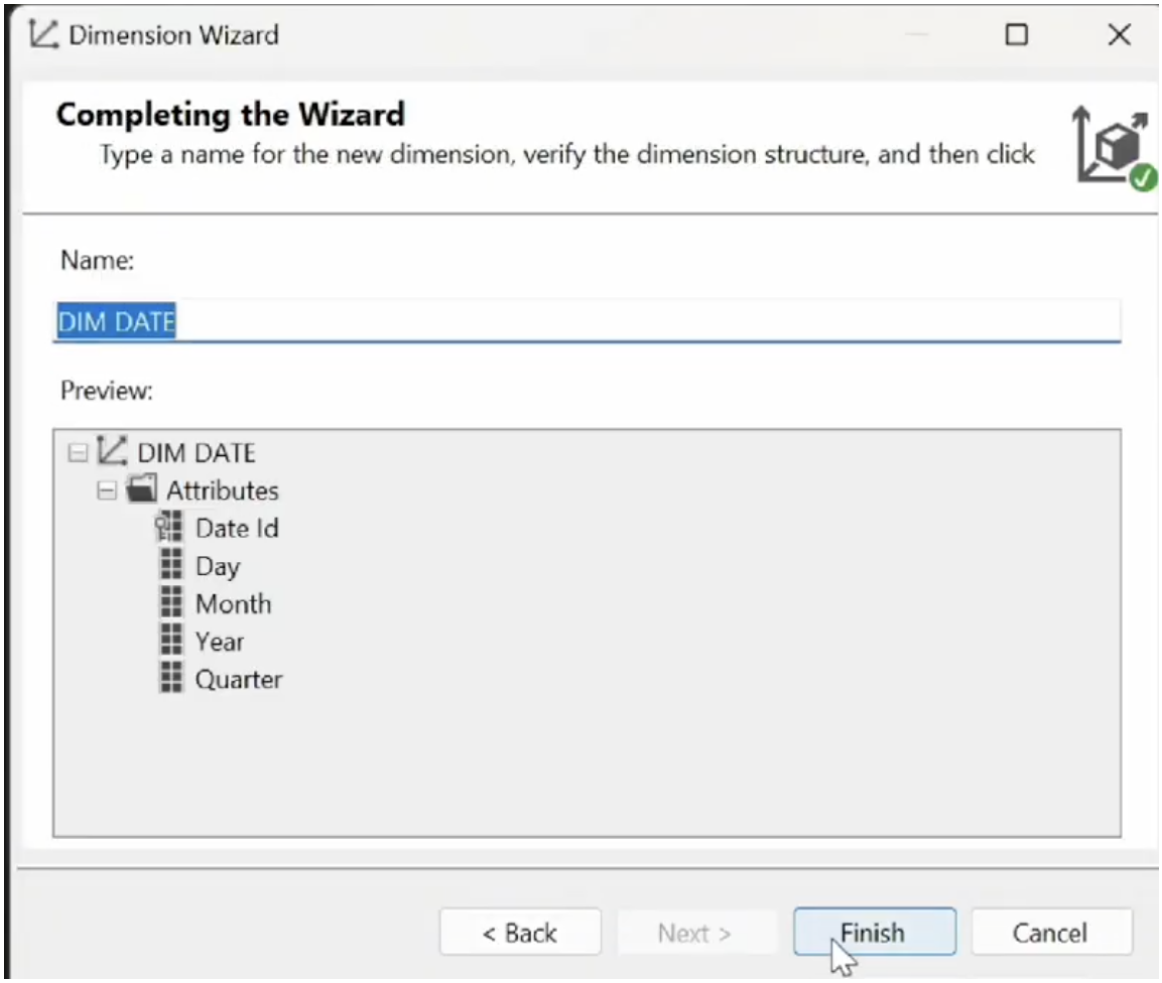
- **DIM_DATE**




*Hình 91. DIM_DATE*

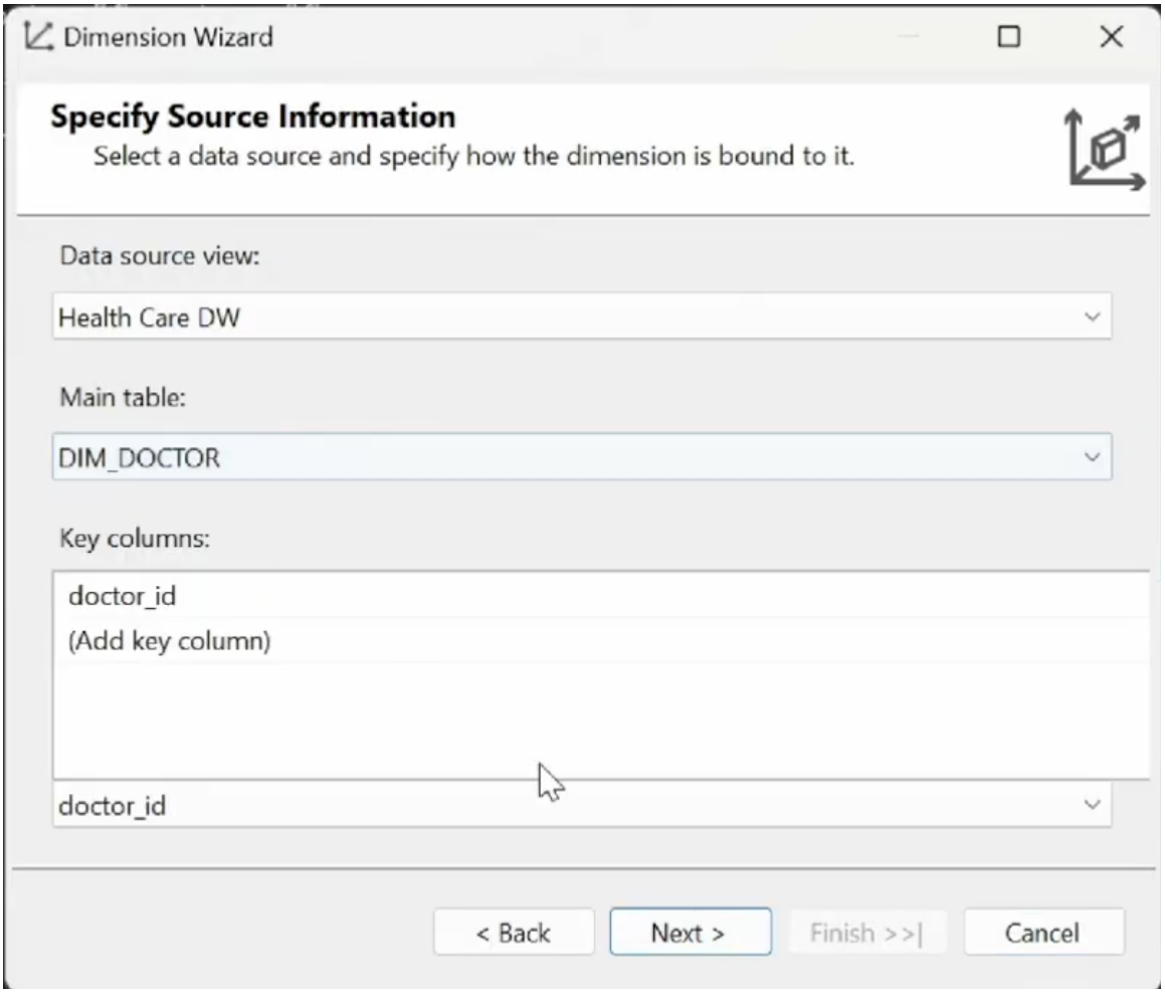



*Hình 92. Chọn thuộc tính cho DIM_DATE*

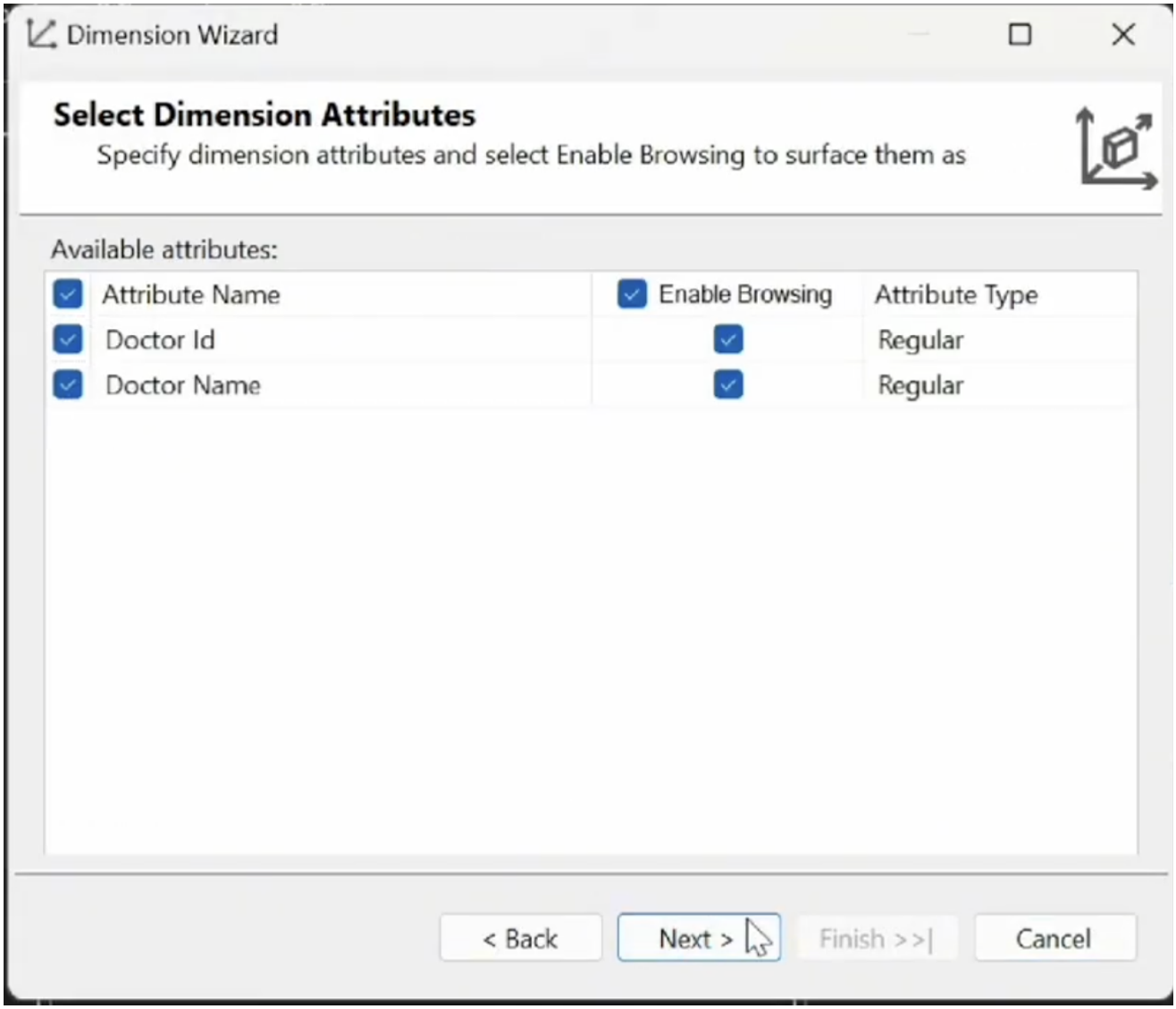



*Hình 93 . Completing*

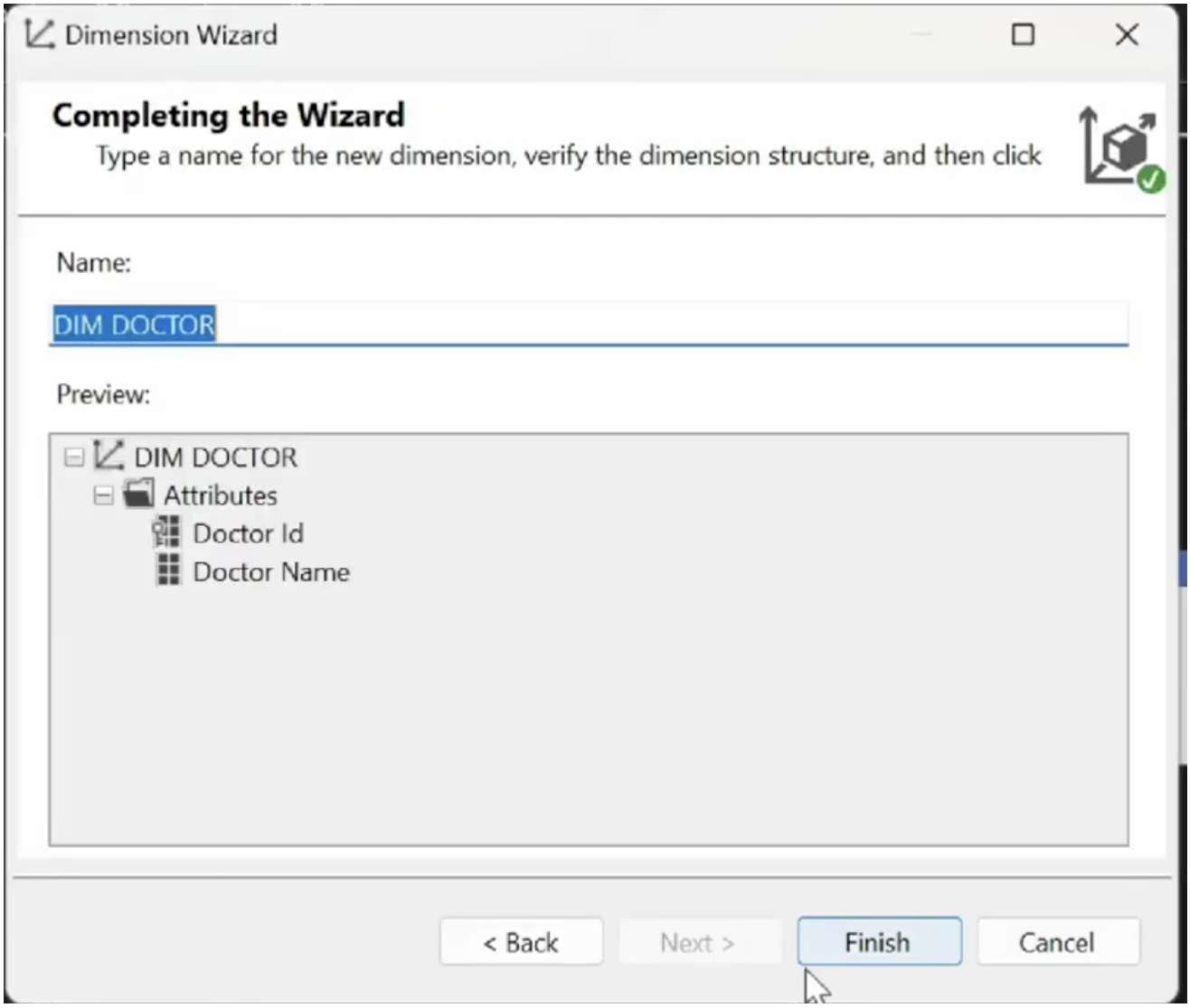
- **DIM_DOCTOR**




*Hình 94. DIM_DOCTOR*

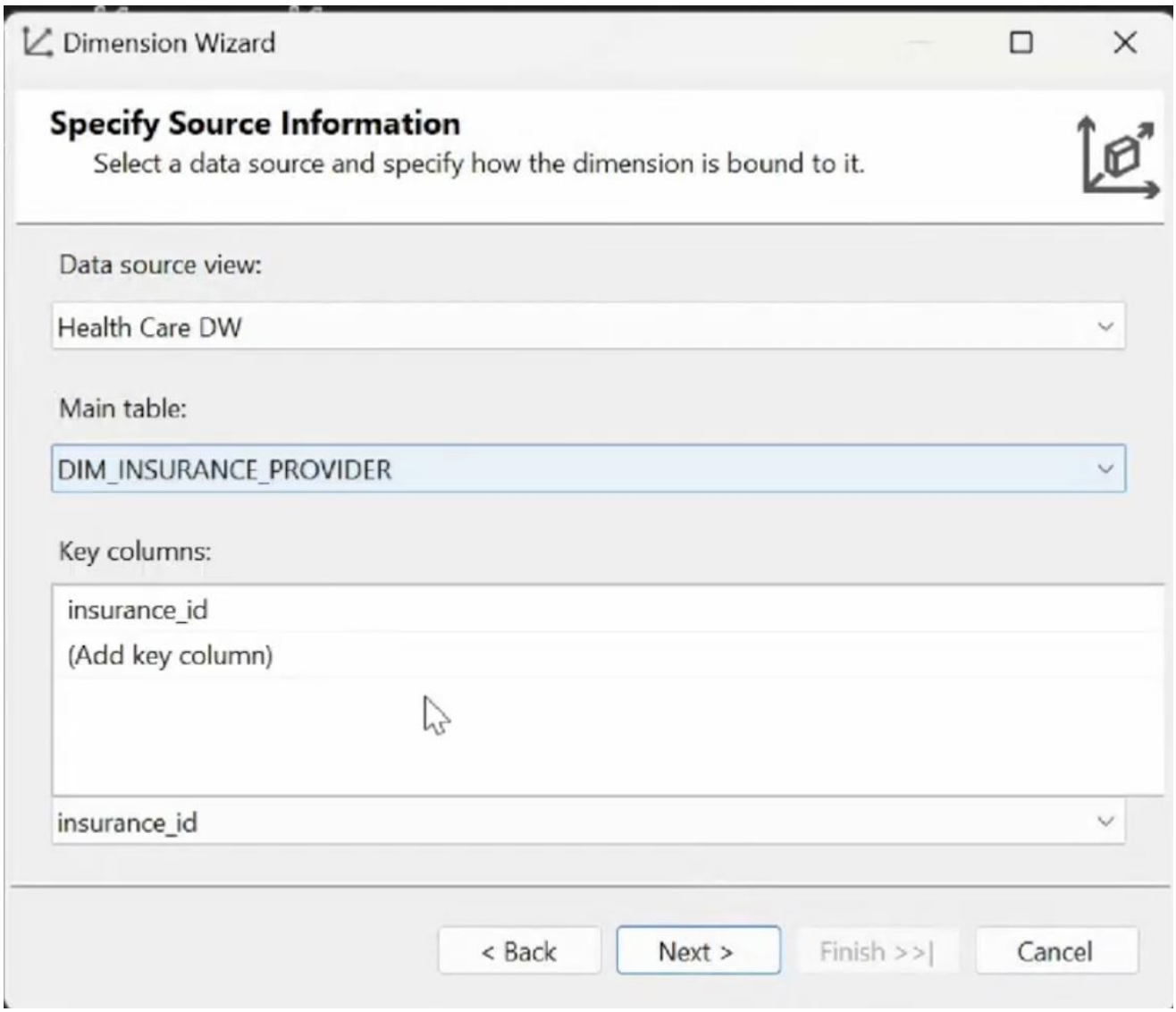



*Hình 95. Chọn thuộc tính cho DIM_DOCTOR*

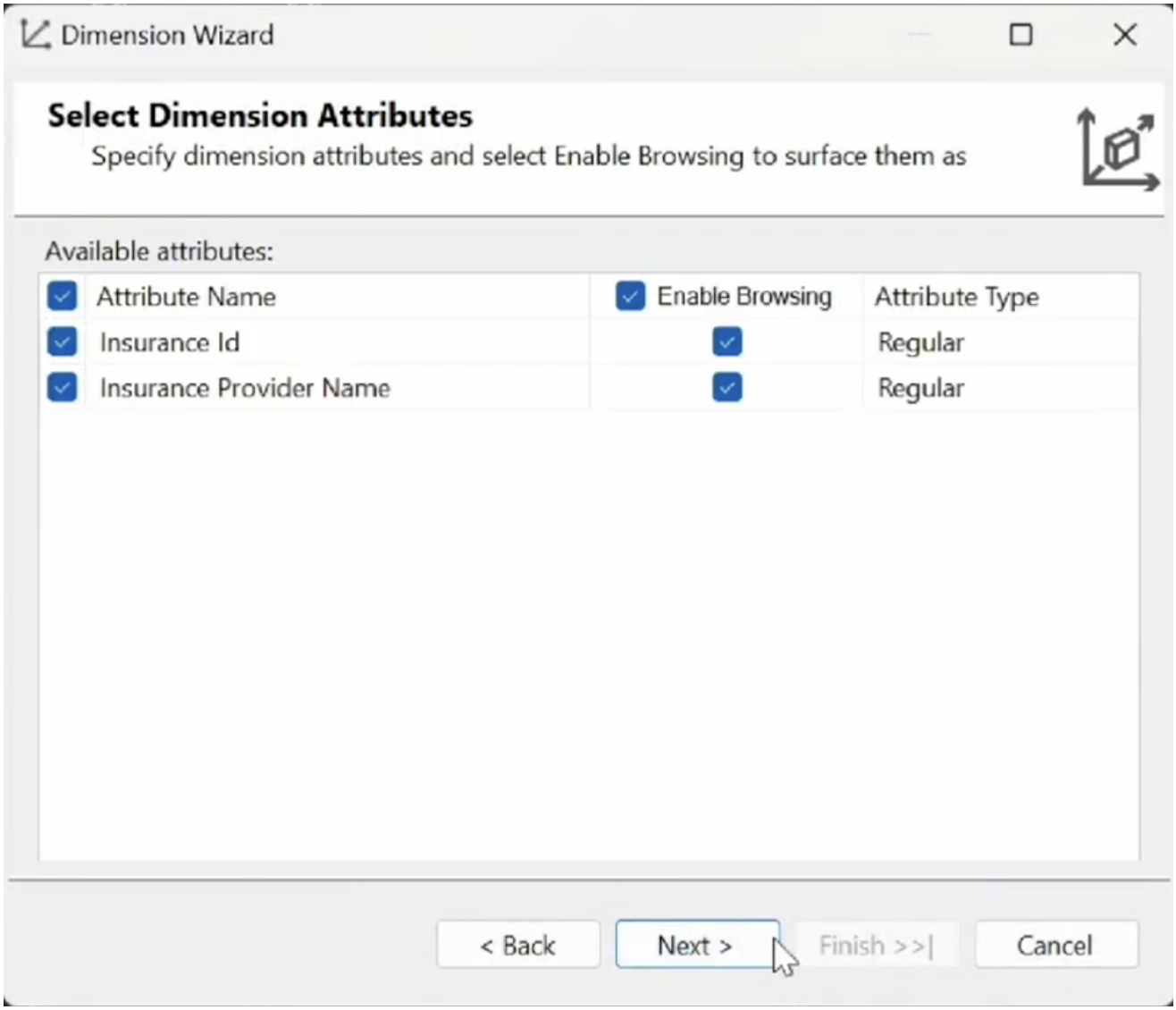



*Hình 96. Hoàn thành tạo DIM_DOCTOR*

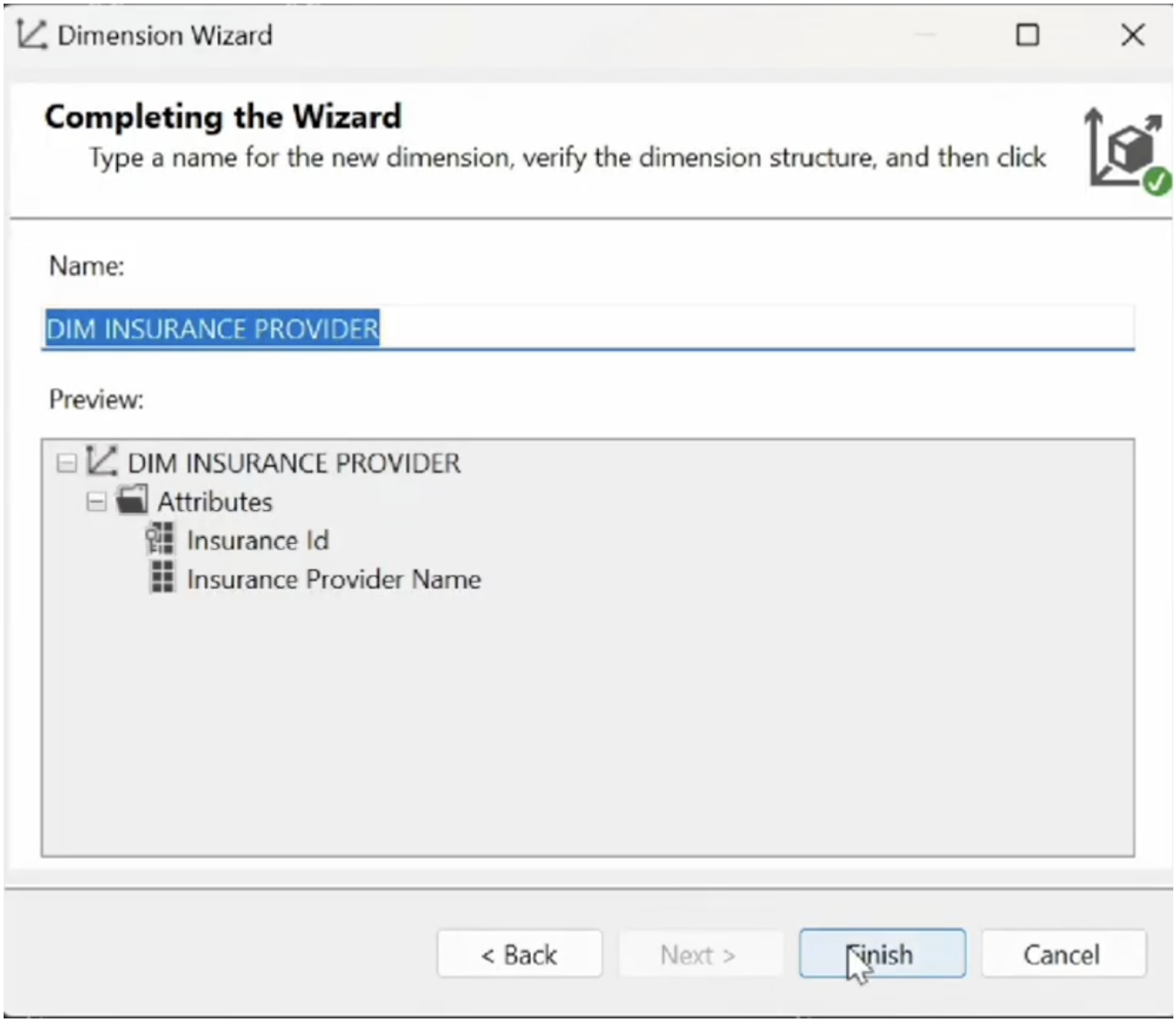
- **DIM_INSURANCE_PROVIDER**




*Hình 97. DIM_INSURANCE_PROVIDER*

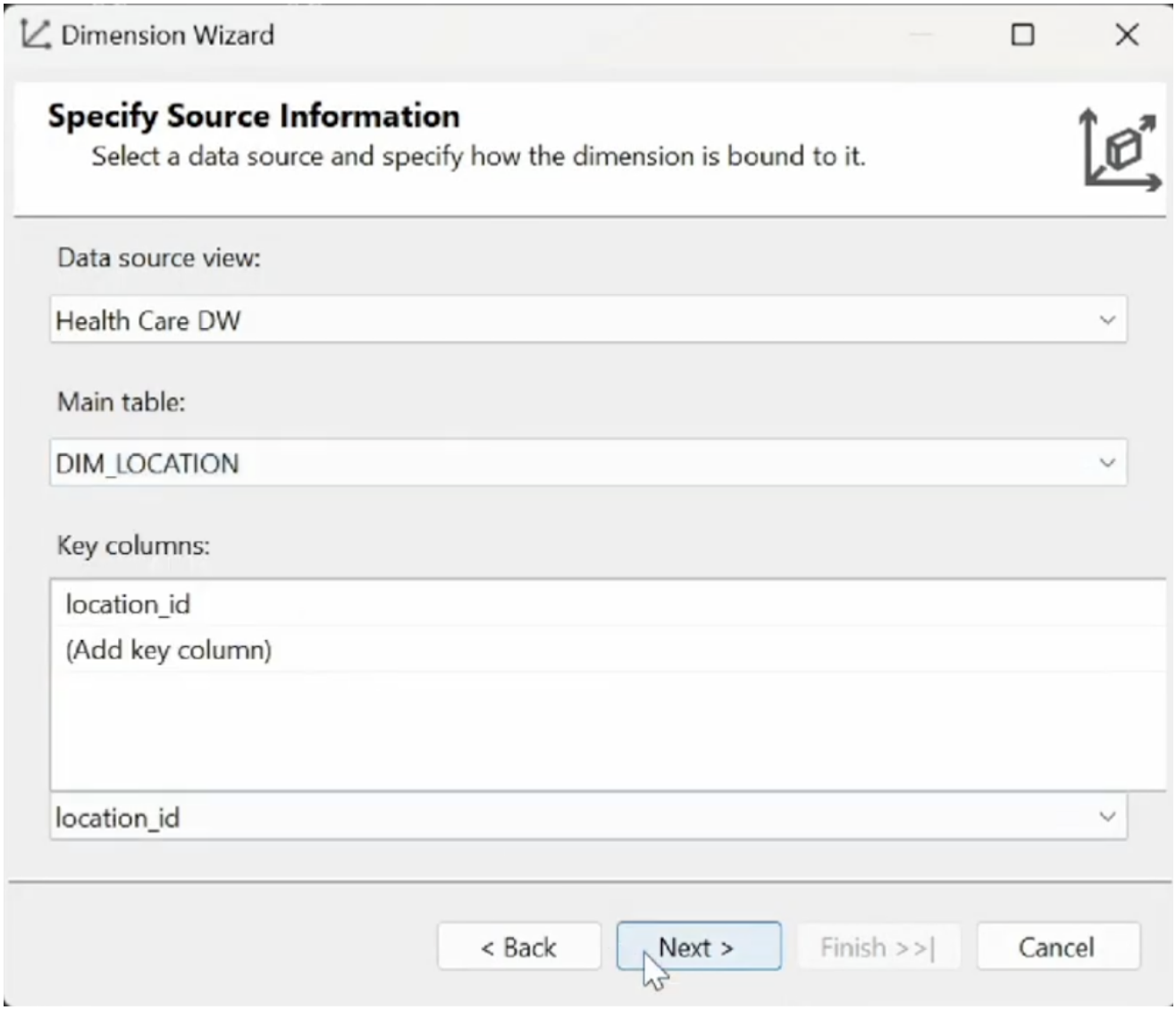



*Hình 98. Chọn thuộc tính cho DIM_INSURANCE_PROVIDER*

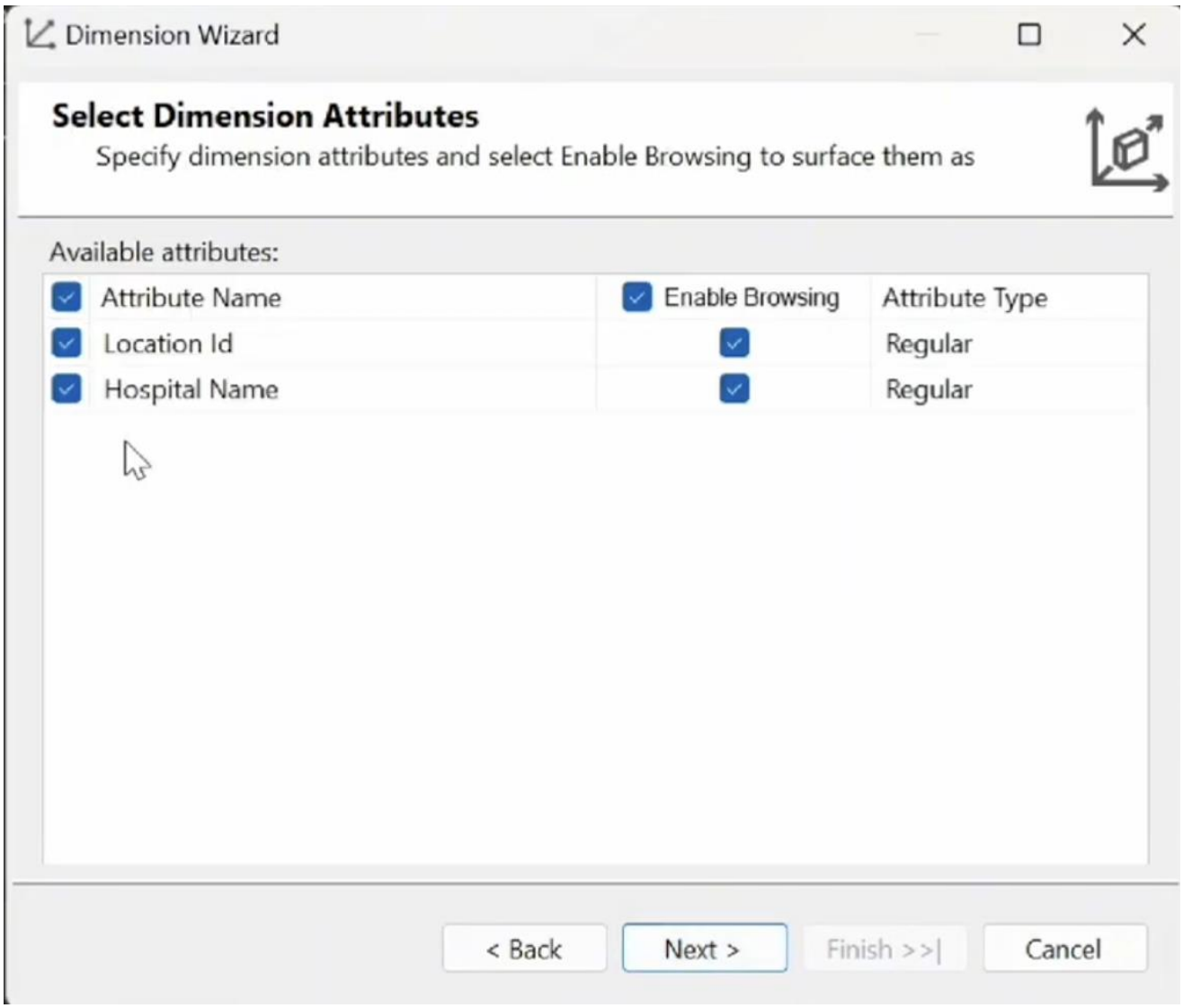



*Hình 99 . Hoàn thành tạo DIM_INSURANCE_PROVIDER*

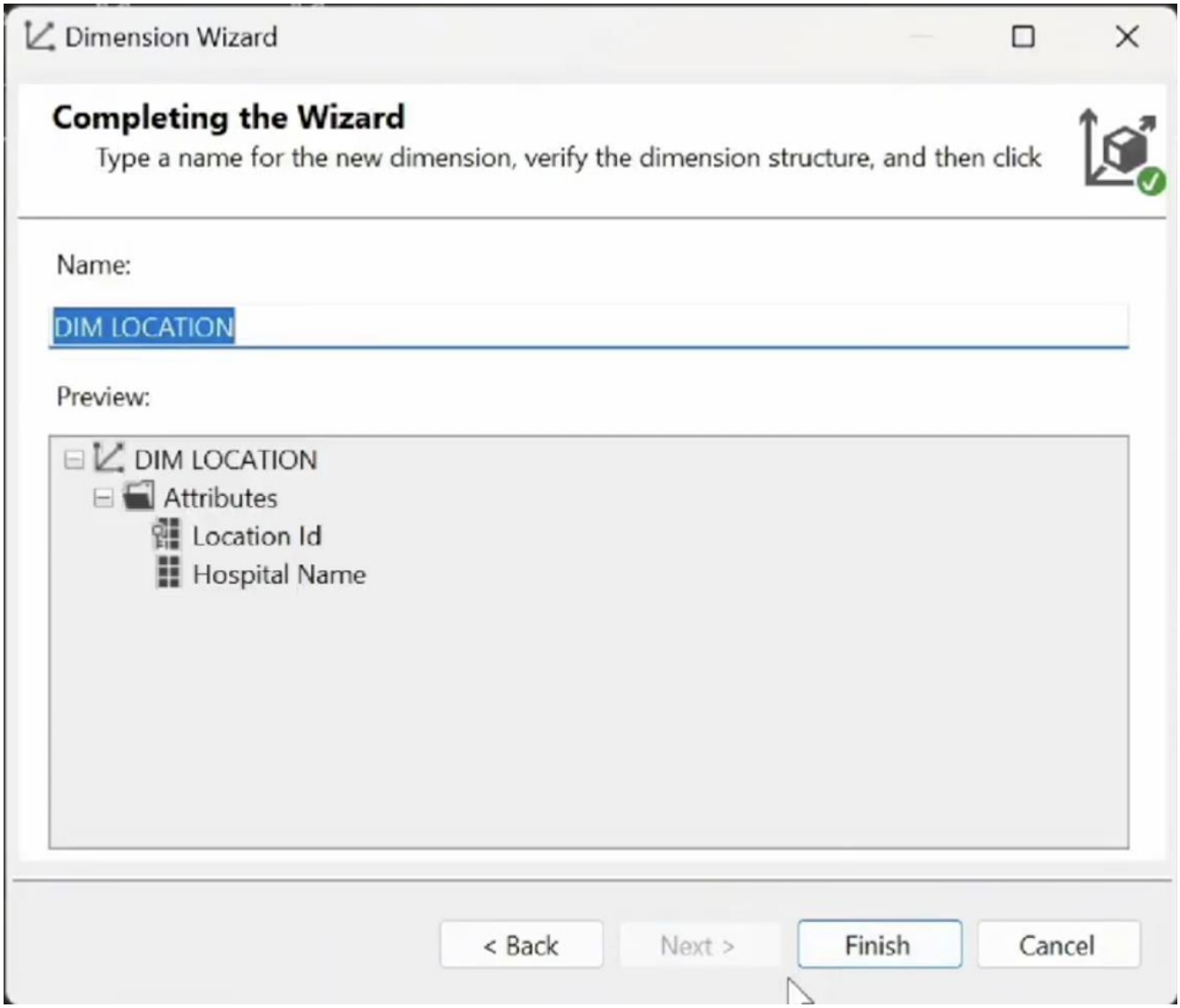
- **DIM_LOCATION**




*Hình 100. DIM_LOCATION*

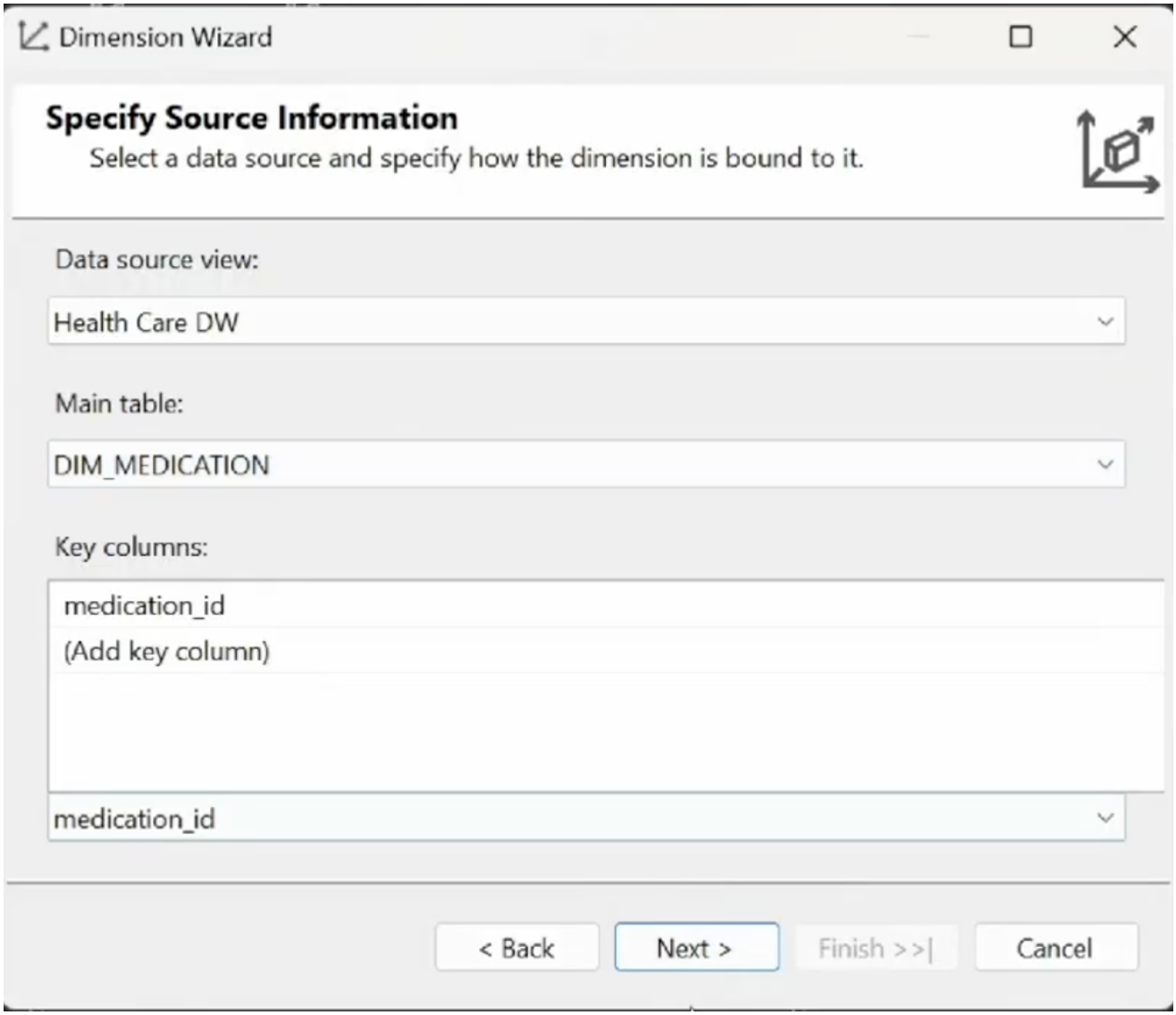



*Hình 101. Chọn thuộc tính cho DIM_LOCATION*

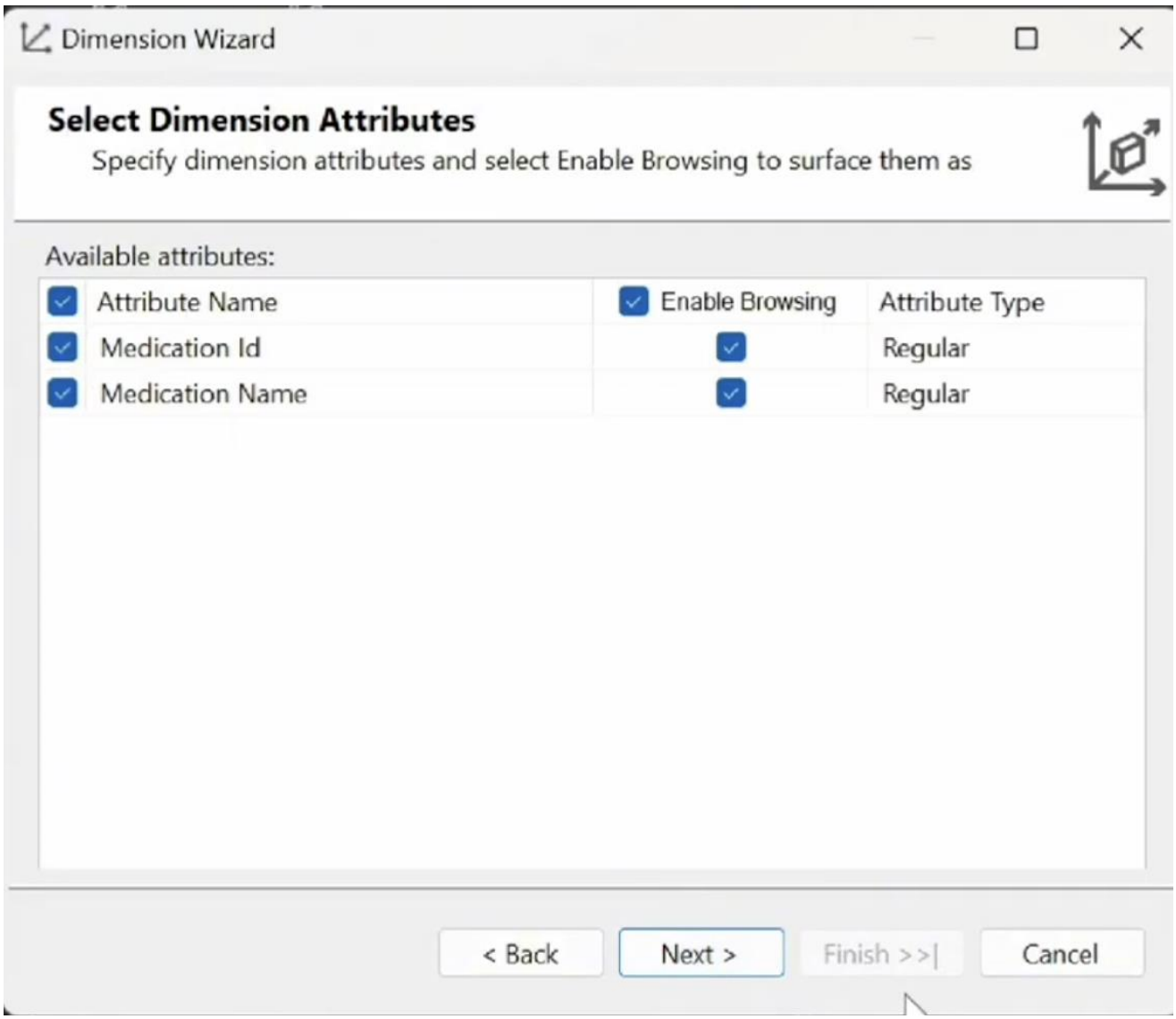



*Hình 102 . Hoàn thành tạo DIM_LOCATION*

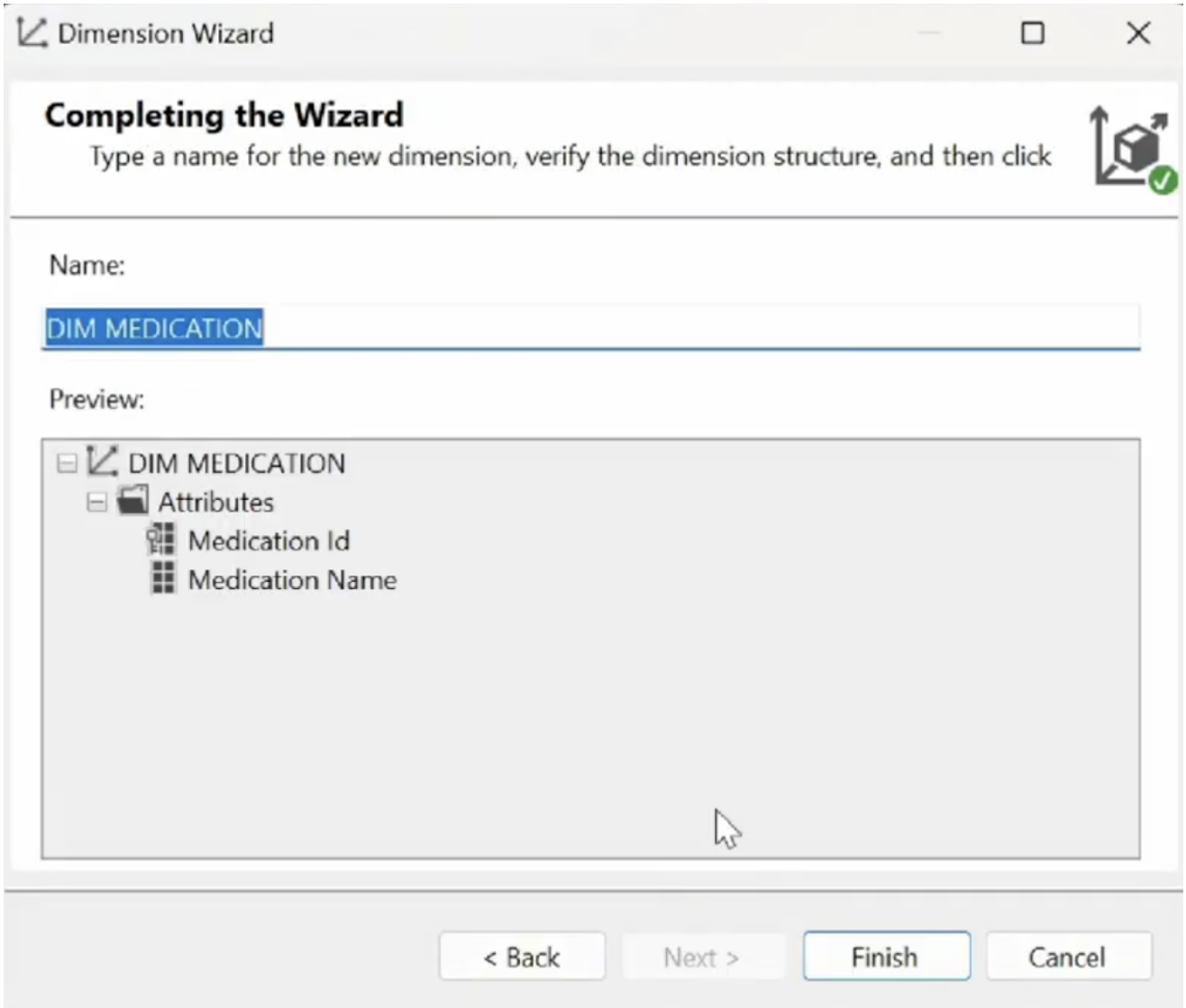
- **DIM_MEDICATION**




*Hình 103. DIM_MEDICATION*

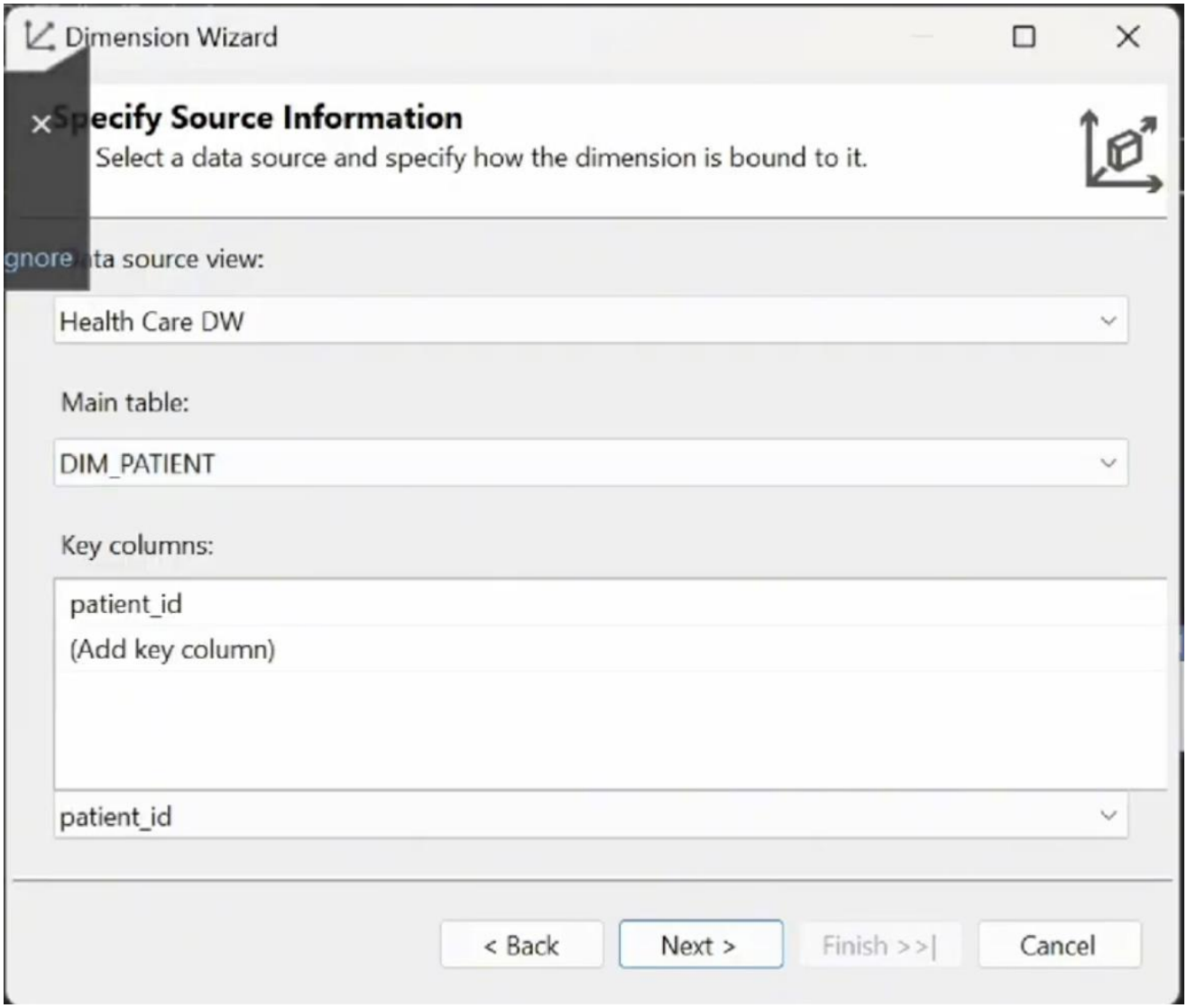



*Hình 104. Chọn thuộc tính cho DIM_MEDICATION*

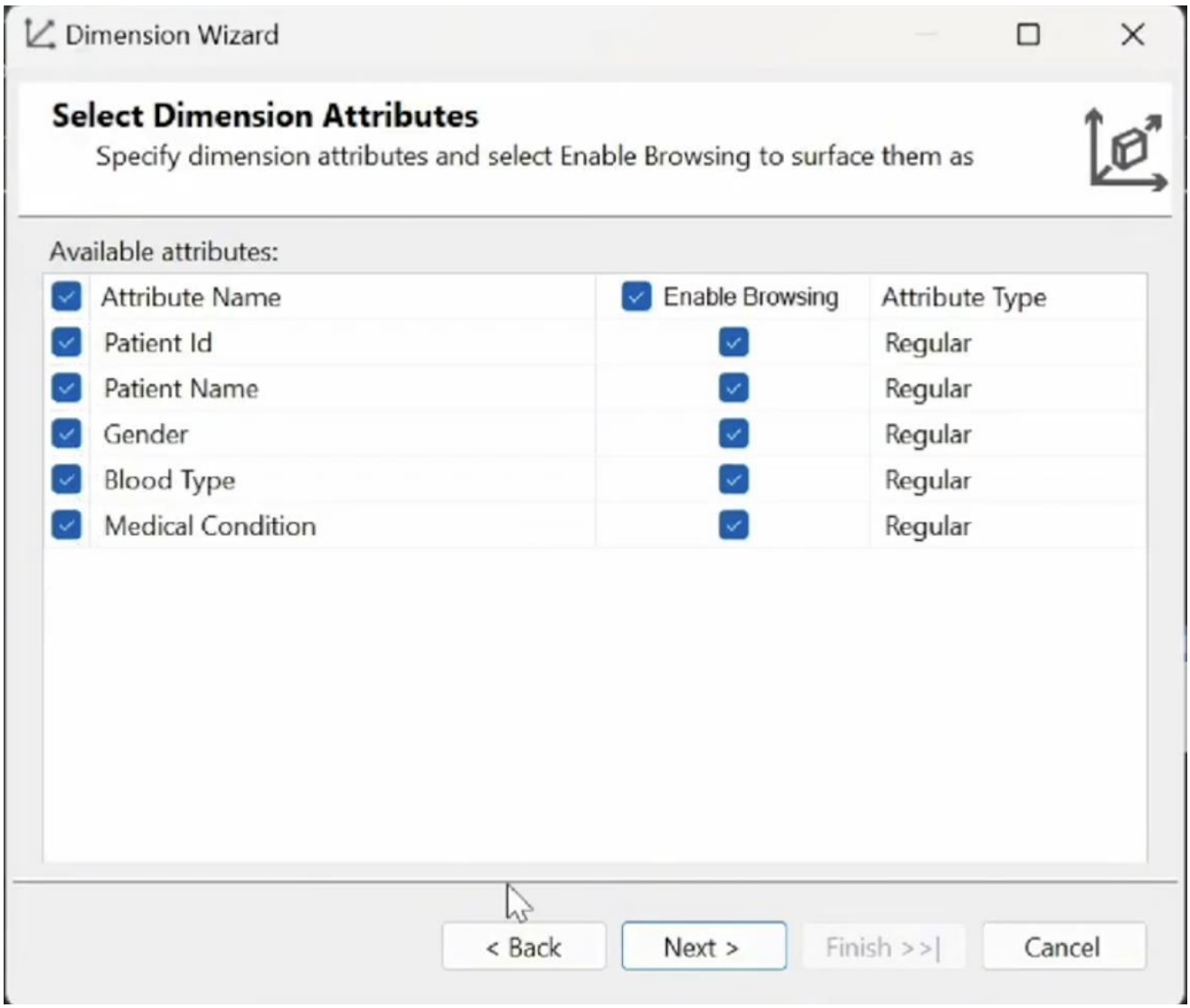



*Hình 105. Hoàn thành tạo DIM_MEDICATION*

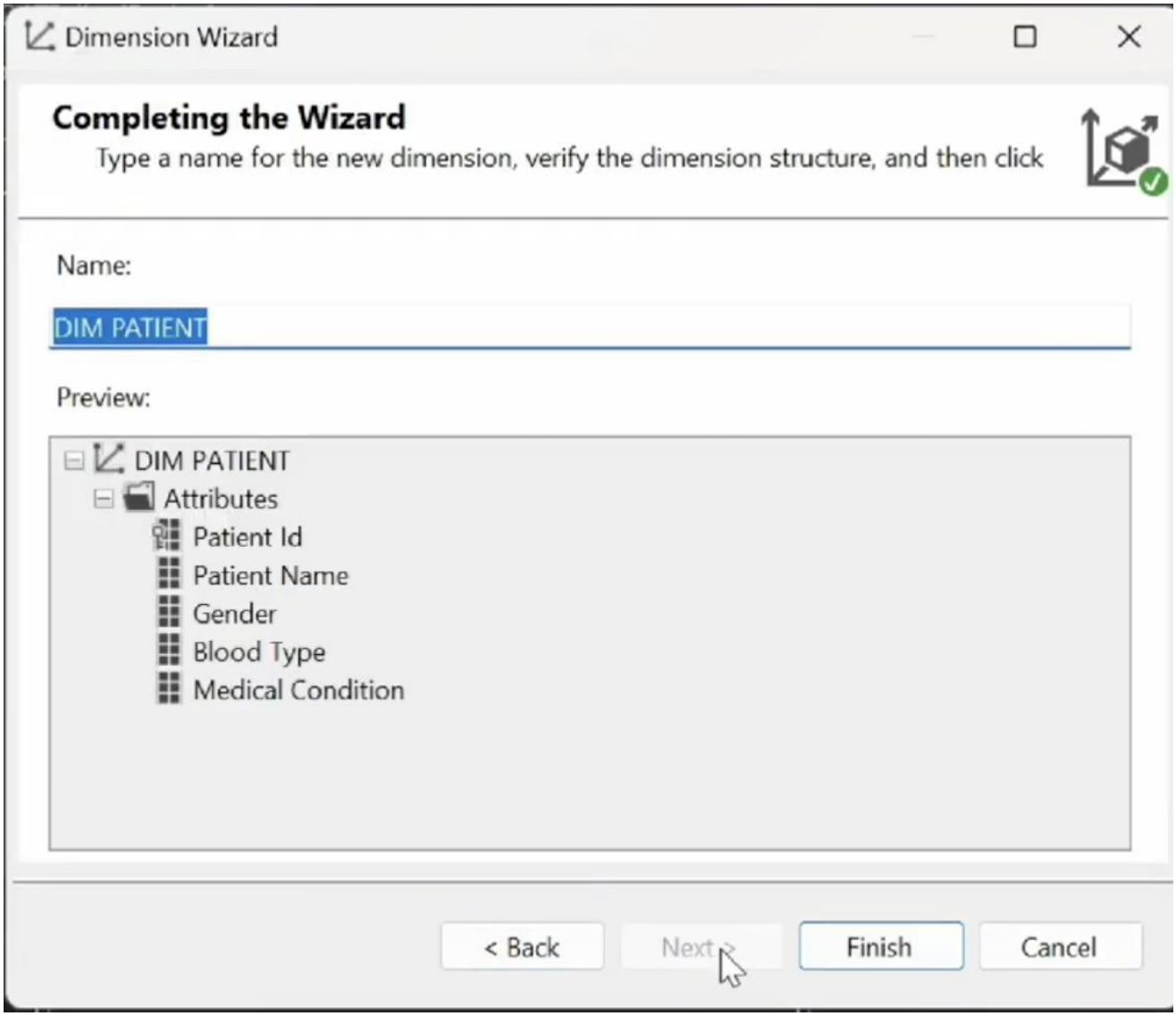
- **DIM_PATIENT**




*Hình 106 . DIM_PATIENT*

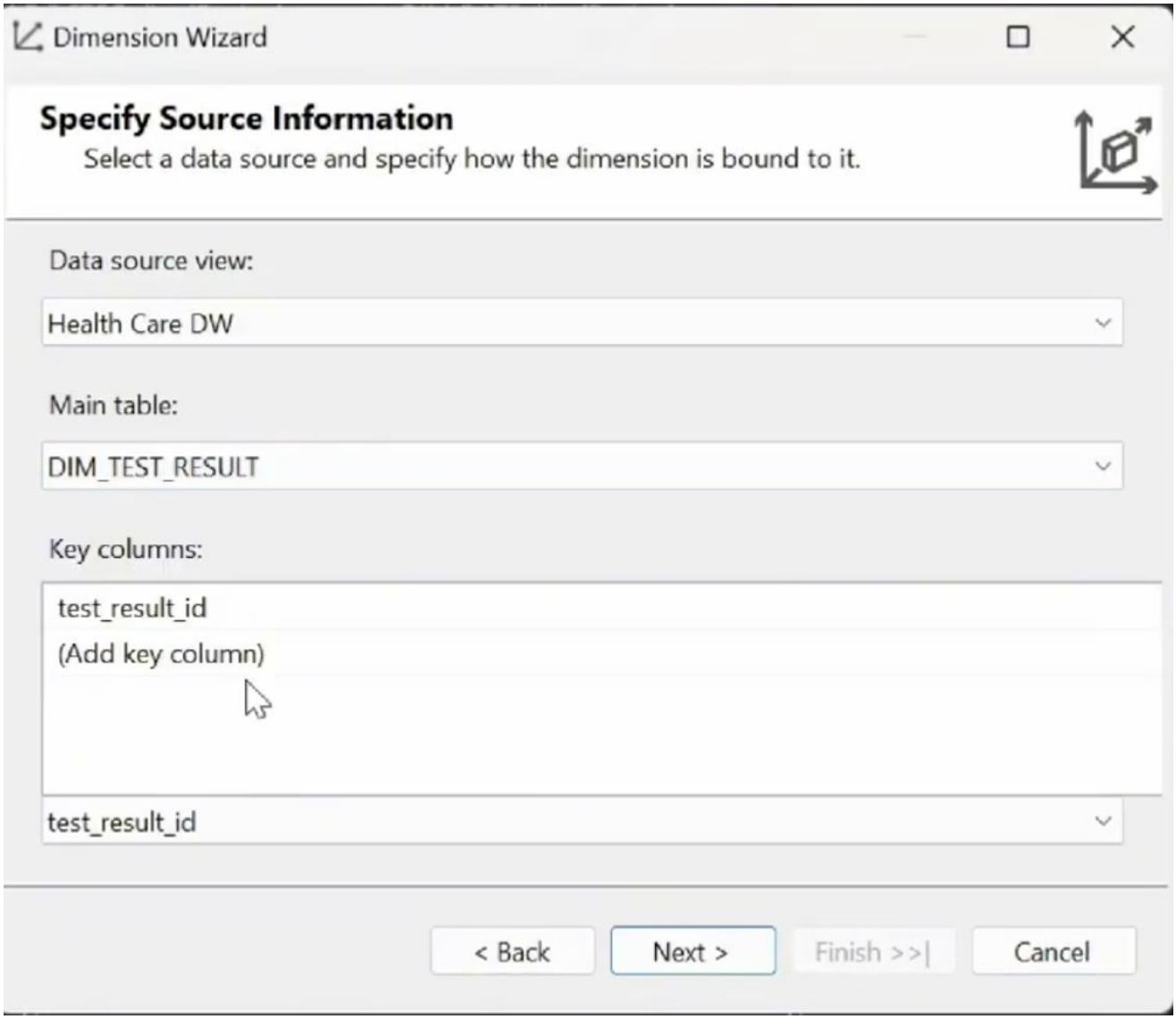



*Hình 107. Chọn thuộc tính cho DIM_PATIENT*

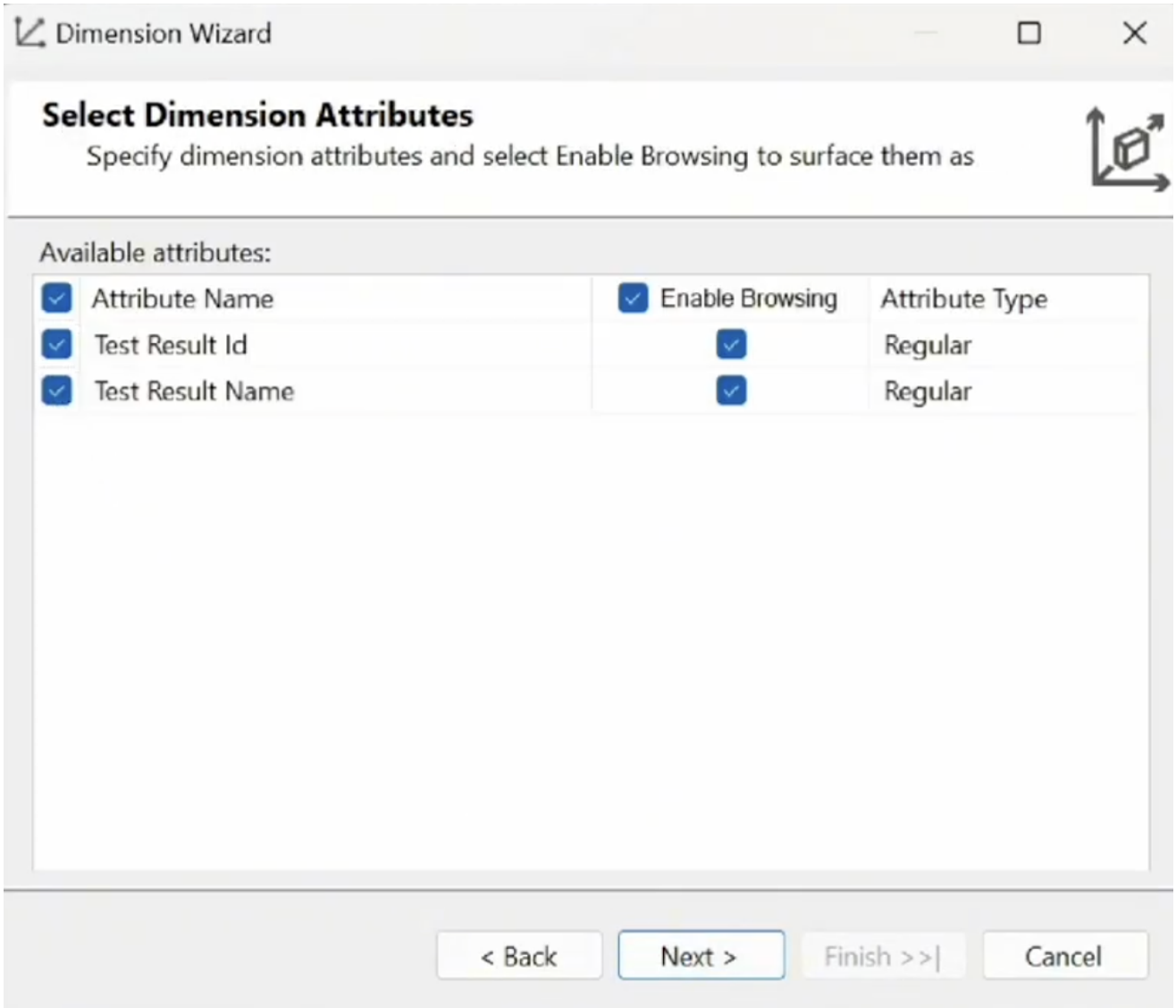



*Hình 108 . Hoàn thành tạo DIM_PATIENT*

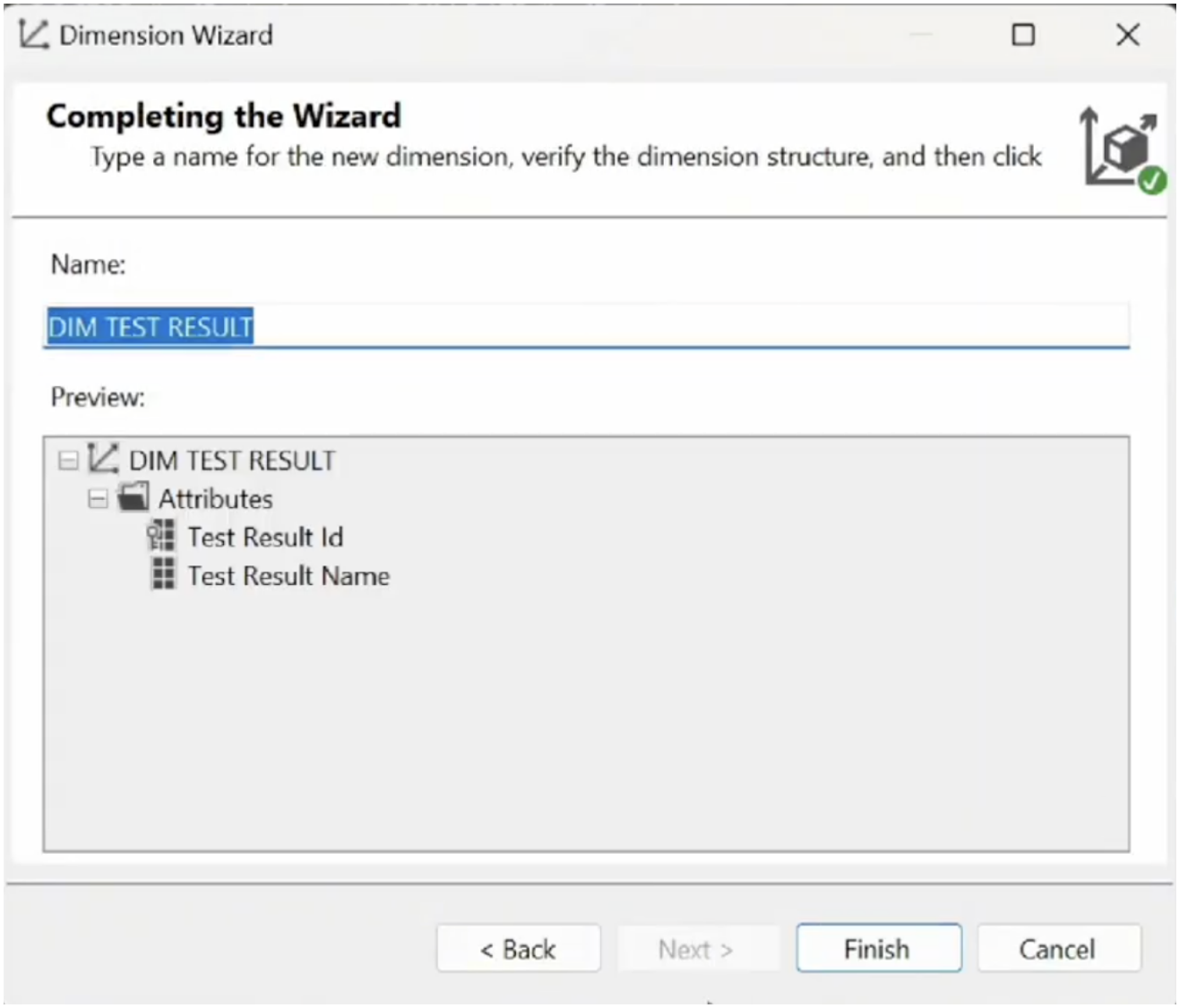
- **DIM_TEST_RESULT**




*Hình 109 . DIM_ TEST_RESULT*



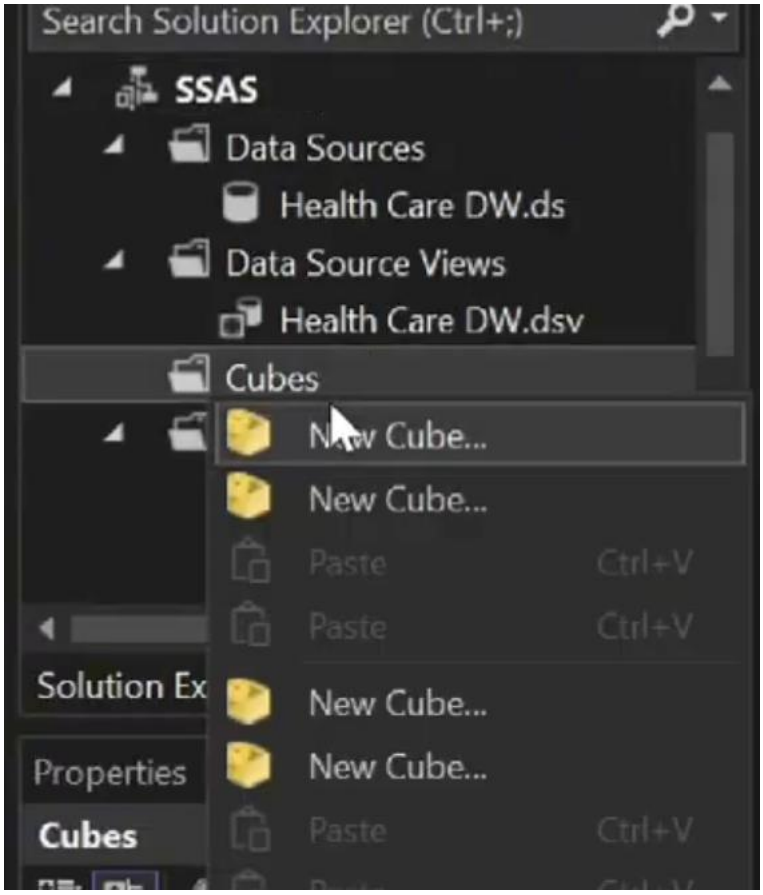

*Hình 110 . Chọn thuộc tính cho DIM_TEST_RESULT*




*Hình 111 . Hoàn thành tạo DIM_TEST_RESULT*

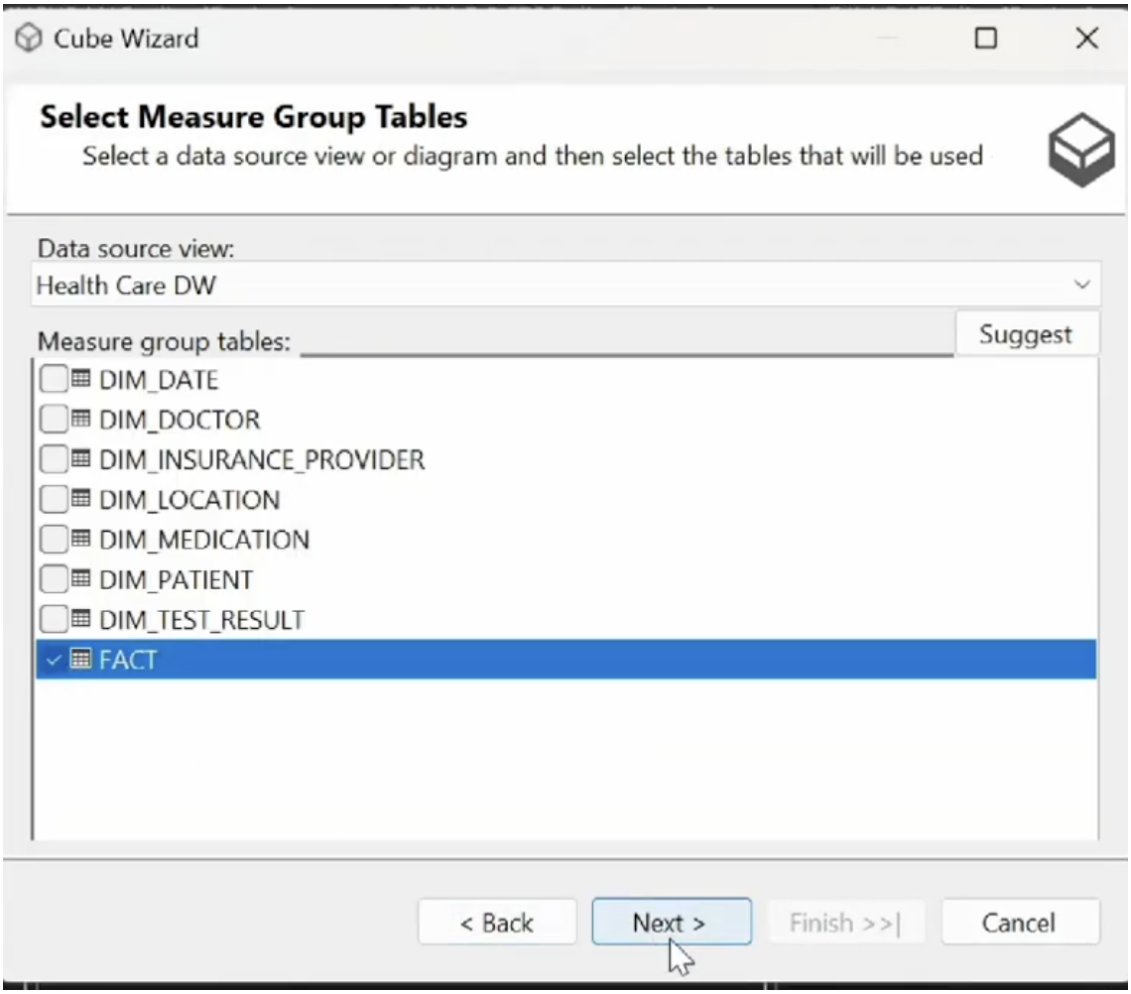
## 5. XÂY DỰNG CÁC KHỐI VÀ CÁC MEASURE

- **Bước 1:** Bên gốc phải phần Solution Explorer nhấn chuột phải vào Cubes và chọn New Cube.



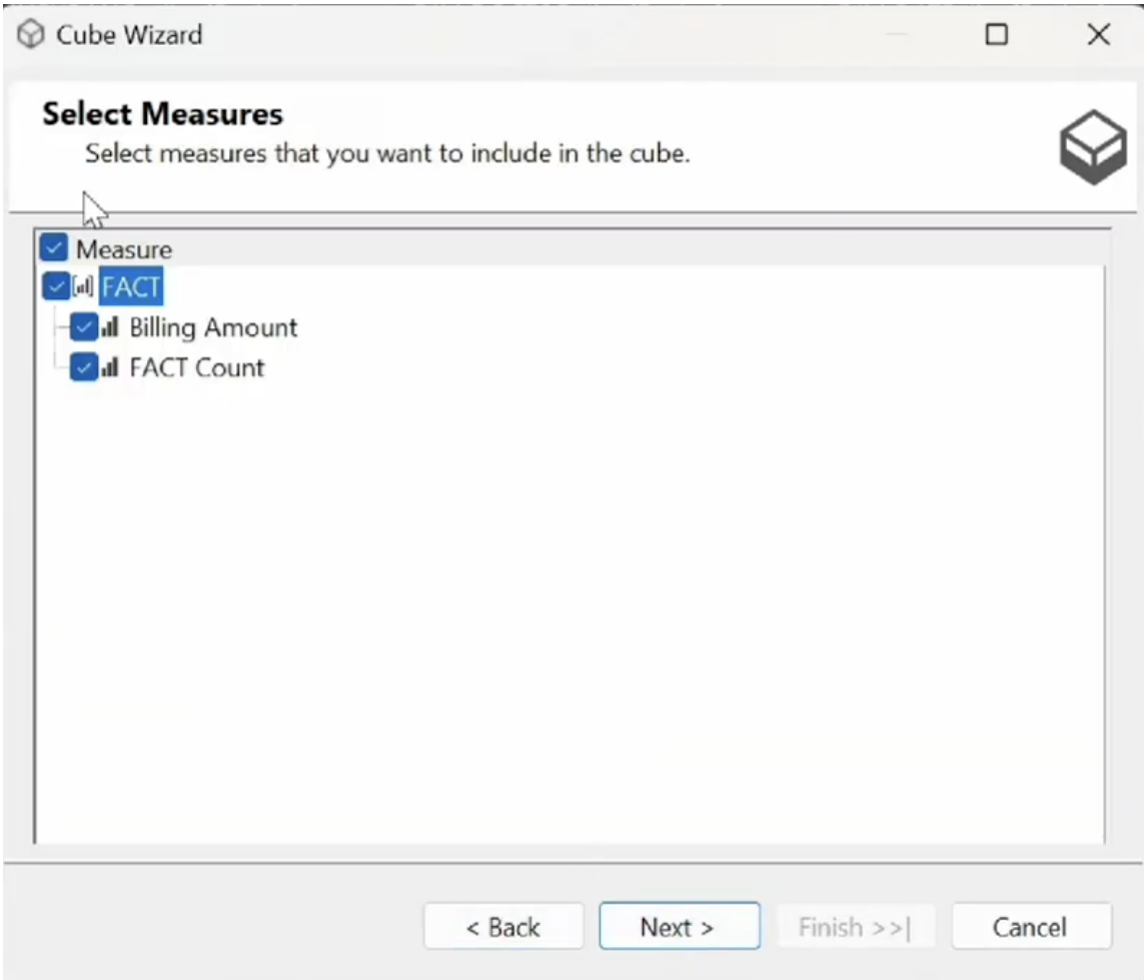

*Hình 112 . New Cube*

- **Bước 2:** Chọn bảng FACT và chọn Next.



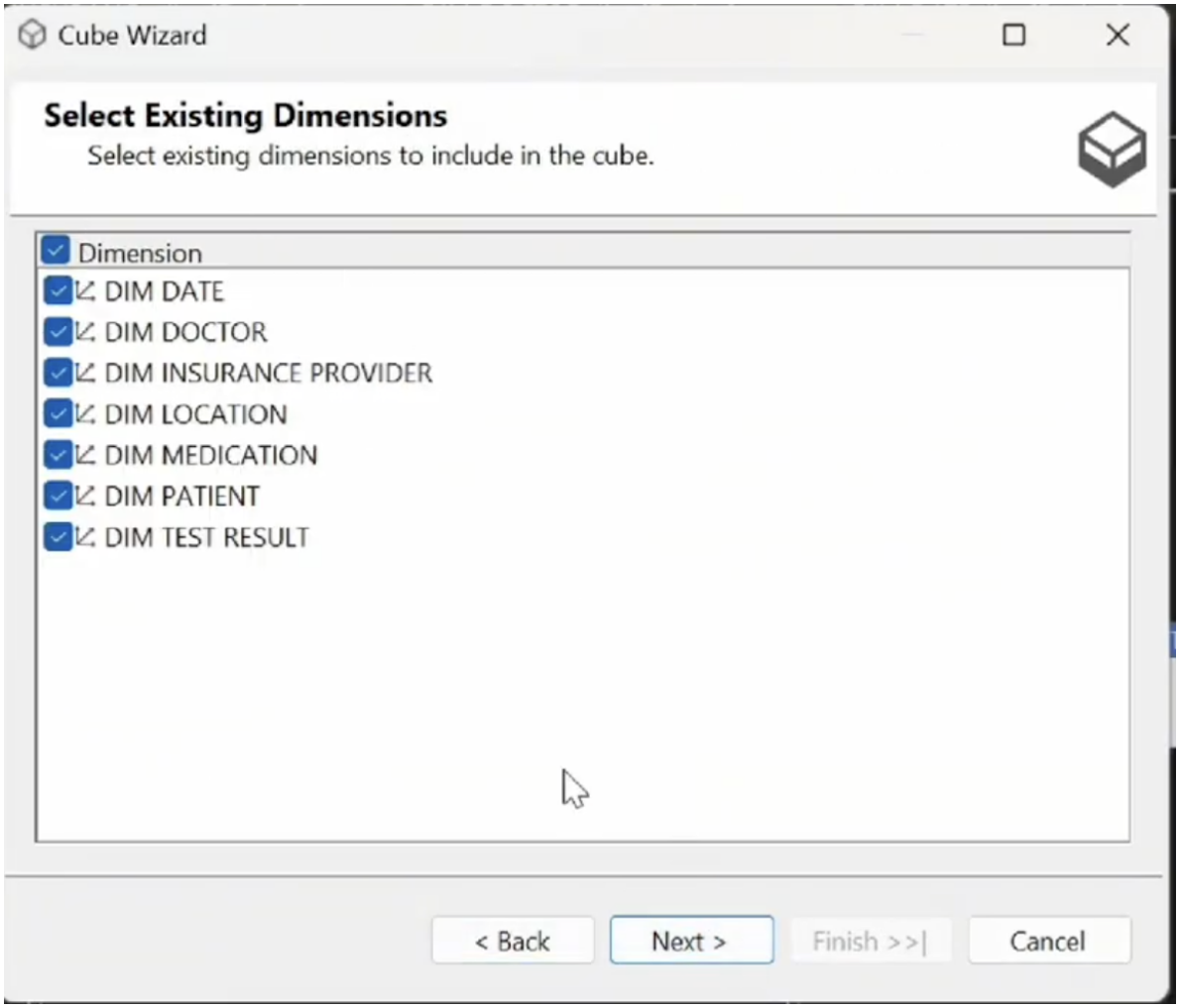

*Hình 113 . Chọn bảng FACT*

- **Bước 3:** Chọn các độ đo và bấm Next




*Hình 114 . Chọn các độ đo*

- **Bước 4:** Chọn các bảng DIM và bấm Next




*Hình 115. Chọn các bảng DIM*

- **Bước 2:** Chọn finish để kết thúc quá trình tạo cube.

*Hình 116 . Kết thúc quá trình tạo Cube*

## 6. ĐỔI TÊN VÀ THUỘC TÍNH CÁC MEASURES

- **Bước 1:** Để đổi tên và thuộc tính các độ đo trong phần Measure ta chọn Show measures grid
- **Bước 2:** Đổi tên các thuộc tính theo ý thích

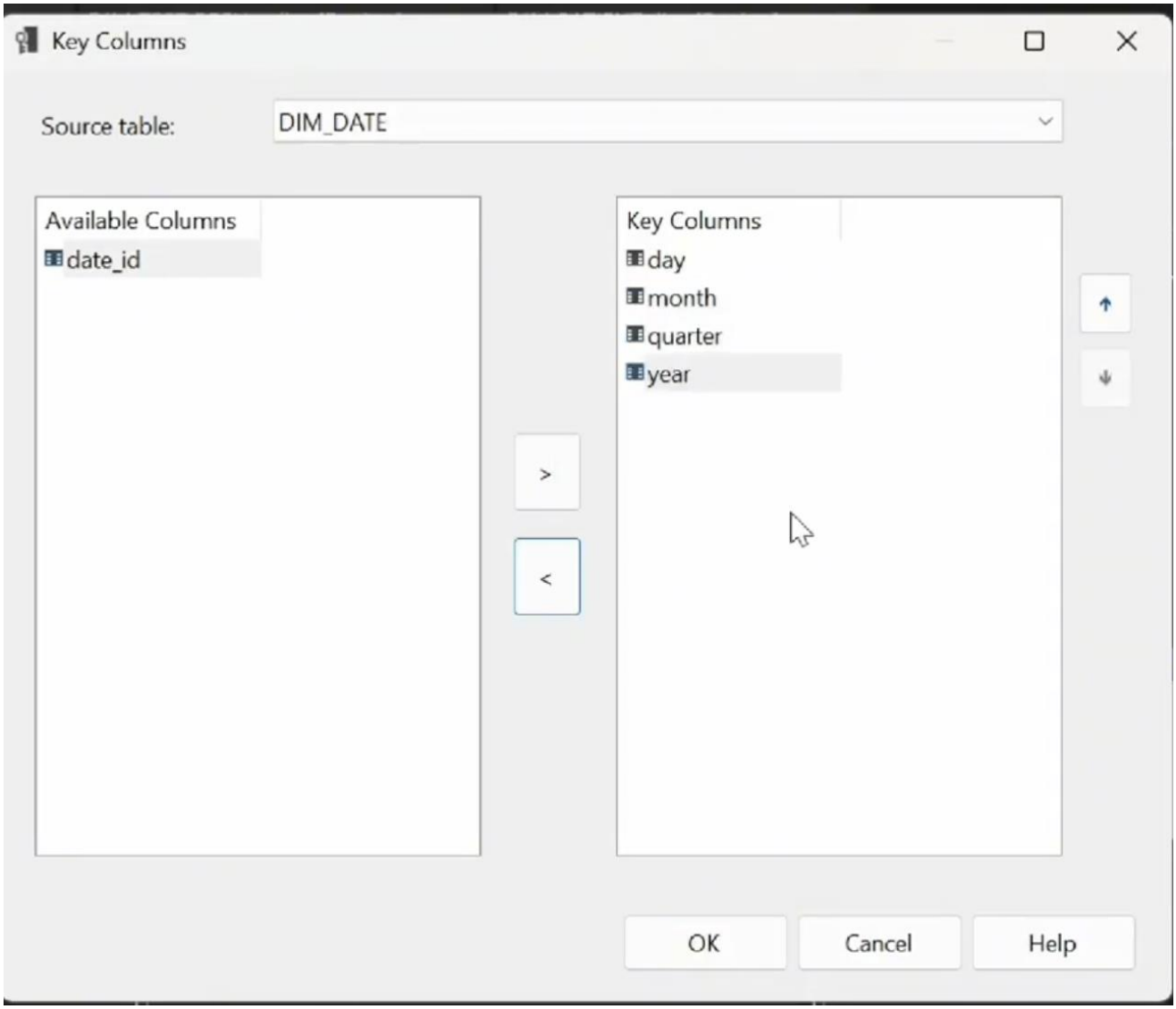
## 7. PHÂN CẤP TRONG BẢNG CHIỀU

### 7.1. Bảng Dim_Date

- **Bước 1:** Kéo những thuộc tính cần thiết sang cửa sổ Hirearchy.

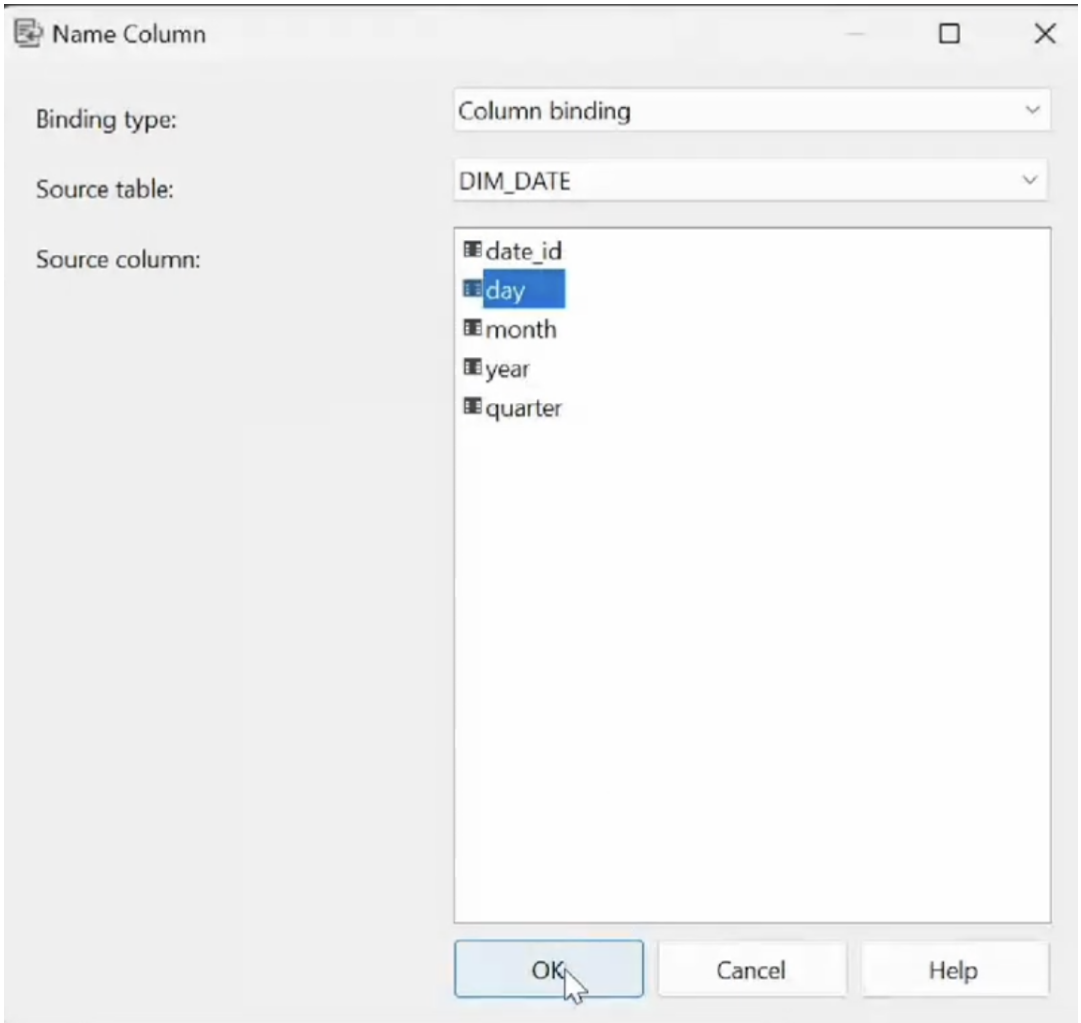
*Hình 117 . Phân cấp bảng DIM_DATE*

- **Bước 2:** Chọn thuộc tính Day và chọn Key Column trong properties của Day. Chọn các thuộc tính có độ phân cấp lớn hơn theo thứ tự lớn dần.



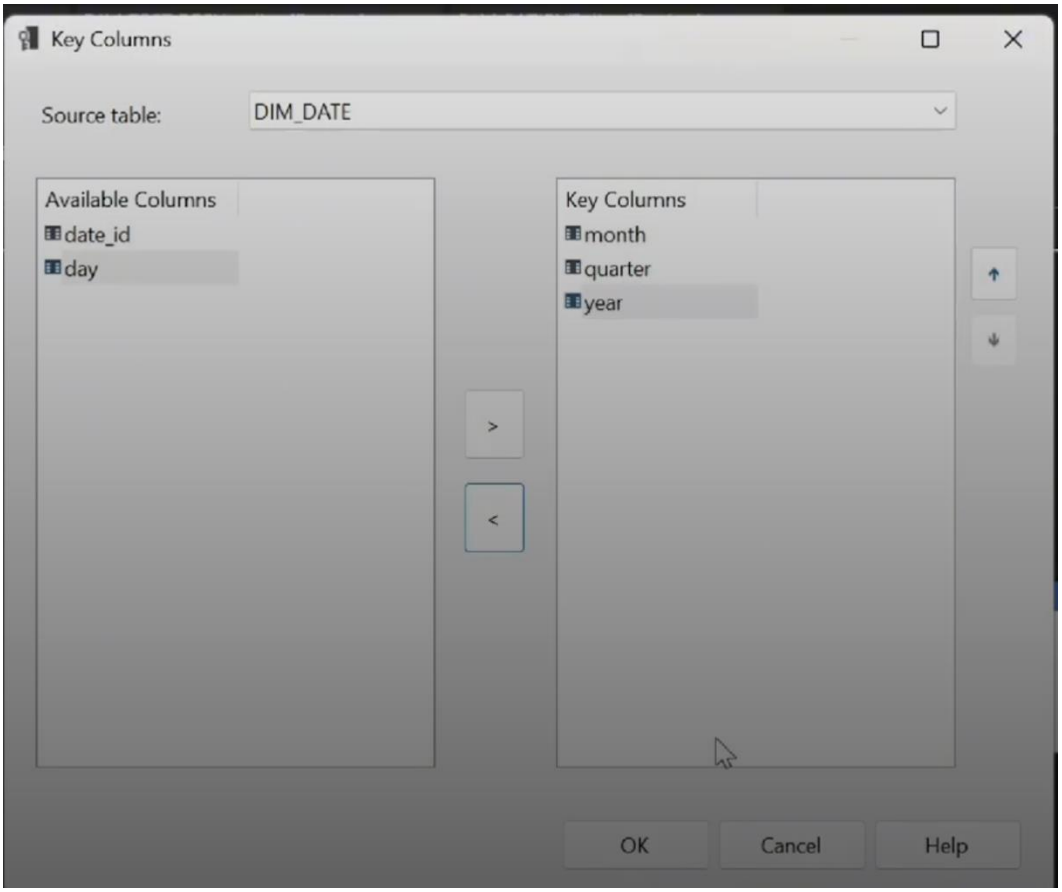

*Hình 118. Phân cấp bảng DIM_DATE*

- **Bước 3:** Đổi Name Column của Day.



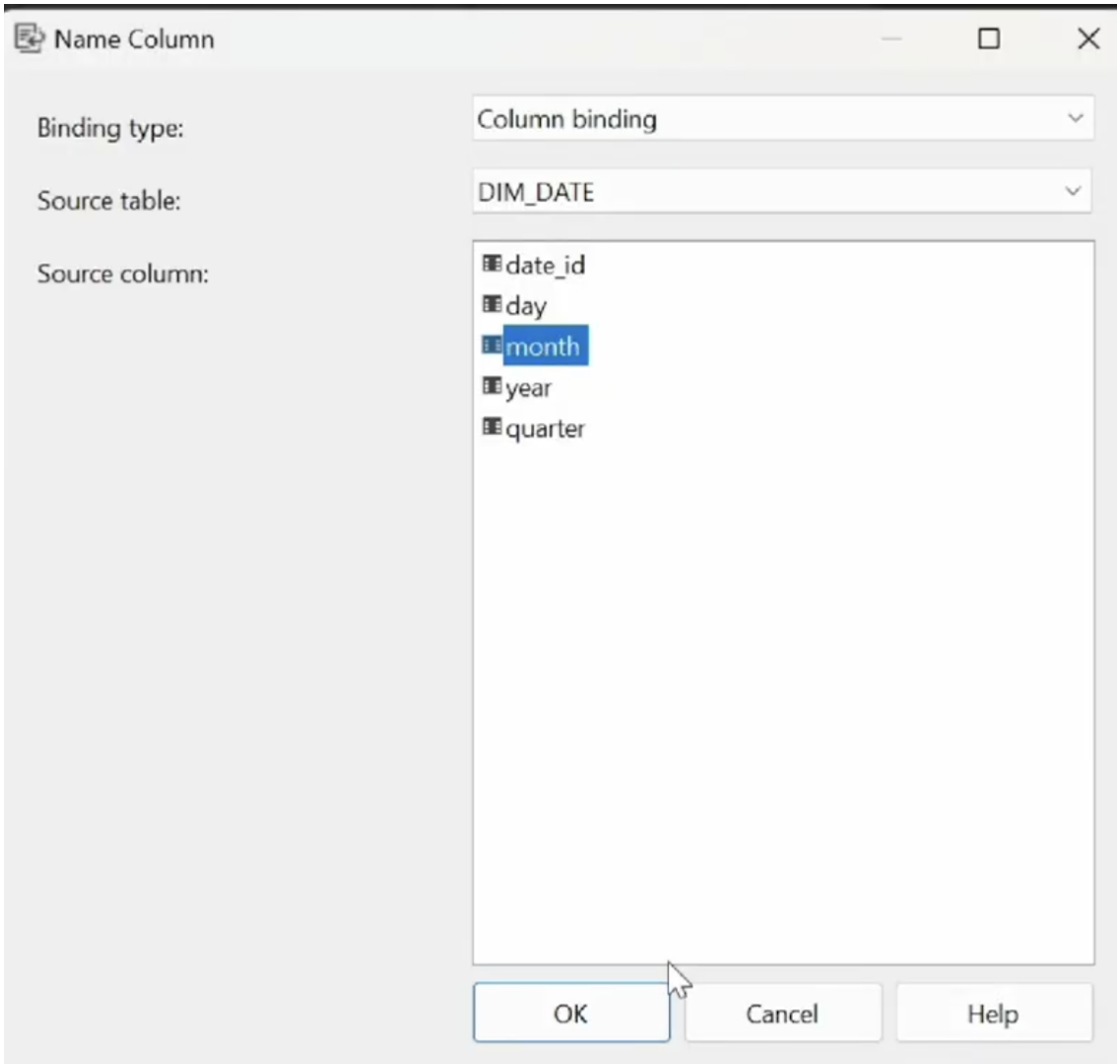

*Hình 119. Đổi Name Column của Day*

- **Bước 4:** Chọn thuộc tính Month và chọn Key Column trong properties của Month. Chọn các thuộc tính có độ phân cấp lớn hơn theo thứ tự lớn dần.



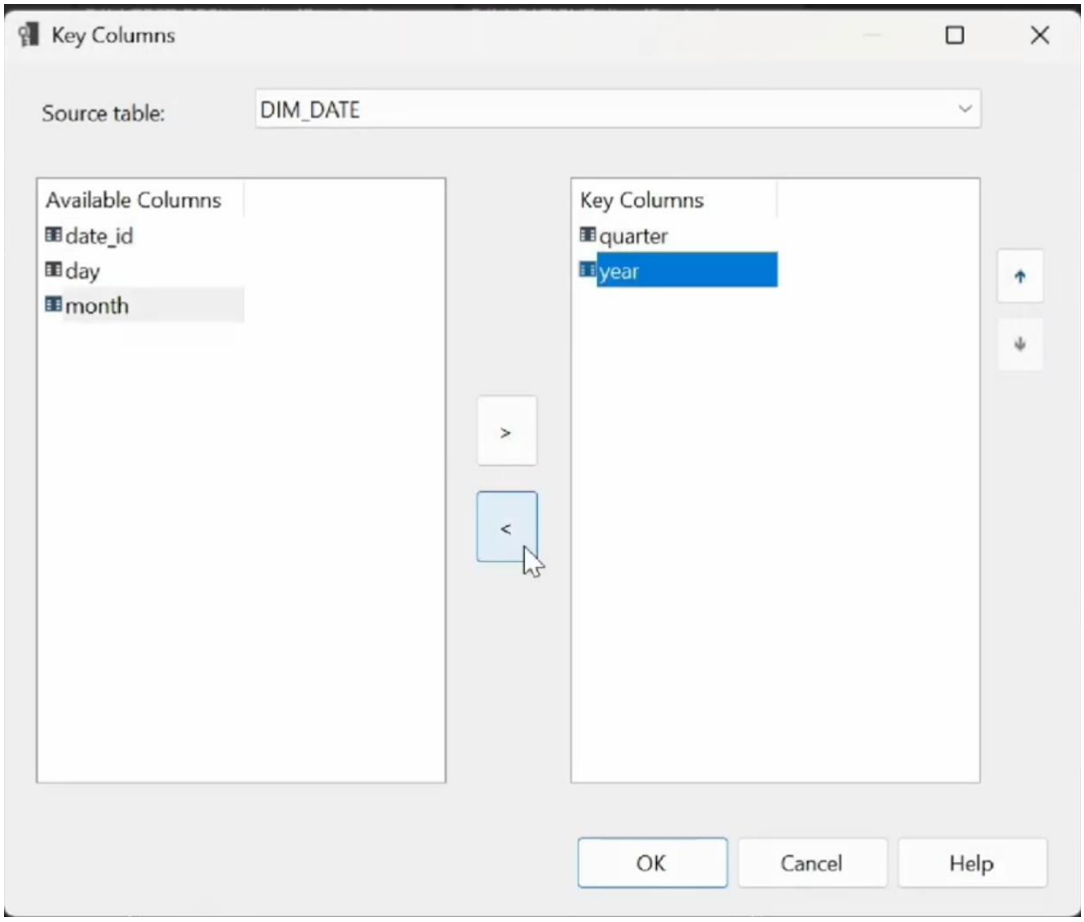

*Hình 120 . Phân cấp bảng DIM_DATE*

- **Bước 5:** Đổi Name Column của Month



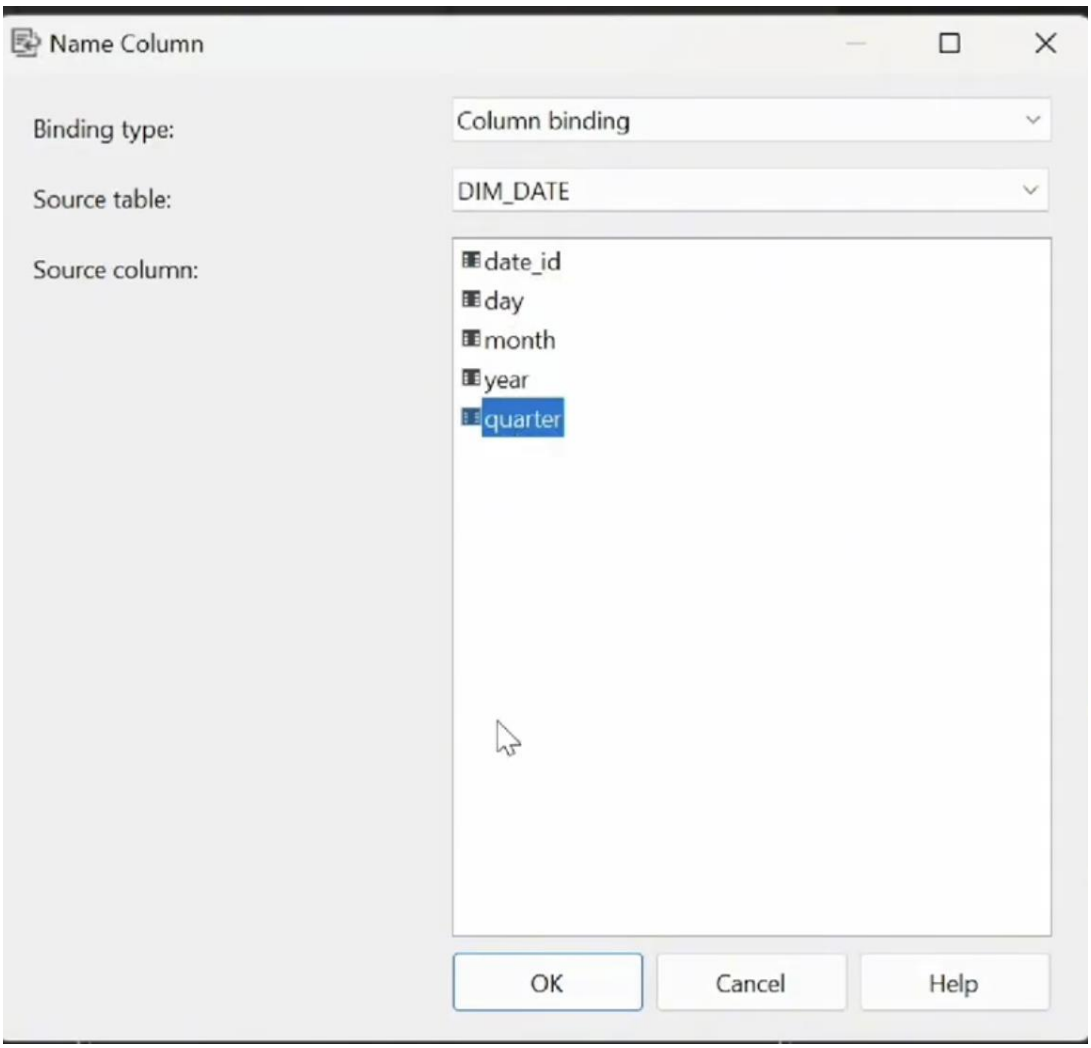

*Hình 121. Đổi Name Column của Month*

- **Bước 6:** Chọn thuộc tính Quarter và chọn Key Column trong properties của Quarter. Chọn các thuộc tính có độ phân cấp lớn hơn theo thứ tự lớn dần.



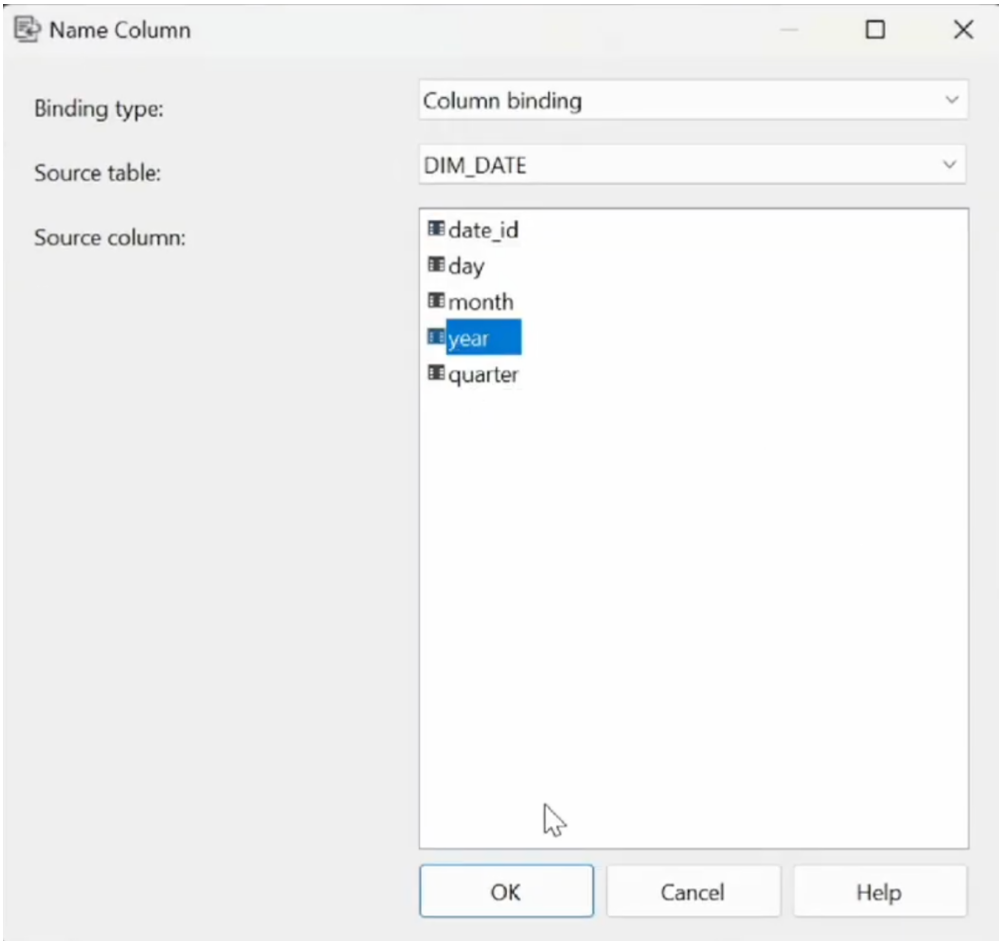

*Hình 122. Phân cấp bảng DIM_DATE*

- **Bước 7:** Đổi Name Column của Quarter.




*Hình 123 . Đổi Name Column của Quarter.*

- **Bước 8:** Đổi Name Column của Year.



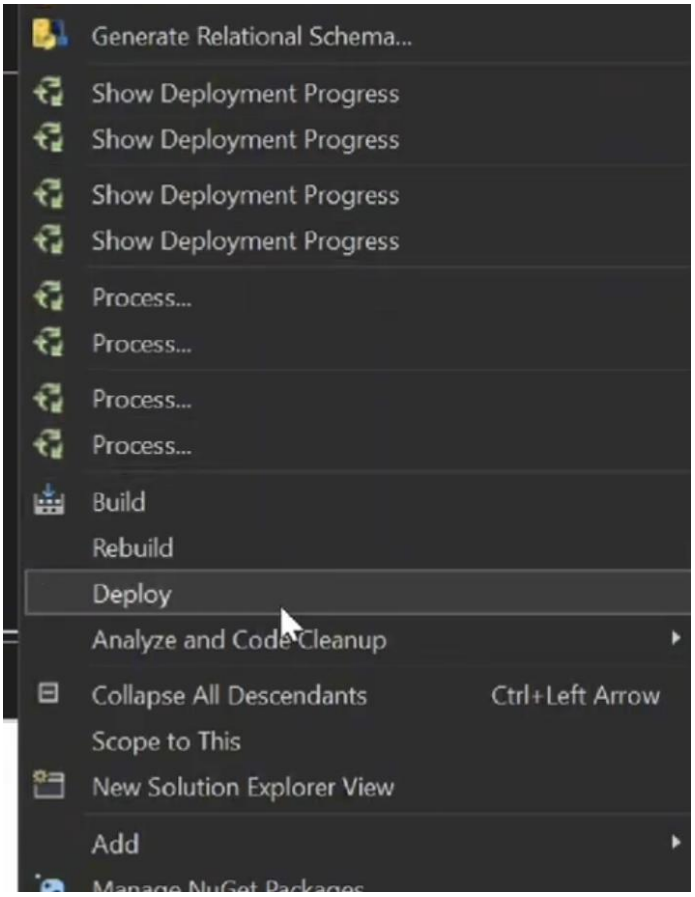

*Hình 124. Đổi Name Column của Year.*

- **Bước 9:** Qua tab Attribute Relationship để liên kết các mối quan hệ phân cấp và tiến hành phân cấp như hình.

*Hình 125. Attribute Relationship*

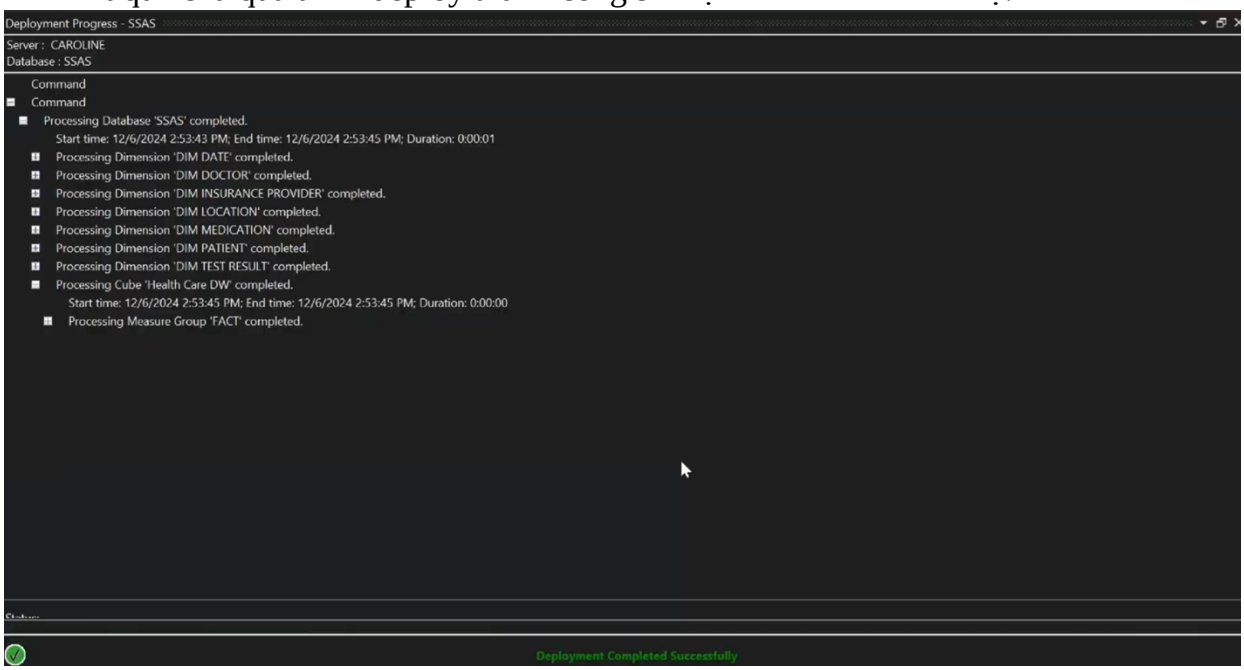
## 8. CHẠY PROJECT SSAS

- Nhấn chuột phải vào Project SSAS và chọn deploy.




*Hình 126. Deploy Project*

- Kết quả của quá trình deploy thành công sẽ hiện như hình minh hoạ.

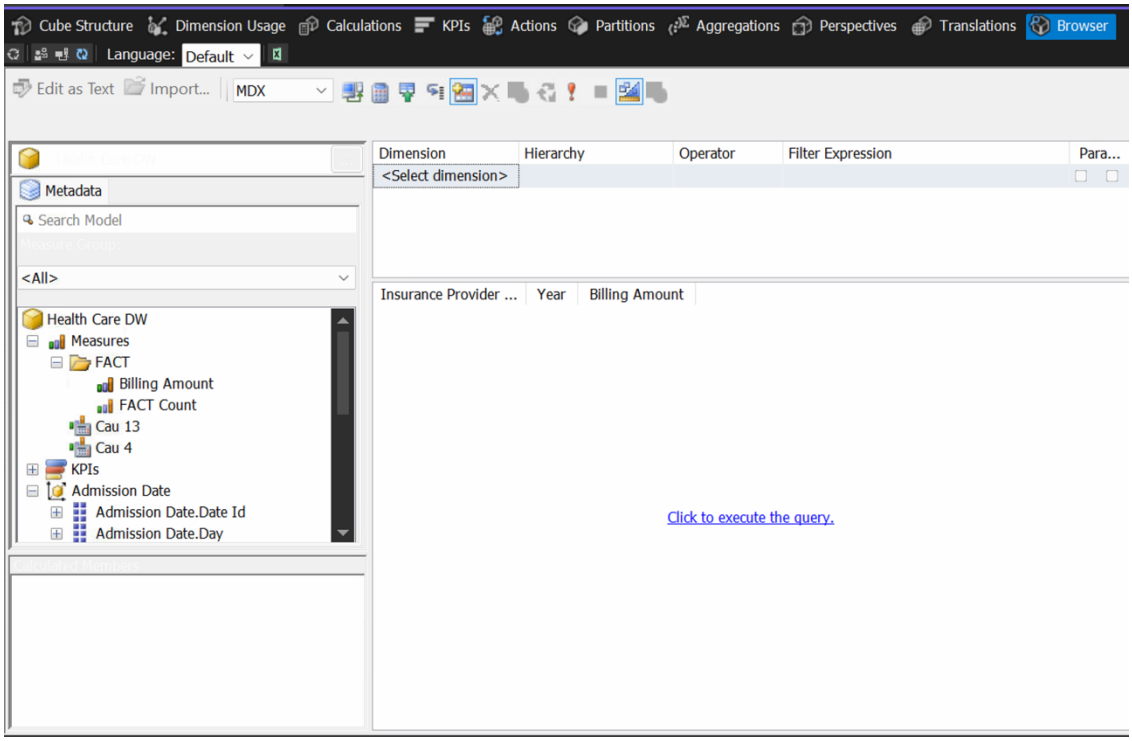



*Hình 127 . Kết quả quá trình Deploy*

## 9. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ SSAS TRÊN CÁC KHỐI CUBE

### 9.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm

- **Ý nghĩa câu truy vấn:** Dùng để xác định tổng chi phí điều trị mà bệnh nhân của từng nhà cung cấp bảo hiểm chi trả mỗi năm, giúp bệnh viện đánh giá sự phụ thuộc vào các nhà cung cấp bảo hiểm cụ thể và lập kế hoạch tài chính.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Kéo thuộc tính Insurance Provider Name từ bảng DIM_INSURANCE_PROVIDER và độ đo Billing Amount của bảng Fact.




*Hình 128. Câu truy vấn 1*

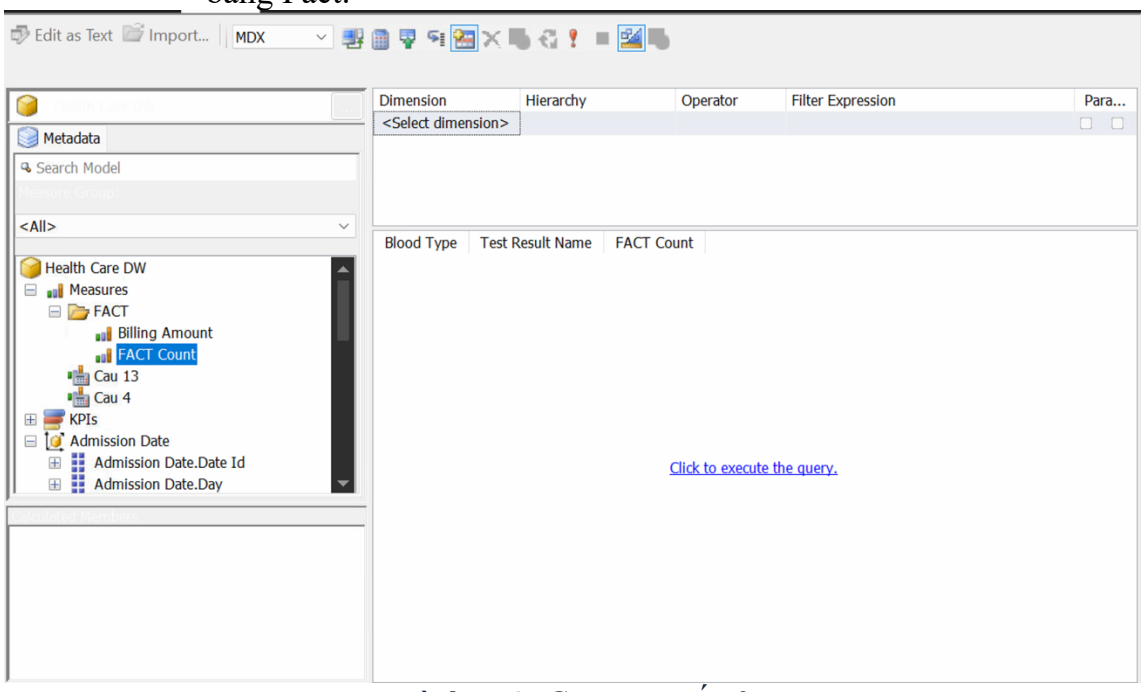
- **Kết quả của câu truy vấn:**

| Insurance Provider ... | Year | Billing Amount |
|---|---|---|
| Aetna | 2019 | 36781597.73... |
| Aetna | 2020 | 57538811.78... |
| Aetna | 2021 | 56450900.93... |
| Aetna | 2022 | 55721455.00... |
| Aetna | 2023 | 52529136.07... |
| Aetna | 2024 | 19841201.41... |
| Blue Cross | 2019 | 38262541.91... |
| Blue Cross | 2020 | 58272187.43... |
| Blue Cross | 2021 | 56734083.72... |
| Blue Cross | 2022 | 54172231.14... |
| Blue Cross | 2023 | 55889840.48... |
| Blue Cross | 2024 | 19923409.50... |
| Cigna | 2019 | 39838339.69... |
| Cigna | 2020 | 56463627.16... |
| Cigna | 2021 | 56308685.47... |
| Cigna | 2022 | 57019211.50... |
| Cigna | 2023 | 57808356.13... |
| Cigna | 2024 | 19701125.30... |

*Hình 129. Kết quả câu truy vấn 1*

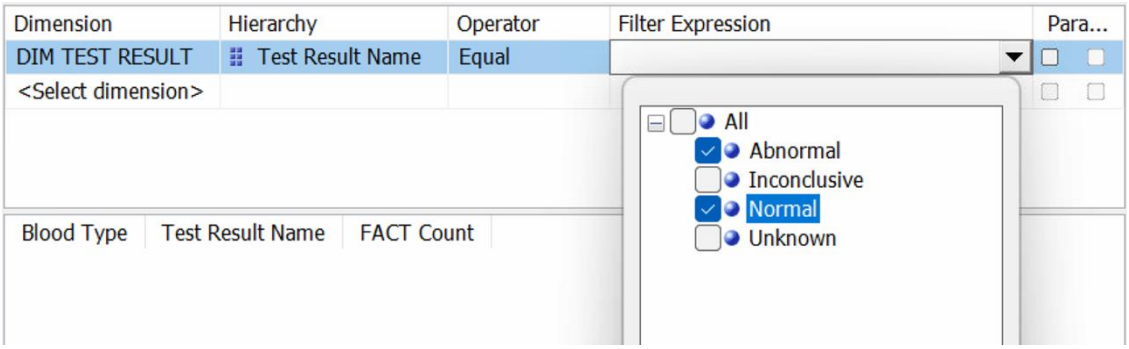
## 9.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu.

- **Ý nghĩa câu truy vấn:** Giúp phân tích tỷ lệ kết quả xét nghiệm "Normal" và "Abnormal" theo từng nhóm máu, hỗ trợ xác định các nhóm máu có nguy cơ bất thường cao.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Kéo thuộc tính Blood Type từ bảng DIM_PATIENT và Test Result Name từ bảng DIM_TEST_RESULT và độ đo FACT Count của bảng Fact.




*Hình 130. Câu truy vấn 2*

  - **Bước 2**: Tại cửa sổ lọc trường Filter Expression chọn Abnormal và Normal.




*Hình 131 . Filter Expression ( Abnormal ; Normal )*

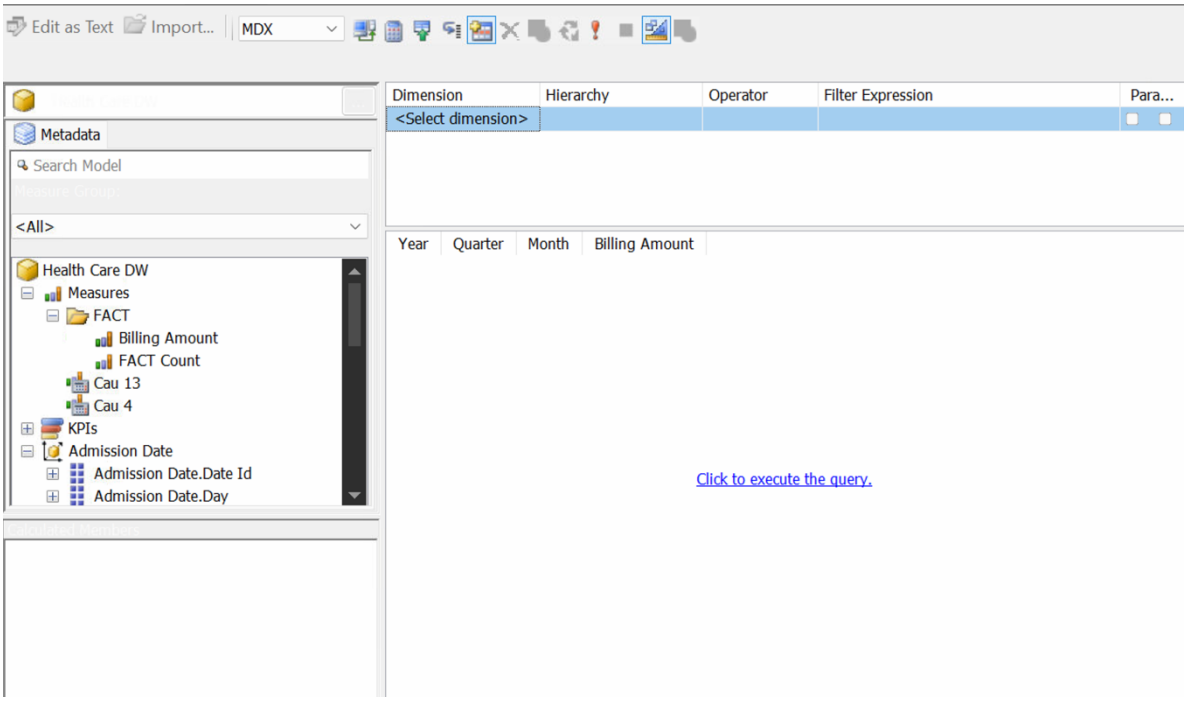
- **Kết quả của câu truy vấn:**

| Blood Type | Test Result Name | FACT Count |
|---|---|---|
| A- | Abnormal | 2336 |
| A- | Normal | 2335 |
| A+ | Abnormal | 2333 |
| A+ | Normal | 2293 |
| AB- | Abnormal | 2333 |
| AB- | Normal | 2362 |
| AB+ | Abnormal | 2308 |
| AB+ | Normal | 2314 |
| B- | Abnormal | 2348 |
| B- | Normal | 2328 |
| B+ | Abnormal | 2303 |
| B+ | Normal | 2276 |
| O- | Abnormal | 2319 |
| O- | Normal | 2301 |
| O+ | Abnormal | 2347 |
| O+ | Normal | 2308 |

*Hình 132. Kết quả câu truy vấn 2*

## 9.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm.

- **Ý nghĩa câu truy vấn:** Cung cấp dữ liệu chi tiết về tổng số tiền thanh toán theo tháng, quý, và năm, hỗ trợ bệnh viện lập ngân sách, dự báo doanh thu, và tối ưu hóa chiến lược tài chính trong tương lai.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Kéo thuộc tính Year, Quarter, Month từ bảng DIM_ADMISSION_DATE và độ đo Billing Amount của bảng Fact.




*Hình 133. Câu truy vấn 3*

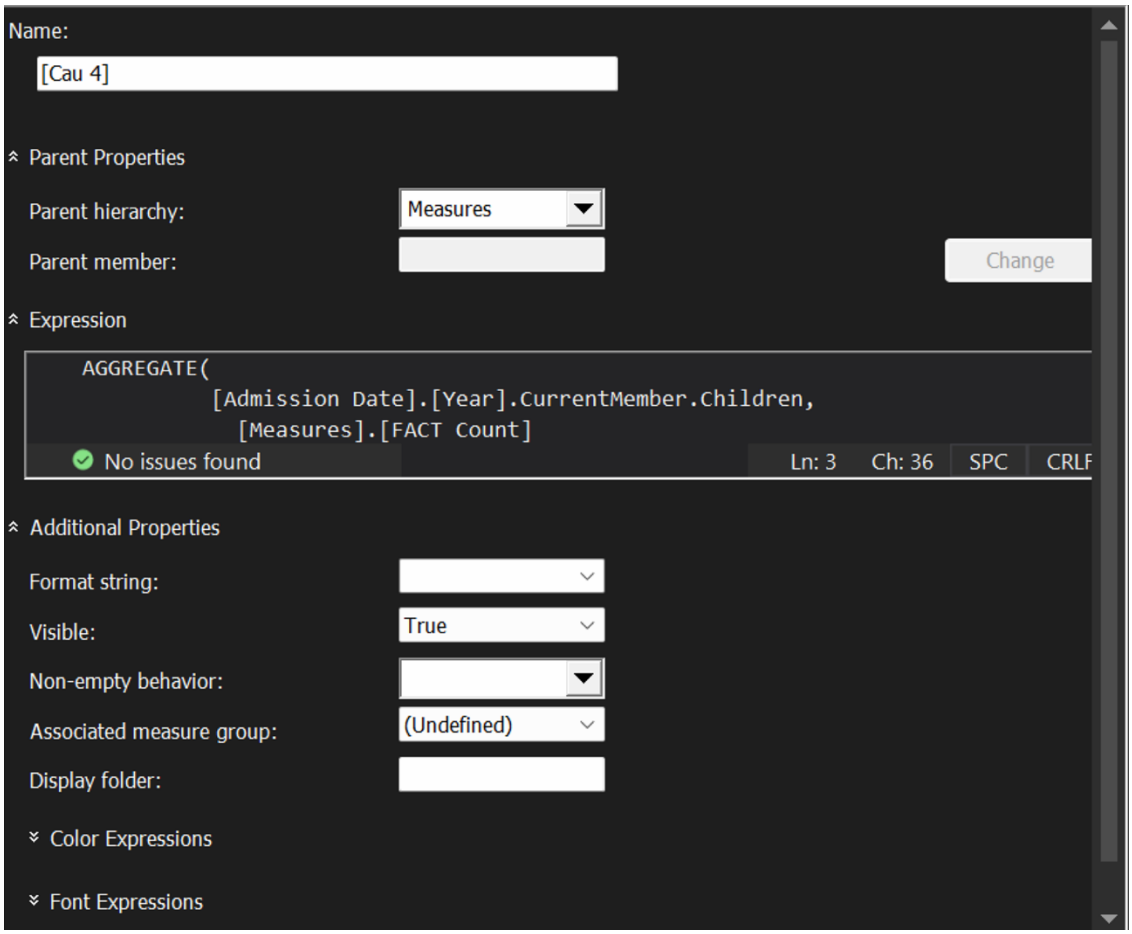
- **Kết quả của câu truy vấn:**

| Year | Quarter | Month | Billing Amount |
|---|---|---|---|
| 2019 | 2 | 5 | 17653145.13... |
| 2019 | 2 | 6 | 23885179.51... |
| 2019 | 3 | 7 | 24771559.19... |
| 2019 | 3 | 8 | 25397958.59... |
| 2019 | 3 | 9 | 23334079.87... |
| 2019 | 4 | 10 | 26276190.68... |
| 2019 | 4 | 11 | 24784507.52... |
| 2019 | 4 | 12 | 23746704.60... |
| 2020 | 1 | 1 | 24371975.18... |
| 2020 | 1 | 2 | 22324346.51... |
| 2020 | 1 | 3 | 23907855.08... |
| 2020 | 2 | 4 | 23129509.10... |
| 2020 | 2 | 5 | 24967566.12... |
| 2020 | 2 | 6 | 24303521.30... |
| 2020 | 3 | 7 | 24993721.59... |
| 2020 | 3 | 8 | 25817225.28... |
| 2020 | 3 | 9 | 22422187.28... |
| 2020 | 4 | 10 | 24103665.02... |

*Hình 134 . Kết quả câu truy vấn 3*

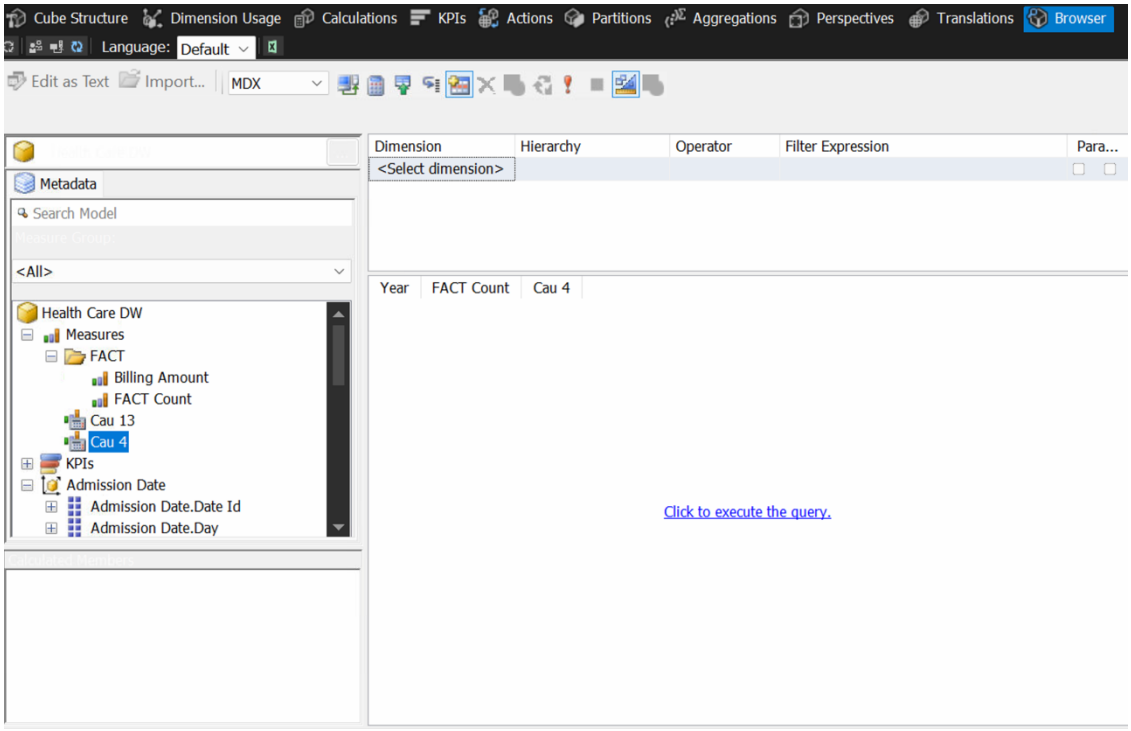
### 9.4. Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm

- **Ý nghĩa câu truy vấn:** Cung cấp số lượng bệnh nhân mắc bệnh "Cancer" theo từng năm, giúp các bệnh viện hiểu rõ xu hướng gia tăng hoặc giảm bớt số lượng bệnh nhân mắc bệnh ung thư qua các năm.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Tạo một Caculated Member mới để tính tổng số lượng bệnh nhân của các năm.



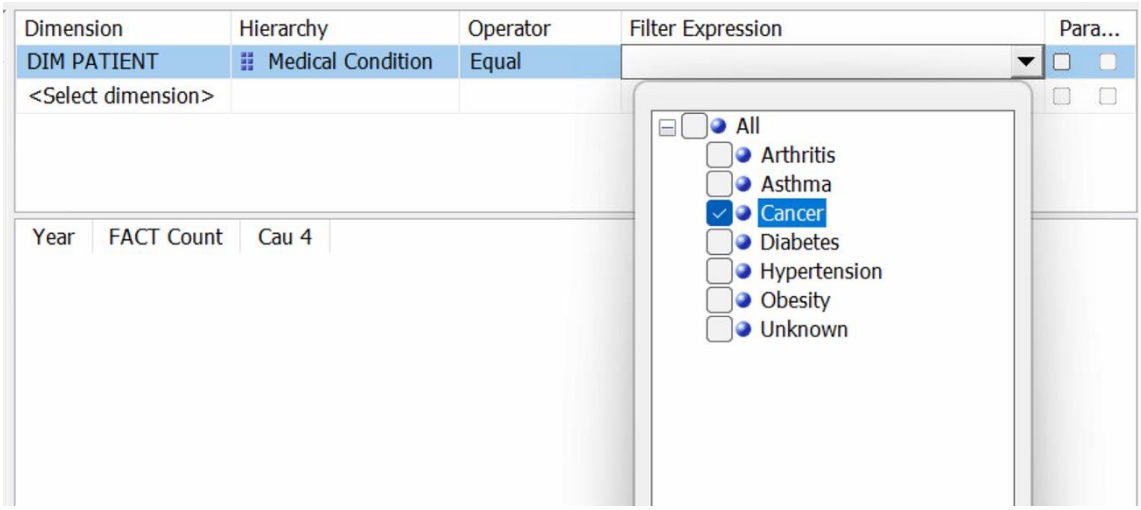

*Hình 135 .Tạo calculated Member cho  Câu truy vấn 4*

- **Bước 2**: Kéo thuộc tính Year từ bảng DIM_ADMISSION_DATE và độ đo FACT Count, và độ đo Cau 4 vừa tạo của bảng Fact.




*Hình 136 . Kéo thả các thuộc tính*

- **Bước 3**: Tại cửa sổ lọc trường Filter Expression chọn Cancer.




*Hình 137 . Filter Expression chọn Cancer.*

- **Bước 4:** Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode và xoá dòng tô đậm đã chọn.

```
SELECT NON EMPTY { [Measures].[FACT Count], [Measures].[Cau 4] } ON COLUMNS, NON EMPTY {
([Admission Date].[Year].[Year].ALLMEMBERS ) } DIMENSION PROPERTIES MEMBER_CAPTION,
MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( { [DIM PATIENT].[Medical Condition].&[Cancer] } ) ON
COLUMNS FROM [Health Care DW]) WHERE ( [DIM PATIENT].[Medical Condition].&[Cancer] ) CELL PROPERTIES
VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE,
FONT_FLAGS
```

| Year | FACT Count | Cau 4 |
|---|---|---|

Click to execute the query.

*Hình 138 . Design Mode sang Script Mode*

- **Kết quả của câu truy vấn:**

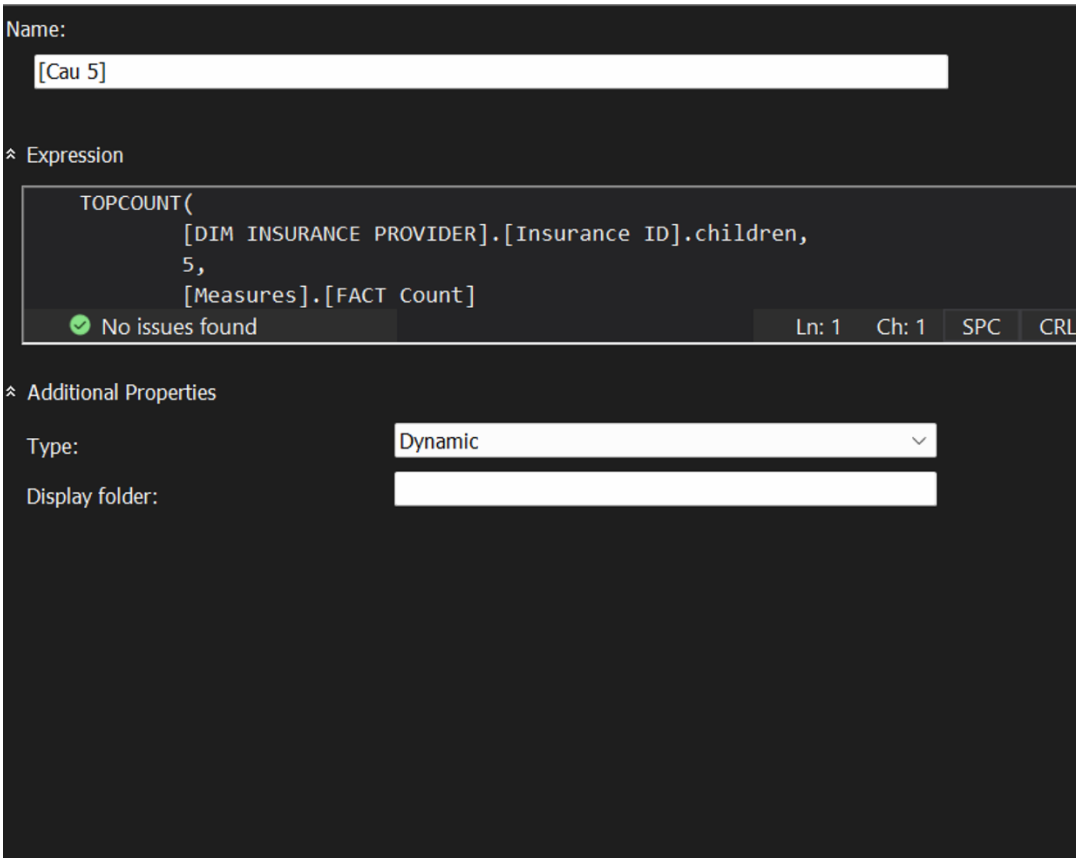
| Year | FACT Count | Cau 4 |
|---|---|---|
| (null) | 9227 | 9227 |
| 2019 | 1283 | (null) |
| 2020 | 1900 | (null) |
| 2021 | 1748 | (null) |
| 2022 | 1825 | (null) |
| 2023 | 1849 | (null) |
| 2024 | 622 | (null) |

*Hình 139. Kết quả câu truy vấn 4*

### 9.5. Đưa ra top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023.

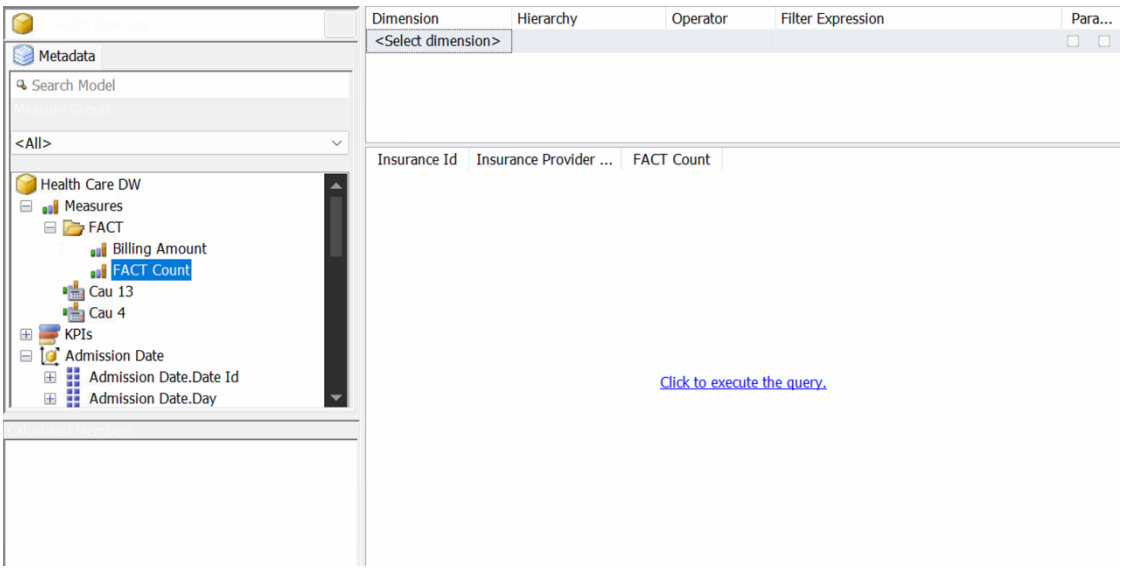
- **Ý nghĩa câu truy vấn:** Giúp bệnh viện xác định những nhà cung cấp bảo hiểm phổ biến nhất, từ đó tối ưu hóa mối quan hệ hợp tác và đảm bảo dịch vụ tốt hơn cho bệnh nhân có bảo hiểm từ các nhà cung cấp này.

- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Tạo một Named Set mới để lấy top 5 nhà cung cấp bảo hiểm




*Hình 140. Tạo Named Set mới cho câu truy vấn 5*

- **Bước 2**: Kéo thuộc tính Insurance ID và Insurance Provider Name từ bảng DIM_INSURANCE_PROVIDER và độ đo FACT Count.




*Hình 141. Kéo thả các thuộc tính*

- **Bước 3**: Tại cửa sổ lọc trường Filter Expression chọn năm 2023 và chọn bộ lọc named set Cau 5 vừa mới tạo.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| Admission Date | Admission Date.... | Equal | { 2023 } | |
| DIM INSURANCE ... | Insurance Id | In | Cau 5 | |
| <Select dimension> | | | | |

*Hình 142. Filter Expression chọn 2023 và bộ lọc named set Cau5*

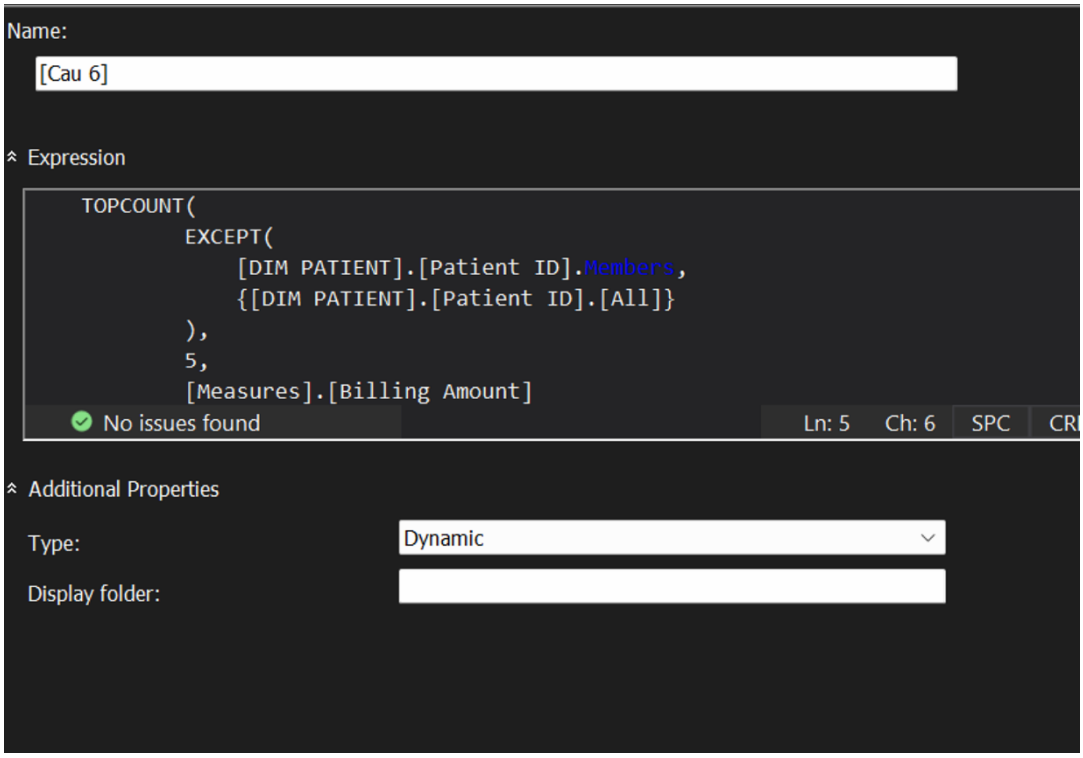
- **Kết quả của câu truy vấn:**

| Insurance Id | Insurance Provider ... | FACT Count |
|---|---|---|
| 21 | Aetna | 2086 |
| 22 | Blue Cross | 2166 |
| 23 | Cigna | 2268 |
| 24 | Medicare | 2258 |
| 25 | UnitedHealthcare | 2248 |

*Hình 143. Kết quả câu truy vấn 5*

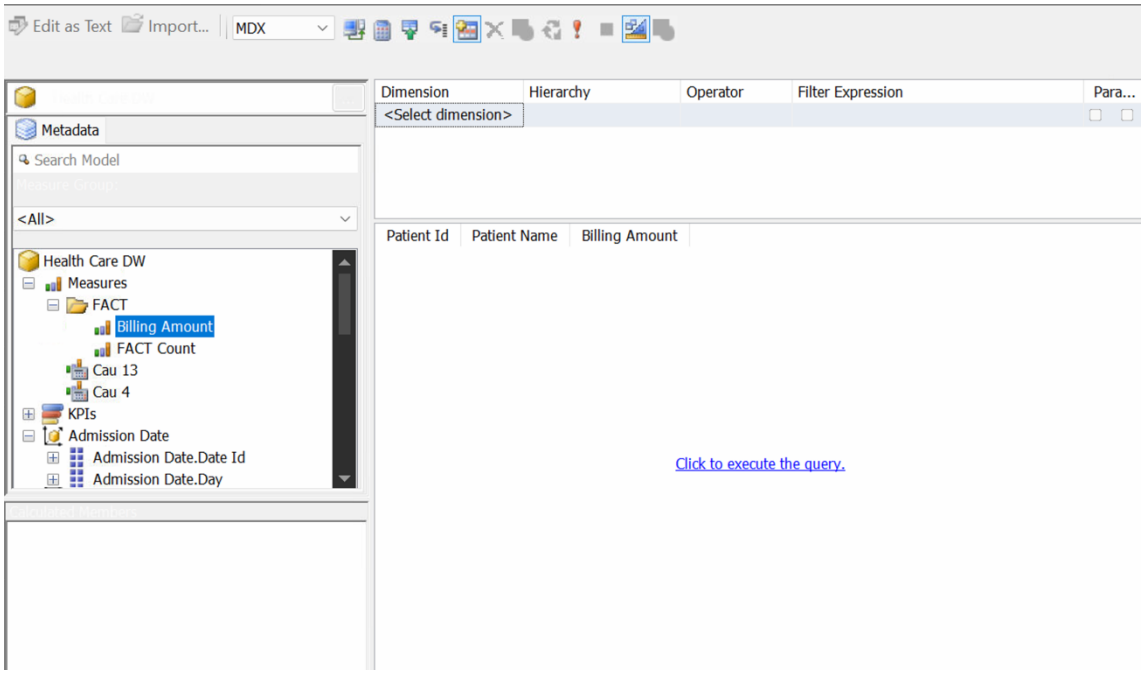
## 9.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc top 5

- **Ý nghĩa câu truy vấn:** Hỗ trợ bệnh viện theo dõi nhóm bệnh nhân chi tiêu cao trong điều trị, phục vụ cho việc nghiên cứu sức khỏe liên quan đến nhóm máu "O+".
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Tạo một Named Set mới để lấy top 5 viện phí của bệnh nhân




*Hình 144. Tạo Named Set mới cho câu truy vấn 6*

- **Bước 2**: Kéo thuộc tính Patient ID và Patient Name từ bảng DIM_PATIENT và độ đo Billing Amount từ bảng FACT.




*Hình 145. Kéo thả các thuộc tính*

- **Bước 3**: Tại cửa sổ lọc trường Filter Expression chọn nhóm máu O+ và chọn bộ lọc named set Cau 6 vừa mới tạo.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| DIM PATIENT | Blood Type | Equal | { O+ } | |
| DIM PATIENT | Patient Id | In | Cau 6 | |
| <Select dimension> | | | | |

*Hình 146 .Filter Expression chọn 0+ ; Named Set Cau 6*

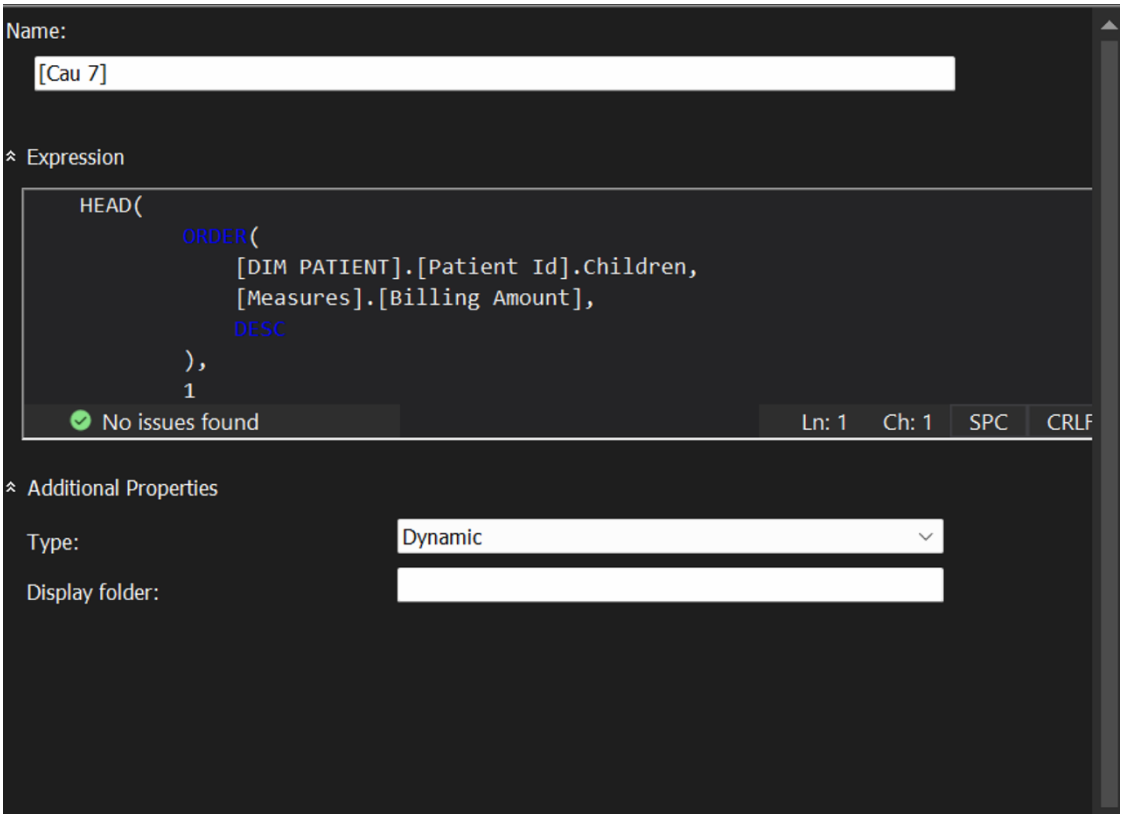
- **Kết quả của câu truy vấn:**

| Patient Id | Patient Name | Billing Amount |
|---|---|---|
| 11836 | davId ADams | 101187.3778... |
| 16115 | Erin HaNSen | 102863.9543... |
| 20778 | jeNNA SULL... | 101859.2016... |
| 32990 | MaTthew Ty... | 101525.0566... |
| 41194 | rOgEr sWA... | 101244.4517... |

*Hình 147. Kết quả câu truy vấn 6*

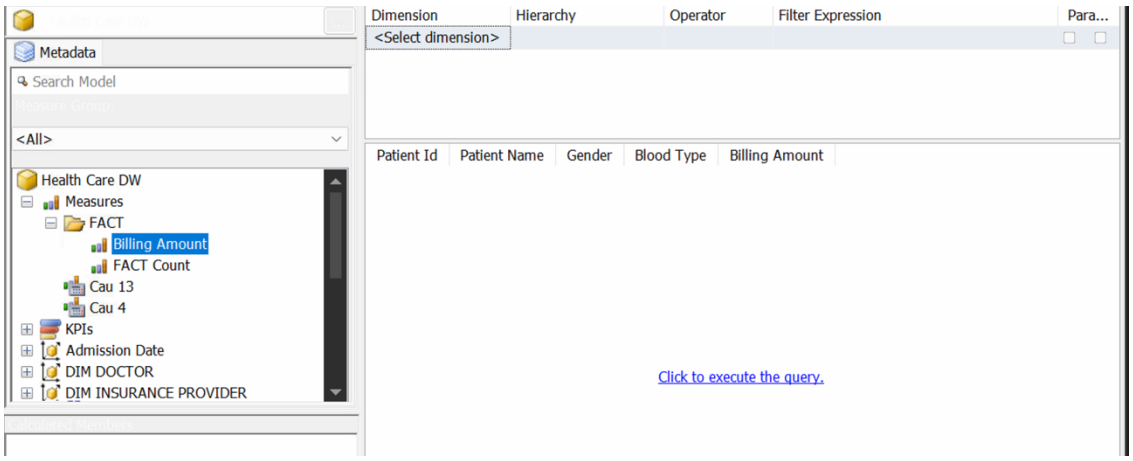
### 9.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024

- **Ý nghĩa câu truy vấn:** Giúp bệnh viện hiểu các trường hợp chi phí cao, từ đó tối ưu hóa quy trình quản lý chi phí và lập kế hoạch tài chính hiệu quả hơn.
- **Thực hiện quá trình phân tích:**
  - **Bước 1:** Tạo một Named Set mới để lấy bệnh nhân có viện phí cao nhất




*Hình 148. Tạo Named Set mới cho câu truy vấn 7*

- **Bước 2:** Kéo thuộc tính Patient ID, Patient Name, Gender, Blood Type từ bảng DIM_PATIENT và độ đo Billing Amount từ bảng FACT.




*Hình 149. Kéo thả các thuộc tính*

- **Bước 3:** Tại cửa sổ lọc trường Filter Expression chọn quý 2 của năm 2024 và chọn bộ lọc named set Cau 7 vừa mới tạo.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| DIM PATIENT | Patient Id | In | Cau 7 | |
| Admission Date | Admission Date.... | Equal | { 2 } | |
| <Select dimension> | | | | |

| Patient Id | Patient Name | Gender | Blood Type | Billing Amount |
|---|---|---|---|---|

*Hình 150. Filter Expression chọn quý 2 của năm 2024 và chọn bộ lọc named set Cau 7*

- **Kết quả của câu truy vấn:**

| Patient Id | Patient Name | Gender | Blood Type | Billing Amount |
|---|---|---|---|---|
| 36200 | mR. matTH... | Male | O+ | 99243.65035... |

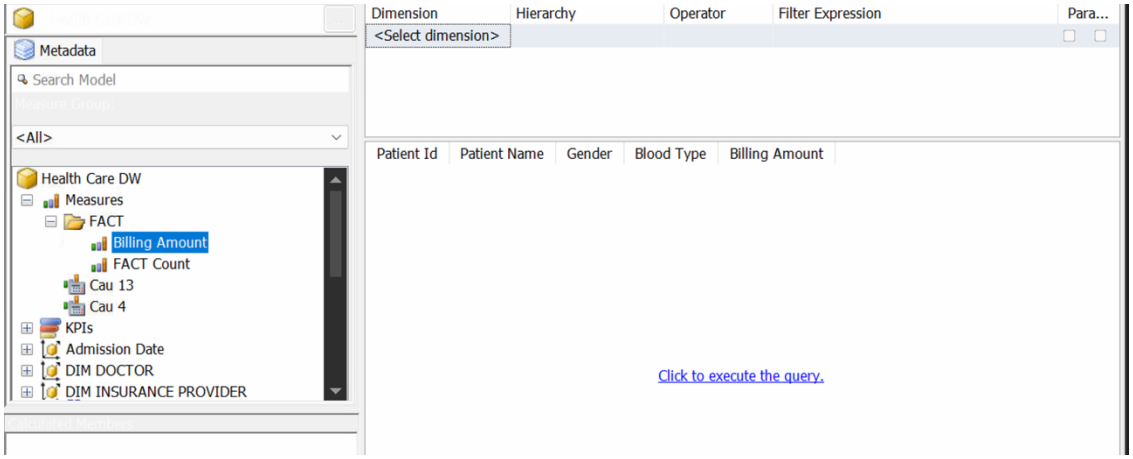
*Hình 151 . Kết quả câu truy vấn 7*

**9.8. Liệt kê các bệnh nhân(mã bệnh nhân, tên bệnh nhân, giới tính, nhóm máu) có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024.**

- **Ý nghĩa câu truy vấn:** Phân tích nhóm bệnh nhân ở mức chi tiêu trung bình, hỗ trợ xây dựng các gói chăm sóc phù hợp với khả năng tài chính của họ.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Tạo một Named Set mới để các bệnh nhân có viện phí từ 1000$ đến 10000$

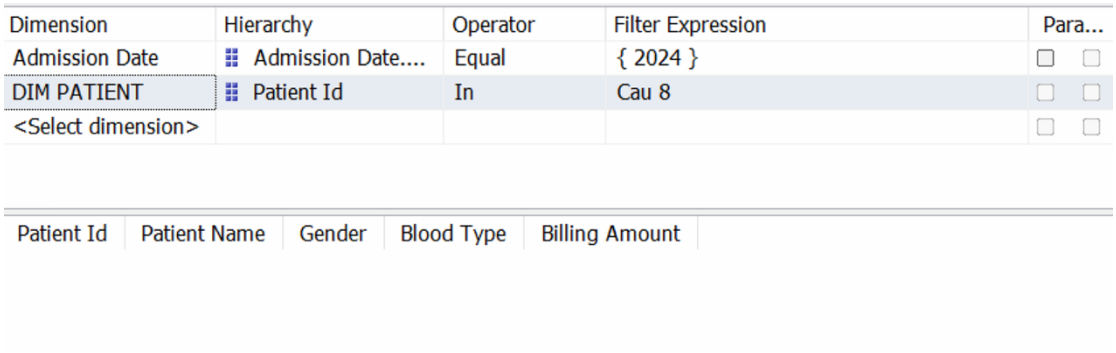
*Hình 152 . Tạo Named Set mới cho câu truy vấn 8*

- **Bước 2**: Kéo thuộc tính Patient ID, Patient Name, Gender, Blood Type từ bảng DIM_PATIENT và độ đo Billing Amount từ bảng FACT.




*Hình 153. Kéo thả các thuộc tính*

- **Bước 3**: Tại cửa sổ lọc trường Filter Expression chọn năm 2024 và chọn bộ lọc named set Cau 8 vừa mới tạo.




*Hình 154 . Filter Expression chọn năm 2024 và chọn bộ lọc named set Cau 8*

- **Kết quả của câu truy vấn:**

| Patient Id | Patient Name | Gender | Blood Type | Billing Amount |
|---|---|---|---|---|
| 115 | AarOn sALi... | Female | B+ | 9713.473812... |
| 121 | aaroN smIth | Female | B- | 8340.967112... |
| 133 | aAROn wEbeR | Female | AB+ | 8767.413833... |
| 229 | aDAm cOLLi... | Male | O+ | 4194.769773... |
| 376 | Adrian hOW... | Male | AB+ | 5361.394980... |
| 426 | ALAn BECk | Female | B- | 1927.733398... |
| 447 | alAN kelleR | Male | A- | 6125.982794... |
| 503 | AlbeRt SuLli... | Male | A+ | 7648.838950... |
| 532 | alEJandro A... | Female | B+ | 9071.639662... |
| 644 | AlExanDer F... | Female | AB+ | 4606.609431... |
| 658 | aleXAnDer ... | Female | B- | 9116.430635... |
| 681 | aleXAndeR ... | Male | A+ | 6988.694683... |
| 784 | AlEXAndra s... | Male | B- | 8355.431835... |
| 884 | AlEXiS mOore | Female | AB+ | 7162.991126... |
| 906 | aLeXIs Tyler | Female | O- | 3355.137901... |
| 1090 | ALIEN dENNIS | Female | A- | 7732.367661... |
| 1179 | alLiSon mC... | Female | AB+ | 7987.502383... |
| 1202 | alLIsoN SaN | Female | AB- | 6019.021389 |

*Hình 155 . Kết quả câu truy vấn 8*

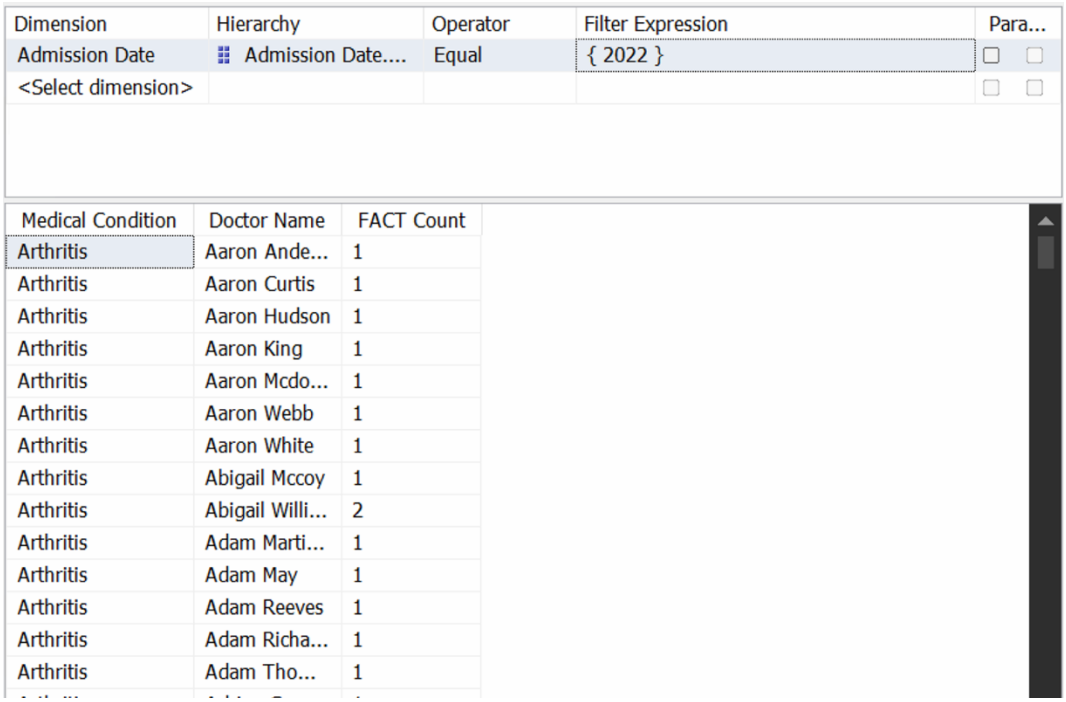
### 9.9. Trong mỗi loại bệnh, in ra bác sĩ  có số lượng bênh nhân lớn nhất trong năm 2022

- **Ý nghĩa câu truy vấn:** Xác định các bác sĩ nổi bật nhất cho từng loại bệnh, giúp bệnh viện phân bổ tài nguyên y tế hợp lý và cải thiện chất lượng điều trị.
- **Thực hiện quá trình phân tích:** Với câu truy vấn này không sử dụng công cụ BI và name set bình thường mà kết hợp công cụ BI và một số kiến thức liên quan đến ngôn ngữ MDX.
    - Dùng hàm Generate và Current member cho câu hỏi này:

```
GENERATE(
    [DIM PATIENT].[Medical Condition].children,
    TOPCOUNT(
        CROSSJOIN(
            {[DIM PATIENT].[Medical Condition].CurrentMember},
            [DIM DOCTOR].[Doctor Name].children
    ),
    1,
    ([Measures].[FACT Count])
)
```

)

- **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua Browser của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:

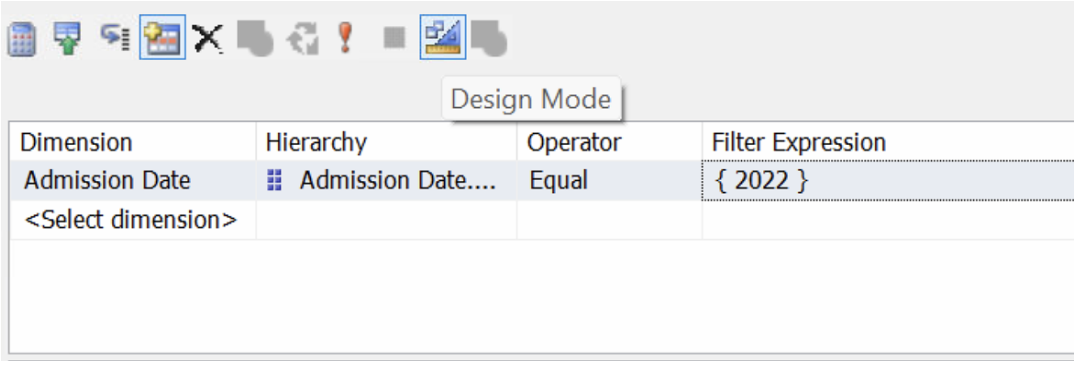


| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| Admission Date | Admission Date.... | Equal | { 2022 } | |
| <Select dimension> | | | | |

| Medical Condition | Doctor Name | FACT Count |
|---|---|---|
| Arthritis | Aaron Ande... | 1 |
| Arthritis | Aaron Curtis | 1 |
| Arthritis | Aaron Hudson | 1 |
| Arthritis | Aaron King | 1 |
| Arthritis | Aaron Mcdo... | 1 |
| Arthritis | Aaron Webb | 1 |
| Arthritis | Aaron White | 1 |
| Arthritis | Abigail Mccoy | 1 |
| Arthritis | Abigail Willi... | 2 |
| Arthritis | Adam Marti... | 1 |
| Arthritis | Adam May | 1 |
| Arthritis | Adam Reeves | 1 |
| Arthritis | Adam Richa... | 1 |
| Arthritis | Adam Tho... | 1 |

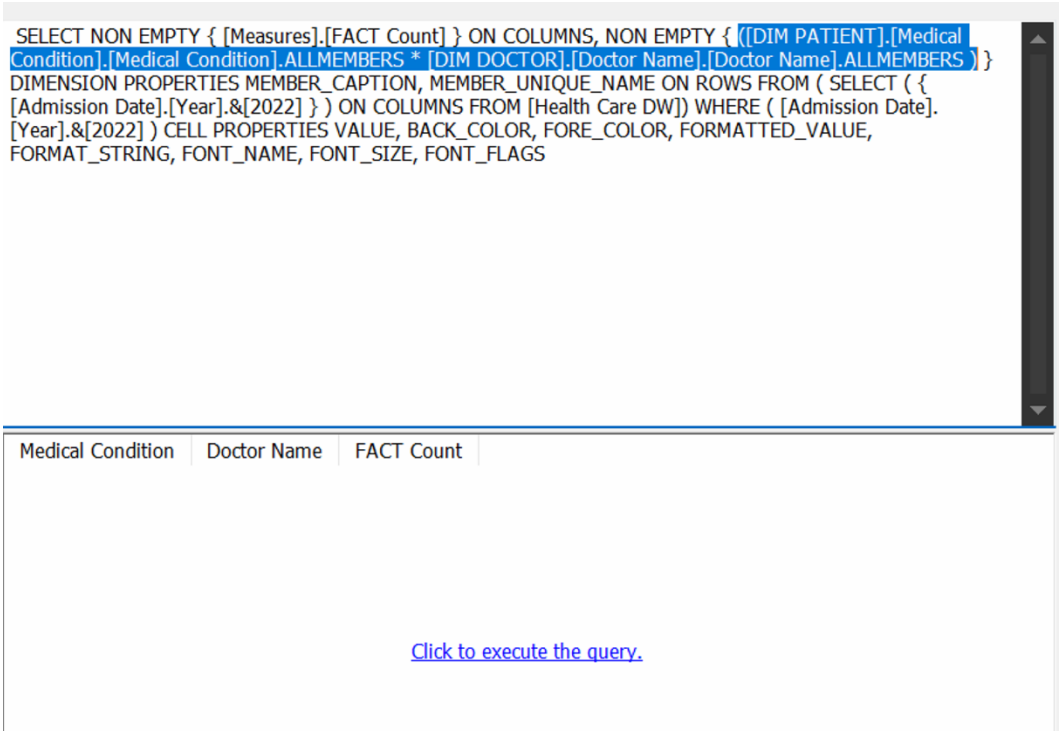
*Hình 156 .Tại browser của công cụ BI Kéo thả các thuộc tính cần thiết*

- **Bước 2:** Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode như hình minh họa:




*Hình 157. Chuyển từ Design Mode sang Script Mode*

- **Bước 3:** Xóa bỏ dòng tô xanh như hình minh họa :




*Hình 158 . Xóa bỏ bớt dòng code dư*

- **Bước 4:** Dán đoạn code trên vào dòng màu xanh vừa xóa và nhấn Execute Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS, NON EMPTY { GENERATE(
      [DIM PATIENT].[Medical Condition].children,
      TOPCOUNT(
            CROSSJOIN(
                {[DIM PATIENT].[Medical Condition].CurrentMember},
                [DIM DOCTOR].[Doctor Name].children
         ),
         1,
         ([Measures].[FACT Count])
      )
   )
 } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( {
[Admission Date].[Year].&[2022] } ) ON COLUMNS FROM [Health Care DW]) WHERE ( [Admission Date].
[Year].&[2022] ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 159. Dán code và Execute Query*

- **Kết quả của câu truy vấn:**

| Medical Condition | Doctor Name | FACT Count |
|---|---|---|
| Arthritis | Ryan Ward | 4 |
| Asthma | Amanda Wil... | 3 |
| Cancer | Carl Williams | 3 |
| Diabetes | David Smith | 4 |
| Hypertension | Lori Smith | 4 |
| Obesity | Amanda Go... | 3 |

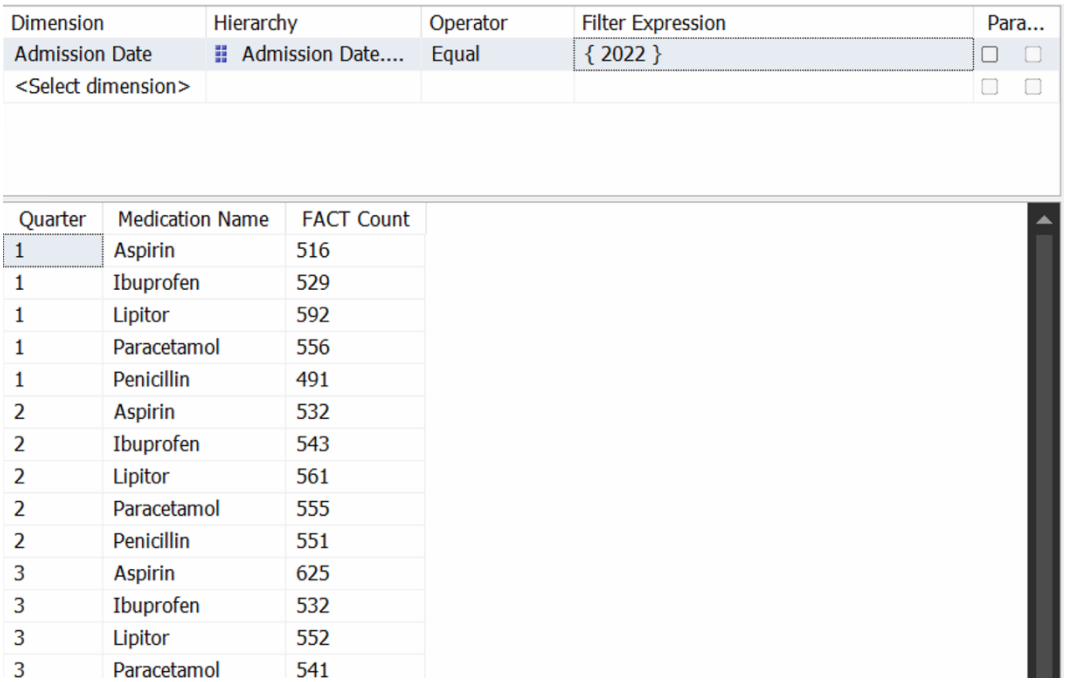
*Hình 160. Kết quả câu truy vấn 9*

**9.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.**

- **Ý nghĩa câu truy vấn:** Theo dõi xu hướng sử dụng thuốc để quản lý kho hiệu quả, tối ưu hóa mua sắm thuốc và tránh lãng phí.
- **Thực hiện quá trình phân tích:** Với câu truy vấn này không sử dụng công cụ BI và name set bình thường mà kết hợp công cụ BI và một số kiến thức liên quan đến ngôn ngữ MDX.
    - Dùng hàm Generate và Current member cho câu hỏi này:

```
GENERATE(
    DESCENDANTS( [Admission Date].[Quarter].children),
     TOPCOUNT(
        {[Admission Date].[Quarter].CurrentMember*
        [DIM MEDICATION].[Medication name].Children},
        3,
        [Measures].[Fact Count]
     )
  )
```

- **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua browser của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:

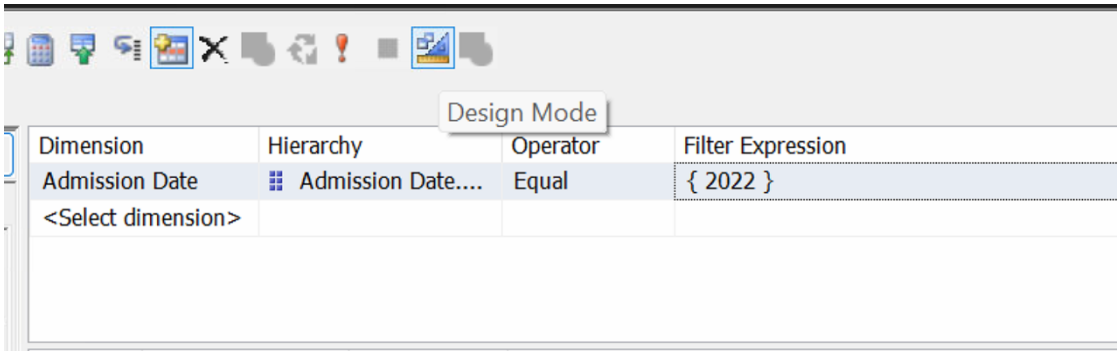


| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| Admission Date | Admission Date.... | Equal | { 2022 } | |
| <Select dimension> | | | | |

| Quarter | Medication Name | FACT Count |
|---|---|---|
| 1 | Aspirin | 516 |
| 1 | Ibuprofen | 529 |
| 1 | Lipitor | 592 |
| 1 | Paracetamol | 556 |
| 1 | Penicillin | 491 |
| 2 | Aspirin | 532 |
| 2 | Ibuprofen | 543 |
| 2 | Lipitor | 561 |
| 2 | Paracetamol | 555 |
| 2 | Penicillin | 551 |
| 3 | Aspirin | 625 |
| 3 | Ibuprofen | 532 |
| 3 | Lipitor | 552 |
| 3 | Paracetamol | 541 |

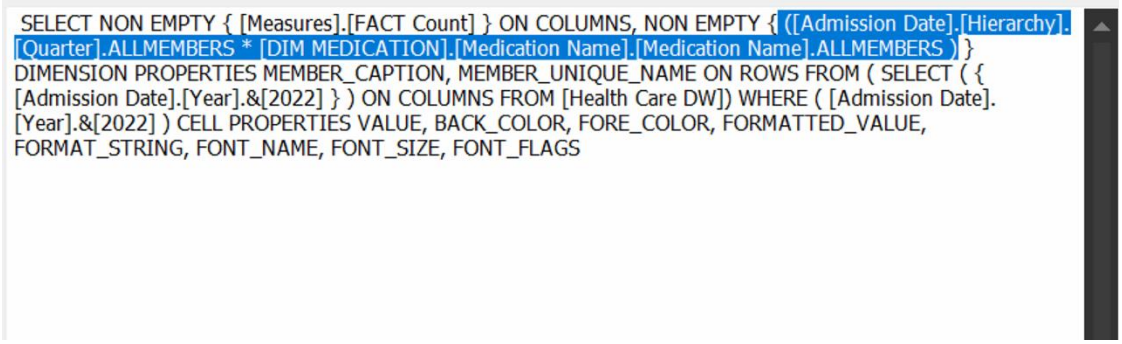
*Hình 161. Qua browser của công cụ BI kéo thả những thuộc tính cần thiết*

- **Bước 2:** Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode như hình minh họa:




*Hình 162. Chuyển từ Design Mode sang Script Mode*

- **Bước 3:** Xóa bỏ dòng tô xanh như hình minh họa :



```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS, NON EMPTY { ([Admission Date].[Hierarchy].
[Quarter].ALLMEMBERS * [DIM MEDICATION].[Medication Name].[Medication Name].ALLMEMBERS ) }
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( {
[Admission Date].[Year].&[2022] } ) ON COLUMNS FROM [Health Care DW]) WHERE ( [Admission Date].
[Year].&[2022] ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 163. Xóa bỏ bớt dòng code dư*

- **Bước 4:** Dán đoạn code trên vào dòng màu xanh vừa xóa và nhấn Execute Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS, NON EMPTY {GENERATE(

    DESCENDANTS( [Admission Date].[Quarter].children),
     TOPCOUNT(
        {[Admission Date].[Quarter].CurrentMember*
        [DIM MEDICATION].[Medication name].Children},
        3,
        [Measures].[Fact Count]
     )
  )
} DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( {
[Admission Date].[Year].&[2022] } ) ON COLUMNS FROM [Health Care DW]) WHERE ( [Admission Date].
[Year].&[2022] ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE,
FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 164. Dán code và Execute Query*

- **Kết quả của câu truy vấn:**

| Quarter | Medication Name | FACT Count |
|---|---|---|
| 1 | Lipitor | 592 |
| 1 | Paracetamol | 556 |
| 1 | Ibuprofen | 529 |
| 2 | Lipitor | 561 |
| 2 | Paracetamol | 555 |
| 2 | Penicillin | 551 |
| 3 | Aspirin | 625 |
| 3 | Penicillin | 575 |
| 3 | Lipitor | 552 |
| 4 | Ibuprofen | 605 |
| 4 | Aspirin | 557 |
| 4 | Paracetamol | 556 |

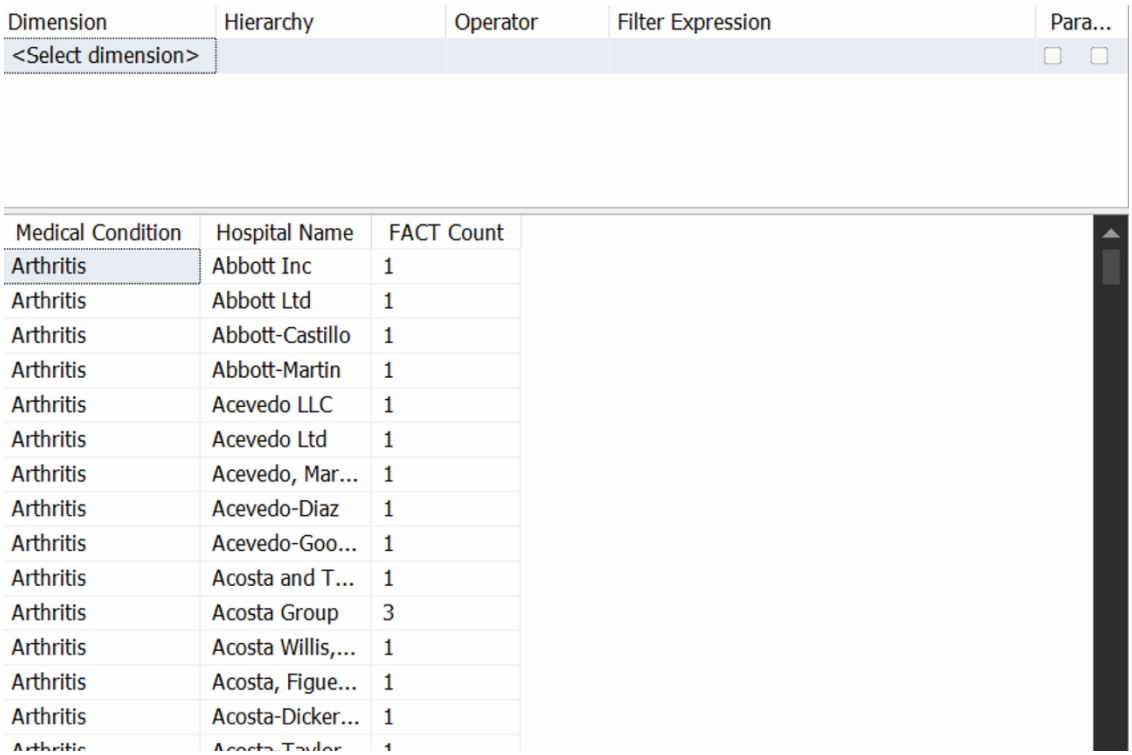
*Hình 165 . Kết quả câu truy vấn 10*

## 9.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể

- **Ý nghĩa câu truy vấn:** Giúp bệnh viện hiểu rõ năng lực điều trị của mình đối với các loại bệnh khác nhau, đồng thời hỗ trợ định vị thương hiệu trong lĩnh vực y tế.
- **Thực hiện quá trình phân tích:** Với câu truy vấn này không sử dụng công cụ BI và name set bình thường mà kết hợp công cụ BI và một số kiến thức liên quan đến ngôn ngữ MDX.
    - Dùng hàm Generate và Current member cho câu hỏi này:

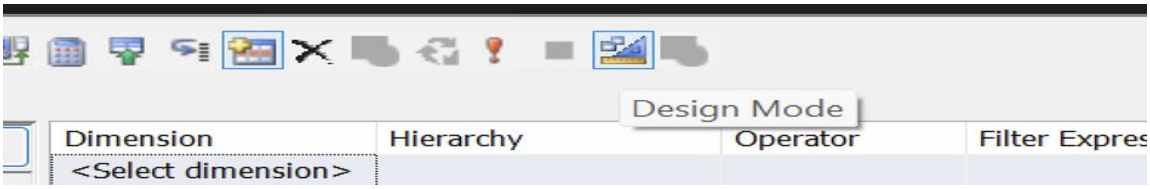
```
GENERATE(
    [DIM PATIENT].[Medical Condition].Children,
    TOPCOUNT(
        [DIM LOCATION].[Hospital Name].Children,
        1,
        [Measures].[FACT Count]
    ) *
    {[DIM PATIENT].[Medical Condition].CurrentMember}
)
```

- **Bước 1**: Copy hàm trên vào bộ nhớ tạm. Qua brower của công cụ BI kéo thả những thuộc tính cần thiết cho câu truy vấn như hình sau đây:




*Hình 166 . Qua browser của công cụ BI kéo thả những thuộc tính cần thiết*

- **Bước 2:** Chuyển từ Design Mode sang Script Mode bằng cách nhấn chuột vào biểu tượng Design Mode như hình minh họa:




*Hình 167. Chuyển từ Design Mode sang Script Mode*

- **Bước 3:** Xóa bỏ dòng tô xanh như hình minh họa :

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS, NON EMPTY { ([DIM PATIENT].[Medical
Condition].[Medical Condition].ALLMEMBERS * [DIM LOCATION].[Hospital Name].[Hospital Name].ALLMEMBERS
) } DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Health Care DW]
CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
```

| Medical Condition | Hospital Name | FACT Count |
|---|---|---|

*Hình 168. Chuyển từ Design Mode sang Script Mode*

- **Bước 4:** Dán đoạn code trên vào dòng màu xanh vừa xóa và nhấn Execute Query:

```
SELECT NON EMPTY { [Measures].[FACT Count] } ON COLUMNS, NON EMPTY {GENERATE(
        [DIM PATIENT].[Medical Condition].Children,
        TOPCOUNT(
           [DIM LOCATION].[Hospital Name].Children,
           1,
           [Measures].[FACT Count]
        ) *
        {[DIM PATIENT].[Medical Condition].CurrentMember}
     )
} DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Health Care DW]
CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING,
FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 169 . Dán code và Execute Query*

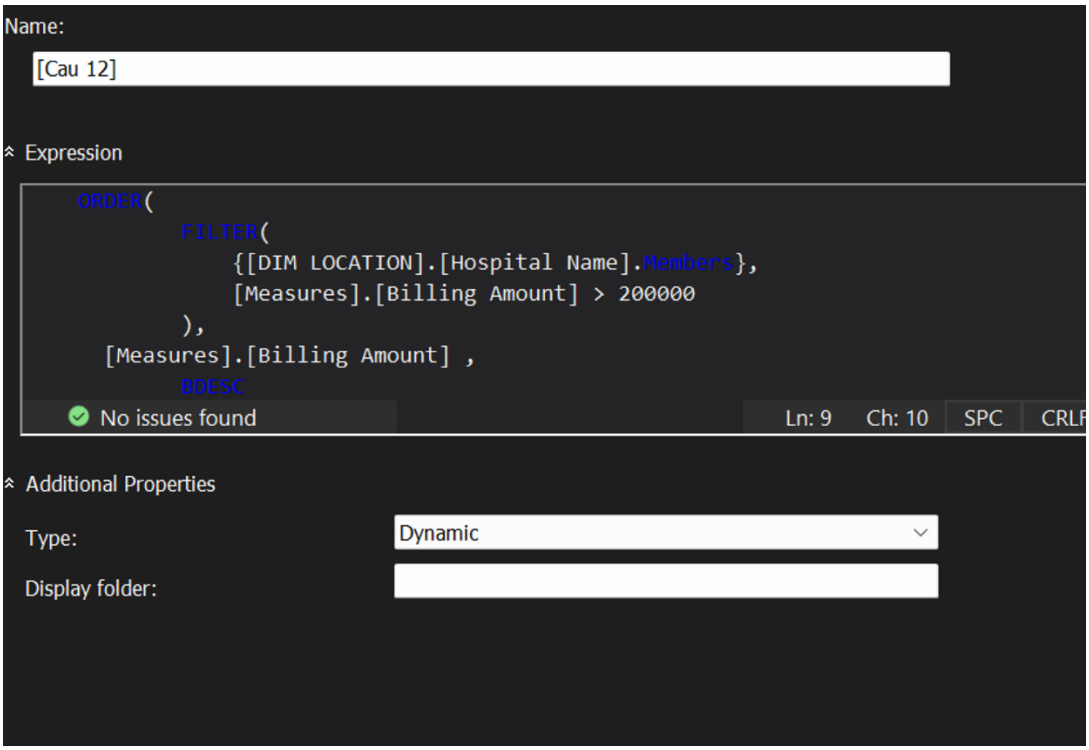
- **Kết quả của câu truy vấn:**

| Hospital Name | Medical Condition | FACT Count |
|---|---|---|
| Smith PLC | Arthritis | 12 |
| Johnson Inc | Asthma | 9 |
| Johnson Inc | Cancer | 10 |
| LLC Smith | Diabetes | 11 |
| LLC Smith | Hypertension | 11 |
| Ltd Smith | Obesity | 11 |

*Hình 170. Kết quả câu truy vấn 11*

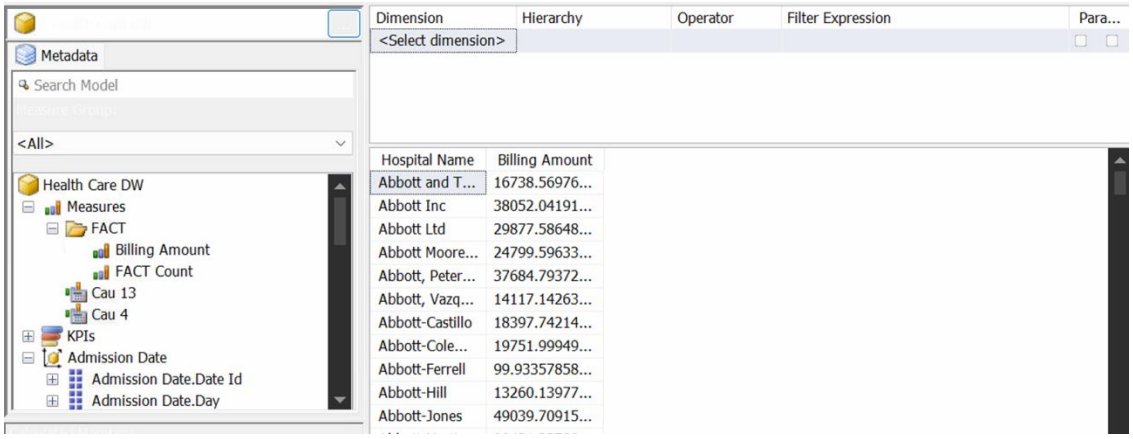
**9.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $100000, sau đó sắp xếp theo tổng chi phí giảm dần.**

- **Ý nghĩa câu truy vấn:** Đánh giá hiệu quả tài chính của các bệnh viện trong điều trị bệnh tiểu đường, phục vụ cho việc quản lý và cải thiện chất lượng dịch vụ.
- **Thực hiện quá trình phân tích:**
  - **Bước 1**: Tạo một Named Set mới để lấy bệnh nhân có viện phí cao nhất




*Hình 171. Tạo một Named Set mới Câu 12*

- **Bước 2**: Kéo thuộc tính Patient ID, Patient Name, Gender, Blood Type từ bảng DIM_PATIENT và độ đo Billing Amount từ bảng FACT.




*Hình 172. Kéo thả các thuộc tính cần thiết*

- **Bước 3**: Tại cửa sổ lọc trường Filter Expression chọn Diabetes và chọn bộ lọc named set Cau 12 vừa mới tạo.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| DIM PATIENT | Medical Condition | Equal | { Diabetes } | |
| DIM LOCATION | Hospital Name | In | Cau 12 | |
| <Select dimension> | | | | |

| Hospital Name | Billing Amount |
|---|---|

*Hình 173. Filter Expression chọn Diabetes và chọn bộ lọc named set Cau 12*

- **Bước 4**: Chuyển từ Design Mode sang Script Mode và xoá các đoạn code sau:

```
SELECT NON EMPTY { [Measures].[Billing Amount] } ON COLUMNS, NON EMPTY { ([DIM LOCATION].[Hospital
Name].[Hospital Name].ALLMEMBERS ) } DIMENSION PROPERTIES MEMBER_CAPTION,
MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( { [DIM PATIENT].[Medical Condition].&[Diabetes] } ) ON
COLUMNS FROM ( SELECT ( [Cau 12] ) ON COLUMNS FROM [Health Care DW])) WHERE ( [DIM PATIENT].
[Medical Condition].&[Diabetes] ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR,
FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 174 175.*

```
SELECT NON EMPTY { [Measures].[Billing Amount] } ON COLUMNS, NON EMPTY { } DIMENSION PROPERTIES
MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM ( SELECT ( { [DIM PATIENT].[Medical
Condition].&[Diabetes] } ) ON COLUMNS FROM ( SELECT ( [Cau 12] ) ON COLUMNS FROM [Health Care DW]))
WHERE ( [DIM PATIENT].[Medical Condition].&[Diabetes] ) CELL PROPERTIES VALUE, BACK_COLOR,
FORE_COLOR, FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 175 . Chuyển từ Design Mode sang Script Mode*

- **Bước 5**: Thêm Named Set Cau 12 vào để hoành chỉnh câu truy vấn

```
SELECT NON EMPTY { [Measures].[Billing Amount] } ON COLUMNS, NON EMPTY {[Cau 12] } DIMENSION
PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME ON ROWS FROM [Health Care DW] WHERE ( [DIM
PATIENT].[Medical Condition].&[Diabetes] ) CELL PROPERTIES VALUE, BACK_COLOR, FORE_COLOR,
FORMATTED_VALUE, FORMAT_STRING, FONT_NAME, FONT_SIZE, FONT_FLAGS
```

*Hình 176 . Thêm Named Set Cau 12 vào*

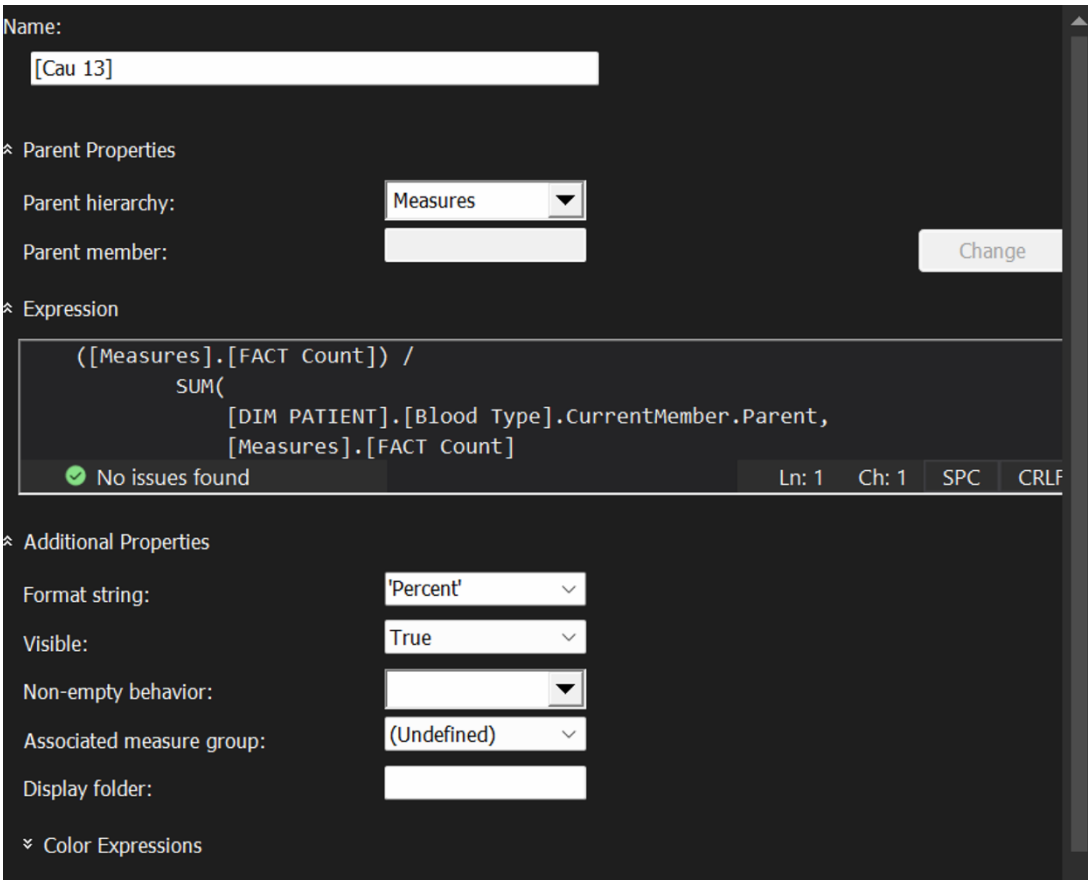
- **Kết quả của câu truy vấn:**

| Hospital Name | Billing Amount |
|---|---|
| (null) | 238539725.4... |
| PLC Brown | 304480.7953... |
| Group Smith | 291522.6529... |
| Smith Ltd | 247134.6458... |
| Rodriguez Inc | 243599.1289... |
| Smith Inc | 227727.1305... |
| LLC Smith | 218458.2625... |
| Martin Inc | 217248.2123... |

*Hình 177 . Kết quả câu truy vấn 12*

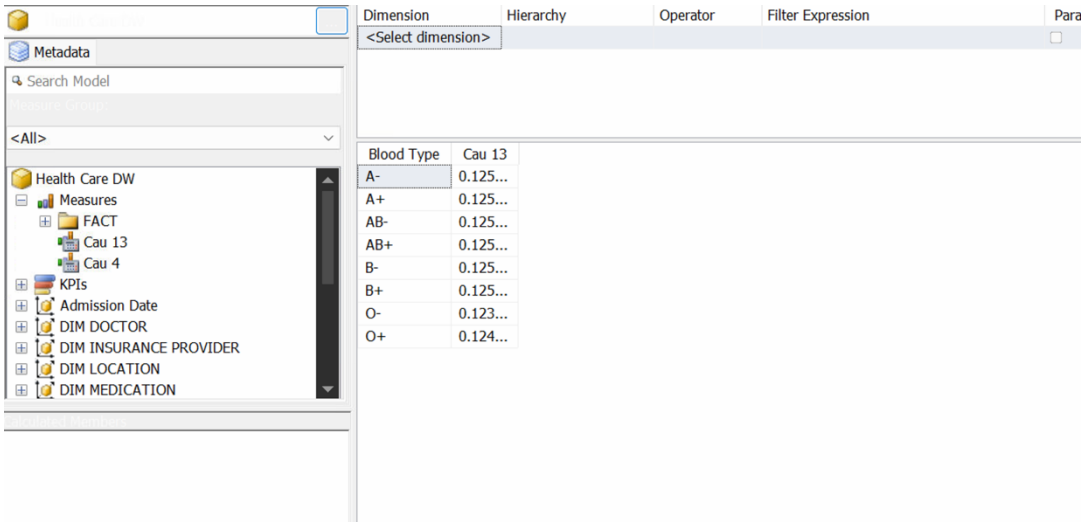
### 9.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.

- **Ý nghĩa câu truy vấn:** Phân bố nhóm máu giúp bệnh viện chuẩn bị nguồn máu phù hợp cho điều trị và phẫu thuật.
- **Thực hiện quá trình phân tích:**
  - **Bước 1:** Tạo độ đo tính toán phần trăm bệnh nhân của từng nhóm máu.



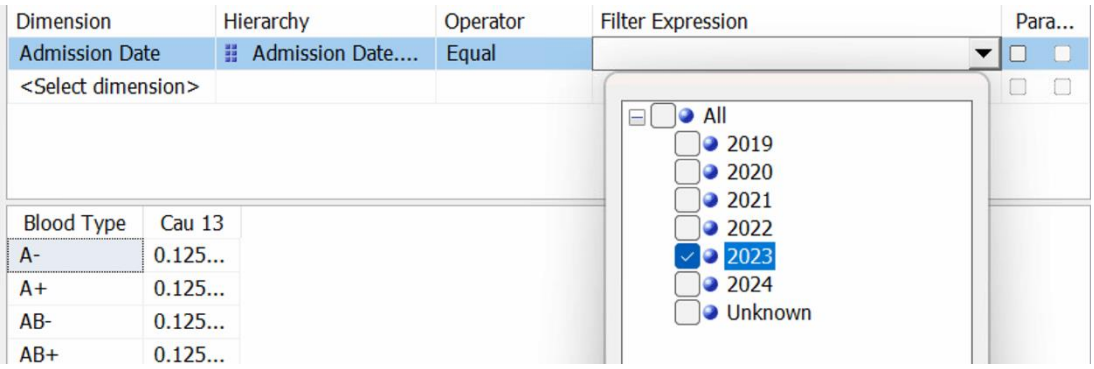

*Hình 178. Tạo độ đo tính toán*

- **Bước 2:** Kéo thuộc tính Blood type và độ đo Cau 13 sang cửa sổ thực thi truy vấn.




*Hình 179. Kéo thả các thuộc tính cần thiết*

- **Bước 3:** Lọc theo năm 2023.




*Hình 180 . Lọc theo năm 2023*

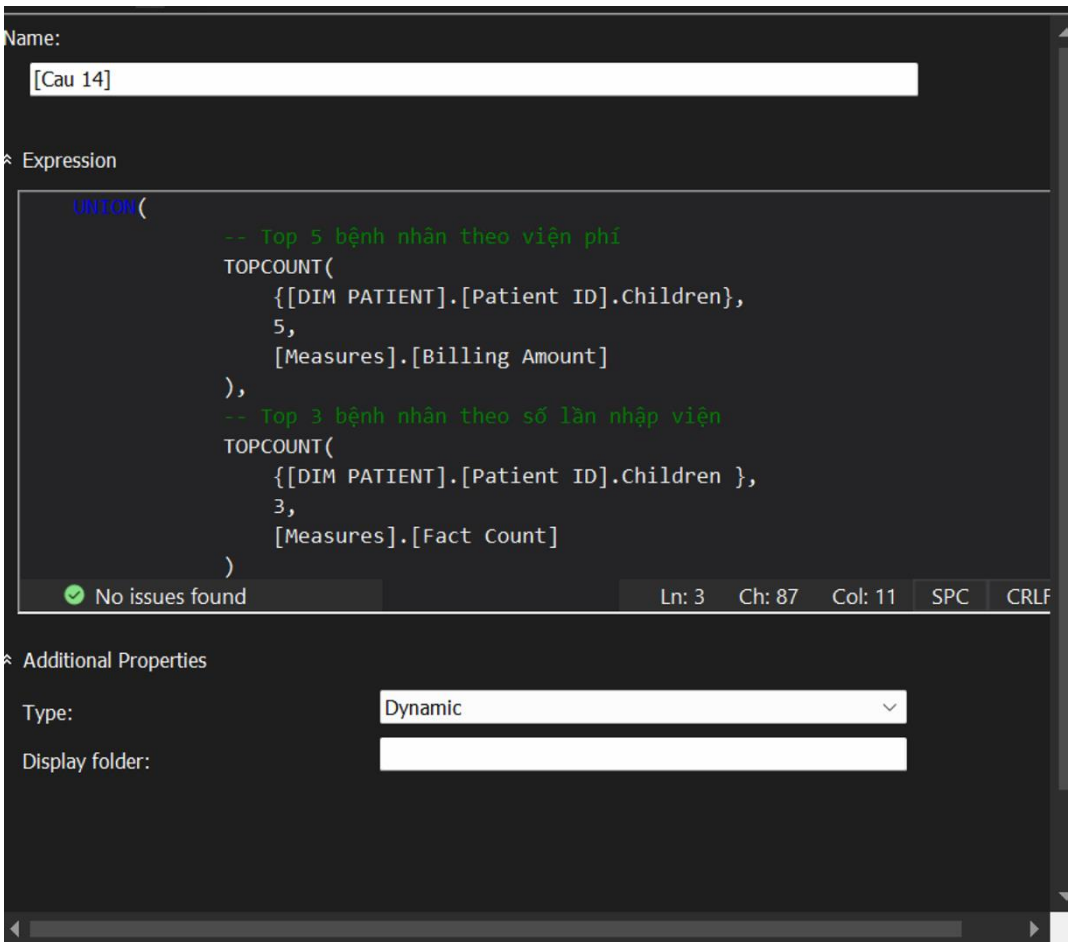
- **Kết quả của câu truy vấn:**

| Blood Type | Cau 13 |
|---|---|
| A- | 0.128423725739162 |
| A+ | 0.12588427353528 |
| AB- | 0.126428441864683 |
| AB+ | 0.129239978233267 |
| B- | 0.11908216941774 |
| B+ | 0.123072737166697 |
| O- | 0.122256484672592 |
| O+ | 0.125612189370579 |

*Hình 181. Kết quả câu truy vấn 13*

## 9.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất / top 3 số lần nhập viện nhiều nhất.

- **Ý nghĩa câu truy vấn:** Theo dõi các trường hợp đặc biệt để nghiên cứu nguyên nhân và cải thiện trải nghiệm bệnh nhân.
- **Thực hiện quá trình phân tích:**
  - **Bước 1:** Tạo Named Set lấy top 5 và top 3 bệnh nhân..




*Hình 182. Tạo một Named Set mới*

- **Bước 2:** Kéo thuộc tính Patient Id và Patient Name và độ đo Billing Amount, FACT Count sang cửa sổ thực thi truy vấn.

| Patient Id | Patient Name | Billing Amount | FACT Count |
|---|---|---|---|
| 1 | aAroN ADaMS | 26052.10640... | 1 |
| 2 | aAROn aguI... | 27087.56055... | 1 |
| 3 | AArOn And... | 39804.65862... | 1 |
| 4 | aaRON And... | 16846.41579... | 1 |
| 5 | AAron ArC... | 21204.15437... | 2 |
| 6 | AaROn bAkEr | 6826.677362... | 1 |
| 7 | AaROn bAkEr | 10135.88544... | 1 |
| 8 | AARON bAl... | 29740.96019... | 1 |
| 9 | aaROn bAR... | 4783.444339... | 1 |
| 10 | aaROn bauT... | 49827.51976... | 1 |
| 11 | aarOn bEnN... | 12740.88681... | 1 |
| 12 | AArON BeN... | 5411.611013... | 1 |
| 13 | AarOn bRA... | 50132.99578... | 1 |
| 14 | AarOn bRA... | 16342.36373... | 1 |
| 15 | AaRon bRA... | 4436.082623... | 1 |
| 16 | aARON BRock | 36225.56927... | 1 |
| 17 | aarOn BrookS | 24756.15064... | 1 |

*Hình 183. Kéo thả các thuộc tính cần thiết*

- **Bước 3:** Lọc theo Named Set Cau 14.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| DIM PATIENT | Patient Id | In | Cau 14 | |
| <Select dimension> | | | | |

*Hình 184. Lọc theo Named Set Cau 14.*

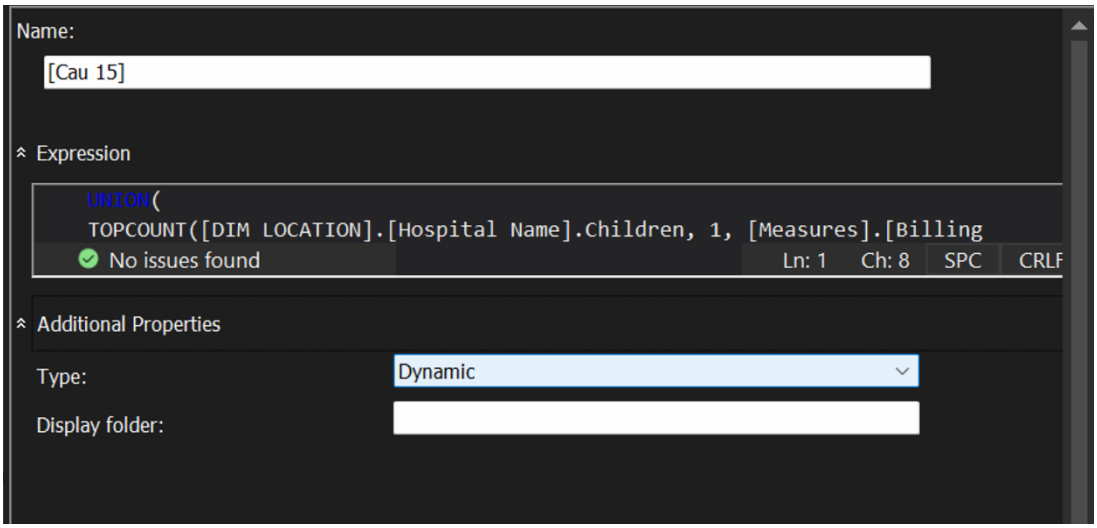
- **Kết quả của câu truy vấn:**

| Patient Id | Patient Name | Billing Amount | FACT Count |
|---|---|---|---|
| 5 | AAron ArC... | 21204.15437... | 2 |
| 24 | aarOn cArR | 58085.92365... | 2 |
| 32 | aaRON DaltoN | 4140.518953... | 2 |
| 16115 | Erin HaNSen | 102863.9543... | 2 |
| 16569 | gaBrIEL HA... | 102561.9858... | 2 |
| 25959 | kARen klInE | 104746.0647... | 2 |
| 29591 | LAuRie hOod | 104340.0737... | 2 |
| 37840 | pAmElA No... | 103951.9362... | 2 |

*Hình 185. Kết quả câu truy vấn 14*

**9.15. Cho biết bệnh viện có doanh thu  cao nhất và bệnh viện có doanh thu thấp nhất.**

- **Ý nghĩa câu truy vấn:** Phân bố nhóm máu giúp bệnh viện chuẩn bị nguồn máu phù hợp cho điều trị và phẫu thuật.
- **Thực hiện quá trình phân tích:**
  - **Bước 1:** Tạo Named Set lấy top 1 bệnh viện có doanh thu cao nhất và thấp nhất.




*Hình 186. Tạo một Named Set mới*

- **Bước 2:** Kéo thuộc tính Hospital Name và độ đo Billing Amount sang cửa sổ thực thi truy vấn.

| Hospital Name | Billing Amount |
|---|---|
| Abbott and T... | 16738.56976... |
| Abbott Inc | 38052.04191... |
| Abbott Ltd | 29877.58648... |
| Abbott Moore... | 24799.59633... |
| Abbott, Peter... | 37684.79372... |
| Abbott, Vazq... | 14117.14263... |
| Abbott-Castillo | 18397.74214... |
| Abbott-Cole... | 19751.99949... |
| Abbott-Ferrell | 99.93357858... |
| Abbott-Hill | 13260.13977... |
| Abbott-Jones | 49039.70915... |
| Abbott-Martin | 90431.22788... |
| Abbott-Rios | 35798.34654... |

*Hình 187. Kéo thả các thuộc tính cần thiết*

- **Bước 3:** Lọc theo Named Set Cau 15.

| Dimension | Hierarchy | Operator | Filter Expression | Para... |
|---|---|---|---|---|
| DIM LOCATION | Hospital Name | In | Cau 15 | |
| <Select dimension> | | | | |

*Hình 188. Lọc theo Named Set Cau 15*

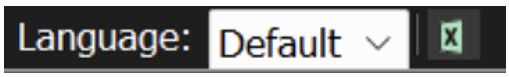
- **Kết quả của câu truy vấn:**

| Hospital Name | Billing Amount |
|---|---|
| Johnson PLC | 1084202.69296492 |
| Medina and Elliott Stewart, | -2633.23716258996 |

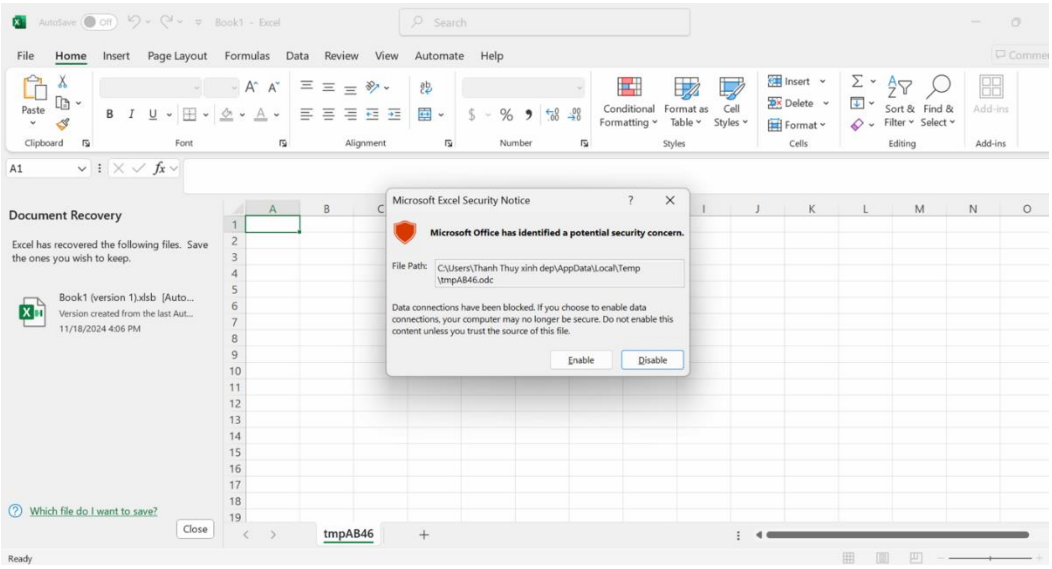
*Hình 189. Kết quả câu truy vấn 15*

## 10. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG CÔNG CỤ PIVOT EXCEL

- Tại màn hình làm việc của công cụ SSAS chọn biểu tượng Excel:




*Hình 190. Chọn biểu tượng Excel*

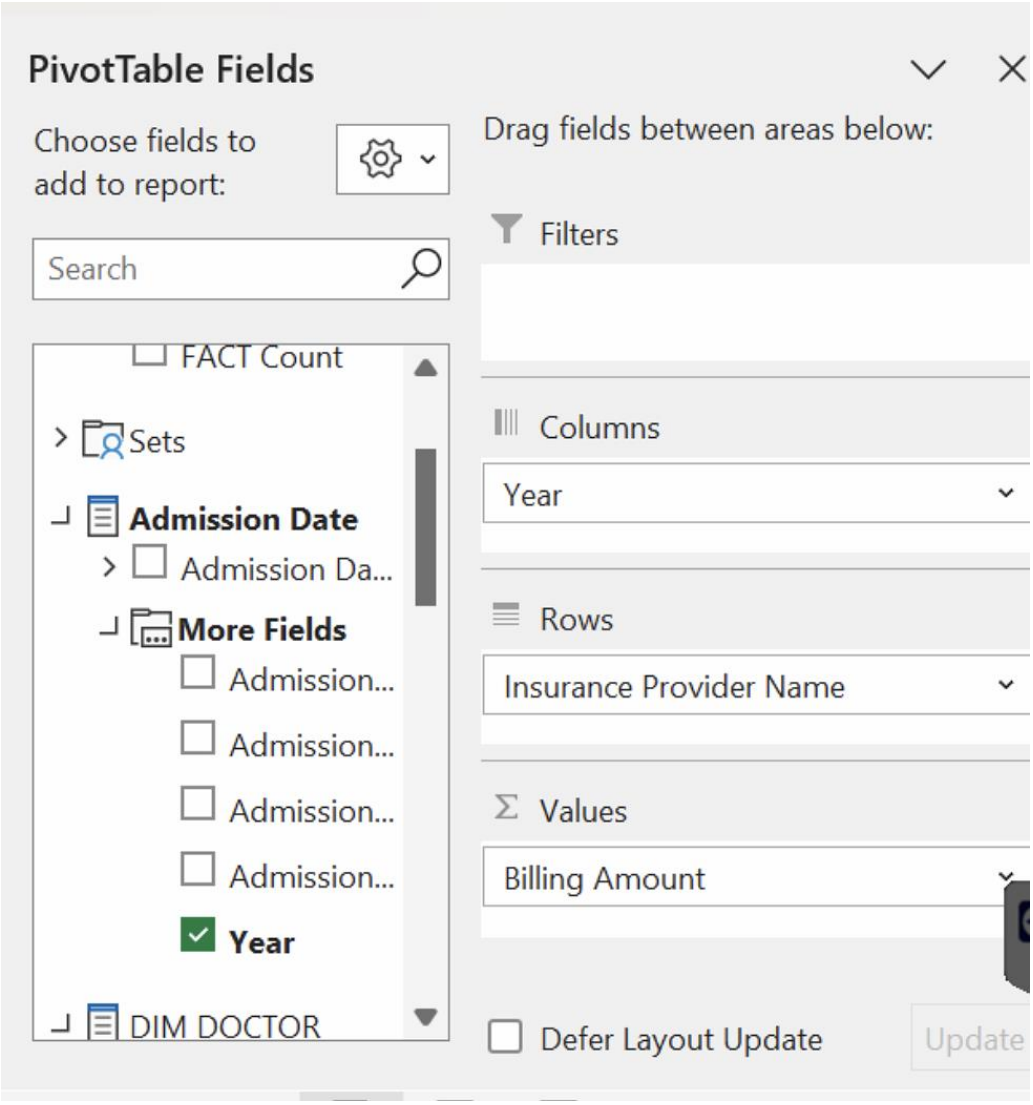



*Hình 191 . Chọn biểu tượng Excel*

- Nhấn bật để mở công cụ Pivot Excel

### 10.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm.

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Insurance Provider Name vào Rows và thuộc tính Year vào Columns, kéo độ đo Billing Amount vào cột Values.



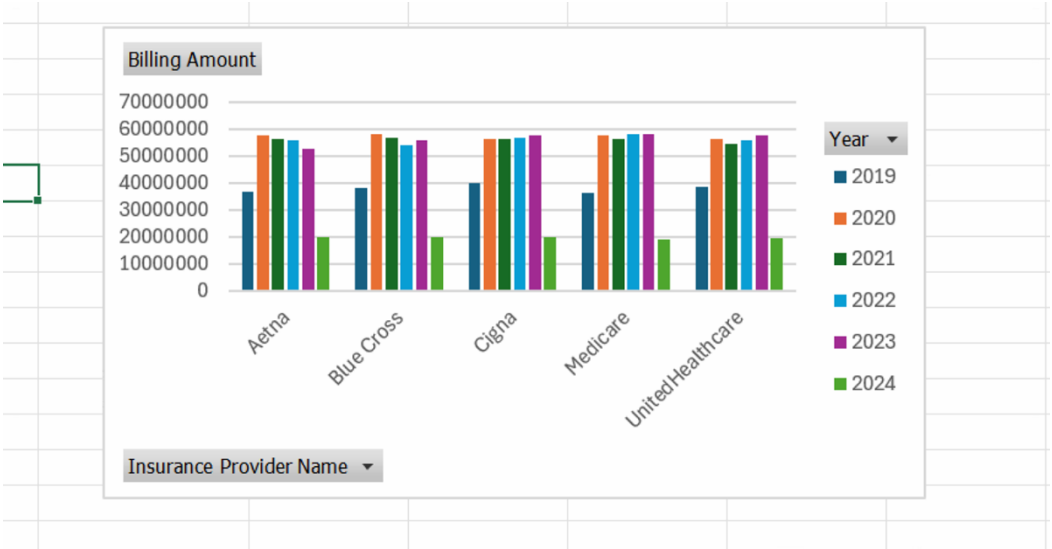

*Hình 192 . Chọn các thuộc tính cần thiết*

- **Kết quả câu truy vấn dạng Matrix:**

| Billing Amount | Column Labels | | | | | | |
|---|---|---|---|---|---|---|---|
| Row Labels | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | Grand Total |
| Aetna | 36781597.74 | 57538811.78 | 56450900.93 | 55721455 | 52529136.07 | 19841201.42 | 278863102.9 |
| Blue Cross | 38262541.91 | 58272187.43 | 56734083.73 | 54172231.15 | 55889840.49 | 19923409.5 | 283254294.2 |
| Cigna | 39838339.69 | 56463627.16 | 56308685.47 | 57019211.5 | 57808356.13 | 19701125.3 | 287139345.3 |
| Medicare | 36443188.96 | 57772568.16 | 56328362.55 | 58186485.98 | 58010543.11 | 18979609.36 | 285720758.1 |
| UnitedHealthcare | 38523656.84 | 56504394.66 | 54333848.6 | 55982953.92 | 57739377.42 | 19370311.39 | 282454542.8 |
| **Grand Total** | **189849325.1** | **286551589.2** | **280155881.3** | **281082337.6** | **281977253.2** | **97815656.98** | **1417432043** |

*Hình 193 . Kết quả câu truy vấn 1 dạng Matrix*

- **Kết quả câu truy vấn dạng Chart:**

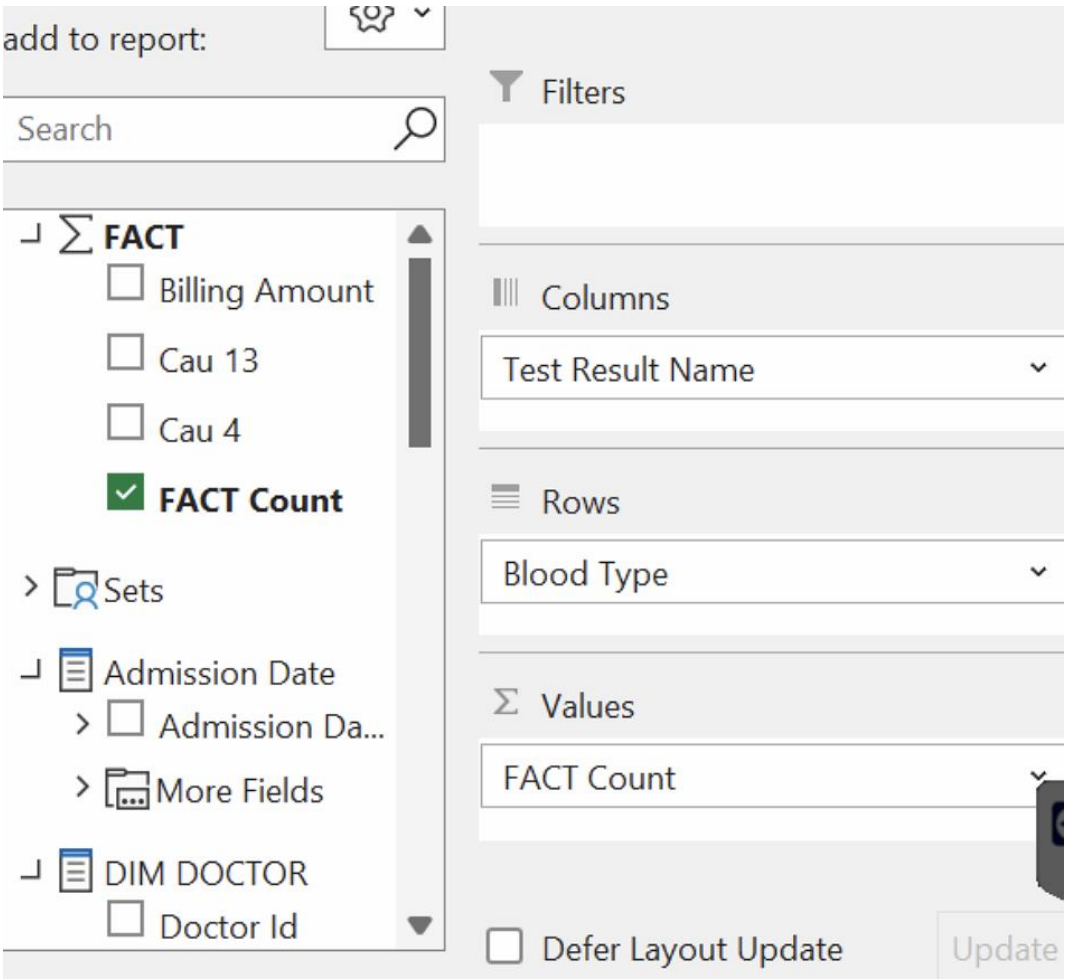



*Hình 194. Kết quả câu truy vấn 1 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.1 (Kéo thả bằng công cụ BI)

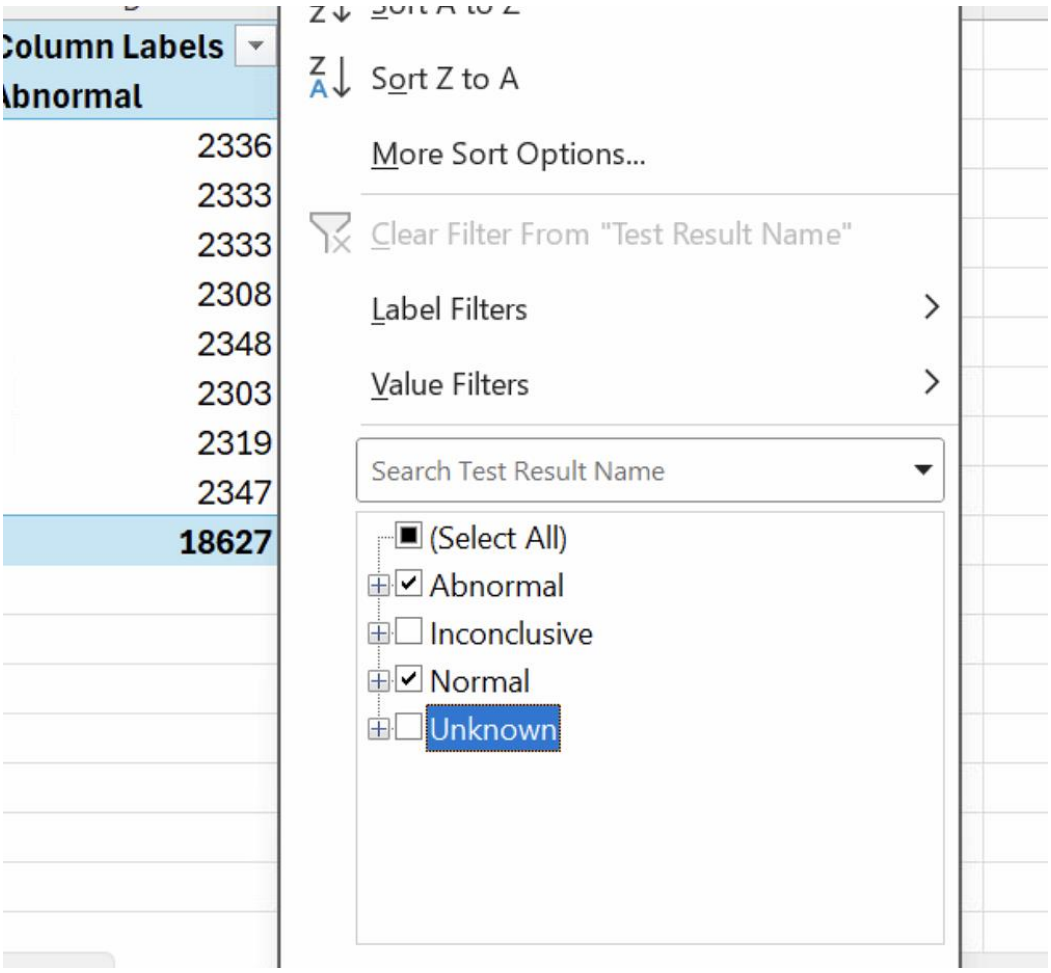
**10.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Blood Type vào Rows và thuộc tính Test Result Name vào Columns, kéo độ đo FACT Count vào cột Values.




*Hình 195 . Chọn các thuộc tính cần thiết*

- **Bước 2:** Chọn Row Labels, chọn 2 giá trị “Abnormal” và “Normal”.



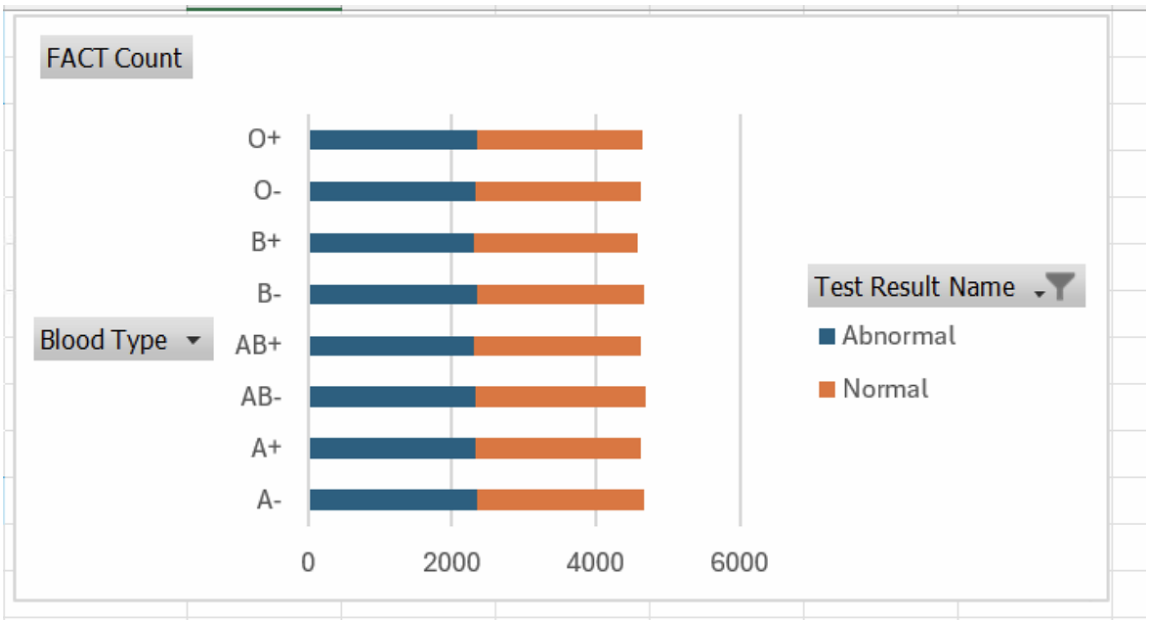

*Hình 196. Filter ( chọn Abnormal và Normal)*

- **Kết quả câu truy vấn dạng Matrix:**

| FACT Count | Column Labels | | |
|---|---|---|---|
| Row Labels | Abnormal | Normal | Grand Total |
| A- | 2336 | 2335 | 4671 |
| A+ | 2333 | 2293 | 4626 |
| AB- | 2333 | 2362 | 4695 |
| AB+ | 2308 | 2314 | 4622 |
| B- | 2348 | 2328 | 4676 |
| B+ | 2303 | 2276 | 4579 |
| O- | 2319 | 2301 | 4620 |
| O+ | 2347 | 2308 | 4655 |
| **Grand Total** | **18627** | **18517** | **37144** |

*Hình 197. Kết quả câu truy vấn 2 dạng Matrix*

- **Kết quả câu truy vấn dạng Chart:**

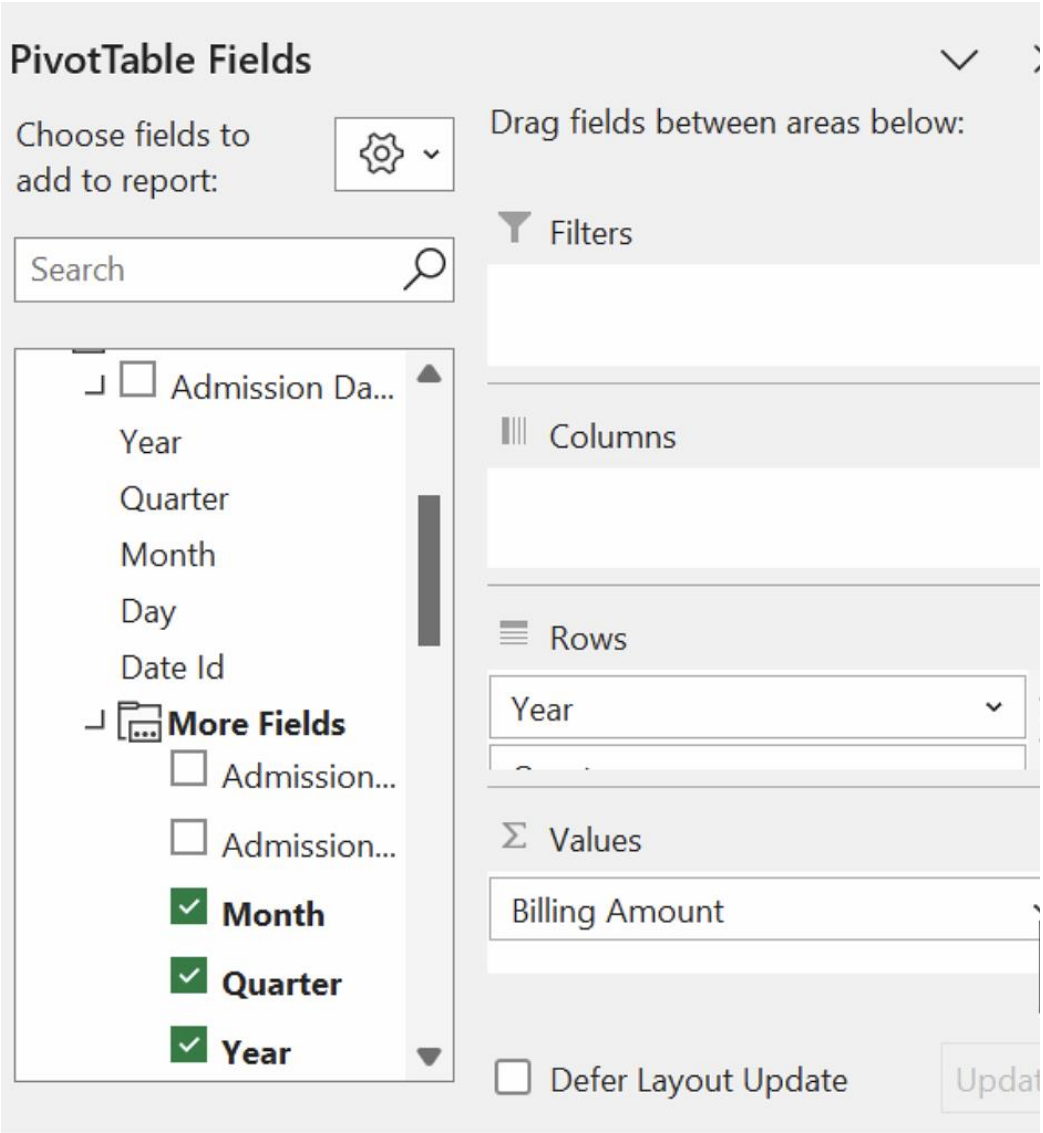



*Hình 198. Kết quả câu truy vấn 2 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.2 (Kéo thả bằng công cụ BI)

**10.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Year, Quarter, Month vào Rows và độ đo Billing Amount vào cột Values.




*Hình 199. Chọn các thuộc tính cần thiết*

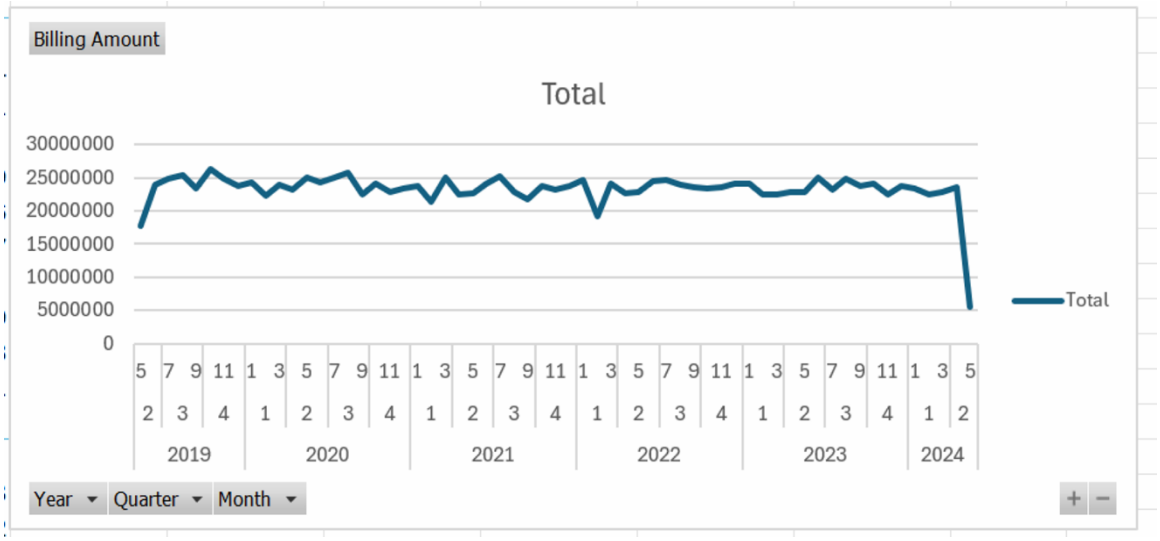
- **Kết quả câu truy vấn dạng Table:**

| | A | B | C |
|---|---|---|---|
| 1 | Row Labels | Billing Amount | |
| 2 | 2019 | | |
| 3 | 2 | | |
| 4 | 5 | 17653145.14 | |
| 5 | 6 | 23885179.51 | |
| 6 | 3 | | |
| 7 | 7 | 24771559.19 | |
| 8 | 8 | 25397958.6 | |
| 9 | 9 | 23334079.87 | |
| 10 | 4 | | |
| 11 | 10 | 26276190.69 | |
| 12 | 11 | 24784507.53 | |
| 13 | 12 | 23746704.61 | |
| 14 | 2020 | | |
| 15 | 1 | | |
| 16 | 1 | 24371975.18 | |
| 17 | 2 | 22324346.52 | |
| 18 | 3 | 23907855.08 | |
| 19 | 2 | | |

Câu 1 Câu 2 Câu 3

*Hình 200. Kết quả câu truy vấn 3 dạng Matrix*

- **Kết quả câu truy vấn dạng Chart:**

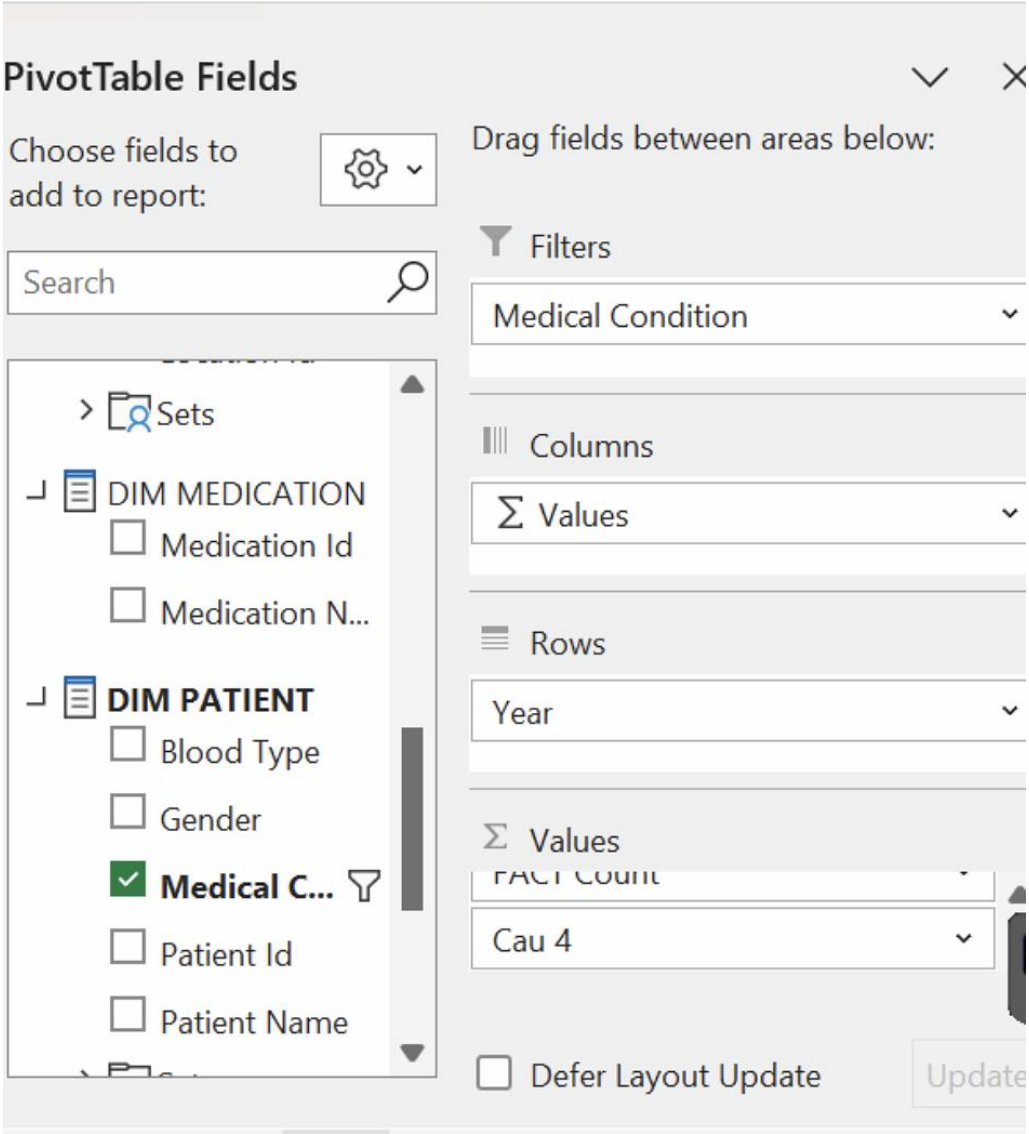



*Hình 201. Kết quả câu truy vấn 3 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.3 (Kéo thả bằng công cụ BI)

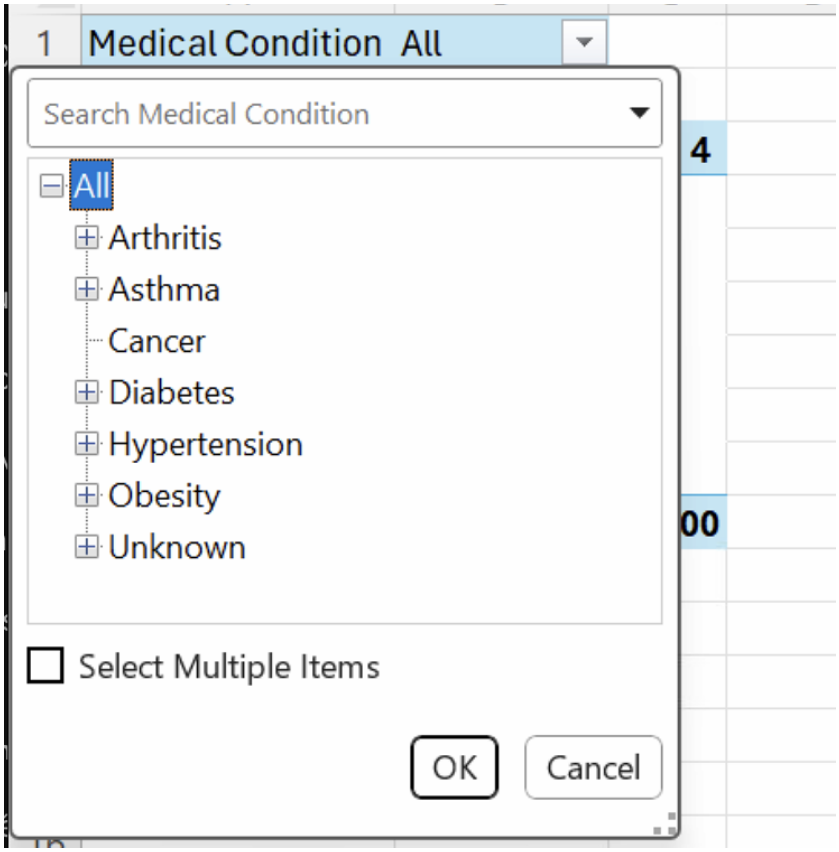
### 10.4. Đưa ra số lượng bệnh nhân bị mắc bệnh 'Cancer' theo từng năm.

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Year vào Rows, độ đo FACT Count, Cau 4 vào cột Values và Medical Condition vào Filters.




*Hình 202. Chọn các thuộc tính cần thiết*

  - **Bước 2:** Lọc theo bệnh Cancer




*Hình 203. Lọc bệnh theo Cancer*

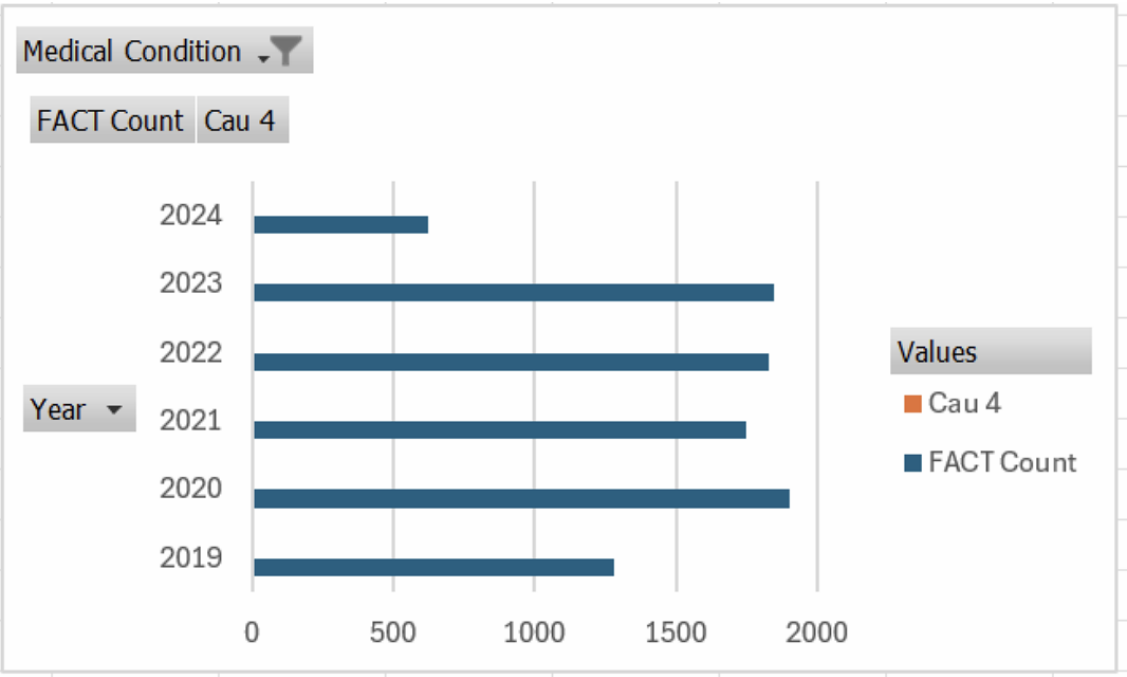
- **Kết quả câu truy vấn dạng Table:**

Medical Condition Cancer

| Row Labels | FACT Count | Cau 4 |
|---|---|---|
| 2019 | 1283 | |
| 2020 | 1900 | |
| 2021 | 1748 | |
| 2022 | 1825 | |
| 2023 | 1849 | |
| 2024 | 622 | |
| **Grand Total** | **9227** | **9227** |

*Hình 204. Kết quả câu truy vấn 4 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**

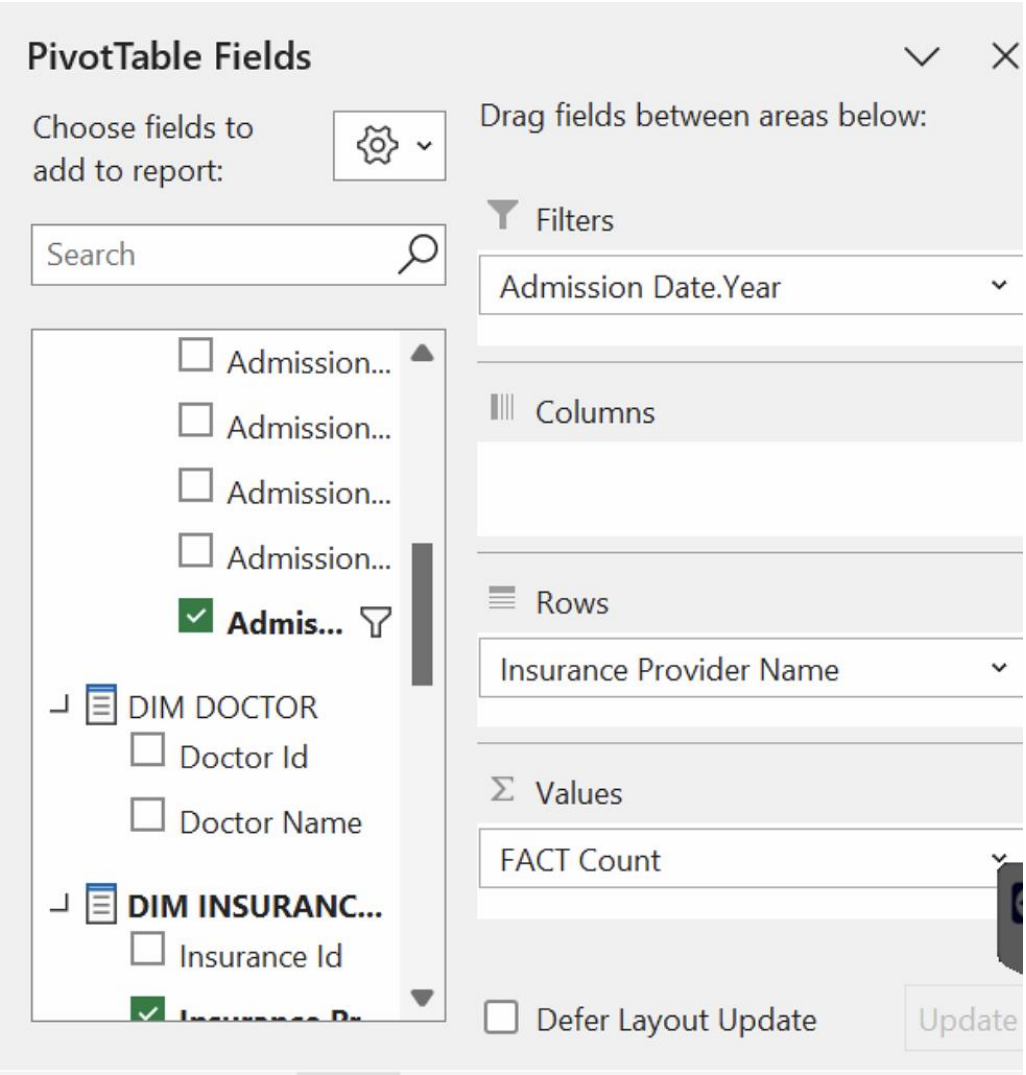



*Hình 205. Kết quả câu truy vấn 4 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.4 (Kéo thả bằng công cụ BI)

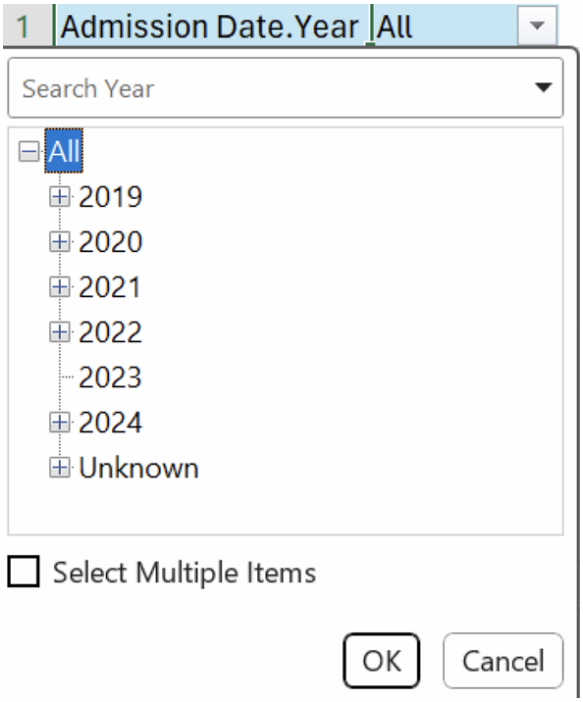
**10.5. Đưa ra Top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Insurance Provider Name vào Rows, độ đo FACT Count vào cột Values và Year vào Filters.




*Hình 206. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo năm 2023.



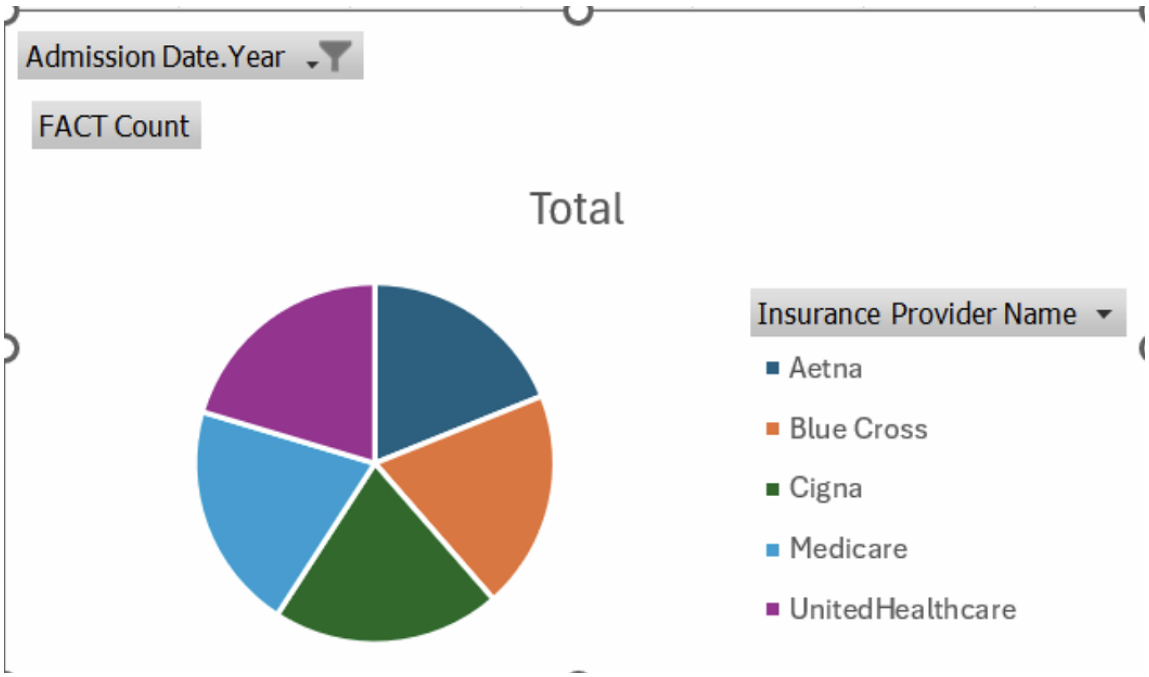

*Hình 207. Lọc theo 2023*

- **Kết quả câu truy vấn dạng Table:**

| Row Labels | FACT Count |
|---|---|
| Aetna | 2086 |
| Blue Cross | 2166 |
| Cigna | 2268 |
| Medicare | 2258 |
| UnitedHealthcare | 2248 |
| **Grand Total** | **11026** |

*Hình 208. Kết quả câu truy vấn 5 dạng Table.*

- **Kết quả câu truy vấn dạng Chart:**

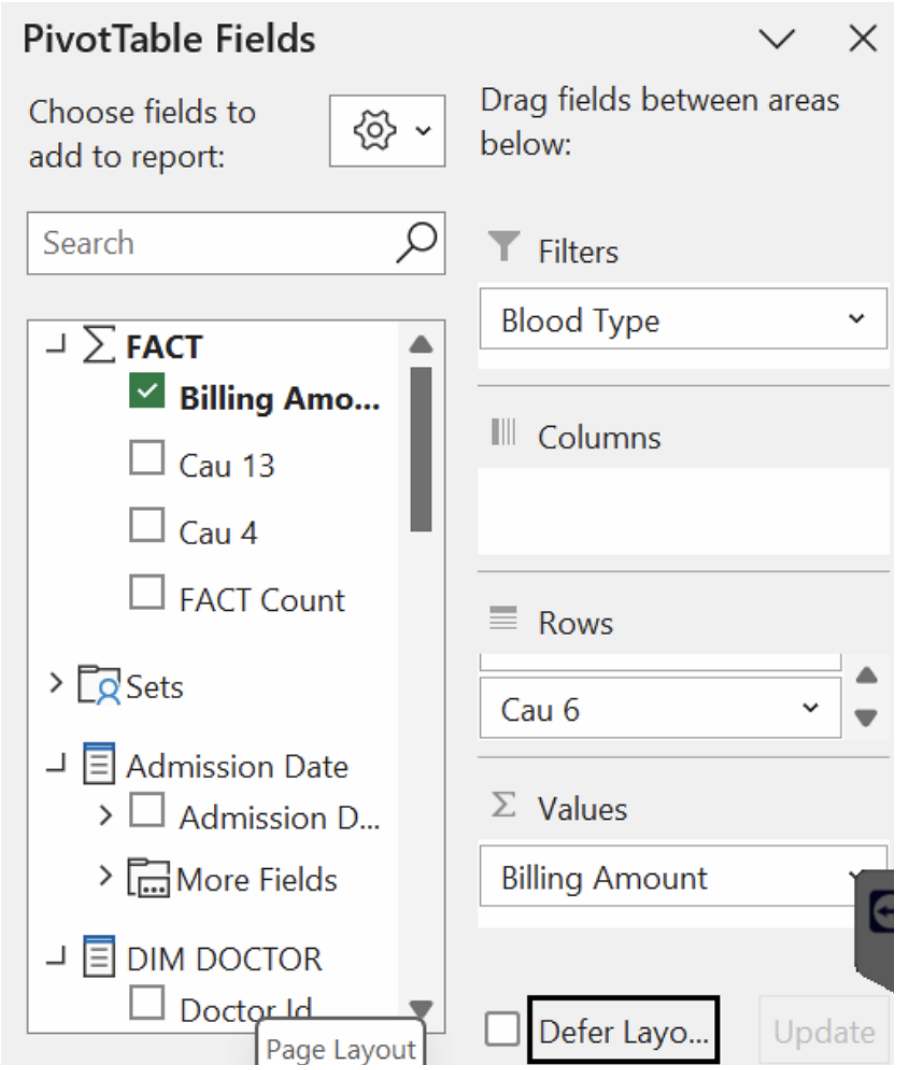



*Hình 209. Kết quả câu truy vấn 5 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.5 (Kéo thả bằng công cụ BI)

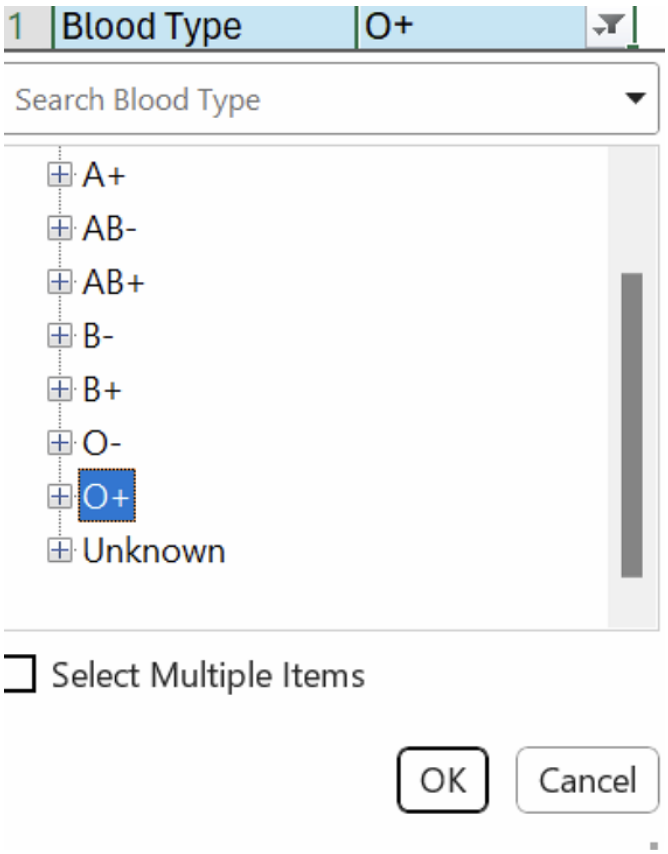
**10.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc Top 5.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Patient Name, Cau 6 vào Rows, độ đo Billing Amount vào cột Values và Blood Type vào Filters.




*Hình 210. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo nhóm máu O+



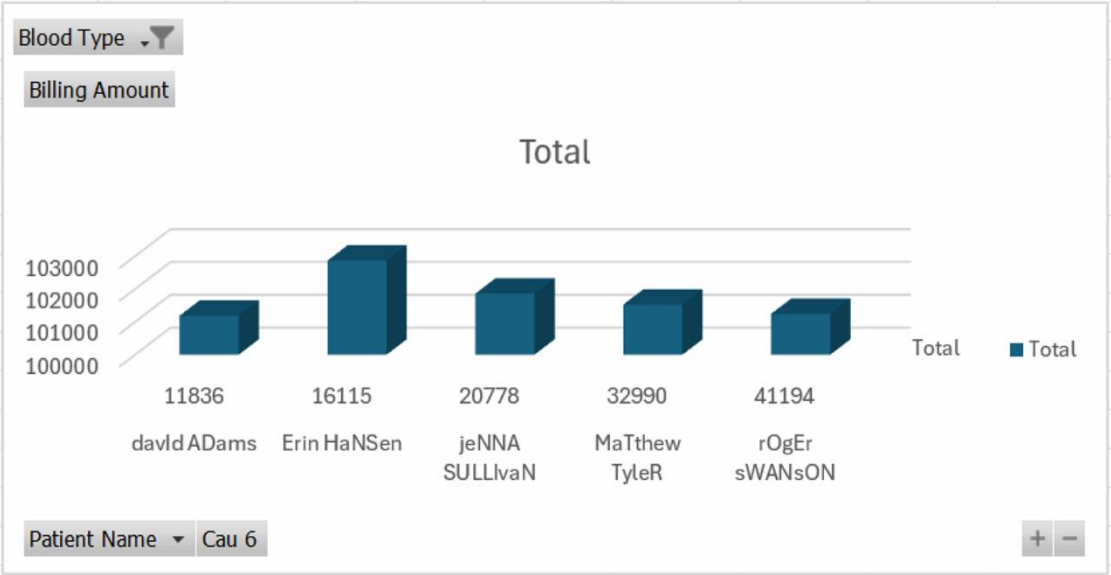

*Hình 211. Lọc theo nhóm máu 0+*

- **Kết quả câu truy vấn dạng Table:**

| Blood Type | O+ |
|---|---|
| | |
| **Row Labels** | **Billing Amount** |
| **davId ADams** | |
| 11836 | 101187.3778 |
| **Erin HaNSen** | |
| 16115 | 102863.9544 |
| **jeNNA SULLIvaN** | |
| 20778 | 101859.2016 |
| **MaTthew TyleR** | |
| 32990 | 101525.0567 |
| **rOgEr sWANsON** | |
| 41194 | 101244.4518 |

*Hình 212. Kết quả câu truy vấn 6 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**

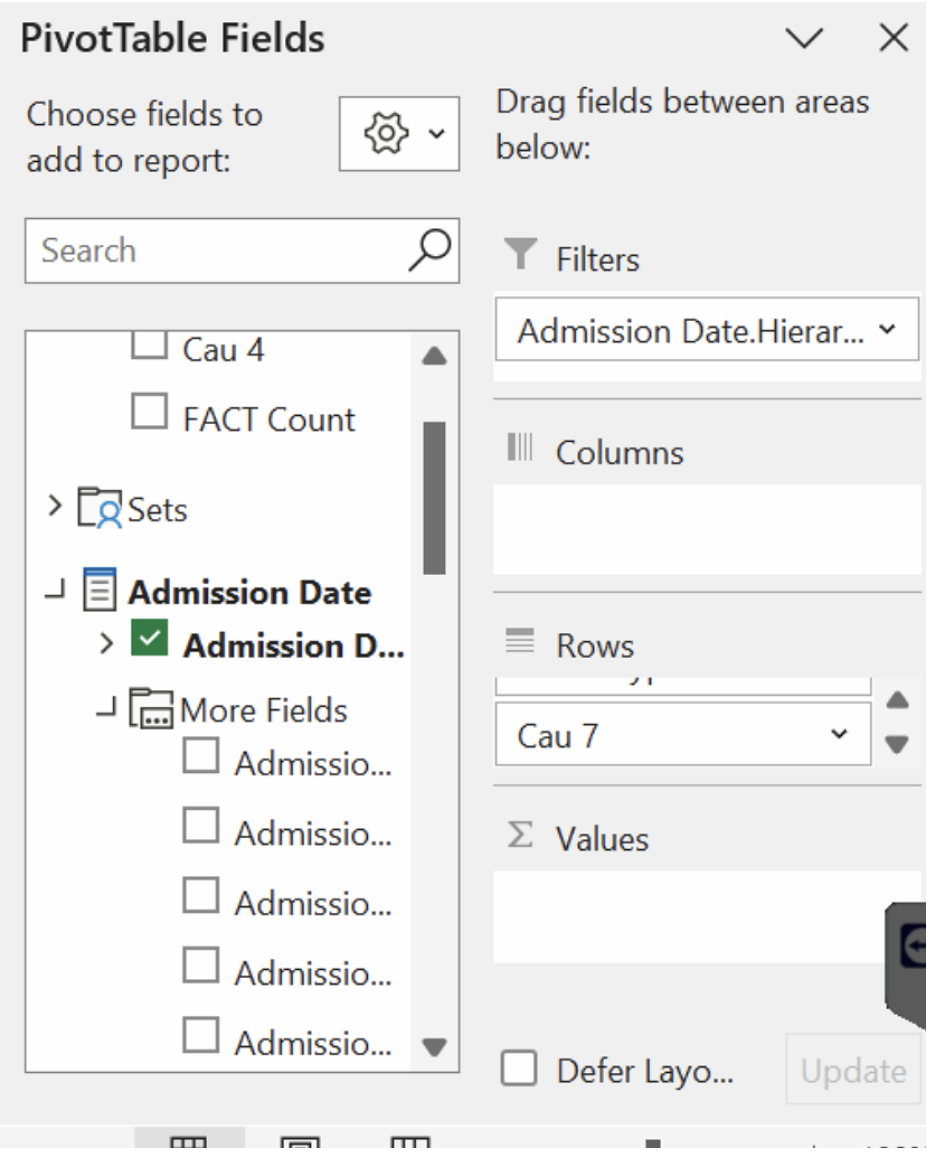



*Hình 213. Kết quả câu truy vấn 6 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.6 (Kéo thả bằng công cụ BI)

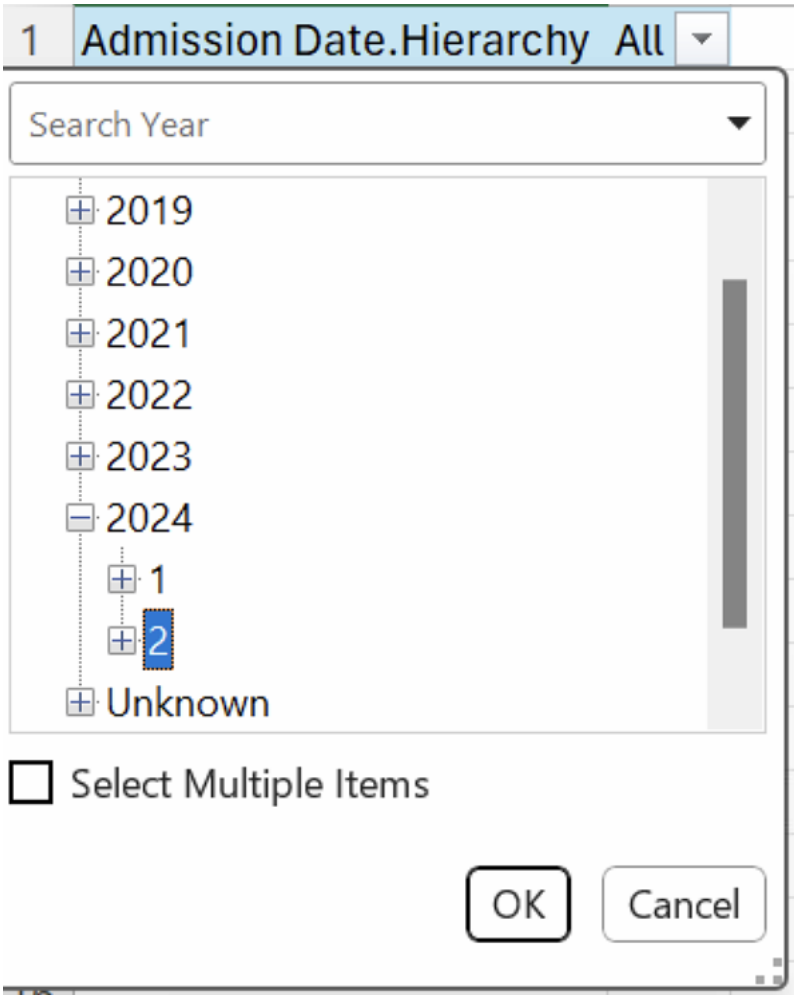
## 10.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024.

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Cau 7, Patient Name, Blood Type, Gender vào Rows, độ đo Billing Amount vào cột Values và Hierarchy của Admission Date vào Filters.



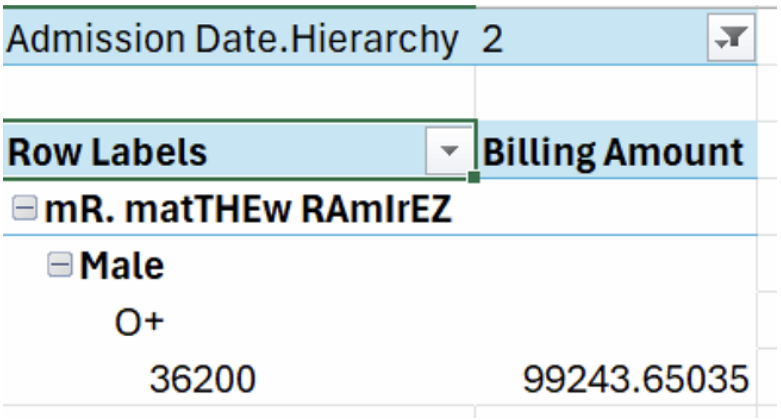

*Hình 214. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo quý 2 năm 2024.




*Hình 215. Lọc theo quý 2 năm 2024*

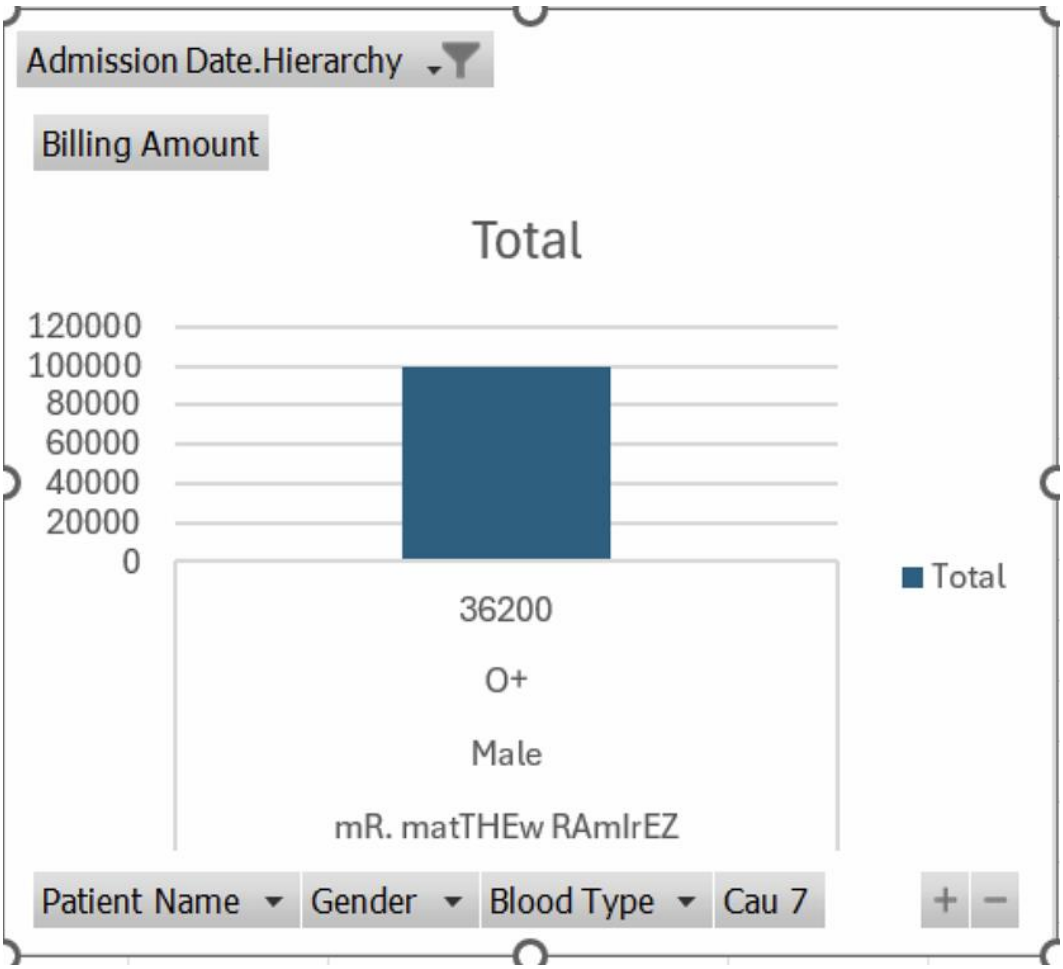
- **Kết quả câu truy vấn dạng Table:**



| Admission Date.Hierarchy | 2 |
|---|---|
| **Row Labels** | **Billing Amount** |
| **mR. matTHEw RAmIrEZ** | |
| **Male** | |
| O+ | |
| 36200 | 99243.65035 |

*Hình 216. Kết quả câu truy vấn 7 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**

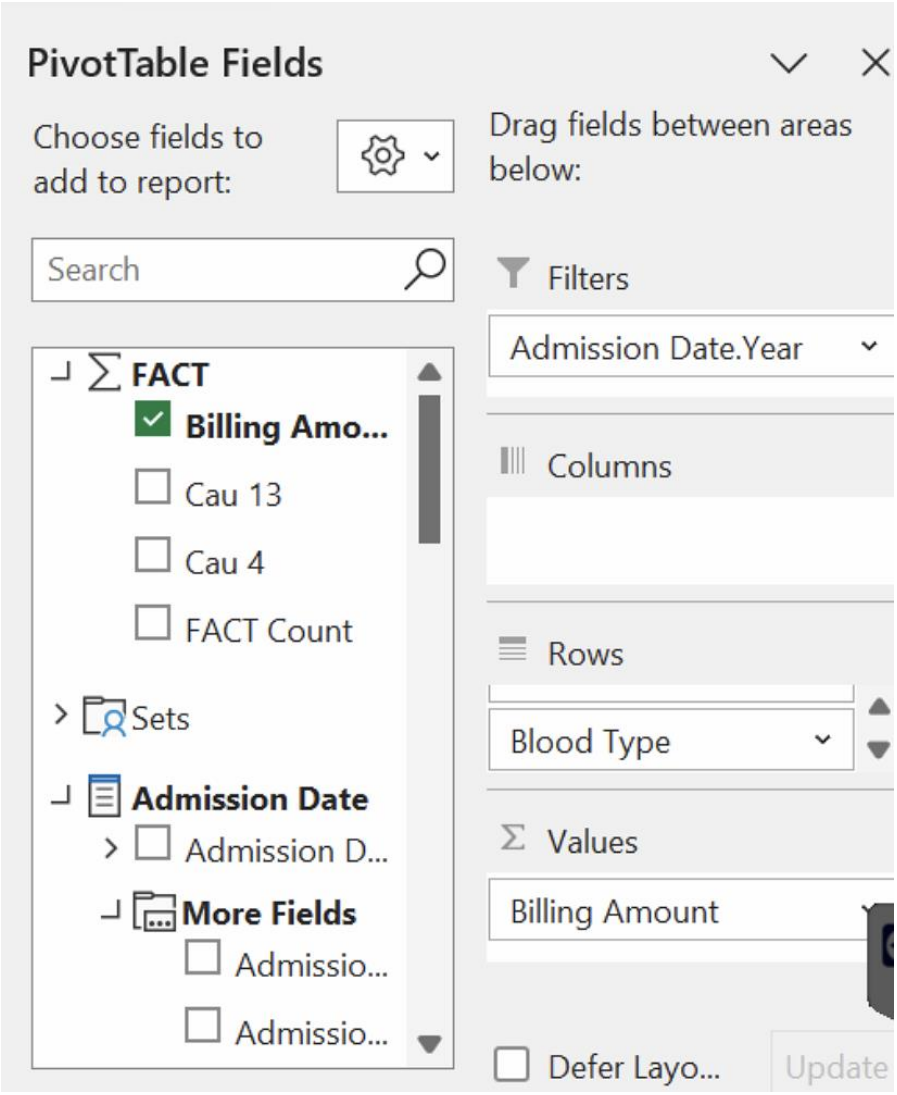



*Hình 217. Kết quả câu truy vấn 7 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.7 (Kéo thả bằng công cụ BI)

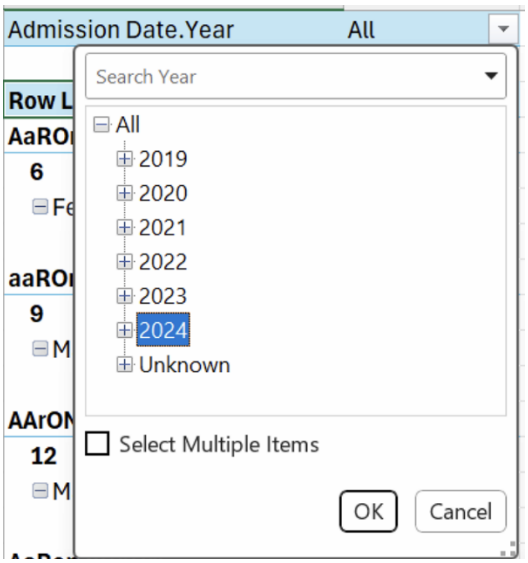
**10.8. Liệt kê các bệnh nhân có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Cau 8, Patient Name, Blood Type, Gender vào Rows, độ đo Billing Amount vào cột Values và Year của Admission Date vào Filters.




*Hình 218. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo năm 2024.




*Hình 219 . Lọc theo 2024*

- **Kết quả câu truy vấn dạng Table:**

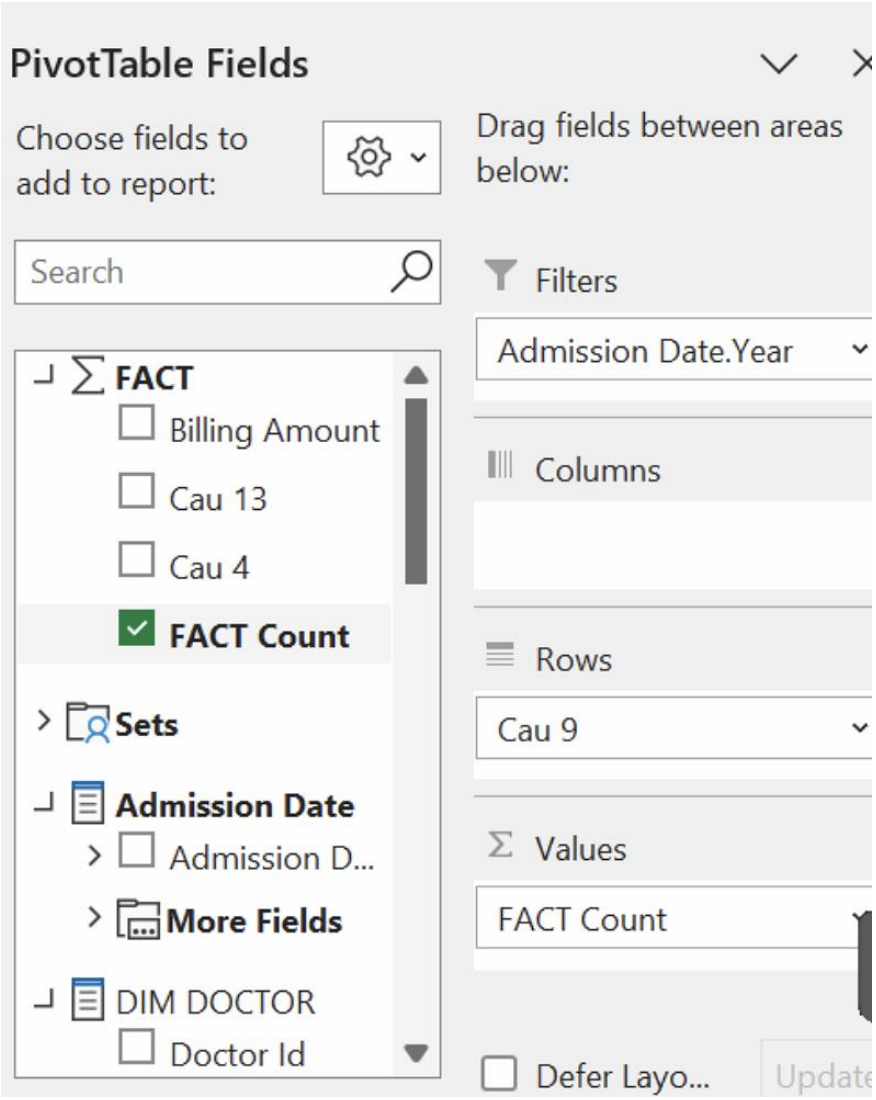
| Admission Date.Year | 2024 |
|---|---|
| | |
| **Row Labels** | **Billing Amount** |
| **AarOn sALiNas** | |
| **115** | |
| Female | |
| B+ | 9713.473813 |
| **aaroN smIth** | |
| **121** | |
| Female | |
| B- | 8340.967113 |
| **aAROn wEbeR** | |
| **133** | |
| Female | |
| AB+ | 8767.413834 |
| **aDAm cOLLins** | |
| **229** | |
| Male | |
| O+ | 4194.769773 |

*Hình 220. Kết quả câu truy vấn 8 dạng Table*

- **Nhận xét:** Kết quả giống câu 9.8 (Kéo thả bằng công cụ BI)

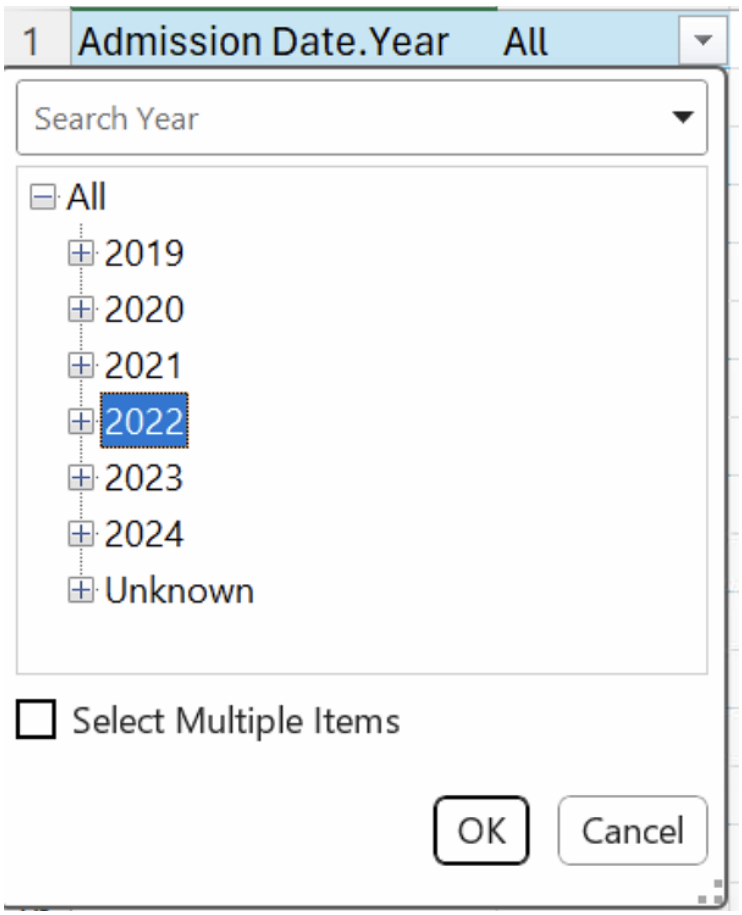
**10.9. Trong mỗi loại bệnh, in ra bác sĩ có số lượng bênh nhân lớn nhất trong năm 2022.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Cau 9 vào Rows, độ đo FACT Count vào cột Values và Year của Admission Date vào Filters.




*Hình 221. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo năm 2022.



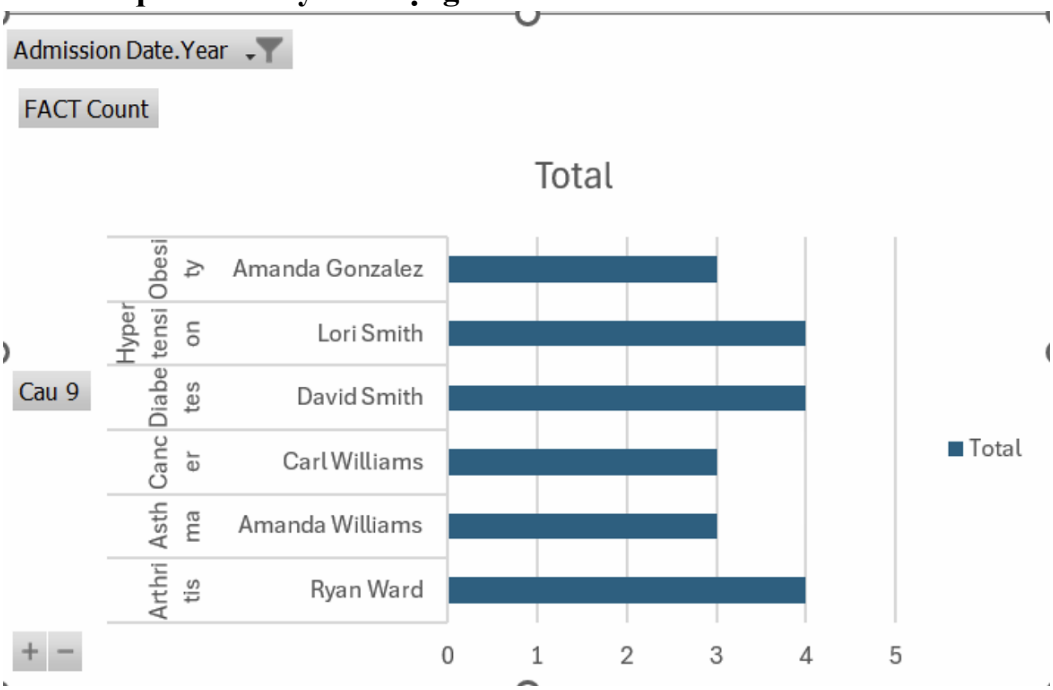

*Hình 222. Lọc theo năm 2022*

- **Kết quả câu truy vấn dạng Table:**

| Admission Date.Year | 2022 |
|---|---|
| | |
| **Row Labels** | **FACT Count** |
| **Arthritis** | |
| Ryan Ward | 4 |
| **Asthma** | |
| Amanda Williams | 3 |
| **Cancer** | |
| Carl Williams | 3 |
| **Diabetes** | |
| David Smith | 4 |
| **Hypertension** | |
| Lori Smith | 4 |
| **Obesity** | |
| Amanda Gonzalez | 3 |

*Hình 223. Kết quả câu truy vấn 9 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**

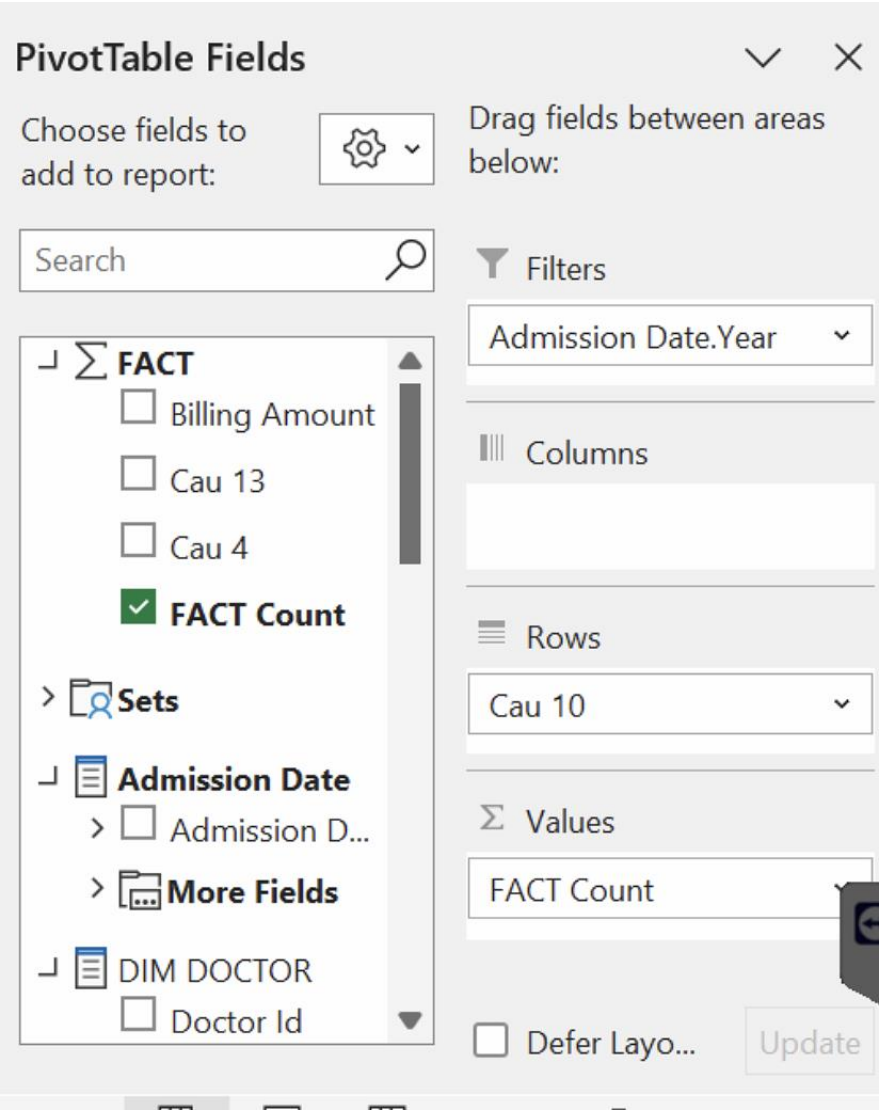



*Hình 224. Kết quả câu truy vấn 9 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.9 (Kéo thả bằng công cụ BI)

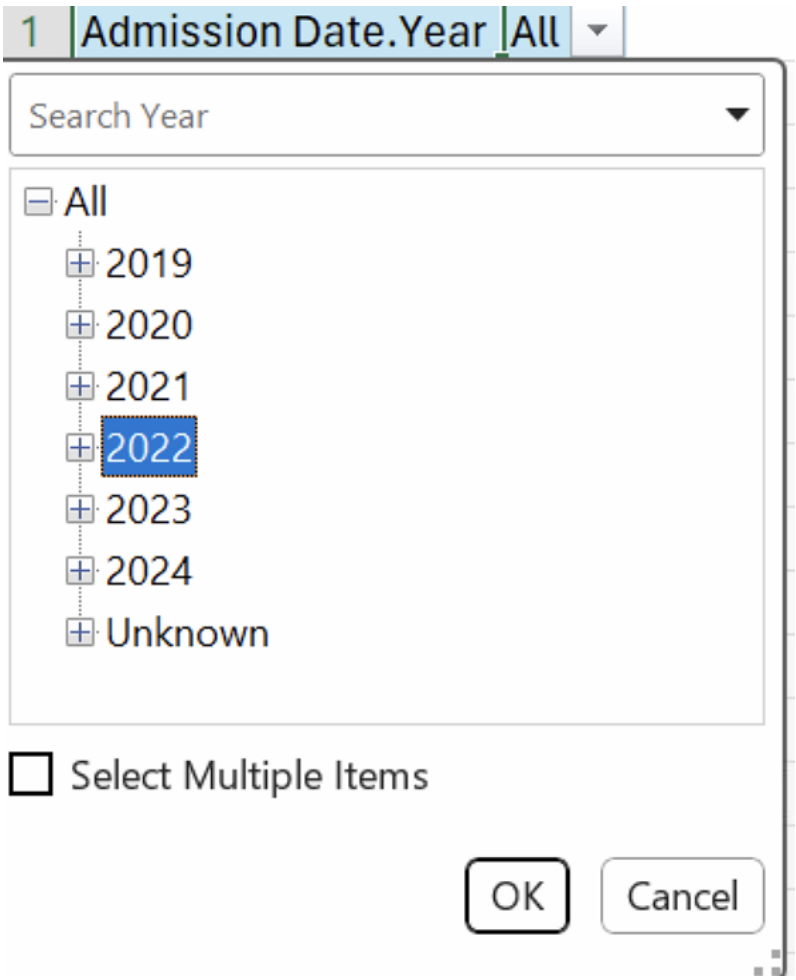
**10.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Cau 10 vào Rows, độ đo FACT Count vào cột Values và Year của Admission Date vào Filters.




*Hình 225. Chọn các thuộc tính cần thiết*

  - **Bước 2:** Lọc theo năm 2022.



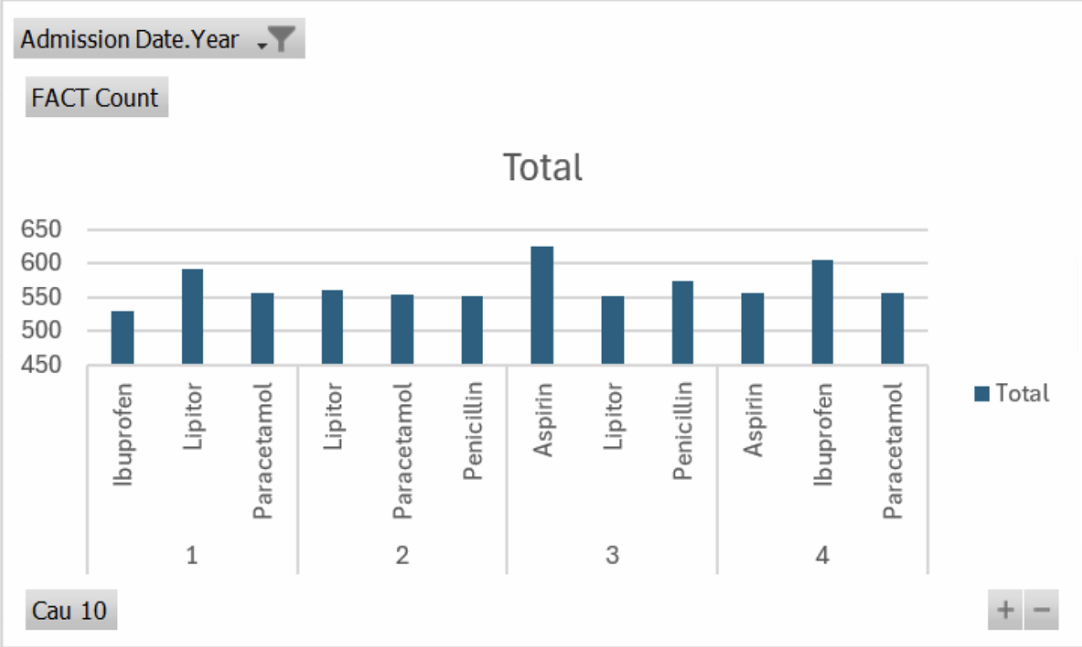

*Hình 226. Lọc theo năm 2022*

- **Kết quả câu truy vấn dạng Table:**

| Admission Date.Year | 2022 |
|---|---|
| | |
| **Row Labels** | **FACT Count** |
| **1** | |
| Ibuprofen | 529 |
| Lipitor | 592 |
| Paracetamol | 556 |
| **2** | |
| Lipitor | 561 |
| Paracetamol | 555 |
| Penicillin | 551 |
| **3** | |
| Aspirin | 625 |
| Lipitor | 552 |
| Penicillin | 575 |
| **4** | |
| Aspirin | 557 |
| Ibuprofen | 605 |
| Paracetamol | 556 |

*Hình 227 . Kết quả câu truy vấn 10 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**



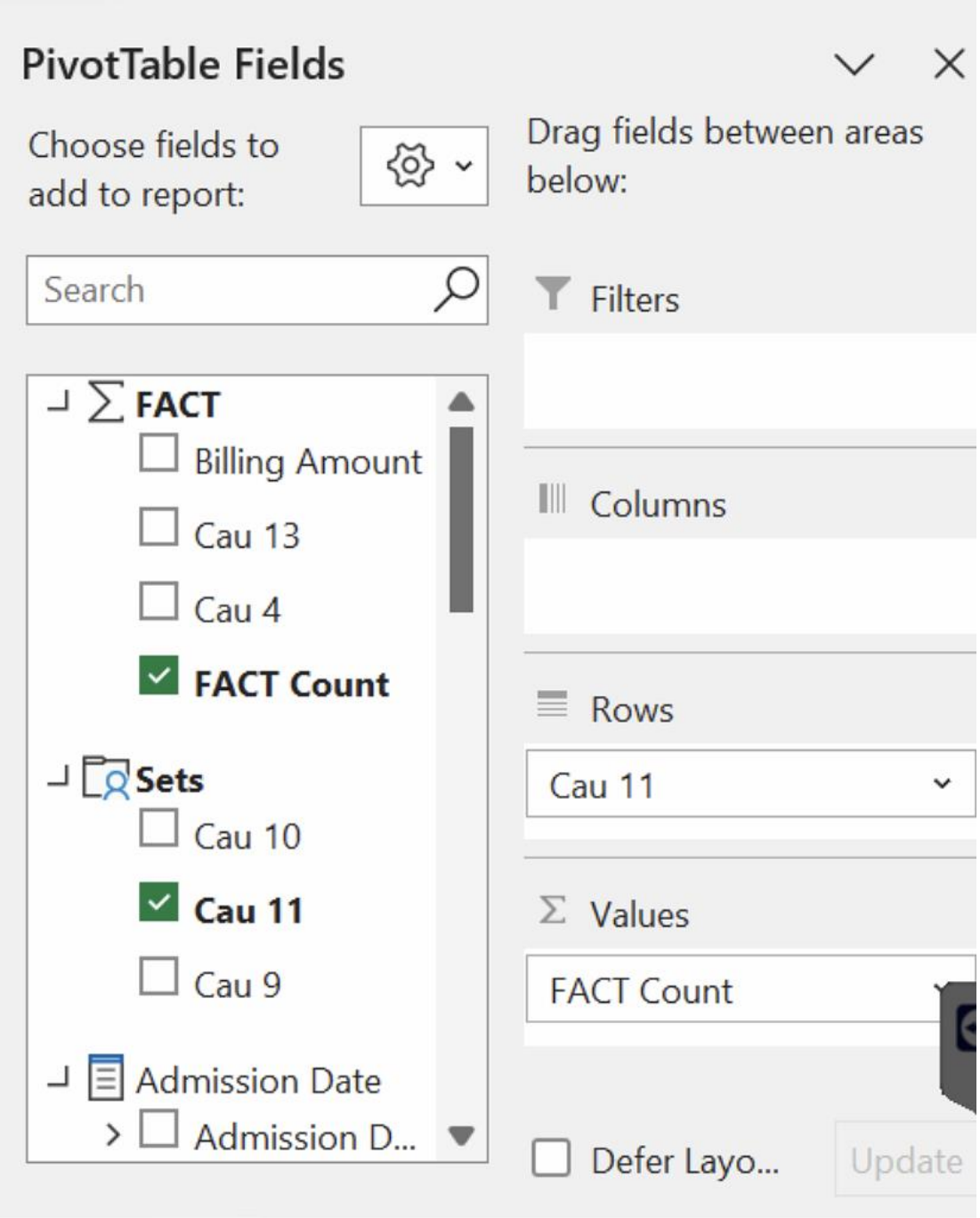

*Hình 228. Kết quả câu truy vấn 10 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.10 (Kéo thả bằng công cụ BI)

**10.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể.**

- **Thực hiện quá trình phân tích dữ liệu:**

- **Bước 1:** Kéo thuộc tính Cau 11 vào Rows, độ đo FACT Count vào cột Values.



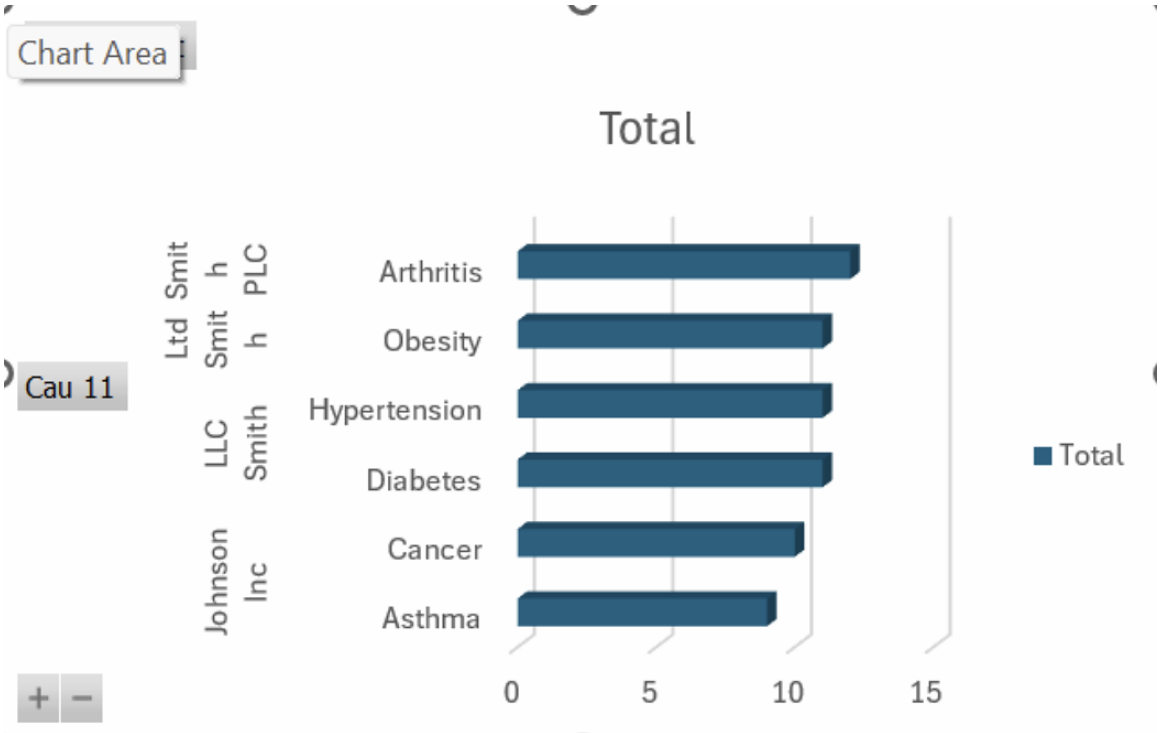

*Hình 229. Chọn các thuộc tính cần thiết*

- **Kết quả câu truy vấn dạng Table:**

| Row Labels | FACT Count |
|---|---|
| **Johnson Inc** | |
| Asthma | 9 |
| Cancer | 10 |
| **LLC Smith** | |
| Diabetes | 11 |
| Hypertension | 11 |
| **Ltd Smith** | |
| Obesity | 11 |
| **Smith PLC** | |
| Arthritis | 12 |

*Hình 230. Kết quả câu truy vấn 11 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**



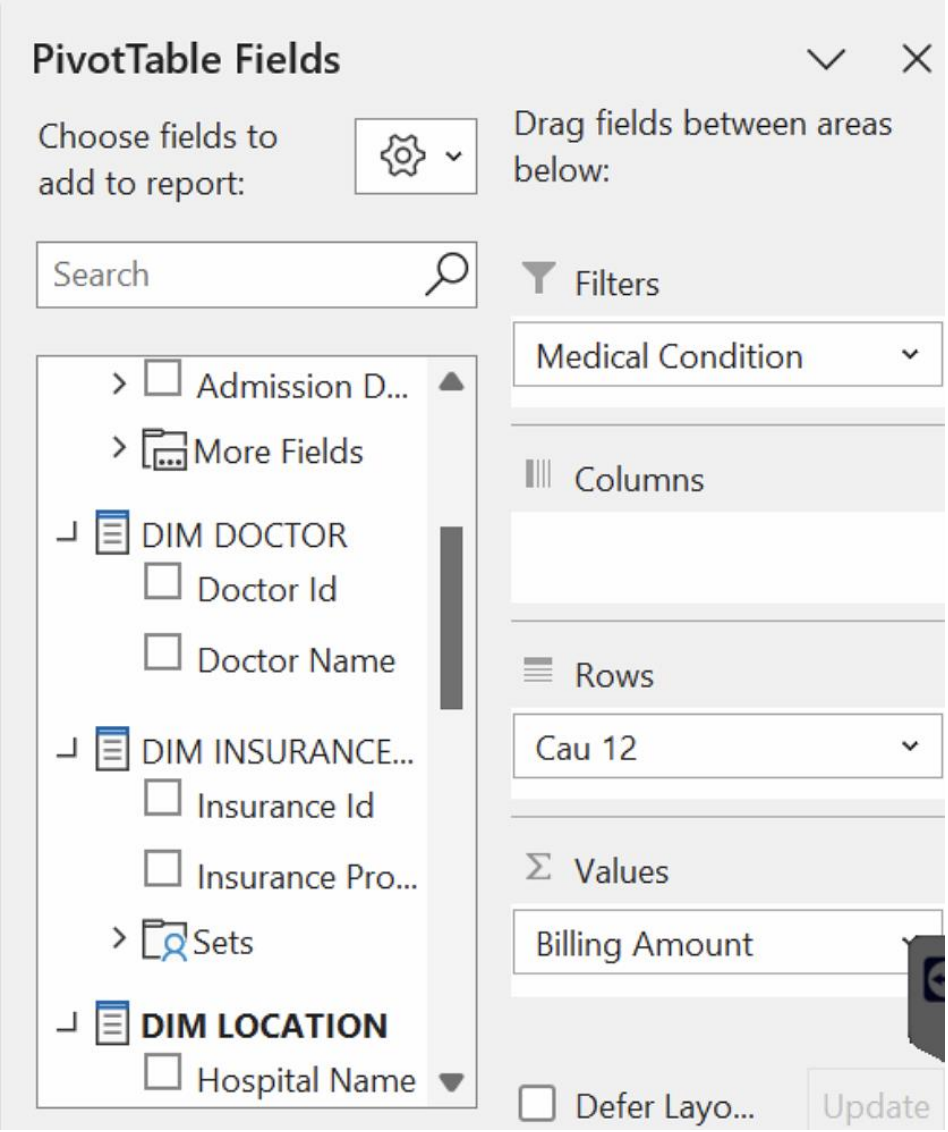

*Hình 231. Kết quả câu truy vấn 11 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.11 (Kéo thả bằng công cụ BI)

**10.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần.**

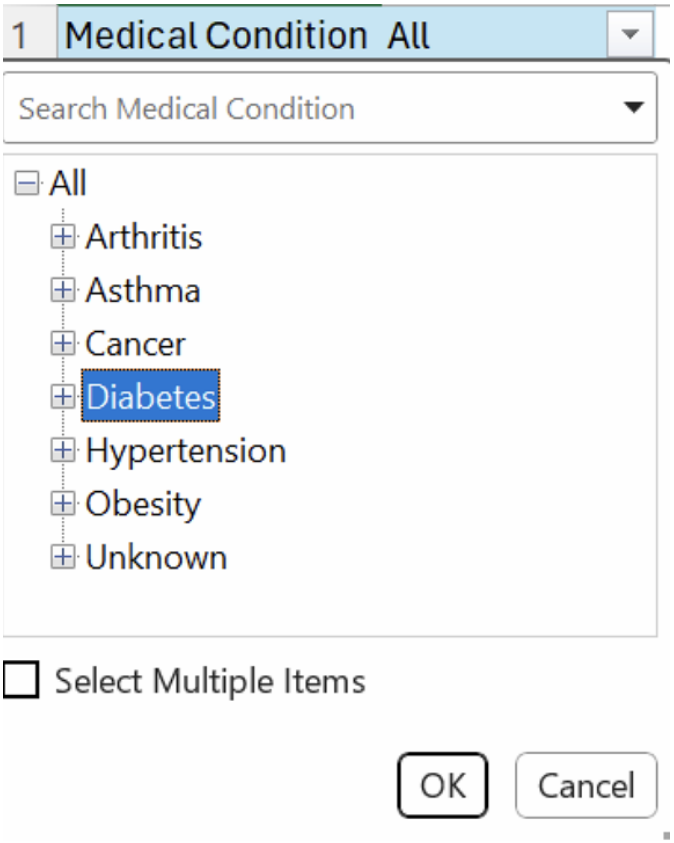
- **Thực hiện quá trình phân tích dữ liệu:**

- **Bước 1:** Kéo thuộc tính Cau 12 vào Rows, độ đo Billing Amount vào cột Values và Medical Condition vào Filters.




*Hình 232. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo bệnh Diabetes.



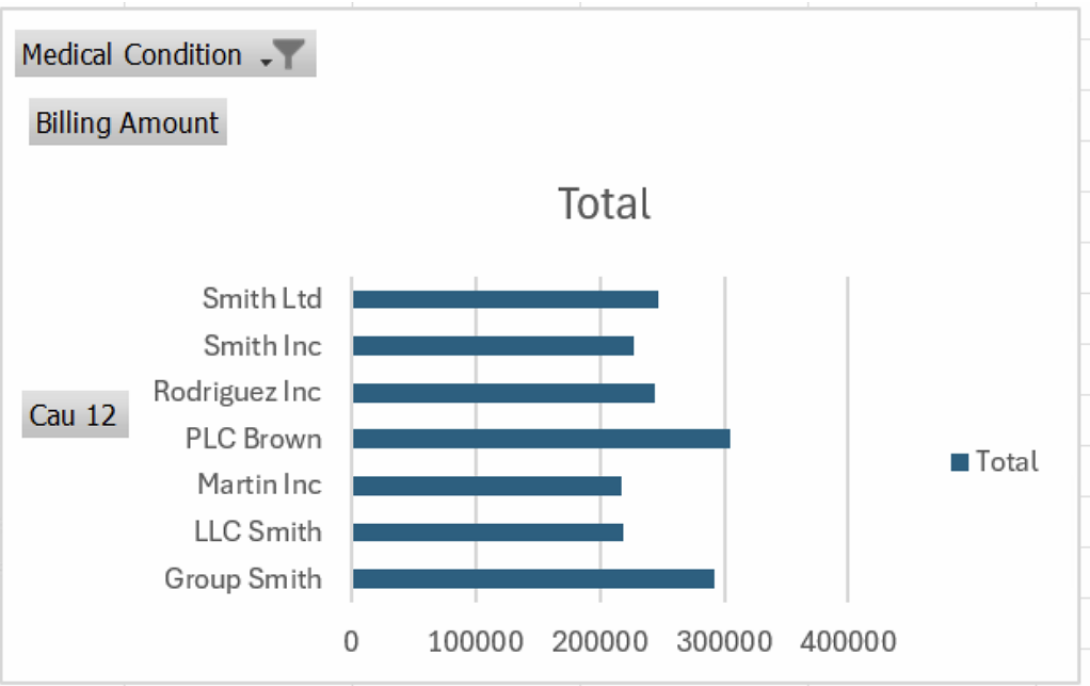

*Hình 233. Lọc theo căn bệnh Diabetes*

- **Kết quả câu truy vấn dạng Table:**

| Medical Condition | Diabetes |
|---|---|
| | |
| **Row Labels** | **Billing Amount** |
| Group Smith | 291522.6529 |
| LLC Smith | 218458.2625 |
| Martin Inc | 217248.2123 |
| PLC Brown | 304480.7954 |
| Rodriguez Inc | 243599.129 |
| Smith Inc | 227727.1306 |
| Smith Ltd | 247134.6459 |
| **Grand Total** | **238539725.5** |

*Hình 234. Kết quả câu truy vấn 12 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**

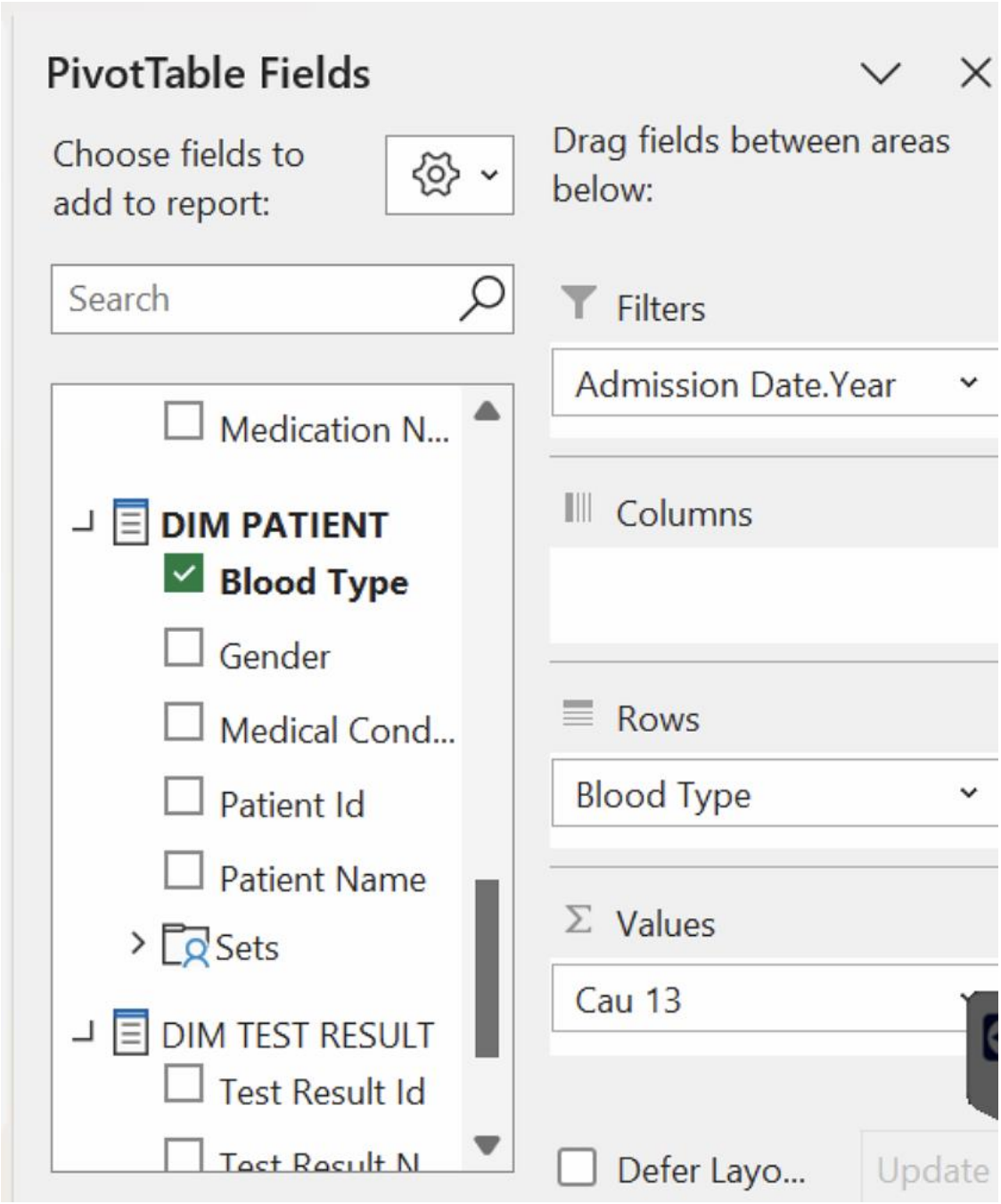



*Hình 235 . Kết quả câu truy vấn 12 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.12 (Kéo thả bằng công cụ BI)

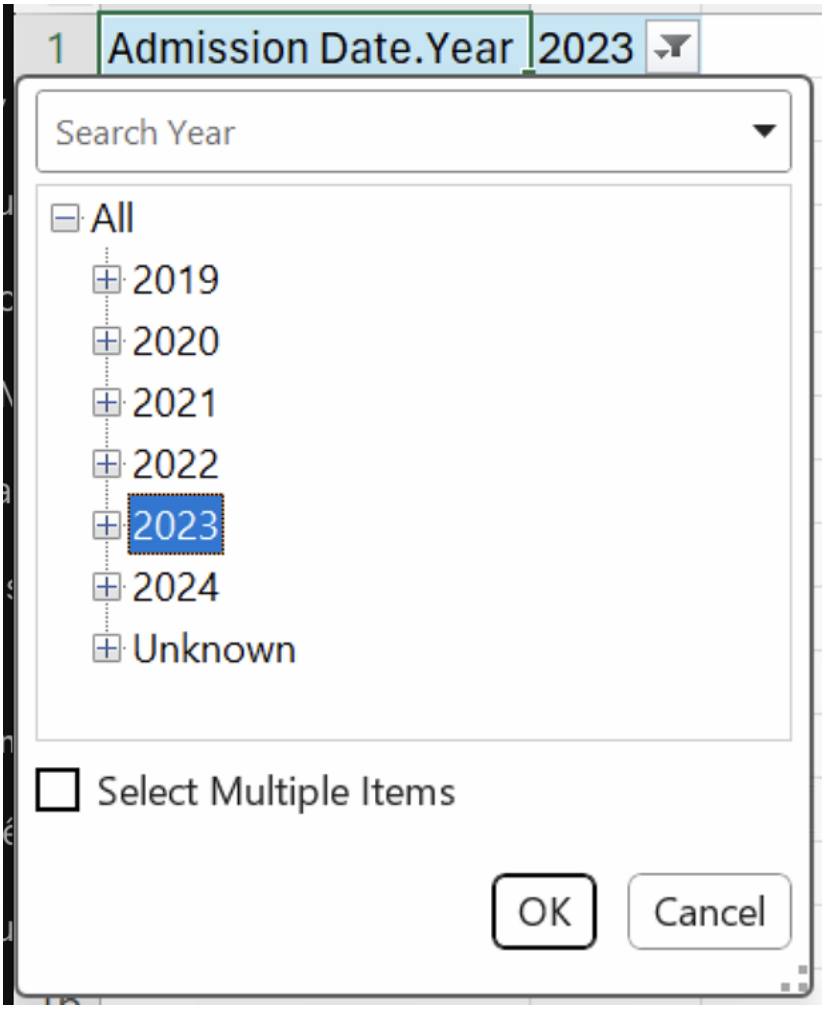
**10.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.**

- **Thực hiện quá trình phân tích dữ liệu:**
  - **Bước 1:** Kéo thuộc tính Blood Type vào Rows, độ đo Cau 13 vào cột Values và Year của Admission Date vào Filters.




*Hình 236. Chọn các thuộc tính cần thiết*

- **Bước 2:** Lọc theo năm 2023



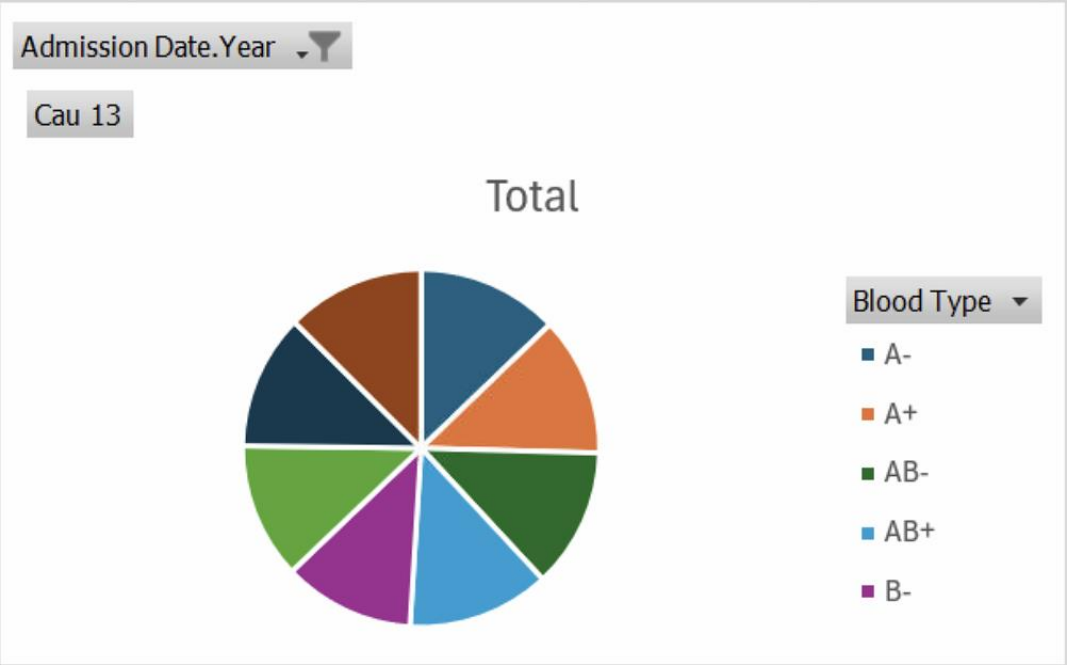

*Hình 237. Lọc theo năm 2023*

- **Kết quả câu truy vấn dạng Table:**

| Admission Date.Year | 2023 |
|---|---|
| | |
| **Row Labels** | **Cau 13** |
| A- | 12.84% |
| A+ | 12.59% |
| AB- | 12.64% |
| AB+ | 12.92% |
| B- | 11.91% |
| B+ | 12.31% |
| O- | 12.23% |
| O+ | 12.56% |
| **Grand Total** | **#NUM!** |

*Hình 238. Kết quả câu truy vấn 13 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**



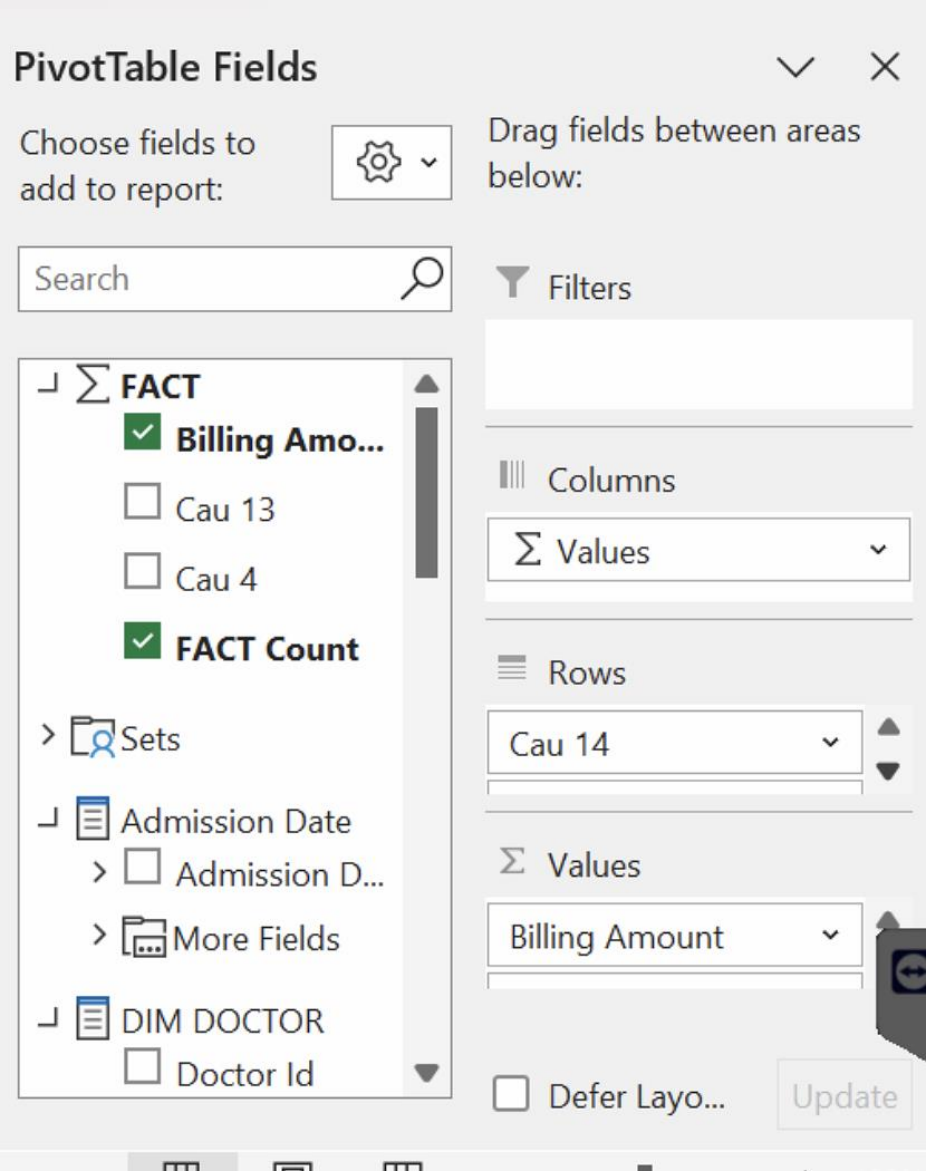

*Hình 239. Kết quả câu truy vấn 13 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.13 (Kéo thả bằng công cụ BI)

**10.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất hoặc top 3 số lần nhập viện nhiều nhất.**

- **Thực hiện quá trình phân tích dữ liệu:**

- **Bước 1:** Kéo thuộc tính Cau 14, Patient Name vào Rows, độ đo FACT Count, Billing Amount vào cột Values.



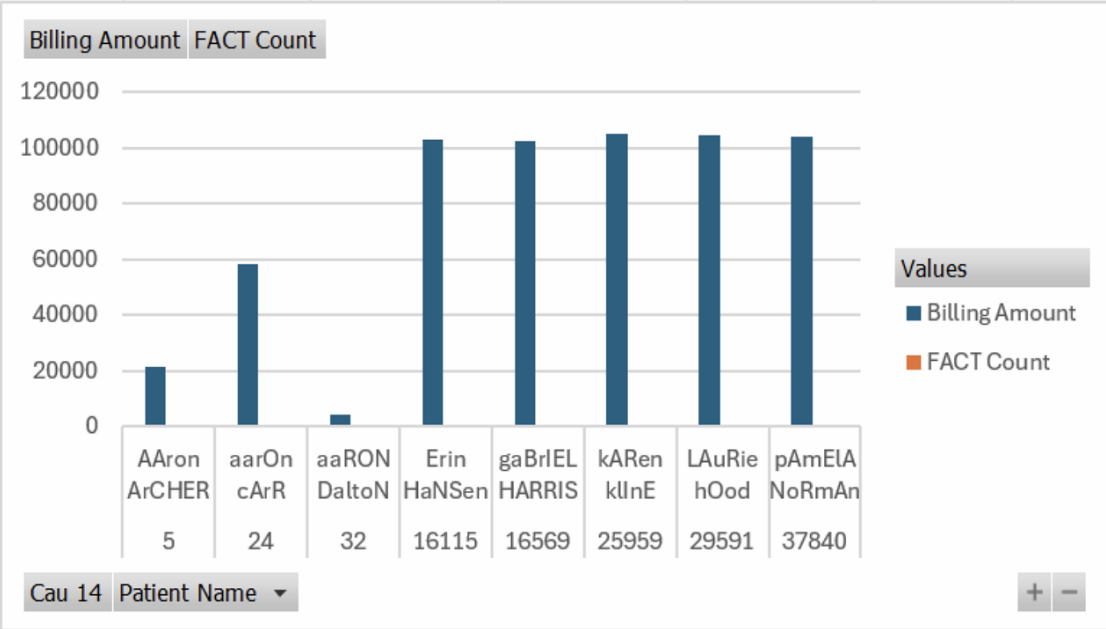

*Hình 240. Chọn các thuộc tính cần thiết*

- **Kết quả câu truy vấn dạng Table:**

| Row Labels | Billing Amount | FACT Count |
|---|---|---|
| **5** | | |
| AAron ArCHER | 21204.15437 | 2 |
| **24** | | |
| aarOn cArR | 58085.92366 | 2 |
| **32** | | |
| aaRON DaltoN | 4140.518953 | 2 |
| **16115** | | |
| Erin HaNSen | 102863.9544 | 2 |
| **16569** | | |
| gaBrIEL HARRIS | 102561.9859 | 2 |
| **25959** | | |
| kARen klInE | 104746.0647 | 2 |
| **29591** | | |
| LAuRie hOod | 104340.0737 | 2 |
| **37840** | | |
| pAmElA NoRmAn | 103951.9363 | 2 |

*Hình 241. Kết quả câu truy vấn 14 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**



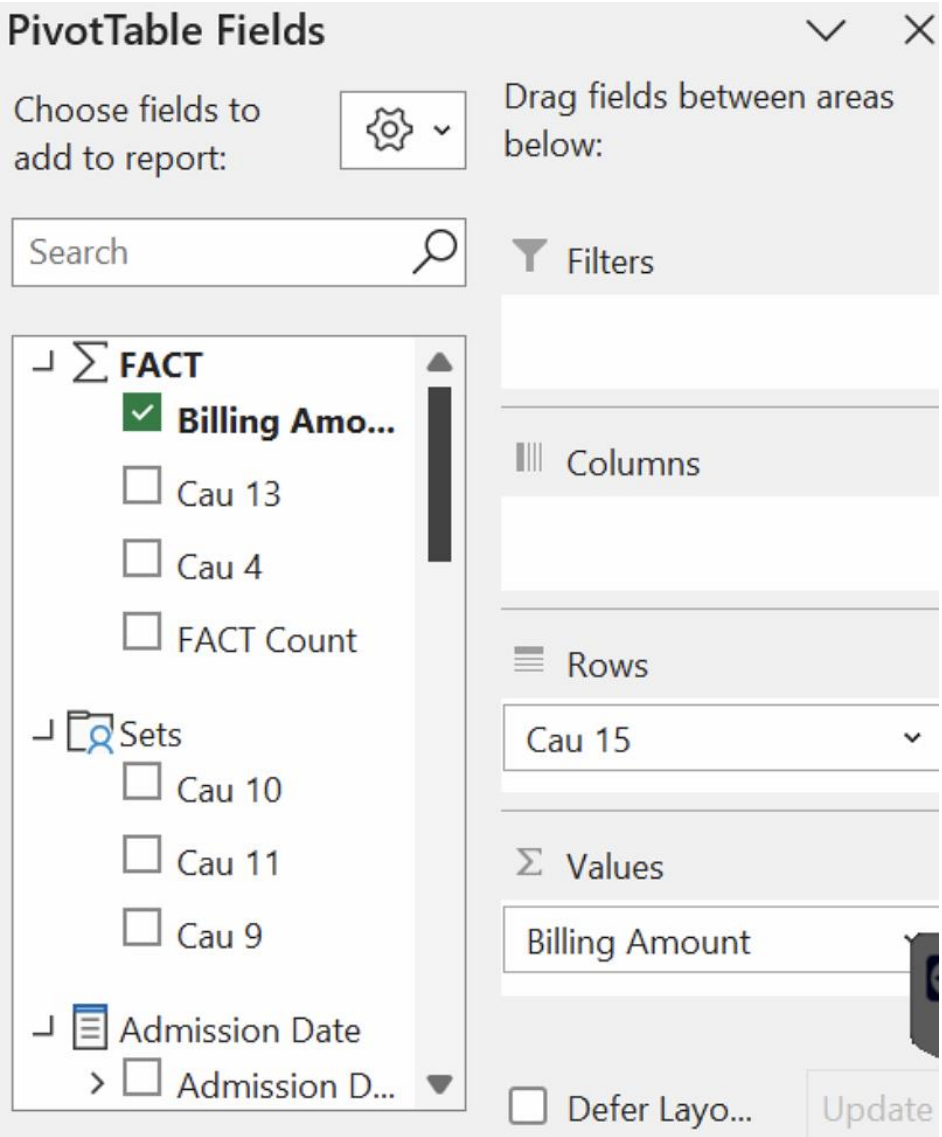

*Hình 242 . Kết quả câu truy vấn 14 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.14 (Kéo thả bằng công cụ BI)

**10.15. Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất.**

- **Thực hiện quá trình phân tích dữ liệu:**

- **Bước 1:** Kéo thuộc tính Cau 15 vào Rows, độ đo Billing Amount vào cột Values.



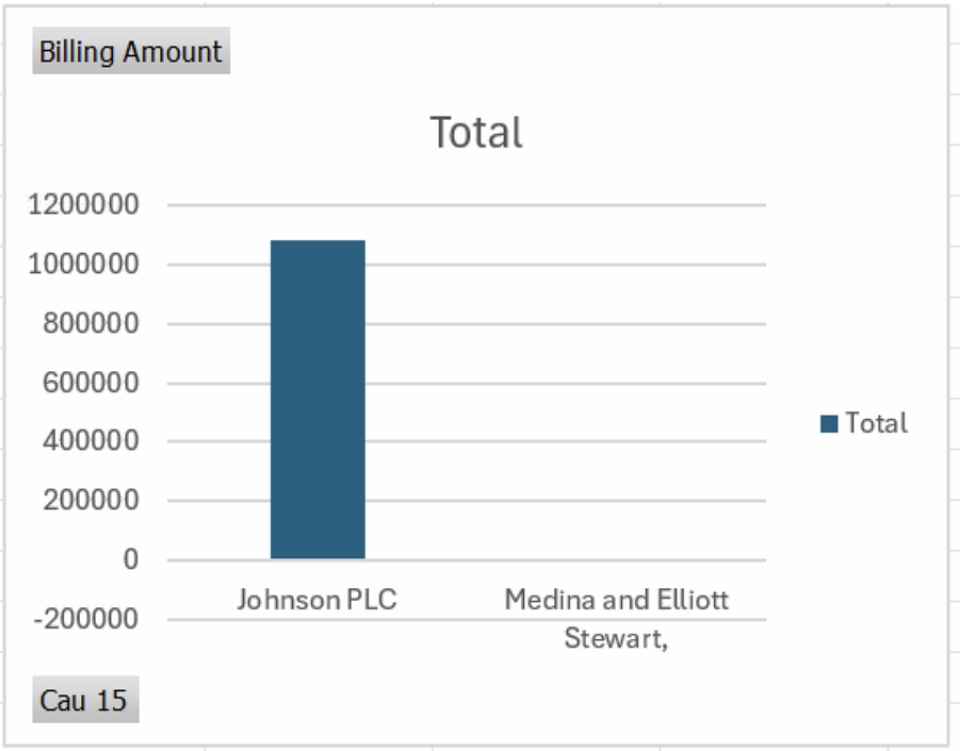

*Hình 243 . Chọn các thuộc tính cần thiết*

- **Kết quả câu truy vấn dạng Table:**

| Row Labels | Billing Amount |
|---|---|
| Johnson PLC | 1084202.693 |
| Medina and Elliott Stewart, | -2633.237163 |

*Hình 244. Kết quả câu truy vấn 15 dạng Table*

- **Kết quả câu truy vấn dạng Chart:**




*Hình 245 . Kết quả câu truy vấn 15 dạng Chart*

- **Nhận xét:** Kết quả giống câu 9.15 (Kéo thả bằng công cụ BI)

## 11. QUÁ TRÌNH PHÂN TÍCH DỮ LIỆU BẰNG NGÔN NGỮ MDX

### 11.1. Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[Billing Amount]} ON COLUMNS,
    NONEMPTY(
        CROSSJOIN(
            [DIM INSURANCE PROVIDER].[Insurance Provider Name].Children,
            [Admission Date].[Year].Children
        )
    ) ON ROWS
FROM [Health Care DW]
```

*Hình 246 . Câu truy vấn ngôn ngữ MDX 1*

- **Kết quả câu truy vấn:**

Messages | Results

| | | Billing Amount |
|---|---|---|
| Aetna | 2019 | 36781597.7361678 |
| Aetna | 2020 | 57538811.7834771 |
| Aetna | 2021 | 56450900.9348828 |
| Aetna | 2022 | 55721455.0015989 |
| Aetna | 2023 | 52529136.0707986 |
| Aetna | 2024 | 19841201.4182693 |
| Blue Cross | 2019 | 38262541.9103262 |
| Blue Cross | 2020 | 58272187.431879 |
| Blue Cross | 2021 | 56734083.729961 |
| Blue Cross | 2022 | 54172231.1490649 |
| Blue Cross | 2023 | 55889840.489461 |
| Blue Cross | 2024 | 19923409.5015492 |
| Cigna | 2019 | 39838339.691158 |
| Cigna | 2020 | 56463627.1646772 |
| Cigna | 2021 | 56308685.4707296 |
| Cigna | 2022 | 57019211.5040334 |
| Cigna | 2023 | 57808356.1331765 |

*Hình 247 . Kết quả câu truy vấn MDX 1*

- **Nhận xét:** Kết quả giống câu 9.1 (Kéo thả bằng công cụ BI) và 10.1 (Công cụ Pivot)

## 11.2. Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[Fact Count]} ON COLUMNS,
  NONEMPTY(
    CROSSJOIN(
        [DIM PATIENT].[Blood Type].Members,
        {[DIM TEST RESULT].[Test Result Name].&[Normal],
        [DIM TEST RESULT].[Test Result Name].&[Abnormal]}
    ))ON ROWS
FROM [Health Care DW]
```

*Hình 248. Câu truy vấn ngôn ngữ MDX 2*

- **Kết quả câu truy vấn:**

Messages | Results

| | | FACT Count |
|---|---|---|
| All | Normal | 18517 |
| All | Abnormal | 18627 |
| A- | Normal | 2335 |
| A- | Abnormal | 2336 |
| A+ | Normal | 2293 |
| A+ | Abnormal | 2333 |
| AB- | Normal | 2362 |
| AB- | Abnormal | 2333 |
| AB+ | Normal | 2314 |
| AB+ | Abnormal | 2308 |
| B- | Normal | 2328 |
| B- | Abnormal | 2348 |
| B+ | Normal | 2276 |
| B+ | Abnormal | 2303 |
| O- | Normal | 2301 |
| O- | Abnormal | 2319 |
| O+ | Normal | 2308 |
| O+ | Abnormal | 2347 |

*Hình 249. Kết quả câu truy vấn MDX 2*

- **Nhận xét:** Kết quả giống câu 9.2 (Kéo thả bằng công cụ BI) và 10.2 (Công cụ Pivot)

### 11.3. Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    [Measures].[Billing Amount] ON COLUMNS,
   {
DrillDownLevel(
DrillDownLevel(
DrillDownLevel(
[Admission Date].[Hierarchy]))
)}
ON ROWS
FROM [Health Care DW]
```

*Hình 250. Câu truy vấn ngôn ngữ MDX 3*

- **Kết quả câu truy vấn:**

Messages | Results

| | Billing Amount |
|---|---|
| All | 1417432043.39526 |
| 2019 | 189849325.140056 |
| 2 | 41538324.6527066 |
| 5 | 17653145.1384676 |
| 6 | 23885179.514239 |
| 3 | 73503597.6643556 |
| 7 | 24771559.1924619 |
| 8 | 25397958.5975822 |
| 9 | 23334079.8743115 |
| 4 | 74807402.8229941 |
| 10 | 26276190.6875709 |
| 11 | 24784507.5287908 |
| 12 | 23746704.6066324 |
| 2020 | 286551589.206669 |
| 1 | 70604176.7813866 |
| 1 | 24371975.1803323 |
| 2 | 22324346.5168337 |
| 3 | 23907855.0842206 |
| 2 | 72400596.5311321 |

*Hình 251 . Kết quả câu truy vấn MDX 3*

- **Nhận xét:** Kết quả giống câu 9.3 (Kéo thả bằng công cụ BI) và 10.3 (Công cụ Pivot)

### 11.4. Đưa ra số lượng bệnh nhân bị mắc bệnh 'Cancer' theo từng năm

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
WITH
    MEMBER [Measures].[Cancer Patient Count] AS
        AGGREGATE(
            [Admission Date].[Year].CurrentMember.Children,
            [Measures].[FACT Count]
        )
SELECT
    {[Measures].[FACT Count],[Measures].[Cancer Patient Count]}
    ON COLUMNS,
    [Admission Date].[Year].Members
    ON ROWS
FROM [Health Care DW]
WHERE ([DIM PATIENT].[Medical Condition].&[Cancer])
```

*Hình 252. Câu truy vấn ngôn ngữ MDX 4*

- **Kết quả câu truy vấn:**

Messages | Results

| | FACT Count | Cancer Patient Count |
|---|---|---|
| All | 9227 | 9227 |
| 2019 | 1283 | (null) |
| 2020 | 1900 | (null) |
| 2021 | 1748 | (null) |
| 2022 | 1825 | (null) |
| 2023 | 1849 | (null) |
| 2024 | 622 | (null) |
| Unknown | (null) | (null) |

*Hình 253. Kết quả câu truy vấn MDX 4*

- **Nhận xét:** Kết quả giống câu 9.4 (Kéo thả bằng công cụ BI) và 10.4 (Công cụ Pivot)

### 11.5. Đưa ra Top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[FACT Count]} ON COLUMNS,
    TOPCOUNT(
        [DIM INSURANCE PROVIDER].[Insurance ID].children,
        5,
        [Measures].[FACT Count]
    )
    * [DIM INSURANCE PROVIDER].[Insurance Provider Name].children ON ROWS
FROM [Health Care DW]
WHERE ([Admission Date].[Year].&[2023])
```

*Hình 254. Câu truy vấn ngôn ngữ MDX 5*

- **Kết quả câu truy vấn:**

Messages | Results

| | | FACT Count |
|---|---|---|
| 23 | Cigna | 2268 |
| 24 | Medicare | 2258 |
| 25 | UnitedHealthcare | 2248 |
| 22 | Blue Cross | 2166 |
| 21 | Aetna | 2086 |

*Hình 255 . Kết quả câu truy vấn MDX 5*

- **Nhận xét:** Kết quả giống câu 9.5 (Kéo thả bằng công cụ BI) và 10.5 (Công cụ Pivot)

### 11.6. Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc Top 5.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[Billing Amount]} ON COLUMNS,
    TOPCOUNT(
        EXCEPT(
            [DIM PATIENT].[Patient ID].Members,
            {[DIM PATIENT].[Patient ID].[All]}
        ),
        5,
        [Measures].[Billing Amount]
    )
    * [DIM PATIENT].[Patient Name].Children ON ROWS
FROM [Health Care DW]
WHERE ([DIM PATIENT].[BLOOD TYPE].&[O+])
```

*Hình 256 . Câu truy vấn ngôn ngữ MDX 6*

- **Kết quả câu truy vấn:**

| | | Billing Amount |
|---|---|---|
| 16115 | Erin HaNSen | 102863.954358062 |
| 20778 | jeNNA SULLIvaN | 101859.201610885 |
| 32990 | MaTthew TyleR | 101525.056684488 |
| 41194 | rOgEr sWANsON | 101244.451786228 |
| 11836 | davld ADams | 101187.37781958 |

*Hình 257. Kết quả câu truy vấn MDX 6*

- **Nhận xét:** Kết quả giống câu 9.6 (Kéo thả bằng công cụ BI) và 10.6 (Công cụ Pivot)

### 11.7. Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    [Measures].[Billing Amount] ON COLUMNS,
    HEAD(
        ORDER(
            [DIM PATIENT].[Patient Id].Children,
            [Measures].[Billing Amount],
            DESC
        ),
        1
    )
    * [DIM PATIENT].[Patient Name].Children
    * [DIM PATIENT].[Gender].Children
    * [DIM PATIENT].[Blood Type].Children
    ON ROWS
FROM [Health Care DW]
WHERE [Admission Date].[Hierarchy].[Quarter].&[2]&[2024]
```

*Hình 258. Câu truy vấn ngôn ngữ MDX 7*

- **Kết quả câu truy vấn:**

Messages | Results

| | | | | Billing Amount |
|---|---|---|---|---|
| 36200 | mR. matTHEw RAmIrEZ | Male | O+ | 99243.6503506949 |

*Hình 259 . Kết quả câu truy vấn MDX 7*

- **Nhận xét:** Kết quả giống câu 9.7 (Kéo thả bằng công cụ BI) và 10.7 (Công cụ Pivot)

**11.8. Liệt kê các bệnh nhân có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024.**

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    [Measures].[Billing amount] ON COLUMNS,
    NON EMPTY
    FILTER(
    [DIM PATIENT].[Patient Id].Children,
        [Measures].[Billing Amount] > 1000
        AND [Measures].[Billing Amount] < 10000
    ) *
    [DIM PATIENT].[Patient Name].Children*
    [DIM PATIENT].[Gender].Children *
    [DIM PATIENT].[BLOOD TYPE].Children
    ON ROWS
FROM [Health Care DW]
WHERE [Admission Date].[YEAR].&[2024]
```

*Hình 260. Câu truy vấn ngôn ngữ MDX 8*

- **Kết quả câu truy vấn:**

Messages | Results

| | | | | Billing Amount |
|---|---|---|---|---|
| 115 | AarOn sALiNas | Female | B+ | 9713.47381295019 |
| 121 | aaroN smIth | Female | B- | 8340.96711266305 |
| 133 | aAROn wEbeR | Female | AB+ | 8767.4138339262 |
| 229 | aDAm cOLLins | Male | O+ | 4194.76977330724 |
| 376 | Adrian hOWARd | Male | AB+ | 5361.39498054342 |
| 426 | ALAn BECk | Female | B- | 1927.73339820995 |
| 447 | alAN kelleR | Male | A- | 6125.98279470799 |
| 503 | AlbeRt SuLliVAn | Male | A+ | 7648.83895041113 |
| 532 | alEJandro AGuIRre | Female | B+ | 9071.63966291391 |
| 644 | AlExanDer FosTeR | Female | AB+ | 4606.60943184244 |
| 658 | aleXAnDer HernaNdeZ | Female | B- | 9116.43063592819 |

*Hình 261. Kết quả câu truy vấn MDX 8*

- **Nhận xét:** Kết quả giống câu 9.8 (Kéo thả bằng công cụ BI) và 10.8 (Công cụ Pivot)

**11.9. Trong mỗi loại bệnh, in ra bác sĩ có số lượng bệnh nhân lớn nhất trong năm 2022.**

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    [Measures].[FACT Count] ON COLUMNS,
    NON EMPTY
    GENERATE(
        [DIM PATIENT].[Medical Condition].children,
        TOPCOUNT(
            CROSSJOIN(
                {[DIM PATIENT].[Medical Condition].CurrentMember},
                [DIM DOCTOR].[Doctor Name].children
            ),
            1,
            ([Measures].[FACT Count])
        )
    ) ON ROWS
FROM [Health Care DW]
WHERE ([Admission Date].[Year].&[2022])
```

*Hình 262. Câu truy vấn ngôn ngữ MDX 9*

- **Kết quả câu truy vấn:**

Messages | Results

| | | FACT Count |
|---|---|---|
| Arthritis | Ryan Ward | 4 |
| Asthma | Amanda Williams | 3 |
| Cancer | Carl Williams | 3 |
| Diabetes | David Smith | 4 |
| Hypertension | Lori Smith | 4 |
| Obesity | Amanda Gonzalez | 3 |

*Hình 263. Kết quả câu truy vấn MDX 9*

- **Nhận xét:** Kết quả giống câu 9.9 (Kéo thả bằng công cụ BI) và 10.9 (Công cụ Pivot)

### 11.10. Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    [Measures].[Fact Count] ON COLUMNS,
    GENERATE(

        DESCENDANTS( [Admission Date].[Quarter].children),
         TOPCOUNT(
            {[Admission Date].[Quarter].CurrentMember*
            [DIM MEDICATION].[Medication name].Children},
            3,
            [Measures].[Fact Count]
         )
    ) ON ROWS
FROM [Health Care DW]
WHERE [Admission Date].[Year].&[2022]
```

*Hình 264. Câu truy vấn ngôn ngữ MDX 10*

- **Kết quả câu truy vấn:**

Messages | Results

| | | FACT Count |
|---|---|---|
| 1 | Lipitor | 592 |
| 1 | Paracetamol | 556 |
| 1 | Ibuprofen | 529 |
| 2 | Lipitor | 561 |
| 2 | Paracetamol | 555 |
| 2 | Penicillin | 551 |
| 3 | Aspirin | 625 |
| 3 | Penicillin | 575 |
| 3 | Lipitor | 552 |
| 4 | Ibuprofen | 605 |
| 4 | Aspirin | 557 |
| 4 | Paracetamol | 556 |

*Hình 265. Kết quả câu truy vấn MDX 10*

- **Nhận xét:** Kết quả giống câu 9.10 (Kéo thả bằng công cụ BI) và 10.10 (Công cụ Pivot)

### 11.11. Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[FACT Count]} ON COLUMNS,
    NONEMPTY(
        GENERATE(
            [DIM PATIENT].[Medical Condition].Children,
            TOPCOUNT(
                [DIM LOCATION].[Hospital Name].Children,
                1,
                [Measures].[FACT Count]
            ) *
            {[DIM PATIENT].[Medical Condition].CurrentMember}
        )
    ) ON ROWS
FROM [Health care DW]
```

*Hình 266. Câu truy vấn ngôn ngữ MDX 11*

- **Kết quả câu truy vấn:**

Messages | Results

| | | FACT Count |
|---|---|---|
| Smith PLC | Arthritis | 12 |
| Johnson Inc | Asthma | 9 |
| Johnson Inc | Cancer | 10 |
| LLC Smith | Diabetes | 11 |
| LLC Smith | Hypertension | 11 |
| Ltd Smith | Obesity | 11 |

*Hình 267. Kết quả câu truy vấn MDX 11*

- **Nhận xét:** Kết quả giống câu 9.11 (Kéo thả bằng công cụ BI) và 10.11 (Công cụ Pivot)

**11.12. Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần.**

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    ORDER(
        FILTER(
            {[DIM LOCATION].[Hospital Name].Members},
            [Measures].[Billing Amount] > 200000
        ),
  [Measures].[Billing Amount] ,
        BDESC
    ) ON ROWS,
    {[Measures].[Billing Amount]} ON COLUMNS
FROM [Health Care DW]
WHERE [DIM PATIENT].[Medical Condition].&[Diabetes]
```

*Hình 268. Câu truy vấn ngôn ngữ MDX 12*

- **Kết quả câu truy vấn:**

Messages | Results

| | Billing Amount |
|---|---|
| All | 238539725.491964 |
| PLC Brown | 304480.795360712 |
| Group Smith | 291522.652933003 |
| Smith Ltd | 247134.645889826 |
| Rodriguez Inc | 243599.128950034 |
| Smith Inc | 227727.130594221 |
| LLC Smith | 218458.262502847 |
| Martin Inc | 217248.212306658 |

*Hình 269. Kết quả câu truy vấn MDX 12*

- **Nhận xét:** Kết quả giống câu 9.12 (Kéo thả bằng công cụ BI) và 10.12 (Công cụ Pivot)

### 11.13. Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
WITH
    MEMBER [Measures].[Blood Type Patient Percentage] AS
        ([Measures].[FACT Count]) /
        SUM(
            [DIM PATIENT].[Blood Type].CurrentMember.Parent,
            [Measures].[FACT Count]
        ),
        FORMAT_STRING = 'Percent'

SELECT
    {[Measures].[Blood Type Patient Percentage]} ON COLUMNS,
    {[DIM PATIENT].[Blood Type].Children} ON ROWS
FROM [Health Care DW]
WHERE ([Admission Date].[Year].&[2023])
```

*Hình 270. Câu truy vấn ngôn ngữ MDX 13*

- **Kết quả câu truy vấn:**

Messages | Results

| | Blood Type Patient Percentage |
|---|---|
| A- | 12.84% |
| A+ | 12.59% |
| AB- | 12.64% |
| AB+ | 12.92% |
| B- | 11.91% |
| B+ | 12.31% |
| O- | 12.23% |
| O+ | 12.56% |
| Unknown | (null) |

*Hình 271. Kết quả câu truy vấn MDX 13*

- **Nhận xét:** Kết quả giống câu 9.13 (Kéo thả bằng công cụ BI) và 10.13 (Công cụ Pivot)

**11.14. Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất hoặc top 3 số lần nhập viện nhiều nhất.**

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
    {[Measures].[Billing Amount], [Measures].[Fact Count]}
    ON COLUMNS,
    {
        UNION(
            TOPCOUNT(
                {[DIM PATIENT].[Patient ID].Children
                * [DIM PATIENT].[Patient Name].Children},
                5,
                [Measures].[Billing Amount]
            ),
            TOPCOUNT(
                {[DIM PATIENT].[Patient ID].Children
                * [DIM PATIENT].[Patient Name].Children},
                3,
                [Measures].[Fact Count]
            )
        )
    } ON ROWS
FROM [Health Care DW]
```

*Hình 272 . Câu truy vấn ngôn ngữ MDX 14*

- **Kết quả câu truy vấn:**

Messages | Results

| | | Billing Amount | FACT Count |
|---|---|---|---|
| 25959 | kARen klInE | 104746.064748484 | 2 |
| 29591 | LAuRie hOod | 104340.073707113 | 2 |
| 37840 | pAmElA NoRmAn | 103951.936270533 | 2 |
| 16115 | Erin HaNSen | 102863.954358062 | 2 |
| 16569 | gaBrIEL HARRIS | 102561.985891421 | 2 |
| 5 | AAron ArCHER | 21204.1543708376 | 2 |
| 24 | aarOn cArR | 58085.9236555436 | 2 |
| 32 | aaRON DaltoN | 4140.51895309554 | 2 |

*Hình 273. Kết quả câu truy vấn MDX 14*

- **Nhận xét:** Kết quả giống câu 9.14 (Kéo thả bằng công cụ BI) và 10.14 (Công cụ Pivot)

**11.15. Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất.**

- **Câu lệnh truy vấn bằng ngôn ngữ MDX:**

```
SELECT
  {[Measures].[Billing Amount]} ON COLUMNS,
  UNION(
  TOPCOUNT(
    [DIM LOCATION].[Hospital Name].Children,
    1,
    [Measures].[Billing Amount]),
  BOTTOMCOUNT(
    [DIM LOCATION].[Hospital Name].Children,
    1,
    [Measures].[Billing Amount]) )
  ON ROWS
FROM [Health Care DW]
```

*Hình 274. Câu truy vấn ngôn ngữ MDX 15*

- **Kết quả câu truy vấn:**

Messages | Results

| | Billing Amount |
|---|---|
| Johnson PLC | 1084202.69296492 |
| Medina and Elliott Stewart, | -2633.23716258996 |

*Hình 275 . Kết quả câu truy vấn MDX 15*

- **Nhận xét:** Kết quả giống câu 9.15 (Kéo thả bằng công cụ BI) và 10.15 (Công cụ Pivot)

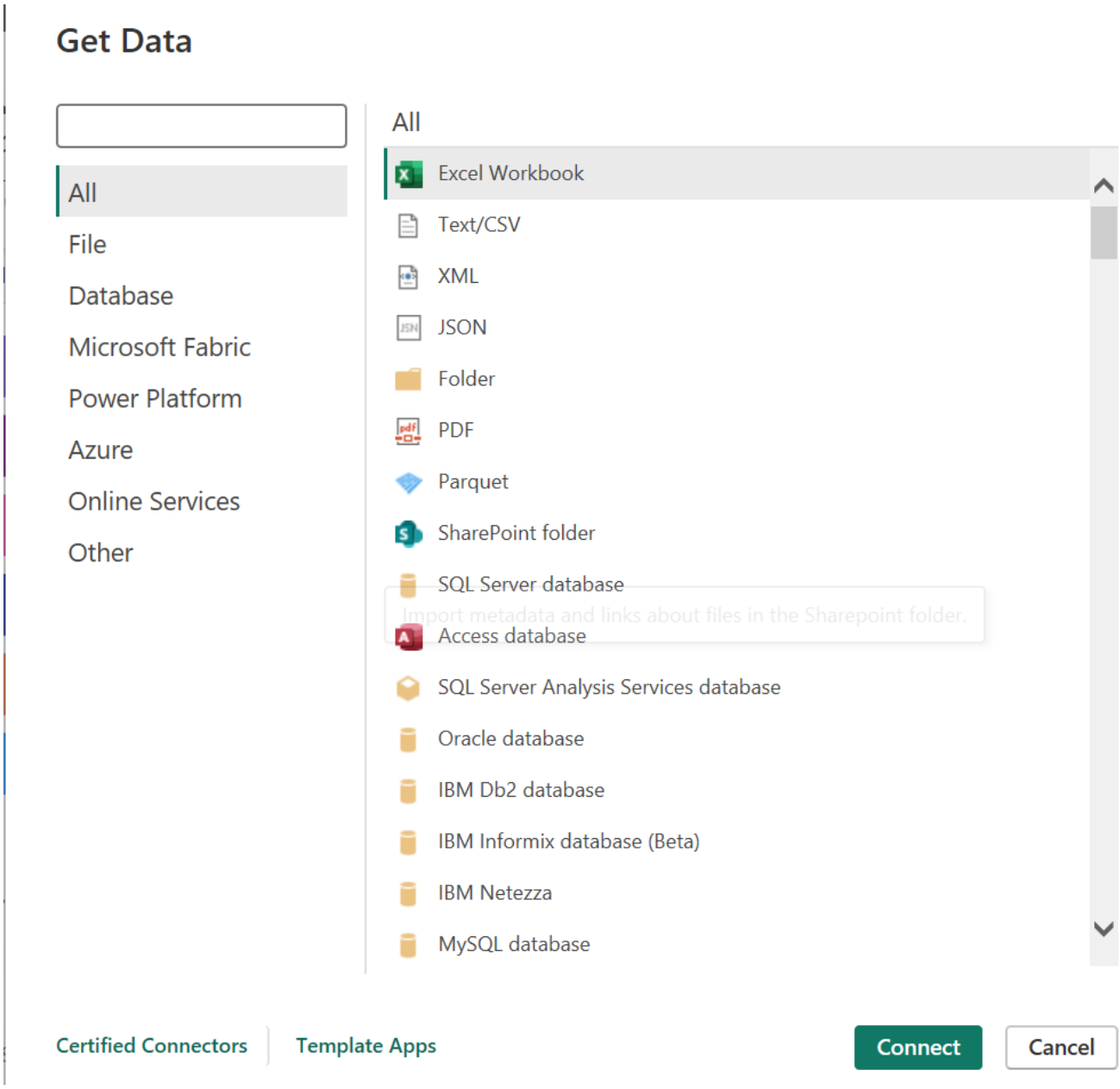
# CHƯƠNG 4. REPORT VỚI POWER BI

## 1. QUÁ TRÌNH LẬP BÁO BIỂU BẰNG CÔNG CỤ POWER BI

Tạo mới một project và lưu vào thư mục.

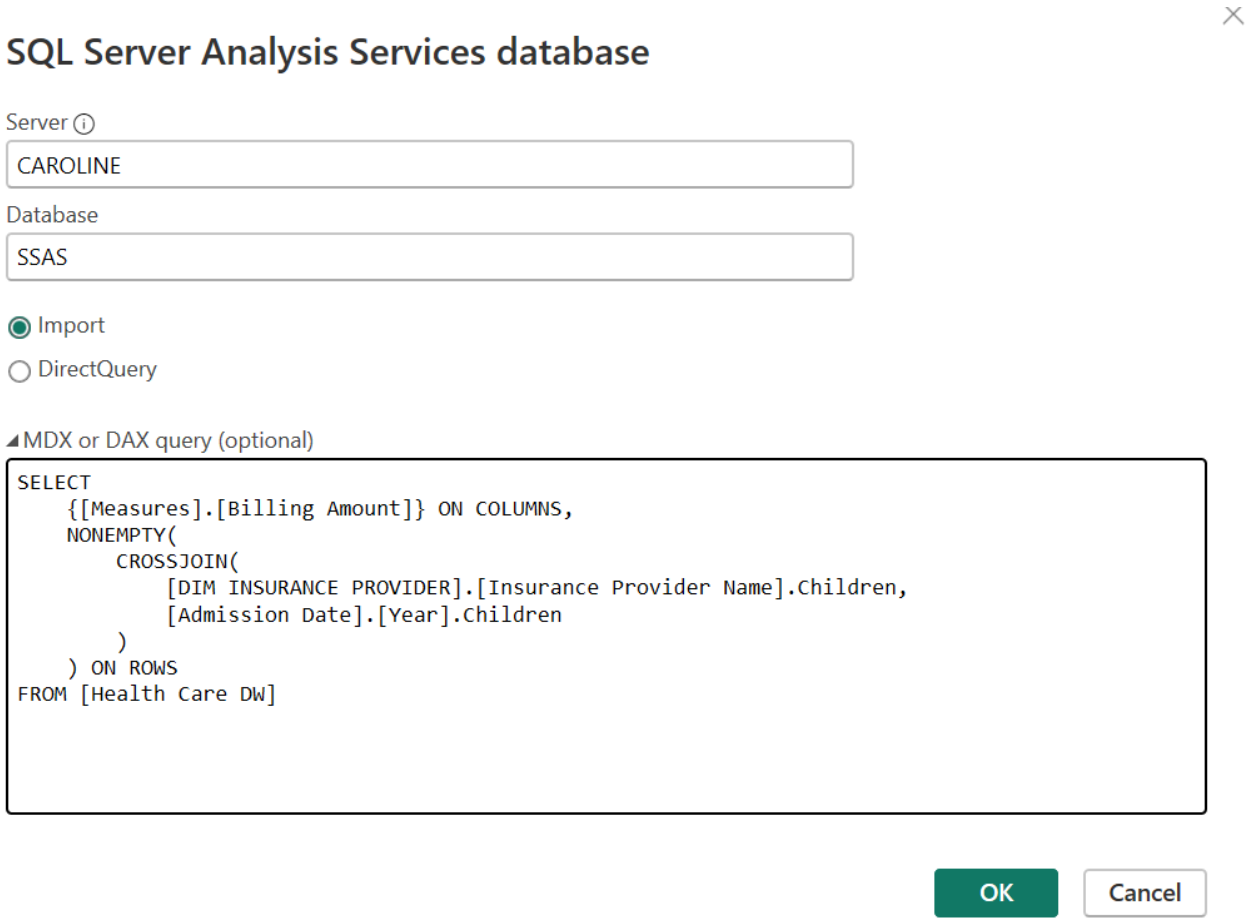
### 1.1. REPORT 1

- **Nội dung ( câu truy vấn 1 )** : Tổng chi phí điều trị của bệnh nhân có bảo hiểm từ các nhà cung cấp bảo hiểm khác nhau theo từng năm
- **Các bước thức hiện :**
  - Bước 1 : Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 276 . Get Data Report 1*

- Bước 2: Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.




*Hình 277. Import câu lệnh MDX 1*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

CAROLINE: SSAS

| [DIM INSURANCE PROVIDER].[Insurance Provider Name] | [Admission Date].[Year].[Year].[MEMBER_CAPTION] | [Measures].[Billing Amount] |
|---|---|---|
| Aetna | 2019 | 36781597.74 |
| Aetna | 2020 | 57538811.78 |
| Aetna | 2021 | 56450900.93 |
| Aetna | 2022 | 55721455 |
| Aetna | 2023 | 52529136.07 |
| Aetna | 2024 | 19841201.42 |
| Blue Cross | 2019 | 38262541.91 |
| Blue Cross | 2020 | 58272187.43 |
| Blue Cross | 2021 | 56734083.73 |
| Blue Cross | 2022 | 54172231.15 |
| Blue Cross | 2023 | 55889840.49 |
| Blue Cross | 2024 | 19923409.5 |
| Cigna | 2019 | 39838339.69 |
| Cigna | 2020 | 56463627.16 |
| Cigna | 2021 | 56308685.47 |
| Cigna | 2022 | 57019211.5 |
| Cigna | 2023 | 57808356.13 |
| Cigna | 2024 | 19701125.3 |
| Medicare | 2019 | 36443188.96 |
| Medicare | 2020 | 57772568.16 |

The data in the preview has been truncated due to size limits.

Load | Transform Data | Cancel

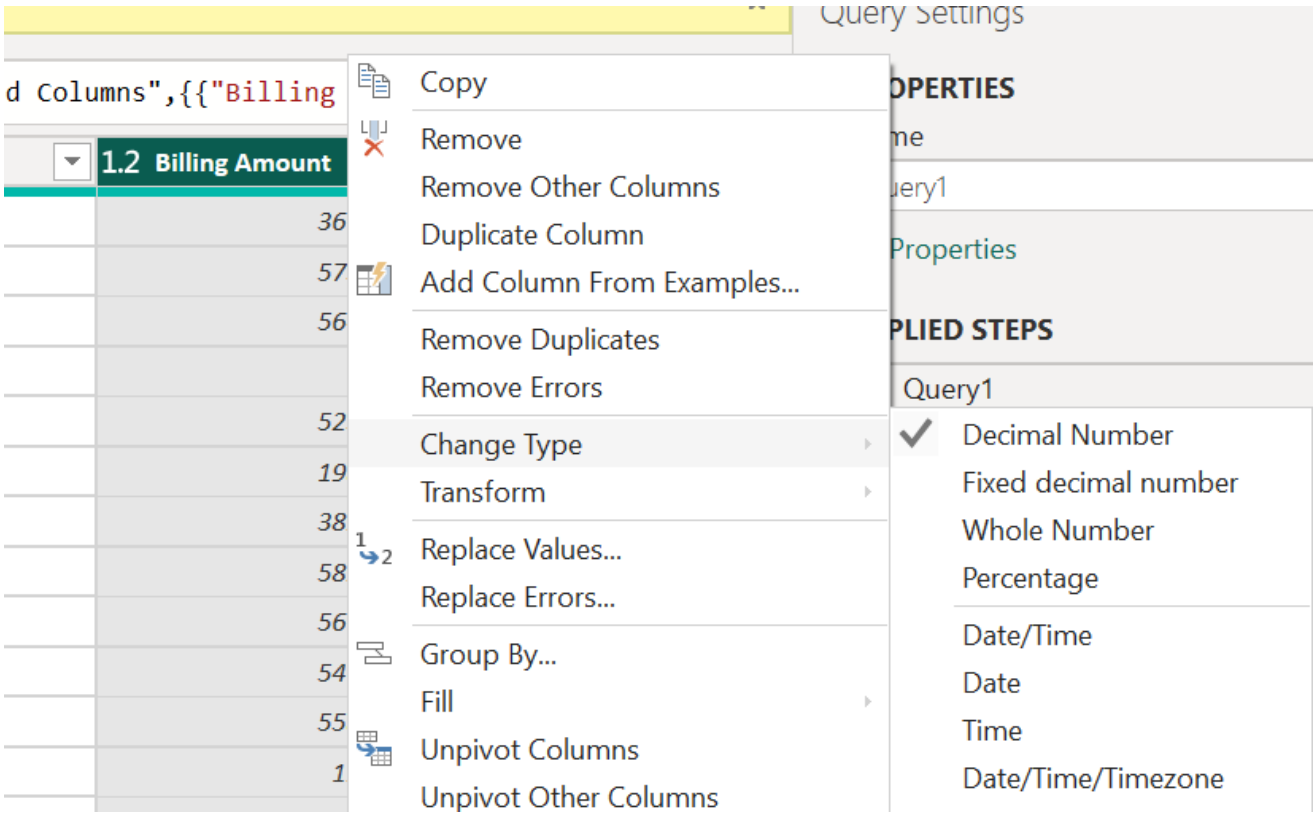
*Hình 278 . Transform Data*

- Bước 4: Thay đổi tên cho các cột thuộc tính

| | Insurance Provider Name | [Admission Date].[Year] | Billing Amount |
|---|---|---|---|
| 1 | Aetna | 2019 | 36781597.74 |
| 2 | Aetna | 2020 | 57538811.78 |
| 3 | Aetna | 2021 | 56450900.93 |
| 4 | Aetna | 2022 | 55721455 |
| 5 | Aetna | 2023 | 52529136.07 |
| 6 | Aetna | 2024 | 19841201.42 |
| 7 | Blue Cross | 2019 | 38262541.91 |
| 8 | Blue Cross | 2020 | 58272187.43 |
| 9 | Blue Cross | 2021 | 56734083.73 |

*Hình 279. Thay đổi tên cho các thuộc tính*

- Bước 5: Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu



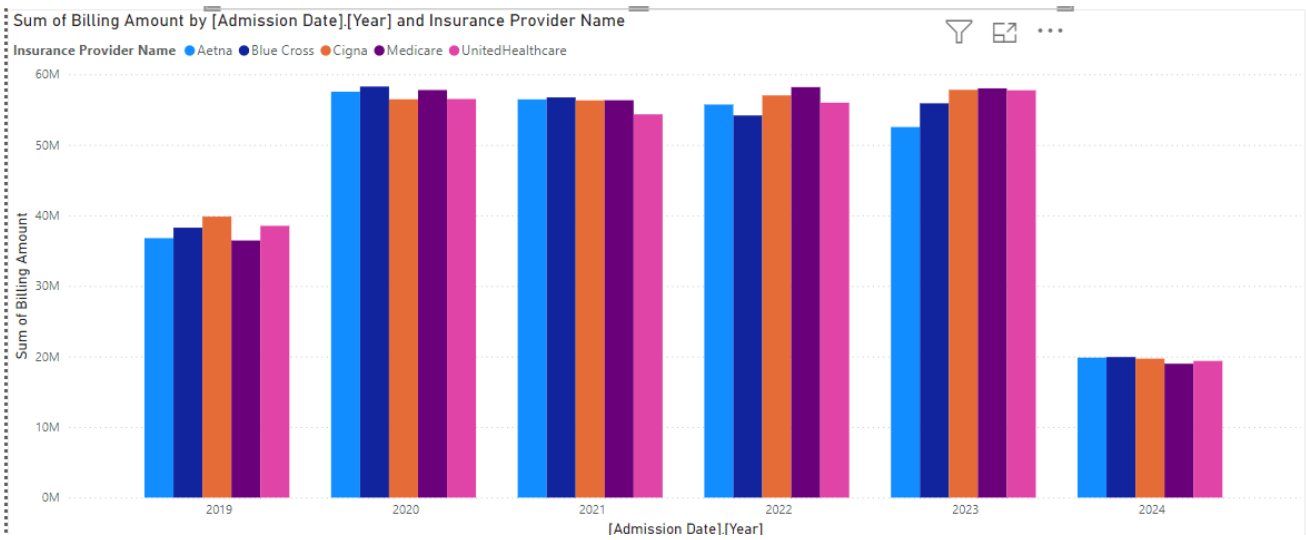

*Hình 280. Close & Apply*

- Bước 6 : Chọn Visualization matrix và phân bố các thuộc tính như hình dưới

| [Admission Date].[Year] | Aetna | Blue Cross | Cigna | Medicare | UnitedHealthcare | Total |
|---|---|---|---|---|---|---|
| 2019 | 36,781,597.74 | 38,262,541.91 | 39,838,339.69 | 36,443,188.96 | 38,523,656.84 | **189,849,325.14** |
| 2020 | 57,538,811.78 | 58,272,187.43 | 56,463,627.16 | 57,772,568.16 | 56,504,394.66 | **286,551,589.21** |
| 2021 | 56,450,900.93 | 56,734,083.73 | 56,308,685.47 | 56,328,362.55 | 54,333,848.60 | **280,155,881.29** |
| 2022 | 55,721,455.00 | 54,172,231.15 | 57,019,211.50 | 58,186,485.98 | 55,982,953.92 | **281,082,337.56** |
| 2023 | 52,529,136.07 | 55,889,840.49 | 57,808,356.13 | 58,010,543.11 | 57,739,377.42 | **281,977,253.22** |
| 2024 | 19,841,201.42 | 19,923,409.50 | 19,701,125.30 | 18,979,609.36 | 19,370,311.39 | **97,815,656.98** |
| **Total** | **278,863,102.95** | **283,254,294.21** | **287,139,345.27** | **285,720,758.13** | **282,454,542.84** | **1,417,432,043.40** |

*Hình 281 . Visualization matrix của Report 1*

- Bước 7 : Chọn Visualization Clustered Column Chart và phân bố các thuộc tính như hình dưới



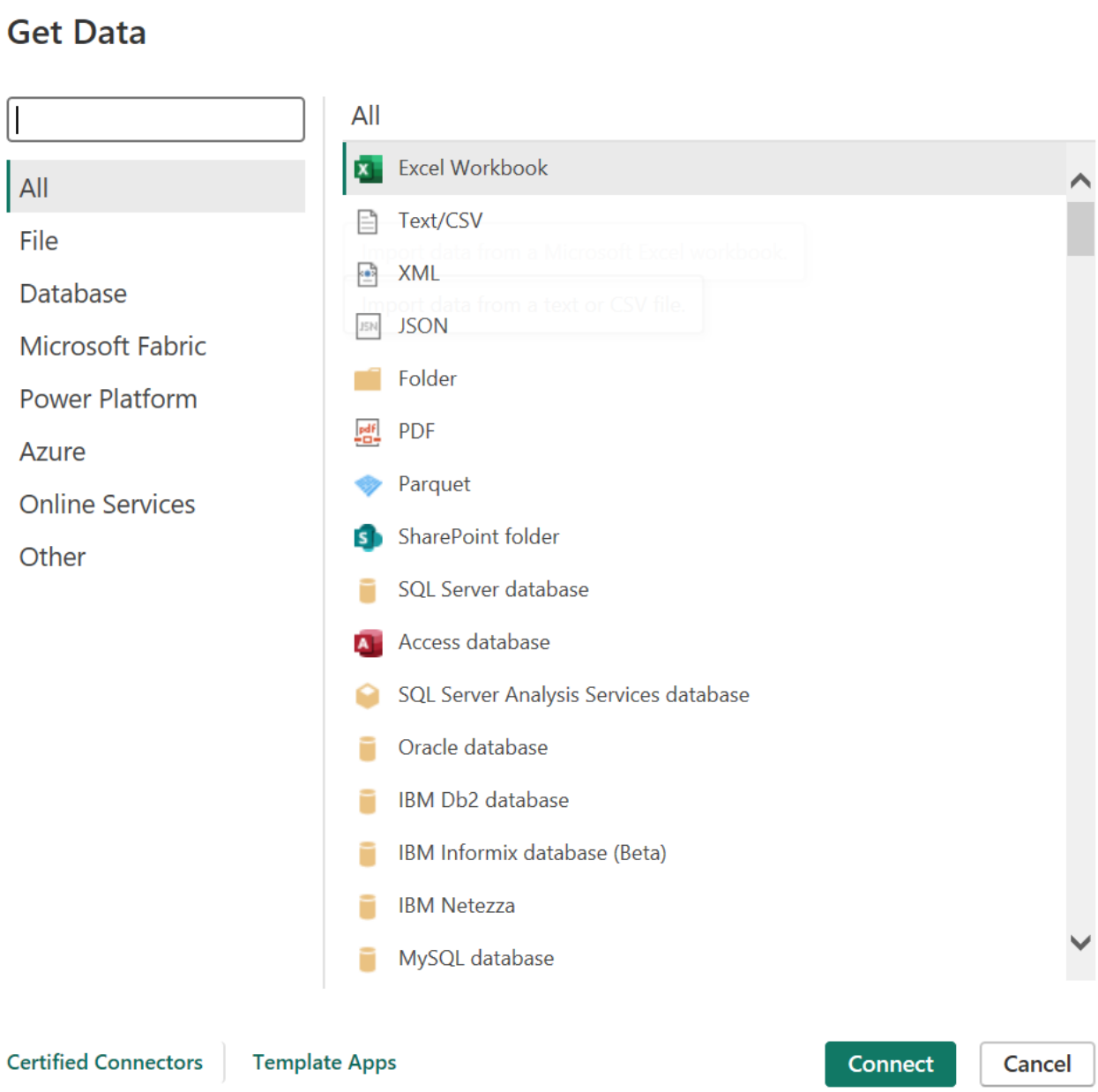

*Hình 282 . Visualization Clustered Column Chart của Report 1*

**Ý nghĩa thực tế:**

1. **Quản lý hệ thống bảo hiểm y tế:**
   - Cigna và Medicare đóng vai trò quan trọng nhất, chiếm tỷ trọng lớn trong hệ thống bảo hiểm. Điều này đòi hỏi cần có chiến lược duy trì chất lượng dịch vụ và tối ưu hóa chi phí cho hai nhà cung cấp này.
2. **Xu hướng giảm chi phí năm 2024:**
   - Sự giảm mạnh chi phí vào 2024 có thể liên quan đến các yếu tố:
     - Dữ liệu chưa đầy đủ cho năm 2024.
     - Thay đổi trong chính sách y tế hoặc nhu cầu y tế giảm.
     - Ảnh hưởng của các yếu tố bên ngoài như kinh tế hoặc thay đổi cấu trúc bảo hiểm.
3. **Cân bằng giữa các nhà cung cấp bảo hiểm:**
   - Blue Cross và Aetna có sự tăng trưởng ổn định nhưng vẫn còn cách xa Medicare và Cigna. Điều này có thể tạo cơ hội để các nhà cung cấp này mở rộng phạm vi hoặc cải thiện dịch vụ.

## 1.2. REPORT 2

- **Nội dung ( Câu truy vấn 2) :** Thống kê số lượng bệnh nhân có kết quả xét nghiệm " Normal" và " Abnormal" theo từng nhóm máu.
- **Các bước thực hiện :**
  - Bước 1: Tạo dataset bằng cách chọn “Get data” trên menu sau đó chọn “**SQL Server Analysis Services database**”.




*Hình 283 . Get Data Report 2*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    {[Measures].[Fact Count]} ON COLUMNS,
        NONEMPTY(
    CROSSJOIN(
        [DIM PATIENT].[Blood Type].Members,
        {[DIM TEST RESULT].[Test Result Name].&[Normal], [DIM TEST RESULT].[Test Result Name].&[
    ))ON ROWS
FROM [Health Care DW]
```

OK Cancel

*Hình 284. Import câu lệnh MDX 2*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM PATIENT].[Blood Type].[Blood Type].[MEMBER_CA | [DIM TEST RESULT].[Test Result Name].[Test Result Nam | [Measures].[FACT Count] |
|---|---|---|
| null | Normal | 18517 |
| null | Abnormal | 18627 |
| A- | Normal | 2335 |
| A- | Abnormal | 2336 |
| A+ | Normal | 2293 |
| A+ | Abnormal | 2333 |
| AB- | Normal | 2362 |
| AB- | Abnormal | 2333 |
| AB+ | Normal | 2314 |
| AB+ | Abnormal | 2308 |
| B- | Normal | 2328 |
| B- | Abnormal | 2348 |
| B+ | Normal | 2276 |
| B+ | Abnormal | 2303 |
| O- | Normal | 2301 |
| O- | Abnormal | 2319 |
| O+ | Normal | 2308 |
| O+ | Abnormal | 2347 |

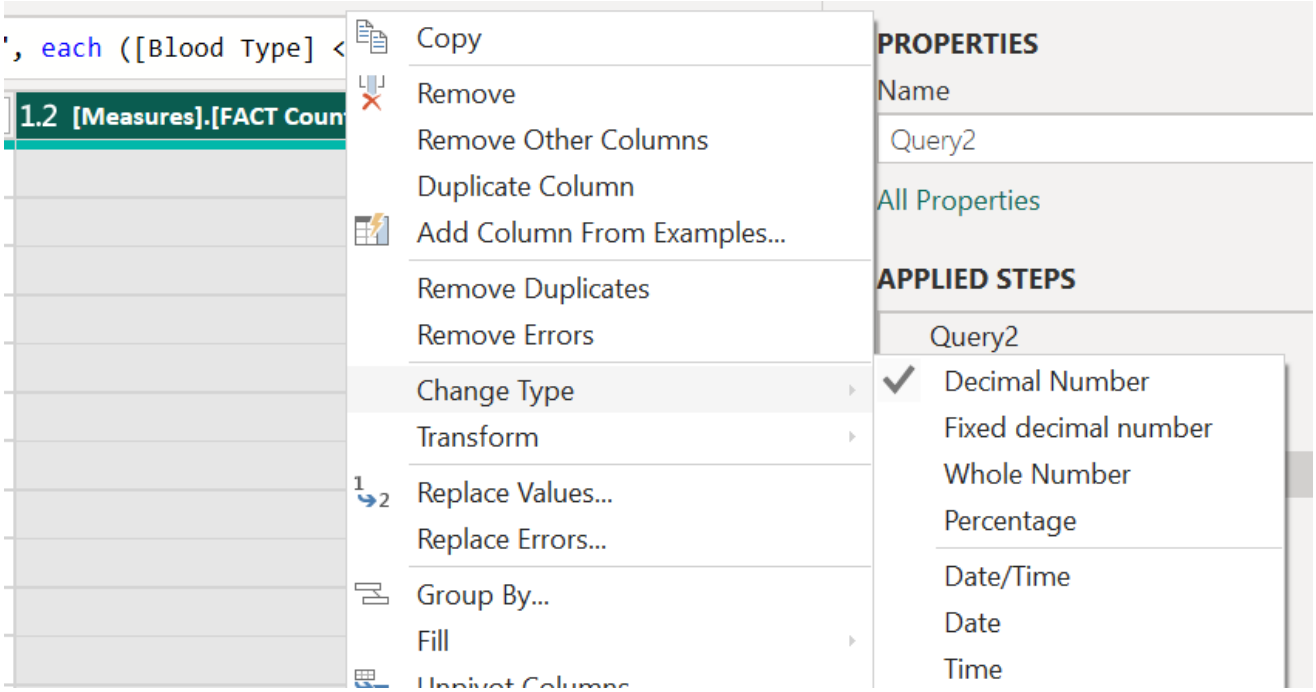
*Hình 285 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

| | Blood Type | Test Result Name | 1.2 [Measures].[FACT Count] |
|---|---|---|---|
| 1 | A- | Normal | 2335 |
| 2 | A- | Abnormal | 2336 |
| 3 | A+ | Normal | 2293 |
| 4 | A+ | Abnormal | 2333 |
| 5 | AB- | Normal | 2362 |
| 6 | AB- | Abnormal | 2333 |
| 7 | AB+ | Normal | 2314 |
| 8 | AB+ | Abnormal | 2308 |
| 9 | B- | Normal | 2328 |
| 10 | B- | Abnormal | 2348 |
| 11 | B+ | Normal | 2276 |
| 12 | B+ | Abnormal | 2303 |

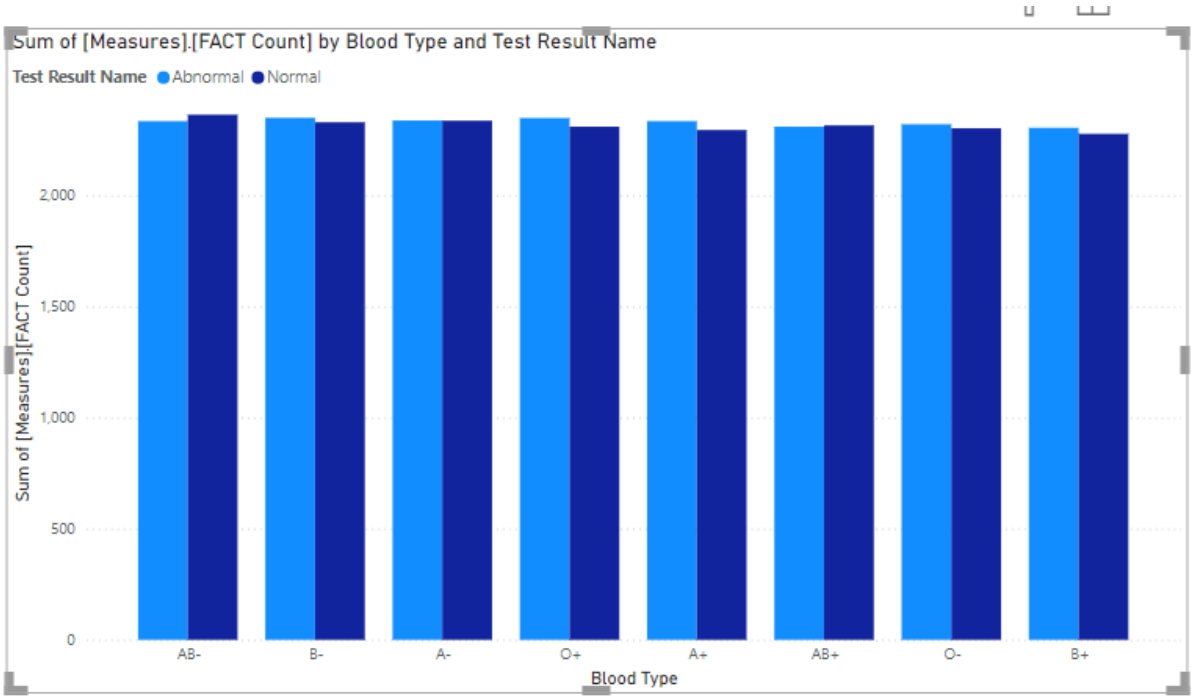
*Hình 286 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Measure.FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 287 . Close & Apply*

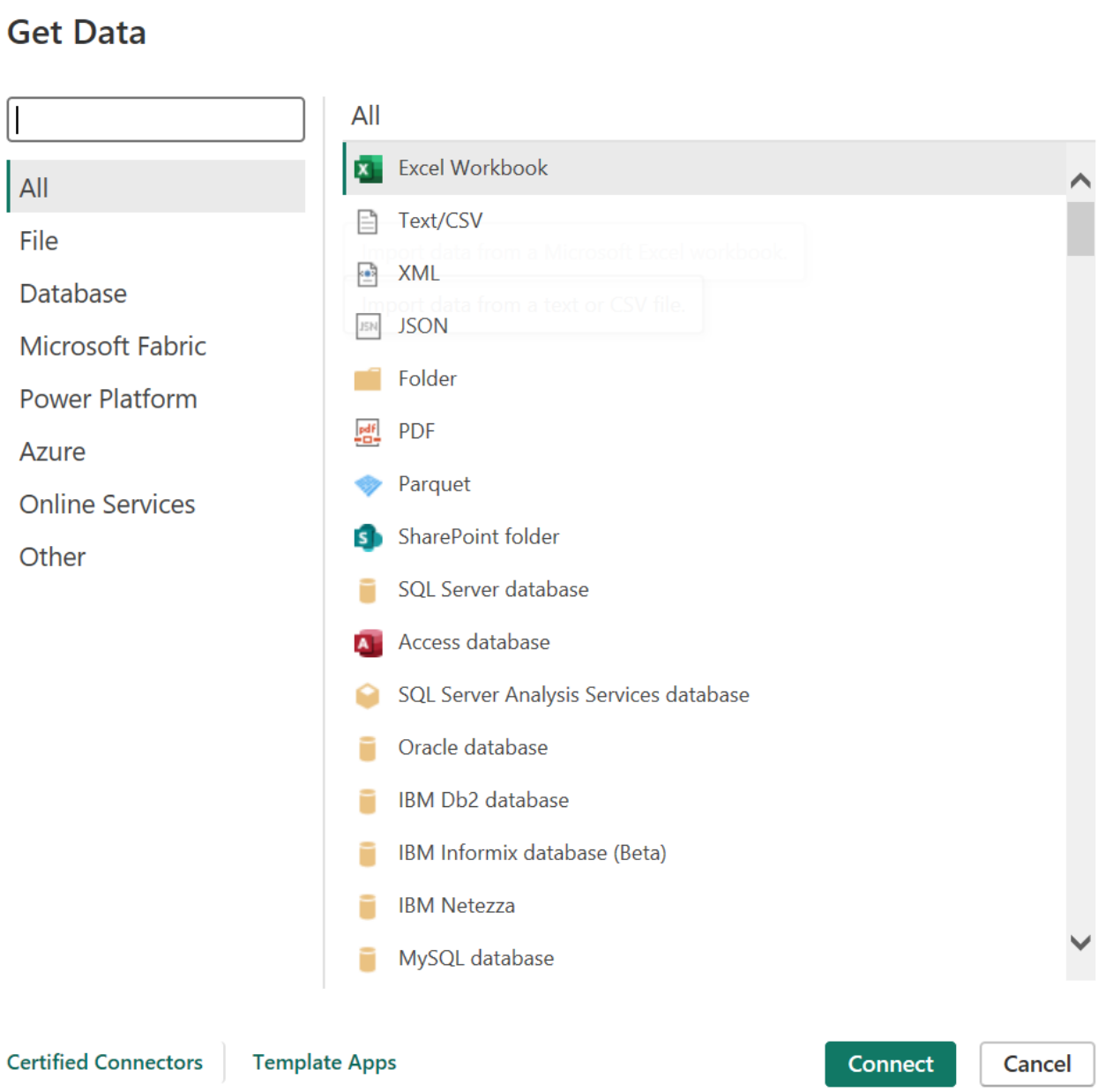
- Bước 6 : Chọn Visualization Clustered Column Chart và phân bố các thuộc tính như hình dưới




*Hình 288 . Visualization Clustered Column Chart của Report 2*

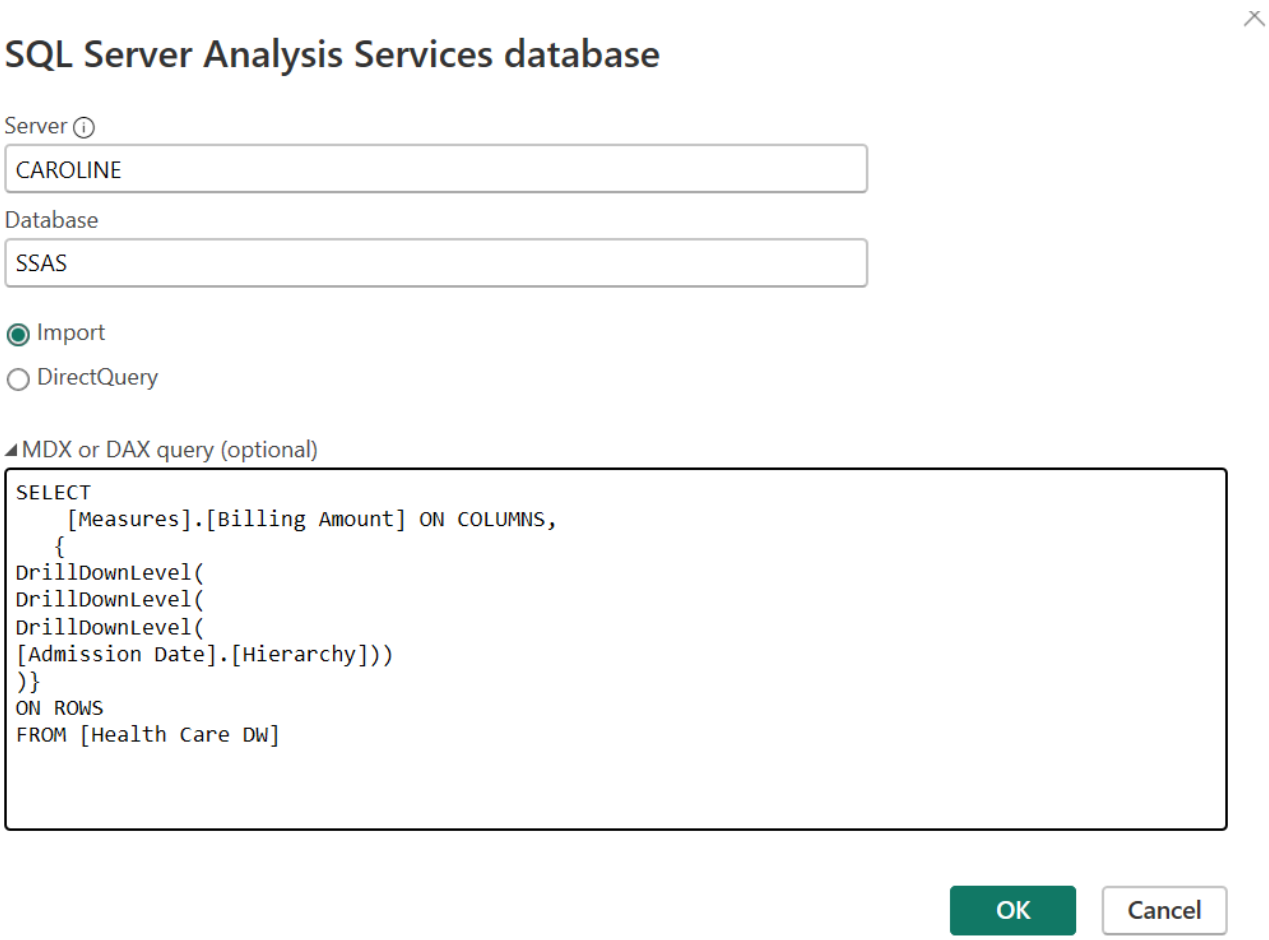
## 1.3. REPORT 3

- **Nội dung ( Câu truy vấn 3 )** : Liệt kê tổng số tiền thanh toán theo từng tháng từng quý theo từng năm.
- **Các bước thực hiện :**
  - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 289 . Get Data Report 3*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.




*Hình 290. Import câu lệnh MDX 3*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

CAROLINE: SSAS

| [Admission Date].[Hierarchy].[Year].[MEMBER_CAPTION | [Admission Date].[Hierarchy].[Quarter].[MEMBER_CAPT | [Admission Date].[Hierarchy].[Month].[MEMBER |
|---|---|---|
| null | null | |
| 2019 | null | |
| 2019 | 2 | |
| 2019 | 2 | 5 |
| 2019 | 2 | 6 |
| 2019 | 3 | |
| 2019 | 3 | 7 |
| 2019 | 3 | 8 |
| 2019 | 3 | 9 |
| 2019 | 4 | |
| 2019 | 4 | 10 |
| 2019 | 4 | 11 |
| 2019 | 4 | 12 |
| 2020 | null | |
| 2020 | 1 | |
| 2020 | 1 | 1 |
| 2020 | 1 | 2 |
| 2020 | 1 | 3 |
| 2020 | 2 | |
| 2020 | 2 | 4 |

The data in the preview has been truncated due to size limits.

Load | Transform Data | Cancel

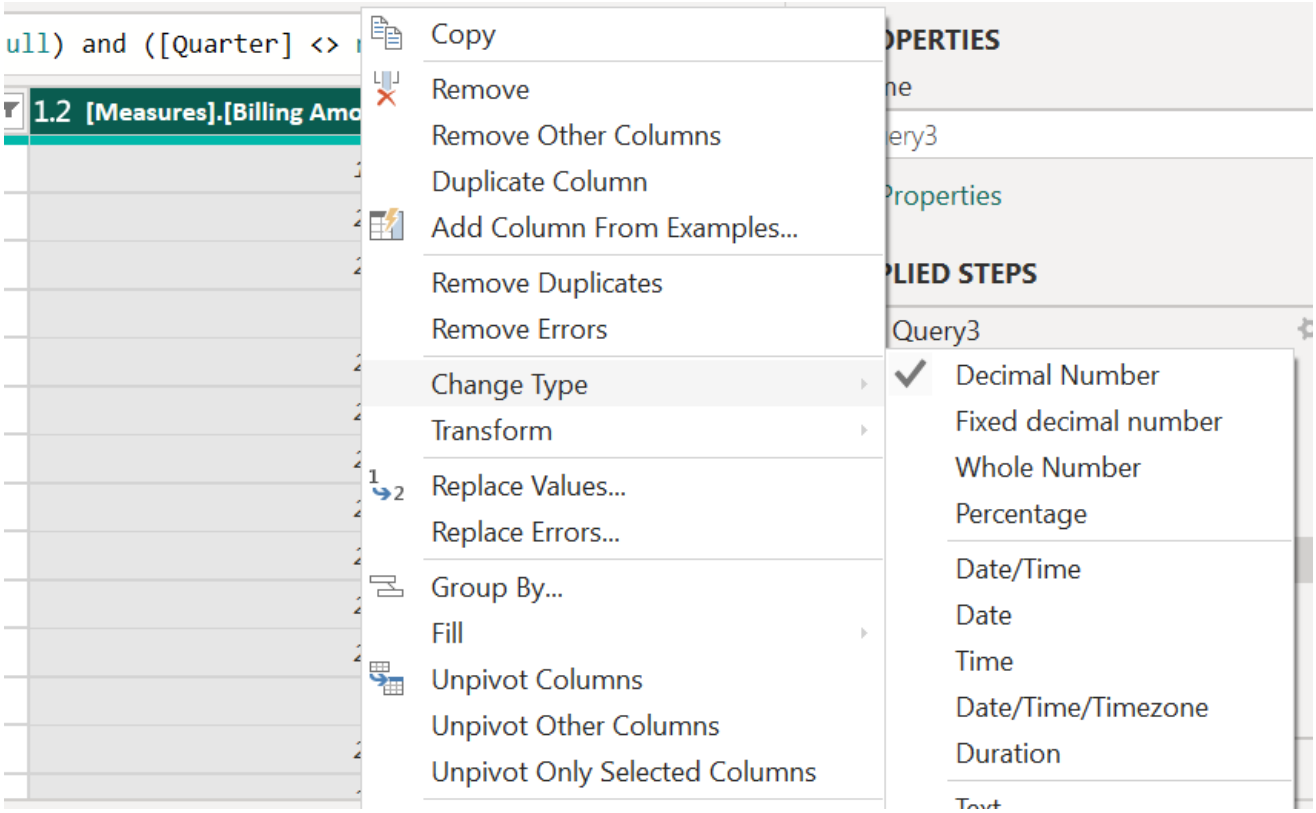
*Hình 291 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

| Year | Quarter | Month | [Measures].[Billing Amount] |
|---|---|---|---|
| 2019 | 2 | 5 | 17653145.14 |
| 2019 | 2 | 6 | 23885179.51 |
| 2019 | 3 | 7 | 24771559.19 |
| 2019 | 3 | 8 | 25397958.6 |
| 2019 | 3 | 9 | 23334079.87 |
| 2019 | 4 | 10 | 26276190.69 |
| 2019 | 4 | 11 | 24784507.53 |
| 2019 | 4 | 12 | 23746704.61 |
| 2020 | 1 | 1 | 24371975.18 |
| 2020 | 1 | 2 | 22324346.52 |
| 2020 | 1 | 3 | 23907855.08 |
| 2020 | 2 | 4 | 23129509.1 |

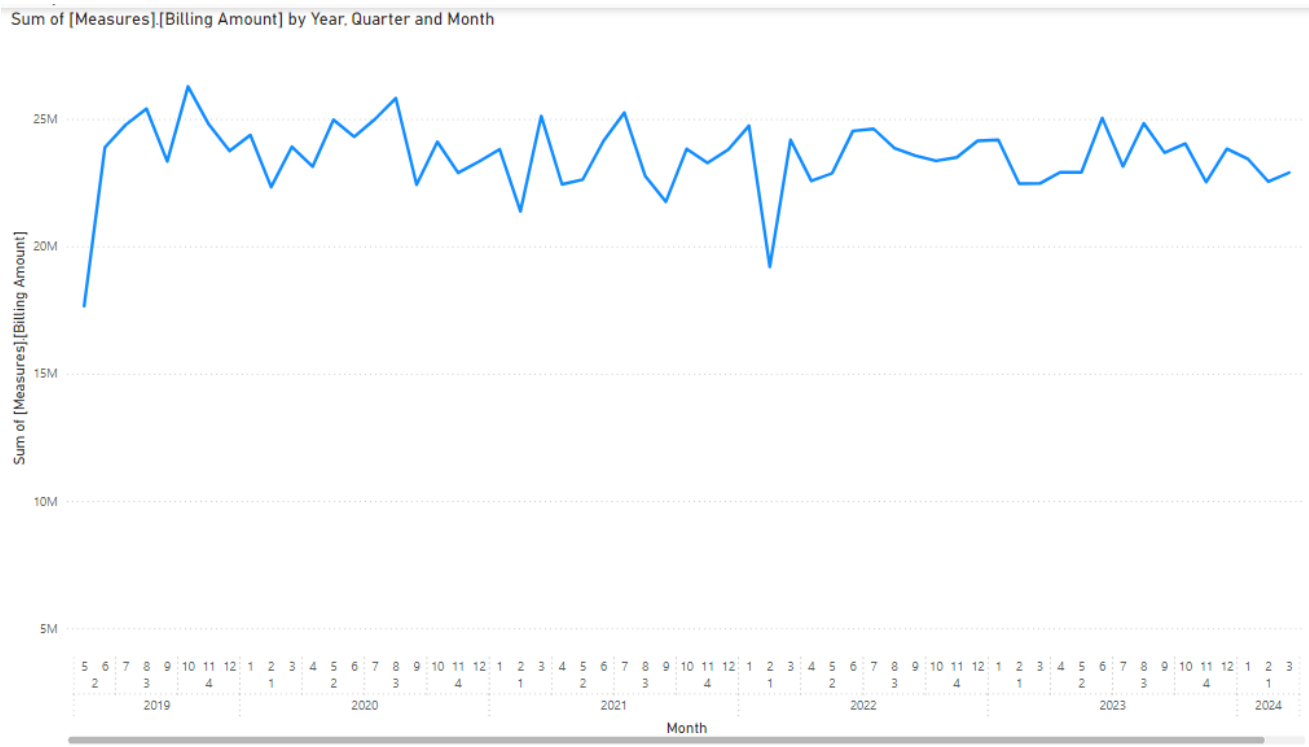
*Hình 292 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 293 . Close & Apply*

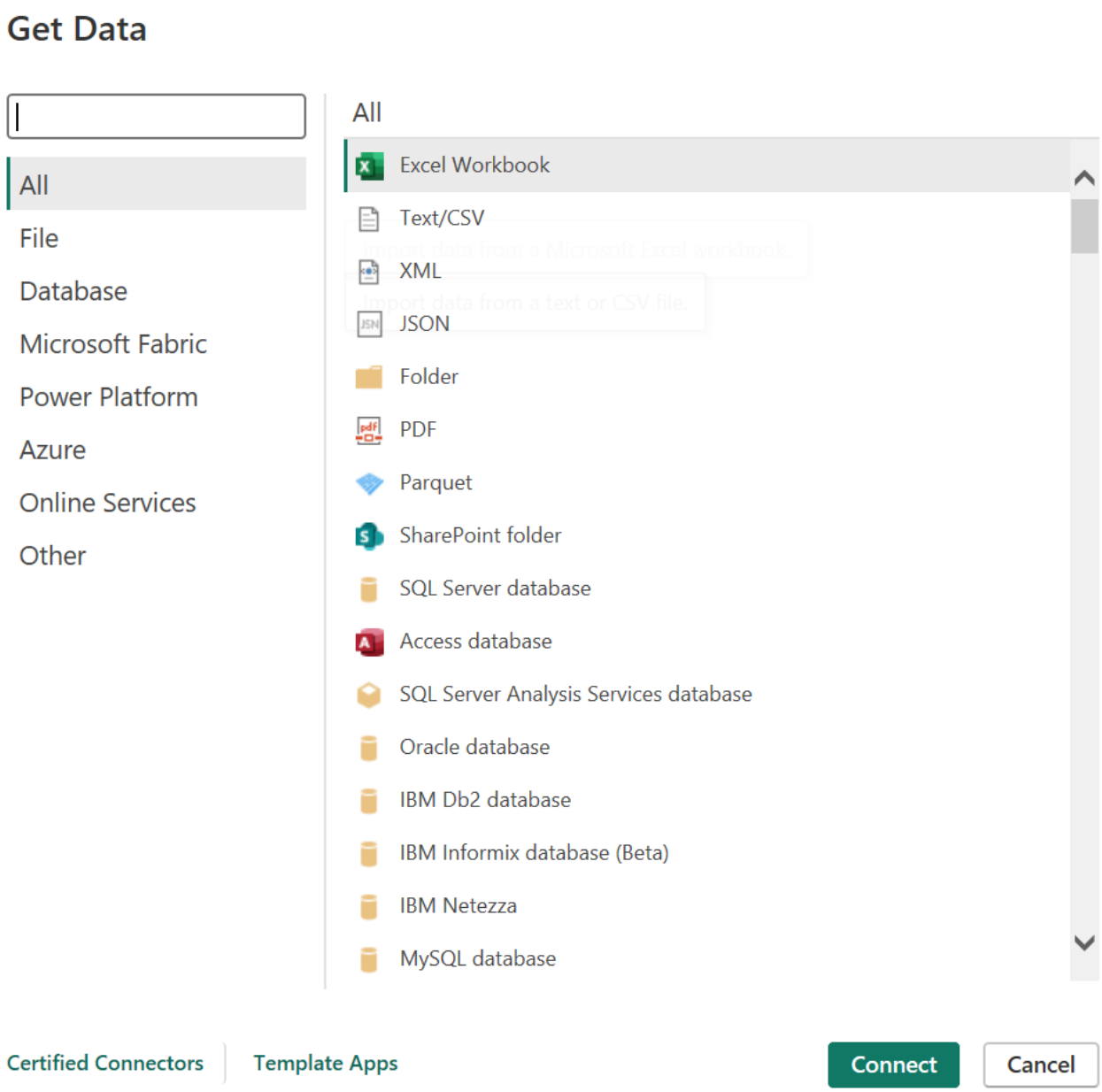
- Bước 6 : Chọn Visualization Line Chart và phân bố các thuộc tính như hình dưới




*Hình 294 . Visualization Line  Chart của Report 3*

## 1.4. REPORT 4

- Nội dung (Câu truy vấn 4 ) : Đưa ra số lượng bệnh nhân bị mắc bệnh ‘Cancer' theo từng năm
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn “Get data” trên menu sau đó chọn “**SQL Server Analysis Services database**”.




*Hình 295 . Get Data Report 4*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
WITH
    MEMBER [Measures].[Cancer Patient Count] AS
        AGGREGATE(
          [Admission Date].[Year].CurrentMember.Children,
            [Measures].[FACT Count]
        )
SELECT
    {[Measures].[FACT Count],[Measures].[Cancer Patient Count]} ON COLUMNS,
    [Admission Date].[Year].Members ON ROWS
FROM [Health Care DW]
WHERE ([DIM PATIENT].[Medical Condition].&[Cancer])
```

OK Cancel

*Hình 296 . Import câu lệnh MDX 4*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [Admission Date].[Year].[Year].[MEMBER_CAPTION] | [Measures].[FACT Count] | [Measures].[Cancer Patient Count] |
|---|---|---|
| null | 9227 | 9227 |
| 2019 | 1283 | null |
| 2020 | 1900 | null |
| 2021 | 1748 | null |
| 2022 | 1825 | null |
| 2023 | 1849 | null |
| 2024 | 622 | null |
| Unknown | null | null |

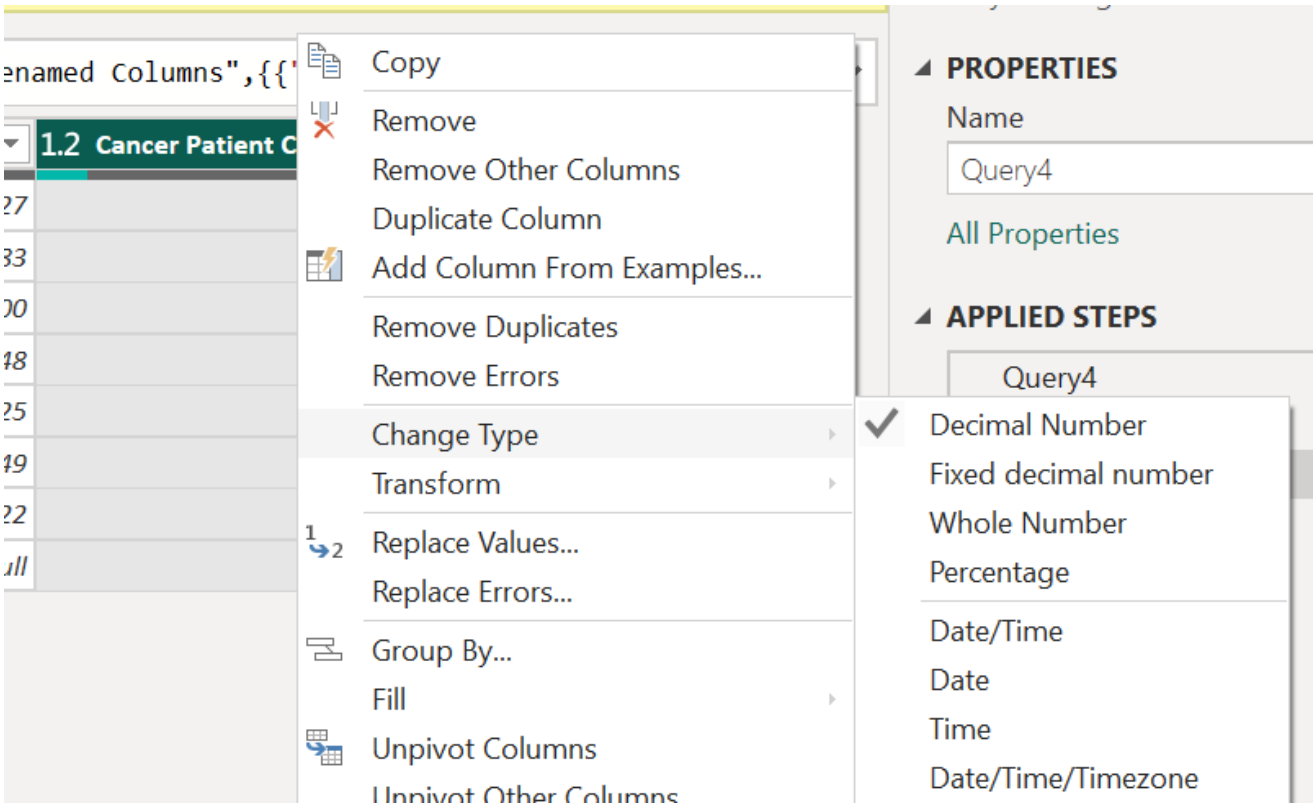
*Hình 297 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

| | Year | FACT Count | Cancer Patient Count |
|---|---|---|---|
| 1 | null | 9227 | 9227 |
| 2 | 2019 | 1283 | null |
| 3 | 2020 | 1900 | null |
| 4 | 2021 | 1748 | null |
| 5 | 2022 | 1825 | null |
| 6 | 2023 | 1849 | null |
| 7 | 2024 | 622 | null |
| 8 | Unknown | null | null |

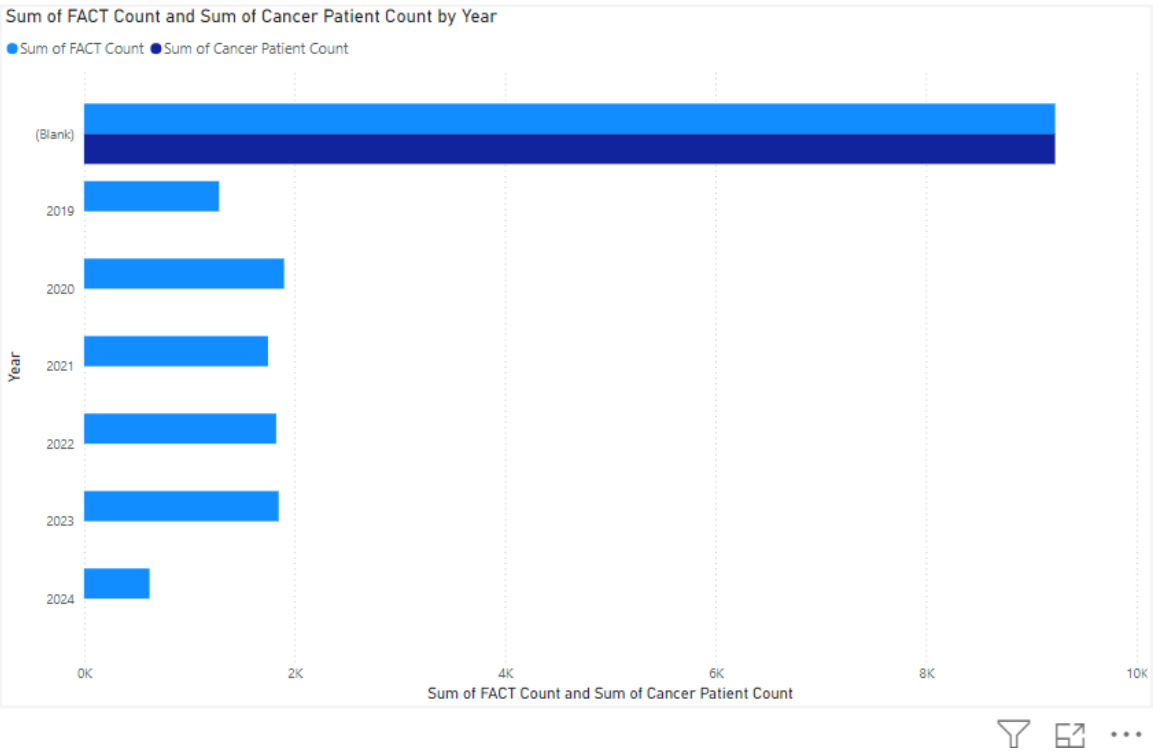
*Hình 298. Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Cancer Patient Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 299 . Close & Apply*

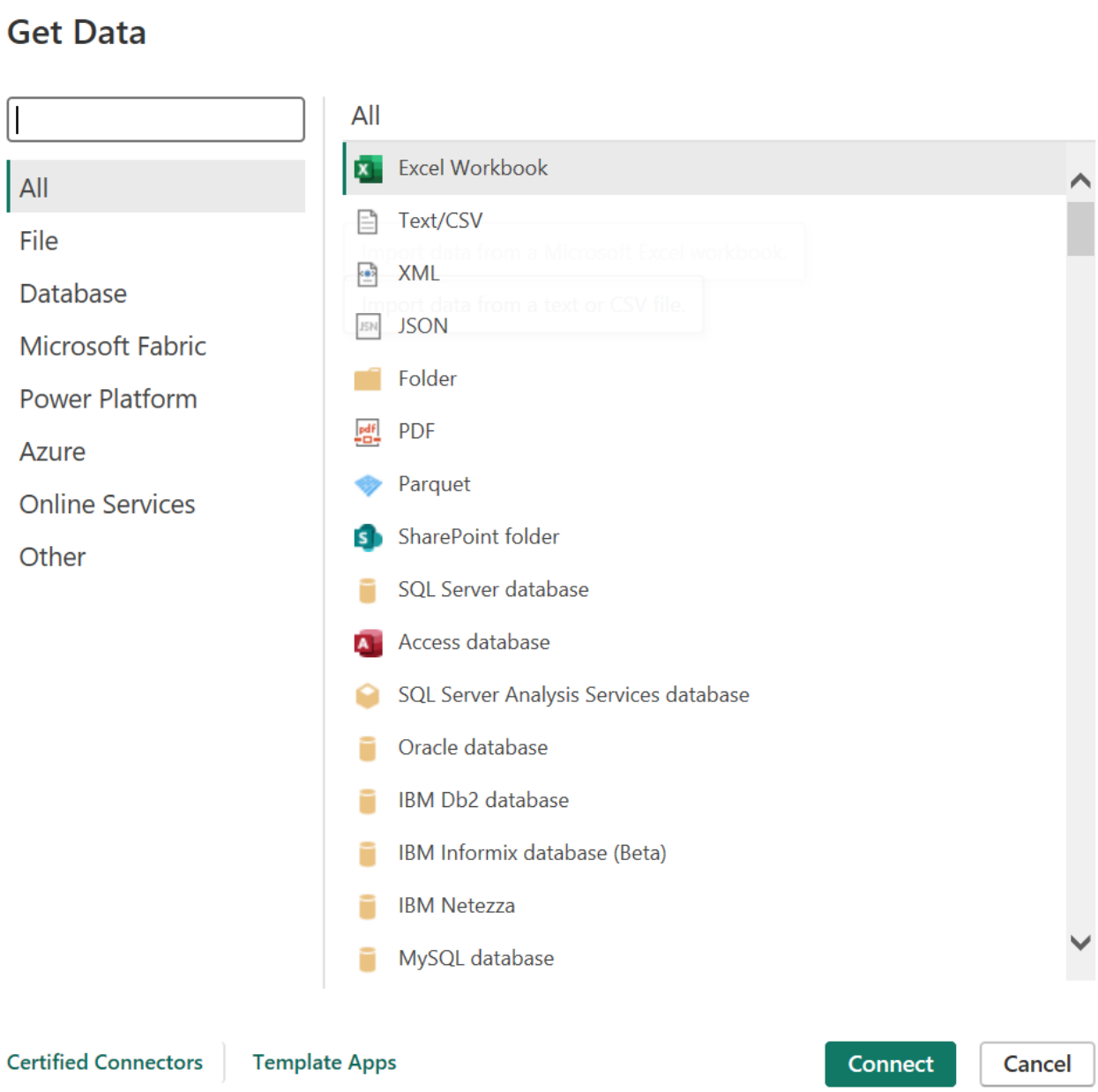
- Bước 6 : Chọn Visualization Clustered Bar Chat và phân bố các thuộc tính như hình dưới




*Hình 300 . Visualization Clustered Bar Chart của Report 4*

## 1.5. REPORT 5

- Nội dung (Câu truy vấn 5 ) : Đưa ra top 5 nhà cung cấp bảo hiểm được sử dụng nhiều nhất trong năm 2023.
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 301. Get Data Report 5*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    {[Measures].[FACT Count]} ON COLUMNS,
    TOPCOUNT(
        [DIM INSURANCE PROVIDER].[Insurance ID].children,
        5,
        [Measures].[FACT Count]
    )
    * [DIM INSURANCE PROVIDER].[Insurance Provider Name].children ON ROWS
FROM [Health Care DW]
WHERE ([Admission Date].[Year].&[2023])
```

OK Cancel

*Hình 302. Import câu lệnh MDX 5*

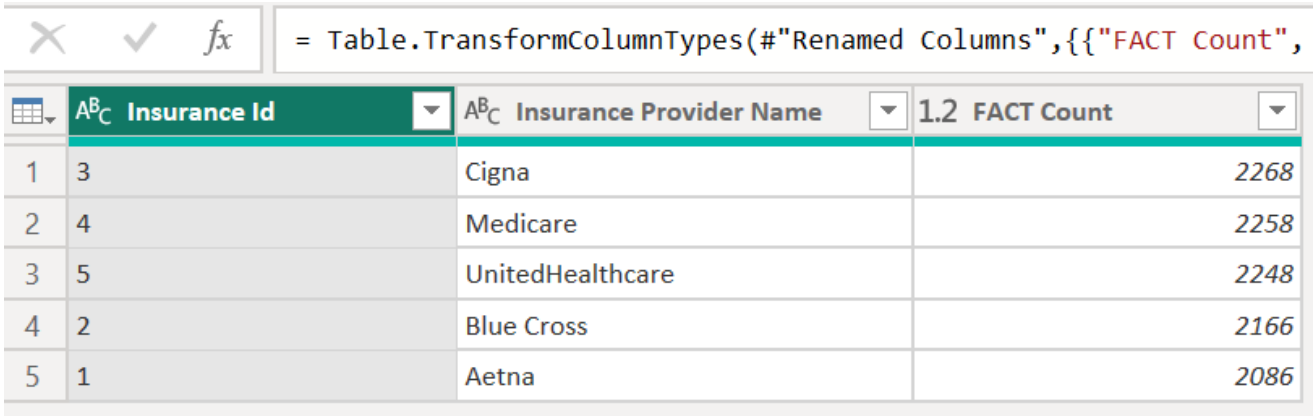
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM INSURANCE PROVIDER].[Insurance Id].[Insurance I | [DIM INSURANCE PROVIDER].[Insurance Provider Name] | [Measures].[FACT Count] |
|---|---|---|
| 3 | Cigna | 2268 |
| 4 | Medicare | 2258 |
| 5 | UnitedHealthcare | 2248 |
| 2 | Blue Cross | 2166 |
| 1 | Aetna | 2086 |

*Hình 303 . Transform Data*

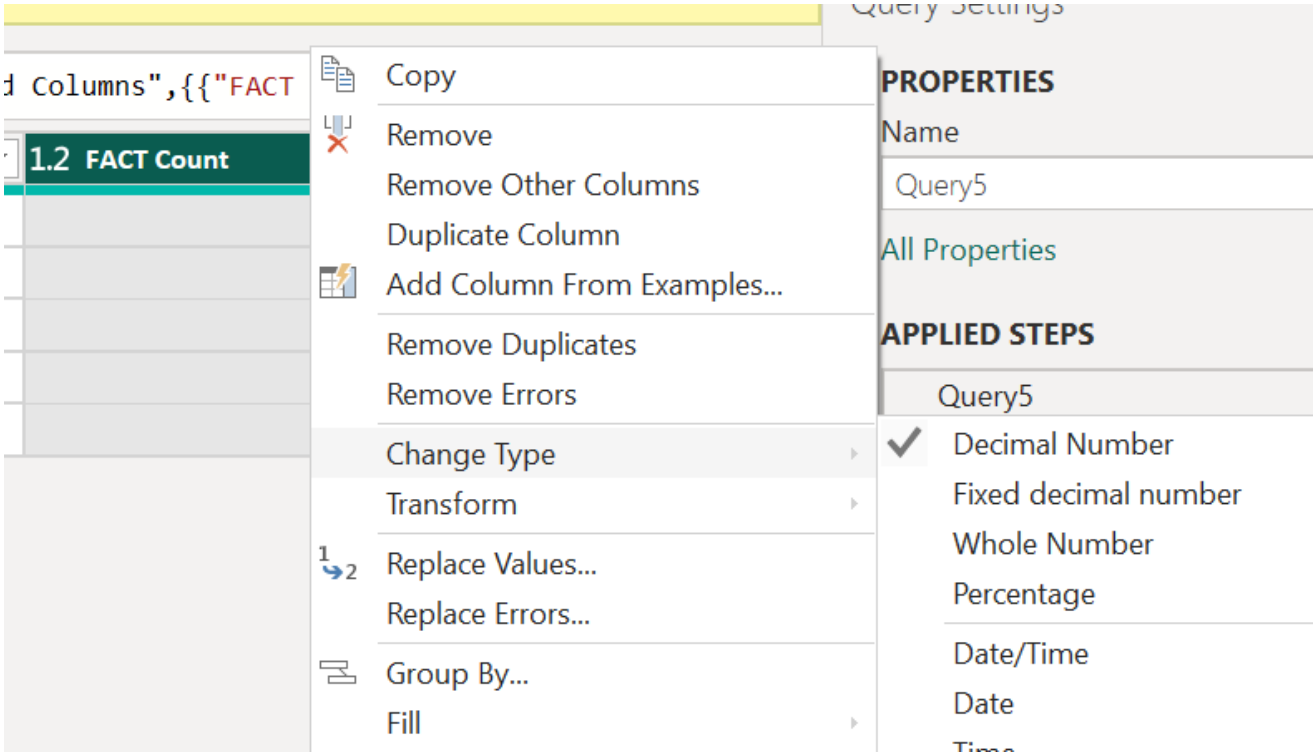
- Bước 4 : Thay đổi tên cho các cột thuộc tính




| | Insurance Id | Insurance Provider Name | 1.2 FACT Count |
|---|---|---|---|
| 1 | 3 | Cigna | 2268 |
| 2 | 4 | Medicare | 2258 |
| 3 | 5 | UnitedHealthcare | 2248 |
| 4 | 2 | Blue Cross | 2166 |
| 5 | 1 | Aetna | 2086 |

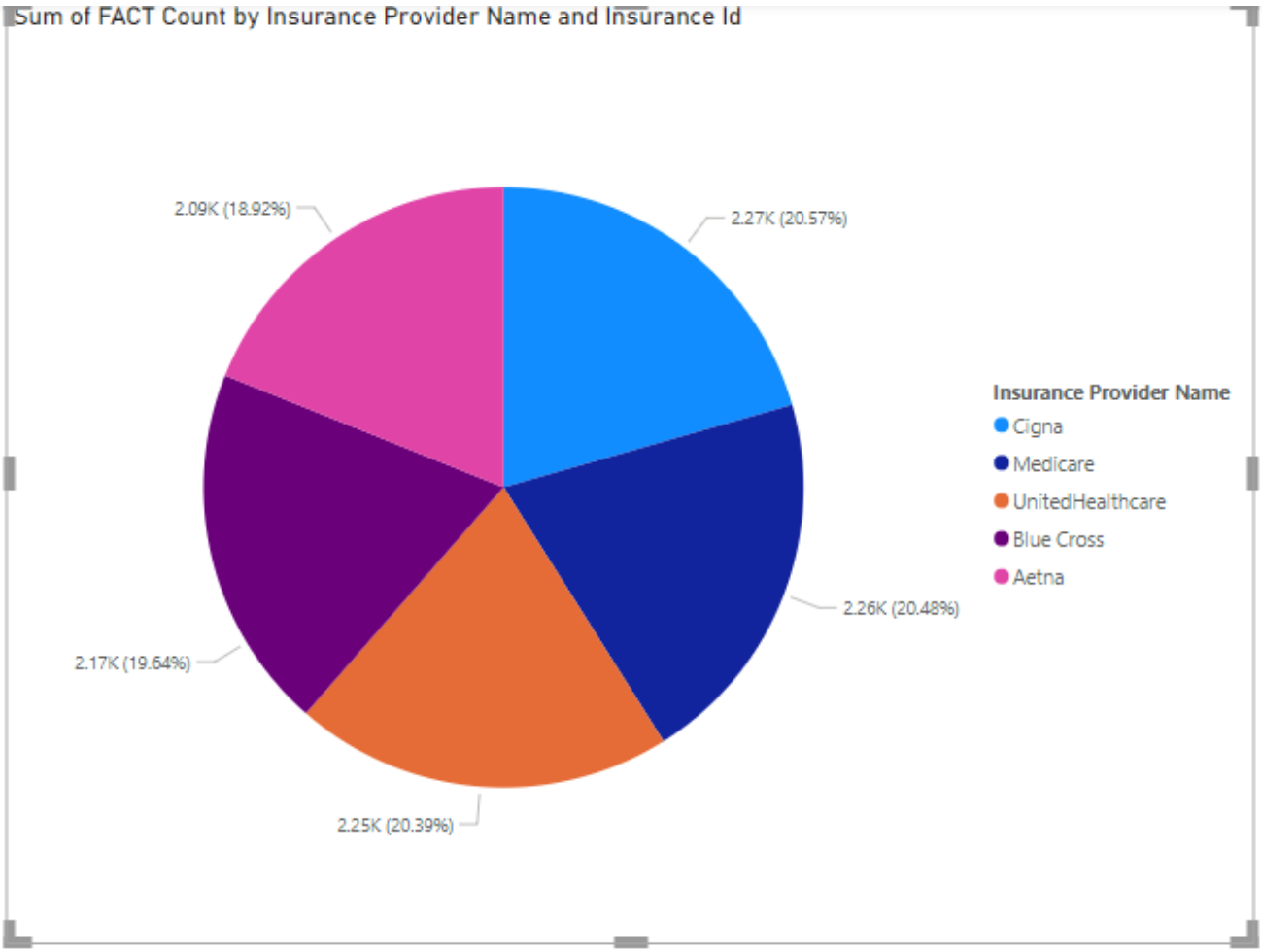
*Hình 304. Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 305. Close & Apply*

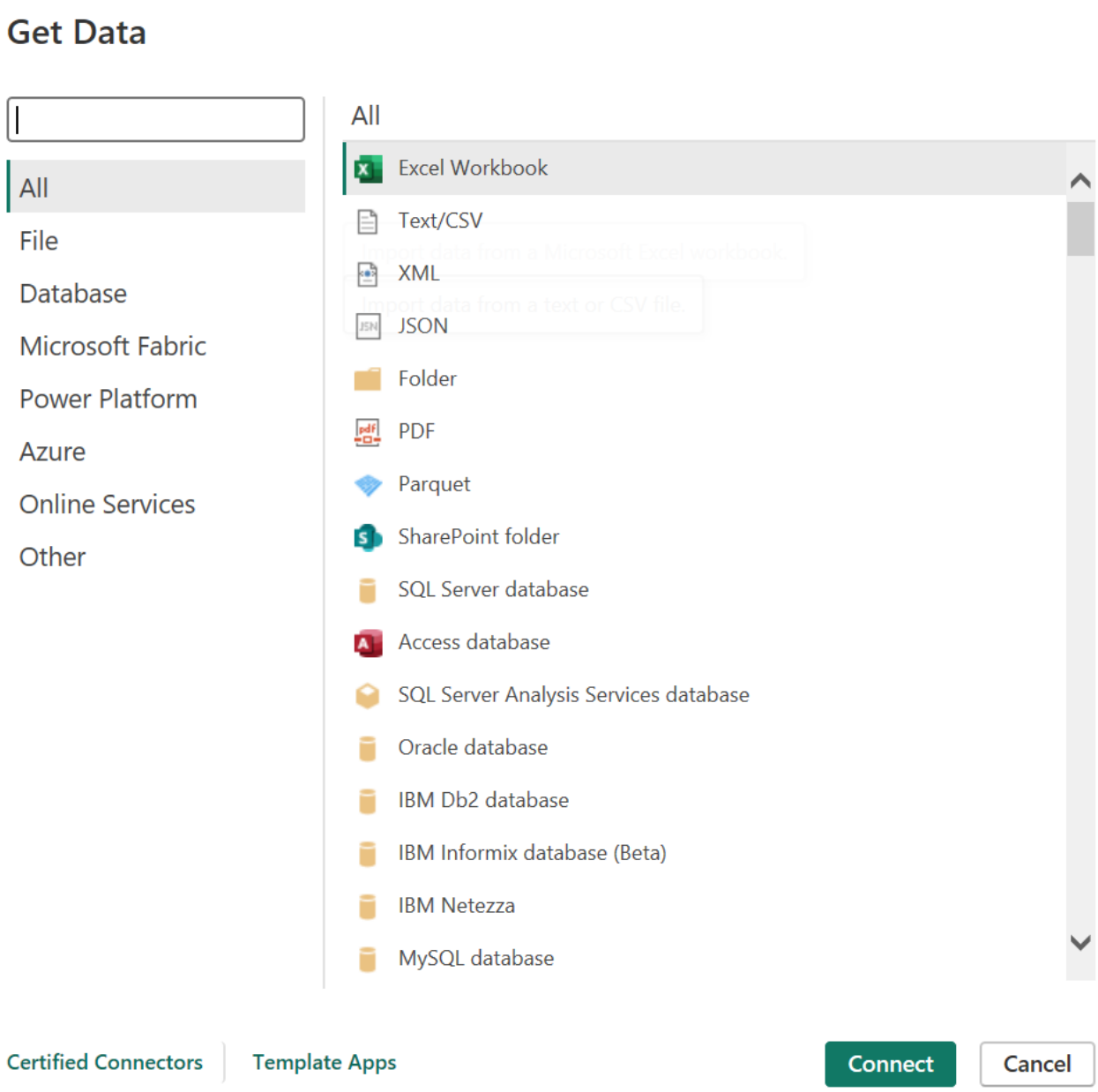
- Bước 6 : Chọn Visualization Pie Chart và phân bố các thuộc tính như hình dưới




*Hình 306. Visualization Pie Chart của Report 5*

## 1.6. REPORT 6

- Nội dung (Câu truy vấn 6 ) : Liệt kê những bệnh nhân thuộc nhóm máu "O+" có tổng số tiền thanh toán thuộc top 5
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn “Get data” trên menu sau đó chọn “**SQL Server Analysis Services database**”.




*Hình 307 . Get Data Report 6*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    {[Measures].[Billing Amount]} ON COLUMNS,
    TOPCOUNT(
        EXCEPT(
            [DIM PATIENT].[Patient ID].Members,
            {[DIM PATIENT].[Patient ID].[All]}
        ),
        5,
        [Measures].[Billing Amount]
    )
    * [DIM PATIENT].[Patient Name].Children ON ROWS
FROM [Health Care DW]
WHERE ([DIM PATIENT].[BLOOD TYPE].&[O+])
```

OK Cancel

*Hình 308 . Import câu lệnh MDX 6*

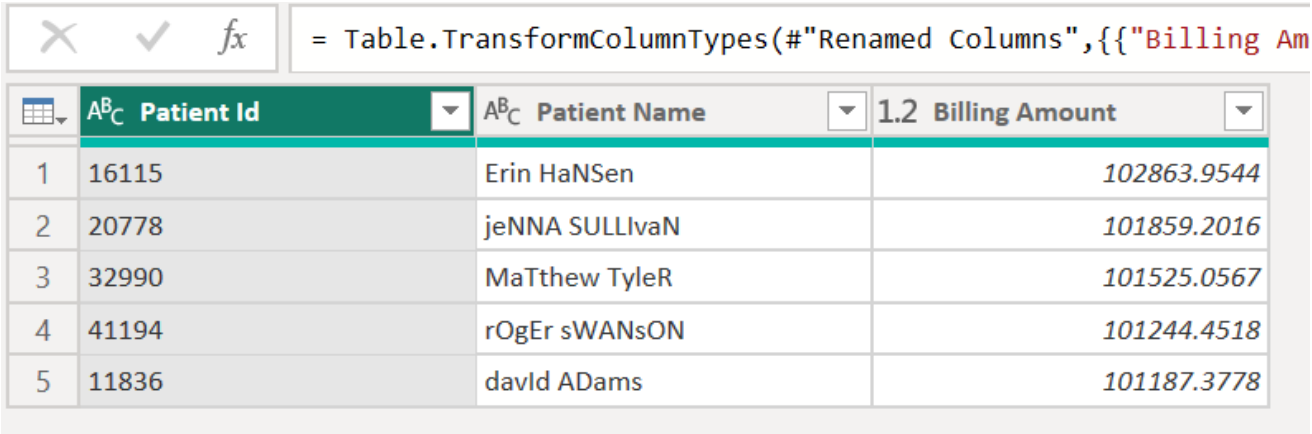
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM PATIENT].[Patient Id].[Patient Id].[MEMBER_CAPTI | [DIM PATIENT].[Patient Name].[Patient Name].[MEMBE | [Measures].[Billing Amount] |
|---|---|---|
| 16115 | Erin HaNSen | 102863.9544 |
| 20778 | jeNNA SULLIvaN | 101859.2016 |
| 32990 | MaTthew TyleR | 101525.0567 |
| 41194 | rOgEr sWANsON | 101244.4518 |
| 11836 | davld ADams | 101187.3778 |

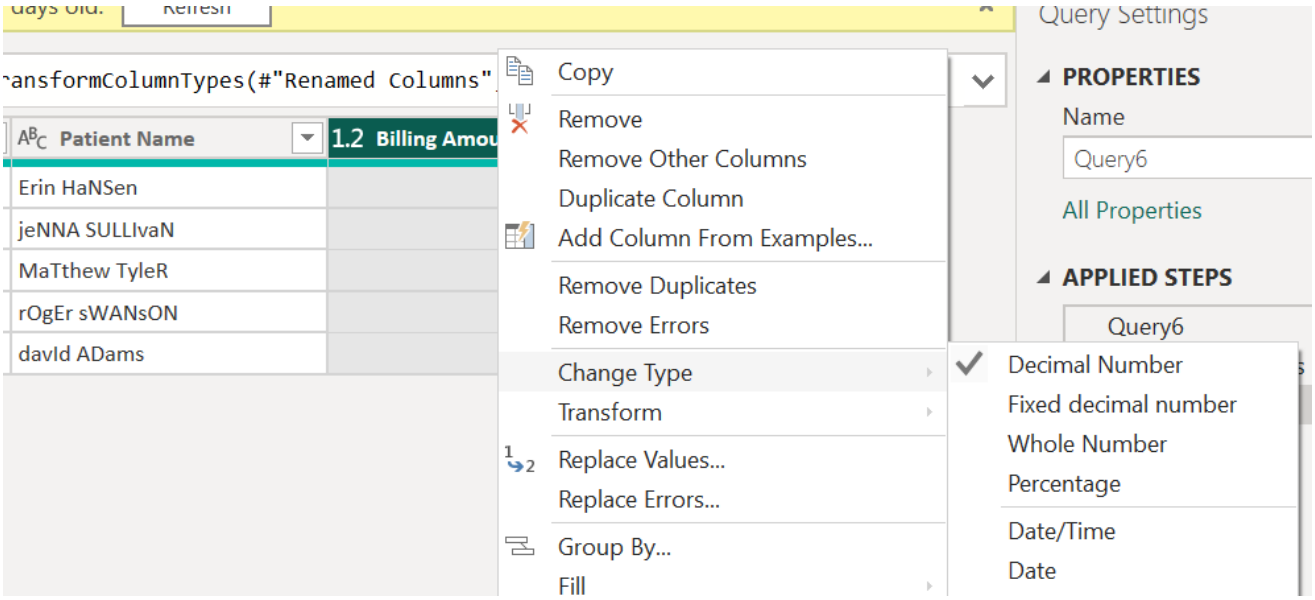
*Hình 309. Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính



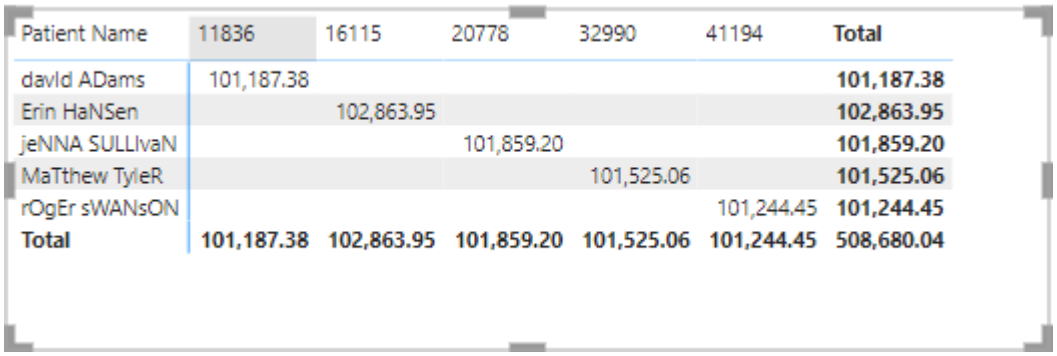

*Hình 310. Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu.




*Hình 311 . Close & Apply*

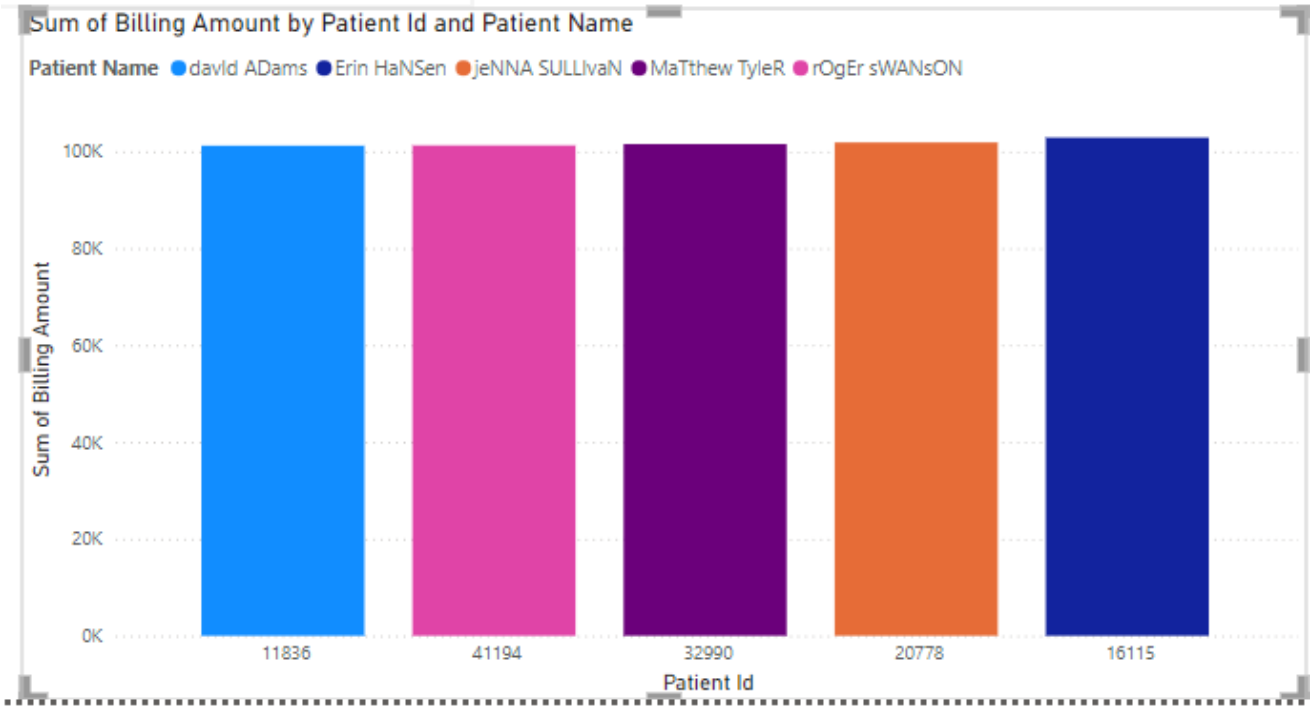
- Bước 6 : Chọn Visualization Matrix và phân bố các thuộc tính như hình dưới



| Patient Name | 11836 | 16115 | 20778 | 32990 | 41194 | Total |
|---|---|---|---|---|---|---|
| davId ADams | 101,187.38 | | | | | **101,187.38** |
| Erin HaNSen | | 102,863.95 | | | | **102,863.95** |
| jeNNA SULLIvaN | | | 101,859.20 | | | **101,859.20** |
| MaTthew TyleR | | | | 101,525.06 | | **101,525.06** |
| rOgEr sWANsON | | | | | 101,244.45 | **101,244.45** |
| **Total** | **101,187.38** | **102,863.95** | **101,859.20** | **101,525.06** | **101,244.45** | **508,680.04** |

*Hình 312 . Visualization matrix của Report 6*

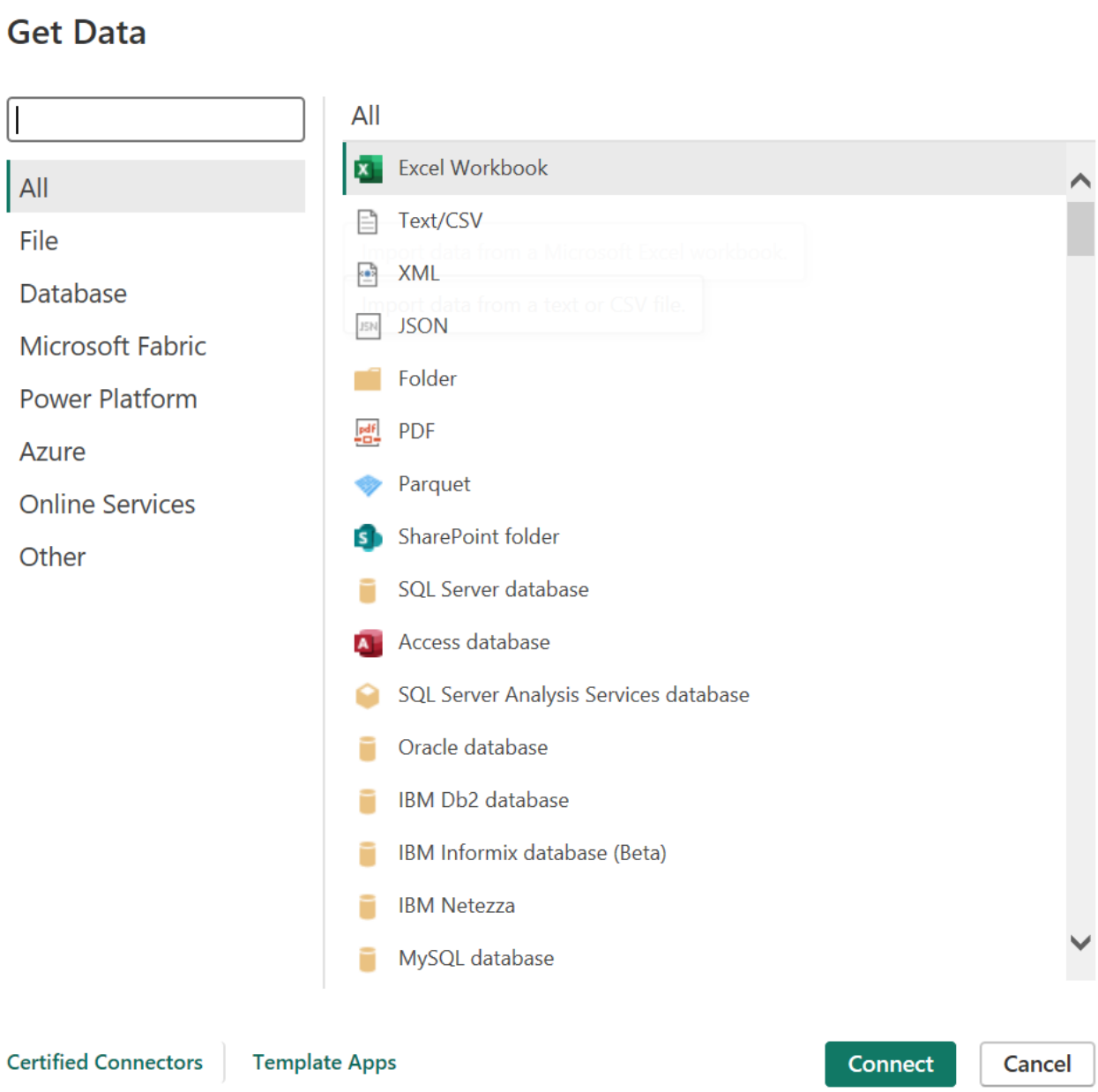
- Bước 7 : Chọn Visualization Stacked Column Chart và phân bố các thuộc tính như hình dưới




*Hình 313 . Visualization Stacked Column Chart của Report 6*

## 1.7. REPORT 7

- Nội dung (Câu truy vấn 7 ) : Thông tin bệnh nhân có số tiền thanh toán cao nhất trong quý 2 năm 2024
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 314 . Get Data Report 7*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    [Measures].[Billing Amount] ON COLUMNS,
    HEAD(
        ORDER(
            [DIM PATIENT].[Patient Id].Children,
            [Measures].[Billing Amount],
            DESC
        ),
        1
    )
    * [DIM PATIENT].[Patient Name].Children
    * [DIM PATIENT].[Gender].Children
    * [DIM PATIENT].[Blood Type].Children
```

OK Cancel

*Hình 315 . Import câu lệnh MDX*

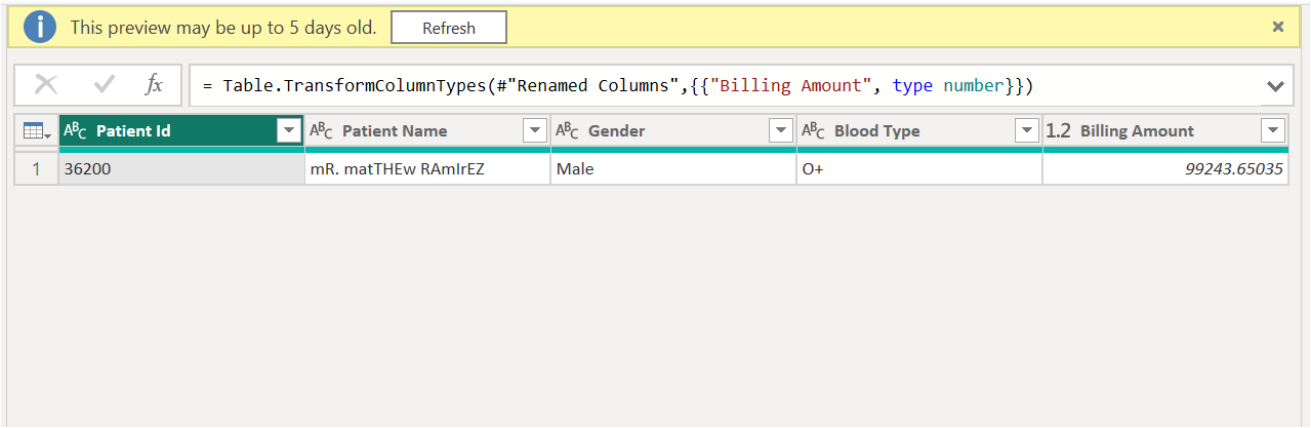
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

CAROLINE: SSAS

| [DIM PATIENT].[Patient Id].[Patient Id].[MEMBER_CAPT | [DIM PATIENT].[Patient Name].[Patient Name].[MEMBE | [DIM PATIENT].[Gender].[Gender].[MEMBER_CA |
|---|---|---|
| 36200 | mR. matTHEw RAmIrEZ | Male |

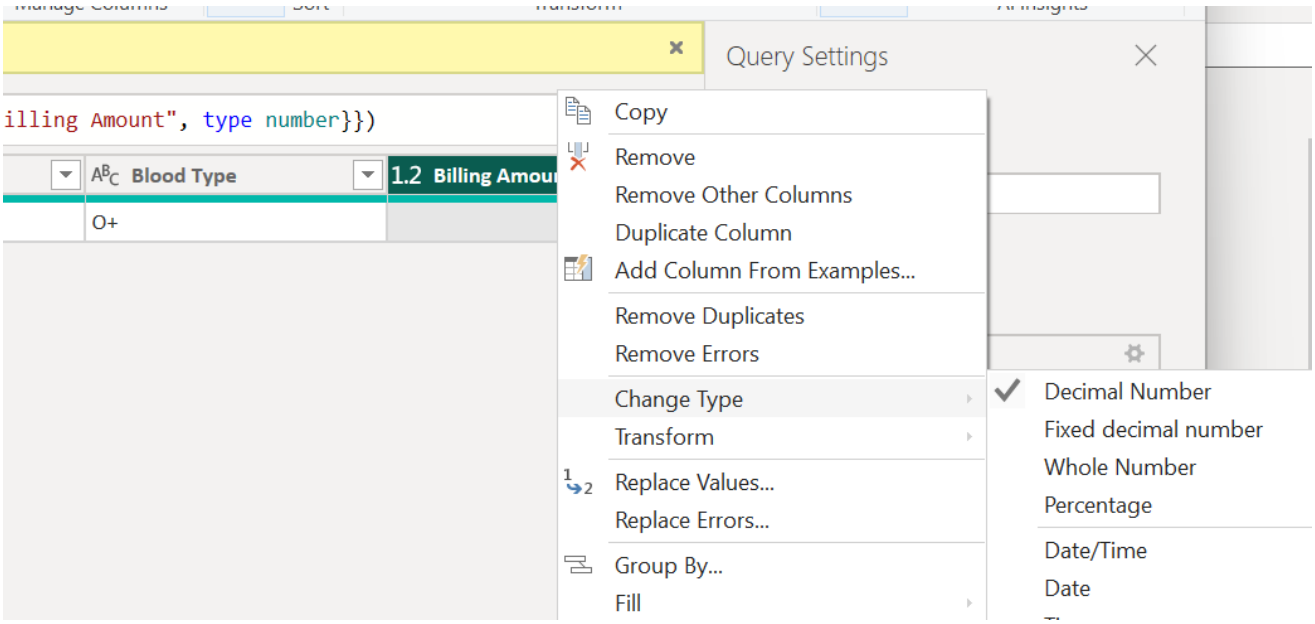
*Hình 316 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính



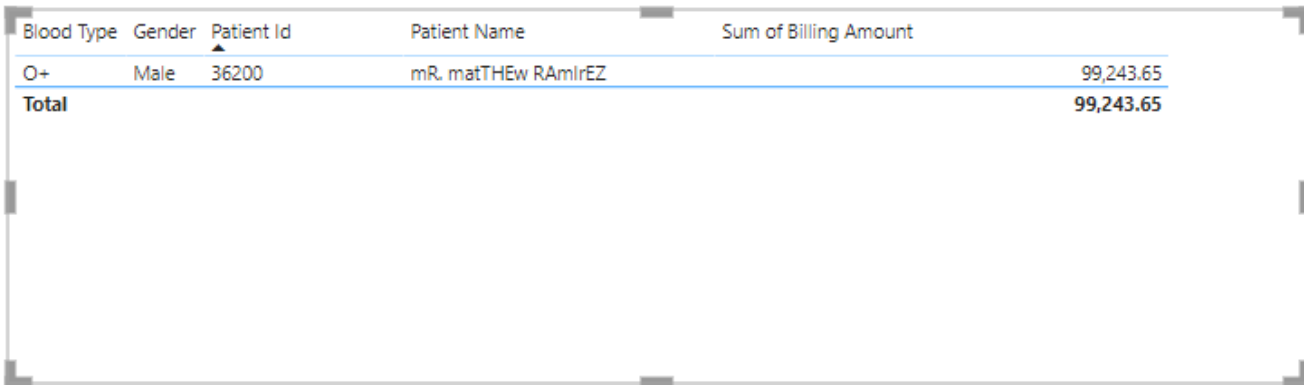

*Hình 317. Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu



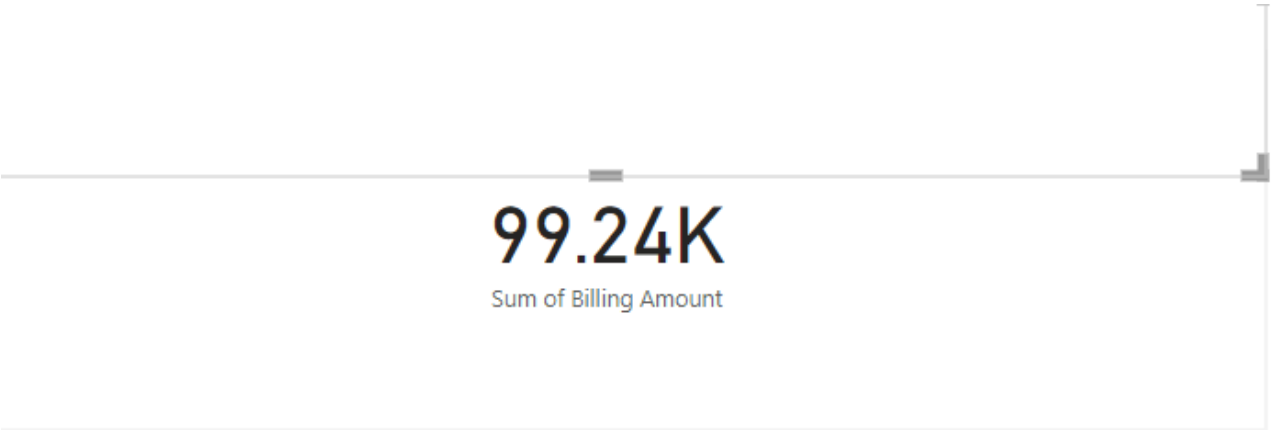

*Hình 318 . Close & Apply*

- Bước 6 : Chọn Visualization Table và phân bố các thuộc tính như hình dưới




*Hình 319 . Visualization Table của Report 7*

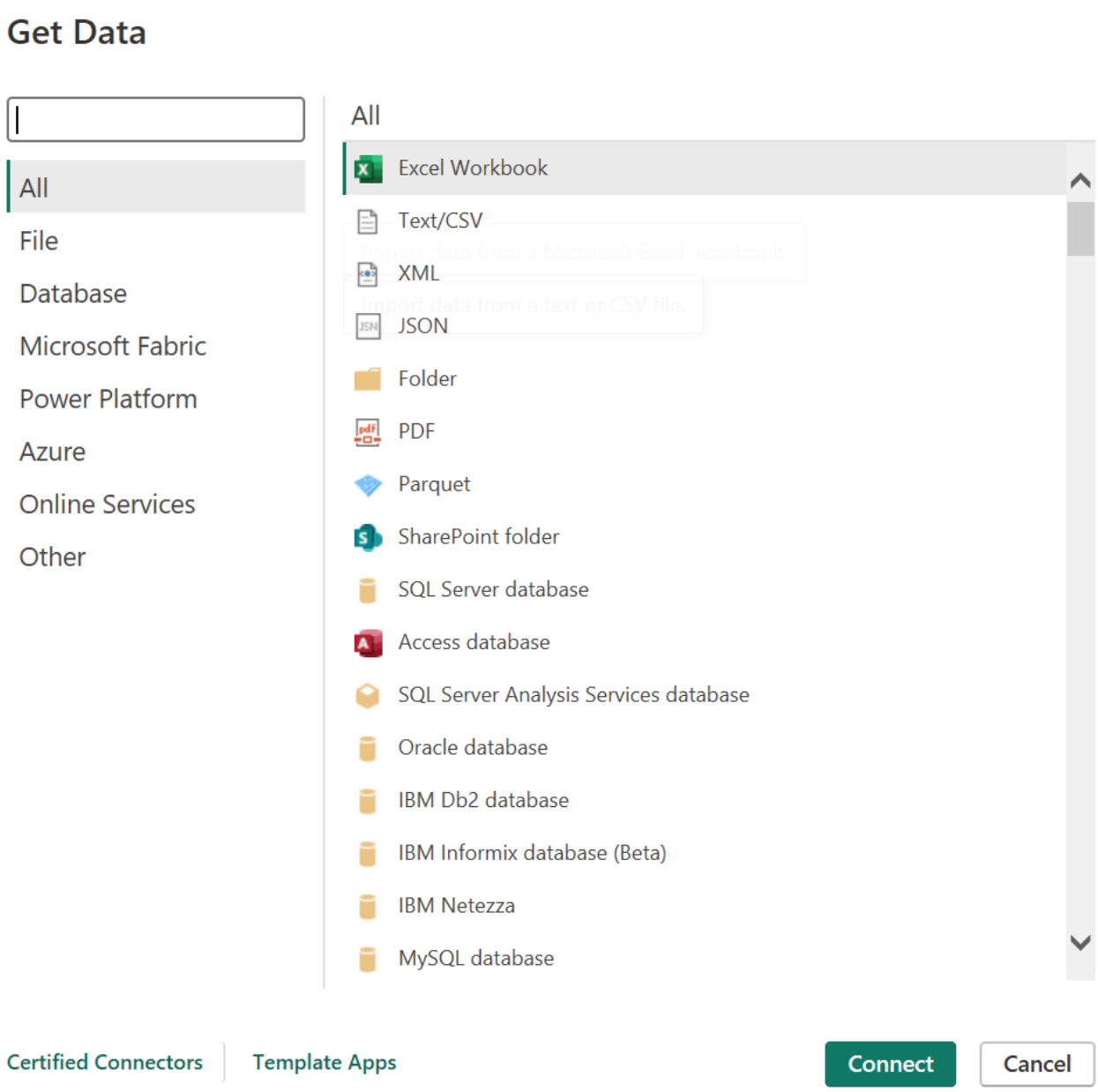
- Bước 7 : Chọn Visualization Card và phân bố các thuộc tính như hình dưới




*Hình 320. Visualization Card  của Report 7*

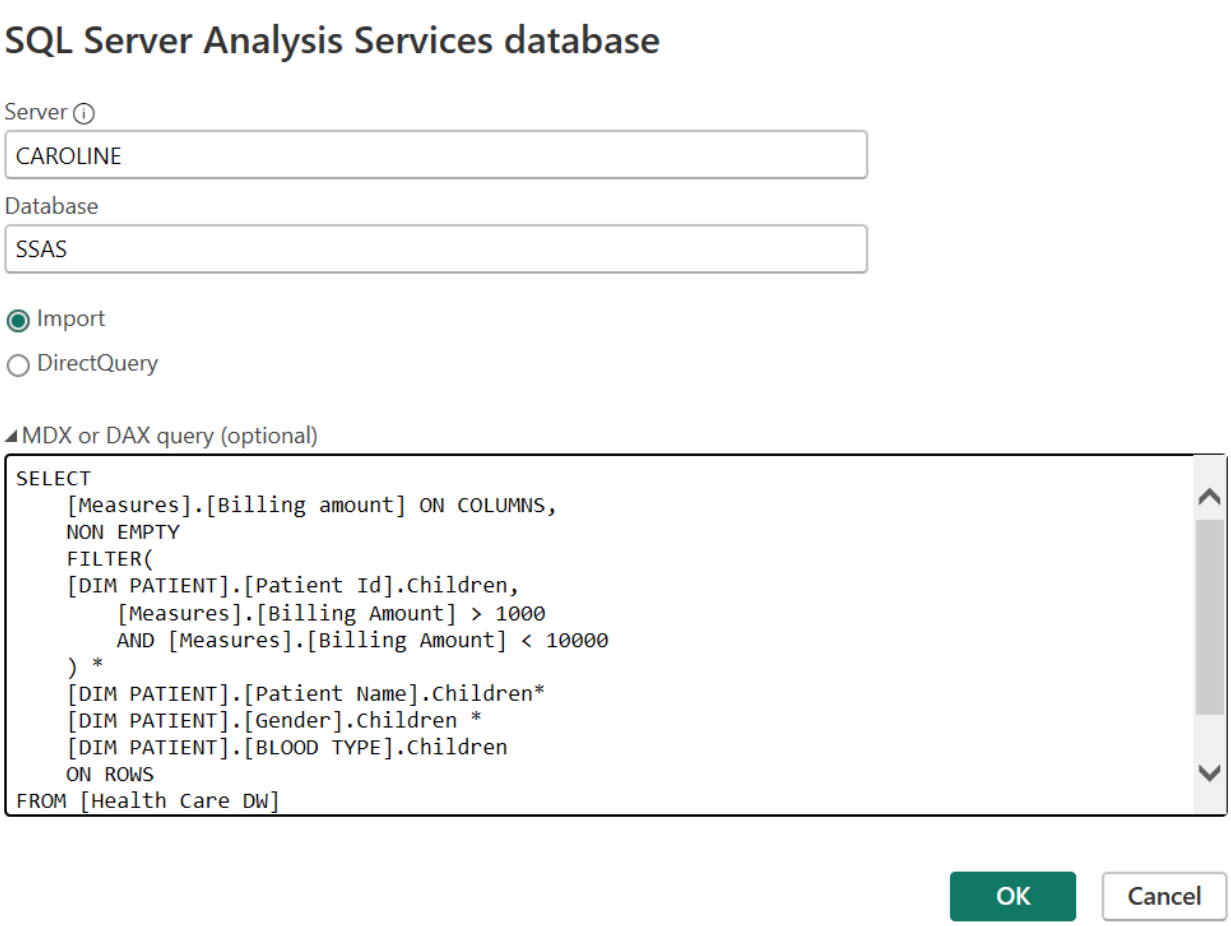
## 1.8. REPORT 8

- **Nội dung (Câu truy vấn 8 )** : Liệt kê các bệnh nhân(mã bệnh nhân, tên bệnh nhân, giới tính, nhóm máu) có tổng số tiền thanh toán trong khoảng từ $1000 đến $10000 trong năm 2024
- **Các bước thực hiện :**
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 321 . Get Data Report 8*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.




*Hình 322 . Import câu lệnh MDX 8*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

CAROLINE: SSAS

| [DIM PATIENT].[Patient Id].[Patient Id].[MEMBER_CAPTI | [DIM PATIENT].[Patient Name].[Patient Name].[MEMBE | [DIM PATIENT].[Gender].[Gender].[MEMBER_CA |
|---|---|---|
| 115 | AarOn sALiNas | Female |
| 121 | aaroN smIth | Female |
| 133 | aAROn wEbeR | Female |
| 229 | aDAm cOLLins | Male |
| 376 | Adrian hOWARd | Male |
| 426 | ALAn BECk | Female |
| 447 | alAN kelleR | Male |
| 503 | AlbeRt SuLliVAn | Male |
| 532 | alEJandro AGuIRre | Female |
| 644 | AlExanDer FosTeR | Female |
| 658 | aleXAnDer HernaNdeZ | Female |
| 681 | aleXAndeR mItCHELL | Male |
| 784 | AlEXAndra sANTIAgO | Male |
| 884 | AlEXiS mOore | Female |
| 906 | aLeXIs Tyler | Female |
| 1090 | ALIEN dENNIS | Female |
| 1179 | alLiSon mCGRATh | Female |
| 1202 | alLIsoN SaNtiAGO | Female |
| 1229 | aLYSsa ALExaNDEr | Female |
| 1362 | aMAnda bRyAnT | Female |

The data in the preview has been truncated due to size limits.

Load | Transform Data | Cancel

*Hình 323. Transform Data*

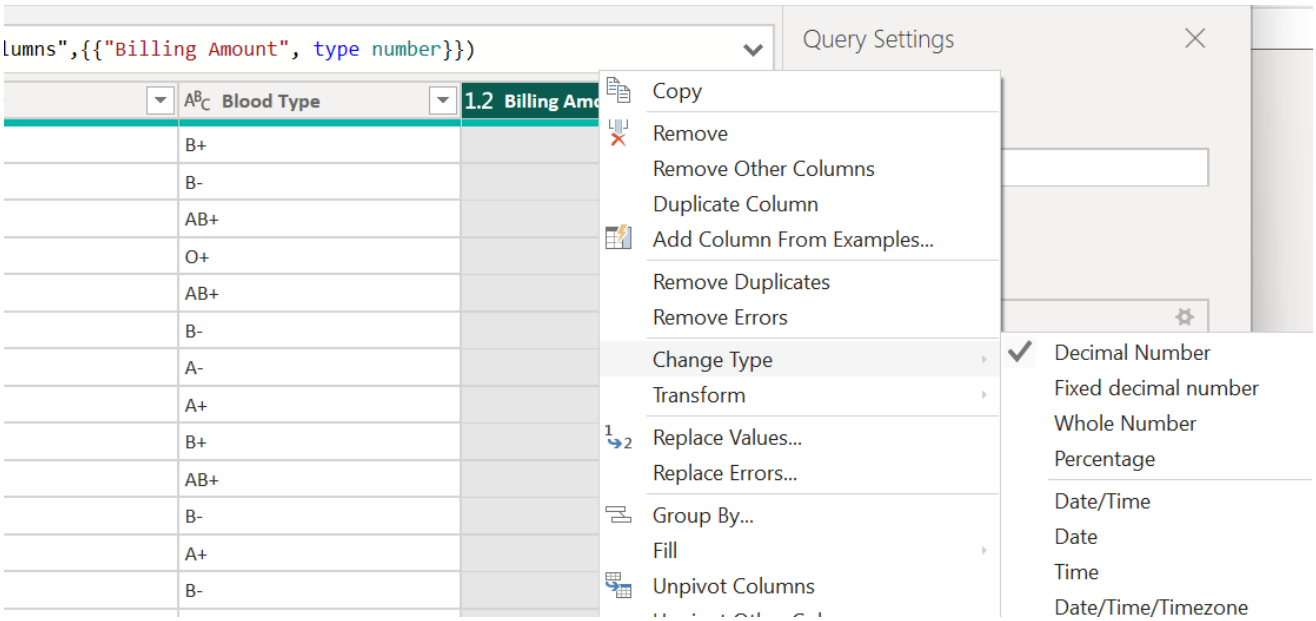
- Bước 4 : Thay đổi tên cho các cột thuộc tính

```
= Table.TransformColumnTypes(#"Renamed Columns",{{"Billing Amount", type number}})
```

| Patient Id | Patient Name | Gender | Blood Type | Billing Amount |
|---|---|---|---|---|
| 115 | AarOn sALiNas | Female | B+ | 9713.4738 |
| 121 | aaroN smIth | Female | B- | 8340.9671 |
| 133 | aAROn wEbeR | Female | AB+ | 8767.4138 |
| 229 | aDAm cOLLins | Male | O+ | 4194.7697 |
| 376 | Adrian hOWARd | Male | AB+ | 5361.3949 |
| 426 | ALAn BECk | Female | B- | 1927.7333 |
| 447 | alAN kelleR | Male | A- | 6125.9827 |
| 503 | AlbeRt SuLliVAn | Male | A+ | 7648.838 |
| 532 | alEJandro AGuIRre | Female | B+ | 9071.6396 |
| 644 | AlExanDer FosTeR | Female | AB+ | 4606.6094 |
| 658 | aleXAnDer HernaNdeZ | Female | B- | 9116.4306 |
| 681 | aleXAndeR mItCHELL | Male | A+ | 6988.6946 |
| 784 | AlEXAndra sANTIAgO | Male | B- | 8355.4318 |
| 884 | AlEXiS mOore | Female | AB+ | 7162.9911 |
| 906 | aLeXIs Tyler | Female | O- | 3355.1379 |
| 1090 | ALIEN dENNIS | Female | A- | 7732.3676 |
| 1179 | alLiSon mCGRATh | Female | AB+ | 7087.5023 |

*Hình 324 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu.




*Hình 325 . Close & Apply*

- Bước 6 : Chọn Visualization Matrix và phân bố các thuộc tính như hình dưới

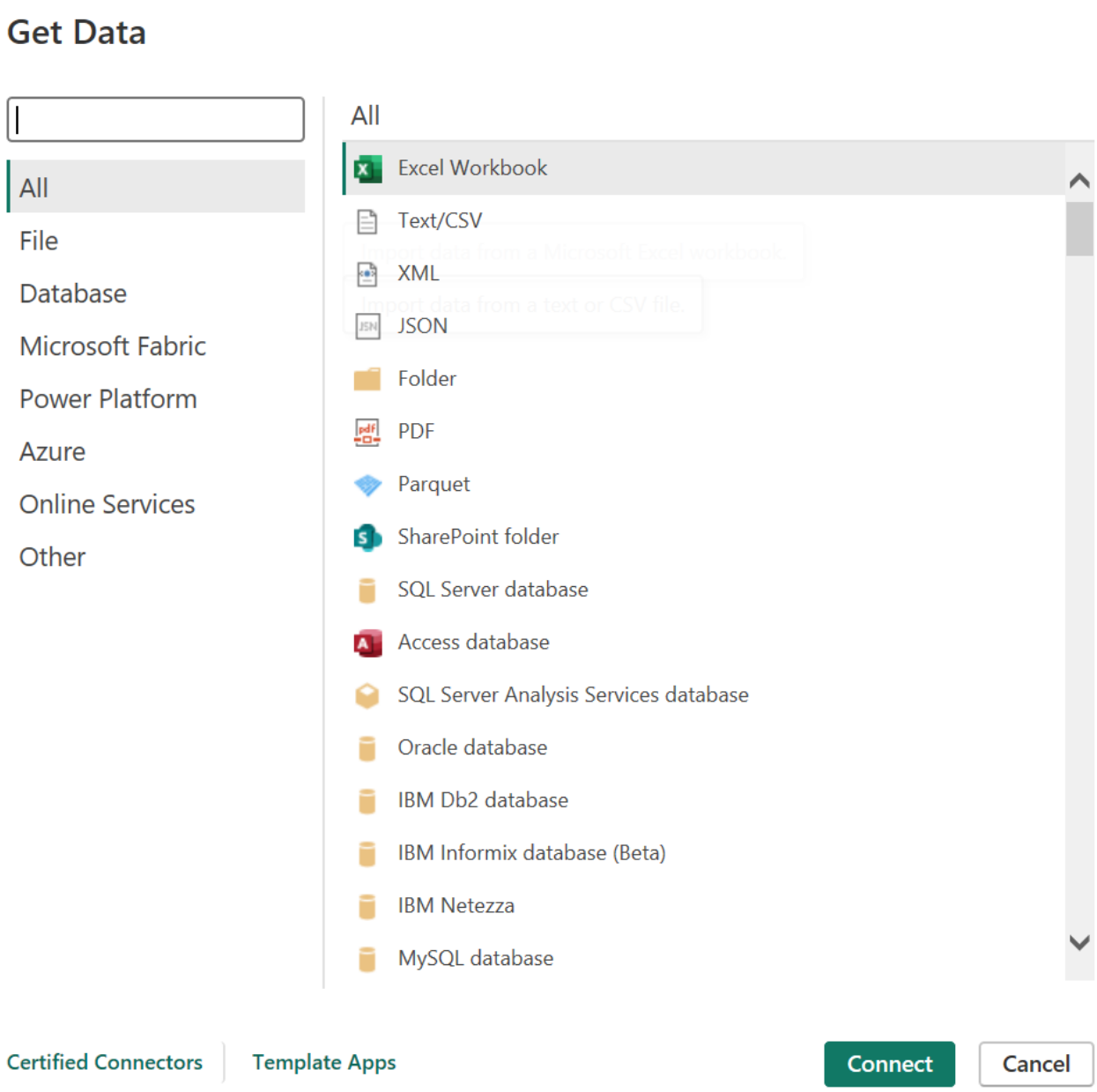
| Patient Name | Patient Id | Gender | Blood Type | Sum of Billing Amount |
|---|---|---|---|---|
| aIAN kelleR | 447 | Male | A- | 6,125.98 |
| ALIEN dENNIS | 1090 | Female | A- | 7,732.37 |
| AManDa MoNTES | 1551 | Female | A- | 2,403.29 |
| AndREW kLEin | 2548 | Female | A- | 2,254.26 |
| anDreW mCCaRThy | 2572 | Female | A- | 5,377.45 |
| aUtumN aILEn | 4538 | Male | A- | 8,645.20 |
| BARbara dIaZ | 4599 | Male | A- | 6,191.39 |
| cHaRIES ANDerSON | 7942 | Female | A- | 8,365.92 |
| chaRLes mccANN | 8089 | Male | A- | 9,842.31 |
| cHeRYl oLIveR | 8401 | Male | A- | 9,342.49 |
| cHRIsTy rUsH | 9686 | Male | A- | 2,307.72 |
| CliFFoRd cAMPBeLL | 9841 | Male | A- | 3,852.79 |
| conNIe Anderson | 10057 | Female | A- | 1,682.52 |
| CouRtnEy NguYEn | 10307 | Female | A- | 5,263.70 |
| CrYsTAL SoTo | 10572 | Male | A- | 5,571.47 |
| dAnA WILLIAmS | 11044 | Male | A- | 8,371.71 |
| DEsTiNY mARtiN | 13390 | Female | A- | 3,148.46 |
| DwaYNE LITTle | 14565 | Female | A- | 5,191.49 |
| EIIzAbeth SPeArs | 15241 | Female | A- | 4,566.56 |
| eRIN AvIla | 16069 | Female | A- | 3,187.62 |
| erIN ILoyD | 16136 | Female | A- | 2,651.14 |
| GILBeRT StEPhENs | 16951 | Male | A- | 3,041.14 |
| GInA raNDALI Dvm | 16996 | Male | A- | 9,025.63 |
| hANnah rAy | 17509 | Female | A- | 5,446.87 |
| hOwARd toRreS | 18119 | Female | A- | 8,689.87 |
| jEssiCa rODRiGuEZ | 22295 | Female | A- | 7,380.73 |
| JoSHUA Peters | 24945 | Male | A- | 7,810.92 |
| jUdy COLEMAN | 25187 | Male | A- | 3,760.71 |
| kARI ROSe | 26098 | Female | A- | 6,181.27 |
| katHRYN gRIffIN | 26502 | Male | A- | 6,885.95 |
| keIth WhITE | 26993 | Female | A- | 1,398.88 |
| **Total** | | | | **3,310,097.91** |

*Hình 326 . Visualization matrix của Report 8*

## 1.9. REPORT 9

- Nội dung (Câu truy vấn 9 ) : Trong mỗi loại bệnh, in ra bác sĩ  có số lượng bênh nhân lớn nhất trong năm 2022

- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn “Get data” trên menu sau đó chọn “**SQL Server Analysis Services database**”.




*Hình 327 . Get Data Report 9*

    - Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    [Measures].[FACT Count] ON COLUMNS,
    NON EMPTY
    GENERATE(
        [DIM PATIENT].[Medical Condition].children,
        TOPCOUNT(
                CROSSJOIN(
                    {[DIM PATIENT].[Medical Condition].CurrentMember},
                    [DIM DOCTOR].[Doctor Name].children
            ),
            1,
            ([Measures].[FACT Count])
        )
```

OK Cancel

*Hình 328 . Import câu lệnh MDX 9*

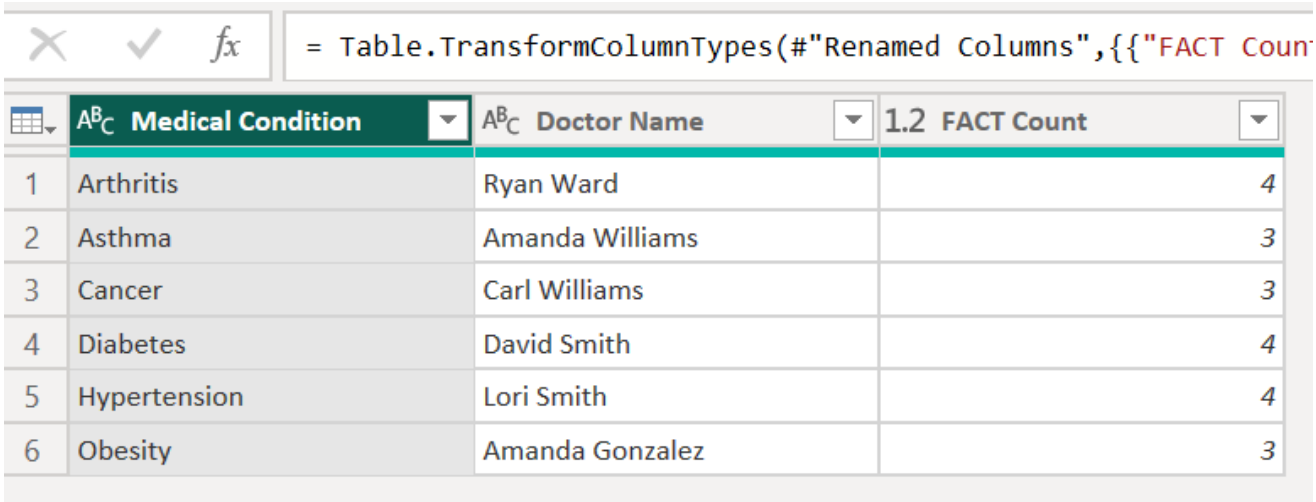
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM PATIENT].[Medical Condition].[Medical Condition]. | [DIM DOCTOR].[Doctor Name].[Doctor Name].[MEMBER | [Measures].[FACT Count] |
|---|---|---|
| Arthritis | Ryan Ward | 4 |
| Asthma | Amanda Williams | 3 |
| Cancer | Carl Williams | 3 |
| Diabetes | David Smith | 4 |
| Hypertension | Lori Smith | 4 |
| Obesity | Amanda Gonzalez | 3 |

*Hình 329 . Transform Data*

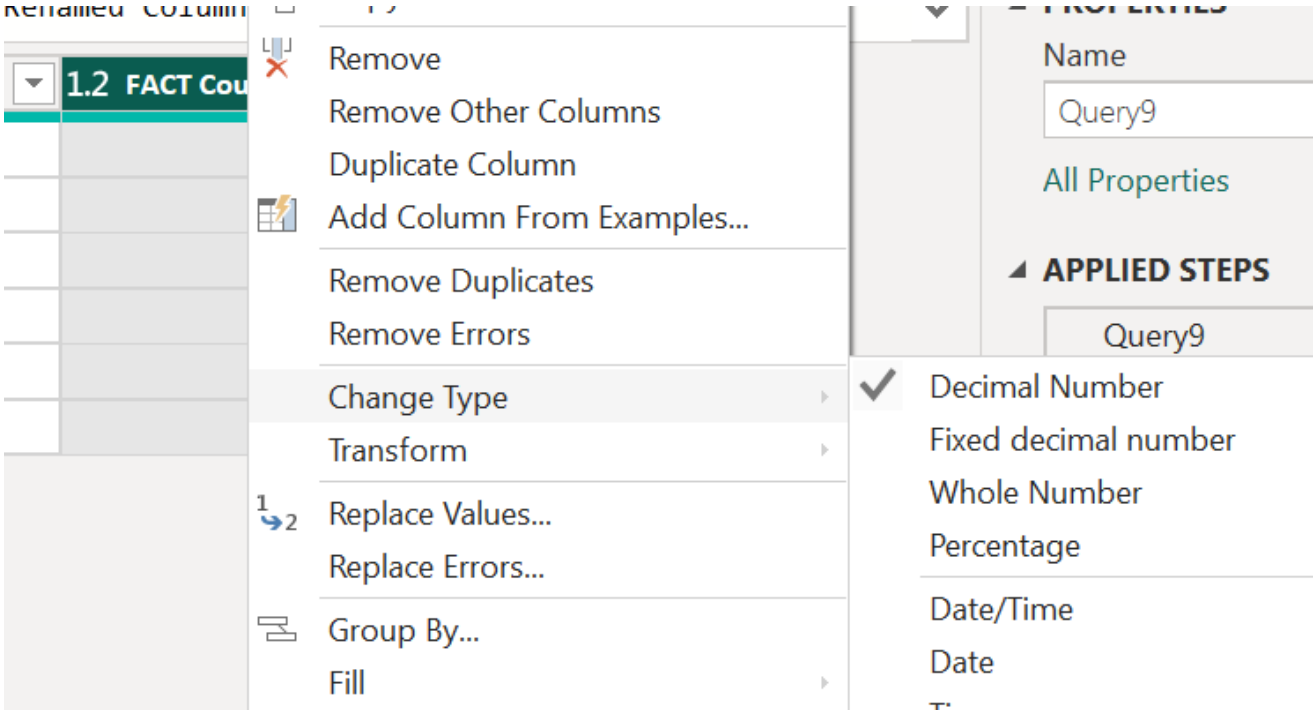
- Bước 4 : Thay đổi tên cho các cột thuộc tính




| | Medical Condition | Doctor Name | FACT Count |
|---|---|---|---|
| 1 | Arthritis | Ryan Ward | 4 |
| 2 | Asthma | Amanda Williams | 3 |
| 3 | Cancer | Carl Williams | 3 |
| 4 | Diabetes | David Smith | 4 |
| 5 | Hypertension | Lori Smith | 4 |
| 6 | Obesity | Amanda Gonzalez | 3 |

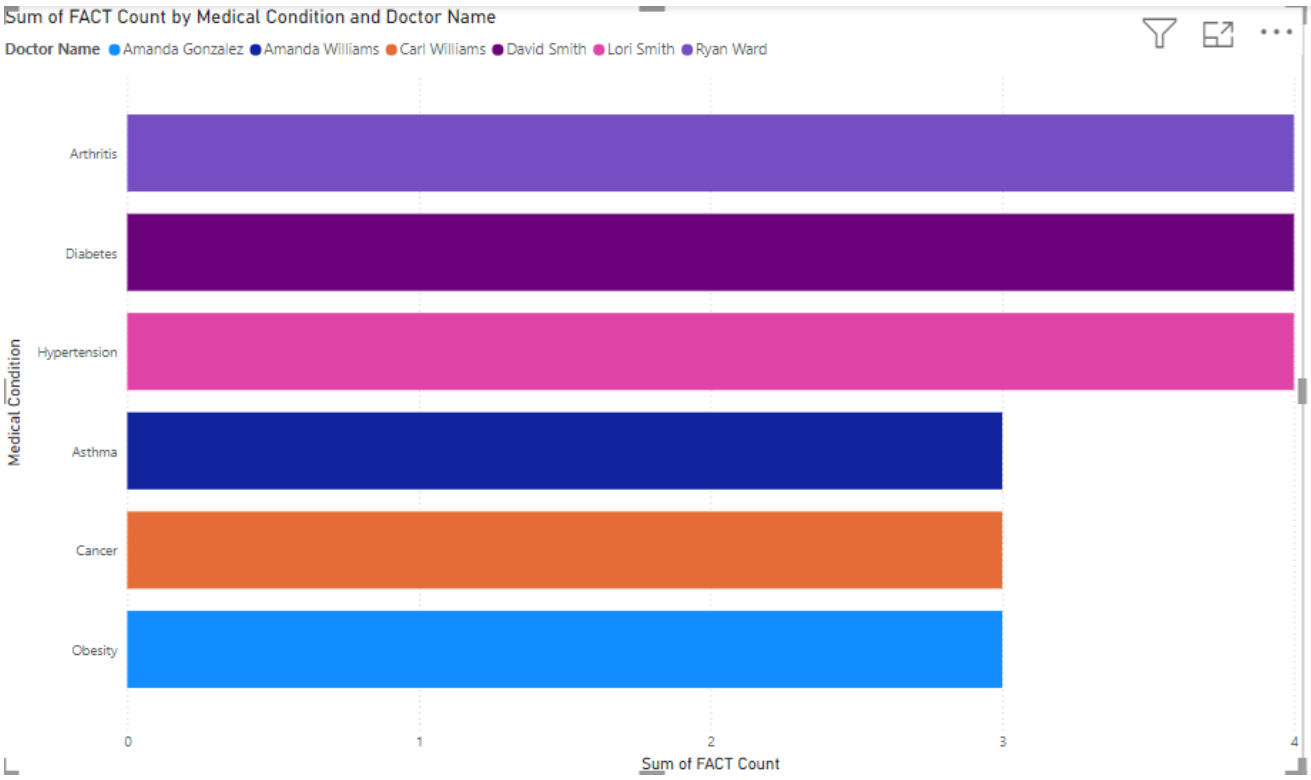
*Hình 330 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 331 . Close & Apply*

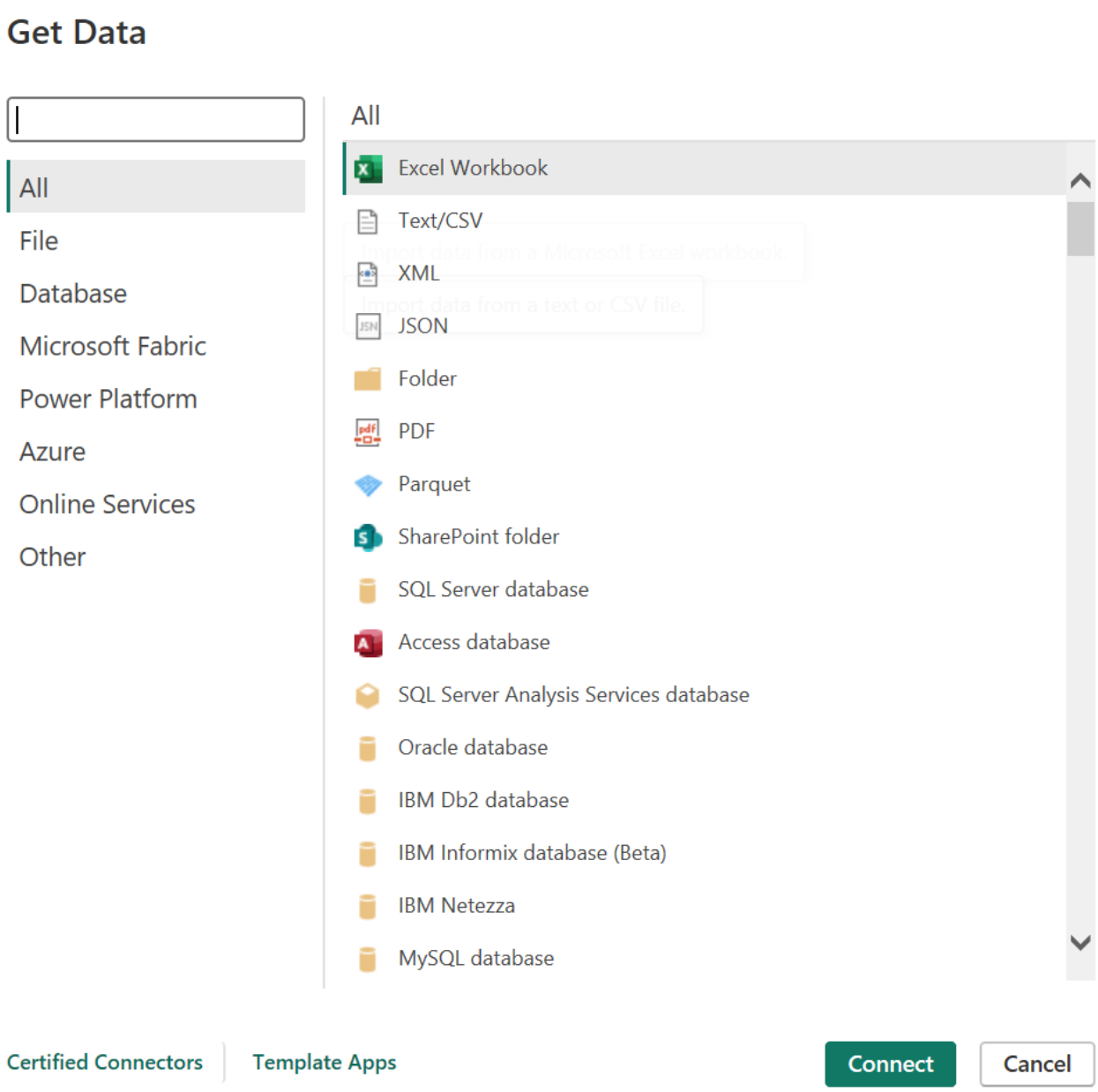
- Bước 6 : Chọn Visualization Stacked Bar Chart và phân bố các thuộc tính như hình dưới




*Hình 332 . Visualization Stacked Bar Chart của Report 9*

## 1.10. REPORT 10

- Nội dung (Câu truy vấn 10 ) : Mỗi quý trong năm 2022, liệt kê 3 loại thuốc được kê đơn nhiều nhất.
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 333 . Get Data Report 10*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    [Measures].[Fact Count] ON COLUMNS,
    GENERATE(
       DESCENDANTS( [Admission Date].[Quarter].children),
        TOPCOUNT(
            {[Admission Date].[Quarter].CurrentMember*
            [DIM MEDICATION].[Medication name].Children},
            3,
            [Measures].[Fact Count]
        )
    ) ON ROWS
FROM [Health Care DW]
WHERE [Admission Date].[Year].&[2022]
```

OK Cancel

*Hình 334 . Import câu lệnh MDX 10*

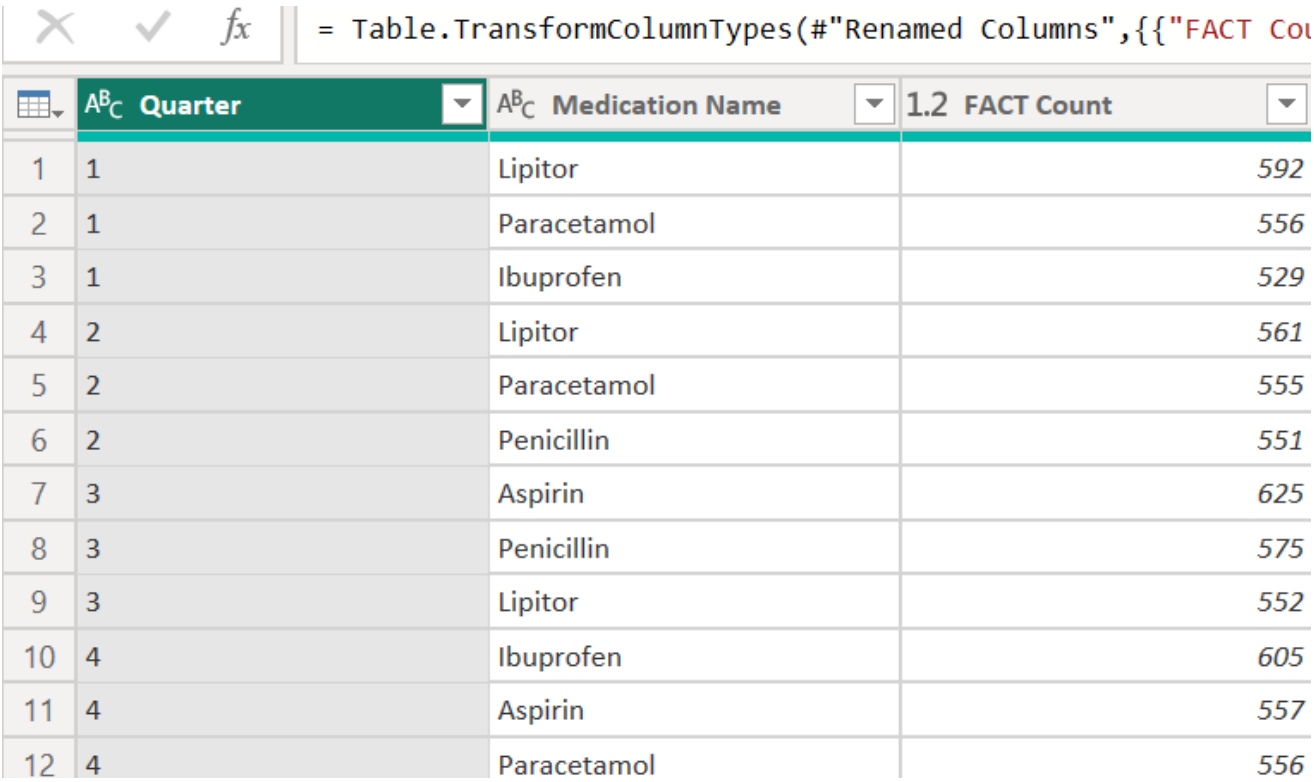
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [Admission Date].[Quarter].[Quarter].[MEMBER_CAPTIC | [DIM MEDICATION].[Medication Name].[Medication Na | [Measures].[FACT Count] |
|---|---|---|
| 1 | Lipitor | 592 |
| 1 | Paracetamol | 556 |
| 1 | Ibuprofen | 529 |
| 2 | Lipitor | 561 |
| 2 | Paracetamol | 555 |
| 2 | Penicillin | 551 |
| 3 | Aspirin | 625 |
| 3 | Penicillin | 575 |
| 3 | Lipitor | 552 |
| 4 | Ibuprofen | 605 |
| 4 | Aspirin | 557 |
| 4 | Paracetamol | 556 |

*Hình 335 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

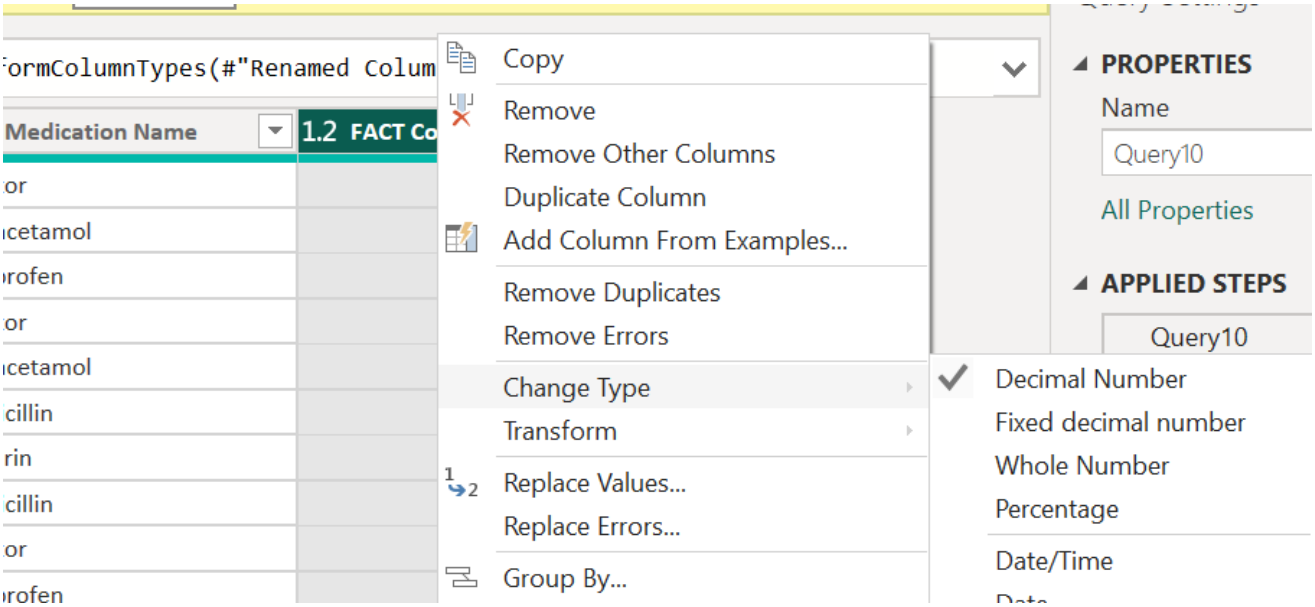


= Table.TransformColumnTypes(#"Renamed Columns",{{"FACT Cou

| | Quarter | Medication Name | FACT Count |
|---|---|---|---|
| 1 | 1 | Lipitor | 592 |
| 2 | 1 | Paracetamol | 556 |
| 3 | 1 | Ibuprofen | 529 |
| 4 | 2 | Lipitor | 561 |
| 5 | 2 | Paracetamol | 555 |
| 6 | 2 | Penicillin | 551 |
| 7 | 3 | Aspirin | 625 |
| 8 | 3 | Penicillin | 575 |
| 9 | 3 | Lipitor | 552 |
| 10 | 4 | Ibuprofen | 605 |
| 11 | 4 | Aspirin | 557 |
| 12 | 4 | Paracetamol | 556 |

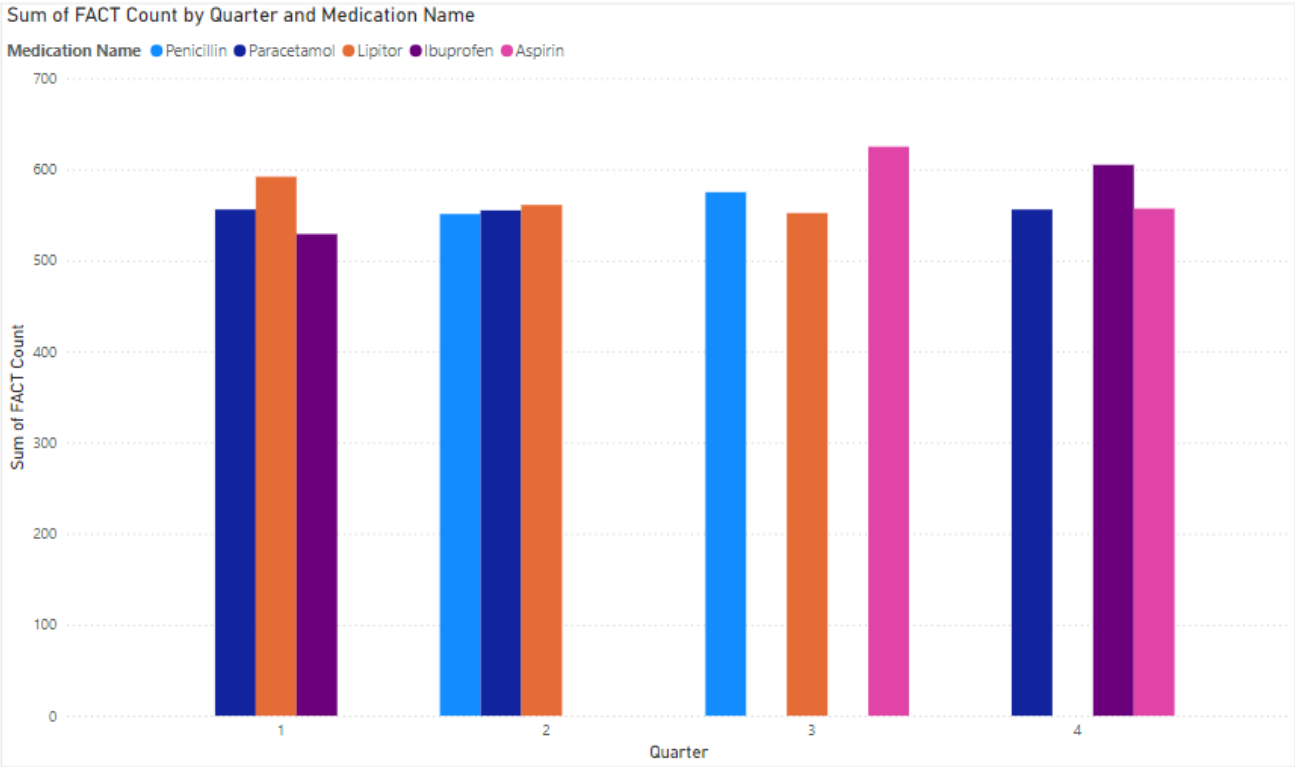
*Hình 336 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 337 . Close & Apply*

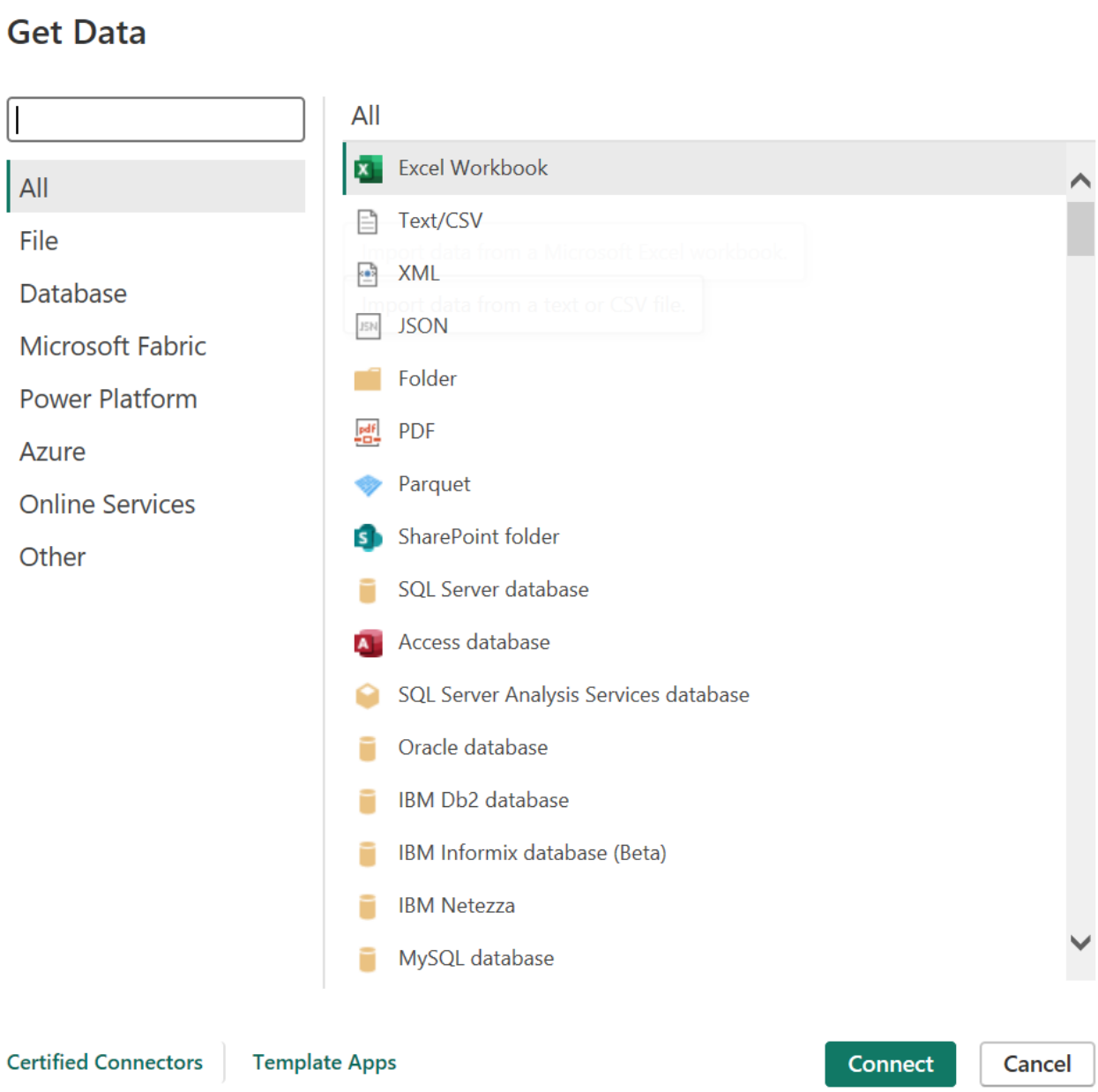
- Bước 6 : Chọn Visualization Clustered Column Chart và phân bố các thuộc tính như hình dưới




*Hình 338 . Visualization Clustered Column Chart của Report 10*

## 1.11. REPORT 11

- **Nội dung** ( Câu truy vấn 11) : Tìm bệnh viện có số ca điều trị nhiều nhất cho từng loại bệnh cụ thể
- **Các bước thực hiện :**
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 339 . Get Data Report 11*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    {[Measures].[FACT Count]} ON COLUMNS,
    NONEMPTY(
        GENERATE(
            [DIM PATIENT].[Medical Condition].Children,
            TOPCOUNT(
                [DIM LOCATION].[Hospital Name].Children,
                1,
                [Measures].[FACT Count]
            ) *
            {[DIM PATIENT].[Medical Condition].CurrentMember}
        )
    ) ON ROWS
```

OK Cancel

*Hình 340 . Import câu lệnh MDX 11*

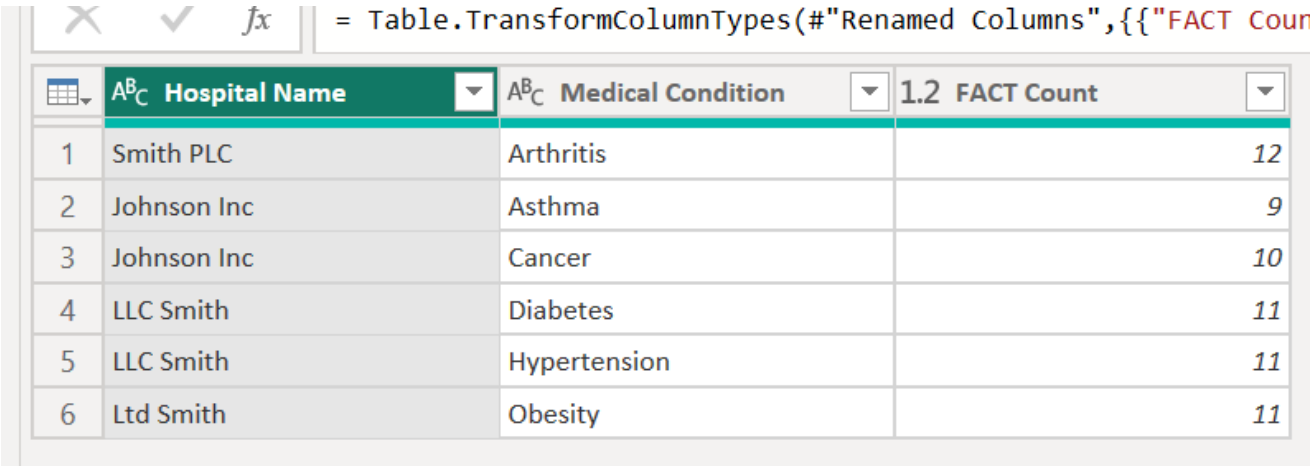
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM LOCATION].[Hospital Name].[Hospital Name].[MEI | [DIM PATIENT].[Medical Condition].[Medical Condition]. | [Measures].[FACT Count] |
|---|---|---|
| Smith PLC | Arthritis | 12 |
| Johnson Inc | Asthma | 9 |
| Johnson Inc | Cancer | 10 |
| LLC Smith | Diabetes | 11 |
| LLC Smith | Hypertension | 11 |
| Ltd Smith | Obesity | 11 |

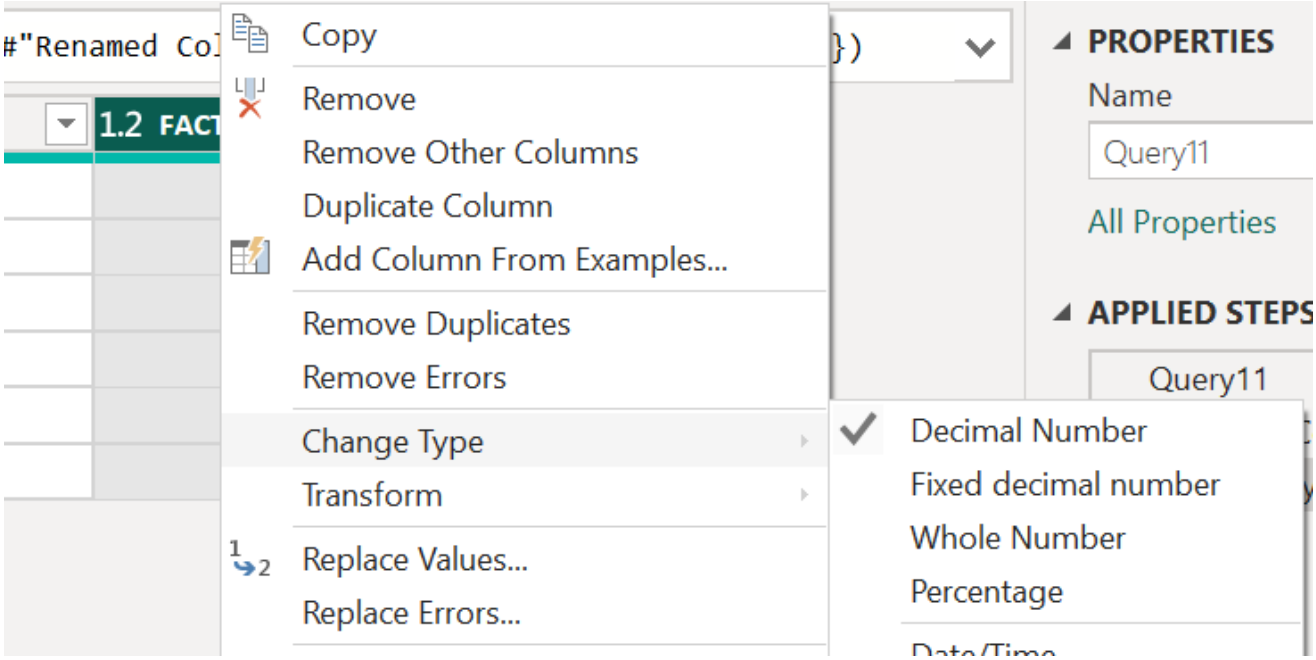
*Hình 341 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính



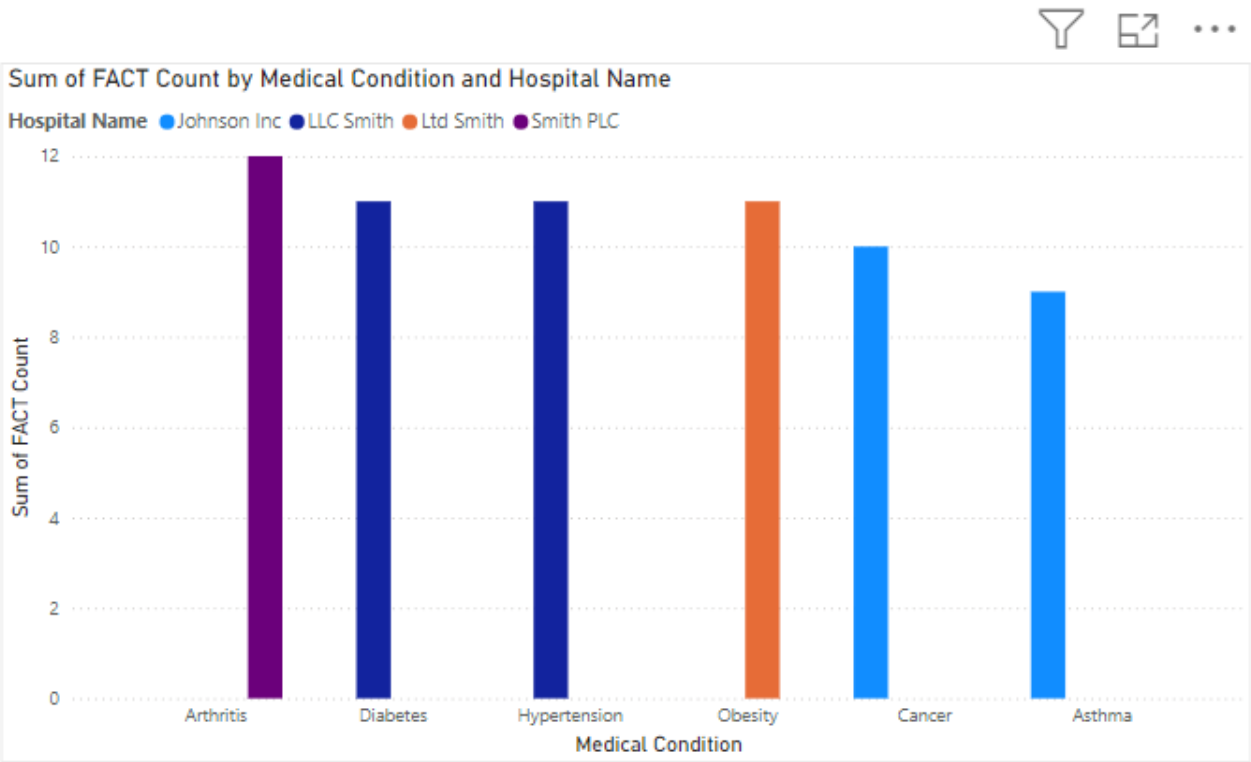

*Hình 342 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 343 . Close & Apply*

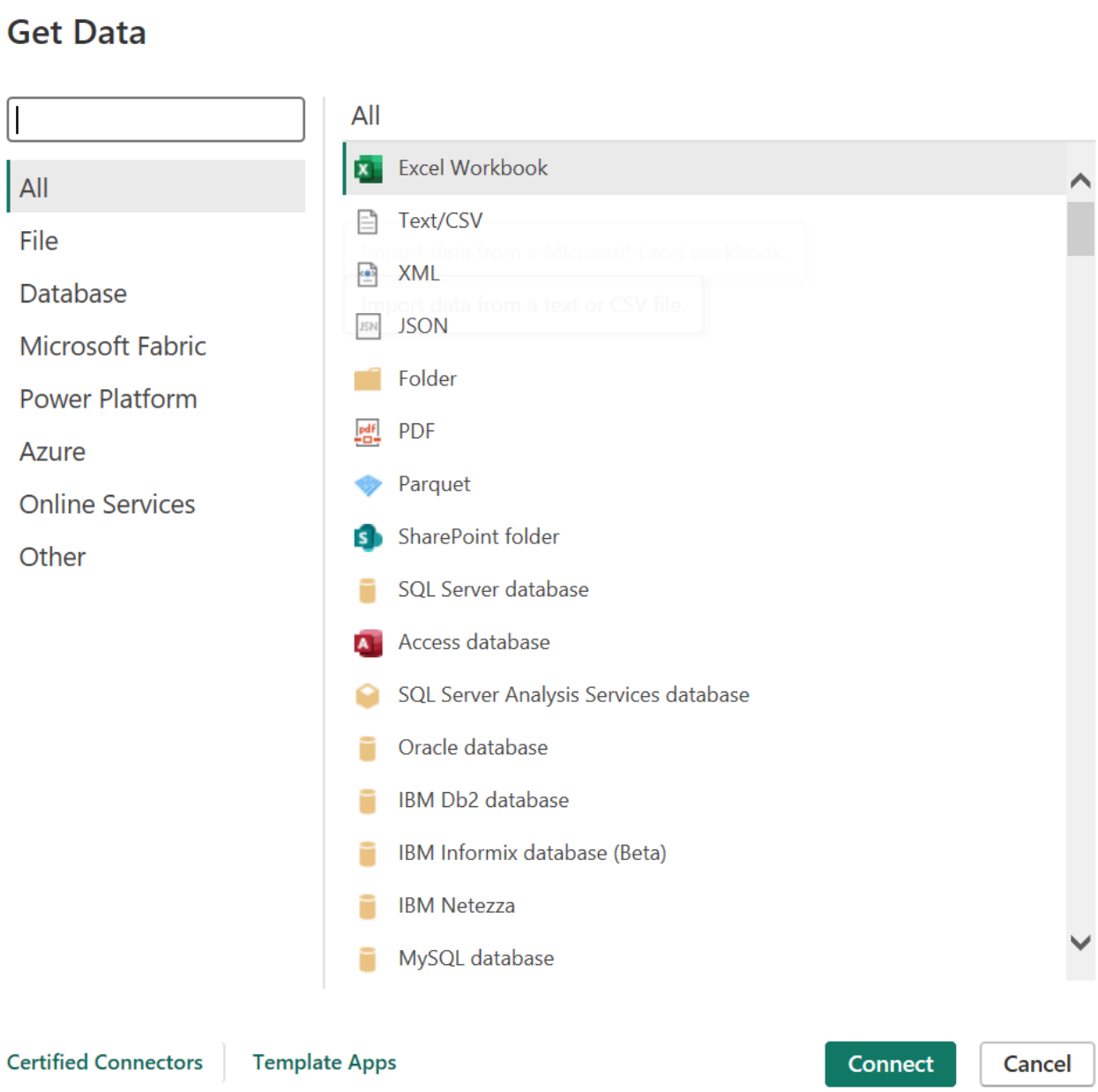
- Bước 6 : Chọn Visualization Clustered Column Chart và phân bố các thuộc tính như hình dưới




*Hình 344 . Visualization Clustered Column Chart của Report 11*

## 1.12. REPORT 12

- **Nội dung** ( Câu truy vấn 12) : Tính tổng chi phí hoá đơn của từng bệnh viện của những bệnh nhân có tình trạng bệnh "Diabetes" và có tổng viện phí trên $200000, sau đó sắp xếp theo tổng chi phí giảm dần.
- **Các bước thực hiện :**
    - Bước 1: Tạo dataset bằng cách chọn “Get data” trên menu sau đó chọn “**SQL Server Analysis Services database**”.




*Hình 345 . Get Data Report 12*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
SELECT
    ORDER(
        FILTER(
            {[DIM LOCATION].[Hospital Name].Members},
            [Measures].[Billing Amount] > 200000
        ),
  [Measures].[Billing Amount] ,
       BDESC
    ) ON ROWS,
    {[Measures].[Billing Amount]} ON COLUMNS
FROM [Health Care DW]
WHERE [DIM PATIENT].[Medical Condition].&[Diabetes]
```

OK Cancel

*Hình 346 . Import câu lệnh MDX 12*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM LOCATION].[Hospital Name].[Hospital Name].[MEI | [Measures].[Billing Amount] |
|---|---|
| *null* | *238539725.5* |
| PLC Brown | *304480.7954* |
| Group Smith | *291522.6529* |
| Smith Ltd | *247134.6459* |
| Rodriguez Inc | *243599.129* |
| Smith Inc | *227727.1306* |
| LLC Smith | *218458.2625* |
| Martin Inc | *217248.2123* |

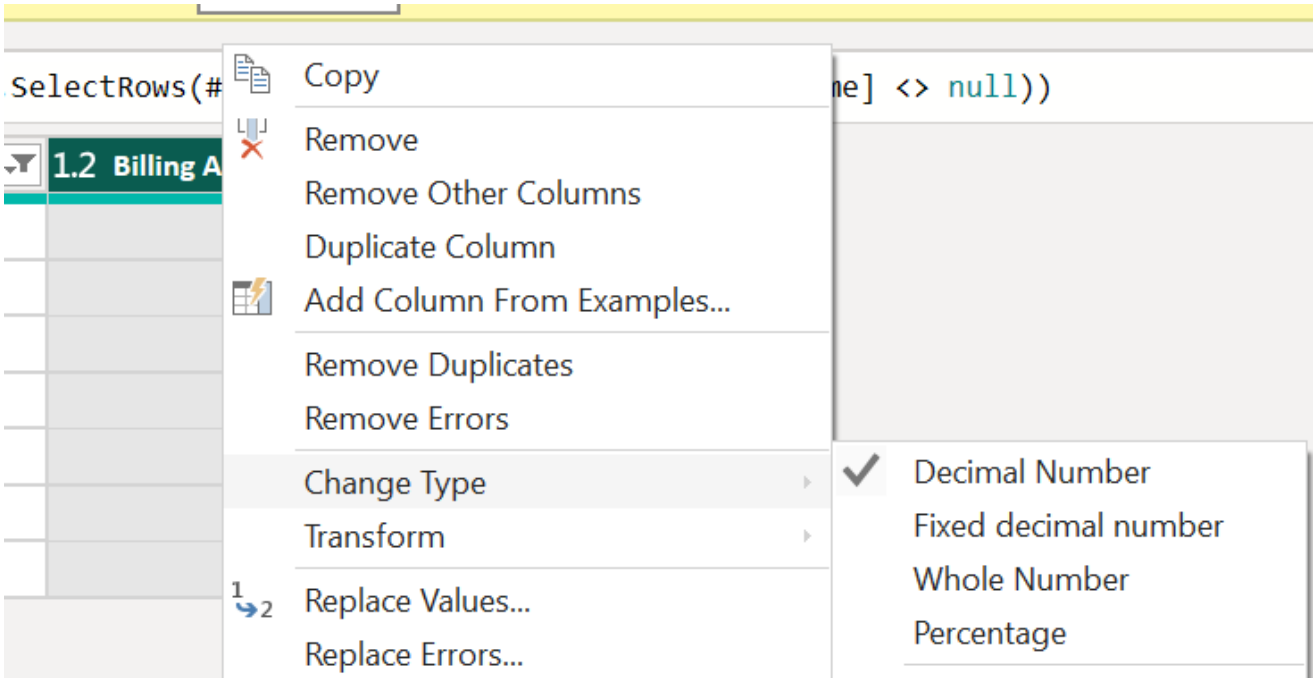
*Hình 347 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

| | Hospital Name | 1.2 Billing Amount |
|---|---|---|
| 1 | PLC Brown | 304480.7954 |
| 2 | Group Smith | 291522.6529 |
| 3 | Smith Ltd | 247134.6459 |
| 4 | Rodriguez Inc | 243599.129 |
| 5 | Smith Inc | 227727.1306 |
| 6 | LLC Smith | 218458.2625 |
| 7 | Martin Inc | 217248.2123 |

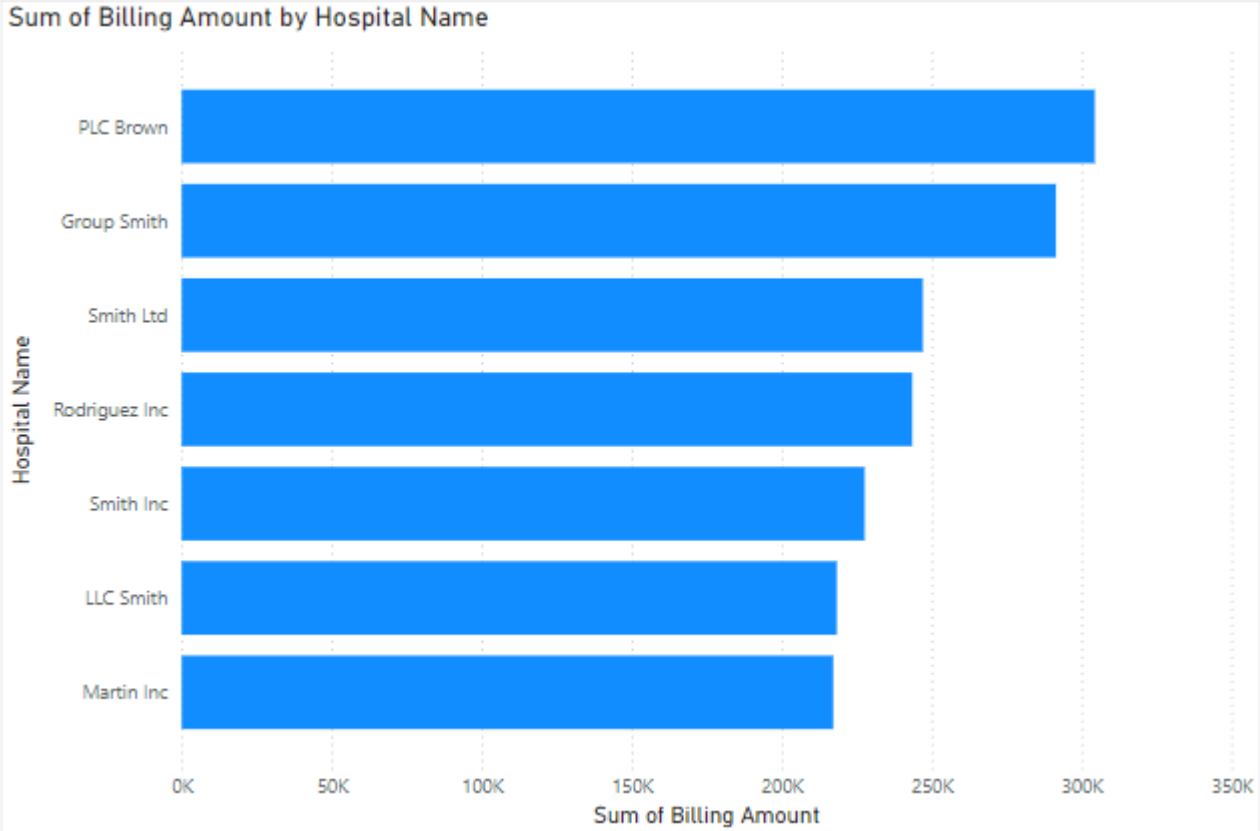
*Hình 348 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 349. Close & Apply*

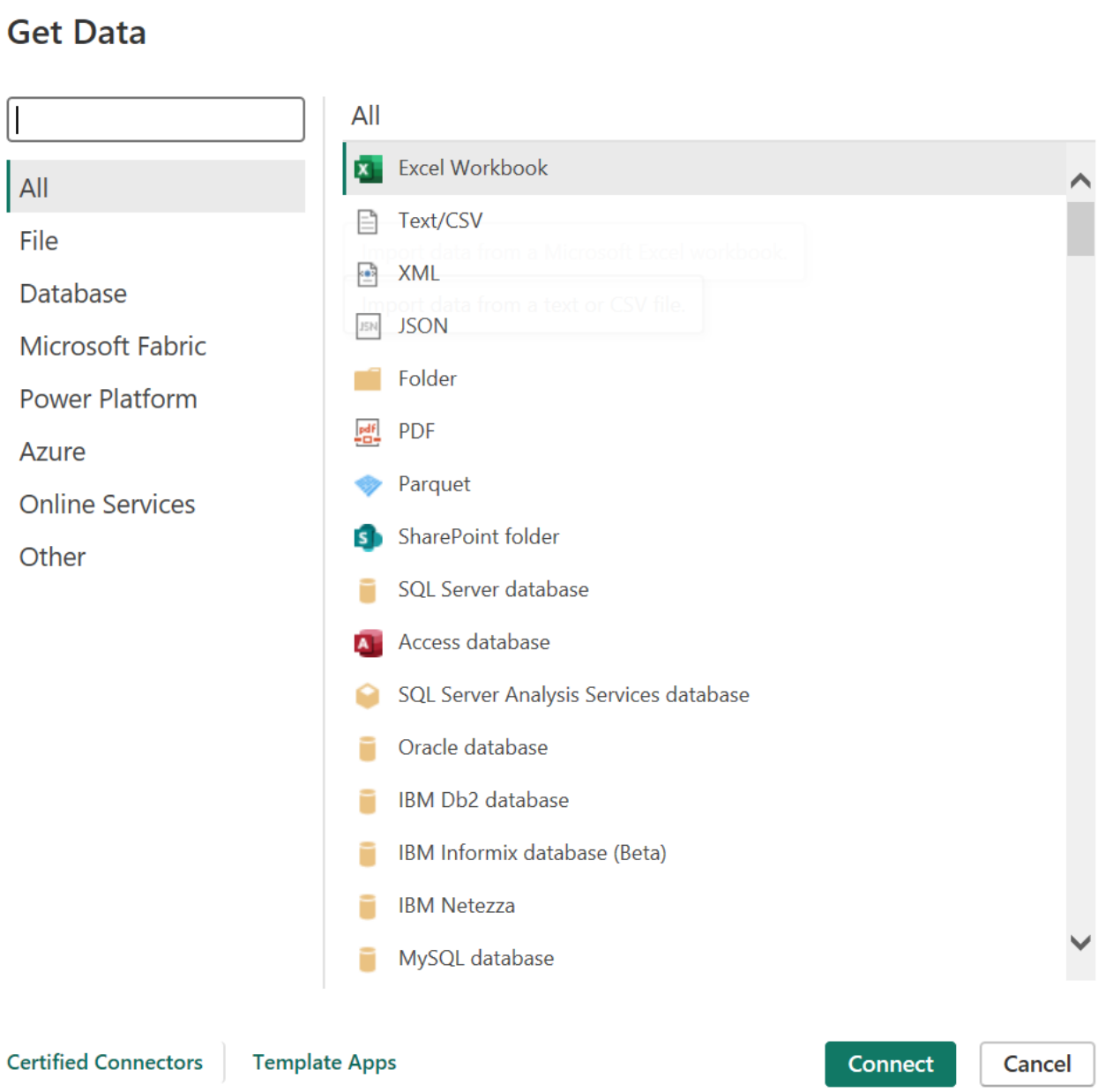
- Bước 6 : Chọn Visualization Stacked Bar Chart và phân bố các thuộc tính như hình dưới




*Hình 350 . Visualization Stacked Bar Chart của Report 12*

## 1.13. REPORT 13

- **Nội dung** ( Câu truy vấn 13) : Đưa ra phần trăm bệnh nhân của từng nhóm máu trong năm 2023.
- **Các bước thực hiện :**
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".




*Hình 351. Get Data Report 13*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.

## SQL Server Analysis Services database

Server

CAROLINE

Database

SSAS

Import

DirectQuery

MDX or DAX query (optional)

```
WITH
    MEMBER [Measures].[Blood Type Patient Percentage] AS
        ([Measures].[FACT Count]) /
        SUM(
            [DIM PATIENT].[Blood Type].CurrentMember.Parent,
            [Measures].[FACT Count]
        ),
        FORMAT_STRING = 'Percent'

SELECT
    {[Measures].[Blood Type Patient Percentage]} ON COLUMNS,
    {[DIM PATIENT].[Blood Type].Children} ON ROWS
FROM [Health Care DW]
```

OK Cancel

*Hình 352 . Import câu lệnh MDX 13*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu

## CAROLINE: SSAS

| [DIM PATIENT].[Blood Type].[Blood Type].[MEMBER_CA | [Measures].[Blood Type Patient Percentage] |
|---|---|
| A- | 0.128423726 |
| A+ | 0.125884274 |
| AB- | 0.126428442 |
| AB+ | 0.129239978 |
| B- | 0.119082169 |
| B+ | 0.123072737 |
| O- | 0.122256485 |
| O+ | 0.125612189 |
| Unknown | null |

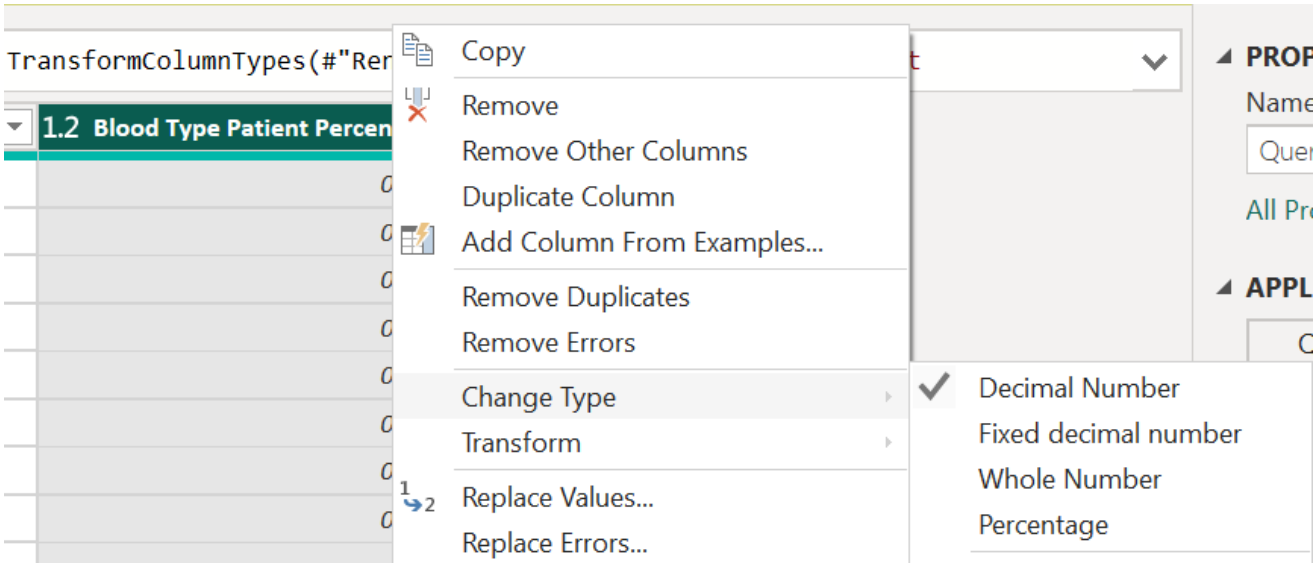
*Hình 353 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính

| | Blood Type | 1.2 Blood Type Patient Percentage |
|---|---|---|
| 1 | A- | 0.128423726 |
| 2 | A+ | 0.125884274 |
| 3 | AB- | 0.126428442 |
| 4 | AB+ | 0.129239978 |
| 5 | B- | 0.119082169 |
| 6 | B+ | 0.123072737 |
| 7 | O- | 0.122256485 |
| 8 | O+ | 0.125612189 |
| 9 | Unknown | null |

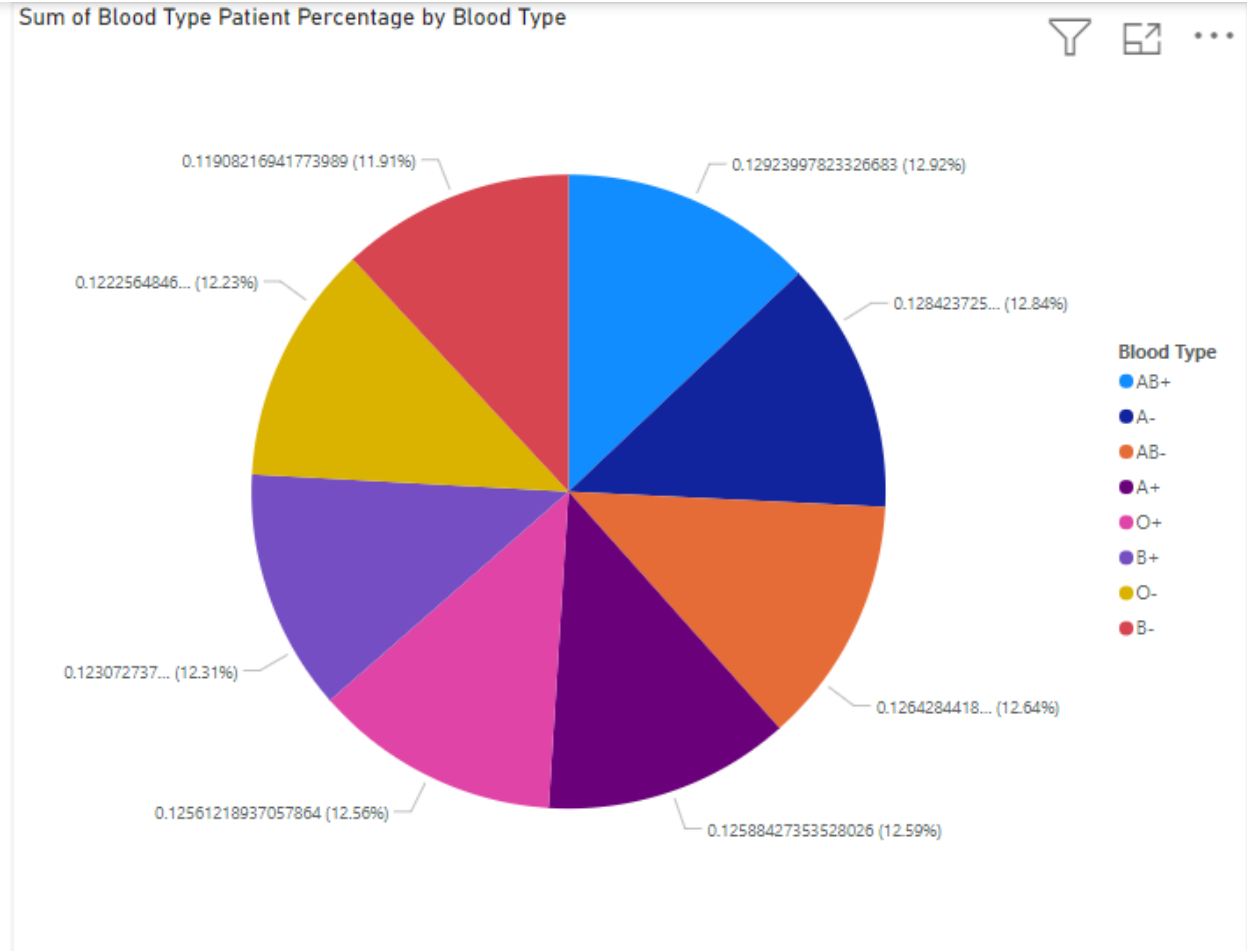
*Hình 354 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Blood Type Patient Percentage sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 355 . Close & Apply*

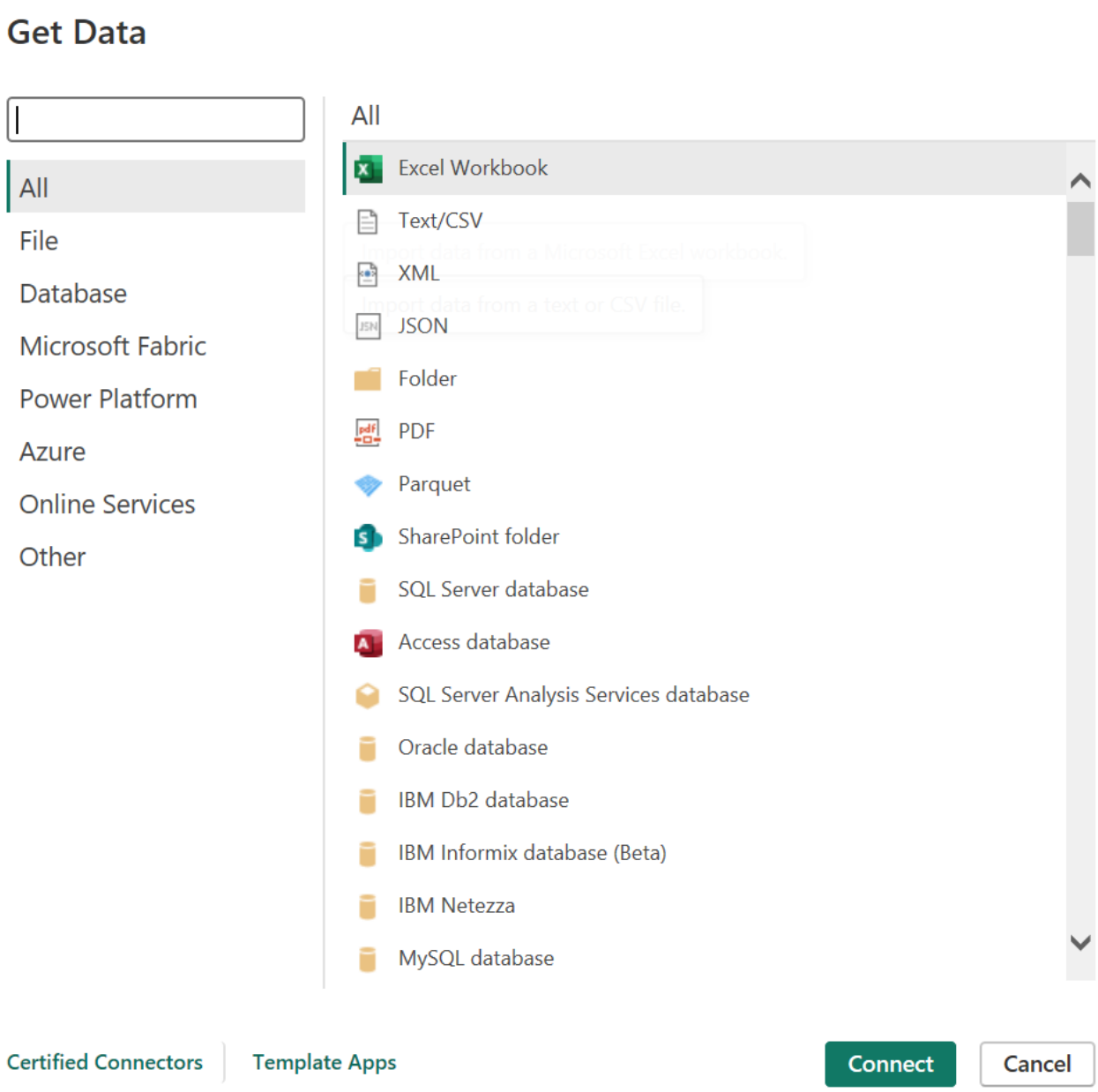
- Bước 6 : Chọn Visualization Pie Chart và phân bố các thuộc tính như hình dưới




*Hình 356 . Visualization Pie Chart của Report 13*

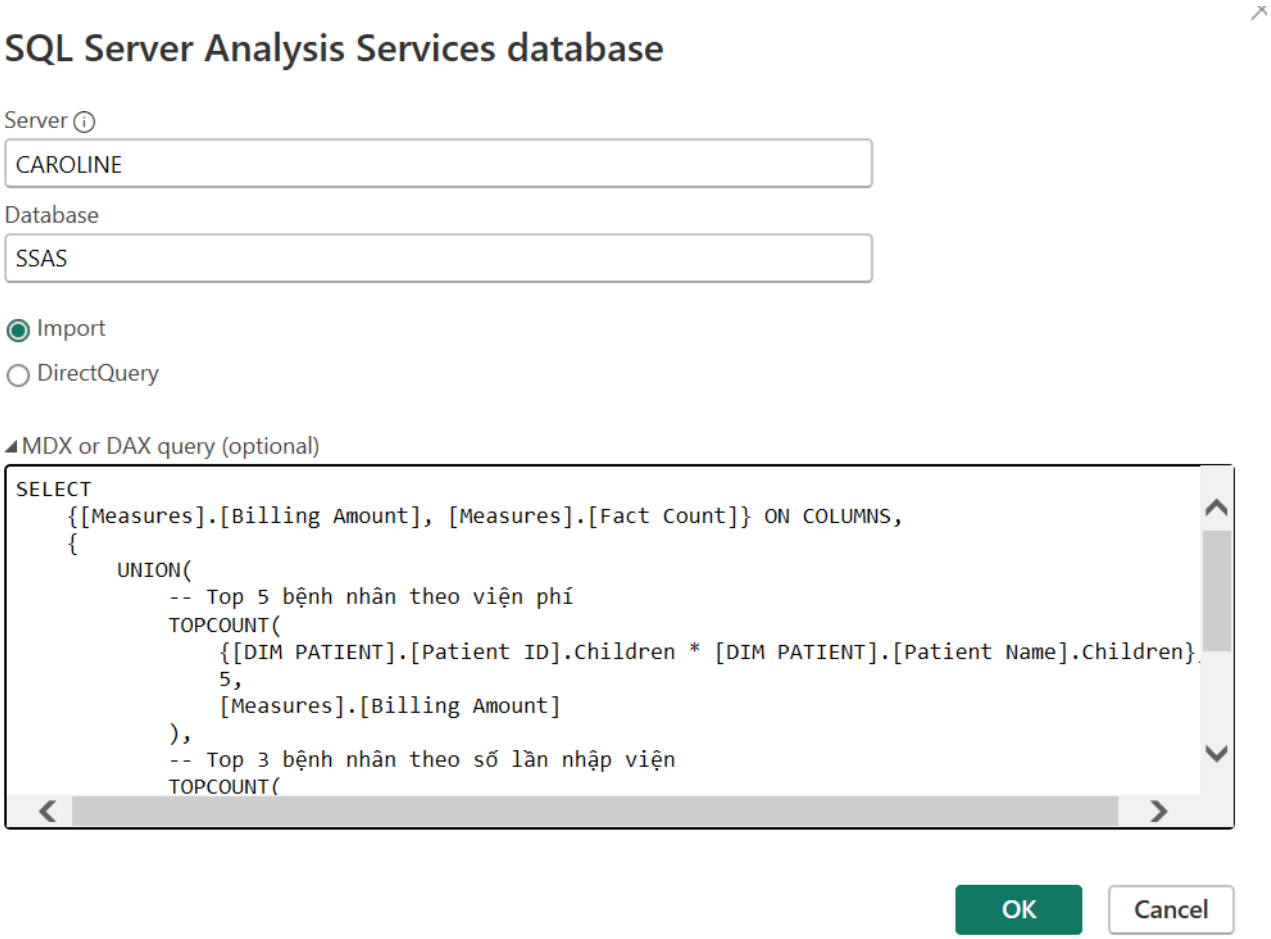
## 1.14. REPORT 14

- Nội dung ( Câu truy vấn 14) : Liệt kê những bệnh nhân thuộc top 5 viện phí cao nhất / top 3 số lần nhập viện nhiều nhất.
- Các bước thực hiện :
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".



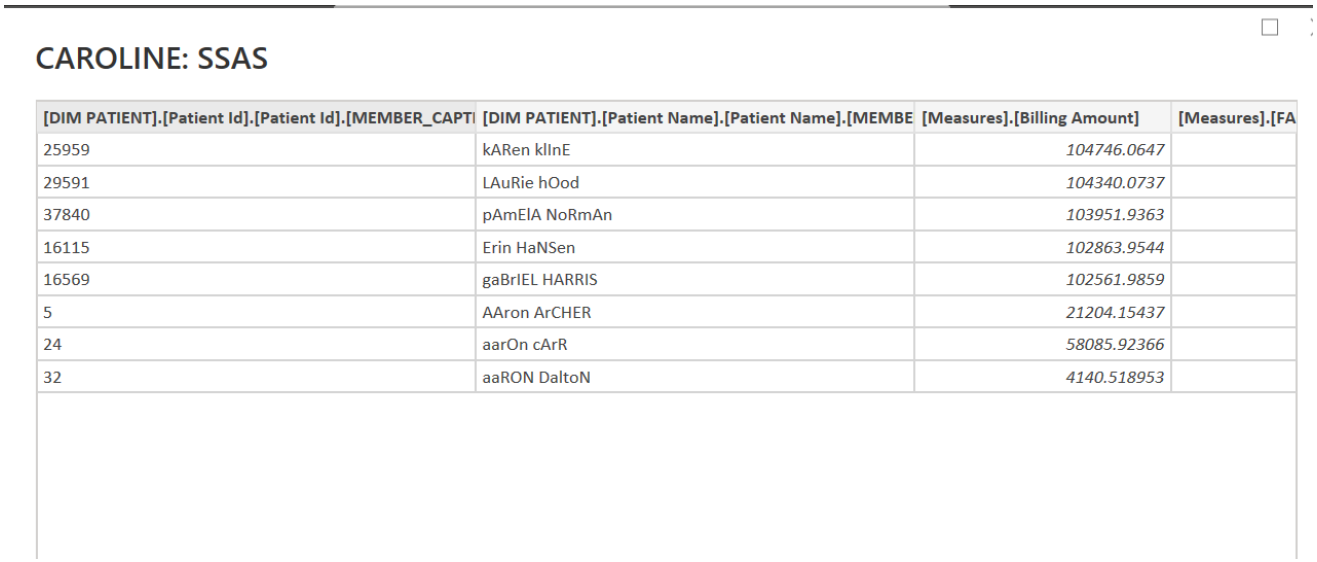

*Hình 357 . Get Data Report 14*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.




*Hình 358 . Import câu lệnh MDX 14*

- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu



CAROLINE: SSAS

| [DIM PATIENT].[Patient Id].[Patient Id].[MEMBER_CAPTI | [DIM PATIENT].[Patient Name].[Patient Name].[MEMBE | [Measures].[Billing Amount] | [Measures].[FA |
|---|---|---|---|
| 25959 | kARen klInE | 104746.0647 | |
| 29591 | LAuRie hOod | 104340.0737 | |
| 37840 | pAmElA NoRmAn | 103951.9363 | |
| 16115 | Erin HaNSen | 102863.9544 | |
| 16569 | gaBrIEL HARRIS | 102561.9859 | |
| 5 | AAron ArCHER | 21204.15437 | |
| 24 | aarOn cArR | 58085.92366 | |
| 32 | aaRON DaltoN | 4140.518953 | |

*Hình 359. Transform Data*

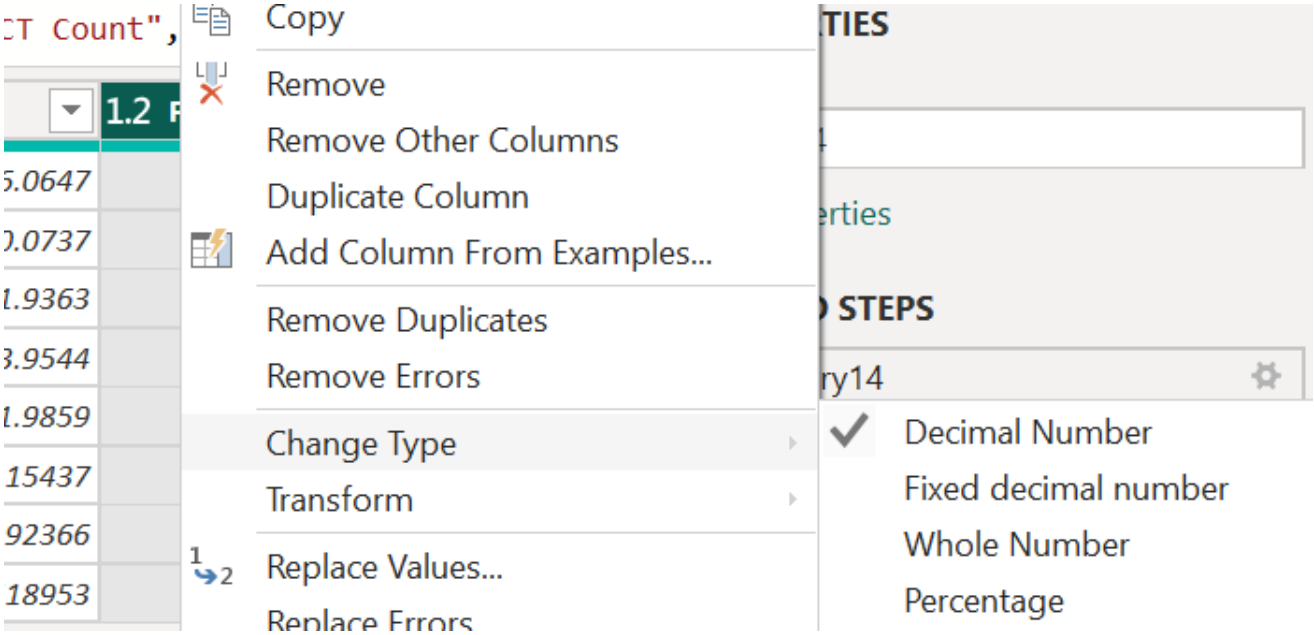
- Bước 4 : Thay đổi tên cho các cột thuộc tính

```
= Table.TransformColumnTypes(#"Renamed Columns",{{"FACT Count", type number},
```

| | Patient Id | Patient Name | Billing Amount | FACT Count |
|---|---|---|---|---|
| 1 | 25959 | kARen klInE | 104746.0647 | 2 |
| 2 | 29591 | LAuRie hOod | 104340.0737 | 2 |
| 3 | 37840 | pAmElA NoRmAn | 103951.9363 | 2 |
| 4 | 16115 | Erin HaNSen | 102863.9544 | 2 |
| 5 | 16569 | gaBrIEL HARRIS | 102561.9859 | 2 |
| 6 | 5 | AAron ArCHER | 21204.15437 | 2 |
| 7 | 24 | aarOn cArR | 58085.92366 | 2 |
| 8 | 32 | aaRON DaltoN | 4140.518953 | 2 |

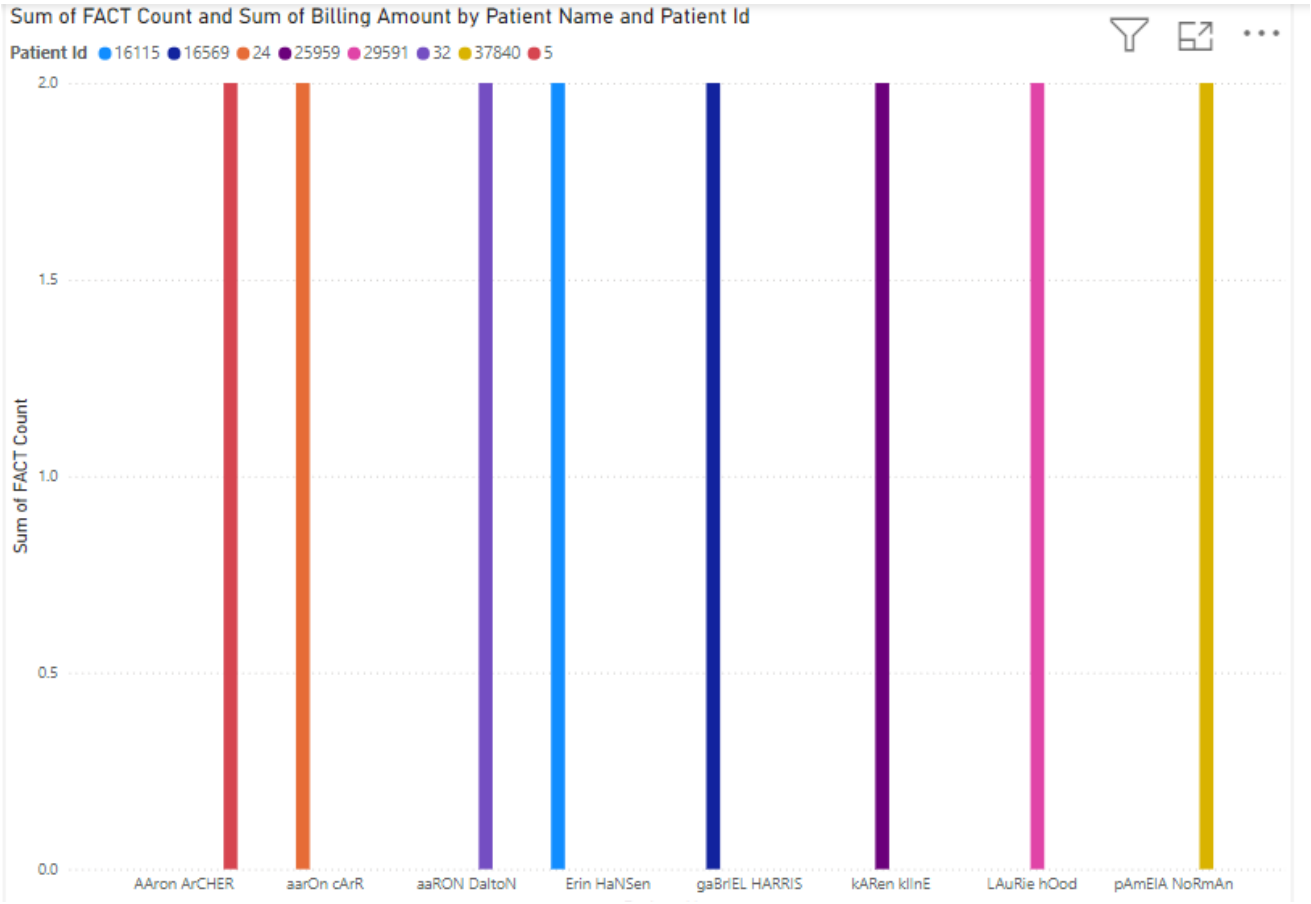
*Hình 360 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột FACT Count sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu.




*Hình 361 . Close & Apply*

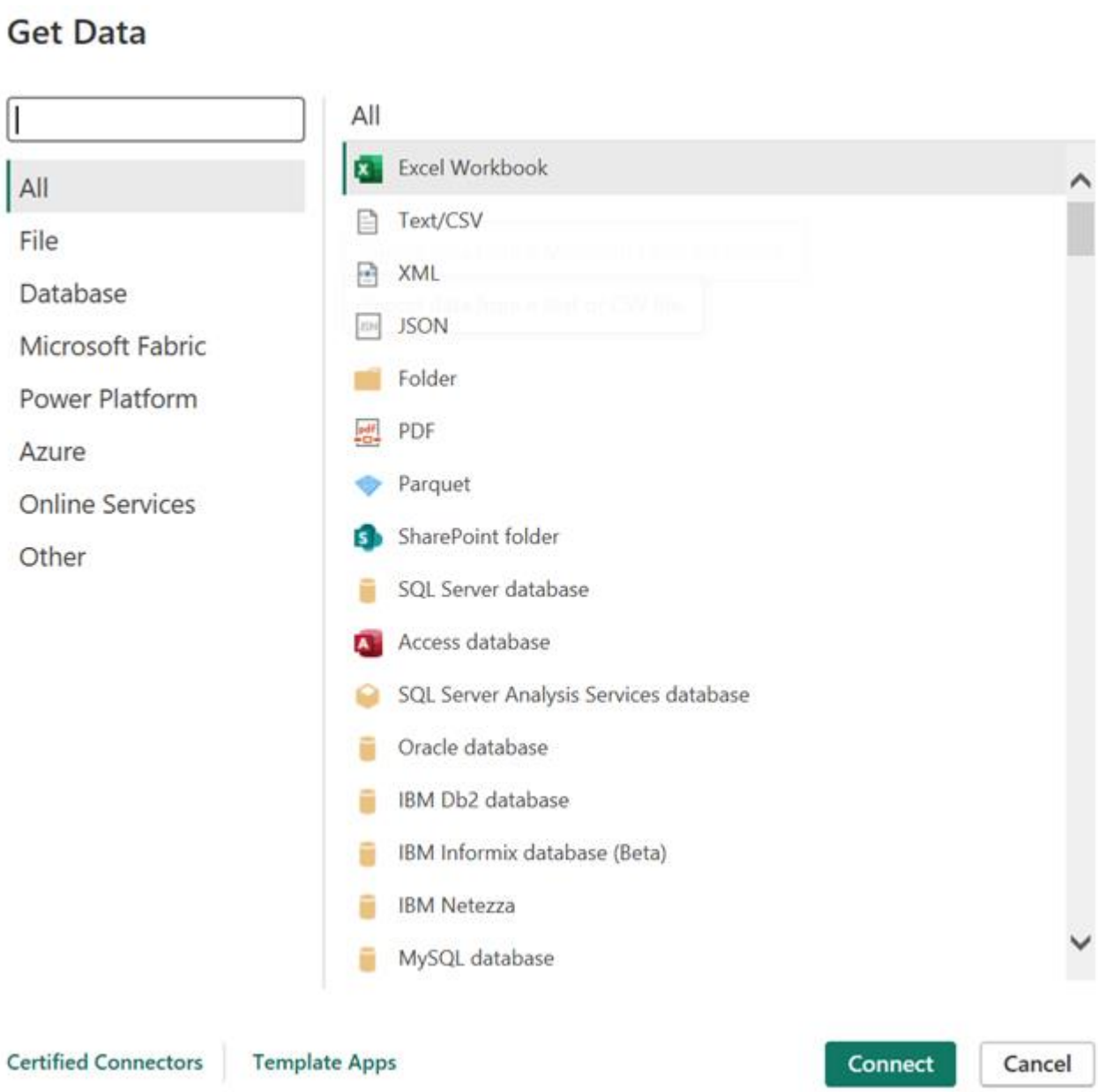
- Bước 6 : Chọn Visualization Clustered Column Chart và phân bố các thuộc tính như hình dưới




*Hình 362 . Visualization Clustered Column Chart của Report 14*

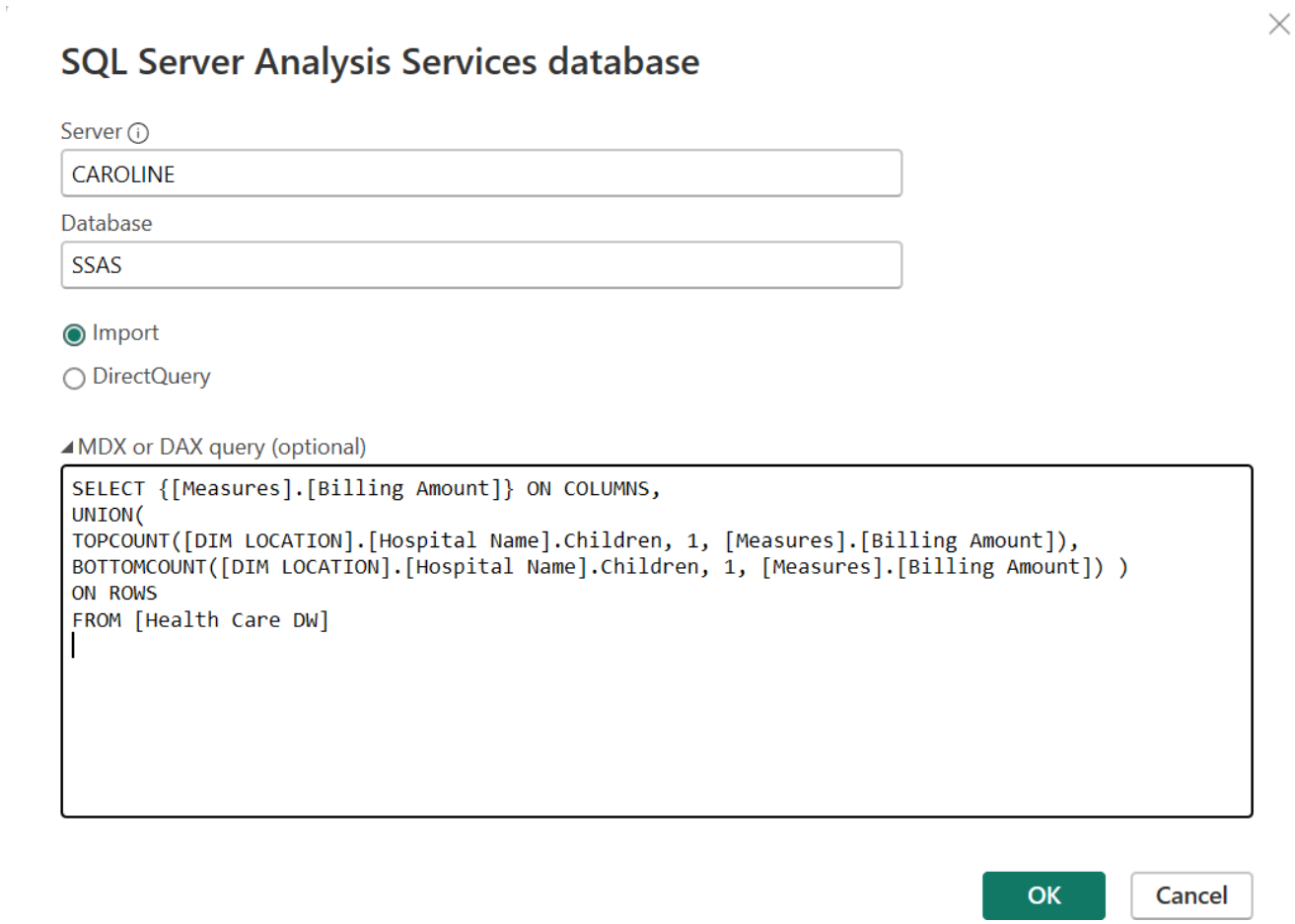
## 1.15. REPORT 15

- **Nội dung** ( Câu truy vấn 15) : Cho biết bệnh viện có doanh thu cao nhất và bệnh viện có doanh thu thấp nhất.
- **Các bước thực hiện** :
    - Bước 1: Tạo dataset bằng cách chọn "Get data" trên menu sau đó chọn "**SQL Server Analysis Services database**".



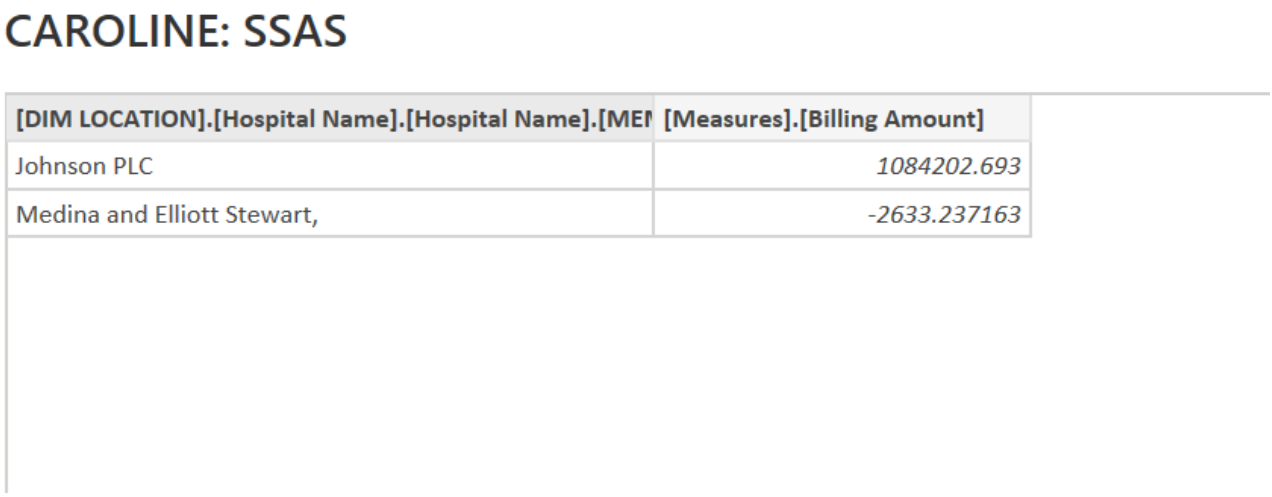

*Hình 363 . Get Data Report 15*

- Bước 2 : Trong cửa sổ mở ra, nhập Server là CAROLINE, Data là SSAS. Chọn mục Import và insert câu lệnh vào ô trống.




*Hình 364 . Import câu lệnh MDX 15*

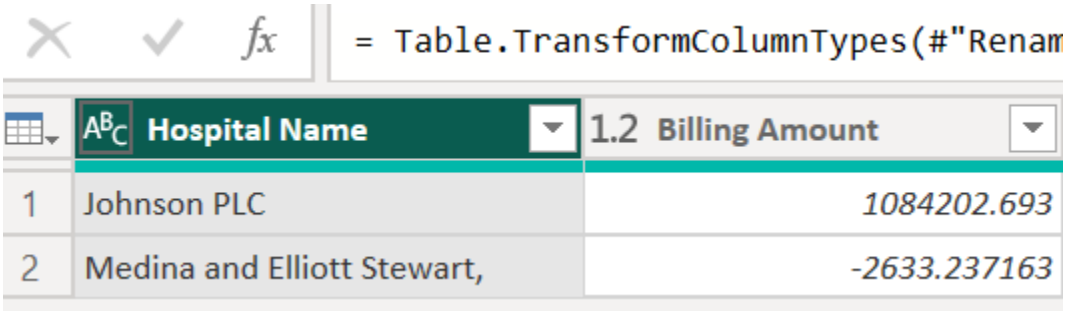
- Bước 3 : Chọn Transform Data để tiến hành biến đổi dữ liệu




| [DIM LOCATION].[Hospital Name].[Hospital Name].[MEI | [Measures].[Billing Amount] |
|---|---|
| Johnson PLC | 1084202.693 |
| Medina and Elliott Stewart, | -2633.237163 |

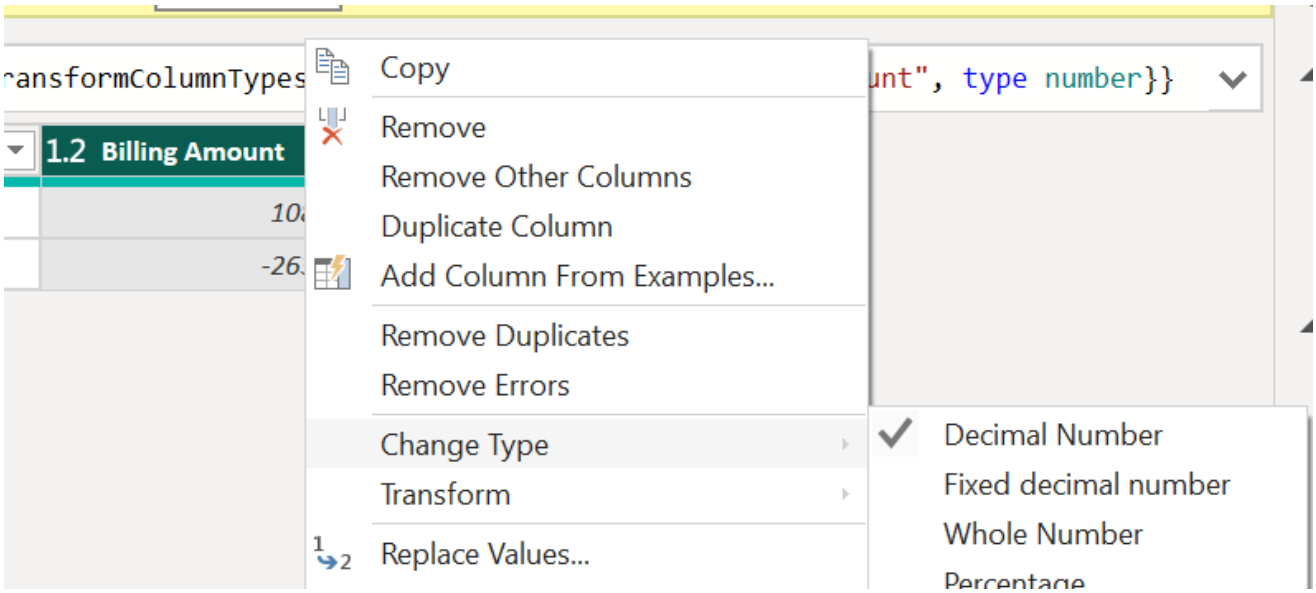
*Hình 365 . Transform Data*

- Bước 4 : Thay đổi tên cho các cột thuộc tính



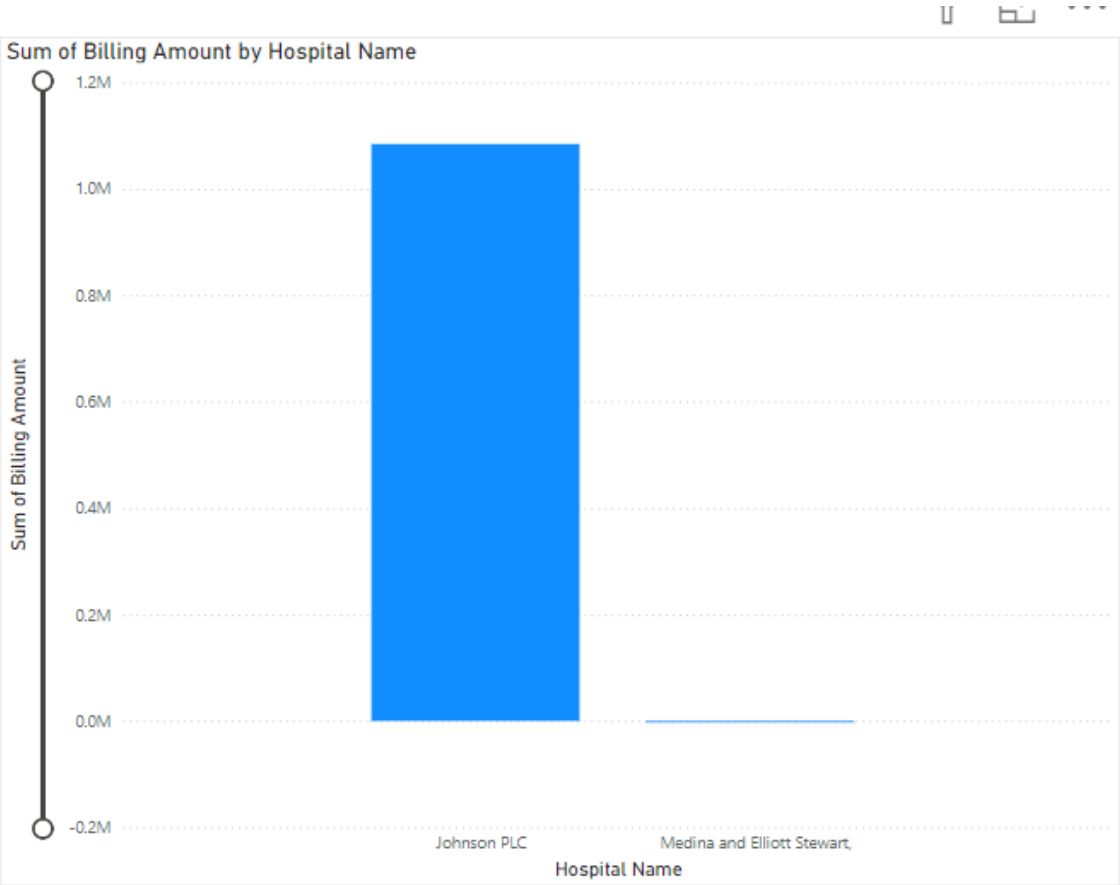

*Hình 366 . Thay đổi tên cho các thuộc tính*

- Bước 5 : Thay đổi kiểu dữ liệu cho cột Billing Amount sang Decimal number. Sau đó chọn Close & Apply để xác nhận dữ liệu




*Hình 367 . Close & Apply*

- Bước 6 : Chọn Visualization Stacked Column Chart  và phân bố các thuộc tính như hình dưới.




*Hình 368 . Visualization Stacked Column Chart của Report 15*

# CHƯƠNG 5. QUÁ TRÌNH DATAMINING

## 1. Tổng quan quá trình

Khai phá dữ liệu (Data Mining) là quá trình phân tích một lượng lớn dữ liệu để phát hiện ra các mẫu, mối quan hệ hữu ích. Mục tiêu là đưa ra những thông tin có giá trị phục vụ các quyết định kinh doanh hoặc dự đoán trong tương lai.

## 2. Giới thiệu bộ dữ liệu

- Kho dữ liệu chăm sóc sức khỏe  được thiết kế để mô phỏng dữ liệu chăm sóc sức khỏe trong thế giới thực, cho phép người dùng thực hành, phát triển và thể hiện kỹ năng xử lý và phân tích dữ liệu của họ trong bối cảnh ngành chăm sóc sức khỏe.
- Thông qua kho dữ liệu người dùng có thể biết được các thông tin về ngày nhập viện, ngày xuất viện của một bệnh nhân ; thông tin bảo hiểm, bệnh lý của một bệnh nhân; đơn thuốc; bác sĩ; viện phí;…
- Kho dữ liệu được xây dựng với hướng chủ đề chăm sóc sức khỏe (Healthcare)
- Kho dữ liệu gồm 55.500 dòng và 15 thuộc tính

## 3. Công cụ khai thác dữ liệu

Các công cụ sử dụng trong Notebook bao gồm:

- **Python:** Công cụ chính để xử lý và khai phá dữ liệu.
- **Thư viện Scikit-learn:** Cung cấp các thuật toán K-Means, Decision Tree và Random Forest.
- **Matplotlib & Seaborn:** Để trực quan hóa dữ liệu.
- **Pandas:** Xử lý dữ liệu dạng bảng.

## 4. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng trong quá trình khai thác dữ liệu, nhằm chuẩn bị dữ liệu ở dạng sạch và thích hợp để phân tích. Các bước tiền xử lý dữ liệu bao gồm:

- **Mã hóa:** Sử dụng LabelEncoder để chuyển đổi cột phân loại Medical Condition thành dạng số.
- **Điền giá trị thiếu:** Sử dụng giá trị trung bình của cột tương ứng.
- **Chuẩn hóa:** Chuẩn hóa dữ liệu bằng StandardScaler để đưa tất cả các đặc trưng về cùng thang đo.

```python
# Bước 2: Tiền xử lý dữ liệu
# Mã hóa cột 'Medical Condition' (chuyển thành số)
label_encoder = LabelEncoder()
data_cleaned['Medical Condition'] = label_encoder.fit_transform(data_cleaned['Medical Condition'])

# Điền giá trị thiếu (nếu có)
data_cleaned.fillna(data_cleaned.mean(), inplace=True)

# Chuẩn hóa dữ liệu
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_cleaned)
```

```
<ipython-input-6-93340d519191>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data_cleaned['Medical Condition'] = label_encoder.fit_transform(data_cleaned['Medical Condition'])
<ipython-input-6-93340d519191>:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  data_cleaned.fillna(data_cleaned.mean(), inplace=True)
```

*Hình 369. Tiền xử lý dữ liệu*

## 5. Khai thác dữ liệu

### 5.1. Thuật toán K-Means

- **K-Means** là thuật toán phân cụm:
  - **Số cụm tối ưu:** Xác định bằng Elbow Method.
  - **Kết quả:** Các cụm dữ liệu được phân nhóm dựa trên các thuộc tính Age, Billing Amount, và Medical Condition.
- **Các bước thực hiện** :
  - Bước 1 : Đọc dữ liệu và Chọn các cột đầu vào

```python
# Import các thư viện cần thiết
import pandas as pd
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns
```

+ Code  + Markdown

```python
# Bước 1: Đọc dữ liệu và chọn cột cần thiết
file_name = '/content/drive/MyDrive/ĐẠI HỌC/NĂM 3/OLAP/healthcare_dataset.csv'  # Thay bằng tên file của bạn
dataset = pd.read_csv(file_name)

# Chọn các cột đầu vào (ví dụ: Age, Billing Amount, Medical Condition)
data_cleaned = dataset[['Age', 'Billing Amount', 'Medical Condition']]
```

*Hình 370. Bước 1_Thuật toán K- Means*

  - Bước 2 : Tiền xử lý dữ liệu

```python
# Bước 2: Tiền xử lý dữ liệu
# Mã hóa cột 'Medical Condition' (chuyển thành số)
label_encoder = LabelEncoder()
data_cleaned['Medical Condition'] = label_encoder.fit_transform(data_cleaned['Medical Condition'])

# Điền giá trị thiếu (nếu có)
data_cleaned.fillna(data_cleaned.mean(), inplace=True)

# Chuẩn hóa dữ liệu
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_cleaned)
```

*Hình 371. Tiền xử lý dữ liệu*

- Bước 3 : Xác định số cụm bằng Elbow Method

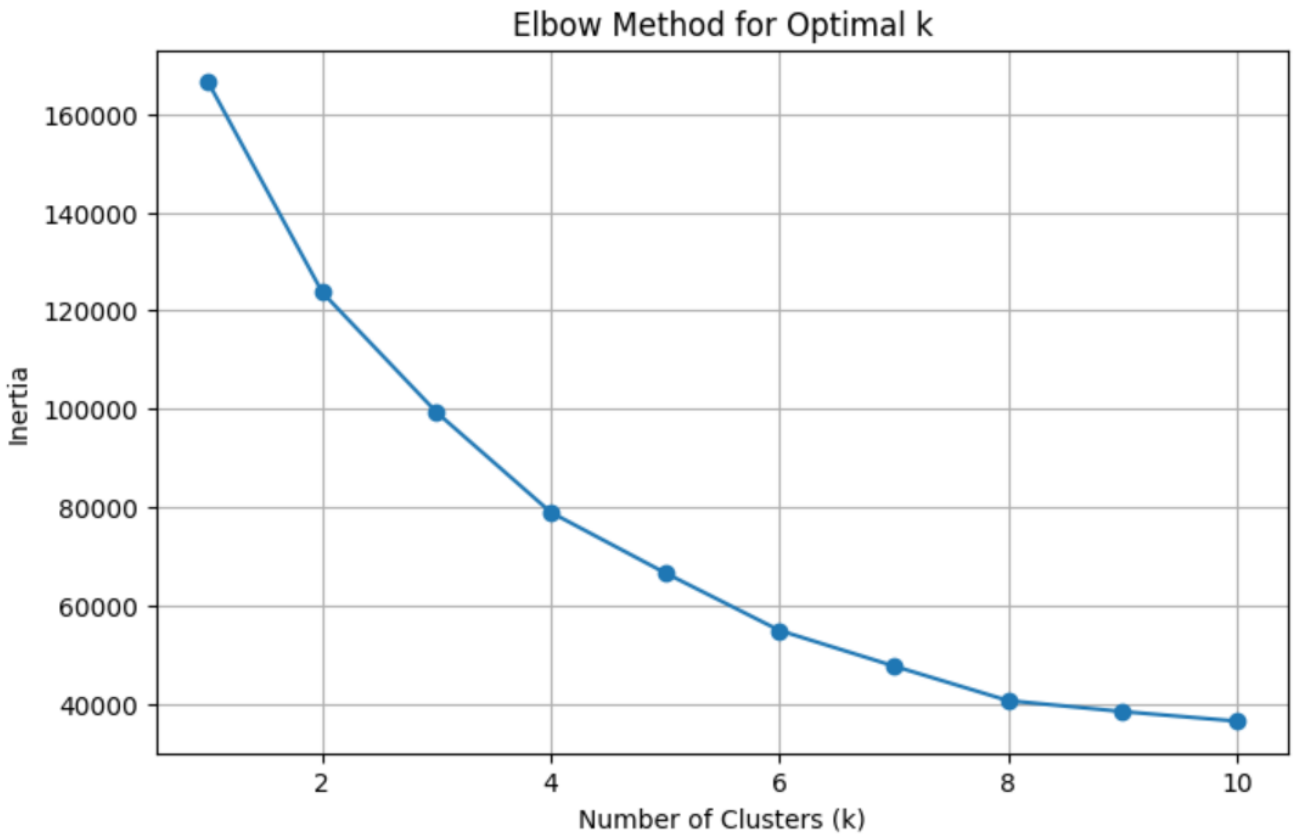
```python
inertia = []
k_range = range(1, 11)
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_scaled)
    inertia.append(kmeans.inertia_)
```

*Hình 372. Xác định số cụm*

- Bước 4 : Vẽ biểu đồ Elbow

```python
plt.figure(figsize=(8, 5))
plt.plot(k_range, inertia, marker='o')
plt.title('Elbow Method for Optimal k')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.grid()
plt.show()
```

*Hình 373. Vẽ biểu đồ Elbow*




*Hình 374. Biểu đồ Elbow*

- Bước 5 : Thực hiện K-Means Clustering

```
# Chọn số cụm dựa trên Elbow Method
k = 4
kmeans = KMeans(n_clusters=k, random_state=42)
clusters = kmeans.fit_predict(data_scaled)

# Thêm cột 'Cluster' vào dataset gốc
dataset['Cluster'] = clusters
```

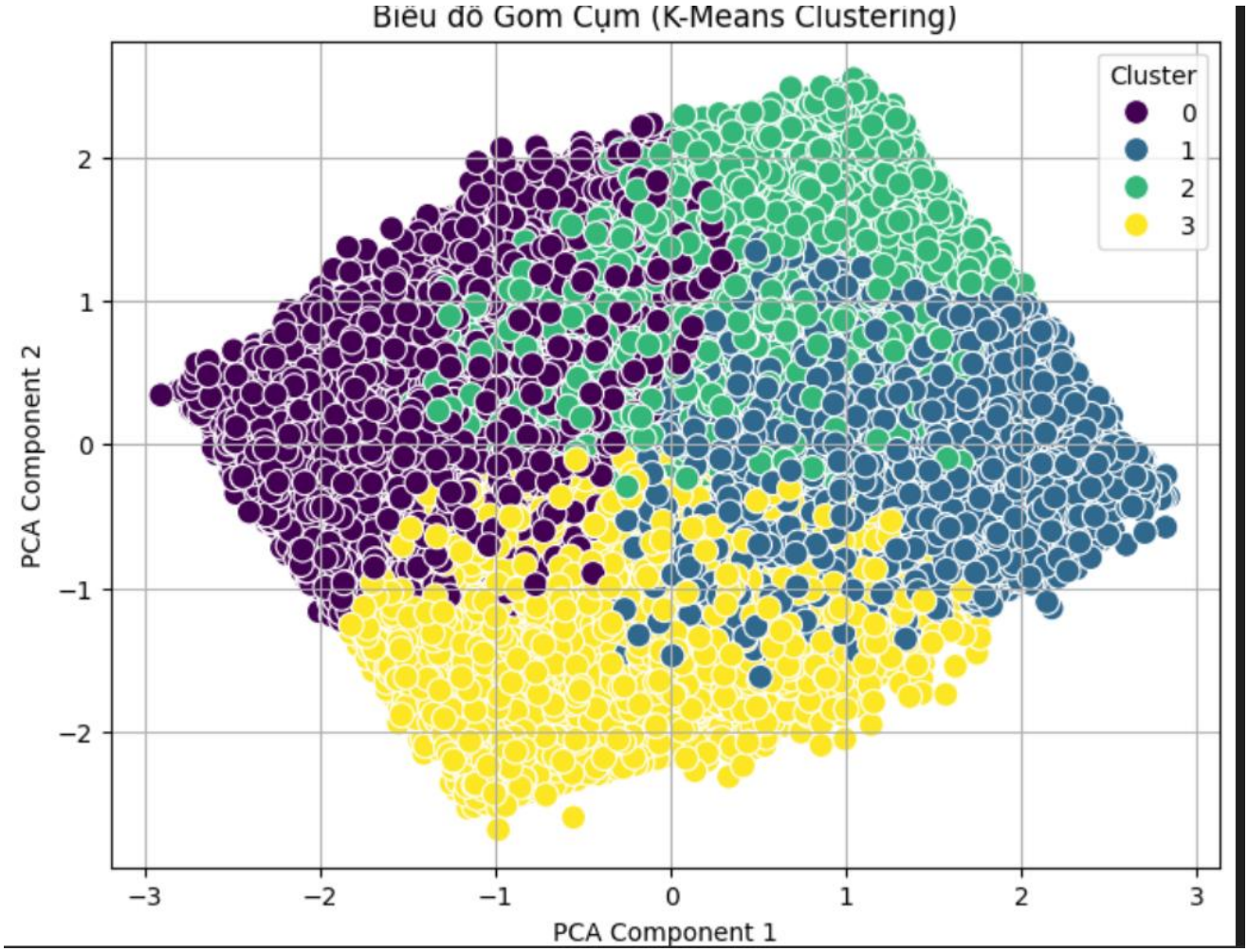
*Hình 375. K-Means Clustering*

- Bước 6 : Vẽ sơ đồ gom cụm

```python
# Giảm chiều dữ liệu xuống 2D bằng PCA
pca = PCA(n_components=2)
data_2d = pca.fit_transform(data_scaled)

# Vẽ biểu đồ các cụm
plt.figure(figsize=(8, 6))
sns.scatterplot(
    x=data_2d[:, 0],
    y=data_2d[:, 1],
    hue=dataset['Cluster'],
    palette='viridis',
    s=100
)
plt.title('Biểu đồ Gom Cụm (K-Means Clustering)')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.legend(title='Cluster', loc='best')
plt.grid()
plt.show()
```

*Hình 376. Vẽ sơ đồ gom cụm*




*Hình 377. Biểu đồ gom cụm*

- Bước 7 : Xuất dữ liệu ra file Excel

```
output_file = 'healthcare_clustered_output.xlsx'
dataset.to_excel(output_file, index=False)
```

*Hình 378. Xuất file Excel*

### 5.2. Thuật toán Decision Tree và Random Forest

- Thuật toán Decision Tree:

- **Ứng dụng:** Phân loại dữ liệu.
- **Phân tách:** Sử dụng Entropy hoặc Gini Index để tối ưu hóa việc chia nhánh.

- Các bước thực hiện :
  - Bước 1 : Import các thư viện cần thiết

```python
# Import các thư viện cần thiết
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
import matplotlib.pyplot as plt
import seaborn as sns
```

*Hình 379. Import thư viện*

- Bước 2 : Xử lý dữ liệu

```python
# ----- 1. Xử lý dữ liệu -----

# Đọc dataset
file_name = '/content/drive/MyDrive/ĐẠI HỌC/NĂM 3/OLAP/healthcare_dataset.csv'
dataset = pd.read_csv(file_name)

# Chuyển đổi các cột chuỗi sang dạng số bằng Label Encoding
label_columns = [
    'Blood Type',  'Doctor',
     'Hospital'
]

le = LabelEncoder()
for col in label_columns:
    dataset[col] = le.fit_transform(dataset[col])

# Loại bỏ các cột không cần thiết
dataset = dataset.drop(columns=['Name', 'Date of Admission', 'Discharge Date','Test Results', 'Medication','Admission Type', 'Insurance Provider', '

# Tách dữ liệu thành X (các thuộc tính đầu vào) và y (biến mục tiêu)
X = dataset.drop(columns=['Billing Amount'])  # Thuộc tính đầu vào
y = dataset['Billing Amount']  # Biến mục tiêu
```

*Hình 380. Xử lý dữ liệu (1)*

```python
# Tách dữ liệu thành X (các thuộc tính đầu vào) và y (biến mục tiêu)
X = dataset.drop(columns=['Billing Amount'])  # Thuộc tính đầu vào
y = dataset['Billing Amount']  # Biến mục tiêu

# Nếu còn cột dạng object, sử dụng LabelEncoder để mã hóa
categorical_cols = X.select_dtypes(include=['object']).columns.tolist()
if categorical_cols:
    for col in categorical_cols:
        X[col] = le.fit_transform(X[col])

# Chuẩn hóa dữ liệu số
numerical_cols = X.select_dtypes(include=np.number).columns.tolist()
scaler = StandardScaler()
X_scaled = X.copy()
X_scaled[numerical_cols] = scaler.fit_transform(X[numerical_cols])

# Chia tập dữ liệu thành training set và testing set
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)
# Chuyển đổi Billing Amount thành nhãn phân loại
y_train = (y_train > y_train.median()).astype(int)
y_test = (y_test > y_test.median()).astype(int)
```

*Hình 381. Xử lý dữ liệu (2)*

- Bước 3 :  Huấn luyện mô hình

```python
# Decision Tree Classifier
dt_model = DecisionTreeClassifier(max_depth=5, random_state=42)
dt_model.fit(X_train, y_train)

# Random Forest Classifier
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

*Hình 382. Huấn luyện mô hình*

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

*Hình 383. Kết quả huấn luyện mô hình*

- Bước 4 : Đánh giá mô hình

```python
# Dự đoán trên tập kiểm tra
y_pred_dt = dt_model.predict(X_test)
y_pred_rf = rf_model.predict(X_test)

# Đánh giá Decision Tree
print("Decision Tree Accuracy:", accuracy_score(y_test, y_pred_dt))
print("Decision Tree Classification Report:\n", classification_report(y_test, y_pred_dt))

# Đánh giá Random Forest
print("\nRandom Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("Random Forest Classification Report:\n", classification_report(y_test, y_pred_rf))
```

*Hình 384. Đánh giá mô hình*

```
...  Decision Tree Accuracy: 0.5018018018018018
     Decision Tree Classification Report:
                   precision    recall  f1-score   support

               0       0.50      0.83      0.62      8325
               1       0.51      0.18      0.26      8325

        accuracy                           0.50     16650
       macro avg       0.50      0.50      0.44     16650
    weighted avg       0.50      0.50      0.44     16650


     Random Forest Accuracy: 0.5693093093093093
     Random Forest Classification Report:
                   precision    recall  f1-score   support

               0       0.57      0.59      0.58      8325
               1       0.57      0.54      0.56      8325

        accuracy                           0.57     16650
       macro avg       0.57      0.57      0.57     16650
    weighted avg       0.57      0.57      0.57     16650
```

*Hình 385. Kết quả đánh giá mô hình*

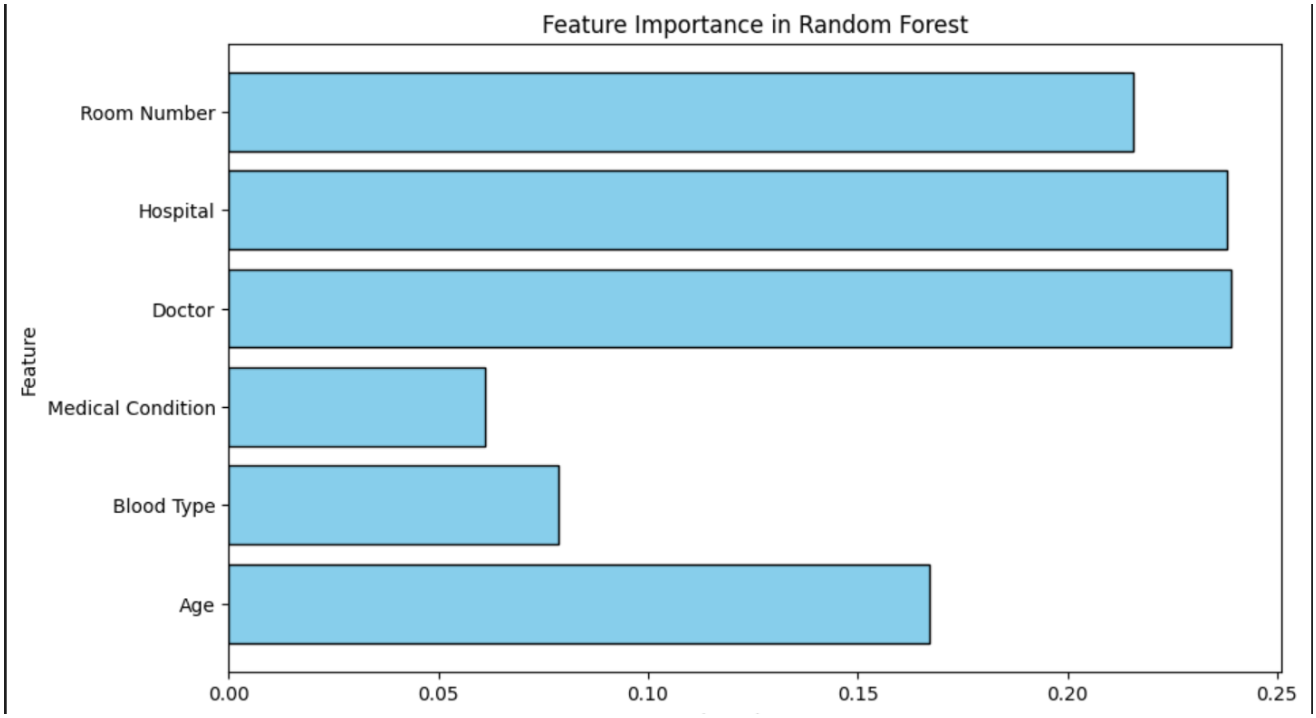
- Bước 5 : Phân tích Feature Importance

```
# ----- 4. Phân tích Feature Importance -----

# Tầm quan trọng của thuộc tính với Random Forest
feature_importances = rf_model.feature_importances_
features = X.columns

# Vẽ biểu đồ tầm quan trọng của thuộc tính
plt.figure(figsize=(10, 6))
plt.barh(features, feature_importances, color='skyblue', edgecolor='black')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Feature Importance in Random Forest')
plt.show()
```

*Hình 386. Phân tích Feature Importance*




*Hình 387. Biểu đồ tầm quan trọng với thuộc tính*

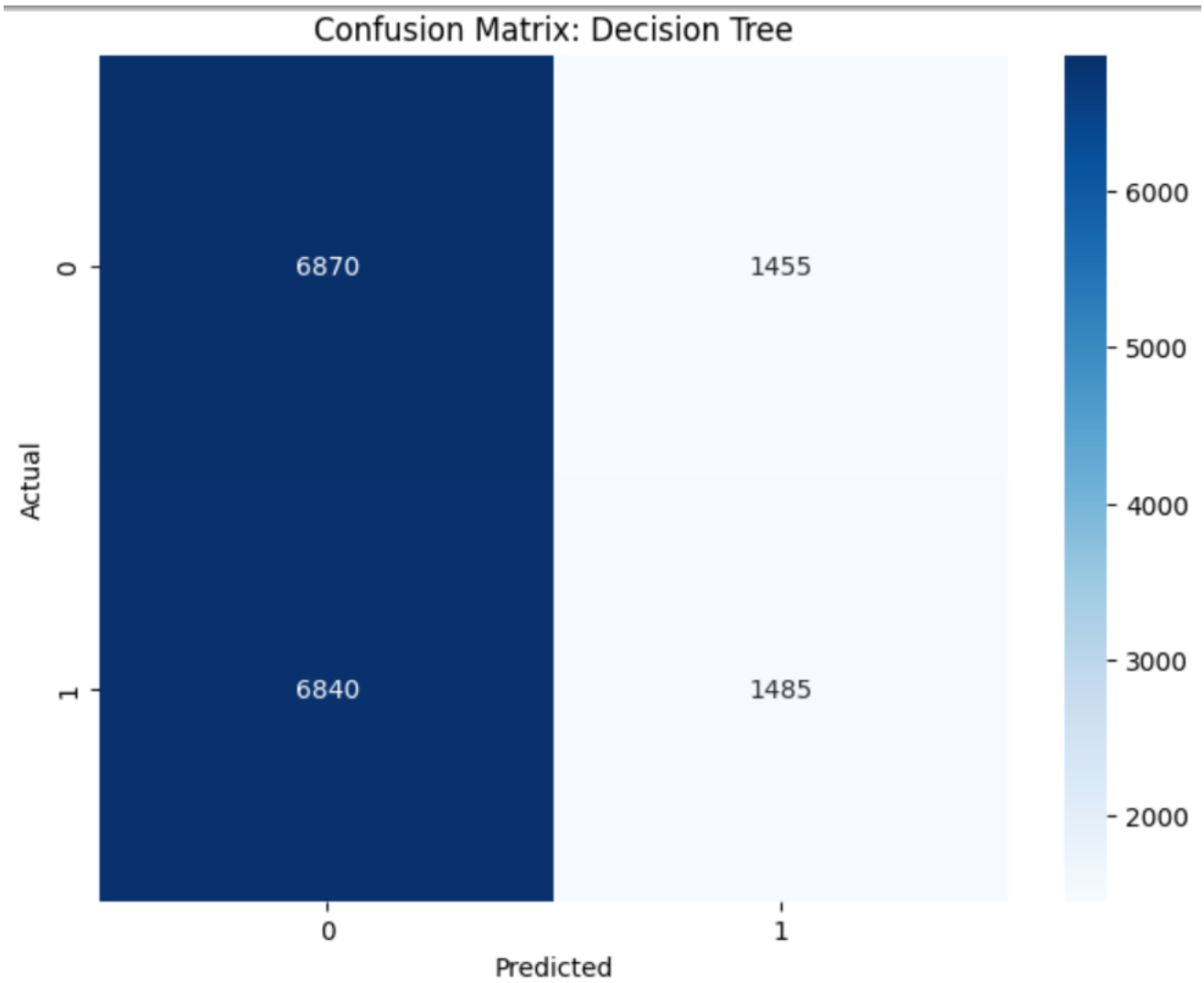
- Bước 6 : Trực quan hóa ma trận nhầm lẫn

```
# ----- 5. Trực quan hóa Ma trận Nhầm Lẫn -----

# Ma trận nhầm lẫn cho Decision Tree
conf_matrix_dt = pd.crosstab(y_test, y_pred_dt, rownames=['Actual'], colnames=['Predicted'])
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_dt, annot=True, fmt="d", cmap="Blues")
plt.title("Confusion Matrix: Decision Tree")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()

# Ma trận nhầm lẫn cho Random Forest
conf_matrix_rf = pd.crosstab(y_test, y_pred_rf, rownames=['Actual'], colnames=['Predicted'])
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_rf, annot=True, fmt="d", cmap="Greens")
plt.title("Confusion Matrix: Random Forest")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()
```

*Hình 388. Trực quan hóa ma trận nhầm lẫn*



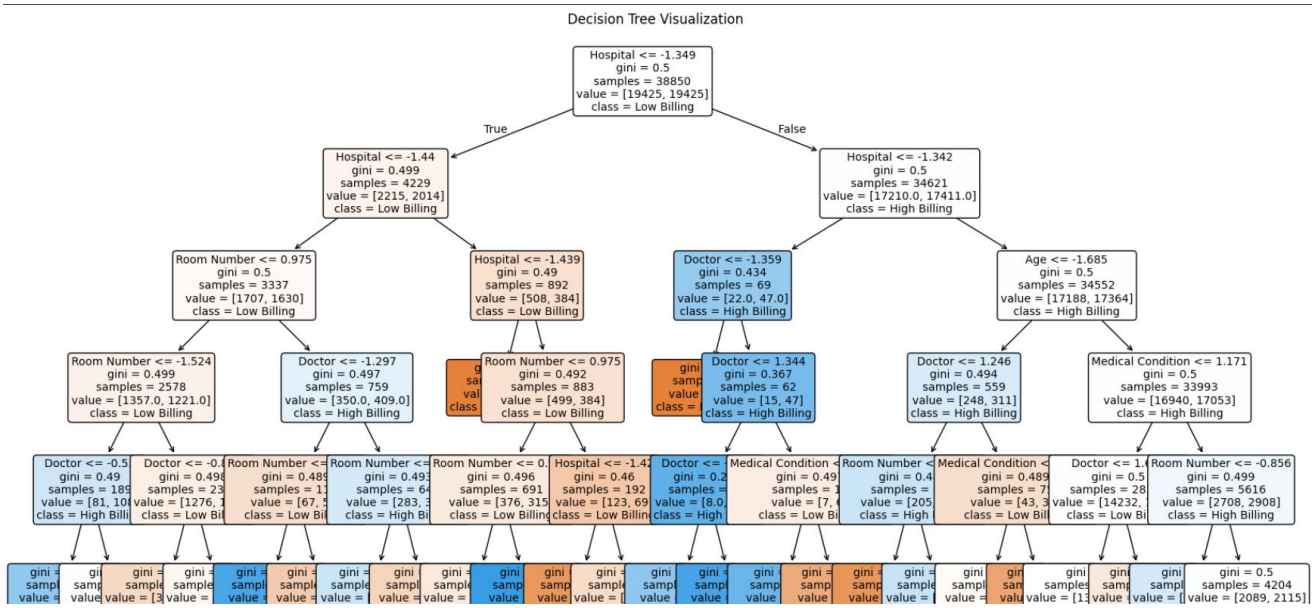

*Hình 389. Ma trận nhầm lẫn Decision Tree*

- Bước 7 : In ra cây quyết định

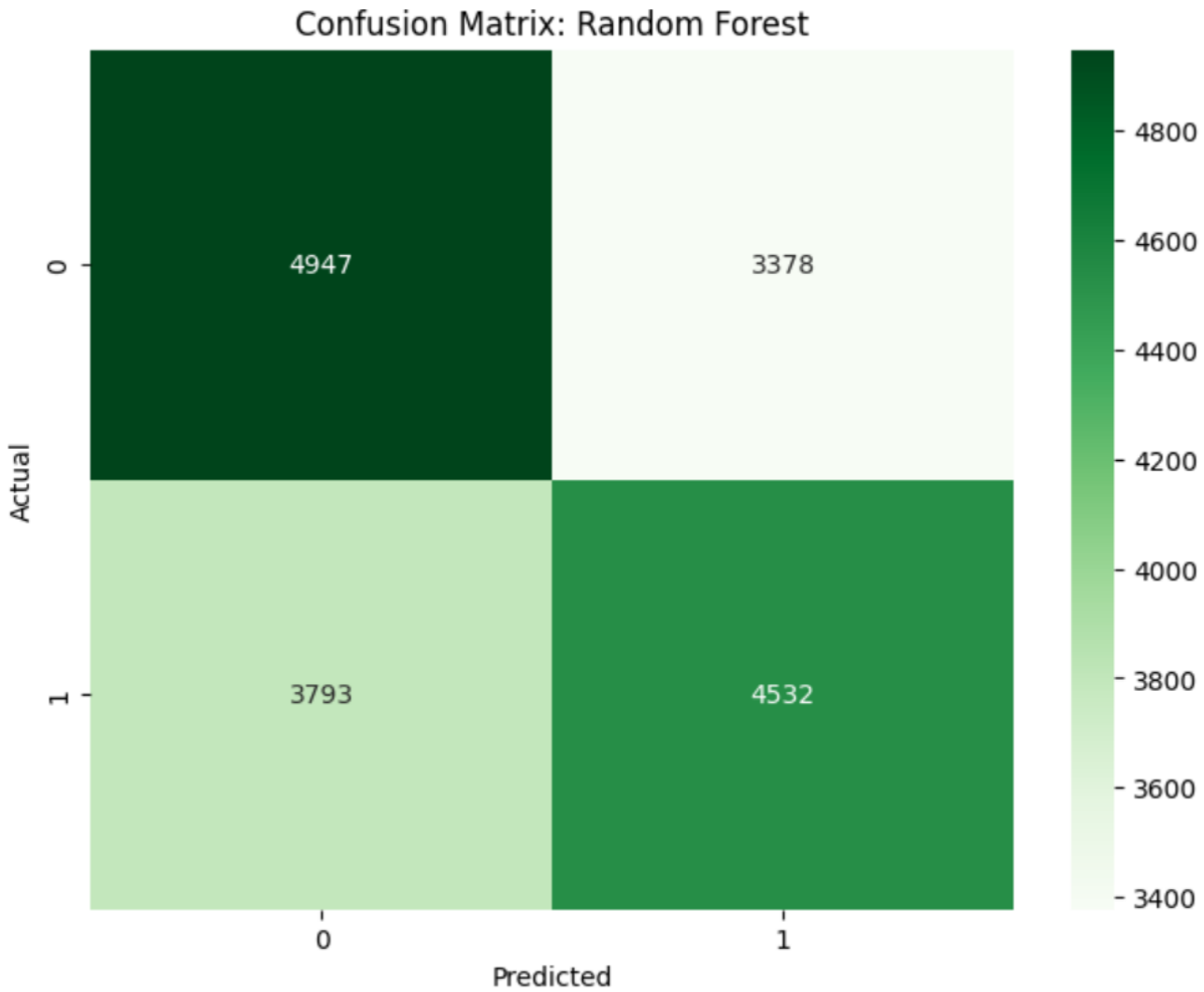
```
from sklearn import tree
import matplotlib.pyplot as plt

# Trực quan hóa Decision Tree bằng cách sử dụng plot_tree
plt.figure(figsize=(20, 10))  # Tùy chỉnh kích thước hình vẽ
tree.plot_tree(
    dt_model,
    feature_names=X.columns,
    class_names=["Low Billing", "High Billing"],
    filled=True,
    rounded=True,
    fontsize=10
)
plt.title("Decision Tree Visualization")
plt.show()
```

*Hình 390. In ra cây quyết định*




*Hình 391. Decision Tree Visualization*




*Hình 392. Ma trận nhầm lẫn Random Forest*

## 6. Phân tích kết quả

### 6.1. Phân tích kết quả thuật toán K-Means

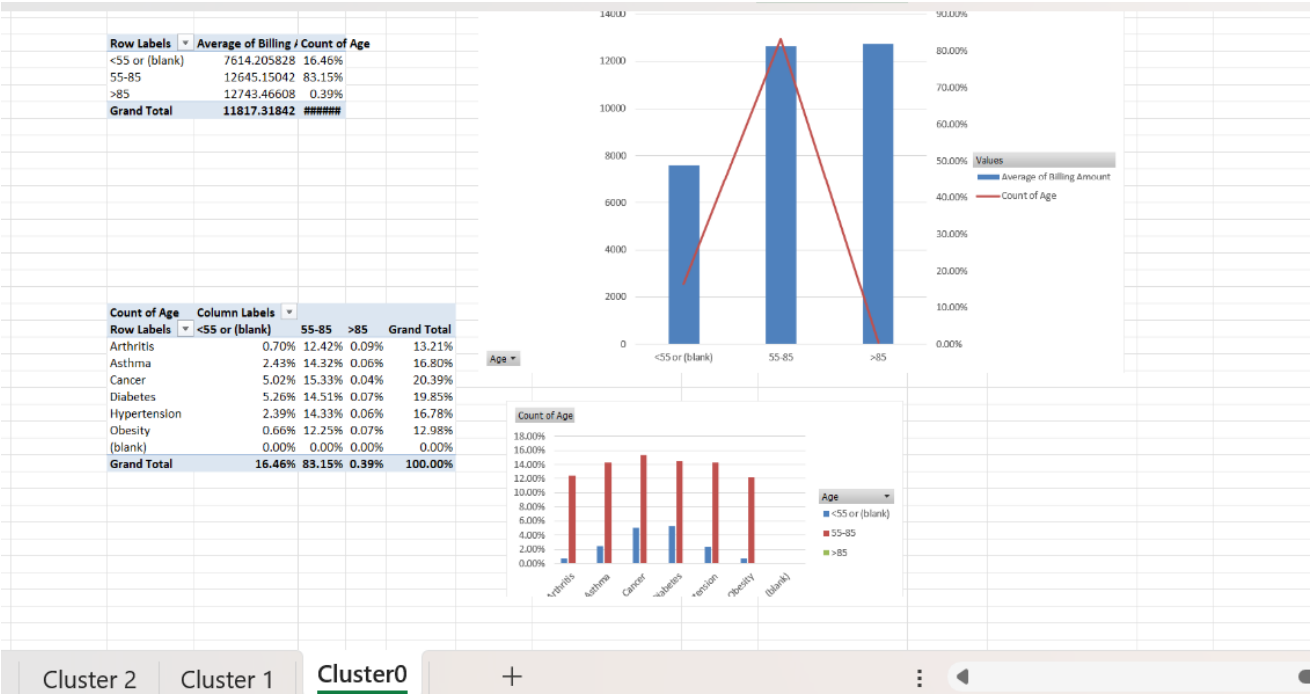
#### *6.1.1. Tổng quan về phân cụm K-Means*

- **Mục tiêu:** Phân chia dữ liệu thành các cụm dựa trên các đặc trưng như Age, Billing Amount, và Medical Condition.
- **Kết quả phân cụm:** Phân thành 4 cụm ( 4 cluster). Dữ liệu được chia thành các nhóm với đặc điểm riêng biệt, mỗi cụm đại diện cho một nhóm bệnh nhân có hành vi và đặc trưng y tế khác nhau.

#### *6.1.1. Đặc điểm từng cụm*

- **Cụm 0:**

- Nhóm tuổi chính: 55-85 tuổi chiếm phần lớn (83.15%).
- Tình trạng y tế nổi bật:
  - Cancer (20.39%)
  - Diabetes (19.85%)
- Chi phí khá cao: 12,743.47 USD.
- Kết luận: Đây là nhóm bệnh nhân cao tuổi với các bệnh mãn tính, có chi phí y tế khá cao.



| Row Labels | Average of Billing A | Count of Age |
|---|---|---|
| <55 or (blank) | 7614.205828 | 16.46% |
| 55-85 | 12645.15042 | 83.15% |
| >85 | 12743.46608 | 0.39% |
| Grand Total | 11817.31842 | ###### |

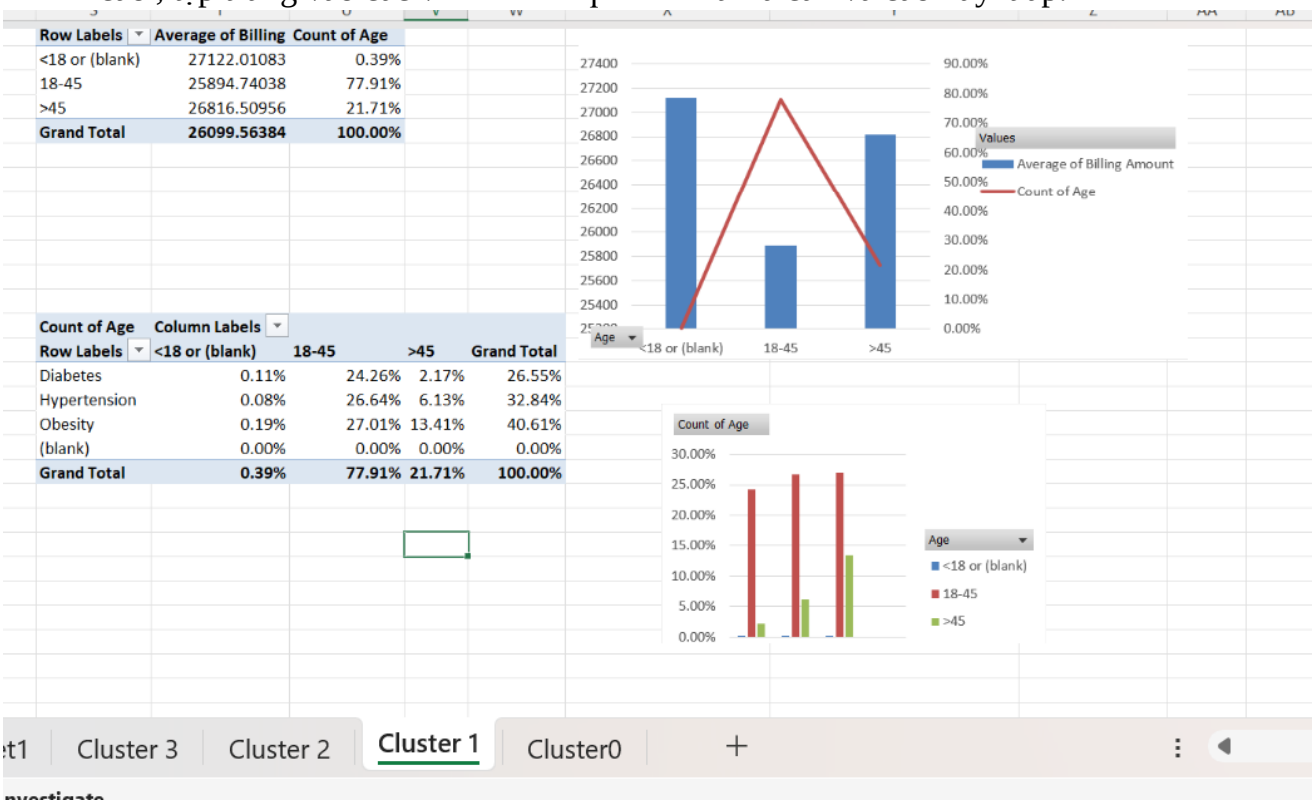
| Count of Age | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | <55 or (blank) | 55-85 | >85 | Grand Total |
| Arthritis | 0.70% | 12.42% | 0.09% | 13.21% |
| Asthma | 2.43% | 14.32% | 0.06% | 16.80% |
| Cancer | 5.02% | 15.33% | 0.04% | 20.39% |
| Diabetes | 5.26% | 14.51% | 0.07% | 19.85% |
| Hypertension | 2.39% | 14.33% | 0.06% | 16.78% |
| Obesity | 0.66% | 12.25% | 0.07% | 12.98% |
| (blank) | 0.00% | 0.00% | 0.00% | 0.00% |
| Grand Total | 16.46% | 83.15% | 0.39% | 100.00% |



*Hình 393. Cluster 0*

- **Cụm 1 :**
- Nhóm tuổi chính: 18-45 tuổi (77.91%).
- Tình trạng y tế nổi bật:

    - Obesity (40.61%)
    - Hypertension (32.84%)
- Chi phí khá cao: 25,894.74 USD.
- Kết luận: Nhóm này bao gồm bệnh nhân trẻ tuổi hơn trung niên, chi phí điều trị khá cao , tập trung vào các vấn đề liên quan đến thừa cân và cao huyết áp.



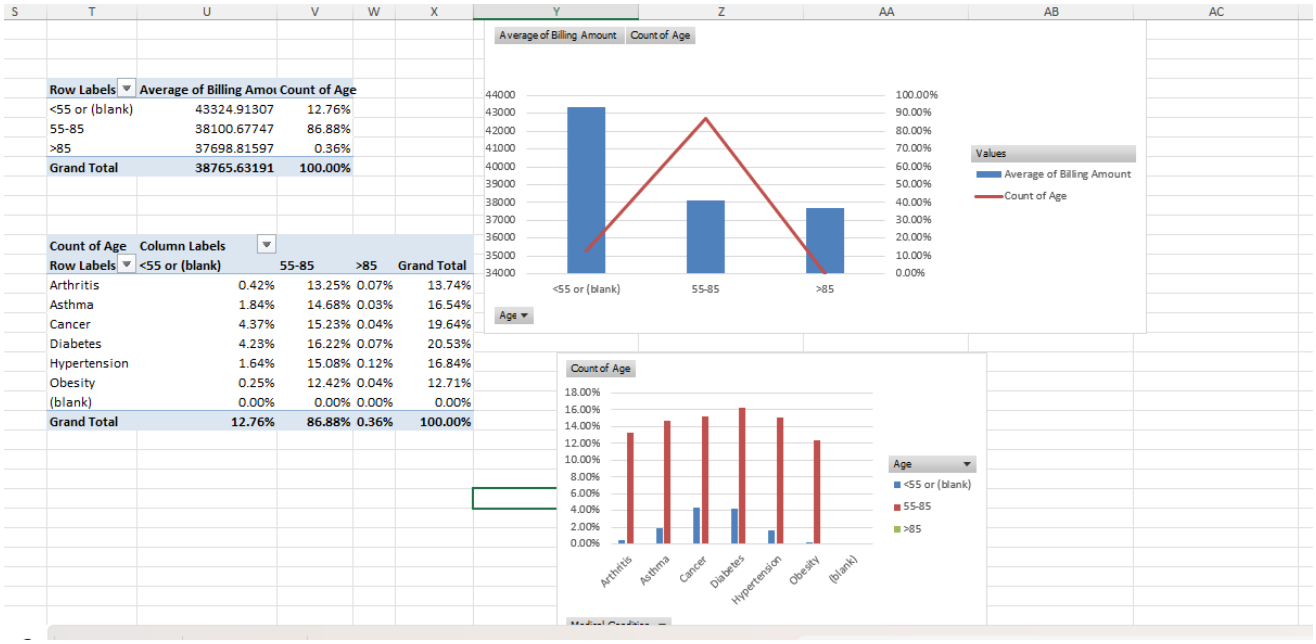
| Row Labels | Average of Billing | Count of Age |
|---|---|---|
| <18 or (blank) | 27122.01083 | 0.39% |
| 18-45 | 25894.74038 | 77.91% |
| >45 | 26816.50956 | 21.71% |
| **Grand Total** | **26099.56384** | **100.00%** |

| Count of Age | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | <18 or (blank) | 18-45 | >45 | Grand Total |
| Diabetes | 0.11% | 24.26% | 2.17% | 26.55% |
| Hypertension | 0.08% | 26.64% | 6.13% | 32.84% |
| Obesity | 0.19% | 27.01% | 13.41% | 40.61% |
| (blank) | 0.00% | 0.00% | 0.00% | 0.00% |
| **Grand Total** | **0.39%** | **77.91%** | **21.71%** | **100.00%** |



*Hình 394. Cluster 1*

- Cụm 2 :
    - Nhóm tuổi chính: 55-85 tuổi chiếm phần lớn ( 86.88%)
    - **Tình trạng y tế nổi bật:**
        - Diabetes(16,22%)
        - Cancer( 15,23%)
    - **Chi phí cao:** 25,798.00 USD.
    - **Kết luận:** Đây là nhóm bệnh nhân cao tuổi ,nhóm này có chi phí cao và tập trung vào bệnh nhân tiểu đường và ung thư.



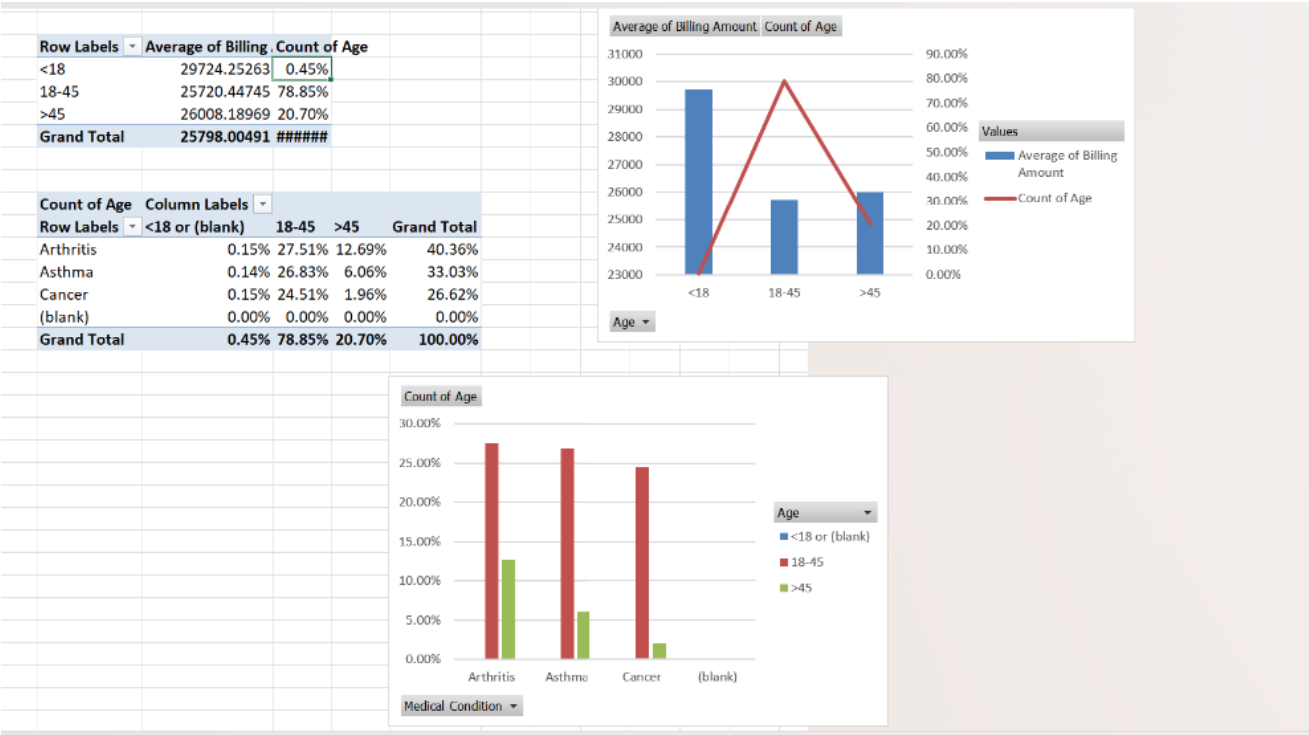
| Row Labels | Average of Billing Amou | Count of Age |
|---|---|---|
| <55 or (blank) | 43324.91307 | 12.76% |
| 55-85 | 38100.67747 | 86.88% |
| >85 | 37698.81597 | 0.36% |
| Grand Total | 38765.63191 | 100.00% |

| Count of Age | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | <55 or (blank) | 55-85 | >85 | Grand Total |
| Arthritis | 0.42% | 13.25% | 0.07% | 13.74% |
| Asthma | 1.84% | 14.68% | 0.03% | 16.54% |
| Cancer | 4.37% | 15.23% | 0.04% | 19.64% |
| Diabetes | 4.23% | 16.22% | 0.07% | 20.53% |
| Hypertension | 1.64% | 15.08% | 0.12% | 16.84% |
| Obesity | 0.25% | 12.42% | 0.04% | 12.71% |
| (blank) | 0.00% | 0.00% | 0.00% | 0.00% |
| Grand Total | 12.76% | 86.88% | 0.36% | 100.00% |



*Hình 395. Cluster 2*

- **Cụm 3 :**
  - **Phân bố độ tuổi và chi phí y tế:**
    - Đa phần bệnh nhân thuộc độ tuổi 18-45 với chi phí y tế trung bình khoảng 25,702.45 USD.
    - Bệnh nhân <18 và >45 chiếm tỷ lệ thấp hơn, nhưng vẫn có mức chi phí cao (26,008.18 USD cho nhóm >45).
  - **Tình trạng y tế theo độ tuổi:**
    - Cancer là bệnh chiếm tỷ lệ lớn nhất, đặc biệt phổ biến ở nhóm tuổi 18-45 và >45.
    - Các bệnh lý khác như Asthma và Arthritis cũng có mặt, nhưng không chiếm ưu thế.



| Row Labels | Average of Billing | Count of Age |
|---|---|---|
| <18 | 29724.25263 | 0.45% |
| 18-45 | 25720.44745 | 78.85% |
| >45 | 26008.18969 | 20.70% |
| Grand Total | 25798.00491 | ###### |

| Count of Age | Column Labels | | | |
|---|---|---|---|---|
| Row Labels | <18 or (blank) | 18-45 | >45 | Grand Total |
| Arthritis | 0.15% | 27.51% | 12.69% | 40.36% |
| Asthma | 0.14% | 26.83% | 6.06% | 33.03% |
| Cancer | 0.15% | 24.51% | 1.96% | 26.62% |
| (blank) | 0.00% | 0.00% | 0.00% | 0.00% |
| Grand Total | 0.45% | 78.85% | 20.70% | 100.00% |



*Hình 396. Cluster 3*

## 6.2. Phân tích kết quả thuật toán Decision Tree và Random Forest

### *6.2.1. Decision Tree*

#### Ma trận nhầm lẫn (Confusion Matrix)

- Ma trận nhầm lẫn của Decision Tree thể hiện:
- True Positive (1 đúng): 1,485 (Nhóm 1 được dự đoán đúng).
- True Negative (0 đúng): 6,870 (Nhóm 0 được dự đoán đúng).
- False Positive (0 sai): 1,455.
- False Negative (1 sai): 6,840.
- **Kết luận:**

Độ chính xác tổng thể ở mức trung bình, do tỷ lệ False Negative và False Positive còn khá cao.

Decision Tree có xu hướng overfitting với tập huấn luyện, dẫn đến khả năng tổng quát không tốt trên tập kiểm tra.

#### Trực quan hóa cây quyết định

- Biểu đồ cây cho thấy:
- Hospital là đặc trưng đầu tiên được sử dụng để phân chia.
- Cây tiếp tục phân nhánh dựa trên Room Number, Doctor, và các thuộc tính khác.
- Cấu trúc cây dễ hiểu, giúp giải thích cách thuật toán đưa ra quyết định.

### *6.2.2. Random Forest*

#### Quan trọng đặc trưng (Feature Importance)

- Từ biểu đồ, các đặc trưng quan trọng trong việc dự đoán bao gồm:
- Hospital: Đóng vai trò quan trọng nhất trong việc phân loại.

- Room Number và Doctor: Cũng ảnh hưởng lớn đến kết quả phân loại.
- Medical Condition và Age: Ảnh hưởng vừa phải.

Điều này cho thấy việc phân loại chi phí y tế phụ thuộc nhiều vào địa điểm điều trị và bác sĩ, hơn là độ tuổi hoặc tình trạng y tế.

Ma trận nhầm lẫn (Confusion Matrix)

- Ma trận nhầm lẫn của Random Forest:
- True Positive (1 đúng): 4,532.
- True Negative (0 đúng): 4,947.
- False Positive (0 sai): 3,378.
- False Negative (1 sai): 3,793.
- **Kết luận:**

Random Forest cải thiện đáng kể khả năng tổng quát so với Decision Tree, với tỷ lệ True Positive và True Negative cao hơn.

Tuy nhiên, tỷ lệ nhầm lẫn (False Positive và False Negative) vẫn còn.

## 7. Đánh giá thuật toán

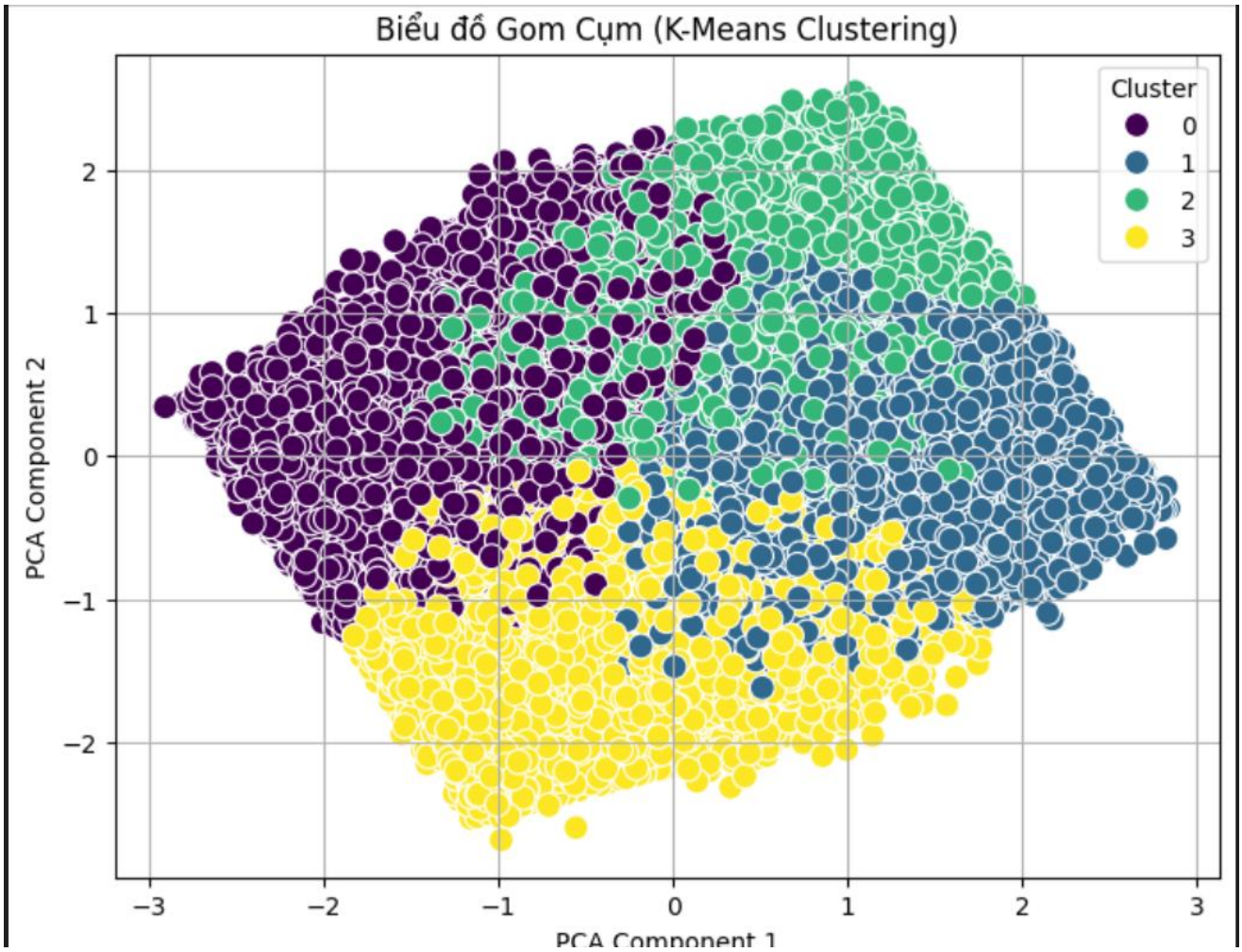
### 7.1. Thuật toán K- Means

DỰA TRÊN TÍNH SILHOUETTE

- **Silhouette Score** cho K-Means Clustering với k=4k = 4k=4 là **0.3044**.
- **Silhouette Score** nằm trong khoảng từ -1 đến 1:

- **0.3** là mức trung bình, cho thấy các cụm được xác định khá tốt nhưng vẫn có sự chồng chéo giữa một số cụm.
- Điều này phản ánh rằng K-Means hoạt động khá tốt, nhưng có thể cải thiện bằng cách:
    - Chọn số cụm tối ưu khác dựa trên phân tích kỹ hơn.
    - Thử các phương pháp phân cụm khác như DBSCAN hoặc Hierarchical Clustering.

DỰA TRÊN BIỂU ĐỒ PHÂN CỤM




- **Phân cụm rõ ràng:**

- Biểu đồ cho thấy các cụm được phân biệt khá rõ ràng (Cluster 0, 1, 2, 3), mặc dù có một số giao thoa nhỏ giữa các cụm, đặc biệt là giữa Cluster 1 và Cluster 2.
- Điều này chứng tỏ thuật toán K-Means đã hoạt động tốt trong việc phân chia dữ liệu thành các nhóm có điểm tương đồng cao.

- **Đặc điểm cụm:**

- Cluster 0 (tím): Có vẻ bao phủ một phần lớn dữ liệu, thể hiện nhóm có đặc điểm phổ biến.
- Cluster 3 (vàng): Một cụm nhỏ hơn, tập trung dữ liệu, cho thấy các điểm dữ liệu có nhiều điểm tương đồng.
- Cluster 1 (xanh biển) và Cluster 2 (xanh lục): Phân chia tốt, nhưng giao thoa giữa hai cụm có thể phản ánh sự chồng chéo đặc điểm.

### 7.2. Thuật toán Decision Tree

- Độ chính xác (Accuracy): 50.18%

- Đây là một mức độ chính xác thấp, gần như tương đương với việc đoán ngẫu nhiên trong bài toán phân loại nhị phân.
- Báo cáo phân loại (Classification Report):
- Precision:
  - Nhãn 0: 50%
  - Nhãn 1: 51%
- Precision thấp cho thấy nhiều dự đoán nhầm lẫn giữa hai lớp.
- Recall:
- Nhãn 0: 83% (tốt hơn khi dự đoán đúng lớp 0).
- Nhãn 1: 18% (rất thấp, nghĩa là Decision Tree không phân loại tốt nhãn 1).
- F1-Score:
- Nhãn 0: 62% (tốt hơn so với nhãn 1).
- Nhãn 1: 26% (thể hiện hiệu suất kém khi xử lý nhãn 1).

**Kết luận Decision Tree:**

Decision Tree bị mất cân bằng trong việc phân loại hai lớp, tập trung chủ yếu vào lớp 0.

Hiện tượng overfitting có thể xảy ra do sự phụ thuộc quá mức vào dữ liệu huấn luyện.

### 7.3. Thuật toán Random Forest

- **Độ chính xác (Accuracy):** 56.93%

Random Forest cải thiện độ chính xác so với Decision Tree, nhưng vẫn chỉ ở mức trung bình.

- **Báo** cáo phân loại (Classification Report):
  - Precision:0.57
    - Nhãn 0: 57%
    - Nhãn 1: 57%
- Precision của cả hai nhãn đã cân bằng hơn, nhưng vẫn còn thấp.
- **Recall:**
- Nhãn 0: 59%
- Nhãn 1: 54%
- Random Forest cải thiện khả năng phân loại cả hai nhãn so với Decision Tree.
- **F1-Score:**
- Nhãn 0: 58%
- Nhãn 1: 56%
- Điểm F1 tổng thể cải thiện, nhưng vẫn chưa đạt mức tối ưu.

**Kết luận Random Forest:**

Random Forest giảm hiện tượng overfitting nhờ sử dụng nhiều cây quyết định.

Hiệu suất tổng thể tốt hơn Decision Tree, nhưng vẫn có thể cải thiện.

# DANH MỤC TÀI LIỆU THAM KHẢO


**Tài liệu tiếng Việt**

[1] Nguyễn Văn Hiệp, *Cơ sở dữ liệu và khai phá dữ liệu*, Nhà xuất bản Đại học Quốc gia TP.HCM, 2020

[2] Nguyễn Đình Phúc, *OLAP và các phương pháp tối ưu hóa truy vấn dữ liệu*, Nhà xuất bản Đại học Quốc gia TP.HCM, 2019.

[3] Trần Quang Hải, "Phân tích dữ liệu y tế sử dụng K-Means và Random Forest," *Tạp chí Công nghệ Thông tin và Ứng dụng*, Đại học Quốc gia TP.HCM, số 5, 2022.

**Tài liệu tiếng Anh**

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2011.

[5] UCI Machine Learning Repository, "Healthcare Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/healthcare. [Accessed: 24-Dec-2024].

[6] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.

[7] Scikit-learn Documentation, "Clustering — K-Means," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means. [Accessed: 24-Dec-2024].

[8] M. Berry and G. Linoff, *Mastering Data Mining: The Art and Science of Customer Relationship Management*, Wiley, 2000.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[10] "K-Means Clustering Algorithm in Healthcare," Towards Data Science, [Online]. Available: https://towardsdatascience.com. [Accessed: 24-Dec-2024].