

ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THÔNG THÔNG TIN



BÁO CÁO ĐỒ ÁN MÔN HỌC
KHO DỮ LIỆU VÀ OLAP
ĐỀ TÀI
PHÂN TÍCH KHO DỮ LIỆU
DỰ ĐOÁN TÌNH TRẠNG CHUYẾN BAY

Giảng viên hướng dẫn: ThS. Nguyễn Thị Kim Phụng

GV. Lê Võ Đình Kha

Lớp: IS217.N21

Nhóm sinh viên thực hiện – Nhóm 21

Trần Thị Ngọc Ánh

MSSV: 20521083

Hà Danh Tuấn

MSSV: 20522109

TP. HỒ CHÍ MINH, 6/2023

LỜI CẢM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, nhóm chúng em xin gửi lời cảm ơn đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – ĐHQG TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này. Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới Cô Nguyễn Thị Kim Phụng – giảng viên lý thuyết môn Kho dữ liệu và OLAP và anh Lê Võ Đình Kha – trợ giảng của môn đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn nhóm trong suốt quá trình làm đồ án môn học. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đồ án. Nếu không có những lời hướng dẫn, dạy bảo của Cô và Anh thì nhóm chúng em nghĩ đồ án này của nhóm rất khó có thể hoàn thiện được.

Ngoài ra, để đồ án được hoàn thành thì không nào không cảm ơn những người đã làm ra nó, cảm ơn các thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ. Xuất phát từ mục đích học tập tìm hiểu về kho dữ liệu, phương pháp xây dựng và phân tích dữ liệu trên kho dữ liệu dự đoán tình trạng chuyến bay. Dựa trên những kiến thức được Cô và Anh cung cấp trên trường kết hợp với việc tự tìm hiểu những công cụ và kiến thức mới, nhóm đã cố gắng thực hiện đồ án một cách tốt nhất. Từ đó, nhóm chúng em vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm chúng em không tránh khỏi những thiếu sót. Chính vì vậy, nhóm chúng em rất mong nhận được những sự góp ý từ phía Cô và Anh nhằm hoàn thiện những kiến thức mà nhóm em đã học tập và là hành trang để nhóm chúng em thực hiện tiếp các đề tài khác trong tương lai.

Sau cùng, chúng em xin kính chúc Cô và Anh thật dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp là truyền đạt kiến thức cho các bạn sinh viên.

TP. Hồ Chí Minh, tháng 06, năm 2023

Nhóm thực hiện

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

....., ngày tháng năm 2023

Người nhận xét

(Ký tên và ghi rõ họ tên)

BẢNG PHÂN CÔNG CÔNG VIỆC

Họ và tên	MSSV	Phân công	Đánh giá
Trần Thị Ngọc Ánh	20521083	<p>Tuần 2: Lựa chọn đề tài và dataset phù hợp.</p> <p>Tuần 3: Soạn form file báo cáo.</p> <p>Tuần 4: Lý do chọn đề tài, mô tả thuộc tính dataset, sơ đồ dữ liệu xây dựng.</p> <p>Tuần 5: Thông nhất lại các thuộc tính được phân tích, các bảng dim, fact.</p> <p>Tuần 6 + 7: Quá trình xây dựng kho dữ liệu SSIS.</p> <p>Tuần 8 + 9: Xây dựng SSAS</p> <p>Tuần 10: Tìm hiểu quá trình Mining</p> <p>Tuần 11 + 12: Thực hiện quá trình Mining</p>	Hoàn thành tốt.

Hà Danh Tuấn	20522109	<p>Tuần 2: Lựa chọn đề tài và dataset phù hợp.</p> <p>Tuần 3: Tiến hành tiền xử lý dữ liệu.</p> <p>Tuần 4: Xác định thuộc tính được phân tích, xây dựng kho dữ liệu, Dictionary.</p> <p>Tuần 5: Thông nhất lại các thuộc tính được phân tích, các bảng dim, fact.</p> <p>Tuần 6 + 7: Tìm hiểu 10 câu truy vấn</p> <p>Tuần 8 + 9 + 10: Thực hiện MDX cho 10 câu truy vấn</p> <p>Tuần 11 + 12: Thực hiện Excel và PowerBI cho 10 câu truy vấn</p>	Hoàn thành tốt
-------------------------	-----------------	---	----------------

Bảng 1. Bảng phân công, đánh giá thành viên

MỤC LỤC

LỜI CẢM ƠN.....	2
NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN	3
BẢNG PHÂN CÔNG CÔNG VIỆC	4
MỤC LỤC	6
DANH MỤC HÌNH ẢNH.....	10
DANH MỤC BẢNG	17
DANH MỤC TỪ VIẾT TẮT	18
1. CHƯƠNG 1 – GIỚI THIỆU KHO DỮ LIỆU	19
1.1. Lý do chọn đề tài	19
1.2. Mô tả dataset.....	20
1.2.1. Danh sách thuộc tính được phân tích	21
1.3. Dictionary	23
1.4. Sơ đồ dữ liệu xây dựng.....	26
1.4.1. Lược đồ hình sao (Star schema)	26
1.4.2. Lược đồ hình bông tuyết (Snowflake schema)	27
1.5. Xây dựng kho dữ liệu	28
1.5.1. Sơ đồ bông tuyết minh họa	28
1.5.2. Mô tả chi tiết các bảng dữ liệu.....	28
2. CHƯƠNG 2 – QUÁ TRÌNH XÂY DỰNG KHO DỮ LIỆU (SSIS).....	33
2.1. Chuẩn bị công cụ và Data warehouse.....	33
2.2. Tạo new database trong Microsoft SQL Server Management Studio.....	33
2.3. Quá trình SSIS	35
2.3.1. Khởi tạo project	35

2.3.2.	Khởi tạo Sequence Container	36
2.3.3.	Import Data	37
2.3.4.	Tạo bảng Dim_Time	41
2.3.5.	Tạo bảng Dim_Plane	46
2.3.6.	Tạo bảng Dim_State	50
2.3.7.	Tạo bảng Dim_City	57
2.3.8.	Tạo bảng Dim_Airline	67
2.3.9.	Tạo bảng Dim_Cancelled	70
2.3.10.	Tạo bảng Dim_Diverted	74
2.3.11.	Tạo bảng Fact.....	77
2.3.12.	Import dữ liệu vào database.....	82
3.	CHƯƠNG 3 – PHÂN TÍCH DỮ LIỆU VÀ BÁO CÁO	84
3.1.	Tạo Project SSAS trong Visual Studio 2022.....	84
3.2.	Connect đến Data Source	85
3.3.	Tạo Data Source View.....	88
3.4.	Tạo Cube và dimensions	92
3.5.	Thao tác trên các bảng Dim.....	95
3.5.1.	Bảng Dim_Time	95
3.5.2.	Bảng Dim_Plane	96
3.5.3.	Bảng Dim_City	97
3.5.4.	Bảng Dim_Airline.....	98
3.5.5.	Bảng Dim_Cancelled.....	99
3.5.6.	Bảng Dim_Diverted.....	100
3.6.	Deploy và process project.	101

3.7. Thực hiện các câu truy vấn dữ liệu.....	103
3.7.1. Thống kê top 5 thành phố ở bang California có số lượng chuyến bay đến nhiều nhất.	103
3.7.2. Thống kê tổng các chuyến bay có khoảng cách lớn hơn 800 dặm vào ngày 27-04-2022.	106
3.7.3. Thống kê số lượng các chuyến bay bị hủy của hãng hàng không Southwest Airlines Co. từ ngày 24-04-2022 đến 27-04-2022 theo từng ngày.....	109
3.7.4. Thống kê các thành phố ở bang Florida có ít nhất trên 100 chuyến bay đến trong ngày 27-04-2022.....	112
3.7.5. Thống kê tổng số chuyến bay khởi hành của hãng hàng không Envoy Air theo từng thành phố.	115
3.7.6. Thống kê tổng số chuyến bay bị hủy theo từng bang của hãng hàng không Horizon Air	118
3.7.7. Thống kê số lượng chuyến bay khởi hành của hãng hàng không Frontier Airlines Inc. bị điều hướng theo từng máy bay.	122
3.7.8. Thống kê tổng thời gian cát cánh trễ và tổng thời gian khởi hành trễ của các chuyến bay đến thành phố Orlando tính theo các ngày của hãng hàng không Alaska Airline Inc. 125	
3.7.9. Thống kê khoảng cách và mã chuyến bay theo từng chiếc máy bay....	128
3.7.10. Thống kê tổng khoảng cách và tổng thời gian bay theo từng máy bay từ ngày 24-04-2022 đến ngày 26-04-2022.....	130
4. CHƯƠNG 4 – DATA MINING.....	134
4.1. Sơ đồ mining.....	134
4.2. Quá trình thực hiện	134
4.2.1. Xử lý dữ liệu	134
4.2.2. Decision Tree	136

4.2.3. Random Forest.....	137
4.3. Giải thích kết quả.....	138
4.3.1. Thuật toán Decision Tree.....	138
4.3.2. Thuật toán Random Forest.....	142
4.4. Nhận xét.....	146
DANH MỤC TÀI LIỆU THAM KHẢO	147

DANH MỤC HÌNH ẢNH

Hình 1.1. Thông tin dataset	21
Hình 1.2. Sơ đồ bông tuyết minh họa.....	28
Hình 2.1. Tạo New Database.....	33
Hình 2.2. Điền thông tin New Database.....	34
Hình 2.3. Kết quả sau khi tạo Database.....	34
Hình 2.4. Tạo mới project SSIS	35
Hình 2.5. Điền thông tin project.....	36
Hình 2.6. Khởi tạo Sequence Container.....	37
Hình 2.7. Tạo một Data Flow Task tên Import Data Source.....	37
Hình 2.8. Tạo Excel Source.....	38
Hình 2.9. Khởi tạo Database	38
Hình 2.10. Tạo OLE DB Destination	39
Hình 2.11. Tạo OLE DB Connection Manager.....	39
Hình 2.12. Điền thông tin Connection Manager	40
Hình 2.13. Kết quả sau khi Create Table Flight_Src.....	40
Hình 2.14. Kết quả sau khi import data.....	41
Hình 2.15. Tạo Data Flow Task Dim_Time	41
Hình 2.16. Tạo OLE DB Source chứa Dataset.....	42
Hình 2.17. Chọn Connection và bảng	42
Hình 2.18. Chọn thuộc tính Time	43
Hình 2.19. Thêm Conditional Split	43
Hình 2.20. Sort dữ liệu và Remove rows	44
Hình 2.21. Tạo 1 OLE Destination tên Dim_Time	44
Hình 2.22. Tạo bảng Dim_Time.....	45
Hình 2.23. Qua Mapping để kiểm tra	45
Hình 2.24. Luồng thực hiện của bảng Dim_Time.....	46
Hình 2.25. Tạo Data Flow Task Dim_Plane	46
Hình 2.26. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	47

Hình 2.27. Thêm Conditional Split	47
Hình 2.28. Sort dữ liệu và Remove rows	48
Hình 2.29. Tạo 1 OLE Destination tên Dim_Plane.....	48
Hình 2.30. Tạo bảng Dim_Plane	49
Hình 2.31. Qua Mapping để kiểm tra	49
Hình 2.32. Luồng thực hiện của bảng Dim_Plane	50
Hình 2.33. Tạo Data Flow Task Dim_State	50
Hình 2.34. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	51
Hình 2.35. Dùng công cụ Multicast để chia tập dữ liệu	51
Hình 2.36. Thêm Conditional Split cho đối tượng Origin	52
Hình 2.37. Thêm Conditional Split cho đối tượng Dest.....	52
Hình 2.38. Sort dữ liệu cho đối tượng Origin và Remove rows.....	53
Hình 2.39. Sort dữ liệu cho đối tượng Dest và Remove rows.....	53
Hình 2.40. Dùng công cụ Union All để kết hợp 2 tập đối tượng Origin và Dest.....	54
Hình 2.41. Kết 2 đối tượng và đổi tên thuộc tính	54
Hình 2.42. Sort dữ liệu và Remove rows	55
Hình 2.43. Tạo 1 OLE Destination tên Dim_State.....	55
Hình 2.44. Tạo bảng Dim_State	56
Hình 2.45. Qua Mapping để kiểm tra	56
Hình 2.46. Luồng thực hiện của bảng Dim_State	57
Hình 2.47. Tạo Data Flow Task Dim_City	57
Hình 2.48. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	58
Hình 2.49. Dùng công cụ Multicast để chia dữ liệu thành nhiều đối tượng	58
Hình 2.50. Thêm Conditional Split cho đối tượng Origin	59
Hình 2.51. Thêm Conditional Split cho đối tượng Dest.....	59
Hình 2.52. Sort dữ liệu cho đối tượng Origin và Remove rows	60
Hình 2.53. Sort dữ liệu cho Dest và Remove rows	60
Hình 2.54. Dùng công cụ Union All để kết hợp 2 tập đối tượng Origin và Dest.....	61
Hình 2.55. Kết 2 đối tượng và đổi tên thuộc tính	61

Hình 2.56. Sort dữ liệu và Remove rows	62
Hình 2.57. Tạo Lookup để kết dữ liệu trên với bảng Dim_State	62
Hình 2.58. Cài đặt Lookup	63
Hình 2.59. Chọn bảng Dim_State.....	63
Hình 2.60. Kết nối Source và bảng Dim_State	64
Hình 2.61. Sort dữ liệu và Remove rows	64
Hình 2.62. Tạo 1 OLE Destination tên Dim_City.....	65
Hình 2.63. Tạo bảng Dim_City	65
Hình 2.64. Qua Mapping để kiểm tra	66
Hình 2.65. Luồng thực hiện của bảng Dim_City	66
Hình 2.66. Tạo Data Flow Task Dim_Airline.....	67
Hình 2.67. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	67
Hình 2.68. Thêm Conditional Split	68
Hình 2.69. Sort dữ liệu và Remove rows	68
Hình 2.70. Tạo bảng Dim_Airline.....	69
Hình 2.71. Qua Mapping để kiểm tra	69
Hình 2.72. Luồng thực hiện của bảng Dim_Airline	70
Hình 2.73. Tạo Data Flow Task Dim_Cancelled	70
Hình 2.74. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	71
Hình 2.75. Thêm Conditional Split	71
Hình 2.76. Sort dữ liệu và Remove rows	72
Hình 2.77. Tạo bảng Dim_Cancelled	72
Hình 2.78. Qua Mapping để kiểm tra	73
Hình 2.79. Luồng thực hiện của bảng Dim_Cancelled	73
Hình 2.80. Tạo Data Flow Task Dim_Diverted	74
Hình 2.81. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng	74
Hình 2.82. Thêm Conditional Split	75
Hình 2.83. Sort dữ liệu và Remove rows	75
Hình 2.84. Tạo bảng Dim_Diverted	76

Hình 2.85. Qua Mapping để kiểm tra	76
Hình 2.86. Luồng thực hiện của bảng Dim_Diverted	77
Hình 2.87. Tạo Data Flow Task Fact	77
Hình 2.88. Khởi tạo OLE DB Source.....	78
Hình 2.89. Tạo Lookup kết Source với Dim_Cancelled.....	78
Hình 2.90. Kết nối Source và bảng Dim_Cancelled	79
Hình 2.91. Dùng Lookup để kết các bảng Dim lại	79
Hình 2.92. Sort dữ liệu và Remove rows	80
Hình 2.93. Tạo bảng Fact	80
Hình 2.94. Qua Mapping để kiểm tra	81
Hình 2.95. Luồng thực hiện của bảng Fact	81
Hình 2.96. Thông tin bảng Execute SQL Task	82
Hình 2.97. Tạo câu truy vấn để xóa dữ liệu trước đó.....	82
Hình 2.98. Luồng thực hiện của quá trình SSIS.....	83
Hình 2.99. Diagram được tạo ra trong SQL Server.....	83
Hình 3.1. Tạo mới project SSAS	84
Hình 3.2. Điện thông tin project	85
Hình 3.3. New Data Source	85
Hình 3.4. Create a data source based on existing on new connection.....	86
Hình 3.5. Điện thông tin Connection Manager	86
Hình 3.6. Test Connection.....	87
Hình 3.7. Use a specific Windows user name and password	87
Hình 3.8. Chọn Data Source và Finish	88
Hình 3.9. New Data Source View	88
Hình 3.10. Kết nối với Database	89
Hình 3.11. Chọn bảng Fact.....	89
Hình 3.12. Chọn Add Related Tables.....	90
Hình 3.13. Bấm lại Add Related Tables.....	90
Hình 3.14. Kiểm tra thông tin.....	91

Hình 3.15. Kết quả Data Source View vừa tạo	91
Hình 3.16. New Cube	92
Hình 3.17. Use an existing table.....	92
Hình 3.18. Chọn Measure table.....	93
Hình 3.19. Measure table.....	93
Hình 3.20. Chọn Dimension	94
Hình 3.21. Kiểm tra thông tin và nhấn finish.....	94
Hình 3.22. Hiển thị sơ đồ	95
Hình 3.23. Chọn Dim Time.dim.....	95
Hình 3.24. Kéo các thuộc tính Dim_Time	96
Hình 3.25. Chọn Dim Plane.dim	96
Hình 3.26. Kéo các thuộc tính Dim_Plane	97
Hình 3.27. Chọn Dim City.dim	97
Hình 3.28. Kéo các thuộc tính Dim_City và Dim_State	98
Hình 3.29. Chọn Dim Airline.dim.....	98
Hình 3.30. Kéo các thuộc tính Dim_Airline	99
Hình 3.31. Chọn Dim_Cancelled	99
Hình 3.32. Kéo các thuộc tính Dim_Cancelled.....	100
Hình 3.33. Chọn Dim Diverted.dim	100
Hình 3.34. Kéo các thuộc tính Dim_Diverted.....	101
Hình 3.35. Thực hiện Process.....	101
Hình 3.36. Chọn Yes	102
Hình 3.37. Chọn Run.....	102
Hình 3.38. Thông báo kết quả	103
Hình 3.39. SSAS - TOP_5_CITY	104
Hình 3.40. SSAS – Kết quả	104
Hình 3.41. MDX - Kết quả	105
Hình 3.42. Kết quả thực hiện Excel	106
Hình 3.43. Kết quả thực hiện Power BI	106

Hình 3.44. SSAS – Kết quả	107
Hình 3.45. MDX - Kết quả	108
Hình 3.46. Kết quả thực hiện Excel	108
Hình 3.47. Kết quả thực hiện Power BI	109
Hình 3.48. SSAS - Câu truy vấn.....	110
Hình 3.49. MDX - Kết quả	111
Hình 3.50. Kết quả thực hiện Excel	111
Hình 3.51. Kết quả thực hiện Power BI	112
Hình 3.52. SSAS - Thực hiện truy vấn.....	113
Hình 3.53. MDX - Kết quả	114
Hình 3.54. Kết quả thực hiện Excel	114
Hình 3.55. Kết quả thực hiện Power BI -1	115
Hình 3.56. Kết quả thực hiện Power BI -2	115
Hình 3.57. SSAS - Thực hiện truy vấn.....	116
Hình 3.58. MDX - Kết quả	117
Hình 3.59. Kết quả thực hiện Excel	117
Hình 3.60. Kết quả thực hiện Power BI -1	118
Hình 3.61. Kết quả thực hiện Power BI -2	118
Hình 3.62. SSAS - Thực hiện truy vấn.....	119
Hình 3.63. MDX - Kết quả	120
Hình 3.64. Kết quả thực hiện Excel	121
Hình 3.65. Kết quả thực hiện Power BI -1	121
Hình 3.66. Kết quả thực hiện Power BI -2	122
Hình 3.67. SSAS - Thực hiện truy vấn.....	122
Hình 3.68. MDX - Kết quả	123
Hình 3.69. Kết quả thực hiện Excel	124
Hình 3.70. Kết quả thực hiện Power BI -1	124
Hình 3.71. Kết quả thực hiện Power BI -2	125
Hình 3.72. SSAS - Thực hiện truy vấn.....	125

Hình 3.73. MDX - Kết quả	126
Hình 3.74. Kết quả thực hiện Excel	127
Hình 3.75. Kết quả thực hiện Power BI - 1	127
Hình 3.76. Kết quả thực hiện Power BI - 2	128
Hình 3.77. SSAS - Thực hiện truy vấn.....	128
Hình 3.78. MDX - Kết quả	129
Hình 3.79. Kết quả thực hiện Excel	130
Hình 3.80. Kết quả thực hiện Power BI	130
Hình 3.81. SSAS - Thực hiện truy vấn.....	131
Hình 3.82. MDX - Kết quả	132
Hình 3.83. Kết quả thực hiện Excel	132
Hình 3.84. Kết quả thực hiện Power BI	133
Hình 4.1. Lựa chọn những thuộc tính sử dụng	135
Hình 4.2. Phân loại giá trị CRSDepTime thành các khoảng	135
Hình 4.3. Tạo cột DelayGroup để phân loại tình trạng chuyến bay	135
Hình 4.4. Lọc giá trị OriginStateFips và DestStateFips	136
Hình 4.5. Xác định thuộc tính features và labels.....	136
Hình 4.6. Huấn luyện mô hình Decision Tree.....	137
Hình 4.7. Xác định X và y	137
Hình 4.8. Xây dựng mô hình Random Forest	138
Hình 4.9. Ma trận nhầm lẫn Decision Tree	139
Hình 4.10. Cây quyết định Decision Tree	140
Hình 4.11. Tập luật - Decision Tree	141
Hình 4.12. Dữ liệu đầu vào	142
Hình 4.13. Kết quả dự đoán - Decision Tree.....	142
Hình 4.14. Cây quyết định Random Forest	144
Hình 4.15. Tập luật từ cây quyết định - Random Forest	145
Hình 4.16. Dữ liệu đầu vào	146
Hình 4.17. Kết quả dự đoán - Random Forest.....	146

DANH MỤC BẢNG

Bảng 1. Bảng phân công, đánh giá thành viên	5
Bảng 2. Danh mục từ viết tắt.....	18
Bảng 1.1. Thuộc tính được phân tích	23
Bảng 1.2. Dictionary thuộc tính Cancelled.	23
Bảng 1.3. Dictionary thuộc tính Diverted.	23
Bảng 1.4. Dictionary thuộc tính AilineName.	24
Bảng 1.5. Dictionary thuộc tính Deptime.....	24
Bảng 1.6. Dictionary thuộc tính OriginCityName.	25
Bảng 1.7. Dictionary thuộc tính OriginStateName.	25
Bảng 1.8. Dictionary thuộc tính ArrDelay.	26
Bảng 1.9. Mô tả chi tiết bảng Dim_City.	28
Bảng 1.10. Mô tả chi tiết bảng Dim_State.	29
Bảng 1.11. Mô tả chi tiết bảng Dim_Time.....	29
Bảng 1.12. Mô tả chi tiết bảng Dim_Cancelled	29
Bảng 1.13. Mô tả chi tiết bảng Dim_Diverted	30
Bảng 1.14. Mô tả chi tiết bảng Dim_Plane	30
Bảng 1.15. Mô tả chi tiết bảng Dim_Airline	30
Bảng 1.16. Mô tả chi tiết bảng Fact.....	32

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Từ đầy đủ	Ý nghĩa
SSIS	SQL Server Integration Services	Công cụ để tích hợp dữ liệu từ nhiều nguồn khác nhau vào cơ sở dữ liệu, cho phép thực hiện các hoạt động tiền xử lý và định dạng dữ liệu trước khi lưu trữ.
SSAS	SQL Server Analysis Services	Công cụ để phân tích dữ liệu và tạo các báo cáo, đồ thị và biểu đồ để giúp người dùng hiểu rõ hơn về dữ liệu. SSAS cho phép tạo các Cube (Khối) dữ liệu để phân tích dữ liệu từ nhiều góc độ khác nhau và dễ dàng hiển thị các mô hình dữ liệu trực quan.
MDX	Multidimensional Expressions	Ngôn ngữ truy vấn đa chiều được sử dụng trong các hệ thống cơ sở dữ liệu đa chiều.

Bảng 2. Danh mục từ viết tắt

CHƯƠNG 1 – GIỚI THIỆU KHO DỮ LIỆU

Chương 1 giới thiệu cho ta cái nhìn tổng quan về kho dữ liệu, giúp cho chúng ta có thể hiểu được quá trình xây dựng và phân tích dữ liệu cho đề tài PHÂN TÍCH KHO DỮ LIỆU DỰ ĐOÁN TÌNH TRẠNG CHUYẾN BAY.

1.1. Lý do chọn đề tài

Hiện nay, tình trạng chậm trễ chuyến bay là vấn đề xảy ra khá phổ biến. Tình trạng chậm trễ, hủy bỏ, hoặc đến sớm của các chuyến bay có thể gây ra rất nhiều khó khăn và phiền toái cho hành khách và các đơn vị liên quan trong ngành hàng không, đặc biệt là những người đang đi công việc hoặc đi chuyến du lịch. Do đó, việc xây dựng một hệ thống dự đoán tình trạng chuyến bay có thể giúp quản lý chuyến bay hiệu quả hơn và giảm thiểu sự bất tiện cho hành khách.

Tình trạng chuyến bay là một vấn đề rất quan trọng trong ngành hàng không. Xây dựng kho dữ liệu dự đoán tình trạng chuyến bay là một đề tài có tính ứng dụng cao trong thực tế, đặc biệt là trong lĩnh vực hàng không. Việc dự đoán tình trạng chuyến bay sẽ giúp các hãng hàng không có thể phát hiện và xử lý các vấn đề kỹ thuật trước khi chuyến bay diễn ra, đồng thời giúp đảm bảo an toàn cho hành khách.

Khi có thông tin dự đoán tình trạng chuyến bay, các bộ phận quản lý và điều hành hàng không có thể đưa ra quyết định nhanh chóng để giảm thiểu tối đa tác động tiêu cực đến hành khách và chuyến bay, giúp tiết kiệm thời gian và tăng hiệu quả trong hoạt động hàng không. Do đó, việc phân tích, đánh giá và dự đoán tình trạng chuyến bay trước khi khởi hành sẽ giúp cho các hãng hàng không có kế hoạch tốt hơn trong việc điều phối tài nguyên và giảm thiểu tình trạng chậm trễ của chuyến bay.

Ngoài ra, việc xây dựng mô hình dự đoán tình trạng chuyến bay có thể giúp các công ty hàng không tối ưu hóa quá trình hoạt động của mình, như quản lý lịch trình, tài nguyên, nhân sự và tài chính.

Vì vậy, nhóm chúng em quyết định chọn nghiên cứu bộ dữ liệu về tình trạng chuyến bay nhằm mục đích có thể đưa ra những thông kê và phần nào dự báo nhằm giúp cho hành

khách cũng như hãng hàng không có thể giảm thiểu được những sự bất tiện do sự chậm trễ chuyến bay cũng như nâng cao hiệu quả và sự hài lòng của khách hàng cho hãng hàng không.

1.2. Mô tả dataset

Tên dataset: Flight Status Prediction [1]

Link dataset: https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?sizeStart=70%2CMB&page=13&select=Combined_Flights_2022.csv

Mô tả dataset:

- Tập dữ liệu "Flight Delay Dataset 2018-2022" trên Kaggle là một bộ dữ liệu về các chuyến bay trong năm 2018-2022 tại Hoa Kỳ.
- Tập dữ liệu này bao gồm thông tin về hàng triệu chuyến bay, bao gồm cả chuyến bay nội địa và quốc tế, với các trường dữ liệu như điểm đến, điểm khởi hành, thời gian khởi hành, thời gian dự kiến, thời gian thực tế, thời gian trễ chuyến bay, và một số thông tin khác.
- Tập dữ liệu này được chia thành các file CSV cho từng năm, mỗi file CSV chứa thông tin về các chuyến bay trong năm tương ứng.
- Việc nghiên cứu về dataset này có thể giúp phân tích và đưa ra dự đoán về việc chuyến bay sẽ bị trễ, bị hủy, hay diễn ra đúng giờ. Điều này có thể giúp các hãng hàng không cải thiện quá trình hoạt động của mình, cũng như tăng trải nghiệm của hành khách.

Đơn vị cung cấp: TranStats data library

Kích thước: 30 thuộc tính và 135640 dòng dữ liệu được thu thập vào tháng 4/2022.

The screenshot shows a Kaggle dataset page for 'Flight Status Prediction'. The left sidebar includes links for Create, Home, Competitions, Datasets (selected), Models, Code, Discussions, Learn, and More. The main content area features the dataset title 'Flight Status Prediction' and a brief description: 'Can you predict which flights will be delayed or cancelled in 5 years of data?'. Below this is a 'Data Card' section with tabs for Data Card, Code (5), and Discussion (6). To the right, there's a 'About Dataset' section with a bulleted list of questions: 'Can you predict which flights will be cancelled or delayed?', 'Can you predict the delay time?', and 'Can you explore how different airlines compare?'. It also notes that the dataset makes all these possible and is perfect for school projects, research, or resumes. On the far right, there are sections for 'Usability' (10.00), 'License' (CC0: Public Domain), and 'Expected update frequency' (Annually). A thumbnail image of a person looking at an airport departure board is displayed.

Hình 1.1. Thông tin dataset

1.2.1. Danh sách thuộc tính được phân tích

STT	Thuộc tính	Tên đầy đủ	Ý nghĩa
1	FlightDate	Flight date	Ngày bay
2	AirlineName	Airline name	Tên hãng hàng không
3	Origin	Origin city	Mã thành phố xuất phát
4	Dest	Destination city	Mã thành phố đích
5	Cancelled	Cancelled	Có bị hủy hay không
6	Diverted	Diverted	Có bị điều hướng hay không
7	CRSDepTime	Computer reservation system departure time	Thời gian khởi hành theo máy tính (Computer Reservation System) – được định dạng theo hhmm
8	DepTime	Actual departure time	Thời gian khởi hành (theo thực tế) – được định dạng theo hhmm

9	DepDelay	Departure time delay	Chênh lệch số phút giữa thời gian khởi hành theo máy tính và thời gian khởi hành theo thực tế, nếu đến sớm thì là số âm
10	CRSArrTime	Computer reservation system arrival time	Thời điểm đến theo máy tính
11	ArrTime	Actual Arrival Time	Thời gian đến (theo thực tế)
12	ArrDelay	Arrival delay	Chênh lệch số phút giữa thời gian đến theo lịch và thời gian đến theo thực tế
13	CRSElapsedTime	Computer reservation system elapsed time of flight	Thời gian bay dự đoán (phút)
14	ActualElapsedTime	Actual elapsed time of flight	Thời gian bay thực tế (phút)
15	Distance	Distance	Khoảng cách giữa 2 sân bay (dặm)
16	Year	Year	Năm
17	Quarter	Quarter	Quý (1 – 4)
18	Month	Month	Tháng
19	DayofMonth	Day of month	Ngày trong tháng
20	DayOfWeek	Day of week	Ngày trong tuần
21	Marketing_Airline_Network	Marketing airline network	Mã hàng hàng không bán vé máy bay này
22	Operating_Airline	Operating airline	Mã hàng hàng không, là hàng hàng không thực hiện chuyến bay

23	Tail_Number	Tail number	Số đuôi của máy bay
24	Flight_Number_Market_Airline	Flight number marketing airline	Số hiệu chuyến bay
25	OriginCityName	Origin city name	Tên thành phố khởi hành
26	OriginStateFips	Origin state Fips	Số hiệu bang khởi hành
27	OriginStateName	Origin state name	Tên bang đích đến
28	DestCityName	Dest city name	Tên thành phố đích đến
29	DestStateFips	Dest state Fips	Số hiệu bang đích đến
30	DestStateName	Dest state name	Tên bang đích đến

Bảng 1.1. Thuộc tính được phân tích

1.3. Dictionary

Dictionary thuộc tính Cancelled.

Thuộc tính Cancelled thể hiện chuyến bay có bị hủy hay không.

Cancelled		
STT	Tên	Ý nghĩa
1	FALSE	Chuyến bay không bị hủy
2	TRUE	Chuyến bay bị hủy

Bảng 1.2. Dictionary thuộc tính Cancelled.

Dictionary thuộc tính Diverted.

Thuộc tính Diverted thể hiện chuyến bay có bị điều hướng hay không.

Diverted		
STT	Tên	Ý nghĩa
1	FALSE	Chuyến bay không bị điều hướng
2	TRUE	Chuyến bay bị điều hướng

Bảng 1.3. Dictionary thuộc tính Diverted.

Dictionary thuộc tính AirlineName.

Thuộc tính AirlineName thể hiện tên hãng hàng không. Dưới đây là 1 số miêu tả được trích từ dữ liệu:

AirlineName		
STT	Tên	Ý nghĩa
1	Air Wisconsin Airlines Corp	Hãng hàng không Air Wisconsin
2	United Air Lines Inc.	Hãng hàng không United Airlines
3	Delta Air Lines Inc.	Hãng hàng không Delta Air Lines
4	Southwest Airlines Co.	Hãng hàng không Southwest Airlines
5	Republic Airlines	Hãng hàng không Republic Airlines
6	Horizon Air	Hãng hàng không Horizon Air

Bảng 1.4. Dictionary thuộc tính AilineName.

Dictionary thuộc tính DepTime.

Thuộc tính DepTime thể hiện thời gian khởi hành thực tế, được định dạng hhmm.

DepTime		
STT	Tên	Ý nghĩa
1	1123	Chuyến bay khởi hành vào 11 giờ 23 phút
2	704	Chuyến bay khởi hành vào 7 giờ 4 phút
3	2137	Chuyến bay khởi hành vào 21 giờ 37 phút

Bảng 1.5. Dictionary thuộc tính DepTime.

Dictionary thuộc tính OriginCityName.

Thuộc tính OriginCityName thể hiện tên thành phố khởi hành và mã bang. Dưới đây là 1 số miêu tả được trích từ dữ liệu.

OriginCityName		
STT	Tên	Ý nghĩa
1	Grand Junction, CO	Khởi hành từ thành phố Grand Junction, bang Colorado
2	Honolulu, HI	Khởi hành từ thành phố Honolulu, Hawaii
3	Chicago, IL	Khởi hành từ thành phố Chicago, Illinois
4	Cleveland, OH	Khởi hành từ thành phố Cleveland, Ohio
5	Washington, DC	Khởi hành từ thành phố Washington, Virginia
6	Orlando, FL	Khởi hành từ thành phố Orlando, Florida

Bảng 1.6. Dictionary thuộc tính OriginCityName.

Dictionary thuộc tính OriginStateName.

Thuộc tính OriginStateName thể hiện tên bang khởi hành. Dưới đây là 1 số miêu tả được trích từ dữ liệu

OriginStateName		
STT	Tên	Ý nghĩa
1	Colorado	Khởi hành từ bang Colorado
1	Ohio	Khởi hành từ bang Ohio
1	Florida	Khởi hành từ bang Florida
1	Virginia	Khởi hành từ bang Virginia
1	Iowa	Khởi hành từ bang Iowa

Bảng 1.7. Dictionary thuộc tính OriginStateName.

Dictionary thuộc tính ArrDelay.

Thuộc tính ArrDelay thể hiện số phút mà chuyến bay đến đích trễ hơn dự tính.

Diverted		
STT	Tên	Ý nghĩa
1	-17	Chuyến bay đến đích sớm 17 phút so với dự tính
2	6	Chuyến bay đến đích trễ 6 phút so với dự tính
3	0	Chuyến bay đến đích đúng như dự tính

Bảng 1.8. Dictionary thuộc tính ArrDelay.

1.4. Sơ đồ dữ liệu xây dựng

Các lược đồ đa chiều:

- Lược đồ đa chiều (multidimensional schema) là một cấu trúc dữ liệu được sử dụng trong việc tổ chức và lưu trữ dữ liệu trong hệ thống kho dữ liệu (data warehouse). Nó được thiết kế để hỗ trợ việc truy vấn và phân tích dữ liệu hiệu quả bằng cách cho phép các dữ liệu được tổ chức theo nhiều chiều khác nhau (dimensions).
- Trong lược đồ đa chiều, các dữ liệu được tổ chức thành các bảng (table) hay các cube, trong đó mỗi chiều của dữ liệu được tạo thành từ một bảng dữ liệu. Các bảng này được liên kết với nhau bằng cách sử dụng các khóa ngoại, tạo thành một cấu trúc đa chiều.
- Có 2 loại lược đồ đa chiều:^[2]

1.4.1. Lược đồ hình sao (Star schema)

Lược đồ sao là loại lược đồ Kho dữ liệu đơn giản nhất. Nó được gọi là lược đồ sao vì cấu trúc của nó giống như một ngôi sao. Trong lược đồ hình sao, tâm của ngôi sao có thể có một bảng sự kiện (fact) và số lượng bảng chiều (dimension) được liên kết. Nó còn được gọi là Star Join Schema và được tối ưu hóa để truy vấn các tập dữ liệu lớn.

Ưu điểm:

- + Các truy vấn đơn giản hơn – logic nối lược đồ sao thường đơn giản hơn logic nối được yêu cầu để truy xuất dữ liệu từ lược đồ giao dịch được chuẩn hóa cao.
- + Tăng hiệu suất truy vấn – các lược đồ sao có thể cung cấp các cải tiến hiệu suất cho các ứng dụng báo cáo chỉ đọc khi so sánh với các lược đồ được chuẩn hóa cao.
- + Tổng hợp nhanh – các truy vấn đơn giản hơn đối với lược đồ sao có thể dẫn đến hiệu suất được cải thiện cho các hoạt động tổng hợp.

Nhược điểm:

- + Tính toàn vẹn dữ liệu không được thực thi tốt vì nó không ở trạng thái chuẩn hóa cao.
- + Lược đồ hình sao không linh hoạt về mặt nhu cầu phân tích như mô hình dữ liệu chuẩn hóa
- + Khả năng mở rộng hạn chế: Nếu số chiều dữ liệu tăng lên, lược đồ hình sao có thể trở nên quá phức tạp và khó mở rộng.
- + Cấu trúc lưu trữ phức tạp: Lược đồ hình sao yêu cầu phải tạo các bảng trung tâm để kết nối các bảng dữ liệu khác nhau, điều này dẫn đến cấu trúc lưu trữ phức tạp.

1.4.2. Lược đồ hình bông tuyết (Snowflake schema)

Lược đồ Bông tuyết là một phần mở rộng của Lược đồ hình sao nhưng các bảng chi tiết được chia nhỏ thành các bảng nhỏ hơn để tạo ra cấu trúc phân cấp. Nó được gọi là bông tuyết vì sơ đồ của nó giống như một Bông tuyết.

Ưu điểm:

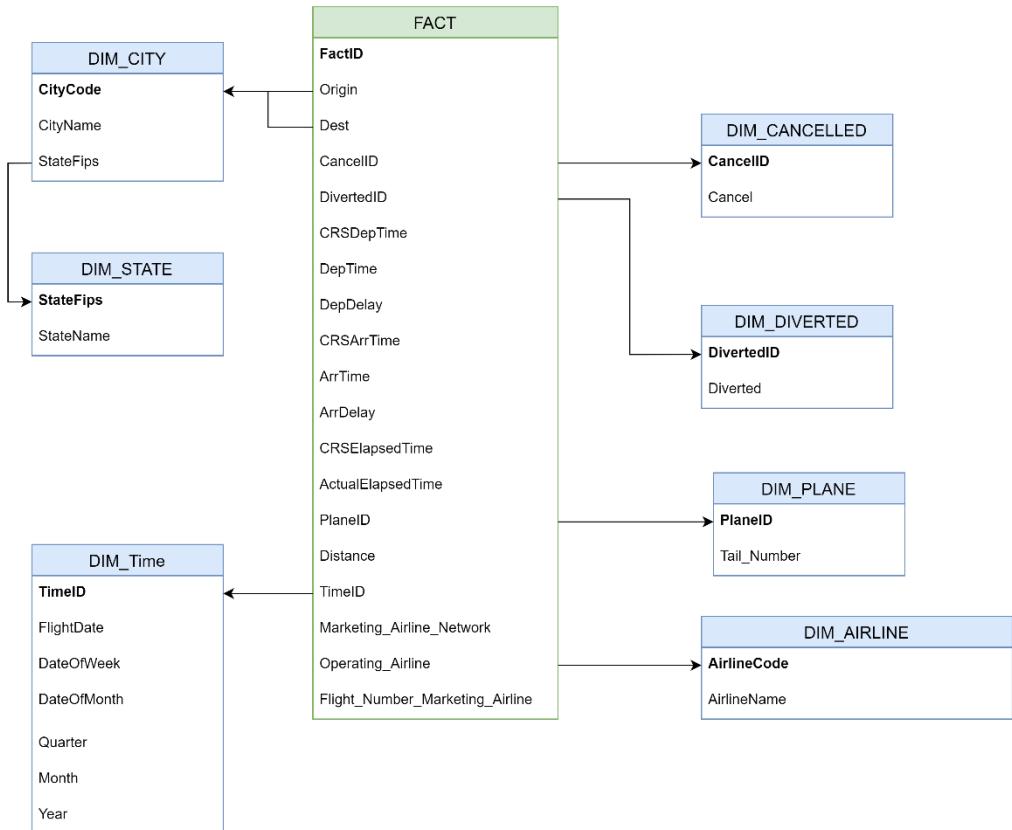
- + Một số công cụ mô hình hóa cơ sở dữ liệu đa chiều (OLAP) được tối ưu hóa cho các lược đồ bông tuyết.
- + Đơn giản hóa các thuộc tính dẫn đến sự tiết kiệm, nhưng đánh đổi là sự phức tạp bổ sung trong các truy vấn nguồn.
- + Một số chiều được phân cấp để thể hiện rõ ràng định chuẩn của bảng chiều.

Hạn chế:

- + Mức chuẩn hóa thuộc tính bổ sung thêm độ phức tạp cho các phép truy vấn nguồn so với lược đồ hình sao.

1.5. Xây dựng kho dữ liệu

1.5.1. Sơ đồ bông tuyết minh họa



Hình 1.2. Sơ đồ bông tuyết minh họa

1.5.2. Mô tả chi tiết các bảng dữ liệu

DIM_CITY

	Tên cột	Kiểu dữ liệu	NULL	Mô tả
Khóa chính	CityCode	nvarchar(255)	NOT	Mã thành phố
	CityName	nvarchar(255)	NOT	Tên thành phố
	StateFips	int	NOT	Mã bang

Bảng 1.9. Mô tả chi tiết bảng Dim_City.

DIM_STATE

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính				
	StateFips	int	NOT	Mã thành phố
	StateName	nvarchar(255)	NOT	Tên thành phố

Bảng 1.10. Mô tả chi tiết bảng Dim_State.

DIM_TIME

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính				
	FlightDate	Date	NOT	Ngày
	DateOfWeek	int	NOT	Thứ trong tuần
	Quarter	int	NOT	Quý
	Month	int	NOT	Tháng
	Year	int	NOT	Năm

Bảng 1.11. Mô tả chi tiết bảng Dim_Time

DIM_CANCELLED

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính				
	CancelId	int	NOT	Mã tình trạng hủy của chuyến bay
	Cancel	int	NOT	Có hủy chuyến bay hay không

Bảng 1.12. Mô tả chi tiết bảng Dim_Cancelled

DIM_DIVERTED

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính	DivertedId	int	NOT	Mã tình trạng điều hướng của chuyến bay
	Diverted	int	NOT	Có điều hướng chuyến bay hay không

Bảng 1.13. Mô tả chi tiết bảng Dim_Diverted

DIM_PLANE

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính	Tail_Number	nvarchar(255)	NOT	Số đuôi máy bay

Bảng 1.14. Mô tả chi tiết bảng Dim_Plane

DIM_AIRLINE

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính	AirlineCode	nvarchar(255)	NOT	Mã hãng hàng không
	Name	nvarchar(255)	NOT	Tên hãng hàng không

Bảng 1.15. Mô tả chi tiết bảng Dim_Airline

FACT

Tên cột		Kiểu dữ liệu	NULL	Mô tả
Khóa chính	FactID	int	NOT	Mã fact
	FlightDate	Date	NOT	Ngày bay
	Origin	nvarchar(255)	NOT	Mã thành phố xuất phát
	Dest	nvarchar(255)	NOT	Mã thành phố đích
	CancelID	int	NOT	Mã tình trạng hủy của chuyến bay
	DivertedID	int	NOT	Mã tình trạng điều hướng của chuyến bay

	CRSDepTime	int	NOT	Thời gian khởi hành theo máy tính (Computer Reservation System) – định dạng kiểu hhmm
	DepTime	int	NOT	Thời gian khởi hành (theo thực tế) – được định dạng theo hhmm
	DepDelay	int	NOT	Chênh lệch số phút giữa thời gian khởi hành theo máy tính và thời gian khởi hành theo thực tế, nếu đến sớm thì là số âm
	CRSArrTime	int	NOT	Thời điểm đến theo máy tính
	ArrTime	int	NOT	Thời gian đến (theo thực tế)
	ArrDelay	int	NOT	Chênh lệch số phút giữa thời gian đến theo lịch và thời gian đến theo thực tế
	CRSElapsedTime	int	NOT	Thời gian bay dự đoán (phút)
	ActualElapsedTime	int	NOT	Thời gian bay thực tế (phút)
	Distance	int	NOT	Khoảng cách giữa 2 sân bay (dặm)

	Marketing_Airline_Network	nvarchar(255)	NOT	Mã hàng hàng không bán vé vé máy bay
	Operating_Airline	nvarchar(255)	NOT	Mã hàng hàng không, là hàng hàng không thực hiện chuyến bay
	Tail_Number	nvarchar(255)	NOT	Số đuôi
	Flight_Number_Operation_Airline	nvarchar(255)	NOT	Số hiệu chuyến bay

Bảng 1.16. Mô tả chi tiết bảng Fact.

CHƯƠNG 2 – QUÁ TRÌNH XÂY DỰNG KHO DỮ LIỆU (SSIS)

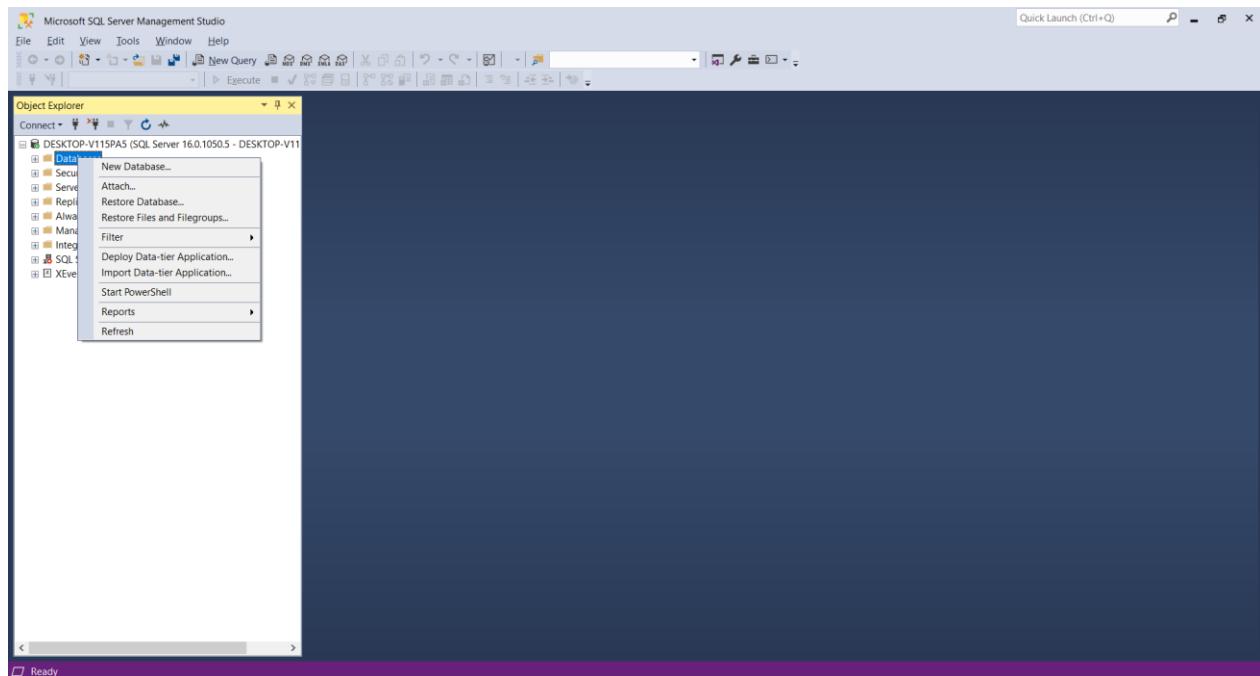
Đến với chương 2, ta sẽ tập trung vào quá trình xây dựng kho dữ liệu (SSIS) để phục vụ cho việc phân tích dữ liệu dự đoán tình trạng chuyến bay. Quá trình này đòi hỏi sự chuẩn bị và tích hợp dữ liệu từ nhiều nguồn khác nhau, sau đó tiến xử lý để dữ liệu trở nên chuẩn mực và sẵn sàng cho việc phân tích.

2.1. Chuẩn bị công cụ và Data warehouse

- Tải Microsoft Visual Studio 2022.
- Tải Microsoft SQL Server Management Studio 2022.
- Tải công cụ SQL Server Data Tools cho phiên bản Visual Studio 2022.

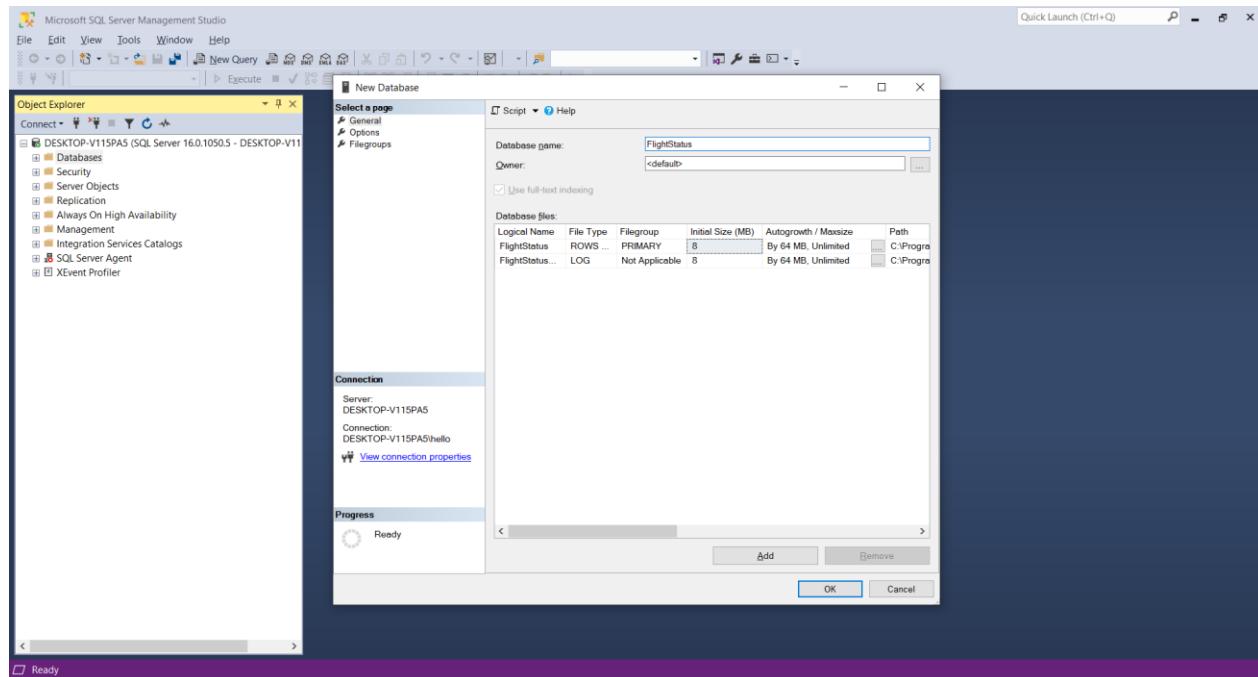
2.2. Tạo new database trong Microsoft SQL Server Management Studio

Bước 1: Nhấn chuột phải vào Database, sau đó chọn New Database để thực hiện tạo.



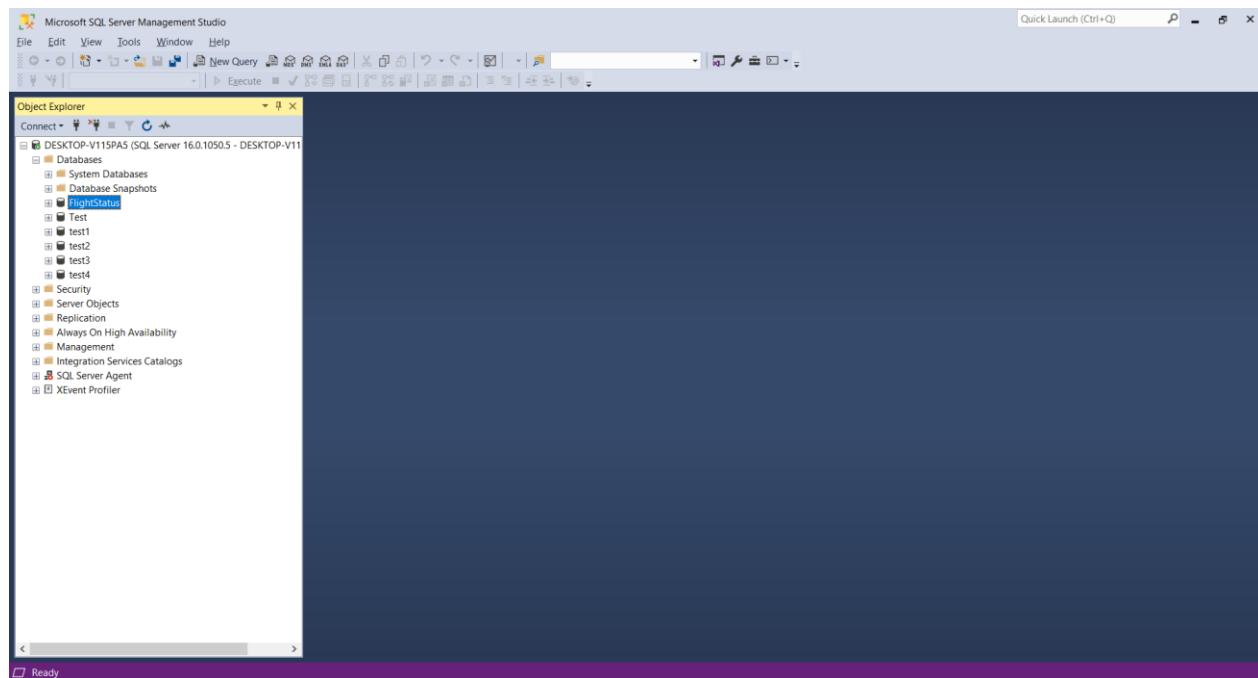
Hình 2.1. Tạo New Database

Bước 2: Đặt tên cho cơ sở dữ liệu mới tạo, sau đó nhấn OK.



Hình 2.2. Điện thoại New Database

Kết quả:

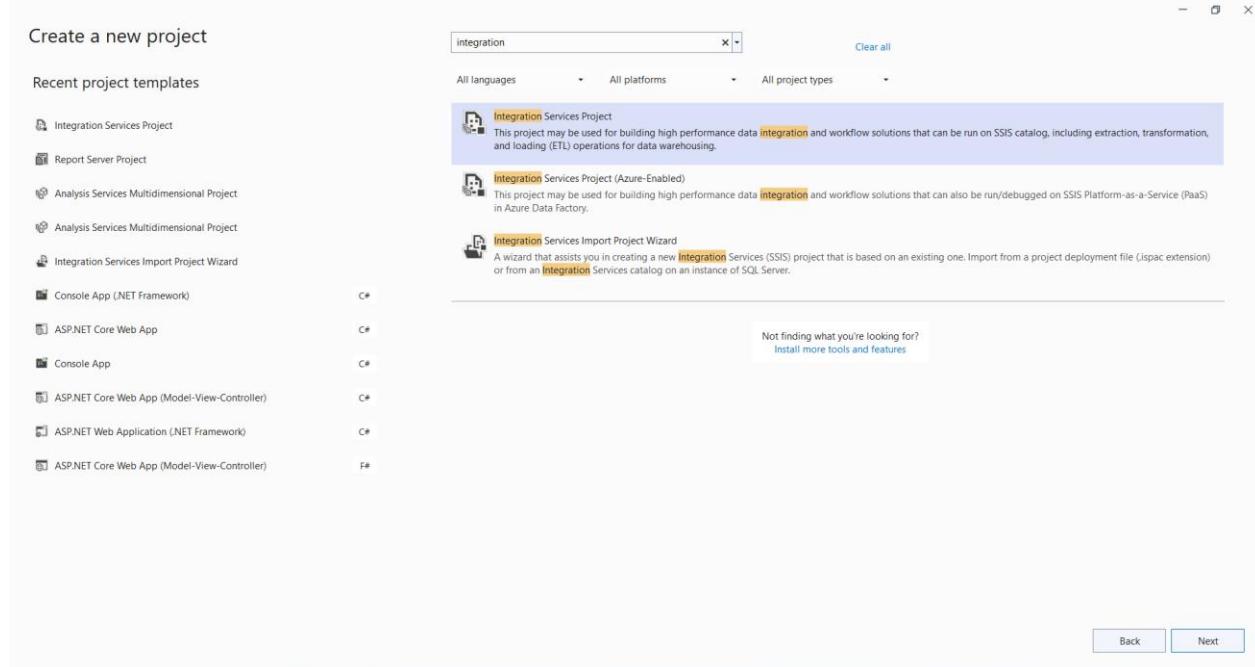


Hình 2.3. Kết quả sau khi tạo Database

2.3. Quá trình SSIS

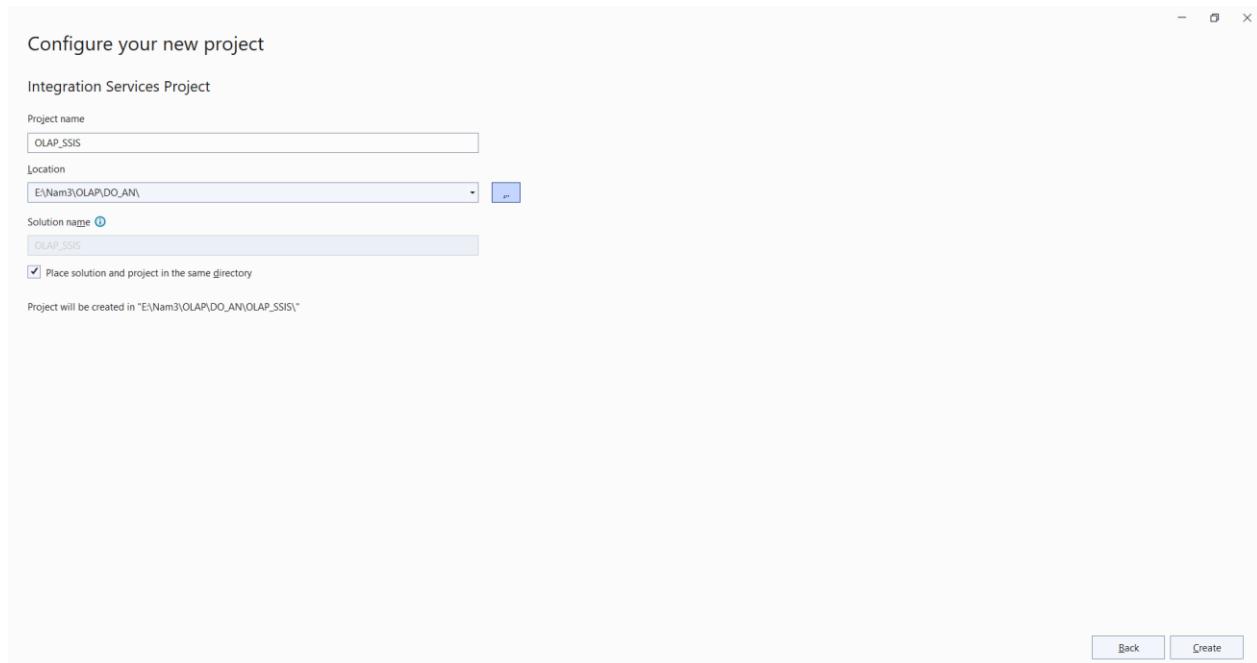
2.3.1. Khởi tạo project

- Bước 1: Tìm từ khóa integration, sau đó chọn Integration Services Project, chọn tiếp Next.



Hình 2.4. Tạo mới project SSIS

- Bước 2: Điền thông tin project.

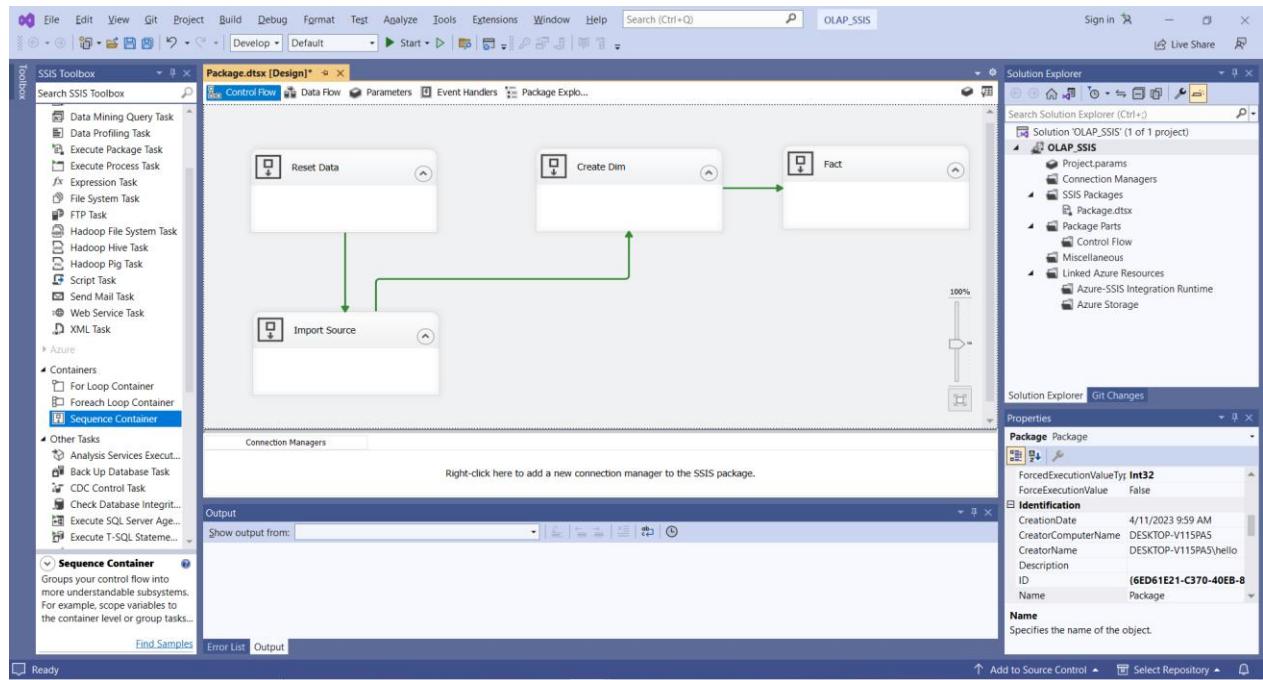


Hình 2.5. Điền thông tin project

2.3.2. Khởi tạo Sequence Container

Tạo các Sequence Container:

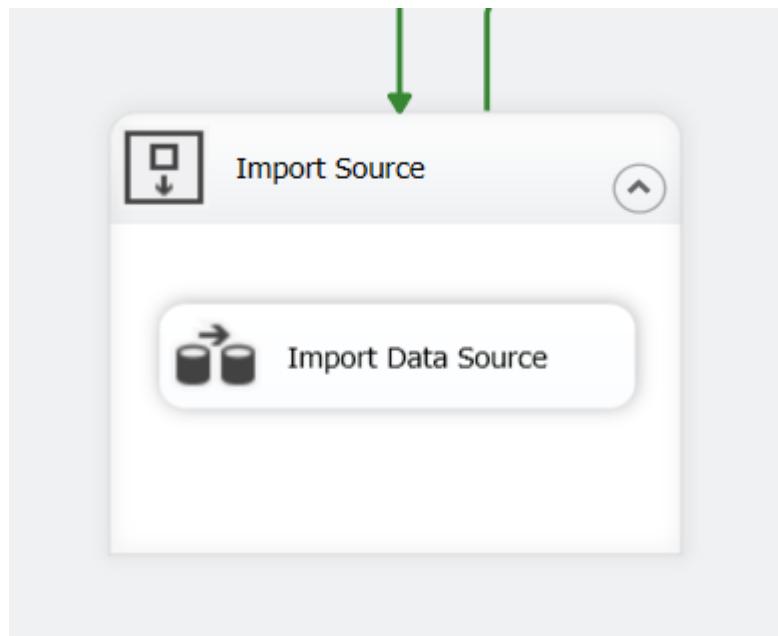
- Reset Data
- Create Dim
- Import Source
- Fact



Hình 2.6. Khởi tạo Sequence Container

2.3.3. Import Data

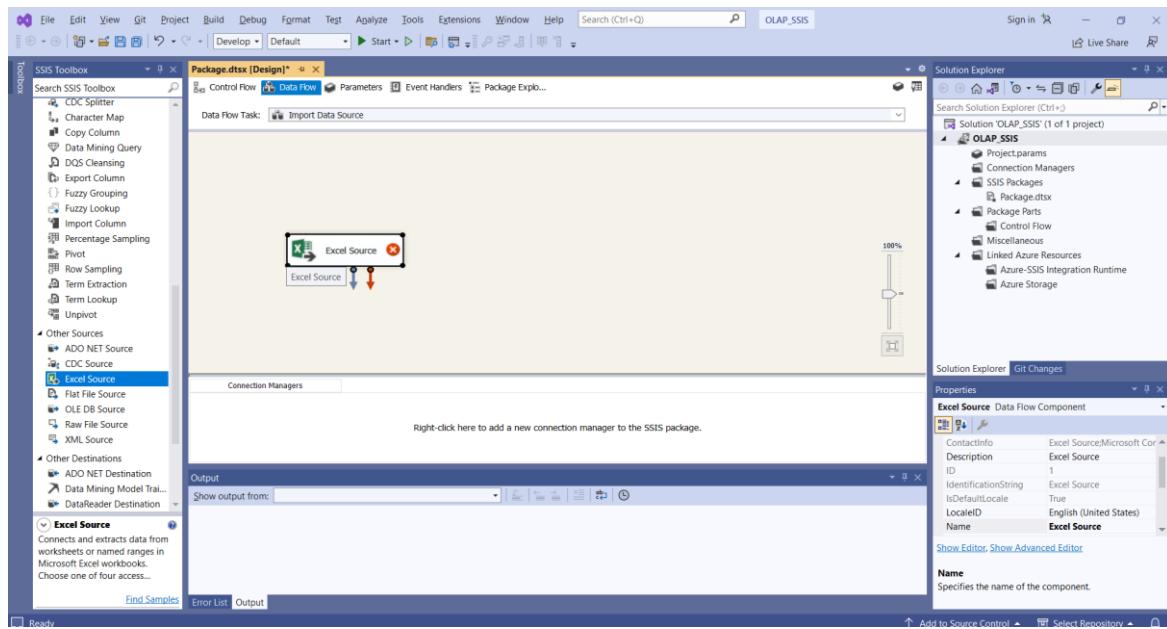
- Bước 1: Tạo một Data Flow Task và đổi tên thành Import Data Source.



Hình 2.7. Tạo một Data Flow Task tên Import Data Source

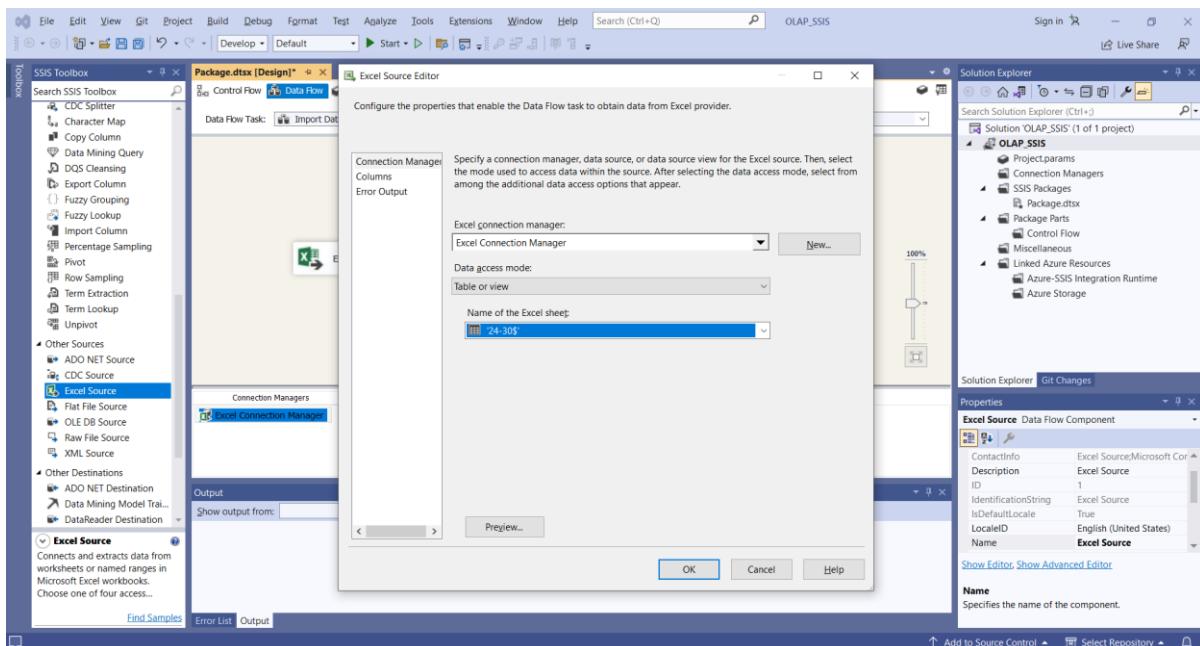
IS217 – Kho dữ liệu và OLAP

- Bước 2: Tạo Excel Source và để chuẩn bị để dữ liệu file data_raw.xlsx vào Excel Source đó.



Hình 2.8. Tạo Excel Source

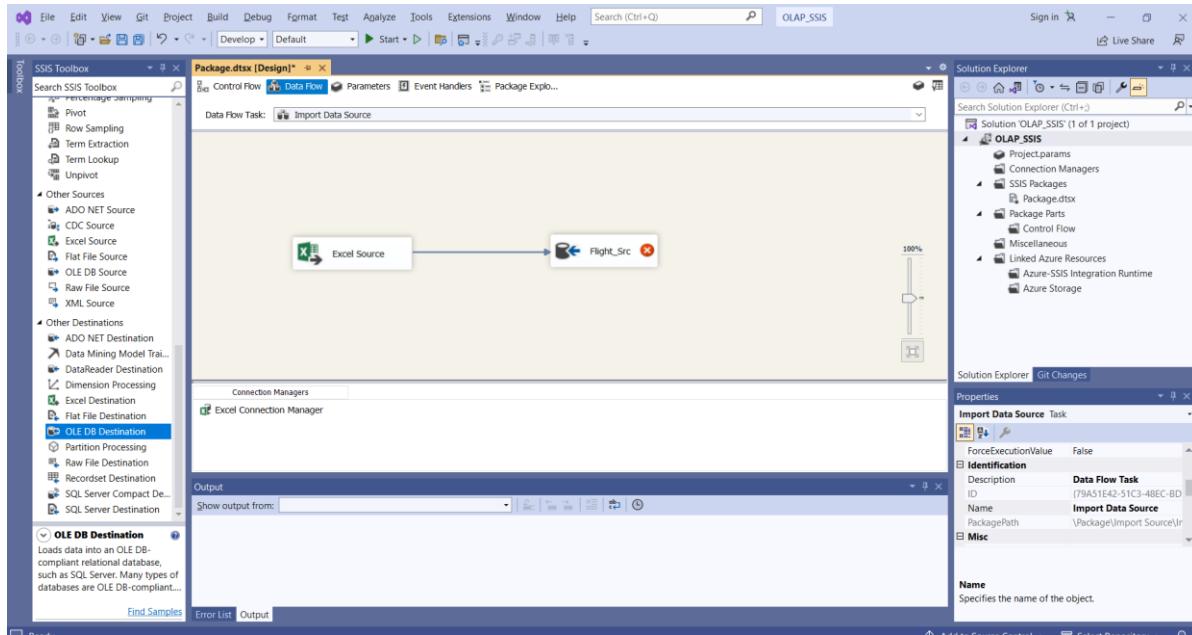
- Bước 3: Nhấn đúp vào Excel Source vừa tạo để chọn sheet excel mình muốn.



Hình 2.9. Khởi tạo Database

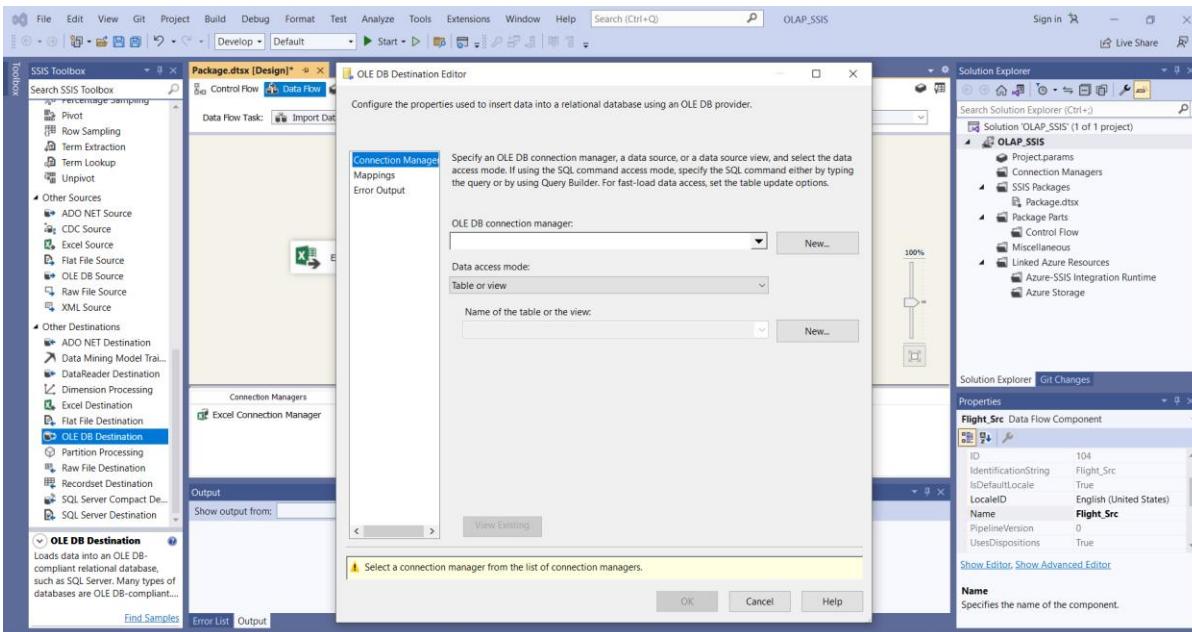
IS217 – Kho dữ liệu và OLAP

- Bước 4: Tạo 1 OLE DB Destination và đổi tên Flight_Src để chứa dữ liệu đó từ Excel Source.



Hình 2.10. Tạo OLE DB Destination

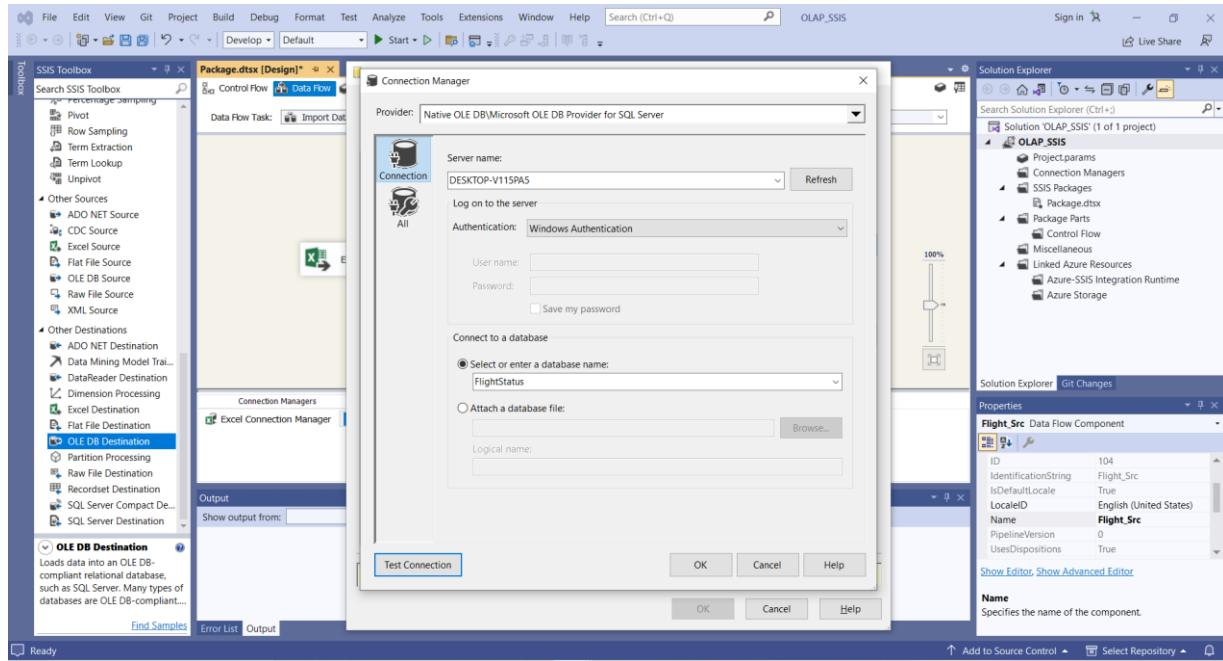
- Bước 5: Nhấn đúp vào OLE DB Destination để bắt đầu tạo một OLE DB Connection manager mới.



Hình 2.11. Tạo OLE DB Connection Manager

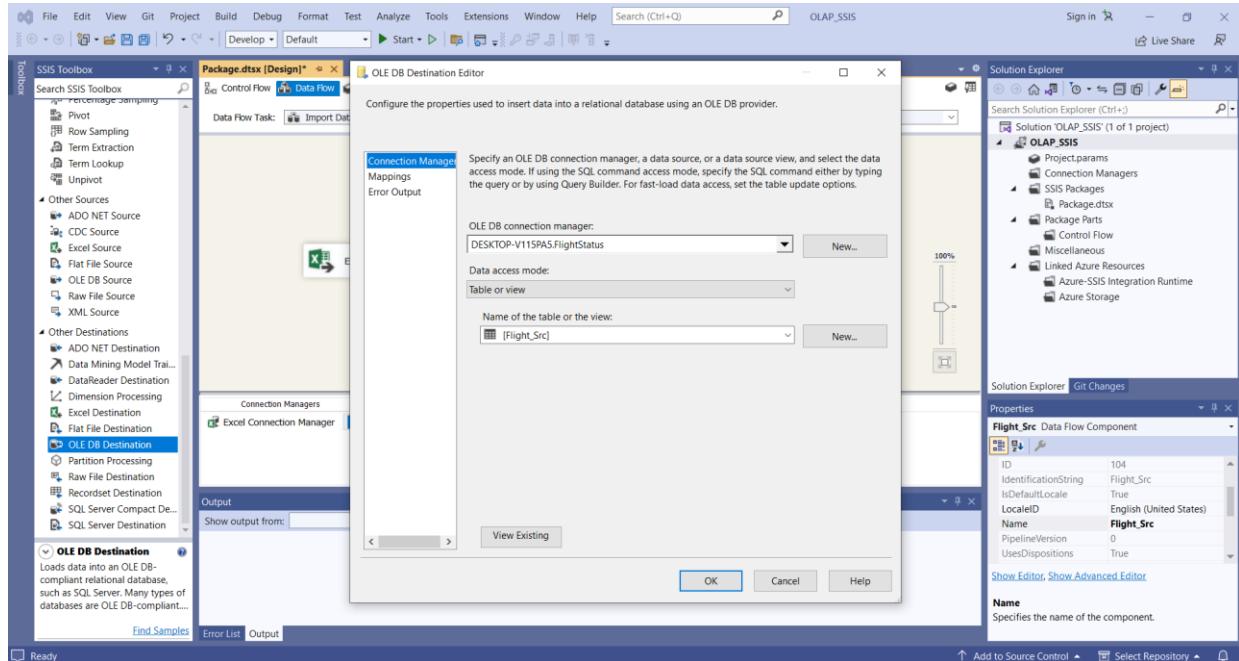
IS217 – Kho dữ liệu và OLAP

- Bước 6: Chọn New để tạo Connection manager mới, điền Server name sau đó chọn database muốn đổ vào, sau đó chọn OK.

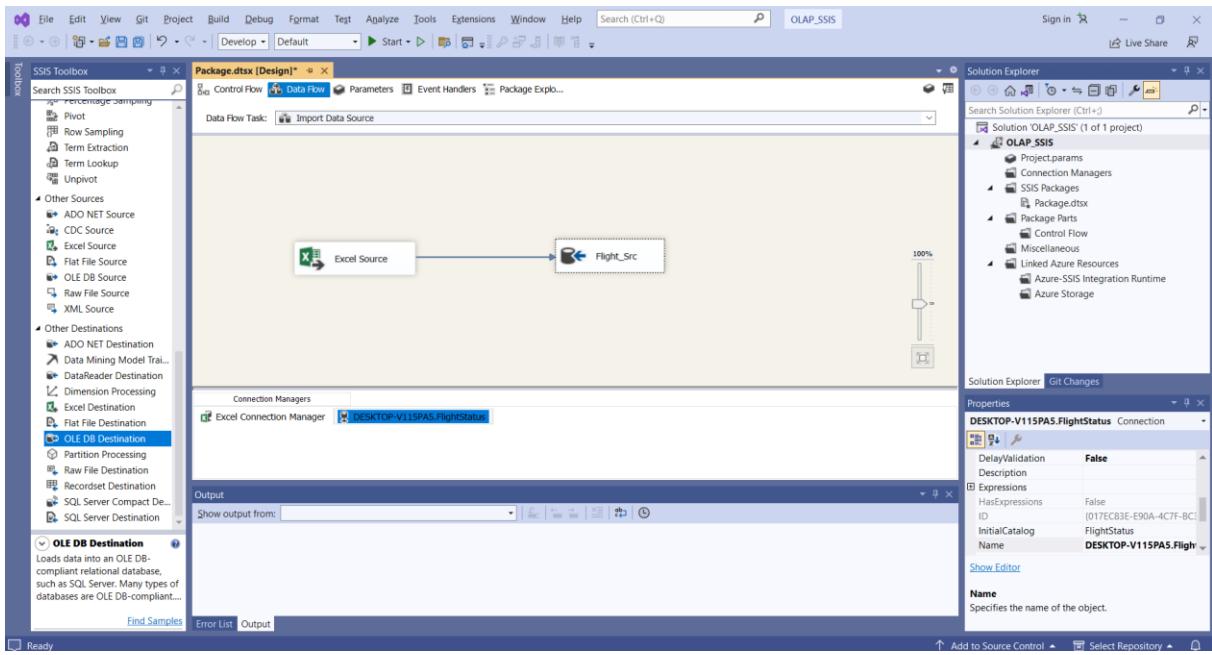


Hình 2.12. Điều thông tin Connection Manager

- Bước 7: Chọn New ở Name of table or view để tạo table.



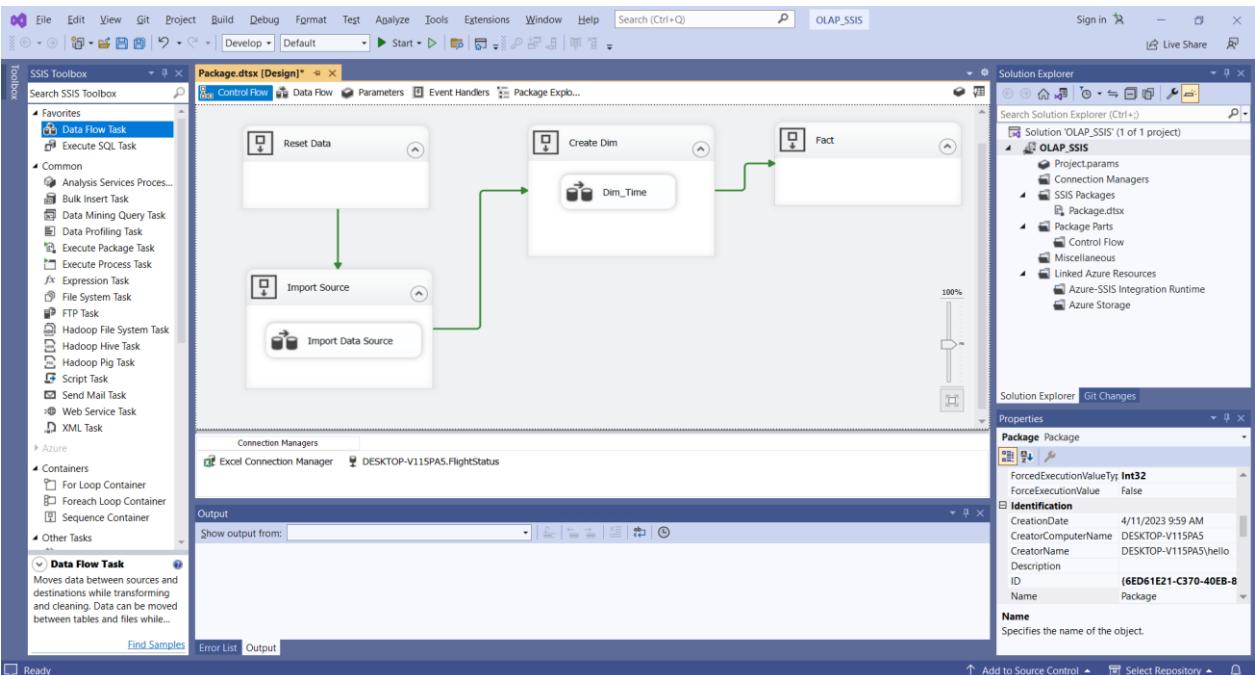
Hình 2.13. Kết quả sau khi Create Table Flight_Src



Hình 2.14. Kết quả sau khi import data.

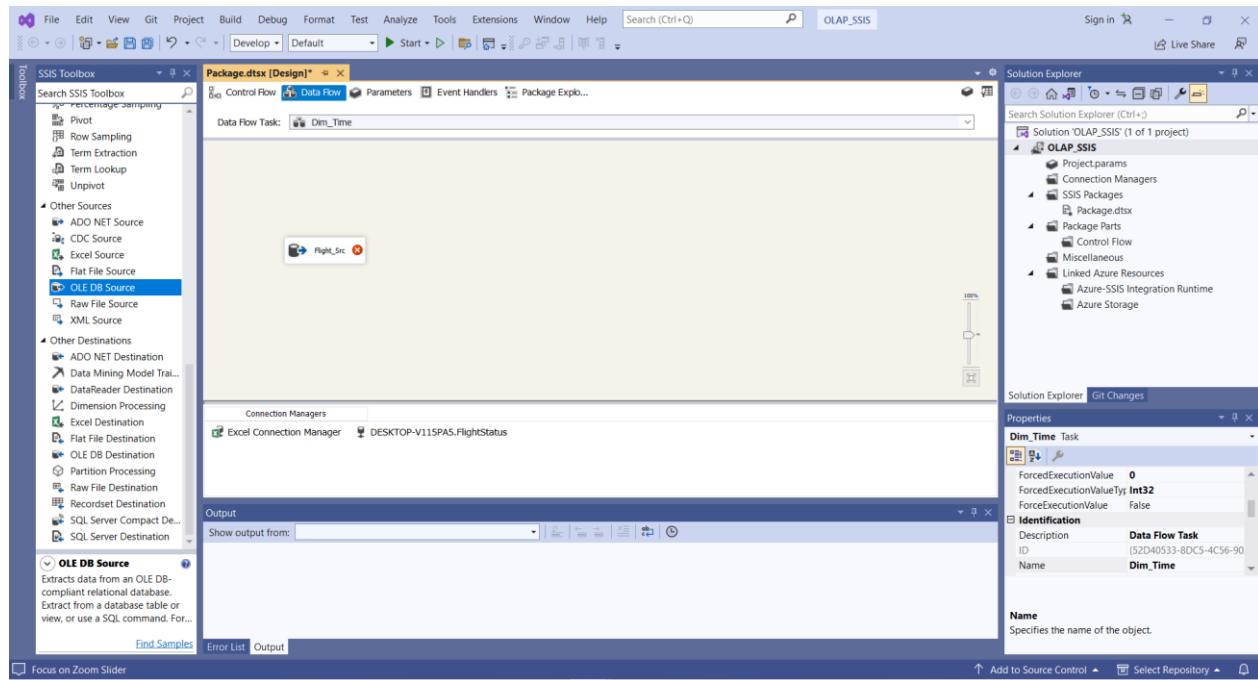
2.3.4. Tạo bảng Dim_Time

- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_Time



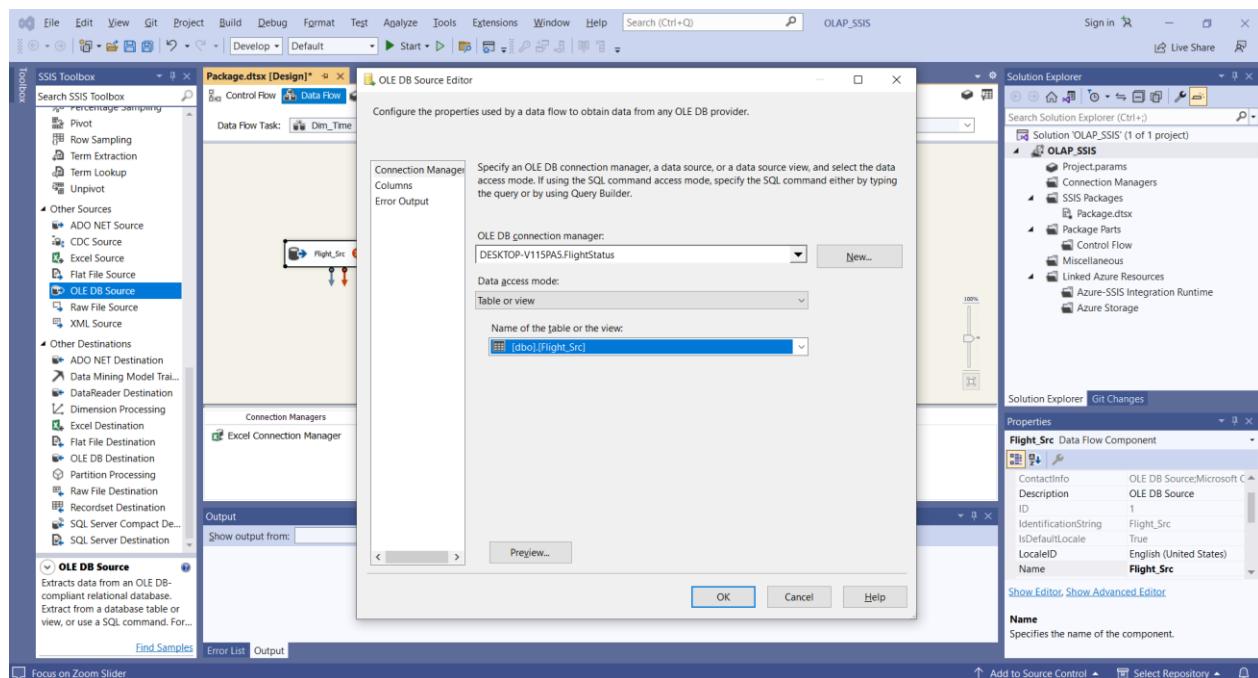
Hình 2.15. Tạo Data Flow Task Dim_Time

- Bước 2: Khởi tạo OLE DB Source chứa Dataset



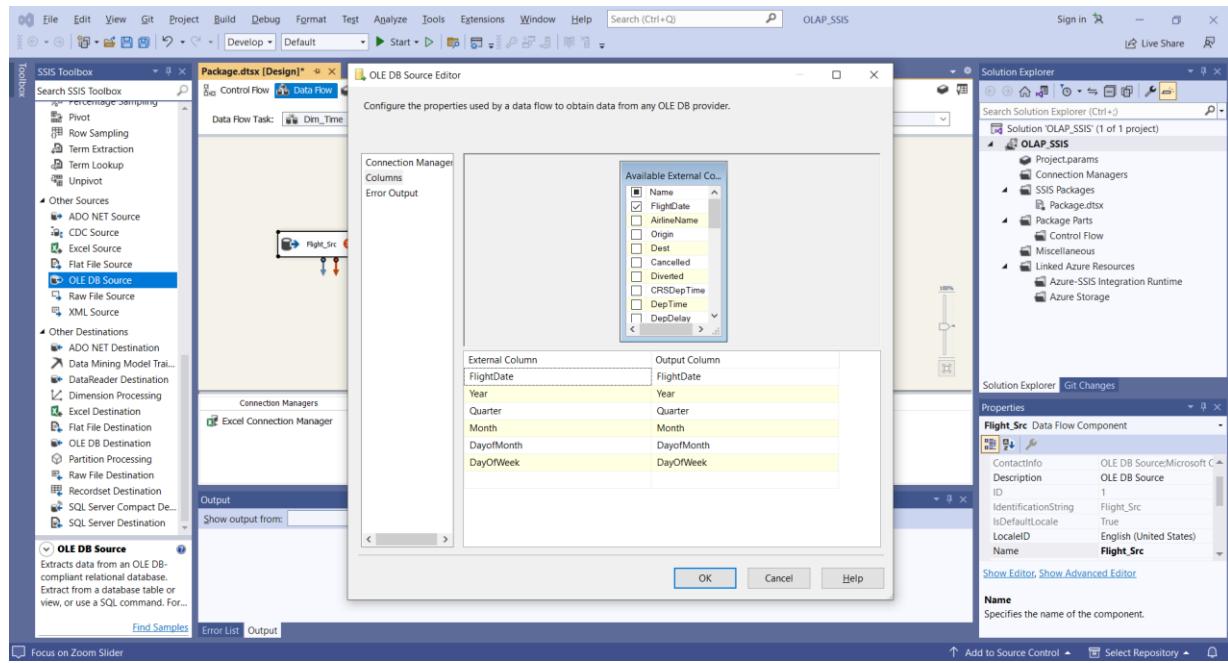
Hình 2.16. Tạo OLE DB Source chứa Dataset

- Bước 3: Chọn Connection FlightStatus và chọn bảng Flight_Src



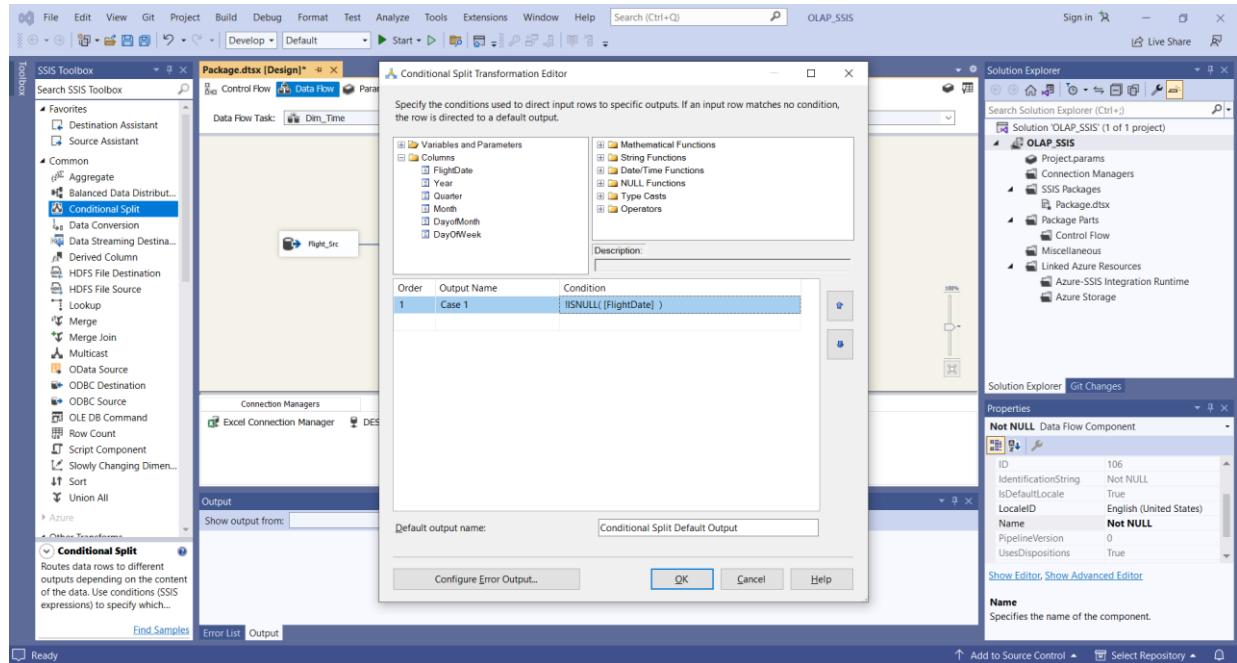
Hình 2.17. Chọn Connection và bảng

- Bước 4: Preview bảng. Ở mục tab Column chọn thuộc tính Time sau đó chọn OK



Hình 2.18. Chọn thuộc tính Time

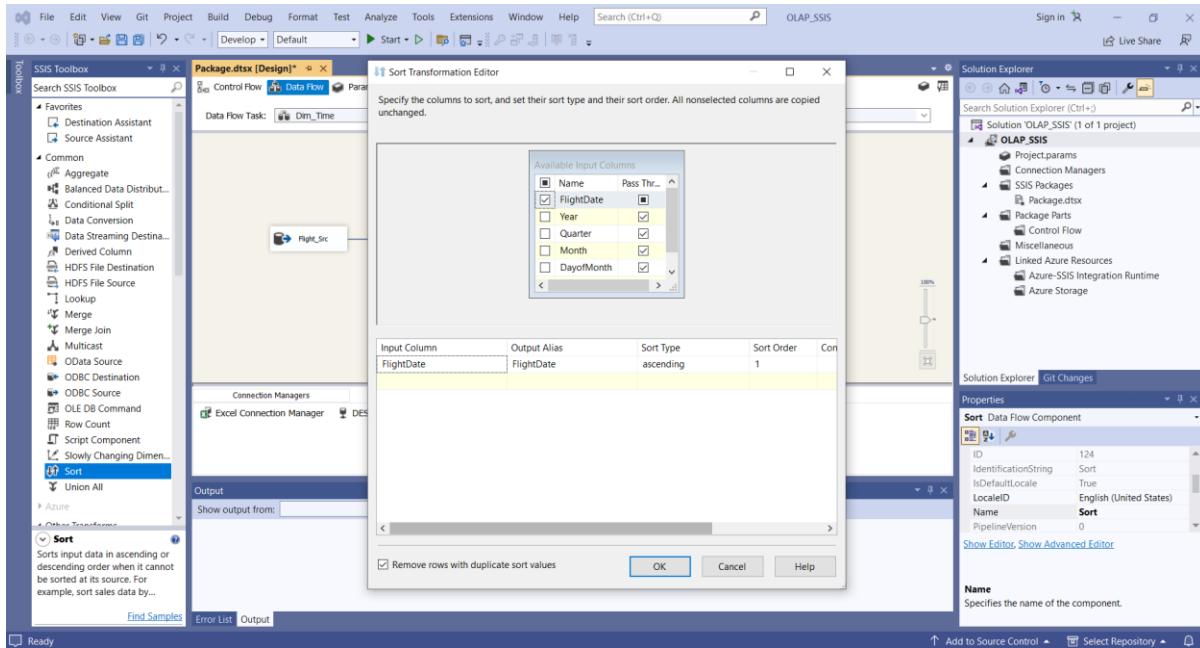
- Bước 5: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



Hình 2.19. Thêm Conditional Split

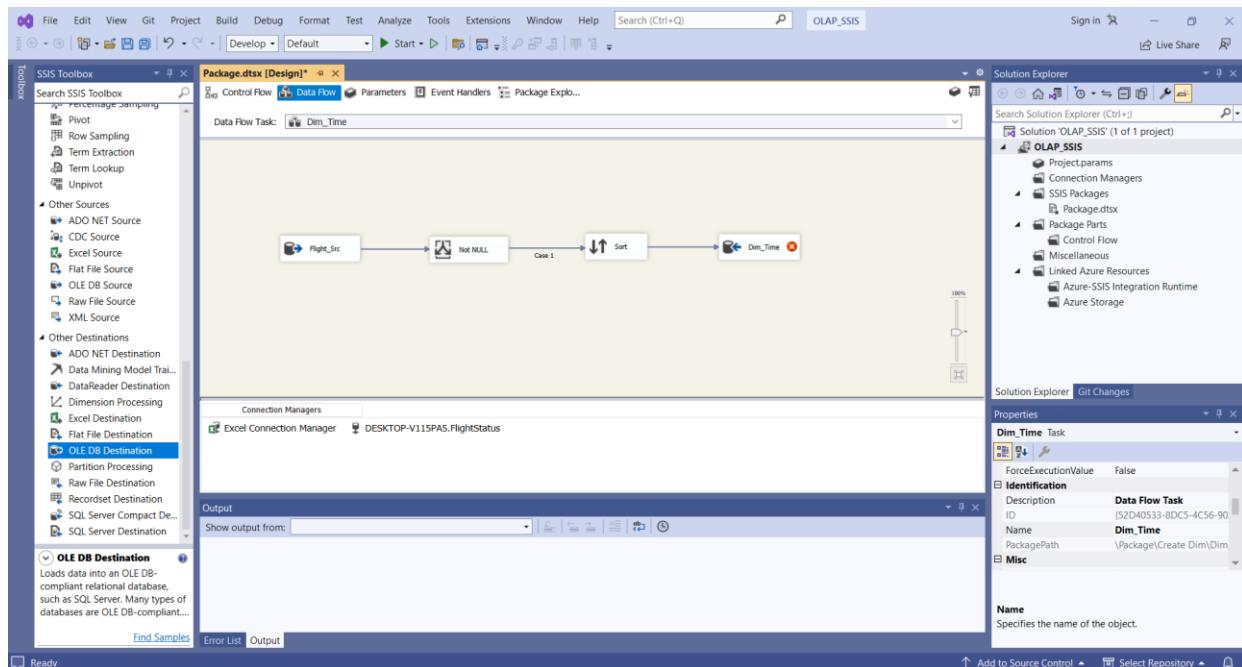
IS217 – Kho dữ liệu và OLAP

- Bước 6: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu FlightDate bị trùng.



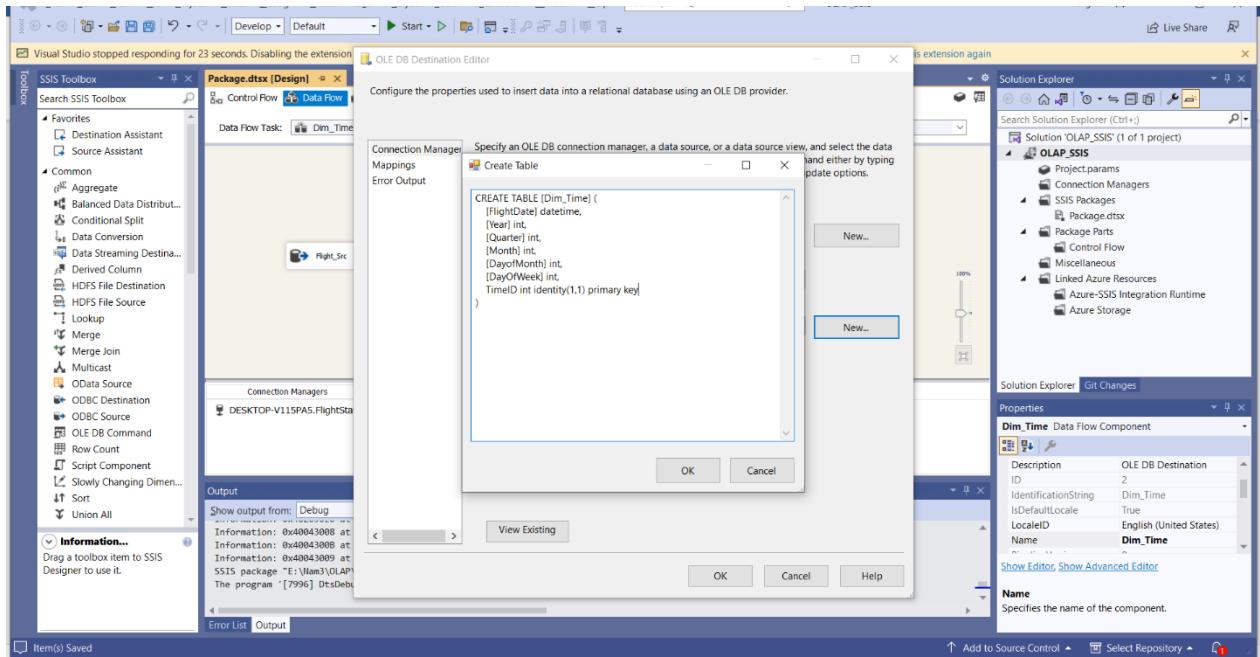
Hình 2.20. Sort dữ liệu và Remove rows.

- Bước 7: Tạo 1 OLE DB Destination sau đó tạo bảng Dim_Time.



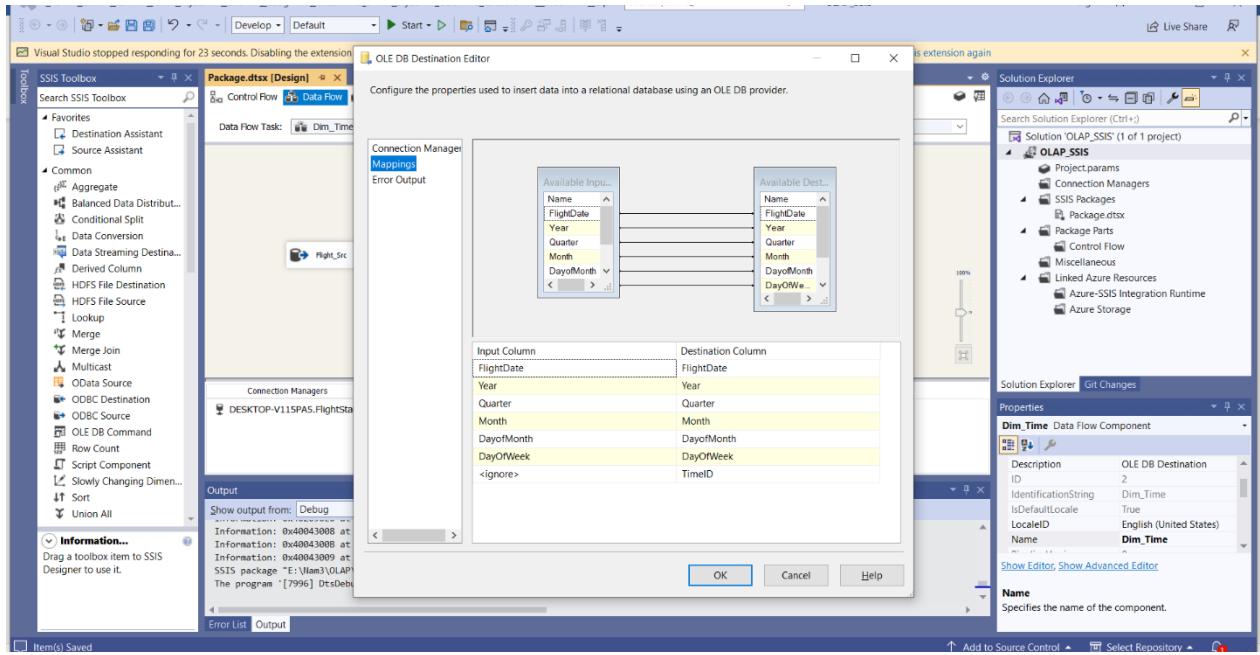
Hình 2.21. Tạo 1 OLE Destination tên Dim_Time

IS217 – Kho dữ liệu và OLAP



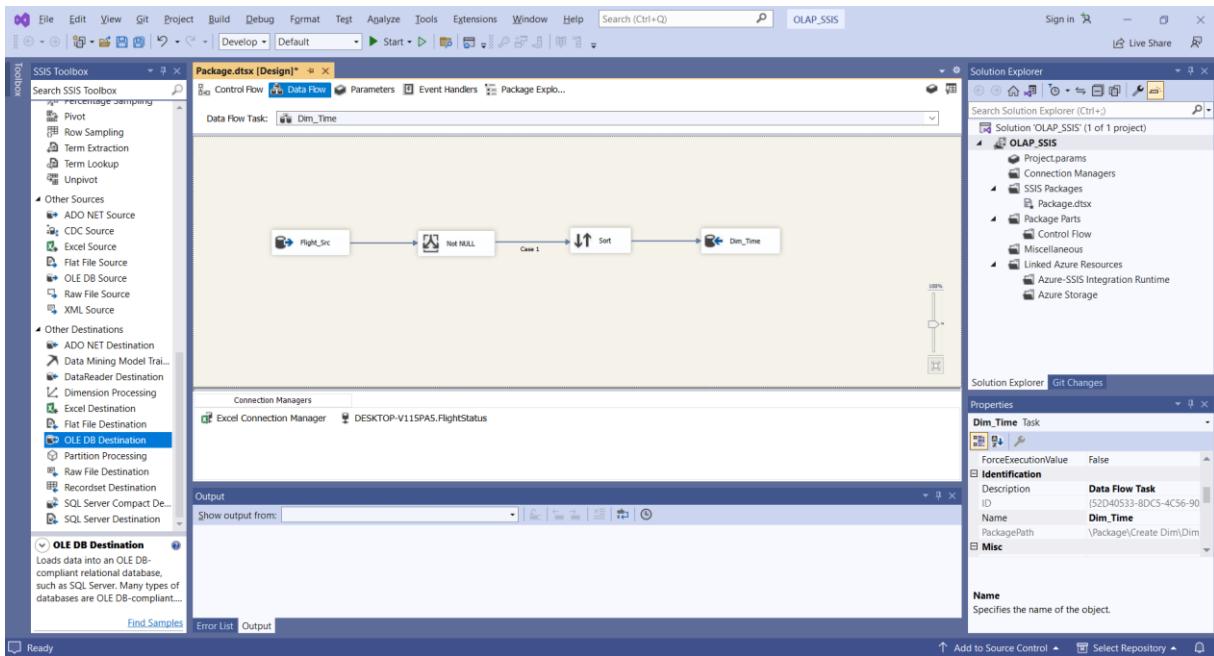
Hình 2.22. Tạo bảng Dim_Time

- Bước 8: Qua tab Mapping để kiểm tra.



Hình 2.23. Qua Mapping để kiểm tra

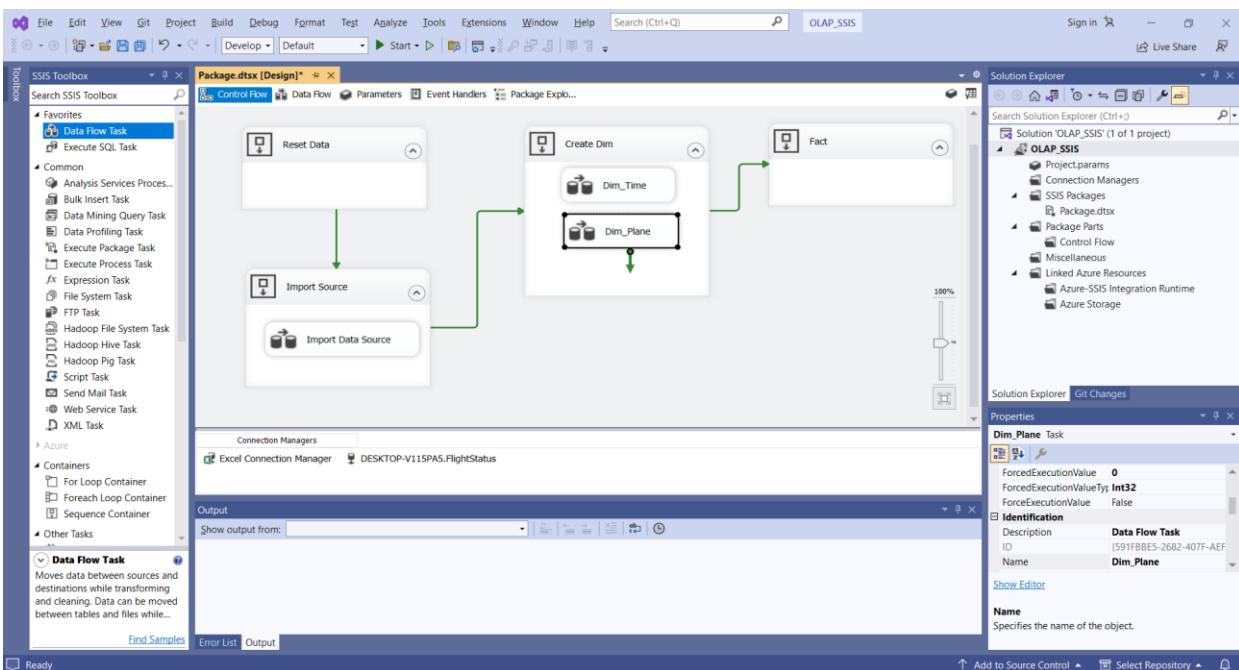
- Kết quả luồng thực hiện của bảng Dim_Time



Hình 2.24. Luồng thực hiện của bảng Dim_Time

2.3.5. Tạo bảng Dim_Plane

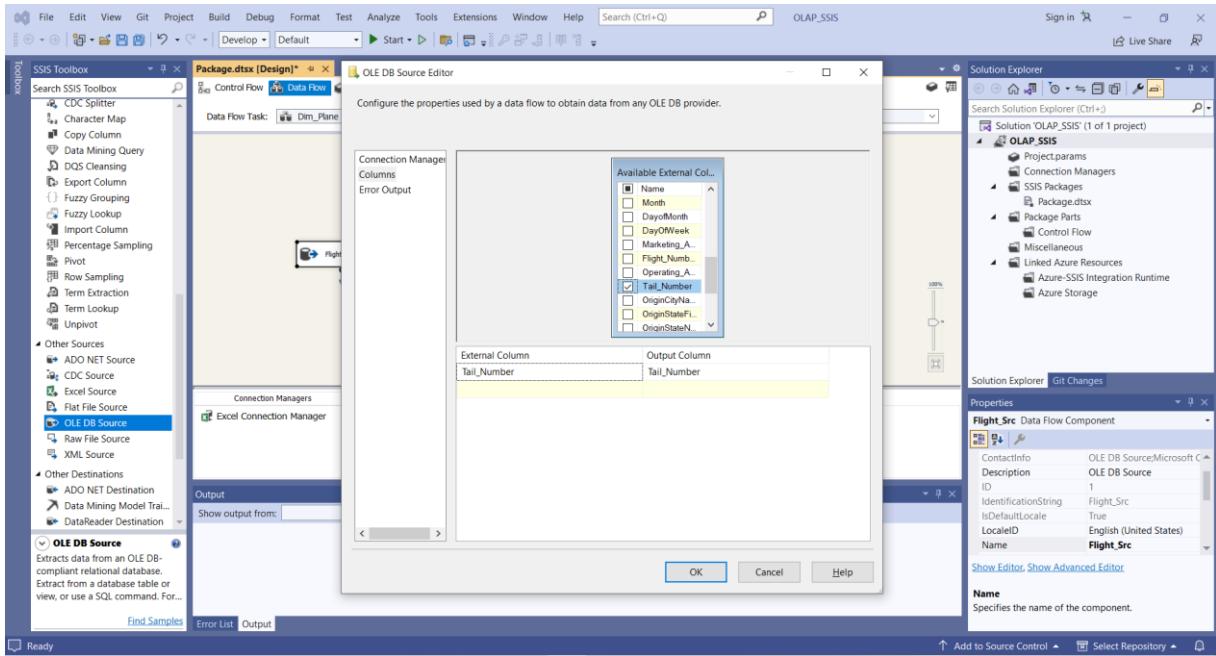
- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_Plane



Hình 2.25. Tạo Data Flow Task Dim_Plane

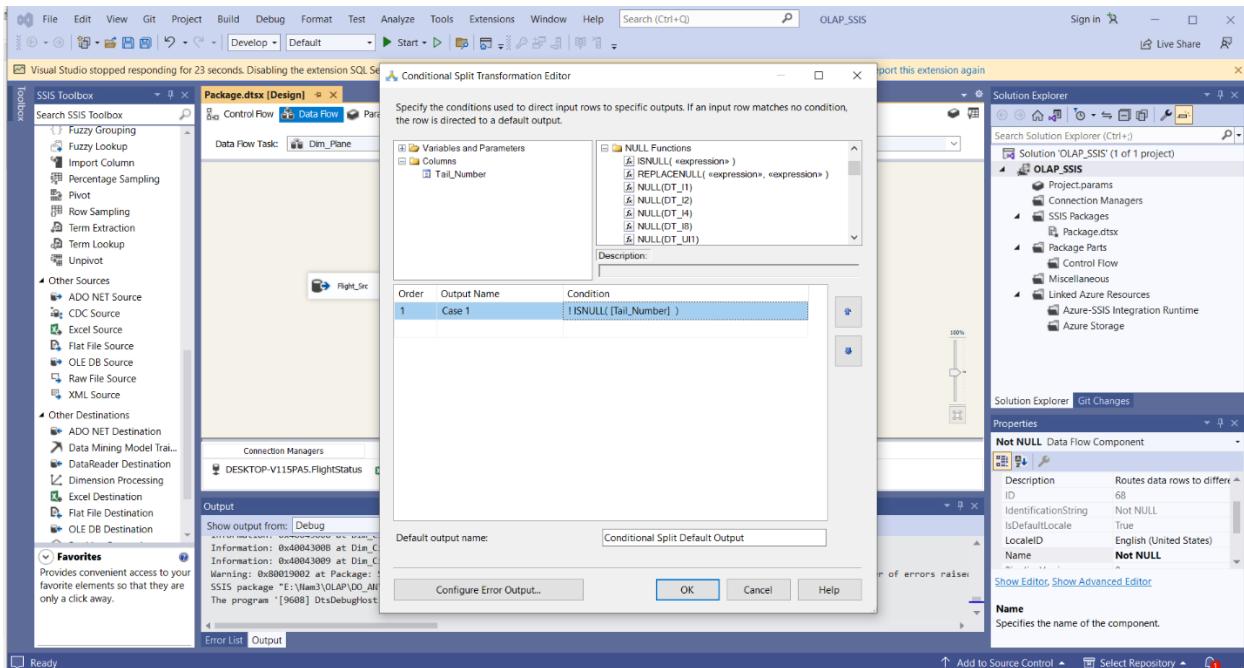
IS217 – Kho dữ liệu và OLAP

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng.



Hình 2.26. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

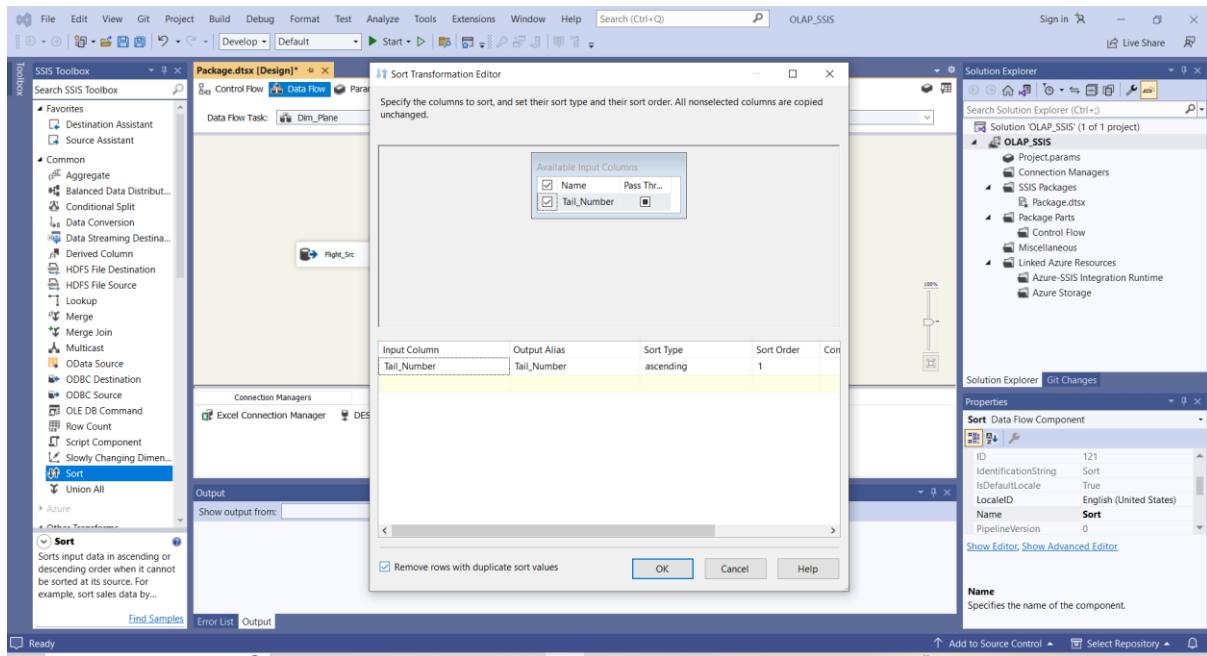
- Bước 3: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



Hình 2.27. Thêm Conditional Split

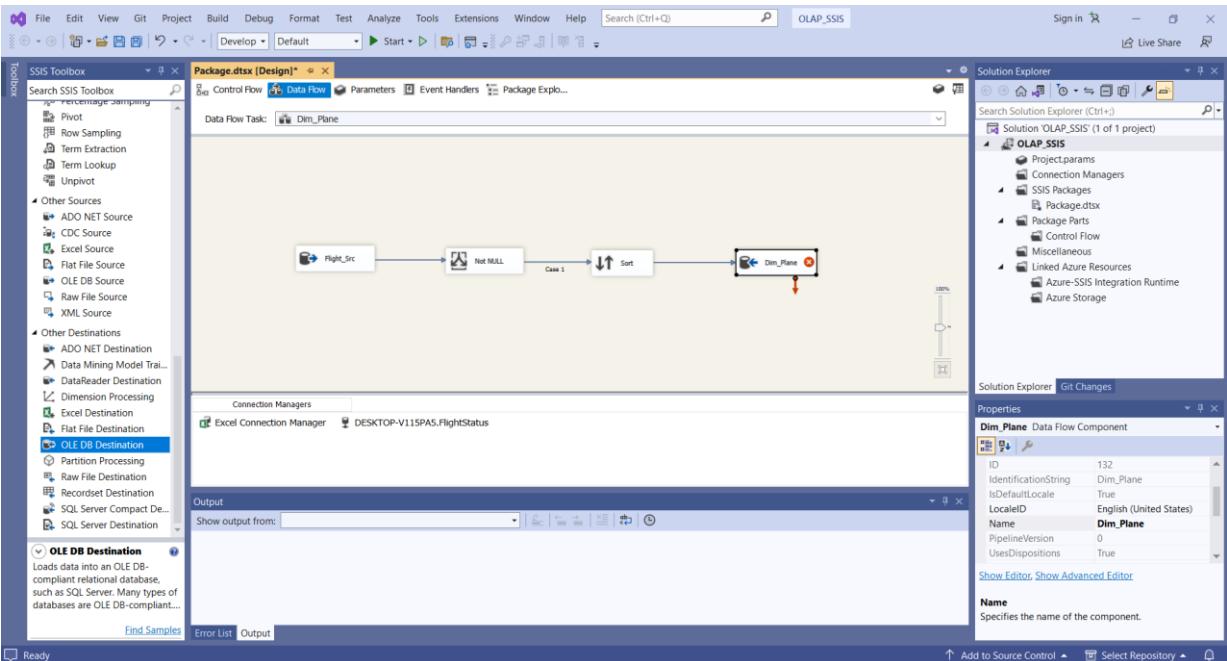
IS217 – Kho dữ liệu và OLAP

- Bước 4: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu Tail_Number bị trùng.



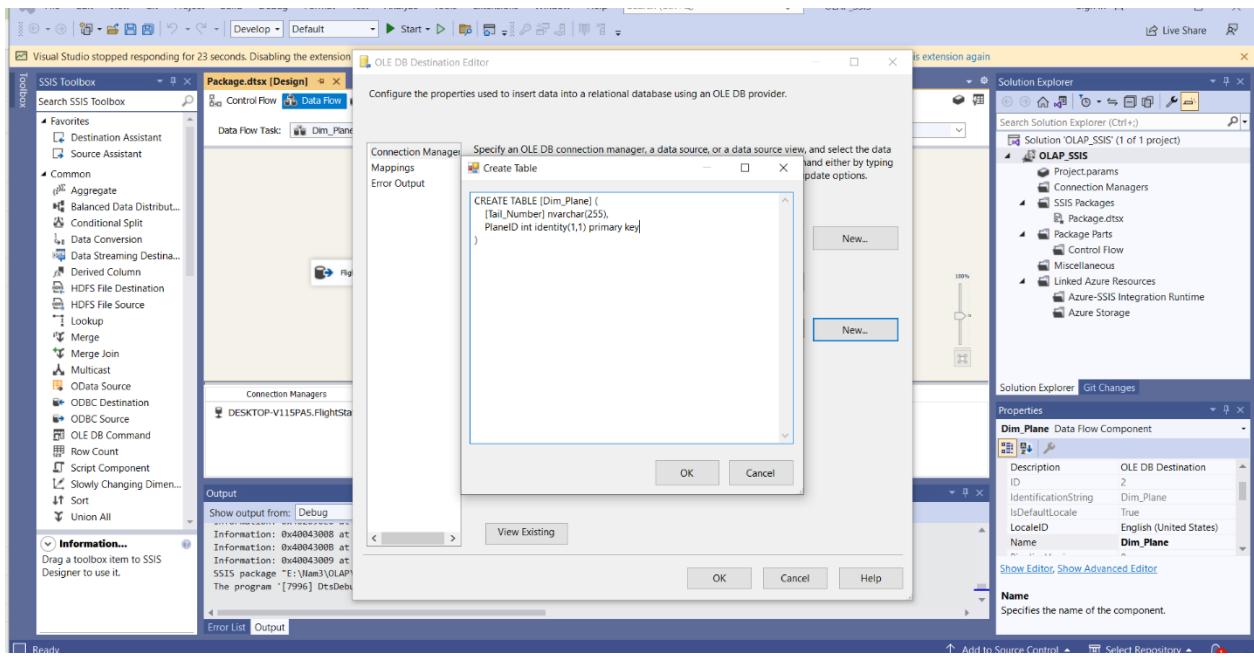
Hình 2.28. Sort dữ liệu và Remove rows

- Bước 5: Tạo 1 OLE DB Destination sau đó tạo bảng Dim_Plane.



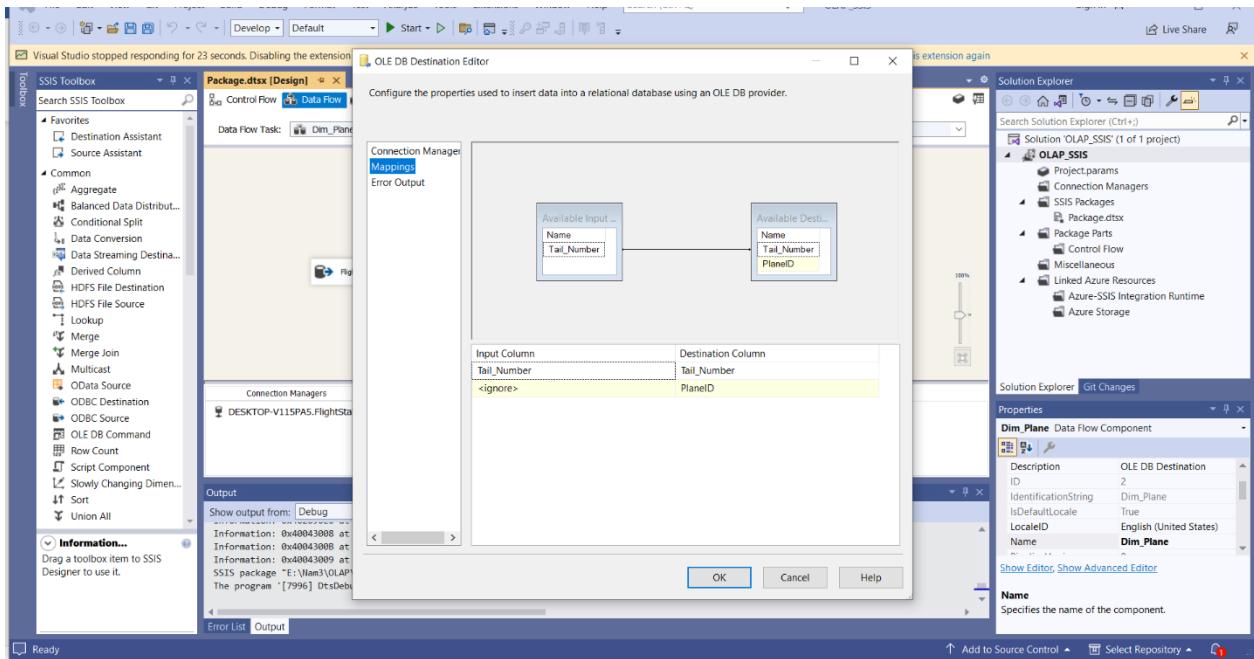
Hình 2.29. Tạo 1 OLE Destination tên Dim_Plane

IS217 – Kho dữ liệu và OLAP



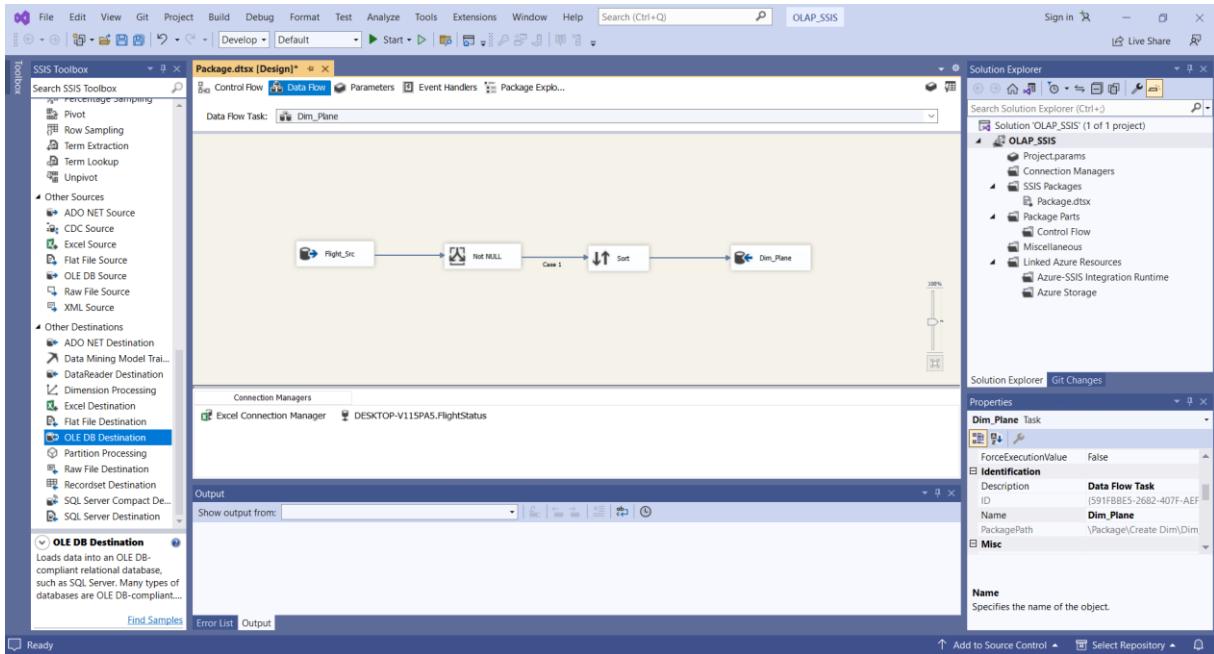
Hình 2.30. Tạo bảng Dim_Plane

- Bước 6: Qua tab Mapping để kiểm tra.



Hình 2.31. Qua Mapping để kiểm tra

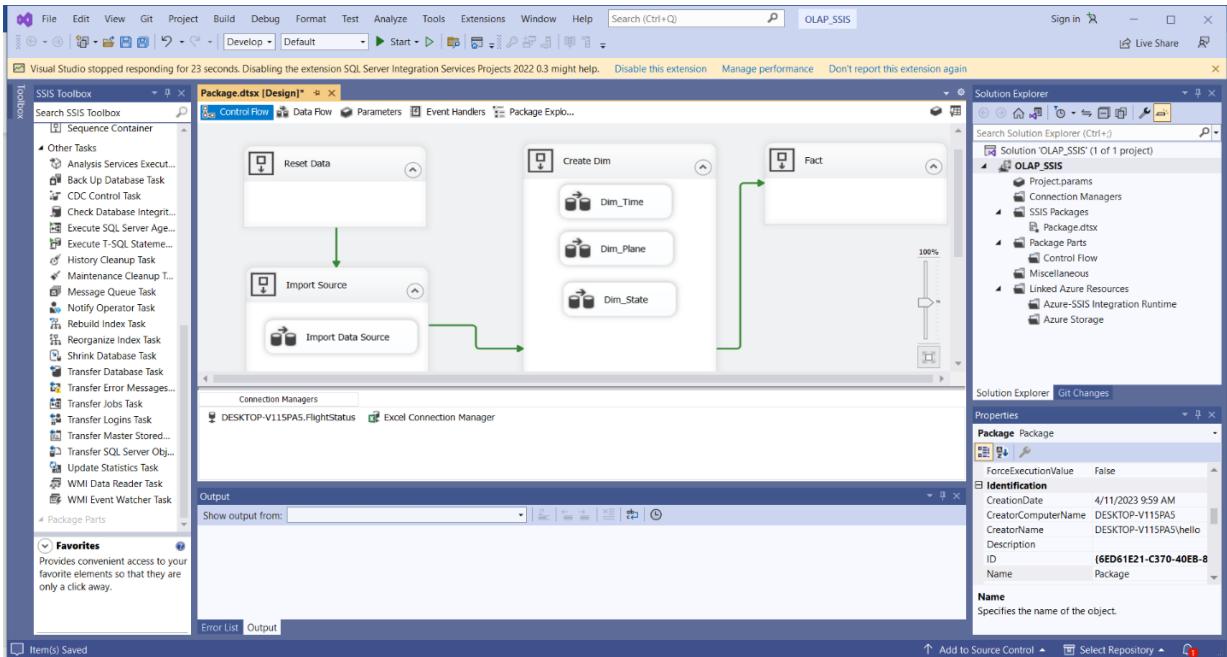
- Kết quả luồng thực hiện của bảng Dim_Plane



Hình 2.32. Luồng thực hiện của bảng Dim_Plane

2.3.6. Tạo bảng Dim_State

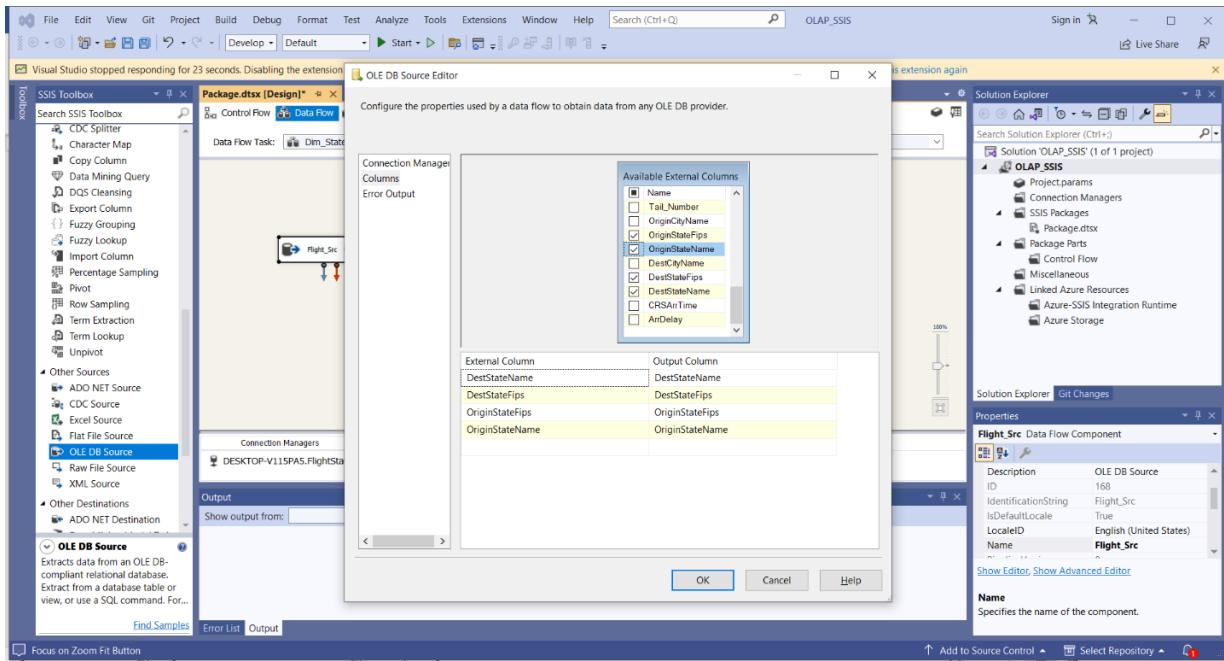
- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_State



Hình 2.33. Tạo Data Flow Task Dim_State

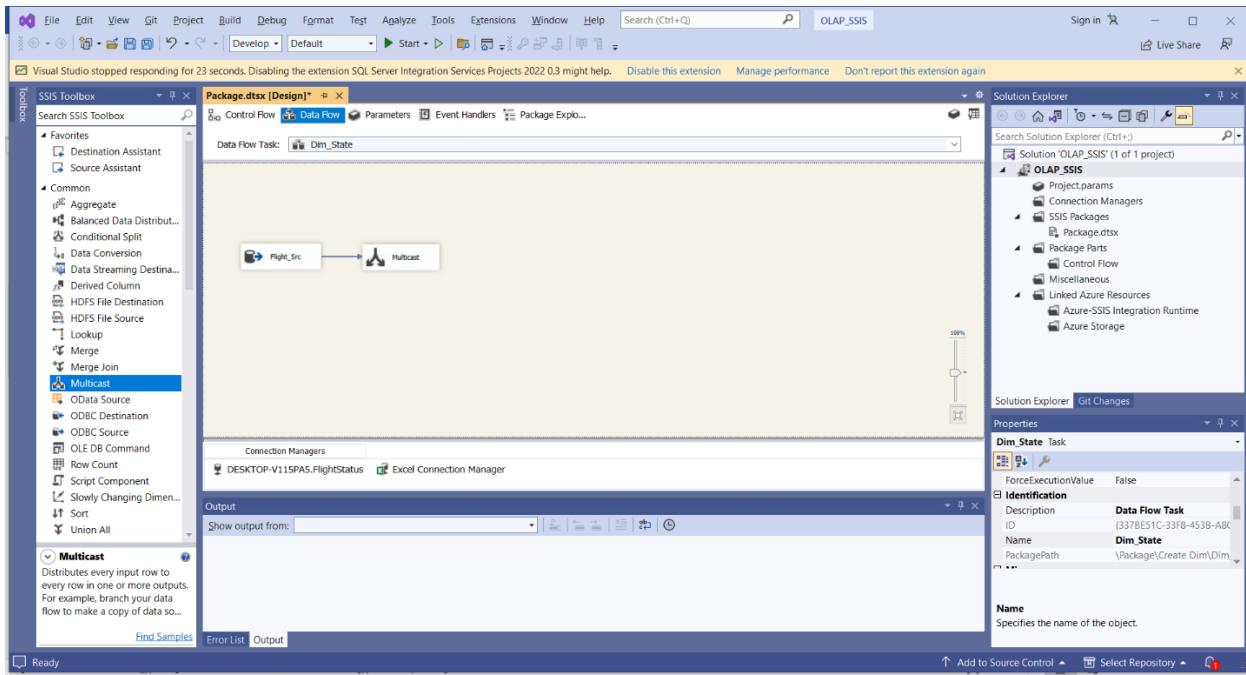
IS217 – Kho dữ liệu và OLAP

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng.



Hình 2.34. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

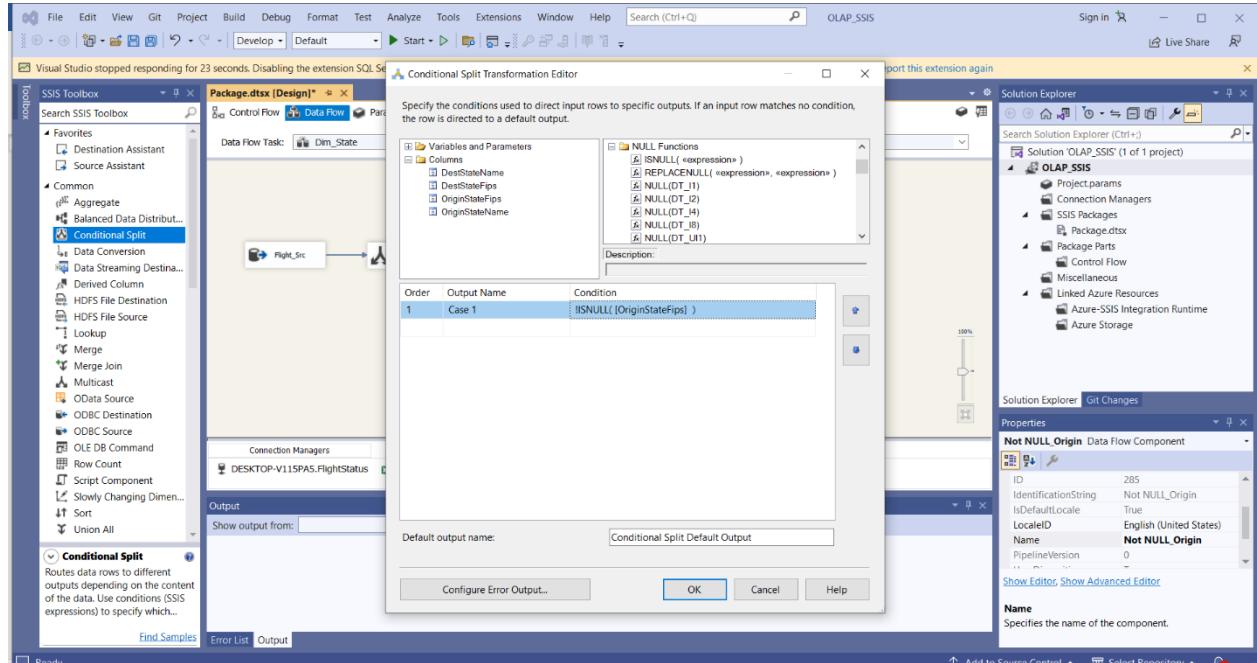
- Bước 3: Dùng công cụ Multicast để chia dữ liệu thành nhiều đối tượng đầu ra khác nhau.



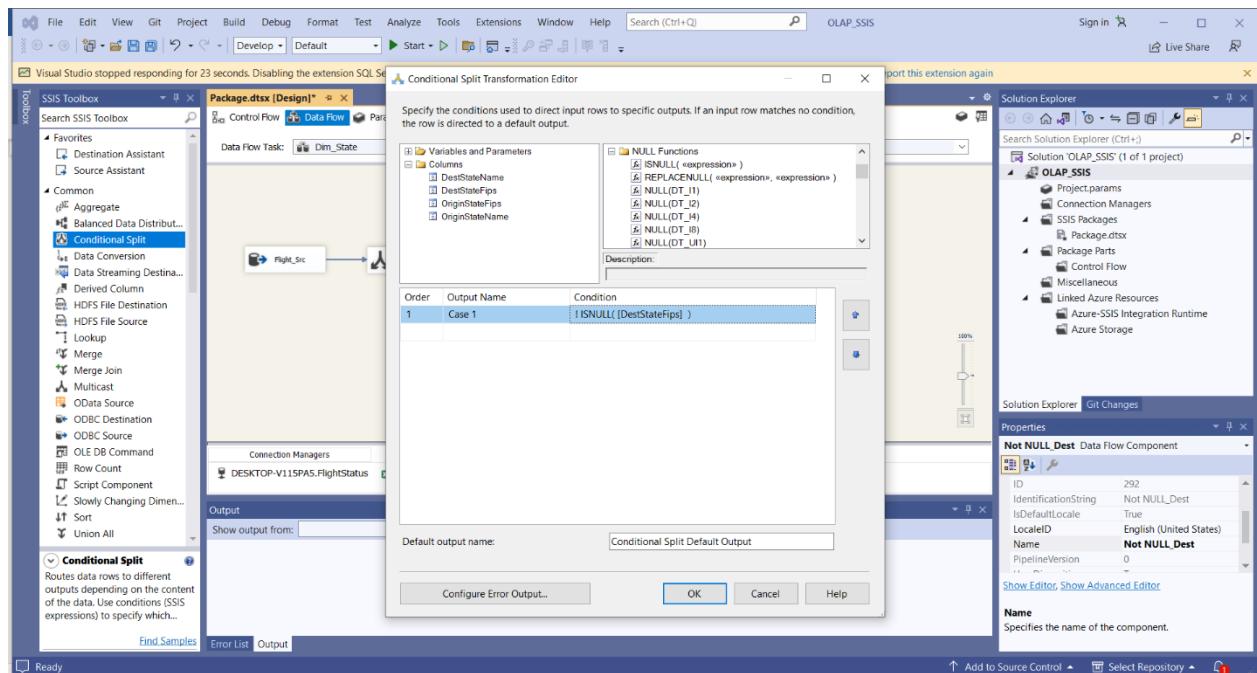
Hình 2.35. Dùng công cụ Multicast để chia tập dữ liệu

IS217 – Kho dữ liệu và OLAP

- Bước 4: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu với 2 Conditional Split lần lượt của 2 đối tượng của Origin và Dest.



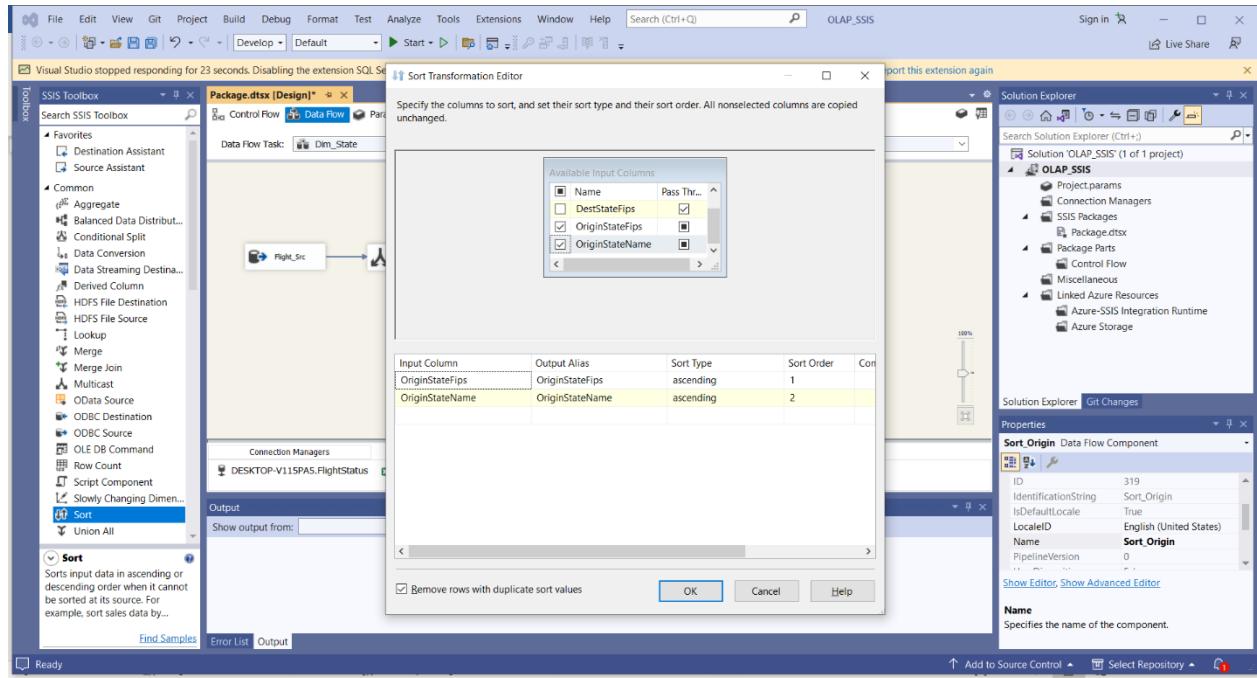
Hình 2.36. Thêm Conditional Split cho đối tượng Origin



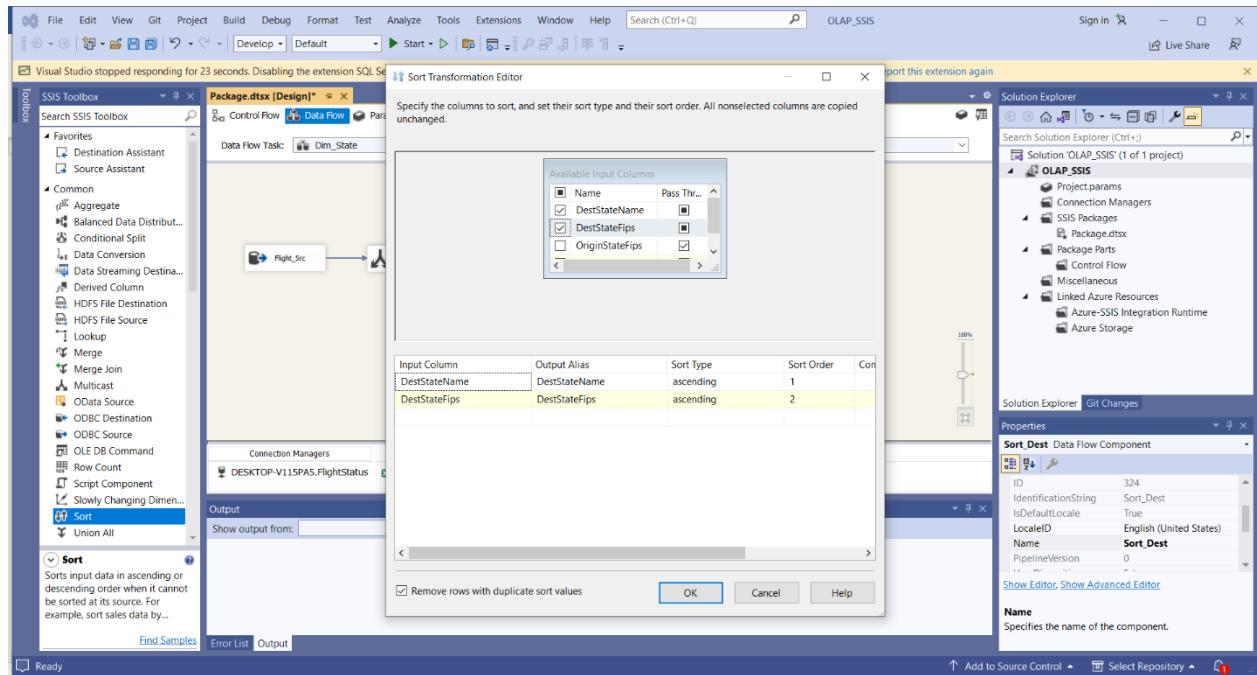
Hình 2.37. Thêm Conditional Split cho đối tượng Dest

IS217 – Kho dữ liệu và OLAP

- Bước 5: Dùng Sort để sắp xếp lại dữ liệu với 2 tập đối tượng của Origin và Dest.
- Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



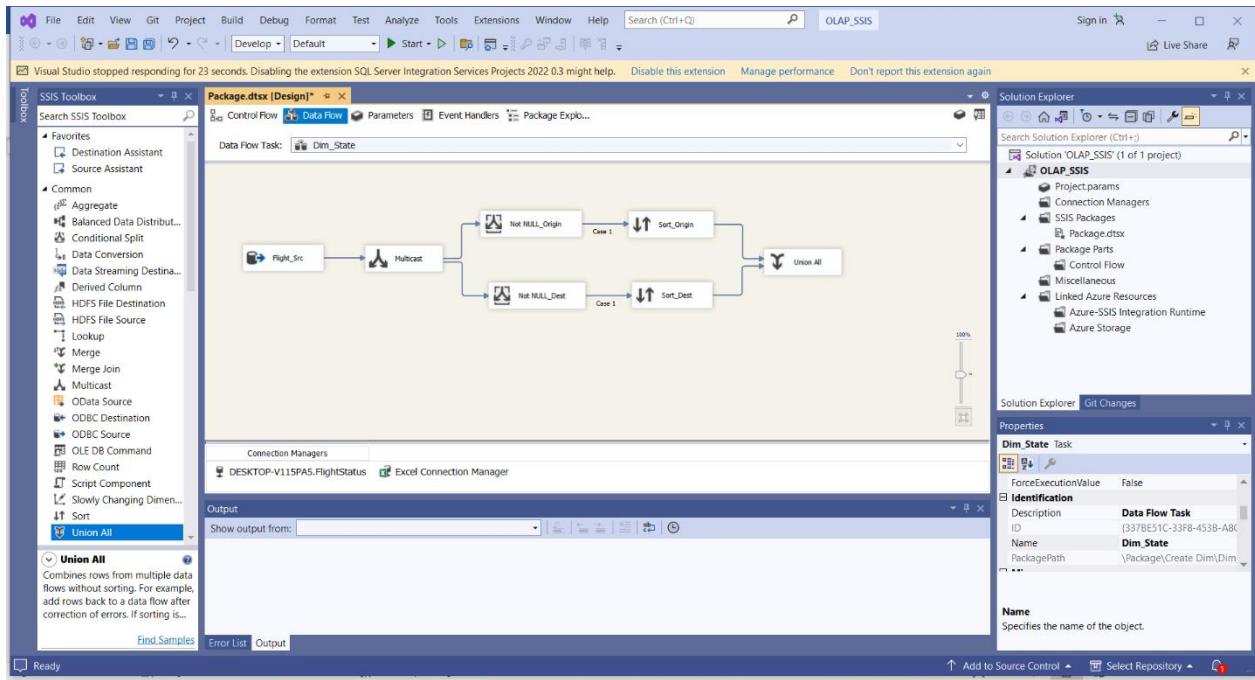
Hình 2.38. Sort dữ liệu cho đối tượng Origin và Remove rows



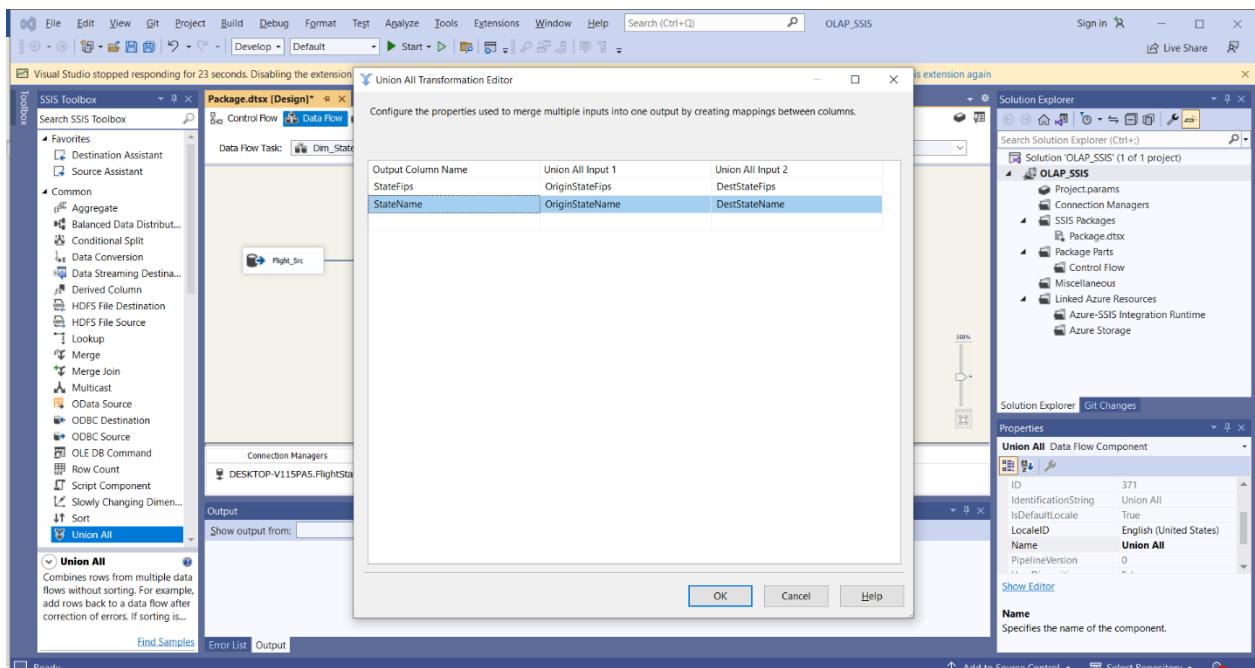
Hình 2.39. Sort dữ liệu cho đối tượng Dest và Remove rows

IS217 – Kho dữ liệu và OLAP

- Bước 6: Dùng công cụ Union All để kết hợp 2 tập đối tượng vừa chia lại với nhau (của nhóm Origin và Dest). Sau đó ta tiến hành đổi lại tên thuộc tính cho phù hợp.



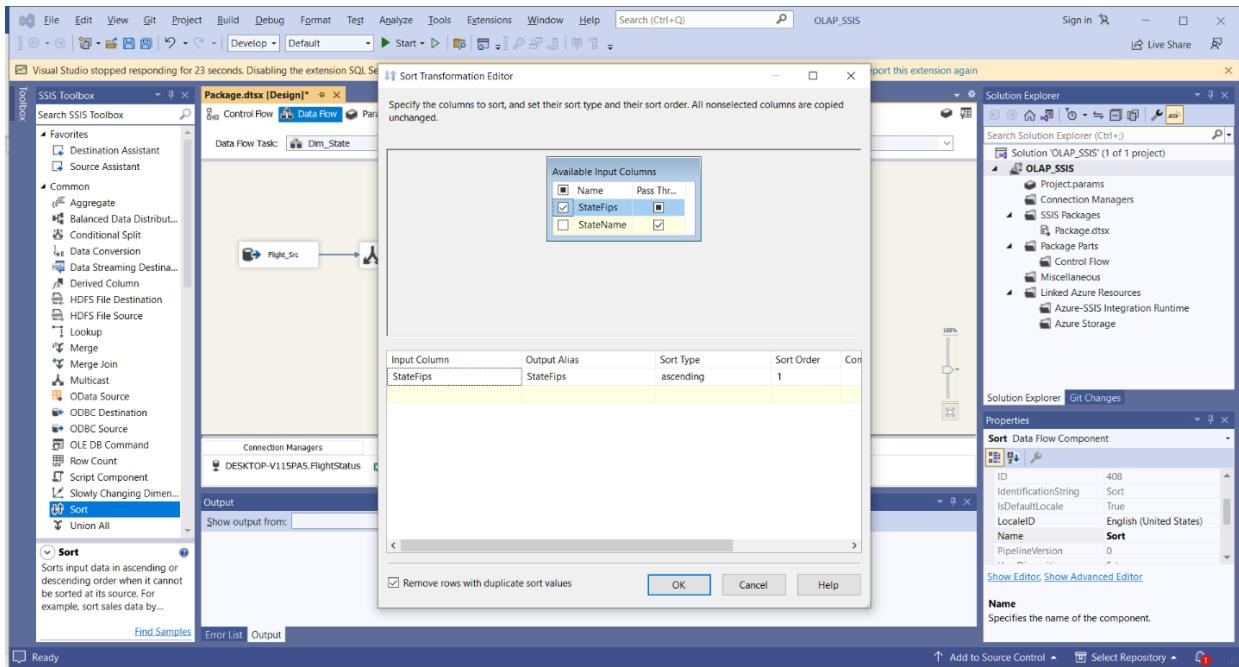
Hình 2.40. Dùng công cụ Union All để kết hợp 2 tập đối tượng Origin và Dest



Hình 2.41. Kết 2 đối tượng và đổi tên thuộc tính

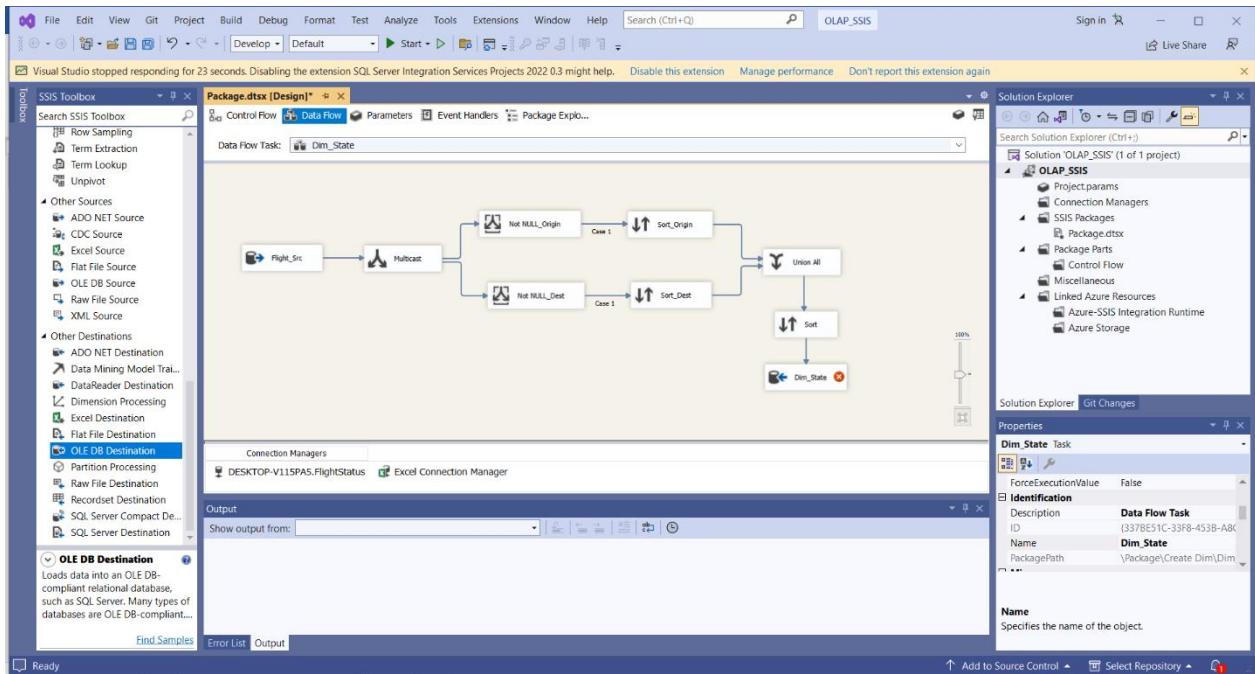
IS217 – Kho dữ liệu và OLAP

- Bước 7: Dùng Sort để sắp xếp dữ liệu dữ liệu vừa kết hợp. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



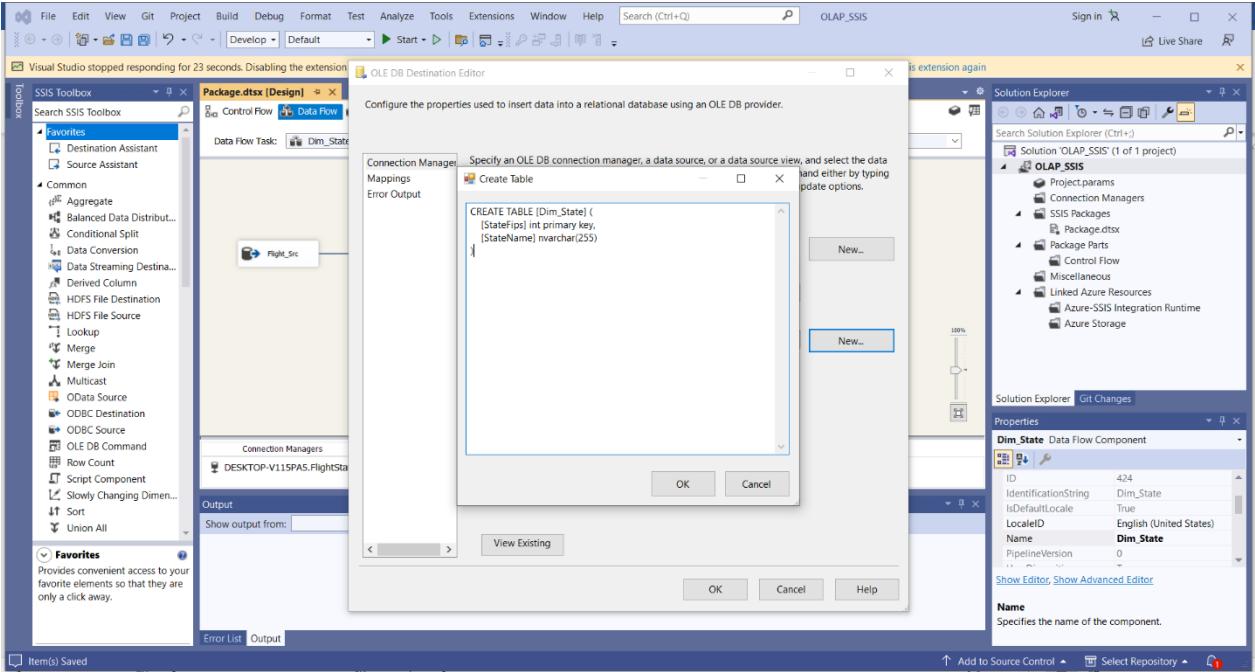
Hình 2.42. Sort dữ liệu và Remove rows

- Bước 8: Tạo 1 OLE Destination sau đó tạo bảng Dim_State.



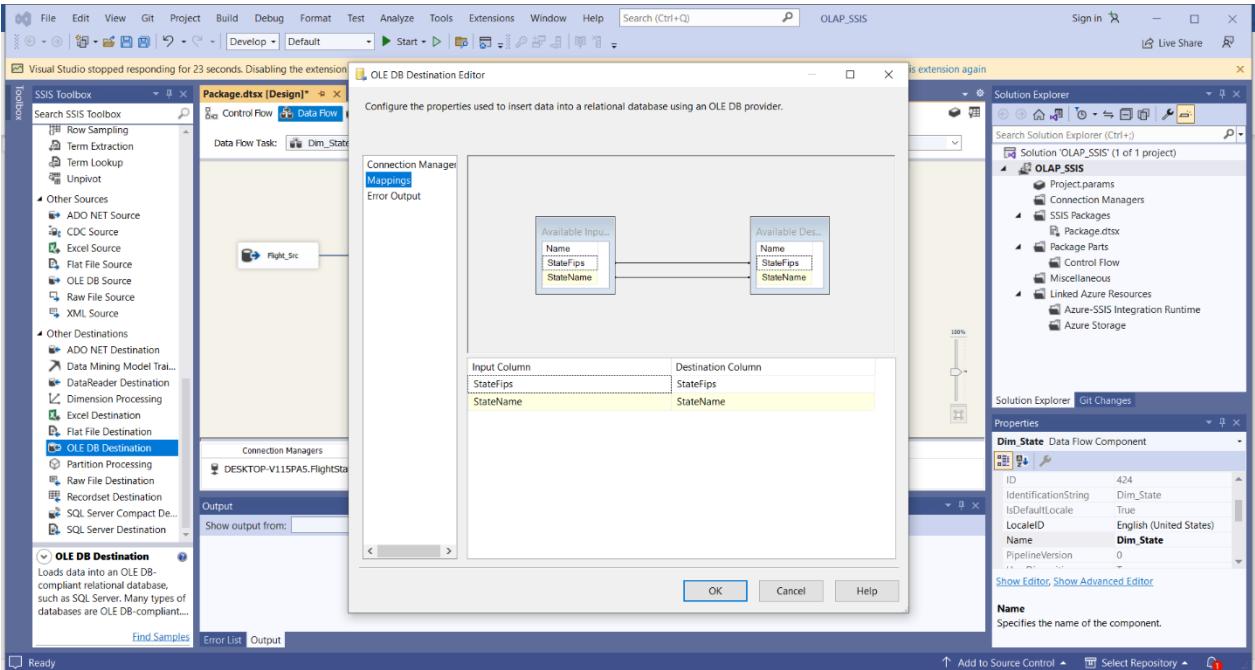
Hình 2.43. Tạo 1 OLE Destination tên Dim_State

IS217 – Kho dữ liệu và OLAP



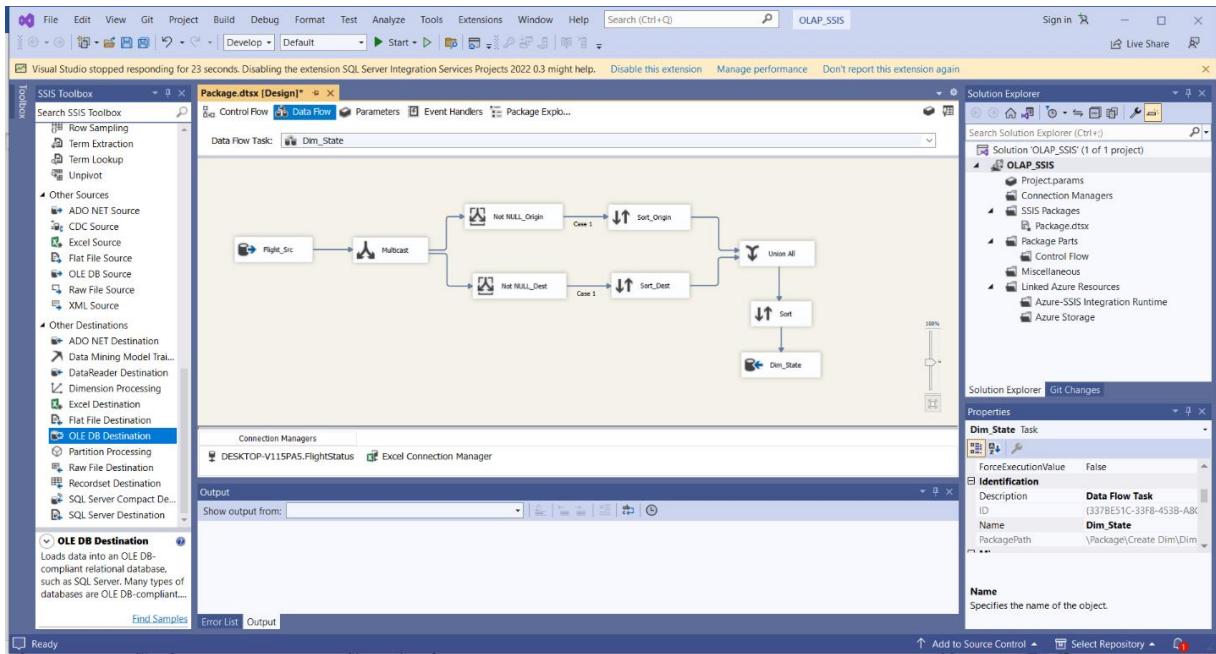
Hình 2.44. Tạo bảng Dim_State

- Bước 9: Qua tab Mapping để kiểm tra.



Hình 2.45. Qua Mapping để kiểm tra

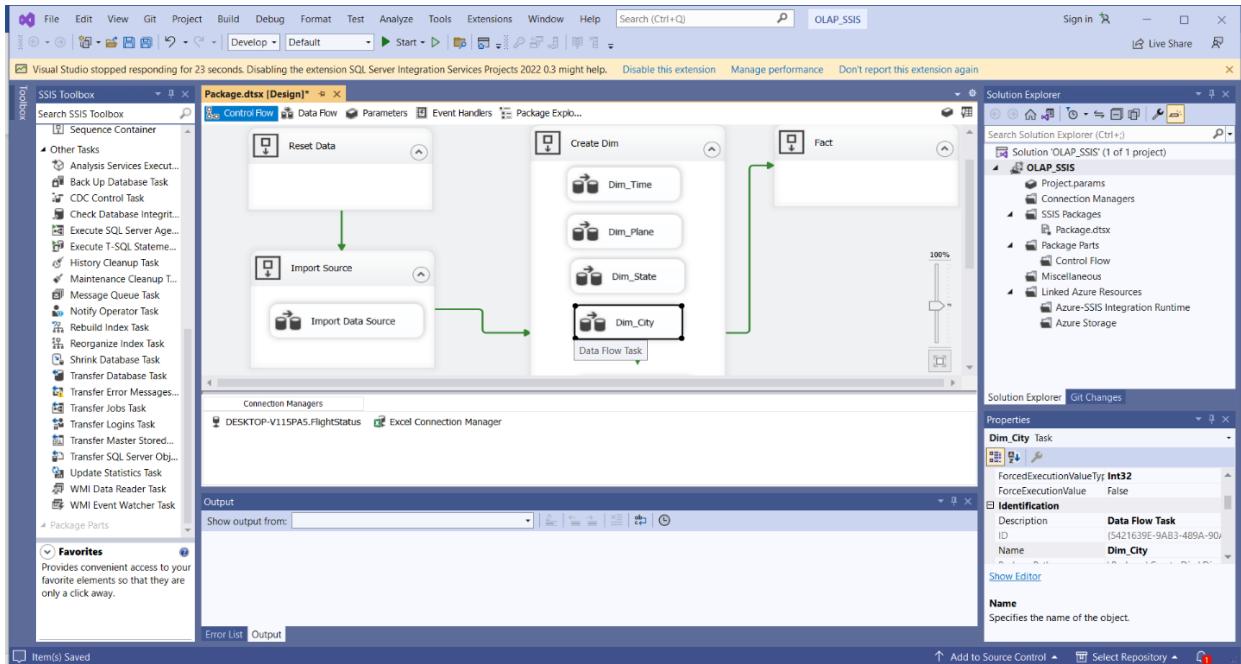
- Kết quả luồng thực hiện của bảng Dim_State



Hình 2.46. Luồng thực hiện của bảng Dim_State

2.3.7. Tạo bảng Dim_City

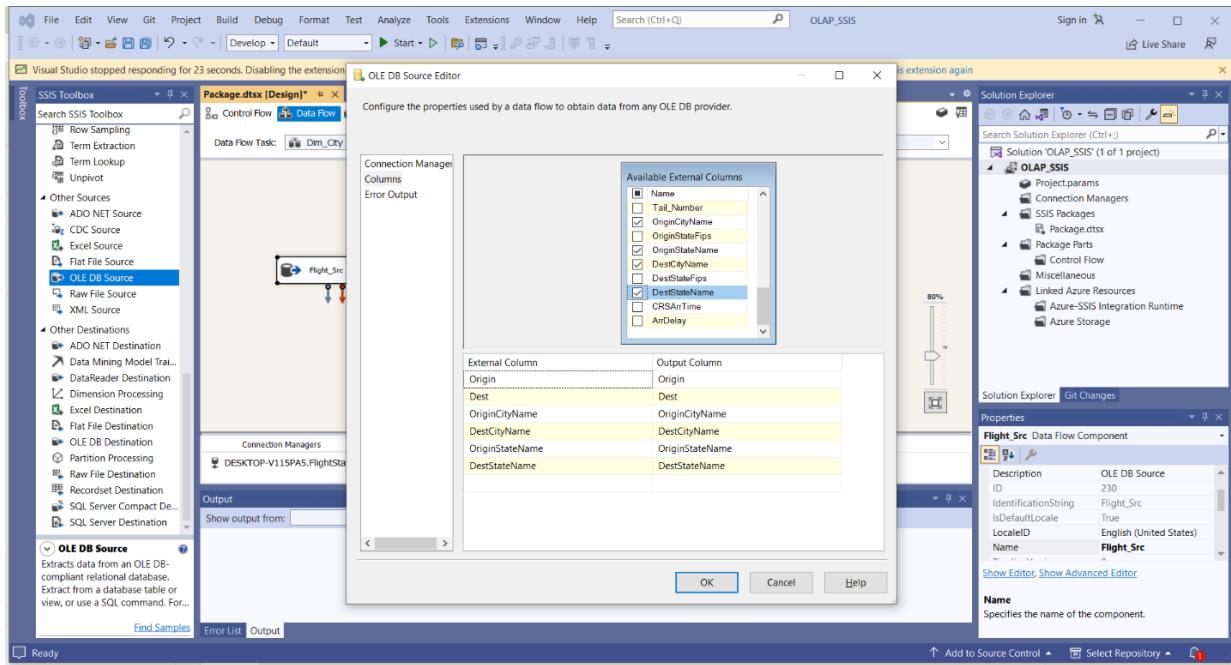
- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_City



Hình 2.47. Tạo Data Flow Task Dim_City

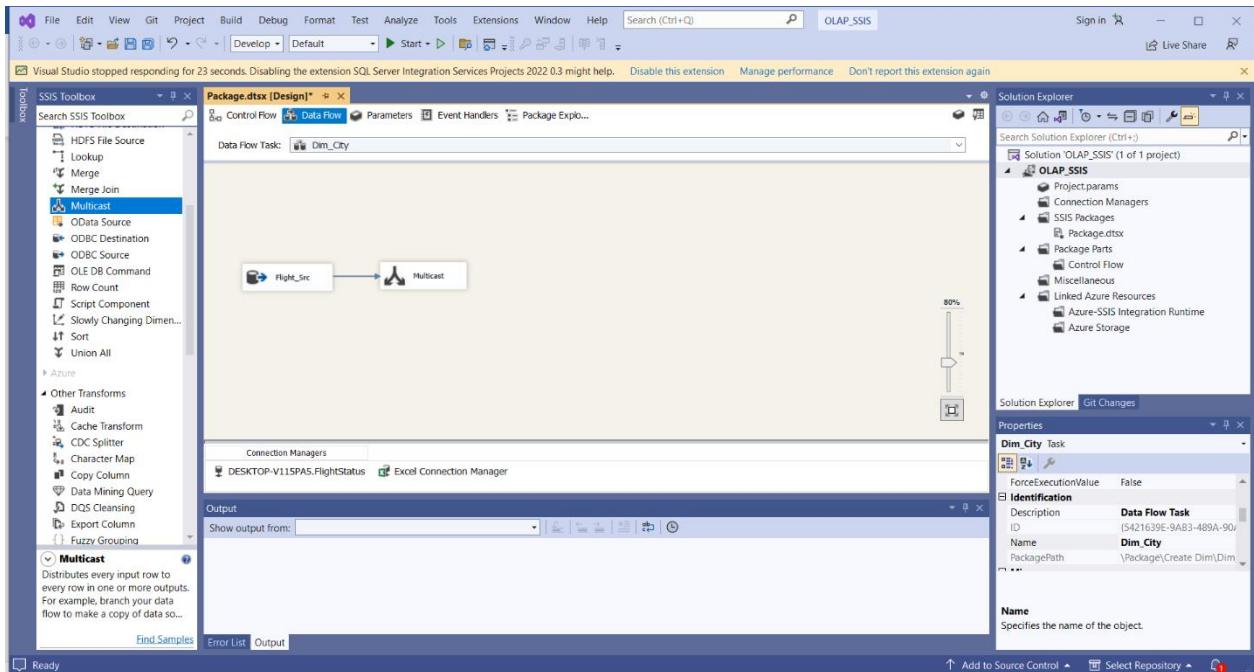
IS217 – Kho dữ liệu và OLAP

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng.



Hình 2.48. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

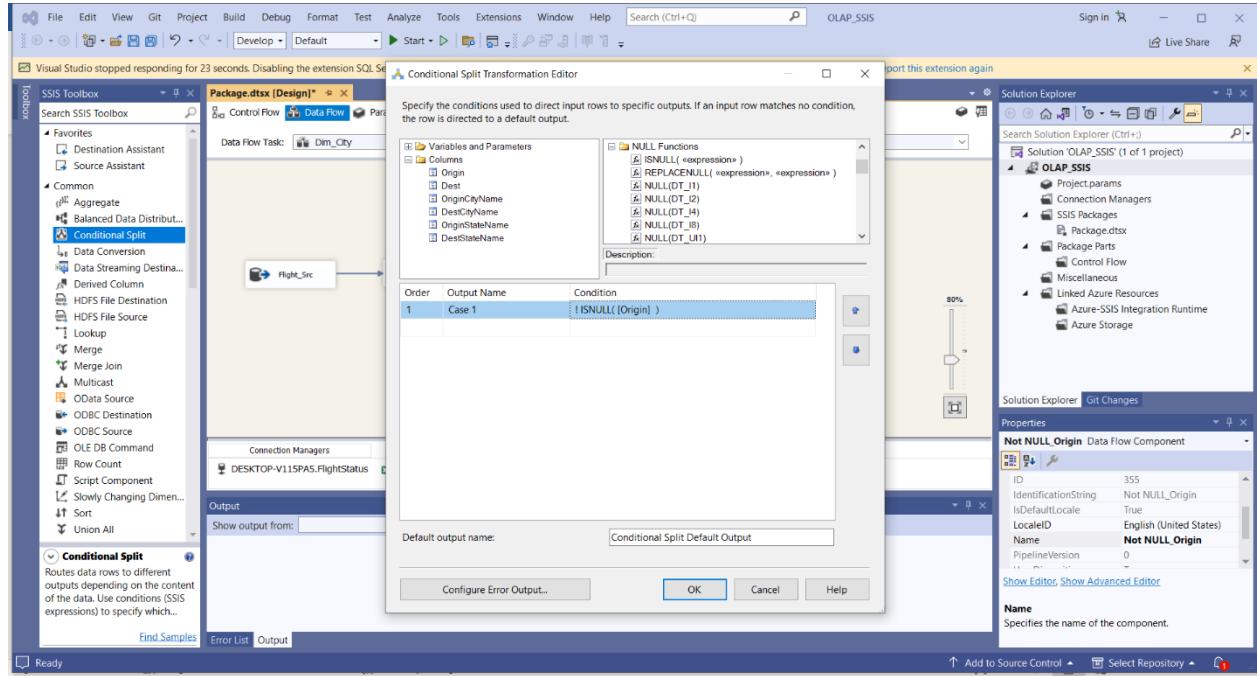
- Bước 3: Dùng công cụ Multicast để chia dữ liệu thành nhiều đối tượng đầu ra khác nhau.



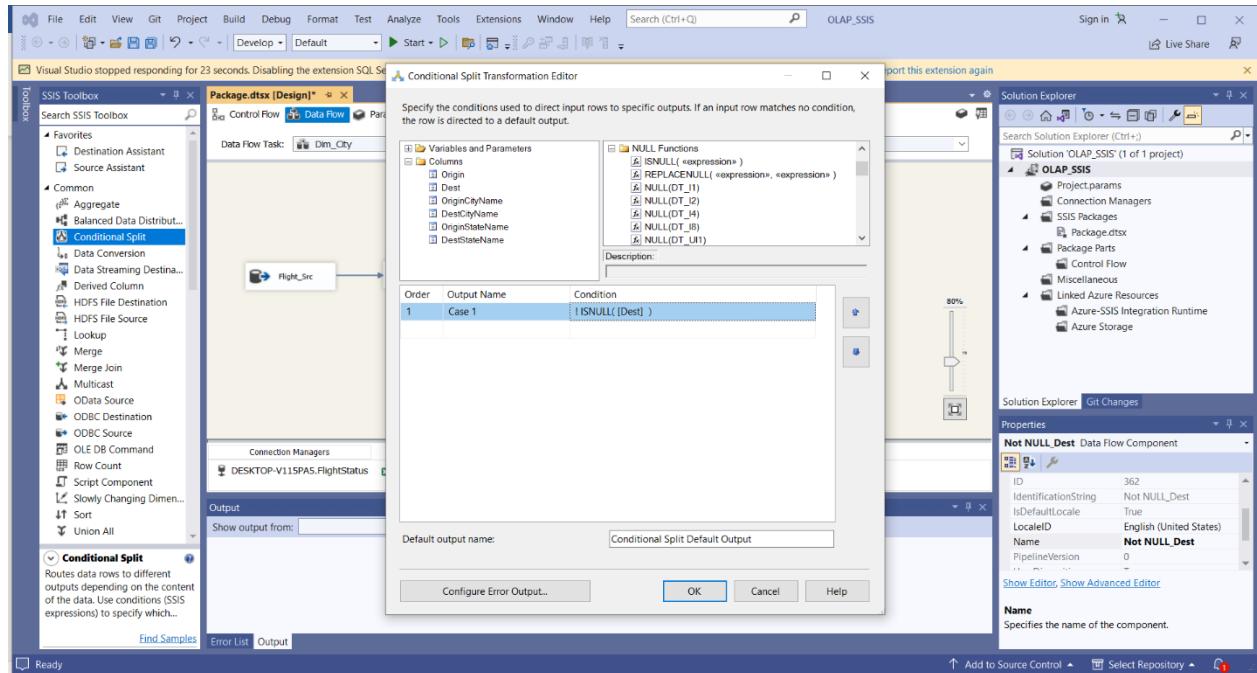
Hình 2.49. Dùng công cụ Multicast để chia dữ liệu thành nhiều đối tượng

IS217 – Kho dữ liệu và OLAP

- Bước 4: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu với 2 Conditional Split lần lượt của 2 đối tượng của Origin và Dest.



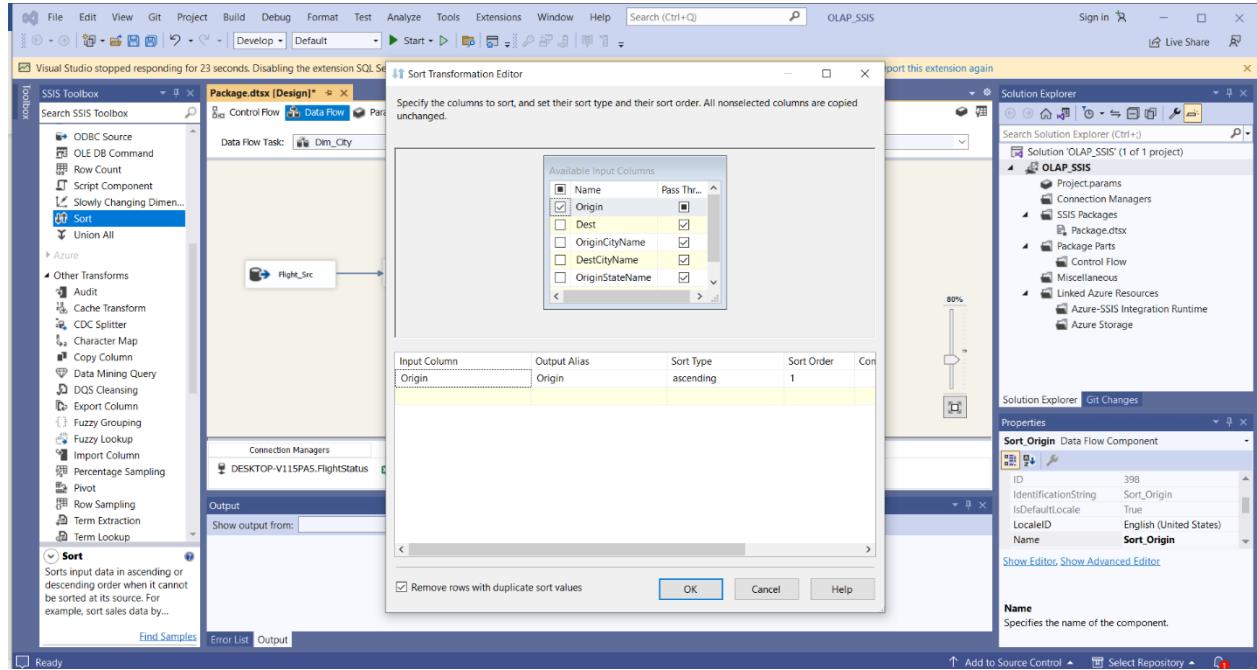
Hình 2.50. Thêm Conditional Split cho đối tượng Origin



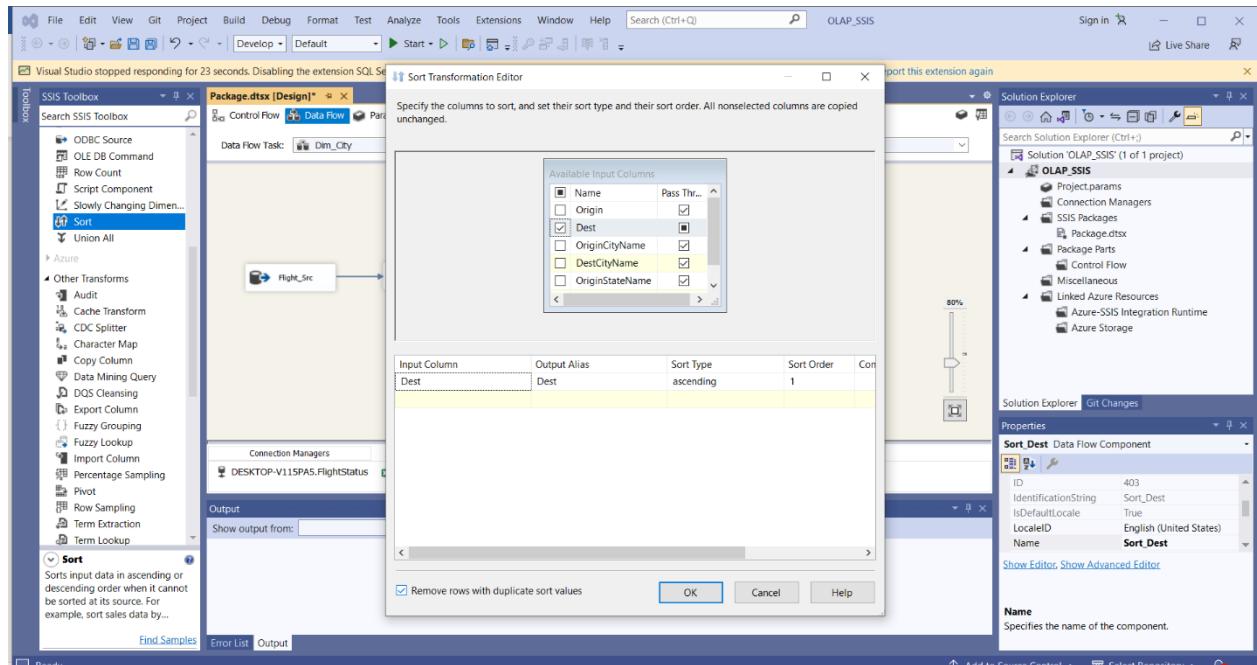
Hình 2.51. Thêm Conditional Split cho đối tượng Dest

IS217 – Kho dữ liệu và OLAP

- Bước 5: Dùng Sort để sắp xếp lại dữ liệu với 2 tập đối tượng của Origin và Dest. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



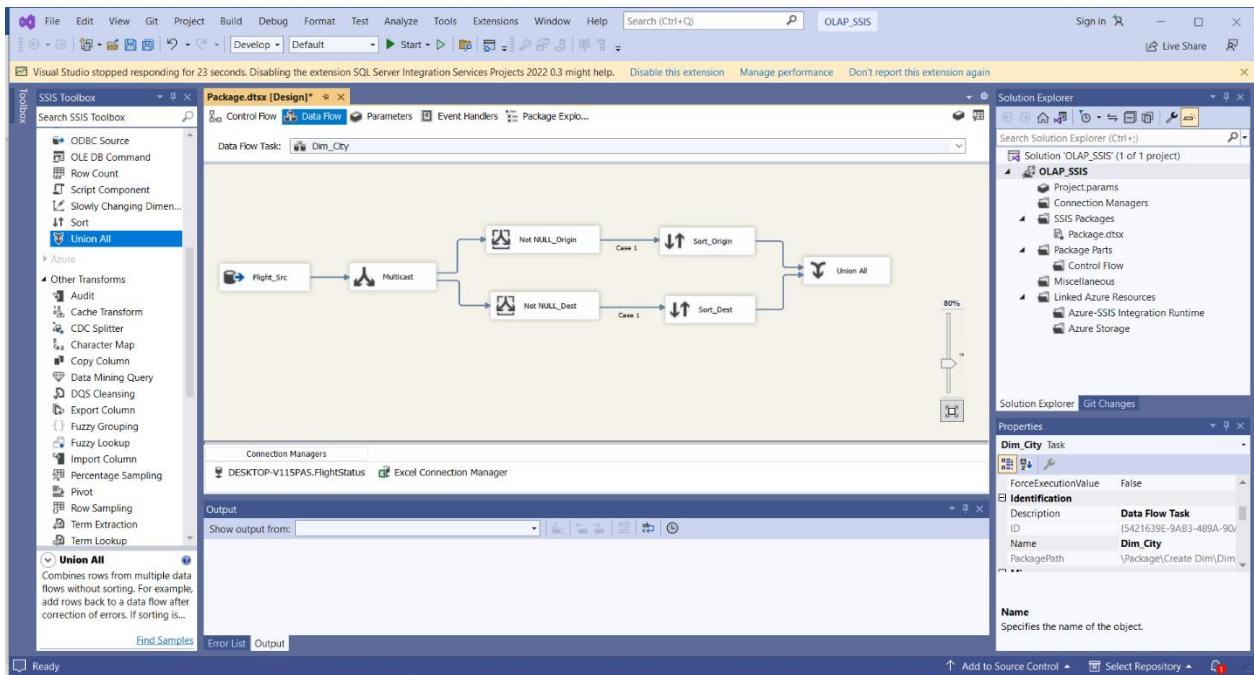
Hình 2.52. Sort dữ liệu cho đối tượng Origin và Remove rows



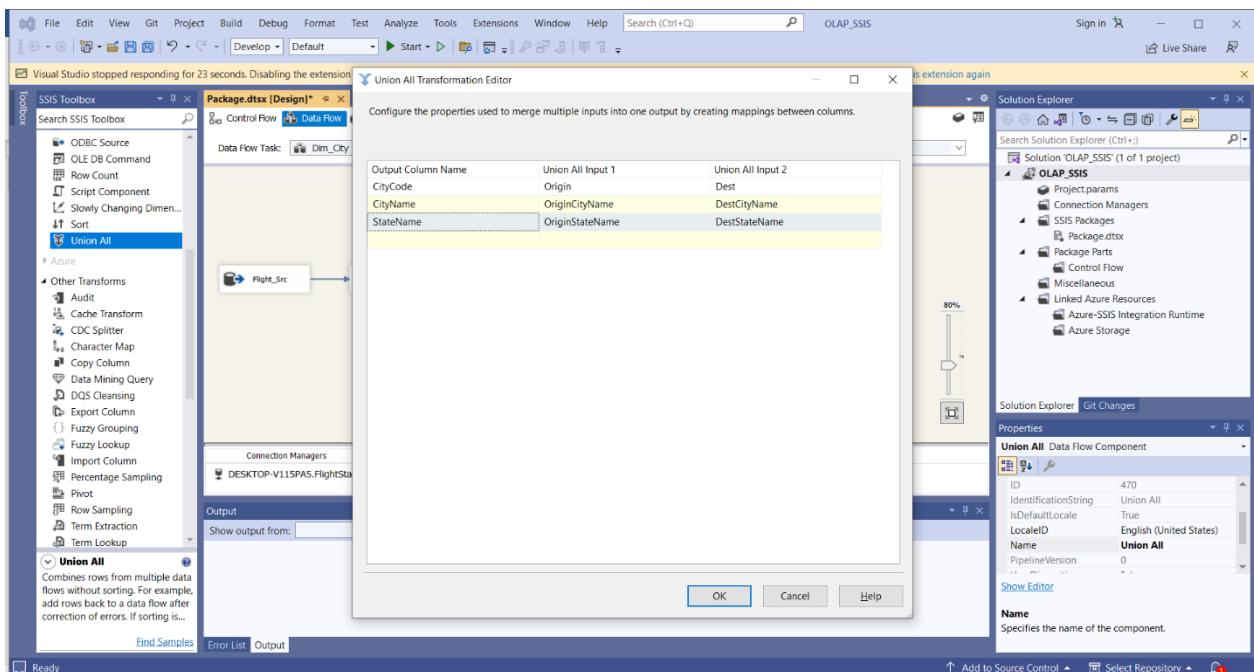
Hình 2.53. Sort dữ liệu cho Dest và Remove rows

IS217 – Kho dữ liệu và OLAP

- Bước 6: Dùng công cụ Union All để kết hợp 2 tập đối tượng vừa chia lại với nhau (của nhóm Origin và Dest). Sau đó ta tiến hành đổi lại tên thuộc tính cho phù hợp.



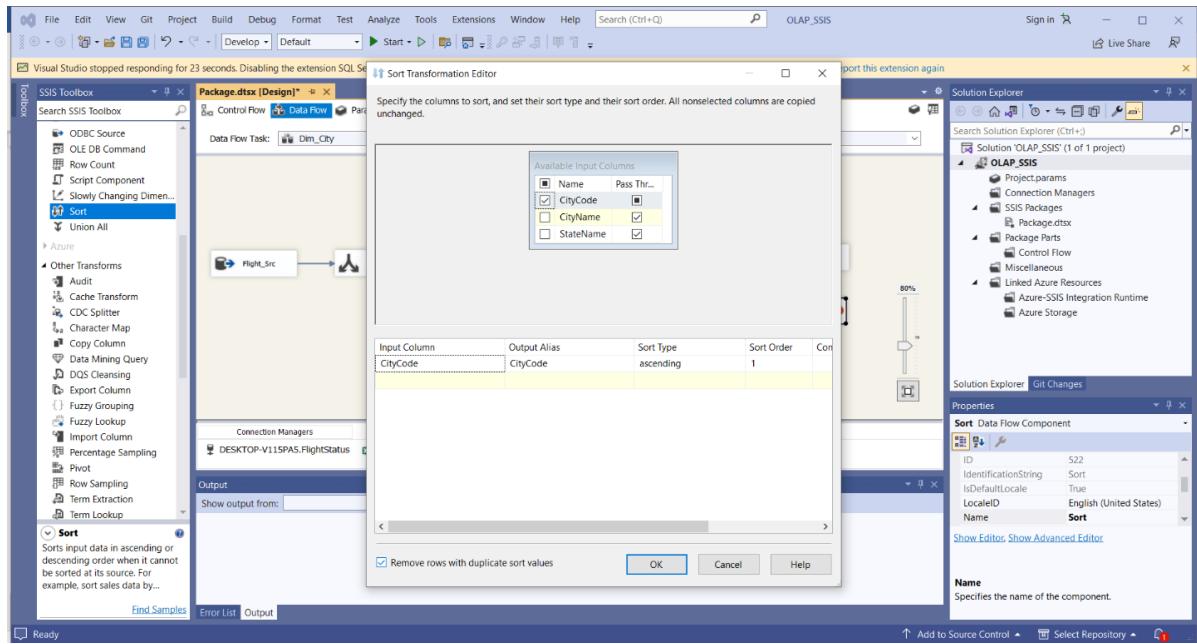
Hình 2.54. Dùng công cụ Union All để kết hợp 2 tập đối tượng Origin và Dest



Hình 2.55. Kết 2 đối tượng và đổi tên thuộc tính

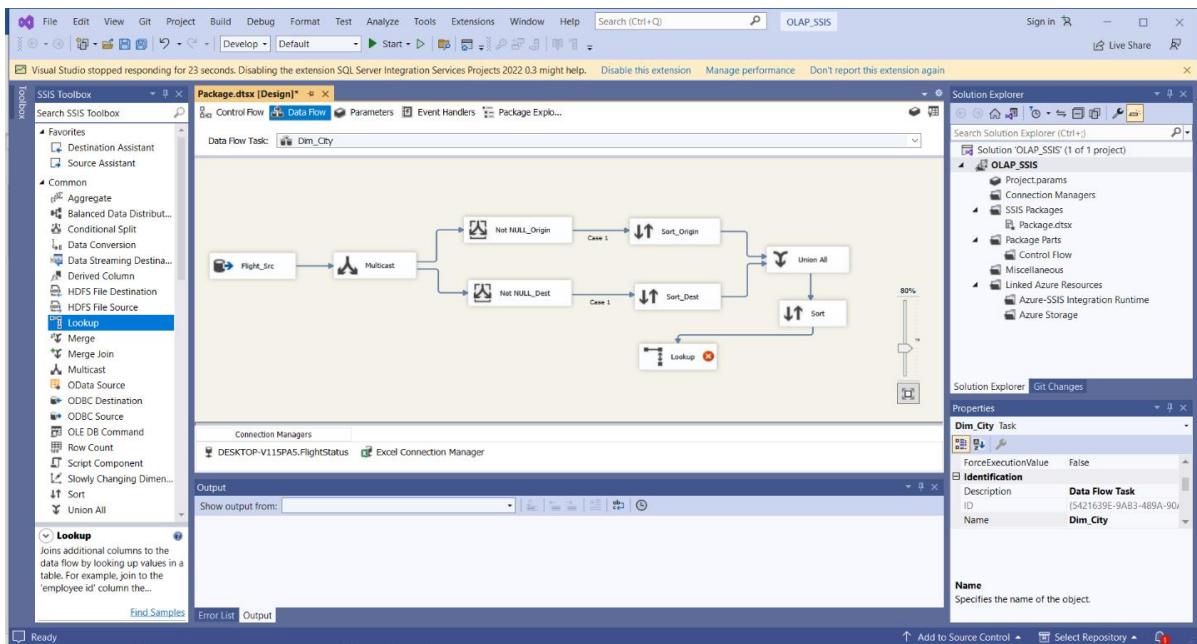
IS217 – Kho dữ liệu và OLAP

- Bước 7: Dùng Sort để sắp xếp dữ liệu dữ liệu vừa kết hợp. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



Hình 2.56. Sort dữ liệu và Remove rows

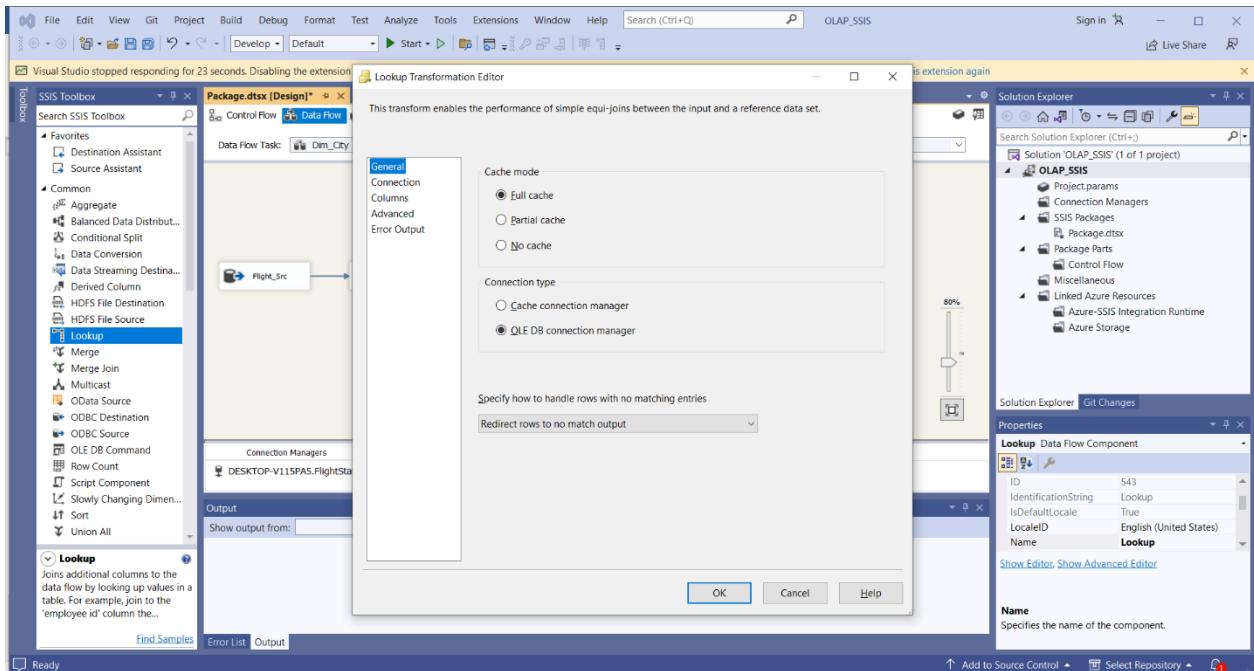
- Bước 8: Dùng Lookup để kết dữ liệu trên với bảng Dim_State thông qua thuộc tính StateName.



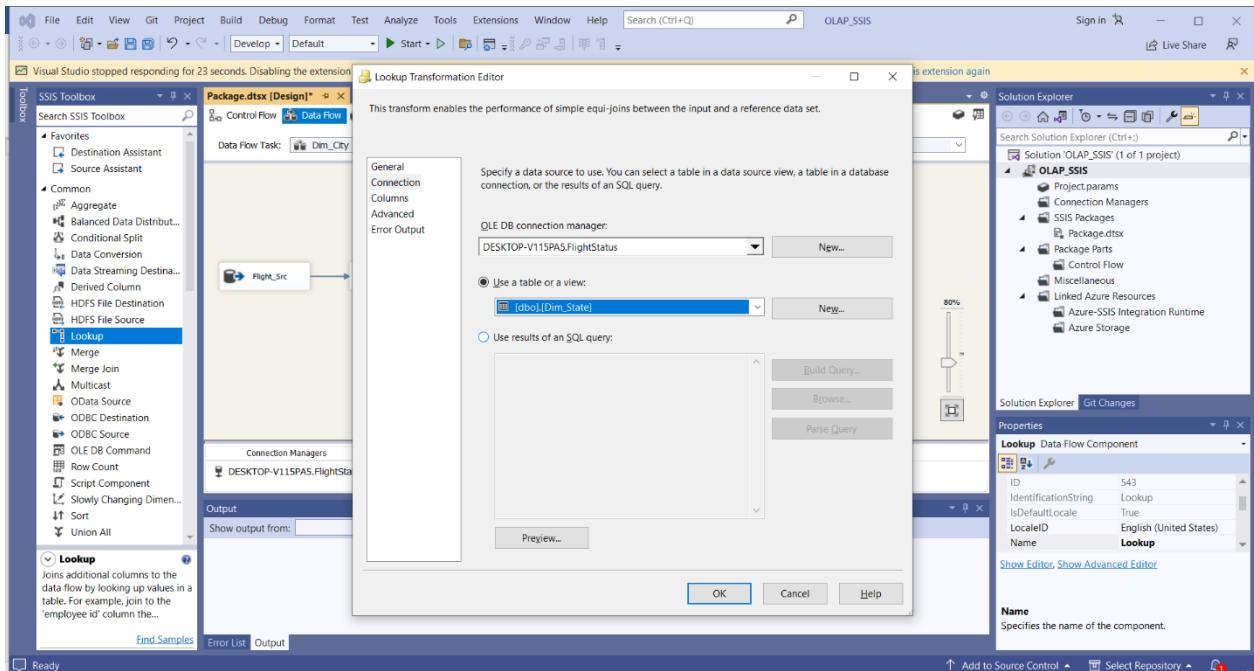
Hình 2.57. Tạo Lookup để kết dữ liệu trên với bảng Dim_State

IS217 – Kho dữ liệu và OLAP

Chú ý là ở hộp thoại Specify how to handle... chọn “Redirect rows to no match output”.

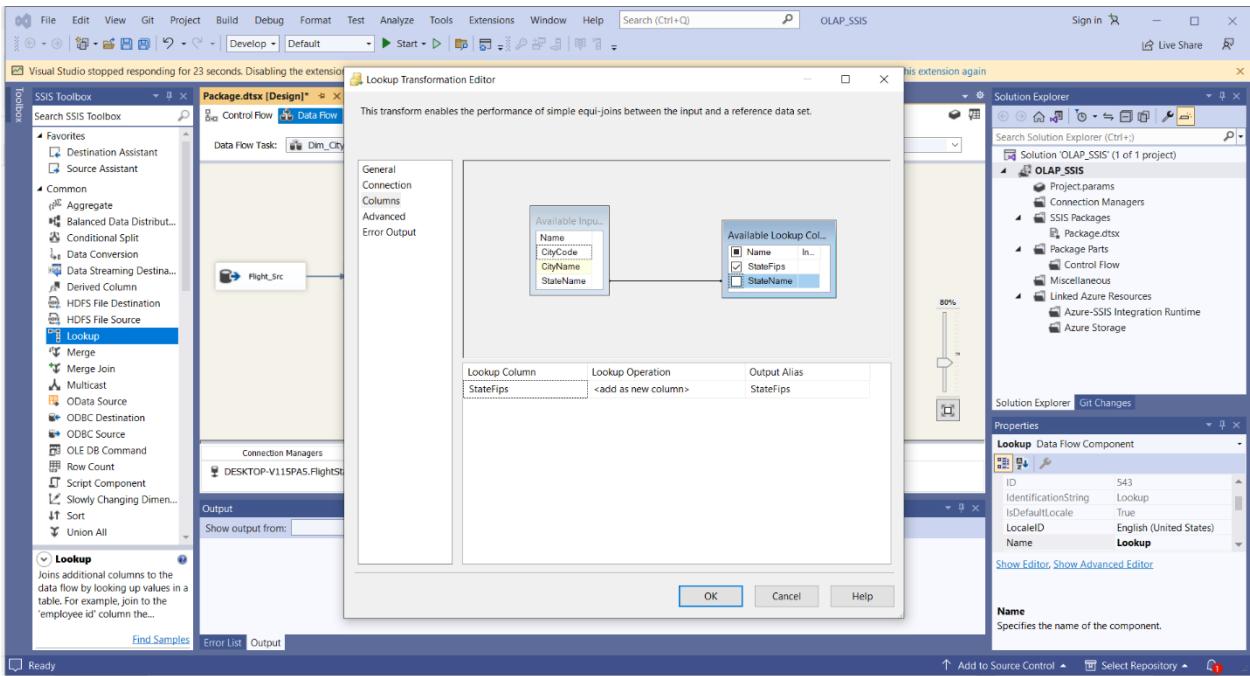


Hình 2.58. Cài đặt Lookup



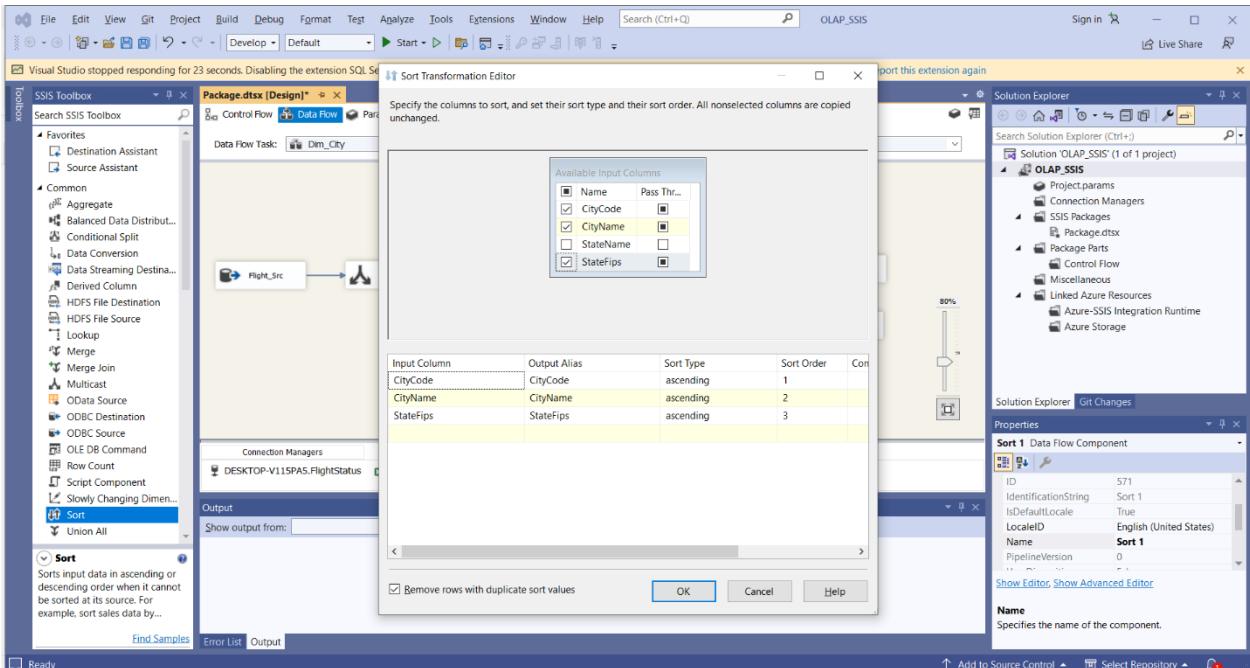
Hình 2.59. Chọn bảng Dim_State

IS217 – Kho dữ liệu và OLAP



Hình 2.60. Kết nối Source và bảng Dim_State

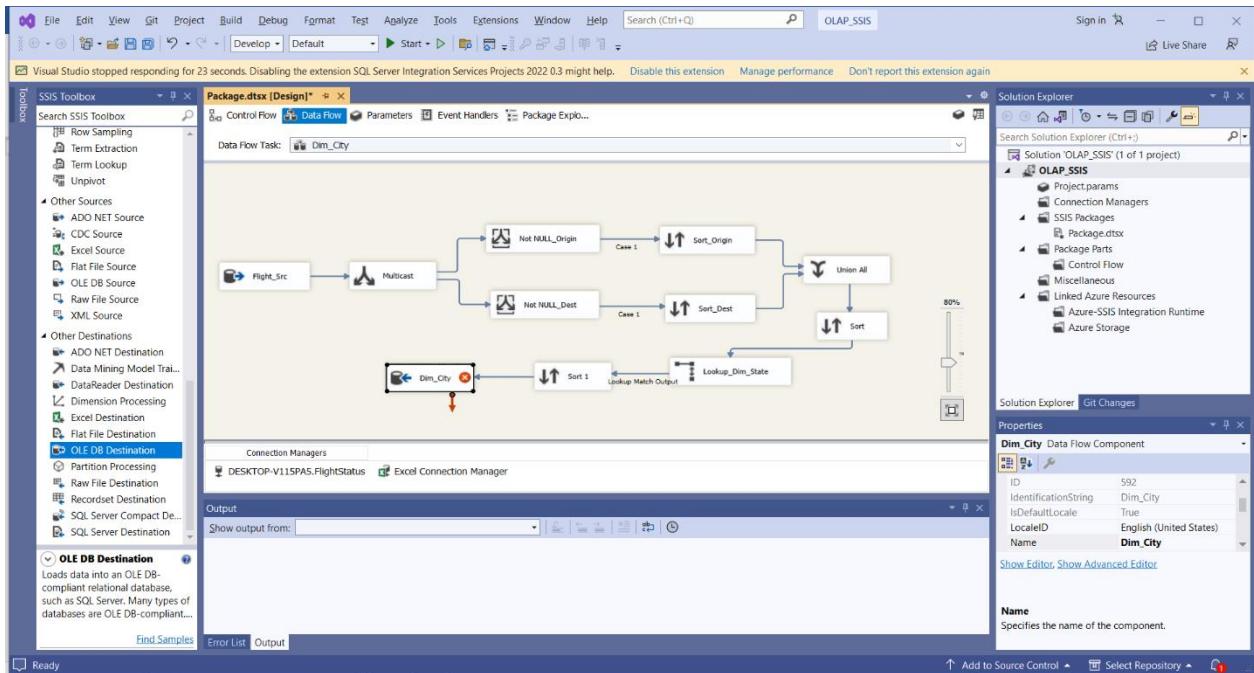
- Bước 9: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



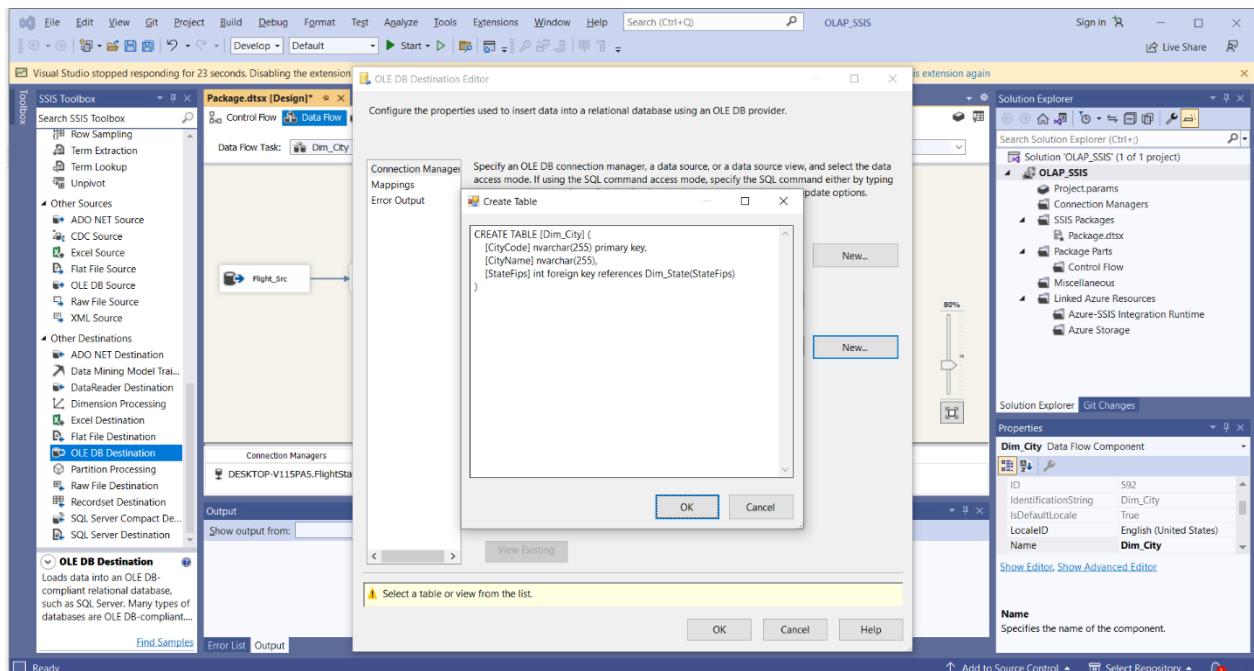
Hình 2.61. Sort dữ liệu và Remove rows

IS217 – Kho dữ liệu và OLAP

- Bước 10: Tạo 1 OLE Destination sau đó tạo bảng Dim_City.

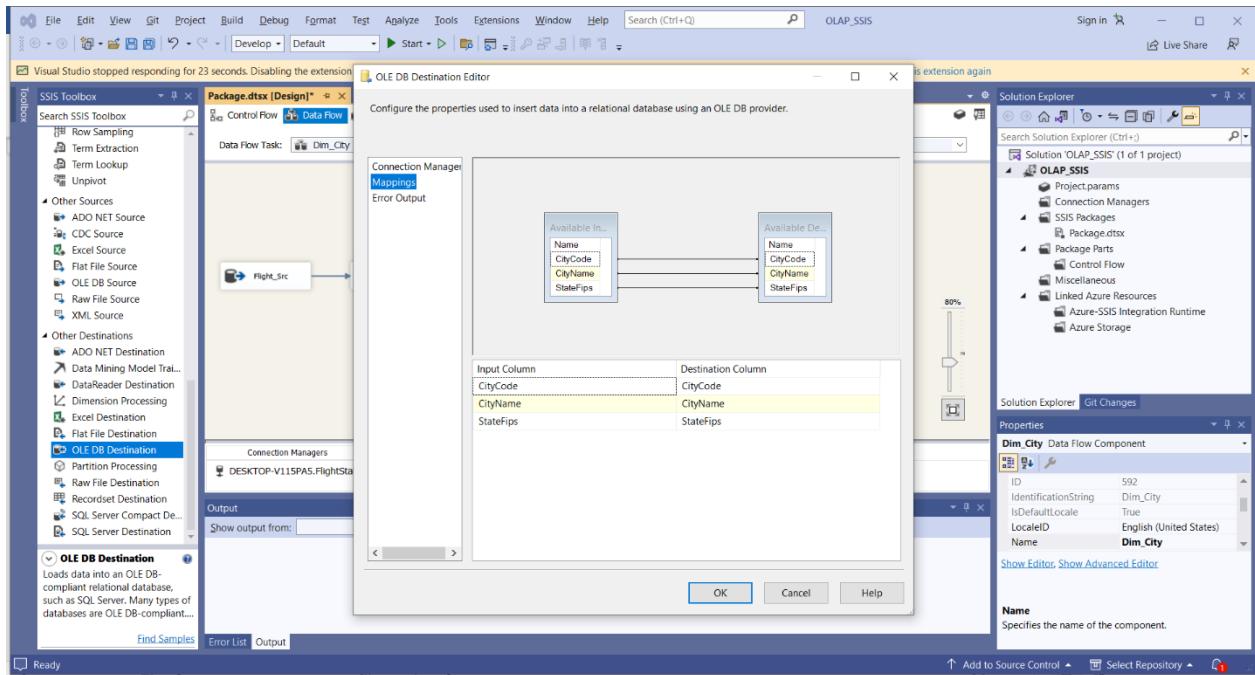


Hình 2.62. Tạo 1 OLE Destination tên Dim_City



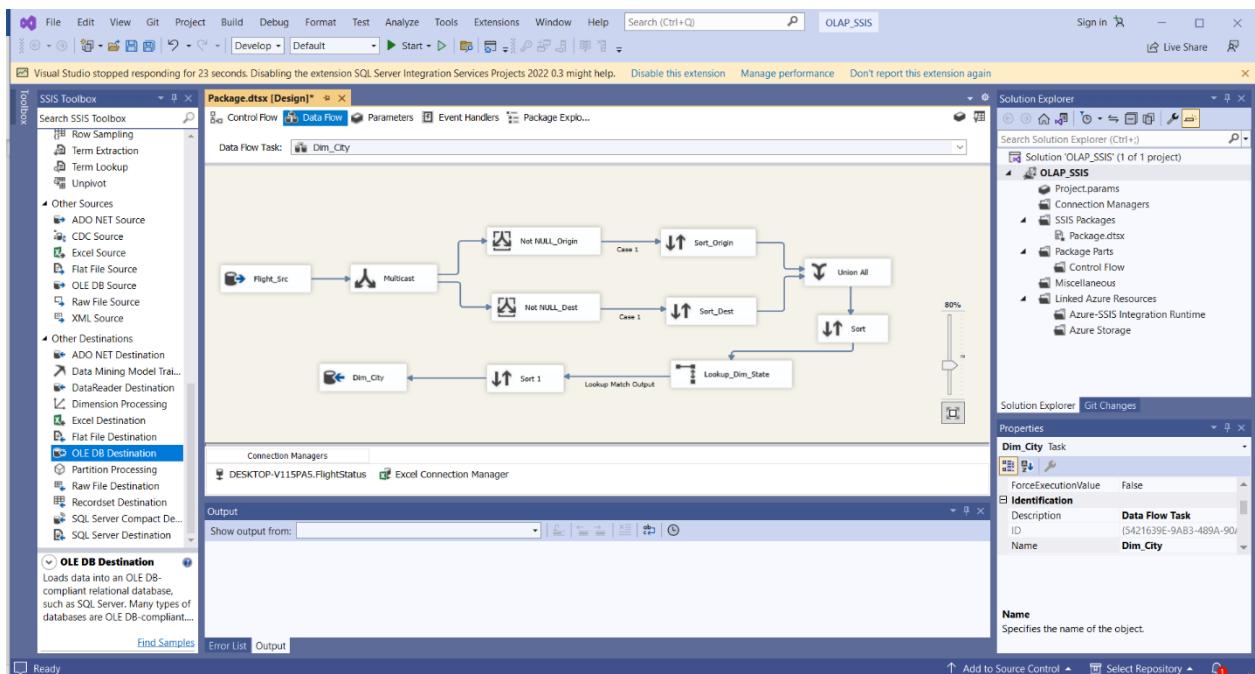
Hình 2.63. Tạo bảng Dim_City

- Bước 11: Qua tab Mapping để kiểm tra.



Hình 2.64. Qua Mapping để kiểm tra

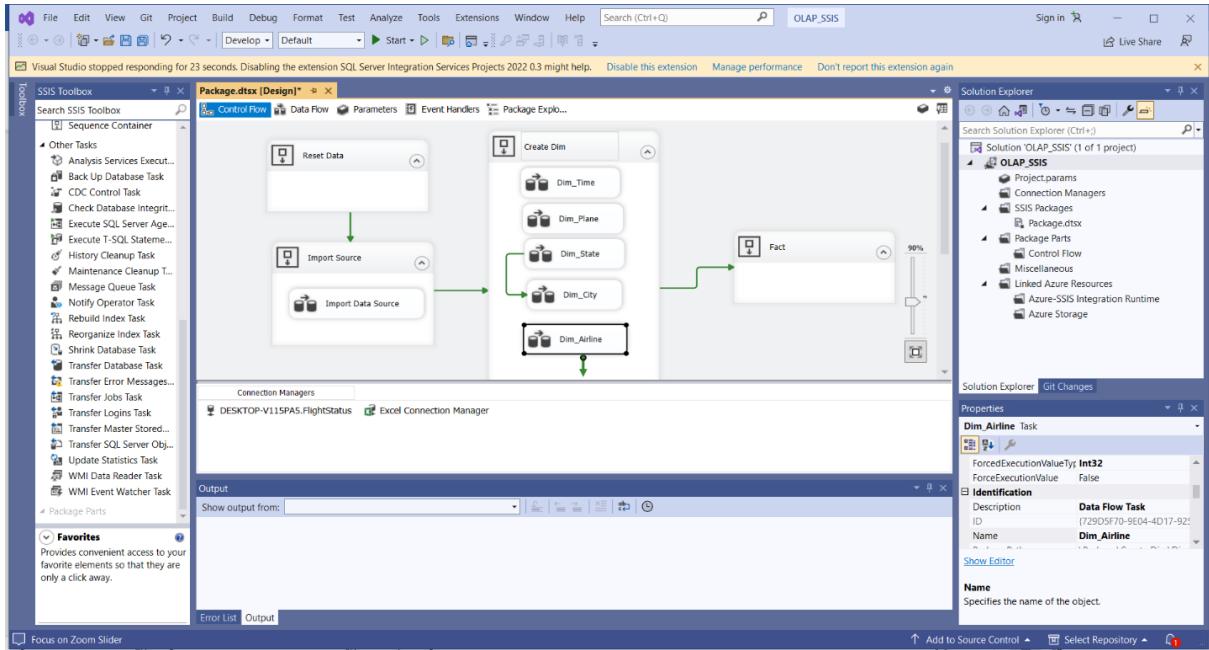
- Kết quả luồng thực hiện của bảng Dim_City



Hình 2.65. Luồng thực hiện của bảng Dim_City

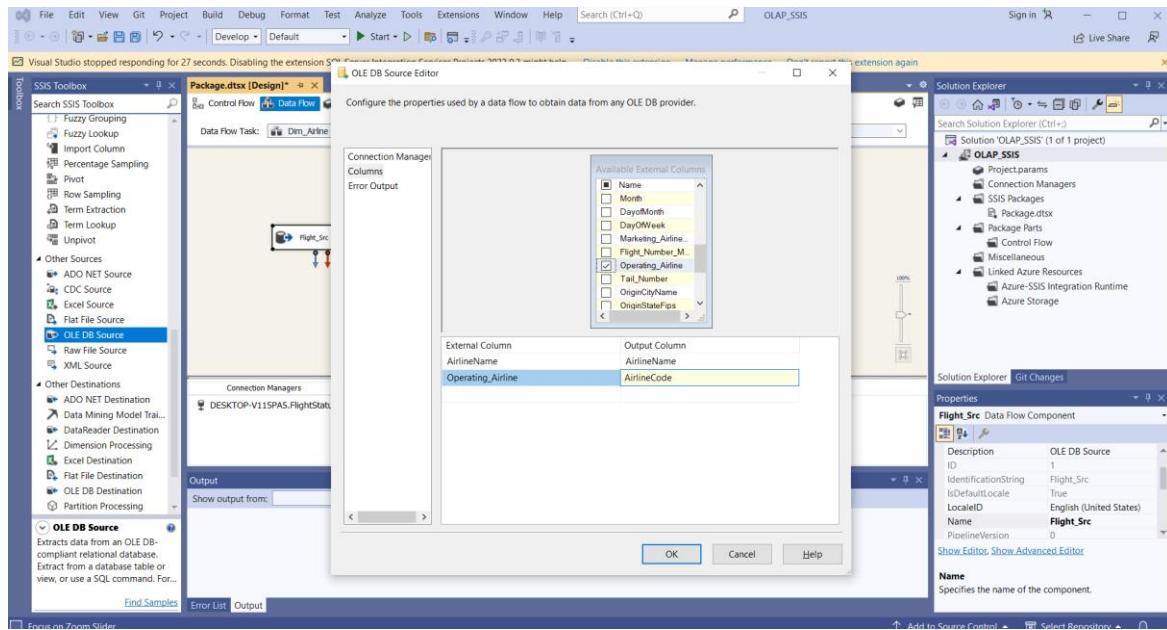
2.3.8. Tạo bảng Dim_Airline

- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_Airline



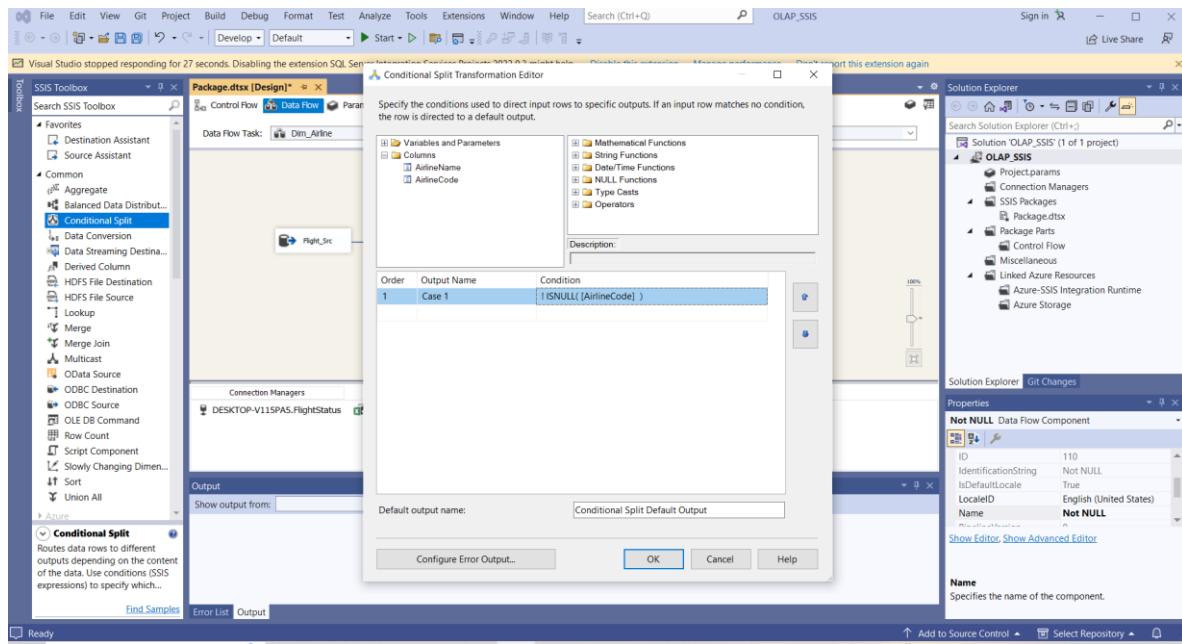
Hình 2.66. Tạo Data Flow Task Dim_Airline

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng, đổi tên cho các thuộc tính (nếu có).



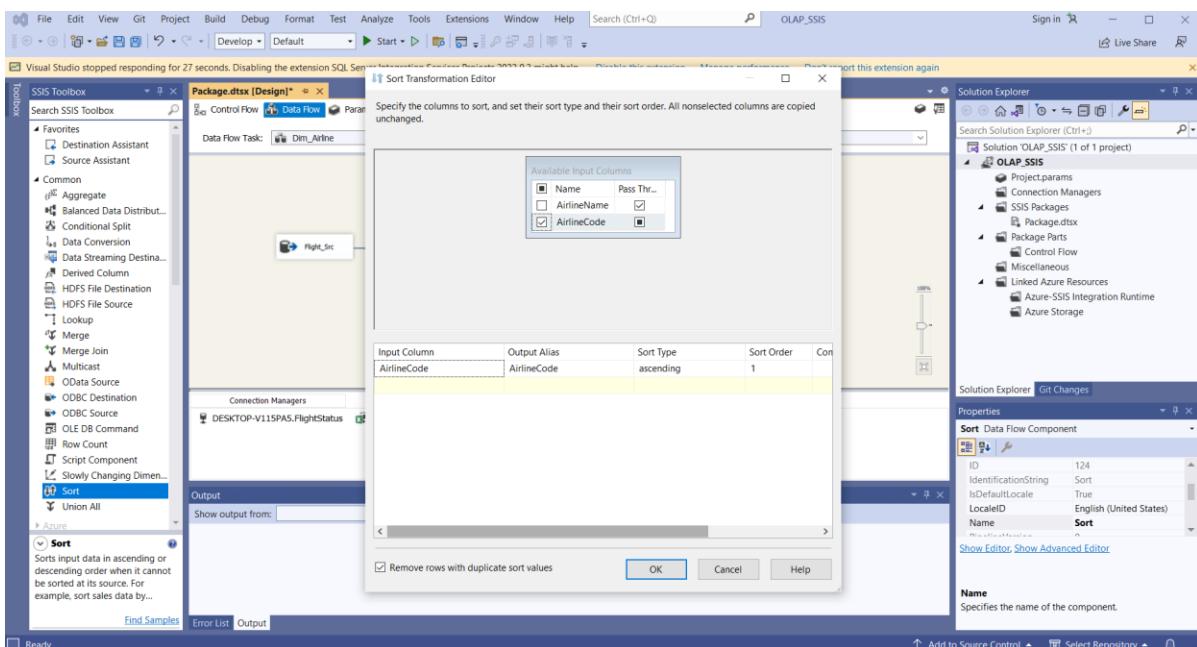
Hình 2.67. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

- Bước 3: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



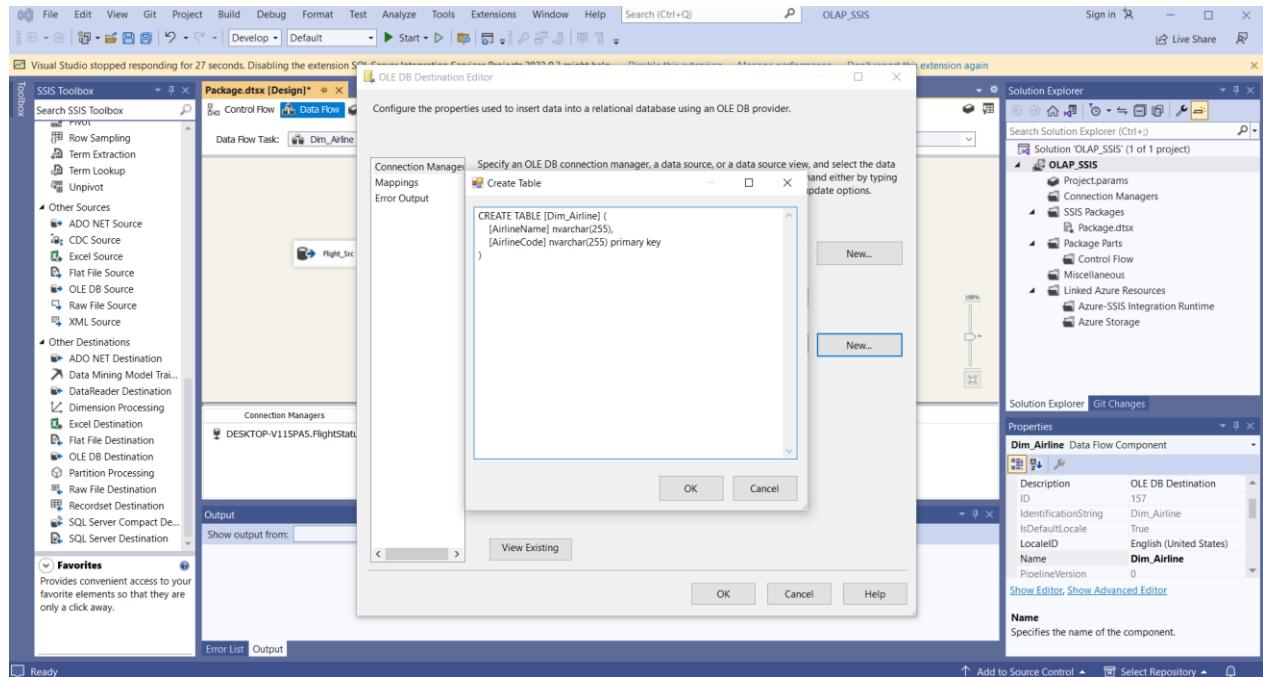
Hình 2.68. Thêm Conditional Split

- Bước 4: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu AirlineCode bị trùng.



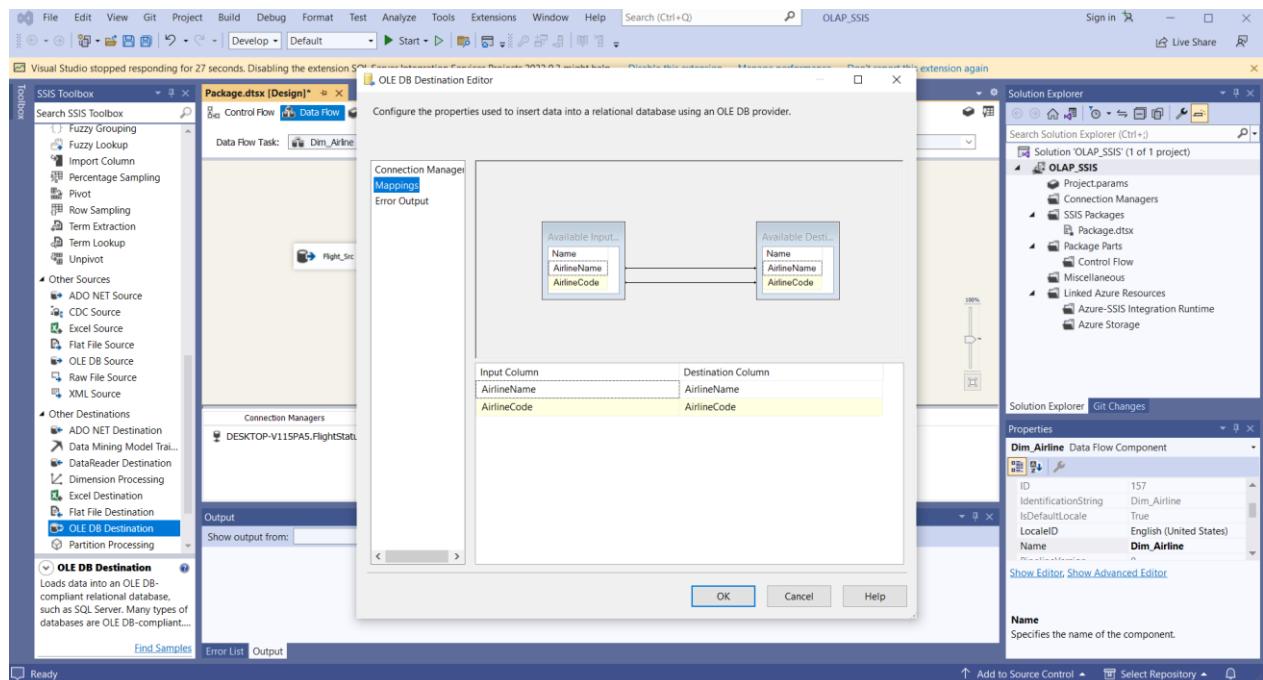
Hình 2.69. Sort dữ liệu và Remove rows

- Bước 5: Tạo 1 OLE Destination sau đó tạo bảng Dim_Airline.



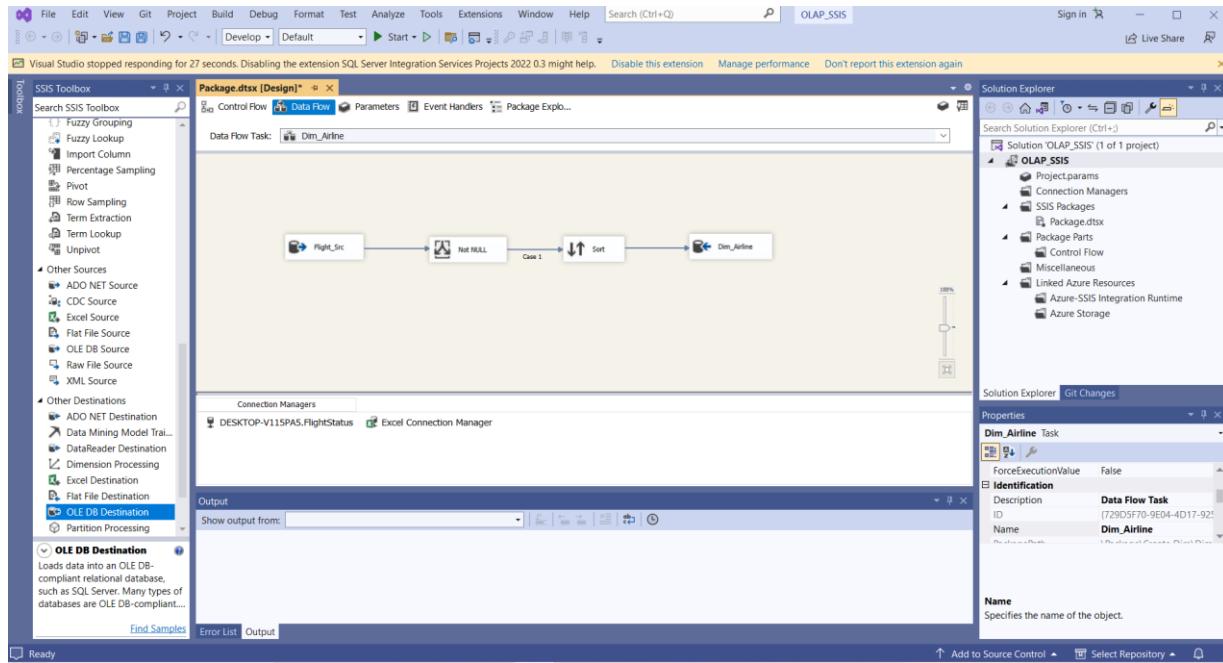
Hình 2.70. Tạo bảng Dim_Airline

- Bước 6: Qua tab Mapping để kiểm tra.



Hình 2.71. Qua Mapping để kiểm tra

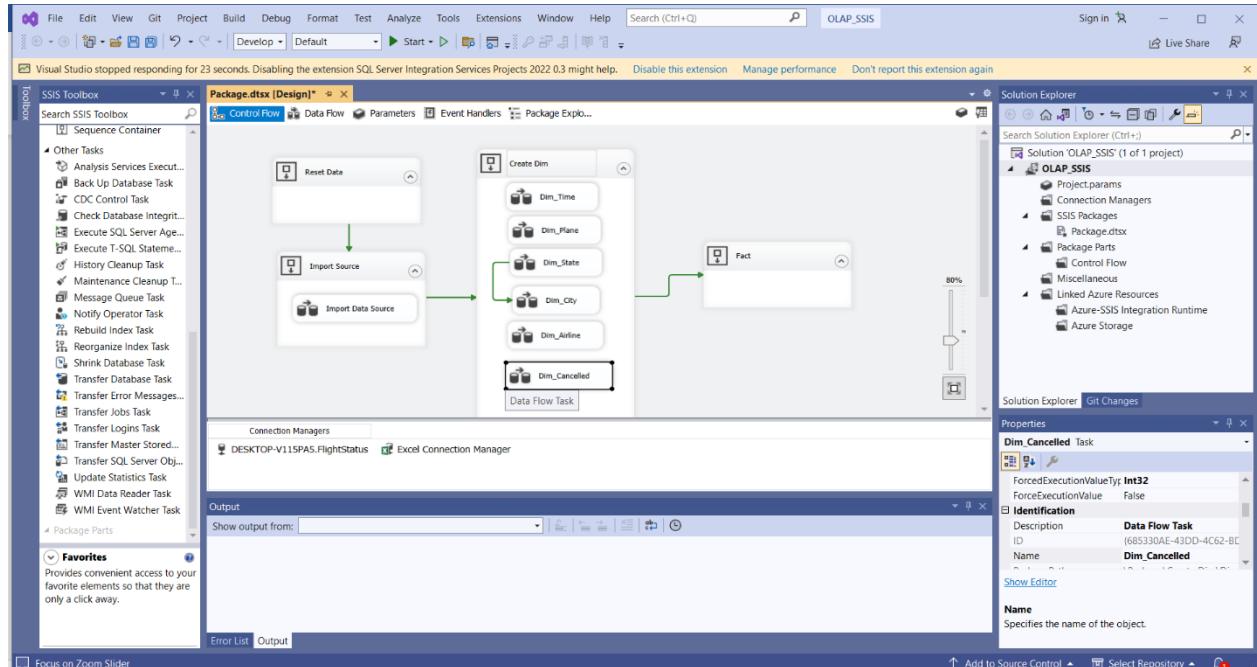
- Kết quả luồng thực hiện của bảng Dim_Airline



Hình 2.72. Luồng thực hiện của bảng Dim_Airline

2.3.9. Tạo bảng Dim_Cancelled

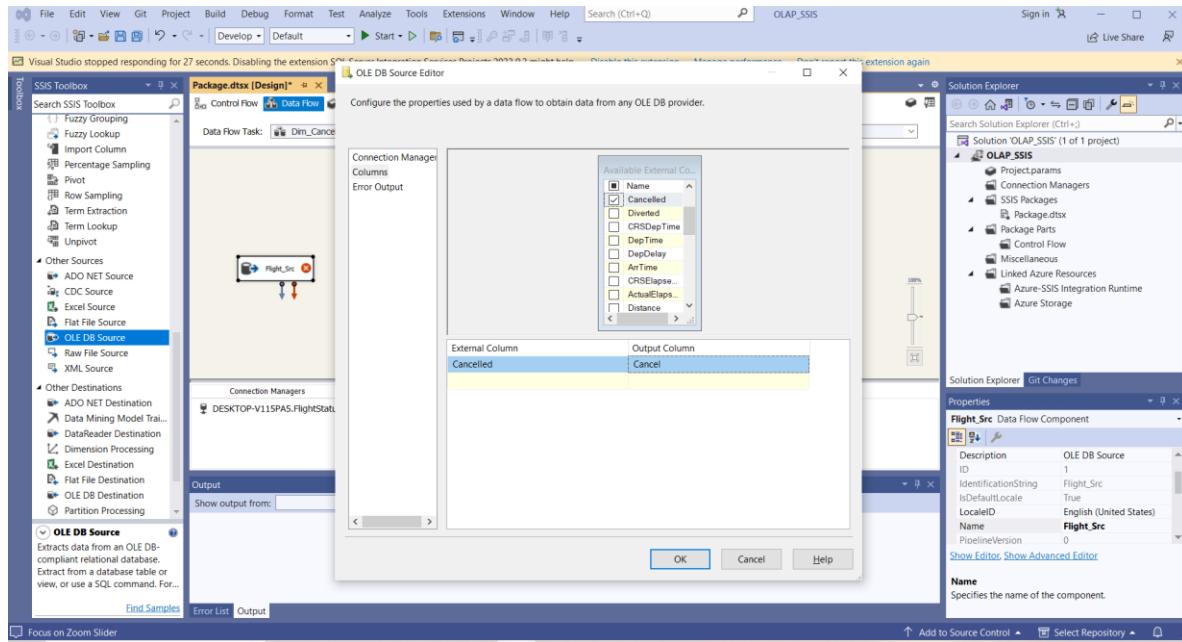
- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_Cancelled



Hình 2.73. Tạo Data Flow Task Dim_Cancelled

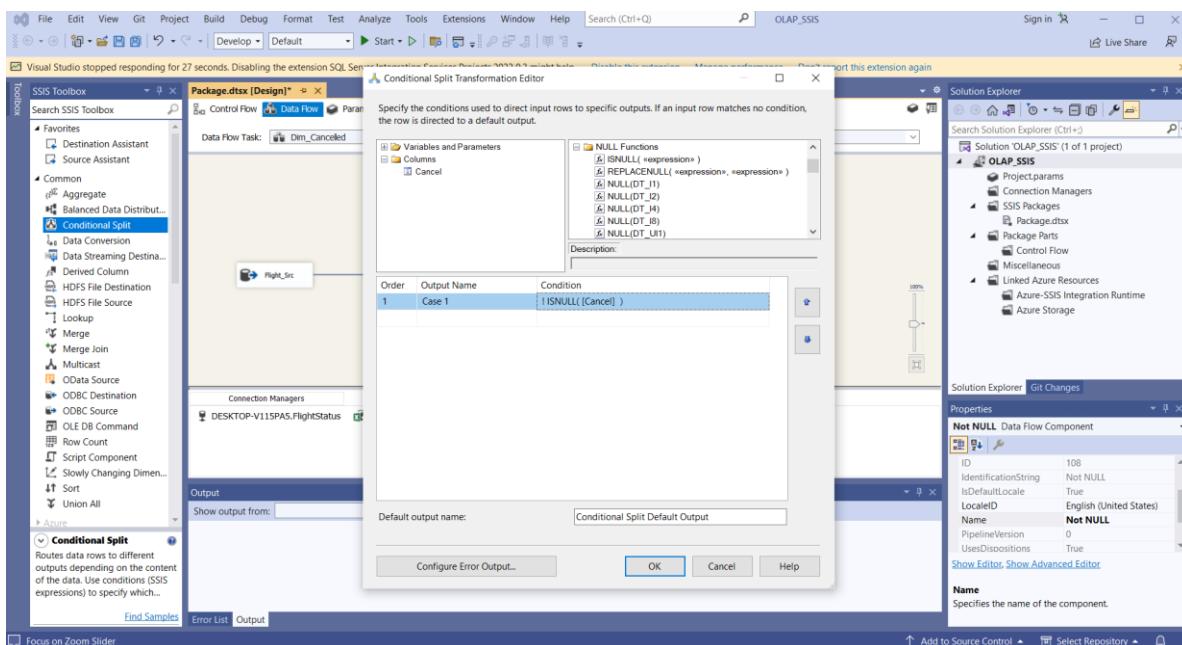
IS217 – Kho dữ liệu và OLAP

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng đổi tên cho các thuộc tính (nếu có).



Hình 2.74. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

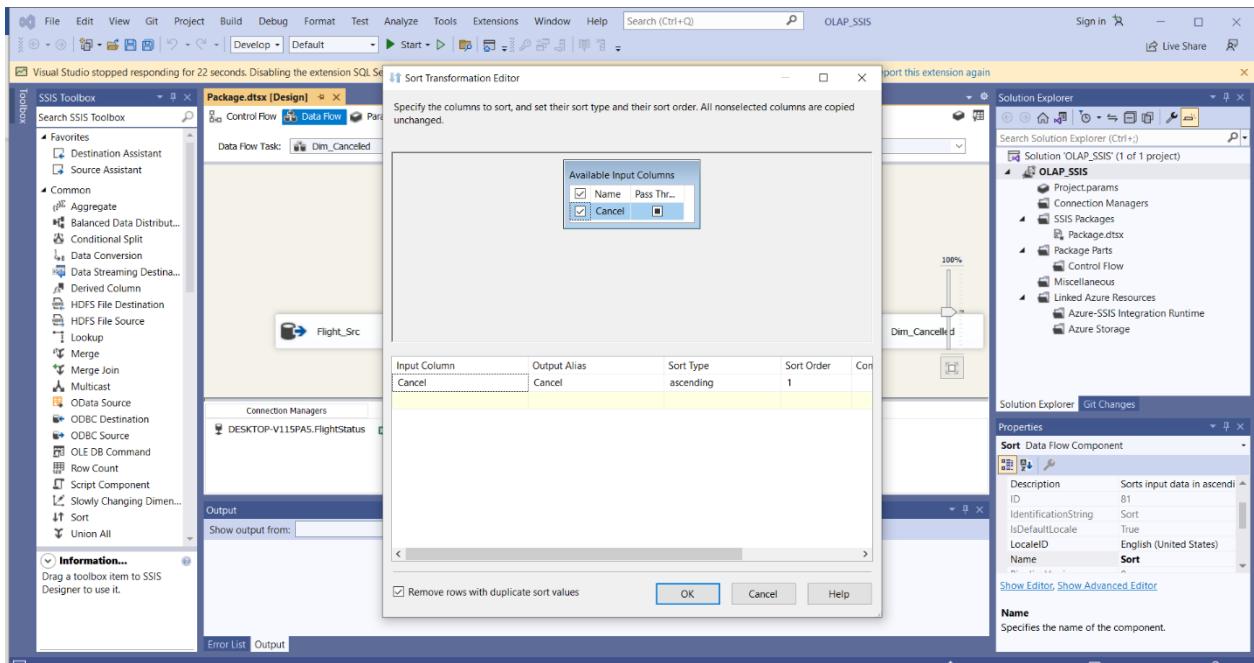
- Bước 3: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



Hình 2.75. Thêm Conditional Split

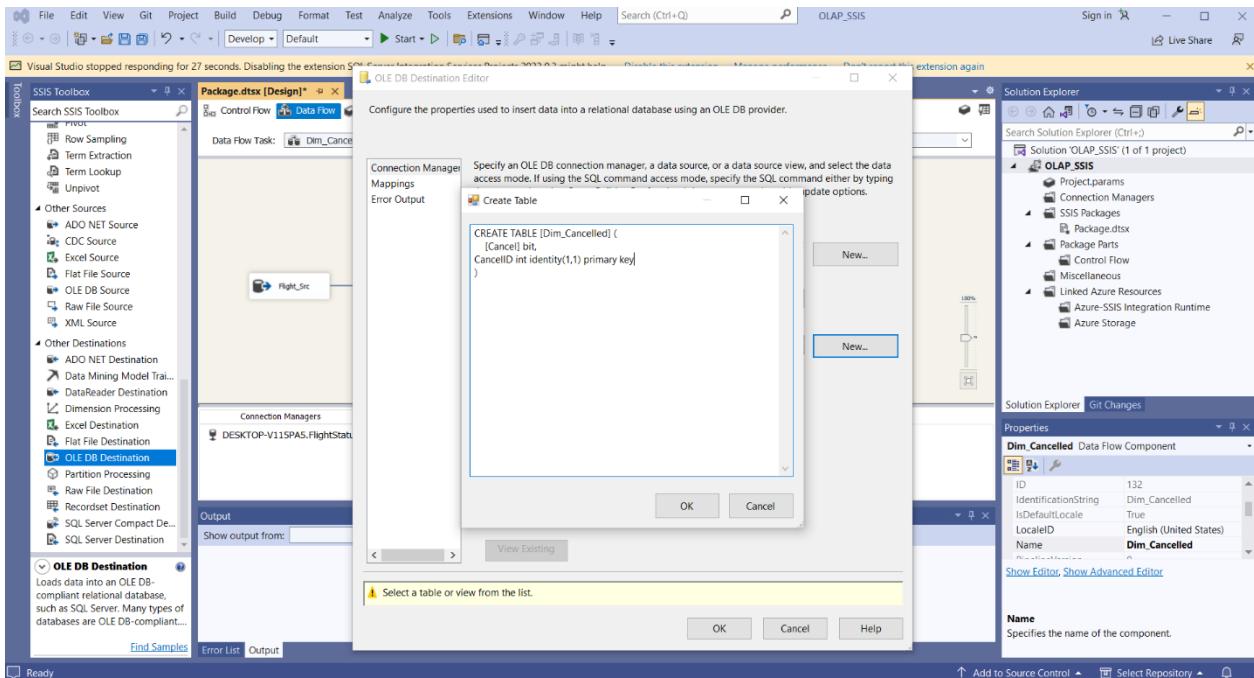
IS217 – Kho dữ liệu và OLAP

- Bước 4: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu bị trùng.



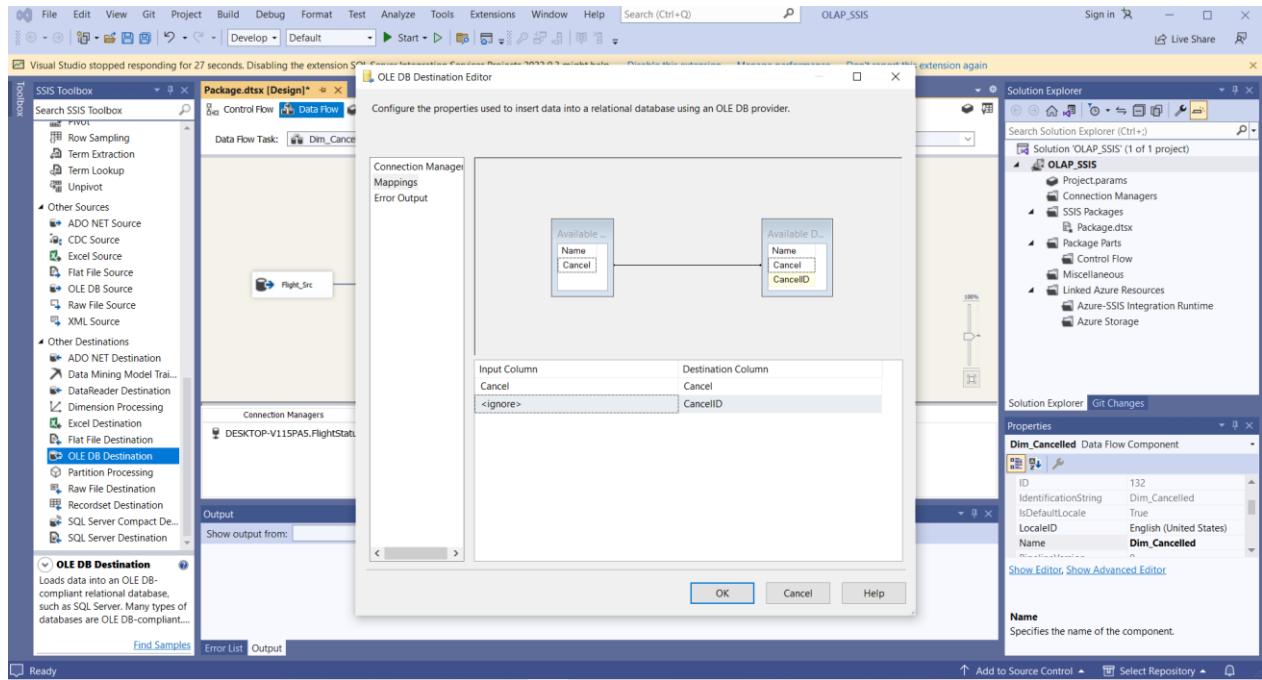
Hình 2.76. Sort dữ liệu và Remove rows

- Bước 5: Tạo 1 OLE Destination sau đó tạo bảng Dim_Cancelled.



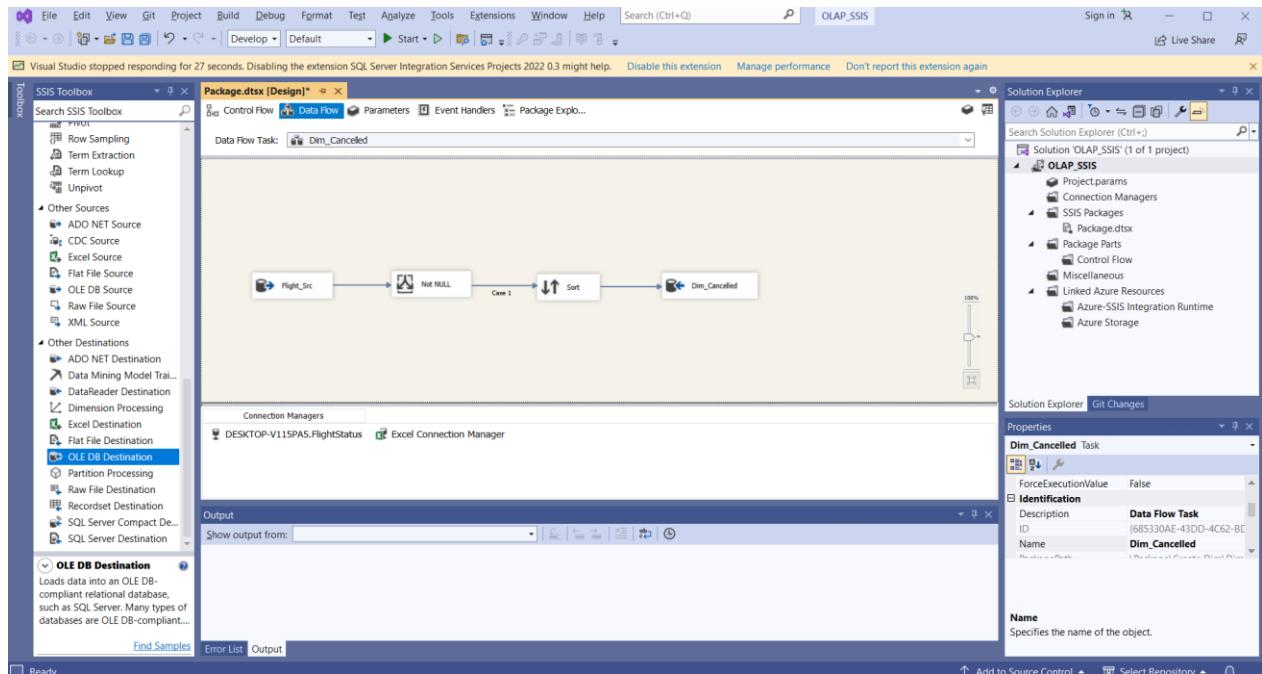
Hình 2.77. Tạo bảng Dim_Cancelled

- Bước 6: Qua tab Mapping để kiểm tra.



Hình 2.78. Qua Mapping để kiểm tra

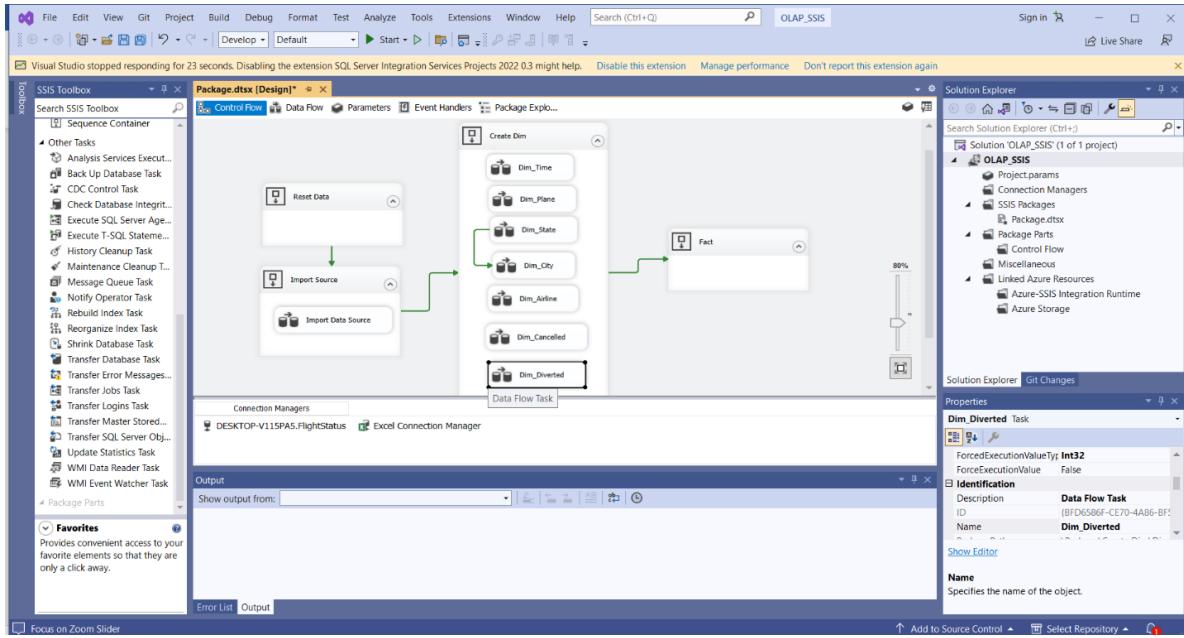
- Kết quả luồng thực hiện của bảng Dim_Cancelled



Hình 2.79. Luồng thực hiện của bảng Dim_Cancelled

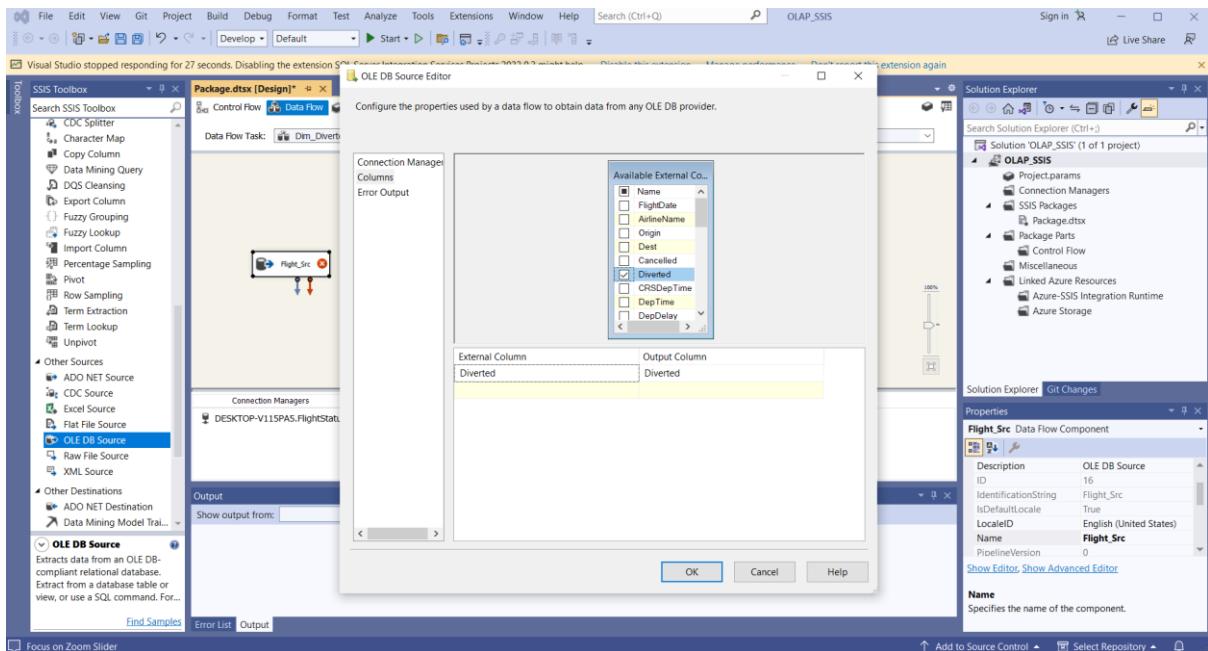
2.3.10. Tạo bảng Dim_Diverted

- Bước 1: Kéo Data Flow Task vào Container đặt tên là Dim_Diverted



Hình 2.80. Tạo Data Flow Task Dim_Diverted

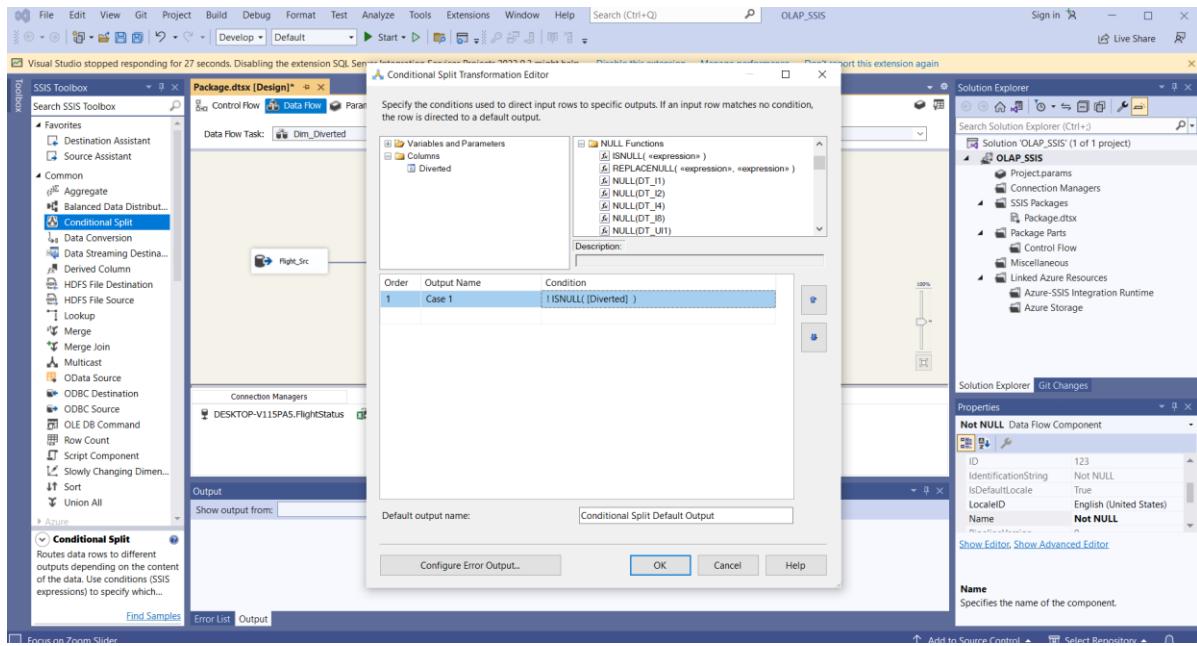
- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng đổi tên cho các thuộc tính (nếu có).



Hình 2.81. Tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng

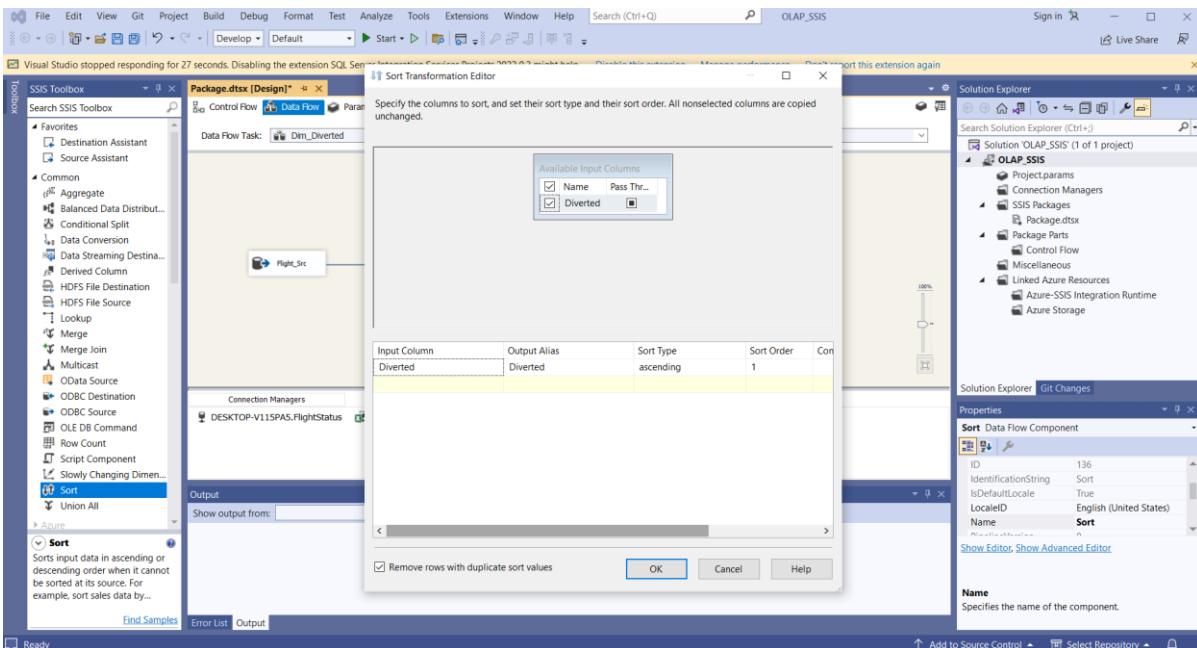
IS217 – Kho dữ liệu và OLAP

- Bước 3: Dùng công cụ Conditional Split và kết nối với Source để bắt đầu cắt dữ liệu có điều kiện, lọc những dòng NULL ra khỏi trước khi đưa vào kho dữ liệu.



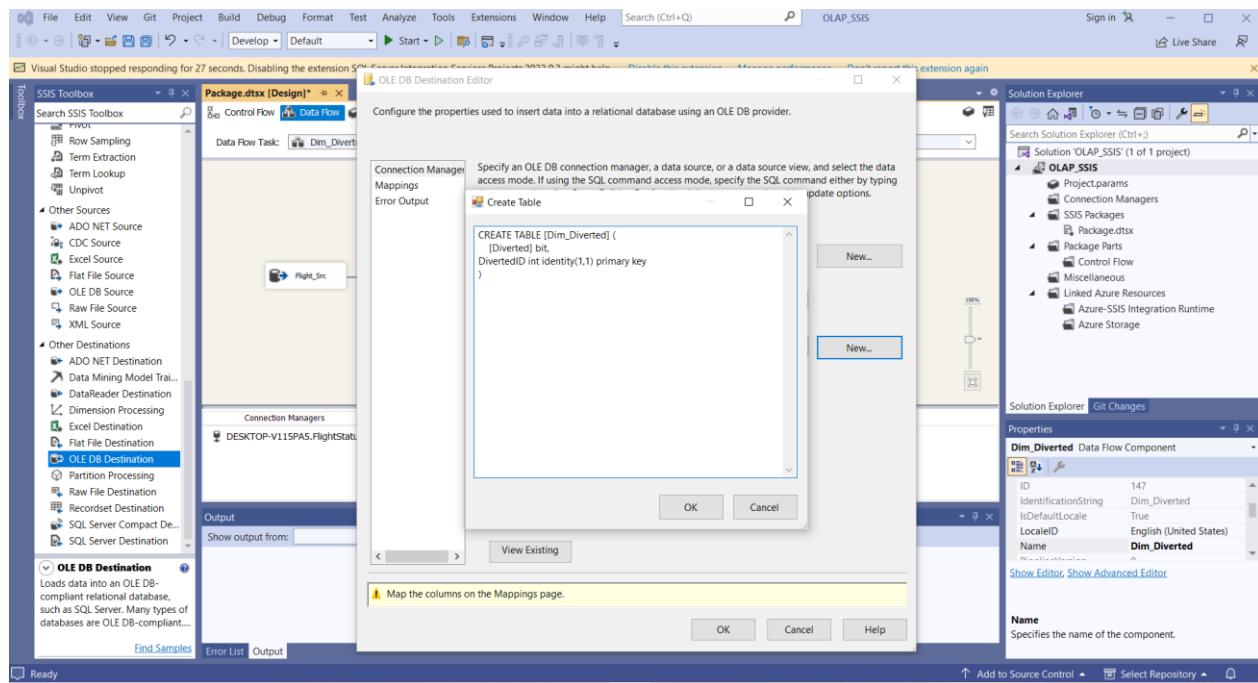
Hình 2.82. Thêm Conditional Split

- Bước 4: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu Diverted bị trùng.



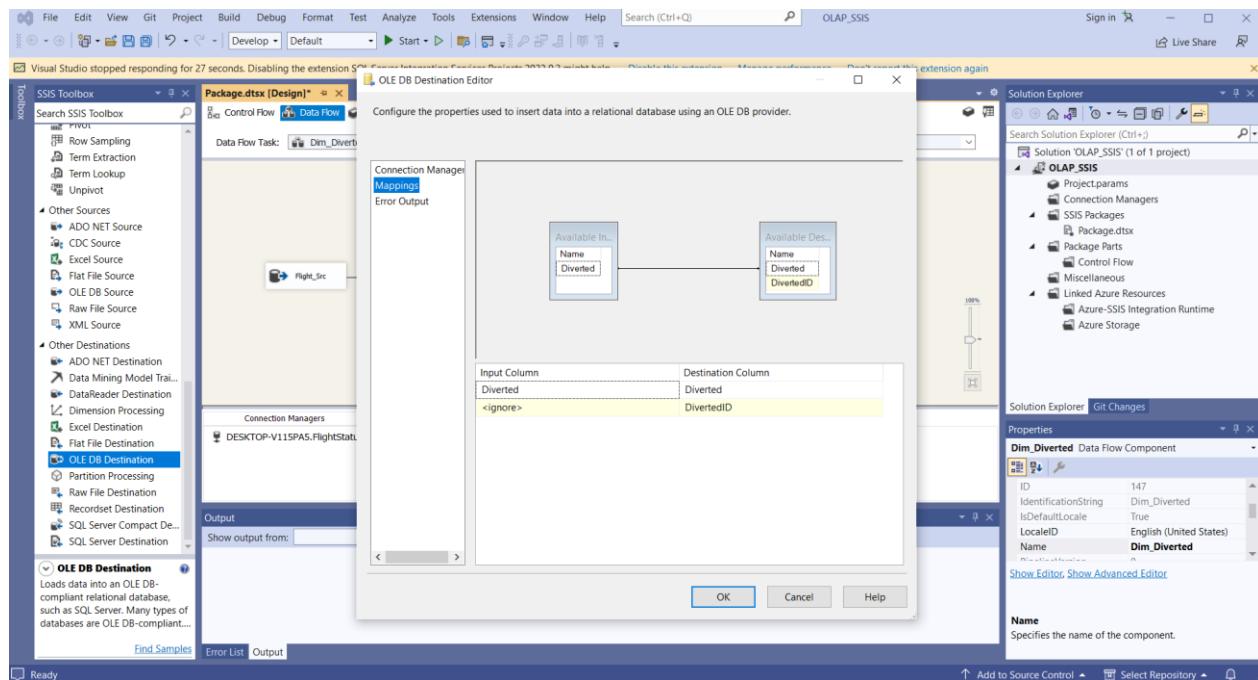
Hình 2.83. Sort dữ liệu và Remove rows

- Bước 5: Tạo 1 OLE Destination sau đó tạo bảng Dim_Diverted.



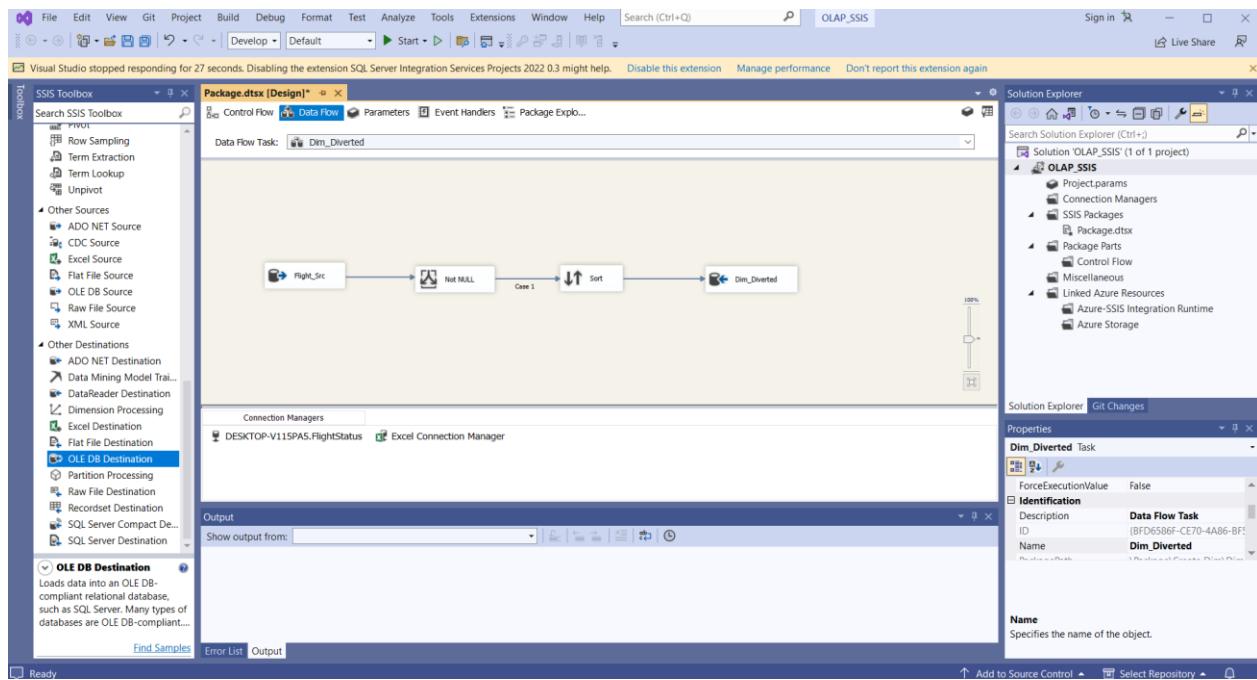
Hình 2.84. Tao bảng Dim_Diverted

- Bước 6: Qua tab Mapping để kiểm tra.



Hình 2.85. Qua Mapping để kiểm tra

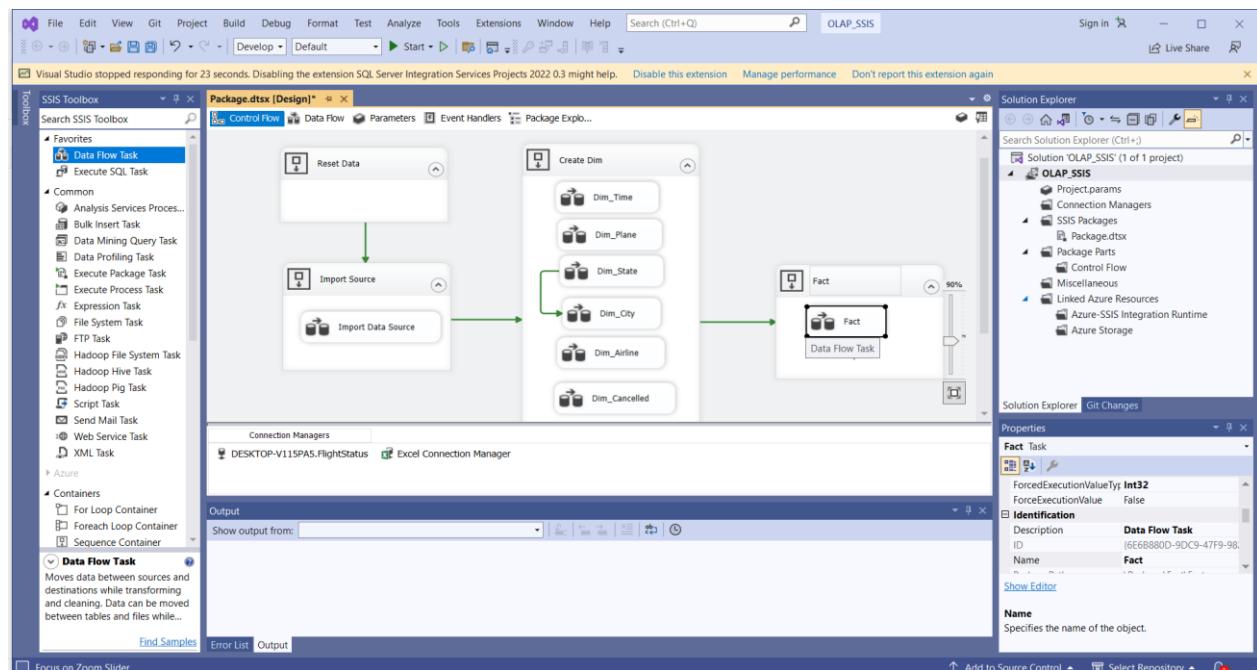
- Kết quả luồng thực hiện của bảng Dim_Diverted



Hình 2.86. Luồng thực hiện của bảng Dim_Diverted

2.3.11.Tạo bảng Fact

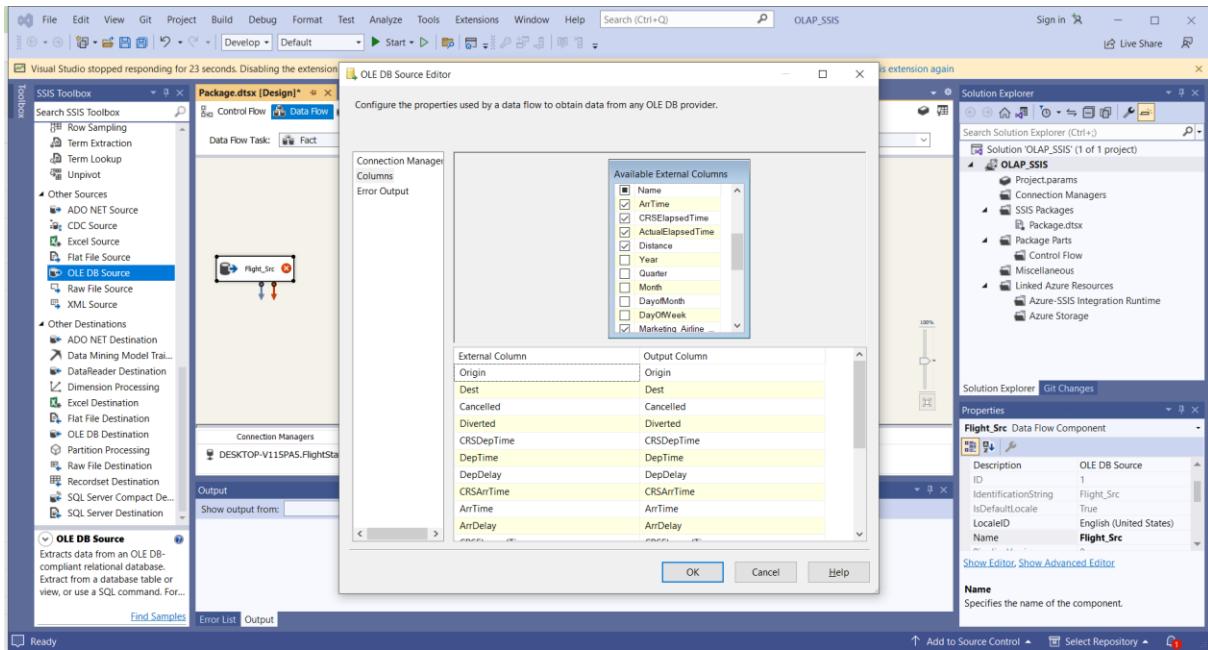
- Bước 1: Kéo Data Flow Task vào Container đặt tên là Fact



Hình 2.87. Tạo Data Flow Task Fact

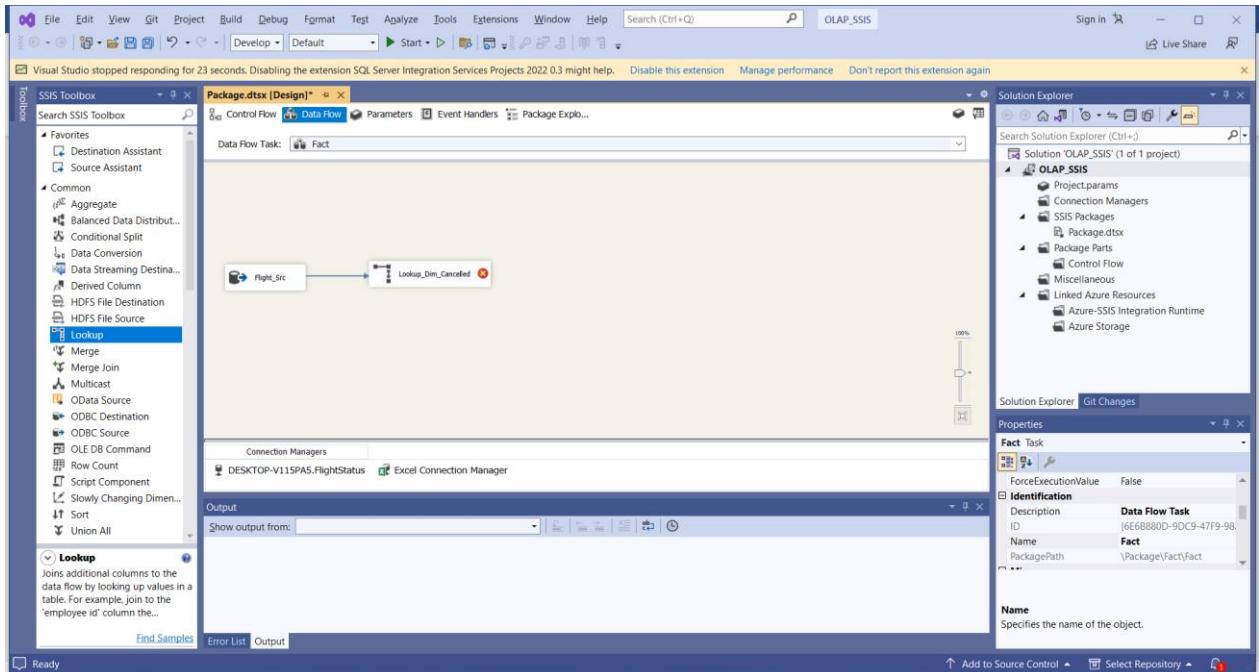
IS217 – Kho dữ liệu và OLAP

- Bước 2: Khởi tạo OLE DB Source chứa Dataset và chọn thuộc tính cho bảng.



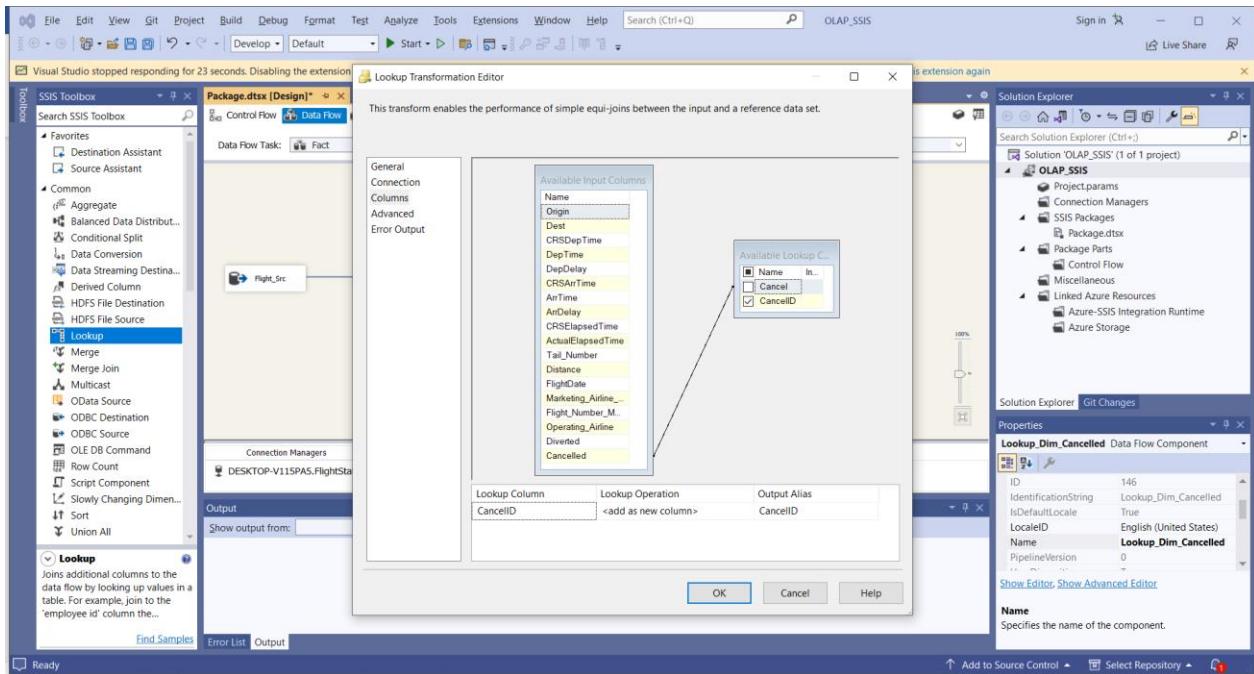
Hình 2.88. Khởi tạo OLE DB Source

- Bước 3: Dùng Lookup để kết bảng Flight_Src và bảng Dim_Cancelled thông qua thuộc tính Cancel



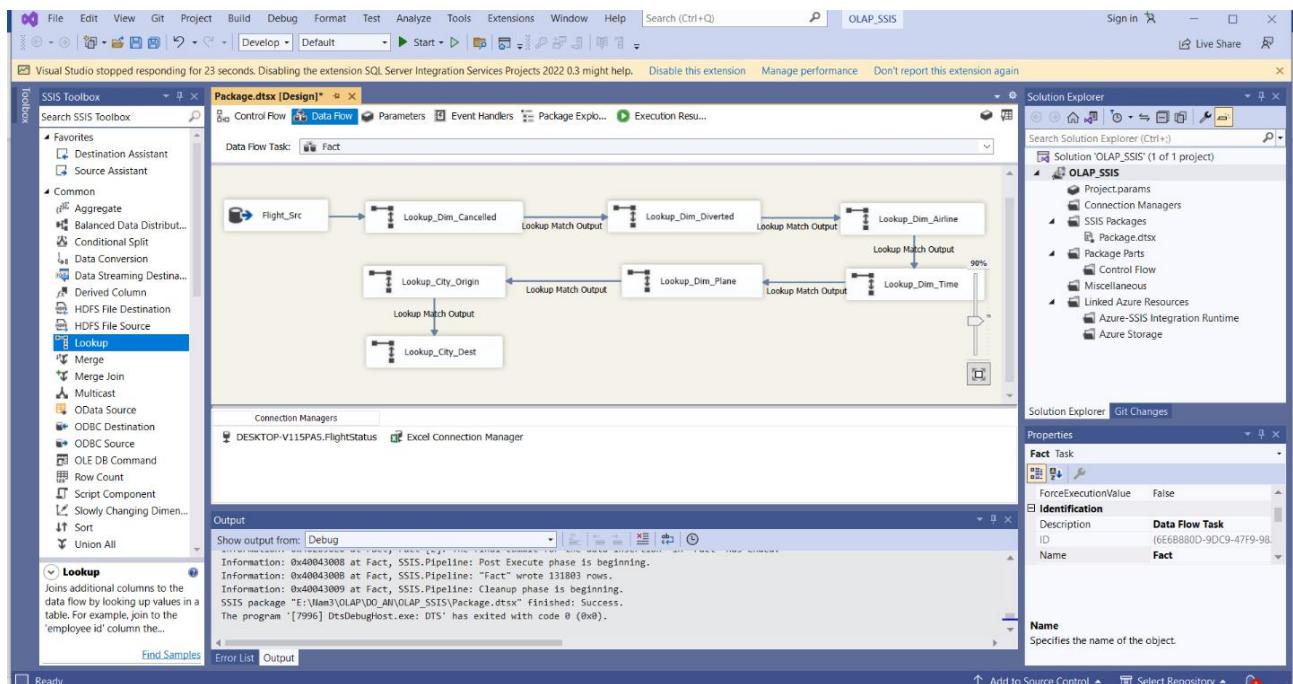
Hình 2.89. Tạo Lookup kết Source với Dim_Cancelled

IS217 – Kho dữ liệu và OLAP



Hình 2.90. Kết nối Source và bảng Dim_Cancelled

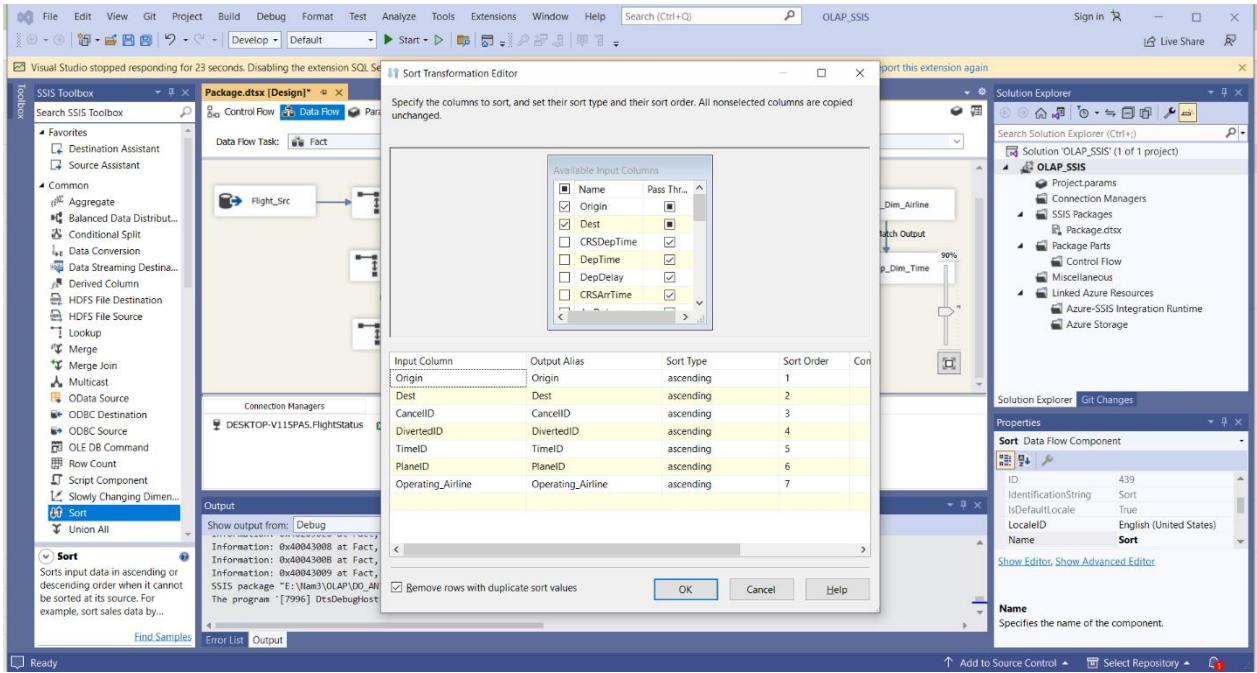
- Bước 4: Tương tự, ta dùng Lookup để kết các bảng Dim lại với nhau.



Hình 2.91. Dùng Lookup để kết các bảng Dim lại

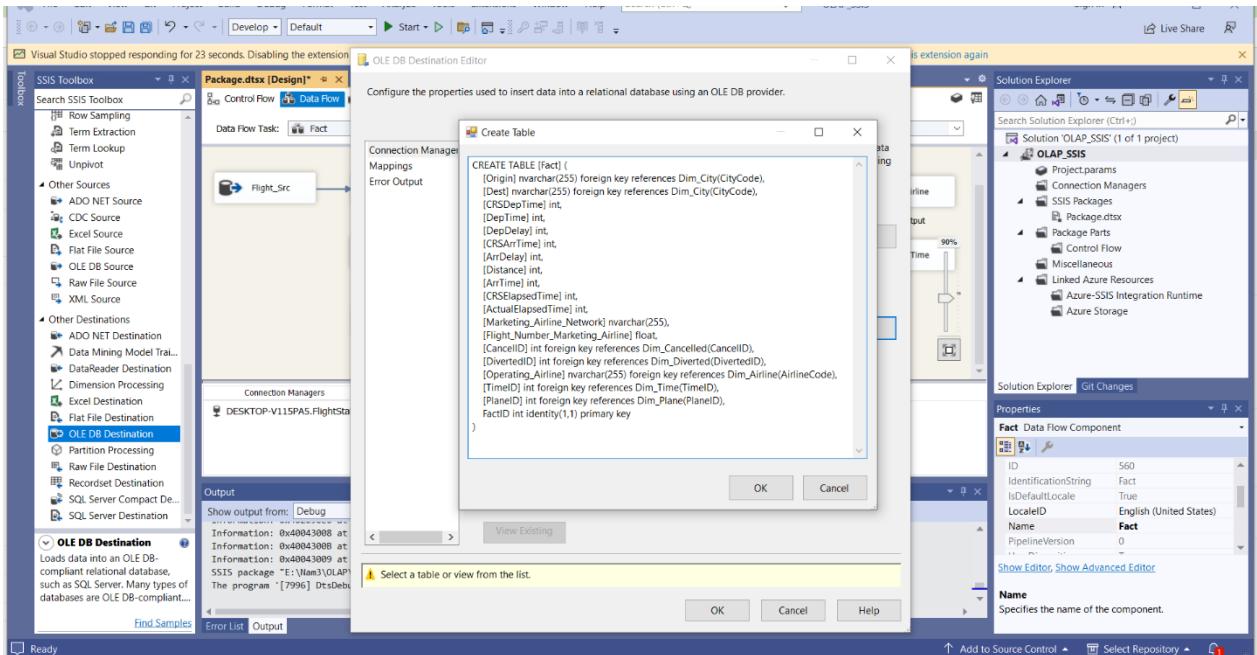
IS217 – Kho dữ liệu và OLAP

- Bước 5: Dùng Sort để sắp xếp lại dữ liệu. Dùng Sort tick vào ô Remove rows with duplicate sort values để xóa những dữ liệu trùng.



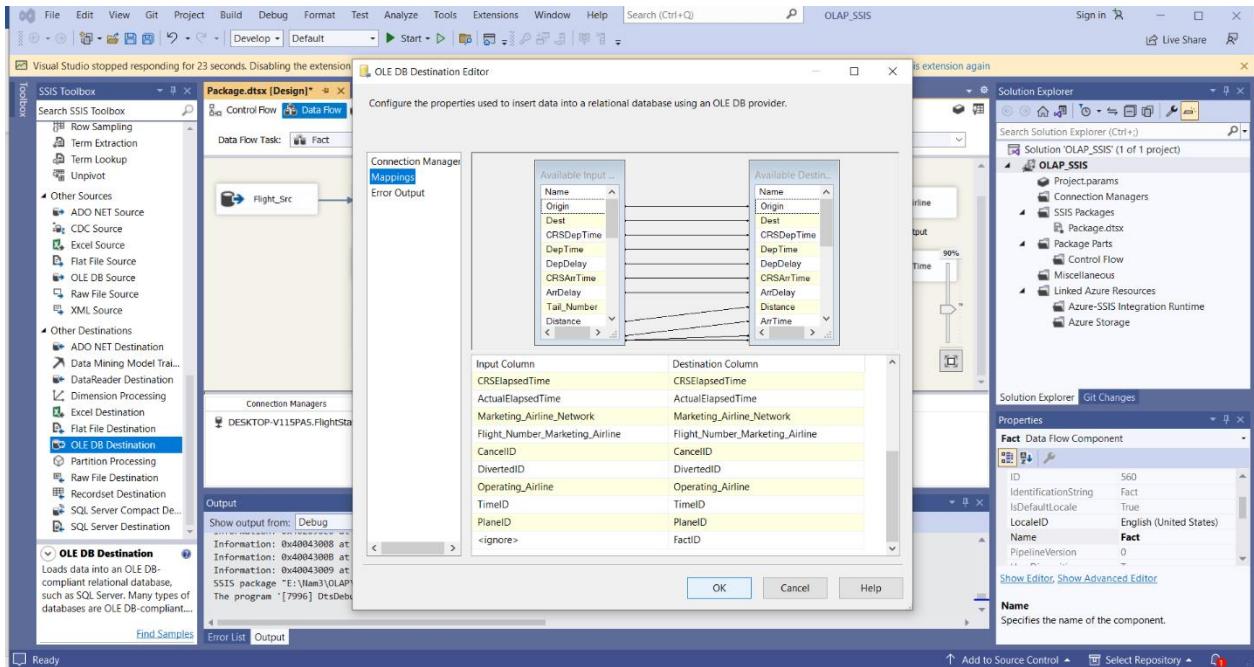
Hình 2.92. Sort dữ liệu và Remove rows

- Bước 6: Tạo bảng Fact.



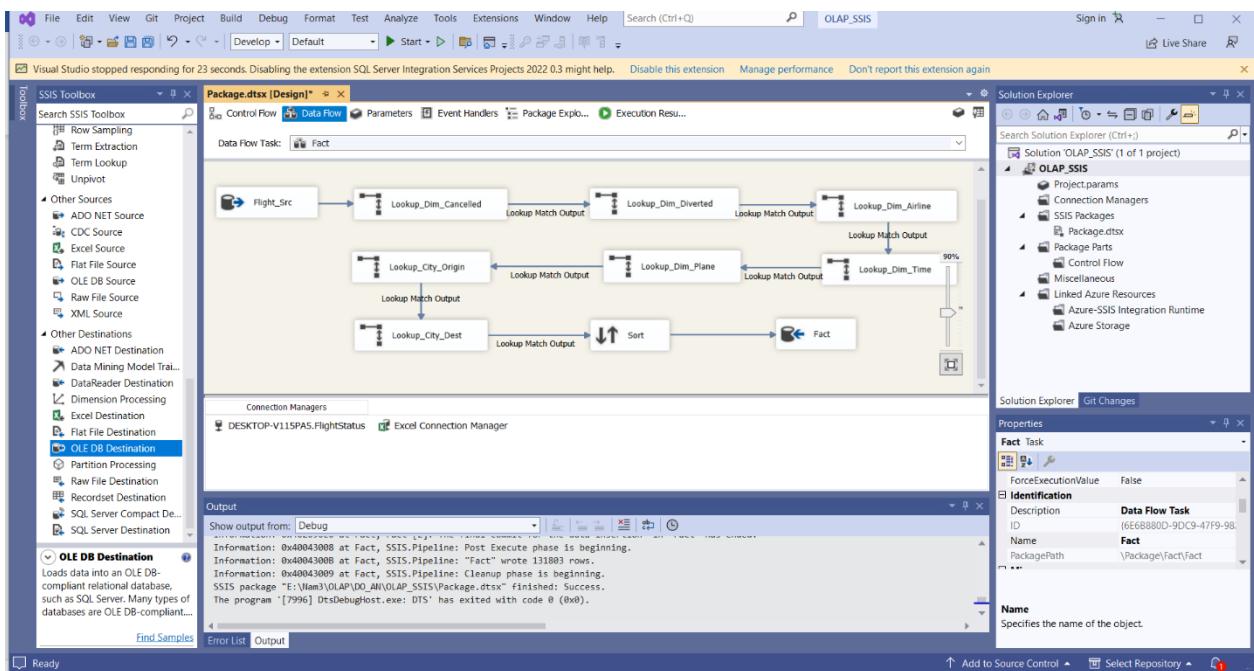
Hình 2.93. Tạo bảng Fact

- Bước 7: Qua Mappings để kiểm tra



Hình 2.94. Qua Mapping để kiểm tra

- Kết quả luồng thực hiện bảng Fact

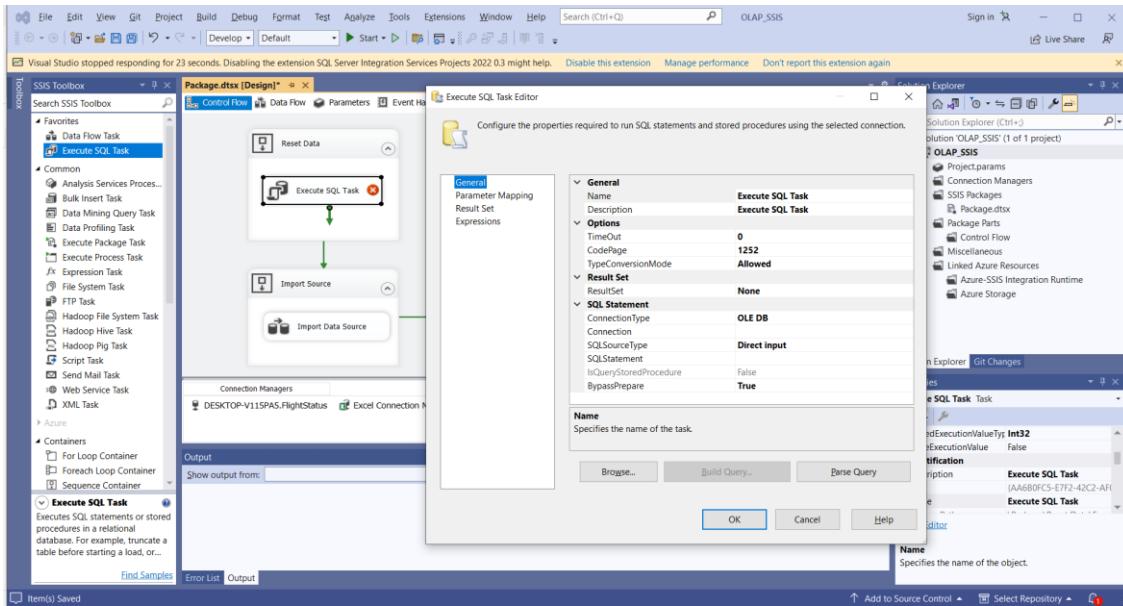


Hình 2.95. Luồng thực hiện của bảng Fact

2.3.12. Import dữ liệu vào database.

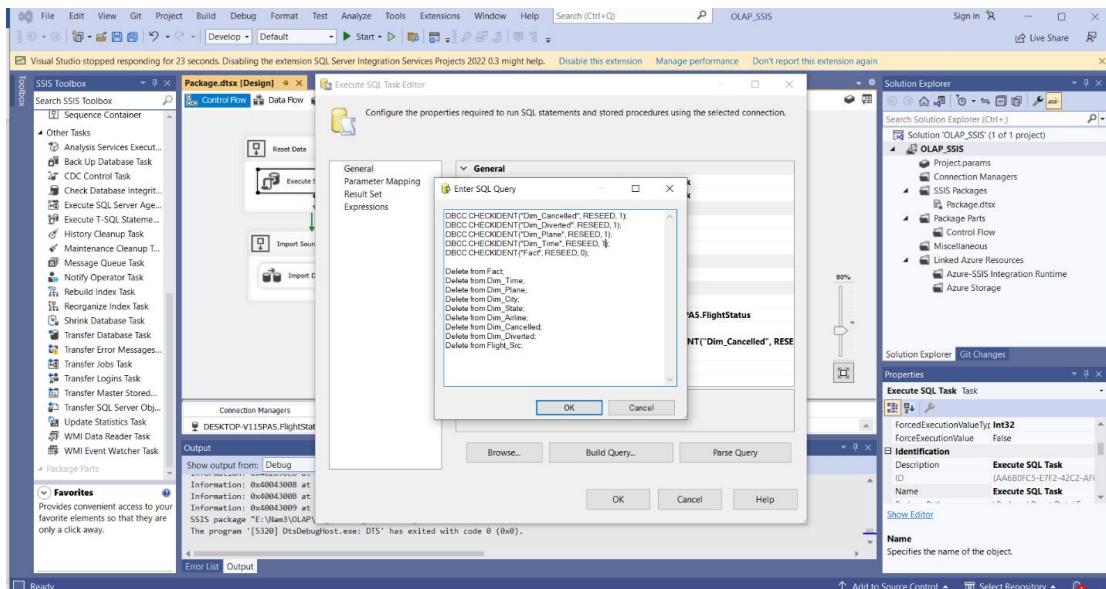
Để thực hiện quá trình import dữ liệu vào database, ta thực hiện các bước sau:

- Bước 1: Ở SSIS Toolbox kéo thả Execute SQL Task vào Control Flow để thực thi import dữ liệu vào database.



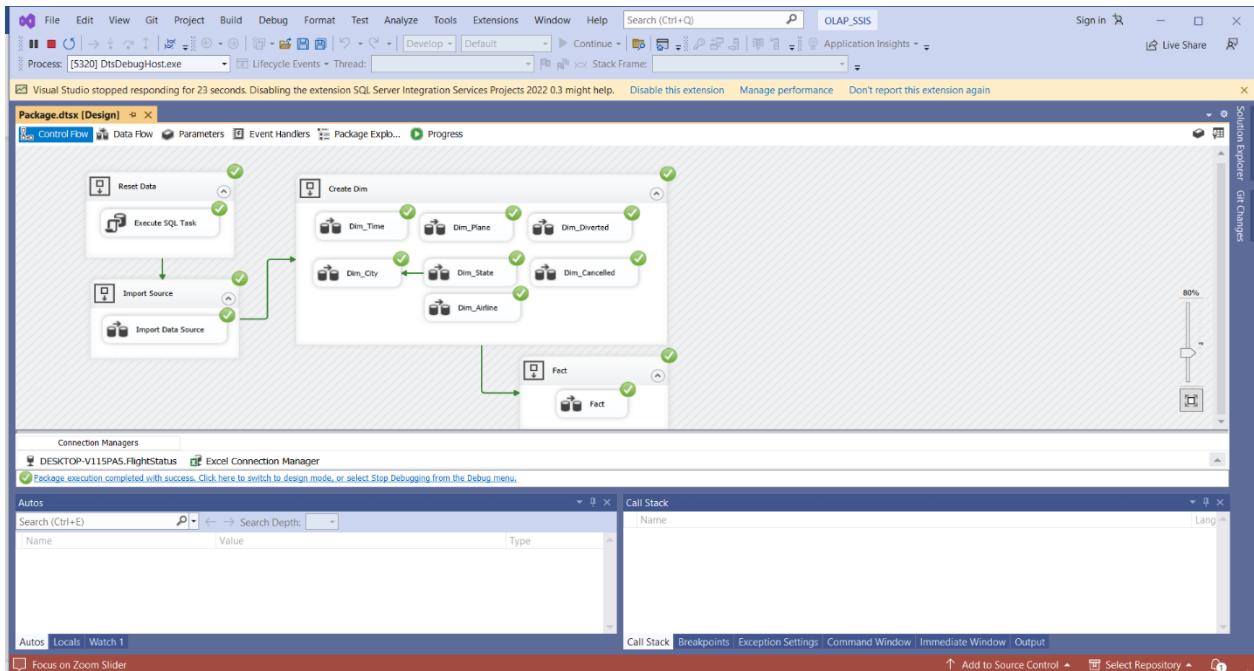
Hình 2.96. Thông tin bảng Execute SQL Task

- Bước 2: Thực hiện câu truy vấn để xóa hết dữ liệu ban đầu (nếu có).

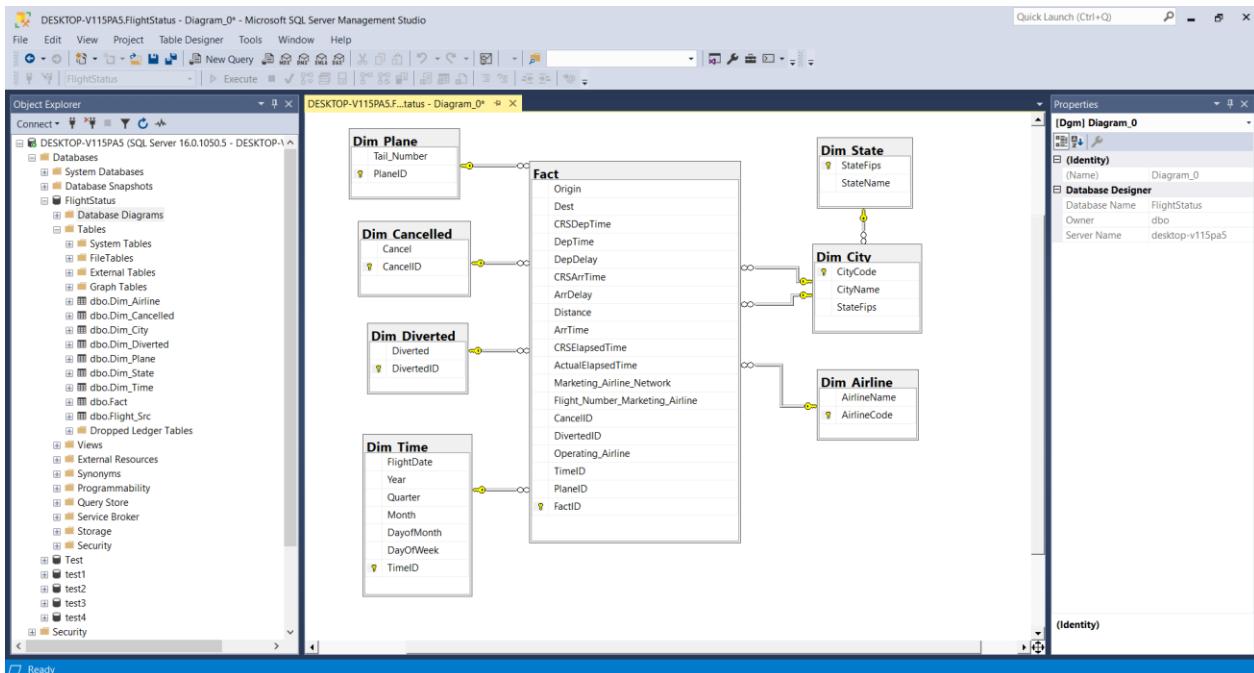


Hình 2.97. Tạo câu truy vấn để xóa dữ liệu trước đó

- Kết quả luồng thực hiện của quá trình SSIS



Hình 2.98. Luồng thực hiện của quá trình SSIS



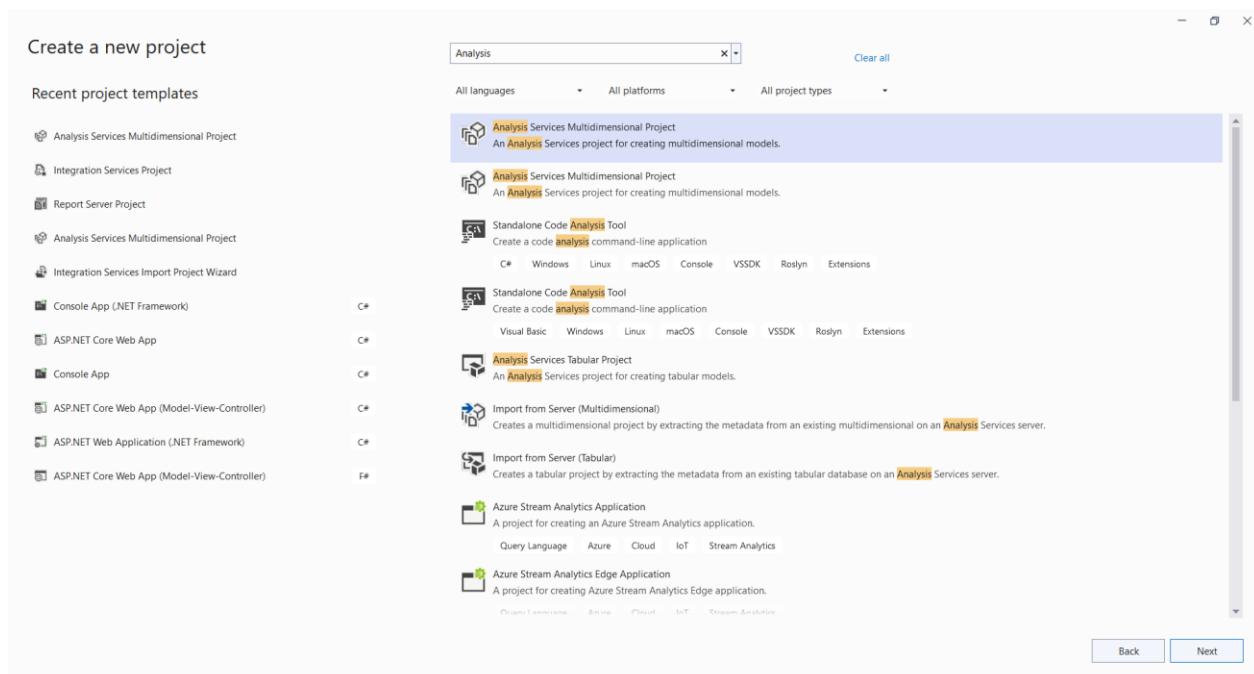
Hình 2.99. Diagram được tạo ra trong SQL Server

CHƯƠNG 3 – PHÂN TÍCH DỮ LIỆU VÀ BÁO CÁO

Ở chương 3, ta sẽ tập trung vào việc thực hiện quá trình SSAS để tạo ra một kho dữ liệu đa chiều, cung cấp khả năng phân tích nhanh chóng và linh hoạt. Sau khi xây dựng kho dữ liệu, truy vấn MDX được sử dụng để truy xuất và phân tích dữ liệu trong cube. MDX cung cấp các cú pháp và hàm mạnh mẽ để thực hiện các phép tính, lọc dữ liệu và tạo các báo cáo tùy chỉnh. Trong phần này, dữ liệu cũng được trực quan hóa và trình bày nhờ công cụ Power BI và Excel giúp ta có cái nhìn cụ thể hơn về tập dữ liệu.

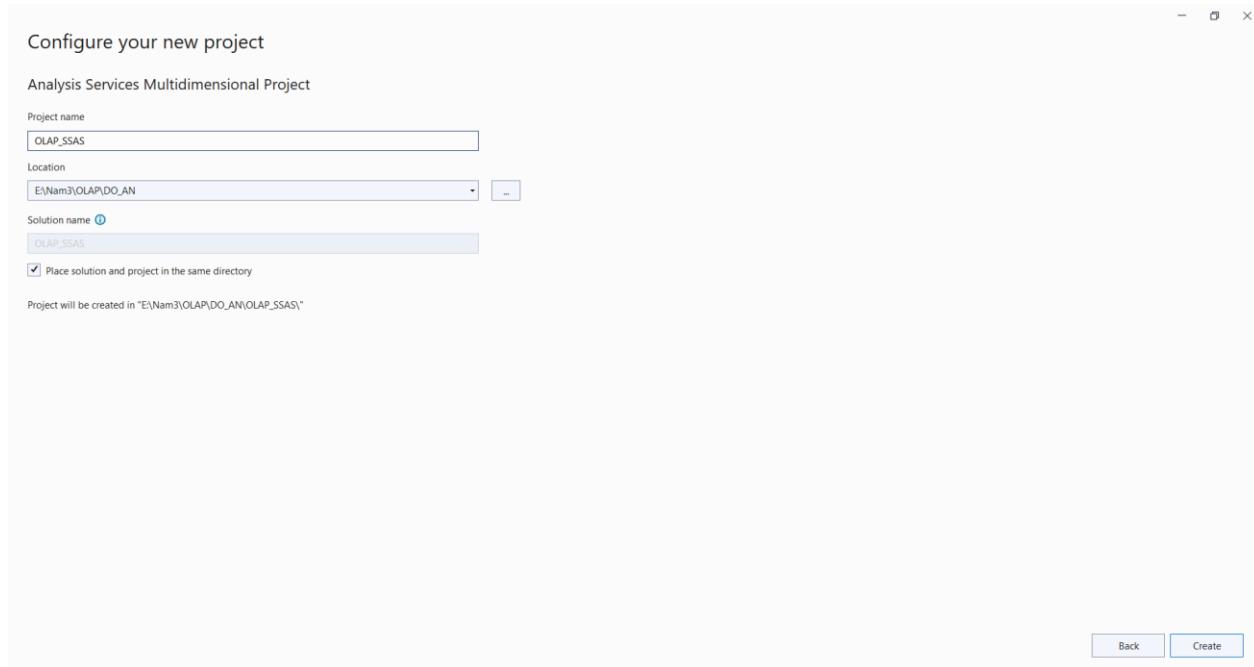
3.1. Tạo Project SSAS trong Visual Studio 2022

- Bước 1: Khởi động Visual Studio 2022
- Bước 2: Tìm từ khóa Analysis, sau đó chọn **Analysis Services Multidimensional Project**, sau đó nhấn Next.



Hình 3.1. Tạo mới project SSAS

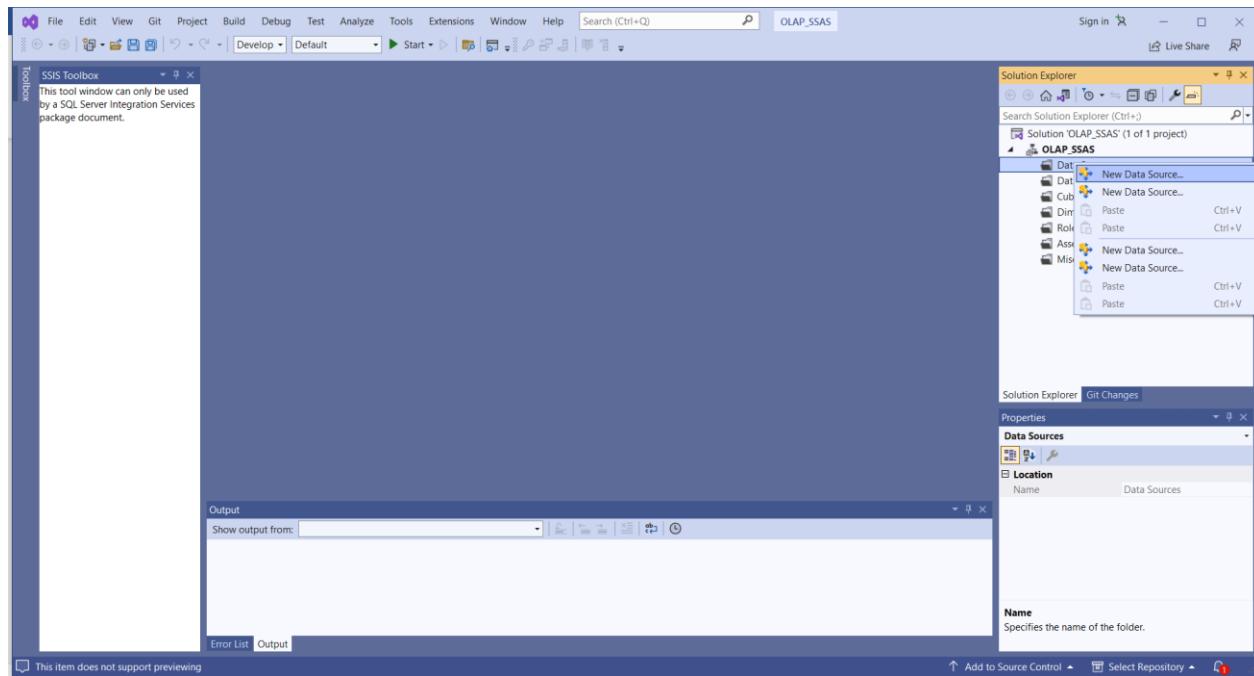
- Bước 3: Đặt tên cho Project và nhấn OK. Giao diện quá trình SSAS hiện ra.



Hình 3.2. Điện thoại tin project

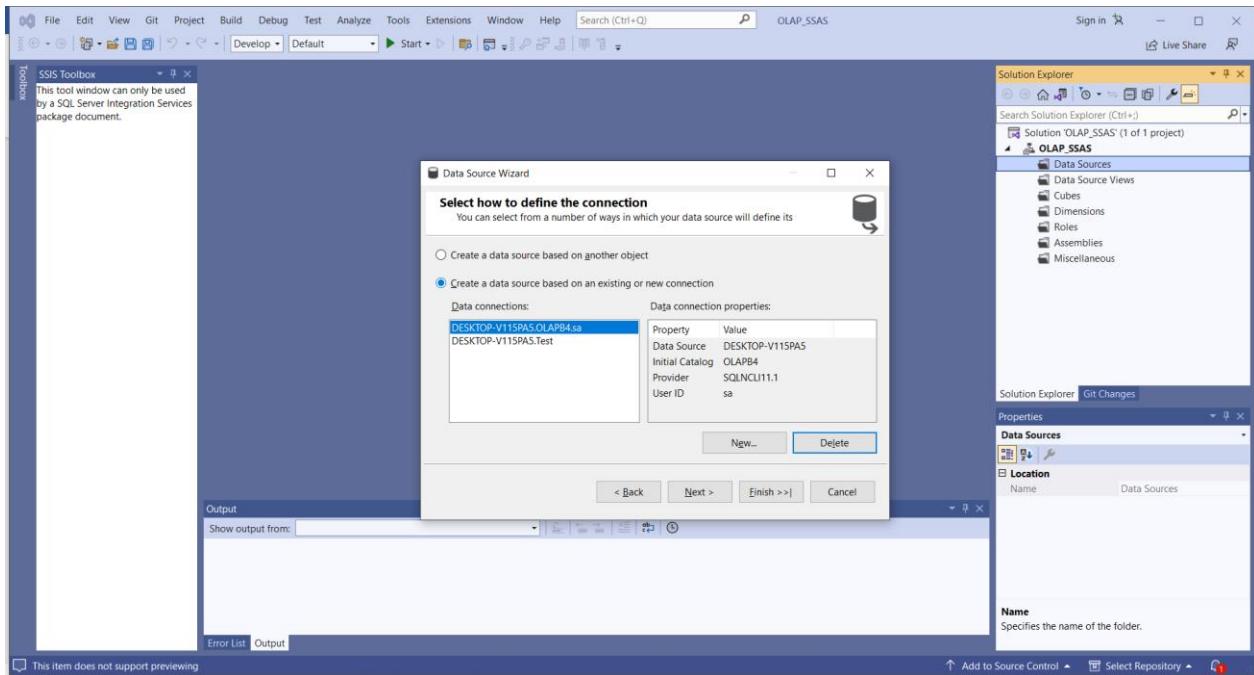
3.2. Connect đến Data Source

- Bước 1: Chọn Data Sources → New Data Source → Next.



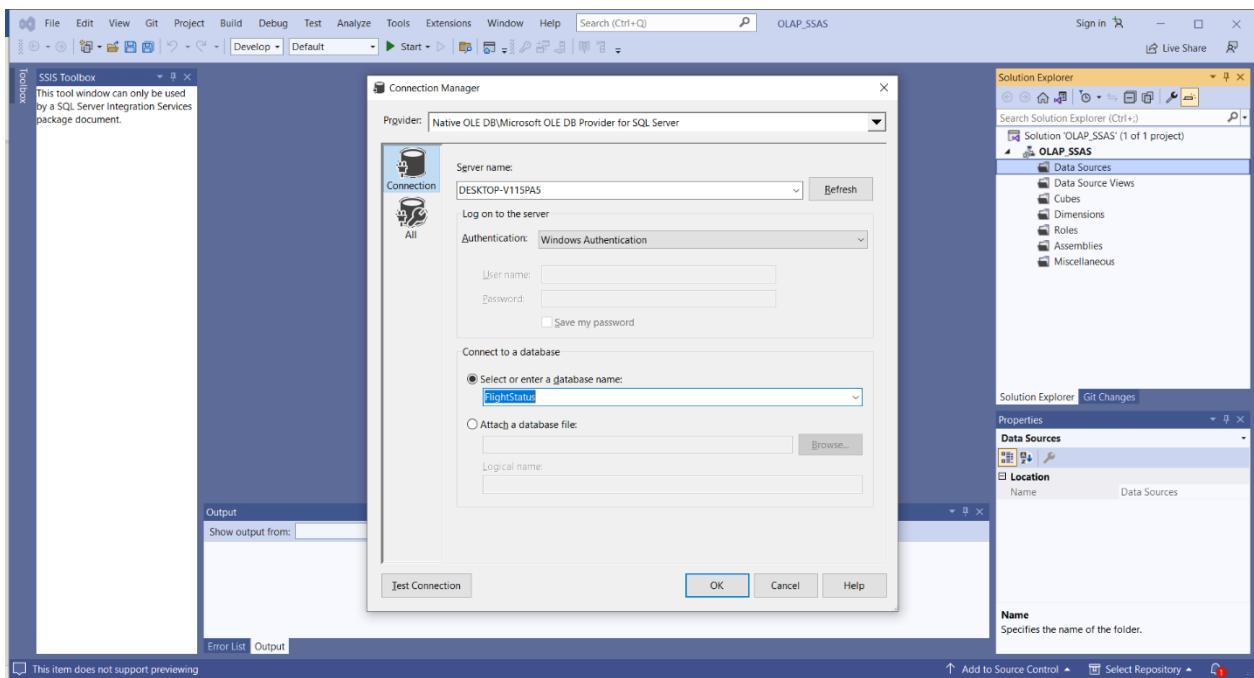
Hình 3.3. New Data Source

- Bước 2: Chọn *Create a data source based on an existing or new connection*.



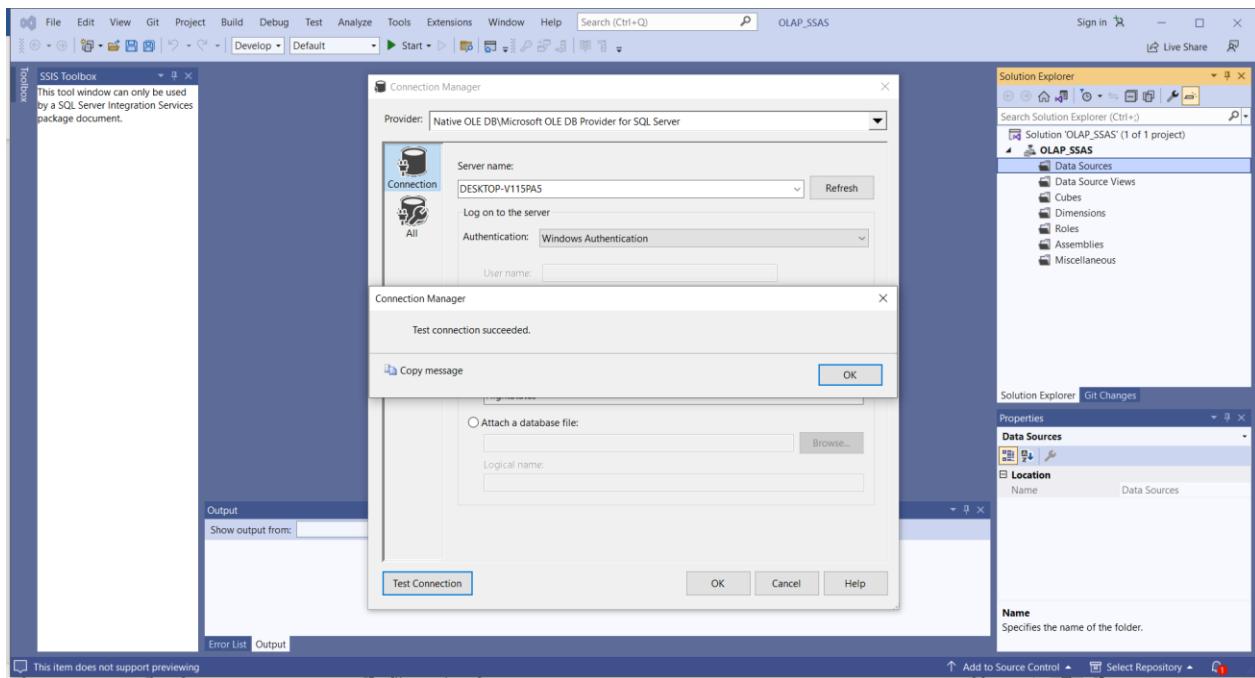
Hình 3.4. Create a data source based on existing on new connection

- Bước 3: Click New, sau đó điền thông tin của Connection sau đó Test Connection



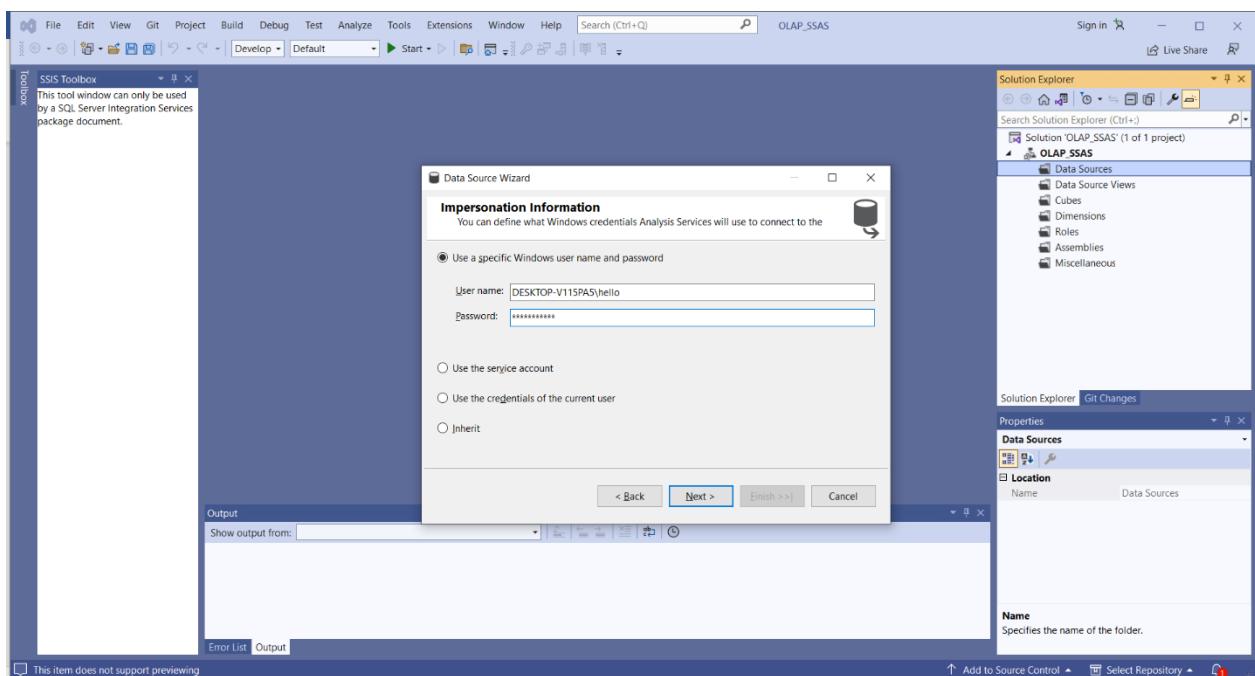
Hình 3.5. Điền thông tin Connection Manager

IS217 – Kho dữ liệu và OLAP



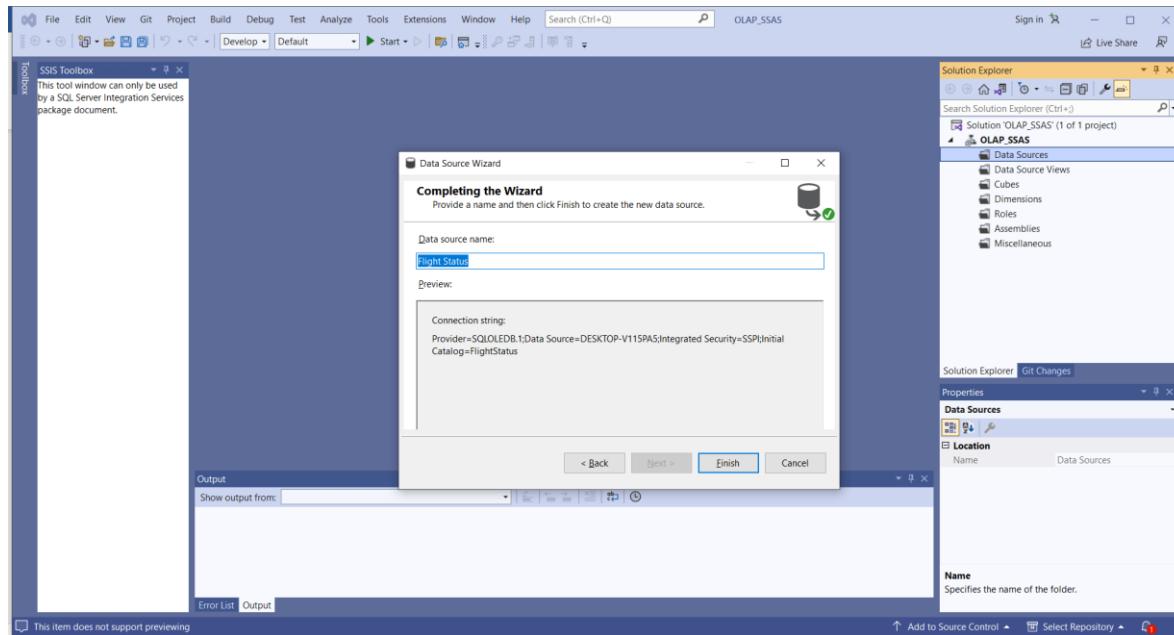
Hình 3.6. Test Connection

- Bước 4: Chọn Use a specific Windows username and password, sau đó nhập thông tin.



Hình 3.7. Use a specific Windows user name and password

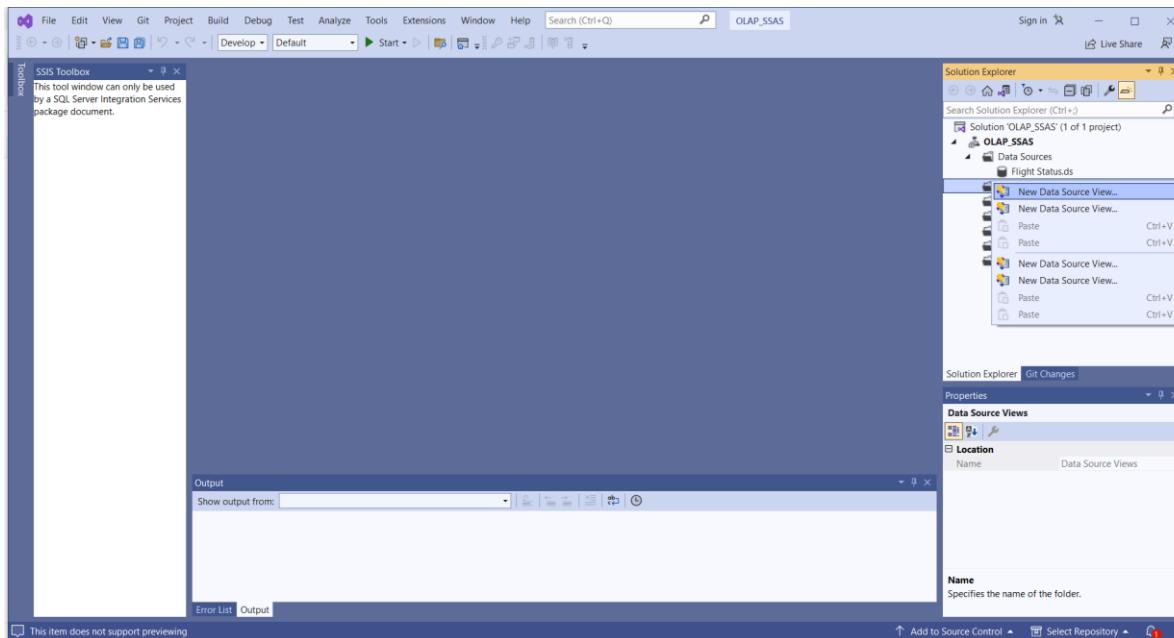
- Bước 5: Chọn Next sau đó chọn Data source và chọn Finish để hoàn tất.



Hình 3.8. Chọn Data Source và Finish

3.3. Tạo Data Source View

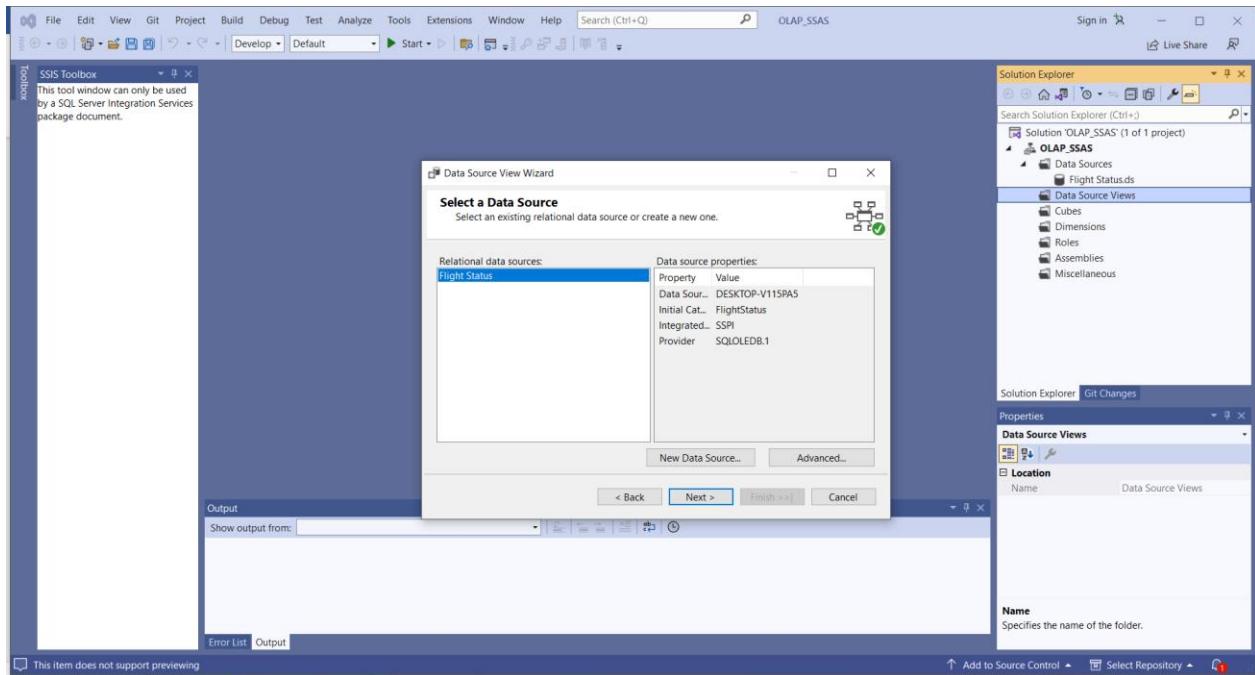
- Bước 1: Bên góc phải phần Solution Explorer, click chuột phải Data Source View → New Data Source View → Next.



Hình 3.9. New Data Source View

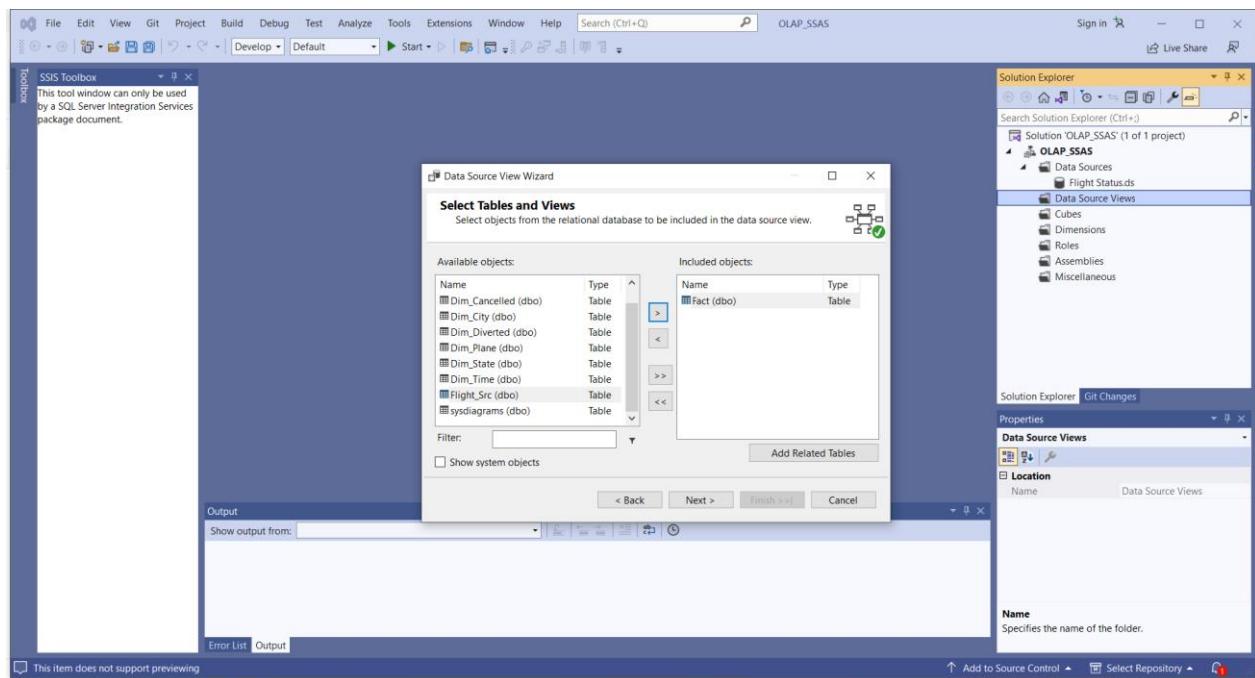
IS217 – Kho dữ liệu và OLAP

- Bước 2: Kết nối với Database đã kết nối trước rồi nhấn Next để tiếp tục.



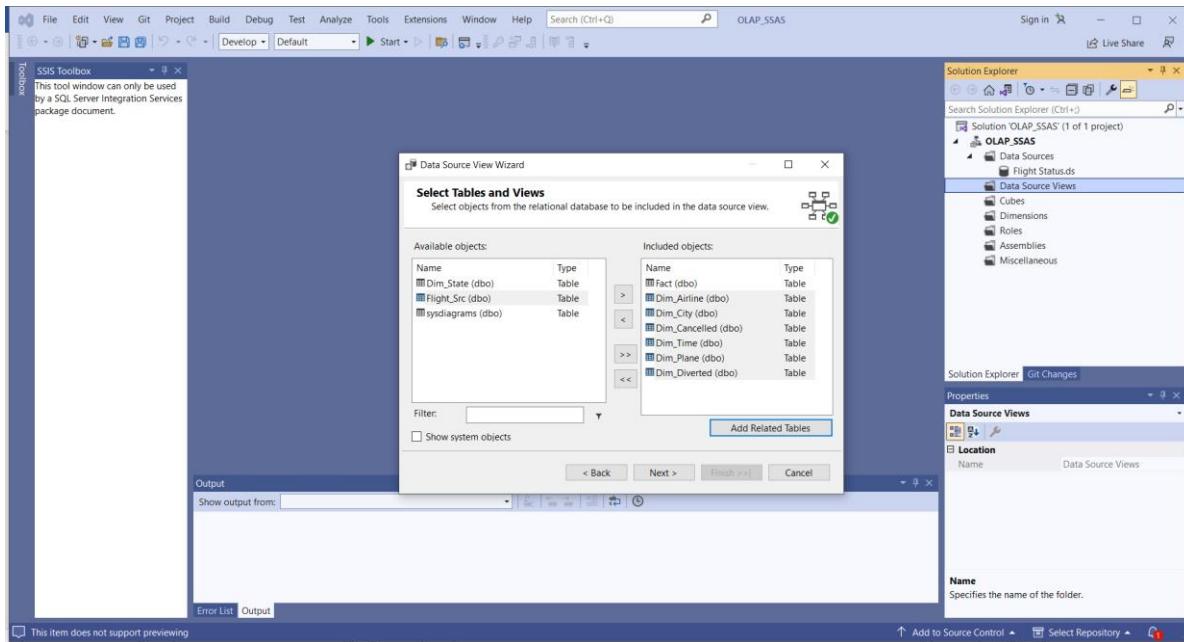
Hình 3.10. Kết nối với Database

- Bước 3: Chọn bảng Fact sau đó click “>”



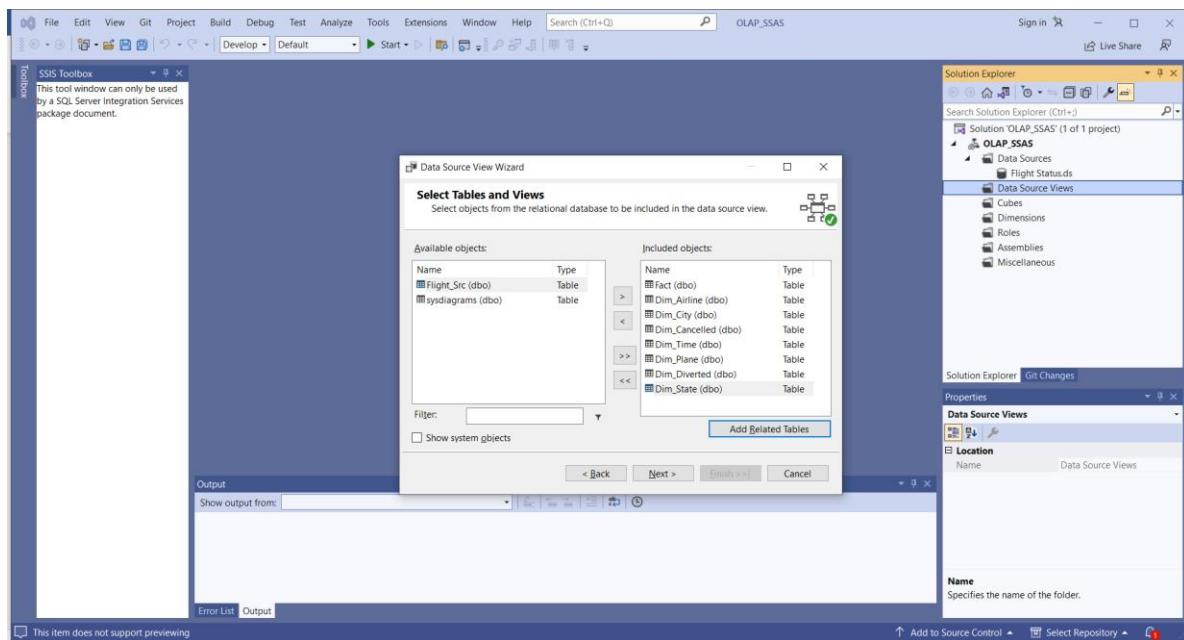
Hình 3.11. Chọn bảng Fact

Bấm Add Related Tables để đưa các Dim có quan hệ khóa ngoại với bảng Fact qua.



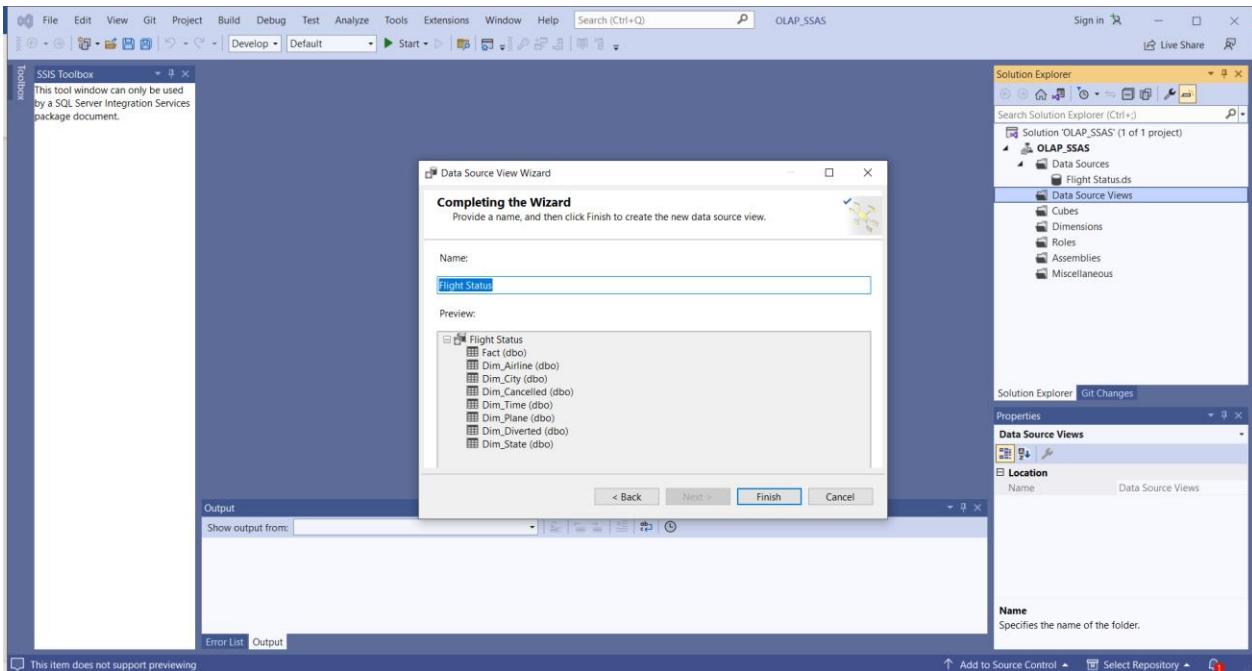
Hình 3.12. Chọn Add Related Tables

Bấm thêm một lần nữa để các bảng Dim có quan hệ với các Dim đã qua ban nãy qua luôn. Sau đó bấm next.



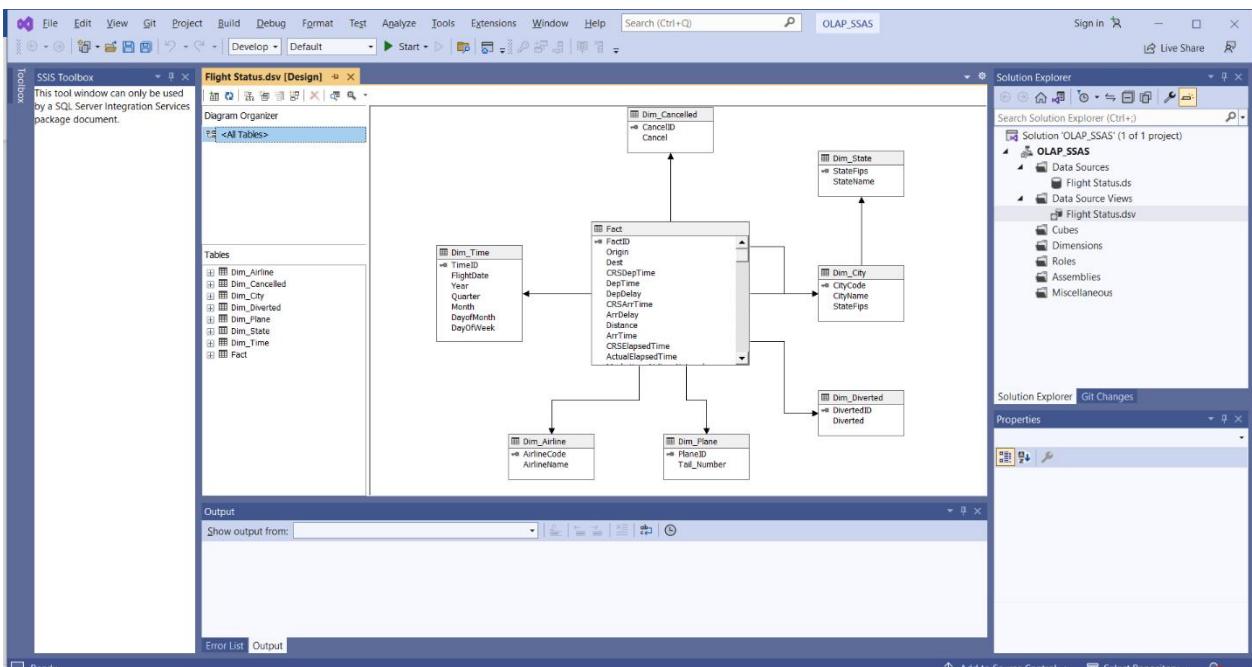
Hình 3.13. Bấm lại Add Related Tables

- Bước 4: Kiểm tra Name và các bảng, sau đó nhấn Finish



Hình 3.14. Kiểm tra thông tin

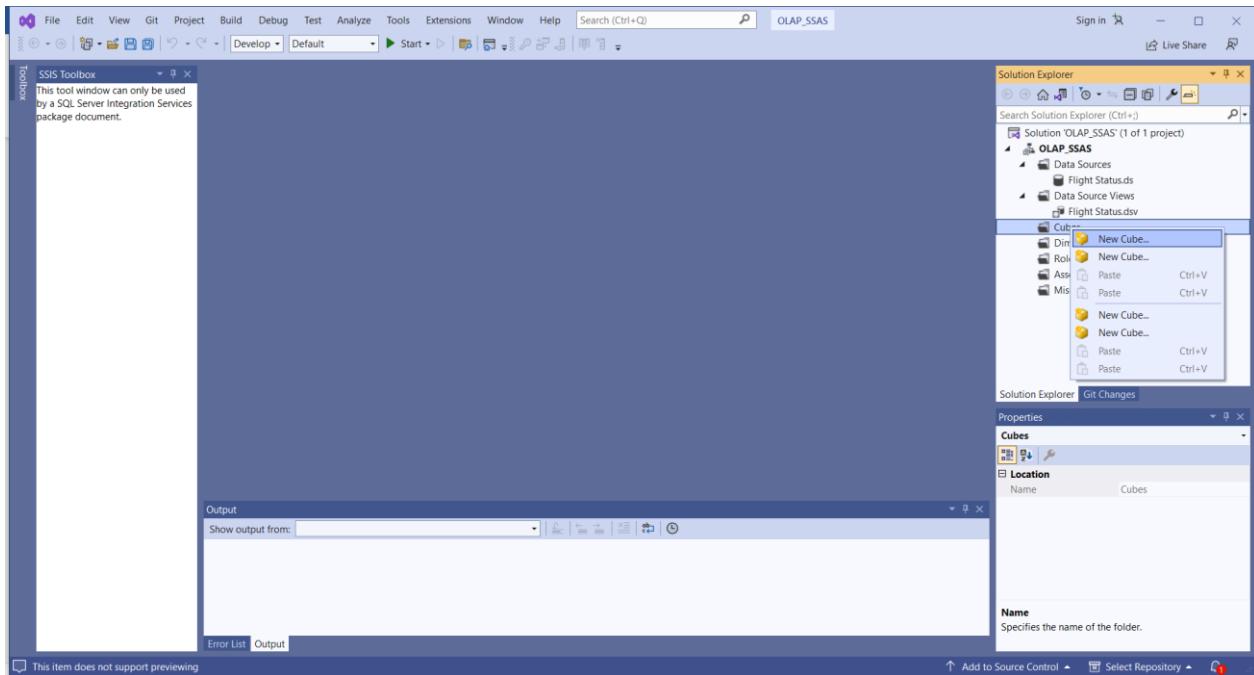
- Nhấn click chuột vào file data source view vừa tạo ta được kết quả



Hình 3.15. Kết quả Data Source View vừa tạo

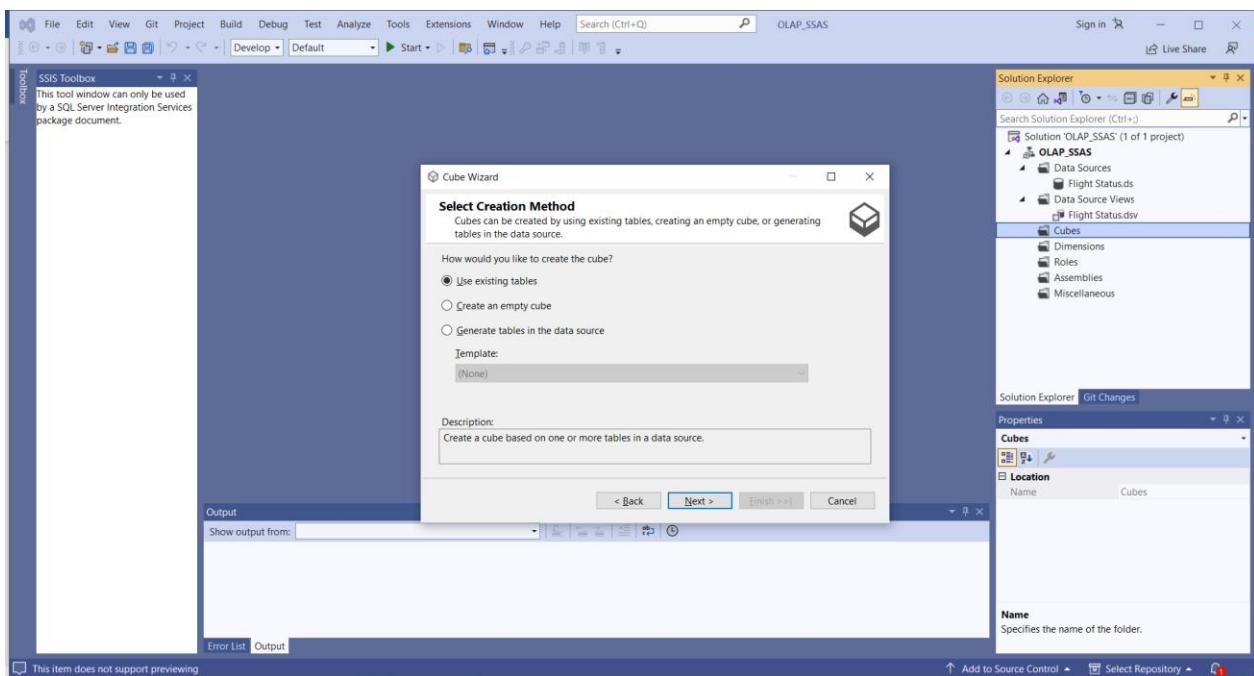
3.4. Tạo Cube và dimensions

- Bước 1: Chọn Cube sau đó chọn tiếp New Cubes → Next



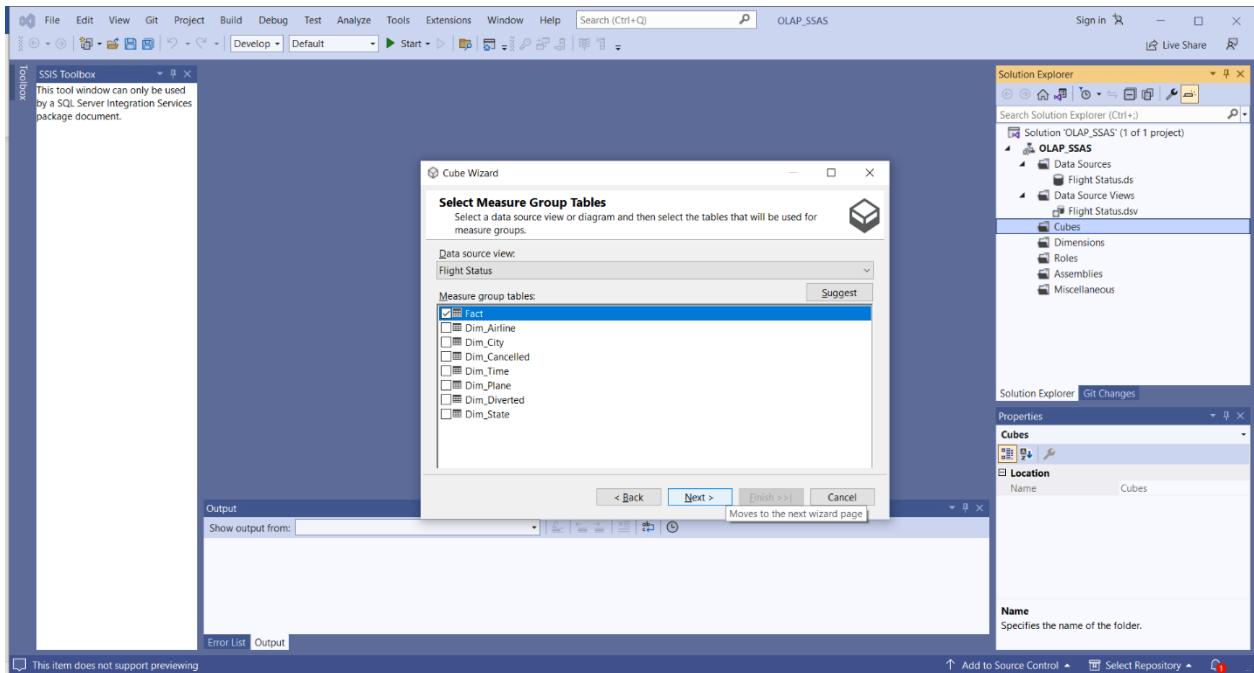
Hình 3.16. New Cube

- Bước 2: Chọn Use an existing table. Nhấn Next

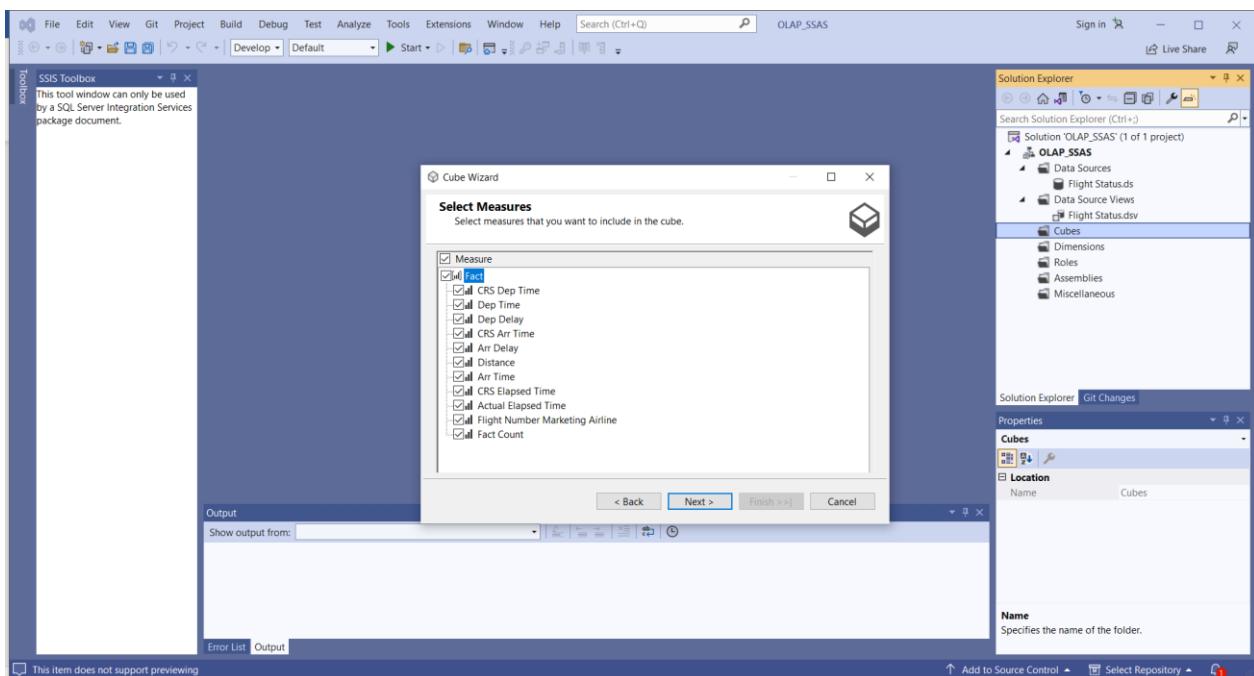


Hình 3.17. Use an existing table

- Bước 3: Chọn Measure group table là bảng Fact và nhấn Next



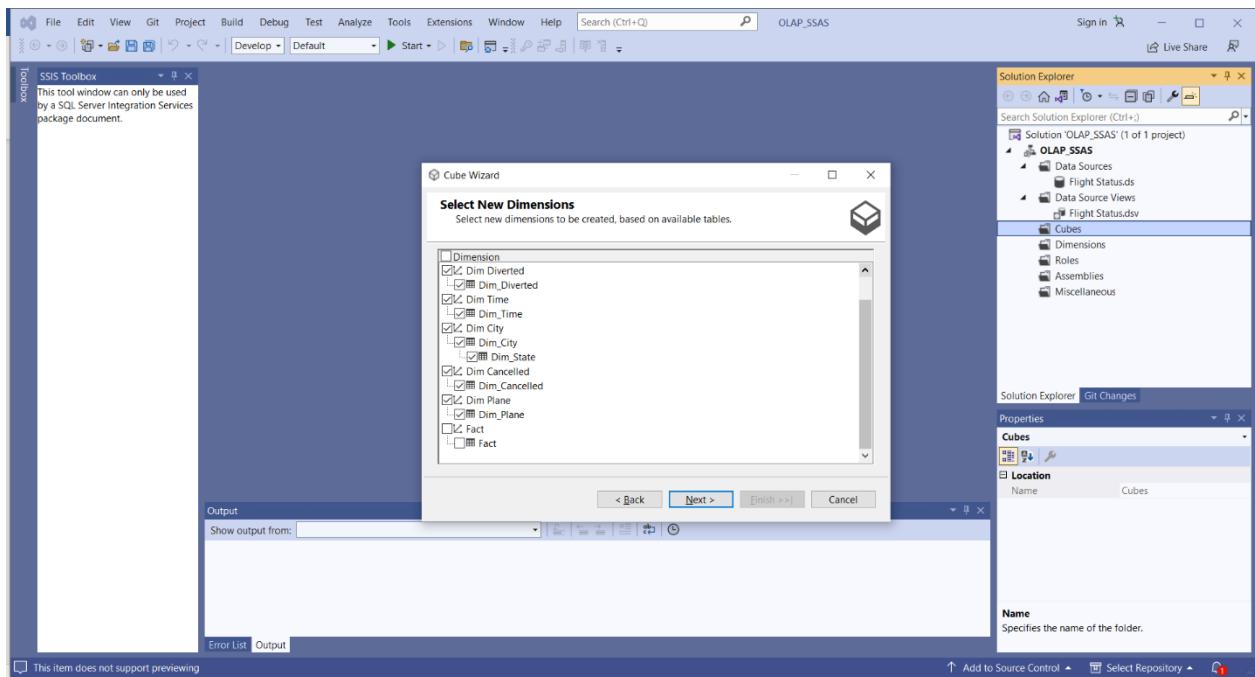
Hình 3.18. Chọn Measure table



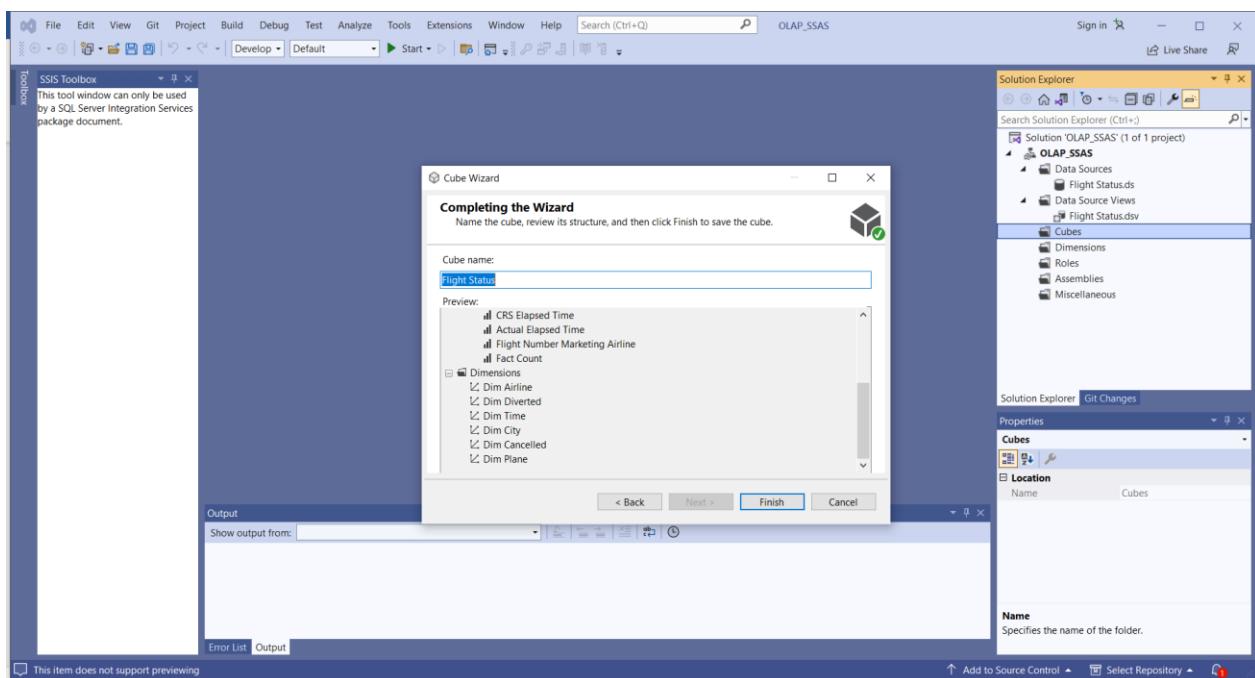
Hình 3.19. Measure table

IS217 – Kho dữ liệu và OLAP

- Bước 4: Chọn những Dimension cần add vào sau đó chọn Next. Kiểm tra và nhấn finish.

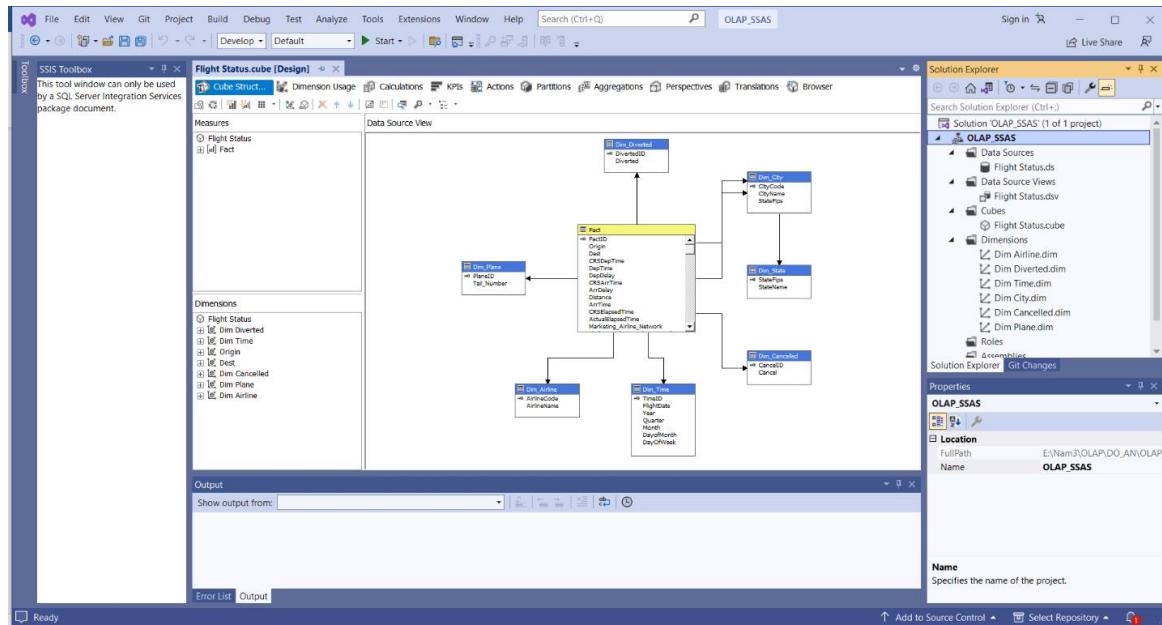


Hình 3.20. Chọn Dimension



Hình 3.21. Kiểm tra thông tin và nhấn finish

- Bước 5: Sau đó các bảng Dim sẽ được tạo tự động.

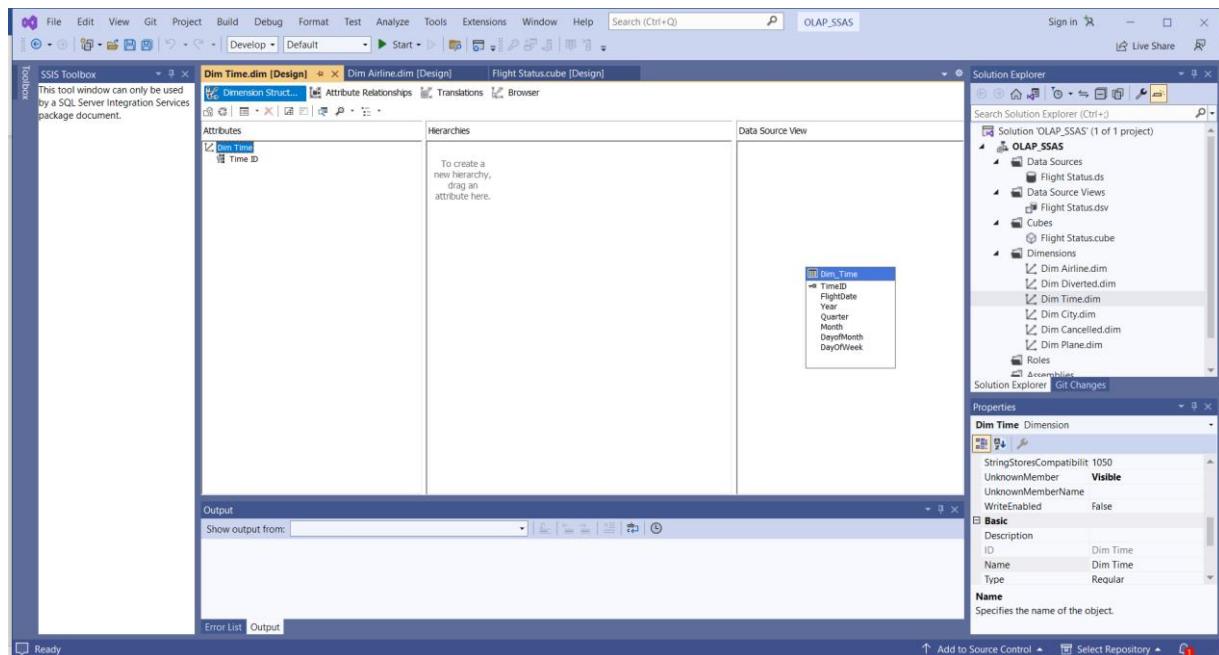


Hình 3.22. Hiển thị sơ đồ

3.5. Thao tác trên các bảng Dim

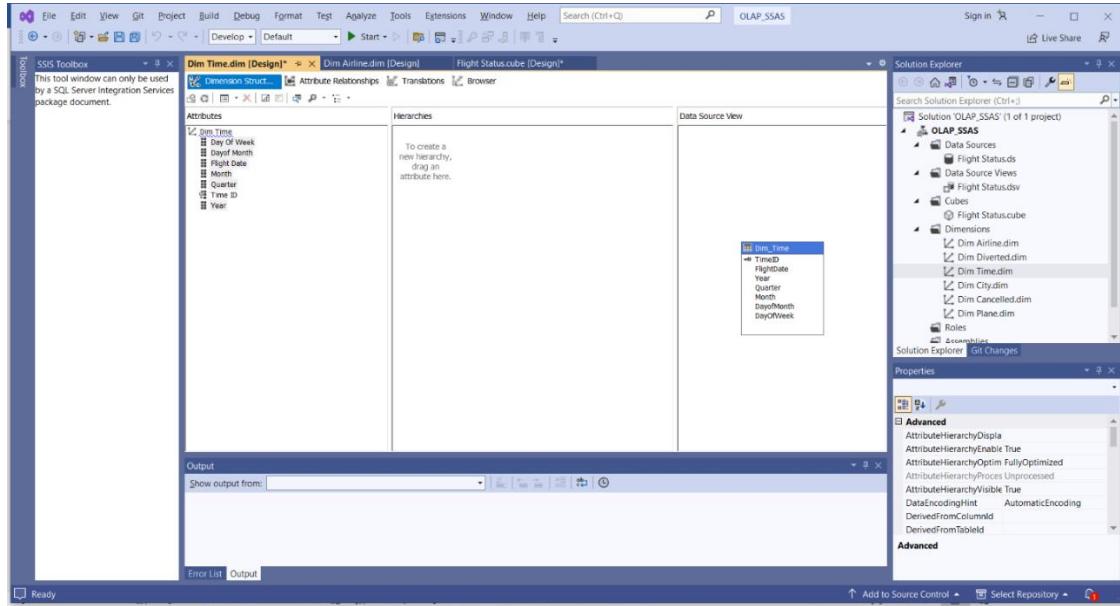
3.5.1. Bảng Dim_Time

- Bước 1: Chọn bảng Dim Time.dim từ khung Solution Explorer



Hình 3.23. Chọn Dim Time.dim

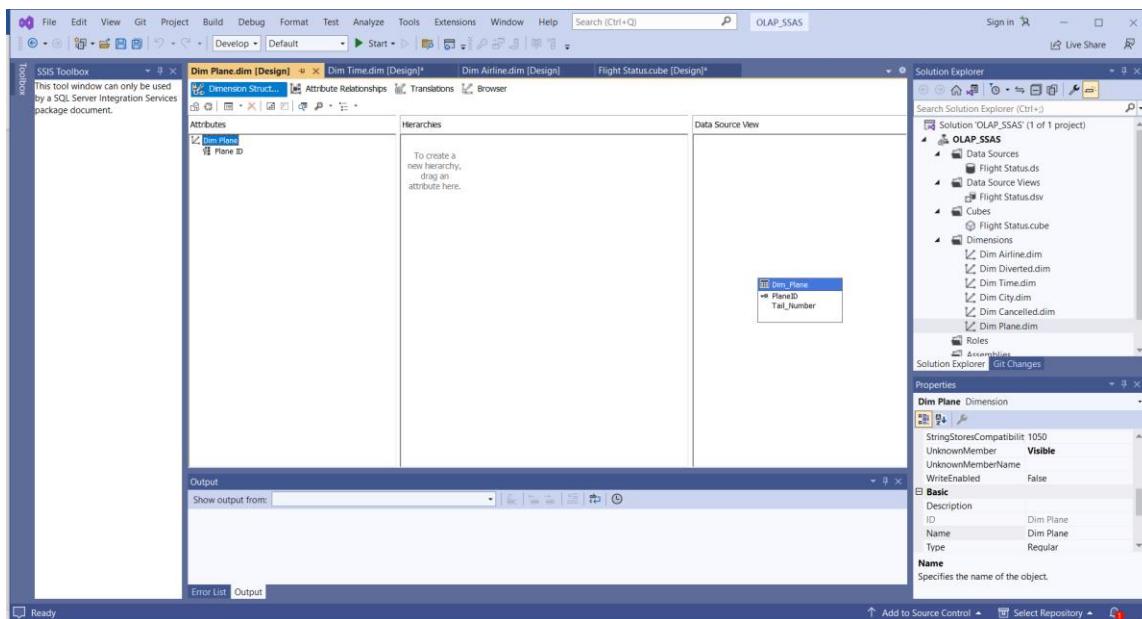
- Bước 2: Chọn những thuộc tính của bảng Dim_Time chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.24. Kéo các thuộc tính Dim_Time

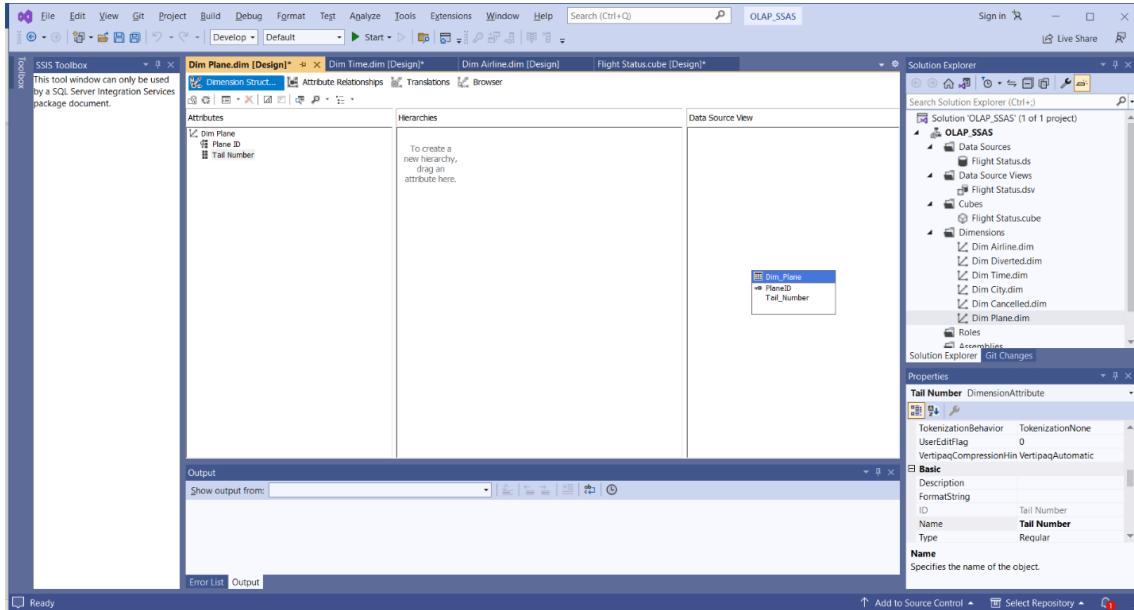
3.5.2. Bảng Dim_Plane

- Bước 1: Chọn bảng Dim Plane.dim từ khung Solution Explorer



Hình 3.25. Chọn Dim_Plane.dim

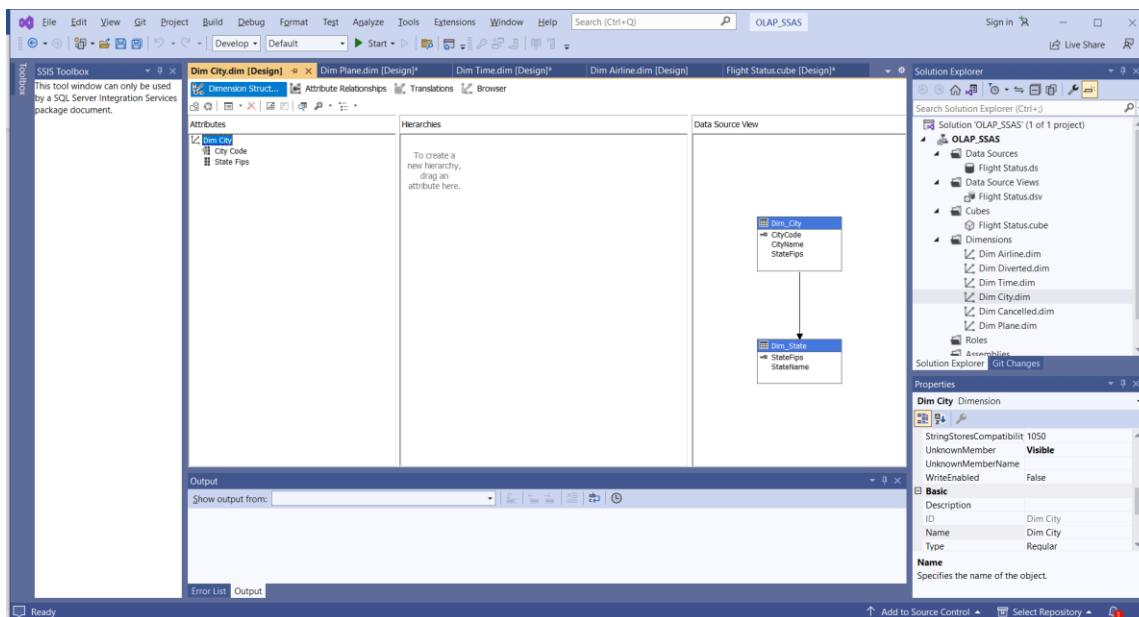
- Bước 2: Chọn những thuộc tính của bảng Dim_Plane chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.26. Kéo các thuộc tính Dim_Plane

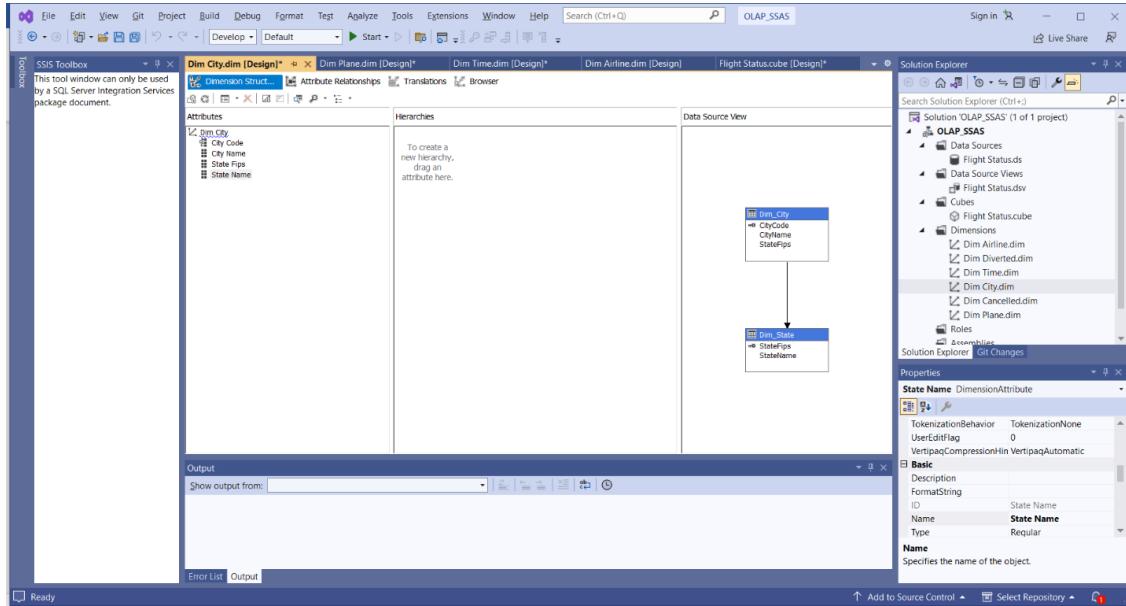
3.5.3. Bảng Dim_City

- Bước 1: Chọn bảng Dim City.dim từ khung Solution Explorer



Hình 3.27. Chọn Dim City.dim

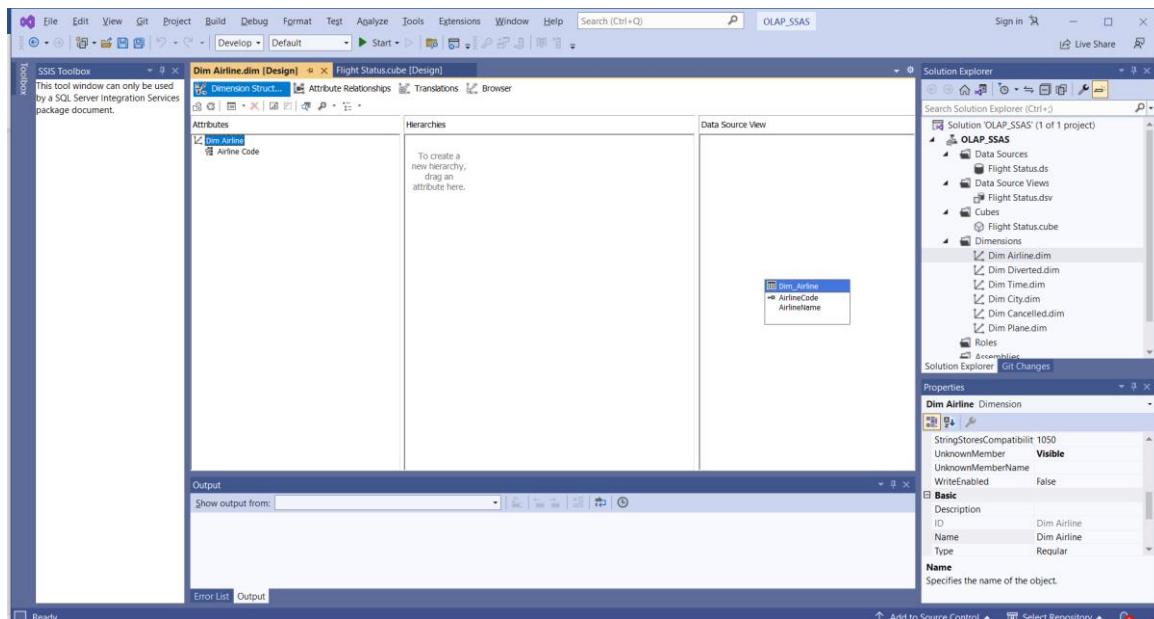
- Bước 2: Chọn những thuộc tính của bảng Dim_City và Dim_State chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.28. Kéo các thuộc tính Dim_City và Dim_State

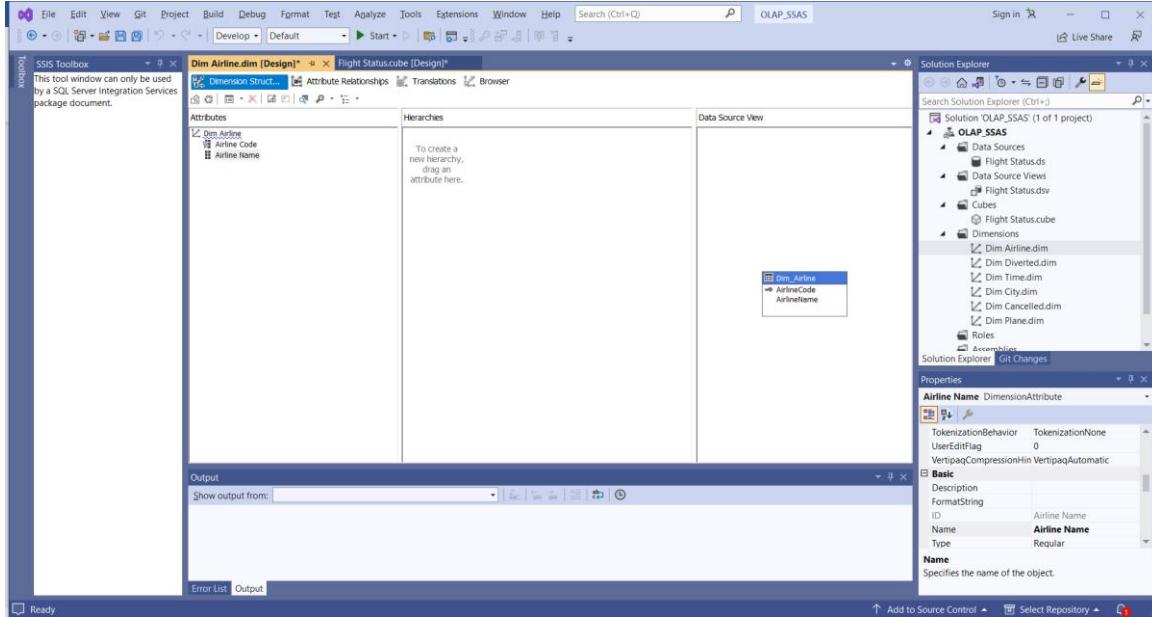
3.5.4. Bảng Dim_Airline

- Bước 1: Chọn bảng Dim Airline.dim từ khung Solution Explorer



Hình 3.29. Chọn Dim Airline.dim

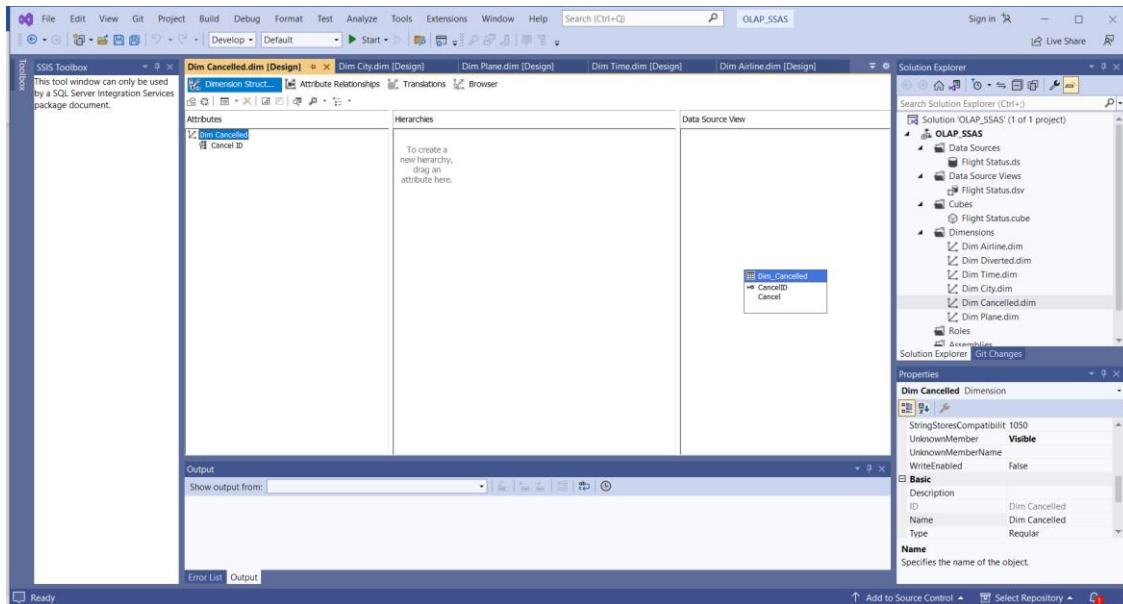
- Bước 2: Chọn những thuộc tính của bảng Dim_Airline chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.30. Kéo các thuộc tính Dim_Airline

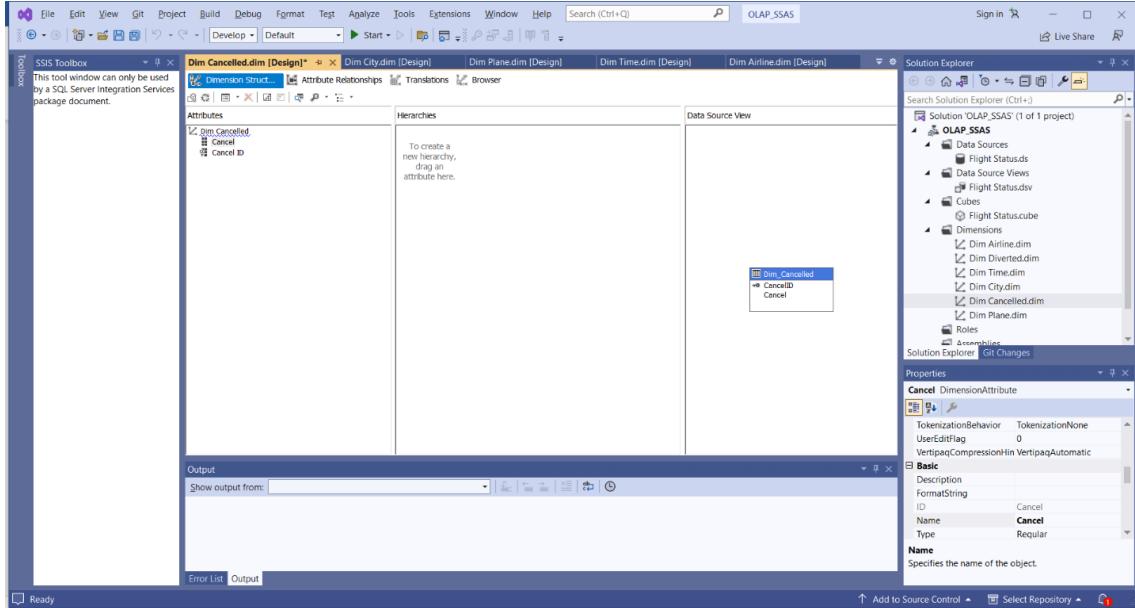
3.5.5. Bảng Dim_Cancelled

- Bước 1: Chọn bảng Dim Cancelled.dim từ khung Solution Explorer



Hình 3.31. Chọn Dim_Cancelled

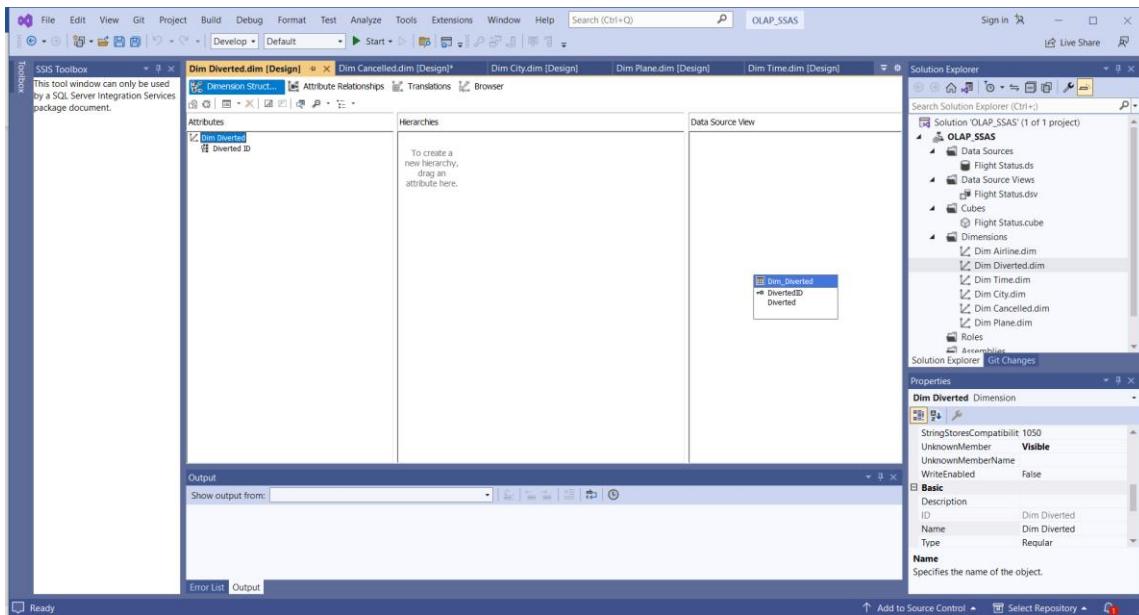
- Bước 2: Chọn những thuộc tính của bảng Dim_Cancelled chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.32. Kéo các thuộc tính Dim_Cancelled

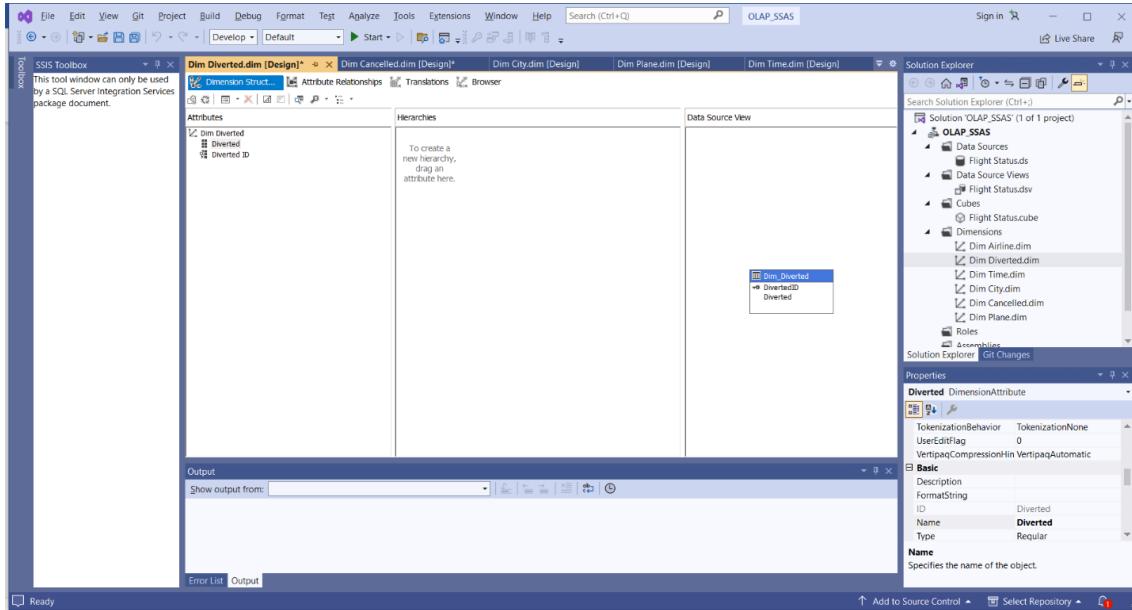
3.5.6. Bảng Dim_Diverted

- Bước 1: Chọn bảng Dim Diverted.dim từ khung Solution Explorer



Hình 3.33. Chọn Dim Diverted.dim

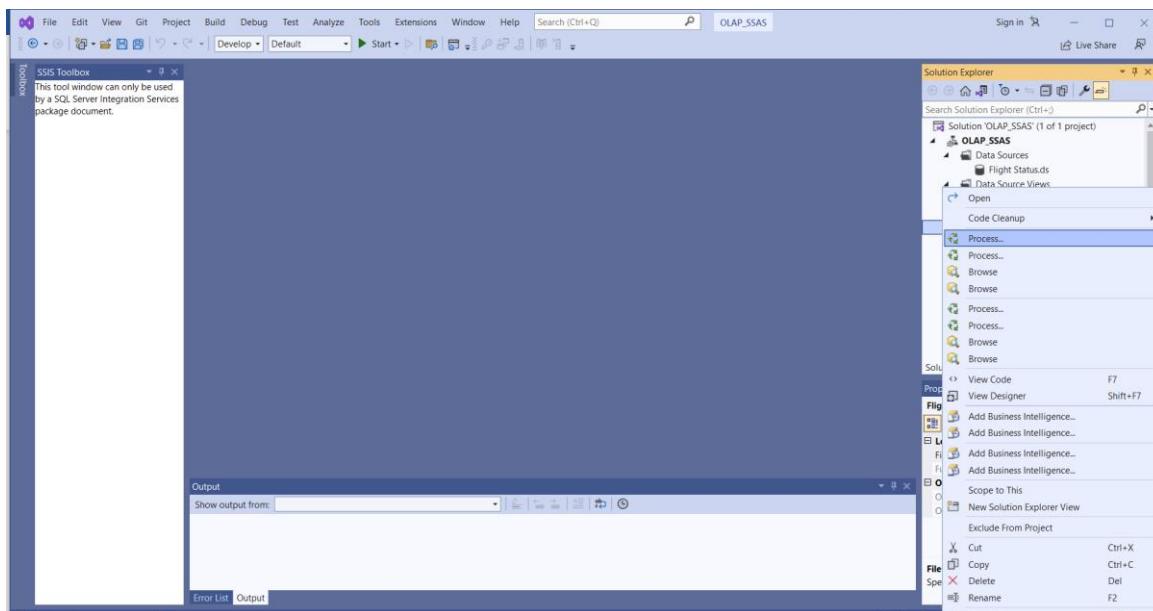
- Bước 2: Chọn những thuộc tính của bảng Dim_Diverted chưa xuất hiện ở cột Attributes. Nhấn giữ chuột trái và kéo từ cột Data Source View sang cột Attributes.



Hình 3.34. Kéo các thuộc tính Dim_Diverted

3.6. Deploy và process project.

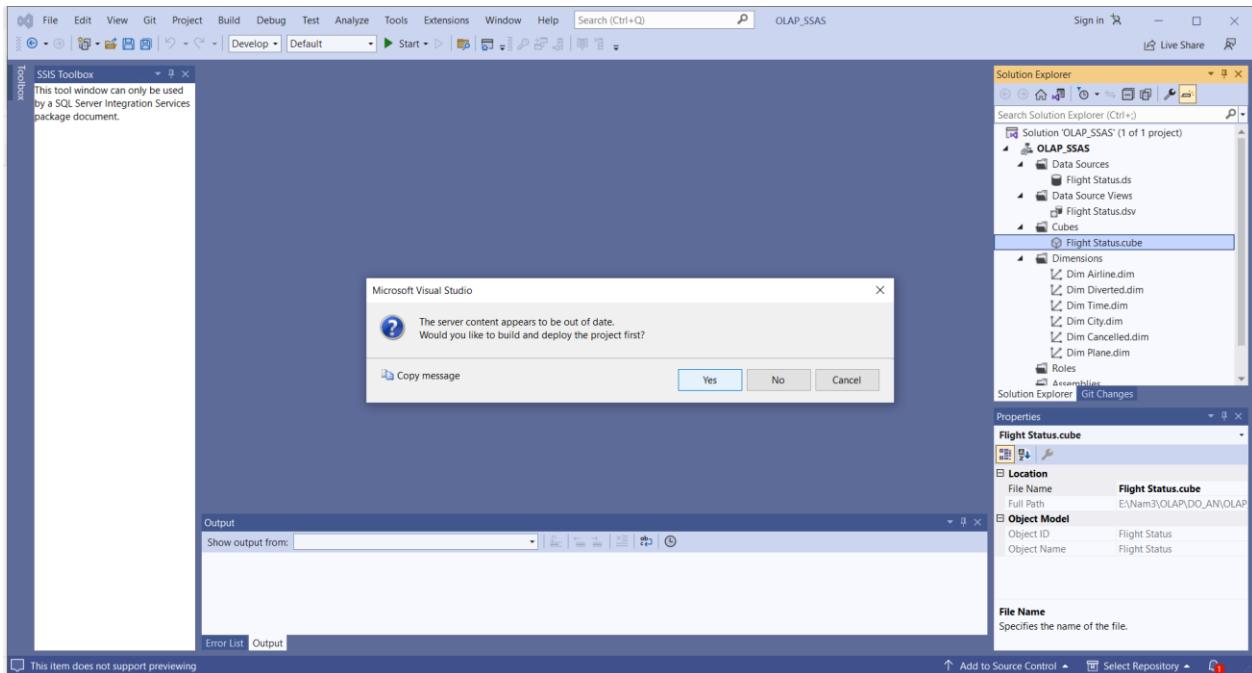
- Bước 1: Tại Cubes, chuột phải vào Flight Status(cube) chọn Process



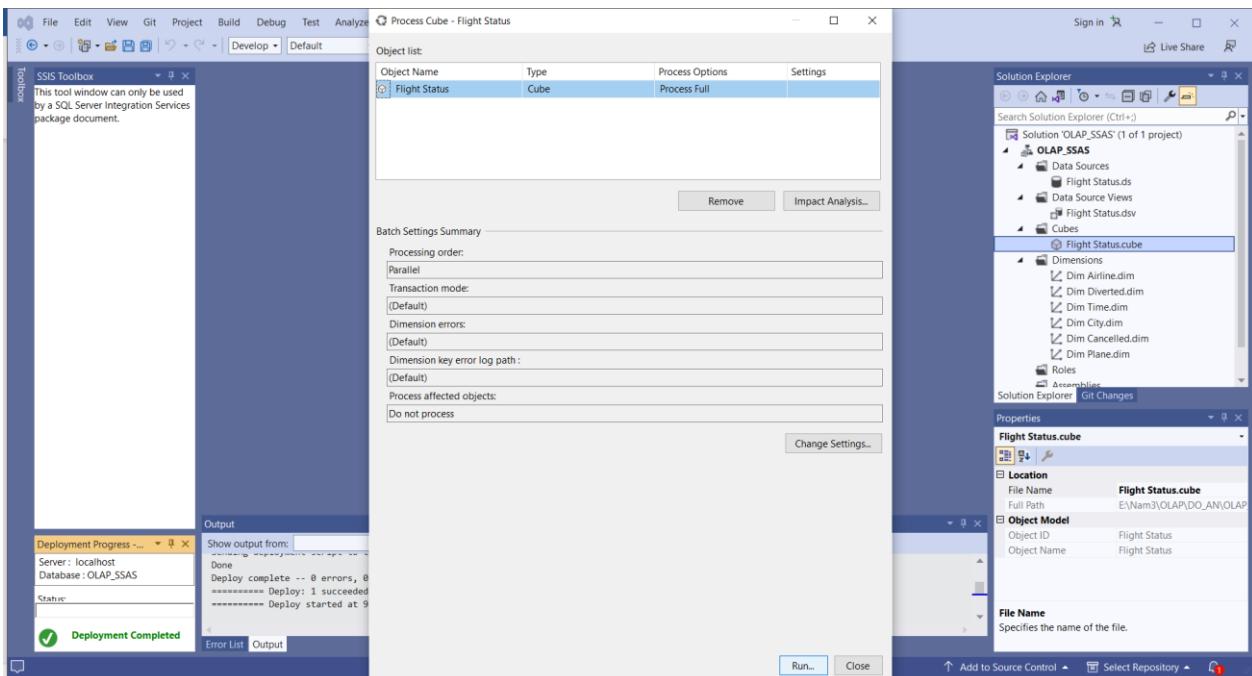
Hình 3.35. Thực hiện Process

IS217 – Kho dữ liệu và OLAP

- Bước 2: Chọn Yes. Sau đó chọn Run để bắt đầu thực hiện.

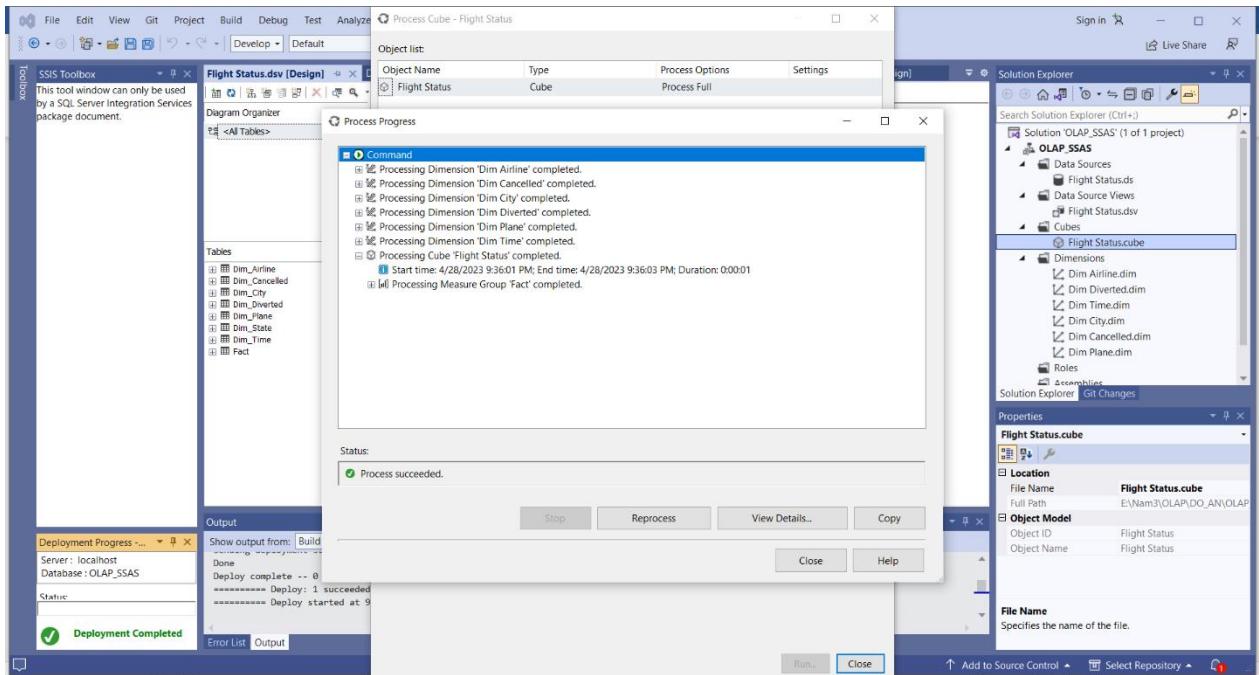


Hình 3.36. Chọn Yes



Hình 3.37. Chọn Run

- Bước 3: Sau khi chạy thành công án Close để hoàn tất



Hình 3.38. Thông báo kết quả

3.7. Thực hiện các câu truy vấn dữ liệu

3.7.1. Thống kê top 5 thành phố ở bang California có số lượng chuyến bay đến nhiều nhất.

SSAS

- Bước 1: Kéo thuộc tính City Name và Fact Count từ Origin và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dest cho Dimension, Dest.State Name cho Hierarchy, Equal cho Operator và California cho Filter Expression.
- Bước 3: Ở tab Calculation tạo thêm 1 Script Organizer có tên TOP_5_CITY

IS217 – Kho dữ liệu và OLAP

```
generate(
    drilldown([Dest].[StateName]),
    topcount(
        [Dest].[CityName].Children, 5,
        [Measures].[Fact Count]
    )
)
No issues found
```

Hình 3.39. SSAS - TOP_5_CITY

- Bước 4: Chọn Click to execute the query.

City Name	Fact Count
Los Angel...	3551
Sacrame...	1012
San Dieg...	1500
San Fran...	2435
San Jose...	983

Hình 3.40. SSAS – Kết quả

MDX

- Query:

```
SELECT [Measures].[Fact Count] ON 0,
TOPCOUNT(
    [Origin].[City Name].CHILDREN,
    5,
    [Origin].[City Code]
) ON 1
FROM [Flight Status]
WHERE [Origin].[State Name].&[California]
```

- Kết quả:

The screenshot shows the MDXQuery1.mdx query editor interface. On the left, the cube 'Flight Status' is selected. The 'Measure Group' dropdown is set to '<All>'. The 'Origin' dimension is expanded, showing 'Origin.City Code', 'Origin.City Name', 'Origin.State Fips', and 'Origin.State Name'. The main pane displays the following MDX query:

```

SELECT [Measures].[Fact Count] ON 0,
TOPCOUNT(
[Origin].[City Name].CHILDREN,
5,
[Origin].[City Code]
) ON 1
FROM [Flight Status]
WHERE [Origin].[State Name].&[California]
  
```

The results pane shows a table titled 'Fact Count' with the following data:

	Fact Count
Los Angeles, CA	3521
San Francisco, CA	2410
San Diego, CA	1530
San Jose, CA	1002
Sacramento, CA	1006

A green status bar at the bottom indicates: 'Query executed successfully.'

Hình 3.41. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn City Name, ở Filter chọn Dest.State Name.
- Bước 3: Chọn Value Filter, chọn Top 5

IS217 – Kho dữ liệu và OLAP

	A	B	C	D	E	F	G	H	I	J
1	Dest.State Name	California								
2										
3	Row Labels	Fact Count								
4	Los Angeles, CA	3551								
5	San Francisco, CA	2435								
6	San Diego, CA	1500								
7	Sacramento, CA	1012								
8	San Jose, CA	983								
9	Grand Total	9481								
10										
11										
12										
13										
14										
15										
16										
17										
18										

Hình 3.42. Kết quả thực hiện Excel

Power BI

- Bước 1: Click vào State Name, City Name, Fact Count
- Bước 2: Filter top 5 ở City Name, Sort ở Fact Count

State Name	City Name	Fact Count
California	Los Angeles, CA	3521
California	San Francisco, CA	2410
California	San Diego, CA	1530
California	Sacramento, CA	1006
California	San Jose, CA	1002
Total		1002 → 69

Hình 3.43. Kết quả thực hiện Power BI

3.7.2. Thống kê tổng các chuyến bay của hãng hàng không JetBlue Airways bị điều hướng vào ngày 27-04-2022.

SSAS

- Bước 1: Kéo thuộc tính Airline Name và Fact Count từ Origin và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dim Diverted, Dim Time, Dim Airline cho Dimension.
- Bước 3: Bấm Click to execute the query.

Airline Name	Fact Count
JetBlue Air...	3

Hình 3.44. SSAS – Kết quả

MDX

- Query:

```

SELECT {[Measures].[Fact Count]} ON COLUMNS,
       [Dim Airline].[Airline Name].&[JetBlue Airways] ON ROWS
  FROM [Flight Status]
 WHERE ([Dim Time].[Flight Date].&[2022-04-27T00:00:00],
        [Dim Diverted].[Diverted].&[True])
    
```

- Kết quả:

IS217 – Kho dữ liệu và OLAP

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. On the left, the 'Cube' dropdown is set to 'Flight Status'. Below it, the 'Measure Group' dropdown shows '<All>'. The main pane displays an MDX query:

```
SELECT {[Measures].[Fact Count]} ON COLUMNS,
       [Dim Airline].[Airline Name].&[JetBlue Airways] ON ROWS
  FROM [Flight Status]
 WHERE ([Dim Time].[Flight Date].&[2022-04-27T00:00:00],
        [Dim Diverted].[Diverted].&[True])
```

Below the query, the 'Results' tab is selected, showing the output:

Fact Count
JetBlue Airways 3

Hình 3.45. MDX - Kết quả

Excel (Pivot Table)

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right indicates the following fields:

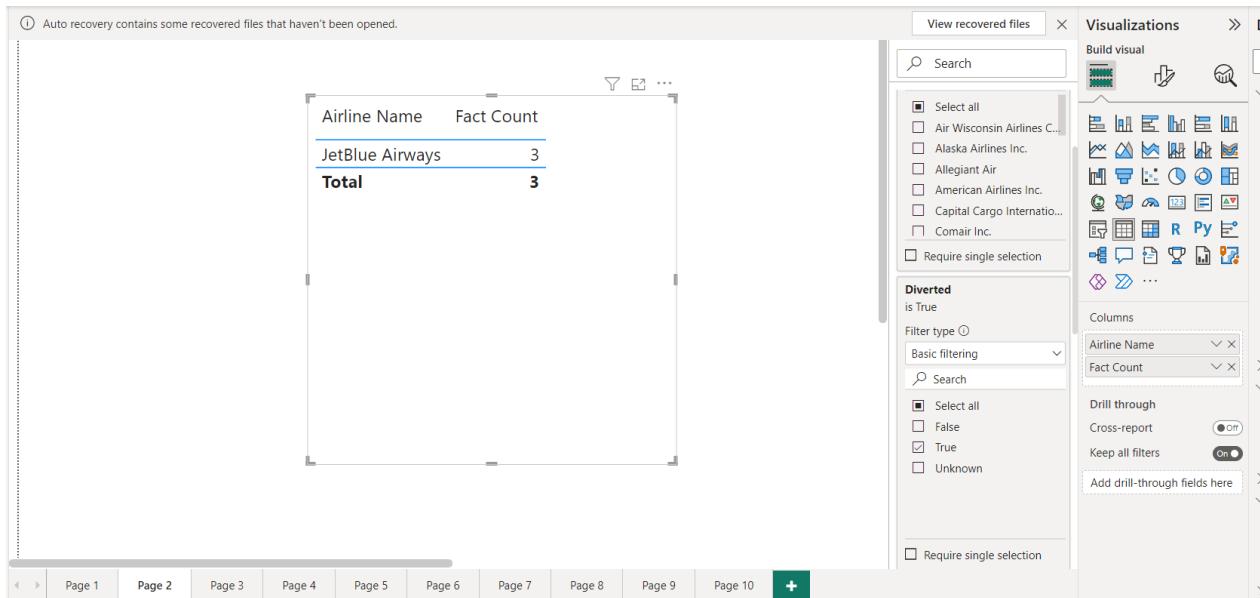
- Choose fields to add to report:
 - Search
 - Quarter
 - Time ID
 - Year
- Origin (selected)
- More Fields

The PivotTable itself has the following structure:

Row Labels	Fact Count
JetBlue Airways	3
Grand Total	3

Hình 3.46. Kết quả thực hiện Excel

Power BI



Hình 3.47. Kết quả thực hiện Power BI

3.7.3. Thông kê số lượng các chuyến bay bị hủy của hãng hàng không Southwest Airlines Co. từ ngày 24-04-2022 đến 27-04-2022 theo từng ngày.

SSAS

- Bước 1: Kéo thuộc tính Flight Date và Fact Count từ Dim Time và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dim Cancelled, Dim Airline, Dim Time cho Dimension, chọn Cancel, Airline Name, Flight Date cho Hierarchy, Equal cho Operator và Range cho Filter Expression.
- Bước 3: Chọn Click to execute the query.

The screenshot shows the Microsoft Analysis Services (SSAS) Management Studio interface. On the left, there's a navigation pane with 'Flight Status' and 'Metadata' sections, and a 'Calculated Members' section below. The main area has a 'Dimension' table with columns: Dimension, Hierarchy, Operator, Filter Expression, and Parameters. It shows filters for 'Dim Cancelled' (Hierarchy: 'Cancel', Operator: Equal, Filter Expression: '{ True }'), 'Dim Airline' (Hierarchy: 'Airline Name', Operator: Equal, Filter Expression: '{ Southwest Airlines Co. }'), and 'Dim Time' (Hierarchy: 'Flight Date', Operator: Range (Inclusive), Filter Expression: '2022-04-24 00:00:00 : 2022-04-27 00:00:00'). Below this is a results grid titled 'Flight Date' with columns 'Flight Date' and 'Fact Count'. The data shows four rows: 2022-04-24 00:00:00 (Fact Count: 16), 2022-04-25 00:00:00 (Fact Count: 95), 2022-04-26 00:00:00 (Fact Count: 9), and 2022-04-27 00:00:00 (Fact Count: 3).

Hình 3.48. SSAS - Câu truy vấn

MDX

- Query:

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Time].[Flight Date].[Flight Date].ALLMEMBERS ) }
ON ROWS
FROM ( SELECT ( [Dim Time].[Flight Date].&[2022-04-24T00:00:00] :
[Dim Time].[Flight Date].&[2022-04-27T00:00:00] )
ON COLUMNS FROM ( SELECT ( { [Dim Airline].[Airline Name].&[Southwest
Airlines Co.] } )
ON COLUMNS FROM ( SELECT ( { [Dim Cancelled].[Cancel].&[True] } )
ON COLUMNS FROM [Flight Status])) )
WHERE ( [Dim Cancelled].[Cancel].&[True], [Dim Airline].[Airline
Name].&[Southwest Airlines Co.] )

```

- Kết quả:

IS217 – Kho dữ liệu và OLAP

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. On the left, the 'Cube' pane displays a cube named 'Flight Status' with various dimensions like Flight Status, Measures, KPIs, Dest, Dim Airline, Dim Cancelled, Dim Diverted, Dim Plane, Dim Time, and Origin. The 'Measure Group' dropdown is set to '<All>'. The main area shows an MDX query:

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Time].[Flight Date].[Flight Date].ALLMEMBERS ) }
ON COLUMNS FROM ( SELECT ( [Dim Time].[Flight Date].&[2022-04-24T00:00:00] : [Dim Time].[Flight Date].&[2022-04-27T00:00:00] )
ON COLUMNS FROM ( SELECT ( { [Dim Airline].[Airline Name].&[Southwest Airlines Co.] } )
ON COLUMNS FROM ( SELECT ( { [Dim Cancelled].[Cancel].&[True] } )
ON COLUMNS FROM [Flight Status])))
WHERE ( [Dim Cancelled].[Cancel].&[True], [Dim Airline].[Airline Name].&[Southwest Airlines Co.] )
```

Below the query, the 'Messages' tab shows the results of the execution:

	Fact Count
2022-04-24 00:00:00	16
2022-04-25 00:00:00	95
2022-04-26 00:00:00	9
2022-04-27 00:00:00	3

Hình 3.49. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn Flight Date, ở Filter chọn Cancel và Airline Name.
- Bước 3: Click chọn giá trị cho Row Labels từ 24-04-2022 đến 27-04-2022.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right is configured as follows:

- Choose fields to add to report:** Flight Date (selected), Month, Quarter, Time ID, Year.
- Filters:** Cancel (selected), Airline Name (selected).
- Rows:** Flight Date (selected).
- Values:** Fact Count (selected).

The main table displays the following data:

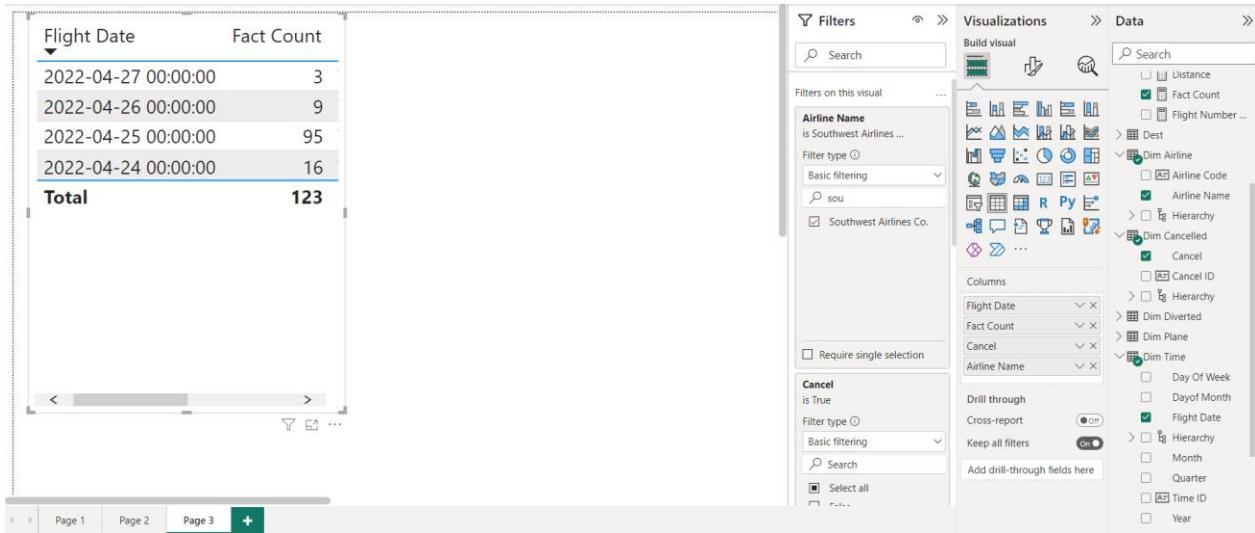
Flight Date	Fact Count
2022-04-24 00:00:00	16
2022-04-25 00:00:00	95
2022-04-26 00:00:00	9
2022-04-27 00:00:00	3
Grand Total	123

Hình 3.50. Kết quả thực hiện Excel

Power BI

- Bước 1: Click chọn Filght Date, Fact Count, Cancel, Airline

- Bước 2: Filter theo yêu cầu



Hình 3.51. Kết quả thực hiện Power BI

3.7.4. Thống kê các thành phố ở bang Florida có ít nhất trên 100 chuyến bay đến trong ngày 27-04-2022.

SSAS

- Bước 1: Kéo thuộc tính City Name, State Name và Fact Count từ Dest và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dest cho Dimension, Dest.State Name, Dest.City Name cho Hierarchy
- Bước 3: Tạo script cho City Name

```
FILTER(
    [Dest].[City Name].[City Name].ALLMEMBERS,
    [Measures].[Fact Count]>=100
)
```

- Bước 4: Chọn Click to execute the query.

IS217 – Kho dữ liệu và OLAP

The screenshot shows the Microsoft Analysis Services (SSAS) MDX query editor interface. On the left, there's a navigation pane with 'Flight Status' selected, followed by 'Metadata' and 'Search Model'. Below these are sections for 'Measure Group' and 'Calculated Members'. The main workspace contains an MDX query editor with tabs for 'Edit as Text' and 'Import...'. The query itself is:

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON  
EMPTY { ([Origin].[City Name].[City Name].ALLMEMBERS ) }  
ON ROWS FROM ( SELECT ( GENERATE(  
DrillDownLevel([Origin].[State Name].&[California]),  
  
ORDER(  
FILTER(  
[Origin].[City Name].[City Name].members,  
[Measures].[Fact Count]>=100  
,  
[Measures].[Fact Count], ASC  
)  
) ) ON COLUMNS FROM [Flight Status])
```

On the right side of the editor, there are sections for 'Dimension', 'Hierarchy', 'Operator', 'Filter Expression', and 'Parameters'. A 'Filter Expression' table is shown with rows for 'Dest', 'Dim Time', and 'Dest'. The 'Dest' row has 'Dest.State Name' as the hierarchy, 'Equal' as the operator, and '{ Florida }' as the filter expression. The 'Dim Time' row has 'Flight Date' as the hierarchy, 'Equal' as the operator, and '{ 2022-04-27 00:00:00 }' as the filter expression. The 'Dest' row has 'Dest.City Name' as the hierarchy, 'Custom' as the operator, and 'FILTER([Dest].[City Name].[City Name].MEMBERS, [Measur...' as the filter expression. Below the editor, a preview table titled 'City Name' shows fact counts for various cities in Florida:

City Name	Fact Count
Fort Laud...	239
Fort Myer...	111
Miami, FL	301
Orlando, FL	376
Tampa, FL	198

Hình 3.52. SSAS - Thực hiện truy vấn

MDX

- Query:

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON  
EMPTY { ([Origin].[City Name].[City Name].ALLMEMBERS ) }  
ON ROWS FROM ( SELECT ( GENERATE(  
DrillDownLevel([Origin].[State Name].&[California]),  
  
ORDER(  
FILTER(  
[Origin].[City Name].[City Name].members,  
[Measures].[Fact Count]>=100  
,  
[Measures].[Fact Count], ASC  
)  
) ) ON COLUMNS FROM [Flight Status])
```

- Kết quả:

IS217 – Kho dữ liệu và OLAP

The screenshot shows the Microsoft SQL Server Management Studio (SSMS) interface. On the left, the 'Cube' dropdown is set to 'Flight Status'. Below it, the 'Measure Group' dropdown shows '<All>'. The main pane displays an MDX query:

```
-- 4.3.7.4. Thống kê các thành phố ở bang Alabama có ít nhất trên 100 chuyến bay
-- khởi hành trong ngày 27-04-2022 sắp xếp tăng dần theo từng thành phố(slice and dice, drill down)
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY { ([Origin].[City Name].[City Name].ALLMEMBERS ) }
ON ROWS FROM ( SELECT ( GENERATE(
DrillDownLevel([Origin].[State Name].&[California]),
ORDER(
FILTER(
[Origin].[City Name].[City Name].members,
[Measures].[Fact Count])>=100
),
[Measures].[Fact Count], ASC
)
) ) ON COLUMNS FROM [Flight Status]
```

Below the query is a table titled 'Results' showing flight counts for various California cities:

City Name	Fact Count
Burbank, CA	605
Fresno, CA	204
Long Beach, CA	288
Los Angeles, CA	3521
Oakland, CA	843
Ontario, CA	462
Palm Springs, CA	343
Sacramento, CA	1006
San Diego, CA	1530
San Francisco, CA	2410
San Jose, CA	1002
Santa Ana, CA	874
Santa Barbara, CA	148

Hình 3.53. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn City Name, ở Filter chọn Dest.State Name và chọn Sort Largest to Smallest cho Fact Count.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right is configured as follows:

- Fields chosen for report: Dest.State Name, Flight Date, City Name, Fact Count.
- Filters: Dest.State Name (set to Florida), Flight Date (set to 2022-04-27 00:00:00).
- Rows: City Name.
- Values: Fact Count.

The PivotTable data is as follows:

Dest.State Name	Flight Date	Row Labels	Fact Count
Florida	2022-04-27 00:00:00	Orlando, FL	376
		Miami, FL	301
		Fort Lauderdale, FL	239
		Tampa, FL	198
		Fort Myers, FL	111
		Grand Total	1225

Hình 3.54. Kết quả thực hiện Excel

Power BI

The screenshot shows a Power BI interface with a table and a filter pane.

Table Data:

State Name	City Name	Fact Count	Flight Date
Florida	Fort Lauderdale, FL	239	2022-04-27 00:00:00
Florida	Fort Myers, FL	111	2022-04-27 00:00:00
Florida	Miami, FL	301	2022-04-27 00:00:00
Florida	Orlando, FL	376	2022-04-27 00:00:00
Florida	Tampa, FL	198	2022-04-27 00:00:00
Total		1225	

Filter Pane:

- Filters on this visual:**
 - City Name is (All)
 - Fact Count is greater than 100
 - Flight Date is 2022-04-27 00:00:00
 - State Name is Florida
- Columns:** State Name, City Name, Fact Count, Flight Date
- Drill through:** Cross-report (Off), Keep all filters
- Add data fields here**
- Filters on this page:**

Hình 3.55. Kết quả thực hiện Power BI -1

The screenshot shows a Power BI interface with a pie chart and a filter pane.

Pie Chart Data:

State Name	Fact Count	Percentage
Florida	198	(16.16%)
Orlando, FL	376	(30.69%)
Miami, FL	301	(24.57%)
Fort Lauderdale, FL	239	(19.51%)
Fort Myers, FL	111	(9.06%)

Filter Pane:

- Filters on this visual:**
 - City Name is (All)
 - Fact Count is greater than 100
 - Flight Date is 2022-04-27 00:00:00
 - State Name is Florida
- Legend:** State Name (Florida)
- Values:** Fact Count
- Details:** City Name
- Toolips:** Add data fields here
- Drill through:** Cross-report (Off), Keep all filters
- Add data fields here**
- Filters on this page:**

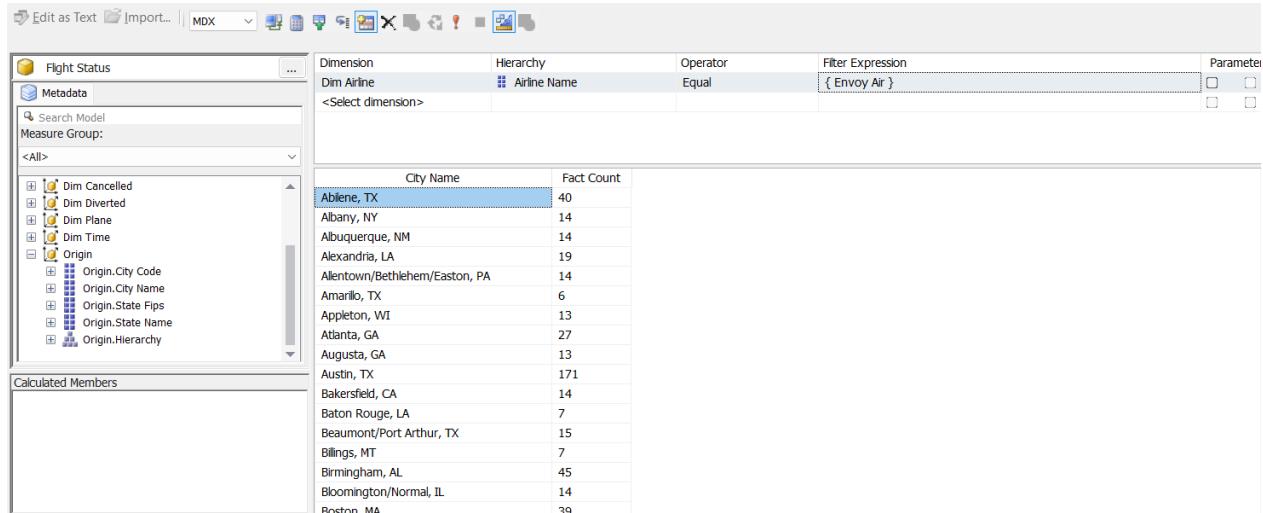
Hình 3.56. Kết quả thực hiện Power BI -2

3.7.5. Thống kê tổng số chuyến bay khởi hành của hãng hàng không Envoy Air theo từng thành phố.

SSAS

- Bước 1: Kéo thuộc tính City Name và Fact Count từ Origin và Fact sang khung truy vấn.

- Bước 2: Ở khung Filter, chọn Airline cho Dimension, Airline Name cho Hierarchy, Equal cho Operator và Envoy Air cho Filter Expression.
- Bước 3: Chọn Click to execute the query.



Hình 3.57. SSAS - Thực hiện truy vấn

MDX

- Query

```
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY
{[Origin].[City Name].[City Name] }
ON ROWS
FROM [Flight Status]
WHERE [Dim Airline].[Airline Name].&[Envoy Air]
```

- Kết quả

The screenshot shows the Microsoft Analysis Services Management Studio interface. On the left, the 'Cube' dropdown is set to 'Flight Status'. Below it, the 'Measure Group' dropdown is set to '<All>'. The 'Flight Status' node is expanded, showing various dimensions like 'Dest', 'Dim Airline', etc. In the center, an MDX query is displayed:

```
-- câu 5 Thông kê tổng số chuyến bay khởi hành của hãng hàng không Envoy Air theo từng thành phố.
SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS, NON EMPTY {[Origin].[City Name].[City Name]} 
ON ROWS
FROM [Flight Status]
WHERE [Dim Airline].[Airline Name].&[Envoy Air]
```

Below the query, the results are shown in a table:

	Fact Count
Abilene, TX	40
Albany, NY	14
Albuquerque, NM	14
Alexandria, LA	19
Allentown/Bethlehem/Easton, PA	14
Amarillo, TX	6
Appleton, WI	13
Atlanta, GA	27
Augusta, GA	13
Austin, TX	171
Bakersfield, CA	14
Baton Rouge, LA	7
Beaumont/Port Arthur, TX	15

Hình 3.58. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn City Name, ở Filter chọn Airline Name và Envoy Air cho giá trị.

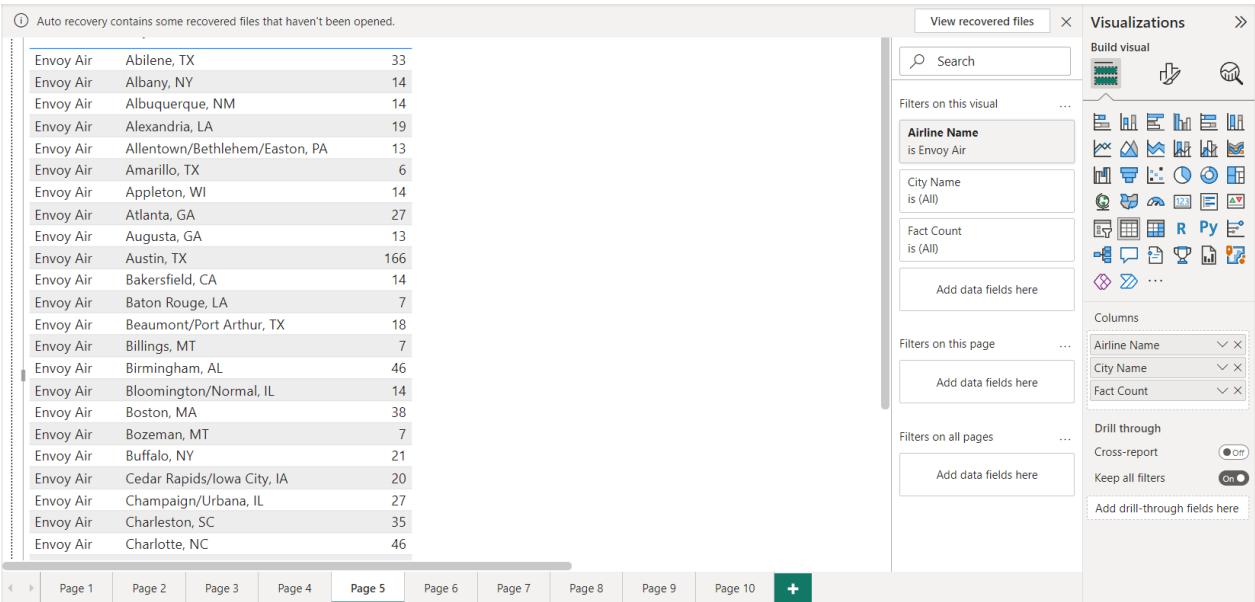
The screenshot shows a Microsoft Excel spreadsheet with a PivotTable. The table has 'Airline Name' in column A and 'Envoy Air' in row 1. The rows are labeled 'Row Labels' and 'Fact Count'. The data includes:

Airline Name	Envoy Air
Abilene, TX	40
Albany, NY	14
Albuquerque, NM	14
Alexandria, LA	19
Allentown/Bethlehem/Easton, PA	14
Amarillo, TX	6
Appleton, WI	13
Atlanta, GA	27
Augusta, GA	13
Austin, TX	171
Bakersfield, CA	14
Baton Rouge, LA	7
Beaumont/Port Arthur, TX	15

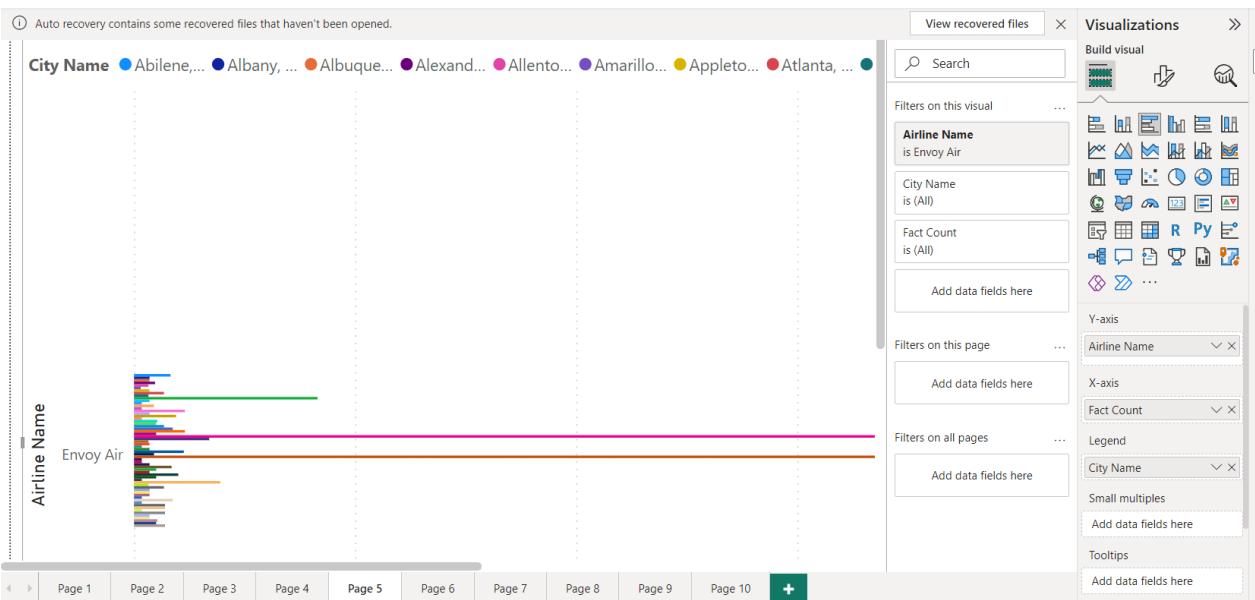
The 'PivotTable Fields' pane on the right shows 'Airline Name' selected under 'Values' and 'City Name' selected under 'Rows'.

Hình 3.59. Kết quả thực hiện Excel

Power BI



Hình 3.60. Kết quả thực hiện Power BI -1

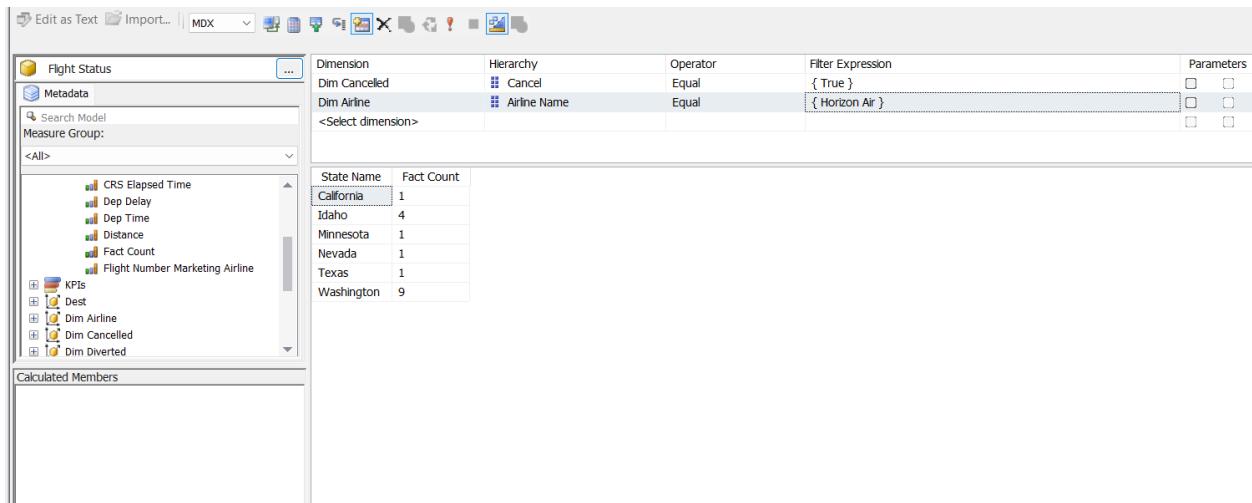


Hình 3.61. Kết quả thực hiện Power BI -2

3.7.6. Thống kê tổng số chuyến bay bị hủy theo từng bang của hãng hàng không Horizon Air .

SSAS

- Bước 1: Kéo thuộc tính State Name và Fact Count từ Origin và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dim Cancelled cho Dimension, Cancel cho Hierarchy, true cho Filter Expression.
- Bước 3: Chọn Click to execute the query.



Hình 3.62. SSAS - Thực hiện truy vấn

MDX

- Query:

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Origin].[State Name].[State Name].ALLMEMBERS ) } ON
ROWS
FROM [Flight Status]
WHERE ([Dim Cancelled].[Cancel].&[True], [Dim Airline].[Airline
Name].&[Horizon Air])

```

- Kết quả:

The screenshot shows the Microsoft Analysis Services MDX query editor interface. On the left, there's a navigation pane with a tree view of the cube structure under 'Flight Status'. The root node 'Flight Status' has children like 'Metadata', 'Functions', and 'Search Model'. Below that, 'Measure Group' is set to '<All>'. The main area contains an MDX query:

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Origin].[State Name].[State Name].ALLMEMBERS ) } ON ROWS
FROM [Flight Status]
WHERE ([Dim Cancelled].[Cancel].&[True], [Dim Airline].[Airline Name].&[Horizon Air])
  
```

Below the query, there's a preview window titled 'Results' showing a table with state names and their corresponding fact counts:

State	Fact Count
California	1
Idaho	4
Minnesota	1
Nevada	1
Texas	1
Washington	9

Hình 3.63. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn State Name, ở Filter chọn Cancel và True cho giá trị.

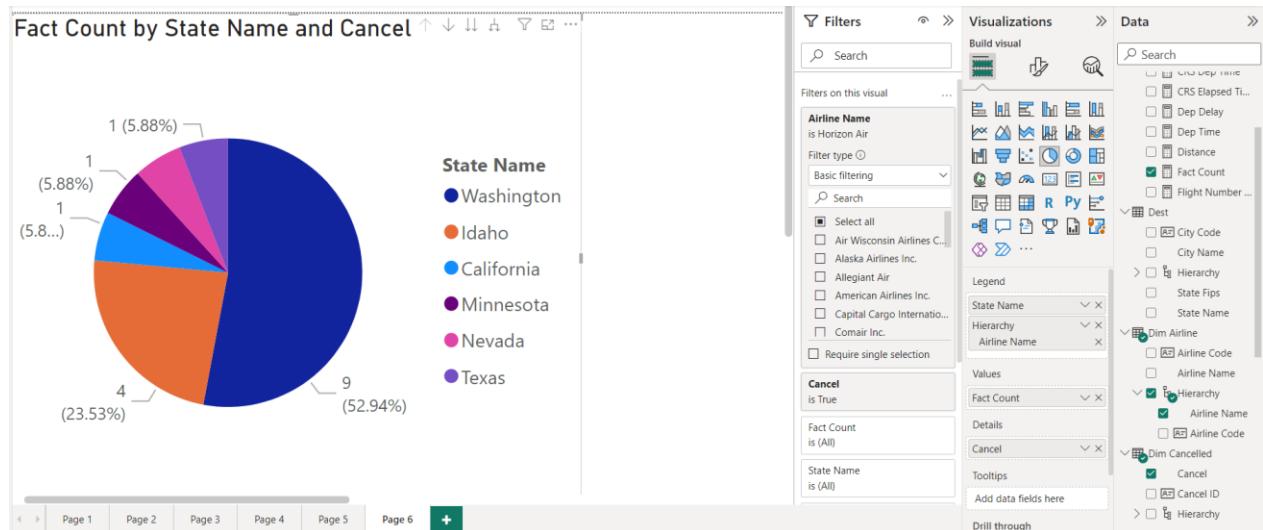
	A	B	C	D	E	F	G	H
1	Cancel	True						
2	Airline Name	Horizon Air						
3								
4	Row Labels	Fact Count						
5	California	1						
6	Idaho	4						
7	Minnesota	1						
8	Nevada	1						
9	Texas	1						
10	Washington	9						
11	Grand Total	17						
12								
13								
14								
15								
16								
17								
18								

Hình 3.64. Kết quả thực hiện Excel

PowerBI

State Name	Fact Count	Cancel	Airline Name
California	1	True	Horizon Air
Idaho	4	True	Horizon Air
Minnesota	1	True	Horizon Air
Nevada	1	True	Horizon Air
Texas	1	True	Horizon Air
Washington	9	True	Horizon Air
Total	17		

Hình 3.65. Kết quả thực hiện Power BI - I

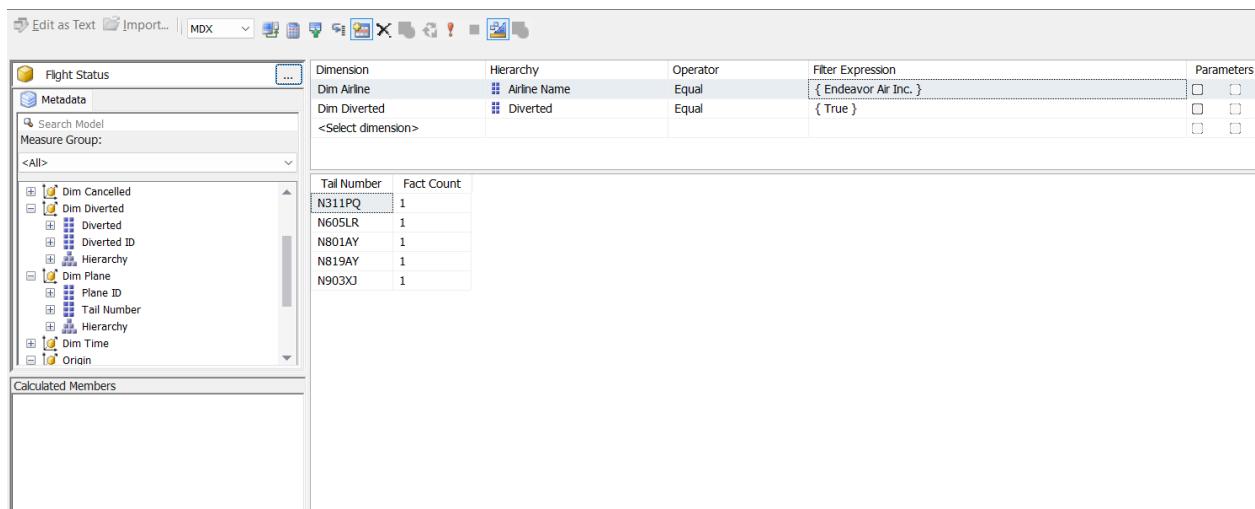


Hình 3.66. Kết quả thực hiện Power BI -2

3.7.7. Thống kê số lượng chuyến bay khởi hành của hãng hàng không Frontier Airlines Inc. bị điều hướng theo từng máy bay.

SSAS

- Bước 1: Kéo thuộc tính Tail Number và Fact Count từ Dim Plane và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dim Diverted và Dim Airline cho Dimension, Diverted và Airline cho Hierarchy.
- Bước 3: Chọn Click to execute the query.



Hình 3.67. SSAS - Thực hiện truy vấn

MDX

- Query:

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Plane].[Tail Number].[Tail Number].ALLMEMBERS ) }
} ON ROWS
FROM [Flight Status]
WHERE      (      [Dim       Diverted].[Diverted].&[True],[Dim
Airline].[Airline Name].&[Endeavor Air Inc.] )
    
```

- Kết quả

The screenshot shows the Microsoft Analysis Services Management Studio (SSMS) interface. On the left, there's a navigation pane with a tree view of the cube structure, including nodes like Flight Status, Measures, KPIs, Dest, Dim Airline, Dim Cancelled, Dim Diverted, Dim Plane, Dim Time, and Origin. The main area has two panes: the top pane displays the MDX query, and the bottom pane shows the results of the query execution.

```

SELECT NON EMPTY { [Measures].[Fact Count] } ON COLUMNS,
NON EMPTY { ([Dim Plane].[Tail Number].[Tail Number].ALLMEMBERS ) } ON ROWS
FROM [Flight Status]
WHERE      (      [Dim       Diverted].[Diverted].&[True],[Dim
Airline].[Airline Name].&[Endeavor Air Inc.] )
    
```

Results:

Fact Count	
N311PQ	1
N605LR	1
N801AY	1
N819AY	1
N903XJ	1

Hình 3.68. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Fact Count, ở mục Rows chọn Tail Number, ở Filter chọn Airline Name và Diverted.

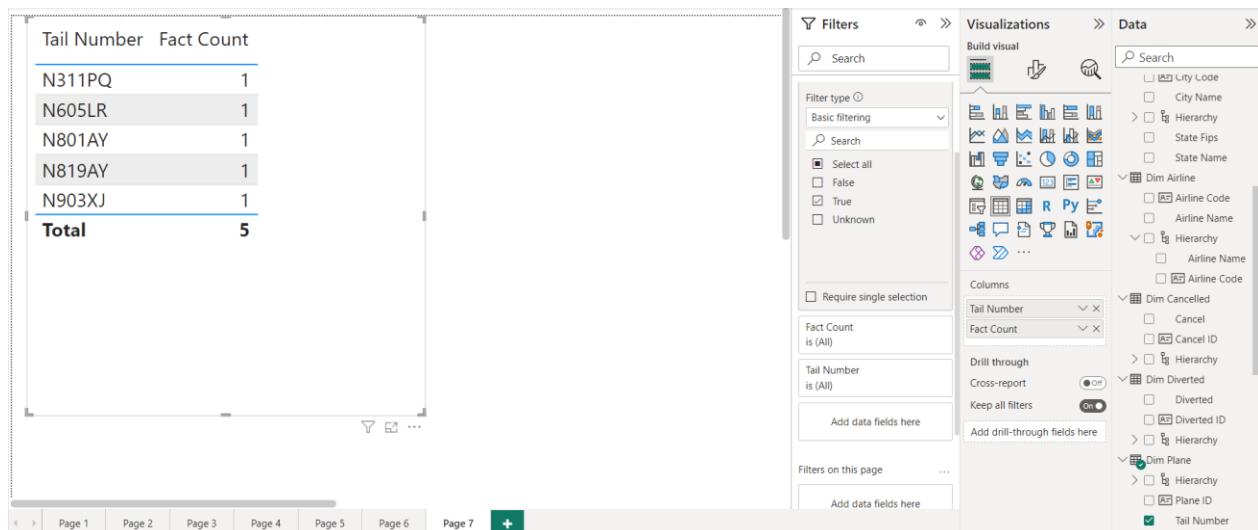
IS217 – Kho dữ liệu và OLAP

The screenshot shows a Microsoft Excel spreadsheet with a PivotTable. The PivotTable Fields pane on the right lists fields from the Dim Diverted and Fact Count tables. The main table displays the following data:

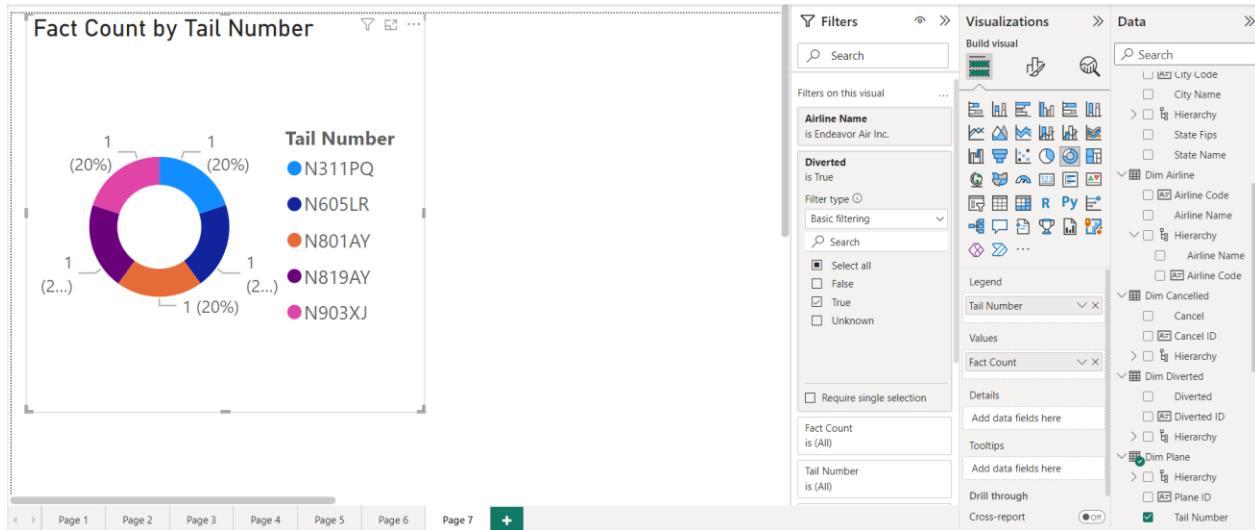
	Airline Name	Diverted	Fact Count
1	Endeavor Air Inc.	True	
2			
3			
4	Row Labels	Fact Count	
5	N311PQ		1
6	N605LR		1
7	N801AY		1
8	N819AY		1
9	N903XJ		1
10	Grand Total		5
11			
12			
13			
14			
15			
16			
17			

Hình 3.69. Kết quả thực hiện Excel

PowerBI



Hình 3.70. Kết quả thực hiện Power BI - I

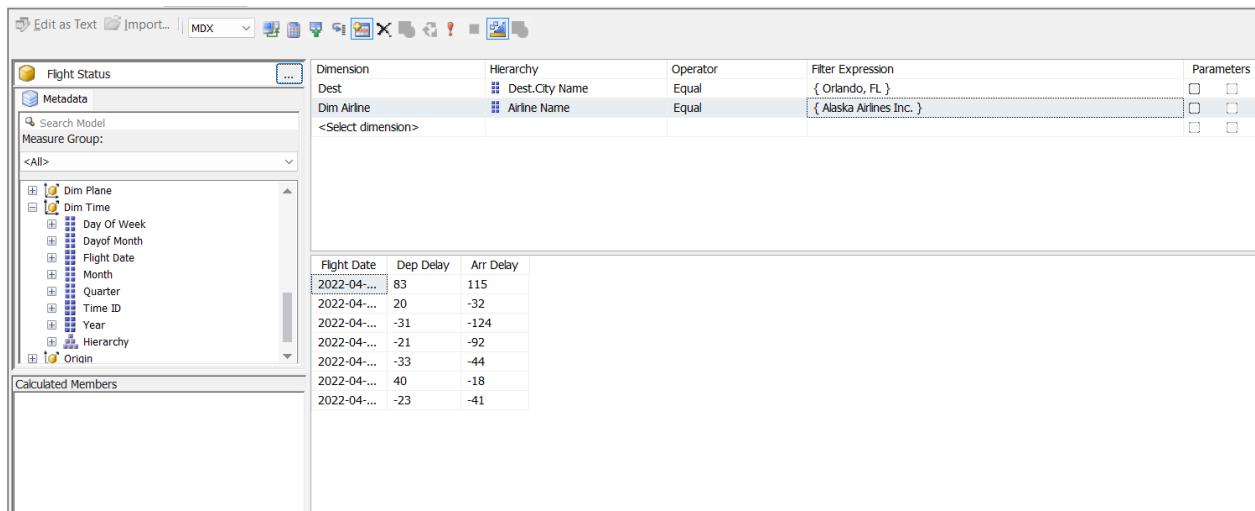


Hình 3.71. Kết quả thực hiện Power BI -2

3.7.8. Thống kê tổng thời gian cất cánh trễ và tổng thời gian khởi hành trễ của các chuyến bay đến thành phố Orlando tính theo các ngày của hãng hàng không Alaska Airline Inc.

SSAS

- Bước 1: Kéo thuộc tính Flight Date, Dep Delay, Arr Delay sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Equal cho Operator, chọn Orlando cho Dest.City Name và chọn Alaska Airlines Inc cho Airline Name
- Bước 3: Chọn Click to execute the query



Hình 3.72. SSAS - Thực hiện truy vấn

MDX

- Query:

```

SELECT
    {[Measures].[Dep Delay], [Measures].[Arr Delay]} ON COLUMNS,
    {[Dim Time].[Flight Date].Members} ON ROWS
FROM [Flight Status]
WHERE ([Dim Airline].[Airline Name].[Alaska Airlines Inc.],
       [Dest].[City Name].[Orlando, Fl])
  
```

- Kết quả

The screenshot shows the SSMS interface with an MDX query results window. The query is:

```

SELECT
    {[Measures].[Dep Delay], [Measures].[Arr Delay]} ON COLUMNS,
    {[Dim Time].[Flight Date].Members} ON ROWS
FROM [Flight Status]
WHERE ([Dim Airline].[Airline Name].[Alaska Airlines Inc.],
       [Dest].[City Name].[Orlando, Fl])
  
```

The results table displays flight delays for various dates in April 2022, categorized by flight date. The table has three columns: Dep Delay, Arr Delay, and a third column which appears to be a timestamp or date.

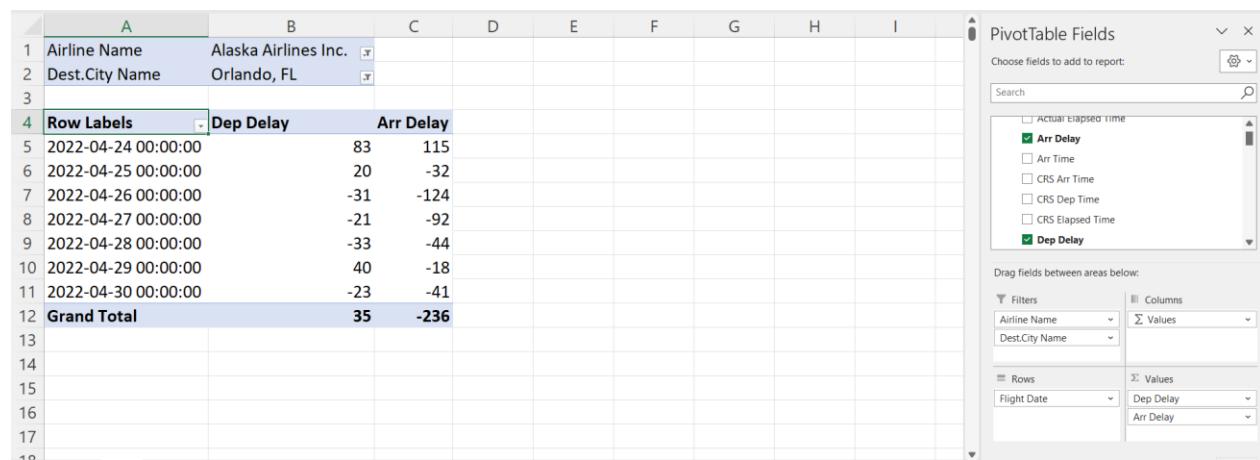
	Dep Delay	Arr Delay
All	35	-236
2022-04-24 00:00:00	83	115
2022-04-25 00:00:00	20	-32
2022-04-26 00:00:00	-31	-124
2022-04-27 00:00:00	-21	-92
2022-04-28 00:00:00	-33	-44
2022-04-29 00:00:00	40	-18
2022-04-30 00:00:00	-23	-41
Unknown	(null)	(null)

Hình 3.73. MDX - Kết quả

Excel (Pivot Table)

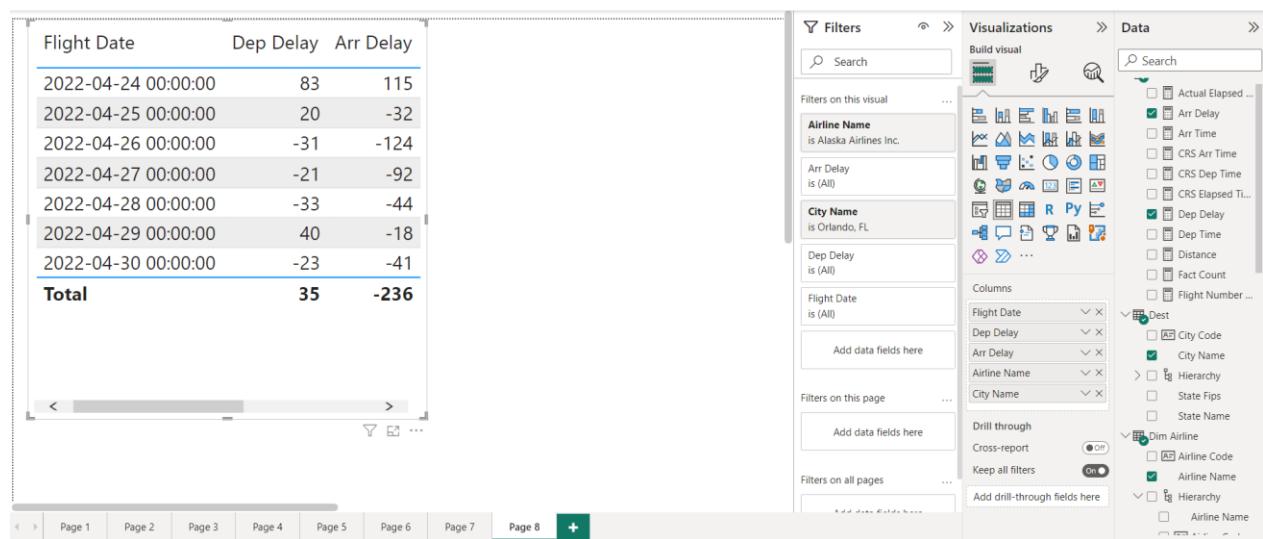
- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Dep Delay và Arr Delay, ở mục Rows chọn Flight Date, ở Filter chọn Airline Name và Dest.City Name.
- Bước 3: Chọn giá trị cho filter Airline Name, Dest.City Name

IS217 – Kho dữ liệu và OLAP

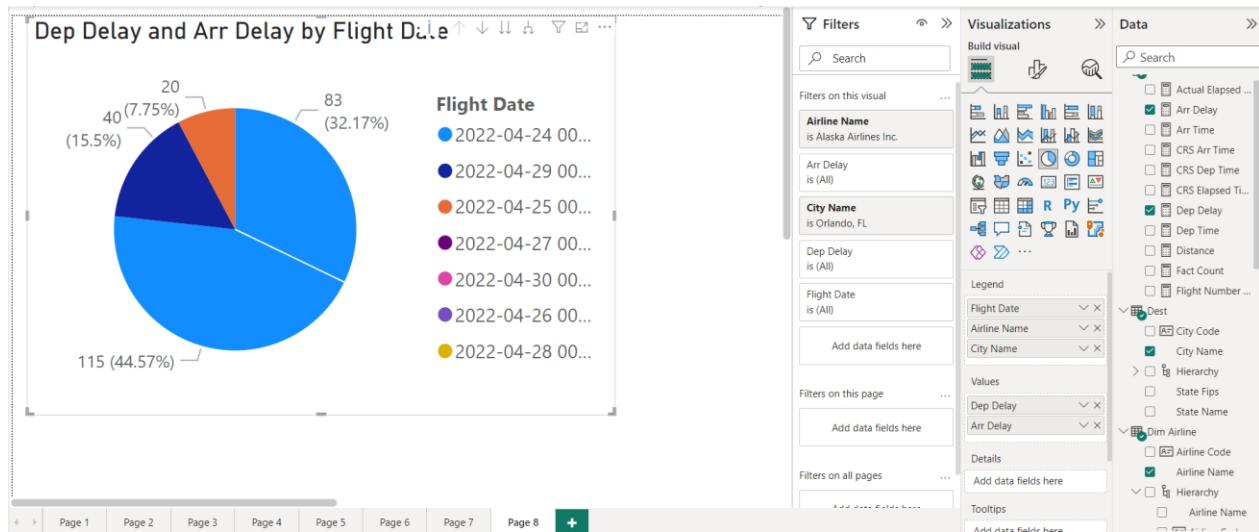


Hình 3.74. Kết quả thực hiện Excel

PowerBI



Hình 3.75. Kết quả thực hiện Power BI -1



Hình 3.76. Kết quả thực hiện Power BI -2

3.7.9. Thống kê khoảng cách và mã chuyến bay theo từng chiếc máy bay

SSAS

- Bước 1: Kéo thuộc tính Flight Number Marketing Airline, Tail Number, Distance sang khung truy vấn.
 - Bước 2: Ở khung Filter, chọn Equal cho Operator, chọn 2 cho Quarter.
 - Bước 3: Chọn Click to execute the query.

The screenshot shows the SSAS Management Studio environment. On the left, the 'Flight Status' cube is selected. A dimension filter for 'Dim Time' is applied with the operator 'Equal' and the value '{ 2 }', which corresponds to 'Quarter'. The 'Calculated Members' pane is expanded, displaying measures such as Arr Delay, Arr Time, CRS Arr Time, CRS Dep Time, CRS Elapsed Time, Dep Delay, Dep Time, Distance, Fact Count, Flight Number Marketing Airline, and KPIs.

Tail Number	Flight Number	Marketing Airline	Distance
202NV	19233		23878
203NV	48597		18222
204NV	10030		7780
205NV	61380		24700
206NV	47544		24177
207NV	1472		8018
215NV	17615		18884
216NV	6223		15032
217NV	58091		18600
218NV	1388		12618
219NV	11153		17246
220NV	15888		18714
221NV	26062		25224

Hình 3.77. SSAS - Thực hiện truy vấn

MDX

- Query:

```

SELECT
    {[Measures].[Flight Number] Marketing Airline},
    {[Measures].[Distance]} ON COLUMNS,
    {[Dim Plane].[Tail Number].Members} ON ROWS
FROM [Flight Status]
WHERE [Dim Time].[Quarter].[2]
  
```

- Kết quả:

The screenshot shows the SSAS Management Studio interface. On the left, the 'Cube' pane displays the 'Flight Status' cube with its dimensions and measures. The 'Measure Group' dropdown is set to '<All>'. The 'Fact' dimension is expanded, showing various time-related measures like Actual Elapsed Time, Arr Delay, Arr Time, CRS Arr Time, CRS Dep Time, CRS Elapsed Time, Dep Delay, Dep Time, Distance, Fact Count, and Flight Number Marketing Airline. Other dimensions like Dest, Dim Airline, Dim Cancelled, Dim Diverted, Dim Plane, Dim Time, and Origin are also listed. On the right, the 'Messages' tab shows the executed MDX query:

```

SELECT
    {[Measures].[Flight Number Marketing Airline], [Measures].[Distance]} ON COLUMNS,
    {[Dim Plane].[Tail Number].Members} ON ROWS
FROM [Flight Status]
WHERE [Dim Time].[Quarter].[2]
  
```

The 'Results' tab displays the query results in a table:

	Flight Number Marketing Airline	Distance
All	335961482	105735592
202NV	19233	23878
203NV	48597	18222
204NV	10030	7780
205NV	61380	24700
206NV	47544	24177
207NV	1472	8018
215NV	17615	18884
216NV	6223	15032
217NV	58091	18600
218NV	1388	12618
219NV	11153	17246
220NV	15888	18714
221NV	26062	25284
222NV	73764	23710
223NV	46029	18276

Hình 3.78. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Distance và Flight Number Marketing Airline, ở mục Rows chọn Tail Number, ở Filter chọn Quarter và 2 cho giá trị.

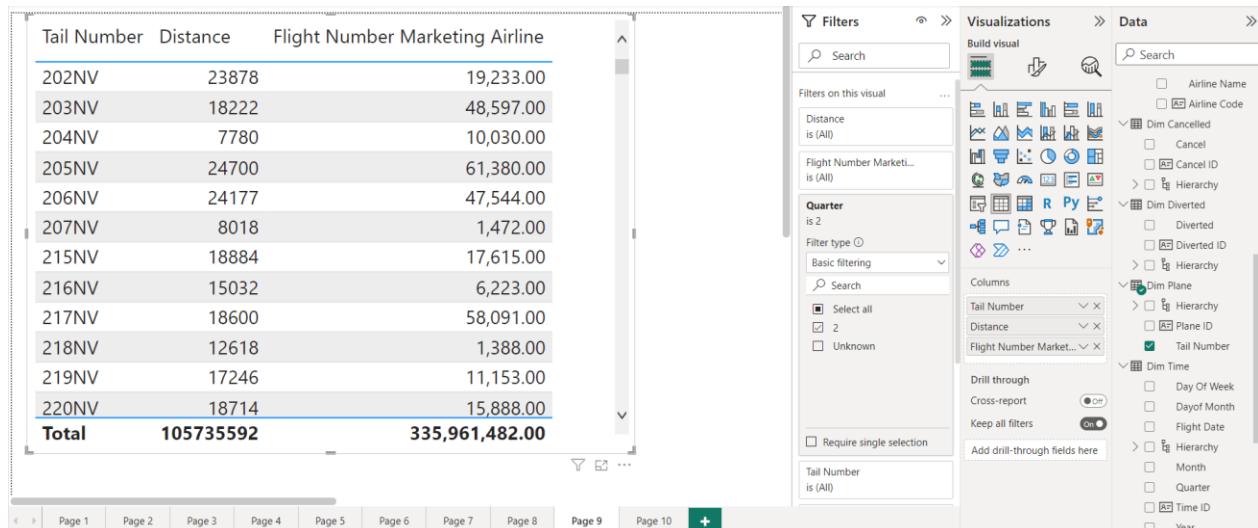
IS217 – Kho dữ liệu và OLAP

A screenshot of Microsoft Excel showing a PivotTable report. The PivotTable Fields pane on the right lists various dimensions and measures. The main table displays flight data with columns for Tail Number, Distance, and Marketing Airline.

Quarter	Distance	Flight Number Marketing Airline
202NV	23878	19233
203NV	18222	48597
204NV	7780	10030
205NV	24700	61380
206NV	24177	47544
207NV	8018	1472
215NV	18884	17615
216NV	15032	6223
217NV	18600	58091
218NV	12618	1388
219NV	17246	11153
220NV	18714	15888
221NV	25284	26062
Total	105735592	335,961,482.00

Hình 3.79. Kết quả thực hiện Excel

PowerBI



Hình 3.80. Kết quả thực hiện Power BI

3.7.10.Thống kê tổng khoảng cách và tổng thời gian bay theo từng máy bay từ ngày

24-04-2022 đến ngày 26-04-2022

SSAS

- Bước 1: Kéo thuộc tính Tail Number, Distance và Actual Elapsed Time từ Dim Plane và Fact sang khung truy vấn.
- Bước 2: Ở khung Filter, chọn Dim Time cho Dimension, Flight Date cho Hierarchy, Range cho Operator.

- Bước 3: Chọn Click to execute the query.

Tail Number	Distance	Actual Elapsed Time
202NV	13922	2158
203NV	8206	1248
205NV	8388	1338
206NV	9348	1430
207NV	8018	1312
215NV	8140	1356
216NV	7292	1182
217NV	7440	1125
218NV	10198	1566
219NV	7168	1188
220NV	9020	1456
221NV	6446	1059

Hình 3.81. SSAS - Thực hiện truy vấn

MDX

- Query:

```

SELECT {[Dim Plane].[Tail Number].[Tail Number].Members} ON ROWS,
{[Measures].[Distance], [Measures].[Actual Elapsed Time]} ON COLUMNS
FROM [Flight Status]
WHERE ([Dim Time].[Flight Date].&[2022-04-24T00:00:00]: [Dim Time].[Flight Date].&[2022-04-26T00:00:00])
    
```

- Kết quả:

The screenshot shows the MDX query interface with the following details:

- Cube:** Flight Status
- Measure Group:** <All>
- Fact:** Actual Elapsed Time, Arr Delay, Arr Time, CRS Arr Time, CRS Dep Time, CRS Elapsed Time, Dep Delay, Dep Time, Distance, Fact Count, Flight Number Marketing Airline
- Filter:** Flight Date between 2022-04-24T00:00:00 and 2022-04-26T00:00:00
- MDX Query:**

```

SELECT
    {[Dim Plane].[Tail Number].[Tail Number].Members} ON ROWS,
    {[Measures].[Distance], [Measures].[Actual Elapsed Time]} ON COLUMNS
FROM [Flight Status]
WHERE ([Dim Time].[Flight Date].&[2022-04-24T00:00:00]: [Dim Time].[Flight Date].&[2022-04-26T00:00:00])
  
```
- Results:** A table showing Distance and Actual Elapsed Time for various flight tail numbers.

Tail Number	Distance	Actual Elapsed Time
202NV	13922	2158
203NV	8206	1248
204NV	(null)	(null)
205NV	8388	1338
206NV	9348	1430
207NV	8018	1312
215NV	8140	1356
216NV	7292	1182
217NV	7440	1125
218NV	10198	1566
219NV	7168	1188
220NV	9020	1456
221NV	6446	1059
222NV	9182	1466
223NV	3970	453
224NV	7588	1188

Hình 3.82. MDX - Kết quả

Excel (Pivot Table)

- Bước 1: Chọn tab Phân tích Pivot
- Bước 2: Ở mục Value chọn Distance và Actual Elapsed Time, ở mục Rows chọn Tail Number, ở Filter chọn Flight Date và giá trị từ 24-04-2022 đến 26-04-2022

The screenshot shows an Excel PivotTable with the following configuration:

- Filters:** Flight Date (Multiple Items) is selected.
- Row Labels:** Tail Number
- Values:** Distance and Actual Elapsed Time (Values)
- PivotTable Fields pane:**
 - Choose fields to add to report: Flight Date
 - Drag fields between areas below:
 - Filters: Flight Date
 - Columns: Values
 - Rows: Tail Number
 - Values: Distance, Actual Elapsed Time

Tail Number	Distance	Actual Elapsed Time
202NV	13922	2158
203NV	8206	1248
205NV	8388	1338
206NV	9348	1430
207NV	8018	1312
215NV	8140	1356
216NV	7292	1182
217NV	7440	1125
218NV	10198	1566
219NV	7168	1188
220NV	9020	1456
221NV	6446	1059
222NV	9182	1466
223NV	3970	453
224NV	7588	1188

Hình 3.83. Kết quả thực hiện Excel

PowerBI

The screenshot shows a PowerBI interface with a table visualization on the left and the PowerBI ribbon on the right.

Table Visualization:

Tail Number	Distance	Actual Elapsed Time
202NV	13922	2158
203NV	8206	1248
205NV	8388	1338
206NV	9348	1430
207NV	8018	1312
215NV	8140	1356
216NV	7292	1182
217NV	7440	1125
218NV	10198	1566
219NV	7168	1188
220NV	9020	1456
221NV	6446	1059
Total	45641228	7584819

PowerBI Ribbon:

- Filters**: Shows filters applied to the visual, including "Actual Elapsed Time is (All)" and "Distance is (All)".
- Visualizations**: Shows a list of available visualizations like Matrix, Gauge, and Map.
- Data**: Shows the data source structure with categories like Dim Cancelled, Dim Diverted, Dim Plane, Dim Time, and Origin.

Hình 3.84. Kết quả thực hiện Power BI

CHƯƠNG 4 – DATA MINING

Đến với phần Data mining sẽ tập trung vào việc sử dụng Decision Tree và Random Forest để phân tích và dự đoán tình trạng chuyến bay. Decision Tree giúp chúng ta hiểu các quy luật và mối quan hệ trong dữ liệu, trong khi Random Forest kết hợp các dự đoán từ nhiều cây quyết định để đưa ra kết quả chính xác. Điều này giúp chúng ta tìm ra mô hình phù hợp và đưa ra dự đoán tin cậy về tình trạng chuyến bay.

4.1. Sơ đồ mining

Các thuộc tính sẽ sử dụng trong quá trình Mining: OriginStateFips, DestStateFips, CRSDepTime, DepDelay, Cancelled.

Thuộc tính quyết định để thực hiện Mining: DelayGroup. Thuộc tính này được tạo ra bằng cách so sánh các giá trị của DepDelay, Cancelled. Trong DelayGroup sẽ có 3 giá trị cho biết tình trạng của chuyến bay đó là: chuyến bay có bị hủy hay không, chuyến có đến đúng giờ hay không và chuyến bay có bị trễ hay không.

Các thuộc tính input đầu vào để thực hiện Mining

- CRSDepTime: Thời gian khởi hành theo dự kiến
- OriginStateFips: Số hiệu bang khởi hành.
- DestStateFips: Số hiệu bang đích đến.

Lựa chọn thuật toán:

- Decision Tree
- Random Forest

4.2. Quá trình thực hiện

4.2.1. Xử lý dữ liệu

- Lựa chọn những thuộc tính sử dụng để thực hiện quá trình Mining

```
#Lựa chọn những thuộc tính sử dụng
df = df[['OriginStateFips', 'DestStateFips', 'CRSDepTime', 'DepDelay', 'Cancelled']]
✓ 0.0s
```

Hình 4.1. Lựa chọn những thuộc tính sử dụng

- Phân loại giá trị CRSDepTime thành các khoảng. Ở đây, ta sẽ dựa vào thời gian khởi hành của máy bay mà chia thời gian này thành các khoảng: sáng sớm, sáng, chiều và tối và được ký hiệu lần lượt theo thứ tự là 1, 2, 3 và 4.

```
#chuyển đổi giá trị trong cột 'CRSDepTime' sang kiểu dữ liệu chuỗi (string) thêm các ký tự '0' vào đầu chuỗi (nếu cần) để đảm bảo rằng chiều dài của
df['CRSDepTime'] = df['CRSDepTime'].astype(str).str.zfill(4)
```

✓ 0.2s

Python

```
# Phân loại giá trị CRSDepTime thành các khoảng
df.loc[df['CRSDepTime'].astype(str)[:2].astype(int) <= 6, 'CRSDepTime'] = 1 # Early Morning
df.loc[(df['CRSDepTime'].astype(str)[:2].astype(int) > 6) & (df['CRSDepTime'].astype(str)[:2].astype(int) <= 12), 'CRSDepTime'] = 2 # Morning
df.loc[(df['CRSDepTime'].astype(str)[:2].astype(int) > 12) & (df['CRSDepTime'].astype(str)[:2].astype(int) <= 18), 'CRSDepTime'] = 3 # Afternoon
df.loc[df['CRSDepTime'].astype(str)[:2].astype(int) > 18, 'CRSDepTime'] = 4 # Evening

# Chuyển đổi kiểu dữ liệu của cột 'CRSDepTime' thành 'int64'
df['CRSDepTime'] = df['CRSDepTime'].astype('int64')

unique_CRSDepTime = df['CRSDepTime'].unique()

print("Các giá trị khác nhau của cột 'CRSDepTime':", unique_CRSDepTime)
```

✓ 0.6s

Python

Các giá trị khác nhau của cột 'CRSDepTime': [3 2 1 4]

Hình 4.2. Phân loại giá trị CRSDepTime thành các khoảng

- Sau đó, ta sẽ tạo mới cột DelayGroup để phân loại tình trạng chuyến bay như đã đề cập trước đó.

```
#Tạo cột DelayGroup để phân loại tình trạng chuyến bay
df['DelayGroup'] = None
df.loc[df['DepDelay'] <= 0, 'DelayGroup'] = 'OnTime_Early'
df.loc[(df['DepDelay'] > 0), 'DelayGroup'] = 'Delay'
df.loc[df['Cancelled'], 'DelayGroup'] = 'Cancelled'
df
```

✓ 0.0s

Hình 4.3. Tạo cột DelayGroup để phân loại tình trạng chuyến bay

- Vì số lượng các bang đến và đi ở đây khá nhiều, nên ta chỉ chọn lại 3 bang có số lượng khởi hành và có đích đến nhiều nhất để thực hiện quá trình Mining.

```
# Lấy top 3 giá trị của cột 'OriginStateFips'  
top_origin = df['OriginStateFips'].value_counts().nlargest(3).index.tolist()  
  
# Lấy top 3 giá trị của cột 'DestStateFips'  
top_dest = df['DestStateFips'].value_counts().nlargest(3).index.tolist()  
  
# Chỉ giữ lại các hàng có giá trị trong top 3 của cả hai cột  
df = df[df['OriginStateFips'].isin(top_origin) & df['DestStateFips'].isin(top_dest)]  
:] ✓ 0.0s
```

Hình 4.4. Lọc giá trị OriginStateFips và DestStateFips

4.2.2. Decision Tree

- Như đã đề cập trước đó, 2 giá trị DepDelay và Cancelled được dùng để xác định cho giá trị DelayGroup nên ta loại bỏ 2 giá trị này.
- Xác định các thuộc tính features và labels.

```
#Tách các cột dữ liệu vào 2 biến features (chứa các thuộc tính bình thường) và biến labels (chứa riêng thuộc tính quyết định)  
features = dt.drop('DelayGroup', axis=1)  
labels = dt['DelayGroup']  
:] ✓ 0.0s
```

Hình 4.5. Xác định thuộc tính features và labels.

- Chia tập dữ liệu train và test (8-2), xây dựng cây ID3, tiến hành huấn luyện mô hình và cuối cùng là đánh giá mô hình.

```
X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.2, random_state=42)
] ✓ 0.0s
```

```
from sklearn import tree
# Xây dựng cây ID3
clf = tree.DecisionTreeClassifier(criterion="entropy", random_state=0)

# Huấn luyện mô hình
clf.fit(X_train, y_train)

] ✓ 0.0s
```

DecisionTreeClassifier
 DecisionTreeClassifier(criterion='entropy', random_state=0)

```
#Predict the response for test dataset
tree_pred = clf.predict(X_test)
#Model Accuracy, how often is the classifier correct?
from sklearn import metrics
tree_score = metrics.accuracy_score(y_test, tree_pred)
print("Accuracy: ", tree_score)
print("Report: ", metrics.classification_report(y_test, tree_pred))
] ✓ 0.1s
```

	precision	recall	f1-score	support
Cancelled	0.00	0.00	0.00	40
Delay	0.57	0.26	0.36	1081
OnTime_Early	0.65	0.88	0.75	1731
accuracy	0.63	0.63	0.63	2852

Hình 4.6. Huấn luyện mô hình Decision Tree

4.2.3. Random Forest

- Tương tự như thuật toán Decision Tree, ta sẽ loại bỏ 2 thuộc tính DepDelay và Cancelled và xác định các thuộc tính X và y. Sau đó khi tập dữ liệu train và test.

```
# Chia thành input features (X) và target variable (y)
X = rf.drop('DelayGroup', axis=1)
y = rf['DelayGroup']
```

```
# Chia tập dữ liệu thành training set và test set
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Hình 4.7. Xác định X và y

- Xây dựng mô hình Random Forest và tiến hành huấn luyện mô hình

```
# Xây dựng mô hình Random Forest
model = RandomForestClassifier(n_estimators=100, max_depth=5, random_state=42)

# Huấn luyện mô hình trên tập huấn luyện
model.fit(X_train, y_train)

# Dự đoán tình trạng chuyến bay trên tập kiểm tra
y_pred = model.predict(X_test)

from sklearn.metrics import accuracy_score

# Đánh giá độ chính xác của mô hình
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

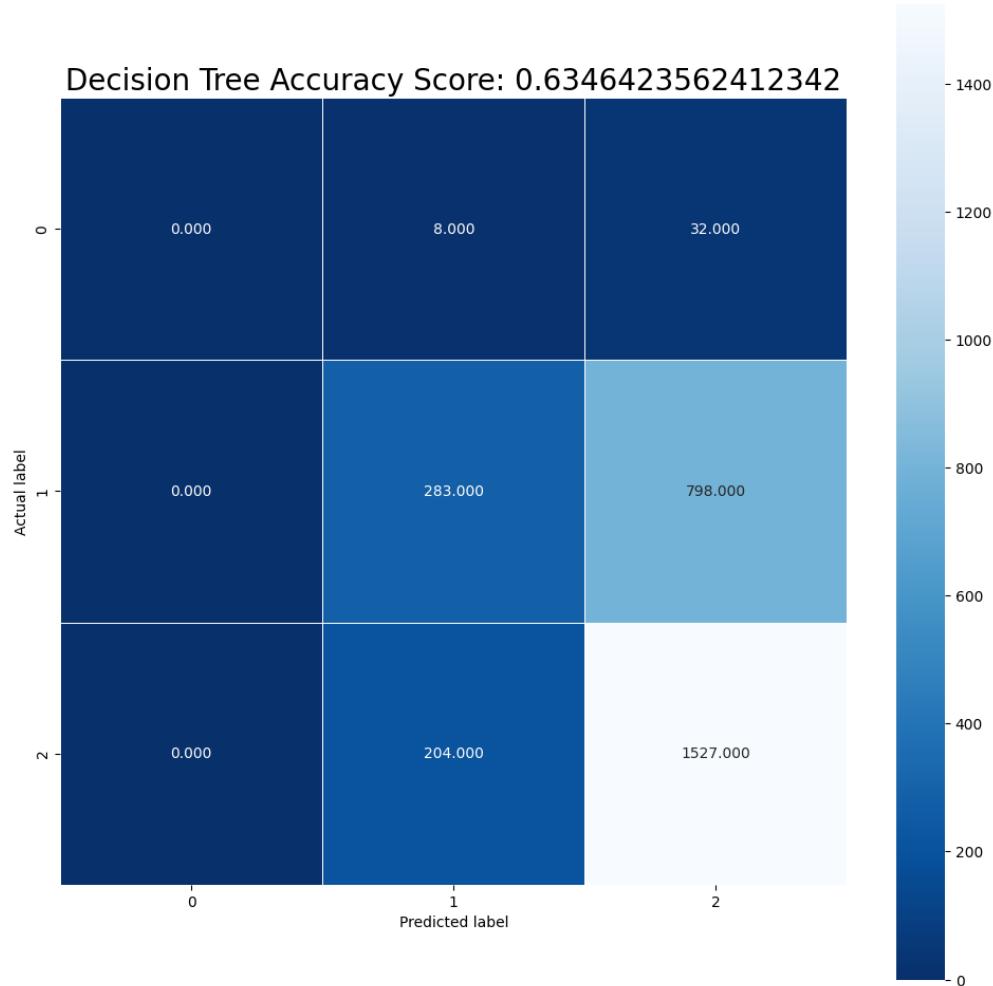
Accuracy: 0.6388499298737728

Hình 4.8. Xây dựng mô hình Random Forest

4.3. Giải thích kết quả

4.3.1. Thuật toán Decision Tree

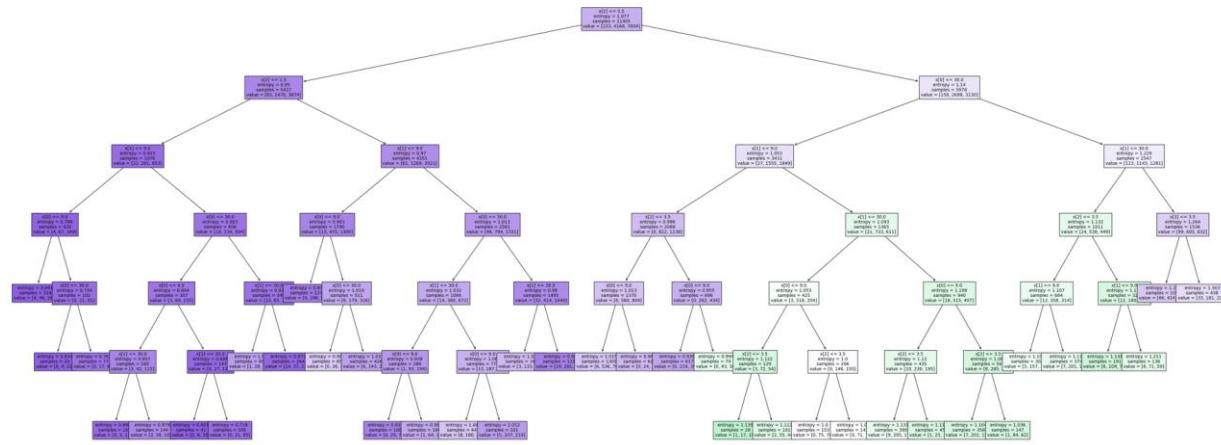
- Ma trận nhầm lẫn



Hình 4.9. Ma trận nhầm lẫn Decision Tree

- + Độ chính xác của mô hình cây quyết định là 63.46%.
- + Các màu trong ma trận nhầm lẫn đại diện cho mức độ tương quan giữa các giá trị trong ô. Các ô có màu tối (xanh đậm) đến màu nhạt (xanh nhạt) thể hiện giá trị từ thấp đến cao.
- + Hàng (cột) đầu tiên của ma trận đại diện cho lớp "OnTime_Early". Hàng (cột) thứ hai của ma trận đại diện cho lớp "Delay". Hàng (cột) thứ ba của ma trận đại diện cho lớp "Cancelled".
- + Có 0 dự đoán chính xác là 'OnTime_Early', 8 dự đoán sai lầm được gán nhãn là 'Delay', và 32 dự đoán sai lầm được gán nhãn là 'Cancelled'.

- + Có 0 dự đoán chính xác là "Delay", 283 dự đoán sai lầm được gán nhãn là "OnTime_Early", và 798 dự đoán sai lầm được gán nhãn là "Cancelled".
- + Có 0 dự đoán chính xác là "Cancelled", 204 dự đoán sai lầm được gán nhãn là "OnTime_Early", và 1527 dự đoán sai lầm được gán nhãn là "Delay".
- Cây ID3



Hình 4.10. Cây quyết định Decision Tree

- Rút ra tập luật từ cây quyết định

```

|--- DestStateFips <= 2.50
|   |--- DestStateFips <= 1.50
|   |   |--- OriginStateFips <= 9.00
|   |   |   |--- CRSDepTime <= 9.00
|   |   |   |   |--- class: OnTime_Early
|   |   |   |--- CRSDepTime >  9.00
|   |   |   |   |--- CRSDepTime <= 30.00
|   |   |   |   |   |--- class: OnTime_Early
|   |   |   |   |--- CRSDepTime >  30.00
|   |   |   |   |   |--- class: OnTime_Early
|   |   |   |--- OriginStateFips >  9.00
|   |   |   |--- CRSDepTime <= 30.00
|   |   |   |   |--- CRSDepTime <= 9.00
|   |   |   |   |   |--- OriginStateFips <= 30.00
|   |   |   |   |   |   |--- class: OnTime_Early
|   |   |   |   |--- OriginStateFips >  30.00
|   |   |   |   |   |--- class: OnTime_Early
|   |   |   |   |--- CRSDepTime >  9.00
|   |   |   |   |   |--- OriginStateFips <= 30.00
|   |   |   |   |   |   |--- class: OnTime_Early
|   |   |   |   |--- OriginStateFips >  30.00
|   |   |   |   |   |--- class: OnTime_Early
|   |   |   |--- CRSDepTime >  30.00
|   |   |   |   |--- OriginStateFips <= 30.00
|   |   |   |   |   |--- class: OnTime_Early
...
|   |   |   |   |--- class: OnTime_Early
|   |   |   |--- DestStateFips >  3.50
|   |   |   |   |--- class: OnTime_Early

```

Hình 4.11. Tập luật - Decision Tree

- + Từ tập luật trên, ta có thể thấy rằng:
 - Nếu $\text{DestStateFips} \leq 1.50$ và $\text{OriginStateFips} \leq 9.00$ và $\text{CRSDepTime} \leq 9.00$ thì tình trạng chuyến bay là `OnTime_Early`.

- Tương tự với các luật còn lại.
- Ở bài toán này, nhóm cũng tiến hành dự đoán tình trạng của chuyến bay dựa vào các giá trị features cho trước. Sau khi xử lý dữ liệu và tiến hành dự đoán dựa trên mô hình đã được huấn luyện trước đó, ta được kết quả tình trạng của chuyến bay.

```
#Dự đoán tình trạng chuyến bay (labels) dựa trên các giá trị thuộc tính (features)
#OriginStateFips    DestStateFips    CRSDepTime      -> DelayGroup
new_data = {
    'CRSDepTime': '2120',                      #Thời gian khởi hành theo máy tính: 21:20
    'OriginStateFips': 12,                       #Fort Myers, Florida
    'DestStateFips': 48                          #Newark, New Jersey
}
```

Hình 4.12. Dữ liệu đầu vào

```
# Lấy danh sách tên cột của quá trình huấn luyện
train_columns = X_train.columns.tolist()

# Sắp xếp lại thứ tự cột trong new_data_encoded
new_data_dt = new_data_dt.reindex(columns=train_columns)

# Dự đoán labels cho dữ liệu mới đã được mã hóa
predicted_labels = clf.predict(new_data_dt)

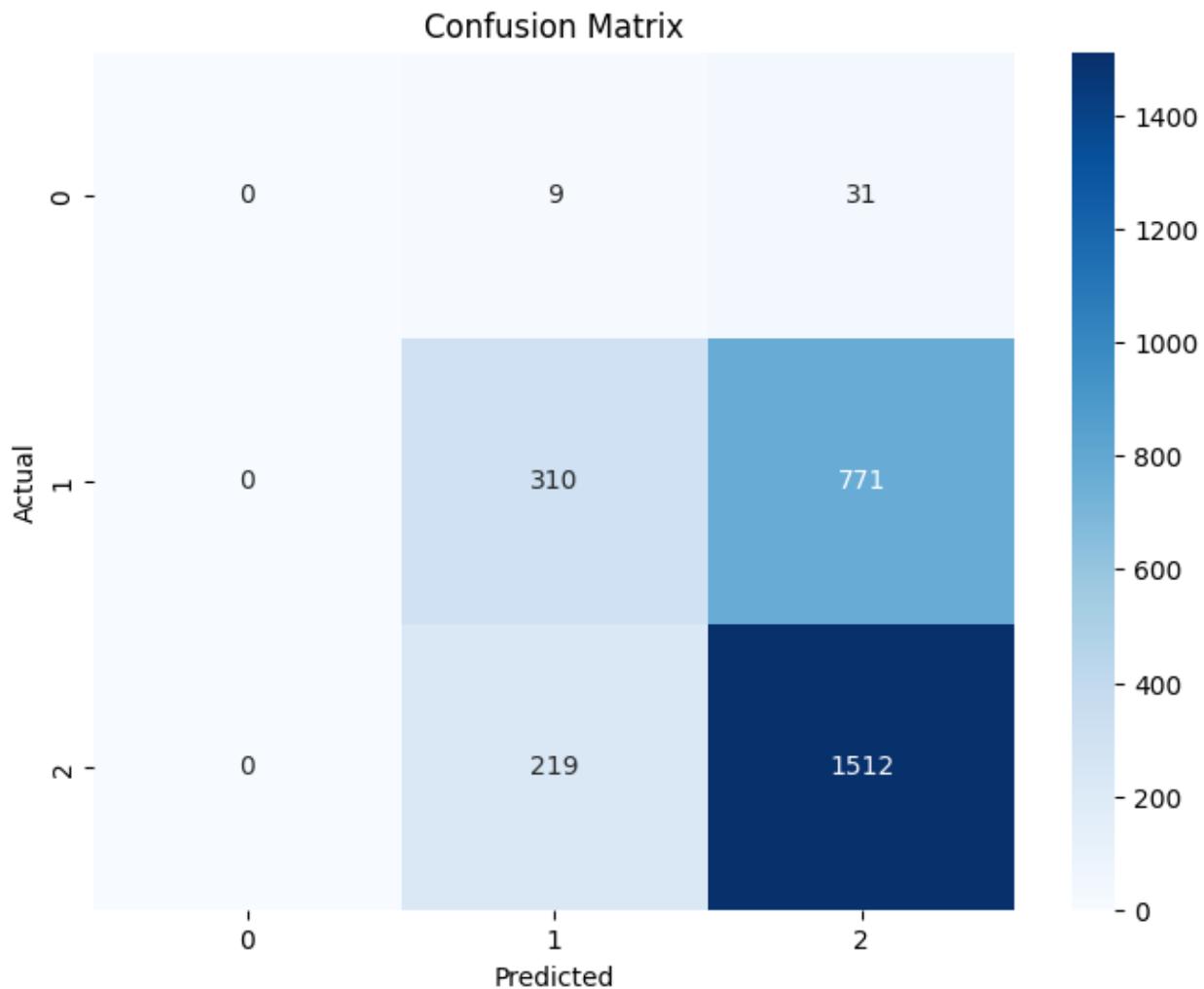
# In ra nhãn dự đoán
print(predicted_labels)

['Delay']
```

Hình 4.13. Kết quả dự đoán - Decision Tree

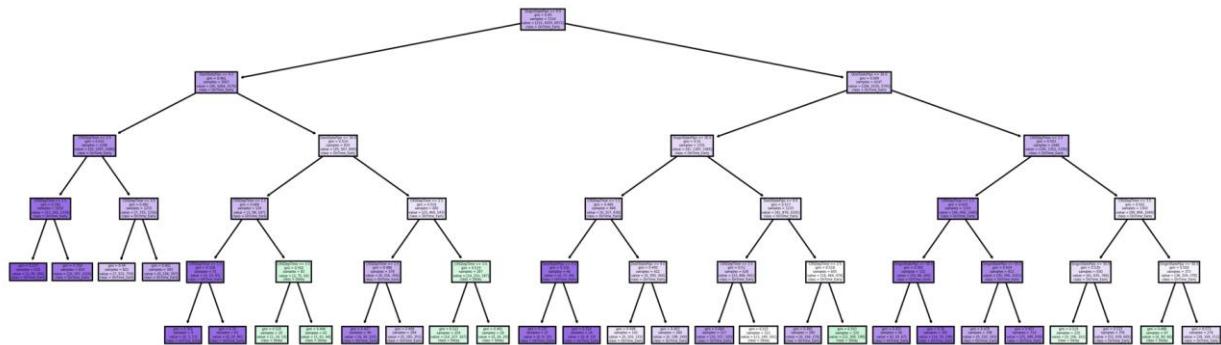
4.3.2. Thuật toán Random Forest

- Ma trận nhầm lẫn



- + Các màu trong ma trận nhầm lẫn đại diện cho mức độ tương quan giữa các giá trị trong ô. Các ô có màu tối (xanh đậm) đến màu nhạt (xanh nhạt) thể hiện giá trị từ thấp đến cao.
- + Hàng (cột) đầu tiên của ma trận đại diện cho lớp "OnTime_Early". Hàng (cột) thứ hai của ma trận đại diện cho lớp "Delay". Hàng (cột) thứ ba của ma trận đại diện cho lớp "Cancelled".
- + Có 0 dự đoán chính xác là 'OnTime_Early', 9 dự đoán sai lầm được gán nhãn là 'Delay', và 31 dự đoán sai lầm được gán nhãn là 'Cancelled'.
- + Có 0 dự đoán chính xác là "Delay", 310 dự đoán sai lầm được gán nhãn là "OnTime_Early", và 771 dự đoán sai lầm được gán nhãn là "Cancelled".

- + Có 0 dự đoán chính xác là "Cancelled", 219 dự đoán sai làm được gán nhãn là "OnTime_Early", và 1512 dự đoán sai làm được gán nhãn là "Delay".
- Biểu đồ quyết định



Hình 4.14. Cây quyết định Random Forest

- Rút ra tập luật

```
Tập luật từ cây quyết định:  
|--- OriginStateFips <= 9.00  
|   |--- DestStateFips <= 9.00  
|   |   |--- CRSDepTime <= 2.50  
|   |   |   |--- CRSDepTime <= 1.50  
|   |   |   |   |--- class: 2.0  
|   |   |   |   |--- CRSDepTime > 1.50  
|   |   |   |   |   |--- class: 2.0  
|   |   |--- CRSDepTime > 2.50  
|   |   |   |--- CRSDepTime <= 3.50  
|   |   |   |   |--- class: 2.0  
|   |   |   |   |--- CRSDepTime > 3.50  
|   |   |   |   |   |--- class: 2.0  
|--- DestStateFips > 9.00  
|   |--- DestStateFips <= 30.00  
|   |   |--- CRSDepTime <= 2.50  
|   |   |   |--- CRSDepTime <= 1.50  
|   |   |   |   |--- class: 2.0  
|   |   |   |   |--- CRSDepTime > 1.50  
|   |   |   |   |   |--- class: 2.0  
|   |   |--- CRSDepTime > 2.50  
|   |   |   |--- CRSDepTime <= 3.50  
|   |   |   |   |--- class: 1.0  
|   |   |   |   |--- CRSDepTime > 3.50  
|   |   |   |   |   |--- class: 1.0  
...  
|   |   |   |   |   |--- class: 1.0  
|   |   |   |--- OriginStateFips > 30.00  
|   |   |   |   |--- class: 2.0
```

Hình 4.15. Tập luật từ cây quyết định - Random Forest

- + Nếu $\text{OriginStateFips} \leq 9.00$ và $\text{DestStateFips} \leq 9.00$ và $\text{CRSDepTime} \leq 1.50$ thì class: 2.0 (Cancelled)

- + Nếu $\text{OriginStateFips} \leq 9.00$ và $(\text{DestStateFips} > 9.00 \text{ và } \text{DestStateFips} \leq 30.00)$ và $\text{CRSDepTime} \leq 3.50$ thì class: 1.0 (Delay)
- + Tương tự với các luật còn lại...
- Tương tự Decision Tree, ta cũng tiến hành dự đoán tình trạng chuyến bay đối với thuật toán Random Forest.

```
#Dự đoán tình trạng chuyến bay (labels) dựa trên các giá trị thuộc tính (features)
#OriginStateFips    DestStateFips    CRSDepTime      -> DelayGroup
new_data = {
    'CRSDepTime': '2120',                      #Thời gian khởi hành theo máy tính: 21:20
    'OriginStateFips': 12,                      #Fort Myers, Florida
    'DestStateFips': 48                         #Newark, New Jersey
}
✓ 0.0s
```

Hình 4.16. Dữ liệu đầu vào

```
# Lấy danh sách tên cột của quá trình huấn luyện
train_columns = X_train.columns.tolist()

# Sắp xếp lại thứ tự cột trong new_data_rf
new_data_rf = new_data_rf.reindex(columns=train_columns)

# Dự đoán tình trạng chuyến bay
new_data_pred = model.predict(new_data_rf)

# In kết quả dự đoán
print(new_data_pred)
[44] ✓ 0.0s
... ['Delay']
```

Hình 4.17. Kết quả dự đoán - Random Forest

4.4. Nhận xét

- Từ kết quả của 2 mô hình Decision Tree và Random Forest, dựa vào giá trị Accuracy, ta thấy mô hình Random Forest hiệu quả hơn mô hình Decision Tree (Random Forest có Accuracy là 63.89%, Decision Tree có Accuracy là 63.46%)

DANH MỤC TÀI LIỆU THAM KHẢO

- [1]. Dataset Flight Status Prediction,
https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?sizeStart=70%2CMB&page=13&select=Combined_Flights_2022.csv
- [2]. [Nhập môn Data Warehouse] Mô hình dữ liệu đa chiều, <https://faditek.com/nhap-mon-data-warehouse-mo-hinh-du-lieu/>
- [3]. Seminar hướng dẫn sử dụng công cụ Lookup,
https://www.youtube.com/watch?v=KTTmhRSwxY&ab_channel=Nguy%E1%BB%85nT%E1%BA%A5nTh%C3%A0nh
- [4]. Trình thiết kế Truy vấn MDX Query Designer (Power Pivot),
<https://support.microsoft.com/vi-vn/office/tri%CC%80nh-thi%C3%AA%CC%81t-k%C3%AA%CC%81-truy-v%C3%A2n%CC%81n-mdx-query-designer-power-pivot-30ab91f9-82a6-4f95-a5ec-2b6b7ab5cbcf>
- [5]. Tổng quan về ngôn ngữ truy vấn Kho dữ liệu MDX,
<http://bis.net.vn/forums/p/560/1083.aspx>
- [6]. Cách kết nối với các mô hình đa chiều SSAS trong POWER BI DESKTOP,
<https://www.bacs.vn/vi/blog/cong-cu-ho-tro/cach-ket-noi-voicac-mo-hinh-da-chieu-ssas-trong-power-bi-desktop-15617.html>
- [7]. Decision Tree algorithm,
https://machinelearningcoban.com/tablml_book/ch_model/decision_tree.html
- [8]. Cây Quyết Định (Decision Tree), <https://trituenhantao.io/kien-thuc/decision-tree/>
- [9]. Tự học ML | Triển khai cây quyết định bằng Python, <https://cafedev.vn/tu-hoc-ml-trien-khai-cay-quyet-dinh-bang-python/>
- [10]. Random Forest algorithm,
https://machinelearningcoban.com/tablml_book/ch_model/random_forest.html
- [11]. Phân lớp bằng Random Forests trong Python, <https://viblo.asia/p/phan-lop-bang-random-forests-trong-python-djeZ1D2QKWz>
- [12]. Diễn giải blackbox model | Phần 1: Random Forest,
<https://rpubs.com/lengockhanhi/343200>