**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**



**BÁO CÁO BÀI TẬP LỚN**

***Học phần: Hệ hỗ trợ ra quyết định và kinh doanh thông minh***

*Đề tài:*

**XÂY DỰNG KHO DỮ LIỆU NÂNG CAO HIỆU QUẢ QUẢN LÝ VÀ**

**BÁO CÁO PHÂN TÍCH THỐNG KÊ CHO CÔNG TY ÂM NHẠC CHINOOK**

Giảng viên hướng dẫn : ThS. Ngô Thùy Linh

Mã lớp học phần : 231IS31A03

Nhóm sinh viên thực hiện : Nhóm 14

Hà Tuấn Phong – 23A4040112

Hoàng Minh Phương – 23A4040114

Đào Thu Hằng – 23A4040035

Nguyễn Thu Hằng – 23A4040037

Ngô Thị Huyền – 23A4040053

**Hà Nội - 2023**

**DANH SÁCH THÀNH VIÊN NHÓM 14:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Thành viên** | **MSV** | **Nội dung công việc** | **Đóng góp(%)** |
| 1 | Hà Tuấn Phong | 23A4040112 | Tìm hiểu dữ liệu, câu 1-15, làm word | 20% |
| 2 | Hoàng Minh Phương | 23A4040114 | Tìm hiểu dữ liệu, câu 1-15 | 20% |
| 3 | Đào Thu Hằng | 23A4040035 | Tìm hiểu dữ liệu, câu 1-15 | 20% |
| 4 | Nguyễn Thu Hằng | 23A4040037 | Tìm hiểu dữ liệu, câu 1-15 | 20% |
| 5 | Ngô Thị Huyền | 23A4040053 | Tìm hiểu dữ liệu, câu 1-15, làm word | 20% |

**MỤC LỤC**

[**DANH MỤC HÌNH VẼ** 5](#_Toc153934469)

[**DANH MỤC CÁC CHỮ VIẾT TẮT** 7](#_Toc153934470)

[1. Trình bày bối cảnh công ty mà nhóm đang tìm hiểu: Kinh doanh sản phẩm dịch vụ gì? Đó là một chuỗi cửa hàng hay chỉ một chi nhánh? Bán hàng online hay offline? (0.5 điểm) 8](#_Toc153934471)

[2. Mô tả usecase của nhóm: có mấy phân hệ nghiệp vụ ở mức vận hành (có mấy nguồn), ở các phòng ban nào (0.5 điểm)?Trình bày mô hình cơ sở dữ liệu ở các hệ thống nghiệp vụ này? (0.5 điểm) 8](#_Toc153934472)

[3. Thiết kế 1 số báo cáo vận hành trong ngày (operational reports): câu hỏi và đáp án. (1 điểm) 10](#_Toc153934473)

[4. Thiết kế câu hỏi về một số báo cáo phân tích, thống kê (analytic reports) mà bộ phận quản lý cần? (1 điểm). Nếu chỉ dựa vào phân hệ vận hành thì có đáp ứng được không? Giải thích (0.5 điểm) 14](#_Toc153934474)

[5. Phát biểu sự cần thiết xây dựng Kho dữ liệu trong bối cảnh này. Mô tả kiến trúc kho dữ liệu dự kiến của nhóm? (1 điểm) 19](#_Toc153934475)

[6. Nghiệp vụ mà nhóm quan tâm? Xác định độ chi tiết của dữ liệu – dựa vào cơ sở nào để đưa ra được mức độ chi tiết này (1 điểm) 21](#_Toc153934476)

[7. Thiết kế các bảng chiều (bảng cắt lớp) DIM: giải thích tại sao lại có các bảng DIM này, mô tả chi tiết từng bảng, nguồn của bảng DIM lấy từ bảng nào? (2 điểm) 24](#_Toc153934477)

[7.1. Xây dựng bảng cắt lớp thời gian DIM\_DATE 25](#_Toc153934478)

[7.2. Xây dựng bảng cắt lớp sản phẩm nhạc DIM\_TRACK 29](#_Toc153934479)

[7.3. Xây dựng bảng cắt lớp khách hàng DIM\_CUSTOMER 31](#_Toc153934480)

[7.4. Xây dựng bảng cắt lớp nhân viên DIM\_EMPLOYEE 32](#_Toc153934481)

[8. Thiết kế bảng FACT. Mô hình lược đồ trên Data mart? Mô hình sao hay bông tuyết? (2 điểm) 33](#_Toc153934482)

[8.1. Thiết kế bảng FACT 33](#_Toc153934483)

[8.2. Mô hình lược đồ trên Data mart 34](#_Toc153934484)

[9. Đổ dữ liệu từ các bảng nguồn vào các bảng DIM, minh họa bằng hình ảnh việc đổ dữ liệu thành công (2 điểm) 37](#_Toc153934485)

[10. Đổ dữ liệu vào bảng Fact, minh họa đổ dữ liệu thành công (2 điểm) 40](#_Toc153934486)

[11. Sử dụng kiểu SCD nào để tải dữ liệu (data loading) từ hệ thống NGUỒN vào bảng DIM? Minh họa (1 điểm) 42](#_Toc153934487)

[12. Thiết kế các báo cáo phân tích thống kê (tương ứng với câu 4 trong bài kiểm tra 1) sử dụng bất cứ tool nào (SQL Server Business Intelligence, Tableau, Power BI, Olap Excel,…) để minh họa các thao tác roll-up; drill-down; slice; dice (2 điểm) 48](#_Toc153934488)

[13. Sau quá trình thử nghiệm cài đặt hệ thống và khai thác dữ liệu trên Cube, nhóm hãy nêu một số đánh giá về hiệu quả của DW về: khả năng truy vấn, khai thác dữ liệu, … ? (0.5 điểm). Loại DB nào phù hợp để thiết kế DW? Giải thích? (0.5 điểm) 57](#_Toc153934489)

[14. Khi dữ liệu từ NGUỒN được tải vào Data mart (tại bảng Dim, Fact) thì nhóm em có sử dụng kỹ thuật học máy, học sâu, luật kết hợp… để xây dựng mô hình dự báo, dự đoán (trên tập dữ liệu ở data mart) hỗ trợ ra quyết định mức chiến lược, chiến thuật không? (1 điểm) 60](#_Toc153934490)

[15. Nhóm hãy phát huy trí tưởng tượng và thiết kế kịch bản để case study mà nhóm đang nghiên cứu hướng đến việc triển khai data lake: nêu lý do/thách thức (làm thế nào gây xúc động cho người nghe là được) và trình bày giải pháp (sử dụng dịch vụ nào của nhà cung cấp nào) có minh họa kiến trúc bằng hình ảnh, lợi ích (1 điểm) 65](#_Toc153934491)

[15.1.Những điều cần biết về Data Lake 66](#_Toc153934492)

[15.2. Tình huống giả lập 68](#_Toc153934493)

[**DANH MỤC THAM KHẢO** 70](#_Toc153934494)

**DANH MỤC HÌNH VẼ**

[Hình 1. Mô hình dữ liệu quan hệ 9](#_Toc153934571)

[Hình 2. Tổng doanh thu bán hàng 11](#_Toc153934572)

[Hình 3. Doanh thu bán hàng theo khách hàng 12](#_Toc153934573)

[Hình 4. Doanh thu bán hàng theo khu vực 13](#_Toc153934574)

[Hình 5. Doanh thu bán hàng theo loại sản phẩm âm nhạc 14](#_Toc153934575)

[Hình 6. Tổng doanh số bán hàng của tất cả các sản phẩm trong tháng 7/2010 17](#_Toc153934576)

[Hình 7. Doanh số bán hàng của từng sản phẩm trong tháng 7/2010 18](#_Toc153934577)

[Hình 8. Mô hình kiến trúc kho dữ liệu (Ralph Kimball, 2013, p. 54) 21](#_Toc153934578)

[Hình 9. Độ chi tiết dữ liệu thời gian 22](#_Toc153934579)

[Hình 10. Độ chi tiết dữ liệu khách hàng 23](#_Toc153934580)

[Hình 11. Độ chi tiết dữ liệu Track 23](#_Toc153934581)

[Hình 12. Mô hình Datamart dự kiến 24](#_Toc153934582)

[Hình 13. Bảng D\_Date 25](#_Toc153934583)

[Hình 14. Bảng Track 30](#_Toc153934584)

[Hình 15. Bảng D\_Track 31](#_Toc153934585)

[Hình 16. Bảng D\_Customer 32](#_Toc153934586)

[Hình 17. Bảng D\_Employee 33](#_Toc153934587)

[Hình 18. Bảng F\_Invoice 34](#_Toc153934588)

[Hình 19. Mô hình Datamart hình sao 36](#_Toc153934589)

[Hình 20. Mô hình đổ dữ liệu bảng Dim\_Employee 38](#_Toc153934590)

[Hình 21. Kết quả đổ bảng Dim\_Employee 38](#_Toc153934591)

[Hình 22. Mô hình đổ dữ liệu bảng Dim\_Customer 39](#_Toc153934592)

[Hình 23. Kết quả đổ bảng Dim\_Customer 39](#_Toc153934593)

[Hình 24. Mô hình đổ dữ liệu bảng Dim\_Track 40](#_Toc153934594)

[Hình 25. Kết quả đổ dữ liệu bảng Dim\_Track 40](#_Toc153934595)

[Hình 26. Mô hình đổ dữ liệu bảng Fact\_Invoice 41](#_Toc153934596)

[Hình 27. Kết quả đổ dữ liệu bảng Fact\_Invoice 42](#_Toc153934597)

[Hình 28.Cài đặt SCD cho bảng Dim\_Employee 43](#_Toc153934598)

[Hình 29. Mô hình SCD bảng Dim\_Employee 44](#_Toc153934599)

[Hình 30. Kết quả SCD bảng Dim\_Employee 44](#_Toc153934600)

[Hình 31. Cài đặt SCD cho bảng Dim\_Customer 45](#_Toc153934601)

[Hình 32. Kết quả SCD bảng Dim\_Customer 46](#_Toc153934602)

[Hình 33. Cài đặt SCD bảng Dim\_Track 47](#_Toc153934603)

[Hình 34. Mô hình SCD bảng Dim\_Track 47](#_Toc153934604)

[Hình 35. Kết quả SCD bảng Dim\_Track 48](#_Toc153934605)

[Hình 36. Báo cáo thống kê về doanh thu bán lẻ theo thời gian 49](#_Toc153934606)

[Hình 37. Báo cáo thống kê doanh số bán lẻ theo quốc gia, bang, thành phố. 50](#_Toc153934607)

[Hình 38. Báo cáo thống kê doanh số bán lẻ theo loại sản phẩm, thể loại nhạc 51](#_Toc153934608)

[Hình 39. Báo cáo thống kê khách hàng có xu hướng mua hàng thường xuyên 52](#_Toc153934609)

[Hình 40. Báo cáo thống kê tỷ lệ tăng trưởng doanh số bán hàng tất cả các năm 53](#_Toc153934610)

[Hình 41. Báo cáo thống kê các sản phẩm bán chạy nhất hoặc bán chậm nhất theo tháng, theo năm 54](#_Toc153934611)

[Hình 42. Báo cáo thống kê về các sản phẩm bán chạy nhất hoặc bán chậm nhất theo năm 55](#_Toc153934612)

[Hình 43. Bộ lọc cho TrackID 55](#_Toc153934613)

[Hình 44. Bộ lọc cho Year 56](#_Toc153934614)

[Hình 45. Báo cáo thống kê doanh số bán hàng của từng nhân viên theo tháng, quý, năm 57](#_Toc153934615)

[Hình 46. Khai báo mô hình và đọc dữ liệu. 61](#_Toc153934616)

[Hình 47. Chuyển đổi dữ liệu dạng chữ sang dạng số 62](#_Toc153934617)

[Hình 48. Xác định các thuộc tính mô tả. 63](#_Toc153934618)

[Hình 49. Chia bộ dữ liệu thành hai tập train và test 64](#_Toc153934619)

[Hình 50. Khai báo mô hình và thực thi. 65](#_Toc153934620)

[Hình 51. Kiểm thử mô hình 65](#_Toc153934621)

[Hình 52. AWS Cloud 69](#_Toc153934622)

# DANH MỤC CÁC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **STT** | **Chữ viết tắt** | **Ý nghĩa** |
| 1 | TBD | To be determined |
| 2 | ERP | Enterprise Resource Planning |

# Trình bày bối cảnh công ty mà nhóm đang tìm hiểu: Kinh doanh sản phẩm dịch vụ gì? Đó là một chuỗi cửa hàng hay chỉ một chi nhánh? Bán hàng online hay offline? (0.5 điểm)

Mô hình dữ liệu Chinook đại diện cho một cửa hàng âm nhạc, bao gồm các thông tin về nhạc sĩ, album, bài hát, thể loại, hóa đơn, nhân viên và khách hàng. Thông tin các giao dịch mua sản phẩm âm nhạc trong khoảng thời gian từ năm 2009 đến 2013.

Chinook không có chuỗi cửa hàng. Chinook bán hàng chủ yếu là online và offline chủ yếu là để cho khách hàng trải nghiệm. Công ty bán hàng online thông qua trang web của mình. Công ty bán hàng online thông qua các kênh phân phối, bán lẻ sản phẩm dịch vụ là hình thức chính của Database Chinook.

Lĩnh vực kinh doanh: Bán lẻ nhạc trực tuyến

* Sản phẩm dịch vụ: Album, đĩa đơn, bài hát, video âm nhạc, dịch vụ thuê nhạc, dịch vụ streaming nhạc
* Có chuỗi cửa hàng hay chỉ một chi nhánh: Không có chuỗi cửa hàng
* Bán hàng online hay offline: Bán hàng cả online và chỉ có 1 cửa hàng offline ở Canada.

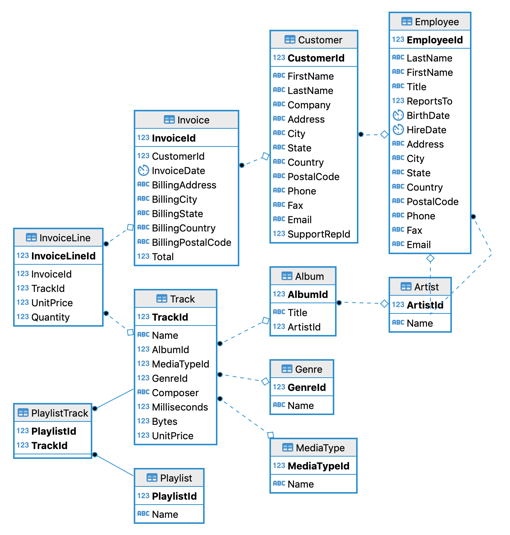
# Mô tả usecase của nhóm: có mấy phân hệ nghiệp vụ ở mức vận hành (có mấy nguồn), ở các phòng ban nào (0.5 điểm)?Trình bày mô hình cơ sở dữ liệu ở các hệ thống nghiệp vụ này? (0.5 điểm)

Có 3 phân hệ nghiệp vụ ở mức vận hành, bao gồm:

* Phân hệ bán hàng: Quản lý dữ liệu về các hóa đơn, khách hàng và chi tiết hóa đơn. Nguồn dữ liệu chính của phân hệ này là hệ thống POS của cửa hàng nhạc.
* Phân hệ marketing: Quản lý dữ liệu về các danh sách phát, khách hàng và mối quan hệ giữa khách hàng và danh sách phát.
* Phân hệ sản phẩm: Quản lý dữ liệu về sản phẩm, giá cả sản phẩm.

Các phân hệ nghiệp vụ này nằm ở các phòng ban sau:

* Phòng bán hàng: Quản lý phân hệ bán hàng.
* Phòng marketing: Quản lý phân hệ marketing.
* Phòng sản phẩm: Quản lý phân hệ sản phẩm.



Hình 1. Mô hình dữ liệu quan hệ

Các mô hình hệ thống nghiệp vụ:

\* Hệ thống bán hàng sử dụng các bảng sau để lưu trữ dữ liệu:

* Invoices: Bảng này lưu trữ dữ liệu về các hóa đơn, chẳng hạn như ngày hóa đơn, khách hàng và tổng số tiền.
* InvoiceLine: Bảng này lưu trữ dữ liệu về các chi tiết hóa đơn, chẳng hạn như bài hát, số lượng bài hát và tổng giá tiền.
* Customers: Bảng này lưu trữ dữ liệu về khách hàng, chẳng hạn như tên khách hàng, địa chỉ và email.
* Album: Bảng này lưu trữ thông tin về các album nhạc (tên album, ngày phát hành, nghệ sĩ và thể loại).
* Artist: Bảng này lưu trữ thông tin về các nghệ sĩ (tên nghệ sĩ, quốc tịch và tiểu sử).
* Track: Bảng này chứa thông tin về các bản nhạc cụ thể, bao gồm tên bản nhạc, độ dài, và thông tin liên quan đến album.
* Genre: Bảng này lưu trữ thông tin về các thể loại nhạc rock, pop, jazz và classical).
* MediaType: Bảng này lưu trữ thông tin về các loại sản phẩm âm nhạc (MP3, CD và DVD).
* Employees: Bảng này lưu trữ dữ liệu về nhân viên bao gồm tên nhân viên, chức vụ và quản lý.

\* Hệ thống marketing sử dụng các bảng sau để lưu trữ dữ liệu:

* Playlists: Bảng này lưu trữ dữ liệu về các danh sách phát, chẳng hạn như tên danh sách phát và các bài hát trong danh sách phát.
* PlaylistTracks: Bảng này lưu trữ dữ liệu về mối quan hệ giữa danh sách phát và bài hát.

\* Hệ thống sản phẩm sử dụng các bảng sau để lưu trữ dữ liệu:

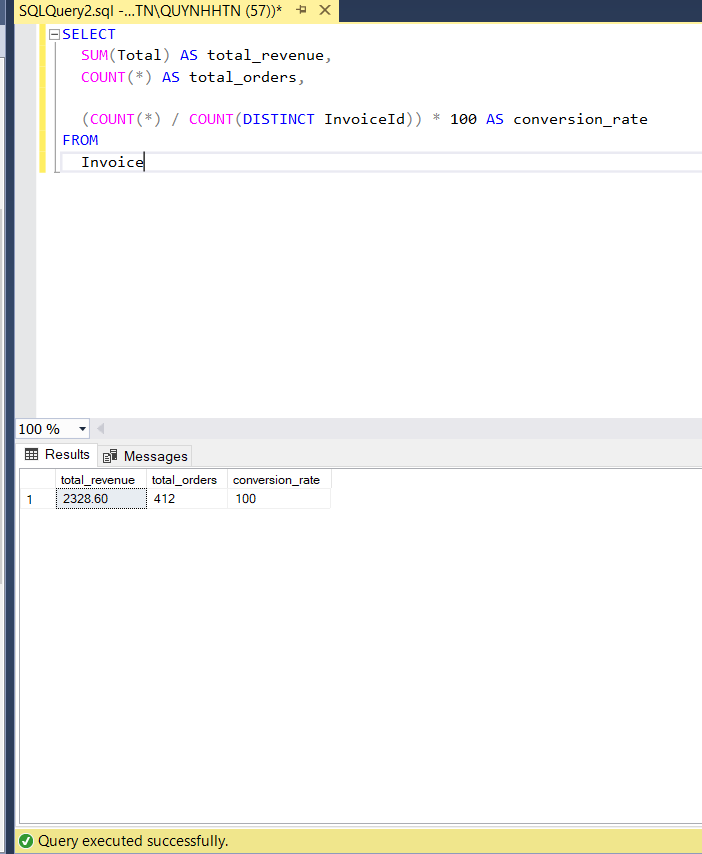
* Album: Bảng này lưu trữ thông tin về các album nhạc (tên album, ngày phát hành, nghệ sĩ và thể loại).
* Artist:Bảng này lưu trữ thông tin về các nghệ sĩ (tên nghệ sĩ, quốc tịch và tiểu sử).
* Track: Bảng này chứa thông tin về các bản nhạc cụ thể, bao gồm tên bản nhạc, độ dài, và thông tin liên quan đến album.
* Genre: Bảng này lưu trữ thông tin về các thể loại nhạc rock, pop, jazz và classical).
* MediaType: Bảng này lưu trữ thông tin về các loại sản phẩm âm nhạc (MP3, CD và DVD).

# Thiết kế 1 số báo cáo vận hành trong ngày (operational reports): câu hỏi và đáp án. (1 điểm)

Báo cáo 1: Tổng doanh thu bán hàng

Báo cáo này cung cấp thông tin tổng quan về doanh thu bán hàng của doanh nghiệp trong ngày. Báo cáo bao gồm các thông tin sau:

* Tổng doanh thu bán hàng
* Tổng số đơn hàng
* Tỷ lệ chuyển đổi

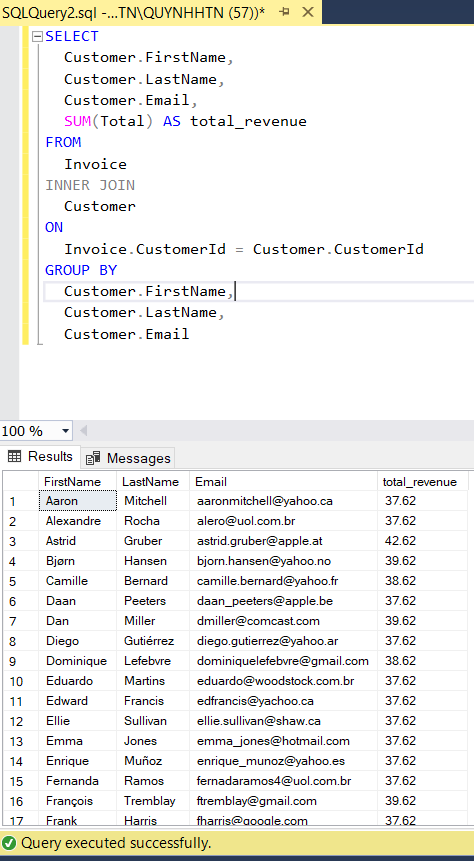


Hình 2. Tổng doanh thu bán hàng

Báo cáo 2: Doanh thu bán hàng theo khách hàng

Báo cáo này cung cấp thông tin về doanh thu bán hàng đạt được của từng khách hàng trong ngày. Báo cáo bao gồm các thông tin sau:

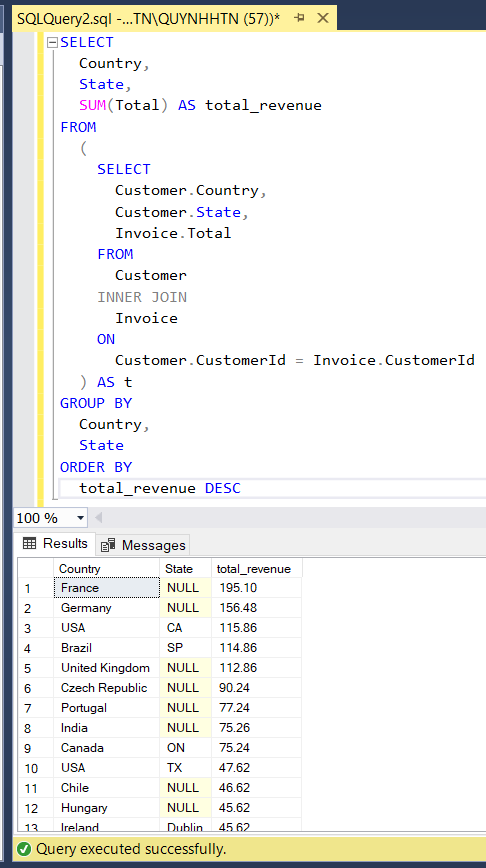
* Tên khách hàng
* Địa chỉ email của khách hàng
* Doanh thu bán hàng của khách hàng



Hình 3. Doanh thu bán hàng theo khách hàng

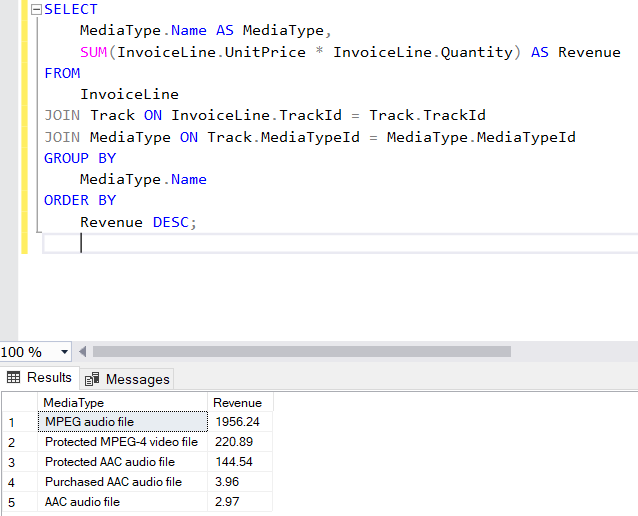
Báo cáo 3: Doanh thu bán hàng theo khu vực

* Tính tổng doanh số bán hàng cho mỗi quốc gia, bang.
* Sắp xếp kết quả theo tổng doanh số bán hàng giảm dần.



Hình 4. Doanh thu bán hàng theo khu vực

Báo cáo 4: Doanh thu bán hàng theo loại sản phẩm âm nhạc



Hình 5. Doanh thu bán hàng theo loại sản phẩm âm nhạc

# Thiết kế câu hỏi về một số báo cáo phân tích, thống kê (analytic reports) mà bộ phận quản lý cần? (1 điểm). Nếu chỉ dựa vào phân hệ vận hành thì có đáp ứng được không? Giải thích (0.5 điểm)

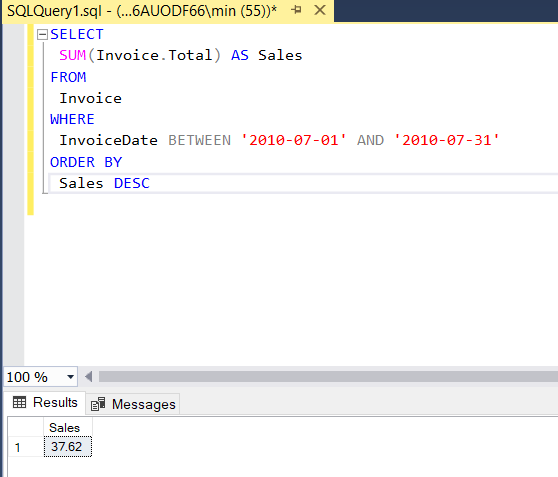
|  |  |  |  |
| --- | --- | --- | --- |
| STT | Nội dung câu hỏi | Tính khả thi | Tầm quan trọng |
| 1 | Thống kê doanh thu bán lẻ theo thời gian để phân tích sự biến động của doanh thu trong ngắn hạn (theo mùa) và trong dài hạn. | Có | Trung bình |
| 2 | Thống kê doanh số bán lẻ theo quốc gia, bang, thành phố của khách hàng để xác định được thị trường tiềm năng. | Có | Cao |
| 3 | Thống kê doanh số bán lẻ theo loại sản phẩm, thể loại nhạc để phân tích xu hướng thị trường về loại sản phẩm và thể loại nhạc được yêu thích. | Có | Cao |
| 4 | Thống kê các sản phẩm được mua nhiều từ những khách hàng mới theo thời gian, để tiếp cận thêm các khách hàng mới khác trên thị trường | Không | Không |
| 5 | Thống kê khách hàng có xu hướng mua hàng thường xuyên để xác định được nhóm khách hàng tiềm năng. | Có | Cao |
| 6 | Thống kê các sản phẩm thường được mua cùng trong một hóa đơn tại một vài thời điểm nhất định để phân tích thói quen, xu hướng mua hàng từ đó xây dựng chiến lược sản phẩm đi kèm | Có | Cao |
| 7 | Thống kê tỷ lệ tăng trưởng doanh số bán hàng tất cả các năm (gồm cả các ngày lễ và sự kiện đặc biệt) để xác định xu hướng thị trường, từ đó đưa ra các chiến lược kinh doanh hiệu quả hơn. | Có | Cao |
| 8 | Thống kê các sản phẩm của những khách hàng đã hủy bỏ đăng ký dịch vụ của Chinook, để hiểu rõ hơn về lý do khách hàng hủy bỏ dịch vụ. Từ đó, đưa ra các biện pháp cải thiện dịch vụ và giữ chân khách hàng. | Không | Không |
| 9 | Thống kê các sản phẩm bán chạy nhất hoặc bán chậm nhất theo tháng, theo năm để đưa ra quyết định sản phẩm mà Chinook cần tiếp tục sản xuất, cần cải thiện hoặc ngừng sản xuất | Có | Trung bình |
| 10 | Thống kê các sản phẩm có doanh số bán hàng cao nhất ở thị trường Châu Á để đưa ra các chiến lược kinh doanh mới. | Không | Không |
| 11 | Thống kê doanh số bán hàng của từng nhân viên theo tháng, quý, năm để phân tích hiệu suất làm việc của từng nhân viên. | Có | Trung bình |

Nếu chỉ dựa vào phân hệ vận hành thì không thể đáp ứng đầy đủ nhu cầu phân tích thống kê số liệu của bộ phận quản lý. Các phân hệ vận hành chủ yếu tập trung vào việc xử lý các nghiệp vụ kinh doanh hàng ngày. Các phân hệ này thường không có khả năng cung cấp các báo cáo phân tích thống kê chi tiết và toàn diện.

Một số lý do tại sao chỉ dựa vào phân hệ vận hành không thể đáp ứng đầy đủ nhu cầu phân tích thống kê số liệu của bộ phận quản lý:

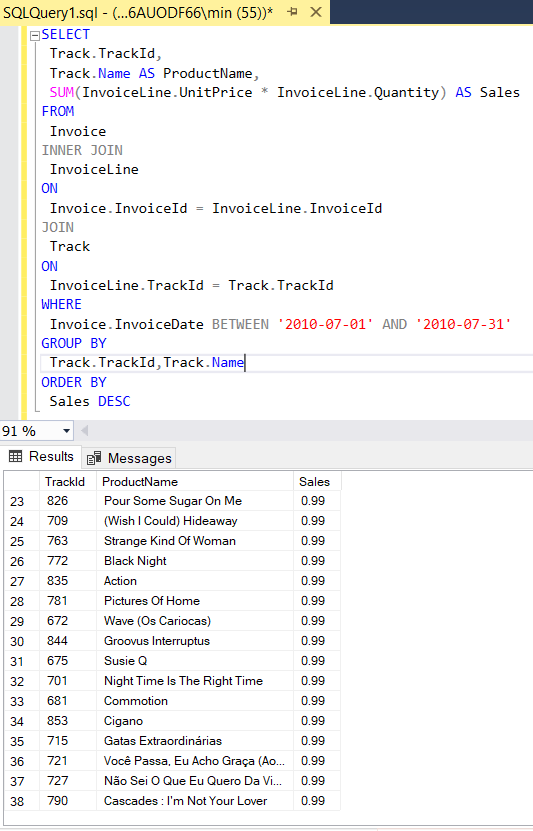
* Các phân hệ vận hành thường không có khả năng cung cấp các báo cáo phân tích thống kê chi tiết. Các báo cáo phân tích thống kê chi tiết thường yêu cầu các phân hệ báo cáo có khả năng truy cập và xử lý dữ liệu từ nhiều nguồn khác nhau.
* Các phân hệ vận hành thường không có khả năng cung cấp các báo cáo phân tích thống kê toàn diện. Các báo cáo phân tích thống kê toàn diện thường yêu cầu các phân hệ báo cáo có khả năng tổng hợp dữ liệu từ nhiều phân hệ khác nhau.

Dựa vào phân hệ vận hành của database Chinook có thể biết được tổng doanh số bán hàng của tất cả các sản phẩm trong tháng 7/2010 là 2328.60 USD.



Hình 6. Tổng doanh số bán hàng của tất cả các sản phẩm trong tháng 7/2010

Tuy nhiên, bộ phận quản lý cần biết doanh số bán hàng của từng sản phẩm trong tháng 7/2010 để có thể đưa ra các quyết định kinh doanh hiệu quả như: Sản phẩm nào đang bán chạy nhất? Sản phẩm nào có doanh số bán hàng tăng trưởng? Ví dụ dưới đây là câu truy vấn về doanh số bán hàng của từng sản phẩm:



Hình 7. Doanh số bán hàng của từng sản phẩm trong tháng 7/2010

Mặc dù, dữ liệu ở mức vận hành vẫn có thể truy vấn được nhưng sẽ rất rườm rà, khó thực hiện và mất nhiều thời gian, gây khó khăn trong việc quản lý và truy vấn dữ liệu, do đó ta cần xây dựng kho dữ liệu để giúp có những báo cáo nhanh chóng, hiệu quả.

# Phát biểu sự cần thiết xây dựng Kho dữ liệu trong bối cảnh này. Mô tả kiến trúc kho dữ liệu dự kiến của nhóm? (1 điểm)

Chinook mô phỏng một hệ thống quản lý bán hàng âm nhạc, với các thông tin về các nghệ sĩ, album, bài hát, khách hàng, đơn đặt hàng, và các yếu tố khác liên quan đến hoạt động bán hàng trong một cửa hàng âm nhạc.Chinook, giống như bất kỳ hệ thống thông tin nào khác, sẽ gặp khó khăn nếu không có một kho dữ liệu (Data Warehouse) để tổ chức, lưu trữ và quản lý thông tin một cách hiệu quả. Dưới đây là mô tả khái quát về những khó khăn mà Chinook có thể đối mặt khi hoạt động mà không có kho dữ liệu. Do dữ liệu của Chinook được phân tán trong nhiều bảng, nếu không có kho dữ liệu, sự đồng bộ giữa các bảng có thể khó khăn. Điều này có thể dẫn đến sự thiếu tính toàn vẹn và chính xác trong thông tin. Khi không có một nơi tập trung chứa lịch sử và dữ liệu lớn, việc dự báo và dự đoán tình hình kinh doanh sẽ trở nên khó khăn và không hiệu quả.

Trong bối cảnh dữ liệu Chinook, việc xây dựng kho dữ liệu là cần thiết vì những lý do sau:

* Tăng khả năng truy cập dữ liệu: Kho dữ liệu cung cấp một cách truy cập dữ liệu thống nhất và toàn diện, giúp người dùng dễ dàng tìm kiếm và phân tích dữ liệu.
* Tăng hiệu quả của việc phân tích dữ liệu: Kho dữ liệu cho phép người dùng phân tích dữ liệu nhanh chóng và dễ dàng hơn, nhờ khả năng lưu trữ và tổ chức dữ liệu một cách hiệu quả.
* Tăng khả năng chia sẻ dữ liệu: Kho dữ liệu cho phép người dùng chia sẻ dữ liệu một cách dễ dàng và an toàn, giúp thúc đẩy hợp tác và cộng tác giữa các bộ phận trong doanh nghiệp.

Cụ thể, dữ liệu Chinook bao gồm một lượng lớn dữ liệu từ nhiều nguồn khác nhau, bao gồm dữ liệu bán hàng, dữ liệu khách hàng, dữ liệu sản phẩm, và dữ liệu nhân viên. Việc xây dựng kho dữ liệu sẽ giúp doanh nghiệp tổ chức và quản lý dữ liệu này một cách hiệu quả hơn, từ đó giúp doanh nghiệp:

* Tăng cường khả năng hiểu khách hàng: Bằng cách phân tích dữ liệu bán hàng, dữ liệu khách hàng, và dữ liệu nhân viên, doanh nghiệp có thể hiểu rõ hơn về nhu cầu và sở thích của khách hàng, từ đó đưa ra các chiến lược marketing và bán hàng hiệu quả hơn.
* Nâng cao hiệu quả hoạt động: Bằng cách phân tích dữ liệu sản phẩm, doanh nghiệp có thể xác định các sản phẩm bán chạy, các sản phẩm cần cải thiện, và các sản phẩm tiềm năng. Điều này giúp doanh nghiệp tối ưu hóa danh mục sản phẩm và cải thiện hiệu quả hoạt động của chuỗi cung ứng.
* Giảm thiểu rủi ro: Bằng cách phân tích dữ liệu tài chính, doanh nghiệp có thể theo dõi tình hình tài chính của doanh nghiệp và đưa ra các quyết định kinh doanh sáng suốt.

Tóm lại, việc xây dựng kho dữ liệu là một bước quan trọng giúp doanh nghiệp khai thác tối đa giá trị của dữ liệu. Trong bối cảnh dữ liệu Chinook, việc xây dựng kho dữ liệu sẽ giúp doanh nghiệp tăng khả năng truy cập dữ liệu, tăng hiệu quả của việc phân tích dữ liệu, và tăng khả năng chia sẻ dữ liệu, từ đó giúp doanh nghiệp tăng cường khả năng hiểu khách hàng, nâng cao hiệu quả hoạt động, và giảm thiểu rủi ro.

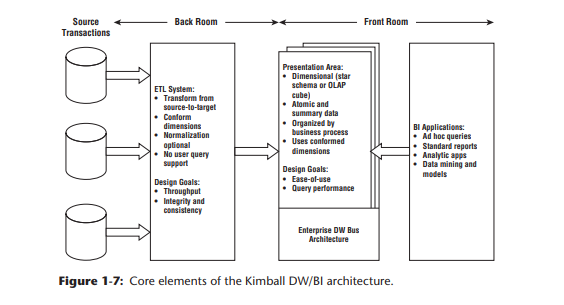
kiến trúc kho dữ liệu dự kiến của nhóm

Kiến trúc kho dữ liệu dự kiến của nhóm gồm ba tầng:

* Tầng nguồn (Source Layer): Tầng này chứa dữ liệu thô từ các hệ thống nguồn dữ liệu Chinook, chẳng hạn như hệ thống ERP
* Tầng thực thể (Dimensional Layer): Tầng này chứa dữ liệu được tổ chức theo các thực thể kinh doanh, chẳng hạn như sản phẩm, khách hàng, đơn hàng, v.v.
* Tầng phân tích (Analytical Layer): Tầng này chứa dữ liệu được phân tích để hỗ trợ cho các hoạt động phân tích dữ liệu.

Kiến trúc kho dữ liệu bao gồm các thành phần lớp sau:

* Lớp nguồn: Lớp nguồn là nơi lưu trữ dữ liệu thô từ các hệ thống khác nhau trong doanh nghiệp. Dữ liệu thô có thể được lưu trữ trong các cơ sở dữ liệu quan hệ, cơ sở dữ liệu NoSQL, các tệp dữ liệu hoặc các nguồn khác.
* Lớp trích xuất, chuyển đổi và tải (ETL): Lớp ETL chịu trách nhiệm trích xuất dữ liệu từ lớp nguồn, chuyển đổi dữ liệu sang định dạng phù hợp với kho dữ liệu và tải dữ liệu vào kho dữ liệu.
* Lớp kho dữ liệu: Lớp kho dữ liệu là nơi lưu trữ dữ liệu đã được xử lý và chuẩn hóa. Dữ liệu trong kho dữ liệu được tổ chức thành các bảng chiều và bảng sự kiện.
* Lớp truy vấn và báo cáo: Lớp truy vấn và báo cáo cung cấp cho người dùng khả năng truy vấn và phân tích dữ liệu trong kho dữ liệu.



Hình 8. Mô hình kiến trúc kho dữ liệu (Ralph Kimball, 2013, p. 54)

Kho dữ liệu Chinook mà nhóm dự kiến xây dựng dựa theo mô hình trên, theo từng tầng và các lớp của kho dữ liệu giúp cho các phân hệ vận hành trở nên dễ dàng và hiệu quả hơn. Cung cấp hiệu suất cao khi thực hiện các truy vấn và phân tích. Tích hợp các kỹ thuật như chỉ số, tối ưu hóa câu truy vấn giúp giảm thời gian truy vấn và tăng trải nghiệm người dùng. Với việc bán hàng chủ yếu bằng hình thức Online thì việc dữ liệu giao dịch được đặt lên hàng đầu và để cải thiện và phát huy các chiến lược từ các thống kê phân tích, và sử dụng kho dữ liệu là biện pháp tối ưu.

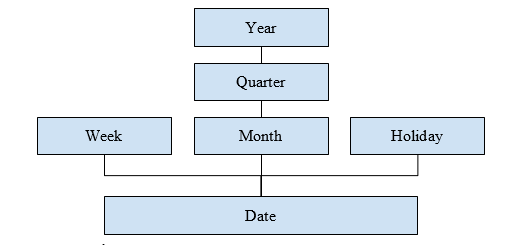
# Nghiệp vụ mà nhóm quan tâm? Xác định độ chi tiết của dữ liệu – dựa vào cơ sở nào để đưa ra được mức độ chi tiết này (1 điểm)

Nhóm quan tâm nghiệp vụ bán hàng của Chinook bởi vì nghiệp vụ đóng vai trò quan trọng trong việc thực hiện các hoạt động mua sắm sản phẩm để đáp ứng nhu cầu của khách hàng.

Độ chi tiết của dữ liệu được xác định dựa trên mục tiêu của câu hỏi.

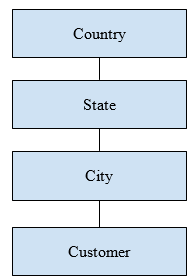
Để phân tích xu hướng bán hàng theo thời gian ở câu hỏi 1, 8, 9, 11 cần dữ liệu cụ thể về ngày đặt hàng, số lượng sản phẩm và giá trị đơn hàng,... (Alejandro Vaisman, 2014) Dữ liệu này sẽ giúp xác định được những giai đoạn nào trong năm có doanh số bán hàng cao, những sản phẩm nào được bán chạy nhất và những khách hàng nào có xu hướng mua sắm nhiều nhất.

Thống kê dữ liệu theo thời gian để phân tích sự biến động của dữ liệu trong ngắn hạn (theo mùa) hay dài hạn thì thiết kế bảng Dim Date là việc không thể thiếu.



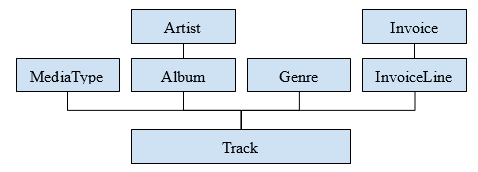
Hình 9. Độ chi tiết dữ liệu thời gian

Dữ liệu trong bảng cắt lớp Date (thời gian) có thể tổng hợp dữ liệu theo week (tuần trong tháng) và week không thể cộng dồn thành tháng nên mức cao nhất của nhánh này là week. Nhánh thứ hai của Date chỉ ra có thể tổng hợp dữ liệu theo month/quarter year (tháng trong năm/quý/năm) và có thể tổng hợp dữ liệu theo month thành quarter (quý) và quarter được tổng hợp thành year(năm). Nhánh thứ 3 là holiday (ngày lễ) và đây là mức cao nhất của nhánh, chúng ta có thể tổng hợp báo cáo theo Holiday (ngày lễ) để đưa ra báo cáo phù hợp.



Hình 10. Độ chi tiết dữ liệu khách hàng

Dữ liệu trong bảng cắt lớp Customer (Khách hàng) có thể tổng hợp dữ liệu theo City/State/Country (Thành phố/Bang/Tỉnh). Để trả lời cho câu hỏi số 2 với việc xác định doanh số bán lẻ theo từng khu vực thì mức dữ liệu chi tiết nhất phải thống kê là City (Thành phố) của khách hàng. Có thể tổng hợp dữ liệu khách hàng theo City(Thành phố), với City(Thành phố) được tổng hợp thành State(Bang), và State(Bang) được tổng hợp thành County(Tỉnh).

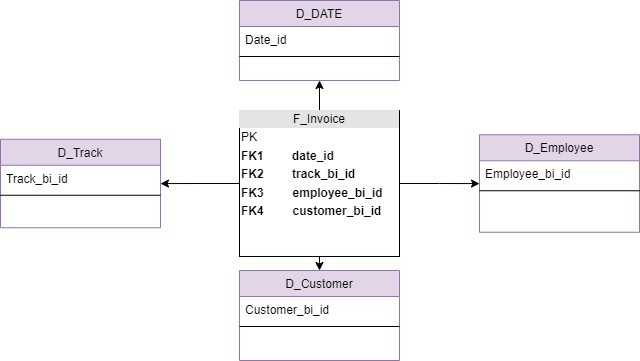


Hình 11. Độ chi tiết dữ liệu Track

Từ yêu cầu của câu hỏi 3, cần các dữ liệu chi tiết từ bảng Track. Dữ liệu trong bảng cắt lớp Track (Sản phẩm) cho thấy chúng ta có thể tổng hợp dữ liệu theo MediaType và MediaType không thể cộng dồn thành phân khúc lớn hơn. Nhánh thứ hai của Track cho thấy có thể tổng hợp dữ liệu theo Album/Artist và có thể tổng hợp dữ liệu theo Album thành Artist. Nhánh thứ 3 của Track biểu thị rằng có thể tổng hợp dữ liệu theo Genre và Genre không thể cộng dồn thành phân khúc lớn hơn. Nhánh thứ tư của Track phản ánh rằng có thể tổng hợp dữ liệu theo InvoiceLine (chi tiết hóa đơn)/Invoice (hóa đơn) và có thể tổng hợp dữ liệu theo InvoiceLine thành Invoice.

# Thiết kế các bảng chiều (bảng cắt lớp) DIM: giải thích tại sao lại có các bảng DIM này, mô tả chi tiết từng bảng, nguồn của bảng DIM lấy từ bảng nào? (2 điểm)

Để phục vụ cho nghiệp vụ bán hàng thì cần lấy dữ liệu từ các bảng: Sản phẩm (Track), Nhân viên, Khách hàng, Hóa đơn bán hàng (Invoice) và không thể không thiếu những dữ kiện về thời gian. Để xây dựng bảng Fact cho nghiệp vụ bán hàng ta xây dựng các dữ liệu xung quanh dữ liệu hóa đơn nên ra cần tạo Bảng Fact\_Invoice với các bảng D\_Track, D\_Employee, D\_Customer, D\_Date.



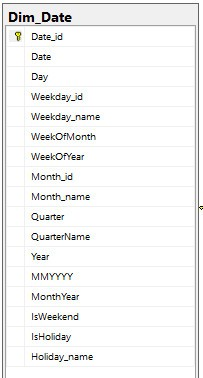
Hình 12. Mô hình Datamart dự kiến

## Xây dựng bảng cắt lớp thời gian DIM\_DATE

Thời gian là một chiều rất quan trọng được dùng trong hầu hết mọi bảng fact, Dim\_Date thường được tổ chức đặc biệt và không có nguồn nhập. Dim thời gian thường được dùng dạng tham chiếu cho nhiều chiều khác nhau. Nhóm xác định Dim\_Date là bảng không thể thiếu trong kho dữ liệu hỗ trợ cho việc quản lý bán hàng của cửa hàng.

Các lựa chọn về thời gian như ngày, tuần, tháng, quý, năm ... giúp dễ dàng quản lý , tạo các báo cáo. Đặc biệt Holiday giúp cửa hàng phân biệt ngày thường ngày lễ dễ dàng đưa ra các chương trình khuyến mãi. (ví dụ như: Giáng sinh, Tết,...)

Từ đó nhóm xây dựng bảng D\_date với những trường: khóa chính PK là date\_id; và các trường thuộc tính week\_day\_id; week\_day\_name; is\_weekend; month\_id; month\_name; year; holiday\_id; holiday\_name.



Hình 13. Bảng D\_Date

Xây dựng bảng Dim\_Date:

CREATE TABLE [dbo].[Dim\_Date] (

[Date\_id] [int] NOT NULL,

[Date] [date] NOT NULL,

[Day] [int] NOT NULL,

[Weekday\_id] [tinyint] NOT NULL,

[Weekday\_name] [varchar](10) NOT NULL,

[WeekOfMonth] [tinyint] NOT NULL,

[WeekOfYear] [tinyint] NOT NULL,

[Month\_id] [tinyint] NOT NULL,

[Month\_name] [varchar](10) NOT NULL,

[Quarter] [tinyint] NOT NULL,

[QuarterName] [varchar](6) NOT NULL,

[Year] [int] NOT NULL,

[MMYYYY] [char](6) NOT NULL,

[MonthYear] [char](7) NOT NULL,

[IsWeekend] BIT NOT NULL,

[IsHoliday] BIT NOT NULL,

[Holiday\_name] NVARCHAR(50) NULL,

PRIMARY KEY CLUSTERED ([Date\_id] ASC)

);

SET NOCOUNT ON

DECLARE @StartDate DATE = '2009-01-01'

DECLARE @EndDate DATE = '2013-12-31'

WHILE @StartDate <= @EndDate

BEGIN

INSERT INTO [dbo].[Dim\_Date] (

[Date\_id],

[Date],

[Day],

[Weekday\_id],

[Weekday\_name],

[WeekOfMonth],

[WeekOfYear],

[Month\_id],

[Month\_name],

[Quarter],

[QuarterName],

[Year],

[MMYYYY],

[MonthYear],

[IsWeekend],

[IsHoliday],

[Holiday\_name]

)

SELECT

Date\_id = YEAR(@StartDate) \* 10000 + MONTH(@StartDate) \* 100 + DAY(@StartDate),

DATE = @StartDate,

Day = DAY(@StartDate),

Weekday\_id = DATEPART(dw, @StartDate),

Weekday\_name = DATENAME(dw, @StartDate),

WeekOfMonth = DATEPART(WEEK, @StartDate) - DATEPART(WEEK, DATEADD(MM, DATEDIFF(MM, 0, @StartDate), 0)) + 1,

WeekOfYear = DATEPART(wk, @StartDate),

Month\_id = MONTH(@StartDate),

Month\_name = DATENAME(mm, @StartDate),

Quarter = DATEPART(q, @StartDate),

QuarterName = CASE

WHEN DATENAME(qq, @StartDate) = 1 THEN 'first'

WHEN DATENAME(qq, @StartDate) = 2 THEN 'second'

WHEN DATENAME(qq, @StartDate) = 3 THEN 'third'

WHEN DATENAME(qq, @StartDate) = 4 THEN 'fourth'

END,

Year = YEAR(@StartDate),

MMYYYY = RIGHT('0' + CAST(MONTH(@StartDate) AS VARCHAR(2)), 2) + CAST(YEAR(@StartDate) AS VARCHAR(4)),

MonthYear = CAST(YEAR(@StartDate) AS VARCHAR(4)) + UPPER(LEFT(DATENAME(mm, @StartDate), 3)),

IsWeekend = CASE WHEN DATENAME(dw, @StartDate) IN ('Sunday', 'Saturday') THEN 1 ELSE 0 END,

IsHoliday = 0,

Holiday\_name = NULL

SET @StartDate = DATEADD(DD, 1, @StartDate)

END

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Tết dương lịch'

WHERE [Month\_id] = 1 AND [Day] = 1;

-- Lễ tình yêu 14/2

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Lễ tình yêu'

WHERE [Month\_id] = 2 AND [Day] = 14;

-- Quốc Khánh 2/9

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Quốc Khánh'

WHERE [Month\_id] = 9 AND [Day] = 2;

-- Thống nhất đất nước 30/4

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Thống nhất đất nước'

WHERE [Month\_id] = 4 AND [Day] = 30;

-- Quốc tế lao động 1/5

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Quốc tế lao động'

WHERE [Month\_id] = 5 AND [Day] = 1;

-- Quốc tế phụ nữ 8/3

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Quốc tế phụ nữ'

WHERE [Month\_id] = 3 AND [Day] = 8;

-- Phụ nữ Việt Nam 20/10

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Phụ nữ Việt Nam'

WHERE [Month\_id] = 10 AND [Day] = 20;

-- Nhà giáo Việt Nam 20/11

UPDATE Dim\_Date

SET [IsHoliday] = 1,

[Holiday\_name] = N'Nhà giáo Việt Nam'

WHERE [Month\_id] = 11 AND [Day] = 20;

## Xây dựng bảng cắt lớp sản phẩm nhạc DIM\_TRACK

Bảng D\_Track được sử dụng để lưu trữ thông tin tổng quan về các bản nhạc, giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn. Ví dụ, nếu muốn tìm hiểu về các bản nhạc được phát nhiều nhất trong một tháng nhất định thì có thể sử dụng bảng dim track để lấy thông tin về tên, album, thể loại, tác giả và các thông tin khác của các bản nhạc đó.Bảng này bao gồm các thông tin sau:

TrackId: ID của bản nhạc

Name: Tên của bản nhạc

AlbumId: ID của album chứa bản nhạc

GenreId: ID của thể loại của bản nhạc

ArtistID: ID của tác giả

UnitPrice: Giá của sản phầm

MediaTypeId: ID của định dạng của bản nhạc

Album\_title: Tiêu đề album

Artist\_name: Tên tác giả

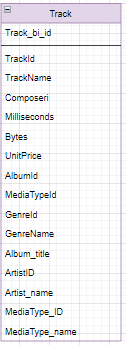
MediaType\_name

Genre\_name: tên thể loại

AlbumId, GenreId và ArtistId có trong bảng D\_Track theo dữ liệu Chinook vì chúng là các thuộc tính của một bản nhạc. Bảng dim track là một bảng tổng hợp, chứa thông tin về tất cả các bản nhạc trong cơ sở dữ liệu. Các thuộc tính AlbumId, GenreId và ArtistId, MediaTypeId giúp chúng ta phân loại và hiểu các bản nhạc tốt hơn.

Cụ thể, AlbumId nhằm xác định album mà bản nhạc thuộc về. GenreId giúp xác định thể loại của bản nhạc. ArtistId giúp xác định nghệ sĩ biểu diễn bản nhạc. MediaTypeId giúp xác định định dạng của bản nhạc.

Quan hệ giữa các bảng dữ liệu trong cây phân cấp cắt lớp D\_Track này đơn giản, chỉ là 1: n. Khi đó cấu trúc bảng cắt lớp có thể thiết kế như sau:



Hình 14. Bảng Track

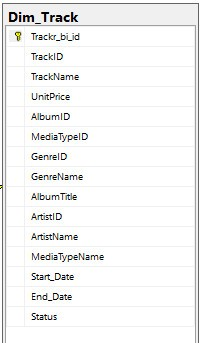
Bảng cắt lớp này là sự kết hợp của 6 bảng Album, Artist, Genre, MediaType, Playlist, PlaylistTrack trong hệ thống nghiệp vụ, cộng thêm 1 trường khoá thay thế làm khoá chính (Track\_bi\_id)

Các trường dữ liệu trong bảng cắt lớp trên không phải là bình đẳng với nhau, có những trường dữ liệu quan trọng, và có những trường dữ liệu ít quan trọng hơn. Với các trường dữ liệu ít quan trọng hơn, mỗi khi có thay đổi giá trị thì chỉ cần cập nhật giá trị mới vào bảng cắt lớp, ghi đè lên giá trị cũ. Nói “ít quan trọng” không có nghĩa nó thật sự chả quan trọng gì, mà nghĩa là giá trị các trường dữ liệu này sẽ không bao giờ thay đổi trong suốt vòng đời dự án, hoặc nếu có thay đổi thì giá trị cũ sẽ không còn ý nghĩa gì nữa. Trong cắt lớp D\_Track, các trường dữ liệu đó được xác định bao gồm Name, UnitPrice, Album\_title, Artist\_name, MediaType\_name, Genre\_name các trường mà sự thay đổi (nếu có) chỉ để sửa lại lỗi chính tả do nhập liệu sinh ra mà thôi. Một ví dụ khác về trường không quan trọng là địa chỉ email và số điện thoại một người nào đó, rõ ràng khi ai đó quyết định dùng địa chỉ email mới thì địa chỉ email cũ sẽ bị bỏ đi không còn tác dụng gì nữa.

Với các trường dữ liệu quan trọng thì khác, mỗi khi có sự thay đổi giá trị thì giá trị mới sẽ được ghi nhận những giá trị cũ vẫn phải được lưu trữ để đảm bảo bản ghi fact ở thời điểm nào sẽ nhận được giá trị cắt lớp tương ứng với thời điểm ấy. Để làm được điều này, nhóm đã thêm vào bảng cắt lớp 3 cột dữ liệu:

* START\_DATE: thời gian ghi bản ghi cắt lớp vào cơ sở dữ liệu
* END\_DATE: thời gian bản ghi hết hiệu lực (do hệ thống nghiệp vụ cập nhật thuộc tính), khi bản ghi còn hiệu lực thì trường dữ liệu này nhận giá trị null
* STATUS: hiệu lực của bản ghi.

Cuối cùng ta có cấu trúc bảng cắt lớp như sau:



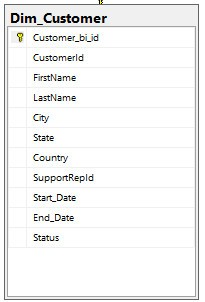
Hình 15. Bảng D\_Track

## Xây dựng bảng cắt lớp khách hàng DIM\_CUSTOMER

Bảng D\_Customer được sử dụng để lưu trữ thông tin tổng quan về khách hàng, giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn. Ví dụ, nếu muốn tìm hiểu về doanh số bán hàng cho từng khách hàng thì có thể sử dụng bảng D\_Customer để lấy thông tin về tên công ty, địa chỉ, và các thông tin khác của các khách hàng đó. Bảng này bao gồm các thông tin sau:

* CustomerId: ID của khách hàng
* City: Thành phố của khách hàng
* State: Tỉnh/thành phố của khách hàng
* Country: Quốc gia/khu vực của khách hàng

Bảng CUSTOMER là nguồn dữ liệu cho cắt lớp khách hàng không có bảng cha (quan hệ 1:n với bảng khác) (tức là đổ nguyên cấu trúc bảng nguồn, thêm một số trường dữ liệu theo kỹ thuật cập nhật cắt lớp kiểu 2). Thoạt nhìn cắt lớp này cũng chỉ có 1 cấp duy nhất là cấp khách hàng. Có thể phân tích địa chỉ khách hàng thành 3 cấp là thành phố, cấp bang quận và cấp quốc gia khi đó cấu trúc bảng D\_CUSTOMER sẽ như sau:



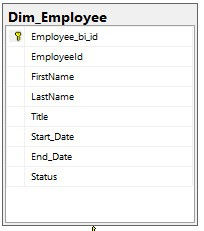
Hình 16. Bảng D\_Customer

## Xây dựng bảng cắt lớp nhân viên DIM\_EMPLOYEE

Bảng D\_Employee được sử dụng để lưu trữ thông tin tổng quan về nhân viên, giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn. Ví dụ, muốn tìm hiểu về số lượng nhân viên trong từng bộ phận thì có thể sử dụng bảng dim employee để lấy thông tin về tên, họ, chức vụ, và các thông tin khác của các nhân viên đó. Bảng này bao gồm các thông tin sau:

* EmployeeId: ID của nhân viên
* FirstName: Tên của nhân viên
* LastName: Họ của nhân viên
* Title: Chức vụ của nhân viên

Do tính chất đơn giản của nghiệp vụ, cắt lớp này chỉ có một cấp duy nhất là cấp nhân viên, tất cả mọi nhân viên đều thuộc cấp này. Qua khảo sát, cắt lớp này không có trường dữ liệu quan trọng do đó không cần áp dụng kỹ thuật cập nhật dimension kiểu số 2. Tuy nhiên để đồng nhất cấu trúc bảng cắt lớp thì vẫn nên thêm 3 cột dữ liệu bi\_start\_date, bi\_end\_date, bi\_status. Khi đó, bảng cắt lớp nhân viên có cấu trúc như sau:



Hình 17. Bảng D\_Employee

# Thiết kế bảng FACT. Mô hình lược đồ trên Data mart? Mô hình sao hay bông tuyết? (2 điểm)

## Thiết kế bảng FACT

Bảng F\_ Invoice được sử dụng để lưu trữ thông tin chi tiết về các hóa đơn bán hàng, giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn. Để đơn giản hoá, phần mềm nghiệp vụ chỉ có 1 loại giao dịch là mua sản phẩm (Track) thì chỉ nên cần Tổng tiền hóa đơn và số lượng mua hàng vậy bảng này bao gồm các thông tin sau:

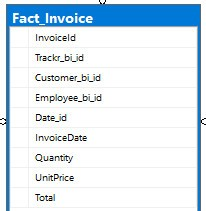
* InvoiceId: ID của hóa đơn
* CustomerId: ID của khách hàng
* EmployeeId: ID của nhân viên
* TrackId: ID của sản phẩm
* DateId: ID của thời gian
* Total: Tổng giá trị hóa đơn
* Quantity: Số lượng sản phẩm trên hóa đơn

Ví dụ, nếu bạn muốn tìm hiểu về doanh số bán hàng cho từng sản phẩm, bạn có thể sử dụng bảng F\_Invoice để lấy thông tin về số lượng sản phẩm được bán, giá của mỗi sản phẩm, và các thông tin khác.

Bảng F\_Invoice là một bảng fact giao dịch. Fact giao dịch là loại bảng fact lưu trữ dữ liệu thực tế về các sự kiện xảy ra trong thế giới thực. Các dữ liệu này thường được cập nhật thường xuyên, chẳng hạn như dữ liệu bán hàng, dữ liệu giao dịch, v.v. Bảng F\_Invoice lưu trữ dữ liệu thực tế về các giao dịch mua hàng của khách hàng. Các dữ liệu này được cập nhật thường xuyên, chẳng hạn như mỗi khi có một giao dịch mua hàng mới được thực hiện.

Dưới đây là một số lý do tại sao bảng F\_Invoice được sử dụng trong dữ liệu Chinook:

* Lưu trữ thông tin chi tiết: Bảng F\_Invoice lưu trữ thông tin chi tiết về các hóa đơn bán hàng, bao gồm ngày đặt hàng, ngày giao hàng, số lượng sản phẩm, tổng giá trị hóa đơn.
* Giúp phân tích dữ liệu dễ dàng hơn: Bảng F\_Invoice cung cấp thông tin chi tiết về các hóa đơn bán hàng, giúp cho việc phân tích dữ liệu trở nên dễ dàng hơn.
* Giúp tạo báo cáo: Bảng F\_Invoice có thể được sử dụng để tạo báo cáo về doanh số bán hàng, lợi nhuận, và các chỉ số khác.

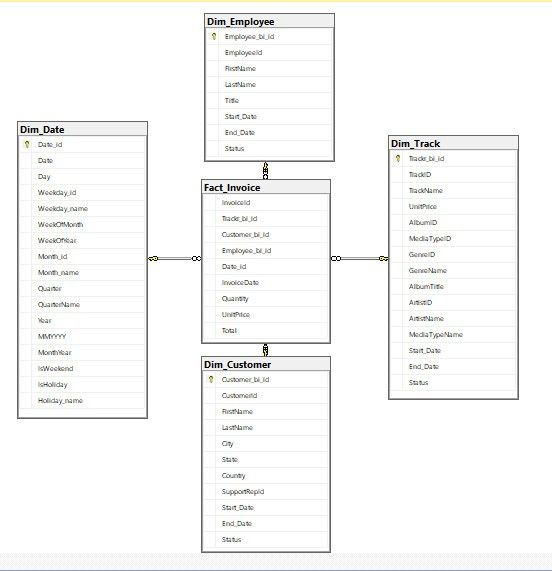


Hình 18. Bảng F\_Invoice

## Mô hình lược đồ trên Data mart

“This collection of a fact table and its dimension tables is called a data mart. There are more complex forms of data marts, such as when you have more than one fact table, but for now let’s consider only one fact table. Bear in mind that this concept of a data mart is applicable only if the data warehouse is in a dimensional model. If the data warehouse is not in a dimensional model—that is, when we are talking about a normalized data store—there is no data mart. A data mart is a group of related fact tables and their corresponding dimension tables containing the measurements of business events, categorized by their dimensions. Data marts exist in dimensional data stores”. (Rainardi, 2008, p. 92)

Data mart, tạm dịch là "kho dữ liệu con" là một phần của data warehouse được tạo ra để phục vụ một phần của tổ chức hoặc một nhóm người dùng cụ thể. Data mart tập trung vào một khía cạnh cụ thể của tổ chức, chẳng hạn như một bộ phận hoặc một dự án. Data mart thường làm việc trên một tập hợp con của dữ liệu trong data warehouse và có thể được tối ưu hóa để phục vụ nhu cầu phân tích và báo cáo cụ thể của nhóm người dùng đó.



Hình 19. Mô hình Datamart hình sao

Nhóm sử dựng mô hình hình sao cho datamart. Sử dụng mô hình hình sao (star schema) cho cơ sở dữ liệu Chinook có thể mang lại nhiều lợi ích trong việc phân tích và truy vấn dữ liệu. Sử dụng mô hình hình sao cho dữ liệu database Chinook:

* Tìm hiểu về hành vi mua hàng của khách hàng: Với mô hình hình sao, chúng ta có thể dễ dàng tìm hiểu về hành vi mua hàng của khách hàng, chẳng hạn như họ mua những loại nhạc nào, họ mua từ đâu và họ mua khi nào.
* Theo dõi doanh số bán hàng: Với mô hình hình sao, chúng ta có thể dễ dàng theo dõi doanh số bán hàng của công ty, chẳng hạn như doanh số bán hàng theo sản phẩm, theo nhân viên và theo thời gian.
* Xây dựng các báo cáo kinh doanh: Với mô hình hình sao, chúng ta có thể dễ dàng xây dựng các báo cáo kinh doanh, chẳng hạn như báo cáo doanh thu, báo cáo lợi nhuận và báo cáo chi phí.

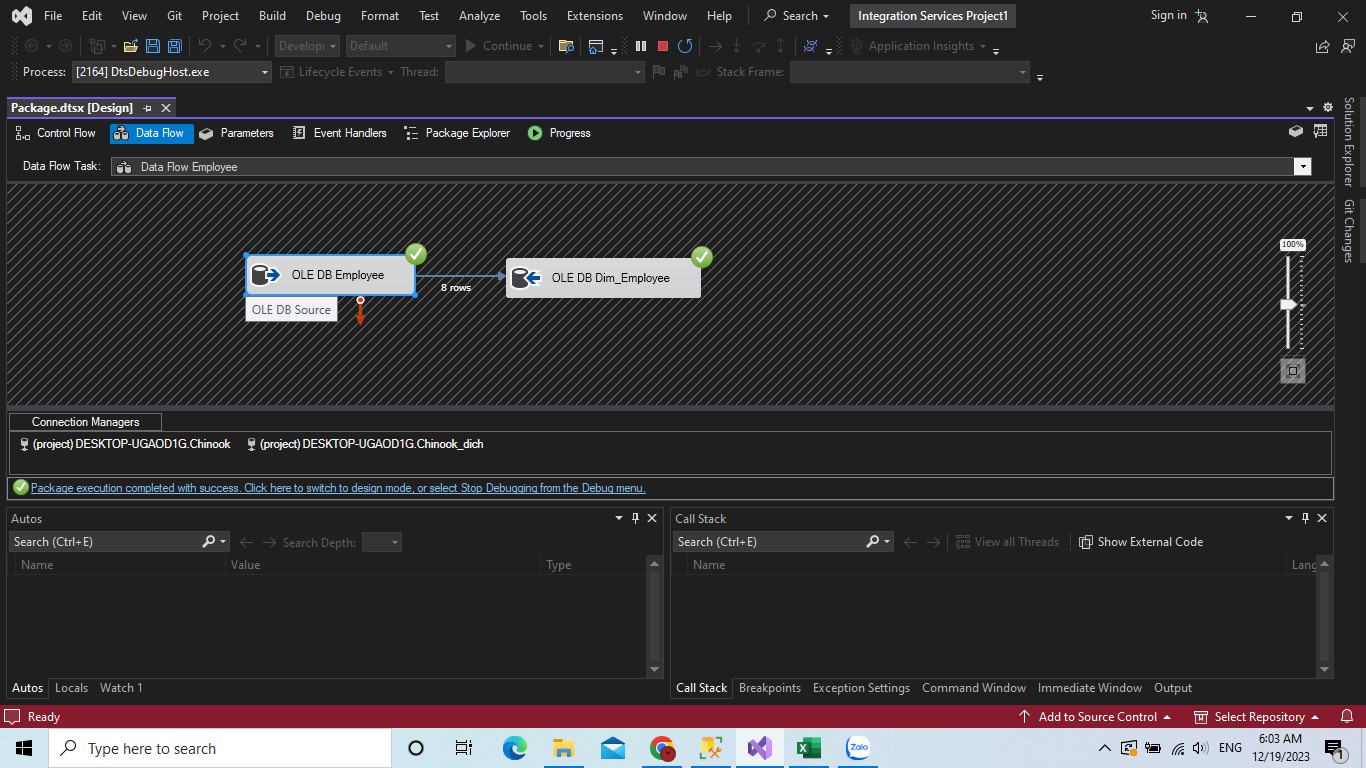
Việc sử dụng mô hình hình sao cho dữ liệu trong database Chinook giúp tạo ra một cấu trúc dữ liệu tổ chức, linh hoạt và thuận tiện cho việc phân tích dữ liệu bán hàng và làm tăng khả năng hiểu biết và đưa ra chiến lược trong lĩnh vực kinh doanh và bán hàng âm nhạc.

# Đổ dữ liệu từ các bảng nguồn vào các bảng DIM, minh họa bằng hình ảnh việc đổ dữ liệu thành công (2 điểm)

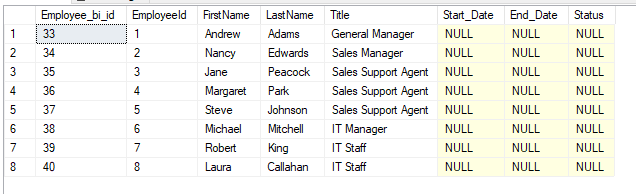
3

Đổ dữ liệu từ nguồn vào các bảng Dim, nhóm sử dụng Microsoft SQL Server Management Studio tích hợp cùng Visual Studio.Với các bảng Dim\_Employee, Dim\_Customer thì không cần dùng nhiều tool để tổng hợp và đổ dữ liệu từ nguồn vào.

* Đổ dữ liệu bảng Dim\_Employee: Đây là bước đổ dữ liệu thô chưa sử dụng các kiểu SCD để phục vụ cho việc cập nhật dữ liệu.

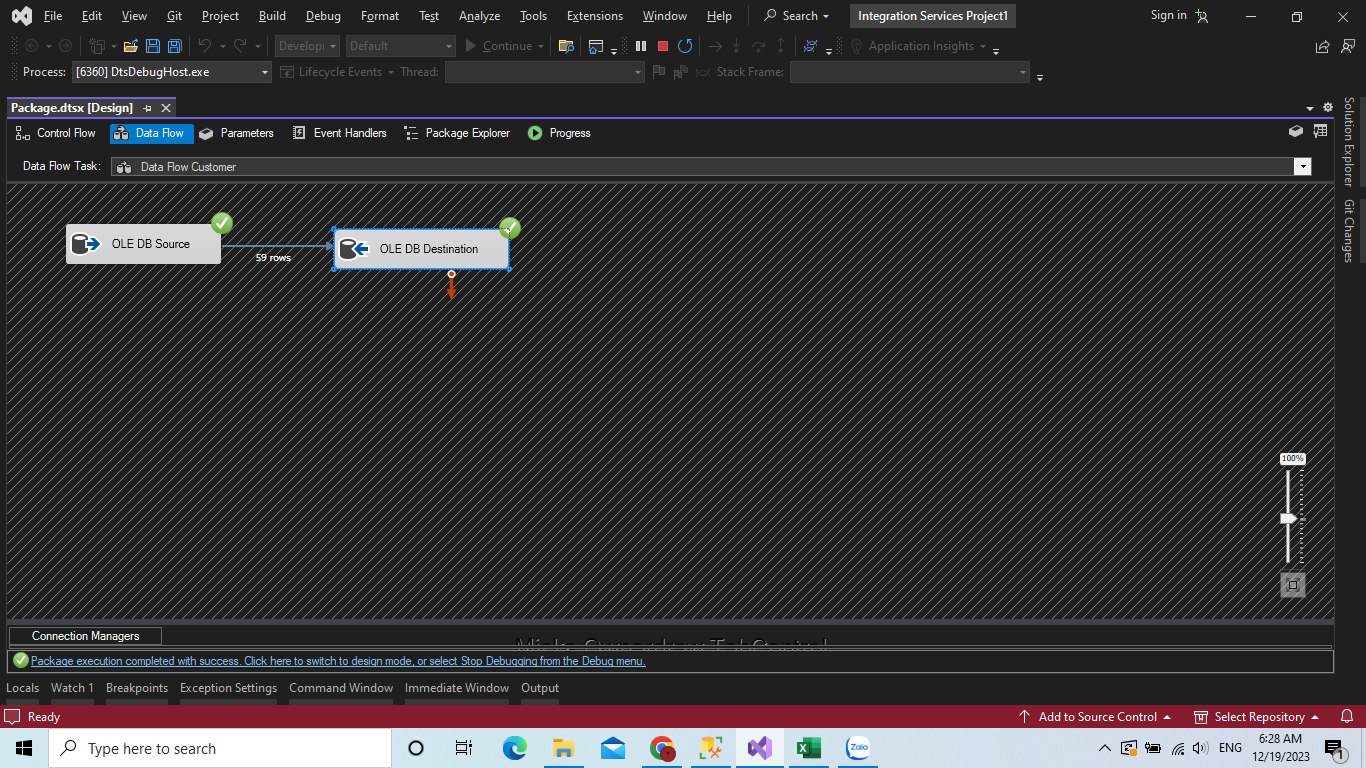


Hình 20. Mô hình đổ dữ liệu bảng Dim\_Employee

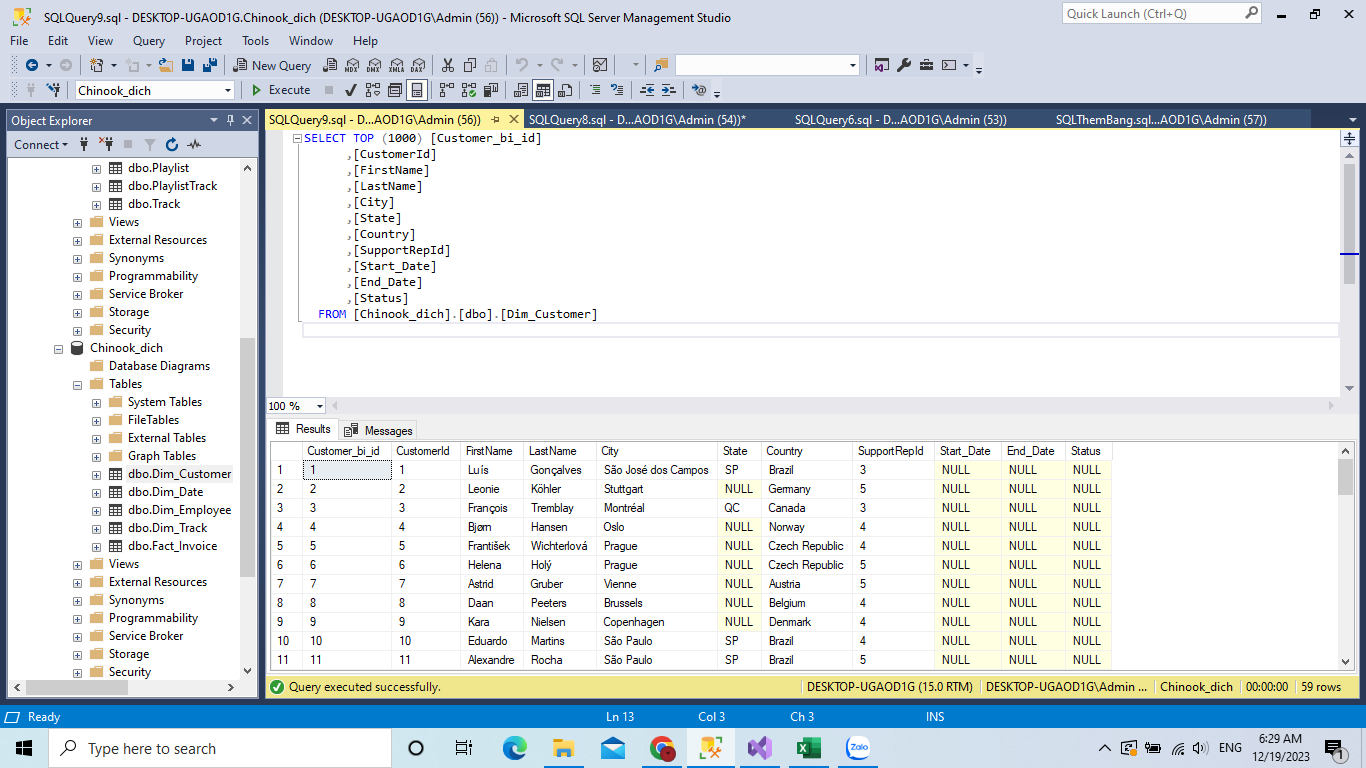


Hình 21. Kết quả đổ bảng Dim\_Employee

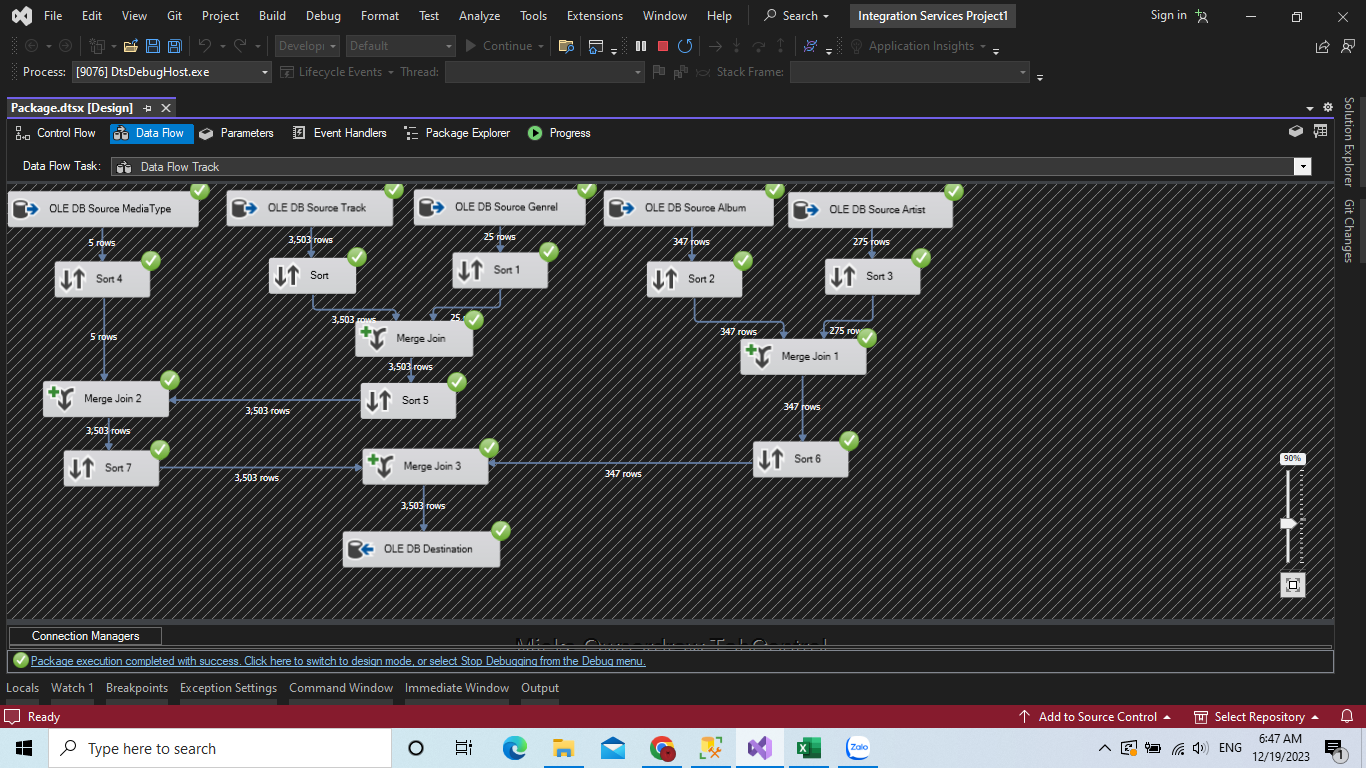
* Đổ dữ liệu bảng Dim\_Custom



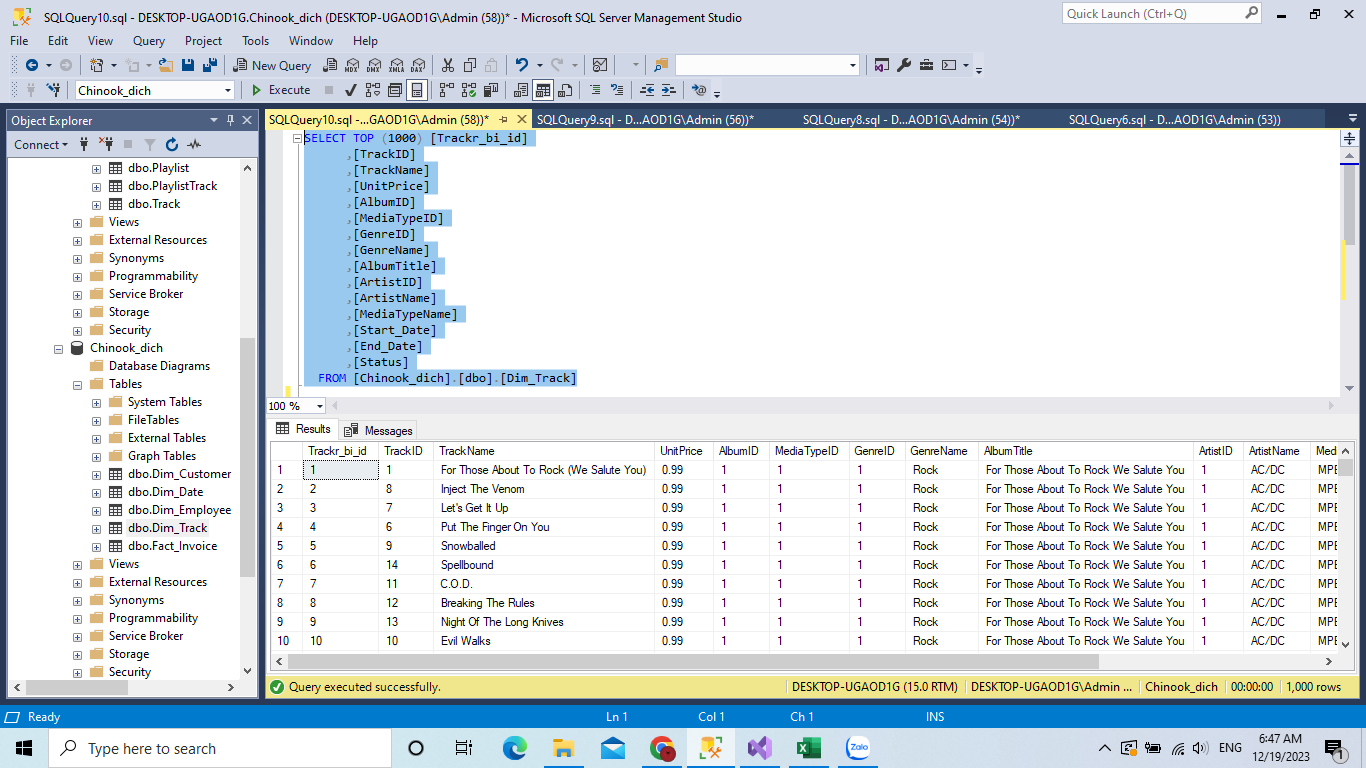
Hình 22. Mô hình đổ dữ liệu bảng Dim\_Customer

Hình 23. Kết quả đổ bảng Dim\_Customer

Đổ dữ liệu bảng Dim\_Track: Do bảng Dim\_Track là tổng hợp dữ liệu từ nhiều bảng khác nhau trong nguồn, MediaType, Track, Genre, Album, Artist. Sử dụng các tool Sort và MergeJoin để kết nối các bảng lại với nhau và đổ lại về bảng Dim\_Track



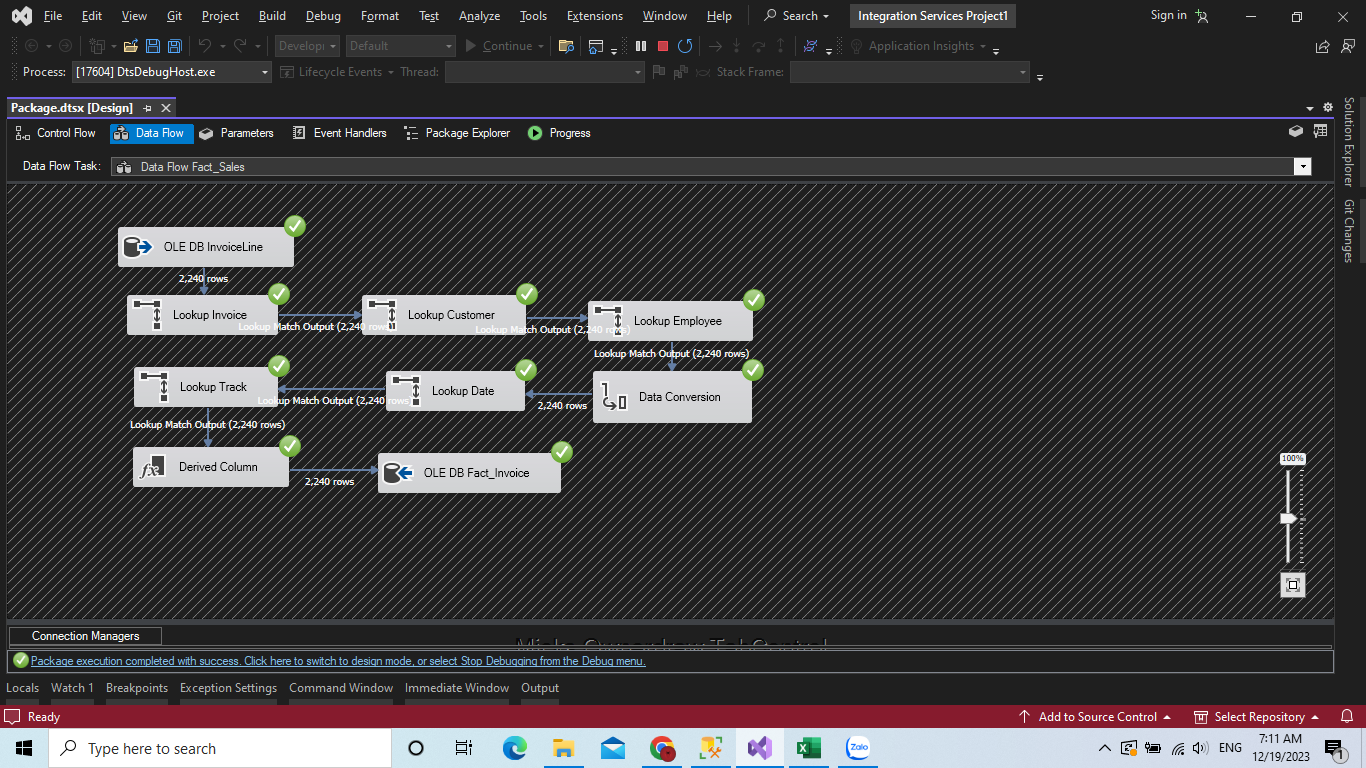
Hình 24. Mô hình đổ dữ liệu bảng Dim\_Track



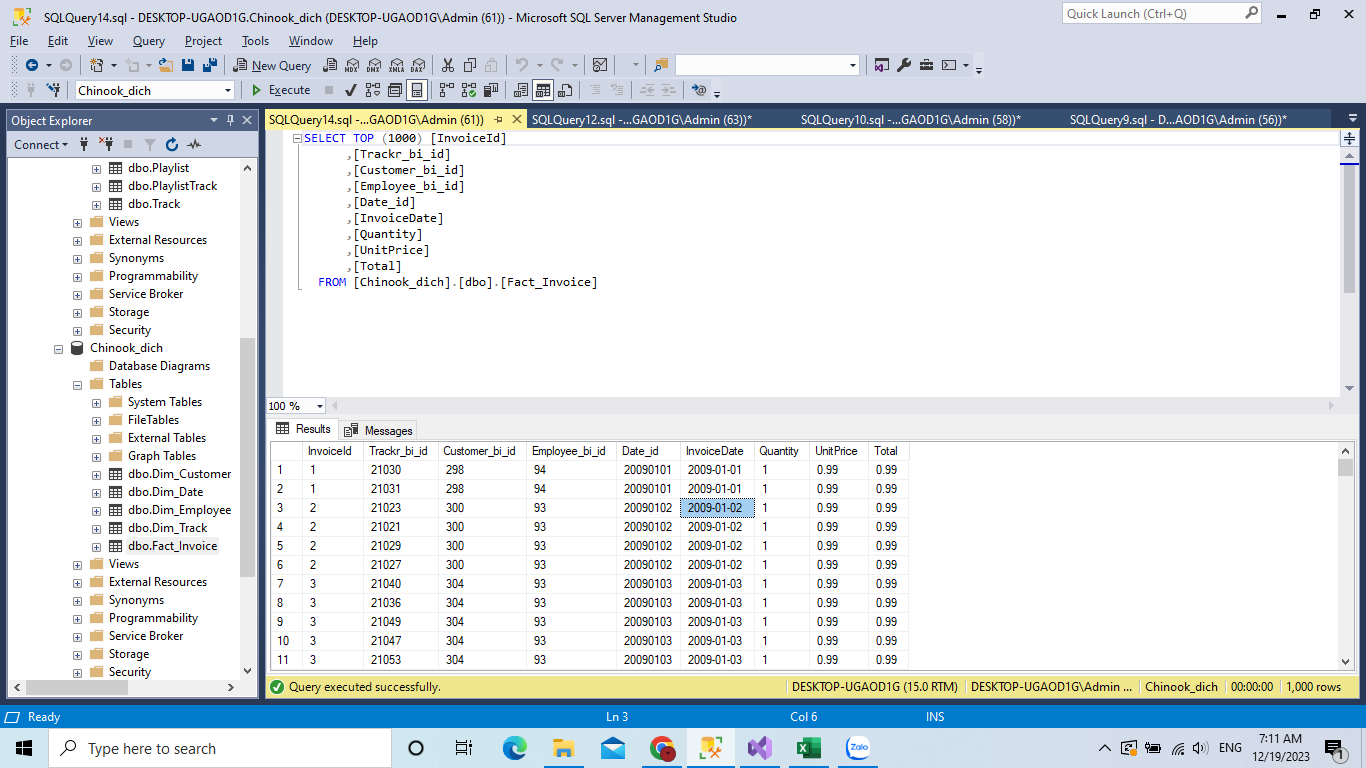
Hình 25. Kết quả đổ dữ liệu bảng Dim\_Track

10. Đổ dữ liệu vào bảng Fact, minh họa đổ dữ liệu thành công (2 điểm)

Bảng Fact\_Invoice, đây là bảng Tổng hợp có trường dữ liệu Total là kết quả của tích giữa UnitPrice và Quantity của từng sản phẩm bán ra trên từng mã hóa đơn. Để đổ dữ liệu từ nguồn vào Fact ta bắt đầu lấy bảng InvoiceLine từ nguồn rồi kết hợp qua các tool Lookup với các bảng Invoice, bảng Dim\_Customer, Dim\_Employee, Dim\_Track, Dim\_Date. Dùng tool Data Conversion để chuyển đổi kiểu dữ liệu phù hợp cho InvoiceDate, và sử dụng tool Derived Column để tạo biến Total = UnitPrice\*Quantity.



Hình 26. Mô hình đổ dữ liệu bảng Fact\_Invoice



Hình 27. Kết quả đổ dữ liệu bảng Fact\_Invoice

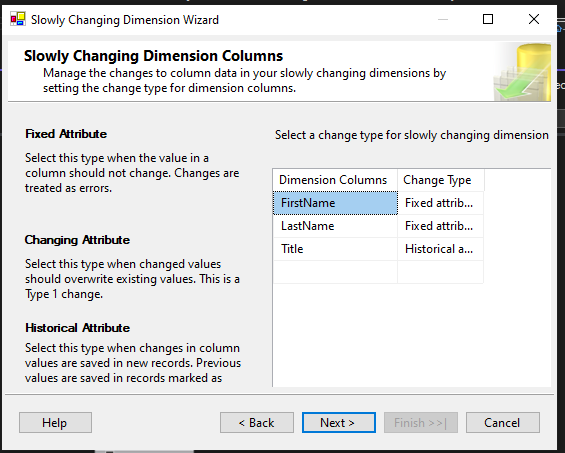
# Sử dụng kiểu SCD nào để tải dữ liệu (data loading) từ hệ thống NGUỒN vào bảng DIM? Minh họa (1 điểm)

Để tải dữ liệu từ hệ thống nguồn vào bảng Dim sử dụng kiểu Slowly Changing Dimension (SCD), một lựa chọn phổ biến là SCD Type 2. SCD Type 2 cho phép lưu trữ lịch sử của các thuộc tính của chiều dữ liệu và theo dõi sự thay đổi trong các thuộc tính đó theo thời gian.

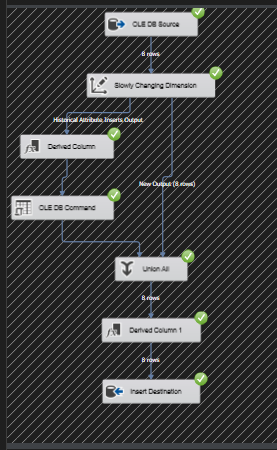
Khi áp dụng SCD Type 2, mỗi bản ghi mới sẽ được thêm vào bảng Dim với một khoá chính mới và ngày bắt đầu hiệu lực mới. Các bản ghi cũ vẫn được giữ lại, nhưng ngày kết thúc hiệu lực của chúng sẽ được cập nhật để phản ánh thời gian chúng ngừng có hiệu lực.

Lợi ích của SCD Type 2 bao gồm khả năng theo dõi lịch sử thay đổi của các thuộc tính, cung cấp khả năng tạo ra các báo cáo lịch sử và phản ánh được sự biến động trong dữ liệu theo thời gian.

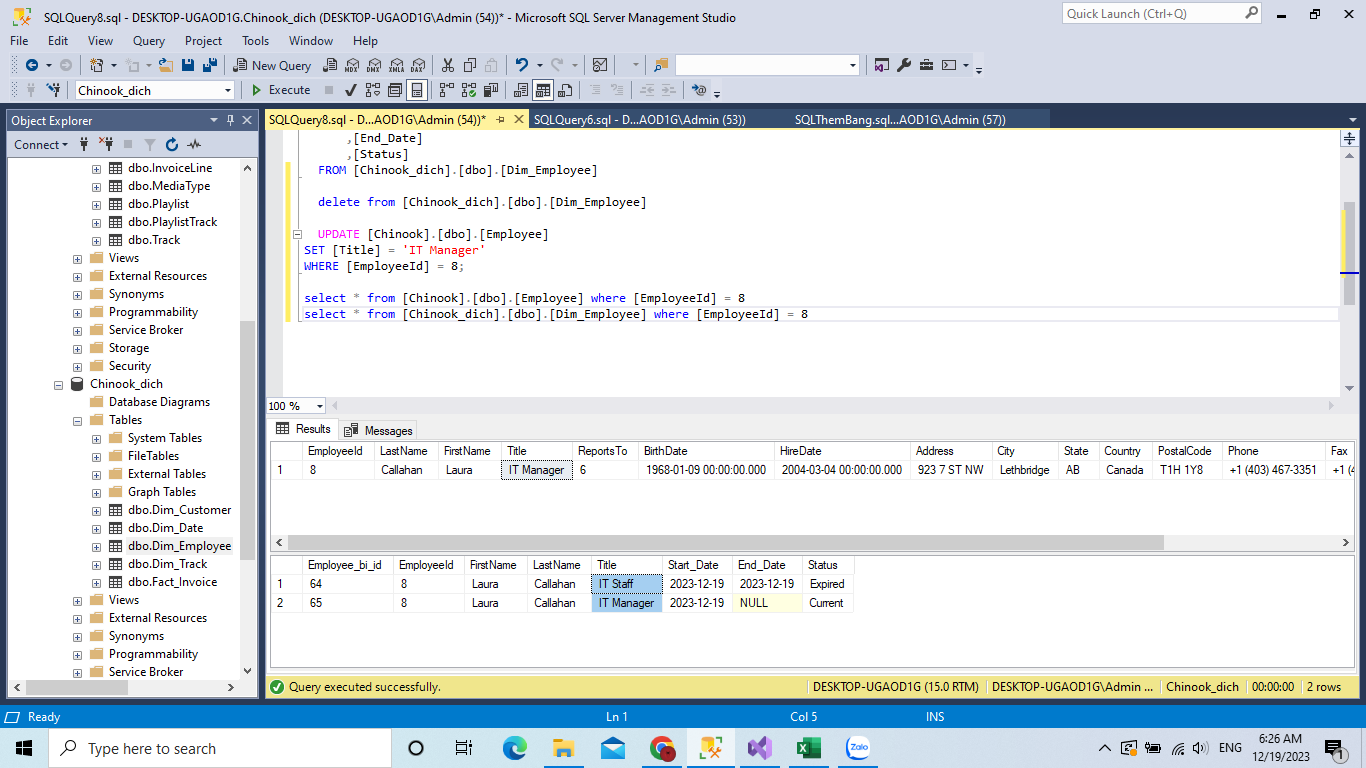
* SCD cho bảng Dim\_Employee: với trường Title ( Chức vụ) áp dụng SCD type



Hình 28.Cài đặt SCD cho bảng Dim\_Employee

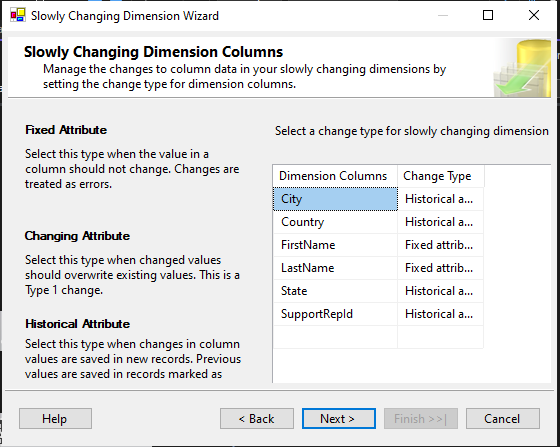


Hình 29. Mô hình SCD bảng Dim\_Employee

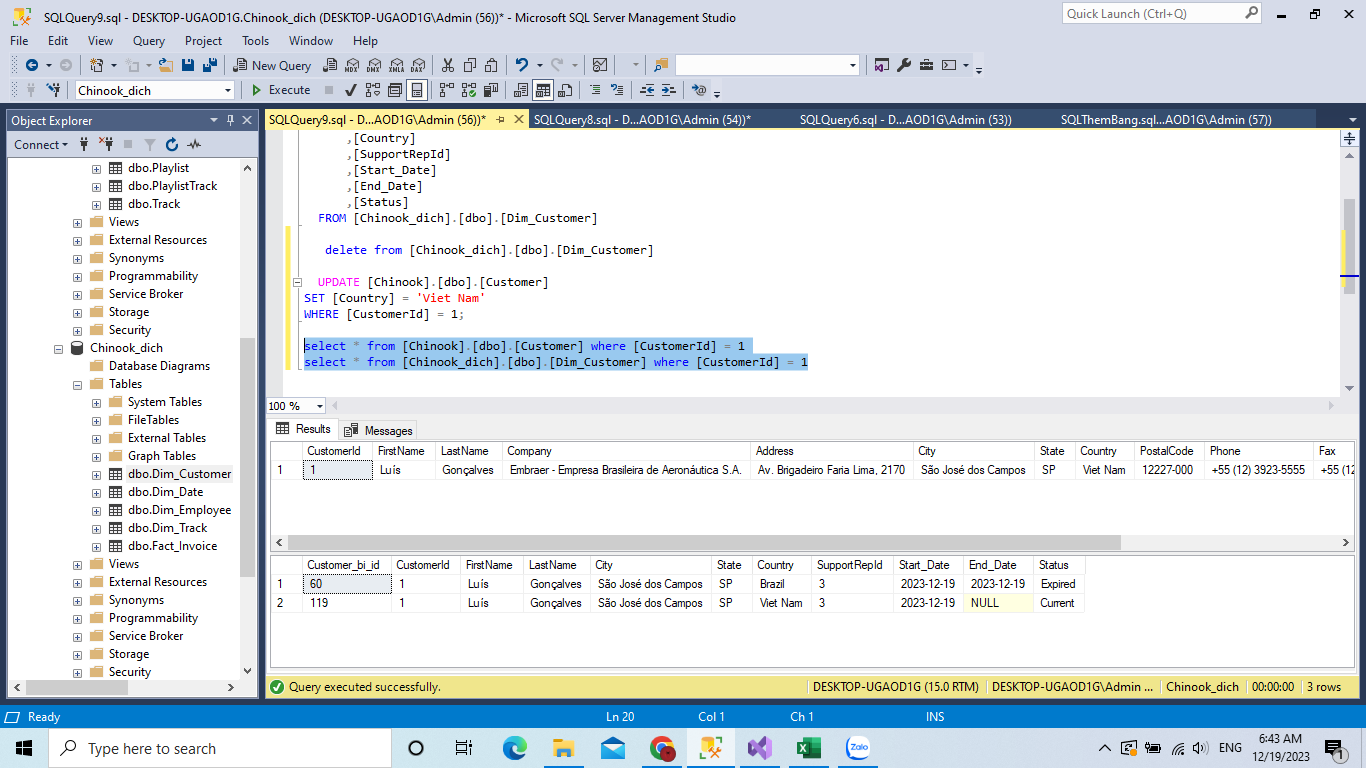


Hình 30. Kết quả SCD bảng Dim\_Employee

SCD cho bảng Dim\_Customer: với việc sử dụng mô hình tương tự như SCD cho bảng Dim\_Employee ta được kết quả như sau. Với các trường City, State, Country và SupportRepId (Mã nhân viên hỗ trợ khách hàng, xử lý hóa đơn) áp dụng SCD type 2

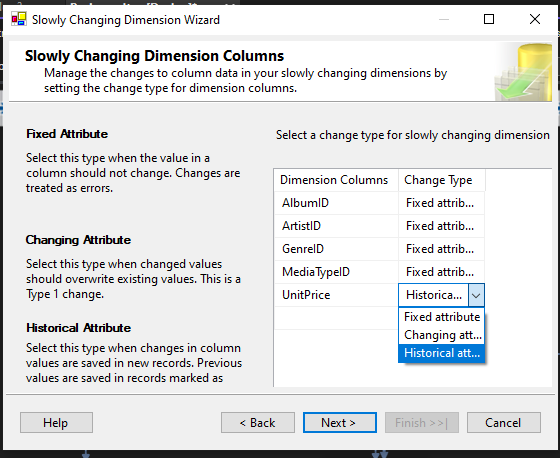


Hình 31. Cài đặt SCD cho bảng Dim\_Customer

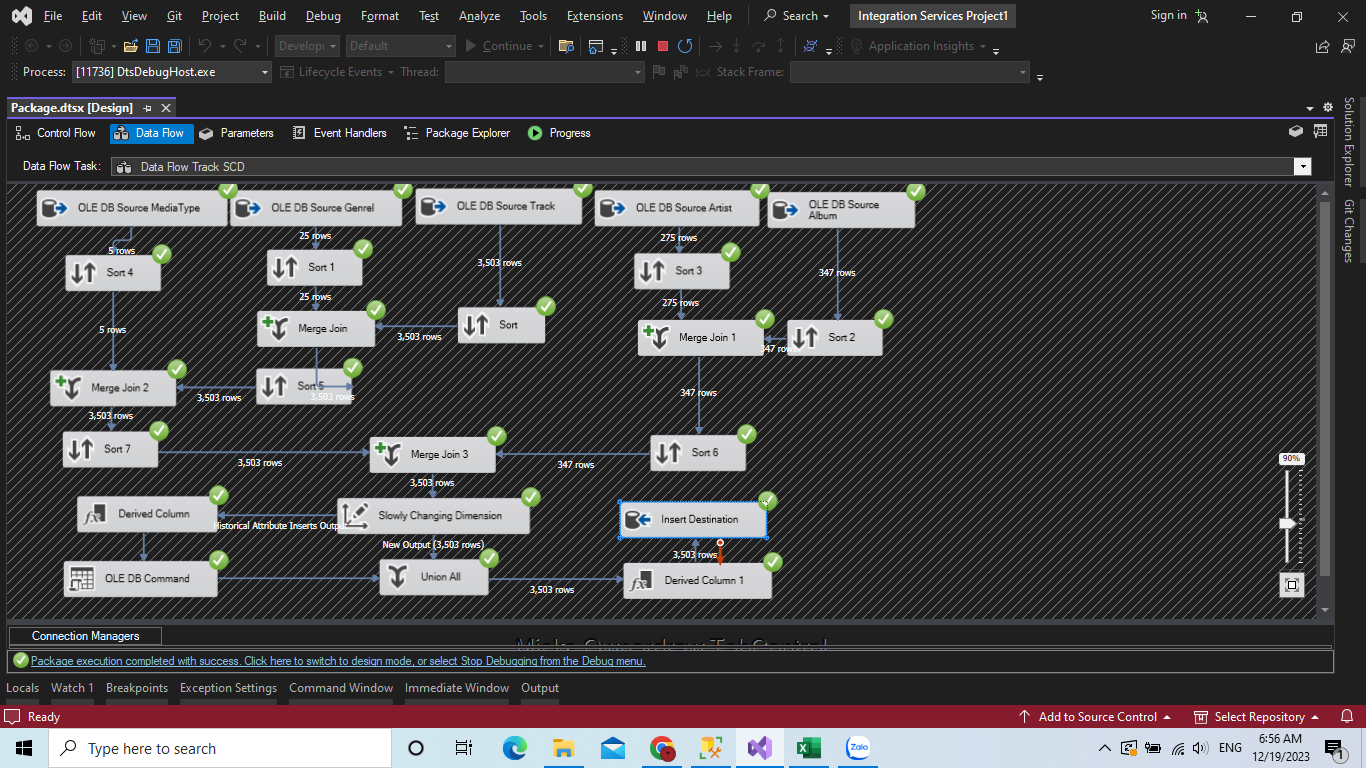


Hình 32. Kết quả SCD bảng Dim\_Customer

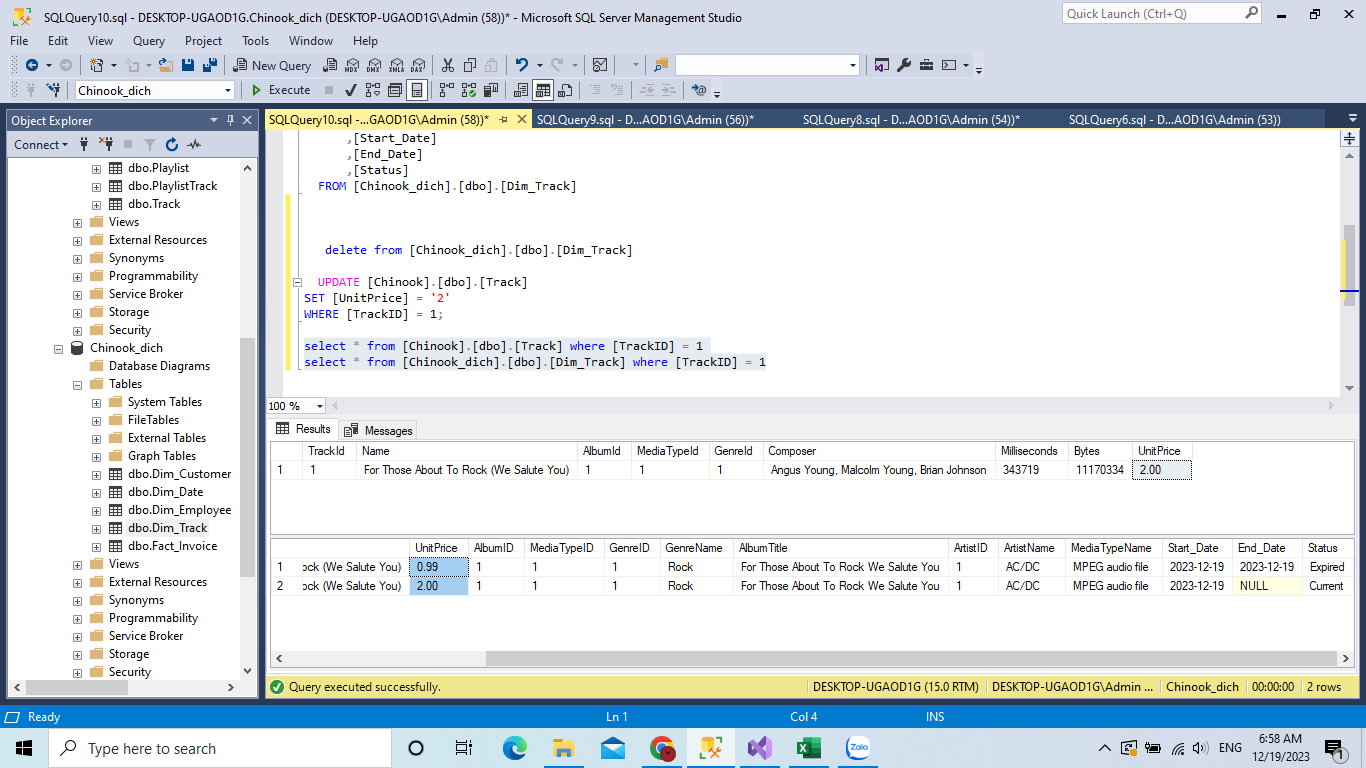
SCD cho bảng Dim\_Track: Tương tụ các bảng trên Dim\_Track cũng sử dụng CSD type 2 cho UnitPrice.



Hình 33. Cài đặt SCD bảng Dim\_Track



Hình 34. Mô hình SCD bảng Dim\_Track



Hình 35. Kết quả SCD bảng Dim\_Track

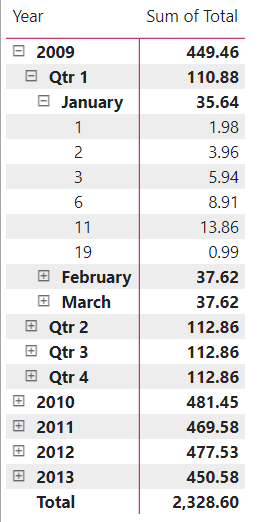
Với việc sử dụng SCD Type 2 trong quá trình tải dữ liệu từ hệ thống nguồn vào bảng Dim giúp duy trì lịch sử và theo dõi sự thay đổi của các thuộc tính trong chiều dữ liệu.

# Thiết kế các báo cáo phân tích thống kê (tương ứng với câu 4 trong bài kiểm tra 1) sử dụng bất cứ tool nào (SQL Server Business Intelligence, Tableau, Power BI, Olap Excel,…) để minh họa các thao tác roll-up; drill-down; slice; dice (2 điểm)

Báo cáo thống kê với dữ liệu từ cơ sở dữ liệu Chinook, bao gồm các bảng Dim và Fact, có tác dụng quan trọng trong việc cung cấp cái nhìn toàn diện và chính xác về hoạt động kinh doanh của hệ thống. Nhóm đã sử dụng tool Power BI để thực hiện thiết kế các báo cáo phân tích thống kê tương ứng với các câu hỏi trong câu 4. Nhóm đã thực hiện như sau:

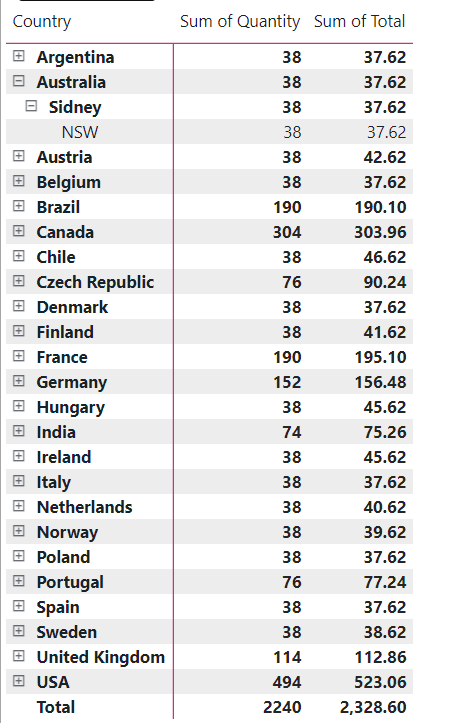
* *Câu hỏi 1: Thống kê doanh thu bán lẻ theo thời gian để phân tích sự biến động của doanh thu trong ngắn hạn (theo mùa) và trong dài hạn.*

Ta theo dõi sự biến động của doanh thu bán hàng theo thời gian, ta thấy thứ tự phân cấp độ chi tiết của Date (Year- Quarter - Month - Week - Day). Phép Toán Drill-Down nhóm đã thực hiện chi tiết hóa thông qua việc chia nhỏ cấp độ Thời gian từ Year cho đến Month, minh họa như ảnh dưới đây.



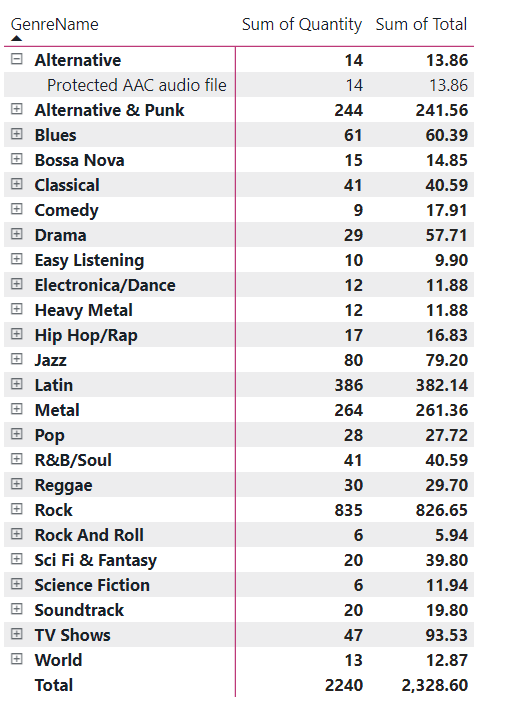
Hình 36. Báo cáo thống kê về doanh thu bán lẻ theo thời gian

* *Câu 2: Thống kê doanh số bán lẻ theo quốc gia, bang, thành phố của khách hàng để xác định được thị trường tiềm năng.*



Hình 37. Báo cáo thống kê doanh số bán lẻ theo quốc gia, bang, thành phố.

* *Câu hỏi 3: Thống kê doanh số bán lẻ theo loại sản phẩm, thể loại nhạc để phân tích xu hướng thị trường về loại sản phẩm và thể loại nhạc được yêu thích.*



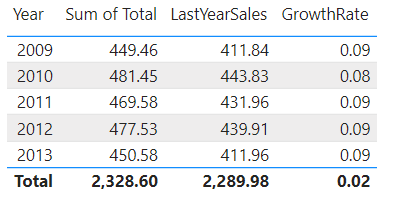
Hình 38. Báo cáo thống kê doanh số bán lẻ theo loại sản phẩm, thể loại nhạc

* *Câu hỏi 5: Thống kê khách hàng có xu hướng mua hàng thường xuyên để xác định được nhóm khách hàng tiềm năng.*



Hình 39. Báo cáo thống kê khách hàng có xu hướng mua hàng thường xuyên

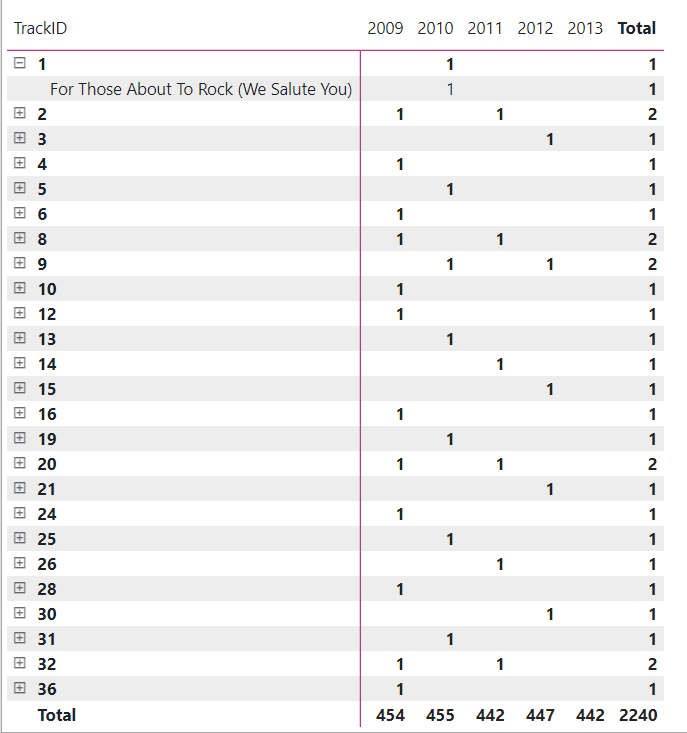
* *Câu hỏi 7: Thống kê tỷ lệ tăng trưởng doanh số bán hàng tất cả các năm để xác định xu hướng thị trường, từ đó đưa ra các chiến lược kinh doanh hiệu quả hơn.*



Hình 40. Báo cáo thống kê tỷ lệ tăng trưởng doanh số bán hàng tất cả các năm

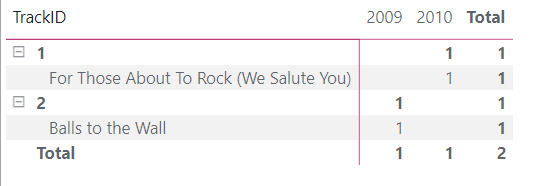
* *Câu hỏi 9: Thống kê các sản phẩm bán chạy nhất hoặc bán chậm nhất theo tháng, theo năm để đưa ra quyết định sản phẩm mà Chinook cần tiếp tục sản xuất, cần cải thiện hoặc ngừng sản xuất*

Để nhìn nhận dữ liệu sản phẩm theo doanh số của từng năm, nhóm đã sử dụng phép toán Slice cho báo cáo thống kê dưới đây:

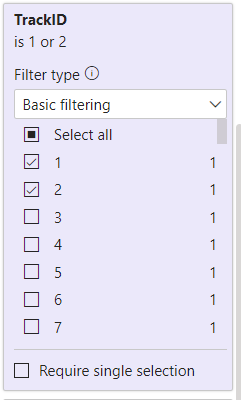


Hình 41. Báo cáo thống kê các sản phẩm bán chạy nhất hoặc bán chậm nhất theo tháng, theo năm

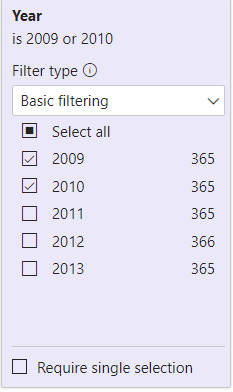
Trong câu hỏi này, phép toán Dice, nhóm sử dụng để minh họa mối quan hệ giữa các chiều dữ liệu. Sau đây là minh họa về phép toán Dice thể hiện mối quan hệ sản phẩm, theo doanh số và  năm.



Hình 42. Báo cáo thống kê về các sản phẩm bán chạy nhất hoặc bán chậm nhất theo năm



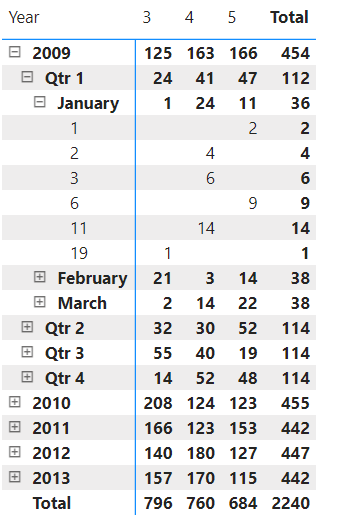
Hình 43. Bộ lọc cho TrackID



Hình 44. Bộ lọc cho Year

* *Câu hỏi 11: Thống kê doanh số bán hàng của từng nhân viên theo tháng, quý, năm để phân tích hiệu suất làm việc của từng nhân viên.*

Ta cũng có thể dùng phép toán Slice, Dice và Drill\_Down và cả Roll-up cho báo cáo thống kê này. Ta thấy độ chi tiết theo thời gian từ cấp độ Year - Quarter - Month. Cũng tương tự chúng ta áp dụng Slice và Dice cho báo cáo thống kê này như hình minh họa dưới đây



Hình 45. Báo cáo thống kê doanh số bán hàng của từng nhân viên theo tháng, quý, năm

# Sau quá trình thử nghiệm cài đặt hệ thống và khai thác dữ liệu trên Cube, nhóm hãy nêu một số đánh giá về hiệu quả của DW về: khả năng truy vấn, khai thác dữ liệu, … ? (0.5 điểm). Loại DB nào phù hợp để thiết kế DW? Giải thích? (0.5 điểm)

**\*Về khả năng truy vấn**

Mức độ phức tạp của truy vấn giữa table nguồn và đích theo dữ liệu Chinook có thể được so sánh như sau:

|  |  |  |
| --- | --- | --- |
| Loại truy vấn | Table nguồn | Table đích |
| Truy vấn đơn giản | Tương đương | Tương đương |
| Truy vấn phức tạp | Có thể phức tạp hơn | Có thể đơn giản hơn |
| Truy vấn kết hợp | Có thể phức tạp hơn | Có thể đơn giản hơn |

**Ví dụ 1:** Truy vấn bảng customers để lấy tất cả khách hàng theo quốc gia:

SQL

SELECT country, COUNT(\*) AS customer\_count

FROM customers

GROUP BY country;

Truy vấn này sử dụng hàm GROUP BY, có thể làm tăng mức độ phức tạp của truy vấn. Tuy nhiên, nếu dữ liệu trong table đích đã được phân tích theo quốc gia trước, thì truy vấn này có thể đơn giản hơn như sau:

SELECT  CustomerId, Country from Dim\_Customer

Ví dụ 2: Truy vấn sau truy vấn bảng customers và bảng orders để lấy tất cả khách hàng đã đặt hàng:

SQL

SELECT customers.\*, orders.order\_id, orders.order\_date

FROM customers

JOIN orders

ON customers.customer\_id = orders.customer\_id;

Truy vấn này sử dụng phép kết hợp JOIN, có thể làm tăng mức độ phức tạp của truy vấn. Tuy nhiên, ở table đích truy vấn có thể đơn giản như sau:

SELECT \* from Fact\_Sales

**\*Về khai thác dữ liệu**

Khai thác dữ liệu ở đích

* Ưu điểm:
  + Dễ dàng truy cập và phân tích dữ liệu
  + Tiết kiệm thời gian và chi phí
  + Đảm bảo tính nhất quán dữ liệu
* Nhược điểm:
  + Có thể mất dữ liệu gốc
  + Khó phát hiện lỗi dữ liệu

Khai thác dữ liệu ở nguồn

* Ưu điểm:
  + Đảm bảo tính toàn vẹn dữ liệu
  + Dễ dàng phát hiện lỗi dữ liệu
* Nhược điểm:
  + Khó truy cập và phân tích dữ liệu
  + Tốn thời gian và chi phí

⇒ Để khai thác dữ liệu Chinook nhằm mục đích phân tích dữ liệu hiện tại, nên lấy thông tin từ đích tốt hơn vì:

* Phân tích dữ liệu lớn: Dữ liệu Chinook lớn do đó lấy thông tin từ đích có thể giúp giảm tải cho hệ thống nguồn.
* Phân tích dữ liệu từ nhiều nguồn: Dữ liệu Chinook được lưu trữ từ nhiều nguồn, vì vậy lấy thông tin từ đích giúp khai thác dữ liệu dễ dàng kết hợp dữ liệu từ các nguồn khác nhau.

Đối với dữ liệu Chinook, có thể sử dụng các loại DB sau:

* DB quan hệ: DB quan hệ là lựa chọn tốt nhất cho dữ liệu Chinook. DB quan hệ có khả năng lưu trữ và truy vấn dữ liệu hiệu quả, cũng như hỗ trợ các truy vấn phức tạp và các công cụ khai thác dữ liệu.
* DB phân tán: DB phân tán có thể được sử dụng nếu cần tăng tốc độ truy vấn cho dữ liệu Chinook.
* DB NoSQL: DB NoSQL có thể được sử dụng nếu cần lưu trữ và truy vấn dữ liệu phi cấu trúc trong dữ liệu Chinook.

Với dữ liệu Chinook DB phù hợp nhất là BD quan hệ vì DW sẽ đáp ứng được nhiều các yêu cầu như:

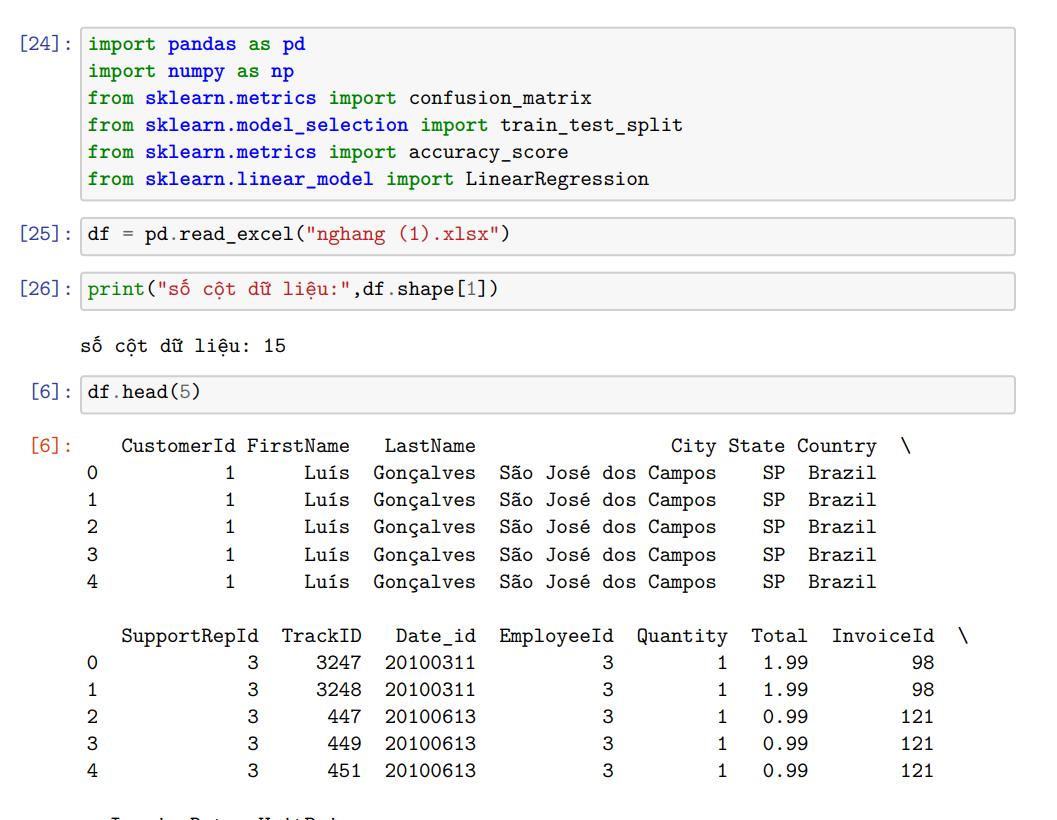
* Khả năng lưu trữ và truy vấn dữ liệu hiệu quả: DW sẽ có khả năng lưu trữ và truy vấn dữ liệu hiệu quả nhờ sử dụng DB quan hệ.
* Khả năng truy vấn phức tạp: DW sẽ có khả năng truy vấn phức tạp nhờ sử dụng DB quan hệ và các bảng dimension.
* Khả năng khai thác dữ liệu: DW sẽ có khả năng khai thác dữ liệu nhờ sử dụng DB quan hệ và các công cụ khai thác dữ liệu.

# Khi dữ liệu từ NGUỒN được tải vào Data mart (tại bảng Dim, Fact) thì nhóm em có sử dụng kỹ thuật học máy, học sâu, luật kết hợp… để xây dựng mô hình dự báo, dự đoán (trên tập dữ liệu ở data mart) hỗ trợ ra quyết định mức chiến lược, chiến thuật không? (1 điểm)

Sau khi đổ xong dữ liệu từ nguồn sang Data mart nhóm sử dụng mô hình hồi quy tuyến tính để dự báo, dự đoán từ đó đưa ra quyết định mức chiến lược, chiến thuật. Câu hỏi của nhóm đưa ra như sau: *Dự đoán doanh thu theo khu vực của khách hàng? Rồi từ đó đưa ra quyết định là có phát triển thêm cơ sở ở khu vực đó hay không.*

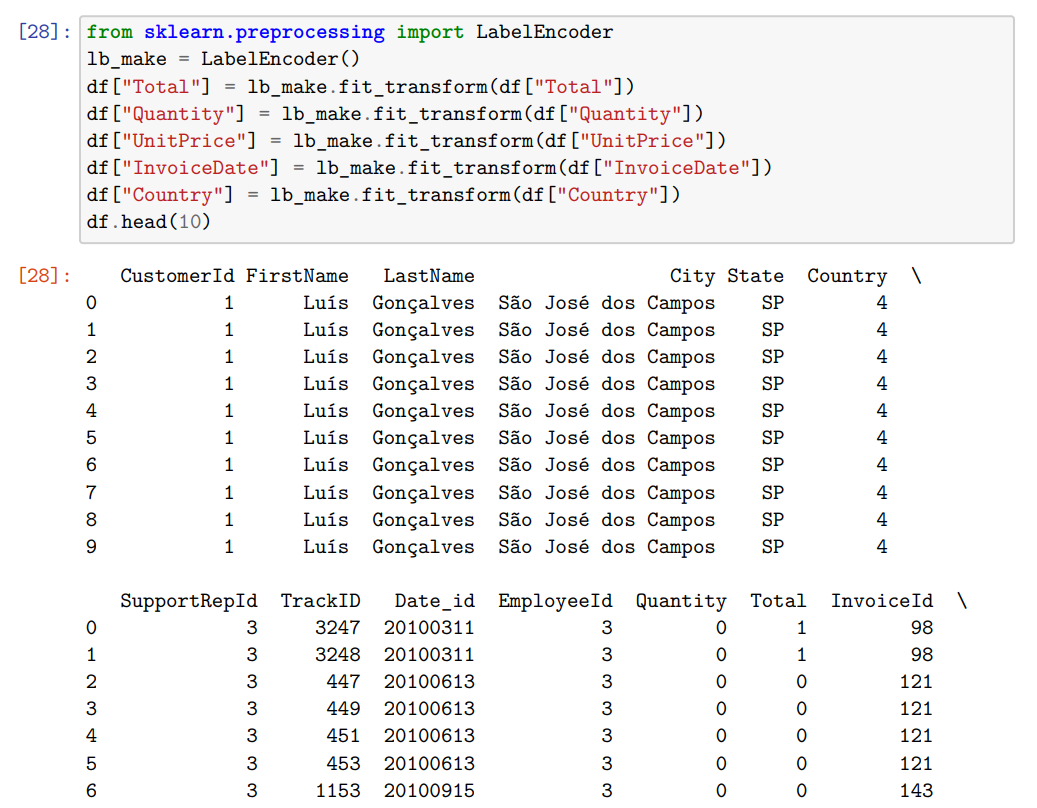
Nhóm sử dụng JupyterLap để thực hiện mô hình dự đoán.

Đầu tiên, cần khai báo mô hình và đọc dữ liệu, nhóm lấy dữ liệu từ bảng Fact của Data mart.



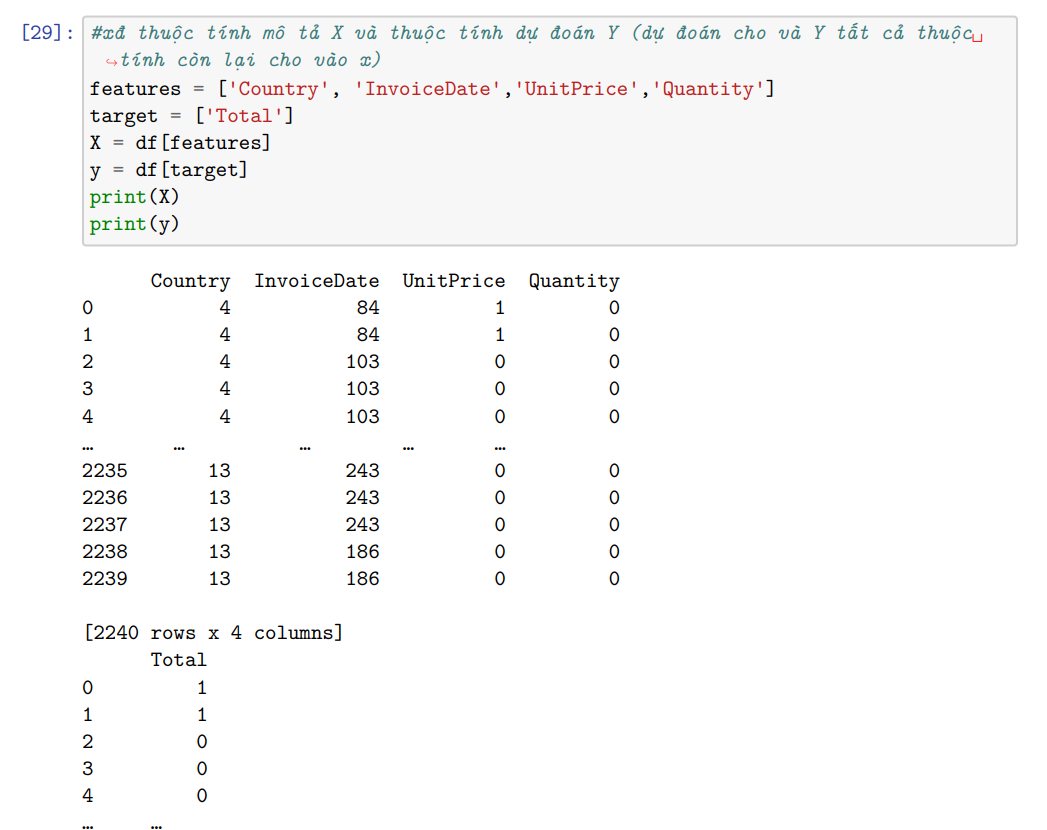
Hình 46. Khai báo mô hình và đọc dữ liệu.

Sau đó, chuyển đổi dữ liệu dạng chữ sang dạng số



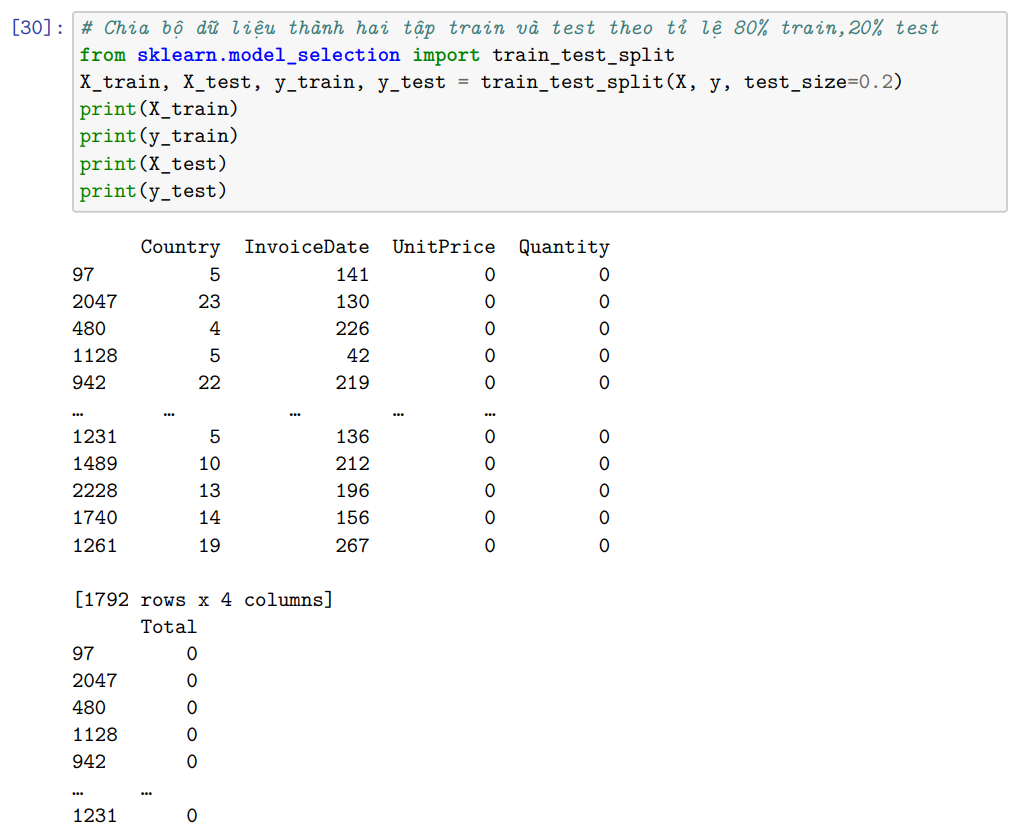
Hình 47. Chuyển đổi dữ liệu dạng chữ sang dạng số

Xác định các thuộc tính mô tả, ta có các biến độc lập X gồm Country, Invoice Date, UnitPrice, Quantity. Biến phụ thuộc Y là Total.



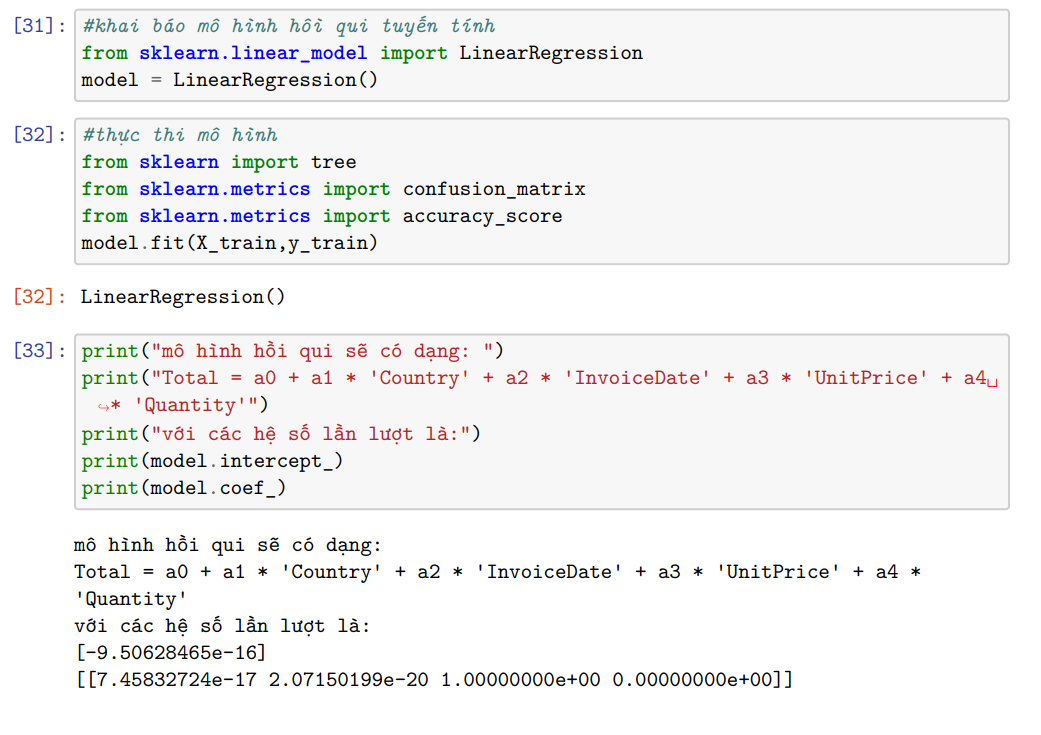
Hình 48. Xác định các thuộc tính mô tả.

Chia bộ dữ liệu thành 2 phần là test và train.



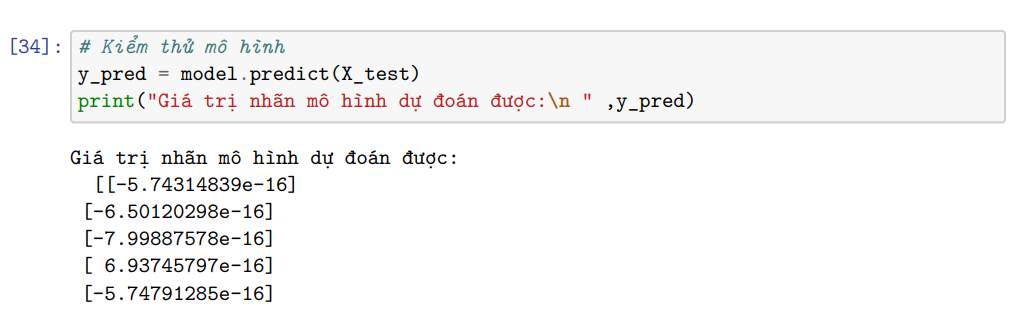
Hình 49. Chia bộ dữ liệu thành hai tập train và test

Khai báo mô hình và thực thi



Hình 50. Khai báo mô hình và thực thi.

Cuối cùng là kiểm thử mô hình



Hình 51. Kiểm thử mô hình

1. Nhóm hãy phát huy trí tưởng tượng và thiết kế kịch bản để case study mà nhóm đang nghiên cứu hướng đến việc triển khai data lake: nêu lý do/thách thức (làm thế nào gây xúc động cho người nghe là được) và trình bày giải pháp (sử dụng dịch vụ nào của nhà cung cấp nào) có minh họa kiến trúc bằng hình ảnh, lợi ích (1 điểm)

## 15.1.Những điều cần biết về Data Lake

**a. Khái niệm**

Data Lake hay hồ dữ liệu là một kho lưu trữ tập trung được thiết kế để lưu trữ, xử lý và bảo mật một lượng lớn dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc. Nó có thể lưu trữ dữ liệu ở định dạng gốc và xử lý bất kỳ loại dữ liệu nào khác nhau, bỏ qua các giới hạn về kích thước.

Data Lake cung cấp một nền tảng có thể mở rộng và an toàn cho phép các doanh nghiệp: nhập bất kỳ dữ liệu nào từ bất kỳ hệ thống nào ở bất kỳ tốc độ nào, ngay cả khi dữ liệu đến từ các hệ thống tại chỗ, đám mây hoặc điện toán biên; lưu trữ bất kỳ loại hoặc khối lượng dữ liệu nào với độ trung thực hoàn toàn; xử lý dữ liệu theo thời gian thực hoặc chế độ hàng loạt; và phân tích dữ liệu bằng SQL,Python, R hoặc bất kỳ ngôn ngữ nào khác, dữ liệu của bên thứ ba hoặc ứng dụng phân tích.

**b. Khi nào cần đến Data Lake?**

Khi xác định xem công ty của bạn có cần một hồ dữ liệu hay không, hãy ghi nhớ các loại dữ liệu bạn đang làm việc, những gì bạn muốn làm với dữ liệu, mức độ phức tạp của quy trình thu thập dữ liệu và chiến lược của bạn để quản lý và quản trị dữ liệu, cũng như các công cụ hiện có trong tổ chức của bạn.

Các công ty ngày nay cũng bắt đầu xem xét giá trị của các hồ dữ liệu thông qua một lăng kính khác - một hồ dữ liệu không chỉ để lưu trữ dữ liệu có độ trung thực, đầy đủ. Đó cũng là việc người dùng hiểu sâu hơn về các tình huống kinh doanh vì họ có nhiều ngữ cảnh hơn bao giờ hết, cho phép họ đẩy nhanh các phân tích thử nghiệm.

Được phát triển chủ yếu để xử lý khối lượng lớn dữ liệu lớn, các công ty thường có thể di chuyển dữ liệu thô vào một hồ dữ liệu mà không cần chuyển đổi nó. Các doanh nghiệp dựa vào hồ dữ liệu để giúp:

* Giảm chi phí
* Đơn giản hóa việc quản lý dữ liệu
* Chuẩn bị để kết hợp trí tuệ nhân tạo và máy học
* Tăng tốc độ phân tích
* Cải thiện bảo mật và quản trị

**c. So sánh giữa Data Warehouse và Data Lake**

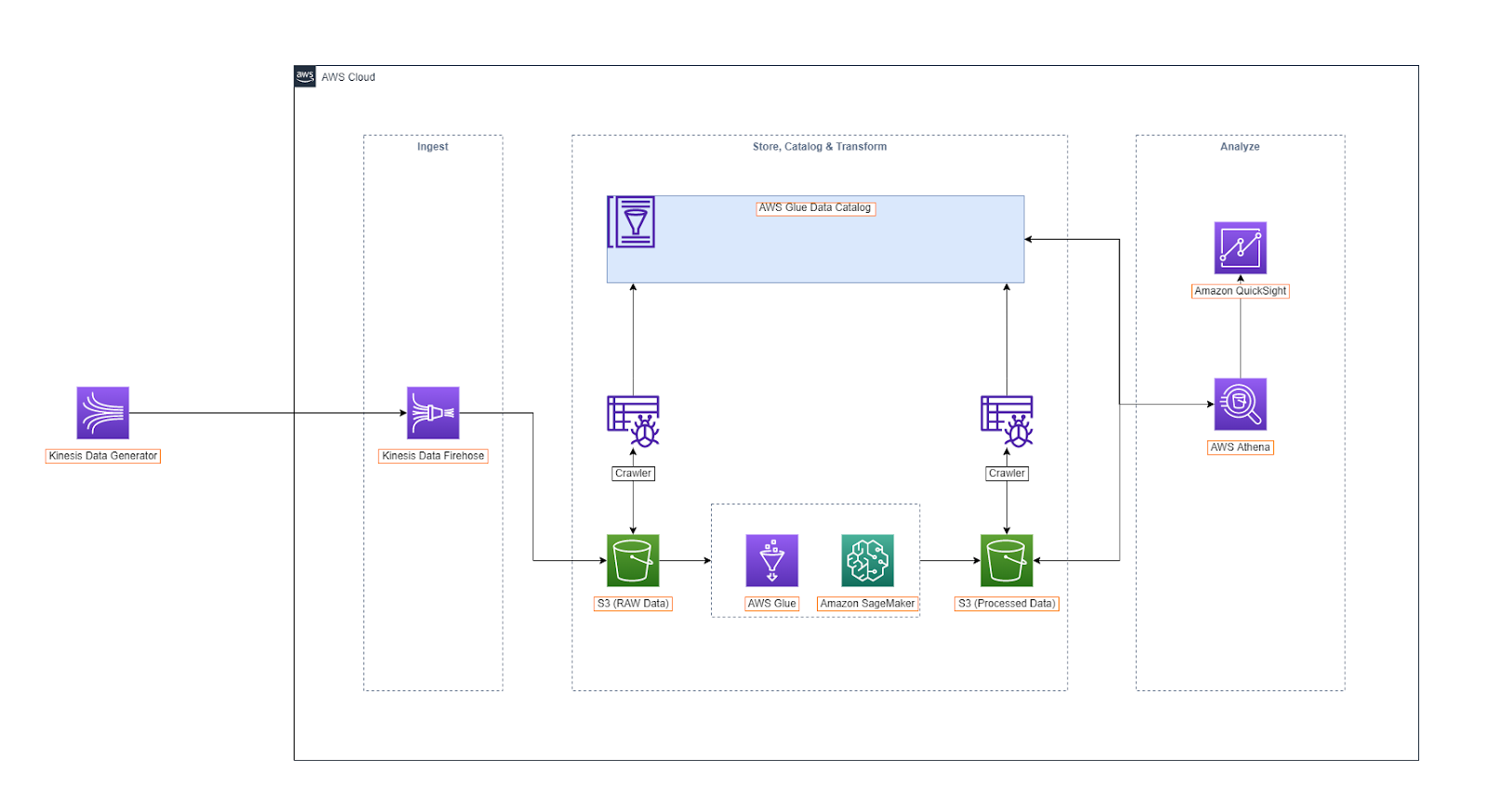
|  |  |  |
| --- | --- | --- |
| Yếu tố | Data Warehouse | Data Lake |
| Dữ liệu | Dữ liệu quan hệ từ các hệ thống giao dịch, cơ sở dữ liệu hoạt động và dòng ứng dụng kinh doanh | Tất cả dữ liệu, bao gồm cả dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc |
| Lược đồ | Thường được thiết kế trước khi triển khai kho dữ liệu nhưng cũng có thể được viết tại thời điểm phân tích   (lược đồ khi ghi hoặc lược đồ khi đọc) | Được viết tại thời điểm phân tích( lược đồ trên đọc) |
| Giá cả/Hiệu Suất | Kết quả truy vấn nhanh nhất bằng cách sử dụng bộ nhớ cục bộ | Kết quả truy vấn ngày càng nhanh hơn bằng cách sử dụng bộ nhớ chi phí thấp và phân tách tính toàn và bộ nhớ |
| Chất lượng dữ liệu | Dữ liệu được quản lý kỹ lưỡng đóng vai trò là phiên bản trung tâm của sự thật | Bất kỳ dữ liệu nào có thể được hoặc không được quản lý( tức là dữ liệu thô) |
| Người dùng | Nhà phân tích kinh doanh, nhà khoa học dữ liệu và nhà phát triển dữ liệu | Nhà phân tích kinh doanh (sử dụng dữ liệu được sắp xếp),nhà khoa học dữ liệu, nhà phát triển dữ liệu, kỹ sư dữ liệu và kiến trúc sư dữ liệu |
| Phân tích | Báo cáo hàng loạt,BI và hình ảnh hóa | Học máy, phân tích khám phá, khám phá dữ liệu, phát trực tuyến, phân tích hoạt động, dữ liệu lớn và lập hồ sơ |

## 15.2. Tình huống giả lập

Hiện nay, với cửa hàng đối diện với một lượng lớn dữ liệu khách hàng và giao dịch ngày càng tăng, việc mở rộng và tăng trưởng trở nên cấp bách. Mục tiêu hàng đầu của cửa hàng là cải thiện trải nghiệm cho khách hàng và tối ưu lợi nhuận. Để đạt được điều này, việc ra quyết định chính xác là cực kỳ quan trọng, và điều này yêu cầu có dữ liệu phù hợp và đầy đủ. Tuy nhiên, việc thu thập dữ liệu thường không dễ dàng khi nguồn dữ liệu đến từ nhiều nguồn khác nhau như website, sàn thương mại điện tử, hệ thống quản lý doanh nghiệp (ERP), hệ thống quản lý mối quan hệ khách hàng (CRM), cũng như từ các cơ sở dữ liệu của các bộ phận khác nhau trong công ty. Hơn nữa, hệ thống cơ sở dữ liệu hiện tại có thể không đủ để lưu trữ toàn bộ thông tin từ những nguồn này, dẫn đến việc thiếu thông tin toàn diện về hành vi của khách hàng. Điều này gây khó khăn trong việc tối ưu hóa tài nguyên có sẵn và điều chỉnh dịch vụ theo nhu cầu của khách hàng. Bên cạnh đó, bảo mật dữ liệu cũng trở thành một vấn đề nan giải. Công ty cũng đang đối mặt với nhiều thách thức cần phải giải quyết, bao gồm...

* Từ dữ liệu trên hệ thống CRM – CareSoft phân tích các cuộc hội thoại của khách hàng để thu thập đánh giá, nhu cầu khách hàng từ đó đưa ra dịch vụ tốt hơn.
* Thu thập liên tục đánh giá trên các sàn TMĐT về từng loại sản phẩm, chất lượng dịch vụ.
* Cập nhật thông tin thời tiết, các chương trình khuyến mãi, ngày lễ, sự kiện đặc biệt để đưa ra ảnh hưởng tới doanh thu, từ đó tính ra target cho  cửa hàng.
* Thống kê lượt truy cập vào các fanpage, trang web của cửa hàng để đánh giá hiệu quả áp dụng các chiến dịch marketing, thu thập hành vi người dùng để đưa ra các mô hình dự đoán cross-selling, ...

**Để giải quyết những vấn đề trên, công ty cần nâng cấp các hệ thống hiện tại lên Cloud và ứng dụng giải pháp Data Lake trên AWS** để phục vụ cho quá trình nghiên cứu triển khai dịch chuyển hệ thống CNTT lên mây với AWS, sử dụng nhiều dịch vụ như: lưu trữ, tính toán, quản lý dữ liệu, quản lý chi phí tối ưu, bảo mật, định danh & tuân thủ.



Hình 52. AWS Cloud

**Lợi ích doanh nghiệp sẽ nhận được nếu nâng cấp hệ thống hiện tại lên Cloud và ứng dụng Data Lake.** Khi doanh nghiệp triển khai sử dụng Cloud/ Data Lake thì họ sẽ nhận được khá nhiều lợi ích từ việc này như: Khả năng sử dụng và khả năng tiếp cận sẽ dễ dàng hơn vì cơ bản các dịch vụ này đều tương tác thông qua giao diện vì vậy không đòi hỏi quá nhiều về kiến thức công nghệ. Ngoài ra khi quản lý trên đây tính bảo mật sẽ được siết chặt vì cơ bản trên đây hầu hết sẽ lưu trữ các dữ liệu dự phòng vì vậy ngay cả khi một trong các trung tâm dữ liệu bị sập, hỏng hóc, mất mát thì dữ liệu của công ty sẽ được an toàn và được giám sát và một dịch vụ lưu trữ Cloud nào cũng được hình thành từ hàng nghìn trung tâm dữ liệu. Ngoài ra, việc này cũng giúp chúng ta tiết kiệm chi phí, chia sẻ tệp thuận tiện hay tự động hóa một số quy trình riêng. Doanh nghiệp trước đây thường xuyên phải đối mặt với thời gian ngừng hoạt động với trung tâm dữ liệu của mình và hệ thống thường gặp sự cố trong thời gian tải cao điểm khi tài nguyên máy tính không thể mở rộng. Sau khi di chuyển, cửa hàng đã triển khai AWS Auto Scaling, tự động điều chỉnh dung lượng để duy trì hiệu suất hệ thống nhất quán.

# DANH MỤC THAM KHẢO

[1] Alejandro Vaisman, E. Z. (2014). *Data Warehouse Systems Design and Implementation.*

[2] Rainardi, V. (2008). *Building a Data Warehouse With Examples in SQL Server.*

[3] Ralph Kimball, M. R. (2007). *The Data Warehouse Toolkit Second Edition.*

[4] Ralph Kimball, M. R. (2013). *The Data Warehouse Toolkit Third Edition.*

[5] <https://www.osam.io/post/data-lake-ho-du-lieu-la-gi>

[6] <https://www.guru99.com/data-lake-architecture.html>