TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

KHOA CÔNG NGHỆ THÔNG TIN

**---🙠**🕮**🙢---**

Diagram

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**KHO DỮ LIỆU**

**CHỦ ĐỀ: NGHIÊN CỨU TỐI ƯU HÓA KHO DỮ LIỆU VÀ ỨNG DỤNG KHO DỮ LIỆU QUẢN LÝ Ô TÔ**

**Giảng viên hướng dẫn: Ts.Đào Thu Vân**

**Sinh viên thực hiện: 1. Phạm Cao Cường – 2121050044**

**2. Phan Văn Khải – 2121050113**

3. Lê Viết Tâm - 2121050051

**4. Lê Việt Dũng - 2021050140**

**HỌC PHẦN: 7080214 - NHÓM 200**

**HÀ NỘI-2023**

**MỤC LỤC**

[DANH MỤC HÌNH ẢNH 3](#_Toc155332687)

[LỜI MỞ ĐẦU 4](#_Toc155332688)

[CHƯƠNG 1: LÝ DO CHỌN ĐỀ TÀI 5](#_Toc155332689)

[1.1 Tổng quan 5](#_Toc155332690)

[1.2 Mục đích của đề tài 5](#_Toc155332691)

[1.3 Ý nghĩa thực tiễn 6](#_Toc155332692)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 7](#_Toc155332693)

[2.1 Giới thiệu về kho dữ liệu 7](#_Toc155332694)

[2.2 Công nghệ kho dữ liệu 8](#_Toc155332695)

[2.2.1 Đặc tính của kho dữ liệu 8](#_Toc155332696)

[2.2.2 Mô hình dữ liệu đa chiều 9](#_Toc155332697)

[2.2.3 Kiến trúc của kho dữ liệu 11](#_Toc155332698)

[2.3 Các công cụ hỗ trợ 13](#_Toc155332699)

[2.3.1 Visual studio code 13](#_Toc155332700)

[2.3.2 SQL Server Management Studio 13](#_Toc155332701)

[2.3.3 Ngôn ngữ lập trình python 13](#_Toc155332702)

[CHƯƠNG 3: TỐI ƯU HOÁ KHO DỮ LIỆU 15](#_Toc155332703)

[3.1 Tại sao cần phải tối ưu hóa kho dữ liệu 15](#_Toc155332704)

[3.2 Các cách tối ưu hóa kho dữ liệu 16](#_Toc155332705)

[CHƯƠNG 4: ỨNG DỤNG KHO DỮ LIỆU VÀO BÀI TOÁN QUẢN LÝ Ô TÔ 19](#_Toc155332706)

[4.1 Phân tích và thiết kế kho dữ liệu 19](#_Toc155332707)

[4.1.1 Thực thể đo lường và số nguyên 19](#_Toc155332708)

[4.1.2 Thứ bậc của kích thước 20](#_Toc155332709)

[4.1.3 Sơ đồ mô hình khái niệm 21](#_Toc155332710)

[4.2 Mô hình logic 23](#_Toc155332711)

[4.2.1.Fact và dimension table (PK,FK) 23](#_Toc155332712)

[4.2.2 Sơ đồ bông tuyết và sơ đồ lai 25](#_Toc155332713)

[4.3 Triển khai kho dữ liệu 30](#_Toc155332714)

[4.3.1 Mô hình SSIS 30](#_Toc155332715)

[4.3.2 Quá trình ETH 31](#_Toc155332716)

[4.4 Đánh giá kết quả 32](#_Toc155332717)

[4.4.1 Kết quả đạt được 32](#_Toc155332718)

[4.4.2 Mặt hạn chế 33](#_Toc155332719)

[4.4.3 Hướng phát triển 33](#_Toc155332720)

[KẾT LUẬN 36](#_Toc155332721)

[TÀI LIỆU THAM KHẢO 37](#_Toc155332722)

DANH MỤC HÌNH ẢNH

[Hình 1.Kiến trúc kho dữ liệu 9](#_Toc155331836)

[Hình 2.Mạng lưới các cuboid 11](#_Toc155331837)

[Hình 3.Lược đồ hình sao 12](#_Toc155331838)

[Hình 4.Lược đồ bông tuyết 12](#_Toc155331839)

[Hình 5.Kiến trúc kho dữ liệu 2 lớp 13](#_Toc155331840)

[Hình 6.Datamart độc lập 13](#_Toc155331841)

[Hình 7.Datamart phụ thuộc 13](#_Toc155331842)

[Hình 8.Mô hình kho dữ liệu 3 lớp 13](#_Toc155331843)

[Hình 9.Sơ đồ bông tuyết quản li ô tô 26](#_Toc155331844)

[Hình 10.Bảng customer 27](#_Toc155331845)

[Hình 11.Bảng Date 27](#_Toc155331846)

[Hình 12.Bảng Payment Method 27](#_Toc155331847)

[Hình 13.Bảng product 28](#_Toc155331848)

[Hình 14.Bảng promotion 28](#_Toc155331849)

[Hình 15.Bảng saleperson 28](#_Toc155331850)

[Hình 16.Bảng store 29](#_Toc155331851)

[Hình 17.Bảng Factsaleorder 29](#_Toc155331852)

[Hình 18.Mô Hình SSIS 31](#_Toc155331853)

[Hình 19.Bảng data product mới 31](#_Toc155331854)

[Hình 20.Import dữ liệu vào excel 32](#_Toc155331855)

LỜI MỞ ĐẦU

Tất cả các lĩnh vực, từ doanh nghiệp đến y tế, khoa học, và công nghệ đều chứng kiến sự gia tăng không ngừng của dữ liệu. Điều này tạo ra một thách thức lớn: làm thế nào để con người có thể hiệu quả hóa việc tổ chức, lưu trữ và sử dụng thông tin một cách thông minh nhất để đáp ứng nhu cầu ngày càng tăng về thông tin và kiến thức.

Nghiên cứu về tối ưu hóa kho dữ liệu đặt ra mục tiêu tối đa hóa hiệu suất của các hệ thống lưu trữ và xử lý dữ liệu. Từ việc thiết kế cơ sở dữ liệu hiệu quả đến việc triển khai các thuật toán và công nghệ mới để tối ưu hóa truy xuất thông tin, mọi nỗ lực đều hướng tới mục tiêu chung tạo ra môi trường lưu trữ dữ liệu linh hoạt, an toàn và có thể mở rộng.

Ứng dụng kho dữ liệu không chỉ dừng lại ở việc lưu trữ thông tin mà còn mở ra cánh cửa cho việc phân tích, dự đoán và học máy. Từ việc tối ưu hóa quy trình sản xuất trong công nghiệp đến việc cải thiện chẩn đoán bệnh trong y học, sức mạnh của dữ liệu khi được tận dụng đúng cách là vô hạn.

Báo cáo này sẽ khám phá sâu hơn vào các phương pháp tối ưu hóa kho dữ liệu, từ các kỹ thuật lưu trữ đến các thuật toán phân tích thông tin, cũng như điều này làm thay đổi cách chúng ta nhìn nhận và tận dụng dữ liệu trong các lĩnh vực khác nhau. Bằng việc xem xét cả mặt lý thuyết và thực tiễn, chúng ta sẽ thấy được tầm quan trọng và tiềm năng to lớn mà tối ưu hóa kho dữ liệu mang lại.

# CHƯƠNG 1: LÝ DO CHỌN ĐỀ TÀI

* 1. **Tổng quan**

Việc nghiên cứu tối ưu hóa kho dữ liệu và áp dụng nó vào quản lý ô tô là một đề tài đầy hứa hẹn vì nhiều lý do. Đầu tiên, ngành công nghiệp ô tô đang phát triển với sự gia tăng đáng kể về dữ liệu từ các xe thông minh, cảm biến và hệ thống GPS. Việc quản lý và phân tích lượng dữ liệu lớn này đòi hỏi một hệ thống kho dữ liệu hiệu quả để giải quyết các thách thức trong việc lưu trữ, truy xuất và phân tích thông tin.

Việc tối ưu hóa kho dữ liệu có thể mang lại hiệu suất tốt hơn cho việc quản lý ô tô. Bằng cách sử dụng các phương pháp và công nghệ tối ưu hóa, chúng ta có thể giảm thiểu thời gian tìm kiếm thông tin, tăng tốc quy trình ra quyết định, và cải thiện khả năng dự đoán về bảo trì và sửa chữa xe hơi.

Việc áp dụng kho dữ liệu trong quản lý ô tô cũng có thể tối ưu hóa việc quản lý kho hàng, lên lịch bảo trì định kỳ, cải thiện trải nghiệm khách hàng qua dịch vụ chăm sóc sau bán hàng và thậm chí tối ưu hóa hoạt động của các dịch vụ gọi xe thông qua phân tích dữ liệu vận chuyển và lộ trình.

* 1. **Mục đích của đề tài**

Tối ưu hóa Quy Trình Quản Lý Dữ Liệu: Nghiên cứu nhằm tạo ra các phương pháp, công cụ và quy trình tối ưu hóa để thu thập, tổ chức, lưu trữ và truy xuất dữ liệu từ các xe ô tô thông minh, cảm biến và hệ thống GPS. Mục tiêu là cải thiện hiệu suất và độ chính xác trong quản lý dữ liệu, từ đó hỗ trợ quyết định và phân tích thông tin liên quan đến ô tô.

Tăng Cường Hiệu Suất Quản Lý Ô Tô: Đề xuất và áp dụng các giải pháp kho dữ liệu hiệu quả để cải thiện quy trình quản lý ô tô. Điều này bao gồm tối ưu hóa việc quản lý kho hàng, dự báo bảo trì và sửa chữa, cải thiện trải nghiệm khách hàng và tối ưu hoá vận hành các dịch vụ liên quan đến xe hơi.

Phát Triển Công Nghệ Thông Minh Trong Ngành Ô Tô: Nghiên cứu nhằm thúc đẩy sự phát triển của các công nghệ thông minh trong lĩnh vực ô tô, bao gồm việc tự động hóa quy trình bảo trì, cải thiện tính hiệu quả của dịch vụ liên quan đến xe hơi và tối ưu hoá các quy trình vận hành.

Nâng Cao Trải Nghiệm Người Dùng: Tối ưu hóa dữ liệu có thể giúp cải thiện trải nghiệm người dùng thông qua việc cung cấp thông tin chính xác và kịp thời, từ quản lý dịch vụ đến thông tin vận hành và bảo trì.

Mục tiêu của đề tài này là tạo ra các giải pháp thông minh dựa trên kho dữ liệu để cải thiện hiệu suất và trải nghiệm trong quản lý ô tô, từ quy trình bảo trì đến cải thiện dịch vụ và quản lý thông tin liên quan đến xe hơi.

* 1. **Ý nghĩa thực tiễn**

Cải Thiện Quy Trình Bảo Trì và Sửa Chữa: Bằng cách sử dụng dữ liệu được quản lý tốt, nhà sản xuất và các cơ sở bảo dưỡng có thể dự đoán được khi nào xe cần bảo trì hoặc sửa chữa. Điều này giúp người dùng tiết kiệm thời gian và tiền bạc, đồng thời cải thiện an toàn khi xe luôn được bảo trì đúng thời điểm.

Tối Ưu Hóa Quản Lý Kho Hàng: Việc quản lý dữ liệu hiệu quả có thể giúp cải thiện việc lưu trữ và điều phối phụ tùng, linh kiện, giảm thiểu lãng phí và chi phí lưu kho không cần thiết.

Nâng Cao Trải Nghiệm Người Dùng: Người dùng có thể nhận được dịch vụ tốt hơn thông qua việc cung cấp thông tin chính xác và kịp thời về vận hành xe, lịch sử bảo trì và thông tin hỗ trợ.

Tạo Ra Cơ Hội Kinh Doanh Mới: Dữ liệu quản lý tốt cung cấp cơ hội cho các doanh nghiệp tạo ra các dịch vụ mới, như dịch vụ gọi xe thông minh, bảo hiểm dựa trên dữ liệu vận hành, hoặc các dịch vụ khác liên quan đến ô tô.

Tiết Kiệm Năng Lượng và Tài Nguyên: Quản lý thông tin vận hành ô tô có thể giúp tối ưu hóa lộ trình, tiết kiệm nhiên liệu và giảm thiểu ô nhiễm môi trường.

Phát Triển Công Nghệ Ô Tô Thông Minh: Nghiên cứu này có thể đóng vai trò quan trọng trong việc phát triển xe ô tô tự lái, ô tô kết nối mạng và các công nghệ thông minh khác.

Việc tối ưu hóa và ứng dụng kho dữ liệu trong quản lý ô tô không chỉ mang lại lợi ích lớn cho các doanh nghiệp trong ngành mà còn cải thiện trải nghiệm của người dùng và có thể góp phần vào việc tiết kiệm tài nguyên và bảo vệ môi trường.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

**2.1 Giới thiệu về kho dữ liệu**

Môn học Kho dữ liệu (Data Warehousing) là một môn học thuộc lĩnh vực Khoa học dữ liệu (Data Science) và Hệ thống thông tin (Information Systems). Môn học này tập trung vào việc xây dựng và quản lý các kho dữ liệu, được sử dụng để lưu trữ các dữ liệu có cấu trúc và không có cấu trúc từ nhiều nguồn khác nhau.

Kho dữ liệu là nơi tập trung tất cả những dữ liệu đã được chuẩn hóa của doanh nghiệp bao gồm cả dữ liệu của quá khứ nhằm mục đích sử dụng lâu dài nên kiến trúc của kho dữ liệu tương đối phức tạp.

Kho dữ liệu là một hệ thống thông tin lưu trữ lịch sử và giao hoán dữ liệu từ một hoặc nhiều nguồn. Kho dữ liệu được thiết kế để phân tích, báo cáo, tích hợp dữ liệu giao dịch từ các nguồn khác nhau.Kho dữ liệu cũng là một phiên bản duy nhất của tổ chức cho quá trình ra quyết định và dự báo.

A green cylinder with arrows

Description automatically generated

Hình 1.Kiến trúc kho dữ liệu

Kho dữ liệu giúp người dùng truy cập dữ liệu từ tất cả các nguồn ở một nơi => tiết kiệm thời gian Kho dữ liệu giúp người dùng tích hợp nhiều nguồn dữ liệu => giảm TAT (total turnaround time: tổng thời gian quay vòng) để phân tích và báo cáo. Kho dữ liệu cho phép lưu trữ một lượng lớn dữ liệu lịch sử => phân tích, dự đoán.

Kho dữ liệu được hướng theo chủ đề vì nó cung cấp thông tin liên quan đến chủ đề thay vì hoạt động liên tục của các công ty. Dữ liệu cũng cần được lưu trữ trong kho dữ liệu theo cách thông thường và được nhất trí chấp nhận. Thời gian cuối cho kho dữ liệu tương đối rộng. Kho dữ liệu không biến động là dữ liệu trước đó không bị xóa khi dữ liệu mới được nhập vào.

**(\*)** Kho dữ liệu được sử dụng cho các lĩnh vực:

* Ngân hàng: Sử dụng trong lĩnh vực ngân hàng cho thông tin khách hàng, các hoạt động liên quan đến tài khoản, thanh toán, tiền gửi, cho vay, thẻ tín dụng, v.v.
* Hàng không: Sử dụng để đặt phòng và thông tin lịch trình.
* Các trường đại học: Để lưu trữ thông tin sinh viên, đăng ký khóa học, cao đẳng, và kết quả.
* Viễn thông: Giúp lưu trữ hồ sơ cuộc gọi, hóa đơn hàng tháng v.v.
* Tài chính: Giúp lưu trữ thông tin liên quan đến cổ phiếu, bán hàng và mua cổ phiếu và trái phiếu.
* Bán hàng & Sản xuất: Sử dụng để lưu trữ chi tiết khách hàng, sản phẩm và bán hàng.
* Chế tạo: Sử dụng để quản lý dữ liệu của chuỗi cung ứng và để theo dõi sản xuất các mặt hàng, tình trạng hàng tồn kho.
* Quản lý nhân sự: Chi tiết về tiền lương của nhân viên, khấu trừ, tạo tiền lương, v.v.

**(\*)** Một vài nhược điểm của kho dữ liệu:

* Thêm các nguồn dữ liệu mới cần có thời gian và nó có liên quan đến chi phí cao.
* Đôi khi các vấn đề liên quan đến kho dữ liệu có thể không bị phát hiện trong nhiều năm.
* Kho dữ liệu là hệ thống bảo trì cao.
* Trích xuất, tải và làm sạch dữ liệu có thể tốn thời gian.
* Kho dữ liệu có thể trông đơn giản, nhưng thực sự, nó quá phức tạp đối với người dùng trung bình.
* Bạn cần đào tạo cho người dùng cuối, những người cuối cùng không sử dụng khai thác dữ liệu và kho.
* Mặc dù có những nỗ lực tốt nhất trong quản lý dự án, phạm vi lưu trữ dữ liệu sẽ luôn tăng.

**2.2 Công nghệ kho dữ liệu**

***2.2.1 Đặc tính của kho dữ liệu***

Dữ liệu hướng chủ thể:

* Dữ liệu hướng theo từng nhóm đối tượng: khách hang, bệnh nhân, sản phẩm, …
* Tập trung vào việc mô hình hóa và phân tích các dữ liệu cho các nhà sản xuất quyết định.
* Chuyển từ hướng ứng dụng sang hướng hỗ trợ quyết định.
* Không dùng cho các hoạt động hang ngày hoặc xử lý giao dịch.

Tính tích hợp:

* Dữ liệu được tập hợp từ nhiều nguồn: có thể khác kiểu, khác cấu trúc,…
* Các nguồn: cơ sở dữ liệu quan hệ, tập tin có cấu trúc, tập tin phẳng,…
* Cần được chuẩn hóa để đảm bảo tính nhất quán trong quy ước đặt tên,…
* Việc chuẩn hóa cần thực hiện trước khi tích hợp.

Dữ liệu biến thời gian:

* Thông tin về quá khứ, hiện tại.
* So sánh dữ liệu theo chiều thời gian.
* Hỗ trợ quyết định cho tương lai.
* Thành phần thời gian có thể tường minh hoặc ngầm định.

Dữ liệu mang tính bền vững, chỉ đọc (non volatile):

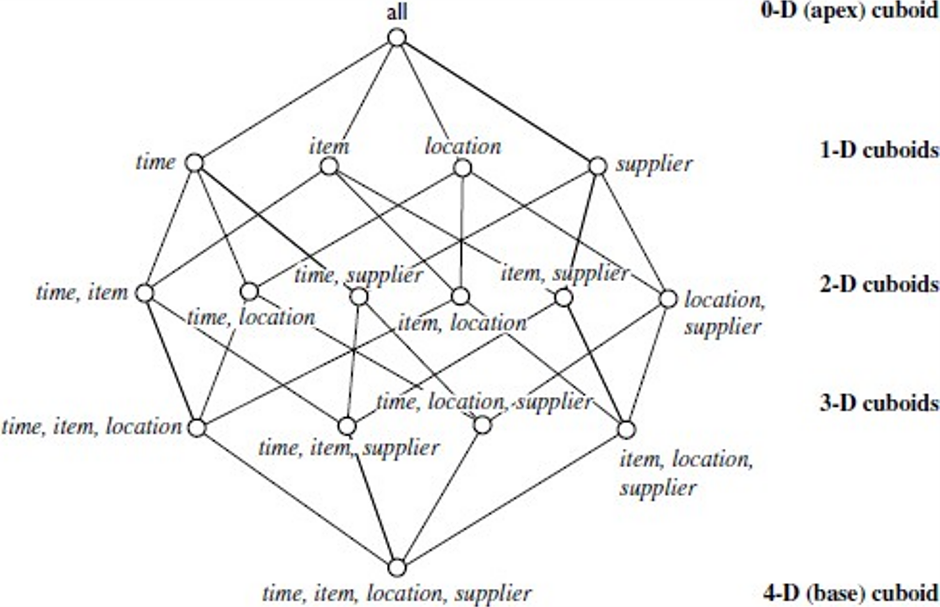
* Có thể thêm vào, nhưng không thay thế.
* Phục vụ việc nghiên cứu, phân tích

***2.2.2 Mô hình dữ liệu đa chiều***

Kho dữ liệu dựa trên mô hình dữ liệu đa chiều cho phép nhìn dữ liệu dưới hình thức của một khối dữ liệu

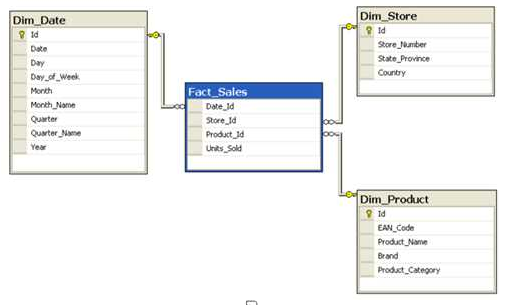
Một khối dữ liệu cho phép dữ liệu được mô hình và được nhìn trong nhiều chiều bởi: Các bản chiều (Dimension Tables) như Item (item\_name, brand, type); time (day, week, month).

Một khối dữ liệu dựa trên n-D (n chiều) được gọi là một cuboid cơ sở.Cube: một lưới các cuboid:



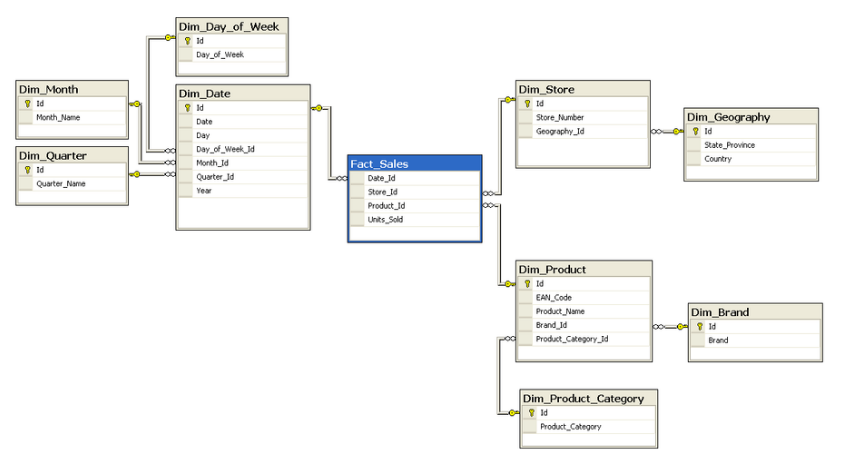
Hình 2.Mạng lưới các cuboid

* Lược đồ hình sao



Hình 3.Lược đồ hình sao

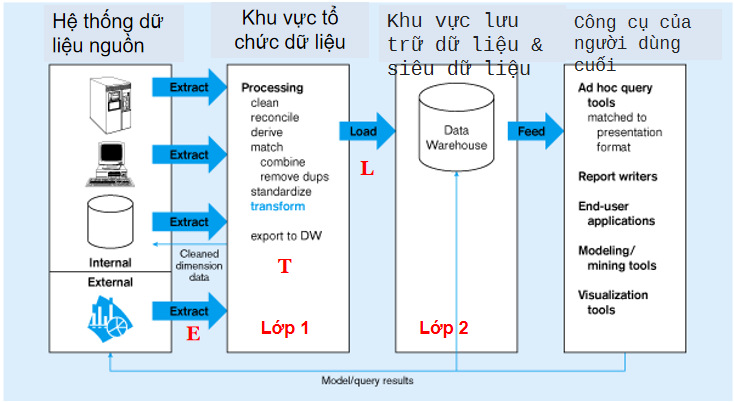
* Lược đồ hình bông tuyết



Hình 4.Lược đồ bông tuyết

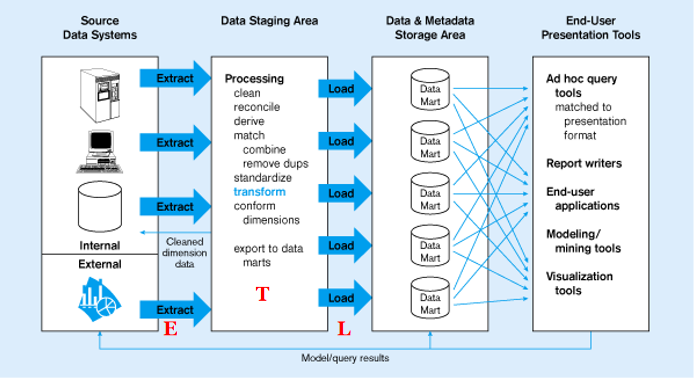
***2.2.3 Kiến trúc của kho dữ liệu***

* Kiến trúc 2 lớp khái quát:



Hình 5.Kiến trúc kho dữ liệu 2 lớp

* Data Mart độc lập:



Hình 6.Datamart độc lập

* Data Mart phụ thuộc và kho lưu trữ dữ liệu tác nghiệp:

A diagram of data storage

Description automatically generated

Hình 7.Datamart phụ thuộc

* Kiến trúc dữ liệu 3 lớp:

A diagram of a warehouse

Description automatically generated

Hình 8.Mô hình kho dữ liệu 3 lớp

**2.3 Các công cụ hỗ trợ**

***2.3.1 Visual studio code***

Visual Studio Code (VS Code) là một trình soạn thảo mã nguồn miễn phí và mã nguồn mở được phát triển bởi Microsoft. Nó được phát hành lần đầu tiên vào năm 2015 và nhanh chóng trở nên phổ biến trong cộng đồng lập trình viên.

VS Code được thiết kế để trở thành một trình soạn thảo mã nguồn toàn diện và linh hoạt, hỗ trợ nhiều ngôn ngữ lập trình và công nghệ khác nhau.

Ngoài ra, VS Code còn có thể được mở rộng bằng các phần mở rộng (extensions), cung cấp thêm các tính năng và chức năng mới. Có hàng nghìn phần mở rộng khác nhau có sẵn cho VS Code, bao gồm các phần mở rộng cho các ngôn ngữ lập trình cụ thể, các công cụ phát triển web, các công cụ kiểm tra lỗi, v.v.

VS Code được đánh giá là một trong những trình soạn thảo mã nguồn tốt nhất hiện nay. Nó được sử dụng bởi các lập trình viên ở mọi cấp độ kỹ năng, từ người mới bắt đầu đến các chuyên gia.

***2.3.2 SQL Server Management Studio***

SQL Server Management Studio (SSMS) là một môi trường tích hợp để quản lý mọi cơ sở hạ tầng SQL. Sử dụng SSMS để truy cập, đặt cấu hình, quản lý, quản trị và phát triển tất cả các thành phần của SQL Server, Cơ sở dữ liệu Azure SQL, Phiên bản quản lý Azure SQL, SQL Server trên Azure VM và Azure SynapseAnalytics. SSMS cung cấp một tiện ích toàn diện duy nhất kết hợp một nhóm công cụ đồ họa rộng rãi với nhiều trình soạn thảo tập lệnh phong phú để cung cấp quyền truy cập vào SQLServer cho các nhà phát triển và quản trị viên cơ sở dữ liệu ở mọi cấp độ kỹ năng.

***2.3.3 Ngôn ngữ lập trình python***

Python là một ngôn ngữ lập trình bậc cao, đa năng và dễ học. Nó được tạo ra vào năm 1991 bởi Guido van Rossum và đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất trên thế giới. Python là một ngôn ngữ mã nguồn mở, có nghĩa là bất kỳ ai cũng có thể tải xuống và sử dụng nó miễn phí.

Python có cú pháp đơn giản và rõ ràng, giúp cho việc viết mã trở nên dễ dàng hơn so với các ngôn ngữ khác. Nó cũng có một cộng đồng lớn và năng động, cung cấp nhiều thư viện và framework cho các ứng dụng khác nhau.

Python được sử dụng rộng rãi trong nhiều lĩnh vực, bao gồm khoa học dữ liệu, trí tuệ nhân tạo, phát triển web, phát triển game, đồ họa máy tính, và nhiều lĩnh vực khác. Python cũng là một trong những ngôn ngữ phổ biến nhất trong các khóa học lập trình cho người mới bắt đầu.

**(\*)** Một số đặc điểm của Python: cú pháp dễ đọc và dễ hiểu, hỗ trợ đa nền tảng (Windows, macOS, Linux, ...), có nhiều thư viện và framework phong phú,hỗ trợ nhiều mô hình lập trình khác nhau (hàm, hướng đối tượng, ...), hỗ trợ kiểu động (dynamic typing) và quản lý bộ nhớ tự động (automatic memory management).

# CHƯƠNG 3: TỐI ƯU HOÁ KHO DỮ LIỆU

**3.1 Tại sao cần phải tối ưu hóa kho dữ liệu**

Tối ưu hóa Kho dữ liệu là cần thiết và quan trọng vì nó mang lại nhiều lợi ích cho tổ chức hoặc doanh nghiệp như:

* Hiệu suất tốt hơn:
  + Tăng tốc độ truy xuất: Tối ưu hóa giúp giảm thời gian truy vấn dữ liệu, cải thiện hiệu suất truy cập thông tin.
  + Tăng cường khả năng xử lý: Cải thiện khả năng xử lý và phân tích dữ liệu nhanh hơn, giúp người dùng có thể làm việc hiệu quả hơn với dữ liệu lớn.
* Tối ưu hóa chi phí:
  + Giảm thiểu tài nguyên: Tối ưu hóa giúp sử dụng tài nguyên phần cứng và hệ thống hiệu quả hơn, giảm chi phí vận hành và quản lý.
  + Tiết kiệm dung lượng lưu trữ: Sử dụng kỹ thuật nén dữ liệu và loại bỏ dữ liệu không cần thiết giúp tiết kiệm dung lượng lưu trữ.
* Nâng cao chất lượng dữ liệu:
  + Đảm bảo tính nhất quán: Tối ưu hóa giúp chuẩn hóa và làm sạch dữ liệu, tăng tính nhất quán và chính xác của thông tin.
  + Phục vụ quyết định tốt hơn: Dữ liệu chất lượng cao hỗ trợ việc ra quyết định kinh doanh một cách chính xác và đáng tin cậy hơn.
* Tăng cường hiệu suất phân tích và dự đoán:
  + Hỗ trợ ra quyết định: Dữ liệu được tối ưu hóa giúp phân tích và dự đoán nhanh chóng, hỗ trợ quyết định kinh doanh một cách linh hoạt và hiệu quả.
  + Phát triển mô hình dự đoán: Tối ưu hóa giúp mô hình dự đoán hoạt động một cách hiệu quả và chính xác hơn.
* Mở rộng và tương thích tốt hơn:
  + Đáp ứng nhu cầu mở rộng: Tối ưu hóa hệ thống để dễ dàng mở rộng khi có nhu cầu lưu trữ và xử lý dữ liệu lớn hơn.
  + Tương thích và linh hoạt: Hệ thống được tối ưu hóa có khả năng tương thích tốt hơn với các công nghệ và hệ thống mới.

**(\*)** Kết luận: Tối ưu hóa data warehouse không chỉ giúp tăng hiệu suất mà còn giúp tổ chức có khả năng đáp ứng nhanh chóng với môi trường kinh doanh đang thay đổi, cũng như cải thiện khả năng cạnh tranh và sự linh hoạt trong việc sử dụng dữ liệu cho các quyết định chiến lược.

**3.2 Các cách tối ưu hóa kho dữ liệu**

Tối ưu hóa Data warehouse là quá trình cải thiện hiệu suất, khả năng truy cập và sử dụng dữ liệu một cách hiệu quả nhất. Dưới đây là một số cách để thực hiện việc tối ưu hóa data warehouse:

* Thiết kế cơ sở dữ liệu:
  + Cấu trúc dữ liệu tối ưu: Xây dựng một cấu trúc dữ liệu phù hợp và tối ưu để lưu trữ thông tin một cách logic và hiệu quả.
  + Quyết định loại cơ sở dữ liệu: Lựa chọn loại cơ sở dữ liệu phù hợp (SQL, NoSQL) tùy thuộc vào loại dữ liệu và yêu cầu của tổ chức.
  + Normalizing & Denormalizing: Ví dụ, trong một cơ sở dữ liệu lưu trữ thông tin khách hàng và đơn hàng, việc tách các bảng riêng biệt cho thông tin khách hàng và đơn hàng (normalization) có thể giúp tránh trùng lặp thông tin và tối ưu hóa việc cập nhật thông tin khách hàng mà không cần sửa đổi tất cả các đơn hàng liên quan.
  + Schema Design Optimization: Nếu một cơ sở dữ liệu lưu trữ dữ liệu từ các thiết bị IoT, việc thiết kế cấu trúc sao cho các dữ liệu thường xuyên truy cập như dữ liệu cảm biến được lưu trữ một cách hiệu quả và có thể truy cập nhanh chóng là quan trọng.
* Quản lý dữ liệu:
  + Xử lý dữ liệu trước khi lưu trữ: Làm sạch, chuẩn hóa và xử lý dữ liệu trước khi lưu vào kho dữ liệu để đảm bảo chất lượng và tính nhất quán của dữ liệu.
  + Tối ưu hóa lưu trữ: Sử dụng các kỹ thuật như nén dữ liệu, phân vùng dữ liệu và chỉ mục để giảm dung lượng lưu trữ và tăng tốc độ truy cập.
  + Query Optimization: Trong trường hợp của một trang web thương mại điện tử, việc viết truy vấn SQL tối ưu để hiển thị sản phẩm theo danh mục, giá hoặc đánh giá có thể tăng trải nghiệm người dùng và giảm thời gian phản hồi.
  + Materialized Views: Trong hệ thống quản lý đặt phòng khách sạn, việc tạo các bản sao về các thống kê tổng hợp như số lượng phòng trống theo ngày để giảm thời gian tính toán khi người dùng tìm kiếm có thể cải thiện hiệu suất.
* Tối ưu hóa truy xuất dữ liệu:
  + Chỉ mục dữ liệu: Tạo các chỉ mục cho các trường dữ liệu thường xuyên được truy vấn để tăng tốc độ truy vấn.
  + Tối ưu hóa các câu truy vấn: Sử dụng câu truy vấn hiệu quả và tối ưu để giảm thời gian xử lý và tăng hiệu suất truy xuất dữ liệu.
  + Query Optimization: Trong trường hợp của một trang web thương mại điện tử, việc viết truy vấn SQL tối ưu để hiển thị sản phẩm theo danh mục, giá hoặc đánh giá có thể tăng trải nghiệm người dùng và giảm thời gian phản hồi.
  + Materialized Views: Trong hệ thống quản lý đặt phòng khách sạn, việc tạo các bản sao về các thống kê tổng hợp như số lượng phòng trống theo ngày để giảm thời gian tính toán khi người dùng tìm kiếm có thể cải thiện hiệu suất.
* Tối ưu hóa về phần cứng và cơ sở hạ tầng:
  + Scale-out và Scale-up: Xây dựng cơ sở hạ tầng linh hoạt để có thể mở rộng khi cần thiết (scale-out) hoặc nâng cấp hiệu suất (scale-up).
  + Sử dụng công nghệ mới: Cân nhắc áp dụng công nghệ mới như bộ nhớ tốc độ cao (in-memory), công nghệ đám mây để cải thiện hiệu suất.
* Bảo mật và quản lý:
  + Áp dụng biện pháp bảo mật: Bảo vệ dữ liệu bằng cách sử dụng các biện pháp bảo mật như mã hóa, kiểm soát truy cập.
  + Quản lý và duy trì: Thực hiện sao lưu định kỳ, kiểm tra và duy trì hệ thống để đảm bảo tính sẵn sàng và an toàn của kho dữ liệu.
* Giám sát và tối ưu hóa liên tục:
  + Giám sát hiệu suất: Sử dụng công cụ giám sát để theo dõi và đo lường hiệu suất của data warehouse, xác định điểm yếu và cải thiện.
  + Tối ưu hóa liên tục: Liên tục thực hiện cải tiến và điều chỉnh dựa trên phản hồi và dữ liệu giám sát.
* Nâng cao hiệu suất:
  + Caching Strategies: Trong ứng dụng xem video trực tuyến, việc lưu trữ tạm thời các phần video đã được xem gần đây giúp người dùng truy cập lại nhanh hơn.
  + Scalability Techniques: Trong hệ thống mạng xã hội, việc sử dụng dọc ngang để mở rộng hệ thống cho số lượng người dùng tăng cao.
* Làm sạch và bảo trì dữ liệu:
  + Data Cleansing: Trong cơ sở dữ liệu khách hàng, việc loại bỏ các bản ghi trùng lặp hoặc không hoàn chỉnh (ví dụ: địa chỉ thiếu thông tin) giúp duy trì tính chính xác và hiệu quả của dữ liệu.
  + Data Archiving: Trong hệ thống quản lý bán lẻ, việc di dời lịch sử đơn hàng cũ vào kho dữ liệu lưu trữ để giảm tải cho cơ sở dữ liệu chính có thể cải thiện hiệu suất.
* Khám phá các công nghệ mới :
  + Distributed Databases: Trong hệ thống đặt vé máy bay, sử dụng cơ sở dữ liệu phân tán để xử lý thông tin đặt chỗ và giữ chỗ trên nhiều máy chủ để đảm bảo sự linh hoạt và độ tin cậy.
  + Cloud-Based Solutions: Trong dịch vụ lưu trữ và chia sẻ tài liệu, việc sử dụng dịch vụ đám mây như Google Drive hoặc Dropbox để lưu trữ và chia sẻ tài liệu giữa nhiều người dùng một cách hiệu quả và an toàn.

**(\*)** Tối ưu hóa data warehouse là quá trình liên tục và đòi hỏi sự theo dõi và điều chỉnh thường xuyên để duy trì hiệu suất và khả năng mở rộng khi cần thiết.

# CHƯƠNG 4: ỨNG DỤNG KHO DỮ LIỆU VÀO BÀI TOÁN QUẢN LÝ Ô TÔ

**4.1 Phân tích và thiết kế kho dữ liệu**

***4.1.1 Thực thể đo lường và số nguyên***

Trong bối cảnh kho dữ liệu và kinh doanh thông minh, "đo lường" và "thứ nguyên" là hai khái niệm cơ bản đóng vai trò quan trọng trong việc tổ chức và phân tích dữ liệu.

Đo lường: Một thước đo đại diện cho một giá trị số hoặc một thuộc tính định lượng được sử dụng để phân tích và tính toán. Các biện pháp thường là các giá trị mà chúng tôi muốn phân tích, chẳng hạn như doanh thu bán hàng, số lượng đã bán, lợi nhuận hoặc giá trung bình. Các thước đo thường được liên kết với các hàm tổng hợp, chẳng hạn như tổng, trung bình, số đếm hoặc tối đa/tối thiểu, cho phép chúng tôi thực hiện các phép tính và rút ra thông tin chi tiết từ dữ liệu.

**(\*)** Ví dụ: trong bối cảnh quản lý doanh số bán ô tô, các biện pháp có thể bao gồm tổng doanh thu được tạo ra, số lượng ô tô đã bán hoặc giá trung bình của ô tô đã bán. Các biện pháp này cung cấp thông tin định lượng có thể được tổng hợp và phân tích để hiểu hiệu suất của quy trình bán hàng.

* Thứ nguyên: Thứ nguyên đại diện cho thuộc tính mô tả hoặc danh mục cung cấp ngữ cảnh và thông tin bổ sung về các biện pháp. Thứ nguyên được sử dụng để cung cấp các khía cạnh "ai, cái gì, khi nào, ở đâu và như thế nào" của dữ liệu. Chúng giúp tổ chức, lọc và nhóm dữ liệu để phân tích.
* Thứ nguyên thường được sử dụng để cắt và chia nhỏ dữ liệu để đạt được các quan điểm khác nhau. Các kích thước phổ biến bao gồm thời gian, địa lý, sản phẩm, khách hàng và nhân viên. Ví dụ: trong kịch bản bán ô tô, các thứ nguyên có thể bao gồm thứ nguyên thời gian (ví dụ: năm, quý, tháng), thứ nguyên sản phẩm (ví dụ: hãng xe, kiểu xe), thứ nguyên khách hàng (ví dụ: tên khách hàng, thông tin nhân khẩu học), và thứ nguyên vị trí (ví dụ: thành phố, vùng).
* Thứ nguyên cung cấp bối cảnh và mức độ chi tiết để phân tích các biện pháp. Chúng giúp phân đoạn dữ liệu, đi sâu vào các tập hợp con cụ thể và lọc dữ liệu dựa trên các tiêu chí khác nhau. Thứ nguyên cũng đóng một vai trò quan trọng trong việc xác định cấu trúc phân cấp, cho phép người dùng điều hướng qua các cấp độ chi tiết khác nhau trong một thứ nguyên.

Tóm lại, thước đo đại diện cho các giá trị số mà chúng tôi muốn phân tích, trong khi thứ nguyên cung cấp các thuộc tính mô tả bổ sung ngữ cảnh và cho phép phân tích từ các quan điểm khác nhau.

Sự kết hợp của các biện pháp và kích thước tạo thành nền tảng cho phân tích và báo cáo đa chiều trong kho dữ liệu và kinh doanh thông minh.

***4.1.2 Thứ bậc của kích thước***

Trong bối cảnh lưu trữ dữ liệu và mô hình hóa thứ nguyên, hệ thống phân cấp thứ nguyên đề cập đến việc tổ chức các thuộc tính thứ nguyên theo cách có cấu trúc và thứ bậc. Thứ nguyên đại diện cho một loại dữ liệu cung cấp ngữ cảnh bổ sung hoặc thông tin mô tả cho các sự kiện trong kho dữ liệu. Ví dụ về thứ nguyên bao gồm thời gian, địa lý, sản phẩm và khách hàng.

Hệ thống phân cấp cho phép nhóm hợp lý và tổ chức các thuộc tính thứ nguyên dựa trên mức độ chi tiết hoặc mức độ chi tiết của chúng. Chúng xác định các mối quan hệ và sự phụ thuộc giữa các thuộc tính trong một thứ nguyên. Một hệ thống phân cấp thường bao gồm nhiều cấp độ, với mỗi cấp độ đại diện cho một cấp độ chi tiết hoặc tổng hợp khác nhau.

**(\*)** Ví dụ, xem xét một chiều thời gian. Nó có thể có một hệ thống phân cấp bao gồm các cấp độ như Năm, Quý, Tháng, Tuần và Ngày. Mỗi cấp độ đại diện cho một mức độ chi tiết thời gian khác nhau, với Năm là cấp độ cao nhất và Ngày là cấp độ thấp nhất.

* Hệ thống phân cấp cung cấp một số lợi ích trong phân tích dữ liệu và báo cáo:
* Truy sâu xuống và cuộn lên: Người dùng có thể điều hướng qua hệ thống phân cấp để phân tích dữ liệu ở các mức độ chi tiết khác nhau. Họ có thể đi sâu từ tóm tắt cấp cao hơn đến dữ liệu chi tiết hơn hoặc cuộn lên từ dữ liệu chi tiết đến tóm tắt cấp cao hơn. Điều này cho phép phân tích linh hoạt và năng động.

**(\*)** Tổng hợp và tóm tắt: Hệ thống phân cấp cho phép tổng hợp và tóm tắt dữ liệu ở các cấp độ khác nhau. Người dùng có thể xem dữ liệu ở mức tổng quan cao hoặc đi sâu vào để kiểm tra các chi tiết cụ thể. Điều này hỗ trợ báo cáo và phân tích hiệu quả.

* Khám phá và điều hướng dữ liệu: Hệ thống phân cấp cung cấp một cách có cấu trúc và trực quan để khám phá và điều hướng dữ liệu. Người dùng có thể dễ dàng điều hướng qua các cấp để hiểu rõ hơn và hiểu mối quan hệ giữa các thuộc tính khác nhau.
* Tính nhất quán và tiêu chuẩn hóa: Hệ thống phân cấp giúp duy trì tính nhất quán và tiêu chuẩn hóa trên khắp kho dữ liệu. Bằng cách xác định hệ thống phân cấp tiêu chuẩn cho thứ nguyên, dữ liệu được tổ chức thống nhất, đảm bảo phân tích chính xác và nhất quán trên các báo cáo và ứng dụng khác nhau.
* Điều quan trọng là phải thiết kế hệ thống phân cấp một cách cẩn thận để phù hợp với yêu cầu kinh doanh và nhu cầu của người dùng. Việc lựa chọn cấp độ và mối quan hệ giữa các thuộc tính phải phản ánh hệ thống phân cấp tự nhiên trong thứ nguyên và cung cấp các khả năng phân tích có ý nghĩa. Hệ thống phân cấp được thiết kế tốt góp phần phân tích dữ liệu, báo cáo và ra quyết định hiệu quả trong môi trường lưu trữ dữ liệu.

***4.1.3 Sơ đồ mô hình khái niệm***

Sơ đồ mô hình khái niệm, còn được gọi là sơ đồ mô hình khái niệm hoặc mô hình dữ liệu khái niệm, là một biểu đồ hình ảnh mô tả các khái niệm, thực thể, mối quan hệ và thuộc tính cấp cao cần thiết để hiểu một lĩnh vực hoặc vấn đề cụ thể. Nó được sử dụng để mô tả cấu trúc tổng thể và tổ chức của dữ liệu ở mức độ khái niệm, mà không tập trung vào các chi tiết cụ thể về triển khai hoặc các khía cạnh kỹ thuật.

Mục đích của sơ đồ mô hình khái niệm là cung cấp một cái nhìn tổng quan rõ ràng và ngắn gọn về các yếu tố chính và mối quan hệ giữa chúng trong một hệ thống hoặc ứng dụng. Nó giúp các bên liên quan, như các nhà phân tích kinh doanh, nhà thiết kế và nhà phát triển, có cùng một hiểu biết về lĩnh vực đang được mô hình hóa và các yêu cầu của hệ thống.

**(\*)** Một số loại sơ đồ mô hình khái niệm phổ biến bao gồm:

* Sơ đồ Thực thể-Mối quan hệ (ERD): ERD biểu thị các thực thể, mối quan hệ và thuộc tính liên quan trong một hệ thống hoặc lĩnh vực vấn đề. Chúng miêu tả cấu trúc tổng quan của dữ liệu và cách các thực thể khác nhau liên quan đến nhau.
* Sơ đồ Lớp: Sơ đồ lớp chủ yếu được sử dụng trong lập trình hướng đối tượng để biểu diễn các lớp, thuộc tính của chúng và mối quan hệ giữa chúng. Chúng cung cấp một biểu đồ hình ảnh về thiết kế hướng đối tượng của một hệ thống.
* Sơ đồ Ngôn ngữ thống nhất (UML): Sơ đồ UML, chẳng hạn như sơ đồ lớp, sử dụng một cú pháp chuẩn để biểu diễn các khái niệm, mối quan hệ và hành vi của một hệ thống. Chúng có thể được sử dụng cho mô hình hóa khái niệm trong các lĩnh vực khác.
* Sơ đồ Lược đồ quan hệ thực thể (ERD): Sơ đồ ERD giúp mô hình hóa cấu trúc của cơ sở dữ liệu quan hệ. Nó biểu diễn các thực thể, mối quan hệ và các ràng buộc giữa chúng dưới dạng các bảng, khóa chính và khóa ngoại.
* Sơ đồ Dự án: Sơ đồ Dự án được sử dụng để biểu diễn cấu trúc tổ chức của một dự án, bao gồm các thành phần, phụ thuộc và quan hệ giữa chúng. Nó giúp hiển thị quy trình công việc, vai trò và trách nhiệm của các thành viên trong dự án.
* Sơ đồ mô hình khái niệm đóng vai trò là cơ sở cho các hoạt động phân tích, thiết kế và triển khai tiếp theo. Nó giúp các bên liên quan xác định các thực thể và mối quan hệ chính, xác định phạm vi và ranh giới của hệ thống, và tạo điều kiện cho việc giao tiếp và cộng tác giữa các thành viên trong dự án.

**(\*)** Lưu ý rằng sơ đồ mô hình khái niệm tập trung vào mức độ trừu tượng cao của hệ thống và có thể không bao gồm tất cả các chi tiết cụ thể hoặc yếu tố kỹ thuật triển khai. Nó cung cấp một khung cơ bản mà có thể được sử dụng làm cơ sở cho việc mô hình hóa và phát triển hệ thống chi tiết hơn.

**4.2 Mô hình logic**

***4.2.1.Fact và dimension table (PK,FK)***

Fact table có thể được hiểu như là bảng chứa các dữ liệu có tính chất đo lường (measurement). Một fact (hay còn gọi là measure) trong data warehouse được dùng để minh họa cho một trường (column) chứa một giá trị đo lường được và đóng một vai trò quan trọng với business. Trên thực tế, ta hay gặp nhất các fact ở dạng số (numeric) và có tính chất cộng (additive). Do cấu trúc của fact table khá đơn giản chỉ chứa duy nhất foreign key và fact, ngoài ra không có bất cứ thông tin nào khác nên tốc độ truy cập bảng khá nhanh. Một kỹ thuật hay dùng là tạo index cho các foreign key. Một kỹ thuật nữa là xây dựng sẵn các aggregate tables để tính toán trước một đại lượng nào đó dựa trên fact table. Fact grain (độ mịn) là một khái niệm để xác định mức độ chi tiết của thông tin chứa trong fact table. Một quy định bắt buộc khi thiết kế fact table là: Tất cả các record trong fact table phải có cùng độ mịn. Fact Grain là thành tố cơ bản của fact table. Xác định đúng fact grain là một công tác quan trọng bậc nhất trong việc xây dựng một mô hình đa chiều. Đây là công việc đòi hỏi người thiết kế phải là những Data Architect giàu kinh nghiệm và có hiểu biết sâu sắc về business domain đang làm.

Nếu như fact table chứa các FK (foreign key) và measure, thì dimesion table chứa các thông tin miêu tả nghiệp vụ. Trong không gian của mô hình đa chiều, các thông tin này được gọi là thuộc tính (attribute) của dimension. Còn khi được lưu trữ cụ thể trong một cơ sở dữ liệu quan hệ, các thông tin này chính là các columns của table. Một nguyên tắc thiết kế của data warehouse là cố gắng đưa càng nhiều thông tin vào dimension thì càng tốt Dimension table đóng một vai trò sống còn trong data warehouse. Do chúng chứa các thông tin miêu tả nghiệp vụ, nên đây là nguồn được dùng để sàng lọc dữ liệu, gộp nhóm, và tạo các nhãn trình bày.

Trong ví dụ ở trên, nếu người dùng muốn xem tổng số tiền bán hàng theo tuần và loại hàng hóa, hiển nhiên fact table chỉ có thể cung cấp số liệu về số tiền thu được. Thông tin về loại hàng hóa chỉ có thể lấy được từ Product dimension, trong khi thông tin về thời gian cũng chỉ có thể tìm được trong Date dimension. Nói cách khác, người dùng tương tác với data warehouse chủ yếu là qua dimension table, còn fact table chỉ cung cấp các số liệu. Tuỳ theo cách nhìn của người dùng đối với nghiệp vụ, họ sẽ dùng các attributes của dimension hoặc là để sàng lọc dữ liệu, hoặc là để tạo ra các kết quả tổng hợp theo nhóm vv. Các attributes của dimension table thường chứa dữ liệu dạng text và có tính rời rạc. Một attribute có thể chứa một miêu tả ngắn (10 - 15 ký tự) hoặc dài (30-50 ký tự), tên hàng, tên loại hàng, kiểu đóng gói, kích cỡ... vv. Mặc dù kích cỡ là một giá trị dạng số và dường như thích hợp cho fact table, nhưng về bản chất mà nói, nó mang tính miêu tả cho một hàng hóa cụ thể chứ không mang nhiều giá trị đo lường. Các attributes trong một dimesion thường tạo thành các cây phân cấp (hierarchy). Có thể nói, trong thiết kế dimension table, ta hy sinh chuẩn 3NF để đổi lấy sự đơn giản và tốc độ. Đây có thể coi là một nguyên tắc vàng của dimensional modeling.

* Có thể chia ra những đặc điểm sau để phân biệt giữa 2 loại table:

|  |  |  |
| --- | --- | --- |
| Đặc điểm | Fact table | Dimension table |
| Định nghĩa | Bao gồm các chỉ số đo lường các hoạt động trong tổ chức. | Bảng thông tin các đối tượng có liên quan tới bảng Fact. |
| Thuộc tính | Trung tâm trong Star schema và Snowflake schema, xung quanh là các bảng Dim. | Nằm xung quanh bảng Fact trong Star schema và Snowflake Schema. |
| Thiết kế | Thiết kế đúng với trọng tâm của vấn đề hay nhu cầu. | Phải bao quát các thông tin của một đối tượng. Tránh nhét vào nhiều thông tin của nhiều đối tượng cùng  một nơi. |
| Phục vụ | Một bảng Fact tập hợp các chỉ số đo lượng có liên quan tới các bảng Dim, dùng cho báo cáo, thống kê. | Thu thập các thông tin liên quan tới một đối tượng nào đó trong tổ chức (như thông tin cửa hàng chẳng hạn). |
| Kiểu dữ liệu | Bảng Fact chứa các cột dữ liệu số liên quan tới các cột đối tượng, khoá ngoại của bảng Dim. | Mỗi bảng Dim chứa các thông tin về một đối tượng. Ví dụ bảng Product gồm Product Id, Category,  SubCategory… |
| Khoá chính | Khoá chính của bảng Fact là tập hợp không lặp lại các khoá ngoại từ các bảng Dim. | Mỗi bảng Dim đều có một khoá chính để xác định rõ tính độc nhất của mỗi dòng trong bảng. |
| Lưu trữ | Được dùng để lưu trữ chỉ số, đồng thời để lọc kết quả trên nó dựa theo thông tin từ các bảng Dim. | Lưu trữ dữ liệu các đối tượng một cách chi tiết. |
| Sự phân cấp | Không có | Có sự phân cấp: Ví dụ đối tượng Địa Điểm có thể chứa Quốc Gia, Khu vực,  Quận/Huyện,MãZip |

***4.2.2 Sơ đồ bông tuyết và sơ đồ lai***

Khi ta liên kết fact table và dimension table lại với nhau dựa trên các Primary Key của dimension và Foreign Key tương ứng của fact, ta được một mô hình dữ liệu dạng hình sao - star schema.

**(\*)**Ví dụ như ta có các bảng như phía dưới:

A diagram of a product

Description automatically generated with medium confidence

Hình 9.Sơ đồ bông tuyết quản li ô tô

A screenshot of a computer

Description automatically generated

Hình 10.Bảng customer

A screenshot of a computer

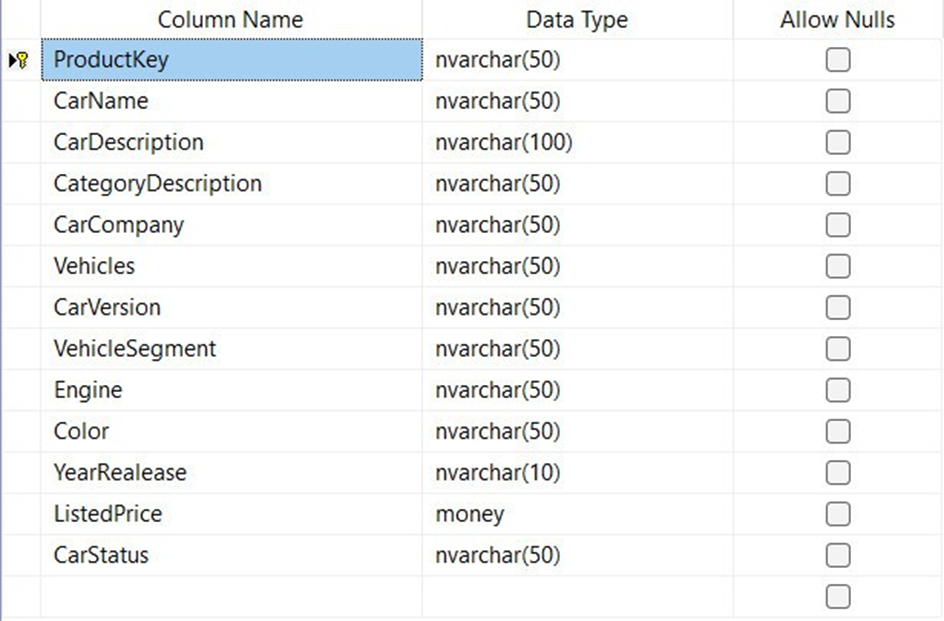
Description automatically generated

Hình 11.Bảng Date

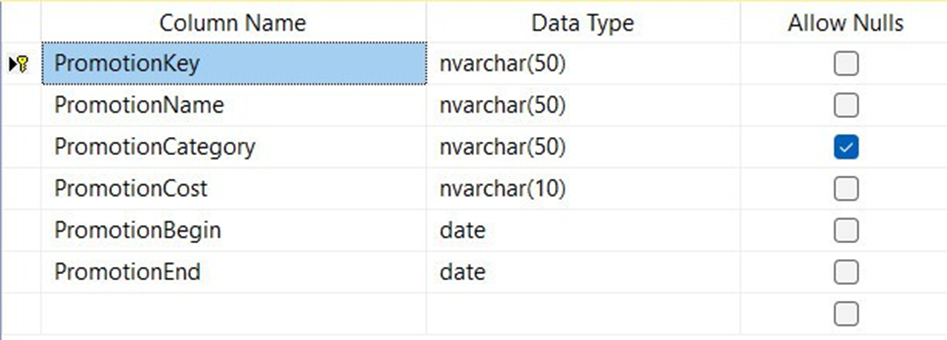
A screenshot of a computer

Description automatically generated

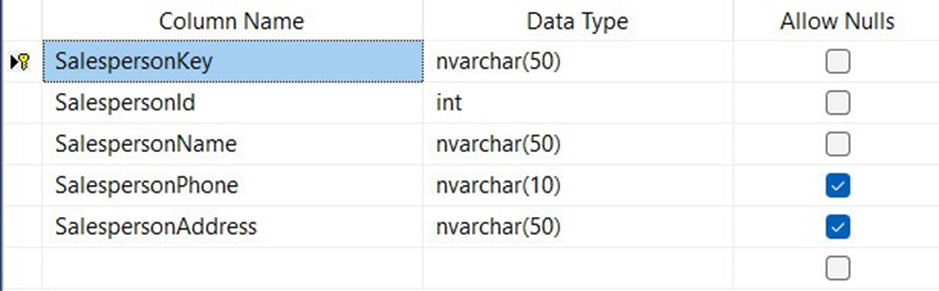
Hình 12.Bảng Payment Method

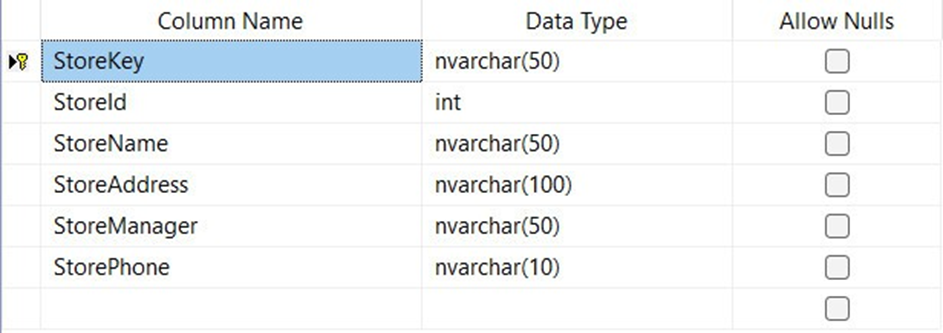


Hình 13.Bảng product

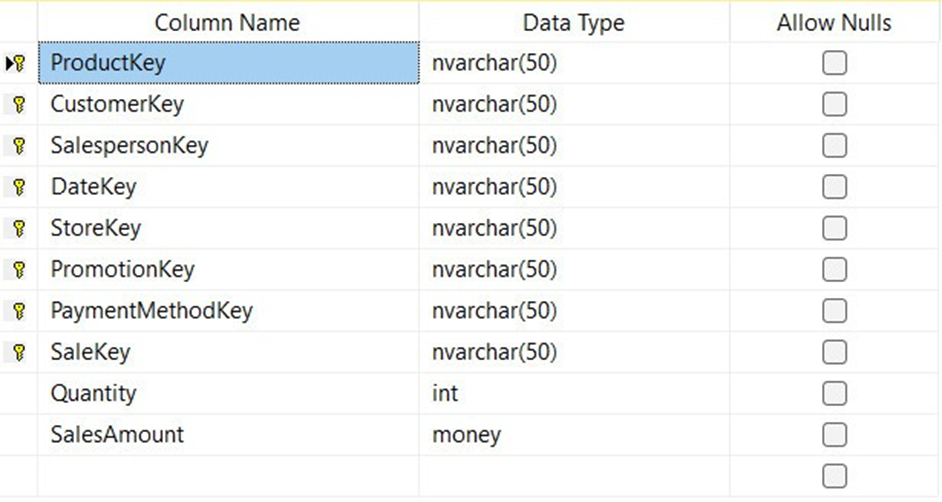


Hình 14.Bảng promotion

Hình 15.Bảng saleperson



Hình 16.Bảng store



Hình 17.Bảng Factsaleorder

Không quá khó để thấy rằng star schema có thiết kế tương đối đơn giản và mang tính đối xứng. Nếu xét từ góc độ thuần túy nghiệp vụ, thiết kế này gần nhất với cái nhìn của người dung cuối về nghiệp vụ của họ. Trong nhiều trường hợp, đây chính là bức tranh chính xác của nghiệp vụ mà người dùng làm việc trong đó.

Thực ra nếu ta suy nghĩ sâu hơn một chút về quá trình xây dựng một cơ sở dữ liệu quan hệ cho một ứng dụng rồi xây dựng data warehouse cho ứng dụng đó, nó gần như tạo thành một vòng tuần hoàn. Những ai đã từng thiết kế cơ sở dữ liệu quan hệ - relational database - đều biết rằng một trong những bước đầu tiên là phải xác định được mô hình quan hệ - thực thể (Entity – Relationship). Sau khi các thực thể và các mối quan hệ giữa chúng đã được xác định, ta mới tiến hành xây dựng các table và column và rồi áp dụng chuẩn 3NF. Trong quá trình này, các thực thể có thể được phân nhỏ ra để tạo nên các table mới với các quan hệ 1-N. Các mối quan hệ đa chiều N-N trong mô hình E-R sẽ được thể hiện bằng một bảng trung gian với các quan hệ 1-N tương ứng. Bằng cách này, một mô hình E-R với một lượng nhỏ thực thể ban đầu có thể phát triển thành một cơ sỡ dữ liệu lớn với hàng trăm bảng được kết nối với nhau qua các mối quan hệ 1-N hoặc 1-1. Khối lượng thông tin như vậy rõ ràng là không hề dễ hiểu với người dùng cuối, vốn là những đối tượng hầu như không có chút khái niệm nào về database.

Khi xây dựng một data warehouse theo dạng star schema, một cách tự nhiên ta đã biến một cơ sở dữ liệu phức tạp quay trở lại một dạng mô hình E-R gần như ban đầu. Dĩ nhiên trong thực tế, một star schema thường tập trung vào một góc nhỏ của relational database. Do đó, từ một cơ sở dữ liệu ban đầu có thể xây dựng nhiều star schema, và mỗi star schema tập trung vào một khía cạnh cụ thể của nghiệp vụ. Trong ví dụ trên, Product là một thực thể. Khi được đưa vào cơ sở dữ liệu quan hệ, nó được chuẩn hóa theo 3NF và tách ra thành 3 bảng Product, SubCategory, Category. Khi đưa vào data warehouse, Product lại quay về hình dạng của thực thể Product ban đầu.

* Snow Flake:

Là dạng mở rộng của lược đồ hình sao bằng các bổ sung các Dim. Bảng Fact như lược đồ hình sao, bảng Dim được chuẩn hoá. Các chiều được cấu trúc rõ ràng. Bảng Dim được chia thành chiều chính hay chiều phụ.

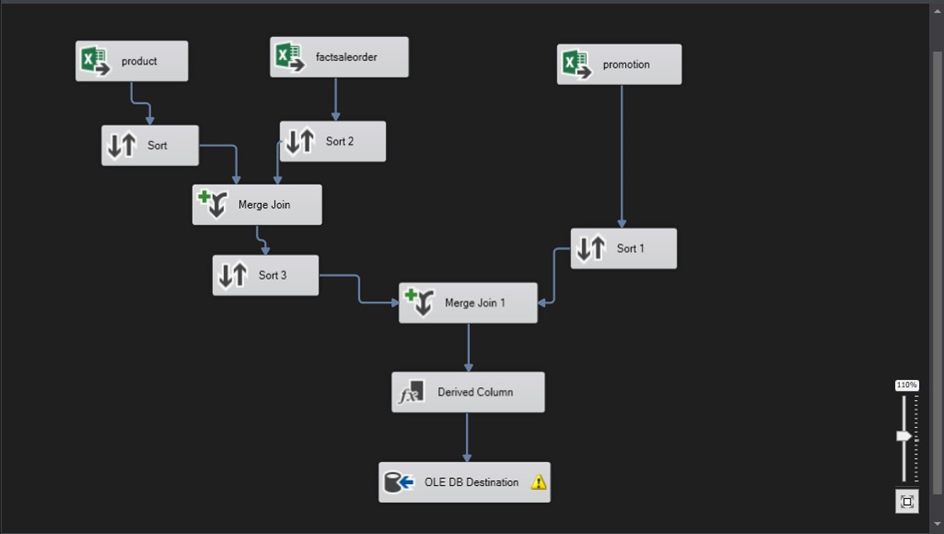
**(\*)**Ưu điểm: Số chiều được phân cấp thể hiện dạng chuẩn của bảng Dim.

**(\*)**Nhược điểm: Cấu trúc phi dạng chuẩn của lược đồ hình sao phù hợp hơn cho việc duyệt các chiều.

**4.3 Triển khai kho dữ liệu**

***4.3.1 Mô hình SSIS***

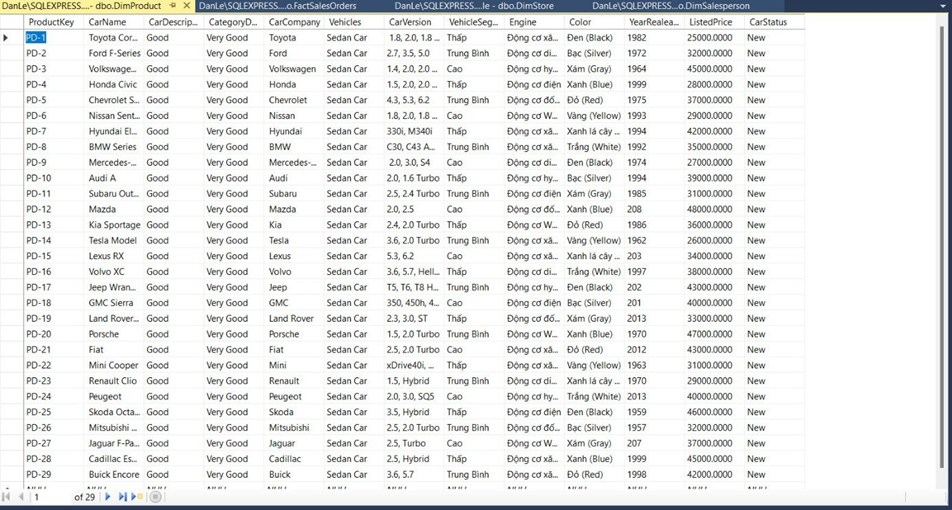
Sử dụng 3 Bảng Product, FactSaleOrder và Promotion để Kết hợp và ra một bảng Product mới:



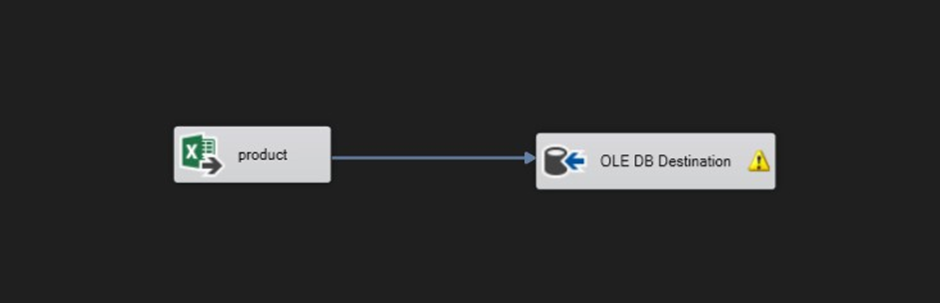
Hình 18.Mô Hình SSIS

***4.3.2 Quá trình ETH***

Sau khi hoàn tất sơ đồ SSIS và khởi chạy quá trình ETH để tạo ra 1 bảng dữ liệu Product mới:



Hình 19.Bảng data product mới



Hình 20.Import dữ liệu vào excel

**4.4 Đánh giá kết quả**

***4.4.1 Kết quả đạt được***

Hiểu biết về cấu trúc và quy trình xây dựng kho dữ liệu: Bạn sẽ hiểu được các khái niệm, phương pháp và quy trình thiết kế, xây dựng và triển khai kho dữ liệu. Điều này giúp bạn có khả năng tổ chức và quản lý dữ liệu một cách hiệu quả trong một hệ thống kho dữ liệu.

Kỹ năng thiết kế và triển khai kho dữ liệu: Bạn sẽ học cách thiết kế mô hình dữ liệu, chuẩn hóa dữ liệu và xác định quy trình ETL để trích xuất, biến đổi và tải dữ liệu vào kho dữ liệu. Kỹ năng này cho phép bạn xây dựng các hệ thống kho dữ liệu chất lượng cao và có khả năng phục vụ nhu cầu phân tích và báo cáo.

Nắm vững công nghệ và công cụ hỗ trợ: Bạn sẽ làm quen với các công nghệ và công cụ quan trọng trong lĩnh vực kho dữ liệu như hệ quản trị kho dữ liệu (Data Warehouse Management System - DWMS), ngôn ngữ truy vấn, công cụ truy vấn và báo cáo, các phương pháp khai thác dữ liệu (data mining) và các công nghệ lưu trữ và tối ưu hóa dữ liệu.

Khả năng phân tích dữ liệu và đưa ra quyết định: Với một hệ thống kho dữ liệu chất lượng, bạn sẽ có khả năng truy vấn và phân tích dữ liệu một cách hiệu quả. Bằng cách sử dụng các công cụ và kỹ thuật phân tích dữ liệu, bạn có thể tìm ra thông tin quan trọng, xu hướng, mẫu đối tượng và quy luật ẩn trong dữ liệu. Điều này hỗ trợ quyết định và phân tích kinh doanh hiệu quả.

Nắm vững khái niệm business intelligence và big data: Môn kho dữ liệu cung cấp cho bạn nền tảng để hiểu về business intelligence (BI) và big data. Bạn sẽ có cái nhìn tổng quan về cách sử dụng dữ liệu để hỗ trợ quyết định kinh doanh và làm việc với dữ liệu lớn và phức tạp.

***4.4.2 Mặt hạn chế***

* Mặt hạn chế của kho dữ liệu bán xe ô tô có thể bao gồm:

Khó khăn trong việc tích hợp dữ liệu: Kho dữ liệu bán xe ô tô thường phải tích hợp dữ liệu từ nhiều nguồn khác nhau như hệ thống bán hàng, hệ thống CRM, trang web bán hàng và các nguồn dữ liệu bên ngoài. Việc tích hợp dữ liệu từ các nguồn khác nhau có thể gặp khó khăn do sự không nhất quán trong cấu trúc dữ liệu, định dạng dữ liệu khác nhau và quá trình ETL phức tạp.

Dữ liệu không chính xác và không đồng nhất: Dữ liệu trong kho dữ liệu bán xe ô tô có thể không chính xác hoặc không đồng nhất. Điều này có thể do dữ liệu nhập liệu sai sót, dữ liệu thiếu sót hoặc không tuân thủ quy trình chuẩn hóa dữ liệu. Dữ liệu không chính xác hoặc không đồng nhất có thể ảnh hưởng đến tính chính xác và tin cậy của các phân tích và báo cáo.

Độ trễ trong quá trình cập nhật dữ liệu: Việc cập nhật dữ liệu trong kho dữ liệu bán xe ô tô có thể gặp độ trễ do quá trình ETL và tích hợp dữ liệu từ các nguồn khác nhau. Điều này có thể làm giảm tính sát thời và độ tin cậy của dữ liệu, đặc biệt khi xử lý các dữ liệu thời gian thực và theo dõi các thay đổi nhanh chóng trong lĩnh vực bán xe ô tô.

Khó khăn trong việc quản lý dữ liệu lớn: Kho dữ liệu bán xe ô tô có thể chứa một lượng lớn dữ liệu, bao gồm thông tin về hàng nghìn xe, khách hàng, giao dịch và dịch vụ. Quản lý dữ liệu lớn đòi hỏi cơ sở hạ tầng lưu trữ và xử lý mạnh mẽ, gây áp lực lên hệ thống và tăng chi phí vận hành.

Hạn chế về quyền riêng tư và bảo mật: Kho dữ liệu bán xe ô tô chứa nhiều thông tin nhạy cảm về khách hàng, bao gồm thông tin cá nhân, lịch sử mua hàng và tài chính. Bảo vệ quyền riêng tư và đảm bảo bảo mật dữ liệu là một thách thức đối với kho dữ liệu bán xe ô tô, đặc biệt là trong bối cảnh các quy định bảo vệ dữ liệu ngày càng nghiêm ngặt.

Khó khăn trong việc phân tích dữ liệu phức tạp: Kho dữ liệu bán xe ô tô có thể chứa dữ liệu phức tạp và đa dạng về xe, khách hàng và giao dịch. Phân tích dữ liệu phức tạp và trích xuất thông tin có giá trị từ kho dữ liệu đòi hỏi sự hiểu biết sâu về ngành công nghiệp ô tô và kỹ năng phân tích dữ liệu phức tạp.

Đối mặt với những hạn chế này, quản lý và phát triển kho dữ liệu bán xe ô tô đòi hỏi sự chú trọng đến việc quản lý dữ liệu, tối ưu hóa quy trình ETL, áp dụng các biện pháp bảo mật và tuân thủ quy định, và đảm bảo tính chính xác và đồng nhất của dữ liệu.

***4.4.3 Hướng phát triển***

* Hướng phát triển của một kho dữ liệu bán xe ô tô có thể bao gồm các yếu tố sau:

Mở rộng và cập nhật nguồn dữ liệu: Để kho dữ liệu bán xe ô tô phát triển, cần liên tục mở rộng và cập nhật nguồn dữ liệu. Điều này có thể bao gồm việc tích hợp dữ liệu từ các nguồn khác nhau như hệ thống bán hàng, hệ thống CRM (quản lý mối quan hệ khách hàng), hệ thống quản lý dịch vụ, trang web bán hàng và các nguồn dữ liệu bên ngoài khác. Cập nhật thông tin về xe mới, giá cả, thông số kỹ thuật và thông tin khách hàng là quan trọng để duy trì tính chính xác và hiện đại của kho dữ liệu.

Mở rộng phạm vi phân tích: Kho dữ liệu bán xe ô tô có thể mở rộng phạm vi phân tích để cung cấp thông tin quan trọng cho các bộ phận trong công ty như bán hàng, marketing, dịch vụ, nghiên cứu và phát triển sản phẩm. Các phân tích có thể bao gồm: xu hướng bán hàng, tình hình thị trường, phân loại khách hàng, đánh giá hiệu quả chiến dịch tiếp thị, sự hài lòng của khách hàng và phân tích dữ liệu về dịch vụ sau bán hàng.

Tích hợp dữ liệu với công nghệ mới: Hướng phát triển của kho dữ liệu bán xe ô tô cũng có thể liên quan đến tích hợp với các công nghệ mới và xu hướng công nghệ như trí tuệ nhân tạo (AI), học máy (machine learning) và phân tích dữ liệu lớn (big data analytics). Các công nghệ này có thể được áp dụng để cải thiện việc phân tích dữ liệu, dự đoán xu hướng, tối ưu hóa chiến dịch tiếp thị và cải thiện trải nghiệm khách hàng.

Tăng cường bảo mật và tuân thủ quy định: Kho dữ liệu bán xe ô tô cần đảm bảo tính bảo mật và tuân thủ các quy định về bảo vệ dữ liệu cá nhân như GDPR (General Data Protection Regulation) và CCPA (California Consumer Privacy Act). Hướng phát triển có thể bao gồm việc áp dụng các biện pháp bảo mật tiên tiến, quản lý quyền truy cập dữ liệu và tuân thủ quy định về quyền riêng tư.

Tích hợp với hệ sinh thái công nghệ: Hướng phát triển của kho dữ liệu bán xe ô tô cũng có thể liên quan đến tích hợp với hệ sinh thái công nghệ khác như hệ thống CRM, hệ thống quản lý tổ chức (ERP), hệ thống báo cáo và dashboard, và các công cụ phân tích dữ liệu.

Điều này giúp tạo ra một hệ thống toàn diện và tích hợp, nâng cao hiệu quả và khả năng đưa ra quyết định trong công ty.

Quan trọng nhất, hướng phát triển của kho dữ liệu bán xe ô tô phụ thuộc vào mục tiêu kinh doanh và yêu cầu cụ thể của công ty. Việc liên tục theo dõi và tương tác với các bộ phận trong công ty sẽ giúp xác định được hướng phát triển tối ưu cho kho dữ liệu bán xe ô tô.

# KẾT LUẬN

**\*\*\*oOo\*\*\***

Qua một thời gian học hỏi và tìm hiểu, nhóm em đã xây dựng được chương trình tối ưu hoá kho dữ liệu và ứng dụng kho dữ liệu quản lý ô tô. Chương trình đã áp dụng các phương pháp tối ưu hoá kho dữ liệu và xây dựng hệ thống quản lí ô tô không dữ liệu, mở ra những cánh cửa mới cho hiệu suất và tính linh hoạt trong ngành công nghiệp ô tô.

Qua việc nghiên cứu và thực hành, chúng ta đã nhìn thấy sức mạnh của dữ liệu thông minh trong việc tối ưu hoá quy trình quản lí, tạo ra trải nghiệm người dùng tốt hơn và đẩy mạnh sự tiến bộ của ngành này.

Em xin chân thành cảm ơn sự hướng dẫn quý báu từ cô Đào Thu Vân, những kiến thức hữu ích đã giúp nhóm chúng em hoàn thành bài thuyết trình theo đúng tiến độ và mở ra những triển vọng tuyệt vời cho tương lai.

## TÀI LIỆU THAM KHẢO

|  |  |
| --- | --- |
| [1] | [Online]. Available: https://aws.amazon.com/vi/what-is/data-warehouse/.. |
| [2] | [Online]. Available: https://bkhost.vn/blog/data-warehouse-kho-du-lieu/... |
| [3] | L. H. Sinh, Kho dữ liệu - Dataware House. |
| [4] | Xây dựng Hệ thống Kho dữ liệu và Khai thác Dữ liệu, Đại Học Bách Khoa TP.HCM. |