HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

PHÂN TÍCH HIỆU NĂNG CỦA CÁC KỸ THUẬT BẢO TRÌ KHUNG NHÌN CỦA KHO DỮ LIỆU

Chuyên ngành: Hệ thống thông tin

Mã số: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – NĂM 2013

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS. Đặng Văn Chuyết**

Phản biện 1: ……

Phản biện 2: …….

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Trong kho dữ liệu việc sử dụng khung nhìn đem lại lợi ích cho các tổ chức doanh nghiệp đó là vấn đề bảo mật dữ liệu**,** đơn giản hoá các thao tác truy vấn dữ liệu, tập trung và đơn giản hoá dữ liệu**,** độc lập dữ liệu. Làm thế nào để bảo trì các khung nhìn thực sao cho chúng vẫn được duy trì khi cập nhật các quan hệ thực tế ở các nguồn dữ liệu thì lúc nào kỹ thuật bảo trì khung nhìn ra đời. Các kỹ thuật bảo trì khung nhìn kho dữ liệu được chia làm hai nhóm lớn: bảo trì theo phương pháp tính lại và phương pháp bảo trì lũy tiến. Tùy thuộc vào việc kho dữ liệu có truy vấn nguồn dữ liệu từ xa để tính lại khung nhìn mới không, các kỹ thuật này lại được phân thành cơ chế tự duy trì và không tự duy trì. Vì vậy, có bốn nhóm kỹ thuật: tính lại có cơ chế tự duy trì, tính lại không có cơ chế tự duy trì, bảo trì lũy tiến có cơ chế tự duy trì, bảo trì lũy tiến không có cơ chế tự duy trì. Nhưng để ứng dụng các kỹ thuật bảo trì khung nhìn này thực tế thì ta phải đánh giá được khả năng của mỗi loại bảo trì khung nhìn. Vì vậy, em chọn nghiên cứu đề tài „*Phân tích hiệu năng của các kỹ thuật bảo trì khung nhìn của kho dữ liệu*“ nghiên cứu về các kỹ thuật bảo trì khung nhìn của kho dữ liệu. Thông qua đó đánh giá được không gian sử dụng trong kho dữ liệu, số hàng truy nhập trong kho dữ liệu để tích hợp và bổ sung kho dữ liệu.

CHƯƠNG 1: TỔNG QUAN

1.1. Khái niệm

Theo John Ladley, **Kỹ thuật kho dữ liệu** (Data Warehouse Technology) là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau.

Khung nhìn (View) là một mối quan hệ ảo được định nghĩa bằng cách sử dụng mối quan hệ thực được lưu trữ trong cơ sở dữ liệu.

Khung nhìn thực (Materialized view) là kết quả mối quan hệ truy vấn đã được lưu trữ trước. Có thể cho phép thực thi các truy vấn phức tạp trên các cơ sở dữ liệu với dung lượng hàng Terabytes trong vài giây hoặc phần nhỏ của giây.

1.2. Triển vọng của kho dữ liệu

Hầu hết các kho dữ liệu đang được dùng cho quản trị doanh nghiệp thông minh làm tăng mối quan hệ khách hàng (CRM - Customer Relationship Management) và khai thác dữ liệu. Một số được sử dụng để báo cáo tổng hợp, một số được sử dụng để tích hợp dữ liệu. Các cách sử dụng này đều tương quan với nhau.

Quản trị doanh nghiệp thông minh (BI)

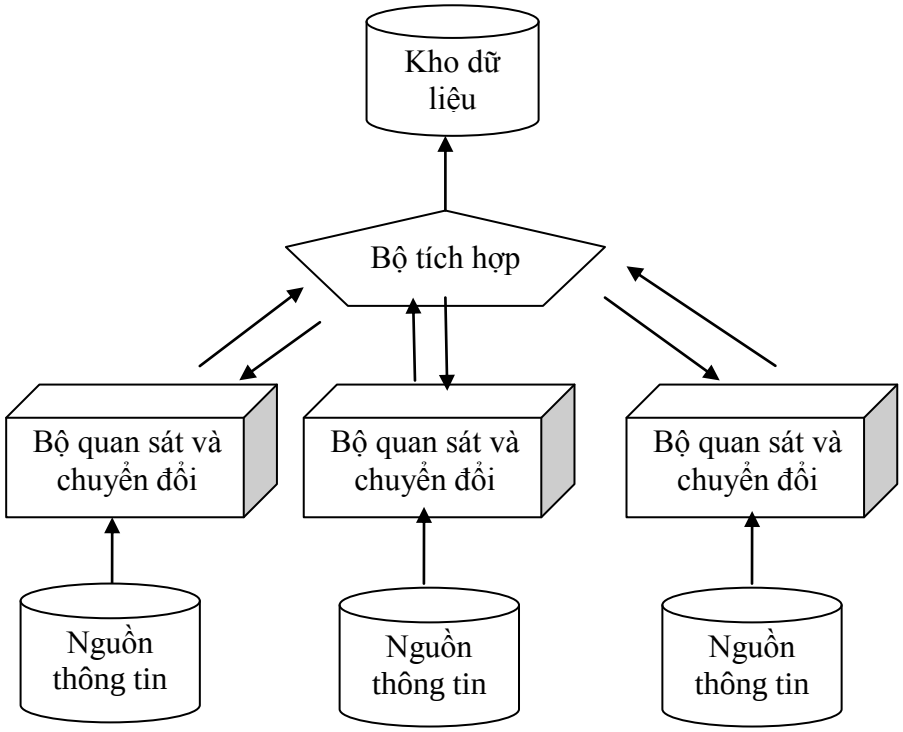
Quản lý mối quan hệ khách hàng (CRM)

Khai phá dữ liệu

Quản lý dữ liệu chủ

Tích hợp dữ liệu khách hàng

1.3. Kiến trúc hệ thống kho dữ liệu



Hình 1.1. Kiến trúc cơ bản của một hệ thống kho dữ liệu.

1.4. Các vấn đề nghiên cứu

1.4.1 Bộ chuyển đổi và giám sát

- **Chuyển đổi:**

Chuyển đổi nguồn thông tin thành mô hình dữ liệu được sử dụng bởi hệ thống kho dữ liệu. Ví dụ, nếu các nguồn thông tin bao gồm một tập hợp các tập tin flat, nhưng mô hình

kho dữ liệu là mô hình quan hệ, do đó Bộ chuyển đổi và giám sát phải hỗ trợ một giao diện để trình bày các dữ liệu nguồn thông tin theo kiểu quan hệ.

- **Quan sát sự thay đổi:**

Để phát hiện sự thay đổi của các dữ liệu nguồn có liên quan đến kho dữ liệu và chuyển những thay đổi này cho Bộ tích hợp. Chức năng này dựa trên bộ chuyển đổi, giống như các dữ liệu chính nó, thay đổi dữ liệu phải được chuyển các định dạng và mô hình của nguồn dữ liệu sang định dạng và mô hình được sử dụng trong hệ thống kho dữ liệu. Một cách khác chuyền bản sao toàn bộ dữ liệu có liên quan từ các nguồn dữ liệu đến kho dữ liệu. Bộ tích hợp có thể kết hợp dữ liệu này với các kho dữ liệu hiện có từ các nguồn khác, hoặc nó có thể yêu cầu thông tin đầy đủ từ tất cả các nguồn dữ liệu và tính lại kho dữ liệu từ đầu. Tuy nhiên phương pháp này đòi hỏi kho dữ liệu phải ngừng hoạt động trong từ khoảng thời gian và tình trạng dữ liệu không đáp ứng kịp thời.

1.4.2. Bộ tích hợp

Việc tiếp theo của Bộ tích hợp nhận được thông báo cập nhật từ Bộ giám sát đối với các nguồn thông tin và phản ánh những thay đổi trong các kho dữ liệu. Chức năng của Bộ tích hợp là bảo trì khung nhìn nơi mà chứa cơ sở dữ liệu tại các nguồn thông tin. Do vây công việc của Bộ tích hợp là thực hiện bảo trì khung nhìn, đó là sự kết nối chặt chẽ giữa bảo trì khung nhìn và kho dữ liệu.

Các nguồn thông tin cập nhật dữ liệu thường hoạt động độc lập với kho dữ liệu và các cơ sở dữ liệu không thể hoặc không muốn tham gia trong việc bảo trì khung nhìn. Hầu hết các kỹ thuật bảo trì dựa trên việc cập nhật cùng với bảo trì khung nhìn và việc thay đổi và cập nhât khung nhìn xảy ra trong cùng một giao dịch. Trong môi trường kho dữ liệu có một số trường hợp xảy ra:

- Hệ thống bảo trì khung nhìn (Bộ tích hợp) không gắn với các hệ thống xử lý cơ sở dữ liệu (các nguồn thông tin).

- Các nguồn thông tin không tham gia trong việc bảo trì khung nhìn, nhưng báo cáo những thay đổi.

- Để xác định khung nhìn và mối quan hệ thực có thể được lưu trữ tại nguồn cơ sở dữ liệu khác nhau ở tại nhiều nguồn khác nhau. Các nguồn này có thể thông báo cho kho dữ liệu khi có cập nhật xảy ra nhưng họ không thể xác định những dữ liệu nào là cần thiết để cập nhật các khung nhìn tại các kho dữ liệu.

Vì vậy họ chỉ có thể gửi dữ liệu cập nhật hiện tại hoặc cập nhật toàn bộ các mối quan hệ đến kho dữ liệu. Khi nhận được thông tin này, các kho dữ liệu có thể bổ sung một số nguồn dữ liệu để cập nhật khung nhìn. Sau đó, truyền một số truy vấn từ một số nguồn để yêu cầu bổ sung nguồn dữ liệu. Một số nguồn có thể cập nhật dữ liệu một lần trước khi họ yêu cầu truy vấn từ các kho dữ liệu. Vì vậy, họ sẽ gửi thêm dữ liệu sai vào kho dữ liệu, sau đó sử dụng dữ liệu không chính xác để tính toán các khung nhìn. Hiện tượng này gọi là phân tán bảo trì khung nhìn bất thường. Giải quyết vấn đề bảo trì khung nhìn trong kho dữ liệu phức tạp hơn các hệ thống cơ sở dữ liệu truyền thống.

1.5. Kết luận

Kho dữ liêu đang phát triển mạnh trong công nghệ cơ sở dữ liệu, chúng ta còn rất nhiều vấn đề cần nghiên cứu để giải quyết những khó khăn, đó là những vấn đề bảo trì tính nhất quán dữ liệu của khung nhìn trên kho dữ liệu mà không làm ngừng việc cập nhật dữ liệu. Trên thực tế, có các kỹ thuật bảo trì khung nhìn để giải quyết các vấn đề đó. Nhưng để lựa chọn, đánh giá khả năng của loại kỳ thuật này thì chúng ta phải xem xét. Đây cũng chính là vấn đề mà luận án này tập trung nghiên cứu. Đó là phân loại kỹ thuật bảo trì khung nhìn để đưa ra đề xuất và tiến hành so sánh các kỹ thuật này trong điều kiện sử dụng không gian và số lượng hàng truy cập bằng cách sử dụng điểm chuẩn TPC (The American Transaction processing performance council) cho các Hệ hỗ trợ truy vấn quyết định.

Chương 2: PHÂN LOẠI KỸ THUẬT BẢO TRÌ KHUNG NHÌN CỦA KHO DỮ LIỆU

2.1. Giới thiệu

2.2. Khái niệm

2.2.1. Khung nhìn (View):

Khung nhìn là một bảng tạm thời, có cấu trúc như một bảng. Khung nhìn không lưu trữ dữ liệu mà nó được tạo ra khi sử dụng, và là đối tượng thuộc cơ sở dữ liệu.

Khung nhìn được định nghĩa như sau:

$$\mathbf{V = \Pi_{proj}(\sigma_{cond}(r_1 \times r_2 \times \ldots \times r_n))} \qquad \text{Công thức (2.1)}$$

Trong đó:

- Proj: là tập hợp các tên thuộc tính
- Cond: là biểu thức logic
- $r_1 \times r_2 \times \ldots \times r_n$ là các quan hệ cơ sở dữ liệu

- **Biểu thức truy vấn**

Trong việc duy trì một khung nhìn về quan hệ r_1, r_2, …, r_n, thuật toán để tạo ra các truy vấn chứa một tập các số hạng mà mỗi số hạng có dạng:

$$\mathbf{T = \Pi_{proj}\,(\sigma_{cond}\,(\overline{r_1} \times \overline{r_2} \ldots \times \overline{r_n}))}$$

Trong đó $\overline{r_i}$ là r_i mối quan hệ hoặc bộ dữ liệu t_i cập nhật của r_i.

Một truy vấn có dạng tổng của các số hạng:

$$\mathbf{Q = \sum_i T_i}$$

2.2.2. Khung nhìn thực (Materialized view)

2.2.3. Bảo trì khung nhìn

Bảo trì khung nhìn là làm thế nào để duy trì khung nhìn thực mà họ có thể lưu giữ đáp ứng với các bộ dữ liệu được cập nhật của cơ sở dữ liệu trong các nguồn dữ liệu từ xa.

Có hai phương pháp bảo trì khung nhìn thực:

Phương pháp tính lại các khung nhìn dẫn đến lượng lưu trữ và chi phí bảo trì bổ sung tăng lên và đôi khi không thể thực hiện do hạn chế về không gian lưu trữ.

Phương pháp bảo trì lũy tiến các khung nhìn nguyên tắc bảo trì lũy tiến khung nhìn là nguồn dữ liệu thông báo những thay đổi của dữ liệu để tích hợp, sau đó tính toán những thay đổi tương ứng và thông báo cho cơ sở dữ liệu với những thay đổi tương ứng. Phương pháp bảo trì lũy tiến khung nhìn tối ưu hơn so với phương pháp tính lại khung nhìn.

Bảo trì khung nhìn có cơ chế tự duy trì.

Một thuật toán có thể xác định thêm thông tin, được gọi là khung nhìn hỗ trợ.

Khung nhìn hỗ trợ được lưu trữ trong kho dữ liệu để duy trì khung nhìn kiểu chọn – tham chiếu – kết nối (**SPJ – Select Project Join**) tức là khung nhìn thực dựa trên truy vấn chỉ chứa các phép chọn, chiếu, và nối mà không cần truy cập vào cơ sở dữ liệu tại nguồn dữ liệu.

Khung nhìn tự duy trì là khi một khung nhìn cùng với một tập hợp các khung nhìn hỗ trợ có thể được duy trì trong kho mà không cần truy cập vào cơ sở dữ liệu. Và cũng có một số khung nhìn không được cập nhật, nhiều thông tin hỗ trợ bắt buộc tự duy trì.

Định nghĩa 2.1 Tự duy trì (Self – Maintenance)

Xét một khung nhìn V được định nghĩa trên một tập các mối quan hệ nguồn R. Gọi δR là những thay đổi được tạo ra trong các mối quan hệ R để đáp ứng cho khung nhìn V được duy trì. Để tính toán được δV (những thay đổi của khung nhìn V) hạn chế sử dụng thêm thông tin. Nếu δV được tính bằng cách sử dụng khung nhìn thực V và tập hợp các thay đổi δR, sau đó khung nhìn V *tự duy trì*.

Cho trước khung nhìn V, chúng ta trình bày thuật toán xác định tập các khung nhìn hỗ trợ A sao cho sự kết hợp V và A là tự duy trì, có nghĩa là có thể được bảo trì căn cứ vào những thay đổi trên các mối quan hệ nguồn mà không cần truy cập vào bất kỳ dữ liệu nào khác. Một khung nhìn hỗ trợ $A_{Ri} \in A$ là một biểu thức có dạng

$$A_{Ri} = (\pi\sigma R_i) \ltimes A_{Rj1} \ltimes A_{Rj2} \ltimes \dots \ltimes A_{Rj2}$$

2.2.4. Cơ chế tự duy trì với khung nhìn SPJ

Khung nhìn định nghĩa bằng cách sử dụng hoạt động chọn và chiếu được gọi là khung nhìn SP (SP- Selection Projection). Còn khung nhìn định nghĩa bằng cách sử dụng hoạt động chọn, chiếu và kết nối gọi là khung nhìn SPJ. Khung nhìn định nghĩa bằng cách sử dụng hoạt động kết nối bên ngoài loại đặc biệt hữu ích cho khung nhìn gọi là khung nhìn OJ (OJ – Outer join).

2.2.4.1.Phép chèn (Insertions)

2.2.4.2. Phép xóa (Deletetions)

2.2.4.3. Phép cập nhật (Updates)

2.3. Phương pháp tính lại có cơ chế tự duy trì

Một lợi thế của các kỹ thuật của loại này là khung nhìn duy trì bất thường tránh tất cả các dữ liệu cần thiết có sẵn tại kho dữ liệu. Kho dữ liệu biết định nghĩa khung nhìn và những gì để làm với các khung nhìn để chúng được cập nhập. Nó giúp loại bỏ truy cập đến các mối quan hệ từ xa, và do đó, nó không cạnh tranh với các nguồn dữ liệu từ xa tài nguyên cục bộ. Các hoạt động của kho dữ liệu duy trì sau đó có thể được tách riêng hoàn toàn các hoạt động OLTP khác. Cho dù một nguồn dữ liệu từ xa có sẵn hay không sẽ không ảnh hưởng đến quá trình duy trì khung nhìn của kho dữ liệu. Tuy nhiên, để làm cho khung nhìn thực tự duy trì, thêm khung nhìn thực cung cấp thông tin cần thiết để cập nhập khung nhìn phải được lưu trữ. Thêm lượng lưu trữ và thời gian như vậy, cần để duy trì các khung nhìn bổ sung.

2.4. Phương pháp tính lại không có cơ chế tự duy trì

Phương pháp tiếp cận tính lại không tự duy trì là đơn giản nhất. Các vấn đề bất thường có thể tránh được một cách dễ dàng. Tuy nhiên, quá trình tính lại mất nhiều thời gian và tốn tài nguyên. Kho dữ liệu gửi các truy vấn trở lại các nguồn và chờ đợi câu trả lời để tính khung nhìn mới. Xử lý các truy vấn này tiêu hao các nguồn tài nguyên nội bộ. Nếu các nguồn không có sẵn, các kho dữ liệu sẽ không nhận được câu trả lời cần thiết.

2.5. Phương pháp bảo trì lũy tiến có cơ chế tự duy trì

Kho dữ liệu không bao giờ phải truy vấn các nguồn dữ liệu từ xa để lấy dữ liệu bổ sung. Các dữ liệu hoạt động cho bảo trì kho có thể tách riêng hoàn toàn các hoạt động khác như ứng dụng xử lý giao dịch trực tuyến (OLTP). Cho dù các nguồn dữ liệu từ xa có sẵn hay không sẽ không ảnh hưởng đến quy trình bảo trì khung nhìn thực trong các kho dữ liệu. Tuy nhiên, để làm cho các khung nhìn thực tự duy trì, khung nhìn hỗ trợ được lưu trong kho dữ liệu để cung cấp các thông tin bổ sung. Thêm lưu trữ và chi phí thời gian là cách để duy trì khung nhìn hỗ trợ. Làm thế nào để thiết kế khung nhìn thực tại các kho dữ liệu để thông tin chỉ cần được lưu trữ tại các kho dữ liệu là một vấn đề lớn.

2.6. Phương pháp bảo trì lũy tiến không có cơ chế tự duy trì

Thay vì mỗi lần khung nhìn tính lại từ đầu, chỉ một phần của kho dữ liệu thay đổi được tính. Tuy nhiên, khi cần thiết các kho dữ liệu muốn truy vấn các nguồn dữ liệu từ xa bởi vì các thông tin tại các kho dữ liệu không đủ để khung nhìn duy trì. Để tiếp cận phương pháp này có truy xuất cơ bản không hạn chế.

2.6.1. Truy xuất cơ bản không hạn chế

Có nhiều thuật toán sử dụng theo phương pháp này. Thuật toán Eager compensating Algorithm (ECA) là thuật toán điển hình. ECA là thuật toán bảo trì khung nhìn lũy tiến. Đó là một phương pháp để sửa các vấn đề bảo trì khung nhìn xảy ra do việc tách giữa cơ sở dữ liệu và quản lý bảo trì khung nhìn tại kho dữ liệu. Phương pháp này không dựa vào trạng thái của các thông tin cơ bản mà tiếp tục cập nhật/sửa đổi tại các nguồn. Và phương pháp này theo dõi các bộ dữ liệu cập nhật nhận được từ nguồn và sau đó lọc ra, bù bất kỳ thông tin sẽ lặp lại các kết quả truy vấn. Bằng cách trừ đi (hoặc thêm vào) kết quả biết rằng sẽ (không) có được truy vấn sau, nó sẽ tạo ra một kết quả cuối cùng chính xác cho khung nhìn.

Trong phương pháp này, các kho dữ liệu có thể phải gửi các truy vấn về nguồn và chờ đợi câu trả lời để tính các bản khung nhìn cập nhật. Vì vậy, phương pháp này có những hạn chế tương tự như phương pháp tiếp cận tính lại không tự duy trì. Việc tính các truy vấn này tiêu thụ các nguồn tài nguyên cục bộ từ xa, và sẽ làm chậm các hoạt động OLTP khác. Nếu các nguồn từ xa không có sẵn, các kho dữ liệu sẽ không nhận được câu trả lời cần.

2.6.2. Tự bảo trì kho dữ liệu tại thời gian chạy chương trình

Một kho dữ liệu gồm tập hợp các khung nhìn . Mỗi khung nhìn được xác định bởi truy vấn trên một số cơ sở dữ liệu D. Các định nghĩa khung nhìn có sẵn trong kho dữ liệu. Mẫu thông tin khác cũng có thể được cung cấp cho các kho dữ liệu, như cơ sở dữ liệu D thỏa mãn tính ràng buộc toàn vẹn. Ban đầu, các khung nhìn phù hợp với cơ sở dữ liệu D. Khi cơ sở dữ liệu D được sửa đổi, cơ sở dữ liệu cập nhật U gửi đến kho dữ liệu. Khung nhìn có thể trở nên không phù hợp với cơ sở dữ liệu mới U(D), Công việc chính của người quản lý kho dữ liệu là cập nhật các khung nhìn để sao cho phù hợp với cơ sở dữ liệu mới.

Để duy trì khung nhìn V từ bước bao gồm:

- A truy vấn Q mà xác định khung nhìn V
- Trường hợp V của khung nhìn riêng
- Cập nhật trường hợp U
- Các thông tin khác (I)
- Ý tưởng cơ bản của " Tự duy trì kho dữ liệu tại thời gian chạy chương trình" là các kho dữ liệu kiểm tra khả năng tư duy trì cho các khung nhìn. Nếu khung nhìn tự duy trì được, nó sẽ được duy trì bằng thông tin cập nhật của chính mình và biểu thức truy vấn xác định khung nhìn. Trong trường hợp này, phương pháp tự

duy trì thời gian thực hiện tương ứng phương pháp tự duy trì kho. Tuy nhiên, các kho dữ liệu không lưu trữ và duy trì bất kỳ khung nhìn hỗ trợ. Nếu khung nhìn không khả năng tự duy trì, thì kho dữ liệu phải truy vấn các quan hệ cần thiết từ nguồn dữ liệu từ xa đối để cập nhật khung nhìn. Trong trường hợp này, phương pháp này giống với truy nhập cơ bản không hạn chế.

Chương 3: PHÂN TÍCH HIỆU NĂNG CỦA KỸ THUẬT BẢO TRÌ KHUNG NHÌN CỦA KHO DỮ LIỆU

3.1. Giới thiệu

3.2. Đo hiệu năng

Trong phân tích, chỉ có khung nhìn SPJ được xem xét. Để đo hiệu năng của các kỹ thuật về không gian và số lượng truy cập hàng, ta sẽ căn cứ vào:

- Không gian: tổng số không gian cần thiết để lưu trữ các dữ liệu trong kho dữ liệu, bao gồm cả không gian đối với khung nhìn hỗ trợ. Trong phần này không xét chỉ số.

- Số hàng truy cập: số lượng hàng được truy cập vào kho dữ liệu và các nguồn dữ liệu để tích hợp bộ dữ liệu cập nhật vào kho dữ liệu.

3.3. Phân tích các tham số

Các thông số và giá trị mặc định đươc liệt kê trong bảng 3.1 được tính toán dựa vào tiêu chuẩn TPC cho các Hệ hỗ trợ truy vấn quyết định.

Ý nghĩa	Ký hiệu	Giá trị mặc định	Phạm vi
Số hàng của khung nhìn V	*Card(V)*	914	0~100.000
Kích thước bộ dữ liệu của khung nhìn V(tính bằng bytes)	*Ts(V)*	43	10~250
Số lượng của khung nhìn hỗ trợ mỗi lần xem	*Nav*	3	1~N
Số lượng của các mối quan hệ cơ sở trong định nghĩa khung nhìn	*N*	3	1~7
Số hàng của mối quan hệ cơ sở r	*Card(r)*	108.000.000	0~1.000.000.000
Kích thước bộ dữ liệu của mối quan hệ cơ sở r (tính bằng byte)	*Ts(r)*	116	100~180
Tính chọn lọc: phần nhỏ của bộ dữ liệu mà đáp ứng điều kiện lựa chọn	σ	0,003	0,00001~1,0
Tính kết nối là giá trị tương đương của bộ dữ liệu trong mối quan hệ liên kết các mối quan hệ khác	*j*	0,73	0,00001~1,0
Số lượng của bộ dữ liệu trong mối quan hệ liên kết các mối quan hệ khác	$J=j \times Card(r)$	Tính toán	Tính toán
Số lượng ảnh hưởng cập nhật cho mỗi truy vấn	*I*	0,5	0~100
Số hàng của cập nhật	*Card (U)*	1	
Số lượng bộ dữ liệu thêm vào trong mối quan hệ cơ sở dữ liệu nguồn	*Nupdate*	1	100

Bảng 3.1. Các thông số

3.4. Căn cứ vào không gian cần thiết trong kho dữ liệu để so sánh các kỹ thuật bảo trì

3.4.1. Phương pháp tính lại có cơ chế tự duy trì

Trong trường hợp thông thường, lượng không gian cần thiết là

$$Card(V)ts(V) + \sum_{i=0}^{Nav} Card(AV_i)ts(AV_i)$$

Trong đó 0 ≤ Nav ≤ N.

3.4.2. Phương pháp tính lại có không cơ chế tự duy trì

Trường hợp thông thường bằng Card(V) ts(V).

3.4.3. Phương pháp bảo trì lũy tiến có cơ chế tự duy trì

Trong trường hợp thông thường, không gian cần thiết như sau:

$$Card(V)ts(V) + \sum_{i=0}^{Nav} Card(AV_i)ts(AV_i)$$

Trong đó: 0 ≤ Nav ≤ N.

3.4.4. Phương pháp bảo trì lũy tiến không có cơ chế tự duy trì

Trong trường hợp bình thường, kích thước của COLLECT cho cập nhật cụ thể với tất cả các cập nhật mà nó can thiệp bằng tổng của số các câu trả lời truy vấn cuối cùng cho tất các các truy vấn. Tổng số truy vấn được gửi đến các nguồn dữ liệu có thể được tính như sau:

$$Nq = \sum_{k=1}^{N-1} \left(I^{k-1}\sigma^k j^{k-1} \sum_{i=0}^{N-k-1} (\sigma^i J^i) \right)$$

3.4.5. So sánh bốn kỹ thuật bảo trì khung nhìn

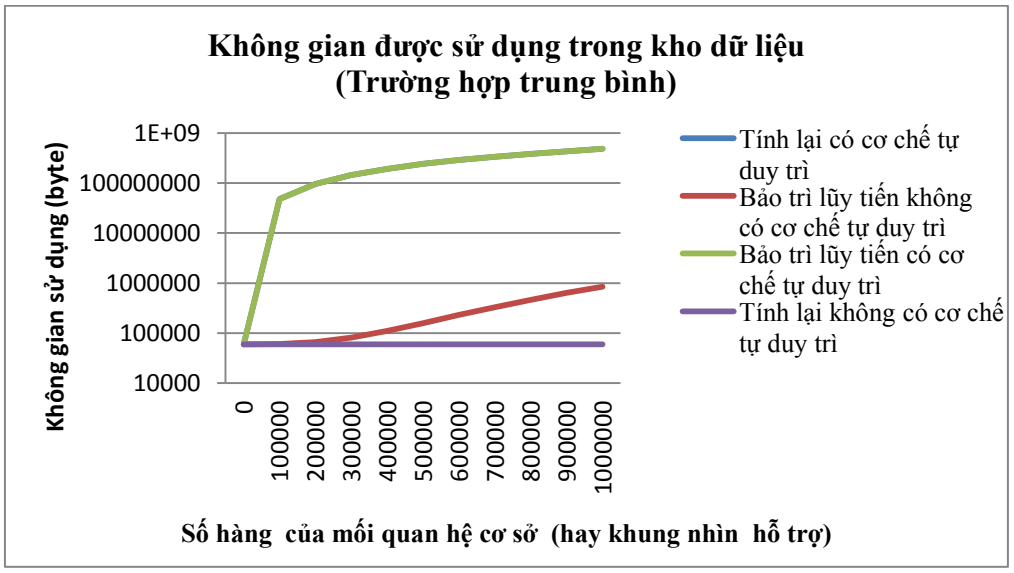
Để so sánh không gian cần thiết trong kho dữ liệu trong các kỹ thuật bảo trì dựa vào các thông số giá trị được liệt kê trong bảng 3.1 và công thức trong bảng 3.2. Ở đây, tôi chỉ xem xét các kết quả của trường hợp bình thường. Trong trường hợp này, phương pháp tính lại không có cơ chế tự duy trì không đòi hỏi không gian thêm tại kho dữ liệu. Tuy nhiên, trong phương pháp tính lại có cơ chế tự duy trì và bảo trì lũy tiến có cơ chế tự duy trì thì không gian thêm là cần tại các kho dữ liệu để lưu trữ các dữ liệu nguồn sao chép như khung

nhìn hỗ trợ. Không gian thêm tỷ lệ thuận với tổng số khung nhìn hỗ trợ *Nav*, bộ dữ liệu khung nhìn hỗ trợ *Card(AV)* và kích thước bộ dữ liệu khung nhìn hỗ trợ ts(AV). Kết quả thể hiện trong bảng 3.3

Số hàng của mối quan hệ cơ sở (hay khung nhìn hỗ trợ)	**SMR**	**SMIM**	**NSMR**	**NSIM**
0	59475	59475	59475	59475
100000	48059475	48059475	59475	60279
200000	96059475	96059475	59475	65839
300000	144059475	144059475	59475	80872
400000	192059475	192059475	59475	110098
500000	240059475	240059475	59475	158236
600000	288059475	288059475	59475	230005
700000	336059475	336059475	59475	330124
800000	384059475	384059475	59475	463312
900000	432059475	432059475	59475	634287
1000000	480059475	480059475	59475	847770

Bảng 3.3. So sánh không gian cần thiết trong kho dữ liệu

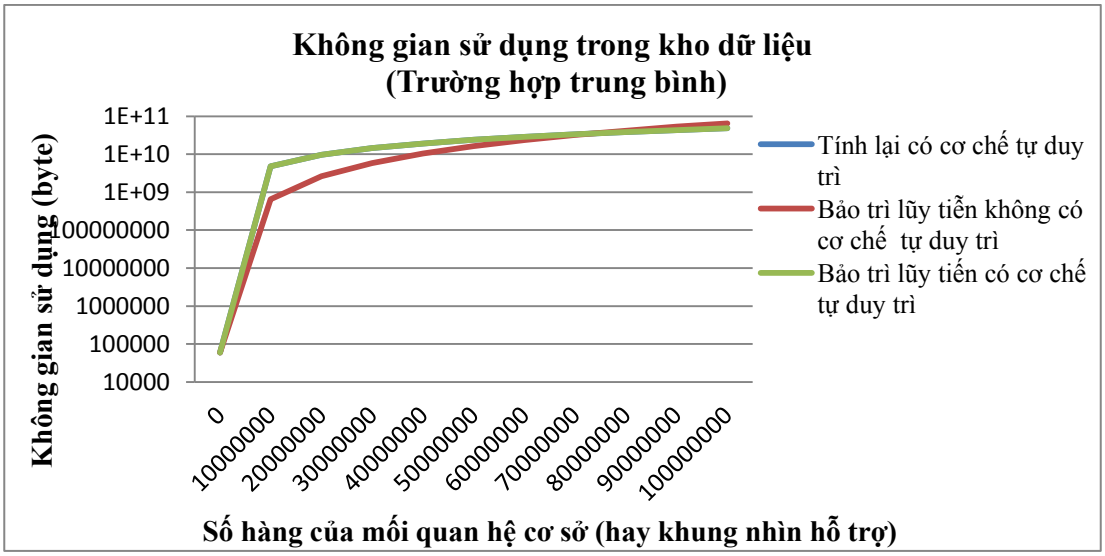
Từ bảng 3.2, cho thấy không gian cần thiết để lưu trữ trong phương pháp bảo trì lũy tiến không có cơ chế tự duy trì (NSIM) nhỏ so với không gian được sử dụng để lưu trữ các khung nhìn thực trong phương pháp tính lại có cơ chế tự duy trì (SMR) và bảo trì lũy tiến có cơ chế tự duy trì (SMIM). Đồ thị được thể hiện trong hình 3.1



Hình 3.1. Không gian được sử dụng trong các kho dữ liệu (Trường hợp trung bình)

Đối với phương pháp bảo trì lũy tiến không có cơ chế tự duy trì, không gian thêm là cần thiết để lưu trữ các bảng COLLECT.

Trong một số trường hợp khác, không gian sử dụng để lưu trữ các bảng COLLECT có thể phát triển hơn so với các khung nhìn thực. Lúc đầu, không gian lưu trữ của phương pháp bảo trì lũy tiến không có cơ chế tự duy trì (NSIM) nhỏ hơn và tăng chậm hơn so với không gian để lưu trữ của phương pháp tính lại có cơ chế tự duy trì (SMR) và bảo trì lũy tiến có cơ chế tự duy trì (SMIM). Nhưng sau đó không gian cần thiết để lưu trữ của phương pháp NSIM đã nhanh hơn, đến thời điểm nhất định, không gian được sử dụng để lưu trữ trong phương pháp NSIM vượt qua các phương pháp khác. Kết quả được thể hiện trong hình 3.2



Hình 3.2. Không gian được sử dụng trong các kho dữ liệu (Trường hợp trung bình)

Tóm lại, trong trường hợp trung bình, không có không gian cần thiết để bổ sung vào tại kho dữ liệu cơ chế không tự duy trì, đó là phương pháp tốt nhất về không gian được sử dụng trong các kho dữ liệu. Số lượng của không gian được sử dụng cho cả cơ chế tự duy trì là phương pháp tính lại và phương pháp bảo trì lũy tiến đều giống nhau. Khi có dữ liệu cập nhật và số hàng của mối quan hệ cơ sở là tương đối nhỏ thì không gian được sử dụng cho các cơ chế không tự duy trì là ít hơn so với cả hai cơ chế tự duy trì. Nhưng nếu nhiều bộ dữ liệu cập nhật hơn và số hàng của mối quan hệ cơ sở lớn thì không gian sử dụng cho cơ chế không tự duy trì lớn hơn so với hai cơ chế tự duy trì.

3.5. Căn cứ vào số hàng truy cập vào kho dữ liệu để so sách các kỹ thuật bảo trì

3.5.1. Phương pháp tính lại có cơ chế tự duy trì

3.5.2. Phương pháp tính lại không có cơ chế tự duy trì

3.5.3. Phương pháp bảo trì lũy tiến có cơ chế tự duy trì

3.5.4. Phương pháp bảo trì lũy tiến không có cơ chế tự duy trì

3.5.5. Căn cứ vào số lượng hàng để so sánh các kỹ thuật bảo trì khung nhìn

Loại	Số lượng hàng truy cập kho dữ liệu	Số lượng hàng truy cập nguồn dữ liệu
SMR	Trường hợp trung bình: Nupdate × (Card(V) + 2Card(AV) + $Card(U) + Card(AV)^2\left(\sum_{i=1}^{N-1}(\sigma^i J^{i-1})\right)$)	0
	Trường hợp xấu: Nupdate ×(Card(V) + Card(U)+ $Card(AV) \times \left(1 + \sum_{i=0}^{N-1}(Card(AV)^i)\right)$)	
NSMR	Nupdate × Card(V)	Trường hợp tốt:0
		Trường hợp trung bình: Nupdate × $(Card(r) + Card(r)^2 \times \left(\sum_{i=1}^{N-1}(\sigma^i J^{i-1})\right)$)
SMIM	Trường hợp trung bình: Nupdate × (Card(V) + Card(AV) + Card(ΔAV) x$\left(2 + Card(AV) \times \sum_{i=0}^{Nav-2}(\sigma^{i+1} J^i)\right)$ 0≤Nav≤N	

	Trường hợp xấu nhất: $\text{Nupdate} \times$ $(\text{Card(V)} + \text{Card(AV)} + \text{Card}(\Delta\text{AV})$ $\times \left(2 + \sum_{i=1}^{N-1} (Card(AV)^i\right))$	
NSMIM	Nupdate/I×Card(V)	Trường hợp tốt nhất: 0 Trường hợp trung bình: $Nupdate \times$ $\sum_{k=1}^{N-1} (I^{k-1}\sigma^k j^{k-1}\,\text{Card(U)}^{\text{k}} \sum_{i=0}^{N-k-1} (\sigma^i J^i)$ $\times (\text{Card(U)} + \sum_{i=0}^{k-2} (\sigma^{i+1} J^i\,\text{Card(U)}^{\text{i+2}})$ $+\sigma^k j^{k-1}\text{Card(U)}^{\text{k}}\, Card(r) \sum_{i=0}^{N-k-1} (\sigma^i J^i)))$
		Trường hợp tồi nhất: $\text{Nupdate} \times$ $\sum_{\text{k=1}}^{\text{N-1}} \left(\text{I}^{\text{k-1}}\text{Card(U)}^{\text{k}} \sum_{\text{i=0}}^{\text{N-k-1}} (\text{Card(r)}^{\text{i}}) \right.$ $\times \left(\text{Card(U)} + \sum_{\text{i=0}}^{\text{k-2}} (\text{Card(U)}^{\text{i+2}}) \right.$ $\left. \left. + \text{Card(U)}^{\text{k}}\text{Card(r)} \sum_{\text{i=0}}^{\text{N-k-1}} (\text{Card(r)}^{\text{i}}) \right) \right)$

Bảng 3.3. Số lượng hàng truy cập

Để kiểm tra việc thực hiện các kỹ thuật bảo trì khung nhìn bằng cách tính số lượng hàng truy cập của từng kỹ thuật bảo trì khung nhìn để bảo trì khung nhìn thực qua các công thức trong bảng 3.4 và sử dụng các giá trị trung bình được liệt kê trong bảng 3.1 để vẽ đồ thị.

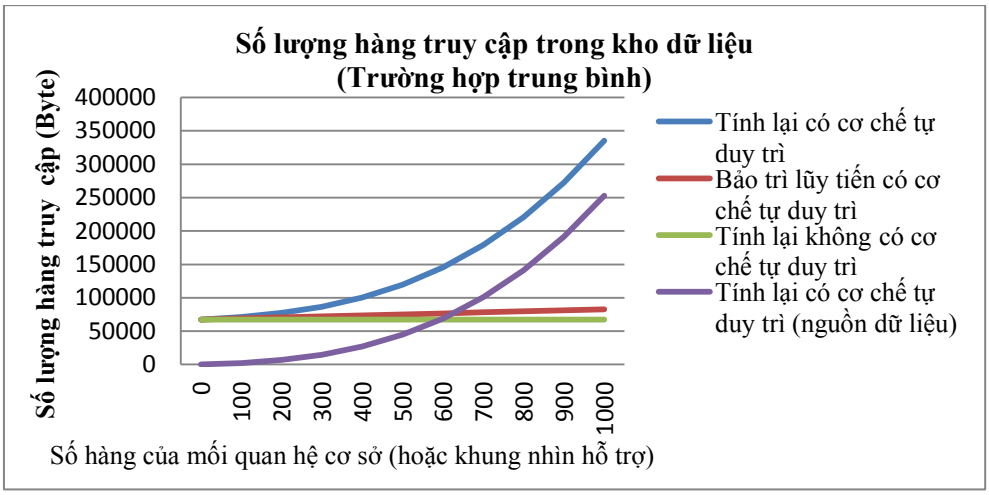
Trong trường hợp trung bình, tổng số lượng hàng truy cập của phương pháp tính lại không có cơ chế tự duy trì (NSMR) và phương pháp tính lại có cơ chế tự duy trì (SMR) phát triển gần tương đồng. Khi mối quan hệ cùng một cơ sở được nhân rộng tại các kho dữ liệu trong phương pháp tính lại không có cơ chế tự duy trì, thực tế những dữ liệu này sẽ được truy cập hai lần. Lần đầu tiên là bảo trì khung nhìn hỗ trợ tại nguồn dữ liệu và lần thứ hai là

bảo trì khung nhìn thực tại các kho dữ liệu. Khi số hàng của các mối quan hệ cơ sở (hay khung nhìn hỗ trợ) Card(r) là nhỏ thì số lượng hàng truy cập của phương pháp tính lại có cơ chế tự duy trì lớn hơn so với những phương pháp tính lại không có cơ chế tự duy trì. Các kết quả được trình bày trong bảng 3.5.

Số hàng của mối quan hệ cơ sở (hay khung nhìn hỗ trợ)	SMR	NSMR	SMIM	NSMR (nguồn dữ liệu)
0	67515	67500	67530	0
100	71293	67500	69038	2278
200	77336	67500	70549	6821
300	86710	67500	72064	14695
400	100481	67500	73582	26966
500	119715	67500	75104	44700
600	145477	67500	76630	68962
700	178832	67500	78159	100817
800	220846	67500	79692	141331
900	272585	67500	81228	191570
1000	335115	67500	82768	252600

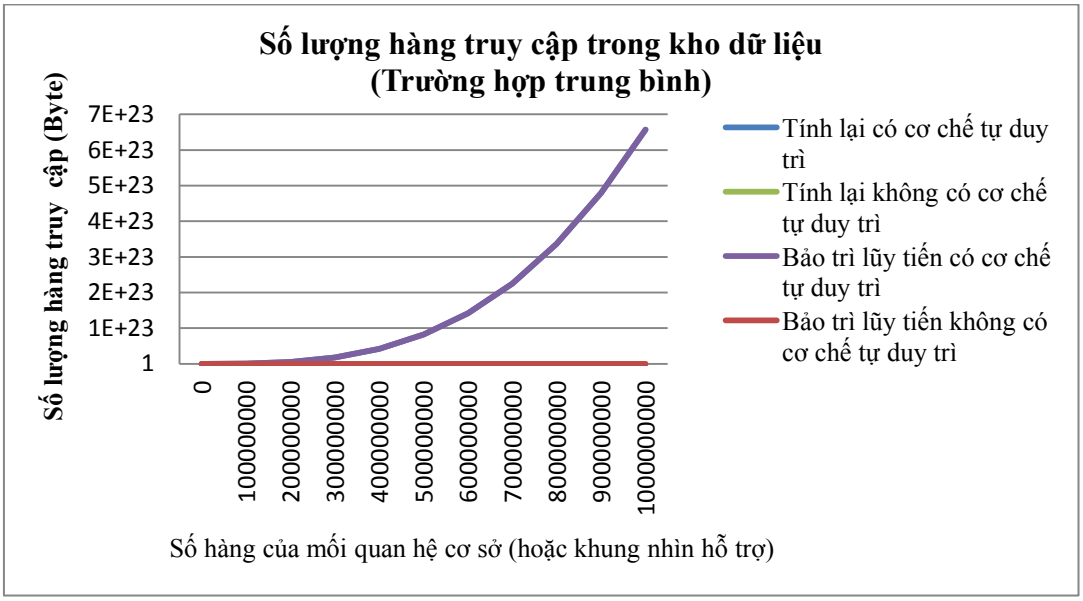
Bảng 3.5. So sánh tổng số hàng truy cập

Đồ thị sử dụng các giá trị bảng 3.5 được thể hiện ở hình 3.3



Hình 3.3. Số lượng hàng truy cập (Trường hợp trung bình)

Khi số hàng của mối quan hệ cơ sở tăng lên, tổng số lượng hàng truy cập của các kỹ thuật bảo trì có sự khác biệt không quá lớn nên có thể bỏ qua (hình 3.4)



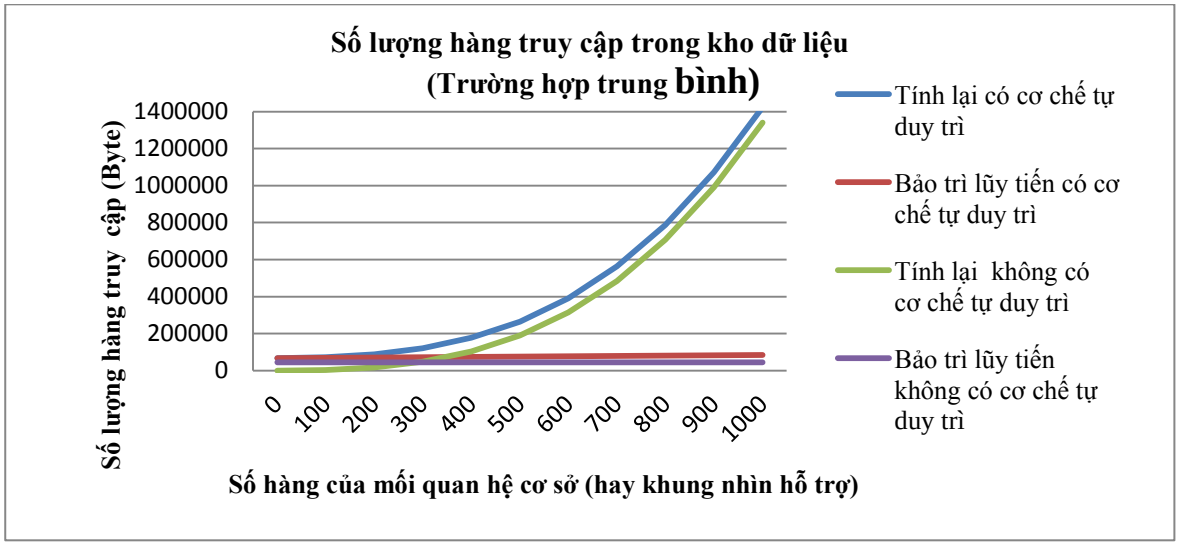
Hình 3.4 Số lượng hàng truy cập trong kho dữ liệu (Trường hợp trung bình)

Khi số hàng của mối quan hệ cơ sở nhỏ, tổng số lượng hàng đã truy cập để phương pháp tính lại có cơ chế tự duy trì và tính lại không có cơ chế tự duy trì tăng nhanh hơn so với phương pháp bảo trì lũy tiến cơ chế tự duy trì và bảo trì lũy tiến không có cơ chế tự duy trì được thể hiện ở bảng 3.6.

Số hàng của mối quan hệ cơ sở (hay khung nhìn hỗ trợ)	SMR	SMIM	NSMR	NSMIM
0	67515	67530	0	45000
100	73056	69055	4041	45000
200	88441	70605	17926	45000
300	120814	72178	48799	45000
400	177320	73775	103805	45000
500	265103	75395	190088	45000
600	391306	77040	314791	45000
700	563075	78708	485060	45000
800	787553	80400	708038	45000
900	1071885	82116	990870	45000
1000	1423215	83856	1340700	45000

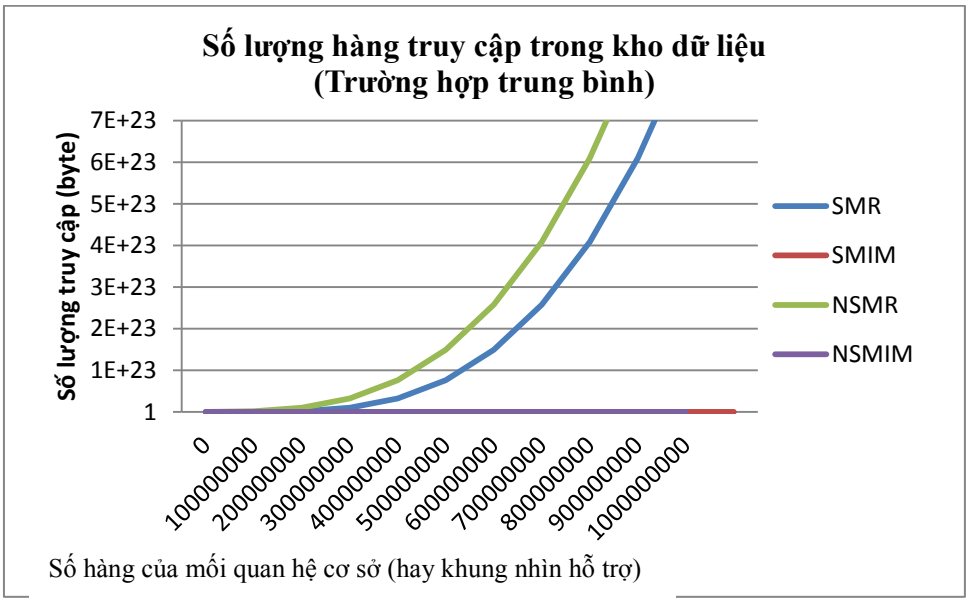
Bảng 3.6. So sánh tổng số hàng truy cập trong kho dữ liệu

Bảng đồ thị được thể hiện trong hình 3.5:



Hình 3.5. Số lượng hàng truy cập trong kho dữ liệu (Trường hợp trung bình)

Khi Card(r) trở nên rất lớn, số lượng hàng truy cập cho cả các phương pháp bảo trì lũy tiến nhỏ hơn so với hai phương pháp tính lại. (hình 3.6)



Hình 3.6. Số lượng hàng truy cập trong kho dữ liệu (Trường hợp trung bình)

Tóm lại, ở trường hợp trung bình, tổng số lượng hàng truy cập trong kho dữ liệu của phương pháp tính lại có cơ chế tự duy trì và tính lại không có cơ chế tự duy trì tăng nhanh hơn so với hai phương pháp của bảo trì lũy tiến.

Chương 4: KẾT LUẬN

Từ những phân tích ở chương 2, 3, chúng ta có bảng tổng hợp để so sánh những ưu điểm, nhược điểm của bốn phương pháp bảo trì khung nhìn:

Thể loại	Ưu điểm	Nhược điểm
SMR	- Hoạt động bảo trì khung nhìn của kho dữ liệu được tách riêng hoàn toàn từ các hoạt động OLTP; - Bất kỳ nguồn nào đều không ngăn quá trình bảo trì khung nhìn kho dữ liệu;	- Dữ liệu được nhân rộng tại kho dữ liệu - Cần thêm dữ liệu lưu trữ cho dữ liệu lặp lại. - Phải thực hiện và bảo trì quy trình truyền dữ liệu để đưa dữ liệu từ các nguồn dữ liệu đến kho dữ liệu.
NSMR	- Thực hiện đơn giản - Không có lặp lại dữ liệu tại kho dữ liệu - Dung lượng dữ liệu không có thêm cho dữ liệu lặp lại. - Được yêu cầu là không có quá trình truyền dữ liệu	-Không có sẵn nguồn để ngăn chăn quá trình bảo trì khung nhìn trong kho dữ liệu -Đánh giá truy vấn các nguồn dữ liệu tiêu thụ các nguồn tài nguyên cục bộ. -Hoạt động bảo trì khung nhìn không được tách ra khỏi các hoạt động OLTP.
SMIM	- Hoạt động bảo trì khung nhìn trong kho dữ liệu hoàn toàn tách khỏi hoạt động OLTP; - Không có nguồn nào mà không không ngăn chặn quá trình bảo trì trong kho dữ liệu.	- Dữ liệu được sao chép tại kho dữ liệu. - Cần thêm dung lượng dữ liệu cho dữ liệu lặp lại; - Thực hiện và bảo trì quá trình truyền dữ liệu.
NSMIM	- Không cần dữ liệu nhân rộng tại các kho dữ liệu	- Không có nguồn để chặn quá trình bảo trì khung nhìn kho dữ liệu

	- Không cần yêu cầu thêm về lưu trữ. - Quá trình truyền dữ liệu để truyền dữ liệu từ các nguồn kho dữ liệu là không cần thiết	- Đánh giá câu hỏi các nguồn dữ liệu làm tiêu hao tài nguyên cục bộ - Hoạt động bảo trì khung nhìn kho dữ liệu không tách khỏi các hoạt động OLTP. -Phải thiết kế quá trình bảo trì khung nhìn cẩn thận để tránh những vấn đề bất thường - Trong trường hợp xấu nhất, số lượng hàng truy cập là cao nhất.

Bảng 4.1. Ưu điểm và nhược điểm của các kỹ thuật bảo trì khung nhìn

Hai phương pháp tính lại có cơ chế tự duy trì và bảo trì lũy tiến có cơ chế tự duy trì hoàn toàn tách biệt hoạt động bảo trì khung nhìn từ các hoạt động OLTP. Vì vậy, hoạt động bảo trì khung nhìn sẽ không tiêu tốn tài nguyên nguồn dữ liệu cục bộ. Các hoạt động này chỉ tiêu hao tài nguyên của các kho dữ liệu. Thậm chí nếu các nguồn dữ liệu từ xa không có sẵn, quá trình bảo trì khung nhìn kho dữ liệu có thể tiếp tục chạy. Tuy nhiên, một phần hoặc tất cả các nguồn dữ liệu được nhân rộng tại kho dữ liệu để được thực hiện quá trình bảo trì khung nhìn có cơ chế tự duy trì. Quá trình truyền dữ liệu được thực hiện để truyền dữ liệu từ các nguồn dữ liệu từ xa đến kho dữ liệu. Các quá trình thiết kế, thực hiện và bảo trì tốn nhiều thời gian. Rất nhiều dữ liệu không cần thiết được lặp lại tại các kho dữ liệu. Tuy nhiên, đây là những kỹ thuật mà nhiều công ty lớn muốn thực hiện nếu họ muốn tách riêng khỏi hoạt động bảo trì khung nhìn của kho dữ liệu từ hoạt động OLTP của họ.

Hai phương pháp tính lại không có cơ chế tự duy trì và bảo trì lũy tiến không có cơ chế tự duy trì có một số nhược điểm phổ biến. Đó là các nguồn dữ liệu từ xa xử lý các truy vấn từ kho dữ liệu mà lại sử dụng tài nguyên của nguồn cục bộ làm hệ thống OLTP sẽ chậm. Một khi một nguồn dữ liệu không có sẵn, tại thời gian đó các nguồn dữ liệu sẽ không thể trả lời truy vấn được gửi từ kho dữ liệu. Nó sẽ chặn quá trình bảo trì khung nhìn của các kho dữ liệu. Phương pháp bảo trì lũy tiến không có cơ chế tự duy trì có một số nhược điểm phụ. Để tránh những vấn đề bất thường, quá trình bảo trì khung nhìn phải được thiết kế một cách cẩn thận. Nếu nhiều bộ dữ liệu cập nhật đưa vào các nguồn dữ liệu, các kho dữ liệu có nhiều truy vấn bù. Rất có thể các kho dữ liệu không thể có được kết quả cuối cùng. Nhưng

phương pháp này đạt hiệu quả cao về số lương hàng truy cập để truyền bộ dữ liệu cho các đối tượng bảo trì khung nhìn trong kho dữ liệu. Cả hai phương pháp này cũng có một số ưu điểm. Vì không có sao chép dữ liệu trong kho dữ liệu, không có quá trình chuyển dữ liệu đã được thực hiện và bảo trì. Không có không gian thêm cho lưu trữ dữ liệu sao chép. Cả hai kỹ thuật này tốt cho các doanh nghiệp vừa nhỏ có hệ thống cơ sở dữ liệu OLTP không quá phức tạp.

Trong số tất cả bốn loại, phương pháp bảo trì lũy tiến có cơ chế tự duy trì tốt nhất về không gian được sử dụng trong các kho dữ liệu và số lượng hàng truy cập để truyền một cập nhật tới khung nhìn thực trong kho dữ liệu. Khi chi phí lưu trữ dữ liệu ngày càng trở nên thấp, đây là phương pháp tốt nhất để thực hiện một kho dữ liệu.

Hiện nay ở Việt Nam, kho dữ liệu đã được sử dụng trong các doanh nghiệp viễn thông, bệnh viện, thương mại điện từ… Nhưng để khai thác được kho dữ liệu tối ưu, chính xác và hiệu quả trong quá trình xử lý dữ liệu mà chi phí thấp là một vấn đề phải xem xét và cân nhắc. Vì vậy, với những vấn đề được đưa ra ở trên, lựa chọn và áp dụng kỹ thuật bảo trì khung nhìn nào thích hợp cho từng doanh nghiệp sẽ giải quyết các vấn đề khó khăn cho các doanh nghiệp.