BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THỰC PHẨM TP. HCM

TIỂU LUẬN MÔN HỌC

ĐỀ TÀI:

NGHIÊN CỨU OLAP VÀ TRUY VẤN ĐA CHIỀU

GVHD:

LỚP:

NHÓM HV:

TP. HCM, tháng …/20….

MỤC LỤC

LỜI MỞ ĐẦU 3
Chương 1. TỔNG QUAN VỀ OLAP 4
1.1. Giới thiệu 4
1.1.1. Giải thích một số thuật ngữ 4
1.1.2. Giới thiệu OLAP 4
1.2. Đặc điểm của OLAP 5
1.2.1. Khung nhìn đa chiều 5
1.2.2. Tính trong suốt 5
1.2.3. Khả năng truy nhập được 5
Chương 2. KIẾN TRÚC KHỐI CỦA OLAP 6
2.1. Giới thiệu kiến trúc khối 6
2.2. Khối (Cube) 7
2.2.1. Xác định khối 7
2.2.2. Xử lý các khối 7
2.2.3. Khối ảo (Virtual Cube) 8
2.3. Chiều (Dimension) 8
2.3.1. Xác định các chiều 8
2.3.2. Chiều có phân cấp 8
2.3.3. Phân cấp chiều 8
2.3.4. Roll_up và Drill_down dựa trên phân cấp chiều 9
2.4. Các phương pháp lưu trữ dữ liệu (MOLAP, ROLAP, HOLAP) 9
2.4.1. MOLAP (Multidimensional OLAP) 9
2.4.2. ROLAP (Relational OLAP) 9
2.4.3. HOLAP (Hybrid OLAP) 10
Chương 3. TRIỂN KHAI OLAP TRONG SQL SERVER 11
3.1. Yêu cầu cài đặt 11
3.2. Các bước thực hiện 11
KẾT LUẬN 16
TÀI LIỆU THAM KHẢO 17

DANH MỤC HÌNH ẢNH

Hình 1.1. Kho dữ liệu và OLAP....4
Hình 1.2. Mô hình dữ liệu khối....5
Hình 2.1. Giản đồ khối hình sao....6
Hình 2.2.Giản đồ khối hình tuyết rơi....6
Hình 2.3. Mô hình dữ liệu nhiều chiều....7
Hình 3.1. Mô hình cơ sở dữ liệu....11
Hình 3.2. Tạo Analysis Services Project....11
Hình 3.3. Tạo bộ kết nối dữ liệu....12
Hình 3.4. Kết nối dữ liệu....12
Hình 3.5. Tạo Data Source View....12
Hình 3.6. Xác định nguồn dữ liệu cần lấy....13
Hình 3.7. Chọn các bảng cần phân tích....13
Hình 3.8. Các bảng Fact và Dimension....14
Hình 3.9. Tạo khối dữ liệu....14
Hình 3.10. Dò tìm Fact và Dimemtion Tables....15

LỜI MỞ ĐẦU

Các hoạt động sản xuất, kinh doanh hiện nay luôn cần có sự đáp ứng nhanh nhạy, tức thời đối với các thay đổi liên tục, vì vậy các nhà quản lý buộc phải thường xuyên ra cùng lúc nhiều quyết định đúng đắn (mà chúng sẽ ảnh hưởng đáng kể đến xu hướng hoạt động và sự cạnh tranh của doanh nghiệp) một cách nhanh chóng. Do đó vấn đề trợ giúp quyết định trở nên rất cần thiết. Người ta cần phải thu thập, tổng hợp và phân tích dữ liệu từ nhiều nguồn khác nhau một cách nhanh và hiệu quả thì mới có thể ra được những quyết định nhanh chóng và phù hợp. Điều này dẫn đến việc cần phát triển những hệ thống tinh thông biết cách làm thế nào để trích chọn và phân tích dữ liệu cho người sử dụng.

Hiện nay có rất nhiều phần mềm cung cấp cho người sử dụng những khả năng truy vấn và lập các báo cáo thông tin, đặc biệt là các hệ quản trị CSDL quan hệ. Tuy nhiên CSDL quan hệ với cấu trúc hai chiều (dòng và cột) không được thiết kế để cung cấp các quan điểm đa chiều trên dữ liệu đầu vào của các phân tích phức tạp. Sử dụng các hệ thống này, chúng ta sẽ gặp rất nhiều khó khăn và bất tiện trong việc tổ chức dữ liệu đa chiều vào các bảng hai chiều, không thể triển khai dữ liệu phân tích với số lượng lớn, công cụ phân tích để tạo ra các dữ liệu quyết định không mạnh, thuận tiện, linh hoạt, nhanh chóng và nhất là không dễ dàng để sử dụng đối với các nhà quản lý, những người ra quyết định.

Như vậy, việc xây dựng một hệ thống mới có khả năng tổ chức dữ liệu đa chiều và có khả năng phân tích dữ liệu linh hoạt để trả lời được các truy vấn đa chiều một cách dễ dàng, nhanh chóng nhằm hỗ trợ cho việc ra quyết định của các nhà quản lý là cần thiết.

Chương 1. TỔNG QUAN VỀ OLAP

1.1. Giới thiệu

1.1.1. Giải thích một số thuật ngữ

Data Warehouse (DW) - Kho dữ liệu: Được xem là tập các cơ sở dữ liệu hướng chủ đề, có tính lịch sử được tích hợp từ nhiều nguồn dữ liệu qua các quá trình trích lọc, hợp nhất, chuyển đổi, làm sạch.

Dữ liệu khối (Data Cube): Dữ liệu trong kho dữ liệu được thể hiện dưới dạng đa chiều (Multi Dimension) gọi là khối (cube). Mỗi chiều mô tả một đặc trưng nào đó của dữ liệu. Ví dụ với Data Cube bán hàng thì chiều hàng hóa (Item) mô tả chi tiết về hàng hóa, chiều thời gian (time) mô tả về thời gian bán hàng, chiều chi nhánh (Branch) mô tả thông tin về các đại lý bán hàng

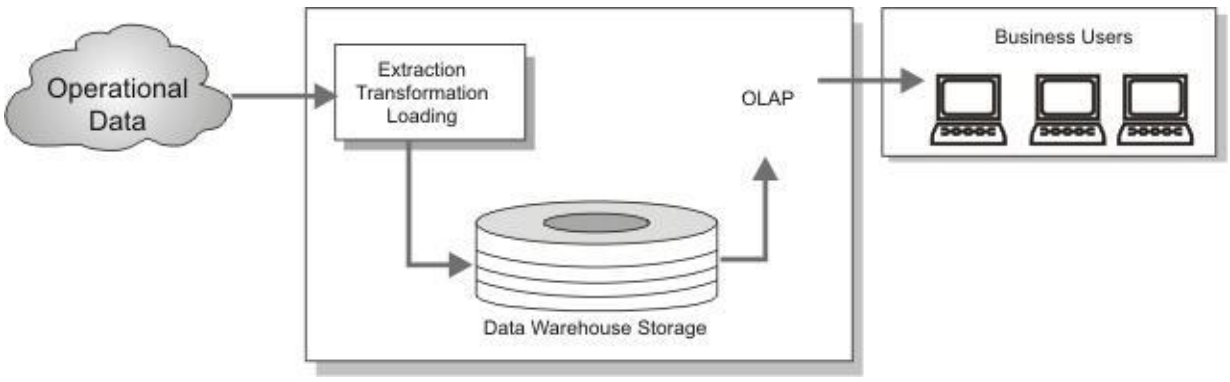
Measure (độ đo): Là đại lượng có thể tính toán được trên các thuộc tính của fact table. Đây là mục tiêu của OLAP và phải xác định trước khi tiến hành phân tích. Ví dụ như tổng tiền bán hàng của một chi nhánh, doanh thu của từng mặt hàng theo quí,…

Phân cấp (Hierarchies): Khái niệm này mô tả sự phân cấp thứ bậc (mức độ chi tiết của dữ liệu). Ví dụ đối với chiều thời gian, ta có thực bậc như sau: day<week<month<quarter<year. Tương tự đối với chiều location ta có thứ bậc street<city<province_or_state<country. Trong khi phân tích dữ liệu chúng ta rất cần khái niệm này để tổng hợp hay chi tiết từng hạng mục dữ liệu trong DW.

OLTP (On-Line Transaction Processing): Xử lý giao dịch trực tuyến

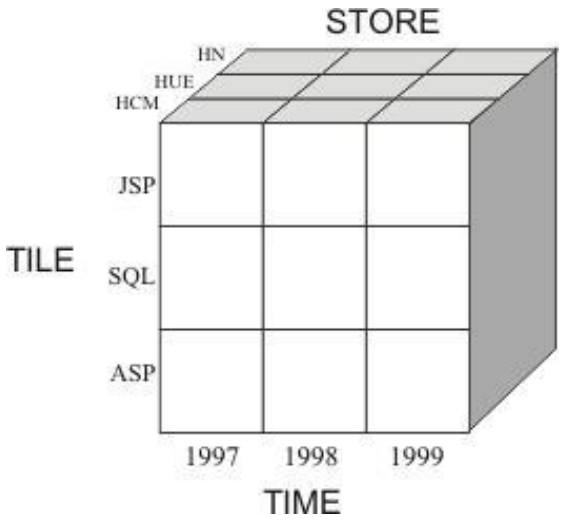
1.1.2. Giới thiệu OLAP

OLAP (***On Line Analysis Processing***) - Xử lý phân tích trực tuyến là một hệ thống xử lý dữ liệu mạnh. Nó cho phép người sử dụng phân tích dữ liệu qua việc cắt lát (slice) dữ liệu theo nhiều khía cạnh khác nhau, khoan xuống (Drill-Down) mức chi tiết hơn hay cuộn lên (Roll-Up) mức tổng hợp hơn của dữ liệu. Bản chất cốt lõi của OLAP là dữ liệu được lấy ra từ kho dữ liệu sau đó được chuyển thành mô hình đa chiều và được lưu trữ trong một kho dữ liệu đa chiều.



Hình 1.1. Kho dữ liệu và OLAP

Đối tượng chính của OLAP là khối (Cube), một sự biểu diễn đa chiều của dữ liệu chi tiết và tổng thể. Một khối bao gồm một bảng sự kiện (Fact), một hoặc nhiều bảng chiều (Dimensions), các đơn vị đo (Measures) và các phân hoạch (Partitions).



Hình 1.2. Mô hình dữ liệu khối

1.2. Đặc điểm của OLAP

1.2.1. Khung nhìn đa chiều

Đối với người thực hiện thì cách nhìn của họ với công việc là nhiều chiều về bản chất. Vì vậy mô hình OLAP phải là đa chiều về bản chất. Những người sử dụng có thể thao tác dễ dàng trên những mô hình dữ liệu đa chiều như vậy.

1.2.2. Tính trong suốt

Công cụ phân tích cần phải trong suốt với người sử dụng. OLAP nên tồn tại trong một kiến trúc hệ thống mở, cho phép các công cụ phân tích có thể được nhúng vào bất kỳ nơi nào mà người sử dụng mong muốn mà không có một sự tác động ngược lại nào với các chức năng của công cụ trên máy chủ.

1.2.3. Khả năng truy nhập được

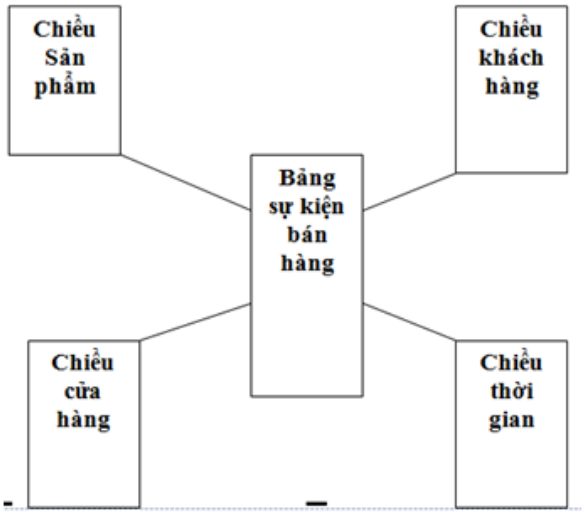
Công cụ OLAP phải ánh xạ được giản đồ Logic của chính nó tới kho dữ liệu vật lý hỗn tạp, truy nhập tới dữ liệu và thực hiện mọi chuyển đổi cần thiết để đưa ra một khung nhìn đơn giản, mạch lạc và đồng nhất cho người sử dụng. Dữ liệu vật lý của hệ thống thuộc kiểu này trở nên trong suốt với người sử dụng và chỉ là mối quan tâm của công cụ.

Chương 2. KIẾN TRÚC KHỐI CỦA OLAP

2.1. Giới thiệu kiến trúc khối

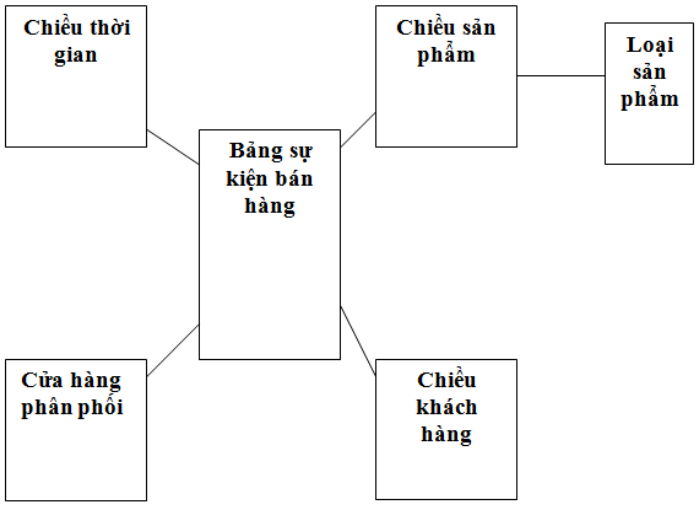
Để mô tả dữ liệu hình khối, chúng ta thử tưởng tượng dữ liệu trong bảng Fact được phân bố như sau: Đối tượng chính của OLAP là khối, một sự biểu diễn đa chiều của dữ liệu chi tiết và tổng thể. Một khối bao gồm một bảng sự kiện (Fact), một hoặc nhiều bảng chiều (Dimensions), các đơn vị đo (Measures) và các phân hoạch (Partitions). Ta có thể thiết kế các khối dựa trên cơ sở các yêu cầu phân tích của người sử dụng. Một kho dữ liệu có thể hỗ trợ nhiều khối khác nhau: khối về lương, khối về hàng tồn kho...

Ví dụ một giản đồ khối hình sao có dạng như sau:



Hình 2.1. Giản đồ khối hình sao

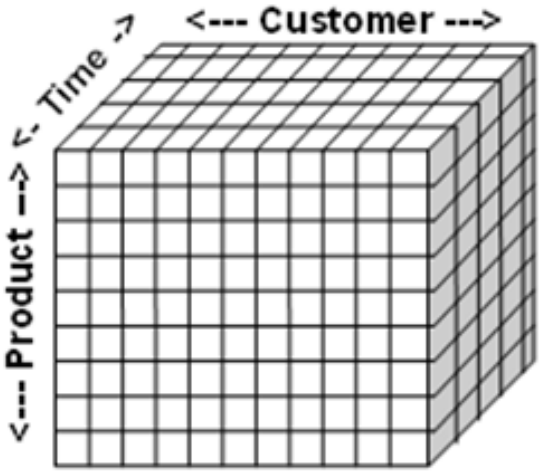
Ở đây nếu muốn, ta có thể mở rộng khối theo nhiều năm bằng cách thêm cột 'Year_ID' vào 'Time_Dimension_Table' và tạo thêm một bảng Dimension là 'Time_Dimension_Table_2' chứa hai cột 'Year_ID' và 'Year' . Lúc này ta có được một giản đồ khối hình tuyết rơi như sau:



Hình 2.2.Giản đồ khối hình tuyết rơi

2.2. Khối (Cube)

Khối là phần tử chính trong xử lý phân tích trực tuyến, một công nghệ cung cấp sự truy cập nhanh tới dữ liệu trong kho dữ liệu. Các khối cung cấp cơ chế truy vấn dữ liệu với thời gian trả lời nhanh và không phụ thuộc vào số lượng dữ liệu trong khối hoặc sự phức tạp của truy vấn.



Hình 2.3. Mô hình dữ liệu nhiều chiều

2.2.1. Xác định khối

Xác định khối là bước đầu tiên trong ba bước tạo khối. Các bước khác là các bước chỉ ra kế hoạch tóm tắt bằng việc thiết kế các khối tập hợp (các thành phần dữ liệu được tính toán trước) và Load khối bằng việc xử lý nó.

Để xác định một khối, ta chọn một bảng Fact và các đơn vị đo lường đồng nhất (các cột số theo sự quan tâm của người dùng khối) trong bảng Fact. Sau đó chọn các chiều, mỗi chiều gồm một hay nhiều cột từ bảng liên quan khác. Các chiều cung cấp mô tả rõ ràng bởi các đơn vị đo lường được chia ra của người dùng khối.

2.2.2. Xử lý các khối

Khi ta xử lý một khối thì các khối liên kết đã thiết kế của nó được tính toán và được Load cùng với khối và dữ liệu. Quá trình xử lý một khối bao gồm việc đọc các bảng Dimentions để xác định các cấp độ dữ liệu hiện tại, đọc bảng Fact, tính toán các liên kết đặc biệt và lưu trữ các kết quả trong khối. Sau khi một khối được xử lý, nó được cung cấp cho yêu cầu của người dùng.

Xử lý là thuật ngữ được dùng chỉ sự tải trọn vẹn dữ liệu của khối. Tất cả các chiều, dữ liệu bảng Fact được đọc và tất cả các khối liên kết đặc biệt được tính toán. Ta phải xử lý một khối khi cấu trúc của nó còn mới hoặc các chiều của nó hay các đơn vị đo lường đã được chọn lọc. Việc xử lý một khối có thể lấy đi một số thời gian thực nếu có một bảng Fact lớn, có nhiều chiều với nhiều cấp độ và nhiều khoản mục trong mỗi cấp độ.

Việc tải thông tin chiều là không cần thiết nếu ta dùng các chiều dùng chung đã được xử lý trong các khối.

2.2.3. Khối ảo (Virtual Cube)

Ta có thể liên kết các khối trong khối ảo giống như các bảng có thể được liên kết với các khung nhìn trong một cơ sở dữ liệu quan hệ. Một khối ảo cung cấp truy cập tới dữ liệu trong các khối kết hợp mà không đòi hỏi xây dựng một khối mới, nó cho phép ta duy trì thiết kế tốt nhất cho mỗi khối riêng biệt.

2.3. Chiều (Dimension)

Các chiều là cách mô tả chủng loại mà theo đó các dữ liệu số trong khối được phân chia để phân tích. Một chiều có thể được dùng bởi nhiều khối khác và được gọi là một chiều dùng chung. Nói chung, các khối cần chia xẻ một hay nhiều hơn các chiều.

Các chiều chia sẻ có thể được dùng trong bất cứ khối nào của cơ sở dữ liệu. Bằng việc tạo ra các chiều chia sẻ và dùng chúng trong đa khối, ta tránh được việc tạo ra các chiều cục bộ giống hệt nhau trong mỗi chiều thuộc các khối.

2.3.1. Xác định các chiều

Khi xác định một chiều, ta chọn một hoặc nhiều cột của một trong các bảng liên kết (bảng chiều). Nếu ta chọn các cột phức tạp thì tất cả cần có quan hệ với nhau, chẳng hạn các giá trị của chúng có thể được tổ chức theo hệ thống phân cấp đơn. Để xác định hệ thống phân cấp, sắp xếp các cột từ chung nhất tới cụ thể nhất.

2.3.2. Chiều có phân cấp

Phân cấp là cột sống của việc gộp dữ liệu hay nói một cách khác là dựa vào các phân cấp mà việc gộp dữ liệu mới có thể thực hiện được. Phần lớn các chiều đều có một cấu trúc đa mức hay phân cấp. Nếu chúng ta làm những quyết định về giá sản phẩm để tối đa doanh thu thì chúng ta cần quan sát ở những dữ liệu về doanh thu sản phẩm được gộp theo giá sản phẩm, tức là chúng ta đã thực hiện một cách gộp. Khi cần làm những quyết định khác thì chúng ta cần thực hiện những phép gộp tương ứng khác. Như vậy có thể có quá nhiều tiến trình gộp nên các tiến trình gộp này cần phải được thực hiện một cách rất dễ dàng, linh hoạt để có thể hỗ trợ những phân tích không hoạch định trước. Điều này có thể được giải quyết trên cơ sở có sự trợ giúp của những phân cấp rộng và sâu.

2.3.3. Phân cấp chiều

Các tham chiếu đến các phần tử trong các ứng dụng đa chiều thường liên quan đến một vài phần tử khác. Tham chiếu liên quan trong một cấu trúc phân cấp thì phức tạp hơn tham chiếu liên quan trong cấu trúc dòng và cột.

Cấu trúc phân cấp thường quan tâm đến hướng mà chúng ta đếm.

Phân cấp chiều như trên gọi là phân cấp bất đối xứng. Phân cấp như trong hình sau gọi là phân cấp đối xứng:

Trong phân cấp đối xứng chúng ta có thể tham khảo đến các phần tử theo mức của nó. Như vậy các 'Quý' là một tập hợp các phần tử một mức từ dưới lên và một mức từ trên xuống.

2.3.4. Roll_up và Drill_down dựa trên phân cấp chiều

Dựa trên phân cấp theo chiều, từ một mức dưới chúng ta có thể cuộn lên (Roll_up) các mức trên, thực hiện một phép gộp để có được kết quả tổng hợp hơn và từ một mức trên có thể khoan sâu xuống (Drill_down) các mức dưới để có các kết quả chi tiết hơn .

2.4. Các phương pháp lưu trữ dữ liệu (MOLAP, ROLAP, HOLAP)

2.4.1. MOLAP (Multidimensional OLAP)

Dữ liệu cơ bản của khối được lưu trữ cùng với dữ liệu kết hợp (Aggregation) trong cấu trúc đa chiều hiệu suất cao. Cách tiếp cận này kết hợp kho dữ liệu đa chiều và các dịch vụ của OLAP trên cùng một Server.

MOLAP là một cấu trúc tối ưu cho việc lưu trữ các sự kiện đã phân loại và cùng với nó là các chiều. Dữ liệu được tổ chức theo khung nhìn dữ liệu và được lưu trữ trong một biểu mẫu được kết hợp và tổng hợp. Tệp Index nhỏ hơn khiến cho việc trả lời những truy vấn phức tạp rất nhanh. Vì dữ liệu được lưu trữ trong các mảng, việc cập nhật các giá trị không ảnh hưởng nhiều tới tệp chỉ số. Điều này khiến cho việc cài đặt những ứng dụng cập nhật hoặc đọc-ghi như dự báo và điều chỉnh trở nên dễ dàng.

MOLAP là sự lựa chọn tốt nhất cho những ứng dụng có đặc điểm:

- Yêu cầu tốc độ truy vấn cao.
- Có khả năng phân tích dữ liệu phức hợp. MOLAP cung cấp môi trường phân tích mạnh hơn ROLAP.
- Dễ sử dụng: bởi dữ liệu đã được tổng hợp từ trước và được lưu trong kho dữ liệu đa chiều. Tất cả những gì người sử dụng cần làm là xác định các chiều và các nhóm nằm trong các chiều đó.

2.4.2. ROLAP (Relational OLAP)

Dữ liệu cơ bản của khối được lưu trữ cùng với dữ liệu kết hợp (Aggregation) trong cơ sở dữ liệu quan hệ. Phương pháp tiếp cận này bao gồm các dịch vụ của OLAP và cơ sở dữ liệu quan hệ. Các dữ liệu được lưu trữ trong những bảng quan hệ và có thể có kích thước hàng trăm Gigabyte. Những hệ ROLAP cung cấp các Engine truy vấn cực kỳ linh động bằng việc "chuẩn bị sẵn sàng" tất cả dữ liệu tác nghiệp cho người sử dụng đầu cuối, dễ dàng trích và tổng hợp dữ liệu theo yêu cầu. Những công cụ ROLAP có thể trích dữ liệu từ rất nhiều nguồn CSDL quan hệ khác nhau.

2.4.3. HOLAP (Hybrid OLAP)

Là kết hợp hai phương pháp MOLAP và ROLAP. Dữ liệu cơ bản của khối được lưu trữ trong cơ sở dữ liệu quan hệ và dữ liệu kết hợp (Aggregation) được lưu trữ trong cấu trúc đa chiều hiệu suất cao. Lưu trữ HOLAP đưa ra những lợi ích của MOLAP cho việc liên kết mà không cần thiết một bản sao chính xác từ dữ liệu chi tiết.

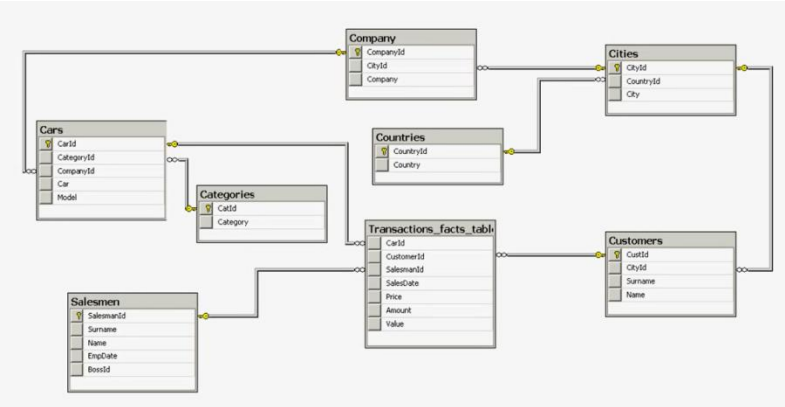
Chương 3. TRIỂN KHAI OLAP TRONG SQL SERVER

3.1. Yêu cầu cài đặt

SQL Server 2008 phiên bản Developer hoặc phiên bản Enterprise Edition đầy đủ và khi cài đặt phải chọn mục "SQLServer Database Services" và "Analysis Services". Công cụ cho phép thực hiện OLAP là "SQL Server Business Intelligence Development Studio - BIDS". Khi cài SQL Server các phiên bản trên thì BIDS sẻ được tự động cài đặt.

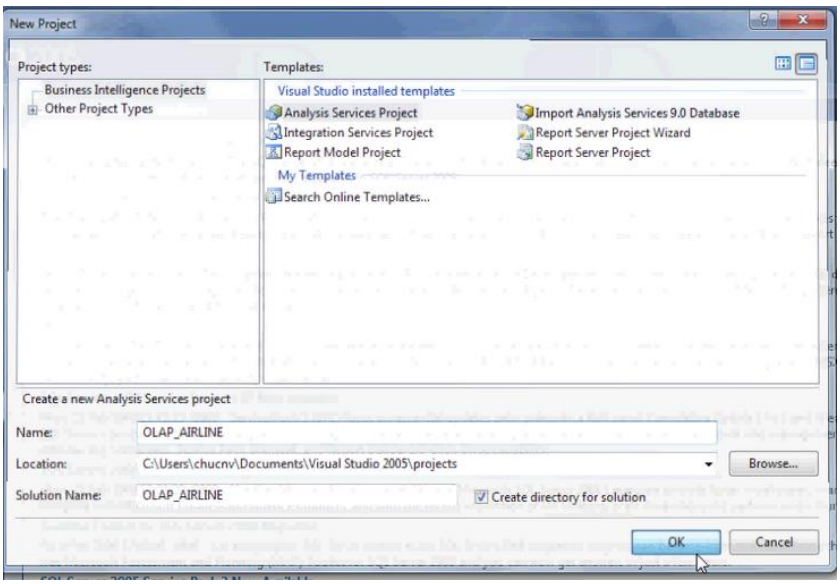
3.2. Các bước thực hiện

Khởi động SQL Server Management Studio và tạo CSDL, nhập vào các bảng một số records để phân tích.



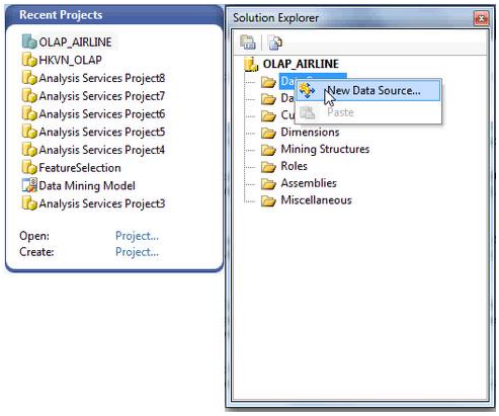
Hình 3.1. Mô hình cơ sở dữ liệu

Khởi động SQL Server Business Intelligence Development Studio, tạo một Analysis Services Project mới có tên "OLAP_AIRLINE"



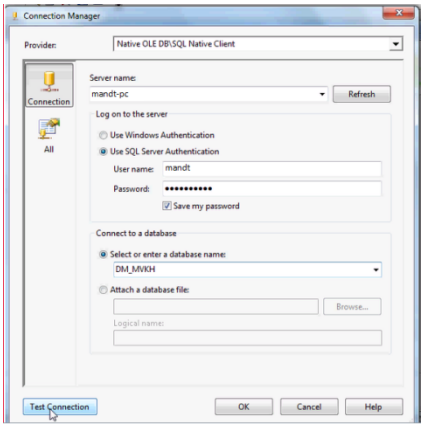
Hình 3.2. Tạo Analysis Services Project

Trong của sổ Solution Explorer của Project OLAP_DW, bấm phím phải chuột vào Data Source để tạo một bộ kết nối đến dữ liệu dùng cho phân tích.



Hình 3.3. Tạo bộ kết nối dữ liệu

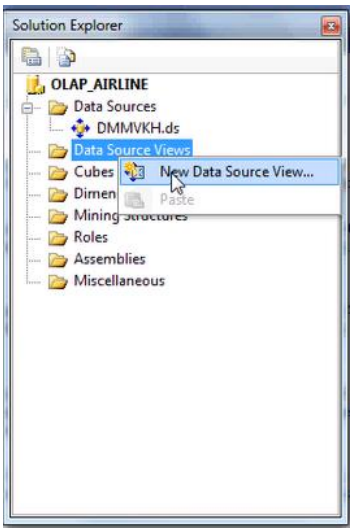
Xác định các tham số kết nối đến kho dữ liệu đã tạo ra trong SQL Server Management Studio.



Hình 3.4. Kết nối dữ liệu

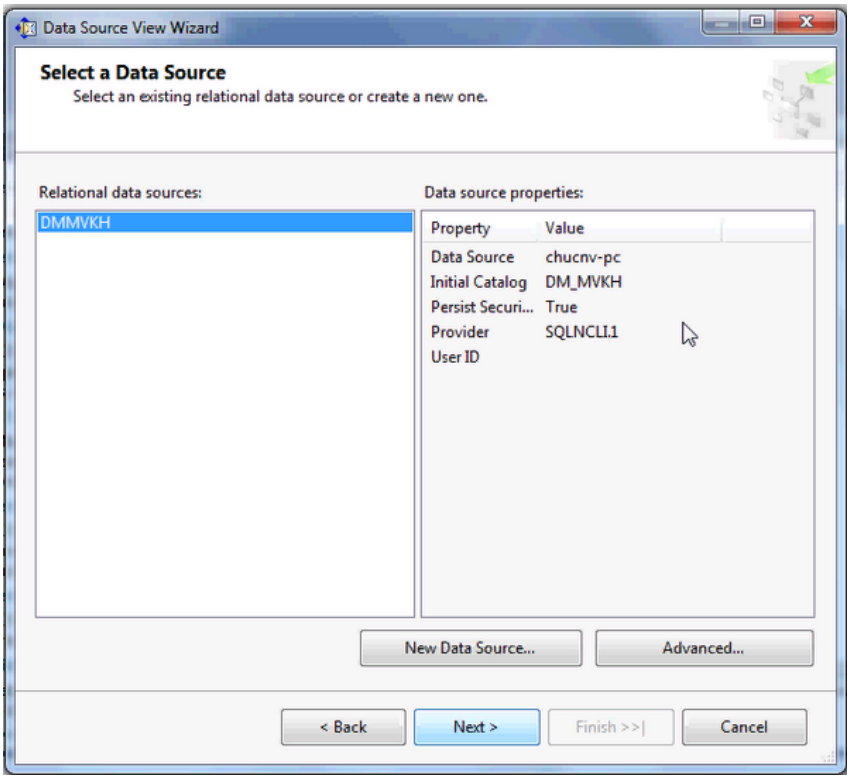
Đặt tên cho Data Source vàm bấm Finish để hoàn thành việc kết nối đến cơ sở dữ liệu.

Tạo Data Source View để lấy các bảng dữ liệu cần thiết cần cho phân tích. Bấm phím phải chuột vào Data Source View trong của sổ Solution Explorer chọn New Data Source View



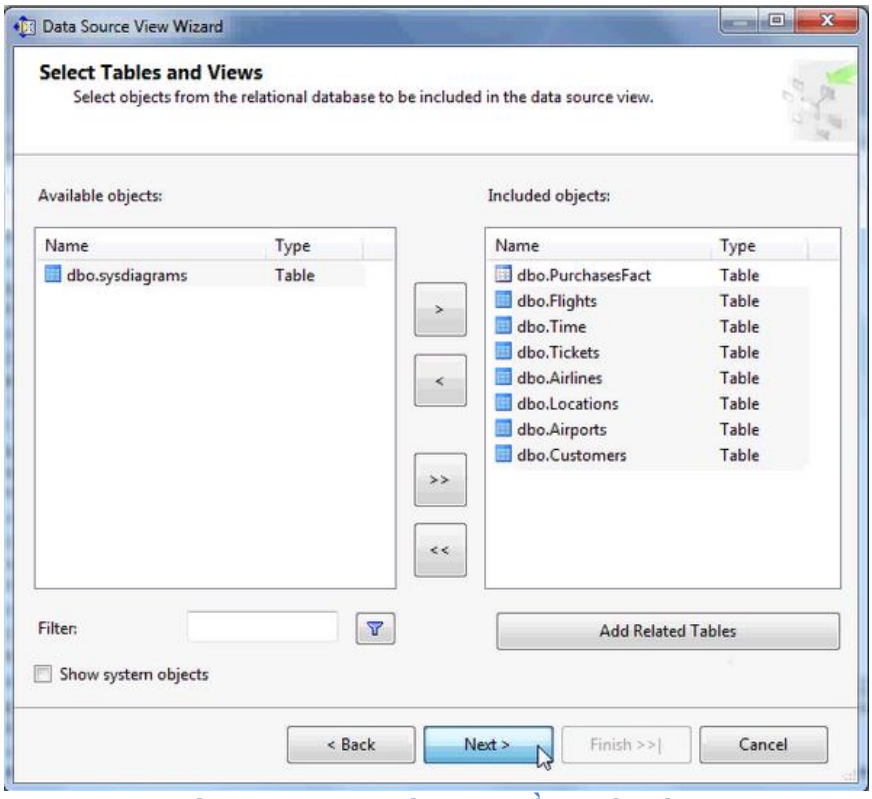
Hình 3.5. Tạo Data Source View

Xác định nguồn dữ liệu (Data Source) cần lấy



Hình 3.6. Xác định nguồn dữ liệu cần lấy

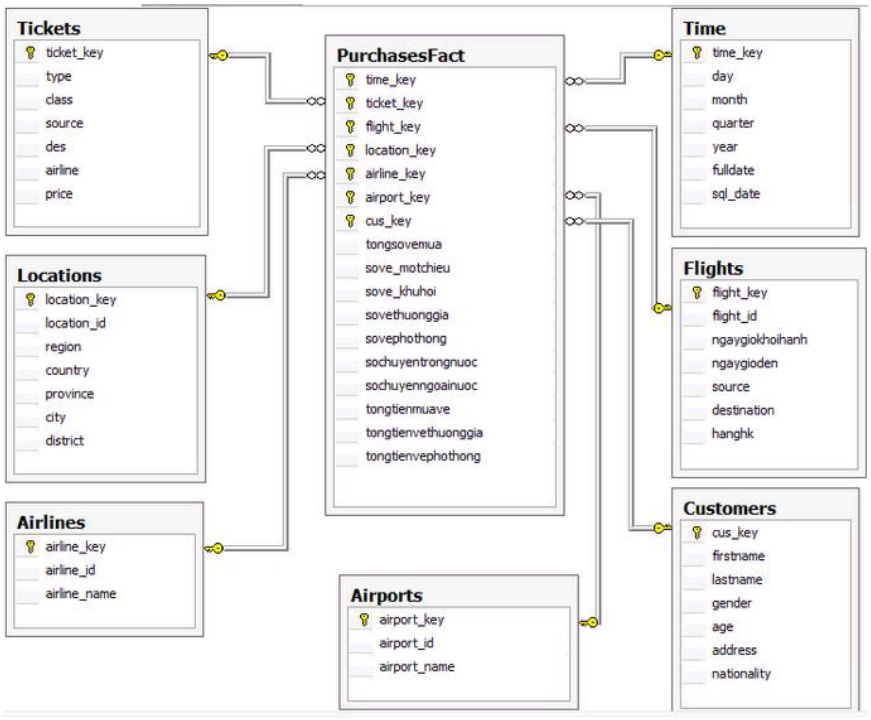
Chọn Next và chọn các bảng cần cho phân tích



Hình 3.7. Chọn các bảng cần phân tích

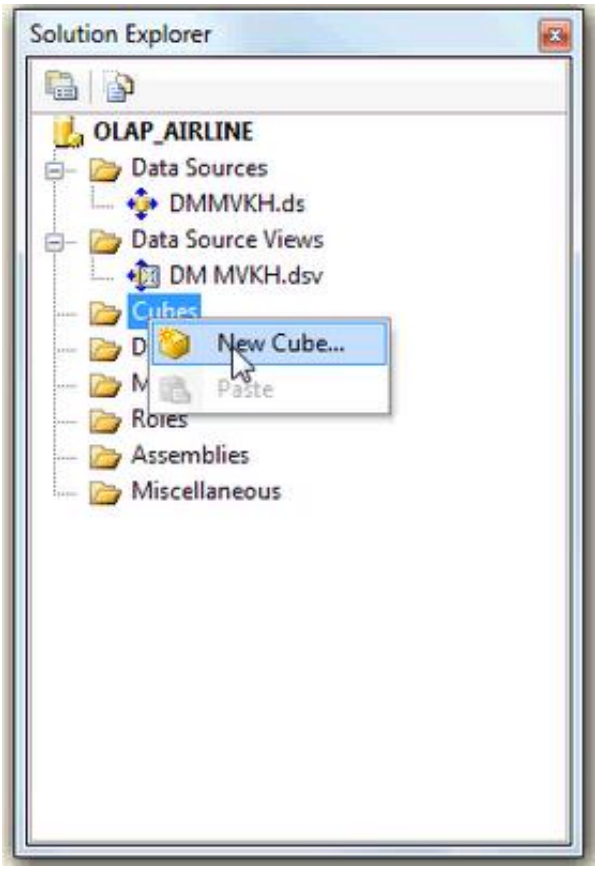
Chú ý: Nếu bạn muốn chọn bảng Fact và các bảng Dimension liên quan đến bảng Fact thì chỉ cần chọn Fact Table đưa qua khung bên phải và bấm nút "Add Related Tables" để tự động lấy các bảng Dimensions liên quan.

Sau khi hoàn thành, các bảng Fact và Dimension như sau:



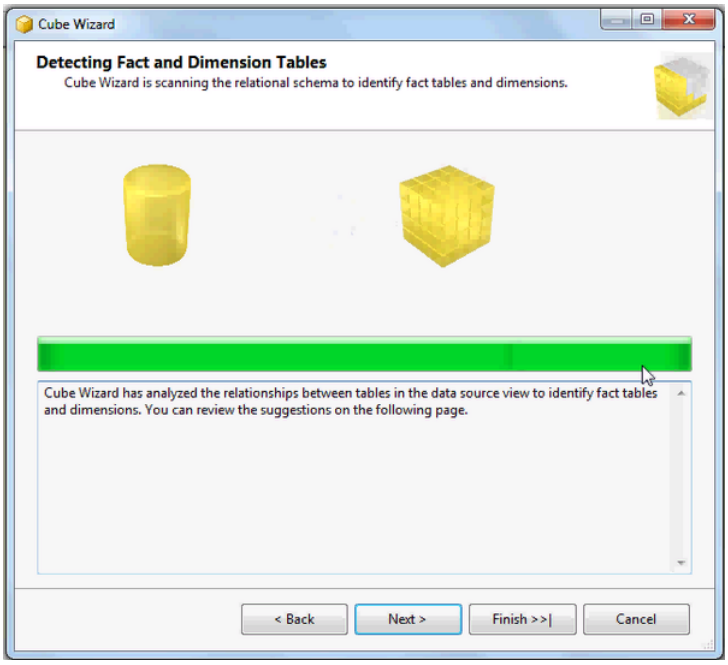
Hình 3.8. Các bảng Fact và Dimension

Sau khi tao Data Source và Data Source View ta tạo dữ liệu khối cho phân tích bằng cách bấm chuột phải lên Cube trong Solution Explorer và chọn New Cube



Hình 3.9. Tạo khối dữ liệu

Chọn Next và chọn nguồn dữ liệu cho Khối, hệ thống sẽ tự động dò tìm fact và Dimension Tables



Hình 3.10. Dò tìm Fact và Dimemtion Tables

KẾT LUẬN

Đề tài Nghiên cứu về OLAP và truy vấn đa chiều đã đạt được các kết quả đạt được bao gồm:

- Nắm bắt được lý thuyết đặc điểm của OLAP.
- Kiến trúc khối OLAP.
- Nắm bắt được cách tiếp cận và phân tích dữ liệu đa chiều.
- Triển khai OLAP trong SQL Server.

Tuy nhiên vẫn còn một số vấn đề mà đề tài chưa đề cập đến: việc tổ chức và quản lý kho dữ liệu trên mạng và thực hiện những liên kết giữa các khối đa chiều với kho dữ liệu hay trực tiếp đến các hệ xử lý tác vụ để tự động hoá việc cập nhật dữ liệu và cấu trúc chiều cho các khối đa chiều; việc phối hợp giữa các khối đa chiều để khai thác tối đa khả năng của các khối đa chiều; nghiên cứu tăng cường khả năng hiển thị kết quả, giúp cho việc mô tả, thay đổi các yêu cầu truy vấn thông tin thuận lợi hơn, linh hoạt hơn.

TÀI LIỆU THAM KHẢO

[1] Viện Công nghệ Thông tin (1997), *Kho dữ liệu - Data Warehouse*, Hà Nội.

[2] Surajit Chaudhuri (1997), *An Overview of Data Warehouse and OLAP Technology*, http://research.microsoft.com.

[3] Ching T.H., Agrawal R., Megiddo N., Srikant R. (1997), *Range Queries in OLAP Data Cubes*, Proceeding ACM SIGMOD.

[4] Alexandre Gachet (2003), *Distributed Decision Support System: A Federalist Model of Cooperation*, University of Fribourg.

[5] William H.Inmon (2005), *Building the Data Warehouse – Fourth Edition*, Wiley Publishing Inc.

[6] Intelligent Science, *Intelligent Decision Support System - IDSS*, http://www.intsci.ac.cn/en/research/idss.html.