

ỨNG DỤNG KỸ THUẬT OLAP PHÂN TÍCH DỮ LIỆU TUYỂN SINH ĐẠI HỌC

APPLYING OLAP TECHNIQUE TO ANALYZE DATA OF UNIVERSITY ENROLLMENT

Nguyễn Văn Chức, Phạm Văn Cường

Trường Đại học Kinh tế, Đại học Đà Nẵng; chuc.nv@due.edu.vn, cuongpv2103@gmail.com

Tóm tắt - Tuyển sinh đại học là công việc rất quan trọng của các trường đại học. Dữ liệu tuyển sinh đại học có khối lượng lớn, cập nhật liên tục, nhất là công tác tổng hợp dữ liệu, truy vấn phức tạp, lập báo cáo dữ liệu nhiều chiều với yêu cầu nhanh chóng và chính xác. Bài báo nghiên cứu kỹ thuật OLAP phân tích dữ liệu đăng ký tuyển sinh của Đại học Đà Nẵng phục vụ cho công tác lưu trữ, tổng hợp, truy vấn dữ liệu để lập báo cáo đa chiều về dữ liệu tuyển sinh đại học, giúp khai thác hiệu quả dữ liệu tuyển sinh nhằm cung cấp thông tin về đăng ký xét tuyển đại học một cách nhanh chóng, kịp thời, chính xác. Các thông tin này cũng rất có ích đối với thí sinh đăng ký xét tuyển, giúp họ có thêm thông tin để quyết định chọn ngành đăng ký phù hợp. Dựa vào kết quả phân tích OLAP, một hệ thống giao tiếp trên nền web được xây dựng để người dùng có thể dễ dàng sử dụng để phân tích dữ liệu tuyển sinh đại học.

Từ khóa - kho dữ liệu; OLAP; tuyển sinh đại học; dữ liệu khối; truy vấn dữ liệu

1. Đặt vấn đề

Với chủ trương của Bộ Giáo dục và Đào tạo (GD&ĐT) nhằm tạo điều kiện thuận lợi và chủ động nhất cho các cơ sở đào tạo và thí sinh, nhất là trong công tác xét tuyển đã làm thay đổi lớn công tác tuyển sinh đại học, cao đẳng. Trong đó công tác xét tuyển là khâu quan trọng, cần phải liên tục cập nhật dữ liệu, tổng hợp và lập báo cáo dữ liệu một cách nhanh chóng, chính xác để ra quyết định trong công bố thông tin kịp thời cho thí sinh.

Kỹ thuật phân tích xử lý trực tuyến OLAP (Online Analytical Processing) là kỹ thuật phân tích dữ liệu chủ yếu của kho dữ liệu, cho phép thực hiện các truy vấn dữ liệu phức tạp trên các khối dữ liệu đa chiều để tổng hợp và lập báo cáo dữ liệu rất nhanh chóng và tiện lợi. [2]

Bài báo tập trung nghiên cứu công nghệ kho dữ liệu, đặc biệt là kỹ thuật OLAP để tổ chức lưu trữ dữ liệu tuyển sinh và phân tích dữ liệu với mong muốn xây dựng một công cụ hữu ích, hỗ trợ công tác quản lý tuyển sinh đại học hiệu quả, giúp cho trường đại học nâng cao hiệu quả trong công tác quản lý và khai thác dữ liệu tuyển sinh, đồng thời giúp thí sinh có được thông tin về quá trình xét tuyển đại học một cách nhanh chóng, chính xác để có quyết định phù hợp trong việc đăng ký xét tuyển.

2. Giải thích một số thuật ngữ

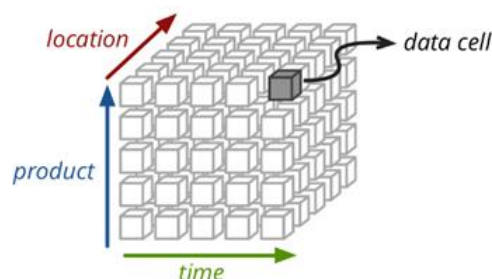
Kho dữ liệu (Data Warehouse, DW): Kho dữ liệu là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người dùng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau. Dữ liệu trong DW được tổ chức dưới dạng khối dữ liệu đa chiều (MultiDimensional Cube). [1]

Khối dữ liệu (Data Cube): Dữ liệu trong kho dữ liệu được thể hiện dưới dạng đa chiều gọi là khối (cube). Mỗi chiều mô tả một đặc trưng nào đó của dữ liệu. [3] Chẳng hạn, với Data Cube bán hàng gồm chiều sản phẩm

Abstract - University enrollment is a very important task of universities. University entrance enrollment data has a huge volume, constantly updated; especially data synthesis, complex queries, building multidimensional data reports are required to be quick and accurate. This paper focuses on OLAP technique to analyze enrollment data of The Da Nang University to facilitate data storage and synthesis, deploying queries to build multidimensional data reports on university enrollment data, helping the university explore data enrollment effectively and providing information about university admission quickly and accurately. This information is also useful for candidates, helping them to have more information to choose a suitable major for enrollment. Based on the results of OLAP analysis, an interface is built on web platform to help users easily analyze data of University enrollment.

Key words - data warehouse; OLAP; university enrollment; data cube; query data

(product) mô tả thông tin về sản phẩm (mã số sản phẩm, tên sản phẩm, giá...), chiều thời gian (time) mô tả về thời gian bán hàng, chiều vị trí (location) mô tả thông tin về địa điểm của các đại lý bán hàng được mô tả như Hình 1.



Hình 1. Data cube bán hàng

Lược đồ hình sao (Star Schema): Đây là mô hình dữ liệu được sử dụng phổ biến để biểu diễn dữ liệu của kho dữ liệu. Lược đồ hình sao về cơ bản gồm có bảng sự kiện (Fact Table) và các bảng chiều (Dimension table). Fact table dùng để theo dõi các biến động của dữ liệu, cấu trúc của Fact table gồm các khóa ngoại, đó là các khóa chính của các Dimension table và các độ đo (Measurement). Dimension Table là các bảng mô tả các đặt trưng của các chiều như chiều thời gian, chiều khách hàng, chiều hàng hóa... [1]

Độ đo (Measurement): Là đại lượng dùng để tính toán, tổng hợp dữ liệu phục vụ cho mục đích phân tích dữ liệu trong Fact table, như tổng tiền bán hàng, tổng số lượng bán... [3]

Phân cấp (Hierarchies): Khái niệm này mô tả sự phân cấp thứ bậc (mức độ chi tiết của dữ liệu) dùng để chi tiết hóa hoặc tổng quát hóa dữ liệu trong quá trình phân tích. Tính phân cấp rất cần thiết trong quá trình phân tích dữ liệu, giúp tổng hợp hay chi tiết từng hạng mục dữ liệu trong DW. [3]

Chẳng hạn đối với chiều thời gian, khi phân tích cần chi

tiết hóa dữ liệu từ năm đến ngày hoặc để tổng hợp dữ liệu từ ngày đến năm, ta có phân cấp như sau: day<week<month<quarter<year. Tương tự, đối với chiều vị trí (location) của các đại lý bán hàng, ta có thể chi tiết hoặc tổng quát hóa dữ liệu theo thứ bậc street<city<province_or_state<country.

Kỹ thuật xử lý phân tích trực tuyến OLAP (OnLine Analytical Processing): OLAP là kỹ thuật phân tích dữ liệu rất quan trọng của kho dữ liệu, cho phép thiết kế các truy vấn dữ liệu phức tạp trên nhiều chiều để tổng hợp và lập báo cáo dữ liệu hiệu quả. [2]

Ngôn ngữ truy vấn dữ liệu đa chiều MDX (MultiDimensional eXpressions): MDX là ngôn ngữ dùng để truy vấn, tính toán, xử lý trên khối dữ liệu đa chiều một cách hiệu quả và thuận tiện. MDX có thể trả lời được các truy vấn phức tạp để lập báo cáo dữ liệu một cách hiệu quả. Các truy vấn phức tạp trong SQL như pivot được thực hiện một cách dễ dàng trong MDX. [4]

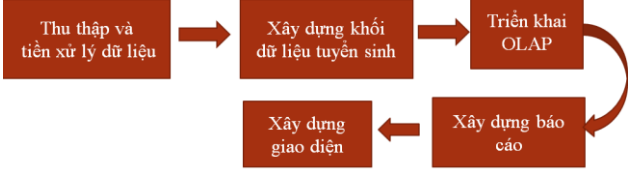
Cấu trúc của MDX cũng giống như SQL nhưng mở rộng hơn để thao tác với các khối dữ liệu đa chiều: [4]

```
SELECT [mô tả chiều thứ nhất], [mô tả chiều thứ hai],...  
FROM [mô tả về khối dữ liệu]  
WHERE [điều kiện cắt lớp dữ liệu]
```

3. Xây dựng hệ thống phân tích dữ liệu tuyển sinh đại học

3.1. Quy trình triển khai hệ thống phân tích dữ liệu tuyển sinh đại học:

Hệ thống phân tích dữ liệu tuyển sinh đại học được tiến hành theo các bước chính như Hình 2.



Hình 2. Quy trình triển khai hệ thống phân tích dữ liệu tuyển sinh đại học

Bước 1. Thu thập và tiền xử lý dữ liệu

Dữ liệu thu thập được hơn 10.000 dòng từ thông tin tuyển sinh năm 2015 vào Trường Đại học Kinh tế - Đại học Đà Nẵng trên website <http://ts.udn.vn> [5]. Dữ liệu được thu thập qua các đợt công bố thông tin tuyển sinh theo các mốc thời gian và được chuyển vào excel để lưu trữ và tiền xử lý.

Dữ liệu ban đầu gồm rất nhiều thuộc tính theo qui định đăng ký xét tuyển đại học, sau quá trình tiền xử lý dữ liệu (tách, gộp dữ liệu, chuyển đổi dữ liệu, loại bỏ các thuộc tính không sử dụng...) sau khi tiền xử lý, các thuộc tính chính dùng trong phân tích được mô tả như Bảng 1:

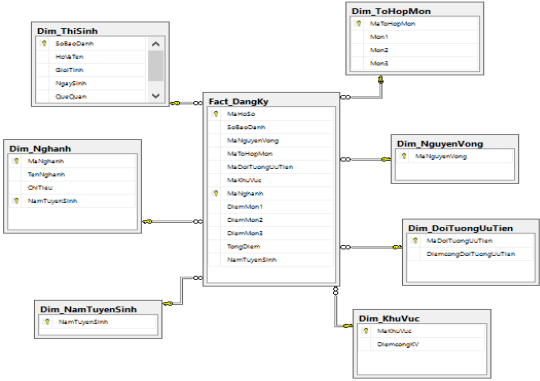
Bảng 1. Cấu trúc dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Độ rộng	Giải thích
1	MaHS	Number	6	Mã hồ sơ
2	SoDB	Text	10	Số báo danh

3	MaNV	Number	1	Mã nguyên vọng (1,2,3,4)
4	MaTHM	Text	3	Mã tổ hợp môn (A00, A01, D00, D01...)
5	MaDTUT	Text	4	Mã đối tượng ưu tiên
6	MaKV	Text	4	Mã khu vực
7	MaNghanh	Text	8	Mã ngành
8	NamTS	Number	4	Năm tuyển sinh
9	DiemMon1	Number	5,2	Điểm môn 1
10	DiemMon2	Number	5,2	Điểm môn 2
11	DiemMon3	Number	5,2	Điểm môn 3
12	TongDiem	Number	5,2	Tổng điểm
13	ChiTieuTS	Number	4	Chỉ tiêu tuyển sinh theo ngành

Bước 2. Xây dựng khối dữ liệu tuyển sinh

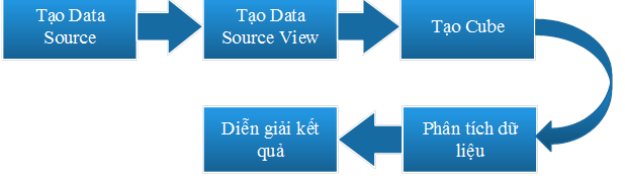
Khối dữ liệu tuyển sinh được xây dựng theo lược đồ hình sao (Star Schema) gồm bảng Fact Table là Fact_DangKy và 7 bảng chiều (Dimension Table) là Dim_ThiSinh, Dim_ToHopMon, Dim_KhuVuc, Dim_NguyenVong, Dim_Nghanh, Dim_NamTuyenSinh, Dim_DoiTuongUuTien như Hình 3:



Hình 3. Khối dữ liệu tuyển sinh

Bước 3. Triển khai OLAP phân tích dữ liệu tuyển sinh

Quy trình triển khai OLAP được thực hiện như Hình 4



Hình 4. Quy trình triển khai OLAP

Triển khai OLAP phân tích dữ liệu tuyển sinh đại học được thực hiện trên công cụ SQL Server Data Tools (SSDT) của Microsoft. SSDT là công cụ rất mạnh trong việc tổ chức dữ liệu khối và triển khai OLAP với giao diện dễ sử dụng, khả năng tích hợp dữ liệu mạnh mẽ và nhất là khả năng biểu diễn kết quả phân tích trực quan, dễ hiểu, dễ sử dụng [4].

Hình 5 minh họa thiết kế truy vấn tổng hợp dữ liệu theo 3 chiều: Ngành học – Nguyên vọng – Khu vực để thống kê số lượng những sinh viên đăng ký ngành Quản trị kinh doanh theo nguyên vọng 1 ở tất cả các khu vực:

Dimension	Hierarchy	Operator	Filter Expression
Dim Nghanh	Ma Nghanh	Equal	{ QTKD }
Dim Nguyen Vong	Ma Nguyen Vong	Equal	{ 1 }
<Select dimension>			
Ma Nghanh	Ma Nguyen Vong	Ma Khu Vuc	Fact Dang Ky Count
QTKD	1	KV01	99
QTKD	1	KV02	81
QTKD	1	KV02NT	93
QTKD	1	KV03	99

Hình 5. Phân tích dữ liệu theo chiều
Ngành học – Nguyễn vọng – Khu vực

Bước 4: Xây dựng các báo cáo dữ liệu

Quy trình xây dựng các báo cáo như Hình 6:

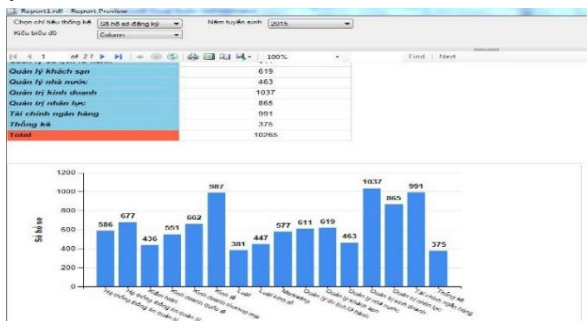


Hình 6. Quy trình xây dựng các báo cáo

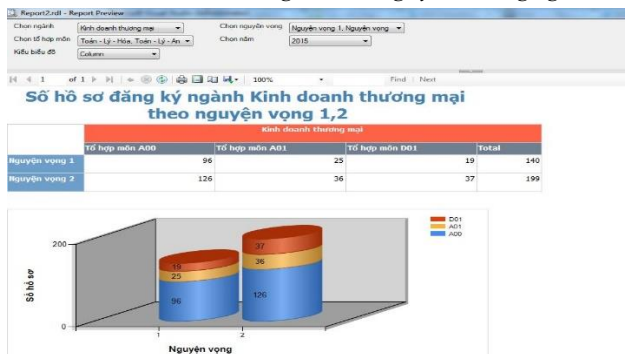
Sau khi triển khai thành công OLAP, tiến hành thiết kế các truy vấn phục vụ cho tổng hợp dữ liệu để lập báo cáo, nhằm cung cấp thông tin về tình hình đăng ký tuyển sinh như báo cáo số hồ sơ tuyển sinh theo từng nguyện vọng, báo cáo điểm chuẩn tham khảo của các ngành tại một thời điểm nhất định...

Sau khi tổng hợp dữ liệu, sử dụng dịch vụ Reporting Services của Microsoft để lập báo cáo và xuất bản dữ liệu lên web cho người dùng sử dụng.

Sau đây là kết quả một số báo cáo dữ liệu về tình hình tuyển sinh.



Hình 7. Báo cáo số lượng hồ sơ đăng ký theo từng ngành



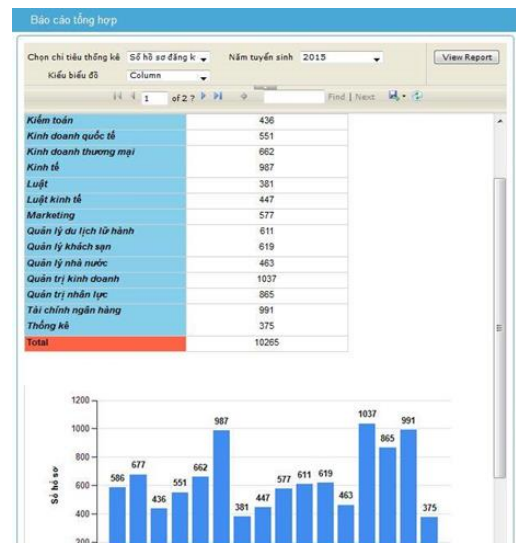
Hình 8. Báo cáo số lượng hồ sơ đăng ký ngành
"Kinh doanh Thương mại" theo nguyện vọng 1 và 2"



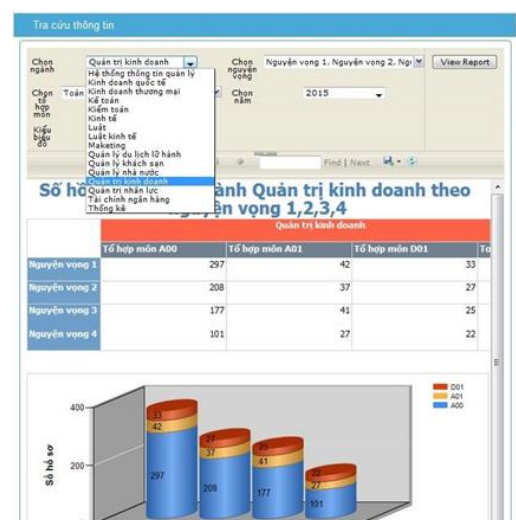
Hình 9. Báo cáo tham khảo điểm chuẩn theo từng ngành

Bước 5: Xây dựng giao diện

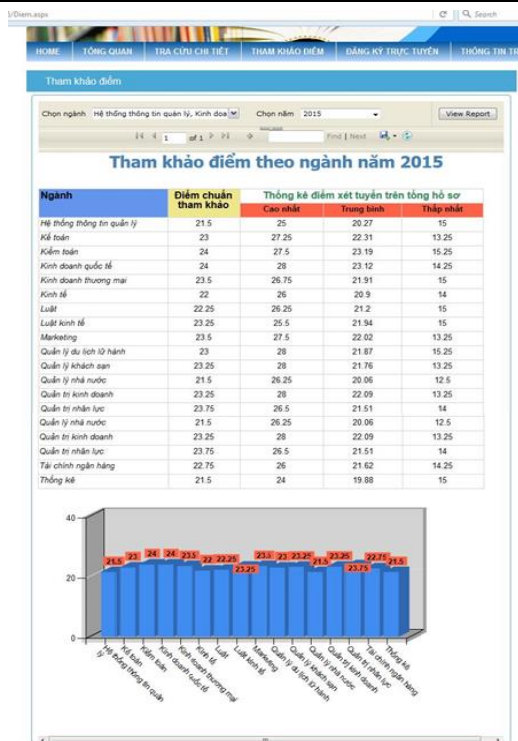
Dựa vào kết quả phân tích từ OLAP, một hệ thống giao tiếp được xây dựng trên nền web cho phép người dùng sử dụng để tra cứu, tham khảo thông tin tuyển sinh.



Hình 10. Trang tra cứu số hồ sơ đăng ký xét tuyển



Hình 11. Trang tra cứu chi tiết



Hình 12. Trang tra cứu điểm chuẩn tham khảo

4. Kết luận và hướng phát triển

Bài báo đã tìm hiểu lý thuyết về kho dữ liệu đặc biệt là kỹ thuật OLAP để phân tích khối dữ liệu đa chiều, từ đó ứng dụng kỹ thuật này vào xây dựng hệ thống phân tích dữ liệu tuyển sinh đại học, minh họa bằng dữ liệu xét tuyển năm 2015 của Trường Đại học Kinh tế - Đại học Đà Nẵng.

Kết quả phân tích giúp trường đại học lập các báo cáo nhanh chóng, chính xác và dễ dàng về tình hình tuyển sinh như tổng số hồ sơ đăng ký theo ngành, nguyện vọng, tổ hợp môn, ..., từ đó ra quyết định kịp thời và hiệu quả trong công tác tuyển sinh. Thêm vào đó, kết quả phân tích cũng giúp cho thí sinh có được thông tin về tình hình xét tuyển đại học một cách kịp thời để lựa chọn ngành đăng ký xét tuyển phù hợp. Nhóm tác giả đã xây dựng một giao tiếp trên nền web giúp người dùng sử dụng kết quả phân tích một cách dễ dàng và thuận tiện. Trong thời gian tới sẽ mở rộng thu thập và phân tích dữ liệu tuyển sinh của các cơ sở đào tạo thuộc Đại học Đà Nẵng, đồng thời với kho dữ liệu tuyển sinh đã xây dựng, sẽ mở rộng nghiên cứu các kỹ thuật khai phá dữ liệu (phân lớp, phân cụm, phát hiện luật kết hợp...) để khai thác tri thức từ kho dữ liệu tuyển sinh, nhằm hỗ trợ tốt hơn cho công tác tuyển sinh của các trường đại học và đăng ký xét tuyển đại học của thí sinh.

TÀI LIỆU THAM KHẢO

- [1] Ralph Kimball and Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling 3rd Edition*, John Wiley & Sons, Inc, 2013.
- [2] Erik Thomsen, *OLAP Solutions: Building Multidimensional Information Systems 2nd Edition*, John Wiley & Sons, Inc, 2002.
- [3] Chris Webb, Alberto Ferrari, Marco Russo, *Expert Cube Development with SSAS Multidimensional Models*, Packt Publishing Ltd, 2014.
- [4] Sivakumar Harinath, Ronald Pihlgren, Denny Guang-Yeu Lee, John Sirmon, Robert M. Bruckner, *Professional Microsoft SQL Server 2012 Analysis Services with MDX and DAX 1st Edition*, John Wiley & Sons, Inc, 2012
- [5] Website tuyển sinh Đại học Đà Nẵng: <http://ts.udn.vn>
- [6] Website của nhóm phát triển ứng dụng Công nghệ thông tin BIS: <http://bis.net.vn>

(BBT nhận bài: 17/01/2017, hoàn tất thủ tục phản biện: 02/03/2017)