

Khai phá dữ liệu trong kinh doanh (Data Mining in Business)

Bộ môn Tin học

Giới thiệu học phần

- ▶ 1. Tên học phần: Khai phá dữ liệu trong kinh doanh (Data Mining in Business)
- ▶ 2. Mã học phần: INFO1831
- ▶ 3. Số tín chỉ: 2 (24,6)

Tài liệu Tham khảo

- ▶ [1] Tập bài giảng Khai phá dữ liệu trong kinh doanh của Bộ môn Tin học.
- ▶ [2] Vincent Rainard, Building a Data Warehouse With Examples in SQL. Apress, 2008.
- ▶ [3] ZhaoHui Tang, Jamie MacLennan. Data Mining with SQL Server 2005. Wiley, 2005.

Tài liệu tham khảo khác

- ▶ [4] Paolo Giudici. Applied data mining statistical methods for business and industry. Wiley, 2003.
- ▶ [5] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Second Edition.
- ▶ [6] Joseph Fong, "Information Systems Reengineering and Integration", Springer Verlag, 2006, ISBN 978-1-84628-382-6.
- ▶ [7]. SQL server 2008 for BI. Website: <https://atdhebuja.files.wordpress.com/2011/03/sql-server-2008-businessintelligence.doc>

Nội dung

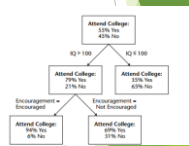
- ▶ Chương 1: Tổng quan về khai phá dữ liệu trong kinh doanh
- ▶ Chương 2: Tiền xử lý và Tổ chức dữ liệu kinh doanh
- ▶ Chương 3: Sử dụng công cụ OLAP trong khai phá dữ liệu
- ▶ Chương 4: Một số bài toán khai phá trong kinh doanh

Chương 1: Tổng quan về khai phá dữ liệu trong kinh doanh

- ▶ 1.1. Giới thiệu chung
 - ▶ Khai phá dữ liệu là gì
 - ▶ Khái niệm kho dữ liệu
 - ▶ Một số mẫu dữ liệu trong kinh doanh cần khai phá
- ▶ Tầm quan trọng của khai phá dữ liệu trong kinh doanh
- ▶ 1.2. Quy trình khai phá dữ liệu trong kinh doanh
 - ▶ Xác định mục tiêu
 - ▶ Chuẩn bị và tổ chức dữ liệu
 - ▶ Lựa chọn phương pháp và mô hình khai phá
 - ▶ Phân tích và đánh giá kết quả trong hỗ trợ kinh doanh

1.1 Giới thiệu chung

- ▶ Khai phá dữ liệu: Là việc phân tích dữ liệu và tìm kiếm các tri thức ẩn giấu bên trong bằng cách sử dụng các kỹ thuật của công nghệ thông tin một cách tự động hoặc bán tự động.
- ▶ Ví dụ: Nếu $IQ > 100$ Thì????

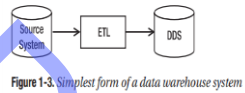


1.1 Giới thiệu chung

- ▶ Kho dữ liệu (Data Warehouse):
- ▶ Là một hệ thống thu lượm và hợp nhất dữ liệu một cách định kỳ từ các hệ thống nguồn vào một kho lưu trữ dữ liệu nhiều chiều hoặc thống thường.
- ▶ Kho này thường giữ các dữ liệu hàng năm và là nguồn cung cấp cho các hệ thống kinh doanh thông minh (BI) hoặc các hoạt động phân tích dữ liệu khác. Việc cập nhật dữ liệu là xử lý theo lô thay vì cập nhật từng giao dịch ở hệ thống nguồn.

Kho dữ liệu (Data Warehouse)

- ▶ Ví dụ:
- ▶ ETL (extract, transform, and load)
- ▶ DDS (dimensional data store)
- ▶ Người dùng trực tiếp truy vấn dữ liệu từ DDS.
- ▶ Trường hợp khác: có thể xây dựng các chương trình ứng dụng để sử dụng dữ liệu từ DDS (dùng spreadsheets, pivot tables,...)



Một số bài toán trong Khai phá dữ liệu KD

- ▶ Phân tích các dữ liệu thị hiếu (Churn analysis): Telecom, banking, and insurance. Ví dụ: để có 1 sản phẩm điện thoại cần 200 USD cho quảng cáo
- ▶ Phân tích DL tham khảo (Cross-selling). VD: Amazon, các website TMĐT sẽ gợi ý sản phẩm liên quan khi người dùng chọn mua sp nào đó
- ▶ Quản lý rủi ro (Risk management): dùng trong chứng minh tài chính của user trong ngân hàng.
- ▶ Phân khúc khách hàng (Customer segmentation): phân tích thông tin khách hàng để phân loại đối tượng khách hàng
- ▶

1.1 Giới thiệu chung

- ▶ Tầm quan trọng của việc khai phá dữ liệu
 - ▶ Sự kiểm tìm tri thức trong một lượng dữ liệu lớn là nhu cầu cấp thiết của nhiều doanh nghiệp.
 - ▶ Tăng khả năng cạnh tranh trên thị trường
 - ▶ Sử dụng nhiều công nghệ sẵn có

- ▶ Liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu...
- ▶ Đặc biệt nó rất gần gũi với lĩnh vực thống kê (phương pháp thống kê để mô hình dữ liệu và phát hiện các mẫu, luật ...)
- ▶ Kho dữ liệu (Data Warehousing) và các công cụ phân tích trực tuyến (OLAP- On Line Analytical Processing) cũng liên quan rất chặt chẽ với data mining

Ứng dụng trong thực tế

- ▶ Bảo hiểm, tài chính và thị trường chứng khoán: phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán. Danh mục vốn và giá, lãi suất, dữ liệu thể tin dụng, phát hiện gian lận, ...
- ▶ Thống kê, phân tích dữ liệu và hỗ trợ ra quyết định.
- ▶ Text mining và Web mining: Phân lớp văn bản và các trang Web, tóm tắt văn bản,...
- ▶ Lĩnh vực khoa học: Quan sát thiên văn, dữ liệu gene, dữ liệu sinh vật học, tìm kiếm, so sánh các hệ gene và thông tin di truyền, mối liên hệ gene và một số bệnh di truyền, ...
- ▶ Mạng viễn thông: Phân tích các cuộc gọi điện thoại và hệ thống giám sát lỗi, sự cố, chất lượng dịch vụ, ...
- ▶

1.2. Quy trình khai phá dữ liệu trong kinh doanh

- ▶ 1.2.1. Xác định mục tiêu
- ▶ 1.2.2. Chuẩn bị và tổ chức dữ liệu
- ▶ 1.2.3. Lựa chọn phương pháp và mô hình khai phá
- ▶ 1.2.4. Phân tích và đánh giá kết quả trong hỗ trợ kinh doanh

1.2.1. Xác định mục tiêu

- ▶ Xác định mục tiêu chung, cụ thể
- ▶ Xác định các nhiệm vụ cần phải hoàn thành.
 - ▶ Bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu

1.2.2. Chuẩn bị và tổ chức dữ liệu

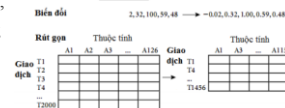
- ▶ Thu thập và tiền xử lý dữ liệu
- ▶ Thu thập: dữ liệu được lấy từ nhiều nguồn khác nhau như: giấy, hệ thống khác, website,...
- ▶ Tiền xử lý DL: DL không đồng nhất, ... có thể gây ra các nhầm lẫn. Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hoá

1.2.2. Chuẩn bị và tổ chức dữ liệu

- ▶ Ví dụ, Một Công ty điện tử đưa ra yêu cầu phân tích dữ liệu bán hàng tại các chi nhánh.
- ▶ → Phải kiểm tra kỹ lưỡng cơ sở dữ liệu bán hàng của toàn công ty cũng như kho xưởng để xác định và lựa chọn các thuộc tính hoặc chiều thông tin đưa vào phân tích như: Chúng loại mặt hàng, mặt hàng, giá cả, chi nhánh bán ra.
- ▶ Lỗi có thể: không ghi lại thông tin, ghi sai thông tin so với quy định, quy chuẩn bình thường.
- ▶ Nhiệm vụ: kiểm soát và kiểm tra dữ liệu đầu vào

1.2.2. Chuẩn bị và tổ chức dữ liệu

- ▶ Quá trình chuẩn bị dữ liệu phục vụ khai phá dữ liệu:
- ▶ - Làm sạch dữ liệu
- ▶ - Tích hợp dữ liệu;
- ▶ - Biến đổi dữ liệu;
- ▶ - Rút gọn dữ liệu



1.2.3. Lựa chọn phương pháp và mô hình khai phá

- ▶ Chọn lựa phương pháp và mô hình sẵn có để trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu.
- ▶ Phương pháp và mô hình dựa trên bài toán:
 - ▶ các bài toán mang tính mô tả - đưa ra tính chất chung nhất của dữ liệu,
 - ▶ các bài toán dự báo - bao gồm cả việc phát hiện các suy diễn dựa trên dữ liệu hiện có.

1.2.3. Lựa chọn phương pháp và mô hình khai phá

- ▶ Các phương pháp chính
 - ▶ Classification: Algorithms: Decision trees, neural network, and Naïve Bayes.

1.2.3. Lựa chọn phương pháp và mô hình khai phá

► Các phương pháp chính

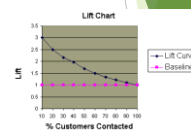
- Classification: Algorithms: Decision trees, neural network, and Naïve Bayes.
- Clustering: K-means, Hierarchical Clustering (BRICH -Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using REpresentatives),...
- Association: Thường được sử dụng trong phân tích thị trường (phân tích các phần tử được sử dụng thường xuyên, và luật kết hợp).
 - find frequent itemsets
 - find association rules.

- Phương pháp hồi quy (Regression): tương tự như phân lớp, điểm khác nhau là ở chỗ thuộc tính để dự báo là liên tục chứ không phải rời rạc.
- Phương pháp dự báo (Forecasting): mô hình ARIMA (AutoRegressive Integrated Moving Average), hoặc Decision Tree



1.2.4. Phân tích và đánh giá kết quả trong hỗ trợ kinh doanh

- Đánh giá mô hình: Sử dụng lift chart (sử dụng a trained model để dự báo giá trị của tập test. Vẽ đồ thị của mô hình với dữ liệu thực và dữ liệu dự báo).
- VD:
- Xem actual lift.
- Để plot the chart: Tính các điểm ở lift curve bằng cách xđ tỉ lệ giữa kết quả dự báo bởi mô hình và kết quả ko dùng mô hình.
- Xem thêm ở: <http://www2.cs.uregina.ca/~dbd/cs831/index.html>



Dự án phần mềm liên quan (Data Mining Project Cycle)

- Step 1: Data Collection
- Step 2: Data Cleaning and Transformation
- Step 3: Model Building: Phụ thuộc rất nhiều vào mục tiêu của bài toán để chọn (classification task, an association task or a segmentation task). Nên xây dựng mô hình với nhiều thuật toán với các tham số khác nhau
- Step 4: Model Assessment: sử dụng các phương pháp như: Lift chart, Confusion matrix,...
- Step 5: Reporting (findings (patterns), and reports about the prediction or forecast)
- Step 6: Prediction (Scoring) - use these models for prediction

- ▶ Step 7: Application Integration - Embedding data mining into business applications
- ▶ Step 8: Model Management (new version updated; security)

Các nhà cung cấp chính

- ▶ SAS.
- ▶ SPSS
- ▶ IBM: Intelligent Miner, developed by an IBM German subsidiary.
- ▶ Microsoft Corporation: Microsoft was the first major database vendor to include data mining features in a relational database. SQL Server 2000.
- ▶ Oracle: Oracle 9i shipped in 2000
- ▶ Angoss: Angoss' KnowledgeSTUDIO : add-in to SQL server
- ▶ KXEN: KXEN is a data mining software founded in 1998 (similar to excel) – SAP solutions

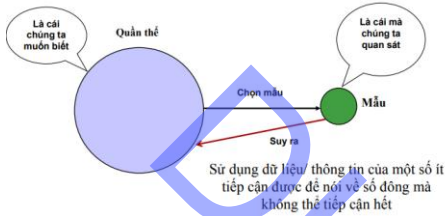
Chương 2: Tiền xử lý và Tổ chức dữ liệu kinh doanh

- ▶ 2.1. Tiền xử lý dữ liệu
 - ▶ 2.1.1. Lựa chọn mẫu
 - ▶ 2.1.2. Làm sạch dữ liệu
 - ▶ 2.1.3. Tích hợp và biến đổi dữ liệu
- ▶ 2.2. Tổ chức dữ liệu
 - ▶ 2.2.1. Tổ chức và cài đặt kho dữ liệu (Data warehouse)
 - ▶ 2.2.2. Tổ chức và cài đặt Data Mart

2.1. Tiền xử lý dữ liệu

- ▶ 2.1.1. Lựa chọn mẫu:
 - ▶ Trong nghiên cứu thống kê và phương pháp nghiên cứu định lượng, một **mẫu dữ liệu** là một tập hợp các dữ liệu thu thập được và/ hoặc được lựa chọn từ một tổng thể thống kê bằng một quy tắc rõ ràng.
 - ▶ Lựa chọn mẫu: Chọn lựa những dữ liệu phù hợp với mục tiêu của bài toán

Chọn mẫu?



Chọn mẫu

- Chọn mẫu là một quy trình lựa chọn cá thể từ quần thể cho quan sát, để có thể coi kết quả quan sát mẫu thành kết quả quan sát quần thể, ở một mức độ chấp nhận mà xác định được.
 - Chú ý: Mẫu là đại diện của một quần thể.
 - Mức độ đại diện phải được xác định/ đo lường được.
 - Có hai cách chọn mẫu:
 - Chọn mẫu ngẫu nhiên
 - Chọn mẫu không ngẫu nhiên

Chọn mẫu ngẫu nhiên

- Cách làm: Trước tiên lập danh sách các đơn vị của tổng thể chung theo một trật tự quy ước nào đó, sau đó đánh số thứ tự các đơn vị trong danh sách. Đầu tiên chọn ngẫu nhiên 1 đơn vị trong danh sách; sau đó cứ cách đều k đơn vị lại chọn ra 1 đơn vị vào mẫu,... cứ như thế cho đến khi chọn đủ số đơn vị của mẫu.
- Ví dụ: Dựa vào danh sách bầu cử tại 1 thành phố, ta có danh sách theo thứ tự của tên chủ hộ, bao gồm 240.000 hộ. Ta muốn chọn ra một mẫu có 2000 hộ. Vậy khoảng cách chọn là: $k = 240000/2000 = 120$, có nghĩa là cứ cách 120 hộ thì ta chọn một hộ vào mẫu.

Phương pháp chọn mẫu không ngẫu nhiên

- Chọn mẫu không ngẫu nhiên (hay chọn mẫu phi xác suất) là phương pháp chọn mẫu mà các đơn vị trong tổng thể chung không có khả năng ngang nhau để được chọn vào mẫu nghiên cứu.
- VD: Tiến hành phỏng vấn các bà nội trợ tới mua hàng tại siêu thị tại một thời điểm nào đó; như vậy sẽ có rất nhiều bà nội trợ do không tới mua hàng tại thời điểm đó nên sẽ không có khả năng được chọn

Công thức tính cỡ mẫu

- Với trường hợp cỡ mẫu lớn và không biết tổng thể.

$$n = \frac{z^2 (p \cdot q)}{e^2}$$

Tính cỡ mẫu

- Trong đó:
 n = là cỡ mẫu
 z = giá trị phân phối tương ứng với độ tin cậy lựa chọn (nếu độ tin cậy 95% thì giá trị z là 1,96, 99% thì $z = 2.58 \dots$)
 p = là ước tính tỷ lệ % của tổng thể
 $q = 1 - p$
 thường tỷ lệ p và q được ước tính 50%/50% đó là khả năng lớn nhất có thể xảy ra của tổng thể.
 e = sai số cho phép ($\pm 3\%$, $\pm 4\%$, $\pm 5\% \dots$)

Ví dụ

- Tính cỡ mẫu của một cuộc trưng cầu ý kiến trước một cuộc bầu cử với độ tin cậy là 95% với giá trị z tương ứng là 1.96, sai số cho phép là nằm trong khoảng $\pm 5\%$. Giả định $p \cdot q$ lớn nhất có thể xảy ra là $0.5 \cdot 0.5$.
- Cỡ mẫu sẽ được tính là:

Tính cỡ mẫu

$$n = \left[\frac{1}{N} + \frac{N-1}{N} \frac{1}{P \cdot Q} \left(\frac{k}{z_{1-\alpha/2}} \right)^2 \right]^{-1}$$

- Ở đó
 N = số lượng đơn vị trong tổng thể.
 P = tỷ lệ tổng thể.
 $Q = 1 - P$,
 k = sai số cho phép.

Tính cỡ mẫu

- Nếu tổng thể nhỏ và biết được tổng thể thì dùng công thức sau:

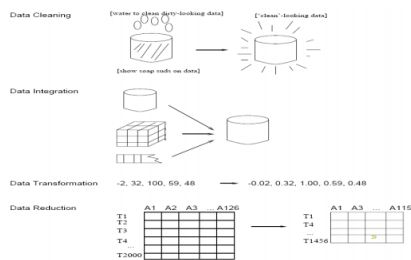
Với n là cỡ mẫu, N là số lượng tổng thể, e là sai số tiêu chuẩn

$$n = \frac{N}{1 + N(e)^2}$$

Ví dụ

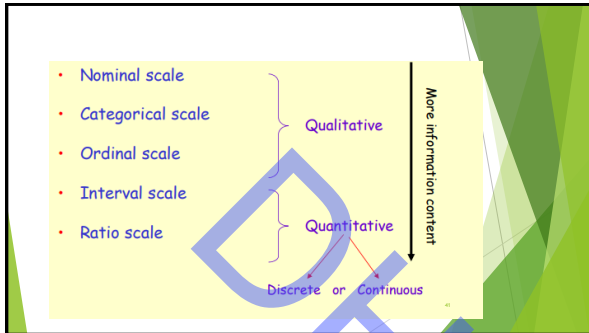
- Tính cỡ mẫu của một cuộc điều tra với Tổng thể là N= 2000, độ chính xác là 95%, sai số tiêu chuẩn là +/- 5%.
- cỡ mẫu sẽ được tính là:

Cách chuẩn bị dữ liệu



Các kiểu dữ liệu

- Nominal scale
- Categorical scale
- Ordinal scale
- Interval scale
- Ratio scale



VD

- Nominal: ID numbers, Names of people
- Categorical: eye color, zip codes
- Ordinal: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval: calendar dates, temperatures in Celsius or Fahrenheit, GRE (Graduate Record Examination) and IQ scores
- Ratio: temperature in Kelvin, length, time, counts

Lí do phải làm sạch DL

DL thực thường được coi là “không sạch”

Không đầy đủ: Thiếu thuộc tính

- e.g., occupation=""

Nhiều: lỗi hoặc ngoại lai

- e.g., Salary="-10"

Không đồng nhất:

- e.g., Age="42" Birthday="03/07/1997"
- e.g., là "1,2,3", nhưng lại đánh "A, B, C"

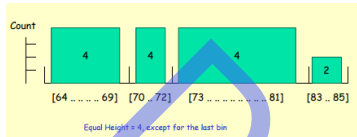
2.1.2. Làm sạch dữ liệu

► Làm sạch dữ liệu :

- Thiếu giá trị: Điền, bỏ qua...
- Dữ liệu nhiều : làm mịn (bin- các dải có cùng kích thước)
- VD: Mảng lưu giá các mặt hàng: 4, 8, 15, 21, 21, 24, 25, 28, 34 Phân thành các bin Bin 1: 4, 8, 15 Bin 2: 21, 21, 24 Bin 3: 25, 28, 34
- Hồi quy

VD

- Temperature values: 64 65 68 69 70 71 72 72 75 75 80 81 83 85



Bài tập

- Rời rạc hóa các giá trị sau, dùng EW (Equal width) and ED (Equal Depth) binning
- 13, 15, 16, 16, 19, 20, 21, 22, 22, 25, 30, 33, 35, 35, 36, 40, 45

2.1.3. Tích hợp và biến đổi dữ liệu

- Tích hợp dữ liệu: kết hợp thành một kho dữ liệu thống nhất
- Biến đổi dữ liệu: biến đổi sang các dạng phù hợp cho việc khai phá dữ liệu.
- Các phương pháp:
 - Chuẩn hóa DL: biến đổi về khoảng nhỏ hơn (dùng phương pháp min- max) ví dụ biến đổi về khoảng [0,1].
 - - Làm mịn: Loại bỏ các trường hợp nhiễu khỏi dữ liệu
 - - Tổng hợp: tổng hợp doanh số bán hàng để tính toán hàng tháng và hàng năm.
 - - Khái quát hóa dữ liệu: sử dụng khi cần tổng quát hóa dữ liệu ở mức cao hơn.

Lí do

- Thuật toán trong Data Mining thường được xây dựng để xử lý với dữ liệu nguyên. Nên các cột dữ liệu Boolean $\rightarrow$ Integer
- Dữ liệu kiểu số liên tiếp (Continuous column) (ví dụ như Income, Age) thường được chia vào các Bin hoặc ánh xạ vào khoảng từ [0,1] hoặc [-1,1].

Tích hợp dữ liệu

- Tích hợp dữ liệu để phát hiện dư thừa thông tin (dữ liệu)
- VD: Thuộc tính doanh thu hàng năm có thể là dư thừa nếu như nó có thể được suy diễn từ các thuộc tính hoặc tập thuộc tính khác
- Một số dư thừa có thể được phát hiện thông qua các phân tích tương quan. Giả sử cho hai thuộc tính, việc phân tích tương quan có thể chỉ ra mức độ một thuộc tính phụ thuộc vào thuộc tính kia, dựa trên các dữ liệu có trong nguồn.

2.1.3. Tích hợp và biến đổi dữ liệu

- Chuẩn hóa DLI:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- VD: Giả sử giá trị nhỏ nhất và lớn nhất cho thuộc tính “thu nhập bình quân” là 500.000 và 4.500.000.

- Ảnh xạ giá trị 2.500.000 về khoảng [0, 1, 0]

$$v' = \frac{2.500.000 - 500.000}{4.500.000 - 500.000} (1,0 - 0) + 0 = \frac{2.000.000}{4.000.000} = 0,5$$

Phương pháp Z-score

- Các giá trị của một thuộc tính A được chuẩn hóa dựa vào độ lệch tiêu chuẩn và trung bình của A.
- $\bar{A}$: trung bình; σ_A : độ lệch tiêu chuẩn
- VD: giả sử độ lệch tiêu chuẩn và trung bình là: 1.000.000 và 500.000, thì giá trị 2.500.000 sẽ là

$$v' = \frac{2.500.000 - 1.000.000}{500.000} = \frac{1.500.000}{500.000} = 3$$

Phương pháp decimal scale

- Di chuyển dấu phân các phần thập phân của các giá trị của thuộc tính A. Số chữ số sau dấu phân cách thập phân được xác định phụ thuộc vào giá trị tuyệt đối lớn nhất có thể có của thuộc tính A.

$$v' = \frac{v}{10^j}$$

Trong đó j là giá trị nguyên nhỏ nhất thỏa mãn $\text{Max}(|v'|) < 1$

- VD: Ví dụ: Giả sử rằng các giá trị của thuộc tính A được ghi nhận nằm trong khoảng -968 đến 917.
- → chia cho 1.000 (j = 3). Như vậy giá trị -986 sẽ chuyển thành - 0.986 và 917 được chuyển thành 0.917

Bài tập

- ▶ 1. Nếu một thuộc tính trong nguồn dữ liệu Điểm: Sinh viên có các giá trị A, B, C, D, F thì kiểu dữ liệu dự kiến của thuộc tính đó trong quá trình tiền xử lý là gì?
- ▶ 2. Cho mảng một chiều $X = \{-5.0, 23.0, 17.6, 7.23, 1.11\}$, hãy chuẩn hóa mảng sử dụng
 - ▶ a. Decimal scaling: trong khoảng $[-1, 1]$.
 - ▶ b. Min-max: trong khoảng $[0, 1]$.
 - ▶ c. Min-max: trong khoảng $[-1, 1]$.
 - ▶ d. Phương pháp độ lệch
 - ▶ e. So sánh kết quả của các dạng chuẩn trên và cho nhận xét về ưu nhược điểm của các phương pháp?

- ▶ 3. Làm mịn dữ liệu sử dụng kỹ thuật làm tròn cho tập sau: $Y = \{1.17, 2.59, 3.38, 4.23, 2.67, 1.73, 2.53, 3.28, 3.44\}$ Sau đó biểu diễn tập thu được với các độ chính xác:

- ▶ a. 0.1 b. 1.

- ▶ 4. Cho tập mẫu với các giá trị bị thiếu

$$X1 = \{0, 1, 1, 2\} \quad X2 = \{2, 1, -, 1\} \quad X3 = \{1, -, -, 0\} \\ X4 = \{-, 2, 1, -\}$$

- ▶ Nếu miền xác định của tất cả các thuộc tính là $[0, 1, 2]$, hãy xác định các giá trị bị thiếu biết rằng các giá trị đó có thể là một trong số các xác trị của miền xác định?

Thu nhỏ dữ liệu

- ▶ - Tổng hợp khối dữ liệu.
- ▶ - Lựa chọn tập thuộc tính con
- ▶ - Rút gọn chiều
- ▶ - Rút gọn số học, trong đó các dữ liệu sẽ được thay thế bằng các dữ liệu phụ nhỏ hơn nhưng cùng biểu diễn vấn đề.
- ▶ - Rời rạc và phân cấp khái niệm

Tổng hợp khối dữ liệu

- ▶ VD: việc khai phá dữ liệu lại quan tâm hơn đến các báo cáo bán hàng theo năm

Year 2004	
Quarter	Sales
Q1	0
Q2	0
Q3	0
Q4	0

Year 2003	
Quarter	Sales
Q1	0
Q2	0
Q3	0
Q4	0

Year 2002	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year	Sales
2002	\$1,568,000
2003	\$2,356,000
2004	\$3,594,000

Lựa chọn tập thuộc tính con

- Rút gọn kích thước dữ liệu bằng cách loại bỏ các thuộc tính không hữu ích hoặc dư thừa (hoặc loại bỏ các chiều)
- Tìm ra tập thuộc tính nhỏ nhất sao cho khi áp dụng các phương pháp khai phá dữ liệu thì kết quả thu được là gần sát nhất với kết quả khi sử dụng tất cả các thuộc tính.

VD

Lựa chọn tăng dần	Loại bỏ	Cây quyết định
Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}	Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}	Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}
Tập rút gọn ban đầu {}		<pre> graph TD A1((A1)) -- Y --> C1_1((Class 1)) A1 -- N --> A2((A2)) A2 -- Y --> C2_1((Class 2)) A2 -- N --> A3((A3)) A3 -- Y --> C1_2((Class 1)) A3 -- N --> C2_2((Class 2)) </pre>
$\Rightarrow$ {A1}	$\Rightarrow$ {A1, A3, A4, A5, A6}	
$\Rightarrow$ {A1, A4}	$\Rightarrow$ {A1, A4, A5, A6}	
$\Rightarrow$ Kết quả {A1, A4, A6}	$\Rightarrow$ Kết quả {A1, A4, A6}	$\Rightarrow$ Kết quả {A1, A4, A6}

2.2. Tổ chức dữ liệu

- 2.2.1. Tổ chức và cài đặt kho dữ liệu (Data warehouse)
- 2.2.2. Tổ chức và cài đặt Data Mart

Kho dữ liệu

- Khái niệm kho dữ liệu (data warehouse): công ty IBM - Barry Devlin và Paul Murphy năm 1988.
- “Kho dữ liệu là hệ thống trích xuất, làm sạch, tổng hợp và cung cấp nguồn dữ liệu vào một kho dữ liệu đa chiều sau đó hỗ trợ, thực hiện truy vấn, hoặc phân tích phục vụ cho mục đích của quá trình ra quyết định”

Đặc trưng của kho dữ liệu

- *Hướng chủ thể* (subject oriented).
- *Biến thiên theo thời gian*.
- *Tính ổn định* (non-volatile).
- *Tính tích hợp*

2.2.1. Tổ chức và cài đặt kho dữ liệu (Data warehouse)

► Hệ thống Data Warehouse

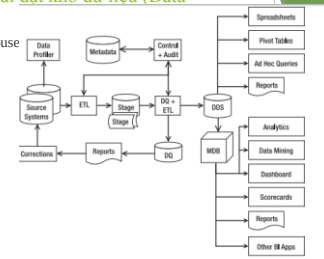


Figure 1-1. A diagram of a data warehouse system

VD về Cube

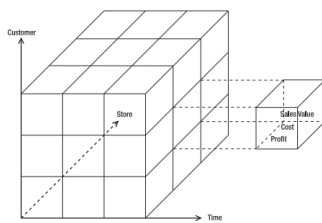
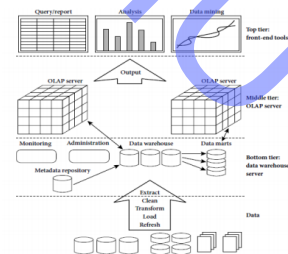


Figure 1-2. A cube with three dimensions

A Three Tier Data Warehouse Architecture



Hệ thống giao dịch vs hệ thống data warehouse

- ▶ Lưu ngắn hạn
- ▶ Không lưu lịch sử của Master data (as products, customers, branches, and vehicles) → khi thay đổi miêu tả sản phẩm nó sẽ xóa việc mô tả cũ.
- ▶ Dùng được truy vấn, dữ liệu được cập nhật bởi người dùng (ví dụ như mua hàng)
- ▶ Chạy truy vấn lâu vì DL đã chuẩn hóa – Normalization- liên kết nhiều bảng)
- ▶ Lấy DL từ 1 nguồn
- ▶ Lưu dài hạn
- ▶ Bắt buộc phải lưu thông tin của Master data
- ▶ User chỉ có thể lấy dữ liệu ở data warehouse thông qua truy vấn → ko thay đổi dữ liệu (chỉ có hệ thống ETL mới cho phép cập nhật vào data warehouse)
- ▶ Chạy truy vấn nhanh do dữ liệu chưa dc chuẩn hóa
- ▶ Truy vấn lấy DL từ nhiều nguồn

Business Intelligence

- ▶ BI là một tập các hoạt động để hiểu trạng thái kinh doanh thông qua các phân tích dữ liệu (trong và ngoài doanh nghiệp) để đưa ra các qđ kinh doanh.
- ▶ VD: Quản lý hiệu quả KD, phân tích lợi nhuận khách hàng,....
- ▶ 3 lĩnh vực chủ yếu:
 - ▶ Reporting
 - ▶ OLAP
 - ▶ Data mining

Tương lai của Kho dữ liệu

- ▶ Sử dụng dữ liệu không có cấu trúc
- ▶ Tìm kiếm trong DL không có cấu trúc
- ▶ Service-Oriented Architecture (SOA).
- ▶ Real-Time Data Warehouse

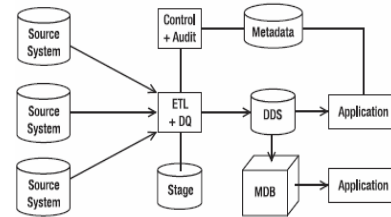
Kiến trúc của Data Warehouse

- ▶ Gồm hai thành phần chính:
 - ▶ Kiến trúc luồng dữ liệu (Data Flow): cách thức lưu trữ (kiến trúc) và cách thức cung cấp dữ liệu cho user
 - ▶ Kiến trúc hệ thống: Cấu hình vật lý của servers, network, software, storage, and clients

Data Flow Architecture

- Kiến trúc luồng dữ liệu thường bao gồm: Cách thức điều khiển, logged, giám sát, và các kỹ thuật để đảm bảo chất lượng của dữ liệu cần lưu trữ.
- Có 4 kiểu kiến trúc:
 - single DDS (dimensional data store),
 - NDS + DDS, (normalized data store + DDS)
 - ODS + DDS, (operational data store + DDS)
 - Và federated data warehouse (liên hợp kho dữ liệu)

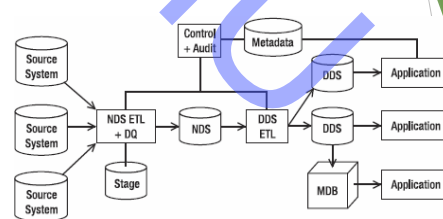
Kiến trúc DDS đơn



Ưu/Nhược

- Ưu điểm:
 - Kiến trúc đơn giản.
 - Ít công đoạn xử lý.
 - Thuận lợi khi xây dựng những kho dữ liệu nhỏ.
- Nhược điểm:
 - Không hỗ trợ việc tạo ra nhiều kho dữ liệu.
 - Không tái sử dụng được gói ETL đã làm..

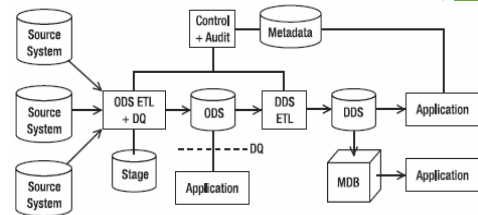
Kiến trúc NDS+DDS



Ưu/Nhược

- **Ưu điểm:**
 - Lưu trữ dữ liệu tập trung đã được làm sạch.
 - Chứa dữ liệu lịch sử.
 - Sẵn sàng cho việc nạp vào nhiều kho dữ liệu đầu cuối.
 - Tái sử dụng được các gói ETL.
- **Nhược điểm:**
 - Kiến trúc phức tạp
 - Tốn thêm không gian lưu trữ.
 - Thời gian thực hiện một chu kì nạp dữ liệu lâu hơn so với kiến trúc DDS đơn.
 - Vùng chứa dữ liệu trung gian không được tận dụng vào mục đích khác.

Kiến trúc ODS + DDS



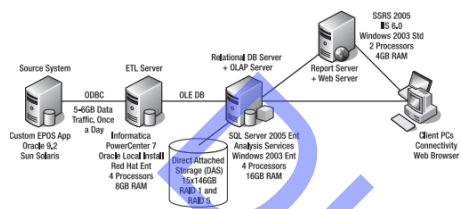
Ưu/Nhược

- **Ưu điểm:**
 - Lưu trữ dữ liệu tập trung đã được làm sạch.
 - Tận dụng làm cơ sở dữ liệu tập trung phục vụ giao tác cho ứng dụng đầu cuối.
- **Nhược điểm:**
 - Không chứa dữ liệu lịch sử.
 - Các gói ETL để đưa dữ liệu từ vùng dữ liệu hoạt động vào kho dữ liệu đầu cuối phức tạp hơn.
 - Vùng dữ liệu hoạt động có thể bị gián đoạn khi nạp kho dữ liệu.
 - Không tái sử dụng được các gói ETL.

Federated Data Warehouse



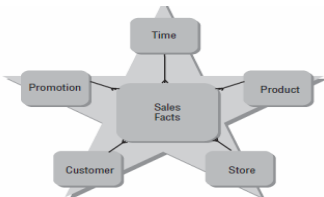
Kiến trúc hệ thống (System Architecture) - VD



Các lược đồ dữ liệu trong data warehouse

- Lược đồ hình sao (StarFlake Schema): *bảng dữ kiện, bảng chiều.*
- Lược đồ bông tuyết (SnowFlake Schema).
- Chú ý: cấu trúc phi chuẩn của các bảng chiều trong lược đồ hình sao có thể thích hợp hơn cho việc duyệt các chiều.

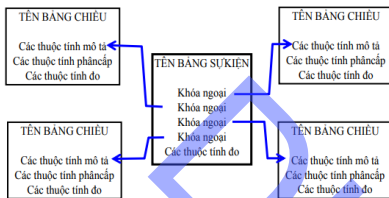
Lược đồ hình sao (StarFlake Schema)



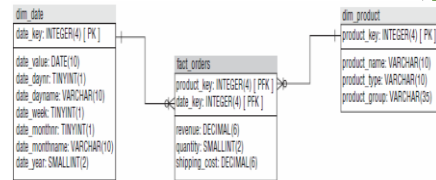
Các thành phần của lược đồ hình sao

- Bảng sự kiện và nội dung của nó: Những thuộc tính và các khóa ngoại cần thiết để liên kết các bảng chiều
- Các bảng chiều và nội dung của nó: các thuộc tính và các thuộc tính tham chiếu. Các bảng chiều đã được chuẩn hóa
- Các đường kết nối bảng chiều với bảng sự kiện

Lược đồ



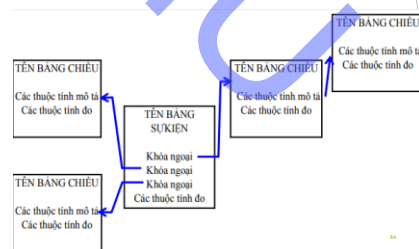
VD



Lược đồ bông tuyết

- 4 thành phần chính:
 - – Bảng sự kiện và nội dung của nó: Những thuộc tính và các khóa ngoại cần thiết để liên kết các bảng chiều
 - – Các bảng chiều và nội dung của nó: các thuộc tính và các thuộc tính tham chiếu, các bảng chiều đã được chuẩn hóa
 - – Các đường kết nối bảng chiều với bảng sự kiện

Lược đồ



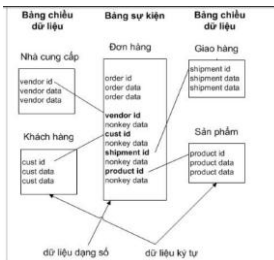
Thiết kế DL

- Bảng sự kiện: là một bảng CSDL quan hệ với các đặc điểm:
- – Thường là bảng lớn nhất
- – Có khóa ngoài là các khóa chính của các bảng chiều.
- – Số lượng bản ghi tăng lên nhanh chóng
- – Khóa chính gồm nhiều phần, mỗi phần tham chiếu đến một bảng chiều mà dữ liệu sự kiện có thể truy nhập.
- – Dữ liệu chuẩn hóa ở mức cao.

Bảng chiều (Dimension Table)

- Là một bảng trong CSDL quan hệ có một phần khóa được kết nối với bảng sự kiện. Các bảng chiều khác nhau kết hợp với bảng sự kiện để trả lời cho các câu hỏi về nghiệp vụ.
- Đặc điểm:
 - – Chứa các dữ liệu có quan hệ tĩnh trong kho dữ liệu.
 - – Lưu trữ các chiều thông tin thường sử dụng trong truy vấn.

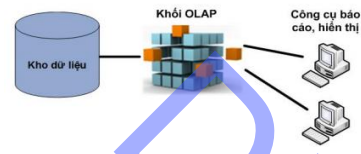
VD



Hoặc lược đồ bông tuyết



OLAP trong kho dữ liệu



OLAP

- ▶ Là hướng tiếp cận để phân tích dữ liệu đa chiều. Nó là một trong các công cụ của BI sử dụng CSDL quan hệ.
- ▶ OLAP gồm có 3 chức năng chính:
 - ▶ Consolidation (Roll-Up)
 - ▶ Drill-Down
 - ▶ Slicing And Dicing

Types of OLAP

- ▶ Relational OLAP (ROLAP).
- ▶ Multidimensional OLAP (MOLAP)
- ▶ Hybrid OLAP (HOLAP)

2.2.2. Tổ chức và cài đặt Data Mart

-Kho dữ liệu chuyên đề

- ▶ Là CSDL có những đặc điểm giống với kho dữ liệu nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chuyên ngành.
- ▶ Các DM thường được xây dựng và điều khiển bởi các phòng ban riêng của một tổ chức.
- ▶ Các DM có thể được kết nối tích hợp lại với nhau tạo thành kho dữ liệu.
- ▶

So sánh DW và DM

► Data Warehouse:

- Phù trong nhiều chủ đề
- Chứa các thông tin chi tiết
- Làm việc từ nhiều nguồn

► Data Mart

- Chỉ chứa 1 chủ đề: Finance, or Sales
- Có thể thống kê dữ liệu
- Tập trung vào việc tích hợp thông tin từ chủ đề cho trước, hoặc nguồn tài nguyên của hệ thống.
- Được xây dựng dựa trên lược đồ sao.

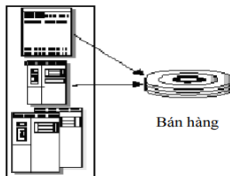
Các dạng cơ bản của Data Mart

- Hai dạng cơ bản của DM là kho dữ liệu chuyên đề độc lập (Independent data marts) và kho dữ liệu chuyên đề phụ thuộc (Dependent data marts)

	Data Warehouse	Data Mart
Scope	Corporate	Line-of-Business (LoB)
Subjects	Multiple	Single Subject
Data Sources	Many	Few
Size (typical)	100 GB-TB+	< 100GB
Implementation Time	Months to years	Months

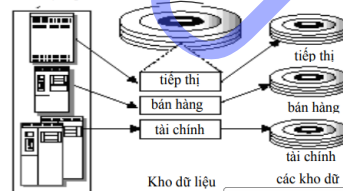
Kho dữ liệu chuyên đề phụ thuộc

Hệ thống tác nghiệp



Kho dữ liệu chuyên đề độc lập (Independent data marts)

Hệ thống tác nghiệp

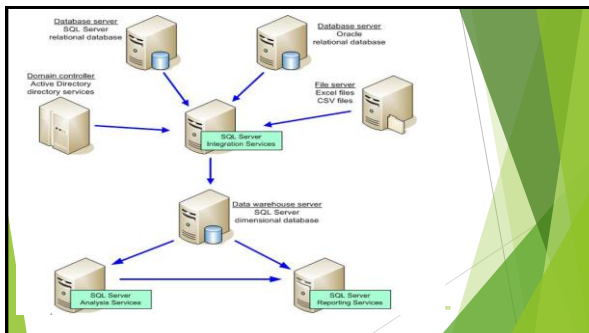


Chương 3: Sử dụng công cụ OLAP trong khai phá dữ liệu

- ▶ 3.1. Tổng quan chung về SQL server analysis service
 - ▶ 3.1.1. Một số khái niệm liên quan
 - ▶ 3.1.2. Tổ chức mô hình cơ sở dữ liệu đa chiều
 - ▶ 3.1.3. Giới thiệu và cài đặt công cụ OLAP
- ▶ 3.2. Xây dựng OLAP Cube
 - ▶ 3.2.1. Xây dựng DB Engine
 - ▶ 3.2.2. Thiết lập nguồn , bảng Fact
 - ▶ 3.2.3. Thiết lập các chiều cho Cube
 - ▶ 3.2.4. Xây dựng Cube
- ▶ 3.3. Xây dựng Report
 - ▶ 3.3.1. Thiết lập nguồn cho Report
 - ▶ 3.3.2. Lập Report từ cơ sở dữ liệu nhiều chiều
 - ▶ 3.3.3. Quản lý Report (bảo mật- security, truyềnsubscription, quản lý thực thi execution).

3.1. Tổng quan chung về SQL server analysis service

- ▶ 3.1.1. Một số khái niệm liên quan
- ▶ Microsoft SQL Server Analysis Services (SSAS):
 - ▶ OLAP
 - ▶ Data Mining Function
- ▶ Ứng dụng của Data Mining: SSAS cho phép thiết kế, tạo, và mô hình hóa mô hình từ các nguồn khác nhau thông qua việc dùng các thuật toán của Data Mining



Yêu cầu về thiết lập môi trường

- ▶ Các thành phần và công cụ
- ▶ SQL Server Database Engine: có trong bộ SQL Server
- ▶ SQL Server Analysis Services: Có trong bộ SQL Server
- ▶ Business Intelligence Studio: Công cụ này sẽ có được khi cài SQL Server phiên bản Full (SQL Server 2008)

Các KN khác

- **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
 - chiều hàng hóa (Item)
 - chiều thời gian (time)
 - chiều chi nhánh (Branch) mô tả thông tin

Branch	Item	January	June	November
Chicago	Q1	100	125	14
Chicago	Q2	100	102	31
Chicago	Q3	112	1023	30
Chicago	Q4	107	1038	38
New York	Q1	100	125	14
New York	Q2	100	102	31
New York	Q3	112	1023	30
New York	Q4	107	1038	38
Toronto	Q1	100	125	14
Toronto	Q2	100	102	31
Toronto	Q3	112	1023	30
Toronto	Q4	107	1038	38

- **Bảng Fact:** bao gồm các độ đo, sự kiện trong kinh doanh
- **Bảng chiều (Dimension Table):** là bảng bao gồm các thuộc tính khác nhau giải thích về khóa chiều ở bảng Fact.
- VD

Branch	Item	January	June	November
Chicago	Q1	100	125	14
Chicago	Q2	100	102	31
Chicago	Q3	112	1023	30
Chicago	Q4	107	1038	38
New York	Q1	100	125	14
New York	Q2	100	102	31
New York	Q3	112	1023	30
New York	Q4	107	1038	38
Toronto	Q1	100	125	14
Toronto	Q2	100	102	31
Toronto	Q3	112	1023	30
Toronto	Q4	107	1038	38

3.1.2. Tổ chức mô hình cơ sở dữ liệu đa chiều

- **Thiết kế Data Mart**
 - Một sự kiện bán hàng xảy ra khi có 1 người khách mua hàng.
 - The roles (vai trò) –who, where, what, của sự kiện này là : customer, product, và store.
 - Còn levels (hay còn gọi là độ đo (measures)) là quantity, unit price, value, direct unit cost, và indirect unit cost.

VD

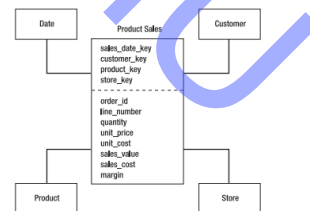


Figure 5-1. Product Sales fact table and its dimensions

Thêm các cột thời gian

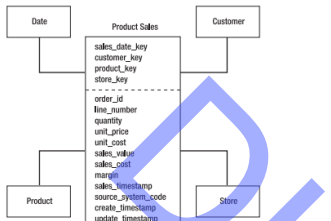


Figure 5-2. Product Sales table (and its dimensions) with timestamp columns

Thiết kế CSDL vật lý

- Xem xét về bộ nhớ lưu trữ: được tính theo kích cỡ của bảng Fact và bảng chiều để ra kích cỡ của DDS. Sau đó, tính kích cỡ của NDS dựa trên số dòng DL và kích thước của mỗi bảng. Tính kích cỡ của toàn CSDL dựa trên bảng nguồn và phương pháp dùng truy vấn.
- VD: Giả sử Product Sales fact table có 8 int columns, 1 decimal column, 10 money columns, và 3 datetime columns (22 cột)
- Integers are 4 bytes, decimals are 5 bytes, and money and datetime are 8 bytes,

- Vì vậy: row size là $8 \times 4 + 1 \times 5 + (10 + 3) \times 8 = 141$ bytes.
- Cho phép 50 % số cột trống cho tương lai thì số bytes sẽ cần là 212 bytes (10 cột x 8 bytes là lớn nhất)- tùy vào người thiết kế.
- Giả sử hệ thống chỉ ra rằng trung bình mỗi ngày có khoảng 500,000 giao dịch.
- Cần phải chuẩn bị lưu DL cho 2 năm thì kích thước của bộ nhớ lưu trữ DL cho bảng Fact sẽ là:
- $600,000 \times 365 \times 2 \times 212 \text{ bytes} = 86\text{GB}$.

Tạo DDS

```
create database DDS
on primary (name = 'dds_fg1'
, filename = 'h:\disk\data1\dds_fg1.ndf'
, size = 30 GB, filegrowth = 5 GB)
, filegroup dds_fg2 (name = 'dds_fg2'
, filename = 'h:\disk\data2\dds_fg2.ndf'
, size = 30 GB, filegrowth = 5 GB)
, filegroup dds_fg3 (name = 'dds_fg3'
, filename = 'h:\disk\data3\dds_fg3.ndf'
, size = 30 GB, filegrowth = 5 GB)
, filegroup dds_fg4 (name = 'dds_fg4'
, filename = 'h:\disk\data4\dds_fg4.ndf'
, size = 30 GB, filegrowth = 5 GB)
, filegroup dds_fg5 (name = 'dds_fg5'
, filename = 'h:\disk\data5\dds_fg5.ndf'
, size = 30 GB, filegrowth = 5 GB)
, filegroup dds_fg6 (name = 'dds_fg6'
, filename = 'h:\disk\data6\dds_fg6.ndf'
, size = 30 GB, filegrowth = 5 GB)
log on (name = 'dds_log'
, filename = 'h:\disk\log1\dds_log.ldf'
, size = 1 GB, filegrowth = 512 MB)
collate SQL_Latin1_General_CP1_CI_AS
```

Creating DDS Database Structure

- ▶ Open SQL Server 2005 Management Studio, connect to your development SQL Server instance, and start creating the tables.
- ▶ create table dim_date
- ▶ Tạo các bảng chiều.
- ▶ VD: create table dim_product (product_key int not null , product_code varchar(10) , name varchar(30))...
- ▶ constraint pk_dim_product primary key clustered (product_key)
- ▶ on dds_fg6)
- ▶ on dds_fg6
- ▶ go

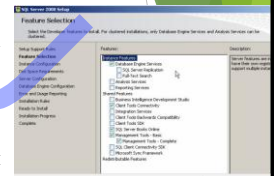
- ▶ Tạo bảng Fact
- ▶ use DDS
- ▶ go
- ▶ if exists (select * from sys.objects where name = 'fact_product_sales' and type = 'U')
- ▶ drop table fact_product_sales go
- ▶ create table fact_product_sales (sales_date_key int not null , customer_key int not null , product_key int not null , store_key int not null , order_id int not null , line_number int not null , quantity int , ...
- ▶ , constraint pk_fact_product_sales primary key clustered (order_id, line_number) on dds_fg4 ,
- ▶ constraint fk_fact_product_sales_dim_date foreign key (sales_date_key) references dim_date(date_key) ,
- ▶ constraint fk_fact_product_sales_dim_customer foreign key (customer_key) references dim_customer(customer_key) , ...
- ▶ go

Tạo các Indexes

- ▶ create index fact_product_sales_sales_date_key on fact_product_sales(sales_date_key)
- ▶ on dds_fg4
- ▶ go
- ▶ create index fact_product_sales_customer_key on fact_product_sales(customer_key) on dds_fg4
- ▶ go
- ▶ create index fact_product_sales_product_key on fact_product_sales(product_key) on dds_fg4
- ▶ go
- ▶ create index fact_product_sales_store_key on fact_product_sales(store_key) on dds_fg4
- ▶ go

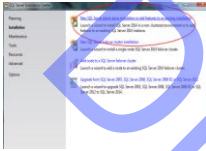
3.1.3. Giới thiệu và cài đặt công cụ OLAP

- ▶ Cài đặt SQL Server 2008 phiên bản Developer hoặc phiên bản Enterprise Edition đầy đủ và khi cài đặt chọn mục "SQL Server Database Services" và "Analysis Services".
- ▶ Công cụ cho phép thực hiện OLAP là "SQL Server Business Intelligence Development Studio - BIDS".
- ▶ Chú ý tại Feature Selection chọn tham số BI và Reporting Services, Analysis Services



Một số chú ý

- ▶ Khi cài SQL nên chọn New SQL Server New SQL Server stand-alone installation
- ▶ Khi không kết nối được với Server thì chọn lại ở Management Studio với TCP/IP là Enable



3.2. Xây dựng OLAP Cube

- ▶ Mở Business Intelligence Studio: Vào Start – All Programs – Microsoft SQL Server 2008 và mở công cụ SQL Server Business Intelligence Development Studio
- ▶ Chọn menu File – New – Project
- ▶ Trong cửa sổ tạo mới Project, chọn Business Intelligence Projects và chọn Project Types, chọn Analysis Service Project trong pane Templates
- ▶ Đặt tên cho Project Name, và thiết lập tương tự cho Solution Name,
- ▶ chọn OK

3.2.1. Xây dựng DB Engine

- ▶ **Định nghĩa một Data Source**
- ▶ Trong khung Solution Explorer, phải chuột Data Sources và chọn New Data Source
- ▶ Tại màn hình Welcome to Data Source Wizard, nhấn Next để chuyển sang cửa sổ Select how to define the connection
- ▶ Tại cửa sổ Select how to define the connection page, có thể định nghĩa một data source bằng cách tạo mới connection hoặc trên một connection đã có sẵn thông qua: **Create a data source based on an existing or new connection**
- ▶ nhấn New

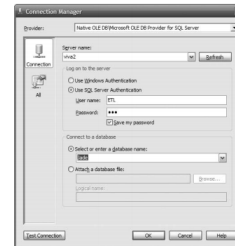
Định nghĩa một Data Source

- ▶ Trong dialog box Connection Manager, định nghĩa giá trị thuộc tính cho Data Source.
- ▶ Trong danh sách Provider, chọn **Native OLE DB\SQL Server Native Client 10.0**
- ▶ Trong Server Name, gõ vào localhost
- ▶ Kiểm tra giá trị **Use Windows Authentication** được chọn, trong danh sách **Select or enter a database name**, chọn AdventureWorksDW2008
- ▶ Nhấn Test Connection để kiểm tra kết nối đến với SQL Server
- ▶ Chọn OK và nhấn Next

Định nghĩa một Data Source

- Tại cửa sổ Impersonation Information, thiết lập Security Credential cho SSAS để kết nối tới Data Sources, chọn **Use the Service account**, và nhấn Next.
- Tại màn hình Completing the wizard, thiết lập tên của Data Source là **Adventure Work DW** và chọn Finish

Kết nối Data Source



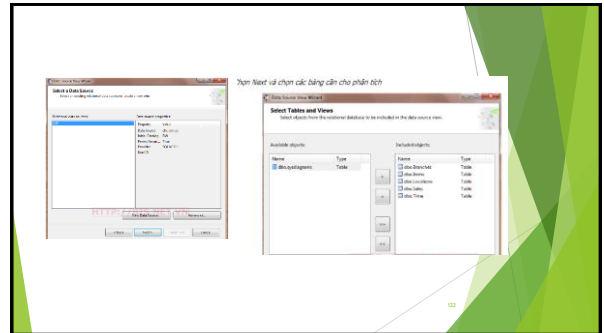
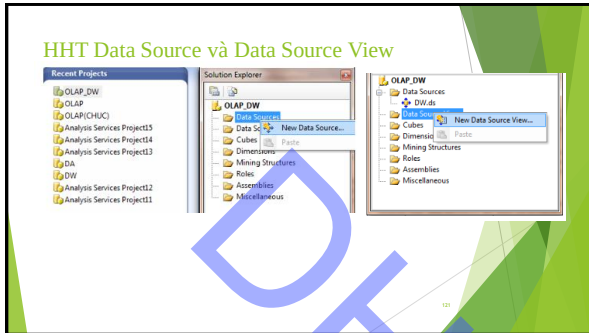
Định nghĩa Data Source View

- Trong khung Solution Explorer, phải chuột Data Sources Views, và chọn New Data Source Views
- Cửa sổ Welcome to Data Source View Wizard xuất hiện, nhấn Next
- Dialog Select data source xuất hiện, tại mục Relational Data Sources, chọn Adventure Work DW mà ta đã tạo như trên rồi chọn Next.

Định nghĩa Data Source View

- Tại Dialog Select Tables and Views, chọn các bảng dữ liệu và views từ danh sách đối tượng,
- VD:
 - DimCustomer (dbo)
 - DimDate (dbo)
 - DimGeography (dbo)
 - DimProduct (dbo)
 - FactInternetSales (dbo)
- Nhấn nút ">>" để thêm các bảng vào danh sách Included Objects
- Nhấn Next

HHT Data Source và Data Source View



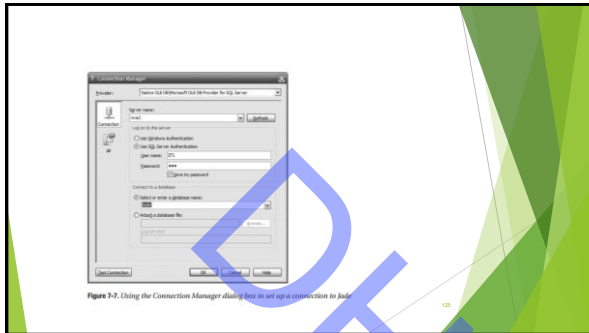
3.2.2. Thiết lập nguồn , bảng Fact

- ▶ Ví dụ:
- ▶ Open Business Intelligence Development Studio. Click File ► New ► Project, select Integrated Services Project, and name the project (vd như Amadeus ETL). Click OK to create the project.



Set up a data source

- ▶ 1. Data Sources, →New Data Source.
- ▶ Click New.
- ▶ 2. Chọn SQL Native Client.
- ▶ 3.Chọn Use SQL Server Authentication,
 - ▶ Gõ ETL tại username và gõ password,
 - ▶ “Save my password”
 - ▶ Có thể chọn Use Windows Authentication.
- ▶ 4. Test Connection, click OK.
- ▶ 5. click Next. click Finish.



Lặp lại các bước trên để chọn nguồn khác (nếu có)

- ▶ 1. Choose SQL Native Client as the provider.
 - ▶ 2. Type the server name, username, and password the same as before.
 - ▶ 3. Enter the database name (VD Stage).
 - ▶ 4. Click Test Connection, and click OK if succeeds.
 - ▶ 5. Make sure servername.Stage.ETL in the Data Connections box is selected, and click Next.
 - ▶ 6. Name it Stage, and click Finish.
- ▶ Chú ý: có thể xem có bao nhiêu Data Source ở Explore Panel.

Chọn Data Source từ OLE DB

- ▶ Nháy kép chuột the Data Flow task in the Toolbox on the left side.
- ▶ A Data Flow task will appear in the top-left corner of the design surface.
- ▶ Nháy kép chuột the Data Flow task in the design surface, and the design surface will change from Control Flow to Data Flow.
- ▶ Nháy kép chuột OLE DB Source under Data Flow Sources in the Toolbox on the left side. The OLE DB Source box will appear in the design surface.

Data Flow Task

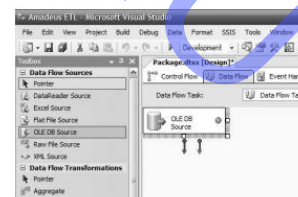


Figure 7-8: Creating the OLE DB source in the Data Flow task

Tạo truy vấn từ OLE_DB Source

- ▶ Từ màn hình Data Flow Task → Kích đúp chuột → Xuất hiện OLE DB Connection Manager, click New.
- ▶ 5. Select tên của nguồn (VD servername.Jade.ETL), OK
- ▶ 6. In the Data Access Mode drop-down list, choose SQL Command.
- ▶ 7. Click Build Query, and Query Builder window opens.
- ▶ 8. Click the rightmost icon, Add Table.
- ▶ 9. Click Close.
- ▶ 10. In the SELECT FROM section of the window, type select count(*) from tên của bảng, and click OK.
- ▶ 11. Click Preview.

Một số chú ý

- ▶ Khi Deploy cube không được mà có thông báo là OLEDB error do không có quyền kết nối. Quay lại SQL server → chọn tên Database → Properties
- ▶ HHT Database Properties → View Server Permission
- ▶ Chọn Tên Account → click chọn các quyền của Grand phía dưới
- ▶ Khi chọn measure cho Cube nên chọn kiểu Number

3.3. Xây dựng Report

- ▶ 3.3.1. Thiết lập nguồn cho Report
- ▶ 3.3.2. Lập Report từ cơ sở dữ liệu nhiều chiều
- ▶ 3.3.3. Quản lý Report (bảo mật- security, truyền- subscription, quản lý thực thi –execution).

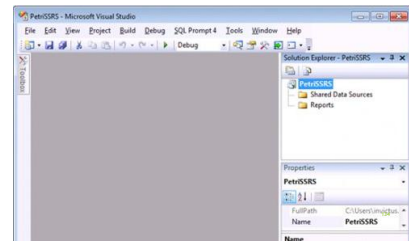
Quy trình



3.3.1. Thiết lập nguồn cho Report

- Mở SQL Server Business Intelligence Studio
- *File > New > Project và Report Server Project* từ danh sách *Business Intelligence Projects*,
- gõ tên cho project và nhấn OK

Tạo mới Report Project



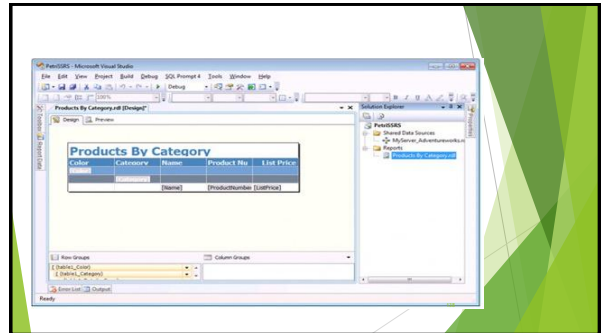
- Trong khung **Solution Explorer** click chuột phải lên **Shared data Sources** chọn **new**.
- Chọn nguồn trong HHT



Thiết kế query cho Report

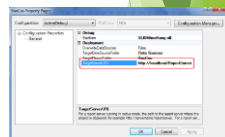
- Màn hình *Design the Query* : gõ các câu lệnh truy vấn.
- Hoặc dùng chức năng *Query Builder* để xây dựng với giao diện đồ họa.
- VD
- ```
SELECT P.Name ,
ProductNumber ,
Color ,
ListPrice ,
SC.Name [Category]
FROM Production.Product P
LEFT OUTER JOIN Production.ProductSubCategory SC
ON P.ProductSubCategoryID = SC.ProductSubCategoryID
WHERE P.ProductSubCategoryID IS NOT NULL
AND P.Color IS NOT NULL
ORDER BY Category, ListPrice ASC
```

- *Select the Report Type*, chọn *Tabular*, *Next*
- *Design the Table*, thêm *Color* và *Category* vào trong mục *Group*, hãy thêm các trường còn lại vào mục *Details*, nhấn *Next*
- chọn *Stepped* tại bước *Choose the Table Layout* và *Next*. Tiếp theo, chọn kiểu cho bảng báo cáo và tiếp tục *Next*. Cuối cùng là đặt tên (ví dụ *Products By Category*) và *Finish*



### Thực thi báo cáo (deploy)

- Nhấp phải chuột lên tên dự án trong khung **Solution Explorer** và chọn **properties**.
- Chọn **targetServerUrl** đánh địa chỉ nơi chứa dữ liệu các báo cáo trên webservice.
- VD: địa chỉ đích được cấu hình ở Report Server là : <http://localhost/reportserver>.
- Cuối cùng nhấp chuột phải vào project
- Chọn **Deploy**.



### Truy xuất báo cáo

- Khởi động **SQL Server Management Studio**.
- Nhấp vào nút **Connect** trong **Object Explorer** và chọn dịch vụ **Reporting Service**.
- Kết nối với máy server của bạn.
- Mở rộng nút **Home** trong cây **Object Explorer** để tìm tên dự án, ví dụ **ProductReport1**.
- Mở rộng tên dự án **ProductReport1** để tìm ra báo cáo có tên **ProductReport1**.
- Nhấp chuột phải vào báo cáo và chọn **View Report**

### 3.3.2. Lập Report từ cơ sở dữ liệu nhiều chiều

- Tương tự như trên, tuy nhiên:
  - Trong khung **Solution Explorer** click chuột phải lên **Shared data Sources** chọn **new**.
  - Trong khung hiện ra chọn kiểu kết nối là **SQL Server Analysis Service**

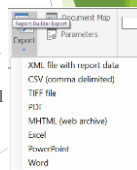


### 3.3.3. Quản lý Report (bảo mật- security, truyền-subscription, quản lý thực thi –excution).

- Có 4 cách thức bảo mật:
  - Login và user: cho phép truy cập vào hệ thống thông qua những login.
  - Server role: một tập hợp 1 nhóm những quyền và được đại diện bằng 1 tên thể có thể thuận tiện trong việc quản lý.
  - Database role
  - Authentication: SQL Server authentication và Windows authentication

### Export báo cáo

- Để export báo cáo ra các định dạng khác nhau như PDF, Excel... :
- Chọn thẻ **Preview**
- Nhấn nút **Export** ở phía bên trên:
- Sau đó chọn định dạng phù hợp, đường dẫn lưu file và nhấn **Save**.
- Chú ý: Có thể Export lên Share Point Hoặc Web Portal



### Chương 4: Một số bài toán khai phá trong kinh doanh \_ Bài tập nhóm

- 4.1. Phân tích dữ liệu Marketing
- 4.2. Quản lý chăm sóc khách hàng
- 4.3. Phân tích dữ liệu bán hàng

#### 4.1. Phân tích dữ liệu Marketing

- ▶ 1. Đặt bài toán, xác định mục tiêu
- ▶ 2. XD CSDL
- ▶ 3. Phân tích với OLAP
- ▶ 4. XD báo cáo

#### 4.2. Quản lý chăm sóc khách hàng

- ▶ 1. Đặt bài toán, xác định mục tiêu
- ▶ 2. XD CSDL
- ▶ 3. Phân tích với OLAP
- ▶ 4. XD báo cáo

#### 4.3. Phân tích dữ liệu bán hàng

- ▶ 1. Đặt bài toán, xác định mục tiêu
- ▶ 2. XD CSDL
- ▶ 3. Phân tích với OLAP
- ▶ 4. XD báo cáo