Trường Đại học CNTT – Khoa Hệ thống thông tin

BÀI 2: TỔNG QUAN KHO DỮ LIỆU

Giảng viên: ThS. Nguyễn Thị Kim Phụng

Nội dung

- Nhu cầu doanh nghiệp, hệ thống hỗ trợ quyết định
- Mục đích xây dựng kho dữ liệu
- Các đặc tính của kho dữ liệu
- Các thành phần kho dữ liệu
- Ứng dụng của kho dữ liệu

Business Intelligence Defined

Business intelligence (BI) is a broad category of applications and technologies for ***gathering, storing, analyzing****, and providing access to data to help enterprise users make better business decisions.*

Bert Brijs, Business Analysis for Business Intelligent

"Create value and competitive advantage through careful mining and analysis of your company's business data"

Philo Janus and Guy Fouché

Dẫn nhập

- Hệ thống OLTP (On-Line Transaction Processing – Xử lý giao dịch trực tuyến)
  - Dữ liệu phát sinh từ các hoạt động hàng ngày.
  - Thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức
  - Thường được gọi là dữ liệu tác vụ và hoạt động thu thập xử lý dữ liệu này

Khái niệm

- Kho dữ liệu, trái lại:
  - Phục vụ cho việc phân tích với kết quả mang tính thông tin cao
  - Xử lý dữ liệu phân tích trực tuyến (OLAP – Online Analytical Processing – Xử lý phân tích trực tuyến)
  - Hỗ trợ ra quyết định trong quản lý

Sự ra đời và phát triển

- Cuối những năm 80, kho dữ liệu bắt đầu xuất hiện.
- Năm 1988, có một bài báo mô tả định nghĩa đầu tiên về kiến trúc kho dữ liệu.
- Đầu thập niên 90, cuộc cách mạng về xử lý dữ liệu không chỉ là phổ cập kho dữ liệu mà còn tạo điều kiện để mở rộng khái niệm kho dữ liệu.

Sự ra đời và phát triển

- Thế kỷ 20 – kỷ nguyên của quản lý dựa trên thông tin.
- Ngày nay, chúng ta chờ đợi và dự đoán tương lai dựa trên những phác thảo quá khứ.

Mục đính xây dựng kho dữ liệu

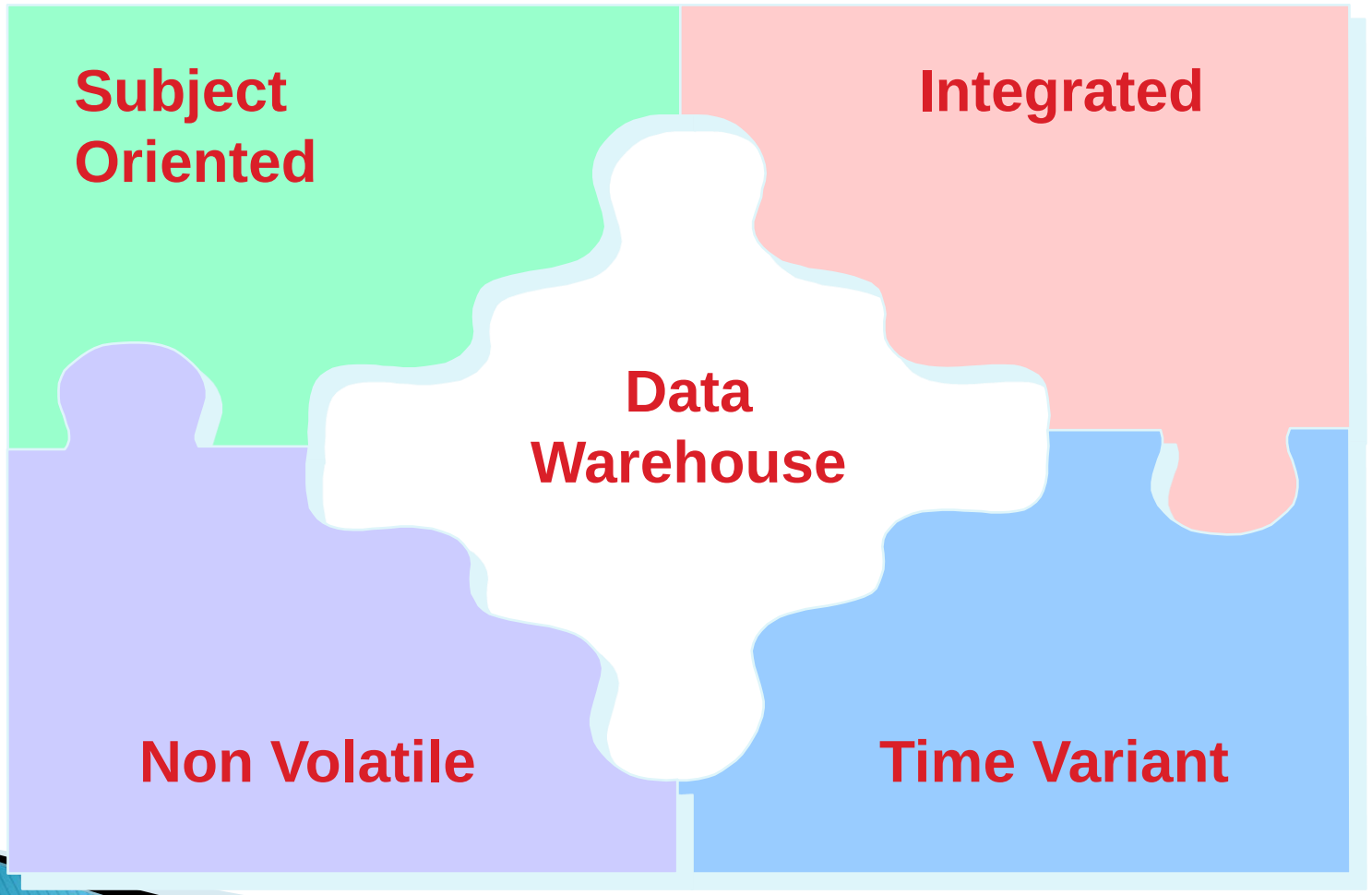
- Kho lưu trữ dữ liệu, thông tin, tri thức, và siêu dữ liệu
  - Tổng hợp toàn bộ thông tin phục vụ cho phân tích sâu, phức tạp
  - Tách việc phân tích dữ liệu ra khỏi xử lý nghiệp vụ

- Chuyển đổi dữ liệu thành thông tin
  - Cung cấp thông tin chính xác đúng thời điểm và đúng định dạng

Mục đính xây dựng kho dữ liệu

- Thi hành các phân tích dữ liệu phức tạp
- Thực hiện phân tích:
  - Phân tích định hướng
  - Phân tích chuỗi thời gian
  - Phân tích rủi ro
  - Thăm dò các hệ hỗ trợ quyết định
  - Khám phá và đưa ra các yếu tố ẩn thông qua các kĩ thuật khai phá dữ liệu

Đặc tính kho dữ liệu (1)



Đặc tính kho dữ liệu (2)

- Kho dữ liệu là nơi dữ liệu được chọn lọc và lưu trữ:
  - Hướng chủ đề
  - Tích hợp
  - Dữ liệu lịch sử (Biến đổi theo thời gian)
  - Ổn định.

1.Hướng chủ đề (Subject Oriented)

- Kho dữ liệu được thiết kế để hỗ trợ trong việc phân tích dữ liệu.
- Được tổ chức xung quanh các chủ đề chính như: khách hàng, sản phẩm, bán hàng, …
- Loại bỏ những dữ liệu không hữu ích cho trình ra quyết định.

→ Việc này giúp cho người dùng xác định được những thông tin cần thiết trong hoạt động của mình.

2.Tích hợp (Integrated) (1)

- Là đặc tính quan trọng nhất của kho dữ liệu.
- Dữ liệu được tập hợp từ nhiều nguồn khác nhau.
  - Cơ sở dữ liệu quan hệ (relational databases), flat files, các bảng ghi toàn tác trực tuyến.

→Điều này sẽ dẫn đến việc quá trình tập hợp dữ liệu phải thực hiện việc làm sạch, sắp xếp, rút gọn dữ liệu.

2.Tích hợp (Integrated) (2)

- Là đặc tính quan trọng nhất của kho dữ liệu.
- Dữ liệu được tập hợp từ nhiều nguồn khác nhau.
  - Cơ sở dữ liệu quan hệ (relational databases), flat files, các bảng ghi toàn tác trực tuyến.

→Điều này sẽ dẫn đến việc quá trình tập hợp dữ liệu phải thực hiện việc làm sạch, sắp xếp, rút gọn dữ liệu.

3.Ổn định (Non Volatile)

- Được lấy từ nhiều nguồn dữ liệu của hệ thống tác nghiệp có sẵn
- Kho dữ liệu tách rời vật lý với môi trường tác nghiệp, nên dữ liệu trong kho dữ liệu là dữ liệu chỉ đọc, không chỉnh sửa hoặc thêm mới được.
- Dữ liệu trong kho dữ liệu rất lớn và không được thêm, xóa, sửa dữ liệu.

4.Biến đổi theo thời gian (Time Variant)

- Biển đổi theo thời gian (Time Variant)
  - Dữ liệu quá khứ và hiện tại
  - Mỗi dữ liệu trong kho dữ liệu đều được gắn với thời gian và có tính lịch sử.

Mục đích của kho dữ liệu (1)

- Kho dữ liệu phải đáp ứng những yêu cầu sau:
  - Phải có khả năng đáp ứng mọi yêu cầu về thông tin của người sử dụng.
  - Hỗ trợ để các nhân viên của tổ chức thực hiện tốt, hiệu quả công việc của mình.
  - Giúp cho tổ chức, xác định, quản lý và điều hành các dự án, các nghiệp vụ một cách hiệu quả và chính xác.

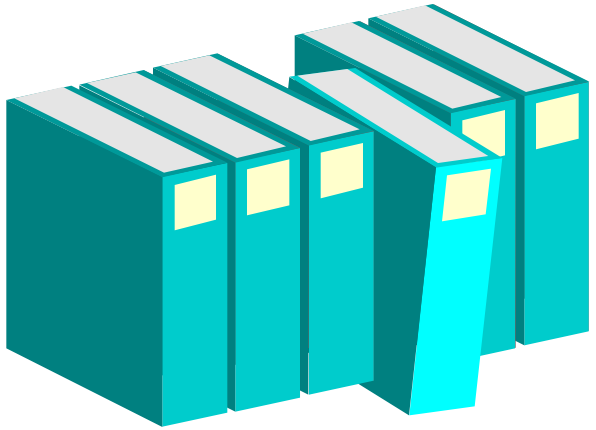
Mục đích của kho dữ liệu (2)

- Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau.
- Dùng trong các hệ thống hỗ trợ quyết định (DSS), các hệ thống thông tin tác nghiệp hoặc hỗ trợ cho các truy vấn đặc biệt

Các thành phần của kho dữ liệu

Methodology – Phương pháp luận

- Đảm bảo sự thành công của KDL
- Thúc đẩy việc phát triển
- Cung cấp một hướng ổn định cho KDL lớn
  - An toàn
  - Quản lí được
  - Kiểm chứng được
  - Ấn tượng tốt

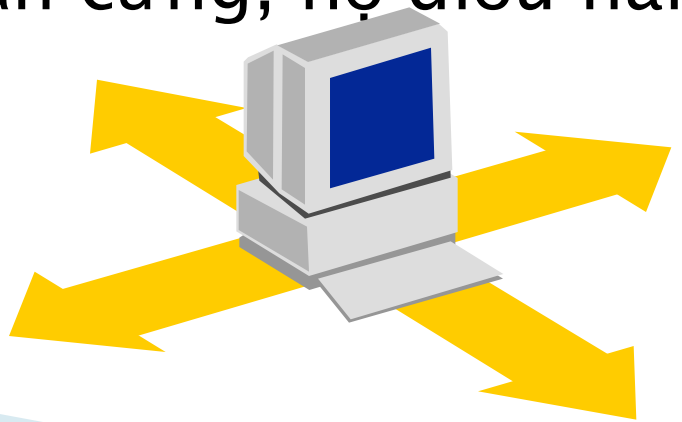


Modeling – Mô hình hóa

- Các điểm khác của KDL so với các hệ thống OLTP
  - Thiết kế các thành phần phục vụ các yêu cầu phân tích
  - Định hướng chủ thể
- Dữ liệu được ánh xạ vào thông tin hướng chủ thể:
  - Nhận dạng các chủ thể kinh doanh
  - Định nghĩa quan hệ giữa các chủ thể
- Mô hình hóa là một quá trình lặp

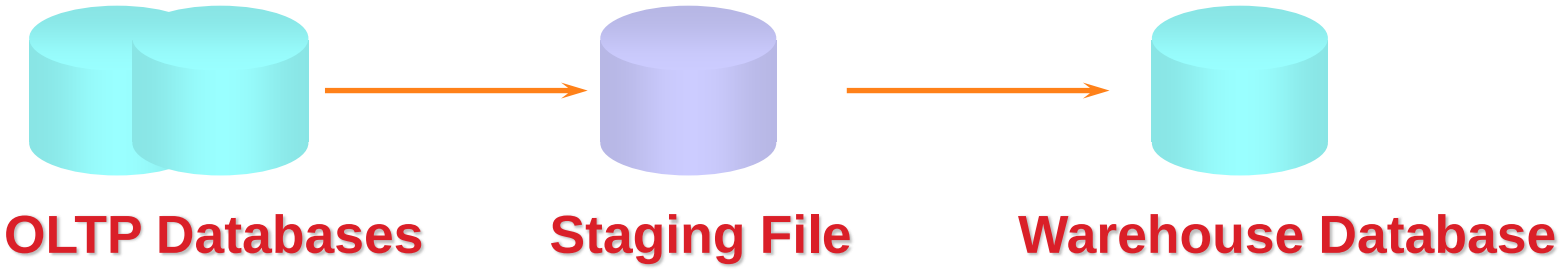
Data Management Tools

- Các công cụ phục vụ cho việc quản lý dữ liệu một cách hiệu quả (ETL, OLAP, SSRS, Data mining)
- Các yêu cầu
  - Mềm dẻo
  - Tự động
  - Hiệu quả
- Quản lí phần cứng, hệ điều hành và mạng

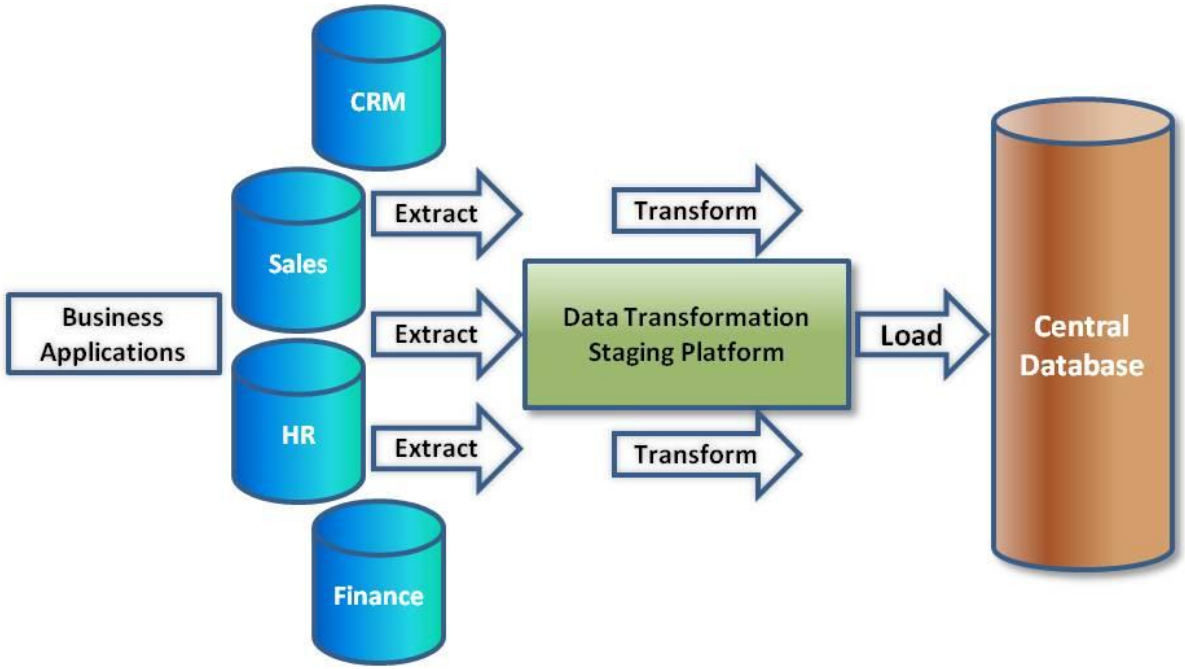


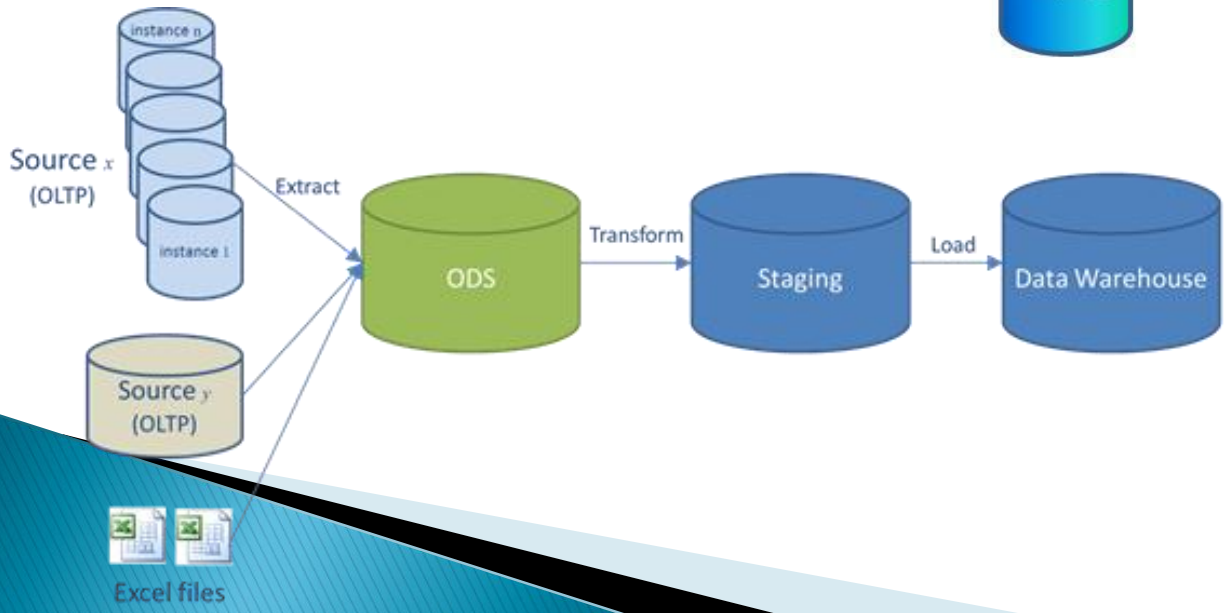
ETL (Extract, Transform, Load) – Dịch vụ tích hợp dữ liệu

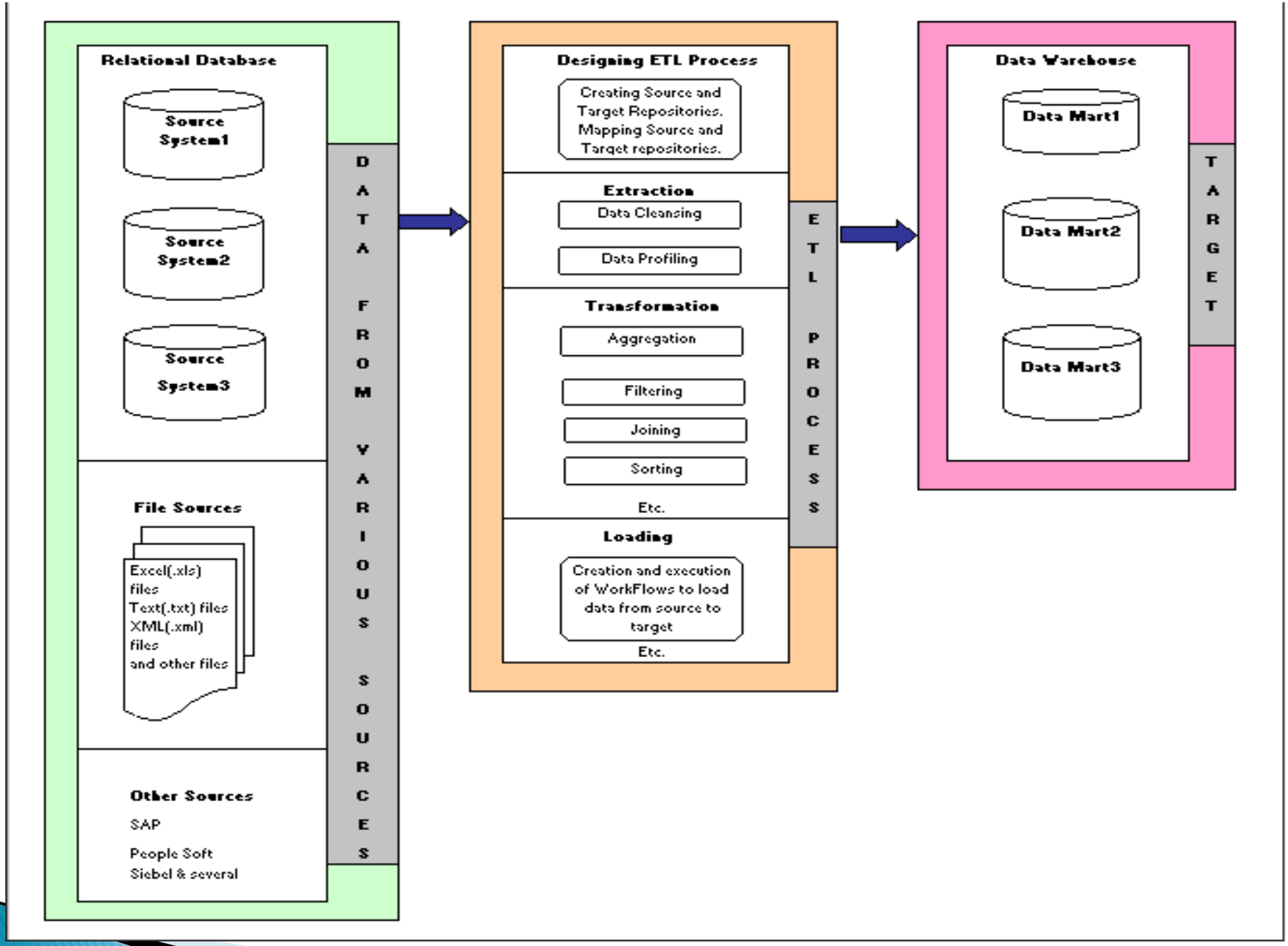
- Extract: chọn lựa dữ liệu bằng nhiều phương thức
- Transform: xác nhận hợp lệ, làm sạch, tích hợp, và dữ liệu nhãn thời gian
- Load: nạp dữ liệu vào KDL



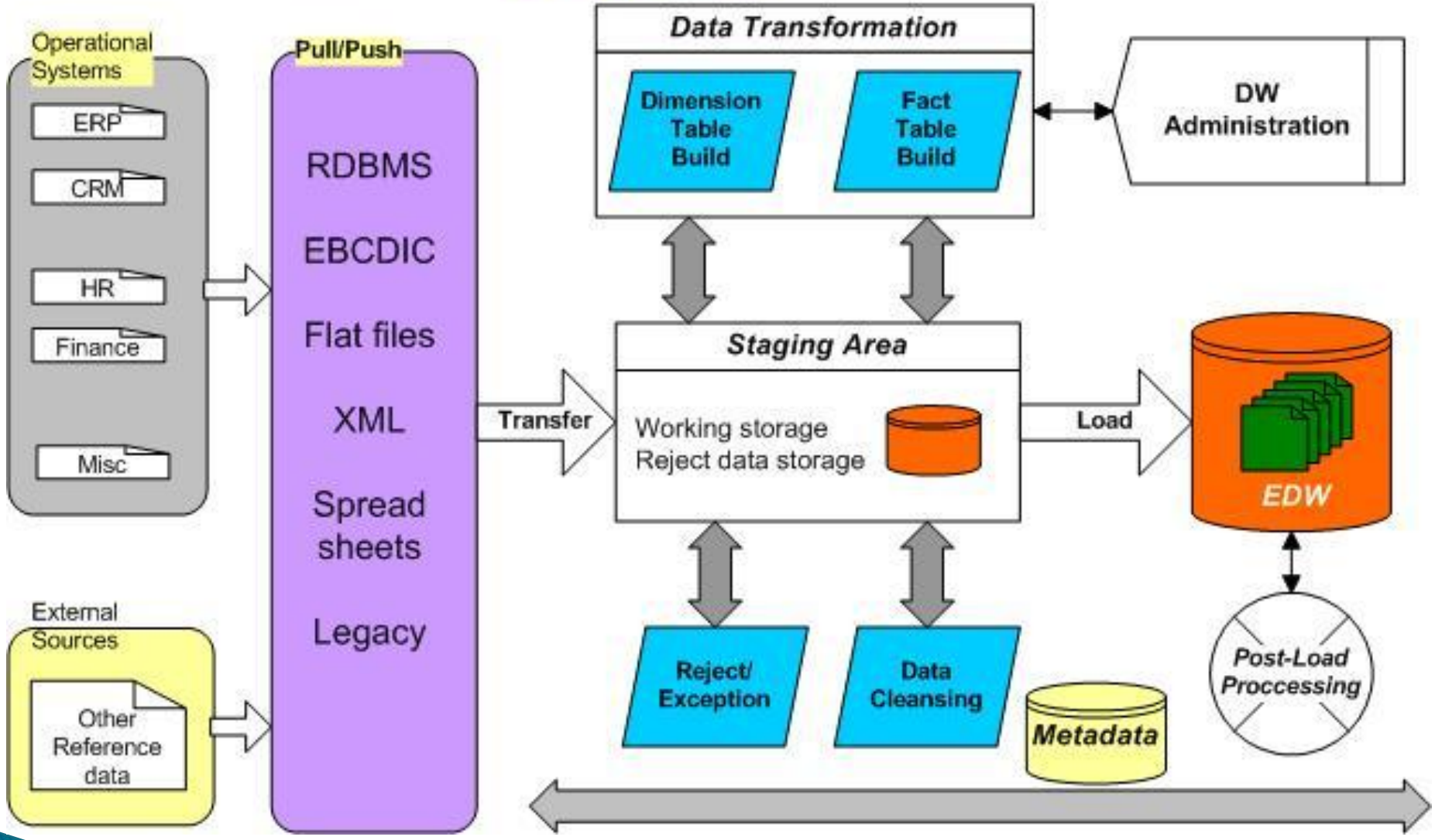
ETL (Extract, Transform, Load)






Relational Database
Source System1
Source System2
Source System3
File Sources
Excel(.xls) files
Text(.txt) files
XML(.xml) files
and other files
Other Sources
SAP
People Soft
Siebel & several
DATA FROM VARIOUS SOURCES
Designing ETL Process
Creating Source and Target Repositories. Mapping Source and Target repositories.
Extraction
Data Cleansing
Data Profiling
Transformation
Aggregation
Filtering
Joining
Sorting
Etc.
Loading
Creation and execution of WorkFlows to load data from source to target
Etc.
ETL PROCESS
Data Warehouse
Data Mart1
Data Mart2
Data Mart3
TARGET

ETL Architect



ETL process

Source data from CEID

	Visitor Arrivals: Total	Visitor Arrivals: Indonesia	Visitor Arrivals: Malaysia
Country	Singapore	Singapore	Singapore
Frequency	Monthly	Monthly	Monthly
Unit	Person	Person	Person
Source	Singapore	Singapore	Singapore
Status	Active	Active	Active
01/01/1999	570444	114115	37231
01/02/1999	517554	73732	36404
01/03/1999	580426	87192	40066
01/04/1999	524455	78634	40327

Import data into JMP

Remove the unnecessary rows and change the column names

Date	Year	Total	China	Indonesia
05/01/2003	2003	177808	3994	54851
06/01/2003	2003	316587	8405	94919
07/01/2003	2003	540914	20989	139957
04/01/2003	2003	203562	23002	38859
03/01/2001	2001	665565	30320	110821
03/01/2000	2000	640922	30644	102620

Using the formula column to generate the Year column based on Date

Aggregation

Aggregate the row based on Year using Summary function

Year	Total	China	Indonesia
2000	7691399	434336	1313316
2001	7522163	497398	1364380
2002	7567039	670099	1393020
2003	6127029	568510	1341747
2004	8328658	880279	1765326
2005	8943041	857820	1813569
2006	9751141	1037201	1922217
2007	10284545	1113956	1962055
2008	10115638	1078637	1765404
2009	9681259	936727	1745057

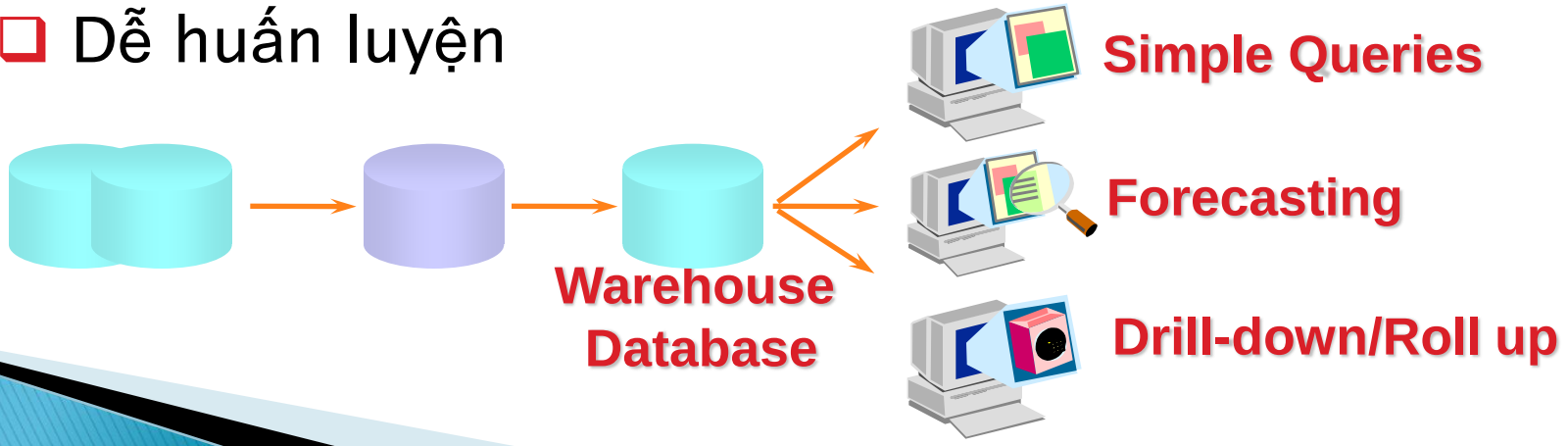
Transformation

Generate the Visitor grwoth rate columns based on the total number of visitors per year

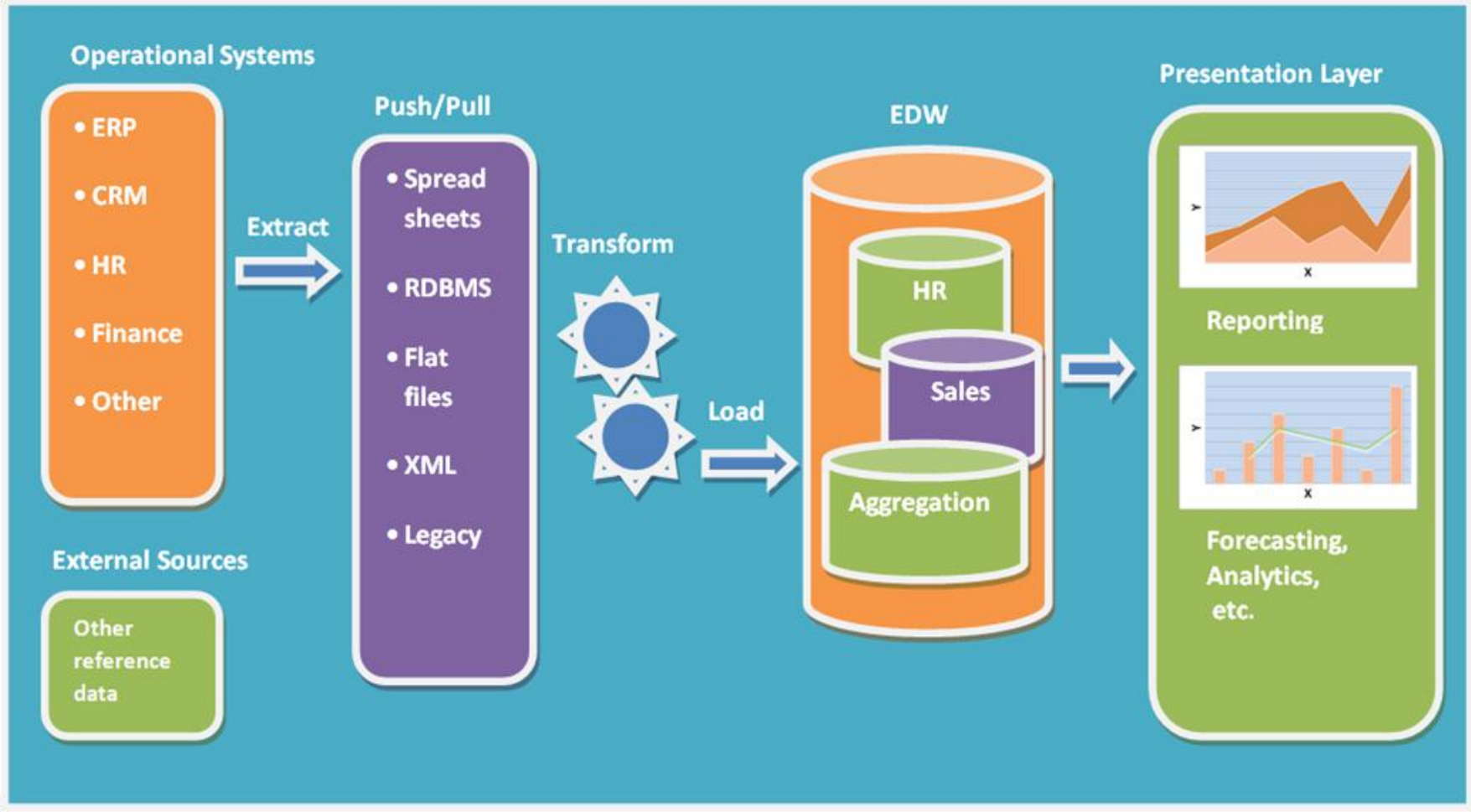
Year	Total 2	Total%	Indonesia	Indonesia %
2000	7691399	10.54%	1313316	8.54%
2001	7522163	-2.20%	1364380	3.89%
2002	7567039	0.60%	1393020	2.10%
2003	6127029	-19.03%	1341747	-3.68%
2004	8328658	35.93%	1765326	31.57%
2005	8943041	7.38%	1813569	2.73%
2006	9751141	9.04%	1922217	5.99%
2007	10284545	5.47%	1962055	2.07%
2008	10115638	-1.64%	1765404	-10.02%
2009	9681259	-4.29%	1745057	-1.15%

Các dịch vụ Phân tích dữ liệu, tạo báo cáo và dự báo (OLAP, SSRS, Data mining)

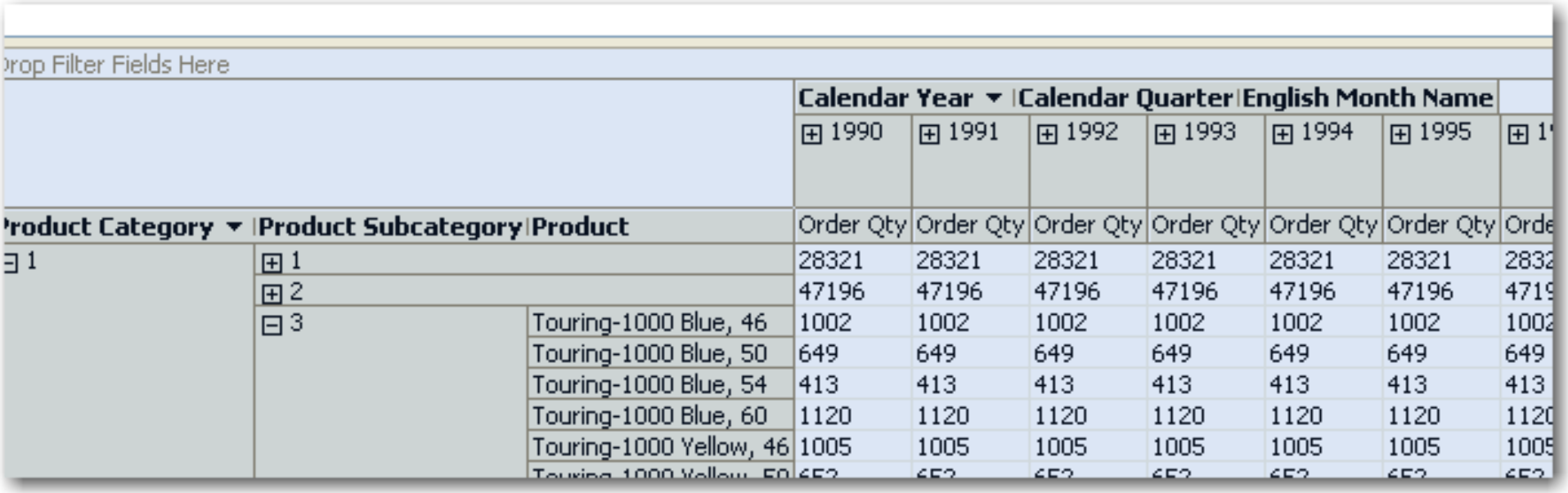
- Các công cụ dùng để phân tích, dự báo xu hướng dữ cho yêu cầu kinh doanh
  - Dễ dùng
  - Trực quan
  - Siêu dữ liệu
  - Dễ huấn luyện



Các dịch vụ Phân tích dữ liệu, tạo báo cáo và dự báo (OLAP, SSRS, Data mining)



OLAP example



rop Filter Fields Here

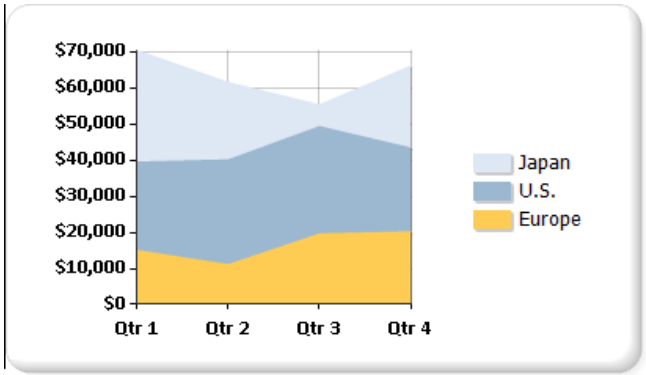
			Calendar Year						
			1990	1991	1992	1993	1994	1995	1
roduct Category	Product Subcategory	Product	Order Qty	Order Qty	Order Qty	Order Qty	Order Qty	Order Qty	Orde
1	1		28321	28321	28321	28321	28321	28321	2832
	2		47196	47196	47196	47196	47196	47196	4719
	3	Touring-1000 Blue, 46	1002	1002	1002	1002	1002	1002	1002
		Touring-1000 Blue, 50	649	649	649	649	649	649	649
		Touring-1000 Blue, 54	413	413	413	413	413	413	413
		Touring-1000 Blue, 60	1120	1120	1120	1120	1120	1120	1120
		Touring-1000 Yellow, 46	1005	1005	1005	1005	1005	1005	1005
		Touring-1000 Yellow, 50	652	652	652	652	652	652	652

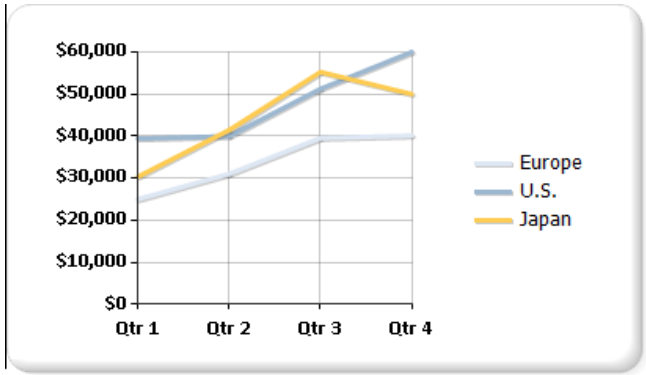
OLAP example

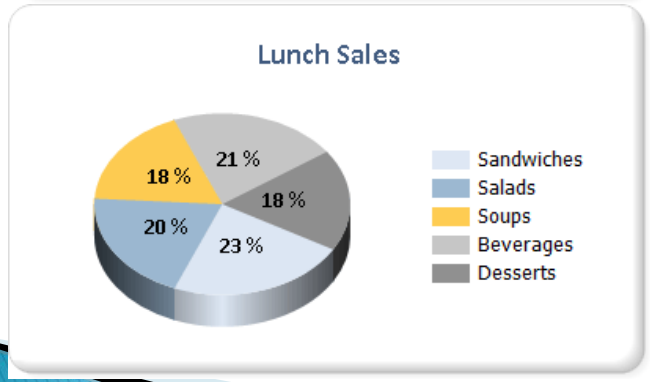
	A	B	C	D	E	F	G
1							
2							
3	Customer Count		Calendar Year				
4	Occupation	Gender	CY 2001	CY 2002	CY 2003	CY 2004	Grand Total
5	Clerical	Female	76	228	712	900	1440
6		Male	86	240	758	907	1488
7	Clerical Total		162	468	1470	1807	2928
8	Management	Female	69	252	790	934	1483
9		Male	90	239	820	1014	1592
10	Management Total		159	491	1610	1948	3075
11	Manual	Female	57	139	556	687	1133
12		Male	59	157	590	765	1251
13	Manual Total		116	296	1146	1452	2384
14	Professional	Female	158	432	1440	1778	2793
15		Male	151	365	1391	1726	2727
16	Professional Total		309	797	2831	3504	5520
17	Skilled Manual	Female	140	323	1123	1331	2284
18		Male	127	302	1129	1335	2293
19	Skilled Manual Total		267	625	2252	2666	4577
20	Grand Total		1013	2677	9309	11377	18484

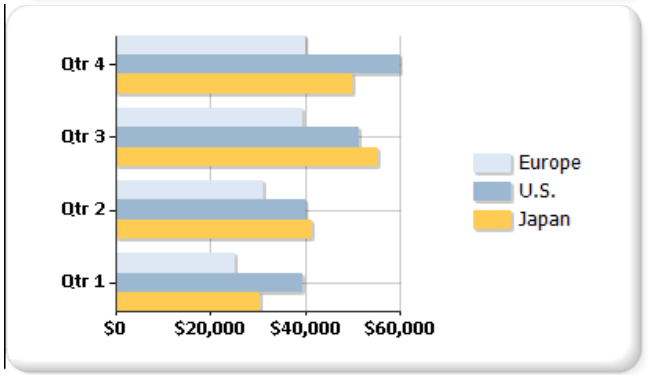
SSRS (Charts)


AUTHORING
Report Definition
MANAGEMENT
Managed Report
DELIVERY
Online Access
Delivery Channels


$70,000
$60,000
$50,000
$40,000
$30,000
$20,000
$10,000
$0
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Japan
U.S.
Europe


$60,000
$50,000
$40,000
$30,000
$20,000
$10,000
$0
Qtr 1
Qtr 2
Qtr 3
Qtr 4
Europe
U.S.
Japan


Lunch Sales
21 %
18 %
18 %
20 %
23 %
Sandwiches
Salads
Soups
Beverages
Desserts


Qtr 4
Qtr 3
Qtr 2
Qtr 1
$0
$20,000
$40,000
$60,000
Europe
U.S.
Japan

So sánh OLTP và kho dữ liệu

Xử lý giao dịch trực tuyến (OLTP)

- Công nghệ: CSDL quan hệ
- Chuẩn hóa, không dư thừa
- Tập trung vào dữ liệu hiện tại
- Trả lời các truy vấn đơn
- Toàn tác: tính toàn vẹn, bảo mật, đồng thời, Locking
- Xử lí giao dịch

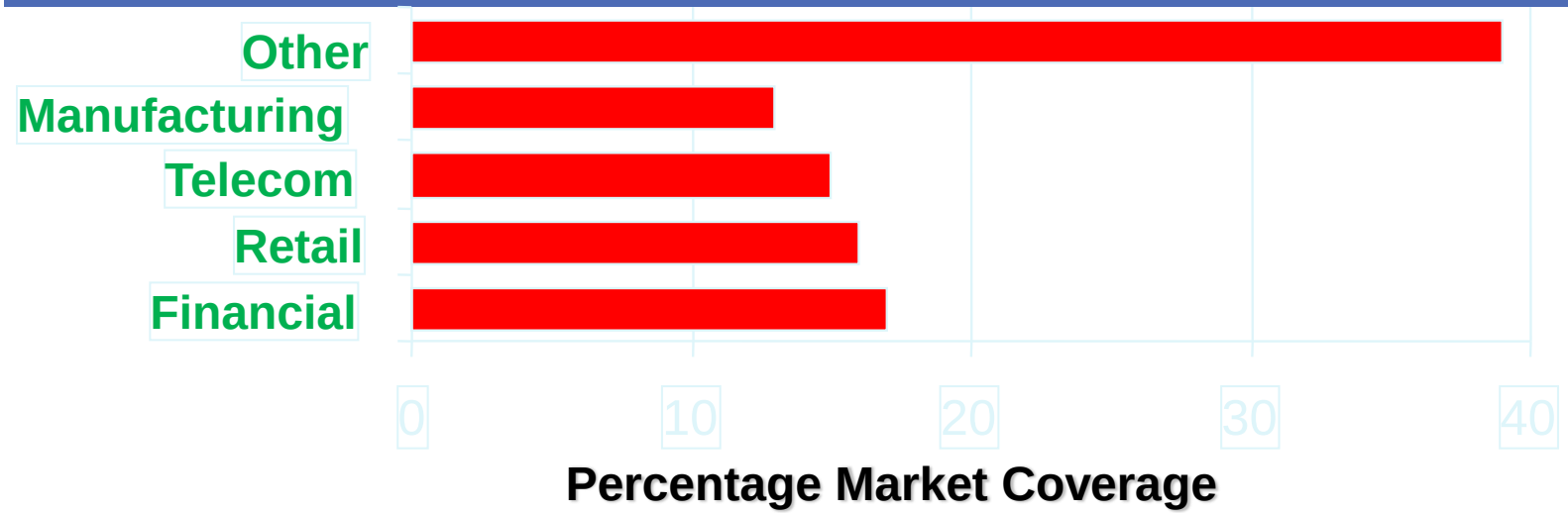
Kho dữ liệu, xử lý phân tích trực tuyến (OLAP)

- CSDL quan hệ, CSDL đa chiều
- Chấp nhận dư thừa
- Tiền tính toán tổng hợp
- Dữ liệu lịch sử
- Hỗ trợ phân tích phức tạp
- Tích hợp dữ liệu từ đa nguồn
- Dữ liệu rất lớn
- Các câu hỏi phức tạp

Các ví dụ về OLTP và OLAP

- Xử lý giao dịch trực tuyến OLTP
  - Số lượng coca cola vừa được bán
- Xử lý phân tích trực tuyến OLAP
  - Số lượng coca cola được bán tháng trước tại các cửa hàng phía bắc tỉnh Thừa thiên Huế
  - Cửa hàng nào phía bắc tỉnh Thừa thiên Huế có số lượng coca cola được bán ra tháng trước lớn nhất
  - Tháng nào trong năm số lượng coca cola được bán ra nhiều nhất tại tỉnh Thừa thiên Huế

Các ứng dụng của KDL



- Hàng không Airline
- Ngân hàng Banking
- Chăm sóc sức khỏe Health care
- Đầu tư Investment
- Bảo hiểm Insurance

- Bán lẻ Retail
- Viễn thông
- Các ngành công nghiệp Manufacturers
- Credit card suppliers
- Clothing distributors

Tóm tắt

- Kho dữ liệu: khái niệm, mục tiêu của kho dữ liệu, các đặc tính và ứng dụng của kho dữ liệu trong thực tế.
- Phân biệt được kho dữ liệu với xử lý phân tích trực tuyến và hệ thống xử lý các giao dịch tác nghiệp trực tuyến