HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 12: Kho dữ liệu

Lê Hải Hà

Kho dữ liệu (Data Warehouse) là gì?

- Nhiều định nghĩa.
  - CSDL hỗ trợ ra quyết định mà được duy trì **tác biệt** từ CSDL vận hành của tổ chức
  - Hỗ trợ **việc xử lý thông tin** bằng cách cung cấp một nền tảng vững chắc các dữ liệu lịch sử và hợp nhất cho việc phân tích.
- "Kho dữ liệu là tập dữ liệu **hướng chủ đề (subject-oriented)**, **tích hợp (integrated)**, **theo thời gian (time-variant)**, và **bền vững (nonvolatile)** hỗ trợ các tiến trình ra quyết định quản lý."—W. H. Inmon
- Data warehousing:
  - Tiến trình xây dựng và sử dụng các kho dữ liệu

Nội dung

❶ Kho dữ liệu: các khái niệm cơ bản

❷ Mô hình hóa DW: Data Cube và OLAP

❸ Thiết kế và sử dụng DW

❹ Thể hiện DW

❺ Tổng quát hóa dữ liệu bằng quy nạp hướng thuộc tính (AOI)

DW – Tích hợp

- Được xây dựng bằng cách tích hợp nhiều nguồn dữ liệu đa dạng
  - Các CSDL quan hệ, các file, các bản ghi giao dịch trực tuyến
- Sử dụng các kỹ thuật làm sạch và tích hợp dữ liệu.
  - Đảm bảo sự nhất quán về các quy ước tên, các cấu trúc mã hóa, các số đo thuộc tính, v.v. trong các nguồn dữ liệu khác nhau
    - Thí dụ: với giá khách sạn: tiền tệ, thuế, bao gồm bữa sáng, v.v.
  - Chuyển đổi dữ liệu khi chuyển dữ liệu từ nguồn tới DW.

DW – Hướng chủ đề

- Được tổ chức quanh các chủ đề như khách hàng, sản phẩm, bán hàng
- Tập trung vào việc mô hình hóa và phân tích dữ liệu của người ra quyết định, không phải là các tác nghiệp hay xử lý giao dịch hàng ngày
- Cung cấp một góc nhìn đơn giản và xúc tích quanh chủ đề quan tâm bằng cách loại bỏ các dữ liệu không hữu dụng trong tiến trình hỗ trợ quyết định

DW – Dữ liệu theo thời gian

- Thời gian trong DW có ý nghĩa dài hơn so với các hệ thống tác nghiệp
  - CSDL tác nghiệp: dữ liệu giá trị hiện tại
  - Dữ liệu DW: cung cấp thông tin từ góc độ lịch sử (thí dụ 5 tới 10 năm qua)
- Mọi cấu trúc khóa trong DW
  - Chứa phần tử thời gian (time) tường minh hay không tường minh
  - Nhưng trong dữ liệu tác nghiệp, khóa có thể hoặc không chứa "phần tử thời gian"

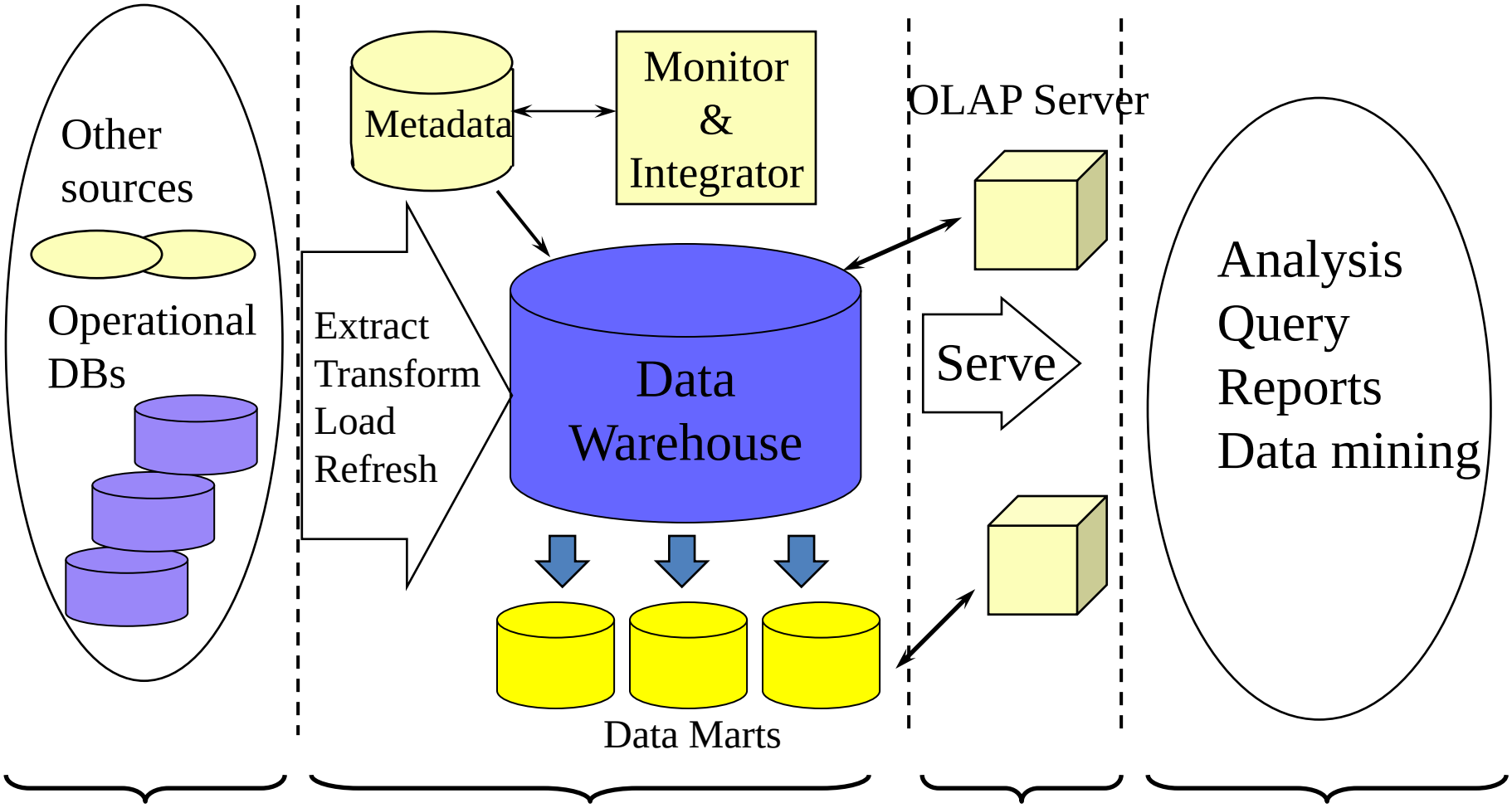
DW – Dữ liệu bền vững

- Kho dữ liệu phân tách vật lý được chuyển đổi từ môi trường tác nghiệp
- Việc cập nhật dữ liệu tác nghiệp không xảy ra trong môi trường DW
  - Không yêu cầu xử lý giao dịch, khôi phục, và các cơ chế kiểm soát đồng thời
  - Yêu cầu chỉ 2 tác vụ trong truy cập dữ liệu:
    - *Tải dữ liệu* và *truy cập dữ liệu*

OLTP và OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

DW: Kiến trúc đa tầng



Ba mô hình DW

- DW doanh nghiệp - Enterprise warehouse
  - Thu thập tất các các thông tin về các chủ đề trải rộng toàn bộ doanh nghiệp
- Data Mart
  - Tập con dữ liệu doanh nghiệp có giá trị cho một nhóm người dùng nhất định. Phạm vi của nó được hạn chế trong các nhóm được chọn, thí dụ Data Mart tiếp thị (marketing data mart)
    - Độc lập hay phụ thuộc (trực tiếp từ DW)
- DW ảo - Virtual warehouse
  - Tập các view trên CSDL tác nghiệp
  - Chỉ một số view tổng hợp có thể được materialized

Tại sao phân tách DW?

- Hiệu năng cao cho cả 2 hệ thống
  - CSDL tác nghiệp - được tối ưu cho OLTP: các phương pháp truy cập, đánh chỉ mục, kiểm soát đồng thời, khôi phục
  - DW – tối ưu cho OLAP: các truy vấn OLAP phức tạp, góc nhìn đa chiều, tổng hợp/hợp nhất
- Chức năng và dữ liệu khác nhau:
  - missing data: Hỗ trợ quyết định đòi hỏi dữ liệu lịch sử mà các CSDL tác nghiệp thường không duy trì
  - data consolidation: Hỗ trợ quyết định đòi hỏi dữ liệu hợp nhất (tổ hợp, tổng hợp) từ nhiều nguồn dữ liệu đa dạng
  - data quality: các nguồn dữ liệu khác nhau thường sử dụng các thể hiện, mã hóa và định dạng dữ liệu không nhất quán với nhau và cần được dàn xếp
- Lưu ý: có nhiều hệ thống thực hiện các phân tích OLAP trực tiếp trên CSDL tác nghiệp

Trích chọn, chuyển đổi, và tải (ETL)

- **Trích chọn dữ liệu - Data extraction**
  - Lấy dữ liệu từ nhiều nguồn định dạng bên ngoài
- **Làm sạch dữ liệu - Data cleaning**
  - Phát hiện các lỗi trong dữ liệu và hiệu chỉnh chúng khi có thể
- **Chuyển đổi dữ liệu - Data transformation**
  - Chuyển dữ liệu từ các định dạng gốc sang định dạng DW
- **Tải - Load**
  - Sắp xếp, tổng hợp, hợp nhất, tính các view, kiểm tra toàn vẹn, và xây dựng các chỉ mục, phân đoạn
- **Làm tươi - Refresh**
  - Lan tỏa các cập nhật từ các nguồn dữ liệu tới DW

Lưu trữ siêu dữ liệu

- **Siêu dữ liệu (Meta data)** là dữ liệu định nghĩa các đối tượng DW. Nó chứa:
- Mô tả cấu trúc của DW
  - Lược đồ, view, chiều, phân cấp, dữ liệu dẫn xuất, các vị trí và nội dung data mart
- Siêu dữ liệu tác nghiệp
  - Nguồn gốc dữ liệu (lịch sử chuyển đổi dữ liệu và đường dẫn chuyển đổi),trạng thái dữ liệu (hiện tại, lưu trữ, hoặc bị xóa), thông tin giám sát (các thống kê sử dụng DW, các báo cáo lỗi, tóm tắt kiểm toán)
- Các giải thuật tổng hợp
- Các ánh xạ từ môi trường tác nghiệp sang môi trường DW
- Dữ liệu liên quan tới hiệu năng hệ thống
  - Lược đồ DW, view, định nghĩa dữ liệu dẫn xuất
- Dữ liệu nghiệp vụ
  - Các thuật ngữ và định nghĩa nghiệp vụ, chủ sở hữu dữ liệu, các chính sách tính phí

Nội dung

❶ Kho dữ liệu: các khái niệm cơ bản

❷ Mô hình hóa DW: Data Cube và OLAP

❸ Thiết kế và sử dụng DW

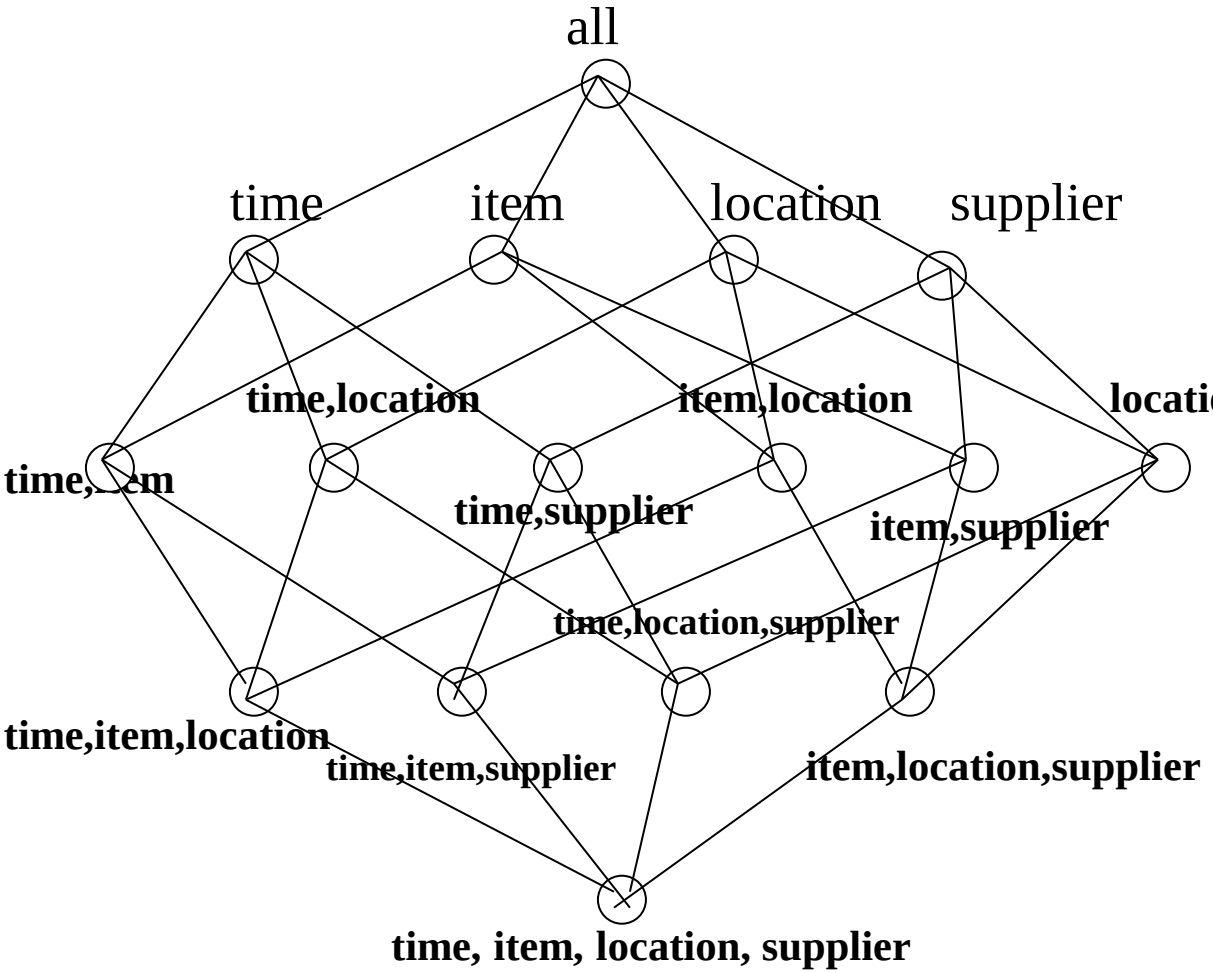
❹ Thể hiện DW

❺ Tổng quát hóa dữ liệu bằng quy nạp hướng thuộc tính

Từ các bảng và bảng tính tới Data Cube

- **Data warehouse (DW)** dựa trên mô hình dữ liệu đa chiều (multidimensional data model) mà xem dữ liệu ở dạng hộp dữ liệu (data cube)
- Data cube, như sales, cho phép mô hình và xem dữ liệu dưới nhiều chiều
  - **Các bảng chiều (Dimension table)**, như item (item_name, brand, type), hay time(day, week, month, quarter, year)
  - **Bảng fact (Fact table)** chứa các số đo (**measure**) (như dollars_sold) và các khóa liên kết tới các bảng chiều tương ứng
- Trong các tài liệu DW, cube n chiều (n-D) được gọi là cube cơ bản (base cuboid). Cuboid mức đỉnh (0-D) ở mức tổng hợp cao nhất gọi là apex cuboid. Lưới các cuboid hình thành một data cube.

Cube: Lưới các Cuboids



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

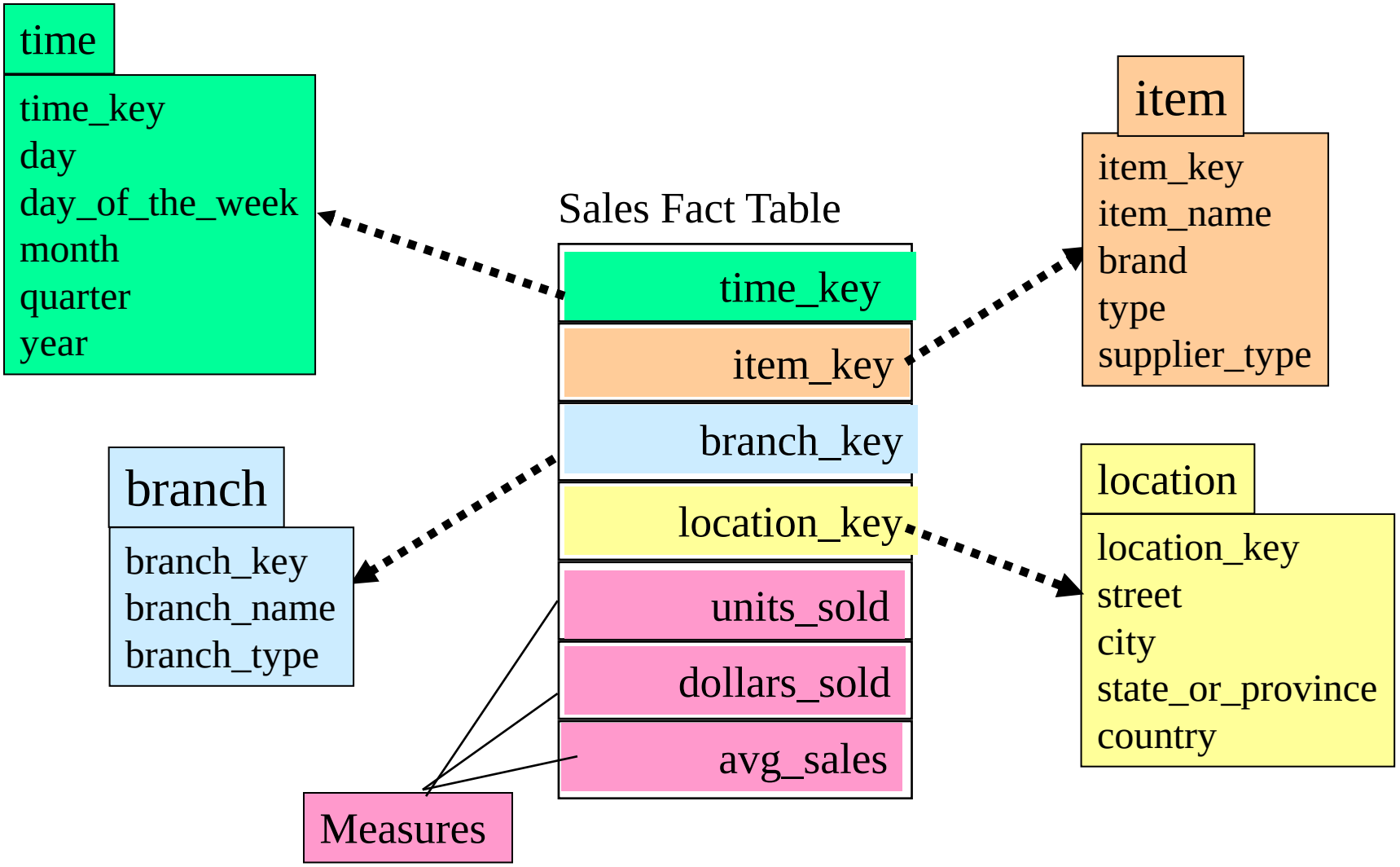
3-D cuboids

4-D (*base*) cuboid

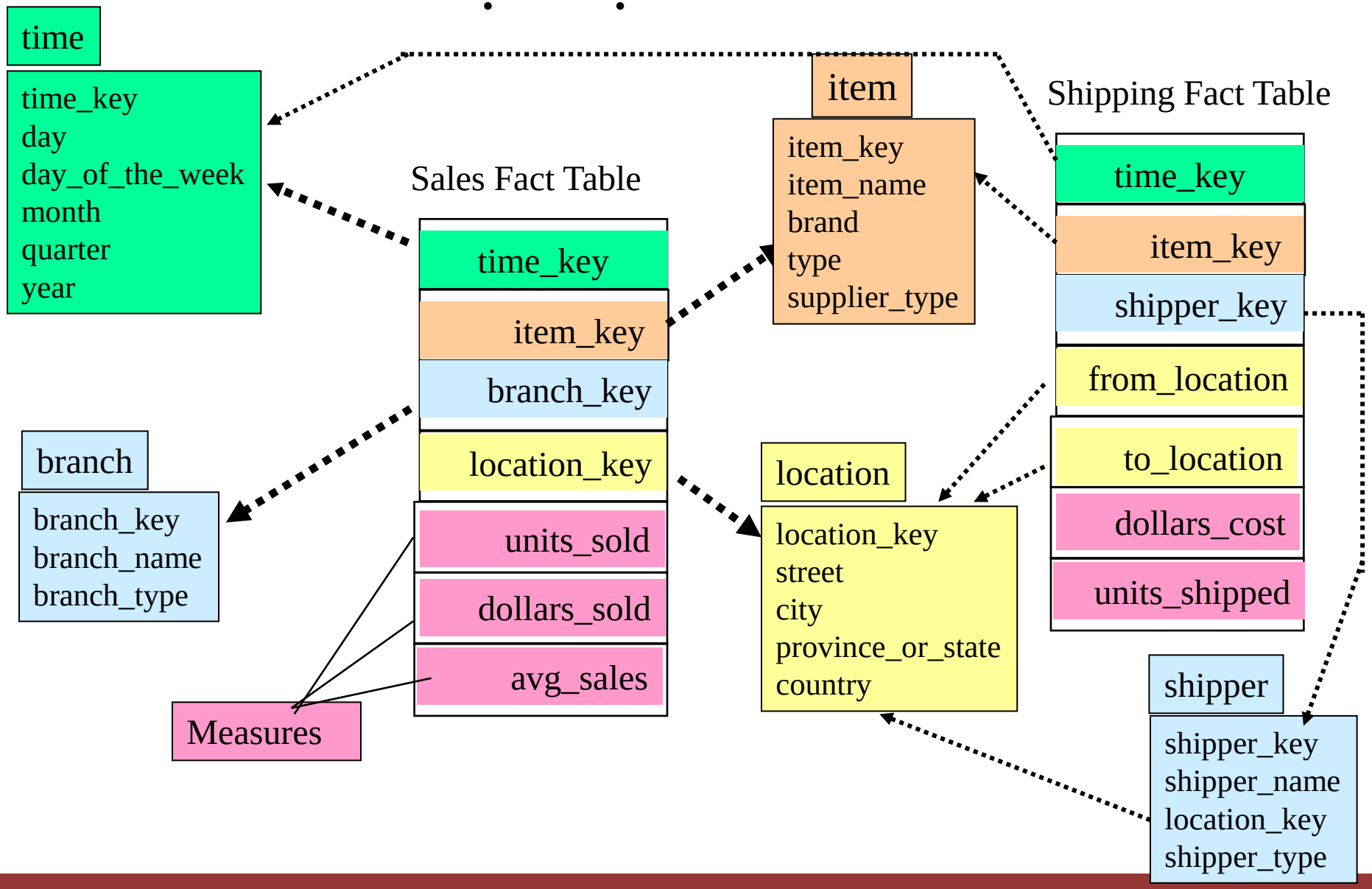
Mô hình khái niệm của DW

- Mô hình hóa các DW: chiều (dimension) và số đo (measure)
  - Lược đồ hình sao - Star schema: Một bảng fact ở trung tâm kết nối với một số bảng chiều
  - Lược đồ bông tuyết - Snowflake schema: chuẩn hóa lược đồ hình sao ở đó các cây phân cấp chiều được chuẩn hóa với tập các bảng chiều nhỏ hơn, hình thành hình dạng như bông tuyết
  - Lược đồ chòm sao - Fact constellations: nhiều bảng fact chia sẻ các bảng chiều, được xem như tập các ngôi sao, do đó gọi là chòm sao

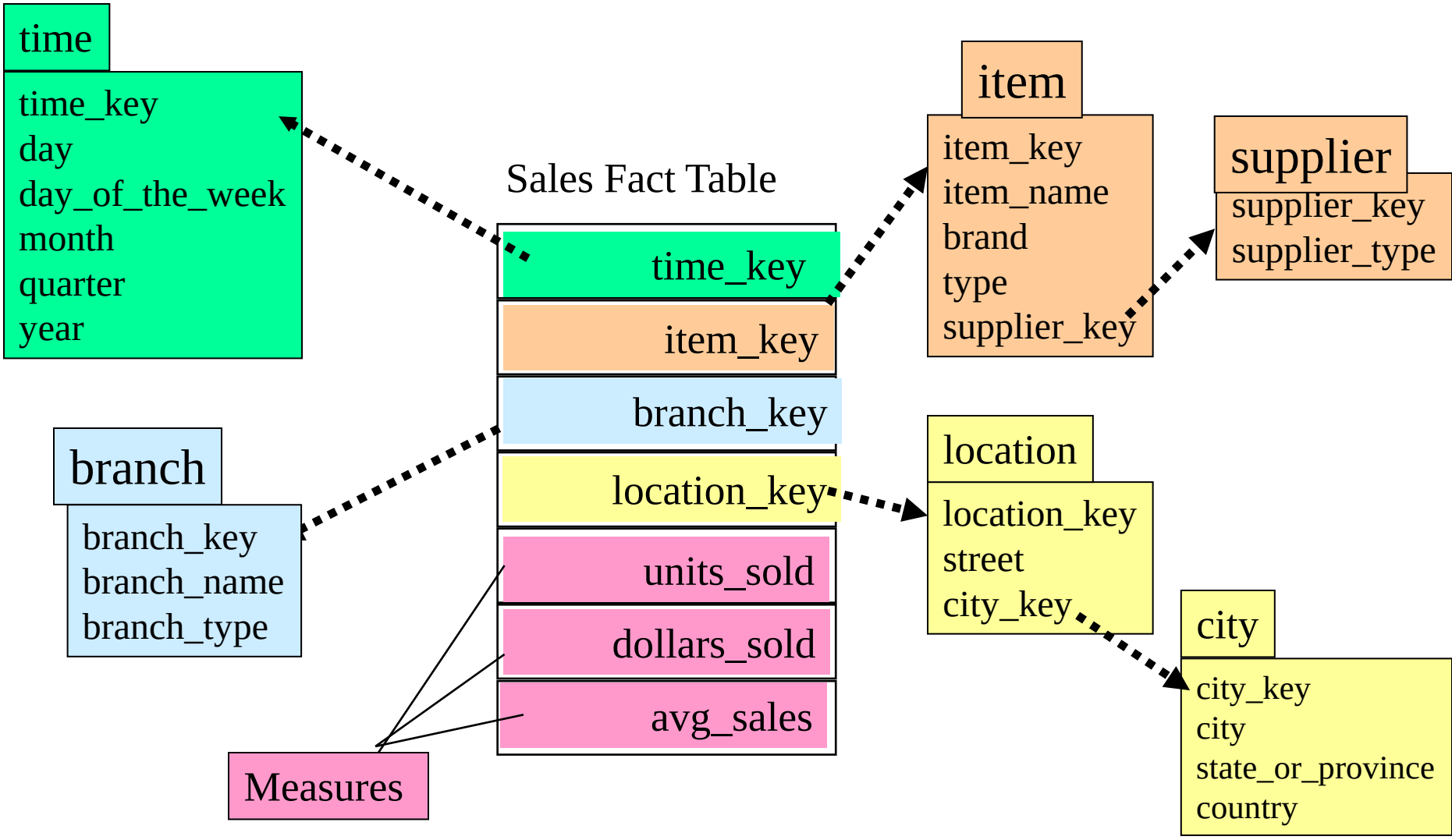
Thí dụ: lược đồ hình sao



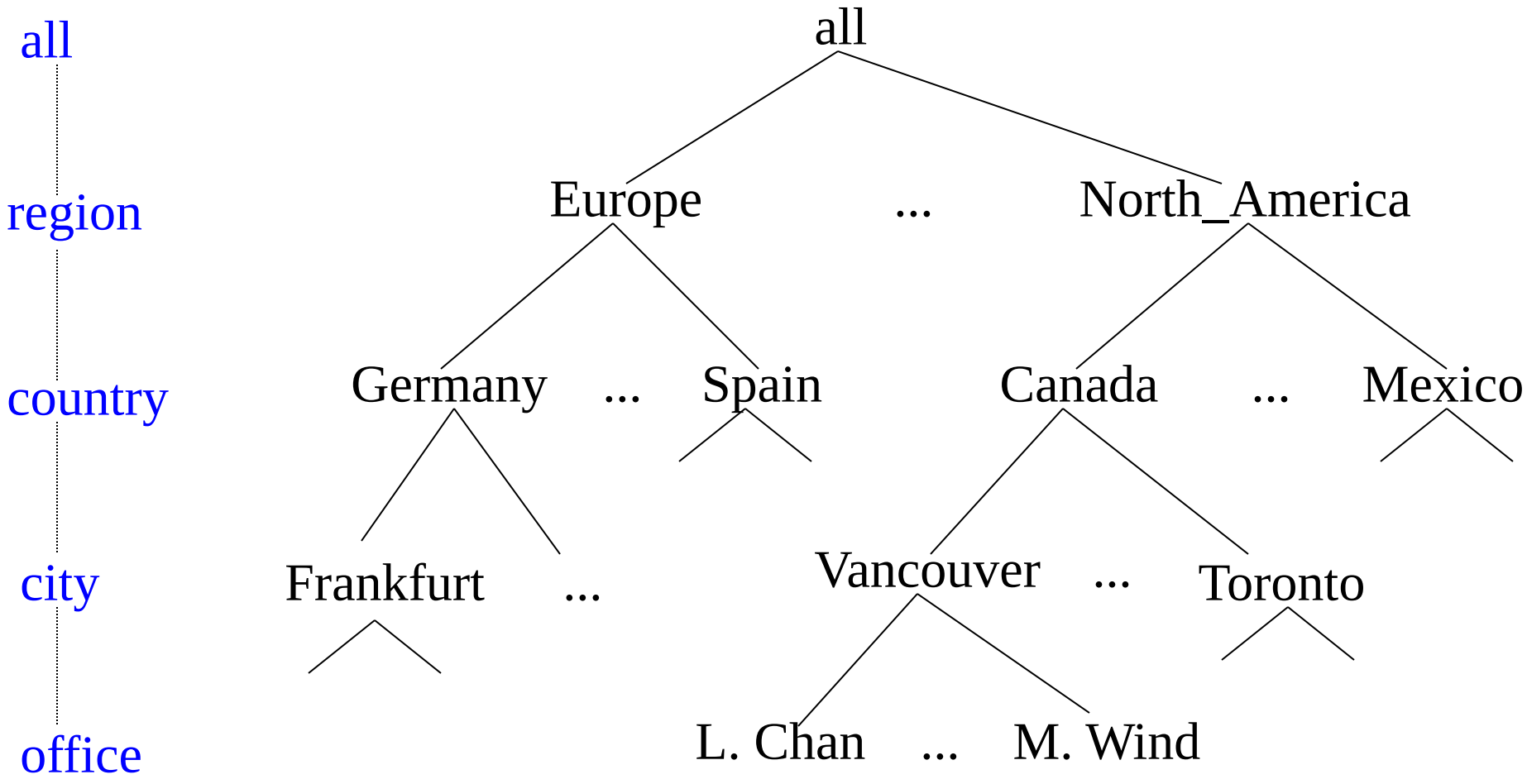
Thí dụ: lược đồ chòm sao



Thí dụ: lược đồ bông tuyết



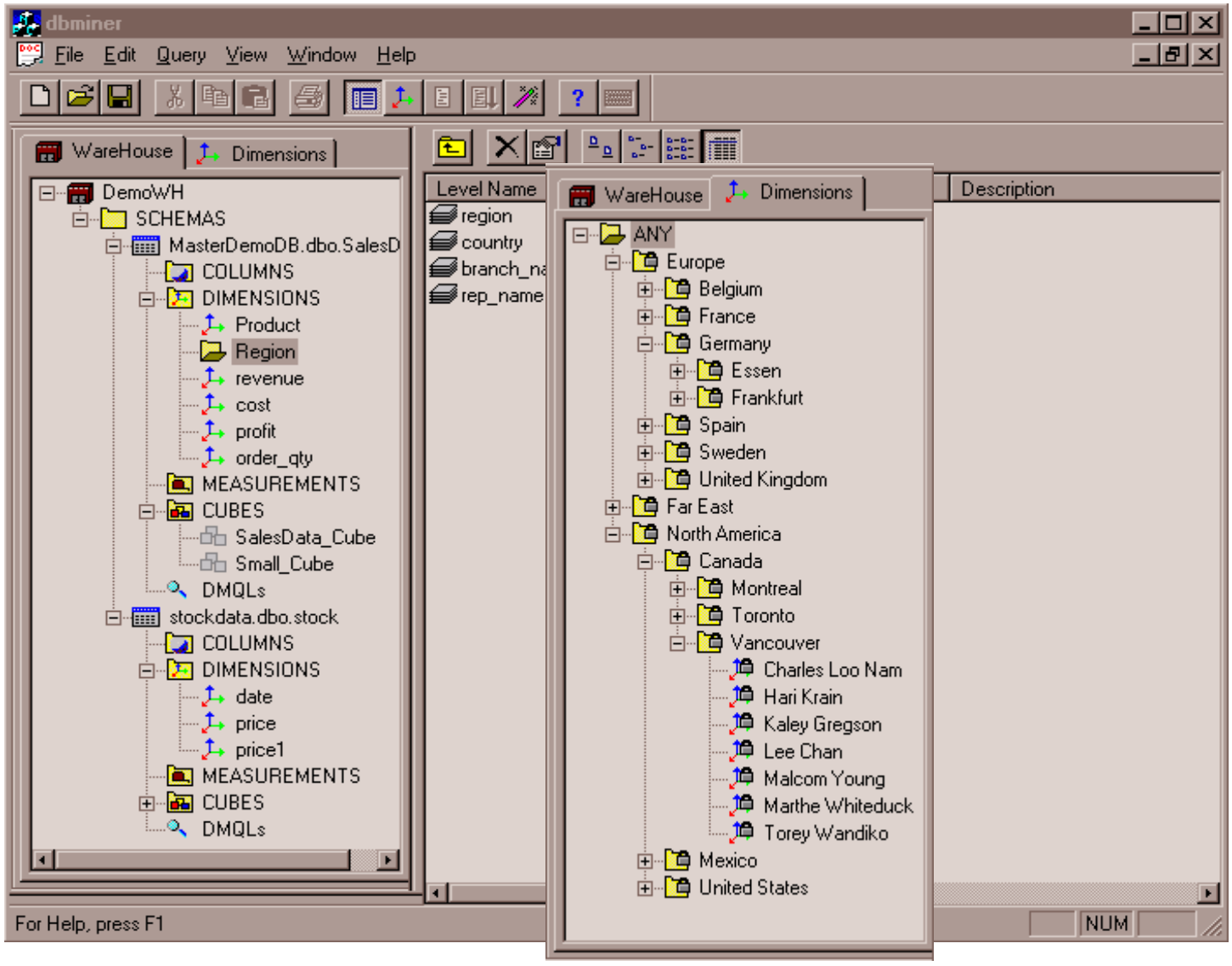
Cây phân cấp khái niệm: Chiều (location)



Các số đo Data Cube: Ba nhóm

- Phân phối - Distributive: nếu kết quả thu được bằng cách áp dụng hàm với *n* giá trị tổ hợp giống như việc áp dụng hàm đó tới tất cả dữ liệu không cần phân hoạch
  - Thí dụ: count(), sum(), min(), max()
- Đại số - Algebraic: nếu nó có thể được tính bằng một hàm đại số với *M* đối số (*M* nguyên, hữu hạn), mỗi đối số có được bằng cách áp dụng hàm tổ hợp phân phối
  - Thí dụ: avg(), min_N(), standard_deviation()
- Thực nghiệm - Holistic: nếu không có ràng buộc (hằng số hữu hạn) về kích thước lưu trữ cần để mô tả một tổ hợp con.
  - Thí dụ: median(), mode(), rank()

View DW và cây phân cấp



Đặc tả cây phân cấp

- Schema hierarchy

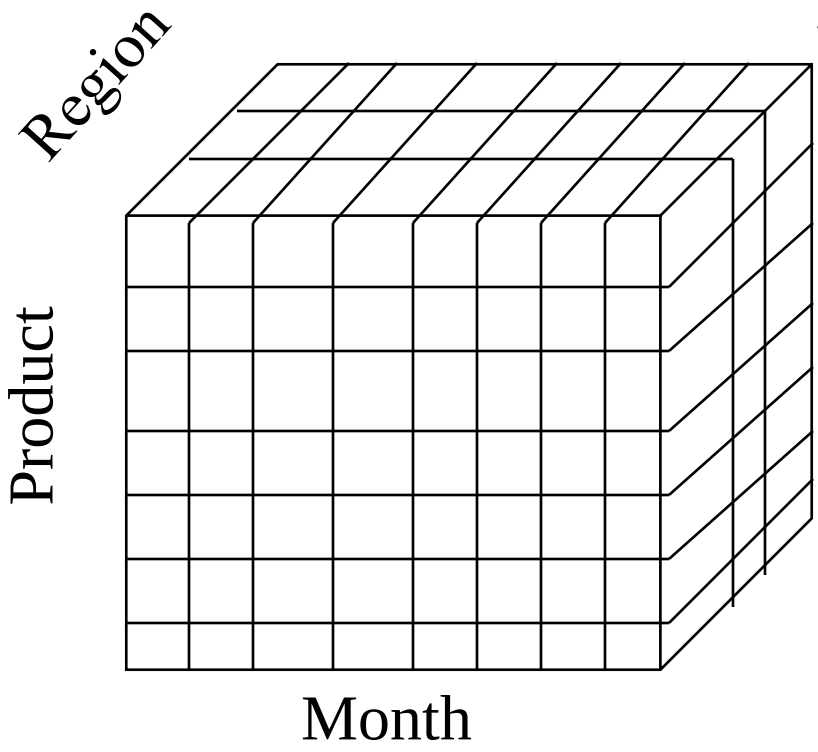
  day < {month < quarter; week} < year

- Set_grouping hierarchy

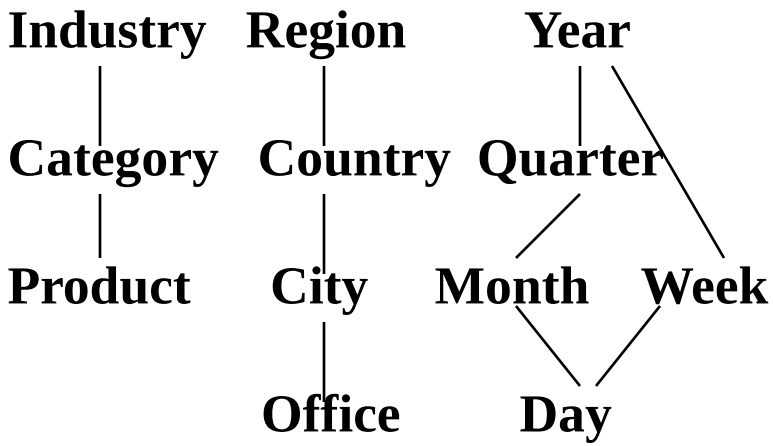
  {1..10} < inexpensive

Dữ liệu đa chiều

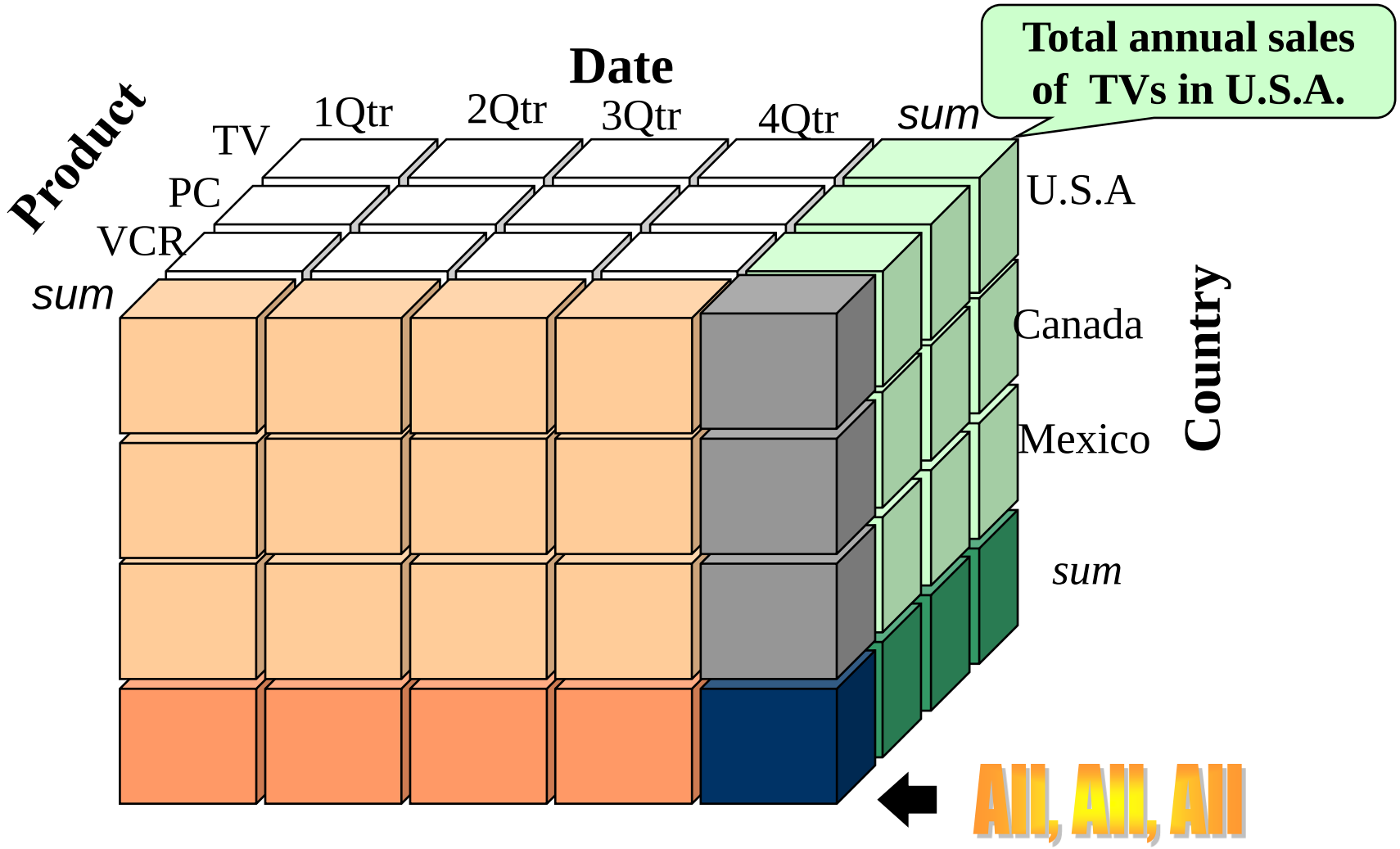
- Số lượng bán được xem là hàm của sản phẩm, tháng, và vùng



Dimensions: *Product, Location, Time*
Hierarchical summarization paths



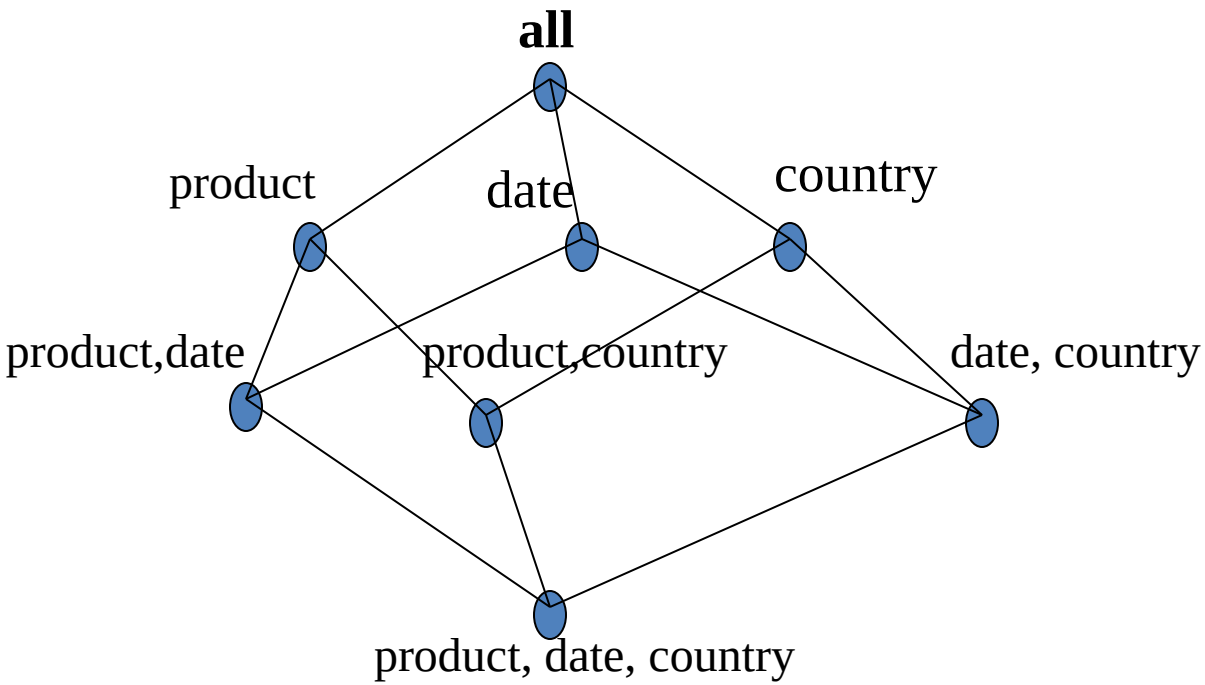
Thí dụ


Total annual sales of TVs in U.S.A.
Date
1Qtr
2Qtr
3Qtr
4Qtr
sum
Product
TV
PC
VCR
sum
U.S.A
Canada
Mexico
sum
Country
All, All, All

Các toán tử OLAP chính

- Roll up (drill-up): tổng hợp dữ liệu
  - *Bằng cách leo lên mức cao của cây phân cấp hay giảm số chiều*
- Drill down (roll down): ngược lại với roll up
  - *Từ mức tổng hợp cao tới mức tổng hợp thấp hay dữ liệu chi tiết, hay đưa vào các chiều mới*
- Slice và dice: *chiếu và chọn*
- Pivot (xoay):
  - *Đổi hướng của cube, trực quan hóa, 3D sang chuỗi mặt 2D*
- Các toán tử khác
  - *drill across: liên quan tới nhiều hơn 1 bảng fact*
  - *drill through: thông qua mức thấp nhất của cube tới các bảng quan hệ back-end của nó (sử dụng SQL)*

Các Cuboid tương ứng với Cube



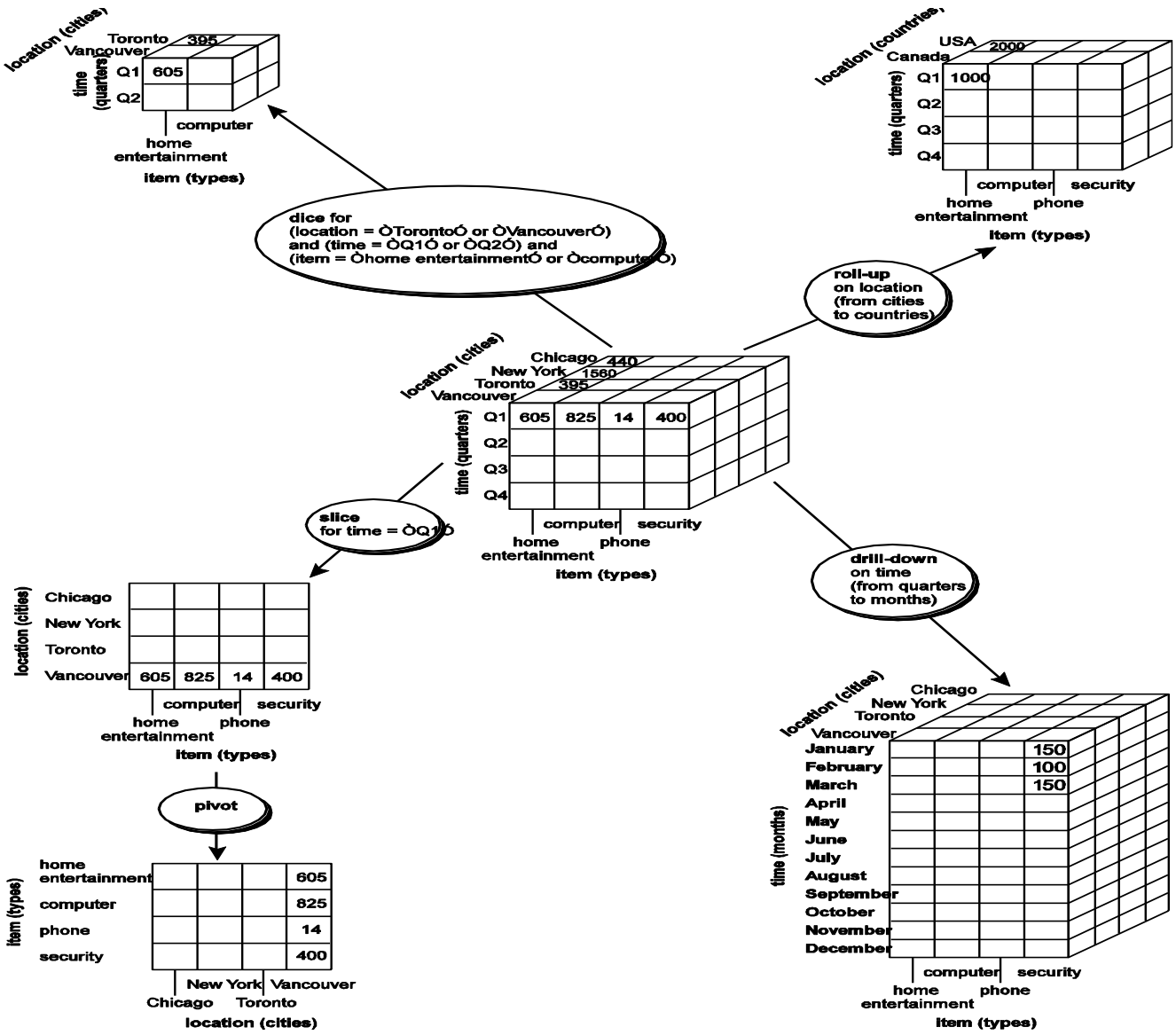
0-D (*apex*) cuboid

1-D cuboids

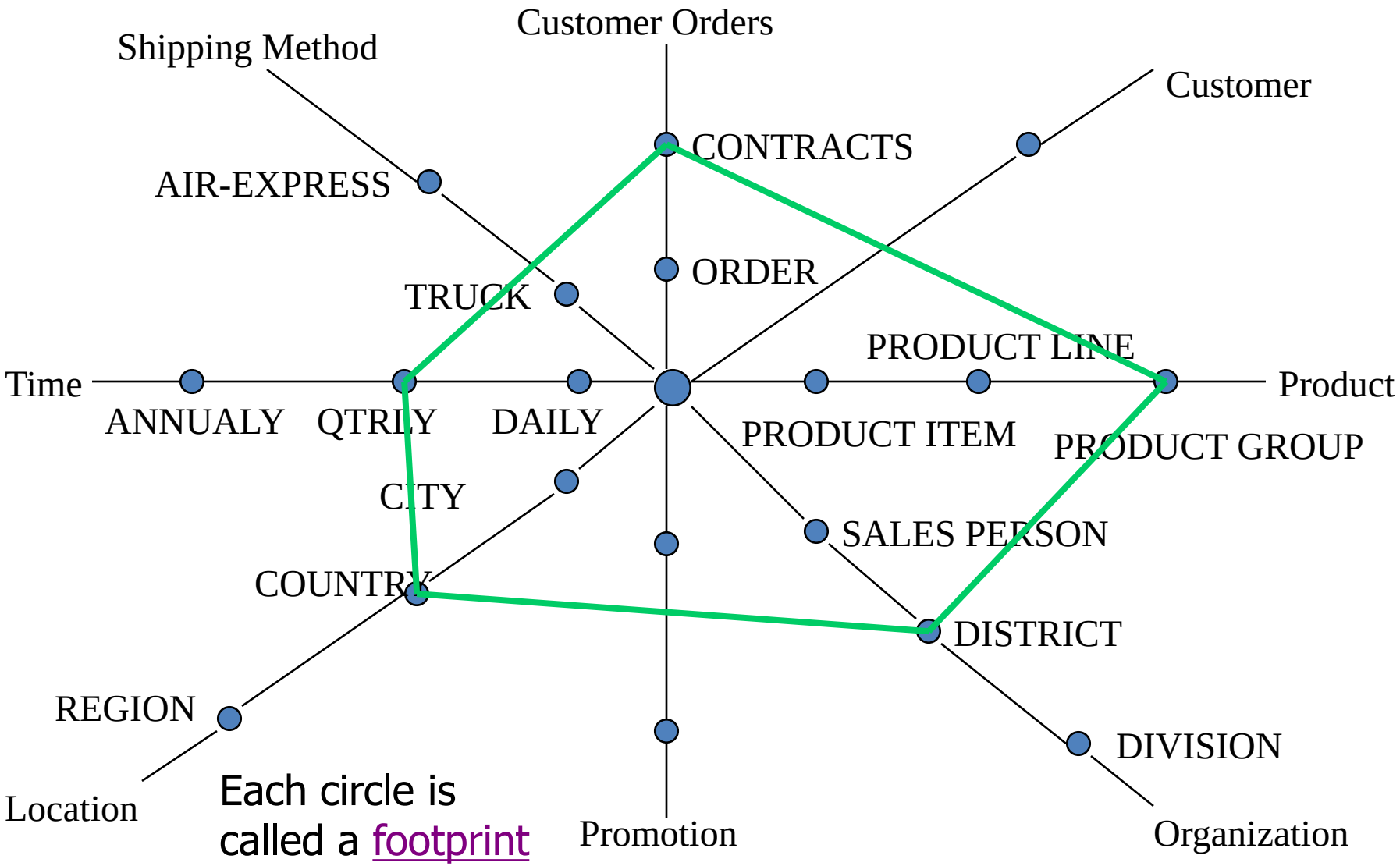
2-D cuboids

3-D (*base*) cuboid

Typical OLAP Operations

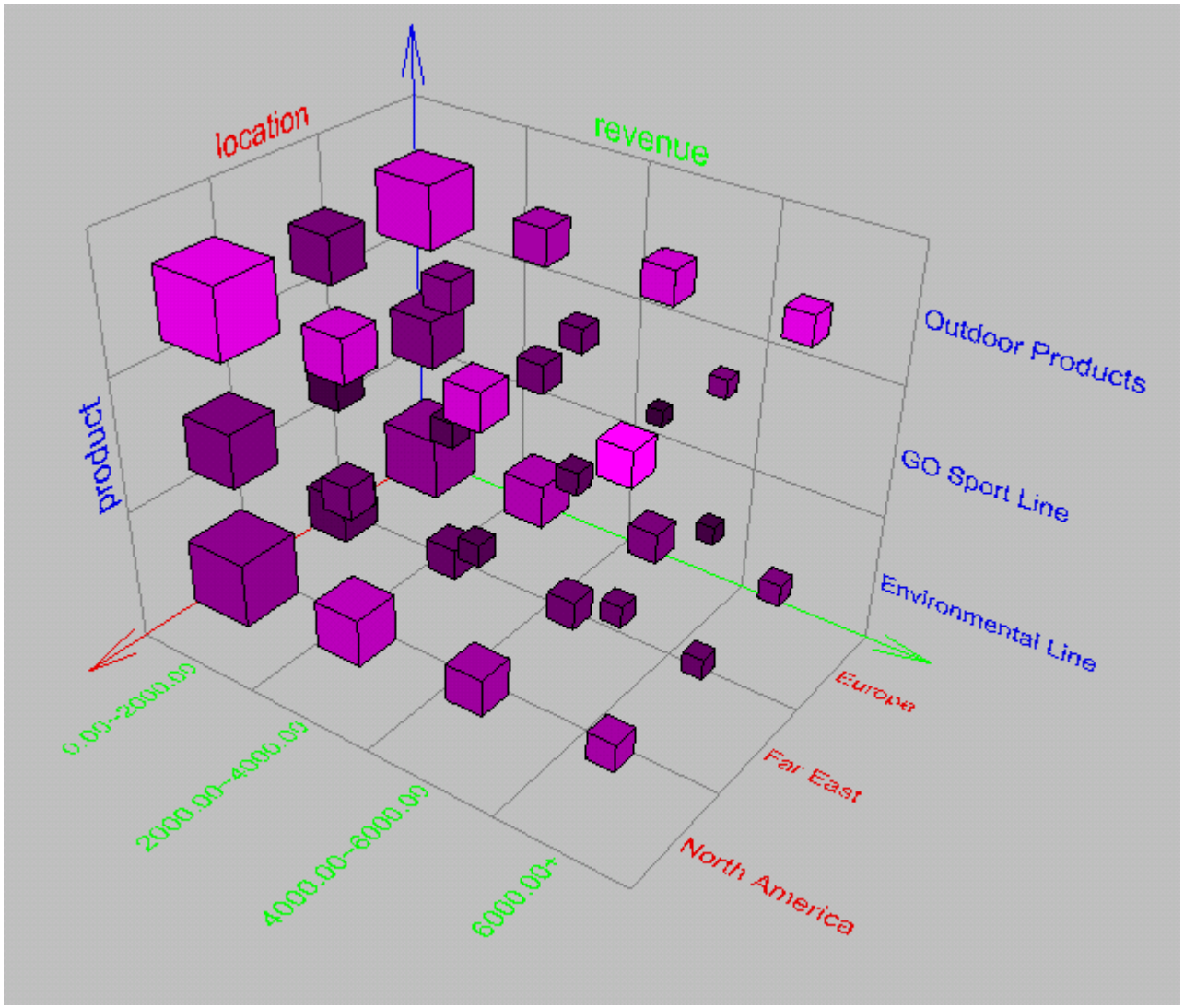


Mô hình truy vấn Star-Net



Each circle is called a footprint

Duyệt/browsing Data Cube



- Visualization
- OLAP capabilities
- Interactive manipulation

Nội dung

❶ Kho dữ liệu: các khái niệm cơ bản

❷ Mô hình hóa DW: Data Cube và OLAP

❸ Thiết kế và sử dụng DW

❹ Thể hiện DW

❺ Tổng quát hóa dữ liệu bằng quy nạp hướng thuộc tính

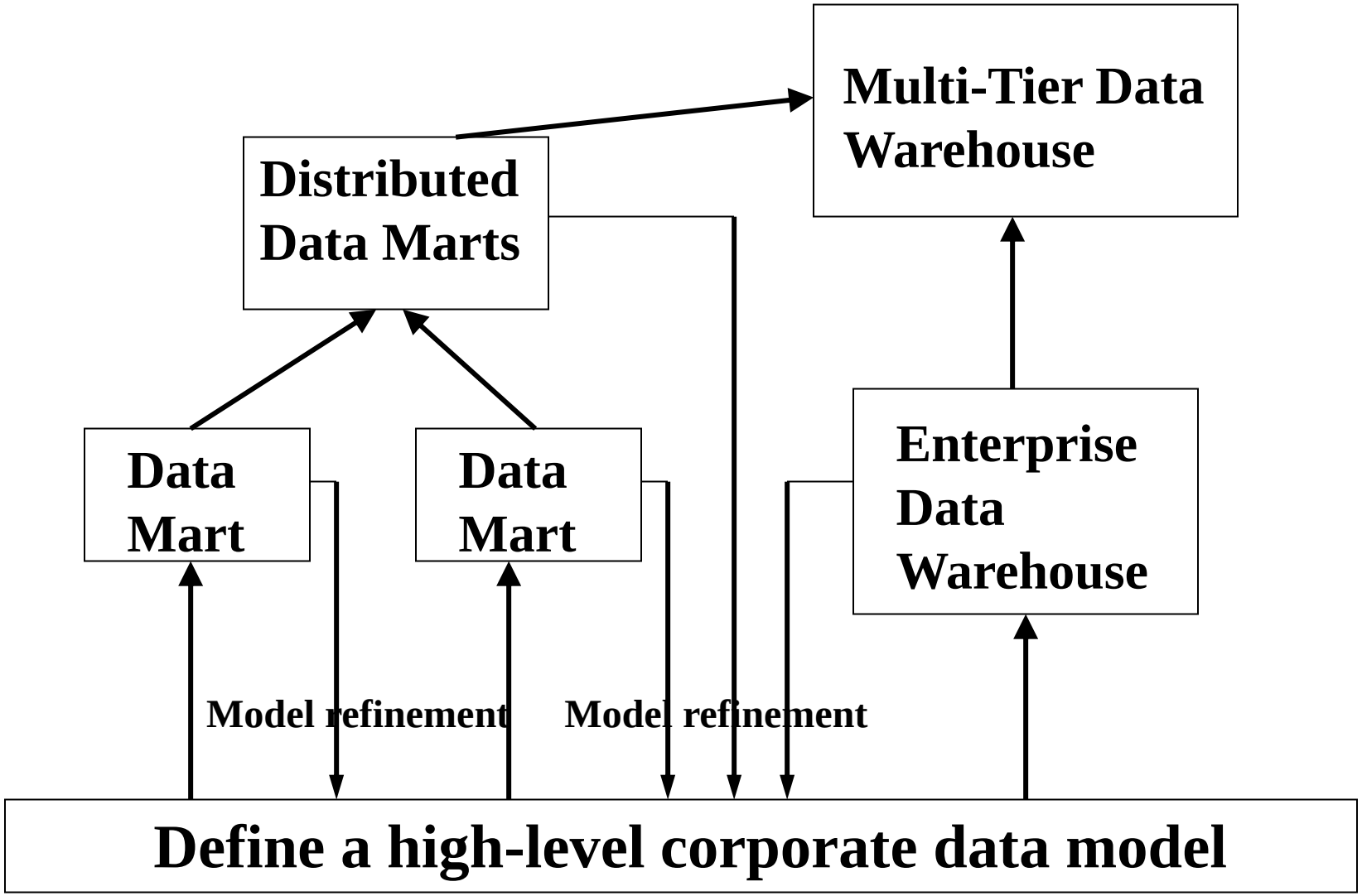
Quy trình thiết kế DW

- **Các tiếp cận trên xuống, dưới lên và tổ hợp cả 2**
  - Trên xuống (Top-down): Bắt đầu với thiết kế chung và lập kế hoạch (trưởng thành)
  - Dưới lên (Bottom-up): Bắt đầu với các thử nghiệm và nguyên mẫu (nhanh)
- **Từ quan điểm kỹ nghệ phần mềm**
  - Thác nước (Waterfall): phân tích cấu trúc và hệ thống tại mỗi bước trước khi tiến hành bước tiếp theo
  - Xoắn ốc (Spiral): xây dựng nhanh hệ thống chức năng tăng dần, vòng quay ngắn, nhanh chóng quay vòng
- **Quy trình thiết kế DW điển hình**
  - Chọn quy trình nghiệp vụ để mô hình, thí dụ: orders, invoices, v.v.
  - Chọn mức hạt/chi tiết (*grain*) (mức dữ liệu nguyên tố) của quy trình nghiệp vụ
  - Chọn các chiều (dimensions) sẽ áp dựng tới mỗi bản ghi của bảng fact
  - Chọn số đo (measure) được tính trên mỗi bảng ghi bảng fact

Thiết kế DW: Khung phân tích nghiệp vụ

- Bốn góc nhìn liên quan tới việc thiết kế 1 DW
  - Góc nhìn từ trên xuống - Top-down view
    - Cho phép chọn các thông tin liên quan cần thiết cho DW
  - Góc nhìn nguồn dữ liệu - Data source view
    - Xem xét thông tin đang được thu thập, lưu trữ và quản lý bởi các hệ tác nghiệp
  - Góc nhìn DW - Data warehouse view
    - Chứa các bảng fact và các bảng chiều
  - Góc nhìn nghiệp vụ - Business query view
    - Xem dữ liệu trong DW từ quan điểm của người dùng cuối

Phát triển DW: Một tiếp cận khuyến nghị



Sử dụng DW

- Ba dạng ứng dụng DW
  - Xử lý thông tin - Information processing
    - Hỗ trợ truy vấn, phân tích thống kê cơ bản, và báo cáo sử dụng bảng chéo, bảng, biểu đồ và đồ thị
  - Xử lý phân tích - Analytical processing
    - Phân tích đa chiều dữ liệu DW
    - Hỗ trợ các phép toán OLAP cơ bản, slice-dice, drilling, pivoting
  - Khai phá dữ liệu - Data mining
    - Khám phá tri thức từ các mẫu ẩn
    - Hỗ trợ các liên kết, thành lập các mô hình phân tích, thực hiện phân lớp và dự báo, và thể hiện các kết quả khai phá sử dụng các công cụ trực quan

Từ xử lý phân tích trực tuyến (OLAP) tới khai phá phân tích trực tuyến (OLAM)

- Tại sao khai phá phân tích trực tuyến (online analytical mining)?
  - Dữ liệu chất lượng cao trong DW
    - DW chứa dữ liệu đã tích hợp, nhất quán và được làm sạch
  - Cấu trúc xử lý thông tin sẵn dùng với các DW
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - Phân tích dữ liệu dựa trên OLAP
    - Khai phá với drilling, dicing, pivoting, v.v.
  - Lựa chọn các chức năng khai phá dữ liệu trực tuyến
    - Tích hợp và hoán đổi nhiều chức năng, giải thuật và công việc khai phá dữ liệu

Nội dung

❶ Kho dữ liệu: các khái niệm cơ bản

❷ Mô hình hóa DW: Data Cube và OLAP

❸ Thiết kế và sử dụng DW

❹ Thể hiện DW

❺ Tổng quát hóa dữ liệu bằng quy nạp hướng thuộc tính

Tính toán Data Cube hiệu quả

- Data cube có thể được xem như một lưới các cuboid
  - Cuboid mức thấp nhất là cuboid cơ sở
  - Cuboid mức cao nhất (apex) chứa chỉ một giá trị
  - Có bao nhiêu cuboid trong một cube n chiều với L mức?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

- Materialization data cube
  - Materialize mọi (cuboid) (**full materialization**), không (không **materialization**), hay một số **(materialization một phần)**
  - Chọn cuboid nào được materialize
    - Dựa trên kích thước, việc chia sẻ và tần suất truy cập, v.v.

Toán tử "Compute Cube"

- Xác định Cube và tính toán trong DMQL

  define cube sales [item, city, year]: sum (sales_in_dollars)

  compute cube sales

- Chuyển đổi nó vào ngôn ngữ giống SQL (với toán tử mới cube by, được giới thiệu bởi Gray et al.'96)

  SELECT item, city, year, SUM (amount)

  FROM SALES

  CUBE BY item, city, year

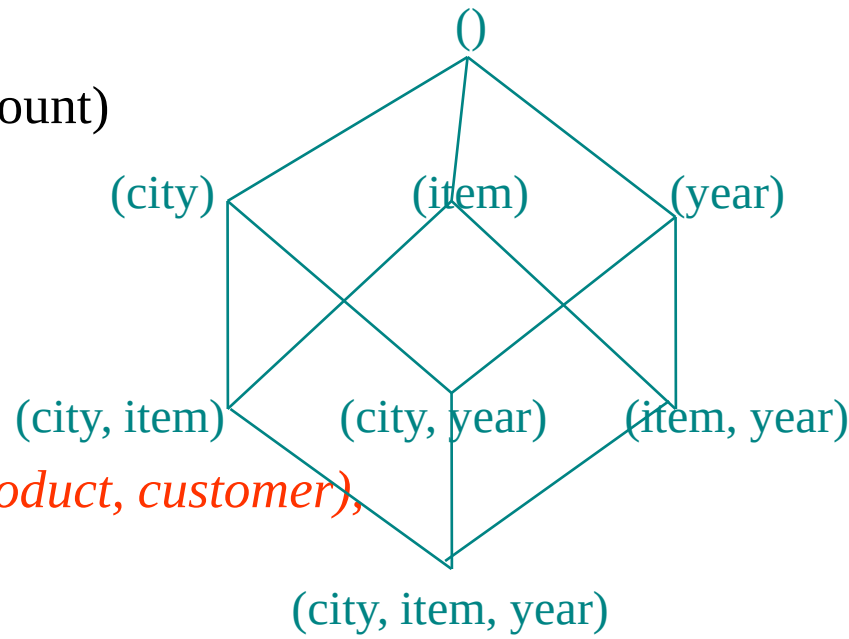
- Cần tính toán các Group-By sau:

  (date, product, customer),

  (date,product),(date, customer), (product, customer),

  (date), (product), (customer)

  ()



Chỉ mục dữ liệu OLAP: Chỉ mục Bitmap

- Chỉ mục trên một cột cụ thể
- Mỗi giá trị của cột có một véc tơ bít: phép toán tên bít thường được thực hiện nhanh hơn
- Chiều dài của véc tơ bít: là số bản ghi của bảng gốc
- Bít thứ *i (i*-th) là 1 nếu hàng thứ *i* của bảng gốc có giá trị của cột chỉ mục
- Không phù hợp đối với các miền có nhiều giá trị
- Kỹ thuật nét bít hiện nay, Word-Aligned Hybrid (WAH), cho phép làm việc với các miền có nhiều giá trị [Wu, et al. TODS'06]

Base table

Cust	Region	Type
C1	Asia	Retail
C2	Europe	Dealer
C3	Asia	Dealer
C4	America	Retail
C5	Europe	Dealer

Index on Region

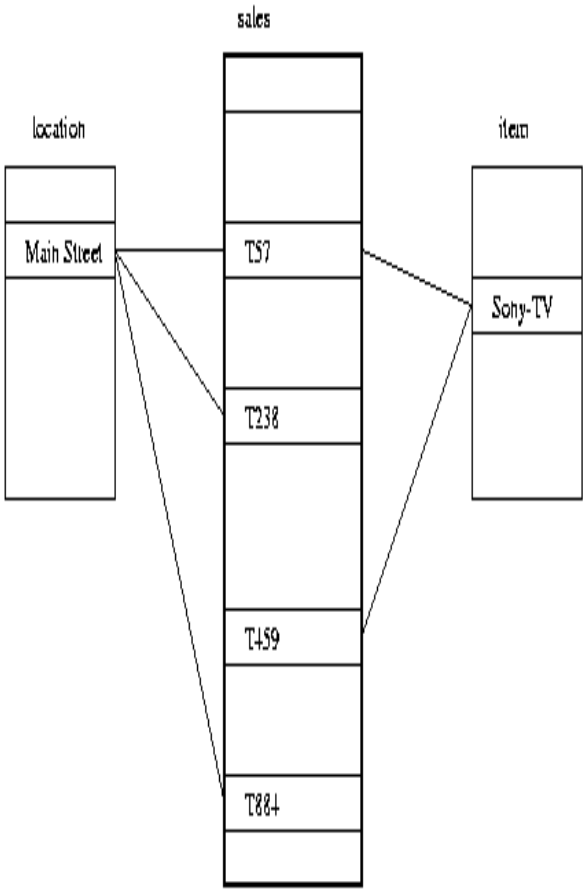
RecID	Asia	Europe	America
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Index on Type

RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

Chỉ mục dữ liệu OLAP: Kết nối các chỉ mục

- Kết nối chỉ mục: JI(R-id, S-id) với R (R-id, …) ⊳⊲ S (S-id, …)
- Các chỉ mục truyền thống ánh xạ các giá trị trong danh sách tới danh sách các định danh bản ghi
  - Nó materialize kết nối quan hệ trong file JI và tăng tốc kết nối quan hệ
- Trong các DW, kết nối chỉ mục liên quan các giá trị của các chiều trong lược đồ hình sao vào các hàng trong bảng fact.
  - Thí dụ bảng fact: *Sales* và 2 bảng chiều *city* và *product*
    - Kết nối chỉ mục trên *city* đảm bảo mỗi thành phố trong danh sách R-ID của các bộ ghi nhận bán hàng ở thành phố xác định
  - Kết nối chỉ mục có thể mở rộng cho nhiều chiều



Xử lý truy vấn OLAP hiệu quả

- **Xác định các toán tử** được thực hiện trên các cuboids tương ứng
  - Chuyển đổi các phép toán drill, roll, v.v. vào các truy vấn SQL và OLAP tương ứng, thí dụ: dice = selection + projection
- **Xác định cuboid materialized nào** được chọn cho toán tử OLAP
  - Cho truy vấn được xử lý trên {*brand*, *province_or_state*} với điều kiện "*year = 2004*", và có 4 materialized cuboid:

    1) {*year, item_name, city*}

    2) {*year, brand, country*}

    3) {*year, brand, province_or_state*}

    4) {*item_name, province_or_state*} where *year = 2004*

    Cuboid nào nên được chọn để thực hiện truy vấn?
- Khám phá các cấu trúc chỉ mục và các cấu trúc mảng nén so với dày đặc trong MOLAP

Các kiến trúc máy chủ OLAP

- Relational OLAP (ROLAP)
  - Sử dụng hệ quản trị CSDL quan hệ hay quan hệ mở rộng để lưu trữ và quản lý DW và kho trung gian OLAP
  - Bao gồm việc tối ưu backend CSDL, thể hiện các logic tổ hợp các các công cụ, dịch vụ bổ sung
  - Khả năng mở rộng lớn hơn
- Multidimensional OLAP (MOLAP)
  - Cơ chế lưu trữ đa chiều dựa trên mảng thưa
  - Chỉ mục nhanh để tổng hợp trước các dữ liệu
- Hybrid OLAP (HOLAP) (thí dụ: Microsoft SQLServer)
  - Mềm dẻo, thí dụ: mức thấp là quan hệ, mức cao là mảng
- Máy chủ SQL đặc biệt (Specialized SQL server) (thí dụ: Redbricks)
  - Hỗ trợ đặc biệt cho các truy vấn SQL trên các lược đồ hình sao và chòm sao

Nội dung

❶ Kho dữ liệu: các khái niệm cơ bản

❷ Mô hình hóa DW: Data Cube và OLAP

❸ Thiết kế và sử dụng DW

❹ Thể hiện DW

❺ Tổng quát hóa dữ liệu bằng quy nạp hướng thuộc tính

Quy nạp hướng thuộc tính (Attribute-Oriented Induction)

- Được đề xuất năm 1989 (trong workshop KDD ‘89)
- Không hạn chế ở dữ liệu phân loại hay các giá trị đo đặc biệt
- Được làm thế nào?
  - Thu thập dữ liệu liên quan (tạo *initial relation*) sử dụng truy vấn CSDL quan hệ
  - Thực hiện việc tổng quát hóa bằng cách loại bỏ thuộc tính (attribute removal) hay tổng quát hóa thuộc tính (attribute generalization)
  - Áp dụng phép tổng bằng cách hợp nhất các bộ đồng nhất hay tổng quát hóa và tích lũy các số đếm tương ứng của chúng
  - Tương tác với người dùng để thể hiện tri thức

Thí dụ

Thí dụ: Mô tả đặc trưng chung của các sinh viên tốt nghiệp của một trường đại học

- Bước 1. Lấy tập dữ liệu liên quan sử dụng câu lệnh SQL, thí dụ:

  Select * (i.e., name, gender, major, birth_place, birth_date, residence, phone#, gpa)

  from student

  where student_status in {“Msc”, “MBA”, “PhD” }

- Bước 2. Thực hiện quy nạp hướng thuộc tính
- Bước 3. Thể hiện các kết quả của quan hệ tổng quát hóa, bảng chéo, hay các dạng luật

Thí dụ: Đặc trưng lớp

Quan hệ ban đầu

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Quan hệ tổng quát hóa

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
…	…	…	…	…	…	…

Gender \ Birth_Region	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

Các nguyên tắc cơ bản của quy nạp hướng thuộc tính

- Tập trung vào dữ liệu - Data focusing: dữ liệu liên quan, bao gồm các chiều, và kết quả là một quan hệ khởi tạo (*initial relation*)
- Loại bỏ thuộc tính - Attribute-removal: loại bỏ thuộc tính *A* nếu có số lớn các giá trị phân biệt đối với *A* mà (1) không có toán tử tổng quát hóa đối với *A*, hay (2) các khái niệm mức cao hơn của *A* được thể hiện thông qua các thuộc tính khác
- Tổng quát hóa thuộc tính - Attribute-generalization: Nếu có tập lớn giá trị phân biệt đối với *A*, và tồn tại tập toán tử tổng quát hóa đối với *A*, thì chọn một toán tử và tổng quát hóa *A*
- Điều khiển ngưỡng thuộc tính - Attribute-threshold control: thường giữ từ 2 tới 8 nhóm
- Điều khiển ngưỡng quan hệ tổng quát hóa - Generalized relation threshold control: điều khiển kích thước quan hệ/luật cuối cùng

Quy nạp hướng thuộc tính: Giải thuật cơ bản

- InitialRel: Truy vấn xử lý các dữ liệu liên quan để tạo quan hệ ban đầu (*initial relation*).
- PreGen: Dựa trên phân tích về số giá trị phân biệt ở mỗi thuộc tích, xác định kế hoạch tổng quát hóa cho mỗi thuộc tính: loại bỏ? hay tổng quát hóa thành mức cao hơn thế nào?
- PrimeGen: Dựa trên kế hoạch PreGen, thực hiện việc tổng quát hóa tới mức hợp lý để tạo quan hệ tổng quát hóa với việc tích lũy các số đếm.
- Presentation: Tương tác người dùng: (1) điều chỉnh các mức chi tiết, (2) xoay, (3) ánh xạ vào các luật, bảng chéo, các thể hiện trực quan.

Thể hiện các kết quả tổng quát hóa

- Quan hệ tổng quát hóa:
  - Các quan hệ mà một số hay tất cả các thuộc tính của nó được tổng quát hóa, với các giá trị số đếm hay tổ hợp khác được tích lũy.
- Bảng chéo:
  - Ánh xạ các kết quả vào dạng bảng chéo.
  - Các kỹ thuật trực quan hóa:
  - Pie charts, bar charts, curves, cubes, và các dạng trực quan khác.
- Các quy tắc đặc trưng định lượng:
  - Ánh xạ kết quả tổng quát hóa thành các luật đặc trưng với vác thông tin định lượng liên kết, thí dụ:

$$grad(x) \wedge male(x) \Rightarrow$$
$$birth\_region(x) = \text{"}Canada\text{"}[t:53\%] \vee birth\_region(x) = \text{"}foreign\text{"}[t:47\%].$$

Khai phá so sánh lớp

- So sánh: So sánh 2 hay nhiều lớp
- Phương pháp:
  - Phân mảnh tập dữ liệu liên quan vào lớp đích và các lớp đối nghịch
  - Tổng quát hóa cả 2 lớp tới cùng mức khái niệm
  - So sánh các bộ với các mô tả cùng mức cao
  - Thể hiện sự mô tả của mọi bộ và 2 số đo
    - support – phân phối trong một lớp
    - comparison – phân phối giữa các lớp
  - Nhấn mạnh (highlight) các bộ với các đặc điểm phân biệt mạnh
- Phân tích mức độ liên quan:
  - Tìm các thuộc tính (đặc trưng) tốt nhất để phân biệt các lớp

Mô tả khái niệm và OLAP dựa trên Cube

- **Giống nhau**:
  - Tổng quát hóa dữ liệu
  - Thể hiện tổng hợp dữ liệu ở nhiều mức trừu tượng
  - Tương tác drilling, pivoting, slicing và dicing
- **Khác nhau**:
  - OLAP thực hiện tiền xử lý một cách hệ thống, độc lập truy vấn và có thể đào sâu tới mức thấp hơn
  - AOI (Attribute Oriented Induction) tự động phân bổ mức mong muốn, và có thể thực hiện các phân tích/xếp hạng liên quan khi có nhiều chiều liên quan
  - AOI làm việc với dữ liệu không thuộc dạng quan hệ

Bài tập

- Cài đặt và học sử dụng MS. Power BI