

Chương 3:

CÔNG NGHỆ KHO DỮ LIỆU VÀ PHÂN TÍCH TRỰC TUYẾN

1. Khái niệm về kho dữ liệu.
2. Mô hình dữ liệu đa chiều
3. Kiến trúc của kho dữ liệu.
4. **Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến.**
5. **Liên hệ công nghệ kho dữ liệu với khai phá dữ liệu.**
6. **Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định.**

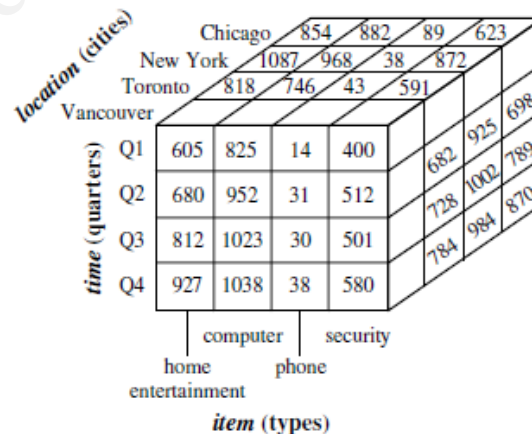
Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – **OLAP**):
 - Làm việc với dữ liệu đã được biến đổi.
 - Sử dụng các bảng chiều (dimension table) và bảng sự kiện (fact table) tạo khối (cube) cho dữ liệu nhằm thể hiện sự đa chiều cho dữ liệu.
 - Hỗ trợ người dùng phân tích dữ liệu qua việc cắt lát (**slice**) dữ liệu theo các khía cạnh khác nhau:
 - ✓ Khoan xuống (**drill down**): khai thác chi tiết của dữ liệu.
 - ✓ Cuộn lên (**drill up**): khai thác dữ liệu qua việc tổng hợp từ mức thấp lên mức cao

Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – OLAP):

| <i>location</i> = "Chicago" | | | | | <i>location</i> = "New York" | | | | | <i>location</i> = "Toronto" | | | | | <i>location</i> = "Vancouver" | | | | |
|-----------------------------|-------------|--------------|--------------|-------------|------------------------------|--------------|--------------|-------------|-----|-----------------------------|--------------|--------------|-------------|------|-------------------------------|--------------|--------------|-------------|--|
| <i>Item</i> | | | | | <i>Item</i> | | | | | <i>Item</i> | | | | | <i>Item</i> | | | | |
| <i>home</i> | | | | | <i>home</i> | | | | | <i>home</i> | | | | | <i>home</i> | | | | |
| <i>time</i> | <i>ent.</i> | <i>comp.</i> | <i>phone</i> | <i>sec.</i> | <i>ent.</i> | <i>comp.</i> | <i>phone</i> | <i>sec.</i> | | <i>ent.</i> | <i>comp.</i> | <i>phone</i> | <i>sec.</i> | | <i>ent.</i> | <i>comp.</i> | <i>phone</i> | <i>sec.</i> | |
| Q1 | 854 | 882 | 89 | 623 | 1087 | 968 | 38 | 872 | 818 | 746 | 43 | 591 | 605 | 825 | 14 | 400 | | | |
| Q2 | 943 | 890 | 64 | 698 | 1130 | 1024 | 41 | 925 | 894 | 769 | 52 | 682 | 680 | 952 | 31 | 512 | | | |
| Q3 | 1032 | 924 | 59 | 789 | 1034 | 1048 | 45 | 1002 | 940 | 795 | 58 | 728 | 812 | 1023 | 30 | 501 | | | |
| Q4 | 1129 | 992 | 63 | 870 | 1142 | 1091 | 54 | 984 | 978 | 864 | 59 | 784 | 927 | 1038 | 38 | 580 | | | |



Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – **OLAP**):
 - Ngôn ngữ truy vấn khai phá dữ liệu (**Data Mining Query Language – DMQL** – Các hàm nguyên thủy):
 - ✓ **define cube** <tên_khối>[<danh_sách_các_chiều>]:
<danh_sách_các_độ_đo>
 - ✓ **Define dimension** <tên_chiều> **as** <tên_chiều_được_khai_báo_lần_đầu> **in cube** <Tên_khối_đầu_tiên_sử_dụng_chiều_đó>

Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – **OLAP**):
 - Ngôn ngữ truy vấn khai phá dữ liệu (**Data Mining Query Language – DMQL** – Các hàm nguyên thủy):
 - ✓ Thuộc tính độ đo: Là một hàm tính toán trên những dữ liệu đã được tích hợp lại dựa trên những cặp giá trị theo chiều cho trước. Có 3 loại như sau:
 - ✓ Phân phối: count(); sum(); min(), max().
 - ✓ Đại số: avg() = sum()/count(), min_N(), standard_deviation().
 - ✓ Khác: median(), mode(), rank().

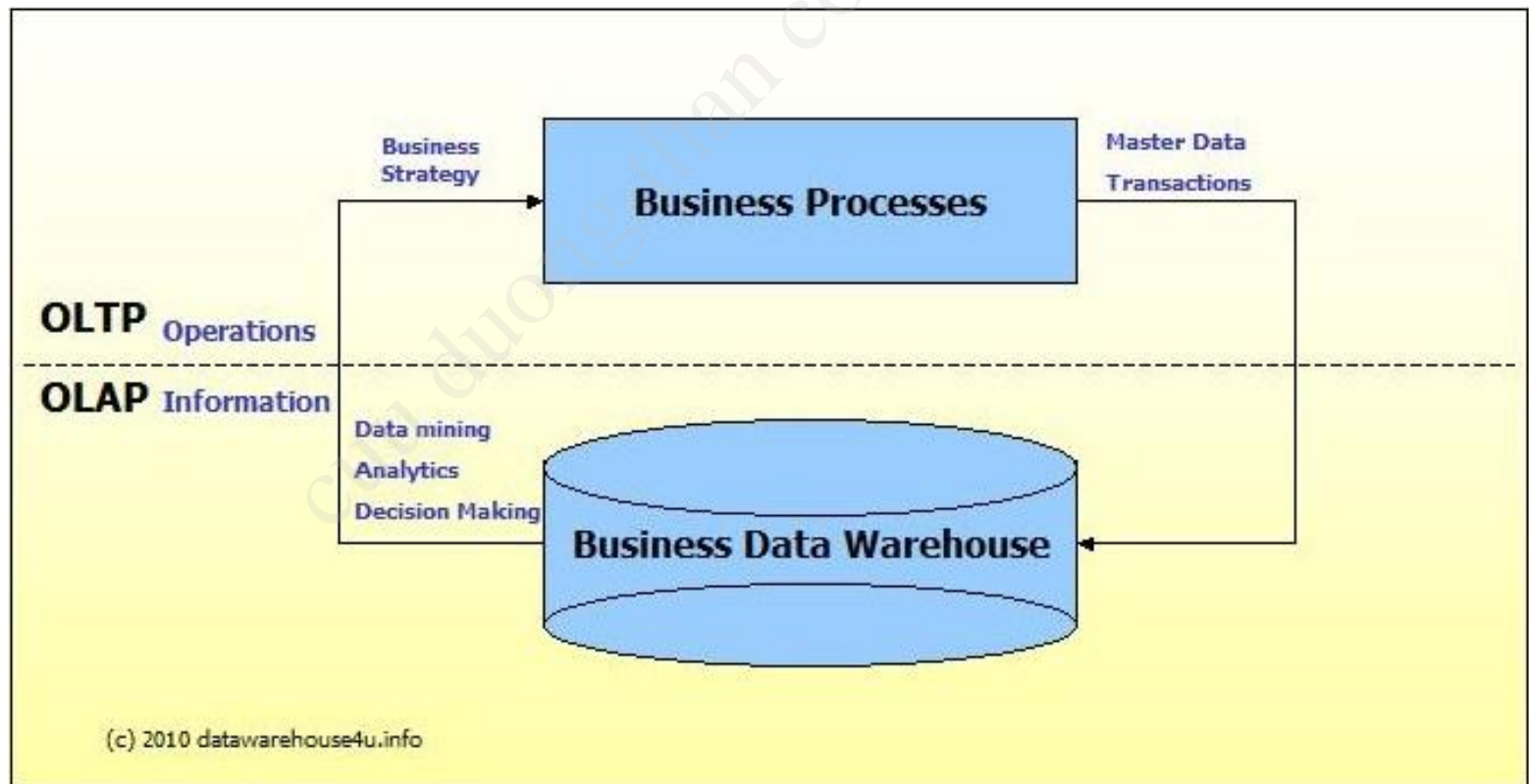
Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – OLAP):

- Ngôn ngữ truy vấn khai phá dữ liệu (Data Mining Query Language – DMQL – Các hàm nguyên thủy):
- Ví dụ:
`define cube sales [time, item, branch, location]: dollars sold = sum(sales in dollars), units sold = count(*)`
`define dimension time as (time key, day, day of week, month, quarter, year)`
`define dimension item as (item key, item name, brand, type, supplier type)`
`define dimension branch as (branch key, branch name, branch type)`
`define dimension location as (location key, street, city, province or state, country)`

Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – **OLAP**):





Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Xử lý phân tích trực tuyến (On-line Transaction Processing – **OLAP**):
 - **OLTP** (Online Transaction Processing) – xử lý giao tác trực tuyến:
 - ✓ Hệ thống có nhiều người dùng đồng thời, thao tác (thêm, xóa, sửa) trên dữ liệu.
 - ✓ Thường dùng cho mục đích thu thập dữ liệu.
 - ✓ Các vấn đề có thể phát sinh:
 - Dữ liệu quá lớn, chi phí về thời gian cao,
 - Vấn đề phân quyền,
 - Sự phức tạp của CSDL quan hệ đối với người phân tích.
 - ✓ Khắc phục sự phức tạp: tạo bản sao để phân tích

Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Các kiến trúc của máy chủ cho việc xử lý phân tích trực tuyến:
 - OLAP quan hệ (Relation OLAP – ROLAP):
 - ✓ Dùng hệ quản trị CSDL quan hệ hoặc quan hệ mở rộng để lưu trữ và quản lý kho dữ liệu.
 - ✓ Bao gồm sự tối ưu hóa các công việc nền tảng của CSDL cũng như các công cụ phụ trợ bổ sung và các dịch vụ.
 - ✓ Có khả năng mở rộng thêm.
 - ✓ Dung lượng Cube chỉ giới hạn bởi dung lượng của cơ sở dữ liệu quan hệ



Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Các kiến trúc của máy chủ cho việc xử lý phân tích trực tuyến (tt):
 - **OLAP đa chiều** (MultiDimensional OLAP – **MOLAP**):
 - ✓ Lưu trữ mảng dữ liệu đa chiều dựa trên cấu trúc mảng (thường dùng kỹ thuật ma trận thưa).
 - ✓ Lập chỉ mục nhanh để tính toán trước khi tổng hợp dữ liệu.
 - ✓ Tồn bộ nhớ
 - ✓ Không xem được dữ liệu mới cho đến khi xây dựng lại Cube.

Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Các kiến trúc của máy chủ cho việc xử lý phân tích trực tuyến (tt):
 - OLAP lai (Hybrid OLAP – HOLAP):
 - ✓ Người dùng sử dụng ROLAP và MOLAP một cách linh hoạt.
 - ✓ Dữ liệu yêu cầu là dạng tổng hợp thì sẽ thực hiện truy vấn tại OLAP.
 - ✓ Dữ liệu yêu cầu là dạng chi tiết thì truy vấn sẽ được dịch và truy vấn tại cơ sở dữ liệu quan hệ.



Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

- ❖ Các kiến trúc của máy chủ cho việc xử lý phân tích trực tuyến (tt):
 - Các máy chủ SQL chuyên dụng:
 - ✓ Chuyên hỗ trợ cho các truy vấn SQL trên lược đồ hình sao hoặc lược đồ hình bông tuyết.

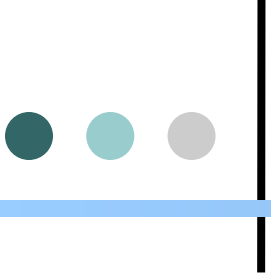
Cài đặt kho dữ liệu và Xử lý phân tích trực tuyến

❖ Công cụ phân tích trực tuyến:

➤ SQL Server Data Tools - Business Intelligence (SSDT-BI):

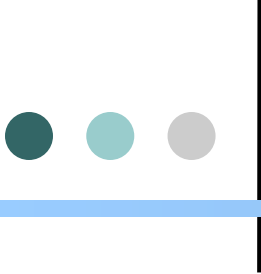
- ✓ Công cụ cho phép thực hiện OLAP là “*SQL Server Business Intelligence Development Studio - BIDS*”.
- ✓ Microsoft SQL Server Data Tools - Business Intelligence for Visual Studio 2013:
SSDTBI_x86_ENU.exe.

➤ ORACLE: Oracle Business Intelligence



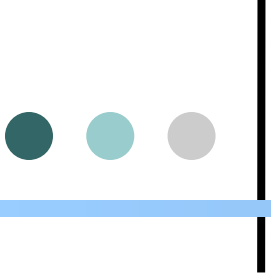
Liên hệ công nghệ kho dữ liệu với khai phá dữ liệu

- ❖ Ứng dụng kho dữ liệu:
 - **Xử lý thông tin**: hỗ trợ việc truy vấn thông tin, phân tích thống kê cơ bản và làm báo cáo sử dụng các bảng tham chiếu chéo, các bảng, các biểu đồ và đồ thị.
 - **Xử lý phân tích**: dùng cho phân tích đa chiều của kho dữ liệu, hỗ trợ các thao tác OLAP cơ bản, cắt ngang, cắt dọc, khoan sâu, xoa.
 - **Khai phá** dữ liệu



Liên hệ công nghệ kho dữ liệu với khai phá dữ liệu

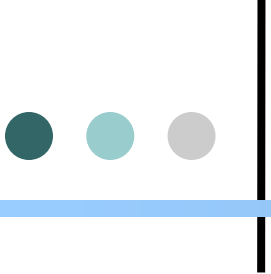
- ❖ Từ xử lý phân tích trực tuyến (**OLAP**) tới khai phá phân tích trực tuyến (**OLAM – Online Analytical Mining**) – Do các yếu tố:
 - Dữ liệu trong kho dữ liệu là loại dữ liệu có chất lượng cao, đã được làm sạch, đồng nhất và tích hợp.
 - Các cấu trúc xử lý thông tin sẵn có xung quanh các kho dữ liệu như ODBC (kết nối dữ liệu), OLEDB (nhúng cơ sở dữ liệu), truy nhập Web, các dịch vụ tiện tích, các công cụ OLAP và báo cáo.
 - Phân tích dữ liệu thăm dò dựa trên OLAP: có thể khai phá với các phép toán khoan sâu, cắt lát, xoay, v.v...
 - Lựa chọn trực tuyến các chức năng khai phá dữ liệu: tích hợp và hoán đổi nhiều chức năng khai thác khác nhau, các thuật toán và nhiệm vụ khác nhau.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

❖ Các giai đoạn xây dựng:

1. Lập kế hoạch
2. Thu thập yêu cầu về dữ liệu và mô hình hóa.
3. Thiết kế và Phát triển cơ sở dữ liệu vật lý.
4. Dữ liệu bản đồ và sự biến đổi
5. Khai thác dữ liệu và tải
6. Tự động hoá việc Quy trình quản lý dữ liệu.
7. Phát triển ứng dụng - Tạo tập khởi đầu của báo cáo.
8. Xác Nhận và kiểm tra dữ liệu.
9. Đào tạo.
10. Triển khai



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

1. Lập kế hoạch

- Xác định phạm vi dự án.
- Tạo ra kế hoạch dự án.
- Xác định các nguồn lực cần thiết, cả trong và ngoài.
- Xác định nhiệm vụ và các sản phẩm phân phối.
- Xác định thời hạn của dự án.
- Xác định sản phẩm phân phối cuối cùng của dự án.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

1. Lập kế hoạch (tt)

- Lập kế hoạch về hiệu năng của dự án:
 - ✓ Tính toán kích cỡ bản ghi cho mỗi bảng.
 - ✓ Ước tính số lượng bản ghi ban đầu cho mỗi bảng
 - ✓ Xem lại các yêu cầu truy cập kho dữ liệu để dự đoán yêu cầu về tập chỉ mục.
 - ✓ Xác định các yếu tố tăng trưởng cho mỗi bảng.
 - ✓ Xác định bảng mục tiêu lớn nhất dự kiến trong một giai đoạn thời gian được lựa chọn và thêm khoảng 25-30% dự trù tới kích thước bảng để xác định kích thước lưu trữ tạm thời.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

2. Thu thập các yêu cầu dữ liệu và mô hình hóa:

➤ Các câu hỏi cần trả lời:

- ✓ Người sử dụng thực hiện các công việc nghiệp vụ như thế nào?
- ✓ Hiệu suất của người dùng được đo như thế nào?
- ✓ Những thuộc tính nào người sử dụng cần?
- ✓ Các phân cấp trong nghiệp vụ kinh doanh của hệ thống là gì?
- ✓ Những dữ liệu nào người dùng hiện nay đang sử dụng và họ muốn có dữ liệu nào trong tương lai?
- ✓ Người dùng cần dữ liệu tổng hợp hay chi tiết ở mức độ nào?

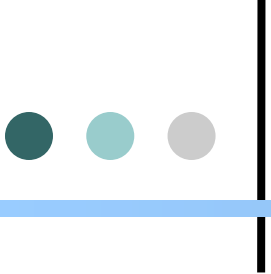


Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

2. Thu thập các yêu cầu dữ liệu và mô hình hóa:

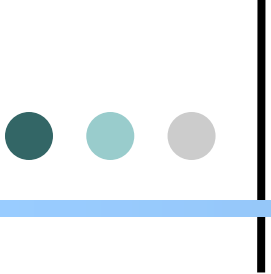
➤ Các dạng mô hình hóa:

- ✓ Mô hình dữ liệu logic bao phủ phạm vi của dự án phát triển bao gồm:
 - Các mối quan hệ,
 - Loại liên kết giữa các quan hệ,
 - Các thuộc tính,
 - Các khóa ứng viên (**candidate keys**).
- ✓ Mô hình nghiệp vụ nhiều chiều được thể hiện qua các bảng Fact, các chiều, các phân cấp, các mối quan hệ và các khóa ứng cử viên cho các phạm vi phát triển của dự án.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

3. Thiết kế và Phát triển cơ sở dữ liệu vật lý:
- Thiết kế cơ sở dữ liệu, bao gồm các bảng Fact, các bảng quan hệ, và các bảng mô tả (dùng cho việc tra cứu).
 - Phi chuẩn dữ liệu,
 - Xác định các khóa,
 - Tạo các chiến lược lập chỉ mục,
 - Tạo các đối tượng cơ sở dữ liệu thích hợp.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

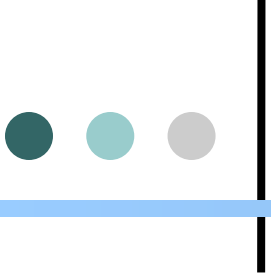
4. Ánh xạ và chuyển đổi dữ liệu:
- Xác định hệ thống nguồn.
 - Xác định cách bố trí tập tin.
 - Phát triển các yêu cầu chi tiết kỹ thuật chuyển đổi bằng văn bản cho các biến đổi phức tạp.
 - Ánh xạ nguồn tới dữ liệu đích.
 - Xem xét lại các kế hoạch về hiệu năng .



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

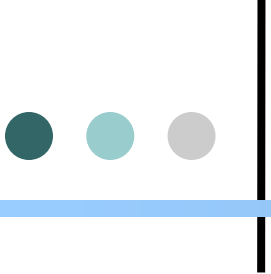
5. Hình thành kho dữ liệu:

- Phát triển các thủ tục để trích xuất và di chuyển dữ liệu vào kho.
- Phát triển các thủ tục để nạp dữ liệu vào kho.
- Phát triển chương trình phần mềm hoặc dùng các công cụ chuyển đổi dữ liệu để chuyển đổi và tích hợp dữ liệu.
- Kiểm thử việc trích xuất, chuyển đổi và các thủ tục tải dữ liệu.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

6. Thủ tục quản lý dữ liệu tự động:
- Tự động hoá và lập lịch cho quá trình tải dữ liệu.
 - Tạo sao lưu dữ liệu và các thủ tục phục hồi.
 - Tiến hành một thử nghiệm đầy đủ của tất cả các thủ tục tự động



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

7. Phát triển ứng dụng – Tạo ra một tập khởi đầu cho các báo cáo:
- Tạo tập khởi đầu cho các báo cáo được định trước.
 - Phát triển các báo cáo cơ bản quan trọng.
 - Kiểm thử tính đúng đắn của các báo cáo.
 - Viết tài liệu cho ứng dụng.
 - Phát triển các đường dẫn để điều hướng.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

8. Xác nhận và kiểm thử dữ liệu:

- Xác nhận dữ liệu bằng cách sử dụng tập khởi đầu cho các báo cáo.
- Xác nhận dữ liệu bằng cách sử dụng các quy trình chuẩn.
- Lặp đi lặp lại thay đổi dữ liệu.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

9. Đào tạo:

Để khai thác hiệu quả, người dùng cần được đào tạo về:

- Phạm vi của dữ liệu trong kho.
- Công cụ truy nhập đầu cuối và cách thức hoạt động nó.
- Việc ứng dụng các DDS hoặc tập khởi tạo các báo cáo bao gồm cả các khả năng ứng dụng và đường dẫn chuyển hướng.
- Liên tục đào tạo và hỗ trợ người sử dụng khi hệ thống thay đổi.



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

10. Triển khai:

- Cài đặt cơ sở hạ tầng vật lý cho tất cả người dùng.
- Phát triển ứng dụng DDS.
- Tạo thủ tục cho việc thêm các báo cáo mới và mở rộng việc áp dụng Hệ hỗ trợ quyết định (**DSS**).
- Thiết lập các thủ tục để sao lưu các ứng dụng **DSS**, không phải chỉ là kho dữ liệu.
- Tạo thủ tục điều tra và giải quyết các vấn đề liên quan tới toàn vẹn dữ liệu



Xây dựng kho dữ liệu với mục đích hỗ trợ quyết định

❖ Thiết kế cơ sở dữ liệu

➤ Lược đồ hình sao:

- ✓ Dễ hiểu đối với những người phân tích và người dùng cuối.
- ✓ Truy vấn nhanh.
- ✓ Bảng Fact: Chứa dữ liệu thực tế định lượng trong doanh nghiệp. Bảng này dữ liệu có thể rất lớn.
- ✓ Bảng theo chiều: Chứa dữ liệu mô tả các yếu tố ảnh hưởng tới doanh nghiệp.

➤ Lược đồ bông tuyết.