DATA WAREHOUSES AND DECISION SUPPORT SYSTEMS

(DW and DSS)

Chương 2
Các vấn đề cơ bản trong kho dữ liệu

Vietnam National University – HCMC
Ho Chi Minh City University of Technology
Faculty of Computer Science & Engineering


Mr. Bui Tien Duc, Meng


tienducut@gmail.com


0769690731

Introduction

1. Các khái niệm cơ bản về kho dữ liệu
2. Kiến trúc kho dữ liệu
3. Các đặc tính về dữ liệu của kho dữ liệu
4. Vấn đề về độ mịn dữ liệu
5. Vấn đề về chuyển đổi dữ liệu
6. Vấn đề về dữ liệu dẫn xuất
7. Vấn đề về siêu dữ liệu
8. Các vấn đề khác về kho dữ liệu

1. Các khái niệm cơ bản về kho dữ liệu

Tổng quan:

Kho dữ liệu (Data Warehouse) là **tập hợp dữ liệu** được tổ chức theo hướng chủ đề, tích hợp, không thay đổi (non-volatile), có tính lịch sử (time-variant) nhằm hỗ trợ quá trình ra quyết định của nhà quản lý. Đây là nền tảng cho các hệ thống DSS (Decision Support Systems) và BI (Business Intelligence) hiện đại

Đặc điểm chính:

Subject-oriented (hướng theo chủ đề)

Integrated (tích hợp dữ liệu từ nhiều nguồn khác nhau)

Non-volatile (dữ liệu không thay đổi sau khi đã nạp vào)

Time-variant (có yếu tố thời gian, phục vụ phân tích lịch sử)

1. Các khái niệm cơ bản về kho dữ liệu

Tổng quan:

1. Các khái niệm cơ bản về kho dữ liệu

Tổng quan

Kho dữ liệu (Data Warehouse) được xem là trung tâm của môi trường dữ liệu có kiến trúc (architected environment). Nó đóng vai trò là nền tảng cho tất cả hoạt động xử lý thông tin hỗ trợ ra quyết định (DSS).

Khái niệm cơ bản về kho dữ liệu (**Inmon Wiley**):

"A Data Warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision-making process."

1. Các khái niệm cơ bản về kho dữ liệu

Tổng quan

Vai trò DW and DSS

1. Các khái niệm cơ bản về kho dữ liệu

Khái niệm cơ bản về kho dữ liệu (**Inmon Wiley**): (thi cuối kỳ)

"A Data Warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision-making process."

"Kho dữ liệu là một tập hợp dữ liệu có định hướng theo chủ đề, được tích hợp, không biến đổi và có tính biến thiên theo thời gian, nhằm hỗ trợ cho quá trình ra quyết định của nhà quản lý."

Trong đó, (4 đặc trưng của Kho dữ liệu)

Subject-oriented (định hướng theo chủ đề): tổ chức dữ liệu quanh các chủ đề phân tích.

Integrated (tích hợp): dữ liệu từ nhiều nguồn được chuẩn hoá thành một hình ảnh thống nhất.

Nonvolatile (không biến đổi): dữ liệu chủ yếu chỉ đọc, không thay đổi sau khi nạp.

Time-variant (biến thiên theo thời gian): dữ liệu gắn với yếu tố thời gian, lưu trữ cả lịch sử để phân tích xu hướng.

1. Các khái niệm cơ bản về kho dữ liệu

Các đặc trưng cốt lõi của kho dữ liệu

1. Subject-Oriented – Định hướng theo chủ đề

Hệ thống tác nghiệp (Operational Systems) thường tổ chức dữ liệu quanh chức năng (bán hàng, nhân sự, bảo hiểm, sản xuất…).

Ngược lại, kho dữ liệu tập trung quanh các chủ đề phân tích:

Bảo hiểm: khách hàng, hợp đồng, phí bảo hiểm, bồi thường.

Sản xuất: sản phẩm, đơn hàng, nhà cung cấp.

Bán lẻ: sản phẩm, SKU, giao dịch, nhà cung cấp.

Điều này giúp nhà quản lý dễ dàng phân tích theo đối tượng quan tâm chính thay vì bị phân mảnh bởi ứng dụng.

1. Các khái niệm cơ bản về kho dữ liệu

Các đặc trưng cốt lõi của kho dữ liệu

2. Integrated – Tích hợp

Quan trọng nhất trong 4 đặc trưng.

Dữ liệu đưa vào kho dữ liệu được trích từ nhiều nguồn khác nhau (ERP, CRM, POS, file Excel, logs web…).

Trong quá trình ETL, dữ liệu sẽ được:

Chuyển đổi (reformat),

Chuẩn hóa (standardize),

Tổng hợp (summarize),

Đồng nhất mã hoá (naming conventions, coding schemes).

Kết quả: hình thành "một bức tranh dữ liệu thống nhất của toàn công ty" – cái mà Inmon gọi là single corporate image..

1. Các khái niệm cơ bản về kho dữ liệu

Các đặc trưng cốt lõi của kho dữ liệu

3. Nonvolatile – Không biến đổi

Dữ liệu trong kho không bị cập nhật trực tiếp như trong hệ thống giao dịch.

Các thao tác chính:

Load (nạp dữ liệu)

Access (truy vấn/đọc dữ liệu)

Không có Update/Delete thường xuyên.

Lợi ích:

Tăng tính ổn định và đáng tin cậy của dữ liệu.

Giúp phân tích lịch sử không bị sai lệch.

1. Các khái niệm cơ bản về kho dữ liệu

Các đặc trưng cốt lõi của kho dữ liệu

4. Time-Variant – Biến thiên theo thời gian

Dữ liệu luôn gắn với yếu tố thời gian (timestamp, ngày hiệu lực, ngày hết hạn, kỳ báo cáo).

Hệ thống OLTP thường chỉ lưu dữ liệu hiện tại.

Kho dữ liệu lưu dữ liệu lịch sử nhiều năm để hỗ trợ phân tích xu hướng, dự báo, và ra quyết định chiến lược.

Ví dụ:

Báo cáo “Doanh thu quý 3 trong 5 năm gần nhất”

Phân tích “Xu hướng hành vi khách hàng từ 2019–2024”.

1. Các khái niệm cơ bản về kho dữ liệu

Kho dữ liệu (Data Warehouse) không chỉ là một tập hợp dữ liệu thông thường, mà dữ liệu trong đó mang những đặc tính riêng biệt để phục vụ cho phân tích và ra quyết định quản trị. Các đặc tính này đã được Inmon khái quát thành bốn trụ cột cốt lõi và *một số khía cạnh mở rộng (Các đặc tính của kho dữ liệu mở rộng khác)*

Các đặc tính của kho dữ liệu mở rộng khác

1. Granularity (Độ chi tiết dữ liệu)

Quyết định mức độ chi tiết hay tổng hợp của dữ liệu lưu trữ.

Dữ liệu càng chi tiết thì càng linh hoạt cho phân tích, nhưng dung lượng càng lớn.

Thực tế: Kho thường lưu song song cả dữ liệu chi tiết và dữ liệu tổng hợp .

1. Các khái niệm cơ bản về kho dữ liệu

Các đặc tính của kho dữ liệu mở rộng khác

2. Partitioning (Phân vùng dữ liệu)

Chia nhỏ dữ liệu để quản lý và tối ưu hiệu suất.

Có thể phân vùng theo thời gian, khu vực địa lý, loại sản phẩm.

Giúp cải thiện tốc độ truy vấn và quản trị dữ liệu dễ dàng hơn.

3. Homogeneity vs. Heterogeneity (Tính đồng nhất / dị thể)

Kho dữ liệu phải xử lý cả dữ liệu đồng nhất (cùng cấu trúc) và dị thể (khác cấu trúc, nguồn gốc).

Ví dụ: dữ liệu từ hệ thống CRM, ERP, file Excel, Web logs.

4. Auditing (Kiểm toán dữ liệu)

Theo dõi nguồn gốc, quá trình biến đổi của dữ liệu từ hệ thống tác nghiệp đến kho.

Đảm bảo tính minh bạch, tin cậy và hỗ trợ tuân thủ (compliance).

1. Các khái niệm cơ bản về kho dữ liệu

Vai trò: Kho dữ liệu được hiểu là một tập hợp dữ liệu có định hướng theo chủ đề, được tích hợp từ nhiều nguồn khác nhau, có tính ổn định (không thay đổi sau khi được nạp), đồng thời phản ánh sự biến thiên/ thay đổi theo thời gian; toàn bộ được tổ chức nhằm phục vụ hiệu quả cho quá trình ra quyết định quản trị.

1. Các khái niệm cơ bản về kho dữ liệu

Ý nghĩa thực tiễn

Nhờ có kho dữ liệu, doanh nghiệp có:

Một phiên bản duy nhất của sự thật (single version of the truth).

Nền tảng cho BI/Analytics.

Khả năng phân tích dài hạn, đa chiều thay vì chỉ nhìn dữ liệu hiện tại.

Đây là bước chuyển từ "**Data → Information → Knowledge → Decision**" mà Inmon nhấn mạnh.

2. Kiến trúc kho dữ liệu

Dẫn nhập

Trong hệ thống Kho dữ liệu (Data Warehouse), việc tổ chức dữ liệu có vai trò then chốt để hỗ trợ truy vấn và phân tích đa chiều. Khác với cơ sở dữ liệu giao dịch (OLTP) vốn tối ưu cho cập nhật, kho dữ liệu tối ưu cho phân tích (OLAP), vì vậy dữ liệu thường được mô hình hóa theo các kiến trúc đa chiều.

Ba kiến trúc phổ biến nhất gồm:

- Star Schema (Sơ đồ Ngôi sao)
- Snowflake Schema (Sơ đồ Bông tuyết)
- Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Các kiến trúc này đều dựa trên hai loại bảng cốt lõi:

- Fact Table (Bảng sự kiện): lưu trữ các giá trị định lượng.
- Dimension Table (Bảng chiều): cung cấp ngữ cảnh phân tích cho fact.

2. Kiến trúc kho dữ liệu

Star Schema (Sơ đồ Ngôi sao)

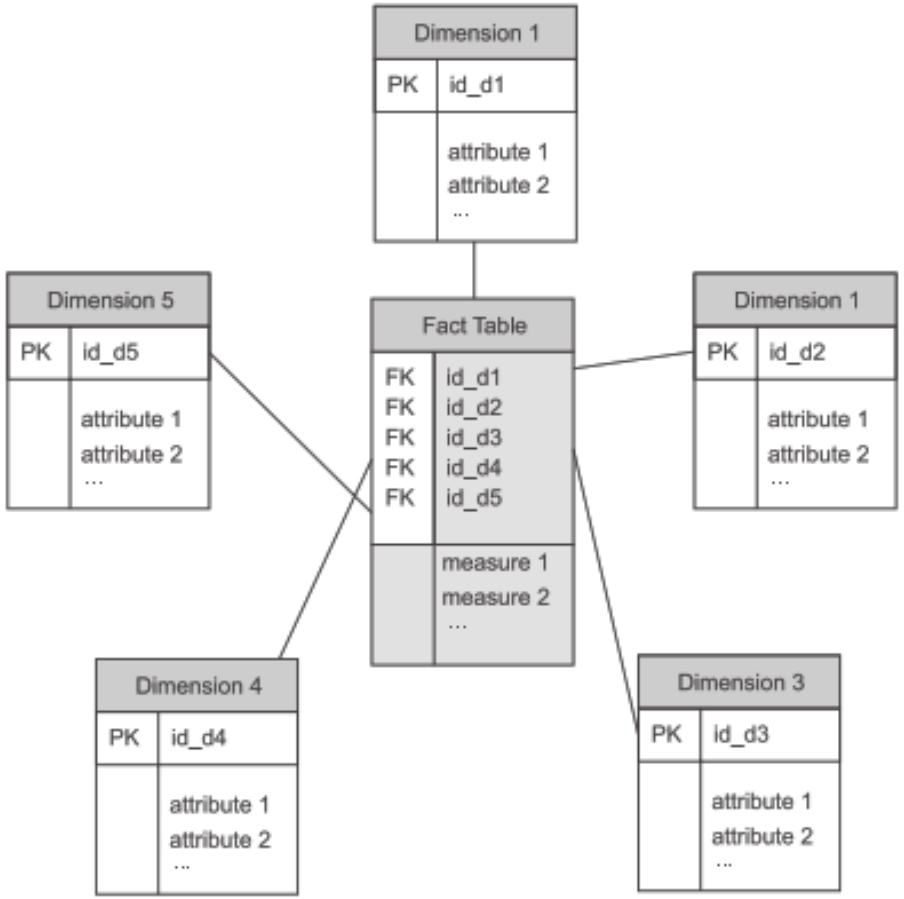

Figure 13.5 Graphical representation of Star schema

2. Kiến trúc kho dữ liệu

Star Schema (Sơ đồ Ngôi sao)

Star Schema (Sơ đồ Ngôi sao) là một mô hình dữ liệu đa chiều phổ biến trong thiết kế kho dữ liệu. Trong mô hình này:

Bảng Fact (Fact Table) nằm ở trung tâm, chứa các số liệu định lượng (measures) như doanh thu, số lượng bán, chi phí,…

Các bảng Dimension (Dimension Tables) bao quanh bảng Fact, lưu trữ dữ liệu mô tả (descriptive attributes) như Thời gian, Khách hàng, Sản phẩm, Khu vực,…

Quan hệ giữa Fact Table và Dimension Tables được thiết kế theo dạng một-nhiều (1–n), và khi vẽ sơ đồ, các bảng dimension toả ra xung quanh bảng fact tạo thành hình ngôi sao.

2. Kiến trúc kho dữ liệu

Star Schema (Sơ đồ Ngôi sao)

Đặc điểm của Star Schema

Đơn giản, dễ hiểu: cấu trúc trực quan, dễ dàng cho người phân tích dữ liệu.

Hiệu suất truy vấn cao: vì các bảng dimension thường được phi chuẩn hoá (denormalized), nên việc join với Fact Table nhanh hơn.

Dư thừa dữ liệu: do phi chuẩn hoá, một số dữ liệu trong bảng dimension có thể bị lặp lại.

Ứng dụng thực tiễn: rất phù hợp cho các hệ thống OLAP và báo cáo doanh nghiệp, ví dụ phân tích doanh thu theo thời gian, theo sản phẩm, hoặc theo khu vực.

2. Kiến trúc kho dữ liệu

Star Schema (Sơ đồ Ngôi sao)

Ví dụ

Giả sử ta phân tích doanh thu bán hàng:

Fact_Sales: chứa các cột Date_ID, Product_ID, Customer_ID, Store_ID, và measure Sales_Amount.

Dim_Date: chứa chi tiết ngày, tháng, quý, năm.

Dim_Product: chứa tên sản phẩm, loại sản phẩm, nhà sản xuất.

Dim_Customer: chứa thông tin khách hàng (tuổi, giới tính, nghề nghiệp).

Dim_Store: chứa thông tin cửa hàng (vị trí địa lý, quy mô).

Khi join các bảng dimension với Fact_Sales, ta có thể trả lời những câu hỏi như: “Doanh thu quý 2 năm 2024 theo từng khu vực là bao nhiêu?” hoặc “Top 10 sản phẩm bán chạy nhất trong tháng 7/2024 là gì?”.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

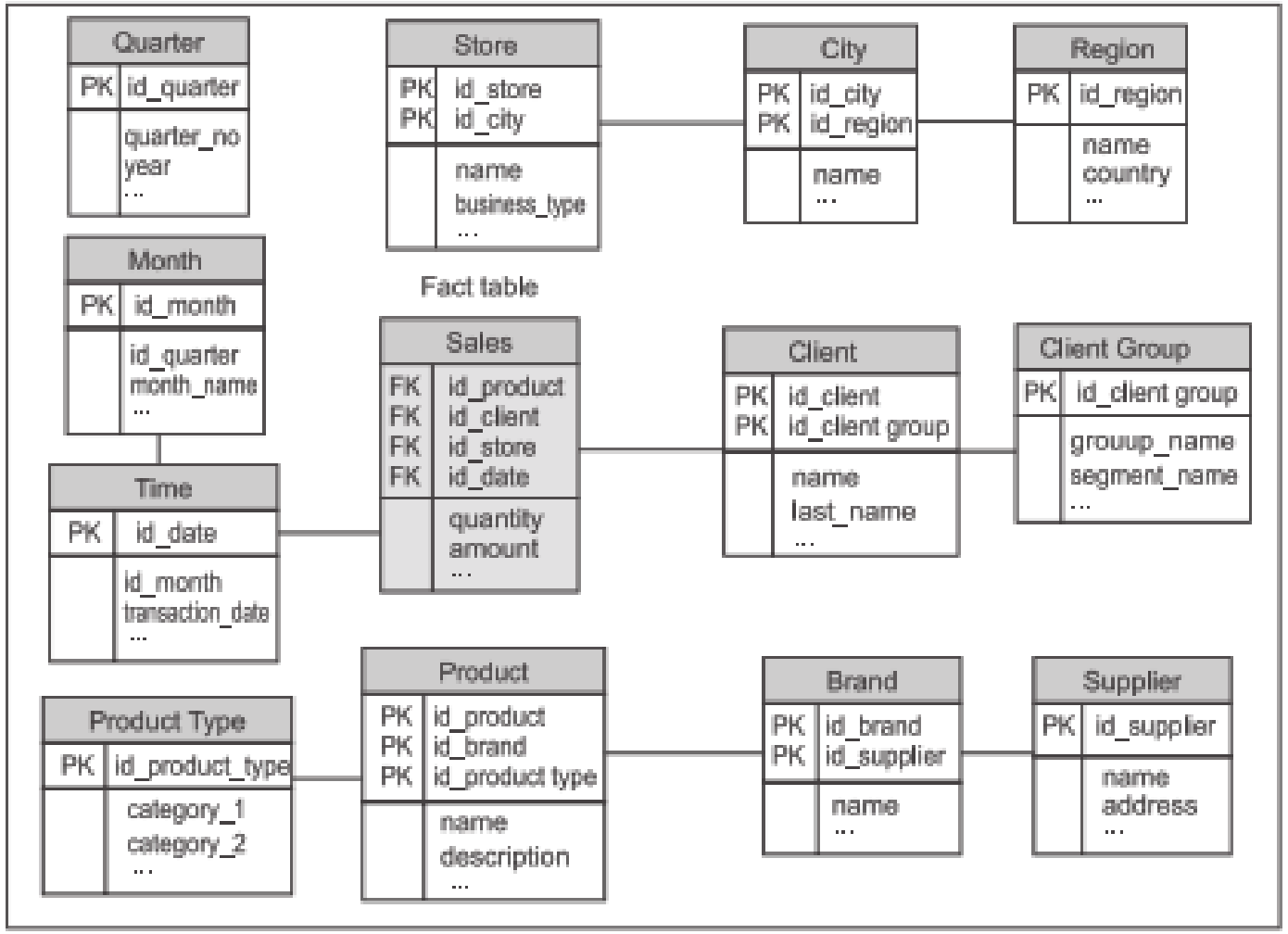

Figure 13.8 Snowflake schema

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

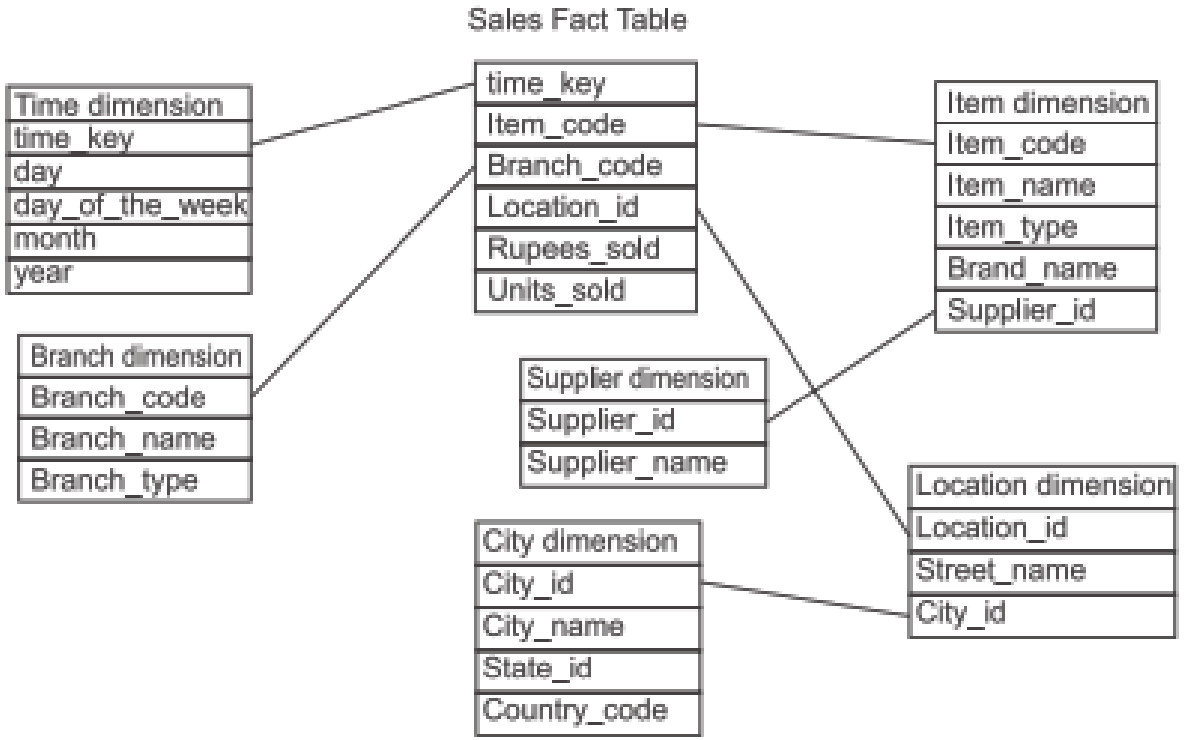

Figure 13.7 Snowflake schema for analysis of sales

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Snowflake Schema là một mô hình dữ liệu đa chiều trong kho dữ liệu, trong đó các bảng chiều (Dimension Tables) được chuẩn hoá thành nhiều bảng con liên kết với nhau. Cấu trúc này tạo thành hình dạng giống bông tuyết, do các nhánh phân cấp của chiều mở rộng ra từ bảng Fact trung tâm.

Ví dụ: Thay vì có một bảng Dim_Product chứa tất cả thông tin về sản phẩm, trong Snowflake Schema, thông tin sẽ được tách ra thành Dim_Product, Dim_Brand, Dim_Category và liên kết với nhau theo quan hệ cha–con.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Cấu trúc:

Fact Table: Ở trung tâm, chứa dữ liệu định lượng (measures) như doanh thu, số lượng bán.

Dimension Tables: Được chuẩn hóa thành nhiều cấp để giảm dư thừa dữ liệu.

Các Dimension liên kết trực tiếp hoặc gián tiếp với Fact Table thông qua các khóa ngoại.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Đặc điểm chính:

Dimension tables được chuẩn hóa (Normalized).

Giảm dư thừa dữ liệu nhờ tách thành nhiều bảng nhỏ.

Tăng tính toàn vẹn dữ liệu vì cập nhật ở một bảng sẽ lan tỏa đến toàn bộ hệ thống.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Ưu điểm

Tiết kiệm dung lượng lưu trữ nhờ loại bỏ dữ liệu lặp.

Dữ liệu dễ bảo trì và cập nhật hơn do được chuẩn hóa.

Hữu ích cho các hệ thống có kích thước dữ liệu chiều rất lớn.

5. Nhược điểm:

Cấu trúc phức tạp hơn Star Schema, khó hiểu đối với người dùng cuối.

Hiệu năng truy vấn chậm hơn, vì phải join nhiều bảng hơn.

Không thân thiện với người dùng phân tích nghiệp vụ, thường đòi hỏi DBA/IT hỗ trợ.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Ví dụ minh họa:

Giả sử ta có Fact_Sales (Doanh số bán hàng) và Dimension Product:

Trong Star Schema, Dim_Product chứa đầy đủ thông tin: Tên sản phẩm, Loại, Nhãn hiệu, Nhà cung cấp.

Trong Snowflake Schema, Dim_Product chỉ giữ ID + tên sản phẩm, các thuộc tính khác được tách ra thành Dim_Category, Dim_Brand, Dim_Supplier, mỗi bảng lại nối với Dim_Product.

2. Kiến trúc kho dữ liệu

Snowflake Schema (Sơ đồ Bông tuyết)

Ví dụ minh họa:

Snowflake Schema

là lựa chọn phù hợp

khi dữ liệu chiều có

tính phân cấp phức tạp

và cần tối ưu dung lượng lưu trữ.

Tuy nhiên, trong thực tế triển khai BI,

nhiều tổ chức vẫn ưu tiên Star Schema

vì dễ sử dụng và hiệu năng tốt hơn cho

truy vấn OLAP.

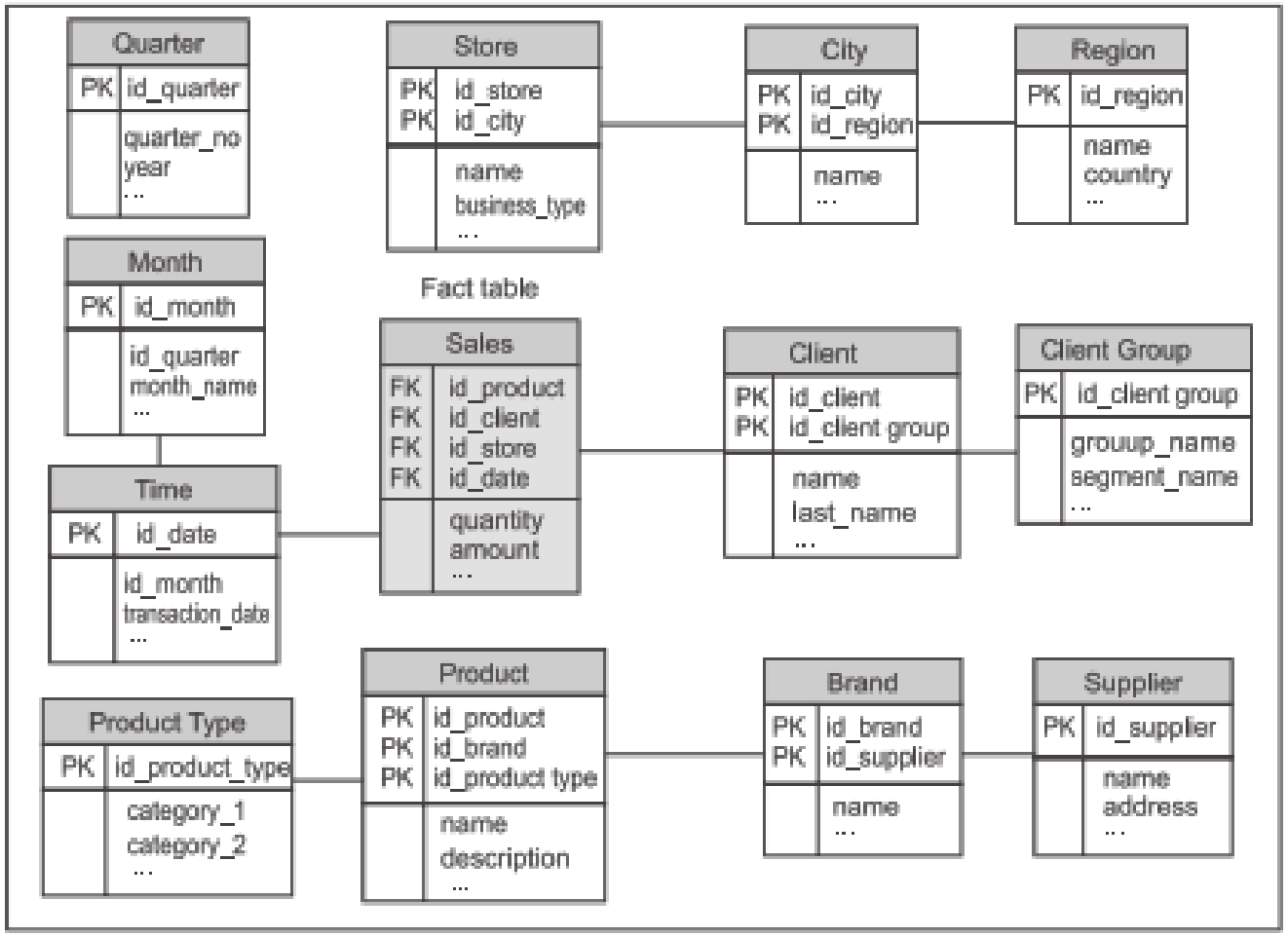

Figure 13.8 Snowflake schema

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

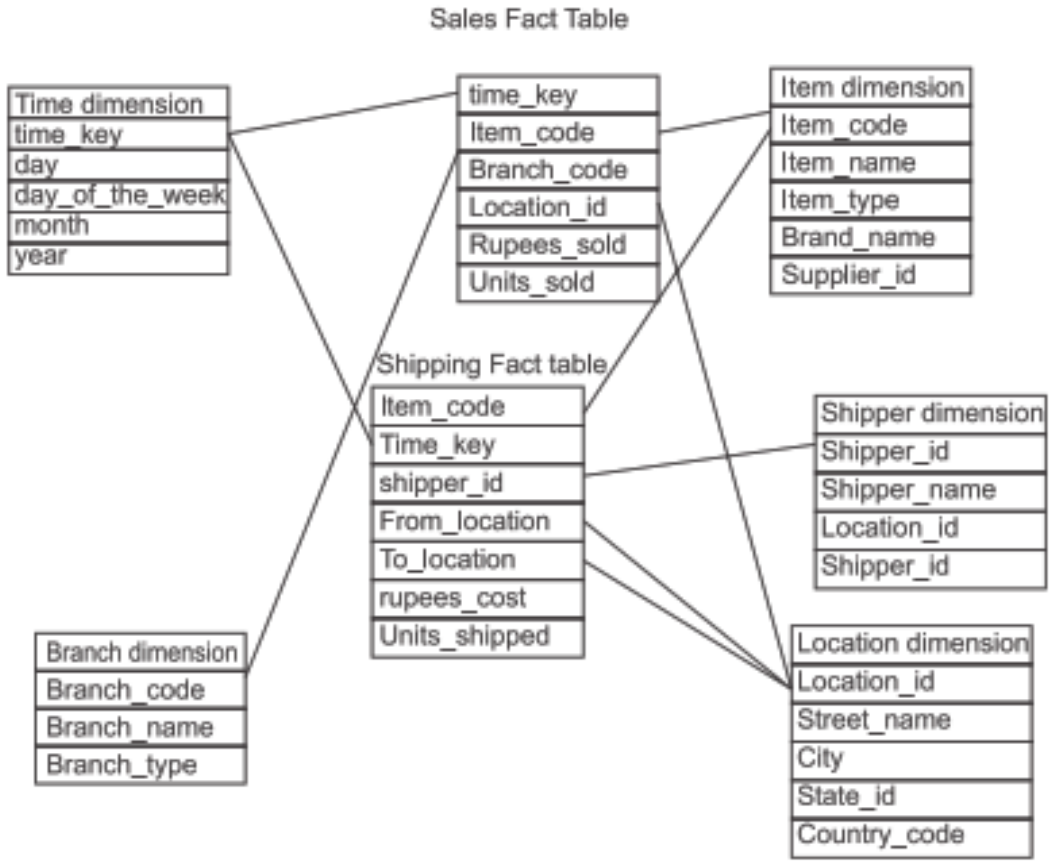

Figure 13.9 Fact constellation schema for analysis of sales

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Khái niệm:

Galaxy Schema – còn gọi là Fact Constellation Schema – là mô hình dữ liệu đa chiều trong đó nhiều bảng Fact cùng chia sẻ các bảng Dimension chung. Cấu trúc này giống một chòm sao hay ngân hà, vì có nhiều bảng Fact kết nối đến các bảng Dimension, tạo thành một mạng lưới phức tạp.

Ví dụ: Một hệ thống phân tích có thể có Fact_Sales (bán hàng) và Fact_Shipping (vận chuyển), cả hai cùng chia sẻ bảng Dimension như Dim_Time, Dim_Product, Dim_Customer.

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Cấu trúc:

Nhiều Fact Table: Mỗi bảng Fact phản ánh một quá trình kinh doanh (bán hàng, vận chuyển, tồn kho…).

Dimension Tables chia sẻ: Các bảng Dimension có thể dùng chung cho nhiều Fact Table (ví dụ: Dim_Time, Dim_Product).

Có thể kết hợp các đặc điểm của Star Schema hoặc Snowflake Schema trong từng phần.

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Đặc điểm chính:

Tổ chức dữ liệu linh hoạt, hỗ trợ phân tích nhiều khía cạnh khác nhau của doanh nghiệp.

Các Fact Table thường liên quan đến nhau qua Dimension chung.

Cho phép phân tích liên miền (cross-functional analysis).

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Ưu điểm:

Hỗ trợ nhiều quy trình kinh doanh cùng lúc.

Giảm trùng lặp Dimension (vì các Fact Table dùng chung).

Mạnh mẽ trong phân tích dữ liệu đa chiều và toàn diện.

5. Nhược điểm:

Cấu trúc rất phức tạp, khó thiết kế và bảo trì.

Người dùng cuối có thể khó hiểu khi viết truy vấn trực tiếp.

Yêu cầu công cụ OLAP/BI tối ưu để xử lý khối lượng dữ liệu lớn.

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

Ví dụ minh họa:

Trong hệ thống siêu thị:

- Fact_Sales: lưu số lượng bán, doanh thu.
- Fact_Inventory: lưu số lượng tồn kho, giá trị tồn kho.
- Fact_Shipping: lưu chi phí và thời gian vận chuyển.

Tất cả cùng chia sẻ các Dimension như:

- Dim_Time (ngày, tháng, năm)
- Dim_Product (sản phẩm, loại, nhãn hiệu)
- Dim_Store (siêu thị, khu vực, quốc gia)
- Dim_Customer (khách hàng, nhóm khách hàng)

2. Kiến trúc kho dữ liệu

Galaxy Schema (Sơ đồ Ngân hà / Fact Constellation)

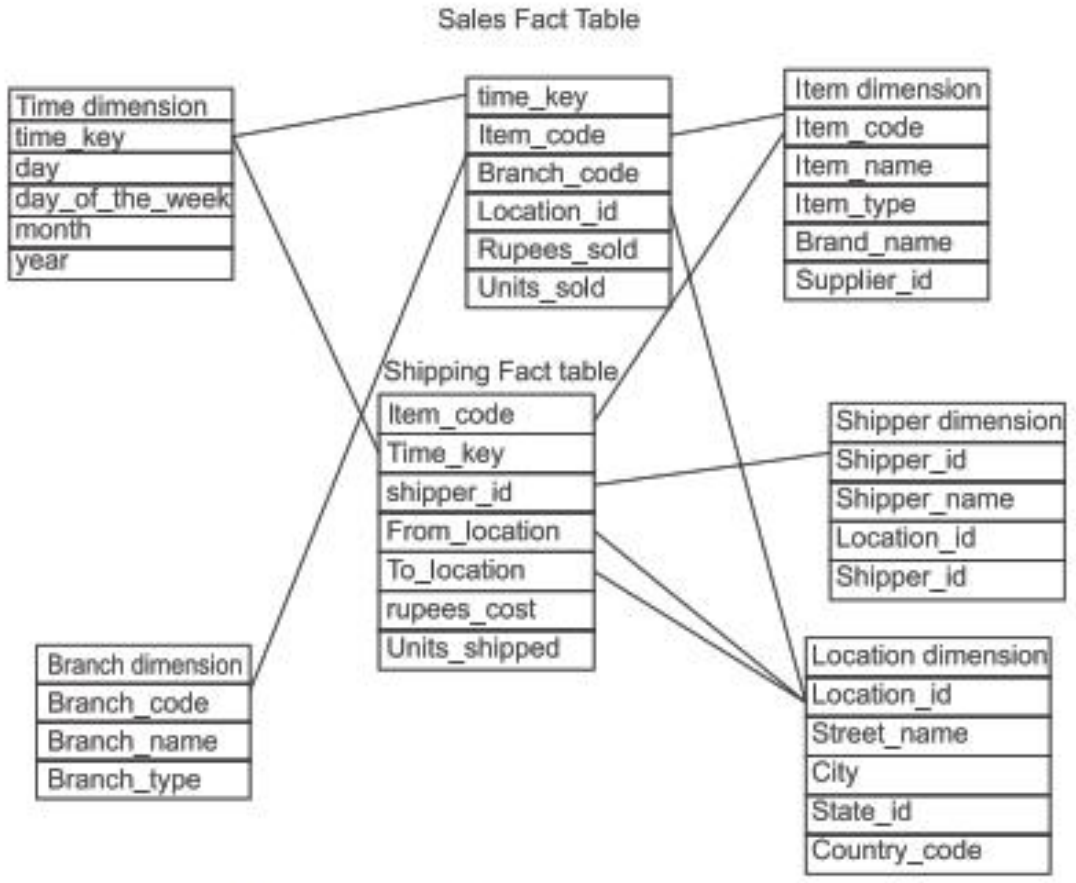

Figure 13.9 Fact constellation schema for analysis of sales

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

1. Khái niệm

Fact Table là bảng trung tâm trong mô hình kho dữ liệu đa chiều (Star, Snowflake, Galaxy).

Chứa các dữ liệu định lượng (measures/facts) – thường là số liệu có thể đo lường, tổng hợp, tính toán (như doanh thu, số lượng, chi phí…).

Các dữ liệu trong Fact Table được tham chiếu bởi khóa ngoại (Foreign Key) liên kết tới Dimension Table (Bảng chiều).

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

2. Đặc điểm chính

Khóa chính (Primary Key) của Fact Table thường là khóa tổng hợp (Composite Key), bao gồm các khóa ngoại từ các bảng Dimension.

Có số lượng bản ghi rất lớn (thường nhiều hơn rất nhiều so với Dimension Table).

Dữ liệu trong Fact Table thường mang tính lịch sử, được ghi nhận theo thời gian (time-variant).

Các số liệu (facts) thường là additive (cộng dồn được) hoặc semi-additive (cộng dồn theo một số chiều nhất định, ví dụ theo thời gian thì không).

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

2. Đặc điểm chính

Khóa chính (Primary Key) của Fact Table thường là khóa tổng hợp (Composite Key), bao gồm các khóa ngoại từ các bảng Dimension.

Có số lượng bản ghi rất lớn (thường nhiều hơn rất nhiều so với Dimension Table).

Dữ liệu trong Fact Table thường mang tính lịch sử, được ghi nhận theo thời gian (time-variant).

Các số liệu (facts) thường là additive (cộng dồn được) hoặc semi-additive (cộng dồn theo một số chiều nhất định, ví dụ theo thời gian thì không).

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

Product_ID	Customer_ID	Time_ID	Store_ID	Quantity_Sold	Sales_Revenue
101	C01	T202401	S1	3	1,500,000
102	C02	T202401	S2	1	500,000

3. Ví dụ minh họa

Fact_Sales (Bảng sự kiện Bán hàng):

Thuộc tính:

Product_ID (FK → Dimension_Product)

Customer_ID (FK → Dimension_Customer)

Time_ID (FK → Dimension_Time)

Store_ID (FK → Dimension_Store)

Quantity_Sold (Số lượng bán)

Sales_Revenue (Doanh thu)

Discount (Chiết khấu).

Fact Table = nơi lưu số liệu định lượng

Dimension Table = nơi lưu ngữ cảnh mô tả cho số liệu đó

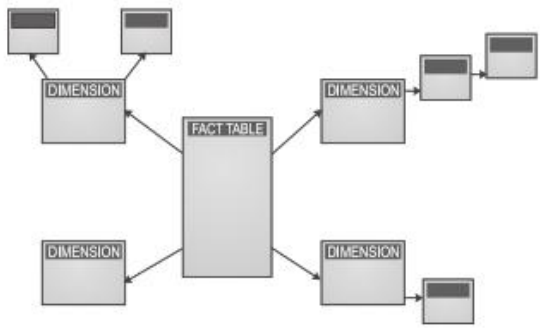

Figure 13.3 Representation of fact and dimension tables

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

3. Ví dụ minh họa

Location Dimension
Locaton_id
Street_name
City
State_id
Country_code

(a)

Item Dimension
Item_code
Item_name
Item_type
Brand_name
Supplier_id

(b)

Figure 13.1 (a) location dimension, (b) item dimension

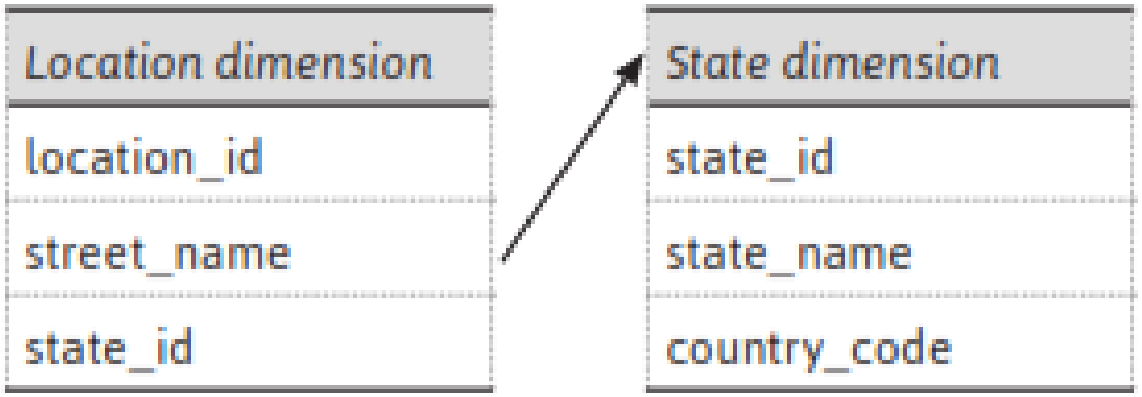

Figure 13.2 Normalized view

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

4. Phân loại Fact Table

Transaction Fact Table: lưu dữ liệu chi tiết từng giao dịch (ví dụ: mỗi hóa đơn).

Snapshot Fact Table: lưu trạng thái tại một thời điểm (ví dụ: số dư tài khoản cuối tháng).

Accumulating Fact Table: lưu dữ liệu gắn với một tiến trình có nhiều giai đoạn (ví dụ: quy trình xử lý đơn hàng).

2. Kiến trúc kho dữ liệu

Fact Table (Bảng sự kiện)

5. Vai trò trong phân tích

Là trung tâm của mô hình dữ liệu, chứa số liệu để phân tích.

Cho phép tính toán các chỉ số kinh doanh: doanh thu, lợi nhuận, số lượng bán,…

Kết hợp với Dimension Table để trả lời các câu hỏi:

“Doanh thu theo tháng/quý/năm như thế nào?”

“Sản phẩm nào bán chạy nhất tại từng khu vực?”

“Khách hàng nhóm nào mang lại doanh thu cao nhất?”).

2. Kiến trúc kho dữ liệu

Dimension Table (Bảng chiều)

1. Khái niệm

Dimension Table là bảng mô tả ngữ cảnh (context) cho các dữ liệu định lượng trong Fact Table.

Chứa các thuộc tính (attributes) dùng để phân tích, lọc, nhóm, sắp xếp hoặc trình bày dữ liệu.

Thường có ít bản ghi hơn Fact Table, nhưng mỗi bản ghi lại chứa nhiều thông tin mô tả.

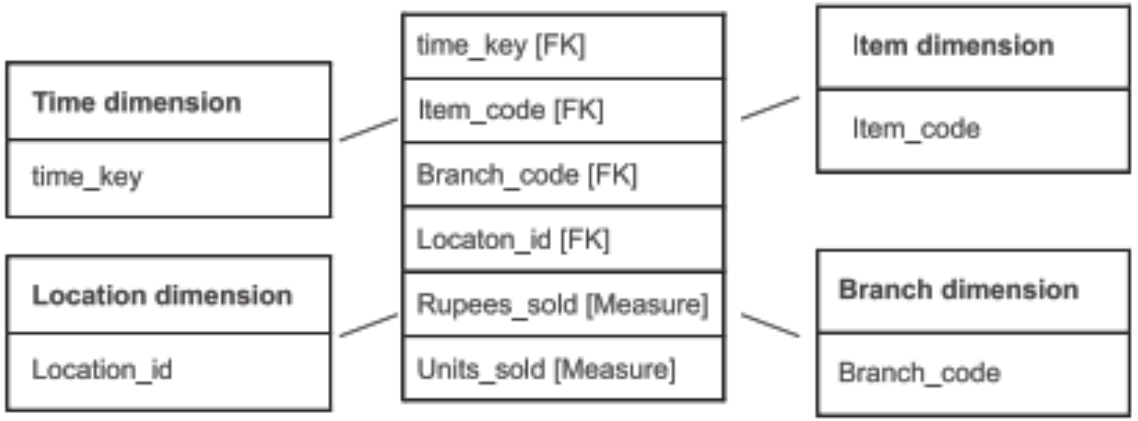

Figure 13.4 The sales fact table

2. Kiến trúc kho dữ liệu

Dimension Table (Bảng chiều)

2. Đặc điểm chính

Khóa chính (Primary Key) của Dimension Table sẽ liên kết với khóa ngoại (Foreign Key) trong Fact Table.

Chứa dữ liệu mô tả, định tính chứ không phải dữ liệu số để tính toán.

Được thiết kế denormalized (ít chuẩn hóa) để thuận tiện cho truy vấn OLAP.

Có thể xây dựng cấu trúc phân cấp (hierarchy) phục vụ drill-down/roll-up trong phân tích đa chiều.

2. Kiến trúc kho dữ liệu

Dimension Table (Bảng chiều)

3. Ví dụ minh họa

Dimension_Product (Sản phẩm)

Thuộc tính: Product_ID, Product_Name, Category, Supplier, Brand, Price_Range.

Dimension_Time (Thời gian)

Thuộc tính: Time_ID, Day, Month, Quarter, Year, Holiday_Flag.

Dimension_Customer (Khách hàng)

Thuộc tính: Customer_ID, Name, Age_Group, Gender, Location, Income_Level.

2. Kiến trúc kho dữ liệu

Dimension Table (Bảng chiều)

4. Vai trò trong phân tích

Cung cấp ngữ cảnh cho số liệu trong Fact Table.

Giúp phân tích dữ liệu theo nhiều góc nhìn: theo thời gian, sản phẩm, khách hàng, địa lý,…

Cho phép thực hiện các thao tác drill-down (chi tiết hơn) hoặc roll-up (tổng hợp) trong OLAP.

2. Kiến trúc kho dữ liệu

Dimension Table (Bảng chiều)

4. Vai trò trong phân tích

Cung cấp ngữ cảnh cho số liệu trong Fact Table.

Giúp phân tích dữ liệu theo nhiều góc nhìn: theo thời gian, sản phẩm, khách hàng, địa lý,…

Cho phép thực hiện các thao tác drill-down (chi tiết hơn) hoặc roll-up (tổng hợp) trong OLAP.

5. Ví dụ kết hợp Fact & Dimension

Fact_Sales có các cột: Product_ID, Customer_ID, Time_ID, Store_ID, Revenue.

Muốn biết “Doanh thu quý 1 năm 2024 theo từng danh mục sản phẩm”, ta JOIN Fact_Sales với Dimension_Time và Dimension_Product.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Dẫn nhập

Trong bất kỳ hệ thống thông tin nào, dữ liệu không chỉ cần được lưu trữ và xử lý mà còn phải đảm bảo chất lượng. Chất lượng dữ liệu được đo lường thông qua nhiều chiều khác nhau, gọi là Data Quality Dimensions. Đây là tập hợp tiêu chuẩn để đánh giá dữ liệu có đủ tin cậy, chính xác và hữu ích cho quá trình ra quyết định hay không.

Điểm cần lưu ý: Các đặc tính chất lượng dữ liệu khác với các đặc tính của kho dữ liệu (Subject-Oriented, Integrated, Nonvolatile, Time-Variant) mà Inmon đề cập.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Dẫn nhập,

Theo Inmon, Kho dữ liệu (Data Warehouse) là một tập hợp dữ liệu định hướng theo chủ đề (subject-oriented), tích hợp (integrated), không biến đổi (nonvolatile), và có tính biến thiên theo thời gian (time-variant) nhằm hỗ trợ quá trình ra quyết định quản trị .

Trong bất kỳ hệ thống thông tin nào, dữ liệu không chỉ cần được lưu trữ và xử lý mà còn phải đảm bảo chất lượng. Chất lượng dữ liệu được đo lường thông qua nhiều chiều khác nhau, gọi là Data Quality Dimensions. Đây là tập hợp tiêu chuẩn để đánh giá dữ liệu có đủ tin cậy, chính xác và hữu ích cho quá trình ra quyết định hay không.

Điểm cần lưu ý: Các đặc tính chất lượng dữ liệu khác với các đặc tính của kho dữ liệu (Subject-Oriented, Integrated, Nonvolatile, Time-Variant) mà Inmon đề cập.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Dẫn nhập,

Theo Inmon, Kho dữ liệu (Data Warehouse) là một tập hợp dữ liệu định hướng theo chủ đề (subject-oriented), tích hợp (integrated), không biến đổi (nonvolatile), và có tính biến thiên theo thời gian (time-variant) nhằm hỗ trợ quá trình ra quyết định quản trị .

*(BỐN đặc tính **CHÍNH** của dữ liệu trong kho dữ liệu)*

Trong bất kỳ hệ thống thông tin nào, dữ liệu không chỉ cần được lưu trữ và xử lý mà còn phải đảm bảo chất lượng. Chất lượng dữ liệu được đo lường thông qua nhiều chiều khác nhau, gọi là Data Quality Dimensions. Đây là tập hợp tiêu chuẩn để đánh giá dữ liệu có đủ tin cậy, chính xác và hữu ích cho quá trình ra quyết định hay không.

Điểm cần lưu ý: Các đặc tính chất lượng dữ liệu khác với các đặc tính của kho dữ liệu (Subject-Oriented, Integrated, Nonvolatile, Time-Variant) mà Inmon đề cập.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Chất lượng các chiều dữ liệu Data Quality Dimensions , (bên cạnh *BỐN đặc tính* ***CHÍNH*** *của dữ liệu trong kho dữ liệu*)

1. Độ chính xác (Accuracy)

Dữ liệu phản ánh đúng thực tế khách quan.

Ví dụ: ngày sinh của khách hàng phải đúng như giấy tờ, không được nhập sai lệch.

2. Tính đầy đủ (Completeness)

Dữ liệu không bị thiếu trường hoặc giá trị quan trọng.

Ví dụ: bản ghi khách hàng phải có đầy đủ tên, số điện thoại, địa chỉ.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Chất lượng các chiều dữ liệu Data Quality Dimensions , (bên cạnh *BỐN đặc tính **CHÍNH** của dữ liệu trong kho dữ liệu*)

3. Tính nhất quán (Consistency)

Dữ liệu phải thống nhất trong toàn bộ hệ thống.

Ví dụ: cùng một khách hàng thì mã khách hàng và thông tin cá nhân không được khác nhau giữa hai bảng khác nhau.

4. Tính hợp lệ (Validity)

Dữ liệu tuân theo chuẩn mực, quy tắc hoặc miền giá trị.

Ví dụ: số điện thoại phải có đúng 10 chữ số, ngày tháng phải hợp lệ theo lịch.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Chất lượng các chiều dữ liệu Data Quality Dimensions , (bên cạnh *BỐN đặc tính **CHÍNH** của dữ liệu trong kho dữ liệu*)

5. Tính kịp thời (Timeliness)

   Dữ liệu phải được cập nhật đúng lúc để phục vụ cho ra quyết định.

   Ví dụ: dữ liệu bán hàng cần cập nhật hàng ngày để quản lý kho chính xác.

6. Tính duy nhất (Uniqueness)

   Không có dữ liệu trùng lặp gây dư thừa hoặc sai lệch.

   Ví dụ: một khách hàng không nên có hai mã ID khác nhau.

7. Tính dễ truy cập và khả dụng (Accessibility & Availability)

   Người dùng có quyền và công cụ để truy cập dữ liệu khi cần.

   Ví dụ: dữ liệu lưu trữ trong cơ sở dữ liệu tập trung với cơ chế phân quyền rõ ràng.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Làm rõ sự khác biệt,

Đặc tính dữ liệu tổng quát (Data Quality Dimensions) thường dùng để đánh giá chất lượng dữ liệu. Data Quality Dimensions *là thước đo chất lượng dữ liệu (bất kể trong hệ thống nào).*

Đặc trưng dữ liệu của kho dữ liệu (theo định nghĩa Inmon: Subject-Oriented (Định hướng theo chủ đề), Integrated (Tích hợp), Nonvolatile (Không biến đổi), Time-Variant (Biến thiên theo thời gian), Inmon's Data Warehouse Characteristics, là nguyên lý thiết kế kiến trúc riêng cho kho dữ liệu.

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Bảng dưới đây so sánh sự khác biệt giữa các Đặc tính dữ liệu tổng quát (Data Quality Dimensions) và các Đặc trưng dữ liệu của Kho dữ liệu (theo định nghĩa Inmon).

Tiêu chí	Đặc tính dữ liệu tổng quát (Data Quality Dimensions)	Đặc trưng dữ liệu của Kho dữ liệu (Inmon)
Mục tiêu	Đảm bảo dữ liệu chính xác, đầy đủ, nhất quán và hữu ích cho nghiệp vụ	Định hình cách tổ chức và lưu trữ dữ liệu trong kho dữ liệu để phục vụ phân tích
Độ chính xác (Accuracy)	Dữ liệu phản ánh đúng thực tế	Không đề cập trực tiếp, nhưng dữ liệu trong kho phải chuẩn hoá và tích hợp để đảm bảo độ tin cậy
Tính đầy đủ (Completeness)	Không thiếu sót dữ liệu cần thiết	Kho dữ liệu lưu trữ dữ liệu lịch sử (time-variant), đảm bảo bức tranh toàn diện
Tính nhất quán (Consistency)	Dữ liệu không mâu thuẫn giữa các nguồn, các hệ thống	Đặc trưng "Integrated" đảm bảo dữ liệu được hợp nhất, loại bỏ mâu thuẫn

CSE 3. Các đặc tính về dữ liệu của kho dữ liệu

Bảng dưới đây so sánh sự khác biệt giữa các Đặc tính dữ liệu tổng quát (Data Quality Dimensions) và các Đặc trưng dữ liệu của Kho dữ liệu (theo định nghĩa Inmon).

Tiêu chí	Đặc tính dữ liệu tổng quát (Data Quality Dimensions)	Đặc trưng dữ liệu của Kho dữ liệu (Inmon)
Tính hợp lệ (Validity)	Dữ liệu tuân thủ quy tắc nghiệp vụ, định dạng, miền giá trị	Kho dữ liệu thiết kế theo chuẩn hoá schema, dữ liệu được kiểm tra trước khi nạp
Tính kịp thời (Timeliness)	Dữ liệu có sẵn đúng thời điểm cần	Đặc trưng "Time-Variant": dữ liệu được lưu kèm dấu thời gian, hỗ trợ phân tích xu hướng
Tính duy nhất (Uniqueness)	Không trùng lặp, không lặp bản ghi	Kho dữ liệu dùng chuẩn ETL để loại bỏ dữ liệu dư thừa
Khả năng truy cập (Accessibility)	Người dùng có thể truy cập và khai thác dễ dàng	Đặc trưng "Nonvolatile": dữ liệu ổn định, chỉ đọc → dễ dàng cho truy vấn phân tích

4. Vấn đề về độ mịn dữ liệu

1. Khái niệm độ mịn dữ liệu

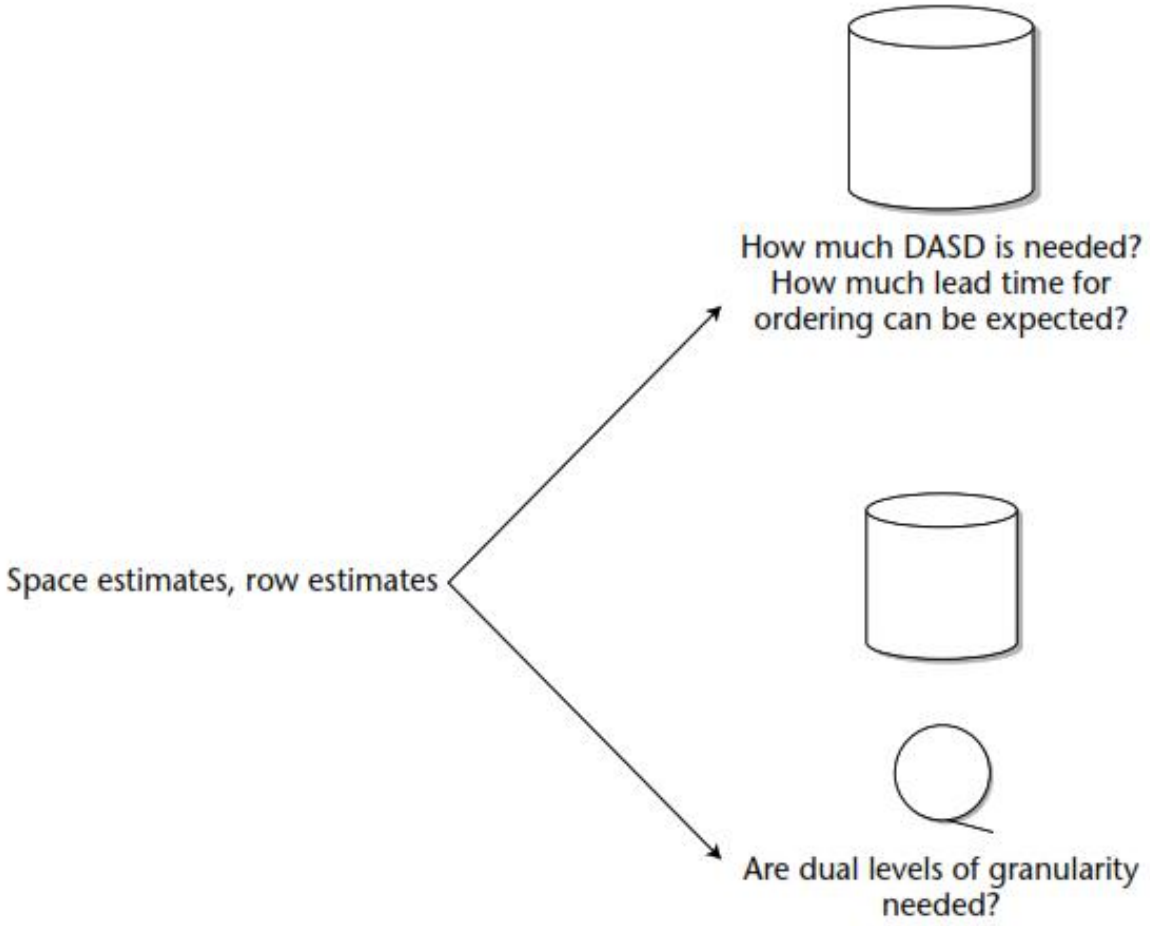

Figure 4-2 Using the output of the space estimates.

4. Vấn đề về độ mịn dữ liệu

1. Khái niệm độ mịn dữ liệu

Độ mịn dữ liệu (Data Granularity) chỉ mức độ chi tiết hay tổng hợp của dữ liệu được lưu trữ trong kho dữ liệu.

Dữ liệu có thể được lưu ở mức chi tiết (fine-grained), ví dụ: từng giao dịch bán hàng, hoặc ở mức tổng hợp (coarse-grained), ví dụ: doanh thu theo tháng.

Việc quyết định chọn mức độ chi tiết nào ảnh hưởng trực tiếp đến khả năng phân tích, dung lượng lưu trữ, và hiệu suất truy vấn .

4. Vấn đề về độ mịn dữ liệu

2. Lợi ích và thách thức của độ mịn dữ liệu

Lợi ích:

Độ mịn chi tiết cao: cho phép phân tích sâu, trả lời nhiều câu hỏi khác nhau, đặc biệt khi cần drill-down (đi sâu vào dữ liệu).

Độ mịn tổng hợp: tiết kiệm không gian lưu trữ, cải thiện tốc độ truy vấn, phù hợp với báo cáo quản trị cấp cao.

Thách thức:

Mức chi tiết cao → cần nhiều dung lượng, phức tạp trong quản lý và xử lý.

Mức tổng hợp → hạn chế khả năng phân tích chi tiết, có thể mất thông tin quan trọng.

4. Vấn đề về độ mịn dữ liệu

3. Ví dụ minh họa

Mức chi tiết (fine-grained): lưu dữ liệu từng giao dịch thẻ ATM (ngày, giờ, số tiền, địa điểm).

Mức tổng hợp (coarse-grained): lưu dữ liệu tổng số tiền rút ATM theo ngày hoặc tháng.

Một ngân hàng có thể chọn cách lưu cả hai mức: dữ liệu giao dịch chi tiết cho phân tích nghiệp vụ, và dữ liệu tổng hợp cho báo cáo chiến lược.

4. Vấn đề về độ mịn dữ liệu

3. Ví dụ minh họa

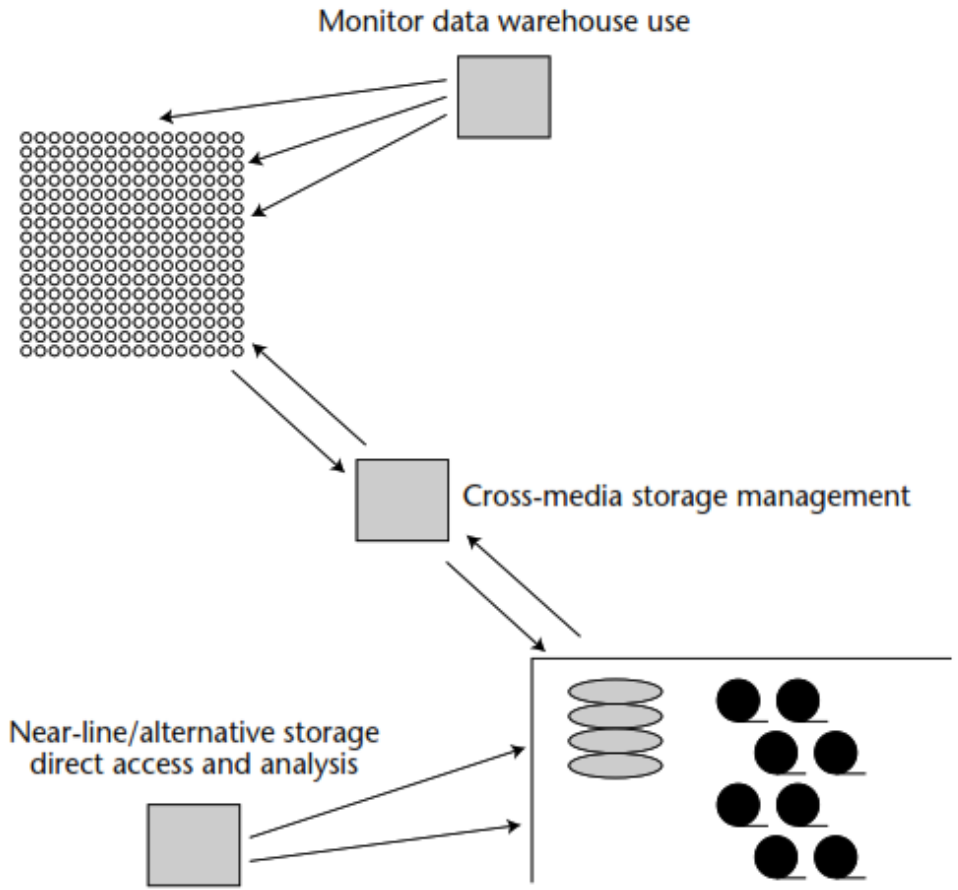

Figure 4-4 The support software needed to make storage overflow possible.

4. Vấn đề về độ mịn dữ liệu

4. Mô hình hai cấp độ (Dual Levels of Granularity)

Inmon nhấn mạnh rằng trong nhiều trường hợp, kho dữ liệu cần kết hợp hai cấp độ:

Atomic data (chi tiết nguyên tử): dữ liệu gốc, chi tiết nhất.

Summary data (dữ liệu tóm tắt): được tính toán/tổng hợp để hỗ trợ báo cáo nhanh .

Cách tiếp cận này giúp cân bằng giữa khả năng phân tích sâu và hiệu quả sử dụng tài nguyên.

4. Vấn đề về độ mịn dữ liệu

5. Ý nghĩa trong thiết kế kho dữ liệu

Quyết định về độ mịn là một trong những quyết định quan trọng nhất khi thiết kế kho dữ liệu.

Nếu thiết kế sai: hoặc hệ thống quá nặng nề do lưu quá nhiều chi tiết, hoặc thiếu tính linh hoạt khi cần phân tích chuyên sâu.

Giải pháp thực tế thường là đa cấp độ độ mịn, kết hợp chi tiết và tổng hợp.

Ý nghĩa: Vấn đề độ mịn dữ liệu là cốt lõi trong kiến trúc kho dữ liệu. Nó quyết định mức độ hữu ích của kho dữ liệu trong việc hỗ trợ ra quyết định, đồng thời ảnh hưởng đến chi phí lưu trữ và hiệu năng hệ thống. Người thiết kế cần cân nhắc kỹ lưỡng giữa nhu cầu phân tích chi tiết và khả năng tối ưu tài nguyên.

5. Vấn đề về chuyển đổi dữ liệu

Buổi 3

6. Vấn đề về dữ liệu dẫn xuất

Buổi 3

7. Vấn đề về siêu dữ liệu

Buổi 3

8. Các vấn đề khác về kho dữ liệu

Buổi 3

Tổng kết nội dung bài học

Buổi 3

Thảo luận 3 phút

Buổi 3

THANK YOU FOR YOUR ATTENTION