DATA WAREHOUSES AND DECISION SUPPORT SYSTEMS

Kho dữ liệu và Hệ hỗ trợ quyết định

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

Chương 1: Tổng quan về kho dữ liệu và hệ hỗ trợ ra quyết định

Nội dung

- Khái niệm
- Ý nghĩa
- Vai trò
- Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

1. Khái niệm

Định nghĩa Kho dữ liệu (Data Warehouse) theo **Bill Inmon**.

Kho dữ liệu là một tập hợp dữ liệu có định hướng theo chủ đề, được tích hợp từ nhiều nguồn, có tính lịch sử, và không thay đổi, nhằm hỗ trợ quá trình ra quyết định trong tổ chức.

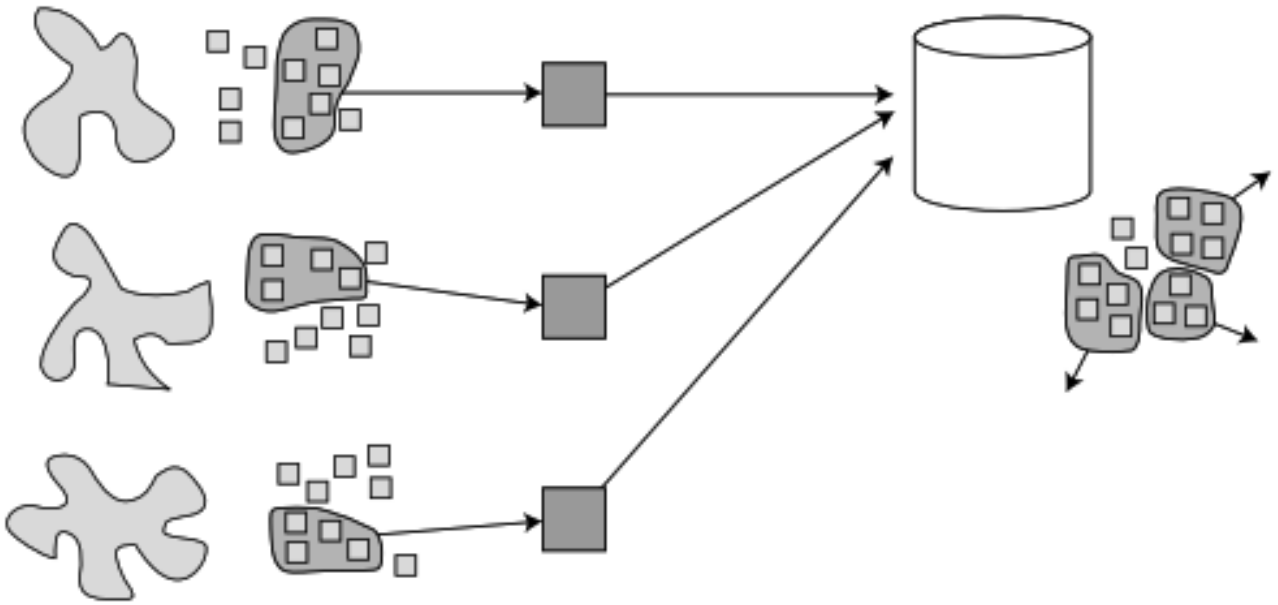
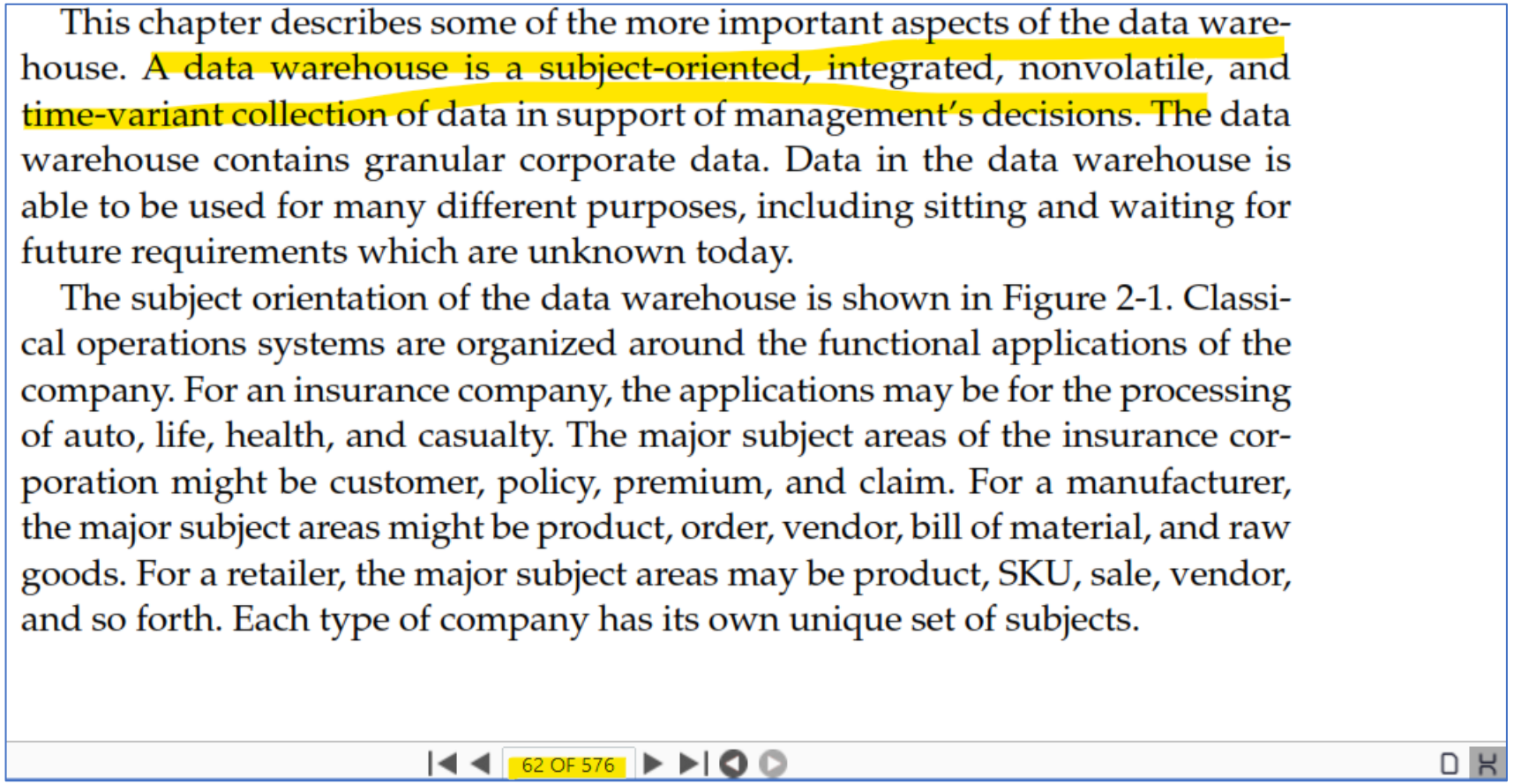


Figure 14-20 The data warehouse becomes the system of record for historical and DSS data.

1. Khái niệm

Định nghĩa Kho dữ liệu (Data Warehouse) theo **Bill Inmon**.

(**Building the Data Warehouse, 4th Edition – Bill Inmon, 2005**)



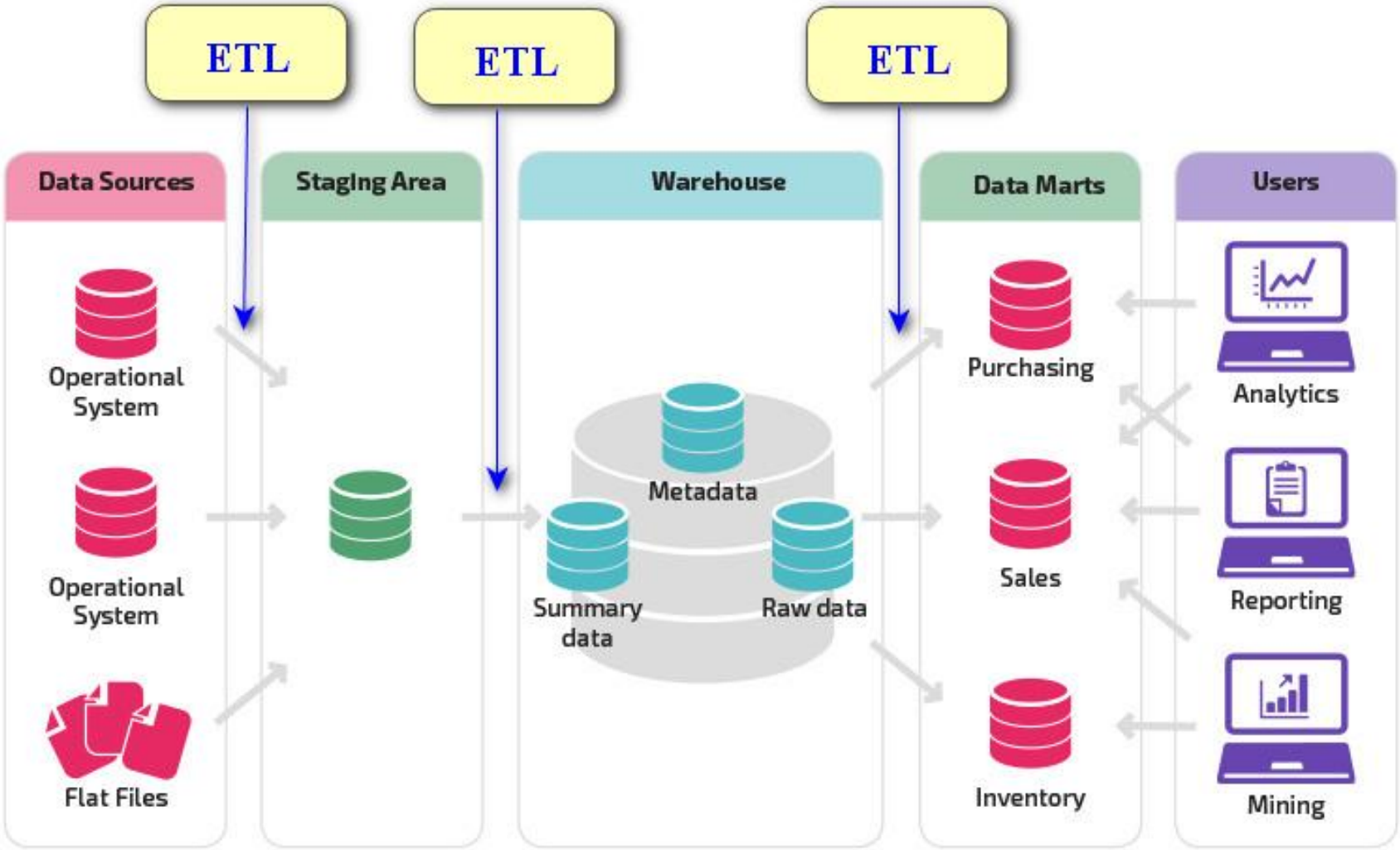
This chapter describes some of the more important aspects of the data warehouse. A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions. The data warehouse contains granular corporate data. Data in the data warehouse is able to be used for many different purposes, including sitting and waiting for future requirements which are unknown today.

The subject orientation of the data warehouse is shown in Figure 2-1. Classical operations systems are organized around the functional applications of the company. For an insurance company, the applications may be for the processing of auto, life, health, and casualty. The major subject areas of the insurance corporation might be customer, policy, premium, and claim. For a manufacturer, the major subject areas might be product, order, vendor, bill of material, and raw goods. For a retailer, the major subject areas may be product, SKU, sale, vendor, and so forth. Each type of company has its own unique set of subjects.

62 OF 576

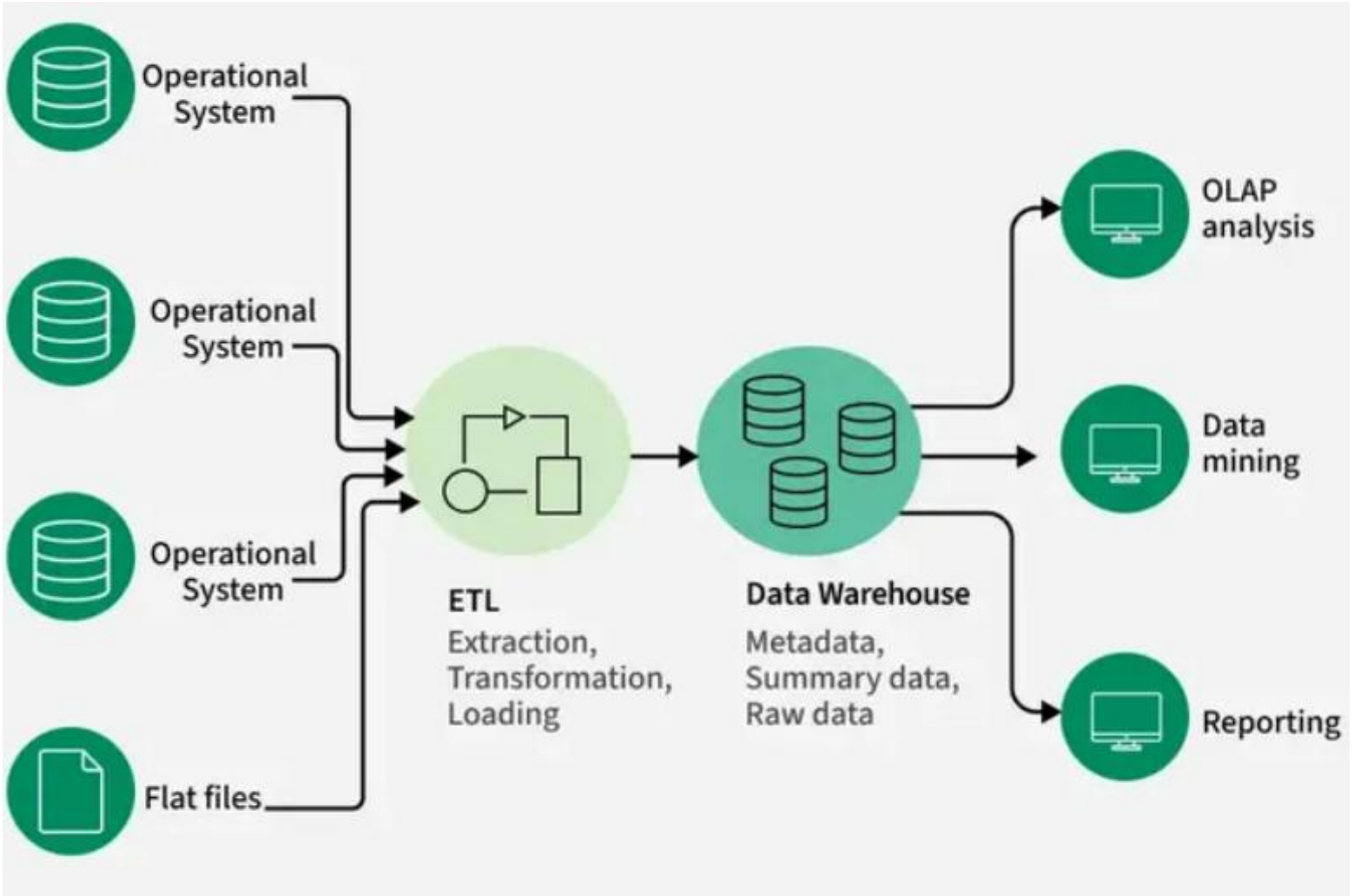
1. Khái niệm

Minh họa tổng Quan về hệ thống Kho dữ liệu (Data Warehouse) theo Bill Inmon.



1. Khái niệm

Minh họa tổng Quan, rút gọn staging area, về hệ thống Kho dữ liệu (Data Warehouse) theo Bill Inmon.



1. Khái niệm

Các đặc điểm chính của kho dữ liệu:

Tích hợp (Integrated)

Chủ đề (Subject-oriented)

Có tính lịch sử (Time-variant)

Không thay đổi (Non-volatile)

Giải thích Các đặc điểm chính của kho dữ liệu:

Chủ đề, có **Hướng Chủ Thể, (Subject-oriented):** tập trung theo chủ đề kinh doanh (khách hàng, sản phẩm, doanh thu…) thay vì giao dịch hằng ngày.

Tích hợp (Integrated): dữ liệu được hợp nhất từ nhiều nguồn khác nhau.

Có tính lịch sử (Time-variant): lưu giữ dữ liệu trong quá khứ *để phân tích xu hướng*, chứa xu hướng và sự thay đổi

Không thay đổi (Non-volatile): dữ liệu trong DW không bị ghi đè hay xóa thường xuyên, chỉ thêm mới.

1. Khái niệm

Sự khác biệt giữa OLTP (tập trung vào vận hành, giao dịch tức thì) và DW (tập trung phân tích, hỗ trợ quyết định).

Ví dụ: hệ thống bán hàng ghi nhận hóa đơn (OLTP) so với phân tích xu hướng doanh thu theo tháng (DW).

1. Khái niệm

Granularity (Hạt, Độ mịn dữ liệu, Độ chi tiết dữ liệu trong kho dữ liệu),

Granularity là mức độ chi tiết hoặc tổng hợp của dữ liệu được lưu trữ trong Data Warehouse.

Nói cách khác, Granularity cho biết dữ liệu trong DW được lưu ở mức chi tiết (fine-grained) hay tổng hợp (coarse-grained).

Đây là một trong những yếu tố thiết kế quan trọng nhất khi xây dựng kho dữ liệu, vì nó ảnh hưởng trực tiếp đến:

Kích thước lưu trữ (storage size)

Tốc độ truy vấn

Khả năng phân tích

1. Khái niệm

Các mức Granularity

a. Fine Granularity (Độ chi tiết cao)

Lưu dữ liệu ở mức chi tiết nhất (transaction-level).

Ví dụ: mỗi hóa đơn bán hàng, từng giao dịch ngân hàng, log truy cập website từng giây.

Ưu điểm: phân tích được chi tiết, linh hoạt.

Nhược điểm: tốn dung lượng, truy vấn chậm hơn.

b. Coarse Granularity (Độ chi tiết thấp – dữ liệu tổng hợp)

Lưu dữ liệu đã được tổng hợp (summary-level).

Ví dụ: doanh số theo ngày, doanh thu theo tháng, số lượng khách hàng theo khu vực.

Ưu điểm: tiết kiệm dung lượng, truy vấn nhanh.

Nhược điểm: mất chi tiết, không drill-down được sâu.

1. Khái niệm

Nguyên tắc thiết kế Granularity

Thông thường, Data Warehouse ưu tiên lưu dữ liệu chi tiết (fine-grained), sau đó tạo các bảng/tầng dữ liệu tổng hợp (summary tables) để tối ưu hiệu suất truy vấn.

Cách tiếp cận này giúp cân bằng:

Lưu giữ toàn bộ dữ liệu gốc để phân tích khi cần.

Dùng dữ liệu tổng hợp cho báo cáo nhanh.

1. Khái niệm

Minh họa Nguyên tắc thiết kế Granularity

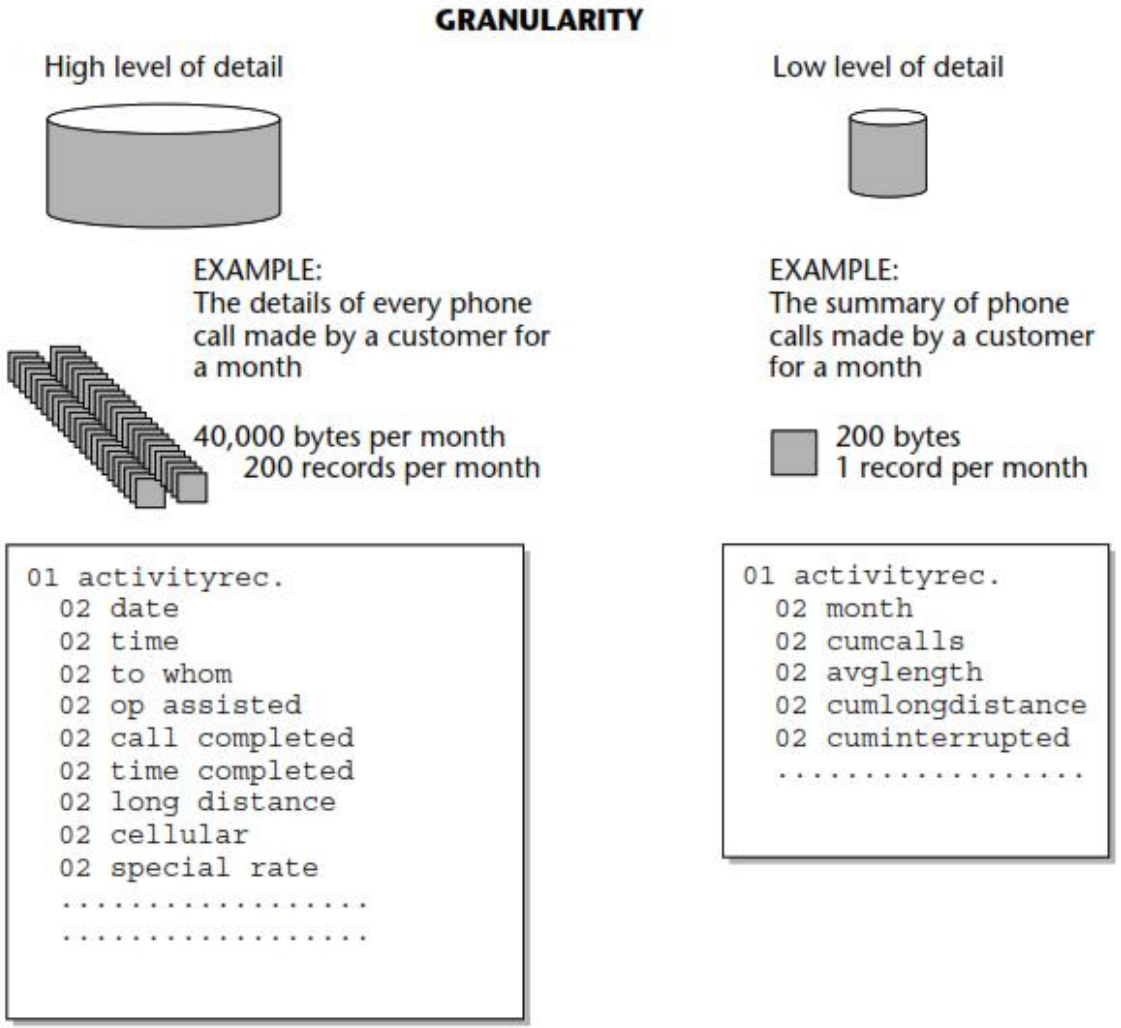


Figure 2-12 Determining the level of granularity is the most important design issue in the data warehouse environment.

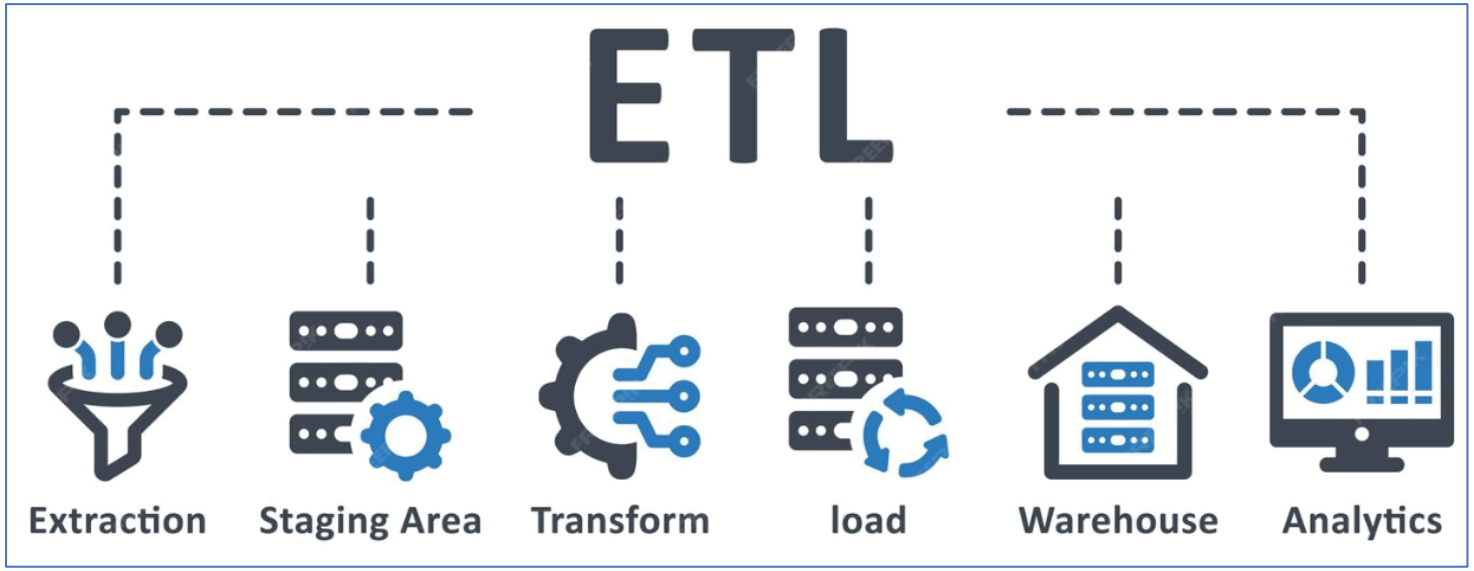
1. Khái niệm

Thuật ngữ: ETL (Extract – Transform – Load)

ETL là viết tắt của Extract – Transform – Load, tức là Trích rút – Biến đổi – Nạp.

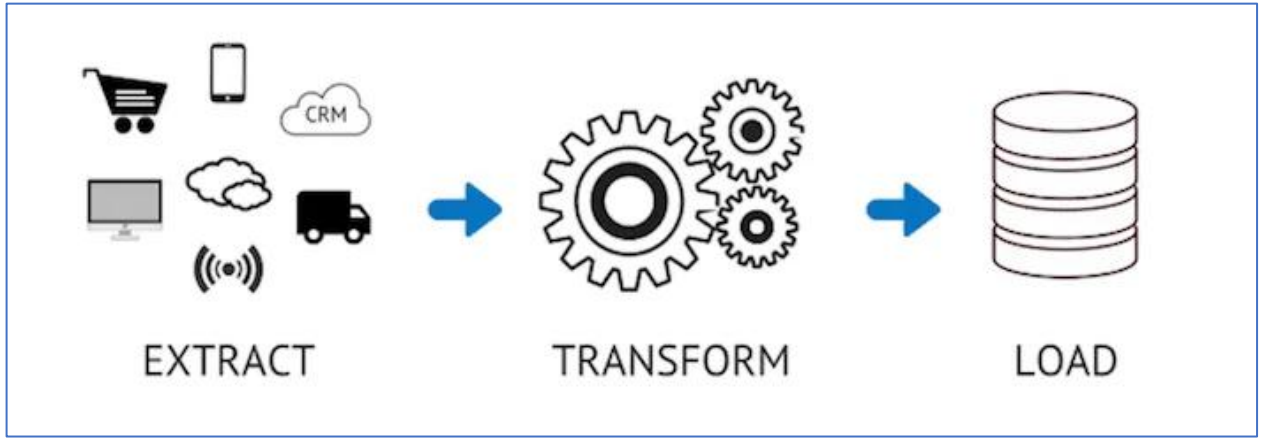
Đây là quy trình cốt lõi trong việc xây dựng Kho dữ liệu (Data Warehouse).

Nhiệm vụ: thu thập dữ liệu từ nhiều nguồn khác nhau, xử lý để đảm bảo chất lượng và đồng nhất, sau đó nạp vào kho dữ liệu trung tâm.



1. Khái niệm

Thuật ngữ: **ETL** (Extract – Transform – Load) có Ba bước chính



1. Extract (Trích rút dữ liệu):

Thu thập dữ liệu từ nhiều nguồn khác nhau: hệ thống giao dịch (OLTP), ERP, CRM, POS, file Excel, log web…

Đảm bảo dữ liệu được lấy đầy đủ, không mất mát.

Có thể trích rút theo lô (batch) hoặc gần thời gian thực (real-time).

1. Khái niệm

2. Transform (Biến đổi dữ liệu):

Là giai đoạn xử lý quan trọng nhất.

Các công việc thường bao gồm:

Chuẩn hóa định dạng (ví dụ: ngày tháng, đơn vị tiền tệ).

Làm sạch dữ liệu (xử lý dữ liệu trùng, thiếu, sai).

Tích hợp dữ liệu từ nhiều nguồn (hợp nhất mã khách hàng từ nhiều hệ thống).

Tính toán, tổng hợp, mã hóa, phân loại.

Mục tiêu: biến dữ liệu thô thành dữ liệu chính xác, nhất quán, phù hợp để phân tích.

1. Khái niệm

3. Load (Nạp dữ liệu):

Nạp dữ liệu đã xử lý vào kho dữ liệu (Data Warehouse) hoặc Data Mart.

Có 2 cách nạp chính:

Full Load: nạp toàn bộ dữ liệu mỗi lần.

Incremental Load: chỉ nạp dữ liệu mới hoặc thay đổi.

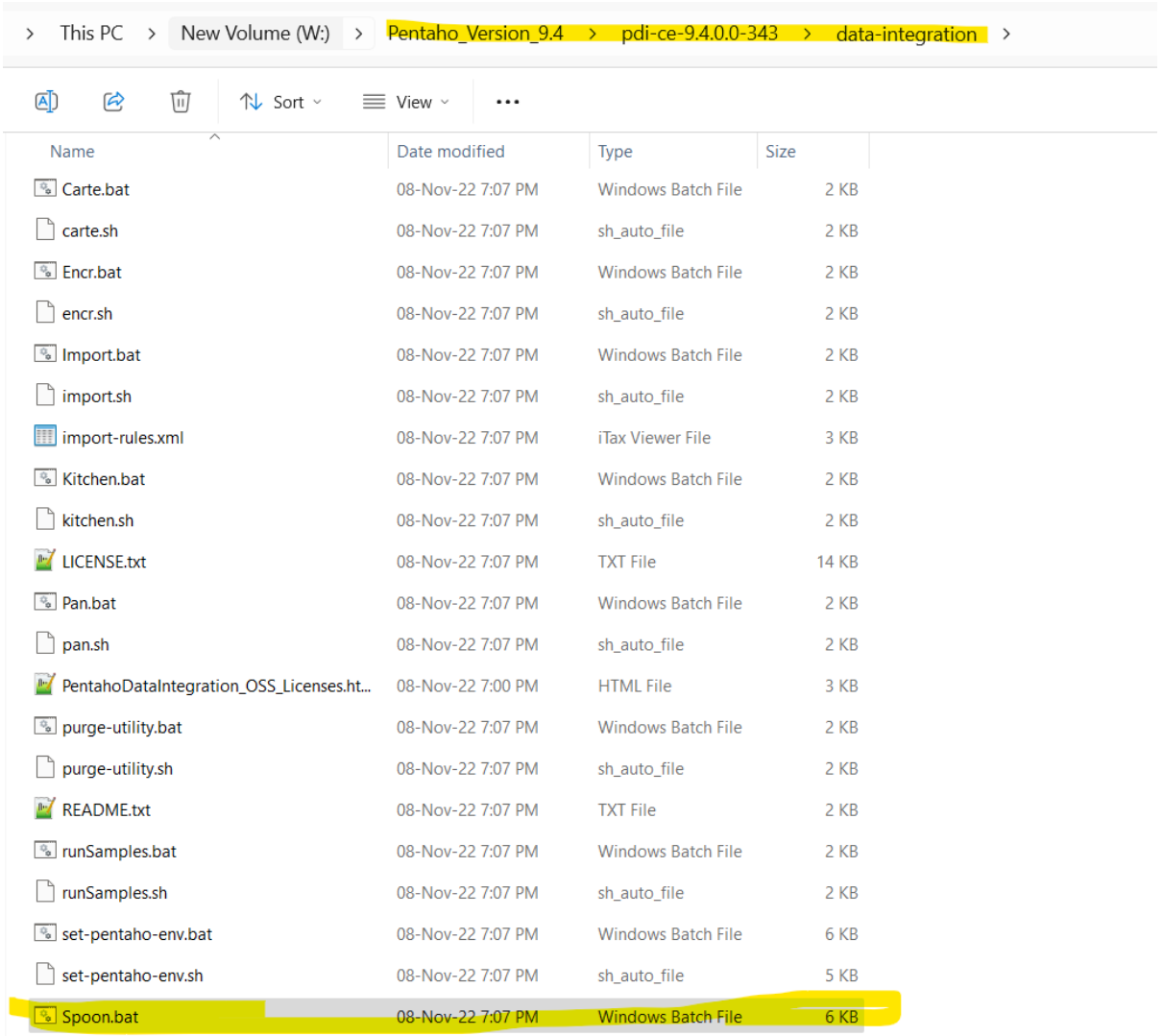
Đảm bảo tốc độ nạp nhanh, không gây ảnh hưởng đến hệ thống phân tích.

Công cụ ETL: Pentaho download | SourceForge.net https://sourceforge.net/projects/pentaho/

https://pentaho.com/pentaho-developer-edition/ (nhớ điền thông tin mới cho downloads về)

1. Khái niệm

https://pentaho.com/pentaho-developer-edition/ (nhớ điền thông tin mới cho downloads về)



1. Khái niệm

Ý nghĩa của ETL

Là cầu nối giúp dữ liệu từ nhiều hệ thống khác nhau trở nên nhất quán và đáng tin cậy.

Đảm bảo kho dữ liệu luôn cập nhật, chính xác, đầy đủ.

Giúp cho việc phân tích, báo cáo và ra quyết định trong BI (Business Intelligence) và DSS (Decision Support System) trở nên hiệu quả.

Nếu không có công cụ ETL, dữ liệu trong kho sẽ rời rạc, sai lệch và không dùng được cho phân tích.

Sinh viên làm ví dụ:

Nguồn A dùng định dạng ngày dd/mm/yyyy.

Nguồn B dùng mm-dd-yyyy.

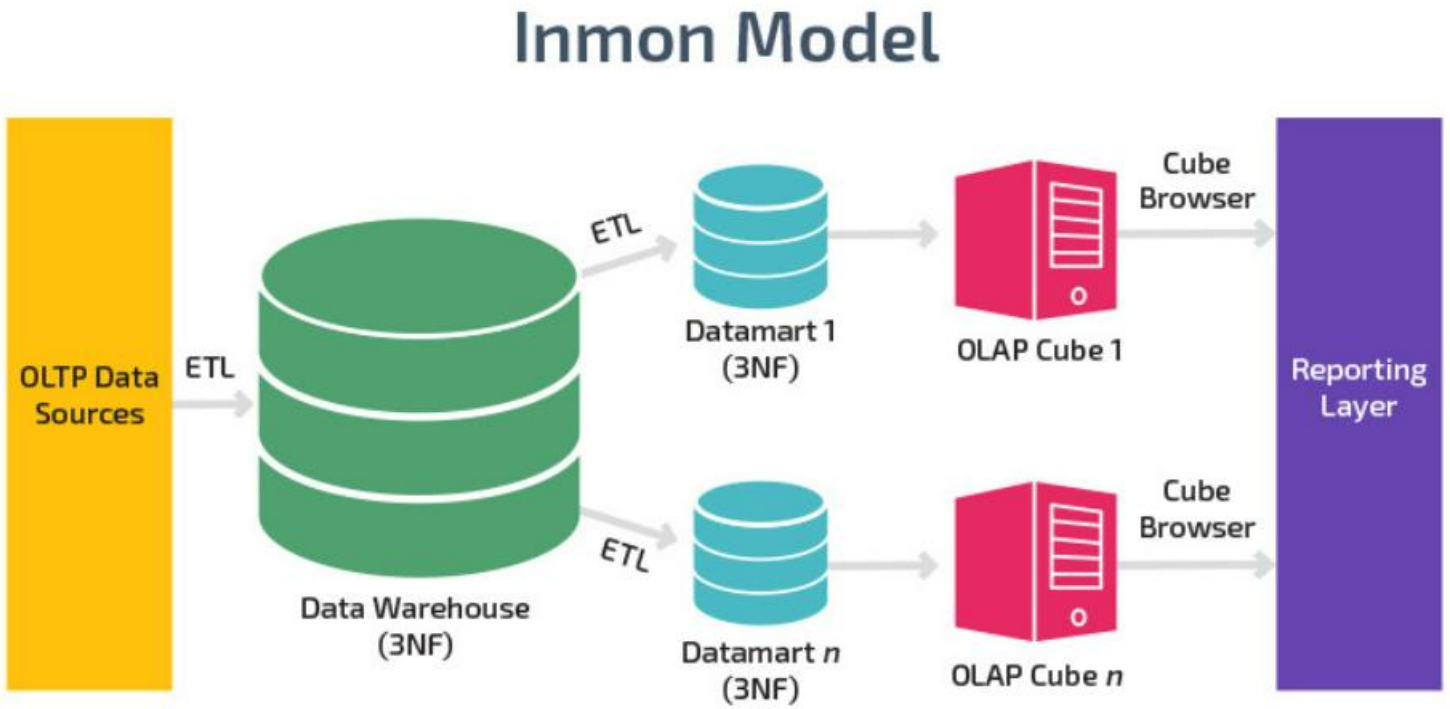
ETL phải biến đổi để thống nhất → nếu không thì báo cáo sai.

1. Khái niệm

Data Mart

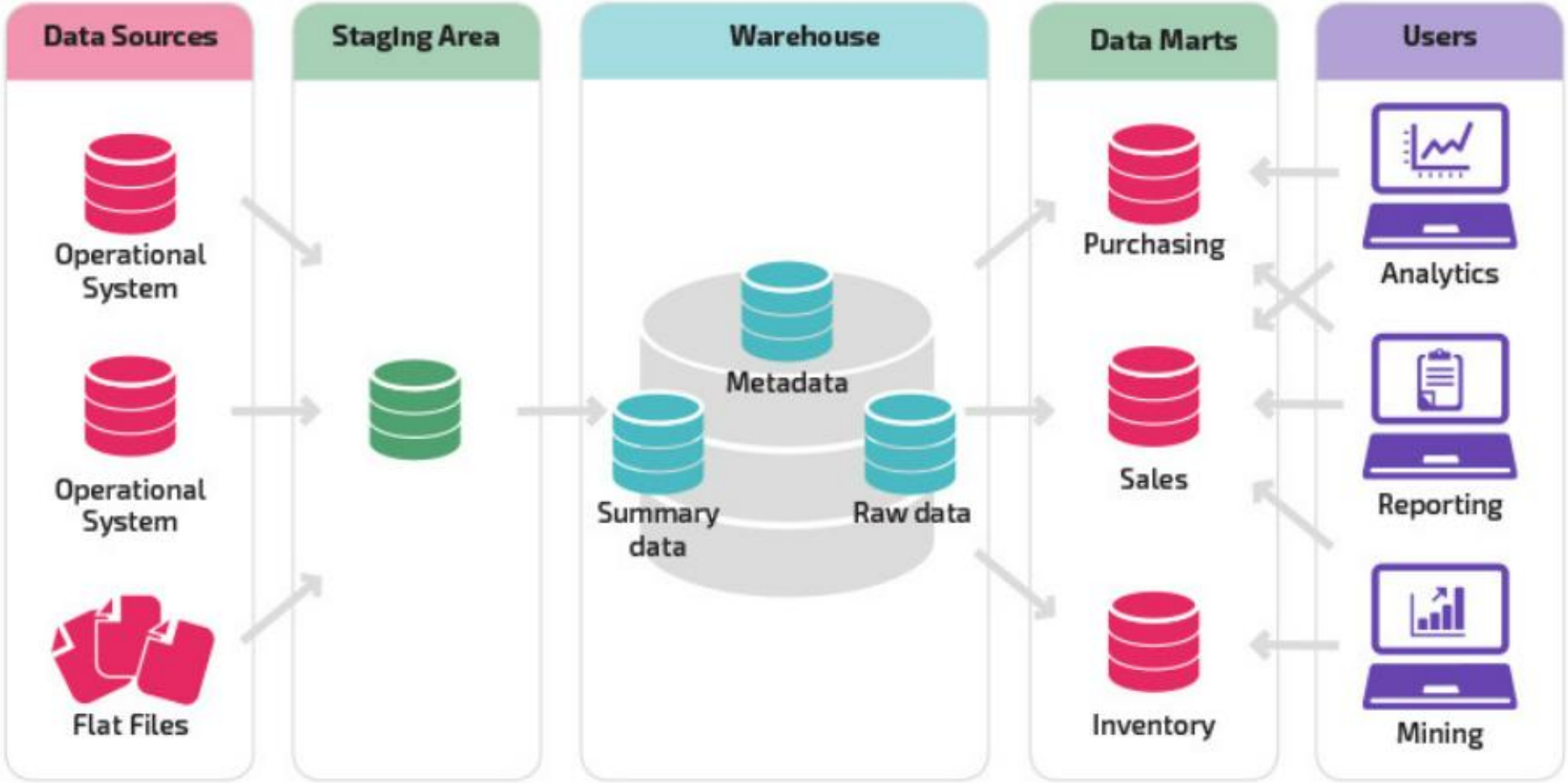
Khái niệm: **Data Mart là một phân vùng nhỏ của kho dữ liệu (Data Warehouse),** được thiết kế để phục vụ một bộ phận, lĩnh vực kinh doanh hoặc nhóm người dùng cụ thể. **VÀ được lấy thêm dữ liệu bên ngoài Data Warehouse**

Minh họa : Nếu Data Warehouse là "thư viện tổng hợp" của doanh nghiệp, thì Data Mart giống như "tủ sách chuyên ngành" cho từng phòng ban.



1. Khái niệm

Data Mart



1. Khái niệm

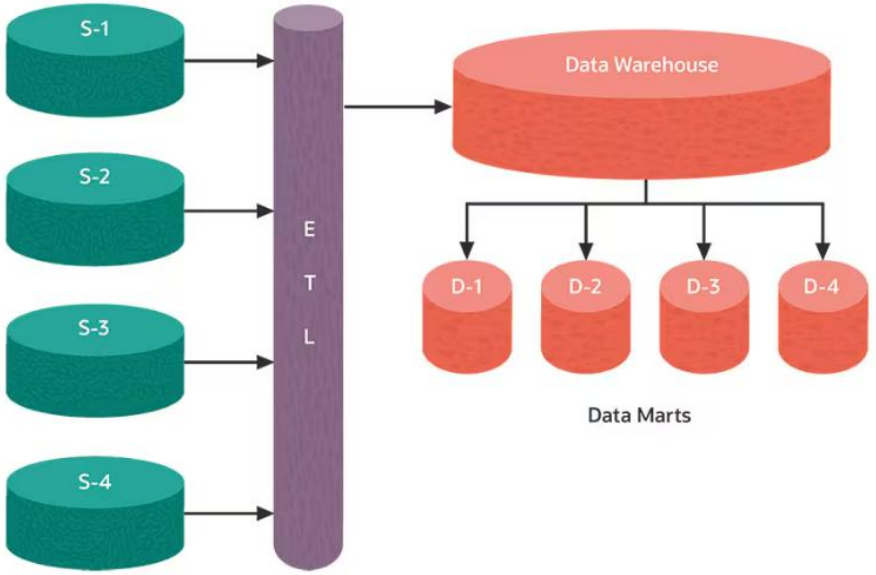
Đặc điểm của **Data Mart**

Phạm vi hẹp: tập trung vào một chủ đề hoặc bộ phận (Marketing, Tài chính, Nhân sự…).

Nguồn dữ liệu: có thể lấy trực tiếp từ kho dữ liệu trung tâm hoặc từ hệ thống giao dịch (OLTP).

Dễ triển khai: thời gian xây dựng nhanh hơn so với Data Warehouse toàn diện.

Tối ưu cho người dùng: dữ liệu được tổ chức theo cách mà bộ phận sử dụng dễ dàng khai thác.



1. Khái niệm

Các loại Data Mart

Dependent Data Mart (Phụ thuộc):

Được trích xuất từ Data Warehouse trung tâm.

Đảm bảo tính đồng nhất và chuẩn hóa.

Ví dụ: Data Mart bán hàng được lấy từ kho dữ liệu tổng hợp của công ty.

Independent Data Mart (Độc lập):

Xây dựng trực tiếp từ nguồn dữ liệu giao dịch mà không qua Data Warehouse.

Nhanh, rẻ, nhưng dễ bị thiếu tính tích hợp.

Hybrid Data Mart (Lai):

Kết hợp cả hai: lấy dữ liệu từ kho trung tâm và từ nguồn ngoài.

Phù hợp khi vừa cần dữ liệu chuẩn hóa, vừa cần dữ liệu nhanh chóng.

1. Khái niệm

Ý nghĩa của Data Mart

Giúp các bộ phận nhanh chóng truy cập dữ liệu phù hợp mà không phải tìm kiếm trong kho dữ liệu khổng lồ.

Giảm tải cho Data Warehouse trung tâm.

Thúc đẩy ra quyết định nhanh và chuyên sâu trong từng lĩnh vực.

Linh hoạt hơn trong việc triển khai các giải pháp BI (Business Intelligence).

1. Khái niệm

Ví dụ minh họa

Bộ phận Marketing: dùng Data Mart để phân tích hành vi khách hàng, hiệu quả quảng cáo.

Bộ phận Tài chính: dùng Data Mart để theo dõi chi phí, lợi nhuận theo quý.

Bộ phận Bán hàng: dùng Data Mart để theo dõi doanh số theo vùng, sản phẩm.

1. Khái niệm

Trình bày mối quan hệ: Data Mart là "con" của Data Warehouse (trong trường hợp dependent).

Giải thích vì sao doanh nghiệp thường triển khai Data Mart trước, rồi mở rộng dần thành Data Warehouse (cách tiếp cận "bottom-up").

Sinh viên làm bài tập nhóm: thiết kế một Data Mart cho bộ phận Marketing của công ty bán lẻ.

1. Khái niệm

Phân biệt giữa:

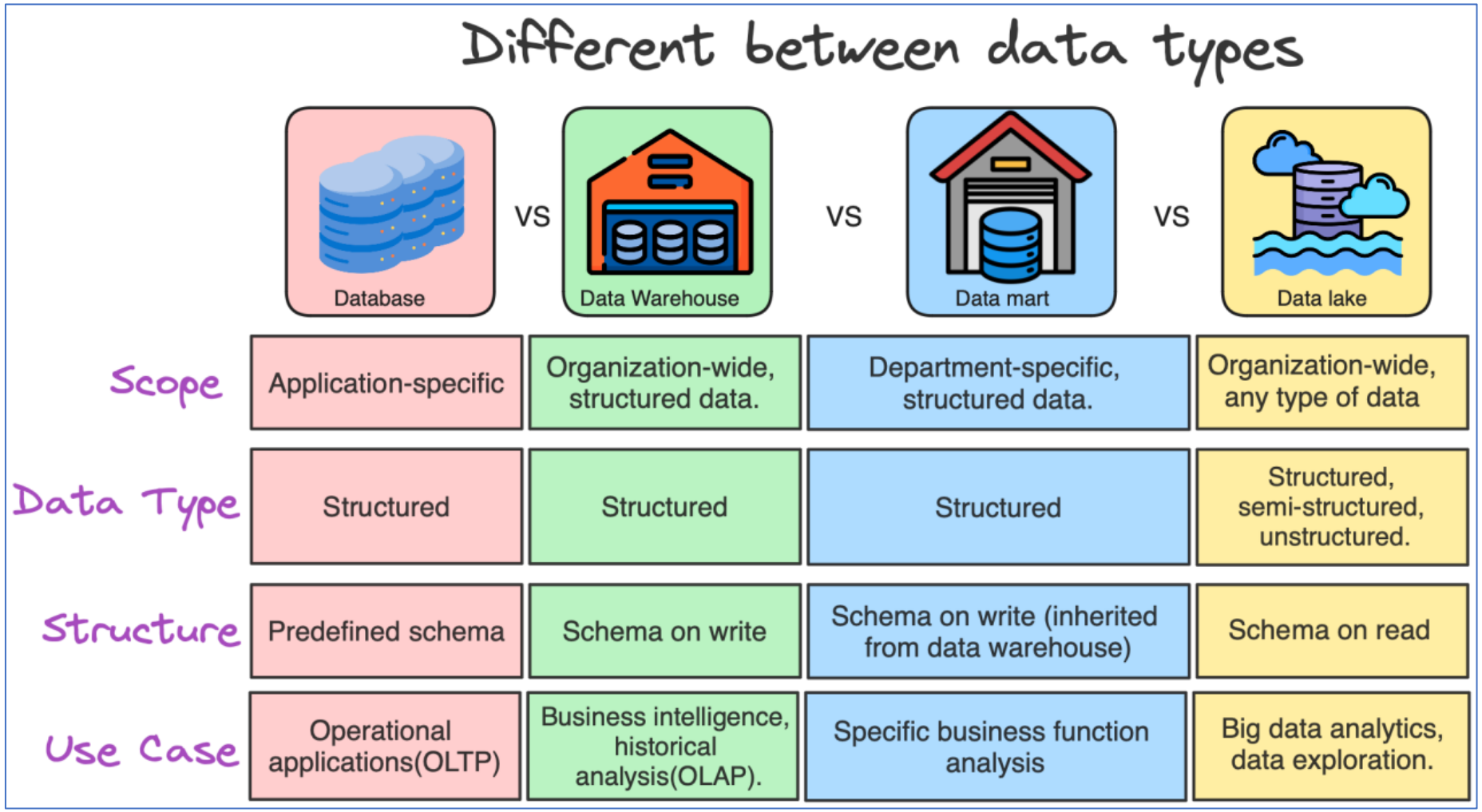
Hệ thống OLTP (Online Transaction Processing), hình bên trái

Hệ thống OLAP (Online Analytical Processing), hình bên **PHẢI**

Ví dụ minh họa: giao dịch bán hàng trong OLTP vs phân tích xu hướng bán hàng trong DW.

1. Khái niệm

So sánh:



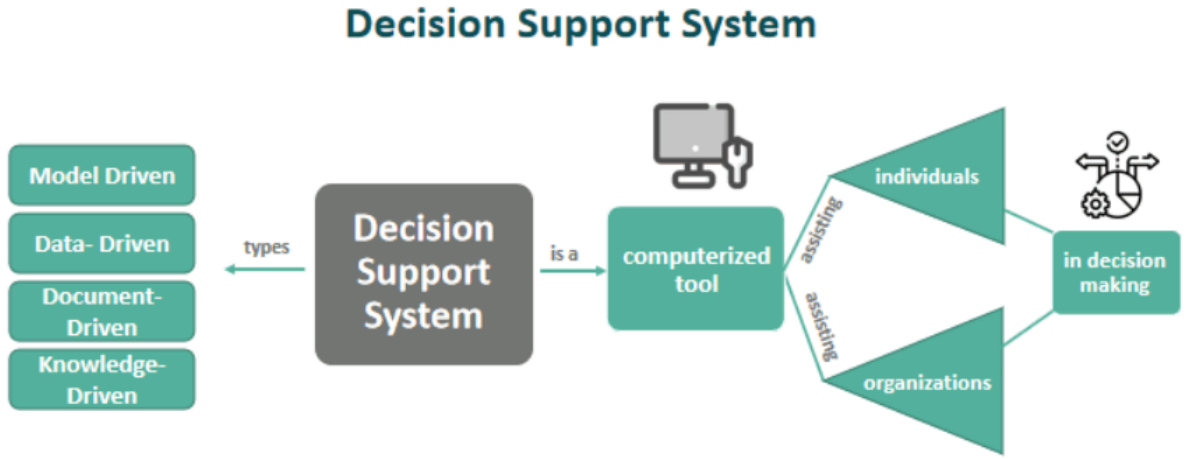
2. Ý nghĩa

Ý nghĩa của kho dữ liệu

Là nền tảng cho các hệ thống Business Intelligence (BI), DSS, Data Mining.

Giúp doanh nghiệp có cái nhìn toàn diện thay vì dữ liệu phân tán.

Tăng tính chính xác, tin cậy của thông tin trong quyết định.



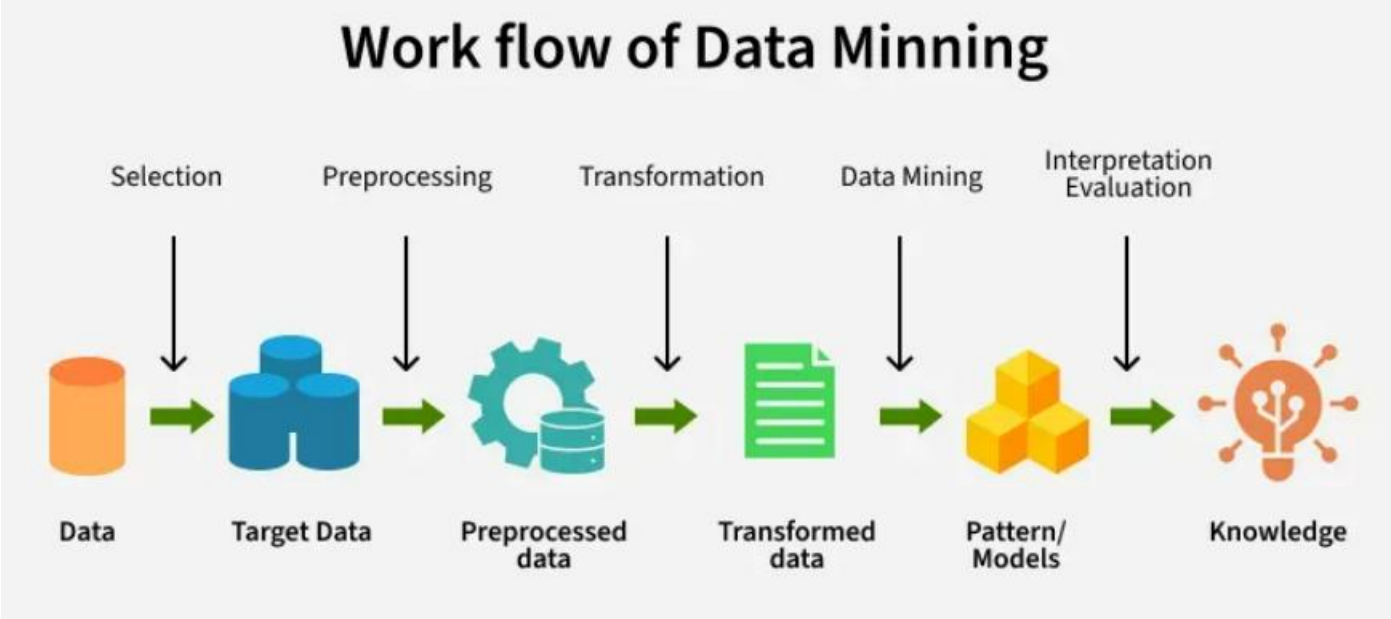
2. Ý nghĩa

Ý nghĩa của kho dữ liệu

Là nền tảng cho các hệ thống Business Intelligence (BI), DSS, Data Mining từ Data Mart

Giúp doanh nghiệp có cái nhìn toàn diện thay vì dữ liệu phân tán.

Tăng tính chính xác, tin cậy của thông tin trong quyết định.

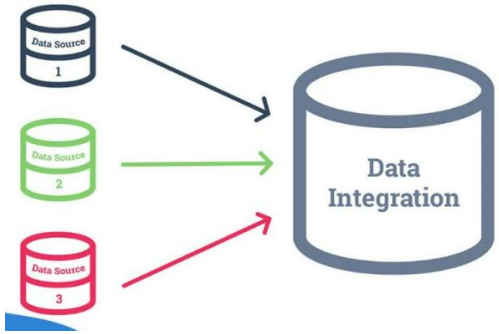


2. Ý nghĩa

Lợi ích cụ thể

Hỗ trợ phân tích xu hướng (ví dụ: dự báo nhu cầu).

Tích hợp dữ liệu từ nhiều nguồn → tránh trùng lặp, mâu thuẫn.



Nâng cao hiệu quả quản trị điều hành và ra quyết định chiến lược.

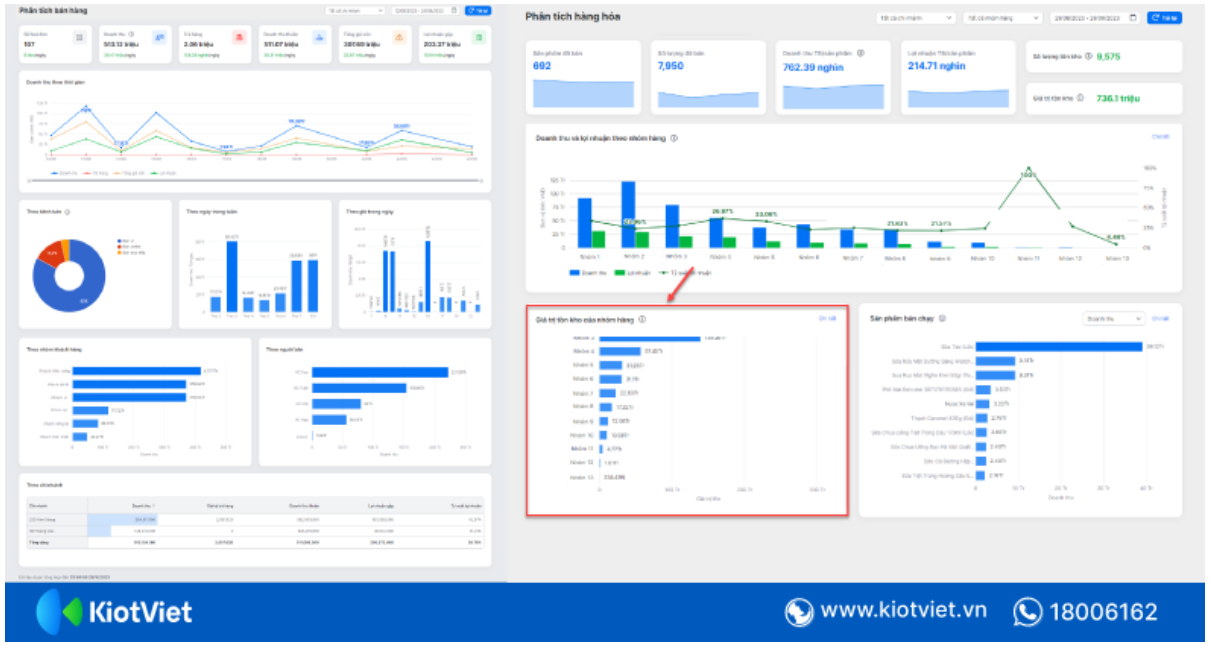
2. Ý nghĩa

Ví dụ minh họa

Một công ty bán lẻ:

OLTP: ghi nhận mỗi giao dịch bán hàng tại quầy.

DW: tổng hợp và phân tích xu hướng mua hàng để quyết định nhập kho trong tháng tới.



2. Ý nghĩa

Vì sao nhiều doanh nghiệp thất bại khi ra quyết định nếu chỉ dựa vào dữ liệu phân mảnh.

Dùng ví dụ thực tế (Amazon, Walmart) để minh họa sức mạnh của DW trong dự báo tồn kho và cá nhân hóa dịch vụ. (chứa xu hướng và sự thay đổi)

2. Ý nghĩa

Vì sao doanh nghiệp cần kho dữ liệu.

Kho dữ liệu như nền tảng của Business Intelligence (BI).

Lợi ích mang lại:

- Cung cấp dữ liệu đáng tin cậy cho phân tích.
- Hỗ trợ đưa ra quyết định chính xác.
- Tích hợp dữ liệu từ nhiều nguồn khác nhau.
- Cải thiện hiệu suất phân tích so với dữ liệu giao dịch.

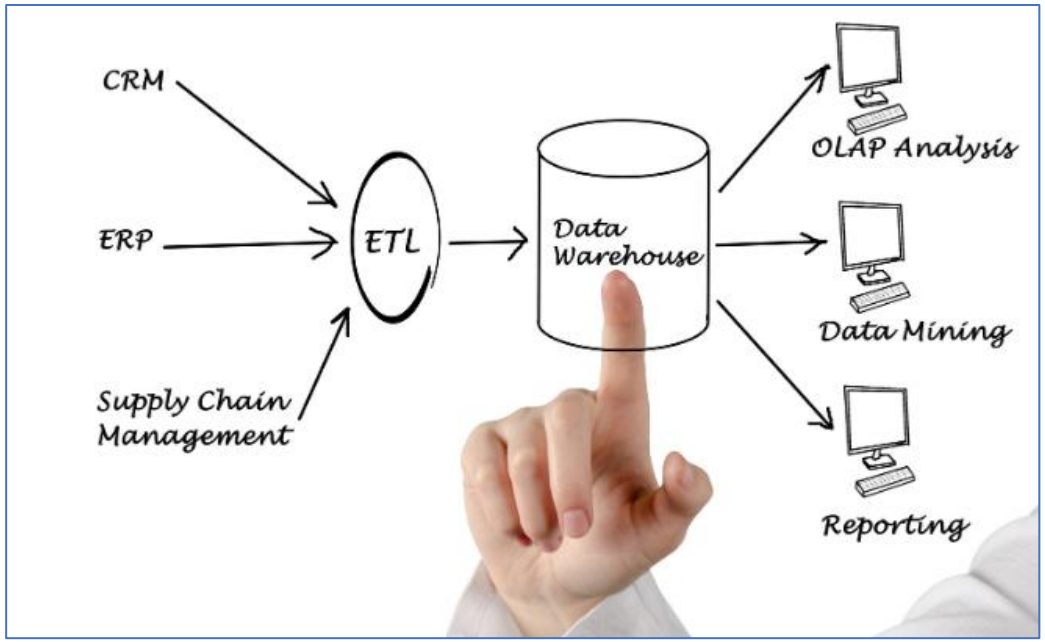
Ví dụ: Tập đoàn bán lẻ dùng DW để dự báo nhu cầu và quản lý tồn kho.

3. Vai trò

Vai trò của Data Warehouse

a. Trong hệ thống thông tin doanh nghiệp:

Cầu nối giữa dữ liệu vận hành và dữ liệu cho ra quyết định.



b. Trong quản lý:

Ban giám đốc: hoạch định chiến lược dài hạn.

Bộ phận tài chính: phân tích lợi nhuận, chi phí.

Marketing: phân khúc khách hàng, hiệu quả chiến dịch.

Sản xuất: tối ưu quy trình, dự báo nhu cầu.

c. Trong DSS (Decision Support System):

Cung cấp dữ liệu nền để DSS hoạt động.

Biến dữ liệu thô thành tri thức phục vụ phân tích.

3. Vai trò

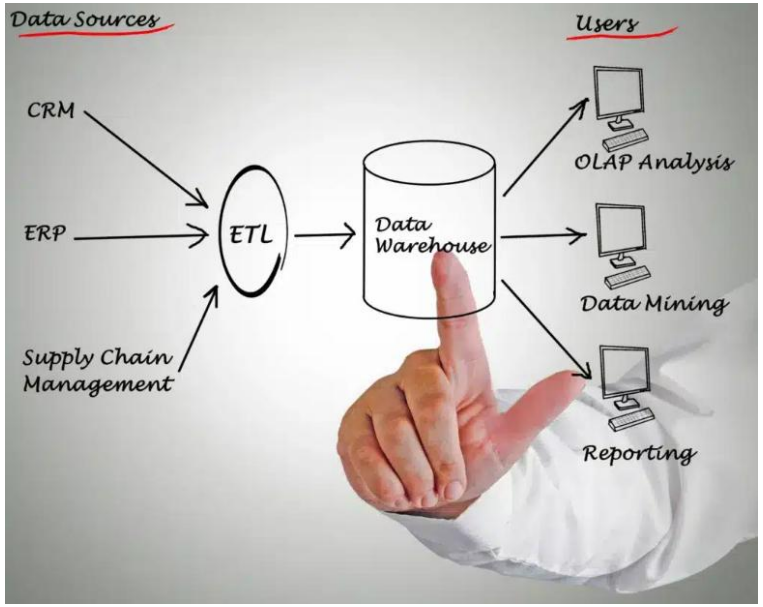
Vai trò trong hệ thống thông tin doanh nghiệp.

Vai trò đối với các bộ phận: quản lý, marketing, tài chính, sản xuất.

Mối quan hệ giữa kho dữ liệu và hệ hỗ trợ ra quyết định (DSS – Decision Support System).

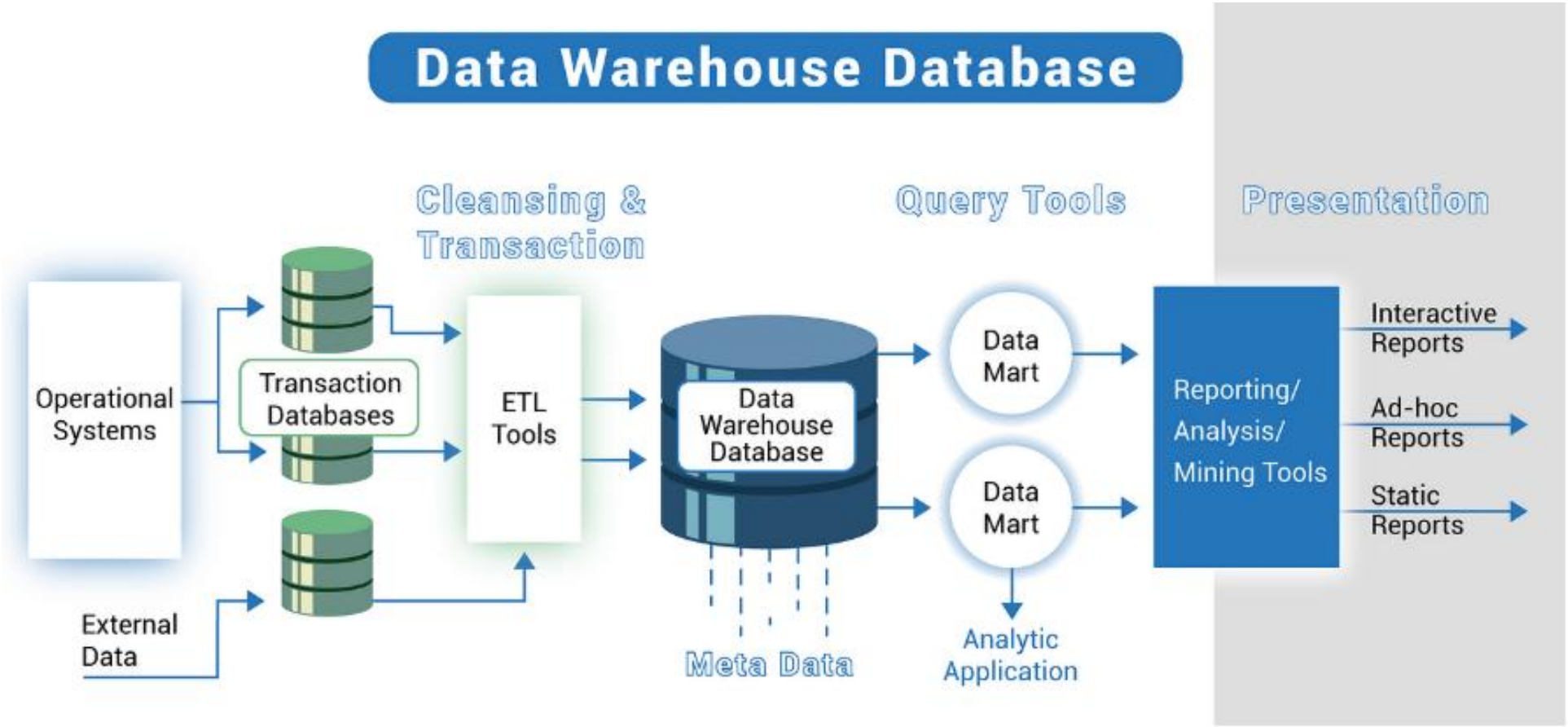
Vai trò chiến lược: biến dữ liệu thô thành tri thức.

Minh họa bằng sơ đồ Vai trò của Data Warehouse và Decision Support System: Data Sources → Data Warehouse → DSS → Quyết định.



3. Vai trò

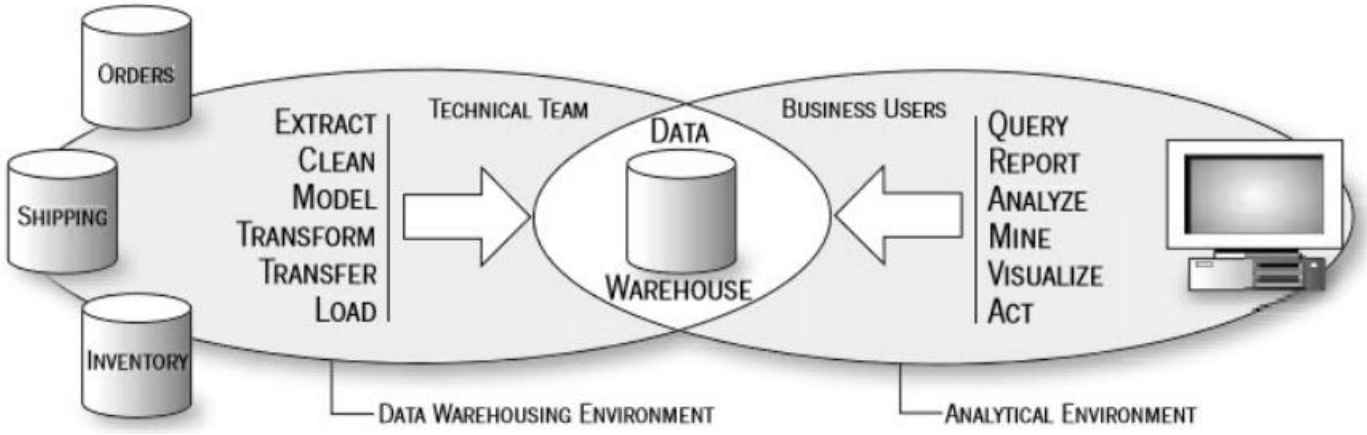
Minh họa bằng sơ đồ: Data Sources → Data Warehouse → DSS → Quyết định.



3. Vai trò

Nhấn mạnh: DSS không thể hoạt động hiệu quả nếu thiếu DW.

Sinh viên thảo luận tình huống: "Công ty không dùng DW thì sẽ gặp khó khăn gì trong việc ra quyết định?".



4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

Ba cấp độ ứng dụng

a. Cấp tác nghiệp (Operational):

Hỗ trợ quyết định trong công việc hằng ngày.

Ví dụ: xác định sản phẩm nào bán chạy nhất trong tuần để bổ sung ngay.

b. Cấp quản lý trung gian (Tactical):

Quyết định ngắn hạn, quản lý bộ phận.

Ví dụ: phân tích hiệu quả chương trình khuyến mãi theo khu vực.

c. Cấp chiến lược (Strategic):

Hoạch định dài hạn, định hướng toàn công ty.

Ví dụ: dự báo xu hướng thị trường 5 năm để mở rộng chi nhánh.

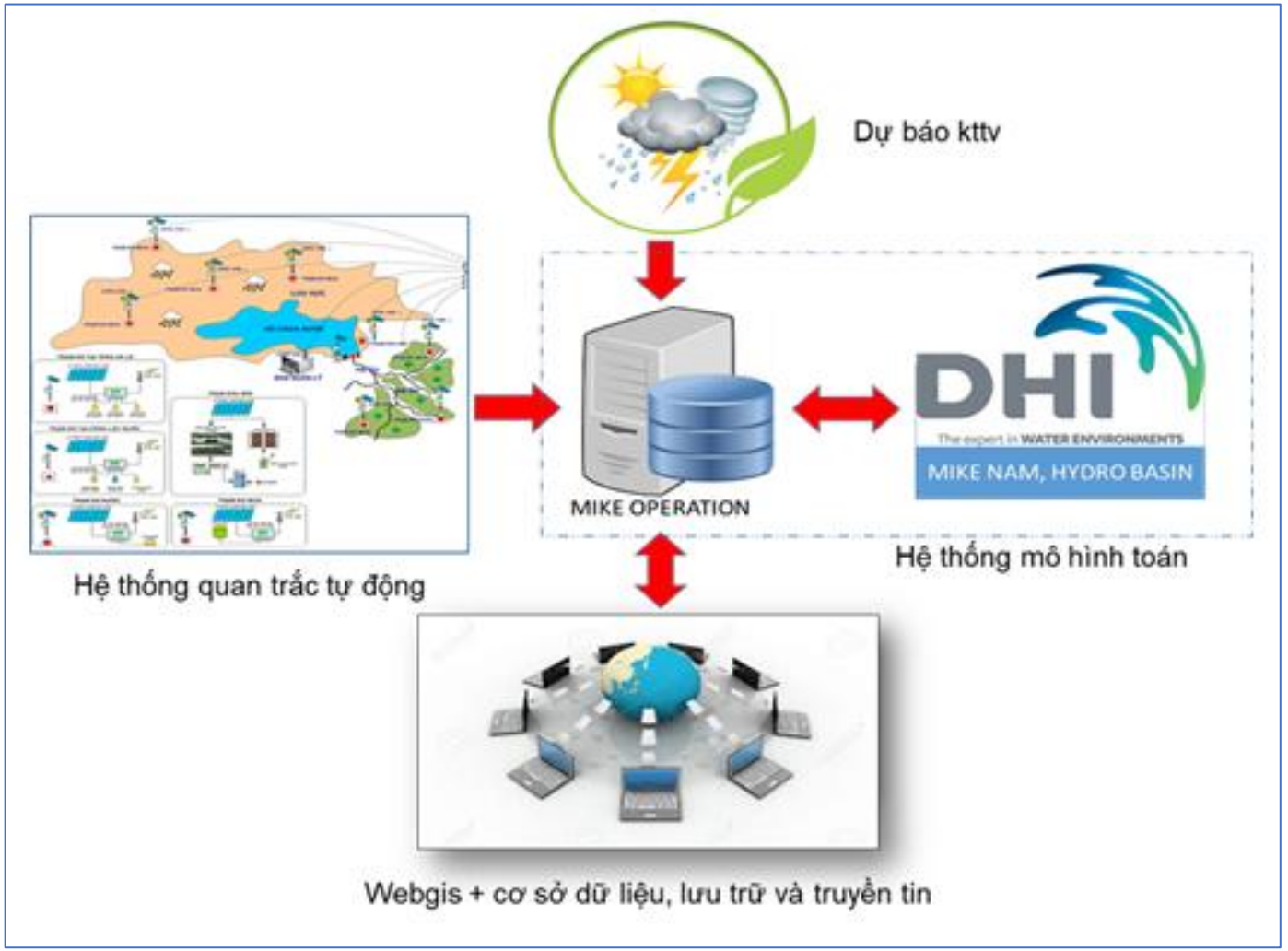
4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

Bảng so sánh nhu cầu thông tin

Mức quản lý	Đặc điểm thông tin	Ví dụ ứng dụng
Tác nghiệp	Chi tiết, thời gian thực	Báo cáo bán hàng ngày
Trung gian	Tổng hợp theo chu kỳ (tuần/tháng)	Đánh giá hiệu quả marketing
Chiến lược	Dữ liệu lịch sử, xu hướng dài hạn	Phân tích thị trường 5 năm

4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

Xây dựng hệ thống hỗ trợ ra quyết định kiểm soát mặn



4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

Giải pháp hỗ trợ nông dân phục hồi, phát triển nông nghiệp

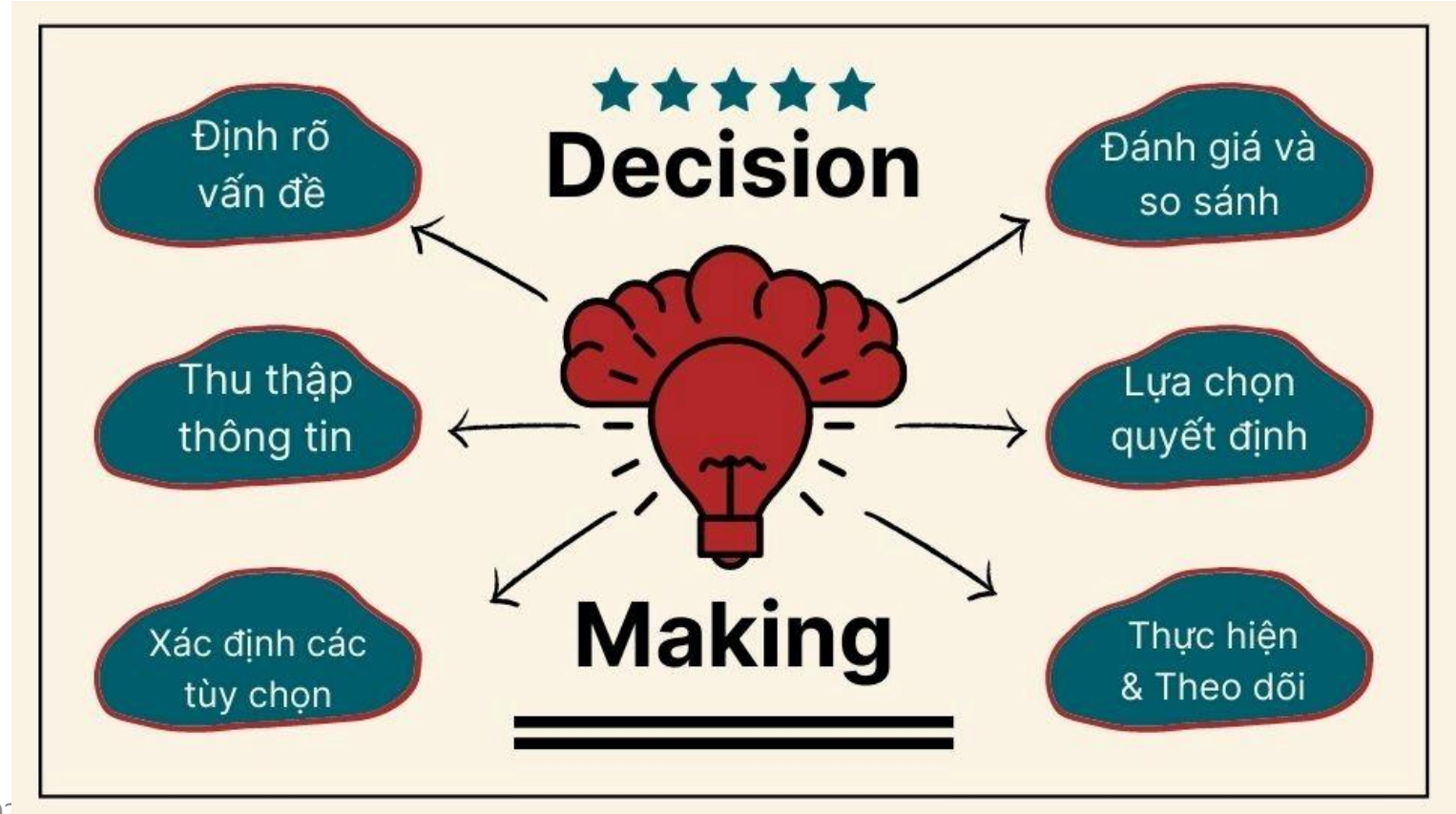
4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

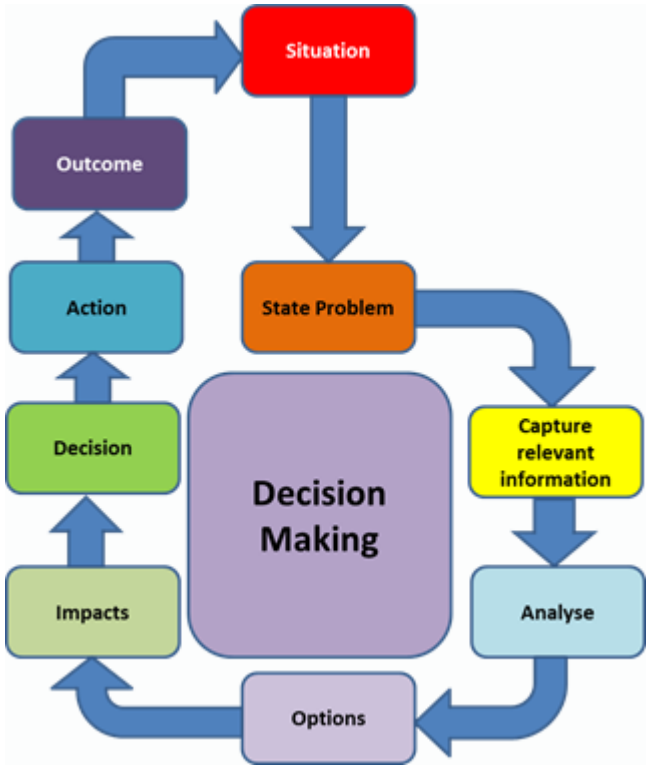
Giải pháp Phát triển Thương mại điện tử xanh

4. Ứng dụng trong việc hỗ trợ ra quyết định ở các mức quản lý khác nhau

Sinh viên làm bài tập nhóm: đưa ra ví dụ thực tế về cách DW hỗ trợ quyết định ở 3 cấp độ.

Nhấn mạnh: cùng một kho dữ liệu nhưng cách dùng sẽ khác nhau theo cấp quản lý.





Tóm tắt nội dung đã học

Khái niệm

Kho dữ liệu (Data Warehouse) theo Bill Inmon: dữ liệu theo chủ đề, tích hợp, có tính lịch sử, không thay đổi, phục vụ hỗ trợ ra quyết định.

Sự khác biệt giữa OLTP (xử lý giao dịch hằng ngày), OLAP là viết tắt của Online Analytical Processing - Hệ thống xử lý phân tích trực tuyến được thiết kế để phân tích khối lượng dữ liệu lớn, thường là dữ liệu lịch sử, và DW (phân tích, hỗ trợ quyết định).

Thuật ngữ ETL (Extract – Transform – Load): quy trình trích rút, biến đổi, và nạp dữ liệu vào DW.

Data Mart: phân vùng nhỏ của DW, phục vụ từng bộ phận/nhóm người dùng.

Tóm tắt nội dung đã học

Ý nghĩa

Kho dữ liệu là nền tảng cho BI, DSS, Data Mining.

Giúp doanh nghiệp có cái nhìn toàn diện, dữ liệu chính xác, đáng tin cậy.

Lợi ích: phân tích xu hướng, tích hợp dữ liệu đa nguồn, cải thiện hiệu quả quản trị.

Ví dụ: công ty bán lẻ dùng DW để dự báo nhu cầu và quản lý tồn kho.

Tóm tắt nội dung đã học

Vai trò

Trong hệ thống thông tin doanh nghiệp: cầu nối giữa dữ liệu vận hành và dữ liệu phân tích.

Trong quản lý:

- Ban giám đốc: hoạch định chiến lược.
- Marketing: phân khúc khách hàng.
- Tài chính: phân tích lợi nhuận, chi phí.
- Sản xuất: tối ưu quy trình, dự báo nhu cầu.

Trong DSS: biến dữ liệu thô thành tri thức.

DSS không thể hoạt động hiệu quả nếu thiếu DW.

Tóm tắt nội dung đã học

Ứng dụng trong các mức quản lý

Tác nghiệp (Operational): quyết định hằng ngày, ví dụ: sản phẩm bán chạy trong tuần.

Quản lý trung gian (Tactical): quyết định ngắn hạn, ví dụ: phân tích hiệu quả marketing theo vùng.

Chiến lược (Strategic): hoạch định dài hạn, ví dụ: dự báo xu hướng thị trường 5 năm.

Bảng so sánh nhu cầu thông tin:

Tác nghiệp → chi tiết, thời gian thực.

Trung gian → tổng hợp tuần/tháng.

Chiến lược → dữ liệu lịch sử, xu hướng dài hạn.

Cảm ơn đã tập trung lắng nghe!