

# KHÁI NIỆM CHUNG VỀ KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

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.



# Khái niệm về kho dữ liệu

- ❖ Kho dữ liệu (**data warehouse**) là nơi lưu trữ dữ liệu.
- ❖ Dữ liệu được tích hợp.
- ❖ Dữ liệu được thu thập từ nhiều nguồn:
  - ✓ Khác về không gian,
  - ✓ Khác về thời gian,
  - ✓ Khác về thể hiện và cấu trúc.
- ❖ Lưu trữ dữ liệu, thông tin, tri thức và siêu dữ liệu phục vụ cho phân tích.
- ❖ Các tổ chức có thể **chết đuối trong dữ liệu** nhưng **đói thông tin**.



# Khái niệm về kho dữ liệu

- ❖ Kho dữ liệu dung cho mục đích riêng biệt, lĩnh vực hẹp gọi là **Data Mart**.
- ❖ Một **Data warehouse** có thể hình thành nhiều **Data Mart**.
- ❖ Thuật ngữ **Data Warehousing**: Quá trình xây dựng và sử dụng một kho dữ liệu.



# Khái niệm về kho dữ liệu

- ❖ Công cụ **ETL** (**E**xtract – **T**ransform – **L**oad):
  - ✓ Rút trích (**E**xtract):
    - Rút trích thông tin từ những nguồn đã có,
    - Những phiên bản phụ thuộc thời gian của dữ liệu,
    - Chọn lựa dữ liệu.
  - ✓ Chuyển đổi (**T**ransform):
    - Chuyển đổi các định dạng khác nhau về định dạng cho trước.
  - ✓ Tải (**L**oad)
    - Sắp xếp, hợp nhất, lập chỉ mục, ... và phân hoạch.



# Các đặc tính của kho dữ liệu

- ❖ Dữ liệu **hướng chủ thể**:
  - ✓ Dữ liệu hướng theo từng nhóm đối tượng: khách hàng, bệnh nhân, sản phẩm, ...
  - ✓ Tập trung vào việc mô hình hóa và phân tích các dữ liệu cho các nhà sản xuất quyết định
  - ✓ Chuyển từ hướng ứng dụng sang hướng hỗ trợ quyết định.
  - ✓ Không dùng cho các hoạt động hàng ngày hoặc xử lý giao dịch.



# Các đặc tính của kho dữ liệu

## ❖ Tính tích hợp:

- ✓ Dữ liệu được tập hợp từ nhiều nguồn: có thể khác kiểu, khác cấu trúc, ...
- ✓ Các nguồn: cơ sở dữ liệu quan hệ, tập tin có cấu trúc, tập tin phẳng, ...
- ✓ Cần được chuẩn hóa để đảm bảo tính nhất quán trong quy ước đặt tên, ...
- ✓ Việc chuẩn hóa cần thực hiện trước khi tích hợp.



# Các đặc tính của kho dữ liệu

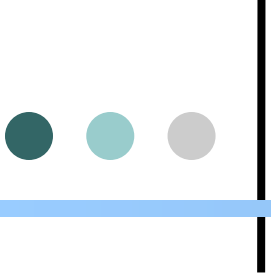
- ❖ Dữ liệu **biến thời gian**.
  - ✓ Thông tin về quá khứ, hiện tại,
  - ✓ So sánh dữ liệu theo chiều thời gian,
  - ✓ Hỗ trợ quyết định cho tương lai.
  - ✓ Thành phần thời gian có thể tương minh hoặc ngầm định.
- ❖ Dữ liệu mang tính **bền vững, chỉ đọc** (non volatile):
  - ✓ Có thể thêm vào, nhưng không thay thế,
  - ✓ Phục vụ việc nghiên cứu, phân tích





# Sự cần thiết của kho dữ liệu

- ❖ Phục vụ các phân tích dữ liệu phức tạp:
  - ✓ Phân tích định hướng,
  - ✓ Phân tích chuỗi thời gian,
  - ✓ Phân tích rủi ro.
- ❖ Hỗ trợ khám phá thông tin, tri thức ẩn.
- ❖ Hỗ trợ ra quyết định.

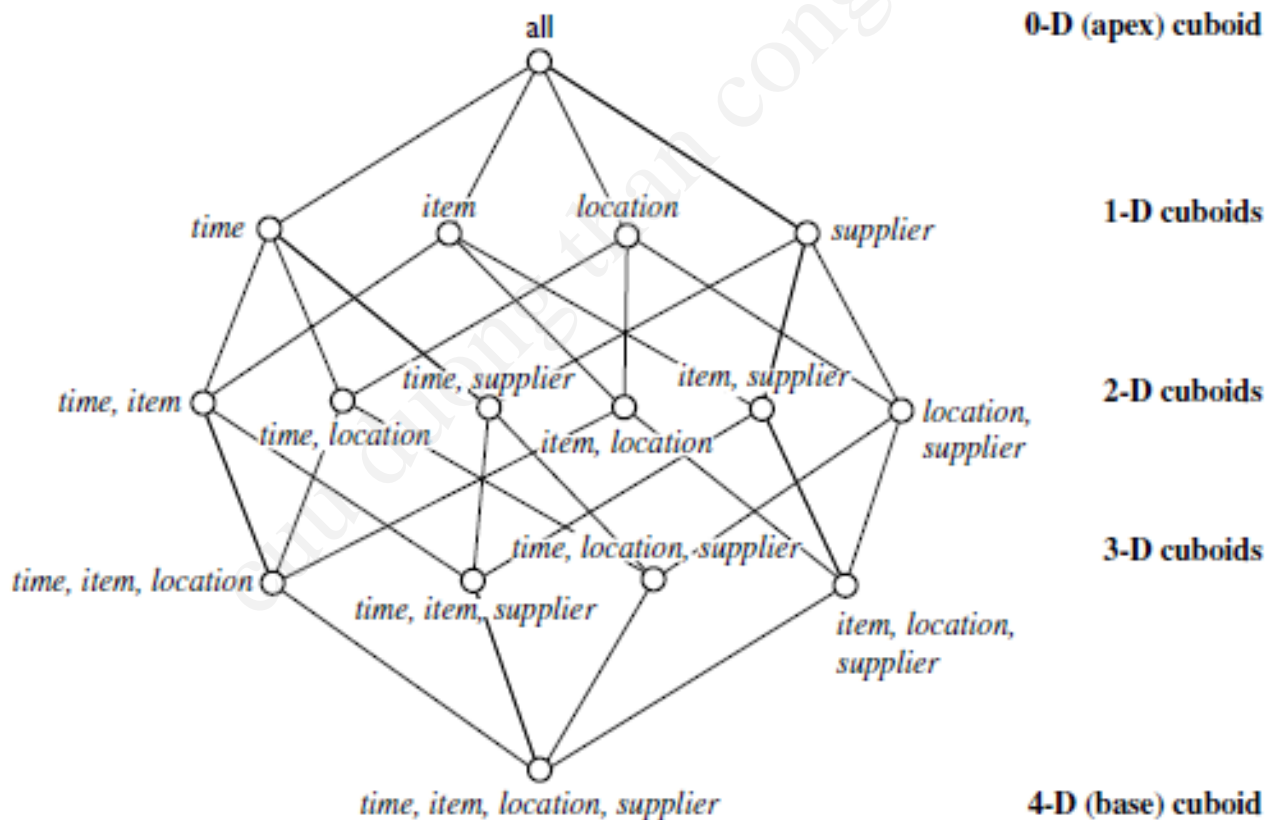


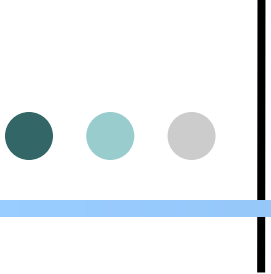
# Mô hình dữ liệu đa chiều

- ❖ Kho dữ liệu dựa trên mô hình dữ liệu đa chiều cho phép nhìn dữ liệu dưới hình thức của một khối dữ liệu
- ❖ Một khối dữ liệu cho phép dữ liệu được mô hình và được nhìn trong nhiều chiều bởi:
  - ✓ Các bản chiều (**Dimension Tables**) như Item (item\_name, brand, type); time(day, week, month).
- ❖ Một khối dữ liệu dựa trên **n-D** (n chiều) được gọi là một **cuboid** cơ sở.

# Mô hình dữ liệu đa chiều

- ❖ Cube: một lưới các cuboid



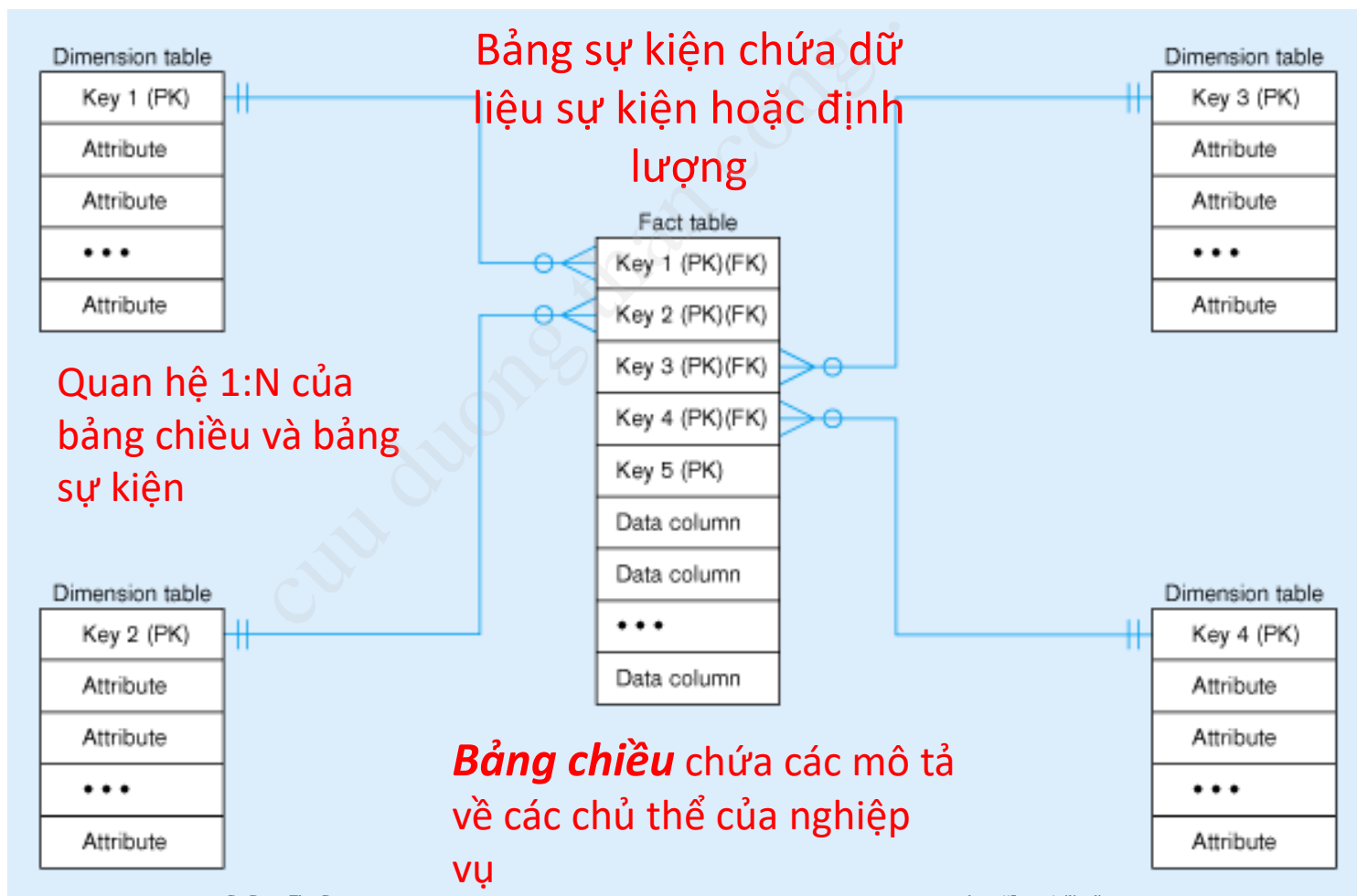


# Mô hình dữ liệu đa chiều

- ❖ Mô hình ý niệm của Kho dữ liệu
  - Lược đồ hình sao (**Star schema**): Một bảng sự kiện ở giữa nối đến một tập bảng chiều
  - Lược đồ hình bông tuyết (**Snowflake schema**): Là lược đồ tinh chế từ lược đồ hình sao (một vài chiều có sự phân cấp được chuẩn hóa thành một tập các bảng chiều nhỏ hơn).
  - Chòm sao sự kiện (**Fact constellation**): Nhiều bảng sự kiện chia sẻ các bảng chiều. Một cách gọi khác cho lược đồ này **Galaxy schema** (lược đồ thiên hà)

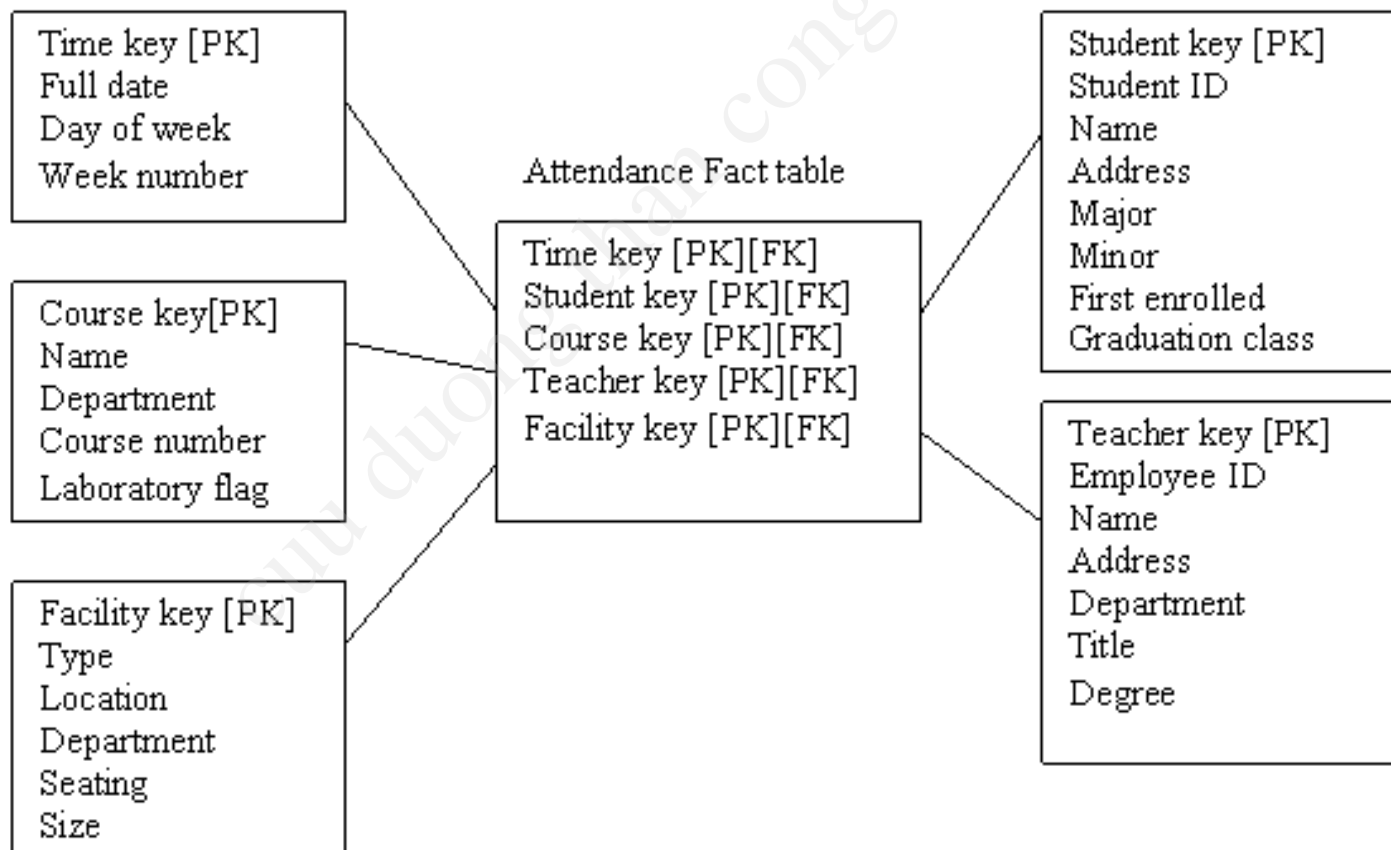
# Mô hình dữ liệu đa chiều

## ❖ Lược đồ hình sao



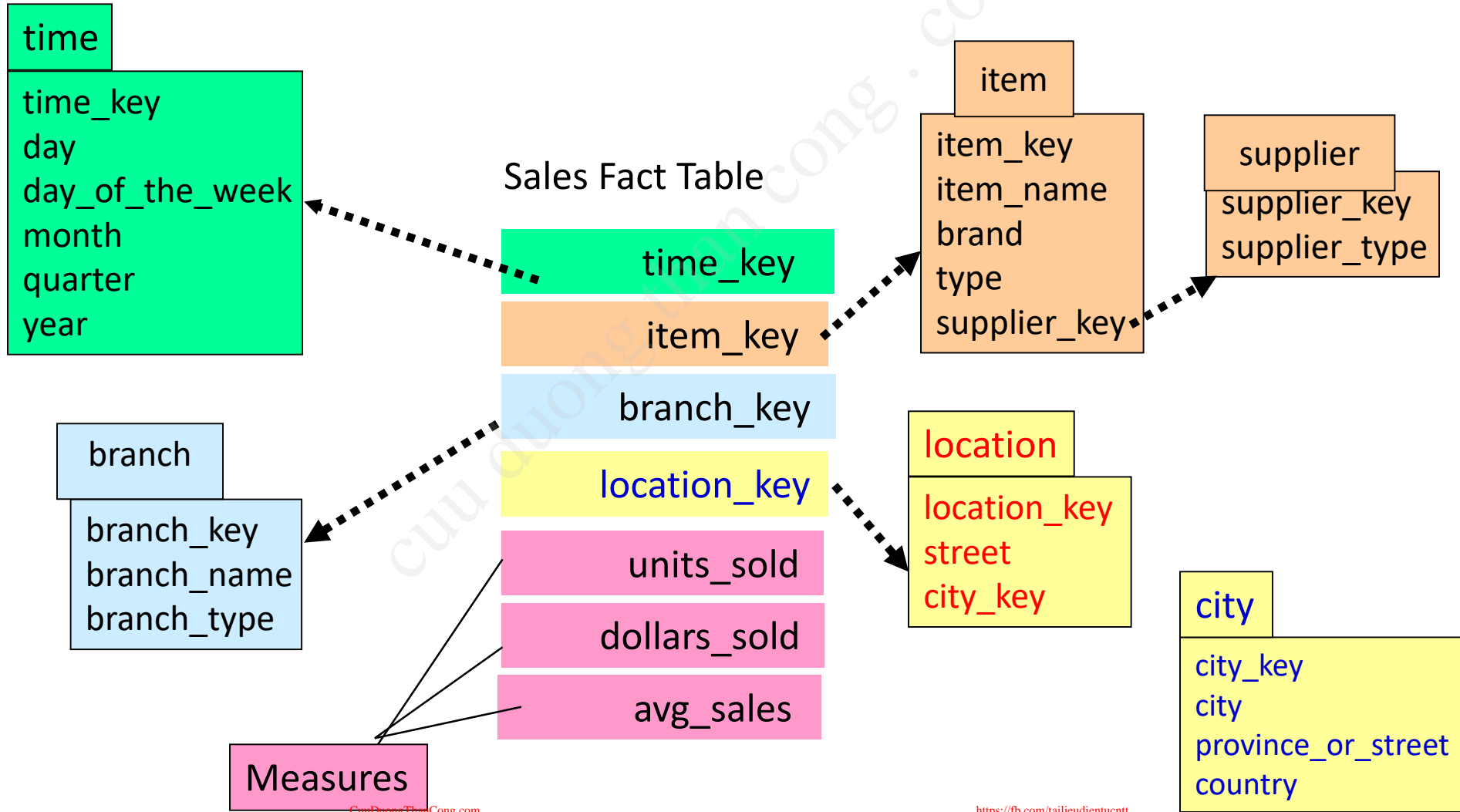
# Mô hình dữ liệu đa chiều

## ❖ Lược đồ chòm sao sự kiện



# Mô hình dữ liệu đa chiều

## ❖ Lược đồ hình bông tuyết





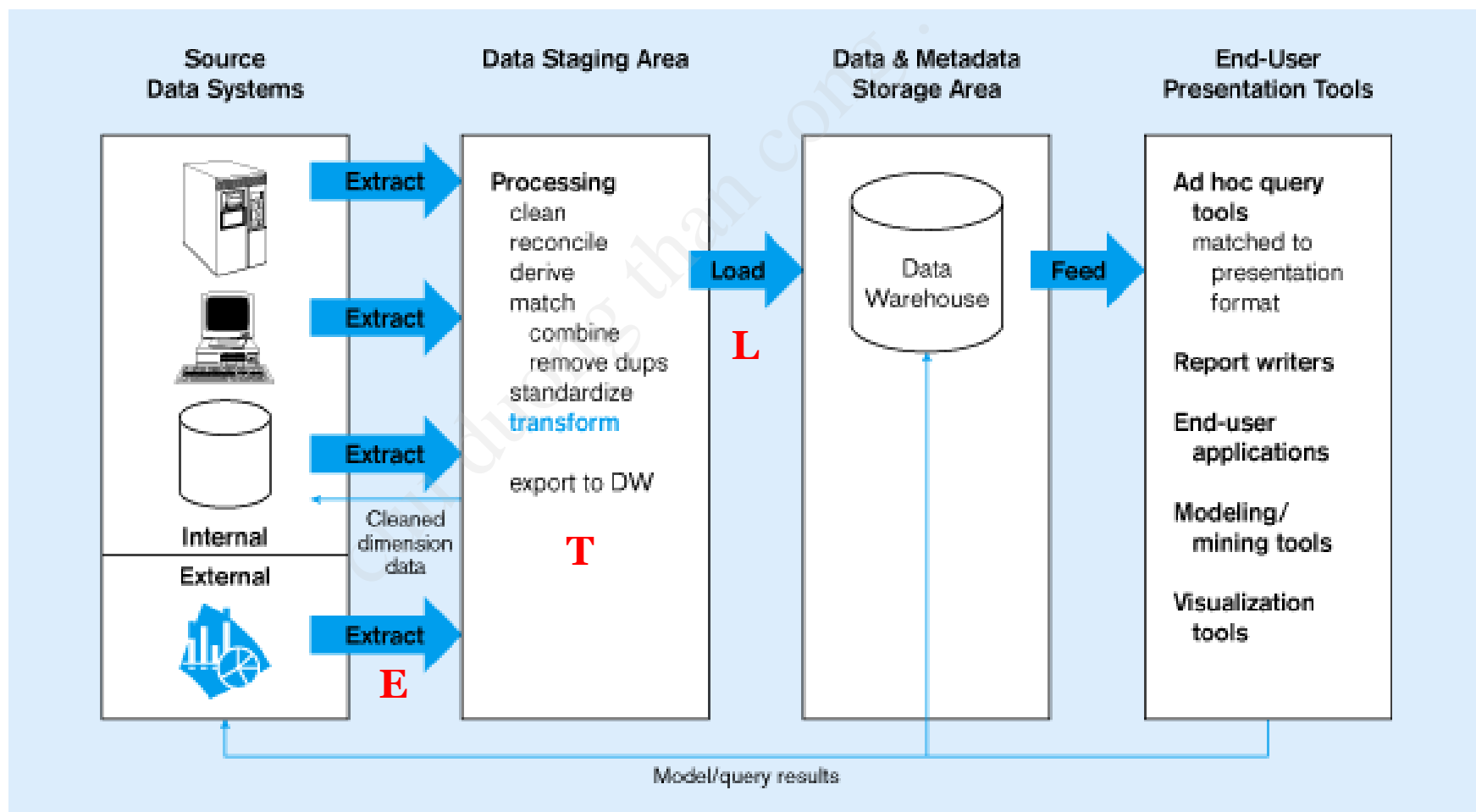
# Kiến trúc của kho dữ liệu

1. Kiến trúc 2 lớp khái quát (**Generic Two-Level Architecture**).
2. Data Mart độc lập (**Independent Data Mart**).
3. Data Mart phụ thuộc và kho lưu trữ dữ liệu tác nghiệp (**Dependent Data Mart and Operational Data Store**).
4. Data Mart luận lý và Kho dữ liệu tích cực (**Logical Data Mart and Active Warehouse**).
5. Kiến trúc dữ liệu ba lớp (**Three-Layer data architecture**)



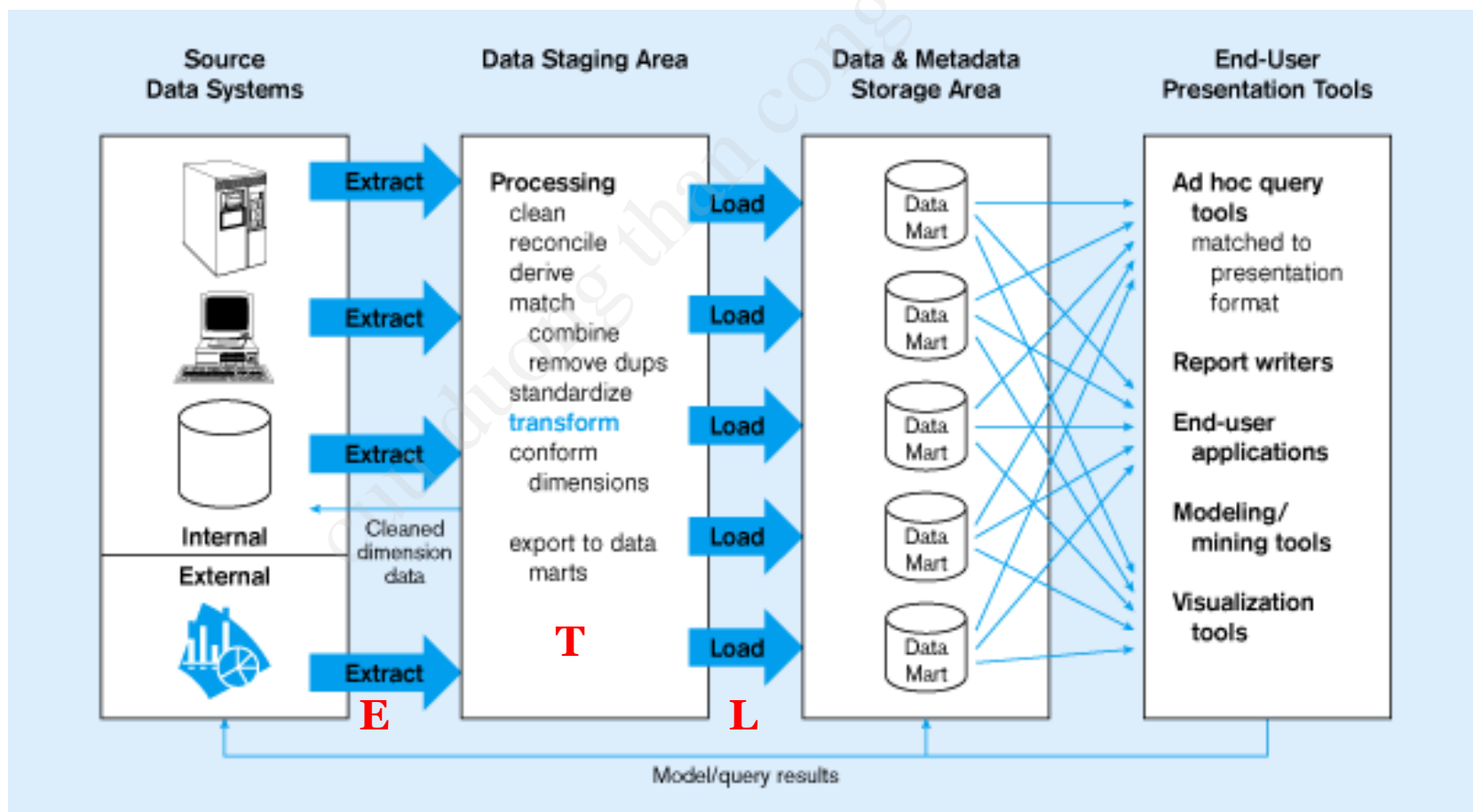
# Kiến trúc của kho dữ liệu

## 2. Kiến trúc 2 lớp khái quát :



# Kiến trúc của kho dữ liệu

## 2. Data Mart độc lập:





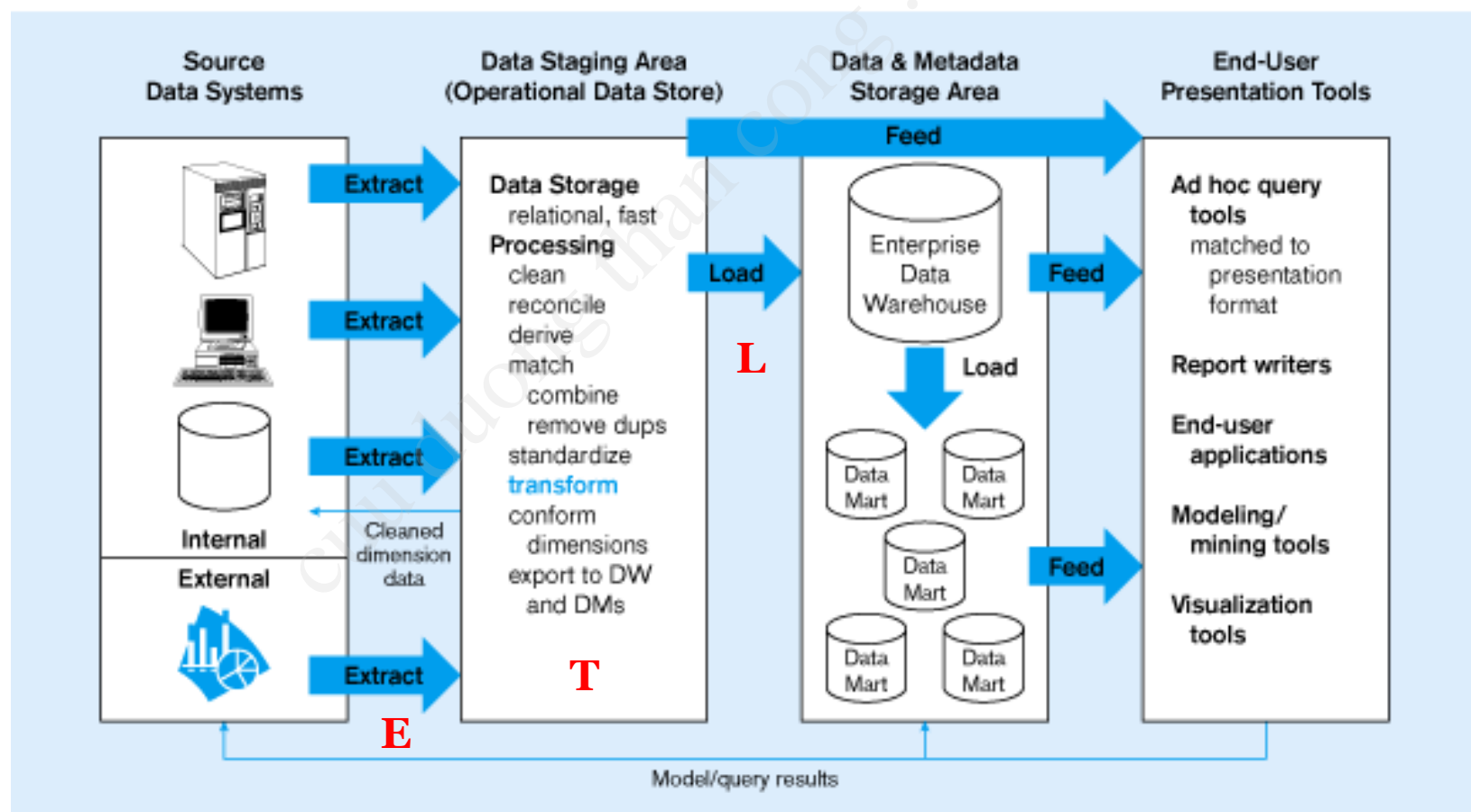
# Kiến trúc của kho dữ liệu

## 2. Data Mart độc lập:

- Dữ liệu được rút trích từ môi trường hoạt động mà không có ảnh hưởng của kho dữ liệu.
- ❖ Hạn chế của Data Mart độc lập:
  - ✓ Mỗi Data Mart độc lập cần một ETL riêng,
  - ✓ Các Data Mart không tương thích nhau,
  - ✓ Tốn nhiều chi phí để có một ứng dụng mới,
  - ✓ Tốn chi phí để làm cho các Data Mart tương thích nhau.

# Kiến trúc của kho dữ liệu

## 3. Data Mart phụ thuộc và kho lưu trữ dữ liệu tác nghiệp:





# Kiến trúc của kho dữ liệu

## 3. Data Mart phụ thuộc và kho lưu trữ dữ liệu tác nghiệp:

### ❖ Data Mart phụ thuộc:

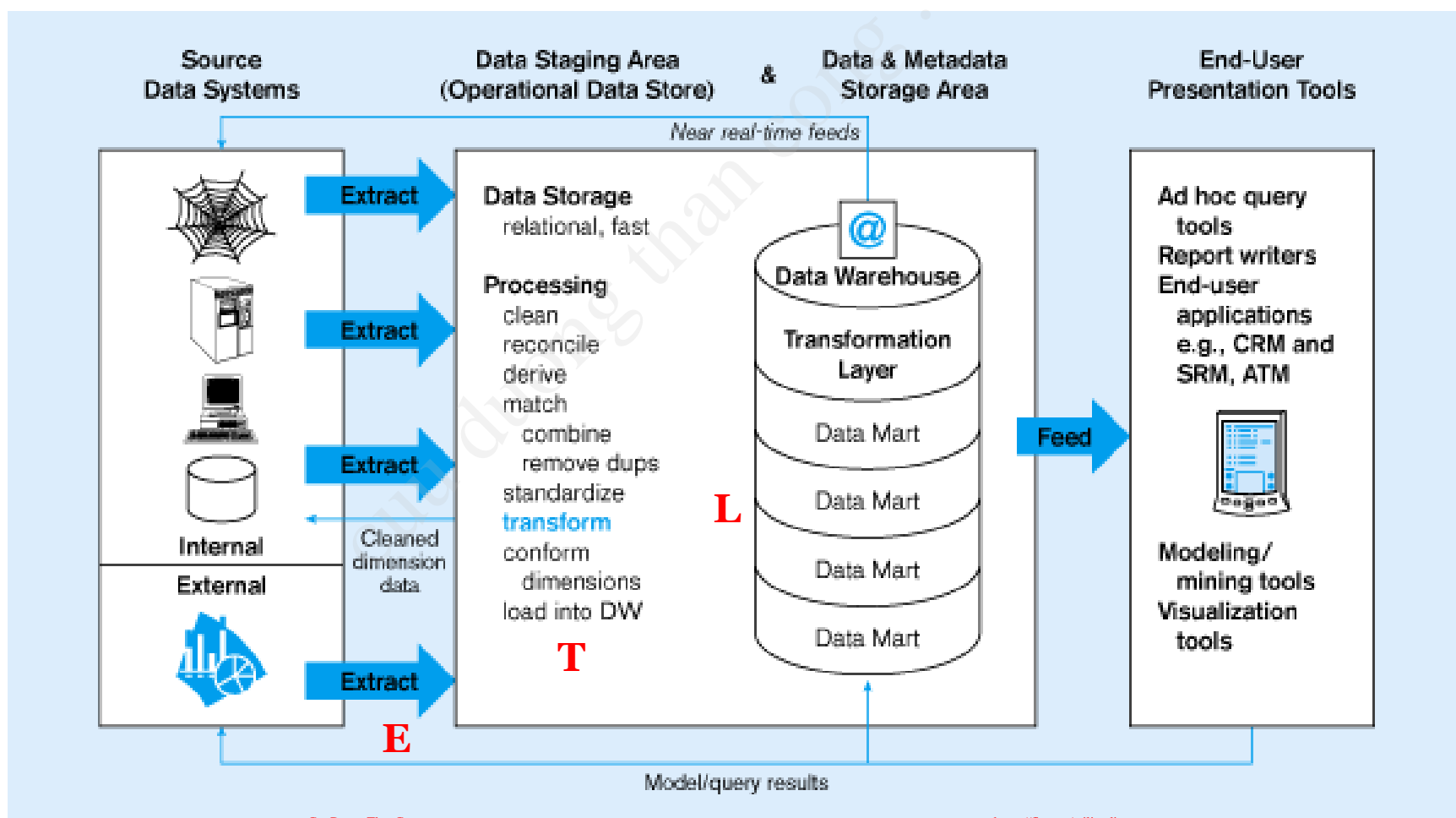
- ✓ Data Mart được nạp dữ liệu dành riêng từ kho dữ liệu doanh nghiệp.

### ❖ Kho lưu trữ dữ liệu hoạt động:

- ✓ Một cơ sở dữ liệu tích hợp hướng chủ thể, có thể cập nhật.
- ✓ Được thiết kế dành cho người dung tác nghiệp trong quá trình làm hỗ trợ quyết định.

# Kiến trúc của kho dữ liệu

## 4. Data Mart luận lý và kho lưu trữ dữ liệu tích cực:





# Kiến trúc của kho dữ liệu

## 4. Data Mart luận lý và kho lưu trữ dữ liệu tích cực:

### ❖ Data Mart luận lý:

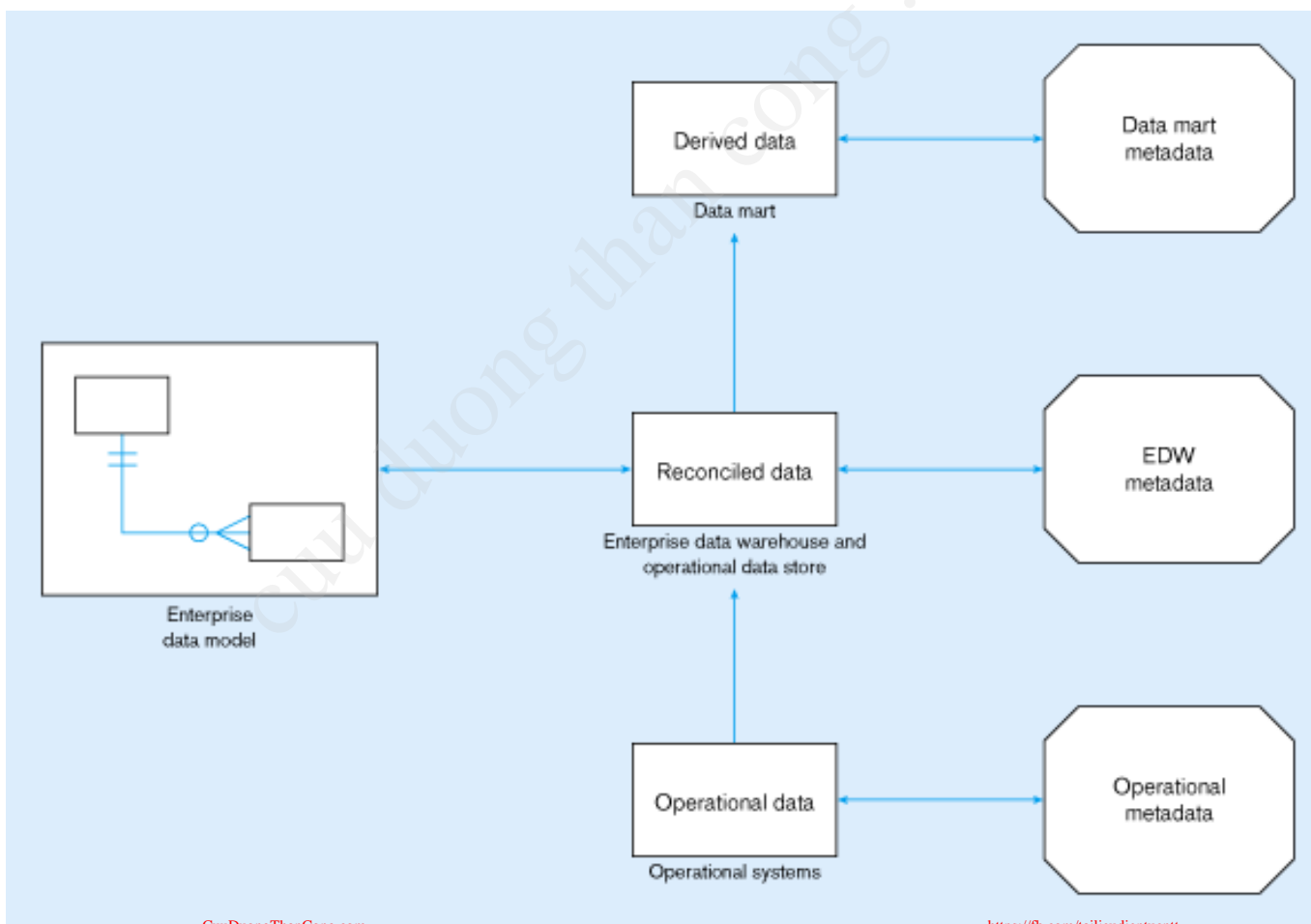
- ✓ Là Data Mart dưới góc nhìn lược đồ quan hệ.

### ❖ Kho dữ liệu tích cực:

- ✓ Chứa dữ liệu gần thời gian thực (near real time) của dữ liệu giao dịch,
- ✓ Ứng dụng dò tìm lỗi.

# Kiến trúc của kho dữ liệu

## 5. Kiến trúc dữ liệu 3 lớp:







# Kiến trúc của kho dữ liệu

## 5. Kiến trúc dữ liệu 3 lớp:

- ❖ Dữ liệu hòa hợp (**Reconcile data**):
  - ✓ Dữ liệu có tính chi tiết.
  - ✓ Dữ liệu chính thức cho tất cả ứng dụng hỗ trợ quyết định.
- ❖ Dữ liệu chuyển giao (**Derived data**):
  - ✓ Dữ liệu được chọn chuyển cho người dùng cuối trong ứng dụng hỗ trợ quyết định.
- ❖ Siêu dữ liệu (**Metadata**):
  - ✓ Dùng để đặc tả dữ liệu khác.