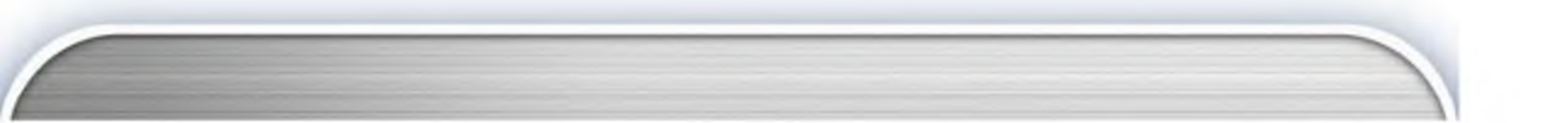




KHO DỮ LIỆU (DATA WAREHOUSE)

ThS.Nguyễn Văn Chức

NỘI DUNG

- Khái niệm về kho dữ liệu
 - Mục đích của kho dữ liệu
 - Đặc tính của kho dữ liệu
 - Kho dữ liệu cục bộ (DataMart)
 - Quy trình xây dựng kho dữ liệu
 - Mô hình kho dữ liệu
 - Quản trị kho dữ liệu
- 

Vì sao phải tìm hiểu kho dữ liệu

Các hệ thống thông tin lớn thường gặp các khó khăn khi khai thác dữ liệu:

- ✓ Dữ liệu lưu trữ phân tán ở nhiều nơi
- ✓ Dữ liệu ở nhiều định dạng khác nhau
- ✓ Không thể tìm thấy dữ liệu cần thiết
- ✓ Không thể lấy ra được dữ liệu cần thiết
- ✓ Không thể hiểu dữ liệu tìm thấy
- ✓ Không thể sử dụng được dữ liệu tìm thấy
- ✓ Yêu cầu dữ liệu ở mức cao (hỗ trợ ra quyết định)
- ✓ Khối lượng dữ liệu tăng lên nhanh chóng

Khái niệm về kho dữ liệu

Kho dữ liệu là tuyển tập các cơ sở dữ liệu tích hợp, hướng chủ đề, được thiết kế để hỗ trợ cho chức năng trợ giúp quyết định.

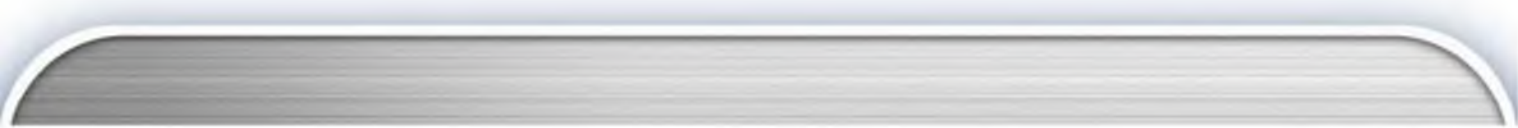
Theo John Ladley, Công nghệ kho dữ liệu (Data Warehouse Technology) là tập các phương pháp, kỹ thuật và các công cụ có thể kết hợp, hỗ trợ nhau để cung cấp thông tin cho người sử dụng trên cơ sở tích hợp từ nhiều nguồn dữ liệu, nhiều môi trường khác nhau.

Kho dữ liệu thường rất lớn tới hàng trăm GB hay thậm chí hàng Terabyte.



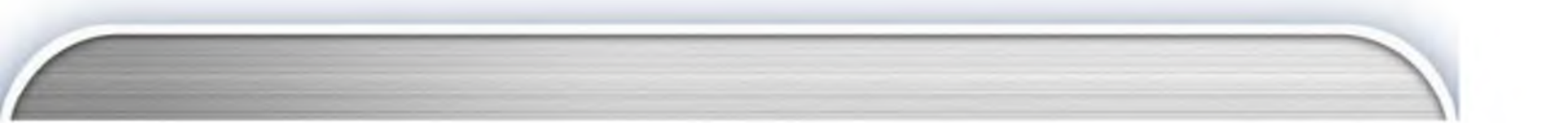
Mục đích của kho dữ liệu

Mục tiêu chính của kho dữ liệu là nhằm đáp ứng các tiêu chuẩn cơ bản sau:

- ✓ Phải có khả năng đáp ứng mọi yêu cầu về thông tin của NSD
 - ✓ Hỗ trợ để các nhân viên của tổ chức thực hiện tốt, hiệu quả công việc của mình, như có những quyết định hợp lý, nhanh và bán được nhiều hàng hơn, năng suất cao hơn, thu được lợi nhuận cao hơn, v.v.
 - ✓ Giúp cho tổ chức, xác định, quản lý và điều hành các dự án, các nghiệp vụ một cách hiệu quả và chính xác.
 - ✓ Tích hợp dữ liệu và các siêu dữ liệu từ nhiều nguồn khác nhau
- 

Đặc tính của kho dữ liệu

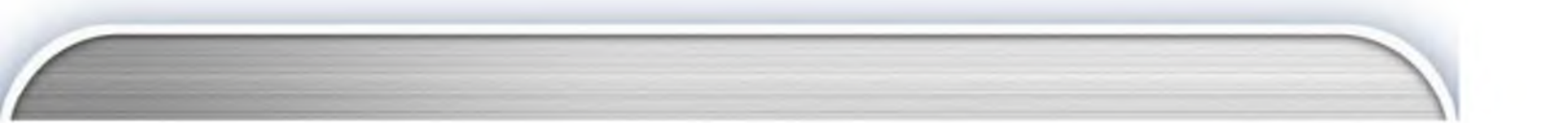
Những đặc điểm cơ bản của Kho dữ liệu (DW) là một tập hợp dữ liệu có tính chất sau:

- ✓ Tính tích hợp (Integration)
 - ✓ Hướng chủ đề
 - ✓ Dữ liệu gần thời gian và có tính lịch sử
 - ✓ Dữ liệu có tính Ổn định (nonvolatility)
 - ✓ Dữ liệu tổng hợp
- 

Kho dữ liệu cục bộ (Data Mart)

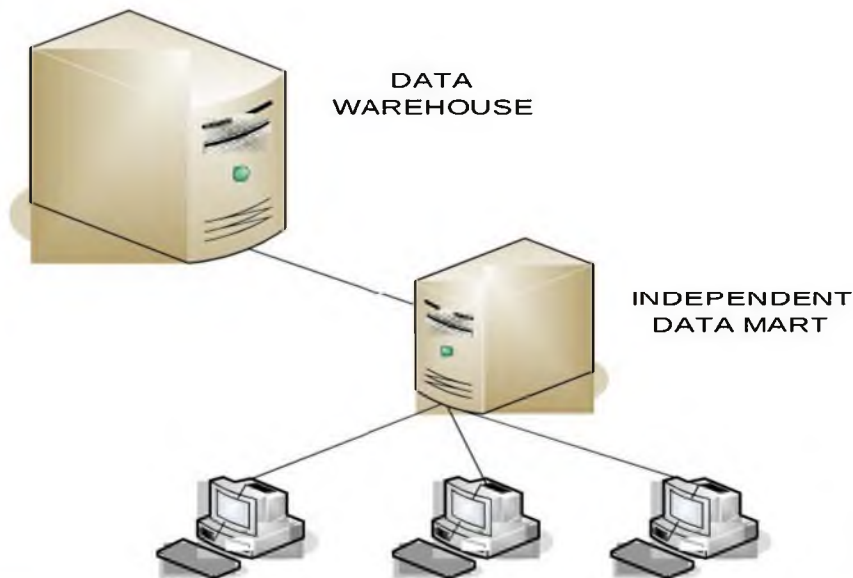
Kho dữ liệu cục bộ (Data Mart - DM) là CSDL có những đặc điểm giống với kho dữ liệu nhưng với quy mô nhỏ hơn và lưu trữ dữ liệu về một lĩnh vực, một chuyên ngành.

Datamart là kho dữ liệu hướng chủ đề. Các DM có thể được hình thành từ một tập con dữ liệu của kho dữ liệu hoặc cũng có thể được xây dựng độc lập và sau khi xây dựng xong, các DM có thể được kết nối tích hợp lại với nhau tạo thành kho dữ liệu. Vì vậy có thể xây dựng kho dữ liệu bắt đầu bằng việc xây dựng các DM hay ngược lại xây dựng kho dữ liệu trước sau đó tạo ra các DM.



Kho dữ liệu cục bộ (Data Mart)

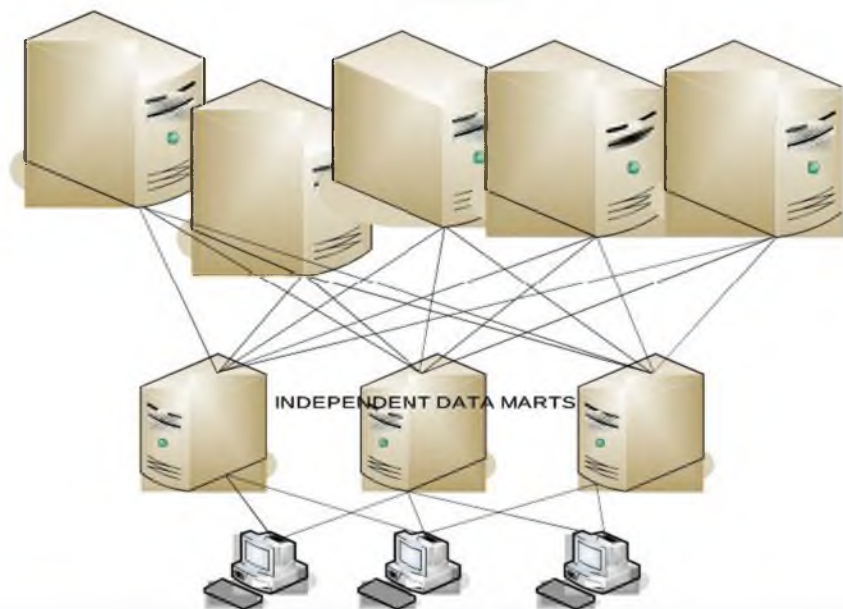
Data mart phụ thuộc (Dependent Data Mart): Chứa những dữ liệu được lấy từ DW và những dữ liệu này sẽ được trích lọc và tinh chế, tích hợp lại ở mức cao hơn để phục vụ một chủ đề nhất định của Datamart



Kho dữ liệu cục bộ (Data Mart)

Data mart độc lập (Independent Data Marts)

Không giống như Data Mart phụ thuộc, Data mart độc lập được xây dựng trước DW và dữ liệu được trực tiếp lấy từ các nguồn khác nhau



Cơ sở dữ liệu phân tán (Distributed Database)

Định nghĩa cơ sở dữ liệu phân tán:

Một cơ sở dữ liệu Phân tán là sự tập hợp dữ liệu phân tán về mặt luận lý chúng cùng một hệ thống nhưng được trải rộng ở nhiều nơi (site) của một mạng máy tính [5].

Định nghĩa này nhấn mạnh hai khía cạnh quan trọng như nhau của một CSDL Phân tán là:

- ✓ Sự Phân tán (Distribution) dữ liệu trên các nơi (site)
- ✓ Sự tương quan luận lý (Logical Correlation)

Tại sao phải sử dụng CSDL phân tán?



Có nhiều lý do tại sao phát triển CSDL Phân tán:

- ✓ Các lý do về tổ chức (organizational) và kinh tế (economic)
- ✓ Kết nối lẫn nhau (interconnection) của các CSDL hiện tại
- ✓ Sự lớn mạnh gia tăng (incremental growth)
- ✓ Giảm chi phí truyền thông (communication overhead)
- ✓ Các nghiên cứu về hiệu suất (performance consideration)
- ✓ Độ tin cậy (reliability) và tính sẵn sàng (availability)



Phân mảnh ngang (*Horizontal Fragmentation*)

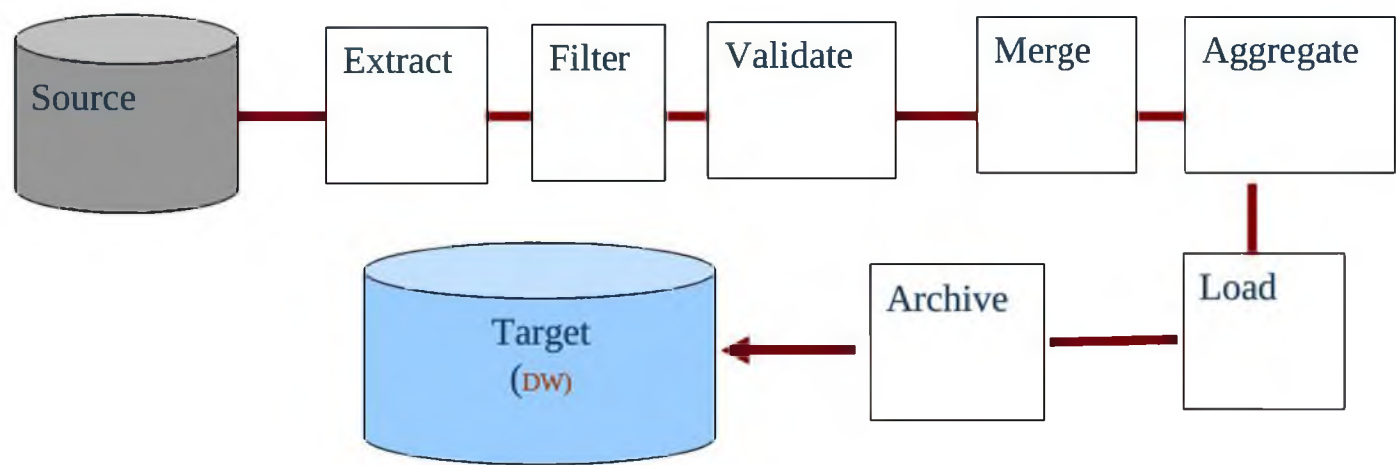
$$\forall u \in R, \exists i \in [1, n] : u \in R_i$$

Phân mảnh dọc (*Vertical Fragmentation*)

$$\forall A \in \text{Attr}(R), \exists i \in [1, n] : A \in \text{Attr}(R_i)$$

Với $\text{Attr}(R)$ là tập thuộc tính của R

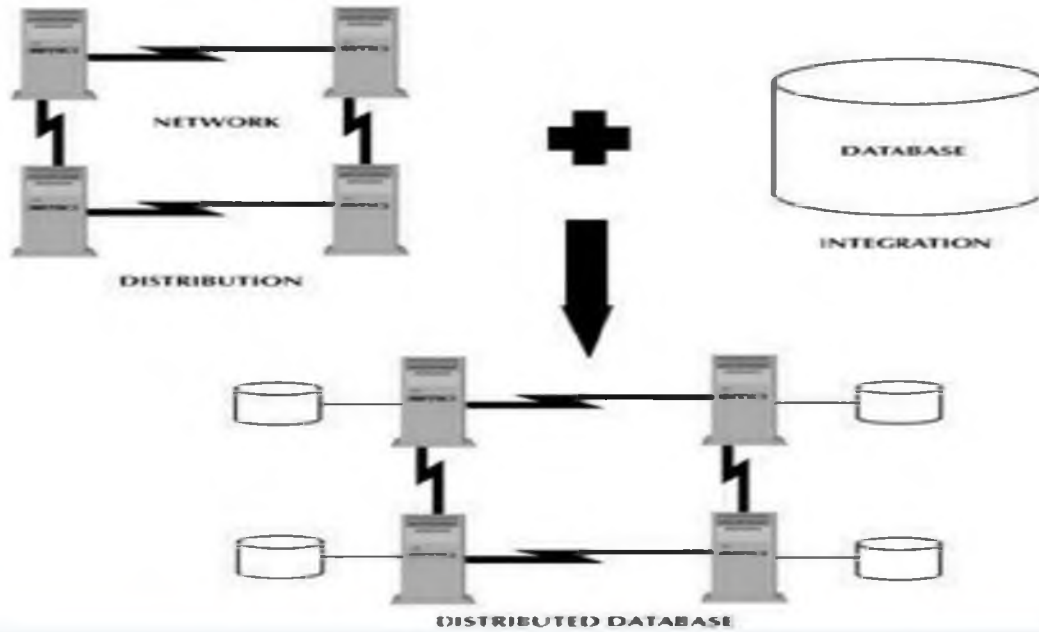
Quy trình xây dựng kho dữ liệu



Quá trình tạo lập kho dữ liệu

Mô hình kho dữ liệu

Kiến trúc kho dữ liệu phân tán bao gồm sự kết hợp của hai khái niệm cơ bản là sự tích hợp(Integration) các thành phần dữ liệu và sự phân tán (Distribution) thông qua các thành phần của mạng như hình sau:

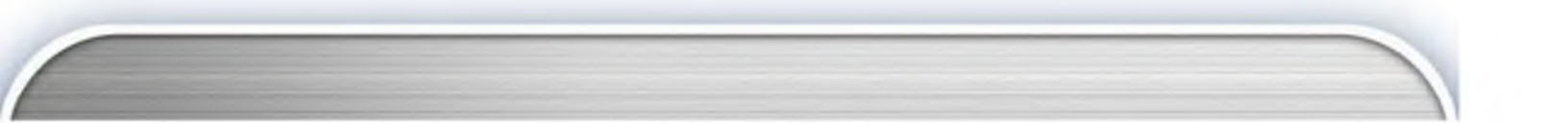


Mô hình kho dữ liệu

Kho dữ liệu phân tán có hai kiến trúc chính là kho dữ liệu phân tán thuần nhất và kho dữ liệu phân tán không thuần nhất.

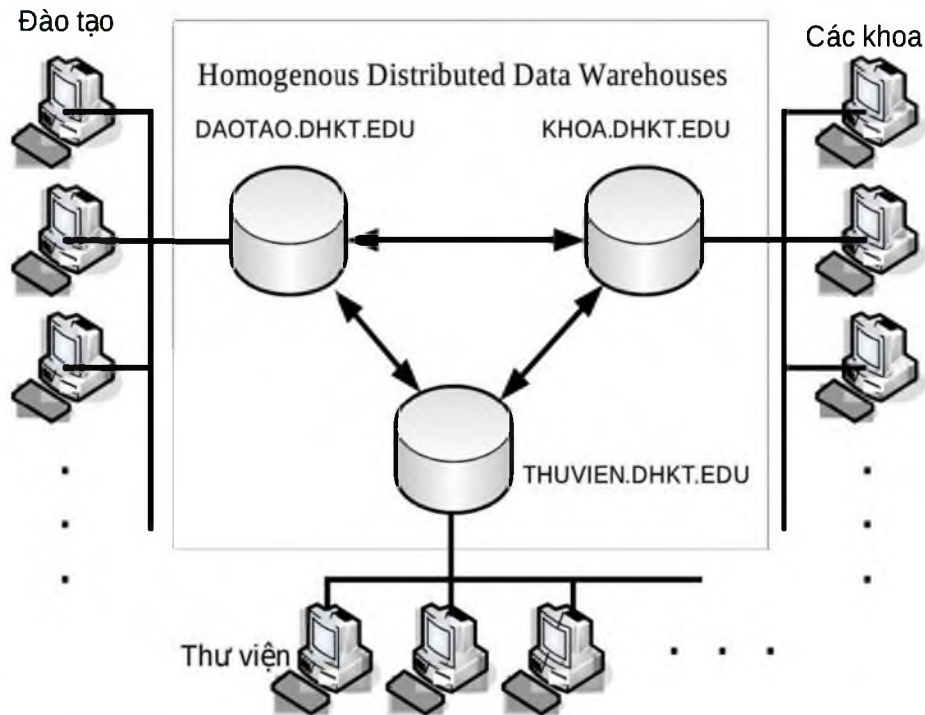
Kho dữ liệu phân tán thuần nhất (Homogenous distributed data warehouses)

Kho dữ liệu phân tán thuần nhất là kho dữ liệu mà trong đó tất cả các kho dữ liệu cục bộ (DM) ở các nơi (Site) đều phải dùng chung một hệ quản trị CSDL.



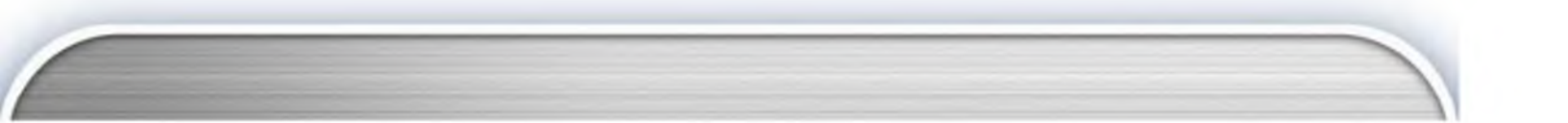
Mô hình kho dữ liệu

Kho dữ liệu phân tán thuần nhất (Homogenous distributed data warehouses)



Mô hình kho dữ liệu

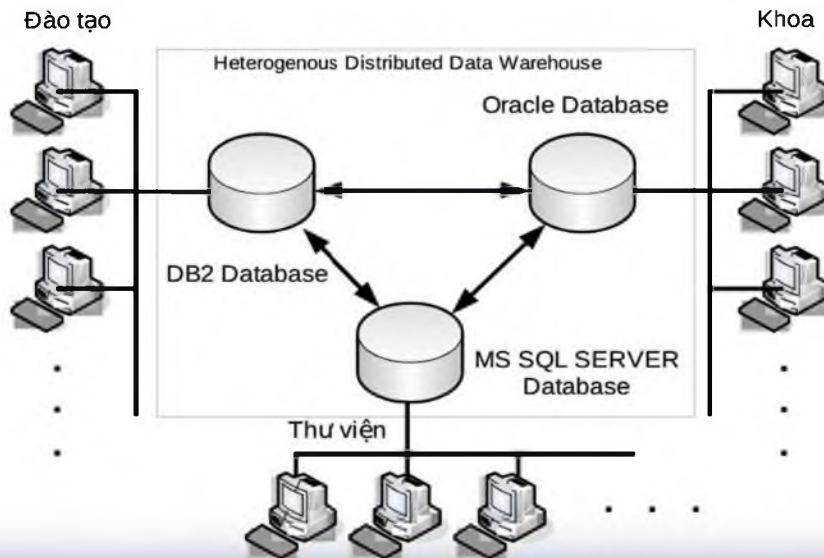
Kiến trúc phân tán thuần nhất có một số ưu điểm sau:

- Do tất cả các DM đều dùng chung DBMS nên công tác quản trị dễ dàng hơn. Người quản trị không cần biết kỹ năng quản trị trong tất cả các DBMS khác nhau như DB2, SQL SERVER,...
 - Công tác chuyển đổi dữ liệu không đòi hỏi cao vì tất cả các DM dùng chung cấu trúc dữ liệu và các ràng buộc dữ liệu.
 - Nhiệm vụ tích hợp dữ liệu từ các nguồn trở nên đơn giản và dễ quản lý
 - Thời gian đáp ứng các truy vấn nhanh (rapid response times)
 - Tuy nhiên, kho dữ liệu phân tán thuần nhất thích hợp nhất đối với những hệ thống xây dựng mới và có chiến lược từ trước, đối với các hệ thống kế thừa dữ liệu từ các nguồn đã có công việc chuyển đổi và tích hợp dữ
- 

Mô hình kho dữ liệu

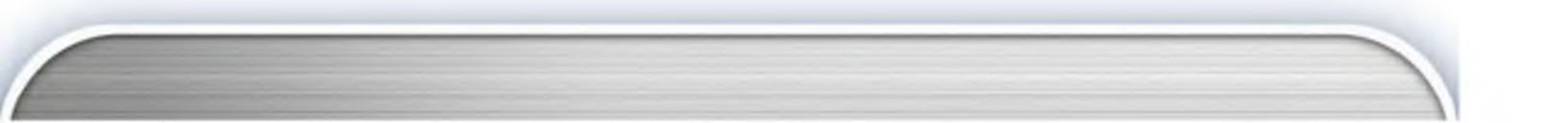
Kho dữ liệu phân tán không thuần nhất (Heterogenous distributed data warehouses)

Kho dữ liệu phân tán không thuần nhất là kho dữ liệu mà trong đó các kho dữ liệu cục bộ (DM) ở các nơi (Site) trong mạng có thể không cùng chung hệ quản trị CSDL [11].



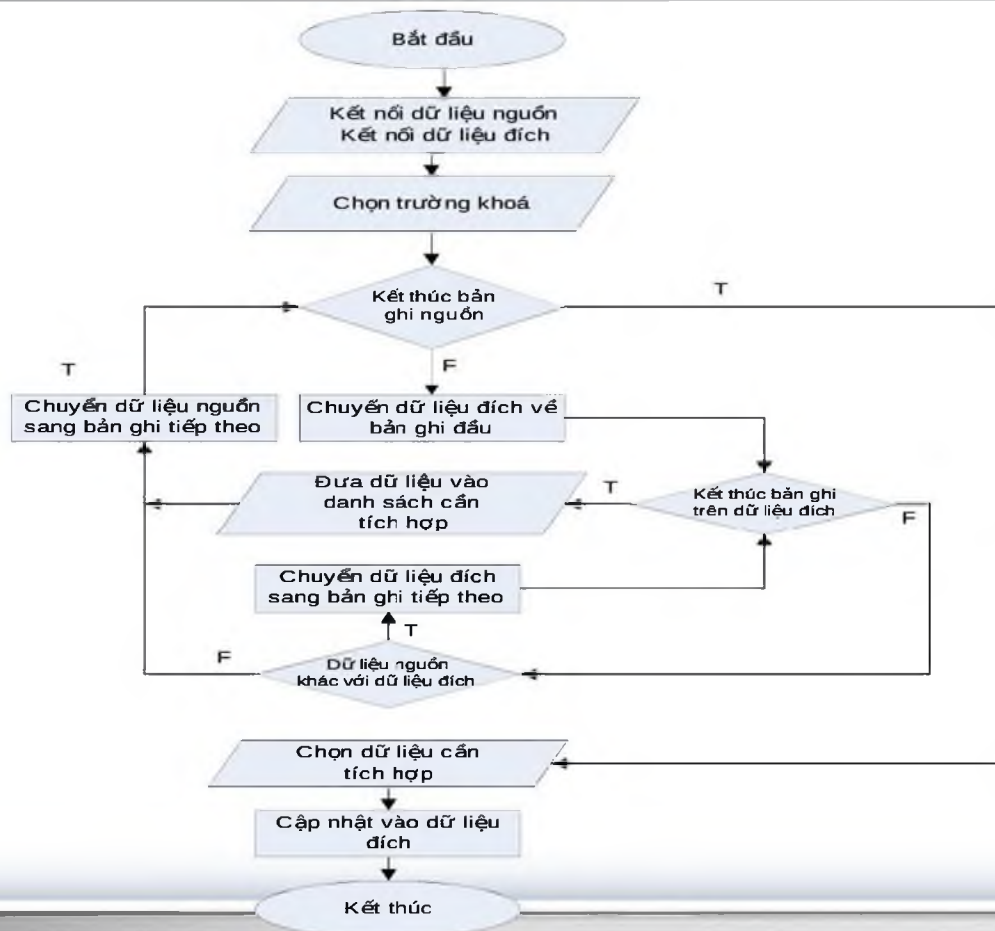
Mô hình kho dữ liệu

Kiến trúc phân tán không thuần nhất có một số ưu điểm sau:

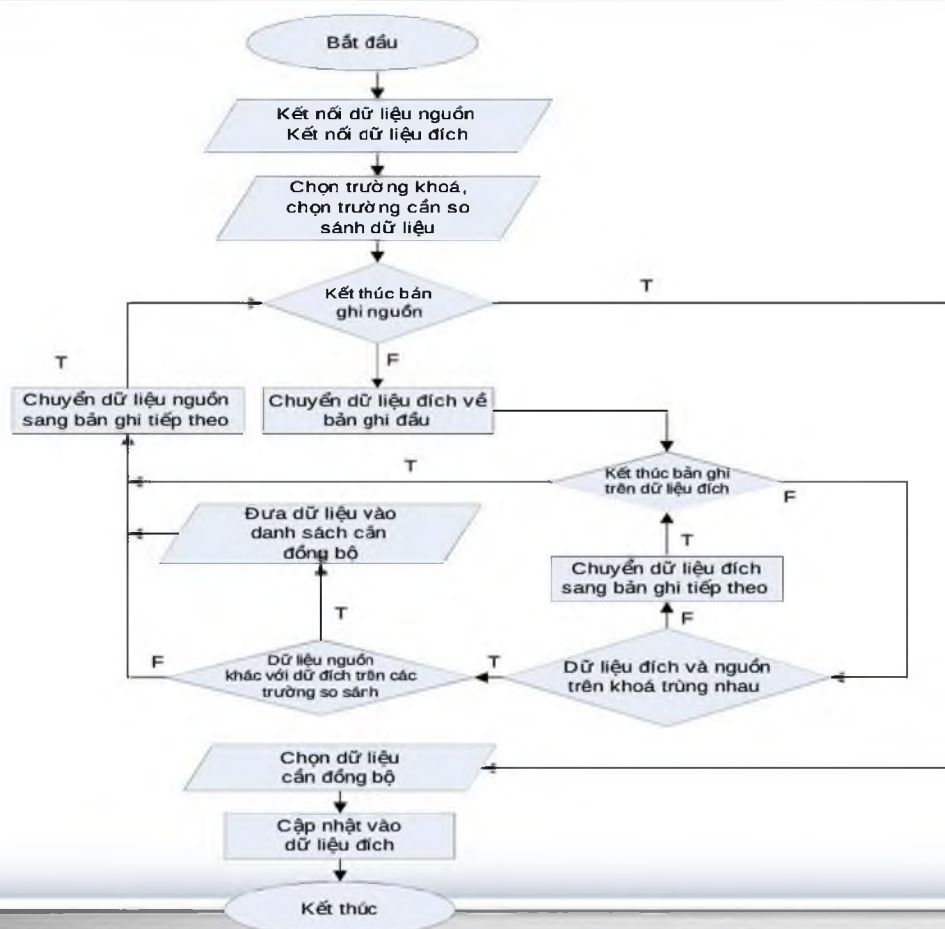
- Kế thừa được các nguồn dữ liệu từ các DM đã tồn tại
 - Thích hợp cho các hệ thống xây dựng trên cơ sở mở rộng hệ thống đã có vì trên thực tế các đơn vị thường bắt đầu với các DM nhỏ cho các phòng ban, sau đó phát triển thành kho dữ liệu lớn hơn cho toàn công ty.
 - Tính tự trị CSDL cao
 - Tuy nhiên, hệ thống phân tán không thuần nhất gặp khó khăn trong việc tích hợp, chuyển đổi dữ liệu cũng như công tác quản trị dữ liệu vì mỗi DBMS có cấu trúc dữ liệu, ràng buộc, cách thức truy vấn, bảo mật dữ liệu khác nhau.
- 

Quản trị kho dữ liệu

- ✓ **Chuyển đổi dữ liệu:** Chuyển đổi dữ liệu giữa các định dạng MS Excel, MS Access, SQL SERVER, XML, Oracle
 - ✓ **Tích hợp dữ liệu:** Trao đổi dữ liệu giữa các Data Mart
 - ✓ **Đồng bộ dữ liệu:** So sánh, làm sạch dữ liệu để dữ liệu giữa các Data Mart thống nhất với nhau
 - ✓ **Phân tán dữ liệu:** Phân tán ngang, phân tán dọc
 - ✓ **Hợp nhất dữ liệu:** Hợp nhất dữ liệu sau khi đã phân tán dọc
 - ✓ **Lọc dữ liệu:** Trích xuất dữ liệu theo điều kiện
-



Thuật toán đồng bộ dữ liệu giữa các DataMart



Thuật toán phân tán dữ liệu giữa các Data Mart

