**Các khái niệm về kho dữ liệu**

Khái niệm cơ bản về Kho dữ liệu là tạo điều kiện thuận lợi cho một phiên bản sự thật duy nhất cho một công ty để ra quyết định và dự báo. Kho dữ liệu là một hệ thống thông tin chứa dữ liệu lịch sử và dữ liệu giao hoán từ một hoặc nhiều nguồn. Các khái niệm về kho dữ liệu

đơn giản hóa quy trình báo cáo và phân tích của các tổ chức.

Trong hướng dẫn này, bạn sẽ học:

Table of Contents

[II. Đặc điểm của Kho dữ liệu 1](#_Toc82510757)

[A. Hướng chủ thể (Subject-Oriented) 2](#_Toc82510758)

[B. Tích hợp 2](#_Toc82510759)

[C. Biến thể thời gian (Time-variant) 4](#_Toc82510760)

[D. Không bay hơi (non-volatile) 4](#_Toc82510761)

[III. Kiến trúc Kho dữ liệu 5](#_Toc82510762)

[Kiến trúc một tầng 6](#_Toc82510763)

[Kiến trúc hai tầng 6](#_Toc82510764)

[Kiến trúc kho dữ liệu ba tầng 6](#_Toc82510765)

[IV. Các thành phần Datawarehouse 7](#_Toc82510766)

[Cơ sở dữ liệu Kho dữ liệu 8](#_Toc82510767)

[Công cụ tìm nguồn cung ứng, mua lại, làm sạch và chuyển đổi (ETL) 9](#_Toc82510768)

[Metadata 10](#_Toc82510769)

[Công cụ truy vấn 11](#_Toc82510770)

[**1. Công cụ truy vấn và báo cáo:** 12](#_Toc82510771)

[**2. Các công cụ phát triển ứng dụng:** 13](#_Toc82510772)

[**3. Các công cụ khai thác dữ liệu:** 13](#_Toc82510773)

[**4. Các công cụ OLAP:** 13](#_Toc82510774)

[Kho dữ liệu Kiến trúc Bus 13](#_Toc82510775)

[V. Data Marts 14](#_Toc82510776)

[VI. Các phương pháp hay nhất về kiến ​​trúc kho dữ liệu 14](#_Toc82510777)

[VII. Tóm lược: 15](#_Toc82510778)

# **Đặc điểm của Kho dữ liệu**

Khái niệm Kho dữ liệu có các đặc điểm sau:

* Hướng chủ đề
* Tích hợp
* Biến thể thời gian
* Không bay hơi

## **Hướng chủ thể (Subject-Oriented)**

Kho dữ liệu được định hướng theo chủ thể vì nó cung cấp thông tin về một chủ thể thay vì các hoạt động đang diễn ra của công ty. Những chủ thể này có thể là bán hàng, tiếp thị, phân phối, v.v.

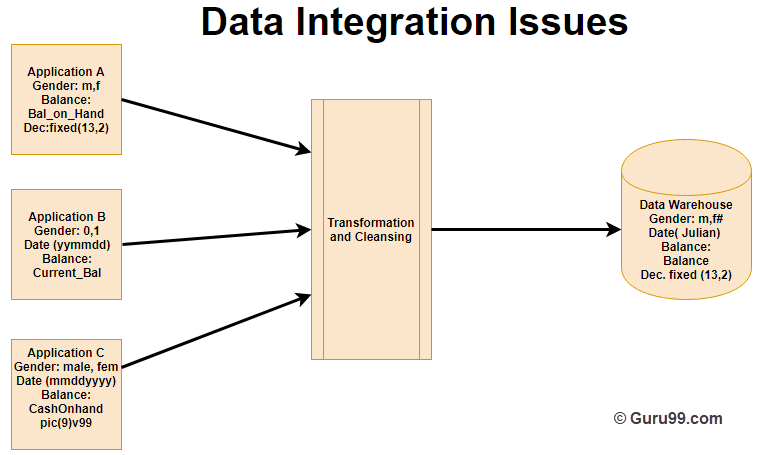
Một kho dữ liệu không bao giờ tập trung vào các hoạt động đang diễn ra. Thay vào đó, nó tập trung vào mô hình hóa và phân tích dữ liệu để ra quyết định. Nó cũng cung cấp một cái nhìn đơn giản và ngắn gọn về chủ thể cụ thể bằng cách loại trừ dữ liệu không hữu ích để hỗ trợ quá trình quyết định.

## **Tích hợp**

Trong Data Warehouse, tích hợp có nghĩa là thiết lập một đơn vị đo lường chung cho tất cả các dữ liệu tương tự từ cơ sở dữ liệu khác nhau. Dữ liệu cũng cần được lưu trữ trong Datawarehouse theo cách phổ biến và được mọi người chấp nhận.

Kho dữ liệu được phát triển bằng cách tích hợp dữ liệu từ các nguồn khác nhau như máy tính lớn, cơ sở dữ liệu quan hệ, tệp phẳng, v.v. Hơn nữa, nó phải giữ các quy ước đặt tên, định dạng và mã hóa nhất quán.

Sự tích hợp này giúp phân tích dữ liệu một cách hiệu quả. Phải đảm bảo tính nhất quán trong các quy ước đặt tên, các thước đo thuộc tính, cấu trúc mã hóa, v.v. Hãy xem xét ví dụ sau:



Trong ví dụ trên, có ba ứng dụng khác nhau có nhãn A, B và C. Thông tin được lưu trữ trong các ứng dụng này là Giới tính, Ngày tháng và Số dư. Tuy nhiên, dữ liệu của mỗi ứng dụng được lưu trữ theo cách khác nhau.

* Trong ứng dụng Một trường giới tính lưu trữ các giá trị logic như M hoặc F
* Trong Ứng dụng B, trường giới tính là một giá trị số,
* Trong ứng dụng Application C, trường giới tính được lưu trữ dưới dạng giá trị ký tự.
* Tương tự là trường hợp với Ngày và số dư

Tuy nhiên, sau quá trình chuyển đổi và làm sạch, tất cả dữ liệu này được lưu trữ ở định dạng chung trong Kho dữ liệu.

## **Biến thể thời gian (Time-variant)**

Khoảng thời gian cho kho dữ liệu khá rộng so với các hệ thống hoạt động. Dữ liệu được thu thập trong kho dữ liệu được ghi nhận với một thời kỳ cụ thể và cung cấp thông tin theo quan điểm lịch sử. Nó chứa một yếu tố thời gian, rõ ràng hoặc ẩn ý.

Một trong những nơi mà phương sai thời gian hiển thị dữ liệu Datawarehouse nằm trong cấu trúc của khóa bản ghi. Mọi khóa chính chứa trong DW phải có một phần tử thời gian ngầm định hoặc rõ ràng. Như ngày, tháng trong tuần, v.v.

Một khía cạnh khác của phương sai thời gian là khi dữ liệu được đưa vào kho, dữ liệu đó không thể được cập nhật hoặc thay đổi.

## **Không bay hơi (non-volatile)**

Kho dữ liệu cũng không thay đổi có nghĩa là dữ liệu trước đó không bị xóa khi dữ liệu mới được nhập vào đó.

Dữ liệu ở chế độ chỉ đọc và được làm mới định kỳ. Điều này cũng giúp phân tích dữ liệu lịch sử và hiểu điều gì & thời điểm đã xảy ra. Nó không yêu cầu quy trình giao dịch, cơ chế phục hồi và kiểm soát đồng thời.

Các hoạt động như xóa, cập nhật và chèn được thực hiện trong môi trường ứng dụng hoạt động được bỏ qua trong môi trường Kho dữ liệu. Chỉ có hai loại hoạt động dữ liệu được thực hiện trong Kho dữ liệu là

* Đang tải dữ liệu
* Truy cập dữ liệu

Dưới đây là một số khác biệt chính giữa Ứng dụng và Kho dữ liệu

|  |  |
| --- | --- |
| **Operational Application** | **Data Warehouse** |
| Chương trình phức tạp phải được mã hóa để đảm bảo rằng các quá trình nâng cấp dữ liệu duy trì tính toàn vẹn cao của sản phẩm cuối cùng. | Loại sự cố này không xảy ra vì cập nhật dữ liệu không được thực hiện. |
| Dữ liệu được đặt ở dạng chuẩn hóa để đảm bảo dư thừa tối thiểu. | Dữ liệu không được lưu trữ ở dạng chuẩn hóa. |
| Công nghệ cần thiết để hỗ trợ các vấn đề về giao dịch, khôi phục dữ liệu, khôi phục dữ liệu và giải quyết vì bế tắc của nó khá phức tạp. | Nó cung cấp sự đơn giản tương đối trong công nghệ. |

# **Kiến trúc Kho dữ liệu**

Kiến trúc kho dữ liệu rất phức tạp vì nó là một hệ thống thông tin chứa dữ liệu lịch sử và dữ liệu giao hoán từ nhiều nguồn. Có 3 cách tiếp cận để xây dựng các lớp Kho dữ liệu: Một lớp, Hai lớp và Ba lớp. Kiến trúc 3 tầng này của Data Warehouse được giải thích như bên dưới.

**Kiến trúc một tầng**

Mục tiêu của một lớp là giảm thiểu lượng dữ liệu được lưu trữ. Mục tiêu này là để loại bỏ dư thừa dữ liệu. Kiến trúc này không được sử dụng thường xuyên trong thực tế.

**Kiến trúc hai tầng**

Kiến trúc hai lớp là một trong các lớp Kho dữ liệu phân tách các nguồn vật lý có sẵn và kho dữ liệu. Kiến trúc này không thể mở rộng và cũng không hỗ trợ một số lượng lớn người dùng cuối. Nó cũng có vấn đề kết nối vì giới hạn mạng.

**Kiến trúc kho dữ liệu ba tầng**

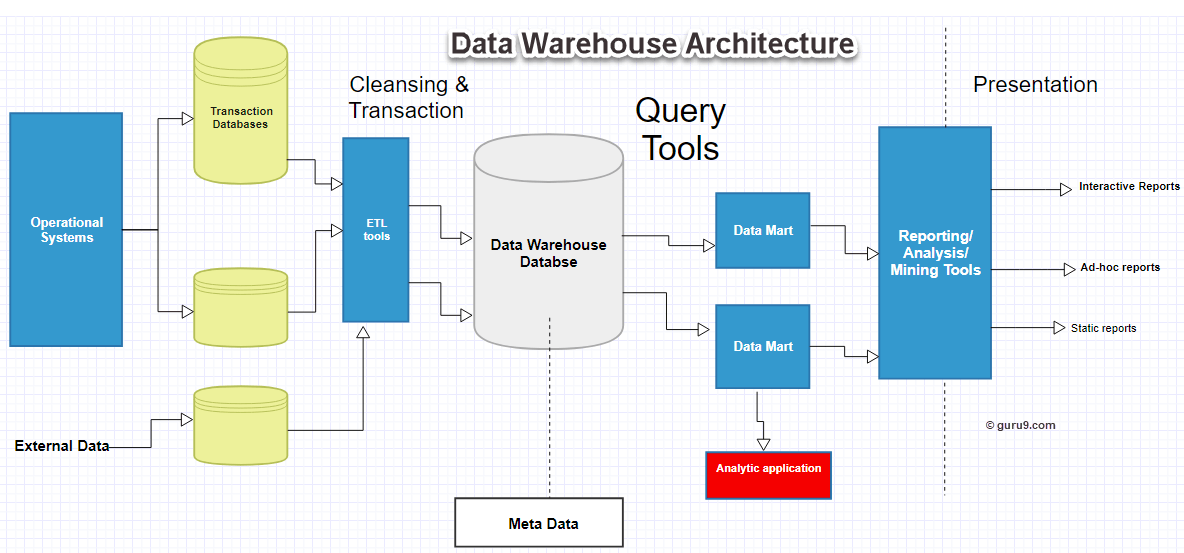
Đây là Kiến trúc Kho dữ liệu được sử dụng rộng rãi nhất.

Nó bao gồm Tầng trên cùng, Tầng giữa và Tầng dưới cùng.

1. Bottom Tier: Cơ sở dữ liệu của các máy chủ Datawarehouse ở tầng dưới cùng. Nó thường là một hệ thống cơ sở dữ liệu quan hệ. Dữ liệu được làm sạch, chuyển đổi và tải vào lớp này bằng các công cụ back-end.
2. Tầng giữa: Tầng giữa trong Kho dữ liệu là một máy chủ OLAP được triển khai bằng cách sử dụng mô hình ROLAP hoặc MOLAP. Đối với người dùng, tầng ứng dụng này trình bày một dạng xem trừu tượng của cơ sở dữ liệu. Lớp này cũng hoạt động như một trung gian giữa người dùng cuối và cơ sở dữ liệu.
3. Lớp trên cùng: Lớp trên cùng là lớp khách hàng giao diện người dùng. Cấp cao nhất là các công cụ và API mà bạn kết nối và lấy dữ liệu ra từ kho dữ liệu. Nó có thể là công cụ truy vấn, công cụ báo cáo, công cụ truy vấn được quản lý, công cụ phân tích và công cụ khai thác dữ liệu.

# **Các thành phần Datawarehouse**

Chúng ta sẽ tìm hiểu về các Thành phần Datawarehouse và Kiến trúc của Data Warehouse với Sơ đồ như hình dưới đây:



Data Warehouse dựa trên một máy chủ RDBMS là một kho thông tin trung tâm được bao quanh bởi một số thành phần Data Warehouse chính để làm cho toàn bộ môi trường hoạt động, có thể quản lý và có thể truy cập được.

Chủ yếu có năm Thành phần Kho dữ liệu:

**Cơ sở dữ liệu Kho dữ liệu**

Cơ sở dữ liệu trung tâm là nền tảng của môi trường kho dữ liệu. Cơ sở dữ liệu này được thực hiện trên công nghệ RDBMS. Mặc dù, kiểu triển khai này bị hạn chế bởi thực tế là hệ thống RDBMS truyền thống được tối ưu hóa để xử lý cơ sở dữ liệu giao dịch chứ không phải để lưu trữ dữ liệu. Ví dụ: truy vấn đặc biệt, liên kết nhiều bảng, tổng hợp tiêu tốn nhiều tài nguyên và làm chậm hiệu suất.

Do đó, các phương pháp tiếp cận thay thế cho Cơ sở dữ liệu được sử dụng như được liệt kê bên dưới-

* Trong một datahouse, các cơ sở dữ liệu quan hệ được triển khai song song để cho phép khả năng mở rộng. Cơ sở dữ liệu quan hệ song song cũng cho phép bộ nhớ được chia sẻ hoặc mô hình không chia sẻ gì trên các cấu hình đa xử lý khác nhau hoặc các bộ xử lý song song lớn.
* Các cấu trúc chỉ mục mới được sử dụng để bỏ qua quá trình quét bảng quan hệ và cải thiện tốc độ.
* Sử dụng cơ sở dữ liệu đa chiều (MDDBs) để khắc phục bất kỳ hạn chế nào được đặt ra do Mô hình Kho dữ liệu quan hệ. Ví dụ: Essbase từ Oracle.

**Công cụ tìm nguồn cung ứng, mua lại, làm sạch và chuyển đổi (ETL)**

Các công cụ tìm nguồn, chuyển đổi và di chuyển dữ liệu được sử dụng để thực hiện tất cả các chuyển đổi, tóm tắt và tất cả các thay đổi cần thiết để chuyển đổi dữ liệu thành một định dạng thống nhất trong datawarehouse. Chúng còn được gọi là Công cụ trích xuất, biến đổi và tải (ETL).

Chức năng của chúng bao gồm:

* Ẩn danh dữ liệu theo quy định của pháp luật.
* Loại bỏ dữ liệu không mong muốn trong cơ sở dữ liệu hoạt động khi tải vào Kho dữ liệu.
* Tìm kiếm và thay thế các tên và định nghĩa phổ biến cho dữ liệu đến từ các nguồn khác nhau.
* Tính toán tóm tắt và dữ liệu dẫn xuất
* Trong trường hợp thiếu dữ liệu, hãy điền chúng bằng các giá trị mặc định.
* Khử trùng lặp dữ liệu lặp lại đến từ nhiều nguồn dữ liệu.

Các công cụ Trích xuất (Extract), Chuyển đổi (Transform) và Tải (Load) này có thể tạo công việc cron, công việc nền, chương trình Cobol, tập lệnh shell, v.v. thường xuyên cập nhật dữ liệu trong datawarehouse. Những công cụ này cũng hữu ích để duy trì Siêu dữ liệu (Metadata).

Các Công cụ ETL này phải đối phó với các thách thức về tính không đồng nhất của Cơ sở dữ liệu & Dữ liệu.

**Metadata**

Tên Meta Data gợi ý một số khái niệm về Kho dữ liệu công nghệ cấp cao. Tuy nhiên, nó khá đơn giản. Siêu dữ liệu là dữ liệu về dữ liệu xác định kho dữ liệu. Nó được sử dụng để xây dựng, duy trì và quản lý kho dữ liệu.

Trong Kiến trúc Kho dữ liệu, siêu dữ liệu đóng một vai trò quan trọng vì nó chỉ định nguồn, cách sử dụng, giá trị và tính năng của dữ liệu kho dữ liệu. Nó cũng xác định cách dữ liệu có thể được thay đổi và xử lý. Nó được kết nối chặt chẽ với kho dữ liệu.

Ví dụ: một dòng trong cơ sở dữ liệu bán hàng có thể chứa:

4030 KJ732 299.90

Đây là một dữ liệu vô nghĩa cho đến khi chúng tôi tham khảo Meta cho chúng tôi biết nó là

* Số mô hình: 4030
* ID đại lý bán hàng: KJ732
* Tổng số tiền bán hàng là $ 299,90

Do đó, dữ liệu siêu dữ liệu là thành phần thiết yếu trong việc chuyển đổi dữ liệu thành kiến ​​thức.

Siêu dữ liệu giúp trả lời các câu hỏi sau

* Kho dữ liệu chứa những bảng, thuộc tính và khóa nào?
* Dữ liệu đến từ đâu?
* Dữ liệu được tải lại bao nhiêu lần?
* Sự biến đổi nào đã được áp dụng với sự tẩy rửa?

Siêu dữ liệu có thể được phân thành các loại sau:

1. Dữ liệu siêu dữ liệu kỹ thuật: Loại siêu dữ liệu này chứa thông tin về kho được sử dụng bởi các nhà thiết kế và quản trị kho dữ liệu.
2. Dữ liệu siêu dữ liệu doanh nghiệp: Loại siêu dữ liệu này chứa thông tin chi tiết giúp người dùng cuối hiểu được thông tin được lưu trữ trong kho dữ liệu một cách dễ dàng.

**Công cụ truy vấn**

Một trong những đối tượng chính của kho dữ liệu là cung cấp thông tin cho doanh nghiệp để đưa ra các quyết định chiến lược. Các công cụ truy vấn cho phép người dùng tương tác với hệ thống kho dữ liệu.

Các công cụ này được chia thành bốn loại khác nhau:

1. Công cụ truy vấn và báo cáo
2. Các công cụ phát triển ứng dụng
3. Các công cụ khai thác dữ liệu
4. Công cụ OLAP

**1. Công cụ truy vấn và báo cáo:**

Các công cụ truy vấn và báo cáo có thể được chia thành

* Công cụ báo cáo
* Các công cụ truy vấn được quản lý

**Công cụ báo cáo:**

Các công cụ báo cáo có thể được chia thành công cụ báo cáo sản xuất và trình viết báo cáo trên máy tính để bàn.

1. Người viết báo cáo: Loại công cụ báo cáo này là những công cụ được thiết kế cho người dùng cuối để phân tích.
2. Báo cáo sản xuất: Loại công cụ này cho phép các tổ chức tạo ra các báo cáo hoạt động thường xuyên. Nó cũng hỗ trợ các công việc hàng loạt khối lượng lớn như in ấn và tính toán. Một số công cụ báo cáo phổ biến là Brio, Business Objects, Oracle, PowerSoft, SAS Institute.

**Các công cụ truy vấn được quản lý:**

Loại công cụ truy cập này giúp người dùng cuối giải quyết các lỗi trong cơ sở dữ liệu và SQL và cấu trúc cơ sở dữ liệu bằng cách chèn siêu lớp (meta-layer) giữa người dùng và cơ sở dữ liệu.

**2. Các công cụ phát triển ứng dụng:**

Đôi khi các công cụ đồ họa và phân tích tích hợp sẵn không thỏa mãn nhu cầu phân tích của một tổ chức. Trong những trường hợp như vậy, các báo cáo tùy chỉnh được phát triển bằng cách sử dụng các công cụ phát triển Ứng dụng.

**3. Các công cụ khai thác dữ liệu:**

Khai thác dữ liệu là một quá trình khám phá các mối tương quan, xu hướng và xu hướng mới có ý nghĩa bằng cách khai thác dữ liệu số lượng lớn. Các công cụ khai thác dữ liệu được sử dụng để làm cho quá trình này tự động.

**4. Các công cụ OLAP:**

Các công cụ này dựa trên các khái niệm về cơ sở dữ liệu đa chiều. Nó cho phép người dùng phân tích dữ liệu bằng cách sử dụng các quan điểm đa chiều kỹ lưỡng và phức tạp.

**Kho dữ liệu Kiến trúc Bus**

Kho dữ liệu Bus xác định luồng dữ liệu trong kho của bạn. Luồng dữ liệu trong kho dữ liệu có thể được phân loại thành luồng vào, luồng lên, luồng xuống, luồng ra và luồng meta.

Trong khi thiết kế Bus dữ liệu, người ta cần phải xem xét các thứ nguyên (chiều) được chia sẻ, dữ kiện giữa các data mart.

# **Data Marts**

Data mart là một lớp truy cập được sử dụng để đưa dữ liệu ra ngoài cho người dùng. Nó được trình bày như một tùy chọn cho kho dữ liệu kích thước lớn vì nó tốn ít thời gian và tiền bạc hơn để xây dựng. Tuy nhiên, không có định nghĩa tiêu chuẩn nào về data mart là khác nhau tùy từng người.

Nói một cách đơn giản Data mart là một công ty con của kho dữ liệu. Data mart được sử dụng cho phân vùng dữ liệu được tạo cho nhóm người dùng cụ thể.

Data mart có thể được tạo trong cùng một cơ sở dữ liệu với Datawarehouse hoặc một Cơ sở dữ liệu riêng biệt về mặt vật lý.

# **Các phương pháp hay nhất về kiến ​​trúc kho dữ liệu**

Để thiết kế Kiến trúc kho dữ liệu, bạn cần làm theo các phương pháp hay nhất dưới đây:

* Sử dụng Mô hình Kho dữ liệu được tối ưu hóa để truy xuất thông tin có thể là chế độ thứ nguyên (chiều), phương pháp tiếp cận không chuẩn hóa hoặc kết hợp.
* Chọn phương pháp thiết kế phù hợp là phương pháp tiếp cận từ trên xuống và từ dưới lên trong Data Warehouse
* Cần đảm bảo rằng Dữ liệu được xử lý nhanh chóng và chính xác. Đồng thời, bạn nên thực hiện một cách tiếp cận hợp nhất dữ liệu thành một phiên bản duy nhất của sự thật.
* Thiết kế cẩn thận quy trình thu thập và làm sạch dữ liệu cho Kho dữ liệu.
* Thiết kế kiến ​​trúc MetaData cho phép chia sẻ siêu dữ liệu giữa các thành phần của Data Warehouse
* Xem xét triển khai mô hình ODS khi nhu cầu truy xuất thông tin ở gần cuối kim tự tháp trừu tượng hóa dữ liệu hoặc khi có nhiều nguồn hoạt động cần được truy cập.
* Người ta nên đảm bảo rằng mô hình dữ liệu được tích hợp chứ không chỉ hợp nhất. Trong trường hợp đó, bạn nên xem xét mô hình dữ liệu 3NF. Nó cũng lý tưởng để mua ETL và các công cụ làm sạch dữ liệu

# **Tóm lược:**

* Kho dữ liệu là một hệ thống thông tin chứa dữ liệu lịch sử và dữ liệu giao hoán từ một hoặc nhiều nguồn. Các nguồn này có thể là Kho dữ liệu truyền thống, Kho dữ liệu đám mây hoặc Kho dữ liệu ảo.
* Kho dữ liệu được định hướng theo chủ đề vì nó cung cấp thông tin liên quan đến chủ đề thay vì các hoạt động liên tục của tổ chức.
* Trong Data Warehouse, tích hợp có nghĩa là thiết lập một đơn vị đo lường chung cho tất cả các dữ liệu tương tự từ các cơ sở dữ liệu khác nhau
* Kho dữ liệu cũng không thay đổi có nghĩa là dữ liệu trước đó không bị xóa khi dữ liệu mới được nhập vào đó.
* Datawarehouse là biến thể theo Thời gian vì dữ liệu trong DW có thời hạn sử dụng cao.
* Chủ yếu có 5 thành phần của Kiến trúc kho dữ liệu: 1) Cơ sở dữ liệu 2) Công cụ ETL 3) Dữ liệu meta 4) Công cụ truy vấn 5) DataMarts
* Đây là bốn loại công cụ truy vấn chính 1. Truy vấn và báo cáo, công cụ 2. Công cụ phát triển ứng dụng, 3. Công cụ khai thác dữ liệu 4. Công cụ OLAP
* Các công cụ tìm nguồn, chuyển đổi và di chuyển dữ liệu được sử dụng để thực hiện tất cả các chuyển đổi và tóm tắt.
* Trong Kiến trúc Kho dữ liệu, siêu dữ liệu đóng một vai trò quan trọng vì nó chỉ định nguồn, cách sử dụng, giá trị và tính năng của dữ liệu kho dữ liệu.