**29.1.What is a data warehouse? How does it differ from a database?**

**Data warehouse** là một hệ thống tập hợp và lưu trữ thông tin từ nhiều nguồn khác nhau trong một tổ chức.

Mục tiêu của kho dữ liệu rõ ràng là theo định hướng kinh doanh: nó được thiết kế để tạo điều kiện thuận lợi cho việc ra quyết định bằng cách cho phép người dùng cuối hợp nhất và phân tích thông tin từ các nguồn khác nhau.

**sự khác nhau của database và data warehouse**

Sự khác biệt chính là database là bộ sưu tập dữ liệu được lưu trữ có tổ chức. Data warehouse là hệ thống thông tin được xây dựng từ nhiều nguồn dữ liệu - chúng được sử dụng để phân tích dữ liệu.

1. Với Database sử dụng giải pháp OLTP ( online transaction processing ) – data wahouse sử dụng giải pháp OLAP (online analytical processing).
2. Cơ sở dữ liệu có thể xử lý hàng nghìn người dùng cùng một lúc. Kho dữ liệu chỉ có thể xử lý một số lượng nhỏ hơn.
3. Cơ sở dữ liệu hữu ích nhất cho các giao dịch nhỏ, nguyên tử. Kho dữ liệu phù hợp nhất cho những câu hỏi lớn hơn đòi hỏi mức độ phân tích cao hơn.
4. Cơ sở dữ liệu cần phải sẵn sàng 24/7, có nghĩa là thời gian ngừng hoạt động rất tốn kém. Kho dữ liệu không bị ảnh hưởng bởi thời gian chết.
5. Cơ sở dữ liệu được tối ưu hóa để nhanh như chớp cho các hoạt động CRUD. Kho dữ liệu được tối ưu hóa cho một số lượng nhỏ các truy vấn phức tạp hơn trên nhiều kho dữ liệu lớn.
6. Cơ sở dữ liệu được cấu trúc hiệu quả nhất có thể, không có thông tin trùng lặp trong nhiều bảng. Kho dữ liệu thường không chuẩn hóa dữ liệu của chúng, ưu tiên các hoạt động đọc hơn các hoạt động ghi.
7. Cơ sở dữ liệu thường chỉ chứa thông tin cập nhật nhất, điều này làm cho các truy vấn lịch sử không thể thực hiện được. Kho dữ liệu đã được thiết kế từ đầu cho các mục đích báo cáo và phân tích.

OLAP (online analytical processing) là một thuật ngữ dùng để mô tả việc phân tích dữ liệu phức tạp từ data warehouse. Các công cụ OLAP cho phép truy vấn nhanh chóng và đơn giản dữ liệu phân tích được lưu trữ trong kho dữ liệu và siêu thị dữ liệu (cơ sở dữ liệu phân tích tương tự như kho dữ liệu nhưng với phạm vi hẹp được xác định).

OLTP ( online transaction processing ) là một kiểu xử lý dữ liệu bao gồm thực hiện một số giao dịch xảy ra đồng thời . Vai trò cơ bản là đảm bảo rằng các giao dịch đồng thời hoạt động chính xác và hiệu quả.

**29.2. Define the following terms: OLAP (online analytical processing), ROLAP (relational OLAP), MOLAP (multidimensional OLAP), and DSS (decision- support systems).**

**OLAP** (online analytical processing) là một thuật ngữ dùng để mô tả việc phân tích dữ liệu phức tạp từ data warehouse. Các công cụ OLAP cho phép truy vấn nhanh chóng và đơn giản dữ liệu phân tích được lưu trữ trong kho dữ liệu và siêu thị dữ liệu (cơ sở dữ liệu phân tích tương tự như kho dữ liệu nhưng với phạm vi hẹp được xác định).

**ROLAP** (Relational Online Analytical Processing) là mô hình quan hệ được triển khai cho data warehouse. **ROLAP** lưu trữ dữ liệu trong các cột và hàng (còn được gọi là bảng quan hệ) và truy xuất thông tin theo yêu cầu thông qua các truy vấn do người dùng gửi. Cơ sở dữ liệu ROLAP có thể được truy cập thông qua các truy vấn SQL phức tạp để tính toán thông tin. ROLAP có thể xử lý khối lượng dữ liệu lớn, nhưng dữ liệu càng lớn thì thời gian xử lý càng chậm.

Vì các truy vấn được thực hiện theo yêu cầu, ROLAP không yêu cầu lưu trữ và tính toán trước thông tin. Tuy nhiên, nhược điểm của việc triển khai ROLAP là các hạn chế về hiệu suất tiềm ẩn và các hạn chế về khả năng mở rộng do các hoạt động nối lớn và không hiệu quả giữa các bảng lớn. Ví dụ về các sản phẩm ROLAP phổ biến bao gồm Metacube của Stanford Technology Group, Red Brick Warehouse của Red Brick Systems và AXSYS Suite của Information Advantage.

**MOLAP** (multidimensional OLAP) là mô hình lập phương đa chiều cho data warehouse. MOLAP phân tách data warehouse thành những khối lập phương đa chiều để truy cập dữ liệu. Dữ liệu được tính toán trước, tóm tắt trước và lưu trữ trước khác với ROLAP là dữ liệu sẽ phục vụ theo nhu cầu của câu truy vấn )

**DSS (decision- support systems)** hệ thống hỗ trợ đưa ra quyết định. DSS là một hệ thống hỗ doanh nghiệp trong các hoạt động ra quyết định. Hệ thống hỗ trợ quản lý cấp trung và cao của một tổ chức bằng cách phân tích khối lượng lớn dữ liệu phi cấu trúc và tích luỹ thông tin để có thể giúp giải quyết các vấn để đưa ra quyết định. DSS có thể do năng lực của con người quyết định, hoặc tự động, hoặc kết hợp cả 2.

29.5 Define the following terms: star schema, snowflake schema, fact constellation, data marts.  
  
Sơ đồ ngôi sao (Star Schema)

Sơ đồ ngôi sao là mô hình đơn giản nhất được sử dụng trong DWH. Bởi vì bảng fact là trung tâm của mô hình với các bảng dimension xung quanh nó, nó nhìn giống như một ngôi sao. Điều này rất rõ ràng khi bảng fact được bao quanh bởi 5 bảng dimension. Một biến thể của sơ đồ ngôi sao là sơ đồ con rết (centipede schema), nơi mà bảng fact được bao quanh bởi số lượng lớn các bảng dimension nhỏ.  
  
Diagram

Description automatically generated  
  
Mô hình thiên hà (Galaxy Schema)

Kiểu mô hình chứa vài bảng fact và chia sẻ các bảng dimension nó được gọi là mô hình galaxy (galaxy schema)

Diagram

Description automatically generated

Mô hình bông tuyết (Snowflake Schema)

Mô bông tuyết là dạng mở rộng của lược đồ hình sao bằng cách bổ sung các Dimension. Bảng Fact giống lược đồ hình sao. Bảng Dim được chuẩn hóa. Các chiều được cấu trúc rõ ràng. Một bảng Dim được chia thành chiều chính hay chiều phụ thì đó là dạng bông tuyết hoặc hình sao mở rộng.  
  
Diagram

Description automatically generated

Mô hình chòm sao (Constellation Schema)

Ở lược đồ này thì Dimension trong lược đồ được tách biệt thành Dimension độc lập dựa trên các cấp độ của hệ thống phân cấp. Ví dụ Geo có 5 cấp Region, Branch, Area, Province, District thay vì chỉ có một Geo chung.

Diagram

Description automatically generated  
  
29.7. **Describe the steps of building a warehouse.  
  
Quy trình triển khai hệ thống Data Warehouse**

1. Đơn vị nghiệp vụ đưa yêu cầu nghiệp vụ, bộ phận BA phân tích các thông tin nghiệp vụ, yêu cầu tiện ích và phương pháp hiển thị

• Đối với các đơn vị mới triển khai Data Warehouse thì nên lựa chọn các báo cáo sau để xây dựng Data Warehouse

• Các báo cáo gây ảnh hưởng đến hiệu năng hệ thống

• Các báo cáo tổng hợp từ nhiều hệ thống trước đây phải tổng hợp thủ công

• Đối với các đơn vị đã triển khai Data Warehouse

• Xây dựng Data Warehouse dựa trên nhu cầu phân tích dữ liệu

• Dựa trên nhu cầu xây dựng hệ thống AI (Cần hiểu chi tiết tại sao Data Warehouse cần thiết thì các bạn tìm hiểu tại …)

2. Phân tích yêu cầu nghiệp vụ của các hệ thống liên quan (Các hệ thống OLTP)

* Đối với các yêu cầu nghiệp vụ có thể lấy và tổng hợp dữ liệu từ nhiều hệ thống khác nhau
* Tính toán có phải chỉnh sửa các hệ thống liên quan để đáp ứng yêu cầu nghiệp vụ

3. Thiết kế Data Warehouse

* Sau khi lựa chọn mô hình Ngôi sao hoặc Thiên hà(Bông tuyết) với các lưu ý
* **Các bảng dimension** được sử dụng để mô tả dữ liệu mà chúng ta muốn lưu trữ. Ví dụ: một nhà bán lẻ muốn lưu trữ thời gian, cửa hàng, và nhân viên tham gia vào một hoá đơn. Mỗi một bảng dimension là một danh mục của chính nó (ngày tháng, nhân viên, cửa hàng) và có thể có một hoặc nhiều thuộc tính (attributes). Với mỗi một cửa hàng, chúng ta lưu chúng các thông tin như vị trí trong thành phố, vùng miền, tỉnh thành và quốc gia. Mỗi một ngày tháng chúng ta lưu năm, tháng, ngày trong tháng, ngày trong tuần…Điều này liên quan đến sự phân cấp của các thuộc tính trong bảng dimension.
* **Bảng Fact** chứa dữ liệu mà chúng ta muốn thêm vào reports, tổng hợp trên các giá trị trong các bảng dimension. Một bảng fact chỉ có các cột lưu giá trị và các cột khóa ngoại tham chiếu đến bảng dimensions. Kết hợp tất cả các khóa ngoại và khoá chính trong bảng fact. Ví dụ, một bảng fact có thể lưu trữ một số lượng các hợp đồng và số lượng các nhân viên bán hàng từ các danh sách hợp đồng.

4. Xây dựng ETL

ELT là từ viết tắt cho Extract (Trích xuất), Load (Tải lên), and Transform (Chuyển đổi).

Data pipeline dành cho ELT sẽ bao gồm quy trình 3 bước được thực hiện ngay trên dữ liệu, gồm có:

**Extract (Trích xuất):** Việc trích xuất các dữ liệu là quá trình xác định và trích xuất các dữ liệu cần thiết, từ một hoặc nhiều nguồn khác nhau, như database, file, archives, ERP, CRM, v.v.

**Load (Tải lên):** Quy trình này sẽ bao gồm việc tải các dữ liệu được trích xuất sẽ được lên các database xác định.

**Transform (Chuyển đổi):** Chuyển đổi dữ liệu chính là quy trình chuyển đổi các dữ liệu từ hình thức cũ trên hệ thống nguồn sang hình thức mới, để phù hợp cho việc phân tích dữ liệu.

Việc chuyển đổi này thường dựa trên nhu cầu phân tích và sử dụng dữ liệu trong kho dữ liệu mục tiêu.

Chart, waterfall chart

Description automatically generated

Mặc dù việc chuyển đổi các dữ liệu có thể diễn ra dưới nhiều hình thức

5. Phát triển báo cáo và sử dụng cac công cụ BI

* Chỉnh sửa và hiệu chỉnh theo yêu cầu sử dụng

6. Lên kế hoạch thực thi ETL

* Dưa vào yêu cầu nghiệp vụ, và ảnh hưởng việc thực thi ETL ảnh hưởng tới các hệ thống OLTP, xây dựng chính sách quy trình thực thi ETL
* Việc này phải cân đối giữa yêu cầu nghiệp vụ và ảnh hưởng tới các hệ thống OLTP trong quá trình chạy ETL

7. Test trên hệ thống DEV hoặc Stag

* Chỉnh sửa và hiệu chỉnh

8. Triển khai hệ thống Production

* Theo quy trình quản lý triển khai của từng đơn vị

Table, calendar

Description automatically generated

29.11. **List the difficulties in implementing a data warehouse.**

Những khó khan trong việc triển khai một data warehouse

* 1. Xây dựng một datawarehouse cho doanh nghiệp lớn là một công việc lớn
  2. Việc kiểm soát chất lượng và tính nhất quán của dữ liệu
  3. Khó khan trong việc thiết kế, xây dựng và triển khai datawarehouse
  4. Độ chính xác của dữ liệu
  5. Khả năng đồng bộ của các database khác nhau trong cùng khoảng thời gian
  6. Khó khan trong việc xử lý dữ liệu trong thời gian thực