ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN

A picture containing logo

Description automatically generated

**BÀI TẬP NHÓM 1**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

*GVHD: Hồ Thị Hoàng Vy*

|  |  |  |
| --- | --- | --- |
| **Mã nhóm** | **MSSV** | **Họ và tên** |
| **TTKD-24** | 18120227 | Phạm Văn Minh Phương |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

**BẢNG PHÂN CÔNG**

|  |  |  |  |
| --- | --- | --- | --- |
| MSSV | Họ và tên | Phân công | Đánh giá % |
| 18120227 | Phạm Văn Minh Phương | Câu hỏi 3, 4 | 100% |
| 18120299 | Trương Công Quốc Cường | Câu hỏi 5, slide tóm tắt | 100% |
| 18120035 | Đoàn Nguyễn Tấn Hưng | Câu hỏi 2, viết báo cáo | 100% |
| 18120534 | Hoàng Công Sơn | Câu hỏi 1 | 100% |

Mục lục

[Câu 1: Khác nhau giữa Data Warehouse và Data Warehousing 2](#_Toc82873662)

[Câu 2: Các thành phần Datawarehouse 3](#_Toc82873663)

[I. Data Warehouse Database 3](#_Toc82873664)

[II. ETL Tools 3](#_Toc82873665)

[III. Metadata 4](#_Toc82873666)

[IV. Công cụ truy vấn 5](#_Toc82873667)

[**1.** **Công cụ truy vấn và báo cáo:** 5](#_Toc82873668)

[**2.** **Công cụ phát triển ứng dụng:** 6](#_Toc82873669)

[**3.** **Công cụ khai thác dữ liệu:** 6](#_Toc82873670)

[**4.** **Công cụ OLAP:** 6](#_Toc82873671)

[Câu 3: Một số đặc điểm của Data Warehouse 7](#_Toc82873672)

[1. Hướng chủ thể (Subject-Oriented) 7](#_Toc82873673)

[2. Tích hợp (Intergrated) 7](#_Toc82873674)

[3. Biến thể thời gian (Time-Variant) 8](#_Toc82873675)

[4. Không bay hơi (Non-volatile) 8](#_Toc82873676)

[Câu 4: Một số tool cho data warehouse và ứng dụng của chúng 9](#_Toc82873677)

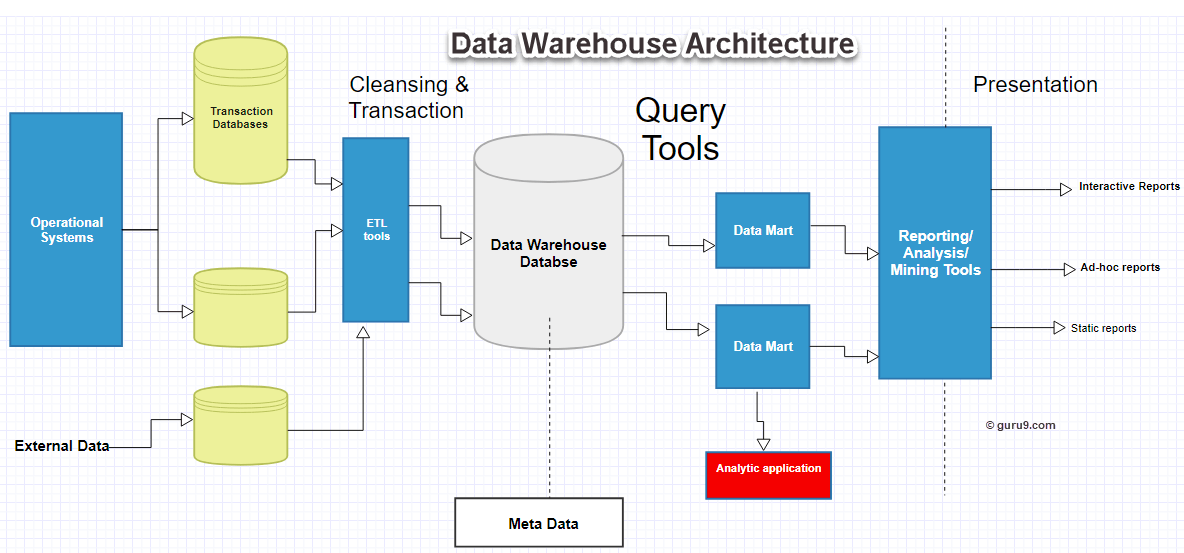
[Câu 5: Business Intelligence là gì? 10](#_Toc82873678)

# **Câu 1: Khác nhau giữa Data Warehouse và Data Warehousing**

* Data warehouse bao gồm một kiến ​​trúc và quy trình hoàn chỉnh (quá trình thiết kế DWH/cơ sở dữ liệu, quá trình ETL), trong khi data warehouse là dữ liệu được lưu trữ trong cơ sở dữ liệu ở dạng thứ nguyên, bảng dữ kiện, bảng tra cứu, bảng dữ kiện tổng hợp.
* Data warehousing là quá trình xây dựng và sử dụng một data warehouse. Một data warehouse được xây dựng bằng cách tích hợp dữ liệu từ nhiều nguồn không đồng nhất mà hỗ trợ báo cáo phân tích, truy vấn có cấu trúc và ra quyết định. Data warehousing liên quan đến việc làm sạch dữ liệu, tích hợp dữ liệu và hợp nhất dữ liệu.

# **Câu 2:** **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:



Trong đó, chủ yếu có năm Thành phần Data Warehouse:

1. **Data Warehouse Database**

Cơ sở dữ liệu trung tâm là nền tảng của môi trường Data Warehouse. Cơ sở dữ liệu này được thực hiện trên công nghệ RDBMS.

1. **ETL Tools**

ETL – Extract, Transform and Load

Các công cụ ETL 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 Data Warehouse.

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 Data Warehouse.
* 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, điền chúng bằng các giá trị mặc định.
* Loại bỏ trùng lặp dữ liệu lặp lại đến từ nhiều nguồn dữ liệu.

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

1. **Metadata**

Metadata là dữ liệu về dữ liệu xác định Data Warehouse. Nó được sử dụng để xây dựng, duy trì và quản lý Data Warehouse.

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

Metadata giúp trả lời các câu hỏi sau:

* Data Warehouse 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 làm sạch?

Metadata có thể được phân thành các loại sau:

1. Technical Meta Data: Loại metadata này chứa thông tin về kho được sử dụng bởi các nhà thiết kế và quản trị Data Warehouse.
2. Business Meta Data: Loại metadata này chứa thông tin chi tiết giúp người dùng cuối hiểu được thông tin được lưu trữ trong Data Warehouse một cách dễ dàng.
3. **Công cụ truy vấn**

Các công cụ truy vấn cho phép người dùng tương tác với hệ thống Data Warehouse.

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

1. Công cụ truy vấn và báo cáo

2. Công cụ phát triển ứng dụng

3. 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
* 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.

**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.

1. **Công cụ phát triển ứng dụng:**

Các báo cáo tùy chỉnh được phát triển bằng cách sử dụng các công cụ phát triển Ứng dụng.

1. **Công cụ khai thác dữ liệu:**

Các công cụ khai thác dữ liệu được sử dụng để làm cho quá trình khai thác dữ liệu được tự động hoá.

1. **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 khung nhìn đa chiều kỹ lưỡng và phức tạp.

1. **Kiến trúc Data warehouse Bus**

Data Warehouse Bus xác định luồng dữ liệu (data flow) trong kho của bạn. Luồng dữ liệu trong Data Warehouse có thể được phân loại thành luồng vào (Inflow), luồng lên (Upflow), luồng xuống (Downflow), luồng ra (Outflow) và luồng meta (Meta flow).

**Câu 3: Một số đặc điểm của Data Warehouse**

1. **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.

1. **Tích hợp (Intergrated)**

* 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.

1. **Biến thể thời gian (Time-Variant)**

* Khoảng thời gian cho kho dữ liệu khá rộng so với các hệ thống vận hành. Dữ liệu được thu thập trong kho dữ liệu được ghi nhận với một giai đoạn cụ thể và cung cấp thông tin theo quan điểm lịch sử. Nó chứa một yếu tố thời gian, một cách rõ ràng hoặc hàm 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.

1. **Không bay hơi (Non-volatile)**

* Kho dữ liệu không bay hơi có nghĩa là dữ liệu trước đó không bị xóa khi dữ liệu mới được nhập vào.
* Dữ liệu ở chế độ chỉ đọc và được làm mới định kỳ. Điều này cũng giúp phân tích dữ liệu lịch sử để hiểu điều gì đã xảy ra khi nào. Nó không yêu cầu quy trình giao dịch, cơ chế phục hồi và kiểm soát đồng thời.
* Các hoạt động như delete, update, insert được thực hiện trong môi trường ứng dụng hoạt động bị bỏ qua trong môi trường Kho dữ liệu. Chỉ có hai loại hoạt động dữ liệu được thực hiện trong Kho dữ liệu là:
* Data loading
* Data access
* Sự khác biệt giữa ứng dụng và Data Warehouse:

|  |  |
| --- | --- |
| Ứng dụng | Data Warehouse |
| Chương trình phải được lập trình để đảm bảo rằng các quá trình update dữ liệu duy trì tính toàn vẹn của sản phẩm cuối cùng. | Loại vấn đề 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 ở dạng chuẩn hóa |
| Cần có các kĩ thuật hỗ trợ giải quyết vấn đề về transaction; khôi phục, rollback dữ liệu (vì có những deadlock khá phức tạp) | Các kĩ thuật xử lý các vấn đề tương tự khá đơn giản |

**Câu 4: Một số tool cho data warehouse và ứng dụng của chúng**

1. **CData Sync (**[**https://www.cdata.com/sync**](https://www.cdata.com/sync)**)**

* Hợp nhất dữ liệu từ bất kỳ ứng dụng hoặc nguồn dữ liệu nào vào Cơ sở dữ liệu hoặc Kho dữ liệu mà bạn lựa chọn.
* Kết nối dữ liệu hỗ trợ doanh nghiệp với BI, Analytics và Machine Learning.
* Tự động sao chép dữ liệu gia tăng thông minh
* Chuyển đổi dữ liệu ETL / ELT có thể custom

1. **BiG EVAL (**[**https://go.bigeval.com/automated-dwh-testing**](https://go.bigeval.com/automated-dwh-testing)**)**

* Đo lường chất lượng dữ liệu và hỗ trợ giải quyết vấn đề
* Autopilot testing để phát triển agile, trích xuất meta-data từ cơ sở dữ liệu hoặc kho meta-data
* Tóm tắt cho bất kỳ loại dữ liệu nào (RDBMS, API, Flatfiles, ứng dụng doanh nghiệp…).

1. **Xplenty (**[**https://www.xplenty.com**](https://www.xplenty.com)**)**

* Tập hợp và chuẩn bị dữ liệu cho BI
* Truyền và chuyển đổi dữ liệu giữa các cơ sở dữ liệu nội bộ hoặc kho dữ liệu
* Gửi dữ liệu bổ sung của bên thứ ba tới Heroku Postgres (và sau đó tới Salesforce qua Heroku Connect) hoặc trực tiếp tới Salesforce.
* Trình kết nối API Rest để lấy dữ liệu từ bất kỳ API Rest nào

1. **Oracle (**[**https://www.oracle.com/autonomous-database/autonomous-data-warehouse**](https://www.oracle.com/autonomous-database/autonomous-data-warehouse)**)**

* Phân phối dữ liệu theo cùng một cách trên các đĩa để mang lại hiệu suất đồng nhất
* Hoạt động với cụm ứng dụng hoặc ứng dụng riêng lẻ
* Test ứng dụng trong môi trường thực
* Kết nối tốc độ cao để chuyển data khối lượng lớn
* Hỗ trợ trực quan hóa
* Cho phép kết nối database, table, view từ xa

1. **Informatica (**[**https://www.informatica.com**](https://www.informatica.com)**)**

* Hệ thống ghi lỗi tập trung tạo điều kiện cho các lỗi ghi nhật ký và từ chối dữ liệu vào các bảng quan hệ
* Khả năng mở rộng quy mô tích hợp dữ liệu

1. **Google BigQuery (**[**https://cloud.google.com/bigquery**](https://cloud.google.com/bigquery)**)**

* Cung cấp tính năng Nhập dữ liệu linh hoạt
* Đọc và ghi dữ liệu thông qua Cloud Dataflow, Hadoop, and Spark.
* Dịch vụ chuyển dữ liệu tự động
* Toàn quyền kiểm soát quyền truy cập vào dữ liệu được lưu trữ
* BigQuery cung cấp cơ chế kiểm soát chi phí

# **Câu 5: Business Intelligence là gì?**

* Business Intelligence (BI) đề cập đến các kỹ năng, quy trình, công nghệ, ứng dụng được sử dụng cho việc hỗ trợ ra quyết định.
* BI là tập các kỹ thuật và công cụ để chuyển đổi dữ liệu thô thành thông tin có nghĩa và có ích cho các mục đích phân tích nghiệp vụ.
* BI là các ứng dụng và công nghệ chuyển đổi dữ liệu doanh nghiệp thành hành động.

**Refs:**

[**https://www.tutorialspoint.com/dwh/dwh\_data\_warehousing.htm#:~:text=A%20data%20warehouse%20is%20constructed,data%20integration%2C%20and%20data%20consolidations**](https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm#:~:text=A%20data%20warehouse%20is%20constructed,data%20integration%2C%20and%20data%20consolidations)

[**http://www.geekinterview.com/question\_details/35430**](http://www.geekinterview.com/question_details/35430)[**https://www.guru99.com/data-warehouse-architecture.html**](https://www.guru99.com/data-warehouse-architecture.html)

[**https://o7planning.org/10355/what-is-business-intelligence**](https://o7planning.org/10355/what-is-business-intelligence)