Các khái niệm về Data Warehouse

# Sự cần thiết của BI

## 1.1) BI (Business Intelligence) là gì?

- BI là hoạt động đóng vai trò trong sự phát triển của công ty.

- BI là hoạt động biến đổi dữ liệu thô/ dữ liệu vận hành thành thông tin có ích cho hoạt động phân tích doanh nghiệp (BA).

## 1.2) Cách mà BI hoạt động

- BI dựa trên công nghệ Data Warehouse để đưa ra **(Extract)** thông tin từ hệ thống vận hành của công ty.

- Dữ liệu được biến đổi **(Transform)** (được làm sạch và tích hợp), và được đưa **(Load)** vào Data Warehouse.

- Giờ dữ liệu đã đáng tin cậy, và được dùng cho các quyết định trong doanh nghiệp.

# Data Warehousing là gì?

## 2.1) Tại sao lại dùng Data Warehouse (DWH)?

Dữ liệu từ các loại Database khác nhau như Oracle, SAP, SQL Server, … liệu được tích hợp và xử lý, sau đó được đưa vào DWH. Từ dữ liệu trong DWH, chúng có thể được sử dụng trong các mục đích sau này, ví dụ như trực quan hóa dữ liệu.

## 2.2) Data Warehouse là gì?

### 2.2.1) Khái niệm về Data Warehouse

- DWH được ví như một nhà kho trung tâm chứa các dữ liệu được hợp lại từ các nguồn (Database) khác.

- DWH được duy trì, vận hành tách biệt với CSDL vận hành của doanh nghiệp. Trong trường hợp DWH gặp sự cố, các dữ liệu vẫn sẽ được phục hồi do các dữ liệu thô được lưu trong các Database ngoài.

- Người dùng cuối truy cập vào DWH bất cứ khi nào họ cần thông tin từ DWH.

- **Lưu ý:** DWH sẽ không load bất cứ khi nào data mới được đưa vào các Database ngoài. Vì vậy Data trong DWH là dữ liệu di sản (hoặc historical data), vì nó không được cập nhật liên tục như data trong các Database ngoài.

### 2.2.2) Lợi ích của DWH

- Các câu hỏi chiến lược có thể được trả lời bằng quá trình phát hiện trends.

- DWH nhanh và chính xác hơn.

- **Lưu ý:** DWH không phải một sản phẩm có thể mua, nó được thiết kế dựa trên yêu cầu của từng doanh nghiệp.

### 2.2.3) Các đặc điểm của DWH

- Hướng chủ đề: Được phân loại và lưu trữ theo chủ đề của doanh nghiệp hơn là theo ứng dụng.

- Tích hợp: Dữ liệu của một chủ đề được thu thập từ các nguồn khác nhau và được lưu trữ lại tại một nơi duy nhất.

- Gán thời gian: Dữ liệu được lưu trữ như một dãy các snapshots được gán thời gian.

- Không biến đổi: Dữ liệu trong DWH không được update hoặc xóa.

# Các thuật ngữ chính liên quan tới Kiến trúc DWH

## 3.1) OLTP (On-line Transactional Processing) vs OLAP (On-line Analytical Processing) cho việc truy cập dữ liệu

|  |  |
| --- | --- |
| **Relational Database (OLTP)** | **Analytical Data Warehouse (OLAP)** |
| Chứa dữ liệu hiện tại | Chứa dữ liệu lịch sử |
| Phục vụ điều hành doanh nghiệp | Phục vụ phân tích doanh nghiệp |
| Dựa trên ERD | Dựa trên Star, Snowflake và Fact Constellation Schema |
| Đưa ra dữ liệu còn thô và mang nhiều chi tiết | Đưa ra dữ liệu được tổng hợp và khái quát hóa. |
| Được dùng để ghi data vào CSDL | Được dùng để đọc data từ DWH |
| Database size từ 100 MB -> 1GB | DWH size từ 100GB -> 1TB |
| Nhanh, hiệu suất cao | Linh hoạt, nhưng k nhanh bằn |
| Số records được truy cập ở hàng chục | Số records được truy cập ở hàng triệu |

## 3.2) ETL (Extract, Transform, Load)

ETL là quá trình lấy dữ liệu từ các nguồn khác nhau, biến đổi chúng để đáp ứng các yêu cầu của doanh nghiệp và tải lên DWH.

## 3.3) Data Mart

- Data Mart là một phiên bản nhỏ hơn của DWH, nó tập trung vào một subject cụ thể.

- Data Mart tập trung vào một lĩnh vực cụ thể, nên chỉ lấy dữ liệu từ một số nguồn dữ liệu giới hạn.

- Thời gian để build 1 Data Mart ngắn hơn nhiều so với 1 DWH.

- *Các loại Data Mart:*

+ Dependent Data Mart: Cần DWH. (OLTP Source -> DWH -> Data Mart)

+ Independent Data Mart: Không cần DWH. (OLTP Source -> Data Mart)

+ Hybrid Data Mart: Lấy dữ liệu từ cả OLTP Source và OLAP Source (DWH). (OLTP Source + DWH -> Data Mart)

## 3.4) Metadata (Siêu dữ liệu)

- Metadata là dữ liệu về dữ liệu. Nó là dữ liệu mô tả về nguồn và bối cảnh của dữ liệu. VD: 1 bức ảnh có Metadata là: Ngày, giờ, tên, …

- Metadata trong DWH định nghĩa nguồn của dữ liệu, VD: File Flat, RD, …

- Metadata dùng để định nghĩa bảng nào là nguồn và đích, và khái niệm nào được dùng để build business logic, quá trình này được gọi là biến đổi sang kết quả cuối cùng.

# Kiến trúc DWH

Data Source (OLTP) - > ETL -> Staging Database -> ETL -> DWH (Metadata + Raw Data + Aggregate Data) -> Data Mart -> End Users.