

# Chapter 5:

# Data Warehouse



1. Data Warehouse
2. Data Mart
3. Master Data Management
4. Dimension and Fact
5. Extract, Transform, Load
6. Style of Data warehouse Schema



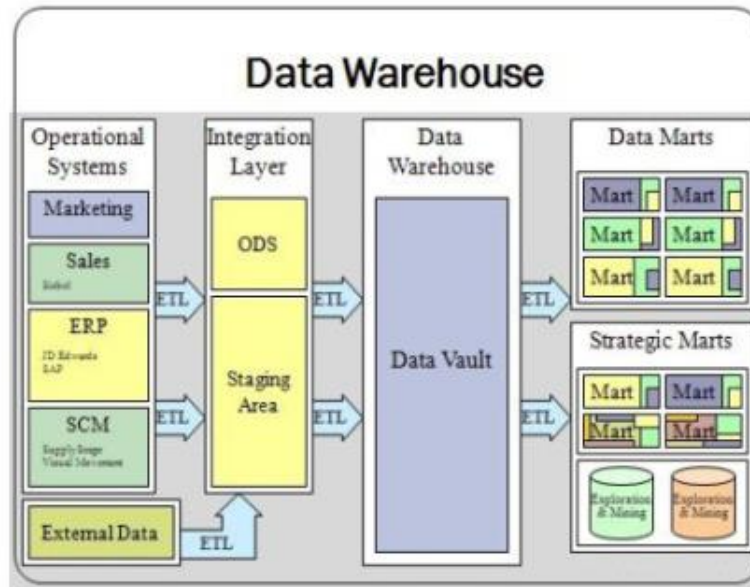
# *1. Data Warehouse*

## 1.1 Định nghĩa

4

Định nghĩa:

- » **Kho dữ liệu (data warehouse)** là khái niệm được giới thiệu lần đầu vào năm 1988 bởi 2 nhà nghiên cứu của IBM là Barry Devlin và Paul Murphy.
- » Data warehouse là nơi lưu trữ dữ liệu bằng thiết bị điện tử của một tổ chức, doanh nghiệp, nhằm hỗ trợ việc phân tích dữ liệu và lập báo cáo.



Data Warehouse Overview

## 1.1 Định nghĩa

5

- » Data warehouse bao gồm tập các phương pháp, kỹ thuật và công nghệ có thể kết hợp, hỗ trợ nhau để thu thập và quản lý dữ liệu từ nhiều nguồn, môi trường khác nhau để cung cấp thông tin cho người sử dụng.
- » Kho dữ liệu còn có thể được gọi bằng những cái tên sau: hệ thống hỗ trợ quyết định, hệ thống điều hành thông tin, hệ thống thông tin quản lý, ứng dụng phân tích. Kho dữ liệu thường có dung lượng lên đến hàng trăm GB thậm chí tính bằng đơn vị TB.

## 1.2 Hệ thống Data Warehouse

6

- » Data mart: là một hình thức đơn giản của Data warehouse, chỉ tập trung vào một chủ đề, một lĩnh vực hoặc một chức năng.
- » OLAP (Online analytical processing):
  - » Khối lượng giao dịch tương đối thấp
  - » Các truy vấn phức tạp
  - » Được sử dụng để phân tích dữ liệu đa chiều từ nhiều nguồn và quan điểm.
  - » Hoạt động chủ yếu là các truy vấn dữ liệu, thường là các truy vấn lớn và mất nhiều thời gian để thực hiện.
  - » Hiệu quả được đo lường bằng thời gian phản hồi.

## 1.2 Hệ thống Data Warehouse

7

- » OLTP (Online transaction processing)
  - » Đặc trưng cho các giao dịch trực tuyến ngắn.
  - » Xử lý truy vấn rất nhanh và duy trì tính toàn vẹn của dữ liệu trong môi trường đa truy cập.
  - » Hiệu quả được đo lường bằng số lượng giao dịch mỗi giây.
- » Phân tích dự đoán (Predictive analysis): là tìm kiếm và định lượng các mẫu ẩn trong dữ liệu bằng cách sử dụng các mô hình toán học phức tạp có thể được sử dụng để dự đoán kết quả trong tương lai.

## 1.3 Đặc điểm của Data warehouse

8

- » Dữ liệu toàn vẹn (Integration)
- » Dữ liệu gắn thời gian và có tính lịch sử (Isolation)
- » Dữ liệu vận hành ổn định (Nonvolatility)
- » Dữ liệu cố định (Durable)
- » Dữ liệu được hướng theo chủ đề (Subject orientation)



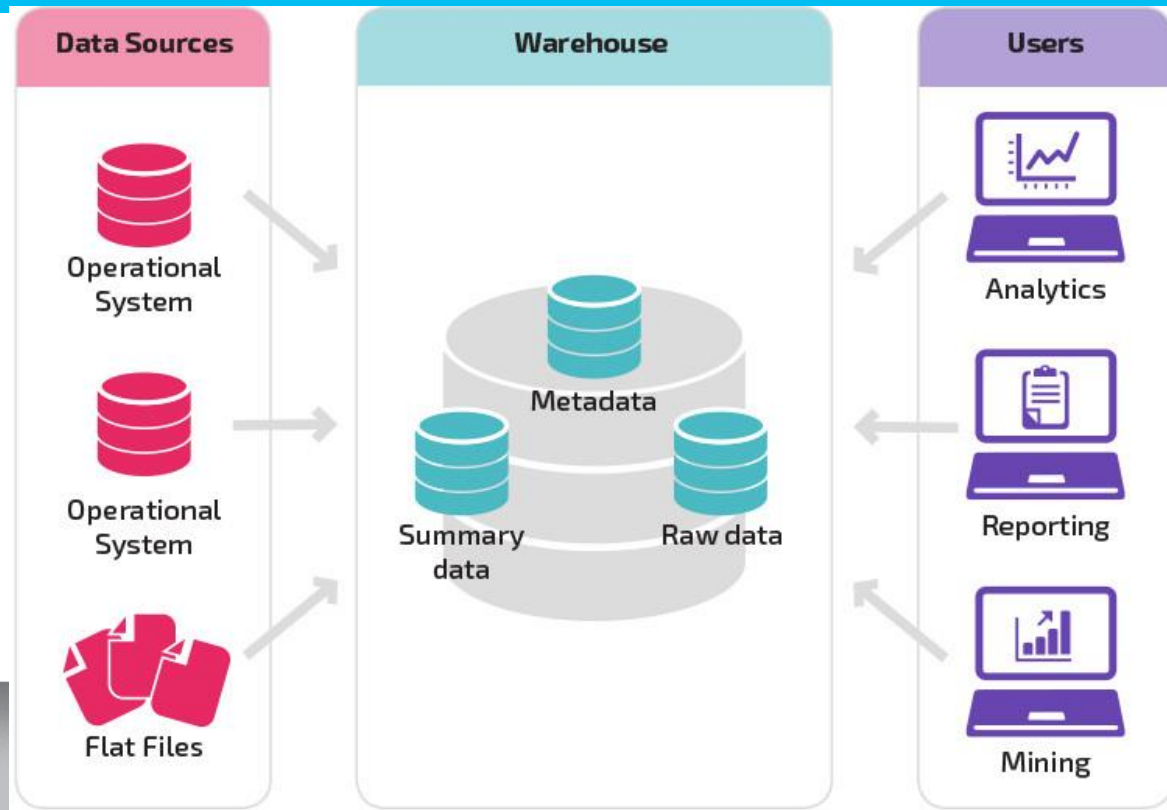
## 1.4 Cấu trúc của Data warehouse

9

- » Tầng đáy (Data Sources): là nơi thu thập, tích hợp dữ liệu từ nhiều nguồn khác nhau sau đó chuẩn hóa về cùng định dạng; làm sạch xử lý dữ liệu để tìm lỗi và sửa và lưu trữ dữ liệu đã tổng hợp
- » Tầng giữa (Warehouse): cung cấp dịch vụ OLAP để thao tác với kho dữ liệu đã được tích hợp. Tầng giữa có thể được cài đặt bằng Relational OLAP, Multidimensional OLAP; hay kết hợp cả hai mô hình trên Hybrid OLAP
- » Tầng trên cùng (Users): nơi thực hiện các thao tác truy vấn, báo cáo, phân tích để tìm ra xu hướng, trung bình, tổng hợp...

## 1.4 Cấu trúc của Data warehouse

10



## 1.5 Các phương pháp thiết kế Data warehouse

11

### **Bottom-up Design**

- » Data warehouse như là tập hợp của các data mart được làm cho phù hợp
- » Data mart được tạo ra trước để cung cấp khả năng phân tích và báo cáo cho những việc kinh doanh chuyên biệt dựa trên mô hình dữ liệu chiều

## 1.5 Các phương pháp thiết kế Data warehouse

12

### Top-down Design

- » Dữ liệu trong Data Warehouse được lưu trữ ở mức “hạt” (granularity) thấp nhất dựa trên mô hình dữ liệu đã chuẩn hóa
- » KDL là trung tâm của “Xưởng thông tin công ty” (Corporate Information Factory-CIF) cung cấp một khung nền luận lý (logical framework) cho việc chuyển giao kinh doanh thông minh đến doanh nghiệp. Các hoạt động kinh doanh cung cấp dữ liệu cho CIF. ☒ KDL được tập trung hóa sẽ cung cấp các data mart phụ thuộc có thể được thiết kế dựa trên mô hình dữ liệu có chiều

## 1.6 Một số lợi ích của Data warehouse

13

- » Tích hợp dữ liệu từ nhiều nguồn vào một database và data model.
- » Tạo ra một cái nhìn tổng quan cho toàn bộ hệ thống doanh nghiệp.
- » Lưu trữ lịch sử dữ liệu ngay cả khi hệ thống giao dịch không lưu.
- » Bằng cách cung cấp mã nhất quán, mô tả và sửa các dữ liệu xấu giúp cải thiện chất lượng dữ liệu.
- » Thể hiện thông tin của tổ chức một cách nhất quán.
- » Cung cấp một mô hình dữ liệu chung duy nhất cho tất cả các nguồn dữ liệu đầu vào.
- » Tái cấu trúc lại dữ liệu để mang lại hiệu suất cao nhất ngay cả đối với các truy vấn phức tạp mà không ảnh hưởng đến hệ thống.

## 1.7 Lợi ích từ việc xây dựng data warehouse cho doanh nghiệp

14

- » Nền tảng để doanh nghiệp phát triển
- » Xây dựng hệ thống truy vấn dữ liệu, nâng cao hiệu quả công việc
- » Tổng hợp các nguồn số liệu hoạt động của doanh nghiệp về một nơi để quản lý
- » Chất lượng dữ liệu và tính thống nhất dữ liệu của toàn doanh nghiệp
- » Cắt giảm thời gian
- » Nâng cao lợi thế cạnh tranh của doanh nghiệp
- » Hệ số thu nhập trên đầu tư (ROI) rất cao
- » Hỗ trợ các nhà quản lý tự tin dự báo các hoạt động của doanh nghiệp
- » Cải thiện quy trình và thời gian để ra quyết định cho các cấp quản lý và lãnh đạo
- » Cung cấp thông tin, dữ liệu hoạt động quá khứ của doanh nghiệp



## *2. Data Mart*

## 2.1 Khái niệm Data mart

16





| Data warehouse                                                                            | Data mart                                                                                              |
|-------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| Chứa nhiều chủ đề                                                                         | Thường chỉ nắm giữ một lĩnh vực                                                                        |
| Chứa thông tin chi tiết                                                                   | Chứa dữ liệu tóm tắt và được chọn                                                                      |
| Hoạt động để tích hợp tất cả các nguồn dữ liệu                                            | Tập trung vào việc tích hợp thông tin từ một lĩnh vực chủ đề nhất định hoặc tập hợp các hệ thống nguồn |
| Data warehouse là <b>tính linh hoạt, định hướng thông tin</b> và có tính chất lâu dài     | Data mart <b>bị hạn chế, định hướng dự án</b> và có thời gian tồn tại ngắn hơn.                        |
| Dữ liệu được lưu trữ trong một Data warehouse <b>tập trung</b> duy nhất trong kho dữ liệu | Lưu trữ dữ liệu trong khu vực người dùng                                                               |

## 2.3 Lợi ích của Data mart

18

- » Dễ dàng truy cập dữ liệu khi cần thiết
- » Tạo một cái nhìn chung cho cùng một nhóm người dùng
- » Cải thiện thời gian phản hồi cho người dùng cuối
- » Chi phí thấp hơn so với việc triển khai một data warehouse đầy đủ
- » Người dùng tiềm năng được xác định rõ ràng hơn so với trong data warehouse
- » Chỉ chứa dữ liệu thiết yếu của doanh nghiệp và gọn gàng hơn

## 2.4 Phân loại Data mart

19

- » **Dependent Data Mart:** Data Mart phụ thuộc chứa những dữ liệu được lấy từ Data Warehouse 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 Data Mart.
- » **Independent Data Mart:** Data Mart độc lập không giống với Data Mart phụ thuộc. Data Mart độc lập được xây dựng trước Data Warehouse và dữ liệu được lấy trực tiếp từ các nguồn khác nhau của chúng ít hơn so với Kho dữ liệu.



### *3. Quản lý dữ liệu chủ*

## *Master Data Management*

- » Trong kinh doanh, quản lý dữ liệu chủ (Master data management viết tắt MDM) là một phương pháp được sử dụng để xác định và quản lý dữ liệu quan trọng của một tổ chức để cung cấp, với tích hợp dữ liệu, một điểm tham chiếu duy nhất.
- » Dữ liệu được làm chủ có thể bao gồm:
  - » Dữ liệu tham chiếu - các đối tượng kinh doanh cho các giao dịch và các thứ nguyên để phân tích
  - » Dữ liệu phân tích - hỗ trợ việc ra quyết định

- » Trong lĩnh vực công nghệ, một công cụ quản lý dữ liệu chính có thể được sử dụng để hỗ trợ quản lý dữ liệu chính bằng cách loại bỏ các bản sao, chuẩn hóa dữ liệu (duy trì hàng loạt) và kết hợp các quy tắc để loại bỏ dữ liệu không chính xác xâm nhập vào hệ thống nhằm tạo ra một nguồn dữ liệu chủ có thẩm quyền
- » Dữ liệu được làm chủ có thể bao gồm:
  - » Dữ liệu tham chiếu - các đối tượng kinh doanh cho các giao dịch và các thứ nguyên để phân tích
  - » Dữ liệu phân tích - hỗ trợ việc ra quyết định

- » **Master Data** là dạng dữ liệu có **mức độ tinh cao**, có nghĩa là nó **hiếm khi thay đổi** và hầu hết có **giá trị dài hạn**. Vì Master Data được sử dụng bởi nhiều phòng ban của công ty nên nó rất liên quan đến tất cả các quy trình nghiệp vụ của doanh nghiệp.
- » Master Data chính là một trong những tiêu chí được sử dụng trong quá trình thống kê và đánh giá dữ liệu. Master Data là chìa khóa, cốt lõi của tổ chức.

## 3.2 Quy trình trong quản lý dữ liệu tổng thể

24

- » Xác định nguồn và thu thập dữ liệu,
- » Chuyển đổi dữ liệu, chuẩn hóa
- » Quản trị quy tắc
- » Phát hiện và sửa lỗi
- » Hợp nhất dữ liệu
- » Lưu trữ, phân loại và phối dữ liệu
- » Tạo mục tổng thể
- » Ánh xạ lược đồ, sản phẩm mã hóa,
- » Làm giàu dữ liệu và quản trị dữ liệu.



### 3.3 Quá trình truyền tải của Master Data

25

Một số cách mà dữ liệu chính có thể được đối chiếu và phân phối đến các hệ thống khác. Bao gồm

- » Hợp nhất dữ liệu: Quá trình thu thập dữ liệu chủ từ nhiều nguồn và tích hợp vào một trung tâm duy nhất (Data warehouse hoạt động) để sao chép sang các hệ thống đích khác.
- » Liên kết dữ liệu : Quá trình cung cấp một chế độ xem ảo duy nhất của dữ liệu chủ từ một hoặc nhiều nguồn tới một hoặc nhiều hệ thống đích.
- » Truyền dữ liệu - Quá trình sao chép dữ liệu chính từ hệ thống này sang hệ thống khác, thường thông qua các kết nối điểm-điểm trong các hệ thống kế thừa.



## *4. Dimension and Fact*

## 4.1 Dimension table

27

- » Bảng dimension cung cấp các thông tin, ngữ cảnh cho bảng fact và do đó cũng là cung cấp cho tất cả số liệu thể hiện trong data warehouse.
- » Chức năng của dimension:
  - » Cung cấp bộ lọc
  - » Phân nhóm
  - » Ghi nhận kết quả truy vấn

## 4.1 Dimension table

28

- » Thành phần của dimension table:
  - » **Khóa thay thế (surrogate key)** : Khoá chính của bảng dimension là trường dữ liệu (thường là kiểu số) lưu những giá trị duy nhất, không có ý nghĩa
  - » **Khóa tự nhiên (natural key)**: khoá chính của dữ liệu trong hệ thống nghiệp vụ, giá trị của khóa tự nhiên thường không phải vô nghĩa.
  - » **Thuộc tính mô tả (descriptive attribute)** : thuộc tính mô tả có thể ở nhiều kiểu dữ liệu khác nhau và số lượng có thể rất lớn.

## 4.1 Dimension table

29

| Loại mặt hàng |                                |
|---------------|--------------------------------|
| <b>PK</b>     | <b><u>Mã loại mặt hàng</u></b> |
|               | Tên loại mặt hàng              |



| Mặt hàng   |                                          |
|------------|------------------------------------------|
| <b>PK</b>  | <b><u>Mã mặt hàng</u></b>                |
|            | Tên mặt hàng<br>Đơn giá<br>Loại mặt hàng |
| <b>FK1</b> | Mã loại mặt hàng                         |

| Mặt hàng |                                          |
|----------|------------------------------------------|
|          | <b><u>Mã mặt hàng</u></b>                |
|          | Tên mặt hàng<br>Đơn giá<br>Loại mặt hàng |

## 4.2 Fact table

30

- » Một bảng sự kiện (Fact table) là một bảng gồm nhiều phép đo (measure), số liệu (metrics) hoặc sự kiện (fact) của quy trình kinh doanh (business process).
- » Mỗi bảng fact được xác định bằng độ chi tiết của bảng dữ liệu .
- » Bảng fact thông thường không có khoá chính riêng, mà chứa một tập các khoá ngoại để kết nối đến bảng dimension
- » Một vài dimension đặc biệt gọi là **dimension thoái hoá (degenerate dimension)**, chính là các dimension có tồn tại trong bảng fact nhưng không nhất thiết phải tạo ra bảng dimension cho nó.

## 4.2 Fact table

31

Phân loại bảng fact:

- » **Bảng fact mức chi tiết (transactional grain fact table):**
  - » Bảng có quy mô lớn nhất (về lượng bản ghi) và chi tiết nhất
  - » Chứa nhiều khóa ngoại hơn các loại khác vì có mối quan hệ với tất cả các thứ nguyên có thể có
  - » Đây là những bảng dữ kiện thường được sử dụng nhất
- » **Bảng fact tổng hợp thường kỳ (periodic snapshot fact table)**
  - » Tóm tắt hoạt động trong một khoảng thời gian, có thể là tháng, năm hoặc tuần.
  - » Lặp lại sau mỗi chu kỳ
- » **Bảng fact lũy kế (accumulating snapshot fact table)**
  - » lưu lại số liệu có thời gian không định trước

## 4.3 Các dạng lưu trữ Dimension

32

Trong DW, Bảng chiều *dimension* cung cấp thông tin, ngữ cảnh cho bảng fact. Dù có quy mô nhỏ hơn fact nhiều lần, nhưng dimension là trọng tâm của DW, vì nếu thiếu nó, các số liệu trong fact sẽ không mang lại nhiều ý nghĩa cho người dùng.

### **Type 0: Retain Original**

- » Đây là phương thức thụ động. Khi có sự tự thay đổi trên dimension, không có hành động nào xảy ra trên DW. Giá trị dimension được giữ nguyên như tại thời thời điểm bản ghi được thêm vào lần đầu tiên.



## 4.3 Các dạng lưu trữ Dimension

33

### Type 1: Over write

» Ghi đè dữ liệu cũ, không theo dõi dữ liệu lịch sử. Ví dụ:

| Supplier_Key | Supplier_Code | Supplier_Name  | Supplier_State |
|--------------|---------------|----------------|----------------|
| 123          | ABC           | Acme Supply Co | CA             |

Khi Nhà cung cấp đổi địa điểm Dữ liệu được cập nhật như sau:

| Supplier_Key | Supplier_Code | Supplier_Name  | Supplier_State |
|--------------|---------------|----------------|----------------|
| 123          | ABC           | Acme Supply Co | IL             |

Ưu điểm: Dễ bảo trì.

Nhược điểm: Không kiểm tra được dữ liệu lịch sử.

## 4.3 Các dạng lưu trữ Dimension

34

### Type 2: Add new row

- » Theo dõi dữ liệu lịch sử bằng cách tạo nhiều bản ghi cho một khoá tự nhiên trong dimension (phân biệt bằng khoá tự nhiên hoặc số phiên bản)
- » Ưu điểm: Lưu trữ giá trị lịch sử không giới hạn.

VD: Dùng version

| Supplier_Key | Supplier_Code | Supplier_Name  | Supplier_State | Version. |
|--------------|---------------|----------------|----------------|----------|
| 123          | ABC           | Acme Supply Co | CA             | 0        |
| 124          | ABC           | Acme Supply Co | IL             | 1        |

Thêm cột ngày có hiệu lực

| Supplier_Key | Supplier_Code | Supplier_Name  | Supplier_State | Start_Date  | End_Date    |
|--------------|---------------|----------------|----------------|-------------|-------------|
| 123          | ABC           | Acme Supply Co | CA             | 01-Jan-2000 | 21-Dec-2004 |
| 124          | ABC           | Acme Supply Co | IL             | 22-Dec-2004 | NULL        |

## 4.3 Các dạng lưu trữ Dimension

35

### Type 3: Add new attribute

- » Phương pháp này theo dõi các thay đổi bằng cách sử dụng các cột riêng biệt và lưu giữ lịch sử hạn chế.
- » Cấu trúc bảng gốc ở Loại 1 và Loại 2 giống nhau nhưng Loại 3 có thêm các cột bổ sung.

VD:

| Supplier_<br>Key | Supplier_<br>Code | Supplier_<br>Name | Original_Supplier_State | Effective_<br>Date | Current_Supplier_<br>State |
|------------------|-------------------|-------------------|-------------------------|--------------------|----------------------------|
| 123              | ABC               | Acme Supply<br>Co | CA                      | 22-Dec-2004        | IL                         |

## 4.3 Các dạng lưu trữ Dimension

36

### Type 4: Add history table

- » Phương pháp này sử dụng các “history table”. Bao gồm:
  - » Một bảng lưu trữ giá trị hiện tại
  - » Một bảng lưu trữ tất cả thay đổi
  - » Cả hai khóa thay thế đều được tham chiếu trong bảng Fact để nâng cao hiệu suất truy vấn.

VD:

| Supplier     |               |                          |                |
|--------------|---------------|--------------------------|----------------|
| Supplier_key | Supplier_Code | Supplier_Name            | Supplier_State |
| 124          | ABC           | Acme & Johnson Supply Co | IL             |

| Supplier_History |               |                          |                |              |
|------------------|---------------|--------------------------|----------------|--------------|
| Supplier_key     | Supplier_Code | Supplier_Name            | Supplier_State | Create_Date  |
| 123              | ABC           | Acme Supply Co           | CA             | 14-June-2003 |
| 124              | ABC           | Acme & Johnson Supply Co | IL             | 22-Dec-2004  |

## 4.3 Các dạng lưu trữ Dimension

37

### Type 6: Hybrid

» Phương pháp kết hợp của 1,2,3

VD: Khi bản ghi lần đầu tiên được thêm vào. Giá trị lịch sử và giá trị hiện tại giống nhau.

| Supplier_<br>Key | Row_<br>Key | Supplier_<br>Code | Supplier_<br>Name | Current_<br>State | Historical_<br>State | Start_<br>Date  | End_<br>Date    | Current_<br>Flag |
|------------------|-------------|-------------------|-------------------|-------------------|----------------------|-----------------|-----------------|------------------|
| 123              | 1           | ABC               | Acme Supply<br>Co | CA                | CA                   | 01-Jan-<br>2000 | 31-Dec-<br>2009 | Y                |

Khi Supplier State thay đổi:

| Supplier_<br>Key | Row_<br>Key | Supplier_<br>Code | Supplier_<br>Name | Current_<br>State | Historical_<br>State | Start_<br>Date  | End_<br>Date    | Current_<br>Flag |
|------------------|-------------|-------------------|-------------------|-------------------|----------------------|-----------------|-----------------|------------------|
| 123              | 1           | ABC               | Acme Supply<br>Co | IL                | CA                   | 01-Jan-<br>2000 | 21-Dec-<br>2004 | N                |
| 123              | 2           | ABC               | Acme Supply<br>Co | IL                | IL                   | 22-Dec-<br>2004 | 31-Dec-<br>2009 | Y                |

## 4.3 Các dạng lưu trữ Dimension

38

Khi Supplier State tiếp tục thay đổi:

| Supplier_<br>Key | Row_<br>Key | Supplier_<br>Code | Supplier_<br>Name | Current_<br>State | Historical_<br>State | Start_<br>Date  | End_<br>Date    | Current_<br>Flag |
|------------------|-------------|-------------------|-------------------|-------------------|----------------------|-----------------|-----------------|------------------|
| 123              | 1           | ABC               | Acme Supply<br>Co | NY                | CA                   | 01-Jan-<br>2000 | 21-Dec-<br>2004 | N                |
| 123              | 2           | ABC               | Acme Supply<br>Co | NY                | IL                   | 22-Dec-<br>2004 | 03-Feb-<br>2008 | N                |
| 123              | 3           | ABC               | Acme Supply<br>Co | NY                | NY                   | 04-Feb-<br>2008 | 31-Dec-<br>2009 | Y                |



## *5. Extract, Transform, Load*

## 5.1 Tổng quát

40

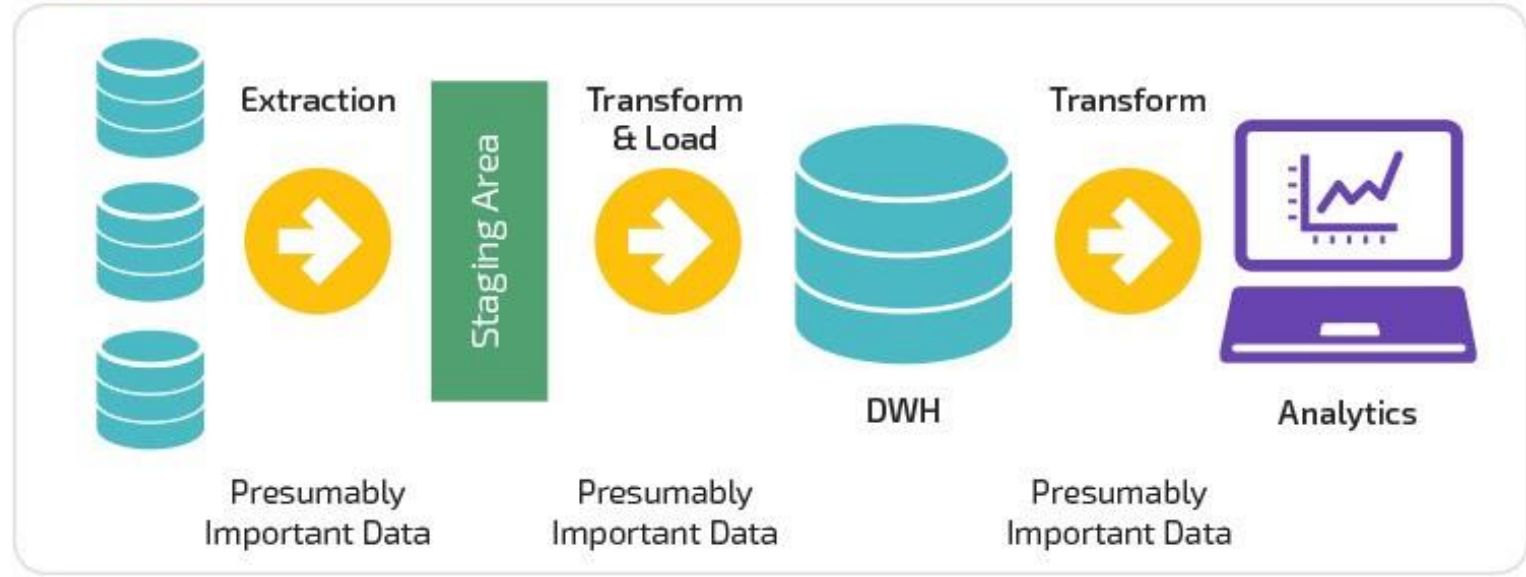
- » Trong máy tính, Extract, Transform, Load (ETL) đề cập đến một quá trình sử dụng cơ sở dữ liệu và đặc biệt là trong kho dữ liệu.
- » Trích xuất dữ liệu là nơi dữ liệu được trích xuất từ các nguồn dữ liệu đồng nhất hoặc không đồng nhất.
- » Chuyển đổi dữ liệu trong đó dữ liệu được chuyển đổi để lưu trữ ở định dạng hoặc cấu trúc thích hợp cho các mục đích truy vấn và phân tích
- » Tải dữ liệu trong đó dữ liệu được tải vào cơ sở dữ liệu đích cuối cùng, cụ thể hơn là kho dữ liệu hoạt động, trung tâm dữ liệu hoặc kho dữ liệu.



## 5.2 Cách thức hoạt động của ETL

41

### ETL



## 5.2.1 Giai đoạn trích xuất - Extract

42

- » Liên quan đến việc trích xuất dữ liệu từ các hệ thống nguồn.
- » Trong bước đầu tiên của quy trình ETL, dữ liệu có cấu trúc và không cấu trúc được nhập và hợp nhất vào một kho lưu trữ duy nhất.

## 5.2.1 Giai đoạn trích xuất - Extract

43

Dữ liệu thô có thể được trích xuất từ nhiều nguồn khác nhau, bao gồm:

- » Cơ sở dữ liệu hiện có
- » Ứng dụng bán hàng và tiếp thị
- » Ứng dụng và thiết bị di động
- » Hệ thống quản lý khách hàng CRM
- » Nền tảng lưu trữ dữ liệu
- » Kho dữ liệu
- » Công cụ phân tích

## 5.2.2 Giai đoạn chuyển đổi - Transform

44

*Ứng dụng ETL trong doanh nghiệp* trong giai đoạn này của quy trình ETL, các quy tắc và quy định có thể được áp dụng để đảm bảo chất lượng dữ liệu và khả năng truy cập.

Quá trình chuyển đổi dữ liệu bao gồm một số quy trình phụ:

- » Data cleansing: nhằm mục đích chỉ truyền dữ liệu “đúng” cho mục tiêu
- » Tiêu chuẩn hóa – quy tắc định dạng được áp dụng cho tập dữ liệu.
- » Loại bỏ dữ liệu trùng lặp – dữ liệu giống nhau được loại trừ hoặc loại bỏ.
- » Xác minh – dữ liệu không thể sử dụng được xóa và dữ liệu dị thường được gắn cờ cảnh báo.
- » Sắp xếp – dữ liệu được sắp xếp theo từng loại.

## 5.2.2 Giai đoạn chuyển đổi - Transform

45

Một hoặc nhiều kiểu chuyển đổi sau có thể được yêu cầu để đáp ứng nhu cầu kinh doanh và kỹ thuật của máy chủ hoặc kho dữ liệu

- » Chỉ chọn một số cột nhất định để tải.
- » Dịch các giá trị được mã hóa.
- » Mã hóa các giá trị dạng tự do.
- » Lấy một giá trị mới được tính toán.
- » Kết hợp dữ liệu từ nhiều nguồn
- » Tổng hợp
- » Tạo giá trị khóa thay thế
- » Chuyển đổi hoặc xoay vòng/Phân tách/Xác nhập các trường dữ liệu

### 5.2.3 Giai đoạn tải - Load

46

- » Tùy thuộc vào yêu cầu của tổ chức, quá trình này rất khác nhau.
- » 3 phương pháp load chính:
  - » **Tải toàn bộ:** toàn bộ dữ liệu đã được trích xuất và chuyển đổi được tải vào khu vực phân tích
  - » **Tải tăng dần:** là quá trình so sánh dữ liệu đã chuyển đổi với dữ liệu ở đích cuối cùng và chỉ tải dữ liệu mới. Tải tăng dần có thể được sử dụng cùng với tải làm mới.
  - » **Tải làm mới:** là quá trình cập nhật dữ liệu ở đích cuối cùng để phản ánh những thay đổi được thực hiện trong nguồn ban đầu. Làm mới có thể đầy đủ hoặc tăng dần.

## 5.3 Chu kỳ của ETL

47

Chu kỳ ETL thực tế diễn hình bao gồm các bước thực hiện sau:

1. Bắt đầu chu kỳ
2. Xây dựng dữ liệu tham khảo
3. Trích (từ các nguồn)
4. Xác thực
5. Chuyển đổi (làm sạch, áp dụng các quy tắc kinh doanh, kiểm tra tính toàn vẹn của dữ liệu, tạo tổng hợp hoặc phân tách)
6. Giai đoạn (tải vào bảng dàn, nếu được sử dụng)

## 5.3 Chu kỳ của ETL

48

7. Báo cáo kiểm toán (ví dụ: về việc tuân thủ các quy tắc kinh doanh. Ngoài ra, trong trường hợp không đạt, giúp chẩn đoán / sửa chữa)
8. Xuất bản (tới các bảng mục tiêu)
9. Lưu trữ



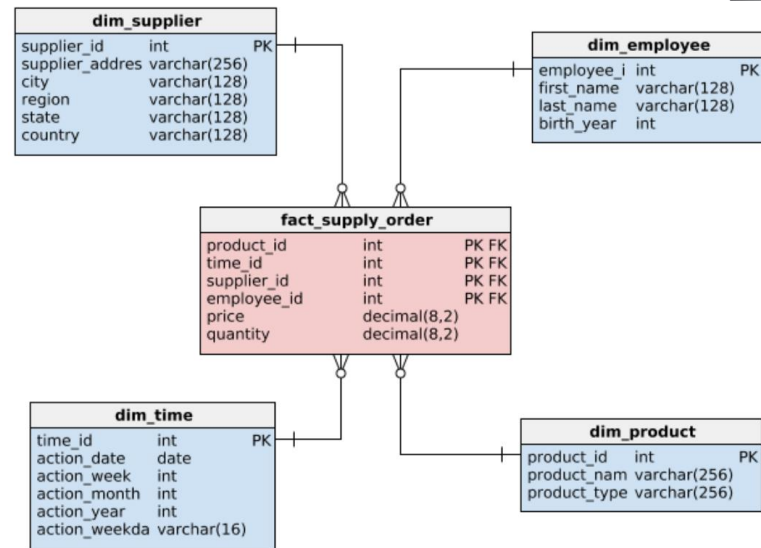


## *6. Style of Data warehouse Schema*

## 6.1 STAR SCHEMA (Lược đồ hình sao)

50

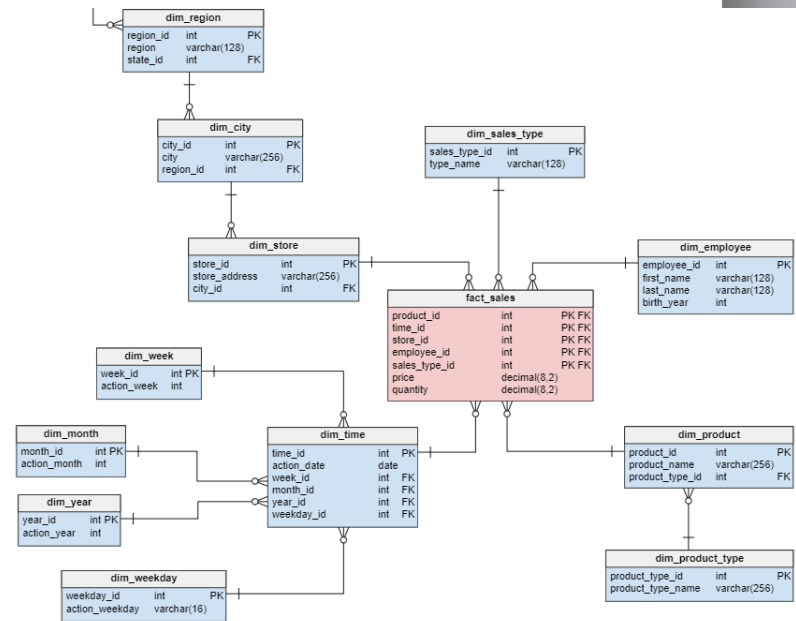
- » Gồm 1 bảng Fact (bảng sự kiện) nằm ở trung tâm và được bao quanh bởi những bảng Dimension (bảng chiều).
- » Là mô hình đơn giản nhất được sử dụng trong DWH



## 6.2 SNOW FLAKE SCHEMA (Lược đồ bông tuyết)

51

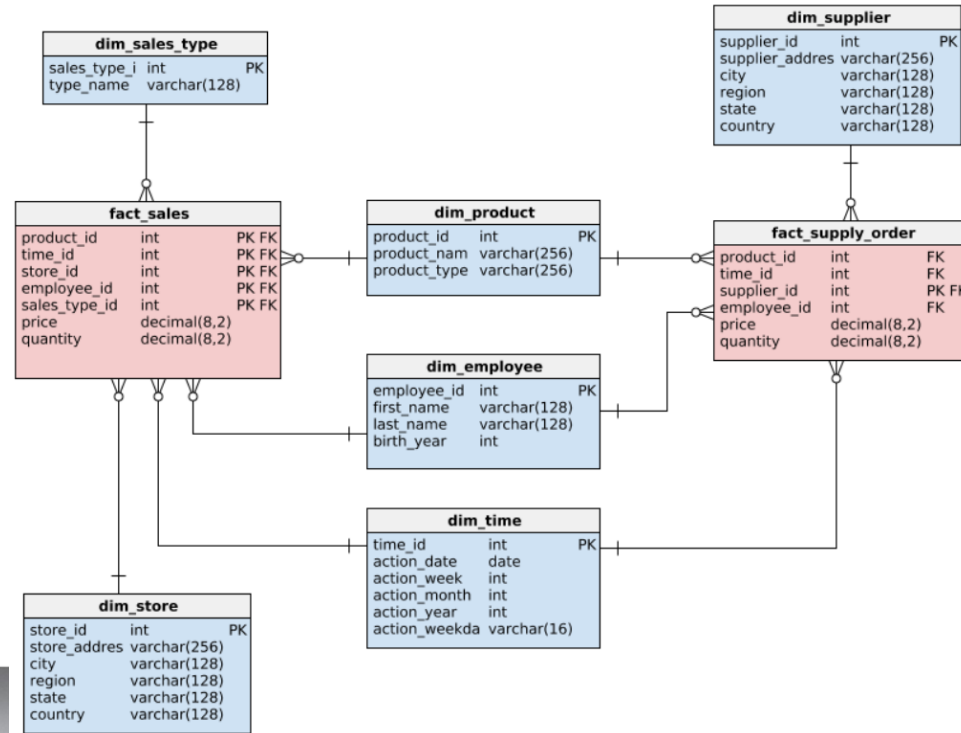
- » Là dạng mở rộng của lược đồ hình sao bằng các bổ sung các Dimension
- » Bảng Fact như lược đồ hình sao, bảng Dimension được chuẩn hoá



## 6.3 GALAXY SCHEMA (Lược đồ thiên hà)

52

- » Chứa nhiều bảng Fact sử dụng chung một số bảng Dimension.
- » Lược đồ là sự kết hợp của nhiều data mart (kho dữ liệu có chủ đề, dạng thu nhỏ của kho dữ liệu, kho dữ liệu được chia thành nhiều phần nhỏ khác nhau).



“