COLE.VN

connecting knowledge

COLE.VN
connecting knowledge

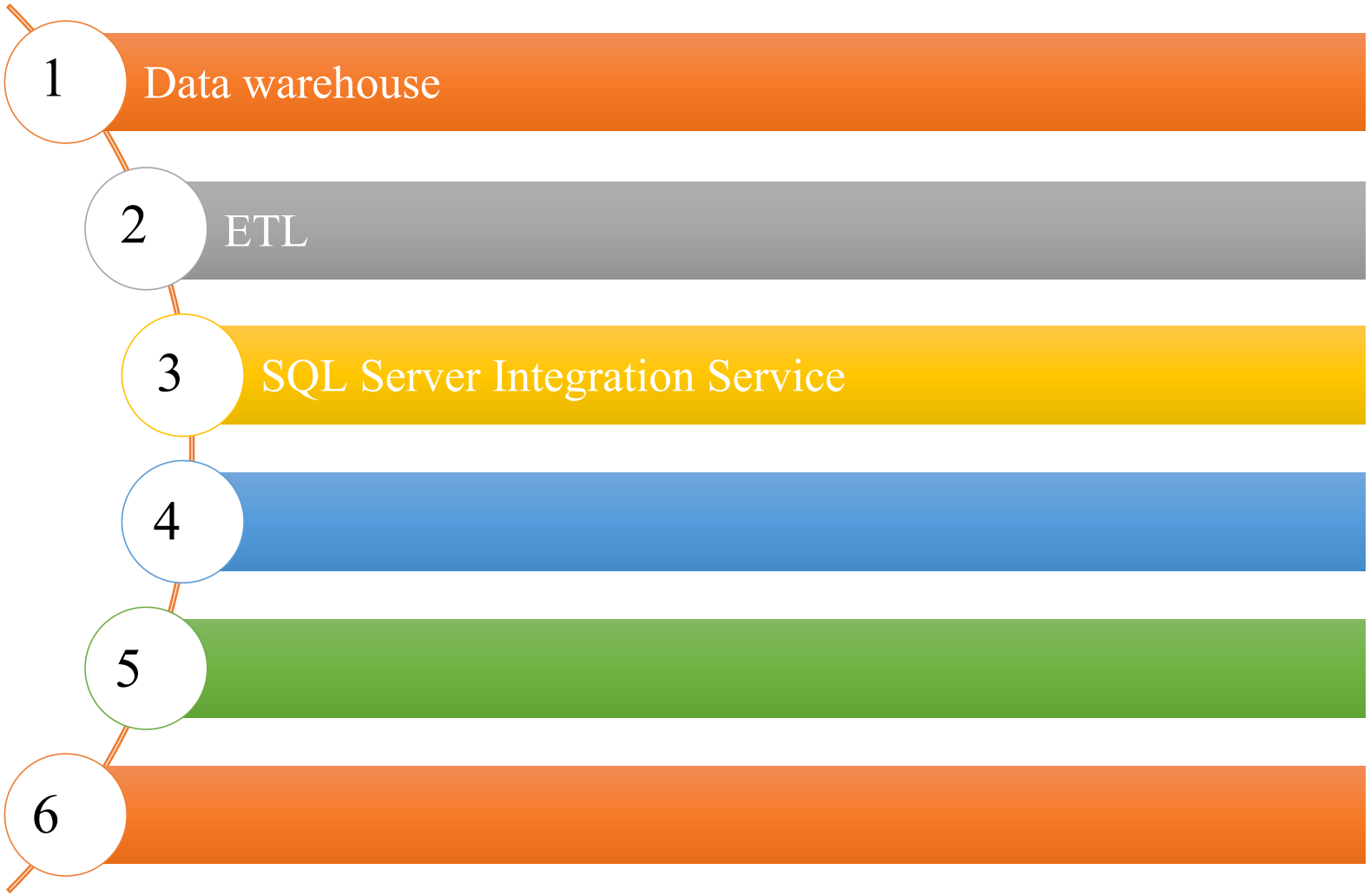
Tổng quan về Data warehouse và ETL

Trình bày: Tạ Minh Tùng
SĐT: 0354.610.796

Module Data warehouse & ETL

- Nội dung
  - Data warehouse
  - ETL
- Mục tiêu
  - Có cái nhìn tổng quan về Data warehouse
  - Xây dựng 1 data warehouse đơn giản
  - Hiểu được cách thức hoạt động của ETL
  - Thực hiện ETL đổ dữ liệu vào Data warehouse đã được xây dựng

NỘI DUNG CHÍNH



Tổng quan Data Warehouse (1)

- Các dữ liệu hoạt động của doanh nghiệp: sản phẩm, bán hàng, mua hàng, tài khoản, khách hàng,… được lưu trữ trong CSDL (thường là CSDL quan hệ)
- Các truy vấn trên dữ liệu : select, insert, update, delete
- Các báo cáo truy vấn trực tiếp vào dữ liệu hoạt động thông qua chuỗi các câu truy vấn
- Chuỗi các câu truy vấn có thể lưu trữ ngay ở CSDL dưới dạng store procedure

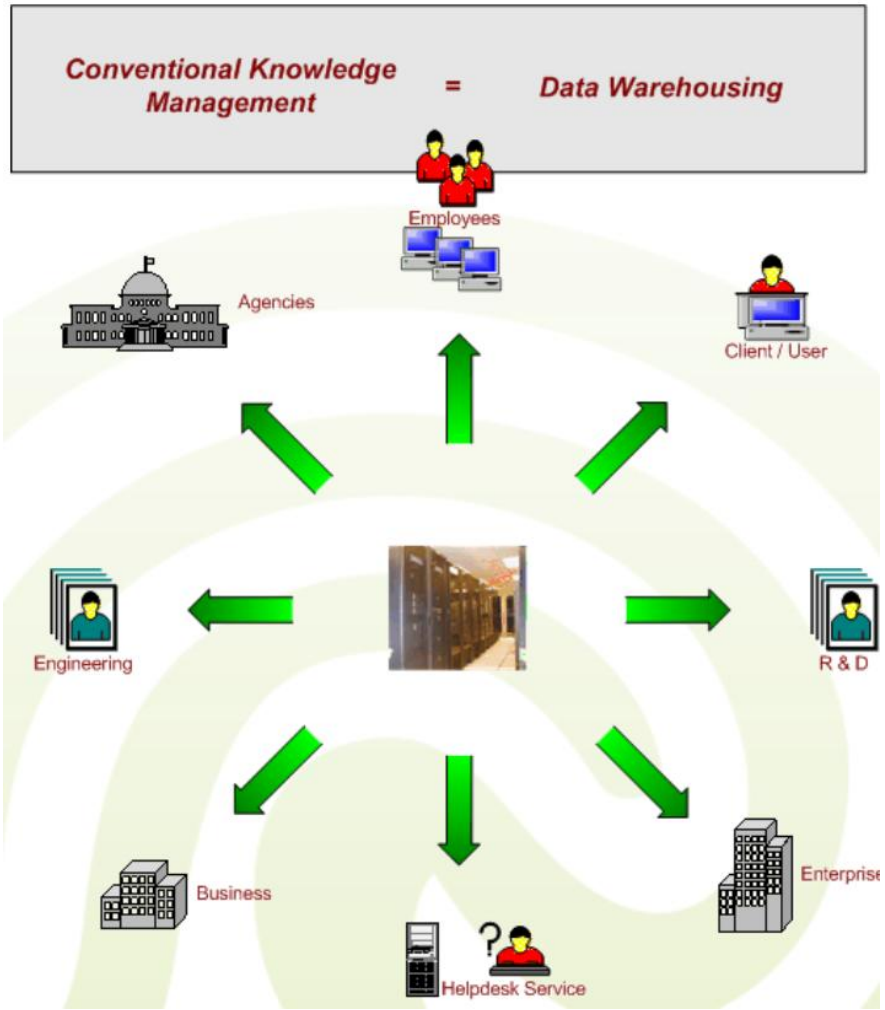
=> Khi dữ liệu nhiều dần lên đến một mức nào đó thì việc truy vấn trực tiếp vào CSDL vận hành doanh nghiệp thì sẽ gặp những vấn đề gì?

Tổng quan Data Warehouse (2)

- Vấn đề:
  - Các câu truy vấn để kết xuất báo cáo thường phải lấy dữ liệu trên nhiều bảng, tổng hợp dữ liệu trên nhiều bản ghi
  - Truy vấn trực tiếp trên dữ liệu hoạt động của doanh nghiệp
  - Doanh nghiệp sử dụng nhiều phần mềm khác nhau

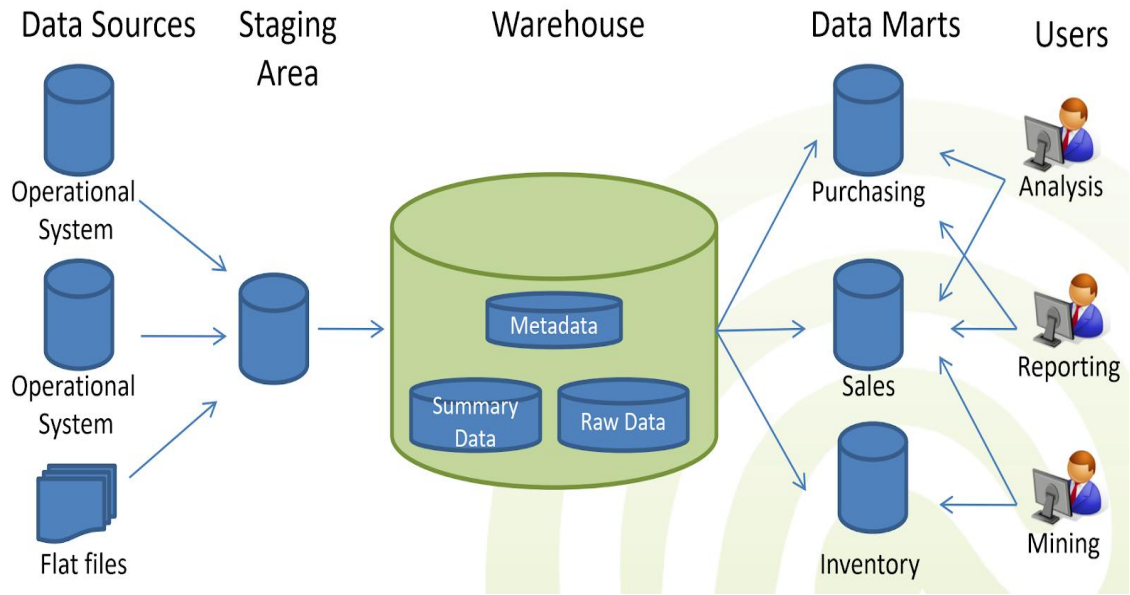
Tổng quan Data Warehouse (3)

- Giải pháp:
  - Tạo CSDL lưu trữ lại kết quả của từng báo cáo
  - Truy vấn trên CSDL mới này
  - **Data Warehouse:**
    - Kho dữ liệu lớn của một tổ chức
    - Được thiết kế đặc biệt cho việc lập báo cáo và phân tích
    - Dữ liệu được tổng hợp từ nhiều nguồn và được đưa vào DWH



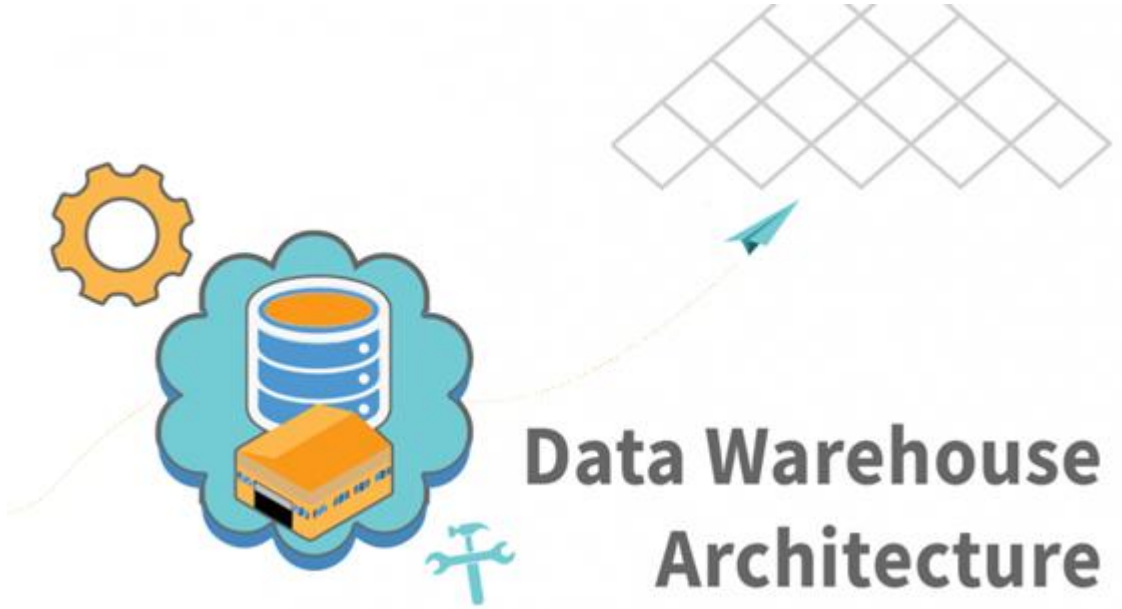
Tổng quan Data Warehouse (4)

- **Nguồn dữ liệu hoạt động** (Operational data store - ODS): Các dữ liệu phục vụ hoạt động hàng ngày của doanh nghiệp như sản phẩm, bán hàng, tài khoản
- **Vùng Stagging**: Chứa các bản copy của dữ liệu được tải vào từ dữ liệu hoạt động
- **Kho dữ liệu** (DWH):
  - Dữ liệu thô đã được làm sạch
  - Dữ liệu dẫn xuất (tổ hợp)
  - Siêu dữ liệu
- **Vùng thể hiện**:
  - Data Mart, dữ liệu được tổ chức theo mục tiêu của một phòng ban
  - Công cụ lập báo cáo và xử lý phân tích



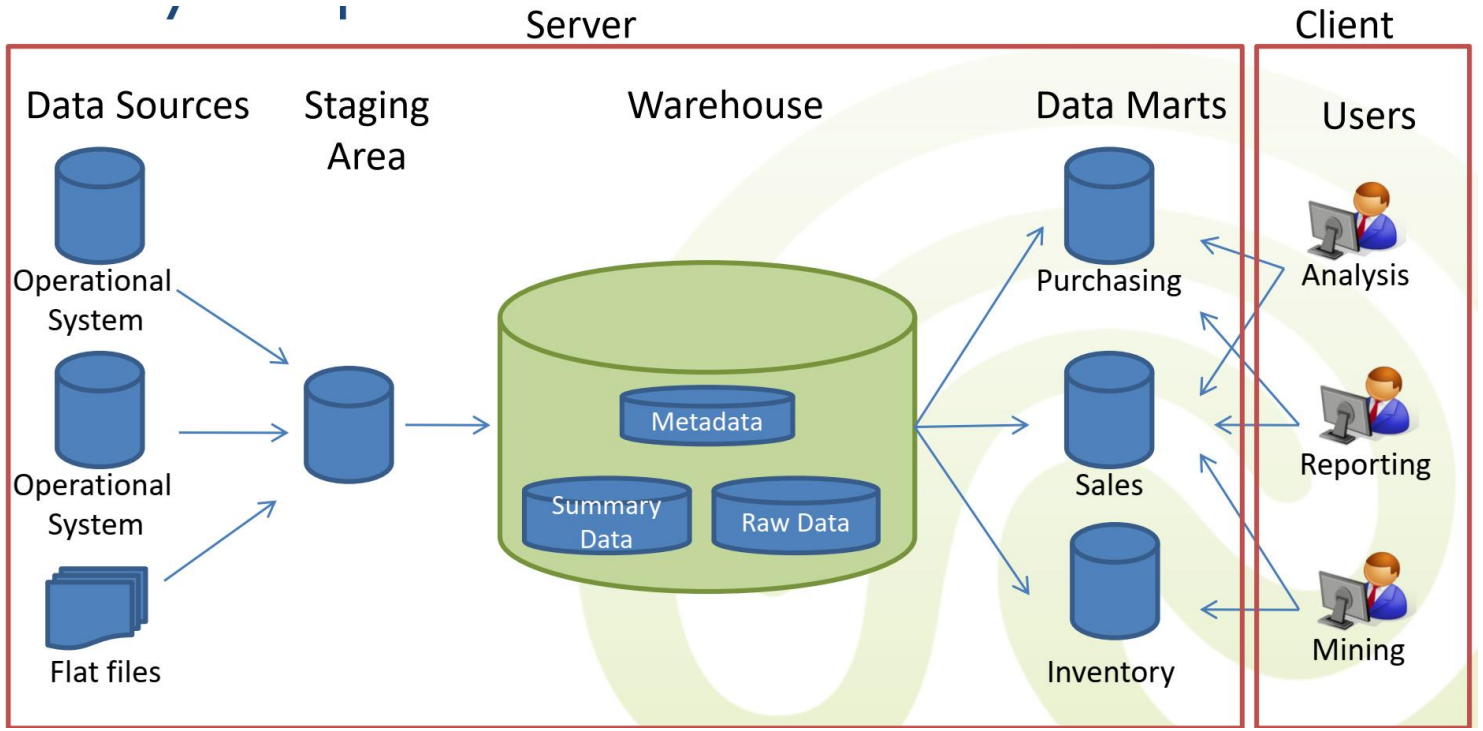
Kiến trúc của Data warehouse (1)

- Các kiến trúc DW phổ biến trong thực tế
  - Kiến trúc 2 tầng
  - Kiến trúc 3 tầng



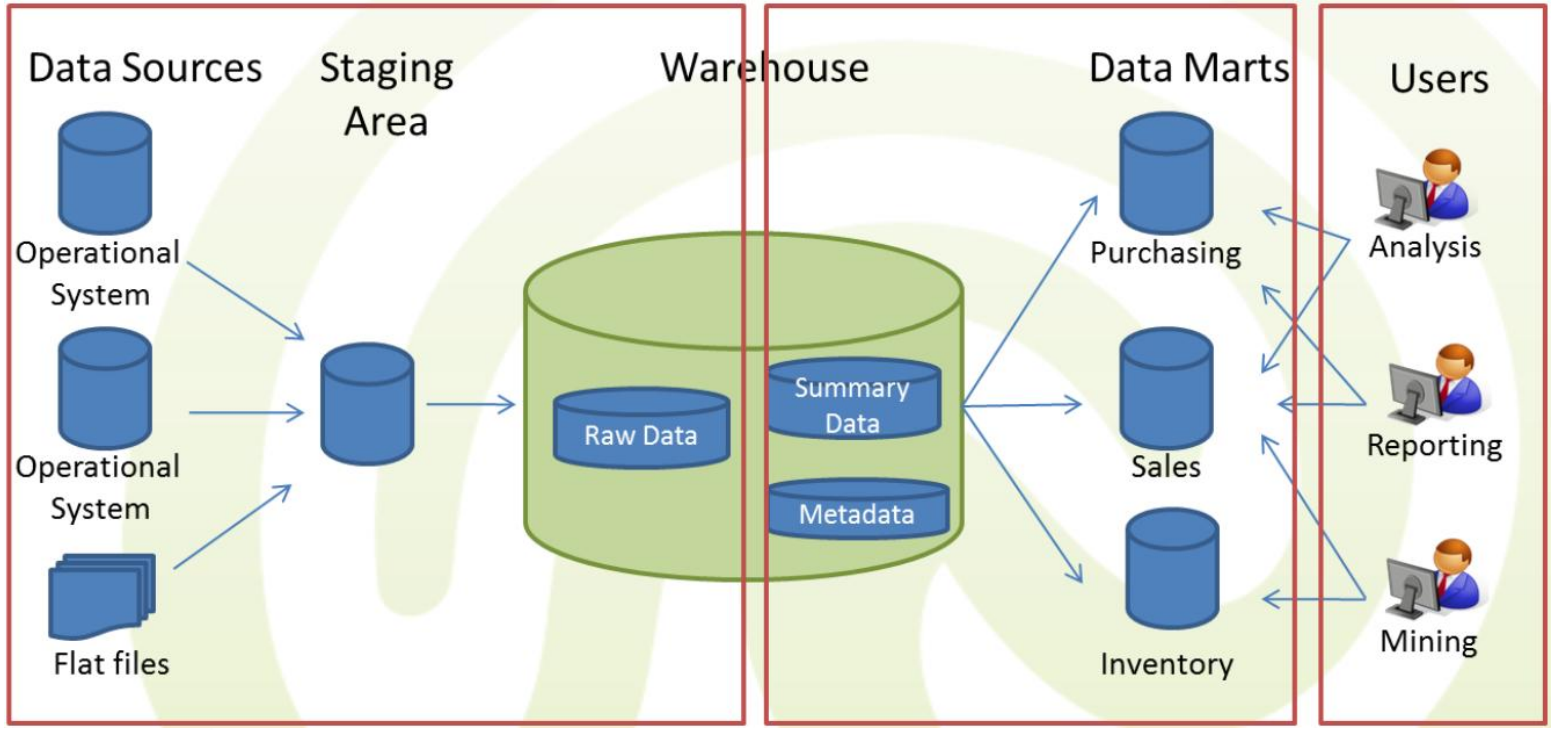
Kiến trúc của Data warehouse (2)

- Kiến trúc 2 tầng Client- Server
  - Thin Client: Server xử lý dữ liệu, client chỉ hiển thị kết quả
  - Fat Client: Server chỉ cung cấp dữ liệu, các phép toán thực hiện trên client



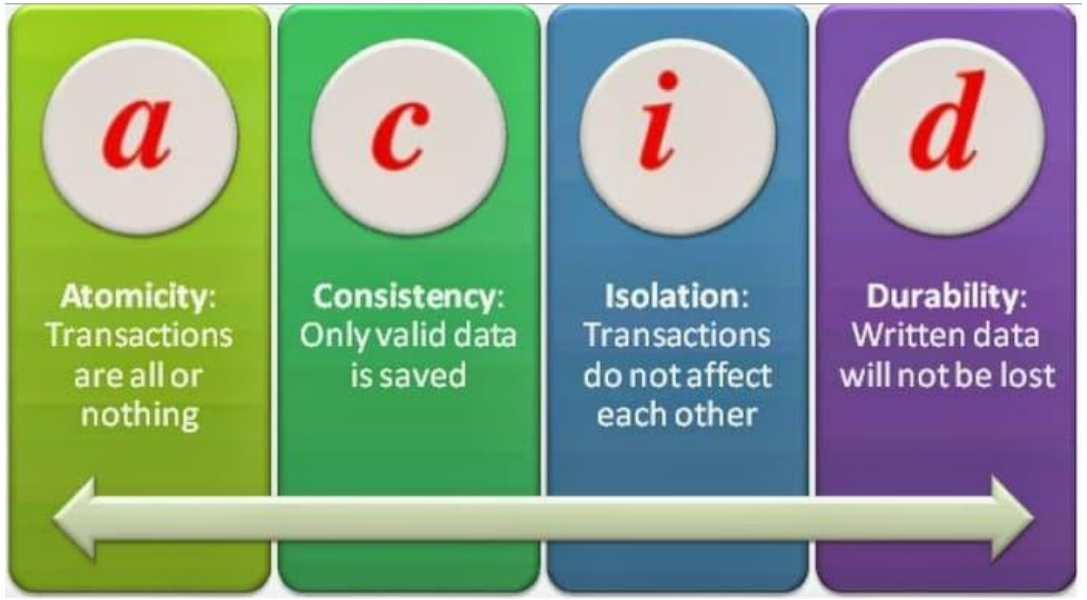
Kiến trúc của Data warehouse (3)

- Kiến trúc 3 tầng
  - Tầng 1: Dữ liệu thô và dữ liệu chi tiết
  - Tầng 2: Dữ liệu dẫn xuất đã được tổ hợp
  - Tầng 3: Công cụ báo cáo và phân tích



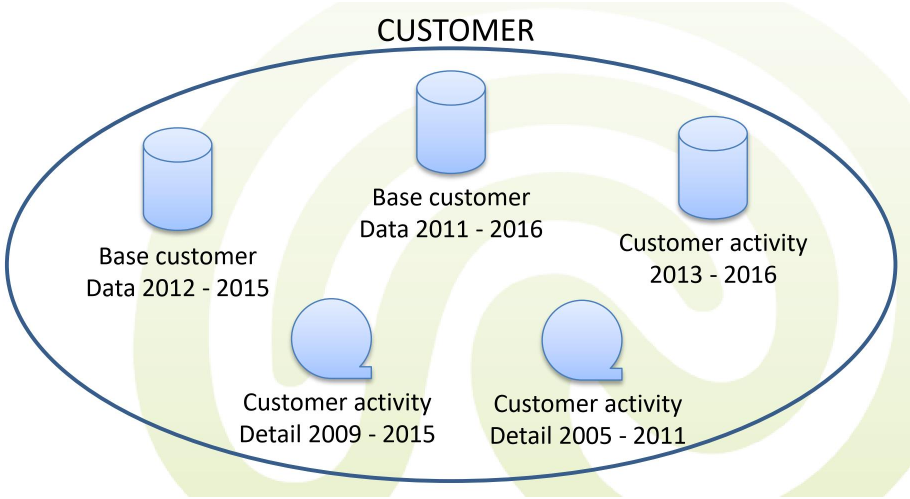
Các đặc tính của Data warehouse

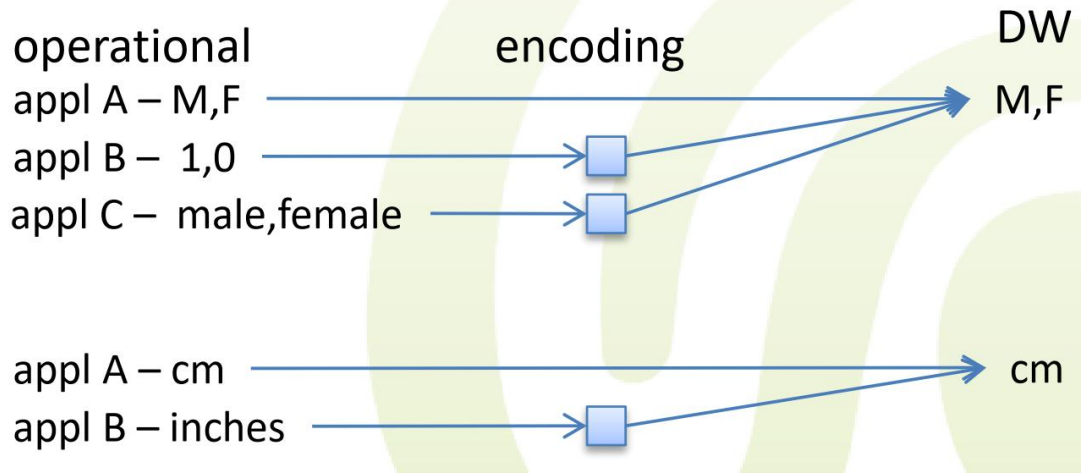
- **Tính nguyên tử** (Atomicity): dữ liệu được tập hợp từ nhiều nguồn khác nhau -> khi tập hợp phải thực hiện làm sạch, sắp xếp, rút gọn dữ liệu
- **Tính nhất quán** (Consistency): chỉ lấy những dữ liệu có ích (các dữ liệu có cùng chủ đề)
- **Tính cô lập** (Isolation): Các dữ liệu truy xuất không bị ảnh hưởng bởi các dữ liệu khác hoặc tác động lên nhau.
- **Tính bền vững** (Durablity): Dữ liệu không thể tạo thêm, xóa hay sửa đổi



Nguyên lý thiết kế Data warehouse (1)

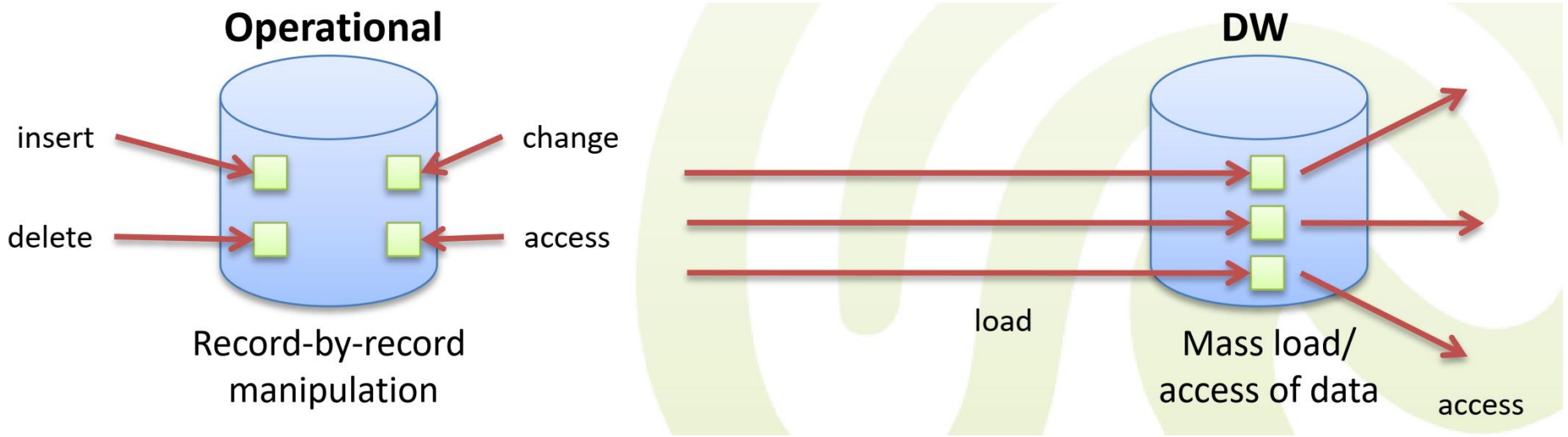
- **Hướng chủ đề**: Loại bỏ các dữ liệu không hữu ích cho quá trình phân tích
- **Tính toàn vẹn**: Tích hợp dữ liệu từ nhiều nguồn khác nhau vào một định dạng thống nhất





Nguyên lý thiết kế Data warehouse (2)

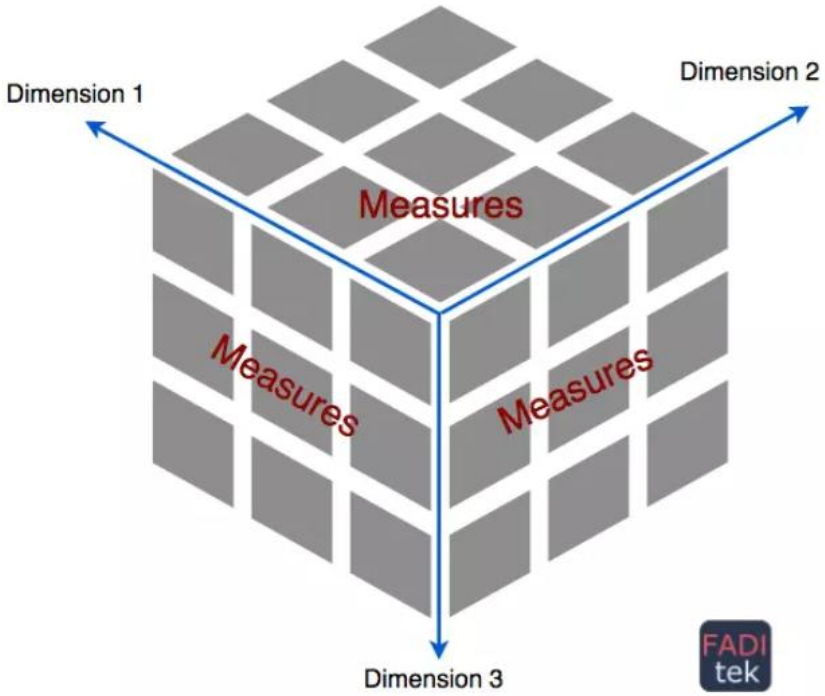
- **Tính bất biến**: Dữ liệu phải thống nhất theo thời gian (Hạn chế đối đa sửa, xoá) >> phân tích thay đổi theo thời gian.
- **Giá trị lịch sử**: Cung cấp dữ liệu tại các thời điểm khác nhau của một thông tin



Cấu trúc Data warehouse (1)

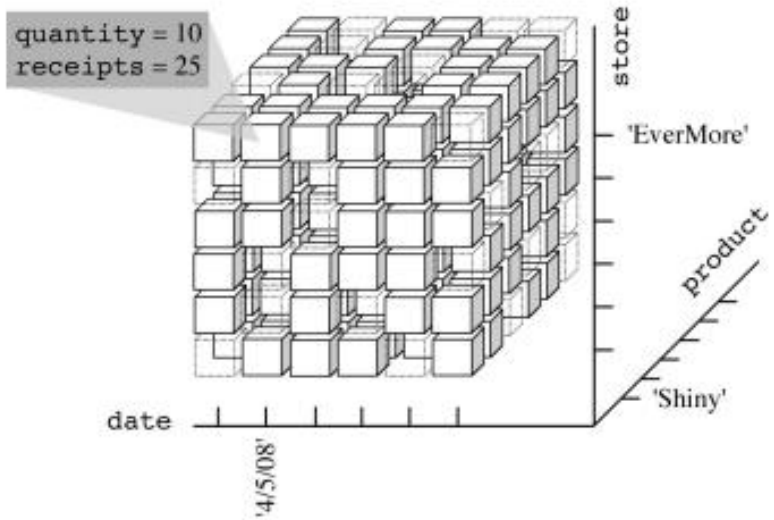
- **Dimension**: Các bảng chứa dữ liệu về các tiêu chí đánh giá
- **Fact/Measure**: Các bảng chứa dữ liệu định lượng cho các tiêu chí

=> Bảng lưu trữ dimension và fact sẽ có gì khác biệt?



Cấu trúc Data warehouse (2)

- Ví dụ: phân tích dimension và fact cho các báo cáo
  - Báo cáo doanh thu theo khách hàng
  - Báo số lượng bán ra theo từng sản phẩm
  - Báo cáo số lượng đơn hàng theo ngày
  - Báo cáo số lượng đơn hàng theo khu vực
  - Báo cáo tỉ lệ đơn hàng online/offline
  - Báo cáo giá trị trung bình của đơn hàng

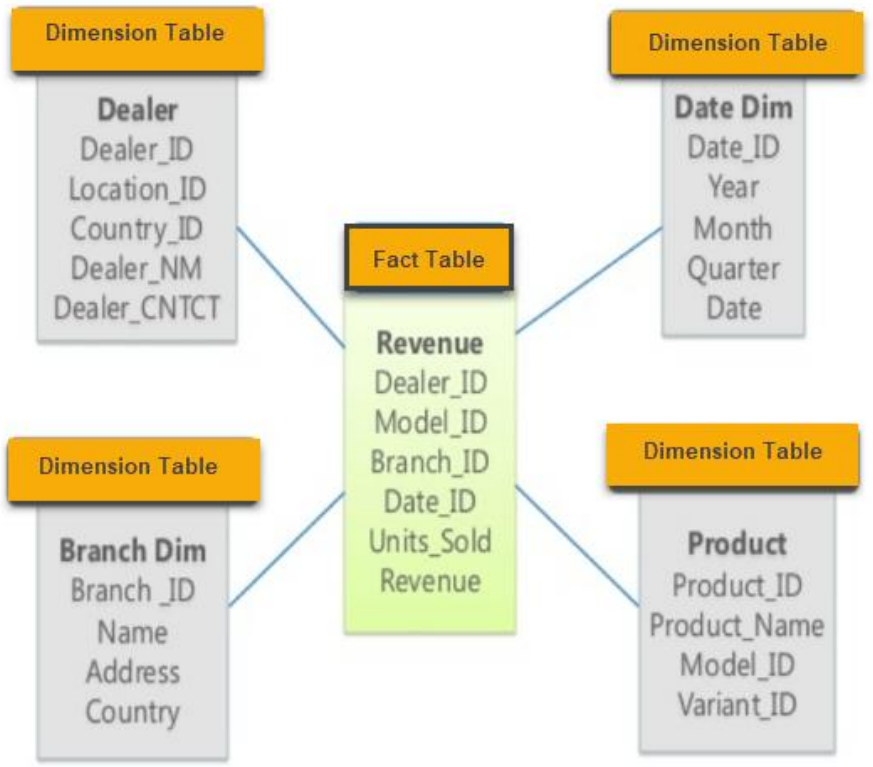


Mô hình hóa Data warehouse (1)

- Các lược đồ
  - Lược đồ hình sao (Star schema)
  - Lược đồ bông tuyết (Snowflake schema)
  - Lược đồ thiên hà (Galaxy Schema)

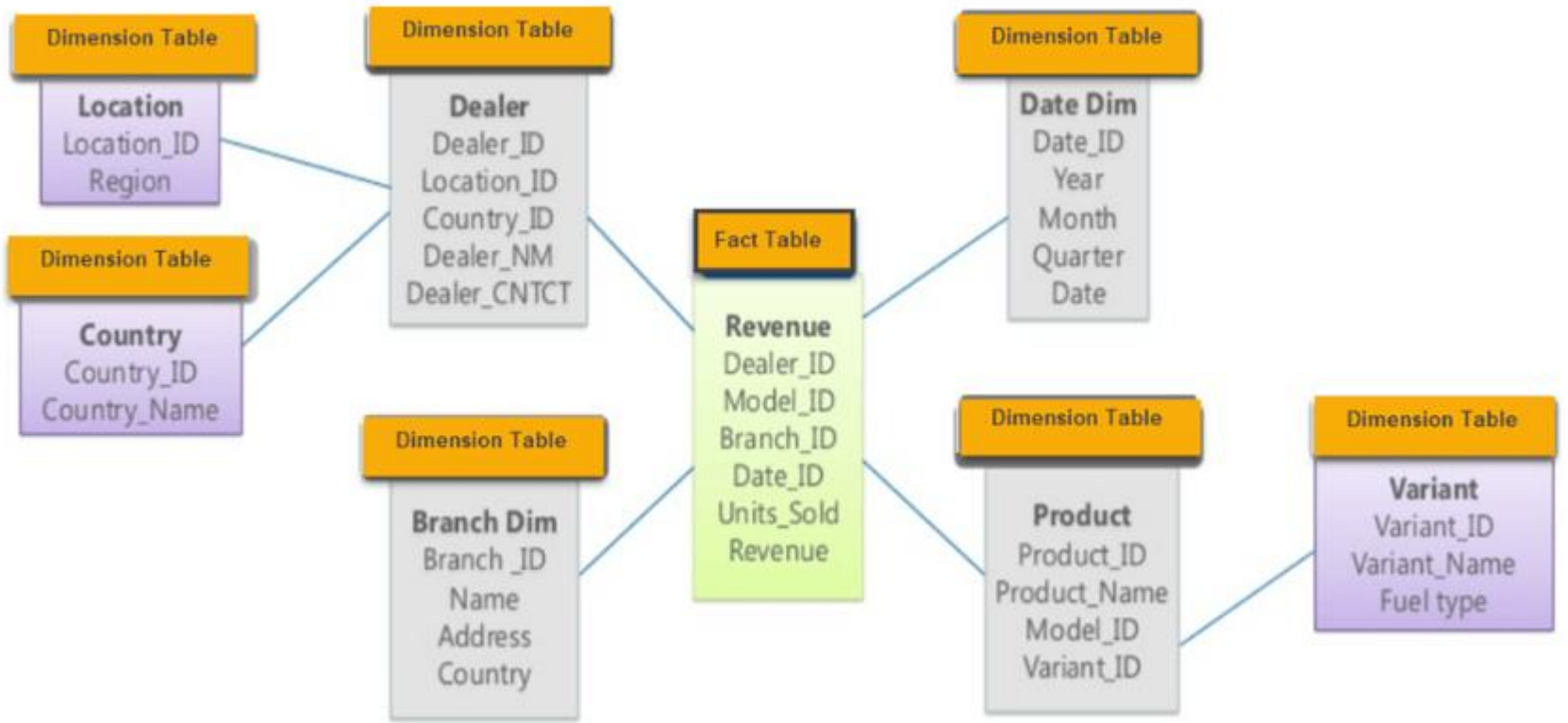
Các lược đồ Data warehouse (2)

- Lược đồ hình sao (star schema)



Các lược đồ Data warehouse (3)

- Lược đồ bông tuyết (snowflake schema)



Các lược đồ Data warehouse (4)

- Lược đồ hình sao và Lược đồ bông tuyết

Ưu nhược điểm của 2 lược đồ này là gì?

- Lưu trữ dữ liệu
- Toàn vẹn dữ liệu
- Mức độ phức tạp của câu truy vấn
- Khả năng mở rộng

Các lược đồ Data warehouse (5)

Lược đồ hình sao	Lược đồ bông tuyết
Bảng dimension không phân cấp	Bảng dimension phân cấp
Ưu điểm: - Cải tiến hiệu năng truy vấn với các dữ liệu thường sử dụng - Ít bảng và cấu trúc đơn giản - Xử lý truy vấn đơn giản trên khía cạnh sử dụng join	Ưu điểm: - Kích thước bảng dimension nhỏ - Dễ bảo trì (tránh dư thừa) - Cho phép các truy vấn phức tạp với các chiều phức tạp, nhiều mức phân lớp
Nhược điểm: - Trong một số trường hợp có sự dư thừa dư thừa lớn	Nhược điểm: - Số lớn các bảng cần được quản lý - Truy cấn có thể cần kết nối nhiều bảng

Các lược đồ Data warehouse (6)

- Snowflake

Product_ID	Description	Brand	Prod_group_ID
10	E71	Nokia	4
11	PS-42A	Samsung	2
12	5800	Nokia	4
	Bold	Berry	4

Prod_group_ID	Description	Prod_categ_ID
2	TV	11
4	Mobile Pho..	11

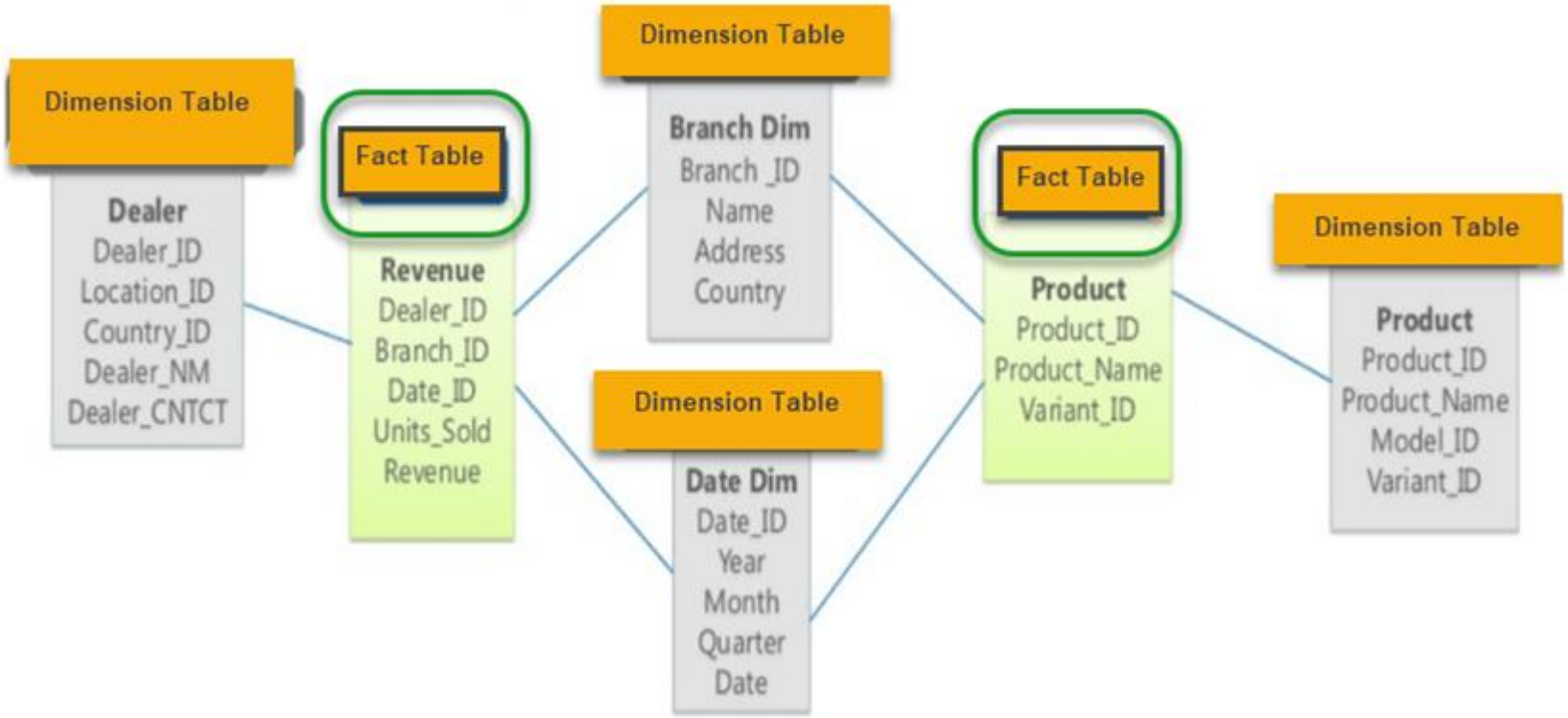
Prod_categ_ID	Description
11	Electronics

- Star

Product_ ID	Description	...	Prod. group	Prod. categ
10	E71	...	Mobile Ph..	Electronics
11	PS-42A	...	TV	Electronics
12	5800		Mobile Ph..	Electronics
13	Bold		Mobile Ph..	Electronics

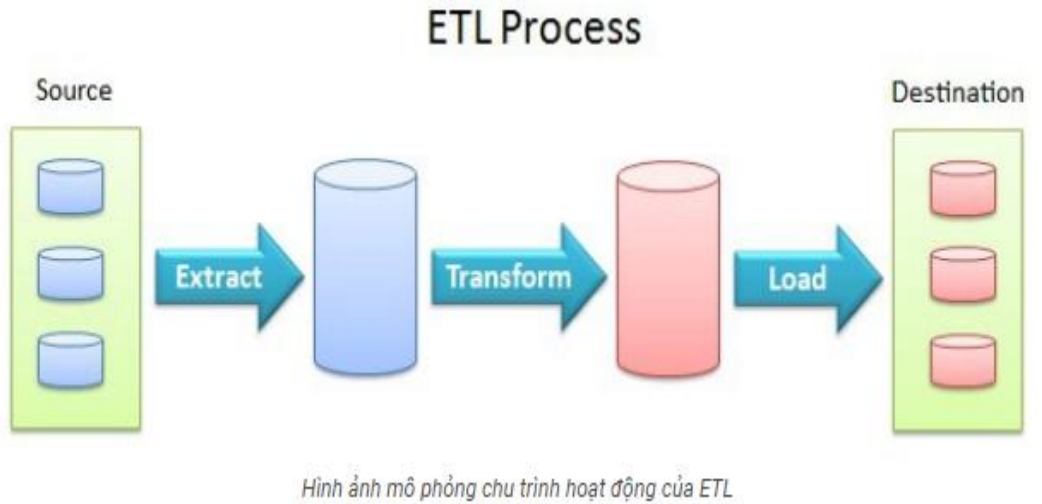
Các lược đồ Data warehouse (7)

- Lược đồ thiên hà (Galaxy schema)



Tổng quan về ETL

- ETL là gì?
  - ETL là quá trình chuyển dữ liệu từ một hay nhiều nguồn đưa vào CSDL đích
  - ETL gồm 3 bước
    - **Trích chọn dữ liệu** (Extract): Quá trình truy cập vào các hệ thống nguồn và trích xuất dữ liệu.
    - **Chuyển đổi dữ liệu** (Transform): quá trình chuyển đổi dữ liệu được trích xuất ở bước extract để phù hợp với các yêu cầu của DWH
    - **Tải dữ liệu** (Load): quá trình ghi chép dữ liệu đã được xử lý ở bước Transform vào cơ sở dữ liệu đích (DWH)



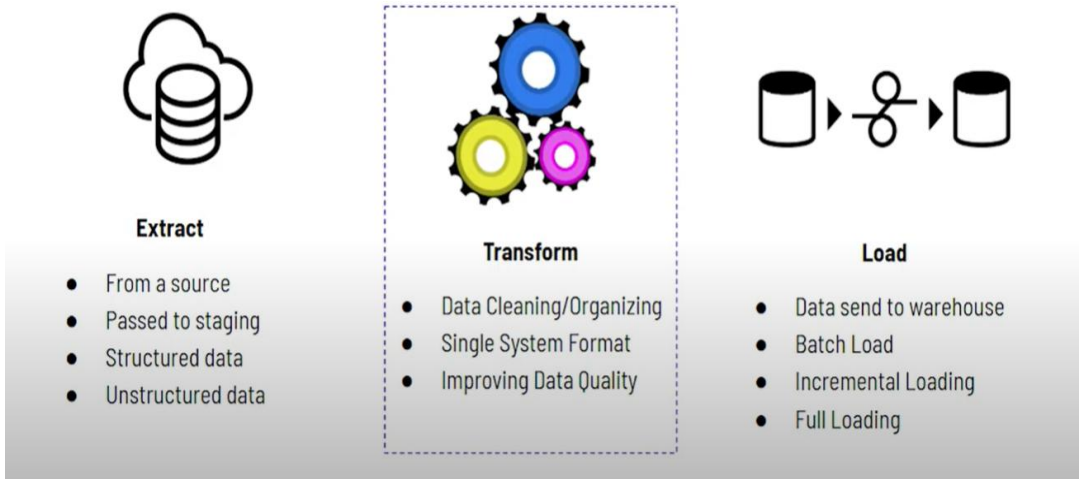
Hình ảnh mô phỏng chu trình hoạt động của ETL

Trích chọn dữ liệu

- Trích chọn dữ liệu (Extract)
  - Dữ liệu được lấy từ các nguồn
  - Sau đó dữ liệu được đẩy vào vùng đệm
  - Trích chọn lần đầu
  - Trích chọn lần sau

Chuyển đổi dữ liệu

- Chuyển đổi dữ liệu
  - Chuẩn hóa dữ liệu
  - Làm sạch dữ liệu
  - Hợp nhất dữ liệu
  - Ghép dữ liệu các cột
  - Lọc dữ liệu
  - Mapping
  - Loại bỏ trùng lặp
  - Tổng hợp dữ liệu



Tải dữ liệu

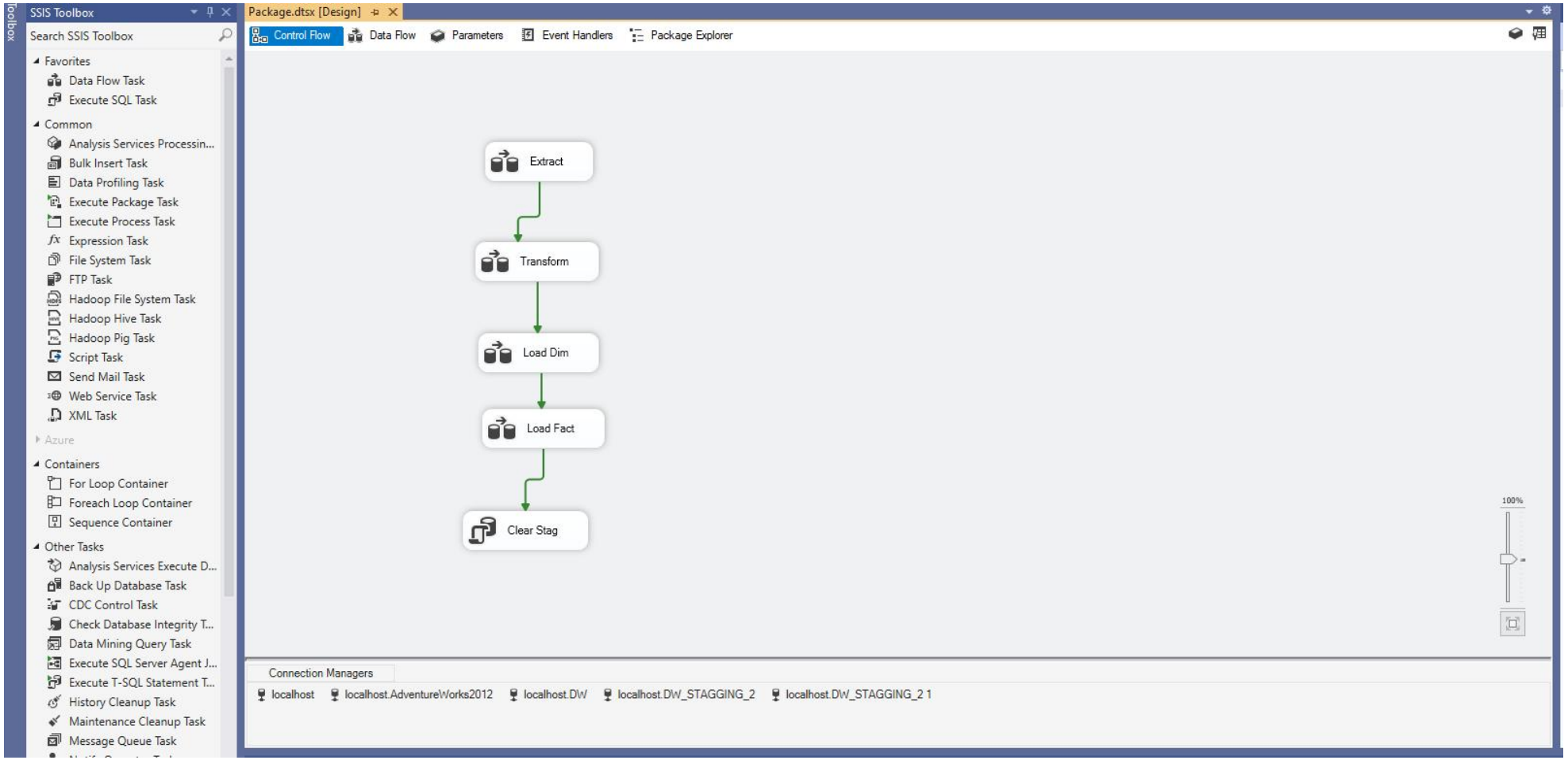
- Tải dữ liệu (Load):
  - Dữ liệu sau khi tổng hợp được đưa vào DWH
  - Tải lần đầu
  - Tải lần sau

ETL Tools

Công cụ ETL	Loại
Pentaho Kettle	Open source
Talend	Open source
Jaspersoft-etl	Open source
Inaplex Inaport	Close source
SQL Server Integration Service	Close source

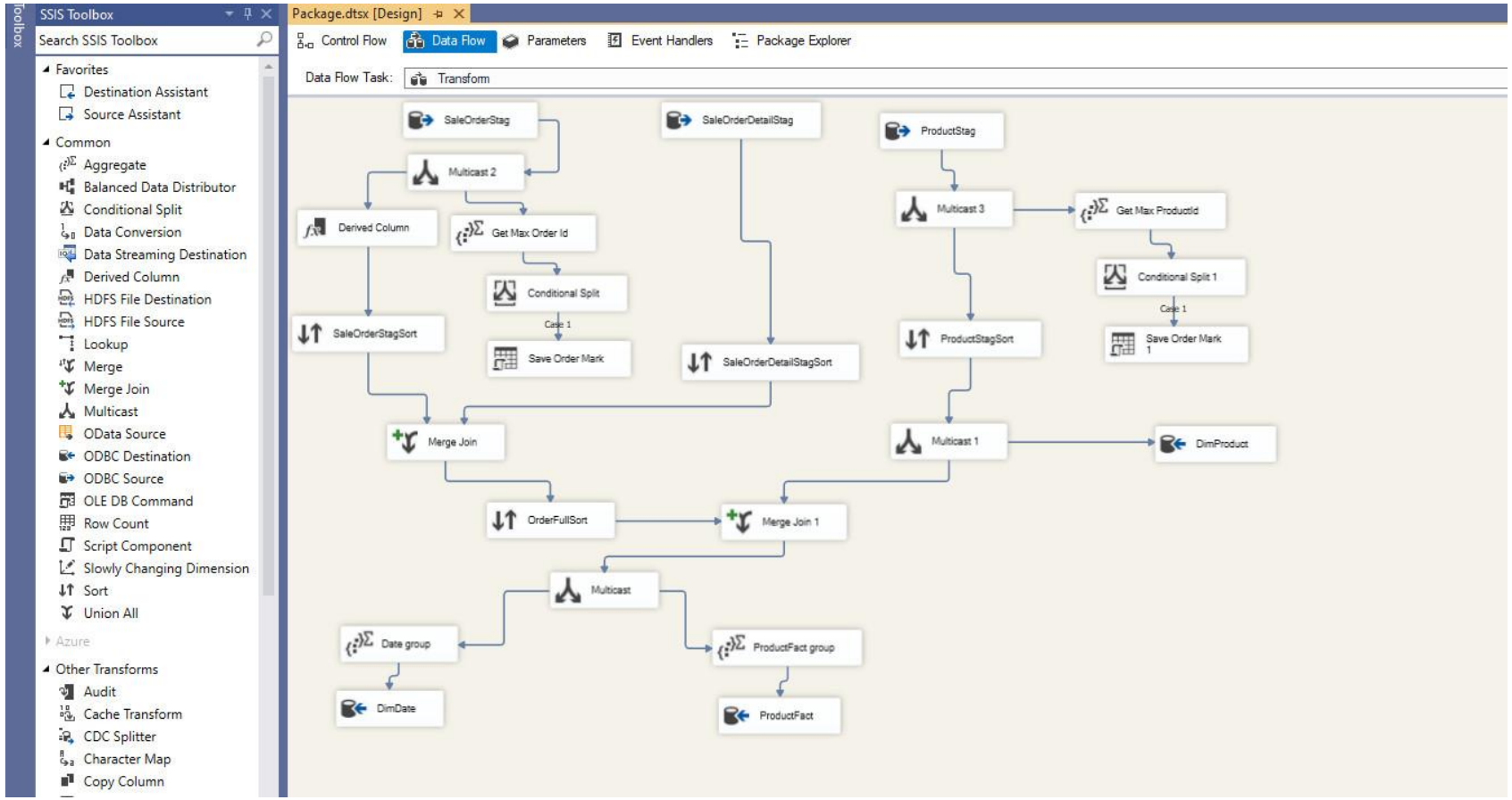
SQL Server Integration Service

- SQL Server Integration Service (SSIS)



SQL Server Integration Service

- SQL Server Integration Service (SSIS)



Cài đặt SQL Server Integration Service

- Cài đặt Visual Studio 2019
- Cài đặt gói SQL Server Data Tools (SSDT)
- https://phanmemnet.com/download-visual-studio-2019-full-professional-enterprise-huong-dan-cai-dat/
- https://www.mssqltips.com/sqlservertip/6481/install-sql-server-integration-services-in-visual-studio-2019/
- https://docs.microsoft.com/en-us/sql/integration-services/install-windows/install-integration-services?view=sql-server-ver16

Bài tập

- Tìm hiểu database AdventureWork2012
  - ✓ Danh sách các bảng dữ liệu, ý nghĩa, số bản ghi có trong từng bảng
  - ✓ Danh sách các trường dữ liệu trong từng bảng:
    - Tên
    - Kiểu dữ liệu
    - Là trường bắt buộc?
    - Ý nghĩa
  - ✓ Diagram thể hiện mối liên hệ giữa các bảng dữ liệu

THANK YOU !

COLE.VN
connecting knowledge