Study

Created time	@April 14, 2025 5:25 PM
Last updated time	@April 23, 2025 7:10 PM

#Các bước để thiết kế datawarehouse cơ bản:

Bước 0: Phân tích chi tiết và hiểu rõ dữ liệu nguồn.

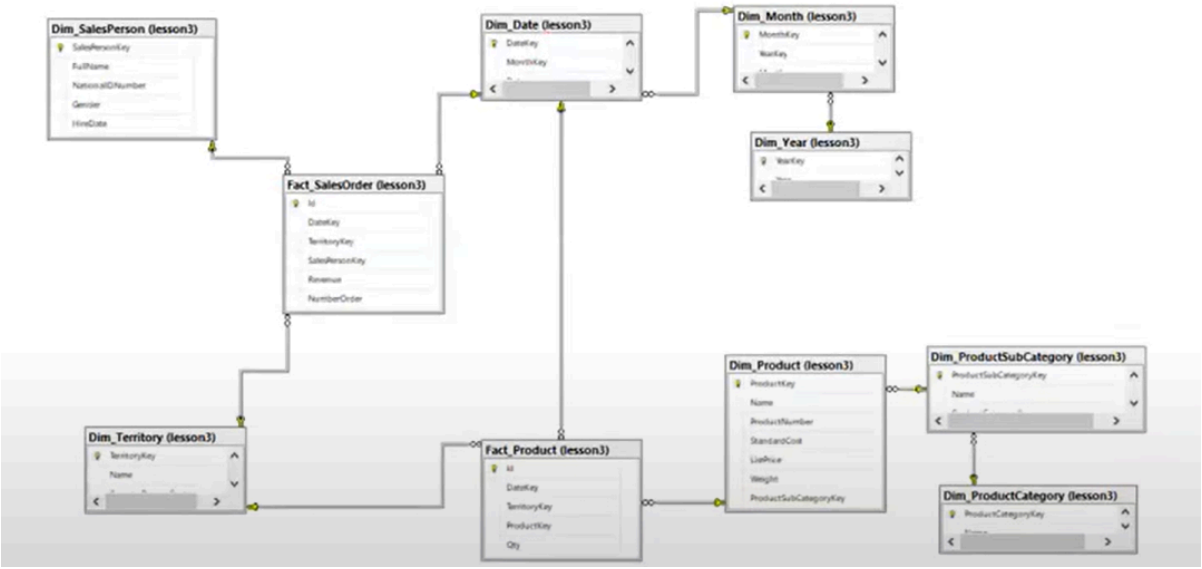
- Có những nguồn dữ liệu nào?
- Các nguồn dữ liệu ấy cung cấp loại dữ liệu nào: Database, Flatfile (CSV, excel,...) hay các file JSON trả về từ API.
- Làm tài liệu đặc tả chi tiết về dữ liệu. Ví dụ: Nếu có dữ liệu nguồn là Database, thì phải phân tích Diagram, ý nghĩa các bảng, các thuộc tính từng bảng, mối liên hệ giữa các bảng ⇒ Mục đích để hiểu dữ liệu.

Bước 1: Khảo sát nghiệp vụ các bên cần dùng Data (Data Analysis, AI Engineer,...) và xác định nhu cầu báo cáo của họ.

Bước 2: Dựa trên nhu cầu mà liệt kê ra các báo cáo cần có và phân tích các báo cáo đó theo 1 số tiêu chí:

- Các bảng Dimension và Fact mà báo cáo đó cần. Lưu ý: Các Dimention cần tạo luôn có Dimention về thời gian, cho dù các báo cáo không chỉ rõ thì vẫn cần thêm Dimention về thời gian vào danh sách.
- Xác định xem các bảng Dimension và Fact này có thể được cấu thành từ những bảng nào trong data gốc (ở bước này có thể xác định được xem, ở thời điểm hiện tại thì có đủ dữ liệu để hoàn thành báo cáo không) ⇒ Và sẽ suy ra được, những báo cáo nào có khả năng hoàn thành\không có khả năng hoàn thành\không có khả năng hoàn thành ở hiện tại nhưng có khả năng hoàn thành trong tương lai - khi dữ liệu được sinh ra trong quá trình vận hành.

- Phân tích xem các bảng Dimension có thể phân cấp không (ví dụ: bảng Dim_salary sẽ có thể là bảng con của bảng Dim_employee).
- Xác định xem các bảng Fact có thể gộp không. Nếu các bảng Fact được cấu thành bởi các bảng chung hoặc các bảng Fact có mục đích tương tự nhau (ví dụ: 2 bảng Fact_revenue_month và Fact_revenue_store có thể xem xét gộp với nhau).
- Sau khi xác định được Dimension và Fact thì lựa chọn xem chọn thiết kế Datawarehouse theo lược đồ Star, Snowflake hay Galaxy schema và xây dựng Diagram cho Datawarehouse. Ví dụ:



- Sau khi, thiết kế sơ bộ Datawarehouse hoàn thành thì đi sâu hơn vào phân tích từng bảng. Ví dụ:
  - Bảng Dim_employee này cần những thuộc tính nào? Các thuộc tính đó lấy từ những bảng nào trong data gốc?
  - Bảng Fact thì lấy dữ liệu từ những bảng Dimension nào?
  - Xác định kiểu dữ liệu cho các trường và xác địnhPrimary key \Foreign key của các bảng.
  - Xác định các trường nào có Null, 0,... để chú ý khi xử lý - vì các phép toán với Null, với 0 có thể bị lỗi (Ví dụ: phép chia cho 0 sẽ lỗi, và các phép toán với Null sẽ luôn trả về Null).