DATA WAREHOUSE

GROUP 8

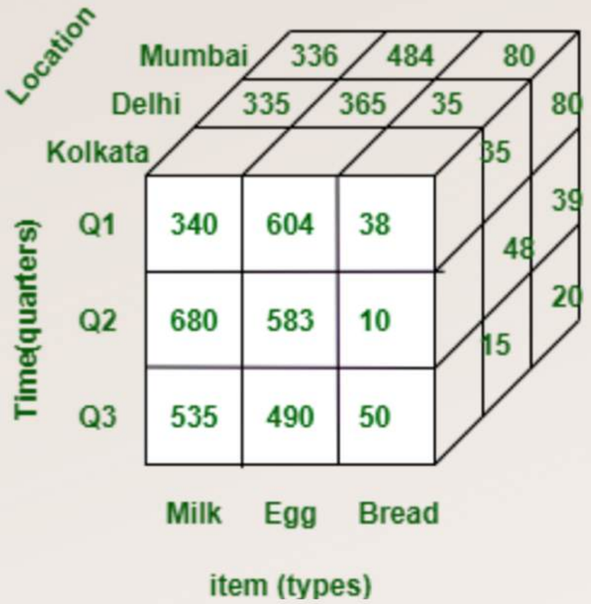
NỘI DUNG

DW & ONLINE ANALYTICAL PROCESSING (OLAP)

MULTIDIMENSIONAL MODEL

- **Mô hình đa chiều** được hiển thị dưới dạng khối dữ liệu **(Data Cube)** (chiều > 3 là **siêu khối**).
- **Dimension** (chiều): chứa thông tin, ngữ cảnh.
- **Fact** (dữ kiện): chứa số liệu, giá trị.

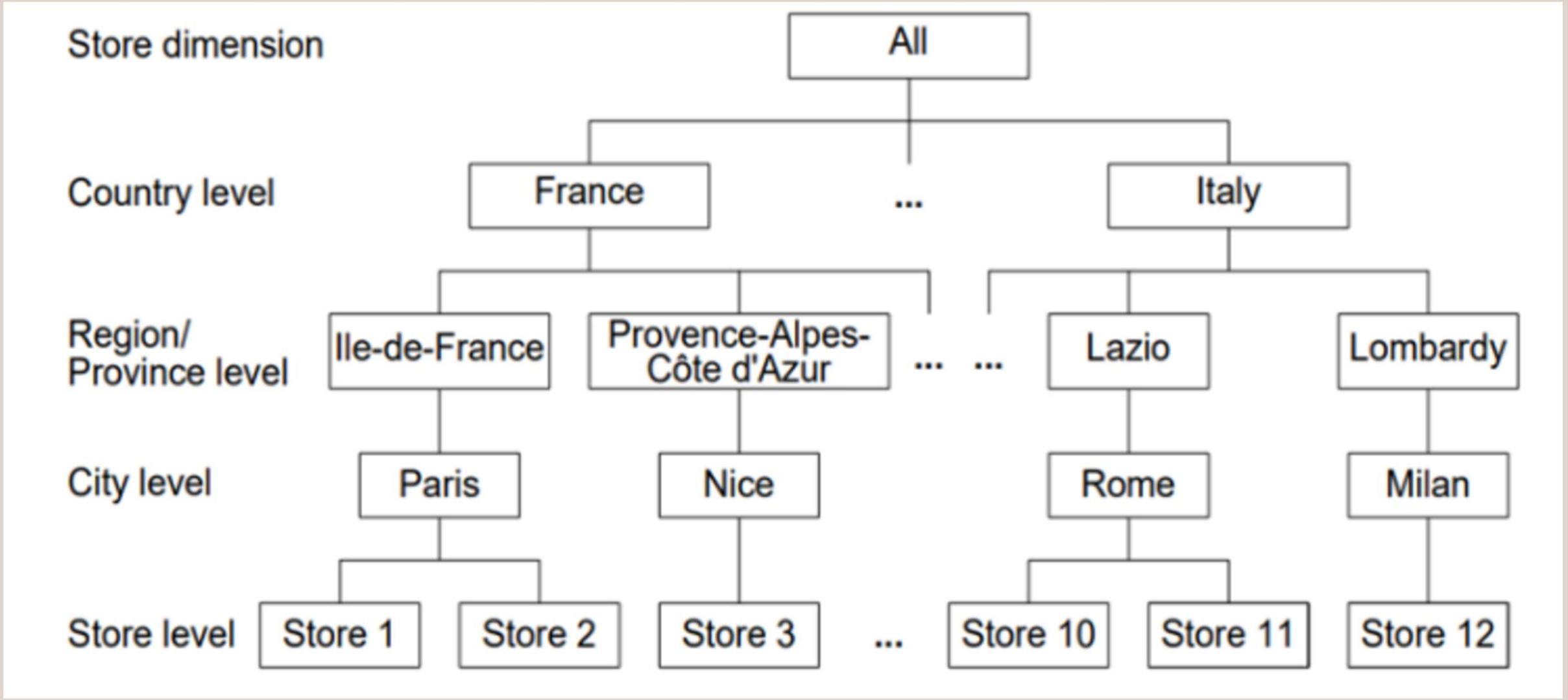
	location=*Kolkata*			location=*Delhi*			location=*Mumbai*		
	Item			Item			Item		
time	*Milk*	*Egg*	*Bread*	*Milk*	*Egg*	*Bread*	*Milk*	*Egg*	*Bread*
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	496	380	48	455	519	39
Q3	535	490	50	489	403	15	483	499	20



HIERARCHIES

- **Data granularity (Độ mịn)**: Mức độ chi tiết của các số liệu.
- Dữ liệu được phân tích ở các mức độ chi tiết khác nhau.
- Cấu trúc phân cấp liên quan đến các khái niệm cấp thấp (chi tiết) với các khái niệm cấp cao hơn (tổng quát).
- Cấp độ thấp hơn được gọi là cấp con, cấp độ cao hơn được gọi là cấp độ được gọi là cấp cha.
- Các trường hợp của các cấp độ này được gọi là thành viên.

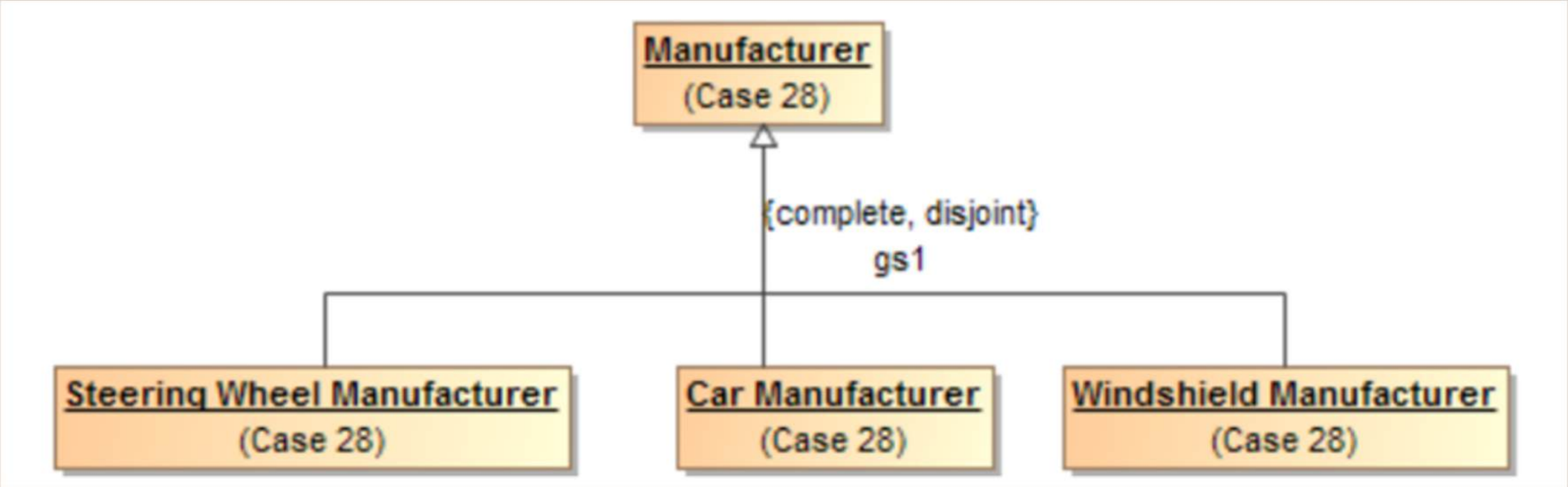
HIERARCHIES



MEASURE AGGREGATION AND SUMMARIZABILITY

Điều kiện tóm tắt để đảm bảo tính đúng đắn của việc tổng hợp

- Sự phân biệt của các trường hợp (*Disjointness of instances*)
- Sự đầy đủ (*Completeness*)
- Sử dụng đúng các hàm (*Correct use of aggregation functions*)



ELEMENTS OF DIMENSIONAL DATA MODEL

FACT

- **Fact** là kết quả của 1 quy trình kinh doanh hoặc sự kiện kinh doanh
- **Granularity/grain:** Biểu thị mức độ chi tiết của dữ liệu được lưu trữ trong hệ thống
  - Mỗi hàng trong bảng đại diện cho 1 thông tin chi tiết
  - **Atomic grain (mức chi tiết nhỏ nhất):** thường linh hoạt nhất

DIMENSION

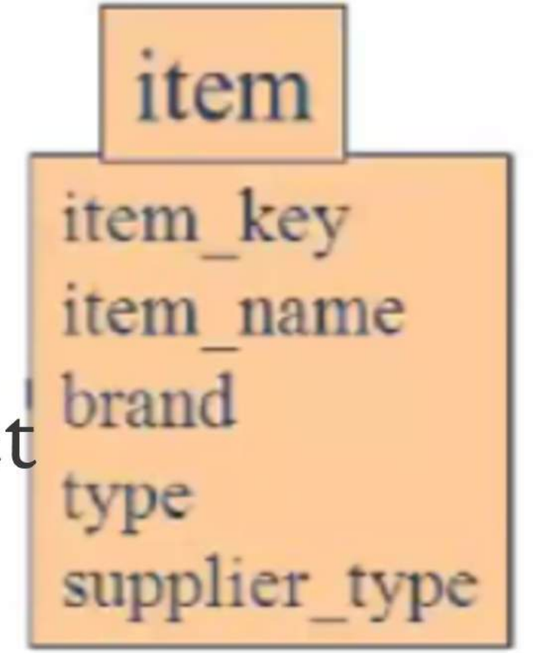
- Cung cấp bối cảnh xung quanh một sự kiện quy trình kinh doanh.
  Ex: Mô hình phân tích doanh số bán hàng có các dimension như thời gian, khách hàng, sản phẩm, khu vực địa lý
- Mỗi hàng chỉ xuất hiện 1 lần duy nhất

DIMENSION ATTRIBUTES

- Nói về khía cạnh nào đó của dimension
  Ex: dimension Sản phẩm thì gồm các thuộc tính phía dưới để mô tả chi tiết cho Sản phẩm đó như tên gì, loại gì
- Tạo nên Hierarchical relationships

DIMENSION TABLE

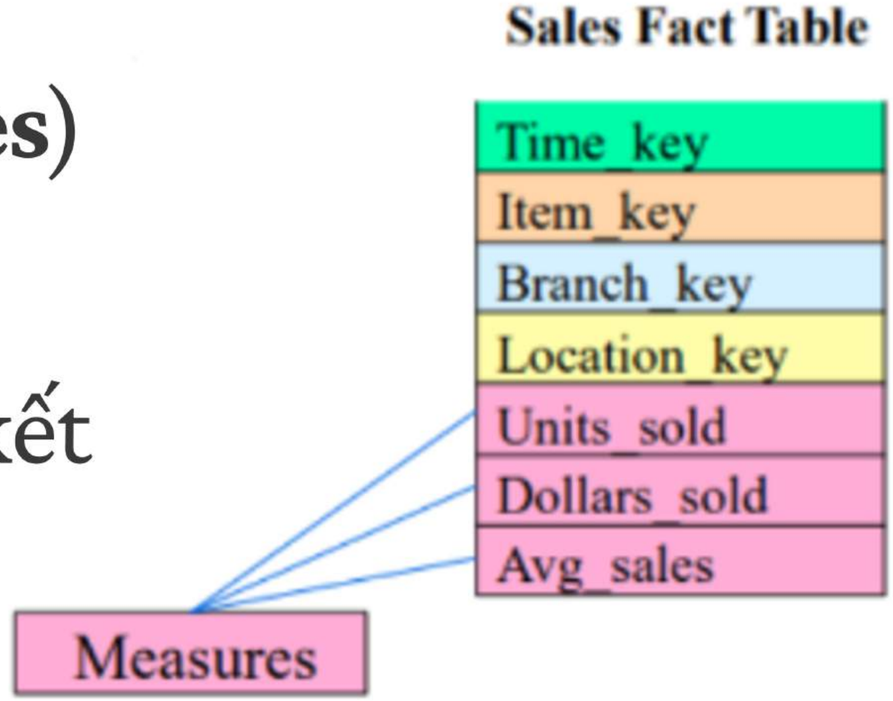
- Chứa các dimension của fact
- Nối vào fact table thông qua foreign key
- Là các bảng không chuẩn hóa
- Cung cấp các đặc điểm mô tả của các fact
- Số lượng dimension không giới hạn
- Chứa hierarchical relationships



ELEMENTS OF DIMENSIONAL DATA MODEL

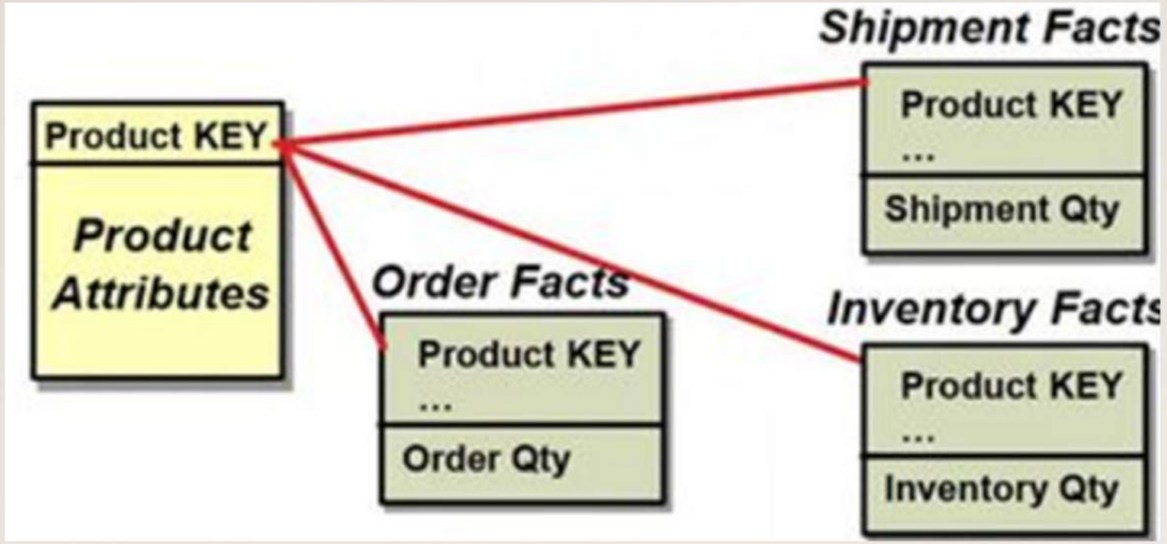
FACT TABLE

- **Bảng chính trong Dimension Modeling**
- Các số liệu (**measures**)
- Các khóa ngoại (**foreign keys**) liên kết đến các bảng **dimension tables**



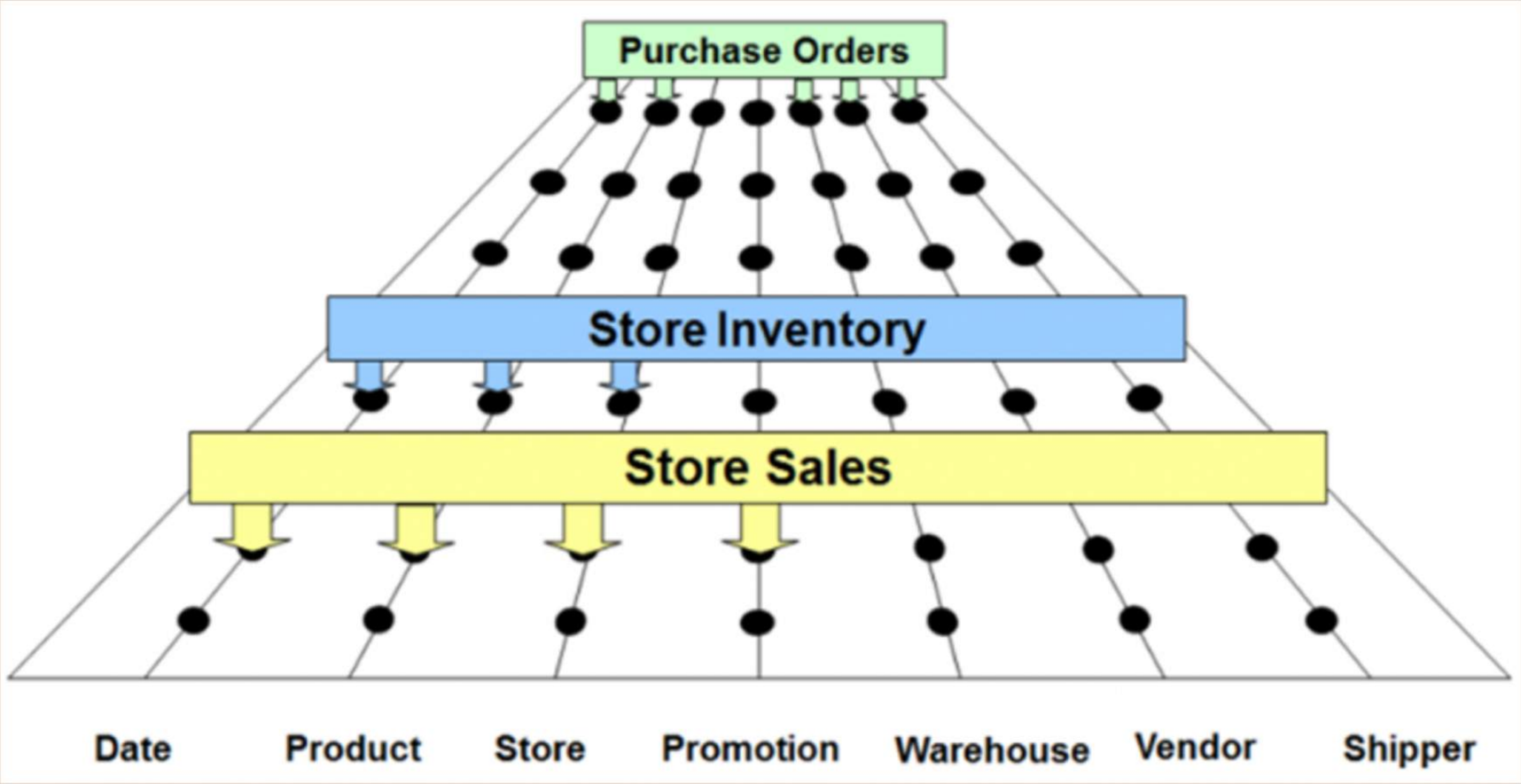
Conformed Dimensions

- Được chia sẻ xuyên suốt các quy trình kinh doanh (Fact Tables) trong kho dữ liệu (DW)
- Tất cả các fact table đều sử dụng chung các standard dimension



- Được thiết lập thông qua Bus Matrix và được đảm bảo thực thi trong quá trình ETL.

ENTERPRISE DW BUS ARCHITECTURE



Cho cái nhìn tổng quan về DW thông qua các fact table và các dimension, thấy được kiến trúc chung của DW

DATA WAREHOUSE BUS MATRIX

Dùng để xác định và quản lý các mối quan hệ giữa các bảng trong DW

	SHARED DIMENSIONS									
BUSINESS PROCESSES	Account	Customer	Data	Department	Employee	Organization	Products	Promotion	Reseller	Sales Territory
Customer Service Calls		✓	✓		✓		✓			
Customer Surveys		✓	✓				✓			
General Ledger	✓		✓	✓		✓				
Internet Sales		✓	✓				✓	✓		✓
Intventory			✓				✓			
Reseller Sales			✓		✓		✓	✓	✓	✓
Sales Plan			✓				✓			✓

MEASURE CLASSIFICATION

ADDITIVITY

01

Additive measures (flow hoặc rate measures)

Những đo lường có thể được tổng hợp một cách có ý nghĩa thông qua phép cộng theo tất cả các chiều.

02

Semi-additive measures(stock hoặc level measures)

Những đo lường có thể được tổng hợp một cách có ý nghĩa thông qua phép cộng theo một số chiều (không phải tất cả)

03

Non-additive measures(value-per-unit measures)

Là những đo lường không thể được tổng hợp một cách có ý nghĩa thông qua phép cộng theo bất kỳ chiều nào.

MEASURE CLASSIFICATION

AGGREGATION COMPLEXITY

01

Distributive(phân tán) measures

Được xác định bởi một hàm tổng hợp có thể tính toán một cách phân tán.

Ex: sum, min, max, count

02

Algebraic (đại số) measures

Được xác định bởi một hàm tổng hợp có thể được biểu diễn dưới dạng một hàm vô hướng của hàm phân tán.

Ex: average (có thể tính bằng cách chia tổng cho số lượng)

03

Holistic (tổng thể) measures

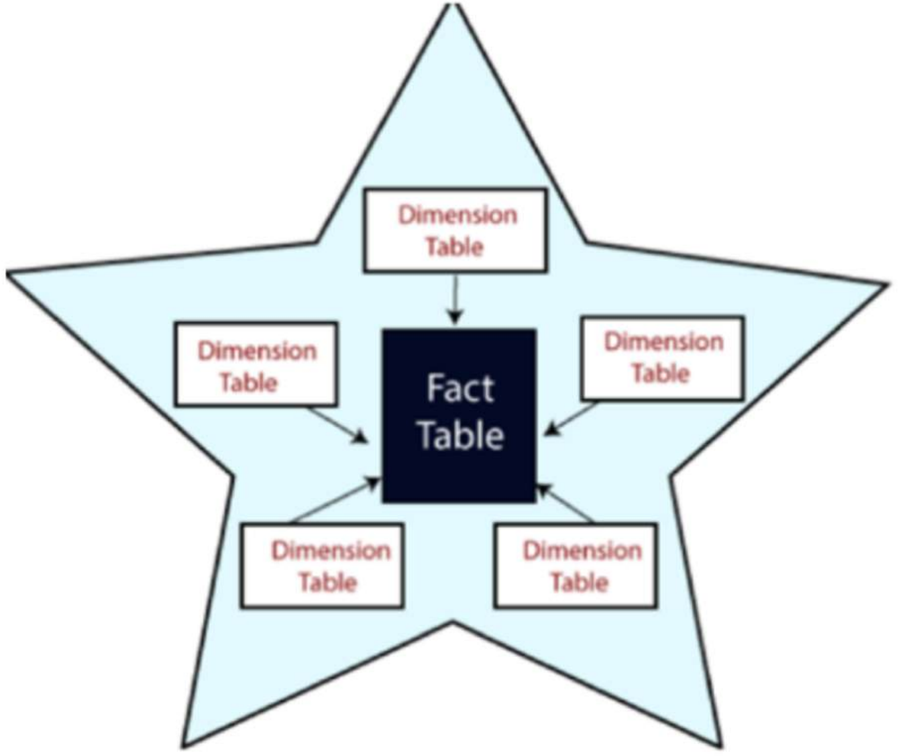
Không có hạn chế hằng số nào về kích thước lưu trữ cần thiết để mô tả một tập dữ liệu con-tổng hợp.

Ex: median, mode, rank

STAR SCHEMA

ĐẶC ĐIỂM CỦA STAR SCHEMA

- Tạo ra một cơ sở dữ liệu được chuẩn hóa nhanh chóng cung cấp các phản hồi truy vấn.
- Cung cấp một thiết kế linh hoạt có thể dễ thay đổi/thêm vào trong suốt chu kỳ phát triển và khi cơ sở dữ liệu phát triển.
- Cung cấp một thiết kế song song với các người dùng cuối thường nghĩ và sử dụng dữ liệu.
- Giảm độ phức tạp của siêu dữ liệu cho cả nhà phát triển và người dùng cuối.



STAR SCHEMA

ƯU ĐIỂM

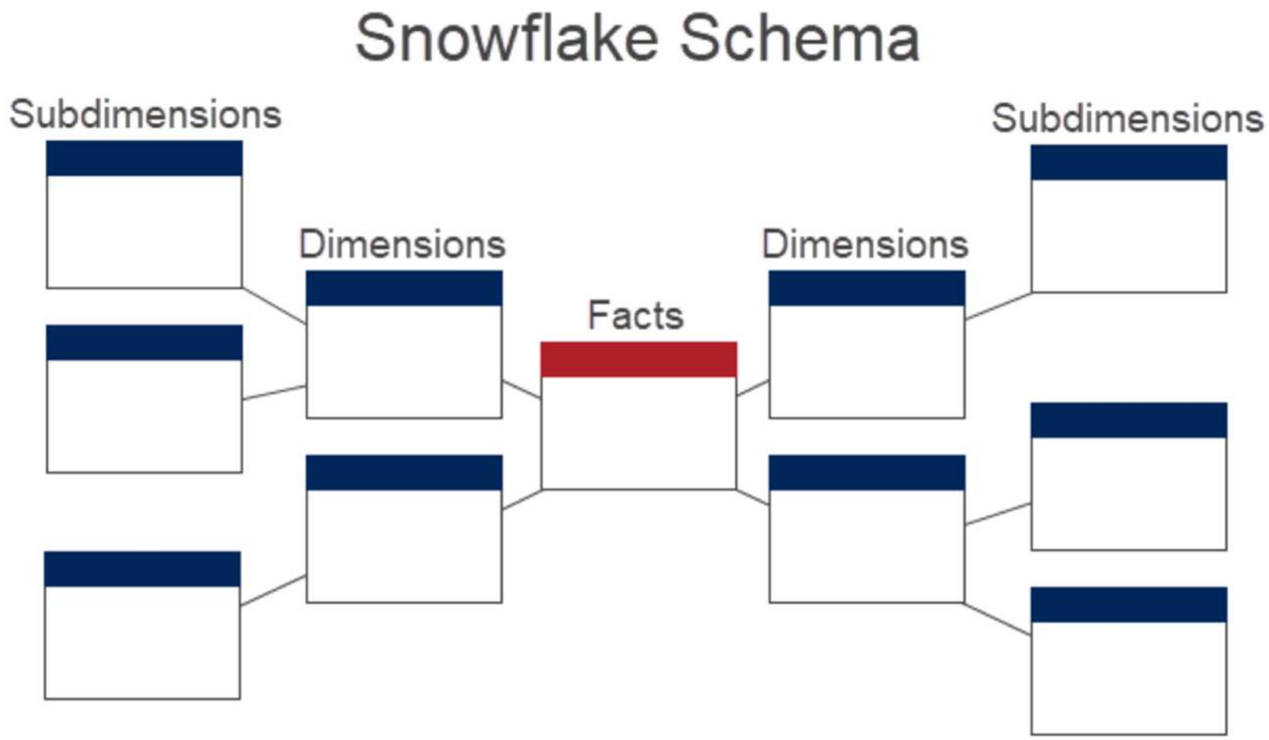
- Dễ hiểu
- Hiệu suất cao hơn
- Mở rộng để xử lý thay đổi

NHƯỢC ĐIỂM

Có một số điều kiện mà Star Schema không thể đáp ứng được.

SNOWFLAKE SCHEMA

- Mô hình dữ liệu được dùng trong DW và Data Mart
- Phần mở rộng của mô hình Star Schema.
- Gồm: **Fact table** và **Dimension Table**



SNOWFLAKE SCHEMA

ƯU ĐIỂM

- Dữ liệu được lưu trữ theo một cấu trúc nhất quán
- Giảm dữ liệu trùng lặp và cải thiện hiệu suất lưu trữ.
- Cải thiện tính bảo mật của data warehouse bằng cách phân tách dữ liệu thành các bảng riêng biệt.

NHƯỢC ĐIỂM

- Khó hiểu, khó thiết kế
- Truy xuất dữ liệu theo nhu cầu sẽ chậm hơn do phải join nhiều bảng.

DEMO

THANK YOU FOR WATCHING!!!

SNOWFLAKE SCHEMA

ƯU ĐIỂM

- Sự phát triển về hiệu suất truy vấn do các yêu cầu lưu trữ đĩa được giảm thiểu và tham gia các bảng tra cứu nhỏ hơn.
- Cung cấp khả năng mở rộng lớn hơn trong mối quan hệ qua lại giữa các mức dimension và các thành phần.
- Không dư thừa nên dễ bảo trì hơn.

NHƯỢC ĐIỂM

- Cần thêm các nỗ lực bảo trì do số lượng bảng tra cứu ngày càng tăng.
- Có nhiều truy vấn phức tạp hơn và do đó, khó hiểu.
- Nhiều bảng tham gia nên thời gian thực hiện truy vấn nhiều hơn.