**MỤC LỤC**

[I- Giới thiệu 2](#_Toc8362467)

[II- Yêu cầu nghiệp vụ 5](#_Toc8362474)

[III- Đặc tả đầu ra – đầu vào kho dữ liệu 6](#_Toc8362475)

[IV- Thiết kế kho dữ liệu 14](#_Toc8362476)

[V- Cài đặt các khối dữ liệu 15](#_Toc8362477)

# Giới thiệu

**Kho dữ liệu** (tiếng Anh: data warehouse) là kho lưu trữ dữ liệu lưu trữ bằng thiết bị điện tử của một tổ chức. Các kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và lập báo cáo.

Định nghĩa cổ điển này về kho dữ liệu tập trung vào việc lưu trữ dữ liệu. Tuy nhiên, các phương tiện cho việc lấy và phân tích, trích rút, biến đổi, nạp dữ liệu, và quản lý dữ liệu từ điển cũng được coi là các thành phần cốt yếu của một hệ thống kho dữ liệu. Nhiều người sử dụng thuật ngữ "kho dữ liệu" với ngữ cảnh rộng hơn. Một định nghĩa mở rộng cho kho dữ liệu bao gồm cả các công cụ thông minh, các công cụ để trích, biến đổi và nạp dữ liệu vào kho, và các công cụ để quản lý và lấy siêu dữ liệu (meta data).

**Giới thiệu**

Trong quá trình hoạt động kinh doanh, các dữ liệu của doanh nghiệp phát sinh ngày càng nhiều. Người ta muốn tận dụng nguồn dữ liệu này để sử dụng cho những mục đích hỗ trợ cho công việc kinh doanh ví dụ như cho mục đích thống kê hay phân tích. Quá trình tập hợp và thao tác trên các dữ liệu này có những đặc điểm sau:

* Dữ liệu tích hợp (Atomicity): Dữ liệu tập hợp từ nhiều nguồn khác nhau. Điều này sẽ dẫn đến việc quá trình tập hợp phải thực hiện việc làm sạch, xắp xếp, rút gọn dữ liệu.
* Theo chủ đề (Consistency): Không phải tất cả các dữ liệu đều được tập hợp, người ta chỉ lấy những dữ liệu có ích.
* Biến thời gian (Isolation): Các dữ liệu truy suất không bị ảnh hưởng bởi các dữ liệu khác hoặc tác động lên nhau.
* Dữ liệu cố định (Durable): Khi một Transaction hoàn chỉnh, dữ liệu không thể tạo thêm hay sửa đổi.

**Khái niệm**

Kho dữ liệu là một tập các dữ liệu có những đặc điểm sau: tập trung vào một chủ đề, tổng hợp từ nhiều nguồn dữ liệu khác nhau, từ nhiều thời gian, và không sửa đổi. Được dùng trong việc hỗ trợ ra quyết định trong công tác quản lý.

**Cấu trúc dữ liệu cho kho dữ liệu**

Vì dữ liệu trong kho dữ liệu rất lớn và không có những thao tác như sửa đổi hay tạo mới nên nó được tối ưu cho việc phân tích và báo cáo. Các thao tác với dữ liệu của kho dữ liệu dựa trên cơ sở là Mô hình dữ liệu đa chiều (multidimensional data model), được mô hình vào đối tượng gọi là data cube. Data cube là nơi trung tâm của vấn đề cần phân tích, nó bao gồm một hay nhiều tập dữ kiện (fact) và các dữ kiện được tạo ra từ nhiều chiều dữ kiện khác nhau (dimension).

Một thống kê doanh số bán hàng dựa trên ba yếu tố là: địa điểm, thời gian và chủng loại hàng. Data cube là vấn đề “Thống kê bán hàng” với ba chiều là ba yếu tố: địa điểm, thời gian và chủng loại hàng. Bảng fact là bảng tổng hợp dữ liệu của mối liên quan của doanh số với 3 yếu tố.

**Ngôn ngữ cho kho dữ liệu**

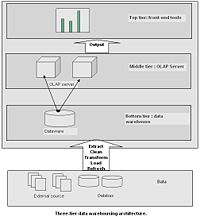
Ngôn ngữ xử lý phân tích trực tuyến (OLAP - On-Line Analytical Prosessing), rất phù hợp với kho dữ liệu, ngôn ngữ này tương tự với ngôn ngữ truy vấn SQL và tập trung vào các câu lệnh sau:

* Thu nhỏ (roll-up): ví dụ: nhóm dữ liệu theo năm thay vì theo quý.
* Mở rộng (drill-down): ví dụ: mở rộng dữ liệu, nhìn theo tháng thay vì theo quý.
* Cắt lát (slice): nhìn theo từng lớp một. Ví dụ: từ danh mục bán hàng của Q1, Q2, Q3, Q4 chỉ xem của Q1.
* Thu nhỏ (dice): bỏ bớt một phần của dữ liệu (tương ứng thêm điều kiện vào câu lệnh WHERE trong SQL).

**Cấu trúc của một hệ thống kho dữ liệu**

Bao gồm ba tầng:

* Tầng đáy: Là nơi cung cấp dịch vụ lấy dữ liệu từ nhiều nguồn khác sau đó chuẩn hóa, làm sạch và lưu trữ dữ liệu đã tập tung
* Tầng giữa: cung cấp các dịch vụ để thực hiện các thao tác với kho dữ liệu gọi là dịch vụ OLAP (OLAP server). Có thể cài đặt bằng Relational OLAP, Multidimensional OLAP hay kết hợp cả hai mô hình trên Hybrid OLAP
* Tầng trên cùng: nơi chứa các câu truy vấn, báo cáo, phân tích



Mô hình data warehouse 3 lớp

**Mục tiêu và phạm vi bài tập lớn**

Thiết kế và cài đặt một kho dữ liệu cho một hệ thống xử lý đặt hàng của khách hàng cho một công ty sử dụng MS SQL Server và Oracle. Xây dựng được CSDL tương đối hoàn chỉnh cho DWH; xây dựng các ánh xạ thu thập số liều từ hệ thống; xây dựng công cụ báo cáo, tra cứu và quản trị hệ thống; xây dựng được một số báo cáo phân tích ban đầu.

# Yêu cầu nghiệp vụ

Để đáp ứng được nhu cầu của khách hàng, hệ thống kho dữ liệu trích lọc dữ liệu từ hai cơ sở dữ liệu hiện có để cho vào kho dữ liệu và cung cấp các xử lý phân tích trực tuyến với các thao tác cuộn lên, khoan xuống, chọn và chiếu dựa trên những yêu cầu chọn lựa của khách hàng và các bảng theo chiều để đáp ứng yêu cầu của người sử dụng. Thiết lập một chiều thời gian, hệ thống sẽ sinh một báo cáo Phân tích trực tuyến cho những yêu cầu sau:

1. Tìm tất cả các cửa hàng cùng với thành phố, bang, số điện thoại, mô tả, kích cỡ, trọng lượng và đơn giá của tất cả các mặt hàng được bán ở kho đó.
2. Tìm tất cả các đơn đặt hàng với tên khách hàng và ngày đặt hàng được thực hiện bởi khách hàng đó
3. Tìm tất cả các cửa hàng cùng với tên thành phố và số điện thoại mà có bán các mặt hàng được đặt bởi một khách hàng nào đó
4. Tìm địa chỉ văn phòng đại diện với tên thành phố, bang của tất cả các cửa hàng lưu kho một mặt hàng nào đó với số lượng trên mức cụ thể.
5. Với mỗi một đơn đặt hàng của khách, liệt kê các mặt hàng được đặt cùng với mô tả, mã cửa hàng, tên thành phố và các cửa hàng có bán mặt hàng đó.
6. Tìm thành phố và bang mà một khách hàng nào đó sinh sống
7. Tìm mức độ tồn kho của một mặt hàng cụ thể tại tất cả các cửa hàng ở một thành phố cụ thể nào đó
8. Tìm các mặt hàng, số lượng đặt, khách hàng, cửa hàng và thành phố của một đơn đặt hàng.
9. Tìm các khách hàng du lịch, khách hàng đặt theo đường bưu điện và khách hàng thuộc cả hai loại.

# Đặc tả đầu ra – đầu vào kho dữ liệu

1. Đặc tả đầu vào

* Cơ sở dữ liệu Văn phòng đại diện bao gồm những quan hệ với các lược đồ như sau:

Khách hàng (Mã KH, Tên KH, Mã Thành phố, Ngày đặt hàng đầu tiên)

Khách hàng du lịch (\*Mã KH, Hướng dẫn viên du lịch, Thời gian)

Khách hàng bưu điện (\* Mã KH, Địa chỉ bưu điện, Thời gian)

* Cơ sở dữ liệu Bán hàng gồm những quan hệ với lược đồ như sau:

Văn phòng đại diện (Mã Thành phố, Tên Thành phố, Địa chỉ VP, Bang, Thời gian)

Cửa hàng (Mã cửa hàng, \* Mã Thành phố, Số điện thoại, Thời gian)

Mặt hàng (Mã MH, Mô tả, Kích cỡ, Trọng lượng, Giá, Thời gian)

Mặt hàng\_được lưu trữ (\* Mã cửa hàng, \* Mã mặt hàng, Số lượng trong kho, Thời gian)

Đơn đặt hàng (Mã đơn, Ngày đặt hàng, Mã Khách hàng)

Mặt hàng được đặt (\* Mã đơn, \* Mã mặt hàng, Số lượng đặt, Giá đặt, Thời gian)

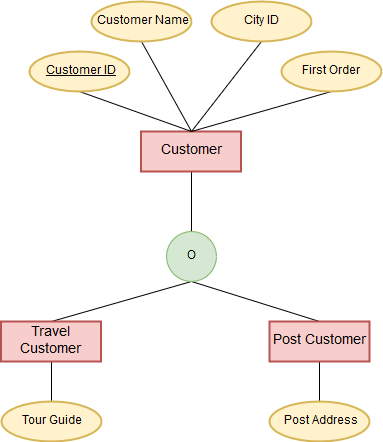
Từ cơ sở dữ liệu Văn phòng đại diện ta thấy:

Primary Relation 1 (PR1): Khách hàng (Customer)

Primary Relation 2 (PR2): Khách hàng du lịch, Khách hàng bưu điện (Tourism Customer, PostCustomer)

Foreign Key Attribute (FKA): Mã KH trong Khách hàng du lịch và Khách hàng bưu điện

Sơ đồ thực thể quan hệ EER:



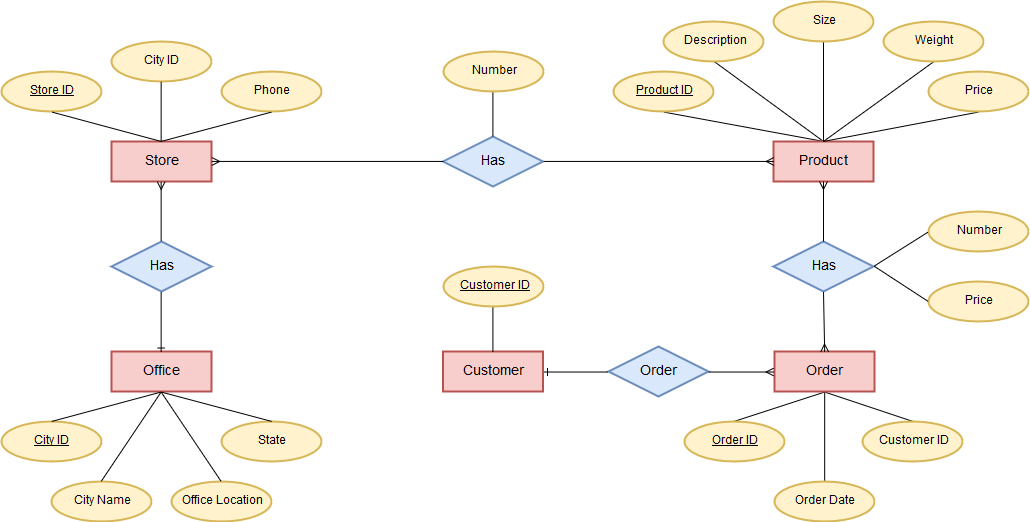
Từ cơ sở dữ liệu Bán hàng ta thấy:

Primary Relation 1 (PR1): Văn phòng đại diện, Cửa hàng, Mặt hàng, Đơn đặt hàng (Office, Store, Product, Order)

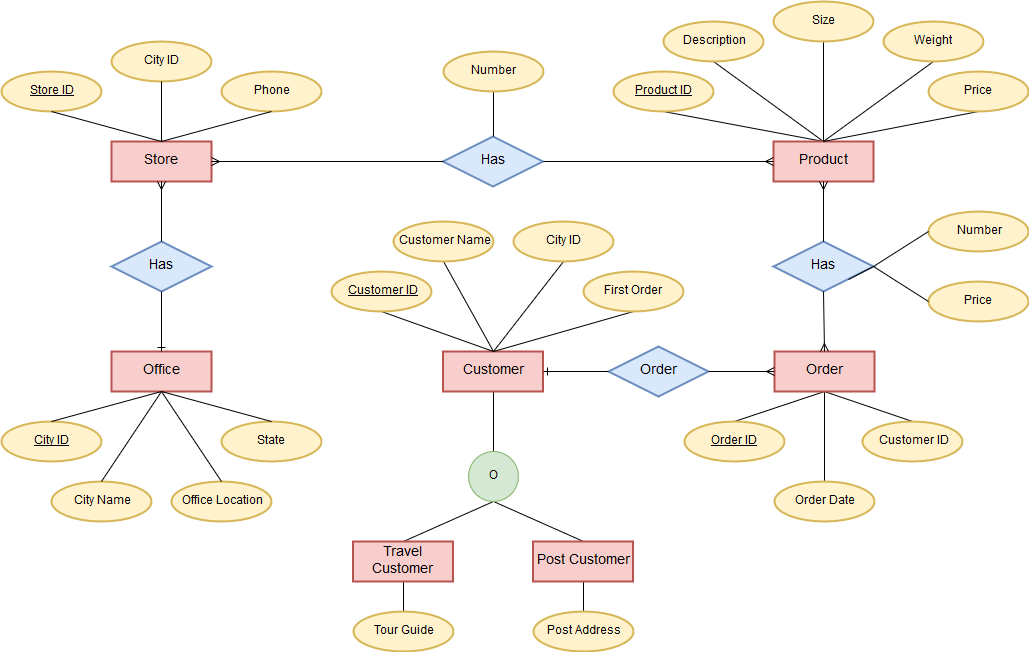
Secondary Relation 1 (SR2): Mặt hàng được lưu trữ, Mặt hàng được đặt

Foreign Key Attribute (FKA): Mã thành phố (CityID trong Store), Mã cửa hàng (StoreID trong Mặt hàng được lưu trữ), Mã mặt hàng (ProductID trong Mặt hàng được lưu trữ và Mặt hàng được đặt), Mã đơn (OrderID trong Mặt hàng được đặt)

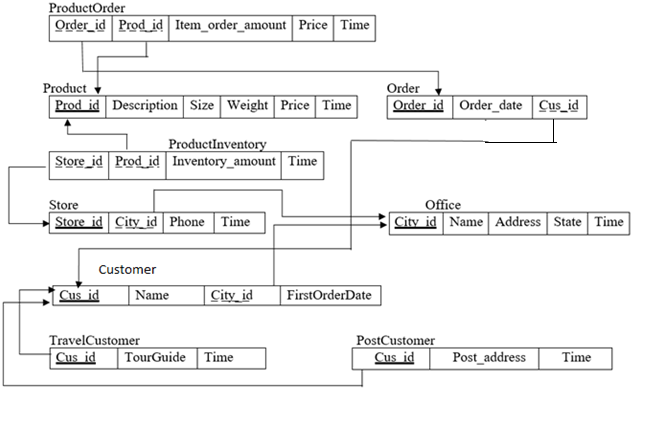
Sơ đồ thực thể quan hệ EER



Tiến hành hợp nhất 2 cơ sở dữ liệu trên ta có IEER:



Sơ đồ quan hệ của cơ sở dữ liệu sau khi hợp nhất:



Các bảng trong database:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên bảng** | **Mô tả** |
| 1 | Customer | Chứa các thông tin về các khách hàng |
| 2 | Office | Chứa các thông tin về các văn phòng đại diện |
| 3 | Order | Chứa các thông tin về các đơn hàng |
| 4 | Store | Chứa các thông tin của cửa hàng |
| 5 | Product | Chứa các thông tin về các mặt hàng |
| 6 | ProductOrder | Chứa các thông tin về các mặt hàng nằm trong các đơn hàng |
| 7 | ProductInventory | Chứa các thông tin về các mặt hàng được lưu trữ trong các cửa hàng |
| 8 | PostCustomer | Chứa mã khách hàng và địa chỉ bưu điện |
| 9 | TravelCustomer | Chứa mã khách hàng và người giới thiệu (hướng dẫn viên du lịch) |

Chi tiết các bảng:  
 - Bảng Customer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Cus\_id** | INT |  | Mã khách hàng (PK) |
| 2 | Name | NVARCHAR(50) | N | Tên của khách hàng |
| 3 | City\_id | INT | N | Mã thành phố (FK) |
| 4 | FirstOrderDate | DATE | N | Ngày đầu tiên đặt hàng |

* Bảng Office:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **City\_id** | INT |  | Mã thành phố được sử dụng làm mã cho mỗi văn phòng đại diện (PK) (do mỗi thành phố có một văn phòng) |
| 2 | name | NVARCHAR(50) | N | Tên thành phố mà văn phòng đại diện nằm tại đó |
| 3 | address | NVARCHAR(50) | N | Địa chỉ của văn phòng đại diện trong thành phố |
| 4 | state | NVARCHAR(50) | N | Bang mà văn phòng đại diện nằm tại đó |
| 5 | time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng Order:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Order\_id** | INT |  | Mã đơn đặt hàng (PK) |
| 2 | Order\_date | DATE | N | Thời gian đặt hàng |
| 3 | Cus\_id | INT | N | Mã khách hàng đặt hàng |

* Bảng Store:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Store\_id** | INT |  | Mã cửa hàng (PK) |
| 2 | City\_id | NVARCHAR(50) | N | Mã thành phố (FK) |
| 3 | Phone | VARCHAR(50) | N | Số điện thoại cửa hàng |
| 4 | Time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng Product:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Prod\_id** | INT |  | Mã mặt hàng (PK) |
| 2 | Description | NVARCHAR(50) | N | Mô tả về mặt hàng |
| 3 | Size | INT | N | Cỡ của mặt hàng |
| 4 | Weight | INT | N | Cân nặng của mặt hàng |
| 5 | Price | FLOAT | N | Giá của mặt hàng |
| 6 | Time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng ProductOrder:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Order\_id** | INT | N | Mã đơn hàng (PK, FK) |
| 2 | **Prod\_id** | INT | N | Mã mặt hàng (PK, FK) |
| 3 | Item\_order\_amount | INT | N | Số lượng hàng đặt |
| 4 | Price | FLOAT | N | Tổng giá của đơn hàng |
| 5 | Time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng ProductInventory:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Store\_id** | INT | N | Mã cửa hàng (PK, FK) |
| 2 | **Prod\_id** | INT | N | Mã mặt hàng (PK, FK) |
| 3 | Inventory\_amount | INT | N | Số lượng lưu trong kho |
| 4 | Time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng PostCustomer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Cus\_id** | INT | N | Mã khách hàng (PK) |
| 2 | Post\_address | VARCHAR(50) | N | Địa chỉ bưu điện |
| 3 | Time | TIME | N | Thời điểm nhập dữ liệu |

* Bảng TravelCustomer:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên trường** | **Kiểu dữ liệu** | **Nullable** | **Mô tả** |
| 1 | **Cus\_id** | INT | N | Mã khách hàng (PK) |
| 2 | TourGuide | NCHAR(10) | N | Tên người dẫn đến |
| 3 | Time | TIME | N | Thời điểm nhập dữ liệu |

1. Đặc tả dữ liệu đầu ra:

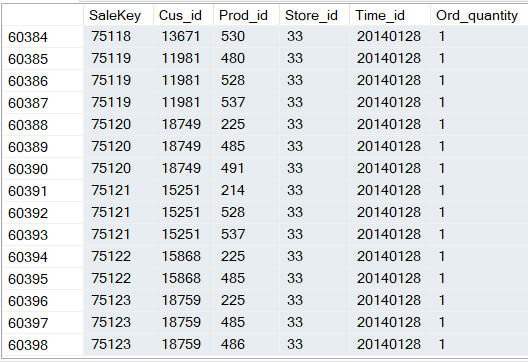
Dữ liệu đầu ra sau khi xây dựng kho dữ liệu cho đề bài trên để đáp ứng các nhu cầu của khách hàng ta có Quanity ( số lượng mặt hàng còn lại trong kho)

# Thiết kế kho dữ liệu

Thiết lập bảng Fact, bảng Dimension và lược đồ dải ngân hà:

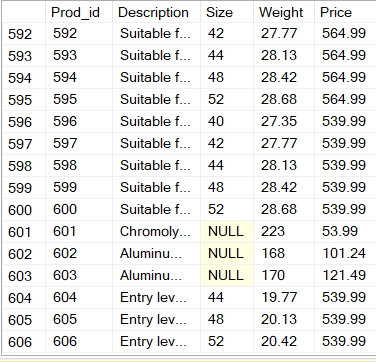
* Bảng DimCustomer

|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| Cus\_id | Mã khách hàng |
| City\_id | Mã thành phố |
| name | Tên khách hàng |
| type | Loại khách hàng. Có 2 loại: Travel và Post |



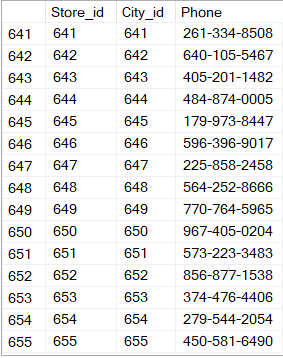
* Bảng DimProduct

|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| Prod\_id | Mã mặt hàng |
| Description | Mô tả mặt hàng |
| Size | Kích cỡ của mặt hàng |
| Weight | Khối lượng của mặt hàng |
| Price | Giá bán của mặt hàng |



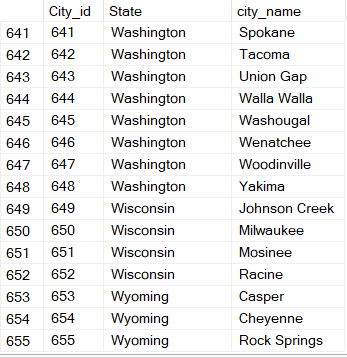
* Bảng DimStore

|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| Store\_id | Mã cửa hàng |
| City\_id | Mã thành phố |
| Phone | Số điện thoại cửa hàng |



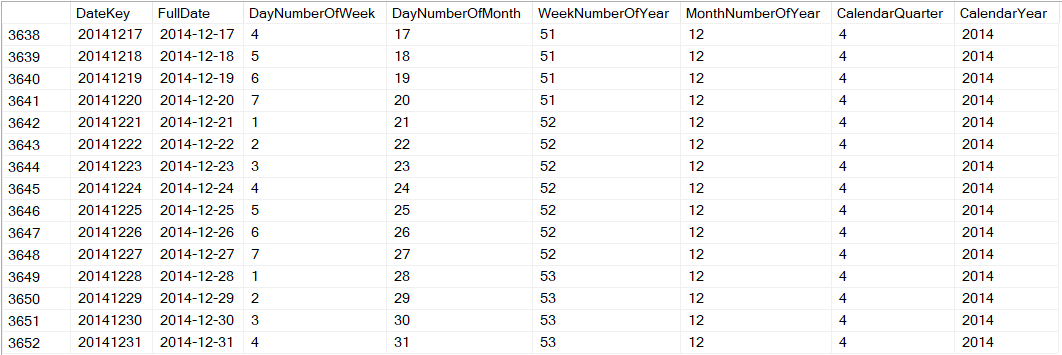
* Bảng DimOffice

|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| City\_id | Mã thành phố |
| City\_name | Tên thành phố |
| State | Tên bang |

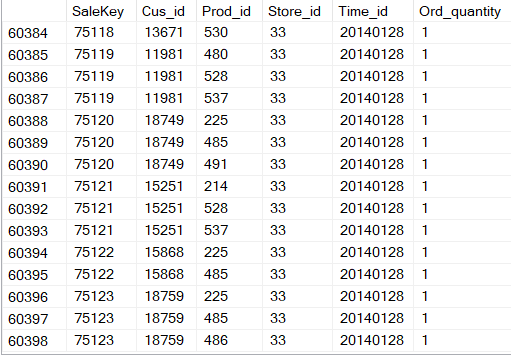


* Bảng DimTime

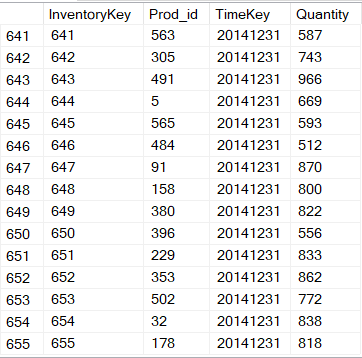
|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| DateKey | Mã ngày |
| FullDate | Ngày đầy đủ |
| DayNumberOfWeek | Ngày trong tuần |
| DayNumberOfMonth | Ngày trong tháng |
| WeekNumberOfYear | Tuần trong năm |
| MonthNumberOfYear | Tháng trong năm |
| CalenderQuarter | Quý |
| CalenderYear | Năm |

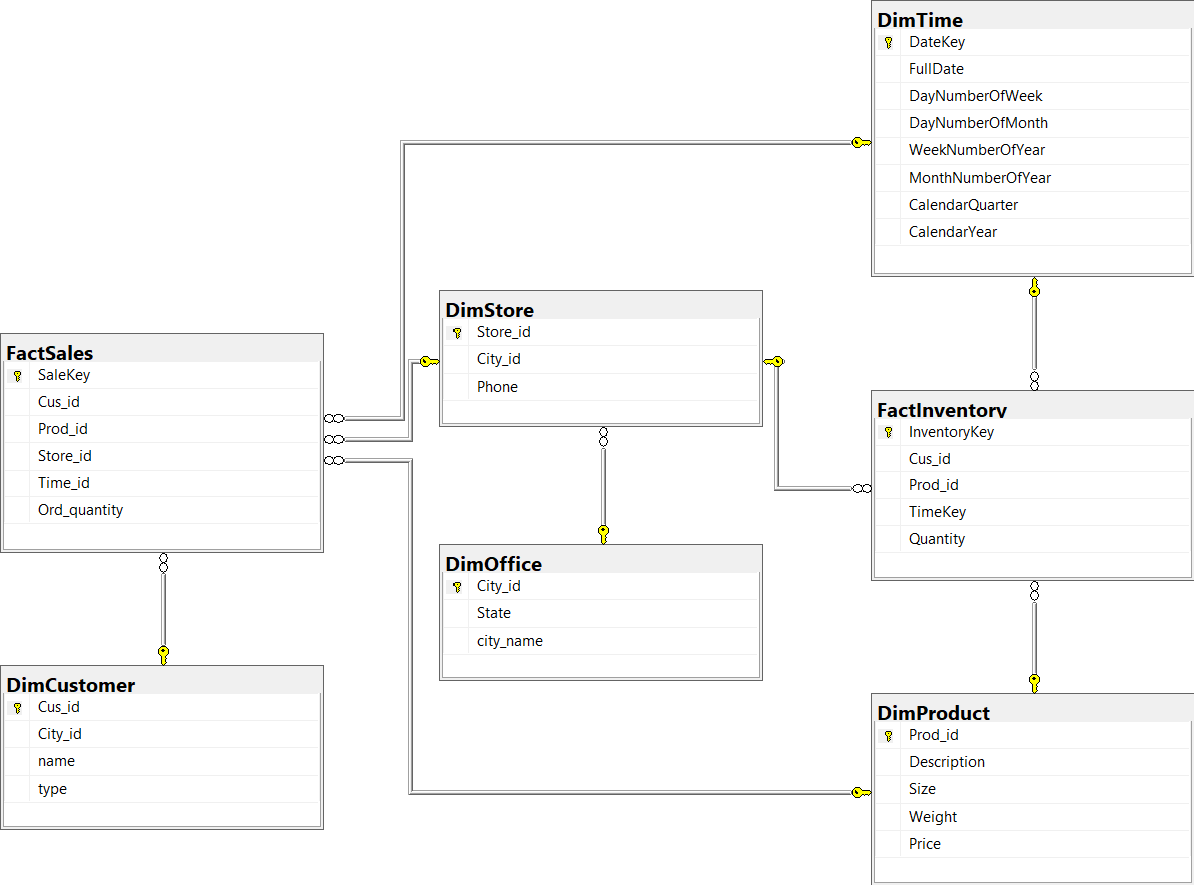


* Bảng FactSales
* Primary Key: SaleKey
* Foreign Key: Cus\_id, Prod\_id, Store\_id, Time\_id
* Độ đo: Ord\_quantity (số lượng đơn hàng)
* Mục đích: Chứa các thông tin và đơn vị đo liên quan đến việc bán mặt hàng.



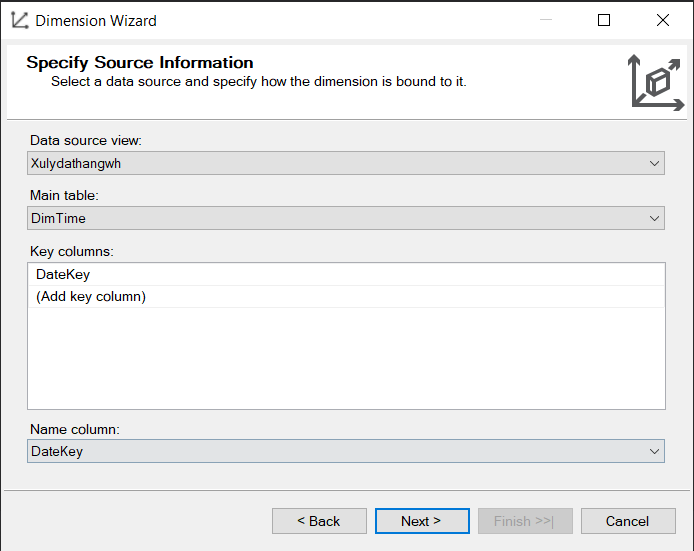
* Bảng FactInventory
* Primary Key: InventoryKey
* Foreign Key: Cus\_id, Prod\_id, TimeKey
* Độ đo: Quantity (số lượng mặt hàng)
* Mục đích: Chứa các thông tin và đơn vị đo liên quan đến việc lưu trữ mặt hàng.

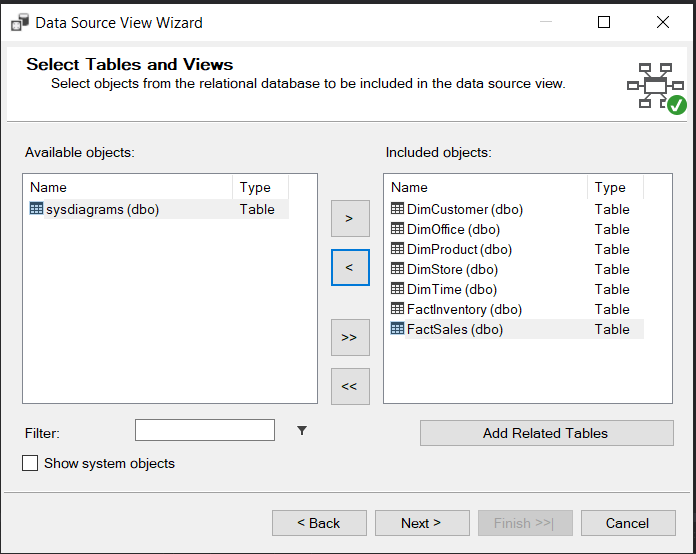


Lược đồ dải ngân hà

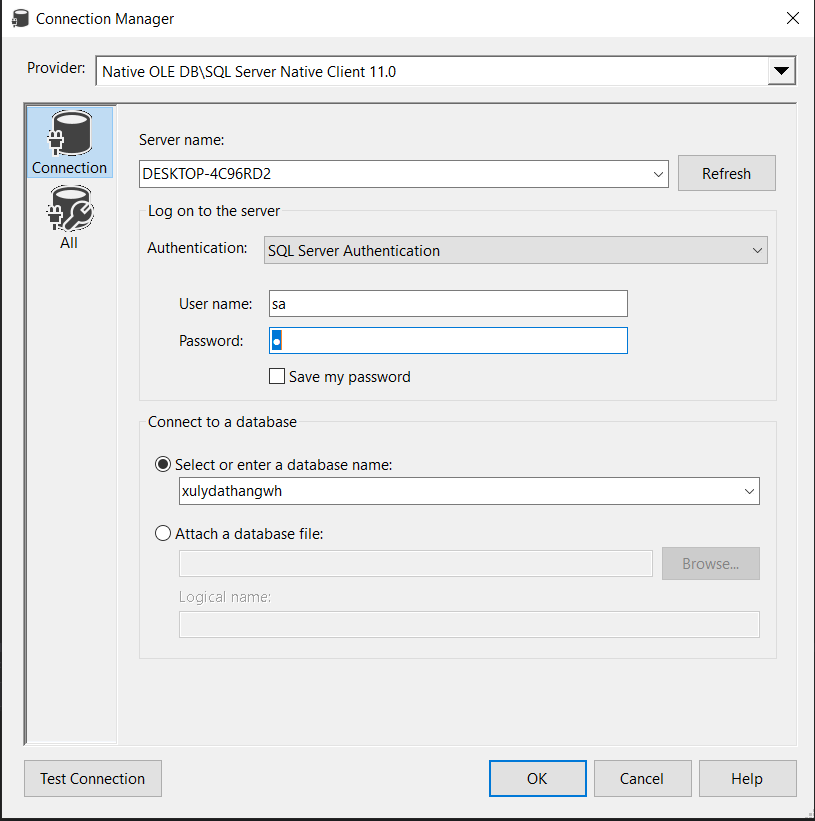
# Cài đặt các khối dữ liệu

Bước 1: Tạo chiều



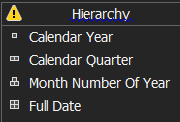


Bước 2: Tạo kết nối đến SQL Server để lấy dữ liệu

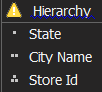


Bước 3: Cài đặt phân cấp cho các chiều có thể phân cấp:

* Chiều Time: Date -> Month -> Quarter -> Year

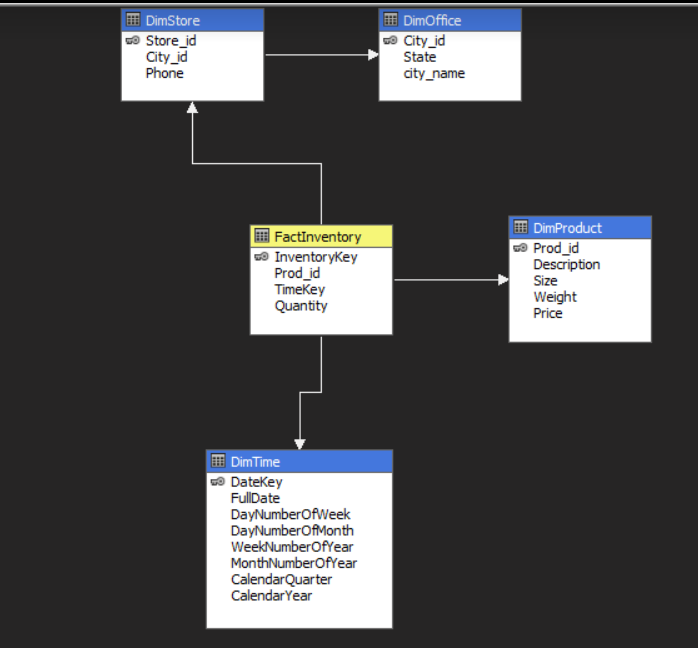


* Chiều Location: Store -> City -> State



Bước 4: Từ yêu cầu khách hàng ta có 2 khối Cube:

* Khối hỗ trợ xử lý kho hàng



* Khối hỗ trợ xử lý giao dịch:

