Dimentional Modeling Lab

I. Giới thiệu

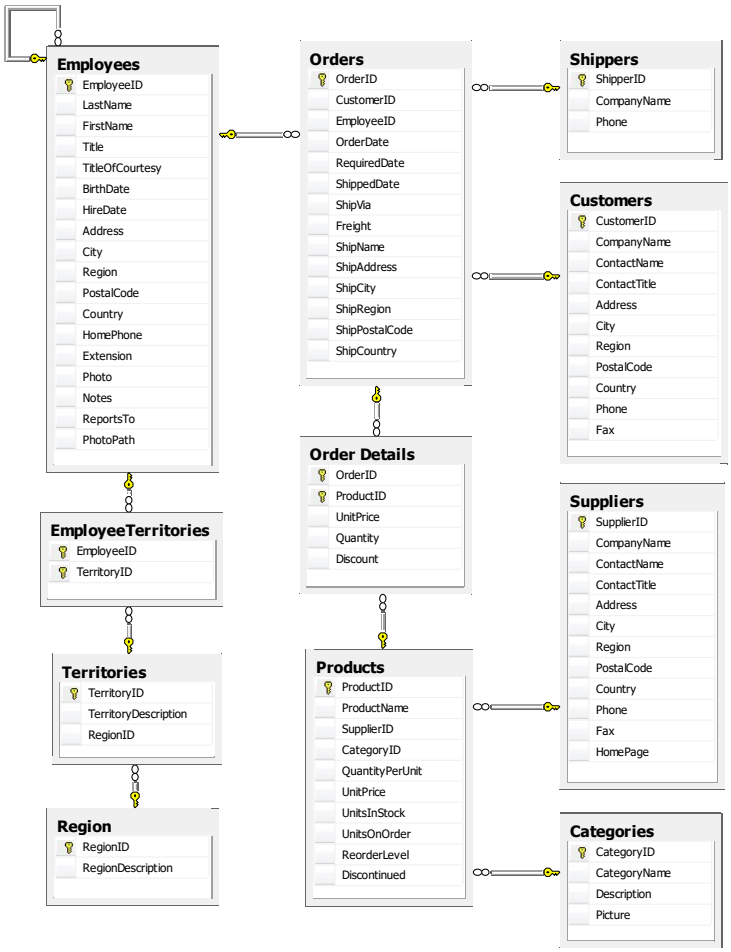
Bài tập này giới thiệu quá trình thiết kế kho dữ liệu dùng kỹ thuật mô hình hóa dữ liệu theo chiều (dimensional modeling).

Chuẩn bị

- Microsoft SQL Server phiên bản 2012 trở đi.
- **Northwind** database
- **High-Level-Dimensional-Modeling-Workbook.xlsx** và **Detailed-Dimensional-Modeling-Workbook-KimballU.xlsm**.
- **Microsoft Excel** phiên bản 2007 trở đi.

Northwind database schema

Bên dưới là Northwind database schema sau khi bỏ đi hai bảng CustomerCustomerDemo và CustomerDemographics vì chúng không chứa dữ liệu.



Bài tập này sẽ sử dụng CSDL Northwind để thực hành quá trình thiết kế mô hình dữ liệu theo chiều (dimensional model design).

Business requirement

Công ty **Northwind** muốn tạo các data mart từ CSDL **Northwind** với các mục đích sau:

Báo cáo doanh số (sales reporting): người quản lý muốn theo dõi doanh số bán hàng theo khách hàng (customer), nhân viên (employee), sản phẩm (product) và nhà cung cấp (supplier) để có thể biết được sản phẩm nào bán chạy nhất, nhân viên nào đặt nhiều hóa (order) đơn nhất, nhà cung cấp nào tốt nhất.

Việc đặt và giao hàng (Order Fulfillment and Delivery): người quản lý muốn phân tích quá trình đặt hàng để biết được khi nào hóa đơn được đặt và giao để cải thiện nó.

Phân tích hàng tồn kho (**Product Inventory Analysis**): người quản lý muốn theo dõi mức độ hàng tồn kho, hàng được đặt theo từng sản phẩm, theo nhà cung cấp và theo loại sản phẩm (category). Mức độ tồn kho nên được lưu lại hằng ngày để phân tích.

Business requirement có thể được tổng hợp bởi **Enterprise Bus Matrix** sau:

Dimension → Bus. Process ↓	**Order Date**	**Shipped Date**	**Customers**	**Employees**	**Shippers**	**Products**	**Suppliers**
Sales Reporting	X	X	X	X		X	X
Order Fulfillment	X	X	X	X	X		
Inventory Analysis	X					X	X

II. Phần hướng dẫn mẫu

1. Thiết kế mức cao (high level design)

Phần này ta sẽ sử dụng **Northwind database** và **High-Level-Dimensional-Modeling Excel Workbook**. Đầu tiên, ta sẽ hoàn thành **Detailed Bus Matrix** worksheet, sau đó sẽ hoàn thành **Attributes and Metrics** worksheet.

Ta sẽ thực hiện các bước mô hình hóa theo phương pháp của Kimball (**Kimball modeling process**) sử dụng **Detailed Bus Matrix** worksheet. Chi tiết về các khái niệm liên quan đến dimensional modeling sinh viên có thể tham khảo file "**Dimentional Modeling - The Data Warehouse Lifecycle Toolkit - Chapter 6, 7.pdf**".

Bước 1: Business Process

Phần hướng dẫn này sẽ minh họa cho **sales reporting**.

Bước 2: Xác định mức độ chi tiết (declare the grain)

Bước này xác định mức độ chi tiết (level of grain) cho fact table, tức xác định một dòng trong **sale reporting fact table** sẽ là gì? Ở đây, mỗi dòng thể hiện cho việc bán một sản phẩm. Thông tin này có trong bảng [**Order Details**]. Đây là một fact table loại **transaction**. Do vậy, ta sẽ điền vào BusMatrix worksheet như hình sau:

	A	B	C	D
1	***Instructions!***			
2	**Business Process Name**	**Fact Table**	**Fact Grain Type**	**Granulairty**
3	sales reporting	sales_fact	Transaction	one row per order detail
4				
5				

Bước 3: Xác định các dimension (identify the dimensions)

Để làm bước này, ta cần xác định các bảng liên quan đến fact table [**Order Details**]. Nhìn vào lược đồ CSDL **Northwind** ta sẽ thấy nó liên quan trực tiếp đến bảng **Products**. Bảng Products lại liên quan đến bảng **Categories** và **Suppliers**. Có nhiều product trong một category và nhiều product ứng với một supplier. Thông tin này có thể giúp ta phát hiện ra các **phân cấp** (**hierarchy**) **theo Product dimension**.

Lặp lại quá trình này để phát hiện ra các dimension khác là **Customer** và **Employee**.

Time dimension là một dimension quan trọng trong data warehouse. Nhìn vào bảng Orders, ta thấy [**Order Date**] và [**Shipped Date**], ta sẽ chọn cả hai vào trong mô hình.

Bước 4: Xác định các fact (identify the facts)

Ở bước này ta xác định các fact, một số fact cần được suy diễn (derived) dựa trên dữ liệu nguồn có sẵn. Một số fact cho sale reporting data mart có thể liệt kê như sau:

- Quantity – (số lượng sản phẩm bán được)
- Unit Price (đơn giá)
- Discount amount (unit price * discount) - chiết khấu
- Sold Amount (Quanity * (Unit Price - discount Amount))
- Freight Amount (tiền cước)

Sau khi hoàn thành xong 4 bước trên, ta sẽ điền vào **Detailed Bus Matrix** worksheet như hình bên dưới.

	A	B	C	D	E	F	G	H	I	J
1	***Instructions!***									
2	**Business Process Name**	**Fact Table**	**Fact Grain Type**	**Granulairty**	**Facts**	**Product**	**Customer**	**Employee**	**Order Date**	**Shipped Date**
3	sales reporting	sales_fact	Transaction	one row per order detail	Quantity, Unit Price , Discount Amount, Sold Amount, Freight Amount	x	x	x	x	x

Sau khi hoàn thành **Detailed Bus Matrix** worksheet, việc hoàn thành **Attributes and Metrics** worksheet là khá dễ. Một số vấn đề cần lưu ý để hoàn thành việc này:

- Bắt đầu với các dimension ta đã xác định trong **Detailed Bus Matrix** worksheet.
- Xác định các **phân cấp** (**hierarchies**) ứng với các dimension (nếu có).
- **Time dimension** thường khá chuẩn, ta chỉ cần xác định chi tiết phân cấp ta cần.
- Ghi chú các loại **fact** (**addictive, semi-additive, non- additive**)
- Nếu fact là **derived**, giải thích cách tính ra nó

2. Thiết kế chi tiết (detailed design)

Phần này ta sẽ sử dụng **Northwind** database và **Detail Dimensional Modeling Workbook**.

Chọn **Home** worksheet và điền vào các trường như sau:

- Database: **NorthwindDW**
- Description: **The Northwind Traders Data Warehouse**
- Gen FK's?: **Y**
- Schema For Views: **(để trống)**

Hoàn thành thiết kế chi tiết cho Customer dimension của Sales Reporting data mart

Để hoàn thành bước này, ta sẽ tham chiếu đến **Attributes & Measures worksheet của High-Level-Dimensional-Modeling Excel Workbook**.

Hình sau mô tả một phần của thiết kế chi tiết cho **Customer** dimension

	A	B	C	D	E
1	***Instructions!***				
2	Dimension / Fact Table	Attribute / Fact Name	Description	Alternate Names	Sample Values
3	Customer	Company Name	The name of the customer's company		Bon App'
4	Customer	Contact Name	The name of the contact at the company		Thomas Hardy
5	Customer	Contact Title	The contact's title at the company		Owner
6	Customer	Customer Country	Country of origin for the customer		France
7	Customer	Customer Region	State or province for the customer (not aval sometimes)		WA
8	Customer	Customer City	Customer's city		London
9	Customer	Customer Postal Code	Customer's postal code		13244

Ở phần này, ta sẽ thực hiện theo các bước sau:

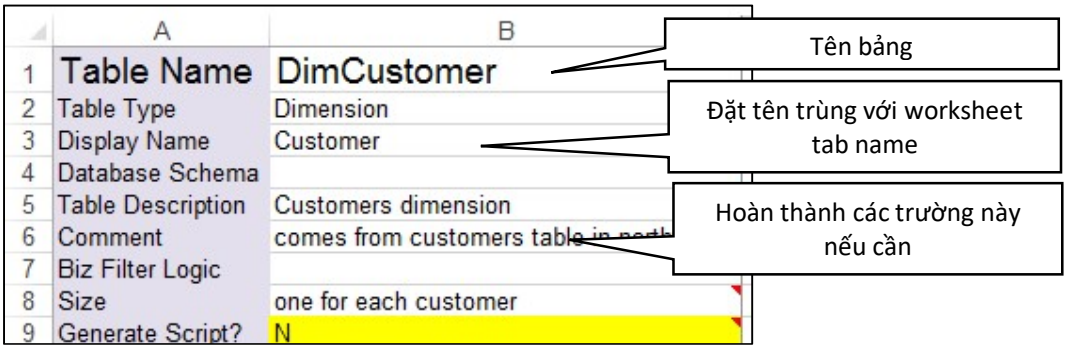
a. Tạo một dimension (hoặc fact) worksheet mới trong workbook.
b. Hoàn thành table definition của worksheet.
c. Hoàn thành basic column information
d. Hoàn thành target table information.
e. Hoàn thành source data information.

a. Tạo một dimension (hoặc fact) worksheet mới trong workbook

Copy **Blank Dimension** worksheet (**Right-click** trên nó, chọn **Move or Copy,** chọn **(move to end),** chọn **Create a copy** checkbox và chọn **OK**). Sau đó, ta đổi tên worksheet bằng cách **Right-click** trên nó, chọn **rename,** gõ tên mới **(Customer)).**

b. Hoàn thành table definition của worksheet

Hoàn thành table definition như sau:



c. Hoàn thành basic column information

Dùng dữ liệu từ **Attributes & Measures** worksheet của **High-Level-Dimensional-Modeling Excel Workbook** để thực hiện bước này. Hoàn thành 9 cột đầu

- **Column Name** → Tên cột (physical) của bảng
- **Display Name** → Tên cột (logical) của bảng (nên giống the physical name)
- **Description** → Giải thích, mô tả cột cho mục đích ghi chú
- **Unknown Member** → Giá trị mặc định cho unknown value (thay vì NULL)
- **Example Values** → Giá trị cụ thể

- **SCD Type** → Slowly changing dimension type: key (không đổi), 1,2,3, hay n/a
- **Display Folder** → Nhóm các attributes / facts tương tự trong một cube.
- **ETL Rules** → Các ETL rule cụ thể nếu có

Hoàn thành basic column information như sau:

Column Name	Display Name	Description	Unknown Member	Example Values	SCD Type	Display Folder	ETL Rules	Comments
CustomerKey	CustomerKey	Surrogate primary key	-1	1, 2, 3…	key			
CustomerID	CustomerID	Business key from source system (aka natural key)		ALFKI	key			
CompanyName	CompanyName	Customer's company Name		Bon app'	2			
ContactName	ContactName	Name of contact at the company		Pedro Alfanso	2			
ContactTitle	ContactTitle	Contact's job title		Owner, Sales Rep.	2			
CustomerCountry	CustomerCountry	Country of origin		USA	2			
CustomerRegion	CustomerRegion	State or province	N/A	WA	2			
CustomerCity	CustomerCity	Customer's City		Seattle	2			
CustomerPostalCode	CustomerPostalCode	Customer's postal code		13244	2			
RowIsCurrent	Row Is Current	Is this the current row for this member (Y/N)?	Y	Y, N	n/a	Exclude from cube	Standard SCD-2	
RowStartDate	Row Start Date	When did this row become valid for this member?	1/1/1900	1/24/2011	n/a	Exclude from cube	Standard SCD-2	
RowEndDate	Row End Date	When did this row become invalid? (12/31/9999 if current row)	12/31/9999	1/14/1998, 12/31/9999	n/a	Exclude from cube	Standard SCD-2	
RowChangeReason	Row Change Reason	Why did the row change last?	N/A		n/a	Exclude from cube	Standard SCD-2	
InsertAuditKey	InsertAuditKey	What process loaded this row?	-1		n/a	Exclude from cube	Standard Audit dim	
UpdateAuditKey	UpdateAuditKey	What process most recently updated this row?	-1		n/a	Exclude from cube	Standard Audit dim	

d. Hoàn thành target table information

Bước này giống với việc định nghĩa bảng.

- **Datatype, Size, Precision** – Chọn datatype (gồm cả size và precision) của thuộc tính.
- **Key?** – Để trống nếu không là key, ghi PK cho primary key, PK ID cho primary key (đại diện - surrogate), hay FK cho foreign key.
- **FK To** – Khi thuộc tính là FK, ta cần thêm dimension table và primary key mà nó tham chiếu.
- **NULL**? – Thuộc tính cho cho phép nhận giá trị null? Nên hạn chế việc này và dùng giá trị mặc định thay cho NULL.
- **Default Value** – Giá trị mặc định nếu thuộc tính không có giá trị.

	Target							
Column Name	Display Name	Datatype	Size	Precision	Key?	FK To	NULL?	Default Value
CustomerKey	CustomerKey	int			PK ID		N	
CustomerID	CustomerID	nvarchar	5				N	
CompanyName	CompanyName	nvarchar	40				N	
ContactName	ContactName	nvarchar	30				N	
ContactTitle	ContactTitle	nvarchar	30				N	
CustomerCountry	CustomerCountry	nvarchar	15				N	
CustomerRegion	CustomerRegion	nvarchar	15				N	N/A'
CustomerCity	CustomerCity	nvarchar	15				N	
CustomerPostalCode	CustomerPostalCode	nvarchar	10				N	
RowIsCurrent	Row Is Current	bit					N	TRUE
RowStartDate	Row Start Date	datetime					N	
RowEndDate	Row End Date	datetime					N	12/31/9999
RowChangeReason	Row Change Reason	nvarchar	200				N	
InsertAuditKey	InsertAuditKey	int			FK	DimAudit.AuditKey	N	
UpdateAuditKey	UpdateAuditKey	int			FK	DimAudit.AuditKey	N	

e. Hoàn thành source data information

Bước này sẽ hỗ trợ cho bước ETL.

- **Source System** – danh sách source system cho thuộc tính. Thuộc tính **Derived** là thuộc tính được tính toán từ các thuộc tính có sẵn.
- **Source Schema** – Nếu thuộc tính thuộc schema cụ thể, liệt kê tên schema.
- **Source Table** – Liệt kê tên table chứa thuộc tính.
- **Source Field Name** – Tên cột ứng với thuộc tính. Nếu cột được tính toán, ghi rõ biểu thức (Ví dụ: OrderQty * Price).

	Column Name	Display Name	Source System	Source Schema	Source Table	Source Field Name	Source Datatype
13	CustomerKey	CustomerKey	Derived				
14	CustomerID	CustomerID	Northwind	dbo	Customers	CustomerID	nvarchar
15	CompanyName	CompanyName	Northwind	dbo	Customers	CompanyName	nvarchar
16	ContactName	ContactName	Northwind	dbo	Customers	ContactName	nvarchar
17	ContactTitle	ContactTitle	Northwind	dbo	Customers	ContactTitle	nvarchar
18	CustomerCountry	CustomerCountry	Northwind	dbo	Customers	Country	nvarchar
19	CustomerRegion	CustomerRegion	Northwind	dbo	Customers	Region	nvarchar
20	CustomerCity	CustomerCity	Northwind	dbo	Customers	City	nvarchar
21	CustomerPostalCode	CustomerPostalCode	Northwind	dbo	Customers	PostalCode	nvarchar
22	RowIsCurrent	Row Is Current	Derived				
23	RowStartDate	Row Start Date	Derived				
24	RowEndDate	Row End Date	Derived				
25	RowChangeReason	Row Change Reason	Derived				
26	InsertAuditKey	InsertAuditKey	Derived				
27	UpdateAuditKey	UpdateAuditKey	Derived				

Các bước tiếp theo

Hoàn thành thiết kế chi tiết bằng cách lặp lại các bước 1-5 cho các dimension và fact table còn lại.

III. Yêu cầu đối với sinh viên

Sau khi hoàn thành thiết kế **sales reporting data mart** được mô tả theo các bước như phần II ở trên, sinh viên thực hành việc thiết kế **Order Fulfillment và Inventory Analysis data mart.**

Ở phần này, sinh viên lặp lại các bước như đã mô tả ở phần II để thiết kế **Order Fulfillment và Inventory Analysis data mart**. Sau đây là một số chỉ dẫn.

1. Thiết kế mức cao (high level design) - hoàn thành **High-Level-Dimensional-Modeling Excel Workbook**.

a. Hoàn thành **Detailed Bus Matrix** worksheet cho business processes mới này (**order fulfillment**). Khi xác định các dimension, chú ý quan sát quan hệ giữa các bảng trong

CSDL **Northwind**. Có thể truy vấn thông tin dùng SQL để quan sát rõ hơn các dimension và các thuộc tính của nó.

b. Hoàn thành **Attributes & Metrics** worksheet dựa trên **Detailed Bus Matrix** worksheet. Tìm cách thiết kế để có thể sử dụng lại các dimension ứng với các business processes khác nhau.
c. Nếu có vấn đề phát sinh, chưa rõ, hay những điều chưa thể thực hiện được, ghi chú nó lại trong **Issues List** worrksheet.

2. Thiết kế chi tiết (detailed design) - hoàn thành **Detailed-Dimensional-Modeling-Workbook**.

a. Cập nhật lại bất cứ thay đổi đối với thông tin của các dimension và fact table. Không nên tạo thêm Customer dimension vì nó đã được tạo cho sales reporting data mart. Thay vì vậy, xem thử có cần thay đổi nó cho data mart mới này. Cố gắng sử dụng lại các dimension ứng với các business processes khác nhau.
b. Tạo thêm các fact và dimension table ứng với yêu cầu mới dựa trên **Attributes & Metrics** worksheet.