Data integration (with SQL) Lab

I. Giới thiệu

Bài tập này giới thiệu quá trình thiết tích hợp dữ liệu vào kho dữ liệu đã xây dựng dựa trên thiết kế chi tiết ở bài tập trước.

Chuẩn bị

- **Microsoft SQL Server** phiên bản 2012 trở đi.
- **Northwind** database (https://northwinddatabase.codeplex.com/)
- **NorthwindSalesDW.sql** (đây là file được điều chỉnh từ script được sinh ra từ file làm mẫu cho **sale reporting** từ bài tập tuần trước **NorthwindSaleDW-Detailed-Dimensional-Modeling-Workbook-KimballU.xlsm**)

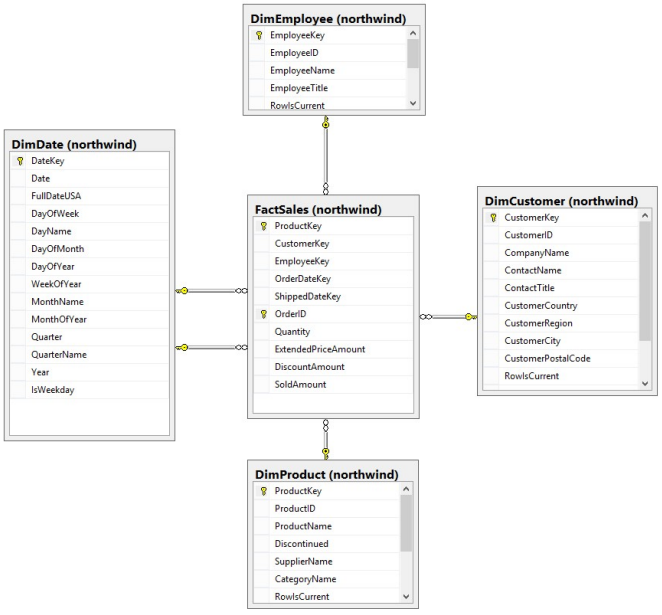
II. Hướng dẫn cho Northwind sales reporting

Ở phần này, ta sẽ hiện thực, load dữ liệu, và thử nghiệm **ROLAP** star schema cho **Northwind sales reporting** dựa trên thiết kế chi tiết đã thực hiện ở bài tập trước.

Thông tin về sales reporting data mart được mô tả ở hình sau.

Business Process Name	Fact Table	Fact Grain Type	Granulairty	Facts	Product	Customer	Employee	Order Date	Shipped Date
Sales Reporting	FactSales	Transaction	one row per order detail / line item.	Quantity, Unit Price , Discount Amount, Sold Amount, Freight Amount	x	x	x	x	x

Dựa trên thiết kế ở bài tập trước, ta dự định sẽ tạo star schema như sau:



1. Tạo star schema cho sales reporting

Ta sẽ tiến hành tạo star schema dùng SQL. Star schema gồm các bảng **DimCustomer**, **DimDate**, **DimEmployee**, **DimProduct**, và **FactSales**. Ta sẽ thực hiện các bước sau:

- Tạo database trong SQL Server, có thể đặt tên là **NorthwindSalesDW**. (phần sau sẽ dùng tên này để minh họa, nếu bạn đặt tên khác thì phải thay đổi cho phù hợp)
- Sử dụng **NorthwindSalesDW.sql** để tạo các câu lệnh tạo bảng (**CREATE TABLE**) dùng SQL.
- Khi xóa bảng, cần xóa **fact table** trước, sau đó mới đến các **dimension table** (do ràng buộc khóa ngoại). Thực hiện ngược lại cho quá trình tạo bảng.

2. Staging

Trước khi **load** dữ liệu vào **star schema**, ta sẽ thực hiện bước **staging** dữ liệu nguồn. Tạo một CSDL staging, có thể đặt tên là **NorthwindSalesDWStage**.

Staging bảng Customers

Thực hiện truy vấn sau

```
select CustomerID
       , CompanyName
       , ContactName
       , ContactTitle
       , Address
       , City
       , Region
       , PostalCode
       , Country
into   [dbo].[NorthwindStageCustomers]
from   [NORTHWND].[dbo].[Customers]
```

Staging bảng Employees

```
select EmployeeID
       , FirstName
       , LastName
       , Title
into   [dbo].[NorthwindStageEmployees]
from   [NORTHWND].[dbo].[Employees]
```

Staging bảng Products

```
select ProductID
       , ProductName
       , Discontinued
       , CompanyName
       , CategoryName
into   [dbo].[NorthwindStageProducts]
from   [NORTHWND].[dbo].[Products] p
       join [NORTHWND].[dbo].[Suppliers] s
              on p.[SupplierID] = s.[SupplierID]
       join [NORTHWND].[dbo].[Categories] c
              on c.[CategoryID] = p.[CategoryID]
```

Staging bảng Date

Bảng **Date** thể hiện **Time dimension** (nó là một **conformed dimension**, dimension được dùng chung cho nhiều fact table) và được dùng chung cho các **data mart** khác nhau. Bảng **Date** cần lưu giữ thông tin về ngày tháng trong khoảng thời gian nào? Ta sẽ truy vấn bảng **Orders** để tìm **min** và **max Order Date** và **Shipped Date**.

```
select min(OrderDate) As StartOrderDate
       , max(OrderDate) As EndOrderDate
       , min(ShippedDate) As StartShippedDate
       , min(ShippedDate) As EndShippedDate
from   [NORTHWND].[dbo].[Orders]
```

Kết quả trả về là bắt đầu từ năm **1996** đến năm **1998** như hình sau.

	StartOrderDate	EndOrderDate	StartShippedDate	EndShippedDate
1	1996-07-04 00:00:00.000	1998-05-06 00:00:00.000	1996-07-10 00:00:00.000	1996-07-10 00:00:00.000

Dựa trên kết quả này ta có thể staging cho bảng **Date** như sau.

```
select *
into [dbo].[NorthwindStageDate]
from [Temp].[dbo].[DateDimension]
where year between 1996 and 1998
```

Ghi chú: Sử dụng thông tin trong file **Ch3-SampleDateDim.xls** (**https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-toolkit/**) để import data cho **[Temp].[dbo].[DateDimension]**. Sử dụng lệnh tạo bảng và dữ liệu mẫu từ **Ch3-SampleDateDim.xls** để tạo bảng và đưa dữ liệu vào [Temp].[dbo].[DateDimension]. Bạn cần sửa lại **dòng đầu tiên** cột **full date** của sheet "**Data dimension**" trong **Ch3-SampleDateDim.xls** để chọn miền ngày tháng phù hợp.

Staging fact table

```
select ProductID
       , od.OrderID
       , CustomerID
       , EmployeeID
       , OrderDate
       , ShippedDate
       , UnitPrice
       , Quantity
       , Discount
into   [dbo].[NorthwindStageSales]
from   [NORTHWND].[dbo].[Order Details] od
       join [NORTHWND].[dbo].[Orders]o
              on od.OrderID = o.OrderID
```

Sau khi tao xong các stage table, ta sẽ có 5 bảng **NorthwindStageCustomers**, **NorthwindStageDates**, **NorthwindStageEmployees**, **NorthwindStageProducts** và **NorthwindStageSales**.

3. Load dữ liệu từ stage table vào data warehouse

Load bảng DimCustomer

```
insert into NorthwindSalesDW.dbo.DimCustomer
       (CustomerId, CompanyName, ContactName, ContactTitle,
       CustomerCountry, CustomerRegion, CustomerCity, CustomerPostalCode)
select
       CustomerID, CompanyName, ContactName, ContactTitle,
       Country, Region, City, PostalCode
from NorthwindSalesDWStage.dbo.NorthwindStageCustomers
```

```
Msg 515, Level 16, State 2, Line 1
Cannot insert the value NULL into column 'CustomerRegion', table
'NorthwindSalesDW.dbo.DimCustomer'; column does not allow nulls. INSERT fails.
```

Đây là lỗi phổ biến khi insert dữ liệu vào CSDL hoặc KDL. Lý do là vì bảng NorthwindStageCustomers trong CSDL NorthwindSalesDWStage cho phép thuộc tính Region nhận giá trị NULL, trong khi bảng DimCustomer trong CSDL NorthwindSalesDW không cho phép thuộc tính CustomerRegion nhận giá trị NULL. Ta có thể xử lý trường hợp này bằng cách thay giá trị NULL bằng giá trị mặc định 'N/A' (Not Available) dùng case when.

```
insert into NorthwindSalesDW.dbo.DimCustomer
       (CustomerId, CompanyName, ContactName, ContactTitle,
       CustomerCountry, CustomerRegion, CustomerCity, CustomerPostalCode)
select
       CustomerID, CompanyName, ContactName, ContactTitle, Country,
       case when Region is null then 'N/A' else Region end,
       City, PostalCode
from NorthwindSalesDWStage.dbo.NorthwindStageCustomers
```

Load bảng DimEmployee

```
insert into NorthwindSalesDW.dbo.DimEmployee
       (EmployeeId, EmployeeName, EmployeeTitle)
select
       EmployeeId, FirstName + ' ' + LastName, Title
from NorthwindSalesDWStage.dbo.NorthwindStageCustomers
```

Load bảng DimProduct

```
insert into NorthwindSalesDW.dbo.DimProduct
       (ProductID, ProductName, Discontinued, SupplierName,CategoryName)
select
       ProductID, ProductName,
       case when Discontinued = 1 then 'Y' else 'N' end,
       CompanyName,CategoryName
from NorthwindSalesDWStage.dbo.NorthwindStageProducts
```

Load bảng DimDate

```
insert into NorthwindSalesDW.dbo.DimDate
       (DateKey, Date, DayOfWeek, DayName, DayOfMonth, DayOfYear,
       WeekOfYear, MonthName, MonthOfYear, Quarter, QuarterName, Year, IsAWeekday)
select
       date_key, full_date, day_of_week, day_name, day_num_in_month,
       day_num_overall, week_num_in_year, month_name, month, quarter,
       case
```

```
            when quarter >= 1 and quarter <= 3 then 'First'
            when quarter >= 4 and quarter <= 6 then 'Second'
            when quarter >= 7 and quarter <= 9 then 'Third'
            when quarter >= 10 and quarter <= 12 then 'Fourth'
      end,
      year, weekday_flag
from NorthwindSalesDWStage.dbo.NorthwindStageDate
```

Load bảng FactSales

Load bảng **FactSales** hơi phức tạp một chút vì bảng **NorthwindStageSales** của CSDL **NorthwindStageDW** chỉ chứa các **business key** (**CustomerID, EmployeeID, ProductID**) trong khi bảng **FactSales** của CSDL **NorthwindDWSales** cần các **surrogate key** (**CustomerKey, EmployeeKey, ProductKey**). Để lấy được các surrogate key, ta cần kết bảng **NorthwindStageSales** của CSDL **NorthwindStageDW** với các bảng **Dimension** của CSDL **NorthwindSalesDW**. Ngoài ra, ta cũng cần chuyển **OrderDate** và **ShippedDate** của bảng **NorthwindStageSales** từ kiểu **datetime** sang kiểu **int** dạng **YYYYMMDD** để phù hợp với kiểu dữ liệu của **OrderDateKey** và **ShippedDateKey** của bảng **FactSales**.

```
insert into NorthwindSalesDW.dbo.FactSales
      (ProductKey, CustomerKey, EmployeeKey, OrderDateKey, ShippedDateKey,
      OrderID, Quantity, ExtendedPriceAmount, DiscountAmount, SoldAmount)
select p.ProductKey, c.CustomerKey, e.EmployeeKey,
      Day(s.OrderDate) + MONTH(s.OrderDate) * 100 + YEAR(s.OrderDate) * 10000 As
OrderDateKey,
      case when s.ShippedDate is null then null
      else Day(s.ShippedDate) + MONTH(s.ShippedDate) * 100 + YEAR(s.ShippedDate) *
10000
      end as ShippedDateKey,
      s.OrderID,
      s.Quantity,
      s.Quantity * s.UnitPrice as ExtendedPriceAmount,
      s.Quantity * s.UnitPrice * s.Discount as DiscountAmount,
      s.Quantity * s.UnitPrice * (1 - s.Discount) as SoldAmount
from NorthwindSalesDWStage.dbo.NorthwindStageSales s
      join NorthwindSalesDW.dbo.DimCustomer c
      on s.CustomerID = c.CustomerId
      join NorthwindSalesDW.dbo.DimEmployee e
      on s.EmployeeID = e.EmployeeId
      join NorthwindSalesDW.dbo.DimProduct p
      on s.ProductID = p.ProductID
```

III. Yêu cầu đối với sinh viên

Sau khi hoàn thành load dữ liệu cho **Sales Reporting** data mart được mô tả theo các bước như phần II ở trên, sinh viên thực hành việc load dữ liệu cho **Order Fulfillment** và **Inventory Analysis** data mart. Cụ thể, SV cần hoàn thành các file sau:

NorthwindOrderFulfillmentDW.sql, NorthwindInventoryAnalysisDW.sql (create table)

NorthwindOrderFulfillmentStage.sql, NorthwindInventoryAnalysisStage.sql (staging)

NorthwindOrderFulfillmentLoad.sql, NorthwindInventoryAnalysisLoad.sql (load)

NorthwindSaleDW-Detailed-Dimensional-Modeling-Workbook-KimballU.xlsm