Bài tập Thực hành

Các hệ thống Thông minh Doanh nghiệp

NỘI DUNG

- **TUẦN 1:**

  Tìm hiểu cấu trúc kho dữ liệu AdventureWorksDW2012

  Tạo Data Warehouse dùng SQL Server Integration Services
- **TUẦN 2-3:**

  Mô hình hóa dữ liệu đa chiều (Multidimensional Modeling)

  Phân tích và Trình bày dữ liệu
- **TUẦN 4-6:**

  Khai phá dữ liệu:

  ☺ Bài toán Phân lớp

  ☺ Bài toán Gom nhóm

  ☺ Bài toán Luật kết hợp
- **TUẦN 7-9:**

  Bài tập tổng hợp (Phân tích và Khai phá dữ liệu

  theo yêu cầu nghiệp vụ cho trước)
- **TUẦN 10:**

  Ôn tập - Kiểm tra thực hành

TUẦN 1

Tìm hiểu cấu trúc kho dữ liệu AdventureWorksDW2012

Tạo Data Warehouse dùng SQL Server Integration Services

Giới thiệu – Kiến thức tổng quan:

1. **Công ty Adventure Works Cycles,** một công ty hư cấu, có cơ sở dữ liệu mẫu AdventureWorksDW2012, là nhà sản xuất xe đạp đa quốc gia lớn chuyên sản xuất và kinh doanh xe đạp kim loại và nhựa tổng hợp, cung cấp cho thị trường Bắc Mỹ, Châu Âu và Châu Á. Công ty có trụ sở đặt tại Bothell, Washington; với khoảng 300 nhân viên làm việc, trong đó 29 người là đại diện bán hàng. Công ty Adventure Works Cycles phân phối các sản phẩm thông qua các cửa hàng bán lẻ của các đại lý. Những đại lý này được đặt tại Úc, Canada, Pháp, Đức, Vương quốc Anh và Hoa Kỳ. Adventure Works Cycles cũng bán hàng cho các khách hàng cá nhân trên toàn thế giới thông qua Internet.

Adventure Works Cycles đang tìm kiếm để mở rộng thị phần của mình bằng cách tập trung bán hàng cho khách hàng tốt nhất, mở rộng sản phẩm thông qua một trang web bên ngoài và giảm chi phí bán hàng nhờ giảm chi phí sản xuất.

Adventure Works Cycles có năm nhóm sản phẩm chính:

- Bikes (Xe đạp) - Ba dòng sản phẩm xe đạp chính: Mountain (xe đạp leo núi), Road (xe đạp đua), và Touring (xe đạp đi tour)
- Accessories (Phụ kiện) - Ví dụ helmets (mũ bảo hiểm), bottles (chai nước)…
- Clothing (Quần áo) - Ví dụ jerseys (áo phông), biking shorts (quần soóc đi xe đạp)…
- Components (Phụ tùng) - Ví dụ bottom brackets (trục giữa), frames (khung sườn)…
- Services (Dịch vụ) - Ví dụ dịch vụ cao cấp, dịch vụ tiêu chuẩn

2. **Data warehouse (kho dữ liệu):**

- **Bảng Fact** thường có 2 loại cột: khóa ngoại ứng với các bảng dimension và dữ liệu của bảng fact gọi là measure (hay các thuộc tính phụ thuộc)- chứa các số liệu dùng để phân tích hiệu quả kinh doanh. Measure đại diện cho 1 cột chứa giá trị định lượng, thường là các số liệu chi tiết hay các con số, có được qua dùng các hàm tổng hợp (gộp) dữ liệu từ các dimension liên kết với bảng fact.
- **Bảng Dimension** thường cấu tạo bởi 1 hay nhiều dữ liệu phân cấp (hierarchies). Khóa chính của các bảng dimension là 1 thành phần của khóa chính ghép trong bảng fact. Các thuộc tính trong bảng dimension dùng để mô tả các giá trị của dimension, chúng thường là giá trị mô tả dạng văn bản. Chức năng chính của các dimension là lọc, gom nhóm và gán nhãn cho dữ liệu.

VD: bảng fact lưu giá trị sales, bảng dimension lưu giá trị về vùng miền địa lý (markets, cities), clients, products, times, channels.

- **Cube** (khối dữ liệu) là đơn vị cơ bản của dữ liệu đa chiều (multidimensional data), trên đó ta có thể phân tích 1 phần kho dữ liệu.

I. TÌM HIỂU CẤU TRÚC KHO DỮ LIỆU ADVENTUREWORKSDW2012

A. **Tìm hiểu các loại sơ đồ:** Dùng data warehouse AdventureWorksDW2012:

1. Chọn các bảng Fact và các bảng Dimension thích hợp để tạo sơ đồ hình sao (star schema)

2. Chọn các bảng Fact và các bảng Dimension thích hợp để tạo sơ đồ hình bông tuyết (snowflake schema)

3. Chọn các bảng Fact và các bảng Dimension thích hợp để tạo sơ đồ chòm sao (constellation schema)

B. **Tìm hiểu cấu trúc các dimension sau:** DimDate**,** DimGeography, DimSalesTerritory. Chỉ ra các thuộc tính phân cấp (hierarchies) trên các dimension trên, liệt kê từ chi tiết đến tổng quát.

C. **Xây dựng các sơ đồ phục vụ khai phá dữ liệu:**

1. Internet Sales diagram: hãy tìm các bảng liên quan đến việc bán trực tiếp các sản phẩm của công ty Adventure Works Cycles cho khách hàng qua Internet (gồm 6 bảng dimension trực tiếp và 2 bảng fact). Đây có phải sơ đồ hình sao? Truy vấn kho dữ liệu để biết:

- Số lượng khách hàng theo phân bố vùng miền (North America, Europe…), đặc trưng khách hàng: thu nhập, khoảng cách đi làm, trình độ học vấn, số con, số xe hơi…
- Các chủng loại sản phẩm (Product Subcategory), mô tả sản phẩm bằng các thứ tiếng, màu sắc, trọng lượng, giá bán…
- Tên các chương trình khuyến mãi, tỉ lệ chiết khấu, thời gian khuyến mãi…
- Thử phân loại sản phẩm theo subcategory, model và product

2. Reseller Sales diagram: hãy tìm các bảng liên quan đến việc bán các sản phẩm của công ty Adventure Works Cycles cho các cửa hàng bán lẻ (gồm 7 bảng dimension trực tiếp và 1 bảng fact). Đây có phải sơ đồ hình bông tuyết? Truy vấn kho dữ liệu để biết:

- Các đặc trưng chính của sản phẩm, thời gian xuất xưởng, giá dành cho đại lý…
- Thông tin đại lý bán lẻ: quốc gia, số lượng, doanh số hàng năm…
- Các chương trình khuyến mãi và qui định khuyến mãi…

D. **Các bước thiết kế sơ đồ hình sao**:

1. Xác định quy trình kinh doanh cần phân tích (VD: Sales - bán hàng)

2. Xác định các measure trong bảng fact (VD: sales dollar) tương ứng với các dimension kết nối với bảng fact (VD: product dimension, location dimension...)

3. Liệt kê các cột mô tả từng dimension (VD: region name, branch dimension...)

4. Xác định mức tổng hợp thấp nhất (lowest level of summary) trong một bảng fact (VD: sales dollar)

Áp dụng các bước thiết kế nêu trên, vẽ diagram liên quan đến bảng **FactInternetSales** và bảng **FactResellerSales** (dùng kết quả câu C nêu trên), chỉ ra các measure ứng với 2 bảng nêu trên. Nêu các dự kiến khai phá dữ liệu trên 2 sơ đồ này?

II. TẠO DATA WAREHOUSE DÙNG SQL SERVER INTEGRATION SERVICES (SSIS):

- ☺ SSIS (SQL Server Integration Services): *thực hiện ETL (Extract, Transform, Loading) trên dữ liệu; kết hợp, chuẩn hóa, phân chia và phân tích dữ liệu.*
- ☺ Nhiệm vụ bài tập:
  - ✓ Tạo Data Warehouse **AdventureWorksDW_Small** từ SQL Scripts.
  - ✓ Phát sinh dữ liệu (populate) Data Warehouse dùng SSIS: Load dữ liệu từ cơ sở dữ liệu **AdventureWorks2012** vào kho dữ liệu AdventureWorksDW_Small vừa tạo.

(Dùng **SQL Server Management Studio** từ câu 1-5)

1. Mở **SQL Server Management Studio** và kết nối vào server

2. Attach (hay Restore) cơ sở dữ liệu **AdventureWorks2012**

3. Tìm hiểu các bảng sau (dùng SELECT TOP 1000 ROWS) :

   a. Production.Product

   b. Production.ProductCategory

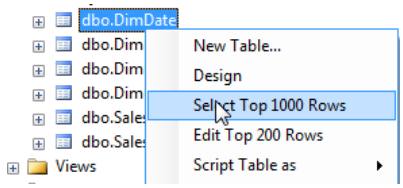
   c. Production.ProductSubCategory

   d. Sales.SalesOrderHeader

   e. Sales.SalesOrderDetail

4. Mở file **DWCreateScript.sql**, xem nội dung script, chạy script để có CSDL và cấu trúc các bảng

5. Tìm hiểu CSDL kho dữ liệu **AdventureWorksDW_Small** vừa tạo: mở bảng DimDate để kiểm tra nội dung; các bảng khác chưa có dữ liệu. Xem lại cách tạo bảng DimDate trong file DWCreateScript.sql



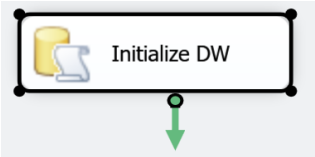
(Dùng **SQL Server Data Tools -Visual Studio 2012** từ câu 6-9)

6. Tạo **SSIS Package**: Từ câu này trở đi dùng **SQL Server Data Tools 2012** để tạo SSIS project (Project này chứa package dùng để populate Data Warehouse). Chọn **Integration Services Project.**

7. Tạo **Connection Objects**: Tại **Solution Explorer**, Right click trên Connection Managers > New Connection Manager, chọn OLEDB > Add, nhập Server Name, chọn CSDL từ danh sách có sẵn, lần lượt là *AdventureWorks2012* và *AdventureWorksDW_Small* (New Connection 2 lần).

8. Khởi tạo **Data Warehouse**:

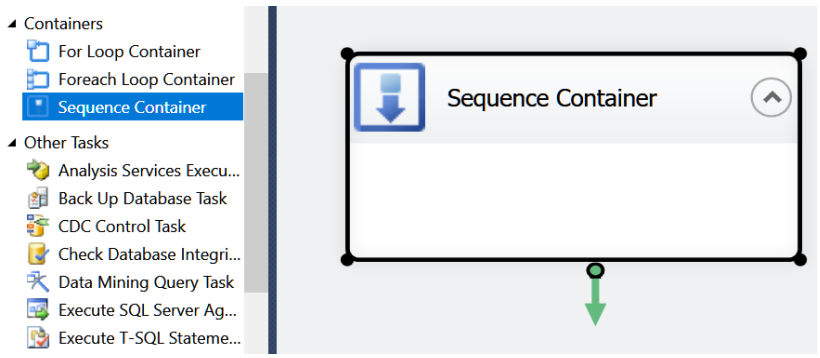
   i. Click View | Toolbox để hiển thị Toolbox.
   ii. Từ **SSIS Toolbox**, kéo **Execute SQL Task** vào **Control Flow tab**.
   iii. Right click trên **Execute SQL Task** và chọn Edit để hiển thị **Execute SQL Task Editor**.
   iv. Nhập Name là **Initialize DW**



   v. Nhập Description là **Clear down DW Tables**
   vi. Chọn Connection đến **AdventureWorksDW_Small**
   vii. Thiết lập (gõ vào) SQL Statement là **procDWInitialize** (thủ tục cần chạy)
   viii. Click OK
   ix. Right Click trên **Initialize DW task** và chọn **Execute Task**
   x. Kiểm tra task chạy thành công (Có dấu ✓ màu xanh)
   xi. Chọn trên menu Debug | Stop Debugging (Shift + F5)

9. Nạp dữ liệu (**Load Data**)

   i. Từ **SSIS Toolbox**, trong **Containers** section, kéo **Sequence Container** vào **Control Flow** tab.



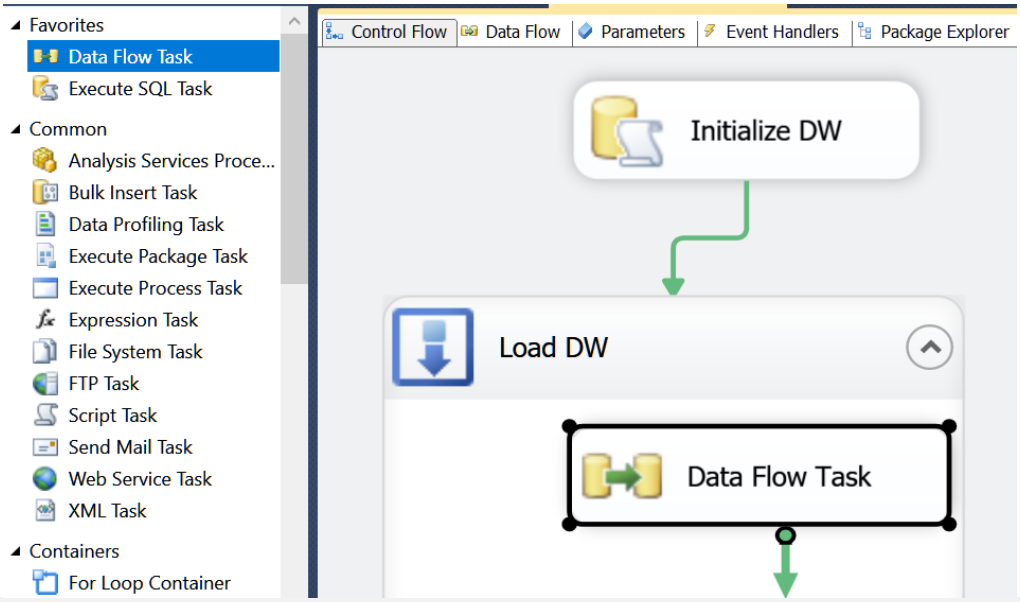
   Sequence Container là nơi chứa các item, giúp dễ quản lý các task có liên quan

   ii. Đổi tên **Sequence Container** thành **Load DW**

iii. Click lên **Initialize DW** task

iv. Kéo **Green Arrow** vào **Load DW** sequence container.

v. **Load dữ liệu từ 1 bảng trong CSDL vào DW**- ở bước này ta thực hiện data task cho bảng **ProductCategory**: Từ **SSIS Tools** kéo **Data Flow Task** vào **Load DW** sequence container:



vi. Đổi tên **Data Flow Task** thành **Load ProductCategory**

vii. Right Click trên **Load ProductCategory** và chọn **Edit** – điều này dẫn bạn đến **Data Flow** tab.

viii. Kéo **OLE DB Source** từ **Other Sources** section trên SSIS Toolbox vào **Data Flow** tab.

ix. Đổi tên **OLE DB Source** thành **ProductCategory DB**

x. Right Click trên **ProductCategory DB** và chọn **Edit**

xi. Chọn **AdventureWorks2012** từ **OLE DB Connection Manager** dropdown.

xii. Chọn **Table or View** từ **Data Access Mode**

xiii. Chọn **[Production].[ProductCategory]** trong **Name of the table or the view**.

xiv. Click **OK**

xv. Kéo **OLE DB Destination** từ **Other Destinations** section trên **SSIS Toolbox** vào **Data Flow** tab.

xvi. Đổi tên **OLE DB Destination** thành **ProductCategory DW**

xvii. Kéo **Blue Arrow** from **ProductCategory DB** vào **ProductCategory DW**

xviii. Right Click trên **ProductCategory DW** và chọn **Edit**

xix. Chọn **AdventureWorksDW_Small** từ **OLE DB connection Manager** dropdown.

xx. Chọn **Table or View – fast load** từ **Data Access Mode**

xxi. Chọn **[dbo].[DimProductCategory]** trong **Name of the table or the view**.

xxii. Click **Mappings** từ cột bên trái để xem các ánh xạ giữa bảng trong CSDL và bảng trong DW

xxiii. Để ý có đường thẳng nối ProductCategoryID trên cả 2 bảng (gọi là ánh xạ).

xxiv. Ánh xạ (Click lên **Name** và kéo vào **EnglishProductCategoryName**)

xxv. Click **OK**

xxvi. Click lên **Control Flow** tab.

xxvii. Right Click lên **Load DW** sequence container

xxviii. Chọn **Execute Container**

xxix. **Lặp lại các bước từ v – xxviii** cho 4 bảng sau, nhớ đổi **tên bảng** cho phù hợp với câu đang làm:

a. **Production.Product**

b. **Production.ProductSubCategory**

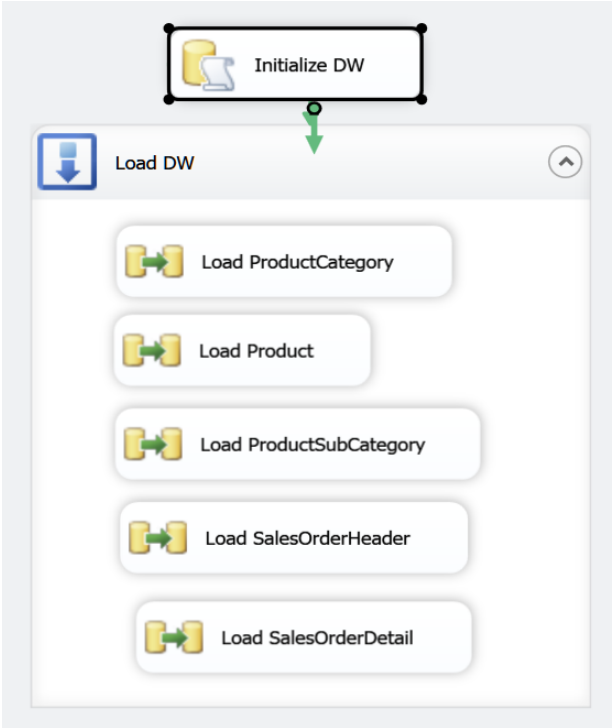
   i. Ánh xạ **Name** đến **EnglishProductSubcategoryName**

   ii. Ánh xạ **ProductCategoryID** đến **ProductCategoryID**

c. **Sales.SalesOrderHeader**

d. **Sales.SalesOrderDetail**

Lưu ý: xóa ánh xạ trên cột **SalesOrderDetail.LineTotal**, vì đây là cột tính toán, không cần import vào DW. Để xóa ánh xạ, right click lên đường nối 2 **Line Total** và chọn **Delete**.



xxx. Ấn **F5** (hay R_Click và chọn Execute Task) để thực thi toàn bộ package.

(Câu 10: **Trở lại dùng SQL Server Management Studio)**

10. Kiểm tra dữ liệu của kho dữ liệu **AdventureWorksDW_Small** vừa sinh ra. Xem thông tin các bảng sau bằng SELECT TOP 1000 ROWS:
    a. DimProduct
    b. DimProductCategory
    c. DimProductSubCategory
    d. SalesOrderHeader
    e. SalesOrderDetail
    f. DimDate

Mô hình hóa dữ liệu đa chiều (Multidimensional Modeling)
Phân tích và Trình bày dữ liệu
TUẦN 2

Các bước thực hiện:

a. **Khởi động** Microsoft Visual Studio 2012.

b. **Tạo Project**: Click New Project > Recent Templates > Business Intelligence> Analysis Services Multidimensional and Datamining Project. Đặt tên Project trong ô Name . Chỉ định đường dẫn trong ô Location. Click OK.

c. **Tạo Data Source**: Trong Solution Explorer, Right click trên Data Source > Click New Data Source > Next > New. Nhập Server name, thông tin xác thực SQL Server. Trong Select or enter a database name: chọn tên CSDL, OK > Next > Inherit > Next > Finish.

d. **Tạo Data Source View**: Trong Solution Explorer, Right Click trên Data Source View > Click New Data Source View > Next > Next. Chọn các bảng đã tạo trước đó, đưa vào Included objects. Click Next> Finish. Data Source View đã sẵn sàng để sử dụng.

e. **Tạo khối dữ liệu (Cube):** Trong Solution Explorer, Right Click trên Cubes > Click New Cube > Next. Chọn Use existing Tables > Next. Chọn bảng Fact từ Measure Group Tables > Next. Chọn Measure từ danh sách, click Next > Next. Đặt tên cho Cube, click Finish. Chú ý quan sát tên các dimensions trong cửa sổ Solution Explorer.

f. **Hiệu chỉnh dimension**: Trong Solution Explorer, double click lên dimension cụ thể, kéo các thuộc tính từ bảng trong vùng Data Source View ở bên phải, thả vào vùng Attributes ở bên trái.

g. **Tạo phân cấp thuộc tính (Attribute Hierarchy)** trong dimension cụ thể: kéo các thuộc tính từ vùng Attributes ở bên trái sang vùng Hierarchies ở giữa.

h. **Triển khai (deploy) khối dữ liệu (Cube)**: Trong Solution Explorer, right click trên tên Project > Click Properties. Trong Configuration Properties, chọn Deployment, chọn các Options tùy ý > Click OK. Trong Solution Explorer, right click trên Project Name > Click Deploy.

i. **Xử lý Cube**: trong Solution Explorer, right click trên tên Project Name > Click Process > Run, rồi click Close.

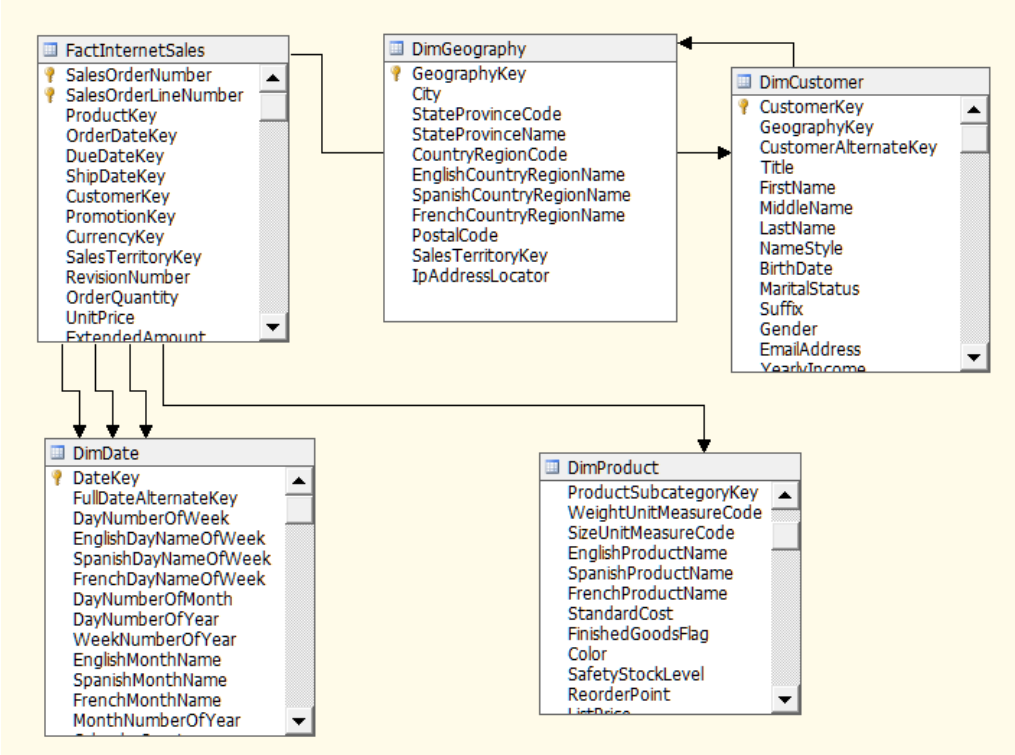
j. **Duyệt Cube để phân tích số liệu**: Trong Solution Explorer, right click trên tên Cube> Click Browse. Kéo và thả các Dimension Attributes và Measures vào vùng trống ở giữa tên "Drag level or measures here to add to the query" .

<u>Bài 1:</u> Tạo mô hình dữ liệu đa chiều liên quan đến mảng bán hàng qua mạng, bao gồm các bảng sau DimCustomer, DimDate, DimGeography, DimProduct, FactInternetSales. Thống kê số liệu theo hình nêu ở mục k của bài tập này.

a. **Creating an Analysis Services Project**

b. **Defining a Data Source**: chọn kho dữ liệu **AdventureWorksDW2012**

c. **Defining a Data Source View**: chọn 5 bảng theo đề bài



d. **Modifying Default Table Names**: dùng **FriendlyName** property để bỏ từ Fact, Dim trước tên các bảng

e. **Defining a Dimension**:

- ✓ Với bảng **Date**: chọn các thuộc tính cần thiết là **Date Key, Full Date Alternate Key, English Month Name, Calendar Quarter, Calendar Year, Calendar Semester**
- ✓ Đổi **Attribute Type** của thuộc tính **Full Date Alternate Key** từ **Regular** sang **Date** (Date > Calendar> Date).
- ✓ Tương tự đổi kiểu thuộc tính của:
  - **English Month Name** thành **Month**
  - **Calendar Quarter** thành **Quarter**
  - **Calendar Year** thành **Year**
  - **Calendar Semester** thành **Half Year**

f. **Defining a Cube:**

- ✓ Chọn **InternetSales** là measure group table (bảng Fact), không chứa các measure: **Promotion Key, Currency Key, Sales Territory Key, Revision Number**
- ✓ Trong **Select New Dimensions** page, giữ lại các dimension **Customer, Geography và Product**, bỏ chọn **InternetSales** check box

g. **Adding Attributes to Dimensions**:

- Từ bảng **Customer** trong Data Source View: BirthDate, MaritalStatus, Gender, EmailAddress, YearlyIncome, TotalChildren, NumberChildrenAtHome, EnglishEducation, EnglishOccupation, HouseOwnerFlag, NumberCarsOwned, Phone, DateFirstPurchase, CommuteDistance

- Từ bảng **Geography** trong Data Source View: City, StateProvinceName, EnglishCountryRegionName, PostalCode

- Từ bảng **Product** trong Data Source View: StandardCost, Color, SafetyStockLevel, ReorderPoint, ListPrice, Size, SizeRange, Weight, DaysToManufacture, ProductLine, DealerPrice, Class, Style, ModelName, StartDate, EndDate, Status

h. **Reviewing Cube and Dimension Properties**

- Mở rộng **Internet Sales** measure group

i. **Deploying an Analysis Services Project**

j. **Browsing the Cube**

k. **Modifying Measures, Attributes and Hierarchies**

- **Order Quantity** có **FormatString** list là #,#

- **Unit Price Discount Pct** có **FormatString** list dạng Percent.

Creating a Hierarchy:

- **Country-Region-> State-Province-> City**

- **Product Line->Model Name->Product Name**

- **Calendar Year->Calendar Semester->Calendar Quarter->**

->English Month Name

l**. Adding a Named Calculation:**

- **FullName**: gõ vào **Expression** box:

```
CASE
    WHEN MiddleName IS NULL THEN
    FirstName + ' ' + LastName
```

```
    ELSE
    FirstName + ' ' + MiddleName + ' ' + LastName
END
```

- **ProductLineName**: gõ vào **Expression** box:

```
CASE ProductLine
    WHEN 'M' THEN 'Mountain'
    WHEN 'R' THEN 'Road'
    WHEN 'S' THEN 'Accessory'
    WHEN 'T' THEN 'Touring'
    ELSE 'Components'
END
```

- **SimpleDate**: gõ vào **Expression** box:

```
DATENAME(mm, FullDateAlternateKey) + '  ' +
DATENAME(dd, FullDateAlternateKey) + ', ' +
DATENAME(yy, FullDateAlternateKey)
```

- **MonthName**: gõ vào **Expression** box:

```
EnglishMonthName+' '+ CONVERT(CHAR (4), CalendarYear)
```

- **CalendarQuarterDesc**: gõ vào **Expression** box:

```
'Q' + CONVERT(CHAR (1), CalendarQuarter) +' '+ 'CY ' +
CONVERT(CHAR (4), CalendarYear)
```

- **CalendarSemesterDesc**: gõ vào **Expression** box:

```
CASE
    WHEN CalendarSemester = 1 THEN 'H1' + '  ' + 'CY' +  '  ' + CONVERT(CHAR(4), CalendarYear)
    ELSE  'H2' + '  ' + 'CY' + '  ' + CONVERT(CHAR(4), CalendarYear)
END
```

m. **Defining Attribute Relationships**

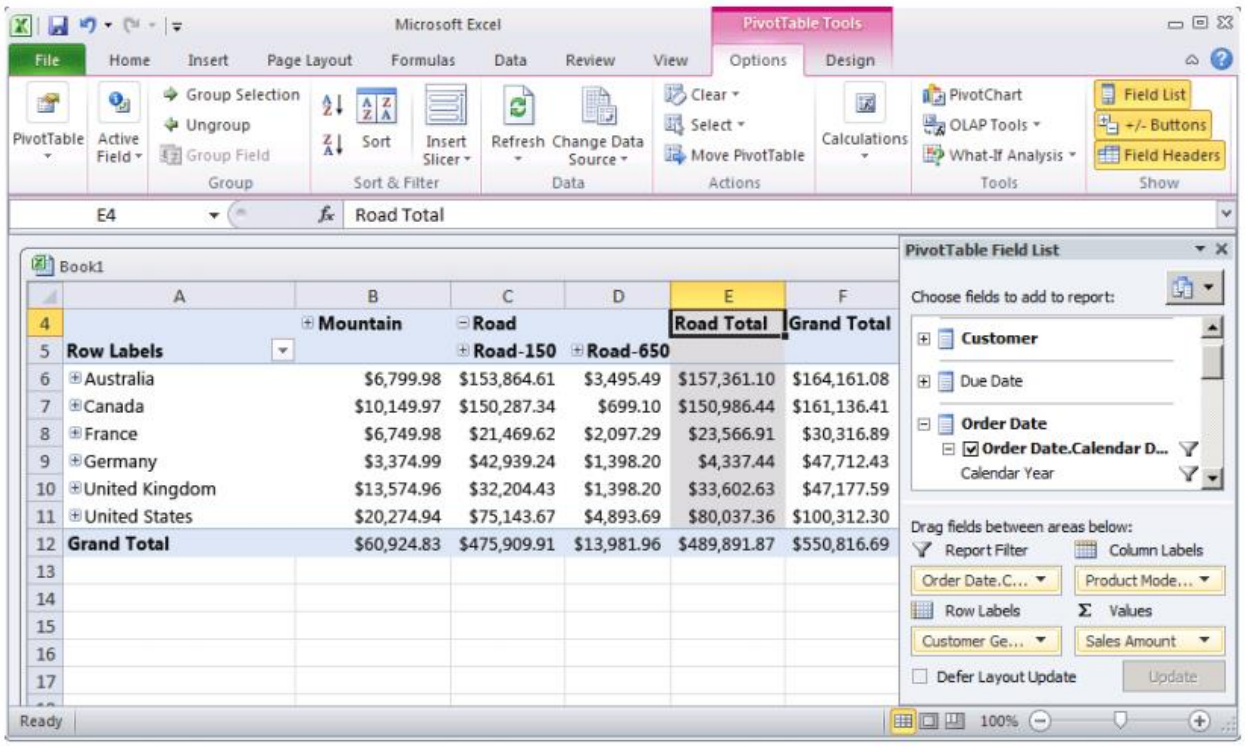
- **City** attribute: **Source Attribute** is **City**, **Related Attribute** is **State-Province**.

- **State-Province** attribute: **Source Attribute** is **StateProvince**, **Related Attribute** is **Country-Region**

- **Model Name** attribute: **Source Attribute** is **Model**, **Related Attribute** is **Product Line**.

- **Calendar Quarter** attribute: **Source Attribute** is **Calendar Quarter**, **Related Attribute** to **Calendar Semester**.

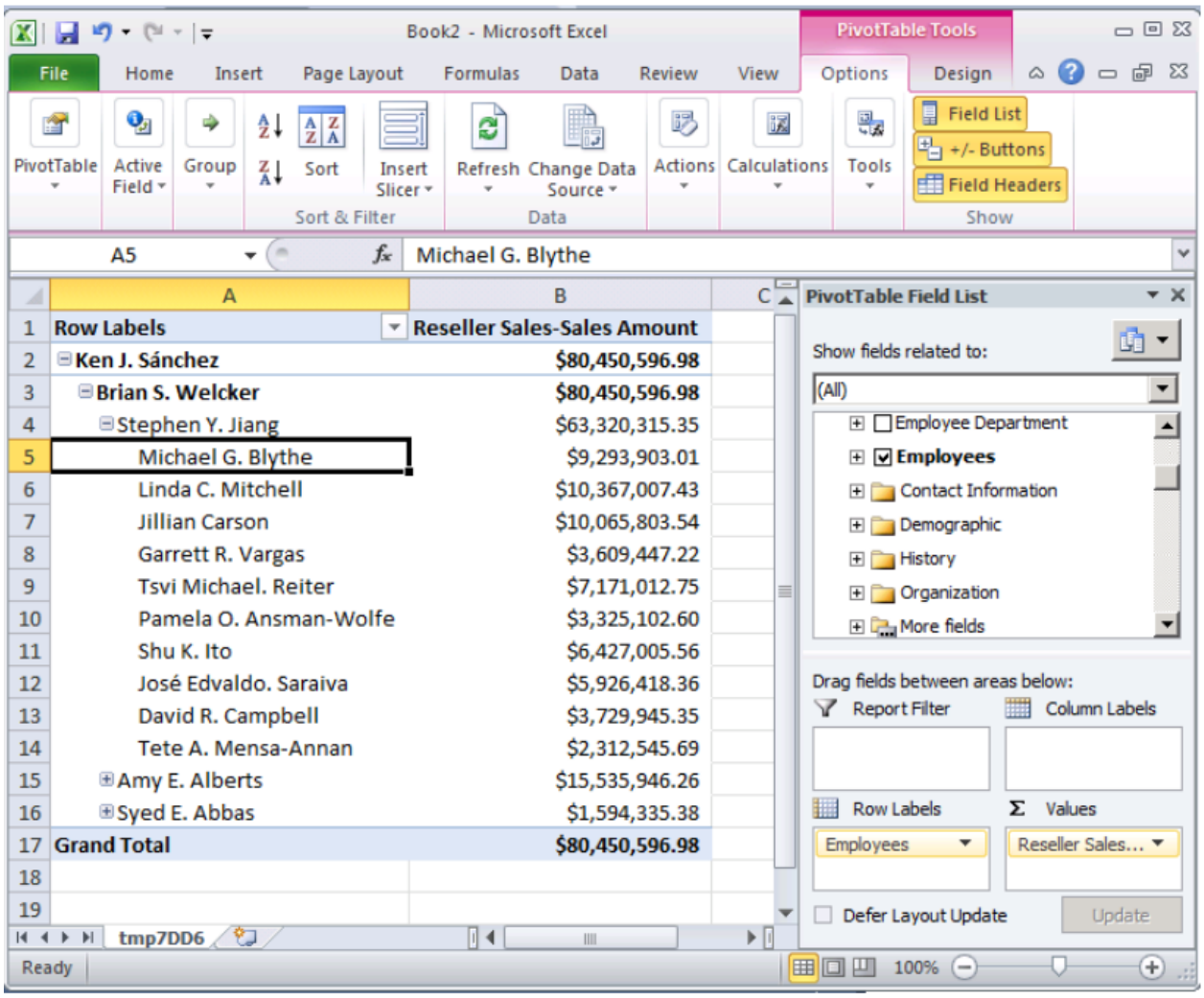
n. **Browsing the Deployed Cube**, chuyển kết quả qua Excel, tạo bảng Pivot trong Excel thống kê **Sales Amount** theo **Location** của **Customer**



	A	B	C	D	E	F
4		Mountain	Road		Road Total	Grand Total
5	Row Labels		Road-150	Road-650		
6	Australia	$6,799.98	$153,864.61	$3,495.49	$157,361.10	$164,161.08
7	Canada	$10,149.97	$150,287.34	$699.10	$150,986.44	$161,136.41
8	France	$6,749.98	$21,469.62	$2,097.29	$23,566.91	$30,316.89
9	Germany	$3,374.99	$42,939.24	$1,398.20	$4,337.44	$47,712.43
10	United Kingdom	$13,574.96	$32,204.43	$1,398.20	$33,602.63	$47,177.59
11	United States	$20,274.94	$75,143.67	$4,893.69	$80,037.36	$100,312.30
12	Grand Total	$60,924.83	$475,909.91	$13,981.96	$489,891.87	$550,816.69
13						
14						
15						
16						
17						

<u>Bài 2:</u> Tạo mô hình dữ liệu đa chiều liên quan đến mảng bán hàng qua các đại diện bán hàng, bao gồm 7 dimension tables và 1 fact table : FactReseller, DimPromotion, DimSalesTerritory, DimGeography, DimDate, DimProduct, DimEmployee, DimResellerSales.

Duyệt khối dữ liệu đã deploy, chuyển kết quả qua Excel, tạo bảng Pivot trong Excel thống kê **Reseller Sales Amount** theo **Employees** hierarchy như sau:



TUẦN 3

Bài 3: Phân tích và trình bày số liệu theo các yêu cầu sau:

a. **Các mặt hàng của công ty Adventure Works Cycles là gì?** Hãy chứng tỏ doanh thu năm 2007 tăng 54% so với năm 2006; số lượng đơn đặt hàng xe đạp tăng 86% trong 3 năm 2006-2008, trong khi đó quần áo và phụ kiện chỉ tăng 3%. Gợi ý bảng số liệu chính để tiến hành pivot: *(Số liệu các bảng dưới đây có thể thay đổi tùy theo version AdventureWorksDW đang dùng)*

TotalSales	Column Labels			
Row Labels	2006	2007	2008	Grand Total
Accessories	36,815	124,433	1,077,065	1,238,313
Bikes	22,090,618	28,179,554	44,350,354	94,620,526
Clothing	66,328	750,716	1,283,474	2,100,517
Components	1,166,765	4,629,101	6,003,210	11,799,077
Grand Total	**23,360,526**	**33,683,805**	**52,714,103**	**109,758,434**

b. **Thị trường của công ty Adventure Works Cycles là ở đâu?** Hãy chứng tỏ doanh thu bán sản phẩm bên ngoài nước Mỹ tăng 33% vào năm 2006 và tăng gần 51% vào năm 2008. Gợi ý bảng số liệu chính để tiến hành pivot:

TotalSales	Column Labels			
Row Labels	2006	2007	2008	Grand Total
United States	15,535,349	20,563,610	25,963,721	62,062,681
United Kingdom	550,507	5,083,062	6,257,260	11,890,829
Canada	3,591,852	2,763,865	5,147,140	11,502,857
Australia	2,568,701	2,099,585	5,805,290	10,473,577
France	414,245	2,021,672	4,714,497	7,150,415
Germany	513,353	593,247	3,574,747	4,681,348
NA	186,518	558,762	1,251,447	1,996,727
Grand Total	**23,360,526**	**33,683,805**	**52,714,103**	**109,758,434**

c. **Các mảng kinh doanh của công ty Adventure Works Cycles là gì?** Hãy dùng số liệu chứng minh số lượng đơn đặt hàng từ 18 đại lý bán lẻ chiếm đến 73% tổng số đơn hàng. Mảng bán hàng qua Internet sa sút trong năm 2007. Gợi ý bảng số liệu chính để tiến hành pivot:

TotalSales	Column Labels			
Row Labels	2006	2007	2008	Grand Total
Internet	7,072,084	5,762,134	16,473,618	29,307,837
Reseller	16,288,442	27,921,671	36,240,485	80,450,597
Grand Total	**23,360,526**	**33,683,805**	**52,714,103**	**109,758,434**

d. **Công ty Adventure Works Cycles bán hàng cho ai**? Mảng bán lẻ gồm 635 khách hàng (tính từ năm 2006-2008), nhưng chỉ có 439 khách tiếp tục đặt hàng trong năm 2008. Mảng bán hàng online tăng mạnh từ 17412 khách lên 17918 khách. Trị giá trung bình đơn hàng bán lẻ là $21.000, so với bán online là $1.098. Lợi nhuận bán hàng qua mạng lại cao hơn bán lẻ, do chi phí giảm hơn so với bán lẻ.

Hãy dùng số liệu chứng minh cho các nhận định trên, gợi ý bảng số liệu chính để tiến hành pivot:

Row Labels	TotalSales	TotalUniqueCustomerCount	TotalOrders	DollarsPerOrder
Internet	**29,307,837**	**17,918**	**26,683**	**1,098**
2006	7,072,084	2,206	2,206	3,206
2007	5,762,134	3,222	3,222	1,788
2008	16,473,618	17,412	21,255	775
Reseller	**80,450,597**	**635**	**3,796**	**21,194**
2006	16,288,442	208	739	22,041
2007	27,921,671	365	1,255	22,248
2008	36,240,485	493	1,802	20,111
Grand Total	**109,758,434**	**18,553**	**30,479**	**3,601**

Bài 4: Dùng kho dữ liệu AdventureWorksDW2012, tạo data source view gồm 1 bảng Fact chưa sử dụng từ trước tới nay và các bảng Dim phù hợp.

1. Tìm hiểu các mối quan hệ và các thuộc tính của bảng Fact này.
2. Tạo các cube bằng cách kết hợp bảng Fact và các bảng có liên quan.
3. Duyệt cube (hoặc đẩy số liệu ra Excel và tiến hành Pivot) để phân tích số liệu theo 3 cách tùy chọn, sao cho thông tin nhận được có ý nghĩa nhất.
4. Sao chép nội dung các bảng phân tích số liệu ở bước 3 và dán vào file Word để nộp lại cho giáo viên. Bên dưới mỗi bảng, hãy diễn giải nội dung của bảng thống kê.
5. Lặp lại các bước từ 1-4 đối với các bảng Fact còn lại.

Khai phá dữ liệu

TUẦN 4

Bài toán phân lớp (Classifying)

Kịch bản kinh doanh:

Phòng tiếp thị của công ty Adventure Works Cycles muốn xác định những đặc trưng của khách hàng đã mua hàng để tiên liệu những khách hàng này có khả năng mua một sản phẩm nào đó trong tương lai hay không. Cơ sở dữ liệu AdventureWorks 2012 lưu trữ thông tin nhân khẩu học mô tả các khách hàng trong quá khứ. Bằng cách sử dụng **giải thuật Cây quyết định** để phân tích thông tin, bộ phận tiếp thị có thể xây dựng mô hình dự đoán xem liệu có một khách hàng cụ thể sẽ mua các sản phẩm nào, dựa trên giá trị của các cột dữ liệu đã biết về khách hàng đó, như thông tin về nhân khẩu học hay các mẫu mua hàng trong quá khứ.

Trong bài toán dùng giải thuật Cây quyết định và giải thuật Gom nhóm, lần lượt thực hiện các bước sau:

- Create the mining model structure.
- Create the mining models.
- Explore the mining models.
- Test the accuracy of the mining models.
- Create predictions from the mining models.

Bài 1: Xét cụ thể tình huống kinh doanh sau:

Công ty Adventure Works Cycles đang giới thiệu một mẫu chiếc xe đạp leo núi mới (Mountain bike). Công ty đang tìm kiếm một giải pháp để tiếp thị sản phẩm của mình và tiếp cận những khách hàng có nhiều khả năng mua mẫu xe đạp này. Giải pháp được đưa ra: xây dựng cấu trúc thư tín cho khách hàng mục tiêu. Vì vậy, công ty lên kế hoạch để tìm hồ sơ của những khách hàng đã mua xe đạp leo núi trong quá khứ, và có được địa chỉ email của khách hàng để gởi email tiếp thị sản phẩm mới.

Hãy **tạo giải pháp khai phá dữ liệu và xây dựng mô hình hỗ trợ chiến dịch thư tín đến khách hàng mục tiêu, nhằm phân tích hành vi mua sắm của khách hàng và người mua hàng tiềm năng.** Giải thuật sử dụng: **cây quyết định**, khai thác bảng **ProspectiveBuyer** – chứa số liệu các khách hàng mua xe đạp tiềm năng và view **vTargetMail** - chứa dữ liệu trong quá khứ của khách hàng mua xe đạp.

✍ Tạo **Data Source View** trong **Analysis Services Project**

a. Creating an **Analysis Services Project**

b. Creating a **Data Source**: Adventure Works DW 2012

c. Creating a **Data Source View** gồm bảng **ProspectiveBuyer** và view **vTargetMail**

- Xây dựng cấu trúc thư tín cho khách hàng mục tiêu nhiều khả năng mua xe đạp từ công ty Adventure Works Cycles

d. Creating a Targeted Mailing Mining Model Structure. Select data mining technique is **Microsoft Decision Trees**. Select case table is **vTargetMail**. **ProspectiveBuyer** table use for testing. Select Predictable column is **Bike Buyer**.

Select columns: **Age, CommuteDistance, EnglishEducation, EnglishOccupation, Gender, GeographyKey, HouseOwnerFlag, MaritalStatus, NumberCarsOwned, NumberChildrenAtHome, Region, TotalChildren, YearlyIncome**

Columns	Content Type	Data Type
Address Line1	Discrete	Text
Address Line2	Discrete	Text
Age	Continuous	Long
Bike Buyer	Discrete	Long
Commute Distance	Discrete	Text
Customer Key	Key	Long
Date First Purchase	Continuous	Date
Email Address	Discrete	Text
English Education	Discrete	Text
English Occupation	Discrete	Text
FirstName	Discrete	Text
Gender	Discrete	Text
Geography Key	Discrete	Text
House Owner Flag	Discrete	Text
Last Name	Discrete	Text
Marital Status	Discrete	Text
Number Cars Owned	Discrete	Long
Number Children At Home	Discrete	Long
Region	Discrete	Text
Total Children	Discrete	Long
Yearly Income	Continuous	Double

e. Specifying the **Data Type** and **Content Type** (Discrete, Continuous, Key)

f. Specifying a **Testing Data Set** (=30%) for the Structure. Allow drill through

- Bổ sung và xử lý mô hình (Adding and Processing Models)
  - ✓ Adding New Models to the Targeted Mailing Structure
  - ✓ Processing Models in the Targeted Mailing Structure

  g. Set the **Holdout Seed**=12

  h. Deploy the project and process all the mining models

- Khám phá mô hình thư tín cho khách hàng mục tiêu

  i. Exploring the **Decision Tree Model** in the **Decision Tree** tab.

  j. Explore the model in the **Dependency Network** tab. Click the **Bike Buyer** node to identify its dependencies. Adjust the **All Links** slider to identify the most influential attribute

- Trả lời các câu hỏi sau:
  - Total number of cases
  - Number of non bike buyer cases
  - Number of bike buyer cases
  - Number of cases with missing values for [Bike Buyer]

  Drill through to case data

Bài 2: Tạo mô hình khai phá dữ liệu, dùng bảng ProspectiveBuyer, DimCustomer, DimAge, DimDate, DimGeography, DimProduct, DimProductCategory, DimProductSubcategory, FactInternetSales.

- Chọn các thuộc tính cần thiết đưa vào mô hình, dùng giải thuật **Cây quyết định**.
- Giải thích kết quả nhận được.

TUẦN 5

Bài toán gom nhóm (Clustering)

Kịch bản kinh doanh:

Các mô hình gom nhóm xác định các mối quan hệ trong một tập dữ liệu mà chúng ta có thể không có được một cách hợp lý thông qua quan sát ngẫu nhiên. Ví dụ: mọi người có thể dễ dàng đoán rằng những người đi làm bằng xe đạp thường không sống xa nơi họ làm việc. Tuy nhiên, **giải thuật gom nhóm** có thể tìm thấy những đặc điểm không hiển nhiên khác về những người đi xe đạp, chẳng hạn tồn tại nhóm người có xu hướng đi xe đạp đến nơi làm việc như một cách rèn luyện thân thể.

Xem xét một nhóm người có thông tin nhân khẩu học tương tự nhau và những người mua các sản phẩm tương tự nhau từ công ty Adventure Works, nhóm người này đại diện cho một nhóm dữ liệu. Bằng cách quan sát các cột dữ liệu của một nhóm, ta có thể thấy rõ hơn cách các bản ghi trong một tập dữ liệu (data set) liên quan đến nhau.

Bài 1:

Hãy tạo mô hình khai phá dữ liệu, dùng bảng **ProspectiveBuyer**s và view **vTargetMail**, áp dụng **giải thuật Gom nhóm**

a. Trước tiên, làm như hướng dẫn các mục a, b, c ở **Bài 1** tuần trước để có mô hình khai phá dữ liệu, rồi làm tiếp phần dưới đây

b. Exploring the **Clustering Model** in the **Cluster Diagram** tab.

In the **Viewer** list, select **Microsoft Cluster Viewer**. In the **Shading Variable** box, select **Bike Buyer**. Select 1 in the **State** box to explore those cases where a bike was purchased. Select the cluster that has the highest density, right-click the cluster, select **Rename Cluster** and type **Bike Buyers High** for later identification. Find the cluster that has the lightest shading (and the lowest density). Right-click the cluster, select **Rename Cluster** and type **Bike Buyers Low**.

c. Exploring the model in the **Cluster Profiles** tab. Set **Histogram** bars to 5.

d. Exploring the model in the **Cluster Discrimination** tab. In the **Cluster 1** box, select **Bike Buyers High**. In the **Cluster 2** box, select **Bike Buyers Low**. Click **Variables** to sort alphabetically.

Bài 2: Làm lại **Bài 2** tuần trước, nhưng dùng giải thuật **Gom nhóm**.

Bài 3: So sánh kết quả nhận được giữa giải thuật **Cây quyết định** và giải thuật **Gom nhóm**. Giải thuật nào là nổi trội và trong trường hợp nào?

TUẦN 6

Bài toán luật kết hợp (Association Rules)

Kịch bản kinh doanh:

Công ty Adventure Works Cycle đang thiết kế lại các chức năng của trang web. Mục tiêu của việc thiết kế lại là tăng doanh số bán sản phẩm. Nhờ có lưu lại thông tin mỗi lần bán hàng trong cơ sở dữ liệu giao dịch, công ty có thể sử dụng **giải thuật luật kết hợp** để xác định các bộ sản phẩm có xu hướng được mua cùng với nhau. Sau đó, công ty có thể dự đoán các mặt hàng bổ sung mà khách hàng có thể quan tâm, dựa trên các mặt hàng đã có trong giỏ hàng của khách hàng.

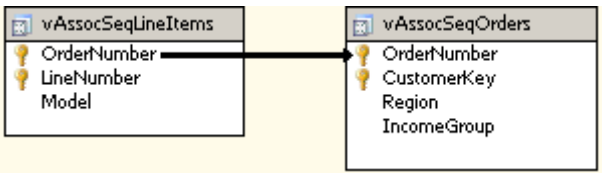
Bài 1:

Phân tích nhóm các sản phẩm khách hàng mua online trên trang thương mại điện tử của công Adventure Works Cycles, dựa trên phân tích giỏ hàng để khuyến cáo sản phẩm cho khách hàng.

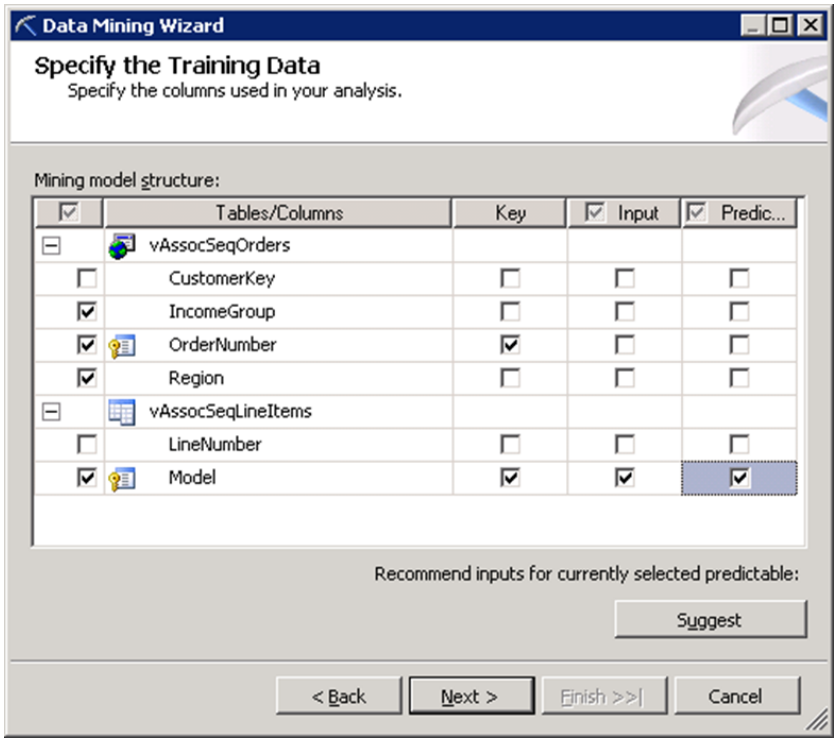
Hãy dùng 2 view **vAssocSeqOrders**, **vAssocSeqLineItems**, tạo **mô hình khai phá dữ liệu để khám phá các luật kết hợp**.

- Creating a Solution and Data Source
- Building a Market Basket Scenario

a. Adding a **Data Source View** with Nested Tables. Create a relationship between tables: the **vAssocSeqOrders** is case table (phía 1), **vAssocSeqLineItems** is nested table (phía nhiều)



b. Creating a Market Basket Structure and Model: create an association mining structure. **Specify the Training Data** page:



Specify Columns' Content and Data Type page:

Columns	**Content Type**	**Data Type**
IncomeGroup	Discrete	Text
Order Number	Key	Text
Region	Discrete	Text
vAssocSeqLineItems		
Model	Key	Text

c. **Create testing set** page, the default value for the option **Percentage of data for testing** is 30 percent. Change this to 0. Select the option **Allow drill through**

d. Modifying and Processing the Market Basket Model: Adjust the parameters of the Association model:

MINIMUM_PROBABILITY = 0.1

MINIMUM_SUPPORT = 0.01

Process the mining model

e. Exploring the Market Basket Models: Open the Association mode in the **Microsoft Association Rules Viewer**. Navigate the dependency graph and locate specific nodes: show only the strongest association. Filter the itemsets that are shown in the viewer by name.

f. View details for an itemset: Locate the item "**Touring Tire**", select **Drill Through**

g. Filter itemsets by support or size: **Minimum support**=100

h. See only rules that include the **Mountain-200** bicycle

i. View details about the rule by using the content viewer: review the value for NODE_TYPE and NODE_DESCRIPTION.

j. Predicting Associations: use **Prediction Query Builder**

Model
Women's Mountain Shorts
Water Bottle
Touring-3000

k. Create a singleton prediction query with nested table inputs

Model
Touring Tire Tube
Sport-100
Water Bottle

l. Create a prediction query using nested table inputs

Bài 2:

Khi mua vỏ và ruột xe (Tires and Tubes), khách hàng có xu hướng mua kèm mặt hàng nào?

- Hãy tạo mô hình khai phá dữ liệu để giải bài toán, dùng các bảng phù hợp.
- Nêu các thông số minh chứng cho khảo sát của bạn.

BÀI TẬP TỔNG HỢP

Khai phá và Phân tích dữ liệu theo yêu cầu nghiệp vụ cho trước

TUẦN 7

Bài 1:

Mary Gibson là phân tích viên kênh bán hàng qua Internet, chuyên trách kênh bán hàng qua mạng và các chương trình khuyến mãi liên quan. Bạn hãy giúp Mary hoàn thành các công việc sau:

1. **Customer Profiling and Target Marketing**

Mary muốn có được số liệu chi tiết hơn, trong việc quảng bá cho khách hàng trên Internet. Cô biết rằng các chương trình khuyến mãi đơn giản nhằm vào những khách hàng phù hợp với hồ sơ nhân khẩu học (demographic) và hồ sơ mua hàng (buying profiles) có thể tạo sự khác biệt lớn trong doanh thu trung bình từ mỗi khách hàng. Mary và một vài nhân viên tiếp thị khác đã nói chuyện với một số nhà cung cấp về các sản phẩm quảng cáo và quản lý chiến dịch tiếp thị. Nhận định đưa ra là Adventure Works Cycles có thể làm tốt hơn gấp nhiều lần trong lĩnh vực này.

Mary xem xét khách hàng theo khu vực, gom nhóm theo bang (state) hay tỉnh (province) nơi họ sinh sống. Mary nhận thấy sự khác biệt đáng kể về các kiểu xe (model) và sở thích màu sắc, dựa trên khu vực sinh sống của khách hàng. John Wood, người quản lý sản phẩm xe đạp leo núi, cũng gom nhóm số liệu mặt hàng xe đạp leo núi (mountain bike) theo các khu vực khác nhau.

Thử làm theo cách của Mary và John Wood. Trình bày cách làm và số liệu bạn có được. Hãy đọc và phân tích rõ về số liệu này.

2. **Customer Loyalty Program**

Mary cũng muốn đánh giá lòng trung thành của khách hàng sử dụng website của Công ty Adventure Works Cycle. Cô nghĩ rằng một số chương trình khuyến mãi hấp dẫn sẽ giúp khách hàng quay trở lại với công ty Adventure Works Cycles, khi họ muốn nâng cấp lên một kiểu (model) xe đạp mới, hoặc muốn có thêm chiếc xe đạp thứ hai (hoặc thứ ba…), hoặc muốn mua thêm chiếc xe đạp cho một thành viên khác trong gia đình. Hãy thử tạo mô hình khai phá dữ liệu theo hướng Mary nêu ra và cho biết Mary có lý hay không?

TUẦN 8

Bài 2:

Brian Welker là trưởng nhóm phụ trách mảng bán lẻ, nhận báo cáo từ 17 nhân viên và 3 giám đốc bán hàng khu vực. Bạn hãy giúp Brian Welker biết các thông tin sau:

1. **Growth analysis**: Tổng quan về thị trường năm 2008: sản phẩm mới, vùng bán hàng mới, người bán hàng mới.

2. **Customer analysis**: Chọn 1 năm tùy ý, hãy xác định ai là người mua hàng hàng đầu của năm (khách hàng có tổng tiền mua hàng lớn nhất)? Mức tiêu dùng của họ đã thay đổi như thế nào ở các năm sau?

3. **Territory analysis**: Khách hàng hàng đầu (top customers) hiện ở đâu? và vùng (territories) bán hàng có tiềm năng nhất ở đâu? (chọn 1 năm xác định)

TUẦN 9

4. **Sales performance**: Brian cũng muốn xem xét đơn đặt hàng, từ quan điểm của một đại diện bán hàng. Nếu Brian phát hiện một vấn đề trong dữ liệu ở mức cao hơn, anh ta muốn xem chi tiết (drill down) các đơn hàng của các đại diện bán hàng riêng lẻ.

Bạn hãy chỉ ra 20 khách hàng hàng đầu và thông tin đơn đặt hàng của họ, so sánh giữa số liệu mảng bán lẻ và mảng bán hàng qua mạng.

5. **Basic sales reporting:** Brian biết rằng 17 phần trăm khách hàng năm 2006 đã không đặt hàng lại vào năm 2007. Và đến năm 2008, Brian vẫn chưa nghe thông tin gì thêm từ con số 17% này.

Hãy kiểm tra số liệu phân tích mà Brian đang nắm giữ và giúp Brian có thêm thông tin cập nhật về nhóm khách hàng này?

6. **Bài tập mở**

Khai thác kho dữ liệu AdventureWorksDW2012, tìm các luật kết hợp/ hoặc xây dựng cây quyết định/ hay giải bài toán gom nhóm, để có các phát hiện có ý nghĩa (khác với phần bài tập ở trên) .

SV tự nêu bài toán, phân tích dữ liệu, khai phá dữ liệu, trình bày kết quả thu nhận được và rút ra các kết luận.

TUẦN 10

Kiểm tra Thực hành