Building a Data Mining Model using Data Warehouse and OLAP Cubes

SS Chung
A data warehouse is a centralized repository that stores data from multiple information sources and transforms them into a common, multidimensional data model for efficient querying and analysis. OLAP and Data Mining are two complementary technologies for Business Intelligence. Online Analytical Processing (OLAP) is a technology that is used to organize large business databases and support business intelligence. OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions. OLAP databases are divided into one or more cubes, and each cube is organized and designed by a cube administrator to fit the way that you retrieve and analyze data. OLAP is used for decision-support systems to analyze aggregated information for sales, finance, budget, and many other types of applications. OLAP is about aggregating measures based on dimension hierarchies and storing these pre-calculated aggregations in a special data structure. With the help of pre-aggregations and special indexes, you can query aggregated data and get decision-support query results back in real time.

OLAP provides us with a very good view of what is happening, but can not predict what will happen in the future or why it is happening. This part is done by data mining. Data Mining is a combination of discovering techniques + prediction techniques.

The sequence of steps that will be followed in this project is

1. Understand the Adventure works database which will be used in this project (fully understand the transactional data available in the database).
2. Getting familiar with SQL Server Analysis Services (SSAS) tools and various datamining algorithms such as Clustering, Association, Time Series etc in the SSAS tool.
3. The next step will be to come up with a list of questions: what questions need to be answered, what metrics will help business managers monitor and grow the business.
4. Based on the business questions that need to be answered, data staging layer in a star schema format will be designed.
5. Then the cube will be built to extract data from the Star schema staging layer and we perform our data mining on the cube.
6. Perform Datamining in the Adventure works database to find hidden patterns and information using DMX and MDX.

1. **Adventure works database:**

   In this section we will try to understand the Adventure works database that will be used as a part of the project. We try to understand scope of the business its various components and products etc.

1.1. **Business Overview:**

Adventure Works Cycles is a large multinational bicycle manufacturer, with headquarters located in Bothell, Washington. The company has approximately 300 employees, 29 of which are sales representatives. The primary distribution channel for Adventure Works Cycles through the retail stores of their resellers. These resellers are located in Australia, Canada, France, Germany, the United Kingdom, and the United States. Adventure Works Cycles also sells to individual customers worldwide by means of the Internet.

Adventure Works Cycles has five major product offerings:

- **Bikes** – Three primary bike product lines: Mountain, Road, and Touring.
- **Accessories** – Examples include helmets and water bottles.
- **Clothing** – Examples include jerseys and biking shorts.
- **Components** – Examples include bottom brackets and frames.
- **Services** – Examples include premium service and standard service.

### Lines of Analysis

Following are anticipated lines of analysis for sales and marketing at Adventure Works Cycles.

<table>
<thead>
<tr>
<th>Line of analysis</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product profitability</td>
<td>Analyzing product sales margin by monitoring costs, discounts, and selling</td>
</tr>
<tr>
<td>Variance analysis</td>
<td>Comparing forecast sales to actual sales</td>
</tr>
<tr>
<td>Sales staff performance</td>
<td>Monitoring variance between sales quotas and actual sales</td>
</tr>
<tr>
<td>Trend and growth analysis</td>
<td>Understanding how the current period compares to prior periods</td>
</tr>
<tr>
<td>Industry market share</td>
<td>Analyzing market share data for the resellers of Adventure Works Cycles</td>
</tr>
<tr>
<td>Promotion effectiveness</td>
<td>Identifying how promotions influence sales performance</td>
</tr>
</tbody>
</table>

The version of Adventure works database that will be used in this project is Adventure works 2012/2014
2. Building Data Mining Project with Data Warehouse and Cube

In the following section we will understand what a star schema is and the various terms in star schema such as fact table, dimension table, measures, groups, etc and then design a star schema based on a list of questions. This understanding will be the first step in our Data mining activity we will be performing as a part of this project.

5.1. Star schema, Fact Tables and Dimension Tables

The star schema architecture is the simplest data warehouse schema. It is called a star schema because the diagram resembles a star, with points radiating from a center. The center of the star consists of fact table and the points of the star are the dimension tables.

A fact table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts. A fact table can contain fact's data on detail or aggregated level.

A dimension is a structure usually composed of one or more hierarchies that categorizes data. If a dimension hasn't got a hierarchies and levels it is called flat dimension or list. The primary keys of each of the dimension tables are part of the composite primary key of the fact table. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Dimension tables are generally small in size then fact table.

Typical fact tables store data about sales while dimension tables data about geographic region (markets, cities), clients, products, times, channels.

5.2. Steps in Star Schema Design:

1. Identify a business process for analysis (like sales)
2. Identify measure or facts (sales dollar)
3. Identify dimensions for facts (product dimension, location dimension, etc)
4. List the columns that describe each dimension (region name, branch dimension, etc)
5. Determine the lowest level of summary in a fact table (sales dollar)

3. Cube AND MDX:

3.1. Cube:
OLAP Cube is the basic unit of storage for Multidimensional data, on which we can do analysis on stored data and study the various patterns.
3.2. **Dimensions:**
The primary functions of dimensions are to provide Filtering, Grouping and Labeling on your data. Dimension tables contain textual descriptions about the subjects of the business. Dimensions in general we can say are the Master entities with related member attributes using which we can study data stored in OLAP Cube Quickly and effectively.

3.3. **Measure & Measure Groups**
Metrics value stored in your Fact Tables is called Measure. Measures are used to analyze performance of the Business. Measure usually contains numeric data, which can be aggregated against usage of associated dimensions. Measure Group holds collection of related Measure

3.4. **Steps in Building and Deploying a cube:**
To build an SSAS cube, you must first start a project by following these steps:

All Programs -> Microsoft SQL Server -> SQL Server BIDS

Create an Analysis Services project.

Name your project FirstCube and click OK.

You’re now presented with an empty window, which seems like a rare beginning to a project with a template; really, you have nothing to start with, so it’s time to start creating. The first component you’ll need is somewhere to retrieve data from: a data source.

**Building a Data Source**

To create the data source you’ll use for your first cube, follow these steps:
Navigate to the Solution Explorer pane on the right, right-click Data Sources, and click New Data Source. This will bring up the Data Source Wizard, which will walk you through the creation process.

The next component you’ll create is the data source view.

Meanwhile, go ahead and click Next to continue creating your data source. In this next screen, it’s time to set up a connection string.

If your AdventureWorksDW database is visible as a selection already, go ahead and choose it; if not, click New.

For your server name, enter (local), and then drop down the box labeled Select or Enter a Database Name and chooseAdventureWorksDW.

Click OK to return to the wizard and then click Next.

You can now enter the user you want SSAS to impersonate when it connects to this data source. Select Use the Service Account and click Next. Using the service account (the account that runs the SQL Server Analysis Server service) is fairly common even in production, but make sure that service account has privileges to read your data source.

For your data source name, type AdventureWorksDW and then click Finish.

**Building a Data Source View**

Follow the below steps:

Right-click Data Source Views and choose New Data Source View. Predictably, up comes the Data Source View Wizard to walk you through the process. Click Next.

Make sure the AdventureWorksDW data source is selected and then click Next.

On the Select Tables and Views screen, choose FactInternetSales under Available objects and then click the right arrow to move it into the Included Objects column on the right.

To add its related dimensions, click the Add Related Tables button as shown in Figure 18-3 and then click Next. Note that one of the related tables is a fact, not a dimension. There’s no distinction made at this level. Later, you will be able to select and edit dimensions individually.
On the last screen, name your data source view according to its contents: Internet Sales.

Click Finish to create the Internet Sales data source view.
Creating an Analysis Services **Cube**. Right-click Cubes in the Solution Explorer and select New Cube to bring up the Cube Wizard. This will walk you through choosing measure groups, the measures within them, and your dimensions for this cube, click Next. On the Select Creation Method screen, make sure Use Existing Tables is selected, and click Next.

The wizard will now want you to tell the wizard where to find measure groups. You could help it out by telling the wizard that those are in your fact tables and then click Next.
Now the wizard would like to know which measures from your measure groups (fact tables) you’d like to store in the cube. By default, it’s got them all selected; go ahead and accept this by clicking Next.

At this point, you have measures, but you still need some dimensions; the wizard will select the dimension tables from your data source view to create as new dimensions. Again, by default they’re all selected, and you can click Next.

The wizard is now ready to complete. Verify everything is done as per above steps. If everything appears to be in order, click Finish.
Deploying the **Cube**

Deploying process can be started by following these steps.

Select Deploy First **Cube** on the **Build** menu. You’ll see a series of status messages as the **cube** is built, deployed, and processed for the first time. You’ll receive a few warnings when you deploy First**Cube**, and if they’re warnings and not errors, you can safely ignore them for now.

When it’s done and you see Deployment Completed Successfully in the lower right, your first **cube** is ready to browse.
4. Querying On Cubes Using MDX-Examples

4.1. Queries:

What is sales amount in all the countries??

```mdx
select non empty([Dim Sales Territory].[Sales Territory Country].children) on rows,
[Measures].[Sales Amount] on columns
from [star_sale_trend_analysis]
```
What is total sales amount in all countries for the year 2007??

select non empty([Dim Sales Territory].[Sales Territory Country].children) on rows,[Measures].[Sales Amount] on columns from [star_sale_trend_analysis] where [Order Date].[Calendar Year].&[2007]
Retrieve all the products in descending order of their Internet sales amount of year 2007

select nonempty([Measures].[Sales Amount]) on columns,
order([Dim Product].[English Product Name].members ,[Measures].[Sales Amount],desc) on rows
from [star_sale_trend_analysis]
where ([Fact Internet Sales - Order Date].[Calendar Year].&[2007])
What is the product wise sales in United States??

select non empty([Dim Product].[English Product Name].children) on rows, [Measures].[Sales Amount] on columns from [star_sale_trend_analysis] where [Dim Sales Territory].[Sales Territory Country].&[United States]
5. References

1. Data-Mining-With-Sql-Server-2008
2. SQL Server 2012 Tutorials - Analysis Services Data Mining
3. Data Mining: Concepts and Techniques by Jiawei Han (Author), Micheline Kamber (Author)
4. Internet resources:
   • http://www.codeproject.com/Articles/658912/Create-First-OLAP-Cube-in-SQL-Server-Analysis-Serv
   • http://www.codeproject.com/Articles/710387/Learn-to-Write-Custom-MDX-Query-First-Time
   • http://marktab.net/datamining/2010/08/21/mining-olap-cubes

5. Software Used:
   • Microsoft SQL Server,
   • Microsoft Visual Studio 2012 or any higher
   • Microsoft SQL Server 2012 Business Intelligence