Building a Data Mining Model using Data Warehouse and OLAP Cubes

IST 734

SS Chung
1. Abstract:

The purpose of this project is learn the application of various database and datamining concpets and its application in current business using Adventure works database and various database tools. We begin by designing a star schema and building a DataWarehouse OLAP cube, for Sales Analysis using Sql Server Analysis Services (SSAS) and query the cube using MultiDimensional eXpressions language(MDX) to find the current business trends. Next we create data mining structure using Data Mining Extensions(DMX) and find the hidden pattern and predict things that are helpful in improving the business and its growth. Some of the data mining techniques that will be considering to accomplish the above tasks are Microsoft Association Mining, Cluster analysis, Time series and Naïve- Bayes Algorithm.

2. Introduction:

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 precalculated aggregations in a special data structure. With the help of preaggregations 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 datamining. 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.

3. **Adventure works database:**

In this section we will try to understand the Adventureworks 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.

3.1. **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.
The version of Adventure works database that will be used in this project is Adventure works 2012.

4. Getting familiar with SQL Server Analysis Services (SSAS) tools and various data mining algorithms

In this section we will learn and understand some data mining algorithms and its applications that will be used as a part of project later.
4.1. **Microsoft Association Algorithm**

The Microsoft Association algorithm is an association algorithm provided by Analysis Services that is useful for recommendation engines. A recommendation engine recommends products to customers based on items they have already bought, or in which they have indicated an interest. The Microsoft Association algorithm is also useful for market basket analysis.

Association models are built on datasets that contain identifiers both for individual cases and for the items that the cases contain. A group of items in a case is called an *itemset*. An association model consists of a series of itemsets and the rules that describe how those items are grouped together within the cases. The rules that the algorithm identifies can be used to predict a customer’s likely future purchases, based on the items that already exist in the customer’s shopping cart.
null
4.2. Microsoft Clustering Algorithm

The Microsoft Clustering algorithm is a segmentation algorithm provided by Analysis Services. The algorithm uses iterative techniques to group cases in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions. Clustering models identify relationships in a dataset that you might not logically derive through casual observation. For example, you can logically discern that people who commute to their jobs by bicycle do not typically live a long distance from where they work. The algorithm, however, can find other characteristics about bicycle commuters that are not as obvious. In the following diagram, cluster A represents data about people who tend to drive to work, while cluster B represents data about people who tend to ride bicycles to work.

- Clustering Model View
Sunnie S Chung IST 734
4.3. **Microsoft Time Series Algorithm**

The Microsoft Time Series algorithm provides regression algorithms that are optimized for the forecasting of continuous values, such as product sales, over time. Whereas other Microsoft algorithms, such as decision trees, require additional columns of new information as input to predict a trend, a time series model does not. A time series model can predict trends based only on the original dataset that is used to create the model. You can also add new data to the model when you make a prediction and automatically incorporate the new data in the trend analysis.
<table>
<thead>
<tr>
<th>Model Region</th>
<th>predict amount</th>
<th>predict quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>N200 Europe</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amount</td>
<td>STIME Quantity</td>
</tr>
<tr>
<td>10/25/2008</td>
<td>337472.7672</td>
<td>10/25/2008</td>
</tr>
<tr>
<td>N200 North Am.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N200 Pacific</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Query execution completed with 12 rows fetched.
4.4. **Microsoft Decision Trees Algorithm**

For discrete attributes, the algorithm makes predictions based on the relationships between input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that you designate as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column. For example, in a scenario to predict which customers are likely to purchase a bicycle, if nine out of ten younger customers buy a bicycle, but only two out of ten older customers do so, the algorithm infers that age is a good predictor of bicycle purchase. The decision tree makes predictions based on this tendency toward a particular outcome.

- **Decision Tree**
For continuous attributes, the algorithm uses linear regression to determine where a decision tree splits.
5. 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)

6. Cube AND MDX:

6.1. 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.
6.2. **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.

6.3. **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.

6.4. **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 choose AdventureWorksDW.

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.
7. Querying On Cubes Using MDX-Examples

7.1. Queries:

Find the product which has been ordered more than 500 in all countries?

select filter(crossjoin([Dim Sales Territory].[Sales Territory Country].children,[Dim Product].[English Product Name].children), [Measures].[Order Quantity]>500) on rows,[Measures].[Order Quantity] on columns from [star_sale_trend_analysis];

Retrieve the products whose sales amount > 5000

select  {[Measures].[Sales Amount],[Measures].[Fact Internet Sales Count]} on columns, filter([Dim Product].[English Product Name].children , [Measures].[Sales Amount]>5000) on rows from [star_sale_trend_analysis];

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]

What is sales amount in all the countries??

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]

7.2. Screenshots:
What is sales amount in all the countries??
What is the product wise sales in United States??

What is total sales amount in all countries for the year 2007??
Find the products in descending order of their Internet sales amount of year 2007
Find the products whose sales amount is greater than 5000?

Select NON EMPTY {[Measures].[Sales Amount]} on columns,
{[Dim Product].[English Product Name].members} on rows From [nnn_cube]
In the following section we perform the main part of our project i.e the data mining activity as a whole where we did the bits and pieces so far.

The result of a star schema is nothing more than a table similar to a database table. Based on the business need we created a table and now we are going to perform our task of finding the hidden activities in the table in the following section.
In the earlier step we set the stage such as creating the table, understanding basically what a data mining is, how to various algorithms work, how to query on the data using various querying languages like MDX, DMX etc and now we start the play.

8. Data Mining and DMX:

**Initial set up:**

Find the tables which contains the data of the activity we are going to perform.

Find the appropriate data mining algorithm which would help us to solve our problem

8.1. Business Scenario 1:
Adventure works is introducing a new Mountain bike. It is looking for a way to market its product and reach customers who are more likely to buy the bike. So it is planning to find the profile of the people who got Mountain Bikes in the past, and get the email address of the customers who has same profile and send email.

**Steps to be followed:**

**Step 1:** Cluster the database using Microsoft Clustering Algorithm

```sql
CREATE MINING STRUCTURE CIS698_Mountain_Bike_Marketing
(
    [Marketingbike Key]   LONG KEY,
    [GeographyKey]  LONG CONTINUOUS,
    [Commute Distance] TEXT DISCRETE,
    [House Owner Flag]   LONG DISCRETE,
    [Marital Status] TEXT DISCRETE,
    [Number Cars Owned]  LONG CONTINUOUS,
    [Number Children At Home] LONG CONTINUOUS,
    [Model] TEXT DISCRETE,
    [Total Children] LONG CONTINUOUS,
    [Yearly Income] LONG DISCRETIZED
)

INSERT INTO CIS698_Mountain_Bike_Marketing
(
    [Marketingbike Key],    [GeographyKey],    [CommuteDistance],    [HouseOwnerFlag],
    [MaritalStatus],    [NumberCarsOwned], [NumberChildrenAtHome], [Model], [TotalChildren],
    [YearlyIncome]
)

OPENQUERY([Adventure Works DW2012],'SELECT [MarketingbikeKey],[GeographyKey], [CommuteDistance], [HouseOwnerFlag],
    [MaritalStatus],[NumberCarsOwned], [NumberChildrenAtHome], [Model], [TotalChildren],
    [YearlyIncome] FROM [dbo].[cis698_marketing_bike]')

ALTER MINING STRUCTURE CIS698_Mountain_Bike_Marketing
ADD MINING MODEL CIS698_Mountain_Bike_MarketingCL
USING Microsoft_Clustering WITH DRILLTHROUGH
```
**Step 2:** Find the cluster which has maximum probability of Mountain Bikes

```sql
SELECT NODE_NAME, NODE_CAPTION, NODE_SUPPORT, NODE_DESCRIPTION
FROM [CIS698_Mountain_Bike_MarketingCL].CONTENT

SELECT FLATTENED PredictHistogram(Cluster())
From [CIS698_Mountain_Bike_MarketingCL]
NATURAL PREDICTION JOIN
(SELECT 'Mountain-500' AS [Model]) AS t
```
Step 3: Use that profile to the company Customer database and get the email address of the customers who are more likely to buy the bike and send them offers on the new product.

Step 2 Result:

Number Cars Owned = 0, Yearly Income < 40419, Commute Distance = 0-1 Miles,

55 <= Geography Key <= 337, Number Children At Home = 0, 0 <= Total Children <= 3, House Owner Flag = 1, Marital Status = M, Model = Road-350-W, Model = Road-750, Model = Touring-3000, Model = Road Bottle Cage, Model = Touring-2000, Model = Touring Tire, Model = Cycling Cap, Model = Road-550-W, Model = Touring Tire Tube, Model = Mountain-400-W, Model = Long-Sleeve Logo Jersey, Model = Mountain-500, Model = Sport-100, Model = ML Road Tire, Model = LL Road Tire, Model = Road-250, Model = Road Tire Tube, Model = Water Bottle

SELECT distinct [LastName], [FirstName], [MiddleName], [EmailAddress] FROM [dbo].[cis698_marketing_bike] where [NumberChildrenAtHome] = 0 and [TotalChildren] BETWEEN 0 and 3 and [NumberCarsOwned] = 0 and [YearlyIncome] BETWEEN 10000 and 40419 and [GeographyKey] BETWEEN 55 and 337 and [CommuteDistance] = '1-2 Miles' or [CommuteDistance] = '0-1 Miles' and [HouseOwnerFlag] = 1 and [MaritalStatus] = 'M'
8.2. Business Scenario 2
Adventure works tries to improve its market on Mountain Bikes further by offering offers or deals on the items which are more likely to be purchased with Mountain Bikes.

CREATE MINING MODEL SubcategoryAssociations
(
   [Customer ID] LONG KEY,
   [Subcategories] TABLE PREDICT
   (
      [Subcategory] TEXT KEY
   )
) USING Microsoft_Association_Rules

Train the association rules model:

INSERT INTO SubcategoryAssociations([Customer ID], [Subcategories])
SHAPE
{OPENQUERY([Adventure Works DW], 'SELECT [OrderNumber] FROM [dbo].[vAssocSeqOrders] ORDER BY [OrderNumber]')}
APPEND
{OPENQUERY([Adventure Works DW], 'SELECT [OrderNumber], [Subcategory] FROM (SELECT DISTINCT vAssocSeqLineItems.OrderNumber, DimProductSubcategory.EnglishProductSubcategoryName AS Subcategory FROM DimProduct INNER JOIN DimProductSubcategory
ON DimProduct.ProductSubcategoryKey = DimProductSubcategory.ProductSubcategoryKey
INNER JOIN vAssocSeqLineItems
ON DimProduct.ModelName = vAssocSeqLineItems.Model)
AS [CustomerSubcategories]
ORDER BY
Determine the top two subcategories a customer is likely to purchase that has purchased a road bike and a jersey:

```sql
SELECT
    Predict([Subcategories], 2) as [Subcategories]
FROM
    [SubcategoryAssociations]
NATURAL PREDICTION JOIN
(SELECT
    (SELECT 'Mountain Bikes' AS Subcategory
     UNION SELECT 'Jersies' AS Subcategory
    ) AS Subcategories
) AS t
```
8.3. Business Scenario3
Adventure works is trying to forecast their sales based on the current trend and based on its results whether it meets its expectations it is planning to change its business strategy.

CREATE MINING MODEL [cis698Forecasting]

( [Reporting Date] DATE KEY TIME,
  [Model Region] TEXT KEY,
  [Quantity] LONG CONTINUOUS PREDICT,
  [Amount] DOUBLE CONTINUOUS PREDICT )

USING Microsoft_Time_Series (AUTO_DETECT_PERIODICITY = 0.8, FORECAST_METHOD = 'MIXED') WITH DRILLTHROUGH

ALTER MINING STRUCTURE [cis698Forecasting_Structure]
ADD MINING MODEL [cis698Forecasting_ARTXP]

( ([Reporting Date],
  [Model Region],
  [Quantity] PREDICT,
  [Amount] PREDICT )

USING Microsoft_Time_Series (AUTO_DETECT_PERIODICITY = .08, FORECAST_METHOD = 'ARTXP') WITH DRILLTHROUGH

INSERT INTO MINING STRUCTURE [cis698Forecasting_Structure]

( [Reporting Date],[Model Region],[Quantity],[Amount] )

OPENQUERY(

[Adventure Works DW2012],
'SELECT [ReportingDate],[ModelRegion],[Quantity],[Amount] FROM vTimeSeries ORDER BY [ReportingDate]' )
9. Conclusion:
In this project we learn the application of various database and data mining concepts and its application. We designed a star schema and then built a Data Warehouse OLAP cube, for Sales Analysis using SQL Server Analysis Services (SSAS) and queries the cube using MultiDimensional eXpressions language (MDX) and found various answers for some business questions. Next we created data mining structure using Data Mining Extensions (DMX) and found the hidden pattern and predicted things that were helpful in improving the business and its growth. We also learned various techniques that were available to create data mining structure. That was demonstrated through both the usage of DMX Queries and a Wizard to create Data mining structure in SSAS tool.

10. 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

Other Useful System Guides:
Remote Desktop

- Go to
  - Start → All Programs → Accessories → Communications → Remote Desktop Connection
  - Give the Remote server address like “winserv1.csuohio.edu”
  - Enter your Remote Desktop Login ID and Password

- Connecting from Home
  - Connect to VPN

In Remote Server

- You can find Microsoft SQL Server 2012/2014 in Programs.

- SQL Server services:
  - Analysis Services
  - sql server management studio.
    - SqlAuthentication
    - UserId and Password Given
    - sql server business Intelligence development studio
Import and Export Data

➢ To import or export data click.
  ➢ Start → Programs → Microsoft SQL Server2005 → right click on database name → tasks → import and export data.

  OR

  ➢ Under SSIS menu in integration services click import and export data and follow the steps
  ➢ Generally our Data Sources will be text files with data separated by commas or tabs etc.
  ➢ Destination database to be used is cis612_2014 for this class.
  ➢ use only SQL Authentication where ever it is asked.

To Create a Data Cube

➢ Start → Programs → Microsoft SQL Server → sql server BIDS → new Analysis service project

➢ Create a new Data Source
➢ Create a Data source view
➢ Create and Deploy the Cube.
Data Cube

- Open new Cube Wizard
- Select Data Source, Measures
- Select or create new Dimension
- Process the cube
- Define Aggregations