Business Intelligence

DCAP606

Edited by:
Sartaj Singh

LOVELY
PROFESSIONAL
UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

BUSINESS INTELLIGENCE

Edited By
Sartaj Singh

Printed by
EXCEL BOOKS PRIVATE LIMITED
A-45, Naraina, Phase-I,
New Delhi-110028
for
Lovely Professional University
Phagwara

CONTENT

SYLLABUS

Business Intelligence

Objectives: To impart the skills needed to manage database of large scale organization, techniques for data mining. Student will learn OLAP and generating quick reports.

S. No.	Description
1.	***Business Intelligence:*** Introduction, Meaning, Purpose and Structure of Business Intelligence Systems. Understanding Multidimensional Analysis Concepts: Attributes, Hierarchies and Dimensions in data Analysis. Understanding Dimensional Data Warehouse: Fact Table, Dimension Tables, Surrogate Keys and alternative Table Structure. What is multi-dimension OLAP?
2.	***Understanding OLAP:*** Fast response, Meta-data based queries, Spread sheet formulas. Understanding Analysis Services speed and meta-data. Microsoft's Business intelligence Platform. Analysis Services Tools. Data Extraction, Transformation and Load. Meaning and Tools for the same.
3.	***Creating your First Business Intelligence Project:*** Creating Data source, Creating Data view. Modifying the Data view. Creating Dimensions, Time, and Modifying dimensions. Parent-Child Dimension.
4.	***Creating Cube:*** Wizard to Create Cube. Preview of Cube. Adding measure and measure groups to a cube. Calculated members. Deploying and Browsing a Cube.
5.	***Advanced Measures and Calculations:*** Aggregate Functions. Using MDX to retrieve values from cube. Calculation Scripting. Creation of KPI's.
6.	***Advanced Dimensional Design:*** Creating reference, fact and many to many dimensions. Using Financial Analysis Cubes. Interacting with a cube. Creating Standard and Drill Down Actions.
7.	***Retrieving Data from Analysis Services:*** Creating Perspectives, MDX Queries, Excel with Analysis Services.
8.	***Data Mining:*** Meaning and purpose. Creating data for data mining. Data mining model creation. Selecting data mining algorithm. Understanding data mining tools. Mapping Mining Structure to Source Data columns. Using Cube Sources. Configuring Algorithm parameters.
9.	***Creating Data mining queries and reports:*** Creation of Prediction queries. Understanding DMX language.
10.	***Reporting Tools:*** Using SQL Server Reporting Services to develop reports for analysis services.

Unit 1: Introduction to Business Intelligence

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss the meaning of Business Intelligence
- Explore history of Business Intelligence
- State the purpose of Business Intelligence Systems
- Construct structure of Intelligence Systems

Introduction

Business Intelligence (BI) is a set of ideas, methodologies, processes, architectures, and technologies that change raw data into significant and useful data for business purpose. Business Intelligence can handle large amounts of data to help identify and evolve new opportunities for the business. Making use of these new opportunities and applying a productive scheme on it can provide a comparable market benefit and long-term stability.

Notes

Business Intelligence (BI) technologies provide chronicled, present and predictive view of business operations. Common functions of enterprise Intelligence technologies are reporting, online analytical processing, analytics, data excavation, process excavation, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

1.1 Meaning of Business Intelligence

BI (Business Intelligence) refers to set of techniques which assist in spotting, digging out and investigating best data from the large amount of data to improve conclusion making. Let us understand the concept better with help of an example.

Example: Suppose we have chronicled data of a Shopping Mart of 3-6 months. Here, in the data we have different products with their respective specifications. Let us select one of the products-say Candles. We have three kinds of Candles in this class say Candle A, Candle B, Candle C. On studying of these data we come to know that sale of Candle C was at peak out of these three classes. Now on afresh and deep study into these data we got the outcome that the sale of this Candle C was maximum between the time intervals of 9 am to 11 am. On further deeper analysis, we came to the conclusion that this specific Candle is the one used in place of worship.

Now, let's apply Business Intelligence for this analysis. What an enterprise firm or the organization can do is, get other material that can be used in church and place them nearby those candles. Now the customers approaching the Shopping Mart to purchase the candles for place of worship can also have a look on the other material and may be tempted to purchase them as well. Now this will surely enhance the sales and hence the income of Shopping Mart.

Self Assessment

Fill in the blanks:

1. .... can handle large amounts of data to help identify and evolve new opportunities for the business.
2. BI (Business Intelligence) refers to set of techniques which assist in ...., digging out and .... best data from the large amount of data to improve conclusion making.

1.2 History of Business Intelligence

Normally, early business applications had their own databases that supported their functions. These databases became "islands of information" in that no other systems had access to them. These islands of information proliferated as more and more departments were automated.

Did u know? Amalgamations and acquisitions aggregated the difficulty since the companies integrated completely distinct systems, numerous of which were doing the similar job.

However, businesses shortly identified the analytical value of the data that they had access to. In fact, as enterprises automated more systems, more data became accessible. However, collecting these data for analysis was a challenge because of the incompatibilities amidst systems.

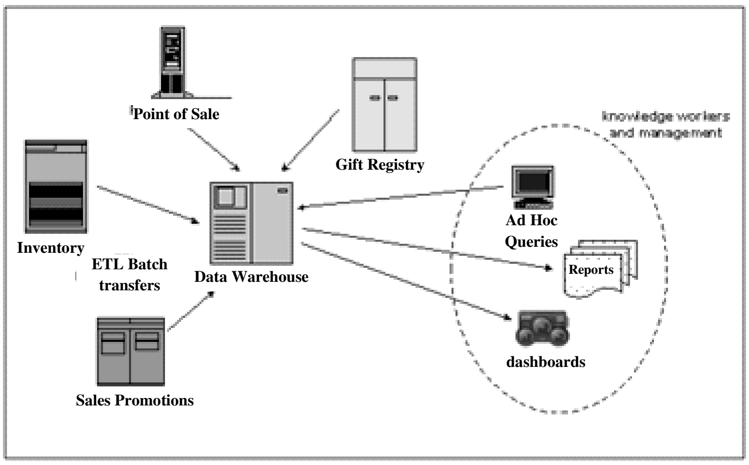
Caution There was no simple way (and often no way) for these systems to interact.

An infrastructure was required for data exchange, collection, and analysis that could supply a unified view of an enterprise's data. The data warehouse evolved to complete this need.

1.2.1 The Data Warehouse

The concept of the data warehouse (Figure 1.1) is a lone scheme that is the repository of all of the organization's data (or simply data) in a pattern that can be competently analysed so that significant accounts can be arranged for administration and other information workers.

Figure 1.1: Data Warehouse Concept



Source: http://www.gravic.com/shadowbase/images/uses/datawarehouse.png

However, meeting this goal requires some challenges:

- Data should be acquired from a variety of incompatible systems.
- The identical piece of data might reside in the databases of distinct systems in distinct types. A specific data item might not only be represented in distinct formats, but the values of this Data piece might be distinct in distinct databases. Which value is the correct one?
- Data is continually altering. How often should the Data warehouse be revised to contemplate a sensibly current view?
- The amount of Data is massive. How is it analysed and presented easily so that it is useful?

To meet these needs, a broad range of powerful tools were developed over the years and became productized. They included:

- Extract, Transform, and Load (ETL) utilities for the moving of data from the diverse data sources to the common data warehouse.
- Data-mining pushes for complex predetermined analysis and ad hoc queries of the Data retained in the Data warehouse.
- Reporting tools to provide management employees with the outcomes of the analysis in very simple to absorb formats.

Notes

1.2.2 Offline Extract, Transform and Load (ETL)

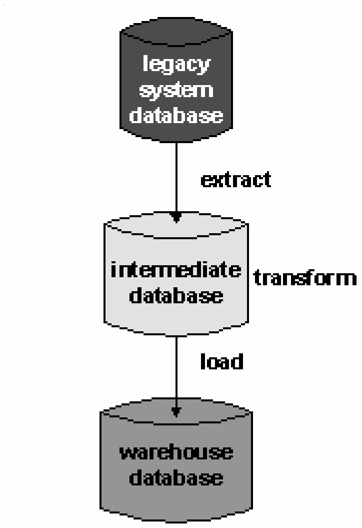
Early on, the one common interface that was provided between the disparate systems in an association was magnetic tape. Tape formats were standardized, and any system could compose tapes that could be read by other systems. Thus, the first data warehouses were fed by magnetic tapes prepared by the various systems inside the association. However, that left the difficulty of data disparity. The data written by the different systems reflected their native data associations.

Notes The data written to tape by one system often had little relation to the similar data written by another system.

Even more important was that the data warehouse's database was designed to support the analytical functions needed for the business intelligence function. This database design was typically a highly organised database with complex indices to support Online Analytical Processing (OLAP). Databases configured for OLAP allowed complex analytical and ad hoc queries with rapid execution time. The data fed to the data warehouse from the enterprise systems was converted to a format significant to the data warehouse.

To explain the difficulty of initially stacking this data into a data warehouse, holding it updated, and resolving discrepancies, Extract, Transform and Load (ETL) utilities were evolved. As their name suggests, these utilities extract data from source databases, change/transform them into the widespread data warehouse format, and load them into the data warehouse, as shown in Figure 1.2.

Figure 1.2: Extract/Transform/Load Structure



ETL – Extract/Transofrm/Load

Source: http://www.gravic.com/shadowbase/images/uses/etl.png

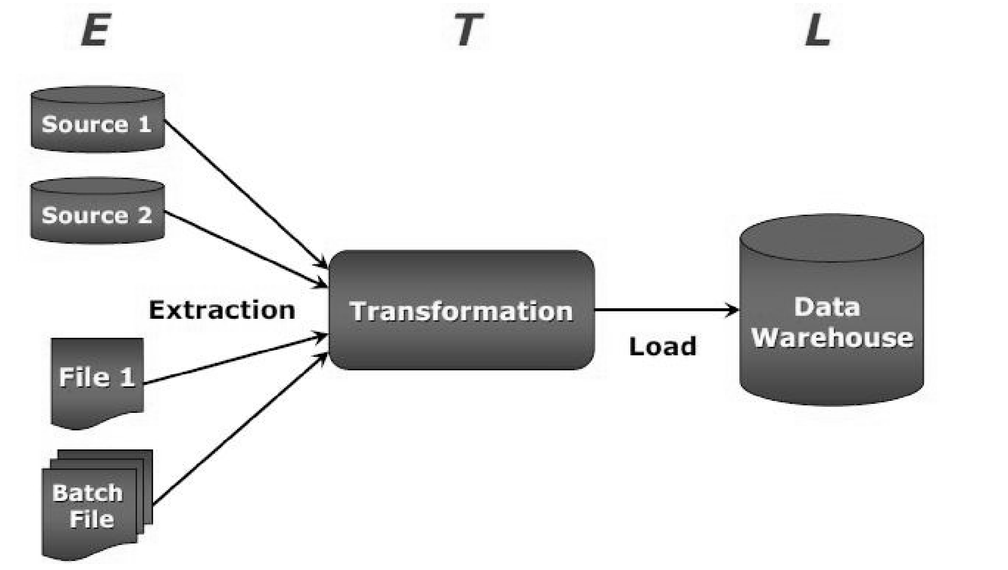
The transform function is the key to the achievement of this approach. Its job is to request a series of rules to extracted data so that it is properly formatted for loading into the data warehouse. An example of transformation rules includes:

- The selection of data to load.
- The translation of encoded items (for example, 1 for male, 2 for female to M, F).

- Deriving new calculated values (sale price = price - discount).
- Merging data from multiple sources.
- Summarizing (aggregating) certain rows and columns.
- Splitting a column into multiple columns.
- Resolving discrepancies between similar data items.
- Validating the data.

Figure 1.3 Shows Representation of ETL Model

Figure 1.3: Representation of ETL Model



Source: http://3.bp.blogspot.com/_tutW43y628U/TL2I-JTIFAI/AAAAAAAAAEI/mir1v2EMiTg/s1600/ETL_Global.jpg

The ETL function permits the consolidation of multiple data sources into a well-structured database for use in complex analysis. The ETL process is performed occasionally, such as daily, weekly, or monthly, depending upon the enterprise needs. This method is called offline ETL because the key database is not relentlessly updated. It is revised on a periodic batch basis. Though offline ETL serves its purpose well, it has some drawbacks as well:

- The data in the data warehouse is not fresh. It could be weeks old. Though, it is useful for strategic functions but is not especially adaptable to tactical use.
- The source database typically should be temporary inactive throughout the extract method. Otherwise, the target database is in an inconsistent state following the load. With this result, the applications must be shutdown, often for hours.

In order to develop to support real-time business intelligence, the ETL function must be relentless and non-invasive, which is called online ETL, and is recounted later. In compare to offline ETL, which supplies data which is not fresh but reliable answers to queries, online ETL supplies present but varying answers to successive queries since the data that it is using is constantly being updated to reflect the current state of the business.

Notes

1.2.3 Data-Mining Engines

The ETL utilities make data collection from numerous diverse systems practical. Then, the data needs to be converted into useful information. Some key points to remember:

- Data are easily facts, figures, and text that can be processed by a computer.

Example: A transaction at retail point-of-sale is data.

- Information is processed data. For example, analysis of point-of-sale transactions yields information of consumer buying behaviour.
- Knowledge represents a pattern that connects information and usually presents a high grade of predictability as to what is recounted or what will happen next.

Example: An example of knowledge is the prediction of promotional efforts on sales of particular items based on buyers' buying behaviour.

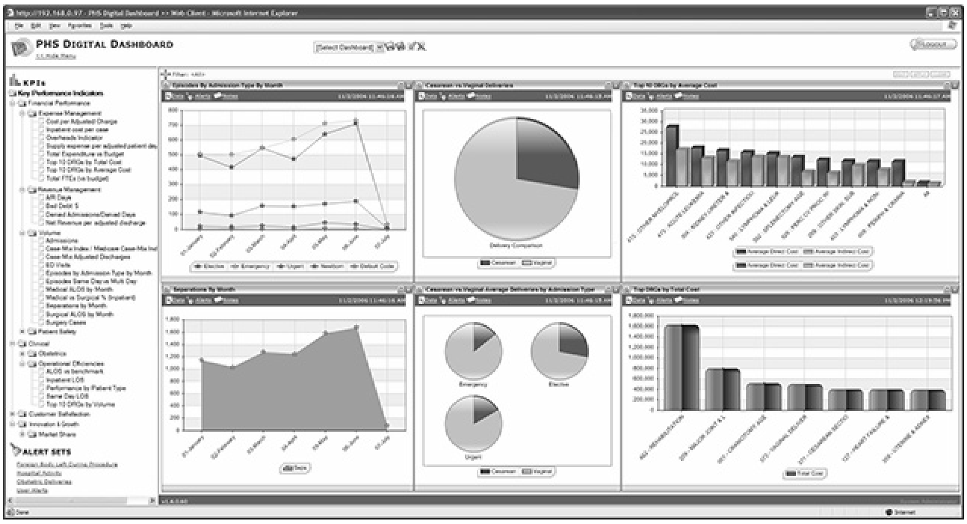
Useful data-mining engines were evolved to support complex analysis and ad hoc queries on a data warehouse's database. Data mining looks for patterns among hundreds of seemingly unrelated fields in a large database, patterns that recognize earlier unknown trends. These trends play a key role in strategic decision making because they disclose localities for process enhancement.

Example: Data-mining engines are those from SPSS and Oracle which are the foundation for OLAP (Online Analytical Processing) systems.

1.2.4 Reporting Tools

The knowledge created by a data-mining engine is not very useful unless it is presented easily and clearly to those who need it. There are many formats for reporting information and knowledge results. One of the common techniques for displaying information is the digital dashboard (shown in Figure 1.4).

Figure 1.4: Digital dashboard



Source: http://www.powerhealthsolutions.com/images/PBR_DigitalDashboard_KPIs.png

It provides a business manager with the input necessary to push the business towards success. It presents the client a graphical view of business processes. The client then drills down the data at will to get more details on a specific process. Today, many versions of digital dashboards are accessible from a kind of software vendors.

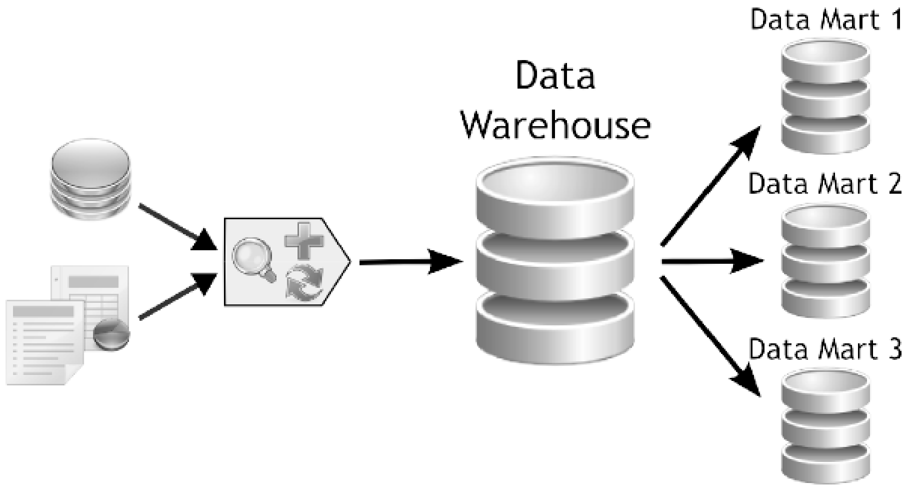
1.2.5 Data Marts

As corporate-wide data warehouses came into use, it was discovered that in many situations a full-blown data warehouse was overkill for applications. Data marts evolved to solve this problem. A data mart is a special type of a data warehouse. It is focused on a single subject (or functional area), such as Sales, Finance, or Marketing. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single department. The primary use for a data mart is Business Intelligence (BI) applications. Implementing a data mart can be less expensive than implementing a data warehouse, thus making it more practical for the small business.

Notes A data mart can also be set up in much less time than a data warehouse.

Figure 1.5 shows the relationship between data warehouse and data mart.

Figure 1.5: Relationship between Data Warehouse and Data Mart



Source: http://www.dataprix.net/files/uploads/250image/HEFESTO%20v2_0/data%20mart%20-%20top%20down.png

Self Assessment

Fill in the blanks:

3. An .... was required for data exchange, collection, and analysis that could supply a unified view of an enterprise's data.

4. .... utilities for the moving of data from the diverse data sources to the common data warehouse.

5. The first data warehouses were fed by .... prepared by the various systems inside the association.

Notes

6. The data fed to the data warehouse from the .... was converted to a format significant to the data warehouse.
7. The job of .... is to request a series of rules to extracted data so that it is properly formatted for loading into the data warehouse.
8. The .... function permits the consolidation of multiple data sources into a well-structured database for use in complex analysis.
9. In compare to offline ETL, ... supplies present but varying answers to successive queries.
10. .... represents a pattern that connects information and usually presents a high grade of predictability as to what is recounted or what will happen next.
11. A .... is a special type of a data warehouse focused on a single subject (or functional area), such as Sales, Finance, or Marketing.
12. The primary use for a data mart is .... applications.

1.3 Purpose of Business Intelligence Systems

'The purpose of business intelligence systems is to utilise all your underlying business data to help executives make better business decisions,' said Tony Banham, technology and solutions director for Oracle Greater China.

Business Intelligence process consists of 3 distinct tasks:

1. The first task BI has to do is to gather the necessary data about the business. The key to this is automating the process. Gathering data was very time and money consuming in the past, but todays with the usage of modern computers, it's much easier to collect data from various sources.
2. The second task is to analyse the collected data and then further extract information from it. The extracted information is then transformed into knowledge.
3. The final task is to use the newly gathered knowledge to improve the business.

There are many business intelligence tools to complete the process of gathering knowledge related to business like Cognos, BizTools, Hummingbird and Informatica.

1.4 Structure of Intelligence Systems

A business intelligence system has three key advantages:

1. It not only supports the latest information technologies, but also provides pre-packaged application solutions.
2. It focuses on the access and delivery of business information to end users, and support both information providers and information consumers.
3. It support access to all forms of business information, and not just the information stored in a data warehouse.

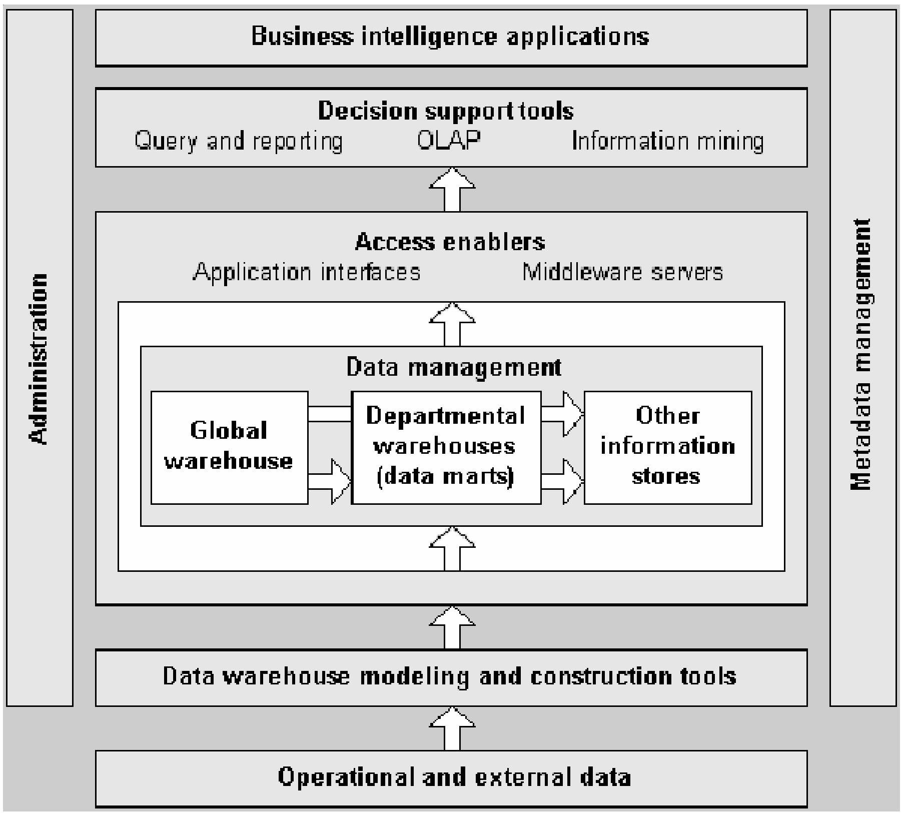
Figure 1.6 gives an overview of Structure of Business Intelligent System.

1.4.1 Business Intelligence Applications

Business intelligence applications provide integrated business applications, hardware, software, and consulting services. Decision Edge from IBM, for example, analyses customer information and based on that assists in the creation of tailored customer marketing programs.

Notes

Figure 1.6: IBM Business Intelligence Product Set



Source: www.redbooks.ibm.com/redbooks/pdfs/sg245415.pdf

1.4.2 Decision Support Tools

Query and Reporting

The two main query and reporting products (in IBM) are the Query Management Facility (QMF) and Lotus approach. QMF has been used for many years as a host-based query and reporting tool by DB2 whereas Lotus approach is a desktop relational DBMS that has gained popularity due to its easy-to-use query and reporting capabilities.

Online Analytical Processing (OLAP)

If we talk about IBM structure, its key product in the OLAP marketplace is the DB2 OLAP Server, which implements three-tier client/server architecture for performing complex data analysis. The value of the DB2 OLAP server lies in its ability to generate and manage relational tables that contain multidimensional data.

Information Mining

Intelligent Miner by IBM is one of the few products in the market to support an external API, allowing resultant data to be collected by other products (for example an OLAP product) for further analysis.

Notes

1.4.3 Access Enablers

Client access to warehouse and operational data from business intelligence tools requires a client database API.

1.4.4 Data Management

Data management offers intelligent data partitioning and parallel query and utility processing of the data.

Example: DB2 for OS/390, DB2 for VM, and DB2 for VSE DB2 Universal Database.

1.4.5 Data Warehouse Modelling

Using Visual Warehouse a data warehouse can be designed and constructed. Tools for developing data warehouse includes components for defining the relationships between the source data and warehouse information, transforming source data and managing warehouse maintenance.

Task Find out the procedure to extract data from each individual database.

Self Assessment

State whether the following statements are true or false:

13. The first task BI has to do is to gather the necessary data about the business.

14. Business intelligence system do not supports the latest information technologies.

15. The value of the DB2 OLAP server lies in its ability to generate and manage relational tables that contain multidimensional data.

Case Study

Business Intelligence Management

You are working for a sporting goods retail company. Overall sales have been declining for the last three quarters and management is very much concerned. Each retail store has its own individual databases that keep track of sales for specific items. However at the overall management level, only sales figures for each store are reported. Management has asked you to prepare reports by sale items (such as specific brand X and model Y of running shoes) and by categories of sale items (such as all running shoes) so that they can make accurate decisions on which product lines to drop to reduce overall inventory costs. As a BI specialist or as a part of a development team, you will need to develop a procedure to extract data from each individual database, reconcile data formats and types, aggregate all data into a data repository, and develop queries based on management requests. You will communicate and work closely with management representatives to ensure that the created data repository and reports meet their needs. This usually involves a series of back and forth discussion during which both sides ask and answer questions. You have to develop a good understanding of the business concerns

Contd....

and be able to frame them into technical requirements for the BI project. You also need to educate management regarding possibilities and constraints of the technology as it translates into business applications. Depending on your level of responsibilities, you will develop or contribute to the development of a realistic estimate of time, resources and cost to achieve the intended goal. Once the project is launched, you will develop full specifications, develop and test subsystems, and integrate into final product. Throughout the development process, you will conduct a series of increasingly rigorous technical, functional and user tests. Based on the results of these tests, you can expect revision of technical specifications, and overall system and unit design, along with project timeline, resources and cost revisions. Of particular importance will be a focus on user needs and requirements, business applications, and easy-to-use user interface. Your BI team may be also invoked in developing and delivering training to managers and other stakeholders in using the new system.

Questions:

1. What are basic technical requirements for the BI project?
2. Explain the procedure to extract data from each individual database.
3. What role does BI specialist play in management company?
4. What all factors will lead to reduction of the inventory cost?

Source: http://www.sdn.sap.com/irj/uac/go/portal/prtroot/docs/library/uuid/d0e7c318-6764-2c10-95b9-d2cf096d8e9d?overridelayout=true&44929653037402

1.5 Summary

- Business Intelligence can handle large amounts of data to help identify and evolve new opportunities for the business.
- BI (Business Intelligence) refers to set of techniques which assist in spotting, digging out and investigating best data from the large amount of data to improve conclusion making.
- An infrastructure was required for data exchange, collection, and analysis that could supply a unified view of an enterprise's data.
- Early on, the one common interface that was provided between the disparate systems in an association was magnetic tape.
- Databases configured for OLAP allowed complex analytical and ad hoc queries with rapid execution time.
- The ETL function permits the consolidation of multiple data sources into a well-structured database for use in complex analysis.
- In order to develop to support real-time business intelligence, the ETL function must be relentless and non-invasive, which is called online ETL, and is recounted later.
- The ETL utilities make data collection from numerous diverse systems practical.
- Useful data-mining engines were evolved to support complex analysis and ad hoc queries on a data warehouse's database.
- There are many formats for reporting information and knowledge results. One of the common techniques for displaying information is the digital dashboard.
- Business intelligence applications provide integrated business applications, hardware, software, and consulting services.

1.6 Keywords

Business Intelligence (BI): Business Intelligence (BI) is a set of ideas, methodologies, processes, architectures, and technologies that change raw data into significant and useful data for business purpose.

Business Intelligence Applications: Business intelligence applications provide integrated business applications, hardware, software, and consulting services.

Data Marts: A data mart is the access layer of the data warehouse environment that is used to get data out to the users.

Data Warehouse: In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis.

Extract, Transform and Load (ETL): The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL.

Magnetic Tape: Magnetic tape is a medium for magnetic recording, made of a thin magnetisable coating on a long, narrow strip of plastic film.

Online Analytical Processing (OLAP): OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining.

Query Management Facility (QMF): Query Management Facility is a query tool invented by IBM, for interfacing with their DB2 system. The most recent version is Version 9.2.

1.7 Review Questions

1. What is Business Intelligence (BI)?
2. Define Business Intelligence. Give some examples.
3. Briefly discuss history of Business Intelligent.
4. Explain the concept of the data warehouse.
5. What is offline Extract, Transform, and Load?
6. What do you understand by data-mining engines?
7. "Implementing a data mart can be less expensive than implementing a data warehouse". Elucidate.
8. What is the purpose of business intelligence systems?
9. What are the key advantages of business intelligence system?
10. Write the brief description of Business Intelligence Applications.

Answers: Self Assessment

1. Business Intelligence
2. Spotting, investigating
3. Infrastructure
4. Extract, Transform, and Load (ETL)
5. Magnetic tapes
6. Enterprise systems
7. Transform function
8. ETL
9. Online ETL
10. Knowledge

Notes

11. Data mart

12. Business Intelligence (BI)

13. True

14. False

15. True

1.8 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

http://data-warehouses.net/

http://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm

http://www.hcltech.com/enterprise-transformation-services/data-warehousing-and-business-intelligence

http://www.techopedia.com/definition/24170/extract-transform-load-etl

Unit 2: Multidimensional Analysis

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize the Dimension Attributes
- Summarize the Dimension Hierarchy
- Identify the Type of Hierarchy

Introduction

In statistics and related fields, multidimensional analysis is a data analysis process that groups data into two or more categories: data dimensions and measurements. To show this, let us take the case of a football game. A data set which comprises of the number of wins for one cricket team every year for many years could be categorized into a single dimensional or longitudinal data set. Another data set which comprises of the number of wins many different cricket groups inside a year can be under a single dimensional or traverse sectional data set. A single data set that comprises of the number of wins for diverse cricket teams over numerous years could be comprised in a two-dimensional data set.

Multi-Dimensional analysis is an Informational analysis on data which takes into account numerous distinct connections, each of which comprises a dimension. For example, a retail analyst may want to understand the connections amidst sales by district, by quarter, by demographic circulation or by product. Multi-dimensional analysis will yield outcomes for these complex relationships.

2.1 Dimension Attributes

A dimension consists of members.

Example: The members of a product dimension are the individual products.

Members have attributes to identify them.

Example: Some possible attributes for a product dimension could be the product code, colour, and size.

If the dimension is defined as a hierarchy, the lower levels of the hierarchy must also have an attribute that identifies the parent of each member. Information about each dimension is stored in one or more dimension tables.

2.1.1 Key Attribute

Each dimension contains a key attribute. Each attribute is bound to have one or more columns in a dimension table. The key attribute is the attribute in a dimension that identifies the columns in the dimension main table that are used in foreign key relationships to the fact table.

Caution Typically, the key attribute represents the primary key column or columns in the dimension table.

An attribute can also be bound to one or more additional columns for a specific task.

Example: An attribute's Name property determines the name that appears to the user for each attribute member and this property can be bound to a calculated column in the data source view.

Table 2.1 shows dimension attribute properties.

Table 2.1: Dimension Attribute Properties

Property	Description
Attribute Hierarchy Display Folder	Identifies the folder in which to display the associated attribute hierarchy to end users.
Attribute Hierarchy Enabled	Determines whether an attribute hierarchy is generated by Analysis Services for the attribute. If the attribute hierarchy is not enabled, the attribute cannot be used in a user-defined hierarchy and the attribute hierarchy cannot be referenced in Multidimensional Expressions (MDX) statements.
Attribute Hierarchy Optimized State	Determines the level of optimization applied to the attribute hierarchy. By default, an attribute hierarchy is Fully Optimized, which means that Analysis Services builds indexes for the attribute hierarchy to improve query performance. The other option, Not Optimized, means that no indexes are built for the attribute hierarchy. Using Not Optimized is useful if the attribute hierarchy is used for purposes other than querying, because no additional indexes are built for the attribute. Other uses for an attribute hierarchy can be helping to order another attribute.
Attribute Hierarchy Ordered	Determines whether the associated attribute hierarchy is ordered. The default value is True. However, if an attribute hierarchy will not be used for querying, you can save processing time by changing the value of this property to False.
Attribute Hierarchy Visible	Determines whether the attribute hierarchy is visible to client applications. The default value is True. However, if an attribute hierarchy will not be used for querying, you can save processing time by changing the value of this property to False.

Contd....

Custom Rollup Column	Specifies the column that defines a custom rollup formula.
Custom Rollup Properties Column	Specifies the column that contains the properties of a custom rollup formula.
Default Member	Specifies a Multidimensional Expressions (MDX) expression that defines the default measure for the attribute.
Description	Contains the description of the attribute.
Discretization Bucket Count	Contains the number of buckets into which to discretize.
Discretization Method	Defines the method to use for discretization.
Estimated Count	Specifies the estimated number of members in the attribute. Until you run the Aggregation Design Wizard, the default value is zero. Either you can allow the wizard to count the number of records or you can enter an estimated value. Enter a value manually if you know the number of members and want to save the time that is required to query the database for the count. If you are working with a test subset of your production data, you can use the counts of your production data so that the aggregation design will be optimized for the production data instead of the test data.
Grouping Behaviour	A user defined value that provides a hint to client applications on how to group attributes.
ID	Contains the unique identifier (ID) of the dimension.
Instance Selection	Provides a hint to client applications about how a list of items should be displayed, based on the expected number of items in the list. The available options are as follows: • ***None*** No hint is provided to the client application. This is the default value. • ***Drop Down*** The number of items is small enough to display in a drop-down list. • ***List*** The number of items is too large for a drop-down list, but does not require filtering. • ***Filtered List*** The number of items is large enough to require users to filter the items to be displayed. • ***Mandatory Filter*** The number of items is so large that the display must always be filtered.
Is Aggregatable	Specifies whether the values of the attribute members can be aggregated. The default value is True, which means that the attribute hierarchy contains an (All) level. If the value for this property is False, the attribute hierarchy does not contain an (All) level.
Key Columns	Contains the column or columns that represent the key for the attribute, which is the column in the underlying relational table in the data source view to which the attribute is bound. The value of this column for each member is displayed to users unless a value is specified for the Name Column property.
Member Names Unique	Determines whether member names in the attribute hierarchy must be unique.
Members With Data	Used by parent attributes to determine whether to display data members for non-leaf members in the parent attribute. This property value is only used when the value of the Usage property is set to Parent. This means that a

Contd....

	parent-child hierarchy has been defined. The available options are as follows: • Non-leaf Data Hidden Non-leaf data is hidden. • Non-leaf Data Visible Non-leaf data is visible.
Members with Data Caption	Provides a template string that is used by parent attributes to create captions for system-generated data members in the parent attribute. This property value is only used when the value of the Usage property is set to Parent. This means that a parent-child hierarchy has been defined.
Name	Contains the user-friendly name of the attribute.
Name Column	Identifies the column that provides the name of the attribute that is displayed to users, instead of the value in the key column for the attribute. This column is used when the key column value for an attribute member is cryptic or not otherwise useful to the user, or when the key column is based on a composite key. The Name Column property is not used in parent-child hierarchies; instead, the Name Column property for child members is used as the member names in a parent-child hierarchy.
Naming Template	Defines how levels are named in a parent-child hierarchy constructed from the parent attribute. This property value is only used when the value of the Usage property is set to Parent. This means that a parent-child hierarchy has been defined.
Order By	Describes how to order the members that are contained in the attribute hierarchy. The default value is Name, which specifies that ordering of the attribute members is based on the value of the Name Column property, if any. Otherwise, members are ordered by the value of the key column. The available options are as follows: • Name Column Order by the value of the Name Column property. • Key Order by the value of the key column of the attribute member. • Attribute Key Order by the value of the member key of a specified attribute, which must have an attribute relationship to the attribute. • Attribute Name Order by the value of the member name of a specified attribute, which must have an attribute relationship to the attribute.
Order By Attribute	Identifies the attribute by which to order the members of the attribute hierarchy.
Root Member If	Determines how the root or topmost members of a parent-child hierarchy are identified. This property value is only used when the value of the Usage property is set to Parent. This means that a parent-child hierarchy has been defined. The default value is Parent Is Blank Self Or Missing, which means that only members that meet one or more of the conditions described for Parent Is Blank, Parent Is Self, or Parent Is Missing are treated as root members. The following values are also available: • Parent Is Blank Only members with a null, a zero, or an empty string in the key column or columns are treated as root members. • Parent Is Self Only members with themselves as parents are treated as root members. • Parent Is Missing Only members with parents that cannot be found are treated as root members.
Type	Contains the type of the attribute.
Unary Operator Column	Specifies the column that provides unary operators. It is a binding of Data Item type that defines the details of a column providing a unary operator.

Contd....

Notes

Usage	Describes how an attribute is used. The available options are as follows: • Regular The attribute is a regular attribute. This is the default value. • Key The attribute is a key attribute. • Parent The attribute is a parent attribute.
Value Column	Identifies the column that provides the value of the attribute. If the Name Column element of the attribute is specified, the same Data Item values are used as default values for the Value Column element. If the Name Column element of the attribute is not specified and the Key Columns collection of the attribute contains a single Key Column element representing a key column with a string data type, the same Data Item values are used as default values for the Value Column element.

Source: http://msdn.microsoft.com/en-us/library/ms174919.aspx

Self Assessment

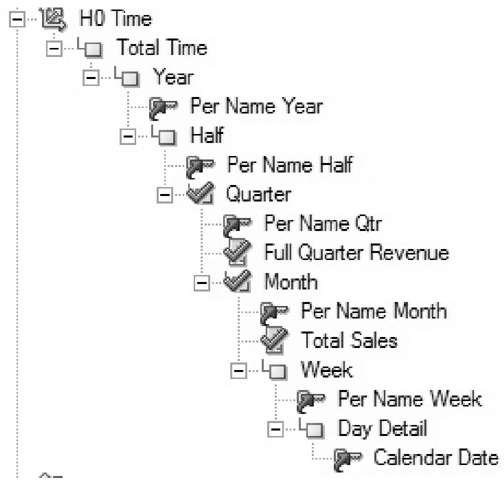
Fill in the blanks:

1. .... is an Informational analysis on data which takes into account numerous distinct connections, each of which comprises a dimension.
2. Information about each dimension is stored in one or more ....
3. The key attribute represents the .... in the dimension table.
4. An .... can also be bound to one or more additional columns for a specific task.

2.2 Dimension Hierarchy

A hierarchy is a set of parent-child relationships between attributes within a dimension. These hierarchy attributes are also known as levels. Example, the Time Dimension can have Total, Year, Quarter, Month and Date as its levels as shown in Figure 2.1:

Figure 2.1: Time Dimension Example



Source: http://oracle-bi.siebelunleashed.com/wp-content/uploads/2011/11/Time-Dim-Hierarchy.jpg

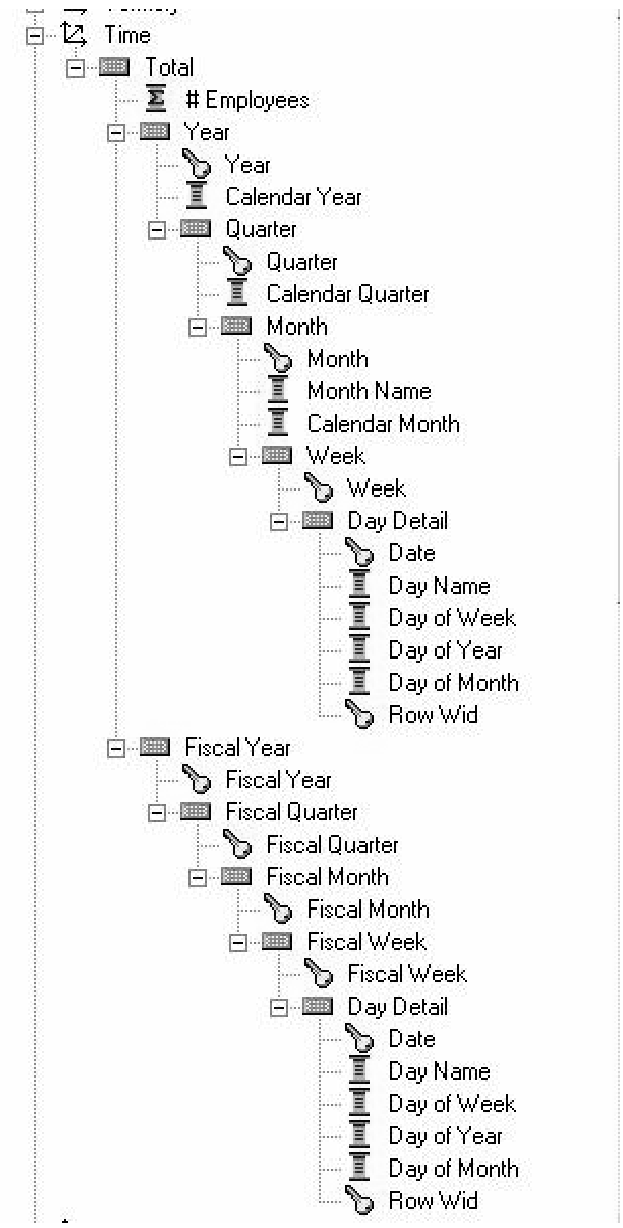
2.2.1 Type of Hierarchy

Level-based: This type of hierarchy consists of an ordered set of two or more levels.

Example: A time hierarchy might have three levels for Year, Quarter, and Month.

Level-based hierarchies can also contain parent-child relationships. This type of dimension hierarchy levels allow to perform aggregate navigation and configure level-based measure calculations.

Figure 2.2: Level based Hierarchy Example



- Time
  - Total
    - # Employees
    - Year
      - Year
      - Calendar Year
      - Quarter
        - Quarter
        - Calendar Quarter
        - Month
          - Month
          - Month Name
          - Calendar Month
          - Week
            - Week
            - Day Detail
              - Date
              - Day Name
              - Day of Week
              - Day of Year
              - Day of Month
              - Row Wid
    - Fiscal Year
      - Fiscal Year
      - Fiscal Quarter
        - Fiscal Quarter
        - Fiscal Month
          - Fiscal Month
          - Fiscal Week
            - Fiscal Week
            - Day Detail
              - Date
              - Day Name
              - Day of Week
              - Day of Year
              - Day of Month
              - Row Wid

Source: http://gerardnico.com/wiki/_media/dat/obiee/bi_server/design/dimension/obiee_multiple_dimension_hierarchie.jpg

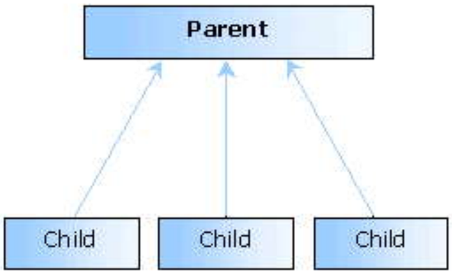
Notes

Also it supports special type of level-based dimension for unbalanced and Skip-level hierarchy. It also supports time dimension to provide special functionality for modelling time series data.

Parent-child: A parent-child hierarchy is a hierarchy in a standard dimension that contains a parent attribute. A parent attribute describes a self-referencing relationship, or self-join, within a dimension main table. It is actually value-based hierarchy. It consists of values that define the hierarchy in a parent-child relationship (Figure 2.3).

Example: An employee hierarchy might have no levels but might just contain names of employees who are managed by other employees. Employees may have titles like Vice President and then Vice Presidents might report to other Vice Presidents and they can be at different depths in the hierarchy.

Figure 2.3: Parent-child Hierarchy



Source: http://3d.recoil.org/nojavascript/Images/Parent-Child.gif

In addition to above discussed two level of hierarchy it can be of following type as well:

Ragged Hierarchy: A hierarchy in which all the lowest-level members do not have the same level of depth is ragged hierarchy.

Example: A time hierarchy might be having current month data at the day level, the previous month's data at the month level, and the previous 10 year's data at the quarter level.

It is also known as an unbalanced hierarchy.

Skip-level: A hierarchy in which certain members do not have values for certain higher levels are known as skip-level hierarchy.

Example: In India, Delhi city does not belong to another state (it belongs to Delhi as a state itself).

What matters is that users can still navigate from the country level (India) to Delhi (city level) and below without the need for a state level.

User-defined: These are user-defined hierarchies of attributes that are used in service of Microsoft SQL Server to arrange the members of a dimension into hierarchical structures and provide navigation paths in a form of cube.

For example, the Table 2.2 defines a dimension table for a time dimension.

Table 2.2: Dimension Table for Time Dimension

Year	Quarter	Month
1999	Quarter 1	Jan
1999	Quarter 1	Feb
1999	Quarter 1	Mar
1999	Quarter 2	Apr
1999	Quarter 2	May
1999	Quarter 2	Jun
1999	Quarter 3	Jul
1999	Quarter 3	Aug
1999	Quarter 3	Sep
1999	Quarter 4	Oct
1999	Quarter 4	Nov
1999	Quarter 4	Dec

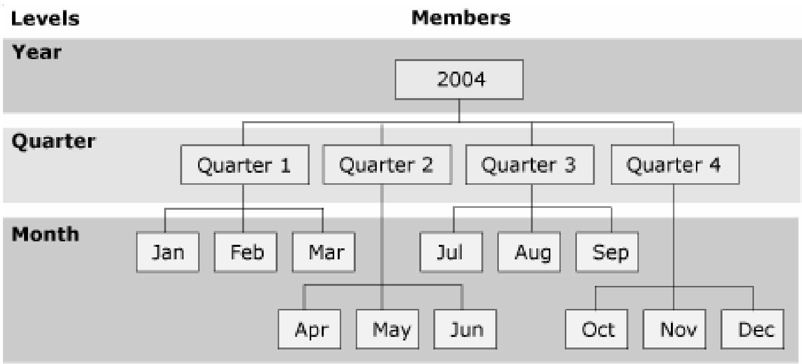
Source: http://msdn.microsoft.com/en-us/library/ms174935.aspx

Did u know? The Year, Quarter, and Month attributes are used to construct a user-defined hierarchy, named Calendar, in the time dimension.

Task Prepare a table hierarchy of employees according to their salary.

The relationship between the levels and members of the Calendar dimension is shown in Figure 2.4.

Figure 2.4: Relationship between the Levels and Members of the Calendar Dimension



Source: http://i.msdn.microsoft.com/dynimg/IC165224.gif

Notes Any hierarchy other than the default two-level attribute hierarchy is called a user-defined hierarchy.

Notes

Self Assessment

State whether the following statements are True or False:

5. A hierarchy is a set of parent-child relationships between attributes within a dimension.
6. Level-based hierarchies can also contain parent-child relationships.
7. A Ragged Hierarchy describes a self-referencing relationship, or self-join, within a dimension main table.
8. Parent attribute is also known as an unbalanced hierarchy.
9. Skip-level is hierarchies of attributes that are used in service of Microsoft SQL Server to arrange the members of a dimension into hierarchical structures.
10. User-defined is a hierarchy in which certain members do not have values for certain higher levels are known as skip-level hierarchy.
11. Ragged Hierarchy consists of values that define the hierarchy in a parent-child relationship.

Case Study

Mining a Star Schema

One of the strengths of Oracle Data Mining is the ability to mine star schemas with minimal effort. Star schemas are commonly used in relational databases, and they often contain rich data with interesting patterns. While dimension tables may contain interesting demographics, fact tables will often contain user behaviour, such as phone usage or purchase patterns. Both of these aspects - demographics and usage patterns - can provide insight into behaviour.

Churn is a critical problem in the telecommunications industry, and companies go to great lengths to reduce the churn of their customer base. One case study describes a telecommunications scenario involving understanding, and identification of, churn, where the underlying data is present in a star schema. That case study is a good example for demonstrating just how natural it is for Oracle Data Mining to analyse a star schema, so it will be used as the basis for this series of posts.

The case study schema includes four tables: CUSTOMERS, SERVICES, REVENUES, and CDR_T. The CUSTOMERS table contains one row per customer, as does the SERVICES table, and both contain a customer id that can be used to join the tables together. Most data mining tools are capable of handling this type of data, where one row of input corresponds to one case for mining. The other two tables have multiple rows for each customer. The CDR_T (call data records) table contains multiple records for each customer which captures calling behaviour. In the case study, this information is already pre-aggregated by type of call (peak, international, etc.) per month, but the information may also be available at a finer level of granularity. The REVENUES table contains the revenue per customer on a monthly basis for a five month history, so there are up to five rows per customer. Capturing the information in the CDR_T and REVENUES table to help predict churn for a single customer requires collapsing all of this fact table information into a single “case” per customer. Most tools will require pivoting the data into columns, which has the drawbacks of densifying data as well as pivoting data beyond column count limitations. The data in a fact table is often stored in sparse form (this case study aggregates it to a denser form, but it need not be this way for other mining activities), and keeping it in sparse form is highly desirable.

Contd....

For fact table data that has a much larger number of interesting groups (such as per-product sales information of a large retailer), retaining the sparse format becomes critical to avoid densification of such high cardinality information. Oracle Data Mining algorithms are designed to interpret missing entries in a sparse fact table appropriately, enabling increased performance and simpler transformation processing.

Some steps in the referenced case study are not completely defined (in my opinion), and in those situations I will take my best guess as to the intended objective. This approximation is sufficient since the intent of this series of posts is to show the power and flexibility of Oracle Data Mining on a real-world scenario rather than to match the case study letter-for-letter.

The following files support reproduction of the results in this series of posts:

telcoddl.sql - SQL which creates the four tables

telcoloadproc.plb - Obfuscated SQL which creates the procedure that can generate data and populate the tables - all data is generated, and patterns are injected to make it interesting and "real-world" like

telcoprep.sql - A SQL create view statement corresponding to the data preparation steps from part 2 of this series

telcomodel.sql - A SQL script corresponding to the steps from part 3 of this series

In order to prepare a schema that can run the above SQL, a user must be created with the following privileges: create table, create view, create mining model, and create procedure (for telcoloadproc), as well as any other privs as needed for the database user (e.g., create session).Once the schema is prepared, telcoddl.sql and telcoloadproc.plb can be run to create the empty tables and the procedure for loading data. The procedure that is created is named telco_load, and it takes one optional argument - the number of customers (default 10000). The results from parts 2 and 3 of this series correspond to loading 10,000 customers.

The sample code in these posts has been tested against an 11gR2 database. Many new features have been added in each release, so some of the referenced routines and syntax are not available in older releases; however, similar functionality can be achieved with 10g. The following modified scripts can be used with 10g (tested with 10gR2):

telcoprep_10g.sql - A SQL create view statement corresponding to the data preparation steps from part 2 of this series, including substitution for the 11g PIVOT syntax and inclusion of manual data preparation for nested columns.

telcomodel_10g.sql - A SQL script corresponding to the steps from part 3 of this series, including substitution of the Generalized Linear Model algorithm for 10g Support Vector Machine, manual data preparation leveraging the transformation package, use of dbms_data_mining.apply instead of 10gR2 built-in data mining scoring functions, explicit commit of settings prior to build, and removal of the EXPLAIN routine from the script flow.

In addition, the create mining model privilege is not available in 10g.

1. ***Handling missing values for call data records:*** The CDR_T table records the number of phone minutes used by a customer per month and per call type (tariff). For example, the table may contain one record corresponding to the number of peak (call type) minutes in January for a specific customer, and another record associated with international calls in March for the same customer. This table is likely to be fairly dense (most type-month combinations for a given customer will be present)

Contd....

due to the coarse level of aggregation, but there may be some missing values. Missing entries may occur for a number of reasons: the customer made no calls of a particular type in a particular month, the customer switched providers during the timeframe, or perhaps there is a data entry problem. In the first situation, the correct interpretation of a missing entry would be to assume that the number of minutes for the type-month combination is zero. In the other situations, it is not appropriate to assume zero, but rather derive some representative value to replace the missing entries. The referenced case study takes the latter approach. The data is segmented by customer and call type, and within a given customer-call type combination, an average number of minutes is computed and used as a replacement value.

In SQL, we need to generate additional rows for the missing entries and populate those rows with appropriate values. To generate the missing rows, Oracle’s partition outer join feature is a perfect fit.

```
select cust_id, cdre.tariff, cdre.month, mins
from cdr_t cdr partition by (cust_id) right outer join
  (select distinct tariff, month from cdr_t) cdre
  on (cdr.month = cdre.month and cdr.tariff = cdre.tariff);
```

I have chosen to use a distinct on the CDR_T table to generate the set of values, but a more rigorous and performant (but less compact) approach would be to explicitly list the tariff-month combinations in the cdre inlined subquery rather than go directly against the CDR_T table itself.

Now that the missing rows are generated, we need to replace the missing value entries with representative values as computed on a per-customer-call type basis. Oracle’s analytic functions are a great match for this step.

```
select cust_id, tariff, month,
  nvl(mins, round(avg(mins) over (partition by cust_id, tariff))) mins
from (<prev query>);
```

We can use the avg function, and specify the partition by feature of the over clause to generate an average within each customer-call type group. The nvl function will replace the missing values with the tailored, computed averages.

2. ***Transposing Call Data Records:*** The next transformation step in the case study involves transposing the data in CDR_T from a multiple row per customer format to a single row per customer by generating new columns for all of the tariff-month combinations. While this is feasible with a small set of combinations, it will be problematic when addressing items with higher cardinality. Oracle Data Mining does not need to transpose the data. Instead, the data is combined using Oracle’s object-relational technology so that it can remain in its natural, multi-row format. Oracle Data Mining has introduced two data types to capture such data - DM_NESTED_NUMERICALS and DM_NESTED_CATEGORICALS.

   In addition, the case study suggests adding an attribute which contains the total number of minutes per call type for a customer (summed across all months). Oracle’s rollup syntax is useful for generating aggregates at different levels of granularity.

```
select cust_id,
  cast(collect(dm_nested_numerical(tariff||’-’||nvl(month,’ALL’),mins))
  as dm_nested_numericals) mins_per_tariff_mon from
 (select cust_id, tariff, month, sum(mins) mins
```

Contd....

```
 from (<prev query>)
 group by cust_id, tariff, rollup(month))
group by cust_id;
```

The above query will first aggregate the minutes by cust_id-tariff-month combination, but it will also rollup the month column to produce a total for each cust_id-tariff combination. While the data in the case study was already aggregated at the month level, the above query would also work on data that is at a finer granularity.

Once the data is generated by the inner query, there is an outer group by on cust_id with the COLLECT operation. The purpose of this step is to generate an output of one row per customer, but each row contains an entry of type DM_NESTED_NUMERICALS. This entry is a collection of pairs that capture the number of minutes per tariff-month combination.

While we performed missing value replacement in the previous transformation step, thereby densifying the data, Oracle Data Mining has a natural treatment for missing rows. When data is presented as a DM_NESTED_NUMERICALS column, it is assumed that any missing entries correspond to a zero in the value - matching the first option for missing value treatment described earlier. If this is the correct interpretation for missing values, then no missing value treatment step is necessary. The data can remain in sparse form, yet the algorithms will correctly interpret the missing entries as having an implicit value of zero.

3. ***Transposing Revenue Records:*** Again, no need to transpose when using Oracle Data Mining. We add an aggregate to produce the total revenue per customer in addition to the per-month breakout coming from the COLLECT.

```
select cust_id, sum(revenue) rev_tot_sum,
 cast(collect(dm_nested_numerical('REV-'||month, revenue))
 as dm_nested_numericals) rev_per_mon
from revenues
group by cust_id;
```

4. ***Creating Derived Attributes:*** The final transformation step in the case study is to generate some additional derived attributes, and connect everything together so that each customer is composed of a single entity that includes all of the attributes that have been identified to this point.

   The PIVOT operation is used to generate named columns that can be easily combined with arithmetic operations. Binning and filtering steps, as identified in the case study, are included in the above SQL.

   The query can execute in parallel on SMPs, as well as MPPs using Oracle's RAC technology. The data can be directly fed to Oracle Data Mining without having to extract it from the database, materialize copies of any parts of the underlying tables, or pivot data that is in a naturally multi-row format.

Questions:

1. How the missing values were handled for call data records?
2. Why do we transposing revenue records?

Source: http://amozes-oracle.blogspot.in/2010/12/mining-star-schema-telco-churn-case.html

Notes

2.3 Summary

- If the dimension is defined as a hierarchy, the lower levels of the hierarchy must also have an attribute that identifies the parent of each member.
- Typically, the key attribute represents the primary key column or columns in the dimension table.
- A hierarchy is a set of parent-child relationships between attributes within a dimension. These hierarchy attributes are also known as levels.
- Level-based hierarchies can contain parent-child relationships.
- A parent-child hierarchy is a hierarchy in a standard dimension that contains a parent attribute.
- A hierarchy in which all the lowest-level members do not have the same level of depth is ragged hierarchy.
- The Year, Quarter, and Month attributes are used to construct a user-defined hierarchy, named Calendar, in the time dimension.

2.4 Keywords

Hierarchy: A hierarchy is a set of parent-child relationships between attributes within a dimension.

Key attribute: The key attribute is the attribute in a dimension that identifies the columns in the dimension main table that are used in foreign key relationships to the fact table.

Level-based: This type of hierarchy consists of an ordered set of two or more levels.

Parent-child: A parent-child hierarchy is a hierarchy in a standard dimension that contains a parent attribute.

Ragged Hierarchy: A hierarchy in which all the lowest-level members do not have the same level of depth is ragged hierarchy.

Skip-level: A hierarchy in which certain members do not have values for certain higher levels are known as skip-level hierarchy.

User-defined: These are user-defined hierarchies of attributes that are used in service of Microsoft SQL.

2.5 Review Questions

1. What is the multi-dimensional analysis?
2. Discuss about the key attribute of dimension attributes.
3. Briefly explain the dimension attribute properties.
4. What is the dimension hierarchy? Explain with example.
5. Discuss are the various type of hierarchy.
6. "Parent-child is actually value-based hierarchy". Comment.
7. What is ragged hierarchy? Give the suitable example.
8. Explain the relationship between the levels and members of the calendar dimension.

Answers: Self Assessment

1. Multi-Dimensional analysis
2. Dimension tables
3. Primary key column or columns
4. Attribute
5. True
6. True
7. False
8. False
9. False
10. False
11. False

2.6 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

dssresources.com/developers/tools/multidimensionalanalysis.html

www.bris.ac.uk/poverty/downloads/socialexclusion/multidimensional.pdf

www.cs.umu.se/education/examina/Rapporter/PerWesterlund.pdf

www.georgehart.com/research/multanal.html

www.learn.geekinterview.com › Data Warehouse › Data Analysis

Sartaj Singh, Lovely Professional University

Unit 3: Dimensional Data Warehouse

CONTENTS

Objectives

After studying this unit, you will be able to:

- Describe about Dimensional Model
- Construct Facts Table
- Demonstrate Dimension Tables
- Discuss about Surrogate Keys and Alternative Table Structure
- Explain Multidimensional OLAP

Introduction

Dimensions are a common way of analysing data. Dimension model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data. Dimensional data modelling is the preferred modelling technique in a BI environment. Knowing the basics of data warehousing and dimensions helps you design a better data warehouse that fits your reporting needs. This unit on data warehousing dimensions explains the importance of dimensions and dimension granularity and stresses the importance of flattening hierarchies—with the goal being to make data more accessible and useful to users. It also focuses on fact and dimension table.

3.1 Dimensional Model

Dimensional model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data. Since Business Intelligence reports are used in assessing the facts (aggregates) across various dimensions, dimensional data modelling prefer the modelling technique in a BI environment.

Facts are normally calculated data like dollars' worth or Sales or income. They correspond to the aim of a conclusion support analysis.

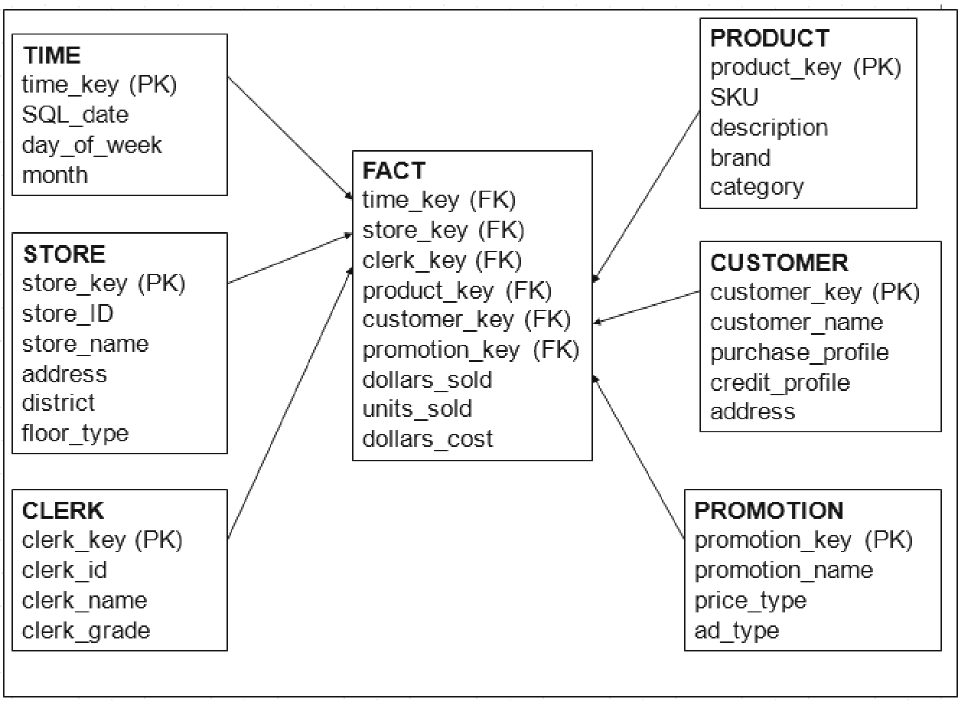
Dimensions define the axis of enquiry of a fact.

Example: For example, Product, Region and Time are the axes of enquiry of the Sales detail.

One such enquiry could be a scenario where the user might require to see the Sales (in dollars) for a specific item in a market over a specific time span of time. In this case, we are calculating the fact (Sales) over three dimensions (Product, Region and Time). Thus we can say that dimensions give different views of the facts. They give structure to the otherwise unstructured facts.

It typically contains the attributes for the SQL answer set. Figure 3.1 shows an example of dimensional model.

Figure 3.1: Example of Dimensional Model



Source: www.oedewaldt.com/movies/dimensional%20modeling.pptý

Notes

Self Assessment

Fill in the blanks:

1. Dimensional model comprises of a .... and numerous dimensional tables and is used for assessing summarized data.
2. .... define the axis of enquiry of a fact.

3.2 Facts Table

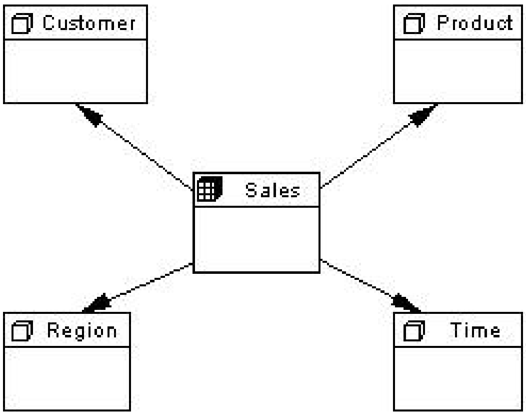
Fact table generally represent a process or reporting environment that is of value to the organization. It is important to determine the identity of the fact table and specify exactly what it represents. A fact table typically corresponds to an associative entity in the E-R model.

They must be listed in a logical fact table. Each measure has its own aggregation rules such as ADD, AVG, MIN or MAX. Aggregation rules define the way by which business would like to contrast standards of a measured value.

Caution Facts are the measurements associated with fact table records at fact table granularity.

The Figure 3.2 displays how Sales detail table is connected in a One-to-Many relationships with other dimension tables.

Figure 3.2: Sales Details Table One-to-many Relationship



Source: http://2.bp.blogspot.com/-JR3HxkgK6w0/Ti_PS_Egw3I/AAAAAAAAAMQ/tOdQBrMG3FU/s1600/Star_Model.JPG

3.2.1 Types of Measure

Various types of measure in a fact table are:

- ***Additive -*** Measures that can be added across any dimensions are additive measure.
- ***Semi Additive -*** Measures that can be added across only some dimensions are semi additive.
- ***Non Additive -*** Measures that cannot be added across any dimension are non-additive.

3.2.2 Types of Fact Table

There are basically three types of fact tables:

- ***Transactional:*** A transactional table is the most basic and fundamental type of fact table. The grain associated with a transactional fact table is usually specified as one row per line in a transaction, e.g., every line on a receipt represents a transaction.
- ***Periodic Snapshots:*** It takes a picture of the moment, where the moment could be anything like performance summary of a salesman over the previous 3 months. A periodic snapshot table is dependent on the transactional table.
- ***Accumulating Snapshots:*** In this type of fact table the activity of a process is shown such that it has a well-defined beginning and end.

Example: The processing of an order where an order moves through specific steps until it is completed.

As steps towards fulfilling the order are completed, the row which is associated with it is updated in the fact table. This type of table often has multiple date columns, each representing a complete step in the process. Therefore, it's important to have an entry in the date dimension that represents an unknown date, as many of the milestone completion time are unknown at the time the row is created.

Self Assessment

Fill in the blanks:

3. A fact table typically corresponds to an associative entity in the ....
4. Measures that can be added across only some dimensions are ....
5. .... take a picture of the moment, where the moment could be anything.
6. In .... table often has multiple date columns, each representing a complete step in the process.

3.3 Dimension Tables

Dimension tables consist of attributes that describe fact records in the fact table. Some of these attributes provide descriptive information; others are used to specify how fact table data should be summarized to provide useful information to the person who is analysing the information. Every dimension has a set of descriptive attributes. Dimension tables contain attributes that describe business entities.

Example: The Client dimension can contain attributes like C_No., Area, State, Country etc.

Did u know? In a dimensional table, columns can be used to categorize the information into hierarchical levels.

Notes

For example, a dimension table for stores in the StandardMart sample database includes the following columns:

Table 3.1: Sample Dimension Table

Column	Description
store_country	Specifies the country or region in which the store is located. This is the country level of the hierarchy.
store_state	Specifies the state in which the store is located. This is the state level of the hierarchy.
store_city	Specifies the city or province in which the store is located. This is the city level of the hierarchy.
store_id	Specifies the individual store. This is the lowest level of the hierarchy. This field contains the primary key of the store dimension table and is used to join the dimension table to the fact table.
store_name	Specifies the name of the store. The values in this column are used to identify the store to users in a readable form.

Source: http://msdn.microsoft.com/en-us/library/aa905979(v=sql.80).aspx

Self Assessment

Fill in the blanks:

7. Dimension tables consist of attributes that describe .... in the fact table.

8. .... contain attributes that describe business entities.

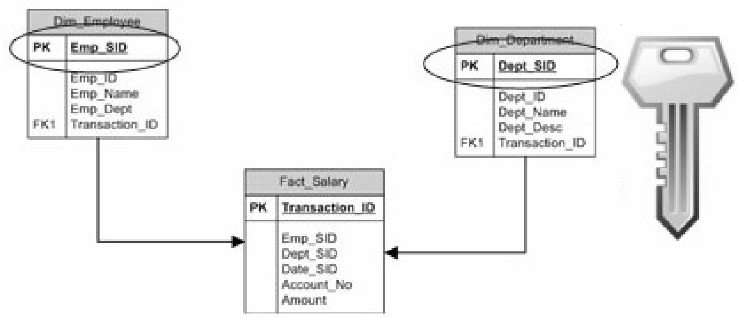
3.4 Surrogate Keys and Alternative Table Structure

A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database. The surrogate key is not derived from application data. Surrogate keys are keys that are maintained within the data warehouse instead of keys taken from source data systems.

Example: Say for the employee 'Emp12 the Business unit changes from B1 to B2. Now, if you use the natural primary key 'Emp12 for your employees within your data warehouse then everything would be allocated to Business unit 'B22 even what actually belongs to 'B1.'

If you use surrogate keys, you could create on the other day a new record for the Employee 'Emp12 in your Employee Dimension with a new surrogate key.

Figure 3.3: Surrogate Key Example



Source: http://mahaveersingh.files.wordpress.com/2012/05/surrogate_key_blog_banner1.jpg

This way, in your fact table, you have your old data (i.e. before the day you added) with the SID of the Employee 'Emp12 >> 'B1.' All new data (i.e. after the day you added) would take the SID of the employee 'Emp12 >> 'B2.'

3.4.1 Advantages of Surrogate Keys

- ***Immutability:*** Surrogate keys do not change while the row exists. Thus applications cannot misplace their reference in the database.
- ***Change in Requirements:*** Attributes that uniquely recognize an entity might change over the time, which might lead to invalidation of the suitability of the compound keys.

Example: An employee's network username is chosen as a natural key. If it is merged with another company, new employees must be inserted. Now, some of the new user names may lead to conflict because their user names were developed independently.

In these cases, usually a new attribute should be added to the natural key (for example, an old_company column). In the case of a surrogate key, only the table that characterizes the surrogate key must be altered. But in the case of natural keys, all tables that use the natural key will have to change.

- ***Performance:*** Surrogate keys tend to be a compact data type, such as a four-byte integer. This allows the database to query the single key column faster than it could multiple columns.
- ***Uniformity:*** When every table has a uniform surrogate key, some tasks can be easily automated by composing the code in a table-independent way.
- ***Validation:*** It is possible to design key-values that are in coordination with a well-known pattern which can be automatically verified.

Example: The keys that are intended to be used in some column of some table might be designed to "look differently from" those that are intended to be used in another column or table, thereby simplifying the detection of application errors in which the keys have been misplaced.

3.4.2 Disadvantages of Surrogate Keys

But surrogate keys also come with some disadvantages. The values of surrogate keys have no relationship with the real world meaning of the data held in a row. Therefore over usage of surrogate keys lead to the problem of disassociation and creates unnecessary ETL burden and performance degradation.

Query optimization also becomes difficult when one disassociates the surrogate key with the natural key. This is because when surrogate key takes the place of primary key, unique index is applied on that column. And any query based on natural key identifier leads to full table scan as that query cannot take the advantage of unique index on the surrogate key.

- ***Referential Integrity:*** Referential integrity must be maintained between all dimension tables and the fact table. Each fact record contains foreign keys which are related to primary keys in the dimension tables.

Caution Every fact record must have a related record in every dimension table used with that particular fact table.

- ***Shared Dimensions:*** To maintain consistency dimension tables that are shared are created. These tables are used by all components and data marts in the data warehouse.

3.4.3 Alternative Tables used in Data Warehousing

Auxiliary Table

This table is created with the SQL statements CREATE AUXILIARY TABLE and is used to hold the data for a column that is defined in a base table.

Base Table

The most common type of table is base table. You can create a base table with the SQL CREATE TABLE statement. All programs and users that refer to this type of table refer to the same description of the table and to the same instance of the table.

Clone Table

A table that is structurally identical to a base table is known as clone table. You can create a clone table by using an ALTER TABLE statement for the base table that includes an ADD CLONE clause.

Example: In the DB2 catalogue, SYSTABLESPACE.CLONE indicates that a clone table exists.

Empty Table

A table with zero rows is an empty table.

History Table

A history table is used by Database to store historical versions of rows from the associated system period temporal table.

Materialized Query Table

Materialized query tables are useful for complex queries that run on large amounts of data.

Notes They are commonly used in data warehousing and business intelligence applications.

Result Table

A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table. A result table is not an object that you can define using a CREATE statement.

Temporal Table

A temporal table is a table that records the period of time when a row is valid.

A table that is defined by the SQL statement CREATE GLOBAL TEMPORARY TABLE or DECLARE GLOBAL TEMPORARY TABLE is temporary table. It is used to hold data temporarily.

XML Table

It is a special table that holds only XML data. When you create a table with an XML column, database implicitly creates an XML table space and an XML table to store the XML data.

Self Assessment

State whether the following statements are true or false:

9. The surrogate key is derived from application data.
10. Surrogate keys change while the row exists.
11. In the case of natural keys, all tables that use the natural key will have to change.
12. When every table has a uniform surrogate key, some tasks can be easily automated by composing the code in a table-independent way.
13. The values of surrogate keys have relationship with the real world meaning of the data held in a row.
14. The most common type of table is base table.
15. A table with zero rows is an empty table.
16. A temporal table is a table that records the period of time when a row is valid.
17. XML table is used to hold data temporarily.

3.5 Multidimensional OLAP

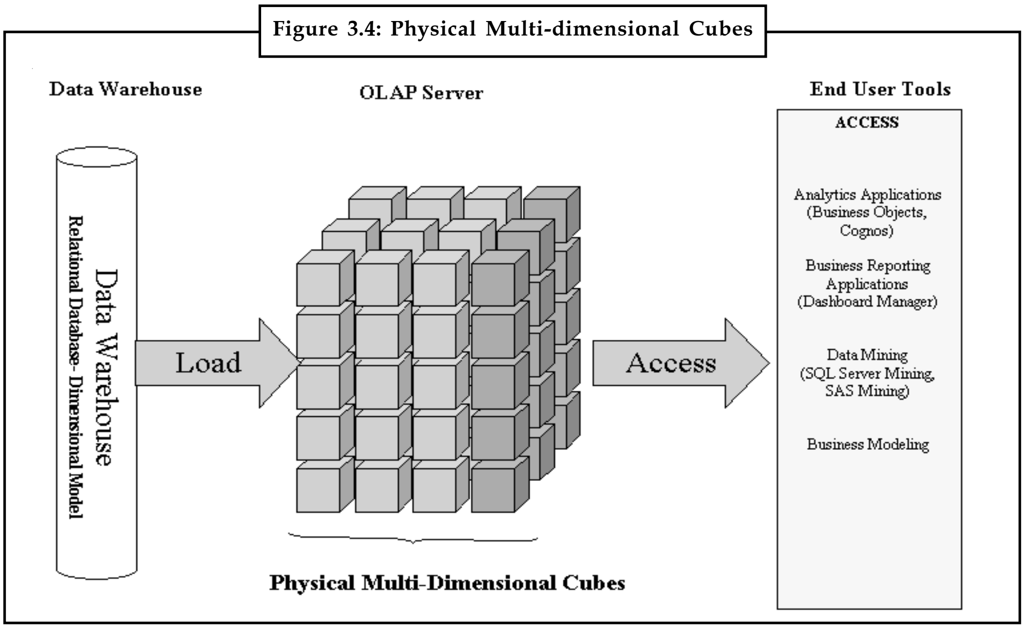
OLAP stands for On-Line Analytical Processing. In computing, OLAP is an approach to answering Multi-Dimensional Analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also includes relational database, report writing and data mining. Depending on the underlying technology used, OLAP can be broadly divided into MOLAP and ROLAP.

Did u know? In the OLAP world, there are mainly two different types: Multidimensional OLAP (MOLAP) and Relational OLAP (ROLAP). Hybrid OLAP (HOLAP) is combination of MOLAP and ROLAP.

3.5.1 MOLAP

In MOLAP, data is stored in a multidimensional cube. It fulfils the requirements for an analytic application, where you require to access only summarized level of data. The storage is not in the relational database, but in proprietary formats. Figure 3.4 shows physical multi-dimensional cubes.

Figure 3.4: Physical Multi-dimensional Cubes


Data Warehouse
OLAP Server
End User Tools
ACCESS
Analytics Applications (Business Objects, Cognos)
Business Reporting Applications (Dashboard Manager)
Data Mining (SQL Server Mining, SAS Mining)
Business Modeling
Data Warehouse
Relational Database- Dimensional Model
Load
Access
Physical Multi-Dimensional Cubes

Source: http://www.executionmih.com/dipm_images/ZCA-MOLAP.GIF

This method stores the data in multi-dimensional arrays which is different from the two dimensional relational structure.

Advantages:

- MOLAP cubes are built for fast data retrieval and are thus optimal for slicing operations.
- MOLAP can perform complex calculations quickly.

Disadvantages:

- MOLAP is limited in the amount of data it can handle because all the calculations are performed when the cube is built.
- Cube technology generally do not already exist in the organization, therefore, to adopt MOLAP technology, chances are additional investments in the form of human and capital is needed.

3.5.2 ROLAP

This methodology depends on manipulating the data stored in the relational database. There are detail level values in relational data warehouse.

Advantages:

- ROLAP can handle large amounts of data.
- ROLAP can leverage functionalities inherent in the relational database as they sit on top of the relational database.

Disadvantages:

- In ROLAP the performance can be slow. As it is known that ROLAP report is essentially a SQL query on the relational database, the query time can be long if the underlying data size is large thus the performance of same can be slow.

- ROLAP can be limited by SQL functionalities. As, ROLAP technology mainly relies on SQL statements and SQL statements do not fit all needs (like it's not easy to do complex queries in SQL), thus what ROLAP can do is traditionally limited by what SQL can do.

3.5.3 HOLAP

HOLAP technologies combine the advantages of MOLAP and ROLAP. The first product to provide HOLAP storage was Holos but with time the technology also became available in other commercial products such as Microsoft Analysis Services (MAS), Oracle Database OLAP Option, MicroStrategy etc.

Task Compare and contrast the MOLAP and ROLAP.

Self Assessment

Fill in the blanks:

18. OLAP stands for ....

19. OLAP can be broadly divided into .... and ....

20. In MOLAP, data is stored in a ....

21. .... can leverage functionalities inherent in the relational database as they sit on top of the relational database.

22. .... technologies combine the advantages of MOLAP and ROLAP.

Case Study

Lolopop: Automated Data Warehouse

The essential concept of a data warehouse is to provide the ability to gather data into optimized databases without regard for the generating applications or platforms. Data warehousing can be formally defined as "the coordinated, architected, and periodic copying of data from various sources into an environment optimized for analytical and informational processing".

The Challenge

Meaningful analysis of data requires us to unite information from many sources in many forms, including: images; text; audio/video recordings; databases; forms, etc. The information sources may never have been intended to be used for data analysis purposes. These sources may have different formats, contain inaccurate or outdated information, be of low transcription quality, be mislabelled or be incompatible.

New sources of information may be needed periodically and some elements of information may be one time only artefacts.

A data warehouse system designed for analysis must be capable of assimilating these data elements from many disparate sources into a common form. Correctly labelling and describing search keys and transcribing data in a form for analysis is critical. Qualifying

Contd....

Notes

the accuracy of the data against its original source of authority is imperative. Any such system must also be able to: apply policy and procedure for comparing information from multiple sources to select the most accurate source for a data element; correct data elements as needed; and check inconsistencies amongst the data. It must accomplish this while maintaining a complete data history of every element before and after every change with attribution of the change to person, time and place. It must be possible to apply policy or procedure within specific periods of time by processing date or event data to assure comparability of data within a calendar or a processing time horizon. When data originates from a source where different policies and procedures are applied, it must be possible to reapply new policies and procedures. Where quality of transcription is low qualifying the data through verification or sampling against original source documents and media is required. Finally, it must be possible to recreate the exact state of all data at any date by processing time horizon or by event horizon.

The analytical system applied to a data warehouse must be applicable to all data and combinations of data. It must take into account whether sufficient data exists at the necessary quality level to make conclusions at the desired significance level. Where possible it must facilitate remediation of data from original primary source(s) of authority.

When new data is acquired from new sources, it must be possible to input and register the data automatically. Processing must be flexible enough to process these new sources according to their own unique requirements and yet consistently apply policy and procedure so that data from new sources is comparable to existing data.

When decisions are made to change the way data is processed, edited, or how policy and procedure is applied, it must be possible to exactly determine the point in time that this change was made. It must be possible to apply old policies and procedures for comparison to old analyses, and new policy and procedure for new analyses.

Defining Data Warehouse Issues

The Lolopop partners served as principals in a data warehouse effort with objectives that are shared by most users of data warehouses. During business analysis and requirements gathering phase, we found that high quality was cited as the number one objective. Many other objectives were actually quality objectives, as well. Based on our experiences, Lolopop defines the generalized objectives in order of importance as:

Quality information to Create data and/or combine with other data sources

In this case, only about one in eight events could be used for analysis across databases. Stakeholders said that reporting of the same data from the same incoming information varied wildly when re-reported at a later date or when it came from another organization's analysis of the same data. Frequently the data in computer databases was demonstrably not contained in the original documents from which they were transcribed. Conflicting applications of policy and procedure by departments with different objectives, prejudices and perspectives were applied inconsistently without recording the changes or their sources, leaving the data for any given event a slave to who last interpreted it.

Timely response to requests for data

Here, the data was processed in time period batches. In some instances, it could take up to four years to finalize a data period. Organizations requiring data for analysis simply went to the reporting source and got their own copies for analysis, entirely bypassing the official data warehouse and analytical sources.

Contd....

Notes

Consistent relating of information

An issue as simple as a name — the information that could be used to connect data events to histories for individuals or other uniting objects — had no consistent method to standardize or simplify naming conventions. Another example, Geographical Information System (GIS) location information had an extravagant infrastructure that was constantly changing. This made comparisons of data from two different time periods extremely difficult.

Easy access to information

Often data warehouse technologies assume or demand a sophisticated understanding of relational databases and statistical analysis. This prevents ordinary stakeholders from using data effectively and with confidence. In some instances, the personnel responsible for analysis lack the professional and technical skills to develop effective solutions. This issue can stultify reporting to a few kinds of reports and variants that have been programmed over time, and reduces data selection for the analyses to kind of magic applied by clerical personnel responsible for generating reports.

Unleash management to formulate and uniformly apply policy and procedure

We found that management decisions and mandates could be hindered by an inability to effectively capture, store, retrieve and analyse data.

In this particular instance, no management controls existed to analyse: source of low quality; work rates; work effort to remediate (or even a concept of remediation); effectiveness of procedures; effectiveness of work effort; etc.

Remediation is a good case in point. Management experienced difficulty with the concept of remedying data transcription from past paper forms — even though the forms existed in images that could be automatically routed. The perception was that quantity of data, not quality, was the objective and that no one would ever attempt to fix data by verifying it or comparing it to original documents.

Manage incoming data from non-integrated sources

Data from multiple, unrelated sources requires a plan to convert electronic data, manage imaging and documents inputs, manage workflow and manage the analysis of data. In this case, every interface required manual intervention. Since there was no system awareness at the beginning of the capture process as to what was needed for analysis at the end, it was very difficult to make rapid and time effective changes to accommodate changing stakeholder needs.

Reproducible Reporting Results

We found that reporting of data was not reproducible and the reasons for differences in reporting were not retrievable, undermining confidence in the data, analysis and reporting. One may essentially summarize these objectives as quality challenges that require a basic systems engineering approach for resolution.

Questions:

1. What were the challenges of lolopop automated data warehouse?
2. What were the data warehouse issues?

Source: http://www.lolopop.net/Lolopop.DWStudy.pdf

Notes

3.6 Summary

- Dimensional model comprises of a fact table and numerous dimensional tables and is used for assessing summarized data.
- Fact table generally represent a process or reporting environment that is of value to the organization.
- A fact table typically corresponds to an associative entity in the E-R model.
- Various types of measure in a fact table are: Additive, Semi Additive, Non-Additive.
- There are basically three types of fact tables: Transactional, Periodic snapshots and accumulating snapshots.
- Dimension tables consist of attributes that describe fact records in the fact table.
- A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database.
- Attributes that uniquely recognize an entity might change over the time, which might lead to invalidation of the suitability of the compound keys.
- But surrogate keys also come with some disadvantages. The values of surrogate keys have no relationship with the real world meaning of the data held in a row.
- Referential integrity must be maintained between all dimension tables and the fact table.
- The most common type of table is base table. You can create a base table with the SQL CREATE TABLE statement.
- A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table.
- OLAP stands for On-Line Analytical Processing. In computing, OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly.
- In MOLAP, data is stored in a multidimensional cube. It fulfils the requirements for an analytic application, where you require to access only summarized level of data.
- HOLAP technologies combine the advantages of MOLAP and ROLAP.

3.7 Keywords

Accumulating Snapshots: In this type of fact table the activity of a process is shown such that it has a well-defined beginning and end.

Auxiliary Table: This table is created with the SQL statements CREATE AUXILIARY TABLE and is used to hold the data for a column that is defined in a base table.

Dimension Tables: Dimension tables consist of attributes that describe fact records in the fact table.

Dimensional Model: Dimensional Modelling (DM) is the name of a set of techniques and concepts used in data warehouse design. It is considered to be different from entity-relationship modelling (ER).

Empty Table: It is a table with zero rows is an empty table.

E-R Model: In software engineering, an Entity-relationship model (ER model) is a data model for describing a database in an abstract way.

Fact Table: Fact table generally represent a process or reporting environment that is of value to the organization.

HOLAP: HOLAP (Hybrid Online Analytical Processing) is a combination of ROLAP (Relational OLAP) and MOLAP (Multidimensional OLAP) which are other possible implementations of OLAP.

Multidimensional Online Analytical Processing (MOLAP): This is the more traditional way of OLAP analysis. In MOLAP, data is stored in a multidimensional cube. The storage is not in the relational database, but in proprietary formats.

Result Table: A table that contains a set of rows that a database selects or generates, directly or indirectly, from one or more base tables in response to an SQL statement is known as result table.

ROLAP: This methodology relies on manipulating the data stored in the relational database to give the appearance of traditional OLAP's slicing and dicing functionality.

Surrogate Key: A surrogate key in a database is a unique identifier for either an entity in the modelled world or an object in the database.

Temporal Table: A temporal table is a table that records the period of time when a row is valid.

Transactional Table: The grain associated with a transactional fact table is usually specified as one row per line in a transaction.

XML Table: It is a special table that holds only XML data.

3.8 Review Questions

1. What is dimension?
2. Explain the dimensional model.
3. “Dimensions define the axis of enquiry of a fact.” Elucidate.
4. Give examples of dimensional model.
5. What do you understand by fact table?
6. Explain the types of measure and fact table.
7. “Dimension tables consist of attributes that describe fact records in the fact table”. Discuss.
8. Define the concept of surrogate key. Also write down the advantages and disadvantages.
9. What do you understand by alternative tables used in data warehousing?
10. Briefly explain about multidimensional OLAP.

Answers: Self Assessment

1. Fact table
2. Dimensions
3. E-R model
4. Semi additive
5. Periodic snapshots
6. Accumulating snapshots
7. Fact records
8. Dimension tables
9. False
10. False

Notes

11. True
12. True
13. False
14. True
15. True
16. True
17. False
18. On-Line Analytical Processing
19. MOLAP, ROLAP
20. Multidimensional cube
21. ROLAP
22. HOLAP

3.9 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

http://sqlmag.com/business-intelligence/data-warehousing-dimension-basics

msdn.microsoft.com/en-us/library/aa905979(v=sql.80).aspx?

www.b-eye-network.com/view/757?

www.fcsm.gov/99papers/yost.pdf?

Unit 4: Understanding OLAP

CONTENTS

Objectives

After studying this unit, you will be able to:

- Recognize Basic Concepts of OLAP
- List the Advantages of OLAP
- Explain about the Fast Response
- Describe the Metadata Based Queries
- Construct the Spreadsheet Formulas
- Discuss about the Analysis Services Speed
- Discuss about Metadata

Introduction

Online Analytical Processing (OLAP) is a technology that is used to create decision support software. OLAP enables application users to quickly analyse information that has been summarized into multidimensional views and hierarchies. By summarizing predicted queries into multidimensional views prior to run time, OLAP tools provide the benefit of increased performance over traditional database access tools. Most of the resource-intensive calculation that is required to summarize the data is done before a query is submitted. This unit on OLAP explains the concepts and advantages of OLAP, spreadsheet formulas. It also covers study of metadata.

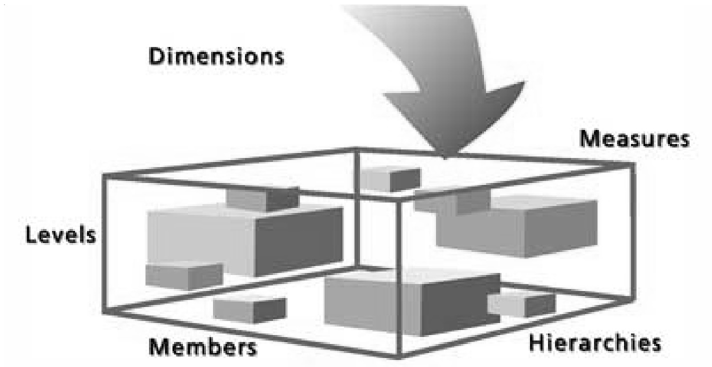
4.1 Basic Concepts of OLAP

OLAP is a database expertise that has been optimized for querying and describing, rather than of processing transactions. OLAP data is drawn from historical data, and it is aggregated into structures that allow complicated analysis. Data in OLAP is also coordinated hierarchically and placed in cubes instead of tables. It is a sophisticated technology that benefits multidimensional structures to supply fast access to data for analysis. This association makes it easy for a PivotTable report or PivotChart report to show high-level abstracts, such as total sales across an entire region, and also show the details for sites where sales are particularly strong.

OLAP databases contain two basic types of data: measures, which are numeric data, the quantities and averages that you use to make informed enterprise decisions, and dimensions, which are the categories that you use to coordinate with these measures. OLAP databases help to coordinate data by many levels of details by utilizing the identical categories that you are familiar with to analyse the data.

4.1.1 Components of OLAP

Figure 4.1: Components of OLAP



Source: http://www.esri.com/news/arcuser/0206/graphics/olap_1.jpg

Figure 4.1 shows the components of OLAP

Let us study about them one by one:

- *Cube:* It is a data structure that aggregates the measures by the levels and hierarchies of each of the dimensions that you want to analyse. Cubes combine some dimensions, such as time and geography, with summarized data, such as sales or inventory figures.
- *Measure:* It is a set of values in a cube that are founded on a column in the cube's detail table and that are generally numeric types. Measures are the centred values in the cube that are pre-processed, aggregated, and analysed.

Example: sales, earnings and charges

- ***Member:*** An item in a hierarchy comprising one or more occurrences of data. A member can be either unique or non-unique.
- ***Calculated Member:*** It is a member of a dimension whose worth is calculated at run time by utilizing an expression.

Example: A calculated member, Profit, can be determined by subtracting the worth of the member, charges, from the worth of the constituent, sales.

- ***Dimension:*** A set of one or more organized hierarchies of levels in a cube that a user understands and benefits as the base for data analysis.

Example: A geography dimension might include levels for Country/Region, State/Province, and town etc.

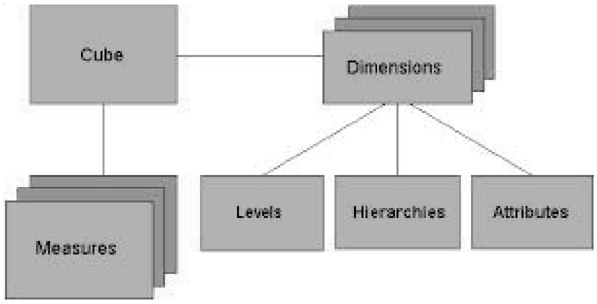
- ***Hierarchy:*** A logical tree structure that organizes the members of a dimension such that each member has one parent member and none or more child members. A child member is a member in the next lower level in a hierarchy that is exactly related to the current member.

Example: In a time hierarchy containing the grades Quarterly, Monthly, and Daily, June is a child member of Quarter2. A parent is a member in the next higher level in a hierarchy that is exactly related to the current member. For example, in a time hierarchy that contains the grades Quarterly, Monthly, and Daily, Quarter1 is the parent of January month.

- ***Level:*** Within a hierarchy, data can be organized into smaller and higher levels of detail, such as Year, Quarter, Month, Week and Day level in case of time hierarchy.

Figure 4.2 shows relationship among OLAP components.

Figure 4.2: Relationship among OLAP Components



Source: http://t3.gstatic.com/images?q=tbn:ANd9GcQmlV7tN_ufWESigi-6Txks5pftS4HO-C7o7iVrHyv4UbrgPF_PJA

4.2 Advantages of OLAP

Following are the advantages of OLAP:

(a) One major advantage of OLAP is consistency of information and computed results. No issue how much or how quick data is processed through OLAP programs or servers, the reporting result is offered in a reliable production, so analysts and executives habitually understand what to gaze for and where.

Notes

Did u know? This is especially helpful when matching data from previous reports to data present in new ones and projected future ones. It avoids the long discussion about who has the correct data.

(b) "*What if*" scenarios are some of the most well liked uses of OLAP programs and are made eminently more possible by multidimensional processing.

(c) Another advantage of multidimensional data presentation is that it allows a supervisor to drag down data from an OLAP database in very broad terms. In other words, reporting can be as easy as comparing a couple of lines of data in one column of a spreadsheet or as complex as viewing all aspects of mountain of data.

(d) OLAP is a technology that can be distributed to many users using a variety of platforms.

(e) Also, multidimensional presentation can create an understanding of connections not before realized.

(f) OLAP creates a single stage for all the information and business requirements; budgeting, forecasting, describing and analysis.

(g) Last but not least, the learning curve to use OLAP is negligible. The most utilised interface to analyse data retained in OLAP technology is the well renowned.

Self Assessment

Fill in the blanks:

1. OLAP databases contain two basic types of data i.e. measures and ....

2. .... is a data structure that aggregates the measures by the levels and hierarchies of each of the dimensions that you want to analyse.

3. .... is a set of values in a cube that are founded on a column in the cube's detail table and that are generally numeric types.

4. .... is a member of a dimension whose worth is calculated at run time by utilizing an expression.

5. .... is a logical tree structure that organizes the members of a dimension such that each member has one parent member and none or more child members.

6. .... is a technology that can be distributed to many users using a variety of platforms.

7. The .... to use OLAP is negligible.

4.3 Fast Response

Hyperion Essbase consistently delivers very quick query response times that make an iterative environment for analytic queries possible. OLAP user's queries are neither predictable nor repairable and the results of one query often frame the obligations of the next. In this natural environment, answers must be forthcoming in seconds— not minutes or hours — or analysts will cut short the analysis process to meet administration deadlines.

Caution To be productive, an analysis session should be interactive and keep pace with the analyst's speed of consideration.

Notes

With Hyperion Essbase, most users obtain responses to their queries in a fraction of a second. Even the most complex queries take only a couple of seconds. In audited OLAP benchmark results, Hyperion Essbase processed more than 6,800 complex queries per minute on a four-processor server — an average answer time of just 0.00876 seconds per query.

While query tools are becoming progressively complicated, their presentation is still limited by the answer time of the data source. It consistently provides very quick answer by permitting designers to optimize performance founded upon an application's unique obligations for query presentation, calculation complexity, assessment window (the allowance of time accessible to load and assess the application), user concurrency and computer disk utilization. Hyperion Essbase accomplishes this flexibility through three calculation choices: precalculate, calc on the go by plane and calc on the go by plane and store. Together, these three assessment schemes let designers maximize flexibility, capacity and performance.

Self Assessment

State whether the following statements are true or false:

8. OLAP user's queries are neither predictable nor repairable and the results of one query often frame the obligations of the next.

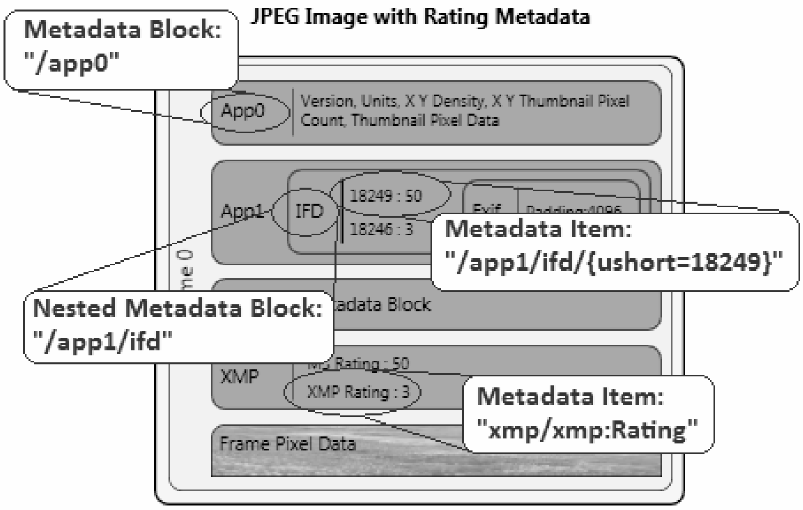
9. To be productive, an analysis session should not be interactive and keep pace with the analyst's speed of consideration.

4.4 Metadata Based Queries

A metadata block is a named group of metadata in a specific format. A metadata block can contain individual metadata pieces such as an author or creation time and additional metadata blocks. A metadata block's name is very resolute by its format.

Example: A metadata block including App1 metadata would be entitled "app1". Common metadata formats includes App1, Exif, IFD, and XMP.

Figure 4.3: JPEG image with Rating Metadata


JPEG Image with Rating Metadata
Metadata Block: "/app0"
App0
Version, Units, X Y Density, X Y Thumbnail Pixel Count, Thumbnail Pixel Data
App1
IFD
18249 : 50
18246 : 3
Exif
Padding:4096
Metadata Item: "/app1/ifd/{ushort=18249}"
me 0
Nested Metadata Block: "/app1/ifd"
adata Block
XMP
Rating : 50
XMP Rating : 3
Metadata Item: "xmp/xmp:Rating"
Frame Pixel Data

Source: http://i.msdn.microsoft.com/dynimg/IC534518.png

A metadata piece is a name/value pair that describes characteristics like author, name, and rating. A route sign contains one or more metadata block titles. It can also identify a metadata piece inside a metadata block. The following path expression comprises an App1 block which comprises an IFD block which comprises the metadata piece:

```
/app1/ifd/{ushort=18249}
```

The Figure 4.3 illustrates the makeup of an example JPEG image with four origin metadata blocks: App0, App1, XMP, and an unidentified block. Each emphasised item notes the type of metadata (block or piece) and the query expression utilised to retrieve the data.

4.4.1 Anatomy of a Path Expression

To get access to metadata, a completely qualified query expression must be used in most cases. So, what is a completely qualified query expression? A completely qualified expression is a string that begins with the path feature slash (/), pursued by a navigation route to a metadata block or a specific metadata piece. Each step within the navigation path is separated by a slash, forming an expression for accessing a metadata block or a metadata item.

Example: The following is a completely trained query expression that accesses the Microsoft photograph ranking in an IFD block that is nested in an App1 block:

```
/app1/ifd/{ushort=18249}
```

When this expression is parsed, it first explore for the App1 metadata block inside the image's metadata. If the App1 block is found, it continues it does seek looking for the nested IFD metadata block. If the IFD block is discovered, it then examines for the specific metadata piece.

Notes If at any time a metadata block or piece is not found, it aborts the query.

4.4.2 Block Selection

The simplest metadata query expression is an expression to get a query reader/writer for an exact metadata block. Getting a query reader/writer enables you to direct subsequent queries exactly to a nested metadata block without considering with its parent block. A block selection query expression is a navigation route to the yearned metadata block. For example, in the preceding example there are five metadata blocks, two of which are nested in other metadata blocks. The following are the path expressions to each metadata block in the JPEG example:

```
/app0
/app1
/app1/ifd
/app1/ifd/exif
/xmp
```

When you use a query reader/writer to execute a query, it comes back a new query reader/writer that services queries inside the scope of the particular metadata block. For instance, if you execute the query "/app1", a new query book reader is got and queries to the new book reader are relation to the App1 block. This means that the query "/ifd" is legitimate for the new reader because the App1 block contains an IFD block. However, "/xmp" would not work because this App1 block does not comprise an XMP metadata block.

For the JPEG example, the following indexed route expression can be utilised:

```
/[0]app1/[0]ifd
```

In the query language, all indexes start at zero. In the previous expression, the first zero queries for the first App1 block and the second zero queries the first nested IFD block. Index notation can still be utilised even when multiple blocks of the identical kind do not live. If the example JPEG includes a second App1 block with an embedded IFD block, the expression "*/[1]app1/ifd*" would be utilised to access the second App1 block.

The following expression accesses the Microsoft Photo ranking in the XMP block:

```
/xmp/xmp: Rating
```

The "xmp:" part of the expression is a schema identifier. XMP is an extensible benchmark and allows third party entities to publish their own schemas which indicate how to shop certain metadata items.

The following data types are accepted by the query dialect:

```
char
uchar
short
ushort
long
ulong
int
uint
longlong
ride high
twice
str
wstr
guid
bool
```

The query dialect is not case sensitive and treats all individual features as lowercase. However, some metadata formats (such as XMP) are case sensitive. When employed with a case-sensitive metadata format, use the backslash (\) feature when you want to identify an uppercase feature.

The following table supplies some example expressions and descriptions of their interpretations by the query dialect parser.

Table 4.1: Examples of Expression and Descriptions

Expression	Description
ifd/xmp/exif:Author	Corresponds to the following navigation path: IFD block -> XMP block -> "Author" property in the "Exif" schema.
/[1]ifd/[0]xmp/exif:Author	Same as the first item in this table except that the [#] prefix describes which item to navigate in event of a name collision.
/ifd/{ushort=700}/Author	Same as the first item in this table except that it uses a data expression to reference the XMP block instead of the block name "xmp" (XMP block is embedded under the unsigned short tag identifier 700). Also, the "Author" property does not specify a schema. The query parser will try to match the property across all schemas and return the first match.
/ifd/xmp	Provides a navigation path to a metadata block. If the block is found, a new metadata reader/writer is returned.
/[*]tEXt/Keyword	Gets or sets the Keyword property for a PNG chunk. Because the PNG metadata specification allows for multiple chunks of a particular type, the [*] notation gets/sets the data PNG chunk with the appropriate property. Per the PNG specification, no two chunks can have the same properties.

Source: http://msdn.microsoft.com/en-us/library/windows/desktop/ee719796(v=vs.85).aspx

Self Assessment

Fill in the blanks:

10. A metadata block is a named group of .... in a specific format.

11. A metadata piece is a .... that describes characteristics like author, name, and rating.

12. The simplest .... is an expression to get a query reader/writer for an exact metadata block.

13. In the query language, all indexes start at ....

14. The query dialect is not case sensitive and treats all individual features as ....

4.5 Spread Sheet Formulas

In OLAP we have about 5 kinds of calculations:

1. Aggregations
2. Matrix Calculations
3. Cross Dimensional Calculations
4. OLAP Aware functions
5. Procedural Calculations

Aggregations are simply addition. Adding days into weeks, weeks into months, and so on or individual customers into customer groups, families etc.

Matrix calculations are like what you would do in a spreadsheet, with arbitrarily complex relationships (+,-,*,/, sum, count, and more) both across the row and down the columns. Calculations like variance, variance %, total, count, inventory balances, etc. for example.

Cross dimensional calculations are like what you would do in linked spreadsheets, or in a multidimensional spreadsheet. In these, computed results can refer to numbers in another sheet in the cube, over distinction dimensions or different hierarchies, not just on the same spreadsheet you are on. Calculation examples include product share, market share, etc.

OLAP Aware Functions are like spreadsheet functions that have been extended to understand OLAP. These encompass statistics, forecasting purposes, economic purposes, time calculations. Just like a spreadsheet, most OLAP servers have some hundred OLAP aware formulas provided.

Example: A single OLAP aware function we could assess a benchmark deviation, a variance, an interest fee, etc. across all dimensions.

Beyond what a spreadsheet does these purposes work across dimensions and realise OLAP notions like parents, young kids, ancestors, time intelligence, etc.

Procedural calculations are the one in which specific calculation rules are defined and executed in a specific order. For example, a per cent of revenue contribution is procedural calculation. Profitability analysis is also an example of procedural calculation. It requires procedural logic to model and execute rules that reflect the business.

Self Assessment

Fill in the blanks:

15. .... are like what you would do in linked spreadsheets, or in a multidimensional spreadsheet.

16. .... are like spreadsheet functions that have been extended to understand OLAP.

4.6 Analysis Services Speed

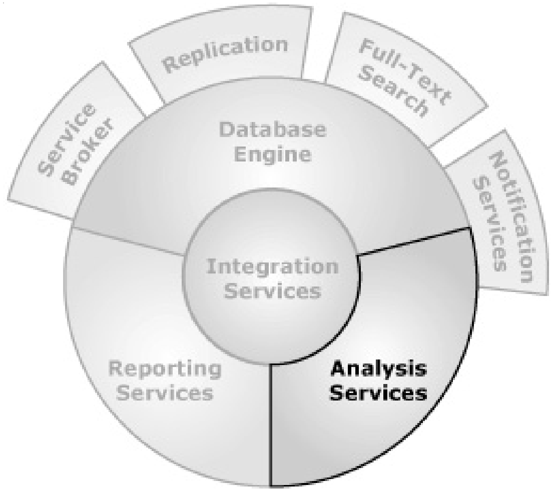
OLAP delivers the simplest form of analysis, permitting any person to slice and dice interrelated subsets of data or "cubes" with the click of a mouse. Users can investigate data using standard OLAP characteristics such as page-by, pivot, sort, filter and drill up/down to flip through a sequence of report views.

Notes OLAP analysis offers users primary access to their data warehouses in lieu of more sophisticated investigation functionality needed by power users and analysts.

Most OLAP vendors supply Multi-dimensional OLAP (MOLAP) solutions to perform this kind of analysis, but restricted cube capacity has burdened numerous IT managers by establishing and organizing hundreds of overlapping cube databases to hold stride with growing organizational claims. To get the most out of BI applications, however, OLAP investigation desires to continue beyond the benchmark MOLAP cube and supply full speed-of-thought interactivity against the whole data warehouse.

Analysis Services carries OLAP by letting you design, create, and manage multidimensional organisations that contain data aggregated from other data sources, such as relational databases. For data mining applications, Analysis Services permits you design, create, and visualize data mining models that are assembled from other data sources by using a broad kind of industry-standard data excavation algorithms. Figure 4.4 shows analysis of services concepts and objects.

Figure 4.4: Analysis of Services Concepts and Objects



Source: http://i.msdn.microsoft.com/dynimg/IC6000.gif

Notes

4.7 Metadata

The term metadata refers to "data about data". The term is ambiguous, as it is utilised for two basically different notions (types). Structural metadata is about the creation and specification of data organisations and is more correctly called "data about the containers of data"; descriptive metadata, on the other hand, is about one-by-one examples of submission data, the data content. In this case, a helpful recount would be "data about data content" or "content about content" thus metacontent.

Metadata (metacontent) are generally found in the business card catalogues of libraries. As data has become progressively digital, metadata are furthermore used to recount digital data utilising metadata measures exact to a particular control and respect. By describing the contents and context of data documents, the value of the original data/files is greatly bigger.

Example: A webpage may contain metadata identifying what language it is in writing in, what tools were utilised to create it, and where to proceed for more on the subject, permitting browsers to automatically improve the know-how of users.

Metadata (metacontent) are characterized as the data providing information about one or more aspects of data, such as:

- Means of creation of the data
- Purpose of the data
- Time and date of creation
- Creator or author of the data
- Location on a computer mesh where the data were created

Example: A digital image may include metadata that describe how large the picture is, the hue deepness, the image resolution, when the image was created, and other data.

A text document's metadata may comprise information about how long the document is, who the author is, when the article was written, and a short abstract of the article.

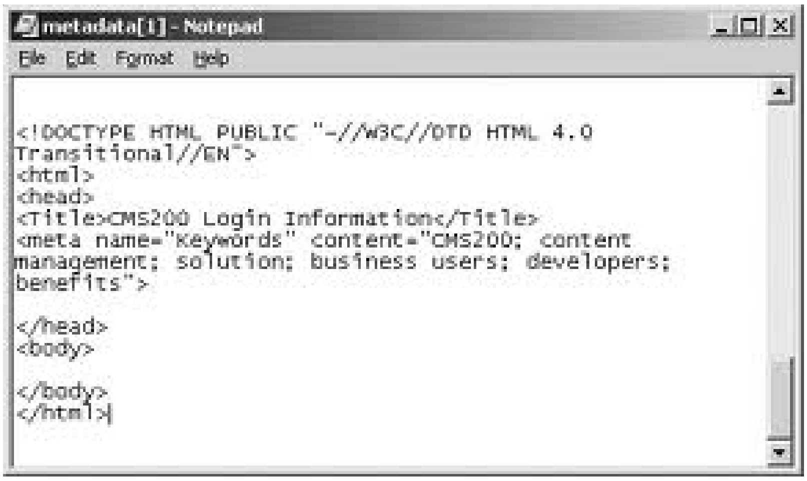
As such, metadata can be stored and organised in a database, often called a Metadata registry or Metadata repository. However, without context and a reference, it might be impossible to identify metadata just by looking at them.

Example: By itself, a database consisting of several figures, all 10 digits long could be the results of computed results or a list of numbers to close into an equation - without any other context, the figures themselves can be seen as the data.

Did u know? The term "metadata" was coined in 1968 by Philip Bagley, in his book "Extension of programming dialect notions".

Following Figure 4.5 show an example of metadata.

Figure 4.5: Example of Metadata



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0
Transitional//EN">
<html>
<head>
<Title>CMS200 Login Information</Title>
<meta name="Keywords" content="CMS200; content
management; solution; business users; developers;
benefits">

</head>
<body>

</body>
</html>
```

Source: http://t3.gstatic.com/images?q=tbn:ANd9GcR1zN_FO3Vk1jHWkuCKQAjxB3O0O41 SbnLZLnNdYmhmg4HWyLM_ZQ

4.7.1 Types of Metadata

There are three main types of metadata:

- ***Descriptive metadata*** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords. Sample elements of it are unique identifiers (PURL, Handle), physical attributes (media, dimensions condition) and bibliographic attributes (title, author/creator, language, keywords).
- ***Structural metadata*** indicates how compound objects are put together.

  Example: How pages are ordered to form chapters.

  Sample elements of it are structuring tags such as title page, table of contents, chapters, parts etc.
- ***Administrative metadata*** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

Did u know? Two further subtypes of administrative metadata:

(i) Rights management metadata, which deals with intellectual property rights, and

(ii) Preservation metadata, which contains information needed to archive and preserve a resource.

Sample elements include technical data such as scanner type and model, resolution, bit depth, colour space, file format, compression, light source, owner, copyright date etc.

Task Give examples for the use of SAS metadata DATA step functions to identify and track metadata that describes data libraries and users.

Notes

4.7.2 Metadata Functions

1. ***Discovery of resources***
   - Allowing resources to be discovered by applicable criteria.
   - Identifying assets;
   - Bringing alike assets simultaneously;
   - Distinguishing dis-similar assets;
   - Giving location information.
2. ***Organizing e-resources***
   - Organizing links to resources based on audience or theme.
   - Creating these pages dynamically from metadata retained in databases.
3. ***Facilitating interoperability***
   - Using characterized metadata designs, shared protocols, and crosswalks between designs, resources across the network can be sought more seamlessly.
   - Cross-system seek, e.g., using Z39.50 protocol;
4. ***Digital identification***
   - Components for standard number like ISBN
   - The location of a digital object may also be granted utilising:

     (i) a file name

     (ii) a URL

4.7.3 Advantage of Metadata

One major advantage of metadata is that redundancy and inconsistencies can be identified more easily as the metadata is centralized.

Example: The system catalogue and data dictionary can help or guide developers at the conceptual or structural phase or for further maintenance.

Self Assessment

State whether the following statements are true or false:

17. By describing the contents and context of data documents, the value of the original data/files is smaller.
18. Metadata (metacontent) are generally found in the business card catalogues of libraries.
19. Metadata can be stored and organised in a database, often called a Metadata registry or Metadata repository.
20. Administrative metadata provides information to help manage a resource.

Notes

Case Study

ABC, Inc.

How a Financial Services Company Developed a Performance Report for Clients, Saved $200K and Sold $167,000,000 of Equity in Just 9 Months.

This story is based on our success at ABC, Inc., the leading outsource collection agency for government debts in the US. We are using the pseudonym "ABC" to protect their confidentiality. Please allow us to recount how we applied OLAP technology to develop a flexible performance report for ABC clients, saved $200K in accounting software expenses and helped ABC sell some equity for $167,000,000 in just 9 months.

Like any other financial services company, ABC must provide regular reports that measure its performance to its clients. ABC measures its performance with what they call their CARE report. The recovery percentages that appear in the CARE report are the primary measure of performance for ABC clients. ABC first deployed their CARE report via a 40-page C program. But, the CARE report for all clients was taking over 24 hours to process and the resulting 500-page report was inflexible. There was no way to quickly focus in on a single client or client contract and there was no way to change the level of detail. There was also no way to further analyse the results, e.g. by loan type, so they could discern what portions of their business are most lucrative. And there was no convenient way to validate or understand a sum by examining the detail records that it represents. What they needed was CARE information delivered in the form of an Excel pivot table.

Merrill Eastman, ex-CEO of Bestfoods and then acting CEO of ABC, suggested that we give Online Analytical Processing (OLAP) a try. Our first assignment was to transform the old CARE report into an OLAP cube. OLAP looked like the answer because it pre-computes numeric aggregations for the cross-product of all relevant dimensions so that summary information for any combination of dimensions can be displayed on demand. If you are familiar with Excel, it suffices to say that OLAP transforms a relational database into a pivot table.

There are a number of OLAP software alternatives out there, but we quickly settled on SQL Server Analysis Services because:

1. ABC already owned Microsoft SQL Server licenses and appreciated its ease of use and administration.
2. Microsoft has bundled Analysis Services with every copy of SQL Server since 1998. So, ABC didn't have to buy anything to give it a try.
3. SQL Server Analysis Services became the OLAP market leader in 2003.
4. SQL Server OLAP Services is tightly coupled with MS Excel. Like most other companies, ABC uses Excel exclusively for all financial reports and analysis.

Developing the OLAP CARE report proceeded slowly at first because it was difficult to reach consensus on CARE Report specifications. Analysis Services is easy to use, but it was still very difficult to figure out how to get the content of the old CARE report out of an OLAP cube. The major challenges we learned to overcome included:

- How to export 80M facts and dimension rows from Informix to SQL Server in less than 4 hours?

Contd....

Notes

- How to transform exported information into a SQL Server data mart with no referential integrity errors?
- How to compute distinct counts within the cube that have a different granularity than the basic revenue facts?
- How to map the same facts to multiple members within the same dimension?
- What ragged hierarchies should be used as dimensions of the cube?
- How to support drillthrough to facts so that cube aggregates can be validated and understood?
- How to tie CARE cube aggregates to the General Ledger so that data integrity could be validated?

It took about 8 weeks to deliver the first CARE cube. A few weeks later, we delivered a sister cube that provided more comprehensive recovery analysis. By then, ABC was a believer in SQL Server OLAP Services and the rush was on to expand its use. We trained three ABC software engineers to build cubes and they set about developing General Ledger, General Ledger Budget, Payroll, Collector Performance and Revenue Forecasting cubes in parallel.

The General Ledger cubes delivered immediate benefits. ABC was using OSAS accounting software. They were not satisfied with the reports that OSAS produced, but was reluctant to invest an estimated $200K to acquire a new package and train accounting personnel to use it. Instead, they purchased an ODBC driver to export OSAS data and we built a cube to generate their reports. Today, their Balance Sheets and Profit and Loss Statements are implemented in an account rollup dimension. They can drill down from a few lines at the top to any level of detail. The drill-down feature is particularly useful in the GL Budget cube. If budget variances are detected at the highest levels, they just double-click on their OLAP pivot table to drill down until they discover the roots of the variance. The OLAP accounting reports reduced the time required to close ABC's books by 5 days. As a result, they can make critical business decisions that much faster.

Meanwhile, ABC's impressive performance attracted outside investors. A venture capital firm became the primary suitor and a team of business analysts set out to understand ABC's business. After exhaustive due diligence, the VCs decided to invest $167,000,000. They did so because ABC has a rock solid business. But, the deal might not have happened without the OLAP cubes. The OLAP cubes answered due diligence questions more quickly and in much more detail than the VC had seen in previous deals. The Billing cube that we developed at the VC's request was fundamental to their belief that future revenues would grow fast enough to support the necessary ROI.

Question:

Analyse the case and provide an alternative solution.

Source: http://www.winmetrics.com/olap_casestudies.html

4.8 Summary

- OLAP is a database expertise that has been optimized for querying and describing, rather than of processing transactions.
- OLAP user's queries are neither predictable nor repairable and the results of one query often frame the obligations of the next.

- A metadata block can contain individual metadata pieces such as an author or creation time and additional metadata blocks.
- To get access to metadata, a completely qualified query expression must be used in most cases.
- The simplest metadata query expression is an expression to get a query reader/writer for an exact metadata block.
- OLAP delivers the simplest form of analysis, permitting any person to slice and dice interrelated subsets of data or "cubes" with the click of a mouse.
- Metadata (metacontent) are generally found in the business card catalogues of libraries.
- Sample elements include technical data such as scanner type and model, resolution, bit depth, colour space, file format, compression, light source, owner, copyright date etc.
- One major advantage of metadata is that redundancy and inconsistencies can be identified more easily as the metadata is centralized.

4.9 Keywords

Calculated member: It is a member of a dimension whose worth is calculated at run time by utilizing an expression.

Cube: It is a data structure that aggregates the measures by the levels and hierarchies of each of the dimensions that you want to analyse.

Descriptive metadata: Descriptive metadata describes a resource for purposes such as discovery and identification.

Dimension: A set of one or more organized hierarchies of levels in a cube that a user understands and benefits as the base for data analysis.

Hierarchy: A logical tree structure that organizes the members of a dimension.

Level: Within a hierarchy, data can be organized into smaller and higher levels of detail.

Measure: It is a set of values in a cube that are founded on a column in the cube's detail table and that are generally numeric types.

Member: An item in a hierarchy comprising one or more occurrences of data.

Metadata block: A metadata block is a named group of metadata in a specific format.

4.10 Review Questions

1. Write down the meaning of OLAP.
2. Discuss the components of OLAP.
3. What are the advantages of OLAP?
4. "Hyperion Essbase consistently delivers very quick query response times that make an iterative environment for analytic queries possible". Elaborate.
5. What are the metadata based queries?
6. Discuss about the anatomy of a Path expression.
7. Describe the block selection.

8. Give some examples of expression and descriptions.
9. Discuss about the spread sheet formulas.
10. What is the analysis services speed?
11. Explain the concept of metadata.
12. What are the various types of metadata?
13. Discuss about the metadata functions.
14. What are the advantages of metadata?

Answers: Self Assessment

1. Averages
2. Cube
3. Measure
4. Calculated member
5. Hierarchy
6. OLAP
7. Learning curve
8. True
9. False
10. Metadata
11. Name/value pair
12. Metadata query expression
13. Zero
14. Lowercase
15. Cross dimensional calculations
16. OLAP Aware Functions
17. False
18. True
19. True
20. True

4.11 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

dublincore.org/metadata-basics/

www.guardian.co.uk › Technology › Data protection

www.newyorker.com/.../verizon-nsa-metadata-surveillance-problem.html

www.sas.com/technologies/bi/olap/

Unit 5: Microsoft Business Intelligence Platform

Notes

CONTENTS

Notes

Objectives

After studying this unit, you will be able to:

- State the Factors affecting Business Intelligence Solutions
- Discuss the Key Benefit Areas of BI
- Explain the Microsoft Business Intelligence Platform Capabilities
- Identify the Business Intelligence Platform Requirements
- Describe the Analysis Services Tools

Introduction

Microsoft Business Intelligence solutions leverage your existing technology investments in .NET, SQL Server and Office to develop rich integrated reporting and analytics experiences that empower users to gain access to accurate, up-to-date information for better, more relevant decision making. If you are seeking comprehensive, server-based reporting service designed to help you author, manage, and deliver both paper-based and interactive Web-based reports, the Microsoft Business Intelligence platform is an ideal service to provide fast, reliable reporting services. This unit provides an introduction to Business Intelligence and to Microsoft BI after discussing factors affecting BI and its key benefit areas.

5.1 Factors Affecting

The following are some key factors that determine a company's choice of BI solutions:

5.1.1 Power

A significant component working out the alternative of an analytical application is its power or functionality. A number of BI tools, such as Cognos PowerPlay, have robust analytical capabilities that allow specialized users in companies to quickly perform complicated database analysis. Usually, the client population for analysis-heavy devices comprises of more accomplished workers such as financial analysts and administration.

5.1.2 Usability

Though some mighty tools are in the market, lack of client friendliness is a barrier to adoption that has often frustrated business users and kept them from utilizing the full worth of analytic applications. If the users of an analytical submission are business staff, picking a solution that is easy to discover and use key to achieving affirmative comes back. Office XP devices such as Excel are high on the usability scale because they enable business users to support BI purposes through the use of desktop tools they are already familiar with.

5.1.3 Compatibility with Existing Technology

If companies are following best-of-breed expertise schemes, the solution must support fast integration with third-party and home-grown applications and databases. In evolving its BI platform, Microsoft has leveraged the compatibility of SQL Server with databases such as Oracle 9i and DB2. Companies can use the SQL Server 2000 for the Oracle Customer kit to link the two databases and alleviate the task of managing the two databases.

Self Assessment

Fill in the blanks:

1. The .... for analysis-heavy devices comprises of more accomplished workers such as financial analysts and administration.
2. Office XP devices such as .... are high on the usability scale.

5.2 Key Benefit Areas

This gives a detailed review of the major areas of benefits of BI.

5.2.1 Improved Technology Management

Microsoft's BI platform can positively influence expertise management by facilitating the task designing and deployment phase of a BI task as well as simplifying the task of ongoing scheme administration. Throughout the task designing phase, devices such as the SQL Server Accelerator for BI can expedite the development of BI schemes by giving developers direct access to existing best practices.

Example: The Analytics Builder Workbook is a Microsoft Excel spreadsheet that can mechanically generate a company's specific.

Specific returns resulting from Microsoft BI solutions include the following:

- Reduced development and integration costs
- Reduced consulting expenditure
- Reduced IT employee training
- Reduced infrastructure costs
- Reduced report development costs
- Reduced ongoing support costs

5.2.2 Improved Information for Decision Making

Improved Information organization and access from BI can influence the bottom line by improving some enterprise undertakings. Companies using Microsoft's BI devices can make it cheaper for users to access data, simplify the task of data mining and analysis, and enable employees to make enterprise decisions that decrease charges or improve profitability.

Some components of the Microsoft BI stage can lower the costs of accessing information by allowing users to efficiently generate queries and reports on their own.

Did u know? Excel is a client-friendly yet precious OLAP tool, enabling employees in various agencies from sales to accounting to present sales analysis on their own.

Self Assessment

Fill in the blanks:

3. Companies using .... devices can make it cheaper for users to access data.

4. Specific returns resulting from Microsoft BI solutions include the reduced development and ....

5.3 Microsoft Business Intelligence Platform Capabilities

Business Intelligence is rather a wide term, thus it wouldn't be superfluous to state what Microsoft BI platform can be utilised for:

- ***Preparing dashboards and scorecards:*** Numbers are significant, although furthermore meaningful it is to know where the numbers begin from. Dashboards and scorecards within the Microsoft solution are completely interactive, thus one can quickly dive into each area of data to watch it from another view and – thereby – find out certain thing more. Even though the data might have its source in differentiated places, it's accessible from within a single dashboard.

- ***Stimulating collaboration:*** The matter of effective collaboration is vital to all up to date associations and Microsoft Business Intelligence platform carries it very well. With well-developed connection capabilities (one can use blogs, wiki sheets, usual documents, and presentations); it allows all users across the association to work on the same "version of data" what increases their presentation. On the other hand, search engine allows users to quickly find exactly what they're looking for, no issue how high volumes of data are.

- ***Self-sufficient reporting and analysis:*** Broad reporting and analytics capabilities in Microsoft Business Intelligence platform permit users to gain the insight as deep as they need. Without any need for IT agencies involvement, they can arrange and change accounts and analysis. The working natural environment is, then, secure and intuitive, so that no specialists need to hold an eye on the system incessantly.

- ***Data mining and predictive analysis:*** The "next grade" of Business Intelligence also is supplied inside the Microsoft platform. One can thus analyse data from different points of view mechanically and find out the tendencies what – if finished without computer support – could last considerably longer or become impossible at all.

- ***Data warehousing:*** Microsoft Business Intelligence platform supplies ETL (Extract, change, burden) methods and – thereby – supports data warehousing.

Self Assessment

Fill in the blanks:

5. .... allows users to quickly find exactly what they're looking for, no issue how high volumes of data are.

6. Microsoft Business Intelligence platform supplies .... methods and thereby supports data warehousing.

5.4 Business Intelligence Platform Requirements

Business intelligence platform should include the following technologies:

- ***Data Warehouse Databases:*** A business intelligence platform should support both relational and multidimensional data warehousing databases.

Caution In supplement, storage forms should support the circulation of data across both and data forms should support clear or near-clear access to data, while it's retained.

- ***OLAP:*** OLAP is a critical business intelligence platform constituent. It is the most broadly utilized approach to analysis.

Caution Business intelligence platforms should provide OLAP support inside their databases, OLAP functionality, interfaces to OLAP functionality, and OLAP construct and manage capabilities.

- ***Data Mining:*** Data mining has reached the mainstream. It is a critical business intelligence platform capability.

Notes Platform should include data mining functionality that boasts a range of algorithms that can operate on data warehouse data.

- ***Interfaces:*** Business intelligence stages should provide open interfaces to data warehouse databases, OLAP, and data mining. Where befitting, interfaces should obey with measures. Open, standards-based interfaces make it simpler both to purchase and to construct applications that use the facilities of a business intelligence stage.

- ***Build and Manage Capabilities:*** Business intelligence stages should supply the capabilities to build and organize data warehouses in their data warehouse databases. Build capabilities should include the implementation of data warehouse forms, the extraction, action, transformation, and cleansing of data from operational sources, and the primary stacking and incremental updating of data warehouses according to their forms.

5.4.1 Uses of Microsoft BI Services

Through Microsoft's Business Intelligence (BI) services:

- Your company can build profitable connections with your customers
- Your company can decrease the risk of non-compliance and financial disasters
- You can competently change your business into a proactive conclusion making business
- You can advance workforce productivity both at the individual employee grade and all through the workforce
- You can optimize the return on current business and IT investments
- You can accomplish greater compliance with government and regulatory guidelines

- You both as an external or interior client will achieve much quicker difficulty explaining and conclusion making ability at all grades strategic, operational and tactical
- You get the right information at the right time to facilitate and expedite key decisions

5.4.2 Microsoft's Business Intelligence Platform Strategy

Microsoft's business intelligence platform strategy is rooted in its database proposing: SQL Server 2000. SQL Server 2000 is the anchor storage and query expertise behind.NET servers. And, one of the key applications for SQL Server is business intelligence. The business intelligence platform scheme for Microsoft leverages SQL Server-based technologies and goods through these components:

- Deliver a comprehensive business intelligence platform with sophisticated data warehousing techniques, large analytic functionality, and very good presentation and scalability across all platform components
- Through Microsoft's business intelligence platform:
  - impel business intelligence to the edges of the business
  - Make business intelligence more pervasive inside the corporation
  - Make business intelligence more reachable for more users and more types of users

Microsoft applies this platform strategy through the application of a well-known and well-proven Microsoft product trading equation. That equation has these key components:

- very quick implementation
- Ease of discovering and ease of use
- reduced cost and high worth
- fast return on investment (ROI)

This is the equation that Microsoft has frequently demonstrated and consistently proven. The business has used it effectively for its Windows platform (now .NET platform), its SQL Server database as utilised for OLTP applications, its e-commerce platform, business Server, and its application development suite, Visual Studio. While all of us have been trained to be sceptical (even very unlikely) about phrases like "fast implementation" and "ease of use," Microsoft has always consigned on them. And, this is not just a vision for little organizations with little allowances. Microsoft consigns data warehousing worth to companies of all dimensions.

Netting it out, Microsoft's business intelligence platform scheme endows companies of any dimensions, of any grade of business intelligence skill and know-how, of any IT budget to establish business understanding all through and to deliver and accomplish the benefits of business intelligence—improved effectiveness, greater efficiency, and higher quality throughout all their business methods.

5.4.3 Partnerships

Microsoft's business intelligence platform is built on Microsoft expertise. Microsoft controls all aspects of it creation, development, product trading, and support. The firm feels that this platform is core to its database scheme and an integral constituent of snare. Control of the whole platform from designing, design, development, and product marketing perspectives is absolutely vital in alignment to supply consistency, integration, and timely technology delivery for both Microsoft's customers and its partners. However, partnerships are critical to Microsoft and to its business intelligence platform. The firm actually values partnerships to create a large groundwork of focused business intelligence tools and applications that supports its platform. These partnerships simplify and accelerate adoption of the platform and make the platform's assets more effortlessly accessible. This is a perfect partnering approach.

> *Notes* If we compare Microsoft's approach with IBM's, in Microsoft it owns its business intelligence platform whereas in IBM OEMs its platform's OLAP expertise from Hyperion, and, as an outcome, IBM doesn't control its business intelligence platform.

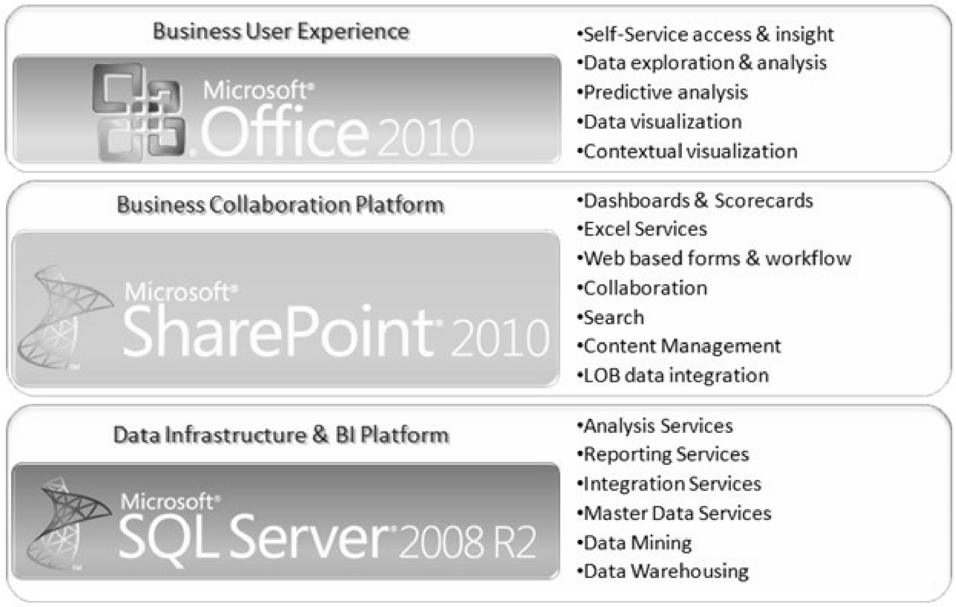
5.4.4 Microsoft's Business Intelligence Platform

Microsoft's business intelligence platform is constructed on SQL Server. SQL Server characteristics supply relational and multidimensional data warehousing, OLAP, data mining, and build and organise capabilities for relational and multidimensional data warehouses. SQL Server also provides an array of submission interfaces all constructed on the flexible and extensible object-oriented COM component model. These interfaces supply the access to all business intelligence assets with the flexibility and control to address any application requirement.

5.4.5 Packaging and Price

Packaging and pricing distance Microsoft's business intelligence platform from the platforms of Oracle, IBM, and Hyperion. For the processor-based permit charge of $19,999 per processor for SQL Server Enterprise Edition, you get the whole business intelligence platform. OLAP, data mining, and build and organise capabilities are included as database characteristics.

Figure 5.1: BI Models in Microsoft SQL Server 2008 R2



Source: http://www.microsoftbiconsultant.com/images/MS-SQL-2008-R2-BI.jpg

Oracle charges $40,000 per processor just for the Enterprise Edition of its relational data. OLAP, data mining, and build and organise capabilities are all individually cost and packaged features

Notes

of this Enterprise Edition of the firm's database. IBM allegations $25,000 per processor just for the Enterprise Server Edition of its relational database with included but basic build and manage capabilities. OLAP, data mining, and sophisticated build and manage are all additional. Hyperion allegations $28,000 per processor just for OLAP with packaged construct and manage capabilities. **Figure 5.1** depicts BI model in Microsoft SQL Server 2008 R2.

5.4.6 Oracle's Business Intelligence Platform Strategy

Oracle is a software business with two foremost lines of business: databases and applications. The present flagship proposing of the firm's database business is Oracle9i. This is an object/relational database administration scheme designed and positioned to support all types of Internet-based applications. Oracle9i integrates what Oracle terms a "complete and integrated infrastructure for building business intelligence applications." So, Oracle's business intelligence platform strategy to supply a comprehensive business understanding platform built on and integrated inside its flagship database system.

Did u know? Oracle Business Intelligence is an integrated, intuitive, and interactive business intelligence solution that provides comprehensive report creation and delivery capabilities, from data preparation to final presentation, against multidimensional OLAP or relational data sources.

Partnerships

Partnerships do not play a major function inside Oracle's business intelligence platform. The "complete and integrated infrastructure" means that every platform component is provided by Oracle and is founded on Oracle9i. Oracle's R&D organization controls the creation, development, and support of the whole platform. Like Microsoft, Oracle uses partnerships for business platform tools and applications, although the firm actually strives against with these partners with its own business intelligence tools and applications. Because Oracle's business intelligence platform does not include colleague technology, Oracle has the significant advantage of control over the platform's components, technology and integration.

Oracle's Business Intelligence Platform

Oracle9i is the foundation of Oracle's business understanding platform. OLAP functionality is supplied by Oracle9i OLAP and data mining functionality is supplied by Oracle9i data mining. Both are characteristics of Oracle9i Business Edition. Build and Manage functionality is provided by two toolsets. The first, Oracle Business supervisor, is the major management structure and DBA toolset as well as the toolset for OLAP construction and organization. The second is Oracle9i Warehouse Builder. This toolset, a component of Oracle Internet Developer Suite, presents capabilities for managing relational data warehousing resources, conceiving relational data warehouse forms, and ETL.

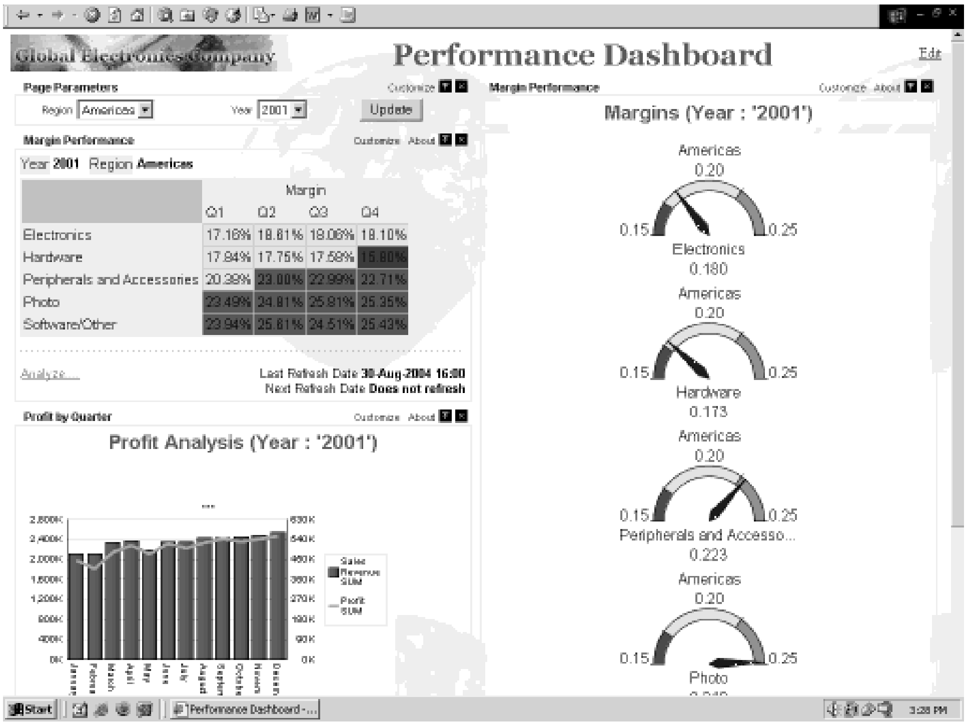
Package and Pricing

From a packaging perspective, Oracle boasts little bundling. All the constituents of its business intelligence platform are individually bundled and cost and the construct and manage components have separately cost and bundled sub-components. Oracle9i Business version is cost at $40,000 per processor. Oracle9i OLAP is cost at $20,000 per processor. Oracle9i data Mining is cost at $20,000 per processor. And, Oracle Warehouse Builder is priced at $5,000 per entitled client. Add

Notes

them up and Oracle's business intelligence platform is at smallest five times higher in price than Microsoft' business understanding platform.

Figure 5.2: Oracle Business Intelligence Discoverer Dashboard Example



Year 2001 Region Americas

	Margin			
	Q1	Q2	Q3	Q4
Electronics	17.16%	18.81%	18.06%	18.10%
Hardware	17.84%	17.75%	17.58%	15.80%
Peripherals and Accessories	20.39%	23.00%	22.99%	22.71%
Photo	23.49%	24.81%	25.81%	25.35%
Software/Other	23.94%	25.81%	24.51%	25.43%

Analyze...

Last Refresh Date **30-Aug-2004 16:00**
Next Refresh Date **Does not refresh**

Source: http://docs.oracle.com/cd/B14099_19/core.1012/b13994/img/dashboard.gif

IBM's Business Intelligence Platform Strategy

From a business perspective, IBM has three businesses: hardware, software, and consulting services. The programs business has four constituents: WebSphere programs, DB2 data administration programs, Lotus (collaboration) programs, and Tivoli (system administration) programs.

Notes Business intelligence is one of two IBM-provided solutions of DB2 data management programs. (The other solution is e-business.)

IBM characterizes business intelligence as "warehousing, data mining, and OLAP." That's precisely our delineation of a business understanding platform. So, IBM's business intelligence answer is a business intelligence platform.

IBM's strategy for business intelligence is to help companies know their customers and to use that information to gain comparable advantages, to maximize revenue, and minimize cost. Business intelligence is implicitly targeted at all of IBM's markets. The firm makes no explicit distinction in the positioning of business understanding for the types or dimensions of businesses or for the kinds of users inside those businesses that can use its business understanding platform. It's a one dimensions aligns all approach.

Notes

Hyperion's Business Intelligence Platform Scheme

In mid-2001, Hyperion changed its business strategy, shifting its focus from business intelligence software infrastructure and applications to business presentation management programs answers. The firm states its target "is to be the premier global provider of business presentation management solutions." These solutions are created to automate the business performance administration method of scheme setting, modelling, planning, performance supervising, describing and investigation. Their target is to improve your profitability.

The expertise platform for Hyperion's performance administration solutions is Essbase, its venerable OLAP Server Within the new strategy, Hyperion states that Essbase technology will be enhanced in the areas of ease of use, ease of application development, interoperability of business performance management applications, scalability, and tighter integration with relational data sources. Missing from Hyperion's enhancement strategy for Essbase are localities such as analytic technology and business intelligence platform technologies. Essbase is evolving away from a general reason OLAP facility and in the direction of a platform for carrying a very exact type of business understanding application.

Hyperion's new scheme and target can play quite well in today's business understanding market where companies' top main concerns are to do business more competently and efficiently. Business performance administration is a classic business intelligence submission. It requires a comprehensive business intelligence platform as its base in order to assemble the data that comprises business presentation, organize that data, investigate it, present it, and use investigation outcomes to advance business presentation.

Packaging and Pricing

Hyperion Essbase has a pricing model for with two elements: a per server charge and a per entitled client charge. Currently, per server charge is $28,000 per processor and per entitled client charge is $1,500. Essbase packaging includes the OLAP server, administrative tools, and construct and organise tools.

⚠
Caution Essbase installations most commonly use relational data warehouses as the data causes for Essbase cubes.

Comparing Business Intelligence Databases

The following table enumerates BI Databases of different companies:

Table 5.1: Different BI Databases

	Interfaces			
	Microsoft	**Oracle**	**IBM**	**Hyperion**
Relational interfaces	SQL and Transact/SQL ODBC and JDBC OLE DB ADO ADO.NET	SQL and PL/SQL ODBC and JDBC	SQL and DB2 SQL ODBC and JDBC	Not applicable
OLAP Interfaces	MDX DSO Pivot Table Service XML for Analysis	OLAP DML Java OLAP API SQL and PL/SQL	Essbase API	Essbase API

Contd....

Notes

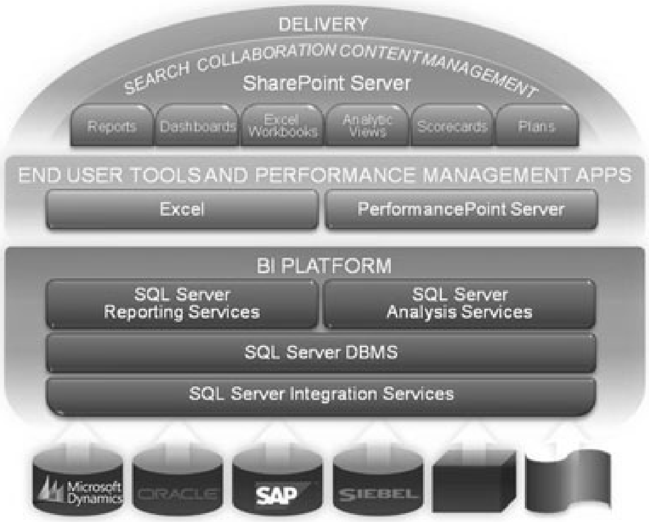
Data mining interfaces	DSO Pivot Table Service Wizards	Oracle9i Data Mining API (Java)	Intelligent Miner • C++ • SQL • Visual tools DB2 OLAP Miner • Essbase API	Not applicable

Source: A Comparison of Business Intelligence Strategies and Platforms by Mitch Kramer.

Microsoft's BI/PM Offering

The following figure illustrates various BI/PM offerings by Microsoft:

Figure 5.3: Various offerings by Microsoft


DELIVERY
SEARCH COLLABORATION CONTENT MANAGEMENT
SharePoint Server
Reports
Dashboards
Excel Workbooks
Analytic Views
Scorecards
Plans
END USER TOOLS AND PERFORMANCE MANAGEMENT APPS
Excel
PerformancePoint Server
BI PLATFORM
SQL Server Reporting Services
SQL Server Analysis Services
SQL Server DBMS
SQL Server Integration Services
Microsoft Dynamics
ORACLE
SAP
SIEBEL

Source: http://www.element61.be/assets/microsoft-business-intelligence-&-performance-management-platform_small.jpg

Self Assessment

Fill in the blanks:

7. A .... platform should support both relational and multidimensional data warehousing databases.

8. Business intelligence stages should provide open interfaces to data warehouse databases, OLAP, and ....

9. Business intelligence stages should supply the capabilities to build and organize data warehouses in their ....

10. SQL Server 2000 is the anchor storage and query expertise behind ....

11. The firm actually values .... to create a large groundwork of focused business intelligence tools and applications that supports its platform.

12. Packaging and pricing distance Microsoft's business intelligence platform from the platforms of Oracle, IBM, and ....

13. .... is a software business with two foremost lines of business: databases and applications.

5.5 Analysis Services Tools

Business Intelligence Development Studio is Microsoft Visual Studio 2008 with project types that are specific to SQL Server business intelligence. Business Intelligence Development Studio is the primary natural environment that you will use to evolve enterprise solutions that includes analysis Services, Integration Services, and Reporting Services tasks. Each project type supplies templates for creating the objects needed for business intelligence answers, and supplies a variety of designers, tools, and wizards to work with the objects.

Features in Business Intelligence Development Studio:

- Start Page
- Tool Windows in Business Intelligence Development Studio
- Menus in Business Intelligence Development Studio
- Toolbars in Business Intelligence Development Studio
- Working with Projects and Solutions
- Customizing Environment, Tools, and Windows
- Using Source Control Services
- Getting More Information
- Analysis Services in Business Intelligence Development Studio
- Integration Services in Business Intelligence Development Studio
- Reporting Services in Business Intelligence Development Studio

5.5.1 Start Page

When you first open Business Intelligence Development Studio, the Start page appears in the centre of the Business Intelligence Development Studio user interface. This page exhibits a register of lately revised projects; help topics, World Wide Web sites and other resources; links to product and updated information from Microsoft; and by default, a register of items from the RSS feed of the specified report.

To display a page other than the Start Page at startup, click Options on the tools menu, expand the Environment node, and in the At Startup menu, select the item to display.

Notes To discover more about the Start page, click within the Start Page and press F1.

5.5.2 Tool Windows in Business Intelligence Development Studio

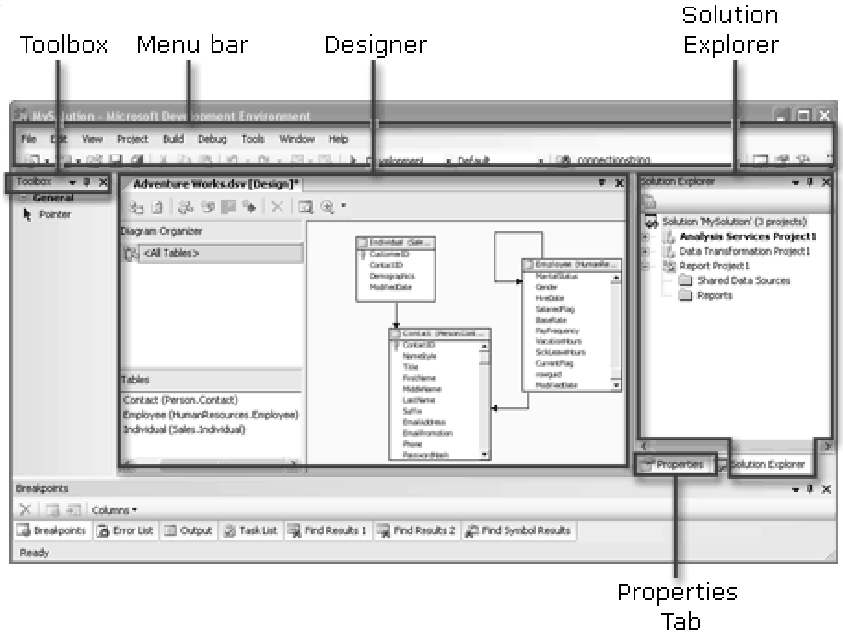
Business Intelligence Development Studio includes a set of windows for all stages of solution development and project administration.

Example: Business Intelligence Development Studio includes windows that let you organise multiple tasks as a unit and view and change the properties of objects in projects. These windows are accessible to all the project types in Business Intelligence Development Studio.

Notes

The Figure 5.4 displays the windows in Business Intelligence Development Studio with the default configuration.

Figure 5.4: Business Intelligence Development Studio with Default Configuration



Source: http://i.msdn.microsoft.com/dynimg/IC146944.gif

Business Intelligence Development Studio comprises of four major windows:

- Solution Explorer
- Properties Window
- Designer Window
- Toolbox Window

Solution Explorer

You can manage all the different tasks in a solution from a single window, Solution Explorer. The Solution Explorer view presents the active solution as a logical container for one or more projects, and includes all the pieces associated with the projects. You can open project pieces for modification and present other administration jobs directly from this view.

In Solution Explorer, you can create empty solutions and then add new or existing tasks to the solution. If you create a new project without first creating a solution, Business Intelligence Development Studio automatically creates the solution too. When the solution includes tasks, the tree view includes nodes for project-specific objects. For example, the analysis Services project includes a Dimensions node, the Integration Services project includes a Packages node, and the Report form project includes a reports node.

Notes To access solution Explorer, click solution Explorer on the View menu.

Notes

Properties Window

The Properties Window menus are the properties of an object. You use this window to view and change the properties of objects, such as packages, that are open in editors and designers. You can furthermore use the Properties window to edit and view document, task, and solution properties.

Notes To access the Properties window, click Properties Window on the view menu.

Designer Window

The Designer window is the tool window in which you create or modify business intelligence objects. The designer provides both a code view and a design view of an object. When you open an object in a project, the object opens inside a specialized designer in this window.

Notes The Designer window is not available until you add a project to a solution and open an object inside that project.

Toolbox Window

The Toolbox window displays a variety of items for use in business intelligence tasks. The tabs and items available in the Toolbox change depending on the designer or editor currently in use. The Toolbox window habitually displays the General tab, and may furthermore display tabs such as Control Flow Items, Maintenance Tasks, Data Flow Sources, or Report Items.

Notes To access the Toolbox, click Toolbox on the view menu.

5.5.3 Menus in Business Intelligence Development Studio

The default menus that emerge in Business Intelligence Development Studio are equal to those in Visual Studio. When you first open Business Intelligence Development Studio, before you change the environment, open a solution, or open any projects, Business Intelligence Development Studio includes the following menus:

- File
- Edit
- View
- Tools
- Window
- Community
- Help

File Menu

The choices on the file menu support file management. When you first open Business Intelligence Development Studio, but before you have created a new project or opened an existing project, some choices are unavailable. These choices become available only when you start to work in the context of a solution, or open a project within a solution.

Edit Menu

The choices on the Edit menu support editing of text and code in documents. This menu supplies commands such as undo and redo; find and replace; enable and manage bookmarks. When you first open Business Intelligence Development Studio, before you have created a new project or opened an existing project, some choices are unavailable. Depending on the project type, some menu options many not be available.

Example: The Undo and Redo options are not supported in Integration Services projects.

View Menu

The choices on the view menu help you manage the client interface of Business Intelligence Development Studio. This menu and its submenus supply the choices to open the diverse windows, toolbox, explorers, and browsers.

Tools Menu

The choices on the Tools menu customize behaviour of the development environment. This menu, its submenus, and the dialog boxes it accesses provide choices to set the following options:

- Select process and a code type for debugging
- Connect to a database. The Database Explorer lists the data connections
- Work with macros
- Select external tools to include in the environment
- Import and export specified environment settings or reset environment settings to their defaults
- Choose the toolbars to display in the user interface and arrange the order of commands
- Set the options that apply to the overall development environment, solutions and projects, source control, debugging, and designers and editors

Window Menu

The choices on the Window menu organise the behaviour of windows, explorers, and browsers in Business Intelligence Development Studio.

Community Menu

The choices on the Community menu lets you ask questions of other users and of technical support, send response to Microsoft, access groups and connect to the developer centre.

Help Menu

The options on the Help menu provide access to Help topics.

5.5.4 Toolbars in Business Intelligence Development Studio

When you first open Business Intelligence Development Studio the Toolbar includes only the Menu Bar toolbar and only a couple of icons that are accessible on the Menu Bar toolbar. To customize the Toolbar, click Customize on the Tools menu, and then select additional toolbars to display, or change options for the toolbar appearance.

5.5.5 Working with Solutions and Projects

In Business Intelligence Development Studio, a solution is a container that organizes the diverse projects that you use when you evolve end-to-end business solutions. A solution permits you handle multiple tasks as one unit and blend one or more associated projects that contribute to a business solution.

When you create a new solution, Business Intelligence Development Studio adds a solution folder to Solution Explorer and creates documents that have the additions .sln and .suo.

- The *.sln document comprises data about solution configuration and registers the tasks in the solution.
- The *.suo file comprises information about your preferences for employed with the solution.

Projects are stored in solutions. You can create a solution first and then add projects to the solution. If no solution exists, Business Intelligence Development Studio technically creates one for you when you first create the project. A solution can contain multiple projects of distinct types. You can furthermore create a new solution and then add subsequent projects.

In Business Intelligence Development Studio you can add projects of the following types:

- Analysis Services projects, for creating analytic objects
- Integration Services projects, for creating ETL packages
- Report Model projects, for creating report models
- Report Server projects, for creating reports

5.5.6 Customizing the Environment, Tools and Windows

Business Intelligence Development Studio can be easily configured to match your employed method. You can configure the general development environment and its behaviour, and make alterations to its tools and windows. When you save a solution, your configurations are kept to a *.suo file in the solution folder.

You can configure the Business Intelligence Development Studio environment with collection of backgrounds customized for SQL Server business intelligence development by choosing the Business Intelligence backgrounds collection. Use import and Export settings on the tools menu to reset all your backgrounds based on the Business Intelligence Settings collection or to trade only the categories of Business Intelligence backgrounds that you choose.

To configure one-by-one choices for the environment and tools, click Options on the tools menu to open choices dialog box. To discover more about the different choices in the dialog carton, click a node in the left pane, and then press F1.

Using Source Control Services

Like Visual Studio, Business Intelligence Development Studio is integrated with source control programs. If source control software is established on the computer, you can add solutions and projects to source control, and then open the solutions and tasks in Business Intelligence Development Studio from the source control application.

Getting More Information

The Visual Studio documentation provides detailed information about the Microsoft application development environment. It help collection provides documentation for the user interface that Business Intelligence Development Studio uses. To access the Visual Studio help topics that are relevant to the user interface shared with Business Intelligence Development Studio, you must install the MSDN Library that is included with SQL Server or configure the Business Intelligence Development Studio help options to access Help. When online Help is enabled, you can obtain context-sensitive help from the Visual Studio windows by pressing F1 or clicking Help.

Analysis Services in Business Intelligence Development Studio

Business Intelligence Development Studio includes the Analysis Services project for developing Online Analytical Processing (OLAP) for business intelligence applications. It includes the templates for cubes, dimensions, mining structures, data sources, data source views etc.

Integration Services in Business Intelligence Development Studio

Business Intelligence Development Studio includes the Integration Services project for developing ETL solutions. It includes the templates for packages, data sources, and data source views, and provides the tools for working with these objects.

Reporting Services in Business Intelligence Development Studio

Business Intelligence Development Studio includes the Report Model and Report projects for developing reporting solutions. The Report Model project type includes the templates for report models; data sources etc. and provide the tools for working with these objects.

Self Assessment

Fill in the blanks:

14. The choices on the .... menu support editing of text and code in documents.

15. The choices on the .... menu help you manage the client interface of Business Intelligence Development Studio.

16. The choices on the .... menu customize behaviour of the development environment.

17. The choices on the .... menu organise the behaviour of windows, explorers, and browsers in Business Intelligence Development Studio.

5.6 Data Extraction, Transformation and Load

In computing, Extract, Transform and Load (ETL) refer to a process that involves:

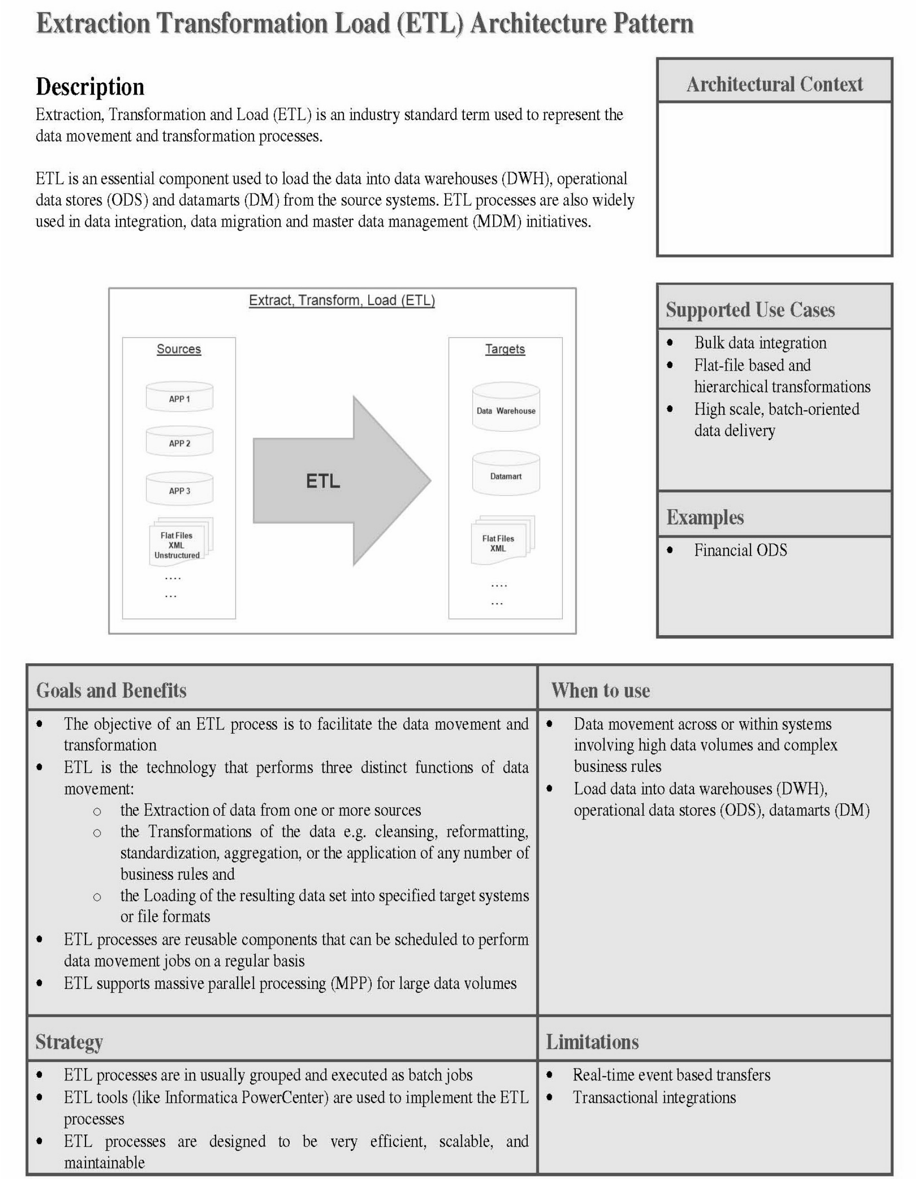
- Extracting data from outside sources

Notes

- Transforming it to fit operational needs, which can include quality levels
- Loading it into the end target (database, more specifically, operational data store, data mart or data warehouse)

Figure 5.5 shows a complete description of ETL Architecture Pattern.

Figure 5.5: ETL Architecture Pattern



Extraction Transformation Load (ETL) Architecture Pattern

Description

Extraction, Transformation and Load (ETL) is an industry standard term used to represent the data movement and transformation processes.

ETL is an essential component used to load the data into data warehouses (DWH), operational data stores (ODS) and datamarts (DM) from the source systems. ETL processes are also widely used in data integration, data migration and master data management (MDM) initiatives.

Architectural Context

Supported Use Cases

- Bulk data integration
- Flat-file based and hierarchical transformations
- High scale, batch-oriented data delivery

Examples

- Financial ODS

Goals and Benefits

- The objective of an ETL process is to facilitate the data movement and transformation
- ETL is the technology that performs three distinct functions of data movement:
  - the Extraction of data from one or more sources
  - the Transformations of the data e.g. cleansing, reformatting, standardization, aggregation, or the application of any number of business rules and
  - the Loading of the resulting data set into specified target systems or file formats
- ETL processes are reusable components that can be scheduled to perform data movement jobs on a regular basis
- ETL supports massive parallel processing (MPP) for large data volumes

When to use

- Data movement across or within systems involving high data volumes and complex business rules
- Load data into data warehouses (DWH), operational data stores (ODS), datamarts (DM)

Strategy

- ETL processes are in usually grouped and executed as batch jobs
- ETL tools (like Informatica PowerCenter) are used to implement the ETL processes
- ETL processes are designed to be very efficient, scalable, and maintainable

Limitations

- Real-time event based transfers
- Transactional integrations

Source: http://upload.wikimedia.org/wikipedia/commons/d/d8/ETL_Architecture_Pattern.jpg

Extract

The first stage of an ETL process involves extracting the data from the source systems. An intrinsic part of the extraction involves the parsing of extracted data, resulting in a check if the data meets an expected pattern or structure. If not, the data may be rejected entirely or in part.

Notes

Transform

The transformation stage applies a series of rules or functions to the extracted data to derive the data for loading into the end target. Some data sources will require very little or even no manipulation of data.

Load

The loading phase loads the data into the end target, usually the Data Warehouse (DW). Some of the data warehouses may overwrite existing information frequently; updating data is done on a daily, weekly, or monthly basis.

5.7 Meaning and Tools of Extraction

Data extraction is the act or process of retrieving data from data sources for further data processing. Usually, the term data extraction is applied when data is first imported into a computer from primary sources like measuring or recording devices.

5.7.1 Tools

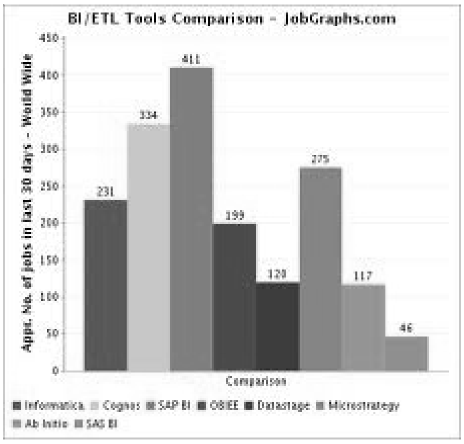
A good ETL tool must be able to communicate with different relational databases and read the various file formats used throughout an organization. Many ETL vendors now have data profiling, data quality, and metadata capabilities. A common use case for ETL tools include converting CSV files to formats readable by relational databases.

Popular ETL Tools are:

- IBM WebSphere Information Integration
- Ab Initio
- Informatica
- Talend

Figure 5.6 depicts an ETL Tools Comparison chart.

Figure 5.6 ETL Tools Comparison



Source: https://sheet.zoho.com/publicgraphs/983047000000039763.png

Notes

ETL Tools are typically used by a broad range of professionals - from students to database architects.

Task "Microsoft enables business users to look no further than Excel for self-service BI". Comment.

Self Assessment

State whether the following statements are true or false:

18. The first stage of an ETL process is transformation stage which involves extracting the data from the source systems.

19. The loading stage applies a series of rules or functions to the extracted data to derive the data for loading into the end target.

20. A common use case for ETL tools include converting CSV files to formats readable by relational databases.

Case Study

Wireless Data Services Provider - Business Intelligence Platform & Reporting

Background

As a growing provider of wireless broadband services, the road to success in this competitive industry is clear: expand coverage as quickly and as efficiently as possible while maintaining the highest level of service quality. In an effort to control rising costs and enable strategic decision making, the company required a method for assembling, analysing, and disseminating network deployment data with maximum visibility and accuracy. During the ongoing network expansion process, numerous sites were quickly acquired, zoned, and permitted simultaneously by the deployment team. With only a small time window to complete work, even the most minor setback could lead to massive cost overruns. A lack of site status visibility was causing problem sites to incur exponentially higher costs due to re-work and redundant permitting. To maintain the aggressive schedule and avoid run-away costs, the network development team needed a system to quickly identify site progress against build-out goals. Concurrently, the market launch team struggled to communicate with the network deployment team. Without accurate market status information, they were unable to make informed decisions and effectively forecast the demand for customer service resources. This lack of communication posed the risk of causing market launch delays and the loss of significant future revenue. Additionally, dependable market coverage information was needed to measure adherence to regulatory requirements.

Giving Visibility to the Network Build Out

The shortage of actionable information was not due to a lack of data. An internally developed system tracked the network build-out process and compiled raw transactional records. Analysts, however, were tasked with manually cleaning and analysing massive "data dumps" from the system. Due to the sheer amount of data, only a small subset of the

Contd....

markets could be realistically analysed and reports were often out of date by completion. Further problems were introduced by discrepancies in the data definitions. To make faster and more accurate decisions, managers needed a way to quickly obtain current and comprehensive reports with maximum accuracy.

Looking forward, the company anticipated expanding their business intelligence capabilities beyond network deployment to optimize business decisions in other units. These new capabilities were expected to yield increasing amounts of data that would require an expansion of storage capacity. A framework was sought for designing and defining future data warehousing and business intelligence implementations.

Creating a Platform to Support Growth and Control Costs

Originally selected to perform an assessment of the organization's overall business intelligence capabilities, Hitachi Consulting was subsequently engaged to provide a comprehensive solution to this pressing issue. By utilizing expertise in BI and in collaboration with the IT, finance, marketing, and network development teams, Hitachi Consulting designed a platform to bring visibility to the company's network deployment efforts. The implementation of a Cognos dashboard, built on an Oracle relational database system, would be linked to a data feed from the existing network monitoring system. Organized by phase and by market, the new dashboards would offer decision makers the option to obtain a high level view or to drill down into the specifics of each. Without the constraints of limited data or the prohibitive costs of manual analysis, reports could be infinitely configured and generated nightly. Key metrics, such as population counts, sites on air, sites leased, and duration calculations, would be integrated while data definitions would be consolidated for increased accuracy.

The future needs of the organization were also considered as the system was designed to serve as the foundation for the development of a more robust and scalable data warehousing platform. Hitachi Consulting helped define the data warehouse framework and design standards by assisting the client in developing an internal "Centre of Excellence" for future implementations.

Cost Savings is realized with Improved Operational Efficiencies

With an aggressive and costly objective of doubling service coverage in less than a year, this young company worked closely with Hitachi Consulting to help them achieve it. Through our business intelligence expertise and implementation, the company realized immediate gains in operational efficiency.

The company is now better positioned to minimize development costs through more accurate financial insights. Reports that formerly required many hours of preparation are near-instantly generated with greater relevance, scope, and accuracy. Managers, in turn, gain increased visibility of schedule slips which has enabled them to prevent them as they occur.

The system is expected to pay for itself through the cost savings of only a single market. Data that was formerly available to only a few is now empowering those making the decisions to manage a network build-out that is on time and on budget.

Questions:

1. Analyse a case find out what is the problem?
2. What solution does wireless broadband services provider provides to Hitachi Consulting?

Source: www.hitachiconsulting.com/files/.../CS_WirelessProviderBI.pdf

Notes

5.8 Summary

- A web-based report, the Microsoft Business Intelligence platform is an ideal service to provide fast, reliable reporting services.
- A number of BI tools, such as Cognos PowerPlay, have robust analytical capabilities that allow specialized users in companies to quickly perform complicated database analysis.
- Office XP devices such as Excel are high on the usability scale because they enable business users to support BI purposes through the use of desktop tools they are already familiar with.
- Companies can use the SQL Server 2000 for the Oracle Customer kit to link the two databases and alleviate the task of managing the two databases.
- Microsoft's BI platform can positively influence expertise management by facilitating the task designing and deployment phase of a BI task as well as simplifying the task of ongoing scheme administration.
- Companies using Microsoft's BI devices can make it cheaper for users to access data, simplify the task of data mining and analysis, and enable employees to make enterprise decisions that decrease charges or improve profitability.
- Microsoft's business intelligence platform strategy is rooted in its database proposing: SQL Server 2000. SQL Server 2000 is the anchor storage and query expertise behind.NET servers.
- Business Intelligence Development Studio is Microsoft Visual Studio 2008 with project types that are specific to SQL Server business intelligence.
- Data extraction is the act or process of retrieving data from data sources for further data processing.

5.9 Keywords

Business Intelligence (BI): Business Intelligence (BI) is a set of theories, methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information for business purposes.

Data Extraction: Data extraction is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration).

Data Mining (DM): Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

Data Warehousing: In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis.

Decision Support Systems (DSS): A decision support system (DSS) is a computer program application that analyses business data and presents it so that users can make business decisions more easily.

5.10 Review Questions

1. What are the key factors that determine a company's choice of BI solutions?
2. What is the Microsoft Business Intelligence Platform?
3. What are the key benefit areas of Microsoft business intelligence?
4. Briefly explain the capabilities of Microsoft business intelligence.
5. What are the requirements in Business intelligence platform?
6. What are the uses of Microsoft BI Services?
7. Write short note on Microsoft's business intelligence platform strategy.
8. Discuss the IBM's business intelligence platform strategy.
9. What is the Hyperion's Business Intelligence Platform Scheme?
10. Write down the features in Business intelligence development studio.
11. Provide details for menus in business intelligence development studio.
12. Discuss the extract transform load.

Answers: Self Assessment

1. Client population
2. Excel
3. Microsoft's BI
4. Integration costs
5. Search engine
6. ETL (Extract, change, burden)
7. Business intelligence
8. Data mining
9. data warehouse databases
10. . NET servers
11. Partnerships
12. Hyperion
13. Oracle
14. Edit
15. View
16. Tools
17. Window
18. False
19. False
20. True

5.11 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Notes

Online links

http://searchcio.techtarget.com/definition/decision-support-system

nfotrove.com/files/GreenHill.pdf?

www.kpmg.com/GR/en/IssuesAndInsights/.../The-MS-Platform-for-BI.pdf

www.microsoft.com/en-in/bi/?

Unit 6: Business Intelligence Project

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Creating Data Source
- Explain Creating the Data Source View
- State the Modifying of the Data View
- Summarize Creating Dimensions, Time and Modifying Dimensions
- Define the Parent-child Dimension

Introduction

Too many times, Business Intelligence (BI) and Data Warehousing project managers are not equipped well to handle their role in guiding a project to success. Often, the person allotted to lead a project is either: (1) a technician who doesn't know the first thing about managing a project, or (2) a project manager who doesn't know the first thing about Business Intelligence. Purpose of BI is making the most of an organization's data assets. By making better data-driven decisions through BI, companies can gain advantages like increasing revenue, reducing costs, or reducing risks. This Unit focus on discussion of data source and data source view. Also, working with dimension is explained in the unit. Finally, you will learn about parent-child dimension.

6.1 Creating Data Source

In an Analysis Services multidimensional model, a data source object represents a connection to the data source from which you are importing data. A multidimensional model must contain at least one data source object, but you can add more to combine data from several data warehouses.

Notes

Choose a Data Provider

You can connect utilizing an organized Microsoft .NET Framework or native OLE DB provider. For Oracle and other third-party data causes, check if the third-party provides a native OLE DB provider and if it does try that first. If you get errors, try one of the other .NET providers or native OLE DB providers listed in connection manager. Be certain that any data provider you use is established on all computers utilized to develop and run the analysis Services solutions.

Set Credentials and Impersonation Options

A data source connection can occasionally use Windows authentication or an authentication service provided by the database administration scheme, such as SQL Server authentication when connecting to SQL Azure databases.

Caution The account you identify should have a login on the isolated database server and read permissions on the external database.

Windows Authentication

Connections that use Windows authentication are specified on the Impersonation data tab of the Data Source Designer. Use this tab to choose the impersonation option that identifies the account under which analysis Services runs when connecting to the external data source. Not all options can be utilised in all scenarios.

Database Authentication

As an alternate to Windows authentication, you can specify a connection that uses an authentication service provided by the database administration scheme. In some situations, utilising database authentication is required. Scenarios that call for utilising database authentication includes SQL Server authentication to connect to a Windows Azure SQL Database, or accessing a relational data source that runs on a different operating system.

For a data source that uses database authentication, the username and password of a database login is particular on the attachment string. Credentials are supplemented to the attachment string when you enter a client name and password in connection manager when setting up the data source connection in your Analysis Services model. Remember to identify a client identity that has read permissions to the data.

When retrieving data, the client library making the connection formulates a connection request that includes the credentials in the connection string.

Notes By default, SQL Server Data Tools (SSDT) does not save passwords with the connection string. If the password is not saved, Analysis Services prompts you to enter the password when it is needed.

Create a Data Source Using the Data Source Wizard

To create data source using the data source wizard follow these steps:

1. In SQL Server Data Tools, open the Analysis Services project or connect to the Analysis Services database in which you want to define the data source.

2. In Solution Explorer, right-click the Data Sources folder, and then click New Data Source to start the Data Source Wizard.

3. On the Select how to define the connection page, choose Create a data source based on an existing or new connection and then click New to open Connection Manager. New connections are created in Connection Manager.

4. Select the Microsoft .NET Framework or native OLE DB provider to use for the connection.

5. Enter the information requested by the selected provider to connect to the data source. If the Native OLE DB\SQL Server Native Client provider is selected for example, then enter the following information:

   (a) Server Name is the network name of the Database Engine instance.

   (b) Log on to the Server specifies how the connection will be authentication. Use Windows Authentication uses Windows authentication. Use SQL Server Authentication specifies a database user login for a Windows Azure SQL databases.

   (c) Select or enter a database name or Attach a database file are used to specify the database.

   (d) In the left side of the dialog box, click All to view additional settings for this connection, including all default settings for this provider.

   (e) Change settings as appropriate for your environment and then click OK. The new connection appears in the Data Connection pane of the Select how to define the connection page of the Data Source Wizard.

⚠

Caution Regardless of whether you clear or select Save my password, Analysis Services will always encrypt and save the password. The password is encrypted and stored in both .abf and data files. This behaviour exists because Analysis Services does not support session-based password storage on the server.

6. Click Next. In Impersonation Information, specify the Windows credentials or user identity that Analysis Services will use when connecting to the external data source.

7. Click Next. In Completing the Wizard, enter a data source name or use the default name. The default name is the name of the database specified in the connection. The Preview pane displays the connection string for this new data source.

8. Click Finish. The new data source appears in the Data Sources folder in Solution Explorer.

View or Edit Connection Properties

The attachment string is formulated based on the properties you choose in the Data Source Designer or the New Data Source Wizard. You can view the attachment string and other properties in SQL Server Data Tools.

To edit the connection string

1. In SQL Server Data Tools, double-click the data source object in Solution Explorer.

2. Click **Edit**, and then click **All** on the left navigation pane.

3. The property grid appears, showing available properties of the data provider you are using.

Did u know? A *data source reference* is an association to another Analysis Services project or data source in the same solution. You can remove the reference by clearing the check box which will end the synchronization between the objects.

Add Multiple Data Sources to a Model

You can create more than one data source object to support connections to additional data sources. To combine data from multiple data sources:

1. Create the data sources in your model.
2. Create a data source view, using a SQL Server relational database as the data source.
3. In Data Source View Designer, using the data source view just created right-click anywhere in the work area and select Add/Remove Tables.
4. Choose the second data source and then select the tables you want to add.
5. Find and select the table you added. Right-click the table and select New Relationship. Choose the source and destination columns that contain matching data.

Supported Data Source Types

Types of data sources that you can use in a multidimensional model are shown in **Table 6.1**.

Table 6.1: Types of Data Sources

Source	Versions	File type	Providers
Access databases	Microsoft Access 2003, 2007, 2010.	.accdb or .mdb	Microsoft Jet 4.0 OLE DB provider
SQL Server relational databases	Microsoft SQL Server 2005, 2008, 2008 R2, 2012, Windows Azure SQL Database	(not applicable)	OLE DB Provider for SQL Server SQL Server Native Client OLE DB Provider SQL Server Native 11.0 Client OLE DB Provider .NET Framework Data Provider for SQL Client
SQL Server Parallel Data Warehouse (PDW)	2008 R2; 2012	(not applicable)	OLE DB provider for SQL Server PDW
Oracle relational databases	Oracle 9i, 10g, 11g.	(not applicable)	Oracle OLE DB Provider .NET Framework Data Provider for Oracle Client .NET Framework Data Provider for SQL Server MSDAORA OLE DB provider OraOLEDB MSDASQL

Contd....

Teradata relational databases	Teradata V2R6, V12	(not applicable)	TDOLEDB OLE DB provider .Net Data Provider for Teradata
Informix relational databases	V11.10	(not applicable)	Informix OLE DB provider
IBM DB2 relational databases	8.1	(not applicable)	DB2OLEDB
Sybase relational databases	V15.0.2	(not applicable)	Sybase OLE DB provider
Other relational databases	(not applicable)	(not applicable)	OLE DB provider or ODBC driver

Source: http://msdn.microsoft.com/en-us/library/ms175608.aspx

Self Assessment

State whether the following statements are true or false:

1. Multidimensional model must contain at least one data source object.
2. A data source connection can occasionally use Windows authentication or an authentication service provided by the database administration scheme.
3. When retrieving data, the client library making the connection formulates a connection request that includes the credentials in the connection string.
4. In completing the Wizard, the new data source appears in the Data Sources folder.
5. In Data Source View Designer, using the data source view just created right-click anywhere in the work area and select Add/Remove Tables.

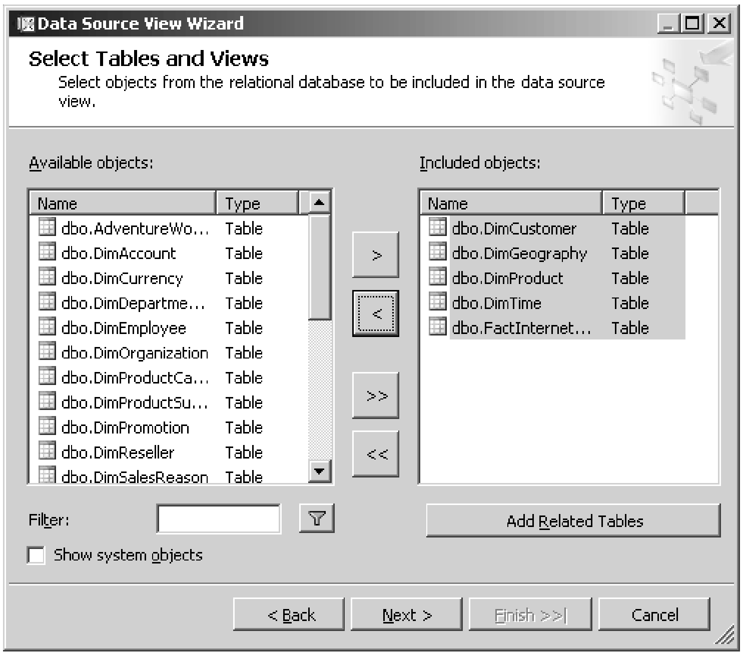
6.2 Creating a Data Source View

After you have defined the data sources that you will use in an Analysis Services project, the next step is generally to define a data source view for the project. To define a new data source view

1. In Solution Explorer, right-click Data Source Views, and then click New Data Source View.
2. On the Data Source View Wizard page, click Next. Then a page appears to select a Data Source.
3. Under Relational data sources, one of the data source will be selected. Click Next.
4. On the Select Tables and Views page, you have to select tables and views from the list of objects that are available from the selected data source. Click > to add the selected tables to the Included objects list. Finally, Click Next and Finish.

Notes

Figure 6.1: Select Tables and Views Page



Select Tables and Views

Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
dbo.AdventureWo...	Table
dbo.DimAccount	Table
dbo.DimCurrency	Table
dbo.DimDepartme...	Table
dbo.DimEmployee	Table
dbo.DimOrganization	Table
dbo.DimProductCa...	Table
dbo.DimProductSu...	Table
dbo.DimPromotion	Table
dbo.DimReseller	Table
dbo.DimSalesReason	Table

Included objects:

Name	Type
dbo.DimCustomer	Table
dbo.DimGeography	Table
dbo.DimProduct	Table
dbo.DimTime	Table
dbo.FactInternet...	Table

Filter:

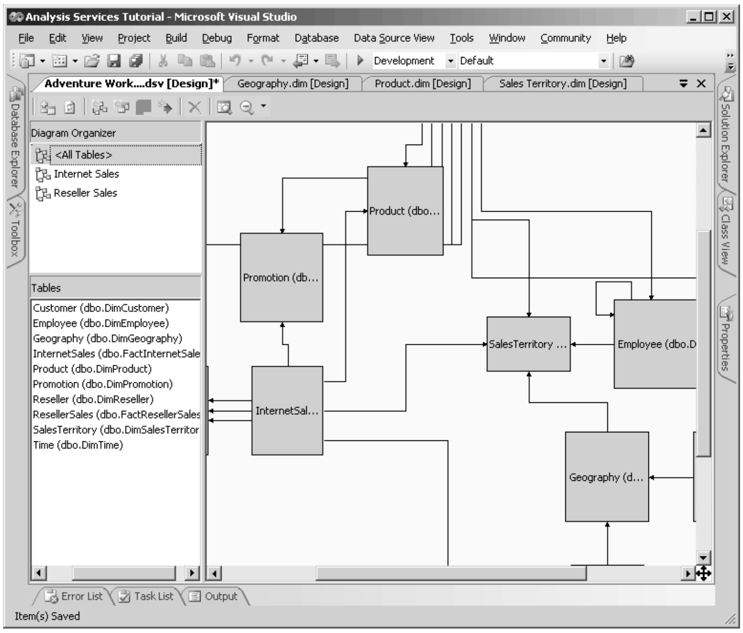
□ Show system objects

Source: http://i.msdn.microsoft.com/dynimg/IC50754.gif

5. To maximize the Microsoft Visual Studio development environment, click the Maximize button.

6. To view the tables in the Diagram pane at 50 per cent, click the Zoom icon on the Data Source View Designer toolbar.

Figure 6.2: Viewing the Tables in a Data Source View



Tables

Customer (dbo.DimCustomer)
Employee (dbo.DimEmployee)
Geography (dbo.DimGeography)
InternetSales (dbo.FactInternetSale
Product (dbo.DimProduct)
Promotion (dbo.DimPromotion)
Reseller (dbo.DimReseller)
ResellerSales (dbo.FactResellerSales
SalesTerritory (dbo.DimSalesTerritor
Time (dbo.DimTime)

Item(s) Saved

Source: http://i.technet.microsoft.com/dynimg/IC102093.gif

7. To hide Solution Explorer, click the Auto Hide button on the title bar. To unhide Solution Explorer, click the Auto Hide button again.

Self Assessment

Fill in the blanks:

6. To maximize the Microsoft Visual Studio development environment, click the .... button.
7. To view the tables in the Diagram pane at 50 per cent, click the .... icon on the Data Source View Designer toolbar.
8. To hide Solution Explorer, click the .... button on the title bar.

6.3 Modifying the Data View

You can use the DataView to add, delete, or modify rows of data in the table. The ability to use the DataView to modify data in the underlying table is controlled by setting one of three Boolean properties of the DataView: AllowNew, AllowEdit, and AllowDelete. They are set to true by default.

Properties are AllowNew, AllowEdit, and AllowDelete. They are set to true by default.

Properties action works in following way:

- If AllowNew is true, you can use the AddNew method of the DataView to create a new DataRowView.
- If AllowNew is false, an exception is thrown if you call the AddNew method of the DataRowView.
- If AllowEdit is true, you can modify the contents of a DataRow via the DataRowView. You can confirm changes to the underlying row using DataRowView.
- If AllowEdit is false, an exception is thrown if you attempt to modify a value in the DataView.
- If AllowDelete is true, you can delete rows from the DataView using the Delete method of the DataView or DataRowView object and the rows will be deleted from the Data table.
- If AllowDelete is false, an exception is thrown if you call the Delete method of the DataView or DataRowView.

6.4 Creating Dimensions, Time and Modifying Dimensions

6.4.1 Creating Dimensions

A database dimension is a collection of associated objects, called attributes, which can be used to supply information about fact data in one or more cubes.

Example: Usual attributes in a product dimension might be product title, product category, product line, product dimensions, and product price. These things are compelled to one or more columns in one or more tables in a data source view. By default, these attributes are visible as attribute hierarchies and can be utilised to realise the detail data in a cube.

Notes

Attributes can be coordinated into user-defined hierarchies that supply navigational routes to assist users when browsing the data in a cube.

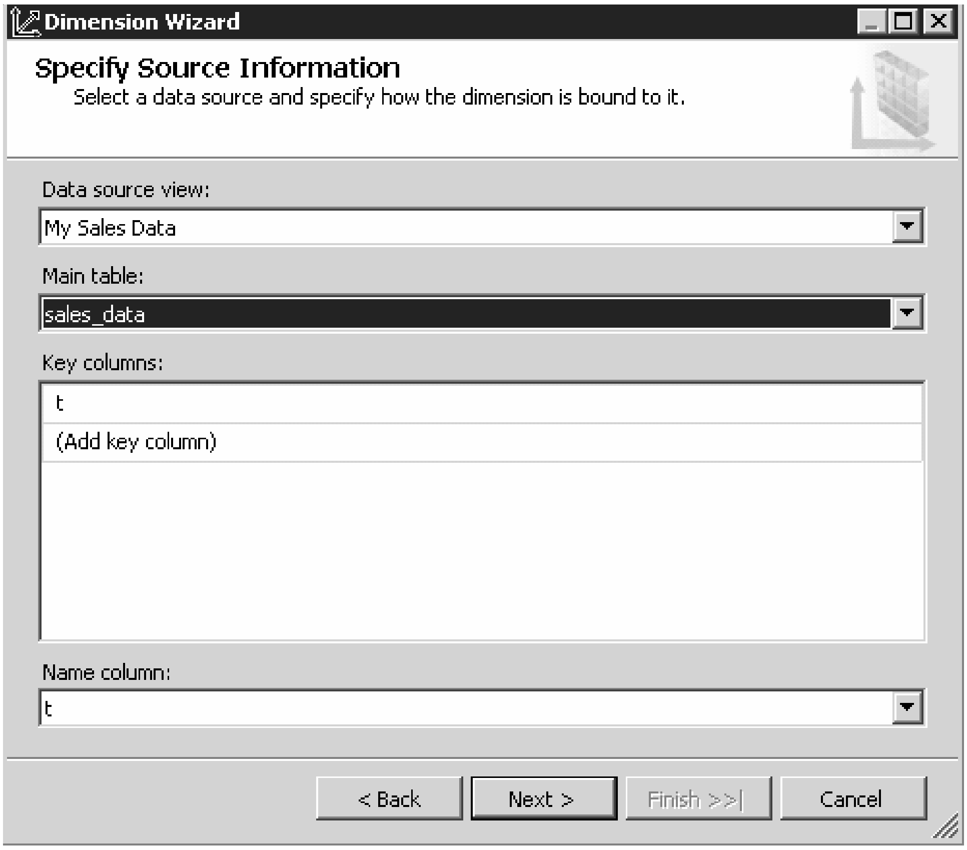
Use the Dimension Wizard in SSDT to create a database dimension in a Microsoft SQL Server Analysis Services project. After a database dimension is created, you can use Dimension Designer to modify its properties. To understand the concept better we will create a DateTime Dimension.

6.4.2 Time Dimension

To Create DateTime Dimension follow these steps:

- Go to Dimensions in project's Solution Explorer. Right-click on Dimensions and click New dimension. A wizard will open. Click Next on that wizard. Choose Use an existing table option and click Next again. On Specify Source Information window, select your Data source view, Main table, Key columns and Name columns.

Figure 6.3: Dimension Wizard: Specify Source Information



Dimension Wizard

Specify Source Information

Select a data source and specify how the dimension is bound to it.

Data source view: My Sales Data

Main table: sales_data

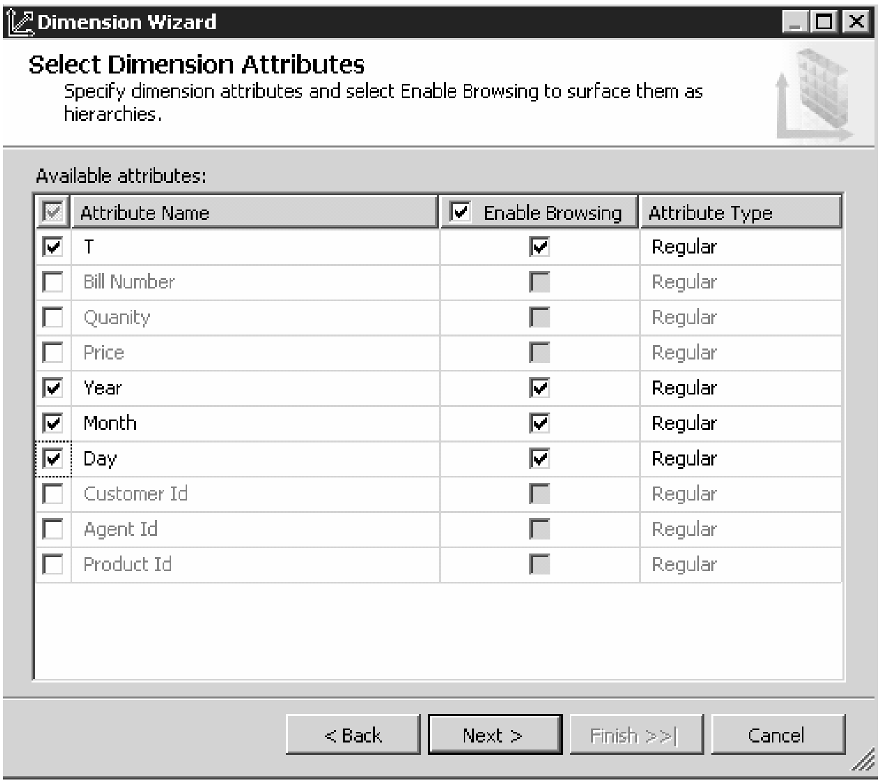
Key columns:
t
(Add key column)

Name column: t

*Source:*http://www.blrf.net/blog/wp-content/uploads/2011/06/visual_studio_2008_dimension_wizard_create_datetime_dimension.png

Click Next and on this window, un-check all related tables offered to you and click Next again. On Select Dimension Attributes only Attribute Names T, Year, Month and Day should be Enabled.

Figure 6.4: Dimension Wizard: Select Dimension attributes



Select Dimension Attributes

Specify dimension attributes and select Enable Browsing to surface them as hierarchies.

Available attributes:

☑	Attribute Name	☑ Enable Browsing	Attribute Type
☑	T	☑	Regular
□	Bill Number	□	Regular
□	Quanity	□	Regular
□	Price	□	Regular
☑	Year	☑	Regular
☑	Month	☑	Regular
☑	Day	☑	Regular
□	Customer Id	□	Regular
□	Agent Id	□	Regular
□	Product Id	□	Regular

Source: http://www.blrf.net/blog/wp-content/uploads/2011/06/visual_studio_2008_dimension_wizard_select_dimension_attributes.png

Click next and on the next window, name this dimension as DateTime and click Finish.

Every dimension has its own Attributes and those attributes can be assigned into Hierarchies, which define how different attributes are related to each other.

In Microsoft SQL Server Analysis Services, you can use the Dimension Wizard in SSDT to create a time dimension when no time table is available in the source database. This can be done by selecting one of the following options on the Select Creation Method page:

- ***Generate a time table in the data source:*** Select this option when you have permission to create objects in the underlying data source.
- ***Generate a time table on the server:*** Select this option when you do not have permission to create objects in the underlying data source.

6.4.3 Modifying the Date Dimension

In this we will create a user-defined hierarchy and change the member names that are displayed for the Date, Month, Calendar Quarter, and Calendar Semester attributes.

Adding a Named Calculation

You can add a named calculation to a table in a data source view. The expression appears and behaves as a column in the table.

To add a named calculation

1. Open the data source view by double-clicking it in the Data Source Views folder in Solution Explorer.
2. In the Tables pane, right-click Date, and then click New Named Calculation.
3. In the Create Named Calculation dialog box, type SimpleDate in the Column name box, and then type the following DATENAMEstatement in theExpression box:

   DATENAME(mm, FullDateAlternateKey) + ‘ ‘ +

   DATENAME(dd, FullDateAlternateKey) + ‘, ‘ +

   DATENAME(yy, FullDateAlternateKey)
4. Click OK, and then expand Date in the Tables pane. On the File menu, click Save All.
5. In the Tables pane, right-click Date, and select Explore Data.
6. Review the last column in the Explore Date Table view. Close the Explore Date Table view.

Self Assessment

Fill in the blanks:

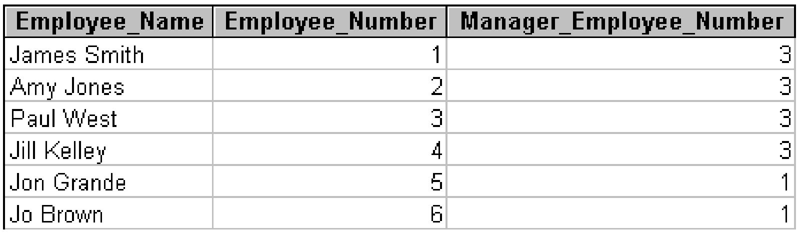
9. We can use the .... to add, delete, or modify rows of data in the table.
10. A database dimension is a collection of associated objects, called................................, which can be used to supply information about fact data in one or more cubes.
11. Uses the Dimension Wizard in .... to create a database dimension in a Microsoft SQL Server Analysis Services project.
12. In Microsoft SQL Server Analysis Services, you can use the .... in SSDT to create a time dimension when no time table is available in the source database.

6.5 Parent-Child Dimensions

A parent-child dimension is based on two dimension table columns that together define the lineage relationships among the members of the dimension. One column, called the member key column, identifies each member; the other column, called the parent key column, identifies the parent of each member.

Example: In the following Employee table, the column that identifies each member is Employee_Number. The column that identifies the parent of each member is Manager_Employee_Number.

Table 6.2: Sample Employee Table

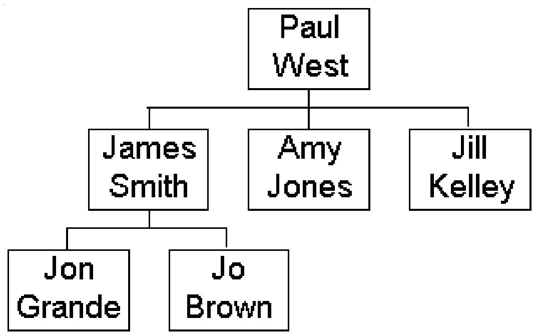


Employee_Name	Employee_Number	Manager_Employee_Number
James Smith	1	3
Amy Jones	2	3
Paul West	3	3
Jill Kelley	4	3
Jon Grande	5	1
Jo Brown	6	1

Source: http://i.msdn.microsoft.com/dynimg/IC553.gif

These columns can be used to define a parent-child dimension that contains the following member hierarchy.

Figure 6.5: Sample Member Hierarchy



Source: http://i.msdn.microsoft.com/dynimg/IC44843.gif

Both columns must have the same data type. Both columns must be in the same table.

When you define a parent-child dimension, you can also select a third column to provide member names, which are displayed to end users as they browse cubes. The depth of a parent-child dimension can vary among its hierarchy's branches.

You can use the Dimension Wizard to create parent-child dimensions. After you create a parent-child dimension, you can edit it in Dimension Editor (if the dimension is shared) or Cube Editor (if the dimension is private).

Task Make report on application of business intelligence and planning in midsize companies.

Self Assessment

State whether the following statements are true or false:

13. A parent-child dimension is based on two dimension table columns that together define the lineage relationships among the members of the dimension.

14. When you define a parent-child dimension, you can also select a third column to provide member names, which are displayed to end users as they browse cubes.

15. You can use the Dimension Wizard to create parent-child dimensions.

Case Study

Managing Data Sources for Input to Data Warehousing and Business Intelligence

Data warehousing and business intelligence effort is only as good as the data that is put into it. The saying "Garbage In, Garbage Out" is all too true. A leading cause of data warehousing and business intelligence project failures is to obtain the wrong or poor quality data.

Contd....

Notes

Managing data warehouse input sources includes a number of steps organized into two phases. In the first phase the following activities are undertaken:

- Manage the Data Source Identification Process
- Identify Subject Matter Experts (SMEs)
- Identify Dimension Data Sources
- Identify Fact Data Sources

When the major data sources have been identified it is time to quickly gain detailed understanding of each one:

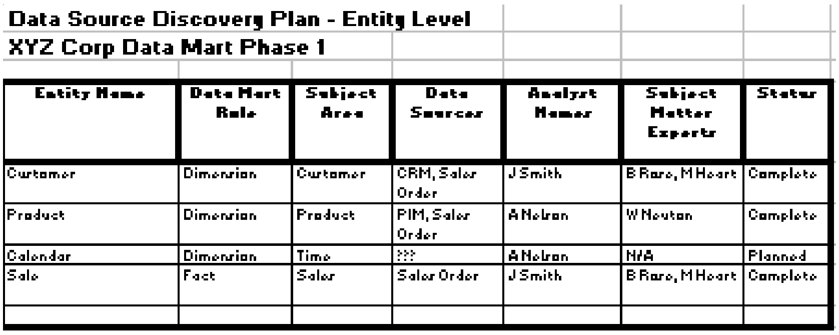
- Obtain Existing Documentation
- Model and Define the Input
- Profile the Input
- Improve Data Quality
- Save Results for Further Reuse

Manage the Data Warehousing Data Source Identification Process

The source identification process is critical to the success of data warehousing and business intelligence projects. It is important to move through this effort quickly, obtaining enough information about the data sources without being bogged down in excess detail while still obtaining the needed information.

Start out with a list of the entities planned for the data warehouse / data mart. This can be managed with a spreadsheet containing these columns:

- Entity name
- Data mart role (Fact, Dimension, Bridge, etc.)
- Subject Area
- Data Source(s)
- Analyst Name(s)
- Subject Matter Expert(s)
- Status



Data Source Discovery Plan - Entity Level

XYZ Corp Data Mart Phase 1

Entity Name	Data Mart Role	Subject Area	Data Sources	Analyst Names	Subject Matter Experts	Status
Customer	Dimension	Customer	CRM, Sales Order	J Smith	B Rose, M Heart	Complete
Product	Dimension	Product	PIM, Sales Order	A Nelson	W Newton	Complete
Calendar	Dimension	Time	???	A Nelson	N/A	Planned
Sale	Fact	Sales	Sales Order	J Smith	B Rose, M Heart	Complete

Complete the entity name, data mart role and subject area entries. Assign an analyst to each entity who will find data sources and subject matter experts for each entity.

Contd....

Identify Data Warehousing Data Source Subject Matter Experts

Consider the following questions when determining the sources and costs of data for the Data Warehouse:

- Where does the data come from?
- What processes are used to obtain the data?
- What does it cost to obtain the data?
- What does it cost to store the data?
- What does it cost to maintain the data?

Identify Dimension Data Sources for the Data Mart

Dimensions enable business intelligence users to put information in context. They focus on questions of: who, when, where and what. Typical dimensions include:

- Time period/calendar
- Product
- Customer
- Household
- Market Segment
- Geographic Area

Master data is a complementary concept and may provide the best source of dimensional data for the data warehouse. Master data is data shared between systems that describe entities like: product, customer and household. Master data is managed using a Master Data Management (MDM) system and stored in an MDM-Hub. Benefits of this approach include:

- It is less expensive to access data from a single source (MDM-Hub) than extracting from multiple sources.
- MDM data is rationalized.
- MDM data is of high quality

If an MDM-Hub does not exist consider creating one. It will have many uses beyond supporting the data warehouse and business intelligence.

If no MDM-Hub is available, you will need to examine source systems and determine which system contains the data most suitable for dimensions. If the data is not stored in a managed database, you may need to define the data locally, in a spreadsheet or desktop database, and then provide to the data warehousing system.

Identify Fact Data Sources for the Data Mart

The Fact contains quantitative measurements while the Dimension contains classification information. The data sources for Fact tend to be transactional software systems. For example:

Contd....

Notes

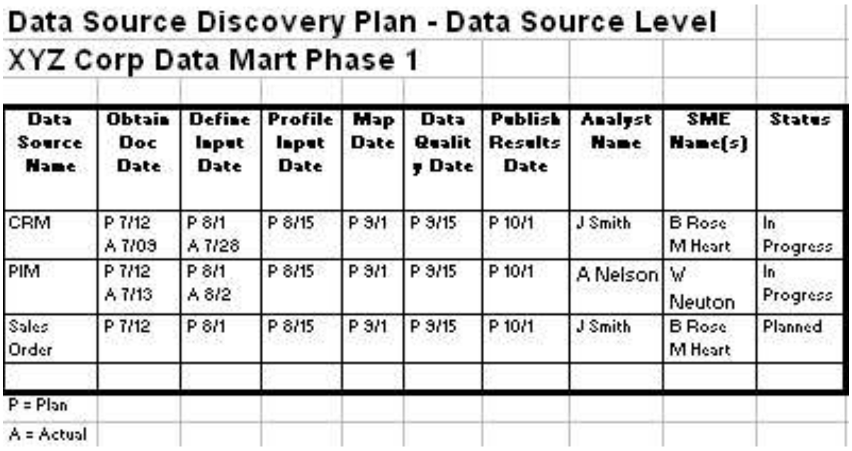
System	Example Fact Data
Sales Order Entry	● Sales Transaction ● Return Transaction
Customer Service	● Service Episode ● Service Result
Accounts Payable	● Payment Transaction
Sales Campaign	● Sales Campaign Event

Larger enterprises may have multiple systems for the same kind data. In that case, you will need to determine the best source of data - the System of Record (SOR) as the source of data warehousing data.

Detailed Data Source Understanding for Data Warehousing

When the major data sources have been identified it is time to quickly gain detailed understanding of each one. Consolidate the spreadsheet developed in the identification phase by data source, then create a new spreadsheet to track and control detailed understanding:

- Data Subject Name
- Obtain Doc Date
- Define Input Date
- Profile Input Date
- Map Date
- Data Quality Date
- Save Results
- Analyst Name
- SME Name(s)
- Status



Data Source Discovery Plan - Data Source Level

XYZ Corp Data Mart Phase 1

Data Source Name	Obtain Doc Date	Define Input Date	Profile Input Date	Map Date	Data Quality Date	Publish Results Date	Analyst Name	SME Name(s)	Status
CRM	P 7/12 A 7/09	P 8/1 A 7/28	P 8/15	P 9/1	P 9/15	P 10/1	J Smith	B Rose M Heart	In Progress
PIM	P 7/12 A 7/13	P 8/1 A 8/2	P 8/15	P 9/1	P 9/15	P 10/1	A Nelson	W Neuton	In Progress
Sales Order	P 7/12	P 8/1	P 8/15	P 9/1	P 9/15	P 10/1	J Smith	B Rose M Heart	Planned

P = Plan

A = Actual

This approach provides an effective workflow as well as a project planning and control method. Due dates are assigned and actual complete dates and status are tracked.

Contd....

Obtain Existing Documentation

When seeking to understand a data source, the first thing to do is look at existing documentation. This avoids "re-inventing the wheel". If a data source is fully documented, data profiled and of high quality most of the job of data source discovery is complete.

Existing documentation may include:

- Data models
- Data dictionary
- Internal/technical documentation
- Business user guides
- Data profiles and data quality assessments

Check through the documentation to assess its completeness and usefulness.

The data source analyst should study the existing documentation before any in depth discussions with the SMEs. This improves the credibility of the data analyst and save time for the SMEs.

Model and Define the Input

The data model is a graphic representation of data structures that improves understanding and provides automation linking database design to physical implementation. This section assumes that the data source is stored in a relational database that modelled using typical relational data modelling tools.

If there is an existing data model, start with that, otherwise use the reverse engineering capability of the data modelling to build a physical data model. Next, group the tables that are of interest into a subject area for analysis. Unless, a large percentage of the data source is needed for the data warehouse avoid studying the entire data source. Stay focused on the current project.

For each selected data source table define:

- Physical Name
- Logical Name
- Definition
- Notes

For each selected data source column define:

- Physical Name
- Logical Name
- Order in Table
- Datatype
- Length
- Decimal Positions
- Nullable/Required
- Default Value

Contd....

Notes

- Edit Rules
- Definition
- Notes

Profile the Data Source

The actual use and behaviour of data sources often tends not to match the name or definition of the data. Sometimes this is called "dirty data" or "unrefined data" that may have problems such as:

- Invalid code values
- Missing data values
- Multiple uses of a single data item
- Inconsistent code values
- Incorrect values such as sales revenue amounts

Data profile is an organized approach to examining data to better understand and later use it. This can be accomplished by querying the data using tools like:

- SQL Queries
- Reporting tools
- Data quality tools
- Data exploration tools

For code values such as gender code and account status code do a listing showing value and count such as this gender code listing:

Code	Count	Notes
F	500	Female
M	510	Male
T	12	Transgender?
Z	5	???
NULL	1000	Missing

Other systems may represent female and male as 1 and 2 rather than F and T, and so may require standardization when stored in the data warehouse. When data from multiple sources is integrated in the data warehouse it is expected that it will be standardized and integrated.

Statistical measures are a good way to better understand numeric information such as revenue amounts. Helpful statistics are**:**

- Mean (average)
- Median
- Mode
- Maximum
- Minimum
- Quartile Averages

Contd....

- Standard Deviation
- Variance

Consistency within a database is another important factor to determine through data profiling. For example, there may be an order table which should only have orders for customers established in the customer table. Perform queries to determine whether this is true.

Improve Data Quality

Data profiling may reveal problems in data quality. For example, it might show invalid values are be entered for a particular column, such as entering 'Z' for gender when 'F' and 'M' are the valid values. Some steps that could be taken to improve data quality include:

- Work with data owners to define the appropriate level of data quality. Build this into a data governance program.
- Determine why there are data quality problems — do a root cause analysis.
- Correct the data in the source system through manual or automated efforts.
- Add edits or database rules to prevent the problem.
- Change business processes to enter correct data.
- Make data quality visible to the business through scorecards, dashboards and reports.

Save Results for Further Reuse

The information gathered during the data source discovery process is valuable metadata that can be useful for future data warehousing or other projects. Be sure to save the results and make available for future efforts. This work can be a great step toward building an improved data resource.

Question:

Discuss the case study in contrast with efficient and effective workflow of obtaining the right source data and using it in the data warehousing and business intelligence project.

Source: http://infogoal.com/datawarehousing/data_sources_2.htm

6.6 Summary

- A multidimensional model must contain at least one data source object, but you can add more to combine data from several data warehouses.
- A data source connection can occasionally use Windows authentication or an authentication service provided by the database administration scheme, such as SQL Server authentication when connecting to SQL Azure databases.
- The attachment string is formulated based on the properties you choose in the Data Source Designer or the New Data Source Wizard.
- After you have defined the data sources that you will use in an Analysis Services project, the next step is generally to define a data source view for the project.
- The ability to use the DataView to modify data in the underlying table is controlled by setting one of three Boolean properties of the DataView: AllowNew, AllowEdit, and AllowDelete.

- A database dimension is a collection of associated objects, called attributes, which can be used to supply information about fact data in one or more cubes.
- You can add a named calculation to a table in a data source view. The expression appears and behaves as a column in the table.
- A parent-child dimension is based on two dimension table columns that together define the lineage relationships among the members of the dimension.

6.7 Keywords

Data Source Reference: A data source reference is an association to another Analysis Services project or data source in the same solution.

OLTP (Online Transaction Processing): Online transaction processing, or OLTP, is a class of information systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

Oracle: The Oracle Database (commonly referred to as Oracle RDBMS or simply as Oracle) is an object-relational database management system produced and marketed by Oracle Corporation.

Parent-child Dimension: A parent-child hierarchy is a hierarchy in a standard dimension that contains a parent attribute.

RDBMS: RDBMS stands for Relational Database Management System. RDBMS data is structured in database tables, fields and records.

SQL: SQL (Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS).

6.8 Review Questions

1. What are the steps of creating data source?
2. Write down the process of creating data source using the data source wizard.
3. What are the types of data sources?
4. Write down the steps in creating a data source view.
5. Give explanation of modifying the data view.
6. Explain about the time dimension options in SQL server 2008 Analysis Services.
7. What are parent-child dimensions?

Answers: Self Assessment

1. True
2. True
3. True
4. False
5. True
6. Maximize
7. Zoom
8. Auto Hide
9. DataView
10. Attributes
11. SSDT
12. Dimension Wizard
13. True
14. True
15. True

6.9 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

https://ojs.hh.se/index.php/jisib/article/download/19/pdf?

site.xavier.edu/sena/info600/businessintelligence.pdf?

www.cgi.com/sites/cgi.com/files/.../business-intelligence-white-paper.pdf?

www.ibm.com/businesscenter/cpe/download0/.../practicalframework.pdf

Notes

Unit 7: Creating Cube

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss the Wizard to Create Cube
- Recall the Concept of Cube
- Explain the Adding Measure and Measure Groups to a Cube
- Demonstrate Calculated Members
- Identify Deploying and Browsing a Cube

Introduction

A data cube is a three (or more) dimensional array of values, commonly used to describe a time series of image data. An OLAP cube is an array of data understood in terms of its 0 or more dimensions. Benefit of building a cube to store your data is that you can centralize the business rules for calculations that you can’t easily store in a relational data mart. The structure of the cube makes it much easier to write queries to compare data year over year and also you gain the ability to transparently manage aggregated data in the cube. In this chapter, you will learn about creating a cube using Wizard. It also discusses adding measure and measure groups to cube. Finally, calculated measures and deploying and browsing of cube are discussed in the unit.

7.1 Wizard to Create Cube

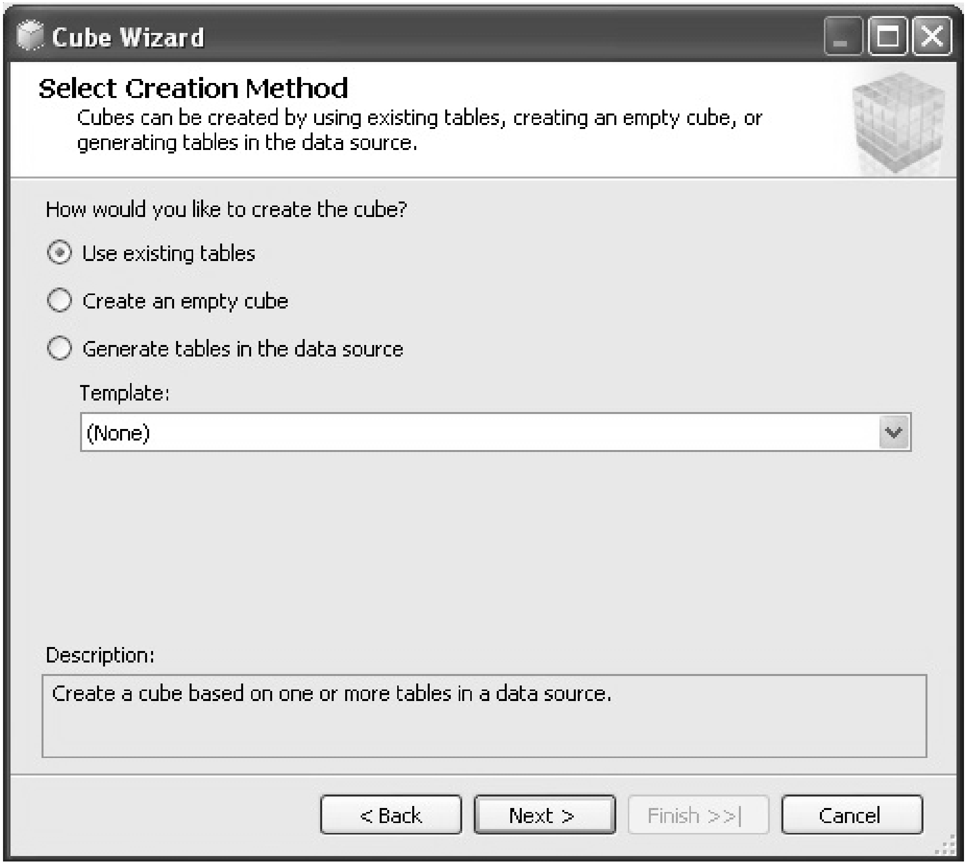
Use the Cube Wizard to create a cube rapidly and effortlessly. When you create the cube, you can add living dimensions or create new dimensions that structure the cube. You can also create dimensions separately, using the Dimension Wizard, and then add them to a cube.

A Cube acts as an OLAP database to the subscribers who need to query data from an OLAP data store. A Cube is the main object of a SSAS solution where the majority of fine tuning, calculations, aggregation design etc. Now, we will create a cube using our dimension and fact tables. We will use SQL Server 2008 here.

7.1.1 Explanation

Right click the Cube folder and select "New Cube", and it will invoke the Cube Wizard. In the first screen select one of the methods of creating a Cube. We assume our dimensions are ready, and schema is already designed to contain dimension and fact tables. So we will select the option of "Use existing tables".

Figure 7.1: Select Creation Method



Cube Wizard

Select Creation Method

Cubes can be created by using existing tables, creating an empty cube, or generating tables in the data source.

How would you like to create the cube?

● Use existing tables

○ Create an empty cube

○ Generate tables in the data source

Template:

(None)

Description:

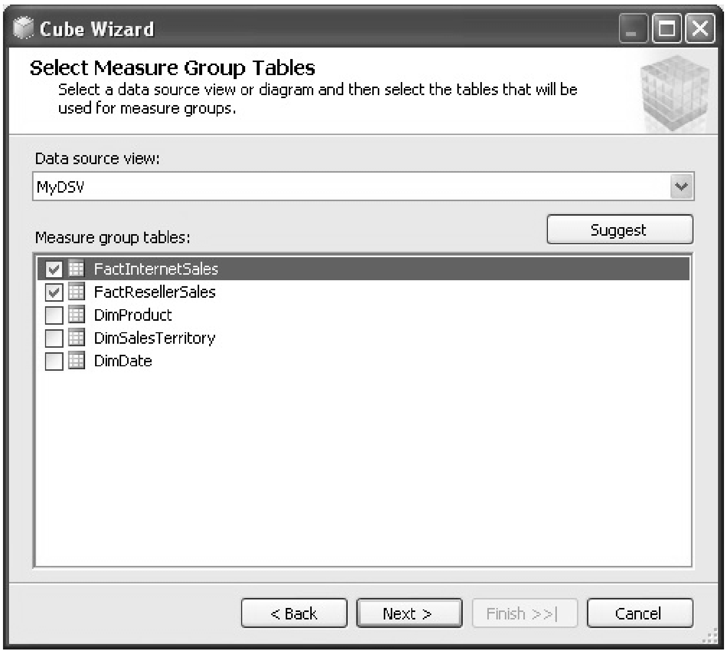
Create a cube based on one or more tables in a data source.

Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_1.jpg

In the next screen, we need to select the tables which will be used to create measure groups. We again assume we have a DSV which has fact tables in the schema. So we will use this as shown in the **Figure 7.2**.

Notes

Figure 7.2: Select Measure Group Table



Cube Wizard

Select Measure Group Tables

Select a data source view or diagram and then select the tables that will be used for measure groups.

Data source view: MyDSV

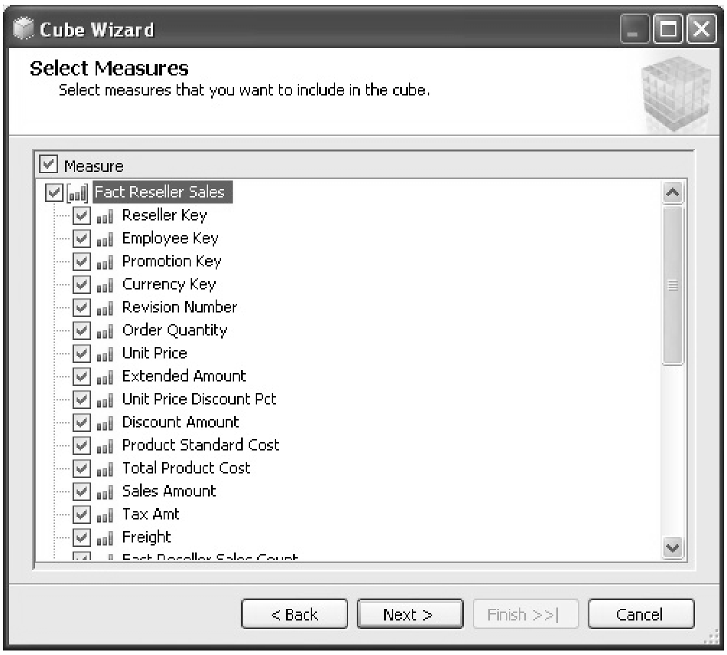
Measure group tables:

☑ FactInternetSales
☑ FactResellerSales
□ DimProduct
□ DimSalesTerritory
□ DimDate

Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_2.jpg

In the next screen, we need to select the measures that we want to create from the fact tables we just selected in the previous screen. For now, select all the fields as shown below and move to the next screen.

Figure 7.3: Select Measures



Cube Wizard

Select Measures

Select measures that you want to include in the cube.

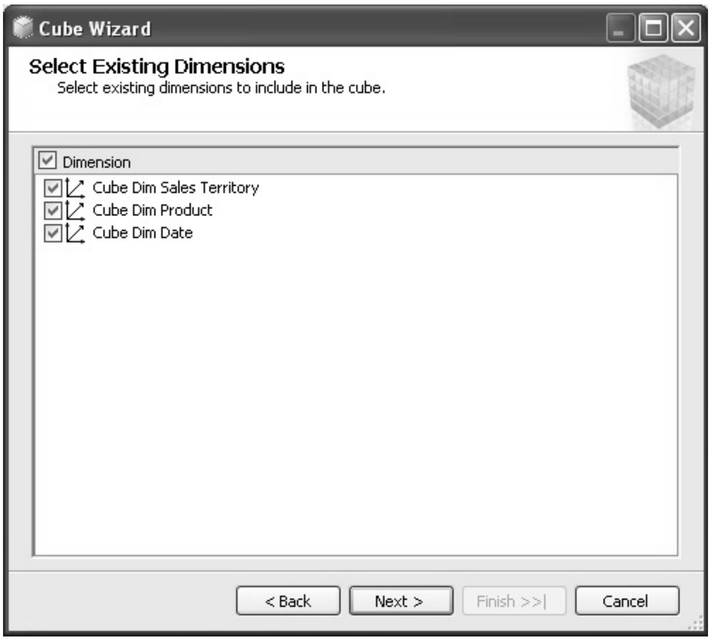
☑ Measure

- ☑ Fact Reseller Sales
  - ☑ Reseller Key
  - ☑ Employee Key
  - ☑ Promotion Key
  - ☑ Currency Key
  - ☑ Revision Number
  - ☑ Order Quantity
  - ☑ Unit Price
  - ☑ Extended Amount
  - ☑ Unit Price Discount Pct
  - ☑ Discount Amount
  - ☑ Product Standard Cost
  - ☑ Total Product Cost
  - ☑ Sales Amount
  - ☑ Tax Amt
  - ☑ Freight
  - ☑ Fact Reseller Sales Count

Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_3.jpg

In this screen you need to select any existing dimensions. We have created three dimensions and we will include all of these dimensions as shown below:

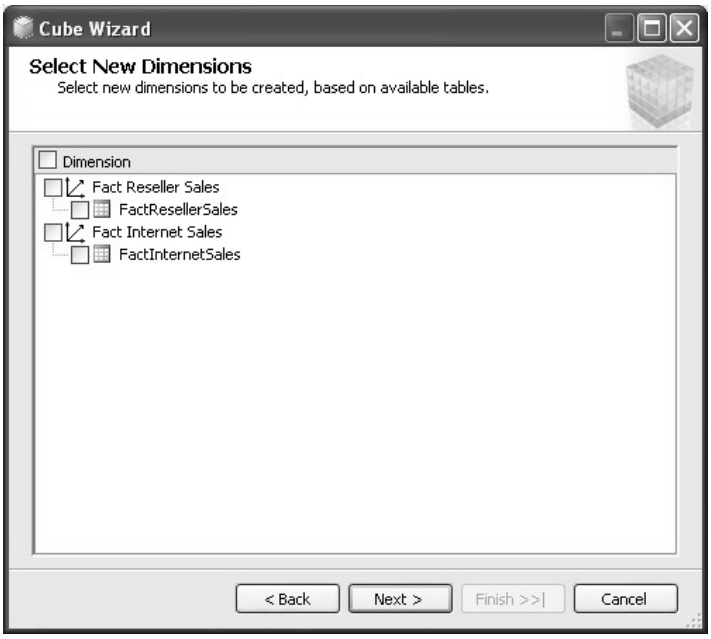
Figure 7.4: Select Existing Dimensions


Cube Wizard
Select Existing Dimensions
Select existing dimensions to include in the cube.
☑ Dimension
☑ Cube Dim Sales Territory
☑ Cube Dim Product
☑ Cube Dim Date
< Back
Next >
Finish >>|
Cancel

Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_4.jpg

In the next screen, we can select if we want to create any additional new dimensions from the tables available in the DSV. We do not want to create any more dimensions, so unselect any selected tables as shown below and move to the next screen.

Figure 7.5: Select New Dimensions

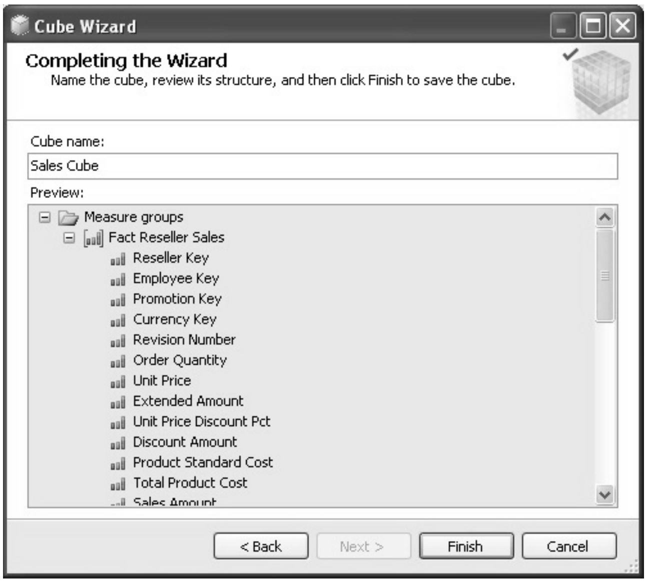


Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_5.jpg

Notes

Finally you need to name your cube, which is the last step of the wizard before your cube is created. Name it something appropriate like "Sales Cube" as shown below.

Figure 7.6: Completing the Wizard



Cube Wizard

Completing the Wizard

Name the cube, review its structure, and then click Finish to save the cube.

Cube name: Sales Cube

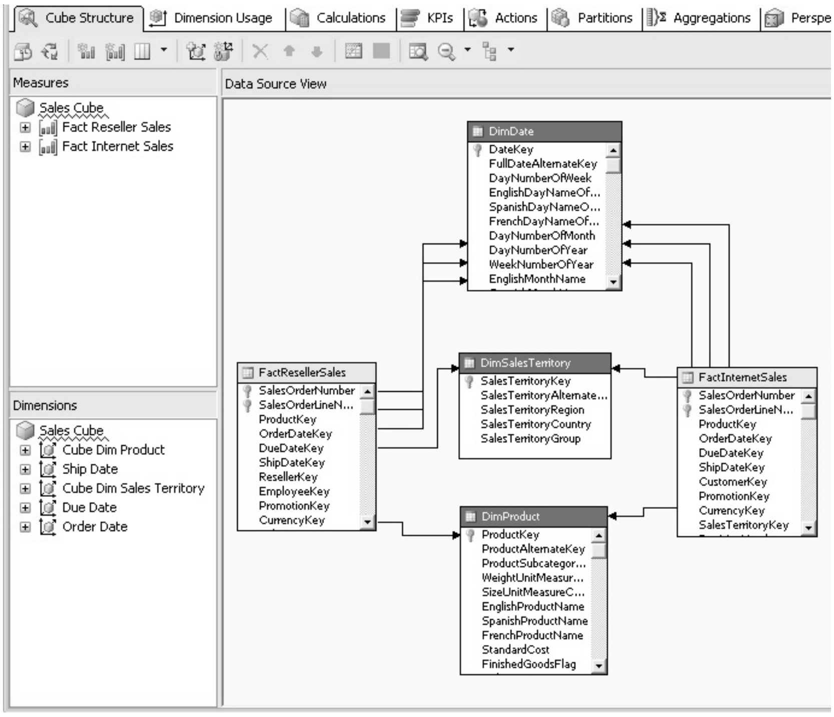
Preview:

- Measure groups
  - Fact Reseller Sales
    - Reseller Key
    - Employee Key
    - Promotion Key
    - Currency Key
    - Revision Number
    - Order Quantity
    - Unit Price
    - Extended Amount
    - Unit Price Discount Pct
    - Discount Amount
    - Product Standard Cost
    - Total Product Cost
    - Sales Amount

Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_6.jpg

Now your cube should have been created and if your cube editor is open you should find different tabs to configure and design various features and aspects of the cube.

Figure 7.7: Final Cube Structure



Source: http://www.mssqltips.com/tutorialimages/2008_Cube_Wizard_Step_7.jpg

Task Find out the defaults used by the create cube wizard.

Self Assessment

Fill in the blanks:

1. A Cube is the main object of a .... where the majority of fine tuning, calculations, aggregation design etc.
2. Right-click the Cube folder and select "New Cube", and it will invoke the ....

7.2 Concept of Cube

A cube is a multidimensional structure that contains information for analytical purposes; the main constituents of a cube are dimensions and measures. Dimensions define the structure of the cube, and measures provide aggregated numerical values of interest to the end user. A cell, in the cube, is defined by the intersection of dimension members and contains the aggregated values of the measures at that specific intersection.

Did u know? A cube provides a single place where all related data, for analysis, is stored.

A cube is composed of:

- Dimensions
- Measures and Measure Groups
- Partitions
- Perspectives
- Hierarchies
- Actions
- Key Performance Indicators (KPI)
- Calculations
- Translations

7.3 Adding Measure and Measure Groups to a Cube

You can use the Define New Measures page to create new measures for a cube that is being created without using a data source.

To get familiar with all related operations of a Cube, read the following context:

- If you select measures from template Options it will displays the measures from the cube template to include in the cube.
- To include a specific measure from the template, select the check box for that measure.
- To include all measures from the template in the cube, select the check box in the header.
- Use *Measure Nam*e to lists the measures that are available in the template.

Notes

- To rename a measure, click on that measure and type a new name.
- Use *Measure Group* to lists the measure group for the measure.
- To change the measure group, click on the measure group, and then either enter a new measure group or select an existing measure group from the list.

To add a measure group to a cube

1. In Solution Explorer, right-click the cube, and then click View Designer.
2. In Cube Designer, click the Cube Structure tab.
3. Either click the New Measure Group button or right-click anywhere in the Measures pane and then click New Measure Group.
4. In New Measure Group, click the table from the data source view that you want to use as the new measure group, and then click OK.

To remove a measure group from a cube

1. In Solution Explorer, right-click the cube, and then click View Designer.
2. In Cube Designer, click the Cube Structure tab.
3. In the Measures pane, click the measure group that you want to remove.
4. Either click the Delete button or right-click the measure group and then click Delete.
5. In the Delete Objects dialog box, review the object to be deleted, and then click OK.

Self Assessment

Fill in the blanks:

3. A .... is a multidimensional structure that contains information for analytical purposes; the main constituents of a cube are dimensions and measures.
4. A cube provides a single place where all related data, for analysis, is ....
5. We use .... to lists the measures that are available in the template.
6. We use .... to lists the measure group for the measure.

7.4 Calculated Members

A calculated member is a dimension member whose worth is calculated at run time using a sign that you specify when you define the calculated member. Calculated member can also be defined as assessees.

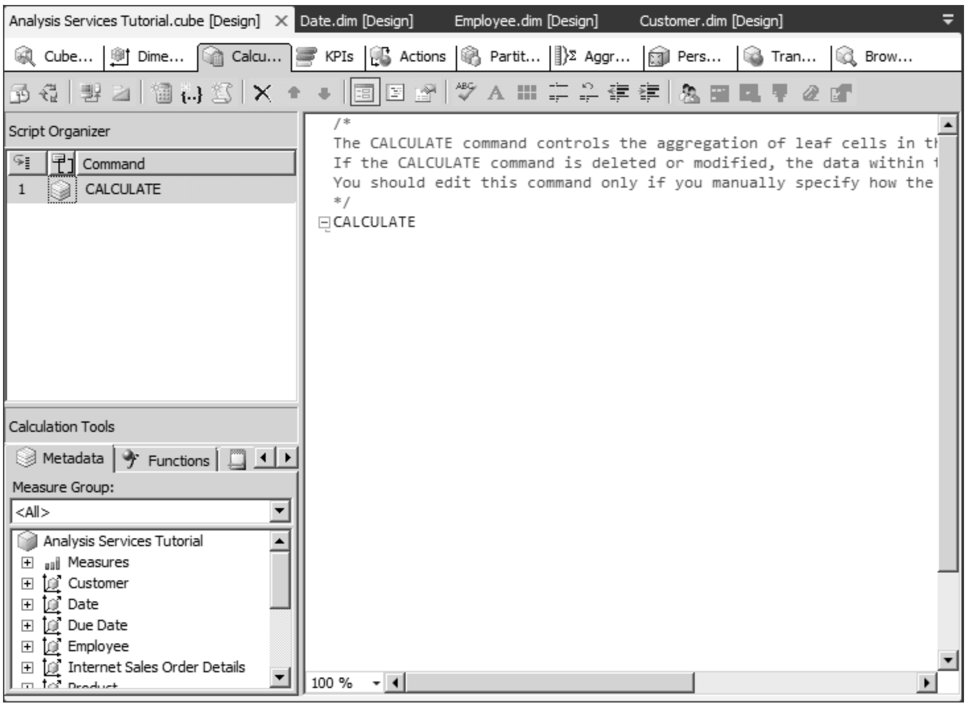
⚠ *Caution* Only the definitions for calculated members are retained; values are calculated in memory when required to answer a query.

Calculated members enable you to add members and measures to a cube without expanding its size. Although calculated members must be based on data (such as constituents) that currently lives in the cube, you can create complex expressions by combining this data with arithmetic operators, figures, and a variety of functions. Analysis Services constitutes a library of over 100 functions and allows you to list and use other function libraries.

Calculated members have a Format String property that controls the format of cell values displayed to end users. This property is accessed in the properties pane of Cube Editor. The Format String property accepts the same values as the Display Format property of measures.

The following image shows the Calculations tab of Cube Designer:

Figure 7.8: Calculations tab of Cube Designer



```
/*
The CALCULATE command controls the aggregation of leaf cells in th
If the CALCULATE command is deleted or modified, the data within t
You should edit this command only if you manually specify how the
*/
CALCULATE
```

Source: http://i.msdn.microsoft.com/dynimg/IC574394.gif

Self Assessment

State whether the following statements are true or false:

7. Only the definitions for calculated members are retained; values are calculated in memory when required to answer a query.

8. Calculated members have a Format String property that controls the format of cell values displayed to end users.

7.5 Deploying and Browsing a Cube

Browsing a deployed cube helps you understand the modifications that you should make to improve the functionality of the cube.

Example: You may have to define dimension member sort orders, delete unnecessary dimension attributes, define new user hierarchies, modify existing user hierarchies, or configure measure properties.

Notes

After you deploy a cube, cube data is viewable on the Browser tab in Cube Designer, and dimension data is viewable on the Browser tab in Dimension Designer.

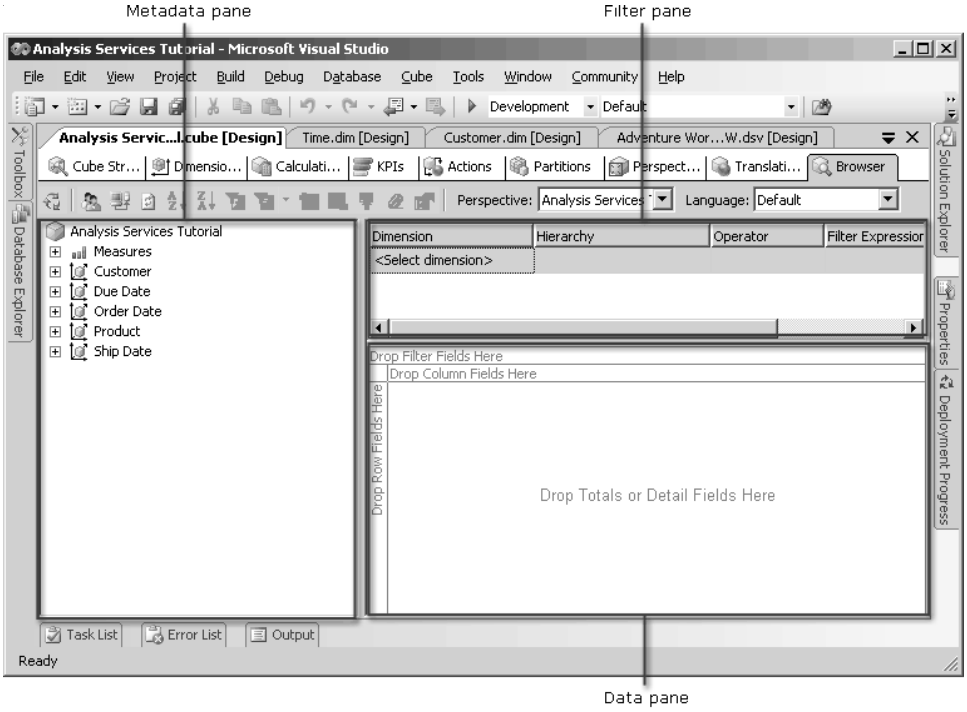
To browse the deployed cube:

1. Switch to Cube Designer in BI Development Studio by clicking the Analysis Services Tutorial cube.

2. Select the Browser tab, and then click Reconnect on the toolbar of the designer.

The following image highlights the individual panes in Cube Designer.

Figure 7.9: Individual Panes in Cube Designer



Source: http://i.technet.microsoft.com/dynimg/IC29955.gif

As you can see, the left pane of the designer shows the metadata for the Analysis Services Tutorial cube. Perspective and Language options are available on the toolbar of the Browser tab.

Notes The Browser tab includes two panes to the right of the metadata pane: the upper pane is the filter pane, and the lower pane is the data pane.

3. In the metadata pane, expand Measures, expand Internet Sales, and then drag the Sales Amount measure to the Drop Totals or Detail Fields Here area of the Data pane.

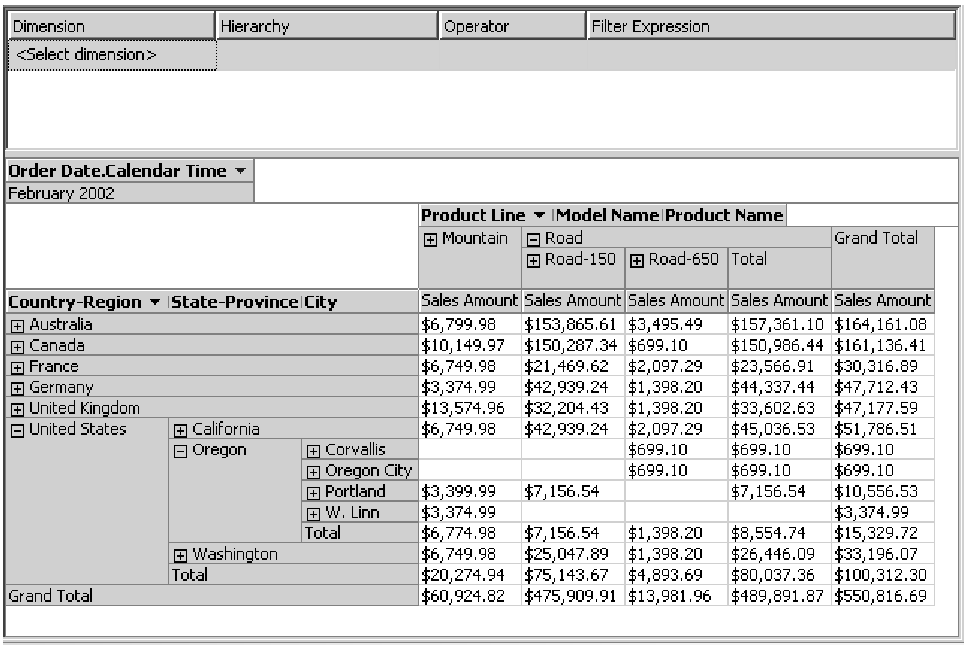
4. In the metadata pane, expand Product. Drag the Product Model Lines user hierarchy to the Drop Column Fields Here area of the data pane, and then expand the Road member of the Product Line level of this user hierarchy.

5. In the metadata pane, expand Customer, expand Location, and then drag the Customer Geography hierarchy from the Location display folder in the Customer dimension to the Drop Row Fields Here area of the data pane.

6. In similar way, expand, United States, Order date, Customer and other headings as required.

Figure 7.10 shows Internet sales by region and product line for the month of February, 2002.

Figure 7.10: Internet sales by region and product line for the month of February, 2002.



Dimension	Hierarchy	Operator	Filter Expression
<Select dimension>			

Order Date.Calendar Time ▾
February 2002

			Product Line ▾ | Model Name | Product Name				
			⊞ Mountain	⊟ Road			Grand Total
				⊞ Road-150	⊞ Road-650	Total	
Country-Region ▾	State-Province	City	Sales Amount	Sales Amount	Sales Amount	Sales Amount	Sales Amount
⊞ Australia			$6,799.98	$153,865.61	$3,495.49	$157,361.10	$164,161.08
⊞ Canada			$10,149.97	$150,287.34	$699.10	$150,986.44	$161,136.41
⊞ France			$6,749.98	$21,469.62	$2,097.29	$23,566.91	$30,316.89
⊞ Germany			$3,374.99	$42,939.24	$1,398.20	$44,337.44	$47,712.43
⊞ United Kingdom			$13,574.96	$32,204.43	$1,398.20	$33,602.63	$47,177.59
⊟ United States	⊞ California		$6,749.98	$42,939.24	$2,097.29	$45,036.53	$51,786.51
	⊟ Oregon	⊞ Corvallis			$699.10	$699.10	$699.10
		⊞ Oregon City			$699.10	$699.10	$699.10
		⊞ Portland	$3,399.99	$7,156.54		$7,156.54	$10,556.53
		⊞ W. Linn	$3,374.99				$3,374.99
		Total	$6,774.98	$7,156.54	$1,398.20	$8,554.74	$15,329.72
	⊞ Washington		$6,749.98	$25,047.89	$1,398.20	$26,446.09	$33,196.07
	Total		$20,274.94	$75,143.67	$4,893.69	$80,037.36	$100,312.30
Grand Total			$60,924.82	$475,909.91	$13,981.96	$489,891.87	$550,816.69

Self Assessment

Fill in the blanks:

9. After you deploy a cube, ... is viewable on the Browser tab in Cube Designer, and .... is viewable on the Browser tab in Dimension Designer.

10. The Browser tab includes two panes to the right of the metadata pane: the upper pane is the ...., and the lower pane is the ....

Case Study

Building a Data Cube

This example uses sales figures from XYZ Co., which makes many kinds of widgets. For each sales transaction, we know four pieces of data:

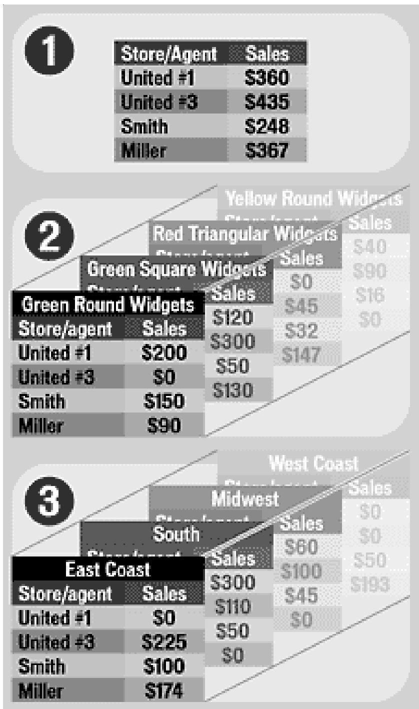
- Which types of widget were involved (style, colour, size and so on)
- Store or sales agent
- Sales amount
- Geographic region or territory

In a real-world situation, we would also know many other data items, including:

- Quantity

Contd....

- Customer
- Cost to XYZ for each widget
- Order date
- Shipment date
- Method and cost of shipping



1

Store/Agent	Sales
United #1	$360
United #3	$435
Smith	$248
Miller	$367

2

Yellow Round Widgets

Sales
$40
$90
$16
$0

Red Triangular Widgets

Sales
$0
$45
$32
$147

Green Square Widgets

Sales
$120
$300
$50
$130

Green Round Widgets

Store/agent	Sales
United #1	$200
United #3	$0
Smith	$150
Miller	$90

3

West Coast

Sales
$0
$0
$50
$193

Midwest

Sales
$60
$100
$45
$0

South

Sales
$300
$110
$50
$0

East Coast

Store/agent	Sales
United #1	$0
United #3	$225
Smith	$100
Miller	$174

Any of these pieces of data can function as a dimension in a data cube. We can take any two dimensions and produce a 2-D table:

1. Thus we can correlate or track sales against individual stores or sales agents. Add in a third factor, such as price, and we can produce a 3-D data cube,
2. That allows us to see how much each store or sales agent is selling in addition to which type of widget. Swap in geography, and
3. We can now see who is selling where.

Questions:

1. What are different pieces of data available for the company here?
2. What can be effect of price change (increase by 25%) in United States on the sales in this company?

Source: http://www.computerworld.com/s/article/91640/Data_Cubes

7.6 Summary

- A Cube is the main object of a SSAS solution where the majority of fine tuning, calculations, aggregation design etc. Now, we will create a cube using our dimension and fact tables. We will use SQL Server 2008 here.

- A cube is a multidimensional structure that contains information for analytical purposes; the main constituents of a cube are dimensions and measures.
- Dimensions define the structure of the cube, and measures provide aggregated numerical values of interest to the end user.
- A calculated member is a dimension member whose worth is calculated at run time using a sign that you specify when you define the calculated member.
- Browsing a deployed cube helps you understand the modifications that you should make to improve the functionality of the cube.
- Perspective and Language options are available on the toolbar of the Browser tab.
- The Browser tab includes two panes to the right of the metadata pane: the upper pane is the filter pane, and the lower pane is the data pane.

7.7 Keywords

Cube: A Cube is the main object of a SSAS solution where the majority of fine tuning, calculations, aggregation design etc.

Cube Wizard: Use this wizard to create a cube. The wizard helps you select the data source, fact table, measures, and dimensions for a new cube.

7.8 Review Questions

1. What is the Cube Wizard?
2. Give the explanation for creating cube.
3. "A Cube is the main object of a SSAS solution". Do you agree? Discuss.
4. Write down the concept of cube.
5. Discuss about adding measure and measure groups to a cube.
6. What are the steps for removing a measure group from a cube?
7. "Calculated member can also be defined as assesses". Elaborate.
8. Give explanation for individual panes in cube designer.

Answers: Self Assessment

1.	SSAS solution	2.	Cube Wizard
3.	Cube	4.	Stored
5.	Measure Name	6.	Measure Group
7.	True	8.	True
9.	Cube data, Dimension data	10.	Filter pane, data pane

7.9 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). " *Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

http://publib.boulder.ibm.com/infocenter/db2luw/v8/topic/com.ibm.dwe.tutorial.doc/tutolaplescube.htm

pic.dhe.ibm.com/infocenter/ctm1/.../t_dsk_model_create_dim.html?

technet.microsoft.com/en-us/magazine/ee677579.aspx?

www.jinfonet.com/kbase/jreport8.1/tutorial/Trail3/cube.htm?

Unit 8: Advanced Measures and Calculations

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain advanced measures and calculations
- Discuss aggregate functions
- Construct retrieving data from a cube using mdx queries
- Learn about Calculation scripting
- Explain creating KPIs

Introduction

In this unit, you will learn advanced measures and calculations. The aggregate functions like SUM, MIN, MAX and COUNT are discussed. It also explains how to use MDX to conditionally apply formatting to a measure or calculated member and retrieve data from cube. As the unit progress you will learn about calculation scripts. Finally, you will learn to create Key Performance Indicators (KPIs) that combine expressions and graphics for actual, target, status, and trend values.

8.1 Advanced Measures and Calculations

The focus of this is the utilization of advanced methods of exploring and surfacing OLAP cube data using Multidimensional Expression Language (MDX), both in the Enterprise Guide viewer and via the PROC SQL interface to OLAP. Before moving to the advanced content, a brief review of OLAP principles and terminology is necessary to provide some context. Since cube navigation within the Enterprise Guide viewer plays an integral part in gaining the MDX knowledge necessary for advance OLAP cube data manipulation, it too will be reviewed.

Aggregate tables store pre-computed results, which are measures that have been aggregated (typically summed) over a set of dimensional attributes. Using aggregate tables is a very popular technique for speeding up query response times in decision support systems. This eliminates the need for run-time calculations and delivers faster results to users. The calculations are done ahead of time and the results are stored in the tables. Aggregate tables should have many fewer rows than the non-aggregate tables, and therefore, processing should be quicker.

Often, a business wants to compare values of a measure and needs a calculation to express the comparison. Oracle BI Server has a calculation engine to perform a multitude of calculations. Calculation measures allow end users to ask business questions like "Show me the accounts receivable balance as of Q3" or "Show me the difference between units ordered and units shipped."

Once the groundwork has been laid, MDX queries and the use of several MDX and SAS functions within those queries will be demonstrated.

Notes The examples provided will allow you to customize the OLAP cube report data and leverage the potential analytical insights made available through this medium.

Self Assessment

Fill in the blanks:

1. .... store pre-computed results, which are measures that have been aggregated (typically summed) over a set of dimensional attributes.
2. .... Server has a calculation engine to perform a multitude of calculations.

8.2 Aggregate Functions

It is very common to sum measures when you aggregate values along dimension hierarchies, but sometimes you need to apply a different aggregation method.

Example: If you want to calculate average sales per customer, you divide total sales by number of customers. You can sum sales amount to get total sales, but to get the number of customers, you need to count customers, making sure to count each customer only once, regardless of how many purchases each customer has made.

Suppose you want to analyse the overall gross margin for every product in your data source. One way to do this is to create a new calculated field called Margin that is equal to the profit divided by the sales. Then you could place this measure on a shelf and use the predefined summation aggregation. Here, Margin is defined as:

Margin = SUM([Profit]/ [Sales])

This formula calculates the ratio of profit and sales for every row in the data source, and then sums the numbers.

Caution However, this is almost certainly not what you would have intended because summing ratios is generally not useful.

Notes

Instead, you probably want to know the sum of all profits divided by the sum of all sales. That formula is shown below:

Margin = SUM([Profit]) / SUM([Sales])

In this case, the division is performed after each measure is aggregated. An aggregate calculation allows you to create formulas like this.

Table 8.1 shows list of aggregate functions with descriptions.

Table 8.1: Common Aggregate Functions

Aggregate Function	Category	Description
Sum	Additive	The value of a parent member is the sum of the values of its children. Sum is the default aggregate function.
Count	Additive	Counts number of rows in a fact table where are particular column is non-empty or counts fact table rows. The value of a parent member can also be calculated by summing of the values of its children.
Min	Pseudo-additive	The value of a parent member is the minimum value of its children.
Max	Pseudo-additive	The value of a parent member is the Maximum value of its children.
Distinct Count	Non-additive	Counts unique values of a column in the fact table. The value of a member is determined by counting unique values for the member.
None	Non-additive	No aggregations are performed.
First Child	Semi-additive	The value of a parent member is the sum of the value of its children, except for a member in the Time dimension. In the Time dimension, the value of a parent member is the value of its first child.
Last Child	Semi-additive	The value of a parent member is the sum of the value of its children, except for a member in the Time dimension. In the Time dimension, the value of a parent member is the value of its last child.
First Non-empty	Semi-additive	The value of a parent member is the sum of the value of its children, except for a member in the Time dimension. In the Time dimension, the value of a parent member is the value of its first non-empty child.
Last Non-empty	Semi-additive	The value of a parent member is the sum of the value of its children, except for a member in the Time dimension. In the Time dimension, the value of a parent member is the value of its last non-empty child.
Average of Children	Semi-additive	The value for a member derived by summing along all dimensions at the lowest level of granularity of the cube's time dimension and then averaging.
ByAccount	Semi-additive	The ByAccount aggregate function is used when the cube contains an account type dimension. The aggregate function applied to the measure is a property of the members of the Account dimension.

Source: http://www.anzmall.com/node/89

Let us understand the aggregate functions using an example. The cube that these examples use has a single measure, Sales, based on the Sales_Amount column in the Sales fact table. The cube has three dimensions:

- Customers, based on the table Customers and containing these levels from highest to lowest:
    - ❖ (All)

Notes

- Customer with Customer_Name as the member name column and Customer_ID as the member key column

- Retail Stores, based on the table Retail_Stores and containing these levels from highest to lowest:
  - (All)
  - Retail Store with Retail_Store_Name as the member name column and Retail_Store_ID as the member key column
- Products, based on the table Products and containing these levels from highest to lowest:
  - (All)
  - Product Category with Product_Category as the member name column and the member key column
  - Product with Product_Name as the member name column and Product_ID as the member key column

The cube's fact table Sales is:

Table 8.2: Fact Table Sales

Transaction_ID	Customer_ID	Product_ID	Retail_Store_ID	Sales_ Amount
1	1	1	1	300
2	1	1	1	250
3	1	1	1	250
4	1	2	1	100
5	1	4	1	700
6	2	1	2	290
7	2	2	2	90
8	2	3	3	510
9	3	1	4	350
10	3	2	3	110
11	4	3	4	550
12	4	4	4	750

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Cube's dimension tables Customers, is:

Table 8.3: Dimension Table Customers

Customer_ID	Customer_Name	Customer_Address_ Line_1	Customer_Address_ Line_2
1	A	1 A Street	Aville, AA 55555
2	B	2 B Street	Bville, BB 55555
3	C	3 C Street	Cville, CC 55555
4	D	4 D Street	Dville, DD 55555

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Notes

Another cube's dimension tables, Retail_Stores, are shown below.

Table 8.4: Dimension Table Retail_Stores

Retail_ Store_ID	Retail_Store_ Name	Retail_Store_ Address_Line_1	Retail_Store_ Address_Line_2
1	A	1 A Avenue	Atown, AA 55555
2	B	2 B Avenue	Btown, BB 55555
3	C	3 C Avenue	Ctown, CC 55555
4	D	4 D Avenue	Dtown, DD 55555

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

The cube's final dimension table, Products, is shown here.

Table 8.5: Dimension Table Products

Product_ID	Product_Name	Product_Description	Product_Category
1	A	aaaa aaaa aaaa	AB
2	B	bbbb bbbb bbbb	AB
3	C	cccc cccc cccc	CD
4	D	dddd dddd dddd	CD

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Sum

If a measure's Aggregate Function property value is Sum, the measure value for a cube cell is calculated by adding the values in the measure's source column from only the rows for the combination of members that defines the cell and the descendants of those members.

The following examples return values that represent accumulated Sales:

- A query on the Sales measure for customer A, retail store A, and product A returns 800.
- A query on the Sales measures for customer A, retail store A, and product category AB returns 900.
- A query on the Sales measure places each retail store on the x-axis, nests products under product categories on the y-axis, and slices by All Customers. It returns the result as shown in Table 8.6.

Table 8.6: Result of Query

			All Retail Stores	A	B	C	D
All Products			4250	1600	380	620	1650
	AB		1740	900	380	110	350
		A	1440	800	290		350
		B	300	100	90	110	
	CD		2510	700		510	1300
		C	1060			510	550
		D	1450	700			750

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Min

If a measure's Aggregate Function property value is Min, the measure value for a cube cell is calculated by taking the lowest value in the measure's source column from only the rows for the combination of members that defines the cell and the descendants of those members.

The following examples return values that represent the lowest Sales price:

- A query on the Sales measure for customer A, retail store A, and product A returns 250.
- A query on the Sales measures for customer A, retail store A, and product category AB returns 100.
- A query on the Sales measure places each retail store on the x-axis, nests products under product categories on the y-axis, and slices by All Customers. It returns the result as shown in Table 8.7.

Table 8.7: Result of Query

			All Retail Stores	A	B	C	D
All Products			90	100	90	110	350
	AB		90	100	90	110	350
		A	250	250	290		350
		B	90	100	90	110	
	CD		510	700		510	550
		C	510			510	550
		D	700	700			750

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Max

If a measure's Aggregate Function property value is Max, the measure value for a cube cell is calculated by taking the highest value in the measure's source column from only the rows for the combination of members that defines the cell and the descendants of those members.

Table 8.8: Result of Query

			All Retail Stores	A	B	C	D
All Products			750	700	290	510	750
	AB		350	300	290	110	350
		A	350	300	290		350
		B	110	100	90	110	
	CD		750	700		510	750
		C	550			510	550
		D	750	700			750

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

The examples return values that represent the highest Sales price.

- A query on the Sales measure for customer A, retail store A, and product A returns 300.
- A query on the Sales measures for customer A, retail store A, and product category AB returns 300.

- A query on the Sales measure places each retail store on the x-axis, nests products under product categories on the y-axis, and slices by All Customers. It returns the result as shown in Table 8.8.

Count

If a measure's Aggregate Function property value is Count, the measure value for a cube cell is calculated by adding the number of values in the measure's source column from only the rows for the combination of members that defines the cell and the descendants of those members.

The following examples return values that represent the number of Sales transactions.

- A query on the Sales measure for customer A, retail store A, and product A returns 3.
- A query on the Sales measures for customer A, retail store A, and product category AB returns 4.
- A query on the Sales measure places each retail store on the x-axis, nests products under product categories on the y-axis, and slices by All Customers. It returns the result as shown in Table 8.9.

Table 8.9: Result of Query

			All Retail Stores	A	B	C	D
All Products			12	5	2	2	3
	AB		8	4	2	1	1
		A	5	3	1		1
		B	3	1	1	1	
	CD		4	1		1	2
		C	2			1	1
		D	2	1			1

Source: http://msdn.microsoft.com/en-us/library/ms365396.aspx

Self Assessment

Fill in the blanks:

3. .... = SUM([Profit]/ [Sales])
4. The division is performed after each measure is ....
5. If a measure's Aggregate Function property value is Sum, the measure value for a cube cell is calculated by adding the values in the ....
6. If a measure's Aggregate Function property value is ...., the measure value for a cube cell is calculated by taking the lowest value in the measure's source column.
7. If a measure's Aggregate Function property value is Max, the measure value for a cube cell is calculated by taking the .... value in the measure's source column.

8.3 Retrieving Data from a Cube Using MDX Queries

To use facts and figures from an analysis Services cube in your report, you should define an Analysis Services data source and create one or more report datasets. When you define the data source definition, you should specify an attachment string and credentials so that you can get access to the data source from your client computer.

Did u know? You can create embedded data source definition for use by a single report or a shared data source definition that can be utilised by multiple reports.

The methods in this topic recount how to create an embedded data source. To create an embedded Microsoft SQL Server Analysis Services data source:

1. On the toolbar in the Report Data pane, click New, and then click Data Source.
2. In the Data Source Properties dialog box, type a name in the Name text box, or accept the default name.
3. Verify that Embedded connection is selected.
4. From the Type drop-down list, select Microsoft Sql Server Analysis Services.
5. Specify a connection string that works with your Analysis Services data source. Contact your database administrator for connection information and for the credentials to use to connect to the data source. The following connection string example specifies the AdventureWorksDW database on the local client:

   Data Source = local host; Initial Catalog = Adventure Works DW
6. Click Credentials and then Click OK. The data source appears in the Report Data pane.

8.4 Calculation Scripting

A script command is an MDX script, included as part of the definition of the cube. Script commands let you perform almost any action that is supported by MDX on a cube, such as scoping a calculation to apply to only part of the cube. In SQL Server 2005 Analysis Services (SSAS), MDX scripts can apply either to the whole cube or to specific sections of the cube, at specific points throughout the execution of the script. The default script command, which is the CALCULATE statement, populates cells in the cube with aggregated data based on the default scope.

The default scope is the whole cube, but you can define a more limited scope, known as a subcube, and then apply an MDX script to only that particular cube space. The SCOPE statement defines the scope of all subsequent MDX expressions and statements in the calculation script until the scope is terminated or redefined. The THIS statement is then used to apply an MDX expression to the current scope. You can use the BACK_COLOR statement to specify a background cell colour for the cells in the current scope, to help you during debugging.

Types of MDX Scripts

There are two types of MDX scripts:

1. ***The default MDX script:*** At the time that you create a cube, Analysis Services creates a default MDX script for that cube. This script defines a calculation pass for the whole cube.
2. ***User-defined MDX script:*** After you have created a cube, you can add user-defined MDX scripts that extend the calculation capabilities of the cube.

Task "Script commands let you perform almost any action that is supported by MDX on a cube". Comment.

Self Assessment

Fill in the blanks:

8. A .... command is an MDX script, included as part of the definition of the cube.

9. The default scope is the whole cube, but you can define a more limited scope, known as a .... and then apply an MDX script to only that particular cube space.

10. There are two types of MDX scripts: .... and ....

8.5 Creating KPIs

KPIs are measurements that define and track specific business goals and objectives that often roll up into larger organizational strategies that require monitoring, improvement, and evaluation. Begin the process of meeting your organization's goals by defining KPIs for monitoring sales and margins.

To create or edit a KPI:

1. ***Open a KPI for editing*** or create a new KPI. To create a new KPI, do one of the following:

   (a) In the global header, hover the mouse pointer over the New menu, select KPI, and from the Select Subject Area dialog, select a subject area for the KPI. The "KPI editor" is displayed.

   (b) From a scorecard, go to the Scorecard Documents pane or the Catalogue pane, click the New Object icon list, select KPI, and from the Select Subject Area dialog, select a subject area for the KPI. The "KPI editor" is displayed.

2. ***On the "KPI editor:*** General Properties page", specify the business owner, actual value, and target value, and indicate if you want to enable trending to determine performance patterns.

3. ***On the "KPI editor:*** Dimensionality page", select the dimensions (for example, Sales by Region and by Financial Quarter) that you want to use to aggregate the KPI's actual and target values.

> *Notes* Note that you should include a time dimension for most KPIs. Exceptions include constants or metrics that are defined as current snapshots, such as "Inventory on Hand" or "Current Phone Support Wait Time."

4. ***On the "KPI editor: Thresholds page"***, indicate the desired goal based on KPI values (for example, "High Values are Desirable"), define the ranges that evaluate KPI values to determine performance status, and associate performance levels with actions.

5. ***On the "KPI editor:*** Related Documents page", add any external links or business intelligence objects to the KPI.

6. ***Save the KPI.*** Note the following items:

   (a) If you are creating a stand-alone KPI, then click Finish to save the KPI.

   (b) If you are creating a new KPI, then the "Save As dialog" is displayed where you specify the KPI's name and where you want to save the KPI. If you want the KPI to

Notes

display within a scorecard's "Scorecard Documents pane", then save the KPI to the scorecard object's folder within the catalogue.

(c) If you are creating a KPI from a scorecard, then click Save from the "Scorecard editor".

Self Assessment

Fill in the blanks:

11. .... are measurements that define and track specific business goals and objectives that often roll up into larger organizational strategies.

12. Open a KPI for .... a new KPI.

Case Study

Business Intelligence Project Pitfalls

Most of us have heard stories of business intelligence failures. I assure you that it is rare for technology to cause the failure. Unfortunately, it is usually the "softer" issues that bring down the project. Here is a list of definite project pitfalls. By understanding these pitfalls, hopefully you will avoid them altogether or at least decrease their effects when confronted with them.

A recent Nucleus Research report lists their top five IT mistakes in generic IT projects. These reasons include:

Customization Overkill

Every business intelligence project is subjected to constantly changing requirements. This can be managed by using a scope document to hold peoples' feet to the fire. Trouble arises when we begin customizing vendor applications or those standard reporting applications that we created to better match individual needs. According to the report, too much customization rarely increases usability, but it surely will increase initial and ongoing costs. You get the one-two punch with overly customized applications. Initially, you will spend more on consulting and projects will take longer to build. Secondly, the applications will require more consulting and personnel to support and maintain. The business must have a good justification for customizing a vendor's application or your in-house analytical capability.

Lack of Training

We would all like to think that our business intelligence applications are so intuitive and easy to use that no training is needed. If that is the case, then why hasn't everyone in your organization embraced your business intelligence environment? There are many other related processes or cultural changes that are needed for full utilization and adoption of your business intelligence applications. The report relates Salesforce.com stories in which the technology was deployed with no user training. The companies ended up either abandoning the tool (which was unfair to the vendor) or worse, investing significantly in changing bad behaviour later because the users made up their own procedures for using the tool. Please listen to your vendor when they offer estimates for the type and length of training needed to use their tool. Then make sure you only train those employees with a "need to know," rather than making the entire company goes through the training.

Contd....

No Executive Involvement

This form of support is especially critical for business intelligence projects that require collaboration between different organizational groups. This is because you are changing the way the business users do their work. The business users become hesitant, maybe even hostile toward the new technology and application. You must have a C-level executive who is willing to be the first guinea pig, to be the first person to adopt the new environment. This executive will make your application the only way they receive such information; which will create incentives for others to adopt the new system and replace workers who flatly refuse to play. A clear list of benefits, along with the metrics that will be used to measure the benefits, will go a long way. Thus, the executive will become comfortable and want to get on board with the project.

Lack of Communication to Consultants

Make sure that you and your consultants understand who is doing what. Agree on specific roles, responsibilities, costs and estimated time frames before you initiate the project. If things get off track, do not wait to call a meeting to determine the problems and potential solutions. A scope document for the project and another for the consultants' roles and responsibilities will eliminate a large heartache later. Specific deliverables assigned to each person on the project, as well as their time frames, will benefit you greatly.

Assuming the Project is finished

The report successfully points out that a project is not over simply because the application has been deployed. It should end when it is being effectively used by the business. This is especially true of business intelligence applications. It may take a while before the business community actually uses the new application. It may be because other processes must be implemented before the analytics can be fully utilized. It is important to understand how the application fits into the business user's workflow before declaring that the project is completed. Your project may not truly end for several months, perhaps even years, after it has been organized.

Question:

Add your own comments about these in terms of specific business intelligence projects.

Source: http://www.b-eye-network.com/view/1519

8.6 Summary

- The focus of this is the utilization of advanced methods of exploring and surfacing OLAP cube data using Multidimensional Expression Language (MDX), both in the Enterprise Guide viewer and via the PROC SQL interface to OLAP.
- Once the groundwork has been laid, MDX queries and the use of several MDX and SAS functions within those queries will be demonstrated.
- It is very common to sum measures when you aggregate values along dimension hierarchies, but sometimes you need to apply a different aggregation method.
- If a measure's Aggregate Function property value is Sum, the measure value for a cube cell is calculated by adding the values in the measure's source column from only the rows for the combination of members that defines the cell and the descendants of those members.
- To use facts and figures from an analysis Services cube in your report, you should define an Analysis Services data source and create one or more report datasets.

Notes

- You can create embedded data source definition for use by a single report or a shared data source definition that can be utilised by multiple reports.
- A script command is an MDX script, included as part of the definition of the cube. Script commands let you perform almost any action that is supported by MDX on a cube, such as scoping a calculation to apply to only part of the cube.
- The default scope is the whole cube, but you can define a more limited scope, known as a sub cube, and then apply an MDX script to only that particular cube space.
- KPIs are measurements that define and track specific business goals and objectives that often roll up into larger organizational strategies that require monitoring, improvement, and evaluation.

8.7 Keywords

Aggregate Functions: In computer science, an aggregate function is a function where the values of multiple rows are grouped together as input on certain criteria to form a single value of more significant meaning or measurement such as a set, a bag or a list.

Calculated Member: A calculated member is a dimension member whose value is calculated at run time using an expression that you specify when you define the calculated member.

Key Performance Indicators (KPIs): Key Performance Indicators are valuable for teams, managers, and businesses to evaluate quickly the progress made against measurable goals.

Multidimensional Expressions (MDX): Multidimensional Expressions is the query language that you use to work with and retrieve multidimensional data in Microsoft SQL Server 2005 Analysis Services (SSAS).

Script: A script command is an MDX script, included as part of the definition of the cube.

8.8 Review Questions

1. Briefly explain the advanced measures and calculations.
2. What is an aggregate function?
3. Discuss about retrieving data from a cube using MDX queries.
4. What is script command?
5. What are the various types of MDX Scripts?
6. Discuss the process of creating or editing a KPI.

Answers: Self Assessment

1. Aggregate tables
2. Oracle BI
3. Margin
4. Aggregated
5. Measure’s source column
6. Min
7. Highest
8. Script
9. Subcube
10. Default MDX script, User-defined MDX script
11. KPIs
12. Editing or create

8.9 Further Readings

Notes

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

msdn.microsoft.com/en-us/library/dd239327(v=sql.100).aspx

quartetfs.com/en/mdx-query-basics-and-usage-example

www.bidn.com › Home › Blogs › DustinRyan

Deepak Mehta,Lovely Professional University

Unit 9: Advanced Dimensional Design

CONTENTS

Objectives

After studying this unit, you will be able to:

- Discuss Creating Dimensions
- Explain Build an Account Dimension to Support Financial Analysis
- Interpret interacting with a Cube

Introduction

In this unit, you will learn about advanced dimensional designs. Multiple hierarchies can be created for a dimension to provide alternate views of dimension members. The account dimension and its associated rules enable you to create and maintain a chart of accounts for various financial models. Here, you will learn about creating dimensions. You will also be able to build an account dimension to support financial analysis. Each base schema design technique brings limitations and implications for aggregate design. Finally, you will learn about interacting with cubes. In this the key focus will be on implementing actions, creating standard actions and creating a drillthrough action.

9.1 Creating Dimensions

Multiple hierarchies can be created for a dimension to provide alternate views of dimension members.

Example: A time dimension that has two hierarchies can comprise of a normal calendar view and a fiscal calendar view.

In Microsoft® SQL Server™ 2000 Analysis Services, a dimension with multiple hierarchies is actually two or more distinct dimensions that can share dimension tables and may share the same aggregations.

Different dimensions with a single hierarchy is called schema. It requires a time span to indicate the presence of more than one hierarchy.

> *Notes* When creating dimensions with multiple hierarchies, the hierarchy part of the name should not be same to any current or future level name or member name in the dimension because queries using the dimension may be ambiguous.

To help minimize disruption to cubes, it is helpful to identify dimensions with multiple hierarchies before they are established. One way to do this is to name the dimension with a time span and a hierarchy title part at the time of creation. Additional hierarchies can then be created by utilising the identical dimension name part followed by the period and the hierarchy title part.

Dimensions that have multiple hierarchies can be created in the Dimension Wizard or Dimension Editor. For each hierarchy that is being created, the method is alike to creating a new dimension.

To create a dimension with a single defined hierarchy using the Dimension Wizard:

1. In the Analysis Manager Tree pane, expand the database in which you want to create a dimension with multiple hierarchies.
2. Right-click the Shared Dimensions folder, point to New Dimension, and then click Wizard.
3. In the second step of the wizard select either Star Schema: A single dimension table or Snowflake Schema: Multiple, related dimension tables.
4. Follow the remaining wizard steps to define levels and various options for the dimension.
5. In the Finish step of the wizard, enter a name in the Dimension name box.
6. Select the Create a hierarchy of a dimension box.
7. Enter a name in the Hierarchy name box.
8. Click Finish to complete the wizard. After you complete the wizard, Dimension Editor appears so that you can further refine the dimension.
9. (Optional.) To create another hierarchy of the dimension, from the File menu in Dimension Editor, point to New Dimension, and then click Wizard. Follow the steps in the next procedure, "To create a dimension with additional defined hierarchies using the Dimension Wizard," beginning with Step 3.

To create a dimension with additional defined hierarchies using the Dimension Wizard:

1. In the Analysis Manager Tree pane, expand the database in which you want to define additional hierarchies for a dimension with at least one named hierarchy.
2. Right-click the Shared Dimensions folder, point to New Dimension, and then click Wizard.
3. In the second step of the Dimension Wizard select either Star Schema: A single dimension table or Snowflake Schema: Multiple, related dimension tables.
4. Follow the remaining wizard steps to define levels and various options for the dimension.
5. Select the Create a hierarchy of a dimension box.

Notes

6. Select a dimension name having a defined hierarchy from the Dimension name box.
7. Enter a name in the Hierarchy name box.
8. Click Finish to complete the wizard. After you complete the wizard, Dimension Editor appears so that you can further refine the dimension.
9. (Optional.) To create another hierarchy of the dimension, from the File menu in Dimension Editor, point to New Dimension, and then click Wizard. Repeat Steps 3 through 8.

Self Assessment

Fill in the blanks:

1. Different dimensions with a single hierarchy are called ....
2. Dimensions that have multiple hierarchies can be created in the ....

9.2 Build an Account Dimension to Support Financial Analysis

The Account dimension and its associated rules enable you to create and maintain a chart of accounts for various financial models.

In Microsoft SQL Server Analysis Services, an account type dimension is a dimension whose attributes represent a journal of accounts for financial reporting reasons. An account dimension permits you selectively manage aggregation across various accounts over time. An account dimension also lets you use a benchmark means to resolve most of the non-standard aggregation issues typically came across in business understanding solutions that handle financial data. If you did not have such a standard mechanism, settling these nonstandard aggregation issues would need Multidimensional Expression (MDX) scripts.

The following table describes the pre-defined properties of Account dimension members:

Table 9.1: Properties of Account Dimension Members

Property	Description
Account Type Member ID	A selectable option that groups the account member into a type, such as *Tax Expense* or *Liability*, for use with business rules. Planning Business Modeler uses this property to determine the aggregation behaviour for accounts. When using this property in a business rule calculation, you must explicitly reference the property as Account Type Member ID. Do not use a substitute such as "Account type".
Debit Credit	Indicates whether, for calculation purposes such as aggregation, Planning Business Modeler treats this account type as a debit entry or a credit entry. When used in calculation, an account type that has a debit name has a negative sign for calculation, and an account type that has a credit entry has a positive sign.
Time Balance	Indicates how Planning Business Modeler handles values in this account type for aggregation. The following values are possible: ● Sum – aggregation value is the sum of all child members. ● End – aggregation value is the last nonempty child along the Time dimension. Also called *Last-child aggregation*. ● Avg – aggregation value is the average value of member children.

Contd....

Notes

Consolidated	Boolean (TRUE-FALSE) value that indicates whether this account type should be included in the Planning Business Modeler consolidation calculations.
Converted	Boolean (TRUE-FALSE) value that indicates whether this account type should be included in the Planning Business Modeler currency conversion calculations.
Inter-company	Boolean (TRUE-FALSE) value that indicates whether this account should be included in intercompany calculations.

Source: http://msdn.microsoft.com/en-us/library/bb795367(v=office.12).aspx

User Modifications in the Account Dimension

You can update the Account Type Member ID property of a member of the Account dimension. You can also add a new dimension member to the Account dimension and create member hierarchies in the dimension. In addition, you can also add member sets, member views, and member properties. The following table shows the ways you can modify user-defined objects in the Account dimension.

Table 9.2: User-defined Objects

User-defined object	Create	Update	Delete
Dimension members	Yes	Yes	Yes
Dimension properties	No	Yes	No
Member sets	Yes	Yes	Yes
Member views	Yes	Yes	Yes
Member properties	Yes	Yes	No

Source: http://msdn.microsoft.com/en-us/library/bb839292(v=office.12).aspx

Self Assessment

State whether the following statements are true or false:

3. In Microsoft SQL Server Analysis Services, an account type dimension is a dimension whose attributes represent a journal of accounts for financial reporting reasons.

4. If you did not have such a standard mechanism, settling these nonstandard aggregation issues would need Multidimensional Expression (MDX) scripts.

9.3 Interacting with a Cube

9.3.1 Implementing Actions

The purpose of an Online Analytical Processing (OLAP) application is to supply users with valuable information to propel business conclusions. Actions supply another means by which users can accumulate information and take steps based on the data they find in cubes. You can add activities to a cube that users will subsequent execute. An action is habitually started by a user or client application and relates to an object in a cube. That object might be a dimension member or a specific cell, which is then used as a parameter for the activity. Not all client applications are able to execute actions, so make certain you realise the capability of the client application before creating actions in your cube.

You can add several kinds of activities to a cube. A URL action is helpful for navigating to a particular World Wide Web location based on cube facts and figures.

Example: You might desire to visit a customer's Web location after examining that customer's data in a cube, or you might desire to get access to information from an internal reporting Web server to get more data about a specific product you're analysing.

9.3.2 Creating Standard Actions

The Cube Designer in Business Intelligence Development Studio (BIDS) includes an activity tab. You can define an action on this tab by specifying the activity name, activity goal, the action type, and the action sign that generates a string used to run the activity. An activity goal is the portion of the cube to which the activity connects and is the object that the client clicks to launch the action. The activity sign is a Multidimensional Expression (MDX) sign that evaluates as a string applicable to the activity kind. Each activity kind has its own syntax requirements, but usually you include the MDX CURRENTMEMBER function in the action expression to link the object to the current cube context.

Did u know? In this method, we will add a new URL activity that opens a World Wide Web page and executes a search for a product category or subcategory.

Create a URL action

1. Use Business Intelligence Development Studio (BIDS) to open the AdventureWorks BI solution. You will need to establish the AdventureWorks SSAS database before you can create an activity. This ensures that the database created is present in BIDS and on the analysis Services server are both the identical.
2. On the build menu, choose Deploy AdventureWorks SSAS. If the AdventureWorks SSAS database currently lives on the server, a dialog carton may emerge warning that the data-groundwork will be overwritten. If the warning appears, click Yes. The prior version of the facts and figures-groundwork will be deleted and the current deployment will extend.
3. After the database has been successfully deployed and processed, expand the Cubes folder in Solution Explorer, right-click the AdventureWorks.cube, and select View Designer.
4. In the Cube Designer, click the activities tab. On the activities tab toolbar, click New Action.
5. In the Action Editor, change the name of the action to Internet Search.
6. An action target is the location in the cube where the action can be executed. An action target has a target type and a target object. You can choose from several target types.

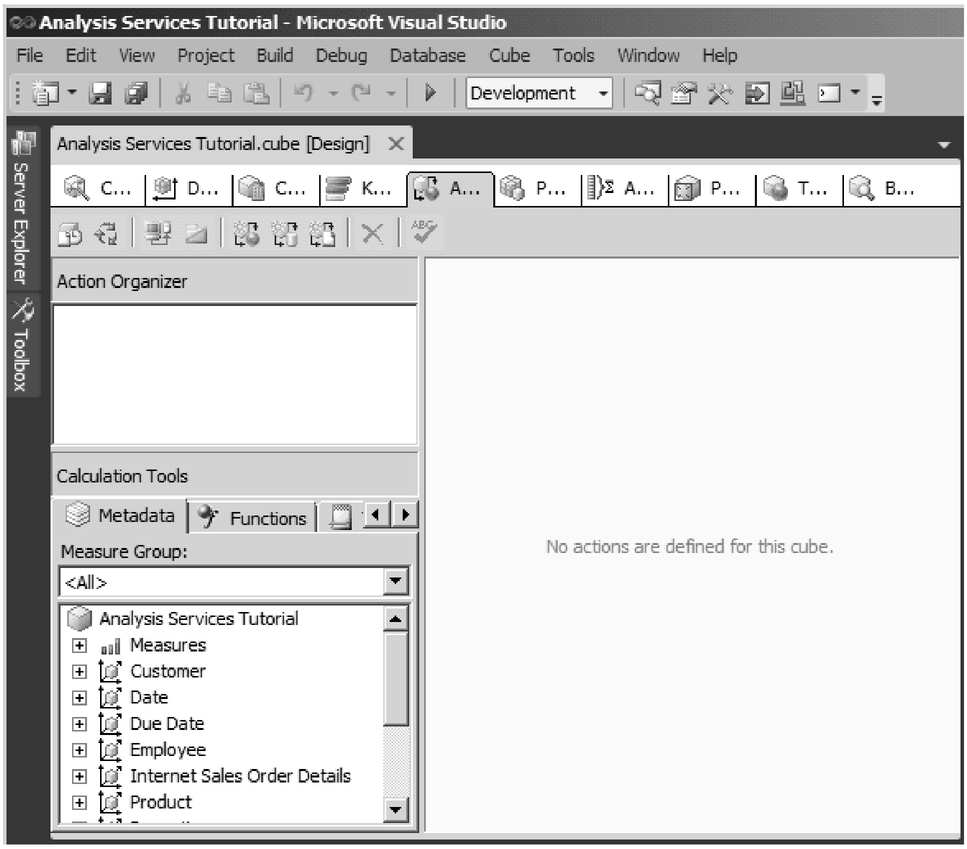
Example: If you select Cube, the action is available for all cube objects—every dimension, hierarchy, level, and member.

7. Expand the Target Object List, expand the product dimension, choose product by Category, and click OK. You can enter an MDX sign in the status text box to further limit the scope of the target.
8. In the Condition text box, enter the following MDX expression:

```
[Product].[Product by Category].Level IS [Product].[Product by Category].[Category] OR
[Product].[Product by Category].Level IS [Product].[Product by Category].[Subcategory]
```

You need to select the type of action that you want to create.

Figure 9.1: Activity Tab Toolbar


Analysis Services Tutorial - Microsoft Visual Studio
File Edit View Project Build Debug Database Cube Tools Window Help
Development
Server Explorer
Toolbox
Analysis Services Tutorial.cube [Design]
C... D... C... K... A... P... A... P... T... B...
Action Organizer
Calculation Tools
Metadata
Functions
Measure Group:
<All>
Analysis Services Tutorial
Measures
Customer
Date
Due Date
Employee
Internet Sales Order Details
Product
No actions are defined for this cube.

Source: http://i.msdn.microsoft.com/dynimg/IC574400.gif

9. In the Action Content section of the editor, verify that URL is selected from the Type list. The Action Expression text box includes the string that will be passed to the application that is begun by the action.

10. In the Action Expression text box, enter the following MDX expression:

```
"http://search.live.com/results.aspx?q="
+ [Product].[Product by Category].CurrentMember.Name
+ "&form=QBLH"
```

Now before you can execute the action, you need to deploy your project. After the project is successfully deployed, you can browse the cube and execute the URL action that you just created.

9.3.3 Creating Drillthrough Actions

Drillthrough actions supply fast access to lowest grade of details stored in a cube. When you create a drillthrough action, you choose dimension attributes and assesses that are returned as columns of data when the action is performed. When a client examining summary value executes

Notes

the activity, the client application executes the drillthrough query supplied by analysis Services to return and display a set of rows containing the detailed data behind the summary value. Contrary to the action name, a drillthrough action does not get access to data stored in the source relational database.

Caution Any data that you want to be available for drillthrough should be comprised in the cube's dimensions and measures.

9.3.4 Linking to Reporting Services Reports

When Reporting Services is part of your Business Intelligence (BI) infrastructure, you can effortlessly create activities that execute these reports. After the report is established, you can create an action that executes the report.

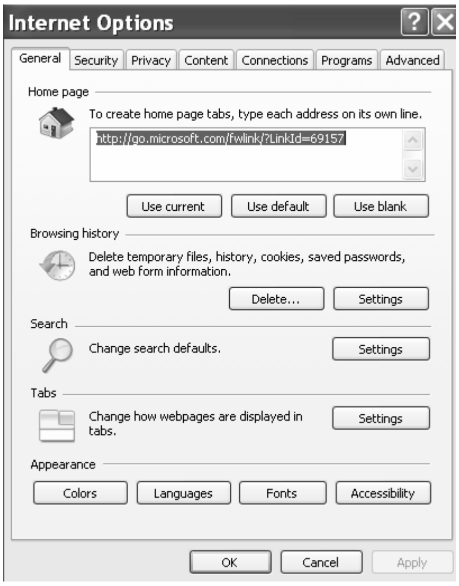
The enhanced security environment of Windows 7 needs that you modify the default Internet Explorer security configuration if you want to deploy a report to the local instance of reporting Services.

Notes You must furthermore add yourself to the Reporting Services Content Manager security role.

We will now configure Internet Explorer Security to allow reporting service.

Configure Internet Explorer Security to allow Reporting Services administration

1. On the Microsoft Windows task bar, click Start, select All Programs, right-click Internet Explorer, and select Run as Administrator.
2. In the User Account Control dialog box, select Allow.
3. From the Internet Explorer Tools menu, select Internet Options.



To create home page tabs, type each address on its own line.

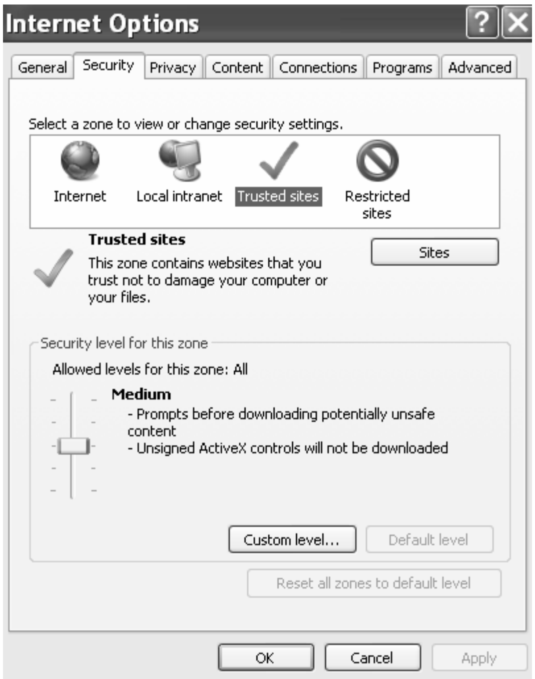
http://go.microsoft.com/fwlink/?LinkId=69157

Delete temporary files, history, cookies, saved passwords, and web form information.

Change search defaults.

Change how webpages are displayed in tabs.

4. In the Internet Options dialog box, click the Security tab, select Trusted Sites, and then click Sites.



Select a zone to view or change security settings.

Trusted sites

This zone contains websites that you trust not to damage your computer or your files.

Security level for this zone

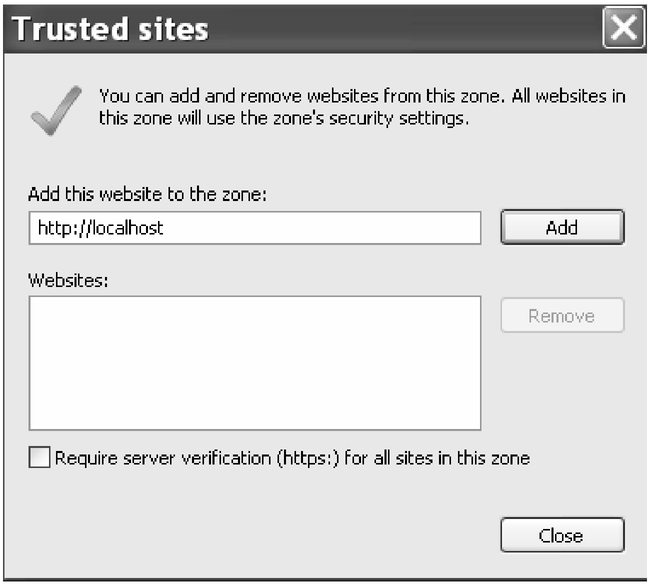
Allowed levels for this zone: All

Medium

- Prompts before downloading potentially unsafe content
- Unsigned ActiveX controls will not be downloaded

5. In the Trusted Sites dialog box, clear the Require Server Verification (https:) For All Sites in This Zone check box.

6. In the Add This Website to the Zone text box, type http://localhost.



You can add and remove websites from this zone. All websites in this zone will use the zone's security settings.

Add this website to the zone: http://localhost

Websites:

□ Require server verification (https:) for all sites in this zone

7. Click Add and then click Close. In the Internet Options dialog box, click OK.

Task Prepare a dimension wizard to create the account dimension.

Notes

Self Assessment

Fill in the blanks:

5. The purpose of an .... application is to supply users with valuable information to propel business conclusions.

6. A URL action is helpful for navigating to a particular World Wide Web location based on .... and figures.

7. The Cube Designer in Business Intelligence Development Studio (BIDS) includes an ....

8. .... actions supply fast access to lowest grade of details stored in a cube.

9. Contrary to the action name, a drillthrough action does not get access to data stored in the source ....

10. After the report is established, you can create an action that executes the ....

Case Study

MITS Provides Custom Solution for Florida Healthcare Maintenance Organization (HMO)

Company Profile:

Name:

Preferred Medical Plan, Inc.

Distinction:

Licensed HMO provider, boasting the largest individual HMO membership in Florida.

Vertical Software Provider:

DST Systems, Inc.

Database:

IBM UniVerse

Company Background

Preferred Medical Plan, Inc. (PMP) is an independently owned, licensed HMO in Southern Florida with 170 employees providing individuals with a range of comprehensive medical, surgical and hospital benefits including dental and optical riders. PMP, Inc. has the largest individual HMO membership in Florida and is committed to providing its members with more choices and benefits than any other individual health plan. In September of 2002, Consumer Reports awarded PMP, Inc. with its highest coverage index for offering the most comprehensive set of benefits.

The Challenge

Due to the nature of their industry, PMP, Inc. has many internal and external information reporting needs. Since healthcare insurance is so highly regulated, they must comply with external state reporting requirements such as enrolment data, HEIDIS, Child Health Check Up and HIV/AIDS statistics, as well as fraud detection and prevention efforts. Internally,

Contd....

there is a need to utilize information to predict and manage expenses, making it feasible to provide the most comprehensive benefits packages possible to members at a reasonable cost. PMP Inc's Information Services Project Manager, John J. Burns, PMP describes how the company relies on efficient access to data: "From a production point of view, we're in the claim adjudication business. We need to determine which type of claim or which types of providers are submitting claims requiring the most effort to adjudicate."

To access their data, PMP, Inc. relied on traditional reporting mechanisms and practices which are heavily dependent on Information Services (IS) resources. John J. Burns provided insight into the common maintenance and development of reports. "Reports that were developed internally by IS staff or canned system reports not customized to our specific needs were run using the same transactional database that was being used in production. There were times when this would severely affect system performance." This impact on the production system was most evident at the beginning of the month, when most reports are run, and resulted in the production system being brought to a crawl. The drain on the IS resources was also significant due to the amount of time needed to create custom reports or modify existing reports, decreasing the amount of time IS staff had available to devote to other projects.

The Solution

When looking into reporting solutions, PMP, Inc. was concerned about some of the issues associated with maintaining large sets of data in a data warehouse environment. Through his experience with the MultiValue community, Burns was aware of MITS products and the advanced reporting and analytics MITS Discover offered. MITS Discover presented a way to access near real time data without need for a dedicated data warehouse. During further review of MITS Discover Burns realized "[the] Discover hypercube concept eliminated these concerns. We [also] anticipated that the browser-based interface would reduce training time and combined with the ability to export directly into a spreadsheet or Adobe PDF documents would empower the end user and not burden the limited resources of our IS Department."

PMP, Inc. wanted to find out what types of claims were costing the most to process and who was filing them. Additionally, they needed to perform comparative analysis, such as differences between geographical regions, average cost per claim type and cost per claim by each of their submitting providers. Burns worked with a MITS Product Specialist to design and implement their first set of Hypercubes to provide this analysis.

As a result of past experience, Burns mentioned, "often a solution based on how a product functions is offered or even forced on a client." He did not find this to be the case with MITS and was impressed with the amount of time the MITS staff took to understand PMP's business processes and unique needs. He went on to say, "MITS took the time to thoroughly comprehend what we hoped to accomplish and how MITS Discover could meet most of those needs. [The MITS Product Specialist] also clearly stated what the tool was not designed to do, and knowing that actually helped us get exactly what we required out of the design of our Hypercubes."

The Results

While the rollout of MITS Discover to the majority of PMP's users is still in process, they are currently using MITS within their Claims, Fraud and Abuse, Utilization, and IS departments. One example of improved access to data is in their Turnaround Hypercube. Claims have a 30 day turnaround for payment and with this Hypercube they are able to identify where outstanding claims stand in their queue. Most claims are adjudicated

Contd....

between day 14 and day 17, resulting in a bell curve distribution of the data. This metric has proved to be so useful that they will be including it, in its graphical form, as a key performance indicator (KPI) on their MITS Discover Dashboard, allowing them to instantly view shifts in the curve and attend to underlying issues. In its beginning stages the response to MITS within PMP, Inc. has been very positive, with users already providing feedback for additional Hypercubes. Burns followed up on this by commenting, "To me, this is a very encouraging sign. It signals that with very little exposure to the tool, end users are seeing its potential in ways that would benefit them and how it can allow them to produce the analysis and reports that they require." He has also found that "[MITS] Discover easily uncovers inaccuracies and inconsistencies that can exist in the underlying data, sometimes to the doubt of the user. Verification, however, proves [MITS] Discover to be correct and our assumptions of the data to be wrong. I like to say that if you shine a flashlight in a darkened room and the roaches scurry, don't blame the flashlight!"

Questions:

1. Analyse the case and provide a solution to Preferred Medical Plan, Inc. (PMP).
2. What is Turnaround Hypercube?

Source: http://www.mits.com/solutions/success-stories/healthcare.html

9.4 Summary

- In Microsoft® SQL Server™ 2000 Analysis Services, a dimension with multiple hierarchies is actually two or more distinct dimensions that can share dimension tables and may share the same aggregations.
- Dimensions that have multiple hierarchies can be created in the Dimension Wizard or Dimension Editor.
- The Account dimension and its associated rules enable you to create and maintain a chart of accounts for various financial models.
- You can update the AccountTypeMemberID property of a member of the Account dimension. You can also add a new dimension member to the Account dimension and create member hierarchies in the dimension.
- The purpose of an Online Analytical Processing (OLAP) application is to supply users with valuable information to propel business conclusions.
- The Cube Designer in Business Intelligence Development Studio (BIDS) includes an activity tab.
- Drillthrough actions supply fast access to lowest grade of details stored in a cube. When you create a drillthrough action, you choose dimension attributes and assesses that are returned as columns of data when the action is performed.

9.5 Keywords

AccountTypeMemberID: A selectable option that groups the account member into a type.

Business Intelligence Development Studio (BIDS): Business Intelligence Development Studio (BIDS) is the IDE from Microsoft used for developing data analysis and Business Intelligence solutions utilizing the Microsoft SQL Server Analysis Services, Reporting Services and Integration Services.

Notes

Cube Designer: The Cube Designer in Business Intelligence Development Studio (BIDS) includes an activity tab.

Multidimensional Expression sign (MDX): Multidimensional Expressions (MDX) is a query language for OLAP databases, much like SQL is a query language for relational databases. It is also a calculation language, with syntax similar to spreadsheet formulas.

URL actions: A URL action is a hyperlink that points to a Web page, file, or other web-based resource outside of Tableau.

9.6 Review Questions

1. What is a data dimension?
2. Why do we create the data dimensions?
3. Discuss the process for creating a dimension with a single defined hierarchy using the dimension wizard.
4. Explain the steps for creating a dimension with additional defined hierarchies using the dimension wizard.
5. Describe the pre-defined properties of account dimension members.
6. Write about the user modifications in the account dimension.
7. How do we implement the actions in OLAP?
8. How do we create a URL action?
9. Write down the process for creating drillthrough actions.
10. Discuss how to configure internet explorer security to allow reporting services administration?

Answers: Self Assessment

1. Schema
2. Dimension Wizard or Dimension Editor
3. True
4. True
5. Online Analytical Processing (OLAP)
6. Cube facts
7. Activity tab
8. Drillthrough
9. Relational database
10. Report

9.7 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Notes

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). *"Business Intelligence"*. John Wiley & Sons.

Swain Scheps (2013). *"Business Intelligence for Dummies"*. Wiley.

Online links

http://www.oracle.com/technetwork/articles/bi/advanced-dimensional-design-359613.html

www.atlantamdf.com/Presentations/AtlantaMDF_091106.pdf?

www.mhprofessional.com/downloads/.../01-ch01_0071770445.pdf?

Unit 10: Retrieving Data from Analysis Services

Notes

CONTENTS

Objectives

After studying this unit, you will be able to:

- Provide insight into creating perspectives
- Explain the Multidimensional Expressions (MDX) queries
- Discuss the usage of Excel with analysis services

Introduction

Your cube comprises numerous dimensions and measures from some subject areas: sales, finance, output, and inventory. It can assist as a source of information for multiple workgroups or agencies across your association. Although this centralization is beneficial for users, application developers, and IT administrators, it might be difficult for users to find their way to the data they need. To help make your cube simpler to comprehend and navigate, you can create perspectives that limit the number of dimensions, calculations, actions, and KPIs that users see when they are browsing or querying a cube. A perspective allows you to create a view of a subset of a cube that can be more easily comprehended by users. This unit will provide you an insight into creating perspectives. You will learn about Multidimensional Expression (MDX) queries. Finally, you will learn about connecting Excel client to Analysis Services environment.

10.1 Creating Perspectives

A cube may contain multiple viewpoints. Generally each viewpoint displays a subset of the cube that is related to a widespread subject area or that is needed by a group of users. To describe and visualize applications, a viewpoint appears just like a cube. Although, it is important to note that you cannot request security to a viewpoint.

> *Notes* If a client has access to a cube, the client has access to all of the perspectives in that cube.

10.1.1 Create a Perspective

For models that contain many subject areas, for example, Sales, Manufacturing, and Supply data, it might be helpful to Report Builder users if you create perspectives of the model. A perspective is a sub-set of a model.

Did u know? Creating perspectives can make navigating through the contents of the model easier for your model users.

To create a perspective:

1. In the Tree view, right-click Model, point to New, and then click Perspective.
2. In the Edit Perspective dialog box, click Clear All.
3. Locate the Purchase Order Detail entity, and then select its check box.
4. To add all the attributes of the Purchase Order Header entity to the perspective, clear the check box, and then select the check box again.
5. Locate the Product entity, clear the check box, and then select the check box again.
6. Click OK.

To rename the perspective

1. To see the new perspective, scroll down to the bottom of the List view. The last item listed is called New Perspective.
2. Right-click New Perspective, and then click Rename.
3. Type Products and Purchases and on the File menu, click Save All.

Self Assessment

Fill in the blanks:

1. To describe and visualize applications, a viewpoint appears just like a ....
2. A perspective is a .... of a model.

10.2 Multidimensional Expressions (MDX) Queries

An MDX query is different from an MDX expression. An expression is a formula that calculates a single value. A query is a command that can get numerous values from a cube, generally to create a report. The cube browser in the Cube Designer, like other describing and data visualization applications utilised with Analysis Services, allows you to drag and drop dimensions and measures and generates MDX queries behind the scenes to get values from a cube. You can use these tools to create accounts without composing any MDX queries of your own. Learning MDX will permit you to take advantage of some of the more advanced features of Analysis Services to create precisely the dataset you need. You can use MDX queries to understand and debug complex MDX signs.

The most widespread use of an MDX query is to extract values from an OLAP cube to populate a report. A cube has dimensions, but a report does not. Reports have axes. An axis can include members from more than one dimension.

Caution A report generally doesn't show all of the data comprised in a cube.

The basic Multidimensional Expressions (MDX) query is the SELECT statement.

10.2.1 SELECT Statement Syntax

The following syntax shows a basic SELECT statement that includes the use of the SELECT, FROM, and WHERE clauses:

```
[ WITH <SELECT WITH clause> [ , <SELECT WITH clause> ... ] ]
SELECT [ * | ( <SELECT query axis clause>
 [ , <SELECT query axis clause> ... ] ) ]
FROM <SELECT subcube clause>
[ <SELECT slicer axis clause> ]
[ <SELECT cell property list clause> ]
```

The MDX SELECT statement supports optional syntax, such as the WITH keyword, the use of MDX functions and the ability to return the values of specific cell properties as part of the query.

10.2.2 SELECT Statement Example

The following example shows a basic MDX query that uses the SELECT statement. This query returns a result set that contains the 2010 and 2011 sales and tax amounts for the North sales territories.

```
SELECT
  { [Measures].[Sales],
  [Measures].[Tax] } ON COLUMNS,
  { [Date].[Fiscal].[Fiscal Year].&[2010],
  [Date].[Fiscal].[Fiscal Year].&[2011] } ON ROWS
FROM [Adventure Works]
WHERE ( [Sales Territory].[North] )
```

In this example, the query defines the following result set information:

- The SELECT clause sets the query axes as the Sales and Tax members of the Measures dimension, and the 2010 and 2011 members of the Date dimension.
- The FROM clause indicates that the data source is the Adventure Works cube.
- The WHERE clause defines the slicer axis as the North member of the Sales Territory dimension.

Notice that the query example also uses the COLUMNS and ROWS axis aliases. The ordinal positions for these axes could also have been used.

Example: The following example shows how the MDX query could have been written to use the ordinal position of each axis:

```
SELECT
  { [Measures].[Sales],
```

Notes

```
  [Measures].[Tax] } ON 0,
  { [Date].[Fiscal].[Fiscal Year].&[2010],
  [Date].[Fiscal].[Fiscal Year].&[2011] } ON 1
FROM [Adventure Works]
WHERE ( [Sales Territory].[North] )
```

Self Assessment

State whether the following statements are true or false:

3. An MDX query is different from an MDX expression.
4. Learning MDX will permit you to take advantage of some of the more advanced features of Analysis Services to create precisely the dataset you need.
5. The MDX SELECT statement supports optional syntax.

10.3 Excel with Analysis Services

You can use Analysis Services as a data source for the Office Excel 2007 PivotTable and PivotChart characteristics. The Excel 2007 PivotTable characteristic permits you to create reports and crosstab reports that will let you pivot, filter, add and remove dimensions, drill down, drill up, and drillthrough data. You can use the Excel 2007 PivotChart feature to create powerful data visualizations. Excel 2007 presents various features you can use to format and analyse data with PivotTable:

- Improve the appearance of your report by hiding field headers and using the expand/ collapse buttons.
- View empty rows and columns.
- Execute Analysis Services actions.
- Sort and filter data.
- Display member properties.
- Display KPIs.
- View and modify data sources.

Analysis Services provides dimensional data that is well-suited for data exploration in PivotTables and Power View reports. You can get Analysis Services data from:

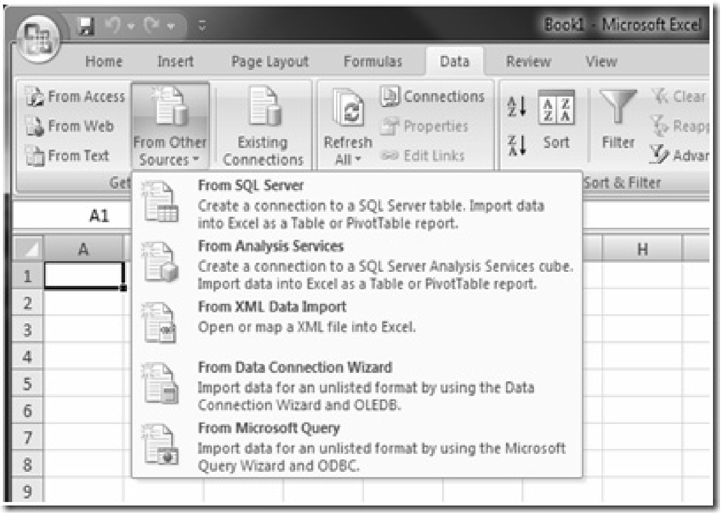
- OLAP cubes on an Analysis Services multidimensional server.
- Tabular models on an Analysis Services tabular server.
- Excel 2013 workbooks on SharePoint 2013, if the workbook contains a data model.
- PowerPivot workbooks on SharePoint 2010.

10.3.1 Connecting Excel Client to Analysis Services Environment

From within Excel, select the Analysis Services drop down from the Data tab -> From Other Sources drop down, and then walk through the data connection wizard to identify location, cube, and credentials.

Notes

Figure 10.1: Excel Accesses Analysis Services



From SQL Server
Create a connection to a SQL Server table. Import data into Excel as a Table or PivotTable report.

From Analysis Services
Create a connection to a SQL Server Analysis Services cube. Import data into Excel as a Table or PivotTable report.

From XML Data Import
Open or map a XML file into Excel.

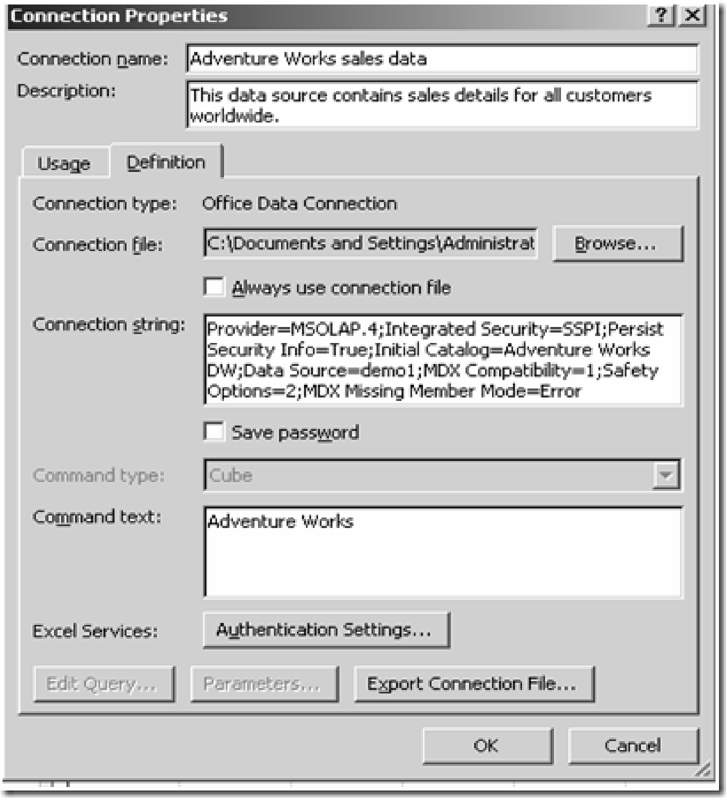
From Data Connection Wizard
Import data for an unlisted format by using the Data Connection Wizard and OLEDB.

From Microsoft Query
Import data for an unlisted format by using the Microsoft Query Wizard and ODBC.

*Source:*http://blogs.msdn.com/blogfiles/excel/WindowsLiveWriter/UsingExcelExcelServiceswithSQLServerAnal_B11A/image_thumb.png

You should now have a connection to a cube within Excel and a pivot table ready.

Figure 10.2: Excel Connection Property



Connection name: Adventure Works sales data

Description: This data source contains sales details for all customers worldwide.

Connection type: Office Data Connection

Connection file: C:\Documents and Settings\Administrat

□ Always use connection file

Connection string: Provider=MSOLAP.4;Integrated Security=SSPI;Persist Security Info=True;Initial Catalog=Adventure Works DW;Data Source=demo1;MDX Compatibility=1;Safety Options=2;MDX Missing Member Mode=Error

□ Save password

Command type: Cube

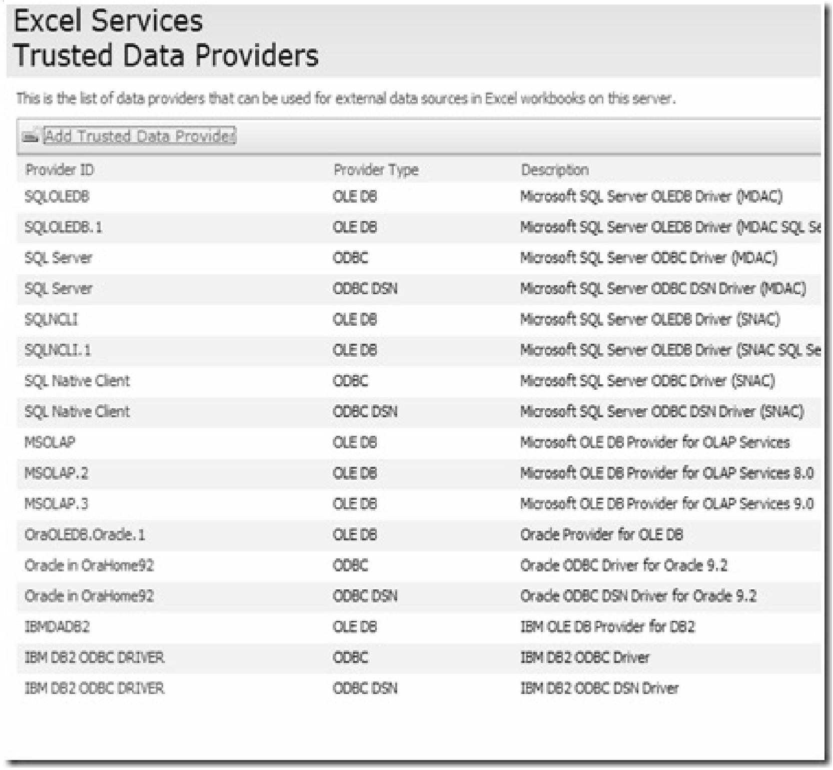
Command text: Adventure Works

Excel Services: Authentication Settings...

Source: http://blogs.msdn.com/blogfiles/excel/WindowsLiveWriter/UsingExcelExcelServiceswithSQLServerAnal_B11A/image_thumb_1.png

You’ll notice that connecting to Analysis Services 2008 uses MSOLAP.4. To add the provider to the list of approved providers in Excel Services, go to Central Administration -> Shared Services Administration for Excel Services. Select Trusted Data Providers from the Excel Services Settings of the Shared Services Administration page. Add MSOLAP.4 to the trusted list in order for the connection to work in Excel Services.

Figure 10.3: Excel Services Data Provider



Excel Services
Trusted Data Providers

This is the list of data providers that can be used for external data sources in Excel workbooks on this server.

Provider ID	Provider Type	Description
SQLOLEDB	OLE DB	Microsoft SQL Server OLEDB Driver (MDAC)
SQLOLEDB.1	OLE DB	Microsoft SQL Server OLEDB Driver (MDAC SQL Se
SQL Server	ODBC	Microsoft SQL Server ODBC Driver (MDAC)
SQL Server	ODBC DSN	Microsoft SQL Server ODBC DSN Driver (MDAC)
SQLNCLI	OLE DB	Microsoft SQL Server OLEDB Driver (SNAC)
SQLNCLI.1	OLE DB	Microsoft SQL Server OLEDB Driver (SNAC SQL Se
SQL Native Client	ODBC	Microsoft SQL Server ODBC Driver (SNAC)
SQL Native Client	ODBC DSN	Microsoft SQL Server ODBC DSN Driver (SNAC)
MSOLAP	OLE DB	Microsoft OLE DB Provider for OLAP Services
MSOLAP.2	OLE DB	Microsoft OLE DB Provider for OLAP Services 8.0
MSOLAP.3	OLE DB	Microsoft OLE DB Provider for OLAP Services 9.0
OraOLEDB.Oracle.1	OLE DB	Oracle Provider for OLE DB
Oracle in OraHome92	ODBC	Oracle ODBC Driver for Oracle 9.2
Oracle in OraHome92	ODBC DSN	Oracle ODBC DSN Driver for Oracle 9.2
IBMDADB2	OLE DB	IBM OLE DB Provider for DB2
IBM DB2 ODBC DRIVER	ODBC	IBM DB2 ODBC Driver
IBM DB2 ODBC DRIVER	ODBC DSN	IBM DB2 ODBC DSN Driver

Source: http://blogs.msdn.com/blogfiles/excel/WindowsLiveWriter/UsingExcelExcelServiceswithSQLServerAnal_B11A/image_thumb_2.png

Click Add Trusted Data Provider at the top of the list, and enter

```
Provider ID = MSOLAP.4
Data Provider Type = OLE DB
Description = Microsoft OLE DB Provider for OLAP Services 10.0.
```

You are now set and can publish your workbook to Excel Services and you will be able to view, interact and refresh data from Analysis Services 2008 in Excel Services.

Task Compare the Syntax of the MDX SELECT Statement to SQL.

Self Assessment

Fill in the blanks:

6. You can use Analysis Services as a data source for the .... PivotTable and PivotChart characteristics.

7. Analysis Services provides dimensional data that is well-suited for data exploration in PivotTables and ....

8. Connecting to Analysis Services 2008 uses ....

Case Study

Data Services Firm Uses Microsoft BI and Hadoop to Boost Insight into Big Data

Klout wanted to give consumers, brands, and partners faster, more detailed insight into hundreds of terabytes of social-network data. It also wanted to boost efficiency. To do so, Klout deployed a business intelligence solution based on Microsoft SQL Server 2012 Enterprise and Apache Hadoop. As a result, Klout processes data queries in near real time, minimizes costs, boosts efficiency, increases insight, and facilitates innovation.

Solution

In 2011, Klout decided to implement a BI solution based on Microsoft SQL Server 2012 Enterprise data management software and the open-source Hive data warehouse system. "When it comes to BI and analytics, open-source tool sets are just ineffective and there's really not a good choice," Mariani says. "Instead, Klout chose the best of both worlds by marrying the Microsoft BI platform with Hadoop and Hive." Based on employees' previous experience with the Microsoft BI platform, Klout also knew that SQL Server offers excellent compatibility with third-party software and it can handle the data scale and query performance needed to manage big-data sets.

In August 2011, engineers implemented a data warehouse with Hive, which consolidates data from all of the network silos hosted by Hadoop. In addition, Klout deployed SQL Server 2012 on a system that runs the Windows Server 2008 R2 Enterprise operating system to take advantage of Microsoft SQL Server 2012 Analysis Services. Engineers use it to manage all business logic required to facilitate Multidimensional Online Analytical Processing (MOLAP). Data is stored in multidimensional cubes, which helps preserve detail and speed analysis. To provide high availability, Klout replicates the database to a secondary system using SQL Server 2012 AlwaysOn.

At the time that Klout was initially deploying its solution, SQL Server 2012 and Hive could not communicate directly. To work around this issue, engineers set up a temporary relational database that runs MySQL 5.5 software. It includes data from the previous 30 days and serves as a staging area for data exchange and analysis. Klout engineers are currently working to implement the new open database connectivity driver in SQL Server 2012 to directly join Hive with SQL Server 2012 Analysis Services. In addition, to enhance insight Klout plans to work with Microsoft to incorporate other Microsoft BI tools into its solution, such as Microsoft SQL Server Power Pivot for Microsoft Excel.

Questions:

1. Analyse the case and provide any other solution to the problem.

2. Discuss the benefits of implementing business intelligence solution based on Microsoft SQL Server 2012 by Klout.

Source: http://www.microsoft.com/en-us/sqlserver/product-info/case-studies/klout.aspx

10.4 Summary

- A cube may contain multiple viewpoints. Generally each viewpoint displays a subset of the cube that is related to a widespread subject area or that is needed by a group of users.
- For models that contain many subject areas, for example, Sales, Manufacturing, and Supply data, it might be helpful to Report Builder users if you create perspectives of the model.
- An MDX query is different from an MDX expression. An expression is a formula that calculates a single value. A query is a command that can get numerous values from a cube, generally to create a report.
- You can use Analysis Services as a data source for the Office Excel 2007 PivotTable and PivotChart characteristics.
- Analysis Services provides dimensional data that is well-suited for data exploration in PivotTables and Power View reports.
- From within Excel, select the Analysis Services drop down from the Data tab -> From Other Sources drop down, and then walk through the data connection wizard to identify location, cube, and credentials.
- Select Trusted Data Providers from the Excel Services Settings of the Shared Services Administration page. Add MSOLAP.4 to the trusted list in order for the connection to work in Excel Services.

10.5 Keywords

MDX SELECT: The MDX SELECT statement supports optional syntax, such as the WITH keyword, the use of MDX functions and the ability to return the values of specific cell properties as part of the query.

Multidimensional Expressions (MDX) query: Multidimensional Expressions (MDX) is a query language for OLAP databases, much like SQL is a query language for relational databases. It is also a calculation language, with syntax similar to spreadsheet formulas.

Multidimensional Expressions (MDX): Multidimensional Expressions is the query language that you use to work with and retrieve multidimensional data in Microsoft SQL Server 2005 Analysis Services (SSAS).

Perspective: A perspective is a sub-set of a model.

10.6 Review Questions

1. What is perspective?
2. Write down the steps for creating a perspective.
3. Briefly explain the Multidimensional Expressions (MDX) Queries.
4. Discuss about SELECT Statement Syntax.
5. Write down the various features you can use to format and analyse data with PivotTable in Excel 2007.
6. Write short note on excel services data provider.

Notes

Answers: Self Assessment

1. Cube
2. Sub-set
3. True
4. True
5. True
6. Office Excel 2007
7. Power View reports
8. MSOLAP.4

10.7 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

quartetfs.com/en/mdx-query-basics-and-usage-example?

www.oracle.com/technetwork/.../bi.../mdx-complex-queries-130019.pdf?

www.slideshare.net/.../mdx-multi-dimensional-expressions-introduction?

wwwbayer.in.tum.de/lehre/SS2001/DAWA-bayer/DWH-Ch10-1.pdf?

Notes

Unit 11: Data Mining

CONTENTS

Objectives

After studying this unit, you will be able to:

- Define data mining
- State data mining approaches
- Identify data mining uses
- Discuss data mining issues
- Demonstrate applications of data mining
- Explain limitations of data mining
- Recognize data mining models
- Discuss basics of data mining algorithms

Introduction

Data mining refers to the extraction of hidden predictive information from large databases. Data mining techniques can yield the benefits of automation on existing software and hardware platforms. Data mining tools can answer business questions that traditionally were too time consuming to resolve. In this unit, you will learn about data mining approaches, uses and its related issues. Also, applications of data mining will be discussed. As the unit progress, you will learn about data mining models – predictive, summary, network and association. Finally, data mining algorithms basics will be introduced.

11.1 Data Mining

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. It uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Some key terms to know before going further detail in Data Mining.

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.

Did u know? This includes operational or transactional data (such as, sales, cost, inventory, payroll, and accounting), non-operational data (such as industry sales, forecast data etc.) and metadata i.e. data about data.

Information

The patterns, associations, or relationships among all types of data can provide information.

Example: Analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into knowledge.

Example: Summary information on supermarket sales can be analysed in view of promotional efforts to provide knowledge of consumer buying behaviour.

11.1.1 Process of Knowledge Discovery

Let us have an overview of the steps one by one:

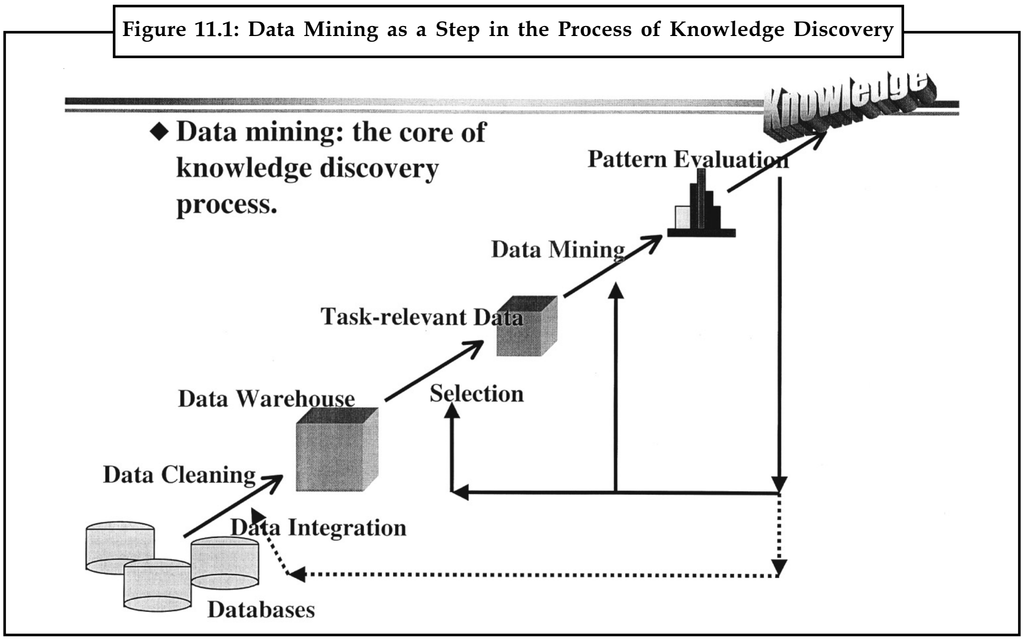
1. ***Data cleaning:*** It refers to removal noise and inconsistent data.
2. ***Data integration:*** In this step, multiple data sources may be combined.
3. ***Data selection:*** In this step, data relevant to the analysis task are retrieved from the database.
4. ***Data transformation:*** In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. ***Data mining:*** This is an essential process where intelligent methods are applied in order to extract data patterns.

6. ***Pattern evaluation:*** This step is used to identify the truly interesting patterns representing knowledge based on some interestingness measures.

7. ***Knowledge presentation:*** In this step, visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Figure 11.1 shows data mining as a step in process of knowledge discovery.

Figure 11.1: Data Mining as a Step in the Process of Knowledge Discovery



Source: http://www.emeraldinsight.com/content_images/fig/0670330804027.png

11.1.2 Types of Data Mining Tasks

Two types of Data mining task are:

1. ***Prediction Methods:*** This type of method uses some variable to predict unknown or future values of other variables.

2. ***Description Methods:*** This type of method finds human-interpretable patterns that describe the data.

11.1.3 Purpose of Data Mining

Data mining enables people to discover information that they can act on to better understand, selectively market to and retain their best customers, or sharply cut consumer fraud.

Caution In this we use pattern recognition logic to identity trends within a sample data set.

Self Assessment

Fill in the blanks:

1. .... is the practice of automatically searching large stores of data to discover patterns and trends that goes beyond simple analysis.

2. .... are any facts, numbers, or text that can be processed by a computer.
3. The patterns, associations, or relationships among all types of data can provide ....

11.2 Data Mining Approaches

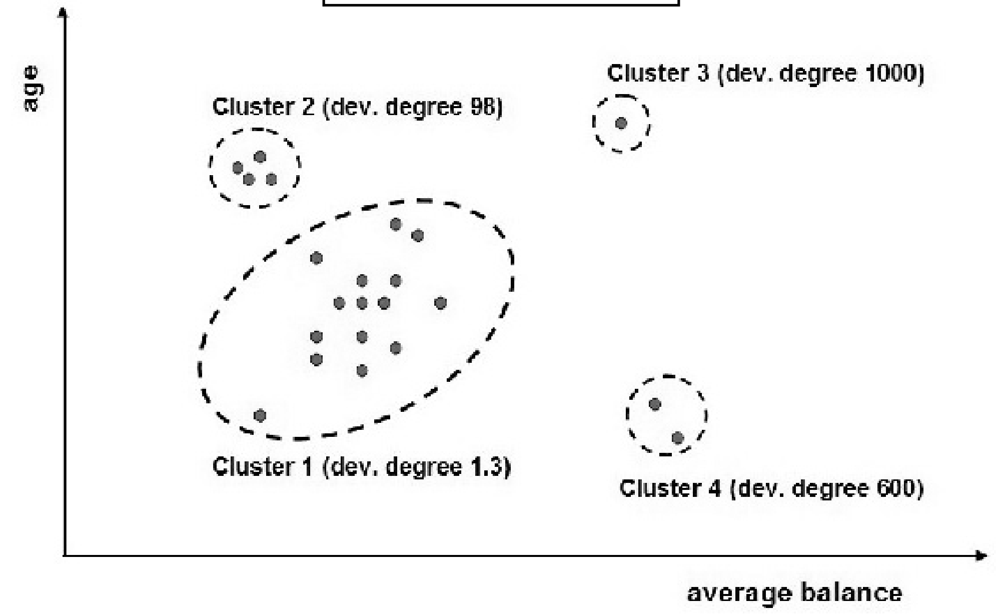
Two widespread data mining methods for finding concealed patterns in data are clustering and classification analysis. Although classification and clustering are often cited in the identical breath, they are different analytical advances.

Imaging a database of customer records, where each record represents a customer's attributes. These can encompass identifiers such as name and address, demographic data such as gender and age, and financial attributes such as income and revenue spent.

Clustering is an automated method to group associated records together. Related records are grouped together on the basis of similar values for attributes.

Notes This approach of segmenting the database via clustering analysis is often used as an exploratory technique because it is not necessary for the end-user/analyst to identify ahead of time how records should be associated simultaneously.

Figure 11.2: Clustering Example



Source: http://www.ibm.com/developerworks/data/library/techarticle/dm-0811wurst/outlier_by_clustering.jpg

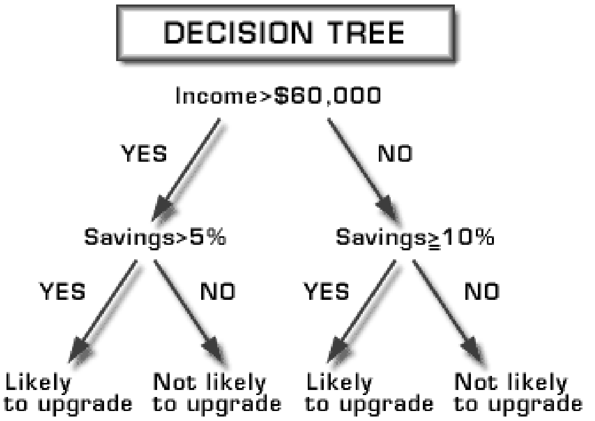
Records inside a cluster are more alike to each other, and more different from records that are in other clusters. Counting on the specific implementation, there is a kind of measure of likeness that is used, but the general aim is for the approach to converge to groups of associated records.

Classification is a different method than clustering. Unlike clustering, a classification analysis requires that the end-user/analyst understand ahead of time how classes are characterised.

Example: Classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No).

A common approach for classifiers is to use decisions trees to partition and segment records. New records can be classified by traversing the tree from the origin through branches and nodes, to a leaf representing a class.

Figure 11.3: Classification Example



Source: http://www.siggraph.org/education/materials/HyperVis/applicat/data_mining/images/tree.gif

Regression

The dependency of one or more independent predictor variables on a single response variable is modelled using regression.

Task Compare and contrast the clustering and classification analysis.

11.3 Data Mining Uses

Data mining is utilised for a variety of reasons in both the personal and public parts. Industries such as banking, insurance, medicine, and retailing commonly use data mining to decrease charges, enhance research, and increase sales.

Example: The protection and banking industries use data mining applications to notice deception and aid in risk evaluation.

Using customer data assembled over several years, businesses can evolve forms that predict if a customer is a good credit risk, or if whether misfortune claim may be fraudulent and should be investigated more neatly. The medical community sometimes utilises data mining to help forecast the effectiveness of a procedure or surgery.

Pharmaceutical companies use data mining of chemical compounds and genetic material to help direct study on new treatments for infections. Retailers can use data assembled through affinity programs (e.g., shoppers' club cards, common flyer points, contests) to consider the effectiveness of product selection and placement conclusions, coupon offers etc. Companies such as phone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are expected to stay as subscribers and which ones are likely to switch to a competitor.

In the public sector, data mining applications were initially used as a means to detect fraud and waste, but now they are used for purposes such as measuring and improving program performance.

Self Assessment

Fill in the blanks:

4. Information can be converted into ....

5. .... refers to removal noise and inconsistent data.

6. .... method uses some variable to predict unknown or future values of other variables.

11.4 Data Mining Issues

Privacy

One of the key matters raised by data mining technology is not an enterprise or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to investigate routine enterprise transactions and glean a significant amount of information about persons buying habits and preferences.

Data Integrity

Clearly, data analysis can only be as good as the data that is being analysed. A key implementation dispute is integrating inconsistent or redundant data from distinct sources.

Example: A bank may sustain credit cards accounts on several distinct databases. The addresses (or even the titles) of a single cardholder may be different in each. Software should convert data from one system to another and choose the recently entered address.

Confusion

Another issue of concern is to decide is if it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is retained in tables, allowing ad hoc queries. In a multidimensional structure, on the other hand, groups of cubes are arranged in arrays, with subgroups created according to category. While multidimensional organisations facilitate multidimensional data mining, relational structures perform better in client/server environments.

Cost

Finally, there is the issue of cost. While system hardware costs have fallen spectacularly inside the past five years, data mining and data warehousing are inclined to be self-reinforcing. The more mighty the data mining queries, the larger the utility of the data being gleaned from the data, and the larger the force to increase the amount of data being assembled and sustained, which increases the pressure for much quicker, more mighty data mining queries. This raises pressure for bigger, much quicker systems, which are more expensive.

Self Assessment

State whether the following statements are true or false:

7. Data analysis can only be as good as the data that is being analysed.
8. Data mining makes it possible to investigate routine enterprise transactions.
9. In a relational structure, data is retained in tables, allowing ad hoc queries.

11.5 Data Mining Applications

Data Mining is a relatively new concept that has not completely matured. Regardless of this, there are a number of industries that are already using it on a normal basis. Some of these companies include retail shops, banks, and insurance firms.

Many of these companies are using data mining for statistics, pattern acknowledgement, and other significant tasks. Data mining can be utilised to find patterns and associations that would else be difficult to find. This concept is popular with many businesses because it permits them to discover more about their customers and make intelligent trading conclusions.

There are a number of applications that data mining has. The first is called market segmentation. With market segmentation, you will be able to find behaviours that are common among your customers. You can look for patterns among customers that appear to buy the same products at the same time.

Another application of data excavation is called customer churn. It will permit you to estimate which customers are most likely to stop purchasing your products or services and proceed to one of your competitors. In addition to this, a company can use data mining to find out which purchases are the most likely to be fraudulent.

Example: By using data mining a retail shop may be able to determine which goods are stolen the most.

By finding out which products are stolen the most, steps can be taken to protect those goods and notice those who are stealing them. You can furthermore use data mining to determine the effectiveness of interactive trading. Some of your customers will be more interested to buy your products online than offline, and you should recognise them.

While many use data mining to boost their profits, many of them don't realize that it can be used to create new businesses.

Example: Assume that you are the owner of a latest gadgets manufacturing company, and you are able to accurately predict the next large-scale latest tendency based on the buying patterns of your customers.

It is very simple to say that you will become very wealthy in a short span of time. You will have an advantage over your competitors. For long-term thinking rather than easily guessing what the next large-scale trend will be, you will be able work out it based on statistics, patterns, and reasoning.

Another example of automatic prediction is to use data mining to look at your past marketing schemes. Which one worked the best? Why did it work the best? Who were the customers that answered most favourably to it? Data mining will allow you to answer these queries, and once you have the responses, you will be able to avoid making any errors that you made in your previous trade.

Data mining can allow you to become better at what you do. A financial organisation such as a bank can predict the number of defaults that will happen among their customers inside a given time, and they can also forecast the amount of deception that will occur as well based on the past records overview.

Another useful application of data mining is the self-acting recognition of patterns that are not discovered before.

Example: If you have a tool that can automatically search your database to look for patterns which are created. If you have access to this technology, you will be able to find relationships that could allow you to make strategic conclusions.

This can lead to growth of organization based on reasoning.

Notes While data mining is a very important tool, it is important to note that it is not a full proof thing. It cannot guarantee the success of you or your business but, it will tilt the odds in your favour.

Self Assessment

Fill in the blanks:

10. .... can be utilised to find patterns and associations that would else be difficult to find.

11. With ...., you will be able to find behaviours that are common among your customers.

11.6 Limitations of Data Mining

Privacy Issues

The concerns about the individual privacy have been increasing enormously recently particularly when internet is booming with social networks, e-commerce, forums, blogs etc. Because of privacy issues, persons are afraid that their personal information is collected and utilised in unethical way that possibly make them face a lot of problems. Businesses collect data about their customers in numerous ways for understanding their buying behaviours trends. A stage may come when the organization may be acquired by other organization or is vanished. At that time the individual data they own likely is be sold to other party or may be leaked.

Security Matters

Security is a large-scale issue. Companies own data about their workers and customers including social security number (like in US), anniversary, payroll etc. although how properly this information is taken care is still doubtful. There have been situation that hackers accessed and robbed big database of customers from large companies.

Notes

Misuse of Data/Inaccurate Data

Data assembled through data mining using ethical purposes can be misused. This information may be exploited by unethical people or companies to take advantage in various ways which not only limits to fake identity proof creation and pass confidential data to competitors. Also if incorrect data is used for decision-making, it will affect the results of the company.

11.7 Data Mining Models

There are several different types of data mining models:

1. ***Predictive models:*** These types of models predict how likely an event is to occur. Usually, higher the score, the more likely the event is to occur.

Example: How likely an online transaction is to be fraudulent, or how likely a railway passenger is to be a thief, or how likely a company is to go bankrupt.

2. ***Summary models:*** These models are used to summarize data.

Example: It can be used to divide online transactions or railway passengers into different groups depending upon their characteristics.

3. ***Network models:*** This type of model represents data by nodes and links.

Example: In a network model describing Facebook friends, nodes might be individuals and directed edges with weights might represent the likelihood that one friend will contact another friend in the next 24 hours.

4. ***Association models:*** Association models are used to find and characterize co-occurrences.

Example: Purchases of certain items, such as soft drink and pizza together will be represented by association models.

Self Assessment

Fill in the blanks:

12. .... type of model represents data by nodes and links.

13. .... are used to find and characterize co-occurrences.

11.8 Data Mining Algorithms

A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. The algorithm first analyses the data you provide, looking for specific types of patterns or trends and then creates the data.

Analysis Services includes the following algorithm types:

- ***Classification algorithms:*** This type of algorithm predicts one or more discrete variables, based on the other attributes in the dataset.
- ***Regression algorithms:*** This type of algorithm predicts one or more continuous variables, such as profit or loss, based on other attributes in the dataset.

Notes

- *Segmentation algorithms:* This type of algorithm divides data into groups, or clusters, of items that have similar properties.
- *Association algorithms:* This type of algorithm finds correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market analysis.
- *Sequence analysis algorithms:* This type summarize frequent sequences or episodes in data, such as a Web path flow.

Self Assessment

Fill in the blanks:

14. A .... is a set of heuristics and calculations that creates a data mining model from data.

15. .... type of algorithm finds correlations between different attributes in a dataset.

Case Study

Logic-ITA Student Data

We have performed a number of queries on datasets collected by the Logic-ITA to assist teaching and learning. The Logic-ITA is a web-based tool used at Sydney University since 2001, in a course taught by the second author. Its purpose is to help students practice logic formal proofs and to inform the teacher of the class progress.

Context of Use

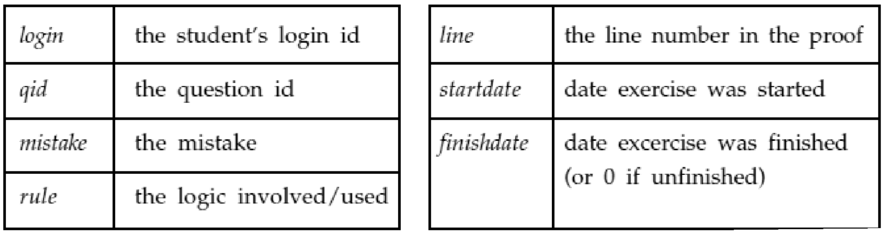
Over the four years, around 860 students attended the course and used the tool, in which an exercise consists of a set of formulas (called premises) and another formula (called the conclusion). The aim is to prove that the conclusion can validly be derived from the premises. For this, the student has to construct new formulas, step by step, using logic rules and formulas previously established in the proof, until the conclusion is derived. There is no unique solution and any valid path is acceptable. Steps are checked on the fly and, if incorrect, an error message and possibly a tip are displayed. Students used the tool at their own discretion. A consequence is that there is neither a fixed number nor a fixed set of exercises done by all students.

Data Stored

The tool's teacher module collates all the student models into a database that the teacher can query and mine. Two often queried tables of the database are the tables mistake and correct_step. The most common variables are shown in Table 1.

Contd....

Table 1: Common Variables in Table's Mistake and Correct_step



login	the student's login id	*line*	the line number in the proof
qid	the question id	*startdate*	date exercise was started
mistake	the mistake	*finishdate*	date excercise was finished (or 0 if unfinished)
rule	the logic involved/used		

Data Mining Performed

Each year of data is stored in a separate database. In order to perform any clustering, classification or association rule query, the first action to take is to prepare the data for mining. In particular, we need to specify two aspects:

1. What element we want to cluster or classify: students, exercises, mistakes?
2. Which attributes and distance do we want to retain to compare these elements?

An example could be to cluster students, using the number of mistakes they made and the number of correct steps they entered. Tada-ed provides a pre-processing facility which allows making the data minable. For instance, the database contains lists of mistakes. If we want to group that information so that we have one vector per student, we need to choose how the mistakes should be aggregated. For instance we may want to consider the total number of mistakes, or the total number of mistakes per type of mistake, or a flag for each type of mistake, and so on.

Data Exploration

Simple SQL queries and histograms can really allow the teacher get a first overview of the class: what were the most common mistakes, the logic rules causing the most problems? What was the average number of exercises per student? Are there any student not finishing any exercise? The list goes on. To understand better how students use the tool, how they practice and how they come to master both the tool and logical proofs, we also analysed data, focussing on the number of attempted exercises per student. In SODAS, the population is partitioned into sets called symbolic objects. Our symbolic objects were defined by the number of attempted exercises and were characterized by the values taken for these newly calculated variables: the number of successfully completed exercises, the average number of correct steps per attempted exercise, the average number of mistakes per attempted exercise. We obtained a number of tables to compare all these objects.

Association Rules

We used association rules to find mistakes often occurring together while solving exercises. The purpose of looking for these associations is for the teacher to ponder and, may be, to review the course material or emphasize subtleties while explaining concepts to students. Thus, it makes sense to have a support that is not too low.

Clustering and Visualization

We applied clustering to try and characterize students with difficulties. We looked in particular at those who attempted an exercise without completing it successfully. To do so, we performed clustering using this subpopulation, both using (i) k-means in TADA- Ed, and (ii) a combination of k-means and hierarchical clustering of Clementine. Because there is neither a fixed number nor a fixed set of exercises to compare students, determining

Contd....

a distance between individuals was not obvious. We calculated and used a new variable: the total number of mistakes made per student in an exercise. As a result, students with similar frequency of mistakes were put in the same group. Histograms showing the different clusters revealed interesting patterns. There are three clusters: 0 (red, on the left), 1 (green, in the middle) and 4 (purple, on the right). From other windows (not shown), we know that students in cluster 0 made many mistakes per exercise not finished, students in cluster 1 made few mistakes and students in cluster 4 made an intermediate number of mistakes. Students making many mistakes use also many different logic rules while solving exercises; this is shown with the vertical, almost solid lines.

Classification

We built decision trees to try and predict exam marks (for the question related to formal proofs). The Decision Tree algorithm produces a tree-like representation of the model it produces. From the tree it is then easy to generate rules in the form IF condition THEN outcome. Using as a training set the previous year of student data (mistakes, number of exercises, difficulty of the exercises, number of concepts used in one exercise, level reached) as well as the final mark obtained in the logic question), we can build and use a decision tree that predicts the exam mark according to the attributes.

Supporting Teachers and Learners

Pedagogical Information Extracted

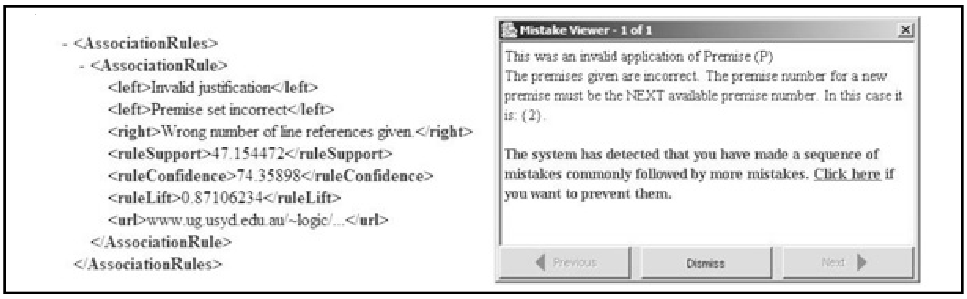
The information extracted greatly assisted us as teachers to better understand the cohort of learners. Whilst SQL queries and various histograms were used during the course of the teaching semester to focus the following lecture on problem areas, the more complex mining was left for reflection between semesters. Symbolic data analysis revealed that if students attempt at least two exercises, they are more likely to do more (probably overcoming the initial barrier of use) and complete their exercises. In subsequent years we required students to do at least 2 exercises as part of their assessment. Mistakes that were associated together indicated to us that the very concept of formal proofs (i.e. the structure of each element of the proof, as opposed to the use of rules for instance) was a problem. In 2003, that portion of the course was redesigned to take this problem into account and the role of each part of the proof was emphasized. After the end of the semester, mining for mistakes associations was conducted again. Surprisingly, results did not change much (a slight decrease in support and confidence levels in 2003 followed by a slight increase in 2004). However, marks in the final exam continued increasing. This leads us to think that making mistakes, especially while using a training tool, is simply part of the learning process and was supported by the fact that the number of completed exercises per student increased in 2003 and 2004. The level of prediction seems to be much better when the prediction is based on exercises (number, length, variety of rules) rather than on mistakes made. This also supports the idea that mistakes are part of the learning process, especially in a practice tool where mistakes are not penalized.

Using data exploration and results from decision tree, one can infer that if students do successfully 2 to 3 exercises for the topic, then they seem to have grasped the concept of formal proof and are likely to perform well in the exam question related to that topic. This finding is coherent with correlations calculated between marks in the final exam and activity with the Logic Tutor and with the general, human perception of tutors in this course. Therefore, a sensible warning system could look as follows: Report to the lecturer-in-charge students who have completed successfully less than 3 exercises. For those students, display the histogram of rules used. Be proactive towards these students, distinguishing those who use out the pop-up menu for logic rules from the others.

Contd....

ITS with Proactive Feedback

Data mining findings can also be used to improve the tutoring system. We implemented a function in Tada-Ed allowing the teacher to extract patterns with a view to integrate them in the ITS from which the data was recorded. Presently this functionality is available for Association Rule module. That is, the teacher can extract any association rule. Rules are then saved in an XML file and fed into the pedagogical module of the ITS. Along with the pattern, the teacher can specify an URL that will be added to the feedback window and where the teacher can design his/her own proactive feedback for that particular sequence of mistakes C (which the student has not yet made).



```
- <AssociationRules>
  - <AssociationRule>
      <left>Invalid justification</left>
      <left>Premise set incorrect</left>
      <right>Wrong number of line references given.</right>
      <ruleSupport>47.154472</ruleSupport>
      <ruleConfidence>74.35898</ruleConfidence>
      <ruleLift>0.87106234</ruleLift>
      <url>www.ug.usyd.edu.au/~logic/...</url>
    </AssociationRule>
  </AssociationRules>
```

Mistake Viewer - 1 of 1

This was an invalid application of Premise (P)
The premises given are incorrect. The premise number for a new premise must be the NEXT available premise number. In this case it is: (2).

The system has detected that you have made a sequence of mistakes commonly followed by more mistakes. Click here if you want to prevent them.

(a) XML encoded patterns **(b) Screen shot of mistake viewer**

The structure of the XML file is fairly simple and is shown in (a). For instance, using our logic data, we extracted the rule saying that if a student makes the mistakes "Invalid justification" followed by "Premise set incorrect" then she/he is likely to make the mistake "Wrong number of references lines given" in a later step (presently there is no restriction on the time window). This rule has a support of 47% and a confidence of 74%. The teacher, when saving the pattern, also entered an URL to be prompted to the student. The pedagogical module of the Logic Tutor then reads the file and adds the rule to its knowledge base. Then, when the student makes these two initial mistakes, she/he will receive, in addition to the relevant feedback on that mistake, an additional message in the same window (in a different colour) advising him/her to consult the web page created by the teacher for this particular sequence of mistakes.

Support for Student Reflection

Extracting information from a group of learners is also extremely relevant to the learner themselves. The fact that learner reflection promotes learning is widely acknowledged. The issue is how to support it well. A very useful way to reflect on one's learning is to look up what has been learned and what has not yet been learned according to a set of learning goals, as well as the difficulties currently encountered. We are seeking here to help learners to compare their achievements and problems in regards to some important patterns found in the class data. For instance, using a decision tree to predict marks, the student can predict his/her performance according to his/her achievements so far and have the time to rectify if needed. Here more work needs to be done to assess how useful this prediction is for the student.

Questions:

1. Discuss how the discovery of different patterns through different data mining algorithms and visualisation techniques suggest you a simple pedagogical policy?

2. Also discuss the behaviour of clustering and cluster visualisation.

Source: books.google.co.in/books?isbn=1586035304

11.9 Summary

- Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases.
- Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis.
- Two widespread data mining methods for finding concealed patterns in data are clustering and classification analysis.
- A common approach for classifiers is to use decisions trees to partition and segment records.
- Data mining is utilised for a variety of reasons in both the personal and public parts.
- One of the key matters raised by data mining technology is not an enterprise or technological one, but a social one.
- Data Mining is a relatively new concept that has not completely matured. Regardless of this, there are a number of industries that are already using it on a normal basis.
- The concerns about the individual privacy have been increasing enormously recently particularly when internet is booming with social networks, e-commerce, forums, blogs etc.
- A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data. The algorithm first analyses the data you provide, looking for specific types of patterns or trends and then creates the data.

11.10 Keywords

Association algorithms: This type of algorithm finds correlations between different attributes in a dataset.

Association models: Association models are used to find and characterize co-occurrences.

Classification algorithms: This type of algorithm predicts one or more discrete variables, based on the other attributes in the dataset.

Data: Data are any facts, numbers, or text that can be processed by a computer.

Data mining: Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis.

Data Mining Algorithm: A data mining algorithm is a set of heuristics and calculations that creates a data mining model from data.

Network models: This type of model represents data by nodes and links.

Predictive models: These types of models predict how likely an event is to occur.

Regression: It is a statistical technique for estimating the relationships among variables.

Regression algorithms: This type of algorithm predicts one or more continuous variables.

Sequence analysis algorithms: This type summarizes frequent sequences or episodes in data.

Summary models: These models are used to summarize data.

Notes

11.11 Review Questions

1. What is data mining?
2. Define data, information and knowledge.
3. Explain the process of knowledge discovery.
4. What are the types of data mining tasks?
5. Discuss purpose of data mining.
6. Briefly explain the data mining approaches.
7. How you can use data mining in Pharmaceutical companies?
8. Elaborate the issues of data mining.
9. Describe the application of data mining.
10. What are the limitations of data mining?
11. Discuss in brief the data mining models.
12. What are data mining algorithms?

Answers: Self Assessment

1. Data mining
2. Data
3. Information
4. Knowledge
5. Data cleaning
6. Prediction
7. True
8. True
9. True
10. Data mining
11. Market segmentation
12. Network models
13. Association models
14. Data mining algorithm
15. Association algorithms

11.12 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Notes

Online links

www.cs.uiuc.edu/~hanj/pdf/ency99.pdf?

www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf

www.stanford.edu/class/stats315b/Readings/DataMining.pdf

Sartaj Singh, Lovely Professional University

Notes

Unit 12: Understanding Data Mining Tools

CONTENTS

Objectives

After studying this unit, you will be able to:

- Identify the data mining tools in SQL server
- Discuss about the mining structure
- Explain the configuring algorithm parameters

Introduction

Data mining is the process of analysing data from different perspectives and summarizing it into useful information. This information can be used to increase revenue, cuts costs, or both. Data mining software analyses relationships and patterns in stored transaction data depending on the open-ended user queries. In this Unit, you will learn about data mining tools. Data mining tools in SQL server will be discussed. Also the unit covers mining structure and models of mining structure. Finally, configuration algorithm parameters will be defined.

12.1 Data Mining Tools in SQL Server

Microsoft SQL Server Analysis Services provides the following tools to create data mining solutions:

- The Data Mining Wizard in SQL Server Data Tools (SSDT) makes it very simple to create mining structures and mining models, using either relational data or multidimensional data in cubes.

Notes In the wizard, you select data to use, and then request exact data mining techniques, such as clustering, neural networks, or time sequence modelling.

- Model viewers are supplied in both SQL Server administrations Studio and SQL Server Data Tools (SSDT), for exploring your mining models after they are created.

Did u know? You can browse models using viewers tailored to each algorithm, or go deeper into analysis by using the model content viewer.

- The Prediction Query Builder is provided in both SQL Server administration Studio and SQL Server Data Tools (SSDT) to help you create prediction queries.

Notes You can also test the accuracy of models against a holdout data set or external data, or use cross-validation to assess the value of your data set.

- SQL Server Management Studio is the interface where you organize living data mining solutions that have been established to an example of Analysis Services. You can reprocess structure and models to update the data in them.
- SQL Server Integration Services includes tools that you can use to clean data, to automate jobs such as creating predictions and updating models, and to create text mining solutions.

Let us discuss more about the data mining tools in SQL Server.

12.1.1 Data Mining Wizard

The Data Mining Wizard in Microsoft SQL Server Analysis Services starts every time when you add a new mining structure to a data mining project. It helps you choose a data source and set up a data source model view that defines the data to be used for analysis, and then helps you create an initial model.

Starting the Data Mining Wizard

To use the Data Mining Wizard, you must have opened a solution in SQL Server Data Tools (SSDT) that contains at least one data mining or OLAP project. Follow these steps:

- If your solution is ready for data mining, you can right-click the Mining Structures node in Solution Explorer and select New Mining Structure to start the wizard.
- If your solution does not contain any existing projects, you can add a new data mining project. To add a new project, from the File menu, select New, and then select Project. Then chose the template, Analysis Services Multidimensional and Data Mining Project.

Relational or OLAP Mining Model

The next decision to make is whether to use a relational data source, or to use an OLAP mining model.

Notes The Data Mining Wizard branches into two paths at this point, depending on whether your data source is relational or in a cube.

Choosing an Algorithm

Next step is to decide on which algorithm to use in processing your data.

Caution Each algorithm provided in Analysis Services has different features and produces different results, so this decision can be difficult to make.

You can experiment and try several different models before determining which is most appropriate for your business problem.

Features of Data Mining Wizard

Data Mining Wizard provides these following features:

- ***Auto-detection of data types:*** The wizard will examine the uniqueness and circulation of column values and then suggest the best data type, and suggest a usage type for the data. You can override these proposals by choosing values from a list.
- ***Suggestions for variables:*** You can click on a dialog box and start an analyser that calculates correlations over the columns included in the model, and determine if any columns are expected predictors of outcome attribute, given the configuration of the form so far.
- ***Feature selection:*** Most algorithms will automatically detect columns that are good predictors and use those preferentially. In columns that comprise too many values, feature selection will be applied, to decrease the cardinality of the data and improve the possibilities for finding a meaningful pattern.
- ***Automatic cube slicing:*** If your mining model is based on an OLAP data source, the ability to slice the model by using cube attributes is automatically provided.

12.1.2 Data Mining Designer

After you have created a data mining structure and mining model by utilising the Data Mining Wizard, you can use the data mining Designer from either SQL Server Data Tools (SSDT) or SQL Server administration Studio to work with living models and structures. The designer includes devices for these tasks:

- Change the properties of mining structures, add columns and create column aliases, change the binning procedure or expected distribution of values.
- Add new models to an existing structure; replicate models, change model properties or metadata, or define filters on a mining model.
- Browse the patterns and rules inside the model; discover associations or decision trees.
- Validate forms by creating lift charts, or analyse the profit curve for models. Compare models using classification matrices, or validate a data set and its models by using cross-validation.

- Create propositions and content queries against existing mining models. Build one-off queries, or set up queries to develop propositions for whole benches of external data.

12.1.3 SQL Server Management Studio

After you create and deploy mining forms to a server, you can use SQL Server administration Studio to organise the Analysis Services database that hosts the data mining items. You can also continue to continue jobs that use the model, such as exploring the models, processing new data, and creating propositions.

Notes Management Studio also comprises query editors that you can use to design and execute Data Mining Extensions (DMX) queries.

12.1.4 Integration Services Data Mining Tasks and Transformations

SQL Server Integration Services provides many components that support data mining.

Some tools in Integration Services are designed to help automate common data mining tasks such as prediction, model building, and processing.

Example:

- Create an Integration Services package that automatically updates the model every time the addition of new customers updates the dataset.
- Perform custom sampling of case records.
- Generate models passed on parameters automatically.

You can also use data mining in a package workflow, as an input to other processes.

Example:

- To weight score for text mining use probability values generated by the model.
- Automatically generate predictions based on prior data and use those values to assess the validity of new data.
- To segment incoming customers by risk use logistic regression.

Self Assessment

Fill in the blanks:

1. Model viewers are supplied in both SQL Server administrations Studio and ....

2. The .... is provided in both SQL Server administrations Studio and SQL Server Data Tools (SSDT)

3. The .... in Microsoft SQL Server Analysis Services starts every time when you add a new mining structure to a data mining project.

4. .... Services provides many components that support data mining.

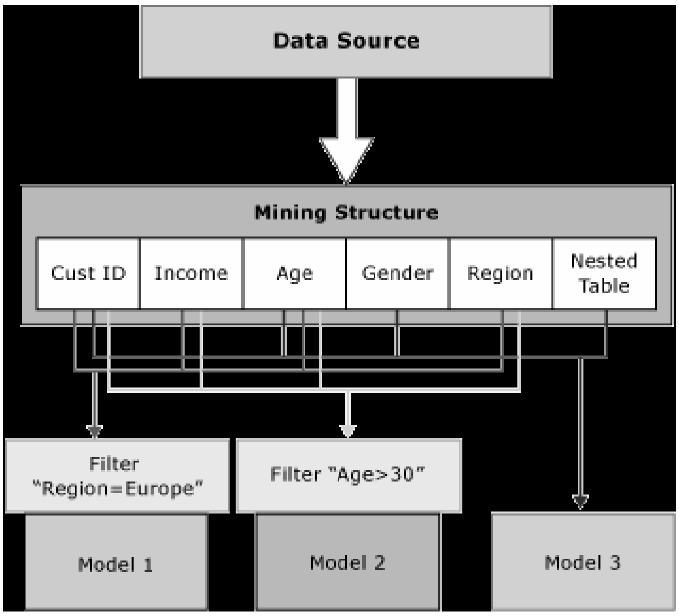
Notes

Task Find out the difference between Mining Model Viewer and Mining Accuracy Chart Tab?

12.2 Mining Structure

The mining structure defines the data from which mining models are constructed: it specifies the source data outlook, the number and kind of columns, and an optional partition into training and testing groups. A single mining structure can support multiple mining models that share the identical domain. The following diagram shows the connection of the data mining structure to the data source, and to its constituent data mining models.

Figure 12.1: Connection of the Data Mining Structure to the Data Source



Source: http://i.technet.microsoft.com/dynimg/IC13488.gif

The mining structure in the Figure 12.1 is based on a data source that comprises multiple tables or views, connected on the CustomerID field. One table contains data about customers, such as the region, age, earnings and gender, while the associated nested table comprises multiple rows of additional data about each customer, such as goods the customer has bought. The Figure 12.1 shows that multiple models can be constructed on one mining structure, and that the models can use different columns from the structure.

Here:

Model 1 uses CustomerID, Income, Age, Region, and filters the data on Region.

Model 2 uses CustomerID, Income, Age, Region and filters the data on Age.

Model 3 uses CustomerID, Age, Gender, and the nested table, with no filter.

As the models use different columns for input, and because two of the models also restrict the data that is used in the model by applying a filter, the models might have very different results even though they are based on the same data.

Caution Here, the CustomerID column is required in all models because it is the only available column that can be used as the case key.

12.2.1 Defining a Mining Structure

Setting up a data mining structure includes the following steps:

- Defining a data source.
- Selecting columns of data to include in the structure and defining a case key.
- Define a key for the structure (including the key for the bested table, if applicable).
- Specify whether the source data should be separate into a training set and testing set or not (Optional).
- Process the structure.

Data Sources for Mining Structures

When you define a mining structure, you use columns that are available in an existing data source view. A data source view is a distributed object that permits you combines multiple data sources and uses them as a single source. The initial data sources are not evident to consumer applications, and you can use the properties of the data source view to modify data kinds, create aggregations, or alias columns.

If you develop multiple mining models from the same mining structure, the models can use distinct columns from the structure.

Example: You can create a single structure and then build distinct decision tree and clustering models from it, with each form using distinct columns and forecasting distinct attributes.

Also, each model can use the columns from the structure in different ways.

Example: Your data source view might comprise an Income column, which you can bin in different ways for different models.

Mining Structure Columns

The building blocks of the mining structure are the mining structure columns, which recount the data that the data source comprises. These columns comprise data such as data type, content type, and how the data is distributed. The mining structure does not comprise data about how columns are used for an exact mining model, or about the kind of algorithm that is used to build a model; this data is characterised in the mining model itself.

A mining structure can also comprise nested tables. A nested table comprises a one-to-many relationship between the entity of a case and its associated attributes.

Example: If the information that recounts the customer resident in one table, and the customer's purchase resides in another table, you can use nested tables to blend the information into a single case.

The customer identifier resident is the entity, and the purchases are the related attributes.

Dividing the Data into Training and Testing Sets

When you define the data for the mining structure, you can also specify that some of the data be used for training, and some for testing. Therefore, it is no longer necessary to separate your data in advance of creating a data mining structure. Instead, while you create your model, you can specify that a certain percentage of the data be held out for testing, and the rest used for training.

Enabling Drillthrough

You can add columns to the mining structure even if you do not plan to use the column in a specific mining model. This is helpful if, for example, you desire to get the e-mail locations of customers in a clustering model, without using the e-mail address throughout the analysis method.

Processing Mining Structures

When you process a mining structure, Analysis Services creates a cache that stores statistics about the data, information about how any continuous attributes are discretized and other information that is later used by mining models.

Viewing Mining Structures

In SQL Server Data Tools (SSDT), you can use the Mining Structure tab of Data Mining Designer to view the structure columns and their definitions.

Self Assessment

State whether the following statements are true or false:

5. A single mining structure can support multiple mining models that share the identical domain.
6. When you define a mining structure, you use columns that are available in an existing data source view.
7. The building blocks of the mining structure are the mining structure columns, which recount the data that the data source comprises.
8. A mining structure cannot comprise nested tables.

12.3 Configuring Algorithm Parameters

You can change the parameters supplied with the algorithms that you use to construct data mining models to customize the results of the model. The algorithm parameters supplied in Microsoft SQL Server Analysis Services change much more than just properties on the model, they can be used to fundamentally adjust the way that data is processed, grouped, and displayed.

Example: You can use algorithm parameters to:

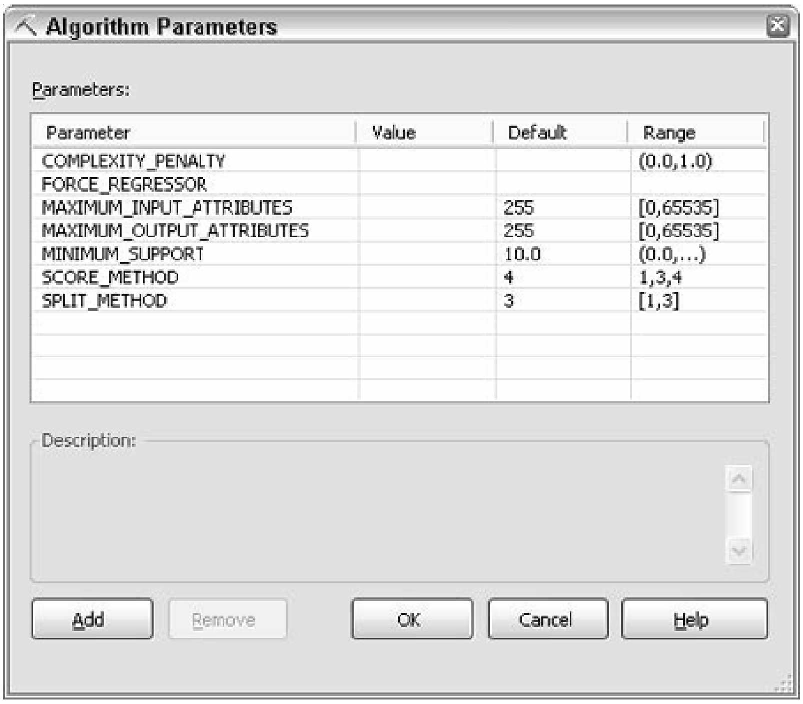
- Change the procedure of analysis, such as the clustering procedure.
- Control feature selection behaviour.

- Identify the dimensions of datasets or the probability of rules.
- Control branching and deepness of decision trees.

12.3.1 Change an Algorithm Parameter

1. On the Mining Models tab of Data Mining Designer in SQL Server Data Tools (SSDT), right-click the algorithm type and select Set Algorithm Parameters. The Algorithm Parameters dialog box will open.

Figure 12.2: An algorithm Parameters dialog box



Parameter	Value	Default	Range
COMPLEXITY_PENALTY			(0.0,1.0)
FORCE_REGRESSOR			
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES		255	[0,65535]
MINIMUM_SUPPORT		10.0	(0.0,...)
SCORE_METHOD		4	1,3,4
SPLIT_METHOD		3	[1,3]

Source: http://media.techtarget.com/digitalguide/images/Misc/dm_6.gif

2. In the Value column, set a new value for the algorithm that you want to change. If you do not enter a value in the Value column, Analysis Services uses the default parameter value. The Range column describes the possible values that you can enter.

3. Click OK. The algorithm parameter is set with the new value. The parameter change will not be reflected in the mining model until you reprocess the model.

12.3.2 View the Parameters

To view the parameters used in an existing model:

- In SQL Server Management Studio, open a DMX Query window.
- Type a query like this one:

```
select MINING_PARAMETERS
from $system.DMSCHEMA_MINING_MODELS
WHERE MODEL_NAME = '<model name>'
```

Notes

Self Assessment

Fill in the blanks:

9. The .... supplied in Microsoft SQL Server Analysis Services change much more than just properties on the model.

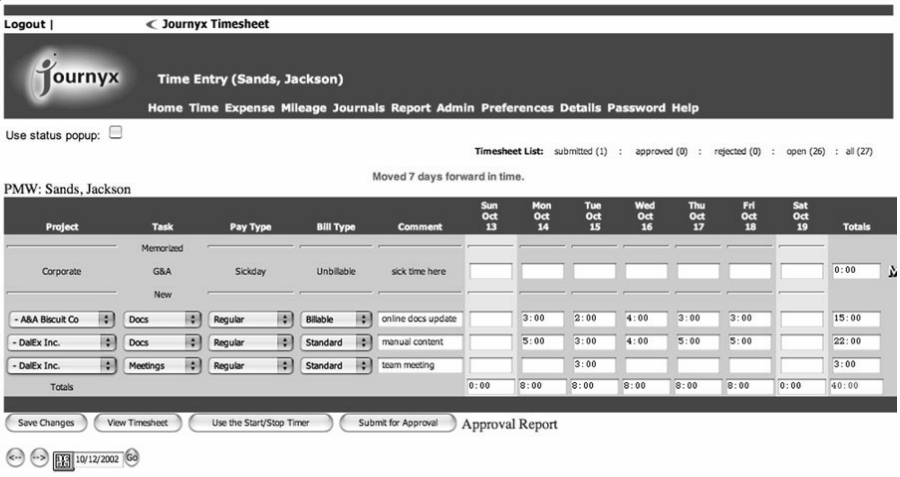
10. On the Mining Models tab of Data Mining Designer in SQL Server Data Tools (SSDT), ... the algorithm type and select Set Algorithm Parameters.

Case Study

Python Powers Journyx Timesheet

Journyx Timesheet (tm) is a commercial application that provides time, expense, and project tracking. In 1996, Curt Finch, Journyx CEO and founder, was working in the staffing industry when he saw an opportunity to use the web to accurately collect and store employee timesheet information.

Figure 1: Journyx Time Entry Screen



Time Entry (Sands, Jackson)

Use status popup: □

Timesheet List: submitted (1) : approved (0) : rejected (0) : open (26) : all (27)

Moved 7 days forward in time.

PMW: Sands, Jackson

Project	Task	Pay Type	Bill Type	Comment	Sun Oct 13	Mon Oct 14	Tue Oct 15	Wed Oct 16	Thu Oct 17	Fri Oct 18	Sat Oct 19	Totals
	Memorized											
Corporate	G&A	Sickday	Unbillable	sick time here								0:00
	New											
- A&A Biscuit Co	Docs	Regular	Billable	online docs update		3:00	2:00	4:00	3:00	3:00		15:00
- DalEx Inc.	Docs	Regular	Standard	manual content		5:00	3:00	4:00	5:00	5:00		22:00
- DalEx Inc.	Meetings	Regular	Standard	team meeting			3:00					3:00
Totals					0:00	8:00	8:00	8:00	8:00	8:00	0:00	40:00

10/12/2002

The first version of Timesheet focused on collecting accurate cost information, with an eye towards applying that data in the formulation of new project cost projections. Since then, Timesheet has expanded considerably to facilitate tracking of time, mileage, and expenses, not just for project management but also for billing and payroll purposes. Optional modules are available for paperless expense reporting, advanced user role management, automated billing and payroll, and to facilitate system access for disconnected travelling users. Today, Timesheet is platform-independent, flexible enough to be reconfigured by customers to fit unique organizational needs and scales to tens of thousands of users for the large enterprise.

From the Start

Journyx Timesheet has been using Python from the beginning. Curt Finch chose Python initially on the recommendation of a friend, Steve Madere, who had founded Dejanews.com (now a part of Google). Describing the rationale for his choice, Curt said, "I looked at Java and C and came to the conclusion that 1 line of Python is 10 lines of Java or 100 lines of C.

Contd...

Notes

Developers write code at basically a constant rate so we chose Python which was (and is) the highest level language I've ever seen that is also flexible enough to be generally useful."

Architecture

From the beginning, Timesheet was designed and implemented as a web application. It uses three-tiered web application architecture with separate layers for web presentation, business logic, and data storage. As time has progressed, the application's functionality has advanced considerably, and Curt's decision to use Python for an implementation language has proven to be good choice. Python is currently used for all application logic in the Timesheet application. This includes all code between the initial Apache dispatch, where mod_python is employed to expedite interpreter instantiation, through the application logic, and down to the point of call out to the database transport layer. Timesheet uses not only the Python standard library but also several independently developed open source Python subsystems, such as PyXML and ActZero's SOAP support. PyXML is used to implement certain business rules and to develop jxAPI, which is a SOAP-based API into the application logic. Work is in progress to extend this API to define Web Services Description Language templates for the jxAPI functions. The application currently builds and ships with Python 2.1.1. Timesheet also incorporates several non-Python technologies. The Unix and Linux distributions are packaged with the Apache HTTP server and PostgreSQL database. The Timesheet distribution for Windows ships with an optional Microsoft Desktop Engine (MSDE) database and integrates with Microsoft IIS. Timesheet can be configured to use a variety of third-party databases.

Results

The Timesheet project has succeeded spectacularly, generating millions in revenue and allowing Journyx to grow every year, even under the current economic conditions. Journyx, like many of our customers, uses Timesheet internally as a mission critical part of the company infrastructure. It is used extensively for project tracking, billing, and payroll. To date, approximately 11 person-years have gone into the Journyx Timesheet product, resulting in over one hundred fifty thousand lines of Python code. In developing Journyx, the two greatest benefits of Python were the speed with which features could be written and deployed, and its true write-once-run-anywhere cross-platform capabilities. Journyx developers have found that the simplicity and clarity of Python combine with its object-oriented properties to make it a very powerful and productive language. Python's rich standard library, which includes modules for things like string manipulation and HTML generation, further supports programmers in meeting aggressive development schedules.

Because of these properties of the language, Python has enabled Journyx to add features more quickly than our competitors. We've been able to implement SOAP/XML and WSDL support and extended other aspects of the application's functionality well ahead of competitive products. One of the key enablers of this efficiency in maintenance and improvement is the inherent clarity and readability of the Python language. Other important factors are the vibrant and responsive Python development community, and the high degree of backwards compatibility and stability we have seen as the language design evolves over time. Python's cross-platform standard library and platform-independent byte code file format allow the deployment of Python modules to any platform, regardless of which platform the module was prepared on. This helped not only in avoiding per-platform development overhead but also facilitates customer support for the Timesheet software product. For example, a patch module built on a Redhat 6.2 system can be sent to a customer for installation on Windows XP or any other operation system without the need for cross-compilation or translation of any kind.

Contd....

Notes

Questions:

1. How Python made it possible for Journyx to produce a flexible, feature-rich product for multiple platforms in less time?
2. Explain how Python becomes an important competitive advantage for Journyx Timesheet (tm)?

Source: http://www.python.org/about/success/journyx/

12.4 Summary

- The Data Mining Wizard in SQL Server Data Tools (SSDT) makes it very simple to create mining structures and mining models, using either relational data or multidimensional data in cubes.
- The Data Mining Wizard in Microsoft SQL Server Analysis Services starts every time when you add a new mining structure to a data mining project.
- After you create and deploy mining forms to a server, you can use SQL Server administration Studio to organise the Analysis Services database that hosts the data mining items.
- The mining structure defines the data from which mining models are constructed: it specifies the source data outlook, the number and kind of columns, and an optional partition into training and testing groups.
- You can change the parameters supplied with the algorithms that you use to construct data mining models to customize the results of the model.
- On the Mining Models tab of Data Mining Designer in SQL Server Data Tools (SSDT), right-click the algorithm type and select Set Algorithm Parameters.

12.5 Keywords

Data Mining Extensions (DMX): Data Mining Extensions (DMX) is a language that you can use to create and work with data mining models in Microsoft SQL Server Analysis Services.

Data Mining Wizard: The Data Mining Wizard in Microsoft SQL Server 2005 Analysis Services (SSAS) starts every time that you add a new mining structure to a data mining project.

Mining structure: It defines the data from which mining models are constructed; it specifies the source data outlook, the number and kind of columns, and an optional partition into training and testing groups.

SQL Server Data Tools: SQL Server Data Tools (SSDT) transforms database development by introducing a ubiquitous, declarative model that spans all the phases of database development and maintenance/update inside Visual Studio.

12.6 Review Questions

1. What tools does Microsoft SQL Server Analysis Services provides to create data mining solutions?
2. What is Data Mining Wizard? Discuss the process in starting the Data Mining Wizard.
3. What are the features of Data Mining Wizard?

4. What is the Data Mining Designer?
5. Explain SQL Server Management Studio.
6. Write short note on integration services data mining tasks and transformations.
7. What is mining structure?
8. Discuss the connection of the data mining structure to the data source.
9. What are the data sources for mining structures?
10. Give steps for changing algorithm parameter.

Answers: Self Assessment

1. SQL Server Data Tools (SSDT)
2. Prediction Query Builder
3. Data Mining Wizard
4. SQL Server Integration
5. True
6. True
7. True
8. False
9. Algorithm parameters
10. Right-click

12.7 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

www.ibm.com/developerworks/library/ba-data-mining-techniques/?

www.icsti.org/IMG/pdf/VTTDataMiningTools.pdf?

www.web-datamining.net/tools/?

Notes

Unit 13: Creating Data Mining Queries and Reports

CONTENTS

Objectives

After studying this unit, you will be able to:

- Explain prediction queries
- Discuss about data mining extensions

Introduction

Data Mining Extensions (DMX) is a query language used for Data Mining Models supported by Microsoft's SQL Server Analysis Services product. Like SQL, it supports a data definition language, data manipulation language and a data query language, all three with SQL-like syntax. Difference is that SQL statements operate on relational tables while DMX statements operate on data mining models. In this unit, you will learn about prediction queries like adding, singleton and batch queries. Later on in this unit you will learn about data mining extensions and statements-data definition statements, data manipulation statements and data query statements.

13.1 Prediction Queries

The aim of a usual data mining task is to use the mining model to make predictions.

Example: You might want to forecast the amount of expected downtime for a certain cluster of servers, or develop a score that shows if segments of customers are expected to reply to an advertising campaign or not. To do all these things, you need to create a prediction query.

Functionally, there are distinct types of prediction queries supported in SQL Server, depending on the type of inputs to the query. They are shown in **Table 13.1.**

Table 13.1: Types of Prediction Queries

Query Type	Query Options
Singleton prediction queries	Use a singleton query when you want to predict outcomes for a single new case, or multiple new cases. You provide the input values directly in the query, and the query is executed as a single session.
Batch predictions	Use batch predictions when you have external data that you want to feed into the model, to use as the basis for predictions. To make predictions for an entire set of data, you map the data in the external source to the columns in the model, and then specify the type of predictive data you want to output. The query for the entire dataset is executed in a single session, making this option much more efficient than sending multiple repeated queries.
Time Series predictions	Use a time series query when you want to predict a value over some number of future steps. SQL Server Data Mining also provides the following functionality in time series queries: • You can extend an existing model by adding new data as part of the query, and make predictions based on the composite series. • You can apply an existing model to a new data series by using the REPLACE_MODEL_CASES option. • You can perform cross-prediction.

Source: http://technet.microsoft.com/en-us/library/hh213169.aspx

13.1.1 Basic Prediction Query Design

When you create a prediction, you normally supply some piece of new information and ask the model to develop a prediction based on the new data.

- In a batch prediction query, you map the model to an external source of data by using a prediction join.
- In a singleton prediction query, you type one or more values to use as inputs.

Did u know? You can create multiple propositions using a singleton prediction query. However, if you need to create many propositions, performance is better when you use a batch query.

Both singleton and batch prediction queries use the PREDICTION JOIN syntax to define the new data. The difference is in how the input side of the prediction join is specified.

- In a batch prediction query, the data comes from an external data source that is specified by using the OPENQUERY syntax.
- In a singleton prediction query, the data is supplied inline as part of the query.

Notes

13.1.2 Adding Prediction Functions

In addition to predicting a value, you can customize a prediction query to return various types of information that are related to the proposition.

Example: If the prediction creates a list of goods to recommend to a customer, you might furthermore desire to return the probability for each proposition, so that you can rank them and present only the peak recommendations to the client.

To do this, you add prediction purposes to the query. Each model or query type supports specific functions.

Example: Clustering models support special prediction purposes that supply extra details about the clusters created by the model, while time series models have functions that assess difference over time.

13.1.3 Singleton Query

The first step is to use the *SELECT FROM <model> PREDICTION JOIN (DMX)* in a singleton prediction query.

Example: The following is an example of the singleton statement:

```
SELECT <select list> FROM [<mining model name>]
NATURAL PREDICTION JOIN
(SELECT '<value>' AS [<column>], ...)
AS [<input alias>]
```

Here, the first line of the code defines the columns from the mining model that the query should return, and specifies the mining model that is used to generate the prediction:

```
SELECT <select list> FROM [<mining model name>]
```

The next lines of the code define the characteristics of the customer that you use to create a prediction:

```
NATURAL PREDICTION JOIN
(SELECT '<value>' AS [<column>], ...)
AS [<input alias>]
ORDER BY <expression>
```

If you specify NATURAL PREDICTION JOIN, the server matches each column from the model to a column from the input, based on column names. If column names do not match, the columns are ignored.

13.1.4 To Create a Singleton Prediction Query

1. In Object Explorer, right-click the instance of Analysis Services, point to New Query, and then click DMX. Query Editor opens and contains a new, blank query.
2. Copy the example of the singleton statement into the blank query.
3. Replace the following:

   ```
   <select list>
   with:
   ```

```
[Car Buyer] AS Buyer, PredictHistogram([Car Buyer]) AS Statistics
```

The AS statement is used to alias columns returned by the query. The PredictHistogram function returns statistics about the prediction, including the probability and the support.

4. Replace the following:

```
[<mining model>]

with:
[Decision Tree]
```

5. Replace the following:

```
(SELECT '<value>' AS [<column name>], ...)   AS t

with:

(SELECT 35 AS [Age],
   '5-10 Miles' AS [Commute Distance],
   '1' AS [House Owner ],
   2 AS [Number Bikes Owned],
   2 AS [Total Children]) AS t
```

The complete statement should now be as follows:

```
SELECT
    [Car Buyer] AS Buyer,
    PredictHistogram([Car Buyer]) AS Statistics
FROM
    [Decision Tree]
NATURAL PREDICTION JOIN
(SELECT 35 AS [Age],
    '5-10 Miles' AS [Commute Distance],
    '1' AS [House Owner],
    2 AS [Number Bikes Owned],
    2 AS [Total Children]) AS t
```

6. On the File menu, click Save DMXQuery1.dmx As.
7. In the Save As dialog box, browse to the appropriate folder, and name the file Singleton_Query.dmx.
8. On the toolbar, click the Execute button.

Caution The query returns a prediction about whether a customer with the specified characteristics will purchase a car, as well as statistics about that prediction.

13.1.5 Batch Query

The next step is to use the SELECT FROM <model> PREDICTION JOIN (DMX) in a batch prediction query.

Example: The following is an example of a batch statement:

```
SELECT TOP <number> <select list>
FROM [<mining model name>]
PREDICTION JOIN
OPENQUERY([<datasource>],'<SELECT statement>')
   AS [<input alias>]
ON <on clause, mapping,>
WHERE <where clause, boolean expression,>
ORDER BY <expression>
```

Let us see which line of code is useful for what purpose. Here, the first two lines of the code define the columns from mining model that the query returns, as well as the name of the mining model that is used to generate the prediction in the saw as singleton query return.

> *Notes* The TOP <number> statement specifies that the query will only return the number or the results specified by <number>.

The next lines of the code define the source data that the predictions are based on:

```
OPENQUERY([<datasource>],'<SELECT statement>')
   AS [<input alias>]
```

The next line defines the mapping between the source columns in the mining model and the columns in the source data:

```
ON <column mappings>
```

The WHERE clause filters the results returned by the prediction query:

```
WHERE <where clause, boolean expression,>
```

The last (and optional) line of the code specifies the column that the results will be ordered by:

```
ORDER BY <expression> [DESC|ASC]
```

Use ORDER BY in combination with the TOP <number> statement, to filter the results that are returned.

Example: In this prediction you will return the top ten car buyers, ordered by the probability of the prediction being correct. You can use [DESC|ASC] to control the order in which the results are displayed.

Self Assessment

Fill in the blanks:

1. The aim of a usual data mining task is to use the mining model to make ....
2. When you create a prediction, you normally supply some piece of new information and ask the model to develop a prediction based on the ....
3. Both .... and .... prediction queries use the PREDICTION JOIN syntax to define the new data.
4. In addition to predicting a value, you can customize a prediction query to return various types of information that are related to the ....

5. In .... -, right-click the instance of Analysis Services, point to New Query, and then click DMX.

13.2 Data Mining Extensions (DMX)

Data Mining Extensions (DMX) is a language that you can use to create and work with data mining models in Microsoft SQL Server Analysis Services. You can also use DMX to create the structure of new data mining models, to train these models, and to browse, manage, and predict against them. It is composed of Data Definition Language (DDL) statements, Data Manipulation Language (DML) statements, and functions and operators.

13.2.1 DMX Statements

You can use DMX statements to create, process, delete, copy and predict against data mining models. There are three types of statements in DMX: data definition statements, data manipulation statements and data query statements.

Task Compare and contrast the data manipulation statements and data query statements.

13.2.2 Data Definition Statements

Data definition statements are used in DMX to create and define new mining structures and models, to import and export mining models and mining structures, and to drop existing models from a database. Data definition statements in DMX are part of the data definition language (DDL). You can perform the following tasks with the data definition statements in DMX:

- Create a mining structure by using the CREATE MINING STRUCTURE statement
- Add a mining model to the mining structure by using the ALTER MINING STRUCTURE statement.
- Create a mining model and associated mining structure simultaneously by using the CREATE MINING MODEL
- Export a mining model and associated mining structure to a file by using the EXPORT statement.
- Import a mining model and associated mining structure from a file that is created by the EXPORT statement by using the IMPORT statement.
- Copy the structure of an existing mining model into a new model, and train it with the same data, by using the SELECT INTO statement.
- Remove a mining model from a database by using the DROP MINING MODEL statement.

13.2.3 Data Manipulation Statements

Data manipulation statements are used in DMX to work with existing mining models, to browse the models and to create predictions against them. Data manipulation statements in DMX are part of the data manipulation language (DML). You can perform the following tasks with the data manipulation statements in DMX:

- Train a mining model by using the INSERT INTO statement.

Notes

- Extend the SELECT statement to browse the information that is calculated during model training and stored in the data mining model, such as statistics of the source data.
- Create predictions that are based on an existing mining model by using the PREDICTION JOIN clause of the SELECT statement.
- Remove all the trained data from a model or a structure by using the DELETE (DMX) statement.

13.2.4 DMX Query Fundamentals

DMX functions can be used to obtain information that is discovered during the training of your models, and to calculate new information. One can also use these functions for many purposes, including to return statistics that describe the underlying data or the accuracy of a prediction, or to return an expanded explanation of a prediction.

Self Assessment

State whether the following statements are true or false:

6. Data manipulation language (DML) is a language that you can use to create and work with data mining models in Microsoft SQL Server Analysis Services.
7. You can use DMX statements to create process, delete, copy and predict against data mining models.
8. Data definition statements are used in DMX to create and define, to import and export, and to drop existing models from a database.
9. Data manipulation statements in DMX are not a part of the Data Manipulation Language (DML).
10. DMX functions can be used to obtain information that is discovered during the training of your models, and to calculate new information.

Case Study

Federal Agency Data Mining Reporting

(a) Short title

This section may be cited as the "Federal Agency Data Mining Reporting Act of 2007".

(b) Definitions

In this section:

(1) ***Data mining:*** The term "data mining" means a program involving pattern-based queries, searches, or other analyses of 1 or more electronic databases, where—

(A) A department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;

Contd....

(B) The queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and

(C) The purpose of the queries, searches, or other analyses is not solely—

(i) The detection of fraud, waste, or abuse in a Government agency or program; or

(ii) The security of a Government computer system.

(2) ***Database:*** The term "database" does not include telephone directories, news reporting, and information publicly available to any member of the public without payment of a fee, or databases of judicial and administrative opinions or other legal research sources.

(c) Reports on data mining activities by Federal agencies

(1) ***Requirement for report:*** The head of each department or agency of the Federal Government that is engaged in any activity to use or develop data mining shall submit a report to Congress on all such activities of the department or agency under the jurisdiction of that official. The report shall be produced in coordination with the privacy officer of that department or agency, if applicable, and shall be made available to the public, except for an annex described in subparagraph (C).

(2) ***Content of report:*** Each report submitted under subparagraph (A) shall include, for each activity to use or develop data mining, the following information:

(A) A thorough description of the data mining activity, its goals, and, where appropriate, the target dates for the deployment of the data mining activity.

(B) A thorough description of the data mining technology that is being used or will be used, including the basis for determining whether a particular pattern or anomaly is indicative of terrorist or criminal activity.

(C) A thorough description of the data sources that are being or will be used.

(D) An assessment of the efficacy or likely efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the data mining activity.

(E) An assessment of the impact or likely impact of the implementation of the data mining activity on the privacy and civil liberties of individuals, including a thorough description of the actions that are being taken or will be taken with regard to the property, privacy, or other rights or privileges of any individual or individuals as a result of the implementation of the data mining activity.

(F) A list and analysis of the laws and regulations that govern the information being or to be collected, reviewed, gathered, analysed, or used in conjunction with the data mining activity, to the extent applicable in the context of the data mining activity.

(G) A thorough discussion of the policies, procedures, and guidelines that are in place or that are to be developed and applied in the use of such data mining activity in order to—

Contd....

Notes

(i) protect the privacy and due process rights of individuals, such as redress procedures; and

(ii) ensure that only accurate and complete information is collected, reviewed, gathered, analysed, or used, and guard against any harmful consequences of potential inaccuracies.

(3) ***Annex***

(A) *In general:* A report under sub-paragraph (A) _shall include in an annex any necessary—

(i) classified information;

(ii) law enforcement sensitive information;

(iii) proprietary business information; or

(iv) trade secrets (as that term is defined in section 1839 of title 18).

(B) *Availability:* Any annex described in clause (i)—

(i) shall be available, as appropriate, and consistent with the National Security Act of 1947 [50 U.S.C. 3001 et seq.], to the Committee on Homeland Security and Governmental Affairs, the Committee on the Judiciary, the Select Committee on Intelligence, the Committee on Appropriations, and the Committee on Banking, Housing, and Urban Affairs of the Senate and the Committee on Homeland Security, the Committee on the Judiciary, the Permanent Select Committee on Intelligence, the Committee on Appropriations, and the Committee on Financial Services of the House of Representatives; and

(ii) shall not be made available to the public.

(4) ***Time for report:*** Each report required under sub-paragraph (A) - shall be—

(A) submitted not later than 180 days after August 3, 2007; and

(B) updated not less frequently than annually thereafter, to include any activity to use or develop data mining engaged in after the date of the prior report submitted under sub-paragraph (A).

Questions:

1. Define data mining according to the act.
2. What is the content of report?

Source: http://www.law.cornell.edu/uscode/text/42/2000ee-3

13.3 Summary

- The aim of a usual data mining task is to use the mining model to make predictions.
- Functionally, there are distinct types of prediction queries supported in SQL Server, depending on the type of inputs to the query.
- When you create a prediction, you normally supply some piece of new information and ask the model to develop a prediction based on the new data.
- Both singleton and batch prediction queries use the PREDICTION JOIN syntax to define the new data.

- In addition to predicting a value, you can customize a prediction query to return various types of information that are related to the proposition.
- Data Mining Extensions (DMX) is a language that you can use to create and work with data mining models in Microsoft SQL Server Analysis Services.
- There are three types of statements in DMX: data definition statements, data manipulation statements and data query statements.
- Data definition statements in DMX are part of the Data Definition Language (DDL).
- Data manipulation statements are used in DMX to work with existing mining models, to browse the models and to create predictions against them.
- DMX functions can be used to obtain information that is discovered during the training of your models, and to calculate new information.

13.4 Keywords

Data Definition Language (DDL): Data Definition Language (DDL) describes the portion of SQL that allows you to create, alter, and destroy database objects.

Data Manipulation Language (DML): A data manipulation language (DML) is a family of syntax elements similar to a computer programming language used for inserting, deleting and updating data in a database.

Data Manipulation Statements: Data manipulation statements are used in DMX to work with existing mining models, to browse the models and to create predictions against them.

Data Mining Extensions (DMX): Data Mining Extensions (DMX) is a language that you can use to create and work with data mining models in Microsoft SQL Server Analysis Services.

13.5 Review Questions

1. What is prediction?
2. What are the types of prediction queries?
3. Discuss about basic prediction query design.
4. Write short note on adding prediction functions.
5. Give examples for singleton query.
6. Elaborate the steps to create a singleton prediction query.
7. What is batch query?
8. Explain the concept of Data Mining Extensions (DMX).
9. What is the use of DMX statements?
10. Write short note on data definition statements.
11. Discuss about data manipulation statements.
12. Provide fundamentals of DMX query.

Answers: Self Assessment

1.	Predictions	2.	New data

Notes

3. Singleton, batch
4. Proposition
5. Object Explorer
6. False
7. True
8. True
9. False
10. True

13.6 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). "*Business Intelligence for Dummies*". Wiley.

Online links

docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm

www.personal.psu.edu/glh10/ist110/topic/topic07/topic07_09.html

technet.microsoft.com/en-us/library/ms174949.aspx

Unit 14: Reporting Tools

CONTENTS

Objectives

After studying this unit, you will be able to:

- Identify the functionalities of reporting tools
- Discuss the reporting services using SQL Server
- Explain the analysis services using SQL Server

Introduction

Reporting software is used to generate human-readable reports from various data sources. Business operations reporting and dashboard are the most common applications for a reporting tool. Market share-leading reporting solution is from Microsoft Microsoft SQL Server Reporting Services is the solution of choice for many businesses that require enterprise reporting capabilities. In this unit, you will learn about functionalities of reporting tools. Later on in this unit, reporting services using SQL server will be discussed with appropriate screenshots. Finally, analysis services using SQL server will be introduced.

14.1 Reporting Tool Functionalities

Following are the reporting tool functionalities:

1. ***Front End***: Data is ineffective if all it does is sitting in the data warehouse. As an outcome, the production layer is of very high significance.
2. ***Data source connection capabilities:*** There are two types of data sources:
    - ❖ The relationship database
    - ❖ The OLAP multidimensional data source

3. ***Scheduling and distribution capabilities:*** The reporting tool must have scheduling and distribution capabilities.

> *Notes* Weekly reports are scheduled to run on Monday morning, and the resulting reports are distributed to the senior executives either by email or web publishing.

4. ***Security Features:*** Because reporting tools are aimed towards a number of users, making sure people see only what they are supposed to see is important.

Did u know? Security can reside at the report level, folder level, column level or row level. Generally all established reporting tools have these capabilities.

5. ***Customization:*** Provide easy way to pre-set the reports to look exactly the way that adheres to the corporate standard.

6. ***Export capabilities***: The most common export needs are to Excel, to a flat file, and to PDF.

Caution For Excel, if the situation wants, you will want to verify that the reporting format, not just the data itself, will be exported out to Excel.

Example: The BIRT reporting features like report layout, data access, and scripting support are used to create reports that use the custom reporting URLs from Rational Asset Manager.

Self Assessment

Fill in the blanks:

1. There are two types of data sources: the .... and the OLAP multidimensional data source.

2. .... is used to generate human-readable reports from various data sources.

14.2 Reporting Services Using SQL Server

Microsoft has come up with its own reporting service, in conjunction with SQL server database to insert the Microsoft SQL Server Reporting Services [SSRS]. It supplies projects of type Business Intelligence task therefore endowing not only large companies but also medium-sized and small businesses also to earn from its advantages. This helps in better business decisions too.

SSRS supplies some extensions towards the data rendering, consignment and security of reports thereby allowing it to have a higher programmable ability. This innovative approach enables reports to be created with lesser development effort [compared to other reporting services], along with customized security choices.

SSRS is a comprehensive reporting platform whereby accounts are retained on a centralized web server (or set of servers). Because accounts are centralized, users run accounts from one location. Having centralized accounts also means that report deployment is rather simplified.

14.2.1 Architecture

After using SSRS, the architecture is just like a small operating system. The Report Manager is the central person who acts as a manager to decide when the reports will be scheduled to run along with maintaining the user profiles on the report server.

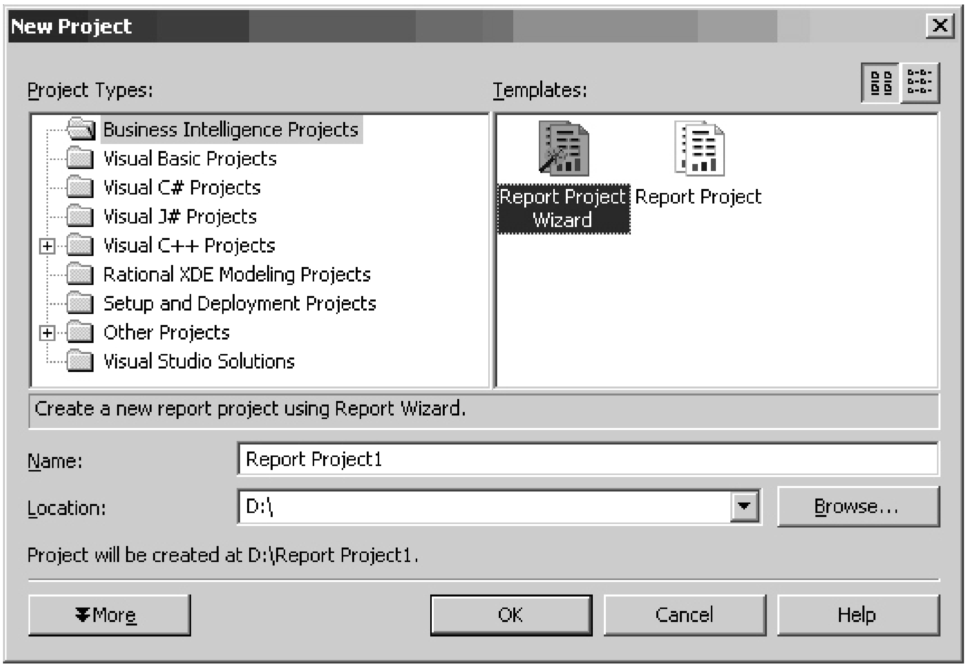
On Report Server all the reports reside. All other activities pertaining to SSRS is done at report server.

Did u know? Report Designer is a graphical tool that is hosted within the Microsoft Visual Studio IDE. It provides tabbed windows for Data, Layout, and Preview that allow you to design a report interactively.

14.2.2 Working with SQL Server Reporting Services

After installing SQL server reporting services on your system, start the Visual studio IDE. Go to File -> New Project, and you will be shown a prompt with 'New project'. Select Business Intelligence Projects from the Project Types. As this is our first project, use Report Project Wizard.

Figure 14.1: New Project Window



Create a new report project using Report Wizard.

Name: Report Project1

Location: D:\

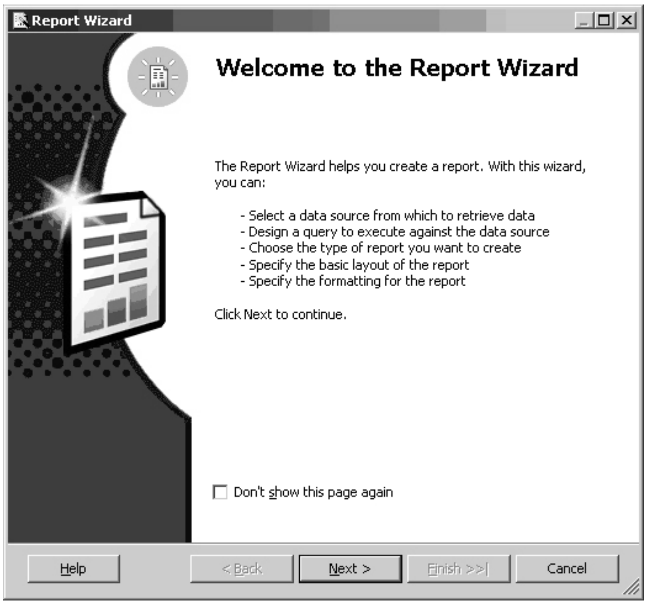
Project will be created at D:\Report Project1.

Source: http://www.codeproject.com/KB/books/Start_SSRS/1.jpg

Specify the name of the project as well as the location where the project will be placed. On click of OK, you will be prompted with a report wizard screen as shown in **Figure 14.2**. Click on Next to follow up to the next screen.

Notes

Figure 14.2: Report Wizard Screen



Report Wizard

Welcome to the Report Wizard

The Report Wizard helps you create a report. With this wizard, you can:

- Select a data source from which to retrieve data
- Design a query to execute against the data source
- Choose the type of report you want to create
- Specify the basic layout of the report
- Specify the formatting for the report

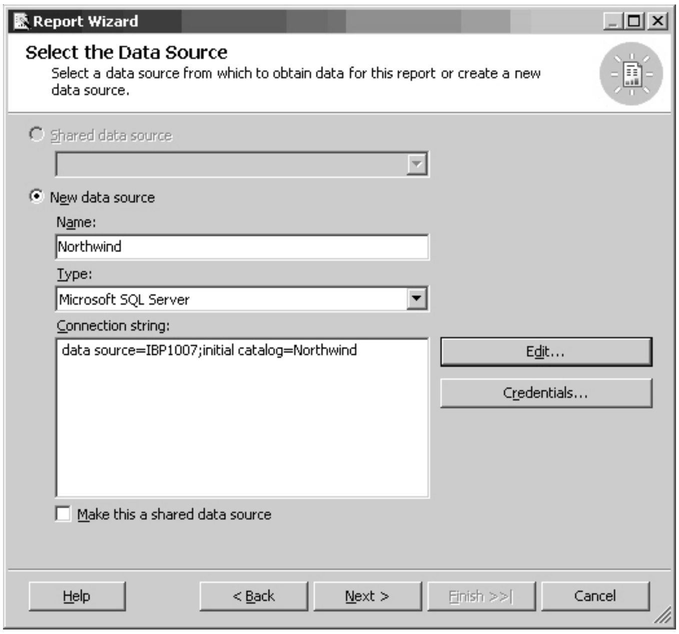
Click Next to continue.

□ Don't show this page again

Source: http://www.codeproject.com/KB/books/Start_SSRS/2.jpg

On the next screen, you will need to create a data source for the report. Just click on Edit to specify the server name and the database then will be used from that server. The connection string is automatically created.

Figure 14.3: Select Data Source



Report Wizard

Select the Data Source

Select a data source from which to obtain data for this report or create a new data source.

○ Shared data source

● New data source

Name: Northwind

Type: Microsoft SQL Server

Connection string: data source=IBP1007;initial catalog=Northwind

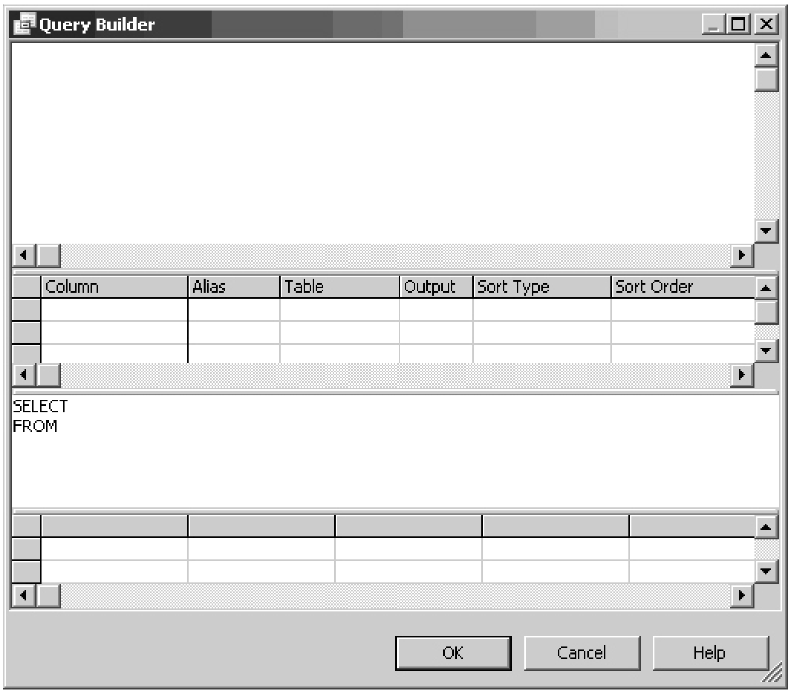
□ Make this a shared data source

Source: http://www.codeproject.com/KB/books/Start_SSRS/3.jpg

On click of Next, you will be prompted with the Query Builder screen. Here you can add tables, select columns as well as execute the SQL statements.

Notes

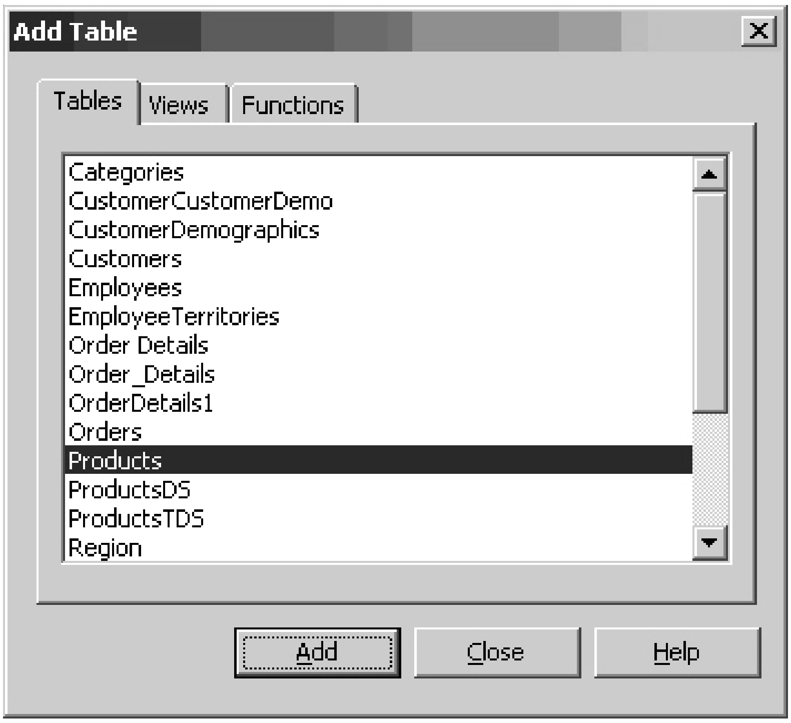
Figure 14.4: Query Builder



```
SELECT
FROM
```

Source: http://www.codeproject.com/KB/books/Start_SSRS/5.jpg

Figure 14.5: Add a table

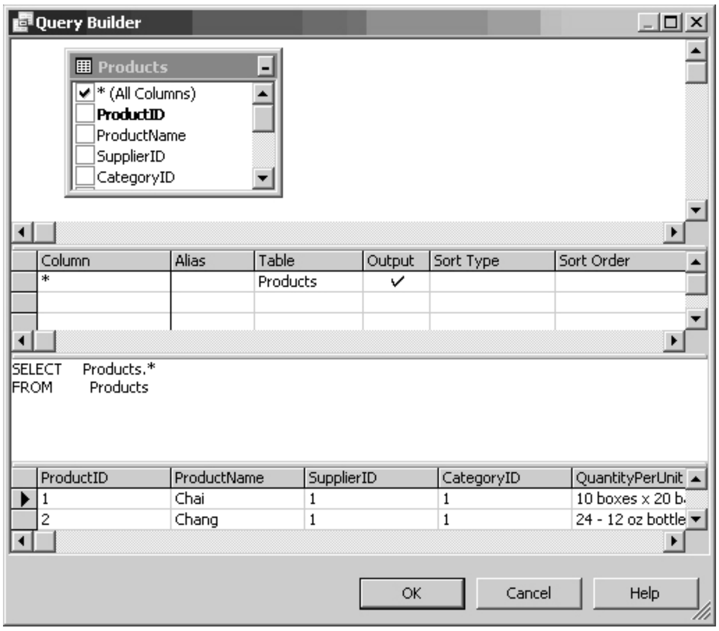


Categories
CustomerCustomerDemo
CustomerDemographics
Customers
Employees
EmployeeTerritories
Order Details
Order_Details
OrderDetails1
Orders
Products
ProductsDS
ProductsTDS
Region

Source: http://www.codeproject.com/KB/books/Start_SSRS/6.jpg

Notes

Figure 14.6: Query Builder with sample data



Query Builder

Products

- ☑ * (All Columns)
- □ **ProductID**
- □ ProductName
- □ SupplierID
- □ CategoryID

Column	Alias	Table	Output	Sort Type	Sort Order
*		Products	✓		

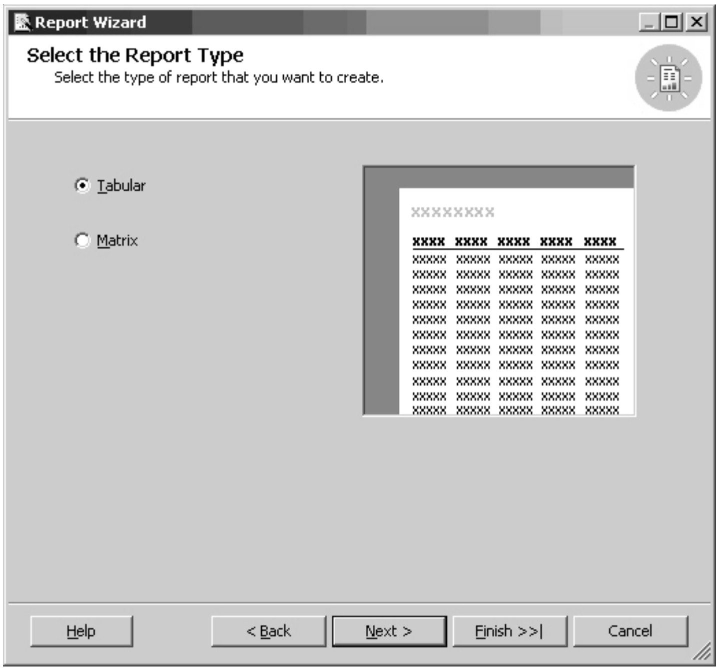
```
SELECT    Products.*
FROM      Products
```

ProductID	ProductName	SupplierID	CategoryID	QuantityPerUnit
1	Chai	1	1	10 boxes x 20 b.
2	Chang	1	1	24 - 12 oz bottle

Source: http://www.codeproject.com/KB/books/Start_SSRS/7.jpg

Based on what query suits your report, create the SQL statement and proceed forward. On next click, you will be prompted with the report type screen. You can choose as Tabular or matrix. To make things simpler, use the Tabular format.

Figure 14.7: Chose Report Type



Report Wizard

Select the Report Type

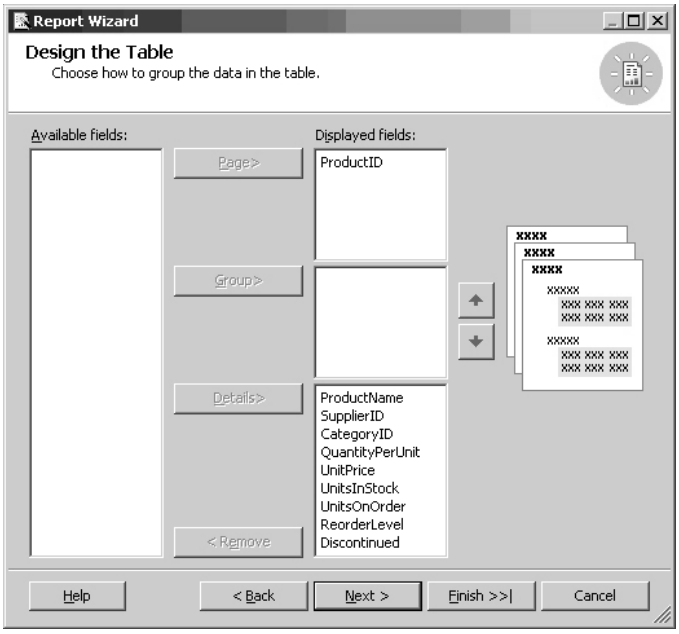
Select the type of report that you want to create.

- ● Tabular
- ○ Matrix

Source: http://www.codeproject.com/KB/books/Start_SSRS/8.jpg

On next click, you will come to the table designing screen, wherein you will be prompted to display the fields as Page, Group or Details.

Figure 14.8: Design the table



Report Wizard

Design the Table

Choose how to group the data in the table.

Available fields:

Displayed fields:

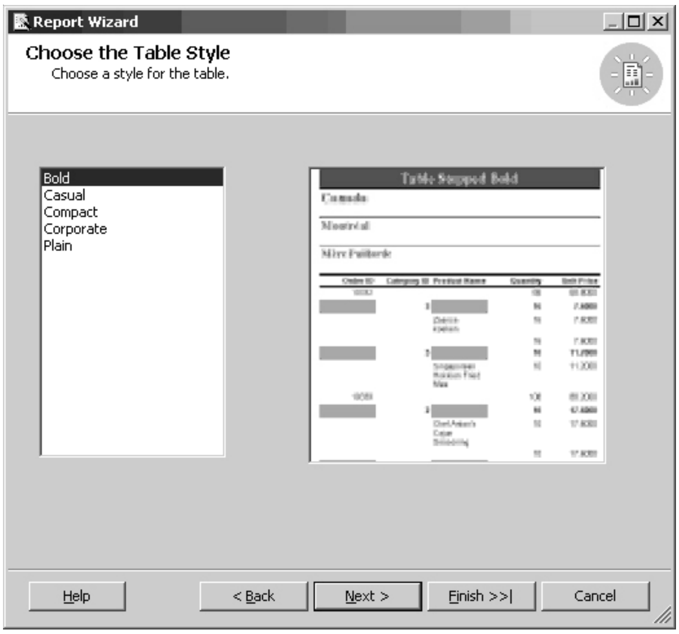
Page: ProductID

Details: ProductName, SupplierID, CategoryID, QuantityPerUnit, UnitPrice, UnitsInStock, UnitsOnOrder, ReorderLevel, Discontinued

Source: http://www.codeproject.com/KB/books/Start_SSRS/9.jpg

On next click, you will be prompted with the Table Style prompt, which contains a list to choose. Select any one from them.

Figure 14.9: Choose table style



Report Wizard

Choose the Table Style

Choose a style for the table.

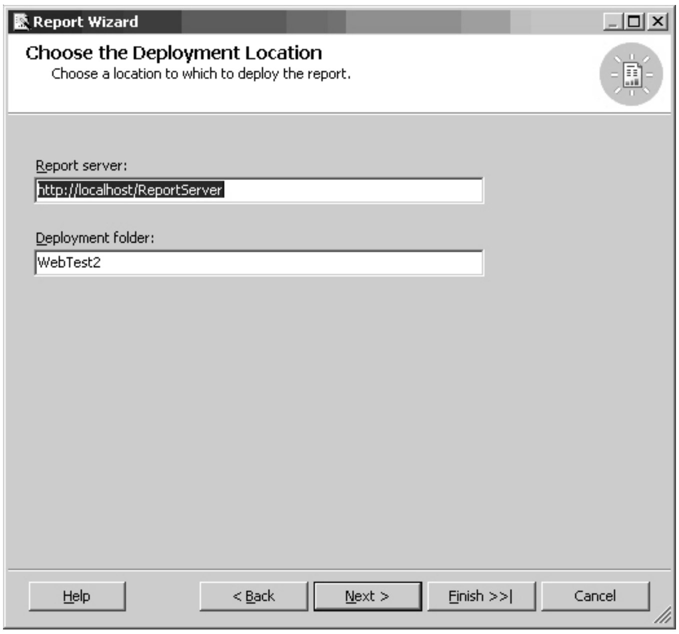
Bold (selected)
Casual
Compact
Corporate
Plain

Source: http://www.codeproject.com/KB/books/Start_SSRS/10.jpg

Notes

On next, you will be prompted with the deployment details screen. Specify the report server name (normally it is http://localhost/ReportServer). If you are using another server then you can specify the location as http://servername/ReportServer . Also give the deployment folder.

Figure 14.10: Choose Deployment Location



Choose the Deployment Location

Choose a location to which to deploy the report.

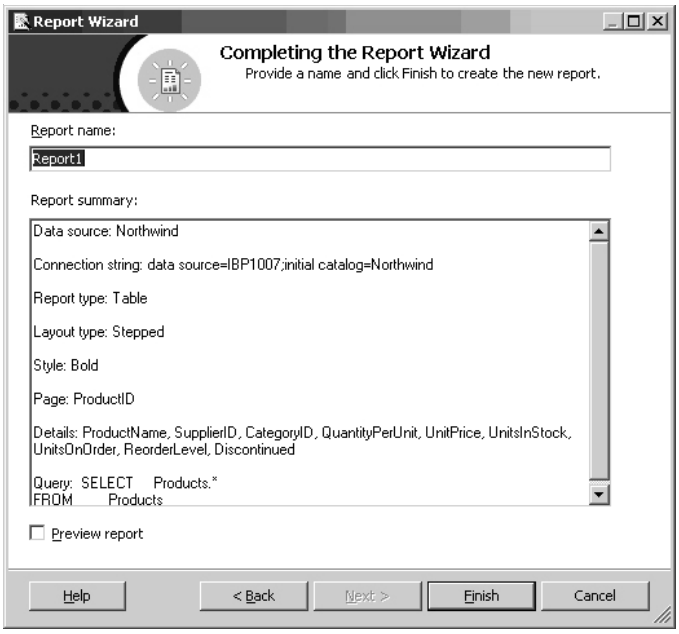
Report server: http://localhost/ReportServer

Deployment folder: WebTest2

Source: http://www.codeproject.com/KB/books/Start_SSRS/11.jpg

Finally, the Report name needs to be entered and done; you got your first report in place.

Figure 14.11: Complete the report



Completing the Report Wizard

Provide a name and click Finish to create the new report.

Report name: Report1

Report summary:

Data source: Northwind

Connection string: data source=IBP1007;initial catalog=Northwind

Report type: Table

Layout type: Stepped

Style: Bold

Page: ProductID

Details: ProductName, SupplierID, CategoryID, QuantityPerUnit, UnitPrice, UnitsInStock, UnitsOnOrder, ReorderLevel, Discontinued

```
Query: SELECT    Products.*
FROM       Products
```

☐ Preview report

Source: http://www.codeproject.com/KB/books/Start_SSRS/12.jpg

You can preview the report to change the data specifications if required. Once done, press Ctrl+F5 and the deployment of the report will occur.

Task Prepare a presentation on the steps of generating report using SQL server 2008.

Self Assessment

State whether the following statements are true or false:

3. SSRS supplies some extensions towards the data rendering, consignment and security of reports thereby allowing it to have a higher programmable ability.

4. SSRS is a comprehensive reporting platform whereby accounts are retained on a centralized web server (or set of servers).

5. The Report Manager is the central person who acts as a manager to decide when the reports will be scheduled to run along with maintaining the user profiles on the report server.

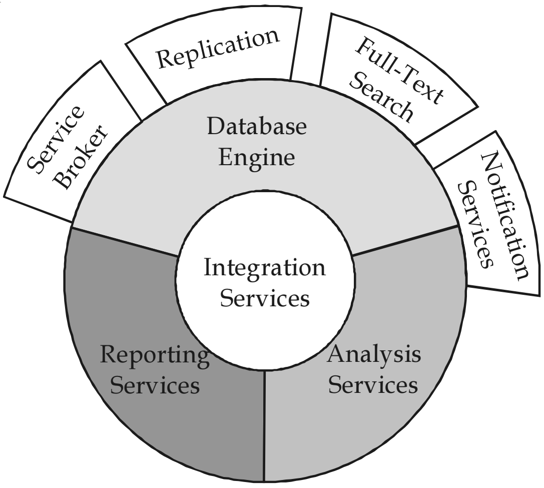
14.3 Analysis Services Using SQL Server

For business intelligence applications Microsoft SQL Server Analysis Services (SSAS) delivers Online Analytical Processing (OLAP) and data mining functionality. OLAP is supported by Analysis Services by letting you design, create, and manage multidimensional structures that contain data aggregated from other data sources, such as relational databases.

Analytical solutions are rapidly evolving mission critical for many companies. This has led to an explosion of data retained in these systems and a need to support bigger and quicker solutions that can be created and evolved quickly and effectively.

An Analysis Services instance is deployed in a mode that supports a memory and storage architecture that is optimized for all kinds of solutions. You can also install PowerPivot for SharePoint, which has its own internally set server mode. Each server mode is independent of the other mode; each one supports a type of analytical database that only runs in its modality.

Figure 14.12: Analysis Service



Source: http://i.msdn.microsoft.com/dynimg/IC6000.gif

Notes

Notes You cannot use SQL Server Management Studio to develop, manage, or query multidimensional data sets that were built by using PowerPivot for Excel.

Self Assessment

Fill in the blanks:

6. OLAP is supported by .... by letting you design, create, and manage multidimensional structures.

7. An Analysis Services instance is deployed in a mode that supports a .... architecture that is optimized for all kinds of solutions.

Case Study

Microsoft SQL Server 2008

Global aviation fuel companies supplying airlines in the current fragile economic climate are highly sensitised to potential airline collapses. The liability of unpaid invoices or even un-invoiced deliveries can significantly harm the profit margins within an extremely competitive market. Getting this information into the hands of the people who can make some sound commercial decisions was paramount for one such company.

Microsoft® SQL Server® 2008

Following a successful pilot in 2007 and the rapidly changed economy in Q4 2008, BMN was engaged to turn around an analysis and reporting tool within a matter of weeks. By using the right tools and approach, the results have been astonishingly productive with a solution that puts the power in the hands of the end user.

Solution:

The principal project objective was to enable the creation, management, and delivery of business intelligence reports for the analysis of customer account open debt and priced un-invoiced deliveries from all account receivable systems.

This required the consolidation of data feeds from multiple regional accounting systems into a single data warehouse against which data cubes and their dependent reports could be developed.

A SQL Server 2005 Business Intelligence solution was implemented to meet the customer requirements of accuracy, flexibility and performance.

Features:

- Centralised credit exposure reporting environment allowing clear identification and analysis of account information by account holders
- A complex data consolidation procedure to ensure data is up-to-date for global account managers

Contd....

- Reports run daily and are easily available to users who are travelling for them to down load and view
- Timing of the extracts takes into account when updates are posted to ensure the data is as up to date as possible
- Reports can be exported in multiple file formats to provide accessibility for further analysis offline
- Optimized data extract procedures and schedules to ensure reporting does not impact on other accounting system functions

Deliverables:

- SQL Server Integration Service components to manage the data extracts and scheduling from multiple disparate systems
- SQL Server Analysis Services (SSAS) components providing a foundation for Online Analytical Processing (OLAP) analysis and data mining. Used to create data-cubes of summarised, pre-calculated data for complex reporting solutions
- SQL Server Reporting Services components for report generation based upon data gathered from the data-cubes. This is the means by which users select, parameterise and run pre-defined reports

Question:

Analyse the case and provide any other solution to the problem.

Source: http://www.bmn.ltd.uk/CaseStudy_SQL_Reporting.aspx

14.4 Summary

- Following are the reporting tool functionalities: front end, data source connection capabilities, scheduling and distribution capabilities, security features, customization and export capabilities.
- There are two types of data sources: The relationship database and The OLAP multidimensional data source.
- Microsoft has come up with its own reporting service, in conjunction with SQL server database to insert the Microsoft SQL Server Reporting Services [SSRS].
- SSRS is a comprehensive reporting platform whereby accounts are retained on a centralized web server (or set of servers).
- After using SSRS, the architecture is just like a small operating system. The Report Manager is the central person who acts as a manager to decide when the reports will be scheduled to run along with maintaining the user profiles on the report server.
- Report Designer is a graphical tool that is hosted within the Microsoft Visual Studio IDE. It provides tabbed windows for Data, Layout, and Preview that allow you to design a report interactively.
- After installing SQL server reporting services on your system, start the Visual studio IDE. Go to File -> New Project, and you will be shown a prompt with 'New project'.

Notes

14.5 Keywords

Ad Hoc Reporting: Ad Hoc Reporting allows end users to easily build their own reports and modify existing ones with little to no training.

Embedded Reporting: It enables organizations to embed reports directly into business applications and web portals, enabling users to consume reports within the context of their business process.

Interactive Sorting: Applying sort capabilities to a report enable users to sort the data by any of the columns the report contains in ascending or descending order.

Matrix: A format that supports row and column groups, and which can display aggregated summary data in the cells where row groups and column groups intersect one another, similarly to a pivot table or crosstab.

Report Designer: Report Designer is a graphical tool that is hosted within the Microsoft Visual Studio IDE.

Table: A tabular format in which data is displayed in rows and columns. You can create a hierarchy of rows to reflect groupings in your data and display group totals.

UDM: It provides an intermediate logical layer between the physical relational database used as the data source and the proprietary cube and dimension structures that are used to resolve user queries.

14.6 Review Questions

1. Discuss about reporting tool functionalities.
2. "After using SSRS, the architecture is just like a small operating system". Elaborate.
3. Discuss about working with SQL Server Reporting Services.
4. Explain about the analysis Services Using SQL Server.

Answers: Self Assessment

1. Relationship database
2. Reporting software
3. True
4. True
5. True
6. Analysis Services
7. Memory and storage

14.7 Further Readings

Books

Carlo Vercellis (2011). "*Business Intelligence: Data Mining and Optimization for Decision Making*". John Wiley & Sons.

David Loshin (2012). "*Business Intelligence: The Savvy Manager's Guide*". Newnes.

Elizabeth Vitt, Michael Luckevich, Stacia Misner (2010). "*Business Intelligence*". O'Reilly Media, Inc.

Rajiv Sabhrwal, Irma Becerra-Fernandez (2010). "*Business Intelligence*". John Wiley & Sons.

Swain Scheps (2013). *"Business Intelligence for Dummies"*. Wiley.

Online links

msdn.microsoft.com/en-us/library/ms159106.aspx

ww.microsoft.com/en-us/sqlserver/solutions.../business.../reporting.aspx

www.accelebrate.com/sql_training/ssrs_tutorial.htm

LOVELY PROFESSIONAL UNIVERSITY
Jalandhar-Delhi G.T. Road (NH-1)
Phagwara, Punjab (India)-144411
For Enquiry: +91-1824-521360
Fax.: +91-1824-506111
Email: odl@lpu.co.in

978-93-90164-20-2
9 789390 164202