DATA MINING

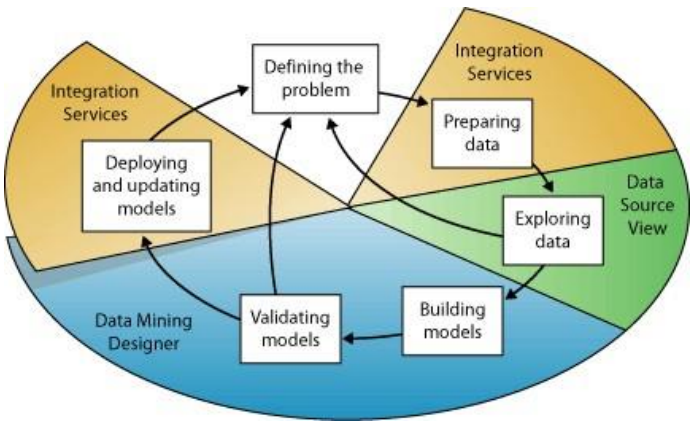
Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data. These patterns and trends can be collected and defined as a *data mining model*. Mining models can be applied to specific scenarios, such as:

- **Forecasting**: Estimating sales, predicting server loads or server downtime
- **Risk and probability**: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes
- **Recommendations**: Determining which products are likely to be sold together, generating recommendations
- **Finding sequences**: Analyzing customer selections in a shopping cart, predicting next likely events
- **Grouping**: Separating customers or events into cluster of related items, analyzing and predicting affinities

Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment. This process can be defined by using the following six basic steps:

1. Defining the Problem
2. Preparing Data
3. Exploring Data
4. Building Models
5. Exploring and Validating Models
6. Deploying and Updating Models

The following diagram describes the relationships between each step in the process, and the technologies in Microsoft SQL Server that you can use to complete each step.



Defining the Problem

The first step in the data mining process is to clearly define the problem, and consider ways that data can be utilized to provide an answer to the problem. This step includes analyzing business requirements, defining the scope of the problem, defining the metrics by which the model will be evaluated, and defining specific objectives for the data mining project. These tasks translate into questions such as the following:

- What are you looking for? What types of relationships are you trying to find?
- Does the problem you are trying to solve reflect the policies or processes of the business?
- Do you want to make predictions from the data mining model, or just look for interesting patterns and associations?
- Which outcome or attribute do you want to try to predict?
- What kind of data do you have and what kind of information is in each column? If there are multiple tables, how are the tables related? Do you need to perform any cleansing, aggregation, or processing to make the data usable?
- How is the data distributed? Is the data seasonal? Does the data accurately represent the processes of the business?

Preparing Data

- The second step in the data mining process is to consolidate and clean the data that was identified in the Defining the Problem step.
- Data can be scattered across a company and stored in different formats, or may contain inconsistencies such as incorrect or missing entries.
- Data cleaning is not just about removing bad data or interpolating missing values, but about finding hidden correlations in the data, identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis.

Exploring Data

Exploration techniques include calculating the minimum and maximum values, calculating mean and standard deviations, and looking at the distribution of the data. For example, you might determine by reviewing the maximum, minimum, and mean values that the data is not representative of your customers or business processes, and that you therefore must obtain more balanced data or review the assumptions that are the basis for your expectations. Standard deviations and other distribution values can provide useful information about the stability and accuracy of the results.

Building Models

The mining structure is linked to the source of data, but does not actually contain any data until you process it. When you process the mining structure, SQL Server Analysis Services generates aggregates and

other statistical information that can be used for analysis. This information can be used by any mining model that is based on the structure.

Exploring and Validating Models

Before you deploy a model into a production environment, you will want to test how well the model performs. Also, when you build a model, you typically create multiple models with different configurations and test all models to see which yields the best results for your problem and your data.

Deploying and Updating Models

After the mining models exist in a production environment, you can perform many tasks, depending on your needs. The following are some of the tasks you can perform:

- Use the models to create predictions, which you can then use to make business decisions.
- Create content queries to retrieve statistics, rules, or formulas from the model.
- Embed data mining functionality directly into an application. You can include Analysis Management Objects (AMO), which contains a set of objects that your application can use to create, alter, process, and delete mining structures and mining models.
- Use Integration Services to create a package in which a mining model is used to intelligently separate incoming data into multiple tables.
- Create a report that lets users directly query against an existing mining model
- Update the models after review and analysis. Any update requires that you reprocess the models.
- Update the models dynamically, as more data comes into the organization, and making constant changes to improve the effectiveness of the solution should be part of the deployment strategy.

DATA WAREHOUSING

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Characteristics of data warehouse

The main characteristics of a data warehouse are as follows:

- **Subject-Oriented:** A data warehouse is subject-oriented since it provides topic-wise information rather than the overall processes of a business. Such subjects may be sales, promotion, inventory, etc
- **Integrated:** A data warehouse is developed by integrating data from varied sources into a consistent format. The data must be stored in the warehouse in a consistent and universally acceptable manner in terms of naming, format, and coding. This facilitates effective data analysis.

- **Non-Volatile** Data once entered into a data warehouse must remain unchanged. All data is read-only. Previous data is not erased when current data is entered. This helps you to analyze what has happened and when.
- **Time-Variant** The data stored in a data warehouse is documented with an element of time, either explicitly or implicitly. An example of time variance in Data Warehouse is exhibited in the Primary Key, which must have an element of time like the day, week, or month.

Database vs. Data Warehouse

Although a data warehouse and a traditional database share some similarities, they need not be the same idea. The main difference is that in a database, data is collected for multiple transactional purposes. However, in a data warehouse, data is collected on an extensive scale to perform analytics. Databases provide real-time data, while warehouses store data to be accessed for big analytical queries.

Data Warehouse Architecture

Usually, data warehouse architecture comprises a three-tier structure.

Bottom Tier

The bottom tier or data warehouse server usually represents a relational database system. Back-end tools are used to cleanse, transform and feed data into this layer.

Middle Tier

The middle tier represents an OLAP server that can be implemented in two ways. The ROLAP or Relational OLAP model is an extended relational database management system that maps multidimensional data process to standard relational process. The MOLAP or multidimensional OLAP directly acts on multidimensional data and operations.

Top Tier

This is the front-end client interface that gets data out from the data warehouse. It holds various tools like query tools, analysis tools, reporting tools, and data mining tools.

How Data Warehouse Works

Data Warehousing integrates data and information collected from various sources into one comprehensive database. For example, a data warehouse might combine customer information from an organization's point-of-sale systems, its mailing lists, website, and comment cards. It might also incorporate confidential information about employees, salary information, etc. Businesses use such components of data warehouse to analyze customers.

Data mining is one of the features of a data warehouse that involves looking for meaningful data patterns in vast volumes of data and devising innovative strategies for increased sales and profits.

Types of Data Warehouse

There are three main types of data warehouse.

Enterprise Data Warehouse (EDW)

This type of warehouse serves as a key or central database that facilitates decision-support services throughout the enterprise. The advantage to this type of warehouse is that it provides access to cross-organizational information, offers a unified approach to data representation, and allows running complex queries.

Operational Data Store (ODS)

This type of data warehouse refreshes in real-time. It is often preferred for routine activities like storing employee records. It is required when data warehouse systems do not support reporting needs of the business.

Data Mart

A data mart is a subset of a data warehouse built to maintain a particular department, region, or business unit. Every department of a business has a central repository or data mart to store data. The data from the data mart is stored in the ODS periodically. The ODS then sends the data to the EDW, where it is stored and used.

BASIC STATISTICAL DESCRIPTIONS OF DATA

Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

Measuring the Central Tendency: Mean, Median, and Mode

The attribute X, like salary, is recorded for a set of objects. Let x1,x2,...,xN be the set of N observed values or observations for X. Here, these values may also be referred to as the data set (for X). To plot the observations for salary, where most of the values fall gives us an idea of the central tendency of the data.

Measures of central tendency include the mean, median, mode, and midrange. The most common and effective numeric measure of the "center" of a set of data is the (arithmetic) mean. Let x1,x2,...,xN be a set of N values or observations, such as for some numeric attribute X, like salary. The mean of this set of values is

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

This corresponds to the built-in aggregate function, average (avg() in SQL), provided in relational database systems.

Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation and Interquartile Range

The measures such as range, quantiles, quartiles, percentiles, and the interquartile range are used to assess the dispersion or spread of numeric data. The five-number summary which can be displayed as a

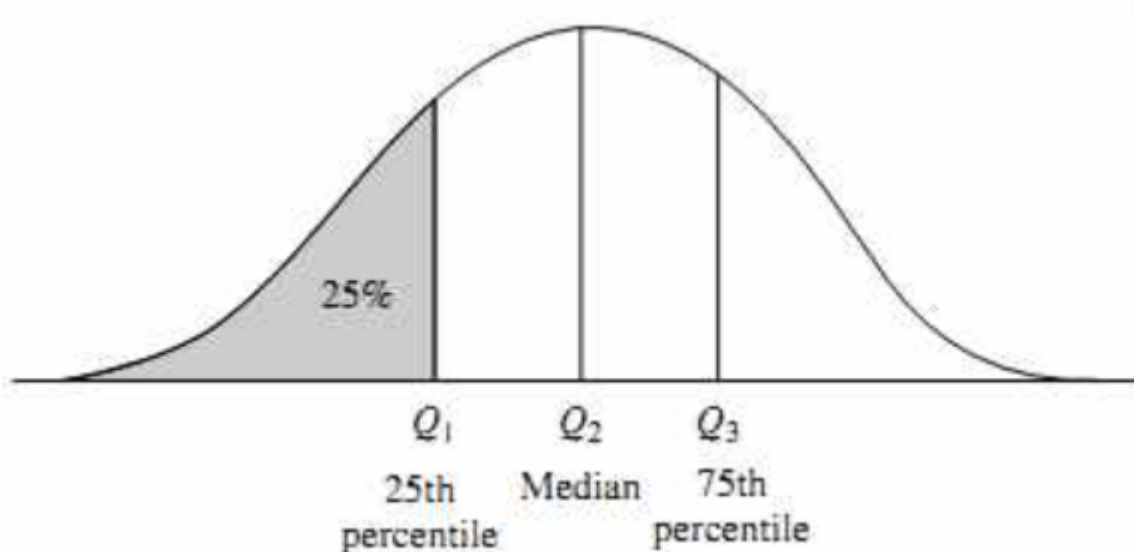
boxplot is useful in identifying outliers. Variance and standard deviation also indicate the spread of a data distribution.

Range, Quartiles and Interquartile Range

Let x1,x2,...,xN be a set of observations for some numeric attribute, X. The range of the set is the difference between the largest (max()) and smallest (min()) values. Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.

The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.

The 100-quantiles are more commonly referred to as percentiles, they divide the data distribution into 100 equal-sized consecutive sets. The median, quartiles, and percentiles are the most widely used forms of quantiles.



The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median. The quartiles give an indication of a distribution's center, spread and shape.

The first quartile, denoted by Q1, is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by Q3, is the 75th percentile. It cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

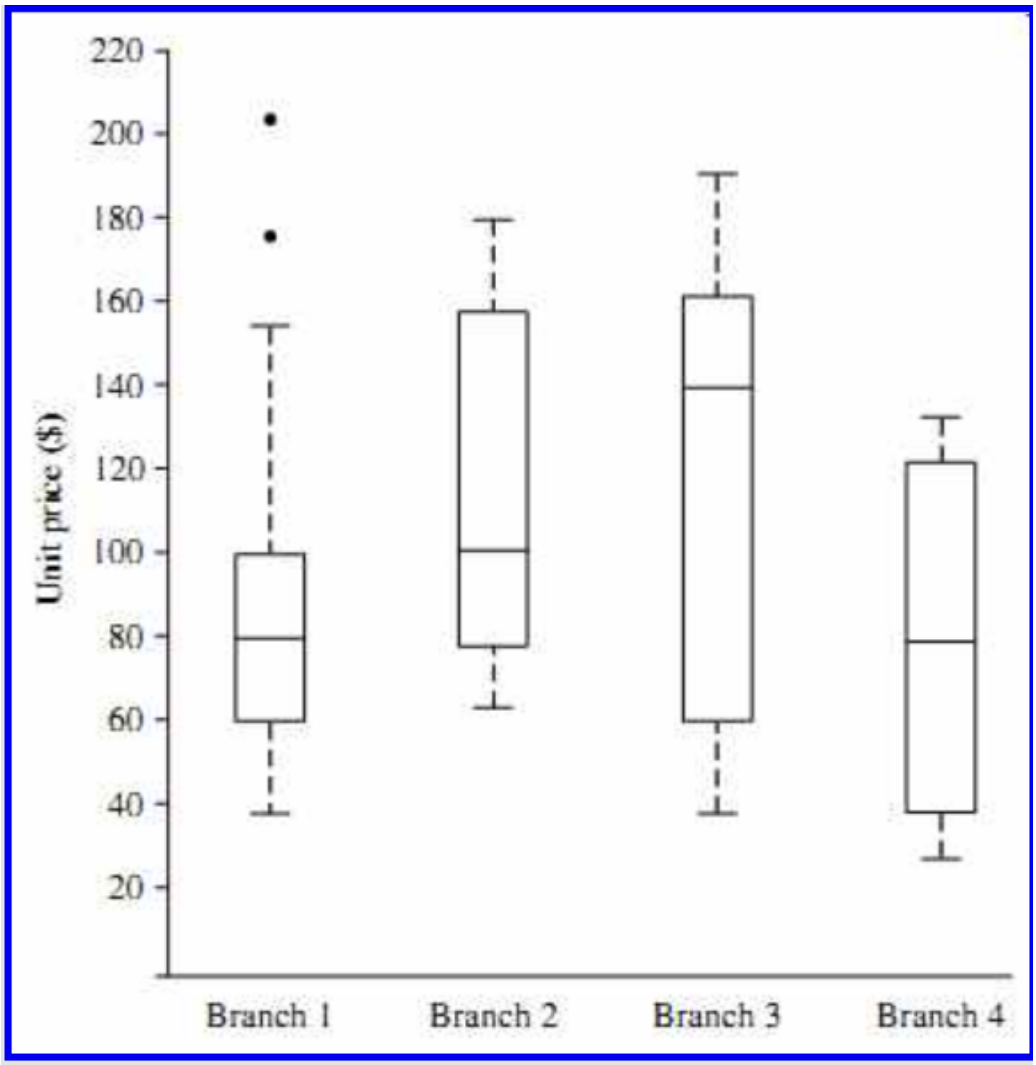
The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the inter quartile range (IQR) and is defined as

$$IQR = Q3 - Q1$$

Five-Number Summary, Boxplots and Outliers

The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, Maximum. Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

- Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.
- Two lines called whiskers outside the box extend to the smallest (Minimum) and largest (Maximum) observations.



Variance and Standard Deviation

Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

The variance of N observations, x1,x2,...,xN, for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

where x is the mean value of the observations, as defined in Equation. The standard deviation, σ, of the observations is the square root of the variance, σ2.

Graphic Displays of Basic Statistical Descriptions of Data

Quantile plots, quantile–quantile plots, histograms, and scatter plots. These graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).