Data Warehouse

Introduction

Data warehouses are special types of database, which are built for the specific purpose of getting certain kinds of information. In general terms, data warehousing involves the collection of data from potentially multiple sources and then managing that data in a single database.

The emphasis is on supporting questions of a strategic nature, to assist the managers of organizations in planning for the future. Thus a data warehouse is used as an aid to understand the trends within a business and to respond, with the relevant data, to broad queries to enable decisions to be made about large areas of the business.

A data warehouse is

- Subject oriented; data organized around subjects
- Non-volatile; data, once placed, is not usually subject to change
- Integrated; data is consistent (same format/semantic, comparable)
- Time variant; records historical data

The Comparison: data warehouse is almost always kept separate from the operational databases, which support the day-to-day operations in a business because:

1. Operational databases are optimized to execute queries that update the database whereas data warehouses are optimized to execute queries that leave the database unchanged;
2. Operational databases are 'twinkling' (constantly changing) whereas data warehouses are quiet (and accumulating data slowly);
3. Operational databases have large and complex schemas whereas data warehouses are simplified structures which have not necessarily been normalized;
4. Data warehouses need historical information and this is usually missing (or difficult to access) from operational.

Dimensional Analysis

Dimensional analysis is a technique used in identifying the requirements of a data warehouse and this is often depicted using a star schema. The star schema identifies the facts and the dimensions of analysis. A fact, such as sales value or call duration, is an attribute which is analyzed across dimensions. Dimensions are things like 'Customers' and 'products' over which the facts are analyzed. A typical query might be:
Show me the sales value of products by customer for this month and last month.

Time is always a dimension of analysis.

Activities: see the dimensional analysis example for Wine Club in p77, subject is sale, dimensions are product, member, area and time. Ready? Then turn to p.85 to see how it is built and how it can support decision-making.

There are five main components to a data warehouse:

1. **extraction component** extracts source data from a variety of application databases. These source applications often use very difficult technology.
2. **integration component** ensures consistency of data from diverse sources. There are two types of integration. First there is format integration where logically similar data types (e.g. dates) are converted so that they have the same physical data type. Second, semantic interaction so that the meaning of the information is consistent.
3. **database component** stores and manages the data. The data warehouse database can become enormous as a new layer of fact data is added each day. The star schema is implemented as a number of tables. The fact table (the center of the star), is long and thin in that it usually has a large number of rows and a small number of columns. The fact columns must be summable. The dimensions tables (the points of the star) are joined to be the fact table through foreign keys.

4. **aggregate navigation** component enables the users to have their queries automatically directed at aggregate table without being aware that is happening. This is very important for query performance.
5. **presentation component** presents data to the users of the data warehouse. Most implementations opt for a client-server approach with client tools that give them the capability to view information in a variety of tabular or graphical formats.

Data Mining

The Operations

Affinity grouping; identify simple relationships within data; e.g. Cheese+Wine always appear in one bill;
Sequence discovery; discover repeated patterns of behavior; e.g. Rice/ToiletPaperRolls… could be given discounts on bulk purchase;
Classification; find a model that classifies a case into one of server predefined classes (case – a row; classification – a column); e.g. JC card holder -> Jusco loyal customer group
Estimation; Classification deals with discrete outcomes, Estimation deal with continuous. Typically it is used prior to Classification; e.g. Estimation -> a model to assign probability that JC card holder will be wine consumer and vulnerable to promotional offer; Classification (PromoteVulnerable) with threshold score -> target customer
Regression; get the model to predict unknown or future value of a set of attribute from another set of attributes (the trend); e.g. see "Annual spend vs. Income" in Fig 4.4 on page 113
Clustering; assigning cases into one of several clusters; unlike Classes, it is not predefined but natural grouping of cases sharing common attributes; e.g. identify a particular region of country (Chinese) -> (live Chicken) is popular

The KDD process

Data selection: creating a target data set
A database contains a variety of diverse data not all of which needs to be analyzed to achieve a particular knowledge discovery goal. The first step to determine the data that is required is by applying some selection criteria so that subsets of the data can be produced. Including irrelevant data will slow down the knowledge discovery process because it considers all the data that is presented when searching for patterns within the data. For example, the supermarket database contains data describing customer purchases, demographic and lifestyle data. To identify how products should be distributed on the shelves in the store possibly only purchase data is needed.

Data cleaning and preprocessing
The data must be cleaned to remove any incorrect or inconsistent data where possible. The analyst will need o decide on the strategy for dealing with missing data. If the selected data is stored in several related tables they may be joined at this stage to create a single table which is more amenable to data mining methods.

Data transformation
It is usually necessary to perform certain transformations on the data. The type of transformation is dictated by the type of data mining operation performed and the data mining method used. Transformations vary from conversions of one type of data to another (for example, enumerating string values, such as 1=yes, 2-No, 3=n/a), grouping continuous values into ranges (for example, grouping salaries into income groups), defining new or derived attributes by performing arithmetic or logical transformations (for example, deriving ratios of two attributes).

Data mining
The transformed data is searched using one or more data mining methods to try to find patterns in the data. The search attempts to locate significant patterns y formulating hypothesis about the data, building models by applying them to the data to affirm or negate the hypotheses. It may be necessary to access additional data from a data warehouse, for example, when performing the data mining step and / or perform further transformations on the originally selected data.

Integration and evaluation
The patterns identified by the data mining step are integrated into knowledge which can be used to support the decision-making process, summarizing the contents of the database or explaining observed phenomena. The information identified may be presented to the analyst through a decision-support system. Therefore, the purchase of the

interpretation of the patterns discovered is not only to visualize the output of the data mining operation, but also to filter the information that will be presented to the analyst through a decision-support system.