## 1 Data mining

### 1.1 What is data mining?

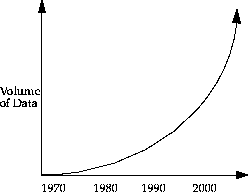
The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. The increase in use of electronic data gathering devices such as point-of-sale or remote sensing devices has contributed to this explosion of available data. Figure 1 from the Red Brick company illustrates the data explosion.  


Figure 1: The Growing Base of Data

Data storage became easier as the availability of large amounts of computing power at low cost ie the cost of processing power and storage is falling, made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc. in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power.

Having concentrated so much attention on the accumulation of data the problem was what to do with this valuable resource? It was recognised that information is at the heart of business operations and that decision-makers could make use of the data stored to gain valuable insight into the business. Database Management systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction processing systems, OLTPs, are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analysing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. This is where Data Mining or Knowledge Discovery in Databases (KDD) has obvious benefits for any enterprise.

The term data mining has been stretched beyond its limits to apply to any form of data analysis. Some of the numerous definitions of Data Mining, or Knowledge Discovery in Databases are:

Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency net works, analysing changes, and detecting anomalies.

William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database.

Marcel Holshemier & Arno Siebes (1994)

The analogy with the mining process is described as:

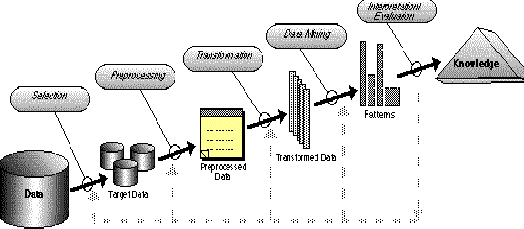
Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful"

Clementine User Guide, a data mining toolkit

Basically data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data. The idea is that it is possible to strike gold in unexpected places as the data mining software extracts patterns not previously discernable or so obvious that no-one has noticed them before.

Data mining analysis tends to work from the data up and the best techniques are those developed with an orientation towards large volumes of data, making use of as much of the collected data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data. Again this is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value.

The following diagram summarises the some of the stages/processes identified in data mining and knowledge discovery by Usama Fayyad & Evangelos Simoudis, two of leading exponents of this area.



The phases depicted start with the raw data and finish with the extracted knowledge which was acquired as a result of the following stages:

* Selection - selecting or segmenting the data according to some criteria e.g. all those people who own a car, in this way subsets of the data can be determined.
* Preprocessing - this is the data cleansing stage where certain information is removed which is deemed unnecessary and may slow down queries for example unnecessary to note the sex of a patient when studying pregnancy. Also the data is reconfigured to ensure a consistent format as there is a possibility of inconsistent formats because the data is drawn from several sources e.g. sex may recorded as f or m and also as 1 or 0.
* Transformation - the data is not merely transferred across but transformed in that overlays may added such as the demographic overlays commonly used in market research. The data is made useable and navigable.
* Data mining - this stage is concerned with the extraction of patterns from the data. A pattern can be defined as given a set of facts(data) F, a language L, and some measure of certainty C a pattern is a statement S in L that describes relationships among a subset Fs of F with a certainty c such that S is simpler in some sense than the enumeration of all the facts in Fs.
* Interpretation and evaluation - the patterns identified by the system are interpreted into knowledge which can then be used to support human decision-making e.g. prediction and classification tasks, summarizing the contents of a database or explaining observed phenomena.

### 1.2 Data mining background

Data mining research has drawn on a number of other fields such as inductive learning, machine learning and statistics etc.

#### 1.2.1 Inductive learning

Induction is the inference of information from data and inductive learning is the model building process where the environment i.e. database is analysed with a view to finding patterns. Similar objects are grouped in classes and rules formulated whereby it is possible to predict the class of unseen objects. This process of classification identifies classes such that each class has a unique pattern of values which forms the class description. The nature of the environment is dynamic hence the model must be adaptive i.e. should be able learn.

Generally it is only possible to use a small number of properties to characterise objects so we make abstractions in that objects which satisfy the same subset of properties are mapped to the same internal representation.

Inductive learning where the system infers knowledge itself from observing its environment has two main strategies:

* supervised learning - this is learning from examples where a teacher helps the system construct a model by defining classes and supplying examples of each class. The system has to find a description of each class i.e. the common properties in the examples. Once the description has been formulated the description and the class form a classification rule which can be used to predict the class of previously unseen objects. This is similar to discriminate analysis as in statistics.
* unsupervised learning - this is learning from observation and discovery. The data mine system is supplied with objects but no classes are defined so it has to observe the examples and recognise patterns (i.e. class description) by itself. This system results in a set of class descriptions, one for each class discovered in the environment. Again this similar to cluster analysis as in statistics.

Induction is therefore the extraction of patterns. The quality of the model produced by inductive learning methods is such that the model could be used to predict the outcome of future situations in other words not only for states encountered but rather for unseen states that could occur. The problem is that most environments have different states, i.e. changes within, and it is not always possible to verify a model by checking it for all possible situations.

Given a set of examples the system can construct multiple models some of which will be simpler than others. The simpler models are more likely to be correct if we adhere to Ockhams razor, which states that if there are multiple explanations for a particular phenomena it makes sense to choose the simplest because it is more likely to capture the nature of the phenomenon.

#### 1.2.2 Statistics

Statistics has a solid theoretical foundation but the results from statistics can be overwhelming and difficult to interpret as they require user guidance as to where and how to analyse the data. Data mining however allows the expert's knowledge of the data and the advanced analysis techniques of the computer to work together.

Statistical analysis systems such as SAS and SPSS have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models. Statistics have a role to play and data mining will not replace such analyses but rather they can act upon more directed analyses based on the results of data mining. For example statistical induction is something like the average rate of failure of machines.

#### 1.2.3 Machine Learning

Machine learning is the automation of a learning process and learning is tantamount to the construction of rules based on observations of environmental states and transitions. This is a broad field which includes not only learning from examples, but also reinforcement learning, learning with teacher, etc. A learning algorithm takes the data set and its accompanying information as input and returns a statement e.g. a concept representing the results of learning as output. Machine learning examines previous examples and their outcomes and learns how to reproduce these and make generalisations about new cases.

Generally a machine learning system does not use single observations of its environment but an entire finite set called the training set at once. This set contains examples i.e. observations coded in some machine readable form. The training set is finite hence not all concepts can be learned exactly.

#### 1.2.4 Differences between Data Mining and Machine Learning

Knowledge Discovery in Databases (KDD) or Data Mining, and the part of Machine Learning (ML) dealing with learning from examples overlap in the algorithms used and the problems addressed.

The main differences are:

* KDD is concerned with finding understandable knowledge, while ML is concerned with improving performance of an agent. So training a neural network to balance a pole is part of ML, but not of KDD. However, there are efforts to extract knowledge from neural networks which are very relevant for KDD.
* KDD is concerned with very large, real-world databases, while ML typically (but not always) looks at smaller data sets. So efficiency questions are much more important for KDD.
* ML is a broader field which includes not only learning from examples, but also reinforcement learning, learning with teacher, etc.

KDD is that part of ML which is concerned with finding understandable knowledge in large sets of real-world examples. When integrating machine learning techniques into database systems to implement KDD some of the databases require:

* more efficient learning algorithms because realistic databases are normally very large and noisy. It is usual that the database is often designed for purposes different from data mining and so properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Databases are usually contaminated by errors so the data mining algorithm has to cope with noise whereas ML has laboratory type examples i.e. as near perfect as possible.
* more expressive representations for both data, e.g. tuples in relational databases, which represent instances of a problem domain, and knowledge, e.g. rules in a rule-based system, which can be used to solve users' problems in the domain, and the semantic information contained in the relational schemata.

Practical KDD systems are expected to include three interconnected phases

* Translation of standard database information into a form suitable for use by learning facilities;
* Using machine learning techniques to produce knowledge bases from databases; and
* Interpreting the knowledge produced to solve users' problems and/or reduce data spaces. Data spaces being the number of examples.

### 1.3 Data Mining Models

IBM have identified two types of model or modes of operation which may be used to unearth information of interest to the user.

#### 1.3.1 Verification Model

The verification model takes an hypothesis from the user and tests the validity of it against the data. The emphasis is with the user who is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis.

In a marketing division for example with a limited budget for a mailing campaign to launch a new product it is important to identify the section of the population most likely to buy the new product. The user formulates an hypothesis to identify potential customers and the characteristics they share. Historical data about customer purchase and demographic information can then be queried to reveal comparable purchases and the characteristics shared by those purchasers which in turn can be used to target a mailing campaign. The whole operation can be refined by `drilling down' so that the hypothesis reduces the `set' returned each time until the required limit is reached.

The problem with this model is the fact that no new information is created in the retrieval process but rather the queries will always return records to verify or negate the hypothesis. The search process here is iterative in that the output is reviewed, a new set of questions or hypothesis formulated to refine the search and the whole process repeated. The user is discovering the facts about the data using a variety of techniques such as queries, multidimensional analysis and visualization to guide the exploration of the data being inspected.

#### 1.3.2 Discovery Model

The discovery model differs in its emphasis in that it is the system automatically discovering important information hidden in the data. The data is sifted in search of frequently occurring patterns, trends and generalisations about the data without intervention or guidance from the user. The discovery or data mining tools aim to reveal a large number of facts about the data in as short a time as possible.

An example of such a model is a bank database which is mined to discover the many groups of customers to target for a mailing campaign. The data is searched with no hypothesis in mind other than for the system to group the customers according to the common characteristics found.

### 1.4 Data Warehousing

Data mining potential can be enhanced if the appropriate data has been collected and stored in a data warehouse. A data warehouse is a relational database management system (RDMS) designed specifically to meet the needs of transaction processing systems. It can be loosely defined as any centralised data repository which can be queried for business benefit but this will be more clearly defined later. Data warehousing is a new powerful technique making it possible to extract archived operational data and overcome inconsistencies between different legacy data formats. As well as integrating data throughout an enterprise, regardless of location, format, or communication requirements it is possible to incorporate additional or expert information. It is,

the logical link between what the managers see in their decision support EIS applications and the company's operational activities

John McIntyre of SAS Institute Inc

In other words the data warehouse provides data that is already transformed and summarized, therefore making it an appropriate environment for more efficient DSS and EIS applications.

#### 1.4.1 Characteristics of a data warehouse

According to Bill Inmon, author of Building the Data Warehouse and the guru who is widely considered to be the originator of the data warehousing concept, there are generally four characteristics that describe a data warehouse:

* subject-oriented: data are organized according to subject instead of application e.g. an insurance company using a data warehouse would organize their data by customer, premium, and claim, instead of by different products (auto, life, etc.). The data organized by subject contain only the information necessary for decision support processing.
* integrated: When data resides in many separate applications in the operational environment, encoding of data is often inconsistent. For instance, in one application, gender might be coded as "m" and "f" in another by 0 and 1. When data are moved from the operational environment into the data warehouse, they assume a consistent coding convention e.g. gender data is transformed to "m" and "f".
* time-variant: The data warehouse contains a place for storing data that are five to 10 years old, or older, to be used for comparisons, trends, and forecasting. These data are not updated.
* non-volatile: Data are not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

#### 1.4.2 Processes in data warehousing

The first phase in data warehousing is to "insulate" your current operational information, ie to preserve the security and integrity of mission-critical OLTP applications, while giving you access to the broadest possible base of data. The resulting database or data warehouse may consume hundreds of gigabytes - or even terabytes - of disk space, what is required then are efficient techniques for storing and retrieving massive amounts of information. Increasingly, large organizations have found that only parallel processing systems offer sufficient bandwidth.

The data warehouse thus retrieves data from a variety of heterogeneous operational databases. The data is then transformed and delivered to the data warehouse/store based on a selected model (or mapping definition). The data transformation and movement processes are executed whenever an update to the warehouse data is required so there should some form of automation to manage and execute these functions. The information that describes the model and definition of the source data elements is called "metadata". The metadata is the means by which the end-user finds and understands the data in the warehouse and is an important part of the warehouse. The metadata should at the very least contain;

* the structure of the data;
* the algorithm used for summarization;
* and the mapping from the operational environment to the data warehouse.

Data cleansing is an important aspect of creating an efficient data warehouse in that it is the removal of certain aspects of operational data, such as low-level transaction information, which slow down the query times. The cleansing stage has to be as dynamic as possible to accommodate all types of queries even those which may require low-level information. Data should be extracted from production sources at regular intervals and pooled centrally but the cleansing process has to remove duplication and reconcile differences between various styles of data collection.

Once the data has been cleaned it is then transferred to the data warehouse which typically is a large database on a high performance box either SMP, Symmetric Multi-Processing or MPP, Massively Parallel Processing. Number-crunching power is another important aspect of data warehousing because of the complexity involved in processing ad hoc queries and because of the vast quantities of data that the organisation want to use in the warehouse. A data warehouse can be used in different ways for example it can be used as a central store against which the queries are run or it can be used to like a data mart. Data marts which are small warehouses can be established to provide subsets of the main store and summarised information depending on the requirements of a specific group/department. The central store approach generally uses very simple data structures with very little assumptions about the relationships between data whereas marts often use multidimensional databases which can speed up query processing as they can have data structures which are reflect the most likely questions.

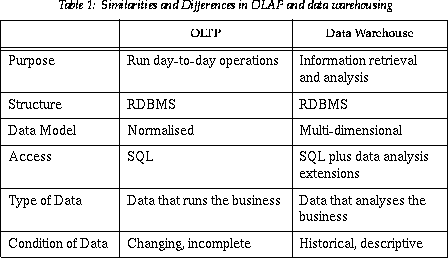
Many vendors have products that provide one or more of the above described data warehouse functions. However, it can take a significant amount of work and specialized programming to provide the interoperability needed between products from multiple vendors to enable them to perform the required data warehouse processes. A typical implementation usually involves a mixture of products from a variety of suppliers.

Another approach to data warehousing is Parsaye's Sandwich Paradigm put forward by Dr. Kamran Parsaye, CEO of Information Discovery, Hermosa Beach, CA. This paradigm or philosophy encourages acceptance of the probability that the first iteration of a data-warehousing effort will require considerable revision. The Sandwich Paradigm advocates the following approach:

* pre-mine the data to determine what formats and data are needed to support a data-mining application;
* build a prototype mini-data warehouse i.e the meat of the sandwich, with most of the features envisaged for the end product;
* revise the strategies as necessary;
* build the final warehouse.

#### 1.4.3 Data warehousing and OLTP systems

A database which is built for on line transaction processing, OLTP, is generally regarded as unsuitable for data warehousing as they have been designed with a different set of needs in mind ie maximising transaction capacity and typically having hundreds of tables in order not to lock out users etc. Data warehouses are interested in query processing as opposed to transaction processing.

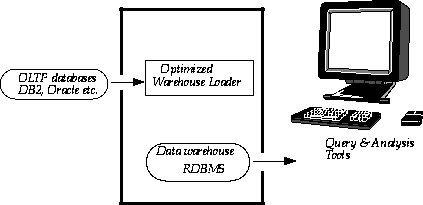
OLTP systems cannot be repositories of facts and historical data for business analysis. They cannot quickly answer ad hoc queries and rapid retrieval is almost impossible. The data is inconsistent and changing, duplicate entries exist, entries can be missing and there is an absence of historical data which is necessary to analyse trends. Basically OLTP offers large amounts of raw data which is not easily understood. The data warehouse offers the potential to retrieve and analyse information quickly and easily. Data warehouses do have similarities with OLTP as shown in the table below.  


The data warehouse serves a different purpose from that of OLTP systems by allowing business analysis queries to be answered as opposed to "simple aggregations" such as `what is the current account balance for this customer?' Typical data warehouse queries include such things as `which product line sells best in middle-America and how does this correlate to demographic data?'

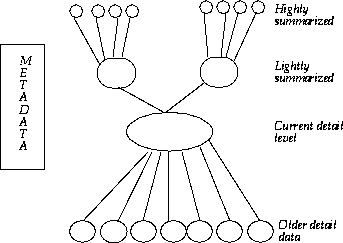
#### 1.4.4 The Data Warehouse model

Data warehousing is the process of extracting and transforming operational data into informational data and loading it into a central data store or warehouse. Once the data is loaded it is accessible via desktop query and analysis tools by the decision makers.

The data warehouse model is illustrated in the following diagram.

Figure 2: A data warehouse model  


The data within the actual warehouse itself has a distinct structure with the emphasis on different levels of summarization as shown in the figure below.

Figure 3: The structure of data inside the data warehouse  


The current detail data is central in importance as it:

* reflects the most recent happenings, which are usually the most interesting;
* it is voluminous as it is stored at the lowest level of granularity;
* it is always (almost) stored on disk storage which is fast to access but expensive and complex to manage.

Older detail data is stored on some form of mass storage, it is infrequently accessed and stored at a level detail consistent with current detailed data.

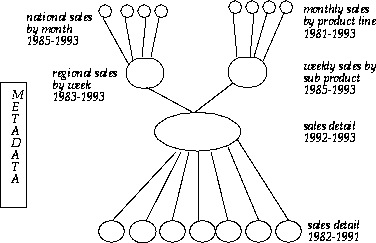
Lightly summarized data is data distilled from the low level of detail found at the current detailed level and generally is stored on disk storage. When building the data warehouse have to consider what unit of time is summarization done over and also the contents or what attributes the summarized data will contain.

Highly summarized data is compact and easily accessible and can even be found outside the warehouse.

Metadata is the final component of the data warehouse and is really of a different dimension in that it is not the same as data drawn from the operational environment but is used as:

* a directory to help the DSS analyst locate the contents of the data warehouse,
* a guide to the mapping of data as the data is transformed from the operational environment to the data warehouse environment,
* a guide to the algorithms used for summarization between the current detailed data and the lightly summarized data and the lightly summarized data and the highly summarized data, etc.

The basic structure has been described but Bill Inmon fills in the details to make the example come alive as shown in the following diagram.

Figure 4: An example of levels of summarization of data inside the data warehouse  


The diagram assumes the year is 1993 hence the current detail data is 1992-93. Generally sales data doesn't reach the current level of detail for 24 hours as it waits until it is no longer available to the operational system i.e. it takes 24 hours for it to get to the data warehouse. Sales details are summarized weekly by subproduct and region to produce the lightly summarized detail. Weekly sales are then summarized again to produce the highly summarized data.

#### 1.4.5 Problems with data warehousing

One of the problems with data mining software has been the rush of companies to jump on the band wagon as

these companies have slapped `data warehouse' labels on traditional transaction-processing products, and co-opted the lexicon of the industry in order to be considered players in this fast-growing category.

Chris Erickson, president and CEO of Red Brick (HPCwire, Oct. 13, 1995)

Red Brick Systems have established a criteria for a relational database management system (RDBMS) suitable for data warehousing, and documented 10 specialized requirements for an RDBMS to qualify as a relational data warehouse server, this criteria is listed in the next section.

According to Red Brick, the requirements for data warehouse RDBMSs begin with the loading and preparation of data for query and analysis. If a product fails to meet the criteria at this stage, the rest of the system will be inaccurate, unreliable and unavailable.

#### 1.4.6 Criteria for a data warehouse

The criteria for data warehouse RDBMSs are as follows:

* Load Performance - Data warehouses require incremental loading of new data on a periodic basis within narrow time windows; performance of the load process should be measured in hundreds of millions of rows and gigabytes per hour and must not artificially constrain the volume of data required by the business.
* Load Processing - Many steps must be taken to load new or updated data into the data warehouse including data conversions, filtering, reformatting, integrity checks, physical storage, indexing, and metadata update. These steps must be executed as a single, seamless unit of work.
* Data Quality Management - The shift to fact-based management demands the highest data quality. The warehouse must ensure local consistency, global consistency, and referential integrity despite "dirty" sources and massive database size. While loading and preparation are necessary steps, they are not sufficient. Query throughput is the measure of success for a data warehouse application. As more questions are answered, analysts are catalysed to ask more creative and insightful questions.
* Query Performance - Fact-based management and ad-hoc analysis must not be slowed or inhibited by the performance of the data warehouse RDBMS; large, complex queries for key business operations must complete in seconds not days.
* Terabyte Scalability - Data warehouse sizes are growing at astonishing rates. Today these range from a few to hundreds of gigabytes, and terabyte-sized data warehouses are a near-term reality. The RDBMS must not have any architectural limitations. It must support modular and parallel management. It must support continued availability in the event of a point failure, and must provide a fundamentally different mechanism for recovery. It must support near-line mass storage devices such as optical disk and Hierarchical Storage Management devices. Lastly, query performance must not be dependent on the size of the database, but rather on the complexity of the query.
* Mass User Scalability - Access to warehouse data must no longer be limited to the elite few. The RDBMS server must support hundreds, even thousands, of concurrent users while maintaining acceptable query performance.
* Networked Data Warehouse - Data warehouses rarely exist in isolation. Multiple data warehouse systems cooperate in a larger network of data warehouses. The server must include tools that coordinate the movement of subsets of data between warehouses. Users must be able to look at and work with multiple warehouses from a single client workstation. Warehouse managers have to manage and administer a network of warehouses from a single physical location.
* Warehouse Administration - The very large scale and time-cyclic nature of the data warehouse demands administrative ease and flexibility. The RDBMS must provide controls for implementing resource limits, chargeback accounting to allocate costs back to users, and query prioritization to address the needs of different user classes and activities. The RDBMS must also provide for workload tracking and tuning so system resources may be optimized for maximum performance and throughput. "The most visible and measurable value of implementing a data warehouse is evidenced in the uninhibited, creative access to data it provides the end user.
* Integrated Dimensional Analysis - The power of multidimensional views is widely accepted, and dimensional support must be inherent in the warehouse RDBMS to provide the highest performance for relational OLAP tools. The RDBMS must support fast, easy creation of precomputed summaries common in large data warehouses. It also should provide the maintenance tools to automate the creation of these precomputed aggregates. Dynamic calculation of aggregates should be consistent with the interactive performance needs.
* Advanced Query Functionality - End users require advanced analytic calculations, sequential and comparative analysis, and consistent access to detailed and summarized data. Using SQL in a client/server point-and-click tool environment may sometimes be impractical or even impossible. The RDBMS must provide a complete set of analytic operations including core sequential and statistical operations.

### 1.5 Data mining problems/issues

Data mining systems rely on databases to supply the raw data for input and this raises problems in that databases tend be dynamic, incomplete, noisy, and large. Other problems arise as a result of the adequacy and relevance of the information stored.

#### 1.5.1 Limited Information

A database is often designed for purposes different from data mining and sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Inconclusive data causes problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain. For example cannot diagnose malaria from a patient database if that database does not contain the patients red blood cell count.

#### 1.5.2 Noise and missing values

Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes which rely on subjective or measurement judgements can give rise to errors such that some examples may even be mis-classified. Error in either the values of attributes or class information are known as noise. Obviously where possible it is desirable to eliminate noise from the classification information as this affects the overall accuracy of the generated rules.

Missing data can be treated by discovery systems in a number of ways such as;

* simply disregard missing values
* omit the corresponding records
* infer missing values from known values
* treat missing data as a special value to be included additionally in the attribute domain
* or average over the missing values using Bayesian techniques.

Noisy data in the sense of being imprecise is characteristic of all data collection and typically fit a regular statistical distribution such as Gaussian while wrong values are data entry errors. Statistical methods can treat problems of noisy data, and separate different types of noise.

#### 1.5.3 Uncertainty

Uncertainty refers to the severity of the error and the degree of noise in the data. Data precision is an important consideration in a discovery system.

#### 1.5.4 Size, updates, and irrelevant fields

Databases tend to be large and dynamic in that their contents are ever-changing as information is added, modified or removed. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time-sensitive as some data values vary over time and the discovery system is affected by the `timeliness' of the data.

Another issue is the relevance or irrelevance of the fields in the database to the current focus of discovery for example post codes are fundamental to any studies trying to establish a geographical connection to an item of interest such as the sales of a product.

### 1.6 Potential Applications

Data mining has many and varied fields of application some of which are listed below.

#### 1.6.1 Retail/Marketing

* Identify buying patterns from customers
* Find associations among customer demographic characteristics
* Predict response to mailing campaigns
* Market basket analysis

#### 1.6.2 Banking

* Detect patterns of fraudulent credit card use
* Identify `loyal' customers
* Predict customers likely to change their credit card affiliation
* Determine credit card spending by customer groups
* Find hidden correlations between different financial indicators
* Identify stock trading rules from historical market data

#### 1.6.3 Insurance and Health Care

* Claims analysis - i.e which medical procedures are claimed together
* Predict which customers will buy new policies
* Identify behaviour patterns of risky customers
* Identify fraudulent behaviour

#### 1.6.4 Transportation

* Determine the distribution schedules among outlets
* Analyse loading patterns

#### 1.6.5 Medicine

* Characterise patient behaviour to predict office visits
* Identify successful medical therapies for different illnesses

## 2 Data Mining Functions

Data mining methods may be classified by the function they perform or according to the class of application they can be used in. Some of the main techniques used in data mining are described in this section.

### 2.1 Classification

Data mine tools have to infer a model from the database, and in the case of supervised learning this requires the user to define one or more classes. The database contains one or more attributes that denote the class of a tuple and these are known as predicted attributes whereas the remaining attributes are called predicting attributes. A combination of values for the predicted attributes defines a class.

When learning classification rules the system has to find the rules that predict the class from the predicting attributes so firstly the user has to define conditions for each class, the data mine system then constructs descriptions for the classes. Basically the system should given a case or tuple with certain known attribute values be able to predict what class this case belongs to.

Once classes are defined the system should infer rules that govern the classification therefore the system should be able to find the description of each class. The descriptions should only refer to the predicting attributes of the training set so that the positive examples should satisfy the description and none of the negative. A rule said to be correct if its description covers all the positive examples and none of the negative examples of a class.

A rule is generally presented as, if the left hand side (LHS) then the right hand side (RHS), so that in all instances where LHS is true then RHS is also true, are very probable. The categories of rules are:

* exact rule - permits no exceptions so each object of LHS must be an element of RHS
* strong rule - allows some exceptions, but the exceptions have a given limit
* probablistic rule - relates the conditional probability P(RHS|LHS) to the probability P(RHS)

Other types of rules are classification rules where LHS is a sufficient condition to classify objects as belonging to the concept referred to in the RHS.

### 2.2 Associations

Given a collection of items and a set of records, each of which contain some number of items from the given collection, an association function is an operation against this set of records which return affinities or patterns that exist among the collection of items. These patterns can be expressed by rules such as "72% of all the records that contain items A, B and C also contain items D and E." The specific percentage of occurrences (in this case 72) is called the confidence factor of the rule. Also, in this rule, A,B and C are said to be on an opposite side of the rule to D and E. Associations can involve any number of items on either side of the rule.

A typical application, identified by IBM, that can be built using an association function is Market Basket Analysis. This is where a retailer run an association operator over the point of sales transaction log, which contains among other information, transaction identifiers and product identifiers. The set of products identifiers listed under the same transaction identifier constitutes a record. The output of the association function is, in this case, a list of product affinities. Thus, by invoking an association function, the market basket analysis application can determine affinities such as "20% of the time that a specific brand toaster is sold, customers also buy a set of kitchen gloves and matching cover sets."

Another example of the use of associations is the analysis of the claim forms submitted by patients to a medical insurance company. Every claim form contains a set of medical procedures that were performed on a given patient during one visit. By defining the set of items to be the collection of all medical procedures that can be performed on a patient and the records to correspond to each claim form, the application can find, using the association function, relationships among medical procedures that are often performed together.

### 2.3 Sequential/Temporal patterns

Sequential/temporal pattern functions analyse a collection of records over a period of time for example to identify trends. Where the identity of a customer who made a purchase is known an analysis can be made of the collection of related records of the same structure (i.e. consisting of a number of items drawn from a given collection of items). The records are related by the identity of the customer who did the repeated purchases. Such a situation is typical of a direct mail application where for example a catalogue merchant has the information, for each customer, of the sets of products that the customer buys in every purchase order. A sequential pattern function will analyse such collections of related records and will detect frequently occurring patterns of products bought over time. A sequential pattern operator could also be used to discover for example the set of purchases that frequently precedes the purchase of a microwave oven.

Sequential pattern mining functions are quite powerful and can be used to detect the set of customers associated with some frequent buying patterns. Use of these functions on for example a set of insurance claims can lead to the identification of frequently occurring sequences of medical procedures applied to patients which can help identify good medical practices as well as to potentially detect some medical insurance fraud.

### 2.4 Clustering/Segmentation

Clustering and segmentation are the processes of creating a partition so that all the members of each set of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and/or mutually exclusive set of clusters.

Clustering according to similarity is a very powerful technique, the key to it being to translate some intuitive measure of similarity into a quantitative measure. When learning is unsupervised then the system has to discover its own classes i.e. the system clusters the data in the database. The system has to discover subsets of related objects in the training set and then it has to find descriptions that describe each of these subsets.

There are a number of approachs for forming clusters. One approach is to form rules which dictate membership in the same group based on the level of similarity between members. Another approach is to build set functions that measure some property of partitions as functions of some parameter of the partition.

#### 2.4.1 IBM - Market Basket Analysis example

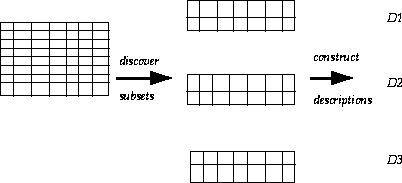
IBM have used segmentation techniques in their Market Basket Analysis on POS transactions where they separate a set of untagged input records into reasonable groups according to product revenue by market basket i.e. the market baskets were segmented based on the number and type of products in the individual baskets.

Each segment reports total revenue and number of baskets and using a neural network 275,000 transaction records were divided into 16 segments. The following types of analysis were also available, revenue by segment, baskets by segment, average revenue by segment etc.

## 3 Data Mining Techniques

### 3.1 Cluster Analysis

In an unsupervised learning environment the system has to discover its own classes and one way in which it does this is to cluster the data in the database as shown in the following diagram. The first step is to discover subsets of related objects and then find descriptions e.eg D1, D2, D3 etc. which describe each of these subsets.

Figure 5: Discovering clusters and descriptions in a database  


Clustering and segmentation basically partition the database so that each partition or group is similar according to some criteria or metric. Clustering according to similarity is a concept which appears in many disciplines. If a measure of similarity is available there are a number of techniques for forming clusters. Membership of groups can be based on the level of similarity between members and from this the rules of membership can be defined. Another approach is to build set functions that measure some property of partitions ie groups or subsets as functions of some parameter of the partition. This latter approach achieves what is known as optimal partitioning.

Many data mining applications make use of clustering according to similarity for example to segment a client/customer base. Clustering according to optimization of set functions is used in data analysis e.g. when setting insurance tariffs the customers can be segmented according to a number of parameters and the optimal tariff segmentation achieved.

Clustering/segmentation in databases are the processes of separating a data set into components that reflect a consistent pattern of behaviour. Once the patterns have been established they can then be used to "deconstruct" data into more understandable subsets and also they provide sub-groups of a population for further analysis or action which is important when dealing with very large databases. For example a database could be used for profile generation for target marketing where previous response to mailing campaigns can be used to generate a profile of people who responded and this can be used to predict response and filter mailing lists to achieve the best response.

### 3.2 Induction

A database is a store of information but more important is the information which can be inferred from it. There are two main inference techniques available ie deduction and induction.

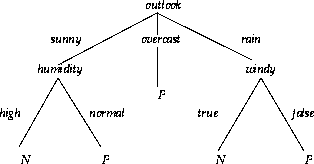
* Deduction is a technique to infer information that is a logical consequence of the information in the database e.g. the join operator applied to two relational tables where the first concerns employees and departments and the second departments and managers infers a relation between employee and managers.
* Induction has been described earlier as the technique to infer information that is generalised from the database as in the example mentioned above to infer that each employee has a manager. This is higher level information or knowledge in that it is a general statement about objects in the database. The database is searched for patterns or regularities.

Induction has been used in the following ways within data mining.

#### 3.2.1 decision trees

Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labelled with attribute names, the edges are labelled with possible values for this attribute and the leaves labelled with different classes. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object.

The following is an example of objects that describe the weather at a given time. The objects contain information on the outlook, humidity etc. Some objects are positive examples denote by P and others are negative i.e. N. Classification is in this case the construction of a tree structure, illustrated in the following diagram, which can be used to classify all the objects correctly.

Figure 6:   
  
Decision tree structure

#### 3.2.2 rule induction

A data mine system has to infer a model from the database that is it may define classes such that the database contains one or more attributes that denote the class of a tuple ie the predicted attributes while the remaining attributes are the predicting attributes. Class can then be defined by condition on the attributes. When the classes are defined the system should be able to infer the rules that govern classification, in other words the system should find the description of each class.

Production rules have been widely used to represent knowledge in expert systems and they have the advantage of being easily interpreted by human experts because of their modularity i.e. a single rule can be understood in isolation and doesn't need reference to other rules. The propositional like structure of such rules has been described earlier but can summed up as if-then rules.

### 3.3 Neural networks

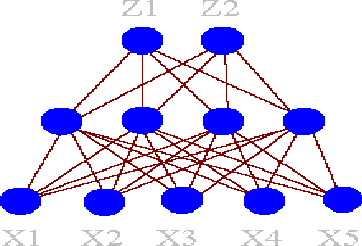
Neural networks are an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyse. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Neural networks have broad applicability to real world business problems and have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including:

* sales forecasting
* industrial process control
* customer research
* data validation
* risk management
* target marketing etc.

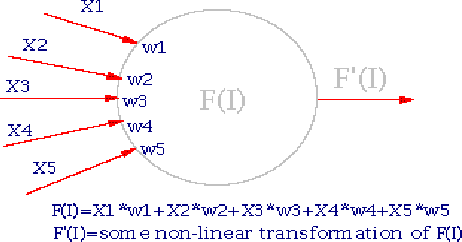
Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs, that simply follow instructions in a fixed sequential order.

The structure of a neural network looks something like the following:

Figure 7: Structure of a neural network  


The bottom layer represents the input layer, in this case with 5 inputs labels X1 through X5. In the middle is something called the hidden layer, with a variable number of nodes. It is the hidden layer that performs much of the work of the network. The output layer in this case has two nodes, Z1 and Z2 representing output values we are trying to determine from the inputs. For example, predict sales (output) based on past sales, price and season (input).

Each node in the hidden layer is fully connected to the inputs which means that what is learned in a hidden node is based on all the inputs taken together. Statisticians maintain that the network can pick up the interdependencies in the model. The following diagram provides some detail into what goes on inside a hidden node.

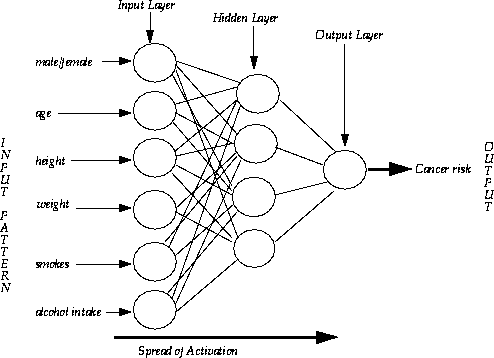
Figure 8: Inside a Node   


Simply speaking a weighted sum is performed: X1 times W1 plus X2 times W2 on through X5 and W5. This weighted sum is performed for each hidden node and each output node and is how interactions are represented in the network.

The issue of where the network get the weights from is important but suffice to say that the network learns to reduce error in it's prediction of events already known (ie, past history).

The problems of using neural networks have been summed by Arun Swami of Silicon Graphics Computer Systems. Neural networks have been used successfully for classification but suffer somewhat in that the resulting network is viewed as a black box and no explanation of the results is given. This lack of explanation inhibits confidence, acceptance and application of results. He also notes as a problem the fact that neural networks suffered from long learning times which become worse as the volume of data grows.

The Clementine User Guide has the following simple diagram to summarise a neural net trained to identify the risk of cancer from a number of factors.

Figure 9:   
  
Example Neural network from Clementine User Guide

### 3.4 On-line Analytical processing

A major issue in information processing is how to process larger and larger databases, containing increasingly complex data, without sacrificing response time. The client/server architecture gives organizations the opportunity to deploy specialized servers which are optimized for handling specific data management problems. Until recently, organizations have tried to target relational database management systems (RDBMSs) for the complete spectrum of database applications. It is however apparent that there are major categories of database applications which are not suitably serviced by relational database systems. Oracle, for example, has built a totally new Media Server for handling multimedia applications. Sybase uses an object-oriented DBMS (OODBMS) in its Gain Momentum product which is designed to handle complex data such as images and audio. Another category of applications is that of on-line analytical processing (OLAP). OLAP was a term coined by E F Codd (1993) and was defined by him as;

the dynamic synthesis, analysis and consolidation of large volumes of multidimensional data

Codd has developed rules or requirements for an OLAP system;

* multidimensional conceptual view
* transparency
* accessibility
* consistent reporting performance
* client/server architecture
* generic dimensionality
* dynamic sparse matrix handling
* multi-user support
* unrestricted cross dimensional operations
* intuitative data manipulation
* flexible reporting
* unlimited dimensions and aggregation levels

An alternative definition of OLAP has been supplied by Nigel Pendse who unlike Codd does not mix technology prescriptions with application requirements. Pendse defines OLAP as, Fast Analysis of Shared Multidimensional Information which means;

Fast in that users should get a response in seconds and so doesn't lose their chain of thought;

Analysis in that the system can provide analysis functions in an intuitative manner and that the functions should supply business logic and statistical analysis relevant to the users application;

Shared from the point of view of supporting multiple users concurrently;

Multidimensional as a main requirement so that the system supplies a multidimensional conceptual view of the data including support for multiple hierarchies;

Information is the data and the derived information required by the user application.

One question is what is multidimensional data and when does it become OLAP? It is essentially a way to build associations between dissimilar pieces of information using predefined business rules about the information you are using. Kirk Cruikshank of Arbor Software has identified three components to OLAP, in an issue of UNIX News on data warehousing;

* A multidimensional database must be able to express complex business calculations very easily. The data must be referenced and mathematics defined. In a relational system there is no relation between line items which makes it very difficult to express business mathematics.
* Intuitative navigation in order to `roam around' data which requires mining hierarchies.
* Instant response i.e. the need to give the user the information as quick as possible.

Dimensional databases are not without problem as they are not suited to storing all types of data such as lists for example customer addresses and purchase orders etc. Relational systems are also superior in security, backup and replication services as these tend not to be available at the same level in dimensional systems. The advantages of a dimensional system are the freedom they offer in that the user is free to explore the data and receive the type of report they want without being restricted to a set format.

#### 3.4.1 OLAP Example

An example OLAP database may be comprised of sales data which has been aggregated by region, product type, and sales channel. A typical OLAP query might access a multi-gigabyte/multi-year sales database in order to find all product sales in each region for each product type. After reviewing the results, an analyst might further refine the query to find sales volume for each sales channel within region/product classifications. As a last step the analyst might want to perform year-to-year or quarter-to-quarter comparisons for each sales channel. This whole process must be carried out on-line with rapid response time so that the analysis process is undisturbed. OLAP queries can be characterized as on-line transactions which:

* Access very large amounts of data, e.g. several years of sales data.
* Analyse the relationships between many types of business elements e.g. sales, products, regions, channels.
* Involve aggregated data e.g. sales volumes, budgeted dollars and dollars spent.
* Compare aggregated data over hierarchical time periods e.g. monthly, quarterly, yearly.
* Present data in different perspectives e.g. sales by region vs. sales by channels by product within each region.
* Involve complex calculations between data elements e.g. expected profit as calculated as a function of sales revenue for each type of sales channel in a particular region.
* Are able to respond quickly to user requests so that users can pursue an analytical thought process without being stymied by the system.

#### 3.4.2 Comparison of OLAP and OLTP

OLAP applications are quite different from On-line Transaction Processing (OLTP) applications which consist of a large number of relatively simple transactions. The transactions usually retrieve and update a small number of records that are contained in several distinct tables. The relationships between the tables are generally simple.

A typical customer order entry OLTP transaction might retrieve all of the data relating to a specific customer and then insert a new order for the customer. Information is selected from the customer, customer order, and detail line tables. Each row in each table contains a customer identification number which is used to relate the rows from the different tables. The relationships between the records are simple and only a few records are actually retrieved or updated by a single transaction.

The difference between OLAP and OLTP has been summarised as, OLTP servers handle mission-critical production data accessed through simple queries; while OLAP servers handle management-critical data accessed through an iterative analytical investigation. Both OLAP and OLTP, have specialized requirements and therefore require special optimized servers for the two types of processing.

OLAP database servers use multidimensional structures to store data and relationships between data. Multidimensional structures can be best visualized as cubes of data, and cubes within cubes of data. Each side of the cube is considered a dimension.

Each dimension represents a different category such as product type, region, sales channel, and time. Each cell within the multidimensional structure contains aggregated data relating elements along each of the dimensions. For example, a single cell may contain the total sales for a given product in a region for a specific sales channel in a single month. Multidimensional databases are a compact and easy to understand vehicle for visualizing and manipulating data elements that have many inter relationships.

OLAP database servers support common analytical operations including: consolidation, drill-down, and "slicing and dicing".

* Consolidation - involves the aggregation of data such as simple roll-ups or complex expressions involving inter-related data. For example, sales offices can be rolled-up to districts and districts rolled-up to regions.
* Drill-Down - OLAP data servers can also go in the reverse direction and automatically display detail data which comprises consolidated data. This is called drill-downs. Consolidation and drill-down are an inherent property of OLAP servers.
* "Slicing and Dicing" - Slicing and dicing refers to the ability to look at the database from different viewpoints. One slice of the sales database might show all sales of product type within regions. Another slice might show all sales by sales channel within each product type. Slicing and dicing is often performed along a time axis in order to analyse trends and find patterns.

OLAP servers have the means for storing multidimensional data in a compressed form. This is accomplished by dynamically selecting physical storage arrangements and compression techniques that maximize space utilization. Dense data (i.e., data exists for a high percentage of dimension cells) are stored separately from sparse data (i.e., a significant percentage of cells are empty). For example, a given sales channel may only sell a few products, so the cells that relate sales channels to products will be mostly empty and therefore sparse. By optimizing space utilization, OLAP servers can minimize physical storage requirements, thus making it possible to analyse exceptionally large amounts of data. It also makes it possible to load more data into computer memory which helps to significantly improve performance by minimizing physical disk I/O.

In conclusion OLAP servers logically organize data in multiple dimensions which allows users to quickly and easily analyse complex data relationships. The database itself is physically organized in such a way that related data can be rapidly retrieved across multiple dimensions. OLAP servers are very efficient when storing and processing multidimensional data. RDBMSs have been developed and optimized to handle OLTP applications. Relational database designs concentrate on reliability and transaction processing speed, instead of decision support need. The different types of server can therefore benefit a broad range of data management applications.

### 3.5 Data Visualisation

Data visualisation makes it possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well along side data mining. Data mining allows the analyst to focus on certain patterns and trends and explore in-depth using visualisation. On its own data visualisation can be overwhelmed by the volume of data in a database but in conjunction with data mining can help with exploration.

## 4 Software - past and present developments

This section outlines the historic background or the evolution of database systems in terms of parallel processing and data mining with reference to the part played by some of the main vendors and their successes.

### 4.1 New architectures

The best of the best commercial database packages are now available for massively parallel processors including IBM DB2, INFORMIX-OnLine XPS, ORACLE7 RDBMS and SYBASE System 10. This evolution, however, has not been an easy road for the pioneers.

HPCwire by Michael Erbschloe, contributing editor Oct. 6, 1995

The evolution described by Michael Erbschloe is detailed and expanded on in the following sections.

#### 4.1.1 Obstacles

What were the problems at the start?

* the typical scientific user knew nothing of commercial business applications and gave little attention or credence to the adaptation of high performance computers to business environments.
* the business database programmers, who, although well versed in database management and applications, knew nothing of massively parallel principles.

The solution was for database software producers to create easy-to-use tools and form strategic relationships with hardware manufacturers and consulting firms.

#### 4.1.2 The key

The key is the retooling database software to maspar environments. Parallel processors can easily assign small, independent transactions to different processors. With more processors, more transactions can be executed without reducing throughput. This same concept applies to executing multiple independent SQL statements. A set of SQL statements can be broken up and allocated to different processors to increase speed.

Multiple data streams allow several operations to proceed simultaneously. A customer table, for example, can be spread across multiple disks, and independent threads can search each subset of the customer data. As data is partitioned into multiple subsets performance is increased. I/O subsystems then just feed data from the disks to the appropriate threads or streams.

An essential part of designing a database for parallel processing is the partitioning scheme. Because large databases are indexed, independent indexes must also be partitioned to maximize performance. There are five partitioning methods used to accomplish this:

1. Hashing, where data is assigned to disks based on a hash key

2. Round-robin partitioning, which assigns a row to partitions in sequence.

3. Allocating rows to nodes based on ranges of values.

4. Schema partitioning (Sybase Navigation Server), which lets you tie tables to specific partitions.

5. User-defined roles (Informix).

#### 4.1.3 Oracle was first

Oracle was the first to market parallel database packages with their flagship product, ORACLE7 RDBMS having been installed at over 100 user sites. Oracle began beta support for the IBM SP platform in July 1994.

Ease of use is an important factor in the success of any commercial application and by design the Oracle Parallel Server hides the complexities of data layout from the users. Users who wish to add disks or processor nodes can do so without complex data reorganization and application re-partitioning. In addition, Oracle Parallel Server software uses the same SQL interface as the Oracle7 database. Since no new commands or extensions to existing commands are needed, previously developed tools and applications will run unchanged.

The Oracle Parallel Server technology performs both the parallelization and optimization automatically, eliminating the need to re-educate application developers and end users. It is also easy for user organizations to deploy because it eliminates many traditional implementation burdens.

Reference - http://www.oracle.com.

#### 4.1.4 Red Brick has a strong showing

Red Brick Systems, based in Los Gatos, Calif., specializes in software products used for fast and accurate business decisions where large client/server databases, usually tens to hundreds of gigabytes in size with hundreds of millions of records, are the norm. These applications require historical context, but timely analysis of complex data relationships for both consolidated and detailed business information.

Red Brick Warehouse VPT, (Very large data warehouse support, Parallel query processing, Time based data management), is a DBMS tuned for data warehouse applications. It employs specialized indexing techniques which are designed to facilitate data warehousing. The join accelerator STARjoin uses a special index to multiple tables that participate in a join. With its parallel capability it can run applications that can handle up to 500 GB or more of data It is a parallel database product that significantly improves the organization, availability, administration, and performance of data warehouse applications.

Unlike RDBMS products optimized for on-line transaction processing, Red Brick Warehouse VPT allows business management applications to be developed and deployed quickly;

* to query very large databases of information gathered from disparate sources;
* to provide the best access to both consolidated and detailed business information;
* and to simply run fast.

Red Brick's server-based relational engine is accessible by several popular front-end client application environments which support Microsoft ODBC, Sybase Open Client, and Information Builders, Inc. EDA/SQL interfaces.

Reference - http://www.redbrick.com.

#### 4.1.5 IBM is still the largest

IBM is the world's largest producer of database management software. Eighty percent of the FORTUNE 500, including the top 100 companies, rely on DB2 database solutions to manage data on mainframes, minicomputers, RISC workstations and personal computers. The availability of the new DB2 Parallel Edition, extends the functionality and reliability of the DB2 to IBM's high-performance parallel systems SP2. With DB2 Parallel Edition running on the SP2, users can access very large databases, process huge amount of data, and perform complex queries in minutes.

DB2 Parallel Edition is packaged with the SP2 running AIX and a set of services to help users speed their transactions and quickly and easily derive the benefits of parallel computing. The turnkey solution, called POWERquery, provides a relatively cost-effective, large-scale decision support.

DB2 Parallel Edition is a member of the IBM DB2 family of databases, therefore users do not have to rewrite any applications or retrain their staffs. To a user, the database appears to be a single database server, only faster. It is faster because all functions are performed in parallel, including data and index scans, index creation, backup and restore, joins, inserts, updates and deletes.

Reference - http://www.ibm.com.

#### 4.1.6 INFORMIX is online with 8.0

Informix has been supporting SMP with Informix Parallel Data Query (PDQ) as part of its Dynamic Scalable Architecture (DSA) and through DSA/XMP by extending PDQ functions to work in loosely coupled parallel environments, including clusters. Online 8.0 is the latest high-performance, scalable database server based on Informix's industry-leading DSA. OnLine XPS extends DSA to loosely coupled, shared-nothing computing architectures including clusters of symmetric multiprocessing (SMP) systems and (MPP) systems.

One key to Informix's success on SMP is a joint development agreement with Sequent Computer Systems (Beaverton, Ore. that resulted in a rebuild of the core of Informix OnLine to a multithreaded system with small-grained, lightweight threads. Virtual processors are pooled and the DBMS allocates them dynamically to CPUs, based on processing requirements. OnLine XPS' high availability, systems management based on the Tivoli Management Environment (TME), data partitioning, enhanced parallel SQL operations, and other features are designed to simplify and economize VLDB applications. OnLine XPS also offers a significant improvement in performance for mission-critical, data-intensive tasks associated with data warehousing, decision support, imaging, document management and workflow, and other VLDB operational environments.

Although Informix databases, such as OnLine XPS and INFORMIX-OnLine Dynamic Server, are at the heart of data warehousing solutions, other products and services must integrate with the databases to ensure a successful data warehouse implementation, a critical component of a data warehouse architecture is online analytical processing (OLAP).

Informix delivers relational multidimensional capabilities through strategic partnerships with Information Advantage, MicroStrategy, and Stanford Technology Group. Informix also has proven partnerships with technology providers, such as Business Objects, Coopers & Lybrand, Evolutionary Technologies, KPMG, Price Waterhouse, Prism, and SHL Systemhouse, to provide capabilities such as data modelling, data extraction, data access, multidimensional analysis, and systems integration.

Reference - http://www.informix.com.

#### 4.1.7 Sybase and System 10

Sybase has improved multithreading with System 10 which has been designed to handle interquery and transaction parallelizing on SMP computers with very large, heavyweight threads. Up to 64 processors can be utilized as SQL servers configured into a single system image. This was accomplished in part by the use of the Sybase Navigation server which takes advantage of parallel computers. Parallelism is achieved by an SQL Server on a processor and control servers, which manages parallel operations. Sybase's Navigation Server partitions data by hashing, ranges, or schema partitioning. Reports indicate that the partitioning scheme and keys chosen impact parallel performance.

Sybase IQ was delivered to 24 beta customers in July, providing predictable interactive access to large amounts of data directly in the warehouse. While offering up to 100-fold query performance improvement over standard relational databases, Sybase IQ slashes warehouse query costs by orders of magnitude, requiring up to 80 percent less disk, up to 98 percent less I/O, and utilizing existing hardware, according to Sybase.

An optional extension for the SYBASE SQL Server, SYBASE IQ includes patent-pending Bit-Wise indexing that allows significantly more data to be processed in each instruction, resulting in up to thousands of times faster performance without adding hardware. Beyond simple bit maps, Bit-Wise indexing makes it possible to index every field in the database --including character and numeric fields not supported by other bit-map indexing schemes -- in less than the size of the raw data, substantially reducing disk costs. SYBASE IQ indexes provide a complete map of the data, eliminating table scans and directly accessing just the information required, reducing I/O by up to 98 percent and resulting in fast, predictable answers to any query.

Reference - http://www.sybase.com.

#### 4.1.8 Information Harvester

Information Harvester software on the Convex Exemplar offers market researchers in retail, insurance, financial and telecommunications firms the ability to analyse large data sets in a short time.

The flexibility of the Information Harvesting induction algorithm enables it to adapt to any system. The data can be in the form of numbers, dates, codes, categories, text or any combination thereof. Information Harvester is designed to handle faulty, missing and noisy data. Large variations in the values of an individual field do not hamper the analysis. Information Harvester claims unique abilities to recognize and ignore irrelevant data fields when searching for patterns. In full-scale parallel-processing versions, Information Harvester can handle millions of rows and thousands of variables.

The Exemplar series, based on HP's high performance PA-RISC processors, is the first supercomputer-class family of systems to track the price/performance development cycle of the desktop. They are being used for a range of applications including automotive, tire and aircraft design, petroleum research and exploration, seismic processing, and university, scientific and biomedical research.

Reference - http://www.convex.com

### 4.2 Vendors and Applications

This section examines some of the major vendors of siftware with supporting case studies.

#### 4.2.1 Information Harvesting Inc

The problem of deriving meaningful information from enormous amounts of complex data is being handled by the data mining software produced by Information Harvesting Inc. (IH), founded in 1994 and based in Cambridge, Mass. It makes use of conventional statistical analysis techniques by building upon a proprietary tree-based learning algorithm similar to CART, ID3 and Chaid that generates expert-system-like rules from datasets, initially presented in forms such as numbers, dates, categories, codes, or any combination.

The proprietary Information Harvesting algorithm operates by creating a set of bins for each field in the data, with groups of values within a field ultimately determining the rules. According to the distribution of values the algorithm delineates bin boundaries via fuzzy logic to determine where a given value falls within a bin and thus how the values may be grouped.

A binary tree then generates rules from the data. At the uppermost node the algorithm analyses all data rows, and at each lower level subsets created by the node above are analysed. Each node arrives at a set of rules categorizing the data reviewed at that level. Each rule may include multiple variables (combined with ANDs) or multiple clauses (combined with ORs) and derives from the way variables fall into various bins. A prediction can be based on one or more rules.

Rule quality, the amount of error for each rule, and importance, how often each rule is used for making predictions, are also assessed by the software. This avoids the effect of simply memorizing historical data or misunderstanding the relevance of a given rule. Design rows are used to extract the rules per se, but test rows are utilized to determine the rules level of accuracy.

In addition, the program is set to optimize results by running over the same datasets again and again while adjusting the internal parameters for the best result. Optimization can be achieved with either a rapid hill-climbing algorithm or completely with a modified genetic algorithm.

The data mining modules are written in ANSI C and thus can be ported to a wide range of platforms: on client/server architecture (where the application uses TCP/IP), parallel processing machines, or mainframe supercomputers.

Two examples of companies using the software are:

Healthcare - Michael Reese Medical Associates (MRMA) employed data mining software from Information Harvesting and Vantage Point as a tool for gaining advantage in contract negotiations. The 28-doctor group had to predict trends in type, price, location, and use of service, since they must negotiate with insurance companies to provide certain services at a set monthly fee, doctors must accurately predict their per member/per month cost to break even or make a profit. Normally physicians could only make an intuitive estimate roughly based on after-the-fact evaluations of prior estimates when determining this critical figure whereas data mining offered a new approach.

Finance - The Philadelphia Police and Fire Federal Credit Union (PFFCU) used data mining to maximize their membership base by cultivating multiple relationships (e.g. consumer loans, annuities, credit cards, etc.) with members. Because the membership base is extremely homogeneous (police and fire dept. employees and their families), data had to be deeply drilled to identify segmented groups. Used in conjunction with software such as InterGlobal Financial Systems' Credit Analyzer, Information Harvester identified members most and least profitable to the organization as well as those who would make attractive loan candidates. Data mining often led PFFCU to accurate but counter-intuitive results. For example, members who had filed for bankruptcy were more inclined to clear debts with the Credit Union than outside lenders. Thus, PFFCU identified members with imperfect credit histories but a strong tendency to pay, whereas these individuals would be ignored by large conventional lenders.

#### 4.2.2 Red Brick

Red Brick have a number of cases to present in support of the use of their data mining technology, two of which are H.E.B. of San Antonio, and Hewlett-Packard.

H.E.B.- Category management in retailing

H.E.B. of San Antonio, Texas (sales of approx. $4.5 billion, 225 stores, 50,000 employees) was able to bring a category management application from design to roll out in under nine months because it kept the requirements simple and had database support from Red Brick and server support from Hewlett-Packard Company.

Previously, the marketing information department would take ad hoc requests for information from users, write a program to extract the information, and return the information to the user a week or so later - not timely enough for most business decisions and in some cases not what the user really wanted in the first place.

The organizational change to category management was implemented in 1990. The category manager is characterized as the "CEO" of the category with profit and loss responsibilities, final decision over which products to buy and which to delete, and where the products are to be located on the shelves. The category manager also decides which stores get which products. Although H.E.B. stores are only within the state of Texas, it is a diverse market where some stores near Mexico are 98% Hispanic while suburban Dallas stores may be only 2% Hispanic. The change to category management centralized all merchandising and marketing decisions, removing these decisions from the stores.

As category managers built up their negotiating skills, technical skills, and partnering skills over three years, the need for more timely decision-support information grew. An enterprise-wide survey of users to determine requirements took until September 1993. The company then benchmarked three database management systems - Red Brick, Teradata and Time Machine - and picked Red Brick. The group leased the hardware, a Hewlett-Packard 9000 model T500 (2-processor, with 768M of RAM, and 100GB of disk space--the system now has 200 GB). For a user interface, the company contracted for a custom graphical front-end based on Windows. Also, a COBOL programmer was used to write data extraction programs to take P.O.S. data from the mainframe, format the data properly, and transfer the data to the Red Brick database.

The model was delivered in March 1994 and the application has been up and running without problems since then. The company maintains two years of data by week, by item (257,000 UPCs), by store. This is about 400 million detail records. Summary files are only maintained by time and total company, which can be an advantage.

The goal was to have all queries answered in 4 seconds, but some trends reports with large groups of items over long time periods take 30 - 40 seconds. The users are not always technically oriented, so the design intentionally aimed for simplicity. The system is ad hoc to the extent that the user can specify time, place, and product.

H.E.B. feels that category managers are now making fact-based decisions to determine which products to put in which stores, how much product to send to a store, and the proper product mix. Historically, buyers usually were promoted from the stores and had considerable product knowledge whereas now category managers are coming from other operational areas such as finance and human resources. This is possible because the system give people with limited product knowledge the equivalent of years experience.

Hewlett-Packard: "Discovering" Data To Manage Worldwide Support

Hewlett-Packard, a premier, global provider of hardware systems is known for manufacturing high quality products but to maintain its reputation they depended on delivering service and support through and after product delivery.

The Worldwide Customer Support Organization (WCSO) within Hewlett-Packard is responsible for providing support services to its hardware and software customers. For several years, WCSO has used a data warehouse of financial, account, product, and service contract information to support decision making. WCSO Information Management is responsible for developing and supporting this data warehouse.

Until 1994, WCSO Information Management supported business information queries with a data warehouse architecture based on two HP3000/Allbase systems and an IBM DB2 system. This was a first attempt at collecting, integrating, and storing data related to customer support for decision-making purposes. As they increasingly relied upon the data warehouse, they began to demand better performance, additional data coverage, and more timely data availability.

The warehouse architecture did not keep pace with the increased requirements from WCSO users. Users wanted to get information quickly. Both load and query performance were directly impacted as more data was added. It was to decided to investigate other warehouse alternatives with the aim of finding a new data warehouse that would significantly improve load/query performance, be more cost effective, and support large amounts of data without sacrificing performance. To help select the best combination of hardware and software for the new warehouse, benchmarks were conducted using Red Brick and two other RDBMS products. They did not look at Oracle or Sybase because they were promoting OLTP data functionality and weren't focused upon data warehousing.

Benchmarks included tests simulating some of HP's most demanding user queries, testing the load times for tables in the five to eight million row range. Tests also were conducted to verify that performance did not degrade as data was added into the warehouse. "The Red Brick product performed head and shoulders above the rest," recalls Ryan Uda, Program Manager for WCSO's Information Management Program. Benchmark results showed Red Brick loading data in one hour against ten hours for other systems. Red Brick's query performance was consistently five to ten times faster. Red Brick returned consistently superior performance results even when large amounts of data were added to the warehouse.

HP chose to use Red Brick software on an HP9000 and the project began with the consolidation of the existing three databases into a single data warehouse named "Discovery." This downsizing provided significant cost savings and increased resource efficiencies in managing and supporting the warehouse environment. Today, Discovery supports approximately 250 marketing, finance, and administration users in the Americas, Europe, and Asia-Pacific regions. They pull query results into their desktop report writers, load information into worksheets, or use the data to feed Executive Information Systems. User satisfaction has risen dramatically due to Discovery's vastly improved performance and remodelled business views.

#### 4.2.3 Oracle

For large scale data mining, Oracle on the SP2 offers customers robust functionality and excellent performance. Data spread across multiple SP2 processor nodes is treated as a single image affording exceptionally fast access to very large databases. Oracle Parallel Query allows multiple users to submit complex queries at the same time. Individual complex queries can be broken down and processed across several processing nodes simultaneously. Execution time can be reduced from overnight to hours or minutes, enabling organizations to make better business decisions faster.

Oracle offers products that help customers create, administer and use their data warehouse. Oracle has a large suite of connectivity products that provide transparent access to many popular mainframe databases. Through the use of these products, customers can move data from legacy mainframe applications into the data warehouse on the SP2.

Some of the examples of their technology at work are as follows:

John Alden Insurance based in Miami, Fla., is using Oracle Parallel Query on the SP2 to mine healthcare information and they have seen orders-of-magnitude improvements in response time for typical business queries.

ShopKo Stores, a $2 billion, Wisconsin-based mass merchandise chain which operates 128 stores throughout the Midwest and Northeast, chose the SP2 to meet their current and projected needs for both data mining and mission-critical merchandising applications.

Pacific Bell and U.S. West, both telecommunications providers, have are using the Oracle Warehouse to improve their ability to track customers and identify new service needs. The solutions are based on the Oracle Warehouse, introduced in June, 1995.

* Pacific Bell's data warehouse provides a common set of summarized and compressed information to base decision support systems. The first system is designed to analyse product profitability, and similar decision support systems are in development for marketing, capital investment and procurement, and two additional financial systems.
* U.S. West has implemented a warehousing system to analyse intra-area code calling data from its three operating companies. Running Oracle7 Release 7.2 on a 9-CPU symmetric multiprocessing system from Pyramid, US West's initial centralized architecture supports use by 20 executives and marketing specialists. The next phase will deliver warehouse access to more than 400 service representatives, which will ultimately be expanded up to 4,500 service representatives.

#### 4.2.4 Informix - Data Warehousing

As a major player in the field of data mining Informix have a number of success stories to quote some of which are:

Informix and Associated Grocers (retail example)

Associated Grocers, one of the leading cooperative grocery wholesalers in the northwest United States, with revenues of $1.2 billion, is replacing its traditional mainframe environment with a three-tiered client/server architecture based on Informix database technology. The new system's advanced applications have cut order-fulfilment times in half, reduced inventory carrying costs, and enabled the company to offer its 350 independent grocers greater selection at a lower cost. The details are -

* Hardware: Hewlett-Packard, IBM, AT&T GIS
* Partners: Micro Focus and Lawson Associates
* Applications: Inventory management, post billing, radio frequency, POS scanning, and data warehousing
* Key Informix Products: INFORMIX-OnLine Dynamic Server

In 1991, Associated Grocers embarked on a phased transition from its mainframe-based information system to open systems. The company initially used IBM RS/6000 hardware, and has since included Hewlett-Packard and NCR. In evaluating relational database management systems, Associated Grocers developed a checklist of requirements including education/training, scalability, technical support, solid customer references, and future product direction.

After selecting Informix as its company wide database standard, Associated Grocers then assembled the rest of its system architecture using a three tier model. On tier one, the "client" presentation layer, graphical user interfaces are developed using Microsoft(R) Windows(TM) and Visual Basic(TM). Tier two, based on Hewlett-Packard hardware, runs Micro Focus COBOL applications on top of the OEC Developer Package from Open Environment Corporation. This helps Associated Grocers develop DCE-compliant applications. The third layer, the data layer, is the INFORMIX-OnLine database.

Associated Grocers' pilot Informix-based application provides real-time inventory information for its deli warehouse. In the past, merchandise was received manually, and pertinent product information was later keyed into Associated Grocers' financial system. In contrast, the new system recognizes merchandise right at the receiving dock. Hand-held radio frequency devices allow merchandise to be immediately scanned into the Informix database. Product is assigned to a warehouse location and its expiration date is noted. When orders are filled, products with the earliest expiration dates are shipped first.

An extension to the deli warehouse system is a new post billing system, which is the ability to separate physical and financial inventory. Previously, merchandise could not be released for sale until the financial systems had been updated, which typically occurred over night. The new Informix-based system allows for immediate sale and distribution of recently received merchandise.

A third Informix-based application enables Associated Grocers to economically sell unique items-slow moving merchandise which is ordered monthly versus daily. Rather than incurring the high cost to warehouse these items, Associated Grocers created a direct link to outside speciality warehouses to supply the needed items on demand. Independent stores simply order the merchandise from Associated Grocers. The order goes into Associated Grocers' billing system then gets transmitted to the speciality warehouse, which immediately ships the merchandise to Associated Grocers. The speciality items are loaded onto Associated Grocers' delivery trucks and delivered along with the rest of an independent store's order.

Host Marriott (retail example)

Host Marriott has revenues of $1.1 billion and is a leading provider of food, beverage, and merchandise concession outlets located at airports, travel plazas, and toll roads throughout the United States. The company is streamlining its information systems to develop better cost controls and more effectively manage operations. To accomplish this, Host Marriott selected Informix database technology as its strategic IS foundation, which includes the development of a data warehouse using INFORMIX-OnLine Dynamic Server(TM) and INFORMIX-NewEra(TM). The new system will deliver valuable information throughout the organization, from field operators to corporate analysts. Details of the solution are:

* Hardware: IBM, Hewlett-Packard
* Applications: Sales and marketing, inventory management, labor productivity, and data warehousing
* Informix Products: INFORMIX-OnLine Dynamic Server, INFORMIX-NewEra, INFORMIX-ESQL/C

The company split into two separate companies; Host Marriott and Marriott International, and as the company grew more diverse, so did its computer systems. Unique and more advanced information systems were coupled with inadequate ones. As a result, financial consolidation was primarily done manually, with sales information from each outlet keyed into individual computer systems every night. The information was then sent to Host Marriotts corporate office, where it was posted to the mainframe accounting system, which had no analysis capabilities. Any analysis had to be completed via a second system, proving to be a labor-intensive and slow process.

In an effort to streamline operations and improve system flexibility, Host Marriott is replacing its manually-intensive system with a series of new client/server-based applications using Informix development tools and relational database products running on an IBM RS/6000 and Hewlett-Packard Vectra PCs.

The first of Host Marriotts new Informix-based applications automates its sales and marketing functions. It was developed using INFORMIX-HyperScript(R) Tools--a visual programming environment used to create client/server applications for Windows(TM), UNIX(R), and Macintosh(R) systems, and INFORMIX-ESQL/C--a database application development tool which is used to embed SQL statements directly into C code. Instead of waiting for individual end-of-day reports, the system automatically polls sales data from the point-of-sale terminals at each outlet and consolidates it in the INFORMIX-SE relational database.

This information is used to consolidate and speed up end-of-day reporting, analyse sales, and monitor regulatory compliance. It has reduced a 10 hour process to less than one hour, and enables corporate and concession management to perform the kind of in-depth analysis that allows them to fine tune their product mix, reduce administrative overhead, and ultimately increase profit margins.

Focus is now on a data warehouse to leverage its existing businesses and generate new growth opportunities in the future. The data warehouse is a separate database that Host Marriott is designing explicitly for its data-intensive, decision-support applications. Building a data warehouse will allow them to optimize query times and eliminate impact on the company's production systems. The warehouse is being developed with INFORMIX-NewEra, an open, graphical, object-oriented development environment especially suited for creating enterprise wide client/server database applications.

The foundation of Host Marriotts data warehouse will be INFORMIX-OnLine Dynamic Server, which takes advantage of multiprocessing hardware to perform multiple database functions in parallel. The data warehouse will help the company determine which brands will succeed in which market. It will also help Host Marriott develop more proprietary brands, and deliver better products and services at lower cost.

By pooling sales data, market research, customer satisfaction ratings, etc., Host Marriott will be able to perform detailed analysis in order to eliminate unnecessary costs from operations, and fully leverage new business opportunities. Relying on Informix products and services is enabling Host Marriott to make the important shift from simple data processing to strategic business analysis.

#### 4.2.5 Sybase

There is a lot of interest and activity in data warehousing, recent surveys show that more than 70 percent of Fortune 1000 companies have Data Warehousing projects budgeted or underway at an average cost of $3 million and a typical development time of 6 to 18 months (Meta Group Inc.).

Conventional warehousing applications today extract basic business data from operational systems, edit or transform it in some fashion to ensure its accuracy and clarity, and move it by means of transformation products, custom programming, or "sneaker net" to the newly deployed analytical database system. This extract, edit, load, query, extract, edit, load, query system might be acceptable if business life were very simple and relatively static but that is not the case, new data and data structures are added, changes are made to existing data, and even whole new databases are added.

Sybase Warehouse WORKS

Sybase Warehouse WORKS was designed around four key functions in data warehousing:

* Assembling data from multiple sources
* Transforming data for a consistent and understandable view of the business
* Distributing data to where it is needed by business users
* Providing high-speed access to the data for those business users

The Sybase Warehouse WORKS Alliance Program provides a complete, open, and integrated solution for organizations building and deploying data warehouse solutions. The program addresses the entire range of technology requirements for data warehouse development, including data transformation, data distribution, and interactive data access. The alliance partners have made commitments to adopt the Warehouse WORKS architecture and APIs, as well as to work closely with Sybase in marketing and sales programs.

#### 4.2.6 SG Overview

The advances in data analysis realized through breakthroughs in data warehousing are now being extended by new solutions for data mining. Sophisticated tools for 3D visualization, coupled with data mining software developed by Silicon Graphics, make it possible to bring out patterns and trends in the data that may not have been realized using traditional SQL techniques. These "nuggets" of information can then be brought to the attention of the end user, yielding bottom-line results.

Using fly-through techniques, you can navigate your models on consumer purchasing and channel velocity to follow trends and observe patterns. In response to what you see, you can interact directly with the data, using visual computing to factor critical "what-if" scenarios into your models. By making it possible to go through many such iterations without resorting to over-burdened IS staff for analytical assistance, you can eliminate days - even months - from the review process.

#### 4.2.7 IBM Overview

IBM provides a number of decision support tools to give users a powerful but easy-to-use interface to the data warehouse. IBM Information Warehouse Solutions offer the choice of decision support tools that best meet the needs of the end users in keeping with their commitment to provide open systems implementations.

IBM has announced, a Customer Partnership Program, to work with selected customers to gain experience and validate the applicability of the data mining technology. This offers customers the advantage of IBM's powerful new data mining technology to analyse their data looking for key patterns and associations. Visa and IBM announced an agreement on 30 May 1995 signalling their intention to work together. This will change the way in which Visa and its member banks exchange information worldwide. The proposed structure will facilitate the timely delivery of information and critical decision support tools directly to member financial institutions' desktops worldwide.

IBM Visualizer provides a powerful and comprehensive set of ready to use building blocks and development tools that can support a wide range of end-user requirements for query, report writing, data analysis, chart/graph making, business planning and multimedia database. As a workstation based product, Visualizer is object-oriented and that makes it easy to plug-in additional functions such as those mentioned. And, Visualizer can access databases such as Oracle and Sybase as well as the DB2 family.

There are a number of other decision support products available from IBM based on the platform, operating environment and database with which you need to work. For example, the IBM Application System (AS) provides a client/server architecture and the widest range of decision support functions available for the MVS and VM environments. AS has become the decision support server of choice in these environments because of its capability to access many different data sources. IBM Query Management Facility (QMF) provides query, reporting and graphics functions in the MVS, VM, and CICS environments. The Data Interpretation System (DIS) is an object-oriented set of tools that enable end users to access, analyse and present information with little technical assistance. It is a LAN-based client/server architecture that enables access to IBM and non-IBM relational databases as well as host applications in the MVS and VM environment. These and other products are available from IBM to provide the functions and capabilities needed for a variety of implementation alternatives.