

1.0 Introduction

Data mining is the process of intelligently extracting hidden trends and information from corporate databases. The information is usually buried too deep to extract using a conventional analysis tool such as OLAP. The power of finding new information helps corporate decision makers to learn more about their customers by perform tasks such as market segmentation, customer profiling, trend forecasting, cross-selling and fraud detection.

OLAP (On-Line Analytical Processing) and data warehousing are two data mining related tools. Traditional query and report tools describe what is in a database. OLAP goes further by answering why certain things are true. The user forms a hypothesis about a relationship and verifies it with a series of queries against the data. Data mining is different from OLAP because rather than verify a hypothesis, it is used to generate a hypothesis.

Data warehousing is the process of cleaning up corporate data and put it in a consistent format. However, data warehousing is not a requirement for data mining. Data mining can be done directly from one or more operational or transactional databases by simply extracting it into a read-only database. [2]

The details of how data mining works as well as how it relates to data warehousing and OLAP can be found in the appendix.

2.0 Focus

This report focuses on the current leading data mining contenders in the market. It is both a summarization and a comparison of commercially available data mining products. Every business problem is different. A product that can solve one problem might be ineffective for another. Data mining is a relatively new market and can be quite confusing. The purpose of this study is to help customers who are in the market to buy data mining products make an informed choice.

To begin, bases of comparison must be established. Using the bases of comparison, companies and their products are summarized free of marketing jargon and then compared. These bases of comparison are things a customer should think about before purchasing. For convenient comparison, summaries tables of the products discussed in the report are included at the end. Each table contains the comparison of all products against each other based on one criteria.

3.0 Bases of Comparison

The spectrum of the data mining industry is very wide. At one end, there are expensive enterprise-level products incorporating multiple techniques capable of performing many tasks. At the other end, there are inexpensive desktop products incorporating only one technique with limited capabilities.

There are several bases of comparison. Some of them are technical, such as technology used, systems requirements and size of the database, ease of use, and ease of integration. Some of the above factors decide which market segment the product occupies. Other bases deal with the company and the market position, such as company profile, past experience, other lines of products, training and support. All of the above considerations are reflected in the price of each product.

3.1 Market Position and Prices

The data mining market can be divided into three segments, high-end, mid-end, and low-end. Each segment is defined by the cost of investment, the techniques used, the systems requirements, and the scale of the problems addressed. The scale of a problem refers to the complexity and amount of data to be analyzed.

3.11 High-end Products

High-end products, also known as enterprise-level data mining, are designed for complex, large-scale problems. These products are characterized as flexible, complete, powerful, scalable, complex, and expensive.

Customers interested in high-end products are typically financial service providers, telecommunications companies and large retailers: organizations that spend large sums of money on directed marketing campaign. They operate in highly competitive markets with high cost and low profit margin. These customers have enough data, money and most importantly, incentive, to engage in enterprise-level data mining. For a large telecommunications company, an increase of 1% in profit or a decrease of 1% in cost usually means millions of dollars.

Business problems of this caliber require accurate and precise analysis. To accommodate this requirement, data mining vendors offer highly customizable software. Incorporating several techniques (3-5 techniques) into a software package makes it very flexible. Depending on the problem, the customer can pick and choose among the selection of techniques in the software to create the most suitable solution. For example, in order to customize the book club's marketing campaign, different types of customers are distinguished by applying segmentation analysis. Segments are typically created by a

clustering detection technique. Then the segments are characterized by constructing a classification tree using the segment identifiers as the dependent variables.

Most business problems fall into five categories and each category is a type of problem (also known as task). By using a product that offers several techniques, customers can be certain that all five types of tasks can be solved, depending on how they use the product. They have purchased a complete solution which not only solves their current problems but can also solve unforeseen problems in the future.

High-end products support multithreaded parallel processors running in client/server environments. Most algorithms used in data mining require multiple passes over the same set of. When the database gets large, multiple passes become very computationally intensive. Using multiprocessors ensure that data can be mined within reasonable time.

As the size of the database gets bigger, scalability becomes an increasingly big problem. Since high-end products are designed to run on multithreaded parallel processors, the problem can be solved by simply adding more processors. Like all systems, eventually the data mining operation will exhibit problems of linearity, where increasing the capability of the hardware by 100% only improves the performance by 90%, for example. As the database grows, less and less improvement in performance is achieved by more processor upgrades. In practice, high-end products can scale up to the range of 20 terabytes without any linearity problems.

More sophisticated a product is, the more complex it becomes. Since these high-end products are individually tailored to each customer, significant consultation and support is needed to deal with the complexity. Specialists who possess both data mining and systems expertise are required to maintain the software as well as the network it runs on.

The advantages of high-end products do not come cheap. Prices are from \$100,000 and up (see Table 10 for price summary). In addition, there are overhead costs including the cost of buying and maintaining expensive servers, and the cost of hiring and training qualified specialists. Here is a good rule of thumb: for every dollar spent on the software, 2 or 3 dollars have to be spent on overhead.

HyperParallel, Thinking Machine, NeoVista, and IBM (see Table 3 for the market segmentation) develop products (see Table 1 for a mapping of companies to their products) that compete in the high-end enterprise-level market segment.

3.12 Mid-end Products

Mid-end products, also known as toolbox data mining, typically lack some of the functionality of high-end products but still suit many corporate needs.

The customers in this segment tend to be medium-to-large organizations who have significant amounts of data to mine and complex analysis problems to solve. Yet they do not require or are unwilling to invest in high-end data mining products.

Each product in this range offers more than one technique but is not as complete as high-level products. Some of the common data mining tasks are covered but not all. Client/server environment is support but the underlying server is single threaded. Mid-end products cannot not support large databases like the high-end products do, so they do not scale as well. Companies in this segment offer less support and consultation than in the high-end segment. The price tag is between \$10,000 to \$100,000.

DataMind and SAS Institute are some of the well-known competitors in this market segment.

3.13 Low-end Products

Low-end products, also called desktop products, have big market potential. They provide limited capabilities but are very cost effective. Used mostly by business managers dealing with data of limited complexity, low-end products require very little technical expertise.

Low-end products typically only contain one technique (either decision trees or neural nets, since they are the most common techniques used in data mining) and run on PC's supporting windows.

The customers using products in this segment put emphasis on the exploratory power of data mining rather than the need to build robust predictive models. Results from low-end products are not very precise. Since desktop users are assumed to have little technical background, sophisticated graphical user interface are employed in order to make the results appealing and easily understandable.

The best of all, prices are just around \$1,000. They are certainly cheaper than the high-end solution. Meta estimates 60% of data mining costs come from data preparation, and it is suggested steering clear of corporate commitment to the technology until the data is cleaned up and nicely data warehoused. Though it's perfectly OK in its view to dip ones' toes into data mining with desktop solutions which can work on the imported flat files which many ad hoc query tools can easily generate. Competitors in this market segment are Business Objects, Cognos, and SPSS.

Sometimes the lines between two market segments are not clear-cut, especially between low-end and mid-end. Certain products border both segments. For example, SPSS' Neural Connection is capable of performing classification, prediction, and clustering. But since it is designed for desktop use and is therefore in the low-end segment.

3.2 Technology Used

This section is intended to be used as a reference for the rest of the paper. It briefly summarizes the tasks and the techniques used in data mining. Refer back to this section

for a quick reference of certain terminology. For a more complete description of data mining, please refer to the appendix.

3.21 Tasks

Some terms in data mining are often used loosely. Different vendors use different terms to describe the same concept. Do not be confused. This section identifies each concept by all the names that are commonly used.

Data mining can be used to tackle five types of tasks: classification, prediction, estimation, clustering, and association (see Table 4 for a listing of products and tasks each products can perform). Most business problems can be phrased in terms of one of the five types of tasks. Each task is performed using one or more underlying techniques. In commercial products, whether or not a technique can perform all the tasks it is capable of depends on the actual implementation. For example, even though decision trees can be used to perform classification, prediction, and clustering, in commercial software such as Scenario and BusinessMiner, its application is limited to just classification.

[1] defines classification as examining the features of a newly presented object and assigning it to one of a predefined set of classes. An example of classification is when banks classify each loan applicant as low, medium or high risk. Decision trees, neural nets, genetic algorithms and memory-based reasoning are techniques well suited for this task. Link analysis can also apply in certain cases. Data mining companies usually mean classification when they use terms such as customer profiling, targeted marketing, and churn analysis.

Prediction, sometimes called time-series forecasting is similar to classification or estimation except that the records are classified according to some predicted future behavior or estimated future value [1]. The emphasis here is the dependence of these values on time. Market basket analysis, memory based reasoning, decision trees, and neural nets are all suitable for use in prediction.

Estimation deals with continuously valued outcomes whereas classification only deals with discrete outcomes. Given some inputs, estimation outputs values for some unknown continuous variables. For example, estimating the income of a particular family. Neural nets are well suited for estimation.

Clustering is the task of segmenting heterogeneous population into a number of more homogeneous subgroups, or clusters [1]. Unlike classification, clustering does not depend on predefined classes, which differentiates it from classification and prediction. Clustering can be implemented using market basket analysis, memory-based reasoning, cluster detection, decision trees and neural algorithms.

Association, sometimes called affinity grouping involve items that occur together in a given event. A rule such as: if item A is part of an event, then x percent of the time, item B is part of the event. Sequence association also falls into this category (if surgical

procedure X is performed, then x percent of the time infection Y will occur). Market basket analysis, memory-based reasoning, and link analysis are often used to perform association tasks.

The techniques used in data mining are not new. They are borrowed and adapted from other disciplines such as AI (artificial intelligence), graph theory, probability and statistics, and even genetics. Every data mining product incorporates one or more techniques, depending on the level of sophistication of the product, which is not surprisingly, directly related to the price (see Table 5 for a listing of products and techniques each product offers). Often, the more techniques a product incorporates, the greater its capability and flexibility, and the higher the price.

Below is a very brief introduction of techniques including its origin, its strengths and weaknesses. These characteristics are referenced in the rest of the paper. For a more thorough discussion, refer to the appendix.

3.22 Neural Nets

Neural net is one of the oldest and most frequently used techniques in data mining. The idea came from AI originally.

The strengths of neural nets include their flexibility (they can handle a wide range of problems), and their ability to handle both categorical and continuous variables.

Even though neural nets arguably offer the most advanced data mining power, it has several weaknesses. The most serious weakness is that it lacks explicitness - the process is often not explained. It is almost as if the pattern discovery process is handled within a "black box" procedure. The interesting aspect is that neural nets' weaknesses are also their strongest assets. The difficulty that lies in explaining the process means that the results are often new and non-intuitive. They have never been thought of before, so the results could potentially lead to a radical change of view or a revolutionary new idea.

3.23 Decision Trees

Decision trees are borrowed from AI and statistics. The decision trees algorithms commonly used are CART, CHAID, and C4.5. CART can only build binary trees and it grows the full tree before pruning it, causing overfitting. C4.5 is similar to CART except that it can produce varying numbers of branches per node. While CART and C4.5 can accept both categorical and continuous values, CHAID is restricted to categorical variables. Continuous variables will have to be broken down into categories when using CHAID. Unlike CART and C4.5, CHAID does pruning while it builds the tree so that overfitting does not occur.

There are several major advantages as well as disadvantages in using decision trees. Decision trees generate understandable rules no matter how complicated the inputs are. It is generally easy to follow any one path through the tree, so explaining the decisions along the way is easy. Computation cost for at each split is inexpensive. In practice, algorithms tend to produce decision trees with a low branching factor with simple tests at each node, so the tree does not grow out of hand and these tests translate into simple Boolean and integer operations that are fast and inexpensive. Using decision trees, the field which is the best at splitting the training records can be singled out for analysis. This will enable the user to figure out which variable influences their data the most.

Even though decision trees are great at classifying data, they are not appropriate for estimation. It is also hard to use decision trees for problems involving time series data unless a lot of effort is put into presenting the data in such a way that trends are made visible.

3.24 Memory-Based Reasoning

Memory-based reasoning (MBR) is also known as nearest neighbors. Like decision trees and neural nets, the idea came from AI. MBR looks for the nearest neighbors in the known instances and combines their values to assign classification or prediction values.

Because MBR generates a list of the nearest neighbors, the results produced can be readily understood. MBR is applicable to arbitrary data types, even non-relational data since the technique does not depend on the underlying representation of the data. The performance of MBR does not depend on the number of fields in the records which makes it practical to use when other techniques such as neural nets cannot make sense of the data.

However, MBR requires costly computation to perform when doing classification and prediction since finding the nearest neighbors involves applying the distance function to all the fields in the record and all the records in the training set.

3.25 Genetic algorithms

Genetic algorithms applies the mechanics of genetics and natural selection to a search used for finding the optimal sets of parameters that describe a predictive function.

A common application is to apply genetic algorithms in training neural nets. Many neural nets packages incorporate genetic algorithms as an option for training.

Genetic algorithms produce results that are explainable. The results can be easily applied because they take the form of parameters in the fitness functions. In many cases, genetic

algorithms are used for finding optimal values. They are not limited to the types of input data - as long as the data can be represented as a string of bits of a fixed length.

Even though genetic algorithms are suitable in optimization, there is no guarantee of optimality. They may struck one local optima and never find the best solution (global optima). Genetic algorithms can be quite computationally intensive, therefore products incorporating them tend to be enterprise-level products that run on powerful servers.

3.26 Link Analysis

Link analysis is referred to as association analysis sometimes. The technique came from the field of graph theory. Link analysis follows relationships between records to develop models based on patterns in the relationships. It is particularly useful for crime solving and the telephone industry because these areas naturally involve links. Link analysis also offers powerful visualization and makes the process easily understandable.

However, link analysis is not appropriate for many types of problems. It does not apply to classification and prediction like neural nets that take data in and produce an answer. Link analysis finds specific patterns which can then be applied to data.

3.27 Market Basket Analysis

Market basket analysis applies probability and statistics measure to the database. The statistic methods used here are different from the standard statistic methods such as regression. Market basket analysis is a form of clustering used for finding groups of items that tend to occur together in a transaction.

The output using market basket analysis are in the forms of rules, which are clear and understandable. When approaching a large set of database without knowing where to begin (i.e. no predefined notions or conjectures), market basket analysis is very useful because it looks for the pattern inherent in the data. It is also simple to use compared to more complex techniques such as genetic algorithms or neural nets.

However, the computational effort grows exponentially as the problem size grows. Even though methods are available to generalize the items analyzed, important rules might be eliminated in the process of generalizing. Market basket analysis works best when all items have approximately the same frequency in the data. Items that rarely occur are in very few transactions and will be pruned.

3.28 Cluster Detection

In clustering detection, there is no pre-classified data and no distinction between independent and dependent variables. Instead, the algorithms search for groups of records - the clusters that are similar to one another, with the assumption that similar records represent similar customers who will behave in similar ways. Cluster detection is rarely used by in isolation because finding clusters is not an end in itself. Once clusters have been detected, other methods must be applied in order to figure out what the cluster mean.

Cluster detection is undirected data mining - it can be applied even when there is no prior knowledge of the internal structure of a database. Hidden structures are uncovered to improve the performance of more directed techniques. Since there is no need to specify independent values, dependent values, input and outputs, cluster detection is very easy to apply.

The flip side of undirected knowledge discovery is that when you don't know what you are looking for, you may not recognize it when you find it. The clusters you discover are not guaranteed to have any practical value.

3.3 Systems Requirements and the Size of the Database

While more sophisticated data mining products require the use of high-end hardware, they also require more memory. Potential buyers of data mining products can compare the platform and systems requirements needed for each product and estimate if their companies have the necessary resources. High-end products support multithreaded parallel architecture and incorporate complex techniques so they need very powerful servers to run on. The clients usually run on PCs or workstations (see Table 6 and 7 for a summary of hardware requirements for each high-end product).

Low-end products can be run on a desktop PC with reasonable memory requirement. The requirements match the capability of the users they target - mainstream users with mainstream personal computing power (see Table 8 for a summary of hardware requirements for each low-end product).

Regardless of the market segment, it is better if the software can be run on industry standard computers, such as PC, IBM, HP, and Sun, so that service is readily available in case of a hardware failure.

As a company grows, be it large, medium, or small, its database grows as well. High-end products offer scalability. Literally, scalability means that as a system gets larger, its performance improves correspondingly. For data mining, scalability means that by taking advantage of parallel database management systems and additional CPUs, the user can solve a wide range of problems without needing to upgrade the underlying data mining environment. Scalable features are only available in systems running in multi-processor

environment and therefore a distinct feature of the high-end products. Scalable features are important to look for when shopping for a product, especially for customers who anticipate database growth in the near future.

3.4 Ease of Use

It is important to know how easy it is to use a product before purchasing. The ease of use takes two forms, the ease of use of the technology and the ease of use of the interface.

High-end products are harder to use because they incorporate more sophisticated methods and because the products are tailored for specific use. Knowledgeable specialists are needed to run the process and interpret the results. They also must possess systems expertise since high-end products run on multithreaded parallel server/client networks. In comparison, low-end products can easily be used by any business manager with very little technical knowledge.

Regardless of which end a certain product falls in, a good user interface should include clear explanation of results in tabular and graphical form. Some products automatically generate results in html form for web viewing. Some products use a decision tree to further illustrate the results. Easily understandable results can help a company reap maximum benefit from its data mining operation.

3.5 Ease of Integration

Data mining is not performed alone. Data warehousing, data mining and OLAP almost always go hand in hand. Some potential buyers who already have data warehousing and OLAP solutions need to think about how the new data mining product will interact with the existing system. Those buyers that wish to implement data warehousing, OLAP and data mining need to figure out what data warehousing and OLAP products to purchase along with the data mining products.

To this end, companies that offer data warehousing and OLAP along with data mining products have a huge advantage (see Table 2 for a listing of companies and the types of products they offer). They offer buyers a one-stop shop with less problems in integration. They have more expertise to deal with their own products should problems arise.

3.6 Experiences

Even though the main focus of this report is on the products, some focus should be placed on the companies that make the products. How experienced is the company? How can a potential buyer be confident that the product will live up to the standards the company claims? A look in the background of the company as well as its past customers provides a good clue (see Table 9 for company profiles).

A background study is based on the assumptions that companies that are consistently successful financially usually mean that their products have withstood the test of many customers. An indication of success can be reflected by the number of employees, the number of offices, and the awards it has received.

A company's past clients is also a good indication. A company that can boast a long history of successful clients most likely has a very good product. Looking at the type of clients can also give away the industries that the company specializes in.

Since we will not be able to know for sure if a product is good unless we purchase it and use it, to a large extent we need to rely on the reputation of the company. A company with a sketchy history might just go out of business very soon leaving their customers without service and support

Also, the location of a company can be very important even after a purchase is made. In case service and support is needed, it would be much more convenient to have the company (or a branch office) near by than to have to travel to a geographically remote location, especially high-end and mid-end products that require more consulting and support. To this end, large companies with many offices and support centers are more desirable.

3.7 Training Programs and Support

The data mining process is typically complex. Even though companies try to make their products as simple as possible, some difficulties in using data mining technology still remain.

High-end products should have more elaborate training programs such as on-site visits. For products in this segment, continuous on-site visits and support are necessary throughout the life of the products. Customers usually need to retain a support team specifically for the system. Vendors typically charge 15-18% service fee for support and upgrade.

Mid-end products need initial training including on-site visits. As customers become more familiar with the product, the amount of consultations decreases.

Low-end products usually come in a small nice package very much like the software that can be purchased in any computer store. They require minimum training and support since they are less complicated and does not require much technical knowledge.

To ensure that customers can use the product and mine with ease, some companies offer day seminars. Regardless of which segment the product falls in, customers should only purchase from vendors who offer life time product support.

Regardless of the type of products, a training program as well as a knowledgeable customer support staff is a must.

4.0 Companies and Products

4.1 Business Objects

Targeted towards mainstream end users, the data mining product offered by Business Objects, BusinessMiner, occupies the desktop segment of the market (low-end).

BusinessMiner uses decision trees to perform classification of data. Because of the particular decision trees algorithm it uses (CART), BusinessMiner is only capable of splitting into two branches at each node and the resulting tree is binary. BusinessMiner uses a special technique, called Intelligent Binning. Binning is choosing the number of bins into which a numeric range is split. For example, if salaries range from \$20,000 to \$100,000, the values must be binned into some number of groups. Often users are required to set binning manually. As a result, certain values maybe lost or improperly binned. Intelligent binning automatically and intelligently bins numeric values based on the range, values and distribution of the data. Both texts and numbers are supported in the data mining process which makes it suitable for business applications. BusinessMiner can build decision trees at the rate of 1,000 rows of data per second.

Running BusinessMiner requires PC 486 or higher, 12 MB RAM (16 MB is recommended), and 30 MB free hard disk. BusinessMiner runs on Windows 95, Windows NT, or Windows 3.1.

Business Objects offers data mining, data warehousing, and OLAP solutions. BusinessMiner is offered as an option to BusinessObjects 4.0, the OLAP software, even though BusinessMiner can be used as a standalone product. BusinessMiner by itself is compatible with ASCII and Excel files.

Since only decision trees are used, learning to use BusinessMiner is easy. Decision trees are known for presenting results in clear and understandable rules. The interface has a distinct Microsoft-Office look, which makes using it easier for users who are familiar with Windows. Outputs are organized in 3D graphs, charts, tables, and trees to help visualization. Online help library and wizards that guide users through mining are also

available. Business Objects offers training courses, customer-specific consulting services, and technical support services.

BusinessMiner is priced at \$995 per license (\$495 if purchased with BusinessObjects 4.0).

Business Objects was founded in 1990. It has 700 employees worldwide and have sold 700,000 licenses to over 5,800 organizations around the world. Business Objects generated \$34.5 million in revenue in the most recent fiscal quarter, a 48% increase from the same fiscal quarter last year. Its most well-known data mining customers are Bank of America, British Airways, British Petroleum, British Telecom, Chevron, Citibank, EDS, Fannie Mae, ITT, Mastercard International, Pacific Bell, Pepsi, Solomon Brothers, Victoria Secret, Western Digital. Recently Business Objects moved its headquarters from Paris to San Jose to concentrate on strengthening relationships with U.S. Partners and clients. Product development and marketing have remained in Paris.

4.2 Cognos

The two lines of data mining products offered by Cognos are: Scenario and 4Thought. Scenario and 4Thought both belong to the desktop market segment (low-end).

Scenario employs a decision tree algorithm to perform classification. It is particularly good at identifying and ranking high impact factors. Inputs of both categorical and continuous values can be used in Scenario. An estimated 40,000 records can be analyzed by Scenario in roughly 3 minutes.

Scenario is designed to run on desktops. It requires a 486 PC or higher, minimum of 8MB of RAM, and 20MB free disk space. It only runs on Microsoft Windows 95 or Windows NT.

The user interface for Scenario has a distinctive windows feel. 2-D graphs, tables and statistical information are used to illustrate the analysis of the data and the results. The user can choose which inputs (factors) to emphasize on by clicking on that input in the table. Tutorials and wizards are also available within the package. The presentation of the results is intuitive and therefore makes Scenario user-friendly.

Scenario is part of the COGNOSuite which also includes OLAP and data warehousing products although it is not necessary to purchase the whole COGNOSuite in order to run

Scenario. Full integration is available with the rest of the COGNOSuite so that a click of a button will start the data mining operations. Scenario supports data from text files, Excel, Lotus 1-2-3 worksheets, and dBase tables.

Priced at \$695, Scenario's ease of use and ease of integration makes it a good buy even though the method it uses is not as powerful as some other products. Scenario received the PlugIn Datamation 1998 Product of the Year Award, as well as the PC Week's Analyst's Choice Award.

The other data mining product offered by Cognos is 4Thought. It has more capability than Scenario. 4Thought was originally developed at Right Information Systems. Cognos acquired Right Information Systems in April, 1997.

4Thought uses a combination of neural nets and statistical tools and is therefore, very useful for "number crunching". It is especially good for the financial industry where large quantity of data is dealt with. Because of the nature of neural nets, 4Thought can be used to perform prediction. Statistical tools can support optimization analysis.

4Thought offers a familiar spreadsheet interface in which to collect and prepare data for analysis. User can type values directly onto the spreadsheet, cut and paste, or import data directly from other sources, such as Excel, Lotus, popular relational and non-relational databases, and text files. Both categorical and continuous data can be inputted into 4Thought.

A variety of line, bar, and area charts let users graphically view data and interpret the results. Scattergrams and overlaid charts show the strength of relationships between factors, or even the strength of the model's predictability.

4Thought comes with a higher price tag - \$20,000. Even though it occupies the low-end segment of the market, it provides more powerful features than most low-end products.

Cognos offers standard or customized training classes through regularly-scheduled public classes either in one of Cognos' classrooms or on-site. And since Cognos has 32 offices in 12 countries, the training programs are relatively accessible. Support is available in the form of on-line, telephone, and in person from Cognos' six support centers around the world.

Founded in 1969, Cognos is an international corporation with corporate headquarters in Ottawa, Canada and U.S. sales headquarters in Burlington, Massachusetts. The company employs more than 1,400 people worldwide. Revenue for the most recent fiscal quarter (1998) is US\$70.7 million, a 21% increase from the same period last year (1997). Some of the more well-known customers include ADP, Mead Johnson, Consolidated Edison, Vanguard, and Deutsche Bank.

4.3 DataMind

DataCruncher occupies the mid-end segment of the data mining market. It offers the Agent Network Technology that uses a belief network. A belief network is essentially a hybrid of neural nets, decision trees, and market basket analysis. The hybrid software compensates the weaknesses of each technique with the strength of others. Neural nets is hard to understand, but very powerful. Decision trees is very easy to understand but lacks some of the capabilities of neural nets. Combining them makes the hybrid powerful and easy to understand even though some of the capability of neural nets is comprised, such as the capability to perform estimation. DataCruncher can perform classification, association, and clustering. DataCruncher typically can support a system with 50 million users and hundreds of fields per user.

DataCruncher supports client/server computing. The server component runs on Unix or Windows NT systems and performs operations including mining the data, reading data sources and building models. The client component that runs on Windows 95 or Windows NT, is responsible for initiating server-side data mining operations, viewing results, , and building reports.

The server requires the use of Hewlett-Packard HP-UX, IBM AIX, Silicon Graphics IRIX, or Sun Microsystems Solaris. The systems requirements are 64 MB of RAM, 15 MB of free disk space, and additional working storage space dependent on volume and complexity of data. The client can run on desktop PCs with 16 MB of RAM, 15 MB of free disk space, and additional working storage space dependent on volume and complexity of data. A special version of Data Cruncher, called PowerPak, can be used on a standalone PC and performs all data mining operations on a single platform.

In addition to the usual reporting formats (graphs, tables, etc.), DataCruncher is capable of generating HTML-standard reports which allows information sharing on corporate intranets through web-viewing.

DataMind specializes in data mining products and therefore does not have any data warehousing or OLAP software. However, DataCruncher has direct connection to Oracle and Informix, delimited ASCII files and ODBC compatible databases.

A five user system (1 server, 5 clients) costs \$80,000. The standalone product, PowerPak costs \$25,000

DataMind was founded in 1994 and has headquarters in San Mateo, California. Regional sales centers are located in Atlanta, Boston, Chicago, San Mateo, and Paris, France. It employs approximately 42 people and is a privately-held company. Some well-known

DataMind customers are ADP, 360 Communications, Engage Technologies, and Chase Manhattan Bank. DataCruncher received the fifth annual Crossroads A-List Awards.

Training programs are presented either on-site or at DataMind locations in California. An introduction to data mining class and a business data mining class are offered. The introduction course lasts one half day and is designed for anyone interested in the basic concepts of data mining and how to use DataMind products. The business data mining is a two-day course with hands-on work tailored towards individuals who use DataCruncher client side tools. On-site consulting services are available from DataMind.

4.4 HyperParallel

HyperParallel offers a suite of data mining tools. //Discovery (HyperParallel uses "/" as a prefix pronounced "Parallel" for all of its software) occupies the high-end segment of the data mining market. The software focuses on retail, banking, and telecommunications, areas that have the most demand for enterprise-level data mining products.

//Discovery tool suite includes a portfolio of data mining techniques: //Induction, //Cluster, //Affinity, //Sequence, //Neural, and //Spatial. //Induction uses decision trees and can be used in target marketing, customer retention, credit risk and authorization, fraud detection, and sales forecasting. //Cluster uses a cluster detection algorithm for customer and market segmentation, and target marketing. //Affinity performs association using market basket analysis to help an enterprise to understand product substitution. //Sequence is very similar to //Affinity except that //Sequence deals with time-oriented data and can detect sequential patterns. //Neural uses neural nets for fraud detection and target marketing where accuracy is important, but understanding the precise means used to generate the results is not. //Spatial is a geographic algorithm that determines the impact of travel-time on customer purchase behavior. Each technique can be used separately based on corporate needs.

HyperParallel uses the concept of //Recipe. Each recipe is essentially a individualized portfolio containing different sets of techniques suitable for a particular industry. Each recipe typically include two or three pre-selected techniques designed to solve a particular problem within an industry. The nine recipes for retail are Local Store Assortment, Markdown Management, Recency-Frequency-Monetary, Segment-of-One Marketing, New Product Introduction, Promotional Forecasting, Seasonal Forecasting, Site Selection, and Ad Item Selection. There are also ten recipes for banking and four recipes for telecommunications. By providing five different techniques, HyperParallel users can mix and match to build a customized and hence more effective data mining strategy. The whole spectrum of common data mining tasks are covered. Each recipe includes business rules specific for the application. These rules are written in the HyperParallel Recipe Language (HRL). Customers can use the HRL to customize the recipe for individual application requirements.

//Discovery tool suite uses multithreaded parallelism. //Enabler is a multithread manager that takes care of problem of concurrency and scheduling and sits on top of the operating systems. The use of multithreaded parallelism gives it true scalability. The server can be used to build models and models can then be applied on the client. //Discovery server runs on IBM RS/6000, NCR Worldmark, Pyramid, and Sun Microsystems servers. The client can run on simple Unix workstation or a Windows PC supporting Windows 95.

HyperParallel does not offer data warehousing and OALP products. //Discovery has ODBC (Open Database Connectivity) access to SQL databases.

Enterprises initially spend about \$120,000 on a basic template which includes //Enabler. Pricing for each //Recipe is based on the number of CPUs and techniques used in the framework. A //Recipe typically uses two or three techniques. For 1-16 Unix processors, each algorithm costs \$40,000, scaling to \$50,000 per technique for 17-32 Unix processors, \$70,000 pre technique for 33-64 processors, and \$90,000 per technique for 65 or more processors.

Because of the complexity of //Discovery in the sense of the software and the underlying systems, specialists with general systems expertise are needed to use the software. HyperParallel offers week long training courses in their testing lab for people with general systems background.

HyperParallel was founded in 1994 and now employs 33 people. It is headquartered in San Francisco. Some of the well-known customers are Bank America and Wal-Mart.

4.5 IBM

IBM's data mining product, Intelligent Miner, is considered to be enterprise-level (high-end). It offers decision trees, neural nets, link analysis, clustering algorithms to perform all of the five common data mining tasks: classification, prediction, estimation, association, and clustering.

Intelligent Miner runs in client/server environment with powerful servers. Multithreaded parallelism is supported which enables the software to be scalable. The servers can run on

IBM AIX, AIX/SP, OS/390, Solaris, OS/400, and Windows NT. Clients run on AIX, Windows NT/Windows 95, and OS/2.

Output of the data mining results are presented in graphs, bar charts, and tables. Results can be put into DB2 format and viewed across the entire network. Intelligent Miner supports an API (Application Programming Interface) which allows analysts to write application specific code. It enables development of customized mining applications by users.

IBM also develops data warehousing, and OLAP tools that are integrated with Intelligent Miner. In addition, Intelligent Miner can support data sources from DB2 databases, flat files, and DRDA-connected platforms. Intelligent Miner costs between \$30,000 to \$150,000 per server, depending on the server.

Founded in 1914, IBM is now an international corporation employing approximately 240,000 people worldwide. The revenue generated in the most recent fiscal quarter is \$17.6 billion, an increase of 1.7% from the same fiscal quarter last year. Customers of the database products include Japan Airlines, Boeing, Yamaha, EMI Music Publishing, NHL, L.L. Bean, Nynex and Merrill Lynch.

4.6 NeoVista

Decision Series is a multi-algorithmic tool set that occupies the high-end of the market segment. Decision series offer DecisionNet, DecisionCluster, DecisionGA, and DecisionAR.

DecisionNet is based on an extended back propagation neural network algorithm. It overcomes limitations common to conventional neural network technologies with innovations that improve generalization, accuracy and reproducibility, and reduce training time, thus reducing the "black box" effect. DecisionCluster performs clustering on the database by using distance-based clustering algorithm. It can also use DecisionNet to predict whether two items are the same. DecisionGA is a genetic algorithm that is used for breeding potential cases based on a loosely constructed model - the cases can converge to the best example. DecisionAR uses link analysis algorithms to perform association tasks in order to determine the likelihood of events occurring at one instant in time or sequentially. Depending on the goals of the application, user has the freedom to

choose one or any combination of the above techniques. Collectively, these techniques cover the five common tasks performed in data mining.

Decision Series runs in client/server computing environment. The server platforms such as HP, Sun, or DEC. The client side runs on Microsoft Windows 95 or Windows NT. The use of parallel computing environment gives Decision Series scalability.

NeoVista does not offer data warehousing or OLAP. DecisionAccess provides the framework that integrates each of the knowledge discovery tools with each other and relational databases to ensure that the discovery environment is compatible to the data warehouse. Decision Series can use data source from Informix, Oracle, Sybase, and Microsoft SQL servers. DecisionAccess allows automated translation between relational databases and discovery tools.

Since Decision Series uses multi-algorithms and runs in a client/server environment, using the software requires technical expertise. NeoVista offers training programs ranging from seminars to on-site visits.

4.7 SAS Institute

Enterprise Miner occupies the mid-end market segment. Decision trees, neural nets and traditional statistics are used to perform classification, prediction, and clustering. The decision trees technique uses a CHAID variant, but a version of CART is under development.

The software supports client/server computing for use with IBM, SPARC, and Sun workstations. Minimum requirement of 32 MB RAM and 8 MB free hard disk.

Like other mid-end and high-end products, Enterprise Miner is not a ready-to-use solution for a particular business problem. Applying the system requires substantial human effort from both business and technical sides.

The graphical user interface uses 3D graphs, charts, tables, trees, and texts to help visualizing the results. All results can be viewed in multiple ways by the click of a button.

The greatest advantage in using Enterprise Miner is the concept of end-to-end solution. Since SAS develops other lines of products for data warehousing and OLAP. Full integration of these two products with data mining is available, which provides users who are in the market to buy data warehousing, OLAP and data mining a one-stop-shop solution.

Prices range from \$8,000 to \$16,000 depending on the size and the capability of the workstation.

Founded in 1976, SAS employs 4,500 people worldwide in over 50 countries. 3.5 million licenses have been sold to over 120 countries. Some well-known customers are Pfizer Inc., Ford Motor Company, and Dun & Bradstreet. Training and support services are provided at each of the 50 locations.

4.8 SPSS

SPSS specializes in developing statistical tools for business intelligence. The two data mining products it offers are: AnswerTree and Neural Connection. AnswerTree and Neural Connection each offers one algorithm and are considered to occupy the low-end of the market.

AnswerTree uses decision trees to perform classification. It uses 4 different decision tree algorithms: CHAID, CART, Exhaustive CHAID and QUEST. Quest is a statistical algorithm that selects variables without bias and build a binary tree. Unlike other algorithms, QUEST performs variable selection and split point selection in separate stages.

AnswerTree runs on Windows 95 or Windows NT. It requires a 486DX processor or higher (includes math co-processor), 40 MB hard drive space, 12 MB RAM (although 16 MB is strongly recommende), and VGA monitor (SVGA recommended).

AnswerTree is compatible with other SPSS statistical packages. The user interface for AnswerTree uses 2D graphs, diagrams and tables to present the results.

Neural Connections employs neural nets to perform classification, prediction, and clustering. Approximately 10 records per input variable are needed to train the neural nets. Neural Connections can handle up to 32,000 records and 750 inputs. Inputs are equivalent to variables with the exception that every level of a categorical variable counts as one input.

Systems requirements for Neural Connections are Windows 3.1, Windows 95, or Windows NT on a 386 or better PC (math co-processor strongly recommended), as well as 4 MB memory (8 MB recommended), 4 MB free hard drive space.

Neural Connections and AnswerTrees can be both launched from an SPSS menu. However, they are also compatible with other types of files such as ASCII and Excel.

Neural Connections has one of the better output formats. It uses 3-D contour plots that could be rotated three dimensionally as well as tables and texts.

The pricing for both Neural Connection and AnswerTree is based on number of licenses. There are two types of licenses offered by SPSS, an annual license and a perpetual license. An annual License is a lease transaction and allows for use of the product for one year. An initial fee is paid and a yearly renewal fee is required for the continuing use of the product. A perpetual license allows for indefinite use of the software. A higher initial fee is required and a service fee is optional on a yearly basis. The user can choose not to pay the service fee and can still keep on using the software. For either Neural Connection or AnswerTree, an annual license for 1 user requires an initial fee of \$375 and renewal fee of \$203 yearly. A perpetual license for 1 user requires an initial fee of \$665 and a Maintenance fee of \$35 yearly, which is optional.

SPSS was founded in 1975. It now employees 535 people. Revenue generated in the most recent fiscal quarter was \$28.5 million, a 4% increase from the same fiscal quarter last year. SPSS has sold approximately 250,000 licenses worldwide. It was ranked No. 11 among the 200 Best Small Companies in America by Forbes for 1997, and was ranked No. 73 in Business Week's Top 100 Growth Companies for 1997.

4.9 Thinking Machine Corporation

Thinking Machine Corp specializes in data mining products. Some of the well-known customers are GTE and Credit Suisse. The multi-algorithmic tool suite of tools it offers, collectively known as Darwin, occupies the high-end of the market segment. Darwin is a collection of three different techniques, StarNet, StarTree, and StarMatch.

StarNet uses neural nets. The user has the option to adjust the activation functions training algorithms and cost functions. A genetic algorithm, StarGene, can be used to automatically optimize the weights used in the training set. StarTree uses decision tree technique, more specifically CART. StarMatch uses memory-based reasoning technique. Even though each technique alone is prone to the inherent weaknesses of the technique, When all three techniques are used in combination, some of the weaknesses are covered up to produce a well-rounded solution package.

Darwin runs on a Windows 95/NT client/Unix server, parallel computing architecture. The client/server architecture allows the users to mine data on servers while accessing Darwin's graphical user interface from the Windows desktop. Users may run Darwin on single or multiple CPU servers. The server requires a minimum of 24 MB RAM (32 MB or higher is recommended) and runs on Sun Solaris 2.5.1 or IBM AIX 4.1.4. The client runs on Windows 95/NT. Minimum of 24 MB RAM is required (32 MB or higher is recommended) and 27MB Disk Space is required as well. Because of the use of multi-

processor computing environment, Darwin has good scaleable features that exhibit almost linear scaling factors.

All three techniques are integrated into one common user interface. The interface displays lift charts, return on investment (ROI) and margin charts, 2D scatter plots, histograms, and line graphs. Models generated using Darwin can be exported as C, C++ or Java code. This is a distinctive feature of high-end products which is lacking in other levels.

The use of client/server parallel architecture, the multi-algorithmic tool suite and along with some of more sophisticated features make Darwin more capable and flexible. At the same time, it makes the learning curve more difficult. More general technical expertise as well as an intuitive understanding of the data mining techniques are needed to set up and maintain the product. Specialists are needed to run the data mining process.

Darwin also has an interface that accesses data from data warehouses, relational databases via ODBC with support for SQL queries and text files. However, since Thinking Machine Corporation does not offer any data warehousing or OLAP products, the process of setting up an OLAP server, data warehouse and then apply Darwin might be complicated.

The pricing is based on the number of processors on the server. The first process license costs \$50,000 and each additional processor is \$20,000. An annual maintenance fee of 15% is charged for support and updates. Each client costs \$995. Since using Darwin requires technical expertise and most companies only have a few specialists, the main portion of the cost comes from server, not clients.

Thinking Machines offers training services to aid in the process of data mining discovery which includes a 3-day educational course on using Darwin held at Thinking Machines or on-site. Consulting service is also available for Darwin.

Thinking Machine Corporation was founded in 1983. It now employs 175 people. Headquartered in Burlington, MA., it also has 3 other offices in the U.S., located in New York City, Dallas, and Washington DC. An international office is located in Japan.

5.0 Conclusion

The data mining market is very complicated. Choosing a right product involves many factors. Corporate needs and resources should be thoroughly evaluated before deciding which market segment of products to choose from. Once that is decided, products within the market segment must be compared. Data mining is not an off-the-shelf software, so it takes a lot of effort to find just the right one.

6.0 Appendix

6.1 Appendix A - Bibliography

[1] Michael J. A. Berry, Gordon Linoff. Data Mining Techniques. Wiley Computer Publishing. 1997.

[2] Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation. 1997.

[3] Company materials and their web sites.