Mining Very Large Databases to Support Knowledge Exploration

Exploitation of data mining may be increased by moving away from dependence on statistically trained experts, and by making data mining, both in terms of its application and its results, more easily deployed within the business. Full volume data mining avoids the constraints of sample based approaches. A data mining framework, enabling software applications to be developed with simplified interfaces, increases the usability of these techniques. Encapsulating resultant predictive models as components enables easy deployment of the results within a business. A workshop style environment is appropriate, where business managers need to work quickly with a powerful data mining capability.

Who Should Read This White Paper?

This white paper has been written for technically adept end users and data mining professionals. It would also be of interest to interview candidates for consulting positions at WhiteCross.
1.1 Applying Data Mining to the Business ................................................................. 5
1.2 Data Mining Using Sampling .............................................................................. 7
   Technical Sampling Issues .................................................................................. 7
   Practical Sampling Issues .................................................................................. 8
   Business Users Not Supported by Sampling ..................................................... 9
1.3 In-Place Mining of Data with VLDB ................................................................. 11
   4. Integration of Multiple Data Mining Algorithms ........................................... 14
   5. Deployment of Data Mining Through Applications ..................................... 16
   6. Deployment of Data Mining Through Predictive Models ......................... 19
   7. Exploration Workshops for Knowledge Discovery ..................................... 21
References ........................................................................................................... 22
1.1 Applying Data Mining to the Business

Data mining is routinely prioritized as a major force in garnering information, in order to gain competitive advantage. However, the promise of delivering a deep understanding of the drivers behind customer behavior, and developing the predictive models that can be deployed within the organization, is often not realized. The following are three of the main reasons why data mining projects fail.

♦ There is a disconnect between the business managers and the quantitative experts in understanding the problem to be addressed.

♦ Data mining tools are often unable to scale appropriately to the size of the data mining task.

♦ Tools don’t support deployment of data mining results into operational systems.

In a competitive environment, successful businesses are driven forward by visionary and creative managers.

In data mining it must be remembered that the business manager is the problem owner—it is he or she who has to make decisions, set prices and design customer propositions, based upon the results of the data mining. It is he/she who has the responsibility for driving forward the business, and whose personal skill set is likely to be business administration or specific tasks such as marketing.

In contrast to the requirements of the business manager, most data mining packages available today are tools designed for advanced users with a strong statistical background. This bias is due to the skills needed to use the tools, and the background required to correctly translate a business problem into a format suitable for analysis through a data mining tool.

Commonly, businesses find themselves in a paradigm of relying upon highly skilled mathematical experts to perform Data Mining studies. This is unfortunate, and perhaps indicative of the relative immaturity of data mining, because the consequences of this mode of working include:

♦ Delay in delivering the solution, as the problem is transferred between the manager and the analyst

♦ The risk that a language barrier between the manager and analysts leads to misinterpretation of the business problem and/or results.
♦ Loss of creativity in exploring the possible solutions to the business problem.

This final point is perhaps most significant, because data mining projects are not easily structured into distinct phases, such as a requirements specification, functional specification, analysis and results. They are inherently far more fluid and iterative. As initial results are found, the business problem is often changed as managers gain insight into their data, and refine or redesign the goals of the exercise. This close interaction with the data and repeated redefinition of the business problem are key success factors in data mining work. It is akin to asking questions, and based on the knowledge gained in the answers, asking further or new questions.

If the paradigm of transferring the analysis between the business manager and the quantitative expert is maintained, then the number of iterations may be reduced, the time expended fruitlessly exploring dead ends increased, and the overall quality of the results compromised. This is exacerbated by any urgency attached to the project—and most worthwhile projects in a competitive environment require urgent results.

A major area of technical complication in data mining is the use of statistical sampling to perform data consolidation. This arises because the volumes of data involved in business problems are often well in excess of those that can be handled on data mining platforms. A frequently taken compromise is to sample the original dataset in order to give one, two or more orders of magnitude reduction in the data volumes. Problems associated with this kind of sampling are discussed below. The disconnect between the business problem owners and the data mining analysis may be overcome by

♦ Simplifying the interface to data mining technologies—embedding data mining within vertical business applications, rather than deploying only horizontal data mining workbenches

♦ Scaling up the data mining to the full volume of data to avoid the complexities and risks introduced by sampling

♦ Using workshop style interaction sessions that enable business managers to exploit an extended data mining capability.
1.2 Data Mining Using Sampling

Data mining tools sometimes use sampling to get around the fundamental computational limitations of most hardware and software environments. Traditional hardware environments can’t compute the necessary number of iterations within reasonable time periods. The use of client-side processing, where a data-cube is brought back from the database server onto the client workstation for processing, inherently constrains the data set size to that which can be handled by the workstation; typically only tens of megabytes in memory. Many of the algorithms involved in data mining are not scaleable with respect to the size of the source data set. As the size of the data set increases, the computation time of the algorithm rises exponentially.

While sampling has been necessary in order to make complex problems tractable, it must be noted that there are costs associated with sampling, which can be divided into technical and practical problems.

Technical Sampling Issues

The confidence with which results can be generalized is related to the sample size. As the sample size is reduced, so the ability to predict the behavior of the population being modeled is reduced. This is a problem for all data mining algorithms, but it is particularly acute for decision trees; each new level of the tree decreases the number of rows in the terminal nodes exponentially.

As you go deeper into a decision tree, the number of rows represented by a node reduces until the point where there are insufficient rows to enable any more specific rules to be discovered. Taking a naïve mode, where each node represents the penetration of a Boolean objective, and the tree segments the population into regions of significantly high and low penetration, then a Binomial probability model may be used to predict the necessary node sizes, in order to generalize at a given confidence interval. For example, a 500 row node expressing a penetration at 20% would generalize at the 95% confidence interval to a penetration of [16.5%, 23.5%] in the population or a 7% interval size [1].

In order to maintain the ability to generalize confidently, it is usual to limit the sophistication of models; a simple method is to restrict the number of variable interactions that are considered.

A further problem with sample-based data mining is that the results of the project are only valid to the extent that the sample is representative of the population being considered. In terms of classification-based decision trees, where the implication is
that once the tree has been generated and before a rule can be used, the data at every separate terminal node in the tree should be verified as being representative of the similarly constrained data in the whole population. In the realm of knowledge discovery, in order to be confident of a sample, you must verify the sample in light of the discovered knowledge. This necessarily introduces considerable delay and repetition into the data mining project.

**Practical Sampling Issues**

The use of samples as a basis for any decision model necessarily results in that decision model being merely an estimate rather than the absolute truth. Wherever practicable, it is desirable to know the absolute truth rather than an estimate thereof. This is especially important when working in the financial/auditing arena.

Handling rare events is difficult with samples. Events such as fraudulent behavior may only occur in a very small percentage of the base; taking a simple sample is very likely to over-represent or under-represent this behavior pattern in the data.

![Figure 1](image_url)  
*Figure 1 – A simple sample is likely to over- or under-represent rare behavior.*

Data mining is only one part of exploring data in order to investigate a business question. Other common operations include “slicing and dicing” data—that is, repeatedly drilling down into the data making increasingly specific qualifications on
the data. For example, in the travel industry one might initially view the data by year, then by year and resort, then by year, resort and week. This style of drilling into the data very quickly hits limits in samples. Examining the most common values, or TopN analysis, for example identifying the top 1000 telephone numbers called in a given month, is also severely restricted by sampling.

However carefully the initial sampling may be undertaken, and whatever scheme for selecting a sample is used, it is inevitable that discarding data introduces a risk that significant information about the business problem is also discarded.

Aggregation techniques may be employed as a data reduction technique. However, the loss of information through aggregation should be considered. By way of example, consider retail data for sales indicating the volume of sales per item per hour. This could be aggregated by summing all the data items on a weekly basis, which would significantly reduce the data volume. However, this method removes any pattern of sales with respect to the day of the week, or the hour of the day.

While sampling may be seen as a technique for accelerating the development of data mining models by generating them over a smaller dataset, it is important to note that the model generation is only one part of the overall data mining process [2]. Building a dataset for mining is a key preceding step, which is often time consuming when a sampling process is involved. This will have to be repeated many times, if new fields that should be brought into the dataset are identified during the mining activity.

**Business Users Not Supported by Sampling**

Many of the issues raised above, which relate to sampling, can in some part be mitigated against by very careful handling of the analysis. Indeed, a skilled, statistically trained analyst may review the above list of sampling issues, and consider immediate work-arounds to each problem. However, a key business issue for deploying data mining is to move away from the dependence on a small number of key statistical gurus in an organization, and enable the business managers to interact with their data using intelligent tools.

The alternative to reliance on highly skilled statistical analysts is to increase the ability of the business manager to engage with the data mining techniques directly.
Figure 2 – The alternative to reliance on highly skilled statistical analysts is to increase the ability of the business manager to engage with the data mining techniques directly.
1.3 In-Place Mining of Data with VLDB

Data mining applications tend to push the limits of the data warehouse. Problems with data integrity are often undiscovered until the Data Mining exercise.

A number of approaches to mining data in place in databases have been described by Freitas [3]. These can be summarized as:

♦ Parallelizing sequential code automatically—which is a sub-optimal because opportunities for speeding up the algorithm may be missed.

♦ Manually parallelizing existing algorithm—which may be impossible for some algorithmic approaches.

♦ Re-designing algorithms from scratch for parallel computing architectures—which involves considerable rework.

The computational requirements associated with data mining algorithms, when run over very large data sets, is high. Data sets commonly comprise millions of rows and several hundred variables or columns.

Data mining algorithms require multiple passes over the data set; each pass uses the information from the previous pass to target or narrow down the subsequent iterations. Each pass is potentially computationally intensive. For a telecommunications project, each run of a decision tree may require about 6,000 queries on a database consisting of many millions of records, each containing several hundred variables.

The application of data mining algorithms to very large data sets requires the algorithm itself to scale, and the computational environment to be of sufficient speed to provide the underlying processing power. There are two important components of scalability here

♦ increases in the number of rows within the database

♦ increases in the number of variables or columns.

Given careful choices in algorithms, parallel computing environments make it feasible to apply data mining algorithms to these full volumes of data.

There is no one, perfect data mining tool or technique. Frequently, one needs to mix and match tools to achieve all the dimensions of information that are required to answer one question. One of the real challenges is how to use multiple data mining tools on the same data set, without restructuring the data set. There are two viable
alternatives to this challenge. The first is to use a data mining tool kit, such as those provided by IBM, SGI or Thinking Machines. The second alternative is to select a group of tools that have a generic back end interface, such as an ODBC connection. We present here the use of the second alternative; the power of a database supercomputer coupled with MiningSTUDIO, which is a data mining workbench developed by integrating the ANGOSS KnowledgeSTUDIO tool [4] with extensions to handle very large datasets, using specific drivers for a WhiteCross computer.

The approach taken by the WhiteCross data mining tool, MiningSTUDIO, to generate decision trees, is one of a classic client/server tool. The client acts as a query generator, submitting SQL requests to the database server and interpreting the query result-sets. The algorithms employed are based on the information theory approaches of ID3/C4.5 [5,6] and IT-Rule [7]. A major advantage of the decision tree is that we can easily take the generated rules and use them proactively.

The WhiteCross computer has been designed for very fast online analytical processing, rather than transactional processing. The server has a shared nothing, massively parallel processing (MPP) architecture. The very tight integration of the database server hardware and software enables considerable performance improvements over conventional, portable database implementations. The server is easily connected to user applications that contain an ODBC (open database connectivity) back-end. [ODBC is a standard that was introduced by Microsoft in 1993. It provides an open interface, based on Structured Query Language (SQL). ODBC consists of a standard, universal library of functions for building database-independent applications.]

A typical installation for a WhiteCross server may have 100 processors and approximately 50Gb of RAM for interactive analysis of data. This data can be queried as though every single column was pre-indexed—which in turn enables ad hoc queries against billions of records to be executed in a few seconds. Using this capability, MiningSTUDIO's algorithms understand the data by repeatedly drilling into it with increasing levels of granularity, until a sufficiently accurate understanding is attained.

As part of the decision tree process, the user must create an objective or business goal. As MiningSTUDIO creates a decision tree, it finds records that match the goals, and then presents a breakdown of characteristics or profiles for the records. The business goals are expressed in terms of SQL clause fragments. A strong feature of this particular implementation is that multiple tables can be used to support both the SQL query, and later tree generation. The user doesn’t have to join the relevant tables prior to the generation of the decision tree. The SQL query calculates how many rows match that the goal.
An interesting point to note about the client/server implementation of data mining in WhiteCross In-Place Mining is that the time required to generate a decision tree is not significantly affected by the number of rows over which the tree is grown. However, the number of columns evaluated as possible splits at every node in the decision tree is a major factor. A user-settable option in WhiteCross’ MiningSTUDIO controls the number of variables considered at every split. Pruning the search space enables even the largest datasets, involving several hundreds of columns, to be reduced to a tractable problem, and executed within minutes.

There are a number of methods that can then be used to refine the model, so that the user can obtain a better result. The main methods include; changing the decision tree parameters, performing statistical attribute analysis so that the default categorisations can be modified, adding or deleting business attributes, refining the business goal, filtering the data set, and using relational joins.
1.4 Integration of Multiple Data Mining Algorithms

Powerfully generating decision trees from data residing in the database is a useful technique in its own right. In the MiningSTUDIO tool, it is integrated with additional exploration and interpretation tools, as well as a number of further algorithms that together form a powerful toolkit.

♦ **Data exploration.** An interactive charting tool that allows the user to flexibly graph and explore the distribution of variables across the full volume of data.

♦ **Interpreting results.** A wide range of graphical views of the decision tree show the variable distributions for each node’s sub-population; many text-based formats for rules such as English prose, Java or SQL code; analysis of the decision tree’s efficacy in terms of a lift or ROI curve and scoring the tree against validation data sets.

♦ **Clustering Algorithms.** These are unsupervised techniques (in that they are not specified a particular goal to search for, but rather uncover hidden structure) that can improve the performance of supervised techniques. They work well with categorical, numeric, and textual data.

♦ **Neural Network Algorithms.** These are supervised techniques when used to generate predictive models. A number of architectures and algorithms are included in the MiningSTUDIO tool.

- **MLN** – multi-layer perceptron with a single layer of hidden neurons, typically used to train a multi-dimensional data set with a continuous underlying model.

- **PNN** – probabilistic neural network using a Memory Based Reasoning technique to create a predictive model such that each training record may be considered as a neuron in the network.

- **RBF** – radial basis function predictive, an optimized form of the PNN which embeds clustering to predict based on cluster centers.

The MiningSTUDIO workbench also implements a three-tier client/server architecture for data mining. Users may choose which computer to use to execute the computationally intensive algorithms. The users’ own workstations can be used, or alternately a remote NT or Unix/Solaris server. DCOM is used as the protocol, or client/server middleware, to facilitate the communication between the client and the compute server.
Figure 3 – MiningSTUDIO's three tier client server architecture
1.5 Deployment of Data Mining Through Applications

The implementation of data mining against the full volume of data frees the business manager from wrestling with statistical issues relating to sampling. However, the increasing complexity of data mining, with workbenches such as MiningSTUDIO including multiple algorithms and considerable flexibility, works against the business manager. This can be overcome by providing simplified interfaces to data mining, which are specific to the needs of the business. Rather than describing data in technical terms and forcing the user to browse through SQL datasource catalogs, an application can simply present the user with the dataset for their specific needs.

This data mining application development is enabled through componentization of the data mining workbench. This use of components is common place in business application software. When you have a spreadsheet or graph inserted into a document, there are two principle components; the document and the embedded object, using COM as the underlying mechanism to facilitate communication.

Every major element of MiningSTUDIO is implemented as a component with both a visible and a hidden interface. The visible interface provides the user interface, for example, the decision tree browser from MiningSTUDIO could be hosted within a simple web page, if that is what a business application requires. The hidden interface enables complete embedding of the data mining technology within an application, so that the user needn’t be aware of its existence. This would be comparable to embedding bayesian algorithms to support the spell checking in word processors.

Data mining applications can be considered in terms of their intended life cycle. A distinction may be readily made between software developments that are intended to be long-lived—such as a Pay-Roll application, and those that are intended to have a short and limited life-span. Indeed, it is commonplace for software developers to undertake maintenance programming on systems where the code was written before they were even born. Within this context, it is illuminating to compare and contrast the two modes of development, and review the place of data mining applications.

Long-lived software is generally the result of a long design cycle, expending many, many man-years of effort, and an extensive period of testing. Such projects may fail, due to their complexity. Their costs may overrun, risking cancellation, or they may miss the business goals for which they were initiated, because the business has moved on at such a pace that the application is obsolete at the time of delivery.
In contrast, short lived software is generally the result of a very short development cycle—indeed ‘Rapid Prototyping’ is a common expression in this regard. The design is often rigid, and extension to the application may be very hard to undertake without a complete re-write.

MiningSTUDIO, and the data mining toolkit therein, supports both modes of development. The components with STUDIO have been crafted so that they can be incorporated within bullet-proof applications. Conversely, they integrate excellently with rapid application development environments, such as Visual Basic. It is conceivable that the result of a short data mining engagement may no longer be just a static hard copy report—it may be a data mining application that the user can interact with, and use to explore the relationships in the data for themselves. This potential for “throw-away” applications means that their development must be very fast and efficient. Two example data mining applications are described below.

**Example 1 – Enhancing Account Management Reports**

Consider a sales account manager working for a telecommunications company with a client base of small businesses. Before visiting or calling a client, the manager would want to review their past and present activity. This may well take the form of generating a standard report showing number of lines, total bill, average calls per day, or average calls per hour.

A data mining application could be used to enhance this report by interpreting the data, by analyzing what was distinctive about the client’s activity, in contrast to the overall base and to other companies in the same business sector. Time series analysis of phone activity would support predictive modeling, looking for future usage trends. This knowledge would support the manager in better serving the client’s telecommunication needs.

**Example 2 – Early Warning Systems for Churn**

Consider the same account manager, but this time his concern is to identify clients who might be about to leave for a competitor. The benefits of keeping a profitable customer are of course immediate, but the real value of a maintaining a high retention rate is only realized over time [8]. Customer attrition must be understood on an on-going basis. Many factors add up to make customer retention a long-term requirement and, perhaps, merit the effort of building a permanent churn application.
The cost of customer acquisition is high. Many companies lose money on first year customers. This can be calculated by adding up the costs of all customer acquisition, versus revenue earned from first year customers.

The cost of maintaining experienced customers is less. They understand the services, and are less likely to burden technical and customer assistance lines.

Existing customers are targets for cross-selling opportunities and a source of referrals, and therefore more profitable.

As new competitors enter the market, your own product offerings change and your competitor’s offerings change, so the reasons for attrition are also likely to change. This reinforces the need to build permanent churn applications.

More specifically, the problem is one of identifying valuable customers who have a high propensity to switch to a competitor – the italicized words indicating the key concepts that must be modeled in a predictive application.

Valuable customers – by estimating the potential future spend and future cost to service from historical transaction records for each customer, a measure of their future worth to the business may be derived.

Propensity to churn – or more specifically the probability that the customer will defect to a competitor within a specific time period.

This problem may be modeled by relating account information to call detail records, and using a mechanism to identify the date of churn from a sequence of calling patterns over successive time periods.
1.6 Deployment of Data Mining Through Predictive Models

The predictive models that result from data mining activity are corporate assets. They can be produced as part of an application, as described above, or produced by analysts using a Data Mining tool. However, knowledge management should ensure their effective deployment.

One method of using a predictive model is to undertake 'Bulk scoring'. This is the process of reading a set of records, and writing the likelihood of an outcome, (for example, response) for each record. Once this is obtained, the population can be sorted, and the best candidates can be included in the marketing list. This is undertaken using code generation—whereby the mining tool generates code to encapsulate the rules of the predictive model. Decision trees can be easily expressed in the form of code. MiningSTUDIO can generate a model in the form of a SQL Case statement. This allows the predictive model to be deployed directly into the database engine, thus solving scalability issues. This is often the most efficient way to deploy a model, and is relevant for both batch and real time model deployment.

Another alternative to bulk scoring that some application developers have embraced is to read the rules output, and convert it into another form. This approach is applicable to virtually all data mining tools that can generate rules. This leads to real time model deployment, whereas batch deployment is only effective for campaign management, mail shots, etcetera, that are not time critical. Other applications require models to be run interactively, because batch scoring can be cumbersome with huge data volumes. Imagine a call centre application that requires a predictive model to indicate which tele-service to pitch a customer. Batch scoring could be used to write the results of the model for each customer record. When the customer phones the center, the model results could be looked up. When the model changes, all the customers must be re-scored.

Clearly, it would be preferable to call the model on a real-time basis, for example, when the customer is on the phone. Using SQL Case rules is a natural way to deploy decision tree models on a real time environment.

An alternative to rule generation is to use the KnowledgeSERVER interface to call predictive models on a real time basis. This has the advantage of abstracting the model type, so that an analyst could switch between a neural network and a decision tree without any changes to the client code, by simply saving a new model. The following code-segment shows how a client application can open a project and call a predictive model.
Another distinct advantage of real time scoring versus batch scoring is the ability to use dynamic inputs, such as time of day or events, during the conversation.
1.7 Exploration Workshops for Knowledge Discovery

A final way in which a business might exploit the advanced data mining approaches discussed above is through the use of Exploration Workshops. These workshop-style sessions for knowledge discovery bring together business managers and their business problems, along with a powerful data exploration capability and data mining analysts to support them.

As a precursor to an exploration workshop, a preliminary brainstorming session would be undertaken to identify:

1. The scope of business problem.
2. Initial ideas as to how it might be addressed.
3. The necessary data to support the analysis.

Ideally, the value of addressing the business problem should be specified—this may be a quantitative assessment (for example, if we improve our fraud detection rate by 1% it will result in a $150m/year saving) or it may be more qualitative (for example, if we improve the market segmentation then it will assist in campaign design).

Following on from the initial brainstorming session, the necessary data for the analysis would be taken and loaded into the analytical environment. The data should be checked in two ways: verification and validation. Data verification checks that the data loaded is what was intended. The data experts who understand the source systems would be needed to clarify the meaning of any columns and provide any additional meta-data necessary to understand the data. Validation checks the data with respect to the business problem to provide assurance that it is suitable for the analysis intended. Data may be unsuitable because it is of poor quality, having many missing or erroneous values.

The Exploration Workshop sessions would then follow, each being a facilitated session of 2 to 4 hours, where one or more business managers engage with their problems and with analysts supporting their work, using the data and the data mining toolkits. A very important aspect is the interactivity—allowing business managers to hypothesize about possible solutions to their problems and using the support of the analysts study these interactively.

This close coupling of data exploration power—very large databases avoiding the complexities of sampling, data mining tools along with analytical skills—provides the business manager an ideal environment to very quickly explore and resolve business problems.
References


