# DATA WAREHOUSING AND DATA MINING: STEPPING FORWARD

**Adil Hussain** (BCS)
FAST-NU, Karachi Campus

**Ovais Ahmad Khan** (BCS)
FAST-NU, Karachi Campus

**Raza Ali** (BCS)
FAST-NU, Karachi Campus

## 1. Abstract

The emerging technologies of data warehousing, OLAP, and data mining have changed the way that organizations utilize their data. Data warehousing, OLAP, and data mining have created a new framework for organizing corporate data, delivering it to business end users, and providing algorithms for more powerful data analysis. These information technologies are defined and described, and approaches for integrating them are discussed. A generalized look at these emerging technologies has been given in this paper. Enterprises are accumulating piles of data, which is providing no practical or business knowledge but only increasing the cost of managing it. Making this data useful and productive is the basic goal of data warehousing and mining technologies. These technologies are about decision making through data analysis, forcasting and identifying data segments of importance. Few case studies elaborating the practical implications of these technologies are also part of this paper.

## 2. Introduction

In the early 1990s, William Inmon introduced a concept called a data warehouse to address many of the decision support needs of managers[1]. A data warehouse contains diverse data collected from across an enterprise and is integrated into a consistent format and useful for efficient querying and analysis. With the data warehouse, query execution does not need to involve data translation and communications with multiple remote sources, thus speeding up the process analysis process.

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions.

## 3. Foundations of Data Mining and Warehousing

The data warehouse came about when the need for data analysis and decision support was felt. Operational databases were already in use but were not very effective for decision support. Data collection, cleaning and organizing along with technologies such as OLAP for analysis of data were developed

and the idea of data warehouses came about.

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers. The software and hardware requirements also restricted the developments of data mining application. But later developments in database technology and high data processing systems paved the way for its introduction. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

| Evolutionary Step | Business Question | Enabling Technologies | Product Providers | Characteristics |
|---|---|---|---|---|
| Data Collection (1960s) | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retrospective, static data delivery |
| Data Access (1980s) | "What were unit sales in New England last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retrospective, dynamic data delivery at record level |
| Data Warehousing & Decision Support (1990s) | "What were unit sales in New England last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retrospective, dynamic data delivery at multiple levels |
| Data Mining (Emerging Today) | "What's likely to happen to Boston unit sales next month? Why?" | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Prospective, proactive information delivery |

Steps in the Evolution of Data Mining.

# 4. Data Warehousing

## 4.1. Defined

The primary concept of data warehousing is that the data stored for business analysis can most effectively be accessed by separating it from the data in the operational systems.

A data warehouse is a structured extensible environment designed for the analysis of non-volatile data, logically and physically transformed from multiple source applications to align with business structure, updated and maintained for a long time period, expressed in simple business terms, and summarized for quick analysis.

Typically, the data warehouse is maintained separately from the organization's operational databases. There are many reasons for doing this. The data warehouse supports OLAP, the functional and performance requirements of which are quite different from those of the OLTP applications traditionally supported by the operational databases.

| Operational | Data Warehouse |
|---|---|
| Current value of time. | Snapshot data. |
| Time horizon: 60-90 days. | Time horizon: 5-10 years. |
| Key may or may not have an element of time. | Key contains an element of time. |
| Data can be updated. | Once snapshot is made, data can't be updated |

## 4.2. Characteristics and Concepts

A data warehouse is a:

- subject-oriented,
- integrated,
- time-variant,
- nonvolatile

collection of data in support of management's decision making process[2]. Apart from these there are some other characteristics of data warehouse,

### 4.2.1. Subject Oriented

Subject oriented data management means that all data related to a subject are extracted from wherever they resides in the organization and brought together into the data warehouse. The data-driven, subject orientation is in contrast to the more classical process/functional orientation of applications, which older operational systems are organized around[2] This transformation of data leads to much more useful categorizations for analysis for a business decision-maker.

### 4.2.2. Integrated

An integrated approach ensures that the data are stored in a common data model that represents the business view of the data. Integrating data into one location and one data model is one of the main tasks of data warehousing.

### 4.2.3. Non Volatile

By non-volatile we mean that data in the warehouse does not change or get updated. In an operational database, records can be inserted, updated or deleted to represent the existing state of the world. This also ensures that the data in the warehouse will remain stable over a long period of time.

### 4.2.4. Time Variant

In the operational environment data is accurate as of the moment of access. In a data warehouse, time is an important element because it allows the end user to conduct trend analysis and historic comparisons.

## 4.3. Architecture

A data warehouse Architecture is a way of representing the overall structure of data, communication, processing and presentation that exists for end-user computing within the enterprise. For data warehousing, the architecture is a description of the elements and services of the warehouse, with details showing how the components will fit together and how the system will grow over time. The architecture is made up of a number of interconnected parts, which are described below:

*External Database Layer* – Operational systems process data to support critical operational needs. In order to do that, operational databases have been historically created to provide an efficient processing structure for a relatively small number of well-defined business transactions.

*Information Access Layer* – The Information Access layer of the Data Warehouse Architecture is the layer that the end-user deals with directly. In particular, it represents the tools that the end-user normally uses day to day.

*Data Access Layer* – The Data Access Layer of the Data Warehouse Architecture is involved with allowing the Information Access Layer to talk to the Operational Layer.

*Data Directory (Metadata) Layer* – In order to provide for universal data access, it is absolutely necessary to maintain some form of data directory or repository of metadata information. Metadata is the data about data within the enterprise.

*Process Management Layer* – The Process Management Layer is involved in scheduling the various tasks that must be accomplished to build and maintain the data warehouse and data directory information.

*Application Messaging Layer* – The Application Message Layer has to do with transporting information around the enterprise computing network. Application Messaging is also referred to as "middleware", but it can involve more than just networking protocols.

*Data Warehouse (Physical) Layer* – The (core) Data Warehouse is where the actual data used primarily for informational uses occurs. In some cases, one can think of the Data Warehouse simply as a logical or virtual view of data.

*Data Staging Layer* – The final component of the Data Warehouse Architecture is Data Staging. It includes all of the processes necessary to select, edit, summarize, combine and load data warehouse and information access data from operational and/or external databases.

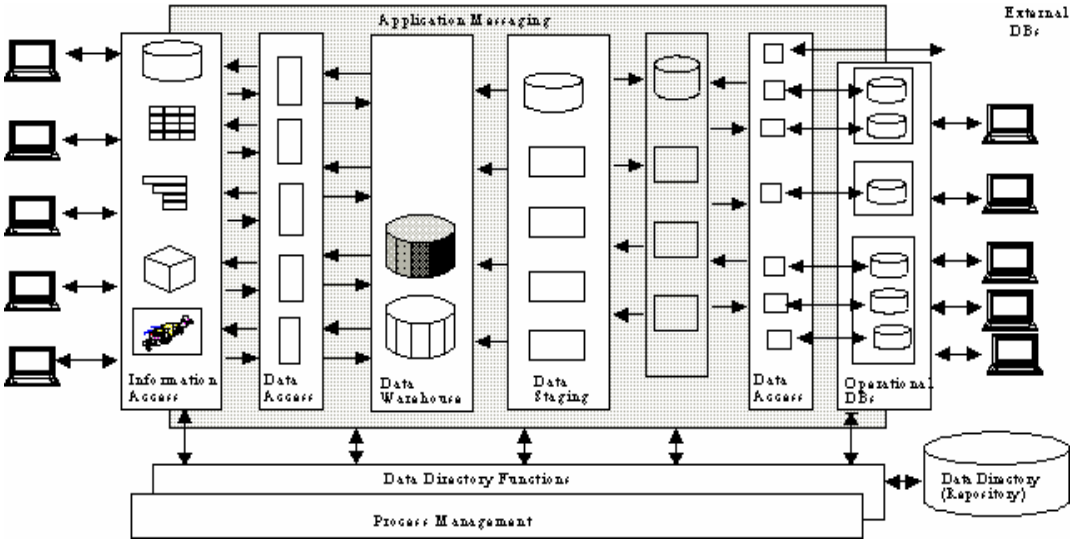
## 4.4. Development

Developing a good data warehouse requires careful planning, requirements definition, design, prototyping and implementation. The first and most important element is a planning process that determines what kind of data warehouse strategy the organization is going to start with.

Before developing a data warehouse, first answer these questions. Who is the audience? What is the scope? What type of data warehouse should we build? Different Types of Architectural Approaches are:

1. Operational storage versus using copies of operational data
2. Data warehouse only
3. Data marts only
4. Data warehouse and data marts
5. Platform and infrastructure partitioning
6. Two-tier client/server architecture
7. Three-tier client/server architecture

One way is to establish a "Virtual Data Warehouse" environment. A second strategy is simply to build a copy of the operational data from a single operational system The optimal data warehousing strategy is to select a user population based on value to the enterprise and do an analysis of their issues, questions and data access needs. Based on these needs, prototype data warehouses are built and populated so the end-users can experiment and modify their requirements. Once there is general agreement on the needs, then the data can be acquired from existing operational systems across the enterprise and/or from external data sources and loaded into the data warehouse.



A Data Warehouse

### 4.4.1. Designing Data Warehouses

Designing data warehouses is very different from designing traditional operational systems. For one thing, data warehouse users typically don't know nearly as much about their wants and needs as operational users. Second, designing a data warehouse often involves thinking in terms of much broader, and more difficult to define, business concepts than does designing an operational system. The ideal design strategy for a data warehouse is often outside-in as opposed to top-down.

## 4.5. Managing and Maintenance

The responsibility for administrating the data warehouse (or data marts) rests with the IS department in an organization. We can form (or use) a *data architect group*[3], whose duties are listed below:

- Monitoring Data Warehouse Operations
- Adding and Deleting Subject Areas
- Supporting End Users
- Maintaining the Metadata
- Updating the Warehouse Content
- Upgrading the Warehouse
- Updating the Data Model
- Capacity Planning
- Managing Security

## 4.6. Data Warehouse and Business

The ultimate goal of a data warehouse is to provide decision support for management. The characteristics described above help resolve many problems related to using operational data as a source for decision support.

Decision support systems (DSS) are systems that use models and data to solve managerial problems ranging in complexity from budget decisions using simple spreadsheets to optimal site location. “An EIS is used by senior managers to find problems; the DSS is used by the staff people to study them and alternatives”[4]. Although extremely useful, the EISs and DSSs often lacked a strong database component. Thus, the builders of these systems had to create their own databases. The vendors of DSS and EIS software were quick to pick up on the opportunity offered by data warehousing.

Along with data mining, it can be used to predict what the organization wants to know.

## 4.7. Advantages over operational databases

- *Performance* – warehouses are optimized for the decision support and are non-volatile, both of which increases the performance.
- *Data access* – Organizations usually maintain multiple databases. Warehouse combines the data from all these sources in a central location.
- *Data formats* – Warehouse contains summary data as well as time-based data, which help the decision-makers. Both are not there in the normal databases.
- *Data quality* – The data in the warehouse is clean, validated and properly aggregated. The warehouse provides "the single version of the truth." [5]
- *Accessibility* – Creating a separate physical location for storing the warehouse data insures that data is available anytime, even if the original sources are not available.
- *Easing Burden* – Giving business users a separate warehouse for analysis also eases the processing burden at the local data sources.

## 4.8. Problems and Issues

As with any design approach, there are trade-offs in the data warehousing approach that must be considered.

Having a separate warehouse also means that there must some systematic mechanism to detect changes in the data sources and to update the warehouse [6].

Data warehousing systems can complicate business processes significantly which have not been satisfied by the operational database.

Your warehouse users will develop conflicting business rules [7]. Possible business rules is so large that you will not be able to incorporate all rules.

Data warehousing systems can require a great deal of "maintenance".

Sometimes the cost to capture data, clean it up, and deliver it in a format and time frame that is useful for the end users is too much of a cost to bear.

Also, the business end users can only query data stored at the warehouse, so determining what this data is in advance may result in the users not being able to perform certain analyses.

Since the data in the warehouse is updated periodically, if the analytical needs of the user are for current information, the warehouse approach may not provide up-to-date information.

Data warehousing failure rates are between 10% and 90%. Thus, if your organization does not know how to manage risky projects, then data warehousing may not be for you[8].

## 4.9. Tools & Techniques

The data warehousing tools can be classified into front-end and back-end. Back-end tools include those meant for cleaning, loading etc. while front-end tools are usually meant for displaying, reporting. etc.

### 4.9.1. Cleaning

Since a data warehouse is used for decision-making, it is important that the data in the warehouse be correct. However, since large volumes of data from multiple sources are involved, there is a high probability of errors and anomalies in the data.. Therefore, tools that help to detect data anomalies and correct them can have a high payoff.

### 4.9.2. Load

After extracting, cleaning and transforming, data must be loaded into the warehouse. Additional preprocessing may still be required: checking integrity constraints, sorting, summarization, aggregation and other

such tasks. Typically, batch load utilities are used for this purpose.

### 4.9.3. Refresh

Refreshing a warehouse consists in propagating updates on source data to correspondingly update the base data and derived data stored in the warehouse. There are two sets of issues to consider: *when* to refresh, and *how* to refresh.

### 4.9.4. OLAP (on-line analytical processing)

Data warehousing and OLAP are essential elements of decision support, which has increasingly become a focus of the database industry. The data warehouse supports OLAP, the functional and performance requirements of which are quite different from those of the OLTP applications traditionally supported by the operational databases.

OLAP operations include *rollup* (increasing the level of aggregation) and *drill-down* (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, *slice and dice* (selection and projection), and *pivot* (re-orienting the multidimensional view of data).

OLAP and data warehousing are very much complementary. In order for the end-user to be able to conduct analysis with the data warehouse, there needs to be an interface. While the data warehouse stores and manages the analytical data, OLAP can be the strategic tool to conduct the actual analysis. It is used as a common methodology for providing the interface between the user and the data warehouse.

### 4.9.5. Data Mining

Data Mining is the next step in this evolution is that of data mining. Data Mining and its relation to data warehousing is discussed in the data mining section of this paper.

## 4.10. Data Marts

The failure of data warehousing to address the knowledge worker's culture and the practical technical difficulties associated with EDM development and warehouse maintenance prompted Forrester Research in 1991 to declare that data warehousing was dead. It had been replaced by what Forrester called "data marting"[9]. In particular, a data mart is a data warehouse that is created for a specific department within an organization[10]. Data marting has advantages over data warehousing [4 (gray, watson)]:

- The data marting model supports individual knowledge worker communities quite well.
- The cost of building a data mart is quite low as compared to a data warehouse solution.
- The lead-time to implementation is short.
- They are controlled locally.
- Contains less information, thus has a high response time.
- The business unit can itself build its own decision support systems without relying on a central IS department.

Users expect more rapid response from a data mart than from a data warehouse. Unfortunately, performance decreases as data marts grow in size over time and as the number of interconnected data marts increases. Multiple data marts also complicate database administration. Data marts have proven to be difficult to build quickly. Also, DM technology is unable to scale-up of the data marts to support increased numbers of users[11].

# 5 Data Mining

## Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of stored product data. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

**Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Typical examples of predictive problem are targeted marketing, predicting bankruptcy, default, identifying responsive population segments.

- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. Examples of pattern discovery are the analysis of sales data to identify related products, detecting fraudulent credit card transactions and anomalous data representing data entry errors.

### *Scope of Pattern Discovery*

Data mining tools are implemented on high performance parallel processing systems; they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyze huge quantities of data. Larger databases, in turn, yield improved predictions. Databases can be larger in both depth and breadth:

- **More columns**. Analysts does not have to limit the number of variables they examine, data mining allows users to explore the full depth of a database, without pre-selecting a subset of variables.
- **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

### *Data Mining Techniques*

The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

These techniques and capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

## How Data Mining Works

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. . For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. You'd like to concentrate on those prospects that have large amounts of long distance usage.

The goal in prospecting or forecasting is to make some calculated guesses about the information in the target areas based on the model that we build. For instance, a simple model for a telecommunications company might be:

98% of my customers who make more than $60,000/year spend more than $80/month on long distance.

With this model in hand new customers can be selectively targeted.

### Verifying Mining Result

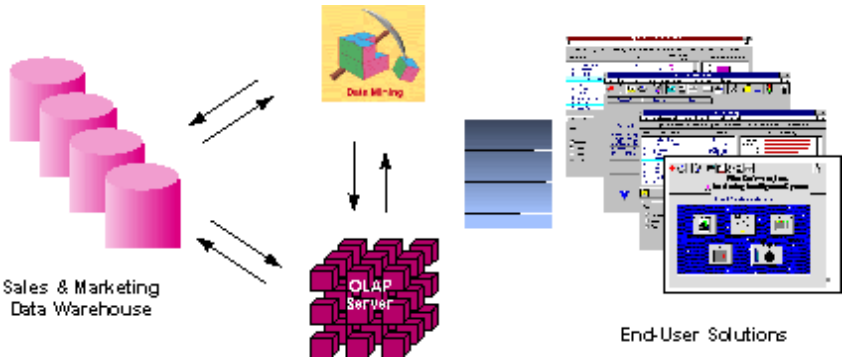
If someone told you that he had a model that could predict customer usage how would you know if he really had a good model? The first thing you might try would be to ask him to apply his model to your customer base - where you already knew the answer. With data mining, the best way to accomplish this is by setting aside some of your data in a vault to isolate it from the mining process. Once the mining is complete, the results can be tested against the data held in the vault to confirm the model's validity. If the model works, its observations should hold for the vaulted data

## Data Warehouses for Data Mining

### *Example architecture for Data Mining*

To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. The following Figure illustrates an architecture for advanced analysis in a large data warehouse.

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.




An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans.

## Problems and Issues

- Noisy data. Typing errors.
- Missing values. Incomplete and improper data.
- Static data. Data must be time dependent.
- Sparse data. Good results are extracted from detailed data.
- Dynamic data. Data taken at anytime should not be modified.
- Relevance. Irrelevant data can misdirect the process of KDD.
- Interestingness. Data of interest should go through the KDD process.
- Heterogeneity. All data sources must be combined and replicate names for one entity should be flattened.
- Algorithm efficiency. Inefficient algorithms and unbalanced data selection can bog down the systems.
- Size and complexity of data. Data complexity requires proportional processing capabilities.

## Learning and Knowledge Discovery process

### *Stages*

- Data pre-processing. Data is selected from operational databases

and stored in a separate database. Data is made uniform and all replicate names resolved.
- Data cleansing. Removal of duplicate records and inconsistent data.
- Data warehousing. Construction of a data warehouse from the data. Enrichment and coding is also performed.
- Data Mining Tools applied. Extraction of patterns from the pre-processed data.
- Reporting. Analysis of the results and application to new data. User can also direct the process and experiment with data.

## Applications of Data Mining

A wide range of companies has deployed successful applications of data mining. While early adopters of this technology have tended to be in information-intensive industries such as financial services and direct mail marketing, the technology is applicable to wider variety of bus inesses.

- Medicine - drug side effects, hospital cost analysis, genetic sequence analysis, epidemic prediction etc.
- Finance - stock market prediction, credit assessment, fraud detection etc.
- Marketing/sales - product analysis, buying patterns, sales prediction, target mailing, identifying ‘unusual behavior' etc.
- Knowledge Acquisition
- Scientific - superconductivity research, Natural disasters prediction etc.
- Engineering - automotive diagnostic expert systems, fault detection etc.

# 6 Case Studies

### 6.1. Verisign Payflow™ Fraud Screen.

**The Problem.** The anonymity of Internet commerce makes the incidence of fraud in e-commerce transactions higher than in in-person transactions - but without online fraud protection services, your e-business must bear the burden of "chargebacks" for the full value of any fraudulent, Web-based credit-card purchases. This results in the risk of losing customers, goods, and even your merchant account.

**The Solution**. To expand VeriSign's commitment to enable trusted commerce over the Internet for merchants, VeriSign worked with the world leader in credit card fraud protection to offer a service which would help merchants protect themselves from losses associated with fraud on the Internet. VeriSign's Payflow Fraud Screen provides merchants wit h intelligent fraud detection and risk management technology. Payflow Fraud Screen is based on the eFalcon fraud-scoring model from HNC Software, Inc.

Integrated with Payflow Payment Services, Payflow Fraud Screen enables merchants to complete authorization and fraud evaluation of Internet credit card purchases in a single transaction request.

#### The HNC eFalcon Fraud Scoring Model

HNC is a publicly held company (NASDAQ: HNCS) located in San Diego, CA. For 10 years, HNC has been protecting issuing banks from fraud using its Falcon model. HNC currently protects over 300 million cardholders, and is used by 9 of the top 10 U.S. issuing banks and 18 of the top 25 worldwide issuing banks. Using patented modeling and profiling

techniques, as well as large amounts of transaction data, HNC developed eFalcon as a fraud scoring model dedicated to Internet transactions. Payflow Fraud Screen allows you to obtain eFalcon scores through the Payflow platform.

**Payflow Fraud Screen Provides**
Payflow Fraud Screen works with Payflow Pro, a robust and fully customizable "real time" payment solution for e-commerce storefronts, to enable merchants to accept online payments while screening for the probability of fraud within those transactions.

When a customer visits the merchant's Web site and makes a purchase, the credit card and transaction data is passed from the merchant's storefront to the Payflow Pro client, which then securely passes the payment transaction data to VeriSign's payment servers for processing. VeriSign securely routes the transaction through the financial network to the appropriate banks, to verify if that credit card transaction is approved. After the credit-card processor has authorized the transaction, VeriSign sends the transaction in real time to the eFalcon fraud model to be scored. Once a score is derived, eFalcon sends the score and other results back to VeriSign, which in turn sends the merchant the score along with the credit-card processor's response. Payflow Fraud Screen allows merchants to receive fraud scores between 1 and 999, with a higher score indicating a higher likelihood that that transaction is fraudulent. Reason and exception codes accompany the score to help explain the scores. You can use the score to decide whether to fulfill or reject the transaction.

**The Payflow Fraud Screen Score**
The eFalcon model evaluates dozens of factors to generate the Payflow Fraud Screen score. eFalcon examines factors such as geographical consistency, velocity (frequency of purchase), general fraud patterns, typical behavior by the individual, and typical behavior of purchasers at the merchant's store. One of the most powerful aspects of the model is the ability to track purchasing behavior across a large number of merchants and banks participating in HNC's Fraud Data Consortium. Some examples of factors considered by the model include:

- E-mail address activity
- Ship-to/bill-to activity.
- Shipping method activity.
- Product purchasing pattern.
- Payment methods history.
- Work/home telephone number patterns.
- Hour of day.
- Purchasing velocity.
- Geographic location of the consumer.
- Typical purchasing patterns at the merchant's site.

Considering these and other factors, over 65 percent of a merchant's e-commerce fraud can typically be prevented with eFalcon, while impacting less than 5 percent of the merchant's e-commerce transactions. eFalcon's performance is continuously improving.

**6.2. NorthRidge Earthquake.**

The data collected during the Northridge, California earthquake occupied several warehouses, and ranged from magnetic media to bound copies of printed reports. Nautilus Systems personnel sorted, organized, and cataloged the materials. Document were scanned and converted to text. Data were organized chronologically and according to situation reports, raw data, agency data, and agency reports. For example, the Department of Transportation had information on highways, street structures, airport

structures, and related damage assessments.

Nautilus Systems applied its proprietary data mining techniques to extract and refine data. Geography was used to link related information, and text searches were used to group information tagged with specific names (e.g., Oakland Bay Bridge, San Mateo, Marina). The refined data were further analyzed to detect patterns, trends, associations and factors not readily apparent. At that time, there was not a seismographic timeline, but it was possible to map the disaster track to analyze the migration of damage based upon geographic location. Many types of analyses were done. For example, the severity of damage was analyzed according to type of physical structure, pre- versus post- 1970 earthquake building codes, and off track versus on track damage. It was clear that the earthquake building codes limited the degree of damage.

Nautilus Systems also looked at the data coming into the command and control center. The volume of data was so great that a lot was filtered out before it got to the decision support level. This demonstrated the need for a management system to build intermediate decision blocks and communicate the information where it was needed. Much of the information needed was also geographic in nature. There was no ability to generate accurate maps for response personnel, both route maps including blocked streets and maps defining disaster boundaries. There were no interoperable communications between local police, the fire department, utility companies, and the disaster field office. There were also no predefined rules of engagement between FEMA and local resources, resulting in delayed response (including such critical areas as firefighting)

**Benefits**

Nautilus Systems identified recurring data elements, data relationships and metadata, and assisted in the construction of the Emergency Information Management System (EIMS). The EIMS facilitates rapid building and maintenance of disaster operations plans, and provides consistent, integrated command (decision support), control (logistics management), and communication (information dissemination) throughout all phases of disaster management. Its remote GIS capability provides the ability to support multiple disasters with a central GIS team, conserving scarce resources.

# 7 Conclusion

Data warehousing is a science that continues to evolve. After the rapid acceptance of data warehousing systems during past three years, there will continue to be many more enhancements and adjustments to the data warehousing system model. A flexible enterprise data warehouse strategy can yield significant benefits for a long period.

Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough.

A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining can make this leap possible.

# 8 Glossary

| | |
|---|---|
| **Analytical model** | A structure and process for analyzing a dataset. For example, a decision tree is a model for the classification of a dataset. |
| **Anomalous data** | Data that result from errors (for example, data entry keying errors) or that represents unusual events. Anomalous data should be examined carefully because it may carry important information. |
| **Artificial Neural Networks** | Non-linear predictive models that learn through training and resemble biological neural networks in structure. |
| **CART** | Classification and Regression Trees. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dat<br>aset to predict which records will have a given outcome. Segments a dataset by creating 2-way splits. Requires less data preparation than CHAID. |
| **CHAID** | Chi Square Automatic Interaction Detection. A decision tree technique used for classification of a dataset. Provides a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. Segments a dataset by using chi square tests to create multi-way splits. Preceded, and requires more data preparation than, CART. |
| **Classification** | The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to specific variable(s) you are trying to predict. For example, a typical classification problem is to divide a database of companies into groups that are as homogeneous as possible with respect to a creditworthiness variable with values "Good" and "Bad." |
| **Clustering** | The process of dividing a dataset into mutually exclusive groups such that the members of each group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables. |
| **Data cleansing** | The process of ensuring that all values in a dataset are consistent and correctly recorded. |
| **Data mining** | The extraction of hidden predictive information from large databases. |
| **Data navigation** | The process of viewing different dimensions, slices, and levels of detail of a multidimensional database. See OLAP. |
| **Data visualization** | The visual interpretation of complex relationships in multidimensional data. |
| **Data warehouse** | A system for storing and delivering massive quantities of data. |
| **Decision tree** | A tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset. See CART and CHAID. |
| **Dimension** | In a flat or relational database, each field in a record represents a dimension. In a multidimensional database, a dimension is a set of similar entities; for example, a multidimensional sales database might include the dimensions Product, Time, and City. |
| **Exploratory data analysis** | The use of graphical and descriptive statistical techniques to learn about the structure of a dataset. |
| **Genetic algorithms** | Optimization techniques that use processes such as genet ic combination, mutation, and natural selection in a design based on the concepts of natural evolution. |
| **Linear model** | An analytical model that assumes linear relationships in the coefficients of the variables being studied. |
| **Linear regression** | A statistical technique used to find the best-fitting linear relationship between a target (dependent) variable and its predictors (independent variables). |
| **Logistic regression** | A linear regression that predicts the proportions of a categorical target variable, such as type of customer, in a population. |
| **Multidimensional** | A database designed for on-line analytical processing. Structured as a |

| | |
|---|---|
| **database** | multidimensional hypercube with one axis per dimension. |
| **Multiprocessor computer** | A computer that includes multiple processors connected by a network. See parallel processing. |
| **Nearest neighbor** | A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called a k-nearest neighbor technique. |
| **Non-linear model** | An analytical model that does not assume linear relationships in the coefficients of the variables being studied. |
| **OLAP** | On-line analytical processing. Refers to array-oriented database applications that allow users to view, navigate through, manipulate, and analyze multidimensional databases. |
| **Outlier** | A data item whose value falls outside the bounds enclosing most of the other corresponding values in the sample. May indicate anomalous data. Should be examined carefully; may carry important information. |
| **Parallel processing** | The coordinated use of multiple processors to perform computational tasks. Parallel processing can occur on a multiprocessor computer or on a network of workstations or PCs. |
| **Predictive model** | A structure and process for predicting the values of specified variables in a dataset. |
| **Prospective data analysis** | Data analysis that predicts future trends, behaviors, or events based on historical data. |
| **RAID** | Redundant Array of Inexpensive Disks. A technology for the efficient parallel storage of data for high-performance computer systems. |
| **Retrospective data analysis** | Data analysis that provides insights into trends, behaviors, or events that have already occurred. |
| **Rule induction** | The extraction of useful if-then rules from data based on statistical significance. |
| **SMP** | Symmetric multiprocessor. A type of multiprocessor computer in which memory is shared among the processors. |
| **Terabyte** | One trillion bytes. |
| **Time series analysis** | The analysis of a sequence of measurements made at specified time intervals. Time is usually the dominating dimension of the data. |

# 9 References


[1] Pine Cone Systems, http://www.pine-cone.com/, 1997

[2] Inmon, W.H., "What is a Data Warehouse?" Prism Solutions, Inc, *http://www.cait.wustl.edu/cait/papers/prism/vol1_no1/*, 1995.

[3] Inmon, W.H. and Hackathorn, R.D., "Using the Data Warehouse" *New York: Wiley*, 1994

[4] Rocart, J.F. and Delong, D.W., "Executive Support Systems: The Emergence of Top Management Computer Use" *Homewood, IL: Dow Jones Irwin*, 1988

[5] Gray, P. and Watson, H.J., "Decision Support in the Data Warehouse" *Prentice Hall*, 1998

[6] Widom, J., "Research Problems in Data Warehousing" *Proceedings of the 4th International Conference on Information and Knowledge Management (CIKM)*, November 1995.

[7] Greenfield, L., "The Case Against Data Warehousing" LGI Systems, Inc, *http://www.dwinfocenter.org/gotchas.html*, June 2001.

[8] Greenfield, L., "Data Warehousing Gotchas" LGI Systems, Inc, *http://www.dwinfocenter.org/against.html*, June 2001.

[9] Demarest, M., "Building The Data Mart" *DBMS Magazine*, July 1994

[10] Wiener, J.L., "What is data warehousing and what is Stanford doing about it?" *An Overview talk given in the Stanford DB Seminar series*, Fall 1997.

[11] Alur, N., "The Enterprise Data Warehouse and Data Mart Debate" *InfoDB*, November 1996

[12] Pieter Adriaans, Dolf Zantinge, "*Data Mining*", Addison-Wesley, 1999.


## Other Hyperlinks

An Introduction to Data Mining
http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm

Data Mining and Customer Relationships
http://www3.shore.net/~kht/text/whexcerpt/whexcerpt.htm

From Data Mining to Database Marketing
http://www3.shore.net/~kht/text/wp9502/wp9502.htm