Data Warehousing Concepts

In Support Of

Data Mining Processes

Created By:
Marco Casale
Will DiGrazio
Doug Ramsey
Joar Rodriguez

Week 10 – Monday August 11th, 2003

TABLE OF CONTENTS

ABSTRACT

Data warehousing is a relatively new data management process receiving widespread acceptance in business and education. The benefits of analyzing enterprise wide, historical data in order to incorporate effective, cost saving decision support methods have fueled the recent expansion. However, developing a data warehouse is not an easy task. There are many aspects related to the successful implementation of a data warehouse. The data warehouse cannot be designed in one step and the implementation process is complex, dynamic, and subjective. Before any data warehouse project begins, the responsible personnel should be familiar with the basic processes, terminology, and technologies. This paper will serve as an effective primer to support such an endeavor as the topics of metadata, data marts, analytical processing, data mining, as well as a general overview of the data warehouse lifecycle model, will be presented in detail.

DATA WAREHOUSING OVERVIEW

The data warehouse lifecycle model depicts the process of an organization's attempt to transform operational data into information for the purpose of generating business knowledge. Data can be defined as the representation of facts regarding concepts or objects. Information is dependent on accurate data, clear data definitions, and an understandable presentation regarding the format represented to someone seeking knowledge. Thus, the following formula can be used to represent information: Information = f(Data + Definition + Presentation). (English, 1999) Knowledge, however, is more than just information. The individual seeking knowledge must understand the significance of the information within its applicable context. For

comparison, the following formula can be used to represent knowledge: Knowldege = f(People + Information + Significance). (English, 1999) It is the application of knowledge, which lies at the heart of every business decision. Thus, an organization needs a vehicle that can transform enterprise data into applicable knowledge and the data warehousing process is such a medium.

Data Warehouse Model

Figure 1 represents an overview of the most common data warehousing environment model. The model represents the initial stages where the data from disparate operational transaction processing systems are extracted and transported to the operational data store (ODS).

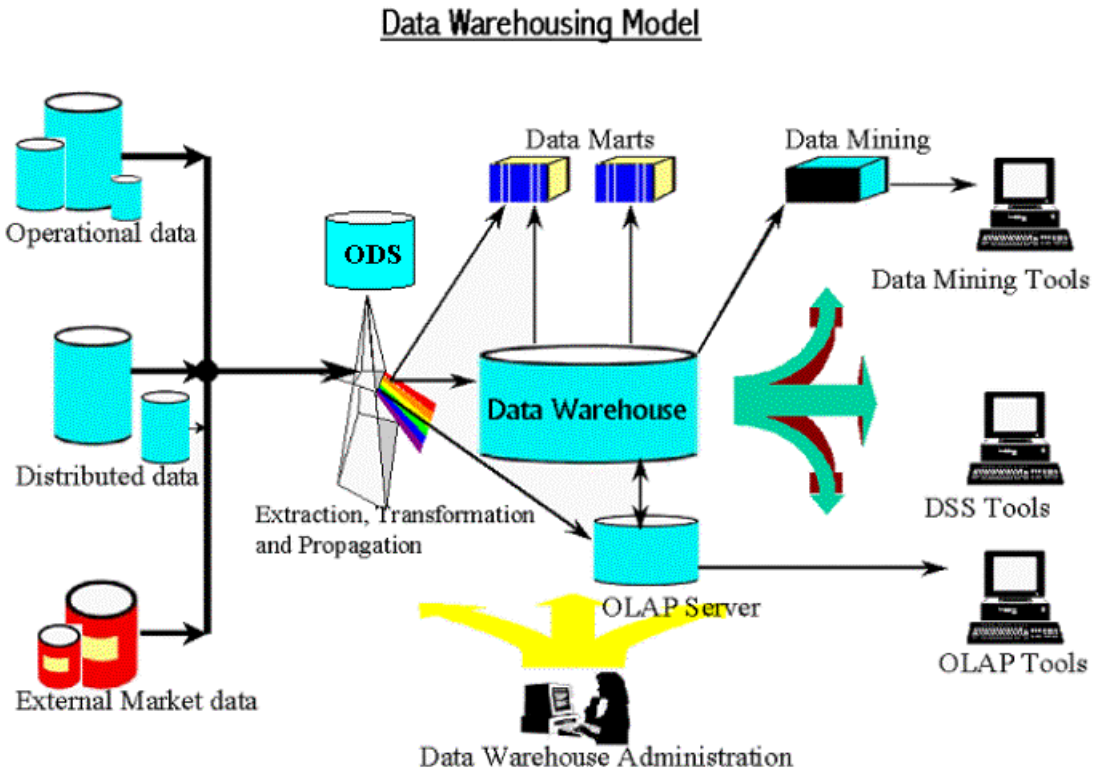


Figure 1 – Data Warehouse Environment Model

The need for this step is derived from the fact that the data within operational systems exist in real time. Since one important goal of the data warehousing process is to allow decision support and data mining tools access to the underlying information, the idea of querying operational systems would not only represent a performance nightmare but it would be unrealistic to directly gain knowledge from the data within operational systems. Therefore, in order to generate the information systems that will allow for such knowledge acquisitions, the data must be extracted from the operational systems. Once the data has been loaded into the ODS, the data is cleansed, formatted, and transformed from its original operational status into a configuration more acceptable to the environment of the data warehouse. Consequently, after the data has been prepared, it is loaded into the de-normalized schema of the data warehouse and resides there in a fine grain level of detail. The logical design of a data warehouse is usually composed of the star schema. "A star schema is a simple database design (particularly suited to ad-hoc queries) in which dimensional data (describing how data are commonly aggregated) are separated from fact or event data (describing individual business transactions)." (Hoffer, Prescott & McFadden, 2002, 421) "Often up to 80 percent of the work in building a data warehouse is devoted to the extraction/transformation/load (ETL) process: locating the data; writing programs to extract, filter, and cleanse the data; transforming it into common encoding scheme; and loading it into the data warehouse." (Hobbs, Hillson & Lawande, 2003, 6)

Now that the transformed data resides in the underlying fact and dimension tables of the data warehouse, the next step entails deriving the necessary aggregates that will define the associated data marts. "An aggregate is simply the rollup of an existing fact

table along one of its dimensions, which more often than not is time." (Scalzo, 2003, 147) For example, if a fact table holds daily automobile fatalities, then location-based aggregates might be town, county, and state. Thus, since the overall purpose of a data warehouse is to support the online analytical processing (OLAP) capabilities associated with data mining, the data in the warehouse must be propagated, as referred to in Figure 1, into information residing within data marts. Data marts serve to hold specialized, codified, dimensional, aggregated data roll-ups in support of the OLAP tools that will be used to discover business relationships and organization insights. However, slicing a cube does not always yield the desired knowledge. As a result, there are times when exploratory data mining processes must be initiated against the data warehouse itself, as opposed to a specialized multidimensional array. This drill down technique can also be accomplished with OLAP tools, specialized data mining software or even customized ad-hoc programmatic queries running against the finer level of detail. The OLAP server in Figure 1, processes all of these requests and coordinates the communication exchange between the warehouse or marts and the end user's tools.

Standing distinctly from OLAP and data mining tools, decision support software represents a unique implementation of applied knowledge. Data mining and OLAP processes serve to answer a question generated by an end user by digging through the data and returning a result that will elicit an insightful response. Afterwards the end user can then act on the newfound knowledge for the betterment of the organization. However, decision support software strives to incorporate this final step by automating the application of knowledge into reality. For example, a divisional manager may want to know when the average inventory for the east coast has dropped below 5% in order to

renegotiate a supply contract. Every week the manager runs a report from the company website using underlying OLAP tools connected to the east coast inventory data mart. Instead of this timely investigation, decision support software aims to automatically notify this manager of the under 5% inventory event. It can also be extended further by triggering the forwarding of a legal document to the current inventory supply contractor indicating that their services have failed to meet the conditions of the original contract and reparations will be in order. Thus, the successful data warehouse implementation will save time and money. Nevertheless, to get there, project stakeholders must understand the basic data warehousing concepts and communicate effectively with those that will build it. A firm knowledge of the data warehouse lifecycle model will permit any business person the ability to communicate effectively with the data warehouse architects that will allow for such OLAP, data mining, and decision support tools the capacity to increase the organization's return of investment.

Purpose & Benefits

A data warehouse can be an effective tool in resolving business related problems. Unsuccessful marketing, lost revenue, high employee turnover rates, and low customer satisfaction levels are all but a few examples of issues that can be identified, correlated, and strategically planned through the data warehousing process. By setting short-term and long-term data warehouse objectives the realization of an effective data warehouse implementation will be evident throughout the organization. Short-term objectives are those you can realize with every data warehouse iteration, which provide immediate benefits to the users, and long-term objectives are achieved over time. (Adelman & Terpeluk-Moss, 2000) The integration of data from multiple sources, improvement of

data quality, minimization of inconsistent reports, improved capability for data sharing, and the merging of historical data with current data are some of the short-term objectives will benefit the organization with improved information management efficiency. "Examples of long-term objectives are reconciling different views of the same data, providing a consolidated picture of the enterprise data, and creating a virtual one-stop shopping data environment." (Adelman & Terpeluk-Moss, 2000, 50) Through the achievement of these goals, the enterprise will surely find itself in a better position than before the existence of an enterprise data warehouse.

Metadata Defined

What is the meaning of metadata? The simple definition of metadata is "data about data". The problem with this definition is that it is generic. Metadata is extremely complex. Unfortunately, there is little theoretical understanding of metadata. Why do we need metadata? Does a corporation need to understand the meaning and use of data? Metadata managed properly can answer such questions. In the past decade, a new breed of applications known as data warehouses and decision support systems have opened a doorway into the information locked within the bits and bytes of data stored within a corporation's data center. How does metadata complement a data warehouse? All of the above questions deserve further investigation.

What is the meaning of Metadata?

What does 'data about data' really mean? "Literally, metadata can be taken to mean any form or description of data other than the actual data itself." (Inmon, 2001 (b), 3) The management of data has been around for a very long time. Metadata is the shadow

that keeps one step behind the progress of data itself. For example, if a report identified a value of 128, what information can be ascertained? However, if the report stated that the gross earnings for the month of June were 128 million dollars, it would be extremely useful. Metadata helps to turn data into information. According to Bill Inmon, the father of data warehousing, "It is the context supplied by metadata that gives understanding to data." (Inmon, 2001 (a), 2) Therefore, data and metadata go hand in hand. Data is meaningless if not accompanied by its vital partner, metadata. In addition, it is helpful to understand the meaning of metadata through an analogy. For example, "in order for an organization to become a world class information processing organization there is a need for a card catalog. In the world of information systems that card catalog is called metadata." (Inmon, 1997, 18) In summary, metadata provides an information system with the ability to quickly search for a variable that otherwise would seem impossible.

Why do we need Metadata?

Metadata has been a very necessary but unpopular goal. It is typically seen as an enormous burden. Inmon states that "The importance of metadata is inversely proportional to its commercial success." (Inmon, 2001 (a), 2) The need for metadata arose as a result to the proliferation of applications. Once an organization had several applications to integrate, it became difficult to manage the integration between applications. "Every application had its own interpretation of what information should be. It was like Elvis singing rock, Garth Brooks singing country, James Brown singing soul, and Lawrence Welk on the accordion, all at once. Combined together, the result was simply a bunch of noise when all played at the same time." (Inmon, 2001 (a), 3)

How was this problem resolved? Metadata was first tracked in the form of a data dictionary. Unfortunately, the data dictionary was difficult and expensive to maintain. The following figure 1M displays the transfer of metadata from several applications into a data dictionary.

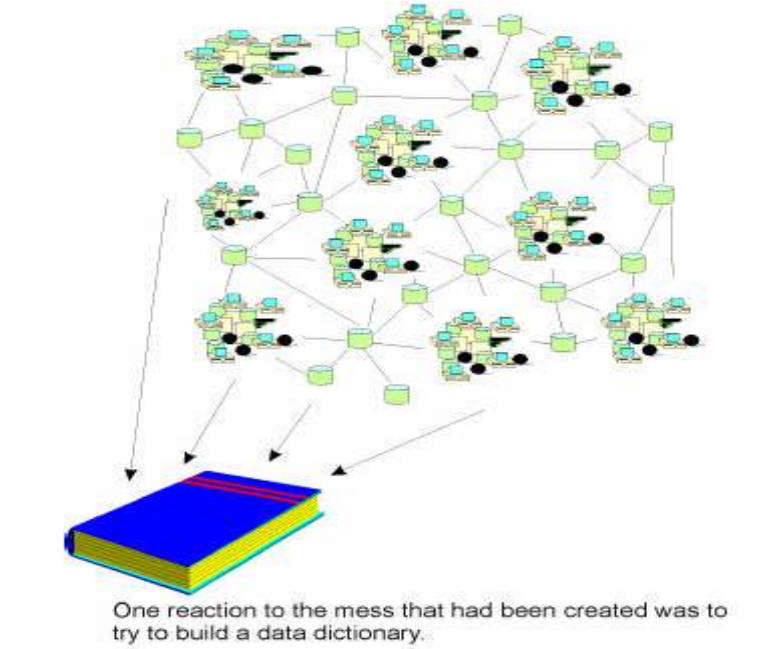


Figure 1M Metadata as the data dictionary (Inmon, 2001 (a), 7)

Metadata then progressed into a data model or entity relationship diagram. The data model was useful because it was metadata about all of the entities and relationships within an organization. It depicted how the applications integrated with one another from a relational point of view. The model could easily be deciphered and shared with developers and analysts. The following figure 2M displays how application metadata was maintained within a data model.

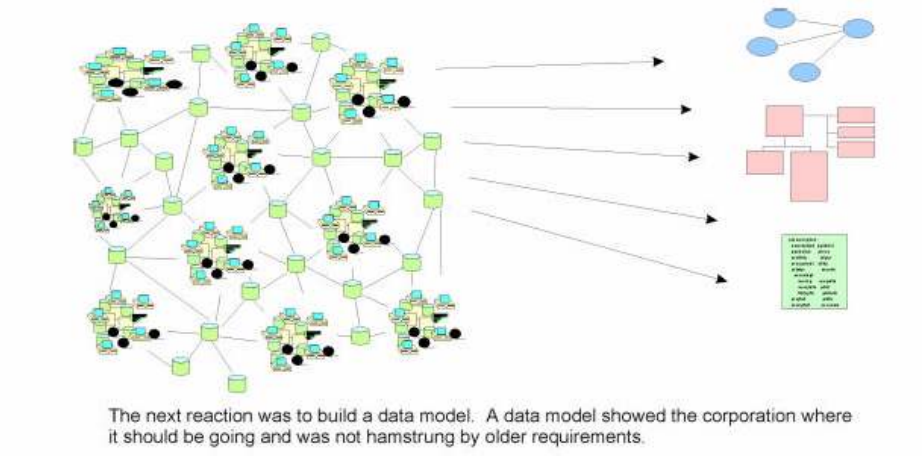


Figure 2M Metadata as a Data Model (Inmon, 2001 (a), 9)

However, there was still a great deal more about metadata that was not captured in a data model. Therefore, the next step was to develop a central data repository that housed metadata. The concept behind a central metadata repository was to extract all of the metadata from each application and store them into one central database. This concept was similar to the idea of the data dictionary but different as well. The central metadata repository was automated. The burden of loading the central metadata repository was reduced and less costly. Extracts of metadata could be siphoned off from the applications residing near by. The central metadata repository also recognized that there are other types of metadata, such as data models, that need to be stored in the metadata repository. The following figure depicts the central metadata repository.

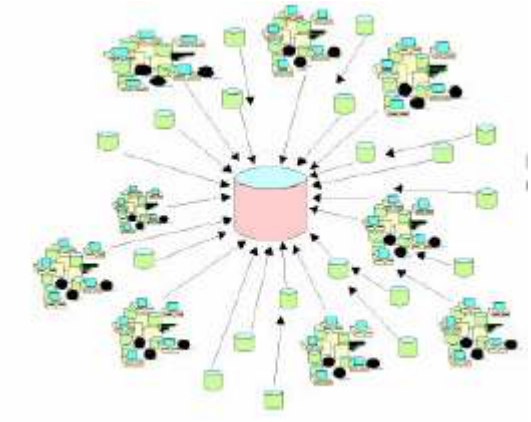


Figure 3M Metadata as a Centralized Metadata Repository (Inmon, 2001 (a), 11)

There were several problem encumbered by this approach. It was still costly to develop and maintain. In addition, it was also difficult to gain sponsorship for such a project. The metadata repository was considered a passive technology. It was challenging to demonstrate a tangible result from such an application.

Today's corporate information infrastructure presents new challenges. Currently, a corporate environment contains a distributed architecture. There are many different flows and transformations of data along the corporate intranet and internet. One approach to maintaining metadata in a distributed environment is the use of web services technology and a data format known as RDF. "Resource Description Framework, as its name implies, is a framework for describing and interchanging metadata." (Bray, 2001, 2). The RDF data format standard holds promise for how metadata can be transmitted over the Internet. Web sites would be able to house metadata repositories and share vocabularies amongst other metadata repositories. In summary, metadata could be used to help improve the usefulness of the Internet as an 'easily' searchable library.

How can metadata complement a Data Warehouse?

The purpose of the data warehouse is to bring together all of the operational systems within an organization and contain their data in one location. The data warehouse becomes a central repository for business analysis. Why do we need metadata if we already have a data warehouse? Metadata plays a crucial role for the organization in a data warehouse. "In order to achieve harmony and unity across the different components of the architected environment, there must be a well-defined and disciplined approach to metadata." (Inmon, 1997, 1)

What is it like for an end user to mine a data warehouse? If there is no metadata available, the end user must poke and probe the data warehouse to find out what data is available. This process can waste considerable time. There is no guarantee that the end user will find the right data or correctly interpret the data encountered. However, metadata comes to the rescue. The end user can quickly go to the necessary data or determine why the data is not there. Metadata then acts like an index to the contents of the data warehouse. It sits above the warehouse and keeps track of what is where in the warehouse. Typically, items the metadata repository tracks are as follows:

- Structure of data as known to the programmer
- Structure of data as known to the DSS analyst
- Source data feeding the data warehouse
- Transformation of data as it passes into the data warehouse
- Data model
- Data warehouse
- History of extracts

(Inmon, 2002)

The data warehouse architect must first consider how to support and design the metadata infrastructure. The metadata should be both shareable amongst the different components of the data warehouse. In contrast, the metadata should also be completely autonomous for each individual component. For example, a decision support analyst may have a report extracted from a warehouse on their local computer. The analyst develops several formulas on the local computer to analyze the data. The formulas are considered private metadata. However, the metadata found in the warehouse that describes the data elements is public or sharable.

Once the metadata infrastructure is in place, the decision support analyst can develop queries into the data warehouse by first querying the metadata itself. This is extremely powerful. The following is a list of example questions that can be answered by querying metadata.

- Where is the data element - QTYONHAND - found?
- What is the source of EMPLTAX?
- How has BALONHAND been converted from its origination in the operational environment?
- If a change occurs in CURRBAL, what source systems are affected?
- There is a data element in the source system called XVT-235J. If there is a change made in the source system, what data warehouse data elements are affected?
- The business person calls it CURRENT BALANCE. What does the technician call it?
- The finance department calls it RECEIVABLE. Is it the same thing in the marketing department?

(Inmon, 1997, 22)

What are the different types of Metadata used in a Data Warehouse?

Metadata can be found in many forms and types. There are a few specific metadata types used within a data warehouse. For example, data elements must be mapped from

operational systems into the data warehouse. The process of mapping requires the need for mapping metadata. Some examples of mapping metadata are:

- mapping from one attribute to another
- conversions
- changes in naming conventions
- changes in physical characteristics of data
- filtering of data, etc

Figure 4M shows the metadata captured during mapping transformations.

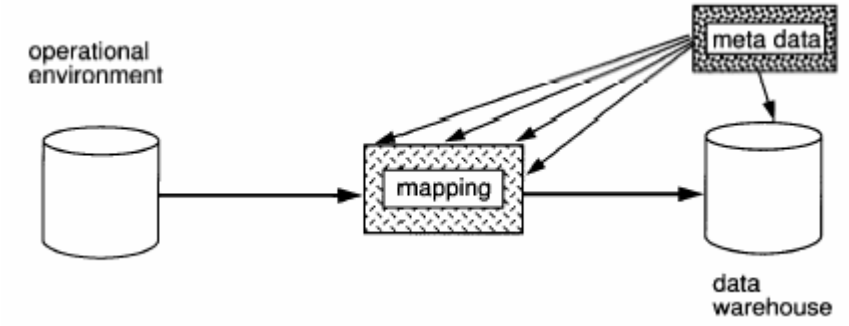


Figure 4M Metadata is used in mapping transformations (Inmon, 2002, 4)

Another important type of metadata is versioning metadata. This type of metadata reflects changes in metadata over time. A version of the metadata is accomplished by storing an effective from date as an attribute to the metadata. According to Bill Inmon, the reason why versioning is so important is that when the DSS analyst wants to interrogate a calculation or report made in the past, the versioned metadata allows the DSS analyst to understand what data was and where it came from as it entered the warehouse. Without versioning, the only meaningful data a data warehouse environment has is data written for and managed under the most current definition and structure of data. (Inmon, 2002)

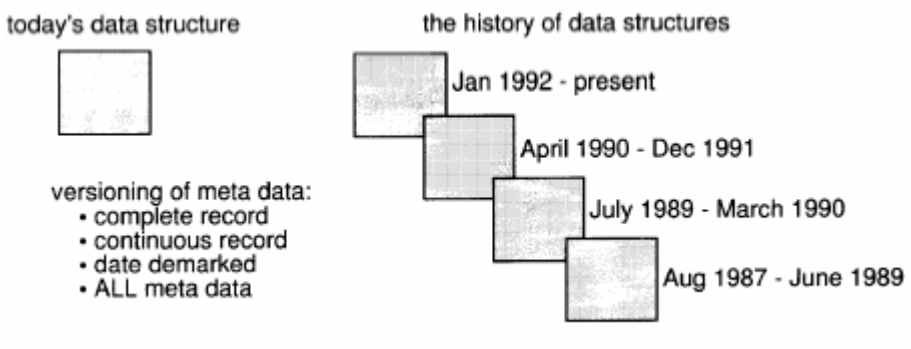


Figure 5M Metadata is used in versioning of data (Inmon, 2002, 6)

Finally, there are metadata types known as metadata components that complement the data warehouse. Basic components describe the physical tables, indexes and attributes within the data warehouse. Extract history components describe the movement and history of data from the operational systems into the data warehouse. This is extremely useful for the DSS analyst who wants to know when the data was last refreshed. Relationship artifact components store the data models and entity relationship diagrams regarding the data warehouse within the metadata repository. User access components describe how the data warehouse is being used for auditing purposes. Volumetric components provide metadata about the growth of tables and indexes within the data warehouse. Summary / calculation data between levels of the data warehouse is metadata that helps the DSS analyst understand how the granular data was aggregated. (Inmon, 2002) In short, there are many forms and types of metadata applicable to the data warehouse environment. It is apparent that metadata plays a significant role in the data warehouse architecture. The decision support analyst relies upon the quality and structure of metadata found within the data warehouse. Metadata provides the context to understand data. Metadata has lurked in the shadows while data has lived in the spotlight. It appears that metadata's moment of fame has arrived.

OLTP vs. OLAP

Transaction processing and decision-making are considerably different processes. As a result, the requirements of systems supporting these processes are also significantly different. Experience has shown that systems that support day-to-day operation, referred as *Online Transaction Processing* (OLTP), do not provide the flexibility and performance required to support analytical processing. Other types of systems, *Online Analytical Processing* (OLAP), are needed to rearrange the data in OLTP systems to optimize it for querying and to provide the required analysis capabilities.

In a data warehouse environment, OLTP systems usually are the main source of internal business data that is periodically loaded into the warehouse. On the other hand, OLAP systems form part of the presentation layer of the warehouse. Unlike OLTP systems, OLAP systems are characterized by "dynamic multidimensional analysis of consolidated enterprise data that supports end user analytical and navigational activities" (Jarke, Lenzerini, Vassiliou & Vassiliadis, 2003, 87-88). The following table, adapted from Meyer and Cannon (1998), presents other key differences between typical OLTP and OLAP environments:

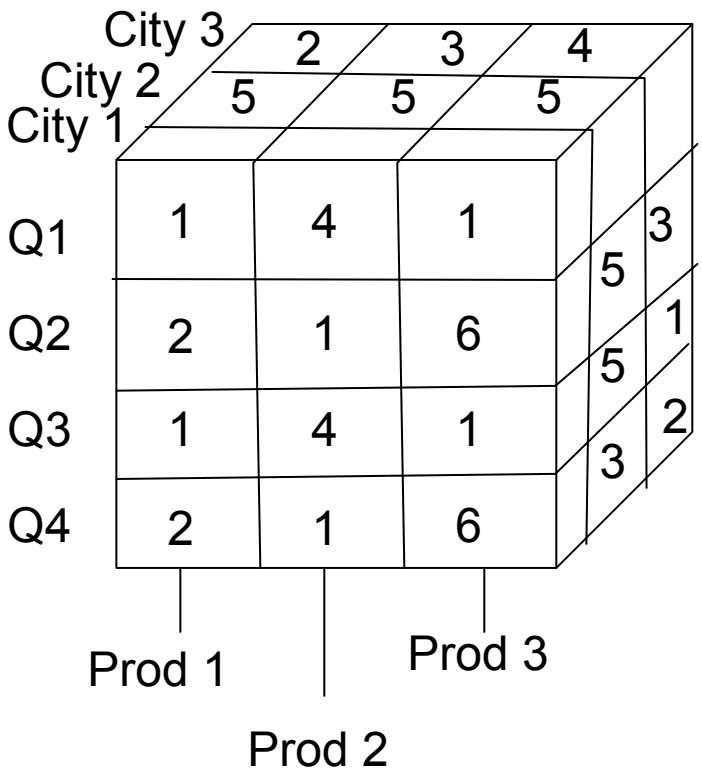
System Characteristics	OLTP	OLAP
Type of User Interaction	Transactions only	Throughout the entire database
Amount of Affected data	Individual records	Groups of Information
Response Time	Seconds	Seconds to minutes
Machine Utilization	Consistent	Dynamic

Data characteristics	Detailed data	Summary Data
Access to data	In a predefined way	Any way the user wants
Overall System Priority	Optimized for performance and availability	Optimized for flexible end-user interaction
Database configuration priority	Transaction updates	Queries
Database optimization priority	Bulk Transaction	Analysis
Maintenance required	Updated frequently	Minimal updates

Dimensional Models

As mentioned above, OLAP applications are based on the multidimensional arrangement of data; this arrangement is called a **data cube**. Data cubes are characterized by dimensions and facts. Dimensions are any aspect of the business for which the organization wants to keep records. Facts are the numerical measures that are associated to the dimensions. Unlike the geometric structures, data cubes are not limited to three dimensions; they can be n-dimensional.

The following simple example of a sales data cube aims to clarify the idea of dimensions and facts. The example cube contains three dimensions: products, cities and time (quarters). Every cell or intersection of these dimensions contains a measure or fact expressed in quantities. For example, if we need to know the quantities of “Product one” sold on “City one” during the second quarter of the year, we would only need to view the intersecting cell to determine that one item has been sold.



The following table is a tabular representation of the data presented in the above cube. Notice that the numbers in red are not visible in the above cube.

	City 1			**City 2**			**City 3**		
Time	Prod 1	Prod 2	Prod 3	Prod 1	Prod 2	Prod 3	Prod 1	Prod 2	Prod 3
Q1	1	4	1	5	5	5	2	3	4
Q2	2	1	6	2	3	5	2	3	3
Q3	1	4	1	7	5	5	2	4	2
Q4	2	1	6	5	6	3	2	5	1

To represent more dimensions in a cube, multiple cubes are needed. In order to represent n-D data series (n-1)-D cubes are needed (Han & Kambler, 2001).

Data Navigation

One of the advantages of dimensional modeling is that it allows an easy navigation through the data. The following are examples of navigations that OLAP systems can make available to users:

- *Aggregation (or Roll-Up).* This is simply the summarization of data. Users can summarize data across any dimension they want to see.
- *Roll down (or Drill down).* This allows users to see more detail data by moving within hierarchies of dimensions.
- *Screening (or Selection).* This allows users to select a criterion that is applied to the data or dimensions to restrict the data retrieved.
- *Slicing.* This allows user to "select all the data satisfying a fix condition along a particular dimension while navigating. A slice is a subset of a multidimensional array where a single value for one or more members of a dimension has been specified." (Jarke, Lenzerini, Vassiliou & Vassiliadis, 2003, 90).
- *Scoping.* This is very similar concept to screening, allowing the users to select a subset of the data.
- *Pivot (or Rotate).* This allows the user to change the orientation of the cube, permitting the same report to be presented in different ways.

In addition to navigation, some OLAP tools provide a feature called *closed loop decision support.* This feature provides the possibility to update operational systems based on the values that resulted from the analysis of the data.

Types of OLAP systems

According to the way OLAP systems are implemented, they can be classified as:

- *ROLAP (Relational OLAP).* As its name indicates, the implementation is based on a relational database. Every dimension and fact becomes a de-normalized table. This is one of the more popular types of solutions since most companies already have a relational database in place. Unlike OLTP systems, ROLAP systems are based on de-normalization, producing large tables. Usually one dimension table is linked to several dimension tables. If there is total de-normalization, a star schema is produced. On the other hand, if there is some normalization, a Snowflake Schema is produced.

- *MOLAP (Multidimensional OLAP).* This approach uses proprietary multidimensional databases (MDDs) to store data, which provides a very high

performance. The drawback is that since no standard exists, the query tool must be bought from the same manufacturer and migrating to others manufactures can prove to be a challenge.

- *HOLAP (Hybrid OLAP).* This is a combination of ROLAP and MOLAP typically having a MOLAP solution with a ROLAP back end. This approach is implemented as a three-tier environment. The back end server contains the relational database, the middle tier is comprised of a MOLAP application and the third tier is the end client application. This solution allows data to be stored both in the back end and in the middle tier. It also provides mechanisms to feed data into the system that do not originate from operational systems.

Data Mining Processes

With recent advances in technology, the use of the Internet, and virtually every facet of life being tracked through means of the computer, electronic data storage has exponentially grown. If you consider that data storage during the middle of the twentieth century averaged around 10 megabytes, compared to today's corporation data storages measured in terabytes, corporate data has grown by a factor of 100,000. (Barry & Linoff, 2000) Thanks to efficient storage methods that have been developed over the years, businesses and industries have gained the ability to efficiently store and scale their data to magnitudes one could not imagine. However, the current capabilities of database systems are limited, and another facet of database technology has emerged, data mining.

Data mining has a broad meaning, and can be perceived in many different perspectives. Data mining can be considered the process of querying data, or perhaps analyzing trends within data. However, given the context of this report, data mining specifically focuses on the concepts and techniques surrounding the ability to extract useful information from large collections of data, known as data warehouses. This portion of the report will specifically define the concept of data mining within the context of a

data warehouse lifecycle, and attempt to educate the reader on data mining's use, current state, and future capabilities.

What is Data Mining?

If one were to think of the natural progression of databases, data mining seems to fall in line with one's expectations. After all, take for example the technique of categorizing books. In the early years of printing books, there were no means to index and store these works in a coherent manner; that is, until methods such as the Dewey Decimal System emerged. From then, librarians were able to effectively store, search, and retrieve books in an efficient manner. This is similar to the growing concept of data warehousing and data mining.

As one begins to examine the deeper techniques and models of data mining, one realizes there are two actionable items to data mining. The first process of data mining is to identify, or discover, the useful data within a warehouse. The second process is to exploit the useful information to build effective applications or models around this data.

Data mining is often used for two primary functions: descriptive modeling and predictive modeling. As is the case with most businesses, there is a constant need to understand its past history, current state and potential future growth. By trying to understand its past history and current state, businesses perform queries (i.e. execute data mining) against current data, so that existing information is pieced together in a coherent and easy-to-read fashion.

On the other hand, predictive modeling allows businesses to execute data mining against existing information in the hopes of identifying patterns that would be useful to ultimately understanding future trends. These two broad types of modeling not only take

a database system and data mining tools, but also requires advanced high-processing technology, use of statistics, and the analytical resources to understand the data. With these components, different techniques of gathering and presenting data can be utilized. (Han & Kamber, 2001)

Concerning descriptive modeling, it is of the utmost importance that the end user easily understands the data being displayed. Although this may be partly true for predictive models, reports generating future trends do not have to be necessarily understood easily; so much as the data being portrayed is indeed a business trend. It is important to discern the difference. Descriptive models are reports often shown directly to the end-user, or customer. Remember, these reports allow the ability to take large amounts of data and classify, cluster, and organize itself in such a way that information is conveyed accurately. With descriptive modeling, in order to effectively display information, understanding the classification and integration of data is most important. Since the results will likely reveal themselves in quite an array of forms and structures, the integration of data and from what angle (i.e. subject matter) it is viewed from is most important.

With regards to predictive modeling, the most important aspect is to identify patters within the data. Identifying these patterns is done without classification and clustering; rather, patterns are identified through methods such as association. Within the art of predictive modeling, identifying such patters can be quite an enormous task. Literally thousands of various patterns can be identified within a large data warehouse. However, it is the handful of patters that prove to be extremely useful, given a particular business within a specific context. This is certainly not to say that the handful of patters identified

within one business will be the tell-all for another business. Actually, quite the opposite, businesses need to determine which patterns tell the truth about their current business situation. Factors such as size, subject area, points-of-interest, consumer, or manufacturer all play a part in identifying which pattern should be used for modeling purposes.

Business Context for Data Mining

Now that one has a fundamental understanding of data mining, and the potential techniques used within data mining, it becomes even more important to understand the business need for such a concept. As described already data mining allows the ability to scour massive amounts of data, seamlessly and effectively, in order to identify trends and potential cost-savings. Industries such as financial services and state governments are finding ways to implement data mining to good use.

Within the financial services industry, most banks make their profit from only the top 5% of their customers. Therefore, it becomes very critical to understand this market segment, and drive many of the opportunities for growth and retain as much of this segment as possible. Data mining plays an important role in this business process. Through use of data mining, data can be dissected to try and understand current buying behaviors, as well as identify potential trends for the future. With this kind of analysis, financial institutions can market to these specific data points, and hope to recruit and retain more of their top 5% profit customers. (Berson et. al., 2000)

Another business context in which data mining can prove to be useful is state governments; more specifically, department of revenue services. Unfortunately, it is commonplace in today's society for various individuals to try and cheat on their taxes. A huge cost savings to a taxpayer, along with an increasing deficit on the part of the state,

can ensue with such actions. Therefore, data mining might prove to be useful in such a scenario.

In the hopes of trying to prevent large deficits, state governments are beginning to use data mining, by examining large amount of tax data, payments, and returns, to understand the human behavior with regards to filing taxes. Often times, this analysis points collection agencies in the right direction for some of their worst offenders; thereby, obtaining the fees owed and effectively decreasing state deficits.

Overall, data mining can be used in many business contexts. Although the initial cost of implementing a data warehouse and operating data mining may be large, the return on investment has the potential to pay for itself many times over.

The Future of Data Mining

As for the future of data mining, two words should come to mind: infancy and limitless. Currently, most data mining techniques are not being used as an efficient method of predicting sales or business maneuvers, at least with any sort of significant accuracy. Rather, the process of data mining is most often used as a business practice to understand and improve on underlying business operations. Each day, enormous amounts of data are still being collected; yet remain untapped and unutilized to their fullest potential. However, it will not be too long before these large data banks allows markets, industries, and businesses all around the world to predict trends and individual preferences.

With the hopes of data mining prediction increasing, one must also be careful not to fall into the classic mistakes of data mining. Simply because data is being analyzed, does not necessarily mean the data is being presented in the correct context. (Delmater &

Hancock, 2001) If action were to be taken on results from data mining within the wrong business context, serious repercussions could be felt through the particular business. Therefore, it is extremely important to classify data in a manner that is appropriate to the business context being analyzed.

Ultimately, data mining is a sensitive, carefully executed process that requires extreme thought and analysis. However, used correctly, data mining can provide incredible benefits to any business looking for a competitive advantage.

CONCLUSION

The successful implementation of a data warehouse can provide many benefits to an organization. Although, for this to occur, a firm knowledge of the data warehouse lifecycle and its underlying terms and technologies must be attained. The understanding of why OLTP will not support the same goals as OLAP, the power of data mining, the importance of metadata, and the foresight to incorporate decision support software into the business process; are all necessary prerequisites to any data warehouse initiative. However, once a knowledgeable foundation has been established, the road to better business decisions, grounded in quality enterprise data, will improve proportionally to the confidence experienced by the organization that indeed better information leads to better decisions.

References

Adelman, Sid & Terpeluk-Moss, Larissa (2000). *Data Warehouse Project Management*. Boston: Addison Wesley.

Berry, M. J. A. & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. Hoboken: John Wiley & Sons.

Berson, A. & Smith, S. & Thearling, K. (2000). *Building Data Mining Applications for CRM.* New York: Computing McGraw-Hill.

Bray, T. (2001). *What is RDF*
Referenced July 26, 2003 from
http://www.xml.com/pub/a/2001/01/24/rdf.html?page=2

Delmater, R. & Hancock, M. (2001). *Data Mining Explained: A Manager's Guide to Customer-Centric Business Intelligence.* St. Louis: Digital Press.

English, Larry P. (1999). *Improving Data Warehouse and Business Information Quality.* New York: Johm Wiley & Sons.

Han, Jiawei & Kamber, Micheline (2001). *Data Mining: Concepts and Techniques.* USA: San Francisco: Morgan Kaufmann.

Hobbs, Lilian; Hillson, Susan & Lawande, Shilpa (2003). *Oracle9iR2 Data Warehousing.* Burlington: Digital Press.

Hoffer, Jeffrey A.; Prescott, Mary B. & McFadden, Fred R. (2002). *Modern Database Management.* Upper Saddle River: Prentice Hall.

Inmon, W. H. (2001)(a). *A Brief History of Metadata*
Retrieved July 20, 2003 from
http://www.inmoncif.com//library/whiteprs/wp.asp#Metadata

Inmon, W. H. (2001)(b). *An Illustrated Taxonomy of Metadata*
Retrieved July 20, 2003 from
http://www.inmoncif.com//library/whiteprs/wp.asp#Metadata

Inmon, W. H. (2002). *Building the Data Warehouse.*
New York: John Wiley & Sons, Inc.

Inmon, W. H. (2000). *Metadata in the Data Warehouse*
Retrieved July 20, 2003 from
http://www.inmoncif.com//library/whiteprs/wp.asp#Metadata

Inmon, W. H. (1997). *Metadata in the Data Warehouse: A Statement of Vision*

Retrieved July 20, 2003 from
http://www.inmoncif.com//library/whiteprs/wp.asp#Metadata

Jarke, Matthias; Maurizio, Lenzerini; Vassiliou, Yannis & Vassiliadis, Panos (2003). *Fundamentals of Data Warehouses.* (2nd. Edition). Berlin: Springer.

Johnson, Diane (1999). Implementing a Hybrid Online Analytical Processing (HOLAP) solution. In Purba, Sanjiv (Ed.). *Data Management Handbook.* (3rd. edition, pp. 747-753). Boca Raton: Auerbach.

Meyer, Don, Cannon, Casey (1998). *Building a Better Data Warehouse.* Upper Saddle River: Prentice Hall.

Scalzo, Bert (2003). *Oracle DBA Guide to Data Warehousing and Star Schemas.* Upper Saddle River: Prentice Hall PTR.