Effective Data Warehouse Modeling for

Turner Broadcasting Sales, Inc.

Robert Milton Underwood, Jr.

Table of Contents

Effective Data Warehouse Modeling for Turner Broadcasting Sales, Inc.

Introduction

Many companies have realized that having a data warehouse can be a very beneficial tool, both for management and for end users. According to information presented in the article required for this assignment, *Case Study: Turner Broadcasting Sales*, plans have been made by executives at Turner Sales to develop a fully integrated Web application based on the entire suite of Microsoft® Office 97 applications. By choosing a suite of products from a singular vendor, Turner Broadcasting Sales, Inc. (Turner Sales) has determined that consistency, integration, and flexibility among applications are important, thereby setting a good foundation for a successful web-based data warehouse.

By using Microsoft® products, Turner Sales will be able to make the transition to a Web based format much easier, especially since the Microsoft® applications in Office97 are so well integrated. Once this application platform is implemented, it will be easier for Turner Sales to move into a data warehouse than if they had chosen numerous different applications from several vendors. Having a consistent data application base will facilitate development of and transition to a data warehouse.

Turner Broadcasting has already proven itself to be a very strong player in its industry. Long-term success is of utmost importance to the company, especially during the current boom of companies with an Internet presence. According to Formichelli (2000), the long-term success of an organization comes from its intangible assets—

innovative processes and services and its customer relationships. The willingness of Turner Sales to have an Internet presence based upon a consistent application base indicates that they do indeed care about a high quality representation of data. Having a successful data warehouse can help management and employees know that they are doing the best that they can do to ensure long-term business success accompanied by positive customer perception.

Having a Web-based data warehouse will allow Turner Sales to realize the benefits of Web and Internet technology. For instance, Web browser software is very easy to use, requiring less training than even user-friendly database query tools (Laudon & Laudon, 2000). The Web interface requires no changes to the legacy system database.

According to the article for this case study, Turner Sales saw their Web advertising revenue increase by 1200 percent over the previous year. Advertisers realize that the Web presence of Turner Sales allows a highly targeted market exposure. It therefore is seen as prudent to advertise on Web sites that Turner Sales produces. It is helpful that Turner Sales' Web development is created with complimentary components. Using all Microsoft® applications will allow for as much consistency as is reasonably possible, and consistency is an important quality of a successful data warehouse.

Overview of Data Warehousing Process

According to Paul Hessinger in the book, *Data Warehouse: Practical Advice from the Experts*, there are three broad stages of the data warehousing process, *modeling, building,* and *deployment. Modeling* entails the understanding of the corporate culture of the enterprise, the whole business process, the information requirements of the business process, and the capabilities and processes of the decisions support system.

Building encompasses the establishment of a data model and a tool set for creating queries, forms, and reports. *Building* can be done for either a data mart or a data warehouse.

Deployment encompasses the implementation of the data warehouse project. The project is not usually considered "finished," but rather *evolves* as updates, revisions, expansions, and reengineering may take place at any time in the future of the data warehouse. In the early stages of the *deployment* phase, Turner Sales will want to allow users to explore the features of the data warehouse so that revisions can be made as soon as are reasonably possible.

As Turner Sales acknowledges the basic data warehouse procedural outline of *modeling, building,* and *deployment*, they can then get down to the production procedures of the project, as discussed in the sections that follow. The sections below contain major issues that Turner Sales should pay close attention to so that their data warehouse project can be successfully completed.

Planning

It is important that planners at Turner Sales clearly understand the definition of a data warehouse. A data warehouse is not just a database, nor is it the same as a database management system (DBMS). According to Bischoff and Alexander (1997), "The data warehouse is a collection of integrated, subject-oriented databases designed to support the DSS (decision support) function, where each unit of data is relevant to some moment of time." According to Laudon and Laudon (2000), "A data warehouse extracts current and historical data from operational systems inside the organization. These data are combined

with data from external sources and reorganized into a central database designed for management reporting and analysis."

William Edward[1] (2000) stated that one benefit that managers from Turner Sales can have upon deployment of the data warehouse is that they can get snapshots of old and current sales information. Effective use of this sales data could allow them to make future plans that provide for improved service to clients. Another practical use of the data warehouse is that they can get historical accounting data without having to request that data from the accounting department. This direct access to data will allow them to quicken their receipt of the needed data.

Turner Sales is ready for a data warehouse. Their plans to have an Internet presence have been successfully implemented. They already have a detailed plan that outlines the expected benefits from using all Microsoft® applications. According to Bischoff & Alexander (1997), it is very important that a company knows exactly what a data warehouse is and what to expect from one. The four main feature requirements of a data warehouse (Bischoff & Alexander, 1997) are that it be (a) subject-oriented, (b) integrated, (c) time-variant, and (d) nonvolatile. *Subject-oriented* databases are organized or designed around the important entities of the firm and require a data-driven design.

Having *integrated* data means that conventions, methods, relationships, and keys are integrated through a design process or methodology. Integrated data is very important for the purposes of consistency. The same conventions should be used throughout the organization, if possible. However, if different naming conventions are used in different departments of the organization, they should be integrated for the data warehouse project.

[1] William Edward is the Chief Operating Officer of Superior Views Real Estate. The company has over 60 offices nationwide, and Mr. Edward has overseen the development of their data warehouse.

A data field named *CustomerID* should not be found anywhere else in the data warehouse as *Cust_Number*.

The requirement of being *time-variant* means that the data warehouse is organized into various time periods, such as daily, weekly, monthly, quarterly and annually. This requirement subsequently allows employees to make future queries based on time periods. For instance, the following question could be answered to help the sales force determine progress in a certain zip code area: *how much has sales volume increased in the 78731 zip code area during the past month?*

The feature of *nonvolatility* means that the data warehouse is not updated in real time, but rather on a periodic basis. *When* the data warehouse is updated is usually a decision based on the timing required for updating other functional systems, like accounting records. For instance, if payroll is met on every Friday and the related records are updated to the data warehouse each Saturday evening, an employee would expect the new data to be available after the update and remain the same until the following update cycle.

Planning for the data warehouse project should not encompass the total scope of the enterprise (Bischoff & Alexander, 1997). Planning should be completed in phases. It is not wise to tackle a project too large. It is more practical to approach the project in an incremental fashion so that it will be easier to evaluate the systems for progress, and to make corrections that will apply on a larger scale.

According to Greenspun (1999), it is important that the designers of the project put themselves into the users' shoes before planning the site's content. Designers need to understand what the end user will expect. The end user, on the other hand, must

understand the reasons for, and the benefits of, the data warehouse project. Greenspun also suggests that the data might best be organized according to expected user interest rather than by mimicking the internal structure of the organization. It would be up to Turner Sales to determine if organizing data by expected user interest would best suit their needs.

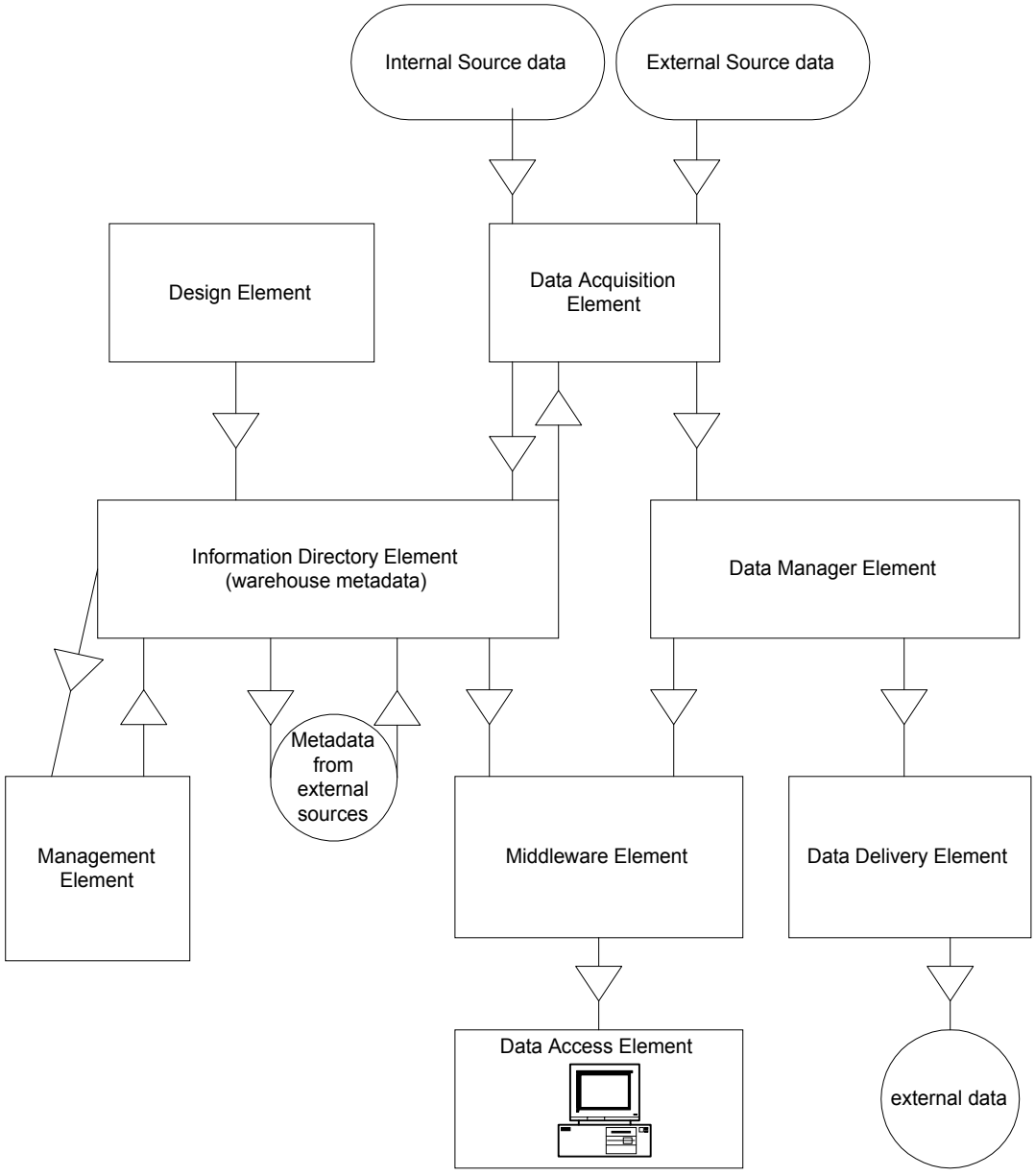
Technical Architecture

There are eight key elements of data warehouse architecture (see Figure 1), and Turner Sales' IT department (and/or outsourced specialists) will be responsible for making sure that each element is installed properly and functions correctly.

The *design element* is used for designing the data warehouse databases. It is used by the warehouse designers and administrators, who may also use work group data modeling tools in its application.

The *data acquisition element* is used for gathering data from databases and other source files. There are five main categories of tools that assist in the process of data acquisition: (a) code generators, (b) data replication tools, (c) data pumps, (d) data reengineering tools, and (e) other generalized data acquisition utility tools.

Figure 1
Main Elements of Data Warehouse Architecture



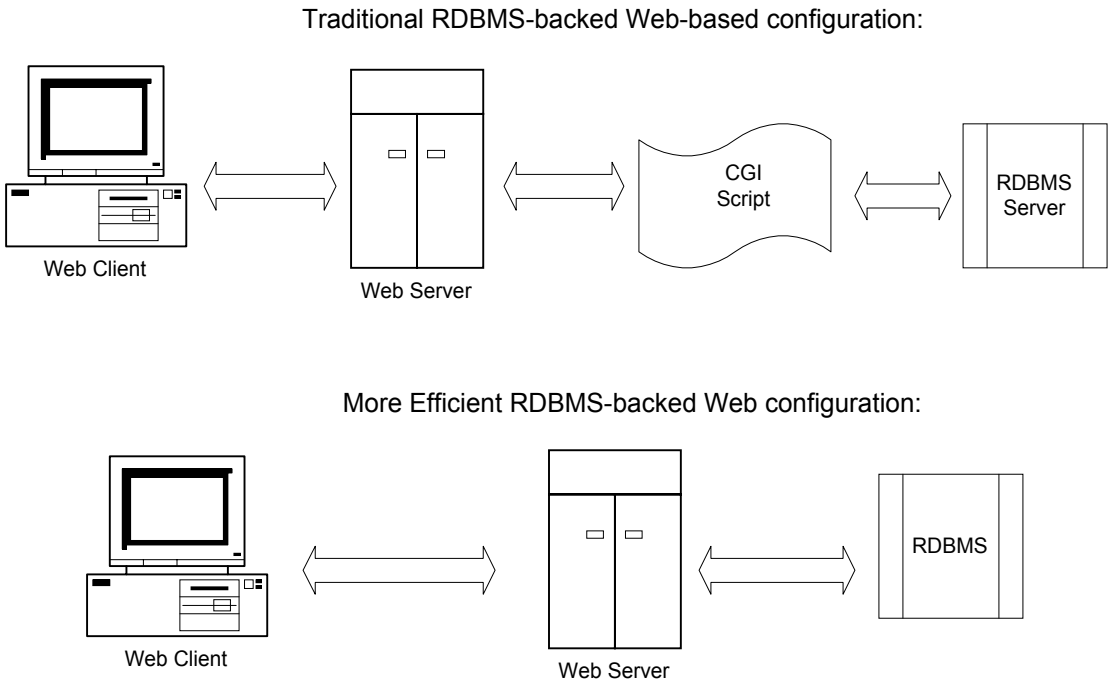
The *data manager element* is used by other elements in the data warehouse system for creating, managing, and accessing warehouse data. Either a relational database management system (RDBMS) or a multidimensional database management system

(MDBMS) is used. RDBMSs can be used to build either large company-wide or small departmental data warehouses, while MDBMSs are mainly used for building small departmental databases.

Turner Sales will probably decide to make their data warehouse a Web-based data warehouse. As mentioned earlier, they already use applications that are well integrated (i.e. the Microsoft® products) and that can make documents Web-ready. According to Greenspun (1999), the most practical database management software for Web-based systems is a relational database management system with a full-text indexer. The five factors that Greenspun suggests that should be considered when choosing a RDBMS to sit behind a Web-based site are (a) cost/complexity to administer, (b) lock management system, (c) full-text indexing option, (d) maximum length of VARCHAR data type, and (e) vendor support.

According to Greenspun (1997), a properly engineered RDBMS-backed Web-based site will have the *client* as the Web server program. The user would type something into a form on the Web client that gets transmitted to the Web server, which has an already-established connection to an RDBMS server (e.g. Microsoft® SQL Server). The data then goes back to the Web client. The scripts that are written inside the Web server program's process preclude the need for common gateway interface (CGI) processes. This configuration allows the saving of (a) the cost of starting up a CGI process, (b) the cost of starting up a new database server process, and (c) the cost of establishing the connection to the database. This configuration also allows a quicker return of data to the user via a *view*. Figure 2 shows a comparison of this configuration compared to a configuration utilizing a CGI script.

Figure 2
RDBMS-backed Web configurations



The *management element* allows for the administering of data warehouse operations. Those operations include managing data acquisition operations, backing up data, and securing the access to the data warehouse.

The *information directory element* allows business users to have information about the contents and meaning of data (viz. metadata) that is available in the data warehouse. The metadata is often created by the data warehouse designers during the design and development phases. Metadata essentially offers end users information *about* data in the warehouse, and is discussed in more detail later in this report.

The *data access element* provides the various tools that users need to access and analyze data. The tools include those for querying, creating reports, and analyzing data.

Multidimensional data analysis tools may also be included, as well as decision support system (DSS) application development tools.

The seventh element of the data warehouse architecture is the *middleware element*, which provides end user tools with access to warehouse databases. There are two main types of middleware: intelligent data warehousing middleware, and analytical servers.

The eighth element of the data warehouse architecture is the *data delivery element*, which is responsible for delivering warehouse data to other systems and to other warehouses. Users can usually define *when* they want data delivered.

Data

Quality of Data

Since managers and other users from Turner Sales will use data from a data warehouse to make decisions, it is very important that the data be of very high quality. Higher quality data will support higher quality decision-making.

There are a number of things to consider to ensure that data is of high quality. For instance, the data must be accurate. Inadvertent data entry errors can lead to problems. If a data entry clerk enters *Farrett* instead of *Garrett*, then it may be difficult to get all relevant data about employee Garrett.

Turner Sales must be aware that data should be stored according to data type. For instance, a numeric field should only contain numeric values.

Referential integrity rules should be followed so that data has the highest level of integrity. Referential integrity rules should be adequately defined in the data models.

The form of the data should be consistent. Naming conventions should be consistent as well. If data is consistent, then different departments can know exactly what to expect when they query for data and use the data. For instance, when querying for a certain sales total by date range, the date field should always be in the same format, for instance, *dd-mm-yyyy*.

Another indicator that data is of high quality is that it is minimally redundant. Keeping redundancy at a minimum helps reduce the chance of errors. However, some data may need to be held in more than one location in an enterprise if different departments need full access to it at all times.

Data must follow the business rules of an organization. For instance, if Turner Sales has added a new regional office, there cannot be a negative number of employees at that office. The number of employees must be zero (i.e. until the office opens) or greater than zero (i.e. once the office opens).

High quality data is also timely data. If the data warehouse is updated weekly, then users should expect all updates to be included and usable. Real-time data is not to be expected from a data warehouse, but rather from operational databases in departments that need real-time information.

Having data that is integrated is an indicator of high quality data. For instance, the Sales Department of Turner Sales may need information about its sales force, as well as about the customers represented by that sales force. Certain salespeople may be responsible for ongoing relationships with certain corporate customers, and relevant data from all databases should be integrated so that management has the best availability to use that data in a meaningful way.

In the movie *Wall Street*, the character of corporate raider Gordon Gekko remarked that he considered *information* the most valuable of all commodities. Information *is* a valuable commodity, but only if it is of high quality. If business decisions are made that rely on the accuracy of data, then it is crucial that all relevant data be of the highest possible quality. If the data is or poor quality, it can be an unfortunate liability instead of an beneficial asset.

Metadata

Metadata is considered to be an integral part of a data warehouse. Metadata is simply defined as *data about data*. The data about the data can take many forms, but it basically is descriptive about the data itself, identifies where the data came from, and describes how the data can be used.

While metadata is not always needed for small data warehouse projects, especially those that only contain a subset of the possible useful data that a company *could* include, it will be needed for the data warehouse designed for Turner Sales. The inclusion of metadata in the data warehouse will improve management's high-level decision support systems, by making it easier for users to understand the data in the warehouse.

Turner Sales should be aware of the three stages of the metadata life cycle: collection, maintenance, and deployment. Collection is concerned with the gathering of metadata into a central repository. Dave Gleason, in *Data Warehouse: Practical Advice from the Experts*, suggests that the collecting of metadata be as automated as possible. The use of automated scanners that can parse the data structures and populate the

metadata repository is preferred over the manual typing of analysts of data into the repository.

The second stage of the metadata life cycle is maintenance. In the maintenance stage, metadata is kept up to date. Automation is important for this stage so that revisions and additions of data about data are made with the highest possible integrity.

The third stage of the metadata life cycle is deployment. Metadata can only be valuable to Turner Sales if it is effectively deployed. Metadata can be deployed differently to different user groups, but the important thing is that it *is* deployed, and in a timely manner.

Turner Sales will want metadata to be integrated in the data warehouse so that it is as easy to use as is using the actual historical data. This integration of metadata into the data warehouse can be done with data access and analysis tools. The highest integration is realized by having full interconnectivity between the metadata tool and the query tool. This level of integration allows users to go directly to a query tool once they have found what they need with the metadata tool.

Data Directory

Since a data warehouse can contain very large quantities of data, it is important that the user not be lost or sidetracked in a search for relevant data. A data directory should be created as the data warehouse is built so that ease of use is facilitated. There are three main types of data in a data warehouse: physical data, semantic data, and environmental data. Physical data is the actual data that is moved from legacy or operational systems to the data warehouse. Semantic data is described as descriptions

about the data that has been moved. Environmental data includes access details, end user profiles, and security information.

A data directory is like a catalog, in that it allows users to circumvent problems accessing data that they need to do their jobs by allowing them access to information that is helpful. The five basic categories or functions of a data directory refer to (a) administration, (b) browser, search, and request, (c) movement of metadata and physical data, (d) event management, and (e) application programming interfaces. These categorizations make it easier for users to understand how the data is organized in the warehouse.

Turner Sales will benefit by having a comprehensive data directory in that end users will be able to take data from the data warehouse and use it in a meaningful way. For instance, the data directory will allow users to ascertain where data came from and where its operational counterpart is. Users can also track usage trends and changes to the warehouse environment. One of the most helpful management reasons for having a data directory is that it allows for more efficient data warehouse administration.

Data Transformation

Prior to being loaded into a data warehouse from legacy systems and/or operational systems, data is typically "transformed." The modification of the data is the essence of data transformation. There are two reasons that Turner Sales should perform data transformation. The first reason is that transforming the data will improve the quality of the data in the data warehouse. The second reason is that data that is located in the data warehouse will be easier to use after it is transformed. What Turner Sales will realize is

that effective data transformation actually verifies and improves the quality and usability of the data as it goes into the data warehouse.

Dave Gleason, in *Data Warehouse: Practical Advice from the Experts*, lists four types of data transformation: (a) *simple transformation*, (b) *cleansing and scrubbing*, (c) *integration*, and (d) *aggregation and summarization*. Some companies use only simple transformation, and may experience problems because of failure to recognize the need to use one or more of the other three types. Turner Sales should therefore be aware of the value of using each of the four types of data transformation, as discussed below.

Simple transformation contains the simplest forms of data transformation, and is mainly concerned with modifying the content of one data attribute at a time, without considering the context of that attribute or other data related to that attribute. Data type conversions, date/time format conversions, and field decoding are three examples of the use of simple transformation.

Cleansing and scrubbing represent a second type of data transformation. This type of data transformation is more complex than *simple transformation*. What distinguishes this type from *simple transformation* is that it is concerned with the actual content of the fields or groups of fields rather than just the storage formatting. One type of cleansing that Turner Sales should perform is checking for valid values in data fields. For example, a typographical error (e.g. a purchase date entered as year *1798* rather than *1998*) would be corrected. A second type of cleansing that Turner Sales should perform is complex reformatting, as necessary. For instance, it is possible that data entries in different departments could be entered as *1998*, *98*, and ‘*98* to represent the year 1998.

The process of complex reformatting would consistently format all entries for the *year* data field.

Integration is a third type of data transformation. It involves integrating data from multiple sources into one consistent data model. Since data from different sources may require different accompanying business rules, the process of integration provides for the eventual cohesiveness of the system. One example of an *integration* process is simple field-level mappings, where a field in an operational database is moved to a field in a data warehouse. Along the way, simple transformation and cleansing may occur to sharpen the integrity of the data for the requirements of the data warehouse.

The fourth type of data transformation that Turner Sales will perform is *aggregation and summarization*. This type of data transformation will minimize the number of queries needed with regards to summarization and aggregation of data. An example of summarization is a query for total sales from the sales department for one particular month. An example of aggregation of data is a query that shows total income from various sources over a period of time. Aggregation can also be used to remove old and unnecessary data from the data warehouse.

The real importance realized by the process of data transformation is that it will allow Turner Sales to convert raw operational data into data that is suitable for the data warehouse. They should not be concerned that the process is often complex and time-consuming. What is important that data delivered to the data warehouse is of high quality and is easily usable by those needing access to the data.

Design and Implementation

Physical Design

Designing a data warehouse should be completed in steps. Each step builds upon the prior step, and modifications are made as needed at each stage of development. Turner Sales should adhere to the following steps for design of their data warehouse.

- Step One: Consider all user requirements and the broad scope of the project.
- Step Two: A subject area data model should be developed. As suggested already, it is best if the project is developed on a small scale at first. One particular subject area, such as *Sales*, should be focused on. Within this subject area, relationships should be identified and diagrammed. Metadata should be documented as the process develops.
- Step Three: A data warehouse logical model should then be developed. Its development grows from the information that is gathered in step two.
- Step Four: The data warehouse architecture should be considered and developed.
- Step Five: The design of the physical database should begin. By this stage, there should be a clear and straightforward structure based on effective planning in the prior four stages.
- Step Six: A data directory should be created for end users.
- Step Seven: The various sources of data to be copied to the data warehouse will be identified from legacy systems and/or operational systems.
- Step Eight: Tools are purchased or developed to get the information from the legacy and operational systems.

- Step Nine: The actual data from the legacy and operational systems is populated into the data warehouse.
- Step Ten: Users should be interviewed in person or via questionnaire that will help the developers recognize problems that exist.
- Step Eleven: Work can be organized to correct any problems recognized from the prior step. Improvements can then be made to the system.

It is important that Turner Sales make reviews at the end of each step listed above. The process of data warehouse design and implementation *evolves* and it is valuable to the company to assess both progress and problems at each step.

Data Warehouse Administration

Systems Administration

Completion of the data warehouse is not the end of the project. It is actually the official beginning for the operational stage of the project, for which system administrators take on a very active role. While design specialists may have moved on to other projects upon completion of the physical data warehouse, many individuals will remain to make sure that the system continues to operate as expected, especially as the data warehouse grows over time. The IT department of the company may need to be expanded to accommodate the maintenance and administrative needs of the data warehouse.

Availability of data

Administrators at Turner Sales need to determine how often the data warehouse will be available for use. Users must rely on its availability during the expected hours of

operation. Potential users should be notified well in advance if the system needs to be down for maintenance procedures.

Data Currency

How often should data be updated to the data warehouse? How *current* should the data be? These questions will be answered as they relate to Turner Sales and their need for up-to-date data. Turner Sales will determine how often the data needs to be updated. Some data, like quarterly reports, would be updated quarterly. If employees are paid every two weeks, then payroll data may need to be updated biweekly. Since the function of a data warehouse does not include the accommodation of real-time information, only historical data is therefore to be expected. However, historical data does not have to be "old." It may be accounting data for the last quarter, sales totals from the last month, or demographical data about a new sales territory.

Retention of data

System administrators and IT professionals must determine how long data will be retained on the data warehouse. Some data will be need to kept indefinitely. But other data will have decreasing importance as time goes on.

Security

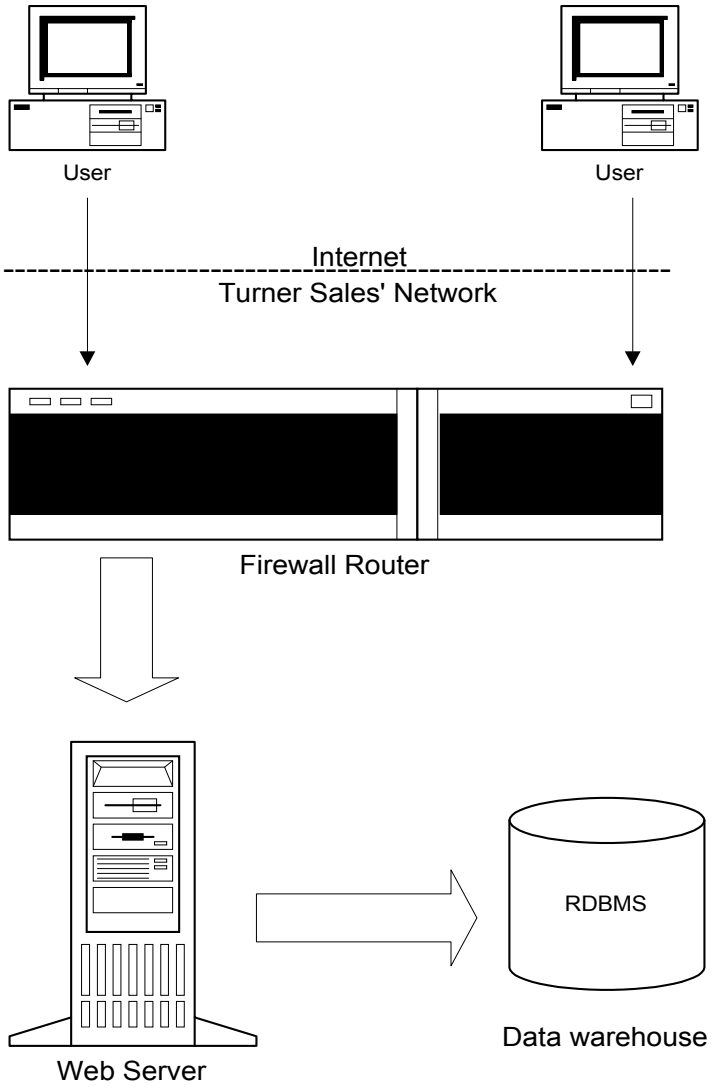
Security measures are required to ensure that only authorized users are able to obtain the data that they seek from the data warehouse. Management at Turner Sales should determine the level of access that each individual has to the data warehouse. That level should depend upon what the individuals need to successfully perform their jobs.

Security issues to consider include (a) how to safely administer passwords, (b) how best to secure back doors, (c) how to manage the availability of remote access, (d)

how best to create user sign-on procedures, and (e) how to ensure the physical security of the server room, to protect from both possible physical damage and unauthorized access.

Turner Sales may decide to allow remote access to the data warehouse. If they decide to create a Web-based data warehouse, their data would be more secure if the Web

Figure 3
Typical Web Security Configuration



server is inside the firewall. This configuration is diagrammed in Figure 3, and would apply to Turner Sales if they allow Internet access to employees. (If they have a virtual private network, then the chances of unauthorized use by outsiders are diminished.) Having the Web server within the secure confines of the firewall computer provides more security from a hacker who may attempt to breach the system. However, if the Web server is placed outside the firewall, it is important to at least program the firewall to not let anyone make a direct connection to the port that provides access to the data warehouse.

Conclusion

The plans that Turner Sales has to create a data warehouse have stemmed from the realization that data from legacy and operational systems can be of value at some point in the future for the purposes of decision-making. As companies grow over time, it becomes necessary to have some system for retaining data that may be needed in the future. To store all of this data in one of the operational databases would slow down the functioning of the database, and would fill it with data that may be underutilized. The data warehouse then becomes the chosen system to maintain historical data.

Having a Web-based data warehouse will allow Turner Sales to have a browser-based front end. This will provide an ease of use that most people are already familiar with. The Web-based platform will also be easier to structure since they use Microsoft® applications that allow ease of saving documents as hypertext markup language (HTML) files.

Data warehouses not only offer improved and refined information, but they make it simple for decision makers to obtain the information. They even include the ability to

model and remodel the data. Turner Sales will be able to use the data warehouse to analyze patterns and trends

The data warehouse project at Turner Sales will be successful if the plan is strategic, carefully measured, and appropriate for their needs. If Turner Sales spends time wisely in the early planning stages, then they will have set the foundation for a successful project. It would then be necessary to make sure that the steps that follow are performed iteratively, and that each step builds systematically from the prior step. Success should be demonstrative, and they must also realize that once the data warehouse system is operational, sound management and system administration will ensure its effectiveness and success over the coming years.

References

Bischoff, J. and Alexander, T. 1997. Data warehouse: practical advice from the experts. Upper Saddle River, NJ: Prentice Hall.

Edward, William. (2000, July 18). [personal interview, Austin, TX].

Formichelli, L. (2000, July/August). Nurturing Knowledge Capital. Business Advisor. 13.

Greenspun, P. 1999. Philip and Alex's guide to Web publishing (pp. 23,140). San Francisco: Morgan Kaufmann Publishers.

Laudon, K. C., & Laudon, J. P. (2000). Management information systems: organization and technology in the networked enterprise (6^{th} ed.) (pp. 247 – 249). Upper Saddle River, NJ: Prentice Hall.

Microsoft Corporation. 1996. Case Study: Turner Broadcasting Sales. 1-8.