Web-enabled Data Warehouse and Data Webhouse

Anca MEHEDINTU, Ion BULIGIU, Cerasela PÎRVU
Department of Informatics in Economy, Faculty of Economics and Business Administration, University of Craiova
ancamehedintu@yahoo.com, buligiu_ion@yahoo.com, ccpirvu2006@yahoo.com

In this paper, our objectives are to understanding what data warehouse means examine the reasons for doing so, appreciate the implications of the convergence of Web technologies and those of the data warehouse and examine the steps for building a Web-enabled data warehouse. The web revolution has propelled the data warehouse out onto the main stage, because in many situations the data warehouse must be the engine that controls or analysis the web experience. In order to step up to this new responsibility, the data warehouse must adjust. The nature of the data warehouse needs to be somewhat different. As a result, our data warehouses are becoming data webhouses. The data warehouse is becoming the infrastructure that supports customer relationship management (CRM). And the data warehouse is being asked to make the customer clickstream available for analysis. This rebirth of data warehousing architecture is called the data webhouse.

Keywords: *data warehouse, web-enabled, data mart, Internet, intranet, extranet*

Introduction

End-user application developers are increasingly building applications around Web pages. The user interface development environment of choice is now a web page development environment. Data warehousing is one of the core responsibilities of information technology. In many ways Data Warehousing fulfils the promise of "getting the data out" after the Online Transaction Processing (OLTP) based system "gets the data in". The web revolution has certainly not replaced the need for the data warehouse. In fact the web revolution has raised everyone's expectations much higher that all sorts of information will be seamlessly published through web browser interface. The audience for data warehouse data has grown from internal management to encompass customers, partners and much larger pool of internal employees. The web's focus on the customer experience has made many organization aware of learning about the customer and giving the customer useful information.

Web-enabled data warehouse

In order to transform our data warehouse into a Web-enabled data warehouse, we first have to bring the data warehouse to the Web, and secondly we need to bring the Web to your data warehouse. Furthermore, we will discuss these two distinct aspects of a Web-enabled data warehouse.

In early implementations, the corporate data warehouse was intended for managers, executives, business analysts, and a few other high-level employees as a tool for analysis and decision making. Information from the data warehouse was delivered to this group of users in a client/server environment. But today's data warehouses are no longer confined to a select group of internal users. Under present conditions, corporations need to increase the productivity of all the members in the corporation's value chain. Useful information from the corporate data warehouse must be provided not only to the employees but also to customers, suppliers, and all other business partners.

So in today's business climate, you need to open your data warehouse to the entire community of users in the value chain, and perhaps also to the general public.

This new delivery method will radically change the ways your users will retrieve, analyze, and share information from your data warehouse. The components of your information delivery will be different. The Internet interface will include browser, search engine, push technology, home page, information content, hypertext links, and downloaded Java or ActiveX applets.

When you bring your data warehouse to the Web, from the point of view of the users, the key requirements are: self-service data access, interactive analysis, high availability and performance, zero-administration client (thin client technology such as Java applets), tight security, and unified metadata.

Bringing the Web to the warehouse essentially involves capturing the clickstream of all the visitors to your company's Web site and performing all the traditional data warehousing functions. And you must accomplish this, near real-time, in an environment that has now come to be known as the data Webhouse. Your effort will involve extraction, transformation, and loading of the clickstream data to the Webhouse repository. You will have to build dimensional schemas from the clickstream data and deploy information delivery systems from the Webhouse.

Clickstream data tracks how people proceeded through your company's Web site, what triggers purchases, what attracts people, and what makes them come back. Clickstream data enables analysis of several key measures including:

- Customer demand
- Effectiveness of marketing promotions
- Effectiveness of affiliate relationship among products
- Demographic data collection
- Customer buying patterns
- Feedback on Web site design

A clickstream Webhouse may be the single most important tool for identifying, prioritizing, and retaining e-commerce customers. The Webhouse can produce the following useful information:

- Site statistics
- Visitor conversions
- Referring partner links
- Site navigation resulting in orders
- Site navigation not resulting in orders
- Pages that are session killers
- Relationships between customer profiles and page activities
- Best customer and worst customer analysis

Architectural configuration

Figure 1 indicates an architectural configuration for a Web-enabled data warehouse. Notice the presence of the essential functional features of a traditional data warehouse. In addition to the data warehouse repository holding the usual types of information, the Webhouse repository contains clickstream data.

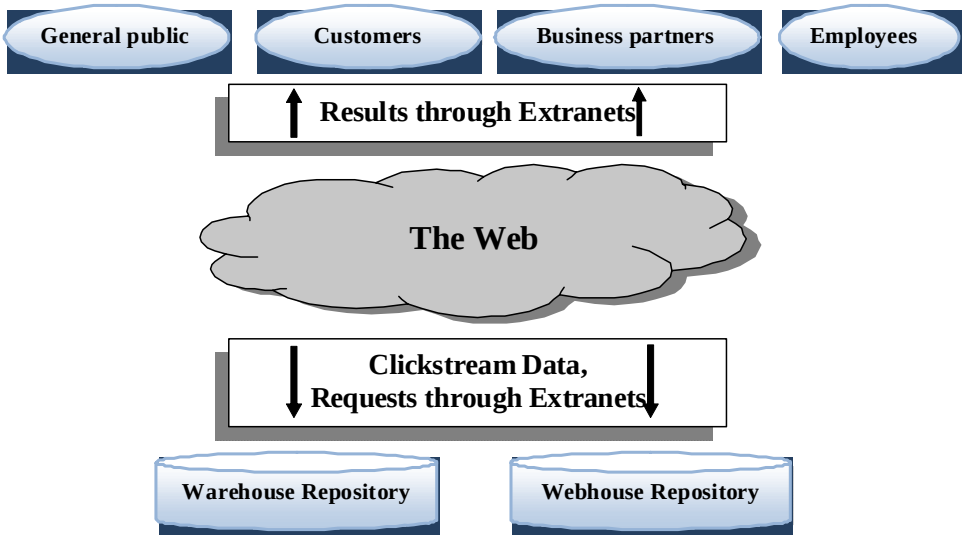

Fig.1. Simplified View of Web-enabled Data Warehouse

A Web-enabled data warehouse uses the Web for information delivery and collaboration among users. As months go by, more and more data warehouses are being connected to the Web. Essentially, this means an increase in the access to information in the data warehouse. Increase in information access, in turn, means increase in the knowledge level of the enter-

prise. It is true that even before connecting to the Web, you could give access for information to more of your users, but with much difficulty and a proportionate increase in communication costs. The Web has changed all that. It is now a lot easier to add more users. The communications infrastructure is already there. Almost all of your users have Web browsers. No additional client software is required. You can leverage the Web that already exists. The exponential growth of the Web, with its networks, servers, users, and pages, has brought about the adoption of the Internet, intranets, and extranets as information transmission media. The Web-enabled data warehouse takes center stage in the Web revolution.
Now consider our data warehouse in relation to the Web. Users need the data warehouse for information. The business partners can use some of the specific information from the data warehouse. What do all of these have in common? Familiarity with the Web and ability to access it easily. These are strong reasons for a Web-enabled data warehouse.
How can we exploit the Web technology for our data warehouse? How can we connect the warehouse to Web? Let's quickly review three information delivery mechanisms that companies have adopted based on Web technology. In each case, users access information with Web browsers.
Internet. The first medium is, of course, the Internet, which provides low-cost transmission of information. You may exchange information with anyone within or outside the company. Because the information is transmitted over public networks, security concerns must be addressed.
Intranet. An intranet is a private computer network based on the data communications standards of the public Internet. The applications posting information over the intranet all reside within the firewall and, therefore, are more secure. You can have all the benefits of the popular Web technology. In addition, you can manage security better on the intranet.
Extranet. The Internet and the intranet have been followed by the extranet. An extranet is not completely open like the Internet, nor is it restricted just for internal use like an intranet. An extranet is an intranet that is open to selective access by outside parties. From your intranet, in addition to looking inward and downward, you could look outward to your customers, suppliers, and business partners.
Figure 2 illustrates how information from the data warehouse may be delivered over these information delivery mechanisms. Note how your data warehouse may be deployed over the Web. If you choose to restrict your data warehouse to internal users, then you adopt the intranet. If it has to be opened up to outside parties with proper authorization, you go with the extranet. In both cases, the information delivery technology and the transmission protocols are the same. The intranet and the extranet come with several advantages. Here are a few:

- With a universal browser, your users will have a single point of entry for information.
- Minimal training is required to access information. Users already know how to use a browser.
- Universal browsers will run on any systems.
- Web technology opens up multiple information formats to the users. They can receive text, images, charts, even video and audio.
- It is easy to keep the intranet/extranet updated so that there will be one source of information.
- Opening up your data warehouse to your business partners over the extranet fosters and strengthens the partnerships.
- Deployment and maintenance costs are low for Web-enabling your data warehouse.
- Primarily, the network costs are less. Infrastructure costs are also low.

Adapting the Data Warehouse for the Web

Much is expected of a Web-enabled data warehouse. That means you have to reinvent your data warehouse. You have to carry out a number of tasks to adapt your data warehouse for the Web. Let us consider the specific provisions for Web-enabling your data warehouse.
First, let's get back to the discussion of the three stages following the introduction of a new technology. Apart from reducing costs from the substitution, demand for data warehouse information has increased. Most compa-

nies seem to be stuck at the end of the second stage. Only a few companies have moved on to the next stage and have realized third-order results. What are these results? Some of such results include extranet and consumer data marts, management by exception, and automated supply and value chains. When you adapt your data warehouse for the Web, make sure that you do not stay put at the second stage. Make plans to exploit the potential of the Web and move on to the third stage where the real benefits are found.

Study the following list of requisites for adapting the data warehouse to the Web.

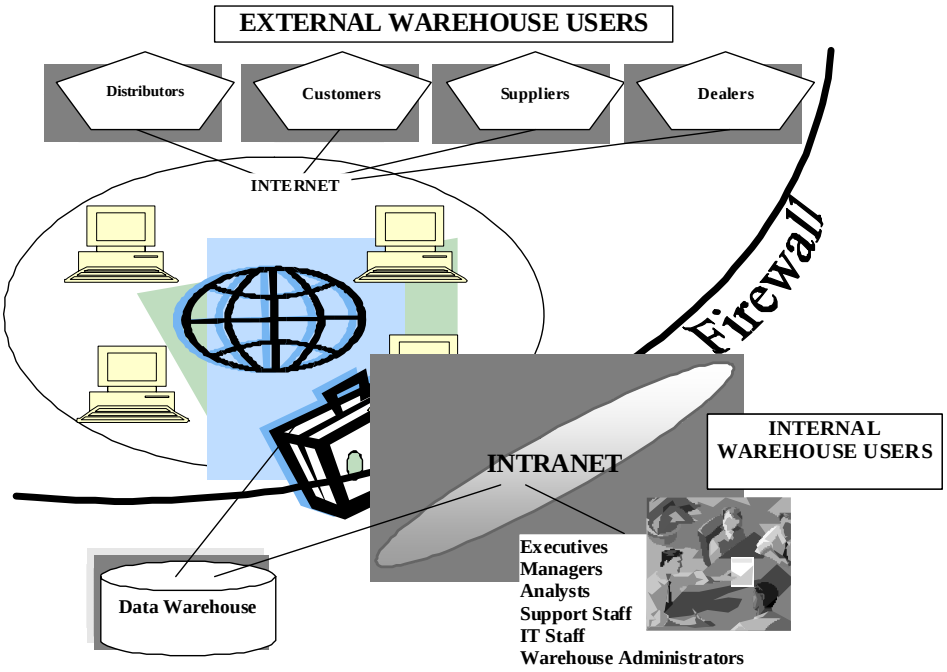

Fig.2. The web and Data Warehouse

Information "Push" Technique. The data warehouse was designed and implemented using the "pull" technique. The information delivery system pulls information from the data warehouse based on requests, and then provides it to the users. The Web offers another technique. The Web can "push" information to users without their asking for it every time. Your data warehouse must be able to adopt the "push" technique.

Ease of Usage. With the availability of click stream data, you can very quickly check the behavior of the user at the site. Among other things, click stream data reveals how easy or difficult it is for the users to browse the pages. Ease of use appears at the top of the list of requirements.

Speedy Response. Some data warehouses allow jobs to run long to produce the desired results. In the Web model, speed is expected and cannot be negotiated or compromised.

No Downtime. The Web model is designed so that the system is available all the time. Similarly, the Web-enabled data warehouse has no downtime.

Multimedia Output. Web pages have multiple data types - textual, numeric, graphics, sound, video, animation, audio, and maps. These types are expected to show as outputs in the information delivery system of the Web-enabled data warehouse.

Market of One. Web information delivery is tending to become highly personalized, with dynamically created XML pages replacing static HTML coding. Web-enabled data warehouses will have to follow suit.

Scalability. More access, more users, and more data - these are the results of Web-enabling the data warehouse. Therefore, scalability becomes a primary concern.

The Web as a Data Source

When we talk about Web-enabling the data warehouse, the first, and perhaps the only thought that comes to mind is the use of Web technology as an information delivery mechanism. Ironically, it rarely crosses your mind that Web content is a valuable and potent data source for your data warehouse. You may hesitate before extracting data from the Web for your Web-enabled data warehouse. Information content on the Web is so disparate and fragmented. You need to build a special search and extract system to sift through the mounds of information and pick up what is relevant for your data warehouse. Assume that your project team is able to build such an extraction system, then selection and extraction consists of a few distinct steps. Before extraction, you must verify the accuracy of the source data. Just because data was found on the Web, you cannot automatically assume it is accurate. You can get clues to accuracy from the types of sources. Please refer to Figure 3 showing an arrangement of components for data selection and extraction from the Web.

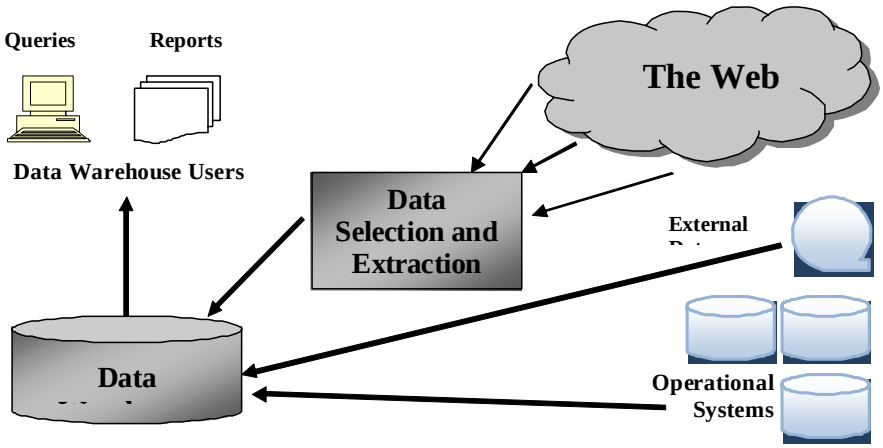

Fig.3. Web data for the data warehouse

Building a web-enabled data warehouse

In 1999, Dr. Ralph Kimball popularized a new term, "data Webhouse," which included the notion of a Web-enabled data warehouse. He declared that the data warehouse is taking central stage in the Web revolution. He went on to state that this requires restating and adjusting our data warehouse thinking.

In attempting to formulate the principles for building a Web-enabled data warehouse, let us first review the nature of the data Webhouse. We will use this knowledge to define the implementation considerations. Now let's review the features. Here is a list of the principal features of the data Webhouse:

- It is a fully distributed system. Many independent nodes make up the whole. As Kimball would say, there is no center to the data Webhouse.
- It is a Web-enabled system; it is beyond a client/server system. The distribution of tasks and the arrangement of the components are radically different.
- The Web browser is the key to information delivery. The system delivers the results of requests for information through remote browsers.
- Because of its openness, security is a serious concern.
- The Web supports all data types including textual, numeric, graphical, photographic, audio, video, and more. It therefore follows that the data Webhouse supports many forms of data.
- The system provides results to information requests within reasonable response times.
- User interface design is of paramount importance for ease of use and for effective publication on the Web. Unlike the interfaces in other configurations, the Web has a definite method to measure the effectiveness of the user interface. The analysis of the clickstream data tells you how good the interface is.
- By nature, the data Webhouse necessitates a nicely distributed architecture comprising small-scale data marts.

• Because the arrangement of the components is based on a "bus" architecture of linked data marts, it is important to have fully conformed dimensions and completely conformed or standardized facts.
• The Web never sleeps. Your data Webhouse is expected to be up all the time.
• Finally, remember that the data Webhouse is meant to be open to all groups of users, both inside and outside the enterprise - employees, customers, suppliers, and other business partners.

The major features described above lead us to factors you need to consider for implementing a Web-enabled data warehouse. Each feature listed above demands readjustment of the implementation principles. Mostly, by going through a list of features, you can derive what is required. We want to highlight just a few implementation considerations that are crucial.

• In order to achieve basic architectural coherence among the distributed units, fervently adopt dimensional modeling as the basic modeling technique.
• Use the data warehouse bus architecture. This architecture, with its fully conformed dimensions and completely standardized facts, is conducive to the flow of correct information.
• In a distributed environment, a suggestion is to centralize the definitions of the conformed dimensions and the conformed facts. This need not be physical centralization; logical centralization will work. This centralization gives the semblance of a center to the data Webhouse.
• Still the question remains: who actually conforms the dimensions and facts? The answer depends on what will work for your environment. If feasible, assign the task of conforming the dimensions and facts to the local groups of participants. Each group gets the responsibility to come up with the definitions for a dimension or a set of facts.
• Well, how do all units become aware of the complete set of definitions for all dimensions and facts? This is where the Web comes in handy. You can have the definitions published on the Web; they then become the standards for conformed dimensions and facts.
• How do you physically implement the conformed dimension tables and the conformed fact tables? Dimension tables are often physically duplicated. Again, see what is feasible in your environment. Total physical centralization of all the dimension tables may not be practical, but the conformed fact tables are rarely duplicated. Generally, fact tables are very large in comparison to the dimension tables.
• We understand the data Webhouse as a distributed set of dimensions and facts based on possibly dissimilar database technologies. How can you make such a distributed collection work as a cohesive whole? This is what the query tool or the report writer is required to do in such a distributed configuration. Let us say one of the remote users executes a specific query. The query tool must establish connections to each of the required fact table providers and retrieve result sets that are constrained by the conformed dimensions. Then the tool must combine all the retrieved result sets in the application server using a single-pass sort-merge. The combination will produce the correct final result set simply because the dimensions are all conformed.

Conclusions

Let's look at a Web architecture configuration, and we can observe that the architecture is more complex that a two-tier or three-tier client/server architecture. We need additional tiers to accommodate the requirements of Web computing. At a minimum, you need to have a Web server between the browser clients and the database. Also, note the firewall to protect the corporate applications from outside intrusions.

This covers the overall architecture. See Figure 4 what show a model for delivering information. The model illustrates how HTML pages are translated into SQL queries passed on to the Database Management System (DBMS) using Common Gateway Interface (CGI) scripts. This model shows the components for information delivery through HTML pages. The model may be generalized to illustrate other technologies.

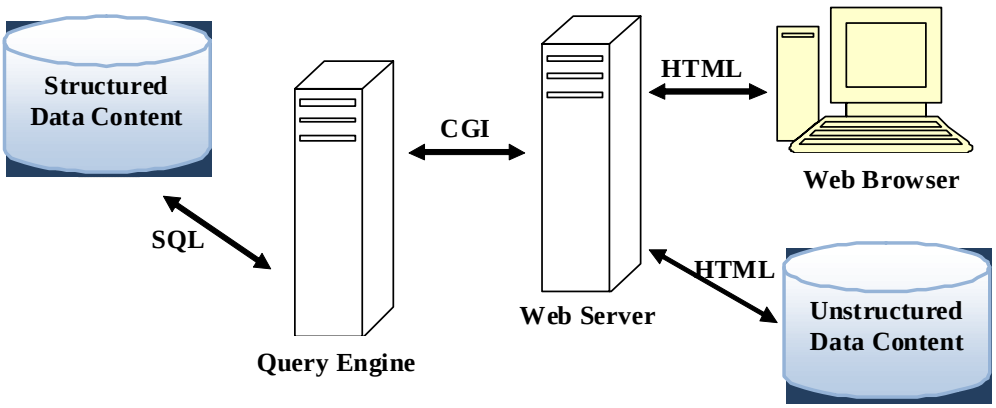

Fig.4. Web processing model

The web and the data warehouse are drawn together like two powerful magnets. The web needs the warehouse for many of it's customer centric functions and the warehouse is been transformed by the demands of the web to Data Webhouses. Data Webhouses will play a major role in the corporate world in the very near future.

References

1.Giovinazzo, W. A., *The Web-Enabled Data Warehouse*, Prentice Hall PTR, 2002
2.Grant, G., *ERP and Data Warehousing in Organizations: Issues and Challenges*, Carleton University, Canada, IRM Press, 2003
3.Kimball, R., Merz, R., *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse,* New York, John Wiley & Sons, Inc., 2000
4.Paulraj P., *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*, New York, John Wiley & Sons, Inc., 2001
5.Shaw, M., Blanning, R., Strader, T., Whinston, A. *Handbook on Electronic Commerce*, Springer, 1999