

JOURNAL

OF DATA WAREHOUSING

A 101communications Publication

Volume 7
Number 2
Spring 2002

CONTENTS

Editor's Note and Statement of Purpose 3
Hugh J. Watson

Delivery of E-

syncsort

E-Business and the Corporate Information Factory

Vic Werner, Craig Abramson, and Kenny Kistler

**Identifying and Removing
and Maintaining First-Class Organizational Data** 15
Stacey Herdlein

E-Business and the Corporate Information Factory 21
Vic Werner, Craig Abramson, and Kenny Kistler

Hosted Data Warehouse 27
Mike Thornton and Mike Lampa

Identifying Meta Data Requirements 35
Adrienne Tannenbaum

Information Quality: The Quest for Justification 44
Frank Dravis

**Today's Intelligent Data Warehouse
Demands Quality Data** 50
Jeff Canter

Editorial Calendar and Instructions for Authors 2002 54

About The Data Warehousing Institute 57

E-Business and the Corporate Information Factory

Vic Werner, Craig Abramson, and Kenny Kistler

Abstract

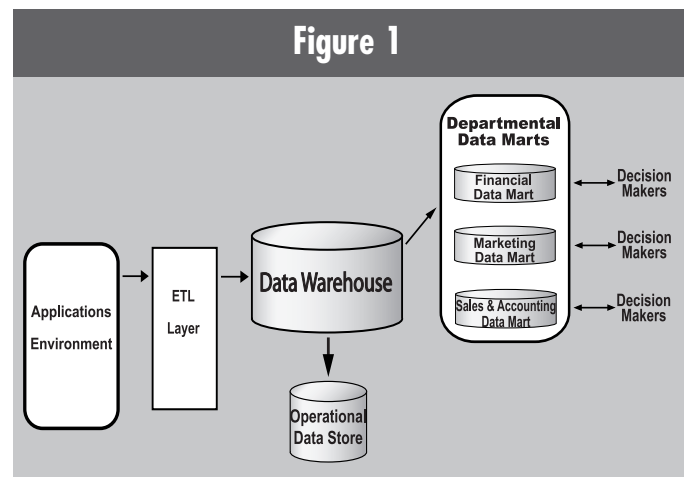
When one thinks of a successful e-business, one usually pictures a flashy Web site with plenty of advertising that helps drive traffic. What many companies fail to realize is that the architecture behind a Web site is just as critical to the success of an e-business. This article examines how a company's e-business architecture can be incorporated into the corporate information factory (CIF) for better management and analysis of Web data. A detailed description is provided on the components of the CIF, which include the applications environment; the ETL layer; the data warehouse with current and historical detailed data; the data mart(s); an operational data store (ODS); an Internet and intranet; and a meta data repository. The Web environment of the CIF is then examined, offering insight into the additional architectural components needed to support an e-business. These components include a granularity manager, HTML page manager, a local operational data store, Web logs, a session manager, a cookie cognition manager, and a personalization facility. The article concludes with a look at how the CIF can handle the massive amounts of Web data that is collected.

Introduction

In today's business environment, gaining the competitive edge is a necessity. One way companies are achieving this is by incorporating an e-business into their corporate strategy. By offering their products and services on the Web, these companies are not only expanding their market reach, they're also obtaining valuable insight into their customers. That's because every move a customer makes on a Web site is recorded into Web logs and clickstream records. When combined and analyzed, these records detail where customers come from, what they are looking for, and how they think. The problem is that in order to take advantage of this information, the data must be easily accessible to the decision makers within a company. One proven solution is to integrate the Web environment into the CIF.

The Corporate Information Factory

Designed to deliver business intelligence and business management capabilities, the CIF is a technical architecture that is driven by data from business operations. It has proven to be a stable architecture for any size enterprise building strategic and tactical decision support systems. The CIF is comprised of both producers of data and consumers of information. The producers capture the data from the operational systems and transform it into a useable format. The consumers then access the information, manipulate it, and assimilate into their own environments (Imhoff, 1999).



The Integration of E-Business into the CIF

Data flows into the CIF through the applications environment. This environment contains the input applications that interface with the customer, usually in the form of transaction processing. These applications are able to gather the raw data, edit and adjust the data, and audit the data. The raw data is manipulated into an integrated format as it passes into the ETL layer. This layer prepares the raw data for informational processing by summarizing, reformatting, converting, re-sequencing, and merging it. The data then flows into the ODS and/or the data warehouse layer (Inmon, 1999).

The ODS stores and manages the data that it receives from the ETL layer. Although the ODS only contains detailed data, it is used to support dynamic summary data. Current detailed data is the basis of the data warehouse. This data is historical, containing the most recent history of the organization. The data has already passed through the ETL layer. When the data is customized and summarized, it is usually moved from the data warehouse to a specific data mart. The data marts are typically divided into departments such as finance, marketing, sales, and accounting (Inmon, 1999). Once this data is passed on to the appropriate data marts, it can be accessed and analyzed by the decision makers within a company. Now imagine how much a company would benefit if Web data was also made available in the CIF.

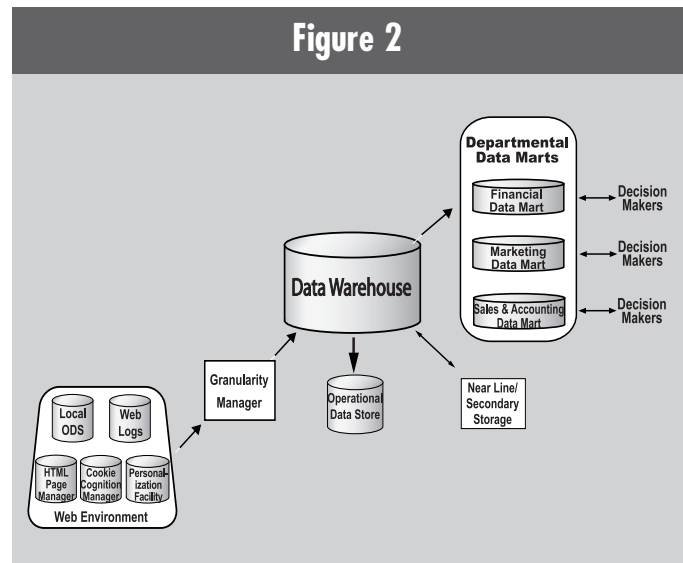
The CIF and the Web

Developers, marketers, and salespeople can gain invaluable knowledge about their most popular and effective products, services, and campaigns from analyzing the raw data collected on Web servers. But more importantly, they can also learn what needs to be changed, improved, or introduced by scrutinizing customer behavior on their Web site. To make this data readily available to the decision makers within the company, the e-business infrastructure can be incorporated into the CIF.

The inclusion of an e-business into the CIF would consist of the following additional architectural components: a granularity manager; the Web environment; HTML page manager; a local ODS; Web logs; a session manager; a cookie cognition manager; and a personalization facility (Inmon, 1999). The components of the Web environment have several key functions. The granularity manager is used to condense the Web data down into a usable size. It edits, deletes, summarizes and prepares the Web data for analytical processing and sends it to the data warehouse. The HTML page manager interfaces directly with users on the Internet. The Web manager coordinates the activities that occur inside the Web. The cookie cognition manager determines if a person previously visited the company's Web site and, if so, the personalization facility provides personalized messages to that person (Inmon, 2001).

Another component is the local ODS, which contains the Web data needed for the immediate operation of the Web environment. Integrated online transaction processing occurs within the ODS. It supports a two- to three-second response time and can be updated directly. The ODS consists of four classes:

- **Class I** is where the data is loaded into the ODS synchronously from the operation environment. There is no



difference in the timing updates between the operation environment and the ODS;

- **Class II** has a four-hour delay between the timing of updates in the operational environment and the ODS;
- **Class III** has an overnight delay in the timing updates; and
- **Class IV** has updates that occur spontaneously and on a non-scheduled basis from the data warehouse environment (Inmon and Terdeman, 2001).

How the CIF Handles All the Web Data

One of the key reasons to place the Web environment in the CIF architecture is that the CIF can accommodate vast amounts of Web data. In fact, the CIF doesn't place any limits on the amount of data that can be collected. To understand this, it's important to understand the flow the data takes from the Web into the CIF. Once the data comes in from the Web site in the form of Web logs and clickstream records, it passes through the granularity manager, which is the software component that edits and manipulates Web data and moves it into the data warehouse. Once this processed data enters the corporate data warehouse, it can be moved to near line/secondary storage once the data becomes inactive (Inmon, 2000).

By moving data through all the different components of the CIF, a company can extend the amount of data collected indefinitely. The data warehouse alone is able to handle significant amounts of data. Its capacity can be expanded greatly by adding near line and/or secondary storage as well. Not only does near line/secondary

storage expand the capacities of the data warehouse, it also greatly reduces costs compared to high performance disk storage. Another benefit to using this type of storage is that performance is enhanced. Also, the data can be stored at a lower level of granularity (Inmon, 2000).

One component of the CIF that handles the enormous amount of Web data is the clickstream data mart. This data mart is used to store information about Web activity that can be analyzed at a later time. Setting up this type of data mart can be completed using a four-step methodology developed by data warehouse expert Ralph Kimball. The first step is to define the source of the data. This includes information such as the date and time a Web page was visited, the user's IP address, the page requested, and any cookie information from a previous visitor. Once this clickstream data has been collected, it must be transformed to provide a clear picture of the user's session. The cookie cognition manager will transform the data source into the following format: date and time of page hit; identity of user; session ID; and page and event requested (Kimball, 1999).

The second step is to choose the grain, or every meaningful event in every individual user session, of the fact table. During the transformation process, automatic events, such as the downloading of a graphic, can be filtered out of the data. The next step is to choose the dimensions of the grain. These dimensions are universal date, universal time, local date, local time, user, page, event, and session. The local and universal versions of the date and time allow a company to determine when clickstream events occurred in absolute time as well as in the user's time. The user dimension offers information that the user provides about their identity. The page dimension describes the location of the user on the Web site, including the type of page. In the session dimension, all the page events of a user are grouped together and labeled. Then the final step is to choose the facts appropriate for the grain. This involves making an accurate estimate of the time a user has spent on a Web page. By following this dimensional design, a company is able to perform numerous powerful queries (Kimball, 1999).

By adding an ETL layer as data is moved into the data warehouse and then into a data mart, the amount of Web data can be significantly reduced. The ETL layer allows the data to be reformatted, re-encoded, cleansed, and restructured to become consistent with the needs of the company (Inmon and Terdeman, 2001). Because a large amount of extraneous data has been removed as it passed through the ETL layer, queries and response times are much quicker.

Choosing the Right ETL Product

There are numerous ETL products to choose from in the marketplace, each with its own distinct method for manipulating data. The vendors for these products vary in size from small companies to large organizations with multiple product lines (Brohan, 2001). Because of the wide range of choices, selecting the right one can be a long process. It's important to consider all the features a product should have in order to complete a specific project, as well as what may be needed in the near future. A number of features that tend to be important for most projects include:

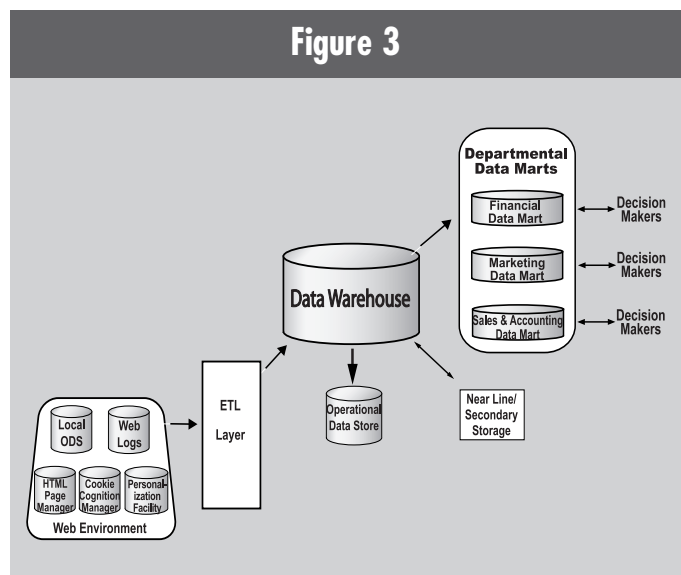
Scalability – benchmark tests as well as independent reviews should indicate that a product can easily handle the gigabytes of data generated from large transaction processing applications.

Support for various data and file formats – since the data may change over time, a product should be able to support a variety of different formats.

Ease-of-use – some products may require complex commands while others perform tasks with the click of a mouse.

Performance – processing time is critical, so it's important to choose a product that not only has all of the key features needed, but can also complete the data processing as fast as possible.

Price – products can range in price from a thousand dollars to over a million. Of course, the higher-end products are not specifically built as ETL products.



The Corporate Information Factory

Recommendations – ask other people in the industry for recommendations on products that they've used.

Corbis Corporation

One company that has developed a unique way to handle the large amount of data in their Web environment is Corbis Corporation. The company is taking advantage of its Web servers to optimize their data gathering process. As a provider of digital images for consumers and creative professionals, Corbis uses its Web site to do business with a growing number of online shoppers as well as the world's most popular publications. Behind the scenes at corbis.com, Web logs record clickstream data, including such details as the number of unique visits to the Web site, the most popular pages, the most purchased products, and the pages that seem to be "session killers," where visitors frequently stop the session and leave the site. These dimensions of the grain are defined by Corbis.

Every day, Corbis experiences nearly half a million visits to its Web site during which customers browse through extensive online art galleries, download pictures, order framed prints, or license specific images for repeated use. Given the dynamic nature of this Web traffic, the massive server logs record more information about the Web site and its visitors than analysts at Corbis can realistically use. Every link, picture, and page that is accessed by each visitor is recorded. At the end of a typical business day, there are literally hundreds of megabytes of information to sort through to access the hidden customer data that analysts covet most.

Identifying the crucial information in these Web logs, isolating it, and preparing it for warehousing and analysis is a task requiring time, resources, and expertise. C.J. Venkataraman, senior software architect at Corbis, supervises this process. For more than two years, Venkataraman has managed the flow of data from the company's operational data sources, between its e-commerce systems, and into its data warehouses. With his team of programmers and designers, he defines the technology architecture at Corbis prior to its implementation by the business systems group. The platform consists of dual proxy Compaq Proliant servers with 1.5 GB of memory and Pentium III 600 processors running Windows 2000 and the Microsoft SQL Server database engine. A key function of this e-business architecture is to facilitate Web log processing.

"We have 26 Web log servers at Corbis, and we copy each server's log files onto one huge server," said Venkataraman. "Basically, we merge them all, filter the resulting file to include only the customer and Web site information that we want to keep for analysis,

and then compress it for storage. Once it is stored and analyzed, we can more clearly see how many visits occur to our Web site, what our customers are doing on the site, and the top domains that our visitors are coming from." In addition to customer analysis, Corbis uses the Web logs to evaluate its partnerships with Yahoo! and AltaVista. "Our clickstream data provides insight into who has been sending us the most Web traffic," said Venkataraman.

Managers and analysts at Corbis can also use Web log information to research how their customers think. For example, clickstream data may reveal that many visitors are filling online shopping carts, but leaving before actually purchasing the items. After further investigation, Corbis may decide that by redesigning the checkout page, the percentage of visitors who follow through with purchases can be increased.

In theory, this method of customer research sounds simple and practical enough. But when the Web log files contain data from some 500,000 hits a day, the resulting numbers can be overwhelming. Each log file at Corbis swells to at least 200 MB in size before it is merged with the other Web logs. The resulting merge can easily exceed five GB daily, considerably slowing down the filtering and compression process. "End to end, it was taking close to five hours a day to complete," said Venkataraman. "We don't need most of what we have in our Web logs, so sifting through all of these files every day is a long process. We were using a tool that we made in Visual Basic, but it wasn't getting the job done."

Concerned that excessive time and resources were being spent converting daily Web logs to a single, compressed flat file for their analysts, Corbis began to search for alternatives to use in their ETL layer. Venkataraman turned to SyncSort, a high-performance sort, merge, and copy tool from Syncsort Incorporated in Woodcliff Lake, New Jersey. "It brought the daily routine down to one hour, which saves us about four hours on average per day."

Corbis has been able to use this specially crafted combination of technological architecture and high-performance tools to turn a potential corporate advantage into an actual one. By overcoming the obstacles in their Web log processing, Corbis can more efficiently use the rich mines of data in its Web logs to better understand customer behavior and, in turn, increase profits.

Summary

The CIF can play a key role in the success of an e-business. It is designed to accommodate the enormous amount of Web data collected each day and make this data readily available to the decision makers within a company. To incorporate an e-business into

the CIF, additional architectural components are needed which include a granularity manager; the Web environment; HTML page manager; a local operational data store; Web logs; a session manager; a cookie cognition manager; and a personalization facility. Also, by including a clickstream data mart in the CIF, a company is able to store information about Web activity that can be quickly accessed and analyzed at a later time. But as Corbis Corporation discovered, there can be an excessive amount of extraneous Web data. They found that by adding an ETL layer between the data warehouse and data mart, they were able to significantly reduce the amount of Web data for analysis.

REFERENCES

Brohan, Mark, "Are Companies Struggling to Evaluate ETL Tools," <http://www.dmreview.com/master.cfm?NavID=55&EdID=3596>, June, 2001.

Imhoff, Claudia, "The Corporate Information Factory," <http://www.dmreview.com/master.cfm?NavID=198&EdID=1667>, DM Review, December, 1999.

Inmon, William H., "Data Warehouse, ODS and Data Marts: The Corporate Information Factory," <http://www.billinmon.com/library/articles/artfacto.asp>, 1999.

Inmon, William H., "Building the Corporate Information Factory from a Blueprint, Part I," <http://www.billinmon.com>, 2001.

Inmon, William H. and Robert H. Terdeman, "The Evolution of the Corporate Information Factory," <http://www.billinmon.com>, 2001.

Inmon, William H., "Ebusiness Infrastructure," <http://www.billinmon.com>, 2000.

Kimball, Ralph, "Clicking with Your Customer," *Intelligent Enterprise Magazine*, http://www.intelligententerprise.com/db_area/archives/1999/990501/warehouse.shtml, January 5, 1999, Volume 2, Number 1.

BIOGRAPHIES

Vic Werner is Director of Marketing at Syncsort Incorporated, focusing on promotions, project management, and tradeshow. With more than 20 years' experience in the technology sector, he has gained an expertise in areas such as database management,

data sorting, clickstream data and Web logs, and backup and restore solutions. Previously, Vic served as a marketing and sales executive at Control Data, Xerox, and Frontec. He earned an MBA at Fordham University.

Vic Werner
Director of Marketing
Syncsort Incorporated
50 Tice Boulevard
Woodcliff Lake, NJ 07677
201.930.8259
Email: vwerner@syncsort.com

Craig Abramson is a technical analyst at Syncsort Incorporated, focusing on the latest data sorting, data aggregation, and backup and restore solutions. He has more than six years' experience in the field working on projects dealing with data warehousing, database management, and Web log processing.

Craig Abramson
Technical Analyst
Syncsort Incorporated
50 Tice Boulevard
Woodcliff Lake, NJ 07677
201.930.9700, Ext. 308
Email: cabramson@syncsort.com

Kenny Kistler is a technical analyst at Syncsort Incorporated, focusing on the latest tools for Web log processing and data quality management. He has more than three years' experience in systems implementation and documentation for a variety of operating platforms.

Kenny Kistler
Technical Analyst
Syncsort Incorporated
50 Tice Boulevard
Woodcliff Lake, NJ 07677
201.930.8233
Email: kkistler@syncsort.com



www.syncsort.com