

**The application of Web technology to data warehousing is an intriguing combination, but does it deliver practical value?**

**by Richard Hackathorn**

# Reaping the Web for Your Data Warehouse

---

*DBMS*, August 1998

---

Amid the chaos of the Web is a diverse collection of ever-changing information, some of which can be highly valuable to your enterprise. The challenge is to wade (with big boots) through the Web, discovering and acquiring those resources that have value for your business. In this article, I will explore the use of information resources from the Web as input to data warehouses. I call this activity Web farming, or the systematic refining of information resources on the Web for business intelligence.

Both the Web and data warehousing are hot technologies receiving considerable attention within the IT industry. In several areas, the combination has proven highly successful. Publishing warehouse data via an intranet is a highly productive approach that combines Web delivery mechanisms with Web-enabled databases. Generating dynamic pages from Web-enabled databases and adding Java applets to manipulate data locally to the browser has made available whole new areas of data analysis and data mining for warehouse users.

In contrast, no one has seriously considered extracting content from the Web and using it as input to the data warehouse. The paradigm of the Web is radically different from the paradigm for the data warehouse. Adapting an old programming term, you might say that Web content is spaghetti data. That is, it links to everywhere with little discipline. Furthermore, Web content is highly volatile and constantly changing. The Web's diversity challenges our imagination and appreciation for new forms of creative expression. The problem is cultivating those few nuggets with real business value from that diversity.

Reactions to using Web content tend to be negative. Web content is too unreliable and unstable for business decisions. The interaction with Web sites is too messy. Transformation of hypertext into a structured database is often impossible. Images and sound contain a lot of hidden content but are not discernible to a machine.

There are increasing instances of various systems integrating Web content into their operations. A simple example is the monitoring of international currency rates by a U.K. financial firm. The system monitors three specific financial sites for changes, retrieves the contents, and parses the retrieved page to obtain the useful data. With a delay of approximately 20 minutes, it retrieves currency exchange rates (such as USD to GBP) along with stock prices and news headlines containing predefined keywords. The data is loaded into several database tables, along with linking information, source, and a date/time stamp. The user can choose the currency for display and ask for recent headlines concerning that currency.

There are hundreds of commercial databases available via the Web, some requiring substantial fees. An example of a recent addition is the IBM Web site ([patent.womplex.ibm.com/respage.html](http://patent.womplex.ibm.com/respage.html)) that offers a database of patents issued in the United States during the last 27 years - more than 2 million patents. You can visually display the results, highlighting clusters of similar patents or ranking patents according to areas of interest. The imperative exists to incorporate the analysis of patent information in firms of all sizes, including R&D, competitive intelligence, licensing management, and strategic planning. Even the investment banking community can determine the intellectual assets of potential acquisition or investment candidates.

New approaches to handling Web content systematically are emerging weekly. The Jungle Corp. ([www.jungle.com](http://www.jungle.com)) has commercialized research at Stanford University to turn the Web into a single virtual database. Using specific wrappers around Web sources, unified metadata drives a database engine to process queries joining multiple sources. This technology has been applied to job classified ads (for example, JobCanopy in the San Jose area) and e-commerce (ShopCanopy for 40 merchants in eight categories).

## Using the Web for Business Intelligence

Many people think that data external to the organization has little value to the business because internal operational systems contain all the required data. Is there a need for companies to integrate external data into their warehouses?

Professor Peter Drucker, a senior guru of management practice, admonishes IT executives to look outside their enterprises for information. He remarked that the single biggest challenge is to "organize outside data because change occurs from the outside." He predicted that the obsession with internal data would lead to organizations being blindsided by external forces.

In the majority of data warehousing efforts, enterprises focus inward. As markets become turbulent, the traditional way of doing business becomes less viable. Data from internal operational systems becomes less relevant to managing your business and planning for its future. Instead, the enterprise should be keenly alert to outside sources.

An enterprise must know more and more about its customers, suppliers, competitors, government agencies, and other external factors. It must enhance the information from internal systems with information about external factors. The synergism of the combination creates the greatest business benefit for the enterprise.

The Web has become the universal and global delivery mechanism for external data. In many ways, the Web is the mother of all data warehouses. The immense resources of the Web, with all of its complexity and dynamics, are largely untapped. Valuable information about external business factors is readily available on the Web and is becoming more so each day.

## Objectives of Web Farming

Web farming is not surfing the Web haphazardly, wandering from one intriguing item to another. Nor is it a one-time search of the Web. On a continuous and systematic basis, a Web farming system must deliver, to the right people at the right time, information highly relevant to the enterprise. In effect, a Web farming system acts as the eyes and ears of the enterprise, focusing externally to be aware of

important changes in the business environment.

Web farming has the objective of refining Web content in a systematic manner. In particular, refining this content involves the processes of discovering, acquiring, structuring, and disseminating, as I'll explain later. Therefore, the specific objectives of Web farming are:

- To discover Web content that is highly relevant to the business
- To acquire that content so it is properly validated within a historical context
- To structure the content into a useful form that's compatible with the data warehouse
- To disseminate the content to the proper people so it has direct and positive impacts on specific business processes
- To manage the previous steps in a systematic manner as part of the production operations of a data center environment.

Web farming is often confused with our personal experiences of surfing the Web -- long periods of frustration with a few moments of elation. However, Web farming is serious business. Many people falsely think that Web farming is like planting a small garden in the backyard. In contrast, Web farming is like managing a large agricultural concern that involves many people and several thousand acres of farmland. There are similarities in the basic concepts, but the scaling of a personal garden to an agricultural business changes the methodology, architecture, tools, and techniques.

## Reliability of Web Content

The reliability of Web content is an important issue that you must manage carefully. Consider the following situation. If you hear, "Buy IBM stock because it will double over the next month," your reaction should depend on who made that statement and in what context. Was it a random conversation overheard on the subway, a chat with a friend over dinner or a phone call from a trusted financial advisor? The same is true with judging the reliability of Web content.

Most people have the "flake free" image of Web content. In reality, the Web is a global bulletin board where the wise and the foolish have equal space. Acquiring content from the Web should not reflect positively or negatively on its quality.

Think of Web resources in terms of quality and coverage. (See [Figure 1](#)) Toward the top are information resources of high quality (for example, accuracy, currency, and validity), while resources toward the right have a wide coverage (for example, scope, variety, and diversity). The interesting aspect of the Web is that its information resources occupy all the quadrants in this [figure](#).

In the upper center of the [figure](#), the commercial online databases from Dialog Information Services and similar vendors have traditionally supplied businesses with high-quality information about numerous topics. However, the complexity of using these services and the infrequent update cycles have limited their usefulness.

To the left, government databases have become tremendously useful in recent years. Public information was often available only by spending many hours of manual labor at libraries or government offices. The Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database maintained by the U.S. Security and Exchange Commission contains extensive information on publicly traded companies and is updated daily.

In the upper left, corporate Web sites often contain vast amounts of useful information in white papers, product demos, and press releases, eliminating the necessity to attend trade exhibits to learn the "latest and greatest" in a marketplace.

Finally, the flaky content occupies the lower half of the [figure](#). Its value is not in the quality of any specific item but in its constantly changing diversity. In combination with the other Web resources, the flaky content acts as a wide-angle lens to avoid tunnel vision of one's marketplace.

## Information Flow

The data warehouse occupies a central position in the information flow of a Web farming system. Like operational systems, the Web farming system provides input to the data warehouse. The result is to disseminate the refined information about specific business subjects to the enterprise.

As the primary source of external perspectives on the business, the Web can be supplemented (but not replaced) by content from the enterprise's intranet. This content is typically in the format of internal Web sites, word processing documents, spreadsheets, and email messages. However, the content from the intranet is usually limited to internal information about the enterprise, missing an important aspect of Web farming.

Most information acquired by the Web farming system will not be in a form suitable for the data warehouse. It will either be unstructured hypertext or unverified tabular values. In either case, you must perform a process of refining that information before loading it into the warehouse. Even in this unrefined state, this information could be highly valuable to the enterprise. It may be useful to disseminate this information directly via textual message alerts or "What's New" bulletins.

## Refining Information

When a data warehouse is first implemented within an enterprise, you need to analyze and reengineer the data from operational systems. The same is true for Web farming. Before you can load Web content into a warehouse, you must refine that information.

There are four processes for refining information: discovery, acquisition, structuring, and dissemination.

*Discovery* is the exploration of available Web resources to find those items that relate to specific topics. Discovery involves considerable "detective" work far beyond searching generic directory services (such as Yahoo) or indexing services (such as AltaVista). Furthermore, the discovery activity must be a continuous process because data sources are continually appearing (and disappearing) from the Web. A business analyst is the central figure in this activity and requires advanced search and indexing tools to be productive.

*Acquisition* is the collection and maintenance of content identified by its source. The main goal of acquisition is to maintain the historical context so you can analyze content in the context of past changes. Acquisition requires a secured server platform with large storage capacity.

*Structuring* is the analysis, validation, and transformation of content into a more useful format and into a more meaningful structure. The formats can be Web pages, spreadsheets, word processing

documents, and database tables. As we move toward loading data into a warehouse, the structures must be compatible with the star-schema design and with key identifier values.

*Dissemination* is the packaging and delivery of information to the appropriate consumers, either directly or through a data warehouse. It requires a range of dissemination mechanisms from predetermined schedules to ad hoc queries. Newer technologies such as information brokering and preference matching may be desirable.

There is a bidirectional flow to the processes. (See [Figure 2](#).) The left-to-right flow refines the content of information, which becomes more structured and validated. The right-to-left flow refines the control of the processes, which become more selective and discriminating.

## **Rendezvous with the Data Warehouse**

The most difficult part of Web farming is the rendezvous with the data warehousing system, especially in matching the data structure of Web content with the data warehouse schema.

Consider a simple data schema for a sales warehouse. (See [Figure 3a](#).) In this warehouse, we have sales data by customer, product, and store aggregated on a weekly basis. Let's assume that we have mostly corporate customers, rather than individuals, as in a large office furniture company.

Web farming would be valuable by enhancing the demographics (for example, quarterly financials) about customers, such as you can find in the EDGAR Web site. (See [Figure 3b](#).) By adding information on customer demographics, you can perform selective marketing based on the profitability and requirements of customers. By knowing what types of customers buy what types of products at which stores, we can promote specific sales and anticipate demand. For example, companies that are expanding are more likely to order office furniture.

Demographic information is added to the customer dimension to enhance analyses. As experience with the demographics matures, data mining techniques can cluster customers into meaningful categories based on demographics. (See [Figure 3c](#).)

Another example of using Web farming to enhance a data warehouse is the addition of demographics on the store. (See [Figure 3d](#).) Using ZIP codes and even the full street address, you can add census data about the communities surrounding the store to your data warehouse as another business dimension. This enhancement can lead to more effective management of stores based on their communities and more effective placement of new stores.

A final example involves adding data that is highly volatile, such as weather. (See [Figure 3e](#).) Seasonal variations have always been an important part of sales analysis. However, a sudden heavy snowstorm or an intense hailstorm can also affect sales of specific products in addition to the seasonal variations. This example shows that timely and continuous flow of Web content into the warehouse can aid in the day-to-day management of the business.

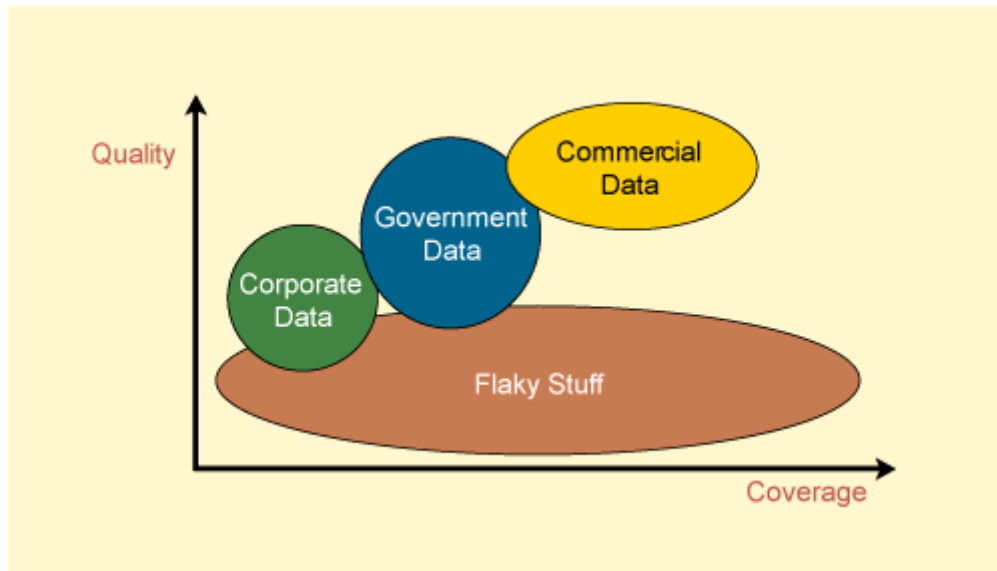
## **Where Are We Heading?**

In many ways, the data warehouse is not a requirement for Web farming. You could successfully farm the Web, reaping tremendous value for the business and bypass the data warehouse entirely. However, establishing the Web farming function is much easier for an enterprise if it has a mature understanding

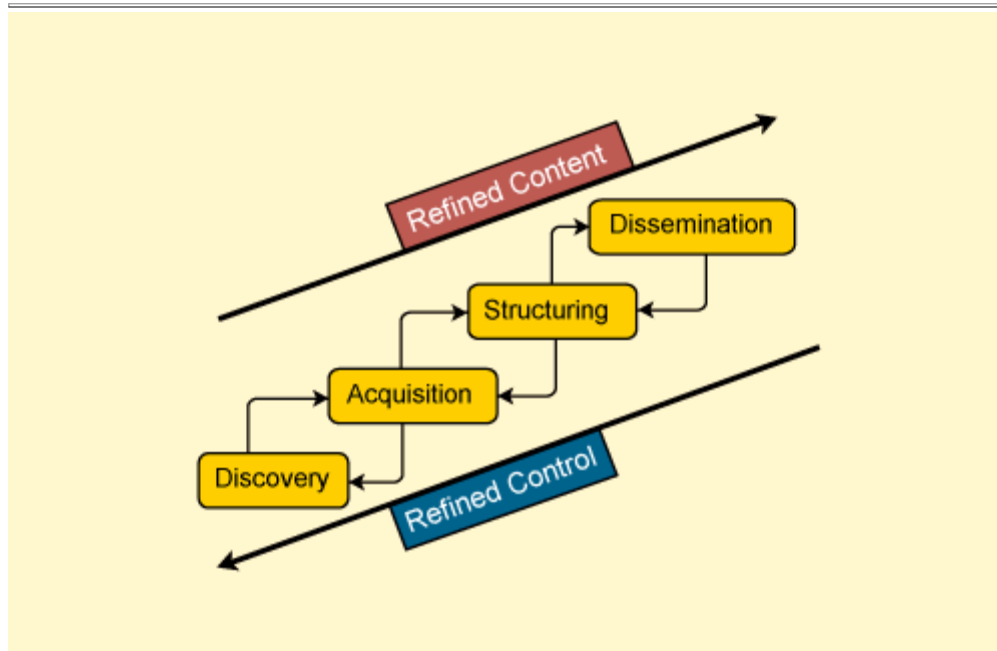
of data warehousing and several successful experiences with data warehousing.

Across the industry, the current practice of data warehousing is fulfilling its promises of business benefits. In retrospect, the current benefits from data warehousing are "low-lying fruit" -- easy accomplishments (relatively speaking) of purging the sins of monolithic legacy systems. Web farming will challenge us with deeper issues concerning information refinement and knowledge management.

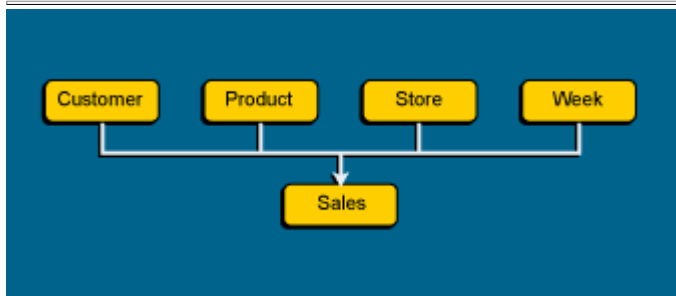
Web farming will be an agent of change (even of a disruptive sort) to the controlled and structured world of data warehousing. This is a necessary change -- a maturing of the basic objectives of data warehousing into a more comprehensive system of knowledge management for the enterprise.



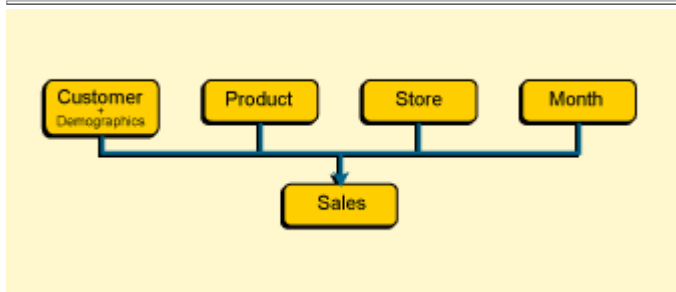
**Figure 1.** Web resources in terms of quality and coverage.



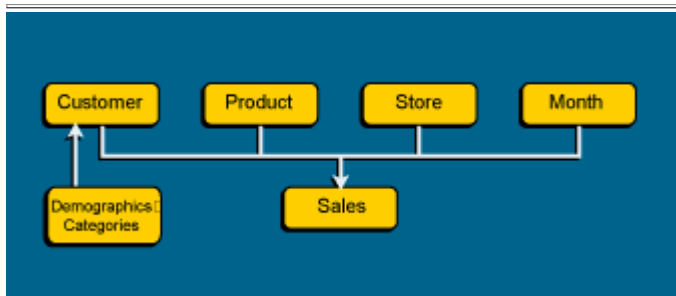
**Figure 2.** Refining information.



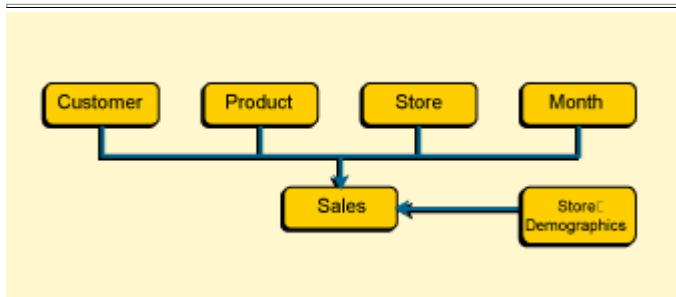
**Figure 3a.** Typical data schema for a sales warehouse.



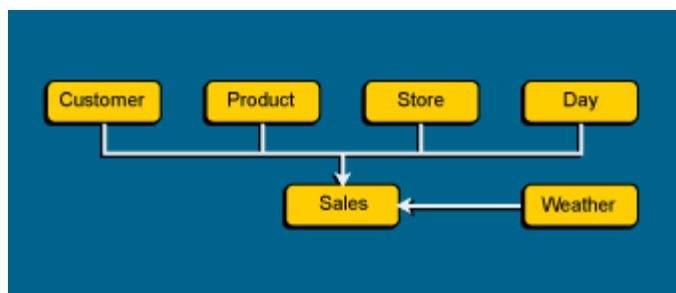
**Figure 3b.** Adding customer demographics - Part I.



**Figure 3c.** Adding customer demographics - Part II.



**Figure 3d.** Adding store demographics.



**Figure 3e.** Adding weather data.

---

Richard Hackathorn is president and founder of Bolder Technology Inc., a firm specializing in enterprise connectivity and data warehousing in Boulder, Colo. You can reach him at [richardh@bolder.com](mailto:richardh@bolder.com). His Web farming resource site is located at [webfarming.com](http://webfarming.com).

*Note: This article is based upon excerpts from the forthcoming book entitled Web Farming for the Data Warehouse to be published by Morgan Kaufmann Publishers this fall.*

---

What did you think of this article? [Send a letter to the editor.](#)

---

[Subscribe to DBMS](#) -- It's **free** for qualified readers in the United States  
[August 1998 Table of Contents](#) | [Other Contents](#) | [Article Index](#) | [Search](#) | [Site Index](#) | [Home](#)

---

DBMS (<http://www.dbmsmag.com>)

Copyright © 1998 Miller Freeman, Inc. ALL RIGHTS RESERVED

**Redistribution without permission is prohibited.**

---

Please send questions or comments to [dbms@mfi.com](mailto:dbms@mfi.com)

Updated July 7, 1998