



Published in the June 1999 issue of *DM Review*.  
Printed from [DMReview.com](http://DMReview.com)

---

## **Farming Web Resources for the Data Warehouse**

By: Richard Hackathorn

*This article explores the use of Web-based information resources as input to data warehouses. This new area is called web farming, which is defined as the systematic refining of Web-based information resources for business intelligence. The intent of Web farming is to enhance the coverage and richness of the data warehouse, motivating its transition to that of an intelligence center or knowledge base for the enterprise.*

**T**he Web has become many things to many people. To those in the data warehousing profession, the Web has been irrelevant or maybe a threat. As a chaotic and unmanageable influence, the Web can be perceived as a threat to the security and tranquility of the warehouse environment.

Although Web technology is used extensively as the delivery mechanism for warehouse data, no one has seriously considered using Web content as input to the data warehouse. The paradigm of the Web is radically different than that of the data warehouse. Adapting an old programming term, one might say that Web content is spaghetti data (i.e., lots of links with little discipline). Web content is highly volatile and diverse, challenging our imagination to discover those nuggets having real business value.

In many ways, the Web is the mother of all data warehouses. The Web is becoming the universal delivery mechanism for global data. However, the immense information resources of the Web are largely untapped by data warehousing systems.

### **Using the Web for Business Intelligence**

Professor Peter Drucker, the senior guru of management practice, has admonished 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 being blindsided by external forces.

The majority of data warehousing efforts result in an enterprise focusing inward, while instead the enterprise should be keenly alert to its externalities. As markets become turbulent, an enterprise must know more about its customers, suppliers, competitors, government agencies and many other external factors. The information from internal systems must be enhanced with external information. It is the synergism of the combination that creates the greatest business benefit.

### **Reliability of Web Content**

Many question the reliability of Web content, as they should. However, few analyze the

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

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.

Think of Web resources in terms of quality and coverage, as shown in Figure 1.

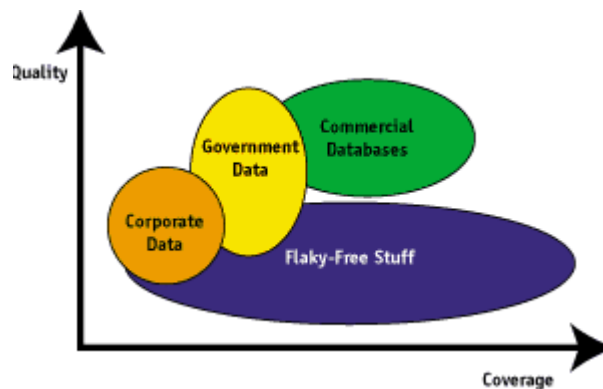


Figure 1: Web-Based Information Resources

Toward the top are information resources of high quality (i.e., accuracy, currency and validity), while resources toward the right have a wide coverage (i.e., scope, variety and diversity). The interesting aspect of the Web is that information resources occupy all quadrants of the figure.

In the upper center, the commercial on-line database 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.

More to the left, governmental 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 EDGAR (Electronic Data Gathering, Analysis and Retrieval) database maintained by the U.S. Securities and Exchange Commission contains extensive information on publicly traded companies and is updated daily.

At the 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-free" 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-free content acts as 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, as shown in Figure 2.

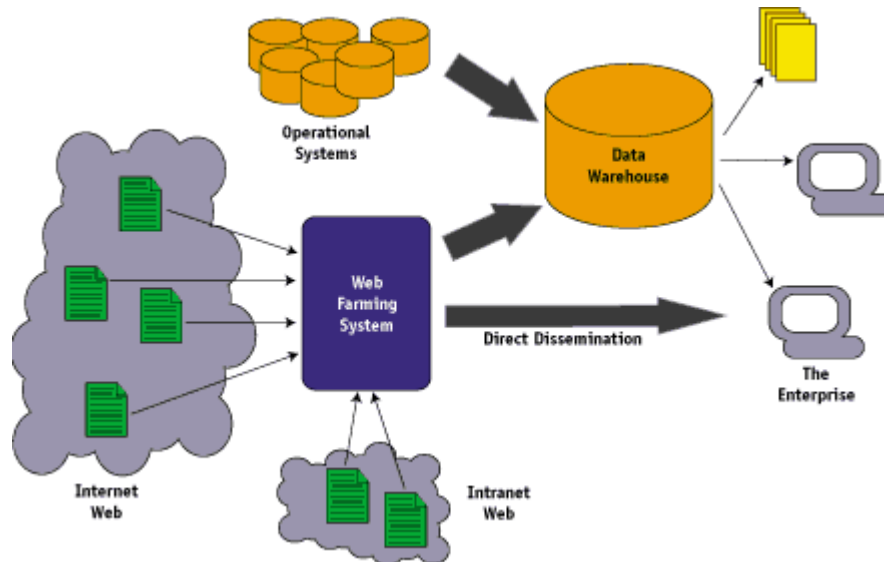


Figure 2: Web Farming System

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.

The primary source of content for the web farming system is the Web because of its external perspectives on the business of the enterprise. As a source of content, the Web can be supplemented (but not replaced) by the intranet web of the enterprise. This content is typically in the format of internal web sites, word processing documents, spreadsheets and e-mail messages. However, the content from the intranet is usually limited to internal information about the enterprise, thus negating 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, a process of refining that information must be performed before loading it into the warehouse. Even in this unrefined state, this information could be highly valuable to the enterprise. The ability to directly disseminate this information may be required via textual message alerts or "What's New" bulletins.

## Refining Information

When a data warehouse is first implemented within an enterprise, a detailed analysis and reengineering of data from operational systems is required. The same is true for web farming. Before Web content can be loaded into a warehouse, there must be a refining of that information.

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

dissemination, as shown in Figure 3.

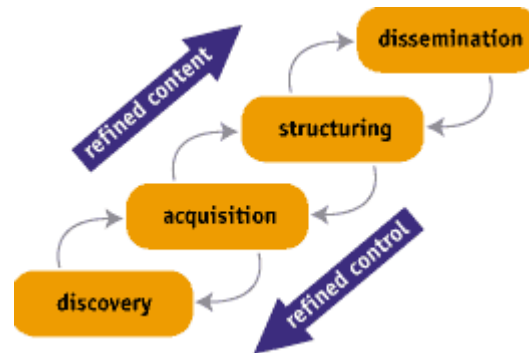


Figure 3: The Four Processes for Refining Information

**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 Alta Vista). Further, the discovery activity must be continuous process since data sources are continually appearing (and disappearing) from the Web.

**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 its past. A mechanism to efficiently use human judgement in the validation of content is another key requirement.

**Structuring** is the analysis 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. A range of dissemination mechanisms is required, from predetermined schedules to ad hoc queries. Newer technologies such as information brokering and preference matching may be desirable.

There is a continuous flow among the process, rather than a step-wise procedure. Further, there is a bi-directional flow among the processes. 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 becomes more selective and discriminating.

### Rendezvous with the Data Warehouse

Let's assume that we have refined (discovered, acquired and structured) some collection of Web-based content so that we have confidence in its validity. The following are various ways in which we can integrate Web content into current warehouses:

1) *Augment the descriptive information about a dimension.* For example, a mailing address can generate a longitude-latitude coordinate, which can retrieve a satellite image of the area around that location.

2) *Add nominal (or ordinal) data about a dimension so that more options for pivoting cross-tabulations are available.* For example, a mailing address can generate a longitude-latitude coordinate, which can classify a customer into a sales region.

3) *Add interval (or ratio) data about a dimension so that correlation analysis with other dimensions can be performed.* For example, a mailing address can generate a census tract ID, which can give family income, household size, population density, age distribution and so on. Or, extensive financial data about publicly traded corporations can be retrieved from the SEC databases.

4) *Create a new dimension table.* For example, recording daily weather as an additional dimension for analyzing sales patterns.

5) *Create a new fact table based on an external event.* For example, a count of the mentions of your product versus competitors' products over week intervals within a set of trade publications.

Consider the table in Figure 4<sup>1</sup>, which lists examples of star schemas for portions of an enterprise data warehouse.

<b>Mart Schema</b>	<b>Fact Table</b>	<b>Dimension Tables</b>
Retailing	Sale	Time, Product, Store, Promotion
Inventory	Item	Time, Product, Supplier, Location
Shipments	Shipment	Time, Product, Customer, Deal, Ship-From, Ship Mode
Banking	Account	Time, Product, Branch, Household, Status
Subscriptions	Subscription	Time, Product, Customer, Status, Promotion
Insurance	Policy	Time, Policy Type, Agent, Coverage, Client, Status
Hospital Procedure	Procedure	Time, Procedure Type, Patient, Hospital, Physician, Assistant, Diagnosis
Frequent Flier	Trip	Time, Customer, Flight, Airports, Fare Class, Sales Channel, Status
Hotel Stays	Visit	Time, Customer, Hotel, Sales Channel, Status
Figure 4: Star Schemas for Different Portions of an Enterprise Data Warehouse		

These examples show some of the ways in which data external from the company can enhance the value of these data marts.

First, all schemas have a time dimension for the period during which the "fact" occurred. Other valuable external information that could be added includes: critical events (economic, political, military, etc.) that happened during that period; the actual weather and weather predictions; holidays and seasonal changes; other events that could affect the flow of commerce. Businesses are not isolated from the effects of natural, social, political and

economic events occurring throughout the world.

Second, most schemas have a product dimension that contains attributes about the company's offerings. Valuable external information to consider includes: recent mentions in the trade press; counting mentions within a set of trade press by week and correlating to sales; links to competitor's products for comparison; prices for same product through alternative distribution channels.

Consider a simple data schema for a sales warehouse as shown in Figure 5. 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.



Figure 5: Sales Warehouse Data Schema

Web farming would be valuable by enhancing the demographics (e.g., quarterly financials) about customers. As shown in Figure 6, by adding information on customer demographics, selective marketing can be performed based upon 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.

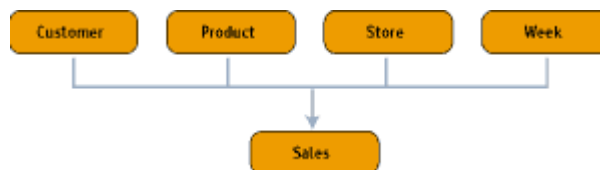


Figure 6: Addition of Customer Demographics

The demographic information is added to the customer dimension so that analyses on specific customers are enhanced by the demographics, as shown in Figure 7. As experience with the demographics matures, data mining techniques can cluster customers into segments based on demographics. Then, demographics can be categorized into meaningful categories and become a separate analysis dimension.



Figure 7: Demographics as a Separate Analysis Dimension

Another example of using web farming to enhance a data warehouse is the addition of demographics on the store, as shown in Figure 8. Using ZIP codes and even the full street address, census data about the communities surrounding the store can be added as another

business dimension. This enhancement can lead to more effective management of stores and to more effective placement of new stores.

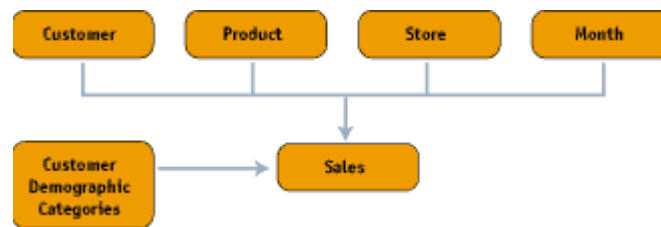


Figure 8: Addition of Store Demographics

A final example involves adding data that is highly volatile, such as weather, as shown in Figure 9. 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.

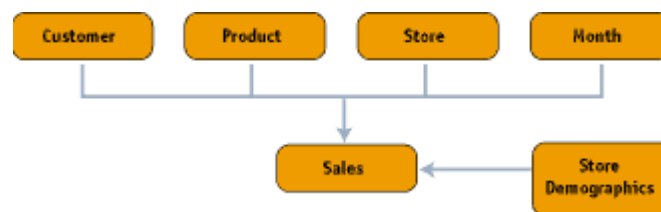


Figure 9: Addition of Weather Statistics

### Where are We Heading?

In many ways, the data warehouse is not a requirement for web farming. You could successfully farm the Web, bypassing the data warehouse and still reap value for the enterprise. However, this approach may achieve short-term gain at the expense of long-term potential. For web farming to be successful in the long-term, it should be integrated into the enterprise data warehouse.

Across the industry, the current practice of data warehousing is fulfilling its promises of delivering real business benefits. With web farming, we are challenged with deeper issues concerning information refinement and knowledge management. Web farming will force change upon the data warehouse. However, this change will evolve the data warehouse into a better system of knowledge management for the enterprise.

Note: Portions are excerpted from the book entitled, **Web Farming for the Data Warehouse**, published by Morgan Kaufmann Publishers (ISBN 1-55860-503-7).

---

*Richard Hackathorn is president and founder of WebFarming.Com, a Boulder Technology company, located in Boulder, Colorado. The firm specializes in design and system integration of web farming systems for corporate clients. He can be contacted at [dick@webfarming.com](mailto:dick@webfarming.com) or through the Web site at <http://www.webfarming.com/>.*