



DB2 UNIVERSAL DATABASE

UPDATED INFORMATION ABOUT THE POWER BEHIND E-BUSINESS

DB2 is now available on the Linux platform

IBM

CLICK HEADLINE FOR MORE



net search:

search

RESULTS: 1 of 1, [www.careercentral.com](http://www.careercentral.com)



# Web Farming

By *Richard Hackathorn*

- [Web Farming Applications](#)

## ***To stay competitive, your business must learn how to exploit the web as an information source for your data warehouse***

The hype surrounding the Web is extreme. Every day brings new revelations about how the Web will bring wonderful benefits to our society and to your company. The current size of the Web is overwhelming. The Internet Archive, a nonprofit organization that archives the Internet for historical purposes, began archiving the entire Web in early 1996. As of March 1999, the Internet Archive has captured more than 13TB of text content, with a new snapshot taken every 30 to 60 days. The total number of web sites is doubling every six months. Each web site contains a median of 300 pages, of which 50 sites have more than 30,000 pages each. The Web is currently estimated to have a total of 80 million pages with an average age of only 44 days.

As an information resource, the Web is the mother of all data warehouses. Its information content is rapidly exceeding the total capacity of traditional libraries and commercial databases. Moreover, its content is often more current.

In a PricewaterhouseCoopers (PWC) Trendsetter Barometer survey last September of more than 400 CEOs of fast-growing companies, the Web was cited as a major source of business intelligence (BI). Of the high-tech companies questioned in the survey, 82 percent use the Web as a “source of competitor information,” while 73 percent use the Web as “a statistical or data resource.” In addition, 68 percent use the Web to “obtain new sales leads” as compared to only 27 percent who use it to “provide direct sales of products.” It would seem that these fast-growing companies use the Web more for BI than for e-commerce!

Unfortunately, in most companies, use of the Web for BI tends to be unproductive. Business professionals surf the Web individually, bouncing from one item to the next, only occasionally finding something of value to the business. There are long periods of frustration interspersed with a few moments of elation. When they come across something of interest, they paste the item into an email or memo, which is shared on a limited basis, certainly not managed, and soon forgotten.

The notion of “web farming” — effectively gleaning and managing business information from the Web — addresses this problem within the context of enterprise architectures. Web farming is defined as the systematic refining of Web-based information for business intelligence. For the IT professional, the term *systematic* should imply secure data centers and managed data warehouses. The objective of web farming is to enhance the data warehouse by integrating externally derived information with data derived from internal operational systems. Although this approach has similarities to data warehousing today, it requires some new skills and several new technologies, including XML structuring, linguistic analysis, and information visualization.

What are the business benefits of web farming? As with data warehousing, the benefits are dependent on how well you enable the information farmed to affect your basic business processes. A mistake many companies may make is investing in web farming technology without linking to the individuals who make a difference and to the business processes that make the profit.

When implemented correctly, a web farming solution yields several benefits. The information farmed can improve the performance of a business activity, such as having better information that will enable you to service a customer faster and more effectively. The information farmed can also change the workflow of a business activity, such as redesigning the customer call center. Furthermore, the information farmed from the Web can lead to the creation of new activities, such as initiating a frequent customer program.

### **Where Is the Value?**

The Web has rapidly become the primary distribution medium for external business information. As the PWC survey indicated, four out of five emerging companies have recognized this fact. However, most people still believe that web-based content is of little value because it is free. Validity, reliability, and accuracy are not characteristics that most people associate with web content.

However, this image of web content is not justified. For each dubious web site, there are many others with highly current and accurate content. Fee-based content with guaranteed validity is becoming more available — and at reasonable prices.

In fact, web content is often cleaner and better organized than the data in your internal operational systems. After all, the data in your operational systems was designed for Cobol programs to support a specific function. As many warehousing efforts have documented, internal data cleaning and validation is a major task that can result in a warehouse of limited usefulness if not done properly.

Web content is driven dynamically from databases (as opposed to static HTML pages). In addition, when technologies (such as XML/RDF) become more widely adopted, you’ll be able to discover and deliver this content more efficiently and reliably.

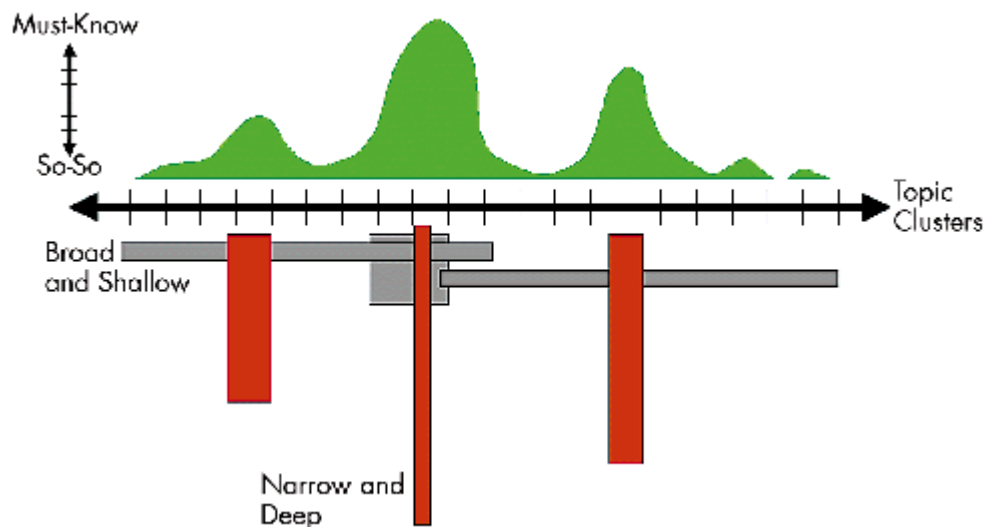
Web technology has shifted the economic trade-offs so that publishing web content from existing databases is simple and cost effective. The major growth areas

for business information are government databases, commercial content providers, business news agencies, and corporate web sites.

### What to Farm

In contrast to *web mining*, the term *web farming* doesn't imply that valuable information exists somewhere on the Web, just waiting to be found and immediately used. Instead, it implies that hard work is involved to prepare the field, seed the crops, cultivate the soil, and then finally harvest the crop. The value of web farming comes from applying effort over time and with patience to the information resources of the Web. Cultivating a few seeds of data will eventually produce a harvest of information.

If you are farming the Web's information resources, what specific information from that huge vastness should you farm? Obviously, just as for any other aspect of your enterprise information architecture, you'll want to concentrate on those information clusters that currently do or potentially can have the most bang for your business's buck.



**Figure 1 Plotting information clusters by importance and availability.**

Figure 1 shows a sample plotting of the various clusters or topics (along the horizontal axis) according to their importance (along the vertical).

If you draw a similar plot chart for your organization, most of the possible information sources will be external to your enterprise and available in some form via the Web. Some examples of the hot business drivers for most enterprises today are the following information areas:

- Customer relationships — understanding the current needs of your customers and anticipating their future needs
- Supply-chain management — managing your value chain from suppliers through distributors to end customers
- Competitive analysis — understanding and monitoring established competitors within your market and detecting emerging competitors
- Technology trends — tracking developments in key technologies to your business and forecasting market impacts of these new technologies
- Deregulation — understanding current and future economic impacts of industry deregulation and exploiting opportunities emerging from the deregulation

- Global economics and politics — monitoring trends and critical events worldwide for threats to existing operations and for opportunities in future markets.

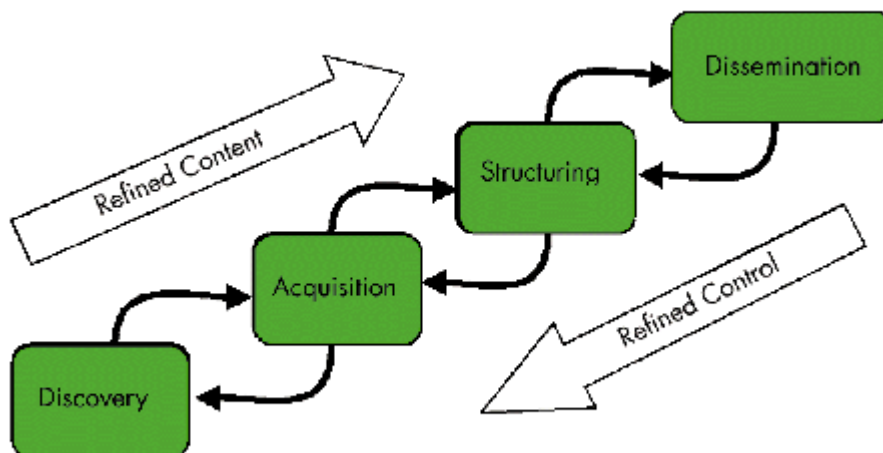
Most of this critical information exists outside of your enterprise. It is not recorded within your accounting files or stored within your warehouse. For most enterprises, the primary sources of external information are marketing reports, industry analyses, *The Wall Street Journal*, the nightly news, and occasional tips from colleagues. Although these sources may have served you well over the years, today's turbulent markets demand more. We suddenly have volumes of external information available to us through the Web, and we must integrate all of our external information with internal information.

There are existing systems that support broad ranges of topics but only to a limited extent. For example, news feeds of press releases and commentaries on certain companies is a broad yet shallow source. The corporate library is also a classic broad yet shallow resource (except when a research library does a special investigation into a well-defined topic).

Your weaknesses are exposed when you need focused, in-depth coverage. These few places are the web farming systems' targets. For example, one of our clients had an urgent need to investigate low-cost tire pressure sensors. Our investigation identified the content creators as research projects in certain universities and companies that are designing and manufacturing a product. A web farming system would then take various links to these content providers and systematically monitor content changes. As experience with a content provider permits, specific data elements would be extracted, validated, transformed, and loaded into a data warehouse.

### Refining Information

The paradigm of web content is radically different from that of the data warehouse. Adapting an old programming term, you might say that web content is *spaghetti data* — it links everywhere with little discipline. Dealing with web content requires disciplined refining of information, transforming raw data into validated information. Although there are similarities to designing data warehousing systems, the discipline of refining web information has some unique requirements.



**Figure 2** The process of refining Web information.

As Figure 2 shows, this discipline consists of four processes:

- Discovery — the exploration of available web resources to find those items that relate to specific topics. Discovery involves considerable detective work that goes beyond searching generic directory services (such as Yahoo) or indexing services (such as AltaVista). The goal is to locate individuals and organizations that create content important to your business. Because news sources are continually appearing on the Web, discovery must be a continual process.

- Acquisition — the collection and maintenance of content identified by its source. The main goal of acquisition is to maintain the information's historical context so you can analyze content in the context of its past. A key requirement is to efficiently use human judgment in the validation of content. For example, a person must be able to review incoming content and record a judgment on its validity.

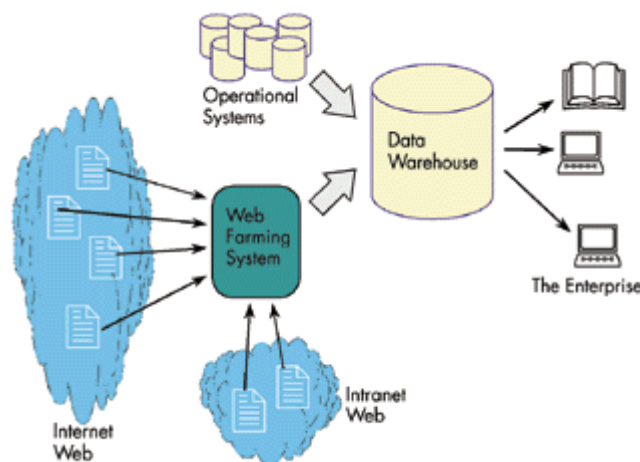
- Structuring — the analysis and transformation of content into a more useful format and into a more meaningful structure. The formats can be web pages, spreadsheets, word processing documents, and database tables. As you move toward loading data into a warehouse, unstructured textual data from the Web must be refined into tabular numeric data in order to support critical business analysis.

- Dissemination — the packaging and delivery of information to the appropriate consumers, either directly or through a data warehouse. An effective web farming solution must support a range of dissemination mechanisms, from predetermined schedules to ad hoc queries. Information portals, content brokering (publish and subscribe mechanisms), and personalization through preferences and usage matching are important components of effective dissemination.

Content flows among these processes bidirectionally. 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.

### Putting It All Together

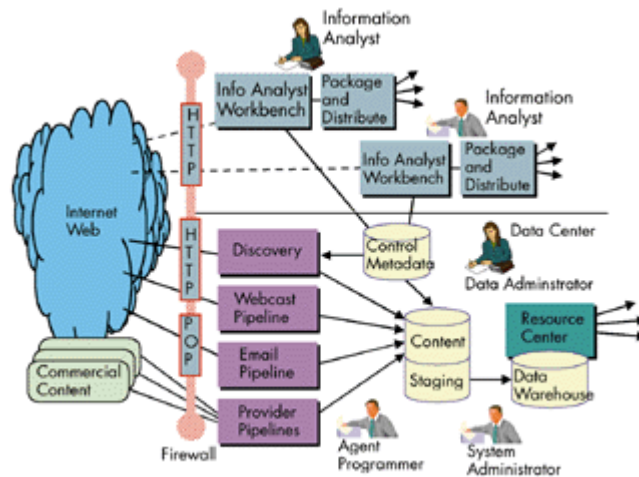
As Figure 3 shows, the data warehouse occupies a central position in a web farming system's information flow. Like operational systems, the web farming system provides input to the data warehouse. The result is to disseminate the refined information about specific business subjects to the enterprise.



**Figure 3** The data warehouse as the center of a Web farming process.

The primary source of content for a web farming system is the global Web. This source can be supplemented (but not replaced) by an enterprise's intranet. Intranet content is limited to internal information about the enterprise, such as internal web sites, word processing documents, spreadsheets, and email messages.

Regardless of its source, most information acquired by the web farming system will not be in a form immediately suitable for incorporation into the data warehouse. It will either be unstructured hypertext or unverified tabular values. In either case, you must perform the refining process before loading the information into the warehouse.



**Figure 4 The players in a web farming process.**

As Figure 4 shows, a robust web farming system spans a variety of roles:

1. One or more information analysts control the activities of web farming through the discovery, acquisition, structuring, and dissemination processes. The information analysts will occasionally probe the Web and distribute information to specific individuals; however, their primary focus is on activities within the data center, which is where most processing is performed. The databases for control metadata, content (in various stages of refining), a staging area, and the data warehouse are all managed environments.

2. The person programming agents creates algorithms for searching and structuring web content based on the control information accumulated by the analysts.

3. The data administrator designs and supervises the web content's flow into the data warehouse.

4. The system administrator ensures the security and reliability of the overall systems.

### Stages of Web Farming

For a successful web farming system implementation, I recommend a four-stage methodology. Each stage builds upon the previous, with the goal of integrating the web farming activity into the data warehouse and eventually into the business intelligence system for the enterprise.

**Stage 1 – Getting Started:** In this stage you establish the business case for web farming based upon the enterprise's objectives and market. You document the critical

external factors (CEF), formulate a discovery plan, identify content providers, disseminate initial information, and compile the business case. It is important to describe clearly those external factors that will have an impact (positive or negative) on the business. For instance, the price of oil has a major impact on an airline business. Monitoring any factor related to oil prices (such as OPEC meeting proceedings) would be important.

**Stage 2 – Getting Serious:** You must legitimize the web farming activity within the organization and create the infrastructure for reliable production operations. You should secure approval of both the budget and staffing plan, build the infrastructure within the data center, refine the CEF list, maintain historical context, and establish an intranet web site.

**Stage 3 – Getting Smart:** This is where you exploit technology (especially for discovery and structuring of information) and build pipelines to primary content providers. You implement selection and extraction filters to tap into various content providers, analyze and structure content, and publish content. The selection and extraction filters are search strategies for finding relevant data and the algorithms for extracting that data from the source sites.

**Stage 4 – Getting Tough:** Here you structure information for the data warehouse by revisiting the business objectives in light of the warehouse schema. In this final stage, you rendezvous with the warehouse, link to other systems, resolve entity mapping, and establish credibility checks.

### **Web Farming in a DB2 Environment**

Currently there is no single product available that provides the complete range of functionality required for web farming systems. Just as in the early days of data warehousing systems, there are many pieces (some of which are quite good), but no complete solution. The pioneers in web farming will initially be forced into extensive integration and development projects before the vendors react to this emerging market segment.

IBM is well-positioned to support web farming. Several tools powered by DB2 are currently available to get you started with web farming. The Intelligent Miner for Text (IMT) and KnowledgeX products include functionality for extracting key concepts and patterns from text documents and organizing the information into a schema that fosters insights and action.

IMT has an impressive array of linguistic functions, including feature extraction, summarization, classification, clustering, and language identification. Feature extraction is necessary to prepare a text document for further analysis. Each word in a document is recognized from a vocabulary, expanded from its abbreviation, reduced to its root form, combined into multiword phrases (such as “computer hardware,” which is different from both “computer” and “hardware”), and categorized into persons, organizations, places, time periods, and so on.

IMT also includes search engine and web crawling capabilities. The search engine allows for easy search of objects and documents plus the creation of summaries, and the automatic discovery of multiword phrases. It also includes knowledge discovery functions such as “clustering” of objects of similar traits and categorization. In other words, IMT can sort text pieces by content according to a user-defined and system-trained taxonomy. The documents can be in many of the popular sources and formats and are accessed through the use of filters.

An interesting feature of IMT’s search engine is its ability to limit its scans to

user-defined sections of documents, which is useful for web pages. For example, you could choose to search only against the headings or summary sections of pages.

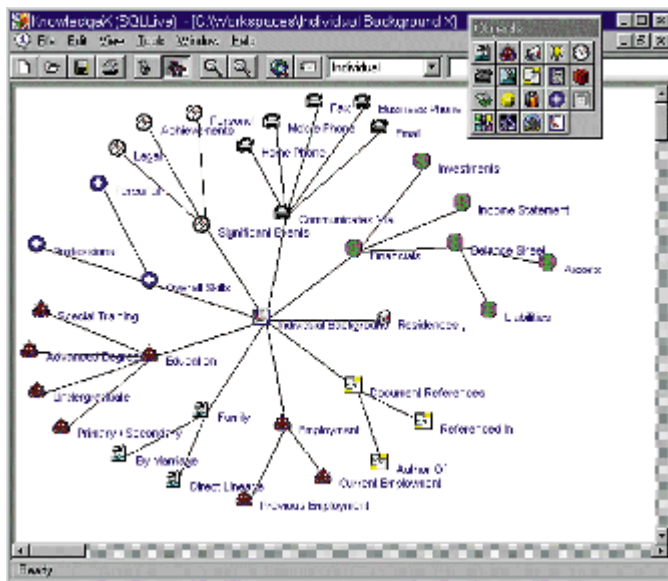
The web crawler stores the collected metadata in DB2 and dumps the pages (you just want to keep the links, not the contents, once indexed). IMT currently supports DB2 on NT, AIX, S/390, and AS/400, with support for Linux to follow shortly.

Acquired by IBM last summer, KnowledgeX is positioned as an application for “acquiring, discovering, publishing, and distributing knowledge across the enterprise.” It graphically reveals relationships among knowledge objects: documents, newsfeeds, people, organizations, and it can use DB2 UDB as its data warehouse. Its graphical capabilities make it easy to define, manipulate, establish relationships, and query knowledge objects.

KnowledgeX is a knowledge organizer that gathers information, sets contexts defined by objects and their relationships, and discovers knowledge that can be distributed across an enterprise. The objects, such as people, companies, and documents, can be organized according to a taxonomy or network of relationships. Each object plays a definite role in each relationship, and complex queries such as “who of the directors that sits on the board of one of our subsidiaries sits also on the board of a competitor or its subsidiaries” can be built and executed through simple drag-and-drop procedures.

KnowledgeX offers two user interfaces, one for knowledge producers (such as market analysts) and another for knowledge consumers (such as executives, traders, and brokers). The interface for knowledge producers works according to a client/server protocol; the interface for knowledge consumers is a web browser interface showing categories of pieces of information. The knowledge producers build the objects and relationship diagrams, known as knowledge maps.

Those knowledge maps are represented in the web browser interface as categorized topics following a hierarchical relationship. The interface for knowledge consumers allows knowledge workers to access multiple documents organized in a category schema or taxonomy. This product is an excellent example of information visualization focused on surfacing relationships among objects.



**Figure 5 Collecting information in KnowledgeX.**

Figure 5 (page 51) shows a template for collecting structured information about the background of a person. Note the use of icons to distinguish object types (people, phone numbers, money, and so on).

Other non-IBM products available today to support web farming in a DB2 environment include:

- Search tools: Web Compass from Quarterdeck, Copernic from Agents Technologies, BullsEye from IntelliSeek, and Enfish Tracker Pro from Enfish Technology
- Text indexing and web crawlers: AltaVista Search from Compaq, Compass Server from Netscape, Search'97 from Verity, SearchServer from PCDOCS/Fulcrum, Webinator from Thunderstone, SmartCrawl from Inktomi, and UltraSeek from InfoSeek
- Knowledge managers: Knowledge Server from Intraspect, Dataware KMS from Dataware, and Agentware i3 from Autonomy.

Although there is currently no complete solution for web farming, I predict that over the next 18 months vendors will recognize web farming as a critical market segment and introduce products specially targeted for web farming. The main beneficiaries of this technology will be industries with rapidly changing markets caused by deregulation, global competition, shifting technologies, and the like. Such industries will be forced to integrate web farming into their data warehousing systems to remain competitive. Doing business as usual is no longer a viable strategy. •

*Note: Portions of this article are adapted from the author's recent book, Web Farming for the Data Warehouse (Morgan Kaufmann, 1998).*

**Richard Hackathorn** is president and founder of WebFarming.Com, a Bolder 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 [www.webfarming.com](http://www.webfarming.com).

## Web Farming Applications

Here are examples of how web farming systems can be used:

Electricité de France, the major electric utility in France, was interested in public opinion about electric vehicles to promote the use of electricity for transportation. With help from IBM France, the press articles from 1992 to the present were analyzed using a thematic text-analysis tool. They concluded that the early journalistic messages were negative, emphasizing the technical difficulties and political issues. However, the media shifted after 1994 to the positive, with messages about advantages, urban ecology, noise-reduction, and an ideal form of transportation in general. As a result, the utility was better able to make decisions regarding partnerships with automobile manufacturers and promotional campaigns for electric vehicles.

A vendor of high-end ERP software packages wanted to improve their customer relationships. To do this, the vendor has linked its DB2 data warehouse to a commercial information provider and is able to monitor timely information about important customers. The data items it monitors include trademarks, patents, litigation, broker reports, Internet domains, SEC filings, stock fluctuations, insider trading, news items, job postings, and USENET discussion forums. This information is captured hourly, transformed into an XML vocabulary, and delivered to the warehouse via a secure web link. Unusual events about customers are also sent via email or pager to the

proper sales representatives. The marketing staff reviews the entire collection periodically to predict sales trends and target marketing campaigns.

A manufacturer of home computers is scanning the web sites of 10 major retail distributors on an hourly basis for pricing data. The prices of its models, along with the prices of competitive models, are extracted from the web site and loaded into the company's data warehouse. Simple statistics of average, minimum, and maximum prices over time are generated. Also performed are correlation analyses with critical events such as advertising campaigns and price reductions. Over time, insights into the market dynamics of actual sales pricing as impacted by marketing efforts were emerging. For example, the price decay of released models was monitored so that their prices remained competitive but at the highest amount.

**Resources:**

**PricewaterhouseCoopers:** [www.pwcglobal.com](http://www.pwcglobal.com)

**Intelligent Miner for Text:** [www.software.ibm.com/data/iminer/fortext/index.html](http://www.software.ibm.com/data/iminer/fortext/index.html)

**KnowledgeX:** [www.software.ibm.com/data/km/knowledgex](http://www.software.ibm.com/data/km/knowledgex)

**World Wide Web Consortium:** [www.w3.org/TR/REC-rdf-syntax](http://www.w3.org/TR/REC-rdf-syntax)



Copyright © 1998 [Miller Freeman, Inc.](http://www.millerfreeman.com) All Rights Reserved

Redistribution without permission is prohibited.

Please send questions or comments to [editor@db2mag.com](mailto:editor@db2mag.com)