



**FREE!**



engineer.com  
mail.com  
journalist.com

**Byte**



**ARTICLES BYTEMARKS FACTS HOTBYTES VPR**

**TALK**

**Feedback**

Write to Byte

**Newsletter**

Sign Up Now

**BYTE Categories**

- Columns
- Features
- Audio

**Print Archives**

By Issue By Topic

**Search BYTE:**



**Resources**

History Of Byte:  
Part I Part II Part III

- Java Books
- More Java Books
- Java One Audio Report

**BYTE Humor**

Ian Shoales' Page

**About Us**

- Byte Editorial Staff
- Feedback
- Sales Staff
- Privacy Policy

**Techweb Sites**

- CMPmetrics
- Data Communications
- File Mine



# Farming the Web

[October 1997 / Core Technologies / Farming the Web](#)

*The Web's content can be harvested for information that's crucial to making strategic decisions.*

*Richard Hackathorn*

The Web and data warehousing (DW) are a powerful combination. Publishing warehouse data via the intranet has become a highly productive approach. By generating dynamic pages from Web-enabled databases, whole new areas of data analysis are supported. No one, however, has seriously considered putting content from the global Internet into the data warehouse. Web content is considered too unreliable, and data external to the organization is often considered to have little business value.

But I would argue to the contrary. As markets become turbulent, the old way of doing business with data only from internal operational systems becomes less relevant. A company must know more about its customers, suppliers, competitors, and government agencies than ever before. Much of this external data is readily available on the Web. The challenge is to wade (with big boots) through the Web, discovering and acquiring those pieces that do have an impact on the business.

The emerging area that is concerned with this challenge is called Web farming (WF). WF is the systematic discovery and acquisition of business-relevant Web content as input to the data warehouse. It has three goals. First, to discover and acquire Web content that is highly relevant to the business

InformationWeek  
 InternetWeek  
 Network Computing  
 Planet IT  
 TechShopper  
 TechWeb News  
 Tele.com  
 WebTools  
 Winmag.com

**Free E-mail**  
 Sign Up Now

Sponsored by:



**TechWeb Sites**

Byte.com  
 CMPmetrics  
 Data Communications

acquire web content that is highly relevant to the business. Second, to structure that data so that it becomes an integral part of the existing data warehouse. Third, to accomplish this in a systematic manner that evolves into a production system. WF must deliver information of value to the business, to the right people at the right time. This is the same objective as the data warehouse. Hence, WF and DW should be closely integrated.

**Getting Started**

The first level of WF documents the external factors that affect the business, and predicts the potential factors that will affect it in the future. Possible avenues of investigation are: analysis of recent company reports and press releases; critiques of your company by news and investment analysts; and observations of typical customers performing transactions. Then, compile a detailed, hierarchically organized list of these external factors. Prioritize the list based on the potential impact (either positive or negative) of each factor upon the business.

Formulate a systematic plan for searching the Web for relevant information, starting with the highest-priority factors. When a useful item is found, format and package it as memo, report, spreadsheet, chart, presentation, or e-mail. Immediately disseminate it to the people who should have a keen interest in it. Then, track the reactions to this information.

In the first level, you're building the foundation for determining what is important to the business. The principal cost item should be a highly skilled business analyst who has a solid understanding of the business. This level should be implemented quickly and cheaply, with feedback expected in one or two months. The end result should be documentation of the business factors associated with an organized list of URL bookmarks.

**Getting Serious**

The second level of WF requires a serious management commitment of resources to pursue WF as a means of expanding coverage for the data warehouse. Its objective is to establish the WF infrastructure within a secure server environment. Under the umbrella of the DW group, the data within the existing data warehouse should be supplemented by expanding its coverage of those external factors impacting the business. The second level involves the transition from a self-contained workstation to a secure

File Mine  
InformationWeek  
InternetWeek  
Network Computing  
Planet IT  
TechShopper  
TechWeb News  
Tele.com  
WebTools  
Winmag.com

server environment, as shown in the figure "Web Farming, Levels 1 and 2." On the client side, the number of analysts should increase as demand of packaged information from Web content increases.

The important changes occur on the server side of the architecture. A database shared among the analysts manages the Web content and various control information such as favorite bookmarks, useful searches, and the like. Data center staff should administer the WF server. Besides the sharing of common data among the analysts, the server takes on the active role of periodically probing those Web pages identified as important. As useful information becomes available on Webcasting channels, e-mail feeds, and newsgroups, you should implement filters to capture, filter, and format that data into the WF server.

### **Get Smart**

The third level of WF builds upon the previous infrastructure to increase the relevance of Web content to your business. Its objective is to get smart about discovering and acquiring new information, and about distributing it. This focus occurs in two places. First, the information acquisition is expanded with intelligent Web searching and with custom information providers. Second, the information distribution is expanded enterprise-wide through the implementation of the publish and subscribe (P&S) mechanism (as shown in the figure "[Web Farming, Levels 3 and 4](#)").

At this level, the objective is to transform the content database into a full-function intranet Web site that serves as a custom resource center for the entire company. The goal is to shift over time from static content of digested Web pages to dynamic content generated from warehouse tables.

Another change is the adoption of a WF workbench environment for analysts. Controlled via a common database, the workbench integrates the browser with other tools, such as linguistic analysis and information visualization. The workbench should increase the productivity of the analysts to discover relevant information. Using P&S, specific channels of information related to important business topics are published. Various people (and applications) can then subscribe to these channels to receive a flow of information on a continuing basis. Finally, you should contract custom information providers to supply reliable data via efficient links using, for example, the Internet Interoperable ORB Protocol (IIOP).

### **Getting Dirty**

The fourth level of WF refines the transformation of Web content into structured data for the DW. As in the previous levels, the WF activity characterizes the business relevance of Web content and establishes the infrastructure to use it.

This level's objective is to exploit the business potential of Web content as input to the data warehouse. Now comes the dirty work of structuring Web content into the proper format. The challenge is twofold: First, adding a reliable time dimension to the detailed facts. Second, linking into the proper fact or dimension tables in the data warehouse. The most frequent application will be augmenting an existing dimension table with an additional attribute. However, the most potential comes from creating new fact tables that allow exploration of external business factors.

Here are some suggestions on how to proceed: Investigate the current data warehouse. Obtain the schema definition. Understand the major fact tables and key dimensions for those tables. Dump some typical data on the main tables. Compare the list of business factors to the warehouse schema. Note the gaps. Next, consider how external data would fit into the schema. Decide if attributes for existing dimensions should be augmented or if new dimensions for existing tables should be added. Finally, prioritize specific business factors that have the greatest potential for extending coverage for the data warehouse.

### **Looking Externally**

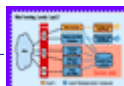
As companies look externally for their next competitive advantage, WF will become a necessary function of all DW systems. Content providers will have an economic incentive to supply reliable and quality information that is prestructured into generic warehouse schemas.

WF requires a new set of skills. It also requires an expanded infrastructure for networking and DW. Both require time to evolve into a production system. It will all come together if you work through the four WF levels I've described.

---

## **Web Farming, Levels 1 and 2**

[illustration link \(15 Kbytes\)](#)



*These levels determine the feasibility of Web farming and build its infrastructure.*

---

### Web Farming, Levels 3 and 4

[illustration link \(19 Kbytes\)](#)



*These levels build the operation into an intranet Web site and integrate it with the data warehouse.*

---

**Dr. Richard Hackathorn ( [richardh@bolder.com](mailto:richardh@bolder.com) ) is president and founder of Bolder Technology, Inc. (Boulder, CO). This article was extracted from a forthcoming book from Morgan Kaufmann Publishers. You can find a resource center for Web farming at <http://www.bolder.com/>.**



|                |
|----------------|
| engineer.com   |
| mail.com       |
| journalist.com |



Byte

**BYTE** ARTICLES BYTEMARKS FACTS HOTBYTES VPR  
**TALK**

**Feedback**

Write to Byte

**Newsletter**

Sign Up Now

**BYTE Categories**

- Columns
- Features
- Audio

**Print Archives**

By Issue By Topic

Search BYTE:

search

**Resources**

History Of Byte:  
Part I Part II Part III

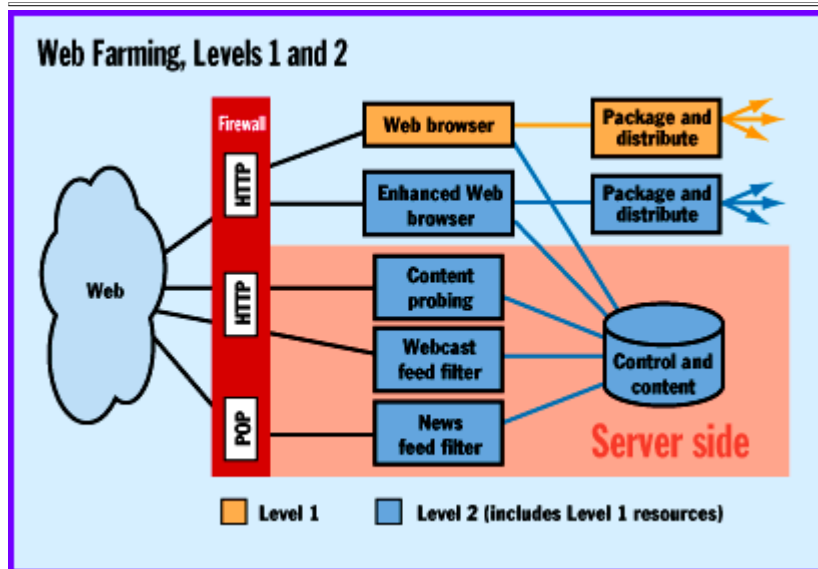
Java Books  
More Java Books  
Java One Audio Report

**BYTE Humor**  
Ian Shoales' Page

**About Us**  
Byte Editorial Staff  
Feedback  
Sales Staff  
Privacy Policy

**Techweb Sites**  
CMPmetrics  
Data Communications  
File Mine  
InformationWeek  
InternetWeek

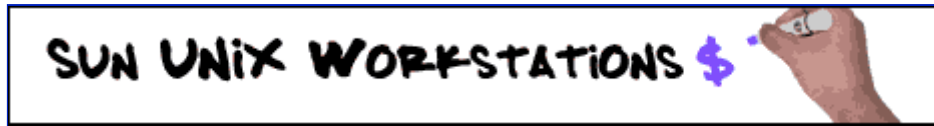
## Web Farming, Levels 1 and 2



*These levels determine the feasibility of Web farming and build its infrastructure.*



**Up Level Search Subscribe**



Byte



ARTICLES BYTEMARKS FACTS HOTBYTES VPR

Feedback

Write to Byte

Newsletter

Sign Up Now

BYTE Categories

- Columns
- Features
- Audio

Print Archives

By Issue By Topic

Search BYTE:



Resources

History Of Byte:  
 Part I Part II Part III

- Java Books
- More Java Books
- Java One Audio Report

BYTE Humor

Ian Shoales' Page

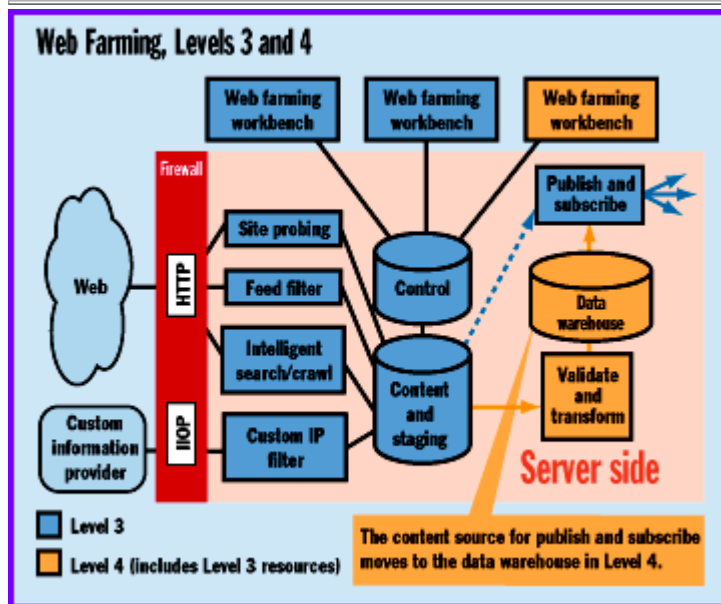
About Us

- Byte Editorial Staff
- Feedback
- Sales Staff
- Privacy Policy

Techweb Sites

- CMPmetrics
- Data Communications
- File Mine
- InformationWeek
- InternetWeek
- Network Computing

### Web Farming, Levels 3 and 4



*These levels build the operation into an intranet Web site and integrate it with the data warehouse.*



Up Level Search Subscribe