



# UNIVERSAL Data Models for Clickstream Analysis

By Len Silverston

The seven sages of ancient Greece advised, “Know thyself.” Successful enterprises not only need to know themselves but know about their prospects, customers, employees, suppliers and other involved parties. Web-site activity analysis or clickstream analysis can provide valuable insights to help enterprises know more about the parties with which they do business and thus act more proactively toward meeting their objectives. Specifically, effective clickstream analysis can reveal:

- Who is potentially interested in your offerings?
- What products are of interest to visitors?
- Where are the sources of your referrals?
- When (which season or time of day) are people most likely to be interested in your offerings?
- What are the patterns of buying? Do people tend to review new releases or references before perusing product offerings?

## Why Reinvent the Wheel?

Clickstream analysis is one of the most universal applications today. All enterprises with a Web presence have access to Web logs that capture Web-site activity. Thus, they have similar information on visitors, referrers (entities that link visitors to your site), user logins, product offerings, Web content, server hits and visits. While enterprises have this information, few take the time to manage and effectively utilize

this information; and even fewer use models for integrating this information.

Why not take advantage of models that have already been developed to manage and integrate this information? The value in using Universal Data Models is that they can point out common pitfalls, provoke thought and provide effective ideas. This article will provide samples of clickstream Universal Data Models (these models are based upon models found in *The Data Model Resource Book Series*, Wiley, 2001).

## Integrating Clickstream Information

While many find value in analyzing clickstream information on its own, there is far greater power in integrating this information. For example, if the enterprise integrates customer Web interactions so they become part of each customer’s profile, there is more complete and accurate customer information. The models presented in this article enable data integration.

## Who is Involved in Clickstream Interactions?

Roles within Web activity include visitors, referrers, Webmasters, ISPs and subscribers, as well as many of the traditional roles that interact on the site such as prospects, customers and employees. Data models often show unrelated entities for each of these roles, for example, by showing separate CUSTOMER, VISITOR, REFERRER and PROSPECT entities. Because each of these entities may maintain similar information such as contact information or demographics,

this information may be maintained redundantly for each role that a person or organization plays.

Furthermore, unrelated “role” entities create obstacles in viewing complete profiles for people or organizations. For example, if the Web-site activity of a visitor is maintained separately from customer contact activities, the enterprise loses the ability to easily view the overall activity for parties that are both visitors and customers.

The model shown in Figure 1 shows each PARTY may be acting in one or more PARTY ROLES. Instead of maintaining separate entities for each role (i.e., CUSTOMER, PROSPECT, VISITOR), these are PARTY-ROLE subtypes. Thus, if Bill Jones is a visitor, subscriber, prospect and customer, there would be a single PARTY instance for Bill Jones and four relationships to PARTY ROLE indicating his various roles. The result is more complete, consistent and accurate information about parties.

## Are PERSON and ORGANIZATION the only PARTY Subtypes?

One reason that PERSON and ORGANIZATION are subtyped within PARTY is that many business transactions involve either an individual (PERSON) or a group of individuals (ORGANIZATION) who are parties within various transactions. For instance, orders, agreements, shipments, invoices, work assignments and many other transactions may be related to either people or organizations.

A prevalent Internet transaction is a Web visit. Very often, visitors of

Web sites are not people or organizations, but automated agents such as Web servers, FTP servers and spiders. The AUTOMATED AGENT subtype in Figure 1 maintains information about computerized Internet entities that may have several roles such as VISITOR or HOSTING SERVER. Hence, if it was desired to determine who visited an enterprise's site, the "who" could be an automated agent.

### A Powerful Product-Object Model

Have you ever noticed that there are often differences in verbiage, descriptions and images of an enterprise's products depending on the medium used to present the information? For instance, the Web-site con-

tent may use a different product image than a brochure, proposal or presentation. This often leads to inconsistent sales messages.

Let's suppose that a computer retailing organization had a product image GIF file for their Ultra 1000 computer on their site. Figure 2 shows a data model with a WEB CONTENT entity that maintains electronic objects stored on Web sites such as product images or product descriptions. This model maintains the text, image or other content for each part of the Web site. Thus, if the enterprise knows which Web contents are being frequently accessed, it can infer what products are of most interest.

Even though the intent of this

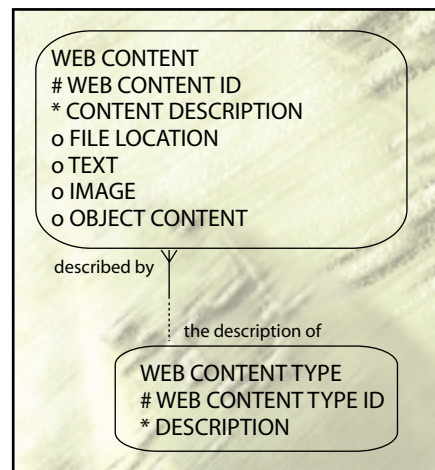


Figure 2: Simplified Web-Content Model

image was to be used specifically for this Web site, isn't it possible that the image could be used in a brochure or within marketing campaign literature? These electronic objects ideally should be managed by their own OBJECT entity, allowing the same OBJECT to be used many different ways and in many circumstances.

Figure 3 shows a model with an intersection entity, OBJECT USAGE, showing which OBJECTS are used in which WEB CONTENT. WEB CONTENT now represents a placeholder of content, pointing to the appropriate object(s) that it references. The model shows that each OBJECT may be used to refer to more than one PRODUCT, and each PRODUCT may be portrayed via one or more OBJECTS through the intersection entity PRODUCT OBJECT. Likewise, other intersection entities may be useful such as PARTY OBJECT (shown in model), FEATURE OBJECT or FACILITY OBJECT.

This model facilitates consistent and complete usage of electronic objects and maintains each object's intended purposes (via OBJECT PURPOSE), applications (via OBJECT USAGE) and which party(s), product(s) or other entities they portray.

### Web Activity

Now that Web parties and products have been addressed, Figure 4 provides a Web-activity model covering information on server hits (requests made to the Web server) and visits (collections of hits within the

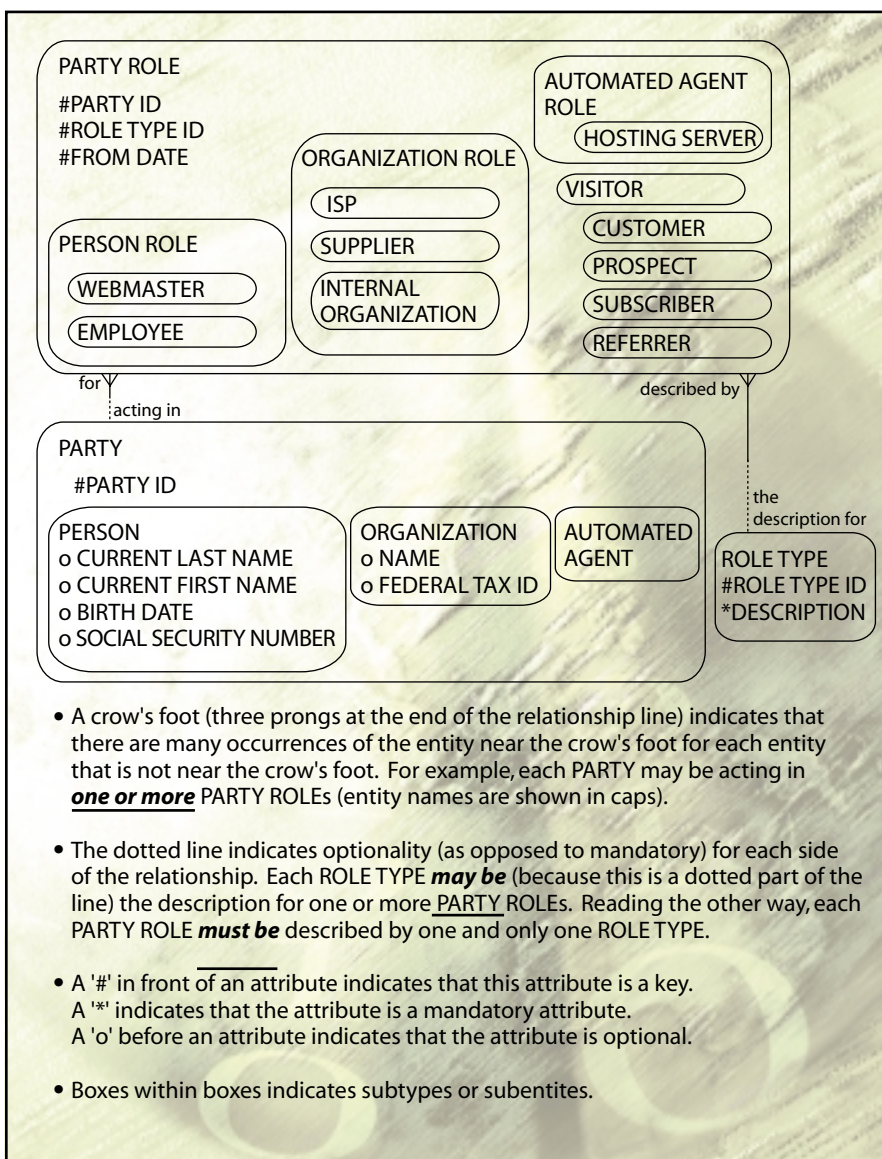


Figure 1: Web Parties

same session). The resulting model enables management and integration of Web activity in order to gain valuable insights into visitors' behavior and interests.

### Server Hits

While the term "clickstream analysis" is used, the enterprise usually does not actually have access to the "clicks;" it only has access to the server hits that have occurred. For example, when a visitor clicks on a certain site icon, there may be several "hits" that are requested from the server. A "click" may spawn a request to the server to go to a different URL, and then bring up five graphic images. In this case, there are six server hits.

Most Web servers have Web logs that maintain information on each hit such as the IP address of the requestor, user login information, the date and time of each hit, what type of request (the protocol used, version, etc.), the status code, the number of bytes transferred, the referring URL and the type of user agent (the type of browser, platform, operating system, etc.).

Figure 4 models this information. The *datetime* attribute of SERVER HIT can help establish when patterns of activity occur and the *num of bytes* attribute can be used to plan load balancing. The relationship from SERVER HIT to SERVER HIT STATUS records the outcome of hits (i.e., successful or unsuccessful) and can identify visitor issues such as failed product-image retrieval.

The relationship from SERVER HIT to WEB CONTENT can be used to monitor visitors' interests. As shown in Figure 2, the WEB CONTENT is linked to various OBJECTS that may be related to various PARTY, PRODUCT or other entities. Thus, we can infer a visitor's interests based upon the objects that are ultimately related to the hit. For instance, if there are 50 server hits from a visitor requesting Web content related to the Superfast 2000 Computer, then the enterprise could infer possible interest in that product.

The hosted-via relationship from SERVER HIT to CONTACT MECHANISM identifies the Web

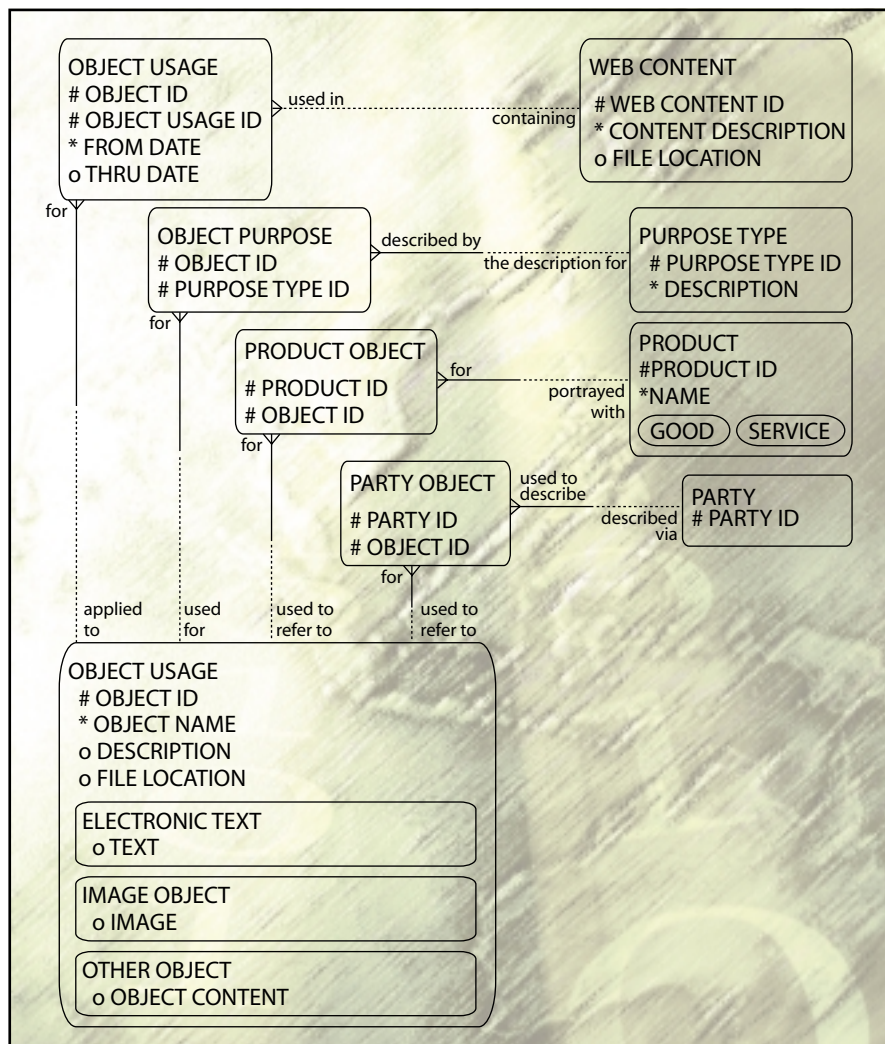


Figure 3: Web Content, Object and Product Model

page that the hit was on. The relationship from SERVER HIT to IP ADDRESS can help identify visitors because it is possible to look up the organization owning an IP address. It is useful to know the sites that are linking and referring visitors to your site, and this can often be found via the "referred via" relationship from SERVER HIT to WEB ADDRESS.

The PLATFORM TYPE (UNIX, Windows), BROWSER TYPE (Microsoft Explorer, Netscape), PROTOCOL TYPE (HTTP, FTP), and USER AGENT METHOD TYPE (GET, PUT, POST) reveal how visitors access the site and may provide insights into more effective Web-site design.

The relationship from SERVER HIT to USER LOGIN maintains the user's login identifier (from the "authuser" field of the Web logs). It is

related to the SERVER HIT versus the VISIT because a visit may contain hits containing user logins as well as hits without user logins identified. An example of this is if the user first enters an open area of the site and then moves to an area requiring a user login.

### Visits

Visit information allows the enterprise to analyze how visitors move around the Web site and thus provides insight into the visitor's activities, interests and buying patterns. To accomplish this, Figure 4 shows that SERVER HITS may be part of a VISIT. The VISIT entity records when the visit started (visit from datetime), when it ended (visit through datetime), what VISITOR was involved and which VISITS resulted in ORDERS. The relationships from

SERVER HIT to VISIT to VISITOR to PARTY allow hit and visit information to be integrated into each party's profile. The PARTY CONTACT MECHANISM shows various ways (CONTACT MECHANISMS) to communicate with a PARTY.

There are several challenges in recording VISIT information. In order to associate hits to visits, the enterprise must know which hits were from the same visitor and which hits occurred within the same session.

### Who Was That Masked Visitor Anyway?

The IP address, cookie and user login ID are three mechanisms for helping to identify the visitor.

A user login ID is the most reliable way to identify visitors. However, it is only available in the Web log for

SERVER HIT DATE	SERVER HIT TIME	USER LOGIN ID (AUTHUSER IN THE WEB LOGS)	COOKIE	IP ADDRESS OF VISITOR	PARTY ID (AND NAME) ASSIGNED
12/30/2001	08:07:35	Jsmith	1001001	10.10.10.10	48958 John Smith
12/30/2001	08:08:20	Jsmith	1001001	10.9.9.9	48958 John Smith
12/31/2001	09:41:45		94949494	10.1.1.1	
12/31/2001	09:42:43		94949494	10.9.9.12	
12/31/2001	23:50:00			10.0.0.1	
01/01/2002	00:10:53		1234567890	10.0.0.1	
01/01/2002	00:41:51	Bjones	8585858123	10.1.0.13	84958 Bob Jones
01/01/2002	00:42:57	Bjones	8585858123	10.1.0.13	84958 Bob Jones
01/01/2002	01:22:44	Bjones	8585858123	10.1.0.13	84958 Bob Jones

Figure 5: Web-Server Hits

hits where the user has logged into the Web site. The first and second rows of the table in Figure 5 show a user login of jsmith. Because these hits occurred

relatively closely, one can conclude that these hits occurred within the same visit.

A cookie is an arbitrary text string

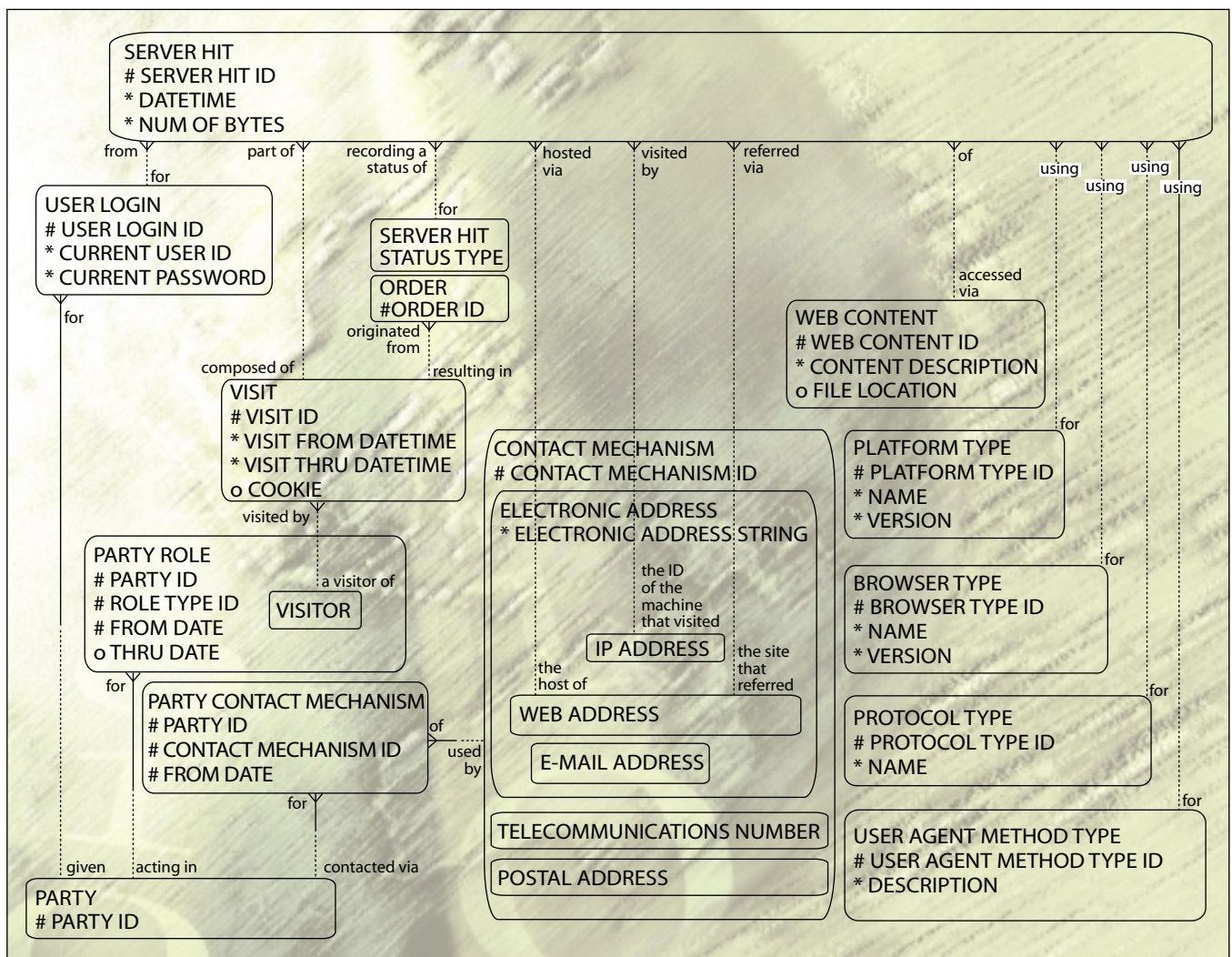


Figure 4: Web Visits

that the Web server will try to place with the visitor's browser (if the visitor's browser accepts cookies) in order to identify if visitors are the same across hits. If hits have the same cookie, then the hits probably came from the same visitor. Thus, the third and fourth hits with the cookie "94949494" were probably from the same visitor. Note that the cookie really establishes that the hits were from the same machine (that stored the cookie in its browser), which infers that it was probably the same visitor, especially if the hits occur within a close time frame.

A third means of identifying a party is by IP address, although it is not as reliable as the first two identifiers. This is because the IP address may be dynamically assigned by the visitor's ISP and may change even within the same visit. However, it is sometimes possible to use the IP address to help identify the visitor if used in conjunction with other information such as a cookie and/or the time of the hits.

Consider the fifth and sixth rows in Figure 5. The fifth record seems to have come in on New Year's Eve on IP address "10.0.0.1," and the sixth record came in approximately 20 minutes later with the same IP address. Is this the same visitor? Probably not given the fact that the first visitor did not accept cookies, but the second visitor did. People tend not to change their browser settings; however, the bottom line is that it is impossible to know if these two hits were from the same visitor. The first two rows of Figure 5 show that the IP addresses could be different, while the visitor may actually be the same. Thus, in many cases, it is difficult to use the IP address to conclude with certainty if the visitor is the same.

### When are Hits Within the Same Visit?

Another factor that determines the hits within a visit is how close hits occur to each other. Consider the last three rows of Figure 5. The enterprise can infer that the *visitor* is the same because the user login, cookie and IP address are the same. However, were

these hits part of the same *visit*? Because the second hit for *bjones* occurred about a minute after the first hit, one could conclude that that they were part of the same visit. The next hit occurred about 40 minutes later. Did Bob Jones go to some other site for a while, or did he just get interrupted, then come back to his computer to continue visiting the site? Most enterprises pick an inactivity time interval to help define a visit. For instance, the enterprise might pick 30 minutes between hits for determining a visit (this is a common standard).

### Modeling Visits?


Knowing that visits (or visitors) cannot be precisely determined, should one even model the entity VISIT? Furthermore, is this derived information and therefore not suitable for a logical data model?

There are many advantages to modeling VISITS. Without a VISIT to link related hits, there may be hundreds of SERVER HITS with the same redundant information. Additionally, the enterprise won't be able to maintain which visits lead to orders. Perhaps most important is the fact that the VISIT entity links together SERVER HITS, which are each related to various WEB CONTENTS, and thus the enterprise could analyze patterns of Web interaction, learning what Web-access patterns ultimately lead to sales.

The model in Figure 4 assumes there is the will and means to capture information about visits. If the enterprise decides that it will not maintain VISIT information, then they could maintain cookie and VISITOR information (via a relationship) with the SERVER HIT entity.

### Valuable Insights

Web activity is available to enterprises, however few enterprises effectively manage this information and even fewer assimilate this information into their knowledge base. Modeling clickstream data offers valuable insights into customer behavior, buying patterns, product interests and business activity. As Web applications continue to become more central to

business operations, it will be essential to model, understand, integrate and utilize this information effectively. 

*Len Silverston is the author of the best-selling **The Data Model Resource Book** series (Wiley, 2001) which describes over 230 reusable generic and industry-specific data models and data warehouse designs. Prebuilt software repository versions of these models are available for license directly and many have been licensed worldwide by Microsoft and included in SQL Server Enterprise Edition. Silverston's company, Universal Data Models, provides consulting, training and software to jumpstart data modeling and data warehouse design efforts. Silverston may be reached at (303) 688-1412 or at [lsilverston@univdata.com](mailto:lsilverston@univdata.com).*

### News from DM Review Interactive

**DMReview.com** would like to welcome Len Silverston as a new online columnist. Beginning in January, Len will write a bimonthly column entitled "National Intelligence and the Integration Gap." This is a call to action for all data management professionals to promote the awareness of the urgent need to appropriately share America's data.

Len Silverston will also be presenting on this subject during **dataWarehouse.com's** online trade show, "Using BI to Unsnarl Data Management Gridlock," beginning January 14. Look for his presentation entitled *Mind the Gap: The Sharing of National Intelligence* at: [dataWarehouse.com/tradeshaw](http://dataWarehouse.com/tradeshaw).

**DMReview.com** announces the return of Richard Kachur as an online columnist. His column, "The Data Warehouse Manager," will focus on the data warehousing process from a number of angles. Kachur is the author of the *Data Warehouse Management Handbook* (Prentice Hall, 2000) and the lead data warehouse architect for TMP Worldwide.