Data Webhouses: Web-Enabled Data Warehouses

Concepts, Techniques and Applications

Prof. Dr. Peter Brezany
Institut für Softwarewissenschaft
Universität Wien

Tel.: 01/4277 38825
E-mail: brezany@par.univie.ac.at
Sprechstunde: Mittwoch 13-14
Motivation
Media That Radically Influenced Society

1500s
Printing Press

1840s
Penny Post

1850s
Telegraph

1920s
Telephone

1930s
Radio

1950s
TV

1990s
Web
The Data Webhouse

Webhouse = the data warehouse, evolved to a new form, because of the existence of the Web

The audience for data warehouse data has grown from internal management to encompass customers, partners, and a much larger pool of internal employees.

The Web’s focus on the “customer experience” has made many organizations much more aware of learning about the customer and giving the customer useful information.

The Web revolution has propelled the data warehouse out onto the main stage, because in many situations the data warehouse must be the engine that controls or analyzes the Web experience. However, the nature of the data warehouse needs to be somewhat different than it has been for the past decade.
The Customer, the Website, and the Webhouse

Interaction \\

Customer \\

Implicit communication \\

Data \\

Website \\

Webhouse
The Focus of This Lecture Part

The Web

The Two Personalities of the Webhouse (This is reflected in the two halves of this lecture.)

1. Bringing the Web to the Webhouse (the clickstream data)
2. Bring the Webhouse to the Web (deploy all Warehouse services to the Web)
Bringing the Web to the Webhouse

The Web is an immense source of behavioral data as individuals interact through their browsers with remote Website.

Although clickstream data in many cases is raw and unvarnished, it has the potential of providing unprecedented detail about every gesture made by every human being using the Web medium.

Data sets carrying clickstream data are becoming the largest known text and number databases.

When we bring the Web to the warehouse, we bring this huge, undisciplined data source into our data Webhouse to be analyzed itself, and alternatively, to be conformed and combined with existing, more conventional data sources.
Bringing the Warehouse to the Web

It means making all data warehouse interfaces available through Web browsers.

It also means addressing the issues of a fully distributed environment.

The data Webhouse is a profound alternative to the fully centralized data warehouse approach.

We must adopt a design philosophy that allows separate islands of data warehousing across the Web to see and communicate with each other in an effective way.
Why Bring the Web to the Warehouse?

- Bringing the Web to the warehouse means bringing behaviour to the warehouse. We are trying to capture analyze, and understand the behavior of users clicking on our Websites.

- The Web presents us with a new data source - clickstream which is a log of every gesture made by every visitor to every Website.

- The clickstream is potentially a much better record of behavior than other traditional detailed data sources:
  - Call detail record (CDR) data from telecomm. companies. It can only show that party A called party B and successfully established a connection for a certain number of minutes. There is no way to know why party A called party B.
  - OLTP data sources usually records only the very last step in a relationship that has been building for some time. E.g., we certainly see that the sale took place, but we have no idea what led up to that sale.
Why Bring the Web to the Warehouse? (2)

- The clickstream is a time series of microscopic actions that can be assembled into sessions.
- The trajectory of actions that led us to a purchase or to other behavior can be analyzed and understood.
- We can be far more confident how the individuals approached us, what their intent was, and what the quality of their experience was. We know what they saw on the screen. We know how long it took them to find the choices they made. We can see direct signs of satisfaction and direct signs of dissatisfaction.
- We are now in a far better position to respond effectively to the individual customer.
The Clickstream Is Not Just Another Data source

- The clickstream is really an evolving collection of data sources (more than 10 log file formats, XML, etc.).
- Clickstream data is often collected simultaneously by different physical data servers - problems with time synchronization (the clocks on different servers are not in synchrony).
- Besides our own log files, we may get clickstream data from referring partners or ISPs (Internet Service Providers).
- Another important form of clickstream data is the search specification given to a search engine that then directs the user to the Web site.
- The ISP sees every click (see the next slide).
The Unique Relationship the ISP/User

User

Personal Computer

User's ISP

The ISP ``owns´´ every click, every gesture

Web
Other Clickstream Data Problems

- The log shows an isolated page retrieval event, but does not provide a clear tie to other page events elsewhere in the log - a **stateless form**.

- **Anonymity of the session.** Unless the user agrees to reveal their anonymity in some way, we often cannot be sure who are they. In certain situations, we may not even be able to distinguish the clicks of two users who are simultaneously browsing our Website.

- There is tremendous pressure to solve these problems. The progress by the Web server vendors and more data-rich descriptions of Web pages is anticipated.
Analyzing Behavior

• A raw clickstream is not a useful description of behavior because you “you can’t see the forest for the trees”.

• A more useful description of behavior is intent.

• The long-term intent of a set of page events may be “Buy a Product.” The short-term intent may be “Get Description of a Product.”

• Behavior types that can be inferred just from the page being viewed: general information gathering, product feature gathering, other information gathering (whether, location, etc.), FAQ list review, specific support question, product order, service order, order status tracking, searching, reading news, bidding, entertainment, downloading, E-mail (responding to something on our Website).

• Higher descriptions of behavior: successful purchase event, cancelled, incomplete, or unsuccessful purchase event, found information being sought, session killer event (user left the site), incomplete display of information, but user stayed on site (i.e., rapid click to next page), wrong path taken, user is angry, user is happy, user is reassured.
The Webhouse Architecture

Visitor with Browser

Visitor’s ISP

The Web

Public Web Server

Directory Server

Private Firewall

Public Application and Business Transaction Server

Hot Response Cache

Data Warehouse Application Server

The data Webhouse Server Suite

Relational DBMS and OLAP Engines

Document, Image, Media Server

(Extract, Transform, Load)

Click Stream Logs
The User and the ISP

The user’s ISP can see everything that the user does, because the user is hardwired through the ISP.
The public Web Server and Business Transactions

All of the various components below the Web cloud in the last figure are assumed to belong to a single commercial entity. One component is the public Web server that everyone gets when the URL www.companyname.com or its equivalent is entered into the user's browser.

Whenever serious business transactions are being performed, a business transaction server is also needed. Its job is to record the business transactions in a legally and financially responsible way, and to never lose these transactions.

All the connections from the public Web server to others servers must be highly secure connections, so that hackers don't gain access to these servers.
The Hot Response Cache

One way to take pressure of the main database engines is to build a powerful hot response cache.

This cache anticipates as many of the predictable and repeated information requests as possible.

The data in this cache is created in a series of jobs running in the main Webhouse application server.

The public Web server makes clickstream logs available. Both these logs and copies of the business transactions are extracted, transformed, and loaded into the company's data Webhouse.
The Data Webhouse System

The four servers in the lower right part of our figure logically comprise the data Webhouse.

Recall our Webhouse definition: The data Webhouse is a Web-enabled data warehouse devoted to publishing the company's data assets appropriately.

Our Webhouse design should be based on maximizing the effectiveness of this publishing mandate.
Tracking Website User Actions

In a retail shop, the customer is identified and his purchases are recorded only after he has completed the entire shopping trip. He will always pass the checkout stand.

On the Web, a customer exists only in a virtual cyberspace. He is one mouse click away from leaving the site or the store, and he may never return to check out his purchase. He can be distracted in a millisecond by another image, a telephone call, or a crying child.

In an E-Store, the visitor enters through a static home or portal page, but once identified, he traverses a store that’s completely dynamic in content. We have an opportunity to identify the customer before he sees even the first page of our site.
Shopping in a retail store.
Shopping in an E-Store.
Tracking Website User Actions (2)

A Brief Catalog of User Actions

Searching, information gathering, entertainment, education, communication, downloading, shopping and ordering, accidental entry

Steps in Product Purchase

Recognition of need
Trying to find what's needed
Searching for information about alternatives
Checkout
Post-order processing
Elements of Tracking

Where did the visitor come from? How did he find our Website? How did he arrive at a particular page or image or order form?

The above answers to these questions are of great importance to the marketing department and Webmasters because they determine the effectiveness of site promotion.

Attracting visitors can be extremely expensive, and new ways of getting visitors are continuously being invented.
User Origin

If we are very, very lucky our site is the default home page for the visitor's browser - this is very unlikely. Unfortunately, there is no easy way to determine from a log whether or not our site is set as a browser's home page.

A visitor may be directed to our site from a search at a portal such as Yahoo or Alta Vista. Such referrals can come either from the portal's index or table of contents, for which you may have paid a placement fee or from a word or content search. For many Websites, the most common source of visitors is from a browser bookmark.

Our site may be reached as a result of a clickthrough - a deliberate click on a text or graphical link from another site. This may be a paid-for referral as via a banner ad or a free referral from an individual or cooperating site. In these cases, the referring site will almost always be identifiable in the Website's referrer data.
Session Identification

Most Web-centric data warehouse applications will require every user session (visit) to have its own identity tag - session ID. The records of every individual user action in a session must contain this tag.

A session ID may not be available immediately, when the events first relating to the session are logged - a temporary session ID will be needed, and this will later be resolved into an enterprise-acceptable session ID that will follow the log information through the data warehouse.

The basic protocol for the World Wide Web, HTTP (hypertext transfer protocol) is stateless - that is, it lacks the concept of a session. There are no intrinsic login or logout actions built into HTTP, so session identity must be established in some other way. There are several ways to do this.
User Identification

Identifying a specific user who logs into our site presents some of the most challenging problems facing a site designer, Webmaster, or manager of data warehousing.

- Web user wish to be anonymous.
- If you request a user's identity he is likely to lie about it.
- You can't be sure which family member is visiting our site.
- You can't assume that an individual is always at the same computer.

Server-provided cookies identify a computer, not an individual.
Using the Clickstream to Make Decisions

Data by itself is virtually useless. When the data has been organized coherently and we can see its patterns, then we have something much more useful. Let us call these patterns information.

But even seeing and describing the patterns does not lead to action. When we can identify cause and effect and correlation, then we have further refined the information into useful knowledge.

The real, final, tangible output from any data warehouse should be decisions made as a result of the knowledge gained.
Decisions About Communicating

- Deciding whether a particular web ad is working
- Deciding if custom greetings are working
- Deciding if a promotion is profitable
- Responding to a customer’s life change
- Improving the effectiveness of our Website

Fundamental Decisions About Our Web Business

- Deciding which products and services we provide over the Web
- Determining if our Web business is profitable
Understanding the Clickstream as a Data Source

• One of the sources of data that will feed our data Webhouse is the HTTP clickstream itself - the log records produced by the Web server each time a request is satisfied.

• A **clickstream post-processor** receives raw log data from a Web server and normalizes it into a format in which it can be combined with application-derived data and piped into the data Webhouse.

• The database volumes required for log processing at an active Website can be equated to the billing system of a large telephone company, both in volume and in complexity.
Web Server Logs

All Web servers have the ability to log client interactions into one or more log files or databases or to pipe the log information to other applications in real time.

The table in the next slide lists some of the typical data elements available from most Web servers.

The original standard for Web servers logs was the Common Log Format (CLF). Two additional elements were added in the Extended Common Log Format Standard (ECLF). Various servers add additional loggable parameters.

The log file presents a particularly difficult analytical challenge. Although an entry is made for each server response, the server may be servicing thousands of user sessions concurrently; the entries for a particular session are not contiguous.
### Web Server Log Data Elements

<table>
<thead>
<tr>
<th>Data Element</th>
<th>CLF*</th>
<th>ECLF*</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host</td>
<td>✓</td>
<td>✓</td>
<td>Fully qualified domain name of the client or its IP address if the name is unavailable.</td>
</tr>
<tr>
<td>Ident</td>
<td>✓</td>
<td>✓</td>
<td>Identity information supplied by the client, if it has identd activated.</td>
</tr>
<tr>
<td>Authuser</td>
<td>✓</td>
<td>✓</td>
<td>If the request was for a password-protected document, then this is the userID used in the request.</td>
</tr>
<tr>
<td>Time</td>
<td>✓</td>
<td>✓</td>
<td>Time the request reached the server in CLF time format (dd/Mmm/yyyy:hh:mm:ss zone).</td>
</tr>
<tr>
<td>Request</td>
<td>✓</td>
<td>✓</td>
<td>The first request line from the client (usually in quotes).</td>
</tr>
<tr>
<td>Status</td>
<td>✓</td>
<td>✓</td>
<td>Three-digit status code returned to the client.</td>
</tr>
<tr>
<td>Bytes</td>
<td>✓</td>
<td>✓</td>
<td>Number of bytes returned to the client excluding HTTP headers.</td>
</tr>
<tr>
<td>Referrer</td>
<td></td>
<td>✓</td>
<td>URL of the referring server.</td>
</tr>
<tr>
<td>User-agent</td>
<td></td>
<td>✓</td>
<td>Name and version of the client (browser).</td>
</tr>
<tr>
<td>Filename</td>
<td></td>
<td></td>
<td>filename.</td>
</tr>
<tr>
<td>Time-to-serve</td>
<td></td>
<td></td>
<td>Time to serve the request (seconds).</td>
</tr>
<tr>
<td>IP-address</td>
<td></td>
<td></td>
<td>IP address of the remote host (see “host” above).</td>
</tr>
<tr>
<td>Server-port</td>
<td></td>
<td></td>
<td>Canonical Port of the server serving the request.</td>
</tr>
<tr>
<td>Process-ID</td>
<td></td>
<td></td>
<td>Process ID of the child that serviced the request.</td>
</tr>
<tr>
<td>Formatted-time</td>
<td></td>
<td></td>
<td>The time, in the specified strftime(3) format.</td>
</tr>
<tr>
<td>URL-requested</td>
<td></td>
<td></td>
<td>The URL path requested.</td>
</tr>
<tr>
<td>Server-name</td>
<td></td>
<td></td>
<td>The canonical name of the server serving the request.</td>
</tr>
<tr>
<td>Cookie</td>
<td></td>
<td></td>
<td>The value of the cookie retrieved from the client’s cookie file.</td>
</tr>
</tbody>
</table>

* Common Log Format (CLF) and the Extended Common Log Format Standard (ECLF).*
Web Server Log Example

P.Brezany

Inst
Cookies

The cookie mechanism provides a Web server with the ability to store a text string on a client computer that can later be read by the server.

Cookies can be *persistent* (stored on disk) or *session* (stored in the client computer’s memory) *level*.

Cookies provide the primary means used today to identify users returning to a Website, and the primary mechanism used by profilers to track users from Website to Website.

The cookie standards exist and cookie mechanisms are included (and activated) in all commercial browsers.
Cookies - Microsoft Internet Explorer

The cookies will reside in the directory: C:\Windows\Cookies

Each cookie has its own .txt file and the following format:

name | value | domain+path | secure (0 or 1) | expiration date | expiration time | last_used date | last_used time | *

Example:

CNNid|cf194799|cnn.com/|0|21899574144|32107986|2410210720|29271805|*
Cookies: Netscape Navigator

The cookies will all reside in the same file, probably in:
C:\ProgramFiles\Netscape\Users\Default\cookies.txt

All of the cookies will be in this file, each on its own line.

Each line will have the following format:

domain | secure (TRUE or FALSE) | path | readable by anyone | expiration datetime | name | value
Implementing the Clickstream Post-Processor

Clickstream postprocessing prepares the clickstream data for loading into the data warehouse.
Bringing the Warehouse to the Web

• The arrival of the Web is hugely beneficial to the data warehouse movement. The Web amplifies and extends the publishing metaphor of the data warehouse.

• But in joining the Web revolution, the data warehouse has to play by a number of new rules. For example:
  - mixing query and update,
  - speed is nonnegotiable,
  - the data Webhouse can no longer go off-line for long periods to reload data,
  - the data Webhouse must deliver its results into a multimedia environment.
  - the Webhouse is profoundly distributed