Clickstream Data Warehouse Initiative

Business Drivers and Enabling Technologies
# Clickstream Data Warehouse Initiative
## Business Drivers and Enabling Technologies

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>1</td>
</tr>
<tr>
<td>Web Analytics Products</td>
<td>1</td>
</tr>
<tr>
<td>The Clickstream Data Warehouse</td>
<td>2</td>
</tr>
<tr>
<td>Clickstream Source Data</td>
<td>2</td>
</tr>
<tr>
<td>Enabling Technologies</td>
<td>2</td>
</tr>
<tr>
<td>DMExpress Clickstream Data Integration Solution</td>
<td>3</td>
</tr>
<tr>
<td>Performance</td>
<td>5</td>
</tr>
<tr>
<td>Customer Examples</td>
<td>5</td>
</tr>
<tr>
<td>Ecommerce Retailer</td>
<td>5</td>
</tr>
<tr>
<td>Marketing Services Provider</td>
<td>6</td>
</tr>
<tr>
<td>Online Gaming Website</td>
<td>7</td>
</tr>
<tr>
<td>DMExpress + Conventional Databases: A Practical Alternative to MPP Databases</td>
<td>7</td>
</tr>
<tr>
<td>Conclusion</td>
<td>8</td>
</tr>
</tbody>
</table>
Executive Summary
Access to timely “clickstream” data from a company’s website provides insight into online visitor behaviour and patterns, which in turn enables companies to be more effective in myriad ways: with improved pricing, more effective campaigns and offers, better visitor segmentation and targeting, optimised website layout and workflow, and so on.

Many companies use off-the-shelf web analytics products, which often make it relatively simple to monitor and analyse website activity, eliminating the need for significant IT efforts. There is no shortage of these web analytics products from which companies can choose. But as companies become more sophisticated in their web analytics requirements, they often need to augment the capabilities of these packaged products. Building and maintaining an internal clickstream data warehouse (CDW) enables these companies to manage, segment and report on data in ways that the packaged products do not.

Until recently, building and maintaining a CDW has been prohibitive for most companies because of the volume and complexity of the warehoused data, as well as the volume and complexity of the raw source data files. Though recent advances in data warehousing technologies, such as columnar databases and data warehousing appliances that deliver high performance with massive and complex data sets, have made CDWs a practical option for many companies, these technologies are not by themselves sufficient.

The raw clickstream source data, which often exists as huge, complex text files, must be parsed, structured, cleansed and loaded on a regular — sometimes daily — basis into the CDW before the CDW can provide value. This process often introduces additional cost, complexity and delay into the process. So the ability to process the raw clickstream data files in a rapid, cost-effective and scalable manner is a critical component for any successful CDW initiative.

Web Analytics Products
Off-the-shelf, software as a service (SaaS) web analytics products have been available for years. One major vendor reports that it has over 5,000 customers, and some of the major search engine companies offer popular web analytics products for free.

These products use a variety of underlying technologies (including page tagging, packet sniffing, and others) to...
collect a company’s website visitor data on the analytics vendors’ servers, and then provide each customer with capabilities to report on their specific website data. The primary benefit of using a SaaS web analytics product is that it requires much less effort than taking the data in-house and building a CDW. For many companies, these products provide sufficient levels of detail and flexibility in their reporting.

However, problems and limitations exist with these products, including:

Lack of user-centric segmentation. Although useful for tracking activity in a page-centric manner, for many companies the specific information and segmentation available with off-the-shelf web analytics products does not satisfy their requirements for user-centric information. So, while companies are able to track the number of visitors, page views, and conversions on their website, they are unable to segment the data by user session to understand what a user does in a particular session, and to track a user’s activity across multiple sessions.

Historical analysis. The pages to be tracked and the tracking criteria must be defined in advance — it is impossible to report on new criteria from previous (historical) website activity. The new criteria must first be defined, and only subsequent activity can be tracked.

Visitor tracking limitations. A limitation with technologies like page tagging is that not all user visits are tracked. For example, for visitors that have deleted cookies, or that don’t have JavaScript enabled (such as on mobile devices), the visits are not recorded.

Object tracking limitations. Activities with certain object types, such as PDF views and file downloads, are not tracked.

Server tracking limitations. Since most of these products rely on code that is executed in the client computer, they cannot report on server responses, such as failed requests, response times, etc.

Confidentiality. Web analytics vendors store all of their customers’ web traffic data on their own servers. For some companies (and government organisations), the risk of their proprietary website analytics information being used without their knowledge or approval is not acceptable.

These issues, and others, are sufficient for some companies to build and maintain their own CDW, and perform analytics on the data with sophisticated Business Intelligence (BI) software.

The Clickstream Data Warehouse

A CDW is used to store all of the historical website activity in a structured format — typically on the company’s own servers — so that sophisticated queries and reports can be run on the data with BI software. Because of the large volume of clickstream data generated on a daily basis, and the large number of fields in the data, the prospect of implementing a CDW can be daunting. Even so, the business advantages of augmenting — or supplanting — packaged SaaS web analytics products is often sufficient justification for companies to undergo the initiative.

Clickstream Source Data

Just as different SaaS web analytics vendors use different underlying technologies — such as page tagging or packet sniffing — to track clickstream data, a CDW can utilise various sources of clickstream data as input. For example, companies that are already using a web analytics product often use their analytics vendors’ source data as input to their CDW. All of the major analytics vendors offer their customers access to their full source data via batch delivery services or APIs. This enables companies to continue to utilise their prior investments in customising the product, and augments many of the limitations of the SaaS offering.

Alternatively, a CDW can be built by processing the raw log files written by the web servers. Since the web server log files contain every transaction, these files provide more complete data to be mined, eliminating many of the limitations of client-side tracking technologies, such as the inability to track visitors running clients without JavaScript.

Some companies choose to build CDWs by combining multiple data sources that use different types of clickstream data. Doing so enables companies to utilise the benefits of multiple underlying technologies — for example, by enriching the batch data files from their web analytics vendor with the web server log data.

Enabling Technologies

Until recently, the massive amount of computing resources required to effectively work with the
Clickstream data made CDW initiatives prohibitive for many companies. However, recent advances in data warehousing technologies, including massively parallel processing (MPP) architectures and columnar databases require less investment in hardware resources and deliver significantly better price-performance ratios than ever before. As a result, many of the newer, successful CDW initiatives rely on these high-performance data warehousing technologies.

But high-performance data warehousing technology alone is not sufficient. A critical component of any CDW implementation is high-performance data transformation technology that can parse, structure, and cleanse the raw clickstream source files to initially populate the CDW, and refresh the CDW on an ongoing basis. Since these source files are typically very large, complex, and require significant processing to extract the desired information from the files, and because of the frequency of the refreshes, processing these files is a common source of bottlenecks, introducing delays and error into the process.

A common technique for processing raw clickstream source files for the CDW is via custom-coded solutions, including shell scripts, SQL programming, Perl scripts, etc. It is rare to encounter a typical data integration or "ETL" product being used for this task, since the requirements are so performance intensive.

But hand-coded applications are fragile and brittle to maintain, and almost impossible for anyone other than the initial developer to extend. Despite low initial costs, over the life of the application the total cost of ownership of a hand-coded application is typically higher than an application created with metadata driven software that utilises a graphical, drag and drop development environment.

However, most packaged data integration software products are not sufficiently optimised for performance. As a result, the techniques that a skilled developer can bring to bear in addressing performance problems by taking advantage of known patterns in the data will often perform more efficiently than a packaged data integration product. Still, the extreme processing requirements around clickstream data files often cause even custom-coded solutions to exhibit performance bottlenecks and cause Service Level Agreements (SLAs) to be missed.

A different type of packaged data integration software — one that provides the benefits of ease-of-use and extensibility while offering extreme levels of performance and scalability in processing large, complex clickstream data files — provides the ideal solution.

DMExpress Clickstream Data Integration Solution

Syncsort’s DMExpress Clickstream Data Integration (DI) solution provides the fastest, most cost-effective, and most scalable technology for processing clickstream data and populating a CDW. DMExpress is to traditional data integration software what MPP-based data warehousing technologies are to traditional data warehousing architectures.

At the heart of the solution is Syncsort’s DMExpress software, which delivers high-performance data integration. Through the use of proprietary sorting algorithms, I/O optimisation, parallel processing, and dynamic environmental monitoring techniques developed and refined over 40 years with over 12,000 customers and more than 90 of the Fortune 100 companies, DMExpress delivers highly efficient data integration solutions with significantly reduced CPU time, elapsed time, and disk I/O activity for data processing applications.

DMExpress provides a simple and straightforward approach to processing clickstream data, especially compared with custom-coded solutions. It utilises a drag and drop graphical development environment that simplifies the initial development of an application, as well as the ongoing maintenance and modification.
of existing applications to handle new data, transformations, and processes.

Clickstream data files, and web server log files in particular, can be difficult to process, in part because they are complex text files containing a tremendous amount of information. The Clickstream DI solution is optimised for processing these clickstream data files with out-of-the-box features that include:

• Sophisticated and flexible text/string processing capabilities. They include specific functions to parse clickstream files to find the first of a given set of text constants contained within a text value, to check if a text value contains any of a given set of text constants, and to check if a text value is equal to any of a given set of text constants. These types of string-specific functions speed and simplify the process of identifying and extracting only the desired data from clickstream data files.

• Built in "URL decode" functions that automatically decode URL encoded expressions to their ASCII equivalents, for example to automatically decode %20 to a space, %2c to a comma, and so on.

• Extensive pattern matching [regular expression] capabilities that make it easy to locate and extract any string patterns in the text files. For example, to find all secure URLs, the process might be to search the entire file for string patterns that begin with "https://" and end with ".htm" or ".html." Powerful and efficient string processing and pattern matching capabilities are provided to easily process and structure various types of clickstream source files prior to loading into the data warehouse.

The Clickstream DI solution provides functionality for processing web server log files as well as vendor-specific clickstream data files provided by the various web analytics providers, including raw page-tagging files, raw packet sniffing logs, etc, providing companies with the flexibility to incorporate a variety of different clickstream source data into the CDW.

DMEXPRESS BENEFITS:
• Massive speed and scale
  - Enables customers to quickly process massive volumes of clickstream source data files to feed the CDW
  - Enables customers to more easily meet overnight batch windows for CDW refreshes
  - Has been proven to be faster than custom-coding and traditional data integration software products

• Runs on a minimal resource footprint on off-the-shelf commodity hardware
  - Saves money by eliminating expensive hardware purchases and maintenance

• Rapid implementation — in days, not months — and easily extensible for future modifications
  - Results in standardised transformation processes, improved maintainability and re-use of the code, and increased transparency via metadata creation and management

"DMExpress is to traditional data integration software what MPP-based data warehousing technologies are to traditional data warehousing architectures. "

Clickstream Data Warehouse Initiative | Business Drivers and Enabling Technologies
Performance
DMExpress is optimised for processing massive clickstream source data files at “world record” speed — using off-the-shelf commodity hardware resources.

DMExpress’ performance results primarily from its approach to the following:
• Algorithms
• Architecture Exploitation
• Dynamic Optimisation

Algorithms
Although there are just a few basic transformation operations (such as sort, join, aggregate, filter, match), there are literally hundreds of different algorithms that can be employed to perform these operations. DMExpress utilises a large library of transformation routines, including its own patented algorithms. Each algorithm has different performance characteristics in terms of the number of compares it does per record, the number of moves of the record, and so on. DMExpress is optimised to select the most appropriate algorithms to apply to the data at different steps in the processing in order to obtain the best overall performance. Synsort’s Research and Development teams constantly benchmark, review, and revise these algorithms and the use cases for each to make sure the software is as efficient as possible.

Architecture Exploitation
Another critical component of performance is to exploit the architecture of the systems being used. Optimal performance is a function of the specific data being processed, the amount of data, and the attributes of the system that the processing is being run on.

For example, different hardware systems have different numbers of processors or processor cores, different instruction cache capacities, different memory cache capacities, and so on. The best transformation algorithm to employ in any given scenario will be partly dependent on the characteristics of the production hardware, since some algorithms are naturally cache efficient, while others are less so, and different levels of parallelisation are appropriate for different systems.

DMExpress Research and Development teams continually study and benchmark these and other issues to deliver the most efficient algorithms based on the specific hardware and architecture being employed.

Dynamic Optimisation
Typically, an application developer will not know key attributes of the production hardware at the time the application is designed and developed. Often, even at the time the application is deployed, the complete system configuration is not known, and data volumes may change radically when run in production environments. DMExpress is optimised to programatically select the most efficient algorithms to fully exploit the data and production environment being employed for the processing tasks at runtime.

DMExpress applications are resource planned at the time they run, not at development or deployment time. As a result, the application behaviour dynamically adapts to the exact system and data attributes at runtime. This unique capability is critically important for achieving optimal performance in data integration processing.

Customer Examples
Ecommerce Retailer
A major online retailer that carries over 60,000 products and generates $1B (approx £0.6B) in annual revenue built their own CDW to augment the reporting capabilities provided by Omniture’s SiteCatalyst.

A key requirement for this company is to segment website activity into individual user sessions, connect multiple sessions over time to each individual user, and — if the user logs in — identify the user and connect all of the user’s online activities with the user’s activities from their various operational systems, including their customer support and provisioning systems. In other words, the company needs to maintain a rich CDW that stores customer information and transactions, and is enriched with the website activity for each user.

“One DMExpress customer has a monthly requirement to process each month’s worth of raw web server log files. The raw monthly web server log files are over 30 TB in size, contain over 50 billion URL events, and are processed in 16 hours on a commodity Dell server with DMExpress.”
Although they are using Omniture’s SiteCatalyst to perform their web analytics, their primary limitation with SiteCatalyst — which is common among all of the SaaS web analytics products — is that it is page-centric in its reporting. So while it is effective at reporting on the number of visits, page views, and conversions, it is unable to segment the website activity based on individual user sessions.

In response, the company developed its own CDW to augment the reporting available with Omniture’s SiteCatalyst. Each day, Omniture delivers the previous day’s complete page-tag data to the customer. These daily Omniture “DataWarehouse” files range from 18 to 30 GB in size, and require cleansing and processing before they can be loaded into the Teradata CDW. The company’s brand and category managers perform reporting and analytics on the data in the CDW via Cognos BI.

Initially the company was using lengthy, complex custom shell scripts to parse, structure, and cleanse the Omniture clickstream data files prior to loading into the CDW. Execution of these scripts was causing performance problems.

In addition to the performance issues, the scripts were tedious and cumbersome to develop, modify and maintain. Their business dictates that they are able to rapidly — and flexibly — bring on new suppliers to sell additional inventory. The inflexible shell scripts were preventing the company from enabling new business partners with new terms or processes in a timely manner.

Finally, one developer in the company created all of these shell scripts, and he was the only employee who could maintain or extend them. This created an unacceptable level of risk for the company.

With the Clickstream DI solution, the company was able to replace their custom shell scripts and process the Omniture DataWarehouse files for their CDW on a daily basis. By replacing their custom-coded processing with DMExpress, the company successfully decreased processing time by 65%, eliminated their reliance on a single developer for maintaining and extending the code, and can now more quickly incorporate new business partners.

---

**Marketing Services Provider**

Another DMExpress customer builds and maintains vast databases of online consumer activity, which it mines for online consumer behaviour in order to design and execute web-based sales and marketing strategies for its clients. The company maintains massive proprietary databases that provide a continuous, real-time measurement of the myriad ways in which the internet is used and the wide variety of activities that are occurring online. To support the varying ways that their clients need to segment this data, the company built their own CDW using raw web server log files as the source data, and utilising a columnar database running on off-the-shelf Windows and Linux Dell servers.

Each week, over 10 billion new master records — about 8.5 terabytes of compressed data — need to be processed and loaded into the CDW. The processing of the raw source data was an overwhelming undertaking due to the volume and unformatted nature of the raw web server log files. Prior to implementing DMExpress, the company had been using custom-coding techniques, which was causing problems in their ability to service their customers in compliance with their SLAs. Without a faster solution to processing the web log files, their analysts did not have timely access to the information they needed to build campaigns for their clients.

Replacing the custom code, the company used DMExpress to parse, structure, and cleanse the raw web log data and perform complex data transformations. The resulting process is 5 times faster than their previous solution, and runs on commodity hardware.

The company also uses DMExpress to perform high-performance data transformations on the structured data already loaded in the database. In one example, for the final aggregation on their domestic data sets, they process a 450 Million row sort (56GB input file to a 48GB output file) in 40 minutes on a Dell 2950. For the international data, they process a 950 Million row sort (137GB input to 103GB output) in 90 minutes on the same hardware platform.

According to the company’s EVP of software engineering, “The performance and ease-of-use of DMExpress technology positively impacts our bottom line. Our analysts are able to deliver timely and accurate solutions to our customers.”

---

“DMExpress makes it simpler and more cost-effective for companies to manage their clickstream data in-house, raising the bar on the insight and benefits that can be obtained via clickstream analysis.”
clients and meet service level agreements, as DMExpress technology is able to convert raw clickstream data into valuable granular information at lightning speed. In fact, we saw a 500% improvement in data integration throughput after deploying DMExpress technology in our environment.”

**Online Gaming Website**

A developer of massively multiplayer online role-playing games (MMORPG) with an active user base in the tens of millions performs analytics on player behaviour via their own CDW. Their requirements — to analyse players’ use of their proprietary web applications [games] — are so unique that they did not implement an off-the-shelf web analytics product.

Their process involves extracting 2 billion rows spread across 3 to 4 primary tables from a transactional Oracle database twice each week. The company then processes and loads the data into an OLAP SQL Server CDW. The analysis and reporting on player behaviour is performed using Cognos BI.

Their initial method of processing the clickstream data was via SQL procedures, with hundreds of lines of SQL code required to execute the end to end process. This approach was causing the elapsed processing time to take two to three days, resulting in delays of one to two days delivering the processed data to the BI software for the subsequent statistical analysis.

**Replacing the SQL code with DMExpress reduced the elapsed time from 2-3 days to 5 hours, enabling the company to consistently meet the requirements of its internal business user community.**

The company also reports that new development and ongoing maintenance is easier, and that more developers, with a wider range of skill levels, can extend the code and create new transformations.

**DMExpress + Conventional Databases: A Practical Alternative to MPP Databases**

In addition to processing raw clickstream data files to load into the CDW, DMExpress is often used for high-performance transformations on the structured clickstream data that is already stored in tables within the CDW. DMExpress is the proven fastest technology for performing transformations such as joins, sorts, and aggregations, and makes it practical to process massive data volumes within short elapsed timeframes on low cost commodity hardware.

For example, one major online retailer collects and stores two years’ worth of data representing every single website visitor’s clicks in their Oracle data warehouse. This retailer maintains a complex set of rules that determines user profiles and user value scores. Further, the company retains a historical record of the value score for each user as it changes continuously over time based on the actions the user performs on the website, the user’s billing history, etc. So, when the criteria for calculating user value scores change, as it sometimes does based on new information or analyses, the company must recalculate every value score for every user as it existed at the various instances over the two year span. This creates tremendous resource requirements for the database and hardware, due to the sheer volume of the data involved and the complexity of the calculations.

This customer determined that a particular change to the calculation for determining the user value score would incur a 1-2 month update using PL/SQL, since they would need to recalculate the various scores for each user as they changed over the two year time period. Using DMExpress to process the data in the Oracle tables, the calculations instead took less than 9 hours, running on a £1,000 workstation.

Instead of adopting new MPP-based data warehousing technologies, for a variety of reasons many organisations prefer to continue to use their existing, conventional database technologies. For these companies, using DMExpress to perform high-performance data integration with their current data warehousing technologies can often be a simpler and more cost-effective alternative to the new MPP-based data warehousing technologies.
Conclusion

Clickstream data is a rich source of information for companies. While many web analytics products are available, limitations associated with these products often drive companies to undertake their own internal CDW initiatives. With a CDW in place, companies can segment individual user sessions, combine customers' online activity with data from other operational systems, and overcome many other limitations of the SaaS web analytics products.

New, high-performance data warehousing technologies are often used for these initiatives, since the massive data volumes and the complexity and cardinality of the data require extreme processing resources. However, these new data warehousing technologies are not sufficient to ensure a successful CDW initiative. High-performance data integration technology that can parse, structure, and cleanse the raw clickstream source files to initially populate the CDW, and to refresh the CDW on a regular basis is also required.

Fortunately, this technology is readily available, making it simpler and more cost-effective than ever before for companies to manage their clickstream data in-house, effectively raising the bar on the insight and benefits that can be obtained via clickstream analysis.
About Syncsort

Syncsort is a global software company that helps the world’s most successful organisations rethink the economics of data. Syncsort provides extreme data performance and rapid time to value through easy to use data integration and data protection solutions. With over 12,000 deployments, Syncsort has transformed decision making and delivered more profitable results to thousands of customers worldwide. Syncsort Incorporated, headquartered at 50 Tice Boulevard, Woodcliff Lake, NJ 07677, USA. www.syncsort.com

Syncsort Ltd., 2 Kings Hill Avenue, Kings Hill, West Malling, Kent, ME19 4AQ, UK, Tel +44 (0)1732 849000 www.syncsort.co.uk

© 2018 Syncsort Incorporated. All rights reserved. DMExpress is a trademark of Syncsort Incorporated. All other company and product names used herein may be the trademarks of their respective companies.