A Primer on Web Analytics
Analyzing clickstreams with additional data sets to move beyond metrics to actionable insights
Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage.

Overview

Creators of web sites and Internet content are faced with the daunting task of understanding how customers interact with their products. Observing customers on the Internet is a very data intensive activity and the web analytics industry has responded with tools to analyze and report on web usage. However the picture is incomplete without looking for deeper insights that involve integrating data from multiple sources including customer relationship management systems (CRM), ad logs and even data from Social media sites.

This paper will address the following:

• Provide a broad overview of web analytics
• Explain the importance of integrating clickstreams with additional data sets
• Review the challenges of moving beyond metrics to insights
• Outline how the Vertica Analytics Platform enables customers to succeed when facing the challenges related to web analytics

The Challenge

The field of web analytics has grown over the last 15 years by creating a set of standards on how to collect, measure and analyze data in order to optimize Web usage. Companies such as Adobe Omniture and comScore provide comprehensive suites of fee-based tools that make it easy for large customers to manage their reporting needs. Smaller sites have only recently seen the barrier to performing web analytics drop as both Google and Yahoo launched services that include rich reporting suites with advanced features like heat maps and conversion funnels for free!

The availability of free tools has somewhat leveled the playing field. Right now the owner of a scrappy start-up can generate metrics and reports to optimize their site in the same way as the largest sites on the Internet. In order to stay ahead of the competition, companies are forced to look beyond simple web metrics and look for ways to gain deeper insights into how customers use their sites. This process involves using sophisticated analytics to create and categorize visitor segments and then derive value either directly based on revenue or interpreted based on less tangible factors such as loyalty.

A Web Analytics Primer

The collection of data on visitor behavior is necessary for reporting and analysis. Concerns of user privacy need to be addressed through the adoption of a privacy policy that ensures customer’s identities are protected. This white paper assumes that companies engaged in web
analytics have done so; for more information on privacy please refer to http://www.webanalyticsassociation.org/?page=privacy

There are three pieces of information used to track visitors:

- **Page URL** – the web address of the page the customer visited.
- **Referrer** – the web address of the previous page the customer visited.
- **Cookie** – a text file that can be stored within a customer’s browser that uniquely identifies their site visit.

Other items commonly tracked include user submitted information such as an email address or information inferred from their IP address such as their location.

Data collection is performed by 2 methods:

- **Javascript tags** – A script that tracks human activity on a page and then forwards the data to a server for analysis. This is the primary method of data collection for 3rd party vendors such as Adobe Omniture and Google Analytics.
- **Server Logfile Analysis** – The transaction data created by web servers. Logfiles contain detailed data such as the visitor IP address, time of visit, status code, etc.

Once data is collected it is then aggregated to create a set of common metrics to provide a high level overview of visitor behavior. The following metrics are common examples used everyday:

- **Visits** - By default in analytics, a visit or session is defined as the period of time during which visitors are interacting with your site and there has been inactivity for less than 30 minutes.²
- **Unique Visitors** – Unique Visitors represents the number of unduplicated (counted only once) visitors to your website over the course of a specified time period.² A unique visitor is determined using cookies or an IP address.
- **Page Views** – A page is defined as any file or content delivered by a web server that would generally be considered a web document.²
- **Page Views / Visit** - The average number of pages viewed per site visit.
- **Bounce Rate** – Bounce rate is the percentage of single-page visits or visits in which the person left your site from the entrance (landing) page. Use this metric to measure visit quality – a high bounce rate generally indicates that site entrance pages aren’t relevant to your visitors.²
- **Conversion Rate** – The percentage of visitors who take a desired action.

Conversions represent visitors taking actions that are meaningful, an example would be tracking the number of sign-ups to an online newsletter. For e-commerce sites conversions could include tracking keywords, banner ads, landing pages and every page in the purchase process.

The need to include advertising as part of conversion reporting may be a turning point where clickstream alone cannot address all web reporting needs. Many analytics tools have sophisticated conversion-tracking capabilities that can trace the step-by-step path from browsing
to buying, however, tying advertising to conversions requires the inclusion of in-house or 3rd party ad server logs with clickstream. Accomplishing this requires organizations to take control of their data and leverage analytical databases to integrate the variety of data types / sources together.

Moving from Metrics to Insights

Conversion tracking is only one case where reporting needs outstrip the capability of a single tool. Companies are becoming focused on understanding outcomes and asking questions similar to the following:

- If my conversion rate is 5% what happened to the rest of my customers?
- What is revenue generated per page view and how do we improve it?
- What segment of my customers are the most loyal and engaged?

Answering these questions requires the combination of multiple data sources - typically web logs and / or click stream data is combined with other sources such as:

- Advertising data logs
- Data extracts from CRM systems such as Salesforce.com
- Results of customer surveys
- Results from multivariate or A/B testing
- Visitor comments from social media sites like Twitter and Facebook

Vertica, An Analytics Platform

Companies must choose a high performance analytical platform that supports sophisticated analytics to generate insights and optimize outcomes. Merging and querying integrated web data severely taxes the scalability of traditional databases; clickstream data is granular and extremely high volume. Adding disparate data sources of moderate to extremely large size increases the complexity, while demanding increased scale and performance. An important side note is the criticality of maintaining the granularity of the original data set - organizations need to understand individual behavior and predict patterns that lead to optimal outcomes (conversions as opposed to abandons), the only way this type of analysis is possible is by keeping the detail level data of individual user’s click paths throughout the entire analytic pipeline.

Performing advanced analytics requires purpose build solutions that are beyond the capability of row based DBMS systems. Traditional RDBMS’s such as MySQL and Oracle do not take advantage of MPP (Massively Parallel Processing) in order to perform the sophisticated analysis required to interpolate data fields and update advanced user segments.

Currently there are a several options available to companies:
• Purchase specialized analytical appliances such as IBM Netezza or Oracle Exadata. These machines are expensive, encounter operational challenges scaling beyond a single rack, and are extremely cost prohibitive.
• Build databases on open source NoSQL technology. This option requires significant in house expertise to design, build and maintain.
• Leverage a purpose built analytics platform coupled with industry standard hardware (could be physical machine, virtual machine, or cloud deployment methodology).

An optimal solution is to use the Vertica Analytics Platform, a True Column Store™, that is based on an MPP architecture offering linear and limitless scalability. Vertica's architecture not only provides parallel load and query with continuous access to your data (analyze while you load) but also leverages sophisticated compression techniques to minimize hardware requirements while maximizing performance. Unlike legacy row store databases where all columns for each row are stored together, Vertica stores the data in columns and subsequently only needs to retrieve the columns required for a specific query, as opposed to all columns in the selected rows. Columnar orientation is an important feature for web analytics as row based databases are not able to perform the queries required to integrate the data or segment customers.

**Key Attributes of the Vertica Analytics Platform**

There are various reasons the Vertica Analytics Platform is superior for web analytics, several of which are listed below:

• Vertica organizes data on disk as columns of values from the same attribute, as opposed to storing it as rows of tabular records. This organization means when a query needs to access only a few columns of a particular table, only those columns need to be read from disk (conserving resources through the entire processing pipeline). Conversely, in a row-oriented database, all values in a table are typically read from disk which creates excessive I/O processing ultimately limiting scale and performance.
• Vertica employs aggressive compression of data on disk as well as a query execution engine that is able to keep data compressed while operating on it. Compression in Vertica is particularly effective, as values within a column tend to be quite similar to each other and compress well, often by up to 90%. In a traditional row-oriented database, values within a row of a table are not likely to be very similar and hence are unlikely to compress well. Columnar compression and direct operation on compressed data shift the bottleneck in query processing from disks (I/O processing is expensive and hasn’t experienced rapid improvements / innovation) to CPU’s (which are less expensive, and are getting faster and more dense all the time).
• Because data is compressed so aggressively, Vertica has sufficient space to store multiple copies of the data to ensure fault tolerance and improve concurrent and ad hoc query performance. Logical tables are decomposed and physically stored as overlapping groups of columns, called “projections” with each projection being sorted on a different attribute or set of attributes. This architecture optimizes the data for answering queries with predicates on its sort attributes. A Vertica database is composed exclusively out of these query
optimized structures on disk, without the overhead of base tables. It’s similar in concept to a database made entirely of materialized views with no base tables (but without the drawbacks associated with materialized views).

- A shared-nothing, massively parallel processing (MPP) architecture. Each node is an industry standard, multi-core server with 4 to 6 GB of RAM per core. Storage can be directly attached to each node or can shared (SAN or NAS). Vertica partitions data across the cluster automatically, the result is an inexpensive and flexible architecture that will scale linearly as data needs grow.

- Advanced In-Database Analytics provide an “out of the box” experience for users to analyze the vast amounts of data associated with web analytics. Vertica provides a library of features and functions to analyze data, integrates natively with various off the shelf tools (SAS, R, etc.), and provides an SDK for users to develop their own functions

- Native BI, ETL, & Hadoop / MapReduce Integration - seamless integration with a robust and ever growing ecosystem of analytic solutions.

Conclusion

The Vertica Analytics Platform provides companies with a competitive advantage in web analytics. It delivers the capabilities required to create insights and optimal outcomes for both the company and customer with unmatched price, performance, flexibility and extensibility.
