Implementing an e-Intelligence Information System in an Extremely Active Web Environment

Qualex Consulting Services, Inc., Apex, NC, USA

Abstract

Businesses are turning to e-intelligence solutions to harness the promise of the Internet to revolutionize the way businesses and their customers interact. The Internet has created an entirely new sphere of customer contact for business. This creates new challenges for the capture and transformation of the data from web logs into meaningful and strategic information. The information garnered from standard analytical and click-stream analysis empowers business decision makers to understand their customers and respond to them in a much more directed and intelligent manner.

Qualex Consulting Services, Inc. has developed an e-intelligence solution for one of the most popular web sites in the United States. With over 35 million lines of data collected daily from customers accessing their web site, the challenge was to collect, store, analyze, and report the data in a consistent and timely manner. These challenges proved insurmountable to other vendor solutions, but not to SAS and Qualex Consulting Services. The e-intelligence information system developed for this project is an integrated SAS end-to-end solution.

This paper will discuss the challenges faced, and the innovative responses used to answer them in delivering a full and complete e-intelligence information system.

Introduction

Washingtonpost.com is the website for one of the United State's most prestigious newspapers, The Washington Post. When they decided to develop a web site to deliver much of the news and editorial content provided by the print edition, the staff at The Washington Post was sure they would have an extremely popular web site.

With diligent planning and careful preparation, they produced Washingtonpost.com. One of the goals of implementing this web site was to track the habits of the site's visitors, to understand how the readers interacted with the web site. This information would be used in numerous ways: to drive advertising revenue, to make improvements to the site based on user navigation, and many other uses.

The primary source of data for this information is contained within the web logs generated by the web server software. The role the SAS System played in this scenario was to enable a solution to transform the data from the web logs into useful information for the managers of the Washingtonpost.com web site. The solution needed to be powerful, flexible and provide the level of information that was required to manage their business. The SAS solution provided the ability to perform ad-hoc queries of the data in an extremely scalable and adaptable environment that will meet the rapid growth of the Washingtonpost.com web site.

Overview

Information is gathered into a central data repository of approximately 150 gigabytes (GB). This repository resides on a Sun 450 Unix-based operating system. The web logs are stored in a GZIP format and is read in directly to the SAS system. This reduces the storage needs for the source data. The SAS warehouse data and compressed source files take up the majority of the space required for this solution. The information generated by this application is provided to internal web and marketing analysts as well as upper management. This information could be reviewed via an Intranet site. Currently it is imported from tab-delimited ASCII files into an Excel spreadsheet.

It can be built up to provide any time duration and provide trending and forecasting capabilities. This information is stored to the hour level to provide the most extensive detailed analysis possible. There is also the ability to access data in an ad-hoc manner via a web-based interface.

Web Servers and Web Logs

Within Washingtonpost.com there are several internal web servers from which the web logs are generated. The web traffic on the web site is balanced across these servers to ensure that visitors to the Washingtonpost.com web site can navigate the site quickly and easily. Each of these servers generates a web log using the extended common log format for web logs. These logs are ASCII flat files or compressed ASCII flat files and provide the bulk of the data processed. These logs track the visitors navigation through the web site's advertising, news stories (AP, staff, and freelance), editorials, etc.

In addition, several outside vendor logs are also processed. The vendor logs are web logs from external web servers. These vendor logs are generated when a visitor to the Washingtonpost.com site clicks on a link that actually points to a page on another web site. This transfer to the other web site is usually transparent to the user. They do not know that they have left the Washingtonpost.com site. These vendor sites have content that the Washingtonpost.com does not maintain, such as stock reports, TV listings, comics, classifieds, etc. The vendors parse their logs and send their web log data that was generated from the Washingtonpost.com site back to Washingtonpost.com to be included in the web log analysis. The vendor web logs are also ASCII flat files and are in a variety of formats, including common log format or extended common log format. Many of these are either a hybrid or totally customized log file.

It is because of this variability of log file formats that the SAS System has proven to be the best choice for developing the web log analysis tool for Washingtonpost.com. The SAS System provides the power and flexibility needed to handle the complexity and volume contained within the web logs.

Page Views vs. Hits

One of the fundamental decisions in the design of the web log analysis tool was to determine exactly what was to be measured. The data from the web logs contains a vast amount of information. The actual HTML pages that are viewed, the hits or individual items (icons, graphic images, text item, links, etc.) contained within an HTML page, parameters passed to/from the HTML page, etc. The analysts at Washingtonpost determined that the information they were most interested in were the actual pages viewed by the visitors to their site, and so that is the key element of data parsed from the web logs. In addition, business rules were established to derive other key indicators, including users and sessions. This was an additional challenge as there were no cookies or user registration in place to give precise information about a specific user. The applied logic required combining a multitude of common fields found in the web logs.

Data Summarization

The Washingtonpost.com web site is available 24 hours a day,7 days a week, 365 days a year. It never closes, and is continuously available for visitors to access. The summarization of the web traffic is based on predefined levels of categorization and time dimensions, i.e. the number of page views per hour, work shift, day, month, etc. The decision was made that the smallest time dimension would be an hour, with the next smallest being a shift (8 hour segment of a day), then day, week, and finally month. The data summarization process starts with a daily deadline, a point in time in which a copy of the internal web logs are pulled from the web servers, and the vendor web logs are collected. Then the SAS processes to read and parse these web logs are started. The SAS application processes the compressed web logs, applies the appropriate categorization and then rolls this daily data up through the various time dimensions and append the data to the historical data for the time series analysis and reporting. For the average daily run of 35 million web log records, the entire process is completed within four to six hours, ensuring timely reporting of the daily web log traffic.

Operational Data Definitions

Web log data and analysis is a perishable commodity; timeliness is key to realizing the value contained within the mountains of data. This is not a significant problem when processing the internal web logs of the Washingtonpost.com web site, but a significant portion of the total web log data that is analyzed comes from external vendor web logs. As this data resides on web servers beyond the control of the Washingtonpost.com, occasionally some of the vendor data does not arrive prior to the processing deadline. The existing data can not be allowed to age due to the tardiness of a vendor, and yet the reports can not show missing data simply because a vendor's data was not available.

The solution to this dilemma was to extrapolate the missing data based on historical vendor data, and annotate the reports to identify those report elements that are estimations. As the vendor data becomes available, the reports and historical data repository are updated, and the accuracy of the reports is improved without impacting the timeliness of the reports as well.

The Outputs

Washingtonpost.com needed several options for its internal reporting needs from their web site. Static reporting formats included HTML and Microsoft Excel reports, plus they needed the ability to perform ad hoc queries online. The static reports served to satisfy the common information requests regarding detailed information by month, week, day, or even hour; session duration; the starting or ending URLs; page views by host, domain, average user, session; popular features, sections, and subsections. The on-line ad hoc query capability satisfied the need to provide highly customized reporting for individuals and special needs.

Strengths Of SAS

The SAS system was chosen over other products for this e-intelligence solution because of the myriad of challenges faced by Washingtonpost.com as they sought to satisfy the business requirements defined for their web site. They needed to be able show activity trends and web traffic loads to support advertising revenue. Management needed to know the effectiveness of their web site, plus the total web activity on their servers. This required a solution that could implement rapid data warehousing, handle large volumes of data, provide exceptional data performance, access data on any platform and in any format with extensive data mining capabilities. SAS easily met these challenges. As an industry leader in decision support, SAS was able to meet all the information needs of the management team for the Washingtonpost.com web site.

Conclusion

An extremely busy web site such as Washingtonpost.com presents unique and formidable challenges to those who wish to understand what is happening within such a web site. The shear volume of data and the myriad of data sources and formats can thwart the ability of some solution providers to make sense of it all. The SAS solution, however, has the tools to meet this challenge. The SAS statistical tools and data mining tools provided all the analysis that was required. The data access tools allowed reading and manipulating data in whatever format they were found. And the SAS language itself proved to be an adaptable programming tool for meeting unanticipated challenges.

Contact Information

Qualex Consulting Services, Inc.
109 Salem Town Court
Apex, North Carolina
USA 25702
E-mail: info@qlx.com

Web site: http://www.qlx.com