Multi-Dimensional Analysis of ISA Server Logs

Salman Ahmed Shaikh and Dr. Manzoor Hashmani

Abstract—Network administrators are always interested in analyzing the activities and bandwidth usage of network users in order to manage network. Usually, third party tools are being used for such analysis, but third party tools suffer from slow response and lack of customization options. In this paper, we present a novel approach for the effective and very flexible multi-dimensional analysis of network usage, by directly analyzing the ISA Server Proxy Logs. In this approach, we accumulate the ISA Server logs into SQL Server and transform it into the dimensional model using SQL Server Integration Services. Dimensional model is a database modeling technique for the efficient analysis of large datasets. In order to evaluate the effectiveness of our proposed approach, we analyze ISA Server Proxy logs generated at a local university.

Keywords—Dimensional Analysis, Dimensional Modeling, ISA Server Logs, Proxy Server Logs, SQL Server Integration Services.

I. INTRODUCTION

In local area networks, usually proxy and/or cache servers are deployed for network resource optimization and internet usage reduction. These servers generate very detailed network usage and access logs for all the requests that pass through them. The sort of information that we can obtain from these logs include information about the requesting host, the resource being requested, and the date and time of the request [12]. Similarly, logs are generated by Web Servers for all the requests made onto it, which can reveal very useful navigation patterns and trends of visitors on the website, commonly termed as web mining [3]. Web mining is being used for long to uncover the search pattern and navigation behavior of the users, to provide the users with improved web site design and target marketing [4], [9], [6]. In the similar fashion, analyses of proxy and/or cache server logs can be utilized to better manage web requests and network resources.

In this paper, we present an approach for the efficient analysis of ISA Server logs. ISA Server generates very detailed security and access logs for all the traffic that passes through its different services [5]. The logs are generated in the form of unstructured flat text files and contain millions of records for a medium size network. Normally, these logs are analyzed using third party tools and tools need to be customized according to the requirements and the type of analysis required. Due to the large number of records, the response of these tools is too slow; furthermore the customization options available in these tools are also limited. In our approach, we connect ISA Server to the SQL Server so that the logs can be accumulated into the database directly. Once the logs are in the structured form, he database model is transformed into the dimensional model using SQL Server Integration Services. Dimensional model is a database design technique used to enhance the query response and for the multi-dimensional analysis of data. The approach is tested on the real time ISA Server proxy logs obtained from a local university and produced very useful multi-dimensional reports and graphs.

II. ISA SERVER LOGS

ISA Server generates comprehensive security and access logs for all traffic that passes through the firewall service and the web caching service. These logs can be generated on daily, weekly, monthly or yearly basis. Depending on the type of ISA Server configuration, there are three types of logs available, 1) Packet Filter, 2) Firewall and 3) Web Proxy.

<table>
<thead>
<tr>
<th>Field</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-ip</td>
<td>172.16.100.5</td>
</tr>
<tr>
<td>cs-username</td>
<td>anonymous</td>
</tr>
<tr>
<td>c-agent</td>
<td>Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)</td>
</tr>
<tr>
<td>Date</td>
<td>2008-04-22</td>
</tr>
<tr>
<td>Time</td>
<td>02:56:22</td>
</tr>
<tr>
<td>s-computername</td>
<td>MUET-03</td>
</tr>
<tr>
<td>cs-referred</td>
<td>-</td>
</tr>
<tr>
<td>r-host</td>
<td><a href="http://www.prizebond.net">www.prizebond.net</a></td>
</tr>
<tr>
<td>r-ipv4</td>
<td>75.126.158.180</td>
</tr>
<tr>
<td>r-port</td>
<td>80</td>
</tr>
<tr>
<td>time-taken</td>
<td>546</td>
</tr>
<tr>
<td>cs-bytes</td>
<td>453</td>
</tr>
<tr>
<td>ac-bytes</td>
<td>28914</td>
</tr>
<tr>
<td>cs-protocol</td>
<td>http</td>
</tr>
<tr>
<td>s-operation</td>
<td>GET</td>
</tr>
<tr>
<td>cs-uri</td>
<td><a href="http://www.prizebond.net/ima">http://www.prizebond.net/ima</a> gent1/style.css</td>
</tr>
<tr>
<td>s-object-source</td>
<td>VCache</td>
</tr>
<tr>
<td>ac-status</td>
<td>364</td>
</tr>
</tbody>
</table>

Fig. 1. Web Proxy log’s attributes.

Packet Filter log contains information regarding the packets that ISA Server examines. Firewall logs possess information regarding all the traffic sent through the firewall service. Web Proxy log contains almost the same fields as that of ISA Server firewall log, but these logs are specific to the web proxy server [5]. In this paper, only the ISA Server Web Proxy logs have been used for the
multi-dimensional analysis of the most frequently visited hosts, for the identification of ports used by futile web sites, for analyzing the protocols used for browsing and for the identification of top bandwidth consumers within the local network. Beside these, hundreds of other useful analysis may be obtained from ISA Server Proxy log which can help the network administrator in managing the network. Figure 1 shows the attributes of the ISA Server Proxy log along with sample values.

One of the major problems when dealing with these types of logs is size and unstructured data [12], [2], [7], [8], [11], [4]. A typical proxy log file of a medium size LAN contains millions of records. Since the log file is generated on daily basis, the number of records grows to tens of millions of records within a month. Analysis of such a large dataset poses a problem as the standard analysis tools are unable to handle huge datasets [12]. Moreover, the available tools are quite inflexible and can only perform some of the predefined analysis e.g. SpeedTracer from IBM [1]. In order to analyze the logs in an efficient and flexible way, the size of the dataset either need to be reduced [4] or we need to adopt the approach which can help us handle and analyze huge datasets and provide us with customizab

III. DIMENSIONAL MODELING

An efficient technique to handle and analyze large datasets is dimensional modeling. Dimensional modeling has proved to provide high performance for queries and speed-of-thought analysis [10]. Dimensional models let us analyze large datasets efficiently by reducing number of joins among tables and by making use of special signed aggregate fact tables and selection of dimensions significantly improves query performance and analysis capability of dimensional models. Moreover, dimensional models help in analyzing facts with respect to multiple dimensions simultaneously and allow us to drill-down, roll-up and slice & dice within and across the dimensions very easily [10].

keys known as surrogate key. Besides, usage of carefullyd

By default ISA Server generates detailed security and access logs in the form of flat text files, but ISA Server can be con-figured to record the logs into any of the supported RDBMS database, we have used SQL Server 2005 for this purpose. We connected the ISA Server to the SQL Server 2005 database in order to accumulate web usage logs directly into the database. The schema of the accumulated logs is then transformed into the dimensional model using the SQL Server 2005 Integration Services as shown in figure 2. Figure 3 depicts the resulting dimensional schema for the ISA Server Proxy logs.

IV. MULTI-DIMENSIONAL ANALYSIS

The dimensional model in figure 3 can be used for the efficient multi-dimensional analysis of ISA Server logs. Some of the useful analyses are as follows.

A. Ports Summary

As listed in the figure 4, ports 80, 21 and 443 are the top three most heavily used whereas port numbers 9816, 8090 and 8101 are the 4th, 5th and 6th most heavily used ports respectively and all three of these ports are unassigned by Internet Assigned Numbers Authority (IANA). Out of the top 15 ports listed in table2, 7portsareunassigned.
B. Protocol Summary

The bar graph in figure 5 shows the ports used by users for browsing. It is obvious from the graph that normally HTTP protocol is being used for browsing. Besides, SSL, HTTPS and FTP are also being used but their usage is quite insignificant. The last bar in figure 5, with caption '-' is used by loopback address 127.0.0.1.

Fig. 5. Protocol Summary.

C. Top hosts on port 80

The multi-dimensional list report in figure 6 summarizes the top 15 hosts with respect to bandwidth consumption in a week on port number 80. As a result of careful analysis of the figure 6, it has been revealed that

Fig. 6. Top hosts on Port 80

Bandwidth utilized by vulgar websites = 20.01 GB
Bandwidth utilized by other websites = 10.62 GB

Hence the ratio of the vulgar versus other sites visited = 2:1. To summarize, approximately 66% of the whole bandwidth is utilized on surfing futile stuff, hence we are left with just 34% bandwidth for other useful surfing.

D Top Bandwidth Consumers

Some users make use of download accelerators; play audio, video and games online, which consumes a lot of network bandwidth and results in problem for other users. Figure 7 lists the top bandwidth consumers of the week within the university network. The multi-dimensional network usage profile can help in identifying the top bandwidth consumers and the amount of traffic they are generating on the network with respect to date which can help the network administrators in defining and/or modifying the existing network policy.

Fig. 7. Top bandwidth consumers.

V. EVALUATION / RECOMMENDATIONS

Based on the evaluation of logs, below are some of the advices and suggestions for the organizations

1) An organization should perform detailed analysis of their network usage logs regularly in order to find network usage anomalies.

2) Using the proposed solution, organizations need not to buy third party analysis tools, but all the tasks i.e. from logs parsing and transformation to multi-dimensional analysis can be done using SQL Server.

3) More flexible and customizable reports can be obtained from the given solution in contrast to rigid pre-defined reports available in third party analysis tools.

4) There is no upper limit on the data size and the number of logs an organization wants to store and analyze, which is one of the very common problems in third party analysis tools.

Eventually, we would like to recommend the following to the university whose logs we have analyzed

1) Heavily used ports should be analyzed regularly for the identification of useless hosts. e.g. approximately 97% of the whole network traffic travels through port number 80 and out of this traffic, 66% of traffic is generated by useless hosts. If
the port 80 is analyzed on regular basis, would help us in saving a lot of network bandwidth. 

From the analysis of different ports traffic, it has been found that unassigned ports are normally used by streaming sites, consuming a lot of network bandwidth, hence they may be blocked. e.g. Port numbers 9816 and 8090. 

3) University should adopt some mechanism to uniformly distribute bandwidth among users, so that some users may not be able to consume whole of the network bandwidth.

VI. CONCLUSION

In this paper we present an approach to analyze the network usage logs generated by ISA Server. ISA Server generates very detailed access and security logs, but these logs are underutilized. One of the main reasons of this underutilization is the unavailability of efficient analysis tools or the inability of the available tools to handle huge datasets. The work presented here is capable of analyzing huge logs datasets in an efficient way by utilizing dimensional model. Using the proposed approach, the ISA Server logs are directly loaded into the SQL Server 2005 database and then transformed into the dimensional model using SQL Server Integration Services. In last, we have shown several multi-dimensional reports and graphs obtained from the analysis of ISA Server Proxy logs of a local university.

REFERENCES


Salman Ahmed Shaikh was born in Hyderabad, Pakistan on 28th February, 1984. He received his B.E. (Computer Systems Engineering) in 2005 and M.E. (Communication Systems and Networks) in 2008 from Mehran University of Engineering & Technology, Jamshoro, Pakistan. He is working as LECTURER in the Department of Computer Systems and Software Engineering, Mehran University of Engineering & Technology, Jamshoro, Pakistan since 2006. He has also worked as SOFTWARE ENGINEER in one of the leading software house InfiniLogic Private Limited, Karachi, Pakistan. His re-search interest includes Data Warehousing, Dimensional Modeling and Multi-Dimensional/OLAP Analysis.

Mr. Shaikh is a member of International Association of Computer Science and Information Technology (IACSIT) and Pakistan Engineering Council.

Prof. Dr. Manzoor Hashmani was born in Hyderabad, Pakistan on 6th March, 1967. He received his B.E. (Computer Systems Engineering) from Mehran University of Engineering & Technology, Jamshoro (Pakistan) in 1991, M.E in 1997 and Ph.D. in 1999 from Nara Institute of Science & Technology, Nara (Japan). He is working as FOREIGN PROFESSOR in the Department of Computer Systems and Software Engineering, Mehran University of Engineering & Technology, Jamshoro, Pakistan. He has authored and co-authored more than 30 research papers published in various journals and conferences of international repute. He has also worked as lead research and development person in a reputable Japanese company for five years. His research areas of interest include High Speed Communication Networks, Software Engineering and Alternative Energy.

Dr. Hashmani is a member of IEICE (Japan), IEEE Communications Society (USA), and Pakistan Engineering Council.