Multi-Dimensional Analysis of ISA Server Logs

Salman Ahmed Shaikh and Dr. Manzoor Hashmani

***Abstract*—Network administrators are always interested in analyz-ing 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.

Salman Ahmed Shaikh is with the Dept. of Computer Systems and Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

**Dr. Manzoor Hashmani is with the Dept. of Computer Systems and Software Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan.

II. SA 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.

Field	Data
c-ip	172.16.100.5
cs-username	anonymous
c-agent	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
Date	2008-04-22
Time	02:58:22
s-computername	MUET-03
cs-referred	-
r-host	www.prizebond.net
r-ip	75.126.158.180
r-port	80
time-taken	546
cs-bytes	453
sc-bytes	28914
cs-protocol	http
s-operation	GET
cs-uri	http://www.prizebond.net/images1/style.css
s-object-source	VCache
sc-status	304

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

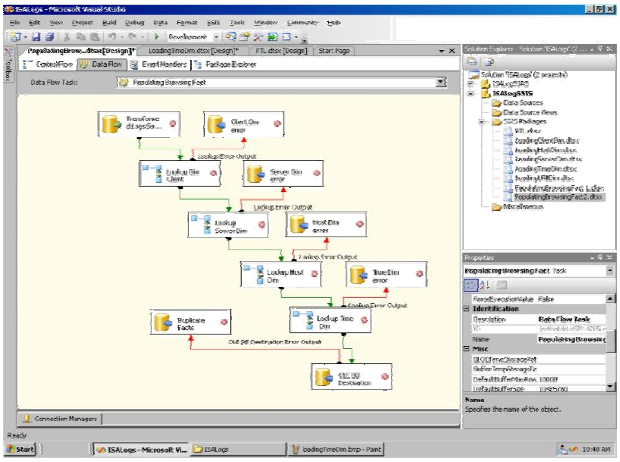

Fig. 2. Schema Transformation using SQL Server Integration Services

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

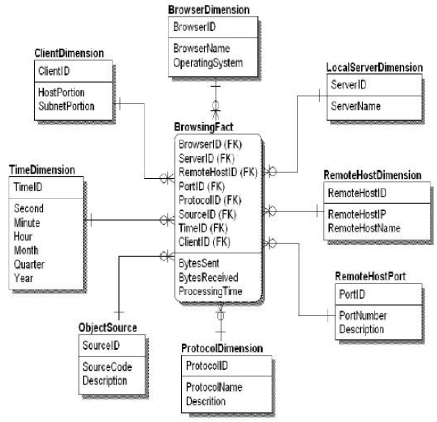

Fig 3: ISA Server Proxy Log Dimen Dimensional Model for Web Proxy Logs

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.sionalMod

Fig. 4. Ports' data traffic summary for a week.

Port Numbers	Total Bytes Received	Rank (Total Bytes Received)
80	292135291770	1
21	4357551151	2
443	1852795449	3
9816	197361889	4
8090	101594380	5
8101	64877760	6
1935	64591183	7
8088	62226851	8
8080	45700594	9
8085	37820535	10
81	18136756	11
8081	16091741	12
8388	9929062	13
9998	8281429	14
88	8057044	15

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.

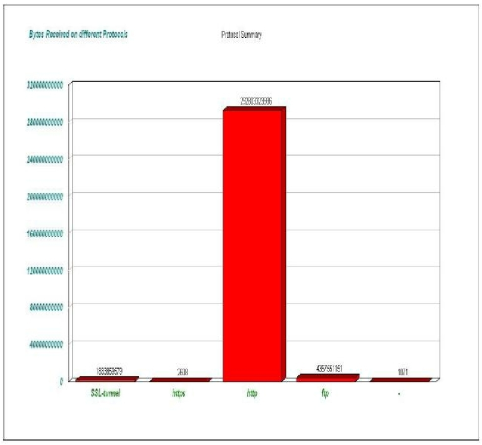

Fig. 5. Protocol Summary.

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

	Port # 80							
	22nd July, Monday	23rd July, Tuesday	24th July, Wednesday	25th July, Thursday	26th July, Friday	21st July, Saturday	Rank	Total Bytes Received
stream.wmlivesvc.vitalstreamcc	599455789	1924913464	2418189466	5219	588248217	1778031183	1	7308843338
download.microsoft.com	363346077	373848379	2710099463	62447363	2500768056	320490735	2	6331000073
www151.megaupload.com	992808316	839161265	1677722752	347969203	0	10891064	3	3868552600
dl.search-download.org	37119737	1766506	0	0	2114992285	110412187	4	2264290715
www.nakedwebtv.com	0	858166444	88107420	0	960850592	0	5	1907124456
dal-v10.dal.youtube.com	223395461	1128276928	144197158	0	0	0	6	1495869547
www.chaltatv.com	5841424	210749376	1228470205	0	93373	0	7	1445154378
software-files.download.com	123540048	502391161	338687944	126911305	208374237	85624341	8	1385529036
www132.megaupload.com	0	0	419430688	419430688	422332292	0	9	1261193668
akmccvideos.metacafe.com	198331722	0	184697646	7409977	707661483	0	10	1098100828
ardownload.adobe.com	109443272	265736107	239856465	84640947	193503064	174859076	11	1068038931
ejang.jang.com.pk	116684057	212018175	148937527	83149530	186265755	169526123	12	916581167
pagead2.googlesyndication.com	77042015	197838866	171018335	86939736	200197391	133130742	13	866167085
definitions.symantec.com	92630302	216845954	225411225	99093471	51826423	157130991	14	842938366
www.videosnstuff.com	2408858	110064108	0	0	729124913	0	15	841597879
Bytes Received / Day	13468227158	33092730918	29595649264	11200908010	31843182990	16822579691	na	136023278031

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 graph 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 policie

	21 July, Monday	22 July, Tuesday	23 July, Wednesda	24 July, Thursday	25 July, Friday	26 July, Saturday	Rank	Total Bytes Received KB
172.16.100.5	15348403	11780615	31354580	28253156	10269956	30454114	1	127460824
172.16.4.1	117141	1014109	1021751	807252	33406	665206	2	3658865
172.16.5.147	0	235	173679	87847	284879	246505	3	793145
172.16.5.6	0	0	1383	104373	271699	381874	4	759329
172.16.99.30	47895	16678	21851	18190	23845	80271	5	208730
172.16.99.4	53458	16774	4981	59443	18673	15729	6	169058
172.16.9.212	0	0	0	12234	82006	58156	7	152396
172.16.100.4	1981	5188	1890	55888	113	80001	8	145061
172.16.98.250	0	62059	10620	16794	14816	18603	9	122892
172.16.5.125	0	47954	45434	0	0	0	10	93388
172.16.10.9	1254	2237	5414	23296	25774	22847	11	80822
172.16.6.101	359	17835	15798	39403	3767	1880	12	79042
172.16.99.2	12744	0	21359	28835	8544	0	13	71482
172.16.4.200	28411	0	5840	11285	14824	4285	14	64645
172.16.6.26	18208	4983	3882	5438	4126	23519	15	60156
172.16.99.1	17750	9148	8406	0	2170	0	16	37474
172.16.4.90	5545	9548	7671	5583	0	0	17	28347
172.16.99.6	2456	0	14036	1216	673	9489	18	27870
Total Bytes Received	16644445	13324500	32878768	29587048	11065423	32092904	na	135593088

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

2) 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

[1] Jesper Andersen, Anders Giversen, Allan H. Jensen, Rune S. Larsen, Torben Bach Pedersen, and Janne Skyt. Analyzing clickstreams using subsessions. In DOLAP '00: Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP, pages 25–32, New York, NY, USA, 2000. ACM.

[2] V. Bacarella, F. Giannotti, M. Nanni, and D. Pedreschi. Discovery of ads web hosts through traffic data analysis. In DMKD '04: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, pages 76–81, New York, NY, USA, 2004. ACM.

[3] Farah Habib Chanchary, Indrani Haque, and Md. Saifuddin Khalid. Web usage mining to evaluate the transfer of learning in a web-based learning environment. In WKDD '08: Proceedings of the First International Workshop on Knowledge Discovery and Data Mining, pages 249–253, Washington, DC, USA, 2008. IEEE Computer Society.

[4] Lim Chinghway, N. Singh, and S. Yajnik. A log mining approach to failure analysis of enterprise telephony systems. Dependable Systems and Networks With FTCS and DCC, pages 398–403, 2008.

[5] SANS Institute. Using isa server logs to interpret network traffic, 2002.

[6] Xin Jin, Yanzan Zhou, and Bamshad Mobasher. Web usage mining based on probabilistic latent semantic analysis. In KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 197–205, New York, NY, USA, 2004. ACM.

[7] Working Paper, Graduate School of Industrial Administration, Carnegie Mellon University, 1999.

[8] Olfa Nasraoui, Osmar R. Za¨ıane, Myra Spiliopoulou, Bamshad Mobasher, Brij Masand, and Philip S. YU. Webkdd 2005: web mining and web usage analysis post-workshop report. SIGKDD Explor. Newsl., 7(2):139–142, 2005.

[9] G. Paliouras, C. Papatheodorou, V. Karkaletsis, C. Spyropoulos, and P. Tzitziras. From web usage statistics to web usage analysis. In IEEE SMC '99: Proceedings of the 1999 IEEE international conference on Systems, Man, and Cybernetics, pages 159–164, Tokyo, Japan, 1999. IEEE Computer Society.

[10] Paulraj Ponniah. Data Warehousing Fundamentals. Replika Press, 2003.

[11] M. Quafafou, S. Naouali, and G. Nachouki. Knowledge datawarehouse: Web usage olap application. In WI '05: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pages 334–337, Washington, DC, USA, 2005. IEEE Computer Society.

[12] Arun Sen, Peter A. Dacin, and Christos Pattichis. Current trends in web data analysis. Commun. ACM, 49(11):85–91, 2006.

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 Hyder-abad, 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 Engery.

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