ISSN 1605-8607

A STUDY RELATED TO PROTOCOLS AND PORTS USAGE FOR THE ISA SERVER

Pardeep Kumar*, Iftikhar A. Koondhar**, Akhtar H. Jalbani***, Agha Shiraz A. Khan****, M. Aamir Bhutto*****

ABSTRACT

Managing and monitoring networks have never been so important as today. Although there are variety of tools available in the market to secure and monitor the network from malicious attacks but these tools do not provide detailed analysis capability of network usage patterns and trends. Analysis of network usage logs to uncover the activities of internal users is a good way of managing the network security and bandwidth related issues. ISA Server generates very detailed access and security logs for all the web requests made through it, but it doesn't provide efficient and interactive analysis capability of the same. This paper aims to provide multidimensional and interactive (providing drill-down, roll-up and slicing & dicing) analysis capability of proxy logs generated by ISA Server.

The solution suggested in this paper can generate very customizable and interactive multidimensional reports, which can help in better understanding the network usage patterns and trends of the users including the details about the website, ports, protocols, bandwidth usage and malicious activities attempted by the users while accessing the network. Moreover, the analysis of ISA Server logs has revealed the ports and protocols commonly being used at the Server. In addition it has been found that the unassigned ports by IANA are never used for any useful browsing, hence they may be blocked.

1. INTRODUCTION

ISA (Internet Security and Acceleration) Server from the Microsoft Corporation is the follow-on release of their Proxy Server 2.0 and is also a part of the .Net Family. Note the all the abbreviations that have been used in this paper are given in Table 1. Web log mining is an important application in the Web Mining research fields. The aim of Web log mining is to find out user access patterns of Web sites. The process mainly includes four steps: data collection, log preprocess, pattern recognition and pattern analysis. First, we extract available information from raw access logs; insert them into databases, certain each user based on heuristic rules, and dig out each user's access sequence [12].

The basic services of ISA Server include an enterprise firewall and a Web proxy (a.k.a. cache server) [1, 2, 10]. ISA Server's firewall monitors all the network traffic that flows through it whereas the Web cache provides storage and faster access to the frequently accessed web pages in order to lessen the network rush. The ISA Server does this by downloading the updates of frequently accessed web pages when the server is in idle state.

Beside this, ISA Server provides very comprehensive security and access logs for all the traffic that passes through it. These logs contain tons of data which can reveal hidden patterns if analyzed properly including:

Table 1: Abbreviations used in the paper

ELT	Extract Load Transform
ETL	Extract Transform Load
FTP	File Transfer Protocol
HTTP	Hyper Text Transform Protocol
IANA	Internet Assigned Network Authority
ICPC	Information Communication Processing Center
IP	Internet Protocol
ISA	Internet Security and Acceleration
LAN	Local Area Network
RDBMS	Relational Database Management System
RPC	Remote Procedure Call
SMB	Server Message Block
SQL	Structured Query Language
SSIS	SQL Server Integration Services
TCP	Transmission Control Protocol
URL	Uniform Resource Locators

*,** Assistant Professor, Department of CSE, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh
*** Associate Professor, Department of IT, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh
****,***** Lecturer, Department of CSE, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Sindh

- URL(s) visited by the users
- Protocol(s) used for communication
- Port(s) used for accessing hosts
- Browser used to made request
- On-Time and Off-Time internet usage details
- Get notified on excessive internet usage

These results can help network administrator in securing the ISA server from unauthorized use through LAN service ports, saving the network bandwidth which may be consumed heavily by the unidentified protocols and providing caches for the most often visited pages.

Logs Collection is the very first phase of this research. Detailed security and network usage logs are generated for all the traffic that passes through the firewall service and the web-caching service of the ISA Server. These logs are in the form of flat text files, which should be parsed and formatted before they may be analyzed. Logs can be generated on daily, weekly monthly or yearly basis [10]. ISA Server generates three types of logs depending on the type of installation and configuration, namely *packet filter*, *firewall* and *proxy logs*.

1.1 PACKET FILTER LOGS

ISA Server logs each and every packet that falls against the packet filtering rules. Packet filtering logs contain only the dropped packet however it may be configured for logging allowed packets.

1.2 FIREWALL LOGS

These logs contain detailed information regarding the data sent through the ISA server's firewall service. Firewall logs include user and host IP address, ports and protocol used for communication, bytes sent and received for any request, time taken to fulfill any request and the date and time of browsing. Beside this firewall service log properties can also be adjusted.

1.3 PROXY LOGS

Proxy logs contain almost the same attributes as that of firewall logs but these logs are specific to the configuration of the web proxy server.
All the above mentioned types of logs can be obtained in two formats, i.e., (1) flat text file and (2) ISA Server can be configured to move the logs to any RDBMS including Oracle, SQL Server.
In this research only the details and usage of proxy logs for Internet are analyzed. These logs are available in the form of flat text files; however ISA Server may also be configured to directly move the logs to some predefined database of any RDBMS. The flat files are generated on daily basis i.e., a single flat text file is generated daily that contains the details of all the hits made on ISA Server on that particular day. The following screen shot depicts a typical flat text proxy log generated by ISA Server.

2. RESEARCH METHODOLOGY

The overall research methodology consists of following four phases as shown in Figure 1.

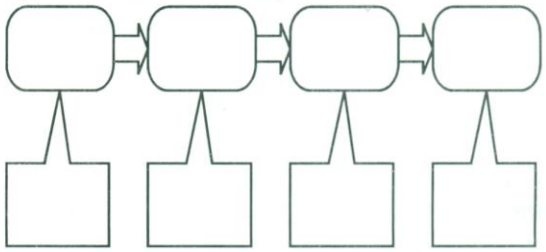

Figure 1: Research Methodology

A. Log collection
B. Data loading
C. Transformation
D. Analysis

2. 1 LOG COLLECTION

During the first phase, log information about the web usage is collected from the ISA server in order to process it further for the data loading phase. Figure 2 shows a snap shot of the ISA log file.

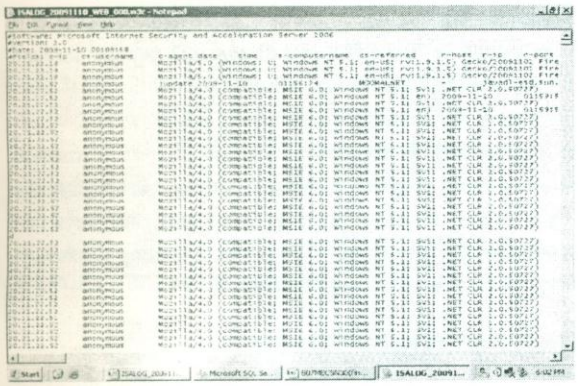

Figure 2: A snap shot of the ISA log file

2. 2 DATA LOADING

The flat text logs generated by ISA Server need to be loaded into database of any RDBMS so that it may be parsed, transformed and analyzed. In this research we have used SQL Server 2005 as the target RDBMS. SQL Server 2005 provides a lot of services for the loading and transformation of data. In this research we have used SQL Server Integration Services (SSIS) for the loading and transformation of ISA Server logs.

For loading the logs, a thorough study of the same is required. The logs need to be studied for the attributes available, their field width and the data types. Since every log file contains millions of records, the loading phase is one of the most time consuming phase of the whole ELT cycle. The logs are initially loaded into the staging area and then transformed and moved to the final database. Figure 3 shows the loading phase of ELT cycle using SQL Server 2005. Here the source is the flat text file and the destination is SQL Server 2005.

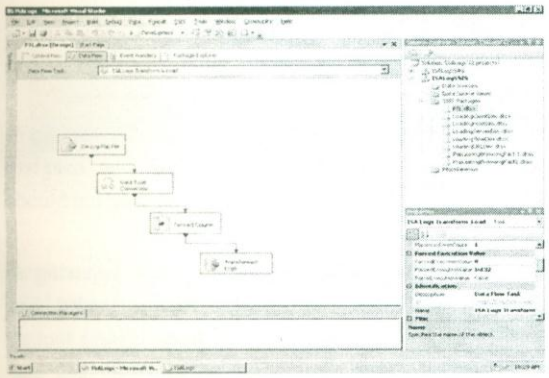

Figure 3: Loading data through SQL server integration service

2. 3 DATA TRANSFORMATION

Once the data is in SQL Server staging database (temporary database), it needs to be transformed for the following reasons.

The data loaded is in a single table in the staging area and it needs to be transformed according to the multi-dimensional schema. Some of the attributes in the staging area are useless, so they need to be dropped. Date and time formats conversion is needed. Aggregates of numerical attributes are needed for fast and quick analysis etc.

Figure 4 depicts the transformation phase used in this process. This transformation will convert the schema of the database and is going to populate the fact and dimension tables.

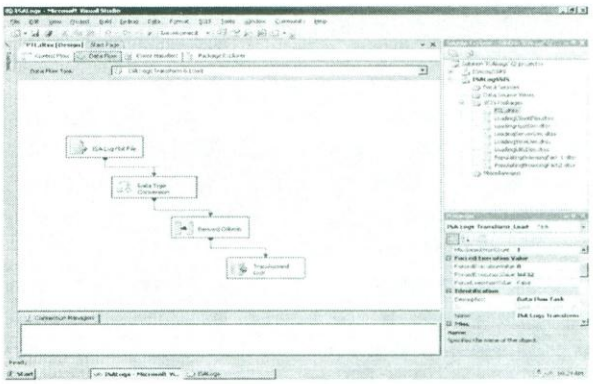

Figure 4: Transformation of Logs using the SQL server integration services

2. 4 ANALYSIS

The final phase of the research methodology is the data analysis, which is achieved by using one of the most powerful data analysis tool named COGNOS. COGNOS is a multi - dimensional reporting tool that delivers managed reporting for consistent, fact-based decision-making. Managed reporting enables report authors to create reports drawn from any data source. These reports can then be delivered to report consumers.

Data Analysis is one of the most important phases. A lot of textual and graphical reports are generated while analyzing the logs which discovered many hidden network usage pattern and trends.

3. SCHEMA TRANSFORMATION

As stated earlier, complete ISA logs are loaded into a single table, means there is no logical schema of the loaded data. Moreover absence of logical model will also result in data redundancy and analysis chaos.

For the analysis of huge amount of data it is always recommended to use dimensional model. Since dimensional model is denormalized, it results in faster analysis of data. The dimensional model used in this paper is as follows:

3. 1 ISA LOG'S DIMENSIONAL MODEL

In the schema above, there is a central fact table containing the foreign keys and numerical attributes surrounded by four dimension tables namely Client, Host, Server and Time dimension. Customarily, the central fact table contains numerical measures or the attributes on which we want to perform calculations. In my case the factual attributes include Bytes Sent, Bytes Received and Time Taken in fulfilling any request. Fact table is always surrounded by dimension tables which contain non-numeric or textual attributes. As the ISA log contains attributes related to client, server and host; hence there are four dimensions client, server, host and time dimension. Time dimension is considered as the mandatory dimension of all dimensional models. Figure 5 depicts the dimensional model of the ISA log.

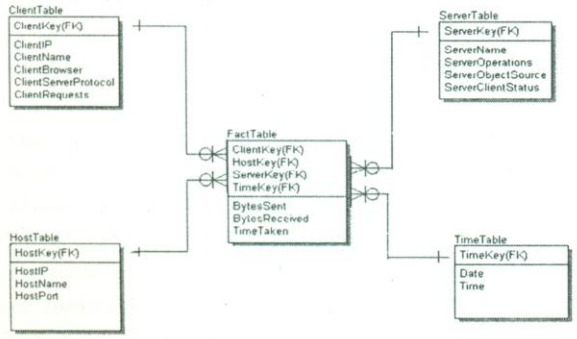

Figure 5: Dimensional model of the ISA log

3. 2 SURROGATE KEY GENERATION

In case of dimensional model, fact and dimension tables are linked by a special type of key known as surrogate key. Surrogate keys are simple sequential keys used by dimension tables. This key helps in linking fact and dimension tables. Due to the simplistic nature of surrogate key, joining between fact and dimensions become fast and data retrieval efficiency increases.

Surrogate key generation is one of the transformation stages. In this stage an additional attribute is added to the dimensions. In this project four surrogate keys are generated for four dimension tables including ClientKey, ServerKey, HostKey and TimeKey

3. 3 DATA TYPE CONVERSION

The data type of the data loaded in the loading stage is usually character type. However for different attributes there need to be different data types, e.g. for numerical attributes there need to be integer data type, for date time related attributes there need to be date data type.
As mentioned above, SQL Server Integration Services (SSIS) may be utilized for all of the above discussed transformation including schema conversion, surrogate key generation and data type conversion. The screenshot shown in Figure 6 shows the schema conversion and surrogate key generation.

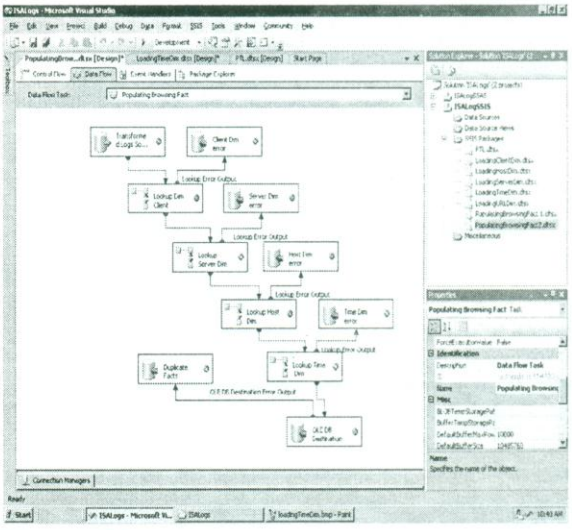

Figure 6: Data type conversion

In the above screenshot, data staging area is behaving as the source. From the staging area data is loaded into different dimensions and fact tables. In each stage lookup table is used for the generation of surrogate key and finally the data is loaded into the dimensional model.

4. LOG ANALYSIS

ISA Server Proxy logs are analyzed for network/internet usage patterns. Some of the analysis includes Bandwidth usage with respect to ports and protocols Sites accessed through unassigned ports popular hosts on popular ports etc.

4. 1 PORT TRAFFIC SUMMARY

The report shown in Table 2 summarizes the network traffic on different ports. Moreover ports are ranked with respect to the bytes received on them while browsing. Port numbers 80, 443 and 1935 are the most traffic bearing ports respectively.

4. 2 PORT CLASSIFICATION

The IANA has divided the port numbers into three ranges well-known ports, registered ports and dynamic (or private) ports.

Table 2: Port summary

Port Number	Bytes Sent	Bytes Received
21	343515399	2872
80	1853719555	43074735
436	0	0
443	8277698	1671041
1935	6392606	316322
8080	220405	11134
8090	0	0
8101	0	0
9816	0	0
Total	2212125663	45076104

Well-known ports: The ports ranging from 0 to 1023 are assigned and controlled by IANA; Well-known ports should not be used without IANA registration.

Registered ports: The Registered ports are those from 1024 through 49151; registered ports should not be used without IANA registration.

Dynamic (or private) ports: The Dynamic and/or Private Ports are those from 49152 through 65535. Dynamic ports are neither controlled nor registered. They can be used by any process [11].

As listed in Table 2, 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 IANA. It is interesting to note that unassigned ports are normally used by useless sites e.g. port number 9816 is used exclusively by host '*streamsolutions.co.uk*' for audio and video streaming, port number 8090 was used by '*videos.asiantsunamivideos.com.nyud.net*'.
Table 3 draws top 15 heavily ports used by the port 80. Out of the 15 ports, 7 ports are unassigned which may be blocked to save some of the network bandwidth.

Table 3: Top 15 hosts for the port 80

Port 80	Received Bytes	Sent Bytes
www.facebook.com	296	2107
www.rapidshare.com	602	33806
www.google.com	536	7932
www.youtube.com	284	117222
www.orkut.com	287	49872
www.wikipedia.com	370	14440
www.onlinewatchmovies.net	415	978
www.islamonline.com	297	6277
www.tagged.com	409	119838
www.manjam.com	359	1239
www.9adultsexgames.com	545	48972
www.hentaicake.com	611	8609
www.hardcartoon.com	507	19034
www.myxxxtoon.com	406	22493
www.thumbparty.com	428	17078
Total	6352	470097

4. 3 POPULAR HOSTS ON PORT 80

The list report presented in Table 3 summarizes the top 15 hosts with respect to bandwidth consumption in a week on port number 80, which concludes that the almost 97% of the whole traffic travels through this port.

4. 4 HOST ON PORT 21

Table 4 presents the list of hosts and their traffic conditions on port 21. This port has specifically been used

only for the FTP communication between the server and clients.

Table 4: Top hosts for the ftp use on port 21

Port 21	Bytes Received	Bytes Sent
ftp://20.21.22.3/indian%movies	371	3686
ftp://20.21.22.3/indian%songs	420	1575
ftp://20.21.22.3/english%movies	371	3668
ftp://20.21.22.3/english%songs	355	3358
ftp://20.21.22.3/ring%tones	402	783
ftp://20.21.22.3/mobile%videos	204	915567
ftp://20.21.22.3/softwares	525	6532
ftp://20.21.22.3/mobile%converters	426	5849
ftp://20.21.22.3/mobile%softwares	356	6954
ftp://20.21.22.3/sindhi%songs	254	2587
ftp://20.21.22.3/sindhi%videos	110	6589
Total	3794	957103

4. 5 HOSTS ON PORT 443

The traffic usage on port 443 is depicted in Table 5.

Table 5: Top hosts for port 443

Port 443	Received Bytes	Sent Bytes
www.urs.microsoft.com	284	4665
www.secure.tagged.com	2164	3447
www.google.com	8682	10904
www.login.wetpaint.com	3590	19358
www.images.wetpaint.com	1416	6842
www.wetpaint.login.rpxnow.com	1230	15073
www.ajax.googleapis.com	836	22260
www.s3.amzonws.com	915	5583
www.sstats.wetpaint.com	2212	4353
www.rpxnow.com	895	361
www.facebook.com	1542	2544
www.client4google.com	9310	1633
www.beliscity.com	859	54599
www.loging.yahoo.com	995	16229
www.s.yimg.com	886	3227
www.login.live.com	2688	5274
Total	3770	179604

4. 6 PROTOCOL SUMMARY

Normally HTTP protocol is used for browsing. Beside http; SSL, HTTPS and FTP are also being used but their usage is not quite significant. Figure 7 presents the overall usage of the protocols that clearly shows that the HTTP protocol is mostly used. Moreover, no any anonymous protocol was found in the two weeks logs.

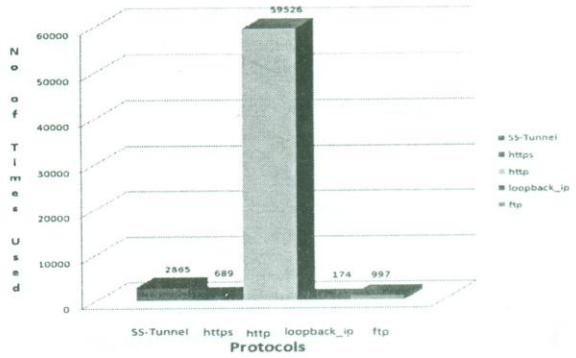

Figure 7: Protocol summary

5. FINDINGS OF THIS RESEARCH

As a result of analysis of list reports and graphs generated from the ISA Logs, a lot of patterns and trends have been discovered. Some of the findings of the analysis are discussed below.

5. 1 PORTS

In TCP/IP networks, a port is a transport layer address that the client program must specify in order to communicate with a specific server application on a computer in the network. In order to communicate on the web, one should specify the port number of the remote host. Normally port number 80 is used as the default port for World Wide Web; however some hosts make use of other assigned and unassigned ports to communicate on TCP/IP network.

5. 2 DATA ON ASSIGNED V/S UNASSIGNED PORTS

From the analysis of data traffic on different ports, it has been discovered that 80, 21 and 443 are the top three most heavily used assigned ports. Port numbers 80 and 8080 are used by a well known protocol HTTP for accessing the World Wide Web, approximately 97% of the whole network traffic travels through these ports. Port numbe

21 is the second most heavily used ports used by FTP. FTP is a file transfer protocol employed for uploading, downloading and manipulating files over a TCP network. Moreover port number 443 is employed by HTTPS-SSL protocol for encrypted communication of data over TCP/IP network for security reasons.

Port numbers 1935 and 8080 are the 4th and 5th most heavily used ports respectively and all four remain of these ports are unassigned by Internet Assigned Numbers Authority (IANA). It is interesting to note that unassigned ports are normally used by useless sites e.g. Port number 9816 is used exclusively by host 'streamsolutions.co.uk' for the audio and video streaming and port number 8090 was used by 'videos.asiantsunamivideos.com.nyud.net' for downloading tsunami videos.

5. 3 PORT NUMBER 80

Port 80 is the default port employed for browsing the World Wide Web. As port 80 is the most heavily used port, data on this port is analyzed for possible bandwidth threats. As a result of careful analysis of port 80 for a week, it has been revealed that approx 66% of bandwidth on this port is wasted on useless surfing and Most of porn sites or links are visited via this port.

5. 4 PORT NUMBER 21

Port number 21 is the second most used port and this is should also be monitored carefully so that any prohibited or restricted link or URL might be accessed via server.

5. 5 PORT NUMBER 9816

9816 is the unassigned port by Internet Assigned Numbers Authority (IANA) and this port is used exclusively by 'streamsolutions.co.uk' for audio and video streaming and which causes the use of bandwidth uselessly.

5. 6 PORT NUMBER 443

Port number 443 is another port by which a very large number of users are accessing the server. The traffic at this port is observed less as compare to port 80 and port 21.

5. 7 PORT 8090

8090 is another unassigned port like port number 8101, 9816 and 436, and these ports might be very dangerous for the attackers and which may be waste of bandwidth usage.

5. 8 PORT NUMBER 8080

Port 8088 is the unassigned port, sometimes used for accessing or downloading the audio and video files or movies.

5. 9 BANDWIDTH HUNGRY HOSTS

From the analysis of data movement on various ports, it has been found that a lot of hosts on different unassigned and assigned ports are useless. Top 15 most heavily visited hosts include:

- *stream.wmlivesvc.vitalstreamcdn.com*
- download.microsoft.com
- *www151.megaupload.com*
- *dl.search-download.org*
- *www.nakedwebtv.com*
- *dal-v10.dal.youtube.com*
- *www.chaltatv.com*
- software-files.download.com
- *www132.megaupload.com*
- *akmccvideos.metacafe.com*
- ardownload.adobe.com
- ejang.jang.com.pk
- pagead2.googlesyndication.com
- definitions.symantec.com
- *www.videosnstuff.com*

In the 15 hosts listed above, almost 9 hosts (*in italic*) are useless / filthy and are used for audio/video streaming and for uploading/downloading similar stuff. These hosts may be blocked to save the bandwidth.

5. 10 PROTOCOLS

In computer networks, a protocol is a set of rules that governs data communications. As a result of analysis of ISA logs, it has been found that the following protocols are commonly used for accessing the network. HTTP is

the mostly used protocol. FTP is the protocol used less than http SSL-Tunnel is the protocol which is used very rarely. Beside these, HTTPS is also being used, but its usage is negligible. Whereas, a hyphen "-" in the protocol field signifies the loopback address 127.0.0.1.

6. RELATED WORK

In [14] similar type of work have been done where the web server log of an academic institute have been analyzed. The obtained results can be helpful for variety of public or private applications. The usage of sever logs have also been analyzed in [16] for the decision making of the organization and the monitoring of user access modes. The research work in [15] describes the regular patterns of system features that describe program and the user behavior, large-scale log processing for analyzing the log records.

However, the study of log files of the ISA server for the ports, protocols and bandwidth usage factors is mainly missing in the literature. We, through this paper, have attempted to fill this gap.

7. RECOMMENDATIONS

Based on the analysis presented in this paper, following recommendations are drawn:

- Heavily used ports should be analyzed regularly for the identification of useless hosts. E.g. approximately 94% of the whole network traffic travels through port number 80 and out of this 94% traffic, 55% of data traffic is generated by useless hosts. If the port 80 is analyzed on regular basis, it would help us in saving a lot of network bandwidth.
- According to Internet Assigned Number Authority, "unassigned port numbers should not be used" [11].
- From the analysis of different ports traffic, it has been discovered that unassigned ports are normally used by streaming sites, which may be blocked. e.g., Port 436 & 8080.
- Some of the clients are given direct access to the ISA Server or in other words they are allowed to bypass server security. These users have no restriction on the type and amount of content they are browsing, hence may choke the network bandwidth and would remain unnoticed. Therefore it is strongly recommended that no one should be allowed to bypass the Symantec security.

The solution given in this research if implemented can be very helpful in analyzing different ports & protocols and identifying bandwidth hungry hosts & clients responsible for generating lot of network.

8. SUMMARY & CONCLUSIONS

ISA Server generates very detailed logs for all the web requests made through it, these logs are in the form of semi structured flat text files. Although these logs contain a lot of data which can be mined or analyzed to gain insight into the network usage trends and patterns but this precious data goes useless. Moreover ISA Server provides built-in logs analysis and reporting features, but with minimal customization options for users.

The solution suggested in this paper is capable of parsing and transforming these semi-structured logs and is capable of generating more customizable multidimensional reports that helps in better understanding and analyzing the network usage patterns & trends.

The paper analyses the ports and protocols of the ISA server being used by the users for accessing the network. As outlined in the recommendation section, some of the bandwidth hungry hosts consuming approximately 55% of the total bandwidth have been identified. Additionally it has been found that the unassigned ports are never used for any useful browsing, hence they may be blocked. Beside this, every user should not be blindly allowed to bypass the server security.

REFERENCES

[1] Arun Sen, Peter A. Dacin and Christos Pattichis, "Current trends in Web Data Analysis", Communications of the ACM, Vol.49, Issue 11, pp. 85 – 91, Nov 2006.

[2] Jia Hu; Ning Zhong; "Clickstram Log Acquisition with Web Farming", The 2005 IEEE/WIC/ACM International Conference on Web Intelligence; pp. 257 – 263; 19-22 Sept. 2005.

[3] I-Hsien Ting, Chris Kimble & Daniel Kudenko, "UBB Mining: Finding Unexpected Browsing

Behaviour in Clickstream Data to Improve a Web Site's Design", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2005.

[4] Dong-Ho Kim, Il Im & Vijayalakshmi Atluri, "A Clickstream-based Collaborative Filtering Recommendation Model for E-Commerce", Proceedings of the Seventh IEEE International Conference on E-Commerce Technology, 2005.

[5] Olfa Nasraoui, Osmar R. Zaïane, Myra Spiliopoulou, Bamshad Mobasher, Brij Masand & Philip S. Y, "Web Mining and Web Usage Analysis" 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2005.

[6] Montgomery, A. L. "Using Clickstream Data to Predict WWW Usage", WebShop, University of Maryland, June 2003.

[7] Alan L. Montgomery and Kannan Srinivasan, "Learning About Customers Without Asking", Nirmal Pal and Arvind Rangawamy (eds.), The Power of One - Leverage Value from Personalizaton Technologies, eBRC Press, Penn State University, 2003.

[8] SANS Institute, "Using ISA Server Logs to Interpret Network Traffic", SANS Institute 2002.

[9] Xiaohua Hu, Nick Cercone; "An OLAM framework for Web Usage Mining and Buiness Intelligence Reporting", IEEE, 2002.

[10] Jesper Anderson, Anders Giversen, Allan H. Jensen, Rune S. Larsen, Torben Bach Pedersen; "Analyzing Clickstreams using Subsessions", Proceedings of the 3rd ACM international workshop on Data warehousing and OLAP; pp. 25 – 32; 2000.

[11] http://www.iana.org/assignments/port-numbers.

[12] Behzad Mortazavi, "Discovering and Mining User WebPage Traversal Patterns," The Requirements for the Degree of Maser of Science in the School of Computing Science, Simon Fraser University, 2001.

[13] Amiya Kumar Tripathy, Siddharth Nimkar, Meenakshi Srivastava "SCAzer - Analyzing Web Content through Users' Access Patterns" International Journal of Recent Trends in Engineering,Vol. 1, Issue 1, 2009.

[14] Dilip Singh Sisodia, Shrish Verma, "Web usage Pattern Analysis Through Web Logs: A review", Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE), 2012.

[15] Kazimierz Kowalski, Mohsen Beheshti, "Analysis of Log Files Intersections for Security Enhancement", Third International Conference on Information Technology: New Generations (ITNG'06), 2006.

[16] DeMin Dong, "Exploration on Web Usage Mining and Its Applications", 978-1-4244-3894-5, IEEE, 2009.