Network-Wide Protection Coverage for Data Loss Prevention

George-Sorin DUMITRU[1], Adrian Florin BADEA[1], Victor CROITORU[2], Daniel GHEORGHICĂ[1]

Rezumat. *Articolul descrie o prezentare generală a tehnologiei de prevenire a scurgerilor de date, o ramură a securităţii informatice, ce a crescut considerabil de la apariția sa, devenind indispensabilă pentru companii datorită numărului tot mai mare de modalități în care angajații pot distribui informații confidențiale în exteriorul organizației. Soluția de prevenire a scurgerilor de date pentru datele (DLP) în tranzit (bazată pe captarea traficului din rețea de la Check Point) a fost configurată cu reguli personalizate şi testată într-un mediu virtual.*

Cuvinte cheie: *prevenirea scurgerilor de date, securitate, sniffing, trafic de rețea, virtualizare, DLP*

Abstract. *The paper presents an overview of the Data Loss Prevention (DLP) technology, a branch of the security industry that has grown considerably since its emergence, meanwhile becoming indispensable for companies, due to the multitude of tools that employees can use to distribute data outside the internal network. The DLP software blade from Check Point for Data in Motion (sniffing of network traffic) was configured with custom rules and tested in a virtual environment.*

Keywords: *data loss prevention, security, sniffing, network traffic, virtualization, DLP*

1. INTRODUCTION

The way networks interact today is controlled by the suite of protocols known as TCP/IP (Transmission Control Protocol/Internet Protocol). The name comes from the two key protocols of the standard which has been in a continuous development and utilization in the last four decades. Meanwhile, TCP/IP has evolved from an experimental technology, used to connect only a handful of research computers, to the core of the most complex network in history, namely the global Internet, linking billions of devices worldwide.

In the world of cybersecurity the term DLP (Data Loss Prevention) first appeared on the market in 2006 and started to become popular in early 2007. Just like other security products, such as firewalls and intrusion detection systems, DLP solutions had a considerable growth in recent years, time consuming deployments, difficult management and high costs quickly becoming history [1].

DLP is the best solution for preventing accidental data leakage by applying a corporate policy which works automatically and which will capture sensitive data before it leaves the organization. It identifies, monitors and protects data transfers by thorough content inspection and analysis of transaction parameters (such as source, destination, data type

[1] Kapsch S.R.L, Bucureşti, mail: adrian.badea@kapsch.net, george.dumitru@kapsch.net
[2] Universitatea Politehnica Bucureşti, Facultatea de Electronică, Telecomunicaţii şi Tehnologia Informaţiei, mail: croitoru@adcomm.pub.ro

and protocol used). A short definition of DLP is that it detects and prevents unauthorized transmissions of confidential information.

The technology is also known under various aliases such as Data Leak Prevention, Information Leak Detection and Prevention, Information Leak Prevention, Content Monitoring and Filtering, and Extrusion Prevention. In simple terms, it is a technology that allows content-level visibility in a network by extracting application level informations from analyzed traffic. It relies on network traffic collection, decoding and processing systems.

2. NETWORK TRAFFIC PROCESSING

2.1. Collection and monitoring

Network analysis is the process of listening and controlling network traffic. It offers a detailed inspection of communications that occur in the network with the purpose of identifying issues that lead to decreased performances, locating security breaches, evaluating applications behavior and planning capacity. Network analysis (also called protocol analysis) is a process used by IT specialists responsible for network security and performances.

The tools used in network analysis are often called sniffers and can be purchased or distributed as hardware plus software or only as software solutions. A very popular open source network analyzer is Wireshark [2]. Wireshark is the most used system for collecting and analyzing network traffic. Available free of charge, Wireshark can be used on many platforms. It has become a standard for analyzing computer network traffic. Wireshark was used increasingly more for collecting, troubleshooting, and helping network administrators to understand the problems they may face.

Visualizing traffic always indicates the origin of the problem, therefore network analysis is indispensable to the functioning of a network. It allows us to look inside the communication system and observe how packets are transferred from side to side. We can see how DNS (Domain Name System) queries are being sent and the answers to these requests. We are able to watch the local system sending a TCP connection request and we can measure how long it takes for the destination to reply consequently it is possible to have an overview of the required round trip time to that site.

During a typical network analysis session the following tasks are performed:

- capturing packets at appropriate locations in the network;
- applying filters for visualizing only the necessary traffic;
- identifying and examining anomalies present in network traffic.

When the network is affected by performance issues, assumptions about the possible causes can become very tedious and time-consuming and can lead to incorrect conclusions that are causing unnecessary costs and resources. A complete understanding of network traffic is required for the correct placement of the analyzer and detection of the issues.

The main reason people avoid network traffic analysis is represented by the total confusion caused by the large number of packets that are constantly crossing the network. In a large network, where many users complain about poor performances, placing the analyzer in the right place is just as important as applying filters to focus on the relevant traffic and correct interpretation of the events.

There are several ways to capture traffic in an IP network:

- hub technology;
- connection to half or full duplex traffic;
- SPAN port configured on switch;
- analyzer installed on the target system.

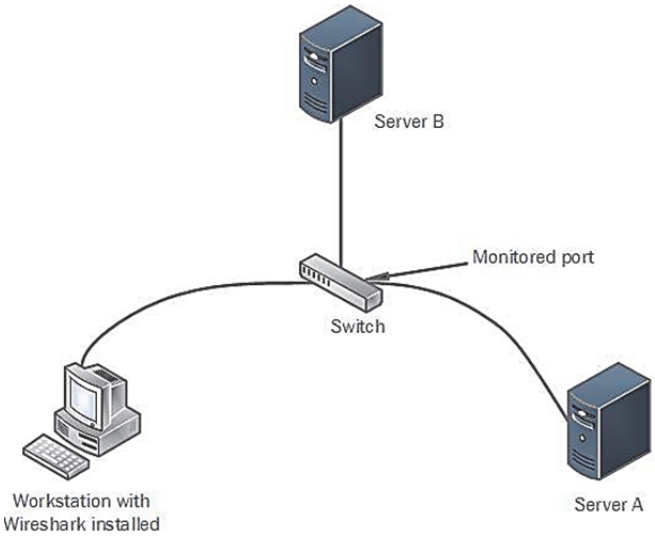

Fig. 1. Capturing network traffic from a monitored port using Wireshark.

An example of capturing network traffic by using Wireshark is shown in figure 1. A monitored port (SPAN port) is configured on the Switch. The traffic is mirrored and also sent to the Workstation that has the Wireshark installed. All the communication between Server A and Server B is duplicated and it can be analyzed by Wireshark.

Another very useful and complex network analyzer is NetFlow. NetFlow is a Cisco proprietary network protocol used for data collection and monitoring network traffic generated by routers and switches which support this technology.

NetFlow can analyze a large volume of traffic to determine from where the traffic came into the internal network, where it goes and the overall generated traffic. NetFlow allowed visibility at the packet level and even visibility at the byte level to understand which IP addresses are the sources for traffic and which applications generated all the traffic. With NetFlow is possible to obtain the following information about the network traffic: network interface, source IP address, destination IP address, IP protocol, source port, destination port, TCP flags, the total number of packets in the flow, the total number of bytes in the flow, the number of packets per second, the number of bits per second, the average number of bits per packet, the duration.

Filters that can collect data are based on tcpdump syntax. This syntax appears in the libpcap/WinPcap library.

Tcpdump is used for capturing data packets transferred over the network in the following cases: to design networks/protocols, to check if some network services are properly running, to troubleshoot, to monitor and to make statistics based on traffic. The captured packets can be displayed selectively in text mode, on the console terminal or saved to a file. In addition, the output of tcpdump application can be viewed using Wireshark application.

Pcap (Packet capture library) provides a high level interface to packet capture systems. All packets in the network, even those destined for other hosts, are available through this mechanism. It also allows saving the captured packets for subsequent analysis [3].

Libcap library is the standard interface at the data link layer for packet capture, used by hosts running UNIX operating systems and WinPcap is its Windows OS ported version.

WinPcap allows applications to capture and send network packets, avoiding the protocol stack and has useful features such as kernel-level packet filtering, a network statistics engine and support for remote packet capture. Due to its characteristics, WinPcap is used by many open source tools including: protocol analyzers, intrusion detection systems, sniffers, traffic generators [4].

Capture filters are used to decrease the number of packets stored in the temporary location during the capture or in a different folder when saving the trace file. They can be set only on ongoing capture processes, not on existing trace files. Capture filters are very useful in limiting the number of captured packets when the network in question is highly loaded

or when following a particular type of traffic. Packets that pass the capture filter criteria are sent to the capture engine of the analyzer, as shown in figure 2.

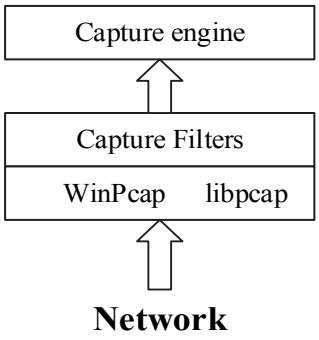

Fig. 2. Capture filters are applied only to incoming network packets.

Wireshark offers the option of tracking a specific network stream. This process lies in the reassembly of communication (except the MAC, network and transport layer headers). Using this technique the commands and data transferred in a conversation can be obtained. The analyzer divides the data link header, IP header and TCP/UDP (User Datagram Protocol) header and uses a color code to differentiate the client generated traffic from the server generated traffic (color red is used to describe traffic from the client – the host initiating conversation and blue for traffic from the server).

UDP, TCP and SSL (Secure Sockets Layer) streams can be reassembled. SSL data flows display the reassembled informations only after they are decrypted. In some communications, such as FTP (File Transfer Protocol), the original file can be reconstructed and transferred by saving the reassembled data. The file identifier can be used to determine what type of file was transferred.

2.2 Decoding and data processing

In the case of HTTP (Hypertext Transfer Protocol) traffic, usual transmissions are using a request/response communication model. Customers make requests to web servers which respond with a status code.

HTTP communication problems may occur due to errors in the process of resolving the IP address from the site name, unsuccessful TCP handshake, requests for non-existent pages, packet loss and also due to server congestion.

Anyone has entered, at least once, the wrong address for the desired website. If the name cannot be resolved to the IP address, the location will be unavailable. DNS traffic is critical when analyzing web browsing issues.

Another problem can occur when the HTTP daemon is not running properly on the web server. When this happens the server replies with a TCP RST/ACK to the customer's SYN. As a result, the connection cannot be established.

If the HTTP client successfully connects to the server, but then calls for a non-existent page, the server will return errors of type 404 Not Found. Some forwarding services will replace this standard message with suggestive links or they can redirect the client to a completely different site. It's important to note that when troubleshooting web traffic, TCP errors should be checked first and then move to HTTP.

Web browsing analysis also includes HTTPS (HTTP Secure) communications. At the start of a secure HTTP session, TCP connection is established first followed by the secure session initiation.

RFC 2818 defines the use of HTTP over TLS (Transport Layer Security) for protected communications. RFC 2246 details TLS, version 1.0, which is based on SSL version 3.0. Although there are minimal differences between them, the two are not interoperable.

SMTP (Simple Mail Transfer Protocol) is the standard application for sending email and, by default, these communications are not secure. Port number

25 is used for SMTP, but it can be configured, as many other protocols, to run on another port. An increasingly large number of Internet providers and firewall configuration block SMTP connection on port 25. This happened in the attempt of stopping spam messages from crossing the supplier's networks.

SMTP communication problems can arise from the TCP handshake, being affected by high latency and packet loss. If the SMTP server responds with code numbers higher than 399, this indicates an issue in the process of sending the message.

In a typical FTP (File Transfer Protocol) session, a control channel is established on port 21 of the server. To transfer data (such as files and folders) another channel is enabled, dynamically allocating the port number, even if the documentation specifies port 20 as used for the data channel.

FTP communication problems start with the TCP handshake. If a server does not have the FTP daemon enabled, it will respond to TCP SYN packets received on port 21 with a TCP RST flag. If the FTP server is configured to use another port than the one set on the client, the FTP connection cannot be established. Moreover, if a firewall is blocking passive transfer mode support, connections attempt will fail.

3. DATA LOSS PREVENTION

In the modern age, data transfers are easier to achieve than ever before, and a large part of the information presents various levels of sensitivity. Certain data is confidential, by default, through their affiliation to certain organizations and are not publicly available. Some data is sensitive due to legal requirements, national laws and international regulations. Very often the value of the information depends on keeping it confidential (intellectual property and the competition must be taken into account) [5].

Data leakages in a company can be embarrassing or, even worse, may lead to losing the competitive advantage and loss of accounts. When an organization acts in non-compliance with the confidentiality standards and other laws, its integrity may be in danger.

Given the increase in data loss incidents, companies have no choice and must take measures to protect sensitive data. Confidential informations regarding the employer and the client, legal documents, intellectual property, these are all exposed. The challenge consists in efficiently addressing the problem at hand, without affecting employee productivity or increasing the number of people in the IT department. Technology is permanently evolving, but it is ineffective in understanding user intentions [1].

With all the information sharing tools available on the Internet, an irreversible mistake can easily be made, such as unintentionally violating the company policy (for example, an employee who takes the laptop home even if he's not allowed to). Moreover, accidental cases of data leakage can show up, which are enhanced by the new solutions possibly to be used for data transmissions: cloud-based servers, Google Docs, Dropbox, etc.

The companies allocate a lot of time and money to educate the employees on the specific corporate policy that is in place, with the desired result of minimizing data leaks caused by careless users. This is not always feasible and the majority of confidential data loss incidents is being caused by users, despite their prior training. Traffic monitoring presents an overview of how well users are compliant with the corporate policy, but it is not enough to prevent data leaks, both unintentional and malicious, therefore imposing the need for DLP-based preventive control.

DLP is different from other security solutions (firewalls, IDS/IPS, etc.) in that, unlike those, (which are searching and identifying any threat to the

organization) it is focused on identifying sensitive data, with critical content for the company. Figure 3 is a chart of the previously described concepts on which DLP technology is based [5].

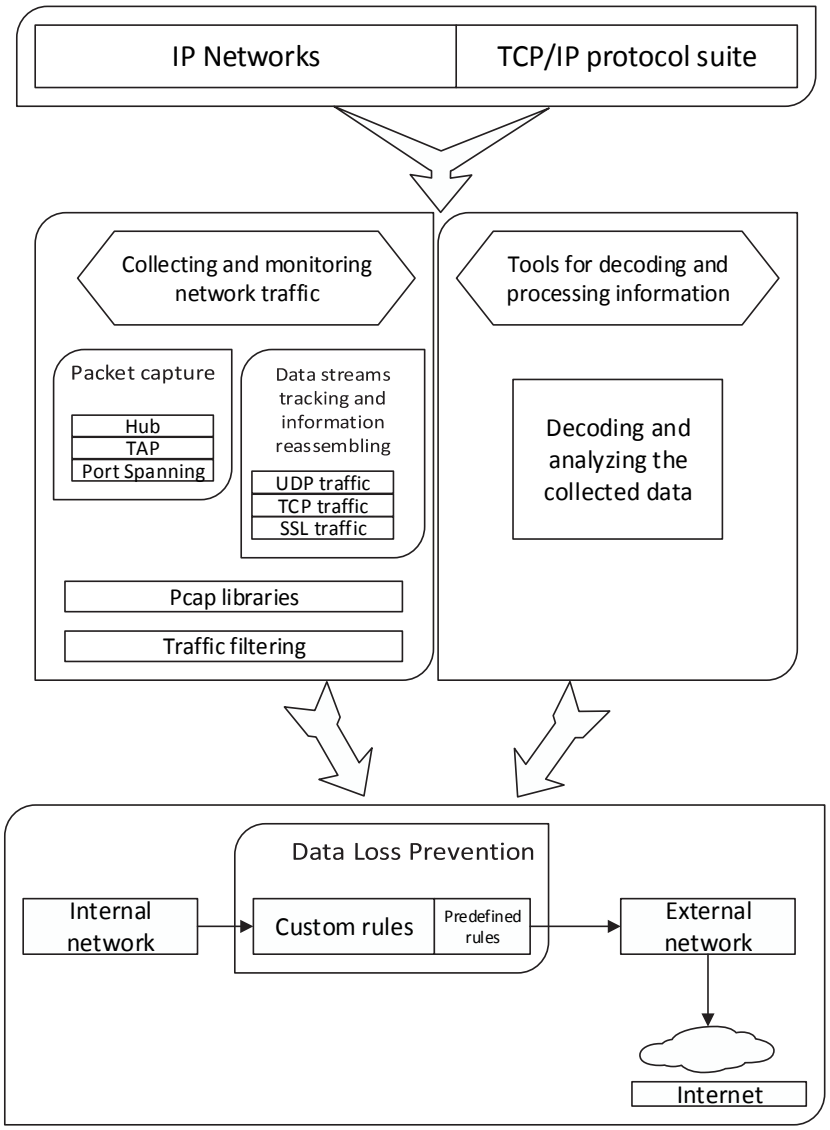

Fig. 3. DLP technology core concepts.

4. NETWORK-WIDE PROTECTION COVERAGE

4.1 DLP deployment with centralized management

Figure 4 presents the global architecture of the virtual environment in which Check Point Security Gateway R77 was integrated (the same configurations are necessary regardless of the environment of the deployment, appliances or open servers). This security solution contains a data loss prevention software blade which combines technologies and processes to remodel this technique, helping businesses to prevent accidental loss of sensitive data and offers incident remediation in real time [6]. In this scenario the management server and the security gateway were installed on the same virtual machine.

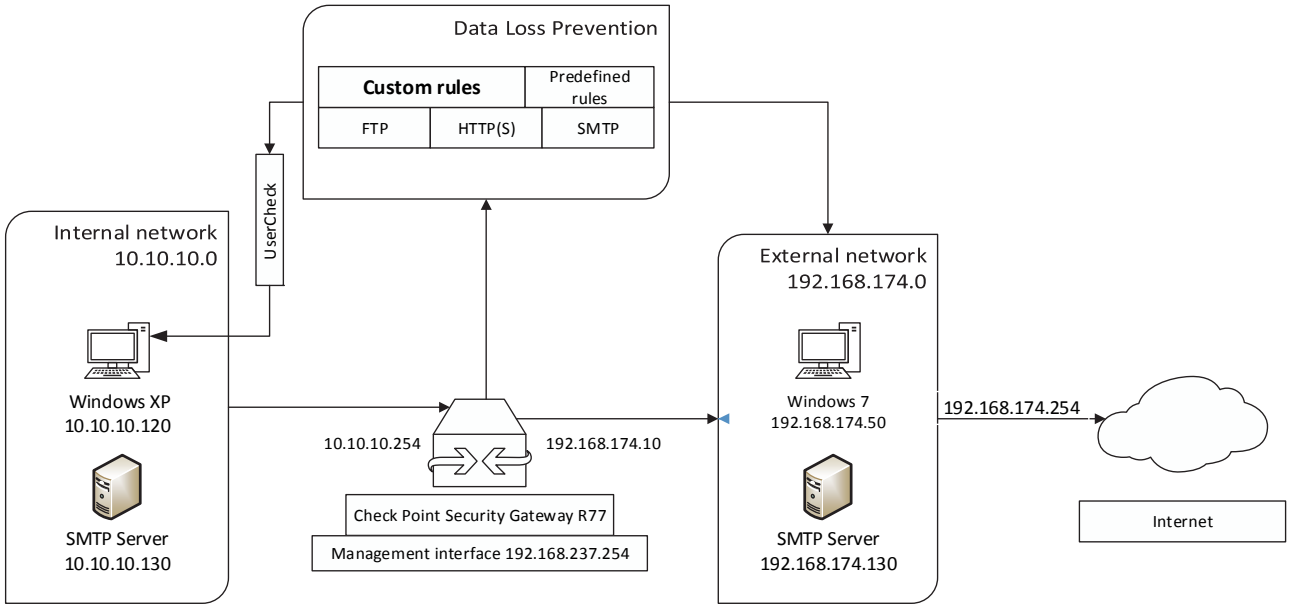

Fig. 4. Global architecture of the data loss prevention system based on network traffic.

Five virtual machines were created in VMware Workstation 10, with the necessary network adapters configured, the Security Gateway requiring three of them: management, internal and external networks.

DLP rules were tested using two VMs having the role of source and destination, the internal one using Windows XP OS and the external one Windows 7. For SMTP (Simple Mail Transfer Protocol) traffic, mail servers were configured in the two networks.

Consoles accessed using the virtual machine credentials and the IP address of the management interface were used to administer the security policies, to monitor network traffic events, etc. Custom DLP rules were created, implemented and tested on the main protocols responsible for data transfer: FTP (File Transfer Protocol), HTTP(S) (Hypertext Transfer Protocol (Secure)) and SMTP using different actions on the incidents and different data types.

SMTP rules were tested on two mail servers installed and configured on Ubuntu 14.04 OS. iRedMail open source solution was used with Postfix MTA (Mail Transfer Agent), and BIND (Berkeley Internet Name Domain). Both machines were configured with static IP addresses and as domain names, test.ro for the internal server and extern.ro for the external one. FTP rules were tested using FileZilla Server and FileZilla Client.

4.2 Pre-emptive DLP with user interaction

Figure 5 is the logic diagram of the DLP module, based on what type of action is being used for a certain defined rule when network traffic matches it, and figure 6 shows the logic diagram for configured SMTP rule that uses fingerprinting. This technique uses a repository, which is a network share that contains files that must not leave the internal network. The DLP software blade scans these files and generates a unique signature for each of them. When a file passes through the gateway, it is scanned and a signature is created which is compared with the signatures of the files in the repository. In case there is a match, the scanned file is allowed/blocked based on the specific chosen action type.

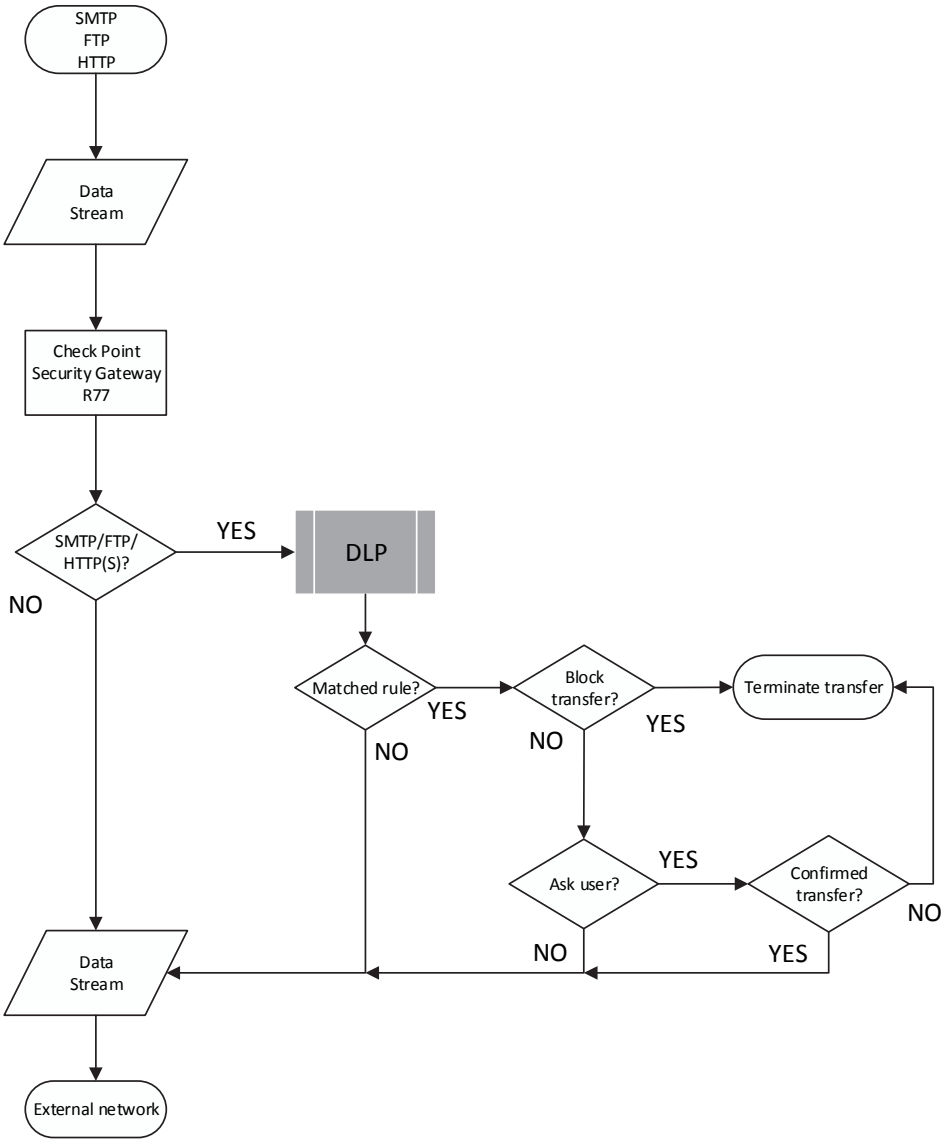

Fig. 5. Logic diagram of the data loss prevention system based on network traffic.

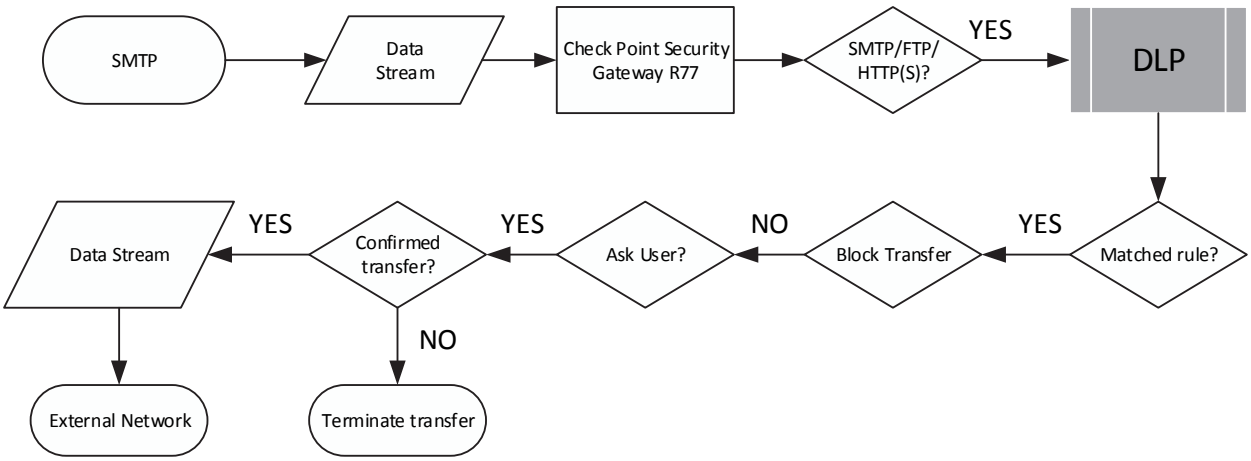

Fig. 6. Logic diagram of the rule with fingerprinting tested on SMTP traffic.

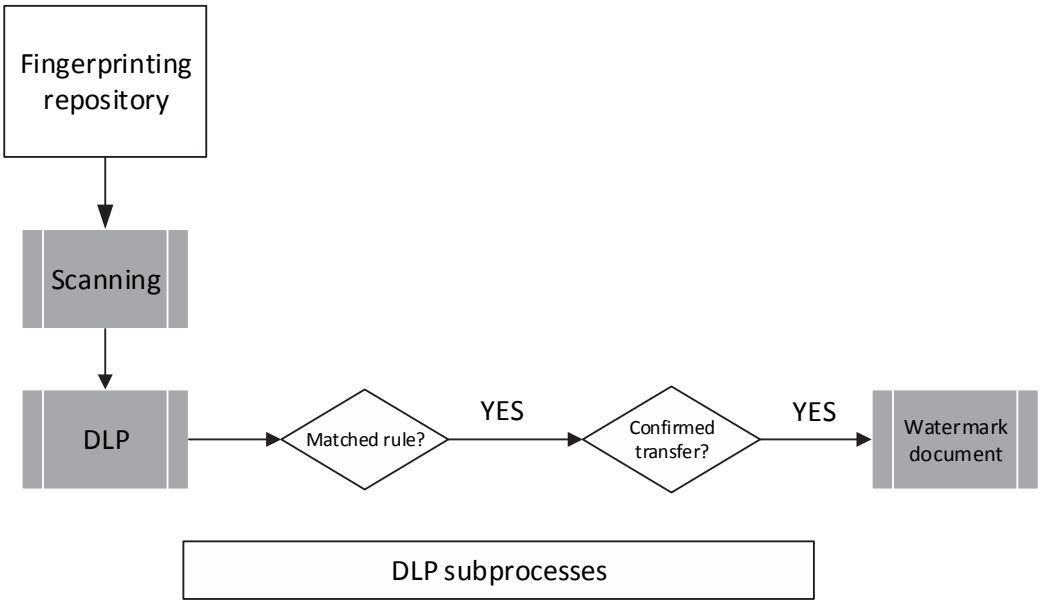

Fig. 6 (continuare)

Figure 7 shows the logic diagram of the rule tested on HTTP(S) traffic, and figure 8 presents the logic diagram of the FTP rule.

For HTTP and HTTPS traffic the custom rule will check the document's properties and will decide, based on these, if the upload to www.transfer.ro and Yahoo mail will be blocked or not. It was decided that .pdf files with a size of over 1000 KB to be blocked. Also, on the internal endpoint, a UserCheck agent was installed which has the purpose of communicating with the gateway and display notifications to users. A choice regarding the transfer can be made in real time. DLP incident notifications can be sent by email (only for SMTP incidents) or can popup through the agent in the system tray (for all protocols) [7].

The rule for FTP traffic specifies that when a document contains at least one of the configured keywords, the user will be notified that he sends confidential information.

5. CONCLUSIONS

DLP technology provides one of the solutions in ensuring the corporate cyber security, proving to be a very efficient tool for data leaks prevention. These products provide a high level of security to companies that plan appropriate deployment and understand how to take full advantage of the solution.

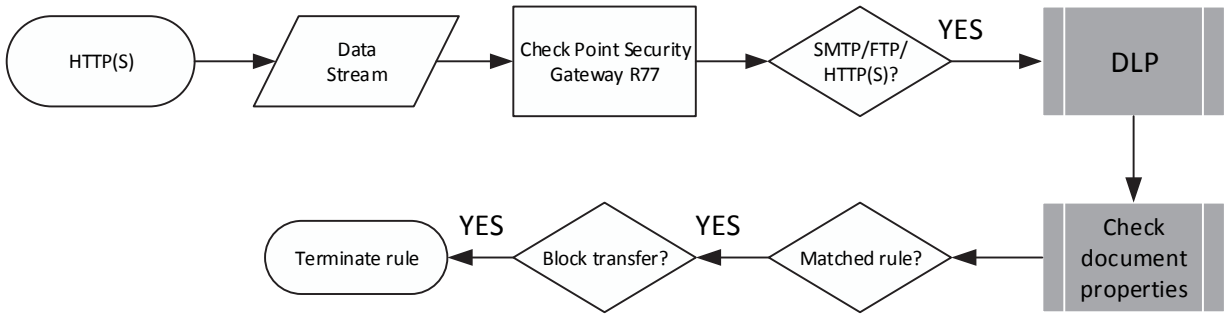

Fig. 7. Logic diagram of the rule applied on HTTP(S) traffic.

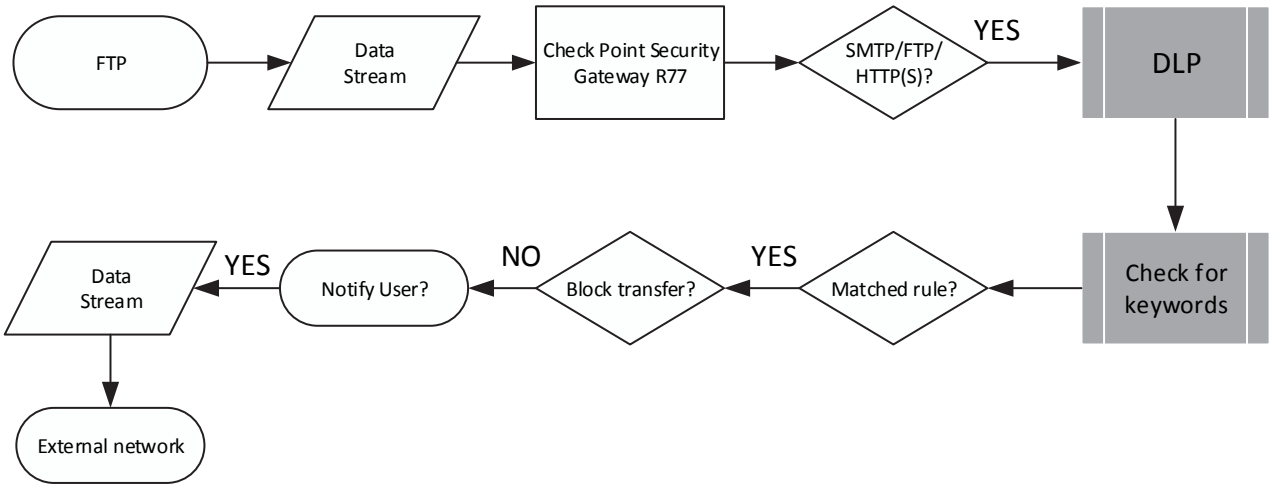

Fig. 8. Logic diagram of the rule applied on FTP traffic.

The conductive tests enable to remark, the ease of integration of a modern data loss prevention system in an existing network and the simplicity of its administration. There is no need for additional employees or expenses, just a system administrator with every user being capable of resolving its own incidents in real time. All these features show the high degree of automation of such a solution, without network delays or blocked false positives which can lead to more serious problems than confidential informations leaving the company.

Initial topology and IP addressing configuration are crucial to the proper functioning of the DLP system. The internal and external networks must be clearly specified and it is recommended to define important nodes in the network with suggestive names, such as username and IP address. In this way, incident management is easier to be achieved, immediately identifying the source of the issue by checking the report.

The level of severity associated with each configured rule plays an important part in the DLP policy. The rules that have a high or critical risk level should have at start an action type of user notification that will then be changed to ask the user and finally to the block mode, but making sure first that the users understood what is expected of them. When several rules are matched, the one with the highest severity level will be applied. It's important to know that after any change in the configuration, the policy must be saved and installed again in order for the changes to have effect.

Fingerprinting is an excellent method of securing important files by simply scanning a directory and afterwards using it to create rules. A whitelist repository can also be used to eliminate false positives and increase the accuracy of the DLP policy. If a document is both in the fingerprint repository and the whitelist repository, the latter takes precedence, therefore it can be sent outside the company.

To monitor Microsoft Office documents that are sent outside the company, watermarks can be added that will clearly establish the fact that the file contains confidential information.

To activate HTTPS inspection, it is necessary to create a CA certificate (certificate authority) which must be exported and distributed to the endpoints. Without the distribution, SSL error messages will be received in the browser, when secure locations on the Internet are accessed. The certificate uses a password to encrypt the private key of the file. The security gateway uses this password to sign certificates for accessed websites.

Based on the implementation and tests that were performed, the crucial importance a DLP system plays in a network has been understood. It is effective due to its simplicity and speed with which it fixes incidents and through its intuitive management interfaces. Identifying data types that were defined as confidential was successfully tested on all the protocols in the scenario, DLP promptly responding to any incident that occurred.

REFERENCES

[1] Kanagasingham P., *Data Loss Prevention*, SANS Institute, 2008.

[2] Chappell L., *Wireshark Network Analysis – The Official Wireshark Certified Network Analyst Study Guide 2nd Edition*, Chappell University, 2012.

[3] URL: http://www.tcpdump.org/

[4] URL: http://www.winpcap.org/

[5] Dumitru G., *Sistem de prevenire a scurgerilor de date, bazat pe informaţiile colectate din reţea* (Diploma Project, scientific advisers: prof. dr. ing. Croitoru V., dr. ing. Gheorghică D.), UPB, 2014.

[6] Check Point Documentation, *R77 Data Loss Prevention Admin Guide*, 2014.

[7] Check Point Documentation, *Gaia Installation and Upgrade Guide, 2014.*