

# BIG DATA CYBERSECURITY

## LAB 10 APACHE SPOT

### APACHE SPOT OVERVIEW

**Lab Description:** This assignment introduces Apache Spot—a cybersecurity analytics tool running on a big data cluster powered by Apache Hadoop. In modern enterprise environments, networks experience billions of flow, DNS and proxy data artifacts, which require a scalable solution for storage and subsequent analysis [1].

Apache Sport supports three types of network data: NetFlow data, Domain Name System (DNS) server logs and proxy logs [2]. NetFlow is instrumentation embedded within Cisco software, which is widely-used by network administrators for incident response and network security forensics, monitoring and analysis of network usage and productivity, security vulnerabilities, and other issues [1, 3].

Since flows contain information on all network activity, it can be thought of as an eye on what happens within the network boundaries. As well, machine learning algorithm may be applied to flow data to detect the latest attacks for which signatures are not yet available [4]. Other NetFlow data applications include detecting anomalous activity in the network, identification of compromised hosts, nefarious activity by insiders, etc. [5]

DNS logs record connections made by a client to a domain. This information allows to reveal a number of threats and malicious activities, e.g. botnets and malware connecting to Command & Control (C&C) servers, Distributed Denial of Service (DDoS), NXDOMAIN, phantom domain, random subdomain, Domain Generation Algorithm (DGA) attacks, etc. [6]

C&C Activity. DNS protocol is commonly allowed to cross enterprise network perimeters. Thus, it allows malicious actors to set up communication channels between infected devices and C&C servers [7].



DDoS via DNS flood attack. DNS server is overloaded by a large number of requests sent by hundreds or thousands of hosts aiming to make the system stop serving DNS requests.

Distributed Reflection Denial of Service (DRDoS) is a variation of DDoS, when requests are sent from servers spoofing the source address so that machines affected by the requests reply back to the targeted victim and overwhelm it [8].

In the NXDOMAIN attack, a threat actor sends a flood of queries to a DNS server to resolve a domain name, which does not exist. While trying to resolve the non-existent domain, the DNS server fills up its cache with NXDOMAIN results, thus slowing response to legitimate requests [9].

In the phantom domain attack, a DNS server is hit with requests to resolver so called phantom domains that do not exist. Consequently, the DNS server runs out of resources since the queries are never answered [8].

When random subdomain is used as an attack vector, a threat actor sends a large number DNS queries against a valid domain name prepended with non-existing subdomains. The goal is to overload the DNS server, which hosts the targeted domain, in order to disrupt its operation [8].

Domain Generation Algorithms (DGAs) are the type of malware which generate up to tens of thousands of new commonly meaningless domain names, such as `nwongtaqfnau.com` used by Cryptolocker. These new domains are used for communication between the malware-infected systems and C&C servers [8, 10, 11, 12].

Other common DNS attacks include domain hijacking involving changes in DNS server and domain registrars in order to redirect traffic from genuine to malicious hosts; cache poisoning or DNS spoofing by injecting malicious data into DNS server cache in order to redirect visitors to different servers; DNS tunneling includes encoded data from other applications into DNS queries and responses; and DNS hijacking alters TCP/IP configuration to point to malicious DNS serves [8].

Proxy log analysis commonly assumes Squid log files due to the wide use of Squid proxy server. All web traffic to and from company devices would go through an instance of the Squid proxy, which allows a full view of network connections. The following topics are commonly focused upon



when investigating proxy logs: internal users or appliances scanning or attacking outside systems generally resulting in numerous 404 or 403 HTTP error codes; internal devices infected with malware, e.g. botnet zombies, worms, trojan horses or other malware connecting to C&C centers or other infrastructure are commonly detected by matching against block lists such as Spamhaus [14], internal users trying to send external emails or to access web resources prohibited by company policies, e.g. gaming or entertainment websites, etc. [13]

In this assignment, we will use Docker Desktop to perform an overview of Apache Spot running in a container. Among other features, Docker Desktop allows to run containerized applications in lightweight virtual machines. Docker Hub is a platform for sharing container images.

The Apache Spot version enclosed in the Docker container has four top menu options: Flows, DNS, Proxy—each containing three submenu items, Suspicious, Threat Investigation and Storyboard—and Ingest Summary. The key features of Apache Spot are: DNS traffic inspection for profiling probable and improbable DNS payloads, visualization, normalization and pattern searches of DNS data; analysis of connections per IP address; machine learning applied for finding most likely threat patterns in data; a Storyboard presenting plain English explanation of detected network activity of interest supplemented with interactive visualizations; Open Data Models—a format for exchange of enriched cybersecurity data; and Hadoop-supported scalability for big datasets [15].

Apache Spot consists of several components: the Ingestion component for data collection from various sources with the help of Hadoop Distributed File System (HDFS), Kafka, and Spark Streaming; the Machine Learning component applying the Latent Dirichlet Allocation (LDA) for finding suspicious connections; and the Visualization component—a user interface with panels for studying suspicious network activity, network connections visualization, the Notebook panel for manual assignment of threat scores; the Details panel with additional information; the Threat Investigation allowing an analyst to enter a custom review for a threat; the Storyboard aiming to present an executive summary of incident progression, impact analysis, geolocation and incident timeline; and the Ingest Summary visualizing the amount of ingested data [16-20]. The Open Data Models component is intended for sharing and reuse of data models and analytics across the community. This feature is marked “under development.”



Apache Spot has not graduated the Apache Incubator, which it entered on September 23<sup>rd</sup>, 2016. This means that the Spot software is in its unstable development phase. Version 1.0 was released on August 7, 2017 and is available for download [21, 22].

**Lab Files that are Needed:** n/a

---

## LAB EXERCISE/STEP 1

Download Docker Desktop for your operating system from [24]. It has different system requirements for Windows and Mac OS. Windows requirements and installation steps are described at [25]. Installation instructions for Mac are given at [26]. Windows users need to ensure that virtualization is enabled in BIOS. Typically, this is done by default. To troubleshoot virtualization, navigate to this URL [27] and locate the Virtualization section.

After Docker installation completes, verify that Docker commands are available from a system terminal. Start a terminal, type the command below and press Enter:

```
docker -v
```

The output should contain Docker version and build values.

---

## LAB EXERCISE/STEP 2

Run the Docker container with Apache Spot application installed by executing the command below in a terminal window on your computer:

```
docker run -it -p 8889:8889 apachespot/spot-demo
```

When running the command for the first time, the container will be downloaded from the Docker Hub. Download progress will be displayed in the terminal window. Make sure the download comes to 100%. In case of unsuccessful download, incorrect local Docker image and container for Apache Spot must be cleaned.

A Docker Hub page describing this container with Apache Spot demo is located at [28]. It contains additional data about the project, indicator when the container was updated last time, the number of times it was downloaded, and other information.

---


## LAB EXERCISE/STEP 3



Start a web browser and navigate it to

<http://localhost:8889/files/ui/flow/suspicious.html#date=2016-07-08>

This opens the Flows tab displaying data for July 8<sup>th</sup>, 2016. Apache Spot :: Netflow :: Suspicious should be displayed in the top left corner. The Flows view consists of four panels: Suspicious, Network View, Notebook and

Details. Maximize the Suspicious panel by clicking  icon in its top right corner. This view displays top 250 suspicious connections ranked from zero to two hundred forty nine by the Spot machine learning block. Notice that the table contains the following columns:

Column Title	Column Description
Rank	ML output rank
Time	Time received field for Netflow record
Source IP	Netflow Record Source IP Address
Destination IP	Netflow Record Destination IP Address
Source Port	Netflow Record TCP/UDP Source Port Number
Destination Port	Netflow Record TCP/UDP Destination Port Number
Protocol	Text format for Protocol contained within Netflow Record (Ex. TCP/UDP)
Input Packets	Reported Input Packets for the Netflow Record
Input Bytes	Reported Input Bytes for the Netflow Record
Output Packets	Reported Output Packets for the Netflow Record
Output Bytes	Reported Output Bytes for the Netflow Record

Internal IP addresses are highlighted in blue. External source and destination IP addresses have color-coded shield and a globe icons. Placing a mouse pointer above a shield icon would display a GTI score—a



cybersecurity reputation score—obtained from McAfee Global Threat Intelligence (GTI), which can take the following values: minimal, medium, high or unverified [15, 23]. Hovering a mouse pointer above a globe icon would display brief geo location and domain data. The search box above the table allows to filter data by IP address.

A search box with the label "IP:" and the text "0.0.0.0" inside. To the right of the text is a magnifying glass icon inside a dark blue square button.


Using the filter box, make the table display connections to 10.199.250.2 IP address.

*Make a screenshot and store it in the submission document.*

Answer the questions: Is this IP address internal or external? What are the GTI score, geo location and domain data for this IP?

---

#### **LAB EXERCISE/STEP 4**

Keep the Suspicious panel table filtered by 10.199.250.2 IP address. Click the  button to resize the panel to its normal size. Notice that the Network view visualizes only the connections to or from 10.199.250.2. In the Network view, internal IPs are displayed as blue diamonds while the external ones are shown as orange circles.

Highlight a row in the Suspicious view. This will make the nodes corresponding to the source and destination IPs blink in the Network view. All the Netflow records in between source and destination IP addresses from the highlighted row happened within the same minute will be displayed in the Details view.

*Make a screenshot and store it in the submission document.*

---

#### **LAB EXERCISE/STEP 5**

The Notebook panel allows to assign risk scores to IP addresses and ports. Score of one means high risk, two medium or potential risk, and three denotes low risk.

Clicking the Score button finds all entries matching the selected values and updates their scores to the rating selected in the radio button list. The Save button updates the rating of the scored connections, refreshes the



content of the panels in the page, removes the manually scored connections from the suspicious connects page and makes them available for the machine learning component of Apache Spot.

Task: assign the high risk score to the connections where 10.138.235.111.

Note: first, you need to click the Score button and then the Save button.

*Make a screenshot and store it in the submission document.*

## **LAB EXERCISE/STEP 6**

As mentioned above, Apache Spot is in its incubating phase, which means this version is not yet complete. Trying to open the Treat Investigation view produces an input-output error. Thus, we will skip this menu item in Flows, DNS and Proxy views.

Navigate your browser to the Netflow Storyboard by clicking the Flows menu item, then Storyboard. This view provides an executive summary of threats described using the incident titles at the Threat Investigation. In the Executive Threat Briefing panel, click 10.13.77.49 Port 80 Denial of Service. This will populate the rest of the boxes in this view.

Incident Progression shows a tree graph detailing the connections that conform the activity related to the threat. This graph will present more details when available.

Impact Analysis displays a horizontal bar graph representing the number of inbound, outbound and two-way connections related to the threat. Clicking any bar in the graph, will break down that information into its context.

Map View | Globe is rendered when geolocation data are available. In order to represent the global scale of communications, lines are displayed on the globe to depict data flows.

The Timeline panel displays clusters of inbound connections to the IP, grouped by time in order to visualize time during the day with the most activity. This panel may be zoomed in or out into using a mouse scroll.

Task: Maximize the Map View | Globe panel and rotate the globe so the lines connecting IP address locations are visible well.

*Make a screenshot and store it in the submission document.*

Maximize the Timeline panel and determine what port number was used mostly in this threat.



*Make a screenshot and store it in the submission document.*

## **LAB EXERCISE/STEP 7**

Open the Suspicious view for the DNS data. This view is organized similarly to the Netflow page. The Suspicious panel shows top two hundred fifty most suspicious DNS as classified by the machine learning component. Moving the mouse pointer over the Suspicious DNS table will highlight an entire row and highlight the corresponding nodes in the Network View panel.

Hovering a mouse pointer over the shield icon displays reputation scores from McAfee GTI, Facebook Threat Exchange, or both. Threat severity score is coded with the icon color.

Moving a mouse pointer over a DNS record in the Suspicious pane highlights a corresponding node in the Network View frame and the Details frame is populated with information on DNS record containing the same query.

The Network View represents the data in the Suspicious panel graphically. Diamonds represent DNS records while circles depict IP addresses participating in the communication. A primary mouse click on a circle IP Address node shows a diagram in the Details frame, providing all the Domain Name records queried by that particular IP Address. A secondary mouse click applies a filter to the suspicious DNS data.

The Details panel provides additional information about a selected connection. This panel has two modes: (1) Showing a table with detailed information on a record selected in the Suspicious frame and (2) a dendrogram view showing a tree of connections.

The Scoring pane allows to assign scores (1=High risk, 2 = Medium/Potential risk, 3 = Low/Accepted risk) to IP addresses and DNS records. Clicking the Score button finds exact matches for a selected threat (a client IP or a query) and assigns a risk value.

You can score a large set of similar or coincident queries by entering a keyword in the "Quick Scoring" text field and then select a severity value from the radiobutton list. The value entered here will only search for matches on the dns\_qry\_name name column. "Quick Scoring" text field has precedence over any selection made on the lists.

Analysts must use the Save button in order to save the scored records into the database, in the dns\_threat\_investigation table. After you clicking





it, the rest of the frames in the page will be refreshed and the connections that you already scored will disappear on the suspicious connects page. At the same time, the scored connections will be made available for the ML to use as feedback.

Task: in the Suspicious window, click the third line from the top where Query begins from "mairie-rueilmalmaison.a." The Details panel will display data with the same query. Also, hold the mouse pointer above that line in the Suspicious pane to filter the Network View.

*Make a screenshot and store it in the submission document.*

Using the Notebook panel, assign score 1 to the query "mairie-rueilmalmaison.accueil-famille.fr."

---

## LAB EXERCISE/STEP 8

Navigate to the DNS → Storyboard item. This view contains two panels. The Executive Threat Briefing frame lists all the incident titles entered at the Threat Investigation notebook.

Click the DNS Tunnel to volluto.com title. Additional comments will be displayed below the threat title. As well, the Incident Progression panel will display a dendrogram visualizing connections that conform the activity related to the threat.

*Make a screenshot and store it in the submission document.*

---

## LAB EXERCISE/STEP 9

Navigate to the Proxy → Suspicious menu item. Similarly to the Flows and DNS views, the Proxy Suspicious view consists of four panels. The Suspicious view displays top two hundred fifty suspicious connections as classified by the Spot machine learning component. When moving the mouse pointer above a record in the Suspicious table, a node is highlighted in the Network View. In the table, the shield and the list icons display the reputation score and web category of a host as provided by the McAfee threat intelligence service.

Clicking on a row will highlight it, as well as the corresponding node in the Network View, and Details panel will be populated with information on other suspicious connections to the same host. The root node is shown as a star, which depicts the proxy server. The path corresponding to the highlighted will be displayed in orange. Double clicking each node in the

highlighted path will display the host, a path on the host, and the client IP address with additional details.

Task: begin by double-clicking the medium-sized node with the red border to display its contents. This will expand the node and display several nodes, two of which will have red border denoting the high risk GTI score. Pick one of the nodes and completely expand the branch leading to that host. Move the mouse pointer to a leaf node corresponding to a client IP. This will display details on IP, URI, and the number of SC and CS bytes.

*Make a screenshot and store it in the submission document.*

Right click the leaf node to filter the Network View and the Suspicious table. Hover the mouse pointer above the shield and list icons to make sure it has a high GTI score. Click the record in the Suspicious panel to populate the Details view. Keep double-clicking the nodes in the graph branch displayed in the Network view until it displays all its nodes (there should be the root node, a method node, a host node, a path node and a node with client IP and additional information).

*Make a screenshot and store it in the submission document.*

---

## **LAB EXERCISE/STEP 10**

Use the Apache Spot menu to navigate to the Proxy → Storyboard item. The Executive Threat Briefing contains titles of incidents created in the Proxy Threat Investigation view, which is not functioning in this version of Spot.

Click the “firewall.happytohell.com website” title. This action will display a description of the threat, populate the Incident Progression and the Timeline panels. The Incident Progression will display a tree graph with nodes of different types: two nodes showing IP addresses connecting to the suspicious proxy record, a node showing the HTTP request method used to communicate between clients and the host, a node displaying the content type, and the node denoting the suspicious host. Move the mouse pointer to the Threat node. A pop-up message will show a URI accessed on the suspicious host.

*Make a screenshot and store it in the submission document.*



The Timeline window display IP connections to the Proxy Record (URL), grouped by time in order to give an overview of the time of day when suspicious activity happened. Numbers in parentheses next to IP addresses correspond to the number of connections made from an IP to the suspicious host.

---

## LAB EXERCISE/STEP 11

It is time to finish the assignment and stop the Docker container running the demo version of Apache Spot. To do so, switch to the terminal window used to start the container and press the Control C key combination. Docker will display a prompt:

Shutdown this notebook server (y/[n])?

Press y and press Enter. This action will terminate the container.

### What to submit

Submit a Word (or other text editor) document with embedded screenshots made as requested in the assignment and a brief description for each screenshot.

### References

- [1] I. Drago, "Flow monitoring explained: From packet capture to data analysis with NetFlow and IPFIX," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 2037-2064, 2014.
- [2] Apache Incubator, "Apache Spot Product Architecture Overview," [Online]. Available: <https://spot.incubator.apache.org/blog/apache-spot-product-architecture-overview/>. [Accessed Aug 8, 2019].
- [3] Cisco, "Introduction to Cisco IOS NetFlow - A Technical Overview," [Online]. Available: [https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod\\_white\\_paper0900aecd80406232.html](https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html). [Accessed Aug 8, 2019].
- [4] A. Powers, "Top 5 Uses of NetFlow for Network Security," *plexer.com*, [Online]. Available: <https://www.plixer.com/blog/top-5-uses-of-netflow-for-network-security/>. [Accessed Aug 8, 2019].
- [5] R. Stiennon, "Five Critical Uses Of Netflow Data For Security," *Forbes*, May 15, 2013. [Online]. Available: <https://www.forbes.com/sites/richardstiennon/2013/05/15/five-critical->



- [uses-of-netflow-data-for-security/#156ad94a1cdb](#). [Accessed Aug 8, 2019].
- [6] NetSurion EventTracker, "Monitoring DNS Traffic for Security Threats," [Online]. Available: <https://www.eventtracker.com/blog/2016/august/monitoring-dns-traffic-for-security-threats/>. [Accessed Aug 8, 2019].
- [7] S. Muppidi, C. Lee, A. Ron, M.P. Stoecklin and F. Satoh, "How to Use DNS Analytics to Find the Compromised Domain in a Billion DNS Queries," *securityintelligence.com*, [Online]. Available: <https://securityintelligence.com/how-to-use-dns-analytics-to-find-the-compromised-domain-in-a-billion-dns-queries/>. [Accessed Aug 8, 2019].
- [8] SecurityTrails, "The Most Popular Types of DNS Attacks," [Online]. Available: <https://securitytrails.com/blog/most-popular-types-dns-attacks>. [Accessed Aug 8, 2019].
- [9] Infoblox, "NXDOMAIN Attack Methods and Mitigation," Infoblox-SN-0208-00, [Online]. Available: <http://www.infoblox.com/wp-content/uploads/2016/04/infoblox-solution-note-nxdomain-attack-methods-and-mitigation.pdf>. [Accessed Aug 8, 2019].
- [10] D.S. Berman, "DGA CapsNet: 1D Application of Capsule Networks to DGA Detection," *Information*, vol. 10, no. 5, April 2019. [Online serial]. Available: <https://www.mdpi.com/2078-2489/10/5/157/htm>. [Accessed Aug 8, 2019].
- [11] TrendMicro, "Why Domain Generating Algorithms (DGAs)?" [Online]. Available: <https://blog.trendmicro.com/domain-generating-algorithms-dgas/>. [Accessed Aug 8, 2019].
- [12] Bambenek Consultings, "Master Feeds," [Online]. Available: <http://osint.bambenekconsulting.com/feeds/>. [Accessed Aug 8, 2019].
- [13] D.B. Cid, "Log analysis for intrusion detection," *dcide.me*, [Online]. Available: <https://dcid.me/oldtexts/log-analysis-for-intrusion-detection.txt>. [Accessed Aug 8, 2019].
- [14] Spamhaus, "About Spamhaus," [Online]. Available: <https://www.spamhaus.org/organization/>. [Accessed Aug 8, 2019].
- [15] The Apache Software Foundation, "Apache Spot Documents," [Online]. Available: <https://spot.apache.org/doc/>. [Accessed Aug 8, 2019].
- [16] The Apache Software Foundation, "Apache Spot Ingestion," [Online]. Available: <https://spot.apache.org/project-components/ingestion/>. [Accessed Aug 8, 2019].

- [17] The Apache Software Foundation, "Apache Spot Machine Learning," [Online]. Available: <https://spot.apache.org/project-components/machine-learning/>. [Accessed Aug 8, 2019].
- [18] The Apache Software Foundation, "Suspicious Connects Analysis," [Online]. Available: <https://spot.apache.org/project-components/suspicious-connects-analysis/>. [Accessed Aug 8, 2019].
- [19] The Apache Software Foundation, "Visualization," [Online]. Available: <https://spot.apache.org/project-components/visualization/>. [Accessed Aug 8, 2019].
- [20] The Apache Software Foundation, "Apache Spot Open Data Model," [Online]. Available: <https://spot.apache.org/project-components/open-data-models/>. [Accessed Aug 8, 2019].
- [21] The Apache Software Foundation, "Apache Spot Download," [Online]. Available: <https://spot.apache.org/download/>. [Accessed Aug 8, 2019].
- [22] The Apache Software Foundation, "Apache Incubator. Spot Project Incubation Status," [Online]. Available: <https://incubator.apache.org/projects/spot.html>. [Accessed Aug 8, 2019].
- [23] McAfee, "McAfee Global Threat Intelligence," [Online]. Available: <https://www.mcafee.com/enterprise/en-us/threat-center/global-threat-intelligence-technology.html>. [Accessed Aug 8, 2019].
- [24] Docker, "Docker Desktop," [Online]. Available: <https://www.docker.com/products/docker-desktop>. [Accessed Aug 8, 2019].
- [25] Docker, "Install Docker Desktop for Windows," [Online]. Available: <https://docs.docker.com/docker-for-windows/install/>. [Accessed Aug 8, 2019].
- [26] Docker, "Docker Desktop for Mac," [Online]. Available: <https://hub.docker.com/editions/community/docker-ce-desktop-mac>. [Accessed Aug 8, 2019].
- [27] Docker, "Logs and Troubleshooting," [Online]. Available: <https://docs.docker.com/docker-for-windows/troubleshoot/#virtualization-must-be-enabled>. [Accessed Aug 8, 2019].
- [28] Docker, "Logs and Troubleshooting," [Online]. Available: <https://hub.docker.com/r/apachespot/spot-demo>. [Accessed Aug 8, 2019].

