
TECHNICAL REPORT

# Metadata Pilot Shaping Workshop

AUTHORS

Dr. Laura Freeman and Dr. Maegen Nix

Technical Report No: 22-0031
18 December 2022

Distribution Statement: Unclassified

Virginia Tech National Security Institute
1311 Research Center Drive
Blacksburg, VA 24060

Virginia Tech National Security Institute
900 North Glebe Road
Arlington, VA 22203

## Background

Executive Order (EO) 14028, "*Improving the Nation's Cybersecurity"*, issued May 12, 2021, requires agencies to meet cybersecurity standards and requirements that will protect and secure government computer systems across on-premises, cloud-based, and hybrid environments. Section 8 of the Executive Order focuses specifically on the capture and retention of information from network and system logs derived from Federal Information Systems. This data is integral to the investigation and remediation of cyber incidents. Subsequently, the Office of Management and Budget (OMB) issued, M-21-31," Improving the Federal Government's Investigative and Remediation Capabilities Related to Cybersecurity." M-21-31 established specific technical requirements for departments and agencies pertaining to Section 8, "ensuring centralized access and visibility for the highest-level enterprise security operations center (SOC) of each agency" and increasing information sharing "as needed and appropriate, to accelerate incident response efforts and to enable more effective defense of Federal information and executive branch departments and agencies."

Virginia Tech's National Security Institute hosted two workshops on June 22 and September 13, 2022, to evaluate how the log generation, retention, management, and sharing approach outlined in the guidance could provide the federal government with sufficient data to effectively detect, protect and secure federal information. Participants strongly proposed that the analysis of metadata about network traffic and transactions would significantly enhance the federal government's logging functionality and cybersecurity capabilities. Workshop attendees also identified limitations and risks associated with existing logging methodologies. For example, the SolarWinds attack manipulated logs by stopping Sysmon and Splunk device logging and cleared Windows Event Logs to cover their tracks.[1]

Shortly after our workshops, in early October, the Cybersecurity and Infrastructure Agency (CISA) issued a binding operational directive to improve, "asset visibility and vulnerability detection on federal networks." By April 3, 2023, agencies must perform automated iterative asset discovery weekly as well as identification and reporting of suspected vulnerabilities on those assets. This will require CISA to develop a standard that includes an extensible metadata taxonomy for agencies to leverage when detecting and subsequently reporting vulnerability information. The proposed metadata pilot directly supports the development of such a standard.

## Purpose of a Federal Metadata Pilot

Both workshops defined an essential pilot operational goal: the federal government must identify, generate, retain, and analyze the necessary data to rapidly detect and investigate anomalies on its own systems, before third-party discovery and notification. As illustrated by the SolarWinds compromise, when the federal government was notified of the breach by third parties, federal agencies often lack real time situational awareness across networks and, as a result, may be unable to detect and respond to cyber-attacks.

One way to accomplish this goal is to enhance the existing log processes with metadata techniques. Network logging provides a mechanism to capture information about activity on networks at endpoints

[1] Moreover, the threat disabled SysInternals Sysmon and Splunk Forwarders on victim machines accessed via Microsoft Remote Desktop. These log manipulations made the attack difficult to detect forensically. See: https://www.mandiant.com/resources/russian-targeting-gov-business.

at trust boundaries. There are, however, numerous risks associated with relying solely on traditional log-based methodologies for detecting cyber-attacks. First, logs only capture characteristics that have been programmed into the logging tools from known threat activity. Organizations that rely exclusively on logs for detection and incident response may miss attacks that were previously undocumented. Second, traditional logs only reflect activity at the boundary endpoints that generate them; they subsequently miss movement by threat actors between endpoints. In addition, attacks that compromise the components that generate logs have the potential to make those logs unreliable. Network transactions, however, can be externally observed and do not rely upon a potentially compromised host to generate a record of the transaction, or in other words, network communication either exist or don't exist, and if they do, they can be observed. And finally, log review does not provide detection and response at the speed of the cyber threat because traditional logs cannot be easily indexed or rapidly searched. As a result, traditional logs hinder the development and application of machine learning tools that could rapidly detect new anomalies, threat behaviors and other indicators of compromise.

## Potential Benefits of a Federal Metadata Pilot

Workshop participants hypothesized that the capture and analysis of network traffic metadata both north-south (i.e., traffic transiting to and from the network) and east-west traffic (i.e., traffic moving within internal networks) would decrease the time to detect an intrusion in a government network. Metadata generation turns raw, non-content focused, network transaction information into a compact and processable format that may also enhance the privacy of communications. Targeted extraction allows for the identification and preservation of the most important information about data movement through a network. In addition, east-west network transactions can be externally observed and do not rely upon a potentially compromised host to generate a record of the transaction.

This form of network communication can be measured and observed, and its metadata can be easily indexed, searched rapidly, and stored in a cost-effective manner over long periods of time. By rapidly analyzing this data with a variety of existing and emerging cybersecurity tools, detection and incident response, threat hunting, and damage assessments could be accelerated. Because it can be stored over longer periods of time, network traffic metadata can be used to conduct retrospective analysis when new indicators of compromise are discovered. Organizations can also use network traffic metadata to validate the integrity of log records as traffic that is not logged, may be indicative of a compromise. Finally, metadata generated and stored over time, enables the use of ML tools to baseline normal traffic and detect anomalies for large amounts of data. These tools have the potential to reveal patterns of "low and slow" threat activity that may otherwise be undetected. The ability to analyze large pools of metadata quickly would also facilitate the rapid correlation of threat activity across multiple environments.

Participants discussed how to operationalize and innovate the logging of cyber-relevant network transactions. For example, packet capture (PCAP) data in its raw form is difficult to deal with and impossible to index, which means it cannot be searched rapidly or stored at reasonable cost. In addition, current standards of preserving 72 hours of PCAP data cannot support intrusion forensic requirements because threat hunters and incident analysis often requires data over much longer timespans. The use of focused PCAP metadata, on the other hand, would require less storage allowing for cost-effective long-term storage and analytics to facilitate AI/ML threat hunting as well as the detection of anomalous cyber behavior.

## Metadata Pilot Design and Requirements

Workshop participants outlined a proposed pilot design. The pilot must prove out the practice of metadata capture by collecting metadata on a government network and analyzing that data. The pilot should be tied to the transition of network services to the Enterprise Infrastructure Solutions (EIS) contract. Pairing the pilot to a new EIS contract would provide an opportunity to introduce necessary data requirements into the contract from the beginning, streamlining data capture processes. The pilot should evaluate different fidelities of metadata capture to balance cost, storage, privacy and data capture effectiveness to guide future standards for metadata. The pilot should be designed such that it provides credible information to:

- Assess and implement a metadata collection taxonomy to enable technical architecture development.
- Provide information to estimate costs for government wide deployment of metadata capture.
- Provide a baseline understanding of privacy preservation through metadata capture.
- Provide a basis for determining the adequacy of legacy infrastructure, to include storage for capturing, maintaining, and using metadata for detection of threats in networks.
- Provide data to assess the feasibility of scaling and extensibility to all federal networks.

## Pilot Timing and Execution

Workshop attendees agreed that an initial pilot should be planned for 12 months with a 6-month interim report aligning the pilot with several key cybersecurity initiatives underway. It should leverage existing infrastructure and production environments without requiring new procurements. As one workshop participant put it, "we just need to capture the data and then analyze it." This does not require the development of new technologies or a new acquisition program.

The selection of a government agency is critical to a successful pilot. Workshop participants recommended that an ideal partner would include an agency with multiple environments that is rearchitecting its network(s) to better align with EO 14028 cloud and zero trust mandates and transitioning to new services offered through the EIS contract. This agency would work with technical experts to directly scope what metadata should be captured and conduct analysis using existing cybersecurity tools. Resulting analysis, or the raw metadata, could be shared with CISA for further analysis and correlation across the federal enterprise.

## A Baseline for the Future

The overarching consensus identified during both workshops was that a metadata-focused pilot could build a foundation to protect federal assets from cyber-attacks. It could enable rapid attack detection and response, mitigating the consequences of a successful attack. Agencies must have the ability to detect an ongoing attack in near real time and reduce their dependency on third parties for breach notification after the initial compromise.

A pilot will help to uncover best practices and areas of further investigation regarding effective metadata architecture, taxonomies, and standards in the areas of data collection, data query processes and analytics. Once the process and practice of metadata capture is proven out, there could be widespread application of such information. Workshop participants outlined additional potential uses and benefits of metadata:

- Retrospective looks to identify anomalous behavior using unsupervised learning. Anomalous behavior could be examined in aggregation in the network traffic or in a more directed manner by evaluating patterns for individual users or endpoints.
- Development of new ML models to enable real time detection analysis.
- Enhanced data sharing across networks for evaluation. Workshop participants hypothesized that sharing data across government agencies would provide a more comprehensive view of the threat landscape. It would enable analysts to build more accurate anomaly detection and predictive models.

Notably, storing metadata over time will enable retrospective analysis and updates to best practices when new indicators are discovered. Ultimately, answering the question of how to best secure networks against an ever changing and evolving cyber threat requires the ability to rapidly experiment with new threats as they emerge. From this vantage, metadata would also support research and development to test innovative approaches to security, compliance measurement, and threat detection.

## Attendees

| Name | Organization |
|---|---|
| Patty Tatro | Virginia Tech |
| Randy Marchany | Virginia Tech |
| Loren Smith | GSA/FAS/ITC/ETS |
| Scott Midkiff | Virginia Tech, Division of IT |
| Nicholas Andersen | Invictus International Consulting, LLC |
| John Forte | Virginia Tech Applied Research Corporation |
| Ian Farquhar | Gigamon |
| Erin Lanus | VT NSI |
| Selene Ceja | Rep. Ro Khanna (CA-17) |
| Karen Evans | Cyber Readiness Institute |
| William Minarchi | Ocient Inc. |
| Antonio Ibáñez | Ocient, Inc. |
| Orlie Yaniv | Gigamon |
| Michael Daniel | Cyber Threat Alliance |
| Mei-Ling Freeman | FCC |
| Craig Saperstein | Pillsbury Winthrop Shaw Pittman LLP |
| Matthew Plummer | Gigamon |
| Triton Pitassi | JHUAPL |
| Nick Polk | Office of Management and Budget |
| Laura Freeman | VT NSI |
| Maegen Nix | VT ARC |
| Chris Cummiskey | Virginia Teach |
| Ron Ross | NIST |