TECHNICAL REPORT

Metadata Workshop Report - Defining the Pilot

AUTHORS

Dr. Laura Freeman, Deputy Director, Virginia Tech National Security Institute
Dr. Maegen Nix, Director, Decision Science Division, Virginia Tech Applied Research Corporation

Technical Report No: 24-0165
26 February 2024

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

Contents

Background

Recent cyber events and direction from the Executive Branch call for a rethinking of how the Federal Government, in partnership with the private sector, must improve efforts to identify, deter, protect against, detect, and respond to malicious cyber campaigns.1 In response to this call, Virginia Tech's National Security Institute hosted a series of three workshops between June 2022 and August 2023 to evaluate how metadata about network traffic moving across and within federal civilian agencies could be leveraged more effectively to detect, protect and respond to cyber threats and to better secure federal information. Specifically, the workshops explored the concept of generating metadata about network traffic traversing government networks to enhance the government's ability to rapidly detect and respond to cyber threats.

The objectives of the three-part workshop series progressed from concept validation and requirements exploration to pilot opportunities.

1. Workshop 1 validated the technical concept that the analysis of metadata about network traffic would significantly enhance the federal government's cybersecurity capabilities. The workshop also identified associated technology and policy challenges and potential solutions.
2. Workshop 2 explored the requirements for a government pilot that would leverage metadata about network traffic for anomaly detection with the goal of detecting cyber threats, incidents, and attacks before third party discovery and notification.
3. Workshop 3 provided a more specific definition of metadata and refined what data should be generated and how that data could be used to enhance federal civilian agency cybersecurity in a pilot.

The initiative responded to evolving requirements and guidance from the federal government. For example, 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 EO focuses specifically on the capture and retention of information from network and system logs derived from Federal Information Systems.[2] 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*."[3] M-21-31 establishes specific technical requirements for departments and agencies pertaining to Section 8 of EO 14028, ensuring "centralized access and visibility for the highest-level enterprise security operations

[1] https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/

[2] Traditional logs were typically diagnostic in intent and, around security, evolved to report the state from a security device or system. This report evolves and broadens the concept of a log in alignment with NIST SP 800-92r1: (log) that states a log is a record of events occurring within an organization's computing assets, including physical and virtual platforms, networks, services, and cloud environments.

[3] https://www.whitehouse.gov/wp-content/uploads/2021/08/M-21-31-Improving-the-Federal-Governments-Investigative-and-Remediation-Capabilities-Related-to-Cybersecurity-Incidents.pdf. CISA issued further implementation guidance in December 2022 TLP CLEAR - Guidance for Implementing M-21-31 Improving the Federal Governments Investigative and Remediation Capabilities .pdf (cisa.gov).

center (SOC) of each agency"; to increase 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." While M-21-31 raises the bar on data collected and stored by agencies, participants at all three workshops agree that the analysis of metadata about network traffic and transactions would significantly enhance the federal government’s logging functionality and cybersecurity capabilities.

In January 2022, OMB issued additional EO implementation guidance in OMB Memorandum M-22-09, “*Moving the U.S. Government Toward Zero Trust Cybersecurity Principles*”. [4] M-22-09 sets forth a federal zero-trust architecture strategy with specific cybersecurity standards and objectives agencies must meet by the end of FY 2024 to “reinforce the Government’s defenses against increasingly sophisticated and persistent threat campaigns.” The strategy acknowledges that perimeter-based defenses are insufficient and mandates a paradigm shift where every user, device, application, and transaction is continuously verified. Among other things, M-22-09 envisions that agencies will “encrypt all DNS and HTTP traffic within their environment and begin executing a plan to break down their perimeters into isolated environments.” As agencies work towards this vision, they will still need a way to detect adversaries moving laterally between isolated environments via encrypted channels. As the volume of encrypted network traffic grows exponentially, it will become increasingly difficult for agencies to break and inspect all network traffic between and within network segments.[5] Recognizing these challenges, M-22-09 states “[n]etwork traffic that is not decrypted can and should still be analyzed using visible or logged metadata, machine learning techniques, and heuristics for detecting anomalous activity.” The proposed pilot will provide a means for achieving this objective.

In April 2022, and CISA issued the “*Extensible Visibility Reference Framework (eVRF) Program Guidebook*”[6]. The eVRF Guidebook focuses on visibility surface identification and the determination of coverage gaps. Participants noted that it provides a framework for understanding metadata collection concepts and requirements through an exploration of visibility coverage maps[7]. The pilot discussed by participants could be used with this guidance to map out specific metadata requirements.

The EO, OMB Memorandum, and CISA guidance described above drive home the need to think differently about securing sensitive data and creating agility in data extraction and analysis. The focus of this report and the workshop series highlight the potential benefits of moving beyond the static and traditional log generation by network components at end points and instead creating a record of cyber relevant events through the observation of network traffic moving within and between network segments and environments within an enterprise, or in other words, monitoring north-south AND east-west network traffic. This will enable agencies to better detect and respond to cyber threats moving laterally within an environment because they will be able to analyze this data using machine learning tools. Leveraging this capability to analyze high volumes of extracted data could make a game-changing difference to cyber defense.

[4] https://www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf
[5] Break and inspect capabilities require significant compute power and resources, which may be lacking within any specific agency.
[6] eVRF_Guidebook_RFC_508C.pdf (cisa.gov)
[7] https://www.cisa.gov/sites/default/files/publications/eVRF_Guidebook_RFC_508C.pdf

Prior Workshop Summaries

The first workshop sought to explore the concept of a metadata pilot and validate the general concept. The complexity of securing government networks and detecting threats provides numerous challenges. These include the and absence of formal standards for network metadata logging; neglect of critical lateral network traffic metadata within the enterprise perimeter; as well as the intricacies of managing hybrid cloud, mobile endpoints, and remote work configurations. Additionally, agencies must work with telemetry providers like Microsoft to capture metadata while preserving data integrity and authenticity and maintaining affordability. Other challenges include data sharing among organizations, aligning priorities between network engineers and security architects, and developing multi-agency risk evaluation frameworks. Cloud solution providers also tend to lock agencies into a specific product, making it difficult to interoperate with different logging and telemetry formats.

Workshop attendees hypothesized that a solution to these challenges lies in the implementation of a pilot program that can generate and analyze metadata about network traffic. Such a program could establish standardized metadata taxonomies for government networks, ensure consistent data collection and storage practices, and enhance interoperability. It would also ensure that network traffic moving within an enterprise is monitored. By monitoring both north-south (NS) and east west (EW) traffic, the ability of the agencies to detect an adversary moving laterally increases [8] Furthermore, the program would facilitate collaboration with telemetry providers, to shape requirements for affordable metadata generation while maintaining data integrity.

Through this pilot, the government could look at methods for optimizing data storage and transmission, investigate the value of information derived from the correlation of telemetry from multiple logging sources, develop robust data integrity and source assessment mechanisms, and implement user and entity behavioral analytics. It could also seek effective strategies for aligning the priorities of network engineers and security architects (for example merging the functions of the Network Operations Center with the Security Operations Center). The pilot should also explore a structured framework for multi-agency risk evaluations and promoting data sharing incentives.

The second workshop sought to shape the purpose, benefits, design, requirements, and future impact of a metadata pilot. Workshop attendees developed several requirements for the pilot program including:

- 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.

[8] EW traffic is a class of traffic where the source and destination of that traffic are within the same agency or network architecture. North-South traffic (NS) means traffic is coming from or going to a source outside of the network. There is not a single EW environment. Rather, there are tens or even hundreds of them in a typical enterprise. To create full visibility, security tools must be able to see and inspect any traffic that terminates or passes through all virtual or physical network adapters in the organization, any endpoint or infrastructure device, and any physical, virtual or cloud infrastructure.

- Provide data to assess the feasibility of scaling and extensibility to all federal networks.

The workshop highlighted the potential for a pilot 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. A pilot could inform how unsupervised learning methods could identify anomalous behavior; explore the potential to develop new machine learning based models that could be deployed in real time after training; and the ability to pilot enhanced sharing of data across networks and agencies for improved prediction capabilities.

Even so, a critical question remained at the end of the two workshops – what metadata are we talking about? This question led to workshop three which focused on defining metadata, providing examples of capabilities that can be used to capture metadata, and understanding potential challenges that the move towards increased encryption requirements driven by OMB Memo, M-22-09 might pose towards metadata capture.

Defining Metadata

Network metadata is information extracted from network traffic. This can include metadata sourced from any Open Systems Interconnection (OSI) model stack layer (1 to 7) and can include or exclude information sourced from headers or content.

It is important to note that this capture must occur not just at the perimeter, but in physical agency networks, public and private cloud environments, even down to individual devices (in accordance with M-22-09).

During the workshop, Ian Farquhar, Security CTO, Gigamon provided a straightforward example of potential metadata capture. The example included metadata from:

- Layer 1 – the location of the interception point
- Layer 2 – MAC (media access control) addresses, encapsulations, ARP (address resolution protocol) requests
- Layer 3 – IP (internet protocol) addresses, ports
- Layer 4 – size (bytes and packets in each direction), duration, state
- Layer 7 – detailed extraction of protocol metadata (protocol dependent)

The workshop discussed that capture of this information is a choice that could be recorded and discarded to meet specific security or data storage requirements. Additionally, the data could be enriched with context based on variables such as geolocation, identity, telemetry chosen, and threat feeds.

Workshop participants shared several metadata collection experiences as input to the discussion. Randy Marchany, the Virginia Tech (VT) CISO, highlighted that VT uses a central logging service (CLS) on high and moderate risk systems because adversaries delete logs.[9] Implementation of the CLS required the team to work through dealing with sensitive data (e.g., PII and CUI) and capacity. Initially, VT generated hundreds of gigabits per day and did not have the scanning and analysis tools to filter and decide what information to store. Consequently, VT decided to leverage a SOC run out of Indiana University that

[9] MITRE ATT&CK techniques T1070.001, T1070.002, T1070.003, T1562.002, T1562.003, T1562.008.

supports a dozen locations and generates 1.1TB of data a day. These experiences highlight that a metadata pilot must also address data storage, filtering, analysis, and deletion on reasonable timelines.

Michael Daniel, CEO Cyber Threat Alliance (CTA), highlighted the experiences of CTA, where member companies share indicators of compromise through the alliance. He estimated that members shared roughly 350,000 indicators per day. When dealing with data on this scale, formatting inconsistencies are a challenge. He noted that they require the STYX II format, although there is still a lot of variation. The CTA is made up of 35 members split between the US and overseas, organizations have different security needs ranging from operational technology (OT), networks, and telecommunications, and endpoint detection response (EDR). The diversity of organizational needs and language differences also leads to challenges expanding information. He highlighted that most of the indicators of compromise reported were using non-native tooling – but that could also be biased by the fact that those are the easiest indicators to identify. Finally, storage is also an issue for CTA, where they must debate how long to keep the data and have currently settled on 18 months.

Participants noted that government standards could help with variations in data, spelling, and reporting. The cost of both securing data and then storing it was an important consideration in participant minds. Overall, organizations and the government do not want to spend lots of money to collect everything if it does not meet the end objective of enhanced government visibility of cyber-attacks before, during, and after a cybersecurity incident.

Josh Perry, Gigamon, highlighted lessons from a Trusted Internet Connections (TIC) 3.0 pilot in Azure. He noted Azure still leverages log analytics when working with CISA hunt team. A key question is how to collect data and hunt against it. Commercially provided cloud environments operated by different providers offer varying standards, levels of visibility, capability, as well as features and functionality within their toolsets. These differences make incident attack, detection, triage, and response challenging. When coupled with a limited workforce capacity and capability, they compound the problem.

Reflecting on what should go in the base package, participants highlighted that network flow data is an absolute minimum. It provides the dual properties of preserving privacy and being easy to capture. All agreed, however, that it was not the most sensitive indicator of threat detection. Participants noted that there are many tools that analyze flow records, and the government should not drive towards a specific solution. Rather, it should work to normalize stored data across solutions for consistency. Building on flow data, additional elements such as ICS, SMTP, and file carving through a cryptographic hash such as SHA2 should be considered as part of the pilot. Context is also important for triangulation of threats so ports, geolocation, and identify elements should be considered in situations where they do not compromise privacy or security. Industry representatives highlighted best practices in industry include autonomous system number, destination, extension protocols, ports, IP addresses, URL, and domain name should be captured.

Metadata and Encryption

One concern addressed during the workshop regarded the impact of encryption on a metadata approach. Figure 1 shows data from 2022 that looks at traffic across the perimeter boundary (north-south) and within the parameter (east-west). As the government pushes for increased encryption of all messages, we must consider the implications on a pilot.

Participants discussed two approaches – breaking encryption and capturing available metadata. Subject matter experts attending the workshop noted that even with encrypted protocols, useful metadata is usually present, and if it is not present that is an indicator of interest. Meta-data-based approaches for encrypted data include Transport Layer Security (TLS) handshake metadata extraction and TLS data flow analysis. The combination of the available metadata combined with multi-session behavior over time provides a useful analytical basis for detecting malicious behavior.

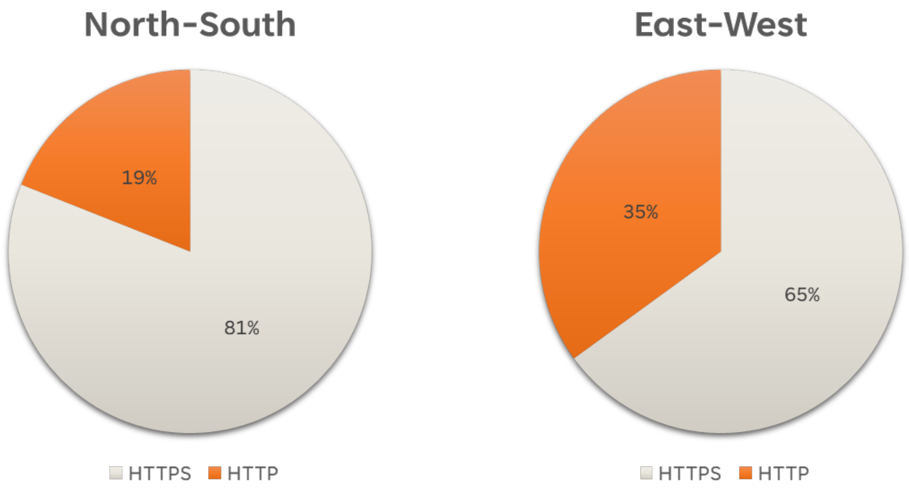

Figure 1. Encryption[10]

Conversation during the workshop reflected that metadata capture is the right path forward for encrypted traffic as it can avoid breaking encryption and still provide relevant information. Participants also noted that protocols which deliberately minimize or evade metadata extraction are themselves usually a significant indicator of concern and could be captured as a data element. An example of this includes the ECH (Encrypted Client Hello) protocol under TLS, a conscious attempt to avoid TLS metadata visibility. Currently, ECH is in front of the IETF as draft 17. Even the use of this, in a typical agency network, would be a significant indicator of risk.

It is important to note that there are multiple metadata extraction libraries, multiple metadata formats, and multiple transport protocols for metadata. The goal of the pilot should not be to determine which one of these should be used, but rather to identify the critical information that needs to be captured and standards so that multiple agencies data can be used in an interoperable way to better find threat actors. A metadata taxonomy is essential for defining what to capture and when.

Pilot Proposal or Next Steps

Mr. Chris Cummiskey and Ms. Karen Evans provided leadership to the workshop series and guided the overarching discussion. They highlighted the U.S. government continuously expands national

[10] https://www.gigamon.com/content/dam/resource-library/english/infographic/in-2022-tls-trends.pdf

capabilities to protect, detect, respond to, and recover from cyber-attacks. They also noted that adversaries will advance in sophistication and require persistent adaptation and evolution of defensive measures. For example, like domestic defensive networks, adversaries share cybersecurity information and coordinate their attack efforts in a production chain. Threat actors know what we can see and how to disrupt and disable visibility when possible (e.g., disabling logging).[11] The defensive position has a fundamentally harder challenge than the attack position because everything must be done perfectly. Once an attacker is on the inside of the network, however, roles are reversed, and the attacker must avoid being seen. Metadata provides a way of capturing more information that is harder to evade. Such behavior underpins the requirement that federal, state, and local governments should partner and coordinate with academia and private industry to find and develop emerging cybersecurity solutions.

The next critical steps to follow the third workshop hinge upon the development of the pilot itself. This includes bringing together the right technical, infrastructure, and government partners. The following sections explore a potential pilot organization and opportunities to bring these ideas to fruition.

Piloting Organization

Karen Evans noted that as a former CIO, it is important that an agency CIO would need to want to volunteer to be part of this effort. Representatives from the FCC noted that they are in the middle of a major modernization effort that could provide useful for testing the concepts of metadata capture discussed in the workshop out to answer key questions. Additionally, an important part of the pilot would be connecting with CISA on architectures, data analytics, and piloting new capabilities in AI.

Pilot Concept

Participating agencies, including the Cybersecurity and Infrastructure Security Agency (CISA), would aggregate certain defined metadata in a cloud-based platform managed by a neutral, non-governmental non-profit organization with experience in managing an information sharing platform. This organization would contribute its expertise and help evaluate the project.

Agencies would send specified metadata to the platform via an Application Programming Interface at regular intervals. CISA and participating agencies, with support from the managed platform, would analyze the shared data to identify anomalies, trends, or other indicators of malicious activity. The pilot project would be designed to answer questions such as:

- What specific data elements need to be captured?
- What analytic tools are needed to understand and take action on the shared data?
- Does aggregating the metadata across agencies enhance the ability to detect malicious behavior?
- How frequently should data be sent to the platform?
- How often should agencies review the shared data?
- How long should data be retained at different levels of availability (hot versus cold storage)?

The managing organization would create operating guidelines for the pilot program; draft a user agreement; develop, deploy, and maintain the sharing platform; and deploy any necessary analytic

[11] Reference example MITRE TTPs where adversaries disable and/or modify logs.

tools. The managing organization would provide appropriate security for the shared data, which would only be available to participants. The pilot would operate for six months to a year, which would provide sufficient time to evaluate the program's utility.

The pilot program would be overseen by a steering committee made up of representatives from participating agencies. A technical committee with representatives from participating agencies and the managing organizations would agree on any technical specifications needed to make the pilot operational.

Project Components

The project would have seven technical elements:

1. *Systems Engineering (SE) and Systems Integration*: The pilot will need to identify existing and emerging requirements for the new concept aligned with participating agencies. SE will be responsible to design and implement the pilot architecture, system design, and milestones in accordance with operational needs and existing technical specifications of system requirements. The Principal Investigator (PI) will be responsible for these technical activities and coordinate other areas of technical responsibility.
2. *User Interface (UI)*: users will need a way to interact with the platform. Although the project can leverage open-source and other existing code, the interface will need to incorporate certain custom components, and it will need refinements to ensure usability across a wide range of skillsets and agency missions.
3. *Infrastructure*: The project will need cloud infrastructure to operate. The managing organization would use a cloud service provider to host the data, UI, user management and analytic tools, and data. How much data is sent to and stored on the platform will drive costs in this component.
4. *User Management*: The pilot will need to manage user accounts, authentication, and access.
5. *Analytic tools*: participants will need analytic tools to understand and make use of the data.
6. *API/Data sources*: Agencies would use standard API endpoints to interact with the platform.
7. *Quality Management/Review and V&V*: The technical team will develop a process supported by agency oversight on quality management and pilot progress.

Cost Estimate

The cost of the pilot will depend on the duration, size, and scope of participating organizations and CISA contributions. Key components of the pilot are identified below and could likely be accomplished in 6-12 months for $600k - $1M in funding based on a rough analysis of similar programs.

Component		**Primary Driver**
User Interface		Engineer Time
Infrastructure		CSP fees

User Management		CSP tool
API/Analytic tools		Engineer Time and license fees
Development Ops		Engineer Time
Systems Engineering & Systems Integration		Engineer Time
Project Management		Cost, Schedule, Performance Management
Quality Review		Engineer Time

Workshop Attendees

Name	**Organization**
Roger Cressey	Mountain Wave Ventures
Chris Cummiskey	Cummiskey Strategic Solutions, LLC
Michael Daniel	Cyber Threat Alliance
Joseph Drummond	CISA
Karen Evans	Cyber Readiness Institute
Ian Farquhar	Gigamon
Laura Freeman	VTNSI
Allen Hill	Federal Communications Commission
Antonio Ibanez	Ocient, Inc.
Danielle Kauffman	VTNSI
Dayton Maco	Ocient, Inc.
Randy Marchany	VT
Maegen Nix	VT-ARC
Dave Otto	CISA
Joshua Perry	Gigamon
Dave Powner	MITRE
Travis Rosiek	Rubrik
John Scott	Exiger/Ion Channel
John Simms	DHS/CISA
Loren Smith	GSA
Brian Vu	GSA
Fred Walton	T-Mobile
Orlie Yaniv	Gigamon