TECHNICAL REPORT

Metadata Workshop Report

AUTHORS

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

Technical Report No: 22-0030
27 July 2022

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

Executive Order (EO) 14028 *Improving the Nation's Cybersecurity*, issued May 12, 2021, requires agencies to meet cyber requirements that will protect and secure government computer systems across resident, cloud-based, and hybrid operational environments. Section 8 of the Executive Order focuses specifically on the capture of information from network and system logs derived from Federal Information Systems. Logging, log retention, and log management are integral to the investigation and remediation of cyber incidents. Following the passage of EO 14028, the Office of Management and Budget (OMB) issued M-21-31 *Improving the Federal Government's Investigative and Remediation Capabilities Related to Cybersecurity* on August 27, 2021. This memorandum establishes a maturity model for event logging in support of section 8 requirements. Its purpose is to ensure "centralized access and visibility for the highest-level enterprise security operations center (SOC) of each agency" and to "increase the sharing of such information, as needed and appropriate, to accelerate incident response efforts and to enable more effective defense of Federal information and executive branch departments and agencies." As agencies continue to implement the Executive Order and comply with OMB's logging maturity model, their integrated approach will determine the success of improving the Unites States' cybersecurity posture. In order to do so, however, we must look more closely at logging architectures and methods that will more securely and effectively acquire, aggregate, filter and share information regarding threats that reside within our enterprise logs and network traffic.

Workshop Motivation & Objectives

Virginia Tech's National Security Institute hosted a workshop on June 22, 2022, to investigate the benefits network traffic metadata could bring to logging functionality and cybersecurity, to include its integration within SOC architectures designed to scale across the Federal ecosystem. (This specifically includes the logging of cyber-relevant network transactions as opposed to traditional logging of cyber-relevant internal state changes within a workload.) Experts across government, academia, and industry met to discuss concrete measures that would further support the fulfillment of EO 14028 Section 8 and M-21-31. Specifically, we focused on "Leveraging Metadata to Better Protect, Detect, and Respond to Cyber Threats on Federal Civilian Networks." The group looked first at existing and emerging challenges and then turned to discuss potential solutions and next steps.

The workshop's critical objective was to elicit feedback and validate a technical concept that would enhance government visibility of cyber-attacks before, during, and after a cybersecurity incident. It explored how the government might leverage existing capabilities and infrastructure to generate the data needed to meet EO 14028 objectives. And it also captured associated technology and policy challenges, potential solutions, and legislative needs to implement an initial concept pilot.

Mr. Chris Cummiskey and Ms. Karen Evans provided leadership to the workshop with opening remarks and initiated the overarching discussion. They highlighted the U.S. government continuously expands national capabilities to protect, detect, respond to, and recover from cyber-attacks. They also noted, however, adversaries will advance in sophistication and require persistent adaptation and evolution of defensive measures. For example, similar to domestic defensive networks, adversaries also share cybersecurity

information and coordinate their attack efforts in a production chain. Such behavior underpins the requirement that federal, state, and local governments, partner and coordinate with academia and private industry to find and develop emerging cybersecurity solutions.

Challenges

Participants noted numerous challenges that must be faced when protecting government networks and detecting threat actions. They agreed that DHS vice industry should have the capability to swiftly brief the legislative branch on malign activity that occurs on federal networks. One speaker stated that CISA and other agencies will always be dependent on incident detection by Microsoft and other service providers unless government organizations can change their architectures, business practices, and solution technologies. A large component of the overarching challenge is sustained by the absence of common standards regarding what should be logged, how it is stored, and protocols for sharing.

From this line of discourse, interconnected themes emerged:

1. There is no formal standard for network metadata logging. In order to maximize the usefulness of metadata and interoperability between organizations, to include storage, a standard and focused metadata taxonomy is needed. The standard should allow for an an absolute minimum and increasing requirements for more critical assets. However, because every government agency cares about a different set of things, parsing across agencies to gain a base will be difficult. Each maintains a different view of risk and threat as well as different levels of technical maturity. However standards provide a basis for connection.
2. Current practice does not routinely analyze traffic inside the network, inside the enterprise perimeter. Such movement is termed "East-West traffic" as opposed to "North-South traffic" that moves across the network boundary. East-West workloads generate significant internal traffic that can be used to assess security.
3. We cannot look across different agency clouds (inter and intra) and additionally may need to think about moving from on premise to off premise data centers. Hybrid, multi-cloud, mobile endpoint (COVID/ work from home, plus normal mobility), bring your own device (BYOD), and ICS/SCADA represent additional complexities.
4. How do we incentivize telemetry providers to enable greater metadata capture for affordable service prices?
5. Once collected and stored, how do we preserve the integrity of the metadata itself? Moreover, how can we assess the integrity of the data log source, endpoint assessment from spoofing, minimizing, or deleting.
6. Foregoing the use of metadata increases storage requirements and prevents transmission of data at speed.
7. We lack logs on information like user and entity behavioral analytics that could be suitable for clustering analysis, correlation, and anomaly detection.
8. We do not have a framework for multi-agency risk evaluations.

9. How might we incentivize organizations to share data rather than storing it in silos?
10. What avenue would best enable implementation – new standards, policy, guidance, or is legislation needed? For example, would having metadata requirements in the federal acquisition regulation (FAR) make contracting for this standardized?
11. States are also important to understand the extended attack surface because the federal structure has connections to states. Can we tap into broader industrial based Information Sharing and Analysis Centers (ISACs) and not just to multistate ISACs to gain visibility?
12. The absence of two-way communication is a large challenge. During Solar Winds, for example, the ability to find out what was going on across the rest of the government during the attack would have been a nice capability.
13. Networks are developed by network engineers who do not necessarily design with security outcomes in mind. Rather, network engineers tend to value reliability, performance, and resilience. They switch frames and packets, seeing session reassembly as something relevant only to endpoints. The techniques network engineers use to achieve reliability, performance, and resilience, such as multipathing or asymmetrical routing, work against the needs of security. Security architects or engineers, on the other hand, are frustrated by network techniques. Security-minded architects try to bottleneck traffic in order to perform session reassembly and analysis of content. These methods conversely compromise reliability, performance, and resilience. Ultimately, network engineer and security engineer requirements are diametrically opposed.
14. The US Government (USG) does not collect the appropriate data to effectively protect and defend the federal information and the executive departments and agencies.

Solutions

Following a robust discussion of cybersecurity challenges regarding logging practices and information sharing, the workshop shifted to focus on a solution frame. Prior to the event, the workshop team developed a notional operational concept and technical architecture to foster the use of metadata and create a federated SOC ecosystem. Both the proposed solution and its associated benefits are captured in the attached presentation, "The Merits of Metadata – Opening Presentation." industry developed the conceptual approach and the team used it to spur creativity and feedback on what can and should be done in the future.

Themes from the solutions discussion included:

1. Sharing log information and desensitized metadata constructed from a taxonomy and lexicon, as needed and appropriate, would improve cyber outcomes.
2. The development of standards for metadata logs will speed analytics and help scale storage. For example, three days of packet capture (PCAP) data can be stored or months/years of metadata depending on the taxonomy and compactness of the representation in storage. Metadata is readily indexable and can be easily kept and stored. It would be much easier to run indicators and compromise and threat indicators against historical data.

3. Metadata logging standards should reflect threat intent. It was highlighted that threats over the years have had several unaltered objectives, for example extracting sensitive information or disrupting operations to achieve a strategic intent. Workshop participants suggested that if we could map out the threat intentions that would define the nature of metadata that needs to be collected.
4. Develop a pilot or use case for a handful of government agencies in order to test the proposed metadata architecture presented. Bring the service providers in early within the pilot to illustrate and signal the direction the market for services will go and then enable them to adjust their business models.
5. Enable and empower agencies and their leadership to get rid of outdated legacy systems.
6. Develop logging tamper resistance at both source and sink. The first thing the solar winds attacker did was turn off logging.
7. Develop policy and legislation to enable the capture of East-West traffic within the perimeter.
8. Rethink data storage architecture and security protocols.
9. Incentivize service providers to enable metadata capture as part of standards.
10. Develop on network logs to capture information like user and entity behavioral analytics that could be suitable for clustering analysis, correlation, and anomaly detection.
11. Automate pathways for organizations to share data rather than storing it in silos.
12. Reach out to and enable a broader range of Information Sharing and Analysis Organizations (ISAOs) and ISACs to become part of the Security Orchestration, Automation, and Response (SOAR) community.
13. Network Operations Centers (NOCs) focus on availability. SOCs focus on integrity and confidentiality. These two forms need to transform into a network operations security center to provide a way to respond to cybersecurity attacks without taking everything offline (e.g., Colonial Pipeline). Such an approach would balance business continuity with resiliency.
14. Establish a federate SOC ecosystem to enable two-way communication across agencies.
15. Close the accountability loop with contractors and federal agencies while also being clear on technical specifications for metadata use.

Legislation and Implementation

Workshop participants generally thought that agency CIOs and CISOs have the authorities to implement such policies already. However, challenges arise with contracting, standardization, and implementation costs. Workshop participants also reflected that without clearly defined metadata and/or logging requirements, contracts can vary both across and within agencies. Additionally, new requirements for metadata require contracting modification which should be included in the government network contract managed by the General Services Administration (GSA). As agencies are transitioning to the enterprise infrastructure solutions (EIS) contract, it provides an opportunity to re-engineer the network building the zero trust network capabilities verses doing a circuit for circuit replacement and layering onto a flawed network design. All these factors result in the very high likelihood that without some guiding body, which could be captured in policy, guidance, or standards, there would be a lack of consistency in any implementation of the concept. Furthermore, without standardization, many of the benefits would be lost. This includes the ability to do

threat hunting across multiple agencies and/or build threat detection models that consider a holistic sense of how adversaries might be attacking the United States.

Many questions were posed. Examples include:

- How can we reengineer or take advantage of contract changes so that we can get ahead of the threat evolution?
- Is legislation needed to close the gap?
- Are we on the right track? We have metadata but is it the right data to improve threat detection?
- What avenue would best enable implementation – new standards, policy, guidance, or is legislation needed? For example, would having metadata requirements in the FAR make contracting for this standardized.

Finally, workshop participants discussed how a common standard could be useful to state and local governments as well as industry. They noted that if tools were built to these standards, all organizations with potentially very small budgets could take advantage of the protection and detection capabilities they would provide.

Next Steps

National critical infrastructure and network challenges revealed by the DHS Solar Winds compromise provided a motivational backdrop to the forum discussion. Participants concluded that we need to change the current metadata paradigm in order to better emphasize *detect* and *protect* equities within network defense. They agreed that East-West metadata could provide new insights regarding the detection and protection of federal enterprise networks while balancing *response* and *recover* responsibilities.

In terms of next steps, multiple workshop participants expressed the need to clearly define the specific aspects of metadata that need to be captured and to begin experimentation on the operational concept.

- Conduct a separate workshop, potentially in collaboration with NIST to define metadata, the specific data elements, and collection requirements that would be useful for cybersecurity purposes. A future workshops focused in metadata should be expanded to include representatives from chief data officers and data privacy officers.
- Pilot metadata collection to pilot technologies, work through implementation processes, and validate assumptions about the benefits of collecting metadata described in this report and supporting slides and white paper.
- Update the overview slides to highlight not only the data flow, but also the feedback loop to the agencies. Clearly articulate how the metadata collection leads to improved agency insights.

Attendees

Name	Organization
Nicholas Andersen	Invictus International Consulting, LLC / Atlantic Council
Selene Ceja	U.S. House of Representatives
Chris Cook	Senate Appropriations Committee
Luis Coronado	DHS
Chris Cummiskey	Virginia Tech
Karen Evans	KE&T Partners, LLC
Ian Farquhar	Gigamon
John Forte	Virginia Tech Applied Research Corporation
Laura Freeman	Virginia Tech
Allen Hill	GSA
Joseph Jablonski	Ocient Inc.
Danielle Kauffman	Virginia Tech
Faith Lowe	Office of the National Cyber Director
Randy Marchany	Virginia Tech
Kirk McConnell	US Senate
Scott Midkiff	Virginia Tech
William Minarchi	Ocient Inc.
Claire Montgomery	
Mark Montgomery	Cyberspace Solarium Commission
Maegen Nix	Virginia Tech Applied Research Corporation
Eric Paterson	Virginia Tech
Dennis Reilly	Gigamon
Craig Saperstein	Pillsbury Winthrop Shaw Pittman LLP
John Simms	DHS/CISA
Loren Smith	GSA
Kevin Stine	NIST
Orlie Yaniv	Gigamon

Attachments/Links

- Solutions Presentation discussed at the workshop, "The Merits of Metadata – Opening Presentation"
- Read ahead material, "Leveraging Metadata to Better Protect, Detect, and Respond to Cyber Threats on Federal Civilian Networks"

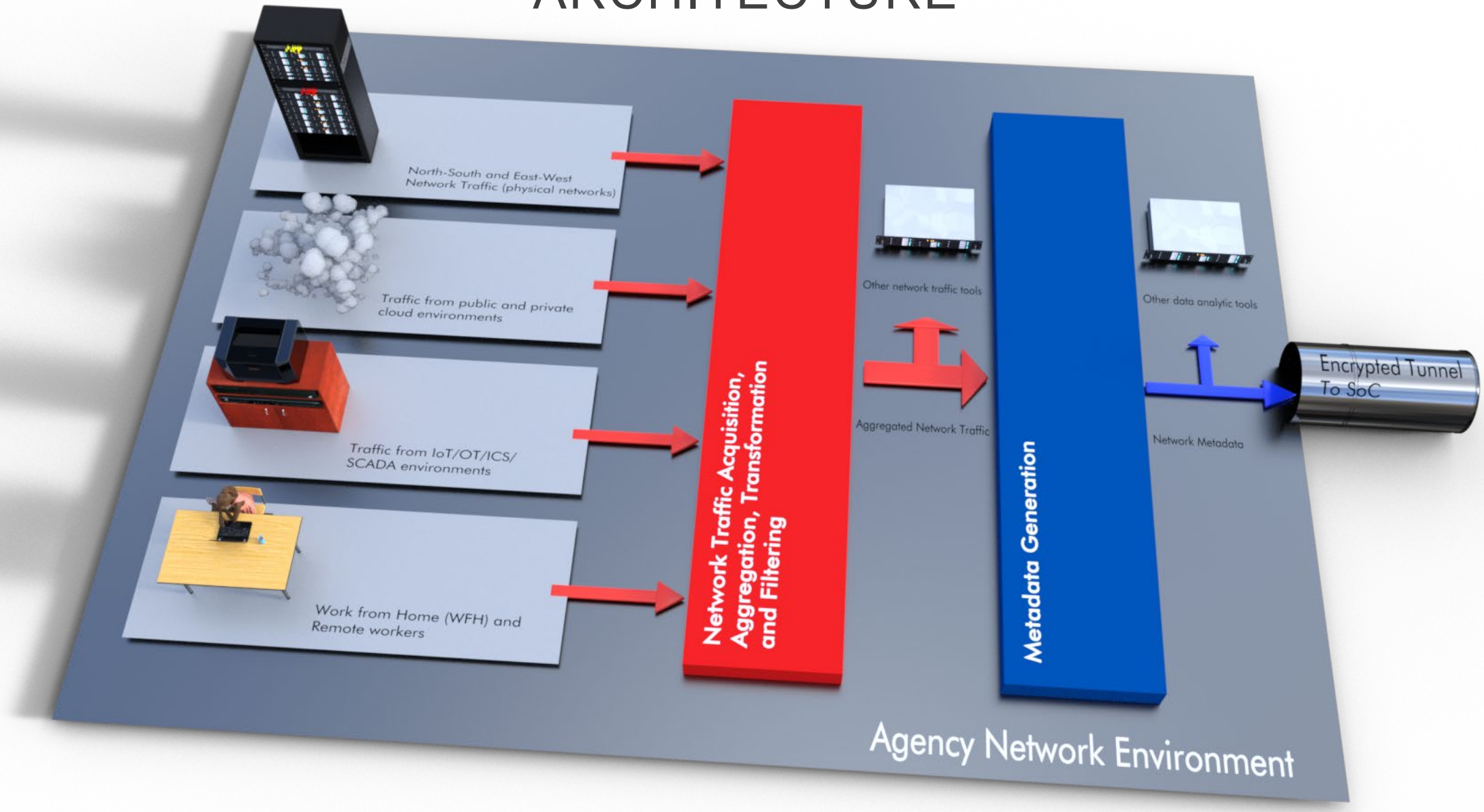
THE MERITS OF METADATA

LEVERAGING NETWORK COMMUNICATIONS TO DETECT, PROTECT, AND DEFEND

WORKSHOP INTRODUCTION

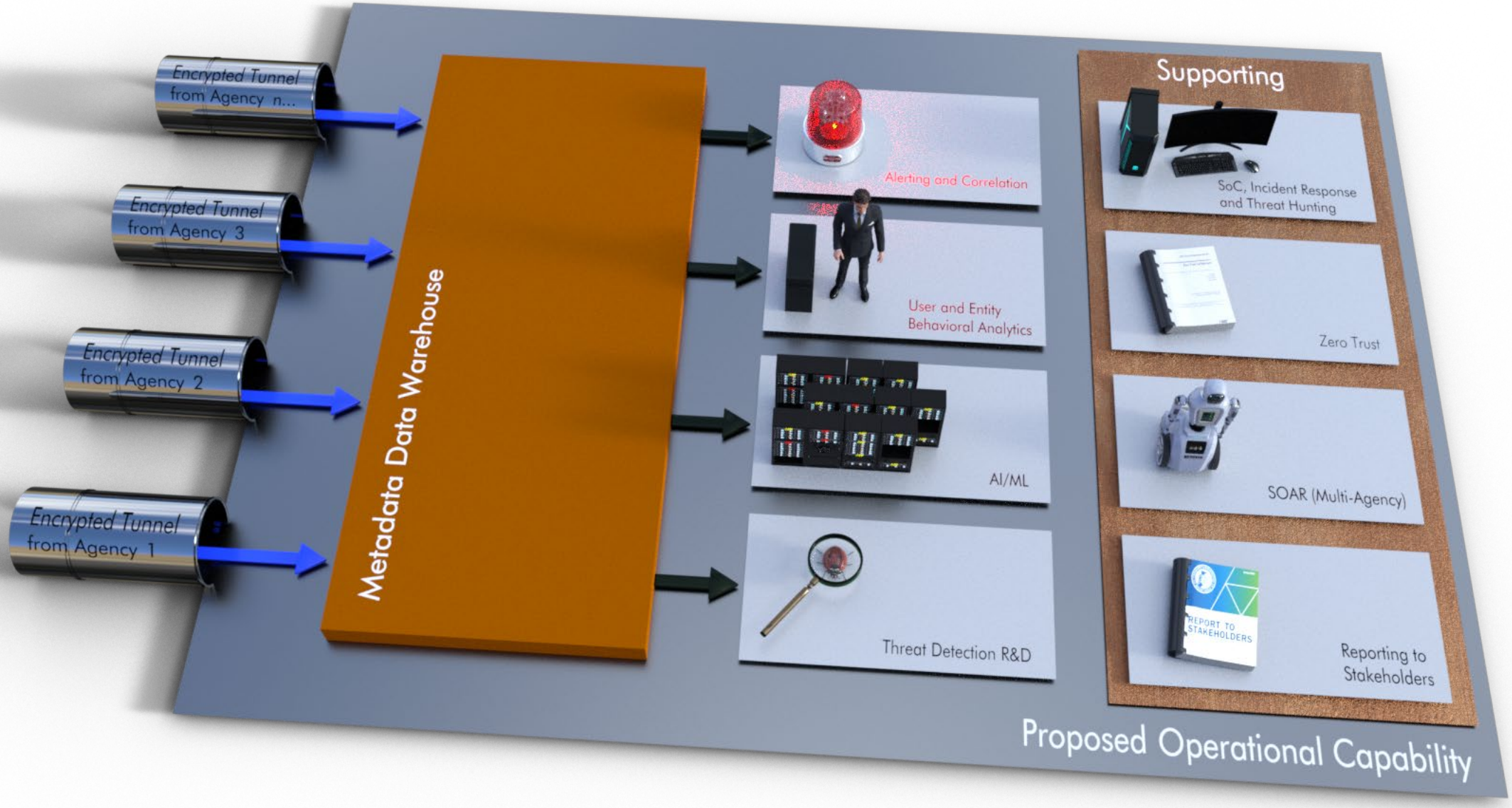
PROPOSED AGENCY METADATA COLLECTION ARCHITECTURE

PROPOSED AGENCY METADATA COLLECTION ARCHITECTURE



PROPOSED SECURITY OPERATIONAL ARCHITECTURE

PROPOSED SOC ARCHITECTURE



BENEFITS

BENEFITS OF THIS PROPOSAL

- **Leverages existing investments** in technology deployed across the federal civilian agencies to:
  - Generate and store rich metadata about **network traffic** over time at a fraction of the file space of full packet capture
  - Enable **rapid search at scale** within minutes
  - Enable **retrospective analysis** when new indicators of compromise are discovered, providing greater insight into threat actor dwell time and activity
  - Rapidly **correlate threat activity** across multiple components, agencies, divisions and/or departments

BENEFITS OF THIS PROPOSAL

- Supports OMB M-21-31 implementation and **strengthens log integrity**
  - Provides log validation and identify communications that are not logged
  - Communications which are not logged may indicate a sophisticated cyber attack (e.g., SolarWinds)
- Compliments the EDR strategy by **increasing visibility into data-in-transit within/between network segments** and endpoints that don't support EDR
- Supports ZTA by providing **network-level context** for access decisions

Leveraging Metadata to Better Protect, Detect, and Respond to Cyber Threats on Federal Civilian Networks

Generating and analyzing metadata about network traffic across the federal civilian enterprise will enhance real time situational awareness, help identify attacks that evade endpoint detection tools, and accelerate incident response, damage assessment and threat hunting. In addition, since metadata may be generated in a compact and standardized structured format, it can be searched rapidly and stored in a cost-effective way.

Problem: The Federal Government Needs Improved Cyber Situational Awareness

The Cybersecurity and Infrastructure Security Agency (CISA) lacks real time cyber situational awareness across its networks and, as a result, may be unable to rapidly detect and respond to cyber adversaries. Several factors contribute to this situation:

- CISA does not have sufficient ongoing access to real-time data that would allow it to rapidly assess the health of the federal civilian enterprise or correlate activity across agencies.
- CISA does not have access to consistent log records going back in time, which hinders threat hunting, incident response, and damage assessments. This limitation also prevents CISA from conducting comprehensive retrospective analysis based on the most up to date information.
- Available log records are likely incomplete because logs are collected only for those items that are pre-programed to be collected. In addition, without a technical strategy for determining when a network component is compromised (and may therefore fail to detect malicious activity or generate misleading logs), the government will be unable to trust the integrity of its logs as sophisticated adversaries routinely turn down or modify them as part of their tradecraft.

Proposal: Federal Pilot to Generate and Analyze Metadata about Network Traffic to Better Protect, Detect, and Respond to Cyber Threats

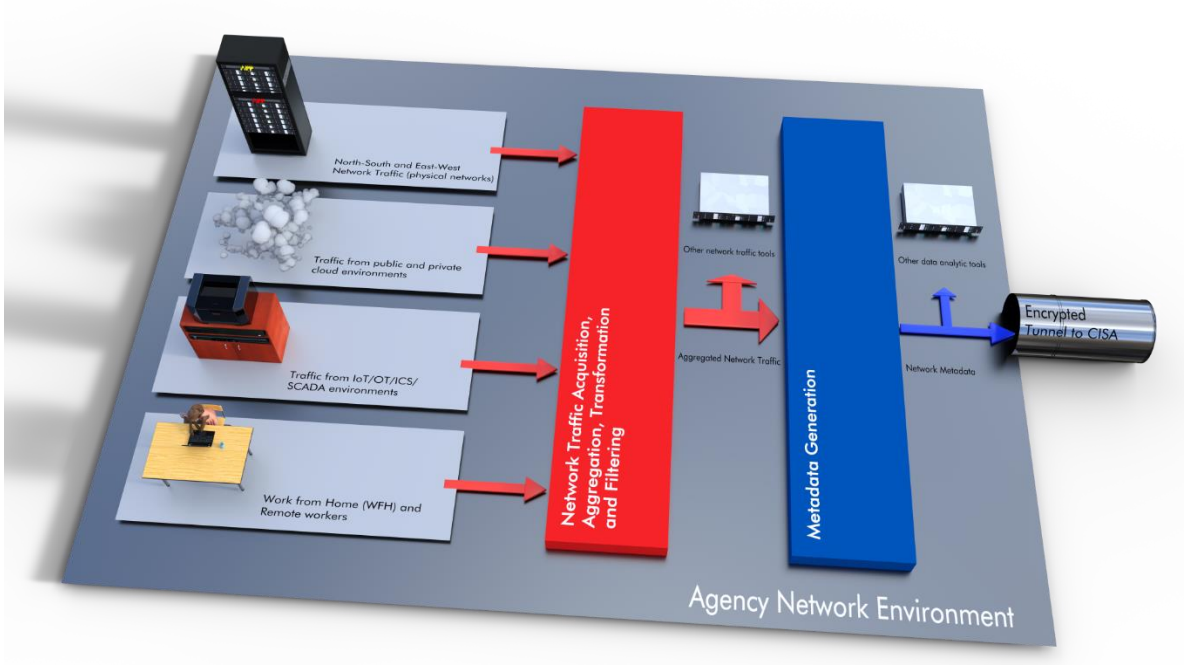
Using technology that is already fully or partially deployed across the federal government, CISA should conduct a pilot in coordination with at least two other federal agencies that:

- Collects all network traffic, including traffic transiting the enterprise, which is a function that is already performed by most departments and agencies today.
- Generates metadata about the network traffic and shares it with CISA for analysis,
- Analyzes the data using AI/ML and other tools as determined by CISA; and
- Utilizes the analysis for cybersecurity purposes.

This information will provide CISA with real time data about the security of federal civilian networks, an ability to rapidly correlate threat actor behavior across multiple agencies, and an ability to detect anomalous communications that evade endpoint-based detection.[1] In addition, the collection and storage of this metadata will supplement the logging requirements described in the Executive Order 14028 *on Improving the Nation's Cybersecurity* and OMB Memorandum M-21-*31 "Improving the Federal Government's*

[1] To mitigate risk in this environment, organizations are leveraging endpoint-based security controls. While this approach does mitigate some risk, it is insufficient. Endpoint Detection and Response (EDR) tools almost all "trust down" to the hardware and are not designed to validate or assure the security of the hardware they run upon. Endpoint tools may also operate in a silo and operate without the benefit of real-time information. EDR tools have a very siloed view of attacks and lack the organization-wide view that network-level visibility provides. In addition, some federal agencies may run devices that cannot support EDR capabilities or allow BYOD devices where running endpoint agents is difficult or impossible.

Investigative and Remediation Capabilities Related to Cybersecurity Incidents." Further, it will help mitigate vulnerability and risk associated with the government's reliance on logging to protect, detect, and respond to cyber threats.[2]



As illustrated in the proposal above, individual agencies collect all network traffic within their environments. This function is already performed by many agencies as part of their existing security and network performance initiatives.

---

[2] Any security strategy that relies upon logs for continuous monitoring and alerting, incident response, and damage assessments needs to mitigate risk associated with adversary exploitation of logging capabilities (e.g., undermining log integrity, reducing or slowing log generation and disabling log generation). The veracity of logging for cybersecurity outcomes depends upon a chain of assumptions which mature adversaries will exploit and break: (1) the log source is coded to produce sufficient log events to detect a threat; (2) the log source is configured to send all relevant log events; (3) the connectivity between log source and SIEM is authenticated and reliable, providing no loss or injection of logging data; and (4) the significance of the event, or sequence of events across the same or multiple hosts, is understood and coded into an alert in the SIEM.

OMB recognizes the risk in M-21-31. In appendix A on p7 under "protecting and validating log information," agencies are required to protect and monitor the integrity of their logs, by among other things, investigating the cessation of logging, and monitoring unexpected file changes. These measures would detect an intruder who has tried to disable or disrupt logging from a system but would not detect a sophisticated adversary that turns logging down instead of off or uses a supply chain exploit to compromise the host. This unaddressed risk is a gap in the security architecture that is slightly mitigated by OMB's requirement that agencies collect full packet capture data and store it for 72 hours. See M-21-31 page 39. While that short time frame likely represents a balance between the huge cost of storing raw network traffic and the security-driven need to do so, it is simply not long enough to assist with incident response. Using the SolarWinds intrusion as an example, based on the latest publicly available information, the initial intrusion into SolarWinds' network was in January 2019, the threat actors deployed their first test code between September 12 and November 4, 2019, and actively exploitation occurred from December 2019 to December 2020. Considering that timeline, storing full packet data for 3 days is completely insufficient.
Network metadata however provides session level visibility and is typically 1-5% the size of the raw network data (depending on the number of data elements included), so 60-300 days of network metadata could be archived in the same amount of storage as three days of full packet data. From a security perspective, network metadata cannot be disabled by an intruder: it is generated externally to the compromised node. A compromised node can only evade network metadata by not generating network traffic, which means it cannot communicate, and the act of network communication becomes evidence for the AI/ML/SIEM infrastructure to detect the presence of the threat. Even extremely sophisticated attacks, such as firmware and hardware implants seeded into the supply chain would be detected by network metadata because of the need to communicate.

This network traffic should include both north-south (i.e., traffic transiting to and from the network) and east-west traffic (traffic moving within the internal network), to detect attacker lateral movement, command and control communications, and insider threats. It should also include traffic in virtual and relevant cloud environments, enterprise user networks, data centers, service networks supporting IOT/OT, management networks and remote working activities from Virtual Private Network (VPN) concentrators.

The traffic will contain a large mixture of information, including some which may be sensitive or classified, that should be excluded from collection and analysis. This traffic would be filtered out pursuant to law and/or policy. Other functions can be performed, such as the identification of packets that have been captured multiple times, so the duplicate packets are discarded. The aggregated visibility into network traffic will be useful for security, operational and other requirements at the agency level.

In addition to supporting agency needs, the raw, aggregated network traffic would be fed to a metadata generation engine, which would re-sessionize[3] the data and convert it into a compact and structured industry standard format such as Common Event Format (CEF) or IPFIX.[4] As part of this process, CISA or the agency would determine what data to evaluate at a detailed level and what data to evaluate at the session level. The metadata feed may be consumed at the agency level by other security and performance tools. It will also be delivered to a consolidated data warehouse[5] at CISA via a unidirectional, encrypted tunnel. There would be an extensible database taxonomy to facilitate rapid analysis and support ingestion of additional sources of metadata.[6] This information flow is depicted in the figure below.

[3] For a definition of "session" in this context, see https://computersciencewiki.org/index.php/Sessions.

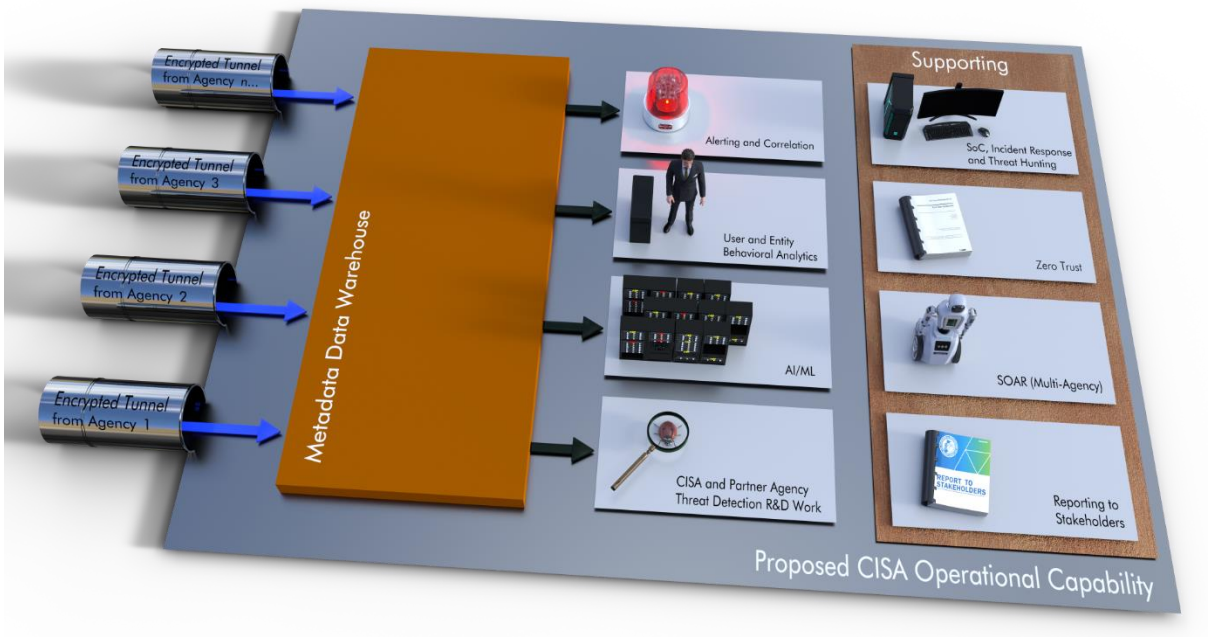
[4] Examples of metadata which can be collected include:

- Flow-based metadata (IPv4/IPv6 addressing, ports, duration, data volumes)
- DNS requests and responses
- HTTP requests and responses
- SSL/TLS connection metadata (TLS negotiation data, SNI for HTTPS, x509 metadata for TLS 1.2 and below, JA3/JA3S)
- RDP, SMTP, SSH, FTP, tunnel
- DHCP, NTLM and Kerberos requests and responses
- SMB and NFS file access requests and responses

This metadata excludes sensitive data. For example, an HTTP URL would be extracted but the query string, which often contains sensitive information, is not included in the metadata. In terms of threat hunting, it is the URL and associated flow-based metadata which is of high value in a cyber security investigation. Configured appropriately, metadata may enhance privacy. This metadata would not necessarily be stored in isolation but could be enriched with (1) mapping of source/destination from flow to a geolocation, AS (and associated metadata on that AS), domain information (and associated metadata especially domain name and entropy), and risk-based metadata on any of these enrichments and (2) additional threat feeds provided by the government.

[5] A data warehouse is a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. See https://www.gartner.com/en/information-technology/glossary/data-warehouse

[6] As all metadata fits within a defined taxonomy, ingestion is relatively straightforward, yet can accommodate agencies making different metadata generation choices driven by mission. It is likely, however, that the majority of incident response and threat hunting would be performed on a common subset of metadata.



Analysis of the metadata supports multiple uses cases and operational requirements, including incident response, threat hunting, automated response to threats, and reporting to relevant stakeholders (including compliance reports).

Alerting and correlation: The evaluation of multiple threat feeds and custom alerting rules will provide the government with the ability to identify and correlate threats in support of incident response activities. New indicators of compromise (IOCs), may be run through historical data to see whether the IOC existed prior to the creation of the IOC, supporting rapid retrospective analysis.

Entity Behavioral Analytics: The metadata may be used to evaluate users or endpoints, devices, business processes, individually or in aggregation, to identify behavior over time and in comparison, to other comparable devices. This capability will allow the government to detect deviations in behavior from peer devices, or over time, which indicates potential risk. This would support both incident response and threat hunting. Threats identified using this approach, once they are confirmed and characterized, would be fed into alerting and correlation tools to identify spread and to determine blast radius.

AI/ML: The metadata would feed AI/ML tools and enable them to better identify known and unknown threats, which can then be moved to incident response for investigation.

R&D around threat detection: The metadata would also support R&D to test innovative approaches to security, compliance measurement and threat detection. As the entire data flow is unidirectional from the agency, it would be possible for the environment to run at a highly classified level with cross-domain solutions deployed to support classified algorithms.

The presence of encrypted traffic does not negate the value of this proposal. Most traffic is encrypted using the TLS (Transport Layer Security) protocol (formerly called SSL), which includes significant quantities of plaintext session metadata that can be extracted and evaluated. Properly deployed, the network traffic acquisition platform shown in the diagram above can also perform break and inspect TLS decryption, providing full access to plaintext in a transparent and secure way.

The viability and utility of the approach has been validated at an exceptionally large scale in a UK government pilot that involved hundreds of gigabits of network ingest.