PROJECT DESCRIPTION

DATA CLASSIFICATION PRACTICES

Facilitating Data-Centric Security Management

Karen Scarfone

Scarfone Cybersecurity

Murugiah Souppaya

National Institute of Standards and Technology

DRAFT

May 2021

data-nccoe@nist.gov

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

NCCoE
NATIONAL CYBERSECURITY
CENTER OF EXCELLENCE

The National Cybersecurity Center of Excellence (NCCoE), a part of the National Institute of Standards and Technology (NIST), is a collaborative hub where industry organizations, government agencies, and academic institutions work together to address businesses' most pressing cybersecurity challenges. Through this collaboration, the NCCoE develops modular, adaptable example cybersecurity solutions demonstrating how to apply standards and best practices by using commercially available technology. To learn more about the NCCoE, visit https://www.nccoe.nist.gov/. To learn more about NIST, visit https://www.nist.gov/.

This document describes a challenge that is relevant to many industry sectors. NCCoE cybersecurity experts will address this challenge through collaboration with a Community of Interest, including vendors of cybersecurity solutions. The resulting reference design will detail an approach that can be incorporated across multiple sectors.

Abstract

As part of a zero trust approach, data-centric security management aims to enhance protection of information (data) regardless of where the data resides or who it is shared with. Data-centric security management necessarily depends on organizations knowing what data they have, what its characteristics are, and what security and privacy requirements it needs to meet so the necessary protections can be achieved. Standardized mechanisms for communicating data characteristics and protection requirements are needed to make data-centric security management feasible at scale. This project will examine such an approach based on defining and using data classifications. The project's objective is to develop technology-agnostic recommended practices for defining data classifications and data handling rulesets and for communicating them to others. This project will inform, and may identify opportunities to improve, existing cybersecurity and privacy risk management processes by helping with communicating data classifications and data handling rulesets. It will not replace current risk management practices, laws, regulations, or mandates. This project will result in a freely available NIST Cybersecurity Practice Guide.

Keywords

data-centric security management; data classification; data labeling; data protection; zero trust architecture; zero trust security

Acknowledgement

We appreciate the experts from JPMorgan Chase, Microsoft, Morgan Stanley, NATO, NIST, and Varonis who presented at the Data Classification workshop and contributed to the development of this project description.

Disclaimer

Certain commercial entities, equipment, products, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST or NCCoE, nor is it intended to imply that the entities, equipment, products, or materials are necessarily the best available for the purpose.

Comments on NCCoE Documents

Organizations are encouraged to review all draft publications during public comment periods and provide feedback. All publications from NIST's National Cybersecurity Center of Excellence are available at https://www.nccoe.nist.gov/.

Comments on this publication may be submitted to data-nccoe@nist.gov

Public comment period: May 19, 2021 to June 21, 2021

TABLE OF CONTENTS

1 Executive Summary

Purpose

A critical factor for achieving success in any business is the ability to share information and collaborate effectively and efficiently while satisfying the security and privacy requirements for protecting that information. Conventional network-centric security measures focus on protecting communications and information systems by providing perimeter-based security with multiple complex layers of security around users, hosts, applications, services, and endpoints. This model is increasingly ineffective for protecting information as systems become more dispersed, mobile, dynamic, and shared across different environments and subject to different types of stewardship.

As part of a zero trust approach [1], data-centric security management aims to enhance protection of information (data) regardless of where the data resides or who it is shared with. Data-centric security management necessarily depends on organizations knowing what data they have, what its characteristics are, and what security and privacy requirements it needs to meet so the necessary protections can be achieved. Standardized mechanisms for communicating data characteristics and protection requirements across systems and organizations are needed to make data-centric security management feasible at scale. The desired approach for this is to define and use data classifications, and this project will examine that approach.

This document defines a National Cybersecurity Center of Excellence (NCCoE) project on which we are seeking feedback. The project focuses on data classification in the context of data management and protection to support business use cases. The project's objective is to define technology-agnostic recommended practices for defining data classifications and data handling rulesets, and communicating them to others. Organizations will also be able to use the recommended practices to inventory and characterize data for other security management purposes, such as preparing for and prioritizing transitions to post-quantum cryptographic algorithms.

This project will focus on communicating and safeguarding data protection requirements through data classifications and labels. Cybersecurity and privacy risk management processes and other sources of data protection requirements are out of scope, as are mechanisms for enforcing data protection requirements. This project will inform, and may identify opportunities to improve, existing risk management processes by helping with communicating data classifications and data handling rulesets. It will not replace current risk management practices, laws, regulations, or mandates.

This project will result in a publicly available NIST Cybersecurity Practice Guide, a detailed implementation guide of the practical steps needed to implement a cybersecurity reference design that addresses this challenge.

Scope

This project will take a layered and modular approach to enable sharing and collaboration within and across organization boundaries. The project will emphasize an evolutionary path through a set of data classification maturity levels that are designed to be adopted at any organizational level (e.g., department, division, or organization) and within/across any geographic locations.

The first phase of this project will define the approach for the solution, independent of the supporting technologies, services, architectures, operational environments, etc. As part of this, a simple proof-of-concept approach implementation of the approach will be attempted. The proof-of-concept will include limited data discovery, analysis, classification, and labeling capabilities, as well as a rudimentary method for expressing how data with a particular label should be handled for each use case scenario. In support of this phase of the project, basic terminology and concepts will be defined based on existing practices and guidance to provide a common language for discussing data classification.

The subsequent phases of the project will build on the first phase by addressing standards, technologies, processes, and recommended practices for discovering and classifying data, and communicating the data classification so the data is properly protected and controlled. This information will span devices and application workloads across on-premises, hybrid, and cloud environments throughout the full data lifecycle. These subsequent phases would primarily focus on the following areas:

- Deployment of additional solutions for information discovery, classification, and labeling, including requirements for secure persistence and binding to content, interoperability, and lifecycle management aligned to the information lifecycle
- Additional labels that address aspects such as provenance and lineage, classification/sensitivity, and releasability, and appropriate mechanisms to define policies and perform lifecycle management aligned to the information lifecycle and sharing. This will cover both regulatory and business policies related to privacy and security. These policies will be driven by the use case scenarios.
- Identification of appropriate controls as recommended in existing cybersecurity and privacy risk management frameworks to manage, monitor, enforce, and demonstrate compliance with the defined classifications for effective, dynamic security and privacy risk management supported by auditing throughout the information lifecycle
- Technologies and industry standards for specifying and implementing classification labels, data handling rulesets, and the corresponding controls such as access control, rights management, and cryptographic protection
- Recommended practices for end-user awareness and training, response to non-compliance or a cybersecurity incident, and continuous improvement of classifications, data handling rulesets, and controls

Assumptions/Challenges

Readers are assumed to understand risk management processes and basic data protection and zero trust concepts.

Background

Data classification and labeling are becoming much more common needs. In the early days of digital computing, data classification was largely associated with the armed forces and defense industry. Classification terms such as TOP SECRET, while well known to the public due to media portrayals, were nearly completely absent outside of certain government and military environments.

A number of forces have come to bear on all organizations that have catapulted data classification and labeling to the forefront and resulted in a sense of urgency regarding establishment of models for use with all data. Laws and regulations such as the California

Consumer Privacy Act (CCPA), Children's Online Privacy Protection Act (COPPA), Fair Credit Reporting Act (FCRA)/Fair and Accurate Credit Transactions Act (FACTA), Family Educational Rights and Privacy Act (FERPA), General Data Protection Regulation (GDPR), Gramm Leach Bliley Act (GLBA), Health Information Portability and Accountability Act (HIPAA), and Payment Card Industry Data Security Standard (PCI DSS) mandate that data containing certain types of information be handled with specific safeguards. As new laws and regulations emerge and as existing ones are augmented, much of the data an organization already has may need to be classified or handled differently.

Organizations are dealing simultaneously with rapid growth in the sheer volume of data stored and in the requirements for protecting and controlling that data, including longer data retention periods. This can be expected to result in larger capital and operational expenditures. Thus, the ability to communicate data classifications and data handling rulesets improves the efficiency of resource expenditure and allocation since the controls used can correlate with the assigned data classification. There is also a need to break down the data silos and enable data sharing across organizational boundaries to support business objectives while still satisfying security, privacy, and regulatory compliance requirements. This need likely varies from sector to sector.

Existing NIST standards and guidance regarding data classification and labeling, such as Federal Information Processing Standard (FIPS) 199 [2] and NIST Special Publication (SP) 800-60 [3], address federal government-specific requirements, but not the many other requirements to which federal agencies and other organizations are subject.

More generally, significant challenges that have hindered effective use of data classification approaches include the following:

- The limited nature of existing standards for data classifications outside of the government and military means that most organizations do not use classifications that are consistent with those of their partners and suppliers. Organizations perform countless transactions with others for which data classification and protection are relevant, and the lack of industry standards impairs organizations' ability to enforce data handling requirements.
- The lack of common definitions for and understanding of classifiers can result in information being classified and labeled inconsistently. Reliance on end users to identify and classify the data they create and receive is particularly error-prone and incomplete.
- Data is everywhere: on devices (e.g., laptops, desktops, mobile devices), in applications running in both on-premises and outsourced environments, and in the cloud. This distributed nature of data complicates the process of establishing and maintaining data inventories.
- Data classifications and data handling requirements often change during the data lifecycle, for example safeguarding the confidentiality of data at first, then subsequently releasing that data to the public. Another example is data being safeguarded and retained for a certain period of time, then being destroyed to prevent further access. This is further complicated with the advancement in quantum computing technology, which introduces a threat to data being protected by current public key algorithms.

This project is intended to address these challenges and to enable organizations of any size and complexity to launch and maintain a solution for defining and communicating data classifications, labels, and data handling rulesets. This project is also intended to inform future updates to FIPS 199, NIST SP 800-60, and other NIST publications.

2 Scenarios

The use case scenarios we are considering for the first phase of the project are as follows:

Scenario 1: Financial sector

This scenario involves a large regulated financial sector organization that is required by regulations and laws to protect its customers' personal phone numbers from unauthorized access and changes. The organization also provides its customer information to certain business partners (e.g., sharing data within contracts) and requires those partners to protect the phone numbers on the organization's behalf. Those partners are located in several jurisdictions.

Scenario 2: Government sector

This scenario involves federation of government agencies from several countries and international and non-governmental organizations that need to collaborate with each other and share information. Supported use cases include writing and editing reports, holding web conferences to discuss the work as a group and to share materials with each other, exchanging emails and chat messages, and sending application-specific data among automated systems. The level of trust between different partners can vary significantly, and there are several independent governing authorities in the federation.

Scenario 3: Manufacturing sector

This scenario involves a small manufacturing company. The manufacturer has trade secrets that it needs only certain employees, contractors, and business partners to be able to access.

Scenario 4: Technology sector

This scenario involves a small technology company that is giving up its office lease and transitioning to 100% work-from-anywhere. As the company makes this transition, it will also be adopting zero trust architecture principles. The focus of this scenario is the integrity of the source code for a particular product. This code is stored in the company's cloud-based code repository.

Scenario 5: Healthcare sector

This scenario involves a small healthcare provider that needs to share protected health information (PHI) with other healthcare providers as authorized by the patient. The healthcare provider also needs to ensure that it retains all PHI for the required period of time, and that it destroys PHI once it no longer needs to be retained.

For each scenario, we will do the following:

1. Document a notional architecture that
   a. indicates people, systems, applications and services, and end user devices directly involved in or affected by data classification activities. These will be representative for the scenario, not comprehensive.
   b. denotes data lifecycle activities such as data creation/capture, processing, storage, transmission/transport/sharing, retention, and destruction. These activities will be representative for the scenario, not comprehensive.

   c. highlights how data classification is foundational for mitigating concerns around protecting data, such as data leakage, in a world where data is distributed across applications hosted in numerous places, processed on many devices, and accessed by different sets of users anytime and from anywhere.
   d. does not necessarily include the implementation of security controls for enforcing data or for system protection. The intent of the scenarios and architectures is to explore challenges specific to classifying data and expressing those classifications, rather than on how expressed classifications may be translated by individual organizations into implemented security controls.
2. Define data classifications that will apply to the sets of data specified in the scenario. The classifications must take into account applicable regulations, laws, and organizational policies.
3. Create a data handling ruleset to specify enforcement requirements for the data in the scenario based on its data classifications. This data handling ruleset must be fully compatible with the data classifications, to include enforcing data protection requirements, secure data sharing requirements, data retention requirements, etc.
4. Implement the notional architecture in the NCCoE lab and cloud environment.
5. Communicate the necessary information (data classifications, data handling rulesets, etc.) to the necessary individuals, systems, and organizations within the implementation in the deployed environment.

3 High-Level Architecture

Component List

The high-level architecture will include, but is not limited to, the following components:

- **Endpoints**:
  - **Client Devices**: Various PCs (desktops or laptops) and mobile devices will be involved in data creation, storage, transmission, retention, and destruction, as well as data-centric security management. Some client devices will be managed by the organization. Some will be used by the organization's employees, while others will be used by people from other organizations.
  - **Client Device Apps**: The client devices will have commercial-off-the-shelf (COTS) apps used for data lifecycle activities, such as word processing software and email client software.
  - **Additional Devices:** Examples of additional types of devices that could be utilized are networked printers and Internet of Things (IoT) devices.
- **Network/Infrastructure Devices** – The architecture will include devices such as firewalls, routers, or switches that are needed for network functionality and network traffic restriction, as well as the software for managing those devices.
- **Services and Applications** – The architecture will include several types of services and applications that are involved in data lifecycle activities for one or more of the scenarios. The following are examples of possible service and application types:
  - **Enterprise Services/Applications**: Email, collaboration, file sharing, web conferencing, file/data backup, code repositories, content management systems

    - **Data Services/Applications**: Data processing, data analytics, artificial intelligence/machine learning services
    - **Business Services/Applications**: A variety of system-to-system and human-to-system business applications, both COTS and custom-written, including those that produce and/or consume data
- **Data Classification Solutions** – The architecture will include several types of components used to perform data classification responsibilities, such as data discovery, inventory, analysis, classification, and labeling.

Desired Security Capabilities

This project seeks to develop a reference design and implementation using commercially available technology that meets the following characteristics:

- All data is discovered and analyzed to determine how it should be classified.
- All data classification and data handling ruleset creation, modification, and deletion is restricted to authorized personnel only, with all actions logged and auditable and with all communications protected.
- For all data classifications and data handling rulesets, there is a mechanism for verifying the integrity of the policy or ruleset.
- Data classification labels or tags are assigned to all data.
- For all data classification labels or tags assigned to data, there is a mechanism for verifying the integrity of the label or tag.

4 Relevant Standards and Guidance

The following resources and references provide additional information to be leveraged to develop this solution:

- National Institute of Standards and Technology (NIST), *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*, April 2018 https://doi.org/10.6028/NIST.CSWP.04162018
- NIST Federal Information Processing Standard (FIPS) 199, *Standards for Security Categorization of Federal Information and Information Systems*, February 2004 https://doi.org/10.6028/NIST.FIPS.199
- NIST Internal Report (IR) 8112, *Attribute Metadata: A Proposed Schema for Evaluating Federated Attributes*, January 2018 https://doi.org/10.6028/NIST.IR.8112
- *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0*, January 2020 https://doi.org/10.6028/NIST.CSWP.01162020
- NIST Special Publication (SP) 800-53 Rev. 5, *Security and Privacy Controls for Information Systems and Organizations*, September 2020 https://doi.org/10.6028/NIST.SP.800-53r5
- NIST SP 800-60 Vol. 1 Rev. 1, *Guide for Mapping Types of Information and Information Systems to Security Categories*, August 2008 https://doi.org/10.6028/NIST.SP.800-60v1r1

- NIST SP 800-154 (Draft), *Guide to Data-Centric System Threat Modeling*, March 2016 https://csrc.nist.gov/CSRC/media/Publications/sp/800-154/draft/documents/sp800_154_draft.pdf
- NIST SP 800-171 Rev. 2, *Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations*, February 2020 https://doi.org/10.6028/NIST.SP.800-171r2
- NIST SP 800-207, *Zero Trust Architecture*, August 2020 https://doi.org/10.6028/NIST.SP.800-207

Appendix A References

[1] National Institute of Standards and Technology (NIST), NIST Special Publication (SP) 800-207, *Zero Trust Architecture*, August 2020 https://doi.org/10.6028/NIST.SP.800-207

[2] National Institute of Standards and Technology (NIST), NIST Federal Information Processing Standard (FIPS) 199, *Standards for Security Categorization of Federal Information and Information Systems*, February 2004 https://doi.org/10.6028/NIST.FIPS.199

[3] National Institute of Standards and Technology (NIST), NIST Special Publication (SP) 800-60 Vol. 1 Rev. 1, *Guide for Mapping Types of Information and Information Systems to Security Categories*, August 2008 https://doi.org/10.6028/NIST.SP.800-60v1r1

Appendix B Acronyms and Abbreviations

CCPA	California Consumer Privacy Act
COPPA	Children’s Online Privacy Protection Act
COTS	Commercial-Off-the-Shelf
FACTA	Fair and Accurate Credit Transactions Act
FCRA	Fair Credit Reporting Act
FERPA	Family Educational Rights and Privacy Act
FIPS	Federal Information Processing Standard
GDPR	General Data Protection Regulation
GLBA	Gramm Leach Bliley Act
HIPAA	Health Information Portability and Accountability Act
IoT	Internet of Things
IR	Internal Report
NCCoE	National Cybersecurity Center of Excellence
NIST	National Institute of Standards and Technology
PC	Personal Computer
PCI DSS	Payment Card Industry Data Security Standard
PHI	Protected Health Information
SP	Special Publication