NIST SPECIAL PUBLICATION 1800-39

Data Classification Practices

William Newhouse
Murugiah Souppaya*
National Institute of Standards and Technology
Gaithersburg, Maryland

John Kent
Kenneth Sandlin*
Ryan Williams*
The MITRE Corporation
McLean, Virginia

Karen Kent
Trusted Cyber Annex

Mark Evans
Jimmy Katz*
ActiveNav

John Dombroski
Harmeet Singh
IBM

Neville Jones
Helen Farrell*
Janusnet

Pablo Blasco
Jane Gilbert
D'Nan Godfrey
Matt Jochim
Rich Johnson
Ludmila Rinaudo
Gina Scinta
Thales Trusted Cyber Technologies

Patrick Greer
Wilson Patton
Jason White*
Trellix

*Former employee; all work for this publication was done while at listed organization.

February 2026

INITIAL PUBLIC DRAFT

This publication is available free of charge from

https://www.nccoe.nist.gov/data-classification

DISCLAIMER

Certain commercial entities, equipment, products, or materials may be identified by name or company logo or other insignia in order to acknowledge their participation in this collaboration or to describe an experimental procedure or concept adequately. Such identification is not intended to imply special status or relationship with NIST or recommendation or endorsement by NIST or NCCoE; neither is it intended to imply that the entities, equipment, products, or materials are necessarily the best available for the purpose.

While NIST and the NCCoE address goals of improving management of cybersecurity and privacy risk through outreach and application of standards and best practices, it is the stakeholder's responsibility to fully perform a risk assessment to include the current threat, vulnerabilities, likelihood of a compromise, and the impact should the threat be realized before adopting cybersecurity measures such as this recommendation.

National Institute of Standards and Technology Special Publication 1800-39, Natl. Inst. Stand. Technol. Spec. Publ. 1800-39 60 pages, (February 2026), CODEN: NSPUE2

Author ORCID iDs

William Newhouse: 0000-0002-4873-7648

Murugiah Souppaya: 0000-0002-8055-8527

John Kent: 0009-0001-9989-2277

Ken Sandlin: 0009-0000-3757-4858

Karen Kent: 0000-0001-6334-9486

FEEDBACK

You can view or download the guide at the NCCoE Data Classification Practices project page.

With this initial public draft, NIST NCCoE is now asking for feedback on this initial public draft.

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

Public comment period: February 12, 2026 - March 30, 2026

All comments are subject to release under the Freedom of Information Act.

NIST NCCoE is particularly interested in your feedback on the following questions:

1. Does this guide influence your plans to use of data classification practices?
2. How well do the data classification practices in this guide relate to your organization's practices or needs?
3. Are there significant gaps this guide should address?
4. What would you like to see added to this guide?
5. Do you have suggestions for future NCCoE data centric cybersecurity or privacy projects?
6. What could the NCCoE build as an example demonstration using commercially available technologies that would help you take action to mitigate some aspects of your cybersecurity and privacy risks?

National Cybersecurity Center of Excellence
National Institute of Standards and Technology
100 Bureau Drive
Mailstop 2002
Gaithersburg, MD 20899
Email: nccoe@nist.gov

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 issues. This public-private partnership enables the creation of practical cybersecurity solutions for specific industries, as well as for broad, cross-sector technology challenges. Through consortia under Cooperative Research and Development Agreements (CRADAs), including technology partners—from Fortune 50 market leaders to smaller companies specializing in information technology security—the NCCoE applies standards and best practices to develop modular, adaptable example cybersecurity solutions using commercially available technology. The NCCoE documents these example solutions in the NIST Special Publication 1800 series, which maps capabilities to the NIST Cybersecurity Framework and details the steps needed for another entity to re-create the example solution. The NCCoE was established in 2012 by NIST in partnership with the State of Maryland and Montgomery County, Maryland.

To learn more about the NCCoE, visit https://www.nccoe.nist.gov/. To learn more about NIST, visit https://www.nist.gov.

NIST CYBERSECURITY PRACTICE GUIDES

NIST Cybersecurity Practice Guides (Special Publication 1800 series) target specific cybersecurity challenges in the public and private sectors. They are practical, user-friendly guides that facilitate the adoption of standards-based approaches to cybersecurity. They show members of the information security community how to implement example solutions that help them align with relevant standards and best practices, and provide users with the materials lists, configuration files, and other information they need to implement a similar approach.

The documents in this series describe example implementations of cybersecurity practices that businesses and other organizations may voluntarily adopt. These documents do not describe regulations or mandatory practices, nor do they carry statutory authority.

ABSTRACT

This guide demonstrates how organizations can discover, identify and label unstructured data using data classification practices. Performing Data Classification Practices allows an organization to know its data and apply technologies that minimize the risk of valuable or sensitive data being lost or mismanaged. Data Classification Practices prepare an organization for the use of emerging security measures—including Zero Trust Architecture, quantum-safe cryptography, and AI model training that requires labeled data. This 1800-series NIST publication documents how the NCCoE and its collaborators created a synthetic dataset and used commercially available data classification tools to discover, identify and label unstructured data.

KEYWORDS

Data classification practices; synthetic unstructured data; data pillar of zero trust; data types, data labels.

ACKNOWLEDGMENTS

We are grateful to the following individuals from our project's technology collaborators for their generous contributions of expertise and time.

- Adobe: Steve Gottwals
- GitLab: Joel Krooswyk
- JPMorgan Chase & Co.: Timothy Brophy*, Lauren Brown, Benjamin Flatgard
- MITRE: Jason Ajmo, Lauren Swan, Spike Dog*
- NIST: Cherilyn Pascoe, Julie Chua
- Seqrite/Quick Heal: Dhruvi Desai, Nachiket Karguppikar, Vinaya Sathyanarayana*
- Virtru: Will Ackerly, Cassandra Zimmerman

* Former employee; all work for this publication was done while at listed organization.

We are also grateful to the experts from JPMorgan Chase, Microsoft, Morgan Stanley, NATO, NIST, and Varonis who presented at an October 2019 NCCoE hosted Data Classification workshop. Their contributions led to the establishment of the NCCoE Data Classification project.

We wish to thank Bill Brunt for his interest in our project and his willingness to share his data classification practices expertise at the start of our project.

The authors are grateful to all who reviewed and provided feedback on this publication.

The Technology Collaborators who participated in this build submitted their capabilities in response to a notice in the Federal Register. Respondents with relevant capabilities or product components were invited to sign a Cooperative Research and Development Agreement (CRADA) with NIST, enabling their participation in a consortium that built this publication's example demonstrations.

DOCUMENT CONVENTIONS

The terms “shall” and “shall not” indicate requirements to be followed strictly to conform to the publication and from which no deviation is permitted. The terms “should” and “should not” indicate that among several possibilities, one is recommended as particularly suitable without mentioning or excluding others, or that a certain course of action is preferred but not necessarily required, or that (in the negative form) a certain possibility or course of action is discouraged but not prohibited. The terms “may” and “need not” indicate a course of action permissible within the limits of the publication. The terms “can” and “cannot” indicate a possibility and capability, whether material, physical, or causal.

CALL FOR PATENT CLAIMS

This public review includes a call for information on essential patent claims (claims whose use would be required for compliance with the guidance or requirements in this Information Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may be directly stated in this ITL Publication or by reference to another publication. This call also includes disclosure, where known, of the existence of pending U.S. or foreign patent applications relating to this ITL draft publication and of any relevant unexpired U.S. or foreign patents.

ITL may require from the patent holder, or a party authorized to make assurances on its behalf, in written or electronic form, either:

a) assurance in the form of a general disclaimer to the effect that such party does not hold and does not currently intend holding any essential patent claim(s); or

b) assurance that a license to such essential patent claim(s) will be made available to applicants desiring to utilize the license for the purpose of complying with the guidance or requirements in this ITL draft publication either:

1. under reasonable terms and conditions that are demonstrably free of any unfair discrimination; or
2. without compensation and under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

Such assurance shall indicate that the patent holder (or third party authorized to make assurances on its behalf) will include in any documents transferring ownership of patents subject to the assurance, provisions sufficient to ensure that the commitments in the assurance are binding on the transferee, and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding each successor-in-interest.

The assurance shall also indicate that it is intended to be binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

Such statements should be addressed to: data-nccoe@nist.gov.

Contents

List of Figures

List of Tables

1 Overview

This publication shows commercially available products performing unstructured data classification practices.

1.1 Challenge

The project's goal is to help organizations learn how to understand its unstructured data and prepare its unstructured data to leverage an organization's security and privacy controls.

An organization's data landscape may seem daunting and uncharted due to the sheer volume and diversity of types, formats, locations, and use cases. It can encompass personally identifiable information (PII) stored in databases and files on endpoints, digital conversations saved in cloud environments, structured and unstructured data in data lakes and file stores, and more.

Since data is so vast and ubiquitous, organizations need a shared understanding of what data assets are to identify and protect them.

The project aims at improving an organization's shared understanding of its data assets as they work to meet regulatory data policies by demonstrating the use data classification tools on unstructured data.

The NCCoE issued an open invitation using a Federal Register Notice (FRN) [1] to organizations interested in bringing their products and technical expertise to support and demonstrate Data Classification Practices.

1.2 Audience

The audience of this practice guide is technical staff and leadership at medium and large organizations who work with, manage or protect data. Specific audience members within these organizations may include:

- Chief Data Officers (CDOs) focused on digital transformation, information sharing with partners, and safeguarding data
- Chief Information Security Officers (CISOs)
- Information security professionals focused on the data lifecycle
- Data management professionals
- Zero Trust Architecture (ZTA) implementers and operators
- Personnel responsible for recognizing that their organization may have information in unstructured data stores that needs to be protected by cryptography
- Data scientists and owners who are discovering and labeling unstructured data for training artificial intelligence systems.
- Regulatory Compliance- support audits, meet data regulation requirements and retention policy who may benefit from understanding data classification tools and are part of an organization's

- Data Protection Officers whose understanding of data classification tools can support their work to understand issues which relate to the protection of personal data as part of business risk and liability programs.
- Cost efficiency experts who wish to safeguard and protect data that are critical and delete data not needed to save storage
- Strategic Planners who seek to understand where data can reside like on-prem vs cloud and if data can be used to support development of AI models or other innovative use cases

1.3 Scope

The scope of this publication is data classification practices, and this publication specifically demonstrates discovery, identification, and labeling of unstructured data via schemas which are frameworks or structures for organizing and categorizing data into different classes or categories based on sensitivity, type, or business function.

The project does not:

- define regulatory and classification policy for specific industry sectors like USG, healthcare, financial services, etc.
- demonstrate a full data management lifecycle
- perform validation of defined policies or defined schema
- demonstrate data protection practices on labelled data post data classification practices
- focus on error handling or data checking during the labeling process

Synthetic data was created to allow the data classification tools to work on unstructured data that was free from privacy and security restrictions. The data classification tools used in this project were installed and configured to look within a known network file location hosting the synthetic data.

1.4 Structure of This Guide

This NIST Cybersecurity Practice Guide provides users with the information they need to replicate the discovery, identification, and labeling of data using commercially available data classification technology. This guide is separated into sections, and is organized as follows:

Section 2 provides a project overview, including the motivation for the project, the challenges in implementing data classification practices, the project approach and the project's collaborators.

Section 3 describes the synthetic data used.

Section 4 describes the demonstrations using steps in a workflow for each of the data classification tools provided by our project collaborators.

Section 5 documents project findings and insights.

Section 6 maps the tools that were used in the project to Cybersecurity Framework 2.0 subcategories.

2 Project Overview

2.1 Motivation for the Project

The objective of the project is to demonstrate how organizations discover, identify and label their unstructured data which is a crucial component of a data management program. Data classification practices enable an organization to better understand and prepare its data to be managed by data protection policies that meet privacy and security regulations, to leverage Zero Trust Architectures, to prioritize what should be protected by post-quantum cryptography, and to support future use of Artificial Intelligence systems.

2.2 Challenges in Implementing Data Classification Practices for Unstructured Data

Organizations may not know the locations of all their data, and they may not understand the sensitivity or value of their unstructured data.

When organizations store and share their data internally and share it with other organizations, ensuring the protection of the data can be challenging. Organizations that are classifying data by discovering and labeling it are improving their security by understanding where their data is located, what the data's sensitivity is, and what its protection needs will be.

Organizations that share data with other organizations lose control of that shared data once it leaves their data protection boundaries. Data classification practices among data sharing organizations that negotiate and use consistent labelling and tagging make it more likely that data protections can be maintained at the receiving organization.

This publication can help organizations explore data classification practices by demonstrating:

- the creation of unstructured data files from synthetic data
- the use of schemas to create a logical data model that allows an organization to label its unstructured data
- the configuration of data classification tools to discover unstructured data that contains sensitive information to be identified and labelled
- the application of schemas to identify sensitive information contained within an email or unstructured data using schemas
- the application of labelling of sensitive information
- the identifying methods to apply data classification practices to unstructured email data

2.2.1 Synthetic Unstructured Data

The project used synthetic data that contains realistic but fictional information. Using synthetic data created from a synthetic population that statistically mirrors a real population establishes an environment for experimenting with large-scale data. Using synthetic data also yields meaningful results that are free from privacy and security restrictions.

. Identifying a source of synthetic, realistic, but not real data was an initial challenge. Creating file types that contained the "not real data" for the data classification demonstrations was also necessary for the project. Section 3 describes the synthetic data source and using that data in over 10,000 unique files.

2.3 Project Approach

In October 2018, the NCCoE hosted an Information Protection and Data-Centric Security Management: Data Classification Workshop [2]. The purpose of workshop was to discuss the challenges and opportunities with data classification in the context of data management and information protection to support various business use cases. The outcome of the workshop was used to develop goals of this demonstration project. In October 2021, NIST posted a Federal Register Notice (FRN) [1] to solicit responses from interested organizations to enter into a Cooperative Research and Development Agreement (CRADA) to provide products and technical expertise to support and demonstrate security platforms for the *Data Classification Practices: Facilitating Data-Centric Security Management* project. Cooperative Research and Development Agreements (CRADAs) were established with selected respondents to form our CRADA consortium for the project.

The consortium came to consensus to demonstrate unstructured data classification practices using synthetic data. The project's demonstration scenarios show how data classification techniques can provide the ability to classify data using discovery, identification, and labeling techniques. Data classification practices were performed on unstructured data located in a network file share location and in emails.

2.3.1 Project Assumptions

This project is guided by the assumption that organizations would benefit from performing data classification practices in a consistent and automated way that works well with the existing tools because:

- managing and protecting data at scale can be challenging, and data classification practices can help alleviate this challenge
- sharing unstructured data among two or more organizations requires coordination to identify data asset types to use to describe the data that is being shared. This coordination may result in an agreed upon schema, a framework or structure for organizing and categorizing data into different classes or categories.
- the use of an agreed upon schema between organizations will improve the likelihood that data security and privacy requirements can be met within each organization even when they are not using the same data classification tools.
- unstructured data is in common use. Research [3] shows that 80-90% of organizational data is unstructured and a good focus area for data classification practices.
    - Performing data classification practices on unstructured data can support an organization's ability to meet its privacy and data retention policies.

2.3.1.1 Related Emerging Technologies

This project is guided by the assumption that organizations should leverage data classification practices to support:

- preparing data for use by artificial intelligence systems
- adoption of zero trust architectures and any security architectures which focus on securing and enforcing access to data at rest and in transit through various methods, including encryption, tagging and labeling, data loss prevention (DLP) strategies, and application of data rights management (DRM) tools.
- prioritization for migration to post-quantum cryptography

3 Synthetic Data

To enable demonstrations of data classification practices, synthetic unstructured data was created.

3.1 Synthetic Data Characteristics

The project collaborators came to consensus on use of synthetic data that has the following characteristics:

- contains realistic but fictional information
- created from a synthetic population that statistically mirrors a real population
- de-identified so that it would not expose any real persons' sensitive information.
- could not be linked back to a real person's name thus avoiding privacy information disclosure concerns.
- includes a range of unstructured data objects that had a known provenance or source of origination
- enables standardized metadata and content
- can be used to create more than 10,000 files containing synthetic data types

3.2 Synthetic Data Source

SyntheticMass [4], a synthetic patient and population health data website for the state of Massachusetts, was chosen as the source for creating this project's synthetic unstructured data.

3.3 Unstructured Synthetic Data Files

A .csv file was downloaded from SyntheticMass and used to create 25,884 files containing synthetic unstructured data containing data types such as fictitious names, addresses, and patient identifiers. Thirteen file types were chosen as representative of the kinds of files one might find in a file folder. While scripted processes were used for creating most of the files, manual use of mail merge was also used to create a small number of Word files. Visit Appendix D Synthetic Data Creation to see the process steps for creating a document formatted synthetic dataset using this source.

Both reactive (containing sensitive data examples) and non-reactive (not containing sensitive data examples) synthetic data is included in the files to enable the data classification tools to demonstrate the ability to distinguish between sensitive and non-sensitive data types.

Table 3-1 File Types Contained in the Synthetic Dataset

File Types	File Types
.doc (document)	.rtf (document)
.docx (document)	.csv (spreadsheet)
.eml (email)	.txt (document)
.html (web)	.wav (audio)
.pdf (document)	.xls (spreadsheet)
.png (image)	.xlsx (spreadsheet)
.pptx (presentation)	.zip (compact file)

Synthetic data types that cover PII, Financial and Health related information were selected, as those are the dominant types found within most organizations.

Table 3-2 Twelve Data Types Contained in the Synthetic Dataset

Data Types	Data Types
Patient ID	Zip
Name	Birthdate
Address	License #
City	Passport #
County	NCCoE Customer #
State	NCCoE Billing #

Table 3-2 shows the data types used in the 25,884 files created for the data classification practice demonstrations. The 12 data types used in the demonstrations include Patient ID, Name, Address, City, County, State, Zip, Birthdate, License #, Passport #, NCCoE Customer #, and NCCoE Billing #. The created files may contain 0-12 of these data types. The NCCoE Customer Billing # data types were added to add financially sensitive data for the data classification tools to discover. Figure 3-1 shows a specific example for each data element used in the demonstrations.

374 **Figure 3-1 Specific Example of Synthetic Unstructured Data for Each Data Element**

4 Demonstrations

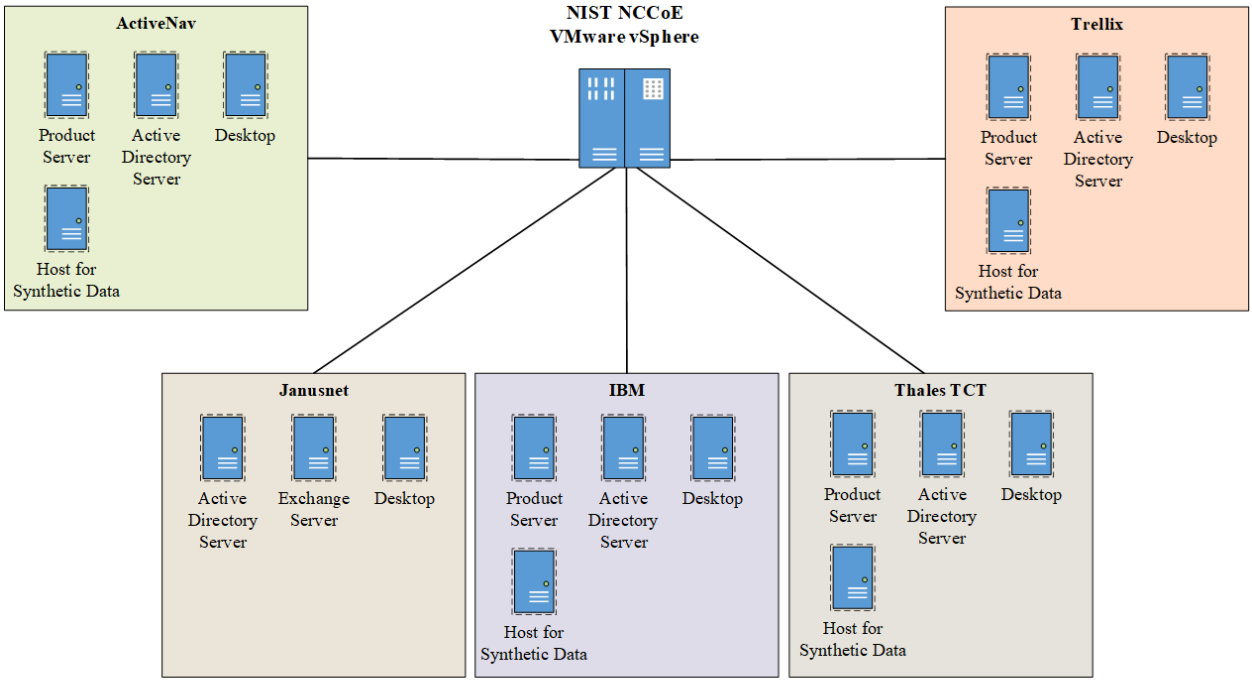
The project's demonstrations show unstructured data being discovered, identified, and labeled using the collaborator tools for potentially sensitive unstructured data files placed in a network file location and unstructured data contained in electronic mail messages.

4.1 Unstructured Data Classification Practice Demonstrations

For the unstructured data classification practice demonstrations, the same 25,884 synthetic files, were placed in a network file folder on host in each of the NCCoE's VMware vSphere virtualization operating environments built for each data classification tool. Each product, a data classification tool, was installed on a server in the same network segment as the host with the synthetic unstructured data. Each data classification tool was then run to scan to discover the synthetic files in their network file location with analysis done to identify the sensitive information contained in the files then apply a schema that labels the data types (Patient ID, Name, Address, City, County, State, Zip, Birthdate, License #, Passport #, NCCoE Customer #, and NCCoE Billing #) in each file.

4.2 Lab Demonstration Environment

A separate VLAN was created for each demonstration as shown in Figure 4-1 High Level Overview of Lab Environment for Each Data Classification Tool. Data Classification Tools were installed on a product server. A host was created with a file storage area where the synthetic unstructured data files were located. An Active Directory Server was used to provide authentication services for the Data Classification Tools product servers and synthetic unstructured data file location servers. A virtual desktop computer was used in each demonstration lab to configure products and provide email client services for the Janusnet demonstration.



396 **Figure 4-1 High Level Overview of Lab Environment for Each Data Classification Tool**

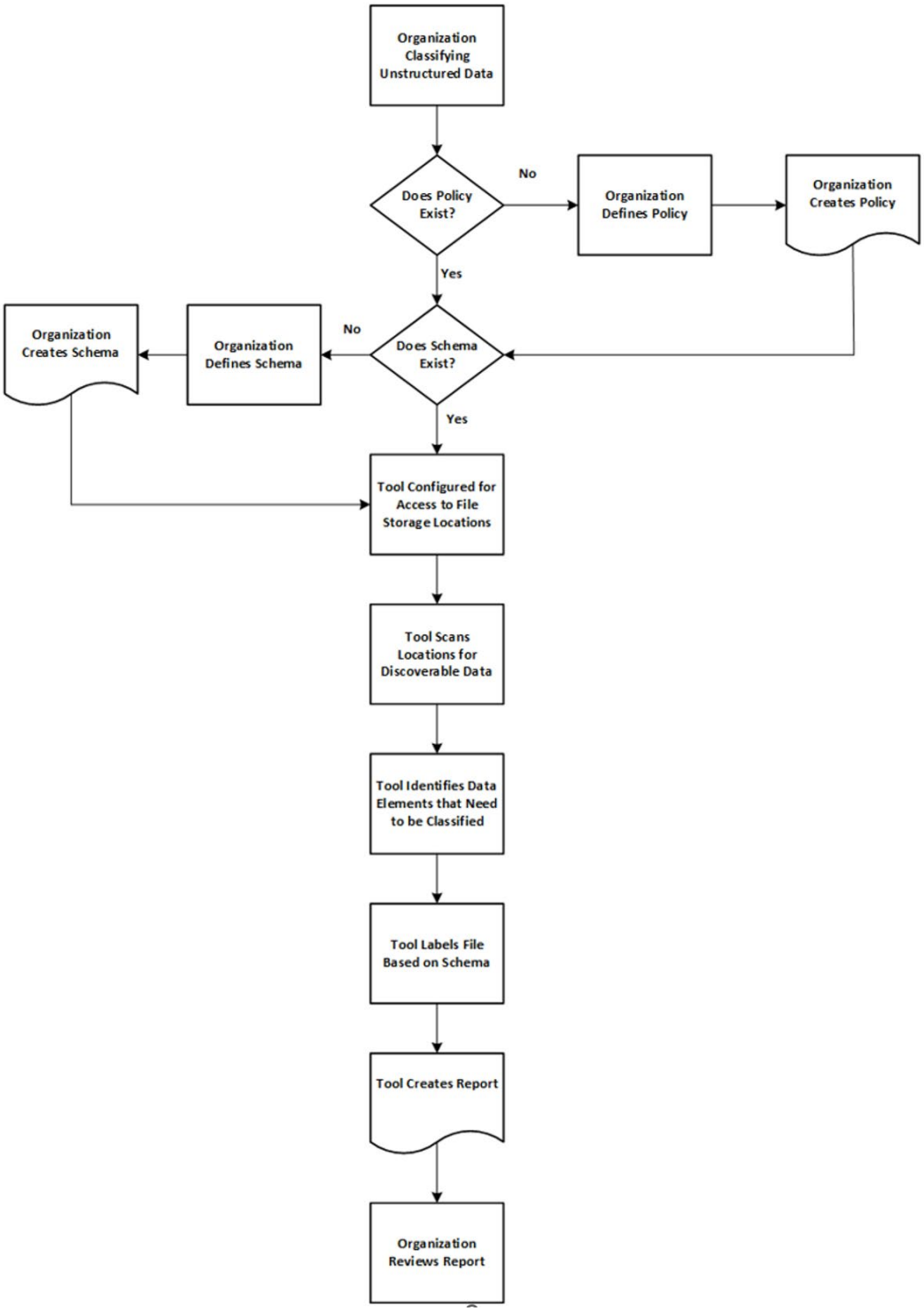
397 4.3 Unstructured Data Classification Practice Demonstration Workflow

398 The following describes an unstructured data classification workflow:

399 400 - **Organization Classifying Unstructured Data -** The lab demonstration shows data classification practices for unstructured data.
401 402 403 - **Does Policy Exist? -** Is the organization using existing policies or creating new classification policies? Data Classification Policy is an organization's data classification schema and the formal description of the data types within the organization.
404 405 406 407 408 - **Does Schema Exist? -** The organization will use existing schema or will create new schema to allow a data classification tool to be programmed with a rules syntax that triggers based on data elements, the sensitive or valuable content and metadata, in the files being analyzed. For example, part of a schema might involve data classifications that characterize high-level business data types of a data asset—for instance, "customer #," or "billing #".
409 410 - Five demonstrations were performed in this project and the documentation of each that follows includes a section on schema creation.
411 412 - **Tool Configured for Access to File Storage Locations -** The classification tool is provided access to where the data is located. For example, a network file share or other location.
413 414 - **Tool Scans Locations for Discoverable Data -** Data discovery takes place as the tool scans the provided locations.
415 416 - **Tool Identifies Data Elements that Need to be Classified -** The tool ingests information about the data. File type, metadata, and content are determined.

- **Tool Labels File Based on Schema -** Labels are assigned to the data based upon its content and in relation to the classification schema.
- **Tool Creates Report** on Data Classification Practice performed on unstructured data**,** and **Organization Reviews Report -** The tool makes the results available, and the organization evaluates them.

The bold text above matches the text in the boxes shown below in Figure 4-2 Unstructured Data Classification Practice Demonstration Workflow.



424 **Figure 4-2 Unstructured Data Classification Practice Demonstration Workflow**

4.3.1 ActiveNav Demonstration

ActiveNav helps organizations take control of their unstructured data**.** ActiveNav Discovery Center finds and classifies unstructured data based on pre-defined and custom classification schemas. Appendix E.1 provides ActiveNav Implementation Details.

4.3.1.1 Schema Creation

For the ActiveNav demonstration scenario, a classification schema was created in ActiveNav Discovery Center Workbench tool. The custom classification schema targeted the discovery of the data types defined in Table 3-2 including patient personal data, financial data, and health information data. The classification schema was created using a rules syntax that allowed the creation of classification nodes that trigger based on the content and metadata of the searched files. For example, a rule identifying all files that contained one or more patient numbers was used. The classification schema could identify and label files with sensitivity levels ranging from low to high based on the content they contain.

The classification schema was then exported from the Workbench tool and imported into ActiveNav Discovery Center Project Suite application. To assign the classification schema to a file discovery scan, an Index Configuration called NIST-DEMO was created that targeted the data we wanted to find in the dataset, such as names, birthdates and patient billing numbers. Based on the schema, ActiveNav then assigned a sensitivity level to each file that was discovered to have privacy, personal or health information. With the Index Configuration created, then an Index (scan) was created that used the NIST-DEMO Index Configuration as a template. The ActiveNav demonstration scenario discovered, identified, and labeled data with privacy, personal credit, and health information schema categories as shown below.

- Privacy High
- Privacy Medium
- Privacy Low
- Personal Credit High
- Personal Credit Medium
- Personal Credit Low
- Health Information High
- Health Information Medium
- Health Information Low

The above schema categories could be part of an organization’s public data classification statement. For example, an organization may have a policy to classify private customer data such as name and birthdate with a “Privacy High” classification and classify customer metadata such as the number of support tickets opened by the customer as “Privacy Medium” classification.

4.3.1.2 Data Discovery

To facilitate the scan, the synthetic dataset was copied to a network file share on a domain controller. The address of the dataset was added to ActiveNav Discovery Center Project Suite as a scan location and

then assigned to the previously created index as a target for the data discovery. With the Index fully configured, the scanning operation was initiated. The scan completed and a total of 25,884 files were discovered and processed for data identification purposes.

4.3.1.3 Data Identification

The ActiveNav Discovery Center Index discovered and analyzed a total of 25,884 files, gathering information about each one, including the file type, metadata, and content. The content of each file was analyzed and compared to the classification schema to identify sensitive information. The results of this process were then used to label and catalogue the files.

4.3.1.4 Data Labeling / Cataloguing

After the content of each discovered file was analyzed, classification labels were assigned to each respective file based on the presence of data types found in the content and in relation to the classification schema. For example, based upon the created schema, 10,363 files were labeled as Health Information HIGH while 12,431 files were labeled with Privacy HIGH. ActiveNav Discovery Center created a catalogue of all discovered files and their content. Reports can be generated to identify which files contained specific pieces of data related to the classification schema.

4.3.1.5 Cross-Product Integration

An additional demonstration activity was performed to look at multiple organizational use cases that can involve more than one product. In the additional demonstration activity, the ActiveNav product read the classification tags contained in the embedded “Tags” and “Categories” custom metadata properties created by Janusseal. ActiveNav used extraction rules to discover the classification tags and store them as metadata associated with the file. ActiveNav then performed actions based on the classification tags that Janusseal applied to the document (low, medium, high, etc.), including re-classifying the document based on ActiveNav’s custom rules. The classification rules in the ActiveNav product could also be used to translate the classification tags found in the Janusnet documents (low, medium, high, etc.) and translate them to a different classification schema such as Restricted, Confidential, and Public or Internal. This demonstration showed the ability of an organization to use two products and read the classification information created by one product and update the classification information using another product.

4.3.2 IBM Demonstration

IBM Guardium Discover and Classify provides discovery and classification of structured and unstructured data. Appendix E.2 provides IBM Guardium Discover and Classify Implementation Details.

4.3.2.1 Schema Creation

Custom classification schemas were created in IBM Guardium Discover and Classify (IGDC) to represent the types of data being discovered and classified. For instance, the custom data element NCCOE_BILLING_NUMBER was created for this demonstration scenario to represent a patient’s account number. This data element was assigned a data type of "sensitive personal" and configured with a custom search pattern using regex.

Additionally, data types can be assigned a sensitivity level such as Restricted, Confidential, and Public or Internal. Once the data types have been created and configured, scans can perform discovery based on the data types that have been created/configured. The IBM demonstration scenario discovered, identified, and labeled data with a sensitive personal schema category across the synthetic data's 12 data types as shown below.

- Sensitive personal schema selected across 12 data types including:
  - Name
  - Address
  - Birthdate

4.3.2.2 Data Discovery

For this demonstration scenario, the synthetic patient dataset was copied to a Server Message Block (SMB) file share located on a Windows Server VM within the lab. The address of the SMB file share was then added as a Data Source Catalog. The Data Source Catalog was configured with credentials to access the SMB file share and assigned an analysis strategy of "baseline," which instructs the tool to perform a full scan of the dataset each time. With everything configured, the scan was initiated, and IBM IGDC successfully accessed the dataset in the SMB file share and processed a total of 25,884 files.

4.3.2.3 Data Identification

During the dataset scan, IBM IGDC processed a total of 25,884 files and analyzed them to determine information about each file and the content they contained. The content of each file was compared to the data types established as part of the classification schema. The results of the scan were then compiled into a report detailing the number of files containing each specific data element from the classification schema.

4.3.2.4 Data Labeling / Cataloguing

After the data in each discovered file was catalogued, a classification label was determined and assigned to the appropriate files. In our schema, all discovered data types were configured as "sensitive personal" data. IBM IGDC also assigned a "criticality score" to each file that helps determine the importance of the data it contains. IBM IGDC also established a searchable database of all discovered files and their content. Each file is assigned one or more classification labels depending on its content. For each file, data element discovered in that file was provided.

4.3.3 Thales Demonstration

Thales CipherTrust Data Discovery and Classification (DDC) solution enables users to locate structured and unstructured data. Appendix E.3 provides CipherTrust DDC Implementation Details.

4.3.3.1 Schema Creation

In the Thales demonstration scenario, a classification profile called Synthetic Data Classification was created and assigned a sensitivity level of restricted within the CipherTrust Data Discovery and Classification (DDC) solution. The classification profile was assigned prebuilt and individual pieces of data to search for during a scan (custom infotypes) that targeted the sensitive information found in the

synthetic dataset (Names, Birthdates, NCCoE_Customer_Number, Patient ID, etc.). The classification profile was then assigned to a custom scan that would be used to discover the synthetic unstructured dataset. The Thales demonstration scenario discovered, identified, and labeled data with a schema that placed data into the following 3 categories as shown below.

- Personal data
- Medical
- Financial

4.3.3.2 Data Discovery

To facilitate the scan, the synthetic patient dataset was copied to a directory on a domain controller in the lab. The domain controller was then added as an available data store to be scanned by CipherTrust DDC. During the creation of the custom scan, the domain controller data store was selected and provided the file path where the synthetic dataset was located. With the custom scan configured, the scanning operation was initiated. The scan completed and a total of 25,884 files were discovered and processed.

4.3.3.3 Data Identification

The Thales DDC scan processed a total of 25,884 files, gathering information about each one, including file type, size, modified date, and other metadata. The content of each file was analyzed and compared to the classification schema to identify sensitive information. The results of this process were then used to label and catalogue the files.

4.3.3.4 Data Labeling / Cataloguing

After the content of each discovered file was analyzed, classification labels were assigned to each respective file based on its content in relation to the classification schema. For example, files containing an NCCoE_Billing_Number or a passport number were labeled to show that the file contained those data types. Thales DDC also created a searchable catalogue of all discovered files. Reports can also be generated to identify which files contained specific pieces of data related to the classification schema.

4.3.4 Trellix Demonstration

Trellix Data Loss Prevention (DLP) Discover allows you to locate, classify, and protect all types of corporate data. Appendix E.4 provides Trellix DLP Discover implementation details.

4.3.4.1 Schema Creation

For the Trellix DLP Discover demonstration scenario, a classification schema was created that targeted custom classifications including privacy, health info, and payment information, each with low, medium, and high variations based on the number of instances found in each document. The classification schema also contained out-of-the-box classifications targeting personally identifiable information and protected health information. The classification policy was then assigned to a network file system (NFS) file share located within the lab and automatically propagated to the DLP Discover scan. The Trellix demonstration scenario discovered, identified, and labeled data with a schema that placed data into privacy, health information, and Payment Card Industry (PCI) categories as shown below.

- Privacy Low
- Privacy Medium
- Privacy High
- Health Info Low
- Health Info Medium
- Health Info High
- PCI Low
- PCI Medium
- PCI High

4.3.4.2 Data Discovery

The synthetic dataset was copied to the NFS file share located in the lab. The address of the dataset was added to the Trellix DLP Discover scan. The scanning operation was then initiated, and the 25,884 files dataset was discovered and processed.

4.3.4.3 Data Identification

The Trellix DLP Discover scan processed the synthetic file dataset, gathering information about each file, including file type, size, modified date, and other metadata. The content of each file was analyzed and compared to the classification schema to identify sensitive information. The results of this process were then used to label and catalogue the files.

4.3.4.4 Data Labeling / Cataloguing

After the content of each discovered file was analyzed, classification labels were assigned to each respective file based on their content in relation to the classification schema. For example, files containing synthetic privacy information were labeled as Privacy Medium while files containing synthetic health record information were labeled with Health Info Low. Trellix DLP Discover also created a searchable catalogue of all discovered files. Reports can also be generated to identify which files contained specific data types related to the classification schema. For example, name, address, birthdate, or account numbers.

4.4 Electronic Mail Message Demonstration

Unstructured data is also commonly found in electronic mail (email) which compelled one demonstration to show that unstructured data in emails can be discovered, identified, and labeled using a collaborator tool.

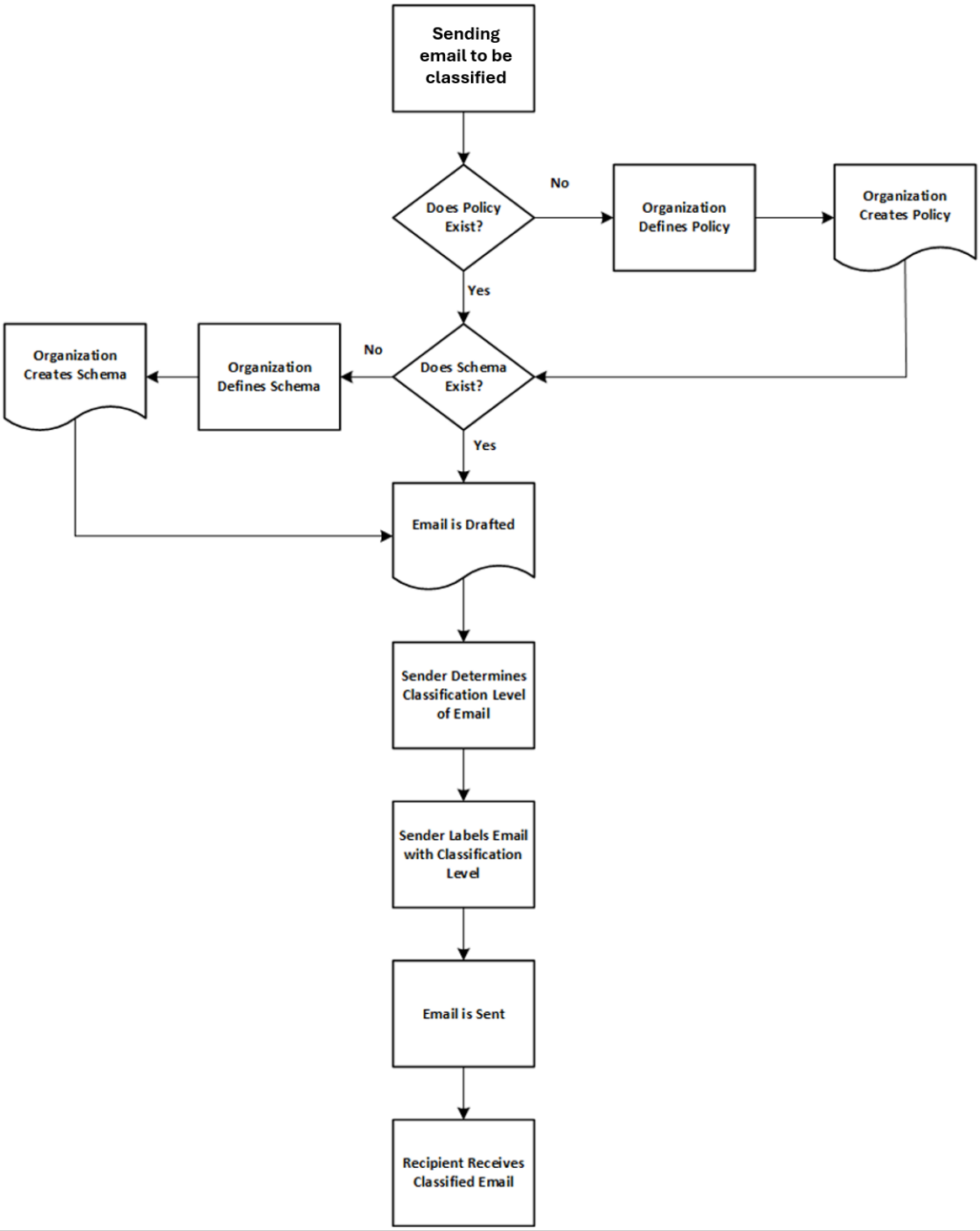
4.4.1 Email Practice Demonstration Workflow

The following describes an email data classification workflow:

- **Sending Email Message to be classified -** A classified email is being sent. In the lab demonstration, email is sent between 2 different accounts.

- **Does Policy Exist?** Is the organization using existing policies or defining and creating their classification policies? The classification policies in this publication categorize data based on the data's sensitivity. For example, an organization's data classification policies related to the sensitivity of the content of an email can require an email to be identified as "publicly releasable" or "not publicly releasable". The organization's email data classification policies define how the email should be controlled.
- **Does Schema Exist? -** Is the organization using existing schema or defining and creating schemas. Using the policy examples above, "Releasable", "Low", "Medium", "High", and "Internal Use Only" schema hierarchy was created.
- **Email is Drafted -** The sender determines the email classification level. The email's classification level is determined by the sensitivity of the message content and attachments.
- **Sender Determines Classification Level of Email and Sender Labels Email with Classification Level -** The user selects the classification level from a drop-down label that is part of the email client. The labels reflect the schema that was created above. The email is labeled with "Releasable", "Low", "Medium", "High", or "Internal Use Only".
- **Email is Sent and Recipient Receives the Classified Email -** Both users can see the classification level that was selected for the email.

The workflow shown in Figure 4-3 was used to perform the demonstration steps in the lab.



625 **Figure 4-3 Email Message Demonstration Workflow**

626 ### 4.4.2 Janusseal Email Demonstration

627 Janusnet products include a utility application to create separate, and different classification schemas
628 for organizations and to perform active classification of an organization's email messages as they were

sent, selecting the most appropriate level of sensitivity based on the content of the email. Appendix E.5 provides Janusseal Implementation Details.

4.4.2.1 Use Case

A common organizational use case was established with email to explore how unstructured data classification needs can be improved. A single organization using a data classification product for email was installed and configured. Data markings were established and communicated within a single organization. Then, a multiple organization demonstration scenario using the same product in both organizations was established using various schemas.

4.4.2.2 Schema Creation

This build demonstrated data classification practices for email. To accomplish this, two Janusnet products were installed, Janusseal Schema and Janusseal for Outlook. Janusseal Schema is a utility application used during the initial configuration that was used to create separate, and different classification schemas for each organization to explore how organizations can interoperate with different classification schemes.

The created schemas were then used to establish a Microsoft Group Policy Object (enterprise configuration) on each organization's domain controller that propagated the Janusseal configuration to users within the organization. On each user workstation, Janusseal for Outlook was installed, a plugin for Microsoft Outlook, that was configured by the group policy object to allow for the user-driven classification of email messages by the email sender.

With Janusseal for Outlook installed, users in each organization had the ability to perform active classification of their email messages as they were sent, selecting the most appropriate level of sensitivity based on the content of the email. The classification schemas for each organization were slightly different, allowing different test scenarios to be performed by modifying the recipient of the email, internal or external, or the assigned classification.

For the Janusnet demonstration scenario, two separate organizations were established with their own domains and email servers. For each organization, Janusseal Schema was installed on a single Windows workstation for use by a system administrator, and the following similar but unique classification schemas were created in each fictitious company that would provide for a more realistic demonstration of two organizations:

- Organization A Schema (hierarchical, lowest to highest sensitivity)
  - Releasable
  - Low
  - Medium
  - High
  - Internal Use Only
- Organizational B Schema (hierarchical, lowest to highest sensitivity)
  - Releasable

- Low
- Medium
- High
- Very High
- Not Publicly Releasable

Each classification schema was exported from Janusseal Schema as two separate Windows Group Policy Administrative Template (ADM) files, one targeting Janusseal for Outlook and one as a Janusseal product-agnostic security classification schema. With the ADM files generated, a custom Group Policy Object (GPO) was created for each organization domain and then both ADM files were imported into their respective domains. Each GPO was then configured to activate the Janusseal product and schemas which were then applied to all Windows workstations in Organizations A and B.

4.4.2.3 Data Discovery, Identification, and Labeling

On each end-user Windows workstation running Microsoft Outlook Janusseal for Outlook was installed. This product integrates directly into Microsoft Outlook as an Outlook add-in and enables functionality based on the configuration embedded in the GPO established in the previous section. For this demonstration scenario, the tool was configured to perform certain actions automatically while also allowing the user to manually select a classification level to label their emails depending on content.

On Windows workstations in Companies 3 and 4, Microsoft Outlook was used to write emails containing synthetic patient data that were then used for data discovery, identification, and labeling. Data discovery and identification was performed by the user authoring the email. For the emails containing the synthetic patient data, the appropriate classification level of the content was manually applied using Janusseal for Outlook. The tool then added the classification markings to the top, bottom, and email header metadata. For a file that is labeled using the sibling product, Janusseal Documents, when added as an attachment to an email Janusseal for Outlook automatically discovers the classification of the file and uses it to determine the minimum allowable classification that can be selected for the email message.

The Janusseal products are also able to discover and write other classification metadata, such as that of Microsoft's Purview labels. The Janusseal products treat its own classification metadata as authoritative, if present, but can use the other metadata structures if Janusseal metadata is not present. When the data has been identified, Janusnet can insert classification tags for labeling purposes into the metadata of office documents and outlook emails. For Microsoft Office documents, Janusnet inserted the classification tags under the “Tags” and “Categories” properties and other Office file metadata areas (custom document properties).

4.5 Tool Summary

The name of the tools used in the Data Classification Practices demonstrations and the functions they supported are shown in Table 4-1.

Table 4-1 Data Classification Practices Tools and Functions

Tool	Data Classification Practice Function	Tool Function
ActiveNav Discovery Center Project Suite 4.18.0.0	Locates and classifies unstructured data, including duplicate and sensitive data discovery. Provides labeling to the data based upon customizable schemas that are based upon extraction rules aligned with organizational information policies.	Discovery Center Project Suite provides discovery, identification, and labeling of files based on the organization's policies and the data's contents and metadata. The data content determines the classification and label, which can be viewed, sorted, and filtered in configurable reports. Discovery Center Project Suite maintains a metadata catalog of all discovered files, and the classification(s) they were assigned. Discovery Center Project Suite provides incremental scans that can discover newly created files and modified content on a periodic basis. Discovery Center Project Suite identifies and records the presence of sensitive content within unstructured data within its metadata catalog. This supports the ability to understand and quantify risk.
ActiveNav Discovery Center Workbench 4.18.0.0	Designs and creates the data classification policies and schemas that are implemented and used by the ActiveNav Discovery Center product.	Discovery Center Workbench supports the data asset selection and life cycle management with the ability to create classification schema for use in tagging files against sensitivity and other lifecycle facets.
Janusnet Janusseal for Outlook 3.7.2.28374	Provides Microsoft Outlook users the ability to apply classification labels to email and create human-readable	Janusseal for Outlook can work with customizable schemas, provides classification and human-readable classification markings within the email

Tool	Data Classification Practice Function	Tool Function
	classification markings within the email subject and body.	subject, body, and header and the ability to block emails sent to unauthorized individuals. Janusseal for Outlook creates human-readable classification markings and the ability to read the classification levels the emails were assigned. Janusseal for Outlook provides the ability to classify email, and then reference the classifications throughout that data's lifecycle.
Janusnet Janusseal Schema 1.4.4.24270	Creates classification schemas that are then applied to emails and files.	Janusnet Janusseal Schema facilitates the creation of classification schemas, which define the rules determining how data will be labeled based on sensitivity and other attributes.
IBM Guardium Discover and Classify (IGDC) 4.0.2	Discovery, Identification, and Labeling	IGDC discovers, identifies, and labels data based on contents and metadata. Files are discovered and their content identified, and then they are labeled according to configurable policies. Results of the classification actions can be viewed, sorted, and filtered in configurable reports. IGDC maintains a catalog of all discovered files, their content, and the classification they were assigned. IGDC classifies data and provides reports that can then be used to manage data throughout its lifecycle.
Thales CipherTrust Manager 2.17.0+12772	Discovery, Identification, and Labeling	CipherTrust Manager provides results of the classification actions can be viewed, sorted,

Tool	Data Classification Practice Function	Tool Function
		and filtered in configurable reports. CipherTrust Manager maintains a catalog of all discovered files, their content, and the classification they were assigned. CipherTrust Manager provides reports that can then be used to manage data throughout its lifecycle.
Thales Data Discovery and Classification (DDC) 3.1.6-316	Discovery, Identification, and Labeling	Data Discovery and Classification discovers, identifies, and labels data based on contents and metadata. Files are discovered and their content identified, and then they are labeled according to configurable policies. Data Discovery and Classification provides incremental scans that can discover newly created files as well as existing files with new modifications.
Trellix ePolicy Orchestrator 5.10.0	Discovery, Identification, and Labeling	ePolicy Orchestrator provides results of the classification actions can be viewed, sorted, and filtered in configurable reports. ePolicy Orchestrator maintains a catalog of all discovered files, their content, and the classification they were assigned. ePolicy Orchestrator provides reports that can then be used to manage data throughout its lifecycle.

Tool	Data Classification Practice Function	Tool Function
Trellix DLP Discover 11.10.500.112	Discovery, Identification, and Labeling	DLP Discover discovers, identifies, and labels data based on contents and metadata. Files are discovered and their content identified, and then they are labeled according to configurable policies. DLP Discover provides incremental scans that can discover newly created files as well as existing files with new modifications.

5 Findings and Insights

A listing of the general findings from the unstructured data classification practices demonstrations follows:

The use of synthetic data objects was useful in our data classification practice demonstrations for the following reasons and may be useful to organizations to gain experience with data classification tools:

- seeing that metadata attributes in the synthetic data files were discovered by the data classification tools - confidence
- the creation of a variety file types with unstructured data provides the ability to test data classification tool - options
- having statistics about the synthetic files with data types that should be discovered by the tools - verification

Performing a **validation of a schema** on a small subset of files before doing the whole set can help avoiding multiple reruns as the classification schema is refined.

Schema versioning is important. Several versions of schemas were developed for our lab. Maintaining a record of schema versions along with results of analysis on how many labels were created can help one select a schema that meets your data classification needs.

Appendix A List of Acronyms

ADM	Administrative Template
CDO	Chief Data Officer
CISO	Chief Information Security Officer
CRADA	Cooperative Research and Development Agreement
DDC	Data Discovery and Classification
DLP	Data Loss Prevention
FRN	Federal Register Notice
GPO	Group Policy Object
IBM	International Business Machines Corporation
IGDC	IBM Guardium Discover and Classify
NCCoE	National Cybersecurity Center of Excellence
NFS	Network File System
NIST	National Institute of Standards and Technology
OCR	Optical Character Recognition
OMB	Office of Management and Budget
PCI	Payment Card Industry
PHI	Protected Health Information
PII	Personally Identifiable Information
SMTP	Simple Mail Transport Protocol
SP	Special Publication
ZTA	Zero Trust Architecture

Appendix B Glossary

For this initial public draft, we are not prescribing terms and definitions. Please submit comments for terms or words for which an explanation would enhance your understanding.

Appendix C References

[1] National Cybersecurity Center of Excellence (NCCoE) Data Classification Practices: Facilitating Data-Centric Security Management, A Notice by the National Institute of Standards and Technology on 8 October 2021. [Online]. Available: https://www.federalregister.gov/documents/2021/10/08/2021-21979/national-cybersecurity-center-of-excellence-nccoe-data-classification-practices-facilitating [Accessed 9 July 2025] .

[2] NCCoE Workshop: Information Protection and Data-Centric Security Management: Data Classification Workshop , October 23, 2019 [Online]. Available: https://www.nccoe.nist.gov/get-involved/attend-events/information-protection-and-data-centric-security-management-data . [Accessed 9 July 2025]

[3] MIT Sloan stories to boost your data analytics strategy," 3 March 2021. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/21-mit-sloan-stories-to-boost-your-data-analytics-strategy#:~:text=Unstructured%20data%20is%20an%20area,out%20how%20to%20use%20it . [Accessed 9 July 2025]

[4] The Centers for Medicare & Medicaid Services (CMS) Alliance to Modernize Healthcare (CAMH) Federally Funded Research and Development Corporation, Operated by The MITRE Corporation, "SyntheticMass," [Online]. Available: https://synthea.mitre.org/ . [Accessed 22 January 2025].

Appendix D Synthetic Data Creation Steps

The following processes were used to create the synthetic data used in the lab demonstrations. The first approach used the Mail Merge feature of Microsoft Word to create multiple Word document outputs from the synthetic data corpus while the second used a Windows PowerShell script to product a wide variety of file outputs including Word documents, Excel workbooks, PowerPoint presentations, Adobe PDFs, Rich Text Format files, HTML files, plain text files, PGN image files, WAV audio files and ZIP files. The steps of the process were performed using a Microsoft Windows based computer with Microsoft Office and the current Microsoft .NET Framework version 4 runtime installed.

D. 1 Download the Data

1. Visit **https://synthea.mitre.org/**
2. Select **Download Data**
3. Under **Previous Versions (2019-2021)** select **1K Sample Synthetic Patient Records, CSV**
4. Extract the files from the downloaded **.zip** file
5. Navigate to the **csv** folder and open the **patients.csv** file in Microsoft Excel. If prompted to convert the file's format, do not convert.

D. 2 Format the Data

6. (Optional) Remove **columns** that will not be part of your dataset from the patients.csv file. For example:
    a. PREFIX
    b. SUFFIX
    c. BIRTHPLACE
    d. LAT
    e. LON
    f. HEALTHCARE_EXPENSES
    g. HEALTHCARE_COVERAGE

D.3 Create Word Document Outputs Using Mail Merge

D.3.1 Set Up Mail Merge

7. Open a new **Microsoft Word** document
8. Click **File**, select **Save**, and name the document "**Mail Merge Template**"
9. Under the **Mailings** tab, click **Start Mail Merge** and select **Letters**
10. Click **Select Recipients** and select **Use an Existing List…**
11. Navigate to the **csv** folder containing the **synthetic patient data** from Synthea
12. Select the **patients.csv** file, click **Open**
13. If a pop-up titled "**File Conversion – patients.csv**" appears, leave the **default settings** and click **OK**

D.3.2 Create Mail Merge Template

14. For the first line of the document, format the text as a **heading** (found under the **Home** tab in the **Styles** section)
    a. NOTE: Formatting the first line as a header isn't necessary for the mail merge itself, but will allow us to save the merged records as individual Microsoft Word documents in a future step

15. Under the **Mailings** tab, click **Insert Merge Field** and select the data you want to include in the template
    a. The following field examples could be selected:
        i. **First** (Name), then **insert** a **space** to separate first name from last name
        ii. **Last** (Name), then press **Enter** to create a carriage return
        iii. Type "Your SSN is:", **insert** a **space**, then select **SSN** (Social Security Number), press **Enter** to create a carriage return
        iv. Type "Your address is:", **insert** a **space**, then select **Address, City, State, Zip**
    b. The template will look like this:

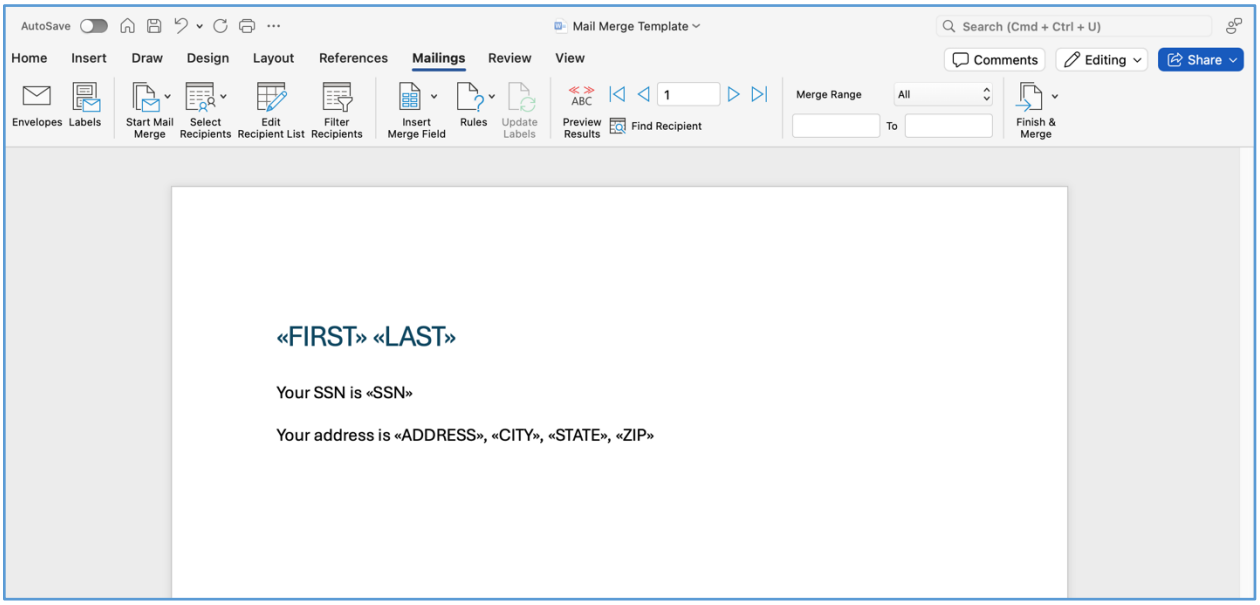


Figure D-1 Synthetic Data Template Example

16. (OPTIONAL) Click **Preview Results** to view what the template will look like with patient data inserted
    a. The results will look like this:

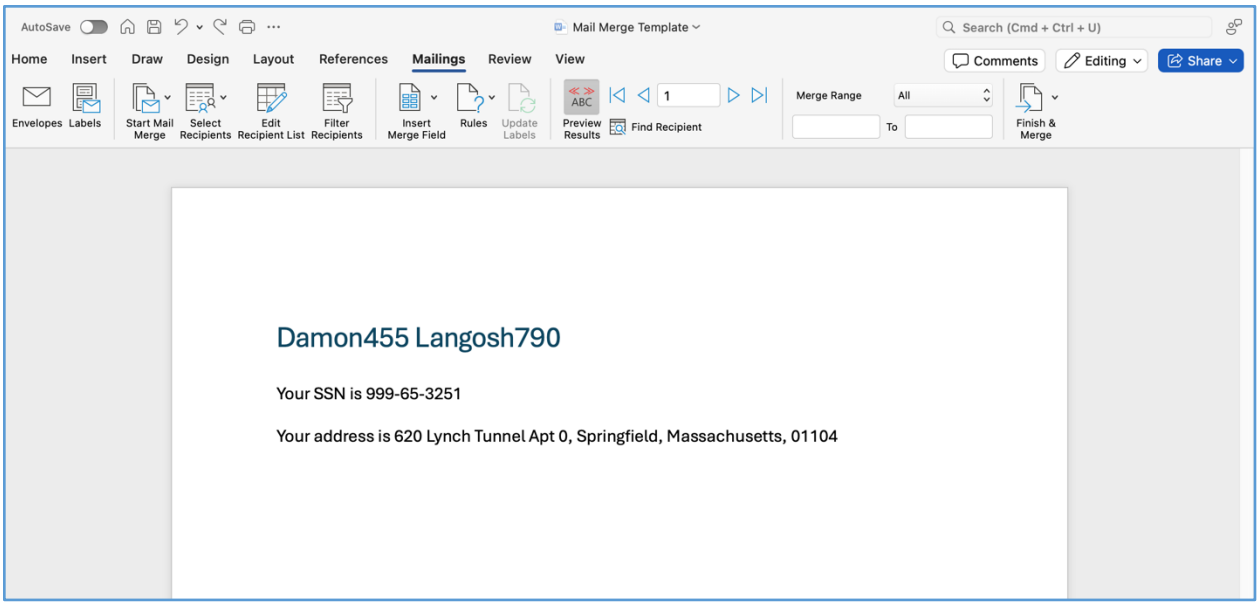


Figure D-2 Results Preview

17. Click **Finish and Merge**, select **Edit Individual Documents**
18. Under **Merge Records**, set the value for **From:** to "**1**", and for **To:** to "**50**"
    a. This value determines which patient records are pulled from the patients.csv file. At this point, only the first 50 records are being exported
19. Click **Ok**
20. A new Microsoft Word document will open that contains each patient record that is exported on their own page
21. In the new document click **File**, select **Save**, and name the document "**All Patients**"

D.3.3 Create Individual Microsoft Word Documents for Each Patient

22. In the "**All Patients**" document, under the **View** tab, click **Outline**
23. Use **CTRL+A** to highlight all **text** in the document
24. With the **text** highlighted, click **Show Document** and then click **Create**
    a. You will see data such as this:

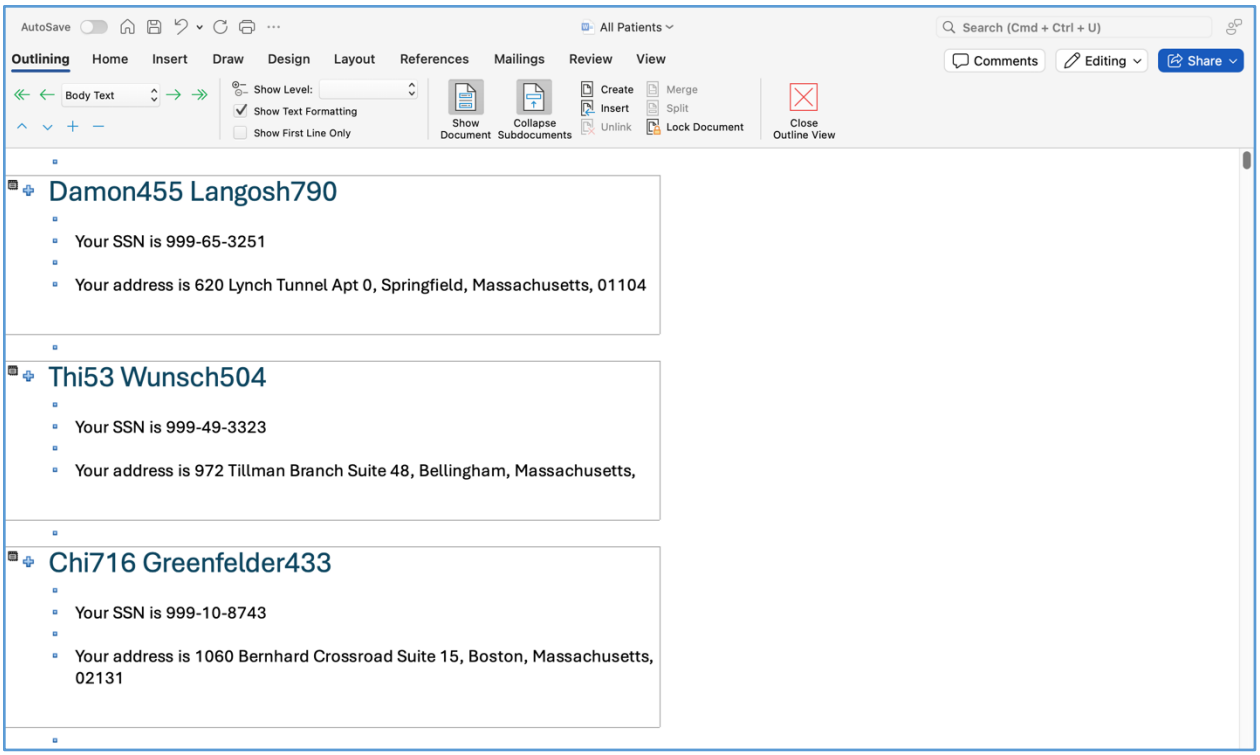


Figure D-3 Synthetic Patient Record

25. In the "**All Patients**" document click **File** and select **Save**
    a. Depending on the speed of your computer, it might take several seconds to create the new individual documents
26. Click **Close Outline View**
27. Navigate to the folder where the "**All Patients**" document is stored
28. Find the individual Microsoft Word documents for each patient in the "**All Patients**" folder

D.4 Create File Outputs Using PowerShell Script

A Windows PowerShell script was developed to automate the creation of the nearly 30,000 file outputs. It is driven by the **csv** file that was obtained in the initial steps and uses various Office automation and other file creation techniques to create the outputs.

```
###########################################
# Author: Neville Jones, Janusnet;
#              neville.jones@janusnet.com
#
# No warranty
#
# Version: 0.8.0
# Date: 2024-09-30 15:10 +11:00
###########################################

###########################################
# Settings for running the script are here #
$extensions = @('doc', 'docx', 'eml', 'html', 'pdf', 'png', 'pptx', 'rtf', 'txt', 'xls', 'xlsx', 'wav', 'zip')      # an array
$includeLabels = @($true, $false) # whether descriptive labels (derived from CSV column headers) should be included in the generated outputs
$correctPersonName = $true
$processRowCount = -1 # the number of rows in the source CSV to process; if set to -1 then will process all rows
$data = Import-Csv -Path .\patients.csv -Delimiter ','
###########################################
```

```
#$extensions = @('txt') # for debugging - can uncomment this line to only do a single (or a subset of) extension

Add-Type -AssemblyName System.Drawing
Add-Type -AssemblyName System.Speech

function ReplaceDigits {
    [OutputType([String])]
    param ([String] $inputString)
    $result = $inputString
    $result = $result.Replace('0','zero')
    $result = $result.Replace('1','one')
    $result = $result.Replace('2','two')
    $result = $result.Replace('3','three')
    $result = $result.Replace('4','four')
    $result = $result.Replace('5','five')
    $result = $result.Replace('6','six')
    $result = $result.Replace('7','seven')
    $result = $result.Replace('8','eight')
    $result = $result.Replace('9','nine')
    return $result
}

$rootDir = Split-Path $MyInvocation.MyCommand.Path -Parent
$outputDirRoot = "output"
if (!(Test-Path .\$outputDirRoot)) {
    New-Item -Path .\$outputDirRoot -ItemType "directory"
}

foreach($extension in $extensions) {
    $msapp = $null
    try {
        Write-Host "-------------------------------------------------------------"
        foreach($includeLabel in $includeLabels) {

            $outputDir = If($includeLabel) {$outputDirRoot + "\" + $extension + "\labels" } Else {$outputDirRoot + "\" + $extension + "\no labels"}
            if (!(Test-Path .\$outputDir)) {
                New-Item -Path .\$outputDir -ItemType "directory"
            }
            Remove-Item $outputDir\*.$extension

            if ($null -eq $msapp) {
                switch -Regex ($extension) {
                    '^pdf$' {
                        $_ = 'docx' # for PDF we will open an existing docx file in Word and use its ability to export PDF
                    }
                    '^(docx?|html|rtf)$' {
                        $msapp = New-Object -ComObject Word.Application
                        $msapp.Visible = $true
                        Write-Host "---- Creating synthetic data using $($msapp.Name) ----"
                    }
                    '^pptx$' {
                        $msapp = New-Object -ComObject PowerPoint.Application
                        $msapp.Visible = 1
                        Write-Host "---- Creating synthetic data using $($msapp.Name) ----"
                    }
                    '^xlsx?$' {
                        $msapp = New-Object -ComObject Excel.Application
                        $msapp.Visible = $true
                        $msapp.ScreenUpdating = $false
                        $msapp.EnableEvents = $false
                        Write-Host "---- Creating synthetic data using $($msapp.Name) ----"
                    }
                }
            }

            $i = 1
            foreach($datum in $data) {
                if (($processRowCount -ne -1) -and ($i -gt $processRowCount)) {
                    break
                }

                $wroteFile = $true
                $firstName = $datum.FIRST
                $lastName = $datum.LAST
```

```
                if ($correctPersonName) {
                    $firstName = ReplaceDigits -inputString $firstName
                    $lastName = ReplaceDigits -inputString $lastName
                }

                $id = $datum.Id
                $outputFilename = "$rootDir\$outputDir\$id.$extension"

                $output = "$($datum.Disclaimer)`r`n`r`n"

                $output += If($includeLabel) {"Name                    : "}
                $output += "$firstName $lastName`r`n"

                $output += If($includeLabel) {"Address                 : "}
                $output += "$($datum.ADDRESS)`r`n"

                $output += If($includeLabel) {"City                    : "}
                $output += "$($datum.CITY)`r`n"

                $output += If($includeLabel) {"County                  : "}
                $output += "$($datum.COUNTY)`r`n"

                $output += If($includeLabel) {"State                   : "}
                $output += "$($datum.STATE)`r`n"

                $output += If($includeLabel) {"Zip                     : "}
                $output += "$($datum.ZIP)`r`n"

                $output += If($includeLabel) {"Birthdate               : "}
                $output += "$($datum.BIRTHDATE)`r`n"

                $output += If($includeLabel) {"License #               : "}
                $output += "$($datum.DRIVERS)`r`n"

                $output += If($includeLabel) {"Passport #              : "}
                $output += "$($datum.PASSPORT)`r`n"

                $output += If($includeLabel) {"NCCoE Customer # : "}
                $output += "$($datum.NCCoE_Customer_Number)`r`n"

                $output += If($includeLabel) {"NCCoE Billing #    : "}
                $output += "$($datum.NCCoE_Billing_Number)`r`n"

                switch -Regex ($extension) {
                    # simple files
                    '^eml$' {
                        $output = "MIME-Version: 1.0`r`nContent-Type: text/plain`r`nFrom: <alice@nccoe.nist.test>`r`nTo: <bob@nccoe.nist.test>`r`nSubject:
test email ID $id`r`n`r`n" + $output

                        Out-File -FilePath "$outputFilename" -InputObject $output
                    }
                    '^txt$' {
                        Out-File -FilePath "$outputFilename" -InputObject $output
                    }
                    # MS Office files
                    '^(docx?|html|rtf)$' {
                        $newDoc = $msapp.Documents.Add()
                        $range = $newDoc.Paragraphs(1).Range
                        $range.Text = $output

                        if ($extension -eq 'docx') {
                            $newDoc.SaveAs("$outputFilename", 16) # https://learn.microsoft.com/en-
us/dotnet/api/microsoft.office.interop.word.wdsaveformat?view=word-pia
                        } elseif ($extension -eq 'doc') {
                            $newDoc.SaveAs("$outputFilename", 0)
                        } elseif ($extension -eq 'html') {
                            $newDoc.SaveAs("$outputFilename", 10)
                        } elseif ($extension -eq 'rtf') {
                            $newDoc.SaveAs("$outputFilename", 6)
                        }
                        $newDoc.Close()
                    }
                    '^pptx$' {
                        $newPres = $msapp.Presentations.Add()
                        $master = $newPres.SlideMaster
```

```
            $customLayout = $null
            foreach($cLayout in $master.CustomLayouts) {
                if ($cLayout.Name -eq "Blank") {
                    $customLayout = $cLayout
                }
            }
            $newSlide = $newPres.Slides.AddSlide(1, $customLayout)
            $textBoxShape = $newSlide.Shapes.AddTextbox(1, 50, 50, 800, 400)
            $textFrame = $textBoxShape.TextFrame
            $textFrame.TextRange.Text = $output

            $newPres.SaveAs("$outputFilename")

            $newPres.Close()
        }
        '^xlsx?$' {
            $newWB = $msapp.Workbooks.Add()
            $newWS = $newWB.Worksheets[1]
            $newWS.Name = $($datum.Disclaimer)
            If ($includeLabel) {
                $newWS.Range("A1").Value2 = "First Name"
                $newWS.Range("A2").Value2 = "Last Name"
                $newWS.Range("A3").Value2 = "Address"
                $newWS.Range("A4").Value2 = "City"
                $newWS.Range("A5").Value2 = "County"
                $newWS.Range("A6").Value2 = "State"
                $newWS.Range("A7").Value2 = "Zip"
                $newWS.Range("A8").Value2 = "Birthdate"
                $newWS.Range("A9").Value2 = "License"
                $newWS.Range("A10").Value2 = "Passport"
                $newWS.Range("A11").Value2 = "NCCoE Customer Number"
                $newWS.Range("A12").Value2 = "NCCoE Billing Number"
            }
            $newWS.Range("B1").Value2 = $firstName
            $newWS.Range("B2").Value2 = $lastName
            $newWS.Range("B3").Value2 = $($datum.ADDRESS)
            $newWS.Range("B4").Value2 = $($datum.CITY)
            $newWS.Range("B5").Value2 = $($datum.COUNTY)
            $newWS.Range("B6").Value2 = $($datum.STATE)
            $newWS.Range("B7").Value2 = $($datum.ZIP)
            $newWS.Range("B8").Value = $($datum.BIRTHDATE)
            $newWS.Range("B9").Value2 = $($datum.DRIVERS)
            $newWS.Range("B10").Value2 = $($datum.PASSPORT)
            $newWS.Range("B11").Value2 = $($datum.NCCoE_Customer_Number)
            $newWS.Range("B12").Value2 = $($datum.NCCoE_Billing_Number)

            if ($extension.EndsWith('x')) {
                $newWS.SaveAs("$outputFilename", 51) # https://learn.microsoft.com/en-us/office/vba/api/excel.xlfileformat
            } else {
                $newWS.SaveAs("$outputFilename", 56)
            }
            $result = $msapp.Workbooks.Close()
        }
        # pdf assumes that the docx extension has already been run!
        '^pdf$' {
            $sourceSubDir = If($includeLabel) {$outputDirRoot + "\docx\labels" } Else {$outputDirRoot + "\docx\no labels"}

            $doc = $msapp.Documents.Open("$rootDir\$sourceSubDir\$id.docx")
            $doc.SaveAs("$outputFilename", 17)
            $doc.Close()
        }
        # multimedia files
        '^wav$' {
            if ($i % 100 -eq 0) { # only make audio file for every hundredth row of the CSV
                $speech = New-Object System.Speech.Synthesis.SpeechSynthesizer
                $speech.SelectVoice("Microsoft Zira Desktop")
                $speech.SetOutputToWaveFile($outputFilename)
                $speech.Speak($output)
                $speech.Dispose()
                $wroteFile = $true
            } else {
                $wroteFile = $false
            }
        }
        '^png$' { # https://stackoverflow.com/questions/2067920/can-i-draw-create-an-image-with-a-given-text-with-powershell
```

```
                                if ($i % 10 -eq 0) { # only make image file for every tenth row of the CSV
                                    $bmp = New-Object System.Drawing.Bitmap 600,400
                                    $font = New-Object System.Drawing.Font Consolas,14
                                    $nistBlue = [System.Drawing.Color]::FromArgb(61,120,169)
                                    $brushBg = New-Object System.Drawing.SolidBrush $nistBlue
                                    $brushFg = [System.Drawing.Brushes]::White
                                    $graphics = [System.Drawing.Graphics]::FromImage($bmp)
                                    $graphics.FillRectangle($brushBg,0,0,$bmp.Width,$bmp.Height)
                                    $graphics.DrawString($output,$font,$brushFg,10,10)
                                    $graphics.Dispose()
                                    $bmp.Save($outputFilename)
                                    $wroteFile = $true
                                } else {
                                    $wroteFile = $false
                                }
                            }
                            # compressed files; any of these assume that the txt extension has already been run!
                            '^zip$' {
                                $sourceSubDir = If($includeLabel) {$outputDirRoot + "\txt\labels" } Else {$outputDirRoot + "\txt\no labels"}
                                Compress-Archive -Path "$rootDir\$sourceSubDir\$id.txt" -CompressionLevel Fastest -DestinationPath $outputFilename
                            }

                            default {
                                $wroteFile = $false
                                throw "Extension $extension not a supported output type"
                            }
                        }

                        if ($wroteFile) {
                            Write-Host "Wrote record number #$i to $($extension.ToUpper()) file at '$outputFilename'"
                        }

                        $i++
                    }
                }
            }
            finally {
                if ($null -ne $msapp) {
                    $msapp.Quit()
                }
            }
        }
```

D.4.2 Set up folder structure

On the Windows desktop computer

1. Open Windows Explorer
2. Find or create a temporary folder (directory)
3. In the temporary folder, copy the patients.csv file to the folder
4. In the temporary folder, copy the Write-SyntheticData.ps1 PowerShell file to the folder

D.4.3 Run the PowerShell script

On the Windows desktop computer

1. Open Windows Explorer
2. Browse to the temporary folder that was created in procedure D.4.2
3. Right-mouse click on the Write-SyntheticData.ps1 file and from the context menu select **Run with PowerShell**
4. The PowerShell script should start and will commence creating all the necessary file outputs; it can take a significant time to run, depending on how many rows are in the source patients.csv file

5. The output files are created in a sub-folder called “output”; within this folder are sub-folders whose name matches the file extension of the output file, so for example, under “output\xlsx” will be all the created Excel workbooks. Under each extension sub-folder there are also two sub-folders called “labels” and “no labels”; the files in the “labels” sub-folder have line delimited values in the form *fieldname: synthetic data field value* while those in the “no labels” sub-folder has line delimited values in the form *synthetic data field value* (no fieldname/label information included).

Appendix E Lab Implementation Details

Five separate example demonstration lab builds were implemented to show unstructured data classification needs. The example demonstrations used collaborator products that were established in isolated virtual labs with over 60 different virtual machines in total. The virtual lab machines hosted collaborator product installations, unstructured data locations, workstations for accessing data and that simulated user desktop systems and included two complete enterprise-level email infrastructures. When deploying the products into the lab environments the product installation instructions were leveraged where possible, and information regarding the implementation details are provided for reference.

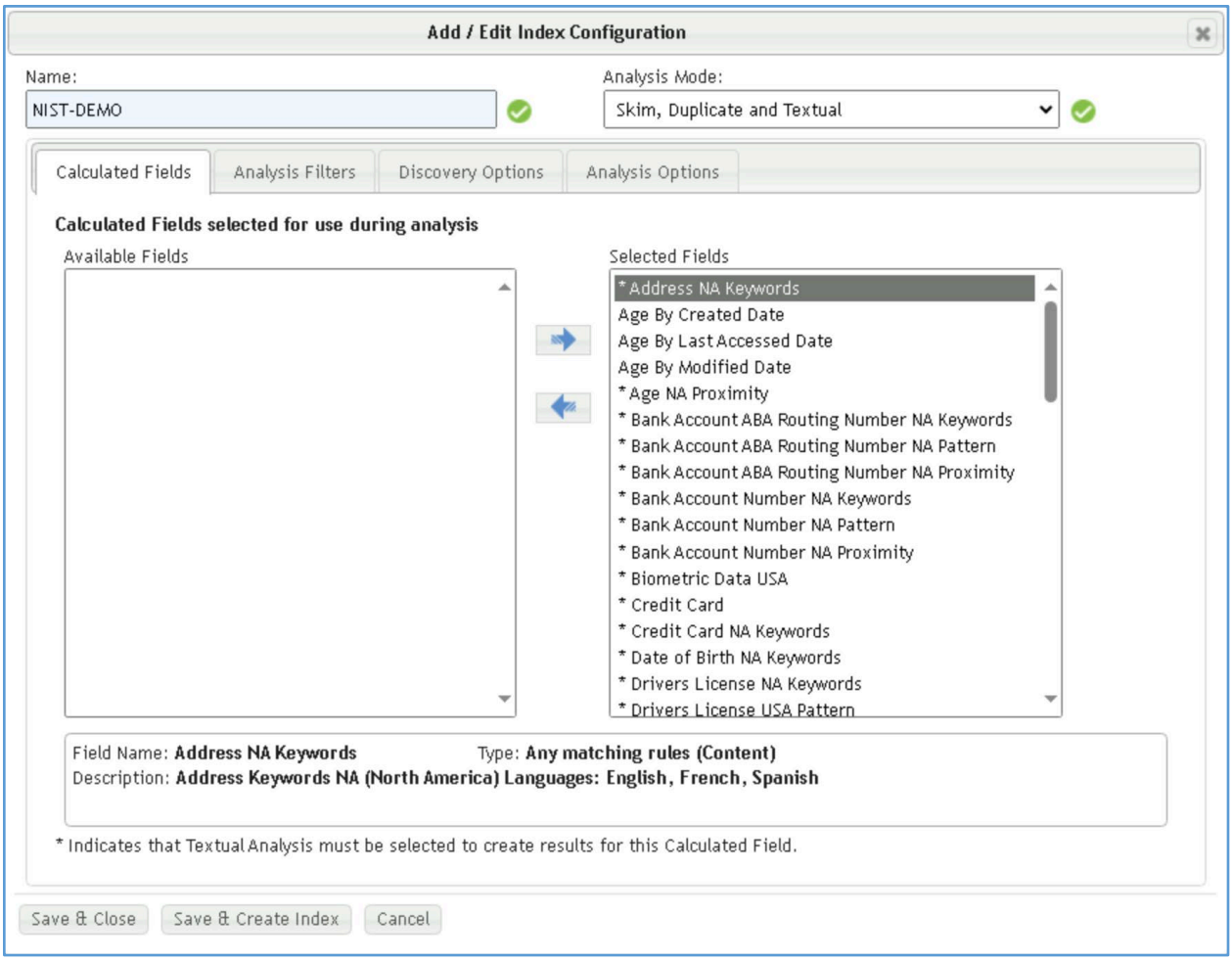
E.1 ActiveNav Implementation Details

ActiveNav Discovery Center finds and classifies unstructured data based on pre-defined and custom classification schemas.

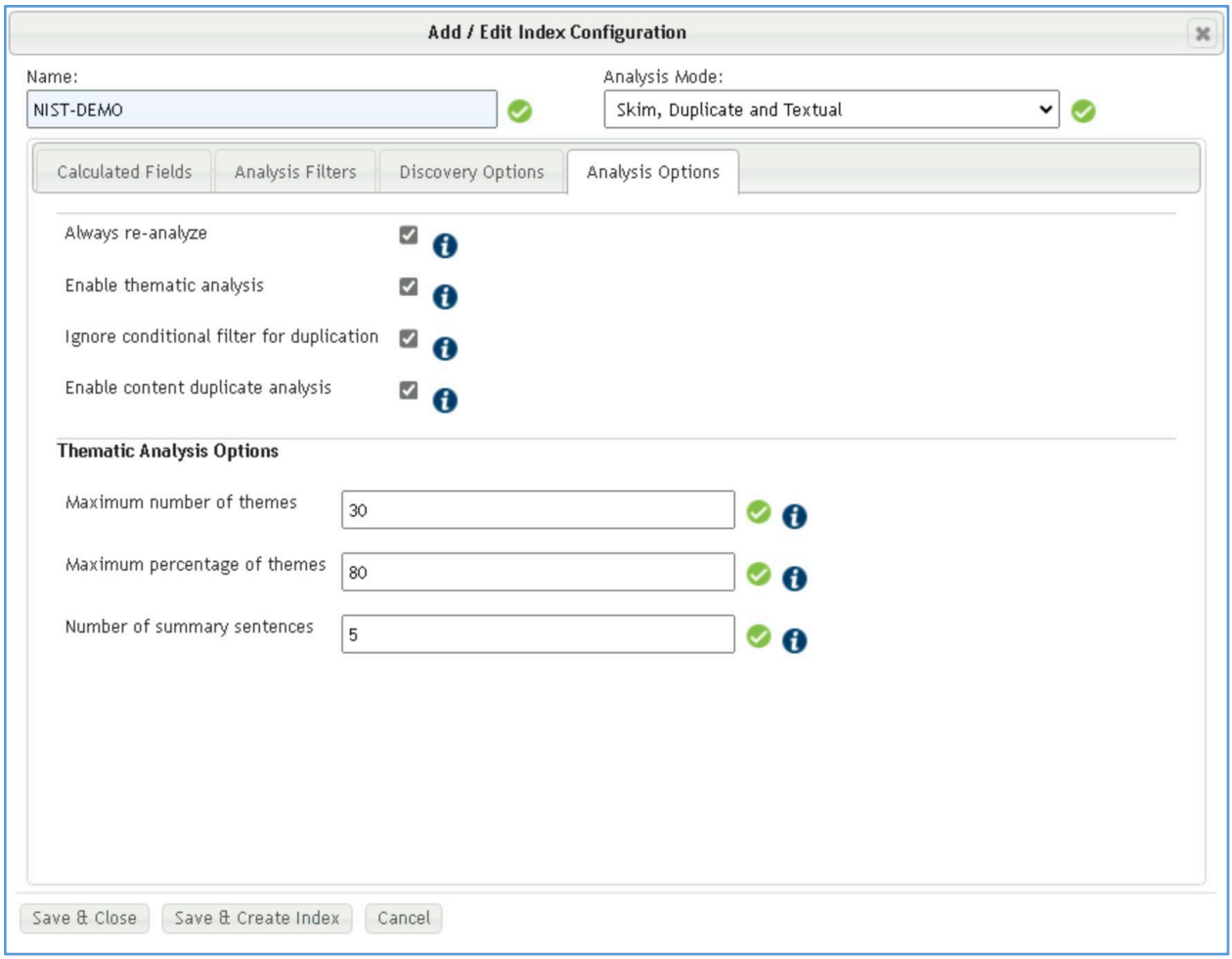
Table E-1 ActiveNav Products

ActiveNav Discovery Center		Installation Notes
Product Information	**Product Name:** Discovery Center Workbench **Product Version:** 4.18.0.0 **Product Name:** Discovery Center Project Suite **Product Version:** 4.18.0.0	Installed per product instructions.
Configuration Information	Installation instructions available at https://support.activenav.com/hc/en-gb/articles/200738982-discovery-center-installation-guide	Synthetic data located in network file share folder.
Classification Schema	For information on creating a custom classification schema, see https://support.activenav.com/hc/en-gb/articles/209409665-discovery-center-starter-presentations	**Schema:** ▪ **Privacy:** Low, Medium, High ▪ **Personal Credit:** Low, Medium, High ▪ **Health Information:** Low, Medium, High

For this build, the on-premises version of ActiveNav Discovery Center Project Suite was used. The data fields selected for discovery and identification were selected using the index configuration tool shown in Figure E-1 Calculated Fields Configuration. In addition to individual field discovery and identification, additional discovery options were configured as shown in Figure E-2 Analysis Options Configuration to enable additional analysis options.



1144 **E-1 Calculated Fields Configuration**



1145 **Figure E-1 Analysis Options Configuration**

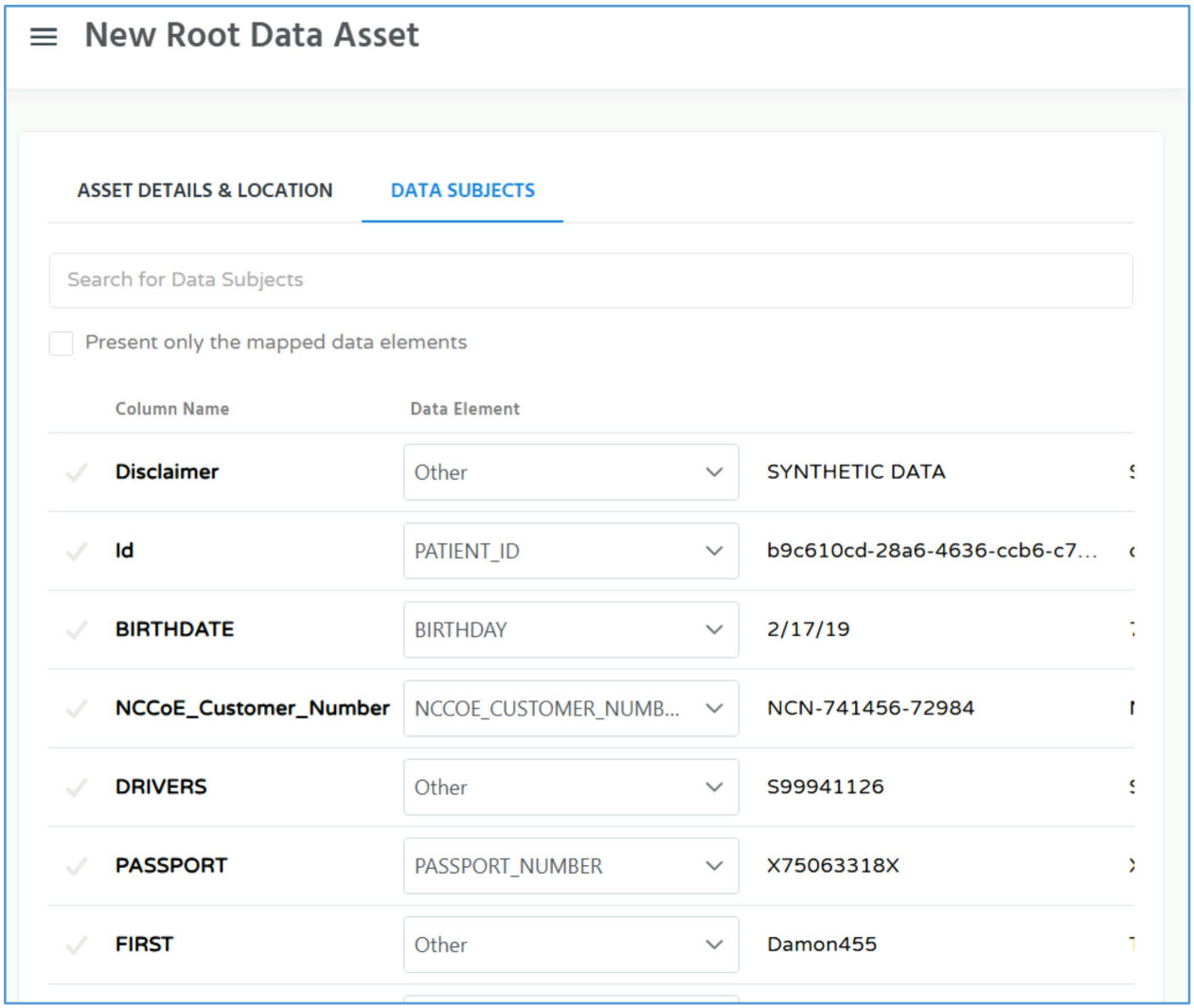
E.2 IBM Implementation Details

IBM Guardium Discover and Classify (IGDC) provides automatic discovery and classification of structured and unstructured data. The demonstration only exercised the discovery and classification of unstructured data.

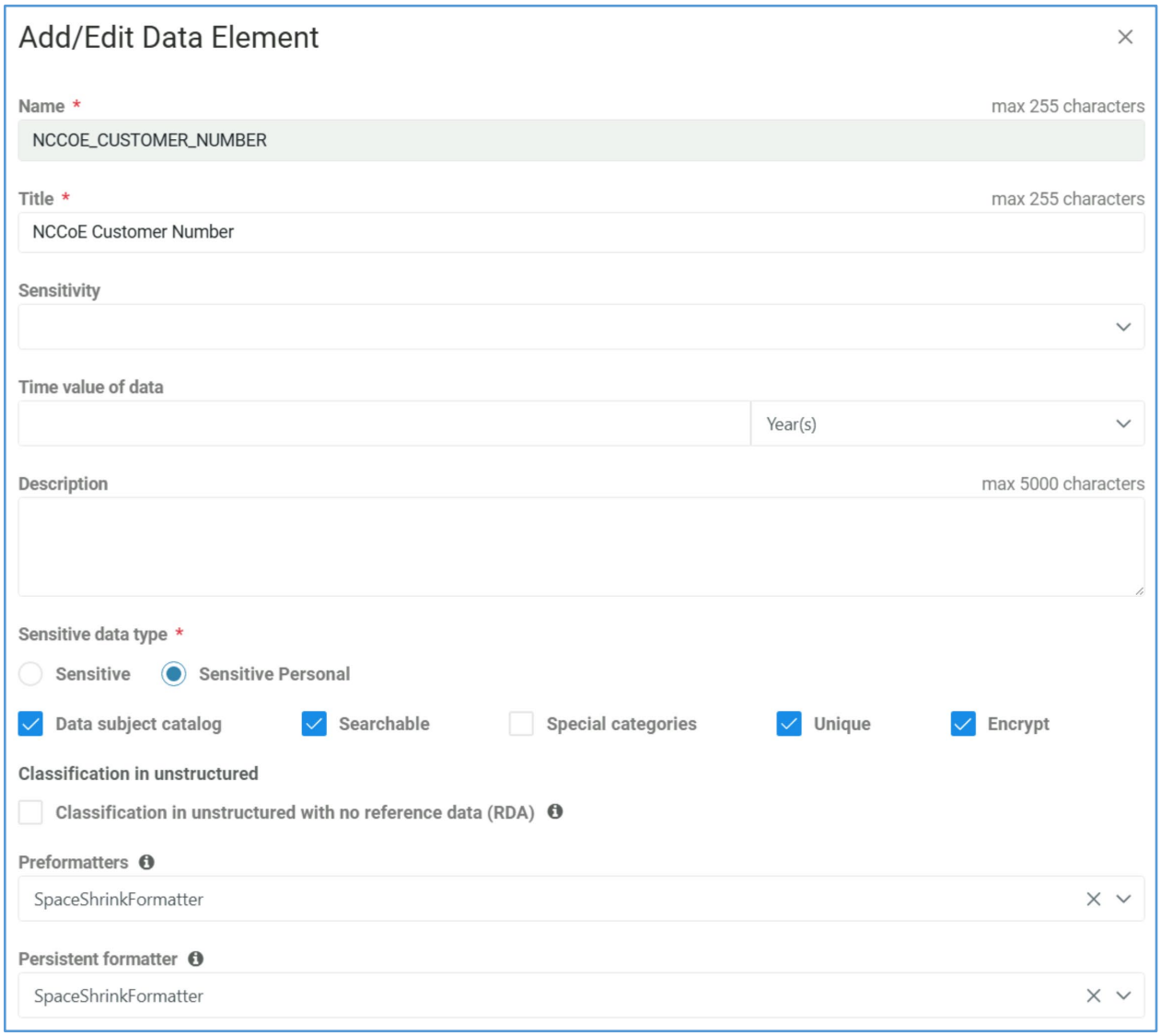
Table E-2 IBM Products

IBM Guardium Discover and Classify		Installation Notes
Product Information	**Product Name:** IBM Guardium Discover and Classify (IGDC) **Product Version:** 4.0.2	Installed per product instructions.
Configuration Information	Installation instructions available at https://www.ibm.com/support/pages/ibm-guardium-discover-and-classify-documentation	Synthetic data located on network storage device.
Classification Schema	For information on creating a custom classification, see https://www.ibm.com/support/pages/ibm-guardium-discover-and-classify-documentation	**Schema:** Sensitive personal across 12 data types including: ▪ Name ▪ Address ▪ Birthdate

The location of the unstructured data, the data fields contained in the files, and the sensitivity levels were configured as shown in Figure E-3 Asset Details and Location and Figure E-4 Add/Edit Data Element.



1154 **Figure E-2 Asset Details and Location**



1155 **Figure E-3 Add/Edit Data Element**

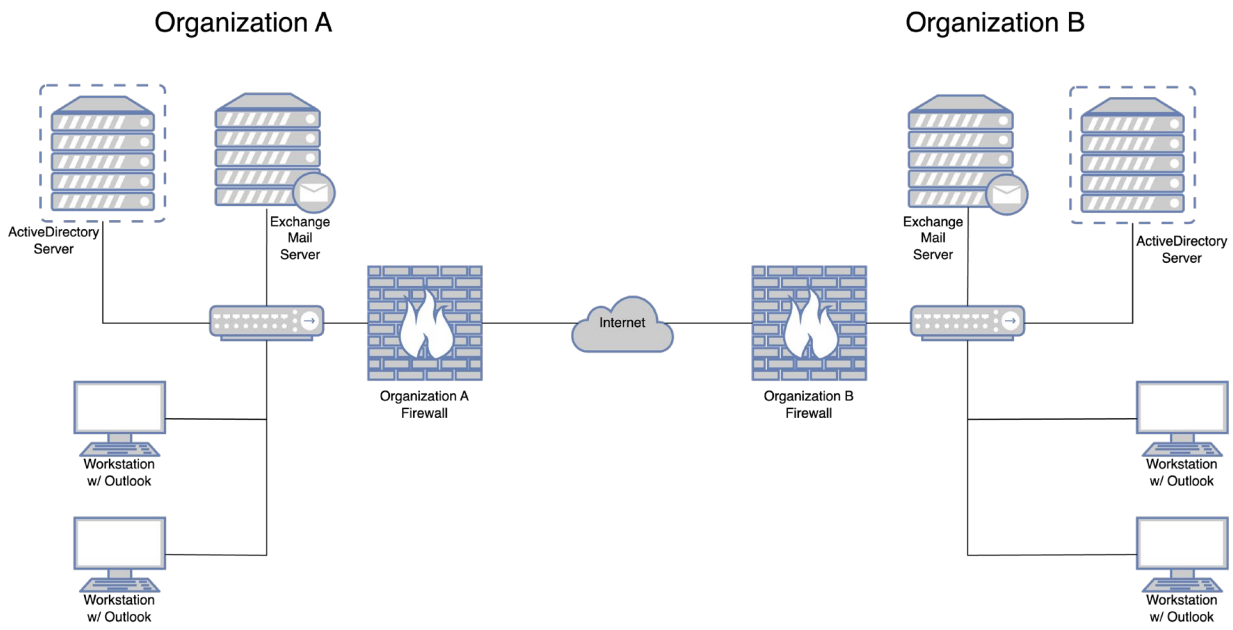
E.3 Janusseal Implementation Details

Janusseal for Outlook identifies sensitive information and enables the application of a classification label to email. Janusseal Schema enables the creation and management of an organization's classification schema. Janusseal was installed in a Microsoft Outlook, Windows and Exchange Server environment.

Table E-3 Janusseal Products

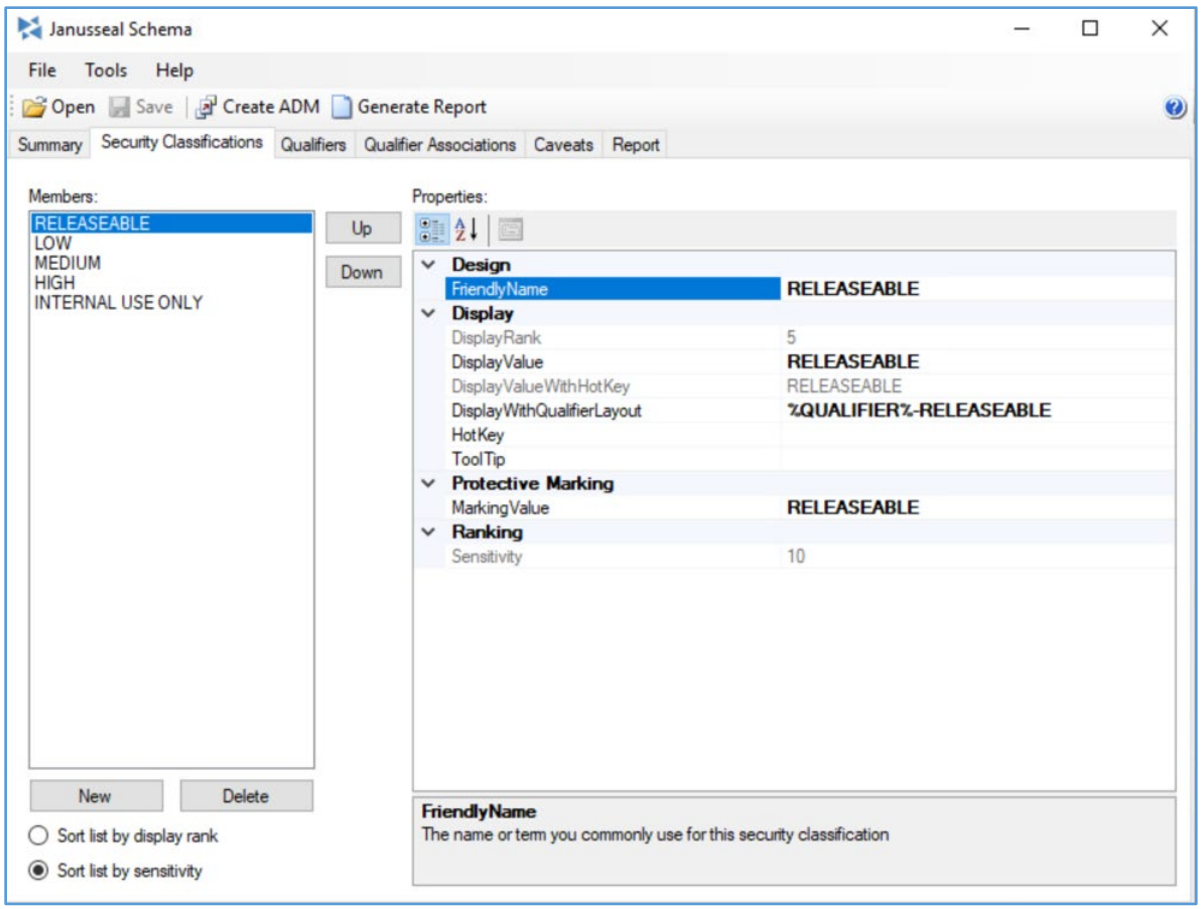
Janusseal Products		Installation Notes
Product Information	**Product Name**: Janusseal Schema **Product Version**: 1.4.4.24270 **Product Name**: Janusseal for Outlook **Product Version**: 3.7.2.28374	Installed per product instructions.
Configuration Information	Installation instructions available at https://www.janusnet.com/	Products installed into two lab environments to create two separate organizations.
Classification Schema	For information on creating a custom classification schema, see https://www.janusnet.com/ Two slightly different schemas were implemented in the lab to create two organizations.	**Organization A Schema:** ▪ Releasable ▪ Low ▪ Medium ▪ High ▪ Internal Use Only **Organization B Schema:** ▪ Releasable ▪ Low ▪ Medium ▪ High ▪ Very High ▪ Not Publicly Releasable

An architecture was developed to demonstrate the classification of unstructured data contained in emails. The architecture enabled both single and multiple organization scenarios, where email could be classified both within a single organization, and between two organizations as shown in Figure E-5 Data Classification Email Reference Architecture.



1165 **Figure E-4 Data Classification Email Reference Architecture**

1166 Using the Janusseal Schema product, separate classification schemas were created for both
1167 organizations as shown in Figure E-6 Janusseal Schema. These schemas were designed to have similar,
1168 but slightly different, naming/ranking values. The schemas implemented are shown in Table E-3
1169 Janusseal Products.



1170 **Figure E-5 Janusseal Schema**

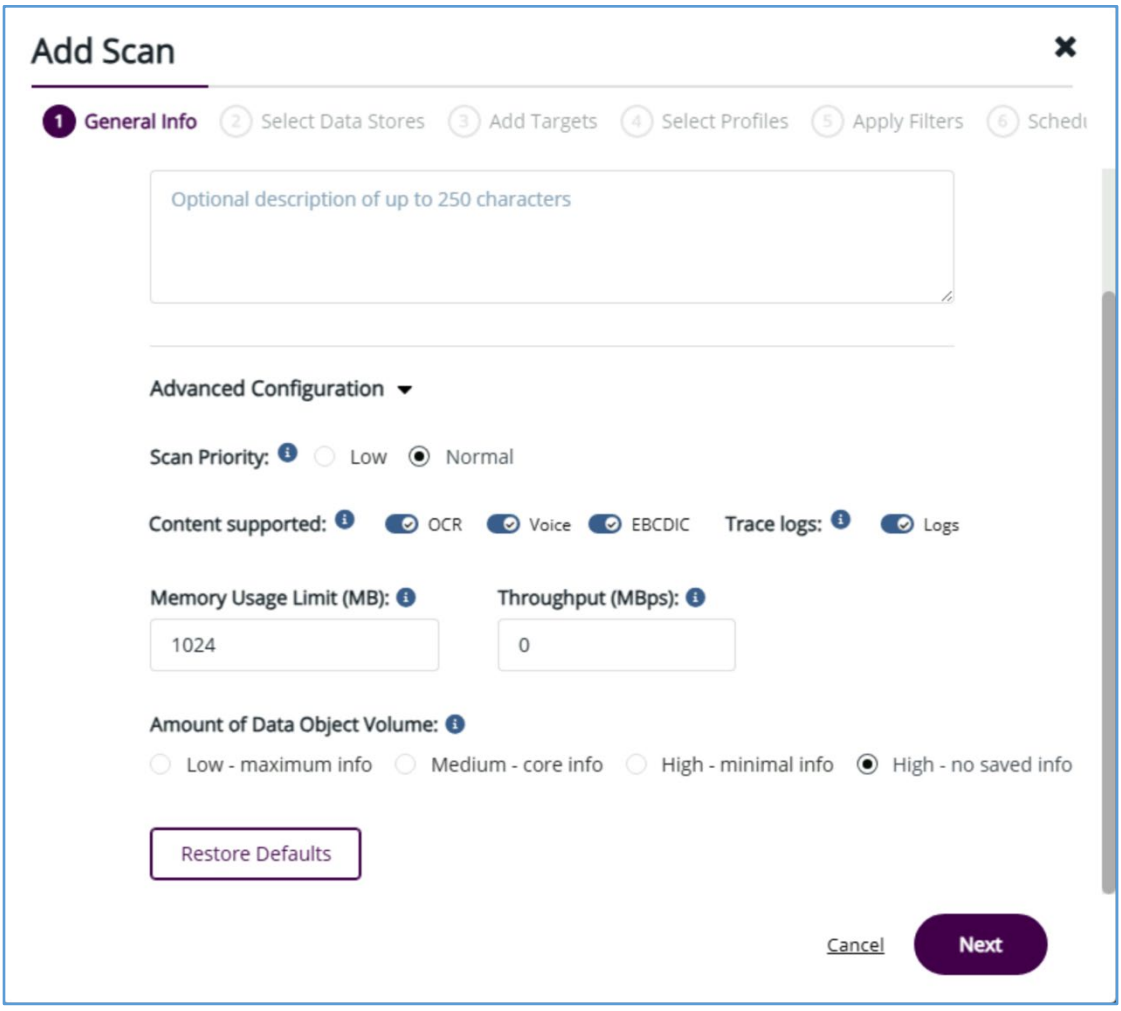
E.4 Thales Implementation Details

The CipherTrust platform includes Thales' data security products.

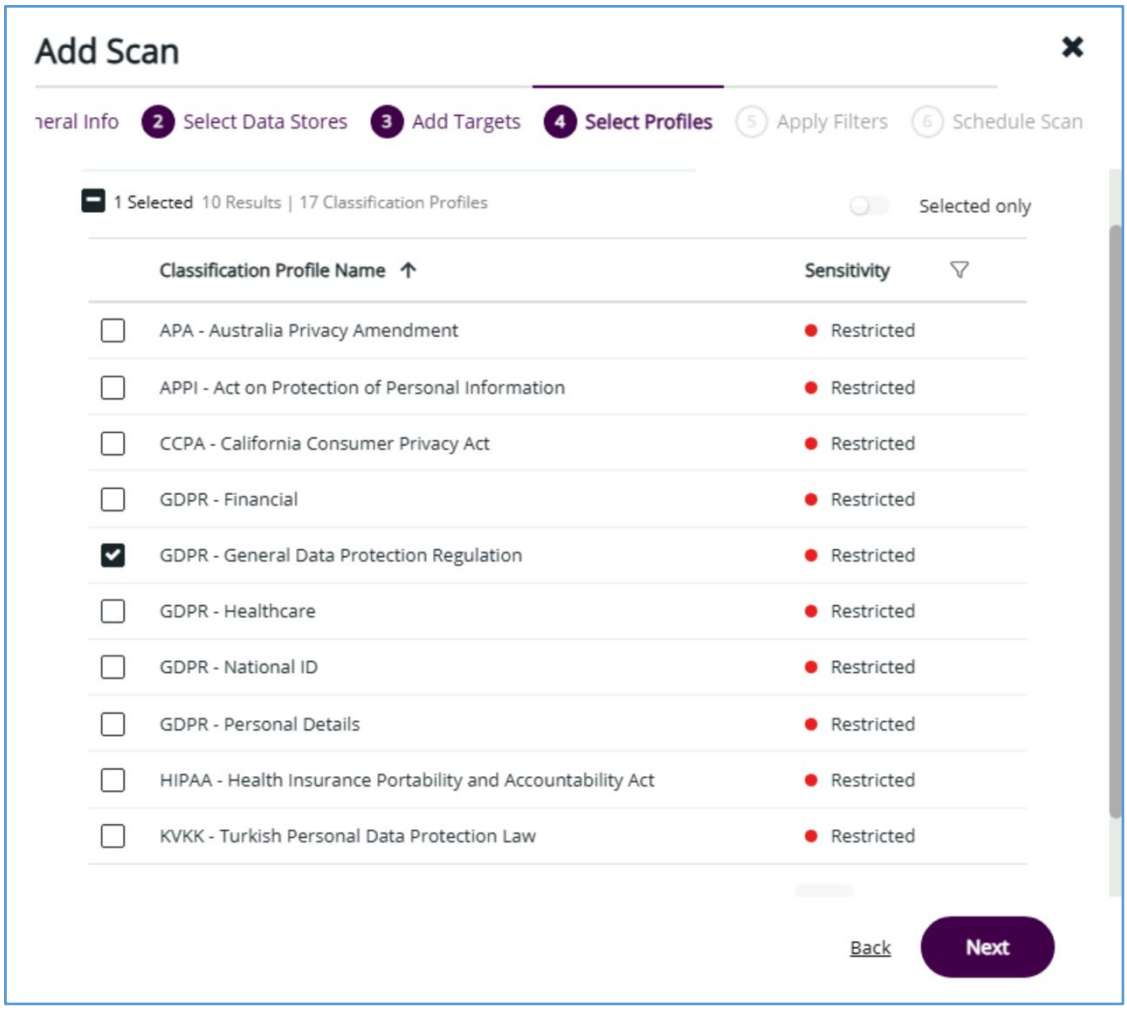
Table E-4 Thales CipherTrust Products

Thales TCT CipherTrust		Installation Notes
Product Information	**Product Name**: Thales CipherTrust Manager **Product Version**: 2.17.0+12772 **Product Name**: Thales Data Discovery and Classification TDP **Product Version**: 3.1.6-316	Installed per product instructions.
Configuration Information	Installation instructions available at https://www.thalestct.com/ciphertrust-data-security-platform/ https://www.thalesdocs.com/ctp/ddc/index.html	Synthetic data located on network storage device.
Classification Schema	For information on creating the classification schema, see https://www.thalestct.com/ciphertrust-data-security-platform/	**Schema:** ▪ Personal data ▪ Medical ▪ Financial

The file types to be scanned by the Thales TCT CipherTrust tools, their classification profiles and sensitivity were selected as shown in Figure E-7 Advanced Configuration Options and Figure E-8 Classification Profile Names and Sensitivity.



1177 **Figure E-6 Advanced Configuration Options**



1178 **Figure E-7 Classification Profile Names and Sensitivity**

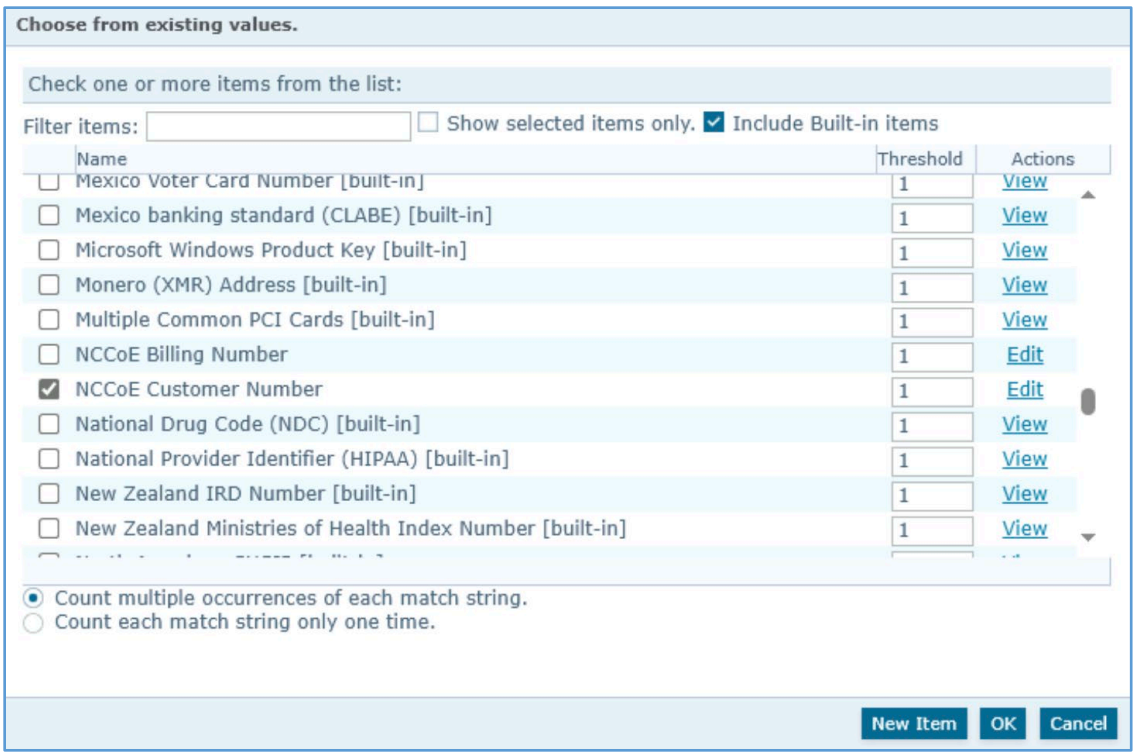
E.5 Trellix Implementation Details

Trellix Data Loss Prevention (DLP) supports agent-based and agentless discovery and classification of data-at-rest on workstation endpoints and network resources.

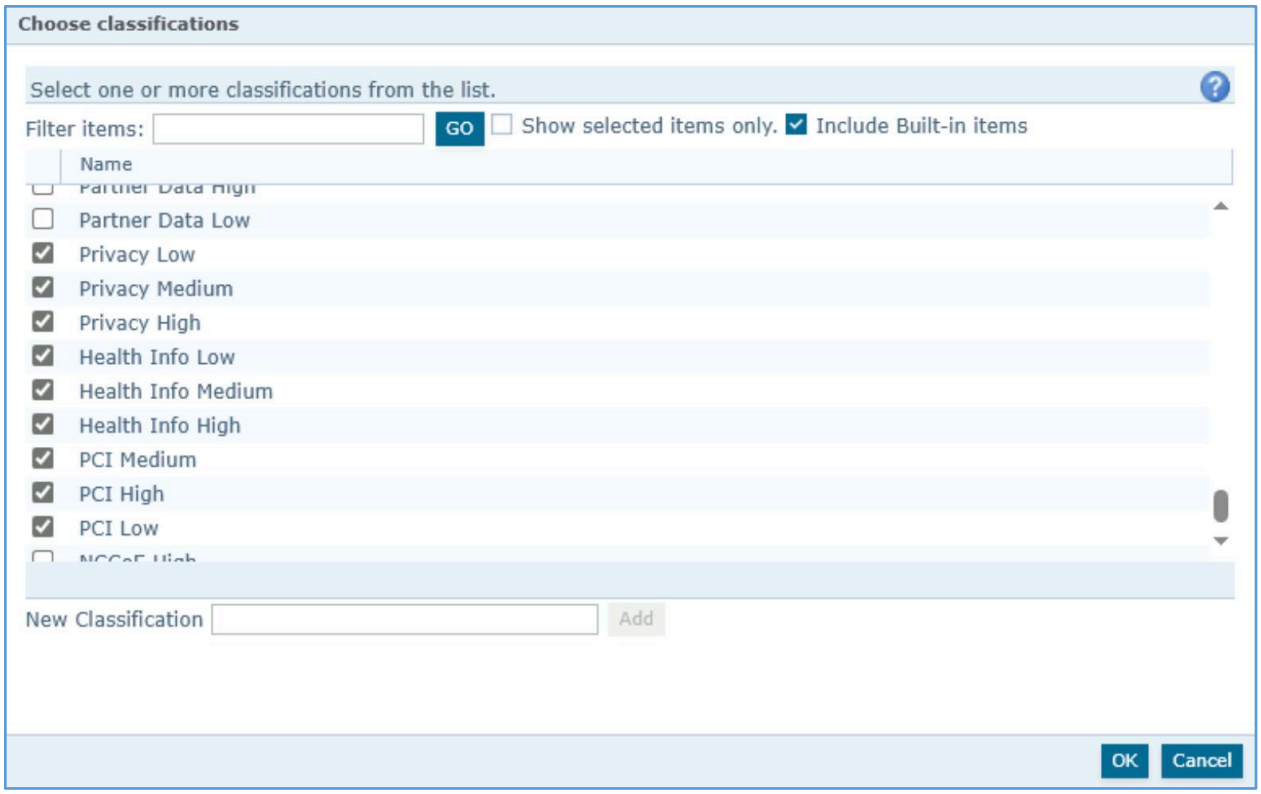
Table E-5 Trellix Products

Trellix Data Loss Prevention		Installation Notes
Product Information	**Product Name**: Data Loss Prevention Discover Server **Product Version**: 11.10.500.112 **Product Name**: Trellix ePolicy Orchestrator **Product Version**: 5.10.0	Installed per product instructions.
Configuration Information	Trellix Data Loss Prevention Discover Server installation instructions can be found at https://docs.trellix.com/bundle Trellix ePolicy Orchestrator installation instructions can be found at https://docs.trellix.com/bundle/epolicy-orchestrator-landing/page/UUID-52b91793-68f9-a8ac-141c-104824763f9a.html	Synthetic data located on network storage device.
Classification Schema	For information on creating a custom classification schema, see https://docs.trellix.com/bundle/data-loss-prevention-11.1.x-product-guide/page/GUID-F7DAE857-DC7C-44DA-AEB4-BFFB1280FD0B.html	**Schema:** ▪ Privacy Low ▪ Privacy Medium ▪ Privacy High ▪ Health Info Low ▪ Health Info Medium ▪ Health Info High ▪ PCI Low ▪ PCI Medium ▪ PCI High

Scan filters were applied to identify individual data fields in the synthetic unstructured data as shown in Figure 0-9 Scan Filter Configuration. Sensitivity of the scanned data was configured using a filter as shown in Figure 0-10 Schema Classification Settings.



1186 **Figure E-8 Scan Filter Configuration**



1187 **Figure E-9 Schema Classification Settings**