Administering Information Protection and Compliance in Microsoft 365

Concise Study Notes

# Implement Information Protection in Microsoft 365

## Introduction to information protection and data lifecycle management

**Defining an Information Protection Strategy**

* Key questions:
  + Do you know the whereabouts and usage of your critical data?
  + Can you control data shared within and outside your organization?
  + Are multiple solutions being used for data protection?

**Microsoft's Data Protection and Governance Solutions**

* Comprehensive approach: Covers data in Microsoft 365 services, on-premises resources, non-Microsoft clouds, and SaaS apps.
* Aims to protect and govern data, wherever it resides or travels.

**Unified Approach to Data Discovery and Classification by Microsoft**

* Consistent classification across all platforms.
* Integrated into applications and services.
* Robust content scanning with over 90 sensitive data types.
* Support for custom data types and trainable classifiers.

**Balancing Security and Productivity**

* **Security Capabilities**: Conditional access, data loss prevention, encryption, and data leakage prevention.
* **Productivity Capabilities**: Seamless experiences in Office apps, flexible search, coauthoring, and built-in labeling.

**Information Protection and Governance Lifecycle**

* **People Involved**:
  + CISO: Heads governance committee.
  + Compliance Officer: Understands regulations.
  + Data Admin: Creates classification policies.
  + Data Owner: Owns content/processes.
  + Help Desk: Assists users.
  + Information Worker: Creates documents and emails.
* **Process Phases**:
  + Define data classification taxonomy.
  + Define classification policy conditions.
  + Test and deploy labels/policies.
  + Monitor and adjust policies.
* **Technology**: Microsoft provides robust solutions for data classification, protection, and Data Lifecycle Management.

**Microsoft's Information Protection and Governance Strategy: An Overview**

Microsoft's approach to data management, protection, and governance is built upon a robust framework that is holistic, flexible, and intelligent. Here's a concise breakdown of the strategy and its key components:

1. **Core Principles**:
   * **Know your data**: Understand the data you have, its types, locations, and relevance.
   * **Protect your data**: Implement encryption, access restrictions, and other protection measures.
   * **Prevent data loss**: Monitor data movement and prevent unauthorized sharing.
   * **Govern your data**: Enforce retention, deletion, and compliant storage rules.

**Enabling Tools and Features**:

* + **Classification**: An integral process to understand your data. Employ sensitive information types, trainable classifiers, labels, and policies.
  + **Sensitive Information Types**: Predefined data patterns like credit card numbers to detect sensitive data.
  + **Trainable Classifiers**: Use AI and machine learning for custom data classification.
  + **Labels**: "Stamps" like "Confidential" to categorize data, offering two forms—sensitivity labels and retention labels.
  + **Policies**: Create custom rules to manage data based on classification inputs.

**User Experience and Automation**:

* + **In-Office Classification**: Directly label and classify data within Office apps like Word, PowerPoint, Excel, and Outlook.
  + **Manual and Automated Labeling**: Users can manually label data or depend on automated processes in Office on Windows, the web, and other platforms.
  + **Auto-Labeling in Microsoft 365 Services**: Implement policies for automatic label application in OneDrive, SharePoint, and Exchange.

**Discovery and Classification Across Platforms**:

* + **Microsoft 365 Content**: Identify and classify data across Exchange emails, SharePoint sites, OneDrive accounts, and Teams messages.
  + **On-Premises Files**: With Azure Information Protection scanner, classify and protect sensitive files on local servers.
  + **Third-Party Clouds and SaaS**: With Microsoft Defender for Cloud Apps, extend protection outside the Microsoft ecosystem, inspecting files in third-party cloud services and applying necessary protections.

**Integrated Approach**:

* + **Unified Policy Management**: Centralize data protection rules and policies.
  + **Analytics Dashboard**: Oversee data movements, access, and issues.
  + **APIs**: Extend protection to custom apps and services.

1. **Continuous Process**: The whole approach is not a one-time task but an ongoing process that needs to be refined and updated based on changing data and threats.

**Protect Your Data**

* **Purpose**: Prevent sensitive data from getting into the wrong hands using information protection.
* **Discover, Classify, Protect**: Information protection aids in identifying, classifying, and safeguarding sensitive content.
* **Differential Protection**: Adjust protection levels based on the sensitivity of the data.
* **Features**:
  + **Built-in experiences**: Integrated into Microsoft 365 apps and services.
  + **Broad coverage**: Protects content in Microsoft 365, devices, on-premises, and third-party services.
  + **Flexible labeling**: Options include automatic, manual, and recommended labeling.

**Prevent Data Loss**

* **Purpose**: Safeguard sensitive information and prevent accidental disclosure.
* **Data Loss Prevention (DLP)**: Identifies sensitive content, stopping inadvertent oversharing.
* **Considerations**:
  + **Data source**: Adjust responses based on user roles.
  + **Data destination**: Modify responses depending on internal or external sharing.
  + **Amount shared**: The quantity of data shared influences reactions.
  + **Exposure impact**: Responses range from tool tips to blocking content sharing.

**Govern Your Data**

* **Purpose**: Retain essential data and discard unnecessary data to comply and reduce risk.
* **Solutions**:
  + **Data Lifecycle Management**: Manages content lifecycle to retain vital and discard unneeded data.
  + **Records Management**: Uses intelligent classification for retention of immutable records.
* **Features**:
  + **Streamlined administration**: Centralized policy management for data inside and outside Microsoft 365.
  + **Automation at scale**: Configures automatic governance policies organization-wide.
  + **Tailored workflows**: Flexible retention policies, including custom triggers and disposition reviews.

## Classify data for protection and governance

**Understanding Sensitive Information Types (SITs) in Microsoft Purview**

Sensitive Information Types (SITs) in Microsoft Purview play a pivotal role in helping organizations identify and manage their sensitive data. The main objective is to help organizations stay compliant with various regulations and to mitigate the risk of data breaches.

Here's a quick summary of what you need to know about SITs:

1. **What is an SIT?** Sensitive Information Types are patterns or classifiers that detect and identify sensitive data like credit card numbers, social security numbers, and more.
2. **Types of SITs:**
   * **Built-in:** Microsoft offers predefined SITs for commonly recognized sensitive data.
   * **Named entity:** These SITs automatically identify specific types of named entities, such as person names or addresses.
   * **Custom:** If the predefined ones don't meet your needs, you can create your own.
   * **Exact data match (EDM):** This allows for precise matching against a database of sensitive info.
3. **How SITs Work:** Each SIT consists of:
   * **Name:** Identifies the type.
   * **Description:** Explains what it detects.
   * **Pattern:** Defines specific criteria indicating the presence of sensitive data. This includes:
     + **Primary element:** The main defining criteria (can be a regular expression, keyword list, etc.).
     + **Supporting element:** Supports the primary element to increase the confidence of a match.
     + **Confidence Level:** Indicates the certainty of a match.
     + **Proximity:** Shows how close the primary and supporting elements are to each other.
4. **Creating Custom SITs:** Microsoft Purview provides options to create tailored SITs:
   * **Using the UI:** A more straightforward method involving regular expressions and keywords.
   * **Using EDM:** Allows the creation of a dynamic SIT using a secure database.
   * **Using PowerShell:** Offers more configuration options but is more complex.
5. **Feedback Mechanism:** Microsoft Purview provides an option to view matches and give feedback on the accuracy of an SIT. This helps in refining and ensuring precise classification.

**Benefits of Using SITs in Microsoft Purview**

* **Data Protection:** Identify sensitive data, thus allowing for its protection through encryption, access restrictions, or other methods.
* **Regulatory Compliance:** Ensure that sensitive data is handled as per regulatory requirements, reducing the risk of penalties or sanctions.
* **Risk Management:** By identifying and managing sensitive data, organizations can reduce the risk of data breaches and associated reputational damage.
* **Operational Efficiency:** Automated identification and management of sensitive data can lead to reduced manual effort and more streamlined operations.

In conclusion, SITs in Microsoft Purview provide a robust framework for organizations to classify and manage their sensitive data, leading to better data governance, protection, and compliance management. It's essential for organizations to leverage these capabilities effectively to secure their data in today's complex digital landscape.

**Understanding Data Classification Using Trainable Classifiers in Microsoft 365**

Data classification is crucial for any organization as it enables the protection of sensitive data and helps in meeting compliance requirements. Microsoft 365 offers a powerful method for classifying data using trainable classifiers, leveraging the capabilities of Artificial Intelligence (AI) and Machine Learning (ML). Here's a guide to understanding and using this feature:

**What is a Trainable Classifier?**

A trainable classifier in Microsoft 365 is a tool that identifies specific types of data. It does this not based on simple pattern matching (like looking for credit card numbers), but by understanding the type of content.

**How Does It Work?**

There are two types of classifiers:

1. **Pre-Trained Classifiers:** These are classifiers provided by Microsoft, already trained on specific types of data and ready to be used right out of the box.
2. **Custom Trainable Classifiers:** These are classifiers that you train on your data, tailored for unique types of content specific to your organization.

**Implementing Custom Trainable Classifiers:**

**1. One-time setup:**

* Before creating classifiers, a scan must be done to understand the content in your organization. This process takes 7-14 days.

**2. Seed:**

* **Prepare sample data:** This is the initial data that you use to "teach" the classifier what to look for. You should provide between 50-500 samples.
* **Create the classifier:** Using the Microsoft Purview compliance portal, create a new classifier, providing the sample data. This process can take up to 24 hours.

**3. Test:**

* **Prepare test data:** This is a different set of data, used to test the classifier's accuracy. This should have at least 200 items, but no more than 10,000.
* **Test the model:** Using the test data, test the classifier's prediction capabilities.
* **Evaluate predictions:** Review the predictions made by the classifier to check its accuracy. Review at least 200 items.

**4. Publish:** Once satisfied with the classifier's accuracy, publish it, making it available for use across Microsoft 365 compliance solutions.

**Practical Applications:**

1. **Legal Documents:** If you often handle contracts, you can train a classifier to detect these even if they're in non-standard formats or have varying language.
2. **Strategic Business Documents:** Documents that detail business strategies or processes can be classified to ensure they're appropriately protected.
3. **Pricing and Financial Information:** Such data is critical and often regulated. Training a classifier ensures you're not accidentally exposing or mishandling this information.

**Licensing Requirements:**

Using trainable classifiers in Microsoft 365 requires specific licenses:

* Microsoft 365 E5
* Microsoft 365 E5 Compliance
* Microsoft 365 E5 Information Protection and Governance

**In Conclusion:**

Trainable classifiers offer a flexible and robust method for data classification, especially when dealing with diverse and nuanced types of content. By leveraging AI and ML, Microsoft 365 provides organizations the tools they need to classify data effectively, ensuring protection and compliance in a dynamic digital landscape.

**Trainable Classifiers in Microsoft 365**:

* **Purpose**: To detect content that might not be identifiable through traditional pattern matching.
* **Types**:
  1. Pre-Trained Classifiers: Ready-made classifiers provided by Microsoft.
  2. Custom Trainable Classifiers: User-created classifiers for organization-specific data.
* **Process**:
  1. **Seed**: Prepare and submit sample data.
  2. **Test**: Use test data to check the predictive model's accuracy.
  3. **Publish**: Make the classifier available for use.

**Review Sensitive Information and Label Usage**:

* The data classification Overview page provides insights into how sensitive information and labels are being utilized in an organization.
* **Cards Featured**:
  1. Top sensitive info types
  2. Top sensitivity labels applied to content
  3. Top retention labels applied to content
  4. Top activities detected
  5. Locations where sensitivity labels are applied
  6. Locations where retention labels are applied
  7. Azure Information Protection labels summary

**Content Explorer**:

* Provides a snapshot of items with sensitivity or retention labels.
* Offers visibility into the amount and context of sensitive data in documents.

**Activity Explorer**:

* Gives insights into document-level activities.
* Monitors actions like label changes, file modifications, etc.
* Useful for understanding the behavior surrounding labeled content over time.

**Interactive Guides**:

* Microsoft offers interactive guides to help users better understand and use these features:
  1. Identifying content using trainable classifiers
  2. Monitoring the use of sensitive information

**Licensing**:

* The features are included with Microsoft 365 E5, Microsoft 365 E5 Compliance, and Microsoft 365 E5 Information Protection and Governance packages. Proper licensing is essential.

## Create and manage sensitive information types

**Built-in Sensitive Information Types**

1. **Definition**: Microsoft 365 has over 100 predefined types to identify commonly found sensitive data like credit card numbers, SSNs, etc.
2. **Implementation**: Uses regular expressions or functions to detect data patterns.
3. **Advantages**:
   * Quick to deploy without needing extensive setup.
   * Backed by Microsoft's vast experience and data matching patterns.
4. **Limitations**: Might not cater to specific organizational needs or unique data types.
5. **Recommendation**: Always start with these for general data protection.

**Custom Sensitive Information Types**

1. **Definition**: Organizations can define their types for unique sensitive data, like employee IDs or proprietary project numbers.
2. **Components**:
   * **Primary pattern**: Uses regex to detect data patterns.
   * **Additional evidence**: Keywords or patterns supporting the primary pattern.
   * **Character proximity**: Determines the closeness of the primary pattern and evidence.
   * **Confidence level**: Indicates accuracy of detection.
3. **Special Features**:
   * **Exact Data Match (EDM)**: Matches exact values from a database.
   * **Document Fingerprinting**: Converts a standard form into a sensitive information type.
   * **Keyword dictionaries**: Manages keyword lists for large-scale detection.
4. **Advantages**:
   * Highly customizable to specific organizational needs.
   * Incorporates advanced detection methods like EDM.
5. **Limitations**: Requires more effort in setup and maintenance.
6. **Recommendation**: Use when built-in types aren't sufficient. Begin with built-in types, then implement custom types for specific needs, and then leverage advanced features for enhanced accuracy and management.

**Creation of Custom Sensitive Information Types with EDM**

* Enables classification based on exact values in a large database.
* Process:
  1. Set up EDM classification.
  2. Hash and upload the data.
  3. Integrate with Microsoft cloud services.

**Document Fingerprinting: A Comprehensive Overview**

**Introduction** Document fingerprinting is an essential feature offered by Microsoft Purview, enabling organizations to detect and prevent the unauthorized transmission of sensitive documents. This article provides insights into the concept, the associated features, the mechanism of action, its applications, and guidelines to create a document fingerprint.

**Features of Document Fingerprinting**

* Utilization by Data loss prevention (DLP) in platforms like Exchange, SharePoint, OneDrive, Teams, and Devices.
* Management via the Microsoft Purview compliance portal.
* Supports both partial and exact matching.
* Enhanced detection accuracy.
* Multi-language detection, inclusive of dual-byte languages.

**Scenario** Consider a blank patent template. A document fingerprint of this template can be created to detect any outgoing filled patent templates. DLP policies can then block or warn senders, ensuring sensitive data, like patents, aren't sent inappropriately. This method is applicable to various forms like government or HIPAA compliance forms, HR forms, and other organizational specific forms.

**Mechanism** The name "fingerprinting" derives from the unique patterns in documents, akin to human fingerprints. Document fingerprinting recognizes these unique word patterns. When a template is uploaded:

* DLP pinpoints the word pattern.
* A unique hash value representing this text pattern is created and stored in a Unicode XML file, known as the document fingerprint. This hash cannot recreate the original document.
* This fingerprint is then associated with DLP policies to detect similar patterns in outbound communications.

**Applications**

1. **Detection and Blocking**: For instance, an organization might wish to block regular employees from sending patents but allow the legal department due to business reasons.
2. **Policy Tips and Overrides**: When a potential breach is detected, senders can be notified, and in certain cases, given an option to override with a justification.

**Supported File Types**

* Document fingerprinting supports numerous file types except for .dotx (Microsoft Word template files). When "template" is mentioned, it typically refers to standard forms and not the file type.

**Limitations** Document fingerprinting does not work for:

* Password-protected files.
* Image-only files.
* Documents not fully aligned with the original template.
* Files larger than 4 MB.

**Partial Matching** Different confidence levels (Low, Medium, High) can be set, dictating the percentage of text matching the fingerprint:

* **Low**: Inclusive of all levels, highest false positives.
* **Medium**: A balanced level.
* **High**: Fewest false positives, more false negatives.

**Exact Matching** To detect files exactly resembling the fingerprint, "Exact" must be chosen for high confidence. This ensures word-for-word matching but might miss slightly deviated documents.

**Creating a Document Fingerprint**

1. Access the Microsoft Purview compliance portal.
2. Navigate to Data classification > Classifiers > Sensitive info types > Create Fingerprint based SIT.
3. Input a name and description.
4. Upload the chosen file.
5. (Optional) Set requirements for confidence levels.
6. Review and confirm.

**Additional Notes**

* Using document fingerprinting for devices necessitates enabling Advanced classification scanning and protection.
* Fingerprints are saved in a rule pack, limited to 150 KB, accommodating approximately 50 fingerprints per tenant.

For further details on creating DLP policies using fingerprints, refer to the appropriate documentation.

**Named Entities in Data Classification**

Named entities are crucial elements in data classification and detection. They refer to identifiable information that has potential sensitivity, like personal names, addresses, or medical conditions. These entities play a vital role in data loss prevention (DLP) and regulatory compliance, helping organizations identify and protect sensitive data.

**Key Takeaways:**

1. **Functionality**: Named entities work by identifying specific types of sensitive information in documents or communications.
2. **Use Cases**:
   * Preventing data leakage by detecting and potentially blocking content with sensitive information.
   * Ensuring compliance with regulations, e.g., HIPAA.
3. **Types of Named Entity SITs**:
   * **Unbundled**: Narrower focus, like addresses from a specific region.
   * **Bundled**: Broader scope, e.g., All physical addresses or all medical terms.
4. **Examples**:
   * All full names: Detects possible name patterns.
   * Australia physical addresses: Specific to addresses from Australia.
   * Blood test terms: Specific to blood test-related terms like 'hCG'.
5. **Enhanced DLP Policies**: DLP offers pre-configured policy templates using named entities which can be customized to fit organizational needs.
6. **Keyword Dictionaries**:
   * Efficient way to manage extensive lists of terms.
   * Supported by files like .txt and .csv.
   * Can be created from the Microsoft Purview compliance portal or using PowerShell.
   * Keyword dictionaries play a vital role in data classification. For instance, an organization can use a keyword dictionary to classify and detect prohibited words in an educational environment or specific industry jargon in a business setting.

**Benefits**:

1. **Precision**: By focusing on named entities, organizations can precisely detect and manage sensitive information.
2. **Flexibility**: Named entities can be bundled for broad detection or unbundled for more precise checks.
3. **Compliance**: Assists organizations in meeting regulatory requirements and industry standards.

**Best Practices**:

1. **Collaboration**: Engage multiple departments, like HR and legal, to ensure comprehensive keyword dictionaries.
2. **Regular Revision**: Modify and update keyword dictionaries as terms and organizational needs evolve.
3. **Integration**: Use named entities in tandem with DLP policies for a multi-layered approach to data protection.