Veritas Data Insight Classification Guide

6.5.1

June 2023

VERITAS™

Veritas Data Insight Classification Guide

Documentation version: 6.5

PN:

Legal Notice

Veritas Technologies LLC
2625 Augustine Drive.
Santa Clara, CA 95054

http://www.veritas.com

.

Technical Support

Technical Support maintains support centers globally. Technical Support's primary role is to respond to specific queries about product features and functionality. The Technical Support group also creates content for our online Knowledge Base. The Technical Support group works collaboratively with the other functional areas within the company to answer your questions in a timely fashion.

Our support offerings include the following:

- A range of support options that give you the flexibility to select the right amount of service for any size organization
- Telephone and/or Web-based support that provides rapid response and up-to-the-minute information
- Upgrade assurance that delivers software upgrades
- Global support purchased on a regional business hours or 24 hours a day, 7 days a week basis
- Premium service offerings that include Account Management Services

For information about our support offerings, you can visit our website at the following URL:

www.veritas.com/support

All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policy.

Contacting Technical Support

Customers with a current support agreement may access Technical Support information at the following URL:

www.veritas.com/support

Before contacting Technical Support, make sure you have satisfied the system requirements that are listed in your product documentation. Also, you should be at the computer on which the problem occurred, in case it is necessary to replicate the problem.

When you contact Technical Support, please have the following information available:

- Product release level
- Hardware information
- Available memory, disk space, and NIC information

- Operating system
- Version and patch level
- Network topology
- Router, gateway, and IP address information
- Problem description:
  - Error messages and log files
  - Troubleshooting that was performed before contacting Technical Support
  - Recent software configuration changes and network changes

Licensing and registration

If your product requires registration or a license key, access our technical support Web page at the following URL:

www.veritas.com/support

Customer service

Customer service information is available at the following URL:

www.veritas.com/support

Customer Service is available to assist with non-technical questions, such as the following types of issues:

- Questions regarding product licensing or serialization
- Product registration updates, such as address or name changes
- General product information (features, language availability, local dealers)
- Latest information about product updates and upgrades
- Information about upgrade assurance and support contracts
- Advice about technical support options
- Nontechnical presales questions
- Issues that are related to CD-ROMs, DVDs, or manuals

Support agreement resources

If you want to contact us regarding an existing support agreement, please contact the support agreement administration team for your region as follows:

Worldwide (except Japan)	CustomerCare@veritas.com
Japan	CustomerCare_Japan@veritas.com

Contents

Chapter 1

About this guide

This chapter includes the following topics:

Introducing this guide

The guide describes the scope of the release and lists the requirements to evaluate the classification feature.

This guide is a supplementary document which must be used in conjunction with the other Data Insight documents.

The *Veritas Data Insight Classification Guide* provides an overview of the workflow to storage and security administrators who want to use the Data Insight classification feature to classify files that Data Insight monitors.

This guide assumes that you have previously worked with Data Insight and are familiar with its features and capabilities.

Chapter 2

Getting Started

This chapter includes the following topics:

About classification

With the continuous growth of unstructured data in the business environment, taking decisions to archive and delete content of business or legal value is a challenge. You can simplify data remediation decisions by categorizing and organizing data based on tags and policies.

Data Insight integrates with Veritas Information Classifier to analyze the files that Data Insight monitors. Veritas Information Classifier uses built-in and user-defined policies to assign classification tags to files in your environment. After the files are classified, users of applications such as Data Insight can use the classification tags to filter the files for searches, reviews, and remediation.

Data Insight integrates with Veritas Information Classifier to help you do the following:

- **Analyze:** Improves the content analytics by focusing on relevant and classified data set to perform risk analysis and remediation.

Classification enables you to identify the type of data being stored in repositories (for example, Personally Identifiable Information), the purpose of the data, and the risk that is associated with it (whether sensitive or otherwise).

- **Decide:** Lets you make informed decisions to retain, secure, move, delete, or monitor data and control permissions based on the classification tags.
- **Regulate:** Ensures that the data complies with the legal requirements.
- **Organize:** Provides the ability to categorize and tag content to make it more accessible, searchable, and usable, specifically for archiving, ediscovery, and audits.
- **Visualize:** Lets you view the classified data in Data Insight and run custom Data Insight Query Language (DQL) reports that group files based on the tags that are assigned to them.

Supported file types for classification

To know more about the file types that Data Insight supports for Classification, check the *Apache Tika 1.27* documentation.

Supported file types for Optical Character Recognition (OCR)

To know more about the file types that Data Insight supports for OCR, check the *Tesseract v5.0.1.20220118* documentation.

Data Insight supports classification of files stored on file servers, SharePoint web applications, SharePoint online sources, OneDrive accounts, Object Storage Sources like Amazon S3, and Box accounts.

Note: For Amazon S3, Data Insight supports classification for files under following storage classes: STANDARD, INTELLIGENT_TIERING, STANDARD_IA, and ONEZONE_IA

Key components of Data Insight integration with Veritas Information Classifier

Table 2-1 Key components of Data Insight classification feature

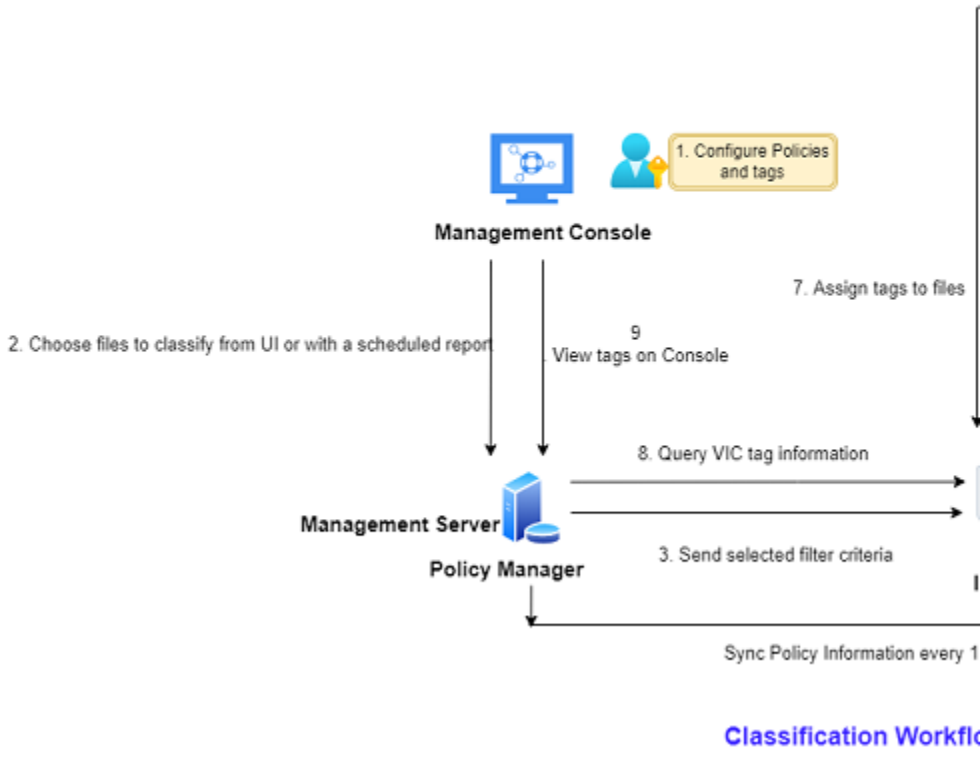
Components	Description
Policy Manager	Invokes the Veritas Information Classifier where you can manage the classification policies. You can configure policies, and enable or disable them from the Policy Manager. For more information, see the *Veritas Information Classifier Online Help*
Classification Server	This is a type of Data Insight server on which Veritas Information Classifier is deployed. This server receives file paths and applies policies to them to classify the files. It evaluates each file against a set of classification policies from the Policy Manager. The tags associated with files are then displayed on the Management Console. You can either install the Classification Server on a standalone server or assign the Classification Server role to any Data Insight server other than a pure Indexer. However, Veritas recommends that you deploy a standalone server, because classification is resource intensive. For more information on installing the Classification server, see the section *Installing the worker node* in the *Data Insight Installation Guide*. For information about installing Data Insight nodes, see the *Veritas Data Insight Installation Guide*.
Data Insight Management Server	Hosts the Policy Manager. It also initiates and monitors the classification requests.
Data Insight Indexer	The Indexer is the central platform for distributing the classification requests among the Classification Servers. It also stores the classification tag information for files.

How Data Insight classifies content

The Data Insight classification feature operates in conjunction with Veritas Information Classifier to assign classification tags to the files which Data Insight monitors.

The interaction between Data Insight and Veritas Information Classifier happens over HTTPS.

Classification with Server Pool



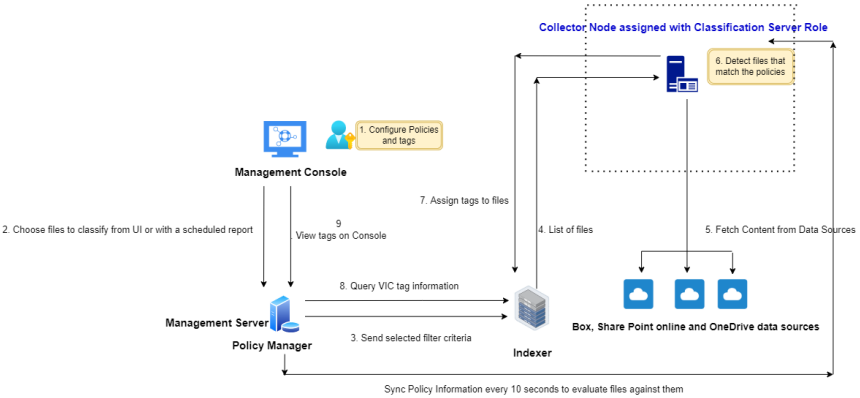
Classification Workflo

Server Pool setup is supported for the following devices / Content Source:

- Amazon S3
- NetApp
- NetApp Cluster file server
- EMC Celera
- EMC Isilon
- EMC Unity filer
- Hitachi NAS

- Windows file server
- Generic

Classification with Cloud Sources



Classification Workflow

Classification is supported for the following cloud sources

- Box
- OneDrive
- SharePoint Online
- Amazon S3

Table 2-2 describes the high-level steps that are involved when you classify files using Data Insight.

Table 2-2 Overview of setting up classification

Steps	Description
Install Classification Server.	Install the Classification Server and associate it to a Collector that is mapped to the data source whose content you want to classify. If you are deploying a Classification Server pool and map the Collector node to the server pool instead of to single Classification Server, see *Associating a Classification Server pool to a Collector* in Veritas Data Insight Administrator's Guide. For information about installing Classification Server, see *Data Insight Installation Guide* For information about mapping a Classification Server to a Collector, see *Assigning Classification Server to a Collector* in Veritas Data Insight Administrator's Guide.
In Veritas Information Classifier, configure policies, patterns, or tags. Enable or disable the policies, as required.	You can access Veritas Information Classifier from **Settings** > **Classification** > **Policy Manager**. For information on how to configure policies, patterns, and tags, see the *Veritas Information Classifier Online Help*.
Configure classification from the Management Console.	See "Configuring classification" on page 20.
From Data Insight, initiate a request to classify files.	You can submit classification request from one of the following pages on the Data Insight Console: ■ **Reports** tab **Note:** You can only submit files for classification using the Data Insight Query Language (DQL) report. It is recommended that you use the DQL classification template to build the query. ■ **Settings** > **Classification** > **Requests** ■ **Workspace** tab See "Initiating classification" on page 31.
Review and monitor the status of the files submitted for classification.	See "Viewing classification status" on page 36.

Table 2-2 Overview of setting up classification *(continued)*

Steps	Description
Review the classified information.	You can view the tags that are applied to the classified files. The tags can be retrieved using the Data Insight Query Language (DQL) report, or can be viewed on the **Workspace** tab. You can filter the tags of your interest from the right-hand panel. See "Viewing classification tags" on page 38.
(Optional) Send the classification tags to Enterprise Vault. These tags are used by Enterprise Vault when it indexes and archives the files.	To archive the files along with the tags, you must push the tags to Enterprise Vault from Data Insight. See "Pushing classification tags while archiving files into Enterprise Vault" on page 41. You can review the tags applied to the files archived in Enterprise Vault, from Enterprise Vault console. For more information about how to view the tags, see the Enterprise Vault documentation.

Prerequisites

To set up classification, make sure that the following requirements are met.

- A Classification Server is installed or a Data Insight server is assigned a classification role.
  Minimum recommended system requirements for *Classification* server if classification or smart classification is enabled:
  - Windows Server 2012 or 2012 R2, 2016 and 2019. The operating system must be 64-bit.
  - 32GB RAM
  - 16 CPUs
- Data Insight base license is installed. Access to Veritas Information Classifier is part of the Data Insight base license.
- Data sources are configured in Data Insight, as required. Make sure that you scan these data sources before initiating a classification request. For information about scanning, see the *Veritas Data Insight Administrator's Guide*.
- The Classification Server is mapped to the Collector responsible for the storage device that you want to content scan.

- The required policies, pattern, and tags are configured in Veritas Information Classifier. Make sure to enable policies before you submit the classification requests.
  For information about how to configure policies, see the *Veritas Information Classifier Online Help*.
- The classification feature is enabled on **Settings** > **Classification** > **Configuration**.
  See "Configuring classification" on page 20.
- If you want to enable Smart Classification, on a Windows Indexer apply the following Windows updates:
  - For Windows Server 2012 R2 install the update available at: https://support.microsoft.com/en-us/help/2999226/update-for-universal-c-runtime-in-windows
- The Classification Server node classifying SharePoint paths should have .Net framework version 4.5 installed on it.
- All downloaded data will be encrypted. This will add another layer of security to the data.
- By default the location for storing the file content while it is being classified is `C:\DataInsight\data`. But, you can change the location by setting the following custom configuration property on the **Settings** > **Data Insight Servers** > **Advanced Settings** page :

Property name	**classify.fetch.content_dir**
Property value	**F:\content**

  Points to remember:
  - Make sure that no classification request is in progress.
  - Restart the DataInsightVICServer, DataInsightVICClient, and DataInsightComm services after changing the location for storing the files.

  For more information, see the *Veritas Data Insight Administrator's Guide*.

About Smart Classification

Smart Classification is disabled by default when classification is enabled. See "Configuring classification" on page 20.

Data Insight prioritizes on-demand classification requests ahead of Smart Classification through machine learning.

See “Initiating classification” on page 31.

Note: Indexer requirements are different if you want to enable Smart Classification.

For Smart Classification, Data Insight automatically starts sending files with high-risk score for content scanning. Once it has enough files with positive classification outcomes from few of these scans, it uses the predictive analysis to generate a list of such files. It then sends the files with the positive predictive outcomes for content scanning with high priority.

How machine learning algorithm prioritizes files

Data Insight uses predictive analysis derived from file attributes with a high information risk score or from file attributes already content scanned by Data Insight.

Data Insight uses machine learning algorithms to create a model that captures patterns (if they exist) in the distribution of sensitive files. In order to help create a model for the machine learning algorithm, Data Insight picks files from shares with high information risk score and sends them for content scanning. It prioritizes shares with a high information risk score because if there are any classified files on such shares which are open and being accessed by a large number of active users, then such files pose a bigger threat to the organization.

Once the model is trained, it starts making predictions on files that have not been scanned by Veritas Information Classifier. It then automatically assigns a higher priority to those files from shares with high information risk score that are predicted to be sensitive and sends them for content scanning.

The following factors affects the accuracy of the prediction:

- The primary analytics attribute that is configured. The more distinctive the attribute values are, the better is the prediction accuracy. For example, the primary analytic attribute, like **Department** or the designation may yield more accuracy rather than **Email**

  For information about configuring the analytics attributes, see the *Veritas Data Insight Administrator's Guide*.
- If the content sources are not configured to receive audit events, the owner of the files may not reflect the true owner, which reduces the accuracy. For example, if most of the file have as have owners as Administrators group, then the predictive algorithm will not have enough data to analyze.

Note: Classification requests submitted using Smart Classifier method cannot be viewed or canceled.

See "Configuring classification" on page 20.

User roles for classification

To administer the classification feature, you require the Server Administrator user role in the Data Insight Console.

Adding Classification Server role to a Data Insight server

Veritas recommends that you deploy a standalone Classification Server because assigning a Classification Server role to an existing Data Insight node will affect its performance.

You can assign a Management server, Collector, or Collector and Indexer node to serve as a Classification Server instead of a designated standalone Classification Server. This indicates that the Data Insight node behaves as a Classification Server. However, you cannot add a Classification Server role to a pure Indexer, Self Service Portal node, and Windows File Server Agent node.

For more information about Classification Server, see the *Veritas Data Insight Installation Guide*.

To add the Classification Server role

1. In the Console, click **Settings** > **Data Insight Servers**.
2. Click the **Select Action** drop-down on the row corresponding to the server that you want to use as the Classification Server, and select **Add Classification role**.

   The option to add classification role is available only for the Management Server, Collector, or Collector and Indexer node. This option is not available when the role is already set to Classification Server.
3. Click **Yes** on the confirmation message to designate the server as the Classification Server. As part of this configuration, the `DataInsightVICClient` and the `DataInsightVICServer` services are installed on the designated server.

   Note that when you add a classification role to a server, it can have a impact on the performance of the server.

For information about Initiating classification and how Data Insight classifies files, see the *Data Insight Classification Guide*.

Chapter 3

Managing content classification from Data Insight

This chapter includes the following topics:

Configuring classification

You can enable and configure classification from the **Settings** > **Configuration** under **Classification**.

Note: If you enable classification, Data Insight automatically adds a rule to exclude the audit events generated by the saved credentials used for content scanning. The exclude rule is added to prevent the accesses by the named credentials from being registered and to ensure that the access time of a file is not modified.

Note: Make sure trust is established between file server domain and Management Server domain.

Note: AmazonS3 service is installed and configured on the **Classification server** by default.

Ensure that all prerequisites are met before you configure classification.

To configure classification

1. In the Management Console, click **Settings** > **Configuration** under **Classification**.
2. On the **Classification Configuration** page, edit all or any of the following settings:

Enable classification	When the check box selected, you can submit files for classification from the **Workspace** tab, **Reports** tab, and **Settings** tab > **Classification** > **Requests**. See “Initiating classification” on page 31.
Enable Smart Classification	Check the box to enable Smart Classification. When enabled, Data Insight intelligently analyzes the files to identify sensitive files and submits them for classification. See “ About Smart Classification” on page 17.
Enable Optical Character Recognition for classification of images	When this check box is selected, you can classify images. To allow classification of images, a software called Tesseract is installed on the Management Server, Collector, and Classification server nodes during the installation of Data Insight. The default location for the Tesseract installation is `C:\Program Files (x86)\Tesseract-OCR`. In case the Tesseract installation fails, refer to the Troubleshooting section in the *Veritas Data Insight Administrator's Guide* to manually install the Tesseract software. **Note:** Optical character recognition (OCR) is a performance-expensive feature. Veritas recommends that you select only those file groups that contain the specific file extensions that you need to classify for OCR. By default, the file group, **Images for OCR** is selected.
Skip classification of files with size greater than	You can set the limit on the size of files which Data Insight submits for classification. Data Insight does not submit those files that exceed the specified size.

3. Check the **Enable MIP Decryption Setting** box if you want to enable MIP decryption.

   For more information, See “MIP Decryption Support” on page 24.

4 Enter **Tenant Id**, **Client Id**, **Client Secret Key**, and **Microsoft Administrator Account** details.

Note: Use Microsoft Administrator Account which is either Global Administrator account or Minimum Privilege Account with Information Protector reader and Sensitivity Label reader privileges. For more information, refer to *Creating Minimum Privilege Account Role in Compliance Center* section in *Data Insight Administrator's Guide*.

---

5 Click **Save**.

If you are deploying a Classification Server pool and map the Collector node to the server pool instead of to single Classification Server, see *Associating a Classification Server pool to a Collector* in

See "Configuring safeguard settings for Classification Server" on page 23.

See "Initiating classification" on page 31.

See "Viewing classification status" on page 36.

Configuring safeguard settings for Classification Server

In order to prevent the disk space on the Classification Server from getting full, you must enable safeguards that stops the services that fetch content from data sources on the server when the disk space falls below the configured value. You must enable the safeguard settings and configure the threshold values that activate and reset the safeguards.

To configure the safeguard settings

1. In the Management Console, click **Settings** > **Configuration**.
2. Edit the following setting:

Specify threshold/reset values in MB / Specify threshold/reset values in percentage	You can specify the threshold for disk utilization in terms of size and percentage. The DataInsightWatchdog service initiates the safeguard mode for the Classification Server node if the free disk space falls under the configured thresholds. The DataInsightWatchdog service automatically resets the safeguard mode when the free disk space is more than the configured thresholds. You can edit the threshold limits as required. If you specify values in terms of both percentage and size, then the condition that is fulfilled first is applied to initiate the safeguard mode. For more information on the DataInsightWatchdog service, see the *Veritas Data Insight Installation Guide*.

See "User roles for classification" on page 19.

MIP Decryption Support

When you add MIP label, which has encryption setting enabled, the file is protected by it. If you want to classify such encrypted files, enable MIP decryption settings in classification configuration. After you edit the required settings,Data Insight will be able to read and classify those files with sensitive information.

Prerequisites for MIP Decryption Support

- Install .net version 4.6 or above
- Check the TLS version on the collector node and classification node by referring to **https://learn.microsoft.com/en-us/azure/active-directory /hybrid/reference-connect-tls-enforcement#powershell-script-to-check-tls-12**
- TLS version should be TLS 1.2 or above. Refer the following link for more details:**https://learn.microsoft.com/en-us/azure/active-directory /hybrid/reference-connect-tls-enforcement#powershell-script-to-enable-tls-12**
- Create a new app registration for the MIP Decryption Support

- For **OneDrive** and **SharePoint**, set the following custom attribute on the Classification server:

  `node.connector.classification.job.timeout` and set the custom attribute value to 7200.

To create a new app registration for the MIP Decryption Support

1. In a new browser window, sign in the Azure portal **https://portal.azure.com** with the Azure AD tenant that you use with Azure Information Protection.
2. Navigate to **Azure Active Directory** > **Manage** > App registrations, and select **New registration**
3. On the **Register an application** pane, specify the following values, and then click **Register**

Option	Value
Name	MIP-Decryption-App Specify a different name as needed. The name must be unique per tenant.
Supported account types	Select Accounts in this organizational directory only

4. On the **AIP-DelegatedUser** pane, copy the value for the Application, that is, **Client ID**. For example: *77c3c1c3-abf9-404e-8b2b-4652836c8c66*. This value is used as a **Client ID**.
5. Similarly, copy the **Tenant ID**
6. Copy and save **Client ID** and **Tenant ID** for future reference.
7. From the sidebar, navigate to **Manage** > **Certificates & secrets**
8. On the *AIP-DelegatedUser - Certificates & secrets* pane, in the **Client secrets** section, select **New client secret**
9. In the **Add a client secret** field, specify the following, and then click **Add**

Field	Value
Description	Application for MIP Decryption Support in DataInsight
Expires	Specify your choice of duration (1 year, 2 years, or never expires)

10. Navigate back to the *AIP-DelegatedUser - Certificates & secrets* pane

11 In the **Client secrets** section, copy the string for the **VALUE**. For example: OAkk+rnuYc/u+]ah2kNxVbtrDGbS47L4.

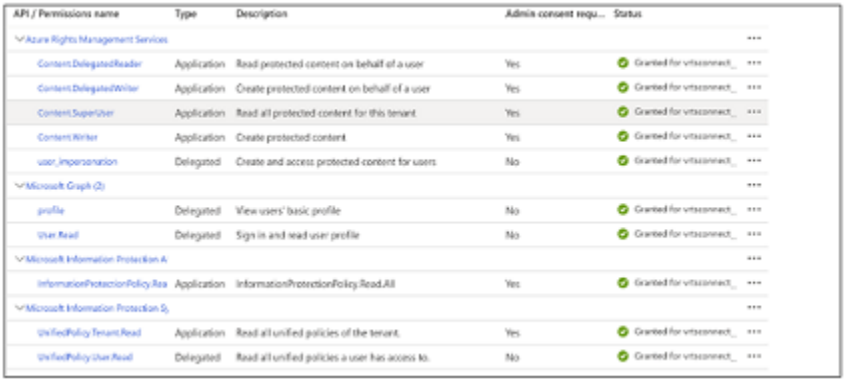
To make sure that all characters are copied, click the **Copy to clipboard** icon.

12 To add **API Permissions**, select **Manage** > **API permissions** from the sidebar.

13 On the *AIP-DelegatedUser - API permissions* pane, select **Add a permission**

- Select *Azure Rights Management Services*
- Select *Delegated permissions*
- Check user_impersonation and select *Add permissions* at the bottom of the page
- Select *Add a permission*
- Select *APIs my organization uses*
- In the search box, type *Microsoft Information Protection Sync Service*, press enter and then select the service
- Select *Delegated permissions*
- Check *UnifiedPolicy.User.Read* then select *Add permissions*

14 After adding permissions specified above, add more permissions shown in the image below:



API / Permissions name	Type	Description	Admin consent requ...	Status
Azure Rights Management Services				
Content.DelegatedReader	Application	Read protected content on behalf of a user	Yes	Granted for vrtxconnect_
Content.DelegatedWriter	Application	Create protected content on behalf of a user	Yes	Granted for vrtxconnect_
Content.SuperUser	Application	Read all protected content for this tenant	Yes	Granted for vrtxconnect_
Content.Writer	Application	Create protected content	Yes	Granted for vrtxconnect_
user_impersonation	Delegated	Create and access protected content for users	No	Granted for vrtxconnect_
Microsoft Graph (2)				
profile	Delegated	View users' basic profile	No	Granted for vrtxconnect_
User.Read	Delegated	Sign in and read user profile	No	Granted for vrtxconnect_
Microsoft Information Protection A				
InformationProtectionPolicy.Rea	Application	InformationProtectionPolicy.Read.All	Yes	Granted for vrtxconnect_
Microsoft Information Protection S				
UnifiedPolicy.Tenant.Read	Application	Read all unified policies of the tenant	Yes	Granted for vrtxconnect_
UnifiedPolicy.User.Read	Delegated	Read all unified policies a user has access to.	No	Granted for vrtxconnect_

Configuring classification policies in Veritas Information Classifier

Veritas Information Classifier lets you create custom policies or use the built-in policies, patterns, and tags. You can access Veritas Information Classifier from the Data Insight Management Console.

Note: For Veritas Information Classifier, supported browsers are Google Chrome, Edge, Internet Explorer, and Mozilla Firefox.

To access Veritas Information Classifier

1. In the Data Insight Management Console, click the **Settings** tab.
2. In the left pane, click **Classification** > **Policy Manager**.

   The **Policy Manager** opens and displays the landing page of the Veritas Information Classifier Console.

The landing page by default displays the list of configured policies. Note that you can navigate to the landing page by clicking in the Veritas Information Classifier icon in the top left corner.

Note: You do not require any special privileges to access Veritas Information Classifier.

Exact Data Matching (EDM)

Exact Data Matching (EDM) is created to protect your sensitive content. You can use EDM to detect structured or tabular data. EDM is designed to find records that are part of an indexed data source in either structured or unstructured targets. Some examples are social security numbers, bank account numbers, and credit card numbers. You can also detect confidential customer and employee records, price list entries, parts from a parts list, and other confidential data stored in a structured data source, such as a database, directory server, or a structured data file such as CSV or spreadsheet.

For information about how to configure policies, patterns, and tags in Veritas Information classifier, see the *Veritas Information Classifier Online Help*.

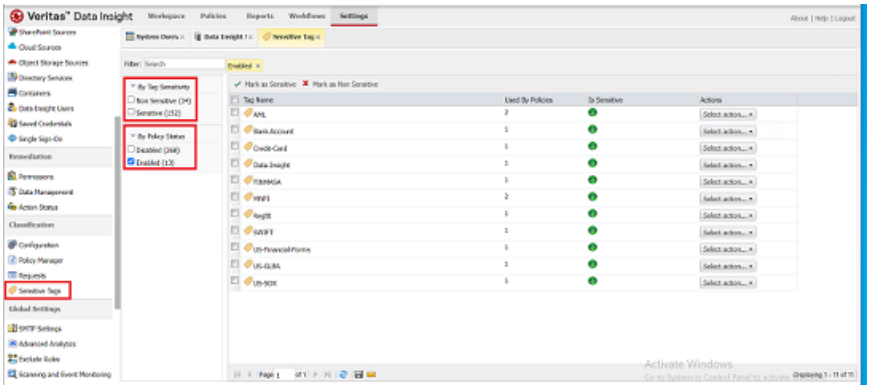
Microsoft Purview Information Protection (MIP) Label

Microsoft Information Protection (MIP) is a built in, intelligent, unified and extensible solution to protect sensitive data. MIP technology integration allows adding labels to the documents. The label might have a policy to restrict access to the sensitive documents.

Fetching MIP Labels using Veritas Information Classifier (VIC)

Note: Classification feature must be enabled in Data Insight to fetch MIP tags associated with the files. Refer *Data Insight Classification guide* and *Data Insight Administrator's guide* to know more about setting up Classification.

To fetch MIP Labels using VIC, navigate to **Settings** and click **Policy Manager** in the left pane.

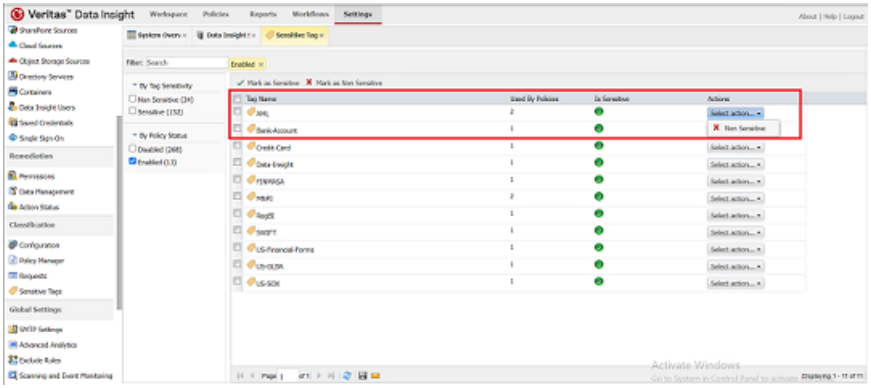


Note: If you want to fetch MIP labels of protected files, enable only MIP related policies, without enabling other policies in Veritas Information Classifier.

Configuring MIP tags

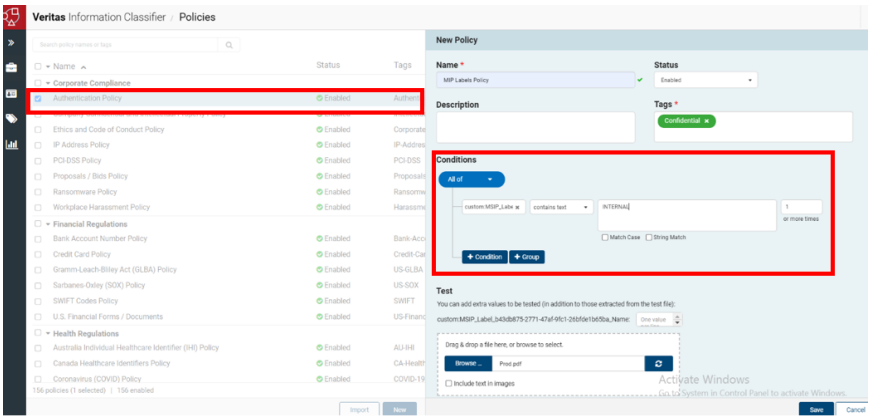
You can create custom tags that you want to associate with the files for which MIP labels are available. These tags will then be available in Data Insight for selection.

Navigate to **Veritas Information Classifier** >> **Tags** to configure new tags.



Configuring policies

After creating new tags, you need to associate those tags with a policy. To create new policy click **New** at the bottom of the **Policies** page. A new window appears.



- Add name of your choice
- Status should be **Enabled**
- Add description (optional field)
- Select the tags that you have created
- Select **custom string field** from the **Conditions** drop-down. You can add multiple conditions as per your requirement.

Note: Data Insight automatically tags all files coming from VIC as Sensitive files. If the files are tagged with MIP label using VIC, it will be tagged as sensitive, affecting the user risk score.

- MIP labels are custom attributes of a file so you will have to provide information in specific format in the Text Box present in VIC Policy Manager.
  - For Office files with **INTERNAL**, label, add the following condition

    ```
    custom:MSIP_Label_b43db875-2771-47af-9fc1-26bfde1b65ba_Name
    "contains Text" INTERNAL
    ```

    Note: You need to add **custom:** before the label name and include GUID present in the label.

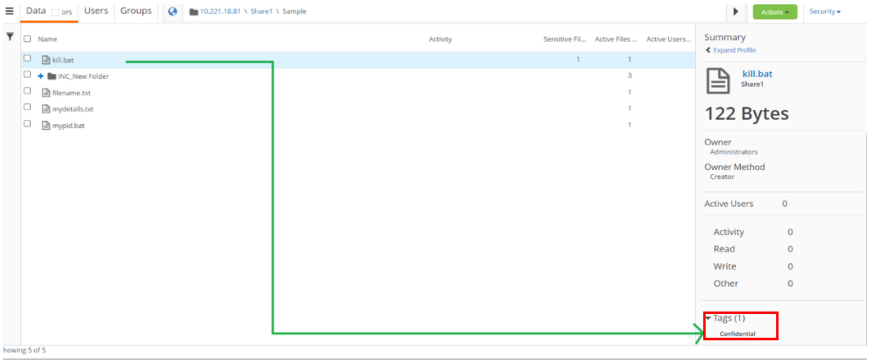
  - For PDF file with **INTERNAL**, label, add the following condition

    ```
    pdf:docinfo:custom:MSIP_Label_b43db875-2771-47af-9fc1-
    26bfde1b65ba_Name "contains Text" INTERNAL
    ```

    Note: You need to add **pdf:docinfo:custom** before the label name and include GUID present in the label.

File Types	Sample Format	Condition	Value
Office Files	custom:MSIP_Label_ <MSIP_label_GUID>_Name	Contains Text	INTERNAL
PDF Files	pdf:docinfo:custom:MSIP_ Label_<MSIP_label_GUID> _Name	Contains Text	INTERNAL

After completing the process, you can see the associated tags in the Data Insight interface.



Note: Newly added MIP labels will be visible only after MIPGetLabelsjob gets executed on the Management Server. You can run it manually if you want to see the labels before the scheduled run, which is at 12 a.m. everyday by default.

Initiating classification

Data Insight lets you scope the classification for a shares/site collections/user accounts, folder, or file. For example, you can prioritize classification of files accessed by a risky user (based on the user risk score) or by a watch-listed user.

Files in the classification request are evaluated against policies configured in Veritas Information Classifier and matching tags are assigned. Note that every policy can have multiple tags and the same tags may be reused in different policies. Thus, you may see multiple files being assigned the same tags even if they are being evaluated against different policies.

You can initiate classification of files in the following ways:

- Using DQL report.
  See "Classifying paths using DQL reports" on page 32.
- From the **Workspace** tab of the Data Insight console.
  See "Initiating classification from the Workspace tab" on page 34.
- Using a CSV file to submit file paths from the **Settings** > **Requests** page under **Classification**.
  See "Uploading paths for classification" on page 35.

You can review the status of the classification request from the **Classification** > **Requests** tab.

See "Viewing classification status" on page 36.

Once the classification of files is complete, you can review the tags assigned to the files and take the appropriate remediation action, such as archiving or deleting.

You can also cancel only in-progress request.

See "Canceling a classification request" on page 37.

See "Viewing classification tags" on page 38.

See "How Data Insight classifies content" on page 12.

See "User roles for classification" on page 19.

Classifying paths using DQL reports

You can create a Data Query Language (DQL) report for the paths that you want to classify and use the report to initiate a classification action.

For more information about creating DQL queries, see the *Veritas Data Insight Programmer's Reference Guide*.

To initiate a classification request from the Reports tab

1. On the Management Console, click the **Reports** tab.
2. From the left-hand side pane, select **Custom Reports** > **DQL Reports**.
3. Configure a DQL report either by using the built-in classification templates or by writing a custom query to suit your requirement.

   For information about creating a DQL report and to review the template queries for classification, see the *Veritas Data Insight User's Guide*.

   In the DQL query, make sure that the **path.device.name**, **path.msu.name**, **type**, and **name** columns are specified.
4. Do one of the following:
   - On the **Remediation** tab of the **Create Report** wizard, select **Classify** to automatically submit all input file paths in the report for classification.
     - If you want to classify only the new or modified files, click **Save and Run**
     - If you want to classify the whole Dataset or if you have changed policy, condition or tags configuration in VIC, uncheck the box and click **Save and Run**.
   - Or, after the report is run, from the DQL reports list page, select the report. Click **Select Action** > **Actions** > **Classify** to submit all files in the report for classification.

- If you want to classify only the new or modified files, click **Classify Now** on the confirmation pop-up.
- If you want to classify the whole Dataset or if you have changed policy, condition or tags configuration in VIC, uncheck the box and click **Classify Now** on the confirmation pop-up.

Once the classification action is complete, you can review the tags assigned to files either by creating a new DQL report for the same paths or on the **Workspace** summary pane.

See "Viewing classification tags" on page 38.

You can also choose to push the classification tags to Enterprise Vault when archiving them.

See "Pushing classification tags while archiving files into Enterprise Vault" on page 41.

DQL templates for classification

Data Insight provides a few out-of-the-box templates that help you create queries for certain use cases. At the time of creating a DQL report, you can select any of the built-in queries, and modify the content to suit your particular reporting needs. Additionally, you can create your own queries and save them to be used later as templates.

Table 3-1 Data Insight Query Language classification templates

Name	Description
Files to send for classification	Creates a report of all files that are accessible to more than 1000 users. Use the DQL report to send file paths in the output for classification.
Classified files with a specific extension	Creates a report of all files with a specific extension (for example, PST) and a specific tag name (for example, US-PII). You can either use the query to identify tags associated with specific files or to push these files to Enterprise Vault for archiving.
All PII files	Creates a report of all files that are tagged as Personally Identifiable Information (PII). These are files that may contain sensitive information such as Social Security, credit card, and drivers' license numbers.

Table 3-1 Data Insight Query Language classification templates *(continued)*

Name	Description
Classify active users files	Creates a report listing all files that have been accessed by users identified as active by Data Insight. You can then use this report to submit these files for classification.
Classified files summary	Creates a report that summarizes all files that have already been classified.
Mis-labeled sensitive files	The query list all the sensitive files (as determined by Veritas Information Classifier) but having sensitivity label as "Public" as labeled by end user. For example, file labeled by user using MIP in OneDrive and SharePoint online.
Un-labeled sensitive files	The query list all the sensitive files (as determined by Veritas Information Classifier) which are not having any sensitivity labels. For example: file not labeled by user using MIP in OneDrive and SharePoint online.
Activity on MIP labeled files	The query list all the activity on "Internal Only" labeled files like the ones labeled by the user using MIP during a provided time range on a particular share. Replace the sensitivity label as defined by your organization.

For more information on Create/Edit DQL report and Creating custom templates for DQL queries, refer to *Data Insight User's guide*

See “Initiating classification” on page 31.

See “Pushing classification tags while archiving files into Enterprise Vault” on page 41.

Initiating classification from the Workspace tab

You can submit files for classification from the Data Insight Management Console.

To initiate content classification from the Workspace tab

1. On the Management Console, click **Workspace** > **Shares** to view the configured shares/site collections/user accounts.

   Select the shares/site collections/user accounts/buckets that you want to classify. Or, expand the shares/site collections/user accounts/buckets to select the folders or files that you want to classify.

   Note: Submitting an entire data source is not supported.

2. From the **Actions** drop-down, select **Classify**.

   - If you want to classify only the new or modified files, click **Classify** on the confirmation pop-up.
   - If you want to classify the whole Dataset or if you have changed policy, condition or tags configuration in VIC, uncheck the box and click **Classify** on the confirmation pop-up.

   A new confirmation pop-up appears informing that the files are submitted for classification.

See "Initiating classification" on page 31.

Uploading paths for classification

Data Insight lets you submit file paths for classification through a CSV file.

To upload a CSV file

1. On the **Settings** > **Requests** page, click **Create Request Using CSV**.
2. In the **Create Request Using CSV** dialog box, select the file that you want to upload. Ensure that the file paths adhere to the format that is provided in the CSV template.

   Alternatively, download the CSV template, make the necessary changes, and upload the file.
3. If you want to classify only the new or modified files, click **Upload**.

   If you want to classify the whole Dataset or if you have changed policy, condition or tags configuration in VIC, uncheck the box and click **Upload**.
4. The confirmation message appears that displays the number of paths submitted for classification, and their status.
5. Click **OK** on the confirmation message.

Viewing classification status

The **Requests** page lets you review and monitor the status of the classification requests. From this page, you can also perform operations such as exporting the details of files for which the classification process failed, and submit file paths for classification using a CSV file.

See "Uploading paths for classification" on page 35.

To view the classification status

- On the Management Console, click **Settings** > **Classification** > **Requests**.

  The **Requests** page displays the following information.

  Certain columns are hidden by default. To view these columns, click on the column header and select **Columns** > <*Column name*>.

  - A unique ID that is assigned to each classification request.
  - The name of the classification request that is submitted for classification.
  - The source (origin) through which the classification request is submitted.
  - The number and size of files that are submitted for classification.
  - The number and size of files that are classified.
  - The number and size of files for which the classification process failed.
  - The number and size of files that are identified as tagged file, earlier referred as sensitive. Data Insight considers a file to be tagged when its contents matches one or more policy definitions.
  - The completed percentage.
  - The time at which the classification request related information was updated.
  - The content fetch status of all the available nodes in a consolidated form. Each node will have one of the following status tag: Completed, On Hold, Paused, or In-Progress. While consolidating the status, if all nodes have same status, it will be reflected in the Consolidated Status column but if not, most prevailing status will be reflected.
    To explain it further, consider that there are 4 nodes,

Node 1	Node 2	Node 3	Node 4	Consolidated Status
Completed	Completed	Completed	Completed	Completed
Completed	On Hold	In-Progress	Paused	On Hold

Node 1	Node 2	Node 3	Node 4	Consolidated Status
In-Progress	In-Progress	Completed	In-Progress	In-Progress
In-Progress	Paused	Completed	In-Progress	Paused

- The status of the classification request for the files that are being classified.
- The time when the Classification Server started processing the classification.
- The total time that the Classification Server has taken to complete the classification operation.
- The user who submitted the classification request.

Data Insight lets you perform certain actions on the classification requests. The actions include cancel the request, submit the requests that are in failed state, and export the request details that failed classification to a CSV file.

See "Resubmitting paths that failed classification" on page 38.

See "Canceling a classification request" on page 37.

Filtering the classification requests

You can use the free-form Filter text box to enter the search criteria. Or, use the group by filter to sort the requests based on their classification status or origin used to submit the files.

To filter the classification requests

- On the **Requests** page, select the filter value in the left pane.

See "Initiating classification" on page 31.

Canceling a classification request

You can cancel the classification request that is currently in **In-progress** state. However, you cannot cancel the requests that are in **Pending** state.

To cancel a classification request

1. On the **Classification** > **Requests** page, on the row corresponding to a request in the pending state, choose **Select Action** > **Cancel**.
2. Click **OK** on the confirmation message.

Resubmitting paths that failed classification

Classification of certain paths may produce a failure. For example, classification may fail when scanning of data sources have failed due to network error or if the Classification Server is unavailable. You can export the list of files on which classification has failed or is partially successful, resolve the issues, and then resubmit them for classification.

You can directly resubmit the failed paths or check, edit, and resubmit the failed paths by uploading a CSV file.

Note: If you resubmit failed paths, *Classify only new or modified files* option will selected by default.

To directly resubmit the failed paths, click **Select Action** and click **Classify failed paths** on the **Classification Requests** page.

To check, edit, and resubmit the failed paths by uploading a CSV file,

1. On the **Classification** > **Requests** page, click **Download failed paths** on the row corresponding to a partially successful or failed classification request.

   A CSV file with the list of files which have not been classified is saved on your machine.

2. Click **Create Request Using CSV**.
3. In the **Create Classification Request** dialog box, select the file that you want to upload.
4. Click **Upload**.

   The confirmation message appears that displays the number of paths submitted for classification and their status.

5. Click **OK** on the confirmation message.

Viewing classification tags

Do the following to review the tags assigned to file paths:

- Run a Data Insight Query Language (DQL) report to review the tags assigned to the files submitted for classification.
  For more information on Create/Edit DQL report, refer to *Data Insight User's guide*
- Navigate to the **Workspace** > **Shares** view and in the **Summary** panel to the right, expand the **Tags** section. The tags assigned to the files under the

shares/site collections/user accounts are listed here. You can use the **Workspace** filters to search for files with specific tags and perform remediation actions.

For more information on Viewing shares summary, refer to *Data Insight User's guide*

If you choose to archive the files that are part of a DQL report in Enterprise Vault, you can also push the classification tags for these files. The tags help you make appropriate retention management decisions on the indexed data.

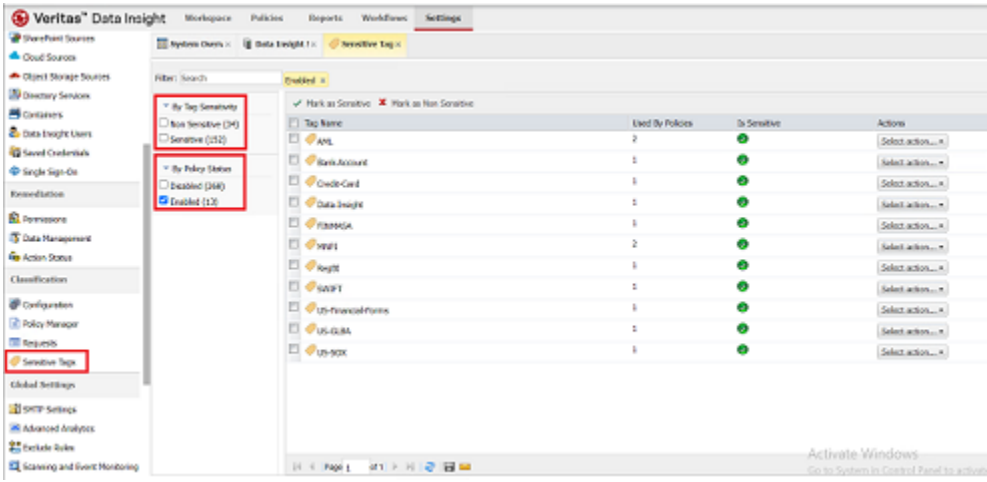
See "Pushing classification tags while archiving files into Enterprise Vault" on page 41.

See "Initiating classification" on page 31.

Sensitive and non-sensitive tags

Data Insight can fetch multiple attributes like *Author, Title, Size, MIP* when files are classified by Veritas Information Classifier (VIC). Data Insight can also differentiate between **Sensitive** and **Non-Sensitive** tags. This feature enables you to have correct sensitivity information and risk score computation.

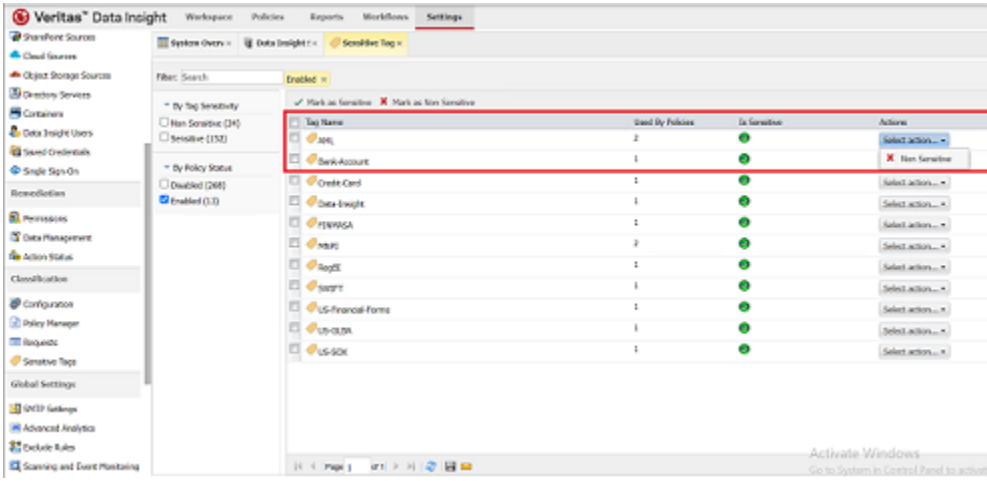
When VIC services are running, go to **Settings** and click **Sensitive Tags** in the left pane.



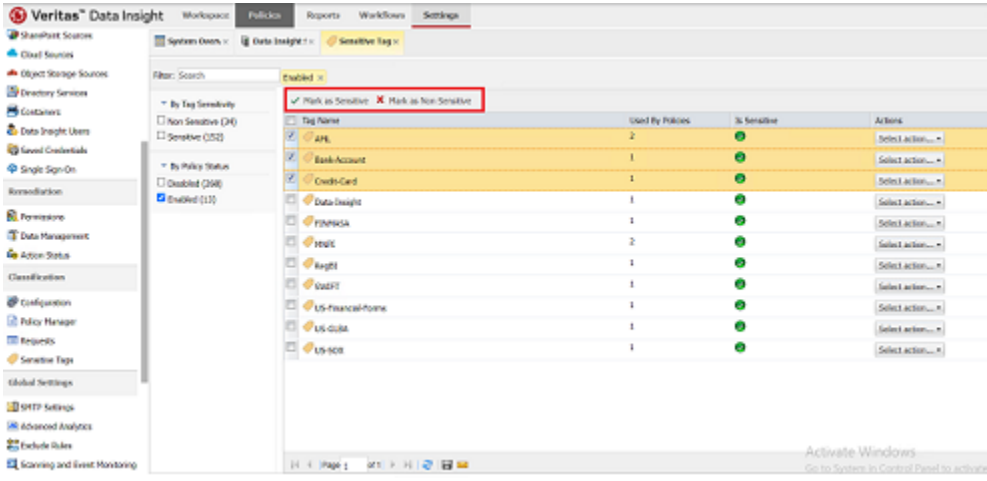
On this page, you can see two sections, **By Tag Sensitivity** and **By Policy Status**. By tag sensitivity filter, you can filter sensitive or non-sensitive tag. If you select **Enabled** in **By Policy Status** section, only tags associated with **Enabled** policy will appear on the screen. If you select **Disabled**, tags associated with disabled

policy will be visible. You can also see the count of policy per tag next to the subsequent field.

You can modify the sensitivity of the VIC policy tag by selecting the tag and clicking the **Select action** drop-down.



You can also modify the sensitivity of the VIC policy tag by selecting multiple tags at a time and using bulk edit option given in the top ribbon.



Pushing classification tags while archiving files into Enterprise Vault

You can add classification tags to a file that you want to archive using Enterprise Vault. When the file is archived and indexed in Enterprise Vault, the classification tag is included in the index of the file. When you search the archive, the search is carried out on the tags that are attached to the file rather than the entire Enterprise Vault database. You can also use the classification tags to reassign retention categories that determine how long archived items are stored in Enterprise Vault. You can add any number of classification tags.

A user with the Server Administrator or a Report Administrator role can add classification tags to files being archived.

To push classification tags

1. Click the **Reports** tab and select the report type that allows data remediation using Enterprise Vault, for example Inactive Data by File Group or DQL report.
2. Click **Create Report**.
3. On the **Create Report** wizard, enter the report input parameters and navigate to the **Remediation** tab.
4. Select the **Take action on data generated by report** check box.
5. Select the **Archiving (Enterprise Vault)** radio button.
6. Select the **Add Custom Index Property** check box.
7. Select a value from the **Property type** drop-down box. The different property types include Text, Integer, Date, and Classification property.

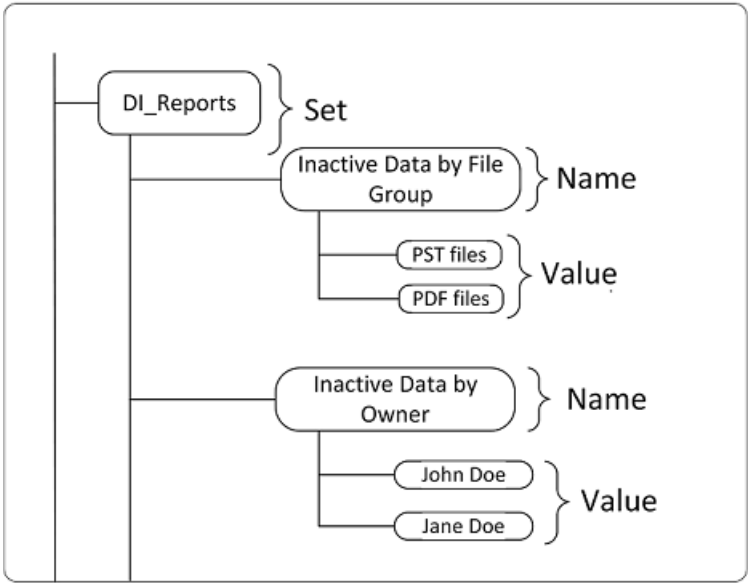
   Note: Classification property is only available for DQL reports.

8. Depending on what you select, three text boxes corresponding to **Set**, **Name** and **Value** appear. Enter the tag in the following format:

Field	Description
Set	Name of the property-set to which the property is added. **Note:** This name can be an already existing property set name or a new one. If it does not already exist, a new property set is created when you make an entry here.
Name	Name of the property. A property set can have a number of unique names.

Field	Description
Value	Value of `Set.Name`. The value can be Text, Integer, or Date type depending on the property type selected. **Note:** This field is not available when the **Property type** is set to Classification property because Data Insight applies the classification tags fetched from Veritas Information Classifier as the value of the property type.

Figure 3-1 An example of Set, Name, and Value



Note: If classification tags are selected, all classification tags applicable to the file being archived are added to the archive request. The tags are in a semi-colon separated list.

You can review the classification tags applied to the files archived in Enterprise Vault from:

- On the Data Insight Management Console, navigate to **Setting** > **Action Status** page.
- On the Enterprise Vault Console.

You can track and manage the progress of the operations that are initiated from the Data Insight Management Console. You can perform the following tasks from the **Action Status** page.

- View the progress of an operation
- Cancel an ongoing operation
- Re-run a completed or canceled operation
- Delete a completed or canceled operation

Note: The Cancel and Re-run options are not supported for Delete Files action.

You can view the progress of the following types of operations.

- Archive operations using Enterprise Vault
- Permission remediation operations
- Remediation operation using custom actions
- Deletion of stale, inactive, or orphan files and folders

The **Action Status** page displays the following information:

- The unique identification number of the operation
- The system-generated name for the triggered operation indicating its origin
- The type of the operation. For Enterprise Vault operations, the type is specified as *EV*. For permission remediation operations, the type is *PR*. For custom action operations, the type is *CUSTOM*. For delete action operations, the type is *DELETE FILES*
- The time when the user triggered the operation from the Management Console
- The user who triggered the operation
- The time when the destination server starts processing the triggered request. The destination server is an external server which is responsible for the actual execution of an operation that is triggered from the Data Insight Management Console.
  For example, in the case of an archive operation, the destination server is the Enterprise Vault server.
- The time when the destination server completes processing the request
- The time it takes to complete the operation
- The status of the operation

- In case of archive operation using Enterprise Vault, the custom index property, including the classification tags applied to the files that are being archived

---

Note: Only some columns are displayed in the default view. You can view the other columns by selecting them from the column header drop-down.

The **Details for Action** panel shows you the step-by-step break-down of the selected operation.

To view the status of an operation

1. In the Management Console navigate to **Settings** > **Action Status**. The Action Status page displays the details of recently triggered operations
2. Use the check box filter to display the operations based on their **Type** or **Status**. Additionally, you can use the search facility to display the operations based on their attributes such as **Origin**, **Type**, or **Status**
3. Click the **Origin** of the selected operation to view granular details of an operation. Alternatively, click the **Select Action** drop-down, and select View.
4. The details of the selected operation are displayed in the **Details for Action** panel.
5. Use the check box filter to display the operations based on the attributes such as: **Status** or **Filer**. Additionally, you can use the search facility to display the details based on attributes such as **Path** or **Status**.

You can cancel an operation that is in progress. Cancelling an operation, pauses all the activities of the operation. You can re-run a canceled operation later.

To cancel an ongoing operation

1. In the Management Console, navigate to **Settings** > **Action Status**. The Action Status page displays the details of recently triggered operations
2. Use the check box filter to display the operations based on their **Type** or **Status**. Additionally, you can use the dynamic filter to display the operations based on their attributes such as **Origin** , **Type**, or **Status**
3. Click **Select Action** for the operation you want to cancel.
4. Click **Cancel**.

You can re-run a canceled or a completed operation.

To re-run a canceled or a completed operation

1. In the Management Console navigate to **Setting** > **Action Status**. The Action Status page displays the details of recently triggered operations.
2. Use the check box filter to display the operations based on their **Type** or **Status**

3. Click **Select Action** for the operation you want to re-run.
4. Select **Run Again**.
5. Select any of the following:
   - **All** - To run all the sub-steps for the operation
   - **Unsuccessful** - To run all the failed sub-steps for the operation

---

Note: For a permission remediation or custom action-related operation that is canceled, the option to run the unsuccessful steps again is not available.

You can delete an operation that is canceled or completed.

To delete a canceled or a completed operation

1. In the Management Console navigate to **Setting** > **Action Status**. The Action Status page displays the details of recently triggered operations.
2. Use the check box filter to display the operations based on their **Type** or **Status**.
3. Click **Select Action** for the operation you want to delete.
4. Select **Delete**.

Appendix A

Classification best practices

This appendix includes the following topics:

General

- Use recommended system configurations for better throughput.
- Use a classification server pool of multiple nodes to achieve higher throughput for large classification tasks.
- Disable smart classification if not required.
  - Smart classification requires significant resources on Indexer and Management Server nodes to automatically generate the list of files to classify.
- Update default disk safeguard thresholds to higher values especially in case of PDF Files where uncompressed files can consume up to 40GB disk space (considering 16 threads and file sizes around 2.5 GB) hence the values given below will safeguard against disk usage reaching maximum limit.
  - Reset at 50 GB (or higher)
  - Stop at 45 GB (or higher)
- As a part of classification, Data Insight does text extraction and uses the data directory for storing temporary files.

Maximum file size supported

- Data Insight has a default maximum file size of 50MB. This limit can be changed in the *Classification Configuration* settings page.

- Text extraction during classification is bounded by the uncompressed size of a file and this uncompressed size dictates whether files can be successfully classified. All Microsoft Office documents since Office 2007 use Office Open XML format (.docx, .pptx etc) which introduced compression.
  - Most Office docs therefore have a degree of compression ranging from 20%-70% depending on the mix of text and images, with pure text compressing to around 80%.
  - Files with a lot of images will compress less as images such as JPEG and PNG are already compressed.
  - PDFs are not compressed by default unless the 'Optimize PDF' option in Adobe Acrobat or similar PDF authoring applications has been used.
- It has been observed that 16 concurrent files of 400MB uncompressed docx files can be classified without any memory exhaustion.
  - This means that 16 concurrent requests of docx files in a range of 100MB-250MB logical sized would probably work fine given the average compression ratio.
  - Note that the compression ratio is impossible to predict unless you analyse each file or have some indication of the type of content within the corpus.
  - These figures do not relate to volume/disk level compression, but the compression that Microsoft Office applies to the content. A .docx file is simply a ZIP container that can be opened in a tool such as 7-Zip to assess the uncompressed size.

The table below shows the file types and sizes tested with the recommended **Classification Server** specification:

- Recommended maximum file sizes for classification without OCR enabled

Document type	Extensions	Maximum compressed file size tested	Maximum uncompressed file size tested
Microsoft Word	doc, docx, docm, dotm, dotx	200 MB	450 MB
Microsoft PowerPoint	ppt, pptx, pps, potm, potx, ppsm, ppsx	200 MB	450 MB
Office Tabular	xls, xlsx, xlt, xltx, xlsb, xlam	50 MB	100 MB

Document type	Extensions	Maximum compressed file size tested	Maximum uncompressed file size tested
Adobe PDF	pdf	1 GB	Compressed PDFs are not yet tested. However, the maximum uncompressed size would mirror the compressed size of 1 GB.

- Server specification used (the recommended Data Insight Classification Server specification)
  - 16 Cores, 32GB RAM
  - 16 classification threads running in parallel
- Using Optical Character Recognition (OCR)
  - OCR usually results in higher memory consumption which eventually affects the classification performance.

Larger File support

- It is possible that larger files than tested could be successfully classified, but it depends on the size of other files being classified at the same time. For example, if a 300MB DOCX is 1GB uncompressed, it could still be classified successfully if all other 15 files running in parallel are relatively small since the total memory used by the classification process would be within limits.
- As there is no way to ensure that a mix of small and large files are classified at the same time, recommend that any DQL reports that are used to select files to classify are not ordered or segregated by file size. This ensures that the files submitted to VIC are done so as ‘randomly’ as possible.
  - For example, do not classify all ‘small DOCX’ files first and leave the largest ones until later. Classifying the very largest files together in one classification Job increases the risk that the total uncompressed size of 16 large files would lead to VIC memory exhaustion. Submitting a mix of file sizes together provides the best chance of large and large uncompressed files being successfully classified.
  - If using DQL to generate a report of files to classify, do not order the output of the report by size as that would lead to VIC processing the largest files together, whether they are sorted to appear at the start or end of the report.

Recommendations for creating classification Jobs

- Use DQL reports which will filter out the files based on the above recommendations and then trigger classification requests accordingly.
- Enable only required policies in VIC configuration.
  - As the number of enabled policies and policy complexity increases (such as using complex regular expressions or hundreds of keywords), the throughput tends to decrease.
- OCR process is generally memory intensive disable the process if not required.
- Configure the content fetch pause window to reduce the potential impact on the source devices.
  - The content fetch job copies files from the source devices to classify them.
  - By default, the job is paused from 7am to 7pm which matches normal working hours.
  - Recommend assessing the load on the devices during the content fetch as many customers have discovered the load does not disrupt any normal activities. If it can run 24-hours a day, that will help ensure that the classification process has a constant feed of files to classify and hence throughput can be increased.

Appendix B

Classification jobs

This appendix includes the following topics:

Classification Jobs

The section explains the function of the classification jobs that run in various services. You can view the status of all Data Insight processes from the **Settings** > **Data Insight Servers** > **Services** tab on the Management Console.

For more information, see the *Veritas Data Insight Administrator's Guide*.

Table B-1 Communication service jobs

Jobs	Description
`FileTransferJob_classify`	Runs on all Data Insight nodes once every minute. Distributes the classification events between Data Insight nodes.
`FileTransferJob_content`	Runs every 10 seconds on the Windows File Server. Routes content file and CSQLite file to the assigned Classification Server.
`ClassifyInputJob`	Runs every 10 seconds on the Management Server. Processes the classification requests from the Data Insight console and from reports for the consumption of the book keeping database.
`ClassifyBatchJob`	Runs every minute on the Indexer. Splits the classification batch input databases for the scanner's consumption, which are later pushed to the Collector.

Table B-1 Communication service jobs *(continued)*

Jobs	Description
`ClassifyIndexJob`	Runs once every minute on the Indexer node. Updates the index with classification tags and also updates the status of the book keeping database.
`ClassifyMergeStatusJob`	Runs once every minute on the Management Server. Calls the files with the classification update status that are received from each indexer. These files are automatically created on the indexer whenever updates are available. It also updates the global book keeping database that is used to show high level classification status on the Console.
`CreateFeaturesJob`	Runs once every week on Sunday at 00.01 A.M. on the Indexer. Checks if sufficient classified data is available for the supervised learning algorithm to create predictions (training sets). This job has a multi-threaded execution framework which executes actions in parallel. The default thread count is 2. You can set the value using the `matrix.classification.sl.features.threads` property at global or node level. Note that the node level property always takes precedence over the global level property.

Table B-2 Classification service jobs

Job	Description
`ClassifyFetchJob`	Runs every minute on the server that is assigned the role of a Classification Server. Searches the `classification/inbox` folder for the input files and adds them to the priority queues. One input file can result in multiple snapshots with the name *<PRIORITY>_<CRID>_<BATCHID>_<NODEID>_<MSUID>_<TIMESTAMP>_snap<N>*`.csqlite`. The input file contains the location where the actual file has been kept in the classification/content folder. The job also keeps a list of files that could not be fetched. **Note:** For On-prem File Servers, error logs are created in the *<Install directory>*`/log/fetch` folder. For cloud sources, check Connector logs.

Table B-2 Classification service jobs *(continued)*

Job	Description
`ClassifyFetchPauseJob`	Runs once every 30 seconds on any node that acts as the Classification Server. Refreshes the pause or resume status of fetch jobs as per the duration configured for content fetching.
`CancelClassifyRequestJob`	Runs every 20 seconds in Communication Service and Classification Service. Fetches the list of classification requests that are cancelled and distributes this request between Data Insight nodes. Before classifying files, all the classification jobs consult this list to identify the requests that are marked for cancellation. If they observe any canceled request in the new request that is submitted for classification, then that request is deleted.
`ClassifyJob`	Runs once every minute on any node that acts as a Classification Server. Checks the `classification/inbox` folder for input files submitted for classification folder and adds them to three separate priority queues. It picks a file from the highest queue in FIFO order, and starts classifying content using Veritas Information Classifier. All files in that input file are submitted for classification. Once all paths in the file have been classified, result of the classification and any resulting errors are written to a database in the `classification/outbox` folder.
`UpdateVICPolicyMapJob`	Runs every ten seconds on the Management Server. Ensures that Data Insight configuration database is in sync with the Classification Policy Manager.
`UpdateConfigJob`	Reconfigures jobs based on the configuration changes made on the Management Server.
`PredictJob`	Runs once every week on Sunday at 05.00 A.M. on the Indexer. Copies the prediction files from the temp output directory to a classification outbox.
`SLCreateBatchesJob`	Runs every 2 hours on the Indexer. Creates batches of files for the consumption of Veritas Information Classifier. These files are classified with high priority.

Table B-2 Classification service jobs *(continued)*

Job	Description
`ClassifyManage WorkloadJob`	Runs every one minute on the server that is assigned the role of a Classification Server. This job is enabled only on master Classification Server. Checks the classification or workload folder on master Classification Server and counts batches based on their priority. If the workload needs to be distributed, the job fetches a list of servers' in its pool and fetches the number of batches based on their priority in the `classification or inbox` folder. If the number of batches on any slave that have priority less than 10, then the job distributes the batches across that slave and copies them to the slave's the `classification or inbox` folder.

Appendix C

Troubleshooting classification

This appendix includes the following topics:

Classification logs and events

To troubleshoot errors, you can download the Data Insight logs relevant to a file server, share, OneDrive, Amazon S3, Box, SharePoint web application, or SharePoint site collection from the **Settings** tab of the Management Console. Or, monitor the events from the **Events** page.

Veritas Data Insight log files are located in the Data Insight installation directory,`<INSTALLDIR>\log`. Typically, the installation directory is located at `C:\Program Files\DataInsight\log`. On Linux, the logs are located at `/INSTALL/DataInsight/log`. For classification logs, see the following files:

- `vic.log`: This file contains log messages of Veritas Information Classifier.
- `vic.0.0.log`: This file contains log messages of the interaction that happens between Veritas Information Classifier and Data Insight.

Potential classification issues

Table C-1 Classification issues

Issue	Description/solution
Data Insight fails to classify content.	Check the following: ■ The correct classification policies are in place and the policies are enabled in Veritas Information Classifier. ■ Classification is enabled in Data Insight. ■ The Classification Server is assigned to the Collector that is monitoring the data source being classified. ■ The `DataInsightVICClient` and the `DataInsightVICServer` services are running on the Classification Server. ■ Appropriate safeguard settings are configured.
Data Insight is not able to fetch content.	Check the following: ■ The safeguard is active (that is the safeguard settings have been activated due to low disk space). If yes, make sure you have enough free disk space. ■ The pause window is not active. If yes, change the pause window or wait for the window to end for the content scan to resume. ■ All jobs are enabled. ■ Make sure respective services are running to fetch content.

Table C-1 Classification issues *(continued)*

Issue	Description/solution
Classification request gives a failure message.	Check the following: ■ If paths submitted for classification are valid. (Especially, when the paths are being uploaded using a CSV file.) Data Insight returns the path as invalid, if the file is offline, has been deleted, or if the shares/site collections/user accounts associated with the path is in the disabled state. ■ A Classification server is assigned to the Collector serving the data source being classified. ■ The `failed paths.csv` file available on the **Failure alert** pop up the details regarding the failure.
The Smart Classification requests that are in-progress are not getting completed for a long time.	Check the following: ■ The Classification Server is associated to the respective collector nodes. ■ At least one policy is enabled. ■ `DataInsightVIC` services are running on respective classification servers. ■ The Indexer node and respective the classification servers are able to communicate with each other. ■ Any of the jobs starting with `Classify*` are disabled on respective the Classification Server. ■ Sufficient free space exists on the respective classification servers.
Prediction job not generating a prediction CSV due to memory error.	Increase memory to at least 16 GB and let the job run one more time
The Smart Classification jobs - `Createfeaturesjob` and the `PredictJob` keep failing without processing anything.	Run `python.exe` located at `C:\Program Files\DataInsight\python`. You should see the following Python prompt: `Python 3.10.4 (tags/v3.10.4:9d38120, Mar 23 AMD64)] on win32`

Error codes

This section describes the error codes that Data Insight returns when you attempt to download the paths for which classification process failed.

Table C-2 Error codes

Error code	Description	Workaround
2	The file submitted for classification was not found.	Ensure that the file name is correct, and resubmit it for classification.
3	The path submitted for classification was not found.	Ensure that the path is correct, and resubmit it for classification.
5	Data Insight cannot fetch the content from the data source because the user does not have sufficient access permissions.	Ensure that the scanner credentials have read data permissions on the file that is submitted for classification.
201	After fetching the content, scanner cannot reset the timestamp due to insufficient privileges on the file path.	Ensure that the scanner credentials have read data and write attributes permissions on the file that is submitted for classification.
205	The input path is deleted.	Review the path to ensure that it is accessible, and resubmit it for classification.
209	The input path does not exist.	Review the path to ensure that it is valid and accessible, and resubmit it for classification.
210	The operation cannot be completed as the connection has timed out.	The operation timed out accessing SharePoint website or site collection. Contact the SharePoint administrator to reset the time out setting. If the issue persists, contact the Data Insight Support.
211	Access to this website is unavailable due to insufficient privileges.	Contact the SharePoint administrator to assign the privileges.

Table C-2 Error codes *(continued)*

Error code	Description	Workaround
215	The file cannot be downloaded due to an error.	Depending on the error description, try to rectify the problem. If the issue persists, contact the Data Insight Support.
256	Data Insight cannot fetch the file content because the shares/site collections/user accounts associated with the path is in the disabled state.	This error may occur when the device or the shares/site collections/user accounts associated with the path is disabled. To enable the device or shares/site collections/user accounts go to **Settings** > Filers/Cloud sources/SharePoint sources or shares/site collections/user accounts. Click the **Select Action** drop-down for the corresponding server in the servers listing table, and select **Enable**.
284	The last accessed timestamp of a file cannot be reset.	Contact the Data Insight Support.
285	Data Insight cannot fetch the file content because the file is offline.	None
286	File is larger than the permissible file size configured for classification.	The default file size that can be sent for classification is 50 MB. To configure the file size for classification, set the following global property: `configdb -O -J matrix.classify.fetch.max_size.mb -j <val_in_mb>`