Get started

NetApp Data Classification

NetApp
February 11, 2026

Table of Contents

Get started

Learn about NetApp Data Classification

NetApp Data Classification is a data governance service for the NetApp Console that scans your corporate on-premises and cloud data sources to map and classify data, and to identify private information. This can help reduce your security and compliance risk, decrease storage costs, and assist with your data migration projects.

> Beginning with version 1.31, Data Classification is available as a core capability within the NetApp Console. There's no additional charge. No Classification license or subscription is required.
> If you've been using legacy version 1.30 or earlier, that version is available until your subscription expires.

NetApp Console

Data Classification is accessible through the NetApp Console.

The NetApp Console provides centralized management of NetApp storage and data services across on-premises and cloud environments at enterprise grade. The Console is required to access and use NetApp data services. As a management interface, it enables you to manage many storage resources from one interface. Console administrators can control access to storage and services for all systems within the enterprise.

You don't need a license or subscription to start using NetApp Console and you only incur charges when you need to deploy Console agents in your cloud to ensure connectivity to your storage systems or NetApp data services. However, some NetApp data services accessible from the Console are licensed or subscription-based.

Learn more about the NetApp Console.

Features

Data Classification uses artificial intelligence (AI), natural language processing (NLP), and machine learning (ML) to understand the content that it scans in order to extract entities and categorize the content accordingly. This allows Data Classification to provide the following areas of functionality.

Learn about use cases for Data Classification.

Maintain compliance

Data Classification provides several tools that can help with your compliance efforts. You can use Data Classification to:

- Identify Personal Identifiable Information (PII).
- Identify a wide scope of sensitive personal information as required by GDPR, CCPA, PCI, and HIPAA privacy regulations.
- Respond to Data Subject Access Requests (DSAR) based on name or email address.

Strengthen security

Data Classification can identify data that is potentially at risk for being accessed for criminal purposes. You can

use Data Classification to:

- Identify all the files and directories (shares and folders) with open permissions that are exposed to your entire organization or to the public.
- Identify sensitive data that resides outside of the initial, dedicated location.
- Comply with data retention policies.
- Use *Policies* to automatically detect new security issues so security staff can take action immediately.

Optimize storage usage

Data Classification provides tools that can help with your storage total cost of ownership (TCO). You can use Data Classification to:

- Increase storage efficiency by identifying duplicate or non-business-related data.
- Save storage costs by identifying inactive data that you can tier to less expensive object storage. Learn more about tiering from Cloud Volumes ONTAP systems. Learn more about tiering from on-premises ONTAP systems.

Supported systems and data sources

Data Classification can scan and analyze structured and unstructured data from the following types of systems and data sources:

Systems

- Amazon FSx for NetApp ONTAP management
- Azure NetApp Files
- Cloud Volumes ONTAP (deployed in AWS, Azure, or GCP)
- On-premises ONTAP clusters
- StorageGRID
- Google Cloud NetApp Volumes

Data sources

- NetApp file shares
- Databases:
  - Amazon Relational Database Service (Amazon RDS)
  - MongoDB
  - MySQL
  - Oracle
  - PostgreSQL
  - SAP HANA
  - SQL Server (MSSQL)

Data Classification supports NFS versions 3.x, 4.0, and 4.1, and CIFS versions 1.x, 2.0, 2.1, and 3.0.

Cost

Data Classification is free to use. No Classification license or paid subscription is required.

Infrastructure costs

- Installing Data Classification in the cloud requires deploying a cloud instance, which results in charges from the cloud provider where it is deployed. See the type of instance that is deployed for each cloud provider. There is no cost if you install Data Classification on an on-premises system.
- Data Classification requires that you have deployed a Console agent. In many cases you already have a Console agent because of other storage and services you are using in the Console. The Console agent instance results in charges from the cloud provider where it's deployed. See the type of instance that is deployed for each cloud provider. There is no cost if you install the Console agent on an on-premises system.

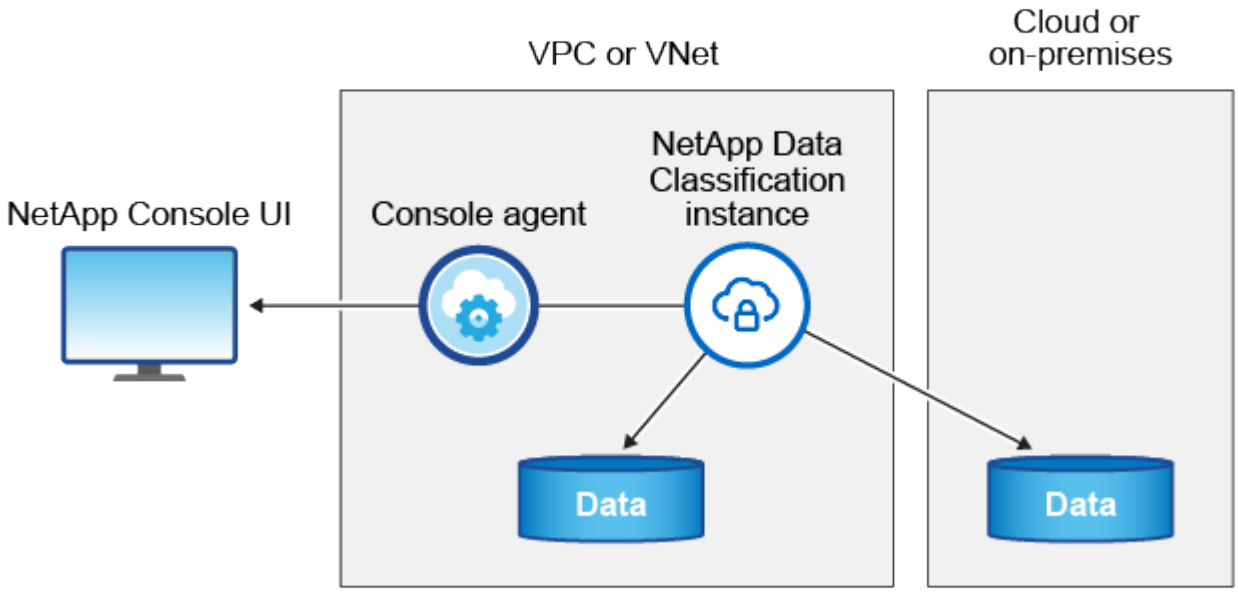
Data transfer costs

Data transfer costs depend on your setup. If the Data Classification instance and data source are in the same Availability Zone and region, then there are no data transfer costs. But if the data source, such as a Cloud Volumes ONTAP system, is in a *different* Availability Zone or region, then you'll be charged by your cloud provider for data transfer costs. See these links for more details:

- AWS: Amazon Elastic Compute Cloud (Amazon EC2) Pricing
- Microsoft Azure: Bandwidth Pricing Details
- Google Cloud: Storage Transfer Service pricing

The Data Classification instance

When you deploy Data Classification in the cloud, the Console deploys the instance in the same subnet as the Console agent. Learn more about the Console agent.



Note the following about the default instance:

- In AWS, Data Classification runs on an m6i.4xlarge instance with a 500 GiB GP2 disk. The operating

system image is Amazon Linux 2. When deployed in AWS, you can choose a smaller instance size if you are scanning a small amount of data.

- In Azure, Data Classification runs on a Standard_D16s_v3 VM with a 500 GiB disk. The operating system image is Ubuntu 22.04.
- In GCP, Data Classification runs on an n2-standard-16 VM with a 500 GiB Standard persistent disk. The operating system image is Ubuntu 22.04.
- In regions where the default instance isn't available, Data Classification runs on an alternate instance. See the alternate instance types.
- The instance is named *CloudCompliance* with a generated hash (UUID) concatenated to it. For example: *CloudCompliance-16bb6564-38ad-4080-9a92-36f5fd2f71c7*
- Only one Data Classification instance is deployed per Console Agent.

You can also deploy Data Classification on a Linux host on your premises or on a host in your preferred cloud provider. The software functions exactly the same way regardless of which installation method you choose. Upgrades of Data Classification software are automated as long as the instance has internet access.

The instance should remain running at all times because Data Classification continuously scans the data.

Deploy on different instance types

Review the following specifications for instance types:

System size	Specs	Limitations
Extra Large	32 CPUs, 128 GB RAM, 1 TiB SSD	Can scan up to 500 million files.
Large (default)	16 CPUs, 64 GB RAM, 500 GiB SSD	Can scan up to 250 million files.

When deploying Data Classification in Azure or GCP, email ng-contact-data-sense@netapp.com for assistance if you want to use a smaller instance type.

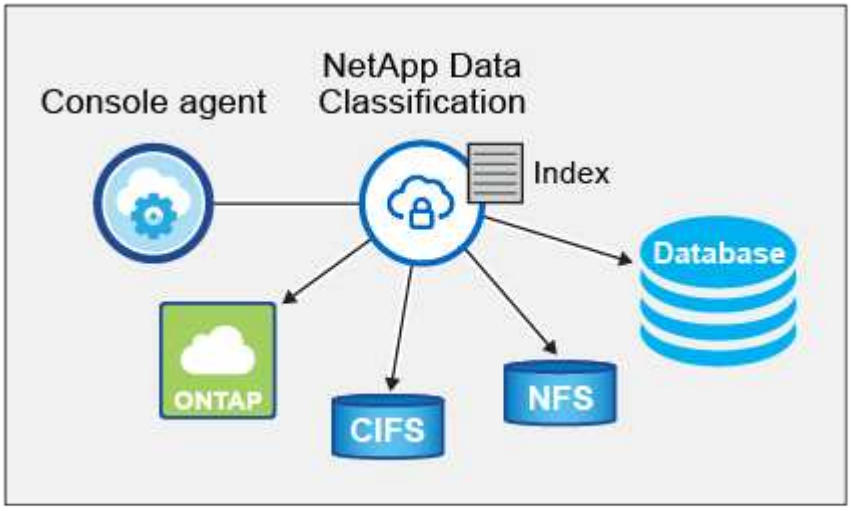
How Data Classification scanning works

At a high-level, Data Classification scanning works like this:

1. You deploy an instance of Data Classification in the Console.
2. You enable high-level mapping (called *Mapping only* scans) or deep-level scanning (called *Map & Classify* scans) on one or more data sources.
3. Data Classification scans data using an AI learning process.
4. You use the provided dashboards and reporting tools to help in your compliance and governance efforts.

After you enable Data Classification and select the repositories that you want to scan (these are the volumes, database schemas, or other user data), it immediately starts scanning the data to identify personal and sensitive data. You should focus on scanning live production data in most cases instead of backups, mirrors, or DR sites. Then Data Classification maps your organizational data, categorizes each file, and identifies and extracts entities and predefined patterns in the data. The result of the scan is an index of personal information, sensitive personal information, data categories, and file types.

Data Classification connects to the data like any other client by mounting NFS and CIFS volumes. NFS volumes are automatically accessed as read-only, while you need to provide Active Directory credentials to scan CIFS volumes.



After the initial scan, Data Classification continuously scans your data in a round-robin fashion to detect incremental changes. This is why it's important to keep the instance running.

You can enable and disable scans at the volume level or the database schema level.

> Data Classification does not impose a limit on the amount of data it can scan. Each Console agent supports scanning and displaying 500 TiB of data. To scan more than 500 TiB of data, install another Console agent then deploy another Data Classification instance.
> The Console UI displays data from a single connector. For tips on viewing data from multiple Console agents, see Work with multiple Console agents.

What's the difference between Mapping and Classification scans

You can conduct two types of scans in Data Classification:

- **Mapping-only scans** provide only a high-level overview of your data and are performed on selected data sources. Mapping-only scans take less time than map and classify scans because they do not access files to see the data inside. You might want to do this initially to identify areas of research and then perform a Map & Classify scan on those areas.
- **Map & Classify scans** provide deep-level scanning of your data.

For details about the differences between Mapping and Classification scans, see What's the difference between Mapping and Classification scans?.

Information that Data Classification categorizes

Data Classification collects, indexes, and assigns categories to the following data:

- **Standard metadata** about files: the file type, its size, creation and modification dates, and so on.
- **Personal data**: Personally identifiable information (PII) such as email addresses, identification numbers, or credit card numbers, which Data Classification identifies using specific words, strings, and patterns in the files. Learn more about personal data.

- **Sensitive personal data**: Special types of sensitive personal information (SPII), such as health data, ethnic origin, or political opinions, as defined by General Data Protection Regulation (GDPR) and other privacy regulations. Learn more about sensitive personal data.
- **Categories**: Data Classification takes the data that it scanned and divides it into different types of categories. Categories are topics based on AI analysis of the content and metadata of each file. Learn more about categories.
- **Name entity recognition**: Data Classification uses AI to extract people's natural names from documents. Learn about responding to Data Subject Access Requests.

Networking overview

Data Classification deploys a single server, or cluster, wherever you choose: in the cloud or on premises. The servers connect via standard protocols to the data sources and index the findings in an Elasticsearch cluster, which is also deployed on the same servers. This enables support for multi-cloud, cross-cloud, private cloud, and on-premises environments.

The Console deploys the Data Classification instance with a security group that enables inbound HTTP connections from the Console agent.

When you use the Console in SaaS mode, the connection to the Console is served over HTTPS, and the private data sent between your browser and the Data Classification instance are secured with end-to-end encryption using TLS 1.2, which means NetApp and third parties can't read it.

Outbound rules are completely open. Internet access is needed to install and upgrade the Data Classification software and to send usage metrics.

If you have strict networking requirements, learn about the endpoints that Data Classification contacts.

Access NetApp Data Classification

You can access the NetApp Data Classification through the NetApp Console.

To sign in to the Console, you can use your NetApp Support Site credentials or you can sign up for a NetApp Console login using your email and a password. Learn more about logging in to the Console.

Specific tasks require specific Console user roles. Learn about Console access roles for all services.

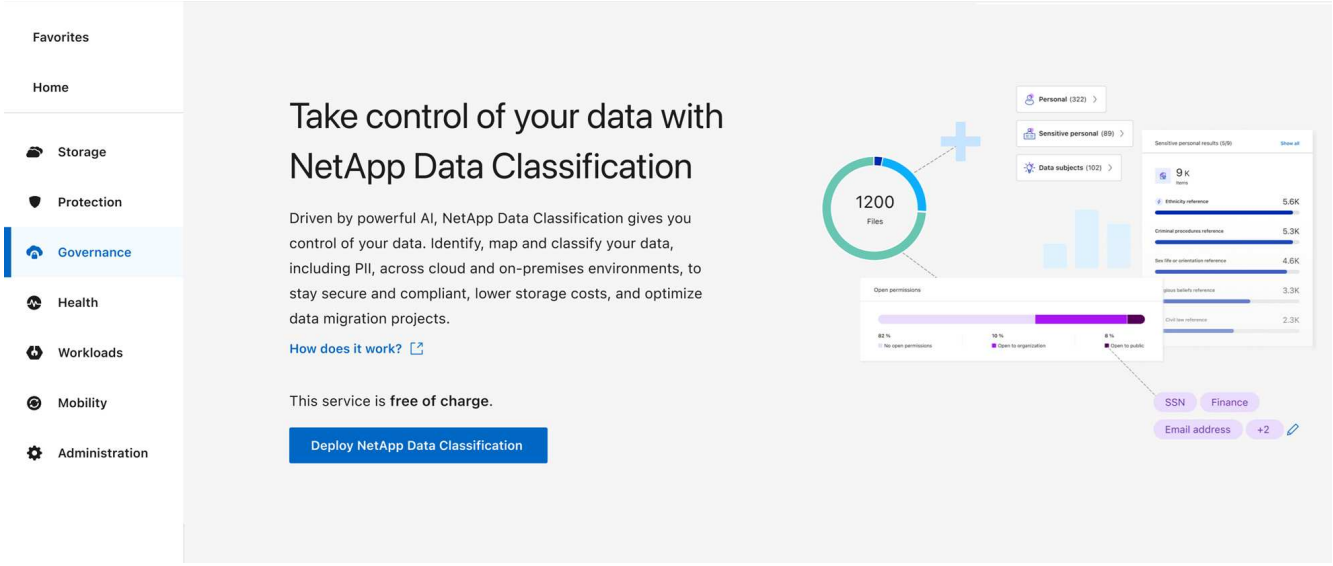
Before you begin

- You should add a Console agent.
- Understand which Data Classification deployment style suits your workload.

Steps

1. In a web browser, navigate to the Console.
2. Log in to the Console.
3. From the main page of the NetApp Console, select **Governance** > **Data Classification**.
4. If this is your first time accessing Data Classification, the landing page appears.

   Select **Deploy Classification On-Premises or Cloud** to begin deploying your classification instance. For more information, see Which Data Classification deployment should you use?



Otherwise, the Data Classification Dashboard appears.

Deploy Data Classification

Which NetApp Data Classification deployment should you use?

You can deploy NetApp Data Classification in different ways. Learn which method meets your needs.

Data Classification can be deployed in the following ways:

- Deploy in the cloud using the Console. The Console deploys the Data Classification instance in the same cloud provider network as the Console agent.
- Install on a Linux host with internet access. Install Data Classification on a Linux host in your network, or on a Linux host in the cloud, that has internet access. This type of installation may be a good option if you prefer to scan on-premises ONTAP systems using a Data Classification instance that's also located on premises, though this isn't a requirement.
- Install on a Linux host in an on-premises site without internet access, also known as *private mode.* This type of installation, which uses an installation script, has no connectivity to the Console SaaS layer.

> BlueXP private mode (legacy BlueXP interface) is typically used with on-premises environments that have no internet connection and with secure cloud regions, which includes AWS Secret Cloud, AWS Top Secret Cloud, and Azure IL6. NetApp continues to support these environments with the legacy BlueXP interface. For private mode documentation in the legacy BlueXP interface, see PDF documentation for BlueXP private mode.

Both the installation on a Linux host with internet access and the on-premises installation on a Linux host without internet access use an installation script. The script starts by checking if the system and environment meet the prerequisites. If the prerequisites are met, the installation starts. If you would like to verify the prerequisites independently of running the Data Classification installation, there is a separate software package you can download that only tests for the prerequisites.

Refer to Check that your Linux host is ready to install Data Classification.

Deploy NetApp Data Classification in the cloud using the NetApp Console

You can deploy NetApp Data Classification in the cloud with the NetApp Console. The Console deploys the Data Classification instance in the same cloud provider network as the Console agent.

Note that you can also install Data Classification on a Linux host that has internet access. This type of installation may be a good option if you prefer to scan on-premises ONTAP systems using a Data Classification instance that's also located on premises — but this is not a requirement. The software functions exactly the same way regardless of which installation method you choose.

Quick start

Get started quickly by following these steps, or scroll down to the remaining sections for full details.

Create a Console agent

If you don't already have a Console agent, create one. See creating a Console agent in AWS, creating a Console agent in Azure, or creating a Console agent in GCP.

You can also install the Console agent on-premises on a Linux host in your network or on a Linux host in the cloud.

Prerequisites

Ensure that your environment can meet the prerequisites. This includes outbound internet access for the instance, connectivity between the Console agent and Data Classification over port 443, and more. See the complete list.

Deploy Data Classification

Launch the installation wizard to deploy the Data Classification instance in the cloud.

Create a Console agent

If you don't already have a Console agent, create a Console agent in your cloud provider. See creating a Console agent in AWS or creating a Console agent in Azure, or creating a Console agent in GCP. In most cases you will probably have a Console agent set up before you attempt to activate Data Classification because most Console features require a Console agent, but there are cases where you'll you need to set one up now.

There are some scenarios where you have to use a Console agent that's deployed in a specific cloud provider:

- When scanning data in Cloud Volumes ONTAP in AWS or Amazon FSx for ONTAP buckets, you use a Console agent in AWS.
- When scanning data in Cloud Volumes ONTAP in Azure or in Azure NetApp Files, you use a Console agent in Azure.
  - For Azure NetApp Files, it must be deployed in the same region as the volumes you wish to scan.
- When scanning data in Cloud Volumes ONTAP in GCP, you use a Console agent in GCP.

On-prem ONTAP systems, NetApp file shares, and databases can be scanned when using any of these cloud Console agents.

Note that you can also install the Console agent on-premises on a Linux host in your network or in the cloud. Some users planning to install Data Classification on-prem may also choose to install the Console agent on-premises.

There might be situations where you need to use multiple Console agents.

> Data Classification does not impose a limit on the amount of data it can scan. Each Console agent supports scanning and displaying 500 TiB of data. To scan more than 500 TiB of data, install another Console agent then deploy another Data Classification instance.
> The Console UI displays data from a single connector. For tips on viewing data from multiple Console agents, see Work with multiple Console agents.

Government region support

Data Classification is supported when the Console agent is deployed in a Government region (AWS GovCloud, Azure Gov, or Azure DoD). When deployed in this manner, Data Classification has the following restrictions:

Learn about deploying the Console agent in a Government region.

Prerequisites

Review the following prerequisites to make sure that you have a supported configuration before you deploy Data Classification in the cloud. When you deploy Data Classification in the cloud, it's located in the same subnet as the Console agent.

Enable outbound internet access from Data Classification

Data Classification requires outbound internet access. If your virtual or physical network uses a proxy server for internet access, ensure that the Data Classification instance has outbound internet access to contact the following endpoints. The proxy must be non-transparent. Transparent proxies are not currently supported.

Review the appropriate table below depending on whether you are deploying Data Classification in AWS, Azure, or GCP.

Required endpoints for AWS

Endpoints	Purpose
https://api.console.netapp.com	Communication with the Console service, which includes NetApp accounts.
https://netapp-cloud-account.auth0.com https://auth0.com	Communication with the Console website for centralized user authentication.
https://cloud-compliance-support-netapp.s3.us-west-2.amazonaws.com https://hub.docker.com https://auth.docker.io https://registry-1.docker.io https://index.docker.io/ https://dseasb33srnrn.cloudfront.net/ https://production.cloudflare.docker.com/	Provides access to software images, manifests, and templates.
https://kinesis.us-east-1.amazonaws.com	Enables NetApp to stream data from audit records.
https://cognito-idp.us-east-1.amazonaws.com https://cognito-identity.us-east-1.amazonaws.com https://user-feedback-store-prod.s3.us-west-2.amazonaws.com https://customer-data-production.s3.us-west-2.amazonaws.com	Enables Data Classification to access and download manifests and templates, and to send logs and metrics.

Required endpoints for Azure

Endpoints	Purpose
https://api.console.netapp.com	Communication with the Console service, which includes NetApp accounts.
https://netapp-cloud-account.auth0.com https://auth0.com	Communication with the Console website for centralized user authentication.
https://support.compliance.api.console.netapp.com/ https://hub.docker.com https://auth.docker.io https://registry-1.docker.io https://index.docker.io/ https://dseasb33srnrn.cloudfront.net/ https://production.cloudflare.docker.com/	Provides access to software images, manifests, templates, and to send logs and metrics.
https://support.compliance.api.console.netapp.com/	Enables NetApp to stream data from audit records.

Required endpoints for GCP

Endpoints	Purpose
https://api.console.netapp.com	Communication with the Console service, which includes NetApp accounts.

Endpoints	Purpose
https://netapp-cloud-account.auth0.com https://auth0.com	Communication with the Console website for centralized user authentication.
https://support.compliance.api.console.netapp.com/ https://hub.docker.com https://auth.docker.io https://registry-1.docker.io https://index.docker.io/ https://dseasb33srnrn.cloudfront.net/ https://production.cloudflare.docker.com/	Provides access to software images, manifests, templates, and to send logs and metrics.
https://support.compliance.api.console.netapp.com/	Enables NetApp to stream data from audit records.

Ensure that Data Classification has the required permissions

Ensure that Data Classification has permissions to deploy resources and create security groups for the Data Classification instance.

- Google Cloud permissions
- AWS permissions
- Azure permissions

Ensure that the Console agent can access Data Classification

Ensure connectivity between the Console agent and the Data Classification instance. The security group for the Console agent must allow inbound and outbound traffic over port 443 to and from the Data Classification instance. This connection enables deployment of the Data Classification instance and enables you to view information in the Compliance and Governance tabs. Data Classification is supported in Government regions in AWS and Azure.

Additional inbound and outbound security group rules are required for AWS and AWS GovCloud deployments. See Rules for the Console agent in AWS for details.

Additional inbound and outbound security group rules are required for Azure and Azure Government deployments. See Rules for the Console agent in Azure for details.

Ensure you can keep Data Classification running

The Data Classification instance needs to stay on to continuously scan your data.

Ensure web browser connectivity to Data Classification

After Data Classification is enabled, ensure that users access the Console interface from a host that has a connection to the Data Classification instance.

The Data Classification instance uses a private IP address to ensure that the indexed data isn't accessible to the internet. As a result, the web browser that you use to access the Console must have a connection to that private IP address. That connection can come from a direct connection to your cloud provider (for example, a VPN), or from a host that's inside the same network as the Data Classification instance.

Check your vCPU limits

Ensure that your cloud provider's vCPU limit allows for the deployment of an instance with the necessary number of cores. You'll need to verify the vCPU limit for the relevant instance family in the region where the

Console is running. See the required instance types.

See the following links for more details on vCPU limits:

- AWS documentation: Amazon EC2 service quotas
- Azure documentation: Virtual machine vCPU quotas
- Google Cloud documentation: Resource quotas

Deploy Data Classification in the cloud

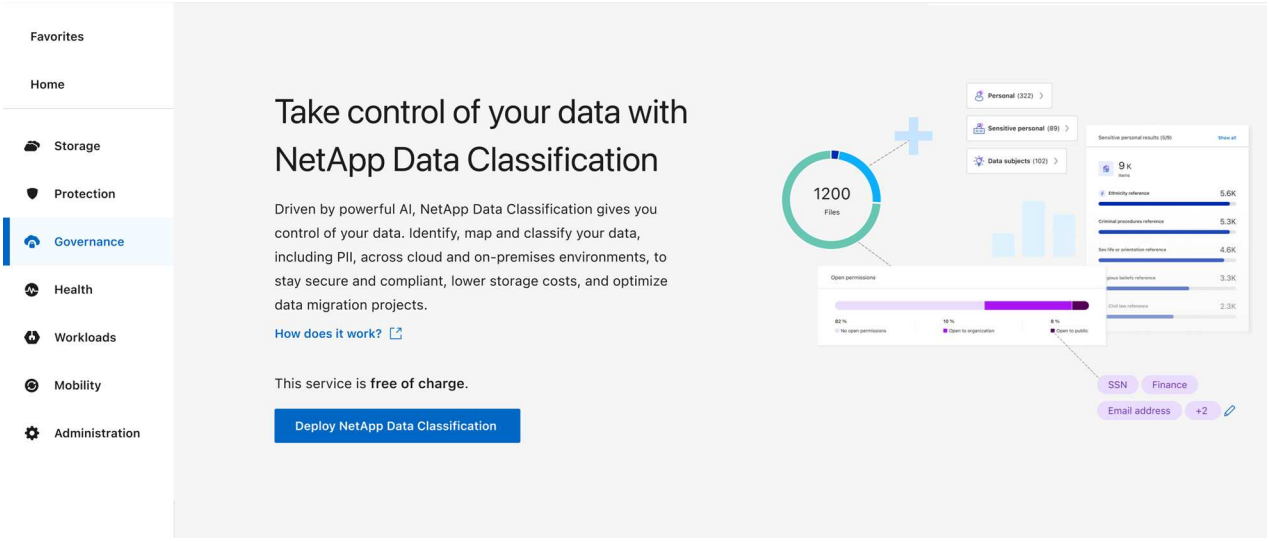
Follow these steps to deploy an instance of Data Classification in the cloud. The Console agent will deploy the instance in the cloud, and then install Data Classification software on that instance.

In regions where the default instance type isn't available, Data Classification runs on an alternate instance type.

Deploy in AWS

Steps

1. From the main page of Data Classification, select **Deploy Classification On-Premises or Cloud**.

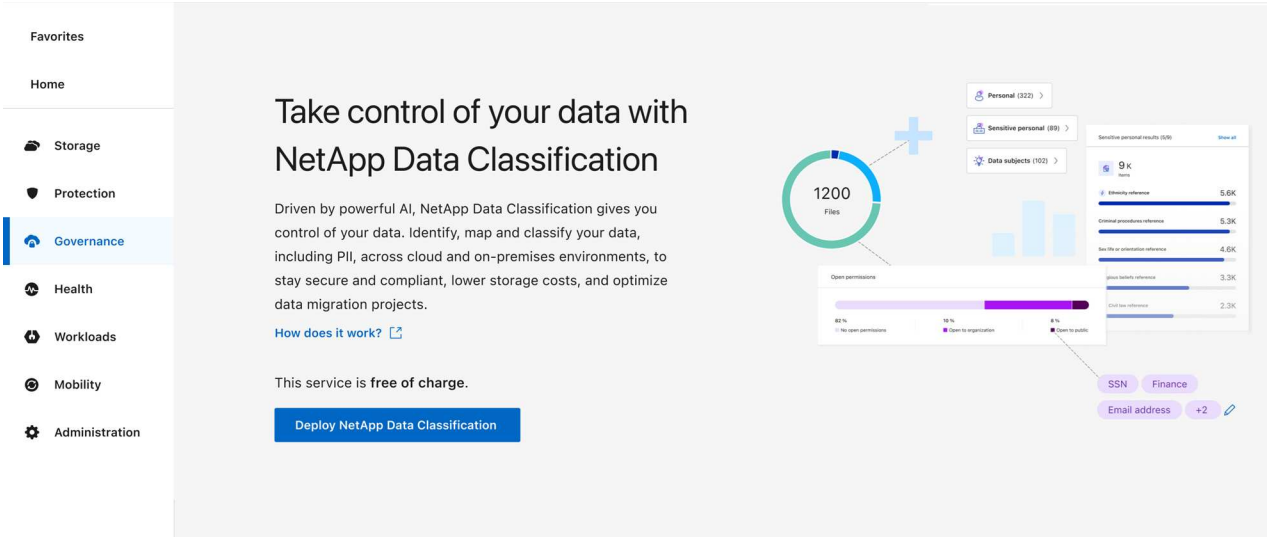


2. From the *Installation* page, select **Deploy > Deploy** to use the "Large" instance size and start the cloud deployment wizard.
3. The wizard displays progress as it goes through the deployment steps. When inputs are required or if it encounters issues, you are prompted.
4. When the instance is deployed and Data Classification is installed, select **Continue to configuration** to go to the *Configuration* page.

Deploy in Azure

Steps

1. From the main page of Data Classification, select **Deploy Classification On-Premises or Cloud**.



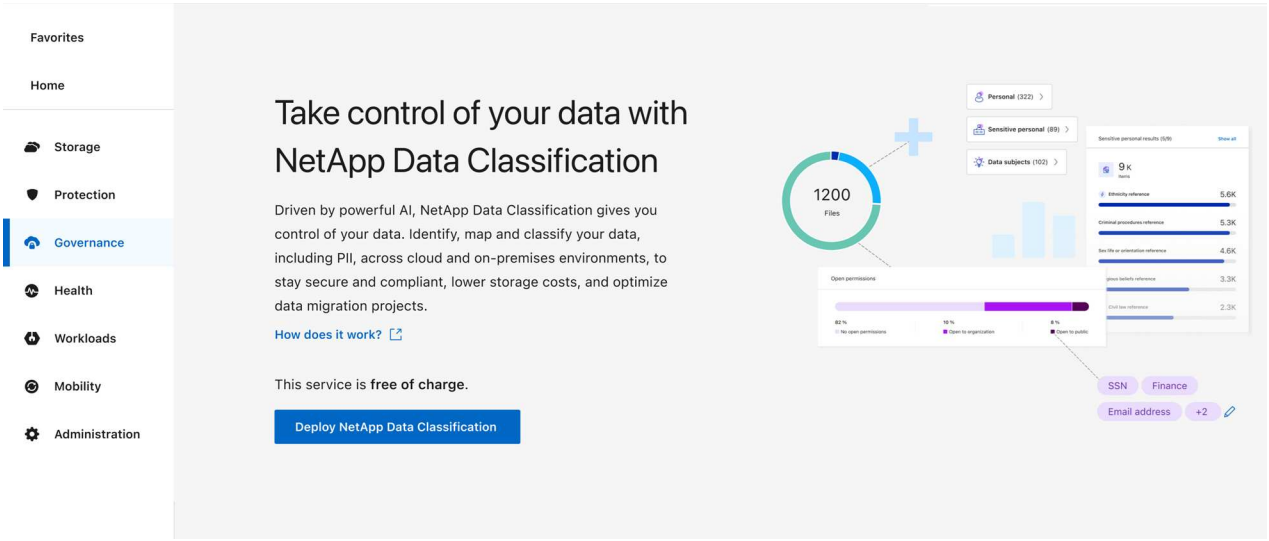
2. Select **Deploy** to start the cloud deployment wizard.
3. The wizard displays progress as it goes through the deployment steps. It will stop and prompt for input if it runs into any issues.

4. When the instance is deployed and Data Classification is installed, select **Continue to configuration** to go to the *Configuration* page.

Deploy in Google Cloud

Steps

1. From the main page of Data Classification, select **Governance > Classification**.
2. Select **Deploy Classification On-Premises or Cloud**.



3. Select **Deploy** to start the cloud deployment wizard.
4. The wizard displays progress as it goes through the deployment steps. It will stop and prompt for input if it runs into any issues.
5. When the instance is deployed and Data Classification is installed, select **Continue to configuration** to go to the *Configuration* page.

Result

The Console deploys the Data Classification instance in your cloud provider.

Upgrades to the Console agent and Data Classification software is automated as long as the instances have internet connectivity.

What's Next

From the Configuration page you can select the data sources that you want to scan.

Install NetApp Data Classification on a host that has internet access

To deploy NetApp Data Classification on a Linux host in your network or on a Linux host in the cloud that has internet access, you need deploy the Linux host manually in your network or in the cloud.

The on-premises installation is a good option if you prefer to scan on-premises ONTAP systems using a Data Classification instance that's also located on premises. This is not a requirement. The software functions the same regardless of which installation method you choose.

The Data Classification installation script starts by checking if the system and environment meet the required

prerequisites. If the prerequisites are all met, then the installation starts. If you would like to verify the prerequisites independently of running the Data Classification installation, there is a separate software package you can download that only tests for the prerequisites. See how to check if your Linux host is ready to install Data Classification.

The typical installation on a Linux host *in your premises* has the following components and connections.

The typical installation on a Linux host *in the cloud* has the following components and connections.

Quick start

Get started quickly by following these steps, or scroll down to the remaining sections for full details.

Create a Console agent

If you don't already have a Console agent , deploy the Console agent on-premises on a Linux host in your network, or on a Linux host in the cloud.

You can also create a Console agent with your cloud provider. See creating a Console agent in AWS, creating a Console agent in Azure, or creating a Console agent in GCP.

Review prerequisites

Ensure that your environment can meet the prerequisites. This includes outbound internet access for the instance, connectivity between the Console agent and Data Classification over port 443, and more. See the complete list.

You also need a Linux system that meets the following requirements.

Download and deploy Data Classification

Download the Cloud Data Classification software from the NetApp Support Site and copy the installer file to the Linux host you plan to use. Then launch the installation wizard and follow the prompts to deploy the Data Classification instance.

Create a Console agent

A Console agent is required before you can install and use Data Classification. In most cases you'll probably have a Console agent set up before you attempt to activate Data Classification because most Console features require a Console agent, but there are cases where you'll you need to set one up now.

To create one in your cloud provider environment, see creating a Console agent in AWS, creating a Console agent in Azure, or creating a Console agent in GCP.

There are some scenarios where you have to use a Console agent that's deployed in a specific cloud provider:

- When scanning data in Cloud Volumes ONTAP in AWS or Amazon FSx for ONTAP, you use a Console agent in AWS.

- When scanning data in Cloud Volumes ONTAP in Azure or in Azure NetApp Files, you use a Console agent in Azure.

  For Azure NetApp Files, it must be deployed in the same region as the volumes you wish to scan.

- When scanning data in Cloud Volumes ONTAP in GCP, you use a Console agent in GCP.

On-prem ONTAP systems, NetApp file shares and database accounts can be scanned using any of these cloud Console agents.

Note that you can also deploy the Console agent on-premises on a Linux host in your network or on a Linux host in the cloud. Some users planning to install Data Classification on-prem may also choose to install the Console agent on-prem.

You'll need the IP address or host name of the Console agent system when installing Data Classification. You'll have this information if you installed the Console agent in your premises. If the Console agent is deployed in the cloud, you can find this information from the Console: select the Help icon then **Support** then **Console agent**.

Prepare the Linux host system

Data Classification software must run on a host that meets specific operating system requirements, RAM requirements, software requirements, and so on. The Linux host can be in your network, or in the cloud.

Ensure that you can keep Data Classification running. The Data Classification machine needs to stay on to continuously scan your data.

- Data Classification must be on a dedicated host. The host can't be shared with other applications or third-party software such as antivirus.
- Choose the size that aligns with the data set you plan to scan with Data Classification.

System size	CPU	RAM (swap memory must be disabled)	Disk
Extra Large	32 CPUs	128 GB RAM	• 1 TiB SSD on /, or 100 GiB available on /opt • 895 GiB available on /var/lib/docker • 5 GiB on /tmp • **For Podman, 30 GB on /var/tmp**
Large	16 CPUs	64 GB RAM	• 500 GiB SSD on /, or 100 GiB available on /opt • 400 GiB available on /var/lib/docker or for Podman /var/lib/containers • 5 GiB on /tmp • **For Podman, 30 GB on /var/tmp**

- When deploying a compute instance in the cloud for your Data Classification installation, it's recommended you use a system that meets the "Large" system requirements above:
  - **Amazon Elastic Compute Cloud (Amazon EC2) instance type**: "m6i.4xlarge". See additional AWS instance types.
  - **Azure VM size**: "Standard_D16s_v3". See additional Azure instance types.
  - **GCP machine type**: "n2-standard-16". See additional GCP instance types.
- **UNIX folder permissions**: The following minimum UNIX permissions are required:

Folder	Minimum permissions
/tmp	`rwxrwxrwt`
/opt	`rwxr-xr-x`
/var/lib/docker	`rwx------`
/usr/lib/systemd/system	`rwxr-xr-x`

- **Operating system**:
  - The following operating systems require using the Docker container engine:
    - Red Hat Enterprise Linux version 7.8 and 7.9
    - Ubuntu 22.04 (requires Data Classification version 1.23 or greater)
    - Ubuntu 24.04 (requires Data Classification version 1.23 or greater)
  - The following operating systems require using the Podman container engine, and they require Data Classification version 1.30 or greater:
    - Red Hat Enterprise Linux version 8.8, 8.10, 9.0, 9.1, 9.2, 9.3, 9.4, 9.5, and 9.6.
  - Advanced Vector Extensions (AVX2) must be enabled on the host system.
- **Red Hat Subscription Management**: The host must be registered with Red Hat Subscription Management. If it's not registered, the system can't access repositories to update required 3rd-party software during installation.
- **Additional software**: You must install the following software on the host before you install Data Classification:
  - Depending on the OS you are using, you need to install one of the container engines:
    - Docker Engine version 19.3.1 or greater. View installation instructions.
    - Podman version 4 or greater. To install Podman, enter (`sudo yum install podman netavark -y`).
- Python version 3.6 or greater. View installation instructions.
  - **NTP considerations**: NetApp recommends configuring the Data Classification system to use a Network Time Protocol (NTP) service. The time must be synchronized between the Data Classification system and the Console agent system.
- **Firewalld considerations**: If you are planning to use `firewalld`, we recommend that you enable it before installing Data Classification. Run the following commands to configure `firewalld` so that it is compatible with Data Classification:

```
firewall-cmd --permanent --add-service=http
firewall-cmd --permanent --add-service=https
firewall-cmd --permanent --add-port=80/tcp
firewall-cmd --permanent --add-port=8080/tcp
firewall-cmd --permanent --add-port=443/tcp
firewall-cmd --reload
```

If you're planning to use additional Data Classification hosts as scanner nodes, add these rules to your primary system at this time:

```
firewall-cmd --permanent --add-port=2377/tcp
firewall-cmd --permanent --add-port=7946/udp
firewall-cmd --permanent --add-port=7946/tcp
firewall-cmd --permanent --add-port=4789/udp
```

Note that you must restart Docker or Podman whenever you enable or update `firewalld` settings.

The IP address of the Data Classification host system can't be changed after installation.

Enable outbound internet access from Data Classification

Data Classification requires outbound internet access. If your virtual or physical network uses a proxy server for internet access, ensure that the Data Classification instance has outbound internet access to contact the following endpoints.

Endpoints	Purpose
https://api.console.netapp.com	Communication with the Console, which includes NetApp accounts.
https://netapp-cloud-account.auth0.com https://auth0.com	Communication with the Console website for centralized user authentication.
https://support.compliance.api.bluexp.netapp.com/ https://hub.docker.com https://auth.docker.io https://registry-1.docker.io https://index.docker.io/ https://dseasb33srnrn.cloudfront.net/ https://production.cloudflare.docker.com/	Provides access to software images, manifests, templates, and to send logs and metrics.
https://support.compliance.api.bluexp.netapp.com/	Enables NetApp to stream data from audit records.
https://github.com/docker https://download.docker.com	Provides prerequisite packages for docker installation.
http://packages.ubuntu.com/ http://archive.ubuntu.com	Provides prerequisite packages for Ubuntu installation.

Verify that all required ports are enabled

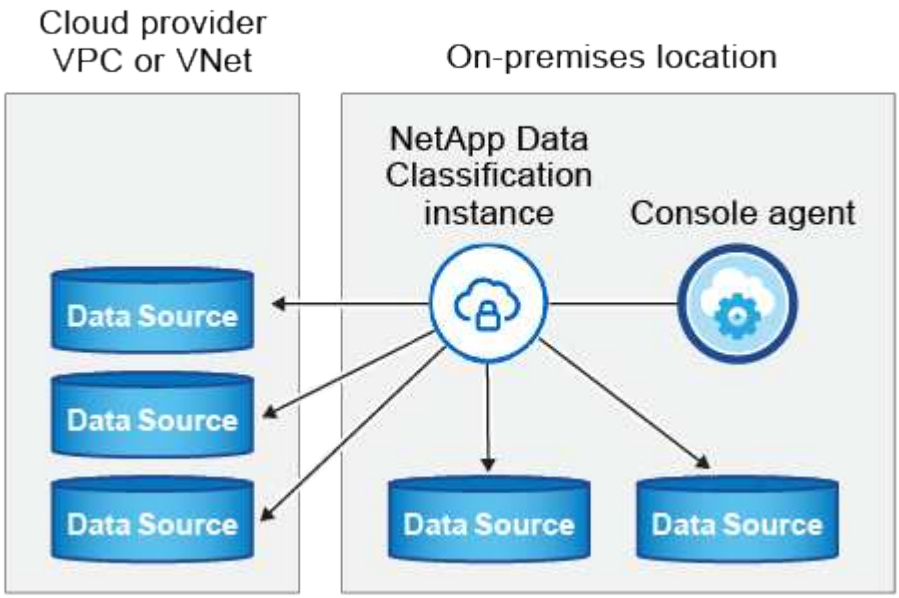
You must ensure that all required ports are open for communication between the Console agent, Data Classification, Active Directory, and your data sources.

Connection Type	Ports	Description
Console agent <> Data Classification	8080 (TCP), 443 (TCP), and 80. 9000	The firewall or routing rules for the Console agent must allow inbound and outbound traffic over port 443 to and from the Data Classification instance. Make sure port 8080 is open so you can see the installation progress in the Console. If a firewall is used on the Linux host, port 9000 is required for internal processes within an Ubuntu server.
Console agent <> ONTAP cluster (NAS)	443 (TCP)	The Console discovers ONTAP clusters using HTTPS. If you use custom firewall policies, they must meet the following requirements: • The Console agent host must allow outbound HTTPS access through port 443. If the Console agent is in the cloud, all outbound communication is allowed by the predefined firewall or routing rules. • The ONTAP cluster must allow inbound HTTPS access through port 443. The default "mgmt" firewall policy allows inbound HTTPS access from all IP addresses. If you modified this default policy, or if you created your own firewall policy, you must associate the HTTPS protocol with that policy and enable access from the Console agent host.
Data Classification <> ONTAP cluster	• For NFS - 111 (TCP\UDP) and 2049 (TCP\UDP) • For CIFS - 139 (TCP\UDP) and 445 (TCP\UDP)	Data Classification needs a network connection to each Cloud Volumes ONTAP subnet or on-prem ONTAP system. Firewalls or routing rules for Cloud Volumes ONTAP must allow inbound connections from the Data Classification instance. Make sure these ports are open to the Data Classification instance: • For NFS - 111 and 2049 • For CIFS - 139 and 445 NFS volume export policies must allow access from the Data Classification instance.

Connection Type	Ports	Description
Data Classification <> Active Directory	389 (TCP & UDP), 636 (TCP), 3268 (TCP), and 3269 (TCP)	You must have an Active Directory already set up for the users in your company. Additionally, Data Classification needs Active Directory credentials to scan CIFS volumes. You must have the information for the Active Directory: • DNS Server IP Address, or multiple IP Addresses • User Name and Password for the server • Domain Name (Active Directory Name) • Whether you are using secure LDAP (LDAPS) or not • LDAP Server Port (typically 389 for LDAP, and 636 for secure LDAP)

Install Data Classification on the Linux host

For typical configurations you'll install the software on a single host system. See those steps here.



See Preparing the Linux host system and Reviewing prerequisites for the full list of requirements before you deploy Data Classification.

Upgrades to Data Classification software is automated as long as the instance has internet connectivity.

Data Classification is currently unable to scan S3 buckets, Azure NetApp Files, or FSx for ONTAP when the software is installed on premises. In these cases you'll need to deploy a separate Console agent and instance of Data Classification in the cloud and switch between Connectors for your different data sources.

Single-host installation for typical configurations

Review the requirements and follow these steps when installing Data Classification software on a single on-premises host.

Watch this video to see how to install Data Classification.

Note that all installation activities are logged when installing Data Classification. If you run into any issues during installation, you can view the contents of the installation audit log. It is written to `/opt/netapp/install_logs/`.

Before you begin

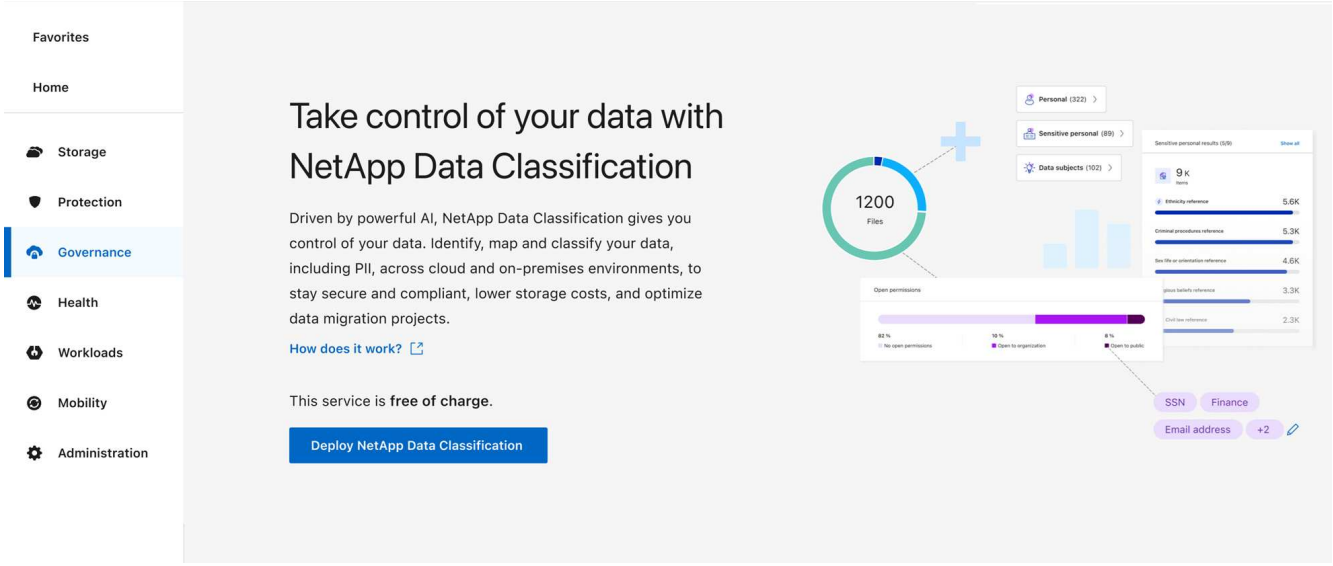
- Verify that your Linux system meets the host requirements.
- Verify that the system has the two prerequisite software packages installed (Docker Engine or Podman, and Python 3).
- Make sure you have root privileges on the Linux system.
- If you're using a proxy for access to the internet:
  - You'll need the proxy server information (IP address or host name, connection port, connection scheme: https or http, user name and password).
  - If the proxy is performing TLS interception, you'll need to know the path on the Data Classification Linux system where the TLS CA certificates are stored.
  - The proxy must be non-transparent. Data Classification does not currently support transparent proxies.
  - The user must be a local user. Domain users are not supported.
- Verify that your offline environment meets the required permissions and connectivity.

Steps

1. Download the Data Classification software from the NetApp Support Site. The file you should select is named **DATASENSE-INSTALLER-<version>.tar.gz**.
2. Copy the installer file to the Linux host you plan to use (using `scp` or some other method).
3. Unzip the installer file on the host machine, for example:

   ```
   tar -xzf DATASENSE-INSTALLER-V1.25.0.tar.gz
   ```

4. In the Console, select **Governance > Classification**.
5. Select **Deploy Classification On-Premises or Cloud**.



6. Depending on whether you are installing Data Classification on an instance you prepared in the cloud or on an instance you prepared in your premises, select the appropriate **Deploy** option to start the Data Classification installation.
7. The *Deploy Data Classification On Premises* dialog is displayed. Copy the provided command (for example: `sudo ./install.sh -a 12345 -c 27AG75 -t 2198qq`) and paste it in a text file so you can use it later. Then select **Close** to dismiss the dialog.
8. On the host machine, enter the command you copied and then follow a series of prompts, or you can provide the full command including all required parameters as command line arguments.

   Note that the installer performs a pre-check to make sure your system and networking requirements are in place for a successful installation. Watch this video to understand the pre-check messages and implications.

Enter parameters as prompted:	Enter the full command:
1. Paste the command you copied from step 7: `sudo ./install.sh -a <account_id> -c <client_id> -t <user_token>` If you are installing on a cloud instance (not on your premises), add `--manual-cloud -install <cloud_provider>`. 2. Enter the IP address or host name of the Data Classification host machine so it can be accessed by the Console agent system. 3. Enter the IP address or host name of the Console agent host machine so it can be accessed by the Data Classification system. 4. Enter proxy details as prompted. If your Console agent already uses a proxy, there is no need to enter this information again here since Data Classification will automatically use the proxy used by the Console agent.	Alternatively, you can create the whole command in advance, providing the necessary host and proxy parameters: `sudo ./install.sh -a <account_id> -c <client_id> -t <user_token> --host <ds_host> --manager-host <cm_host> --manual-cloud-install <cloud_provider> --proxy-host <proxy_host> --proxy-port <proxy_port> --proxy-scheme <proxy_scheme> --proxy -user <proxy_user> --proxy-password <proxy_password> --cacert-folder-path <ca_cert_dir>`

Variable values:

- *account_id* = NetApp Account ID
- *client_id* = Console agent Client ID (add the suffix "clients" to the client ID if it not already there)
- *user_token* = JWT user access token
- *ds_host* = IP address or host name of the Data Classification Linux system.
- *cm_host* = IP address or host name of the Console agent system.
- *cloud_provider* = When installing on a cloud instance, enter "AWS", "Azure", or "Gcp" depending on cloud provider.
- *proxy_host* = IP or host name of the proxy server if the host is behind a proxy server.
- *proxy_port* = Port to connect to the proxy server (default 80).
- *proxy_scheme* = Connection scheme: https or http (default http).
- *proxy_user* = Authenticated user to connect to the proxy server, if basic authentication is required. The user must be a local user - domain users are not supported.
- *proxy_password* = Password for the user name that you specified.
- *ca_cert_dir* = Path on the Data Classification Linux system containing additional TLS CA certificate bundles. Only required if the proxy is performing TLS interception.

Result

The Data Classification installer installs packages, registers the installation, and installs Data Classification. Installation can take 10 to 20 minutes.

If there is connectivity over port 8080 between the host machine and the Console agent instance, you'll see the installation progress in the Data Classification tab in the Console.

What's Next

From the Configuration page you can select the data sources that you want to scan.

Install NetApp Data Classification on a Linux host with no internet access

Installing NetApp Data Classification on a Linux host in an on-premises site that doesn't have internet access is known as *private mode*. This type of installation, which uses an installation script, has no connectivity to the NetApp Console SaaS layer.

BlueXP private mode (legacy BlueXP interface) is typically used with on-premises environments that have no internet connection and with secure cloud regions, which includes AWS Secret Cloud, AWS Top Secret Cloud, and Azure IL6. NetApp continues to support these environments with the legacy BlueXP interface. For private mode documentation in the legacy BlueXP interface, see PDF documentation for BlueXP private mode.

Check that your Linux host is ready to install NetApp Data Classification

Before installing NetApp Data Classification manually on a Linux host, optionally run a script on the host to verify that all the prerequisites are in place for installing Data Classification. You can run this script on a Linux host in your network, or on a Linux host in the cloud. The host can be connected to the internet, or the host can reside in a site

that doesn’t have internet access (a *dark site*).

The Data Classification installation script encompasses a test script to ensure your environment meets the requirements. You can run this script separately to verify the Linux host’s readiness before run the installation script.

Getting Started

You’ll perform the following tasks.

- Optionally, install a Console agent if you don’t already have one installed. You can run the test script without having a Console agent installed, but the script checks for connectivity between the Console agent and the Data Classification host machine - so it is recommended that you have a Console agent.
- Prepare the host machine and verify that it meets all the requirements.
- Enable outbound internet access from the Data Classification host machine.
- Verify that all required ports are enabled on all systems.
- Download and run the Prerequisite test script.

Create a Console agent

A Console agent is required before you can install and use Data Classification. You can, however, run the Prerequisites script without a Console agent.

You can install the Console agent on-premises on a Linux host in your network or on a Linux host in the cloud. You can also install Data Classification on-premises if the Console agent is installed on-premises.

To create a Console agent in your cloud provider environment, see:

- creating a Console agent in AWS
- creating a Console agent in Azure
- creating a Console agent in GCP

You need the IP address or host name of the Console agent system when running the Prerequisites script. You have this information if you installed the Console agent in your premises. If the Console agent is deployed in the cloud, you can find this information from the Console: select the Help icon then **Support**; in the Agent and Audit section, select **Go to the agent**.

Verify host requirements

Data Classification software must run on a host that meets specific operating system requirements, RAM requirements, and software requirements.

- Data Classification must be on a dedicated host. The host can’t be shared with other applications or third-party software such as antivirus.
- Choose the size that aligns with the data set you plan to scan with Data Classification.

System size	CPU	RAM (swap memory must be disabled)	Disk
Extra Large	32 CPUs	128 GB RAM	• 1 TiB SSD on /, or 100 GiB available on /opt • 895 GiB available on /var/lib/docker • 5 GiB on /tmp • **For Podman, 30 GB on /var/tmp**
Large	16 CPUs	64 GB RAM	• 500 GiB SSD on /, or 100 GiB available on /opt • 400 GiB available on /var/lib/docker or for Podman /var/lib/containers • 5 GiB on /tmp • **For Podman, 30 GB on /var/tmp**

- When deploying a compute instance in the cloud for your Data Classification installation, it's recommended you use a system that meets the "Large" system requirements above:
  - **Amazon Elastic Compute Cloud (Amazon EC2) instance type**: "m6i.4xlarge". See additional AWS instance types.
  - **Azure VM size**: "Standard_D16s_v3". See additional Azure instance types.
  - **GCP machine type**: "n2-standard-16". See additional GCP instance types.
- **UNIX folder permissions**: The following minimum UNIX permissions are required:

Folder	Minimum permissions
/tmp	`rwxrwxrwt`
/opt	`rwxr-xr-x`
/var/lib/docker	`rwx------`
/usr/lib/systemd/system	`rwxr-xr-x`

- **Operating system**:
  - The following operating systems require using the Docker container engine:
    - Red Hat Enterprise Linux version 7.8 and 7.9
    - Ubuntu 22.04 (requires Data Classification version 1.23 or greater)
    - Ubuntu 24.04 (requires Data Classification version 1.23 or greater)
  - The following operating systems require using the Podman container engine, and they require Data Classification version 1.30 or greater:
    - Red Hat Enterprise Linux version 8.8, 8.10, 9.0, 9.1, 9.2, 9.3, 9.4, 9.5, and 9.6.

- Advanced Vector Extensions (AVX2) must be enabled on the host system.

- **Red Hat Subscription Management**: The host must be registered with Red Hat Subscription Management. If it's not registered, the system can't access repositories to update required 3rd-party software during installation.
- **Additional software**: You must install the following software on the host before you install Data Classification:
  - Depending on the OS you are using, you need to install one of the container engines:
    - Docker Engine version 19.3.1 or greater. View installation instructions.
    - Podman version 4 or greater. To install Podman, enter (`sudo yum install podman netavark -y`).
- Python version 3.6 or greater. View installation instructions.
  - **NTP considerations**: NetApp recommends configuring the Data Classification system to use a Network Time Protocol (NTP) service. The time must be synchronized between the Data Classification system and the Console agent system.
- **Firewalld considerations**: If you are planning to use `firewalld`, we recommend that you enable it before installing Data Classification. Run the following commands to configure `firewalld` so that it is compatible with Data Classification:

```
firewall-cmd --permanent --add-service=http
firewall-cmd --permanent --add-service=https
firewall-cmd --permanent --add-port=80/tcp
firewall-cmd --permanent --add-port=8080/tcp
firewall-cmd --permanent --add-port=443/tcp
firewall-cmd --reload
```

If you're planning to use additional Data Classification hosts as scanner nodes (in a distributed model), add these rules to your primary system at this time:

```
firewall-cmd --permanent --add-port=2377/tcp
firewall-cmd --permanent --add-port=7946/udp
firewall-cmd --permanent --add-port=7946/tcp
firewall-cmd --permanent --add-port=4789/udp
```

Note that you must restart Docker or Podman whenever you enable or update `firewalld` settings.

Enable outbound internet access from Data Classification

Data Classification requires outbound internet access. If your virtual or physical network uses a proxy server for internet access, ensure that the Data Classification instance has outbound internet access to contact the following endpoints.

This section is not required for host systems installed in sites without internet connectivity.

Endpoints	Purpose
https://api.console.netapp.com	Communication with the Console service, which includes NetApp accounts.
https://netapp-cloud-account.auth0.com https://auth0.com	Communication with the Console website for centralized user authentication.
https://support.compliance.api.console.netapp.com/ https://hub.docker.com https://auth.docker.io https://registry-1.docker.io https://index.docker.io/ https://dseasb33srnrn.cloudfront.net/ https://production.cloudflare.docker.com/	Provides access to software images, manifests, templates, and to send logs and metrics.
https://support.compliance.api.console.netapp.com/	Enables NetApp to stream data from audit records.
https://github.com/docker https://download.docker.com	Provides prerequisite packages for docker installation.
http://packages.ubuntu.com/ http://archive.ubuntu.com	Provides prerequisite packages for Ubuntu installation.

Verify that all required ports are enabled

You must ensure that all required ports are open for communication between the Console agent, Data Classification, Active Directory, and your data sources.

Connection Type	Ports	Description
Console agent <> Data Classification	8080 (TCP), 443 (TCP), and 80. 9000	The firewall or routing rules for the Console agent must allow inbound and outbound traffic over port 443 to and from the Data Classification instance. Make sure port 8080 is open so you can see the installation progress in the Console. If a firewall is used on the Linux host, port 9000 is required for internal processes within an Ubuntu server.
Console agent <> ONTAP cluster (NAS)	443 (TCP)	The Console discovers ONTAP clusters using HTTPS. If you use custom firewall policies, the Console agent host must allow outbound HTTPS access through port 443. If the Console agent is in the cloud, all outbound communication is allowed by the predefined firewall or routing rules.

Run the Data Classification prerequisites script

Follow these steps to run the Data Classification prerequisites script.

Watch this video to see how to run the Prerequisites script and interpret the results.

Before you begin

- Verify that your Linux system meets the host requirements.
- Verify that the system has the two prerequisite software packages installed (Docker Engine or Podman, and Python 3).
- Make sure you have root privileges on the Linux system.

Steps

1. Download the Data Classification Prerequisites script from the NetApp Support Site. The file you should select is named **standalone-pre-requisite-tester-<version>**.
2. Copy the file to the Linux host you plan to use (using `scp` or some other method).
3. Assign permissions to run the script.

   ```
   chmod +x standalone-pre-requisite-tester-v1.25.0
   ```

4. Run the script using the following command.

   ```
   ./standalone-pre-requisite-tester-v1.25.0 <--darksite>
   ```

   Add the option "--darksite" only if you are running the script on a host that doesn't have internet access. Certain prerequisite tests are skipped when the host is not connected to the internet.

5. The script prompts you for the IP address of the Data Classification host machine.
   - Enter the IP address or host name.
6. The script prompts whether you have an installed Console agent.
   - Enter **N** if you do not have an installed Console agent.
   - Enter **Y** if you do have an installed Console agent. And then enter the IP address or host name of the Console agent so the test script can test this connectivity.
7. The script runs a variety of tests on the system and it displays results as it progresses. When it finishes it writes a log of the session to a file named `prerequisites-test-<timestamp>.log` in the directory `/opt/netapp/install_logs`.

Result

If all the prerequisites tests ran successfully, you can install Data Classification on the host when you are ready.

If any issues were discovered, they are categorized as "Recommended" or "Required" to be fixed. Recommended issues are typically items that would make the Data Classification scanning and categorizing tasks run slower. These items do not need to be corrected - but you may want to address them.

If you have any "Required" issues, you should fix the issues and run the Prerequisites test script again.

Activate scanning on your data sources

Scan data sources with NetApp Data Classification

NetApp Data Classification scanns the data in the repositories (the volumes, database schemas, or other user data) that you select to identify personal and sensitive data. Data Classification then maps your organizational data, categorizes each file, and identifies predefined patterns in the data. The result of the scan is an index of personal information, sensitive personal information, data categories, and file types.

After the initial scan, Data Classification continuously scans your data in a round-robin fashion to detect incremental changes. This is why it's important to keep the instance running.

You can enable and disable scans at the volume level or at the database schema level.

What's the difference between Mapping and Classification scans

You can conduct two types of scans in Data Classification:

- **Mapping-only scans** provide only a high-level overview of your data and are performed on selected data sources. Mapping-only scans take less time than map and classify scans because the do not access files to see the data inside. You might want to do this initially to identify areas of research and then perform a Map & Classify scan on those areas.
- **Map & Classify scans** provide deep-level scanning of your data.

The table below shows some of the differences:

Feature	Map & classify scans	Mapping-only scans
Scan speed	Slow	Fast
Pricing	Free	Free
Capacity	Limited to 500 TiB*	Limited to 500 TiB*
List of file types and used capacity	Yes	Yes
Number of files and used capacity	Yes	Yes
Age and size of files	Yes	Yes
Ability to run a Data Mapping Report	Yes	Yes
Data Investigation page to view file details	Yes	No
Search for names within files	Yes	No
Create saved queries that provide custom search results	Yes	No
Ability to run other reports	Yes	No
Ability to see metadata from files**	No	Yes

* Data Classification does not impose a limit on the amount of data it can scan. Each Console agent supports scanning and displaying 500 TiB of data. To scan more than 500 TiB of data, install another Console agent

then deploy another Data Classification instance.
The Console UI displays data from a single connector. For tips on viewing data from multiple Console agents, see Work with multiple Console agents.

** The following metadata is extracted from files during mapping scans:

- System
- System type
- Storage repository
- File type
- Used capacity
- Number of files
- File size
- File creation
- File last access
- File last modified
- File discovered time
- Permissions extraction

Governance dashboard differences:

Feature	Map & Classify	Map
Stale data	Yes	Yes
Non-business data	Yes	Yes
Duplicated files	Yes	Yes
Predefined saved queries	Yes	No
Default saved queries	Yes	Yes
DDA report	Yes	Yes
Mapping report	Yes	Yes
Sensitivity level detection	Yes	No
Sensitive data with wide permissions	Yes	No
Open permissions	Yes	Yes
Age of data	Yes	Yes
Size of data	Yes	Yes
Categories	Yes	No
File types	Yes	Yes

Compliance dashboard differences:

Feature	Map & Classify	Map
Personal information	Yes	No
Sensitive personal information	Yes	No
Privacy risk assessment report	Yes	No
HIPAA report	Yes	No
PCI DSS report	Yes	No

Investigation filters differences:

Feature	Map & Classify	Map
Saved queries	Yes	Yes
System type	Yes	Yes
System	Yes	Yes
Storage repository	Yes	Yes
File type	Yes	Yes
File size	Yes	Yes
Created time	Yes	Yes
Discovered time	Yes	Yes
Last modified	Yes	Yes
Last access	Yes	Yes
Open permissions	Yes	Yes
File directory path	Yes	Yes
Category	Yes	No
Sensitivity level	Yes	No
Number of identifiers	Yes	No
Personal data	Yes	No
Sensitive personal data	Yes	No
Data subject	Yes	No
Duplicates	Yes	Yes
Classification status	Yes	Status is always "Limited insights"
Scan analysis event	Yes	Yes
File hash	Yes	Yes
Number of users with access	Yes	Yes
User/group permissions	Yes	Yes
File owner	Yes	Yes
Directory type	Yes	Yes

Scan Amazon FSx for ONTAP volumes with NetApp Data Classification

Complete a few steps to scan Amazon FSx for ONTAP volumes with NetApp Data Classification.

Before you begin

- You need an active Console agent in AWS to deploy and manage Data Classification.
- The security group you selected when creating the system must allow traffic from the Data Classification instance. You can find the associated security group using the ENI connected to the FSx for ONTAP file system and edit it using the AWS Management Console.

  AWS security groups for Linux instances

  AWS security groups for Windows instances

  AWS elastic network interfaces (ENI)

- Ensure the following ports are open to the Data Classification instance:
  - For NFS – ports 111 and 2049.
  - For CIFS – ports 139 and 445.

Deploy the Data Classification instance

Deploy Data Classification if there isn't already an instance deployed.

You should deploy Data Classification in the same AWS network as the Console agent for AWS and the FSx volumes you wish to scan.

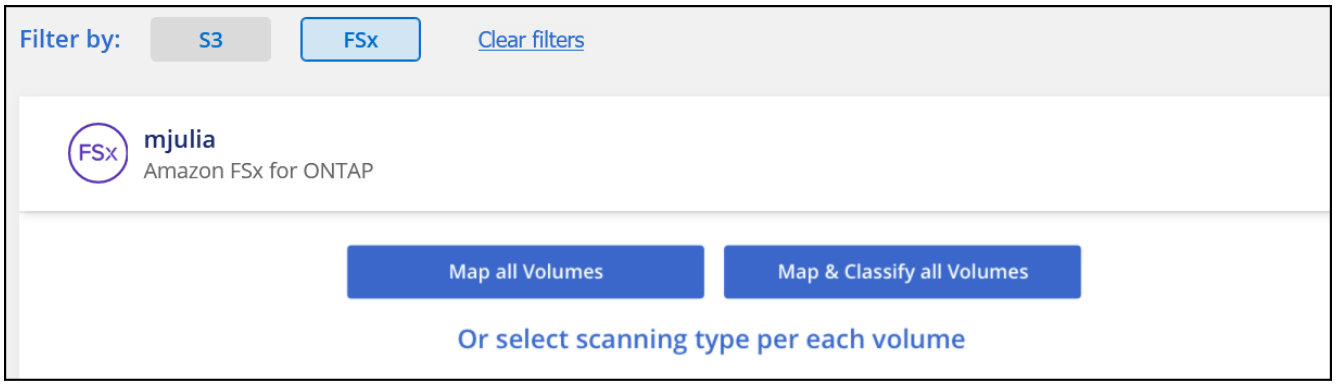
Note: Deploying Data Classification in an on-premises location is not currently supported when scanning FSx volumes.

Upgrades to Data Classification software is automated as long as the instance has internet connectivity.

Enable Data Classification in your systems

You can enable Data Classification for FSx for ONTAP volumes.

1. From NetApp Console, **Governance > Classification**.
2. From the Data Classification menu, select **Configuration**.



3. Select how you want to scan the volumes in each system. Learn about mapping and classification scans:
   - To map all volumes, select **Map all Volumes**.
   - To map and classify all volumes, select **Map & Classify all Volumes**.
   - To customize scanning for each volume, select **Or select scanning type for each volume**, and then

choose the volumes you want to map and/or classify.

4. In the confirmation dialog box, select **Approve** to have Data Classification start scanning your volumes.

Result

Data Classification starts scanning the volumes you selected in the system. Results will be available in the Compliance dashboard as soon as Data Classification finishes the initial scans. The time that it takes depends on the amount of data—it could be a few minutes or hours. You can track the progress of the initial scan by navigating to the **Configuration** menu then selecting the **System configuration**. Track the progress of each scan in the progress bar; you can hover over the progress bar to see the number of files scanned relative to the total files in the volume.

- By default, if Data Classification doesn't have write attributes permissions in CIFS, or write permissions in NFS, the system won't scan the files in your volumes because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, select **Or select scanning type for each volume**. The resulting page has a setting you can enable so that Data Classification will scan the volumes regardless of permissions.
- Data Classification scans only one file share under a volume. If you have multiple shares in your volumes, you'll need to scan those other shares separately as a shares group. See more details about this Data Classification limitation.

Verify that Data Classification has access to volumes

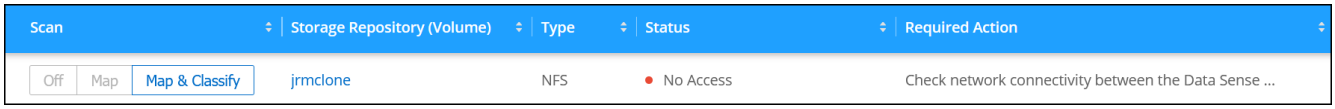
Make sure Data Classification can access volumes by checking your networking, security groups, and export policies.

You'll need to provide Data Classification with CIFS credentials so it can access CIFS volumes.

Steps

1. From the Data Classification menu, select **Configuration**.
2. On the Configuration page, select **View Details** to review the status and correct any errors.

   For example, the following image shows a volume Data Classification can't scan due to network connectivity issues between the Data Classification instance and the volume.



3. Make sure there's a network connection between the Data Classification instance and each network that includes volumes for FSx for ONTAP.

   For FSx for ONTAP, Data Classification can scan volumes only in the same region as the Console.

4. Ensure NFS volume export policies include the IP address of the Data Classification instance so it can access the data on each volume.
5. If you use CIFS, provide Data Classification with Active Directory credentials so it can scan CIFS volumes.
   a. From the Data Classification menu, select **Configuration**.
   b. For each system, select **Edit CIFS Credentials** and enter the user name and password that Data

Classification needs to access CIFS volumes on the system.

The credentials can be read-only, but providing admin credentials ensures that Data Classification can read any data that requires elevated permissions. The credentials are stored on the Data Classification instance.

If you want to make sure your files "last accessed times" are unchanged by Data Classification scans, it's recommended the user has Write Attributes permissions in CIFS or write permissions in NFS. If possible, configure the Active Directory user as part of a parent group in the organization which has permissions to all files.

After you enter the credentials, you should see a message that all CIFS volumes were authenticated successfully.

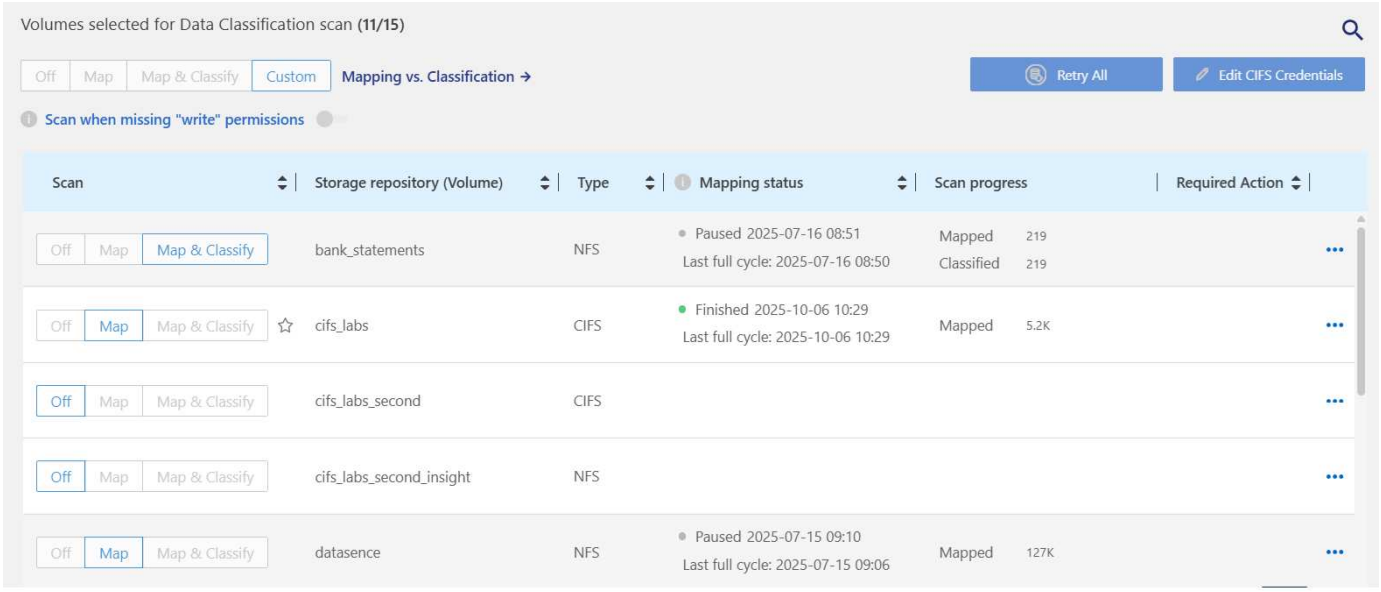
Enable and disable scans on volumes

You can start or stop scans on any system at any time from the Configuration page. You can also switch scans from mapping-only scans to mapping and classification scans, and vice-versa. It's recommended that you scan all volumes in a system.

New volumes added to the system are automatically scanned only when you have selected the **Map** or **Map & Classify** setting in the heading area. When set to **Custom** or **Off** in the heading area, you'll need to activate mapping and/or full scanning on each new volume you add in the system.

The switch at the top of the page for **Scan when missing "write" permissions** is disabled by default. This means that if Data Classification doesn't have write attributes permissions in CIFS or write permissions in NFS, the system won't scan the files because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, turn the switch ON and all files are scanned regardless of the permissions. Learn more.

New volumes added to the system are automatically scanned only when you have set the **Map** or **Map & Classify** setting in the heading area. When the setting for all volumes is **Custom** or **Off**, you need to activate scanning manually for each new volume you add.



Steps

1. From the Data Classification menu, select **Configuration**.
2. Choose a system, then select **Configuration**.
3. To enable or disable scans for all volumes, select **Map**, **Map & Classify**, or **Off** in the heading above all volumes.

   To enable or disable scans for individual volumes, find the volumes in the list then select **Map**, **Map & Classify**, or **Off** next to the volume name.

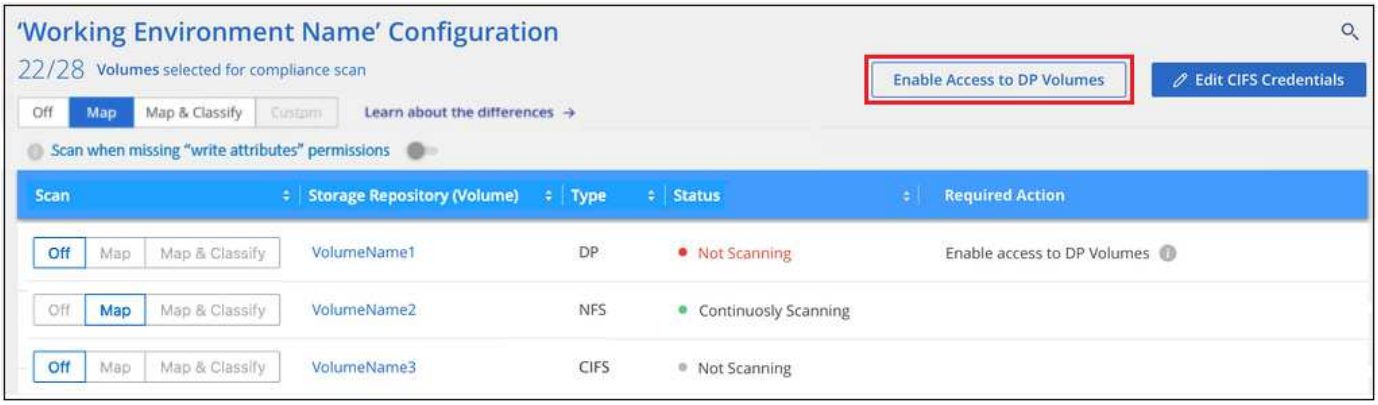
Result

When you enable scanning, Data Classification starts scanning the volumes you selected in the system. Results start to appear in the Compliance dashboard as soon as Data Classification starts the scan. Scan completion time depends on the amount of data, ranging from minutes to hours.

Scan data protection volumes

By default, data protection (DP) volumes are not scanned because they are not exposed externally and Data Classification cannot access them. These are the destination volumes for SnapMirror operations from an FSx for ONTAP file system.

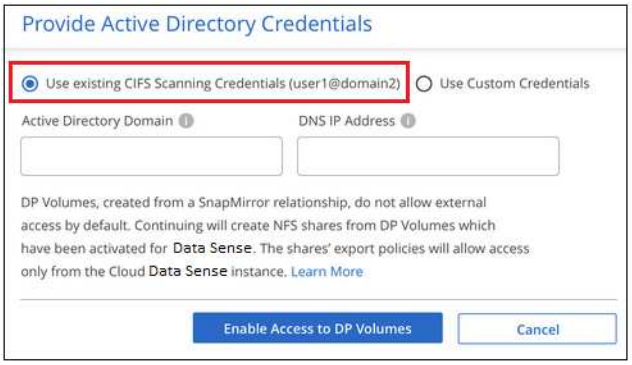
Initially, the volume list identifies these volumes as *Type* **DP** with the *Status* **Not Scanning** and the *Required Action* **Enable Access to DP volumes**.

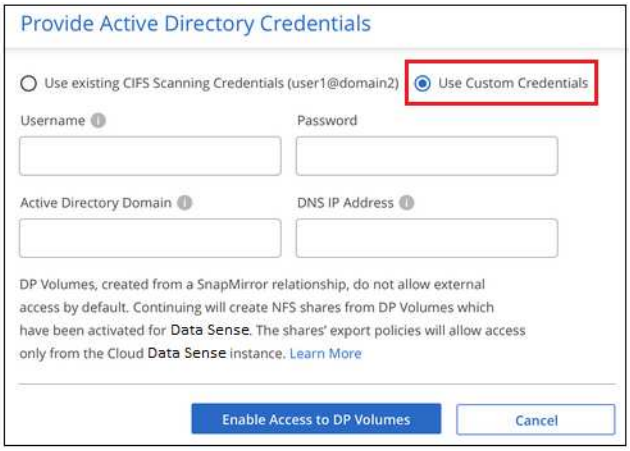


Steps

If you want to scan these data protection volumes:

1. From the Data Classification menu, select **Configuration**.
2. Select **Enable Access to DP volumes** at the top of the page.
3. Review the confirmation message and select **Enable Access to DP volumes** again.
   - Volumes that were initially created as NFS volumes in the source FSx for ONTAP file system are enabled.
   - Volumes that were initially created as CIFS volumes in the source FSx for ONTAP file system require that you enter CIFS credentials to scan those DP volumes. If you already entered Active Directory credentials so that Data Classification can scan CIFS volumes you can use those credentials, or you can specify a different set of Admin credentials.





4. Activate each DP volume that you want to scan.

Result

Once enabled, Data Classification creates an NFS share from each DP volume that was activated for scanning. The share export policies only allow access from the Data Classification instance.

If you had no CIFS data protection volumes when you initially enabled access to DP volumes, and later add some, the button **Enable Access to CIFS DP** appears at the top of the Configuration page. Select this button and add CIFS credentials to enable access to these CIFS DP volumes.

Active Directory credentials are registered only in the storage VM of the first CIFS DP volume, so all DP volumes on that SVM will be scanned. Any volumes that reside on other SVMs will not have the Active Directory credentials registered, so those DP volumes won't be scanned.

Scan Azure NetApp Files volumes with NetApp Data Classification

Complete a few steps to get started with NetApp Data Classification for Azure NetApp Files.

Discover the Azure NetApp Files system that you want to scan

If the Azure NetApp Files system you want to scan is not already in the NetApp Console as a system, add it in the Systems page.

Deploy the Data Classification instance

Deploy Data Classification if there isn't already an instance deployed.

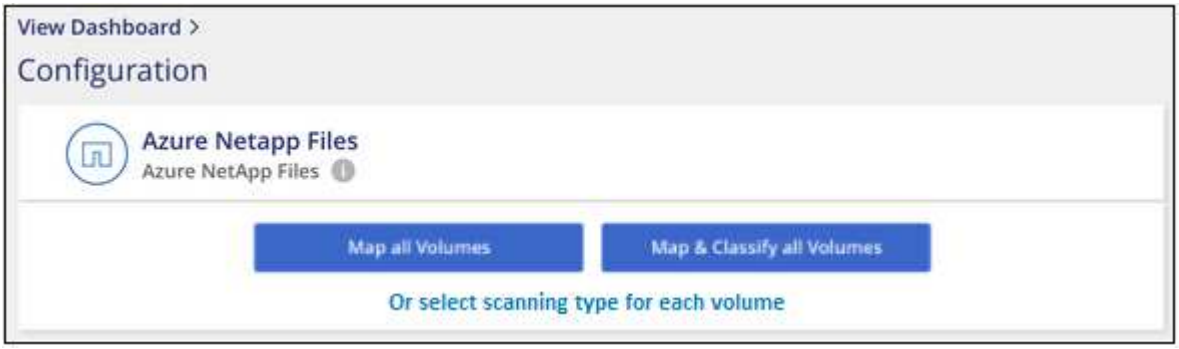
Data Classification must be deployed in the cloud when scanning Azure NetApp Files volumes, and it must be deployed in the same region as the volumes you wish to scan.

Note: Deploying Data Classification in an on-premises location is not currently supported when scanning Azure NetApp Files volumes.

Enable Data Classification in your systems

You can enable Data Classification on your Azure NetApp Files volumes.

1. From the Data Classification menu, select **Configuration**.



2. Select how you want to scan the volumes in each system. Learn about mapping and classification scans:
   - To map all volumes, select **Map all Volumes**.
   - To map and classify all volumes, select **Map & Classify all Volumes**.
   - To customize scanning for each volume, select **Or select scanning type for each volume**, and then choose the volumes you want to map or map and classify.

     See Enable or disable scans on volumes for details.
3. In the confirmation dialog box, select **Approve**.

Result

Data Classification starts scanning the volumes you selected in the system. Results are available in the Compliance dashboard as soon as Data Classification finishes the initial scans. The time that it takes depends on the amount of data—it could be a few minutes or hours. You can track the progress of the initial scan by navigating to the **Configuration** menu then selecting the **System configuration**. Data Classification displays a progress bar for each scan. You can hover over the progress bar to see the number of files scanned relative to the total number of files in the volume.

- By default, if Data Classification doesn't have write attributes permissions in CIFS, or write permissions in NFS, the system won't scan the files in your volumes because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, select **Or select scanning type for each volume**. The resulting page has a setting you can enable so that Data Classification will scan the volumes regardless of permissions.
- Data Classification scans only one file share under a volume. If you have multiple shares in your volumes, you'll need to scan those other shares separately as a shares group. Learn about this Data Classification limitation.

Verify that Data Classification has access to volumes

Make sure that Data Classification can access volumes by checking your networking, security groups, and export policies. You need to provide Data Classification with CIFS credentials so it can access CIFS volumes.

For Azure NetApp Files, Data Classification can only scan volumes in the same region as the Console.

Checklist

- Make sure that there's a network connection between the Data Classification instance and each network that includes volumes for Azure NetApp Files.
- Ensure the following ports are open to the Data Classification instance:

- For NFS – ports 111 and 2049.
- For CIFS – ports 139 and 445.

- Ensure the NFS volume export policies include the IP address of the Data Classification instance so it can access the data on each volume.

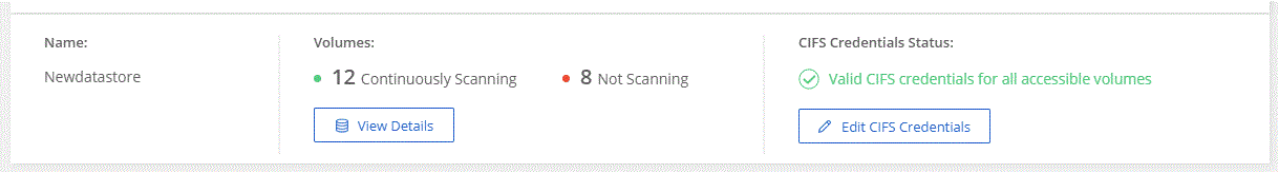
Steps

1. From the Data Classification menu, select **Configuration**.
   a. If you're using CIFS (SMB), ensure the Active Directory credentials are correct. For each system, select **Edit CIFS Credentials** then enter the user name and password that Data Classification needs to access CIFS volumes on the system.

      The credentials can be read-only; providing admin credentials ensures that Data Classification can read any data that requires elevated permissions. The credentials are stored on the Data Classification instance.

      If you want to make sure your files "last accessed times" are unchanged by Data Classification scans, it's recommended the user has Write Attributes permissions in CIFS or write permissions in NFS. If possible, configure the Active Directory user as part of a parent group in the organization which has permissions to all files.

      After you enter the credentials, you should see a message that all CIFS volumes were authenticated successfully.



2. On the Configuration page, select **View Details** to review the status for each CIFS and NFS volume. If necessary, correct any errors such as network connectivity issues.

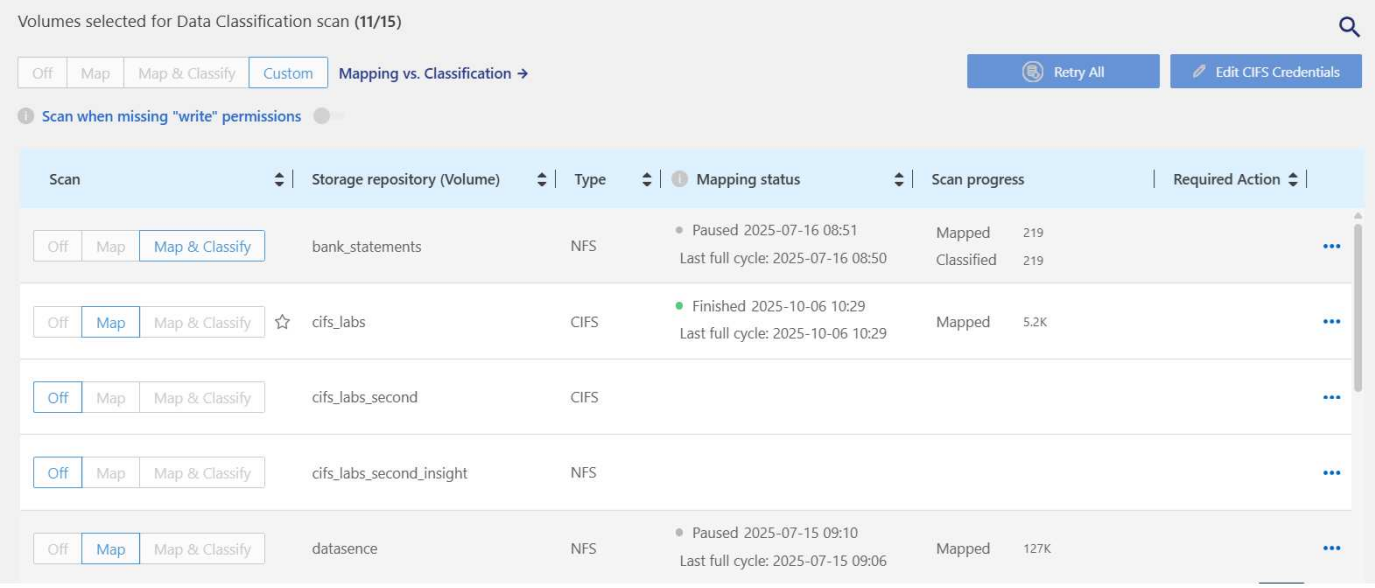
Enable or disable scans on volumes

You can start or stop scans on any system at any time from the Configuration page. You can also switch scans from mapping-only scans to mapping and classification scans, and vice-versa. It's recommended that you scan all volumes in a system.

New volumes added to the system are automatically scanned only when you have selected the **Map** or **Map & Classify** setting in the heading area. When set to **Custom** or **Off** in the heading area, you'll need to activate mapping and/or full scanning on each new volume you add in the system.

The switch at the top of the page for **Scan when missing "write" permissions** is disabled by default. This means that if Data Classification doesn't have write attributes permissions in CIFS or write permissions in NFS, the system won't scan the files because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, turn the switch ON and all files are scanned regardless of the permissions. Learn more.

New volumes added to the system are automatically scanned only when you have set the **Map** or **Map & Classify** setting in the heading area. When the setting for all volumes is **Custom** or **Off**, you need to activate scanning manually for each new volume you add.



Steps

1. From the Data Classification menu, select **Configuration**.
2. Choose a system, then select **Configuration**.
3. To enable or disable scans for all volumes, select **Map**, **Map & Classify**, or **Off** in the heading above all volumes.

   To enable or disable scans for individual volumes, find the volumes in the list then select **Map**, **Map & Classify**, or **Off** next to the volume name.

Result

When you enable scanning, Data Classification starts scanning the volumes you selected in the system. Results start to appear in the Compliance dashboard as soon as Data Classification starts the scan. Scan completion time depends on the amount of data, ranging from minutes to hours.

Scan Cloud Volumes ONTAP and on-premises ONTAP volumes with NetApp Data Classification

Complete a few steps to start scanning your Cloud Volumes ONTAP and on-premises ONTAP volumes using NetApp Data Classification.

Prerequisites

Before you enable Data Classification, make sure you have a supported configuration.

- If you are scanning Cloud Volumes ONTAP and on-premises ONTAP systems that are accessible over the internet, you can deploy Data Classification in the cloud or in an on-premises location that has internet access.
- If you are scanning on-premises ONTAP systems that have been installed in a dark site that has no internet access, you need to deploy Data Classification in the same on-premises location that has no internet access. This requires the Console agent to be deployed in that same on-premises location.

Verify that Data Classification has access to volumes

Make sure that Data Classification can access volumes by checking your networking, security groups, and export policies. You'll need to provide Data Classification with CIFS credentials so it can access CIFS volumes.

Checklist

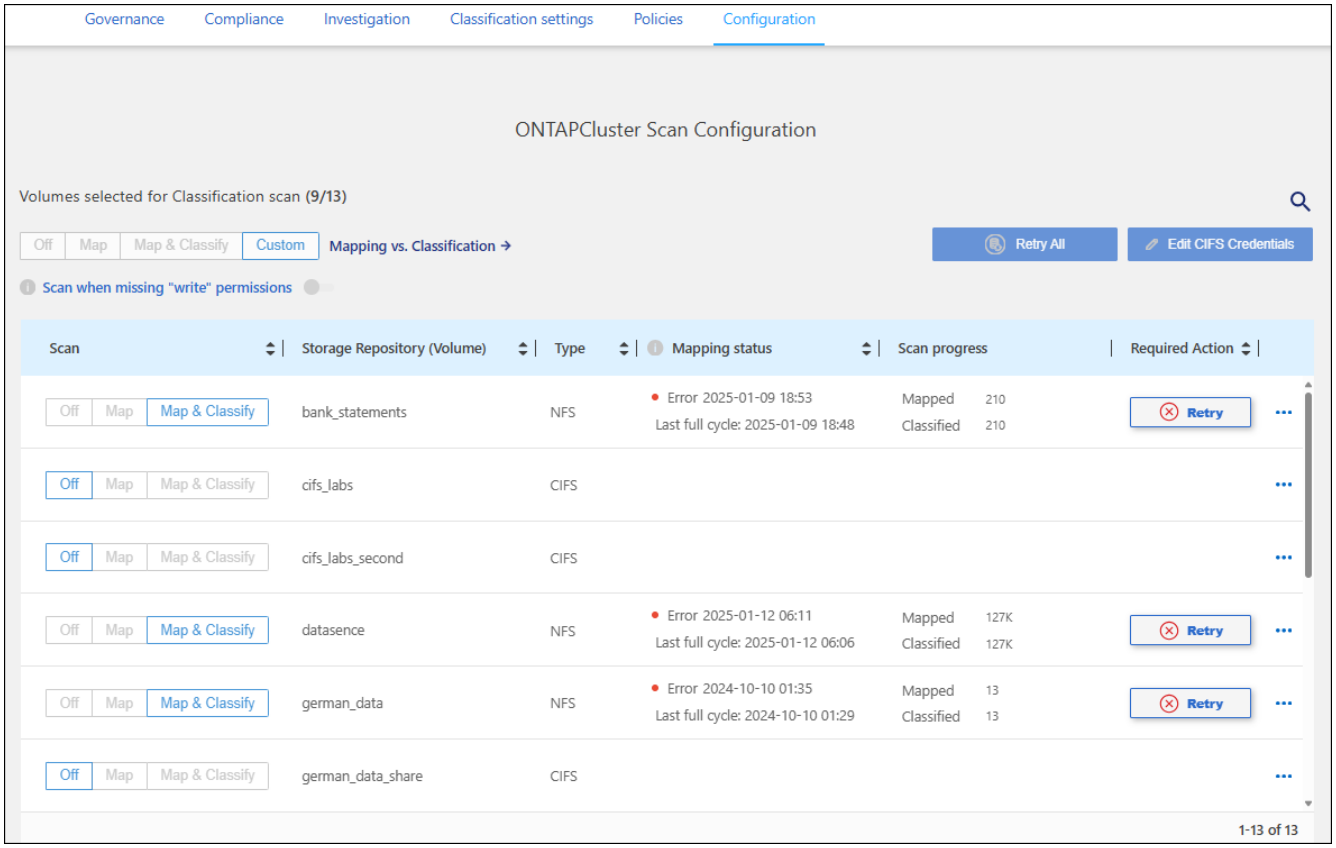
- Make sure that there's a network connection between the Data Classification instance and each network that includes volumes for Cloud Volumes ONTAP or on-prem ONTAP clusters.
- Ensure that the security group for Cloud Volumes ONTAP allows inbound traffic from the Data Classification instance.

  You can either open the security group for traffic from the IP address of the Data Classification instance, or you can open the security group for all traffic from inside the virtual network.

- Ensure that NFS volume export policies include the IP address of the Data Classification instance so it can access the data on each volume.

Steps

1. From the Data Classification menu, select **Configuration**.



2. If you use CIFS, provide Data Classification with Active Directory credentials so it can scan CIFS volumes. For each system, select **Edit CIFS Credentials** and enter the user name and password that Data Classification needs to access CIFS volumes on the system.

   The credentials can be read-only, but providing admin credentials ensures that Data Classification can read any data that requires elevated permissions. The credentials are stored on the Data Classification instance.

   If you want to make sure your files "last accessed times" are unchanged by Data Classification scans, it's recommended the user has Write Attributes permissions in CIFS or write permissions in NFS. If possible,

configure the Active Directory user as part of a parent group in the organization which has permissions to all files.

If you've entered the credentials correctly, a message confirms all CIFS volumes were authenticated successfully.

3. On the Configuration page, select **Configuration** to review the status for each CIFS and NFS volume and correct any errors.

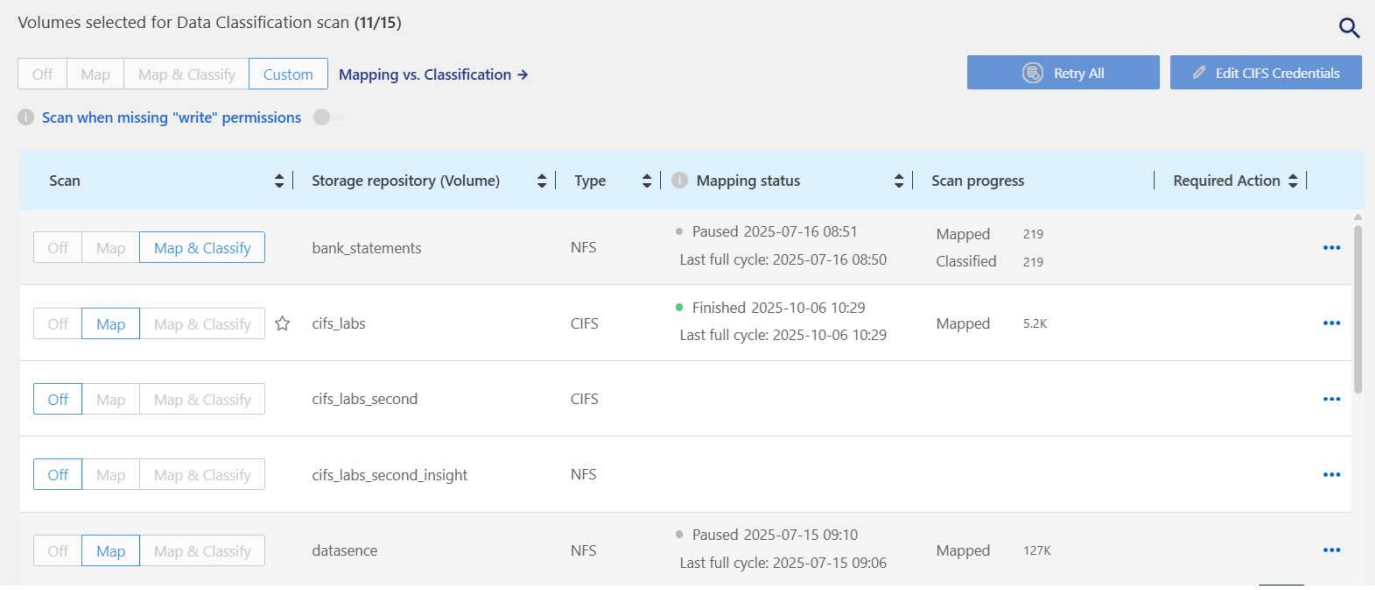
Enable or disable scans on volumes

You can start or stop scans on any system at any time from the Configuration page. You can also switch scans from mapping-only scans to mapping and classification scans, and vice-versa. It's recommended that you scan all volumes in a system.

New volumes added to the system are automatically scanned only when you have selected the **Map** or **Map & Classify** setting in the heading area. When set to **Custom** or **Off** in the heading area, you'll need to activate mapping and/or full scanning on each new volume you add in the system.

The switch at the top of the page for **Scan when missing "write" permissions** is disabled by default. This means that if Data Classification doesn't have write attributes permissions in CIFS or write permissions in NFS, the system won't scan the files because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, turn the switch ON and all files are scanned regardless of the permissions. Learn more.

New volumes added to the system are automatically scanned only when you have set the **Map** or **Map & Classify** setting in the heading area. When the setting for all volumes is **Custom** or **Off**, you need to activate scanning manually for each new volume you add.



Steps

1. From the Data Classification menu, select **Configuration**.
2. Choose a system, then select **Configuration**.
3. To enable or disable scans for all volumes, select **Map**, **Map & Classify**, or **Off** in the heading above all

volumes.

To enable or disable scans for individual volumes, find the volumes in the list then select **Map**, **Map & Classify**, or **Off** next to the volume name.

Result

When you enable scanning, Data Classification starts scanning the volumes you selected in the system. Results start to appear in the Compliance dashboard as soon as Data Classification starts the scan. Scan completion time depends on the amount of data, ranging from minutes to hours.

Data Classification scans only one file share under a volume. If you have multiple shares in your volumes, you'll need to scan those other shares separately as a shares group. See more details about this Data Classification limitation.

Scan database schemas with NetApp Data Classification

Complete a few steps to start scanning your database schemas with NetApp Data Classification.

Review prerequisites

Review the following prerequisites to make sure that you have a supported configuration before you enable Data Classification.

Supported databases

Data Classification can scan schemas from the following databases:

- Amazon Relational Database Service (Amazon RDS)
- MongoDB
- MySQL
- Oracle
- PostgreSQL
- SAP HANA
- SQL Server (MSSQL)

The statistics gathering feature **must be enabled** in the database.

Database requirements

Any database with connectivity to the Data Classification instance can be scanned, regardless of where it is hosted. You just need the following information to connect to the database:

- IP address or host name
- Port
- Service name (only for accessing Oracle databases)
- Credentials that allow read access to the schemas

When choosing a user name and password, it's important to choose one that has full read permissions to all the schemas and tables you want to scan. We recommend that you create a dedicated user for the Data Classification system with all the required permissions.

For MongoDB, a read-only admin role is required.

Deploy the Data Classification instance

Deploy Data Classification if there isn't already an instance deployed.

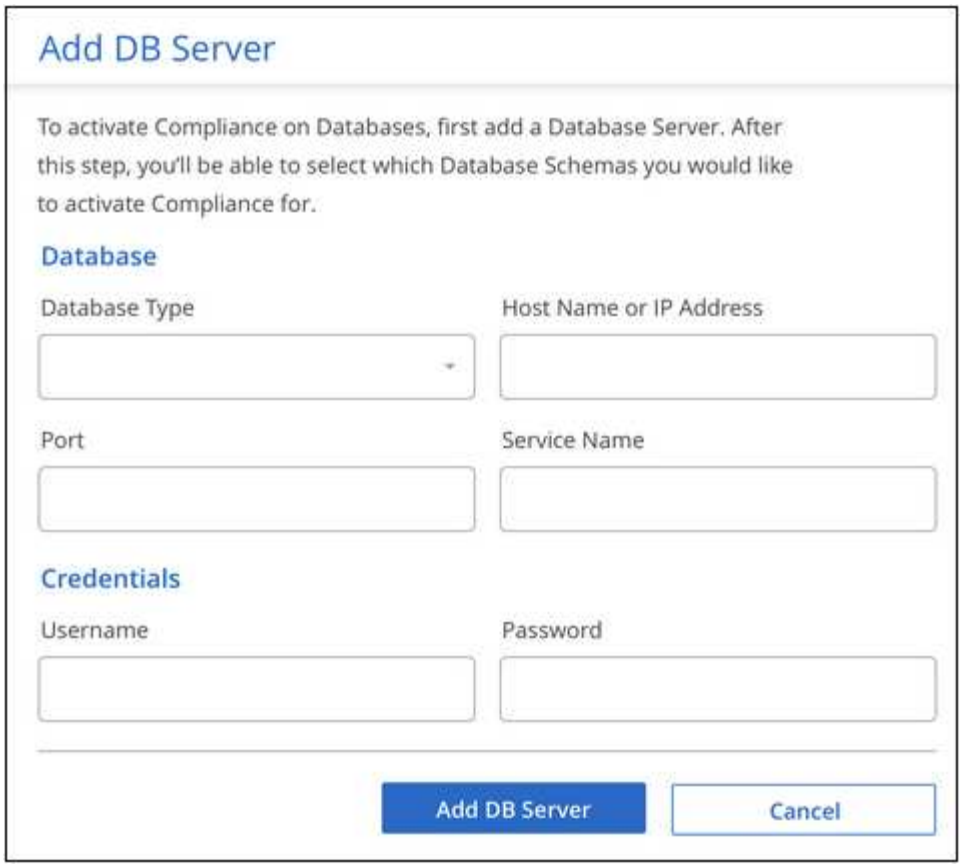
If you are scanning database schemas that are accessible over the internet, you can deploy Data Classification in the cloud or deploy Data Classification in an on-premises location that has internet access.

If you are scanning database schemas that have been installed in a dark site that has no internet access, you need to deploy Data Classification in the same on-premises location that has no internet access. This also requires that the Console agent is deployed in that same on-premises location.

Add the database server

Add the database server where the schemas reside.

1. From the Data Classification menu, select **Configuration**.
2. From the Configuration page, select **Add System** > **Add Database Server**.
3. Enter the required information to identify the database server.
    a. Select the database type.
    b. Enter the port and the host name or IP address to connect to the database.
    c. For Oracle databases, enter the Service name.
    d. Enter the credentials so that Data Classification can access the server.
    e. Select **Add DB Server**.



The database is added to the list of systems.

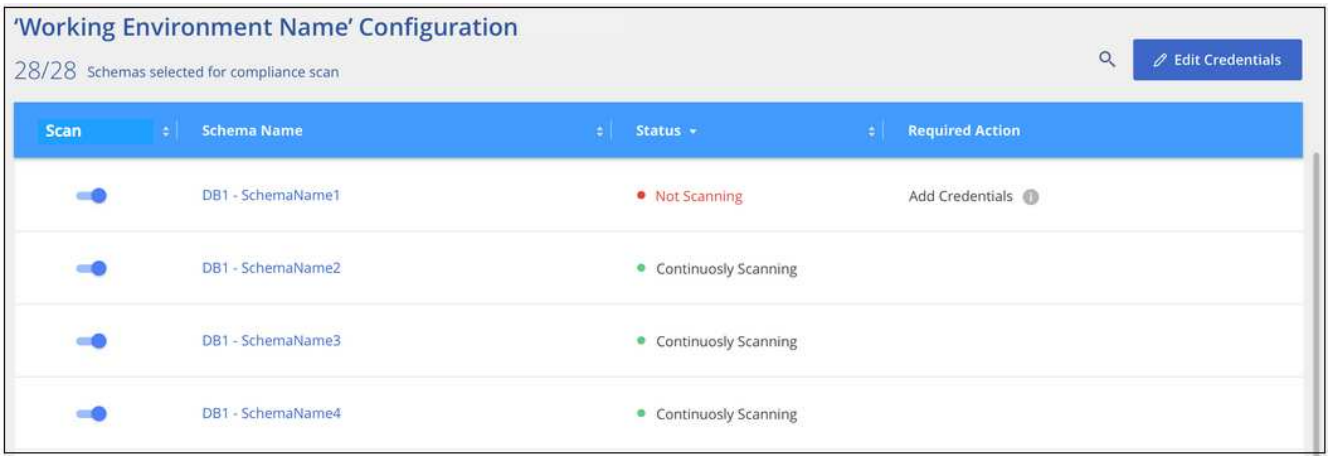
Enable and disable scans on database schemas

You can stop or start full scanning of your schemas at any time.

There is no option to select mapping-only scans for database schemas.

1. From the Configuration page, select the **Configuration** button for the database you want to configure.

2. Select the schemas that you want to scan by moving the slider to the right.



Result

Data Classification starts scanning the database schemas that you enabled. You can track the progress of the initial scan by navigating to the **Configuration** menu then selecting the **System configuration**. The progress of each scan is shown as a progress bar. You can also hover over the progress bar to see the number of files scanned relative to the total number of files in the volume. If there are any errors, they'll appear in the Status column, alongside the required actions to fix the error.

Data Classification scans your databases once per day; databases are not continuously scanned like other data sources.

Scan Google Cloud NetApp Volumes with NetApp Data Classification

NetApp Data Classification supports Google Cloud NetApp Volumes as a system. Learn how to scan your Google Cloud NetApp Volumes system.

Discover the Google Cloud NetApp Volumes system that you want to scan

If the Google Cloud NetApp Volumes system you want to scan is not already in the NetApp Console as a system, add it to the Systems page.

Deploy the Data Classification instance

Deploy Data Classification if there isn't already an instance deployed.

Data Classification must be deployed in the cloud when scanning Google Cloud NetApp Volumes, and it must be deployed in the same region as the volumes you wish to scan.

Note: Deploying Data Classification in an on-premises location is not currently supported when scanning Google Cloud NetApp Volumes.

Enable Data Classification in your systems

You can enable Data Classification on your Google Cloud NetApp Volumes system.

1. From the Data Classification menu, select **Configuration**.
2. Select how you want to scan the volumes in each system. Learn about mapping and classification scans:
   - To map all volumes, select **Map all Volumes**.

- To map and classify all volumes, select **Map & Classify all Volumes**.
- To customize scanning for each volume, select **Or select scanning type for each volume**, and then choose the volumes you want to map and/or classify.

  See Enable and disable scans on volumes for details.

3. In the confirmation dialog box, select **Approve**.

Result

Data Classification starts scanning the volumes you selected in the system. Results are available in the Compliance dashboard as soon as Data Classification finishes the initial scans. The time that it takes depends on the amount of data: a few minutes to a few hours. You can track the progress of the initial scan in the **Configuration** menu's **System configuration** section. Data Classification displays a progress bar for each scan. You can also hover over the progress bar to see the number of files scanned relative to the total files in the volume.

- By default, if Data Classification doesn't have write attributes permissions in CIFS, or write permissions in NFS, the system won't scan the files in your volumes because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, select **Or select scanning type for each volume**. The resulting page has a setting you can enable so that Data Classification will scan the volumes regardless of permissions.
- Data Classification scans only one file share under a volume. If you have multiple shares in your volumes, you need to scan those other shares separately as a shares group. Learn about this Data Classification limitation.

Verify that Data Classification has access to volumes

Ensure Data Classification can access volumes by checking your networking, security groups, and export policies. For CIFS volumes, you need to provide Data Classification with CIFS credentials.

For Google Cloud NetApp Volumes, Data Classification can only scan volumes in the same region as the Console.

Checklist

- Make sure that there's a network connection between the Data Classification instance and each network that includes volumes for Google Cloud NetApp Volumes.
- Ensure the following ports are open to the Data Classification instance:
  - For NFS – ports 111 and 2049.
  - For CIFS – ports 139 and 445.
- Ensure the NFS volume export policies include the IP address of the Data Classification instance so it can access the data on each volume.

Steps

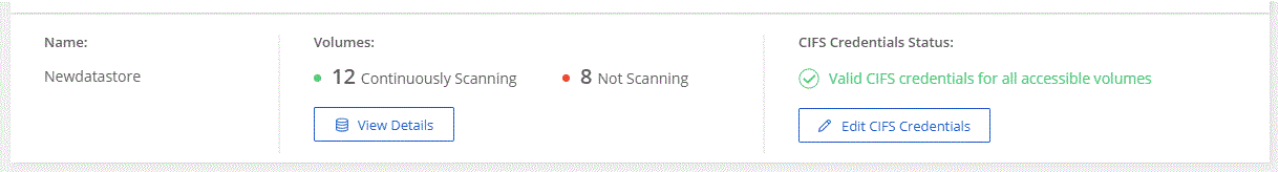
1. From the Data Classification menu, select **Configuration**.
   a. If you're using CIFS (SMB), ensure the Active Directory credentials are correct. For each system, select **Edit CIFS Credentials** then enter the user name and password that Data Classification needs to access CIFS volumes on the system.

      The credentials can be read-only, but providing admin credentials ensures that Data Classification can read any data that requires elevated permissions. The credentials are stored on the Data Classification

instance.

If you want to make sure your files "last accessed times" are unchanged by Data Classification scans, it's recommended the user has Write Attributes permissions in CIFS or write permissions in NFS. If possible, configure the Active Directory user as part of a parent group in the organization which has permissions to all files.

After you enter the credentials, you should see a message that all CIFS volumes were authenticated successfully.



2. On the Configuration page, select **View Details** to review the status for each CIFS and NFS volume and correct any errors.

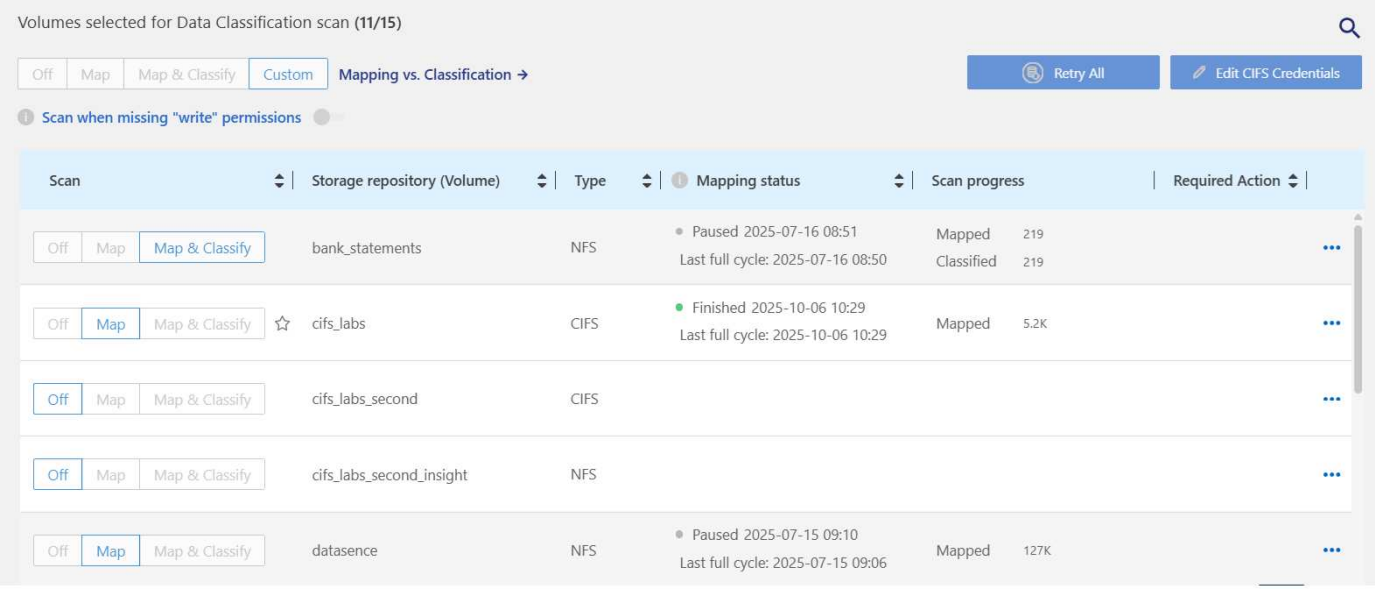
Enable and disable scans on volumes

You can start or stop scans on any system at any time from the Configuration page. You can also switch scans from mapping-only scans to mapping and classification scans, and vice-versa. It's recommended that you scan all volumes in a system.

New volumes added to the system are automatically scanned only when you have selected the **Map** or **Map & Classify** setting in the heading area. When set to **Custom** or **Off** in the heading area, you'll need to activate mapping and/or full scanning on each new volume you add in the system.

The switch at the top of the page for **Scan when missing "write" permissions** is disabled by default. This means that if Data Classification doesn't have write attributes permissions in CIFS or write permissions in NFS, the system won't scan the files because Data Classification can't revert the "last access time" to the original timestamp. If you don't care if the last access time is reset, turn the switch ON and all files are scanned regardless of the permissions. Learn more.

New volumes added to the system are automatically scanned only when you have set the **Map** or **Map & Classify** setting in the heading area. When the setting for all volumes is **Custom** or **Off**, you need to activate scanning manually for each new volume you add.



Steps

1. From the Data Classification menu, select **Configuration**.
2. Choose a system, then select **Configuration**.
3. To enable or disable scans for all volumes, select **Map**, **Map & Classify**, or **Off** in the heading above all volumes.

   To enable or disable scans for individual volumes, find the volumes in the list then select **Map**, **Map & Classify**, or **Off** next to the volume name.

Result

When you enable scanning, Data Classification starts scanning the volumes you selected in the system. Results start to appear in the Compliance dashboard as soon as Data Classification starts the scan. Scan completion time depends on the amount of data, ranging from minutes to hours.

Scan file shares with NetApp Data Classification

To scan file shares, you must first create a file shares group in NetApp Data Classification. File shares groups are for NFS or CIFS (SMB) shares hosted on-premises or in the cloud.

Scanning data from non-NetApp file shares is not supported in the Data Classification core version.

Prerequisites

Review the following prerequisites to make sure that you have a supported configuration before you enable Data Classification.

- The shares can be hosted anywhere, including in the cloud or on-premises. CIFS shares from older NetApp 7-Mode storage systems can be scanned as file shares.
  - Data Classification can't extract permissions or the "last access time" from 7-Mode systems.
  - Because of a known issue between some Linux versions and CIFS shares on 7-Mode systems, you

must configure the share to use only SMBv1 with NTLM authentication enabled.

- There needs to be network connectivity between the Data Classification instance and the shares.
- You can add a DFS (Distributed File System) share as a regular CIFS share. Because Data Classification is unaware that the share is built upon multiple servers/volumes combined as a single CIFS share, you might receive permission or connectivity errors about the share when the message really only applies to one of the folders/shares that is located on a different server/volume.
- For CIFS (SMB) shares, ensure that you have Active Directory credentials that provide read access to the shares. Admin credentials are preferred in case Data Classification needs to scan any data that requires elevated permissions.

  If you want to make sure your files "last accessed times" are unchanged by Data Classification scans, it's recommended the user has Write Attributes permissions in CIFS or write permissions in NFS. If possible, configure the Active Directory user as part of a parent group in the organization which has permissions to all files.

- All CIFS file shares in a group must use the same Active Directory credentials.
- You can mix NFS and CIFS (using either Kerberos or NTLM) shares. You must add the shares to the group separately. That is, you must complete the process twice—once per protocol.
  - You cannot create a file shares group that mixes CIFS authentication types (Kerberos and NTLM).
- If you're using CIFS with Kerberos authentication, ensure the IP address provided is accessible to the Data Classification. The files shares can't be added if the IP address is unreachable.

Create a file shares group

When you add file shares to the group, you must use the format `<host_name>:/<share_path>`.

You can add file shares individually or you can enter a line-separated list of the file shares you want to scan. You can add up to 100 shares at a time.

Steps

1. From the Data Classification menu, select **Configuration**.
2. From the Configuration page, select **Add System** > **Add File Shares Group**.
3. In the Add File Shares Group dialog, enter the name for the group of shares then select **Continue**.
4. Select the protocol for the file shares you are adding.

   a. If you're adding CIFS shares with NTLM authentication, enter the Active Directory credentials to access the CIFS volumes. Although read-only credentials are supported, it's recommended you provide full access with administrator credentials. Select **Save**.

5. Add the file shares that you want to scan (one file share per line). Then select **Continue**.
6. A confirmation dialog displays the number of shares that were added.

   If the dialog lists any shares that could not be added, capture this information so that you can resolve the issue. If the issue pertains to a naming convention, you can re-add the share with a corrected name.

7. Configure scanning on the volume:
   - To enable mapping-only scans on file shares, select **Map**.

- To enable full scans on file shares, select **Map & Classify**.
- To disable scanning on file shares, select **Off**.

> The switch at the top of the page for **Scan when missing "write attributes" permissions** is disabled by default. This means that if Data Classification doesn't have write attributes permissions in CIFS or write permissions in NFS, the system won't scan the files because Data Classification can't revert the "last access time" to the original timestamp.
> If you switch **Scan when missing "write attributes" permissions** to **On**, the scan resets the last accessed time and scans all files regardless of permissions.
> To learn more about the last accessed time stamp, see Metadata collected from data sources in Data Classification.

Result

Data Classification starts scanning the files in the file shares you added. You can Track the scanning progress and view the results of the scan in the **Dashboard**.

> If the scan doesn't complete successfully for a CIFS configuration with Kerberos authentication, check the **Configuration** tab for errors.

Edit a file shares group

After you create a file shares group, you can edit the CIFS protocol or add and remove file shares.

Edit the CIFS protocol configuration

1. From the Data Classification menu, select **Configuration**.
2. From the Configuration page, select the file shares group you want to modify.
3. Select **Edit CIFS Credentials**.

4. Choose the authentication method: **NTLM** or **Kerberos**.
5. Enter the Active Directory **Username** and **Password**.
6. Select **Save** to complete the process.

Add file shares to scans

1. From the Data Classification menu, select **Configuration**.
2. From the Configuration page, select the file shares group you want to modify.
3. Select **+ Add shares**.
4. Select the protocol for the file shares you are adding.

   If you're adding file shares to a protocol you've already configured, no changes are required.

   If you're adding file shares with a second protocol, ensure you've properly configured the authentication as detailed in the prerequisites.

5. Add the file shares you want to scan (one file share per line) using the format `<host_name>:/<share_path>`.
6. Select **Continue** to complete adding the file shares.

Remove a file share from scans

1. From the Data Classification menu, select **Configuration**.
2. Select the system you want to remove file shares from.
3. Select **Configuration**.
4. From the Configuration page, select the Actions ••• for the file share you want to remove.
5. From the Actions menu, select **Remove Share**.

Track the scanning progress

You can track the progress of the initial scan.

1. Select the **Configuration** menu.
2. Select the **System Configuration**.
3. For the storage repository, check the Scan progress column to view its status.

Scan StorageGRID data with NetApp Data Classification

Complete a few steps to start scanning data within StorageGRID directly with NetApp Data Classification.

Review StorageGRID requirements

Review the following prerequisites to make sure that you have a supported configuration before you enable Data Classification.

- You need to have the endpoint URL to connect with the object storage service.
- You need to have the Access Key and Secret Key from StorageGRID so that Data Classification can access the buckets.

Deploy the Data Classification instance

Deploy Data Classification if there isn't already an instance deployed.

If you are scanning data from StorageGRID that is accessible over the internet, you can deploy Data Classification in the cloud or deploy Data Classification in an on-premises location that has internet access.

If you are scanning data from StorageGRID that has been installed in a dark site that has no internet access, you need to deploy Data Classification in the same on-premises location that has no internet access. This also requires that the Console agent is deployed in that same on-premises location.

Add the StorageGRID service to Data Classification

Add the StorageGRID service.

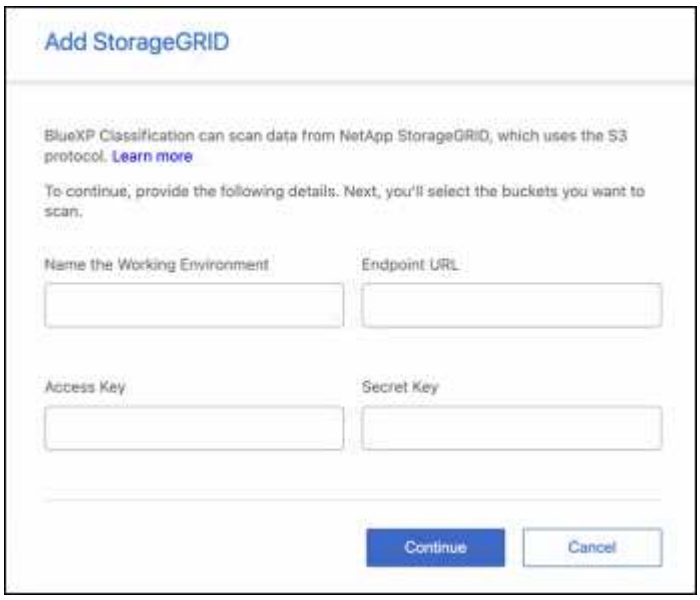
Steps

1. From the Data Classification menu, select the **Configuration** option.
2. From the Configuration page, select **Add System** > **Add StorageGRID**.
3. In the Add StorageGRID Service dialog, enter the details for the StorageGRID service and select **Continue**.
    a. Enter the name you want to use for the System. This name should reflect the name of the

StorageGRID service to which you are connecting.

b. Enter the Endpoint URL to access the object storage service.

c. Enter the Access Key and Secret Key so that Data Classification can access the buckets in StorageGRID.



Result

StorageGRID is added to the list of systems.

Enable and disable scans on StorageGRID buckets

After you enable Data Classification on StorageGRID, the next step is to configure the buckets that you want to scan. Data Classification discovers those buckets and displays them in the system you created.

Steps

1. In the Configuration page, locate the StorageGRID system.
2. On the StorageGRID system tile, select **Configuration**.
3. Complete one of the following steps to enable or disable scanning:
   - To enable mapping-only scans on a bucket, select **Map**.
   - To enable full scans on a bucket, select **Map & Classify**.
   - To disable scanning on a bucket, select **Off**.

Result

Data Classification starts scanning the buckets that you enabled. You can track the progress of the initial scan by navigating to the **Configuration** menu then selecting the **System configuration**. The progress of each scan is show as a progress bar. You can also hover over the progress bar to see the number of files scanned relative to the total files in the volume. If there are any errors, they'll appear in the Status column, alongside the required action to fix the error.

Integrate your Active Directory with NetApp Data Classification

You can integrate a global Active Directory with NetApp Data Classification to enhance the results that Data Classification reports about file owners and which users and groups have access to your files.

When you set up certain data sources (listed below), you need to enter Active Directory credentials in order for Data Classification to scan CIFS volumes. This integration provides Data Classification with file owner and permissions details for the data that resides in those data sources. The Active Directory entered for those data sources might differ from the global Active Directory credentials you enter here. Data Classification will look in all integrated Active Directories for user and permission details.

This integration provides additional information in the following locations in Data Classification:

- You can use the "File Owner" filter and see results in the file's metadata in the Investigation pane. Instead of the file owner containing the SID (Security IDentifier), it is populated with the actual user name.

  You can also view more details about the file owner: account name, email address, and SAM account name, or view items owned by that user.

- You can see full file permissions for each file and directory when you click the "View all Permissions" button.
- In the Governance dashboard, the Open Permissions panel will show a greater level of detail about your data.

Local user SIDs, and SIDs from unknown domains, are not translated to the actual user name.

Supported data sources

An Active Directory integration with Data Classification can identify data from within the following data sources:

- On-premises ONTAP systems
- Cloud Volumes ONTAP
- Azure NetApp Files
- FSx for ONTAP

Connect to your Active Directory server

After you've deployed Data Classification and have activated scanning on your data sources, you can integrate Data Classification with your Active Directory. Active Directory can be accessed using a DNS Server IP address or an LDAP Server IP address.

The Active Directory credentials can be read-only, but providing admin credentials ensures that Data Classification can read any data that requires elevated permissions. The credentials are stored on the Data Classification instance.

For CIFS volumes/file shares, if you want to make sure your files "last accessed times" are unchanged by Data Classification classification scans, the user should have Write Attributes permission. If possible, we recommend making the Active Directory configured user part of a parent group in the organization which has

permissions to all files.

Requirements

- You must have an Active Directory already set up for the users in your company.
- You must have the information for the Active Directory:
  - DNS Server IP address, or multiple IP addresses

    or

    LDAP Server IP address, or multiple IP addresses
  - User Name and Password to access the server
  - Domain Name (Active Directory Name)
  - Whether you are using secure LDAP (LDAPS) or not
  - LDAP Server Port (typically 389 for LDAP, and 636 for secure LDAP)
- The following ports must be open for outbound communication by the Data Classification instance:

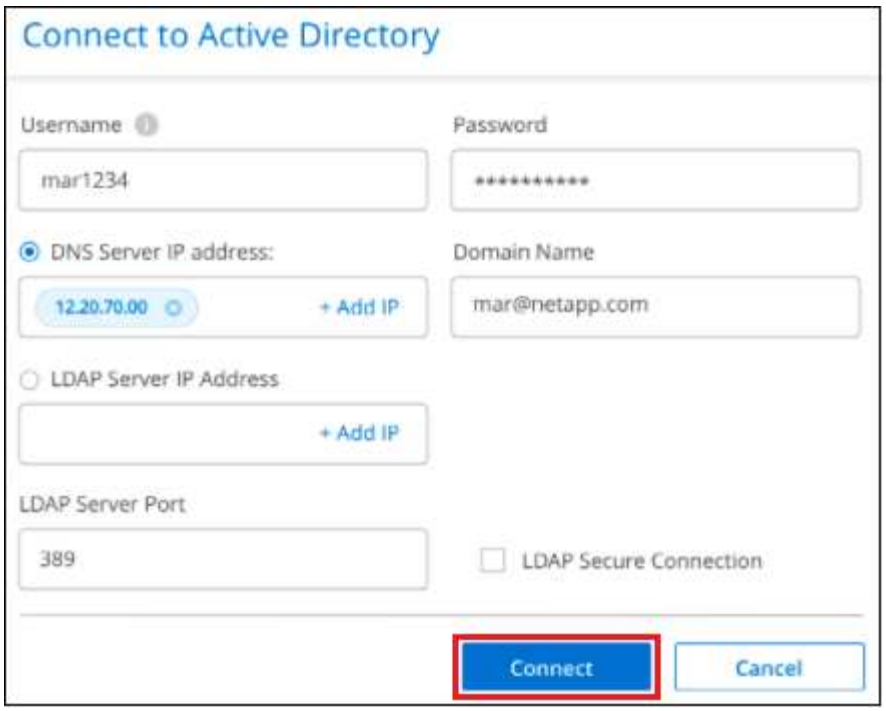
Protocol	Port	Destination	Purpose
TCP & UDP	389	Active Directory	LDAP
TCP	636	Active Directory	LDAP over SSL
TCP	3268	Active Directory	Global Catalog
TCP	3269	Active Directory	Global Catalog over SSL

Steps

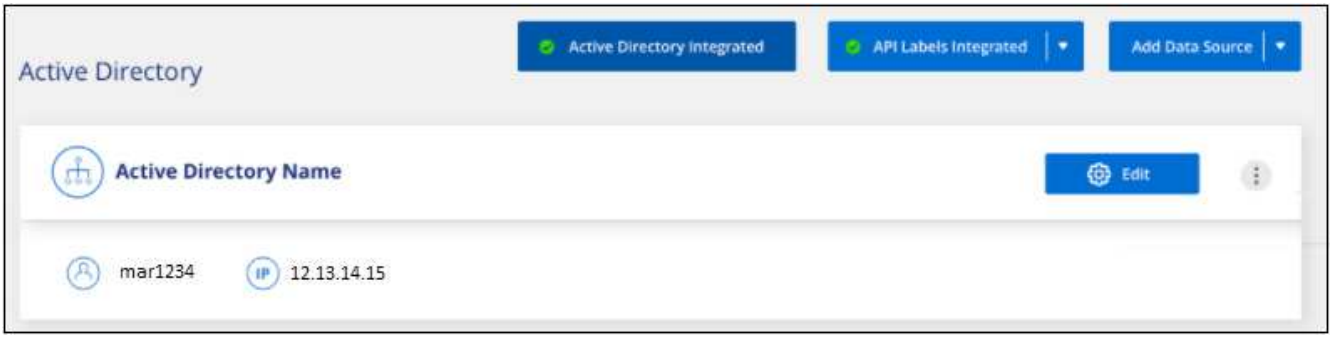
1. From the Data Classification Configuration page, click **Add Active Directory**.

2. In the Connect to Active Directory dialog, enter the Active Directory details and click **Connect**.

   You can add multiple IP addresses, if required, by selecting **Add IP**.



Data Classification integrates to the Active Directory, and a new section is added to the Configuration page.



Manage your Active Directory integration

If you need to modify any values in your Active Directory integration, click the **Edit** button and make the changes.

You can also delete the integration selecting the ⋮ button then **Remove Active Directory**.

Copyright information

Trademark information