AI Data Engine documentation

AI Data Engine

NetApp
March 12, 2026

Table of Contents

AI Data Engine documentation

Release notes

What's new in AI Data Engine

AI Data Engine (AIDE) 9.18.1 is the initial release of NetApp's platform for AI data management. This release introduces a metadata engine and management workflows that enable organizations to catalog and organize unstructured data for AI workloads, providing the foundation for advanced governance and vectorization capabilities. Advanced governance (guardrails) and vectorization are available for customers who have the appropriate AI Data Engine licenses.

What's new in the AIDE 9.18.1 initial release

AIDE 9.18.1 introduces the following foundational capabilities:

Metadata Engine for AI data cataloging

The initial release includes a metadata engine that catalogs files and objects across ONTAP clusters.

Key features include:

- Automated extraction of metadata (core and extended attributes, object tags) from local and remote ONTAP volumes on peered clusters.
- Centralized querying and filtering REST APIs for applications requiring a global view of enterprise data.
- Scalable metadata storage.
- Automatic metadata extraction triggered during workspace creation.

Workspace management

Workspaces provide logical grouping of data sources (volumes) for AI projects.

The initial release supports:

- Creation of workspaces spanning local and remote ONTAP volumes (using cluster peering).
- Assignment of access controls to workspaces, supporting multi-user and multi-tenant environments.
- Automatic metadata extraction and catalog population upon workspace creation.

Data Sync for automated data currency

Data Sync keeps metadata catalogs and data collections current as source data changes, without manual intervention.

Key features include:

- Automated synchronization of data from remote or local ONTAP clusters using policy-driven SnapMirror replication.
- Incremental updates that propagate only modified data, reducing overhead.
- Configurable refresh intervals per workspace.
- Workspace-level monitoring of sync status and activity.

Cluster setup and management

The initial release includes the following workflows:

- Discovery and addition of data compute nodes (DCNs) during cluster setup.
- Creation of dedicated metadata storage VMs for the metadata engine.
- Configuration of Data Engine service interfaces for cluster-wide metadata access.
- Peering with other ONTAP clusters to extend metadata cataloging across the data estate.

OpenID Connect (OIDC) authentication

- OIDC/OAuth-based authentication for secure access to ONTAP System Manager and Data Engine Console with Microsoft Entra ID and Active Directory Federation Services (ADFS).
- Role-based access controls for workspace and metadata management.

Advanced data curation and governance capabilities

The following capabilities are available for customers who have the appropriate AI Data Engine licenses:

- **Vectorization and RAG**: Creation of data collections, embeddings, and retrieval endpoints in the AI Data Engine Console, using metadata from AIDE workspaces.
- **Guardrail-based governance**: Definition of guardrail policies in the AI Data Engine Console and association of those policies with workspaces in ONTAP System Manager.

Supported hardware and platforms

AI Data Engine 9.18.1 runs on ONTAP AI data platform clusters that combine:

- AFX 1K storage nodes
- NetApp data compute nodes

Related information

- Known limitations of AI Data Engine
- Learn about AI Data Engine architecture and components

Known limitations of AI Data Engine

Known limitations identify platforms, devices, or functions that are not supported by this release of the product, or that do not interoperate correctly with it. Review these limitations carefully.

Known limitations for AIDE 9.18.1 initial release

These limitations apply to the Metadata Engine, data compute nodes, and management workflows in AIDE 9.18.1.

Compute node requirements and management

- **Minimum data compute node requirement**

  AIDE clusters require a minimum and maximum of 3 data compute nodes (DCNs) for Metadata Engine functionality. Clusters with fewer than 3 DCNs cannot enable Metadata Engine features.

- **No support for adding DCN nodes using NetApp Console**

  DCN node upgrades and additions must be performed using ONTAP System Manager, not through NetApp Console.

Supported data sources

- **No support for ONTAP S3 buckets or StorageGRID as data sources**

  Only ONTAP volumes (local or remote) are supported as data sources for workspaces and metadata cataloging. ONTAP S3 buckets and StorageGRID objects cannot be added to workspaces and are not indexed by the Metadata Engine in this release.

- **No support for workspace creation with FlexCache volumes**

  FlexCache volumes cannot be added as data sources to workspaces.

Software update and revert limitations

- **Manual software updates for DCN nodes only**

  Automatic software updates for DCN clusters are not supported in AIDE 9.18.1. DCN node software can only be updated by uploading the image from a local client. Downloading images from external servers (HTTP/FTP) is not supported.

- **No revert of DCN cluster software**

  DCN cluster software cannot be reverted to an earlier version. Only upgrades to later versions are allowed.

- **No ONTAP revert of AFX storage clusters**

  AFX storage clusters cannot be reverted to earlier ONTAP versions. Only upgrades to later versions are allowed.

Workspace lifecycle and access configuration

- **No soft delete or restore for workspaces**

  Deleting a workspace is permanent. There is no option to restore deleted workspaces.

- **No support for OIDC configuration during initial cluster setup**

  OIDC/OAuth configuration must be performed after cluster creation using ONTAP System Manager.

Related information

- What's new in AI Data Engine

Get started

Learn about your AI Data Engine system

Learn about AI Data Engine

The NetApp AI Data Engine (AIDE) is an enterprise-grade platform designed to accelerate and simplify AI-driven data processing, management, and governance. AIDE can help transform large amounts of unstructured data into structured, AI-ready datasets. It is engineered to meet the demands of modern machine learning (ML) and generative AI (GenAI) workloads, supporting both traditional IT operations and new AI-centric roles.

AIDE addresses AI challenges

AIDE is designed to help organizations manage data for AI workloads and provides the following key capabilities:

- **Centralized metadata management**: AIDE collects and catalogs metadata from ONTAP volumes, making it possible to search, classify, and apply governance policies to datasets.
- **Automated data processing**: AIDE supports the creation of data pipelines for AI and ML workloads, including the ability to generate vector embeddings for semantic search (with appropriate licensing).
- **Data isolation and access control**: AIDE enforces access controls and basic data isolation for multiple teams or projects.
- **Integration with NetApp tools**: AIDE works with ONTAP System Manager for storage administration and provides a dedicated interface (AI Data Engine Console) for data engineers and scientists to manage data collections and workflows.

High-level design characteristics

The following design characteristics define how AI Data Engine is built to meet the needs of AI workloads:

- **Microservices-based services**: Uses Kubernetes to orchestrate modular, resilient services for metadata cataloging, vector search, and infrastructure management.
- **Enterprise-grade security**: Implements encryption, role-based access control (RBAC), and auditing across all data and metadata.
- **Multi-protocol data access**: Supports NFS and SMB for flexible data ingestion and retrieval.
- **Automated data pipelines**: Tracks data changes, creates embeddings, and manages vector databases for AI applications.

How data flows through AIDE

Understanding how data flows through AIDE helps illustrate the platform’s value for AI/ML teams:

1. **Data ingestion**: Files are stored in ONTAP volumes using standard protocols (NFS and SMB). Data can reside on local AIDE storage (the AFX cluster within your AIDE deployment) or on remote ONTAP clusters. Data from remote clusters is synchronized to the local AFX cluster using ONTAP SnapMirror, so all data processed by AIDE is ultimately stored and accessed locally.

S3 buckets are not supported as data sources for workspaces or data collections.

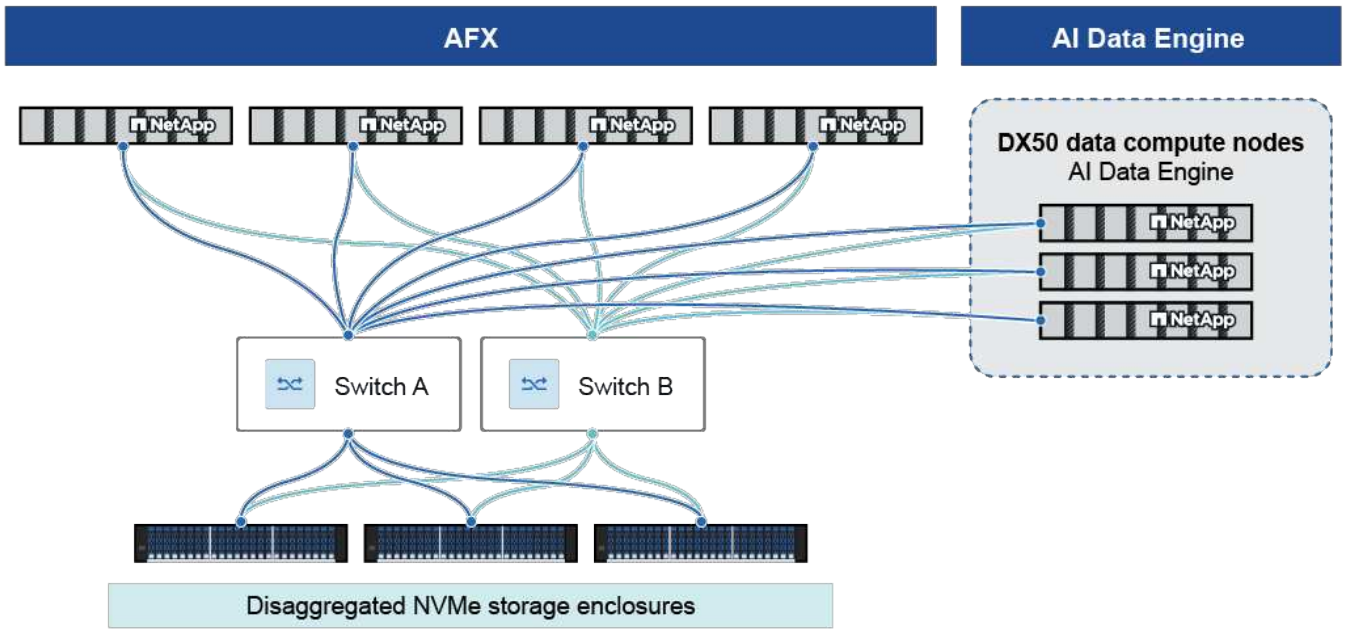
1. **Workspace creation**: Storage administrators define workspaces in ONTAP System Manager, grouping related ONTAP volumes for specific projects, teams, or workflows. Access permissions and governance policies are assigned at the workspace level.
2. **Metadata extraction**: AIDE automatically scans files and objects in workspaces, extracting metadata (file type, size, timestamps, custom attributes) and storing it in a centralized catalog. This happens continuously as data changes.
3. **Classification and governance**: Classifiers scan data for sensitive information (PII, financial data) or document types (legal, HR). Guardrail policies enforce redaction or access restrictions automatically.
4. **Data collection creation**: Data engineers and data scientists use the AI Data Engine Console to query the metadata catalog, filter results, and assemble curated data collections for specific AI tasks.
5. **Vectorization**: For collections requiring semantic search, AIDE generates embeddings using selected AI models. Vectors are stored in the vector database for high-performance retrieval.
6. **AI/ML consumption**: Applications access data through multiple paths:
   - Direct file/object access using NFS or SMB
   - Semantic search queries against the vector database
   - RAG endpoints that combine data retrieval with GenAI model integration
   - REST API access for programmatic workflows

This automated, policy-driven workflow reduces the time and manual effort required to prepare data for AI, enabling teams to focus on model development and insights rather than data wrangling.

AI Data Engine architecture

AIDE is built on a scalable, fault-tolerant architecture that separates storage and compute, enabling high performance and flexibility for AI workloads.

Physical components



AFX controller nodes

AFX controller nodes run a specialized personality of the ONTAP software designed to support the requirements of the AFX environment. Clients access the nodes through multiple protocols, including NFS and SMB. Each node has a complete view of the storage, which it can access based on the client requests. The nodes are stateful with non-volatile memory to persist critical state information and include additional enhancements specific to the target workloads.

At least four AFX controller nodes are required for AIDE deployments to ensure high availability and performance.

Data compute nodes

Data compute nodes (DCNs) are Linux-based servers with high CPU, RAM, and GPU resources, dedicated to AI data processing tasks. They host AI-specific services such as metadata cataloging, vector search, and embedding pipelines.

Exactly *three* DCNs are required for AIDE deployments.

Cluster/storage switches

Redundant, high-speed (100GbE or higher) switches connect ONTAP and DCNs for low-latency data transfer and high availability.

Storage shelves

NVMe-oF shelves with high-density SSDs provide ultra-low latency and redundancy, supporting PB-scale storage.

Networking

All DCNs and ONTAP storage nodes are connected through redundant, high-speed cluster switches (minimum 100GbE). This architecture separates compute and storage resources, allowing each to scale independently and optimizing both performance and resource utilization.

Networking between DCNs and ONTAP nodes is isolated using dedicated VLANs and IPspaces on the cluster switches. This ensures that all communications, such as data access, management APIs, and internal service traffic, remain secure, efficient, and do not interfere with other network operations.

AI Data Engine primary features

The AI Data Engine (AIDE) primary features work together to automate, secure, and accelerate the AI data lifecycle. Each feature is implemented as a set of microservices running on DCNs, integrated with ONTAP storage, and exposed through REST APIs and management interfaces.

Metadata Engine

The Metadata Engine automatically generates a structured, up-to-date, and interactive view of your NetApp data estate.

License and access

The Metadata Engine is included with the base ONTAP One license and is available upon AIDE installation.

You can access it through ONTAP System Manager.

Capabilities

- Catalogs metadata for all data sources, including volumes stored locally on the AFX cluster and those synchronized from remote ONTAP clusters.
- Extracts metadata automatically and populates the catalog as data is ingested or changed.
- Provides REST API access for querying metadata, allowing data practitioners and storage administrators to discover, classify, and understand data.
- Offloads metadata queries from the data path, reducing NFS traffic load on storage systems.
- Supports large metadata records with indexing and search capabilities.
- Integrates with workspace and data collection abstractions to enforce access control and governance.

Data Sync

Data Sync is an automated background service that ensures that the metadata catalog and data collections remain current and consistent with the underlying data sources, even as source data changes.

License and access

Data Sync functionality is not included with the base ONTAP One license and requires a separate AIDE license.

Capabilities

- Synchronizes data from remote or local ONTAP clusters using policy-driven SnapMirror replication. Data from remote clusters is copied to the local AFX cluster for AIDE processing.
- Updates incrementally based on detected changes, propagating only modified data.
- Provides secure, incremental data mobility and synchronization across the data estate.
- Schedules and monitors sync intervals with configurable refresh rates per workspace.
- Integrates with workspace creation workflows to extract and update metadata as new data sources are added.

Data Guardrails

The Data Guardrails service provides continuous, automated governance and protection for sensitive data throughout the AI lifecycle.

License and access

Data Guardrails functionality is not included with the base ONTAP One license and requires a separate AIDE license.

You can access guardrail functionality through the AI Data Engine Console.

Capabilities

- Continuously scans, classifies, and categorizes data.
- Identifies sensitive data and risks using built-in and customizable classifiers for tasks such as PII detection.
- Automates handling of sensitive data through policy-driven redaction, masking, and access restrictions.
- Enforces company and regulatory standards through guardrail policies attached to workspaces.
- Restricts access to sensitive files or volumes as configured, with audit logging and compliance reporting.
- Integrates with workspace and data collection management to apply guardrails consistently across AI data workflows.

Data Curator

The Data Curator service enables fast data discovery, search, vectorization, and retrieval for AI and GenAI applications.

License and access

Data Curator functionality is not included with the base ONTAP One license and requires a separate AIDE license.

You can access data curator through the AI Data Engine Console.

Capabilities

- Searches storage for relevant data using the centralized metadata catalog.
- Provides tools for data scientists to create curated data collections.
- Generates vector embeddings automatically at the storage layer.
- Provides a secure retrieval endpoint for AI applications, supporting vector semantic search and re-ranking.
- Integrates with AI tools and technologies, including Retrieval-Augmented Generation (RAG) pipelines and agentic AI frameworks.
- Provides REST APIs for programmatic access to data collections, vector search, and retrieval endpoints.

Security and multi-tenancy

The platform enforces both role-based access control (RBAC) and resource-level access control lists (ACLs). All API and user actions are audited, and all data is encrypted at rest and in transit. Individual tenants are isolated for data and metadata.

Related information

- Install an AIDE license
- Data-to-RAG quick start

AIDE components and responsibilities by role

AI Data Engine components and role-based interactions

AI Data Engine (AIDE) consists of many core components that work together to provide a comprehensive data management and processing platform for AI workloads. These components include workspaces, data collections, vector databases, guardrails, metadata catalogs, retrieval endpoints, and classifiers. Each component plays a specific role in enabling efficient data discovery, curation, governance, and integration with AI/ML applications.

Each AIDE user interacts with AIDE components differently according to their role.

Storage and data focused user roles

AIDE introduces new user roles while still supporting traditional ONTAP system administration roles:

Storage users

- **Storage administrator**: Manages AFX and AIDE cluster setup, networking, storage provisioning, and user access.

Data users

- **Data engineer**: Builds and optimizes AI/ML pipelines, manages data collections, and integrates AI models.
- **Data scientist**: Discovers, curates, and analyzes datasets, creates data collections, and leverages retrieval endpoints for GenAI applications.

Role (RBAC name)	Description
Storage administrator (`admin`)	Manages AFX and AIDE cluster setup, networking, storage provisioning, and user access. Assigns RBAC roles to users that determine the level of access to AIDE interfaces and features. This admin role has full management access using ONTAP System Manager and AI Data Engine Console.
Data engineer (`data-engineer`)	Builds and optimizes AI/ML pipelines, manages data collections, and integrates AI models. This role has access to the AI Data Engine Console for data engineering workflows.
Data scientist (`data-scientist`)	Discovers, curates, and analyzes datasets, creates data collections, and leverages retrieval endpoints for GenAI applications. This role has access to the AI Data Engine Console for data science workflows.

AIDE system components

Each AIDE user (storage administrators, data engineers, and data scientists) interacts with AIDE components according to their role.

Workspaces

A workspace is a logical segment of data within the cluster, grouping volumes for a specific project, team, or workflow. Workspaces define the scope of data visibility, access, and governance in AIDE.

Metadata catalog

A centralized, scalable database storing metadata records for all files and objects across the local cluster, including data synchronized from remote ONTAP clusters using ONTAP SnapMirror or cluster peering. It enables rich, interactive search and filtering.

Classifiers

Classifiers are tools (built-in or custom) that scan and tag files for specific types of sensitive data (for example, PII, financial, healthcare) or categorize documents by type (for example, legal, HR, sales).

Data collections

A data collection is a curated group of related files or objects from a workspace, defined by a user-specified query for use in GenAI workflows. The content of the files in the data collection, after publication, is available for semantic search by APIs for GenAI applications.

Vector database

The vector database stores embeddings generated from data collections, enabling high-performance semantic search and retrieval for AI and GenAI applications.

Guardrails

Guardrails are policy-driven mechanisms that enforce data governance, classification, and protection (such as redaction or access restrictions) throughout the AI data lifecycle.

Retrieval endpoint (RAG endpoint)

A retrieval endpoint (sometimes called a Retrieval-Augmented Generation or "RAG" endpoint) is a secure API that enables AI and GenAI applications to access relevant data, context, or embeddings from curated collections and the vector database.

RAG endpoints are designed to support advanced AI workflows, such as semantic search and context-aware responses in generative AI models. By connecting your AI applications to a retrieval endpoint, you can enhance model accuracy and relevance by providing real-time access to curated, AI-ready datasets managed by AIDE.

Related information

- How AIDE storage administrators work with AIDE components
- How AIDE data engineers work with AIDE components
- How AIDE data scientists work with AIDE components

AI Data Engine interfaces

AI Data Engine (AIDE) provides three primary interfaces for user interaction and automation. Each role, such as storage administrators, data engineers, and data scientists, utilizes these interfaces based on their specific tasks and responsibilities.

ONTAP System Manager

ONTAP System Manager is a web-based interface designed for storage administrators. It provides workflows for cluster setup, workspace management, DCN monitoring, and attaching guardrail policies.

AI Data Engine Console

The AI Data Engine Console is a dedicated interface for data engineers and data scientists. It enables users to explore data sources, create and manage data collections, configure data pipelines, apply classifiers, and interact with guardrails and vector search features. The console provides advanced tools for data discovery, curation, and integration with AI/ML workflows.

REST API

AIDE exposes the ONTAP REST API for automation, integration, and programmatic access. The API supports cluster setup, workspace and collection management, metadata queries, vector search, and retrieval endpoints.

Learn how AI Data Engine storage administrators work with AIDE components

As a storage administrator, you manage the AIDE infrastructure through ONTAP and the

AIDE Console, provisioning workspaces, attaching guardrail policies, and monitoring system health. Your role focuses on ensuring reliable, secure, and compliant data storage for AI workloads.

Storage administrator component access

Component	Access level	Storage administrator workflow
ONTAP System Manager	Manage (create, edit, delete)	You use ONTAP System Manager as your primary interface for cluster administration, workspace provisioning, guardrail policy management, and monitoring system health.
AI Data Engine Console	Manage (create, edit, delete)	You use the AI Data Engine Console to monitor workspaces, view collection status, and oversee system activity across the AIDE environment.
ONTAP REST API	Manage (create, edit, delete)	You use the REST API to automate infrastructure tasks, manage workspaces and guardrail policies programmatically, and integrate AIDE administration with external tools and workflows.
Workspaces	Manage (create, edit, delete)	You create and manage workspaces using ONTAP System Manager. You select which data sources are included, assign permissions to data engineers and data scientists, and attach guardrail policies to enforce governance and compliance. You also monitor workspace health and access.
Data collections	View (read-only)	You view the status and health of data collections within each workspace using System Manager. You ensure underlying data sources are available and protected, but you do not create or modify collections.
Guardrails	Manage (create, edit, delete)	You define and attach guardrail policies to workspaces using System Manager. You monitor guardrail status and compliance reports. You ensure policies are enforced and updated as needed.
Metadata catalog	Monitor (view health, status, activity)	You ensure the metadata catalog is populated and up to date. You monitor catalog health and support access control.
Vector database	Provision/Monitor (deploy, configure, view status)	You provision and monitor the vector database infrastructure, ensuring data compute nodes with GPU resources and proper licensing are in place. You support the environment but do not manage embeddings or queries directly.

Component	Access level	Storage administrator workflow
Classifiers	Manage (create, edit, delete)	You create, configure, and manage classifiers and their categories. You apply classifiers to workspaces and monitor their effectiveness.

Learn how AI Data Engine data engineers and data scientists work with AIDE components

As a data engineer or data scientist, you use the AI Data Engine Console to explore workspaces you have been granted access to, create and manage data collections, perform semantic searches, and integrate retrieval endpoints into AI/ML workflows.

Data engineers focus on transforming raw data into AI-ready datasets by building collections, configuring embedding pipelines, and controlling which users can access published collections. Data scientists focus on leveraging curated datasets for analysis, model training, and GenAI applications, without managing access control or infrastructure.

Data user component access

Component	Access level	Data engineer workflow	Data scientist workflow
AI Data Engine Console	Manage (create, edit, delete)	The AI Data Engine Console is your primary interface for day-to-day tasks, including data discovery, collection management, pipeline configuration, and publishing RAG or retrieval endpoints, for the workspaces you are authorized to access.	The AI Data Engine Console is your primary interface for data exploration, refining and versioning collections within workspaces you can access, and connecting curated datasets and retrieval endpoints to analysis, modeling, and GenAI workflows.
ONTAP REST API	Manage (create, edit, delete)	You use the REST API to automate collection lifecycle operations, trigger and monitor embedding pipelines, and programmatically integrate data workflows with external tools.	You use the REST API to programmatically access data collections, run vector search queries, and integrate retrieval endpoints into AI/ML applications and agentic frameworks.
Workspaces	View/use (read-only)	You explore your assigned workspaces to identify and understand available data sources before building collections.	You search your assigned workspaces to locate files and objects relevant to specific research or modeling tasks.

Component	Access level	Data engineer workflow	Data scientist workflow
Data collections	Manage (create, edit, delete)	You build data collections by selecting and filtering source data using tags, classification, and other attributes, and you manage the full collection lifecycle from creation and versioning through publishing as RAG endpoints for AI use. You also manage which data scientists and other users can access each collection.	You create, select, annotate, version, and refine data collections within the workspaces you have been given access to. You use these collections as the basis for semantic search and GenAI workflows.
Metadata catalog	Query/use (consume for workflows)	You use the metadata catalog to evaluate and select data sources for ingestion, running queries to locate relevant files and confirm they meet the requirements of the collections you are building within your assigned workspaces.	You search and filter metadata across the workspaces you can access to locate files and objects needed for analysis or model training, relying on the catalog structure that has been built and maintained by data engineers.
Vector database	• Manage embeddings/search (data engineer) • Use/search (data scientist)	You trigger embedding pipelines, monitor vectorization status, configure chunking and embedding parameters, and expose retrieval endpoints backed by vector search. Applications and agents then query these endpoints via the API for semantic search and RAG workflows.	You run semantic search queries against embeddings generated by data engineer-managed pipelines and integrate retrieval results into GenAI or RAG workflows for context-aware model responses. You do not configure chunking, embeddings, or pipeline parameters.
Classifiers	Use (consume classified data)	You use classification results to annotate and tag source data during collection preparation, ensuring that content entering your pipelines is properly labeled for downstream AI workflows.	You consume pre-classified data to ensure that only compliant and relevant content is used in your analysis and modeling.

Quick start for AI Data Engine

To get up and running with your AI Data Engine system, you need to install your hardware components, set up your cluster, set up data access from your hosts to the storage system, and provision your storage.

1 **Will AIDE be installed with a new or existing AFX cluster?**

You need to decide if AIDE and AFX will be installed together at the same time or if AIDE will be integrated with an existing AFX cluster.

2 **Install and set up your hardware**

Install and set up your AIDE cluster compute nodes. Depending on the installation environment, also make sure to install the AFX hardware.

3 **Set up your cluster**

Use ONTAP System Manager to guide you through a quick and easy process to Set up AIDE with an AFX cluster.

4 **Set up workspaces and data access**

Set up a workspace and users for that workspace who can access AI Data Engine data.

What's next?

You can now use ONTAP System Manager to manage your AI Data Engine and get your data engineers and data scientists started with their workspaces and configurations.

Install AIDE

Requirements for installing AI Data Engine

Review the requirements for installing the AI Data Engine. AIDE requires an AFX storage system, a minimum of three data compute nodes, network switches, and cables.

Hardware requirements

The AI Data Engine requires an AFX storage system and a minimum of three data compute nodes. The AFX system provides the storage infrastructure, while the data compute nodes host the AIDE software components that enable data management, curation, and AI capabilities.

- **AFX storage system**: Includes an AFX controller, disk shelf, and network switch. The AFX storage system is required for AIDE deployment.
- **Data compute nodes**: A minimum of three data compute nodes are required. The data compute nodes are NetApp-provided hardware nodes that host the AIDE software, including the Metadata Engine, Data Sync, Data Curator, and Data Guardrails.
  - Each data compute node has I/O slots 4 and 5 available for connectivity. Slot 3 is reserved for the GPU. Slots 1 and 2 are not populated or accessible.
  - Ports e4a and e5a are used for cluster connections.
  - Ports e4b and e5b are used for host network connections.

Network switch requirements

AI Data Engine requires network switches to enable host network connectivity and inter-node communication for the data compute nodes.

- Client switches for host network connectivity (Cisco Nexus 9332D-GX2B or Cisco Nexus 9364D-GX2A)

- Cluster switches for inter-node communication (Cisco Nexus 9332D-GX2B or Cisco Nexus 9364D-GX2A)
- Optional management switches for network management

Cabling requirements

The following cables are required to connect the data compute nodes to the network switches and management network.

- 400-GbE-to-100-GbE (4x100GbE) breakout cables for connecting nodes to client and cluster switches

Multi-cluster support

After AIDE is deployed with AFX, it can connect to and manage data from other ONTAP 9.18.1 and later clusters using SnapMirror and cluster peering.

Install the AFX storage system for AI Data Engine

Install the AFX storage system as the first step in deploying the AI Data Engine. The AFX storage system provides the storage infrastructure foundation and is required before installing the data compute nodes.

Follow the AFX 1K installation documentation to install the AFX storage system.

What's next

After completing the AFX storage system installation, install the data compute nodes.

Install your data compute nodes

Installation and setup workflow for data compute nodes for AI Data Engine

To install and configure your data compute nodes (DCN), you review the hardware requirements, prepare your site, install and cable the hardware components, power on the system, and set up your ONTAP cluster.

Review the hardware installation requirements

Ensure that you have an existing AFX 1K storage system installed, and then review the hardware requirements for installing the data compute nodes for AIDE. For information on installing the AFX 1K storage system, refer to AFX 1K storage system installation documentation.

Prepare to install your data compute nodes

To prepare to install your data compute nodes, you need to get the site ready, check the environmental and electrical requirements, and ensure there's enough cabinet space. Then, unpack the equipment, compare its contents to the packing slip, and register the hardware to access support benefits.

Install the hardware for your data compute nodes

Install the rail kits for your data compute nodes. Secure your data compute nodes within the cabinet. Finally,

attach cable management devices to the rear of the system for organized cable routing.

4 Cable your data compute nodes

To cable the hardware, first connect the data compute nodes to your data and cluster network, then connect the data compute nodes to the cluster switches.

5 Power on your data compute nodes

After you install the rack hardware and cable your data compute nodes, you should power on your DCNs and your controller nodes for the AFX storage system if not already powered on.

Installation requirements for data compute nodes for AI Data Engine

Review the equipment needed and the lifting precautions for your data compute nodes for AI Data Engine.

Prerequisites

Before you install the data compute nodes for AIDE, ensure that you have:

- An AFX 1K storage system.

> For information on installing the AFX 1K storage system, refer to AFX 1K storage system installation documentation.

Equipment needed for install

To install the data compute nodes for AIDE, you need the following equipment and tools.

- Access to a web browser to configure your data compute nodes
- Electrostatic discharge (ESD) strap
- Flashlight
- Laptop or console with a USB/serial connection
- Phillips #2 screwdriver

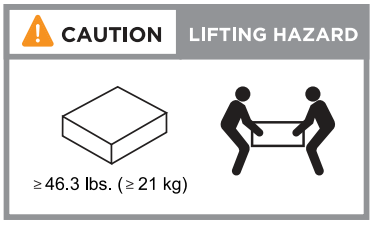
Lifting precautions

The data compute nodes are heavy. Exercise caution when lifting and moving these items.

Data compute node weights

Take the necessary precautions when moving or lifting your data compute node.

A data compute node can weigh up to 46.3 lbs (21 kg). To lift the data compute node, use two people or a hydraulic lift.



Related information

- Safety information and regulatory notices

What's next?

After you've reviewed the hardware requirements, prepare to install your data compute nodes.

Prepare to install your data compute nodes for AI Data Engine

Prepare to install your data compute nodes for AI Data Engine by getting the site ready, unpacking the boxes and comparing the contents of the boxes to the packing slip, and registering the system to access support benefits.

Step 1: Prepare the site

To install your data compute nodes, ensure that the site and the cabinet or rack that you plan to use meet specifications for your configuration.

Steps

1. Use NetApp Hardware Universe to confirm that your site meets the environmental and electrical requirements for your data compute nodes.
2. Make sure you have adequate cabinet space for your data compute nodes within the existing AFX 1K storage system installation.
   - 1U for each data compute node
   - 2U for each AFX 1K controller node
   - 2U for each NX224 shelf
   - 1U or 2U per switch, depending on switch model.

Step 2: Unpack the boxes

After you've ensured that the site and the cabinet or rack that you plan to use for your data compute nodes meet the required specifications, unpack all boxes and compare the contents to the items on the packing slip.

Steps

1. Carefully open all the boxes and lay out the contents in an organized manner.
2. Compare unpacked items against the packing slip and note any discrepancies.

   You can get your packing list by scanning the QR code on the side of the shipping carton.

   The following items are some of the contents you might see in the boxes.

Hardware	Cables

- Bezel
- Rail kits with instructions
- Data compute node(s)
- Power cords

Step 3: Register your data compute nodes

After you've ensured that your site meets the requirements for your data compute node specifications, and you've verified that you have all the parts you ordered, you should register your system.

Steps

1. Locate the serial numbers for your data compute nodes.

   You can find the serial numbers in the following locations:

   - On the packing slip
   - In your confirmation email
   - On each data compute node, or on some systems, on the system management module of each data compute node.

2. Navigate to the NetApp Support Site.
3. Determine whether you need to register your storage system:

If you are a…	Follow these steps…
Existing NetApp customer	a. Sign in with your username and password. b. Select **Systems** > **My Systems**. c. Confirm that the new serial number is listed. d. If the serial number is not listed, follow the instructions for new NetApp customers.
New NetApp customer	a. Click **Register Now** to create an account. b. Select **Systems** > **Register Systems**. c. Enter the storage system's serial number and requested details. Once NetApp approves your registration, you can download the required software. Approval takes up to 24 hours.

What's next?

After you've prepared to install your data compute nodes, install the data compute nodes.

Install your data compute nodes for AI Data Engine

Install and secure your data compute nodes in the cabinet.

Before you begin

- Make sure you have the instructions packaged with the rail kit.
- Understand the safety concerns related to the weight of the data compute node, storage system, and storage shelf.
- Understand that the airflow through the storage system enters from the front where the bezel or end caps are installed and exhausts out the rear where the ports are located.

> In general, the switches should be installed in the center of the cabinet. The storage shelves should be installed below the switch and above a second installed switch. The controller nodes can be installed above or below the switches within the cabinet. Data compute nodes can be installed above or below the controller nodes within the cabinet.

Steps

1. Install the rail kits for your data compute nodes, as needed, using the instructions included with the kits.
2. Install and secure your data compute nodes in the cabinet:
   a. Position the data compute node onto the rails in the middle of the cabinet, and then support the appliance from the bottom and slide it into place.
   b. Secure the data compute node to the cabinet using the included mounting screws.
3. Attach the bezel to the front of the data compute node.

What's next?

After you've installed the data compute nodes, cable the data compute nodes for AIDE.

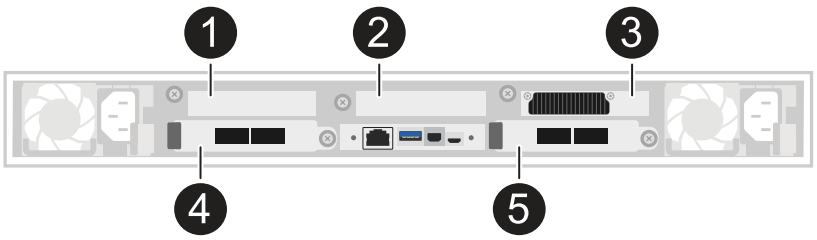
Requirements for cabling data compute nodes for AI Data Engine

Data compute nodes integrate with your AFX 1K storage system through host network and cluster network connections. Review the I/O slot configuration, cable types, and connection requirements for your deployment.

Cabling configuration

Data compute nodes connect to the same cluster switches as the AFX 1K controller nodes, extending your storage system with compute resources optimized for AI and machine learning workloads.

The initial AI Data Engine (AIDE) configuration supports a minimum of three data compute nodes. For comprehensive configuration details and slot priorities, see NetApp Hardware Universe.



1	Unused slot on the data compute node.
2	Unused slot on the data compute node.
3	GPU slot on the data compute node.
4	I/O slot on the data compute node.
5	I/O slot on the data compute node.

I/O slot configuration

The data compute node uses a specific slot numbering scheme that differs from standard server configurations. Understanding the slot layout is essential for proper cabling.

- **Slot 3**: Reserved for GPU (not accessible for I/O cabling)
- **Slots 4 and 5**: I/O slots used for network connections
  - Port a: Cluster network connections
  - Port b: Host network connections
- **Slots 1 and 2**: Unpopulated and inaccessible for use

Network connections

Data compute nodes require two types of network connections to integrate with the AFX 1K storage system.

- **Host network connections**

  Host network connections provide access to client data and enable the data compute nodes to process workloads. Each data compute node uses ports e4b and e5b for redundant connections to separate host network switches.

  Port assignments:

  - e4b: Connects to host network switch A
  - e5b: Connects to host network switch B
- **Cluster network connections**

  Cluster network connections enable communication between data compute nodes and AFX 1K controller nodes within the storage cluster. Each data compute node uses ports e4a and e5a for redundant connections to separate cluster network switches.

  Port assignments:

  - e4a: Connects to cluster network switch A
  - e5a: Connects to cluster network switch B

Supported hardware components

The data compute nodes require specific cables and switches to ensure proper connectivity and performance with the AFX 1K storage system.

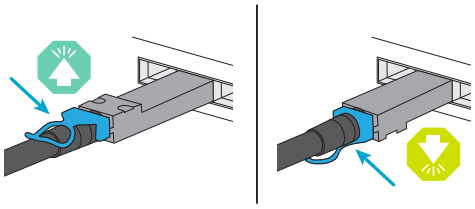
Data Compute Node	Supported Switches	Supported Cables
Data compute nodes (minimum of three required)	• Cisco Nexus 9332D-GX2B (400GbE) • Cisco Nexus 9364D-GX2A (400GbE)	• 400GbE QSFP-DD breakout to 4x100GbE QSFP56 cables for connections to data compute nodes: ◦ 100GbE to data compute node cluster network ports (e4a, e5a) ◦ 100GbE to data compute node host network ports (e4b, e5b) • RJ-45 cables for management connections

Breakout cables provide four 100GbE connections from each 400GbE switch port. Connect the 400GbE end to the switches and the 100GbE end to the data compute node I/O ports.

Cable orientation

When connecting cables to data compute nodes, proper cable orientation ensures reliable connections.

The cabling graphics in the installation procedures show arrow icons indicating the correct orientation (up or down) of the cable connector pull-tab when inserting a connector into a port. As you insert the connector, you should feel it click into place. If you do not feel it click, remove it, turn it over, and try again.



Handle the delicate connector components carefully when clicking them into place.

What's next?

After reviewing the cabling configuration, cable the hardware for your data compute nodes.

Cable your data compute nodes for AI Data Engine

Connect your data compute nodes to the host network and cluster network switches to enable AI workload processing and integration with your AFX 1K storage system. This procedure uses 100GbE connections for both host network access and cluster communication, allowing the nodes to leverage the existing cluster infrastructure without powering down the AFX system.

About this task

These procedures show common configurations. The specific cabling depends on the components ordered for your storage system. For comprehensive configuration details and slot priorities, see NetApp Hardware

Universe.

> You do not need to power off the AFX 1K storage system when cabling the data compute nodes. You can add the data compute nodes to an existing AFX 1K storage system that is already powered on and configured.

Before you begin

- You have an existing AFX 1K storage system installed. For information on installing the AFX 1K storage system, refer to AFX 1K storage system installation documentation.
- You have the required network switches installed and configured. Contact your network administrator for information about connecting the system to your network switches.
- You have reviewed the cabling requirements for the data compute nodes.

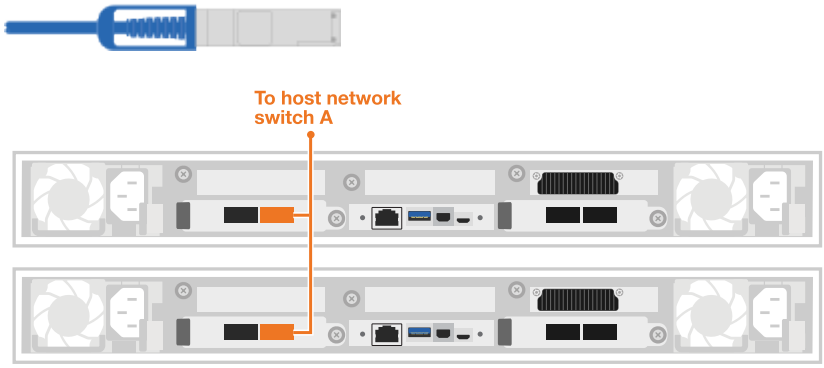
Step 1: Connect the data compute nodes to the host network

You can connect the data compute node ports to your host network.

Steps

1. Connect port e4b from the following data compute nodes to Ethernet data network switch A:
   - Data Compute Node 1, port e4b
   - Data Compute Node 2, port e4b

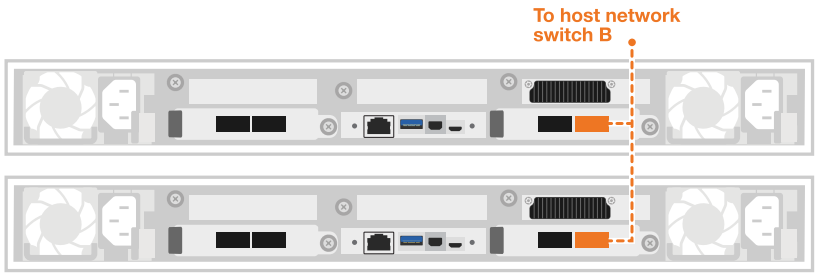
   100GbE cables

2. Connect port e5b from the following data compute nodes to Ethernet data network switch B:
   - Data Compute Node 1, port e5b
   - Data Compute Node 2, port e5b

   100GbE cables



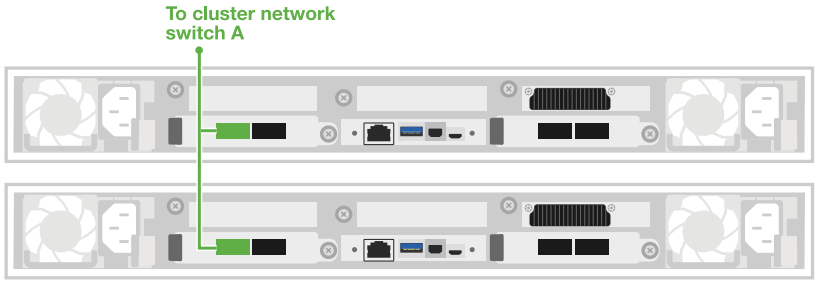
Step 2: Cable the data compute node cluster connections

For data compute nodes, the e4a/e5a ports are used for the cluster connections.

Steps

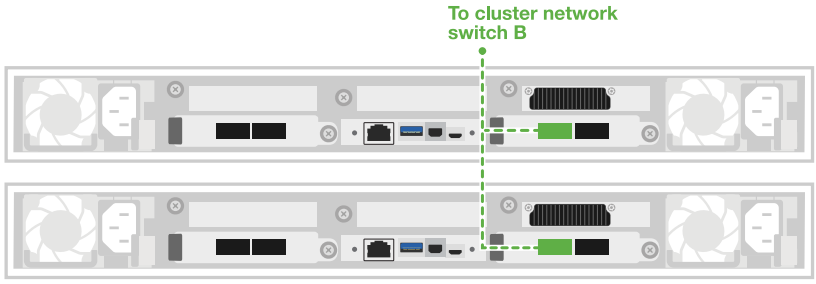
1. Connect port e4a from the following data compute nodes to a non-ISL port on cluster network switch A:
   - Data Compute Node 1, port e4a
   - Data Compute Node 2, port e4a

   100GbE cables



2. Connect port e5a from the following data compute nodes to a non-ISL port on cluster network switch B:
   - Data Compute Node 1, port e5a
   - Data Compute Node 2, port e5a

   100GbE cables



What's next?

After you've cabled the hardware, power on your data compute nodes.

Power on your data compute nodes for AI Data Engine

After you install the rack hardware and cable your data compute nodes, you should power on your DCNs and your controller nodes for the AFX storage system if not already powered on.

Before you begin

- Make sure that your shelves are powered on and each assigned a unique shelf ID. For information on assigning shelf IDs for the AFX storage system, see the documentation about assigning unique shelf IDs.

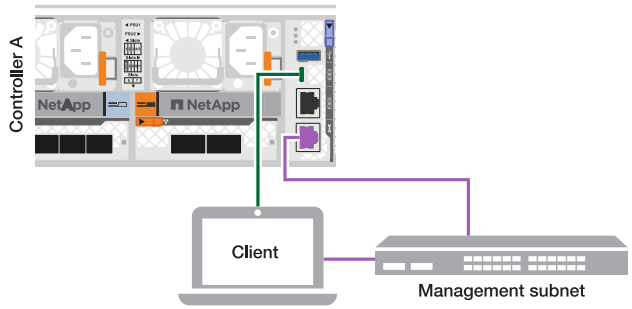
Steps

After you've turned on your storage shelves and assigned the unique IDs, power on your DCNs and power on the storage controller nodes if they are not already powered on.

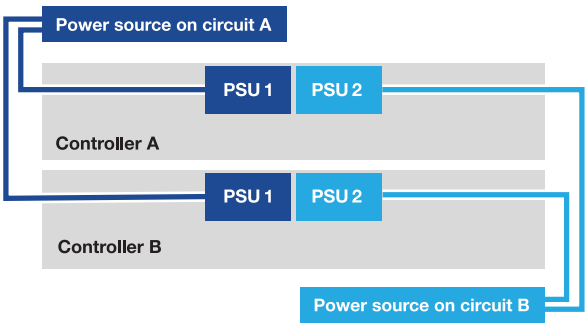
1. Connect your laptop to the serial console port. This allows you to monitor the boot sequence when the controllers are powered on.
   a. Set the serial console port on the laptop to 115,200 baud with N-8-1.

      See your laptop's online help for instructions on how to configure the serial console port.

   b. Connect the console cable to the laptop, and connect the serial console port on the controller using the console cable that came with your storage system.
   c. Connect the laptop to the switch on the management subnet.



2. Assign a TCP/IP address to the laptop, using one that is on the management subnet.
3. Plug the power cords into the controller power supplies, and then connect them to power sources on different circuits.



   - The system begins to boot. Initial booting may take up to eight minutes.
   - The LEDs flash on and the fans start, which indicates that the controllers are powering on.
   - The fans may be noisy at start-up, which is normal.

4. Plug the power cords into the data compute node power supplies, and then connect them to power sources on different circuits.
5. Secure the power cords using the securing device on each power supply.
6. Power on the data compute nodes.

   You might have to remove the bezel to access the power switch; if so, remember to reinstall it afterwards.

What's next?

After you've turned on your data compute nodes, set up an ONTAP AIDE cluster.

Set up your AIDE system

Set up AI Data Engine

As an ONTAP storage administrator, you can integrate AI Data Engine (AIDE) with your AFX system deployment. You can either set up AIDE as part of an initial AFX storage system deployment or add AIDE to an AFX cluster that's already been deployed and serving data. In both cases, you'll also need to complete the AIDE configuration after finishing the basic setup.

1. Set up AIDE and integrate it with an AFX storage system

There are two ways to set up AIDE and integrate it with an AFX storage system. Select the option that's appropriate for your environment.

> Before setting up AIDE, make sure you have the data compute nodes installed and connected to the cluster switches. See Install AIDE for more information.

Set up AIDE with a new AFX cluster

If you're setting up AIDE as part of an initial AFX system deployment, the AIDE setup process is part of the standard AFX cluster setup workflow. During the ONTAP cluster setup, System Manager automatically discovers the connected DCNs and includes them as part of the cluster configuration. After that, you're ready to complete the AIDE configuration, including the required license configuration.

Set up AIDE with an existing AFX cluster

If you're adding a cluster of data compute nodes to an existing AFX storage system, you need to configure the DCNs and finish setting up your AIDE cluster.

Add the data compute nodes to the cluster

ONTAP dynamically detects the new data compute nodes connected to your cluster network and displays them in System Manager. You can add data compute nodes to your AIDE cluster. Also note the following:

- The data compute nodes must be physically connected to the cluster switches. Exactly three DCNs are required when creating an AIDE cluster.
- You must have sufficient IP addresses for the DCN network and other required network configuration information.

Finish setting up your AIDE cluster

After integrating the data compute nodes and completing the initial network configuration, you can finish setting up your new AIDE cluster.

About this task

A welcome setup guidance window is displayed on the right side of the page with the necessary configuration steps. A checkmark is displayed to the left of each step or action item that has already been completed. An arrow (→) is displayed for items that have not yet been completed. For each of the incomplete actions:

- Select the enabled title link and complete the setup action for your environment.
- If the link is disabled, hover over the title to display the action required to enable it and continue with the setup action.

Each step has a link to additional documentation where possible. Some of the configuration might be included as part of the AFX cluster setup. See the AFX documentation for more information.

Steps

1. Select **Configure link aggregation groups and VLANs**.

   If required for your environment, you can set up cluster-level LAGs and VLANs. Selecting this option launches the Ethernet ports page.

2. Select **Configure network protocols**.

   You can configure the storage VM (SVM) network protocols.

3. Select **Update data compute software**.

   If a data compute software image is available for update, this option will be enabled. Select it to launch the data compute software update page. For more information, refer to Update AI Data Engine software.

4. Select **Configure data engine networking**.

   This option displays the page to configure the data engine network address.

5. Select **Create intercluster network interfaces**

   This option launches the page to configure the intercluster network interfaces as part of a two-step procedure if you intend to catalog remote data.

6. Select **Peer with other ONTAP storage systems**.

   After creating the intercluster network interfaces, you can establish peering relationships with other ONTAP storage systems.

7. Select **Add a data container**.

   You can add a volume for use by AIDE and associate it with a specific SVM.

8. Select **Add Workspaces**.

   This launches the Data Engine workspaces page where you can create a workspace.

9. Select **Configure OpenID Connect**.

You'll need to configure OpenID authentication to enable access to the AIDE console.

After you finish

After all the action items are completed, the guidance window is hidden. You can manually close and reopen it as needed.

2. Complete AIDE configuration

After your AIDE cluster is set up and integrated with the AFX storage system, you need to complete the AIDE configuration.

Steps

1. Add required AIDE licenses.

   There is an AI Data Engine license you must install for full AIDE functionality, including vectorization and guardrails capabilities.

2. Configure OIDC authentication.

   Make sure to configure OIDC for all deployments. You must configure OIDC to access the AI Data Engine Console.

Install an AIDE license in your AFX system

You'll need to install a NetApp license to access the full features of the AI Data Engine (AIDE). As an ONTAP cluster administrator, you can perform license management using ONTAP System Manager.

Prepare to license the AI Data Engine

A license is a record of one or more software entitlements. All the AFX licenses are delivered as a NetApp license file (NLF), which is a single file that enables multiple features. There are several things you should consider before installing a license on your AFX storage system in support of AIDE.

Types of licenses

There are two primary types of licenses needed to begin using AIDE through your AFX storage system.

ONTAP One license

The Metadata Engine Basic license is typically factory-installed as part of the ONTAP One license. It enables access to the Metadata Engine functionality which provides the essential foundation for AIDE operations. All the core capabilities needed to administer your ONTAP system are also included.

AI Data Engine

You need to purchase and install the AI Data Engine license to access the premium services needed to activate the full capabilities of AIDE. The license unlocks your data compute nodes, enabling AI features such as vectorization, governance guardrails, inferencing, and an integrated UI experience. The license includes a GPU count as well as an expiration date.

License installation requirements

You'll need to purchase the AIDE license and download the associated NLF file to your local system. You can

then upload the file to your AFX storage system through System Manager. Also make sure you have the following:

- Administrator credentials to sign in to ONTAP System Manager
- An AFX cluster running ONTAP 9.18.1 and later

Install a license on your AFX system

You can install an AIDE license to activate the additional AIDE features needed for your AFX storage system.

Steps

1. In System Manager, select **Cluster** and then **Settings**.
2. Next to **Licenses**, select →.
3. Select the **Features** tab to display the available ONTAP features.
4. To install a license, select the **Installed licenses** tab.
5. Select + Add.
6. Select a local license file and select **Add**.

Related information

- ONTAP licensing overview
- How to download NLF licenses from the NetApp Support Site
- ONTAP CLI: system license add

Configure OpenID Connect for AIDE in ONTAP

As an ONTAP cluster administrator, you can use ONTAP System Manager to configure OpenID Connect (OIDC) authentication for an AI Data Engine (AIDE) cluster. This provides a secure and centralized login through an external identity provider (IdP).

> You must configure OIDC to access the AI Data Engine Console. When configured, all authentication flows through OIDC. If OIDC is not configured, the console will be unavailable to administrators as well as the data engineers and data scientists. In this case, signing in to System Manager reverts to local authentication.

Also note the following about OIDC configuration for AIDE access:

- You cannot modify an existing OIDC configuration. If you need to make a change, first delete the configuration and create a new one with the desired settings.
- If you disable or remove OIDC, System Manager will revert to local ONTAP user authentication.

OIDC overview

OpenID Connect (OIDC) is an authentication protocol built on the OAuth 2.0 framework. It extends OAuth 2.0, which is used primarily for authorization, by adding an identity layer. OIDC introduces the concept of an ID token, which is a JSON Web Token (JWT) containing claims about the authentication event and the identity of the user.

You need to select and configure an external identity provider (IdP) supported by AFX with AIDE. The IdP

authenticates users and issues tokens that AFX, through System Manager, can use to grant access to the AIDE Console.

Configure third-party identity providers

To authenticate using OIDC, you need to first configure an external IdP. The ONTAP implementation of OIDC uses role claims in the tokens to enforce RBAC. When setting up an IdP, make sure it's configured to return role claims in the id token and access token. ONTAP supports two IdPs for OIDC authentication: Entra ID and Active Directory Federation Services (AD FS).

Entra ID

You can configure Entra ID using the following high-level steps:

1. Create a new App Registration at the Entra ID configuration page.
2. Set the Redirect URI (Web) value to `https://$CLUSTER_MGMT_IP/oidc/callback`, substituting the appropriate cluster management IP address or FQDN.
3. Create the required roles under App Roles and assign them to your users.
4. Update the token claims to under Token Configuration to return roles in id-token and access-token.

See Set up an OpenID Connect provider with Entra ID for more information.

Active Directory Federation Services

You can configure AD FS using the following high-level steps:

1. Create new Application Group and select **Server application accessing a web API**.
2. Set the Redirect URI (Web) value to `https://$CLUSTER_MGMT_IP/oidc/callback`, substituting the appropriate cluster management IP address or FQDN.
3. Configure claims to return roles in tokens.

See Add AD FS as an OpenID Connect identity provider for more information.

Configure OIDC in System Manager

After configuring your IdP, you can set up OIDC authentication in System Manager to enable secure access to the AIDE Console.

Before you begin

- You need to have administrator access to System Manager.
- Your OIDC identity provider must be configured and accessible.

Steps

1. In System Manager, select **Cluster** and then **Settings**; locate the OpenID Connect card.
2. If OIDC is already configured, you can edit or disable the configuration. If OIDC is not configured select  to start the setup process.
3. Under Configure OpenID Connect, provide values for the following fields:
    - Provider
    - Issuer

- JSON web key set URI
- Authorization endpoint
- Token endpoint
- End session endpoint
- Access Token Issuer (optional)

4. Under Client configuration, provide values for the following fields:
   - Client ID
   - Remote user claim
   - Refresh interval
5. Under Connection details, provide values for the following fields:
   - Cluster IP address or FQDN
   - Outgoing proxy (optional)
6. Under External Role mapping, select an existing role mapping or define a new role for the ONTAP `admin` user.
7. Select **Enable now** and then **Save**. System Manager will refresh to apply the new authentication settings.
8. Log in with your IdP credentials; after successful authentication you'll be returned to System Manager.

Related information

- OpenID Foundation
- ONTAP OAuth 2.0 implementation

Set up workspaces

Prepare data for AI Data Engine

After you create a cluster, establish a data container that contains data you intend to use with AI Data Engine (AIDE). This data container must be an ONTAP volume, either a local volume or a volume from a peered ONTAP cluster running ONTAP 9.

Rather than manually upload ONTAP cluster data to AIDE, you'll need to peer targeted clusters and SVMs with the AIDE cluster and then decide which NFS volumes you want to use with the AIDE metadata catalog. After you've created the data container, you can create a workspace and associate the data container with that workspace. Users of that workspace can then access and interact with the workspace affiliated data collections and resources for their AI workloads.

About this task

You must peer each SVM that contains data you want to use with AI Data Engine. Peering only the cluster is not sufficient. This ensures the AI Data Engine can access, onboard, and index your data as intended.

You should peer the SVM in the AIDE cluster that will act as the SnapMirror destination.

You don't need to create SnapMirror relationships between the ONTAP clusters and the AIDE cluster. These relationships will be created during workspace creation automatically.

Before you begin

- You need *storage administrator* privileges to peer clusters and SVMs and select data containers.
- You have identified the ONTAP clusters and SVMs that contain the data you want to use with AI Data Engine.
- You've confirmed that data source volumes have met the following requirements:
  - Volumes are online and accessible.
  - NFS protocol is enabled. Only NFS-enabled volumes are supported as data containers for AI Data Engine. SMB and CIFS volumes are not supported.
  - Volumes are not FlexCache volumes.
  - Data sources are read-write volumes. Data protection volumes are not supported.

Steps

1. Peer each ONTAP cluster and SVM that contains data you want to use with AI Data Engine.
2. Select the volumes you want to use with AI Data Engine.

   For each of volume, note the following information:

   - Volume name
   - UUID
   - SVM name and UUID
   - Cluster name and UUID

What's next?

Create a workspace and associate the data containers you created with a workspace.

Related information

- Data migration options

Create a workspace in AI Data Engine

After you set up a cluster, you can create a workspace. Workspaces allow you to segment data on the cluster, control data access for individuals, and exclude data that AI Data Engine (AIDE) should not access.

If you administer storage, you'll use ONTAP System Manager to create and manage workspaces.

Organizations create workspaces based on teams, projects, data sensitivity levels, or other relevant criteria. For example, if you work in healthcare, you might segment clinical data into a workspace but leave out data pertaining to IT, legal, or other departments.

About this task

System processing limits affect workspace creation (typically up to 15 GB per day per cluster). If you create multiple workspaces in parallel or in quick succession, each workspace might take longer to process, and you might experience significant delays.

Monitor the status of workspace creation from the Workspaces inventory page. For best results, avoid creating many workspaces at once if you need immediate access to these features.

Before you begin

- You need *storage administrator* privileges to create workspaces and associate data collections.
- You've determined the remote (peered) and local data sources you intend to use with the workspace and with AI Data Engine.
- You've created at least one data container that the workspace can use, such as a local volume or a volume from a peered cluster.

> Add a volume to a workspace that you won't delete during the expected lifetime of that workspace. If you delete a volume after adding it to a workspace, the workspace will enter a failed state. Confirm the longer-term viability of the volume before establishing a workspace.

- Ensure NFS is enabled on the volume but that CIFS is not enabled. Workspaces only support volumes with NFS. Volumes with CIFS (SMB) are not supported.

Create a workspace

Create a workspace and associate data containers that contain the data you want to use with AI Data Engine.

Steps

1. In ONTAP System Manager, navigate to **Data Engine > Workspaces**.
2. Select **Add**.
3. In the **Add Workspace** dialog, select at least one available data container to associate with the workspace.
4. Configure peered clusters so that the data from those clusters can be accessed within the workspace
5. If you'd like to configure user access to the workspace, you can do that now or wait until after the workspace is created.

6. Configure a refresh interval for how often the workspace synchronizes with the associated data containers to capture new or updated data (for example, six hours).

Choose an interval that balances data freshness with system performance. If you add a data container to multiple workspaces, the system automatically uses the most aggressive (shortest) interval. To learn more, see documentation about workspace refreshes and versioning.

7. Select **Continue**.
8. In the **Finalize workspace** dialog, enter a workspace name and description.
9. Select **Add** to create the workspace.

Result

The workspace creation process takes several minutes to hours to complete, depending on the associated dataset and its file count, file size, and other factors.

The system automatically extracts metadata for all data sources and stores it in a metadata catalog that users can use to locate the files they need for their projects. After you assign users to the workspace, data engineer users can set up and interact with workspace-affiliated components from AI Data Engine Console.

The new workspace appears on the Workspaces page in `Creating` state until the process completes and the state changes to `ready`.

Review workspace details

After workspace creation, review the workspace details.

Steps

1. Review workspace details, including total size, percentage of cluster capacity used, and the date of most recent workspace refresh.
2. Select the workspace name to open the details page.
3. In the Overview tab, view workspace details that include associated data containers, users, and activity.

Workspace refreshes and versioning

Each workspace refresh creates an immutable version that captures the current state of all files and objects in the workspace. Versions include complete metadata, references to snapshots used during extraction, and a job ID for traceability. This supports data lineage, reproducibility, and auditing.

Refreshes occur either on the schedule you configure (such as every six hours) or when you trigger them manually. The minimum supported refresh interval is one hour; the maximum is one year. If a data container is included in multiple workspaces, the system uses the most frequent, shortest duration refresh interval for scheduling metadata extraction.

By default, the system retains previous, current, and next (in-progress) versions. The system retains older versions according to your organization’s policy and can purge them as needed.

You can list all versions of a workspace and view differences between versions to identify which files or objects were added, modified, or deleted. This allows you to track changes over time and understand the evolution of your workspace data.

Assign user access to AI Data Engine workspaces

As a storage administrator, you assign users to a workspace based on their roles: data engineers, data scientists, or other roles depending on your organization's structure and needs. Users log in to the AI Data Engine (AIDE) Console using their credentials and access the data container resources within the assigned workspace.

ONTAP System Manager enables you to manage which users have access to AI Data Engine workspaces. You can add or remove users to control who can view, modify, and interact with workspace data and activities.

Before you begin

- You need *storage administrator* privileges to manage user access to workspaces.
- Confirm that you created the workspace and that it is active in the Workspaces inventory.
- Confirm that all pertinent data containers have been added to the workspace and are accessible.
- Confirm OIDC is enabled and configured for the cluster. Role mapping from IdP to ONTAP roles must be completed for each relevant data engineer and data scientist IdP user or group.

Steps

- Add a user to a workspace:
  a. In ONTAP System Manager, navigate to **Data Engine > Workspaces**.
  b. Select the workspace name to open its details page.
  c. Go to the Users tab.
  d. Select the **Add** button to open the add users dialog.
  e. Enter the details of one or more users. Enter the details as a comma-separated list of OIDC users.
  f. Select **Add** to grant the user access to the workspace.
- Remove a user from a workspace:
  a. In the Users tab of the workspace details page, locate and select the user you want to remove.
  b. Select the **Remove** button.
  c. Confirm the removal in the dialog.
  d. The system immediately removes the user and revokes their access to the workspace.

Result

Only users listed in the workspace's Users tab can access and interact with workspace data and activities.

Administer and monitor

Monitor cluster processes

View the AIDE system and cluster status

As a storage administrator, you can use ONTAP System Manager to access the dashboard and display the cluster status. This is a good first step before beginning your AIDE administrative tasks or if you suspect an operational issue.

Before you begin

- You need *storage administrator* privileges to perform AIDE ONTAP-related administrative tasks.

Monitor AIDE health and capacity from the dashboard

1. Connect to ONTAP System Manager using the cluster management address:

   ```
   https://$FQDN_OR_IP/
   ```

2. Sign in with an administrator account.
3. Select **Dashboard** in the left navigation pane.
4. Review the **Health** tile:
   - Confirm overall cluster health.
   - Verify the **Data compute nodes** count and status.
   - Check for alerts:
     - DCN node issues or connectivity problems
     - Workspaces or data collections in error (for example, collection publishing failures)
5. Review the **Capacity** tile:
   - Note total cluster capacity and used capacity.
   - For AIDE clusters, verify:
     - Capacity used by AIDE metadata and application volumes (metadata Storage VM)
     - Capacity used by workspaces and data collections (if available)
6. Optionally review **Network** and **Performance** tiles to understand cluster-wide behavior that might impact AIDE workloads (for example, network congestion or protection lag).

View data DCN health and utilization

1. In the navigation pane, select **Cluster** and then **Overview**.
2. Select the **Data compute** tab.

   This tab shows all DCN nodes in the cluster with:

   - Node name, model, serial, and software version
   - Overall node state

- CPU and memory utilization
- GPU utilization (if GPUs are present)
- Any node-level error indicators

3. Expand a DCN node to open the detailed view and check:
   - System CPU and memory usage
   - GPU memory usage
   - Reported hardware or service issues
4. Select **Cabling** on the **Cluster > Overview** page to verify that DCN nodes are correctly cabled to the cluster switches and to identify any port or link issues.

Monitor workspaces and metadata footprint

1. In the navigation pane, select **Data engine** and then **Workspaces**.
2. Review the workspace summary at the top of the page:
   - Count of workspaces and their states (for example, `Processing`, `Healthy`, `Error`).
   - Total workspace size.
   - Percentage of cluster capacity consumed by all workspaces.
3. Review the workspace grid:
   - Confirm that critical workspaces show a **Healthy** state.
   - Check workspace sizes and capacity consumption.
   - Look for any workspaces in `Error` or long-running `Processing` states.
4. To review details for a specific workspace, select its name:
   - On the **Overview** tab, confirm:
     - Workspace state and size
     - Data containers (volumes) included and their item counts
     - Last updated time for each data source
   - On the **Data collections** tab, confirm:
     - Which data collections exist for that workspace (data collections are read-only in System Manager)
     - Their state, size, and last updated time
   - On the **Users** tab, check which AI Data Engine Console users have access.

Monitor metadata Storage VM and AIDE-managed protection

1. In the navigation pane, select **Cluster** and then **Storage VMs**.
2. Locate the Storage VM with subtype `data-engine` (the metadata SVM):
   - Confirm that the metadata SVM is online.
   - Optionally open its details to see counts for:
     - Volumes
     - LIFs with type `Data compute network` (used for DCN-ONTAP communication)

3. Select **Protection** and then **Relationships** to view protection for remote data sources used in workspaces:
   - Identify AIDE-created SnapMirror relationships by naming pattern:
     - Destination volume: `<source_volume_name>_dest_<source_volume_UUID>`
     - Policy: `<source_volume_name>_dest_aide_policy_<source_volume_UUID>`
   - Use this view to verify that relationships are healthy and that lag time aligns with workspace refresh expectations.

> Do not modify the metadata Storage VM, AIDE-created SnapMirror relationships, or AIDE-managed snapshots (or their schedules) directly in ONTAP. Changes can disrupt AIDE version history. Adjust workspace refresh settings if you need to adjust refresh behavior.

Review AIDE-related alerts and notifications

1. In the navigation pane, select **Events & Jobs** and then **System alerts**.
2. Review any active alerts related to:
   - DCN node health or connectivity
   - Data engine networking issues
   - Workspace or data collection errors
   - Software version mismatches between ONTAP and DCN cluster
3. As needed, configure notification destinations (for example, email, syslog) in **Cluster > Settings > Notification management** to ensure AIDE-related alerts are forwarded to your operations tooling.

Related information

Prepare to administer your AFX storage system

View Insights to optimize your AIDE system

As a storage administrator, you can use the *Insights* feature of ONTAP System Manager to display suggested configuration updates that align with NetApp best practices. These changes can optimize the security and performance of your AIDE cluster.

About this task

Each of the insights is presented as a separate tile or card on the page that you can choose to implement or dismiss. You can also select the associated documentation link to learn more about a specific technology.

Steps

1. In System Manager, select **Analysis** and then **Insights**.
2. Review the available recommendations.

What's next

Perform any of the recommended actions to implement configuration best practices.

View AIDE system events, jobs, and audit log

As a storage administrator, you can review the events, jobs, and audit log messages generated by AIDE to track internal processing and diagnose potential problems. The

AIDE system can be configured to forward this information, along with other related data, for additional processing and archival.

Before you begin

- You need *storage administrator* privileges to perform AIDE ONTAP-related administrative tasks.

Monitor AIDE activities, events, and jobs

You can use the centralized **Activity** view to monitor AIDE-specific events and jobs across all workspaces, or review activity scoped to individual workspaces.

View cluster-wide AIDE activity

Monitor workspace operations, troubleshoot metadata extraction issues, and track data collection publishing across your entire AIDE deployment.

1. In ONTAP System Manager, in the navigation pane select **Data engine** and then **Activity**.
2. Select the **Events** tab:
   - Review recent AIDE-specific events such as the following:
     - Workspace creation, update, or deletion
     - Data container add/remove operations
     - Data collection publishing (if present)
   - Use filters (by severity, object type, workspace, or time range) to focus on active or critical events.
3. Select an individual event to open and review:
   - Description and timestamp
   - Affected workspace, data collection, or data source
   - Recommended action, if provided
4. Select the **Jobs** tab:
   - Monitor long-running jobs such as:
     - Initial metadata extraction for a workspace
     - Workspace refresh / catalog update jobs
     - Data collection publishing or refresh jobs
   - Check job status and progress.
5. Select a job to open the peek view and review:
   - Start and end time
   - Progress percentage and phase (for example, **Scanning**, **Publishing**)
   - Affected workspace, data collection, or data source
   - Error messages for failed jobs

View workspace-specific activity

To troubleshoot a specific workspace, open the workspace details (**Data engine > Workspaces**, select a workspace) and use the **Activity** tab:

- Review events and jobs scoped to that workspace only.
- Use this view to isolate issues such as a single workspace stuck in `Processing` state.

View cluster-wide events

Review event messages for a valuable record of system activity. Each event includes a description and unique identifier along with a recommended action.

1. In ONTAP System Manager, select **Events & jobs** and then **Events**.
2. Review and respond to the recommended actions at the top of the page, such as enabling automatic update.
3. Select the **Events log** tab to display a list of the messages.
4. Select an event message to examine it in more detail, including the sequence number, description, event, and recommended action.
5. Optionally select the **Active IQ suggestions** tab and register with Active IQ to get detailed risk information for the cluster.

View cluster-wide jobs

View all jobs running on the AIDE cluster, including AIDE-specific jobs and general ONTAP jobs.

1. In ONTAP System Manager, select **Events & Jobs** and then **Jobs**.
2. Customize the display as well as search and download the job information as needed.

View audit log

Use the audit log to review a record of system activity based on the use of access protocols such as HTTP.

1. In ONTAP System Manager, select **Events & jobs** and then **Audit logs**.
2. Select **Settings** to enable or disable the operations that are tracked.

Manage notifications

Configure notification destinations to automatically forward AIDE events and audit logs.

Steps

1. In ONTAP System Manager, select **Cluster** and then **Settings**.
2. Navigate to **Notification management** and select ⋮.
3. Select the appropriate action to view or configure the destinations used by AIDE:
   a. Event destinations: Select **View event destinations**
   b. Audit log destinations: Select **View audit destinations**
4. Select **Add** as appropriate and provide the destination information.
5. Select **Save**.

Related information

- ONTAP Event, performance, and health monitoring

Manage AI Data Engine workspaces

A workspace is a set of data sources (volumes) that the AI Data Engine (AIDE) uses to build and refresh a metadata catalog for a particular project or use case. As a storage administrator, you can use ONTAP System Manager to monitor workspace health, adjust configuration, control data sources, manage users, and delete workspaces when they are no longer needed.

Before you begin

- You need *storage administrator* privileges to manage workspaces.

Review workspace status

Review workspace health, capacity usage, and metadata status to ensure the Metadata Engine is operating as expected and not consuming unexpected resources.

Steps

1. From ONTAP System Manager, in the navigation pane select **Data engine > Workspaces**.
2. Review the summary at the top of the page for total workspaces, overall workspace health, and capacity usage.
3. For workspace-specific information, select a workspace name. On the **Overview** tab confirm:
   - Workspace state and size.
   - Data containers (volumes) included in the workspace.
   - Item counts and last updated time per data source.
   - Any workspace-level warnings.

   If a workspace or data collection shows an error state, verify that all source volumes are online and accessible.
4. Select the **Data collections** tab to see:
   - All data collections associated with this workspace.
   - State (such as `Published` or `Error`), size, and last updated time.

   System Manager is read-only for data collections. Data engineers can create and manage data collections in the AI Data Engine Console.
5. Select the **Users** tab to view:
   - The list of users with access to this workspace.
6. Select the **Activity** tab to see events and jobs related to this workspace only.

Edit workspace properties and refresh schedule

You can adjust a workspace's name, description, refresh interval, and (if licensed) its guardrail policy.

Steps

1. From **Data engine > Workspaces**, select ⋮ next to a workspace and select **Edit**.

2. Edit workspace properties:
   - Update **Name** and **Description** as needed.
   - Adjust the **Refresh interval** (metadata update frequency) within the allowed range (hours and days).
   - If an AIDE license is installed, you can select **Guardrail policy**.
3. Select **Save**.

Changes to the refresh interval or metadata processing might affect how often remote SnapMirror relationships are updated for this workspace.

Add data containers to an existing workspace

You can add additional mounted volumes (local or from a peered remote cluster) so that their metadata is included in the workspace catalog.

1. From **Data engine > Workspaces**, do one of the following:
   - Select ⋮ next to the workspace and select **Add data containers**.
   - Open the workspace, select the **Overview** tab, and then select **Add** in the data containers section.
2. In the **Add data containers to workspace** dialog:
   - Locate local volumes on the AIDE cluster.
   - Expand peered clusters to select remote volumes (remote volumes require cluster and SVM peering).

   Only eligible, online volumes that are not globally excluded and not already part of the workspace can be selected.
3. If you are prompted for remote volume mapping:
   - Select the target Storage VM on the AIDE cluster to receive SnapMirror destinations for the chosen remote volumes.
4. Select **Add**.
5. Use the workspace **Activity** tab or **Data engine > Activity** to track metadata extraction and any SnapMirror initialization for new data sources.

Remove data containers from a workspace

You might remove a data container when it is no longer relevant to the workspace's purpose or if you want to reduce the scope of metadata management for that workspace. Removing a data container stops metadata refresh for that volume and removes its metadata from the metadata catalog.

Do not delete a source volume from ONTAP that has been added to a workspace. If you delete the volume, the workspace will enter a failed state. Always remove the data container from the workspace first before deleting any underlying ONTAP volumes.

Steps

1. Navigate to **Data engine > Workspaces** and select the workspace that contains the data container.
2. On the **Overview** tab, locate the data container to remove.
3. Select **Remove** next to the data container.

4. Review the confirmation dialog and select **Remove**.

> Removing a data container from a workspace does not delete the underlying ONTAP volume or its SnapMirror relationship. It only affects metadata usage within the AI Data Engine.

Manage workspace users

You can grant or revoke access for data engineer and data scientist users to a workspace. These users are defined in your identity provider (OIDC) and mapped to ONTAP roles. See assign users to workspaces documentation to learn how to manage user access.

Delete a workspace

You can delete a workspace to remove the workspace definition and associated AIDE metadata. Any data collections and vector embeddings tied to the workspace are also removed.

> Underlying ONTAP data (volumes, SnapMirror relationships) are not deleted.

Steps

1. In **Data engine > Workspaces**, do one of the following:
   - Delete a single workspace, select ⋮ and select **Delete**.
   - Delete multiple workspaces, select the checkboxes for the workspaces, and then select **Delete**.
2. In the confirmation dialog, review the impacts of the action before proceeding:
   - Workspace metadata is permanently deleted.
   - Data collections and embeddings associated with the workspace are permanently deleted.

   > There is no soft-delete or restore option.
3. Select the checkbox to confirm your understanding, and select **Delete**.

Related information

- Assign users to workspaces

Upgrade and maintain your AIDE system

AI Data Engine system updates and compatibility

Keep AI Data Engine (AIDE) system components up to date to maintain optimal performance and access to new features. Update components after deployment, when new software or firmware becomes available, when you add or replace nodes, or periodically for feature updates.

AIDE system components

There are two main components that require updates in an AIDE system: ONTAP software and AIDE software, which includes DCN firmware.

ONTAP Software

ONTAP is the operating system that runs on NetApp storage systems, including those used in AIDE deployments. Keep ONTAP current to preserve system stability, security, and compatibility with AIDE components. AIDE components are updated separately.

AIDE software updates

AIDE Console software and DCN firmware updates are distributed together as a single package (.tgz) and are not embedded in ONTAP images. Updates ensure the proper functioning of hardware components within the AIDE system and provide new features, performance improvements, and bug fixes.

Understand the update process

AIDE software updates can be managed through ONTAP System Manager.

AIDE does not support ONTAP automatic software updates functionality. You can sign up to receive notifications from NetApp Support Site downloads, but all updates of AIDE software are performed manually by the admin.

Release types and scope:

- ONTAP major releases (9.x.x) and AIDE major releases (9.x.x U0) introduce new features, APIs, or changes that affect ONTAP integration.
- ONTAP patch releases (9.x.x Px) and AIDE update releases (9.x.x Ux) contain fixes and updates that do not impact ONTAP integration.

Compatibility matrix

Ensure compatibility between ONTAP and AIDE software when planning updates.

AIDE software is released as "U" releases. An AIDE major release is a "U0" release and subsequent minor releases will be "U1" and later.

ONTAP and AIDE compatibility

AIDE Releases	Supported ONTAP Versions
9.18.1 U0	9.18.1 GA and all 9.18.1 Px
9.18.1 U1 and later	9.18.1 GA and all 9.18.1 Px

"Px" denotes all ONTAP patch releases in major version (for example, 9.18.1 P1, 9.18.2, and others).

AIDE upgrade paths

The following examples use hypothetical future versions to show permitted upgrade and update paths permitted from AIDE 9.18.1 U0 and 9.18.1 U1.

If your current AIDE release is…	And your target AIDE release is…	Your upgrade or update path is…
9.18.1 U0	9.18.1 U1	Direct

If your current AIDE release is…	And your target AIDE release is…	Your upgrade or update path is…
9.18.1 U0	9.18.1 U3	Direct (you can update from any Ux to any later Ux in 9.18.1)
9.18.1 U1	9.18.1 U3	Direct (you can update from any Ux to any later Ux in 9.18.1)

Revert limitations

AIDE systems do not support revert operations for DCN firmware, AIDE software updates, or ONTAP on AFX storage systems. After you install an update or upgrade, you cannot revert to a previous version. Review compatibility requirements before you update or upgrade.

Related information

- Upgrade ONTAP software for AFX systems
- Update AIDE software

Update AI Data Engine software

As a storage administrator, you can update AI Data Engine (AIDE) software, data compute node (DCN) firmware, and other system files on your AIDE system using ONTAP System Manager.

AIDE does not support ONTAP automatic software updates functionality. You can sign up to receive notifications from NetApp Support Site downloads, but all updates of AIDE software are performed manually by the admin.

About this task

The combined AI Data Engine software package is significantly larger than a typical ONTAP update package (AIDE package is approximately 40GB). Plan for longer upload and installation times when updating AIDE software.

Before you begin

- You need *storage administrator* privileges to update DCN firmware and AI Data Engine software.
- You need your NetApp Support Site credentials for your active account.
- Ensure compatibility between ONTAP, DCN firmware, and AI Data Engine software when planning updates.

> Revert functionality is not supported for DCN firmware or AI Data Engine software updates. After an update is installed, you cannot revert to a previous version.

Steps

1. Download the combined DCN firmware and AIDE software update file to a local client.
2. In System Manager, select **Cluster > Settings > Software updates**.
3. Next to **Software updates**, select →.
4. Under AI Data Engine updates, select **Add AI Data Engine software files** and select the update package.
5. After package upload is complete, select **Update** to begin installing updates on the DCN nodes.

AIDE Console is not available or accessible when the DCN node update is in progress.

Result

DCNs are updated with AI Data Engine software and the updated version is displayed for each node.

Related information

- Upgrade ONTAP software for AFX systems

Add data compute nodes to your AIDE cluster

You can add data compute nodes (DCNs) as part of creating a new AI Data Engine (AIDE) cluster or expanding an existing cluster. The workflow consists of discovering and configuring the nodes using ONTAP System Manager.

Prepare to add nodes

There are several considerations when adding DCN nodes.

When creating a new AIDE cluster

Exactly three DCN nodes are required for a new AIDE cluster.

Hardware installation and addressability

Make sure the following prerequisites are met:

- The new DCN hardware is racked, powered on, and cabled to the cluster switches.
- You have an IP address space range available for the DCN to ONTAP backend subnet.
- The ONTAP cluster has been initialized and is reachable from the cluster management LIF.

System Manager credentials

You need *storage administrator* privileges to perform AIDE cluster creating or expansion tasks.

Software compatibility

Review the following documentation to confirm that your DCN hardware and software versions are compatible with your ONTAP cluster:

- DCN software is compatible with the ONTAP version.

During the node add operation, System Manager will confirm the new nodes run a software version compatible with:

- The ONTAP cluster effective version (ECV) if this is the first DCN join.
- The existing DCN cluster version if DCNs are already present.

If a node is incompatible:

- An error is shown next to the affected DCN in the **Add** dialog.
- You must first update the DCN software (or ONTAP, as appropriate) to a compatible version.

Add the data compute nodes

You add DCN nodes as part of creating a new AIDE cluster or expanding an existing cluster.

Steps

1. In System Manager, select **Dashboard** in the navigation pane and then the **Health** card.
2. Confirm that there are nodes to add and select **View details** to display the list.

   The list contains discovered nodes that are not yet part of the AIDE cluster

3. Alternatively, you can select **Cluster** and **Overview** and the **Data compute** tab to see the list.
4. At the bottom of the data compute page, select **Add** above the list of nodes .
5. In the **Add data compute nodes** dialog, select the DCN nodes you want to add.

   You can optionally rename individual nodes before adding them.

6. If this is the first time you're adding nodes and no backend subnet exists, select **Add subnet** and provide:
   - Subnet name (for internal use)
   - Subnet address and mask
   - IP address range for DCN and ONTAP nodes on this backend network

     System Manager validates the range includes enough free IP addresses for all DCNs being added and all ONTAP nodes in the cluster and additional cluster-level floating IPs used for DCN to ONTAP communication.

7. Whether you added the backend subnet or it already exists:
   a. Review the available IP addresses.
   b. If needed, select **Edit subnet** and extend the IP range.
      - You can only grow the range. Shrinking or changing the subnet is not supported.
      - Changing the subnet or IP range might require recreating the underlying Kubernetes cluster on DCNs and can take several minutes.
8. Optionally configure the Data Engine service interface by providing:
   - Service IP address
   - Network mask
   - Gateway (if required for your environment)

     The single IP will be load-balanced across DCNs and used as the frontend address for the AI Data Engine Console and related APIs.

9. Review the selected nodes, backend subnet, and Data Engine service interface settings.
10. Select **Add** and wait for the operation to complete. System Manager will perform the following actions:
    - Add the selected nodes to the DCN cluster
    - Provision the backend networking and join nodes to the Kubernetes-based DCN cluster
    - Updates internal metadata for DCN discovery
11. After completion, select **Cluster** and **Overview** and confirm:

a. Under **Data compute** the new DCNs appear as part of the cluster
b. All nodes are **Healthy**
c. Verify the dashboard **Health** card shows the updated node count

Replace a node in your AIDE cluster

You need to replace a data compute node (DCN) in your AI Data Engine (AIDE) cluster if it stops functioning or needs to be swapped due to hardware failure, upgrade, or maintenance. This ensures the AIDE cluster remains healthy and operational. The procedure can be performed without disrupting ongoing services.

Prepare to replace a node

There are several things to consider before replacing a node in your AIDE cluster.

System Manager credentials

You need *storage administrator* privileges to perform AIDE cluster node replacement tasks.

Restrictions

You should be aware of the following restrictions when replacing a node in your AIDE cluster:

- Node replacement is only supported using the CLI and optionally the REST API.
- You cannot perform node replacement using System Manager.
- The new node should match the cluster's software version; ONTAP will update it if needed.
- The failed node must not be powered on while connected to the cluster network to avoid IP address conflicts.

Requirements

You'll need the following:

- Serial number for the new replacement node

Replace a DCN node in your AIDE cluster

You can replace a DCN node in your AIDE cluster using the following procedure.

Steps

1. Physically remove the failed node

   Power off and disconnect the node from the cluster network. Make sure the node is not booted up on the network during the replacement process.

2. Delete the failed node from the cluster using the following command:

   ```
   dcn cluster node delete -name <node_name> -force true
   ```

   Provide the actual name for the <node_name> value.

3. Physically attach the new node to the cluster

Make sure the node is cabled in, powered on, and discoverable.

4. View the discoverable and unconfigured nodes to verify the new node is online:

   ```
   dcn cluster node show -membership available
   ```

5. Add the node to the cluster using the following command:

   ```
   dcn cluster node create -serial-number <new_node_serial>
   ```

   ONTAP will allocate an IP address for the new node. If the node's software version does not match the cluster, ONTAP will automatically update the node.

6. Verify cluster health and node integration with either of the following commands:

   ```
   dcn cluster node show
   ```

   ```
   dcn cluster node show -instance
   ```

Related information

- Expand your compute cluster

Manage vectorization and data collections

Data-to-RAG quick start for AI Data Engine

Go from a newly deployed AI Data Engine (AIDE) system to a working retrieval-augmented generation (RAG) endpoint using this workflow. Understand how storage administrators, data engineers, and data scientists collaborate using ONTAP System Manager and AIDE Console.

Before you begin

- You've installed and added Data compute nodes (DCNs) to the ONTAP cluster.
- You've installed and licensed AI Data Engine software for vectorization and guardrails.
- You've configured OpenID Connect (OIDC) and mapped roles for admin, data engineer, and data scientist roles.

Define data scope and governance

As a storage administrator or security administrator, you want to prepare the environment in AIDE Console and ONTAP System Manager:

- Create one or more workspaces from local and remote data sources.
- Configure classifiers and guardrail policies in AIDE Console.
- Assign data engineer and data scientist access to the workspaces.

Explore workspace metadata

As a data engineer or data scientist, you want to explore the workspace metadata using AIDE Console:

- Explore workspace metadata to understand available content.
- Define one or more logical subsets of data that should feed RAG (for example, support articles, product manuals, or anonymized clinical notes).

Create and publish a data collection

As a data engineer or data scientist, you want to turn the chosen subset into a RAG-ready collection:

- Create a data collection from the workspace using selected filters.
- Publish the data collection and monitor indexing until it reaches `Ready` state.
- Copy the retrieval endpoint URI for the chosen collection and provide to data scientists or application developers.
- View data collection status and vector footprint as needed.

What's next?

- Define your data estate and guardrail policies in AI Data Engine
- Explore workspace metadata in AI Data Engine Console

- Create data collections in AI Data Engine Console

Explore workspace metadata in AI Data Engine Console

As a data engineer or data scientist, your first task in AI Data Engine (AIDE) is to understand what data is available in your workspaces. You use AIDE Console to query the metadata catalog, search for relevant files, and identify subsets of data to turn into data collections.

Before you begin

- You need *data engineer* or *data scientist* privileges in AI Data Engine Console and access to at least one workspace.
- A storage administrator has:
  - Created one or more workspaces in ONTAP System Manager.
  - Assigned your user or group access to the relevant workspaces.
- Metadata extraction for the workspace has completed and the workspace is in `Ready` state.
- Classifiers are enabled, so that metadata includes classification tags (for example, PII indicators).

Sign in to AIDE Console as a data engineer or data scientist

Steps

1. In a browser, navigate to the AIDE Console URL:

   ```
   https://<cluster_management_ip>/console
   ```

2. Authenticate through your organization's OIDC provider.
3. Confirm that your role is recognized as a data engineer or data scientist (for example, by the available workspace and data collection actions). For more information, see AIDE role documentation to learn how data engineers and data scientists work with AIDE components.

Result

You are logged into AIDE Console and see only those workspaces to which you have been granted access.

View your accessible workspaces

Steps

1. In AIDE Console, navigate to **Data Curator > Workspaces**.
2. Review the list of workspaces you can access.
3. Select a workspace to open its details.

Result

You now have a workspace-scoped view of the data estate that storage administrators have made available for your projects.

What's next?

- Create data collections for RAG from a workspace

Create data collections in AI Data Engine Console

Data collections are the core RAG building blocks in AI Data Engine (AIDE). As a data engineer or data scientist, you define which files belong in a collection, configure embedding and indexing options, and publish the collection so that applications can query it through a retrieval endpoint.

You'll perform all data collection tasks in the AI Data Engine Console.

Before you begin

- You need *data engineer* or *data scientist* privileges in AI Data Engine Console (`https://<cluster_management_ip>/console`).
- You have access to at least one workspace with metadata extracted and in `Ready` state.
- You have explored the workspace metadata and identified queries or filters that define meaningful subsets of data.
- The AI Data Engine software license is installed and inferencing features are enabled.

Create a data collection from workspace metadata

Steps

1. Navigate to **Data Curator > Workspaces** and select the workspace that contains your target data.
2. Select **Add data collection**.
3. In the **Create new data collection** page, do the following:
   a. Enter a name and description for the collection (for example, `Support_KB_RAG_EN`).
   b. Choose whether the collection should be:
      - **Dynamic**: New files are automatically identified and added to the data collection based on the filtering criteria you define. This happens during workspace refreshes.
      - **Static**: You choose which files are included in the collection. You can edit the files if the data collection is in `draft` state. After the data collection moves to `Published` state, it cannot be edited.
4. Specify the source subset:
   a. Use keywords and filters (file type, timestamps, and other attributes) to find the relevant files to include.

   You can select a file name to open a preview window of the content.

5. Add these files to the data collection.
6. Select **Save** to finalize the collection.

Result

You have defined the scope of the data collection and added the required files to it. AIDE generates embeddings and builds the vector index when you publish the collection.

Create small, focused collections (for example, per use case or domain) rather than a single "everything" collection. This improves retrieval relevance and manageability.

Publish a data collection

Publish the data collection to make it queryable by AI applications through a RAG retrieval endpoint. Publishing generates vector embeddings from your selected files and indexes them for semantic search. After the collection reaches `Ready` state, its endpoint becomes available for data scientists to integrate into notebooks, pipelines, and AI applications for retrieval-augmented generation (RAG) and search.

For large collections, consider scheduling initial publish and major re-publishes during off-peak times to minimize resource contention.

Steps

1. Navigate to **Data Curator > Data collections** and select the options menu (•••) for your data collection.
2. Select **Publish**.
3. Select a default or custom optimization configuration.
4. Select **Publish** to initiate the data transformation.
5. In AIDE Console, open the collection detail view (**Data Curator > Data collections**) for status updates.

Result

The collection reaches the `Ready` state and is available for use by downstream applications and data scientists.

From **Data Curator > Data collections**, you can select **Copy URI** to obtain the information needed to access the data collection using an API.

Update or delete a data collection

Over time you might need to refine or retire data collections. Refining a collection might involve adjusting filters to add or remove files, changing embedding settings, or updating the collection description. Deleting a collection removes it permanently and makes its retrieval endpoint unavailable.

Update a data collection

You can update a data collection when it's in `draft` state.

Steps

1. Navigate to **Data Curator > Data collections**.
2. Select the collection you want to modify.
3. Choose **Edit**.
4. Adjust any of the following:
   - Name and description
   - Filters (paths, file types, classification tags).
   - Embedding and chunking settings.
5. Save your changes.
6. Publish the collection again so that the new definition and embeddings take effect.

Result

A new indexing job runs with the updated configuration, and the collection returns to a `Ready` state when

complete.

Delete a collection

Deleting a collection is permanent. Ensure that no production application still depends on the collection's retrieval endpoint before deleting it.

Steps

1. Navigate to **Data Curator > Data collections**, and select the options menu (•••) for the collection.
2. Choose **Delete**.
3. Confirm the deletion.

Result

The collection definition and its embeddings are removed from AI Data Engine. Applications attempting to query the former retrieval endpoint will fail after the collection is removed.

What's next?

- View data collections

View data collections in AI Data Engine

After data engineers or data scientists create and publish data collections from workspaces, you need visibility into their status, size, and impact on the AI Data Engine cluster.

If you're a storage administrator, data engineer, or data scientist, you can view data collections across ONTAP System Manager and AIDE Console.

Before you begin

- You need either *storage administrator* privileges in ONTAP System Manager or *data engineer* or *data scientist* privileges in AI Data Engine Console (`https://<cluster_management_ip>/console`) to view data collections.
- At least one workspace exists with successfully extracted metadata.
- Data engineers or data scientists have created and published at least one data collection from AI Data Engine Console.
- The AI Data Engine software license is installed and inferencing features are enabled, so that vectorization and retrieval endpoints are active.

View cluster-wide data collections

For storage administrators, ONTAP System Manager provides a cluster-wide view of data collections and their footprint but does not allow admins to create or modify them.

Steps

1. In System Manager, navigate to **Data Engine > Data collections**.
2. Review the inventory summary at the top of the page:
    - Total number of data collections by status

- Total space consumed by the vector database across all collections
- Vector space as a percentage of overall cluster capacity

3. Select an individual data collection and review:
   - Collection name and description
   - UUID
   - Associated workspace
   - Status
   - Collection size
   - Creator
   - Last refresh time

Result

You now have a high-level view of all data collections in the cluster and their storage impact. Use this view to identify collections that are large, stale, or stuck in a non-ready state.

You can also see whether an individual data collection is actively being updated and whether any failures are blocking RAG usage.

Monitor collection-related jobs and events

As a storage administrator, you can monitor jobs that build and update collections from the cluster-wide **Activity** page and from the workspace details.

Steps

1. In System Manager, navigate to **Data Engine > Activity**.
2. On the **Events** tab:
   a. Filter by type (for example, workspace, data collection) or severity.
   b. Expand any event related to data collections (for example, "Data collection publish failed") to see more details.
3. On the **Jobs** tab:
   a. Filter to focus on data collection indexing and publishing jobs.
   b. For each job, open the peek view to see:
      - Progress percentage.
      - Start and end times.
      - Any reported error messages or warnings.
4. Optionally, navigate back to the affected workspace (**Data Engine > Workspaces**) and open its **Activity** tab to see events and jobs scoped only to that workspace.

Result

You can track the lifecycle of data collections, identify stalled or failed jobs, and gather contextual information to pass to data engineers, data scientists, or support.

When a data collection remains in `Publishing` state for an extended period, check for a corresponding long-running job in the Activity page before assuming a failure.

View data collections from AIDE Console

Data engineers and data scientists typically monitor data collections directly from AIDE Console, where they are created and published.

Steps

1. Log in to AIDE Console as a data engineer or data scientist.
2. Navigate to **Data Collections** and select the desired data collection.
3. For each collection:
   a. Check the state (`Draft`, `Publishing`, `Ready`, or `Failed`).
   b. Select the data collection name to review definition details (filters, included file types, classifier options, embedding settings).
   c. Inspect timestamps for last publish or update.
4. If needed, open job details or logs (where available) to understand failures or incomplete runs.

Result

Data engineers and data scientists can iterate on collection definitions and publish them again while monitoring status and health, without involving storage administrators.

What's next?

- Create data collections for RAG in AIDE Console

Implement guardrails

Define your guardrail policies in AI Data Engine for your data estate

As a data or platform owner, you use AI Data Engine (AIDE) Console to define which data is in scope for AI, which data is always off-limits, and what safety rules apply when that data is used for classification and retrieval-augmented generation (RAG).

Use these procedures to define those policies in AIDE Console so that ONTAP System Manager can enforce them on all data in workspaces.

Before you begin

- You need *storage administrator* privileges in AI Data Engine Console (`https://<cluster_management_ip>/console`) to create and manage global policies.
- You have an AIDE cluster with deployed and healthy data compute nodes.
- OpenID Connect (OIDC) is configured, and your IdP role is mapped to an AIDE admin role that allows data policy management.
- The AI Data Engine software license is installed so that guardrail and inferencing features are enabled.
- At least one workspace exists, or you have coordinated with the administrator to understand which data sources (volumes) will be used in workspaces.

Understand policy types

AIDE Console exposes these policy types that shape your data estate:

- **Classifiers**: Enable classifiers to detect PII, security issues, or other patterns across all workspaces.
- **Classifier categories**: Group classifiers into compliance categories for organization and management.
- **Guardrail policies**: Safety and redaction rules applied at the time of retrieval or inference.

You can't use ONTAP System Manager to create or manage these guardrail policies. It only reads them and enforces them when a storage admin applies them to workspaces. All policy definition and maintenance occurs in AIDE Console.

Enable classifiers

Classifiers analyze both metadata and content to annotate files and objects (for example, detecting PII or sensitive categories). Before classifiers can run on workspace data, you must enable them in AIDE Console.

About this task

Classifier behavior is controlled globally in AIDE Console. All enabled classifiers run on every workspace. Because they are globally applied, they cannot be enabled or disabled for an individual workspace. They can be enabled or disabled globally only.

Steps

1. In AIDE Console, navigate to **Data Guardrails > Classifiers**.
2. Select a classifier category to reveal the classifiers it contains.

3. Select the checkboxes for the classifiers you want to enable, or select all rows to enable classifiers in bulk.
4. Select **Enable**.

Use the bulk-select option to enable multiple classifiers at once. Each time you enable a classifier, a workspace refresh is triggered across all workspaces. To minimize unnecessary refreshes, enable multiple classifiers at once rather than one at a time.

Result

All newly created and existing workspaces run the enabled classifiers during metadata processing.

Classification tags are written to the metadata catalog and become available to data engineers for filtering when creating data collections.

Manage classifier categories

Classifiers are organized into categories (such as "PII" or "Financial data"). Categories help you group related classifiers for easier management and compliance visibility. You can use the default categories that AIDE provides or create custom categories to match your compliance requirements.

Steps

1. In AIDE Console, navigate to **Data Guardrails > Classifiers**.
2. View the existing classifier categories. There are two major categories of classification:
   - Content or data: Detects particular types of data within files.
   - Document: Classifies the type of document based on the content.
3. Determine if the default classifier subcategories are sufficient or if you want to create your own subcategory.
   - If you are using a default classifier subcategory (for example, **General Privacy**):
     i. Select the category name in Classifier categories to reveal the associated classifiers.
     ii. Examine the list of classifiers.
     iii. Select **Add** to find and add unlisted classifiers from the complete list of available classifiers.
   - If you want to create a custom category, select + Add .
     i. Add a unique name, description, and assign available classifiers to the category.
     ii. Select **Add**
4. To disable a classifier within a category, select ••• for the classifier and choose **Disable**. You can also select all rows to make state changes in bulk.

Result

Categories organize classifiers for compliance visibility. Data engineers can use classification tags when filtering and creating data collections.

Create and manage guardrail policies

Guardrail policies determine how AIDE responds when classifiers detect sensitive content or when prompts and retrieval results violate content rules.

Typical guardrail behaviors include:

- Masking or redacting PII from retrieved snippets.
- Blocking answers that violate compliance rules.
- Logging or tagging violations for audit.

About this task

You create and manage guardrail policies only in AIDE Console.

You can only associate workspaces in ONTAP System Manager with a single guardrail policy at a time.

Steps

1. In AIDE Console, navigate to **Data guardrails** > **Guardrail policies**.
2. Select **Add**.
3. Enter a name and description that clearly describe the scope (for example, `Customer PII redaction for support KB`).
4. Configure data classifier-driven conditions required for guardrail activation:
   a. Define the conditions for guardrail activation:
      i. Choose the classifier category or classifier type for each condition.
      ii. Add and define additional conditions as needed.
      iii. Define specific search criteria in **Search**, then select **Accept**.
   b. Define actions for the guardrail policy, such as anonymizing content or blocking and removing a file from a data collection.
5. Select the workspace that the guardrail will be applied to.
6. Set the policy state:
   - **Enabled**: Activates the policy immediately.
   - **Test Mode**: Allows you to validate the impact of the policy before activating it.
   - **Disabled**: Saves the guardrail without enforcing it.
7. Select **Add** to save the policy and apply it to the workspace.

Use **Test Mode** with a pilot workspace and a non-production data collection to understand how many responses would be affected before enabling strict enforcement.

Result

The new guardrail policy is active and scoped to the selected workspace.

How policies interact with workspaces

After policies are defined:

- The storage admin uses ONTAP System Manager to create workspaces, select data containers, and associate a guardrail policy.
- Classifiers run automatically on workspace content based on what you've enabled.
- Guardrails attached to the workspace influence how retrieval endpoints behave.

For data engineers and data scientists:

- The visible data estate (workspaces and data collections) is already filtered by role assignment.
- Metadata you query (for example, PII tags) is driven by the classifiers that are enabled.
- The responses your RAG pipelines receive are constrained by the guardrails configured at the workspace level.

Related information

- View data collections in AI Data Engine
- Data-to-RAG quick start for AI Data Engine

FAQ for NetApp AI Data Engine

This FAQ covers common questions about NetApp AI Data Engine (AIDE), including its architecture, deployment, user types, technical features, integration, and licensing.

AIDE basics

What is NetApp AI Data Engine (AIDE)?

NetApp AI Data Engine (AIDE) is a storage-integrated AI data service that spans the entire AI lifecycle from discovering and preparing raw data to providing retrieval endpoints to power generative AI (GenAI), Retrieval-Augmented Generation (RAG), agentic AI, and AI factories. AIDE automates sync and change detection, providing a unified, up-to-date view of selected data for data discovery and curation.

How does AIDE work?

AIDE integrates directly with NetApp ONTAP storage systems to create a global, structured view of the entire NetApp data estate with automated change detection and synchronization. AIDE provides real-time vectorization with compression and deduplication, policy-driven guardrails, and integration with AI tools.

Users and roles

Who uses the AI Data Engine?

Primary users of AIDE include:

1. **ONTAP storage administrators**: Manage infrastructure, AI-specific storage needs, security, and compliance.
2. **Data engineers**: Manage data movement, preparation, and integration across environments.
3. **Data scientists**: Prepare and transform the relevant data for AI consumption.

Requirements and Deployment

What hardware is required?

AIDE requires AFX systems for deployment (including an AFX controller, disk shelf, and network switch), but can use cluster data from clusters running ONTAP 9 using SnapMirror and cluster peering. At least four AFX controller nodes are required for AIDE deployments to ensure high availability and performance.

AIDE runs on a NetApp data compute node (DCN). Three DCNs are required. The DCN hosts the AIDE software, which includes the Metadata Engine, Data Sync, Data Curator, and Data Guardrails.

Can I use my own DCN?

No. The DCN is a NetApp-provided data compute hardware node and is the only deployment mechanism for the AI Data Engine.

What is the minimum number of DCNs required?

Exactly three DCNs are required.

What OS runs on the DCNs?

The DCNs run a NetApp-provided software stack with AIDE.

Can AIDE be deployed without AFX?

No. AIDE requires AFX for deployment. AIDE uses Trident to consume the AFX volumes for internal storage (persistent volumes). The AFX cluster providing storage for AIDE can be peered with an ONTAP 9 system or cluster. It uses cluster peering and SnapMirror to sync data from the remote ONTAP cluster to the AFX system.

Management and Interfaces

Is the AIDE Console part of NetApp Console or a separate interface?

The AIDE Console is a separate management interface that runs on DCNs. You use the AIDE Console to manage AIDE services, such as Data Guardrails and Data Curator. You can also use ONTAP System Manager to monitor the AIDE cluster.

Features and Capabilities

What are the key features of AIDE?

There are four main features of AIDE:

Metadata Engine

- Automatically generates a structured, up-to-date, interactive view of your data.
- Works with data stored on ONTAP.
- Enables data practitioners to collaborate with storage admins to find and understand data.
- APIs query metadata to provide capabilities while reducing NFS traffic load on storage systems.
- Metadata extraction and cataloging capability is built specifically for AIDE and works on a continuous basis and leverages ONTAP capabilities like snapshots.

Data Sync

- Maintains data recency automatically as source data changes without manual intervention.
- Admins define the data refresh interval in days or hours.
- Provides incremental data mobility and sync across the data to eliminate redundant copies of AI data.

Data Guardrails

- Automatically identifies and protects sensitive data throughout the AI lifecycle. It’s accessible through AI Data Engine Console.
- Continuously scans, classifies, and categorizes data.
- Identifies sensitive data (such as PII) and risks.
- Facilitates the creation of policies for automatic handling of sensitive data in line with company and regulatory standards.
- Provides automatic sensitive information redaction for data protection.
- Restricts access to sensitive files as necessary.

Data Curator

- Allows data scientists to search across storage for relevant data.
- Creates curated data collections with data existing on AFX volumes.
- Generates vector embeddings at the storage layer to reduce data bloat and increase performance.
- Provides a retrieval endpoint for AI applications with vector semantic search and re-ranking.

Integration and Interoperability

Does AIDE support federated metadata across multiple ONTAP clusters?

AIDE can connect to multiple ONTAP clusters using SnapMirror and cluster peering, enabling centralized metadata visibility.

Where is the metadata stored?

AIDE stores metadata on the connected AFX cluster using a persistent volume provided by AFX. The DCNs use local storage for internal operations.

Does the AIDE Metadata Engine classify data?

No. The Metadata Engine catalogs filesystem metadata and provides APIs to query this cataloged metadata.

What data sources are supported?

AIDE supports ONTAP volumes (local or remote) as data sources. Remote ONTAP clusters must run ONTAP 9 and be connected via cluster peering and SnapMirror.

ONTAP S3 buckets and StorageGRID objects are not supported as data sources in AIDE 9.18.1.

What types of files can AIDE process for classification, vectorization, and semantic search?

AIDE supports a wide range of file types including PDF, DOCX, PPTX, TXT, and image files with OCR capabilities.

Does AIDE support classification of non-English data?

AIDE supports English-language data only.

What integrations does AIDE support?

AIDE provides a RAG API endpoint accessible through direct API calls or through a Model Context Protocol (MCP) server. This supports integration with agentic AI frameworks and tools.

Deployment and Licensing

What are the deployment options?

AIDE is deployed on-premises on AFX infrastructure with DCNs. It integrates directly with NetApp ONTAP AFX installations.

How is AIDE licensed?

AIDE requires a software license to run Data Guardrails and Data Curator.

If you require only the Metadata Engine, the ONTAP One license, which is included with all AFX systems, provides entitlement for Metadata Engine-only capabilities.

Related information

- Install AIDE licenses in ONTAP System Manager
- Learn about AIDE architecture and components

Legal notices

Legal notices provide access to copyright statements, trademarks, patents, and more.

Copyright

https://www.netapp.com/company/legal/copyright/

Trademarks

NETAPP, the NETAPP logo, and the marks listed on the NetApp Trademarks page are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

https://www.netapp.com/company/legal/trademarks/

Patents

A current list of NetApp owned patents can be found at:

https://www.netapp.com/pdf.html?item=/media/11887-patentspage.pdf

Privacy policy

https://www.netapp.com/company/legal/privacy-policy/

Open source

Notice files provide information about third-party copyright and licenses used in NetApp software.

AI Data Engine

Notice for AIDE 9.18.1

Copyright information

Trademark information