AI Data Engine 文档

AI Data Engine

NetApp
March 13, 2026

目录

AI Data Engine 文档

发行说明

AI Data Engine 的新增功能

AI Data Engine (AIDE) 9.18.1 是 NetApp 的 AI 数据管理平台的初始版本。此版本引入了 Metadata Engine 和管理工作流程，使组织能够为 AI 工作负载编目和组织非结构化数据，为高级治理和矢量化功能提供基础。拥有适当 AI Data Engine 许可证的客户可以使用高级治理（护栏）和矢量化。

AIDE 9.18.1 初始版本中的新增功能

AIDE 9.18.1 引入了以下基础功能：

"用于 AI 数据编目的 Metadata Engine"

初始版本包括一个 Metadata Engine，用于跨 ONTAP 集群对文件和对象进行编目。

主要功能包括：

- 从对等集群上的本地和远程 ONTAP 卷自动提取元数据（核心和扩展属性、对象标签）。
- 为需要企业数据全局视图的应用程序提供集中查询和筛选 REST API。
- 可扩展的元数据存储。
- 在工作区创建期间触发的自动元数据提取。

"工作区管理"

工作区为 AI 项目提供数据源（卷）的逻辑分组。

初始版本支持：

- 创建跨越本地和远程 ONTAP 卷的工作区（使用集群对等）。
- 为工作区分配访问控制，支持多用户和多租户环境。
- 创建工作区时自动元数据提取和目录填充。

"用于自动数据时效性的 Data Sync"

Data Sync 可使元数据目录和数据集在源数据更改时保持最新，无需手动干预。

主要功能包括：

- 使用策略驱动的 SnapMirror 复制自动同步来自远程或本地 ONTAP 集群的数据。
- 仅传播已修改数据的增量更新，从而降低开销。
- 每个工作区可配置的刷新间隔。
- 工作区级别的同步状态和活动监控。

"集群设置和管理"

初始版本包括以下工作流：

- 在集群设置期间发现和添加数据计算节点 (DCN)。
- 为 Metadata Engine 创建专用元数据存储虚拟机。
- 用于集群范围元数据访问的 Data Engine 服务接口配置。
- 与其他 ONTAP 集群对等，以将元数据编目扩展到整个数据资产。

"OpenID Connect (OIDC) 身份验证"

- 基于 OIDC/OAuth 的身份验证，用于使用 Microsoft Entra ID 和 Active Directory Federation Services (ADFS) 安全访问 ONTAP System Manager 和 Data Engine Console。
- 用于工作区和元数据管理的基于角色的访问控制。

"先进的数据管理和治理能力"

以下功能适用于拥有相应 AI Data Engine 许可证的客户：

- 矢量化和 **RAG**：使用 AIDE 工作区的元数据，在 AI Data Engine Console 中创建数据集合、嵌入和检索端点。
- 基于护栏的治理：在 AI Data Engine Console 中定义护栏策略，并将这些策略与 ONTAP System Manager 中的工作区关联。

支持的硬件和平台

AI Data Engine 9.18.1 运行在结合了以下内容的 ONTAP AI 数据平台集群上：

- AFX 1K 存储节点
- NetApp 数据计算节点

相关信息

- "AI Data Engine 的已知限制"
- "了解 AI Data Engine 架构和组件"

AI Data Engine 的已知限制

已知限制可识别此版本产品不支持的平台、设备或功能，或无法与其正确互操作的平台、设备或功能。请仔细查看这些限制。

AIDE 9.18.1 初始版本的已知限制

这些限制适用于 Metadata Engine、数据计算节点和 AIDE 9.18.1 中的管理工作流。

计算节点需求与管理

- 最低数据计算节点要求

  AIDE 群集需要最少和最多 3 个数据计算节点 (DCN) 才能实现 Metadata Engine 功能。少于 3 个 DCN 的群集无法启用 Metadata Engine 功能。

- 不支持使用 **NetApp Console** 添加 **DCN** 节点

DCN 节点升级和添加必须使用 ONTAP System Manager 执行，而不是通过 NetApp Console。

支持的数据源

- 不支持 **ONTAP S3** 存储桶或 **StorageGRID** 作为数据源

  仅支持 ONTAP 卷（本地或远程）作为工作区和元数据编目的数据源。在此版本中，ONTAP S3 存储桶和 StorageGRID 对象无法添加到工作区，也不会被 Metadata Engine 索引。

- 不支持使用 **FlexCache** 卷创建工作区

  无法将 FlexCache 卷作为数据源添加到工作区。

软件更新和还原限制

- 仅针对 **DCN** 节点的手动软件更新

  AIDE 9.18.1 不支持 DCN 群集的自动软件更新。DCN 节点软件只能通过从本地客户端上传映像来更新。不支持从外部服务器（HTTP/FTP）下载映像。

- 无需还原 **DCN** 集群软件

  DCN 集群软件无法还原到早期版本。仅允许升级到更高版本。

- **AFX** 存储集群无 **ONTAP** 还原

  AFX 存储集群无法还原为较早的 ONTAP 版本。仅允许升级到更高版本。

工作空间生命周期和访问配置

- 不对工作区进行软删除或还原

  删除工作区是永久性的。没有恢复已删除工作区的选项。

- 初始集群设置期间不支持 **OIDC** 配置

  必须在使用 ONTAP System Manager 创建集群后执行 OIDC/OAuth 配置。

相关信息

- "AI Data Engine 的新增功能"

开始使用

了解您的 AI Data Engine 系统

了解 AI Data Engine

NetApp AI Data Engine (AIDE) 是一个企业级平台，旨在加速和简化 AI 驱动的数据处理、管理和治理。AIDE 可以帮助将大量非结构化数据转换为结构化、AI 就绪的数据集。它旨在满足现代机器学习 (ML) 和生成式人工智能 (GenAI) 工作负载的需求，支持传统的 IT 运营和新的以人工智能为中心的角色。

AIDE 应对 AI 挑战

AIDE 旨在帮助组织管理 AI 工作负载的数据，并提供以下关键功能：

- 集中式元数据管理：AIDE 从 ONTAP 卷中收集元数据并对其进行编目，从而可以对数据集进行搜索、分类并将治理策略应用于数据集。
- 自动数据处理：AIDE 支持为 AI 和 ML 工作负载创建数据管道，包括为语义搜索生成向量嵌入的能力（需要适当的许可）。
- 数据隔离和访问控制：AIDE 为多个团队或项目实施访问控制和基本数据隔离。
- 与 NetApp 工具集成：AIDE 与 ONTAP System Manager 合作进行存储管理，并为数据工程师和科学家提供专用界面（AI Data Engine Console），以管理数据收集和工作流。

高级设计特征

以下设计特性定义了如何构建 AI Data Engine 以满足 AI 工作负载的需求：

- 基于微服务的服务：使用 Kubernetes 为元数据编目、矢量搜索和基础设施管理编排模块化的弹性服务。
- 企业级安全性：对所有数据和元数据实施加密、基于角色的访问控制 (RBAC) 和审计。
- 多协议数据访问：支持 NFS 和 SMB，实现灵活的数据接收和检索。
- 自动化数据管道：跟踪数据更改、创建嵌入和管理 AI 应用程序的矢量数据库。

数据如何通过 AIDE 传输

了解数据如何通过 AIDE 流动有助于说明该平台对 AI/ML 团队的价值：

1. 数据摄取：文件使用标准协议（NFS 和 SMB）存储在 ONTAP 卷中。数据可以驻留在本地 AIDE 存储（AIDE 部署中的 AFX 集群）或远程 ONTAP 集群上。来自远程集群的数据使用 ONTAP SnapMirror 同步到本地 AFX 集群，因此 AIDE 处理的所有数据最终都会在本地存储和访问。

> 不支持 S3 存储区作为工作区或数据集合的数据源。

1. 工作区创建：存储管理员在 ONTAP System Manager 中定义工作区，为特定项目、团队或工作流程分组相关的 ONTAP 卷。访问权限和治理策略在工作区级别分配。
2. 元数据提取：AIDE 自动扫描工作区中的文件和对象，提取元数据（文件类型、大小、时间戳、自定义属性

）并将其存储在集中式目录中。随着数据的变化，这种情况会不断发生。

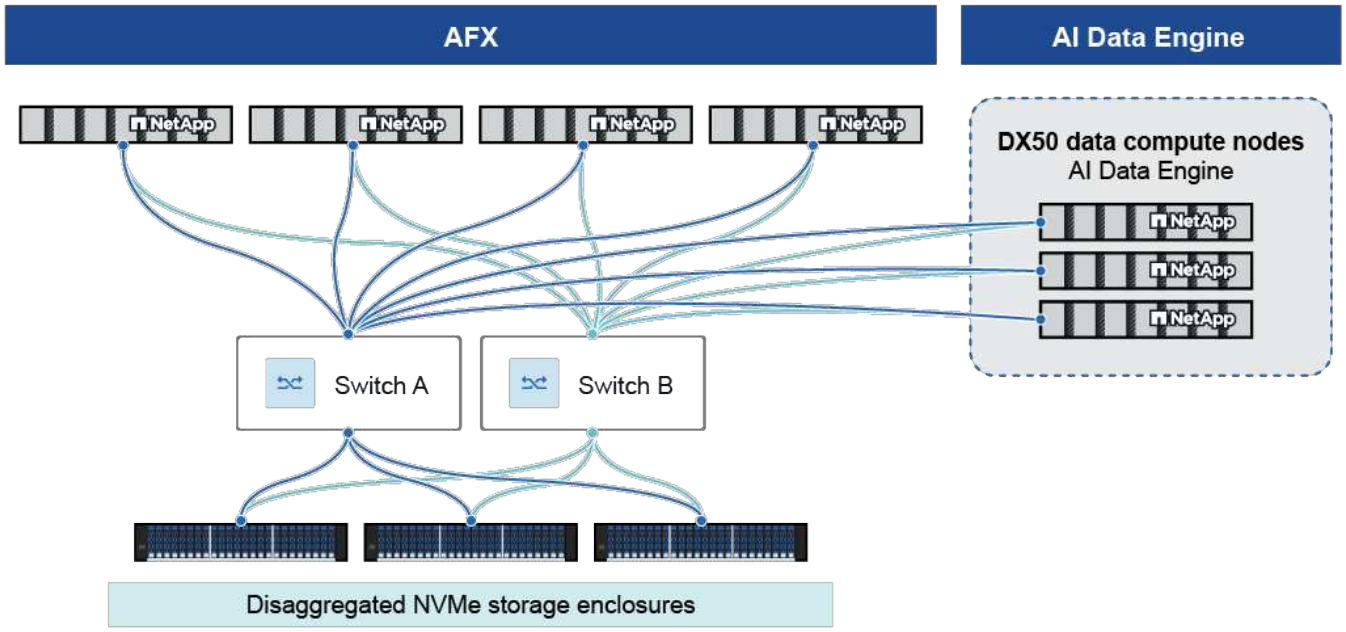
3. 分类和治理：分类器扫描数据以查找敏感信息（PII、财务数据）或文档类型（法律、人力资源）。Data Guardrails 政策自动执行编辑或访问限制。
4. 数据收集创建：数据工程师和数据科学家使用 AI Data Engine Console 查询元数据目录，筛选结果，并为特定的 AI 任务汇编策划的数据集。
5. 矢量化：对于需要语义搜索的集合，AIDE 使用选定的 AI 模型生成嵌入。矢量存储在矢量数据库中，用于高性能检索。
6. **AI/ML** 消耗：应用程序通过多条路径访问数据：
   - 使用 NFS 或 SMB 直接访问文件/对象
   - 针对矢量数据库的语义搜索查询
   - 将数据检索与 GenAI 模型集成相结合的 RAG 端点
   - 用于编程工作流程的 REST API 访问

这种自动化、策略驱动的工作流程减少了为 AI 准备数据所需的时间和手动工作量，使团队能够专注于模型开发和见解，而不是数据争论。

AI Data Engine 架构

AIDE 基于可扩展的容错架构构建，将存储和计算分离，为 AI 工作负载提供高性能和灵活性。

物理组件



AFX 控制器节点

AFX 控制器节点运行专为支持 AFX 环境要求而设计的 ONTAP 软件的专用个性。客户端通过多种协议访问节点，包括 NFS 和 SMB。每个节点都有存储的完整视图，可以根据客户端请求访问存储。节点具有非易失性内存的状态，以保留关键状态信息，并包括特定于目标工作负载的其他增强功能。

AIDE 部署至少需要四个 AFX 控制器节点，以确保高可用性和高性能。

数据计算节点

数据计算节点 (DCN) 是基于 Linux 的服务器，具有高 CPU、RAM 和 GPU 资源，专门用于 AI 数据处理任务。它们托管特定于 AI 的服务，如元数据编目、矢量搜索和嵌入管道。

AIDE 部署需要恰好 三个 DCN。

集群/存储交换机

冗余的高速（100GbE 或更高）交换机连接 ONTAP 和 DCN，实现低延迟数据传输和高可用性。

存储架

配备高密度 SSD 的 NVMe-oF 盘架可提供超低延迟和冗余，支持 PB 级存储。

网络连接

所有 DCN 和 ONTAP 存储节点均通过冗余高速群集交换机（最低 100GbE）连接。该架构将计算和存储资源分开，允许每个资源独立扩展，并优化性能和资源利用率。

DCN 和 ONTAP 节点之间的网络使用集群交换机上的专用 VLAN 和 IPspaces 进行隔离。这可确保所有通信（如数据访问、管理 API 和内部服务流量）保持安全、高效，并且不会干扰其他网络操作。

AI Data Engine 主要功能

AI Data Engine (AIDE) 的主要功能协同工作，以实现 AI 数据生命周期的自动化、安全和加速。每个功能都实现为一组在 DCN 上运行的微服务，与 ONTAP 存储集成，并通过 REST API 和管理接口公开。

Metadata Engine

Metadata Engine 会自动生成 NetApp 数据资源的结构化、最新的交互式视图。

许可和访问

Metadata Engine 包含在基本 ONTAP One 许可证中，可在 AIDE 安装时使用。

您可以通过 ONTAP System Manager 访问它。

功能

- 编录所有数据源的元数据，包括本地存储在 AFX 集群上的卷和从远程 ONTAP 集群同步的卷。
- 自动提取元数据，并在导入或更改数据时填充目录。
- 提供用于查询元数据的 REST API 访问，允许数据从业人员和存储管理员发现、分类和理解数据。
- 从数据路径卸载元数据查询，从而减少存储系统上的 NFS 流量负载。
- 支持具有索引和搜索功能的大型元数据记录。
- 与工作区和数据收集抽象集成，以实施访问控制和治理。

数据同步

Data Sync 是一种自动后台服务，可确保元数据目录和数据集保持最新并与基础数据源保持一致，即使在源数据发生变化时也是如此。

许可和访问

数据同步功能包含在基本 ONTAP One 许可证中，可在 AIDE 安装后使用。

功能

- 使用策略驱动的 SnapMirror 复制同步来自远程或本地 ONTAP 集群的数据。来自远程集群的数据被复制到本地 AFX 集群以进行 AIDE 处理。
- 基于检测到的更改进行增量更新，仅传播已修改的数据。
- 提供安全的增量数据移动和跨数据资产的同步。
- 使用每个工作区可配置的刷新率来安排和监控同步间隔。
- 与工作区创建工作流集成，以便在添加新数据源时提取和更新元数据。

Data Guardrails

Data Guardrails 服务在整个 AI 生命周期中为敏感数据提供持续、自动化的治理和保护。

许可和访问

基本 ONTAP One 许可证不包含 Data Guardrails 功能，需要单独的 AIDE 许可证。

您可以通过 AI Data Engine Console 访问护栏功能。

功能

- 持续扫描、分类和归类数据。
- 使用用于 PII 检测等任务的内置和可自定义分类器来识别敏感数据和风险。
- 通过策略驱动的编辑、屏蔽和访问限制，自动处理敏感数据。
- 通过附加到工作空间的 Data Guardrails 策略执行公司和监管标准。
- 通过审核日志记录和合规性报告，限制对配置的敏感文件或卷的访问。
- 与工作空间和数据收集管理集成，在 AI 数据工作流程中一致地应用护栏。

Data Curator

Data Curator 服务可实现 AI 和 GenAI 应用程序的快速数据发现、搜索、矢量化和检索。

许可和访问

基本 ONTAP One 许可证不包含 Data Curator 功能，需要单独的 AIDE 许可证。

您可以通过 AI Data Engine Console 访问 Data Curator。

功能

- 使用集中式元数据目录搜索存储中的相关数据。
- 为数据科学家提供工具，以创建精心策划的数据集。

- 在存储层自动生成矢量嵌入。
- 为 AI 应用程序提供安全的检索端点，支持矢量语义搜索和重新排名。
- 与 AI 工具和技术集成，包括 Retrieval-Augmented Generation (RAG) 管道和代理 AI 框架。
- 提供 REST API，用于以编程方式访问数据集、矢量搜索和检索端点。

安全与多租户

该平台同时实施基于角色的访问控制 (RBAC) 和资源级别的访问控制列表 (ACL)。审核所有 API 和用户操作，并在静态和传输过程中对所有数据进行加密。针对数据和元数据对单个租户进行隔离。

相关信息

- "安装 AIDE 许可证"
- "Data-to-RAG 快速启动"

按角色划分的 **AIDE** 组件和职责

AI Data Engine 组件和基于角色的交互

AI Data Engine (AIDE) 由许多核心组件组成，它们协同工作，为人工智能工作负载提供全面的数据管理和处理平台。这些组件包括工作区、数据集、矢量数据库、护栏、元数据目录、检索端点和分类器。每个组件都在实现高效的数据发现、管理、治理以及与 AI/ML 应用程序的集成方面发挥着特定的作用。

每个 AIDE 用户根据其角色与 AIDE 组件进行不同的交互。

以存储和数据为重点的用户角色

AIDE 引入了新的用户角色，同时仍然支持传统的 ONTAP 系统管理角色：

存储用户

- 存储管理员：管理 AFX 和 AIDE 集群设置、网络、存储配置和用户访问。

数据用户

- 数据工程师：构建和优化 AI/ML 流水线，管理数据收集，并集成 AI 模型。
- 数据科学家：发现、管理和分析数据集，创建数据集合，并利用 GenAI 应用程序的检索端点。

角色 (**RBAC** 名称)	说明
存储管理员 (`admin`)	管理 AFX 和 AIDE 集群设置、网络、存储配置和用户访问。将 RBAC 角色分配给用户，以确定对 AIDE 接口和功能的访问级别。此管理员角色具有使用 ONTAP System Manager 和 AI Data Engine Console 的完全管理访问权限。
数据工程师 (`data-engineer`)	构建和优化 AI/ML 管道，管理数据收集，并集成 AI 模型。该角色可以访问 AI Data Engine Console 进行数据工程工作流程。

角色 (**RBAC** 名称)	说明
数据科学家 (`data-scientist`)	发现、管理和分析数据集，创建数据集，并利用 GenAI 应用程序的检索端点。该角色可以访问用于数据科学工作流程的 AI Data Engine Console。

AIDE 系统组件

每个 AIDE 用户（存储管理员、数据工程师和数据科学家）根据其角色与 AIDE 组件进行交互。

工作区

工作区是集群内数据的逻辑段，用于对特定项目、团队或工作流的卷进行分组。工作区定义了 AIDE 中数据可见性、访问权限和治理的范围。

元数据目录

一个集中的、可扩展的数据库，存储本地群集中所有文件和对象的元数据记录，包括使用 ONTAP SnapMirror 或群集对等从远程 ONTAP 群集同步的数据。它支持丰富的交互式搜索和过滤。

分类器

分类器是一种工具（内置或自定义），用于扫描和标记特定类型的敏感数据（例如 PII、金融、医疗保健）的文件，或按类型（例如法律、HR、销售）对文档进行分类。

数据收集

数据集是来自工作区的相关文件或对象的精选组，由用户指定的查询定义，用于 GenAI 工作流。发布后，数据集合中文件的内容可用于 GenAI 应用程序的 API 进行语义搜索。

向量数据库

矢量数据库存储从数据集合生成的嵌入，为 AI 和 GenAI 应用程序实现高性能语义搜索和检索。

护栏

Guardrails 是策略驱动的机制，可在整个 AI 数据生命周期中实施数据治理、分类和保护（例如编辑或访问限制）。

检索端点（**RAG** 端点）

检索端点（有时称为检索增强生成或"RAG"端点）是一种安全的 API，使 AI 和 GenAI 应用程序能够从策划的集合和矢量数据库中访问相关数据、上下文或嵌入。

RAG 端点旨在支持高级 AI 工作流程，例如生成式 AI 模型中的语义搜索和上下文感知响应。通过将 AI 应用程序连接到检索端点，您可以提供对由 AIDE 管理的精选 AI 就绪数据集的实时访问，从而提高模型的准确性和相关性。

相关信息

- "AIDE 存储管理员如何使用 AIDE 组件"
- "AIDE 数据工程师如何使用 AIDE 组件"

- "AIDE 数据科学家如何使用 AIDE 组件"

AI Data Engine 接口

AI Data Engine (AIDE) 为用户交互和自动化提供了三个主要界面。每个角色（例如存储管理员、数据工程师和数据科学家）都会根据其特定任务和职责利用这些界面。

ONTAP System Manager

ONTAP System Manager 是为存储管理员设计的基于 Web 的界面。它为集群设置、工作区管理、DCN 监控和附加护栏策略提供工作流。

AI Data Engine Console

AI Data Engine Console 是数据工程师和数据科学家的专用界面。它使用户能够探索数据源、创建和管理数据集合、配置数据管道、应用分类器以及与护栏和矢量搜索功能进行交互。控制台提供高级工具，用于数据发现、管理以及与 AI/ML 工作流程的集成。

REST API

AIDE 公开了用于自动化、集成和编程访问的 ONTAP REST API。API 支持集群设置、工作区和集合管理、元数据查询、矢量搜索和检索端点。

了解 AI Data Engine 存储管理员如何使用 AIDE 组件

作为存储管理员，您可以通过 ONTAP 和 AIDE Console 管理 AIDE 基础架构，配置工作空间，附加护栏策略，并监控系统运行状况。您的角色侧重于确保 AI 工作负载的可靠、安全和合规数据存储。

存储管理员组件访问

组件	访问级别	存储管理员工作流程
ONTAP System Manager	管理（创建、编辑、删除）	您使用 ONTAP System Manager 作为集群管理、工作区配置、护栏策略管理和监控系统运行状况的主要界面。
AI Data Engine Console	管理（创建、编辑、删除）	您可以使用 AI Data Engine Console 来监控工作区、查看收集状态以及监督整个 AIDE 环境中的系统活动。
ONTAP REST API	管理（创建、编辑、删除）	您可以使用 REST API 自动化基础架构任务，以编程方式管理工作空间和护栏策略，并将 AIDE 管理与外部工具和工作流集成。
工作区	管理（创建、编辑、删除）	您可以使用 ONTAP System Manager 创建和管理工作区。您可以选择包含哪些数据源，为数据工程师和数据科学家分配权限，并附加护栏策略以强制实施治理和合规性。您还可以监控工作区运行状况和访问权限。
数据收集	查看（只读）	您可以使用 System Manager 查看每个工作区内数据集合的状态和运行状况。您确保基础数据源可用且受保护，但不会创建或修改集合。

组件	访问级别	存储管理员工作流程
护栏	管理（创建、编辑、删除）	您可以使用 System Manager 定义护栏策略并将其附加到工作区。您监控护栏状态和合规性报告。您确保根据需要执行和更新策略。
元数据目录	监控（查看运行状况、状态、活动）	您确保元数据目录已填充并保持最新。您监控目录运行状况并支持访问控制。
向量数据库	调配/监控（部署、配置、查看状态）	您配置和监控矢量数据库基础设施，确保具有 GPU 资源和适当许可的数据计算节点到位。您支持环境，但不直接管理嵌入或查询。
分类器	管理（创建、编辑、删除）	您可以创建、配置和管理分类器及其类别。您可以将分类器应用于工作区并监控其有效性。

了解 **AI Data Engine** 数据工程师和数据科学家如何使用 **AIDE** 组件

作为数据工程师或数据科学家，您可以使用 AI Data Engine Console 探索已授予访问权限的工作区，创建和管理数据集合，执行语义搜索，并将检索端点集成到 AI/ML 工作流程中。

数据工程师专注于通过构建集合、配置嵌入管道以及控制哪些用户可以访问已发布的集合，将原始数据转换为 AI 就绪数据集。数据科学家专注于利用精选数据集进行分析、模型训练和 GenAI 应用，而无需管理访问控制或基础设施。

数据用户组件访问

组件	访问级别	数据工程师工作流程	数据科学家工作流程
AI Data Engine Console	管理（创建、编辑、删除）	AI Data Engine Console 是您进行日常任务的主要界面，包括数据发现、集合管理、管道配置以及为您有权访问的工作区发布 RAG 或检索端点。	AI Data Engine Console 是您在可以访问的工作区内进行数据探索、优化和版本控制集合的主要界面，并将精选的数据集和检索端点连接到分析、建模和 GenAI 工作流程。
ONTAP REST API	管理（创建、编辑、删除）	您可以使用 REST API 自动化收集生命周期操作，触发和监控嵌入管道，并以编程方式将数据工作流与外部工具集成。	您可以使用 REST API 以编程方式访问数据集，运行矢量搜索查询，并将检索端点集成到 AI/ML 应用程序和代理框架中。
工作区	查看/使用（只读）	在构建集合之前，您可以浏览分配的工作区以识别和了解可用数据源。	您可以搜索分配的工作区，以查找与特定研究或建模任务相关的文件和对象。
数据收集	管理（创建、编辑、删除）	您通过使用标记、分类和其他属性选择和过滤源数据来构建数据集合，并管理从创建和版本控制到发布为 RAG 端点供 AI 使用的完整集合生命周期。您还可以管理哪些数据科学家和其他用户可以访问每个集合。	您可以在已授予访问权限的工作区内创建、选择、注释、版本和优化数据集合。您可以使用这些集合作为语义搜索和 GenAI 工作流程的基础。

组件	访问级别	数据工程师工作流程	数据科学家工作流程
元数据目录	查询/使用（工作流使用）	您可以使用元数据目录来评估和选择要接收的数据源，运行查询以查找相关文件，并确认它们符合您在分配的工作区中构建的集合的要求。	您可以在可访问的工作区中搜索和筛选元数据，以查找分析或模型训练所需的文件和对象，这依赖于数据工程师构建和维护的目录结构。
向量数据库	• 管理嵌入/搜索（数据工程师） • 使用/搜索（数据科学家）	您可以触发嵌入管道，监控矢量化状态，配置分块和嵌入参数，并公开由矢量搜索支持的检索端点。然后，应用程序和代理通过 API 查询这些端点，以获取语义搜索和 RAG 工作流。	您可以针对数据工程师管理的管道生成的嵌入运行语义搜索查询，并将检索结果集成到 GenAI 或 RAG 工作流中，以实现上下文感知模型响应。您不配置分块、嵌入或管道参数。
分类器	使用（消耗分类数据）	您可以使用分类结果在收集准备期间对源数据进行注释和标记，确保进入管道的内容为下游 AI 工作流正确标记。	您使用预先分类的数据，以确保在分析和建模中仅使用合规和相关的内容。

AI Data Engine 快速入门

要启动并运行 AI Data Engine 系统，您需要安装硬件组件、设置群集、设置从主机到存储系统的数据访问以及配置存储。

1 **AIDE** 将与新的或现有的 **AFX** 集群一起安装?

需要决定是将 AIDE 和 AFX 同时安装在一起，还是将 AIDE 与现有 AFX 集群集成。

2 安装和设置硬件

"安装和设置" 您的 AIDE 集群计算节点。根据安装环境的不同，还要确保 "安装" AFX 硬件。

3 设置集群

使用 ONTAP System Manager 引导您快速轻松地完成"使用 AFX 集群设置 AIDE"。

4 设置工作区和数据访问

"为该工作区设置可以访问 AI Data Engine 数据的工作区和用户"。

下一步是什么?

现在，您可以使用 ONTAP System Manager 管理您的 AI Data Engine，让您的数据工程师和数据科学家开始使用他们的工作空间和配置。

安装 AIDE

安装 **AI Data Engine** 的要求

查看安装 AI Data Engine 的要求。AIDE 需要 AFX 存储系统、至少三个 Data Compute Node、网络交换机和电缆。

硬件要求

AI Data Engine 需要 AFX 存储系统和至少三个 Data Compute Node。AFX 系统提供存储基础设施，而 Data Compute Node 托管 AIDE 软件组件，从而实现数据管理、管理和 AI 功能。

- **AFX** 存储系统：包括 AFX 控制器、磁盘架和网络交换机。AIDE 部署需要 AFX 存储系统。
- **Data Compute Node**：至少需要三个 Data Compute Node。Data Compute Node 是 NetApp 提供的硬件节点，用于托管 AIDE 软件，包括 Metadata Engine、Data Sync、Data Curator 和 Data Guardrails。
  - 每个 Data Compute Node 都有可用于连接的 I/O 插槽 4 和 5。插槽 3 预留给 GPU。插槽 1 和 2 未填充或无法访问。
  - 端口 e4a 和 e5a 用于集群连接。
  - 端口 e4b 和 e5b 用于主机网络连接。

网络交换机要求

AI Data Engine 需要网络交换机来为 Data Compute Node 启用主机网络连接和节点间通信。

- 用于主机网络连接的客户端交换机（Cisco Nexus 9332D-GX2B 或 Cisco Nexus 9364D-GX2A）
- 用于节点间通信的集群交换机（Cisco Nexus 9332D-GX2B 或 Cisco Nexus 9364D-GX2A）
- 用于网络管理的可选管理交换机

布线要求

将 Data Compute Node 连接到网络交换机和管理网络需要以下电缆。

- 400-GbE 到 100-GbE (4x100GbE) 分支电缆，用于将节点连接到客户端和集群交换机

多集群支持

使用 AFX 部署 AIDE 后，它可以使用 SnapMirror 和集群对等连接并管理来自其他 ONTAP 9.18.1 及更高版本集群的数据。

为 **AI Data Engine** 安装 **AFX** 存储系统

安装 AFX 存储系统，作为部署 AI Data Engine 的第一步。AFX 存储系统提供存储基础设施基础，并且在安装 Data Compute Node 之前是必需的。

按照 "AFX 1K 安装文档" 安装 AFX 存储系统。

下一步

完成 AFX 存储系统安装后，"安装 Data Compute Node"。

安装您的数据计算节点

AI Data Engine 数据计算节点的安装和设置工作流程

要安装和配置数据计算节点 (DCN)，您需要查看硬件要求，准备站点，安装硬件组件并连接电缆，打开系统电源，并设置 ONTAP 群集。

1. "查看硬件安装要求"

请确保已安装现有的 AFX 1K 存储系统，然后查看用于安装 AIDE 数据计算节点的硬件要求。有关安装 AFX 1K 存储系统的信息，请参阅"AFX 1K 存储系统安装文档"。

2. "准备安装数据计算节点"

要准备安装数据计算节点，需要准备好站点，检查环境和电气要求，并确保有足够的机柜空间。然后，打开设备的包装，将其内容与装箱单进行比较，并注册硬件以获得支持优势。部署 AI Data Engine 至少需要三个数据计算节点。

3. "为您的数据计算节点安装硬件"

为您的数据计算节点安装导轨套件。将您的数据计算节点固定在机柜内。最后，将电缆管理设备连接到系统后部以进行有组织的电缆布线。

4. "连接您的数据计算节点"

要连接硬件，首先将数据计算节点连接到数据网络，然后将数据计算节点连接到集群交换机。

5. "启动您的数据计算节点"

安装机架硬件并连接数据计算节点后，应打开 DCN 和 AFX 存储系统的控制器节点（如果尚未打开）。

AI Data Engine 数据计算节点的安装要求

查看 AI Data Engine 数据计算节点所需的设备和提升预防措施。

前提条件

在安装 AIDE 数据计算节点之前，请确保您已：

- AFX 1K 存储系统。

  有关安装 AFX 1K 存储系统的信息，请参阅"AFX 1K 存储系统安装文档"。

安装所需的设备

要安装 AIDE 的数据计算节点，需要以下设备和工具。

- 访问 Web 浏览器以配置数据计算节点
- 静电放电 (ESD) 腕带
- 手电筒
- 带有 USB/串行连接的笔记本电脑或控制台
- Phillips #2 螺丝刀

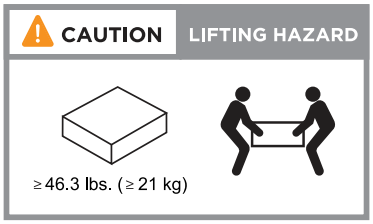
起重注意事项

数据计算节点很重。抬起和移动这些物品时要小心。

数据计算节点权重

移动或提升数据计算节点时，请采取必要的预防措施。

数据计算节点的重量可达 46.3 磅 (21 kg)。要提升数据计算节点，请使用两人或液压提升机。



相关信息

- "安全信息和监管通知"

下一步是什么?

查看硬件要求后，"准备安装您的数据计算节点"。

准备为 **AI Data Engine** 安装数据计算节点

通过准备好现场，打开箱子并将箱子的内容与装箱单进行比较，以及注册系统以获取支持权益，来准备安装 AI Data Engine 的数据计算节点。

步骤 **1**：准备站点

要安装数据计算节点，请确保计划使用的站点和机柜或机架符合配置规范。

步骤

1. 使用 "NetApp Hardware Universe"确认您的网站是否符合数据计算节点的环境和电气要求。
2. 确保在现有 AFX 1K 存储系统安装中为数据计算节点提供足够的机柜空间。
    - 每个数据计算节点 1U
    - 每个 AFX 1K 控制器节点 2U
    - 每个 NX224 机架 2U
    - 每个交换机 1U 或 2U，具体取决于交换机型号。

步骤 2：拆箱

在确保计划用于数据计算节点的站点和机柜或机架符合所需规格后，打开所有箱子，并将内容与装箱单上的项目进行比较。

步骤

1. 仔细打开所有盒子，并有条不紊地布置内容物。
2. 将拆包物品与装箱单进行比较，并记录任何差异。

   您可以通过扫描运输纸箱侧面的二维码来获取装箱单。

   以下项目是您可能会在框中看到的一些内容。

硬件	电缆	
• 挡板 • 带有说明的导轨套件 • 数据计算节点	• 电源线	

步骤 3：注册您的数据计算节点

确保您的站点满足数据计算节点规范的要求，并确认已订购所有零件后，应注册系统。

步骤

1. 找到数据计算节点的序列号。

   您可以在以下位置找到序列号：

   ◦ 在装箱单上
   ◦ 在您的确认电子邮件中
   ◦ 在每个数据计算节点上，或在某些系统上，在每个数据计算节点的系统管理模块上。

2. 导航到 "NetApp 支持站点"。
3. 确定是否需要注册存储系统：

如果您是……	按照下面的步骤进行操作…
现有 NetApp 客户	a. 请使用您的用户名和密码 Sign in。 b. 选择 **Systems** > **My Systems**。 c. 确认已列出新的序列号。 d. 如果序列号未列出，请按照新 NetApp 客户的说明进行操作。

如果您是……	按照下面的步骤进行操作…
新 NetApp 客户	a. 点击 **Register Now** 以创建帐户。 b. 选择 **Systems** > **Register Systems**。 c. 输入存储系统的序列号和所需的详细信息。 一旦 NetApp 批准您的注册后，您可以下载所需的软件。审批最多需要 24 小时。

下一步是什么？

准备好安装数据计算节点后，"安装 Data Compute Node"。

为 **AI Data Engine** 安装数据计算节点

在机柜中安装并保护您的数据计算节点。

开始之前

- 确保您拥有与滑轨套件一起包装的说明书。
- 了解与数据计算节点、存储系统和存储架的重量相关的安全问题。
- 了解通过存储系统的气流从安装挡板或端盖的前部进入，并从端口所在的后部排出。

> 一般来说，交换机应安装在机柜的中心。存储架应安装在交换机下方和第二个已安装的交换机上方。控制器节点可以安装在机柜内交换机的上方或下方。数据计算节点可以安装在机柜内控制器节点的上方或下方。

步骤

1. 根据需要，使用套件附带的说明为您的数据计算节点安装滑轨套件。
2. 在机柜中安装并保护您的数据计算节点：
   a. 将数据计算节点放置在机柜中间的导轨上，然后从底部支撑设备并将其滑入到位。
   b. 使用随附的安装螺钉将数据计算节点固定到机柜上。
3. 将挡板连接到数据计算节点的前面。

下一步是什么？

安装数据计算节点后，"为 AIDE 连接数据计算节点"。

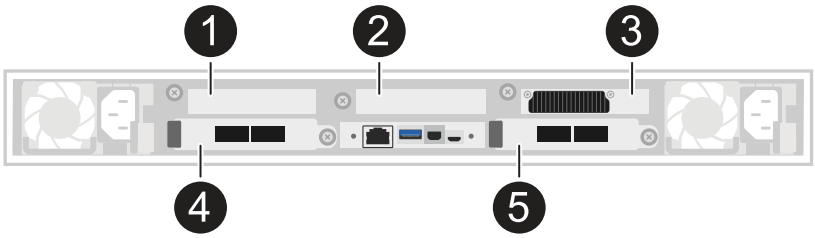
AI Data Engine 数据计算节点布线要求

Data Compute Node 通过主机网络和集群网络连接与 AFX 1K 存储系统集成。查看部署的 I/O 插槽配置、电缆类型和连接要求。

布线配置

Data Compute Node 连接到与 AFX 1K 控制器节点相同的群集交换机，通过针对 AI 和机器学习工作负载优化的计算资源扩展您的存储系统。

初始 AI Data Engine (AIDE) 配置支持至少三个 Data Compute Node。有关全面的配置详细信息和插槽优先级，请参阅"NetApp Hardware Universe"。



1	Data Compute Node 上未使用的插槽。
2	Data Compute Node 上未使用的插槽。
3	Data Compute Node 上的 GPU 插槽。
4	Data Compute Node 上的 I/O 插槽。
5	Data Compute Node 上的 I/O 插槽。

I/O 插槽配置

Data Compute Node 使用与标准服务器配置不同的特定插槽编号方案。了解插槽布局对于正确布线至关重要。

- 插槽 **3**：保留用于 GPU（不可用于 I/O 布线）
- 插槽 **4** 和 **5**：用于网络连接的 I/O 插槽
  - 端口 a：集群网络连接
  - 端口 b：主机网络连接
- 插槽 **1** 和 **2**：未填充且无法使用

网络连接

Data Compute Node 需要两种类型的网络连接才能与 AFX 1K 存储系统集成。

- 主机网络连接

  主机网络连接提供对客户端数据的访问，并使 Data Compute Node 能够处理工作负载。每个 Data Compute Node 使用端口 e4b 和 e5b 进行到独立主机网络交换机的冗余连接。

  端口分配：

  - e4b：连接到主机网络交换机 A
  - e5b：连接到主机网络交换机 B
- 集群网络连接

集群网络连接可实现存储集群内 Data Compute Node 与 AFX 1K 控制器节点之间的通信。每个 Data Compute Node 使用端口 e4a 和 e5a 进行到单独集群网络交换机的冗余连接。

端口分配：

- e4a：连接到集群网络交换机 A
- e5a：连接到集群网络交换机 B

支持的硬件组件

Data Compute Node 需要特定的电缆和交换机，以确保与 AFX 1K 存储系统的正确连接和性能。

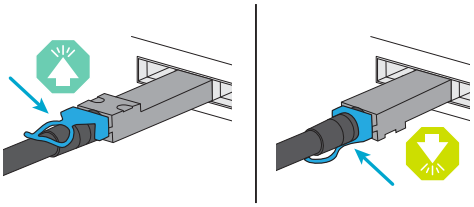
Data Compute Node	支持的交换机	支持的电缆
数据计算节点（至少需要三个）	• Cisco Nexus 9332D-GX2B（400GbE） • Cisco Nexus 9364D-GX2A（400GbE）	• 400GbE QSFP-DD 分支到 4x100GbE QSFP56 电缆，用于连接到数据计算节点： ◦ 100GbE 到数据计算节点集群网络端口（e4a、e5a） ◦ 100GbE 到数据计算节点主机网络端口（e4b、e5b） • 用于管理连接的 RJ-45 电缆

分支电缆从每个 400GbE 交换机端口提供四个 100GbE 连接。将 400GbE 端连接到交换机，将 100GbE 端连接到数据计算节点 I/O 端口。

电缆方向

将电缆连接到数据计算节点时，正确的电缆方向可确保可靠的连接。

安装过程中的电缆图形显示箭头图标，指示将连接器插入端口时电缆连接器拉片的正确方向（向上或向下）。插入连接器时，您应该会感觉到连接器卡入到位。如果感觉不到，请将其移除并翻转，然后重试。



将精致的连接器组件点击到位时，请小心处理。

下一步是什么？

查看布线配置后，"连接硬件电缆"针对您的数据计算节点。

为 **AI Data Engine** 连接数据计算节点

将您的 Data Compute Node 连接到主机网络和集群网络交换机，以实现 AI 工作负载处理

并与 AFX 1K 存储系统集成。此过程使用 100GbE 连接进行主机网络访问和集群通信，允许节点利用现有的集群基础设施，而无需关闭 AFX 系统。

关于此任务

这些程序显示了常见的配置。具体的布线取决于为您的存储系统订购的组件。有关全面的配置详细信息和插槽优先级，请参阅 "NetApp Hardware Universe"。

连接 Data Compute Node 时，无需关闭 AFX 1K 存储系统的电源。您可以将 Data Compute Node 添加到已打开电源并已配置的现有 AFX 1K 存储系统。

开始之前

- 您已安装现有的 AFX 1K 存储系统。有关安装 AFX 1K 存储系统的信息，请参阅"AFX 1K 存储系统安装文档"。
- 您已安装并配置了所需的网络交换机。有关将系统连接到网络交换机的信息，请联系网络管理员。
- 您已审阅 "Data Compute Node 的布线要求"。

部署 AI Data Engine 至少需要三个数据计算节点。

步骤 1：将 **Data Compute Node** 连接到主机网络

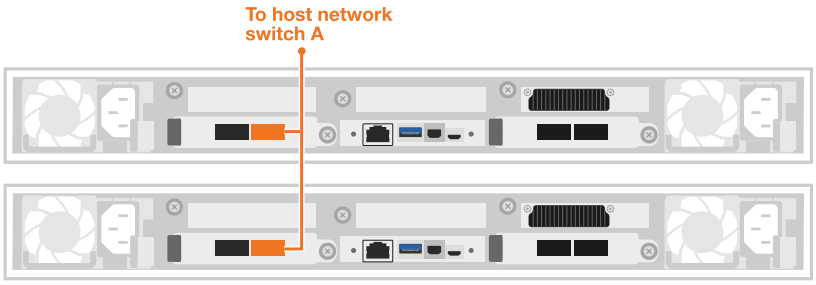
您可以将 Data Compute Node 端口连接到主机网络。

步骤

1. 将以下 Data Compute Node 的端口 e4b 连接到以太网数据网络交换机 A：
   - Data Compute Node 1，端口 e4b
   - Data Compute Node 2，端口 e4b

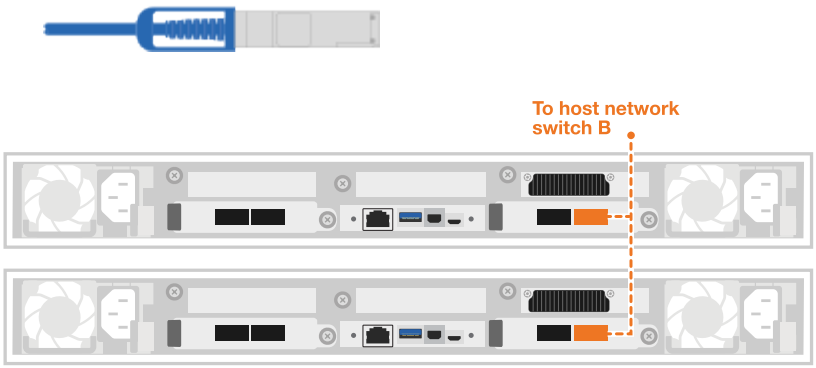
   100GbE 电缆

2. 将以下 Data Compute Node 的端口 e5b 连接到以太网数据网络交换机 B：
   - Data Compute Node 1，端口 e5b
   - Data Compute Node 2，端口 e5b

   100GbE 电缆



步骤 2：为 **Data Compute Node** 集群连接布线

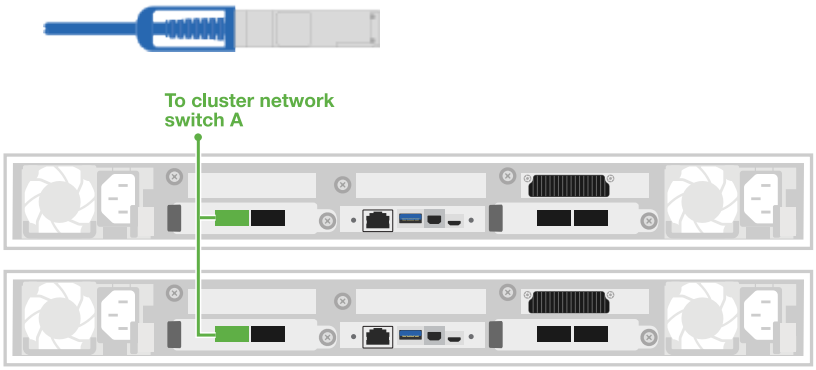
对于数据计算节点，请使用 4x100GbE 分支电缆连接 e4a/e5a 端口以进行集群连接。

步骤

1. 从下列 Data Compute Node 将端口 e4a 连接到集群网络交换机 A 上的非 ISL 端口：

   - Data Compute Node 1，端口 e4a
   - Data Compute Node 2，端口 e4a

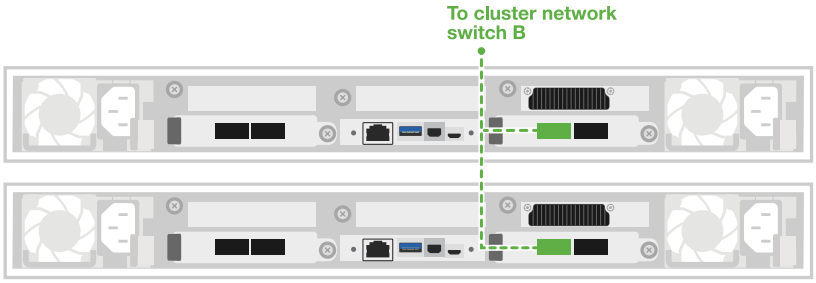
   4x100GbE 分支电缆



2. 从下列 Data Compute Node 将端口 e5a 连接到集群网络交换机 B 上的非 ISL 端口：

   - Data Compute Node 1，端口 e5a
   - Data Compute Node 2，端口 e5a

   4x100GbE 分支电缆



下一步是什么?

在您连接硬件电缆后，"开启您的 Data Compute Node"。

为 **AI Data Engine** 启动数据计算节点

安装机架硬件并连接数据计算节点后，应打开 DCN 和 AFX 存储系统的控制器节点（如果尚未打开）。

开始之前

- 请确保您的盘架已通电，并且每个盘架都分配了唯一的盘架 ID。有关为 AFX 存储系统分配盘架 ID 的信息，请参见"关于分配唯一货架 ID 的文档"。

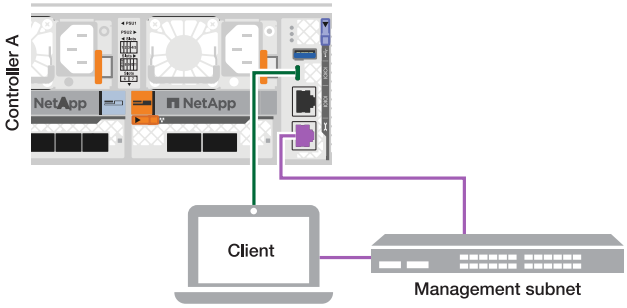
步骤

打开存储盘架并分配唯一 ID 后，打开 DCN 并打开存储控制器节点（如果尚未打开）。

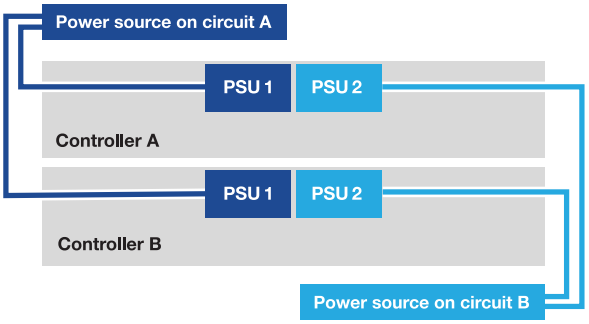
1. 将笔记本电脑连接到串行控制台端口。这允许您在控制器通电时监控引导顺序。
   a. 使用 N-8-1 将笔记本电脑上的串行控制台端口设置为 115,200 波特。

      有关如何配置串行控制台端口的说明，请参见笔记本电脑的在线帮助。

   b. 将控制台电缆连接到笔记本电脑，并使用存储系统附带的控制台电缆连接控制器上的串行控制台端口。
   c. 将笔记本电脑连接到管理子网上的交换机。



2. 使用管理子网上的地址为笔记本电脑分配 TCP/IP 地址。
3. 将电源线插入控制器电源，然后将其连接到不同电路上的电源。



- 系统开始启动。初始启动可能需要长达 8 分钟的时间。
- LED 指示灯闪烁，风扇启动，这表明控制器正在通电。
- 启动时风扇可能会发出噪音，这是正常的。

4. 将电源线插入数据计算节点电源，然后将其连接到不同电路上的电源。
5. 使用每个电源上的固定装置固定电源线。
6. 打开数据计算节点的电源。

   您可能需要卸下挡板才能访问电源开关；如果是这样，请记得之后重新安装。

下一步是什么？

打开数据计算节点后，"设置 ONTAP AIDE 集群"。

设置您的 AIDE 系统

设置 AI Data Engine

作为 ONTAP 存储管理员，您可以将 AI Data Engine (AIDE) 与 AFX 系统部署集成。您可以将 AIDE 设置为初始 AFX 存储系统部署的一部分，或者将 AIDE 添加到已部署并提供数据的 AFX 群集中。在这两种情况下，您还需要在完成基本设置后完成 AIDE 配置。

1.设置 AIDE 并将其与 AFX 存储系统集成

有两种方法可以设置 AIDE 并将其与 AFX 存储系统集成。选择适合您环境的选项。

> 在设置 AIDE 之前，请确保已安装数据计算节点并将其连接到集群交换机。有关详细信息，请参见 "安装 AIDE"。

使用新的 AFX 集群设置 AIDE

如果您将 AIDE 设置为初始 AFX 系统部署的一部分，则 AIDE 设置过程是标准 AFX 集群设置工作流程的一部分。在 ONTAP 集群设置过程中，System Manager 会自动发现连接的 DCN 并将其包含在集群配置中。之后，您就可以完成 AIDE 配置，包括所需的许可证配置。

使用现有 AFX 集群设置 AIDE

如果要向现有 AFX 存储系统添加数据计算节点群集，则需要配置 DCN 并完成 AIDE 群集的设置。

向集群添加数据计算节点

ONTAP 动态检测连接到群集网络的新数据计算节点，并将其显示在 System Manager 中。您可以"将数据计算节点添加到您的 AIDE 集群"。另请注意以下事项：

- 数据计算节点必须物理连接到集群交换机。创建 AIDE 集群时，需要恰好三个 DCN。
- 您必须具有足够的 IP 地址以用于 DCN 网络和其他必需的网络配置信息。

完成 **AIDE** 集群设置

集成数据计算节点并完成初始网络配置后，您可以完成新 AIDE 集群的设置。

关于此任务

页面右侧显示欢迎设置指导窗口，其中包含必要的配置步骤。每个已完成的步骤或操作项目的左侧都会显示一个复选标记。对于尚未完成的项目，将显示箭头 (→)。对于每个未完成的操作：

- 选择已启用的标题链接，并为您的环境完成设置操作。
- 如果链接已禁用，请将鼠标悬停在标题上，以显示启用该链接所需的操作，然后继续执行设置操作。

在可能的情况下，每个步骤都有指向其他文档的链接。部分配置可能包含在 AFX 集群设置中。有关详细信息，请参见 "AFX 文档"。

步骤

1. 选择 **Configure link aggregation groups and VLANs**。

   如果环境需要，您可以设置群集级别的 LAG 和 VLAN。选择此选项将启动以太网端口页面。

2. 选择 **Configure network protocols**。

   您可以配置 Storage VM (SVM) 网络协议。

3. 选择 **Update data compute software**。

   如果数据计算软件映像可用于更新，则将启用此选项。选择它以启动数据计算软件更新页面。有关更多信息，请参阅"更新 AI Data Engine software"。

4. 选择 **Configure data engine networking**。

   此选项显示配置 data engine 网络地址的页面。

5. 选择 **Create intercluster network interfaces**

   如果要编录远程数据，此选项将启动页面以配置群集间网络接口，这是两步过程的一部分。

6. 选择 与其他 **ONTAP** 存储系统对等。

   创建群集间网络接口后，可以与其他 ONTAP 存储系统建立对等关系。

7. 选择 **Add a data container**。

   您可以添加卷以供 AIDE 使用，并将其与特定 SVM 关联。

8. 选择 **Add Workspaces**。

   此时将启动 Data Engine 工作区页面，您可以在其中进行操作"创建工作区"。

9. 选择 **Configure OpenID Connect**。

   您需要"配置 OpenID 身份验证"以启用对 AIDE Console 的访问权限。

完成后

完成所有操作项目后，将隐藏指导窗口。您可以根据需要手动关闭并重新打开。

2.完成 **AIDE** 配置

AIDE 群集设置并与 AFX 存储系统集成后，需要完成 AIDE 配置。

步骤

1. "添加所需的 AIDE 许可证"。

   您必须安装 AI Data Engine 许可证才能获得完整的 AIDE 功能，包括矢量化和护栏功能。

2. "配置 OIDC 身份验证"。

   确保为所有部署配置 OIDC。您必须配置 OIDC 才能访问 AI Data Engine Console。

在您的 **AFX** 系统中安装 **AIDE** 许可证

您需要安装 NetApp 许可证才能访问 AI Data Engine (AIDE) 的全部功能。作为 ONTAP 群集管理员，您可以使用 ONTAP System Manager 执行许可证管理。

准备许可 **AI Data Engine**

许可证是一个或多个软件授权的记录。所有 AFX 许可证都以 NetApp 许可证文件 (NLF) 的形式交付，这是一个启用多个功能的单个文件。在 AFX 存储系统上安装许可证以支持 AIDE 之前，您应该考虑以下几点。

许可证类型

通过您的 AFX 存储系统开始使用 AIDE 需要两种主要类型的许可证。

ONTAP One 许可证

Metadata Engine 基本许可证通常作为 ONTAP One 许可证的一部分出厂安装。它允许访问 Metadata Engine 功能，为 AIDE 操作提供必要的基础。还包括管理 ONTAP 系统所需的所有核心功能。

AI Data Engine

您需要购买并安装 AI Data Engine 许可证，才能访问激活 AIDE 全部功能所需的高级服务。该许可证将解锁您的数据计算节点，启用矢量化、治理护栏、推理和集成 UI 体验等 AI 功能。许可证包括 GPU 计数和到期日期。

许可证安装要求

您需要购买 AIDE 许可证，并将相关的 NLF 文件下载到本地系统。然后，您可以通过 System Manager 将文件上传到 AFX 存储系统。另外，请确保您具备以下条件：

- 登录到 ONTAP System Manager 的管理员凭据
- 运行 ONTAP 9.18.1 及更高版本的 AFX 集群

在您的 **AFX** 系统上安装许可证

您可以安装 AIDE 许可证以激活 AFX 存储系统所需的其他 AIDE 功能。

步骤

1. 在 System Manager 中，选择 **Cluster**，然后选择 **Settings**。
2. 在 **Licenses** 旁边，选择 →。
3. 选择 **Features** 选项卡以显示可用的 ONTAP 功能。
4. 要安装许可证，请选择 **Installed licenses** 选项卡。
5. 选择 + Add。
6. 选择一个本地许可证文件，然后选择 **Add**。

相关信息

- "ONTAP 许可概述"
- "如何从 NetApp 支持站点下载 NLF 许可证"
- "ONTAP CLI：system license add"

在 ONTAP 中为 AIDE 配置 OpenID Connect

作为 ONTAP 集群管理员，您可以使用 ONTAP System Manager 为 AI Data Engine (AIDE) 集群配置 OpenID Connect (OIDC) 身份验证。这通过外部身份提供商 (IdP) 提供安全和集中的登录。

您必须配置 OIDC 才能访问 AI Data Engine Console。配置后，所有身份验证都将通过 OIDC 进行。如果未配置 OIDC，管理员以及数据工程师和数据科学家将无法使用控制台。在此情况下，登录到 System Manager 将恢复到本地身份验证。

还请注意以下有关用于 AIDE 访问的 OIDC 配置的信息：

- 您无法修改现有的 OIDC 配置。如果需要进行更改，请先删除配置，然后创建具有所需设置的新配置。
- 如果禁用或删除 OIDC，System Manager 将恢复为本地 ONTAP 用户身份验证。

OIDC 概述

OpenID Connect (OIDC) 是基于 OAuth 2.0 框架构建的身份验证协议。它通过添加身份层来扩展主要用于授权的 OAuth 2.0。OIDC 引入了 ID 令牌的概念，它是一个 JSON Web Token (JWT)，包含有关身份验证事件和用户身份的声明。

您需要选择和配置 AFX 与 AIDE 支持的外部身份提供程序 (IdP)。IdP 对用户进行身份验证，并发出令牌，AFX 可通过 System Manager 使用这些令牌来授予对 AIDE Console 的访问权限。

配置第三方身份提供程序

要使用 OIDC 进行身份验证，需要首先配置外部 IdP。ONTAP 实现的 OIDC 使用令牌中的角色声明来强制执行 RBAC。设置 IdP 时，请确保将其配置为在 id 令牌和访问令牌中返回角色声明。ONTAP 支持两个 IdPs 进行 OIDC 身份验证：Entra ID 和 Active Directory Federation Services (AD FS)。

Entra ID

您可以通过以下高级步骤配置 Entra ID：

1. 在 Entra ID 配置页面创建新的应用注册。
2. 将重定向 URI（Web）值设置为 `https://$CLUSTER_MGMT_IP/oidc/callback`，替换相应的集群管理 IP 地址或 FQDN。
3. 在应用角色下创建所需的角色，并将其分配给您的用户。
4. 将令牌声明更新到令牌配置下，以返回 id-token 和 access-token 中的角色。

有关更多信息，请参见 "使用 Entra ID 设置 OpenID Connect 提供程序"。

Active Directory Federation Services

您可以使用以下高级步骤配置 AD FS：

1. 创建新的应用程序组并选择 **Server application accessing a web API**。
2. 将重定向 URI（Web）值设置为 `https://$CLUSTER_MGMT_IP/oidc/callback`，替换相应的集群管理 IP 地址或 FQDN。
3. 配置声明以返回令牌中的角色。

有关更多信息，请参见 "将 AD FS 添加为 OpenID Connect 标识提供程序"。

在 **System Manager** 中配置 **OIDC**

配置 IdP 后，您可以在 System Manager 中设置 OIDC 身份验证，以启用对 AIDE Console 的安全访问。

开始之前

- 您需要具有 System Manager 的管理员访问权限。
- 必须配置并访问您的 OIDC 身份提供程序。

步骤

1. 在 System Manager 中，选择 **Cluster**，然后选择 **Settings**；找到 OpenID Connect 卡。
2. 如果已配置 OIDC，则可以编辑或禁用该配置。如果未配置 OIDC，请选择 开始设置过程。
3. 在 Configure OpenID Connect 下，提供以下字段的值：
   - 提供程序
   - 发行人
   - JSON Web 密钥集 URI
   - 授权端点

- Token 端点
- 结束会话端点
- 访问令牌颁发者（可选）

4. 在客户端配置下，为以下字段提供值：
   - 客户端 ID
   - 远程用户声明
   - 刷新间隔
5. 在"连接详细信息"下，为以下字段提供值：
   - 集群 IP 地址或 FQDN
   - 传出代理（可选）
6. 在外部角色映射下，选择现有角色映射或为 ONTAP `admin` 用户定义新角色。
7. 选择 立即启用 ，然后选择 保存 。System Manager 将刷新以应用新的身份验证设置。
8. 使用 IdP 凭据登录；身份验证成功后，您将返回到 System Manager。

相关信息

- "OpenID Foundation"
- "ONTAP OAuth 2.0 实施"

设置工作区

为 AI Data Engine 准备数据

创建集群后，建立一个数据容器，其中包含您打算与 AI Data Engine (AIDE) 一起使用的数据。此数据容器必须是 ONTAP 卷，可以是本地卷或来自运行 ONTAP 9 的对等 ONTAP 集群的卷。

您无需手动将 ONTAP 群集数据上传到 AIDE，而是需要将目标群集和 SVM 与 AIDE 群集对等，然后确定要将哪些 NFS 卷与 AIDE 元数据目录一起使用。创建数据容器后，可以创建一个工作区，并将数据容器与该工作区关联。然后，该工作区的用户可以访问其 AI 工作负载的工作区关联数据集和资源并与之交互。

关于此任务

必须对包含要用于 AI Data Engine 的数据的每个 SVM 进行对等。仅对等集群是不够的。这确保了 AI Data Engine 可以按预期访问、载入和索引您的数据。

您应该对等 AIDE 集群中将充当 SnapMirror 目标的 SVM。

您不需要在 ONTAP 集群和 AIDE 集群之间创建 SnapMirror 关系。这些关系将在工作区创建过程中自动创建。

开始之前

- 您需要 *storage administrator* 权限才能对等集群和 SVM 并选择数据容器。
- 您已确定包含要用于 AI Data Engine 的数据的 ONTAP 集群和 SVM。
- 您已确认数据源卷满足以下要求：
  - 卷处于联机状态且可访问。
  - 已启用 NFS 协议。仅支持启用 NFS 的卷作为 AI Data Engine 的数据容器。不支持 SMB 和 CIFS 卷。
  - 卷不是 FlexCache 卷。
  - 数据源是读写卷。不支持数据保护卷。

步骤

1. "对等每个 ONTAP 集群和 SVM" 包含要用于 AI Data Engine 的数据。
2. "选择要用于 AI Data Engine 的卷"。

   对于每个卷，请注意以下信息：

   - 卷名称
   - UUID
   - SVM 名称和 UUID
   - 集群名称和 UUID

下一步是什么？

"创建工作区" 并将您创建的数据容器与工作区相关联。

相关信息

- "数据迁移选项"

在 AI Data Engine 中创建工作空间

设置集群后，您可以创建工作区。工作区允许您对集群上的数据进行分段，控制个人的数据访问，并排除 AI Data Engine (AIDE) 不应访问的数据。

如果您管理存储，您将使用 ONTAP System Manager 创建和管理工作区。

组织根据团队、项目、数据敏感性级别或其他相关标准创建工作区。例如，如果您从事医疗保健工作，则可能会将临床数据细分为工作区，但遗漏了与 IT、法律或其他部门相关的数据。

关于此任务

系统处理限制会影响工作区创建（通常每个集群每天最多 15 GB）。如果您并行或快速连续创建多个工作区，则每个工作区可能需要更长的时间来处理，并且您可能会遇到严重的延迟。

从 Workspaces 清单页面监控工作区创建的状态。为了获得最佳效果，如果您需要立即访问这些功能，请避免同时创建多个工作区。

开始之前

- 您需要 *storage administrator* 权限才能创建工作区和关联数据集合。
- 您已经确定了要与工作区和 AI Data Engine 一起使用的远程（对等）和本地数据源。
- 您已"已创建至少一个数据容器"工作区可以使用的卷，例如本地卷或来自对等集群的卷。

> 将卷添加到您在该工作区的预期生存期内不会删除的工作区。如果在将某个卷添加到工作区后将其删除，则该工作区将进入失败状态。在建立工作区之前，请确认卷的长期可行性。

- 确保已在卷上启用 NFS，但未启用 CIFS。工作区仅支持使用 NFS 的卷。不支持具有 CIFS (SMB) 的卷。

创建工作区

创建工作区并关联包含要与 AI Data Engine 一起使用的数据的数据容器。

步骤

1. 在 ONTAP System Manager 中，导航到 **Data Engine > Workspaces**。
2. 选择 **Add**。
3. 在 **Add Workspace** 对话框中，至少选择一个可用数据容器以与工作区关联。
4. 配置 "对等集群"，以便可以在工作区内访问这些集群中的数据
5. 如果您想配置用户对工作区的访问权限，可以立即执行此操作或"等到创建工作区后"。
6. 配置刷新间隔，以确定工作区与关联数据容器同步以捕获新数据或更新数据的频率（例如，六小时）。

> 选择一个平衡数据新鲜度与系统性能的间隔。如果将数据容器添加到多个工作区，系统会自动使用最激进（最短）的时间间隔。如需了解更多信息，请参阅有关 工作区刷新和版本控制 的文档。

7. 选择 **Continue**。
8. 在 **Finalize workspace** 对话框中，输入工作区名称和描述。
9. 选择 **Add** 以创建工作区。

结果

工作区创建过程需要几分钟到几小时才能完成，具体取决于相关数据集及其文件数量、文件大小和其他因素。

系统会自动提取所有数据源的元数据，并将其存储在元数据目录中，用户可以使用该目录来查找项目所需的文件。将用户分配给工作区后，数据工程师用户可以从 AI Data Engine Console 设置工作区附属组件并与之交互。

新工作区以 `Creating` 状态显示在工作区页面上，直至流程完成且状态更改为 `ready`。

查看工作区详细信息

创建工作区后，请查看工作区详细信息。

步骤

1. 查看工作区详细信息，包括总大小、使用的集群容量百分比以及最近的工作区刷新日期。
2. 选择工作区名称以打开详细信息页面。
3. 在"概述"选项卡中，查看工作区详细信息，其中包括关联的数据容器、用户和活动。

工作区刷新和版本控制

每次工作区刷新都会创建一个不可变版本，以捕获工作区中所有文件和对象的当前状态。版本包括完整的元数据、对提取过程中使用的快照的引用以及用于可追溯性的作业 ID。这支持数据沿袭、可重复性和审核。

刷新根据您配置的计划（例如每六小时）或在手动触发时进行。支持的最小刷新间隔为一小时；最长为一年。如果数据容器包含在多个工作区中，则系统使用最频繁、最短的持续时间刷新间隔来安排元数据提取。

默认情况下，系统将保留以前、当前和下一个（正在进行的）版本。系统会根据组织的策略保留旧版本，并可以根据需要清除它们。

您可以列出工作区的所有版本并查看版本之间的差异，以确定添加、修改或删除了哪些文件或对象。这使您可以跟踪一段时间内的更改，并了解工作区数据的演变。

分配用户对 AI Data Engine 工作区的访问权限

作为存储管理员，您可以根据其角色将用户分配到工作区：数据工程师、数据科学家或其他角色，具体取决于组织的结构和需求。用户使用其凭据登录到 AI Data Engine (AIDE) Console，并访问分配的工作区内的数据容器资源。

ONTAP System Manager 使您能够管理哪些用户有权访问 AI Data Engine 工作区。您可以添加或删除用户以控制谁可以查看、修改工作区数据和活动并与之交互。

开始之前

- 您需要 *storage administrator* 权限才能管理用户对工作区的访问权限。

- 确认已创建工作区，并且该工作区在 Workspaces 清单中处于活动状态。
- 确认所有相关数据容器已添加到工作区并可访问。
- "确认为集群启用并配置 OIDC"。必须为每个相关数据工程师和数据科学家 IdP 用户或组完成从 IdP 到 ONTAP 角色的角色映射。

步骤

- 将用户添加到工作区：
  a. 在 ONTAP System Manager 中，导航到 **Data Engine > Workspaces**。
  b. 选择工作区名称以打开其详细信息页面。
  c. 转到 Users 选项卡。
  d. 选择 **Add** 按钮以打开添加用户对话框。
  e. 输入一个或多个用户的详细信息。以逗号分隔的 OIDC 用户列表形式输入详细信息。
  f. 选择 **Add** 以授予用户对工作区的访问权限。
- 从工作区中删除用户：
  a. 在工作区详细信息页面的"用户"选项卡中，找到并选择要删除的用户。
  b. 选择 **Remove** 按钮。
  c. 在对话框中确认删除。
  d. 系统立即删除用户并撤销其对工作区的访问权限。

结果

只有在工作区"用户"选项卡中列出的用户才能访问工作区数据和活动并与之交互。

管理和监控

监控集群进程

查看 **AIDE** 系统和集群状态

作为存储管理员，您可以使用 ONTAP System Manager 访问仪表板并显示集群状态。在开始您的 AIDE 管理任务或怀疑存在操作问题之前，这是一个很好的第一步。

开始之前

- 您需要 *storage administrator* 权限才能执行 AIDE ONTAP 相关的管理任务。

从控制面板监控 **AIDE** 状况和容量

1. 使用集群管理地址连接到 ONTAP System Manager：

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

2. 使用管理员帐户 Sign in。
3. 在左侧导航窗格中选择 **Dashboard**。
4. 查看 **Health** 图块：
   - 确认集群整体运行状况。
   - 验证 **Data compute nodes** 计数和状态。
   - 检查警报：
     - DCN 节点问题或连接问题
     - 错误的工作区或数据集（例如，集合发布失败）
5. 查看 **Capacity** 图块：
   - 记录集群总容量和已用容量。
   - 对于 AIDE 集群，请验证：
     - AIDE 元数据和应用程序卷使用的容量（元数据 Storage VM）
     - 工作区和数据集合使用的容量（如果可用）
6. （可选）查看 **Network** 和 **Performance** 磁贴，以了解可能影响 AIDE 工作负载的集群范围行为（例如，网络拥塞或保护延迟）。

查看数据 **DCN** 运行状况和利用率

1. 在导航窗格中，选择 **Cluster**，然后选择 **Overview**。
2. 选择 **Data compute** 选项卡。

   此选项卡显示集群中的所有 DCN 节点，其中包含：

   - 节点名称、型号、序列号和软件版本

- 总体节点状态
- CPU 和内存利用率
- GPU 利用率（如果存在 GPU）
- 任何节点级错误指示器

3. 展开 DCN 节点以打开详细视图并检查：
   - 系统 CPU 和内存使用率
   - GPU 内存使用量
   - 报告的硬件或服务问题
4. 在 集群 **>** 概述 页面上选择 布线，以验证 DCN 节点已正确连接到集群交换机，并识别任何端口或链路问题。

监控工作区和元数据占用空间

1. 在导航窗格中，选择 **Data engine**，然后选择 **Workspaces**。
2. 查看页面顶部的工作区摘要：
   - 工作区及其状态的计数（例如， `Processing`、 `Healthy`、 `Error`）。
   - 工作区总大小。
   - 所有工作区消耗的集群容量百分比。
3. 查看工作区网格：
   - 确认关键工作区显示 **Healthy** 状态。
   - 检查工作空间大小和容量消耗。
   - 查找处于 `Error` 或长期运行 `Processing` 状态的任何工作区。
4. 要查看特定工作区的详细信息，请选择其名称：
   - 在 概览 选项卡上，确认：
     - 工作区状态和大小
     - 包含的数据容器（卷）及其项目计数
     - 每个数据源的上次更新时间
   - 在 **Data collections** 选项卡上，确认：
     - 该工作区存在哪些数据集（数据集在 System Manager 中是只读的）
     - 它们的状态、大小和最近更新时间
   - 在 用户 选项卡上，检查哪些 AI Data Engine Console 用户具有访问权限。

监控元数据 **Storage VM** 和 **AIDE** 管理的保护

1. 在导航窗格中，选择 **Cluster**，然后选择 **Storage VMs**。
2. 找到带有子类型 `data-engine`（元数据 SVM）的 Storage VM：
   - 确认元数据 SVM 已联机。

- （可选）打开其详细信息以查看以下项的计数：
  - 卷
  - 带类型的 LIF `Data compute network`（用于 DCN-ONTAP 通信）

3. 选择 **Protection**，然后选择 **Relationships** 以查看工作区中使用的远程数据源的保护：
   - 通过命名模式识别 AIDE 创建的 SnapMirror 关系：
     - 目标卷： `<source_volume_name>_dest_<source_volume_UUID>`
     - 政策： `<source_volume_name>_dest_aide_policy_<source_volume_UUID>`
   - 使用此视图可验证关系是否正常，以及延迟时间是否与工作区刷新预期保持一致。

> 请勿直接在 ONTAP 中修改元数据 Storage VM、AIDE 创建的 SnapMirror 关系或 AIDE 管理的快照（或其计划）。更改可能会中断 AIDE 版本历史记录。"调整工作区刷新设置"如果需要调整刷新行为。

查看与 **AIDE** 相关的警报和通知

1. 在导航窗格中，选择 **Events & Jobs**，然后选择 **System alerts**。
2. 查看与以下内容相关的任何活动警报：
   - DCN 节点健康或连接
   - 数据引擎网络问题
   - 工作区或数据收集错误
   - ONTAP 和 DCN 集群之间的软件版本不匹配
3. 根据需要，在 **Cluster > Settings > Notification management** 中配置通知目标（例如，email、syslog），以确保将与 AIDE 相关的警报转发到您的操作工具。

相关信息

"准备管理 AFX 存储系统"

查看见解以优化您的 **AIDE** 系统

作为存储管理员，您可以使用 ONTAP System Manager 的 *Insights* 功能显示符合 NetApp 最佳实践的建议配置更新。这些更改可以优化您的 AIDE 群集的安全性和性能。

关于此任务

每个见解都会在页面上显示为单独的图块或卡片，您可以选择实施或取消。您还可以选择相关文档链接，以了解有关特定技术的更多信息。

步骤

1. 在 System Manager 中，选择 **Analysis**，然后选择 **Insights**。
2. 查看可用建议。

下一步

执行任何建议的操作以实施配置最佳实践。

查看 AIDE 系统事件、作业和审核日志

作为存储管理员，您可以查看 AIDE 生成的事件、作业和审核日志消息，以跟踪内部处理并诊断潜在问题。AIDE 系统可以配置为转发此信息以及其他相关数据，以进行额外的处理和存档。

开始之前

- 您需要 *storage administrator* 权限才能执行 AIDE ONTAP 相关的管理任务。

监控 AIDE 活动、事件和作业

您可以使用集中式*Activity*视图来监控所有工作区中特定于 AIDE 的事件和作业，或查看个别工作区的活动。

查看集群范围的 AIDE 活动

监控工作区操作，对元数据提取问题进行故障排除，并跟踪整个 AIDE 部署中的数据收集发布。

1. 在 ONTAP System Manager 中，在导航窗格中选择 **Data engine**，然后选择 **Activity**。
2. 选择 **Events** 选项卡：
   - 查看最近的 AIDE 特定事件，例如：
     - 工作区创建、更新或删除
     - 数据容器添加/删除操作
     - 数据收集发布（如果存在）
   - 使用筛选器（按严重性、对象类型、工作区或时间范围）专注于活动或关键事件。
3. 选择要打开和查看的单个事件：
   - 说明和时间戳
   - 受影响的工作区、数据收集或数据源
   - 建议采取的措施（如有）
4. 选择 **Jobs** 选项卡：
   - 监控长期运行的作业，例如：
     - 工作区的初始元数据提取
     - 工作区刷新 / 目录更新作业
     - 数据收集发布或刷新作业
   - 检查作业状态和进度。
5. 选择作业以打开 peek 视图并查看：
   - 开始和结束时间
   - 进度百分比和阶段（例如 **Scanning**、**Publishing**）
   - 受影响的工作区、数据收集或数据源
   - 失败作业的错误消息

查看工作区特定活动

要对特定工作区进行故障排除，请打开工作区详细信息（**Data engine > Workspaces**，选择工作区），然后使用 **Activity** 选项卡：

- 仅查看范围为该工作区的事件和作业。
- 使用此视图可隔离单个工作区卡在 `Processing` 状态等问题。

查看集群范围的事件

查看事件消息，获取有价值的系统活动记录。每个事件都包括描述和唯一标识符以及建议的操作。

1. 在 ONTAP System Manager 中，选择 **Events & jobs**，然后选择 **Events**。
2. 查看并响应页面顶部的建议操作，例如启用自动更新。
3. 选择 **Events log** 选项卡以显示消息列表。
4. 选择一个事件消息以更详细地检查它，包括序列号、描述、事件和建议操作。
5. （可选）选择 *Active IQ 建议*选项卡并注册 Active IQ，以获取集群的详细风险信息。

查看集群范围的作业

查看 AIDE 集群上运行的所有作业，包括 AIDE 特定作业和常规 ONTAP 作业。

1. 在 ONTAP System Manager 中，选择 **Events & Jobs**，然后选择 **Jobs**。
2. 根据需要自定义显示以及搜索和下载作业信息。

查看审核日志

使用审核日志，根据 HTTP 等访问协议的使用情况，查看系统活动记录。

1. 在 ONTAP System Manager 中，选择 **Events & jobs**，然后选择 **Audit logs**。
2. 选择 **Settings** 以启用或禁用所跟踪的操作。

管理通知

配置通知目标以自动转发 AIDE 事件和审核日志。

步骤

1. 在 ONTAP System Manager 中，选择 **Cluster**，然后选择 **Settings**。
2. 导航到 **Notification management** 并选择 ⋮。
3. 选择适当的操作以查看或配置 AIDE 使用的目标：
   a. 事件目标：选择 **View event destinations**
   b. 审核日志目标：选择 **View audit destinations**
4. 根据需要选择 添加 ，并提供目的地信息。
5. 选择 **Save**。

相关信息

- "ONTAP 事件、性能和运行状况监控"

管理 AI Data Engine 工作区

工作区是 AI Data Engine (AIDE) 用于构建和刷新特定项目或用例的元数据目录的一组数据源（卷）。作为存储管理员，您可以使用 ONTAP System Manager 监视工作区运行状况、调整配置、控制数据源、管理用户，并在不再需要工作区时删除工作区。

开始之前

- 您需要 *storage administrator* 权限来管理工作区。

查看工作区状态

查看工作区运行状况、容量使用情况和元数据状态，以确保 Metadata Engine 按预期运行，并且不会消耗意外资源。

步骤

1. 从 ONTAP System Manager 中，在导航窗格中选择 **Data engine > Workspaces**。
2. 查看页面顶端的摘要，了解工作区总数、整体工作区运行状况和容量使用情况。
3. 对于工作区特定信息，请选择工作区名称。在 概览 选项卡上确认：
   - 工作区状态和大小。
   - 工作区中包含的数据容器（卷）。
   - 每个数据源的项目计数和上次更新时间。
   - 任何工作区级别的警告。

   如果工作区或数据集合显示错误状态，请验证所有源卷均处于联机状态且可访问。
4. 选择 **Data collections** 选项卡以查看：
   - 与此工作区关联的所有数据集。
   - 状态（如 `Published` 或 `Error`）、大小和上次更新时间。

   System Manager 对数据收集是只读的。数据工程师可以在 AI Data Engine Console 中创建和管理数据收集。
5. 选择 **Users** 选项卡以查看：
   - 有权访问此工作区的用户列表。
6. 选择 **Activity** 选项卡以仅查看与此工作区相关的事件和作业。

编辑工作区属性和刷新计划

您可以调整工作区的名称、描述、刷新间隔以及（如果已获得许可）其护栏策略。

步骤

1. 从 **Data engine > Workspaces** 中，选择工作区旁边的 ⋮ 并选择 **Edit**。
2. 编辑工作区属性：
   - 根据需要更新 **Name** 和 **Description**。
   - 在允许的范围内（小时和天）调整 **Refresh interval**（元数据更新频率）。
   - 如果安装了 AIDE 许可证，您可以选择 **Guardrail policy**。
3. 选择 **Save**。

ⓘ 对刷新间隔或元数据处理的更改可能会影响为此工作区更新远程 SnapMirror 关系的频率。

向现有工作区添加数据容器

您可以添加其他已装入的卷（本地或来自对等远程集群），以便其元数据包含在工作区目录中。

1. 从 **Data engine > Workspaces** 中，执行以下操作之一：
   - 选择工作区旁边的 ⋮，然后选择 **Add data containers**。
   - 打开工作区，选择 **Overview** 选项卡，然后在数据容器部分中选择 **Add**。
2. 在*将数据容器添加到工作区*对话框中：
   - 在 AIDE 集群上查找本地卷。
   - 扩展对等集群以选择远程卷（远程卷需要集群和 SVM 对等）。

   ⓘ 只能选择符合条件的在线卷，这些卷未被全局排除且尚未属于工作区。
3. 如果系统提示您进行远程卷映射：
   - 选择 AIDE 集群上的目标 Storage VM 以接收所选远程卷的 SnapMirror 目标。
4. 选择 **Add**。
5. 使用工作区 **Activity** 选项卡或 **Data engine > Activity** 跟踪元数据提取和新数据源的任何 SnapMirror 初始化。

从工作区中删除数据容器

当数据容器不再与工作区的目的相关时，或者如果要缩小该工作区的元数据管理范围，则可以删除该数据容器。删除数据容器会停止该卷的元数据刷新，并从元数据目录中删除其元数据。

ⓘ 请勿从 ONTAP 中删除已添加到工作区的源卷。如果删除卷，工作区将进入失败状态。在删除任何底层 ONTAP 卷之前，请务必先从工作区中删除数据容器。

步骤

1. 导航到 **Data engine > Workspaces**，然后选择包含此数据容器的工作区。
2. 在 **Overview** 选项卡上，找到要删除的数据容器。
3. 选择数据容器旁边的 **Remove**。

4. 查看确认对话框并选择 **Remove**。

> 从工作区中删除数据容器不会删除底层 ONTAP 卷或其 SnapMirror 关系。它仅影响 AI Data Engine 中的元数据使用。

管理工作区用户

您可以授予或撤销数据工程师和数据科学家用户对工作区的访问权限。这些用户在您的身份提供程序 (OIDC) 中定义并映射到 ONTAP 角色。请参见 "将用户分配到工作区" 文档以了解如何管理用户访问。

删除工作区

您可以删除工作区以删除工作区定义和相关的 AIDE 元数据。与工作区相关的任何数据集和矢量嵌入也将被删除。

> 不会删除基础 ONTAP 数据（卷、SnapMirror 关系）。

步骤

1. 在 **Data engine > Workspaces** 中，执行以下操作之一：
   - 删除单个工作区，选择 ⋮ 并选择 **Delete**。
   - 删除多个工作区，选中工作区的复选框，然后选择 **Delete**。
2. 在确认对话框中，在继续之前查看操作的影响：
   - 工作区元数据已永久删除。
   - 与工作区关联的数据收集和嵌入将被永久删除。

   > 没有软删除或还原选项。
3. 选中复选框以确认您的理解，然后选择 **Delete**。

相关信息

- "将用户分配到工作区"

升级和维护您的 AIDE 系统

AI Data Engine 系统更新和兼容性

保持 AI Data Engine (AIDE) 系统组件的最新状态，以保持最佳性能和访问新功能。部署后、新软件或固件可用时、添加或替换节点时或定期进行功能更新时更新组件。

AIDE 系统组件

AIDE 系统中有两个主要组件需要更新：ONTAP 软件和 AIDE 软件，其中包括 DCN 固件。

ONTAP 软件

ONTAP 是在 NetApp 存储系统上运行的操作系统，包括在 AIDE 部署中使用的系统。保持 ONTAP 为最新版本，以维护系统稳定性、安全性以及与 AIDE 组件的兼容性。AIDE 组件单独更新。

AIDE 软件更新

AIDE Console 软件和 DCN 固件更新作为单个软件包 (.tgz) 一起分发，不会嵌入到 ONTAP 映像中。更新可确保 AIDE 系统内硬件组件的正常运行，并提供新功能、性能改进和错误修复。

了解更新流程

AIDE 软件更新可通过 ONTAP System Manager 进行管理。

AIDE 不支持 ONTAP 自动软件更新功能。您可以注册接收来自 "NetApp 支持站点下载" 的通知，但 AIDE 软件的所有更新都由管理员手动执行。

发布类型和范围：

- ONTAP 主要版本（9.x.x）和 AIDE 主要版本（9.x.x U0）引入了影响 ONTAP 集成的新功能、API 或更改。
- ONTAP 修补程序版本 (9.x.x Px) 和 AIDE 更新版本 (9.x.x Ux) 包含不影响 ONTAP 集成的修复和更新。

兼容性矩阵

在计划更新时确保 ONTAP 和 AIDE 软件之间的兼容性。

AIDE software 作为"U"版本发布。AIDE 主要版本是"U0"版本，后续次要版本将是"U1"及更高版本。

ONTAP 和 AIDE 兼容性

AIDE 发布	支持的 ONTAP 版本
9.18.1 U0	9.18.1 GA 和所有 9.18.1 Px
9.18.1 U1 及更高版本	9.18.1 GA 和所有 9.18.1 Px

"Px" 表示主要版本中的所有 ONTAP 补丁版本（例如，9.18.1 P1、9.18.2 等）。

AIDE 升级路径

以下示例使用假设的未来版本来显示从 AIDE 9.18.1 U0 和 9.18.1 U1 允许的升级和更新路径。

如果您当前的 AIDE 版本是…	您的目标 AIDE 版本是…	您的升级或更新路径是…
9.18.1 U0	9.18.1 U1	直接
9.18.1 U0	9.18.1 U3	直接（您可以在 9.18.1 中从任何 Ux 更新到任何以后的 Ux）
9.18.1 U1	9.18.1 U3	直接（您可以在 9.18.1 中从任何 Ux 更新到任何以后的 Ux）

还原限制

AIDE 系统不支持 DCN 固件、AIDE 软件更新或 AFX 存储系统上的 ONTAP 的还原操作。安装更新或升级后，无法还原到以前的版本。请在更新或升级之前查看 兼容性要求。

相关信息

- "升级 AFX 系统的 ONTAP 软件"
- "更新 AIDE 软件"

更新 AI Data Engine software

作为存储管理员，您可以使用 ONTAP System Manager 更新 AIDE 系统上的 AI Data Engine (AIDE) 软件、数据计算节点 (DCN) 固件和其他系统文件。

AIDE 不支持 ONTAP 自动软件更新功能。您可以注册接收来自 "NetApp 支持站点下载" 的通知，但 AIDE 软件的所有更新都由管理员手动执行。

关于此任务

合并的 AI Data Engine software 包比典型的 ONTAP 更新包大得多（AIDE 包约为 40GB）。在更新 AIDE software 时，请规划更长的上传和安装时间。

开始之前

- 您需要 *storage administrator* 权限才能更新 DCN 固件和 AI Data Engine software。
- 您需要使用 NetApp 支持站点凭据才能使用活动帐户。
- "在规划更新时确保 ONTAP、DCN 固件和 AI Data Engine software 之间的兼容性"。

> DCN 固件或 AI Data Engine software 更新不支持还原功能。安装更新后，无法还原到以前的版本。

步骤

1. "将组合的 DCN 固件和 AIDE 软件更新文件下载到本地客户端"。
2. 在 System Manager 中，选择 **Cluster > Settings > Software updates**。
3. 在 软件更新 旁边，选择 →。
4. 在 AI Data Engine 更新下，选择 添加 **AI Data Engine software** 文件，然后选择更新包。
5. 包上传完成后，选择 更新 开始在 DCN 节点上安装更新。

> DCN 节点更新正在进行时，AIDE Console 不可用或不可访问。

结果

DCN 使用 AI Data Engine software 进行更新，并为每个节点显示更新的版本。

相关信息

- "升级 AFX 系统的 ONTAP 软件"

将数据计算节点添加到您的 **AIDE** 集群

您可以在创建新的 AI Data Engine (AIDE) 集群或扩展现有集群时添加数据计算节点 (DCN)。工作流程包括使用 ONTAP System Manager 发现和配置节点。

准备添加节点

添加 DCN 节点时有几个注意事项。

创建新 **AIDE** 集群时

新的 AIDE 集群需要恰好三个 DCN 节点。

硬件安装和可寻址性

确保满足以下先决条件：

- 新的 DCN 硬件安装在机架上，接通电源，并连接到集群交换机。
- 您有一个可用于 DCN 到 ONTAP 后端子网的 IP 地址空间范围。
- ONTAP 集群已初始化，可从集群管理 LIF 访问。

System Manager 凭据

需要 *storage administrator* 权限才能执行 AIDE 群集创建或扩展任务。

软件兼容性

请查看以下文档，以确认您的 DCN 硬件和软件版本与 ONTAP 集群兼容：

- "DCN 软件与 ONTAP 版本兼容"。

在节点添加操作期间，System Manager 将确认新节点运行与以下内容兼容的软件版本：

- 如果这是第一个 DCN 加入，则为 ONTAP 集群有效版本 (ECV)。
- 如果已存在 DCN，则为现有 DCN 集群版本。

如果节点不兼容：

- *添加*对话框中受影响的 DCN 旁边会显示错误。
- 您必须首先将 DCN 软件（或 ONTAP，如适用）更新为兼容版本。

添加数据计算节点

在创建新 AIDE 群集或扩展现有群集时，您需要添加 DCN 节点。

步骤

1. 在 System Manager 中，选择导航窗格中的 **Dashboard**，然后选择 **Health** 卡。
2. 确认有要添加的节点，然后选择 **View details** 以显示此列表。

   该列表包含尚未属于 AIDE 集群的已发现节点

3. 或者，您可以选择 **Cluster** 和 **Overview** 以及 **Data compute** 选项卡来查看列表。
4. 在数据计算页面的底部，选择节点列表上方的 添加 。
5. 在*添加数据计算节点*对话框中，选择要添加的 DCN 节点。

   您可以选择在添加节点之前重命名单个节点。

6. 如果这是您第一次添加节点，并且不存在后端子网，请选择 添加子网 并提供：
   - 子网名称（供内部使用）
   - 子网地址和掩码
   - 此后端网络上 DCN 和 ONTAP 节点的 IP 地址范围

     System Manager 验证该范围包括要添加的所有 DCN 和集群中所有 ONTAP 节点的足够可用 IP 地址，以及用于 DCN 到 ONTAP 通信的其他集群级浮动 IP 地址。

7. 无论您是添加了后端子网还是它已经存在：
   a. 查看可用的 IP 地址。
   b. 如果需要，选择 编辑子网 并扩展 IP 范围。
      - 您只能扩大范围。不支持缩小或更改子网。
      - 更改子网或 IP 范围可能需要在 DCN 上重新创建底层 Kubernetes 集群，这可能需要几分钟的时间。
8. 可选择通过提供以下功能来配置 Data Engine 服务接口：
   - 服务 IP 地址
   - 网络掩码
   - 网关（如果您的环境需要）

     单个 IP 将在 DCN 之间进行负载平衡，并用作 AI Data Engine Console 和相关 API 的前端地址。

9. 查看选定节点、后端子网和 Data Engine 服务接口设置。
10. 选择 添加 并等待操作完成。System Manager 将执行以下操作：
    - 将所选节点添加到 DCN 集群
    - 配置后端网络并将节点加入基于 Kubernetes 的 DCN 集群
    - 更新 DCN 发现的内部元数据
11. 完成后，选择 **Cluster** 和 **Overview** 并确认：
    a. 在 **Data compute** 下，新 DCN 显示为集群的一部分
    b. 所有节点均为 **Healthy**
    c. 验证仪表板 **Health** 卡显示更新的节点计数

替换 **AIDE** 集群中的节点

如果 AI Data Engine (AIDE) 集群中的数据计算节点 (DCN) 停止运行或由于硬件故障、升级或维护需要更换，则需要更换该节点。这确保了 AIDE 集群保持健康和运行。可以在不

中断正在进行的服务的情况下执行该程序。

准备替换节点

在替换 AIDE 集群中的节点之前，需要考虑几个事项。

System Manager 凭据

需要 *storage administrator* 权限才能执行 AIDE 群集节点替换任务。

限制

替换 AIDE 集群中的节点时，应注意以下限制：

- 仅使用 CLI 和可选的 REST API 支持节点更换。
- 您无法使用 System Manager 执行节点替换。
- 新节点应与集群的软件版本相匹配；如果需要，ONTAP 将对其进行更新。
- 连接到集群网络时，不得打开故障节点，以避免 IP 地址冲突。

要求

您需要具备以下条件：

- 新替换节点的序列号

替换 **AIDE** 集群中的 **DCN** 节点

您可以使用以下过程替换 AIDE 集群中的 DCN 节点。

步骤

1. 物理移除故障节点

   关闭电源并断开节点与群集网络的连接。在更换过程中，请确保节点未在网络上启动。

2. 使用以下命令从集群中删除失败的节点：

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

   提供 <node_name> 值的实际名称。

3. 将新节点物理连接到集群

   确保节点已接通电缆、已接通电源且可被发现。

4. 查看可发现和未配置的节点，以验证新节点是否在线：

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

5. 使用以下命令将该节点添加到群集中：

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

ONTAP 将为新节点分配 IP 地址。如果节点的软件版本与集群不匹配，ONTAP 将自动更新节点。

6. 使用以下命令之一验证集群健康和节点集成：

```
dcn cluster node show

dcn cluster node show -instance
```

相关信息

- "扩展计算集群"

管理向量化和数据集合

AI Data Engine 的数据到 RAG 快速入门

使用此工作流从新部署的 AI Data Engine (AIDE) 系统转到工作检索增强生成 (RAG) 端点。了解存储管理员、数据工程师和数据科学家如何使用 ONTAP System Manager 和 AIDE Console 进行协作。

开始之前

- 您已安装数据计算节点 (DCN) 并将其添加到 ONTAP 集群。
- 您已经安装并许可了用于矢量化和护栏的 AI Data Engine software。
- 您已配置 "OpenID Connect (OIDC)" 并映射了管理员、数据工程师和数据科学家角色。


定义数据范围和治理

作为存储管理员或安全管理员，您希望在 AIDE Console 和 ONTAP System Manager 中准备环境：

- "创建一个或多个工作区"从本地和远程数据源。
- "配置分类器和护栏策略" 在 AIDE Console 中。
- "为数据工程师和数据科学家分配工作区访问权限"。


探索工作区元数据

作为数据工程师或数据科学家，您希望使用 AIDE Console 探索工作区元数据：

- "探索工作区元数据" 以了解可用内容。
- 定义应为 RAG 提供数据的一个或多个逻辑子集（例如，支持文章、产品手册或匿名临床备注）。


创建和发布数据收集

作为数据工程师或数据科学家，您希望将所选子集转换为 RAG 就绪集合：

- "创建数据收集"使用选定的过滤器从工作区。
- "发布数据收集" 并监控索引，直到它达到 `Ready` 状态。
- 复制所选集合的检索端点 URI，并提供给数据科学家或应用程序开发人员。
- "查看数据收集状态和向量占用空间" 根据需要。

下一步是什么？

- "在 AI Data Engine 中定义您的数据资产和护栏策略"
- "在 AI Data Engine Console 中探索工作空间元数据"
- "在 AI Data Engine Console 中创建数据集合"

在 AI Data Engine Console 中探索工作空间元数据

作为数据工程师或数据科学家，您在 AI Data Engine (AIDE) 中的首要任务是了解工作区中可用的数据。您可以使用 AIDE Console 查询元数据目录、搜索相关文件并识别要转换为数据集的数据子集。

开始之前

- 您需要在 AI Data Engine Console 中拥有 *data engineer* 或 *data scientist* 权限，并可以访问至少一个工作区。
- 存储管理员有：
  - 已在 ONTAP System Manager 中创建一个或多个工作区。
  - 已将您的用户或组访问权限分配给相关工作区。
- 工作区的元数据提取已完成，工作区处于 `Ready`状态。
- 已启用分类器，以便元数据包括分类标签（例如 PII 指标）。

以数据工程师或数据科学家身份登录 AIDE Console

步骤

1. 在浏览器中，导航到 AIDE Console URL：

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

2. 通过您组织的 OIDC 提供程序进行身份验证。
3. 确认您的角色被认可为数据工程师或数据科学家（例如，通过可用的工作区和数据收集操作）。有关详细信息，请参见 "AIDE 角色文档" 了解数据工程师和数据科学家如何使用 AIDE 组件。

结果

您已登录 AIDE Console，仅查看已授予您访问权限的工作区。

查看可访问的工作区

步骤

1. 在 AIDE Console 中，导航到 **Data Curator > Workspaces**。
2. 查看您可以访问的工作区列表。
3. 选择工作区以打开其详细信息。

结果

现在，您可以看到存储管理员为您的项目提供的数据资源的工作区范围视图。

下一步是什么？

- "从工作区创建 RAG 数据集"

在 AI Data Engine Console 中创建数据集合

数据收集是 AI Data Engine (AIDE) 中的核心 RAG 构建块。作为数据工程师或数据科学家，您可以定义哪些文件属于集合，配置嵌入和索引选项，并发布集合，以便应用程序可以通过检索端点查询它。

您将在 AI Data Engine Console 中执行所有数据收集任务。

开始之前

- 您需要在 AI Data Engine Console (`https://<cluster_management_ip>/console` 中拥有 *data engineer* 或 *data scientist* 权限）。
- 您可以访问至少一个已提取元数据且处于 `Ready` 状态的工作区。
- 您已浏览了工作区元数据，并确定了定义有意义的数据子集的查询或筛选器。
- 已安装 AI Data Engine software 许可证并启用推理功能。

从工作区元数据创建数据收集

步骤

1. 导航到 **Data Curator > Workspaces**，然后选择包含目标数据的工作区。
2. 选择 **Add data collection**。
3. 在 **Create new data collection** 页面中，执行以下操作：
   a. 输入集合的名称和描述（例如， `Support_KB_RAG_EN`）。
   b. 选择集合是否应为：
      - **Dynamic**：根据您定义的过滤标准，自动识别新文件并将其添加到数据集合中。这发生在工作区刷新期间。
      - 静态：您可以选择集合中包含的文件。如果数据集合处于 `draft`状态，则可以编辑文件。数据集合进入 `Published`状态后，无法编辑。
4. 指定源子集：
   a. 使用关键字和筛选器（文件类型、时间戳和其他属性）查找要包含的相关文件。

   您可以选择文件名以打开内容的预览窗口。

5. 将这些文件添加到数据集合中。
6. 选择 **Save** 以完成收集。

结果

您已经定义了数据收集的范围，并向其中添加了所需的文件。当您发布收集时，AIDE 会生成嵌入并构建向量索引。

创建小而有针对性的集合（例如，每个用例或域），而不是单个"全部"集合。这提高了检索的相关性和可管理性。

发布数据收集

发布数据集，使其可由 AI 应用程序通过 RAG 检索端点进行查询。发布从所选文件生成矢量嵌入并将其编入索引以进行语义搜索。集合达到 `Ready`状态后，其端点可供数据科学家集成到笔记本、管道和 AI 应用程序中，以进行检索增强生成 (RAG) 和搜索。

对于大型集合，请考虑在非高峰时段安排初始发布和主要重新发布，以最大限度地减少资源争用。

步骤

1. 导航到 **Data Curator > Data collections**，然后选择数据收集的选项菜单（•••）。
2. 选择 **Publish**。
3. 选择默认或自定义优化配置。
4. 选择 **Publish** 以启动数据转换。
5. 在 AIDE Console 中，打开集合详细信息视图（**Data Curator > Data collections**）以获取状态更新。

结果

集合达到 `Ready` 状态，可供下游应用程序和数据科学家使用。

从 **Data Curator >** 数据集合 中，您可以选择 复制 **URI** 以获取使用 API 访问数据集合所需的信息。

更新或删除数据收集

随着时间的推移，您可能需要优化或淘汰数据集合。优化集合可能涉及调整筛选器以添加或删除文件、更改嵌入设置或更新集合描述。删除集合将永久删除它，并使其检索端点不可用。

更新数据收集

您可以在数据集处于 `draft` 状态时对其进行更新。

步骤

1. 导航到 **Data Curator >** 数据集合。
2. 选择要修改的集合。
3. 选择 **Edit**。
4. 调整以下任意选项：
   - 名称和描述
   - 筛选器（路径、文件类型、分类标签）。
   - 嵌入和分块设置。
5. 保存更改。
6. 重新发布集合，使新定义和嵌入生效。

结果

新的索引作业将使用更新的配置运行，完成后集合将返回到 `Ready` 状态。

删除集合

删除集合是永久性的。在删除集合之前，请确保没有任何生产应用程序仍然依赖于集合的检索端点。

步骤

1. 导航到 **Data Curator > Data collections**，然后选择集合的选项菜单 (•••)。
2. 选择 **Delete**。
3. 确认删除。

结果

集合定义及其嵌入已从 AI Data Engine 中删除。删除集合后，尝试查询前一个检索终结点的应用程序将失败。

下一步是什么?

- "查看数据集合"

在 AI Data Engine 中查看数据收集

数据工程师或数据科学家从工作空间创建和发布数据集合后，您需要了解其状态、大小以及对 AI Data Engine 集群的影响。

如果您是存储管理员、数据工程师或数据科学家，则可以通过 ONTAP System Manager 和 AIDE Console 查看数据收集。

开始之前

- 您需要 ONTAP System Manager 中的 *storage administrator* 权限或 AI Data Engine Console (`https://<cluster_management_ip>/console` 中的 *data engineer* 或 *data scientist* 权限才能查看数据集合。
- 至少存在一个已成功提取元数据的工作区。
- 数据工程师或数据科学家已经从 AI Data Engine Console 创建并发布了至少一个数据集合。
- 已安装 AI Data Engine software 许可证并启用推理功能，因此矢量化和检索端点处于活动状态。

查看集群范围内的数据收集

对于存储管理员，ONTAP System Manager 提供数据集及其占用空间的集群范围视图，但不允许管理员创建或修改它们。

步骤

1. 在 System Manager 中，导航到 **Data Engine > Data collections**。
2. 查看页面顶部的库存摘要：
    - 按状态分类的数据收集总数
    - 矢量数据库在所有集合中消耗的总空间
    - 矢量空间占总集群容量的百分比
3. 选择单个数据收集并查看：

- 集合名称和描述
- UUID
- 关联工作区
- 状态
- 集合大小
- 创作者
- 上次刷新时间

结果

现在，您可以对集群中的所有数据集及其存储影响进行高级查看。使用此视图可识别大型、过时或处于未就绪状态的集合。

您还可以查看是否正在主动更新单个数据集，以及是否有任何故障阻止了 RAG 使用。

监视与集合相关的作业和事件

作为存储管理员，您可以从集群范围的 **Activity** 页面和工作区详细信息监控构建和更新集合的作业。

步骤

1. 在 System Manager 中，导航到 **Data Engine > Activity**。
2. 在 **Events** 选项卡上：
   a. 按类型（例如，workspace、data collection）或严重程度筛选。
   b. 展开与数据集合相关的任何事件（例如"数据集合发布失败"）以查看更多详细信息。
3. 在 **Jobs** 选项卡上：
   a. 筛选以专注于数据收集索引和发布作业。
   b. 对于每个作业，打开 peek 视图以查看：
      - 进度百分比。
      - 开始和结束时间。
      - 任何报告的错误消息或警告。
4. （可选）导航回受影响的工作区（**Data Engine >** 工作区）并打开其*活动*选项卡，以查看仅限于该工作区的事件和作业。

结果

您可以跟踪数据收集的生命周期，识别停滞或失败的作业，并收集上下文信息以传递给数据工程师、数据科学家或支持人员。

当数据收集在较长时间内保持 `Publishing` 状态时，在假定失败之前，请在 Activity 页面中检查相应的长期运行作业。

从 **AIDE Console** 查看数据收集

数据工程师和数据科学家通常直接从 AIDE Console 监控数据集合，数据集合在此处创建和发布。

步骤

1. 以数据工程师或数据科学家身份登录 AIDE Console。
2. 导航到 **Data Collections** ，然后选择所需的数据集。
3. 对于每个集合：
   a. 检查状态（(`Draft`、 `Publishing`、 `Ready` 或 `Failed`）。
   b. 选择数据收集名称以查看定义详细信息（筛选器、包含的文件类型、分类器选项、嵌入设置）。
   c. 检查上次发布或更新的时间戳。
4. 如果需要，打开作业详细信息或日志（如果可用）以了解失败或未完成的运行。

结果

数据工程师和数据科学家可以迭代收集定义并再次发布，同时监控状态和运行状况，而无需存储管理员参与。

下一步是什么？

- "在 AIDE Console 中为 RAG 创建数据集合"

实施护栏

在 AI Data Engine 中为您的数据资产定义 Data Guardrails 策略

作为数据或平台所有者，您可以使用 AI Data Engine (AIDE) Console 来定义哪些数据属于 AI 的范围，哪些数据始终处于禁区，以及当该数据用于分类和检索增强生成 (RAG) 时适用哪些安全规则。

使用这些过程在 AIDE Console 中定义这些策略，以便 ONTAP System Manager 可以对工作区中的所有数据实施这些策略。

开始之前

- 您需要在 AI Data Engine Console (`https://<cluster_management_ip>/console` 中具有_存储管理员_权限才能创建和管理全局策略。
- 您有一个部署了健康数据计算节点的 AIDE 集群。
- "OpenID Connect (OIDC)" 已配置，并且您的 IdP 角色映射到允许数据策略管理的 AIDE 管理员角色。
- 已安装 AI Data Engine software 许可证，以便启用 Data Guardrails 和推理功能。
- 至少存在一个工作区，或者您已与管理员协调以了解将在工作区中使用哪些数据源（卷）。

了解策略类型

AIDE Console 公开了塑造数据资产的以下策略类型：

- 分类器：启用分类器以检测所有工作区的 PII、安全问题或其他模式。
- **Classifier categories**：将分类器分为合规类别，用于组织和管理。
- **Guardrail** 政策：检索或推断时适用的安全和编辑规则。

您无法使用 ONTAP System Manager 创建或管理这些护栏策略。仅当存储管理员将它们应用于工作区时，它才会读取并强制执行它们。所有策略定义和维护都发生在 AIDE Console 中。

启用分类器

分类器分析元数据和内容以注释文件和对象（例如，检测 PII 或敏感类别）。在工作区数据上运行分类器之前，必须在 AIDE Console 中启用它们。

关于此任务

分类器行为在 AIDE Console 中进行全局控制。所有启用的分类器都在每个工作区上运行。由于它们是全局应用的，因此无法为单个工作区启用或禁用它们。它们只能在全局范围内启用或禁用。

步骤

1. 在 AIDE Console 中，导航到 **Data Guardrails > Classifiers**。
2. 选择分类器类别以显示其包含的分类器。
3. 选中要启用的分类器的复选框，或选择所有行以批量启用分类器。
4. 选择 启用。

使用批量选择选项可一次启用多个分类器。每次启用分类器时，都会触发所有工作区的工作区刷新。为了最大限度地减少不必要的刷新，请一次启用多个分类器，而不是一次启用一个分类器。

结果

所有新创建和现有的工作区在元数据处理期间运行启用的分类器。

分类标记会写入元数据目录，并可供数据工程师在创建数据集合时进行筛选。

管理分类器类别

分类器按类别组织（例如"PII"或"财务数据"）。类别可帮助您对相关分类器进行分组，以便更轻松地管理和合规可见性。您可以使用 AIDE 提供的默认类别或创建自定义类别以满足您的合规要求。

步骤

1. 在 AIDE Console 中，导航到 **Data Guardrails > Classifiers**。
2. 查看现有分类器类别。分类分为两大类：
   - 内容或数据：检测文件中的特定类型的数据。
   - 文档：根据内容对文档类型进行分类。
3. 确定默认分类器子类别是否足够，或者您是否要创建自己的子类别。
   - 如果使用默认分类器子类别（例如 **General Privacy**）：
     i. 在分类器类别中选择类别名称以显示关联的分类器。
     ii. 检查分类器列表。
     iii. 选择 **Add** 从可用分类器的完整列表中查找并添加未列出的分类器。
   - 如果要创建自定义类别，请选择 `+ Add` 。
     i. 添加唯一的名称、描述，并为类别分配可用的分类器。
     ii. 选择 **Add**
4. 要禁用类别中的分类器，请为分类器选择 ••• 并选择 禁用。您还可以选择所有行以批量更改状态。

结果

类别组织分类器以实现合规可见性。数据工程师可以在筛选和创建数据集合时使用分类标签。

创建和管理 Data Guardrails 策略

Guardrail 策略确定当分类器检测到敏感内容或当提示和检索结果违反内容规则时 AIDE 如何响应。

典型的 Data Guardrails 行为包括：

- 从检索到的代码段中隐藏或编辑 PII。
- 阻止违反合规性规则的答案。
- 记录或标记违规以进行审核。

关于此任务

您只能在 AIDE Console 中创建和管理护栏策略。

一次只能将 ONTAP System Manager 中的工作区与单个护栏策略相关联。

步骤

1. 在 AIDE Console 中，导航到 **Data Guardrails** > **Guardrail policies**。
2. 选择 **Add**。
3. 输入清楚描述范围的名称和描述（例如， `Customer PII redaction for support KB`）。
4. 配置激活 Data Guardrails 所需的数据分类器驱动条件：
   a. 定义 Data Guardrails 激活条件：
      i. 为每个条件选择分类器类别或分类器类型。
      ii. 根据需要添加和定义其他条件。
      iii. 在 **Search** 中定义特定的搜索条件，然后选择 **Accept**。
   b. 定义 Data Guardrails 策略的操作，例如匿名化内容或从数据集合中阻止和删除文件。
5. 选择要应用 guardrail 的工作区。
6. 设置策略状态：
   - **Enabled**：立即激活策略。
   - **Test Mode**：允许您在激活策略之前验证策略的影响。
   - 已禁用：保存 guardrail 但不强制执行。
7. 选择 **Add** 以保存策略并将其应用于工作区。

在启用严格执行之前，将 **Test Mode** 与试点工作区和非生产数据收集一起使用，以了解有多少响应将受到影响。

结果

新 Data Guardrails 策略处于活动状态，并适用于选定工作区。

策略如何与工作区交互

定义策略后：

- 存储管理员使用 ONTAP System Manager 创建工作区、选择数据容器和关联 Data Guardrails 策略。
- 分类器根据您启用的内容在工作区内容上自动运行。
- 附加到工作区的 Data Guardrails 会影响检索端点的行为。

对于数据工程师和数据科学家：

- 已按角色分配筛选可见数据资源（工作区和数据集合）。
- 您查询的元数据（例如 PII 标签）由启用的分类器驱动。
- 您的 RAG 管道接收的响应受工作区级别配置的 Data Guardrails 的限制。

相关信息

- "在 AI Data Engine 中查看数据收集"
- "AI Data Engine 的数据到 RAG 快速入门"

NetApp AI Data Engine 常见问题解答

本常见问题解答涵盖了有关 NetApp AI Data Engine (AIDE) 的常见问题，包括其架构、部署、用户类型、技术功能、集成和许可。

AIDE 基础知识

什么是 **NetApp AI Data Engine (AIDE)**?

NetApp AI Data Engine (AIDE) 是一种存储集成的 AI 数据服务，涵盖从发现和准备原始数据到为生成式 AI (GenAI)、Retrieval-Augmented Generation (RAG)、代理 AI 和 AI 工厂提供检索端点的整个 AI 生命周期。AIDE 自动同步和更改检测，为数据发现和管理提供所选数据的统一、最新视图。

AIDE 的工作原理是什么?

AIDE 直接与 NetApp ONTAP 存储系统集成，通过自动更改检测和同步，创建整个 NetApp 数据集的全局结构化视图。AIDE 通过压缩和重复数据删除、策略驱动的 Data Guardrails 以及与 AI 工具的集成提供实时矢量化。

用户和角色

谁在使用 **AI Data Engine**?

AIDE 的主要用户包括：

1. **ONTAP** 存储管理员：管理基础设施、AI 特定的存储需求、安全性和合规性。
2. 数据工程师：管理跨环境的数据移动、准备和集成。
3. 数据科学家：准备和转换相关数据以供 AI 使用。

要求和部署

需要哪些硬件?

AIDE 需要 AFX 系统进行部署（包括 AFX 控制器、磁盘架和网络交换机），但可以使用 SnapMirror 和集群对等从运行 ONTAP 9 的集群中使用集群数据。AIDE 部署至少需要四个 AFX 控制器节点，以确保高可用性和性能。

AIDE 在 NetApp 数据计算节点 (DCN) 上运行。需要三个 DCN。DCN 托管 AIDE 软件，其中包括 Metadata Engine、Data Sync、Data Curator 和 Data Guardrails。

我可以使用自己的 **DCN** 吗?

否。DCN 是 NetApp 提供的数据计算硬件节点，是 AI Data Engine 的唯一部署机制。

最少需要多少个 **DCN**?

需要恰好三个 DCN。

哪些操作系统在 **DCN** 上运行?

DCN 运行一个由 NetApp 提供的带有 AIDE 的软件堆栈。

可以在没有 **AFX** 的情况下部署 **AIDE** 吗?

不需要。AIDE 需要 AFX 才能部署。AIDE 使用 Trident 消耗 AFX 卷用于内部存储（持久卷）。为 AIDE 提供存储的 AFX 集群可以与 ONTAP 9 系统或集群对等。它使用集群对等和 SnapMirror 将数据从远程 ONTAP 集群同步到 AFX 系统。

管理和接口

AIDE Console 是 **NetApp Console** 的一部分还是单独的界面?

AIDE Console 是在 DCN 上运行的独立管理界面。您可以使用 AIDE Console 管理 AIDE 服务，例如 Data Guardrails 和 Data Curator。您还可以使用 ONTAP System Manager 来监控 AIDE 集群。

特性和功能

AIDE 的主要功能是什么?

AIDE 有四个主要功能：

Metadata Engine

- 自动生成数据的结构化、最新的交互式视图。
- 适用于存储在 ONTAP 上的数据。
- 使数据从业人员能够与存储管理员协作以查找和理解数据。
- API 查询元数据以提供功能，同时减少存储系统上的 NFS 流量负载。
- 元数据提取和编目功能专为 AIDE 构建，可连续工作，并利用快照等 ONTAP 功能。

数据同步

- 无需人工干预，即可在源数据更改时自动维护数据更新。
- 管理员以天或小时为单位定义数据刷新间隔。
- 提供增量数据移动和跨数据同步，以消除 AI 数据的冗余副本。

Data Guardrails

- 在整个 AI 生命周期中自动识别和保护敏感数据。可通过 AI Data Engine Console 访问。
- 持续扫描、分类和归类数据。
- 识别敏感数据（如 PII）和风险。
- 促进根据公司和监管标准制定自动处理敏感数据的策略。
- 提供自动敏感信息编辑以实现数据保护。
- 根据需要限制对敏感文件的访问。

Data Curator

- 允许数据科学家跨存储搜索相关数据。
- 使用 AFX 卷上的现有数据创建精选数据集。
- 在存储层生成矢量嵌入，以减少数据膨胀并提高性能。
- 通过矢量语义搜索和重新排名为 AI 应用程序提供检索端点。

集成和互操作性

AIDE 是否支持跨多个 **ONTAP** 集群的联合元数据?

AIDE 可以使用 SnapMirror 和集群对等连接到多个 ONTAP 集群，从而实现集中式元数据可见性。

元数据存储在哪里?

AIDE 使用 AFX 提供的持久卷在连接的 AFX 集群上存储元数据。DCN 使用本地存储进行内部操作。

AIDE Metadata Engine 对数据进行分类吗?

否。Metadata Engine 对文件系统元数据进行编目，并提供查询此编目元数据的 API。

支持哪些数据源?

AIDE 支持 ONTAP 卷（本地或远程）作为数据源。远程 ONTAP 集群必须运行 ONTAP 9 并通过集群对等和 SnapMirror 连接。

AIDE 9.18.1 中不支持 ONTAP S3 存储桶和 StorageGRID 对象作为数据源。

AIDE 可以处理哪些类型的文件进行分类、矢量化和语义搜索?

AIDE 支持多种文件类型，包括 PDF、DOCX、PPTX、TXT 和具有 OCR 功能的图像文件。

AIDE 支持非英语数据的分类吗?

AIDE 仅支持英语数据。

AIDE 支持哪些集成?

AIDE 提供可通过直接 API 调用或通过 Model Context Protocol (MCP) 服务器访问的 RAG API 端点。这支持与代理 AI 框架和工具的集成。

部署和许可

有哪些部署选项?

AIDE 在具有 DCN 的 AFX 基础设施上本地部署。它直接与 NetApp ONTAP AFX 安装集成。

AIDE 如何获得许可?

AIDE 需要软件许可证才能运行 Data Guardrails 和 Data Curator。

如果您只需要 Metadata Engine，则所有 AFX 系统都包含的 ONTAP One 许可证提供仅 Metadata Engine 功能的权限。

相关信息

- "在 ONTAP System Manager 中安装 AIDE 许可证"
- "了解 AIDE 架构和组件"

法律声明

法律声明提供对版权声明、商标、专利等信息的访问。

版权

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

商标

NETAPP、NETAPP 标识和 NetApp 商标页面上所列的商标是 NetApp, Inc. 的商标。其他公司和产品名称可能是其各自所有者的商标。

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

专利

当前 NetApp 拥有的专利列表可以在以下网址找到：

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

隐私政策

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

开源

通知文件提供有关 NetApp 软件中使用的第三方版权和许可的信息。

AI Data Engine

"AIDE 9.18.1 通知"