Documentación de NetApp Workload Factory para GenAI

GenAI

NetApp
October 06, 2025

Tabla de contenidos

Documentación de NetApp Workload Factory para GenAI

Notas de la versión

Novedades de NetApp Workload Factory para GenAI

Conozca las novedades sobre la capacidad de cargas de trabajo de IA generativa de Workload Factory.

5 de octubre de 2025

BlueXP workload factory ahora Fábrica de cargas de trabajo de NetApp

BlueXP ha sido renombrado y rediseñado para reflejar mejor el papel que desempeña en la gestión de su infraestructura de datos. Como resultado, la BlueXP workload factory pasó a llamarse Fábrica de carga de trabajo de NetApp .

Compatibilidad para agregar orígenes de datos NFS/SMB genéricos en conectores de NetApp para Amazon Q Business

Con la API de Workload Factory, ahora puede agregar una fuente de datos desde un recurso compartido NFSv3, NFSv4 o SMB genérico a un conector de NetApp para Amazon Q Business. Esto le permite incluir archivos almacenados en volúmenes alojados por sistemas de archivos distintos de Amazon FSx for NetApp ONTAP.

"Cree un conector de NetApp para Amazon Q Business"

"Agregar fuentes de datos a un conector"

Configuración avanzada de chat para bases de conocimiento

Ahora puede configurar ajustes de chat avanzados que se aplican al modelo de chat para la base de conocimiento, como la duración de la respuesta, la temperatura, la configuración de razonamiento y más. Algunas de estas configuraciones, como las configuraciones de actualidad y hora de modificación, las configuraciones de recuperación avanzada y el mensaje del sistema, solo están disponibles mediante la API de Workload Factory.

"Cree una base de conocimientos de GenAI"

La selección del tipo de inferencia ahora es compatible con modelos de incrustación, chat y reclasificación.

Si el modelo de incrustación, chat o reclasificación que eligió tiene configuraciones de inferencia, ahora puede seleccionar un tipo de inferencia. Esto le permite ajustar mejor el rendimiento del chatbot y los requisitos de recursos a sus necesidades.

"Cree una base de conocimientos de GenAI"

03 de agosto de 2025

Almacenamiento seguro para resultados de datos estructurados

Si los resultados de la consulta del chatbot contienen datos estructurados, GenAI puede almacenar los resultados en un bucket de Amazon S3. Cuando estos resultados se almacenan en un bucket S3, puedes

descargarlos utilizando el enlace de descarga dentro de la sesión de chat.

"Cree una base de conocimientos de GenAI"

Disponibilidad del servidor MCP

NetApp ahora proporciona un servidor de Protocolo de contexto de modelo (MCP) con NetApp Workload Factory para GenAI. Puede instalar el servidor localmente para permitir que los clientes MCP externos descubran y recuperen resultados de consultas de una base de conocimiento GenAI.

"Servidor MCP GenAI de NetApp Workload Factory"

29 de junio de 2025

Compatibilidad con fuentes de datos alojadas en sistemas de archivos NFS/SMB genéricos

Ahora puede agregar una fuente de datos desde un recurso compartido SMB o NFS genérico. Esto le permite incluir archivos almacenados en volúmenes alojados por sistemas de archivos distintos de Amazon FSx para NetApp ONTAP.

"Agregar fuentes de datos a una base de conocimientos"

"Agregar fuentes de datos a un conector"

03 de junio de 2025

Rastreador disponible para operaciones de monitoreo y seguimiento

La función de monitoreo de Tracker ya está disponible en GenAI. Puede usar Tracker para monitorear y dar seguimiento al progreso y estado de operaciones pendientes, en curso y completadas, revisar los detalles de las tareas y subtareas de la operación, diagnosticar problemas o fallos, editar parámetros de operaciones fallidas y reintentar operaciones fallidas.

"Supervise las operaciones de carga de trabajo con Tracker en NetApp Workload Factory"

Elija un modelo de reranking para una base de conocimientos

Ahora puede aumentar la relevancia de los resultados de consultas rerankeadas seleccionando un modelo de reranking específico para usar con una base de conocimiento. GenAI es compatible con los modelos Cohere Rerank y Amazon Rerank.

"Cree una base de conocimientos de GenAI"

04 de mayo de 2025

Compatibilidad con NetApp Connector para Amazon Q Business

Esta versión de GenAI presenta compatibilidad con NetApp Connector para Amazon Q Business, lo que le permite crear conectores para Amazon Q Business. Aproveche rápida y fácilmente el asistente de IA de Amazon Q Business con menos configuración inicial que la creación de una base de conocimientos de GenAI para Amazon Bedrock.

"Cree un conector de NetApp para Amazon Q Business"

Compatibilidad mejorada con modelos de chat

GenAI ahora admite los siguientes modelos de chat adicionales para las bases de conocimiento:

- "Modelos de IA Mistral"
- "Modelos de texto de Amazon Titan"
- "Modelos de Meta Llama"
- "Modelos Jamba 1,5"
- "Modelos de comando Cohere"
- "Modelos Deepseek"

GenAI admite los modelos de cada proveedor compatible con Amazon Bedrock: "Modelos de base admitidos en Amazon Bedrock"

"Cree una base de conocimientos de GenAI"

Terminología de permisos actualizada

La interfaz de usuario y la documentación de Workload Factory ahora usan "solo lectura" para referirse a los permisos de lectura y "lectura/escritura" para referirse a los permisos automatizados.

02 de marzo de 2025

Mejoras de chatbot incrustadas

Ahora puede copiar preguntas y respuestas directamente al portapapeles, ajustar el tamaño de la ventana de chat y cambiar su título. Además, las respuestas de chat ahora pueden incluir tablas, que también son copiables.

"Pruebe una base de conocimientos de GenAI"

Compatibilidad con citas de respuesta de chat

Las respuestas de chat ahora incluyen citas que enumeran los archivos y fragmentos de datos que se utilizaron para generar la respuesta.

"Pruebe una base de conocimientos de GenAI"

Compatibilidad con tipos de archivo mejorada

Esta versión de GenAI proporciona soporte de archivos mejorado:

- Los modelos de chat ofrecen soporte CSV mejorado. Esto permite respuestas más útiles al consultar datos de archivos CSV.
- GenAI ahora puede ingerir archivos Apache Parquet de fuentes de datos.
- GenAI ahora admite la ingesta de archivos DOCX de Microsoft Word que incluyen imágenes. Las imágenes incrustadas en los documentos DOCX se escanean, y las perspectivas de texto de las imágenes incrustadas se incluyen en las respuestas a las consultas de la base de conocimientos.

"Formatos de archivo de origen de datos compatibles"

02 de febrero de 2025

Compatibilidad con los modelos de base de Amazon Nova

GenAI ahora es compatible con los modelos básicos de Amazon Nova. Amazon Nova Micro, Amazon Nova Lite y Amazon Nova Pro son compatibles.

"Requisitos de GenAI"

Filtrado de tipo de archivo para orígenes de datos

GenAI ahora admite la selección de tipos de archivo específicos para incluir en el análisis de origen de datos al agregar un origen de datos.

"Agregue orígenes de datos a la base de conocimientos"

Filtrado de fecha de modificación de archivo para orígenes de datos

GenAI ahora admite el filtrado de archivos para incluirlos en el análisis de origen de datos por fecha de modificación al agregar un origen de datos. Puede elegir un rango de fechas de modificación para los archivos incluidos.

"Agregue orígenes de datos a la base de conocimientos"

Soporte para archivos de imagen y soporte mejorado para archivos PDF

GenAI ahora admite la mejora de las respuestas a las consultas de la base de conocimientos con información de las imágenes y descripciones de gráficos, así como el texto de los documentos, lo que conduce a respuestas más ricas y de mayor calidad. GenAI ahora puede escanear archivos de imagen e imágenes dentro de archivos PDF (también conocido como soporte de archivos multimodal). Si decide digitalizar imágenes o archivos PDF, el texto de las imágenes (incluidas las imágenes incrustadas en documentos PDF) se escanea en la fuente de datos y las perspectivas de los escaneos se incluyen en las respuestas a las consultas de la base de conocimientos.

"Agregue orígenes de datos a la base de conocimientos"

Soporte híbrido de búsqueda y reselección

Ahora GenAI puede mejorar significativamente la relevancia y precisión de los resultados de búsqueda mediante la búsqueda híbrida y la re-clasificación de los resultados. La búsqueda híbrida combina los puntos fuertes de la búsqueda tradicional basada en palabras clave con técnicas avanzadas de búsqueda semántica densas basadas en vectores. Los resultados de búsqueda de palabras clave estándar se aumentan con coincidencias cercanas y matices lingüísticos, lo que mejora la relevancia. Luego, GenAI refina aún más estos resultados mediante el uso de modelos avanzados de re-clasificación, como Cohere Rerank y Amazon Rerank, y devuelve los resultados más relevantes. Esta capacidad está disponible para las bases de conocimiento recién creadas.

"Obtenga más información sobre NetApp Workload Factory para GenAI"

05 de enero de 2025

Nombre de snapshot personalizado

Ahora es posible proporcionar un nombre de snapshot para una snapshot ad-hoc.

"Proteger una base de conocimientos con instantáneas"

Nombre de instancia de motor AI personalizado

Ahora puede asignar un nombre personalizado a la instancia del motor de AI durante la implementación.

"Ponga en marcha la infraestructura de GenAI"

Reconstruir la infraestructura de GenAI dañada o faltante

Si su instancia de motor de IA se daña o se elimina de alguna manera, puede dejar que Workload Factory la reconstruya por usted. La fábrica de carga de trabajo vuelve a conectar automáticamente sus bases de conocimiento a la infraestructura una vez finalizada la reconstrucción, de modo que estén listas para usar.

"Resolución de problemas"

01 de diciembre de 2024

Clone una base de conocimientos desde una copia de Snapshot

NetApp Workload Factory para GenAI ahora admite la clonación de una base de conocimiento a partir de una instantánea. Esto permite la recuperación rápida de bases de conocimiento y la creación de nuevas bases de conocimiento con fuentes de datos existentes, y ayuda con la recuperación y el desarrollo de datos.

"Clonar una base de conocimientos"

Detección y replicación de clústeres de ONTAP en las instalaciones

Descubra y replique datos del clúster ONTAP local en un sistema de archivos FSx para ONTAP para que puedan usarse para enriquecer las bases de conocimiento de IA. Todos los flujos de trabajo de descubrimiento y replicación locales son posibles desde el nuevo menú **On-Premises ONTAP** en el inventario de almacenamiento.

"Detectar un clúster de ONTAP en las instalaciones"

3 de noviembre de 2024

Enmascarar la información de identificación personal con barreras de datos

La carga de trabajo de IA generativa presenta la función de protección de datos, impulsada por la clasificación de la consola de NetApp . La función de protección de datos identifica y enmascara la información de identificación personal (PII), lo que le ayuda a mantener el cumplimiento y fortalecer la seguridad de los datos confidenciales de su organización.

"Cree una base de conocimientos de GenAI"

"Obtenga más información sobre la clasificación de la consola de NetApp"

29 de septiembre de 2024

Compatibilidad con Snapshot y restauración para volúmenes de base de conocimientos

Ahora puede proteger los datos de las cargas de trabajo de IA generativas con una copia puntual de una base de conocimientos. Esto le permite proteger los datos contra la pérdida accidental o cambios de pruebas en la

configuración de la base de conocimientos. Puede restaurar la versión anterior del volumen de la base de conocimientos en cualquier momento.

"Tomar una copia Snapshot de un volumen de base de conocimientos"

"Restaure una copia Snapshot de un volumen de base de conocimientos"

Pausar secuencias programadas

Ahora puede pausar las exploraciones de orígenes de datos programadas. De forma predeterminada, las cargas de trabajo de IA generativas analizan cada origen de datos diariamente para incorporar datos nuevos en cada base de conocimientos. Si no desea que se ingieran los cambios más recientes (durante la prueba o al restaurar una instantánea, por ejemplo), puede pausar las exploraciones programadas y reanudarlas en cualquier momento.

"Gestionar bases de conocimientos"

Volúmenes de protección de datos ahora compatibles con las bases de conocimientos

Cuando se selecciona un volumen de base de conocimientos, ahora se puede elegir un volumen de protección de datos que forme parte de una relación de replicación de NetApp SnapMirror. Esto permite almacenar bases de conocimientos en volúmenes que ya están protegidos por la replicación de SnapMirror.

"Identifique las fuentes de datos que desea integrar en su base de conocimientos"

1 de septiembre de 2024

Estrategias de fragmentación adicionales

Las cargas de trabajo de IA generativas ahora admiten la fragmentación de varias frases y la fragmentación basada en superposiciones para los orígenes de datos.

Volumen dedicado para cada base de conocimientos

Las cargas de trabajo de IA generativas ahora crean un volumen de Amazon FSx para NetApp ONTAP dedicado para cada nueva base de conocimientos, lo que permite establecer políticas Snapshot individuales para cada base de conocimientos y mejorar la protección contra errores y envenenamientos de datos.

4 de agosto de 2024

Integración de Amazon CloudWatch Logs

Las cargas de trabajo de IA generativas ahora están integradas con Amazon CloudWatch Logs, lo que le permite supervisar los archivos de registro de las cargas de trabajo de IA generativas.

Aplicación de chatbot de ejemplo

La aplicación de muestra NetApp Workload Factory GenAI le permite probar la autenticación y la recuperación de su base de conocimiento publicada de NetApp Workload Factory interactuando directamente con ella en una aplicación de chatbot basada en la web.

7 de julio de 2024

Lanzamiento inicial de Workload Factory para GenAI

La versión inicial incluye la capacidad de desarrollar una base de conocimientos personalizada mediante la incorporación de los datos de la organización. La base de conocimientos puede ser accedida por una aplicación de chatbot para sus usuarios. Esta capacidad garantiza respuestas precisas y relevantes a preguntas específicas de la organización, mejorando la satisfacción y la productividad de todos sus usuarios.

Obtenga más información sobre NetApp Workload Factory para GenAI

Obtenga más información sobre NetApp Workload Factory para GenAI

NetApp Workload Factory para GenAI le permite integrar los sistemas de archivos Amazon FSx for NetApp ONTAP con los modelos básicos de GenAI. Esto le proporciona almacenamiento de alto rendimiento con un amplio conjunto de funciones de protección, seguridad y optimización de costos para sus conjuntos de datos de IA.

¿Qué es NetApp Workload Factory para GenAI?

NetApp Workload Factory para GenAI le permite utilizar sus fuentes de datos empresariales en Amazon FSx for NetApp ONTAP con aplicaciones de IA generativa. Al utilizar la generación aumentada por recuperación (RAG), puede conectar rápidamente fuentes de datos a modelos base disponibles a través de Amazon Bedrock o Amazon Q Business para desarrollar aplicaciones impulsadas por IA generativa, como asistentes virtuales, chatbots de preguntas y respuestas, resumen de documentos, creación de contenido, etc.

El uso de la IA generativa con los datos de su organización le permite aprovechar sus propios conocimientos y experiencia, en lugar de depender únicamente de la inteligencia del modelo basada en datos públicos en los que se entrenaron los modelos. El uso de RAG para personalizar los modelos garantiza respuestas precisas y relevantes a preguntas específicas de la organización, mejorando la productividad y la eficiencia para los usuarios de sus aplicaciones mediante IA generativa.

El desarrollo de una aplicación de GenAI adaptada a los datos de su organización le permite aprovechar su propio conocimiento y experiencia. Esta capacidad de personalización garantiza respuestas precisas y relevantes a preguntas específicas de la organización, mejorando la satisfacción y la productividad de todos sus usuarios.

Si "crear una base de conocimientos"GenAI ingiere datos de sus orígenes de datos, almacena los resultados vectorizados en una base de datos y le proporciona un control total sobre cómo utilizar los datos ingeridos para responder a las consultas. Este enfoque requiere más configuración inicial, pero le permite elegir diferentes modelos de chat para diferentes resultados. Si usted "definir un conector NetApp para Amazon Q Business", los datos de sus orígenes de datos los consume Amazon Q Business y los almacena en un índice. Este enfoque requiere menos configuración inicial, pero le da menos control sobre los resultados.

Para obtener más información sobre Workload Factory, consulte "Descripción general de Workload Factory" .

Beneficios de usar GenAI para crear aplicaciones de IA generativas

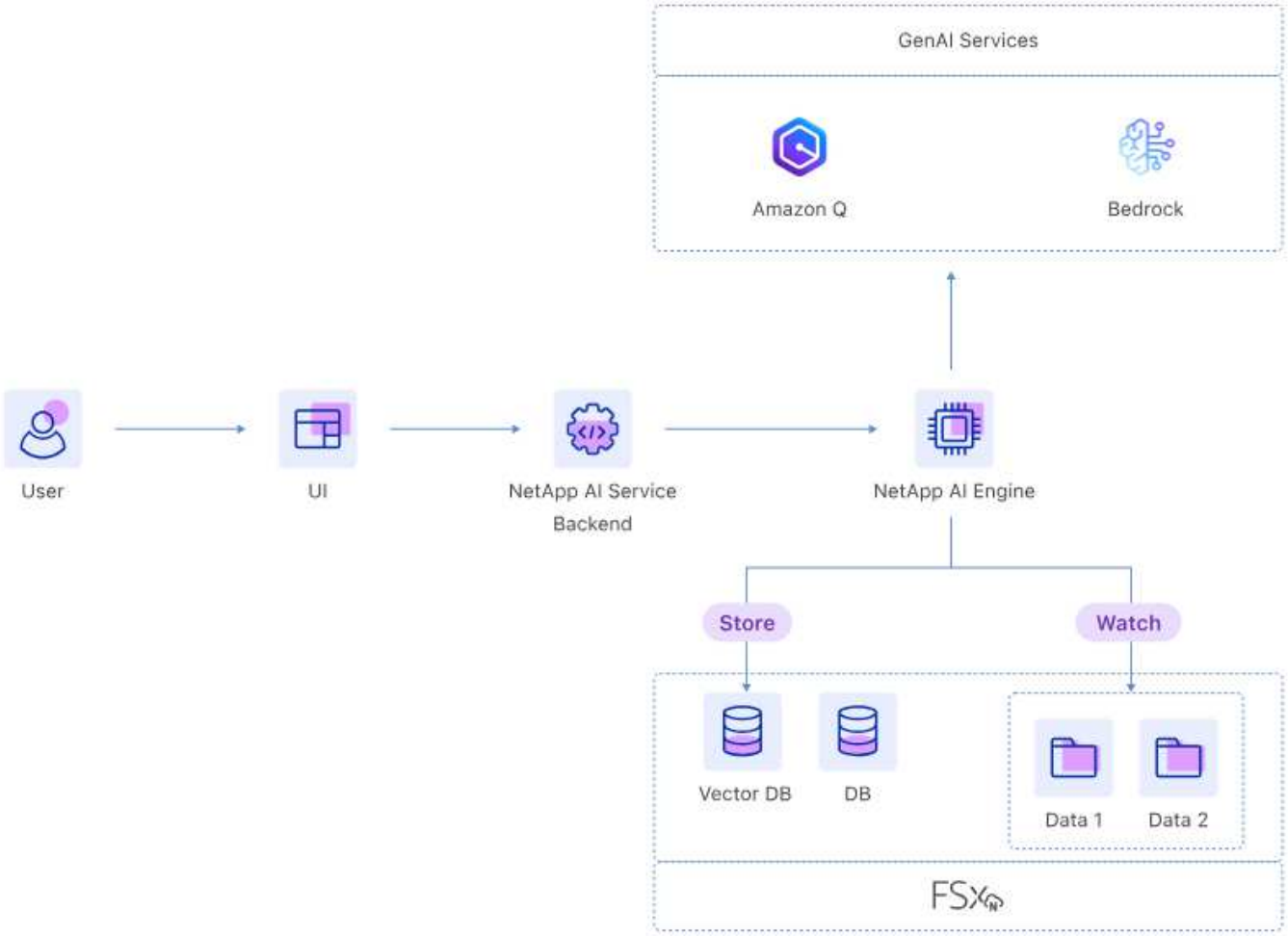
NetApp Workload Factory para GenAI simplifica el proceso de implementación de la infraestructura necesaria para crear aplicaciones de IA generativa mediante generación aumentada por recuperación (RAG). En concreto, GenAI proporciona los siguientes beneficios:

- Sin necesidad de un profundo conocimiento de la infraestructura de datos, los modelos de lenguaje y bases, los administradores y desarrolladores DE TI pueden acelerar el desarrollo de aplicaciones utilizando la automatización que ofrece GenAI. Los administradores y los desarrolladores de datos pueden crear de forma rápida y sencilla bases de conocimientos empresariales que integren los datos no estructurados de su organización para que los usen las aplicaciones de IA generativas.

- Mejore la seguridad al conservar los permisos de los usuarios en los archivos integrados en las bases de conocimiento para garantizar que se mantenga la seguridad y la privacidad de los datos. Una aplicación, como un chatbot, se puede desarrollar para proporcionar solo a los usuarios autenticados respuestas basadas en los datos a los que los usuarios tienen acceso.
- Mantenga la privacidad y seguridad de los datos de su empresa desde su cuenta de cliente de AWS, donde los datos de su organización nunca se exponen externamente.
- Acelere el desarrollo de aplicaciones de GenAI, como un chatbot de preguntas y respuestas mediante marcos de código abierto como LangChain, utilizando la API de GenAI para aprovisionar y gestionar bases de conocimiento y conectores, chatear con una base de conocimientos y almacenar y recuperar el historial de chat.
- Mejora la protección de datos y la política de disponibilidad mediante la puesta en marcha de la infraestructura de datos de IA generativa en sistemas de archivos FSx para NetApp ONTAP y el aprovechamiento de funciones de ONTAP, como la alta disponibilidad, los snapshots para la protección y recuperación de datos locales, SnapMirror para la recuperación ante desastres y SnapVault para hacer backups de tu infraestructura de datos.
- Reduce los costes generales de almacenamiento para la infraestructura de datos de IA generativa aprovechando funciones de eficiencia de datos de ONTAP como deduplicación de datos, compresión y compactación, organización en niveles de los datos y thin provisioning.
- Obtenga resultados de alta calidad de sus datos con las funciones de búsqueda híbrida y reclasificación proporcionadas por GenAI. La búsqueda híbrida combinada con la re-clasificación mejora en gran medida la relevancia de los resultados de búsqueda. Estas funciones están disponibles a través de Amazon AWS y dependen de la región.

Cómo funciona GenAI

GenAI utiliza los datos privados de su organización para complementar la inteligencia del modelo (basada en los datos en los que se entrenó) con el fin de proporcionar respuestas personalizadas a las preguntas formuladas por los usuarios de su organización. En primer lugar, debe implementar la infraestructura necesaria para un marco RAG, luego crear una base de conocimientos o definir un conector utilizando las fuentes de datos y los modelos de base de su organización disponibles a través de Amazon Bedrock o Amazon Q Business y, a continuación, conectar una aplicación (como un chatbot de preguntas y respuestas) a la base de conocimientos o el conector.



Cómo NetApp Workload Factory para GenAI ayuda a crear aplicaciones de IA generativa

GenAI ayuda a construir aplicaciones de IA generativas usando RAG de las siguientes maneras:

- Pone en marcha la infraestructura necesaria para el marco de generación aumentada de recuperación (RAG) con el fin de trabajar con fuentes de datos en los sistemas de archivos FSx para ONTAP y Amazon Bedrock o Amazon Q Business. La infraestructura incluye la instancia del motor GenAI de NetApp para gestionar los datos, una base de datos vectorial incorporada (LanceDB) y almacenamiento en el sistema de archivos FSx para ONTAP para la base de datos vectorial.
- Ayuda a conectar las fuentes de datos con incrustaciones y modelos de lenguaje disponibles a través de Amazon Bedrock o Amazon Q Business para incrustar fuentes de datos y recuperar las respuestas de las consultas de los usuarios. Las fuentes de datos, junto con los modelos y su configuración, se presentan como bases de conocimiento de FSx para ONTAP.
- Procesa datos de origen en el conector o la base de conocimientos para incrustar archivos de origen en recursos compartidos SMB y exportaciones NFS en sistemas de archivos FSx para ONTAP, junto con almacenar permisos de archivos para archivos dentro de recursos compartidos SMB.
- Crea automáticamente preguntas para iniciar conversaciones basadas en el contenido de las bases de conocimientos.
- Proporciona un simulador de chat para que los administradores de datos prueben el chat con bases de conocimientos.
- Proporciona una interfaz de conector simple para que pueda conectar GenAI con Amazon Q Business, de

forma rápida y sencilla utilizando las capacidades de este asistente de IA.

Herramientas para utilizar NetApp Workload Factory

Puede utilizar NetApp Workload Factory con las siguientes herramientas:

- **Consola Workload Factory**: La consola Workload Factory proporciona una vista visual y holística de sus aplicaciones y proyectos.
- *Consola NetApp *: La consola NetApp proporciona una experiencia de interfaz híbrida para que pueda utilizar Workload Factory junto con otros servicios de datos de NetApp .
- **Pregúntame**: utiliza el asistente de IA Pregúntame para hacer preguntas y obtener más información sobre Workload Factory sin salir de la consola de Workload Factory. Accede a Pregúntame desde el menú de ayuda de Workload Factory.
- **CloudShell CLI**: Workload Factory incluye una CLI de CloudShell para administrar y operar entornos de AWS y NetApp en todas las cuentas desde una única CLI basada en navegador. Acceda a CloudShell desde la barra superior de la consola de Workload Factory.
- **API REST**: utilice las API REST de Workload Factory para implementar y administrar sus sistemas de archivos FSx para ONTAP y otros recursos de AWS.
- **CloudFormation**: use el código de AWS CloudFormation para realizar las acciones que definió en la consola de Workload Factory para modelar, aprovisionar y administrar recursos de AWS y de terceros desde la pila de CloudFormation en su cuenta de AWS.
- **Proveedor de Terraform NetApp Workload Factory**: utilice Terraform para crear y administrar flujos de trabajo de infraestructura generados en la consola de Workload Factory.

Coste

No hay ningún costo por utilizar la capacidad GenAI de Workload Factory.

Sin embargo, tendrá que pagar por los recursos de AWS que ponga en marcha para dar soporte a la infraestructura de IA generativa. Por ejemplo, pagará a AWS por Amazon Bedrock o Amazon Q Business, FSX para el sistema de archivos y la capacidad de almacenamiento de ONTAP y la instancia EC2 del motor GenAI.

Algunas operaciones multimodales, como escanear imágenes para obtener información de texto, pueden utilizar más recursos y, por lo tanto, incurrir en un costo mayor. Algunas operaciones de configuración, como cambiar los ajustes de una base de conocimiento, pueden hacer que los orígenes de datos se vuelvan a analizar y los análisis de orígenes de datos también pueden suponer un coste mayor.

Licencia

No se necesitan licencias especiales de NetApp para utilizar las capacidades de IA de Workload Factory.

Regiones

Workload Factory es compatible con todas las regiones comerciales donde se admite FSx para ONTAP ."Ver regiones de Amazon compatibles."

Las siguientes regiones de AWS no son compatibles:

- regiones de China

- Regiones de GovCloud (EE. UU.)
- Nube secreta
- Nube de alto secreto

Componentes del motor NetApp GenAI

Cuando implementa la infraestructura GenAI, Workload Factory crea una instancia EC2 para el motor GenAI. También crea una función de IAM, un grupo de seguridad y puntos finales privados para esta instancia. Es posible que desee comprender más detalles sobre estos componentes que Workload Factory crea en su entorno de AWS.

Tipo de instancia de EC2

m5.large

Rol IAM

La instancia del motor de GenAI necesita permisos para enviar fragmentos de datos al modelo de integración en Amazon Bedrock y para comunicarse con el backend del servicio de IA de NetApp. El rol de IAM incluye los siguientes permisos:

Permisos de rol de IAM

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Action": [
        "iam:CreateRole",
        "iam:CreatePolicy",
        "iam:AttachRolePolicy",
        "iam:PassRole"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:DescribeDocument",
        "ssm:DescribeAssociation",
        "ssm:GetDeployablePatchSnapshotForInstance",
        "ssm:GetManifest",
        "ssm:ListInstanceAssociations",
        "ssm:ListAssociations",
        "ssm:PutInventory",
        "ssm:PutComplianceItems",
        "ssm:PutConfigurePackageResult",
        "ssm:UpdateAssociationStatus",
        "ssm:UpdateInstanceAssociationStatus",
        "ssm:UpdateInstanceInformation",
        "ssmmessages:CreateControlChannel",
        "ssmmessages:CreateDataChannel",
        "ssmmessages:OpenControlChannel",
        "ssmmessages:OpenDataChannel"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ssm:GetParameter"
      ],
      "Resource": "arn:aws:ssm:*:*:parameter/netapp/wlmai/*",
      "Effect": "Allow"
    },
    {
      "Action": [
```

```
            "fsx:DescribeVolumes",
            "fsx:DescribeStorageVirtualMachines",
            "fsx:DescribeFileSystems"
        ],
        "Resource": "*",
        "Effect": "Allow"
    },
    {
        "Action": [
            "fsx:TagResource",
            "fsx:ListTagsForResource"
        ],
        "Resource": [
            "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
            "arn:aws:fsx:*:*:volume/*/*"
        ],
        "Effect": "Allow"
    },
    {
        "Action": [
            "fsx:CreateVolume"
        ],
        "Resource": [
            "arn:aws:fsx:*:*:volume/*/*",
            "arn:aws:fsx:*:*:storage-virtual-machine/*/*"
        ],
        "Effect": "Allow"
    },
    {
        "Condition": {
            "StringLike": {
                "aws:ResourceTag/netapp:wlmai:<id>:kbId": "*"
            }
        },
        "Action": "fsx:DeleteVolume",
        "Resource": [
            "arn:aws:fsx:*:*:volume/*/*",
            "arn:aws:fsx:*:*:backup/*"
        ],
        "Effect": "Allow"
    },
    {
        "Condition": {
            "StringLike": {
                "aws:ResourceTag/netapp:wlmai:<id>:qConnectorId": "*"
            }
```

```
    },
    "Action": "fsx:DeleteVolume",
    "Resource": [
      "arn:aws:fsx:*:*:volume/*/*",
      "arn:aws:fsx:*:*:backup/*"
    ],
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>:kbId": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:volume/*/*",
    "Effect": "Allow"
  },
  {
    "Condition": {
      "StringLike": {
        "aws:ResourceTag/netapp:wlmai:<id>:qConnectorId": "*"
      }
    },
    "Action": "fsx:UntagResource",
    "Resource": "arn:aws:fsx:*:*:volume/*/*",
    "Effect": "Allow"
  },
  {
    "Action": [
      "bedrock:InvokeModel",
      "bedrock:Rerank",
      "bedrock:GetFoundationModel",
      "bedrock:GetInferenceProfile",
      "bedrock:GetModelInvocationLoggingConfiguration",
      "bedrock:PutModelInvocationLoggingConfiguration"
```

```
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ec2messages:GetMessages",
        "ec2messages:GetEndpoint",
        "ec2messages:AcknowledgeMessage",
        "ec2messages:DeleteMessage",
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:ListApplications",
        "qbusiness:GetApplication",
        "qbusiness:CreateDataSource",
        "qbusiness:DeleteDataSource",
        "qbusiness:ListIndices",
        "qbusiness:StartDataSourceSyncJob",
        "qbusiness:StopDataSourceSyncJob",
        "qbusiness:ListDataSourceSyncJobs",
        "qbusiness:BatchPutDocument",
        "qbusiness:BatchDeleteDocument"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogGroups"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogStreams",
        "logs:PutLogEvents",
        "logs:CreateLogStream",
        "logs:CreateLogGroup"
```

```
        ],
        "Resource": [
          "arn:aws:logs:*:*:log-group:/aws/bedrock*",
          "arn:aws:logs:*:*:log-group:/netapp/wlmai/*:log-stream:*",
          "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
        ],
        "Effect": "Allow"
      },
      {
        "Action": [
          "s3:GetObject",
          "s3:PutObject"
        ],
        "Resource": "*",
        "Effect": "Allow"
      },
      {
        "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey"
        ],
        "Resource": "*",
        "Effect": "Allow"
      }
    ]
}
```

Grupo de seguridad

Las reglas salientes están abiertas a todo el tráfico, mientras que las reglas entrantes están completamente cerradas.

Puntos finales privados

Si la VPC de destino aún no los tiene, Workload Factory crea puntos finales privados para la instancia EC2 del motor GenAI para que pueda comunicarse con los siguientes servicios de AWS:

- Roca base de Amazon
  - roca
  - tiempo de ejecución de bedrock
  - bedrock-agent-runtime
- Registro de contenedores elásticos de Amazon (ECR)
  - API
  - docker
- Administrador de sistemas AWS (SSM)
  - ssm

    - ec2messages
    - ssmmensajes
- Amazon FSX para ONTAP de NetApp
- Amazon CloudWatch

Utilice GenAI para crear bases de conocimiento para Amazon Bedrock

Manos a la obra

Inicio rápido para las bases de conocimiento de GenAI

Empieza a crear una base de conocimientos o un conector de Amazon Q Business con los datos de tu organización que existen en los sistemas de archivos de Amazon FSx for NetApp ONTAP. Una aplicación como un chatbot accederá a esta base de conocimientos o conector para proporcionar respuestas centradas en la organización a los usuarios finales.

Iniciar sesión en Workload Factory

Necesitarás "Configurar una cuenta con Workload Factory" e inicie sesión usando uno de los "experiencias de consola" .

Configure su entorno para cumplir los requisitos de GenAI

Necesitará las credenciales de AWS para implementar la infraestructura de AWS, un sistema de archivos FSX for ONTAP implementado y descubierto, la lista de fuentes de datos que desea integrar en su base de conocimientos o conector, el acceso al servicio de IA de Amazon Bedrock o la aplicación de Amazon Q Business y mucho más.

"Obtenga más información sobre los requisitos de GenAI".

Identifique el sistema de archivos FSx para ONTAP que contiene los orígenes de datos

Las fuentes de datos que integrarás en tu base de conocimientos pueden ubicarse en un único sistema de archivos FSx para ONTAP o en varios sistemas de archivos FSx para ONTAP. Si estos sistemas están en VPC diferentes, deben ser accesibles dentro de la misma red o los VPC deben estar conectados y usar la misma región y cuenta de AWS que el motor AI.

"Aprenda a identificar fuentes de datos".

Ponga en marcha la infraestructura de GenAI

Inicie el asistente de implementación de infraestructuras para implementar la infraestructura de GenAI en su entorno AWS. Este proceso pone en marcha una instancia EC2 para el motor NetApp GenAI y un volumen en un sistema de archivos FSx para ONTAP para contener las bases de datos del motor de IA de NetApp. El volumen se utiliza para almacenar la base de datos vectorial utilizada por la base de conocimientos.

"Aprenda a implementar la infraestructura de la base de conocimientos".

El futuro

Ahora puede crear una base de conocimientos para proporcionar respuestas centradas en la organización a

los usuarios finales.

Requisitos de la base de conocimientos de GenAI

Asegúrese de que Workload Factory y AWS estén configurados correctamente antes de crear su base de conocimientos. Esto incluye tener sus credenciales de inicio de sesión de AWS, un sistema de archivos FSx para ONTAP implementado que contenga las fuentes de datos que desea integrar en su base de conocimiento, acceso al servicio Amazon Bedrock AI y más.

Requisitos básicos de GenAI

GenAI tiene requisitos generales que su entorno necesita cumplir antes de empezar.

Inicio de sesión y cuenta de Workload Factory

Necesitarás "Configurar una cuenta con Workload Factory" e inicie sesión usando uno de los "experiencias de consola" .

Credenciales y permisos de AWS

Debe agregar credenciales de AWS a Workload Factory con permisos de lectura y escritura, lo que significa que utilizará Workload Factory en modo de lectura y escritura para GenAI.

Los permisos de modo *Basic* y modo *Read-Only* no se admiten en este momento.

Al configurar sus credenciales, al seleccionar los permisos que se muestran a continuación, tendrá acceso completo para administrar los sistemas de archivos de FSx para ONTAP y para implementar y administrar la instancia de GenAI EC2 y otros recursos de AWS necesarios para su base de conocimientos y chatbot.

"Aprenda a agregar credenciales de AWS a Workload Factory"

Requisitos de la base de conocimientos de GenAI

Si planea trabajar con bases de conocimientos, asegúrese de que su entorno cumpla con los siguientes requisitos.

Roca base de Amazon

Amazon Bedrock le permite utilizar modelos de base y proporciona las funcionalidades necesarias para crear aplicaciones de IA generativas.

Antes de comenzar a utilizar NetApp Workload Factory para GenAI, debe configurar Amazon Bedrock. Su implementación de GenAI debe estar en una región de AWS que tenga Amazon Bedrock habilitado.

- "Documentación de AWS: Configure Amazon Bedrock"
- "Documentación de AWS: Regiones y modelos compatibles para las bases de conocimiento de Amazon Bedrock"

GenAI vuelve a clasificar los resultados de búsqueda de forma predeterminada para mejorar la relevancia de los resultados. Para obtener los mejores resultados, asegúrate de que la configuración del modelo de base de Amazon Bedrock incluya acceso a un modelo de reclasificación, como Cohere Rerank o Amazon Rerank, si está disponible en tu región.

Modelo de incrustación

Debe activar el modelo de incrustación que planea utilizar antes de crear la base de conocimientos. Se admiten los siguientes modelos de inclusión:

- Titan Embeddings G1 - Texto
- Titan Embedding Text v2
- Titán incrustaciones multimodales G1
- Insertar inglés
- Incrustar multilingüe

"Más información sobre Amazon Titan"

Modelo de chat

Debe activar el modelo de chat básico que planea utilizar antes de crear su base de conocimientos. Dado que el soporte de modelos varía según la región de AWS, consulte "La documentación de AWS" para verificar qué modelos puede utilizar en las regiones en las que planea implementar su base de conocimientos.

GenAI soporta varios modelos de Anthropic, Amazon, Mistral AI, Meta, Jamba y cohere.

Obtenga más información sobre el uso de estos modelos en Amazon Bedrock:

- "Claude de Antropic en Amazon Bedrock"
- "Introducción a Amazon Nova en la consola Amazon Bedrock"
- "Modelos de IA Mistral"
- "Modelos de texto de Amazon Titan"
- "Modelos de Meta Llama"
- "Modelos Jamba"
- "Modelos de comando Cohere"

FSX para sistema de archivos ONTAP

Necesita un mínimo de un sistema de archivos FSx para ONTAP:

- El motor GenAI de NetApp utilizará (o creará, si no existe) un sistema de archivos para almacenar la base de datos vectorial utilizada por la base de datos de conocimientos.

  Este sistema de archivos FSx para ONTAP debe utilizar volúmenes de FlexVol. No se admiten los volúmenes de FlexGroup.

- Uno o varios sistemas de archivos contendrán las fuentes de datos que integrará en su base de conocimientos.

  Un sistema de archivos FSx para ONTAP se puede utilizar con ambos fines o puede utilizar varios sistemas de archivos FSx para ONTAP.

- Necesitará conocer la región, la VPC y la subred de AWS donde reside el sistema de archivos AWS FSx para ONTAP. El sistema de archivos debe estar en una región de AWS que tenga habilitado Amazon Bedrock.
- Tendrá que considerar los pares clave/valor de etiqueta que desea aplicar a los recursos de AWS que

forman parte de esta implementación (opcional).

- Necesitará conocer la información del par de claves que le permite conectarse de forma segura a la instancia del motor de NetApp AI.

"Descubre cómo implementar y gestionar sistemas de archivos FSx para ONTAP"

Identifique orígenes de datos para agregarlos a una base de conocimientos o conector

Identifica o crea los documentos (fuentes de datos) que residen en tu sistema de archivos FSx para ONTAP que integrarás en tu base de conocimientos. Estas fuentes de datos permiten a la base de conocimientos proporcionar respuestas precisas y personalizadas a las consultas de los usuarios basadas en datos relevantes para su organización.

Número máximo de orígenes de datos

El número máximo de orígenes de datos admitidos es 10.

Ubicación de los orígenes de datos

Los orígenes de datos pueden almacenarse en un único volumen o en una carpeta dentro de un volumen, en un recurso compartido de SMB o exportación NFS en un sistema de archivos Amazon FSx para NetApp ONTAP. Los orígenes de datos también pueden almacenarse en Amazon FSx para volúmenes de NetApp ONTAP que se encuentran en una relación de protección de datos de NetApp SnapMirror.

No puede seleccionar documentos individuales dentro de un volumen o carpeta, por lo tanto, debe asegurarse de que cada volumen o carpeta que contenga orígenes de datos no contenga documentos extraños que no deberían integrarse con su base de conocimientos.

Puede añadir varias fuentes de datos a cada base de conocimientos, pero todos ellos tienen que residir en sistemas de archivos FSx para ONTAP a los que puede acceder desde su cuenta de AWS.

El tamaño máximo de archivo para cada origen de datos es de 50 MB.

Protocolos compatibles

Las bases de conocimientos admiten datos de volúmenes que usan protocolos NFS o SMB/CIFS. Al seleccionar archivos almacenados con el protocolo SMB, debe introducir la información de Active Directory para que la base de conocimientos pueda acceder a los archivos de esos volúmenes. Esto incluye el dominio de Active Directory, la dirección IP, el nombre de usuario y la contraseña.

Al almacenar su fuente de datos en un recurso compartido (archivo o directorio) al que se accede a través de SMB, los usuarios o grupos de chatbot solo pueden acceder a los datos. Cuando se habilita esta “capacidad de detección de permisos”, el sistema AI comparará el correo electrónico del usuario en auth0 con los usuarios autorizados para ver o usar los archivos en el recurso compartido SMB. El chatbot proporcionará respuestas basadas en los permisos de usuario para los archivos incrustados.

Por ejemplo, si ha integrado 10 archivos (fuentes de datos) en su base de conocimientos, y 2 de los archivos son archivos de recursos humanos que contienen información restringida, solo los usuarios de chatbot que están autenticados para acceder a esos archivos 2 recibirán respuestas del chatbot que incluyen datos de esos archivos.

Formatos de archivo de origen de datos compatibles

Los siguientes formatos de archivos de fuentes de datos actualmente son compatibles con las bases de conocimiento GenAI de Workload Factory.

Formato de archivo	Extensión
Apache Parquet [1]	.parquet
Valores separados por comas file[1]	.csv
Formato de intercambio de gráficos	.gif
JPEG	.jpg or.jpeg
JSON y JSONP[1]	.json
Rebaja	.md
Microsoft Word	.doc o .docx
Texto sin formato	.txt
Formato de documento portátil	.pdf
Gráficos de red portátiles	.png
Imagen WebP	.webp

Ponga en marcha la infraestructura de GenAI

Necesita implementar la infraestructura de GenAI para el marco RAG en su entorno antes de crear bases de conocimientos, conectores y aplicaciones de FSx para ONTAP para su organización. Los principales componentes de la infraestructura son el servicio Bedrock de Amazon, una instancia de máquina virtual para el motor GenAI de NetApp y un sistema de archivos FSx para ONTAP.

La infraestructura implementada puede admitir varias bases de conocimientos, bots conversacionales y conectores, por lo que, por lo general, solo tendrá que realizar esta tarea una vez.

Detalles de la infraestructura

Su implementación de GenAI debe estar en una región de AWS que tenga habilitado Amazon Bedrock. "Vea la lista de regiones admitidas"

La infraestructura consta de los siguientes componentes.

Servicio de Amazon Bedrock

Amazon Bedrock es un servicio totalmente gestionado que te permite usar modelos de base (FMS) de las principales empresas de IA a través de una única API. También proporciona las funcionalidades que necesita para crear aplicaciones de IA generativas seguras.

"Más información sobre Amazon Bedrock"

Amazon Q Business

Amazon Q se basa en Amazon Bedrock para proporcionar un asistente de IA generativa totalmente administrado que puede utilizar para responder preguntas y generar contenido basado en la información de

sus fuentes de datos.

"Obtenga más información sobre Amazon Q Business"

Máquina virtual para el motor NetApp GenAI

El motor NetApp GenAI se pone en marcha durante este proceso. Proporciona la potencia de procesamiento para ingerir los datos de sus orígenes de datos y luego escribir esos datos en la base de datos vectorial.

FSX para sistema de archivos ONTAP

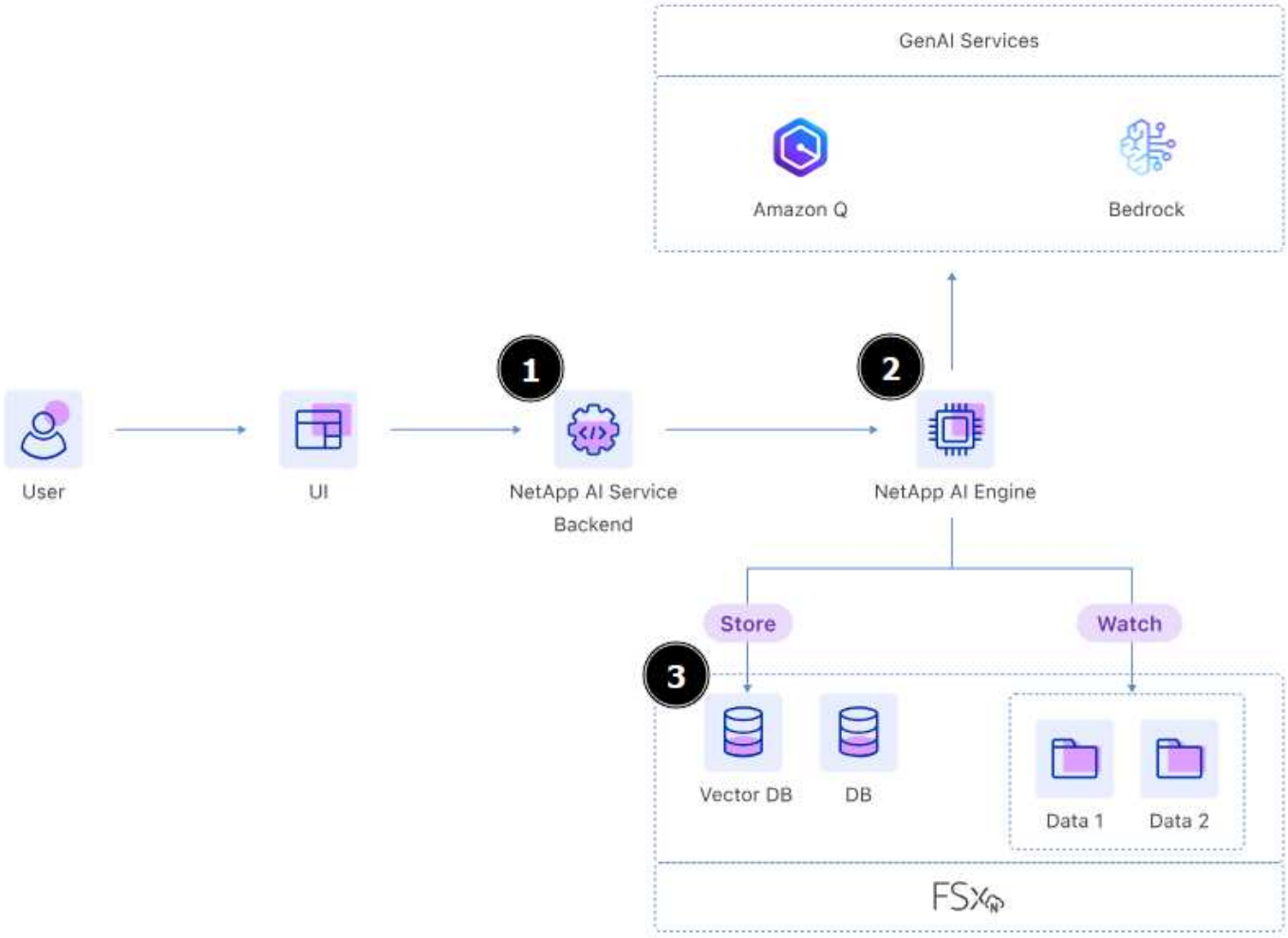
El sistema de archivos FSx para ONTAP proporciona el almacenamiento para su sistema GenAI.

Se implementa un único volumen que contendrá la base de datos vectorial que almacena los datos que ha generado el modelo base basado en sus orígenes de datos.

Las fuentes de datos que integrarás en tu base de conocimientos pueden residir en el mismo sistema de archivos FSx for ONTAP o en un sistema diferente.

El motor GenAI de NetApp supervisa e interactúa con ambos volúmenes.

La siguiente imagen muestra la infraestructura de GenAI. Durante este procedimiento, se implementan los componentes numerados 1, 2 y 3. Los demás elementos deben estar en su lugar antes de iniciar la puesta en marcha.



Ponga en marcha la infraestructura de GenAI

Tendrás que introducir tus credenciales de AWS y seleccionar el sistema de archivos FSx para ONTAP para implementar la infraestructura de generación aumentada de recuperación (RAG).

Antes de empezar

Asegúrese de que su entorno cumple con los requisitos de las bases de conocimientos o conectores, dependiendo de cuál elija, antes de iniciar este procedimiento.

- "Requisitos de la base de conocimientos"
- "Requisitos del conector"

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Revise el diagrama de infraestructura y seleccione **Siguiente**.
4. Complete los elementos en la sección **AWS settings**:
   a. **Credenciales de AWS**: Seleccione o agregue las credenciales de AWS que proporcionan permisos para implementar los recursos de AWS.
   b. **Ubicación**: Seleccione una región, VPC y subred de AWS.

      La implementación de GenAI debe estar en una región de AWS que tenga habilitado Amazon Bedrock. "Vea la lista de regiones admitidas"
5. Complete los elementos en la sección **Configuración de infraestructura**:
   a. **Etiquetas**: Ingrese cualquier par clave/valor de etiqueta que desee aplicar a todos los recursos de AWS que forman parte de esta implementación. Estas etiquetas son visibles en la Consola de administración de AWS y en el área de información de infraestructura dentro de Workload Factory, y pueden ayudarlo a realizar un seguimiento de los recursos de Workload Factory.
6. Complete la sección **Conectividad**:
   a. **Par claves**: Seleccione un par de claves que le permita conectarse de forma segura a la instancia del motor NetApp GenAI.
7. Completa la sección **AI ENGINE**:
   a. **Nombre de instancia**: opcionalmente, seleccione **Definir nombre de instancia** e ingrese un nombre personalizado para la instancia del motor de IA. El nombre de la instancia aparece en la Consola de administración de AWS y en el área de información de infraestructura dentro de Workload Factory, y puede ayudarlo a realizar un seguimiento de los recursos de Workload Factory.
8. Seleccione **Desplegar** para comenzar la implementación.

> Si la implementación falla con un error de credenciales, puede obtener más detalles del error seleccionando los hipervínculos en el mensaje de error. Puede ver una lista de permisos que faltan o están bloqueados, así como una lista de permisos que la carga de trabajo de GenAI necesita para que pueda implementar la infraestructura de GenAI.

Resultado

Workload Factory comienza a implementar la infraestructura del chatbot. Este proceso puede tardar hasta 10 minutos.

Durante el proceso de despliegue, se configuran los siguientes elementos:

- La red se configura junto con los extremos privados.
- Se crean el rol de IAM, el perfil de instancia y el grupo de seguridad.
- Se despliega la instancia de máquina virtual para el motor GenAI.
- Amazon Bedrock está configurado para enviar registros a Amazon CloudWatch Logs, utilizando un grupo de registros con el prefijo `/aws/bedrock/`.
- El motor GenAI está configurado para enviar registros a Amazon CloudWatch Logs, utilizando un grupo de registros con el nombre `/netapp/wlmai/<tenancyAccountId>/randomId` , dónde `<tenancyAccountID>` es el "ID de cuenta de la consola de NetApp" para el usuario actual.

Cree una base de conocimientos de GenAI

Una vez que haya implementado la infraestructura de IA e identificado las fuentes de datos que integrará en su base de conocimiento desde sus almacenes de datos de FSx para ONTAP , estará listo para crear la base de conocimiento utilizando Workload Factory. Como parte de este paso, también definirás las características de la IA y crearás iniciadores de conversación.

Asegúrese de que su entorno cumpla con la "requisitos" para la base de conocimientos antes de continuar.

Acerca de esta tarea

Las bases de conocimiento tienen dos modalidades de integración de datos: *Modo público* y *modo empresarial*.

Modo público

Se puede utilizar una base de conocimientos sin integrar las fuentes de datos de su organización. En este caso, una aplicación integrada con la base de conocimientos solo proporcionará resultados de información disponible públicamente en Internet. Esto se conoce como integración *modo público*.

Modo empresarial

En la mayoría de los casos, querrá integrar las fuentes de datos de su organización en la base de conocimientos. Esto se conoce como una integración *Enterprise mode* porque proporciona conocimiento de su empresa.

Las fuentes de datos de su organización pueden contener información de identificación personal (PII). Para proteger esta información confidencial, puede habilitar *barandillas de datos* al crear y configurar bases de conocimiento. Las barreras de protección de datos, impulsadas por NetApp Data Classification, identifican y enmascaran la información PII, haciéndola inaccesible e irrecuperable.

"Obtenga más información sobre la clasificación de datos de NetApp" .

NetApp Workload Factory para GenAI no enmascara información personal confidencial (SPii). Referirse a"tipos de datos personales confidenciales" para obtener más información sobre este tipo de datos.

Las barandillas de datos se pueden habilitar o deshabilitar en cualquier momento. Si cambia la habilitación de las barandillas de datos, Workload Factory escanea toda la base de conocimientos desde cero, lo que genera un costo.

Crear y configurar la base de conocimientos

La base de conocimientos define características como los modelos de IA Bedrock y el formato de incrustación que desea utilizar para crear su base de conocimientos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En el menú Bases de conocimiento y conectores, seleccione el menú desplegable **Crear nuevo** y elija **Base de conocimiento NetApp GenAI para Bedrock**.
4. En la página Crear base de conocimientos de NetApp GenAI, configure los ajustes de la base de conocimientos:

Detalles de la base de conocimientos

1. **Nombre**: Introduzca el nombre que desea utilizar para la base de conocimientos.
2. **Descripción**: Introduzca una descripción detallada de la base de conocimientos.
3. **Bedrock**: elija la región donde Amazon Bedrock está disponible para su cuenta de AWS.

Ingestión

1. **Modelo de incrustación**:

   - Elija un modelo de inserción para utilizar en la base de conocimientos. El modelo de incrustación define cómo se convertirán sus datos en incrustaciones vectoriales para la base de conocimiento. Workload Factory admite los siguientes modelos:
   - Titan Embeddings G1 - Texto
   - Titan Embedding Text v2
   - Titán incrustaciones multimodales G1
   - Insertar inglés
   - Incrustar multilingüe

     Tenga en cuenta que ya debe haber habilitado el modelo de incrustación de Amazon Bedrock.

   "Más información sobre Amazon Titan"

   - Si corresponde, seleccione el tipo de inferencia que coincida con la configuración del modelo de inserción seleccionado.
2. **Barreras de protección de datos**: elija si desea habilitar o deshabilitar las barreras de protección de datos. "Obtenga información sobre las barreras de seguridad de datos, impulsadas por la clasificación de datos de NetApp" .

   Se deben cumplir los siguientes requisitos previos para activar las barreras de datos.

   - Se requiere una cuenta de servicio para comunicarse con NetApp Data Classification. Debe tener el rol *Administrador de la organización* en su cuenta de inquilino de NetApp Console para la creación de una cuenta de servicio. Un miembro que tenga el rol de administrador de la organización puede completar todas las acciones en el ."Aprenda a agregar un rol a un miembro en la consola de NetApp"
   - El motor de IA debe tener acceso a la"Punto final de la API de la consola de NetApp" .

- Necesitará hacer lo siguiente como se describe en"Documentación de clasificación de datos de NetApp" :
  i. Crear un agente de consola
  ii. Asegúrese de que su entorno pueda cumplir con los requisitos previos
  iii. Implementar la clasificación de datos de NetApp

> La función de barandillas de datos no se admite al ingerir archivos de datos estructurados como CSV, JSON, JSONP o Parquet.

Configuración de chat y recuperación

1. **Modelo de chat**:
   - Elija entre varios modelos de chat que están integrados en Amazon Bedrock. Tenga en cuenta que ya debe haber habilitado el modelo de chat desde Amazon Bedrock.
   - Si corresponde, seleccione el tipo de inferencia que coincida con la configuración del modelo seleccionado.
2. **Configuración del chat**:
   - Elige una temperatura para el chatbot para configurar la aleatoriedad y creatividad de las respuestas. Una temperatura más baja produce respuestas más predecibles, y una temperatura más alta produce respuestas más variadas.
   - Elija una longitud máxima de respuesta para configurar qué tan detalladas serán las respuestas. Las longitudes de respuesta más largas utilizan más tokens de respuesta y pueden generar un costo mayor.
3. **Modo de pensamiento**: cuando el modo de pensamiento está habilitado, el chatbot tardará más tiempo en procesar las consultas y los resultados generalmente serán más precisos. Cuando habilita el modo de pensamiento, puede controlar cuántos tokens de razonamiento se utilizan al generar resultados. El uso de más tokens de razonamiento puede generar respuestas más precisas, pero podría implicar un mayor costo.
4. **Reranking**: habilite o deshabilite el reranking, lo que puede mejorar la relevancia y la calidad de los resultados de la consulta. Elija un modelo de chat estándar o un modelo de reranking especializado para utilizar en el reranking. Las opciones del modelo de reranker solo se muestran si están disponibles en su región. Seleccione el tipo de inferencia que coincida con la configuración del modelo seleccionado.
5. * Iniciadores de conversación *: Elija si desea proporcionar hasta cuatro mensajes de inicio de conversación que se muestran a los usuarios que interactúan con un chatbot que utiliza esta base de conocimientos. Le recomendamos que active esta configuración.

   Si activa los arrancadores de conversación, se selecciona el modo automático de forma predeterminada. El modo manual solo se puede activar después de agregar fuentes de datos a su base de conocimientos. "Aprenda a modificar la configuración de la base de conocimientos".

Definiciones de almacenamiento

1. *Sistema de archivos FSx para ONTAP *: cuando se define una nueva base de conocimiento, Workload Factory crea un nuevo volumen de Amazon FSx for NetApp ONTAP para almacenarla. Elija un nombre de sistema de archivos existente y una SVM (también llamada VM de almacenamiento) donde se creará el nuevo volumen.
2. **Política de instantáneas**: elija una política de instantáneas de la lista de políticas existentes definidas en el inventario de almacenamiento de Workload Factory. Se crearán automáticamente instantáneas

recurrentes de la base de conocimientos con una frecuencia basada en la política de instantáneas que seleccione.

3. **S3 Bucket**: si los resultados de la consulta del chatbot contienen datos estructurados, GenAI puede almacenar los resultados en un bucket S3. Para utilizar esta función, habilite la configuración **Activar depósito S3** y elija un depósito S3 asociado con su cuenta de la lista. Cuando estos resultados se almacenan en un bucket S3, puedes descargarlos utilizando el enlace de descarga dentro de la sesión de chat.

   Si la política de Snapshot que necesita no existe, puede "crear una política de snapshot" hacerlo en la máquina virtual de almacenamiento que contiene el volumen.

4. Seleccione **Crear base de conocimiento** para agregar la base de conocimiento a GenAI.

   Aparece un indicador de progreso mientras se crea la base de conocimientos.

   Después de crear la base de conocimientos, tiene la opción de agregar un origen de datos a la nueva base de conocimientos o de finalizar el proceso sin agregar un origen de datos. Le recomendamos que seleccione **Añadir fuente de datos** y agregue una o más fuentes de datos ahora.

Agregue orígenes de datos a la base de conocimientos

Puede agregar uno o más orígenes de datos para rellenar la base de conocimientos con los datos de la organización.

Acerca de esta tarea

El número máximo de orígenes de datos admitidos es 10.

Pasos

1. Después de seleccionar **Agregar fuente de datos**, seleccione el tipo de fuente de datos que desea agregar:
   - Agregar el sistema de archivos FSx para ONTAP (usar archivos de un volumen FSx para ONTAP existente)
   - Agregar sistema de archivos (utilizar archivos de un recurso compartido SMB o NFS genérico)

Agregar un sistema de archivos FSx para ONTAP

1. **Seleccione un sistema de archivos**: Seleccione el sistema de archivos FSX for ONTAP donde residen sus archivos de origen de datos y seleccione **Siguiente**.
2. **Selecciona un volumen**: Selecciona el volumen en el que residen tus archivos de origen de datos y selecciona **Siguiente**.

   Al seleccionar los archivos almacenados mediante el protocolo SMB, deberá introducir la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.

3. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.
4. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

       La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:
     - En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

       Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

     Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

     Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

   - En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la

inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

5. En la sección **Permission Aware**, que solo está disponible cuando el origen de datos que seleccionó está en un volumen que utiliza el protocolo SMB, puede habilitar o deshabilitar las respuestas con permiso:
   - **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
   - **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.
6. Seleccione **Agregar** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos NFS genérico

1. **Seleccionar un sistema de archivos**: Ingrese la dirección IP o FQDN para el host del sistema de archivos donde residen los archivos de origen de datos, elija el protocolo NFS para el recurso compartido de red y seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre de la exportación NFS y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar la exportación completa o solo carpetas específicas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:
   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

     > La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

> Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

- En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos SMB genérico

1. **Seleccionar sistema de archivos**:
   a. Ingrese la dirección IP o FQDN del host del sistema de archivos donde residen los archivos de origen de datos.
   b. Seleccione el protocolo SMB para el recurso compartido de red.
   c. Ingrese la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.
   d. Seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre del recurso compartido SMB y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar todo el recurso compartido o solo algunas carpetas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:

- * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
- * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

> La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

◦ **Consciente de permisos**: habilitar o deshabilitar respuestas que tengan en cuenta los permisos:

- **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
- **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.

◦ **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

> Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

◦ En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Resultado

El origen de datos comienza a integrarse en su base de conocimientos. El estado cambia de “Embedding” a

“Embedded” cuando el origen de datos está completamente embebido.

Después de agregar una única fuente de datos a la base de conocimientos, puede probarla localmente en la ventana del simulador de chatbot y realizar los cambios necesarios antes de que el chatbot esté disponible para sus usuarios. También puede seguir los mismos pasos para agregar orígenes de datos adicionales a la base de conocimientos.

Pruebe una base de conocimientos de GenAI

Después de crear la base de conocimientos, podrá probarla localmente utilizando el simulador de chatbot y realizar los cambios necesarios antes de poner la base de conocimientos a disposición de los usuarios a través de una aplicación de chatbot.

Acerca de esta tarea

Prueba tu base de conocimientos para asegurarte de que funciona como esperas, y puedes personalizar los temas de inicio de conversación que quieres que estén disponibles de forma predeterminada para los usuarios de chatbot de esta base de conocimientos. El simulador de chatbot se ejecuta contra todas las fuentes de datos que se han incrustado en la base de conocimientos.

Puede probar una base de conocimientos chateando con sus fuentes de datos incrustadas en el simulador de chatbot. Tenga en cuenta que ninguna de las interacciones o conocimientos se captura en la base de datos de vectores de GenAI al probar la base de conocimientos localmente.

Realizará la mayoría de sus pruebas dentro de Workload Factory antes de implementar la base de conocimiento en una aplicación para sus usuarios. Si necesita realizar cambios en su fuente de datos o en el funcionamiento del chatbot, deberá hacerlo ahora, antes de publicar su base de conocimientos.

Puede cambiar el tamaño y volver a abrir la ventana del simulador de chatbot, y copiar las preguntas y las respuestas en el portapapeles.

Algunas de las tareas que querrás realizar para probar tu chatbot son:

- Introduzca un gran número de preguntas relevantes para su organización para asegurarse de que las respuestas sean las esperadas.
- Personaliza los temas de inicio de conversación que quieres que estén disponibles de forma predeterminada para tus usuarios en la aplicación chatbot.
- Asegúrese de que el contenido atribuido que se proporciona en la parte inferior de las respuestas del chatbot contiene las referencias correctas.

Pasos

1. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea probar.

   El simulador de chatbot aparece en el panel derecho. Si se define, también se muestran los iniciadores de conversación existentes.

2. En el campo de entrada de chatbot, ingrese una pregunta o petición de datos y seleccione para ver cómo responde su chatbot con su conocimiento organizativo.

- Puede ver las fuentes utilizadas para producir la respuesta ampliando la lista **Sources** debajo de la respuesta. Esto proporciona una lista de archivos utilizados para generar la respuesta. Puede ver y copiar los fragmentos de datos utilizados en cada ruta de archivo y volumen a cada archivo pasando el cursor sobre el nombre del archivo.
- Si se incluyen tablas en la respuesta, puede ordenar los datos en cada columna y copiar cada tabla al portapapeles.
- Si los resultados de la respuesta contienen datos estructurados y la función **S3 Bucket** está habilitada para la base de conocimiento, GenAI almacena los resultados en un S3 Bucket. Puedes descargar los resultados del bucket usando el enlace **Descargar resultados** dentro de la sesión de chat.

3. Si necesita actualizar alguna de sus fuentes de datos para que su base de conocimientos proporcione respuestas más centradas, realice esos cambios ahora y vuelva a probar la base de conocimientos.

Active la autenticación externa para una base de conocimientos de GenAI

Active la autenticación de una base de conocimientos para que se requiera la validación de tokens y ACL cuando se utilicen los puntos finales de la API para integrar una base de conocimientos con una aplicación de chatbot. Cuando activa la autenticación, configura la configuración de un token web JSON que se utilizará para las solicitudes de API a una base de conocimientos de clientes de chatbot.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos para la que desea activar la autenticación.
4. Selecciona y selecciona **Administrar base de conocimientos**.
5. Seleccione el menú **Acciones** y seleccione **Administrar configuración de autenticación**.
6. Configurar la autenticación:

   a. Selecciona **Activar configuración de autenticación**.

   b. Proporcione la información obligatoria. Se proporcionan ejemplos, pero debe obtener los valores de estos campos del proveedor de autenticación:

      - **Algoritmos**: El algoritmo de firma que utiliza su proveedor de autenticación.
      - **Audiencia** (Opcional): Una cadena que contiene el destinatario previsto del token (a veces una URL).
      - **Emisor**: Una cadena que identifica al proveedor que emitió el token.

        Por ejemplo, Amazon Cognito utiliza cadenas de emisores con el siguiente formato:

        ```
        https://cognito-idp-<region>.amazonaws.com/<UserPoolID>
        ```

Donde `<region>` es la región de AWS que contiene el grupo de usuarios, y `<UserPoolID>` es su ID de grupo de usuarios. Puede recuperar su ID de grupo de usuarios con el siguiente comando:

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

- **JWKS URI**: La cadena URI que proporciona claves públicas necesarias para verificar las firmas de este token.

  Por ejemplo, Amazon Cognito utiliza cadenas URI JWKS con el siguiente formato:

```
https://cognito-idp.<region>.amazonaws.com/<userPoolId>/.well-known/jwks.json
```

+
Donde `<region>` es la región de AWS que contiene el grupo de usuarios, y `<UserPoolID>` es su ID de grupo de usuarios. Puede recuperar su ID de grupo de usuarios con el siguiente comando:

```
aws cognito-idp list-user-pools --max-results=60 --output=table
```

7. Seleccione **Guardar**.

Resultado

La autenticación de la base de conocimientos ahora está activa, y puede utilizar puntos finales de API para interactuar con la base de conocimientos e integrar la base de conocimientos con una aplicación de chatbot.

Publique una base de conocimientos de GenAI y vea el punto final único

Después de crear y probar su base de conocimientos localmente, puede publicar la base de conocimientos para que pueda integrarse con una aplicación de chatbot que permita a sus usuarios consultar la base de conocimientos.

Acerca de esta tarea

La publicación de la base de conocimientos le permitirá utilizarla en aplicaciones de chat. La acción de publicación activa la API de Workload Factory para generar y publicar puntos finales únicos. Después de la publicación, la base de conocimientos se vuelve accesible para las aplicaciones de chat y los puntos finales de la API están listos para la integración.

Cada base de conocimientos que publique tiene puntos finales únicos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea publicar.

4. Selecciona y selecciona **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación y la lista de todos los orígenes de datos incrustados.

5. Selecciona el menú **Acciones** y selecciona **Publicar**.

   Workload Factory publica la base de conocimientos. En la página de detalles de la base de conocimientos, el estado cambia de **No publicado** a **Publicado.**

   Ahora puede obtener detalles sobre el extremo único de la base de conocimientos.

6. Junto al estado publicado, selecciona **Ver**.

   Se muestran detalles sobre cómo acceder a la base de conocimientos mediante la API de Workload Factory.

7. En el cuadro de diálogo **Ver información publicada**, copie los puntos finales de la API que puede utilizar para integrar la base de conocimientos con una aplicación.

   Para obtener más información sobre los puntos finales de la API, vaya a "Documentación de API" y seleccione **AI > Externo**.

   Antes de poder utilizar estos puntos finales, debe obtener un token de usuario de su proveedor de autenticación.

Resultado

Ahora tiene una base de conocimientos publicada y el punto final único que puede utilizar para integrar la base de conocimientos con una aplicación de chatbot.

Utilice la aplicación de chatbot de ejemplo externo de GenAI

Después de configurar, activar y publicar una base de conocimientos, los desarrolladores de aplicaciones externas pueden configurar y ejecutar la aplicación de chatbot de ejemplo de código abierto proporcionada por NetApp para interactuar con su base de conocimientos y aprender a usar la API de Workload Factory para crear sus propias aplicaciones de IA generativa.

Pasos

1. "Crear una base de conocimientos".
2. "Active la autenticación" para la base de conocimientos que ha creado.

   Esto permite que la base de conocimientos autentique las solicitudes de API y hace que la validación de tokens y las ACL sean necesarias cuando se utilizan los extremos de la API.

   > Las aplicaciones de chat externas que se integran con esta base de conocimientos deberán utilizar el mismo proveedor de autenticación (emisor) que configure en la configuración de autenticación de la base de conocimientos.

3. "Publique la base de conocimientos" Para habilitar el acceso API para aplicaciones externas.

Después de publicar una base de conocimientos, los puntos finales de la API son accesibles externamente y puede integrar la base de conocimientos con una aplicación de chat externa (como la aplicación de chatbot de ejemplo).

4. Descargue el paquete de aplicación chatbot de ejemplo de "GitHub".
5. Instale y ejecute la aplicación chatbot siguiendo las instrucciones del archivo README incluido en el paquete.
6. Vaya a "http://localhost:9091" para iniciar sesión en la aplicación.

   Aparece el ejemplo de aplicación de chatbot.

Leer más

"Documentación de la API de Workload Factory"

Cree una aplicación GenAI basada en RAG

Después de construir su base de conocimientos y probar su chatbot, está listo para configurar la aplicación que permitirá a sus usuarios consultar el chatbot.

"Descubre cómo crear una aplicación de IA basada en RAG en FSx para ONTAP"

Qué puede hacer después con GenAI

Ahora que ha creado una base de conocimientos que utiliza sus datos empresariales y los ha puesto en marcha para sus usuarios, puede gestionar la base de conocimientos, las fuentes de datos y la infraestructura RAG, incluidos los sistemas de archivos FSx para ONTAP.

Algunas de las tareas que puede realizar para administrar los componentes de la base de conocimientos son:

- Actualice el contenido de sus fuentes de datos, o agregue nuevas fuentes de datos, y sincronice esos cambios con su base de conocimientos y chatbot.
- Gestione la configuración del origen de datos, incluida la estrategia de fragmentación y el reconocimiento de permisos (para el acceso a archivos SMB).
- Gestione la configuración de su base de conocimientos, incluidos el modelo de chat y los temas de inicio de conversación.
- Anule la publicación de una base de conocimientos o vuelva a publicarla después de realizar cambios.
- Realice backups y proteja los datos importantes en su sistema de archivos FSx para ONTAP para asegurarse de que sus datos de base de conocimientos y otros componentes de la infraestructura estén siempre disponibles.

Para obtener información sobre cómo administrar su sistema de archivos FSx para ONTAP , vaya a "Documentación de Workload Factory para Amazon FSx for NetApp ONTAP" para ver las capacidades de respaldo y protección que puede utilizar.

[1] La función de barandillas de datos no es compatible cuando se ingieren archivos de datos estructurados en bases de conocimiento.

Utilice GenAI para crear conectores para Amazon Q Business

Manos a la obra

Inicio rápido para conectores GenAI

Comience a crear un conector de NetApp para Amazon Q Business utilizando los datos de su organización que existen en Amazon FSx para los sistemas de archivos de NetApp ONTAP. Después de crear un conector, los usuarios finales pueden acceder al asistente de Amazon Q Business para obtener respuestas centradas en la organización a sus preguntas.

Iniciar sesión en Workload Factory

Necesitarás "Configurar una cuenta con Workload Factory" e inicie sesión usando uno de los "experiencias de consola" .

Configure su entorno para cumplir los requisitos de GenAI

Necesitarás credenciales de AWS para implementar la infraestructura de AWS, un sistema de archivos FSX for ONTAP implementado y descubierto, la lista de fuentes de datos que deseas integrar en tu conector, el acceso a la aplicación de Amazon Q Business y mucho más.

"Obtenga más información sobre los requisitos de GenAI".

Identifique el sistema de archivos FSx para ONTAP que contiene los orígenes de datos

Las fuentes de datos que integrarás en tu conector pueden ubicarse en un único sistema de archivos FSx para ONTAP o en varios sistemas de archivos FSx para ONTAP. Si estos sistemas están en VPC diferentes, deben ser accesibles dentro de la misma red o los VPC deben estar conectados y usar la misma región y cuenta de AWS que el motor AI.

"Aprenda a identificar fuentes de datos".

Ponga en marcha la infraestructura de GenAI

Inicie el asistente de implementación de infraestructuras para implementar la infraestructura de GenAI en su entorno AWS. Este proceso pone en marcha una instancia EC2 para el motor NetApp GenAI y un volumen en un sistema de archivos FSx para ONTAP para contener las bases de datos del motor de IA de NetApp. El volumen se utiliza para almacenar información sobre el conector.

"Descubre cómo implementar la infraestructura de GenAI".

El futuro

Ahora puede crear un conector para Amazon Q Business para proporcionar respuestas centradas en la organización a los usuarios finales.

Requisitos del conector GenAI

Asegúrese de que Workload Factory y AWS estén configurados correctamente antes de crear un conector de NetApp para Amazon Q Business.

Requisitos básicos de GenAI

GenAI tiene requisitos generales que su entorno necesita cumplir antes de empezar.

Inicio de sesión y cuenta de Workload Factory

Necesitarás "Configurar una cuenta con Workload Factory" e inicie sesión usando uno de los "experiencias de consola" .

Credenciales y permisos de AWS

Debe agregar credenciales de AWS a Workload Factory con permisos de lectura y escritura, lo que significa que utilizará Workload Factory en modo de lectura y escritura para GenAI.

Los permisos de modo *básico* y de modo *solo lectura* no son compatibles en este momento.

Al configurar sus credenciales, al seleccionar los permisos que se muestran a continuación, tendrá acceso completo para administrar los sistemas de archivos de FSx para ONTAP y para implementar y administrar la instancia de GenAI EC2 y otros recursos de AWS necesarios para su base de conocimientos y chatbot.

"Aprenda a agregar credenciales de AWS a Workload Factory"

Requisitos para el conector NetApp para Amazon Q Business

Asegúrese de que su entorno cumple los siguientes requisitos específicos para los conectores de Amazon Q Business.

Aplicación Amazon Q Business

Debe crear una aplicación de Amazon Q Business o utilizar una existente.

- Asegúrese de que la aplicación existe en una de sus regiones de AWS.
- Asegúrese de que "se ha creado un índice" tiene para la aplicación.
- Asegúrese de que la aplicación no está en un estado fallido.

FSX para sistema de archivos ONTAP

Necesita un mínimo de un sistema de archivos FSx para ONTAP:

- El motor NetApp GenAI utilizará (o creará, si no existe) un sistema de archivos para almacenar información sobre el conector.

  Este sistema de archivos FSx para ONTAP debe utilizar volúmenes de FlexVol. No se admiten los volúmenes de FlexGroup.

- Uno o más sistemas de archivos contendrán las fuentes de datos que agregará al conector.

  Un sistema de archivos FSx para ONTAP se puede utilizar con ambos fines o puede utilizar varios sistemas de archivos FSx para ONTAP.

- Necesitará conocer la región, la VPC y la subred de AWS donde reside el sistema de archivos AWS FSx para ONTAP.

- Tendrá que considerar los pares clave/valor de etiqueta que desea aplicar a los recursos de AWS que forman parte de esta implementación (opcional).
- Necesitará conocer la información del par de claves que le permite conectarse de forma segura a la instancia del motor de NetApp AI.

"Descubre cómo implementar y gestionar sistemas de archivos FSx para ONTAP"

Identifique orígenes de datos para agregarlos a un conector

Identifica o crea los documentos (fuentes de datos) que residen en tu sistema de archivos FSx para ONTAP que integrarás en tu conector. Estas fuentes de datos permiten que Amazon Q Business proporcione respuestas precisas y personalizadas a las consultas de los usuarios basadas en datos relevantes para su organización.

Número máximo de orígenes de datos

El número máximo de orígenes de datos admitidos es 10.

Ubicación de los orígenes de datos

Los orígenes de datos pueden almacenarse en un único volumen o en una carpeta dentro de un volumen, en un recurso compartido de SMB o exportación NFS en un sistema de archivos Amazon FSx para NetApp ONTAP. Los orígenes de datos también pueden almacenarse en Amazon FSx para volúmenes de NetApp ONTAP que se encuentran en una relación de protección de datos de NetApp SnapMirror.

No puede seleccionar documentos individuales dentro de un volumen o carpeta, por lo tanto, debe asegurarse de que cada volumen o carpeta que contenga orígenes de datos no contenga documentos extraños que no deberían integrarse con su base de conocimientos.

Puede agregar varias fuentes de datos a cada conector, pero todos ellos tienen que residir en los sistemas de archivos de FSx para ONTAP a los que se puede acceder desde su cuenta de AWS.

El tamaño máximo de archivo para cada origen de datos es de 50 MB.

Protocolos compatibles

Los conectores admiten datos de volúmenes que usan protocolos NFS o SMB/CIFS. Al seleccionar archivos almacenados con el protocolo SMB, debe introducir la información de Active Directory para que el conector pueda acceder a los archivos de esos volúmenes. Esto incluye el dominio de Active Directory, la dirección IP, el nombre de usuario y la contraseña.

Al almacenar su fuente de datos en un recurso compartido (archivo o directorio) al que se accede a través de SMB, los usuarios o grupos de chatbot solo pueden acceder a los datos. Cuando se habilita esta “capacidad de detección de permisos”, el sistema AI comparará el correo electrónico del usuario en auth0 con los usuarios autorizados para ver o usar los archivos en el recurso compartido SMB. El chatbot proporcionará respuestas basadas en los permisos de usuario para los archivos incrustados.

Por ejemplo, si ha integrado archivos 10 (fuentes de datos) en su conector, y 2 de los archivos son archivos de recursos humanos que contienen información restringida, solo los usuarios de chatbot que están autenticados para acceder a esos archivos 2 recibirán respuestas del chatbot que incluyen datos de esos archivos.

Al agregar orígenes de datos a un conector de Amazon Q Business, solo se aplican permisos de usuario a los archivos de origen de datos. No se aplican los permisos de grupo.

Si un archivo de su fuente de datos carece de texto (por ejemplo, una imagen sin texto), Amazon Q Business no lo indexa, sino que registra una entrada en Amazon CloudWatch Logs y observa la ausencia de texto.

Formatos de archivo de origen de datos compatibles

Los siguientes formatos de archivos de origen de datos son actualmente compatibles con NetApp Connector para Amazon Q Business.

Formato de archivo	Extensión
Archivo de valores separados por comas	.csv
JSON y JSONP	.json
Rebaja	.md
Microsoft Word	.docx
Texto sin formato	.txt
Formato de documento portátil	.pdf
Microsoft PowerPoint	.ppt o .pptx
Lenguaje de marcado de hipertexto	.html
Lenguaje de marcado extensible	.xml
XSLT	.xslt
Microsoft Excel	.xls
Formato de texto enriquecido	.rtf

Ponga en marcha la infraestructura de GenAI

Necesita implementar la infraestructura de GenAI para el marco RAG en su entorno antes de crear bases de conocimientos, conectores y aplicaciones de FSx para ONTAP para su organización. Los principales componentes de la infraestructura son el servicio Bedrock de Amazon, una instancia de máquina virtual para el motor GenAI de NetApp y un sistema de archivos FSx para ONTAP.

La infraestructura implementada puede admitir varias bases de conocimientos, bots conversacionales y conectores, por lo que, por lo general, solo tendrá que realizar esta tarea una vez.

Detalles de la infraestructura

Su implementación de GenAI debe estar en una región de AWS que tenga habilitado Amazon Bedrock. "Vea la lista de regiones admitidas"

La infraestructura consta de los siguientes componentes.

Servicio de Amazon Bedrock

Amazon Bedrock es un servicio totalmente gestionado que te permite usar modelos de base (FMS) de las principales empresas de IA a través de una única API. También proporciona las funcionalidades que necesita para crear aplicaciones de IA generativas seguras.

"Más información sobre Amazon Bedrock"

Amazon Q Business

Amazon Q se basa en Amazon Bedrock para proporcionar un asistente de IA generativa totalmente administrado que puede utilizar para responder preguntas y generar contenido basado en la información de sus fuentes de datos.

"Obtenga más información sobre Amazon Q Business"

Máquina virtual para el motor NetApp GenAI

El motor NetApp GenAI se pone en marcha durante este proceso. Proporciona la potencia de procesamiento para ingerir los datos de sus orígenes de datos y luego escribir esos datos en la base de datos vectorial.

FSX para sistema de archivos ONTAP

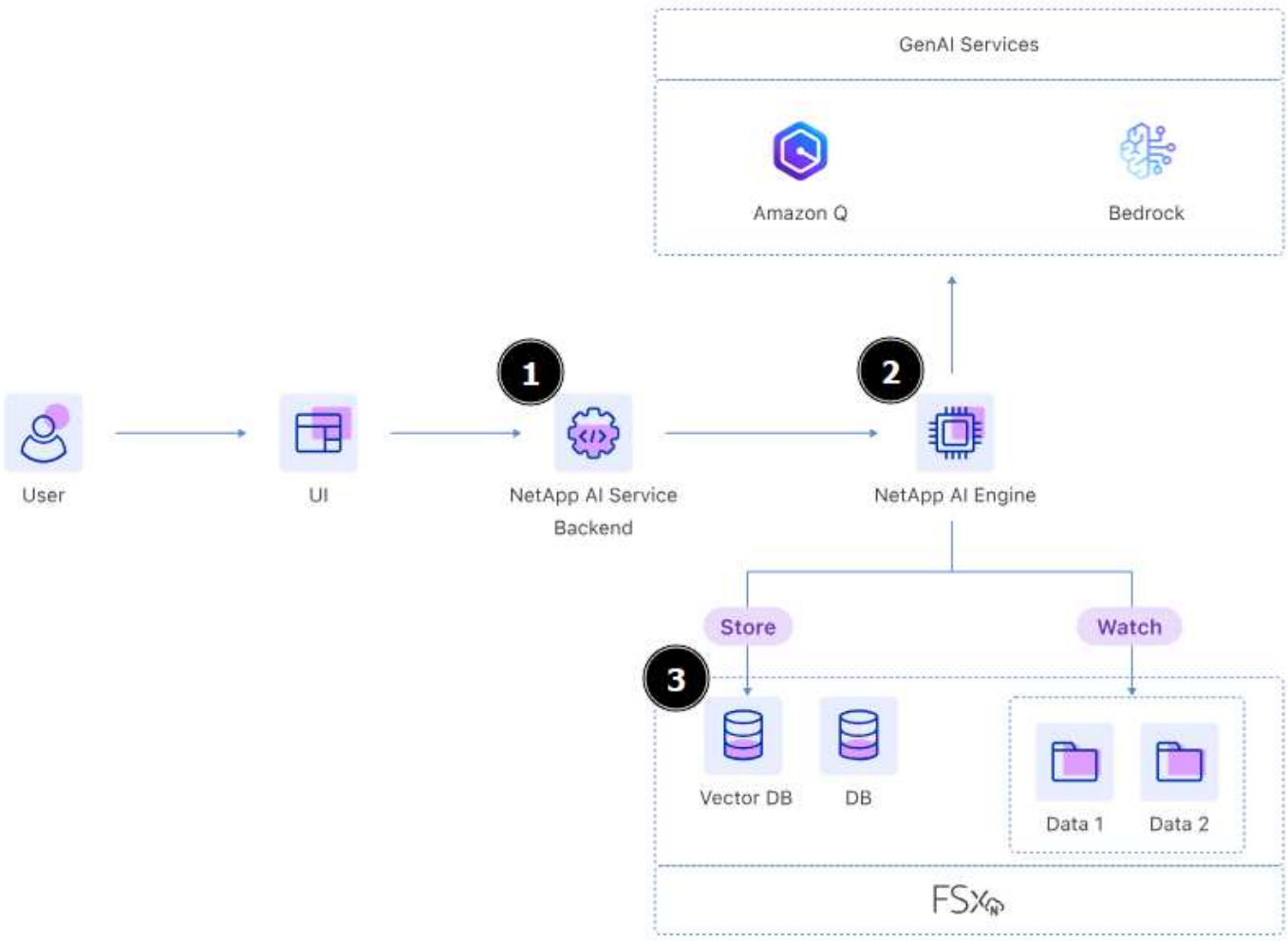
El sistema de archivos FSx para ONTAP proporciona el almacenamiento para su sistema GenAI.

Se implementa un único volumen que contendrá la base de datos vectorial que almacena los datos que ha generado el modelo base basado en sus orígenes de datos.

Las fuentes de datos que integrarás en tu base de conocimientos pueden residir en el mismo sistema de archivos FSx for ONTAP o en un sistema diferente.

El motor GenAI de NetApp supervisa e interactúa con ambos volúmenes.

La siguiente imagen muestra la infraestructura de GenAI. Durante este procedimiento, se implementan los componentes numerados 1, 2 y 3. Los demás elementos deben estar en su lugar antes de iniciar la puesta en marcha.



Ponga en marcha la infraestructura de GenAI

Tendrás que introducir tus credenciales de AWS y seleccionar el sistema de archivos FSx para ONTAP para implementar la infraestructura de generación aumentada de recuperación (RAG).

Antes de empezar

Asegúrese de que su entorno cumple con los requisitos de las bases de conocimientos o conectores, dependiendo de cuál elija, antes de iniciar este procedimiento.

- "Requisitos de la base de conocimientos"
- "Requisitos del conector"

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Revise el diagrama de infraestructura y seleccione **Siguiente**.
4. Complete los elementos en la sección **AWS settings**:
   a. **Credenciales de AWS**: Seleccione o agregue las credenciales de AWS que proporcionan permisos para implementar los recursos de AWS.
   b. **Ubicación**: Seleccione una región, VPC y subred de AWS.

La implementación de GenAI debe estar en una región de AWS que tenga habilitado Amazon Bedrock. "Vea la lista de regiones admitidas"

5. Complete los elementos en la sección **Configuración de infraestructura**:
   a. **Etiquetas**: Ingrese cualquier par clave/valor de etiqueta que desee aplicar a todos los recursos de AWS que forman parte de esta implementación. Estas etiquetas son visibles en la Consola de administración de AWS y en el área de información de infraestructura dentro de Workload Factory, y pueden ayudarlo a realizar un seguimiento de los recursos de Workload Factory.
6. Complete la sección **Conectividad**:
   a. **Par claves**: Seleccione un par de claves que le permita conectarse de forma segura a la instancia del motor NetApp GenAI.
7. Completa la sección **AI ENGINE**:
   a. **Nombre de instancia**: opcionalmente, seleccione **Definir nombre de instancia** e ingrese un nombre personalizado para la instancia del motor de IA. El nombre de la instancia aparece en la Consola de administración de AWS y en el área de información de infraestructura dentro de Workload Factory, y puede ayudarlo a realizar un seguimiento de los recursos de Workload Factory.
8. Seleccione **Desplegar** para comenzar la implementación.

> Si la implementación falla con un error de credenciales, puede obtener más detalles del error seleccionando los hipervínculos en el mensaje de error. Puede ver una lista de permisos que faltan o están bloqueados, así como una lista de permisos que la carga de trabajo de GenAI necesita para que pueda implementar la infraestructura de GenAI.

Resultado

Workload Factory comienza a implementar la infraestructura del chatbot. Este proceso puede tardar hasta 10 minutos.

Durante el proceso de despliegue, se configuran los siguientes elementos:

- La red se configura junto con los extremos privados.
- Se crean el rol de IAM, el perfil de instancia y el grupo de seguridad.
- Se despliega la instancia de máquina virtual para el motor GenAI.
- Amazon Bedrock está configurado para enviar registros a Amazon CloudWatch Logs, utilizando un grupo de registros con el prefijo `/aws/bedrock/`.
- El motor GenAI está configurado para enviar registros a Amazon CloudWatch Logs, utilizando un grupo de registros con el nombre `/netapp/wlmai/<tenancyAccountId>/randomId` , dónde `<tenancyAccountID>` es el "ID de cuenta de la consola de NetApp" para el usuario actual.

Cree un conector de NetApp para Amazon Q Business

Una vez que haya implementado la infraestructura de IA e identificado las fuentes de datos que utilizará de sus almacenes de datos FSx para ONTAP, estará listo para definir un conector de NetApp para Amazon Q Business.

Asegúrese de que su entorno cumple con el "requisitos" para Amazon Q Business antes de continuar.

Acerca de esta tarea

Las fuentes de datos de su organización pueden contener información de identificación personal (PII). Para proteger esta información confidencial, puede habilitar *barandillas de datos* al definir un conector. Las barreras de protección de datos, impulsadas por NetApp Data Classification, identifican y enmascaran la información PII, haciéndola inaccesible e irrecuperable.

"Obtenga más información sobre la clasificación de datos de NetApp" .

NetApp Workload Factory para GenAI no enmascara información personal confidencial (SPII). Referirse a"tipos de datos personales confidenciales" para obtener más información sobre este tipo de datos.

Las barandillas de datos se pueden habilitar o deshabilitar en cualquier momento. Si cambia la habilitación de las barandillas de datos, Workload Factory escanea toda la fuente de datos desde cero, lo que puede generar un costo.

Defina un conector

Cree un conector de NetApp para Amazon Q Business. Este conector permite la comunicación entre GenAI y Amazon Q Business a través de API y fuentes de datos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En el menú Bases de conocimiento y conectores, seleccione el menú desplegable **Crear nuevo** y elija **Conector Amazon Q Business**.
4. En la página Definir conector, configure los ajustes del conector:
   a. **Nombre**: Introduzca el nombre que desea utilizar para el conector.
   b. **Descripción**: Introduzca una descripción detallada del conector.
   c. **Amazon Q**: La región y el nombre de la aplicación para la instancia de Amazon Q Business que desea integrar.
   d. **Barreras de protección de datos**: elija si desea habilitar o deshabilitar las barreras de protección de datos. "Obtenga información sobre las barreras de seguridad de datos, impulsadas por la clasificación de datos de NetApp" .

      Se deben cumplir los siguientes requisitos previos para activar las barreras de datos.

      - Se requiere una cuenta de servicio para comunicarse con NetApp Data Classification. Debe tener el rol *Administrador de la organización* en su cuenta de inquilino de NetApp Console para la creación de una cuenta de servicio. Un miembro que tenga el rol de administrador de la organización puede completar todas las acciones en la consola de NetApp ."Aprenda a agregar un rol a un miembro en la consola de NetApp"
      - El motor de IA debe tener acceso a la"Punto final de la API de la consola de NetApp" .
      - Necesitará hacer lo siguiente como se describe en"Documentación de clasificación de datos de NetApp" :
        A. Crear un agente de consola
        B. Asegúrese de que su entorno pueda cumplir con los requisitos previos
        C. Implementar la clasificación de datos de NetApp

Cuando habilita la función de barandas de datos, GenAI procesa archivos .txt, .md, .csv, .docx y .pdf ingiriendo solo texto sin formato (excluyendo imágenes incrustadas o texto multimedia) y enmascarando cualquier dato privado o confidencial. Todos los demás tipos de archivos se procesan normalmente sin enmascarar datos privados o confidenciales.

e. *Sistema de archivos FSx para ONTAP *: cuando define un nuevo conector de NetApp para Amazon Q Business, Workload Factory crea un nuevo volumen de Amazon FSx for NetApp ONTAP para almacenar la información del conector. Elija un sistema de archivos existente y una SVM (también llamada VM de almacenamiento) donde se creará el nuevo volumen.

f. **Política de instantáneas**: elija una política de instantáneas de la lista de políticas existentes definidas en el inventario de almacenamiento de Workload Factory. GenAI crea automáticamente instantáneas recurrentes del volumen que almacena la información del conector con una frecuencia basada en la política de instantáneas que seleccione.

   Si la política de Snapshot que necesita no existe, puede "crear una política de snapshot" hacerlo en la máquina virtual de almacenamiento que contiene el volumen.

5. Selecciona **Crear conector** para integrar Amazon Q Business con GenAI.

   Aparece un indicador de progreso mientras se crea el conector.

   Después de crear el conector, tiene la opción de agregar un origen de datos al conector para que Amazon Q Business ingiera sus datos y los agregue a su índice. Le recomendamos que seleccione **Añadir fuente de datos** y agregue una o más fuentes de datos ahora.

Agregue orígenes de datos al conector

Puede agregar uno o más orígenes de datos para rellenar el índice de Amazon Q Business con los datos de su organización.

Acerca de esta tarea

- El número máximo de orígenes de datos admitidos es 10.
- Consulte el "Documentación de Amazon Q Business" para obtener información sobre las restricciones de servicio específicas del índice de Amazon Q Business.

Pasos

1. Después de seleccionar **Agregar fuente de datos**, aparece la página **Seleccionar un sistema de archivos**.
2. **Seleccione un sistema de archivos**: Seleccione el sistema de archivos FSX for ONTAP donde residen sus archivos de origen de datos y seleccione **Siguiente**.
3. **Selecciona un volumen**: Selecciona el volumen en el que residen tus archivos de origen de datos y selecciona **Siguiente**.

   Al seleccionar los archivos almacenados mediante el protocolo SMB, deberá introducir la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.

4. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.
5. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos

incluye en los escaneos:

- **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:
  - En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.
  - En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

6. En la sección **Permission Aware**, que solo está disponible cuando el origen de datos que seleccionó está en un volumen que utiliza el protocolo SMB, puede habilitar o deshabilitar las respuestas con permiso:
   - **Activado**: Los usuarios del chatbot que accedan a este conector solo obtendrán respuestas a consultas de fuentes de datos a las que tengan acceso.
   - **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.

> Los permisos de grupo de Active Directory no son compatibles con los orígenes de datos del conector de Amazon Q Business.

7. Selecciona **Añadir** para añadir esta fuente de datos al conector de Amazon Q Business.

Resultado

El origen de datos está incrustado en el índice de Amazon Q Business. El estado cambia de “Embedding” a “Embedded” cuando el origen de datos está completamente embebido.

Después de agregar una única fuente de datos al conector, puede probarlo en el entorno de chatbot de Amazon Q Business y realizar los cambios necesarios antes de poner el servicio a disposición de los usuarios. También puede seguir los mismos pasos para agregar orígenes de datos adicionales al conector.

Administración y supervisión

Gestione la infraestructura de GenAI

Puede ver los detalles sobre su infraestructura RAG de GenAI implementada o eliminar la infraestructura de chatbot si ya no la necesita.

Permite ver información sobre la infraestructura

Puede ver información sobre la infraestructura de chatbot.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Seleccione el menú **Infraestructura**.
4. Vea información sobre la infraestructura, que incluye detalles sobre los siguientes componentes:
   - Configuración de AWS
   - Configuración de infraestructura
   - El motor de la IA
   - La base de datos de vectores

Elimine la infraestructura

Si ya no necesita la infraestructura de chatbot que implementó para uno o más chatbots, puede eliminarla de Workload Factory.

Todos los chatbots que se han implementado en esta infraestructura se deshabilitarán y se eliminará todo el historial de chat.

Esta operación solo elimina los vínculos a la infraestructura de IA de Workload Factory; no elimina todos los componentes de AWS. Necesitará eliminar manualmente los siguientes componentes de infraestructura de AWS:

- La instancia de VM
- Puntos finales privados
- El volumen del sistema de archivos FSx para ONTAP que contiene las bases de datos de IA
- El rol de IAM
- La política
- El grupo de seguridad

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Seleccione el menú **Infraestructura**.

4. Seleccione y seleccione **Eliminar infraestructura de chatbot**.
5. Confirme que desea eliminar la infraestructura y seleccione **Eliminar**.

Resultado

Los componentes de infraestructura del chatbot se eliminan de Workload Factory.

Gestione las bases de conocimientos de GenAI

Después de crear una base de conocimientos, puede ver los detalles de la base de conocimientos, modificar la base de conocimientos, integrar orígenes de datos adicionales o suprimir la base de conocimientos.

Ver información sobre una base de conocimientos

Puede ver información sobre la configuración de una base de conocimientos y el origen de datos que están integrados.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Seleccione la base de conocimientos que desea ver.

   Si se define, los inicios de conversación que se están utilizando se muestran en el panel derecho.

4. Para ver los detalles de la base de conocimientos, seleccione y seleccione **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

   El menú **Acciones** le permite administrar la base de conocimientos si desea realizar cambios.

Editar una base de conocimientos

Puede actualizar una base de conocimientos cambiando algunas configuraciones o puede agregar o eliminar orígenes de datos.

Cada vez que agregue, modifique o elimine orígenes de datos de la base de conocimientos, debe sincronizar el origen de datos para que se vuelva a indexar en la base de conocimientos. La sincronización es incremental, por lo que Amazon Bedrock solo procesa los objetos del volumen FSx para ONTAP que se han agregado, modificado o eliminado desde la última sincronización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea actualizar.
4. Selecciona y selecciona **Administrar base de conocimientos**.

Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

5. Selecciona el menú **Acciones** y selecciona **Editar base de conocimientos**.
6. En la página Editar base de conocimiento, puede cambiar el nombre de la base de conocimiento, la descripción, el modelo de inserción, el modelo de chat, la habilitación de funciones, elegir si los iniciadores de conversación se crean de manera automática o manual y la política de instantáneas utilizada para el volumen que contiene la base de conocimiento.

   Si utiliza el modo Manual para iniciar una conversación, también puede cambiar los temas de inicio de la conversación aquí.

   Cada escaneo de la base de conocimientos, que incluye incrustación, costos. Si las barreras de protección de datos están activadas después de crear una base de conocimientos, la base de conocimientos se vuelve a analizar e incurre en costes. Del mismo modo, si cambia los modelos de chat, GenAI volverá a analizar las fuentes de datos asociadas (lo que supondrá un coste).

7. Seleccione **Guardar** después de haber realizado los cambios.

Proteger una base de conocimientos con instantáneas

Puede proteger los datos de la base de conocimientos realizando y restaurando instantáneas de sus volúmenes de base de conocimientos. Puede restaurar desde una instantánea para revertir a la versión anterior de la base de conocimientos en cualquier momento.

Las copias Snapshot pueden ser más rápidas y eficientes en cuanto al almacenamiento que los backups, y permiten proteger cada base de conocimientos mediante una política de protección diferente. Algunos de los escenarios en los que las instantáneas pueden ser útiles son:

- Pérdida o daños de datos accidentales
- Recuperación de datos incorrectos que se están ingiriendo en la base de conocimientos
- Probar diferentes fuentes de datos o estrategias de fragmentación, y revertir rápidamente una vez finalizada la prueba

Tomar una copia Snapshot de un volumen de base de conocimientos

Puede guardar el estado de una base de conocimientos mediante una instantánea manual del volumen de la base de conocimientos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea proteger.
4. Selecciona y selecciona **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

5. Selecciona el menú **Acciones** y selecciona **Instantánea > Crear nueva instantánea**.

6. Opcionalmente, seleccione **Definir nombre de instantánea** e introduzca un nombre personalizado para la instantánea.

   Definir un nombre personalizado puede ayudarle a determinar mejor el contenido de una instantánea si necesita restaurarla en el futuro.

7. Seleccione **Crear**.

   Se crea una instantánea de la base de conocimientos.

Restaure una copia Snapshot de un volumen de base de conocimientos

Cualquier momento, puede restaurar una snapshot manual o programada de un volumen de base de conocimientos.

> No puede restaurar una snapshot utilizando la interfaz de usuario de cargas de trabajo de IA generativas si la base de datos almacenada en el volumen está dañada o se ha eliminado. Como solución alternativa, puede restaurar la snapshot mediante "CLI de ONTAP" en el clúster de ONTAP donde se aloja el volumen.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea restaurar.
4. Selecciona y selecciona **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

5. Seleccione el menú **Acciones** y seleccione **Instantánea > Restaurar instantánea**.

   Aparece el cuadro de diálogo de selección de instantáneas, donde puede ver una lista de las instantáneas que se han creado para esta base de conocimientos.

6. (Opcional) Deseleccione la opción **Pausa en ejecución y escaneos programados después de restaurar la instantánea** si desea que los escaneos de origen de datos programados y actualmente en ejecución continúen después de restaurar la instantánea.

   Esta opción está activada de forma predeterminada para garantizar que no se realice un análisis mientras la base de conocimientos se encuentre en un estado parcialmente restaurado o que un análisis no actualice una base de conocimientos recién restaurada con datos más antiguos.

7. Seleccione la copia Snapshot que desea restaurar de la lista.
8. Seleccione **Restaurar**.

Clonar una base de conocimientos

Puede crear una nueva base de conocimientos a partir de una instantánea de la base de conocimientos. Esto es útil si la base de conocimientos original está dañada o perdida.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea restaurar.
4. Selecciona y selecciona **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

5. Seleccione el menú **Acciones** y seleccione **Instantánea > Clonar base de conocimientos**.

   Aparece el cuadro de diálogo del clon.

6. Opcionalmente, anule la selección de la opción **Pausa en ejecución y escaneos programados después de clonar la instantánea** si desea que los escaneos de origen de datos programados y actualmente en ejecución continúen después de clonar la instantánea.

   Esta opción está activada de forma predeterminada para garantizar que no se realice un análisis mientras la base de conocimientos se encuentre en un estado parcialmente restaurado o que un análisis no actualice una base de conocimientos recién restaurada con datos más antiguos.

7. Seleccione la copia de Snapshot que desea clonar de la lista.
8. Seleccione **continuar**.
9. Introduzca un nombre para la nueva base de conocimientos.
10. Seleccione una SVM del sistema de archivos y un nombre de volumen para la nueva base de conocimientos.
11. Seleccione **Clonar**.

Agregue orígenes de datos adicionales a una base de conocimientos

Puede incrustar orígenes de datos adicionales en su base de conocimientos para rellenarlos con datos adicionales de la organización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimiento, seleccione la base de conocimientos donde desea agregar el origen de datos.
4. Selecciona y selecciona **Añadir fuente de datos**.
5. Seleccione el tipo de fuente de datos que desea agregar:
   - Agregar el sistema de archivos FSx para ONTAP (usar archivos de un volumen FSx para ONTAP existente)
   - Agregar sistema de archivos (utilizar archivos de un recurso compartido SMB o NFS genérico)

Agregar un sistema de archivos FSx para ONTAP

1. **Seleccione un sistema de archivos**: Seleccione el sistema de archivos FSX for ONTAP donde residen sus archivos de origen de datos y seleccione **Siguiente**.
2. **Selecciona un volumen**: Selecciona el volumen en el que residen tus archivos de origen de datos y selecciona **Siguiente**.

   Al seleccionar los archivos almacenados mediante el protocolo SMB, deberá introducir la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.

3. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.
4. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

       La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:
     - En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

       Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

     Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

     Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

   - En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la

inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

5. En la sección **Permission Aware**, que solo está disponible cuando el origen de datos que seleccionó está en un volumen que utiliza el protocolo SMB, puede habilitar o deshabilitar las respuestas con permiso:
   - **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
   - **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.
6. Seleccione **Agregar** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos NFS genérico

1. **Seleccionar un sistema de archivos**: Ingrese la dirección IP o FQDN para el host del sistema de archivos donde residen los archivos de origen de datos, elija el protocolo NFS para el recurso compartido de red y seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre de la exportación NFS y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar la exportación completa o solo carpetas específicas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:
   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

     > La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

> Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

- En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos SMB genérico

1. **Seleccionar sistema de archivos**:
   a. Ingrese la dirección IP o FQDN del host del sistema de archivos donde residen los archivos de origen de datos.
   b. Seleccione el protocolo SMB para el recurso compartido de red.
   c. Ingrese la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.
   d. Seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre del recurso compartido SMB y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar todo el recurso compartido o solo algunas carpetas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:

- * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
- * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

  La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

◦ **Consciente de permisos**: habilitar o deshabilitar respuestas que tengan en cuenta los permisos:

- **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
- **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.

◦ **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

◦ En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Resultado

La fuente de datos está integrada en su base de conocimientos.

Sincronice sus orígenes de datos con una base de conocimientos

Las fuentes de datos se sincronizan automáticamente con la base de conocimientos asociada una vez al día para que cualquier cambio en la fuente de datos se refleje en el chatbot. Si realiza cambios en cualquiera de sus orígenes de datos y desea sincronizar los datos de inmediato, puede realizar una sincronización a petición.

La sincronización es incremental, por lo que Amazon Bedrock solo procesa los objetos de los orígenes de datos que se han agregado, modificado o eliminado desde la última sincronización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea sincronizar.
4. Selecciona y selecciona **Administrar base de conocimientos**.
5. Seleccione el menú **Acciones** y seleccione **Escanear ahora**.

   Verá un mensaje que indica que se están escaneando las fuentes de datos y un mensaje final cuando se complete el análisis.

Resultado

La base de conocimientos se sincroniza con las fuentes de datos adjuntas y cualquier chatbot activo comenzará a utilizar la información más reciente de sus fuentes de datos.

Pausar o reanudar una sincronización programada

Si desea pausar o reanudar la siguiente sincronización (escaneo) de los orígenes de datos, puede hacerlo en cualquier momento. Es posible que deba pausar la siguiente sincronización programada si va a realizar cambios en un origen de datos y no desea que la sincronización se produzca durante la ventana de cambio.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En el menú Bases de conocimiento y conectores, seleccione la base de conocimiento cuyos análisis desea pausar o reanudar.
4. Selecciona y selecciona **Administrar base de conocimientos**.
5. Seleccione el menú **Acciones** y seleccione **Escanear > Pausar escaneo programado** o **Escanear > Reanudar escaneo programado**.

   Verá un mensaje que indica que el siguiente análisis programado se ha pausado o reanudado.

Evalúe los modelos de chat antes de crear una base de conocimientos

Puede evaluar los modelos de chat básicos disponibles antes de crear una base de conocimientos para poder ver qué modelo funciona mejor para su implementación. Dado que el soporte de modelos varía según la región de AWS, consulte "Esta página de documentación de AWS" para verificar qué modelos puede utilizar en las regiones en las que planea implementar su base de conocimientos.

Esta funcionalidad sólo está disponible cuando no se han creado bases de conocimiento, cuando no existen bases de conocimiento en la página de inventario de bases de conocimiento.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de Bases de Conocimientos, verás la opción de seleccionar el modelo de chat en el lado derecho de la página para el Chatbot.
4. Seleccione el modelo de chat de la lista e introduzca un conjunto de preguntas en el área de mensajes para ver cómo responde el chatbot.
5. Pruebe varios modelos para ver qué modelo es el mejor para su implementación.

Resultado

Utilice ese modelo de chat cuando cree su base de conocimientos.

Anule la publicación de su base de conocimientos

Una vez que hayas publicado tu base de conocimientos para que pueda integrarse con una aplicación de chatbot, puedes anular la publicación si quieres que la aplicación de chatbot no acceda a la base de conocimientos.

La anulación de la publicación de la base de conocimientos impide que las aplicaciones de chat funcionen. El punto final de API único en el que se puede acceder a la base de conocimientos está desactivado.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea anular la publicación.
4. Selecciona y selecciona **Administrar base de conocimientos**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación y la lista de todos los orígenes de datos incrustados.

5. Selecciona el menú **Acciones** y selecciona **Anular publicación**.

Resultado

La base de conocimientos está desactivada y ya no es accesible por una aplicación de chatbot.

Suprimir una base de conocimientos

Si ya no necesita una base de conocimientos, puede eliminarla. Cuando se elimina una base de conocimientos, ésta se elimina de Workload Factory y se elimina el volumen que contiene la base de conocimientos. Cualquier aplicación o chatbot que utilice la base de conocimientos dejará de funcionar. La eliminación de una base de conocimientos no es reversible.

Al suprimir una base de conocimientos, también debe desasociar la base de conocimientos de los agentes a los que está asociada para suprimir por completo todos los recursos asociados a la base de conocimientos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos, seleccione la base de conocimientos que desea suprimir.
4. Selecciona y selecciona **Administrar base de conocimientos**.
5. Selecciona el menú **Acciones** y selecciona **Eliminar base de conocimientos**.
6. En el cuadro de diálogo Eliminar base de conocimientos, confirme que desea eliminarla y seleccione **Eliminar**.

Resultado

La base de conocimientos se elimina de Workload Factory y se elimina su volumen asociado.

Administre conectores de Amazon Q Business

Después de crear un conector para Amazon Q Business, puede ver los detalles del conector, modificar el conector, integrar orígenes de datos adicionales o eliminar el conector.

Ver información sobre un conector

Puede ver información sobre la configuración de un conector y los orígenes de datos integrados.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Desde la página de inventario de Bases de conocimiento y conectores, seleccione el conector que desea ver.
4. Para ver los detalles del conector, seleccione y seleccione **Administrar conector**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

   El menú **Acciones** le permite administrar el conector si desea realizar cambios.

Edite un conector

Puede actualizar un conector cambiando algunos ajustes o puede agregar o eliminar orígenes de datos.

Cada vez que agrega, modifica o elimina fuentes de datos del conector, GenAI necesita enviar la información de origen de datos a Amazon Q Business para que se vuelva a indexar. La sincronización es incremental, por lo que Amazon Q Business solo procesa los objetos de tu volumen de FSx para ONTAP que se han agregado, modificado o eliminado desde la última sincronización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos y conectores, seleccione el conector que desea

actualizar.

4. Selecciona y selecciona **Administrar conector**.

   Esta página muestra el estado Publicado, el estado de incrustación de los orígenes de datos, el modo de incrustación, la lista de todos los orígenes de datos incrustados, etc.

5. Seleccione el menú **Acciones** y seleccione **Editar conector**.
6. En la página Editar conector, puede cambiar el nombre del conector, la descripción, el modelo de incrustación, la habilitación de barandillas de datos y la política de instantáneas utilizada para el volumen que contiene el conector.

   Cada escaneo de la fuente de datos, que incluye la incrustación, incurre en un costo. Si habilita las barandillas de datos después de crear un conector, la fuente de datos se escanea de nuevo y conlleva costes.

7. Seleccione **Guardar** después de haber realizado los cambios.

Agregue orígenes de datos adicionales a un conector

Puede incrustar orígenes de datos adicionales en el conector para rellenarlo con datos adicionales de la organización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de Knowledge Bases & Connectors, seleccione el conector en el que desea agregar el origen de datos.
4. Selecciona y selecciona **Añadir fuente de datos**.
5. Seleccione el tipo de fuente de datos que desea agregar:
   - Agregar el sistema de archivos FSx para ONTAP (usar archivos de un volumen FSx para ONTAP existente)
   - Agregar sistema de archivos (utilizar archivos de un recurso compartido SMB o NFS genérico)

Agregar un sistema de archivos FSx para ONTAP

1. **Seleccione un sistema de archivos**: Seleccione el sistema de archivos FSX for ONTAP donde residen sus archivos de origen de datos y seleccione **Siguiente**.
2. **Selecciona un volumen**: Selecciona el volumen en el que residen tus archivos de origen de datos y selecciona **Siguiente**.

   Al seleccionar los archivos almacenados mediante el protocolo SMB, deberá introducir la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.

3. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.
4. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

       La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:
     - En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

       Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

     Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

     Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

   - En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la

inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

5. En la sección **Permission Aware**, que solo está disponible cuando el origen de datos que seleccionó está en un volumen que utiliza el protocolo SMB, puede habilitar o deshabilitar las respuestas con permiso:
   - **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
   - **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.
6. Seleccione **Agregar** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos NFS genérico

1. **Seleccionar un sistema de archivos**: Ingrese la dirección IP o FQDN para el host del sistema de archivos donde residen los archivos de origen de datos, elija el protocolo NFS para el recurso compartido de red y seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre de la exportación NFS y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar la exportación completa o solo carpetas específicas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:
   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:
     - * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
     - * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

     > La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

   - **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

> Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

- En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Agregar un sistema de archivos SMB genérico

1. **Seleccionar sistema de archivos**:
   a. Ingrese la dirección IP o FQDN del host del sistema de archivos donde residen los archivos de origen de datos.
   b. Seleccione el protocolo SMB para el recurso compartido de red.
   c. Ingrese la información de Active Directory, que incluye el dominio, la dirección IP, el nombre de usuario y la contraseña.
   d. Seleccione **Siguiente**.
2. **Seleccione una fuente de datos**: Seleccione la ubicación de la fuente de datos en función de dónde haya guardado los archivos. Esto puede ser un volumen completo, o simplemente una carpeta o subcarpeta específica en el volumen, y seleccione **Siguiente**.

> En algunos casos, podría ser necesario introducir manualmente el nombre del recurso compartido SMB y seleccionar **Recuperar directorios** para ver los directorios disponibles. Puede seleccionar todo el recurso compartido o solo algunas carpetas.

3. **Configuraciones**: Configure cómo la fuente de datos ingiere información de sus archivos y qué archivos incluye en los escaneos:

   - **Definir fuente de datos**: En la sección **Estrategia de Chunking**, defina cómo el motor GenAI divide el contenido de la fuente de datos en fragmentos cuando la fuente de datos se integra con una base de conocimientos. Puede elegir una de las siguientes estrategias:

- * Fragmentación de varias frases*: Organiza la información de su fuente de datos en fragmentos definidos por frases. Puedes elegir cuántas oraciones componen cada fragmento (hasta 100).
- * Fragmento basado en superposición *: Organiza la información de su fuente de datos en fragmentos definidos por caracteres que pueden superponerse a fragmentos vecinos. Puedes elegir el tamaño de cada fragmento en caracteres, y cuánto se superpone cada fragmento con fragmentos adyacentes. Puede configurar un tamaño de fragmento de entre 50 y 3000 caracteres, y un porcentaje de superposición de entre 1 y 99%.

> La elección de un alto porcentaje de superposición puede aumentar en gran medida los requisitos de almacenamiento con solo pequeñas mejoras en la precisión de la recuperación.

◦ **Consciente de permisos**: habilitar o deshabilitar respuestas que tengan en cuenta los permisos:

- **Activado**: Los usuarios del chatbot que accedan a esta base de conocimientos solo obtendrán respuestas a las consultas de las fuentes de datos a las que tengan acceso.
- **Deshabilitado**: Los usuarios del chatbot recibirán respuestas usando contenido de todas las fuentes de datos integradas.

◦ **Filtrado de archivos**: Configure qué archivos se incluyen en los escaneos:

- En la sección **File Types support**, elija incluir todos los tipos de archivos o seleccionar tipos de archivos individuales para incluirlos en los escaneos de fuentes de datos.

  Si incluye imágenes o archivos PDF, NetApp Workload Factory for GenAI analiza el texto de las imágenes (incluidas las imágenes en documentos PDF), y esto implica un costo mayor.

Al incluir datos de texto de imágenes, GenAI no puede enmascarar la información de identificación personal (PII) de la imagen a medida que los datos de texto escaneados se envían desde su entorno a AWS. Sin embargo, una vez almacenados los datos, todos los PII se enmascaran en la base de datos de GenAI.

> Su elección para incluir archivos de imagen en los escaneos está relacionada con el modelo de chat de la base de conocimientos. Si incluye archivos de imagen en los escaneos, el modelo de chat debe admitir imágenes. Si se seleccionan aquí tipos de archivo de imagen, no puede cambiar la base de conocimientos a un modelo de chat que no admita archivos de imagen.

◦ En la sección **Filtro de tiempo de modificación de archivo**, elija habilitar o deshabilitar la inclusión de archivos en función de su tiempo de modificación. Si activa el filtrado de tiempo de modificación, seleccione un intervalo de fechas de la lista.

> Si incluye archivos basados en un rango de fechas de modificación, en cuanto el rango de fechas no se cumpla (los archivos no se han modificado dentro del rango de fechas especificado), los archivos se excluirán de la exploración periódica y el origen de datos no incluirá estos archivos.

4. Seleccione **Agregar fuente de datos** para agregar esta fuente de datos a su base de conocimientos.

Resultado

El origen de datos está integrado en el conector.

Sincronice sus orígenes de datos con un conector

Las fuentes de datos se sincronizan automáticamente con el conector asociado una vez al día para que cualquier cambio en la fuente de datos se refleje en Amazon Q Business. Si realiza cambios en cualquiera de sus orígenes de datos y desea sincronizar (escanear) los datos inmediatamente, puede realizar una sincronización a petición.

La sincronización es incremental, por lo que Amazon Q Business solo procesa los objetos de los orígenes de datos que se han agregado, modificado o eliminado desde la última sincronización.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Desde el menú Bases de conocimiento y conectores, seleccione el conector que desea sincronizar.
4. Selecciona y selecciona **Administrar conector**.
5. Seleccione el menú **Acciones** y seleccione **Escanear ahora**.

   Verá un mensaje que indica que se están escaneando las fuentes de datos y un mensaje final cuando se complete el análisis.

Resultado

El conector se sincroniza con las fuentes de datos adjuntas y Amazon Q Business comenzará a utilizar la información más reciente de sus fuentes de datos.

Pausar o reanudar una sincronización programada

Si desea pausar o reanudar la siguiente sincronización (escaneo) de los orígenes de datos, puede hacerlo en cualquier momento. Es posible que deba pausar la siguiente sincronización programada si va a realizar cambios en un origen de datos y no desea que la sincronización se produzca durante la ventana de cambio.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de conectores, seleccione el conector para el que desea pausar o reanudar las exploraciones.
4. Selecciona y selecciona **Administrar conector**.
5. Seleccione el menú **Acciones** y seleccione **Escanear > Pausar escaneo programado** o **Escanear > Reanudar escaneo programado**.

   Verá un mensaje que indica que el siguiente análisis programado se ha pausado o reanudado.

Eliminar un conector

Si ya no necesita un conector, puede eliminarlo. Cuando se elimina un conector, se elimina de Workload Factory y se elimina el volumen que contiene el conector. La eliminación de un conector no es reversible.

Al suprimir un conector, también debe desasociar el conector de cualquier agente con el que esté asociado para suprimir completamente todos los recursos asociados al conector.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimientos y conectores, seleccione el conector que desea suprimir.
4. Selecciona y selecciona **Administrar conector**.
5. Seleccione el menú **Acciones** y seleccione **Borrar conector**.
6. En el cuadro de diálogo Eliminar conector, confirme que desea eliminarlo y seleccione **Eliminar**.

Resultado

Se elimina el conector de Workload Factory y se elimina su volumen asociado.

Gestionar orígenes de datos de GenAI

Después de crear una base de conocimientos o un conector utilizando orígenes de datos en el sistema de archivos FSx para ONTAP, puede ver los detalles de la fuente de datos, actualizar o cambiar el contenido de la fuente de datos, editar la configuración de la fuente de datos o eliminar la fuente de datos.

Ver información sobre un origen de datos

Puede ver información sobre el contenido de un origen de datos y puede ver su estado de incrustación con la base de conocimientos o el conector. Dado que las fuentes de datos están asociadas a una base de conocimientos o conector, primero tendrá que elegir la base de conocimientos o el conector para poder ver los detalles de la fuente de datos.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los "experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. Seleccione la base de conocimientos o el conector donde reside la fuente de datos y, a continuación, seleccione * Administrar base de conocimientos * o * Administrar conector *.

   La parte inferior de la página muestra los orígenes de datos asociados.

4. Expanda cada fila seleccionando la para ver información detallada sobre cada origen de datos, como el sistema de archivos FSx para ONTAP, el volumen y la ruta donde reside el origen de datos.

   También muestra la información de incrustación y si ese origen de datos está actualmente incrustado en la base de conocimientos o en el conector.

Editar la configuración del origen de datos

Puede editar información sobre un origen de datos que haya integrado con una base de conocimientos o conector. La mayor parte de la información se corrige después de agregar un origen de datos, pero puede realizar cambios en alguna de la configuración (como la definición de fragmentación o el reconocimiento de permisos).

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de bases de conocimiento, seleccione la base de conocimientos donde reside el origen de datos y, a continuación, ••• seleccione y seleccione **Gestionar base de conocimientos**.

   La parte inferior de la página muestra los orígenes de datos que forman parte de esta base de conocimientos.
4. En la fila de la fuente de datos que desea editar, ••• seleccione y seleccione **Editar fuente de datos**.
5. En la página Editar origen de datos, seleccione ⌄ para ampliar la fila para la definición de fragmentos.
6. Actualice la configuración y la estrategia de fragmentación, así como el reconocimiento de permisos (para volúmenes SMB) y seleccione **Guardar**.

Resultado

La configuración del origen de datos se actualiza y el sistema AI sincroniza el origen de datos para que se vuelva a indexar en la base de conocimientos.

Actualizar el contenido de un origen de datos existente

Puede cambiar el contenido de un origen de datos en cualquier momento para agregar o actualizar los datos de su organización. Si este origen de datos se está utilizando activamente en una base de conocimientos, debe sincronizar el origen de datos para que se vuelva a indexar en la base de conocimientos. La sincronización es incremental, por lo que Amazon Bedrock solo procesa los objetos del volumen FSx para ONTAP que se han agregado, modificado o eliminado desde la última sincronización.

Las fuentes de datos se sincronizan automáticamente con la base de conocimientos una vez al día para que cualquier cambio en la fuente de datos se refleje en el chatbot. Si realiza cambios en un origen de datos y desea sincronizar los datos inmediatamente, puede "realice una sincronización bajo demanda".

Suprimir un origen de datos

Si ya no necesita un origen de datos para formar parte de su base de conocimientos, puede eliminarlo.

Pasos

1. Inicie sesión en Workload Factory utilizando uno de los"experiencias de consola" .
2. En el icono Cargas de trabajo de IA, seleccione **Implementar y gestionar**.
3. En la página de inventario de la base de conocimientos, seleccione la base de conocimientos donde reside el origen de datos y, a continuación, ••• seleccione y seleccione **Gestionar base de conocimientos**.

   La parte inferior de la página muestra los orígenes de datos que forman parte de esta base de conocimientos.
4. En la fila de la fuente de datos que desea eliminar, ••• seleccione y seleccione **Eliminar fuente de datos**.
5. En el cuadro de diálogo Eliminar origen de datos, confirme que desea eliminarlo y seleccione **Confirmar**.

Resultado

El origen de datos se elimina de la base de conocimientos y el sistema AI elimina la información indexada sobre este origen de datos de la base de conocimientos. Cualquier información de esa fuente de datos ya no estará disponible para los chatbots que estén utilizando la base de conocimientos.

Supervise las operaciones de carga de trabajo con Tracker en NetApp Workload Factory

Supervise y realice un seguimiento de la ejecución de las operaciones de carga de trabajo y supervise el progreso de las tareas con Tracker en NetApp Workload Factory.

Acerca de esta tarea

NetApp Workload Factory ofrece Tracker, una función de monitoreo, para que pueda monitorear y rastrear el progreso y el estado de las operaciones de carga de trabajo, revisar detalles de tareas y subtareas de operación y diagnosticar cualquier problema o falla.

Hay varias acciones disponibles en Tracker. Puede filtrar trabajos por intervalo de tiempo (últimas 24 horas, 7 días, 14 días o 30 días), carga de trabajo, estado y usuario; buscar trabajos con la función de búsqueda; y descargar la tabla de trabajos como archivo CSV. Puede actualizar el Tracker en cualquier momento y volver a intentar rápidamente una operación fallida o editar parámetros para una operación fallida y volver a intentar la operación.

Tracker admite dos niveles de supervisión dependiendo de la operación. Cada tarea, como el despliegue del sistema de archivos, muestra la descripción de la tarea, el estado, la hora de inicio, la duración de la tarea, el usuario, la región, el recurso proxy, el identificador de tarea y todas las subtareas relacionadas. Puede ver las respuestas de la API para entender lo que ha pasado durante la operación.

Niveles de tareas de Tracker con ejemplos

- Nivel 1 (tarea): Realiza un seguimiento de la implementación del sistema de archivos.
- Nivel 2 (subtarea): Realiza un seguimiento de las subtareas relacionadas con el despliegue del sistema de archivos.

Estado de la operación

El estado de la operación en Tracker es el siguiente *in progress*, *success* y *failed*.

Frecuencia de operación

La frecuencia de operación se basa en el tipo de trabajo y la programación de trabajo.

Retención de eventos

Los eventos se conservan en la interfaz de usuario durante 30 días.

Seguimiento y supervisión de operaciones

Realice un seguimiento y monitoree las operaciones en la consola de Workload Factory con Tracker.

Pasos

1. Inicie sesión con uno de los "experiencias de consola"botones .
2. Desde el menú de carga de trabajo, seleccione **Administración** y luego seleccione **Rastreador**.
3. En Tracker, utilice los filtros o la búsqueda para limitar los resultados de trabajo. También puedes descargar un informe de trabajos.

Ver solicitud de API

Vea la solicitud API en CodeBox para una tarea en Tracker.

Pasos

1. En Rastreador, seleccione una tarea.
2. Seleccione el menú de acciones y luego seleccione **Ver solicitud de API**.

Vuelva a intentar una operación fallida

Vuelva a intentar una operación fallida en Tracker. También puede copiar el mensaje de error de una operación con un error.

Puede volver a intentar una operación con fallos hasta 10 veces.

Pasos

1. En Tracker, seleccione una operación fallida.
2. Seleccione el menú de acciones y luego seleccione **Reintentar**.

Resultado

Se vuelve a iniciar la operación.

Edite y vuelva a intentar una operación con errores

Edite los parámetros de la operación fallida y vuelva a intentar la operación fuera de Tracker.

Pasos

1. En Tracker, seleccione una operación fallida.
2. Seleccione el menú de acciones y luego seleccione **Editar y reintentar**.

   Se le redirigirá a la página de operaciones, donde puede editar los parámetros y volver a intentar la operación.

Resultado

Se vuelve a iniciar la operación. Vaya a Rastreador para ver el estado de la operación.

Conocimiento y apoyo

Regístrese para obtener soporte para NetApp Workload Factory para GenAI

Es necesario registrarse para recibir soporte técnico específico para NetApp Workload Factory y sus soluciones y servicios de almacenamiento. Debe registrarse para recibir soporte desde la consola de NetApp , que es una consola web independiente de Workload Factory.

Registrarse para recibir soporte no habilita el soporte de NetApp para un servicio de archivos de un proveedor de nube. Para obtener asistencia técnica relacionada con un servicio de archivos de un proveedor de nube, su infraestructura o cualquier solución que utilice el servicio, consulte "Obtener ayuda" en la documentación de Workload Factory para ese producto.

"Amazon FSX para ONTAP"

Información general del registro de soporte

El registro de su ID de cuenta de suscripción de soporte (su número de serie 960xxxxxxxx de 20 dígitos ubicado en la página Recursos de soporte en la consola de NetApp) sirve como su ID de suscripción de soporte única. Cada suscripción de soporte a nivel de cuenta de NetApp debe estar registrada.

El registro habilita capacidades como abrir tickets de soporte y generar casos automáticamente. El registro se completa agregando cuentas del sitio de soporte de NetApp (NSS) a la consola de NetApp como se describe a continuación.

Registre su cuenta de soporte de NetApp

Para registrarse para recibir soporte y activar el derecho a soporte, un usuario de su cuenta debe asociar una cuenta del sitio de soporte de NetApp con su inicio de sesión de la consola de NetApp . La forma de registrarse para el soporte de NetApp depende de si ya tiene una cuenta del sitio de soporte de NetApp (NSS).

Cliente existente con una cuenta de NSS

Si es cliente de NetApp con una cuenta NSS, simplemente necesita registrarse para recibir soporte a través de la consola de NetApp .

Pasos

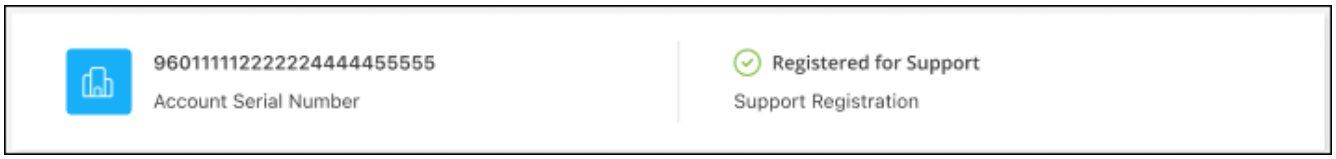
1. En la parte superior derecha de la consola de Workload Factory, seleccione **Ayuda > Soporte**.

   Al seleccionar esta opción, se abre la consola de NetApp en una nueva pestaña del navegador y se carga el panel de soporte.

2. Desde el menú de la consola de NetApp , seleccione **Administración** y, a continuación, seleccione **Credenciales**.
3. Seleccione **Credenciales de usuario**.
4. Seleccione **Agregar credenciales NSS** y siga el aviso de autenticación del sitio de soporte de NetApp (NSS).

5. Para confirmar que el proceso de registro se ha realizado correctamente, seleccione el icono Ayuda y seleccione **Soporte**.

   La página **Recursos** debe mostrar que su cuenta está registrada para soporte.



   Tenga en cuenta que otros usuarios de la consola de NetApp no verán este mismo estado de registro de soporte si no han asociado una cuenta del sitio de soporte de NetApp con su inicio de sesión de la consola de NetApp . Sin embargo, eso no significa que su cuenta de NetApp no esté registrada para recibir soporte. Siempre que un usuario de la cuenta haya seguido estos pasos, su cuenta quedará registrada.

Cliente existente pero no cuenta NSS

Si es un cliente existente de NetApp con licencias y números de serie existentes pero *no* una cuenta NSS, debe crear una cuenta NSS y asociarla con su inicio de sesión de la consola de NetApp .

Pasos

1. Cree una cuenta del sitio de soporte de NetApp completando el "Formulario de registro de usuarios del sitio de soporte de NetApp"
   a. Asegúrese de seleccionar el nivel de usuario adecuado, que normalmente es **Cliente/Usuario final de NetApp**.
   b. Asegúrese de copiar el número de serie de la cuenta de NetApp (960xxxx) utilizado anteriormente para el campo de número de serie. Esto acelerará el procesamiento de la cuenta.
2. Asocie su nueva cuenta NSS con su inicio de sesión de la consola NetApp completando los pasos a continuaciónCliente existente con una cuenta de NSS .

Totalmente nuevo en NetApp

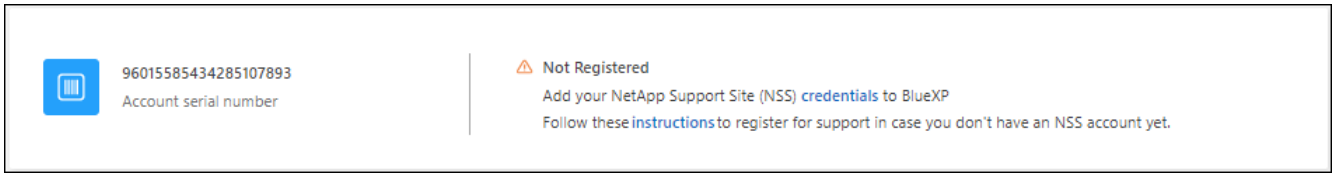
Si es totalmente nuevo en NetApp y no tiene una cuenta de NSS, siga cada paso que se indica a continuación.

Pasos

1. En la parte superior derecha de la consola de Workload Factory, seleccione **Ayuda > Soporte**.

   Al seleccionar esta opción, se abre la consola de NetApp en una nueva pestaña del navegador y se carga el panel de soporte.

2. Busque el número de serie de su ID de cuenta en la página Recursos de soporte.



3. Navegue "Sitio de registro de soporte de NetApp" y seleccione **No soy un Cliente de NetApp Registrado**.

4. Rellene los campos obligatorios (aquellos con asteriscos rojos).
5. En el campo **línea de productos**, seleccione **Cloud Manager** y, a continuación, seleccione el proveedor de facturación correspondiente.
6. Copie el número de serie de su cuenta desde el paso 2 anterior, complete la comprobación de seguridad y confirme que ha leído la Política de privacidad de datos global de NetApp.

   Se envía inmediatamente un correo electrónico al buzón de correo para finalizar esta transacción segura. Asegúrese de comprobar sus carpetas de spam si el correo electrónico de validación no llega en pocos minutos.

7. Confirme la acción desde el correo electrónico.

   Confirmar envía su solicitud a NetApp y recomienda que cree una cuenta en la página de soporte de NetApp.

8. Cree una cuenta del sitio de soporte de NetApp completando el "Formulario de registro de usuarios del sitio de soporte de NetApp"
   a. Asegúrese de seleccionar el nivel de usuario adecuado, que normalmente es **Cliente/Usuario final de NetApp**.
   b. Asegúrese de copiar el número de serie de la cuenta (960xxxx) utilizado anteriormente para el campo de número de serie. Esto agilizará el procesamiento de la cuenta.

Después de terminar

NetApp debería ponerse en contacto con usted durante este proceso. Este es un ejercicio de incorporación puntual para nuevos usuarios.

Una vez que tenga su cuenta del sitio de soporte de NetApp , asocie la cuenta con su inicio de sesión de la consola de NetApp completando los pasos a continuación.Cliente existente con una cuenta de NSS .

Solución de problemas de GenAI

Aprenda a solucionar algunos problemas comunes que puede encontrar.

Problemas y soluciones comunes

Si tiene uno de estos problemas, puede seguir los pasos de la columna Solución alternativa para intentar resolverlo.

Zona	Problema	Causa	Solución alternativa
Puesta en marcha	La implementación falla porque el volumen ya existe.	NetApp Workload Factory for GenAI necesita crear un nuevo volumen durante el proceso de implementación, pero ya existe un volumen que usa el nombre que usted especificó.	Especifique un nombre único que se utilizará para el nuevo volumen e intente implementar de nuevo.

Zona	Problema	Causa	Solución alternativa
Puesta en marcha	La implementación falla porque NetApp Workload Factory for GenAI no puede montar el volumen.	Uno o más de los puertos de entrada necesarios para FSx para NetApp ONTAP están cerrados o filtrados.	Abra los siguientes puertos de entrada:

| Protocolo | Puerto | Específico

| Todos los ICMP | Todo | Hacer ping a la instancia

| HTTPS | 443 | Acceso desde la LIF de gestión de Connector to fsxadmin para enviar llamadas API a FSX

| SSH | 22 | Acceso SSH a la dirección IP de administración del clúster LIF o una LIF de gestión de nodos

| TCP | 111 | Llamada a procedimiento remoto para NFS

| TCP | 139 | Sesión de servicio NetBIOS para CIFS

| TCP | 161-162 | Protocolo simple de gestión de red

| TCP | 445 | Microsoft SMB/CIFS sobre TCP con trama NetBIOS

| TCP | 635 | Montaje NFS

| TCP | 749 | Kerberos

| TCP | 2049 | Daemon del servidor NFS

| TCP | 3260 | Acceso iSCSI mediante la LIF de datos iSCSI

| TCP | 4045 | Daemon de bloqueo NFS

| TCP | 4046 | Supervisor de estado de red para NFS

| TCP | 10000 | Backup con NDMP

| TCP | 11104 | Gestión de sesiones de comunicación de interconexión de clústeres para SnapMirror

| TCP | 11105 | Transferencia de datos de SnapMirror mediante LIF de interconexión de clústeres

| UDP | 111 | Llamada a procedimiento remoto para NFS

| UDP | 161-162 | Protocolo simple de gestión de red

| UDP | 635 | Montaje NFS

| UDP | 2049 | Daemon del servidor NFS

| UDP | 4045 | Daemon de bloqueo NFS

| UDP | 4046 | Supervisor de estado de red para NFS

| UDP | 4049 | Protocolo rquotad NFS

Mantenimiento	El motor AI no se inicia y aparece el error “Error de instancia del motor AI” en la página **Bases de conocimiento**.	La instancia del motor AI está dañada o no existe.	Seleccione el botón **Reconstruir**. NetApp Workload Factory para GenAI reconstruye la infraestructura y muestra el progreso de la reconstrucción. Una vez finalizado, sus bases de conocimiento se volverán a conectar a la infraestructura reconstruida y se mostrará la lista de bases de conocimiento.
Mantenimiento	El motor AI no se inicia y aparece el error que indica que la instancia del motor GenAI se ha detenido en la página de las bases de conocimiento *.	La instancia del motor AI no se está ejecutando.	Utilice la consola de administración de AWS o la CLI de AWS para iniciar la instancia del motor de AI.
Mantenimiento	El motor AI no se inicia y aparece el error que indica que el servidor del motor GenAI no responde en la página de las bases de conocimiento *.	La instancia del motor AI no responde.	Utilice los siguientes pasos de recuperación: **Pasos** 1. Modifique el grupo de seguridad de la instancia del motor de GenAI para habilitar el acceso SSH a la instancia del motor de GenAI. 2. Inicie sesión en la instancia mediante SSH. 3. Ejecute el siguiente comando: `docker-compose up`

Mantenimiento	La instancia de Docker de backend utilizada por NetApp Workload Factory para GenAI no pudo iniciarse.	Se ha suprimido el volumen y se ha reiniciado la instancia de EC2.	Utilice los siguientes pasos de recuperación:
			
			Pasos
			1. Crea un nuevo volumen en FSx para NetApp ONTAP. Por ejemplo, el nombre del volumen puede ser `netapp_ai` y la ruta de volumen puede ser `/netapp_ai`.
			2. SSH a la instancia de Amazon EC2.
			3. Enumere los volúmenes:
			`docker volume list`
			4. Elimine el volumen antiguo:
			`docker volume rm ec2-user_persistent_folder`
			5. Abra el `docker-compose.yml` archivo con un editor de texto.
			6. En `volumes` la sección, cambie la ruta del dispositivo a la nueva ruta del volumen. Por ejemplo:

Mantenimiento	La instancia de Docker de backend utilizada por NetApp Workload Factory para GenAI no pudo iniciarse.	Se eliminó el volumen raíz.	Cree un volumen con un nombre y una ruta y, a continuación, reinicie la instancia de Docker back-end desde Amazon EC2.
Mantenimiento	La instancia de Docker de backend utilizada por NetApp Workload Factory para GenAI no pudo iniciarse.	Se eliminó el volumen raíz.	Cree un volumen con un nombre y una ruta y, a continuación, reinicie la instancia de Docker back-end desde Amazon EC2.

```
"addr=svm-
d8a72.\
fs-
12bc3.\
fsx.us-east-
1.amazonaws.
com,nolock,s
oft,rw"

device:
':/netapp_ai
' # Path to
new volume
```

Obtenga ayuda con NetApp Workload Factory para GenAI

NetApp brinda soporte para Workload Factory y sus servicios en la nube de diversas maneras. Hay amplias opciones de autoayuda gratuitas disponibles las 24 horas, los 7 días de la semana, como artículos de la base de conocimientos (KB) y un foro comunitario. Su registro de soporte incluye soporte técnico remoto mediante tickets web.

Obtenga soporte para FSx for ONTAP

Para obtener soporte técnico relacionado con FSx para ONTAP, su infraestructura o cualquier solución que utilice el servicio, consulte "Obtener ayuda" en la documentación de Workload Factory para ese producto.

"Amazon FSX para ONTAP"

Para recibir soporte técnico específico para Workload Factory y sus soluciones y servicios de almacenamiento, utilice las opciones de soporte descritas a continuación.

Utilice opciones de soporte automático

Estas opciones están disponibles de forma gratuita, las 24 horas del día, los 7 días de la semana:

- Documentación

  La documentación de Workload Factory que estás viendo actualmente.

- "Base de conocimientos"

  Busque en la base de conocimientos de Workload Factory para encontrar artículos útiles para solucionar problemas.

- "Comunidades"

  Únase a la comunidad de Workload Factory para seguir discusiones en curso o crear otras nuevas.

Cree un caso con el soporte de NetApp

Además de las opciones de autosoporte anteriores, puede trabajar con un especialista de soporte de NetApp para resolver cualquier problema después de activar el soporte de.

Antes de empezar

Para utilizar la función **Crear un caso**, primero debe registrarse para recibir soporte. Asocie sus credenciales del sitio de soporte de NetApp con su inicio de sesión de Workload Factory. "Aprenda a registrarse para obtener soporte" .

Pasos

1. En la parte superior derecha de la consola de Workload Factory, seleccione **Ayuda > Soporte**.

   Al seleccionar esta opción, se abre la consola de NetApp en una nueva pestaña del navegador y se carga el panel de soporte.

2. En la página **Recursos**, elija una de las opciones disponibles en Soporte técnico:

   a. Selecciona **Llámanos** si quieres hablar con alguien por teléfono. Se le dirigirá a una página de netapp.com que enumera los números de teléfono a los que puede llamar.

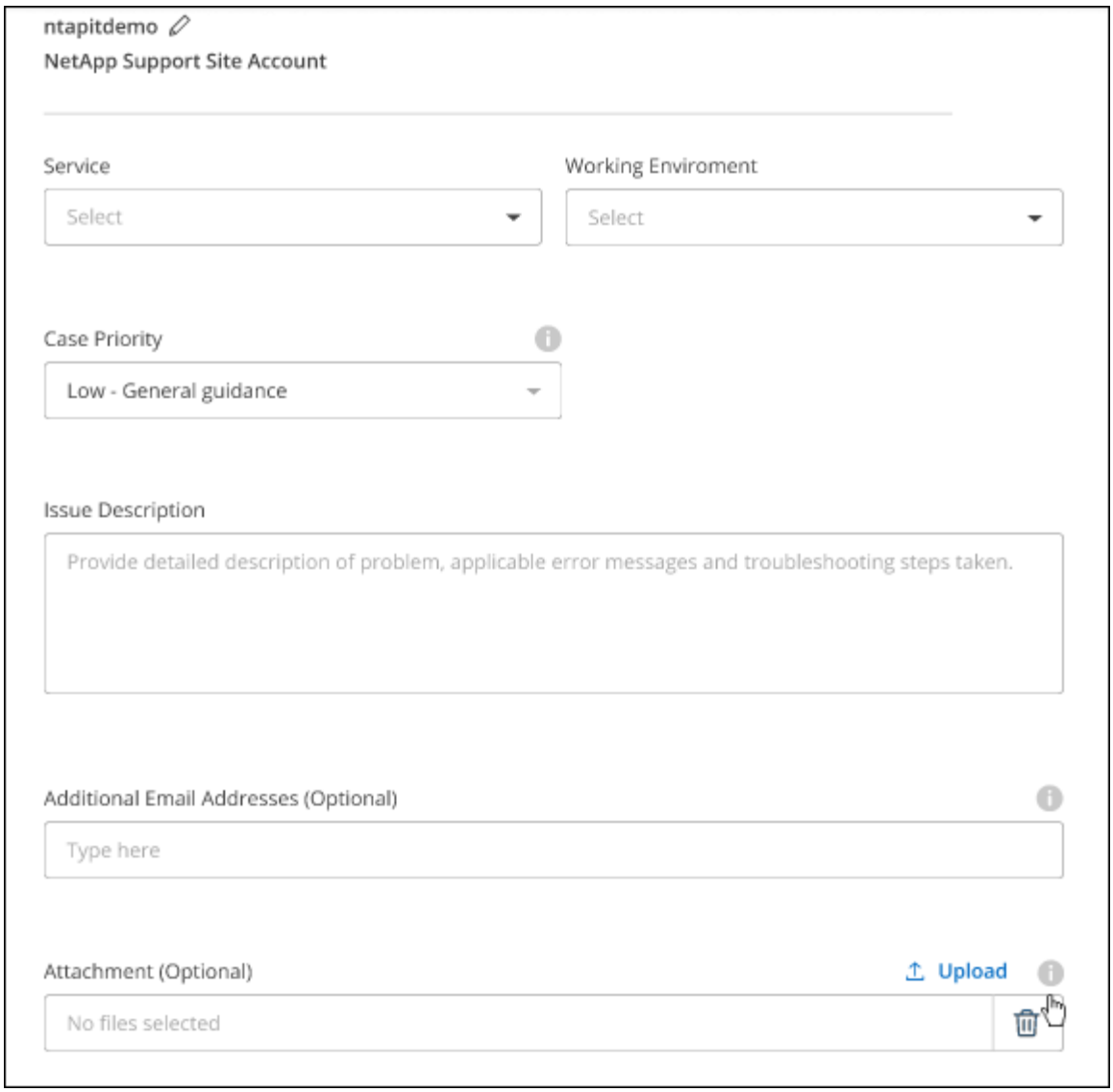
   b. Selecciona **Crear un caso** para abrir un ticket con un especialista en Soporte NetApp:

      - **Servicio**: Selecciona **Fábrica de carga de trabajo**.
      - **Prioridad de caso**: Elija la prioridad para el caso, que puede ser Baja, Media, Alta o crítica.

        Para obtener más información sobre estas prioridades, pase el ratón sobre el icono de información situado junto al nombre del campo.

      - **Descripción del problema**: Proporcione una descripción detallada del problema, incluidos los mensajes de error aplicables o los pasos de solución de problemas que haya realizado.
      - **Direcciones de correo electrónico adicionales**: Introduzca direcciones de correo electrónico adicionales si desea que alguien más conozca este problema.
      - **Accesorio (opcional)**: Cargue hasta cinco archivos adjuntos, uno a la vez.

        Los archivos adjuntos están limitados a 25 MB por archivo. Se admiten las siguientes extensiones de archivo: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx y csv.



Después de terminar

Aparecerá una ventana emergente con el número de caso de soporte. Un especialista de soporte de NetApp revisará su caso y le pondrá en contacto con usted próximamente.

Para obtener un historial de sus casos de soporte, puede seleccionar **Ajustes > Línea de tiempo** y buscar acciones denominadas “Crear caso de soporte”. Un botón situado en el extremo derecho le permite ampliar la acción para ver los detalles.

Es posible que se encuentre el siguiente mensaje de error al intentar crear un caso:

"No está autorizado a crear un caso en el servicio seleccionado"

Este error podría significar que la cuenta NSS y la empresa registrada con la que está asociada no son la misma empresa registrada para el número de serie de la cuenta de la consola de NetApp (es decir, 960xxxx) o el número de serie del sistema. Puede buscar ayuda utilizando una de las siguientes opciones:

- Usar el chat en el producto
- Envíe un caso no técnico a https://mysupport.netapp.com/site/help

Gestione sus casos de soporte (vista previa)

Puede ver y administrar casos de soporte activos y resueltos directamente desde la consola de NetApp . Podrás gestionar los casos asociados a tu cuenta NSS y a tu empresa.

La gestión de casos está disponible como vista previa. Tenemos pensado perfeccionar esta experiencia y añadir mejoras en próximos lanzamientos. Envíenos sus comentarios mediante el chat en el producto.

Tenga en cuenta lo siguiente:

- La consola de gestión de casos en la parte superior de la página ofrece dos vistas:
  - La vista de la izquierda muestra el total de casos abiertos en los últimos 3 meses por la cuenta de usuario NSS que ha proporcionado.
  - La vista de la derecha muestra el total de casos abiertos en los últimos 3 meses a nivel de su compañía en función de su cuenta NSS de usuario.

  Los resultados de la tabla reflejan los casos relacionados con la vista seleccionada.
- Puede agregar o quitar columnas de interés y filtrar el contenido de columnas como prioridad y estado. Otras columnas proporcionan funciones de clasificación.

  Consulte los pasos a continuación para obtener más información.
- En el nivel por caso, ofrecemos la posibilidad de actualizar las notas de un caso o cerrar un caso que no esté ya en estado cerrado o pendiente de cierre.

Pasos

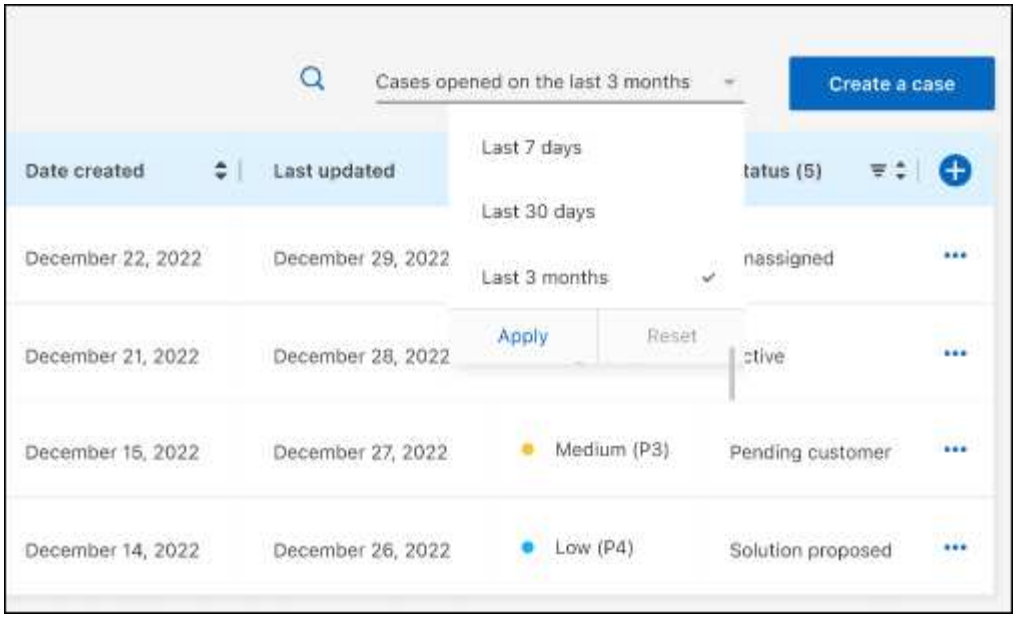
1. En la parte superior derecha de la consola de Workload Factory, seleccione **Ayuda > Soporte**.

   Al seleccionar esta opción, se abre la Consola de NetApp en una nueva pestaña del navegador y se carga el panel de soporte.

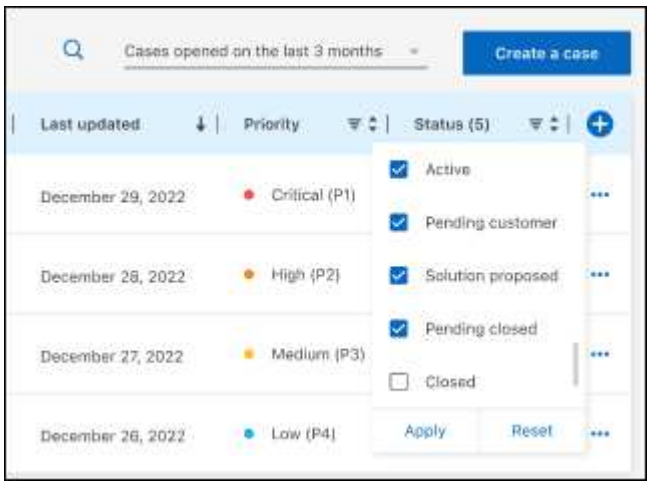
2. Seleccione **Administración de casos** y, si se le solicita, agregue su cuenta NSS a la consola de NetApp .

   La página **Administración de casos** muestra casos abiertos relacionados con la cuenta NSS que está asociada con su cuenta de usuario de la consola de NetApp . Esta es la misma cuenta NSS que aparece en la parte superior de la página de **administración de NSS**.

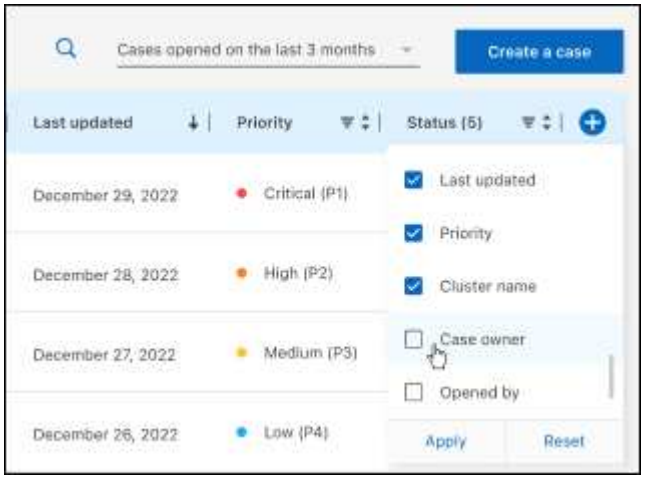
3. Si lo desea, puede modificar la información que se muestra en la tabla:
   - En **Casos de la organización**, selecciona **Ver** para ver todos los casos asociados a tu empresa.
   - Modifique el rango de fechas eligiendo un rango de fechas exacto o eligiendo un marco de tiempo diferente.



- Filtre el contenido de las columnas.



- Para cambiar las columnas que aparecen en la tabla, seleccione [El icono más que se muestra en la tabla] y, a continuación, elija las columnas que desea mostrar.



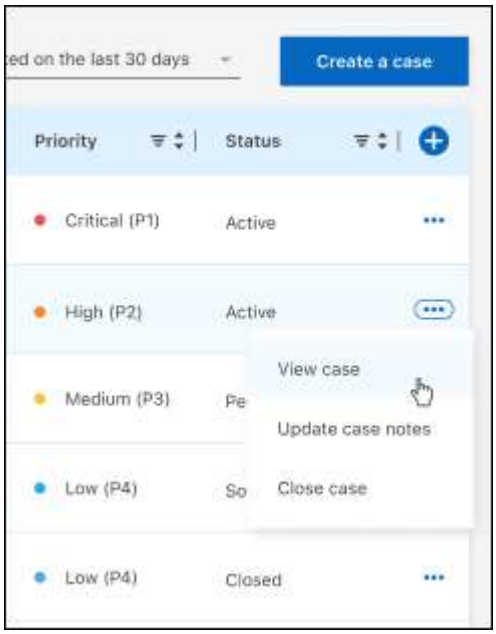
4. Para gestionar un caso existente, [Icono con tres puntos que aparece en la última columna de la tabla]

seleccione y seleccione una de las opciones disponibles:

- **Ver caso**: Ver todos los detalles sobre un caso específico.
- **Actualizar notas de caso**: Proporcione detalles adicionales sobre su problema o seleccione **cargar archivos** para adjuntar hasta un máximo de cinco archivos.

  Los archivos adjuntos están limitados a 25 MB por archivo. Se admiten las siguientes extensiones de archivo: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx y csv.

- **Cerrar caso**: Proporciona detalles sobre por qué estás cerrando el caso y selecciona **Cerrar caso**.



Avisos legales de NetApp Workload Factory para GenAI

Los avisos legales proporcionan acceso a las declaraciones de copyright, marcas comerciales, patentes y mucho más.

Copyright

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

Marcas comerciales

NETAPP, el logotipo de NETAPP y las marcas enumeradas en la página de marcas comerciales de NetApp son marcas comerciales de NetApp, Inc. Los demás nombres de empresas y productos son marcas comerciales de sus respectivos propietarios.

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

Estadounidenses

Puede encontrar una lista actual de las patentes propiedad de NetApp en:

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

Política de privacidad

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

Código abierto

Los archivos de notificación proporcionan información sobre los derechos de autor y las licencias de terceros que se utilizan en software de NetApp.

"Fábrica de cargas de trabajo de NetApp"