Dokumentation zu NetApp Workload Factory für GenAI

GenAI

NetApp
October 06, 2025

Inhalt

Dokumentation zu NetApp Workload Factory für GenAI

Versionshinweise

Was ist neu bei NetApp Workload Factory für GenAI

Informieren Sie sich über die Neuerungen bei der Funktion „Generative AI Workloads“ von Workload Factory.

05. Oktober 2025

BlueXP workload factory jetzt NetApp Workload Factory

BlueXP wurde umbenannt und neu gestaltet, um seine Rolle bei der Verwaltung Ihrer Dateninfrastruktur besser widerzuspiegeln. Aus diesem Grund wurde die BlueXP workload factory in NetApp Workload Factory umbenannt.

Unterstützung für das Hinzufügen generischer NFS/SMB-Datenquellen in NetApp -Konnektoren für Amazon Q Business

Mithilfe der Workload Factory API können Sie jetzt eine Datenquelle aus einer generischen NFSv3-, NFSv4- oder SMB-Freigabe zu einem NetApp Connector für Amazon Q Business hinzufügen. Dadurch können Sie Dateien einschließen, die auf Volumes gespeichert sind, die von anderen Dateisystemen als Amazon FSx for NetApp ONTAP gehostet werden.

"Erstellen Sie einen NetApp Connector für Amazon Q Business"

"Hinzufügen von Datenquellen zu einem Connector"

Erweiterte Chat-Konfiguration für Wissensdatenbanken

Sie können jetzt erweiterte Chat-Einstellungen konfigurieren, die auf das Chat-Modell für die Wissensdatenbank anwendbar sind, wie z. B. Antwortlänge, Temperatur, Argumentationseinstellungen und mehr. Einige dieser Einstellungen, wie etwa Einstellungen für Aktualität und Änderungszeit, erweiterte Abrufeinstellungen und Systemaufforderungen, sind nur über die Workload Factory-API verfügbar.

"Einrichtung einer GenAI Knowledge Base"

Die Auswahl des Inferenztyps wird jetzt für Einbettungs-, Chat- und Neurangmodelle unterstützt.

Wenn Ihr ausgewähltes Einbettungs-, Chat- oder Neurankingmodell über Inferenzeinstellungen verfügt, können Sie jetzt einen Inferenztyp auswählen. Dadurch können Sie die Leistung und den Ressourcenbedarf des Chatbots besser auf Ihre Bedürfnisse abstimmen.

"Einrichtung einer GenAI Knowledge Base"

03. August 2025

Sichere Speicherung strukturierter Datenergebnisse

Wenn die Abfrageergebnisse des Chatbots strukturierte Daten enthalten, kann GenAI die Ergebnisse in einem Amazon S3-Bucket speichern. Wenn diese Ergebnisse in einem S3-Bucket gespeichert sind, können Sie sie über den Download-Link innerhalb der Chat-Sitzung herunterladen.

"Einrichtung einer GenAI Knowledge Base"

MCP-Serververfügbarkeit

NetApp stellt jetzt einen Model Context Protocol (MCP)-Server mit NetApp Workload Factory für GenAI bereit. Sie können den Server lokal installieren, um externen MCP-Clients das Ermitteln und Abrufen von Abfrageergebnissen aus einer GenAI-Wissensdatenbank zu ermöglichen.

"NetApp Workload Factory GenAI MCP-Server"

29 Juni 2025

Unterstützung für Datenquellen, die auf generischen NFS/SMB-Dateisystemen gehostet werden

Sie können jetzt eine Datenquelle aus einer generischen SMB- oder NFS-Freigabe hinzufügen. Dadurch können Sie Dateien einbeziehen, die auf Volumes gespeichert sind, die von anderen Dateisystemen als Amazon FSx für NetApp ONTAP gehostet werden.

"Hinzufügen von Datenquellen zu einer Wissensdatenbank"

"Hinzufügen von Datenquellen zu einem Connector"

03 Juni 2025

Tracker zur Überwachung und Nachverfolgung von Vorgängen verfügbar

Die Tracker-Überwachungsfunktion ist jetzt in GenAI verfügbar. Mit Tracker können Sie den Fortschritt und Status ausstehender, laufender und abgeschlossener Vorgänge überwachen und verfolgen, Details zu Vorgangsaufgaben und Unteraufgaben überprüfen, Probleme oder Fehler diagnostizieren, Parameter für fehlgeschlagene Vorgänge bearbeiten und fehlgeschlagene Vorgänge wiederholen.

"Überwachen Sie Workload-Vorgänge mit Tracker in NetApp Workload Factory"

Auswählen eines Reranking-Modells für eine Wissensdatenbank

Sie können jetzt die Relevanz neu bewerteter Abfrageergebnisse erhöhen, indem Sie ein bestimmtes Reranking-Modell für die Verwendung mit einer Wissensdatenbank auswählen. GenAI unterstützt die Modelle Cohere Rerank und Amazon Rerank.

"Einrichtung einer GenAI Knowledge Base"

04 Mai 2025

Support für NetApp Connector für Amazon Q Business

Diese Version von GenAI bietet Unterstützung für NetApp Connector für Amazon Q Business und ermöglicht Ihnen die Erstellung von Konnektoren für Amazon Q Business. Nutzen Sie den Amazon Q Business KI-Assistenten schnell und einfach mit einer geringeren Erstkonfiguration als mit dem Aufbau einer GenAI Knowledge Base für Amazon Bedrock.

"Erstellen Sie einen NetApp Connector für Amazon Q Business"

Verbesserter Support für Chatmodelle

GenAI unterstützt jetzt die folgenden zusätzlichen Chatmodelle für Wissensdatenbanken:

- "Mistral KI-Modelle"
- "Amazon Titan Textmodelle"
- "Meta Llama Modelle"
- "Jamba 1.5 Modelle"
- "Co-here Command-Modelle"
- "Deepseek-Modelle"

GenAI unterstützt die Modelle von jedem Provider, die von Amazon Bedrock unterstützt werden: "Unterstützte Basismodelle in Amazon Bedrock"

"Einrichtung einer GenAI Knowledge Base"

Terminologie für Berechtigungen wurde aktualisiert

In der Benutzeroberfläche und Dokumentation von Workload Factory wird jetzt „schreibgeschützt“ für Leseberechtigungen und „Lesen/Schreiben“ für automatisierte Berechtigungen verwendet.

02. März 2025

Verbesserungen des integrierten Chatbot

Sie können jetzt Fragen und Antworten direkt in die Zwischenablage kopieren, die Größe des Chat-Fensters anpassen und dessen Titel ändern. Darüber hinaus können Chat-Antworten jetzt Tabellen enthalten, die auch kopierbar sind.

"Testen Sie eine GenAI Knowledge Base"

Unterstützung bei der Chat-Antwort

Die Chat-Antworten enthalten nun Zitate, in denen die Dateien und Datenpakete aufgelistet sind, die zur Generierung der Antwort verwendet wurden.

"Testen Sie eine GenAI Knowledge Base"

Verbesserte Unterstützung von Dateitypen

Diese Version von GenAI bietet erweiterte Dateiunterstützung:

- Chat-Modelle bieten eine verbesserte CSV-Unterstützung. Dies ermöglicht nützlichste Antworten beim Abfragen von Daten aus CSV-Dateien.
- GenAI kann jetzt Apache Parkett Dateien aus Datenquellen aufnehmen.
- GenAI unterstützt jetzt die Aufnahme von Microsoft Word DOCX-Dateien, die Bilder enthalten. In DOCX-Dokumenten eingebettete Bilder werden gescannt, und Texteinblicke aus den eingebetteten Bildern werden in Antworten auf Knowledge-Base-Abfragen berücksichtigt.

"Unterstützte Datenquelldateiformate"

02 Februar 2025

Support für Amazon Nova-Basismodelle

GenAI unterstützt jetzt die Amazon Nova Foundation Modelle. Amazon Nova Micro, Amazon Nova Lite und Amazon Nova Pro werden unterstützt.

"GenAI-Anforderungen"

Dateitypfilterung für Datenquellen

GenAI unterstützt jetzt die Auswahl bestimmter Dateitypen, die beim Hinzufügen einer Datenquelle in den Scan der Datenquelle einbezogen werden sollen.

"Fügen Sie der Wissensdatenbank Datenquellen hinzu"

Dateiänderung nach Datenquellen filtern

GenAI unterstützt jetzt das Filtern von Dateien, die beim Hinzufügen einer Datenquelle in den Scan der Datenquelle nach Änderungsdatum einbezogen werden sollen. Sie können einen Änderungsdatumbereich für die enthaltenen Dateien auswählen.

"Fügen Sie der Wissensdatenbank Datenquellen hinzu"

Unterstützung für Bilddateien und erweiterte Unterstützung für PDF-Dateien

GenAI unterstützt jetzt die Verbesserung der Antworten auf Knowledge-Base-Abfragen durch Einblicke aus Bildern und Diagrammbeschreibungen sowie Dokumenttext, was zu reichhaltigeren und qualitativ hochwertigeren Antworten führt. GenAI kann nun Bilddateien und Bilder innerhalb von PDF-Dateien scannen (auch als Multi-modale Dateiunterstützung bekannt). Wenn Sie Bilder oder PDF-Dateien scannen, wird der Text der Bilder (einschließlich der in PDF-Dokumenten eingebetteten Bilder) in die Datenquelle gescannt und die Ergebnisse der Scans werden in die Antworten auf Knowledge-Base-Abfragen aufgenommen.

"Fügen Sie der Wissensdatenbank Datenquellen hinzu"

Hybride Suche und erneute Unterstützung

GenAI kann jetzt die Relevanz und Genauigkeit von Suchergebnissen durch hybride Suche und Neubewertung der Ergebnisse deutlich verbessern. Die hybride Suche kombiniert die Stärken der herkömmlichen Keyword-basierten Suche mit erweiterten dichten vektorbasierten semantischen Suchtechniken. Die Standard-Keyword-Suchergebnisse werden durch enge Übereinstimmungen und sprachliche Nuance erweitert, was die Relevanz erhöht. GenAI verfeinert diese Ergebnisse dann weiter, indem er erweiterte Re-Ranking-Modelle wie Cohere Rerank und Amazon Rerank verwendet, und gibt die relevantesten Ergebnisse zurück. Diese Funktion ist für neu erstellte Wissensdatenbanken verfügbar.

"Erfahren Sie mehr über NetApp Workload Factory für GenAI"

05 Januar 2025

Benutzerdefinierter Snapshot-Name

Sie können jetzt einen Snapshot-Namen für einen Ad-hoc-Snapshot bereitstellen.

"Schützen Sie eine Wissensdatenbank mit Snapshots"

Individueller Name für die AI Engine Instance

Sie können der KI-Engine-Instanz jetzt während der Implementierung einen benutzerdefinierten Namen geben.

"Implementierung der GenAI-Infrastruktur"

Wiederherstellung beschädigter oder fehlender GenAI-Infrastruktur

Wenn Ihre KI-Engine-Instanz beschädigt oder irgendwie gelöscht wird, können Sie sie von Workload Factory für Sie neu erstellen lassen. Workload Factory fügt Ihre Wissensdatenbanken nach Abschluss des Neuaufbaus automatisch wieder an die Infrastruktur an, sodass sie einsatzbereit sind.

"Fehlerbehebung"

Bis 01. Dezember 2024

Klonen Sie eine Wissensdatenbank aus einem Snapshot

NetApp Workload Factory für GenAI unterstützt jetzt das Klonen einer Wissensdatenbank aus einem Snapshot. Dies ermöglicht die schnelle Wiederherstellung von Wissensdatenbanken und die Erstellung neuer Wissensdatenbanken mit vorhandenen Datenquellen und hilft bei der Datenwiederherstellung und -entwicklung.

"Klonen einer Wissensdatenbank"

Erkennung und Replizierung von lokalen ONTAP Clustern

Entdecken und replizieren Sie lokale ONTAP Clusterdaten in ein FSx für ONTAP -Dateisystem, damit diese zur Anreicherung von KI-Wissensdatenbanken verwendet werden können. Alle lokalen Erkennungs- und Replikations-Workflows sind über das neue Menü **On-Premises ONTAP** im Speicherinventar möglich.

"Ermitteln eines lokalen ONTAP Clusters"

Bis 3. November 2024

Maskieren Sie personenbezogene Daten mithilfe von Datengeländern

Die Generative AI-Workload führt die Datenleitplankenfunktion ein, die durch die NetApp Console-Klassifizierung unterstützt wird. Die Datenleitplankenfunktion identifiziert und maskiert personenbezogene Daten (PII) und hilft Ihnen so, die Compliance aufrechtzuerhalten und die Sicherheit Ihrer sensiblen Unternehmensdaten zu erhöhen.

"Einrichtung einer GenAI Knowledge Base"

"Erfahren Sie mehr über die Klassifizierung der NetApp Konsole"

29 September 2024

Snapshot und Restore-Unterstützung für Knowledge Base Volumes

Sie können jetzt Daten aus generativen KI-Workloads schützen, indem Sie eine zeitpunktgenaue Kopie einer Knowledge Base erstellen. So können Sie Ihre Daten vor versehentlichem Verlust schützen oder Änderungen an den Einstellungen der Wissensdatenbank testen. Sie können jederzeit die vorherige Version des Knowledge Base-Volumes wiederherstellen.

"Erstellen Sie einen Snapshot eines Knowledge-Base-Volumes"

"Stellen Sie einen Snapshot eines Knowledge-Base-Volumes wieder her"

Geplante Scans anhalten

Sie können jetzt geplante Scans der Datenquelle anhalten. Standardmäßig scannt generative KI-Workloads jede Datenquelle täglich, um neue Daten in jeder Knowledge Base zu erfassen. Wenn Sie nicht möchten, dass die letzten Änderungen aufgenommen werden (z. B. während des Tests oder beim Wiederherstellen eines Snapshots), können Sie die geplanten Scans anhalten und jederzeit wieder aufnehmen.

"Management von Wissensdatenbanken"

Datensicherung-Volumes werden jetzt für Wissensdatenbanken unterstützt

Bei der Auswahl eines Knowledge-Base-Volumes können Sie nun ein Datensicherungs-Volume auswählen, das Teil einer NetApp SnapMirror Replizierungsbeziehung ist. So können Sie Wissensdatenbanken auf Volumes speichern, die bereits durch die SnapMirror-Replizierung geschützt sind.

"Ermitteln Sie die Datenquellen, die in Ihre Wissensdatenbank integriert werden sollen"

1 September 2024

Zusätzliche Chunking-Strategien

Generative KI-Workloads unterstützen jetzt Multi-Sentence Chunking und überlappendes Chunking für Datenquellen.

Dediziertes Volumen für jede Wissensdatenbank

Generative KI-Workloads erstellen jetzt für jede neue Knowledge Base ein dediziertes Amazon FSX for NetApp ONTAP Volume, mit dem individuelle Snapshot-Richtlinien für jede Knowledge Base erstellt werden. So wird auch der Schutz vor Ausfällen und Datenvergiftungen verbessert.

4 August 2024

Integration von Amazon CloudWatch Logs

Generative KI-Workloads sind jetzt in Amazon CloudWatch Logs integriert, sodass Sie die Protokolldateien generativer KI-Workloads überwachen können.

Beispiel für eine Chatbot-Anwendung

Mit der NetApp Workload Factory GenAI-Beispielanwendung können Sie die Authentifizierung und den Abruf aus Ihrer veröffentlichten NetApp Workload Factory-Wissensdatenbank testen, indem Sie in einer webbasierten Chatbot-Anwendung direkt mit ihr interagieren.

7 Juli 2024

Erstveröffentlichung der Workload Factory für GenAI

Die erste Version umfasst die Fähigkeit, eine Wissensdatenbank zu entwickeln, die durch die Einbettung der Unternehmensdaten angepasst wird. Auf die Wissensdatenbank kann über eine Chatbot-Anwendung für Ihre Benutzer zugegriffen werden. Diese Funktion gewährleistet präzise und relevante Antworten auf

unternehmensspezifische Fragen und verbessert so die Zufriedenheit und Produktivität aller Benutzer.

Erfahren Sie mehr über NetApp Workload Factory für GenAI

Erfahren Sie mehr über NetApp Workload Factory für GenAI

NetApp Workload Factory für GenAI ermöglicht Ihnen die Integration von Amazon FSx for NetApp ONTAP Dateisysteme mit GenAI-Grundmodellen. Dadurch erhalten Sie Hochleistungsspeicher mit zahlreichen Schutz-, Sicherheits- und Kostenoptimierungsfunktionen für Ihre KI-Datensätze.

Was ist NetApp Workload Factory für GenAI?

NetApp Workload Factory für GenAI ermöglicht Ihnen die Nutzung Ihrer Unternehmensdatenquellen auf Amazon FSx for NetApp ONTAP mit generativen KI-Anwendungen. Mithilfe der Retrieval-Augmented Generation (RAG) können Sie Datenquellen schnell mit über Amazon Bedrock oder Amazon Q Business verfügbaren Basismodellen verbinden, um Anwendungen auf Basis generativer KI-Technologie wie virtuelle Assistenten, Q&A-Chatbots, Dokumentzusammenfassungen, Inhaltserstellung usw. zu entwickeln.

Durch den Einsatz von Generative AI mit Ihren Unternehmensdaten können Sie Ihr eigenes Wissen und Ihre eigenen Fachkenntnisse nutzen und sich nicht nur auf die Intelligenz des Modells verlassen, die auf öffentlichen Daten basiert, auf denen die Modelle trainiert wurden. Mit RAG zur Anpassung der Modelle werden genaue und relevante Antworten auf organisationsspezifische Fragen sichergestellt, wodurch die Produktivität und Effizienz für die Benutzer Ihrer Anwendungen mithilfe von Generative AI verbessert wird.

Durch die Entwicklung einer GenAI-Applikation, die auf Ihre Unternehmensdaten zugeschnitten ist, können Sie Ihr eigenes Wissen und Ihre Fachkenntnisse nutzen. Diese Anpassungsfunktion sorgt für präzise und relevante Antworten auf unternehmensspezifische Fragen und verbessert die Zufriedenheit und Produktivität aller Benutzer.

Falls Sie "Erstellen einer Wissensdatenbank", GenAI nimmt Daten aus Ihren Datenquellen auf, speichert die vektorisierten Ergebnisse in einer Datenbank und gibt Ihnen volle Kontrolle darüber, wie Sie die aufgenommenen Daten zum Beantworten von Abfragen verwenden. Dieser Ansatz erfordert eine stärkere Erstkonfiguration, ermöglicht Ihnen jedoch die Auswahl verschiedener Chatmodelle für unterschiedliche Ergebnisse. Wenn Sie "Definieren Sie einen NetApp Connector für Amazon Q Business", werden Daten aus Ihren Datenquellen von Amazon Q Business aufgenommen und in einem Index gespeichert. Dieser Ansatz erfordert weniger Anfangskonfiguration, bietet Ihnen jedoch weniger Kontrolle über die Ergebnisse.

Weitere Informationen zu Workload Factory finden Sie im "Übersicht über Workload Factory" .

Vorteile des Einsatzes von GenAI zur Erstellung generativer KI-Applikationen

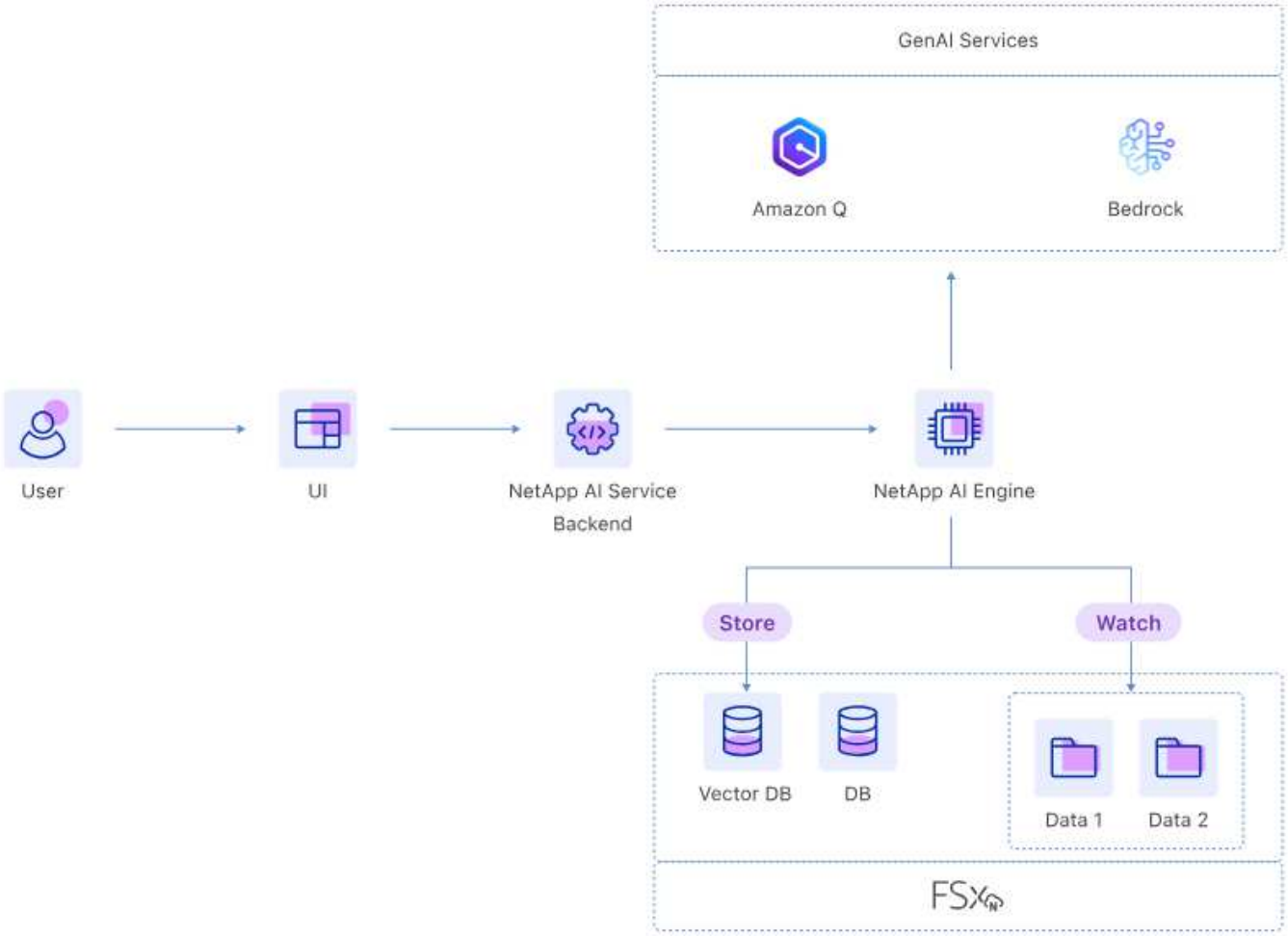
NetApp Workload Factory für GenAI vereinfacht den Prozess der Bereitstellung der Infrastruktur, die zum Erstellen generativer KI-Anwendungen mithilfe von Retrieval-Augmented Generation (RAG) erforderlich ist. Konkret bietet GenAI die folgenden Vorteile:

- Durch die Automatisierung von GenAI können IT-Administratoren und -Entwickler die Applikationsentwicklung beschleunigen, ohne umfassende Kenntnisse über Dateninfrastruktur, Grundlagen und Sprachmodelle benötigen. Datenadministratoren und -Entwickler können auf einfache und schnelle Weise unternehmenseigene Wissensdatenbanken erstellen, in die unstrukturierte Daten Ihres Unternehmens für generative KI-Applikationen eingebettet werden.

- Erhöhen Sie die Sicherheit, indem Sie Benutzerberechtigungen in Dateien erhalten, die in Knowledge-Base eingebettet sind, um sicherzustellen, dass Datensicherheit und Datenschutz erhalten bleiben. Eine Anwendung, wie ein Chatbot, kann so entwickelt werden, dass nur authentifizierte Benutzer Antworten erhalten, die auf Daten basieren, auf die die Benutzer zugreifen können.
- Schützen Sie Ihre Unternehmensdaten in Ihrem AWS Kundenkonto und bewahren Sie sie sicher auf. Dabei werden Ihre Unternehmensdaten nie an Dritte zugänglich gemacht.
- Beschleunigen Sie die Entwicklung von GenAI-Applikationen, wie z. B. einen Q&A-Chatbot, mithilfe von Open-Source-Frameworks wie LangChain. Verwenden Sie die GenAI-API, um Wissensdatenbanken und -Konnektoren bereitzustellen und zu managen, mit einer Knowledge Base zu chatten und den Chat-Verlauf zu speichern und abzurufen.
- Verbessern Sie die Datensicherung und -Verfügbarkeit, indem Sie die generative KI-Dateninfrastruktur auf FSX für NetApp ONTAP-Filesysteme implementieren und ONTAP Funktionen wie Hochverfügbarkeit, Snapshots für lokale Datensicherung und -Recovery, SnapMirror für Disaster Recovery und SnapVault für Backup Ihrer Dateninfrastruktur nutzen.
- Profitieren Sie von den ONTAP Dateneffizienzfunktionen wie Datendeduplizierung, Komprimierung und Data-Compaction, Daten-Tiering und Thin Provisioning und senken Sie so die Storage-Gesamtkosten für die generative KI-Dateninfrastruktur.
- Mit der hybriden Suche und den von GenAI bereitgestellten Neurankungsfunktionen erhalten Sie hochwertige Ergebnisse aus Ihren Daten. Die hybride Suche in Kombination mit einem Re-Ranking verbessert die Relevanz der Suchergebnisse deutlich. Diese Funktionen sind über Amazon AWS verfügbar und regionsabhängig.

So funktioniert GenAI

GenAI verwendet die privaten Daten Ihres Unternehmens, um die Intelligenz des Modells (basierend auf den Daten, auf denen es trainiert wurde) zu ergänzen und personalisierte Antworten auf Fragen von Benutzern in Ihrem Unternehmen zu geben. Sie implementieren zuerst die Infrastruktur, die für ein RAG-Framework erforderlich ist, erstellen dann eine Wissensdatenbank oder definieren einen Konnektor mithilfe der Datenquellen und Basismodelle Ihres Unternehmens, die über Amazon Bedrock oder Amazon Q Business verfügbar sind, und verbinden dann eine Applikation (z. B. einen Q&A-Chatbot) mit der Wissensdatenbank oder dem Connector.



So hilft NetApp Workload Factory für GenAI beim Erstellen generativer KI-Anwendungen

GenAI hilft wie folgt, generative KI-Applikationen mit RAG zu erstellen:

- Implementiert die erforderliche Infrastruktur für das RAG-Framework (Retrieval Augmented Generation), um mit Datenquellen auf FSX für ONTAP-Filesystemen und Amazon Bedrock oder Amazon Q Business zusammenzuarbeiten. Die Infrastruktur umfasst die NetApp GenAI Engine-Instanz zum Managen von Daten, eine eingebettete Vektordatenbank (LanceDB) und den Speicher auf Ihrem FSX für ONTAP-Dateisystem für die Vektordatenbank.
- Hilft, die Datenquellen mit Einbettungen und Sprachmodellen zu verbinden, die über Amazon Bedrock oder Amazon Q Business verfügbar sind, um Datenquellen einzubetten und Antworten für Benutzeranfragen abzurufen. Die Datenquellen werden zusammen mit Modellen und ihrer Konfiguration als FSX for ONTAP-Wissensdatenbanken dargestellt.
- Eingliederung von Quelldaten in die Wissensdatenbank oder den Connector zur Einbindung von Quelldateien in SMB-Freigaben und NFS-Exporte in FSX für ONTAP-Filesysteme sowie Speicherung von Dateiberechtigungen für Dateien in SMB-Freigaben.
- Entwickelt automatisch Einstiegsfragen basierend auf dem Inhalt in Wissensdatenbanken.
- Bietet einen Chat-Simulator für Datenadministratoren, um das Chatten mit Wissensdatenbanken zu testen.
- Über eine einfache Connector-Schnittstelle kann GenAI schnell und einfach mit Amazon Q Business verbunden werden.

Tools zur Verwendung von NetApp Workload Factory

Sie können NetApp Workload Factory mit den folgenden Tools verwenden:

- **Workload Factory-Konsole**: Die Workload Factory-Konsole bietet eine visuelle, ganzheitliche Ansicht Ihrer Anwendungen und Projekte.
- * NetApp Konsole*: Die NetApp Konsole bietet eine hybride Benutzeroberfläche, sodass Sie Workload Factory zusammen mit anderen NetApp -Datendiensten verwenden können.
- **Fragen Sie mich**: Verwenden Sie den KI-Assistenten „Fragen Sie mich“, um Fragen zu stellen und mehr über Workload Factory zu erfahren, ohne die Workload Factory-Konsole zu verlassen. Greifen Sie über das Hilfemenü von Workload Factory auf „Fragen Sie mich“ zu.
- **CloudShell CLI**: Workload Factory enthält eine CloudShell CLI zum Verwalten und Betreiben von AWS- und NetApp -Umgebungen über Konten hinweg von einer einzigen, browserbasierten CLI aus. Greifen Sie über die obere Leiste der Workload Factory-Konsole auf CloudShell zu.
- **REST-API**: Verwenden Sie die Workload Factory REST-APIs, um Ihre FSx für ONTAP Dateisysteme und andere AWS-Ressourcen bereitzustellen und zu verwalten.
- **CloudFormation**: Verwenden Sie AWS CloudFormation-Code, um die Aktionen auszuführen, die Sie in der Workload Factory-Konsole definiert haben, um AWS- und Drittanbieterressourcen aus dem CloudFormation-Stack in Ihrem AWS-Konto zu modellieren, bereitzustellen und zu verwalten.
- **Terraform NetApp Workload Factory-Anbieter**: Verwenden Sie Terraform, um in der Workload Factory-Konsole generierte Infrastruktur-Workflows zu erstellen und zu verwalten.

Kosten

Die Nutzung der GenAI-Funktion von Workload Factory ist kostenlos.

Sie müssen jedoch für die von Ihnen implementierten AWS Ressourcen zahlen, um die generative KI-Infrastruktur unterstützen zu können. Sie zahlen beispielsweise AWS für Amazon Bedrock oder Amazon Q Business, FSX für ONTAP-Dateisystem- und Storage-Kapazität und die EC2-Instanz für die GenAI Engine.

Einige multimodale Operationen, wie das Scannen von Bildern nach Textinformationen, können mehr Ressourcen verwenden und damit höhere Kosten verursachen. Einige Konfigurationsvorgänge, wie das Ändern von Einstellungen für eine Wissensdatenbank, können dazu führen, dass Datenquellen erneut gescannt werden, und Datenquellprüfungen können auch höhere Kosten verursachen.

Lizenzierung

Für die Nutzung der KI-Funktionen von Workload Factory sind keine speziellen Lizenzen von NetApp erforderlich.

Regionen

Workload Factory wird in allen kommerziellen Regionen unterstützt, in denen FSx für ONTAP unterstützt wird."Sehen Sie sich die unterstützten Amazon-Regionen an."

Die folgenden AWS-Regionen werden nicht unterstützt:

- Regionen Chinas
- GovCloud (USA)-Regionen
- Geheime Cloud

- Streng geheime Cloud

Komponenten der NetApp GenAI Engine

Wenn Sie die GenAI-Infrastruktur bereitstellen, erstellt Workload Factory eine EC2-Instanz für die GenAI-Engine. Außerdem werden eine IAM-Rolle, eine Sicherheitsgruppe und private Endpunkte für diese Instanz erstellt. Möglicherweise möchten Sie mehr Details zu diesen Komponenten erfahren, die Workload Factory in Ihrer AWS-Umgebung erstellt.

EC2 Instanztyp

m5.large

IAM-Rolle

Die GenAI-Engine-Instanz benötigt die Berechtigungen, um Datenblöcke an das Einbettungsmodell in Amazon Bedrock zu senden und mit dem NetApp AI Service Back-End zu kommunizieren. Die IAM-Rolle umfasst die folgenden Berechtigungen:

IAM-Rollenberechtigungen

```
{
  "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"
      }
    ]
}
```

Sicherheitsgruppe

Die Regeln für ausgehende Anrufe sind für den gesamten Datenverkehr offen, während die Regeln für eingehende Anrufe vollständig geschlossen sind.

Private Endpunkte

Wenn die Ziel-VPC diese noch nicht hat, erstellt Workload Factory private Endpunkte für die EC2-Instance der GenAI-Engine, damit diese mit den folgenden AWS-Diensten kommunizieren kann:

- Amazonas-Grundgestein
  - Grundgestein
  - Bedrock-Runtime
  - Bedrock-Agent-Runtime
- Amazon Elastic Container Registry (ECR)
  - API
  - docker
- AWS System Manager (SSM)
  - SSM

    - ec2messages
    - Ssmbotschaften
- Amazon FSX für NetApp ONTAP
- Amazon CloudWatch

Mit GenAI können Sie Wissensdatenbanken für Amazon Bedrock aufbauen

Los geht’s

Der schnelle Einstieg in GenAI Knowledge Base

Erste Schritte zum Erstellen einer Wissensdatenbank oder eines Amazon Q Business Connectors unter Verwendung der Daten Ihres Unternehmens, die auf Amazon FSX for NetApp ONTAP-Dateisystemen vorhanden sind. Eine Anwendung wie ein Chatbot greift auf diese Wissensdatenbank oder den Konnektor zu, um den Endbenutzern organisationsbezogene Antworten bereitzustellen.

Melden Sie sich bei Workload Factory an

Sie müssen "Einrichten eines Kontos bei Workload Factory" und melden Sie sich mit einem der "Konsolenerfahrungen" .

Richten Sie Ihre Umgebung so ein, dass sie die GenAI-Anforderungen erfüllt

Für die Implementierung der AWS-Infrastruktur, eines implementierten und erkannten FSX for ONTAP-Filesystems, der Liste der Datenquellen, die Sie in Ihre Knowledge Base oder Connector integrieren möchten, des Zugriffs auf den Amazon Bedrock KI-Service oder die Amazon Q Business-Applikation und vieles mehr benötigen Sie Zugangsdaten von AWS.

"Weitere Informationen zu den GenAI-Anforderungen".

Ermitteln Sie das FSX für ONTAP-Dateisystem, das die Datenquellen enthält

Die Datenquellen, die Sie in Ihre Wissensdatenbank integrieren, können auf einem einzigen FSX für ONTAP-Dateisystem oder auf mehreren FSX für ONTAP-Dateisystemen abgelegt werden. Wenn sich diese Systeme in verschiedenen VPCs befinden, müssen sie entweder innerhalb desselben Netzwerks zugänglich sein oder die VPCs müssen mit derselben Region und demselben AWS-Konto wie die KI-Engine ausgestattet werden.

"Erfahren Sie, wie Sie Datenquellen erkennen".

Implementierung der GenAI-Infrastruktur

Starten Sie den Infrastruktur-Implementierungsassistenten, um die GenAI-Infrastruktur in Ihrer AWS Umgebung zu implementieren. Bei diesem Prozess werden eine EC2-Instanz für die NetApp-GenAI-Engine und ein Volume auf einem FSX für ONTAP-Filesystem implementiert, das die Datenbanken der NetApp AI Engine enthält. Das Volume dient zum Speichern der von der Wissensdatenbank verwendeten Vektordatenbank.

"Erfahren Sie, wie Sie die Knowledge Base Infrastruktur implementieren".

Wie es weiter geht

Sie können jetzt eine Wissensdatenbank aufbauen, um Endbenutzern organisationsbezogene Antworten zu bieten.

Anforderungen der GenAI Knowledge Base

Stellen Sie sicher, dass Workload Factory und AWS ordnungsgemäß eingerichtet sind, bevor Sie Ihre Wissensdatenbank erstellen. Dazu gehören Ihre AWS-Anmeldeinformationen, ein bereitgestelltes FSx für ONTAP -Dateisystem, das die Datenquellen enthält, die Sie in Ihre Wissensdatenbank integrieren möchten, Zugriff auf den Amazon Bedrock AI-Dienst und mehr.

Grundlegende GenAI-Anforderungen

Für GenAI gelten allgemeine Anforderungen, die Ihre Umgebung vor der Inbetriebnahme erfüllen muss.

Workload Factory-Anmeldung und -Konto

Sie müssen "Einrichten eines Kontos bei Workload Factory" und melden Sie sich mit einem der "Konsolenerfahrungen" .

AWS Zugangsdaten und Berechtigungen

Sie müssen Workload Factory AWS-Anmeldeinformationen mit Lese-/Schreibberechtigungen hinzufügen, was bedeutet, dass Sie Workload Factory für GenAI im Lese-/Schreibmodus verwenden.

Basic Mode und *Read-Only* Mode Berechtigungen werden derzeit nicht unterstützt.

Wenn Sie Ihre Anmeldedaten einrichten, erhalten Sie durch die Auswahl der unten aufgeführten Berechtigungen vollständigen Zugriff auf das Management von FSX für ONTAP-Dateisysteme sowie auf die Bereitstellung und das Management der GenAI EC2-Instanz und anderer AWS-Ressourcen, die für Ihre Knowledge Base und Ihren Chatbot benötigt werden.

"Erfahren Sie, wie Sie AWS-Anmeldeinformationen zu Workload Factory hinzufügen"

Anforderungen der GenAI Knowledge Base

Wenn Sie mit Wissensdatenbanken arbeiten möchten, stellen Sie sicher, dass Ihre Umgebung die folgenden Anforderungen erfüllt.

Amazonas-Grundgestein

Mit Amazon Bedrock können Sie Grundmodelle verwenden und es bietet die Möglichkeiten, generative KI-Anwendungen zu erstellen.

Bevor Sie mit NetApp Workload Factory für GenAI beginnen, müssen Sie Amazon Bedrock einrichten. Ihre GenAI-Bereitstellung muss sich in einer AWS-Region befinden, in der Amazon Bedrock aktiviert ist.

- "AWS-Dokumentation: Einrichten von Amazon Bedrock"
- "AWS-Dokumentation: Unterstützte Regionen und Modelle für Knowledge Basen für Amazon Bedrock"

GenAI ordnet die Suchergebnisse standardmäßig neu ein, um die Relevanz der Ergebnisse zu verbessern. Um die besten Ergebnisse zu erzielen, stellen Sie sicher, dass Ihre Konfiguration des Basismodells Amazon Bedrock den Zugriff auf ein neues Modell wie Cohere Rerank oder Amazon Rerank umfasst, falls es in Ihrer Region verfügbar ist.

Einbettungsmodell

Sie müssen das einzubetende Modell aktivieren, das Sie verwenden möchten, bevor Sie Ihre Wissensdatenbank erstellen. Die folgenden Einbettungsmodelle werden unterstützt:

- Titan-Einbauteile G1 - Text
- Titan Einbettung Text v2
- Titan Multimodal Embeddings G1
- Englisch Einbetten
- Mehrsprachig Einbetten

  "Erfahren Sie mehr über Amazon Titan"

Chat-Modell

Sie müssen das grundlegende Chatmodell aktivieren, das Sie verwenden möchten, bevor Sie Ihre Wissensdatenbank erstellen. Da der Modellsupport je nach AWS-Region variiert, finden Sie unter "Die AWS Dokumentation" Informationen dazu, welche Modelle in den Regionen verwendet werden können, in denen Sie Ihre Knowledge Base bereitstellen möchten.

GenAI unterstützt verschiedene Modelle von Anthropic, Amazon, Mistral AI, Meta, Jamba und Cohhere.

Erfahren Sie mehr über die Verwendung dieser Modelle in Amazon Bedrock:

- "Anthropic's Claude in Amazon Bedrock"
- "Erste Schritte mit Amazon Nova in der Amazon Bedrock Konsole"
- "Mistral KI-Modelle"
- "Amazon Titan Textmodelle"
- "Meta Llama Modelle"
- "Jamba-Modelle"
- "Co-here Command-Modelle"

FSX für ONTAP-Dateisystem

Sie benötigen mindestens ein FSX für ONTAP-Dateisystem:

- Ein Dateisystem wird von der NetApp GenAI-Engine verwendet (oder erstellt, falls es nicht existiert), um die von der Wissensdatenbank verwendete Vektordatenbank zu speichern.

  Dieses Dateisystem FSX für ONTAP muss FlexVol-Volumen verwenden. FlexGroup Volumes werden nicht unterstützt.

- Ein oder mehrere Dateisysteme enthalten die Datenquellen, die Sie in Ihre Wissensdatenbank integrieren werden.

  Ein FSX für ONTAP-Dateisystem kann für beide dieser Zwecke verwendet werden, oder Sie können mehrere FSX für ONTAP-Dateisysteme verwenden.

- Sie müssen die AWS-Region, VPC und das Subnetz kennen, in der sich das AWS FSX für ONTAP-Filesystem befindet. Das Filesystem muss sich in einer AWS-Region befinden, für die Amazon Bedrock aktiviert ist.
- Sie müssen die Tag-Schlüssel-/Wertpaare berücksichtigen, die Sie auf die AWS Ressourcen anwenden

möchten, die Teil dieser Implementierung sind (optional).

- Dabei sind die wichtigsten Paar-Informationen wichtig, die eine sichere Verbindung zur NetApp AI Engine-Instanz ermöglichen.

"Erfahren Sie, wie Sie FSX für ONTAP-Dateisysteme implementieren und managen"

Identifizieren von Datenquellen, die zu einer Wissensdatenbank oder einem Konnektor hinzugefügt werden sollen

Identifizieren oder erstellen Sie die Dokumente (Datenquellen), die sich auf Ihrem FSX für ONTAP-Dateisystem befinden, die Sie in Ihre Wissensdatenbank integrieren werden. Mithilfe dieser Datenquellen kann die Wissensdatenbank genaue und personalisierte Antworten auf Benutzeranfragen basierend auf Daten bereitstellen, die für Ihr Unternehmen relevant sind.

Maximale Anzahl von Datenquellen

Die maximale Anzahl unterstützter Datenquellen beträgt 10.

Speicherort der Datenquellen

Datenquellen können auf einem einzelnen Volume oder in einem Ordner innerhalb eines Volumes, auf einer SMB-Freigabe oder auf einem NFS-Export auf einem Amazon FSX für NetApp ONTAP Filesystem gespeichert werden. Datenquellen können auch auf Amazon FSX for NetApp ONTAP Volumes in einer NetApp SnapMirror Datensicherungsbeziehung gespeichert werden.

Sie können keine einzelnen Dokumente in einem Volume oder Ordner auswählen. Daher sollten Sie sicherstellen, dass jedes Volume oder jeder Ordner, das Datenquellen enthält, keine fremden Dokumente enthält, die nicht in Ihre Wissensdatenbank integriert werden sollten.

Sie können jeder Knowledge Base mehrere Datenquellen hinzufügen, allerdings müssen alle auf FSX for ONTAP-Filesystemen liegen, auf die über Ihr AWS-Konto zugegriffen werden kann.

Die maximale Dateigröße für jede Datenquelle beträgt 50 MB.

Unterstützte Protokolle

Wissensdatenbanken unterstützen Daten von Volumes, die entweder NFS- oder SMB/CIFS-Protokolle verwenden. Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, damit die Wissensdatenbank auf die Dateien auf diesen Volumes zugreifen kann. Dazu gehören die Active Directory-Domäne, die IP-Adresse, der Benutzername und das Kennwort.

Wenn Sie Ihre Datenquelle auf einer über SMB zugänglichen Freigabe (Datei oder Verzeichnis) speichern, sind die Daten nur für Chatbot-Benutzer oder -Gruppen zugänglich, die über die Berechtigung zum Zugriff auf diese Freigabe verfügen. Wenn diese "Berechtigungserkennungsfunktion" aktiviert ist, vergleicht das KI-System die Benutzer-E-Mail in auth0 mit den Benutzern, die die Dateien auf der SMB-Freigabe anzeigen oder verwenden dürfen. Der Chatbot liefert Antworten basierend auf Benutzerberechtigungen für die eingebetteten Dateien.

Wenn Sie beispielsweise 10 Dateien (Datenquellen) in Ihre Wissensdatenbank integriert haben und 2 der Dateien Personaldateien mit eingeschränkten Informationen sind, erhalten nur Chatbot-Benutzer, die für den Zugriff auf diese 2 Dateien authentifiziert sind, Antworten vom Chatbot, die Daten aus diesen Dateien enthalten.

Unterstützte Datenquelldateiformate

Die folgenden Datenquellendateiformate werden derzeit von Workload Factory GenAI-Wissensdatenbanken unterstützt.

Dateiformat	Erweiterung
Apache Parquet [1]	.Parkett
Kommagetrennte Werte file[1]	.Csv
Format Für Den Austausch Von Grafiken	.Gif
JPEG	.Jpg or.jpeg
JSON und JSONP[1]	.Json
Markdown	.md
Microsoft Word	.Doc oder .docx
Nur Text	.Txt
Tragbares Dokumentformat	.Pdf
Tragbare Netzwerkgrafiken	.Png
WEBP-Bild	.webp

Implementierung der GenAI-Infrastruktur

Sie müssen die GenAI-Infrastruktur für RAG Framework in Ihrer Umgebung implementieren, bevor Sie FSX für ONTAP-Wissensdatenbanken, -Konnektoren und -Applikationen für Ihr Unternehmen entwickeln können. Die primären Infrastrukturkomponenten sind der Amazon Bedrock Service, eine Virtual Machine-Instanz für die NetApp GenAI Engine und ein FSX für ONTAP Filesystem.

Die implementierte Infrastruktur kann mehrere Wissensdatenbanken, Chatbots und Konnektoren unterstützen, sodass Sie diese Aufgabe in der Regel nur einmal durchführen müssen.

Detaillierte Informationen zur Infrastruktur

Ihre GenAI-Implementierung muss sich in einer AWS-Region mit aktiviertem Amazon Bedrock befinden. "Zeigen Sie die Liste der unterstützten Regionen an"

Die Infrastruktur besteht aus folgenden Komponenten:

Amazon Bedrock Service

Amazon Bedrock ist ein vollständig gemanagter Service, mit dem Sie über eine einzige API Foundation Models (FMS) führender KI-Unternehmen verwenden können. Außerdem bietet sie die Funktionen, die Sie zum Erstellen sicherer generativer KI-Applikationen benötigen.

"Erfahren Sie mehr über Amazon Bedrock"

Amazon Q Business

Amazon Q baut auf Amazon Bedrock auf, um einen vollständig gemanagten generativen KI-Assistenten bereitzustellen, mit dem Sie Fragen beantworten und Inhalte basierend auf Informationen aus Ihren

Datenquellen generieren können.

"Erfahren Sie mehr über Amazon Q Business"

Virtual Machine für die NetApp GenAI Engine

Dabei wird die NetApp GenAI Engine implementiert. Es bietet die Verarbeitungsleistung, um die Daten aus Ihren Datenquellen aufzunehmen und diese dann in die Vektordatenbank zu schreiben.

FSX für ONTAP-Dateisystem

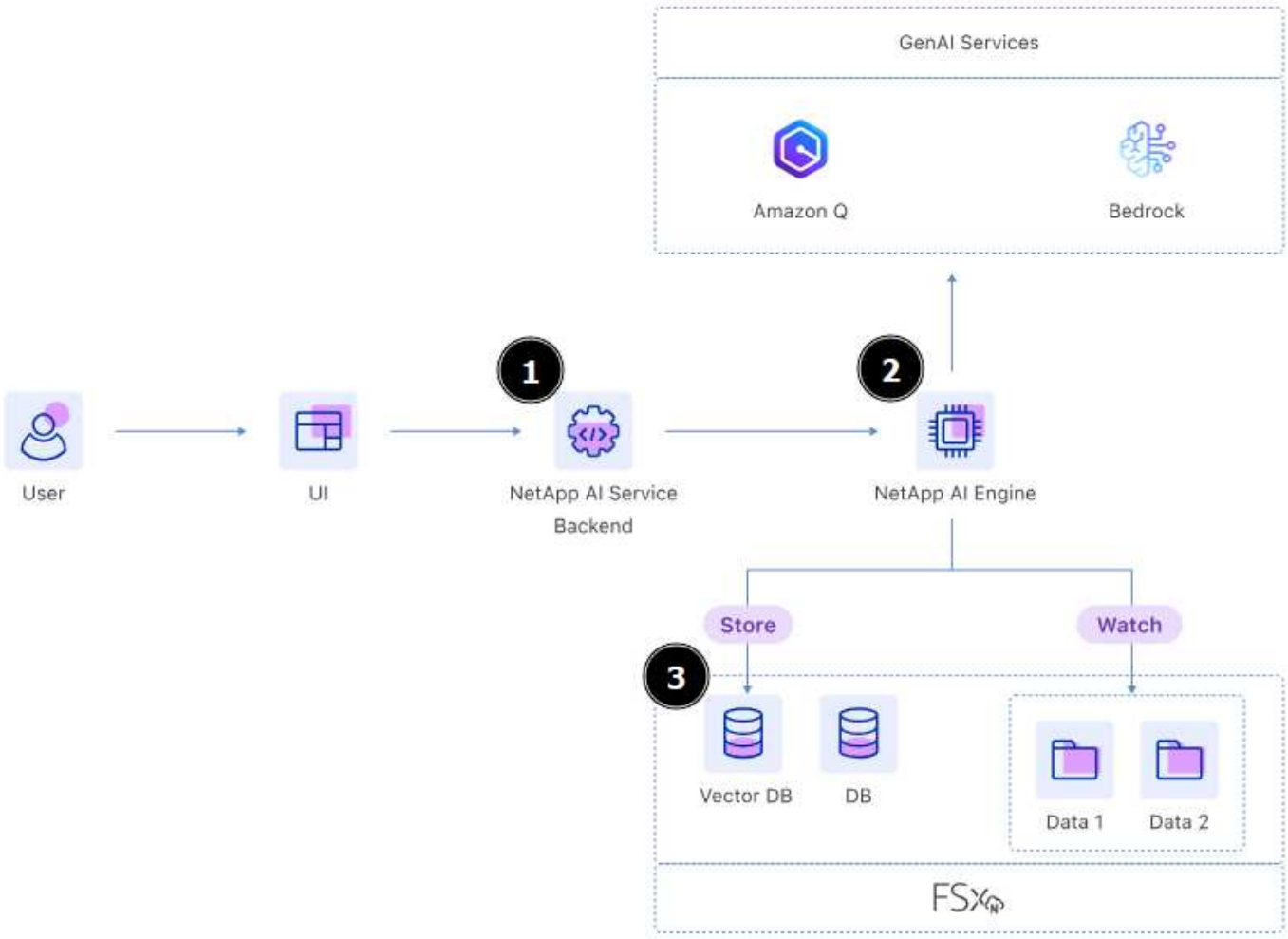
Das Filesystem FSX für ONTAP bietet den Storage für Ihr GenAI-System.

Es wird ein einzelnes Volume bereitgestellt, das die Vektordatenbank enthält, in der die Daten gespeichert werden, die vom Basismodell basierend auf Ihren Datenquellen generiert wurden.

Die Datenquellen, die Sie in Ihre Wissensdatenbank integrieren, können sich auf demselben FSX für ONTAP-Dateisystem oder auf einem anderen System befinden.

Die NetApp GenAI Engine überwacht beide Volumes und interagiert mit ihnen.

Die folgende Abbildung zeigt die GenAI-Infrastruktur. Die Komponenten mit den Nummern 1, 2 und 3 werden während dieses Verfahrens eingesetzt. Die anderen Elemente müssen vorhanden sein, bevor mit der Bereitstellung begonnen werden kann.



Implementierung der GenAI-Infrastruktur

Geben Sie Ihre AWS-Zugangsdaten ein und wählen Sie das Filesystem FSX für ONTAP aus, um die Infrastruktur zur Retrieval-Augmented Generation (RAG) implementieren zu können.

Bevor Sie beginnen

Stellen Sie sicher, dass Ihre Umgebung die Anforderungen für Wissensdatenbanken oder Konnektoren erfüllt, je nachdem, welche Sie wählen, bevor Sie mit diesem Verfahren beginnen.

- "Anforderungen an die Wissensdatenbank"
- "Anforderungen an Steckverbinder"

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Sehen Sie sich das Infrastrukturdiagramm an und wählen Sie **Weiter**.
4. Füllen Sie die Einträge im Abschnitt **AWS-Einstellungen** aus:
   a. **AWS-Anmeldeinformationen**: Wählen Sie die AWS-Anmeldeinformationen aus oder fügen Sie diese hinzu, um die AWS-Ressourcen bereitzustellen.
   b. **Standort**: Wählen Sie eine AWS-Region, VPC und Subnetz.

      Die Implementierung von GenAI muss sich in einer AWS Region befinden, für die Amazon Bedrock aktiviert ist. "Zeigen Sie die Liste der unterstützten Regionen an"
5. Füllen Sie die Einträge im Abschnitt **Infrastruktureinstellungen** aus:
   a. **Tags**: Geben Sie beliebige Tag-Schlüssel/Wert-Paare ein, die Sie auf alle AWS-Ressourcen anwenden möchten, die Teil dieser Bereitstellung sind. Diese Tags sind in der AWS-Managementkonsole und im Infrastrukturinformationsbereich innerhalb von Workload Factory sichtbar und können Ihnen dabei helfen, den Überblick über die Workload Factory-Ressourcen zu behalten.
6. Füllen Sie den Abschnitt **Konnektivität** aus:
   a. **Schlüsselpaar**: Wählen Sie ein Schlüsselpaar aus, mit dem Sie eine sichere Verbindung zur NetApp GenAI Engine-Instanz herstellen können.
7. Füllen Sie den Abschnitt **AI Engine** aus:
   a. **Instanzname**: Wählen Sie optional **Instanznamen definieren** aus und geben Sie einen benutzerdefinierten Namen für die KI-Engine-Instanz ein. Der Instanzname wird in der AWS-Managementkonsole und im Infrastrukturinformationsbereich innerhalb von Workload Factory angezeigt und kann Ihnen dabei helfen, den Überblick über die Workload Factory-Ressourcen zu behalten.
8. Wählen Sie **Deploy**, um die Bereitstellung zu starten.

Wenn die Bereitstellung mit einem Fehler bei den Anmeldeinformationen fehlschlägt, erhalten Sie weitere Fehlerdetails, indem Sie die Hyperlinks in der Fehlermeldung auswählen. Sie sehen eine Liste mit fehlenden oder blockierten Berechtigungen sowie eine Liste mit Berechtigungen, die der GenAI-Workload benötigt, um die GenAI-Infrastruktur bereitstellen zu können.

Ergebnis

Workload Factory beginnt mit der Bereitstellung der Chatbot-Infrastruktur. Dieser Vorgang kann bis zu 10

Minuten dauern.

Während des Bereitstellungsprozesses werden die folgenden Elemente eingerichtet:

- Das Netzwerk wird zusammen mit den privaten Endpunkten eingerichtet.
- Die IAM-Rolle, das Instanzprofil und die Sicherheitsgruppe werden erstellt.
- Die Virtual Machine-Instanz für die GenAI-Engine wird bereitgestellt.
- Amazon Bedrock ist so konfiguriert, dass Protokolle an Amazon CloudWatch Logs gesendet werden, wobei eine Protokollgruppe mit dem Präfix verwendet `/aws/bedrock/`wird.
- Die GenAI-Engine ist so konfiguriert, dass sie Protokolle an Amazon CloudWatch Logs sendet, wobei eine Protokollgruppe mit dem Namen `/netapp/wlmai/<tenancyAccountId>/randomId` , Wo `<tenancyAccountID>` ist die "NetApp -Konsolen-Konto-ID" für den aktuellen Benutzer.

Einrichtung einer GenAI Knowledge Base

Nachdem Sie die KI-Infrastruktur bereitgestellt und die Datenquellen identifiziert haben, die Sie aus Ihren FSx for ONTAP -Datenspeichern in Ihre Wissensdatenbank integrieren möchten, können Sie mit dem Erstellen der Wissensdatenbank mithilfe von Workload Factory beginnen. Im Rahmen dieses Schritts definieren Sie auch die KI-Eigenschaften und erstellen Gesprächseinstiege.

Stellen Sie sicher, dass Ihre Umgebung den für Wissensdatenbanken entspricht"Anforderungen", bevor Sie fortfahren.

Über diese Aufgabe

Wissensdatenbanken haben zwei Datenintegrationsmodalitäten - *public Mode* und *Enterprise Mode*.

Öffentlicher Modus

Eine Wissensdatenbank kann verwendet werden, ohne Datenquellen aus Ihrem Unternehmen zu integrieren. In diesem Fall liefert eine mit der Wissensdatenbank integrierte Anwendung nur Ergebnisse aus öffentlich zugänglichen Informationen im Internet. Dies wird als „*public Mode*“-Integration bezeichnet.

Enterprise-Modus

In den meisten Fällen möchten Sie Datenquellen aus Ihrem Unternehmen in die Wissensdatenbank integrieren. Dies wird als *Enterprise Mode* Integration bezeichnet, da es Wissen aus Ihrem Unternehmen liefert.

Datenquellen Ihrer Organisation können personenbezogene Daten (PII) enthalten. Um diese vertraulichen Informationen zu schützen, können Sie beim Erstellen und Konfigurieren von Wissensdatenbanken *Datenleitplanken* aktivieren. Datenleitplanken, unterstützt durch NetApp Data Classification, identifizieren und maskieren personenbezogene Daten und machen sie so unzugänglich und unwiederbringlich.

"Erfahren Sie mehr über die NetApp Datenklassifizierung" .

NetApp Workload Factory für GenAI maskiert keine sensiblen persönlichen Informationen (SPii). Siehe"Arten sensibler personenbezogener Daten" für weitere Informationen zu dieser Art von Daten.

Datenleitplanken können jederzeit aktiviert oder deaktiviert werden. Wenn Sie die Aktivierung von Datenleitplanken ändern, scannt Workload Factory die gesamte Wissensdatenbank von Grund auf neu, was Kosten verursacht.

Erstellen und Konfigurieren der Wissensdatenbank

Die Wissensdatenbank definiert Eigenschaften wie die Bedrock-KI-Modelle und das Embedding-Format, mit denen Sie Ihre Wissensdatenbank erstellen möchten.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie im Menü „Wissensdatenbanken und Konnektoren“ das Dropdown-Menü „Neu erstellen“ und wählen Sie „NetApp GenAI-Wissensdatenbank für Bedrock“ aus.
4. Konfigurieren Sie auf der Seite „NetApp GenAI-Wissensdatenbank erstellen“ die Einstellungen für die Wissensdatenbank:

Details zur Wissensdatenbank

1. **Name**: Geben Sie den Namen ein, den Sie für die Wissensdatenbank verwenden möchten.
2. **Beschreibung**: Geben Sie eine detaillierte Beschreibung für die Wissensdatenbank ein.
3. **Bedrock**: Wählen Sie die Region, in der Amazon Bedrock für Ihr AWS-Konto verfügbar ist.

Verschlucken

1. **Einbettungsmodell**:
   - Wählen Sie ein Einbettungsmodell zur Verwendung für die Wissensdatenbank aus. Das Einbettungsmodell definiert, wie Ihre Daten in Vektoreinbettungen für die Wissensdatenbank konvertiert werden. Workload Factory unterstützt die folgenden Modelle:
   - Titan-Einbauteile G1 - Text
   - Titan Einbettung Text v2
   - Titan Multimodal Embeddings G1
   - Englisch Einbetten
   - Mehrsprachig Einbetten

     Beachten Sie, dass Sie das Einbettungsmodell von Amazon Bedrock bereits aktiviert haben müssen.

   "Erfahren Sie mehr über Amazon Titan"

   - Wählen Sie gegebenenfalls den Inferenztyp aus, der der Konfiguration des ausgewählten Einbettungsmodells entspricht.
2. **Datenleitplanken**: Wählen Sie, ob Sie Datenleitplanken aktivieren oder deaktivieren möchten. "Erfahren Sie mehr über Datenleitplanken, unterstützt durch NetApp Data Classification" .

   Die folgenden Voraussetzungen müssen erfüllt sein, um die Data Guardrails zu aktivieren.

   - Für die Kommunikation mit NetApp Data Classification ist ein Servicekonto erforderlich. Für die Erstellung eines Dienstkontos müssen Sie in Ihrem NetApp Console-Mandantenkonto über die Rolle

„Organisationsadministrator“ verfügen. Ein Mitglied mit der Rolle des Organisationsadministrators kann alle Aktionen in der ausführen."Erfahren Sie, wie Sie einem Mitglied in der NetApp -Konsole eine Rolle hinzufügen"

- Die KI-Engine muss Zugriff auf die"NetApp Console API-Endpunkt" .
- Sie müssen Folgendes tun, wie in"Dokumentation zur NetApp -Datenklassifizierung" :
  i. Erstellen eines Konsolenagenten
  ii. Stellen Sie sicher, dass Ihre Umgebung die Voraussetzungen erfüllt
  iii. Bereitstellen der NetApp Datenklassifizierung

Die Datenguardrails-Funktion wird bei der Aufnahme strukturierter Datendateien wie CSV, JSON, JSONP oder Parkett nicht unterstützt.

Chat- und Abrufeinstellungen

1. **Chat-Modell**:
   - Wählen Sie aus verschiedenen Chat-Modellen, die in Amazon Bedrock integriert sind. Beachten Sie, dass Sie das Chat-Modell von Amazon Bedrock bereits aktiviert haben müssen.
   - Wählen Sie gegebenenfalls den Inferenztyp aus, der der Konfiguration des ausgewählten Modells entspricht.
2. **Chat-Einstellungen**:
   - Wählen Sie eine Temperatur für den Chatbot, um die Zufälligkeit und Kreativität der Antworten zu konfigurieren. Eine niedrigere Temperatur führt zu vorhersehbareren Reaktionen und eine höhere Temperatur zu vielfältigeren Reaktionen.
   - Wählen Sie eine maximale Antwortlänge, um zu konfigurieren, wie detailliert die Antworten sein sollen. Längere Antwortlängen verbrauchen mehr Antworttoken und können höhere Kosten verursachen.
3. **Denkmodus**: Wenn der Denkmodus aktiviert ist, benötigt der Chatbot mehr Zeit zum Verarbeiten von Anfragen und die Ergebnisse sind in der Regel genauer. Wenn Sie den Denkmodus aktivieren, können Sie steuern, wie viele Argumentationstoken beim Generieren von Ergebnissen verwendet werden. Die Verwendung mehrerer Argumentationstoken kann zu genaueren Antworten führen, ist jedoch möglicherweise mit höheren Kosten verbunden.
4. **Neurangfolge**: Aktivieren oder deaktivieren Sie die Neurangfolge, die die Relevanz und Qualität der Abfrageergebnisse verbessern kann. Wählen Sie für die Neubewertung ein Standard-Chat-Modell oder ein spezielles Reranker-Modell aus. Reranker-Modelloptionen werden nur angezeigt, wenn sie in Ihrer Region verfügbar sind. Wählen Sie den Inferenztyp aus, der der Konfiguration des ausgewählten Modells entspricht.
5. **Gesprächseinstieg**: Wählen Sie aus, ob Sie bis zu vier Gesprächseinstiegstelefonate bereitstellen möchten, die Benutzern angezeigt werden, die mit einem Chatbot interagieren, der diese Wissensdatenbank nutzt. Wir empfehlen, diese Einstellung zu aktivieren.

   Wenn Sie Gesprächseinstieg aktivieren, ist standardmäßig der automatische Modus ausgewählt. Der „manuelle Modus“ kann erst aktiviert werden, nachdem Sie Datenquellen zu Ihrer Wissensdatenbank hinzugefügt haben. "Erfahren Sie, wie Sie die Einstellungen der Wissensdatenbank ändern".

Speicherdefinitionen

1. **FSx für ONTAP -Dateisystem**: Wenn Sie eine neue Wissensdatenbank definieren, erstellt Workload Factory ein neues Amazon FSx for NetApp ONTAP Volume, um sie zu speichern. Wählen Sie einen vorhandenen Dateisystemnamen und eine SVM (auch Speicher-VM genannt), auf der das neue Volume

erstellt wird.

2. **Snapshot-Richtlinie**: Wählen Sie eine Snapshot-Richtlinie aus der Liste der vorhandenen Richtlinien aus, die im Workload Factory-Speicherinventar definiert sind. Es werden automatisch wiederkehrende Snapshots der Wissensdatenbank in einer Häufigkeit erstellt, die auf der von Ihnen ausgewählten Snapshot-Richtlinie basiert.
3. **S3-Bucket**: Wenn die Abfrageergebnisse des Chatbots strukturierte Daten enthalten, kann GenAI die Ergebnisse in einem S3-Bucket speichern. Um diese Funktion zu verwenden, aktivieren Sie die Einstellung **S3-Bucket aktivieren** und wählen Sie aus der Liste einen S3-Bucket aus, der mit Ihrem Konto verknüpft ist. Wenn diese Ergebnisse in einem S3-Bucket gespeichert sind, können Sie sie über den Download-Link innerhalb der Chat-Sitzung herunterladen.

   Wenn die benötigte Snapshot-Richtlinie nicht vorhanden ist, können Sie "Erstellen einer Snapshot-Richtlinie" auf der Storage-VM, die das Volume enthält, arbeiten.

4. Wählen Sie **Create Knowledge Base**, um die Wissensdatenbank zu GenAI hinzuzufügen.

   Während der Erstellung der Wissensdatenbank wird ein Fortschrittsindikator angezeigt.

   Nachdem die Wissensdatenbank erstellt wurde, haben Sie die Möglichkeit, eine Datenquelle zu Ihrer neuen Wissensdatenbank hinzuzufügen oder den Prozess zu beenden, ohne eine Datenquelle hinzuzufügen. Wir empfehlen Ihnen, **Add Data source** auszuwählen und jetzt eine oder mehrere Datenquellen hinzuzufügen.

Fügen Sie der Wissensdatenbank Datenquellen hinzu

Sie können eine oder mehrere Datenquellen hinzufügen, um die Wissensdatenbank mit den Daten Ihres Unternehmens zu füllen.

Über diese Aufgabe

Die maximale Anzahl unterstützter Datenquellen beträgt 10.

Schritte

1. Nachdem Sie **Datenquelle hinzufügen** ausgewählt haben, wählen Sie den Typ der Datenquelle aus, die Sie hinzufügen möchten:
   - FSx für ONTAP-Dateisystem hinzufügen (Dateien aus einem vorhandenen FSx für ONTAP-Volume verwenden)
   - Dateisystem hinzufügen (Dateien aus einer generischen SMB- oder NFS-Freigabe verwenden)

Fügen Sie ein FSx für ONTAP-Dateisystem hinzu

1. **Wählen Sie ein Dateisystem**: Wählen Sie das FSX für ONTAP Dateisystem, in dem sich Ihre Datenquelldateien befinden, und wählen Sie **Weiter**.
2. **Wählen Sie ein Volume**: Wählen Sie das Volume aus, auf dem sich Ihre Quelldateien befinden, und wählen Sie **Weiter**.

   Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, einschließlich Domäne, IP-Adresse, Benutzername und Passwort.
3. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.
4. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

       Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.
   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

       Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

     Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

     Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.
   - Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien

basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

> Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

5. Im Abschnitt **permission aware**, der nur verfügbar ist, wenn sich die von Ihnen ausgewählte Datenquelle auf einem Volume befindet, das das SMB-Protokoll verwendet, können Sie permission-aware Antworten aktivieren oder deaktivieren:
   - **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
   - **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.
6. Wählen Sie **Hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen NFS-Dateisystems

1. **Wählen Sie ein Dateisystem**: Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden, wählen Sie das NFS-Protokoll für die Netzwerkfreigabe und klicken Sie auf **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

> In manchen Fällen müssen Sie den NFS-Exportnamen manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können den gesamten Export oder nur bestimmte Ordner auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:
   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

> Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder

wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

- Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen SMB-Dateisystems

1. **Dateisystem auswählen**:
   a. Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden.
   b. Wählen Sie das SMB-Protokoll für die Netzwerkfreigabe.
   c. Geben Sie die Active Directory-Informationen ein, darunter Domäne, IP-Adresse, Benutzername und Kennwort.
   d. Wählen Sie **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

In manchen Fällen müssen Sie den Namen der SMB-Freigabe manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können die gesamte Freigabe oder nur bestimmte Ordner darin auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:

- **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
- **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

  Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

◦ **Berechtigungsbewusst**: Aktivieren oder deaktivieren Sie berechtigungsbewusste Antworten:

- **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
- **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.

◦ **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:

- Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

  Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

◦ Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Ergebnis

Die Datenquelle wird in Ihre Wissensdatenbank eingebettet. Der Status ändert sich von „Einbetten“ in „eingebettet“, wenn die Datenquelle vollständig eingebettet ist.

Nachdem Sie der Wissensdatenbank eine einzelne Datenquelle hinzugefügt haben, können Sie diese lokal im Chatbot-Simulator testen und alle erforderlichen Änderungen vornehmen, bevor Sie den Chatbot Ihren Benutzern zur Verfügung stellen. Sie können die gleichen Schritte ausführen, um der Wissensdatenbank zusätzliche Datenquellen hinzuzufügen.

Testen Sie eine GenAI Knowledge Base

Nachdem Sie die Wissensdatenbank erstellt haben, können Sie sie mithilfe des Chatbot-Simulators lokal testen und erforderliche Änderungen vornehmen, bevor Sie die Wissensdatenbank Ihren Benutzern über eine Chatbot-Anwendung zur Verfügung stellen.

Über diese Aufgabe

Sie testen Ihre Wissensdatenbank, um sicherzustellen, dass sie funktioniert, wie Sie es erwarten, und Sie können die Konversationsstarter anpassen, die Sie standardmäßig für Chatbot-Benutzer dieser Wissensdatenbank verfügbar sein möchten. Der Chatbot-Simulator läuft gegen alle Datenquellen, die in die Wissensdatenbank eingebettet wurden.

Sie können eine Wissensdatenbank testen, indem Sie mit Ihren eingebetteten Datenquellen im Chatbot-Simulator chatten. Beachten Sie, dass beim lokalen Testen der Wissensdatenbank keine der Interaktionen oder Erkenntnisse in der GenAI-Vektordatenbank erfasst werden.

Sie führen die meisten Ihrer Tests in Workload Factory durch, bevor Sie die Wissensdatenbank in einer Anwendung für Ihre Benutzer bereitstellen. Wenn Sie Änderungen an Ihrer Datenquelle oder dem Chatbot-Betrieb vornehmen müssen, sollten Sie dies jetzt tun, bevor Sie Ihre Wissensdatenbank veröffentlichen.

Sie können die Größe des Chatbot-Simulatorfensters ändern und neu benennen sowie Fragen und Antworten in die Zwischenablage kopieren.

Einige der Aufgaben, die Sie durchführen möchten, um Ihren Chatbot zu testen, sind:

- Geben Sie eine große Anzahl von Fragen ein, die für Ihr Unternehmen relevant sind, um sicherzustellen, dass die Antworten den Erwartungen entsprechen.
- Passen Sie die Gesprächseinstiege an, die Sie standardmäßig für Ihre Benutzer in der Chatbot-Anwendung zur Verfügung stellen möchten.
- Stellen Sie sicher, dass der zugeschriebene Inhalt, der unten in den Chatbot-Antworten angegeben wird, die richtigen Referenzen enthält.

Schritte

1. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, die Sie testen möchten.

   Der Chatbot-Simulator wird im rechten Fensterbereich angezeigt. Falls festgelegt, werden auch bestehende Gesprächseinstiege angezeigt.

2. Geben Sie im Chatbot-Eingabefeld eine Eingabeaufforderung oder Frage ein, und wählen Sie aus , wie Ihr Chatbot mit Ihrem organisatorischen Wissen reagiert.

- Sie können die Quellen sehen, die zur Erstellung der Antwort verwendet werden, indem Sie die Liste **sources** unter der Antwort erweitern. Hier finden Sie eine Liste der Dateien, die zum Generieren der Antwort verwendet werden. Sie können die von jeder Datei und jedem Volume verwendeten Datenblöcke anzeigen und in jede Datei kopieren, indem Sie den Mauszeiger über den Dateinamen bewegen.
- Wenn die Antwort Tabellen enthält, können Sie die Daten in jeder Spalte sortieren und jede Tabelle in die Zwischenablage kopieren.
- Wenn die Antwortergebnisse strukturierte Daten enthalten und die Funktion **S3 Bucket** für die Wissensdatenbank aktiviert ist, speichert GenAI die Ergebnisse in einem S3-Bucket. Sie können die Ergebnisse aus dem Bucket über den Link **Ergebnisse herunterladen** innerhalb der Chat-Sitzung herunterladen.

3. Wenn Sie eine Ihrer Datenquellen aktualisieren müssen, damit Ihre Wissensdatenbank gezielter Antworten liefert, nehmen Sie diese Änderungen jetzt vor und testen Sie die Wissensdatenbank erneut.

Aktivieren der externen Authentifizierung für eine GenAI-Wissensdatenbank

Aktivieren Sie die Authentifizierung für eine Wissensdatenbank, damit Token-Validierung und ACLs erforderlich sind, wenn die API-Endpunkte zur Integration einer Wissensdatenbank in eine Chatbot-Anwendung verwendet werden. Wenn Sie die Authentifizierung aktivieren, konfigurieren Sie Einstellungen für ein JSON-Webtoken, das für API-Anfragen von Chatbot-Clients an eine Wissensdatenbank verwendet wird.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, für die Sie die Authentifizierung aktivieren möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.
5. Wählen Sie das Menü **actions** und dann **Manage Authentication settings**.
6. Authentifizierung einrichten:

   a. Wählen Sie **Authentifizierungseinstellungen aktivieren**.

   b. Geben Sie die erforderlichen Informationen an. Beispiele sind angegeben, aber Sie sollten die Werte für diese Felder von Ihrem Authentifizierungsanbieter beziehen:

      - **Algorithmen**: Der Signierungsalgorithmus, den Ihr Authentifizierungsanbieter verwendet.
      - **Audience** (Optional): Eine Zeichenkette, die den gewünschten Empfänger des Tokens enthält (manchmal eine URL).
      - **Aussteller**: Eine Zeichenfolge, die den Anbieter identifiziert, der das Token ausgegeben hat.

        Amazon Cognito verwendet beispielsweise Emittent Strings mit dem folgenden Format:

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

Wobei <region> die AWS-Region, die den Benutzerpool enthält, und <UserPoolID> die Benutzer-Pool-ID ist. Sie können Ihre Benutzer-Pool-ID mit dem folgenden Befehl abrufen:

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

- **JWKS URI**: Der URI-String, der öffentliche Schlüssel bereitstellt, die benötigt werden, um die Signaturen dieses Tokens zu überprüfen.

  Amazon Cognito verwendet beispielsweise JWKS URI-Strings mit dem folgenden Format:

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

+
Wobei <region> die AWS-Region, die den Benutzerpool enthält, und <UserPoolID> die Benutzer-Pool-ID ist. Sie können Ihre Benutzer-Pool-ID mit dem folgenden Befehl abrufen:

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

7. Wählen Sie **Speichern**.

Ergebnis

Die Authentifizierung für die Wissensdatenbank ist jetzt aktiv, und Sie können API-Endpunkte verwenden, um mit der Wissensdatenbank zu interagieren und die Wissensdatenbank in eine Chatbot-Anwendung zu integrieren.

Veröffentlichen Sie eine GenAI Knowledge Base, und zeigen Sie den eindeutigen Endpunkt an

Nachdem Sie Ihre Wissensdatenbank lokal erstellt und getestet haben, können Sie die Wissensdatenbank veröffentlichen, sodass sie in eine Chatbot-Anwendung integriert werden kann, mit der Ihre Benutzer die Wissensdatenbank abfragen können.

Über diese Aufgabe

Durch die Veröffentlichung der Wissensdatenbank können Sie diese in Chat-Anwendungen verwenden. Die Veröffentlichungsaktion löst die Workload Factory-API aus, um eindeutige Endpunkte zu generieren und zu veröffentlichen. Nach der Veröffentlichung wird die Wissensdatenbank für Chat-Anwendungen zugänglich und die API-Endpunkte sind für die Integration bereit.

Jede von Ihnen veröffentlichte Wissensdatenbank verfügt über eindeutige Endpunkte.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Wissensdatenbanken die Wissensdatenbank aus, die Sie veröffentlichen möchten.

4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus und die Liste aller eingebetteten Datenquellen angezeigt.

5. Wählen Sie das Menü **actions** und dann **Publish**.

   Workload Factory veröffentlicht die Wissensdatenbank. Auf der Detailseite der Wissensdatenbank ändert sich der Status von **Unveröffentlicht** zu **Veröffentlicht**.

   Sie können jetzt Details zum eindeutigen Endpunkt für die Wissensdatenbank abrufen.

6. Wählen Sie neben dem veröffentlichten Status **Ansicht**.

   Es werden Details zum Zugriff auf die Wissensdatenbank mithilfe der Workload Factory-API angezeigt.

7. Kopieren Sie im Dialogfeld **veröffentlichte Informationen anzeigen** die API-Endpunkte, mit denen Sie die Wissensdatenbank in eine Anwendung integrieren können.

   Um mehr über die API-Endpunkte zu erfahren, gehen Sie zu "API-Dokumentation" und wählen Sie **AI > Extern**.

   Bevor Sie diese Endpunkte verwenden können, müssen Sie ein Benutzertoken von Ihrem Authentifizierungsanbieter beziehen.

Ergebnis

Sie verfügen nun über eine veröffentlichte Wissensdatenbank und den einzigartigen Endpunkt, mit dem Sie die Wissensdatenbank in eine Chatbot-Anwendung integrieren können.

Verwenden Sie die externe Anwendung des GenAI-Beispiels Chatbot

Nachdem Sie eine Wissensdatenbank konfiguriert, aktiviert und veröffentlicht haben, können externe Anwendungsentwickler die von NetApp bereitgestellte Open-Source-Beispiel-Chatbot-Anwendung konfigurieren und ausführen, um mit Ihrer Wissensdatenbank zu interagieren und zu lernen, wie sie mit der Workload Factory API ihre eigenen generativen KI-Anwendungen erstellen.

Schritte

1. "Erstellen einer Wissensdatenbank".
2. "Aktivieren Sie die Authentifizierung" Für die von Ihnen erstellte Wissensdatenbank.

   Dies ermöglicht der Wissensdatenbank die Authentifizierung von API-Anforderungen und macht die Token-Validierung und ACLs erforderlich, wenn die API-Endpunkte verwendet werden.

   > ⓘ Externe Chat-Anwendungen, die in diese Wissensdatenbank integriert werden, müssen denselben Authentifizierungsanbieter (Emittent) verwenden, den Sie in den Authentifizierungseinstellungen für die Wissensdatenbank konfigurieren.

3. "Veröffentlichen Sie die Wissensdatenbank" Um den API-Zugriff für externe Anwendungen zu aktivieren.

Nachdem eine Wissensdatenbank veröffentlicht wurde, sind die API-Endpunkte extern zugänglich, und Sie können die Wissensdatenbank mit einer externen Chat-Anwendung (wie der Beispiel-Chatbot-Anwendung) integrieren.

4. Laden Sie das Beispiel-Chatbot-Anwendungspaket von herunter "GitHub".
5. Installieren und starten Sie die Chatbot-Anwendung, indem Sie die Anweisungen in der im Paket enthaltenen README-Datei befolgen.
6. Navigieren Sie zu "http://localhost:9091" , um sich bei der Anwendung anzumelden.

   Die Chatbot-Beispielanwendung wird angezeigt.

Weitere Informationen .

"Workload Factory-API-Dokumentation"

Erstellen Sie eine RAG-basierte GenAI-Anwendung

Nachdem Sie Ihre Wissensdatenbank aufgebaut und Ihren Chatbot getestet haben, können Sie die Anwendung einrichten, mit der Ihre Benutzer den Chatbot abfragen können.

"So erstellen Sie eine RAG-basierte KI-Applikation auf FSX für ONTAP"

Was Sie mit GenAI erreichen können

Nachdem Sie jetzt eine Wissensdatenbank mit Ihren Unternehmensdaten erstellt und für Ihre Benutzer bereitgestellt haben, können Sie die Wissensdatenbank, Datenquellen und die RAG-Infrastruktur, einschließlich FSX für ONTAP-Dateisysteme, verwalten.

Einige der Aufgaben, die Sie zur Verwaltung Ihrer Knowledge-Base-Komponenten durchführen können, sind:

- Aktualisieren Sie den Inhalt Ihrer Datenquellen, oder fügen Sie neue Datenquellen hinzu, und synchronisieren Sie diese Änderungen mit Ihrer Wissensdatenbank und Ihrem Chatbot.
- Managen Sie Ihre Datenquelleinstellungen, einschließlich der Chunking-Strategie und Berechtigungsorientierung (für SMB-Dateizugriff).
- Verwalten Sie Ihre Knowledge Base-Einstellungen, einschließlich Chat-Modell und Gesprächseinstieg.
- Heben Sie die Veröffentlichung einer Wissensdatenbank auf, oder veröffentlichen Sie sie erneut, nachdem Sie Änderungen vorgenommen haben.
- Sichern und schützen Sie die wichtigen Daten auf Ihrem FSX für ONTAP-Dateisystem, um sicherzustellen, dass Ihre Wissensdatenbank-Daten und andere Infrastrukturkomponenten immer verfügbar sind.

Informationen zur Verwaltung Ihres FSx for ONTAP Dateisystems finden Sie im "Workload Factory für Amazon FSx for NetApp ONTAP -Dokumentation" um die Sicherungs- und Schutzfunktionen anzuzeigen, die Sie verwenden können.

[1] das Data Guardrails Feature wird nicht unterstützt, wenn strukturierte Datendateien in Wissensdatenbanken aufgenommen werden.

Mit GenAI können Sie Konnektoren für Amazon Q Business erstellen

Los geht’s

Schnellstart für GenAI-Anschlüsse

Beginnen Sie mit der Erstellung eines NetApp Connectors für Amazon Q Business unter Verwendung der Daten Ihrer Organisation, die auf Amazon FSx für NetApp ONTAP-Dateisysteme vorhanden sind. Nachdem Sie einen Konnektor erstellt haben, können Endbenutzer auf den Amazon Q Business Assistant zugreifen, um auf ihre Fragen auf organisationsbezogene Antworten zuzugreifen.

Melden Sie sich bei Workload Factory an

Sie müssen "Einrichten eines Kontos bei Workload Factory" und melden Sie sich mit einem der "Konsolenerfahrungen" .

Richten Sie Ihre Umgebung so ein, dass sie die GenAI-Anforderungen erfüllt

Für die Implementierung der AWS-Infrastruktur, eines implementierten und erkannten FSX for ONTAP-Dateisystems, der Liste der Datenquellen, die Sie in Ihren Connector integrieren möchten, des Zugriffs auf die Amazon Q Business-Anwendung und vieles mehr benötigen Sie AWS-Anmeldeinformationen.

"Weitere Informationen zu den GenAI-Anforderungen".

Ermitteln Sie das FSX für ONTAP-Dateisystem, das die Datenquellen enthält

Die Datenquellen, die Sie in Ihren Connector integrieren, können sich auf einem einzelnen FSX für ONTAP-Dateisystem oder auf mehreren FSX für ONTAP-Dateisystemen befinden. Wenn sich diese Systeme in verschiedenen VPCs befinden, müssen sie entweder innerhalb desselben Netzwerks zugänglich sein oder die VPCs müssen mit derselben Region und demselben AWS-Konto wie die KI-Engine ausgestattet werden.

"Erfahren Sie, wie Sie Datenquellen erkennen".

Implementierung der GenAI-Infrastruktur

Starten Sie den Infrastruktur-Implementierungsassistenten, um die GenAI-Infrastruktur in Ihrer AWS Umgebung zu implementieren. Bei diesem Prozess werden eine EC2-Instanz für die NetApp-GenAI-Engine und ein Volume auf einem FSX für ONTAP-Filesystem implementiert, das die Datenbanken der NetApp AI Engine enthält. Das Volume dient zum Speichern von Informationen über den Anschluss.

"Erfahren Sie, wie Sie die GenAI-Infrastruktur implementieren".

Wie es weiter geht

Sie können jetzt einen Connector für Amazon Q Business erstellen, um den Endbenutzern organisationsbezogene Antworten bereitzustellen.

Anforderungen an den GenAI-Anschluss

Stellen Sie sicher, dass Workload Factory und AWS ordnungsgemäß eingerichtet sind, bevor Sie einen NetApp Connector für Amazon Q Business erstellen.

Grundlegende GenAI-Anforderungen

Für GenAI gelten allgemeine Anforderungen, die Ihre Umgebung vor der Inbetriebnahme erfüllen muss.

Workload Factory-Anmeldung und -Konto

Sie müssen "Einrichten eines Kontos bei Workload Factory" und melden Sie sich mit einem der "Konsolenerfahrungen" .

AWS Zugangsdaten und Berechtigungen

Sie müssen Workload Factory AWS-Anmeldeinformationen mit Lese-/Schreibberechtigungen hinzufügen, was bedeutet, dass Sie Workload Factory für GenAI im Lese-/Schreibmodus verwenden.

Die Berechtigungen im *Basismodus* und *Nur-Lese-Modus* werden derzeit nicht unterstützt.

Wenn Sie Ihre Anmeldedaten einrichten, erhalten Sie durch die Auswahl der unten aufgeführten Berechtigungen vollständigen Zugriff auf das Management von FSX für ONTAP-Dateisysteme sowie auf die Bereitstellung und das Management der GenAI EC2-Instanz und anderer AWS-Ressourcen, die für Ihre Knowledge Base und Ihren Chatbot benötigt werden.

"Erfahren Sie, wie Sie AWS-Anmeldeinformationen zu Workload Factory hinzufügen"

Anforderungen für NetApp Connector für Amazon Q Business

Stellen Sie sicher, dass Ihre Umgebung die folgenden spezifischen Anforderungen für Amazon Q Business Connectors erfüllt.

Amazon Q Business-Anwendung

Sie müssen eine Amazon Q Business-Anwendung erstellen oder eine vorhandene verwenden.

- Stellen Sie sicher, dass die Applikation in einer Ihrer AWS Regionen vorhanden ist.
- Stellen Sie sicher, dass Sie für die Anwendung haben "Index erstellt".
- Stellen Sie sicher, dass sich die Anwendung nicht in einem fehlerhaften Zustand befindet.

FSX für ONTAP-Dateisystem

Sie benötigen mindestens ein FSX für ONTAP-Dateisystem:

- Ein Dateisystem wird von der NetApp GenAI Engine verwendet (oder erstellt, falls es nicht vorhanden ist), um Informationen über den Connector zu speichern.

  Dieses Dateisystem FSX für ONTAP muss FlexVol-Volumen verwenden. FlexGroup Volumes werden nicht unterstützt.

- Ein oder mehrere Dateisysteme enthalten die Datenquellen, die Sie Ihrem Connector hinzufügen werden.

  Ein FSX für ONTAP-Dateisystem kann für beide dieser Zwecke verwendet werden, oder Sie können mehrere FSX für ONTAP-Dateisysteme verwenden.

- Sie müssen die AWS-Region, VPC und das Subnetz kennen, in der sich das AWS FSX für ONTAP-Filesystem befindet.
- Sie müssen die Tag-Schlüssel-/Wertpaare berücksichtigen, die Sie auf die AWS Ressourcen anwenden möchten, die Teil dieser Implementierung sind (optional).
- Dabei sind die wichtigsten Paar-Informationen wichtig, die eine sichere Verbindung zur NetApp AI Engine-Instanz ermöglichen.

"Erfahren Sie, wie Sie FSX für ONTAP-Dateisysteme implementieren und managen"

Identifizieren von Datenquellen, die einem Konnektor hinzugefügt werden sollen

Identifizieren oder erstellen Sie die Dokumente (Datenquellen), die sich auf Ihrem FSX für ONTAP-Dateisystem befinden, die Sie in Ihren Connector integrieren werden. Mithilfe dieser Datenquellen kann Amazon Q Business genaue und personalisierte Antworten auf Benutzeranfragen basierend auf Daten liefern, die für Ihr Unternehmen relevant sind.

Maximale Anzahl von Datenquellen

Die maximale Anzahl unterstützter Datenquellen beträgt 10.

Speicherort der Datenquellen

Datenquellen können auf einem einzelnen Volume oder in einem Ordner innerhalb eines Volumes, auf einer SMB-Freigabe oder auf einem NFS-Export auf einem Amazon FSX für NetApp ONTAP Filesystem gespeichert werden. Datenquellen können auch auf Amazon FSX for NetApp ONTAP Volumes in einer NetApp SnapMirror Datensicherungsbeziehung gespeichert werden.

Sie können keine einzelnen Dokumente in einem Volume oder Ordner auswählen. Daher sollten Sie sicherstellen, dass jedes Volume oder jeder Ordner, das Datenquellen enthält, keine fremden Dokumente enthält, die nicht in Ihre Wissensdatenbank integriert werden sollten.

Sie können jedem Connector mehrere Datenquellen hinzufügen, diese müssen jedoch alle in FSX for ONTAP-Filesystemen liegen, auf die über Ihr AWS-Konto zugegriffen werden kann.

Die maximale Dateigröße für jede Datenquelle beträgt 50 MB.

Unterstützte Protokolle

Konnektoren unterstützen Daten von Volumes, die entweder NFS- oder SMB/CIFS-Protokolle verwenden. Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, damit der Connector auf die Dateien auf diesen Volumes zugreifen kann. Dazu gehören die Active Directory-Domäne, die IP-Adresse, der Benutzername und das Kennwort.

Wenn Sie Ihre Datenquelle auf einer über SMB zugänglichen Freigabe (Datei oder Verzeichnis) speichern, sind die Daten nur für Chatbot-Benutzer oder -Gruppen zugänglich, die über die Berechtigung zum Zugriff auf diese Freigabe verfügen. Wenn diese "Berechtigungserkennungsfunktion" aktiviert ist, vergleicht das KI-System die Benutzer-E-Mail in auth0 mit den Benutzern, die die Dateien auf der SMB-Freigabe anzeigen oder verwenden dürfen. Der Chatbot liefert Antworten basierend auf Benutzerberechtigungen für die eingebetteten Dateien.

Wenn Sie beispielsweise 10 Dateien (Datenquellen) in Ihren Konnektor integriert haben und 2 der Dateien Personaldateien mit eingeschränkten Informationen sind, erhalten nur Chatbot-Benutzer, die für den Zugriff auf diese 2 Dateien authentifiziert sind, Antworten vom Chatbot, die Daten aus diesen Dateien enthalten.

Wenn Sie Datenquellen zu einem Amazon Q Business Connector hinzufügen, gelten nur Benutzerberechtigungen für Datenquelldateien. Gruppenberechtigungen werden nicht angewendet.

Wenn in einer Datei in Ihrer Datenquelle kein Text vorhanden ist (z. B. ein textfreies Bild), indiziert Amazon Q Business diese Datei nicht, sondern protokolliert einen Eintrag in Amazon CloudWatch Logs, der das Fehlen von Text feststellt.

Unterstützte Datenquelldateiformate

Die folgenden Datenquellendateiformate werden derzeit mit NetApp Connector für Amazon Q Business unterstützt.

Dateiformat	Erweiterung
Durch Kommas getrennte Wertedatei	.Csv
JSON und JSONP	.Json
Markdown	.md
Microsoft Word	.Docx
Nur Text	.Txt
Tragbares Dokumentformat	.Pdf
Microsoft PowerPoint	.Ppt oder .pptx
Hypertext Markup Language	.HTML
Erweiterbare Markup-Sprache	.XML
XSLT	.xslt
Microsoft Excel	.Xls
Rich Text-Format	.Rtf

Implementierung der GenAI-Infrastruktur

Sie müssen die GenAI-Infrastruktur für RAG Framework in Ihrer Umgebung implementieren, bevor Sie FSX für ONTAP-Wissensdatenbanken, -Konnektoren und -Applikationen für Ihr Unternehmen entwickeln können. Die primären Infrastrukturkomponenten sind der Amazon Bedrock Service, eine Virtual Machine-Instanz für die NetApp GenAI Engine und ein FSX für ONTAP Filesystem.

Die implementierte Infrastruktur kann mehrere Wissensdatenbanken, Chatbots und Konnektoren unterstützen, sodass Sie diese Aufgabe in der Regel nur einmal durchführen müssen.

Detaillierte Informationen zur Infrastruktur

Ihre GenAI-Implementierung muss sich in einer AWS-Region mit aktiviertem Amazon Bedrock befinden. "Zeigen Sie die Liste der unterstützten Regionen an"

Die Infrastruktur besteht aus folgenden Komponenten:

Amazon Bedrock Service

Amazon Bedrock ist ein vollständig gemanagter Service, mit dem Sie über eine einzige API Foundation Models (FMS) führender KI-Unternehmen verwenden können. Außerdem bietet sie die Funktionen, die Sie zum Erstellen sicherer generativer KI-Applikationen benötigen.

"Erfahren Sie mehr über Amazon Bedrock"

Amazon Q Business

Amazon Q baut auf Amazon Bedrock auf, um einen vollständig gemanagten generativen KI-Assistenten bereitzustellen, mit dem Sie Fragen beantworten und Inhalte basierend auf Informationen aus Ihren Datenquellen generieren können.

"Erfahren Sie mehr über Amazon Q Business"

Virtual Machine für die NetApp GenAI Engine

Dabei wird die NetApp GenAI Engine implementiert. Es bietet die Verarbeitungsleistung, um die Daten aus Ihren Datenquellen aufzunehmen und diese dann in die Vektordatenbank zu schreiben.

FSX für ONTAP-Dateisystem

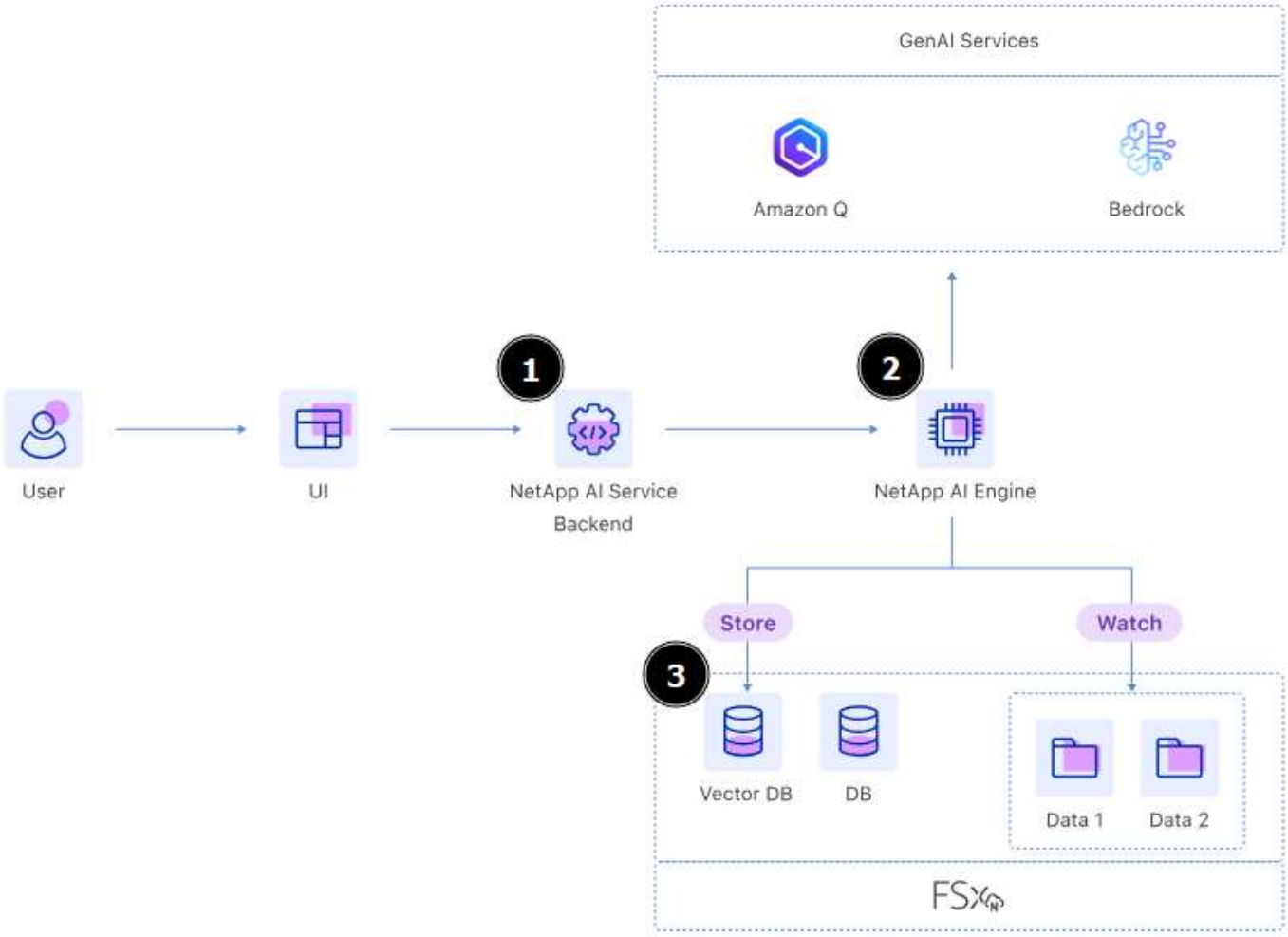
Das Filesystem FSX für ONTAP bietet den Storage für Ihr GenAI-System.

Es wird ein einzelnes Volume bereitgestellt, das die Vektordatenbank enthält, in der die Daten gespeichert werden, die vom Basismodell basierend auf Ihren Datenquellen generiert wurden.

Die Datenquellen, die Sie in Ihre Wissensdatenbank integrieren, können sich auf demselben FSX für ONTAP-Dateisystem oder auf einem anderen System befinden.

Die NetApp GenAI Engine überwacht beide Volumes und interagiert mit ihnen.

Die folgende Abbildung zeigt die GenAI-Infrastruktur. Die Komponenten mit den Nummern 1, 2 und 3 werden während dieses Verfahrens eingesetzt. Die anderen Elemente müssen vorhanden sein, bevor mit der Bereitstellung begonnen werden kann.



Implementierung der GenAI-Infrastruktur

Geben Sie Ihre AWS-Zugangsdaten ein und wählen Sie das Filesystem FSX für ONTAP aus, um die Infrastruktur zur Retrieval-Augmented Generation (RAG) implementieren zu können.

Bevor Sie beginnen

Stellen Sie sicher, dass Ihre Umgebung die Anforderungen für Wissensdatenbanken oder Konnektoren erfüllt, je nachdem, welche Sie wählen, bevor Sie mit diesem Verfahren beginnen.

- "Anforderungen an die Wissensdatenbank"
- "Anforderungen an Steckverbinder"

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Sehen Sie sich das Infrastrukturdiagramm an und wählen Sie **Weiter**.
4. Füllen Sie die Einträge im Abschnitt **AWS-Einstellungen** aus:
   a. **AWS-Anmeldeinformationen**: Wählen Sie die AWS-Anmeldeinformationen aus oder fügen Sie diese hinzu, um die AWS-Ressourcen bereitzustellen.
   b. **Standort**: Wählen Sie eine AWS-Region, VPC und Subnetz.

Die Implementierung von GenAI muss sich in einer AWS Region befinden, für die Amazon Bedrock aktiviert ist. "Zeigen Sie die Liste der unterstützten Regionen an"

5. Füllen Sie die Einträge im Abschnitt **Infrastruktureinstellungen** aus:
   a. **Tags**: Geben Sie beliebige Tag-Schlüssel/Wert-Paare ein, die Sie auf alle AWS-Ressourcen anwenden möchten, die Teil dieser Bereitstellung sind. Diese Tags sind in der AWS-Managementkonsole und im Infrastrukturinformationsbereich innerhalb von Workload Factory sichtbar und können Ihnen dabei helfen, den Überblick über die Workload Factory-Ressourcen zu behalten.
6. Füllen Sie den Abschnitt **Konnektivität** aus:
   a. **Schlüsselpaar**: Wählen Sie ein Schlüsselpaar aus, mit dem Sie eine sichere Verbindung zur NetApp GenAI Engine-Instanz herstellen können.
7. Füllen Sie den Abschnitt **AI Engine** aus:
   a. **Instanzname**: Wählen Sie optional **Instanznamen definieren** aus und geben Sie einen benutzerdefinierten Namen für die KI-Engine-Instanz ein. Der Instanzname wird in der AWS-Managementkonsole und im Infrastrukturinformationsbereich innerhalb von Workload Factory angezeigt und kann Ihnen dabei helfen, den Überblick über die Workload Factory-Ressourcen zu behalten.
8. Wählen Sie **Deploy**, um die Bereitstellung zu starten.

> Wenn die Bereitstellung mit einem Fehler bei den Anmeldeinformationen fehlschlägt, erhalten Sie weitere Fehlerdetails, indem Sie die Hyperlinks in der Fehlermeldung auswählen. Sie sehen eine Liste mit fehlenden oder blockierten Berechtigungen sowie eine Liste mit Berechtigungen, die der GenAI-Workload benötigt, um die GenAI-Infrastruktur bereitstellen zu können.

Ergebnis

Workload Factory beginnt mit der Bereitstellung der Chatbot-Infrastruktur. Dieser Vorgang kann bis zu 10 Minuten dauern.

Während des Bereitstellungsprozesses werden die folgenden Elemente eingerichtet:

- Das Netzwerk wird zusammen mit den privaten Endpunkten eingerichtet.
- Die IAM-Rolle, das Instanzprofil und die Sicherheitsgruppe werden erstellt.
- Die Virtual Machine-Instanz für die GenAI-Engine wird bereitgestellt.
- Amazon Bedrock ist so konfiguriert, dass Protokolle an Amazon CloudWatch Logs gesendet werden, wobei eine Protokollgruppe mit dem Präfix verwendet `/aws/bedrock/`wird.
- Die GenAI-Engine ist so konfiguriert, dass sie Protokolle an Amazon CloudWatch Logs sendet, wobei eine Protokollgruppe mit dem Namen `/netapp/wlmai/<tenancyAccountId>/randomId` , Wo `<tenancyAccountID>` ist die "NetApp -Konsolen-Konto-ID" für den aktuellen Benutzer.

Erstellen Sie einen NetApp Connector für Amazon Q Business

Nachdem Sie die KI-Infrastruktur bereitgestellt und die Datenquellen identifiziert haben, die Sie aus Ihren FSx für ONTAP-Datenspeichern verwenden werden, können Sie einen NetApp Connector für Amazon Q Business definieren.

Stellen Sie sicher, dass Ihre Umgebung den für Amazon Q Business erfüllt"Anforderungen", bevor Sie fortfahren.

Über diese Aufgabe

Datenquellen Ihrer Organisation können personenbezogene Daten (PII) enthalten. Um diese vertraulichen Informationen zu schützen, können Sie beim Definieren eines Connectors *Datenleitplanken* aktivieren. Datenleitplanken, unterstützt durch NetApp Data Classification, identifizieren und maskieren personenbezogene Daten und machen sie so unzugänglich und unwiederbringlich.

"Erfahren Sie mehr über die NetApp Datenklassifizierung" .

NetApp Workload Factory für GenAI maskiert keine sensiblen persönlichen Informationen (SPII). Siehe"Arten sensibler personenbezogener Daten" für weitere Informationen zu dieser Art von Daten.

Datenleitplanken können jederzeit aktiviert oder deaktiviert werden. Wenn Sie die Aktivierung von Datenleitplanken ändern, scannt Workload Factory die gesamte Datenquelle von Grund auf, was Kosten verursachen kann.

Definieren Sie einen Konnektor

Erstellen Sie einen NetApp-Connector für Amazon Q Business. Der Connector ermöglicht die API- und Datenquellenkommunikation zwischen GenAI und Amazon Q Business.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie im Menü „Wissensdatenbanken und Konnektoren“ das Dropdown-Menü „Neu erstellen“ und wählen Sie „Amazon Q Business Connector“ aus.
4. Konfigurieren Sie auf der Seite Konnektor definieren die Verbindungseinstellungen:
   a. **Name**: Geben Sie den Namen ein, den Sie für den Konnektor verwenden möchten.
   b. **Beschreibung**: Geben Sie eine detaillierte Beschreibung für den Konnektor ein.
   c. **Amazon Q**: Die Region und der Anwendungsname für die Amazon Q Business Instanz, die Sie integrieren möchten.
   d. **Datenleitplanken**: Wählen Sie, ob Sie Datenleitplanken aktivieren oder deaktivieren möchten. "Erfahren Sie mehr über Datenleitplanken, unterstützt durch NetApp Data Classification" .

      Die folgenden Voraussetzungen müssen erfüllt sein, um die Data Guardrails zu aktivieren.

      - Für die Kommunikation mit NetApp Data Classification ist ein Servicekonto erforderlich. Für die Erstellung eines Dienstkontos müssen Sie in Ihrem NetApp Console-Mandantenkonto über die Rolle „Organisationsadministrator“ verfügen. Ein Mitglied mit der Rolle „Organisationsadministrator“ kann alle Aktionen in der NetApp Konsole ausführen."Erfahren Sie, wie Sie einem Mitglied in der NetApp -Konsole eine Rolle hinzufügen"
      - Die KI-Engine muss Zugriff auf die"NetApp Console API-Endpunkt" .
      - Sie müssen Folgendes tun, wie in"Dokumentation zur NetApp -Datenklassifizierung" :
        A. Erstellen eines Konsolenagenten
        B. Stellen Sie sicher, dass Ihre Umgebung die Voraussetzungen erfüllt

C. Bereitstellen der NetApp Datenklassifizierung

> Wenn Sie die Funktion „Data Guardrails“ aktivieren, verarbeitet GenAI die Dateien .txt, .md, .csv, .docx und .pdf, indem nur Klartext (ohne eingebetteten Bild- oder Medientext) aufgenommen und private oder sensible Daten maskiert werden. Alle anderen Dateitypen werden normal verarbeitet, ohne dass private oder sensible Daten maskiert werden.

e. **FSx für ONTAP Dateisystem**: Wenn Sie einen neuen NetApp Connector für Amazon Q Business definieren, erstellt Workload Factory ein neues Amazon FSx for NetApp ONTAP -Volume zum Speichern der Connector-Informationen. Wählen Sie ein vorhandenes Dateisystem und eine SVM (auch Speicher-VM genannt), auf der das neue Volume erstellt wird.

f. **Snapshot-Richtlinie**: Wählen Sie eine Snapshot-Richtlinie aus der Liste der vorhandenen Richtlinien aus, die im Workload Factory-Speicherinventar definiert sind. GenAI erstellt automatisch wiederkehrende Snapshots des Volumes, in denen die Connector-Informationen gespeichert sind, und zwar in einer Häufigkeit, die auf der von Ihnen ausgewählten Snapshot-Richtlinie basiert.

   Wenn die benötigte Snapshot-Richtlinie nicht vorhanden ist, können Sie "Erstellen einer Snapshot-Richtlinie" auf der Storage-VM, die das Volume enthält, arbeiten.

5. Wählen Sie **Create Connector**, um Amazon Q Business mit GenAI zu integrieren.

   Während der Verbindungserstellung wird eine Fortschrittsanzeige angezeigt.

   Nachdem der Connector erstellt wurde, haben Sie die Möglichkeit, dem Connector eine Datenquelle hinzuzufügen, sodass Amazon Q Business Ihre Daten aufnimmt und zu seinem Index hinzufügt. Wir empfehlen Ihnen, **Add Data source** auszuwählen und jetzt eine oder mehrere Datenquellen hinzuzufügen.

Fügen Sie dem Connector Datenquellen hinzu

Sie können eine oder mehrere Datenquellen hinzufügen, um den Amazon Q Business Index mit den Daten Ihres Unternehmens zu füllen.

Über diese Aufgabe

- Die maximale Anzahl unterstützter Datenquellen beträgt 10.
- In "Amazon Q Business-Dokumentation" finden Sie spezifische Serviceeinschränkungen des Amazon Q Business Index.

Schritte

1. Nachdem Sie **Datenquelle hinzufügen** ausgewählt haben, wird die Seite **Dateisystem auswählen** angezeigt.
2. **Wählen Sie ein Dateisystem**: Wählen Sie das FSX für ONTAP Dateisystem, in dem sich Ihre Datenquelldateien befinden, und wählen Sie **Weiter**.
3. **Wählen Sie ein Volume**: Wählen Sie das Volume aus, auf dem sich Ihre Quelldateien befinden, und wählen Sie **Weiter**.

   Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, einschließlich Domäne, IP-Adresse, Benutzername und Passwort.

4. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

5. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:
   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.
     - Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

     Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

6. Im Abschnitt **permission aware**, der nur verfügbar ist, wenn sich die von Ihnen ausgewählte Datenquelle auf einem Volume befindet, das das SMB-Protokoll verwendet, können Sie permission-aware Antworten aktivieren oder deaktivieren:
   - **Enabled**: Benutzer des Chatbot, die auf diesen Konnektor zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie Zugriff haben.
   - **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.

     Active Directory-Gruppenberechtigungen werden für Datenquellen des Amazon Q Business Connectors nicht unterstützt.

7. Wählen Sie **Hinzufügen**, um diese Datenquelle zum Amazon Q Business Connector hinzuzufügen.

Ergebnis

Die Datenquelle ist in den Amazon Q Business Index eingebettet. Der Status ändert sich von „Einbetten“ in „eingebettet“, wenn die Datenquelle vollständig eingebettet ist.

Nachdem Sie dem Connector eine einzelne Datenquelle hinzugefügt haben, können Sie diese in der Chatbot-Umgebung von Amazon Q Business testen und alle erforderlichen Änderungen vornehmen, bevor Sie den Dienst Ihren Benutzern zur Verfügung stellen. Sie können auch die gleichen Schritte ausführen, um dem Connector zusätzliche Datenquellen hinzuzufügen.

Administration und Überwachung

Management der GenAI-Infrastruktur

Sie können Details zu Ihrer bereitgestellten GenAI RAG-Infrastruktur anzeigen oder die Chatbot-Infrastruktur entfernen, wenn Sie sie nicht mehr benötigen.

Zeigen Sie Informationen zur Infrastruktur an

Sie können Informationen über die Chatbot-Infrastruktur anzeigen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie das Menü **Infrastruktur**.
4. Hier finden Sie Informationen zur Infrastruktur mit Details zu den folgenden Komponenten:
   - AWS-Einstellungen
   - Einstellungen der Infrastruktur
   - Die KI-Engine
   - Die Vektordatenbank

Entfernen Sie die Infrastruktur

Wenn Sie die Chatbot-Infrastruktur, die Sie für einen oder mehrere Chatbots bereitgestellt haben, nicht mehr benötigen, können Sie sie aus Workload Factory entfernen.

> Alle Chatbots, die in dieser Infrastruktur bereitgestellt wurden, werden deaktiviert und der gesamte Chatverlauf wird gelöscht.

Dieser Vorgang entfernt nur die Links zur KI-Infrastruktur von Workload Factory; er entfernt nicht alle Komponenten von AWS. Sie müssen die folgenden Infrastrukturkomponenten manuell aus AWS löschen:

- Die VM-Instanz
- Private Endpunkte
- Das Volumen des FSX für ONTAP-Filesystems mit den AI-Datenbanken
- Die Rolle des IAM
- Richtlinie
- Die Sicherheitsgruppe

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie das Menü **Infrastruktur**.
4. Wählen Sie ••• und wählen Sie **Chatbot-Infrastruktur entfernen**.

5. Bestätigen Sie, dass Sie die Infrastruktur löschen möchten, und wählen Sie **Entfernen**.

Ergebnis

Die Chatbot-Infrastrukturkomponenten werden aus Workload Factory entfernt.

Management von GenAI Knowledge Base

Nachdem Sie eine Wissensdatenbank erstellt haben, können Sie die Details der Wissensdatenbank anzeigen, die Wissensdatenbank ändern, zusätzliche Datenquellen integrieren oder die Wissensdatenbank löschen.

Informationen zu einer Wissensdatenbank anzeigen

Sie können Informationen zu den Einstellungen für eine Wissensdatenbank und die integrierte Datenquelle anzeigen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie die Wissensdatenbank aus, die Sie anzeigen möchten.

   Falls definiert, werden die derzeit verwendeten Gesprächseinstiege im rechten Fensterbereich angezeigt.

4. Um Details zur Wissensdatenbank anzuzeigen, wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

   Über das Menü **Aktionen** können Sie die Wissensdatenbank verwalten, wenn Sie Änderungen vornehmen möchten.

Wissensdatenbank bearbeiten

Sie können eine Wissensdatenbank aktualisieren, indem Sie einige Einstellungen ändern, oder Sie können Datenquellen hinzufügen oder entfernen.

Jedes Mal, wenn Sie Datenquellen in der Wissensdatenbank hinzufügen, ändern oder entfernen, müssen Sie die Datenquelle synchronisieren, damit sie mit der Wissensdatenbank neu indiziert wird. Die Synchronisierung ist inkrementell, daher verarbeitet Amazon Bedrock nur die Objekte in Ihrem FSX für ONTAP-Volume, die seit der letzten Synchronisierung hinzugefügt, geändert oder gelöscht wurden.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, die Sie aktualisieren möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

5. Wählen Sie das Menü **Aktionen** und dann **Wissensdatenbank bearbeiten**.
6. Auf der Seite „Wissensdatenbank bearbeiten“ können Sie den Namen und die Beschreibung der Wissensdatenbank, das Einbettungsmodell, das Chatmodell und die Funktionsaktivierung ändern, auswählen, ob Gesprächsstarter automatisch oder manuell erstellt werden, und die Snapshot-Richtlinie für das Volume verwenden, das die Wissensdatenbank enthält.

   Wenn Sie den manuellen Modus für Gesprächseinstieg verwenden, können Sie hier auch den Gesprächseinstieg ändern.

   > Jeder Knowledge Base Scan, der Einbettung, Kosten beinhaltet. Wenn die Datenguardrails nach Erstellung einer Wissensdatenbank aktiviert werden, wird die Wissensdatenbank erneut gescannt und verursacht Kosten. Ebenso scannt GenAI die zugehörigen Datenquellen erneut, wenn Sie die Chatmodelle ändern (Kosten entstehen).

7. Wählen Sie **Speichern**, nachdem Sie Ihre Änderungen vorgenommen haben.

Schützen Sie eine Wissensdatenbank mit Snapshots

Sie können Ihre Knowledge Base-Daten schützen, indem Sie Snapshots Ihrer Knowledge Base-Volumes erstellen und wiederherstellen. Sie können jederzeit von einem Snapshot wiederherstellen, um die vorherige Version der Wissensdatenbank wiederherzustellen.

Snapshots können schneller und speichereffizienter als Backups sein und ermöglichen es Ihnen, jede Wissensdatenbank mit einer anderen Sicherungsrichtlinie zu schützen. Einige der Szenarien, in denen Snapshots nützlich sein können, sind:

- Versehentlicher Datenverlust oder -Beschädigung
- Wiederherstellung nach fehlerhaften Daten, die in die Wissensdatenbank aufgenommen werden
- Testen verschiedener Datenquellen oder Chunking-Strategien und schnelles Zurücksetzen nach Abschluss des Tests

Erstellen Sie einen Snapshot eines Knowledge-Base-Volumes

Sie können den Status einer Wissensdatenbank speichern, indem Sie einen manuellen Snapshot des Knowledge Base-Volumes erstellen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Wissensdatenbanken die Wissensdatenbank aus, die Sie schützen möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

5. Wählen Sie das Menü **actions** und dann **Snapshot > Neuen Snapshot erstellen**.
6. Wählen Sie optional **Snapshot-Namen definieren** und geben Sie einen benutzerdefinierten Namen für den Snapshot ein.

   Durch das Definieren eines benutzerdefinierten Namens können Sie den Inhalt eines Snapshots besser

bestimmen, wenn Sie ihn in Zukunft wiederherstellen müssen.

7. Wählen Sie **Erstellen**.

   Es wird eine Momentaufnahme der Wissensdatenbank erstellt.

Stellen Sie einen Snapshot eines Knowledge-Base-Volumes wieder her

Sie können jederzeit einen manuellen oder geplanten Snapshot eines Knowledge-Base-Volumes wiederherstellen.

> Sie können einen Snapshot nicht mithilfe der Benutzeroberfläche Generative AI-Workloads wiederherstellen, wenn die auf dem Volume gespeicherte Datenbank beschädigt ist oder gelöscht wurde. Als Workaround können Sie den Snapshot mithilfe des auf dem ONTAP-Cluster wiederherstellen "CLI VON ONTAP", auf dem das Volume gehostet wird.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, die Sie wiederherstellen möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

5. Wählen Sie das Menü **actions** und dann **Snapshot > Snapshot wiederherstellen**.

   Das Dialogfeld Snapshot-Auswahl wird angezeigt, in dem Sie eine Liste der Snapshots sehen können, die für diese Wissensdatenbank erstellt wurden.

6. (Optional) Wählen Sie die Option **laufende und geplante Scans anhalten nach dem Wiederherstellen des Snapshots** aus, wenn geplante und aktuell ausgeführte Datenquellenscans nach der Wiederherstellung des Snapshots fortgesetzt werden sollen.

   Diese Option ist standardmäßig aktiviert, um sicherzustellen, dass ein Scan nicht stattfindet, während sich die Wissensdatenbank in einem teilweise wiederhergestellten Zustand befindet, oder dass ein Scan keine frisch wiederhergestellte Wissensdatenbank mit älteren Daten aktualisiert.

7. Wählen Sie den zu wiederherstellenden Snapshot aus der Liste aus.
8. Wählen Sie **Wiederherstellen**.

Klonen einer Wissensdatenbank

Sie können eine neue Wissensdatenbank aus einem Momentaufnahme der Wissensdatenbank erstellen. Dies ist nützlich, wenn die ursprüngliche Wissensdatenbank beschädigt ist oder verloren geht.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, die Sie wiederherstellen

möchten.

4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

5. Wählen Sie das Menü **actions** und dann **Snapshot > Clone Knowledge Base**.

   Das Dialogfeld Klonen wird angezeigt.

6. Deaktivieren Sie optional die Option **Anhalten der laufenden und geplanten Scans nach dem Klonen des Snapshots**, wenn geplante und aktuell ausgeführte Datenquellen-Scans nach dem Klonen des Snapshots fortgesetzt werden sollen.

   Diese Option ist standardmäßig aktiviert, um sicherzustellen, dass ein Scan nicht stattfindet, während sich die Wissensdatenbank in einem teilweise wiederhergestellten Zustand befindet, oder dass ein Scan keine frisch wiederhergestellte Wissensdatenbank mit älteren Daten aktualisiert.

7. Wählen Sie den zu klonenden Snapshot aus der Liste aus.
8. Wählen Sie **Weiter**.
9. Geben Sie einen Namen für die neue Wissensdatenbank ein.
10. Wählen Sie einen Dateisystem-SVM- und Volume-Namen für die neue Wissensdatenbank aus.
11. Wählen Sie **Clone**.

Fügen Sie einer Wissensdatenbank zusätzliche Datenquellen hinzu

Sie können zusätzliche Datenquellen in Ihre Wissensdatenbank einbetten, um diese mit zusätzlichen Unternehmensdaten zu füllen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, in die Sie die Datenquelle hinzufügen möchten.
4. Wählen Sie ••• **Datenquelle hinzufügen**.
5. Wählen Sie den Typ der Datenquelle aus, die Sie hinzufügen möchten:
   - FSx für ONTAP-Dateisystem hinzufügen (Dateien aus einem vorhandenen FSx für ONTAP-Volume verwenden)
   - Dateisystem hinzufügen (Dateien aus einer generischen SMB- oder NFS-Freigabe verwenden)

Fügen Sie ein FSx für ONTAP-Dateisystem hinzu

1. **Wählen Sie ein Dateisystem**: Wählen Sie das FSX für ONTAP Dateisystem, in dem sich Ihre Datenquelldateien befinden, und wählen Sie **Weiter**.
2. **Wählen Sie ein Volume**: Wählen Sie das Volume aus, auf dem sich Ihre Quelldateien befinden, und wählen Sie **Weiter**.

   Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, einschließlich Domäne, IP-Adresse, Benutzername und Passwort.

3. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.
4. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

       Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

       Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

     Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

     Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

   - Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien

basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

> Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

5. Im Abschnitt **permission aware**, der nur verfügbar ist, wenn sich die von Ihnen ausgewählte Datenquelle auf einem Volume befindet, das das SMB-Protokoll verwendet, können Sie permission-aware Antworten aktivieren oder deaktivieren:
   - **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
   - **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.
6. Wählen Sie **Hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen NFS-Dateisystems

1. **Wählen Sie ein Dateisystem**: Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden, wählen Sie das NFS-Protokoll für die Netzwerkfreigabe und klicken Sie auf **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

> In manchen Fällen müssen Sie den NFS-Exportnamen manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können den gesamten Export oder nur bestimmte Ordner auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:
   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

     > Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder

wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

- Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen SMB-Dateisystems

1. **Dateisystem auswählen**:
   a. Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden.
   b. Wählen Sie das SMB-Protokoll für die Netzwerkfreigabe.
   c. Geben Sie die Active Directory-Informationen ein, darunter Domäne, IP-Adresse, Benutzername und Kennwort.
   d. Wählen Sie **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

In manchen Fällen müssen Sie den Namen der SMB-Freigabe manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können die gesamte Freigabe oder nur bestimmte Ordner darin auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:

- **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
- **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

  Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

◦ **Berechtigungsbewusst**: Aktivieren oder deaktivieren Sie berechtigungsbewusste Antworten:

- **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
- **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.

◦ **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:

- Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

  Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

◦ Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Ergebnis

Die Datenquelle ist in Ihre Wissensdatenbank integriert.

Synchronisieren Sie Ihre Datenquellen mit einer Wissensdatenbank

Datenquellen werden automatisch einmal täglich mit der zugehörigen Wissensdatenbank synchronisiert, sodass Änderungen der Datenquelle im Chatbot berücksichtigt werden. Wenn Sie Änderungen an einer Ihrer Datenquellen vornehmen und die Daten sofort synchronisieren möchten, können Sie eine On-Demand-Synchronisierung durchführen.

Die Synchronisierung ist inkrementell, daher verarbeitet Amazon Bedrock nur die Objekte in Ihren Datenquellen, die seit der letzten Synchronisierung hinzugefügt, geändert oder gelöscht wurden.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Knowledge Base Inventory die Wissensdatenbank aus, die Sie synchronisieren möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.
5. Wählen Sie das Menü **Aktionen** und dann **Jetzt scannen**.

   Sie sehen eine Meldung, dass Ihre Datenquellen gescannt werden, und eine abschließende Meldung, wenn der Scan abgeschlossen ist.

Ergebnis

Die Wissensdatenbank wird mit den angehängten Datenquellen synchronisiert und jeder aktive Chatbot verwendet die neuesten Informationen aus Ihren Datenquellen.

Eine geplante Synchronisierung anhalten oder fortsetzen

Wenn Sie die nächste Synchronisierung (Scan) der Datenquellen anhalten oder fortsetzen möchten, können Sie dies jederzeit tun. Möglicherweise müssen Sie die nächste geplante Synchronisierung anhalten, wenn Sie Änderungen an einer Datenquelle vornehmen und die Synchronisierung während des Änderungsfensters nicht durchführen möchten.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie im Menü „Wissensdatenbanken und Konnektoren“ die Wissensdatenbank aus, für die Sie Scans anhalten oder fortsetzen möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.
5. Wählen Sie das Menü **Aktionen** und dann **Scan > geplante Messung anhalten** oder **Scan > geplante Messung fortsetzen**.

   Es wird eine Meldung angezeigt, dass der nächste geplante Scan entweder angehalten oder fortgesetzt wurde.

Bewerten Sie Chatmodelle, bevor Sie eine Wissensdatenbank erstellen

Sie können die verfügbaren grundlegenden Chatmodelle bewerten, bevor Sie eine Wissensdatenbank

erstellen, damit Sie sehen können, welches Modell für Ihre Implementierung am besten geeignet ist. Da der Modellsupport je nach AWS-Region variiert, finden Sie unter "Dieser AWS Dokumentationsseite" Informationen dazu, welche Modelle in den Regionen verwendet werden können, in denen Sie Ihre Knowledge Base bereitstellen möchten.

Diese Funktion ist nur verfügbar, wenn keine Wissensdatenbanken erstellt wurden — wenn auf der Bestandsseite der Wissensdatenbanken keine Wissensdatenbanken vorhanden sind.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Auf der Bestandsseite der Wissensbasen sehen Sie die Option, das Chatmodell auf der rechten Seite der Seite für den Chatbot auszuwählen.
4. Wählen Sie das Chatmodell aus der Liste aus, und geben Sie eine Reihe von Fragen in den Eingabebereich ein, um zu sehen, wie der Chatbot reagiert.
5. Testen Sie mehrere Modelle, um herauszufinden, welches Modell sich am besten für Ihre Implementierung eignet.

Ergebnis

Verwenden Sie dieses Chatmodell, wenn Sie Ihre Wissensdatenbank erstellen.

Heben Sie die Veröffentlichung Ihrer Wissensdatenbank auf

Nachdem Sie Ihre Wissensdatenbank veröffentlicht haben, damit sie in eine Chatbot-Anwendung integriert werden kann, können Sie die Veröffentlichung aufheben, wenn Sie den Zugriff der Chatbot-Anwendung auf die Wissensdatenbank deaktivieren möchten.

Durch das Aufheben der Veröffentlichung der Wissensdatenbank werden alle Chat-Anwendungen nicht mehr funktionsfähig. Der eindeutige API-Endpunkt, auf den die Wissensdatenbank zugegriffen werden konnte, ist deaktiviert.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Wissensdatenbanken die Wissensdatenbank aus, die Sie für die Veröffentlichung aufheben möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus und die Liste aller eingebetteten Datenquellen angezeigt.

5. Wählen Sie das Menü **actions** und dann **Unpublish**.

Ergebnis

Die Wissensdatenbank ist deaktiviert und kann nicht mehr von einer Chatbot-Anwendung aufgerufen werden.

Löschen einer Wissensdatenbank

Wenn Sie eine Wissensdatenbank nicht mehr benötigen, können Sie sie löschen. Wenn Sie eine Wissensdatenbank löschen, wird sie aus Workload Factory entfernt und das Volume, das die

Wissensdatenbank enthält, wird gelöscht. Alle Anwendungen oder Chatbots, die die Wissensdatenbank verwenden, funktionieren nicht mehr. Das Löschen einer Wissensdatenbank kann nicht rückgängig gemacht werden.

Wenn Sie eine Wissensdatenbank löschen, sollten Sie auch die Zuordnung der Wissensdatenbank zu allen Agenten aufheben, mit denen sie verknüpft ist, um alle Ressourcen, die der Wissensdatenbank zugeordnet sind, vollständig zu löschen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Wissensdatenbanken die Wissensdatenbank aus, die Sie löschen möchten.
4. Wählen Sie ••• **Wissensdatenbank verwalten**.
5. Wählen Sie das Menü **Aktionen** und dann **Wissensdatenbank löschen**.
6. Bestätigen Sie im Dialogfeld Wissensdatenbank löschen, dass Sie löschen möchten, und wählen Sie **Löschen**.

Ergebnis

Die Wissensdatenbank wird aus Workload Factory entfernt und das zugehörige Volume wird gelöscht.

Verwalten von Amazon Q Business Connectors

Nachdem Sie einen Connector für Amazon Q Business erstellt haben, können Sie die Verbindungsdetails anzeigen, den Connector ändern, zusätzliche Datenquellen integrieren oder den Connector löschen.

Anzeigen von Informationen zu einem Konnektor

Sie können Informationen zu den Einstellungen für einen Konnektor und zu den integrierten Datenquellen anzeigen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Inventarseite „Wissensdatenbanken und Konnektoren“ den Konnektor aus, den Sie anzeigen möchten.
4. Wählen Sie zum Anzeigen von Verbindungsdetails •••**Connector verwalten** aus, und wählen Sie diese Option aus.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

   Über das Menü **actions** können Sie den Konnektor verwalten, wenn Sie Änderungen vornehmen möchten.

Bearbeiten Sie einen Konnektor

Sie können einen Konnektor aktualisieren, indem Sie einige Einstellungen ändern, oder Sie können

Datenquellen hinzufügen oder entfernen.

Jedes Mal, wenn Sie Datenquellen vom Connector hinzufügen, ändern oder entfernen, muss GenAI die Datenquellinformationen an Amazon Q Business senden, damit sie neu indiziert werden. Die Synchronisierung ist inkrementell, daher verarbeitet Amazon Q Business nur die Objekte in Ihrem FSX für ONTAP-Volume, die seit der letzten Synchronisierung hinzugefügt, geändert oder gelöscht wurden.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Knowledge Basen & Connectors den Konnektor aus, den Sie aktualisieren möchten.
4. Wählen Sie **Connector verwalten**.

   Auf dieser Seite werden der veröffentlichte Status, der Einbettungsstatus der Datenquellen, der Einbettungsmodus, die Liste aller eingebetteten Datenquellen und vieles mehr angezeigt.

5. Wählen Sie das Menü **actions** und dann **Connector bearbeiten**.
6. Auf der Seite Connector bearbeiten können Sie den Anschlussnamen, die Beschreibung, das Einbettungsmodell, die Aktivierung von Data Guardrails und die Snapshot-Richtlinie ändern, die für das Volume verwendet wird, das den Konnektor enthält.

   Jeder Scan der Datenquelle, einschließlich Einbettung, verursacht Kosten. Wenn Sie nach der Erstellung eines Konnektors die Data Guardrails aktivieren, wird die Datenquelle erneut gescannt und verursacht Kosten.

7. Wählen Sie **Speichern**, nachdem Sie Änderungen vorgenommen haben.

Fügen Sie einem Konnektor zusätzliche Datenquellen hinzu

Sie können zusätzliche Datenquellen in Ihren Connector einbetten, um sie mit zusätzlichen Unternehmensdaten zu füllen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Knowledge Basen & Connectors den Konnektor aus, zu dem Sie die Datenquelle hinzufügen möchten.
4. Wählen Sie **Datenquelle hinzufügen**.
5. Wählen Sie den Typ der Datenquelle aus, die Sie hinzufügen möchten:
   - FSx für ONTAP-Dateisystem hinzufügen (Dateien aus einem vorhandenen FSx für ONTAP-Volume verwenden)
   - Dateisystem hinzufügen (Dateien aus einer generischen SMB- oder NFS-Freigabe verwenden)

Fügen Sie ein FSx für ONTAP-Dateisystem hinzu

1. **Wählen Sie ein Dateisystem**: Wählen Sie das FSX für ONTAP Dateisystem, in dem sich Ihre Datenquelldateien befinden, und wählen Sie **Weiter**.
2. **Wählen Sie ein Volume**: Wählen Sie das Volume aus, auf dem sich Ihre Quelldateien befinden, und wählen Sie **Weiter**.

   Bei der Auswahl der mit dem SMB-Protokoll gespeicherten Dateien müssen Sie die Active Directory-Informationen eingeben, einschließlich Domäne, IP-Adresse, Benutzername und Passwort.

3. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.
4. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

       Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

       Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

     Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

     Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

   - Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien

basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

> Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

5. Im Abschnitt **permission aware**, der nur verfügbar ist, wenn sich die von Ihnen ausgewählte Datenquelle auf einem Volume befindet, das das SMB-Protokoll verwendet, können Sie permission-aware Antworten aktivieren oder deaktivieren:
   - **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
   - **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.
6. Wählen Sie **Hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen NFS-Dateisystems

1. **Wählen Sie ein Dateisystem**: Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden, wählen Sie das NFS-Protokoll für die Netzwerkfreigabe und klicken Sie auf **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

> In manchen Fällen müssen Sie den NFS-Exportnamen manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können den gesamten Export oder nur bestimmte Ordner auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:
   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:
     - **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
     - **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

     > Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

   - **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:
     - Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder

wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

- Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Hinzufügen eines generischen SMB-Dateisystems

1. **Dateisystem auswählen**:
   a. Geben Sie die IP-Adresse oder den FQDN für den Dateisystem-Host ein, auf dem sich Ihre Datenquelldateien befinden.
   b. Wählen Sie das SMB-Protokoll für die Netzwerkfreigabe.
   c. Geben Sie die Active Directory-Informationen ein, darunter Domäne, IP-Adresse, Benutzername und Kennwort.
   d. Wählen Sie **Weiter**.
2. **Wählen Sie eine Datenquelle**: Wählen Sie den Speicherort der Datenquelle basierend darauf, wo Sie die Dateien gespeichert haben. Dies kann ein ganzes Volume sein, oder nur ein bestimmter Ordner oder Unterordner im Volume, und wählen Sie **Weiter**.

In manchen Fällen müssen Sie den Namen der SMB-Freigabe manuell eingeben und „Verzeichnisse abrufen“ auswählen, um die verfügbaren Verzeichnisse anzuzeigen. Sie können die gesamte Freigabe oder nur bestimmte Ordner darin auswählen.

3. **Konfigurationen**: Konfigurieren Sie, wie die Datenquelle Informationen aus Ihren Dateien einliest und welche Dateien sie in Scans enthält:

   - **Datenquelle definieren**: Definieren Sie im Abschnitt **Chunking-Strategie**, wie die GenAI-Engine den Inhalt der Datenquelle in Blöcke teilt, wenn die Datenquelle in eine Wissensdatenbank integriert ist. Sie können eine der folgenden Strategien wählen:

- **Multi-Sentence Chunking**: Organisiert Informationen aus Ihrer Datenquelle in Satzbausteine. Sie können festlegen, wie viele Sätze jedes Stück ausmachen (bis zu 100).
- **Überlappendes Chunking**: Organisiert Informationen aus Ihrer Datenquelle in zeichendefinierte Blöcke, die benachbarte Blöcke überlappen können. Sie können die Größe jedes Chunk in Zeichen auswählen und festlegen, wie viel sich jeder Chunk mit benachbarten Chunks überschneidet. Sie können eine Chunk-Größe von 50 bis 3000 Zeichen und einen Überlappungsprozentsatz von 1 bis 99 % konfigurieren.

  Die Auswahl eines hohen Prozentsatzes für Überschneidungen kann den Speicherbedarf erheblich erhöhen, da die Abrufgenauigkeit nur geringfügig verbessert wird.

◦ **Berechtigungsbewusst**: Aktivieren oder deaktivieren Sie berechtigungsbewusste Antworten:

- **Enabled**: Benutzer des Chatbot, die auf diese Wissensdatenbank zugreifen, erhalten nur Antworten auf Abfragen aus Datenquellen, auf die sie zugreifen können.
- **Disabled**: Benutzer des Chatbot erhalten Antworten über Inhalte aus allen integrierten Datenquellen.

◦ **Dateifilterung**: Konfigurieren Sie, welche Dateien in Scans enthalten sind:

- Wählen Sie im Abschnitt **Unterstützung für Dateitypen** entweder alle Dateitypen aus oder wählen Sie einzelne Dateitypen aus, um sie in die Datenquellprüfungen einzubeziehen.

  Wenn Sie Bilder oder PDF-Dateien einschließen, analysiert NetApp Workload Factory für GenAI den Text in den Bildern (einschließlich Bildern in PDF-Dokumenten), was höhere Kosten verursacht.

Beim Einbeziehen von Textdaten aus Bildern ist GenAI nicht in der Lage, personenbezogene Daten (PII) aus dem Bild zu maskieren, wenn die gescannten Textdaten aus Ihrer Umgebung an AWS gesendet werden. Sobald die Daten jedoch gespeichert sind, werden alle PII in der GenAI-Datenbank maskiert.

Ihre Wahl, Bilddateien in Scans aufzunehmen, hängt mit dem Chatmodell der Wissensdatenbank zusammen. Wenn Sie Bilddateien in Scans aufnehmen, muss das Chatmodell Bilder unterstützen. Wenn hier Bilddateitypen ausgewählt sind, können Sie die Wissensdatenbank nicht auf ein Chatmodell umschalten, das keine Bilddateien unterstützt.

◦ Wählen Sie im Abschnitt **Dateiänderungszeitfilter** die Option, die Aufnahme von Dateien basierend auf ihrer Änderungszeit zu aktivieren oder zu deaktivieren. Wenn Sie die Zeitfilterung für die Änderung aktivieren, wählen Sie einen Datumsbereich aus der Liste aus.

Wenn Sie Dateien auf Basis eines Änderungsdatums einbeziehen, werden die Dateien aus dem periodischen Scan ausgeschlossen, und die Datenquelle enthält diese Dateien nicht, sobald der Datumsbereich nicht erfüllt ist (die Dateien wurden nicht innerhalb des von Ihnen angegebenen Datumsbereichs geändert).

4. Wählen Sie **Datenquelle hinzufügen**, um diese Datenquelle zu Ihrer Wissensdatenbank hinzuzufügen.

Ergebnis

Die Datenquelle ist in Ihren Connector integriert.

Synchronisieren Sie Ihre Datenquellen mit einem Konnektor

Datenquellen werden automatisch einmal täglich mit dem zugehörigen Connector synchronisiert, sodass Änderungen an der Datenquelle in Amazon Q Business berücksichtigt werden. Wenn Sie Änderungen an einer Ihrer Datenquellen vornehmen und die Daten sofort synchronisieren (scannen) möchten, können Sie eine On-Demand-Synchronisierung durchführen.

Die Synchronisierung ist inkrementell, daher verarbeitet Amazon Q Business nur die Objekte in Ihren Datenquellen, die seit der letzten Synchronisierung hinzugefügt, geändert oder gelöscht wurden.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie im Menü „Wissensdatenbanken und Konnektoren“ den Konnektor aus, den Sie synchronisieren möchten.
4. Wählen Sie **Connector verwalten**.
5. Wählen Sie das Menü **Aktionen** und dann **Jetzt scannen**.

   Sie sehen eine Meldung, dass Ihre Datenquellen gescannt werden, und eine abschließende Meldung, wenn der Scan abgeschlossen ist.

Ergebnis

Der Connector wird mit den angehängten Datenquellen synchronisiert, und Amazon Q Business verwendet die neuesten Informationen aus Ihren Datenquellen.

Eine geplante Synchronisierung anhalten oder fortsetzen

Wenn Sie die nächste Synchronisierung (Scan) der Datenquellen anhalten oder fortsetzen möchten, können Sie dies jederzeit tun. Möglicherweise müssen Sie die nächste geplante Synchronisierung anhalten, wenn Sie Änderungen an einer Datenquelle vornehmen und die Synchronisierung während des Änderungsfensters nicht durchführen möchten.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Seite Connector Inventory den Anschluss aus, für den Sie die Scans anhalten oder fortsetzen möchten.
4. Wählen Sie **Connector verwalten**.
5. Wählen Sie das Menü **Aktionen** und dann **Scan > geplante Messung anhalten** oder **Scan > geplante Messung fortsetzen**.

   Es wird eine Meldung angezeigt, dass der nächste geplante Scan entweder angehalten oder fortgesetzt wurde.

Löschen Sie einen Konnektor

Wenn Sie einen Connector nicht mehr benötigen, können Sie ihn löschen. Wenn Sie einen Connector löschen,

wird er aus Workload Factory entfernt und das Volume, das den Connector enthält, wird gelöscht. Das Löschen eines Konnektors kann nicht rückgängig gemacht werden.

Wenn Sie einen Konnektor löschen, sollten Sie auch die Zuordnung des Connectors zu allen Agenten aufheben, mit denen er verknüpft ist, um alle Ressourcen, die dem Connector zugeordnet sind, vollständig zu löschen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Knowledge Basen & Connectors den Konnektor aus, den Sie löschen möchten.
4. Wählen Sie **Connector verwalten**.
5. Wählen Sie das Menü **actions** und dann **Delete Connector**.
6. Bestätigen Sie im Dialogfeld Connector löschen, dass Sie es löschen möchten, und wählen Sie **Löschen**.

Ergebnis

Der Connector wird aus Workload Factory entfernt und das zugehörige Volume wird gelöscht.

Management von GenAI Datenquellen

Nachdem Sie eine Wissensdatenbank oder einen Konnektor mit Datenquellen auf Ihrem FSX für ONTAP-Dateisystem erstellt haben, können Sie die Details der Datenquelle anzeigen, den Inhalt der Datenquelle aktualisieren oder ändern, die Einstellungen der Datenquelle bearbeiten oder die Datenquelle löschen.

Informationen zu einer Datenquelle anzeigen

Sie können Informationen über den Inhalt einer Datenquelle anzeigen und deren Einbettungsstatus mit der Wissensdatenbank oder dem Konnektor anzeigen. Da Datenquellen mit einer Wissensdatenbank oder einem Konnektor verknüpft sind, müssen Sie zunächst die Wissensdatenbank oder den Konnektor auswählen, bevor Sie die Details der Datenquelle anzeigen können.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden "Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie die Wissensdatenbank oder den Konnektor aus, in der sich die Datenquelle befindet, und wählen Sie dann **Wissensdatenbank verwalten** oder **Konnektor verwalten** aus.

   Im unteren Teil der Seite werden die zugehörigen Datenquellen aufgelistet.

4. Erweitern Sie jede Zeile, indem Sie die auswählen , um detaillierte Informationen zu jeder Datenquelle anzuzeigen, z. B. das FSX für ONTAP-Dateisystem, das Volume und den Pfad, in dem sich die Datenquelle befindet.

   Außerdem werden die einbettenden Informationen aufgelistet und es wird angegeben, ob diese Datenquelle derzeit in die Wissensdatenbank oder den Konnektor eingebettet ist.

Bearbeiten Sie die Einstellungen für die Datenquelle

Sie können Informationen zu einer Datenquelle bearbeiten, die Sie in eine Wissensdatenbank oder einen Konnektor integriert haben. Die meisten Informationen werden behoben, nachdem Sie eine Datenquelle hinzugefügt haben, aber Sie können Änderungen an einigen der Konfiguration vornehmen (z. B. Chunking-Definition oder Berechtigungssicherung).

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite Wissensdatenbanken die Wissensdatenbank aus, in der sich die Datenquelle befindet, und wählen Sie dann **Wissensdatenbank verwalten** aus ••• .

   Im unteren Teil der Seite werden die Datenquellen aufgelistet, die Teil dieser Wissensdatenbank sind.
4. Wählen Sie in der Zeile der Datenquelle, die Sie bearbeiten möchten, **Datenquelle bearbeiten** aus ••• .
5. Wählen Sie auf der Seite Datenquelle bearbeiten ⌄die Option aus, um die Zeile für die Chunk-Definition zu erweitern.
6. Aktualisieren Sie die Einstellungen für die Chunking-Strategie und -Konfiguration sowie Berechtigungssicherung (für SMB-Volumes), und wählen Sie **Speichern**.

Ergebnis

Die Einstellungen der Datenquelle werden aktualisiert und das KI-System synchronisiert die Datenquelle, sodass sie erneut mit der Wissensdatenbank indiziert wird.

Aktualisieren Sie den Inhalt einer vorhandenen Datenquelle

Sie können den Inhalt einer Datenquelle jederzeit ändern, um Ihre Unternehmensdaten hinzuzufügen oder zu aktualisieren. Wenn diese Datenquelle aktiv in einer Wissensdatenbank verwendet wird, müssen Sie die Datenquelle synchronisieren, damit sie mit der Wissensdatenbank neu indiziert wird. Die Synchronisierung ist inkrementell, daher verarbeitet Amazon Bedrock nur die Objekte in Ihrem FSX für ONTAP-Volume, die seit der letzten Synchronisierung hinzugefügt, geändert oder gelöscht wurden.

Datenquellen werden automatisch einmal täglich mit der Wissensdatenbank synchronisiert, sodass Änderungen der Datenquelle im Chatbot berücksichtigt werden. Wenn Sie Änderungen an einer Datenquelle vornehmen und die Daten sofort synchronisieren möchten, können Sie "Führen Sie eine On-Demand-Synchronisierung durch".

Löschen einer Datenquelle

Wenn Sie keine Datenquelle mehr benötigen, um Teil Ihrer Wissensdatenbank zu sein, können Sie sie löschen.

Schritte

1. Melden Sie sich bei Workload Factory mit einem der folgenden"Konsolenerfahrungen" .
2. Wählen Sie in der Kachel AI-Workloads **Deploy & manage** aus.
3. Wählen Sie auf der Bestandsseite der Wissensdatenbank die Wissensdatenbank aus, in der sich die Datenquelle befindet, und wählen Sie dann **Wissensdatenbank verwalten** aus ••• .

   Im unteren Teil der Seite werden die Datenquellen aufgelistet, die Teil dieser Wissensdatenbank sind.

4. Wählen Sie in der Zeile der Datenquelle, die Sie löschen möchten, **Datenquelle löschen** aus ••• .
5. Bestätigen Sie im Dialogfeld Datenquelle löschen, dass Sie es löschen möchten, und wählen Sie **Bestätigen**.

Ergebnis

Die Datenquelle wird aus der Wissensdatenbank entfernt, und das KI-System entfernt die indizierten Informationen zu dieser Datenquelle aus der Wissensdatenbank. Alle Informationen aus dieser Datenquelle werden Chatbots, die die Wissensdatenbank nutzen, nicht mehr zur Verfügung stehen.

Überwachen Sie Workload-Vorgänge mit Tracker in NetApp Workload Factory

Überwachen und verfolgen Sie die Ausführung von Workload-Vorgängen und überwachen Sie den Aufgabenfortschritt mit Tracker in NetApp Workload Factory.

Über diese Aufgabe

NetApp Workload Factory bietet Tracker, eine Überwachungsfunktion, mit der Sie den Fortschritt und Status von Workload-Vorgängen überwachen und verfolgen, Details zu Betriebsaufgaben und Unteraufgaben überprüfen und etwaige Probleme oder Fehler diagnostizieren können.

In Tracker sind mehrere Aktionen verfügbar. Sie können Jobs nach Zeitrahmen (letzte 24 Stunden, 7 Tage, 14 Tage oder 30 Tage), Workload, Status und Benutzer filtern, Jobs mithilfe der Suchfunktion suchen und die Jobtabelle als CSV-Datei herunterladen. Sie können Tracker jederzeit aktualisieren und einen fehlgeschlagenen Vorgang schnell wiederholen oder Parameter für einen fehlgeschlagenen Vorgang bearbeiten und den Vorgang erneut versuchen.

Tracker unterstützt je nach Betrieb zwei Überwachungsebenen. Jede Aufgabe, wie z. B. die Dateisystembereitstellung, zeigt die Aufgabenbeschreibung, den Status, die Startzeit, die Aufgabendauer, den Benutzer, die Region, die Proxyressource, die Aufgaben-ID und alle zugehörigen untergeordneten Aufgaben an. Sie können API-Antworten anzeigen, um zu verstehen, was während des Vorgangs passiert ist.

Verfolgung von Aufgabenebenen mit Beispielen

- Stufe 1 (Aufgabe): Verfolgt die Dateisystembereitstellung.
- Stufe 2 (Unteraufgabe): Verfolgt die Unteraufgaben im Zusammenhang mit der Dateisystembereitstellung.

Betriebsstatus

Der Betriebsstatus in Tracker ist wie folgt: *In Bearbeitung*, *Erfolg* und *Fehlgeschlagen*.

Betriebsfrequenz

Die Vorgangshäufigkeit basiert auf dem Jobtyp und dem Jobplan.

Aufbewahrung von Ereignissen

Ereignisse werden 30 Tage lang in der Benutzeroberfläche gespeichert.

Verfolgen und überwachen Sie Vorgänge

Verfolgen und überwachen Sie Vorgänge in der Workload Factory-Konsole mit Tracker.

Schritte

1. Melden Sie sich mit einem der "Konsolenerfahrungen"an.
2. Wählen Sie im Workload-Menü **Administration** und dann **Tracker** aus.
3. Verwenden Sie im Tracker die Filter oder die Suche, um die Jobergebnisse einzugrenzen. Sie können auch einen Stellenbericht herunterladen.

API-Anforderung anzeigen

Zeigen Sie die API-Anforderung in der Codebox für eine Aufgabe in Tracker an.

Schritte

1. Wählen Sie in Tracker eine Aufgabe aus.
2. Wählen Sie das Aktionsmenü und dann **API-Anforderung anzeigen**.

Wiederholen Sie einen fehlgeschlagenen Vorgang

Wiederholen Sie einen fehlgeschlagenen Vorgang in Tracker. Sie können auch die Fehlermeldung eines fehlgeschlagenen Vorgangs kopieren.

Sie können einen fehlgeschlagenen Vorgang bis zu 10 Mal wiederholen.

Schritte

1. Wählen Sie in Tracker einen fehlgeschlagenen Vorgang aus.
2. Wählen Sie das Aktionsmenü und dann **Wiederholen**.

Ergebnis

Der Vorgang wird erneut initiiert.

Bearbeiten Sie einen fehlgeschlagenen Vorgang und versuchen Sie es erneut

Bearbeiten Sie die Parameter des fehlgeschlagenen Vorgangs, und versuchen Sie den Vorgang außerhalb von Tracker erneut.

Schritte

1. Wählen Sie in Tracker einen fehlgeschlagenen Vorgang aus.
2. Wählen Sie das Aktionsmenü und dann **Bearbeiten und erneut versuchen**.

   Sie werden zur Seite Operation weitergeleitet, auf der Sie die Parameter bearbeiten und den Vorgang wiederholen können.

Ergebnis

Der Vorgang wird erneut initiiert. Rufen Sie Tracker auf, um den Status des Vorgangs anzuzeigen.

Wissen und Support

Registrieren Sie sich für den Support für NetApp Workload Factory für GenAI

Um technischen Support speziell für NetApp Workload Factory und seine Speicherlösungen und -dienste zu erhalten, ist eine Support-Registrierung erforderlich. Sie müssen sich für den Support über die NetApp -Konsole registrieren, eine separate webbasierte Konsole von Workload Factory.

Durch die Registrierung für den Support wird kein NetApp Support für den Dateidienst eines Cloud-Anbieters aktiviert. Technischen Support für den Dateidienst eines Cloud-Anbieters, seine Infrastruktur oder eine Lösung, die den Dienst nutzt, erhalten Sie unter „Hilfe“ in der Workload Factory-Dokumentation für das jeweilige Produkt.

"Amazon FSX für ONTAP"

Übersicht über die Support-Registrierung

Die Registrierung Ihres Support-Abonnements mit der Konto-ID (Ihre 20-stellige Seriennummer 960xxxxxxxx, die Sie auf der Seite „Support-Ressourcen“ in der NetApp Konsole finden) dient als Ihre einzige Support-Abonnement-ID. Jedes Support-Abonnement auf NetApp -Kontoebene muss registriert werden.

Durch die Registrierung werden Funktionen wie das Öffnen von Support-Tickets und die automatische Fallgenerierung ermöglicht. Die Registrierung wird abgeschlossen, indem Sie der NetApp Konsole wie unten beschrieben NetApp Support Site (NSS)-Konten hinzufügen.

Registrieren Sie Ihr Konto für NetApp Support

Um sich für den Support zu registrieren und den Supportanspruch zu aktivieren, muss ein Benutzer in Ihrem Konto ein NetApp Support Site-Konto mit seinem NetApp Console-Login verknüpfen. Wie Sie sich für den NetApp Support registrieren, hängt davon ab, ob Sie bereits über ein NetApp Support Site (NSS)-Konto verfügen.

Bestandskunde mit NSS-Konto

Wenn Sie NetApp -Kunde mit einem NSS-Konto sind, müssen Sie sich lediglich über die NetApp -Konsole für den Support registrieren.

Schritte

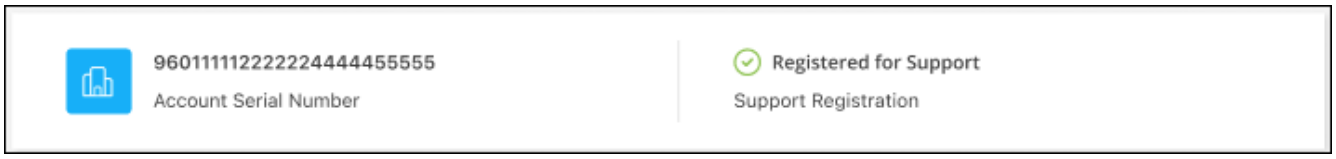
1. Wählen Sie oben rechts in der Workload Factory-Konsole **Hilfe > Support** aus.

   Wenn Sie diese Option auswählen, wird die NetApp Konsole in einem neuen Browser-Tab geöffnet und das Support-Dashboard geladen.

2. Wählen Sie im NetApp -Konsolenmenü **Administration** und dann **Anmeldeinformationen** aus.
3. Wählen Sie **Benutzeranmeldeinformationen**.
4. Wählen Sie **NSS-Anmeldeinformationen hinzufügen** und folgen Sie der Eingabeaufforderung für die NetApp-Support-Website (NSS)-Authentifizierung.

5. Um zu bestätigen, dass die Registrierung erfolgreich war, wählen Sie das Hilfesymbol und dann **Support**.

   Auf der Seite **Ressourcen** sollte angezeigt werden, dass Ihr Konto für Support registriert ist.



   Beachten Sie, dass anderen NetApp -Konsolenbenutzern dieser Support-Registrierungsstatus nicht angezeigt wird, wenn sie ihrem NetApp -Konsolen-Login kein NetApp -Support-Site-Konto zugeordnet haben. Dies bedeutet jedoch nicht, dass Ihr NetApp -Konto nicht für den Support registriert ist. Sofern ein Benutzer des Kontos diese Schritte befolgt hat, wurde Ihr Konto registriert.

Vorhandener Kunde, aber kein NSS-Konto

Wenn Sie bereits NetApp -Kunde mit vorhandenen Lizenzen und Seriennummern, aber *keinem* NSS-Konto sind, müssen Sie ein NSS-Konto erstellen und es mit Ihrem NetApp Konsolen-Login verknüpfen.

Schritte

1. Erstellen Sie ein NetApp Support Site Konto, indem Sie das abschließen "NetApp Support Site-Formular zur Benutzerregistrierung"
   a. Stellen Sie sicher, dass Sie die entsprechende Benutzerebene wählen, die normalerweise **NetApp Kunde/Endbenutzer** ist.
   b. Denken Sie daran, die Seriennummer des NetApp -Kontos (960xxxx) zu kopieren, die oben für das Feld „Seriennummer“ verwendet wurde. Dies beschleunigt die Kontobearbeitung.
2. Verknüpfen Sie Ihr neues NSS-Konto mit Ihrem NetApp Console-Login, indem Sie die folgenden Schritte ausführenBestandskunde mit NSS-Konto .

Neu bei NetApp

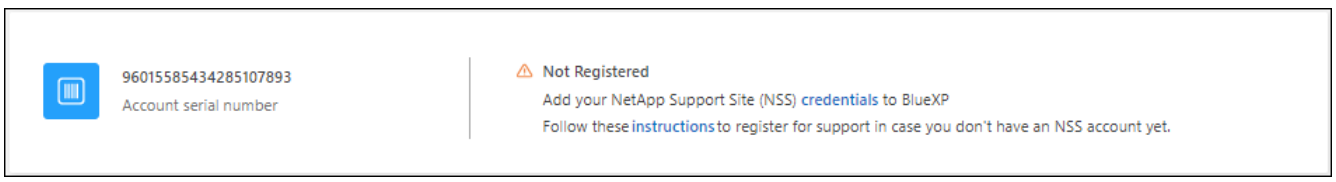
Wenn Sie neu bei NetApp sind und über keinen NSS-Account verfügen, befolgen Sie jeden Schritt unten.

Schritte

1. Wählen Sie oben rechts in der Workload Factory-Konsole **Hilfe > Support** aus.

   Wenn Sie diese Option auswählen, wird die NetApp Konsole in einem neuen Browser-Tab geöffnet und das Support-Dashboard geladen.

2. Suchen Sie auf der Seite „Support Resources“ nach der Seriennummer Ihres Kontos.



3. Navigieren Sie zu "Die Support-Registrierungs-Website von NetApp" und wählen Sie **Ich bin kein registrierter NetApp-Kunde**.
4. Füllen Sie die Pflichtfelder aus (mit roten Sternchen).
5. Wählen Sie im Feld **Product Line** die Option **Cloud Manager** aus, und wählen Sie dann den gewünschten

Abrechnungsanbieter aus.

6. Kopieren Sie die Seriennummer des Kontos von Schritt 2 oben, füllen Sie die Sicherheitsprüfung aus und bestätigen Sie dann, dass Sie die globale Datenschutzrichtlinie von NetApp lesen.

   Zur Fertigstellung dieser sicheren Transaktion wird sofort eine E-Mail an die angegebene Mailbox gesendet. Überprüfen Sie Ihre Spam-Ordner, wenn die Validierungs-E-Mail nicht in wenigen Minuten ankommt.

7. Bestätigen Sie die Aktion in der E-Mail.

   Indem Sie Ihre Anfrage an NetApp senden, wird Ihnen die Erstellung eines NetApp Support Site Kontos empfohlen.

8. Erstellen Sie ein NetApp Support Site Konto, indem Sie das abschließen "NetApp Support Site-Formular zur Benutzerregistrierung"
   a. Stellen Sie sicher, dass Sie die entsprechende Benutzerebene wählen, die normalerweise **NetApp Kunde/Endbenutzer** ist.
   b. Kopieren Sie die oben angegebene Seriennummer (960xxxx) für das Feld „Seriennummer“. Dadurch wird die Kontobearbeitung beschleunigt.

Nachdem Sie fertig sind

NetApp sollte sich bei diesem Prozess mit Ihnen in Verbindung setzen. Dies ist eine einmalige Onboarding-Übung für neue Benutzer.

Sobald Sie über Ihr NetApp Support Site-Konto verfügen, verknüpfen Sie das Konto mit Ihrem NetApp Console-Login, indem Sie die folgenden Schritte ausführenBestandskunde mit NSS-Konto .

GenAI-Fehlerbehebung

Lernen Sie, wie Sie mit einigen häufigen Problemen umgehen können.

Häufige Probleme und Lösungen

Wenn Sie eines dieser Probleme haben, können Sie mithilfe der Schritte in der Spalte „Problemumgehung“ versuchen, das Problem zu beheben.

Werden	Problem	Ursache	Behelfslösung
Einsatz	Bereitstellung schlägt fehl, da das Volume bereits vorhanden ist.	NetApp Workload Factory für GenAI muss während des Bereitstellungsprozesses ein neues Volume erstellen, es ist jedoch bereits ein Volume mit dem von Ihnen angegebenen Namen vorhanden.	Geben Sie einen eindeutigen Namen für das neue Volume an, und versuchen Sie es erneut.

Werden	Problem	Ursache	Behelfslösung
Einsatz	Die Bereitstellung schlägt fehl, da NetApp Workload Factory für GenAI das Volume nicht mounten kann.	Mindestens einer der für FSX for NetApp ONTAP erforderlichen eingehenden Ports wird geschlossen oder gefiltert.	Öffnen Sie die folgenden eingehenden Ports:

| Protokoll | Port | Zweck

| Alle ICMP | Alle | Pingen der Instanz

| HTTPS | 443 | Zugriff vom Connector auf die Verwaltungsschnittstelle fsxadmin, um API-Aufrufe an FSX zu senden

| SSH | 22 | SSH-Zugriff auf die IP-Adresse der Cluster Management LIF oder einer Node Management LIF

| TCP | 111 | Remote-Prozeduraufruf für NFS

| TCP | 139 | NetBIOS-Servicesitzung für CIFS

| TCP | 161-162 | Einfaches Netzwerkverwaltungsprotokoll

| TCP | 445 | Microsoft SMB/CIFS über TCP mit NETBIOS-Framing

| TCP | 635 | NFS-Mount

| TCP | 749 | Kerberos

| TCP | 2049 | NFS-Server-Daemon

| TCP | 3260 | ISCSI-Zugriff über die iSCSI-Daten-LIF

| TCP | 4045 | NFS-Sperr-Daemon

| TCP | 4046 | Netzwerkstatusüberwachung für NFS

| TCP | 10000 | Backup mit NDMP

| TCP | 11104 | Management von interclusterübergreifenden Kommunikationssitzungen für SnapMirror

| TCP | 11105 | SnapMirror Datenübertragung über Cluster-interne LIFs

| UDP | 111 | Remote-Prozeduraufruf für NFS

| UDP | 161-162 | Einfaches Netzwerkverwaltungsprotokoll

| UDP | 635 | NFS-Mount

| UDP | 2049 | NFS-Server-Daemon

| UDP | 4045 | NFS-Sperr-Daemon

| UDP | 4046 | Netzwerkstatusüberwachung für NFS

| UDP | 4049 | NFS rquotad-Protokoll

Wartung	Die AI-Engine startet nicht, und Sie sehen den Fehler "AI Engine Instance error" auf der Seite **Knowledge Basen**.	Die Instanz der KI-Engine wurde beschädigt oder existiert nicht.	Wählen Sie die Schaltfläche **Neu erstellen**. NetApp Workload Factory für GenAI baut die Infrastruktur neu auf und zeigt den Fortschritt des Wiederaufbaus an. Wenn der Vorgang abgeschlossen ist, werden Ihre Wissensdatenbanken wieder mit der neu erstellten Infrastruktur verbunden und die Liste der Wissensdatenbanken wird angezeigt.
Wartung	Die AI-Engine startet nicht, und Sie sehen den Fehler "die GenAI-Engine-Instanz wird gestoppt" auf der Seite **Knowledge Basen**.	Die Instanz der KI-Engine wird nicht ausgeführt.	Starten Sie die KI-Engine-Instanz mithilfe der AWS Management Console oder der AWS CLI.
Wartung	Die AI-Engine startet nicht, und Sie sehen den Fehler "der GenAI-Engine-Server reagiert nicht" auf der Seite **Wissensdatenbanken**.	Die Instanz der KI-Engine reagiert nicht.	Führen Sie die folgenden Wiederherstellungsschritte durch: **Schritte** 1. Ändern Sie die Sicherheitsgruppe der GenAI-Engine, um SSH-Zugriff auf die GenAI-Engine-Instanz zu aktivieren. 2. Melden Sie sich bei der Instanz über SSH an. 3. Führen Sie den folgenden Befehl aus: `docker-compose up`

Wartung	Die von NetApp Workload Factory für GenAI verwendete Back-End-Docker-Instanz konnte nicht gestartet werden.	Das Volume wurde gelöscht und die EC2 Instanz neu gestartet.	Führen Sie die folgenden Wiederherstellungsschritte durch:
			
			Schritte
			1. Erstellen Sie ein neues Volume auf FSX für NetApp ONTAP. Der Volume-Name kann beispielsweise sein `netapp_ai` und der Volume-Pfad kann sein `/netapp_ai`.
			2. SSH zur Amazon EC2 Instanz.
			3. Listen Sie die Volumes auf:
			`docker volume list`
			4. Entfernen Sie das alte Volume:
			`docker volume rm ec2-user_persistent_folder`
			5. Öffnen Sie die `docker-compose.yml` Datei mit einem Texteditor.
			6. `volumes`Ändern Sie im Abschnitt den Gerätepfad zum neuen Volume-Pfad. Beispiel:

Wartung	Die von NetApp Workload Factory für GenAI verwendete Back-End-Docker-Instanz konnte nicht gestartet werden.	Das Root-Volume wurde gelöscht.	Erstellen Sie ein Volume mit einem Namen und Pfad und starten Sie dann die Backend-Docker-Instanz von Amazon EC2 neu.
Wartung	Die von NetApp Workload Factory für GenAI verwendete Back-End-Docker-Instanz konnte nicht gestartet werden.	Das Root-Volume wurde gelöscht.	Erstellen Sie ein Volume mit einem Namen und Pfad und starten Sie dann die Backend-Docker-Instanz von Amazon EC2 neu.

```
0df66b96a890
d8a72.\
                fs-
0d673008aaca
12bc3.\
fsx.us-east-
1.amazonaws.
com,nolock,s
oft,rw"
                docker
':/netapp_ai
' # Path to
new volume
```

Erhalten Sie Hilfe zu NetApp Workload Factory für GenAI

NetApp bietet auf vielfältige Weise Support für Workload Factory und seine Cloud-Dienste. Umfangreiche kostenlose Self-Support-Optionen stehen rund um die Uhr zur Verfügung, beispielsweise Knowledgebase-Artikel (KB) und ein Community-Forum. Ihre Support-Registrierung beinhaltet technischen Remote-Support per Web-Ticketing.

Erhalten Sie Support für FSX für ONTAP

Technischen Support zu FSx for ONTAP, seiner Infrastruktur oder anderen Lösungen, die den Service nutzen, erhalten Sie unter „Hilfe erhalten“ in der Workload Factory-Dokumentation für das jeweilige Produkt:

"Amazon FSX für ONTAP"

Nutzen Sie die unten beschriebenen Support-Optionen, um technischen Support für Workload Factory und seine Storage-Lösungen und -Services zu erhalten.

Nutzen Sie Self-Support-Optionen

Diese Optionen sind kostenlos verfügbar, 24 Stunden am Tag, 7 Tage die Woche:

- Dokumentation

  Die Workload Factory-Dokumentation, die Sie gerade anzeigen.

- "Wissensdatenbank"

  Durchsuchen Sie die Wissensdatenbank von Workload Factory, um hilfreiche Artikel zur Problembehebung zu finden.

- "Communitys"

  Treten Sie der Workload Factory-Community bei, um laufende Diskussionen zu verfolgen oder neue zu starten.

Erstellen Sie einen Fall mit dem NetApp Support

Zusätzlich zu den oben genannten Self-Support-Optionen können Sie gemeinsam mit einem NetApp Support-

Experten eventuelle Probleme nach der Aktivierung des Supports beheben.

Bevor Sie beginnen

Um die Funktion **Fall erstellen** zu verwenden, müssen Sie sich zunächst für den Support registrieren. Verknüpfen Sie Ihre Anmeldeinformationen für die NetApp -Support-Site mit Ihrem Workload Factory-Login. "Erfahren Sie, wie Sie sich für Support registrieren" .

Schritte

1. Wählen Sie oben rechts in der Workload Factory-Konsole **Hilfe > Support** aus.

   Wenn Sie diese Option auswählen, wird die NetApp Konsole in einem neuen Browser-Tab geöffnet und das Support-Dashboard geladen.

2. Wählen Sie auf der Seite **Ressourcen** eine der verfügbaren Optionen unter Technischer Support:

   a. Wählen Sie **Rufen Sie uns an**, wenn Sie mit jemandem am Telefon sprechen möchten. Sie werden zu einer Seite auf netapp.com weitergeleitet, auf der die Telefonnummern aufgeführt sind, die Sie anrufen können.

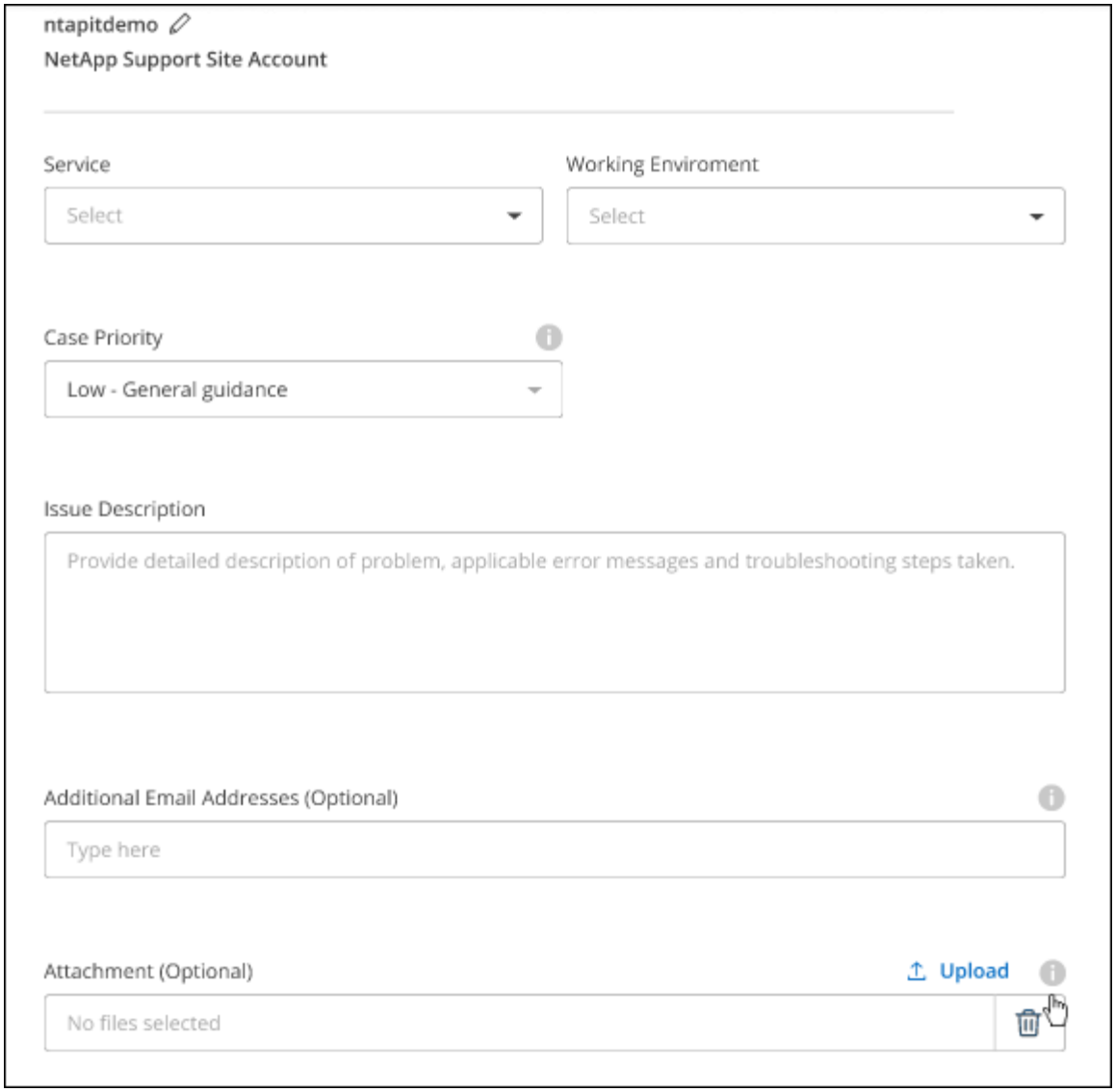
   b. Wählen Sie **Fall erstellen**, um ein Ticket mit einem NetApp-Supportspezialisten zu öffnen:

      - **Service**: Wählen Sie **Workload Factory**.
      - **Case Priority**: Wählen Sie die Priorität für den Fall, der niedrig, Mittel, hoch oder kritisch sein kann.

        Wenn Sie weitere Informationen zu diesen Prioritäten wünschen, bewegen Sie den Mauszeiger über das Informationssymbol neben dem Feldnamen.

      - **Problembeschreibung**: Geben Sie eine detaillierte Beschreibung Ihres Problems an, einschließlich aller anwendbaren Fehlermeldungen oder Fehlerbehebungsschritte, die Sie durchgeführt haben.
      - **Zusätzliche E-Mail-Adressen**: Geben Sie zusätzliche E-Mail-Adressen ein, wenn Sie jemand anderes auf dieses Problem aufmerksam machen möchten.
      - **Anhang (optional)**: Laden Sie bis zu fünf Anhänge nacheinander hoch.

        Anhänge sind auf 25 MB pro Datei begrenzt. Folgende Dateierweiterungen werden unterstützt: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx und csv.



Nachdem Sie fertig sind

Es wird ein Popup-Fenster mit der Support-Fallnummer angezeigt. Ein NetApp Support-Experte prüft Ihren Fall und macht Sie umgehend mit.

Um eine Historie deiner Support-Fälle anzuzeigen, kannst du **Einstellungen > Chronik** auswählen und nach Aktionen mit dem Namen „Support-Case erstellen“ suchen. Mit einer Schaltfläche ganz rechts können Sie die Aktion erweitern, um Details anzuzeigen.

Es ist möglich, dass beim Versuch, einen Fall zu erstellen, möglicherweise die folgende Fehlermeldung angezeigt wird:

„Sie sind nicht berechtigt, einen Fall für den ausgewählten Service zu erstellen.“

Dieser Fehler könnte bedeuten, dass das NSS-Konto und das damit verknüpfte Unternehmen nicht dasselbe Unternehmen sind, für das die Seriennummer des NetApp Konsolenkontos gilt (d. h. 960xxxx) oder die Seriennummer des Systems. Sie können auf eine der folgenden Arten Hilfe anfordern:

- Verwenden Sie den Chat im Produkt
- Übermitteln Sie einen nicht-technischen Case unter https://mysupport.netapp.com/site/help

Managen Ihrer Support-Cases (Vorschau)

Sie können aktive und gelöste Supportfälle direkt von der NetApp Konsole aus anzeigen und verwalten. Sie können die mit Ihrem NSS-Konto und Ihrem Unternehmen verknüpften Fälle verwalten.

Case Management ist als Vorschau verfügbar. Wir planen, diese Erfahrungen weiter zu verbessern und in zukünftigen Versionen Verbesserungen hinzuzufügen. Bitte senden Sie uns Ihr Feedback über den Product-Chat.

Beachten Sie Folgendes:

- Das Case-Management-Dashboard oben auf der Seite bietet zwei Ansichten:
  - Die Ansicht auf der linken Seite zeigt die Gesamtzahl der Fälle, die in den letzten 3 Monaten durch das von Ihnen angegebene NSS-Benutzerkonto eröffnet wurden.
  - Die Ansicht auf der rechten Seite zeigt die Gesamtzahl der in den letzten 3 Monaten auf Unternehmensebene eröffneten Fälle basierend auf Ihrem NSS-Benutzerkonto an.

  Die Ergebnisse in der Tabelle geben die Fälle in Bezug auf die ausgewählte Ansicht wieder.
- Sie können interessante Spalten hinzufügen oder entfernen und den Inhalt von Spalten wie Priorität und Status filtern. Andere Spalten bieten nur Sortierfunktionen.

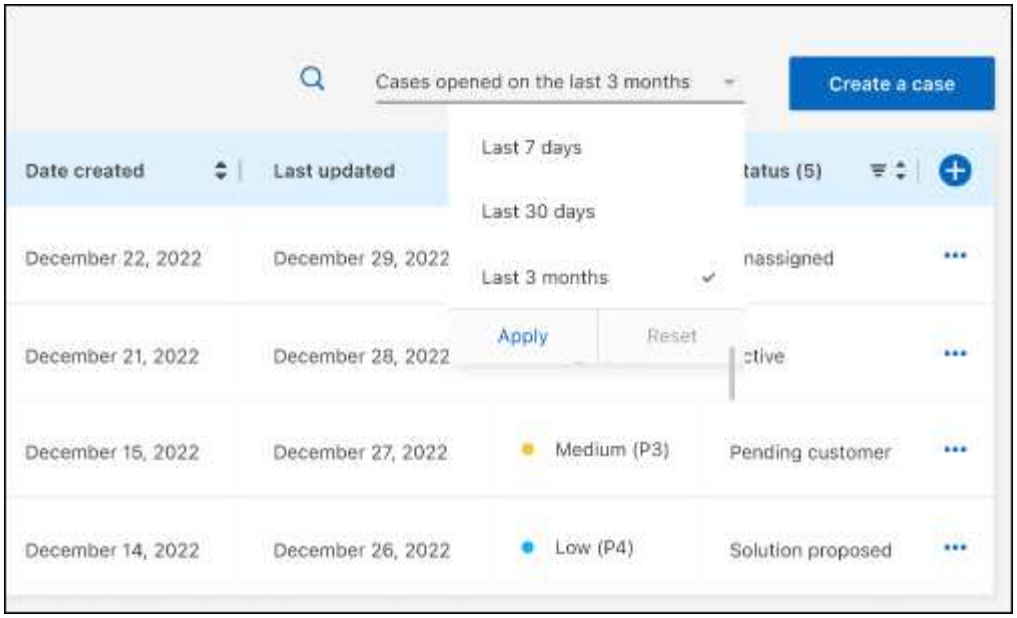
  Weitere Informationen erhalten Sie in den Schritten unten.
- Auf Fallebene bieten wir die Möglichkeit, Fallnotizen zu aktualisieren oder einen Fall zu schließen, der sich noch nicht im Status „Geschlossen“ oder „Geschlossen“ befindet.

Schritte

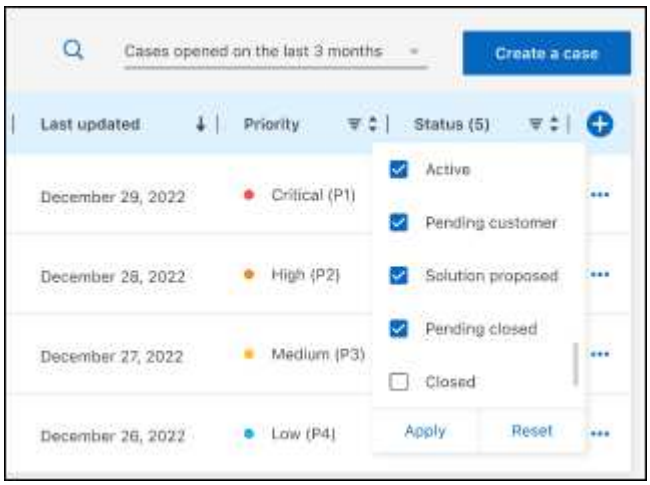
1. Wählen Sie oben rechts in der Workload Factory-Konsole **Hilfe > Support** aus.

   Wenn Sie diese Option auswählen, wird in der NetApp Konsole ein neuer Browser-Tab geöffnet und das Support-Dashboard geladen.
2. Wählen Sie **Case Management** und fügen Sie bei entsprechender Aufforderung Ihr NSS-Konto zur NetApp Konsole hinzu.

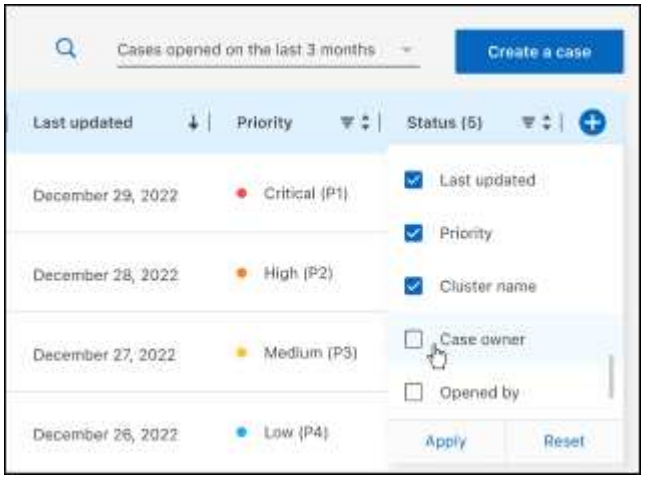
   Auf der Seite **Fallverwaltung** werden offene Fälle angezeigt, die sich auf das NSS-Konto beziehen, das mit Ihrem NetApp Console-Benutzerkonto verknüpft ist. Dies ist dasselbe NSS-Konto, das oben auf der **NSS-Verwaltungsseite** angezeigt wird.
3. Ändern Sie optional die in der Tabelle angezeigten Informationen:
   - Wählen Sie unter **Vorgänge der Organisation Ansicht** aus, um alle mit Ihrem Unternehmen verbundenen Fälle anzuzeigen.
   - Ändern Sie den Datumsbereich, indem Sie einen genauen Datumsbereich oder einen anderen Zeitrahmen auswählen.



- Filtern Sie den Inhalt der Spalten.



- Ändern Sie die Spalten, die in der Tabelle angezeigt werden, indem [Das Plus-Symbol, das in der Tabelle angezeigt wird] Sie die Spalten auswählen und dann auswählen, die Sie anzeigen möchten.



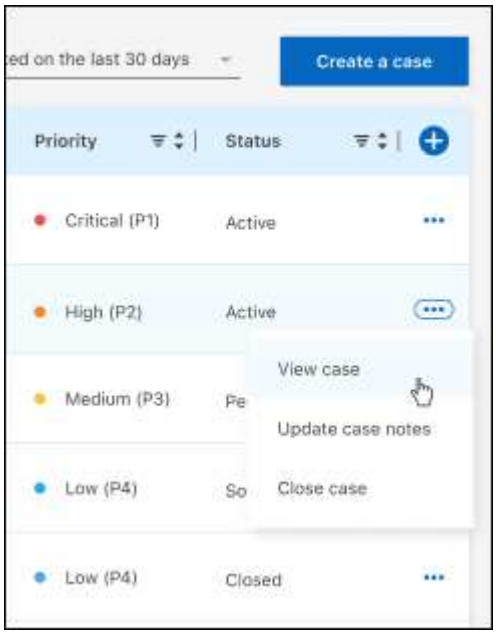
4. Managen Sie einen bestehenden Fall, indem [Ein Symbol mit drei Punkten, das in der letzten Spalte der

Tabelle angezeigt wird] Sie eine der verfügbaren Optionen auswählen:

- **Fall anzeigen**: Vollständige Details zu einem bestimmten Fall anzeigen.
- **Aktennotizen aktualisieren**: Geben Sie zusätzliche Details zu Ihrem Problem an oder wählen Sie **Dateien hochladen**, um maximal fünf Dateien anzuhängen.

  Anhänge sind auf 25 MB pro Datei begrenzt. Folgende Dateierweiterungen werden unterstützt: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx und csv.

- **Fall schließen**: Geben Sie Einzelheiten darüber an, warum Sie den Fall schließen und wählen Sie **Fall schließen**.



Rechtliche Hinweise zu NetApp Workload Factory für GenAI

Rechtliche Hinweise ermöglichen den Zugriff auf Copyright-Erklärungen, Marken, Patente und mehr.

Urheberrecht

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

Marken

NetApp, das NETAPP Logo und die auf der NetApp Markenseite aufgeführten Marken sind Marken von NetApp Inc. Andere Firmen- und Produktnamen können Marken der jeweiligen Eigentümer sein.

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

Patente

Eine aktuelle Liste der NetApp Patente finden Sie unter:

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

Datenschutzrichtlinie

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

Open Source

In den Benachrichtigungsdateien finden Sie Informationen zu Urheberrechten und Lizenzen von Drittanbietern, die in der NetApp Software verwendet werden.

"NetApp Workload Factory"