Los geht's

GenAI

NetApp
October 06, 2025

Inhalt

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

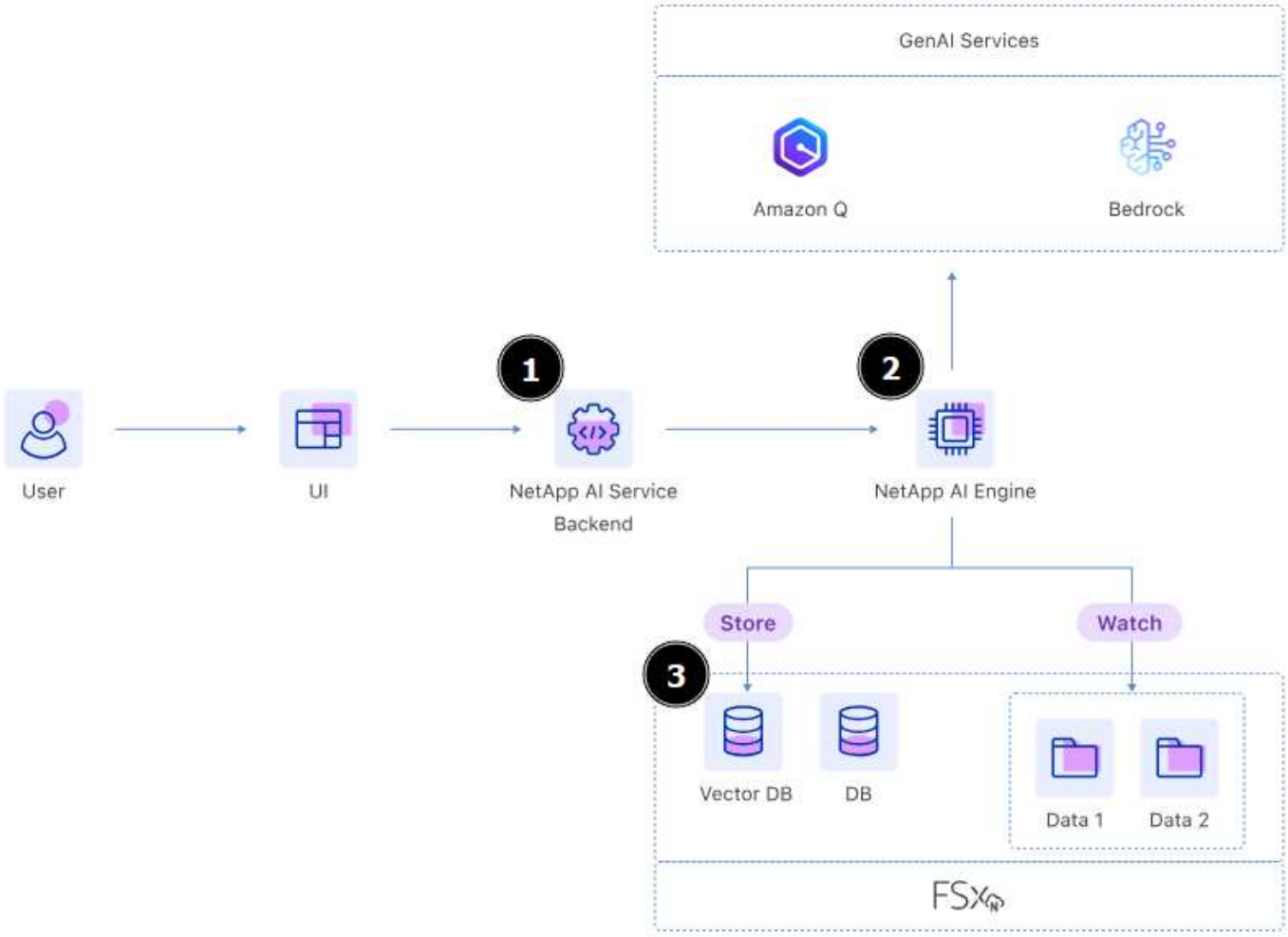
Das Filesystem FSX für ONTAP bietet den Storage für Ihr GenAI-System.

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.