Fabbrica di workload BlueXP per la documentazione GenAI

GenAI

NetApp
May 13, 2025

Sommario

Fabbrica di workload BlueXP per la documentazione GenAI

Note di rilascio

Novità della farm di workload BlueXP per Genai

Scopri le novità della funzionalità per i carichi di lavoro ai generici della farm di workload.

04 maggio 2025

Supporto per NetApp ONTAP Connector per Amazon Q

Questa versione di GenAI introduce il supporto per NetApp ONTAP Connector per Amazon Q, consentendo di creare connettori per Amazon Q Business. Sfrutta in modo rapido e semplice l'assistente ai aziendale Amazon Q con una configurazione iniziale inferiore rispetto alla creazione di una knowledge base Genai per Amazon Bedrock.

"Creare un connettore GenAI per Amazon Q Business"

Supporto di modelli di chat avanzati

GenAI supporta ora i seguenti modelli di chat aggiuntivi per le knowledge base:

- "Modelli Mistral ai"
- "Modelli di testo Amazon Titan"
- "Modelli meta Llama"
- "Modelli Jamba 1,5"
- "Modelli Cohere Command"
- "Modelli Deepseek"

GenAI supporta i modelli di ciascun provider supportati da Amazon Bedrock: "Modelli di base supportati in Amazon Bedrock"

"Creare una knowledge base GenAI"

Terminologia aggiornata per le autorizzazioni

L'interfaccia utente di fabbrica del carico di lavoro e la documentazione ora utilizzano "sola lettura" per fare riferimento alle autorizzazioni di lettura e "lettura-scrittura" per fare riferimento all'automazione delle autorizzazioni.

02 marzo 2025

Miglioramenti del chatbot integrato

A questo punto è possibile copiare domande e risposte direttamente negli Appunti, regolare le dimensioni della finestra di chat e modificarne il titolo. Inoltre, le risposte della chat possono ora includere tabelle, che sono anche copyable.

"Testare una knowledge base GenAI"

Supporto citazione risposta chat

Le risposte della chat ora includono citazioni che elencano i file e i frammenti di dati che sono stati utilizzati per generare la risposta.

"Testare una knowledge base GenAI"

Supporto avanzato dei tipi di file

Questa versione di GenAI fornisce un supporto file migliorato:

- I modelli di chat sono dotati di un supporto CSV migliorato. Ciò consente risposte più utili quando si eseguono query sui dati da file CSV.
- Genai ora può acquisire i file Apache Parquet dalle origini dei dati.
- GenAI ora supporta l'acquisizione di file Microsoft Word DOCX che includono immagini. Le immagini incorporate nei documenti DOCX vengono sottoposte a scansione e le informazioni sul testo delle immagini incorporate vengono incluse nelle risposte alle query della Knowledge base.

"Formati di file origine dati supportati"

02 febbraio 2025

Supporto dei modelli di base di Amazon Nova

Genai ora supporta i modelli di base Amazon Nova. Sono supportati Amazon Nova Micro, Amazon Nova Lite e Amazon Nova Pro.

"Requisiti Genai"

Filtraggio del tipo di file per le origini dati

GenAI ora supporta la selezione di tipi di file specifici da includere nella scansione dell'origine dati quando si aggiunge un'origine dati.

"Aggiungere fonti di dati alla knowledge base"

Filtro della data di modifica del file per le origini dati

GenAI ora supporta il filtraggio dei file da includere nella scansione dell'origine dati per data di modifica quando si aggiunge un'origine dati. È possibile scegliere un intervallo di date di modifica per i file inclusi.

"Aggiungere fonti di dati alla knowledge base"

Supporto per i file di immagine e supporto migliorato per i file PDF

GenAI ora supporta il miglioramento delle risposte alle query della knowledge base con approfondimenti dalle immagini e dalle descrizioni dei grafici, oltre al testo dei documenti, per ottenere risposte più ricche e di qualità superiore. GenAI può ora eseguire la scansione di file e immagini di immagini all'interno di file PDF (noti anche come supporto di file multimodali). Se si sceglie di eseguire la scansione di immagini o file PDF, il testo delle immagini (comprese le immagini incorporate nei documenti PDF) viene sottoposto a scansione nell'origine dati e le informazioni acquisite dalle scansioni vengono incluse nelle risposte alle query della Knowledge base.

"Aggiungere fonti di dati alla knowledge base"

Ricerca ibrida e supporto di nuovo rank

GenAI può ora migliorare significativamente la pertinenza e la precisione dei risultati di ricerca utilizzando la ricerca ibrida e ri-classificando i risultati. La ricerca ibrida combina i punti di forza della ricerca tradizionale basata su parole chiave con tecniche di ricerca semantica avanzate basate su vettori densi. I risultati della ricerca per parola chiave standard sono aumentati con corrispondenze ravvicinate e sfumature linguistiche, migliorando la rilevanza. GenAI perfeziona ulteriormente questi risultati utilizzando modelli avanzati di re-classifica, come Cohere Rank e Amazon Rank, e restituisce i risultati più rilevanti. Questa funzionalità è disponibile per le nuove basi di conoscenza create.

"Scopri la BlueXP workload Factory per Genai"

05 gennaio 2025

Nome istantanea personalizzata

A questo punto è possibile fornire un nome di istantanea per uno snapshot ad-hoc.

"Proteggere una knowledge base con le snapshot"

Nome istanza motore ai personalizzato

Ora puoi assegnare un nome personalizzato all'istanza del motore ai durante l'implementazione.

"Implementare l'infrastruttura GenAI"

Ricostruire un'infrastruttura GenAI danneggiata o mancante

Se l'istanza del motore di ai viene danneggiata o viene eliminata, puoi permettere alla fabbrica del carico di lavoro di ricrearla per te. Workload Factory ricollega automaticamente le knowledge base all'infrastruttura al termine della ricostruzione, in modo che siano pronte all'uso.

"Risoluzione dei problemi"

01 dicembre 2024

Clonare una knowledge base da uno snapshot

La fabbrica di carichi di lavoro BlueXP per Genai ora supporta il cloning di una knowledge base da una snapshot. Ciò consente il rapido recupero delle basi di conoscenza e la creazione di nuove basi di conoscenza con fonti di dati esistenti, e contribuisce al recupero e allo sviluppo dei dati.

"Clonare una knowledge base"

Rilevamento e replica del cluster ONTAP on-premise

Rileva e replica i dati dei cluster ONTAP on-premise in un file system FSX per ONTAP in modo che possano essere utilizzati per arricchire le knowledge base di ai. Tutti i flussi di lavoro di rilevamento e replica on-premise sono possibili dalla nuova scheda **ONTAP on-premise** nell'inventario di archiviazione.

"Scopri un cluster ONTAP on-premise"

3 novembre 2024

Mascherare le informazioni personali identificabili con le barriere di protezione dei dati

Il carico di lavoro ai generativo introduce la funzionalità di data Guardrails, con tecnologia di classificazione BlueXP . La funzione di protezione dei dati identifica e maschera le informazioni personali identificabili (PII, Personal Identifiable Information), contribuendo a mantenere la conformità e a rafforzare la sicurezza dei dati aziendali sensibili.

"Creare una knowledge base GenAI"

"Scopri di più sulla classificazione BlueXP"

29 settembre 2024

Supporto di Snapshot e ripristino per i volumi della Knowledge base

Ora puoi proteggere i dati dei carichi di lavoro ai generici acquisendo una copia point-in-time di una knowledge base. Ciò consente di proteggere i dati da perdite accidentali o di verificare le modifiche alle impostazioni della knowledge base. È possibile ripristinare la versione precedente del volume della Knowledge base in qualsiasi momento.

"Creare un'istantanea del volume di una knowledge base"

"Ripristinare uno snapshot di un volume della Knowledge base"

Sospendere le scansioni pianificate

È ora possibile mettere in pausa le scansioni delle origini dati pianificate. Per impostazione predefinita, i carichi di lavoro ai generativi analizzano ogni origine dati ogni giorno per acquisire nuovi dati in ciascuna knowledge base. Se non si desidera acquisire le ultime modifiche (ad esempio durante il test o il ripristino di uno snapshot), è possibile sospendere le scansioni pianificate e riprenderle in qualsiasi momento.

"Gestire le knowledge base"

Volumi di protezione dei dati ora supportati per le basi di conoscenza

Quando si seleziona un volume della Knowledge base, è ora possibile scegliere un volume di data Protection che fa parte di una relazione di replica NetApp SnapMirror. In questo modo è possibile archiviare knowledge base su volumi già protetti dalla replica SnapMirror.

"Identificare le origini dati da integrare nella propria knowledge base"

1 settembre 2024

Ulteriori strategie di scissione

I carichi di lavoro ai generativi ora supportano la suddivisione di più frasi e la suddivisione in blocchi basata su sovrapposizioni per le origini dati.

Volume dedicato per ogni knowledge base

I workload ai generativi ora creano un volume Amazon FSX per NetApp ONTAP dedicato per ogni nuova knowledge base, abilitando singole policy Snapshot per ogni knowledge base e migliorando la protezione da

guasti e intossicazione dei dati.

4 agosto 2024

Integrazione di Amazon CloudWatch Logs

I carichi di lavoro ai generativi sono ora integrati con Amazon CloudWatch Logs, consentendo di monitorare i file di log dei carichi di lavoro ai generici.

Esempio di applicazione chatbot

L'applicazione di esempio Genai della fabbrica di workload NetApp consente di testare l'autenticazione e il recupero dalla knowledge base di fabbrica dei workload NetApp pubblicati interagendo direttamente con essa in un'applicazione chatbot basata sul Web.

7 luglio 2024

Rilascio iniziale della fabbrica del carico di lavoro per GenAI

La versione iniziale include la capacità di sviluppare una knowledge base personalizzata incorporando i dati dell'organizzazione. È possibile accedere alla knowledge base tramite un'applicazione chatbot destinata agli utenti. Questa funzionalità garantisce risposte accurate e pertinenti alle domande specifiche dell'organizzazione, migliorando la soddisfazione e la produttività di tutti gli utenti.

Scopri la BlueXP workload Factory per Genai

Scopri la BlueXP workload Factory per Genai

La BlueXP workload factory per Genai ti permette di integrare i file system Amazon FSX per NetApp ONTAP con i modelli di base Genai. In questo modo otterrai uno storage dalle performance elevate con un ricco set di funzionalità di protezione, sicurezza e ottimizzazione dei costi per i tuoi set di dati ai.

Che cos'è la fabbrica di carichi di lavoro BlueXP per GenAI?

La BlueXP workload factory per Genai ti permette di usare le tue origini dei dati Enterprise in Amazon FSX per NetApp ONTAP con applicazioni ai generative. Utilizzando la generazione aumentata di recupero (RAG), è possibile collegare rapidamente le origini dati ai modelli di base disponibili tramite Amazon Bedrock o Amazon Q Business per sviluppare applicazioni basate sull'intelligenza artificiale generativa, come assistenti virtuali, chatbot di domande e risposte, sintesi dei documenti, creazione di contenuti, ecc.

L'utilizzo dell'intelligenza artificiale generativa con i dati organizzativi ti consente di sfruttare le tue conoscenze e competenze, non solo l'intelligenza del modello basata sui dati pubblici su cui i modelli sono stati formati. L'utilizzo di RAG per personalizzare i modelli garantisce risposte precise e pertinenti alle domande specifiche dell'organizzazione, migliorando la produttività e l'efficienza per gli utenti delle applicazioni utilizzando l'intelligenza artificiale generativa.

Lo sviluppo di un'applicazione Genai personalizzata in base ai dati della vostra organizzazione vi consente di sfruttare le vostre conoscenze e competenze. Questa funzionalità di personalizzazione garantisce risposte accurate e pertinenti alle domande specifiche dell'organizzazione, migliorando la soddisfazione e la produttività di tutti gli utenti.

Se "creare una knowledge base", GenAI acquisisce i dati dalle origini dati, memorizza i risultati vettorizzati in un database e fornisce il pieno controllo su come utilizzare i dati acquisiti per rispondere alle query. Questo approccio richiede una configurazione iniziale maggiore, ma consente di scegliere modelli di chat diversi per risultati diversi. Se si "Definire un connettore NetApp ONTAP per Amazon Q", i dati provenienti dalle origini dati vengono acquisiti da Amazon Q Business e archiviati in un indice. Questo approccio richiede una configurazione iniziale inferiore, ma consente di avere meno controllo sui risultati.

Per ulteriori informazioni sulla fabbrica dei carichi di lavoro, fare riferimento alla "panoramica della fabbrica del carico di lavoro".

Vantaggi dell'utilizzo di Genai per la creazione di applicazioni di intelligenza artificiale generative

La fabbrica di workload BlueXP per Genai semplifica il processo di implementazione dell'infrastruttura necessaria per creare applicazioni ai generative utilizzando la generazione aumentata di recupero (RAG). In particolare, GenAI offre i seguenti vantaggi:

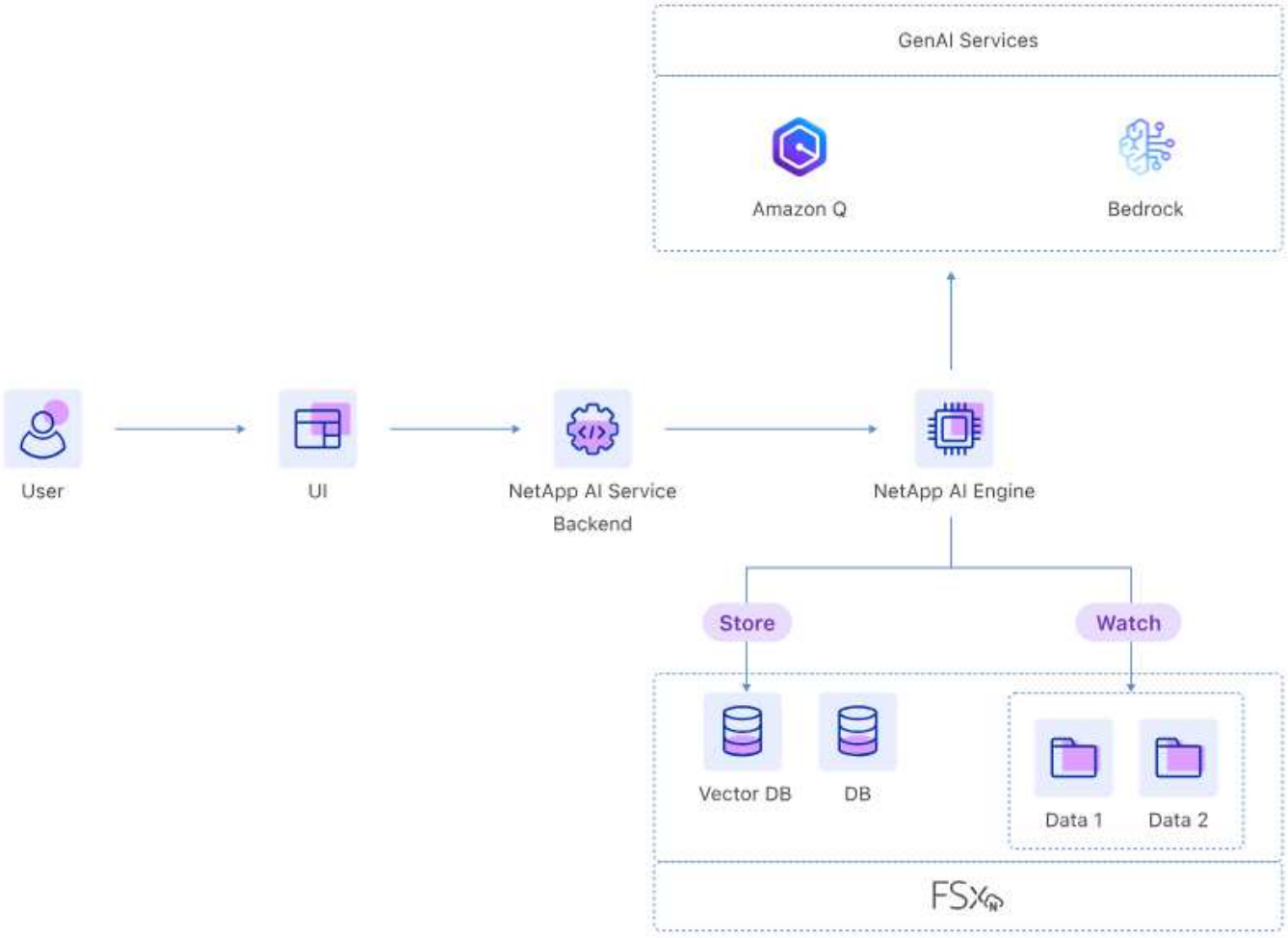
- Senza bisogno di una conoscenza approfondita dell'infrastruttura dati, delle basi e dei modelli linguistici, gli amministratori IT e gli sviluppatori possono accelerare lo sviluppo delle applicazioni utilizzando l'automazione fornita da Genai. Amministratori e sviluppatori dei dati possono creare in modo semplice e rapido knowledge base aziendali che integrano i dati non strutturati dell'organizzazione che possono essere utilizzati da applicazioni ai generative.
- Migliorare la protezione preservando le autorizzazioni degli utenti nei file incorporati nelle knowledge base

per garantire il mantenimento della protezione e della privacy dei dati. Un'applicazione, come un chatbot, può essere sviluppata per fornire solo agli utenti autenticati risposte basate sui dati a cui gli utenti hanno accesso.

- Mantieni privati e protetti i tuoi dati aziendali all'interno dell'account clienti AWS, in cui i dati aziendali non sono mai esposti all'esterno.
- Accelera lo sviluppo di applicazioni Genai come un chatbot Q&A utilizzando framework open source come LangChain che utilizzano l'API GenAI per il provisioning e la gestione di knowledge base e connettori, chattare con una knowledge base e memorizzare e recuperare la cronologia delle chat.
- Migliora la data Protection e il livello di disponibilità implementando l'infrastruttura di dati generativa ai su file system FSX per NetApp ONTAP e sfruttando funzionalità ONTAP come alta disponibilità, snapshot per la data Protection e il recovery locali, SnapMirror per il disaster recovery e SnapVault per il backup della tua infrastruttura dati.
- Ridurre i costi generali di storage per l'infrastruttura dati ai generativa, sfruttando le funzionalità di efficienza dei dati di ONTAP, come deduplica, compressione e compaction dei dati, tiering dei dati e thin provisioning.
- Ottieni risultati di alta qualità dai tuoi dati con la ricerca ibrida e le funzionalità di re-rank fornite da GenAI. La ricerca ibrida combinata con il re-ranking migliora notevolmente l'importanza dei risultati di ricerca. Queste funzionalità sono disponibili tramite Amazon AWS e dipendono da una regione.

Come funziona GenAI

GenAI utilizza i dati privati della tua organizzazione per integrare l'intelligenza del modello (in base ai dati su cui è stata formata) per fornire risposte personalizzate alle domande poste dagli utenti della tua organizzazione. Innanzitutto, è necessario implementare l'infrastruttura necessaria per un framework RAG, quindi creare una knowledge base o definire un connettore utilizzando le origini dati e i modelli di base dell'organizzazione disponibili tramite Amazon Bedrock o Amazon Q Business, quindi collegare un'applicazione (ad esempio un chatbot Q&A) alla knowledge base o al connettore.



In che modo la workload factory BlueXP per Genai aiuta a creare applicazioni ai generative

GenAI aiuta a creare applicazioni ai generative utilizzando RAG nei seguenti modi:

- Implementa l'infrastruttura richiesta per il framework RAG (Return-augmented generation) per lavorare con le origini dati in FSX per i file system ONTAP e in Amazon Bedrock o Amazon Q Business. L'infrastruttura include l'istanza del motore GenAI di NetApp per la gestione dei dati, un database vettoriale incorporato (LanceDB) e lo storage nel file system FSX per ONTAP per il database vettoriale.
- Consente di collegare le origini dati alle incorporazioni e ai modelli linguistici disponibili tramite Amazon Bedrock o Amazon Q Business per l'inclusione delle origini dati e il recupero delle risposte per le query degli utenti. Le origini dei dati, insieme ai modelli e alla loro configurazione, sono presentate come knowledge base FSX per ONTAP.
- Acquisisce i dati di origine nella knowledge base o nel connettore per incorporare i file di origine nelle condivisioni SMB e nelle esportazioni NFS in file system FSX per ONTAP, oltre a memorizzare le autorizzazioni dei file per i file all'interno delle condivisioni SMB.
- Crea automaticamente domande introduttive sulla base del contenuto delle knowledge base.
- Fornisce un simulatore di chat per gli amministratori di dati per testare le chat con le knowledge base.
- Fornisce una semplice interfaccia di connettore che ti consente di collegare Genai con Amazon Q Business, utilizzando in modo rapido e semplice le funzionalità di questo assistente ai.

https://raw.githubusercontent.com/NetAppDocs/workload-family/main/_include/learn-about-tools.adoc

Costo

L’utilizzo della funzionalità GenAI della workload factory non comporta costi.

Tuttavia, dovrai pagare per le risorse AWS implementate per supportare l’infrastruttura ai generativa. Ad esempio, pagherai AWS per Amazon Bedrock o Amazon Q Business, FSX per la capacità di storage e file system ONTAP e l’istanza EC2 del motore Genai.

Alcune operazioni multimodali, come la scansione delle immagini per le informazioni di testo, possono utilizzare più risorse e quindi comportano costi più elevati. Alcune operazioni di configurazione, ad esempio la modifica delle impostazioni per una knowledgebase, possono causare la ripetizione della scansione delle origini dati, mentre le scansioni delle origini dati possono comportare costi più elevati.

Licensing

NetApp non richiede licenze speciali per utilizzare le funzionalità ai della workload Factory.

Componenti del motore NetApp Genai

Quando si implementa l’infrastruttura GenAI, workload Factory crea un’istanza EC2 per il motore GenAI. Crea inoltre un ruolo IAM, un gruppo di protezione e endpoint privati per questa istanza. Potresti voler conoscere più dettagli su questi componenti creati dalla fabbrica dei carichi di lavoro nel tuo ambiente AWS.

Tipo di istanza di EC2

m5.large

Ruolo IAM

L’istanza del motore Genai ha bisogno delle autorizzazioni per inviare blocchi di dati al modello di incorporamento su Amazon Bedrock e per comunicare con il backend del servizio ai di NetApp. Il ruolo IAM include le seguenti autorizzazioni:

Autorizzazioni ruolo IAM

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "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: :<ai-engine-id>:kbId": "*"
        }
      },
      "Action": "fsx:DeleteVolume",
      "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
      ],
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai: :<ai-engine-
id>:qConnectorId": "*"
        }
      },
      "Action": "fsx:DeleteVolume",
      "Resource": [
        "arn:aws:fsx:*:*:volume/*/*",
        "arn:aws:fsx:*:*:backup/*"
      ],
      "Effect": "Allow"
    },
    {
```

```
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai:<ai-engine-id>": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:storage-virtual-machine/*/*",
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai:<ai-engine-id>:kbId": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:volume/*/*",
      "Effect": "Allow"
    },
    {
      "Condition": {
        "StringLike": {
          "aws:ResourceTag/netapp:wlmai:<ai-engine-
id>:qConnectorId": "*"
        }
      },
      "Action": "fsx:UntagResource",
      "Resource": "arn:aws:fsx:*:*:volume/*/*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "bedrock:InvokeModel",
        "bedrock:Rerank",
        "bedrock:GetFoundationModel",
        "bedrock:GetInferenceProfile"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "ec2messages:GetMessages",
        "ec2messages:GetEndpoint",
        "ec2messages:AcknowledgeMessage",
        "ec2messages:DeleteMessage",
```

```
            "ec2messages:FailMessage",
            "ec2messages:SendReply"
        ],
        "Resource": "*",
        "Effect": "Allow"
    },
    {
        "Action": [
            "qbusiness:ListWebExperiences",
            "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:/netapp/wlmai/*:log-stream:*",
            "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
        ],
        "Effect": "Allow"
    }
  ]
}
```

Gruppo di sicurezza

Le regole in uscita sono aperte a tutto il traffico, mentre le regole in entrata sono completamente chiuse.

Endpoint privati

Se il VPC di destinazione non ne dispone già, la fabbrica del carico di lavoro crea endpoint privati per l'istanza GenAI Engine EC2 in modo che possa comunicare con i seguenti servizi AWS:

- Amazon Bedrock
  - bedrock
  - bedrock-runtime
  - bedrock-agent-runtime
- Amazon Elastic Container Registry (ECR)
  - API
  - docker
- AWS Systems Manager (SSM)
  - ssm
  - ec2messages
  - smmessages
- Amazon FSX per NetApp ONTAP
- Amazon CloudWatch

Utilizza GenAI per creare knowledge base per Amazon Bedrock

Inizia subito

Introduzione rapida alle basi di conoscenza GenAI

Inizia a creare una knowledge base o Amazon Q Business Connector utilizzando i dati della tua organizzazione presenti nei file system di Amazon FSX per NetApp ONTAP. Un'applicazione come un chatbot accede a questa knowledge base o a questo connettore per fornire agli utenti finali risposte incentrate sull'organizzazione.

Accedi a workload Factory

Sarà necessario "configurare un account con workload factory" ed effettuare l'accesso utilizzando uno dei "esperienze di console" .

Configura il tuo ambiente per soddisfare i requisiti Genai

Avrai bisogno delle credenziali AWS per implementare l'infrastruttura AWS, un file system FSX per ONTAP implementato e rilevato, l'elenco delle origini dati che desideri integrare nella tua knowledge base o connettore, l'accesso al servizio ai di Amazon Bedrock o all'applicazione Amazon Q Business e altro ancora.

"Ulteriori informazioni sui requisiti GenAI".

Identificare il file system FSX per ONTAP che contiene le origini dati

Le origini dei dati che integrerai nella tua knowledge base possono trovarsi in un singolo file system FSX per ONTAP o in più file system FSX per ONTAP. Se questi sistemi si trovano in VPC diversi, devono essere accessibili all'interno della stessa rete, oppure è necessario eseguire il peering dei VPC e utilizzare la stessa regione e lo stesso account AWS del motore di ai.

"Scopri come identificare le origini dati".

Implementare l'infrastruttura GenAI

Avvia la procedura guidata di implementazione dell'infrastruttura per implementare l'infrastruttura Genai nel tuo ambiente AWS. Questo processo implementa un'istanza EC2 per il motore NetApp Genai e un volume su un file system FSX for ONTAP in modo da contenere i database del motore ai di NetApp. Il volume viene utilizzato per memorizzare il database vettoriale utilizzato dalla knowledge base.

"Scopri come implementare l'infrastruttura della Knowledge base".

Cosa succederà

È ora possibile creare una knowledge base per fornire agli utenti finali risposte incentrate sull'organizzazione.

Requisiti della Knowledge base Genai

Verifica che la workload Factory e AWS siano configurati correttamente prima di costruire la tua knowledge base. Ciò include la disponibilità delle credenziali di accesso ad AWS, un file system FSX per ONTAP distribuito che contiene le origini dei dati da integrare nella tua knowledge base, l'accesso al servizio ai Amazon Bedrock e altro ancora.

Requisiti di base di GenAI

Genai ha requisiti generali che il tuo ambiente deve soddisfare prima di iniziare.

Account e login di fabbrica del carico di lavoro

Sarà necessario "configurare un account con workload factory" ed effettuare l'accesso utilizzando uno dei "esperienze di console" .

Credenziali e autorizzazioni AWS

Devi aggiungere le credenziali AWS alla fabbrica del carico di lavoro con permessi di lettura/scrittura, il che significa che utilizzerai la fabbrica del carico di lavoro in modalità *lettura/scrittura* per Genai.

Le autorizzazioni per la modalità *Basic* e la modalità *Read-only* non sono attualmente supportate.

Quando imposti le tue credenziali, selezionando i permessi come mostrato di seguito avrai accesso completo alla gestione dei file system FSX per ONTAP e alla distribuzione e gestione dell'istanza Genai EC2 e delle altre risorse AWS necessarie per la tua knowledge base e il tuo chatbot.

"Scopri come aggiungere le credenziali AWS alla fabbrica di carichi di lavoro"

Requisiti della Knowledge base Genai

Se si prevede di lavorare con le knowledge base, accertarsi che l'ambiente soddisfi i seguenti requisiti.

Amazon Bedrock

Amazon Bedrock ti consente di utilizzare modelli di base e offre le funzionalità per creare applicazioni ai generative.

Prima di iniziare a utilizzare la fabbrica di carichi di lavoro BlueXP per Genai, è necessario configurare Amazon Bedrock. L'implementazione Genai deve trovarsi in una regione AWS con Amazon Bedrock abilitato.

- "Documentazione AWS: Configurare Amazon Bedrock"
- "Documentazione AWS: Regioni e modelli supportati per le Knowledge base di Amazon Bedrock"

Genai classifica nuovamente i risultati di ricerca per impostazione predefinita per migliorare la pertinenza dei risultati. Per ottenere i migliori risultati, assicurati che la configurazione del modello di base Amazon Bedrock includa l'accesso a un modello di re-rank, come Cohere Rank o Amazon Rank, se disponibile nella tua regione.

Modello di inclusione

È necessario attivare il modello di integrazione che si intende utilizzare prima di creare la propria knowledge base. Sono supportati i seguenti modelli di inclusione:

- Titano embedding G1 - testo

- Titan Embedding testo v2
- Incorporazioni multimodali Titan G1
- Incorporare l'inglese
- Incorpora multilingue

  "Scopri di più su Amazon Titan"

Modello di chat

Devi attivare il modello di chat di base che intendi utilizzare prima di creare la tua knowledge base. Poiché il supporto dei modelli varia in base alla regione di AWS, fare riferimento a "La documentazione di AWS" per verificare quali modelli è possibile utilizzare nelle regioni in cui si prevede di implementare la propria knowledge base.

GenAI supporta vari modelli da antropico, Amazon, Mistral ai, Meta, Jamba e Cohere.

Scopri di più su come utilizzare questi modelli in Amazon Bedrock:

- "Claude di antropic in Amazon Bedrock"
- "Guida introduttiva ad Amazon Nova nella console Amazon Bedrock"
- "Modelli Mistral ai"
- "Modelli di testo Amazon Titan"
- "Modelli meta Llama"
- "Modelli Jamba"
- "Modelli Cohere Command"

File system FSX per ONTAP

Ti serve almeno un file system FSX per ONTAP:

- Un file system verrà utilizzato (o creato, se non esiste) dal motore NetApp GenAI per memorizzare il database vettoriale utilizzato dalla knowledge base.

  Questo file system FSX per ONTAP deve usare FlexVol Volumes. I volumi FlexGroup non sono supportati.

- Uno o più file system conterranno le origini dati che verranno integrate nella knowledge base.

  Un unico file system FSX per ONTAP può essere utilizzato per entrambi questi scopi oppure è possibile utilizzare più file system FSX per ONTAP.

- Dovrai conoscere la regione, il VPC e la subnet di AWS in cui si trova il file system AWS FSX per ONTAP. Il file system deve trovarsi in una regione AWS in cui è abilitato Amazon Bedrock.
- Dovrai prendere in considerazione le coppie tag key/valore che desideri applicare alle risorse AWS che fanno parte di questa implementazione (opzionale).
- È necessario conoscere le informazioni sulla coppia di chiavi che consentono di connettersi in modo sicuro all'istanza del motore NetApp ai.

"Scopri come implementare e gestire i file system FSX per ONTAP"

Identificare le origini dati da aggiungere a una knowledge base o a un connettore

Identificare o creare i documenti (origini dati) che risiedono nel file system FSX per ONTAP da integrare nella knowledge base. Queste fonti di dati consentono alla knowledge base di fornire risposte accurate e personalizzate alle query degli utenti in base a dati rilevanti per l'organizzazione.

Numero massimo di origini dati

Il numero massimo di origini dati supportate è 10.

Ubicazione delle origini dati

Le origini dati possono essere memorizzate in un singolo volume o in una cartella all'interno di un volume, in una condivisione SMB o in un file system NFS su Amazon FSX per NetApp ONTAP. Le origini dei dati possono anche essere memorizzate su Amazon FSX per NetApp ONTAP Volumes che sono in una relazione di data Protection di NetApp SnapMirror.

Non è possibile selezionare singoli documenti all'interno di un volume o di una cartella, pertanto è necessario assicurarsi che ogni volume o cartella contenente origini dati non contenga documenti estranei che non devono essere integrati con la knowledge base.

Puoi aggiungere più origini dati in ciascuna Knowledge base, ma tutte devono risiedere in FSX per ONTAP, accessibile dal tuo account AWS.

La dimensione massima del file per ciascuna origine dati è di 50 MB.

Protocolli supportati

Le Knowledge base supportano i dati dai volumi che utilizzano i protocolli NFS o SMB/CIFS. Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory in modo che la knowledge base possa accedere ai file su tali volumi. Sono inclusi il dominio Active Directory, l'indirizzo IP, il nome utente e la password.

Quando si archivia l'origine dati in una condivisione (file o directory) a cui si accede tramite SMB, i dati sono accessibili solo da utenti o gruppi di chatbot che dispongono delle autorizzazioni per accedere a tale condivisione. Quando questa funzionalità "in base alle autorizzazioni" è attivata, il sistema ai confronta l'e-mail dell'utente in auth0 con gli utenti autorizzati a visualizzare o utilizzare i file nella condivisione SMB. Il chatbot fornirà le risposte in base alle autorizzazioni dell'utente per i file incorporati.

Ad esempio, se nella knowledge base sono stati integrati file 10 (origini dati) e 2 dei file sono file di risorse umane che contengono informazioni riservate, solo gli utenti di chatbot autenticati per accedere a tali file 2 riceveranno risposte dal chatbot che includono dati da tali file.

Formati di file origine dati supportati

I seguenti formati di file di origine dati sono attualmente supportati con le knowledge base Genai della fabbrica di workload.

Formato file	Interno
Apache Parquet [1]	parquet
Valori separati da virgola file[1]	.csv

Formato file	Interno
Formato di interscambio grafico	gif
JPEG	.jpg or.jpeg
JSON e JSONP[1]	json
Valutazione	md
Microsoft Word	.doc o .docx
Testo normale	.txt
Formato documento portatile	in formato .pdf
Grafica di rete portatile	.png
Immagine WebP	.webp

Implementare l'infrastruttura GenAI

Prima di creare knowledge base, connettori e applicazioni FSX per ONTAP per la tua organizzazione, devi implementare l'infrastruttura Genai per il framework RAG nel tuo ambiente. I componenti principali dell'infrastruttura sono il servizio Amazon Bedrock, un'istanza di macchina virtuale per il motore NetApp Genai e un file system FSX per ONTAP.

L'infrastruttura distribuita può supportare più knowledge base, chatbot e connettori, quindi in genere è necessario eseguire questa attività una sola volta.

Dettagli dell'infrastruttura

L'implementazione Genai deve trovarsi in una regione AWS con Amazon Bedrock abilitato. "Consente di visualizzare l'elenco delle regioni supportate"

L'infrastruttura è costituita dai seguenti componenti.

Servizio Amazon Bedrock

Amazon Bedrock è un servizio completamente gestito che ti consente di utilizzare i modelli di base (FMS) di aziende di intelligenza artificiale leader attraverso una singola API. Fornisce inoltre le funzionalità necessarie per creare applicazioni ai generative sicure.

"Scopri di più su Amazon Bedrock"

Amazon Q Business

Amazon Q si basa su Amazon Bedrock per fornire un assistente ai generativo completamente gestito che puoi utilizzare per rispondere alle domande e generare contenuti sulla base delle informazioni provenienti dalle tue origini dati.

"Scopri di più su Amazon Q Business"

Macchina virtuale per il motore NetApp Genai

Il motore NetApp Genai viene implementato durante questo processo. Fornisce la potenza di elaborazione per acquisire i dati dalle origini dati e poi scriverli nel database vettoriale.

File system FSX per ONTAP

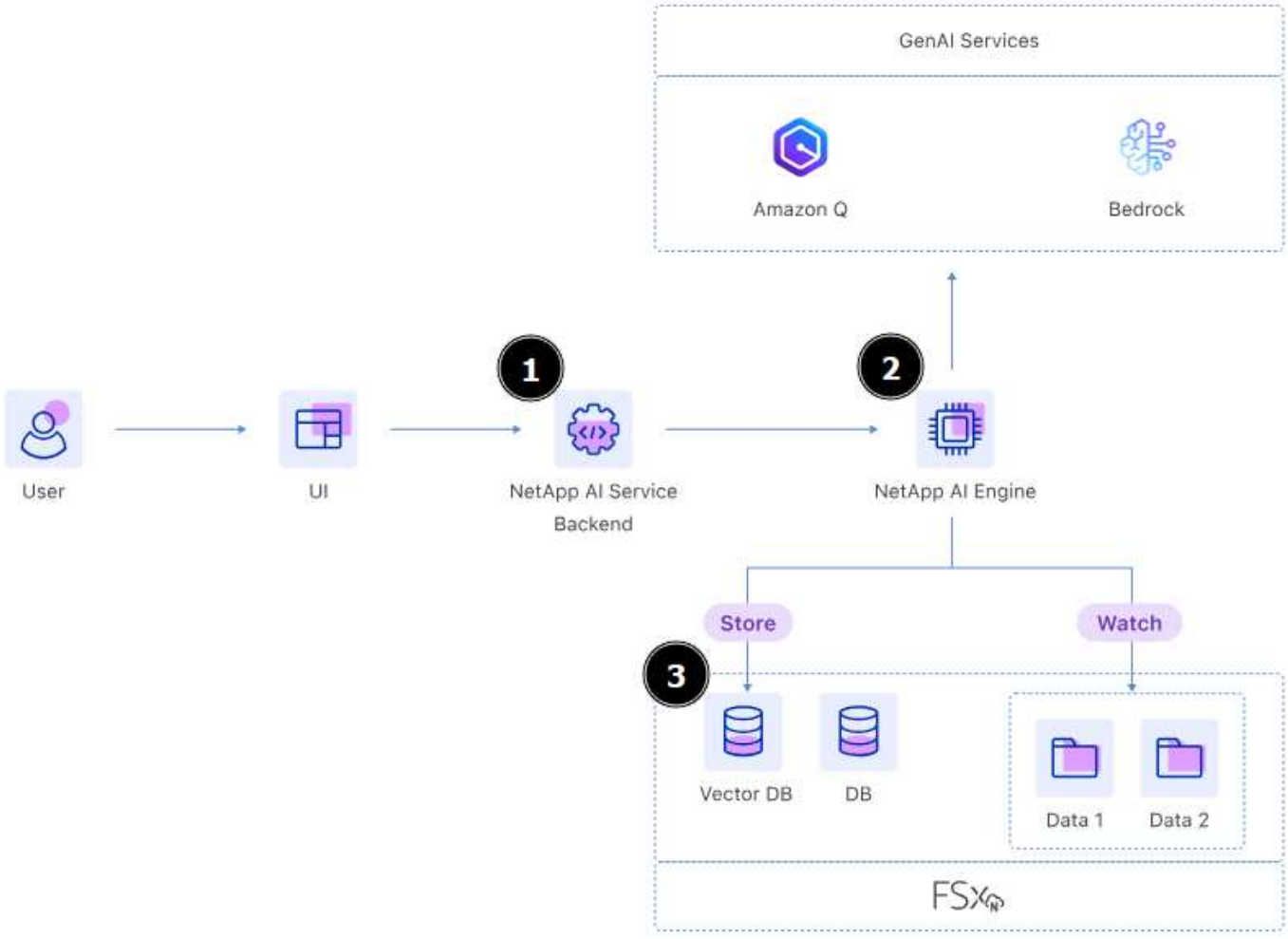
Il file system FSX per ONTAP offre lo storage per il tuo sistema GenAI.

Viene distribuito un singolo volume contenente il database vettoriale che memorizza i dati generati dal modello di base in base alle origini dati.

Le origini dati che integrerai nella tua knowledge base possono risiedere nello stesso file system FSX per ONTAP o su un sistema diverso.

Il motore NetApp Genai monitora e interagisce con entrambi questi volumi.

L'immagine seguente mostra l'infrastruttura GenAI. I componenti numerati 1, 2 e 3 vengono attivati durante questa procedura. Gli altri elementi devono essere presenti prima di iniziare la distribuzione.



Implementare l'infrastruttura GenAI

Dovrai inserire le tue credenziali AWS e selezionare il file system FSX per ONTAP per implementare l'infrastruttura di recupero e generazione aumentata (RAG).

Prima di iniziare

Prima di iniziare la procedura, accertarsi che l'ambiente in uso soddisfi i requisiti delle knowledge base o dei connettori, a seconda di quale scelta si scelga.

- "Requisiti della Knowledge base"

- "Requisiti del connettore"

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella sezione carichi di lavoro ai, selezionare **Deploy & Manage**.
3. Esaminare il diagramma dell'infrastruttura e selezionare **Avanti**.
4. Completare gli elementi nella sezione **Impostazioni AWS**:
   a. **Credenziali AWS**: Selezionare o aggiungere le credenziali AWS che forniscono le autorizzazioni per implementare le risorse AWS.
   b. **Posizione**: Selezionare una regione AWS, un VPC e una subnet.

      L'implementazione Genai deve trovarsi in una regione AWS con Amazon Bedrock abilitato. "Consente di visualizzare l'elenco delle regioni supportate"
5. Completare gli elementi nella sezione **Impostazioni infrastruttura**:
   a. **Tags**: Immettere qualsiasi coppia di chiavi di tag/valori che si desidera applicare a tutte le risorse AWS che fanno parte di questa distribuzione. Questi tag sono visibili in AWS Management Console e nell'area di informazioni sull'infrastruttura all'interno della fabbrica del carico di lavoro e possono aiutarti a tenere traccia delle risorse di fabbrica del carico di lavoro.
6. Completare la sezione **connettività**:
   a. **Coppia di chiavi**: Selezionare una coppia di chiavi che consente di connettersi in modo sicuro all'istanza del motore NetApp Genai.
7. Completare la sezione **motore ai**:
   a. **Nome istanza**: Facoltativamente, selezionare **Definisci nome istanza** e immettere un nome personalizzato per l'istanza del motore ai. Il nome dell'istanza viene visualizzato in AWS Management Console e nell'area di informazioni sull'infrastruttura all'interno della fabbrica dei workload e può aiutarti a tenere traccia delle risorse di fabbrica dei workload.
8. Selezionare **Deploy** per iniziare la distribuzione.

> Se la distribuzione non riesce con un errore di credenziali, è possibile ottenere ulteriori dettagli di errore selezionando i collegamenti ipertestuali all'interno del messaggio di errore. È possibile visualizzare un elenco di autorizzazioni mancanti o bloccate, nonché un elenco di autorizzazioni necessarie al carico di lavoro GenAI per poter implementare l'infrastruttura GenAI.

Risultato

La fabbrica dei workload inizia a implementare l'infrastruttura chatbot. Questo processo può richiedere fino a 10 minuti.

Durante il processo di distribuzione, vengono impostati i seguenti elementi:

- La rete viene impostata insieme agli endpoint privati.
- Vengono creati il ruolo IAM, il profilo di istanza e il gruppo di protezione.
- Viene implementata l'istanza della macchina virtuale per il motore GenAI.
- Amazon Bedrock è configurato per inviare i registri ad Amazon CloudWatch Logs, utilizzando un gruppo di registri con il prefisso `/aws/bedrock/`.

- Il motore GenAI è configurato per inviare i registri ad Amazon CloudWatch Logs, utilizzando un gruppo di log con il nome `/netapp/wlmai/<tenancyAccountId>/randomId`, dove `<tenancyAccountID>` è "ID dell'account BlueXP" per l'utente corrente.

Creare una knowledge base GenAI

Dopo aver implementato l'infrastruttura ai e identificato le origini dei dati che integrerai nella tua knowledge base dai datastore FSX per ONTAP, sarai pronto a costruire la knowledge base utilizzando la farm di carichi di lavoro. Nell'ambito di questo passaggio, definirai anche le caratteristiche dell'ai e creerai degli spunti di conversazione.

Prima di procedere, accertarsi che l'ambiente soddisfi "requisiti"le basi di conoscenza per .

A proposito di questa attività

Le knowledge base hanno due modalità di integrazione dei dati: *Modalità pubblica* e *modalità aziendale*.

Modalità pubblica

È possibile utilizzare una knowledge base senza integrare le origini dati dell'organizzazione. In questo caso, un'applicazione integrata con la knowledge base fornirà i risultati solo dalle informazioni disponibili pubblicamente su Internet. Questa operazione è nota come integrazione *modalità pubblica*.

Modalità Enterprise

Nella maggior parte dei casi, è necessario integrare le origini dati dell'organizzazione nella knowledge base. Questa è nota come integrazione *Enterprise mode* perché fornisce informazioni dall'azienda.

Le fonti di dati dell'organizzazione possono contenere informazioni personali identificabili (PII). Per proteggere queste informazioni sensibili, è possibile abilitare *data Guardrails* durante la creazione e la configurazione di knowledge base. Le barriere di protezione dei dati, basate sulla classificazione BlueXP , identificano e maschera PII, rendendolo inaccessibile e irrecuperabile.

"Scopri di più sulla classificazione BlueXP".

La fabbrica di carichi di lavoro BlueXP per GenAI non maschera le informazioni personali sensibili (SPii). Per ulteriori informazioni su questo tipo di dati, fare riferimento alla "tipi di dati personali sensibili".

Le barriere di sicurezza dei dati possono essere attivate o disattivate in qualsiasi momento. Se si cambia l'abilitazione dei data Guardrails, workload Factory analizza da zero l'intera knowledge base, il che implica un costo.

Creare e configurare la knowledge base

La knowledge base definisce caratteristiche quali i modelli di intelligenza artificiale Bedrock e il formato di inclusione che si desidera utilizzare per creare la propria knowledge base.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella sezione carichi di lavoro ai, selezionare **Deploy & Manage**.
3. Dalla scheda Knowledge base e connettori, selezionare il menu a discesa **Crea nuovo** e scegliere **NetApp Genai Knowledge base per Bedrock**.

4. Nella pagina Definisci Knowledge base, configurare le impostazioni della Knowledge base:
   a. **Nome**: Immettere il nome che si desidera utilizzare per la knowledge base.
   b. **Descrizione**: Inserire una descrizione dettagliata per la base di conoscenza.
   c. **Modello di incorporazione**: Il modello di incorporazione definisce come i dati verranno convertiti in incorporazioni vettoriali per la base di conoscenza. Workload Factory supporta i seguenti modelli:

      - Titano embedding G1 - testo
      - Titan Embedding testo v2
      - Incorporazioni multimodali Titan G1
      - Incorporare l'inglese
      - Incorpora multilingue

      Si noti che è necessario aver già attivato il modello di inclusione da Amazon Bedrock.

      "Scopri di più su Amazon Titan"

   d. **Modello chat**: Scegli tra i vari modelli di chat integrati in Amazon Bedrock. Nota: È necessario aver già attivato il modello di chat da Amazon Bedrock.
   e. **Data Guardrails**: Scegliere se attivare o disattivare le barriere di protezione dei dati. "Scopri le barriere di protezione dei dati, basate sulla classificazione BlueXP ".

      Per abilitare le barriere di protezione dei dati, è necessario soddisfare i seguenti prerequisiti.

      - Per comunicare con la classificazione BlueXP è necessario un account di servizio. Per la creazione dell'account di servizio, è necessario disporre del ruolo *Amministratore organizzazione* nell'account di locazione BlueXP . Un membro che ha il ruolo di amministratore dell'organizzazione può completare tutte le azioni in BlueXP . "Scopri come aggiungere un ruolo a un membro in BlueXP "
      - Il motore ai deve avere accesso a "Endpoint API BlueXP ".
      - È necessario effettuare le seguenti operazioni come descritto in "Documentazione sulla classificazione BlueXP":
        A. Crea un connettore BlueXP
        B. Accertarsi che l'ambiente in uso soddisfi i prerequisiti
        C. Implementare la classificazione BlueXP

      La funzione data Guardrails non è supportata quando si acquisiscono file di dati strutturati come CSV, JSON, JSONP o Parquet.

   f. **Avviatori di conversazione**: Scegliere se si desidera fornire fino a quattro prompt iniziali di conversazione che vengono visualizzati agli utenti che interagiscono con un chatbot che utilizza questa knowledge base. Si consiglia di attivare questa impostazione.

      Se si attivano gli avviatori di conversazione, per impostazione predefinita viene selezionata la modalità automatica. La "modalità manuale" può essere attivata solo dopo aver aggiunto origini dati alla knowledge base. "Scopri come modificare le impostazioni della Knowledge base".

   g. **File system FSX per ONTAP**: Quando si definisce una nuova knowledge base, workload Factory crea un nuovo volume Amazon FSX per NetApp ONTAP per memorizzarlo. Scegli un nome di file system esistente e una SVM (detti anche Storage VM) dove verrà creato il nuovo volume.

h. **Criterio istantanea**: Scegliere un criterio istantanea dall'elenco dei criteri esistenti definiti nell'inventario di archiviazione della fabbrica del carico di lavoro. Gli snapshot ricorrenti della Knowledge base verranno creati automaticamente con una frequenza basata sulla policy di snapshot selezionata.

   Se la policy di snapshot necessaria non esiste, è possibile "creare un criterio di snapshot" sulla VM di storage che contiene il volume.

5. Selezionare **Crea knowledge base** per aggiungere la knowledge base a GenAI.

   Durante la creazione della knowledge base viene visualizzato un indicatore di avanzamento.

   Una volta creata la knowledge base, è possibile aggiungere un'origine dati alla nuova knowledge base o terminare il processo senza aggiungere un'origine dati. Si consiglia di selezionare **Aggiungi origine dati** e aggiungere una o più origini dati.

Aggiungere fonti di dati alla knowledge base

È possibile aggiungere una o più origini dati per popolare la knowledge base con i dati dell'organizzazione.

A proposito di questa attività

Il numero massimo di origini dati supportate è 10.

Fasi

1. Dopo aver selezionato **Aggiungi origine dati**, viene visualizzata la pagina **Seleziona un file system**.
2. **Selezionare un file system**: Selezionare il file system FSX per ONTAP in cui risiedono i file di origine dati e selezionare **Avanti**.
3. **Selezionare un volume**: Selezionare il volume in cui risiedono i file di origine dati e selezionare **Avanti**.

   Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory, che includono il dominio, l'indirizzo IP, il nome utente e la password.

4. **Selezionare un'origine dati**: Selezionare la posizione dell'origine dati in base alla posizione in cui sono stati salvati i file. Può trattarsi di un intero volume o solo di una cartella o sottocartella specifica del volume e selezionare **Avanti**.
5. **Configurazioni**: Configurare il modo in cui l'origine dati acquisisce le informazioni dai file e quali file include nelle scansioni:

   - **Definisci origine dati**: Nella sezione **strategia di Chunking**, definisci il modo in cui il motore GenAI divide il contenuto dell'origine dati in blocchi quando l'origine dati è integrata con una knowledge base. È possibile scegliere una delle seguenti strategie:
     - **Chunking a più frasi**: Organizza le informazioni dalla vostra origine dati in blocchi definiti dalle frasi. È possibile scegliere quante frasi compongono ciascun blocco (fino a 100).
     - **Chunking basato su sovrapposizione**: Organizza le informazioni dall'origine dati in blocchi definiti dai caratteri che possono sovrapporsi a blocchi adiacenti. È possibile scegliere la dimensione di ciascun frammento in caratteri e la sovrapposizione di ciascun frammento con i frammenti adiacenti. È possibile configurare una dimensione del frammento compresa tra 50 e 3000 caratteri e una percentuale di sovrapposizione compresa tra 1 e 99%.

     La scelta di un'elevata percentuale di sovrapposizione può aumentare notevolmente i requisiti di archiviazione con solo lievi miglioramenti nella precisione di recupero.

- **Filtraggio dei file**: Configurare i file da includere nelle scansioni:
  - Nella sezione **supporto tipi di file**, scegliere di includere tutti i tipi di file o selezionare singoli tipi di file da includere nelle scansioni dell'origine dati.

    Se si includono immagini o file PDF, BlueXP workload factory per GenAI analizza il testo nelle immagini (incluse le immagini nei documenti PDF) e ciò comporta un costo maggiore.

  Quando si includono dati di testo da immagini, GenAI non è in grado di mascherare informazioni di identificazione personale (PII) dall'immagine, poiché i dati di testo acquisiti vengono inviati dall'ambiente ad AWS. Tuttavia, una volta memorizzati i dati, tutte le PII vengono mascherate nel database GenAI.

  La scelta di includere i file di immagine nelle scansioni è correlata al modello di chat della Knowledge base. Se si includono file di immagine nelle scansioni, il modello di chat deve supportare le immagini. Se qui sono selezionati i tipi di file immagine, non è possibile passare dalla knowledge base a un modello di chat che non supporta i file immagine.

- Nella sezione **filtro ora modifica file**, scegliere di attivare o disattivare l'inclusione dei file in base al tempo di modifica. Se si attiva il filtraggio dell'ora di modifica, selezionare un intervallo di date dall'elenco.

  Se si includono file basati su un intervallo di date di modifica, non appena l'intervallo di date non è soddisfatto (i file non sono stati modificati entro l'intervallo di date specificato), i file saranno esclusi dalla scansione periodica e l'origine dati non includerà questi file.

6. Nella sezione **Permission Aware**, disponibile solo quando l'origine dati selezionata si trova su un volume che utilizza il protocollo SMB, è possibile attivare o disattivare le risposte in base alle autorizzazioni:
   - **Enabled**: Gli utenti del chatbot che accedono a questa knowledge base riceveranno solo risposte alle query provenienti da origini dati a cui hanno accesso.
   - **Disabled**: Gli utenti del chatbot riceveranno le risposte utilizzando il contenuto di tutte le origini dati integrate.
7. Selezionare **Aggiungi** per aggiungere questa origine dati alla knowledge base.

Risultato

L'origine dati inizia a essere integrata nella tua knowledge base. Lo stato cambia da "incorporazione" a "incorporata" quando l'origine dati è completamente incorporata.

Dopo aver aggiunto una singola origine dati alla knowledge base, puoi testarla localmente nella finestra del simulatore di chatbot ed apportare le modifiche necessarie prima di rendere il chatbot disponibile per gli utenti. È inoltre possibile seguire la stessa procedura per aggiungere ulteriori origini dati alla knowledge base.

Testare una knowledge base GenAI

Dopo aver creato la knowledge base, sarà possibile testarla localmente utilizzando il simulatore di chatbot e apportare le modifiche necessarie prima di rendere disponibile la knowledge base agli utenti tramite un'applicazione chatbot.

A proposito di questa attività

È possibile testare la knowledge base per assicurarsi che funzioni come previsto e personalizzare gli spunti di

conversazione che si desidera siano disponibili per impostazione predefinita per gli utenti di chatbot di questa knowledge base. Il simulatore di chatbot viene eseguito su tutte le origini dati incorporate nella knowledge base.

Puoi testare una knowledge base chattando con le tue origini dati incorporate nel simulatore di chatbot. Tenere presente che nessuna interazione o analisi viene acquisita nel database vettoriale GenAI quando si esegue il test della knowledge base a livello locale.

Prima di distribuire la knowledge base in un'applicazione per gli utenti, eseguirete la maggior parte dei test all'interno della fabbrica dei carichi di lavoro. Se devi apportare modifiche all'origine dati o all'operazione chatbot, devi farlo ora prima di pubblicare la tua knowledge base.

È possibile ridimensionare e modificare il titolo della finestra del simulatore di chatbot e copiare domande e risposte negli Appunti.

Alcune delle attività che si desidera eseguire per testare il chatbot sono:

- Immettere un numero elevato di domande rilevanti per l'organizzazione per assicurarsi che le risposte siano quelle previste.
- Personalizzare gli avviatori di conversazione che si desidera rendere disponibili per impostazione predefinita per gli utenti nell'applicazione chatbot.
- Assicurarsi che il contenuto attribuito fornito nella parte inferiore delle risposte del chatbot contenga i riferimenti corretti.

Fasi

1. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera testare.

   Il simulatore di chatbot viene visualizzato nel riquadro di destra. Se definito, vengono visualizzati anche gli avviatori di conversazione esistenti.

2. Nel campo di immissione chatbot, immettere un prompt o una domanda e selezionare per vedere come il chatbot risponde con le proprie conoscenze organizzative.

   - È possibile visualizzare le fonti utilizzate per produrre la risposta espandendo l'elenco **sources** sotto la risposta. Viene fornito un elenco di file utilizzati per generare la risposta. È possibile visualizzare e copiare i blocchi di dati utilizzati da ciascun percorso di file e volume in ciascun file passando con il mouse sul nome del file.
   - Se le tabelle sono incluse nella risposta, è possibile ordinare i dati in ciascuna colonna e copiare ciascuna tabella negli Appunti.

3. Se è necessario aggiornare una qualsiasi delle origini dati in modo che la knowledge base fornisca risposte più mirate, apportare le modifiche subito e quindi testare nuovamente la knowledge base.

Attivare l'autenticazione esterna per una knowledge base GenAI

Attivare l'autenticazione per una knowledge base in modo che la convalida dei token e gli ACL siano necessari quando si utilizzano gli endpoint API per integrare una knowledge base con un'applicazione chatbot. Quando si attiva l'autenticazione, si configurano le impostazioni per un token Web JSON che verrà utilizzato per le richieste API a una

knowledge base dai client chatbot.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base per la quale si desidera attivare l'autenticazione.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.
4. Selezionare il menu **azioni** e selezionare **Gestisci impostazioni di autenticazione**.
5. Impostazione dell'autenticazione:

   a. Selezionare **attiva impostazioni di autenticazione**.

   b. Fornire le informazioni richieste. Alcuni esempi sono forniti, ma i valori di questi campi devono essere forniti dal provider di autenticazione:

      - **Algorithms**: L'algoritmo di firma utilizzato dal provider di autenticazione.
      - **Pubblico** (facoltativo): Una stringa contenente il destinatario del token (a volte un URL).
      - **Issuer**: Stringa che identifica il provider che ha emesso il token.

        Ad esempio, Amazon Cognito utilizza stringhe dell'emittente con il seguente formato:

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

        Dove `<region>` è la regione AWS contenente il pool di utenti e `<UserPoolID>` rappresenta l'ID del pool di utenti. È possibile recuperare l'ID del pool di utenti utilizzando il seguente comando:

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

      - **JWKS URI**: La stringa URI che fornisce le chiavi pubbliche necessarie per verificare le firme di questo token.

        Ad esempio, Amazon Cognito utilizza stringhe URI JWKS con il seguente formato:

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

        +
        Dove `<region>` è la regione AWS contenente il pool di utenti e `<UserPoolID>` rappresenta l'ID del pool di utenti. È possibile recuperare l'ID del pool di utenti utilizzando il seguente comando:

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

6. Selezionare **Salva**.

Risultato

L’autenticazione per la knowledge base è ora attiva ed è possibile utilizzare gli endpoint API per interagire con la knowledge base e integrare la knowledge base con un’applicazione chatbot.

Pubblicare una knowledge base GenAI e visualizzare l’endpoint unico

Dopo aver creato e testato la knowledge base a livello locale, è possibile pubblicare la knowledge base in modo che possa essere integrata con un’applicazione chatbot che consentirà agli utenti di interrogare la knowledge base.

A proposito di questa attività

La pubblicazione della knowledge base consente di utilizzarla nelle applicazioni di chat. L’azione di pubblicazione attiva l’API di fabbrica del workload per generare e pubblicare endpoint univoci. Dopo la pubblicazione, la knowledge base diventa accessibile per le applicazioni di chat e gli endpoint API sono pronti per l’integrazione.

Ogni knowledge base che pubblichi ha endpoint univoci.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell’inventario delle Knowledge base, selezionare la knowledge base che si desidera pubblicare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione e l’elenco di tutte le origini dati incorporate.

4. Selezionare il menu **azioni** e selezionare **pubblica**.

   La fabbrica del carico di lavoro pubblica la knowledge base. Nella pagina dei dettagli della knowledge base, lo stato cambia da **non pubblicato** a **pubblicato.**

   Ora è possibile ottenere dettagli sull’endpoint unico per la knowledge base.

5. Accanto allo stato pubblicato, selezionare **Visualizza**.

   Vengono visualizzati i dettagli su come accedere alla knowledge base utilizzando l’API di fabbrica del carico di lavoro.

6. Dalla finestra di dialogo **Visualizza informazioni pubblicate**, copiare gli endpoint API che è possibile utilizzare per integrare la knowledge base con un’applicazione.

   Per ulteriori informazioni sugli endpoint API, andare al "Documentazione API" e selezionare **ai > External**.

   Prima di poter utilizzare questi endpoint, è necessario ottenere un token utente dal provider di autenticazione.

Risultato

Ora avete una knowledge base pubblicata e l’endpoint unico che potete utilizzare per integrare la knowledge base con un’applicazione chatbot.

Utilizzare l’applicazione chatbot di esempio GenAI esterna

Dopo aver configurato, attivato e pubblicato una knowledge base, gli sviluppatori di applicazioni esterne possono configurare ed eseguire l’applicazione chatbot open source di esempio fornita da NetApp per interagire con la knowledge base e imparare a utilizzare l’API della farm dei workload per creare le proprie applicazioni ai generative.

Fasi

1. "Creare una knowledge base".
2. "Attivare l’autenticazione" per la knowledge base creata.

   Ciò consente alla knowledge base di autenticare le richieste API e rende necessarie la convalida token e gli ACL quando si utilizzano gli endpoint API.

   > Le applicazioni di chat esterne che si integrano con questa knowledge base dovranno utilizzare lo stesso provider di autenticazione (emittente) configurato nelle impostazioni di autenticazione per la knowledge base.

3. "Pubblicare la knowledge base" Per abilitare l’accesso API per le applicazioni esterne.

   Dopo la pubblicazione di una knowledge base, gli endpoint API sono accessibili dall’esterno ed è possibile integrare la knowledge base con un’applicazione di chat esterna (come l’applicazione chatbot di esempio).

4. Scaricare il pacchetto applicativo chatbot di esempio da "GitHub".
5. Installare ed eseguire l’applicazione chatbot seguendo le istruzioni contenute nel file README incluso nel pacchetto.
6. Accedere a "http://localhost:9091" per accedere all’applicazione.

   Viene visualizzata l’applicazione chatbot di esempio.

Scopri di più

"Documentazione dell’API di fabbrica del carico di lavoro"

Creare un’applicazione GenAI basata su RAG

Dopo aver creato la knowledge base e verificato il chatbot, è possibile configurare l’applicazione che consentirà agli utenti di eseguire query sul chatbot.

"Scopri come creare un’applicazione ai basata su RAG in FSX per ONTAP"

Cosa puoi fare con GenAI

Ora che hai creato una knowledge base utilizzando i dati aziendali e li hai implementati per i tuoi utenti, puoi gestire la knowledge base, le origini dei dati e l’infrastruttura RAG, inclusi i file system FSX per ONTAP.

Di seguito sono elencate alcune delle attività che è possibile eseguire per gestire i componenti della

Knowledge base:

- Aggiorna il contenuto delle tue origini dati o Aggiungi nuove origini dati e sincronizza tali modifiche con la tua knowledge base e il tuo chatbot.
- Gestire le impostazioni dell'origine dati, incluse la strategia di suddivisione dei blocchi e la consapevolezza dei permessi (per l'accesso ai file SMB).
- Gestisci le impostazioni della knowledge base, inclusi il modello di chat e gli avvii di conversazione.
- Annullare la pubblicazione di una knowledge base o ripubblicarla dopo aver apportato modifiche.
- Esegui il backup e proteggi i dati importanti sul file system FSX per ONTAP, in modo da assicurarti che i dati della knowledge base e altri componenti dell'infrastruttura siano sempre disponibili.

Per informazioni sulla gestione del file system FSX per ONTAP, consultare la "Documentazione della fabbrica di carichi di lavoro per Amazon FSX per NetApp ONTAP" per visualizzare le funzionalità di backup e protezione utilizzabili.

[1] la funzione di protezione dei dati non è supportata quando si acquisiscono file di dati strutturati in knowledge base.

Utilizzare GenAI per creare connettori per Amazon Q Business

Inizia subito

Avvio rapido per i connettori GenAI

Inizia a creare un connettore NetApp ONTAP per Amazon Q utilizzando i dati della tua organizzazione presenti nei file system di Amazon FSX per NetApp ONTAP. Dopo aver creato un connettore, gli utenti finali possono accedere all'assistente aziendale Amazon Q per ottenere risposte alle loro domande incentrate sull'organizzazione.

Accedi a workload Factory

Sarà necessario "configurare un account con workload factory" ed effettuare l'accesso utilizzando uno dei "esperienze di console" .

Configura il tuo ambiente per soddisfare i requisiti Genai

Avrai bisogno delle credenziali AWS per implementare l'infrastruttura AWS, un file system FSX per ONTAP implementato e rilevato, l'elenco di origini dati che desideri integrare nel tuo connettore, l'accesso all'applicazione Amazon Q Business e altro ancora.

"Ulteriori informazioni sui requisiti GenAI".

Identificare il file system FSX per ONTAP che contiene le origini dati

Le origini dati che integrerai nel tuo connettore possono trovarsi in un singolo file system FSX per ONTAP o in più file system FSX per ONTAP. Se questi sistemi si trovano in VPC diversi, devono essere accessibili all'interno della stessa rete, oppure è necessario eseguire il peering dei VPC e utilizzare la stessa regione e lo stesso account AWS del motore di ai.

"Scopri come identificare le origini dati".

Implementare l'infrastruttura GenAI

Avvia la procedura guidata di implementazione dell'infrastruttura per implementare l'infrastruttura Genai nel tuo ambiente AWS. Questo processo implementa un'istanza EC2 per il motore NetApp Genai e un volume su un file system FSX for ONTAP in modo da contenere i database del motore ai di NetApp. Il volume viene utilizzato per memorizzare le informazioni sul connettore.

"Scopri come implementare l'infrastruttura GenAI".

Cosa succederà

Ora puoi creare un connettore per Amazon Q Business per fornire risposte incentrate sull'organizzazione agli utenti finali.

Requisiti del connettore GenAI

Prima di creare un NetApp ONTAP Connector per Amazon Q., verifica che la fabbrica dei carichi di lavoro e AWS siano configurati correttamente

Requisiti di base di GenAI

Genai ha requisiti generali che il tuo ambiente deve soddisfare prima di iniziare.

Account e login di fabbrica del carico di lavoro

Sarà necessario "configurare un account con workload factory" ed effettuare l'accesso utilizzando uno dei "esperienze di console" .

Credenziali e autorizzazioni AWS

Devi aggiungere le credenziali AWS alla fabbrica del carico di lavoro con permessi di lettura/scrittura, il che significa che utilizzerai la fabbrica del carico di lavoro in modalità *lettura/scrittura* per Genai.

Le autorizzazioni per la modalità *Basic* e la modalità *Read-only* non sono attualmente supportate.

Quando imposti le tue credenziali, selezionando i permessi come mostrato di seguito avrai accesso completo alla gestione dei file system FSX per ONTAP e alla distribuzione e gestione dell'istanza Genai EC2 e delle altre risorse AWS necessarie per la tua knowledge base e il tuo chatbot.

"Scopri come aggiungere le credenziali AWS alla fabbrica di carichi di lavoro"

Requisiti per NetApp ONTAP Connector per Amazon Q

Assicurati che il tuo ambiente soddisfi i seguenti requisiti specifici per Amazon Q Business Connector.

Applicazione business Amazon Q

È necessario creare un'applicazione Amazon Q Business o utilizzarne una esistente.

- Verifica che l'applicazione esista in una delle tue regioni AWS.
- Assicurarsi di disporre di "ha creato un indice" per l'applicazione.
- Verificare che l'applicazione non sia in uno stato di errore.

File system FSX per ONTAP

Ti serve almeno un file system FSX per ONTAP:

- Un file system verrà utilizzato (o creato, se non esiste) dal motore NetApp Genai per memorizzare le informazioni sul connettore.

  Questo file system FSX per ONTAP deve usare FlexVol Volumes. I volumi FlexGroup non sono supportati.

- Uno o più file system conterranno le origini dati che verranno aggiunte al connettore.

  Un unico file system FSX per ONTAP può essere utilizzato per entrambi questi scopi oppure è possibile utilizzare più file system FSX per ONTAP.

- Dovrai conoscere la regione, il VPC e la subnet di AWS in cui si trova il file system AWS FSX per ONTAP.

- Dovrai prendere in considerazione le coppie tag key/valore che desideri applicare alle risorse AWS che fanno parte di questa implementazione (opzionale).
- È necessario conoscere le informazioni sulla coppia di chiavi che consentono di connettersi in modo sicuro all'istanza del motore NetApp ai.

"Scopri come implementare e gestire i file system FSX per ONTAP"

Identificare le origini dati da aggiungere a un connettore

Identificare o creare i documenti (origini dati) che risiedono nel file system FSX per ONTAP da integrare nel connettore. Queste fonti di dati consentono ad Amazon Q Business di fornire risposte accurate e personalizzate alle domande degli utenti in base ai dati rilevanti per l'organizzazione.

Numero massimo di origini dati

Il numero massimo di origini dati supportate è 10.

Ubicazione delle origini dati

Le origini dati possono essere memorizzate in un singolo volume o in una cartella all'interno di un volume, in una condivisione SMB o in un file system NFS su Amazon FSX per NetApp ONTAP. Le origini dei dati possono anche essere memorizzate su Amazon FSX per NetApp ONTAP Volumes che sono in una relazione di data Protection di NetApp SnapMirror.

Non è possibile selezionare singoli documenti all'interno di un volume o di una cartella, pertanto è necessario assicurarsi che ogni volume o cartella contenente origini dati non contenga documenti estranei che non devono essere integrati con la knowledge base.

Puoi aggiungere più origini dati in ogni connettore, ma tutte devono risiedere in FSX per ONTAP, accessibili dal tuo account AWS.

La dimensione massima del file per ciascuna origine dati è di 50 MB.

Protocolli supportati

I connettori supportano i dati da volumi che utilizzano protocolli NFS o SMB/CIFS. Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory in modo che il connettore possa accedere ai file su tali volumi. Sono inclusi il dominio Active Directory, l'indirizzo IP, il nome utente e la password.

Quando si archivia l'origine dati in una condivisione (file o directory) a cui si accede tramite SMB, i dati sono accessibili solo da utenti o gruppi di chatbot che dispongono delle autorizzazioni per accedere a tale condivisione. Quando questa funzionalità "in base alle autorizzazioni" è attivata, il sistema ai confronta l'e-mail dell'utente in auth0 con gli utenti autorizzati a visualizzare o utilizzare i file nella condivisione SMB. Il chatbot fornirà le risposte in base alle autorizzazioni dell'utente per i file incorporati.

Ad esempio, se avete integrato file 10 (origini dati) nel connettore e 2 dei file sono file di risorse umane che contengono informazioni riservate, solo gli utenti di chatbot che sono autenticati per accedere a questi file 2 riceveranno risposte dal chatbot che includono dati da questi file.

Quando si aggiungono origini dati a un Amazon Q Business Connector, ai file di origine dati vengono applicate solo le autorizzazioni utente. Le autorizzazioni del gruppo non vengono applicate.

Se un file nell'origine dati non contiene testo (ad esempio, un'immagine senza testo), Amazon Q Business non lo indicizza ma registra una voce in Amazon CloudWatch Logs rilevando l'assenza di testo.

Formati di file origine dati supportati

I seguenti formati di file di origine dati sono attualmente supportati con NetApp ONTAP Connector per Amazon Q

Formato file	Interno
File di valori separati da virgola	.csv
JSON e JSONP	json
Valutazione	md
Microsoft Word	.docx
Testo normale	.txt
Formato documento portatile	in formato .pdf
Microsoft PowerPoint	.ppt o .pptx
Hypertext Markup Language	html
Extensible Markup Language	xml
XSLT	xslt
Microsoft Excel	.xls
Formato RTF	rtf

Implementare l'infrastruttura GenAI

Prima di creare knowledge base, connettori e applicazioni FSX per ONTAP per la tua organizzazione, devi implementare l'infrastruttura Genai per il framework RAG nel tuo ambiente. I componenti principali dell'infrastruttura sono il servizio Amazon Bedrock, un'istanza di macchina virtuale per il motore NetApp Genai e un file system FSX per ONTAP.

L'infrastruttura distribuita può supportare più knowledge base, chatbot e connettori, quindi in genere è necessario eseguire questa attività una sola volta.

Dettagli dell'infrastruttura

L'implementazione Genai deve trovarsi in una regione AWS con Amazon Bedrock abilitato. "Consente di visualizzare l'elenco delle regioni supportate"

L'infrastruttura è costituita dai seguenti componenti.

Servizio Amazon Bedrock

Amazon Bedrock è un servizio completamente gestito che ti consente di utilizzare i modelli di base (FMS) di aziende di intelligenza artificiale leader attraverso una singola API. Fornisce inoltre le funzionalità necessarie per creare applicazioni ai generative sicure.

"Scopri di più su Amazon Bedrock"

Amazon Q Business

Amazon Q si basa su Amazon Bedrock per fornire un assistente ai generativo completamente gestito che puoi utilizzare per rispondere alle domande e generare contenuti sulla base delle informazioni provenienti dalle tue origini dati.

"Scopri di più su Amazon Q Business"

Macchina virtuale per il motore NetApp Genai

Il motore NetApp Genai viene implementato durante questo processo. Fornisce la potenza di elaborazione per acquisire i dati dalle origini dati e poi scriverli nel database vettoriale.

File system FSX per ONTAP

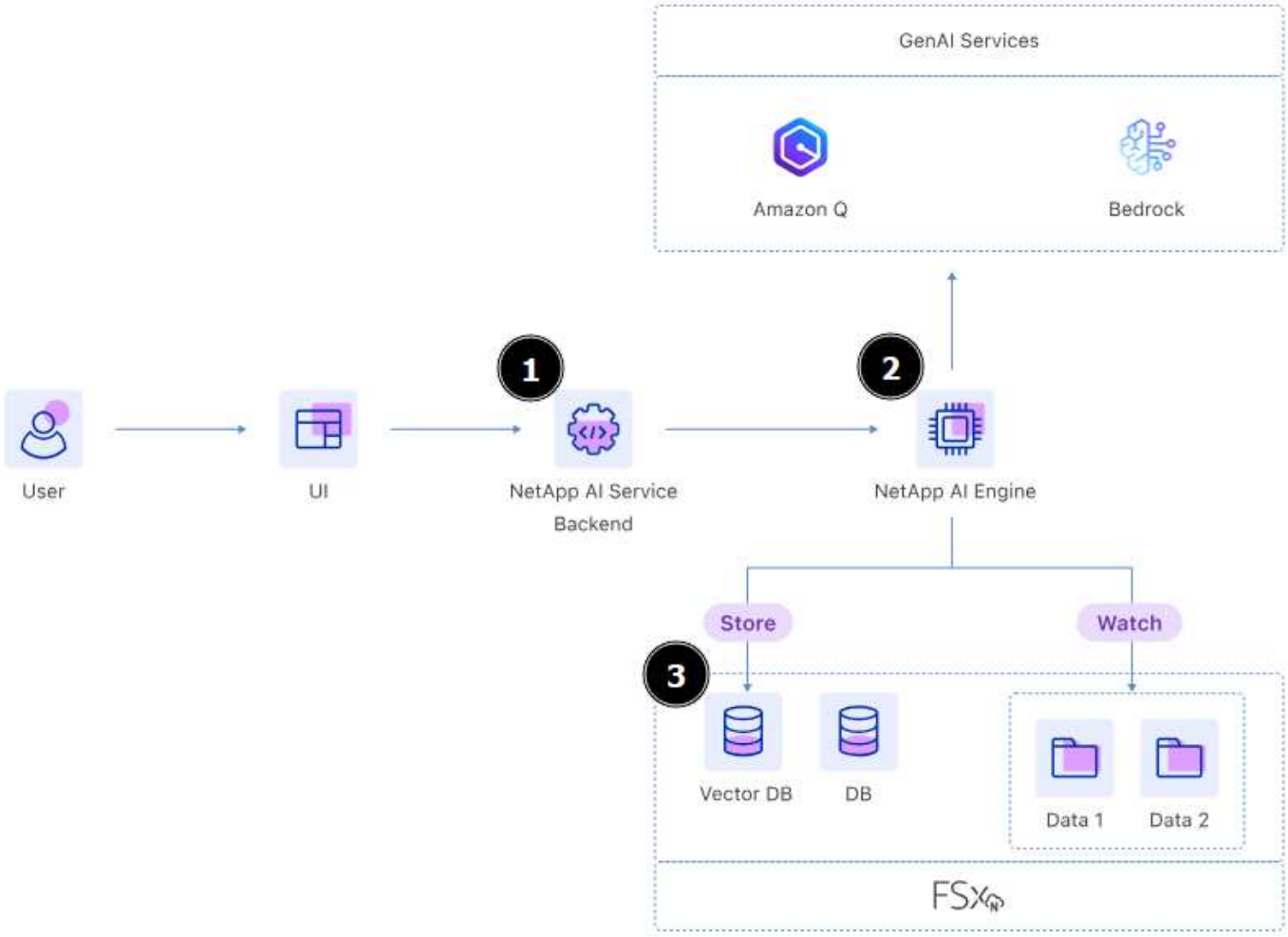
Il file system FSX per ONTAP offre lo storage per il tuo sistema GenAI.

Viene distribuito un singolo volume contenente il database vettoriale che memorizza i dati generati dal modello di base in base alle origini dati.

Le origini dati che integrerai nella tua knowledge base possono risiedere nello stesso file system FSX per ONTAP o su un sistema diverso.

Il motore NetApp Genai monitora e interagisce con entrambi questi volumi.

L'immagine seguente mostra l'infrastruttura GenAI. I componenti numerati 1, 2 e 3 vengono attivati durante questa procedura. Gli altri elementi devono essere presenti prima di iniziare la distribuzione.



Implementare l'infrastruttura GenAI

Dovrai inserire le tue credenziali AWS e selezionare il file system FSX per ONTAP per implementare l'infrastruttura di recupero e generazione aumentata (RAG).

Prima di iniziare

Prima di iniziare la procedura, accertarsi che l'ambiente in uso soddisfi i requisiti delle knowledge base o dei connettori, a seconda di quale scelta si scelga.

- "Requisiti della Knowledge base"
- "Requisiti del connettore"

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella sezione carichi di lavoro ai, selezionare **Deploy & Manage**.
3. Esaminare il diagramma dell'infrastruttura e selezionare **Avanti**.
4. Completare gli elementi nella sezione **Impostazioni AWS**:
   a. **Credenziali AWS**: Selezionare o aggiungere le credenziali AWS che forniscono le autorizzazioni per implementare le risorse AWS.
   b. **Posizione**: Selezionare una regione AWS, un VPC e una subnet.

L'implementazione Genai deve trovarsi in una regione AWS con Amazon Bedrock abilitato. "Consente di visualizzare l'elenco delle regioni supportate"

5. Completare gli elementi nella sezione **Impostazioni infrastruttura**:
   a. **Tags**: Immettere qualsiasi coppia di chiavi di tag/valori che si desidera applicare a tutte le risorse AWS che fanno parte di questa distribuzione. Questi tag sono visibili in AWS Management Console e nell'area di informazioni sull'infrastruttura all'interno della fabbrica del carico di lavoro e possono aiutarti a tenere traccia delle risorse di fabbrica del carico di lavoro.
6. Completare la sezione **connettività**:
   a. **Coppia di chiavi**: Selezionare una coppia di chiavi che consente di connettersi in modo sicuro all'istanza del motore NetApp Genai.
7. Completare la sezione **motore ai**:
   a. **Nome istanza**: Facoltativamente, selezionare **Definisci nome istanza** e immettere un nome personalizzato per l'istanza del motore ai. Il nome dell'istanza viene visualizzato in AWS Management Console e nell'area di informazioni sull'infrastruttura all'interno della fabbrica dei workload e può aiutarti a tenere traccia delle risorse di fabbrica dei workload.
8. Selezionare **Deploy** per iniziare la distribuzione.

> Se la distribuzione non riesce con un errore di credenziali, è possibile ottenere ulteriori dettagli di errore selezionando i collegamenti ipertestuali all'interno del messaggio di errore. È possibile visualizzare un elenco di autorizzazioni mancanti o bloccate, nonché un elenco di autorizzazioni necessarie al carico di lavoro GenAI per poter implementare l'infrastruttura GenAI.

Risultato

La fabbrica dei workload inizia a implementare l'infrastruttura chatbot. Questo processo può richiedere fino a 10 minuti.

Durante il processo di distribuzione, vengono impostati i seguenti elementi:

- La rete viene impostata insieme agli endpoint privati.
- Vengono creati il ruolo IAM, il profilo di istanza e il gruppo di protezione.
- Viene implementata l'istanza della macchina virtuale per il motore GenAI.
- Amazon Bedrock è configurato per inviare i registri ad Amazon CloudWatch Logs, utilizzando un gruppo di registri con il prefisso `/aws/bedrock/`.
- Il motore GenAI è configurato per inviare i registri ad Amazon CloudWatch Logs, utilizzando un gruppo di log con il nome `/netapp/wlmai/<tenancyAccountId>/randomId`, dove `<tenancyAccountID>` è "ID dell'account BlueXP" per l'utente corrente.

Creare un connettore NetApp ONTAP per Amazon Q

Dopo aver implementato l'infrastruttura ai e identificato le origini dei dati che utilizzerai dai datastore FSX per ONTAP, sarai pronto a definire un connettore NetApp ONTAP per Amazon Q

Prima di procedere, assicurati che il tuo ambiente sia conforme a "requisiti"per Amazon Q Business.

A proposito di questa attività

Le fonti di dati dell'organizzazione potrebbero contenere informazioni personali identificabili (PII, Personally Identifiable Information). Per proteggere queste informazioni sensibili, è possibile attivare *data Guardrails* quando si definisce un connettore. Le barriere di protezione dei dati, basate sulla classificazione BlueXP , identificano e maschera PII, rendendolo inaccessibile e irrecuperabile.

"Scopri di più sulla classificazione BlueXP".

La fabbrica del carico di lavoro BlueXP per GenAI non maschera le informazioni personali sensibili (SPII). Per ulteriori informazioni su questo tipo di dati, fare riferimento alla "tipi di dati personali sensibili".

Le barriere di sicurezza dei dati possono essere attivate o disattivate in qualsiasi momento. Se si cambia l'abilitazione dei data Guardrails, workload Factory esegue da zero l'intera origine dati, con un conseguente costo.

Definire un connettore

Crea un connettore NetApp ONTAP per Amazon Q. il connettore abilita la comunicazione delle API e delle origini dati tra GenAI e Amazon Q Business.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella sezione carichi di lavoro ai, selezionare **Deploy & Manage**.
3. Dalla scheda Knowledge base e connettori, seleziona il menu a discesa **Crea nuovo** e scegli **Amazon Q Business Connector**.
4. Nella pagina Definisci connettore, configurare le impostazioni del connettore:
   a. **Nome**: Immettere il nome che si desidera utilizzare per il connettore.
   b. **Descrizione**: Immettere una descrizione dettagliata per il connettore.
   c. **Amazon Q**: La regione e il nome dell'applicazione per l'istanza di Amazon Q Business che si desidera integrare.
   d. **Data Guardrails**: Scegliere se attivare o disattivare le barriere di protezione dei dati. "Scopri le barriere di protezione dei dati, basate sulla classificazione BlueXP ".

      Per abilitare le barriere di protezione dei dati, è necessario soddisfare i seguenti prerequisiti.

      - Per comunicare con la classificazione BlueXP è necessario un account di servizio. Per la creazione dell'account di servizio, è necessario disporre del ruolo *Amministratore organizzazione* nell'account di locazione BlueXP . Un membro che ha il ruolo di amministratore dell'organizzazione può completare tutte le azioni in BlueXP . "Scopri come aggiungere un ruolo a un membro in BlueXP "
      - Il motore ai deve avere accesso a "Endpoint API BlueXP ".
      - È necessario effettuare le seguenti operazioni come descritto in "Documentazione sulla classificazione BlueXP":
        A. Crea un connettore BlueXP
        B. Accertarsi che l'ambiente in uso soddisfi i prerequisiti
        C. Implementare la classificazione BlueXP

Quando si attiva la funzione di protezione dei dati, GenAI elabora i file .txt, .md, .csv, .docx e .pdf acquisendo solo testo normale (escludendo l'immagine incorporata o il testo del supporto) e mascherando qualsiasi dato privato o sensibile. Tutti gli altri tipi di file vengono elaborati normalmente senza mascherare dati privati o sensibili.

e. **FSX per il file system ONTAP**: Quando si definisce un nuovo connettore NetApp ONTAP per Amazon Q, workload Factory crea un nuovo volume Amazon FSX per NetApp ONTAP per memorizzare le informazioni sul connettore. Scegli un file system e una SVM (detti anche Storage VM) dove verrà creato il nuovo volume.

f. **Criterio istantanea**: Scegliere un criterio istantanea dall'elenco dei criteri esistenti definiti nell'inventario di archiviazione della fabbrica del carico di lavoro. Genai crea automaticamente istantanee ricorrenti del volume che memorizzano le informazioni del connettore a una frequenza basata sulla policy di snapshot selezionata.

   Se la policy di snapshot necessaria non esiste, è possibile "creare un criterio di snapshot" sulla VM di storage che contiene il volume.

5. Selezionare **Crea connettore** per integrare Amazon Q Business con GenAI.

   Durante la creazione del connettore viene visualizzato un indicatore di avanzamento.

   Dopo aver creato il connettore, è possibile aggiungere un'origine dati al connettore in modo che Amazon Q Business acquisisca i dati e li aggiunga al relativo indice. Si consiglia di selezionare **Aggiungi origine dati** e aggiungere una o più origini dati.

Aggiungere origini dati al connettore

È possibile aggiungere una o più origini dati per popolare l'indice Amazon Q Business con i dati dell'organizzazione.

A proposito di questa attività

- Il numero massimo di origini dati supportate è 10.
- Fare riferimento a "Documentazione di Amazon Q Business" per le restrizioni specifiche dei servizi dell'indice di Amazon Q Business.

Fasi

1. Dopo aver selezionato **Aggiungi origine dati**, viene visualizzata la pagina **Seleziona un file system**.
2. **Selezionare un file system**: Selezionare il file system FSX per ONTAP in cui risiedono i file di origine dati e selezionare **Avanti**.
3. **Selezionare un volume**: Selezionare il volume in cui risiedono i file di origine dati e selezionare **Avanti**.

   Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory, che includono il dominio, l'indirizzo IP, il nome utente e la password.

4. **Selezionare un'origine dati**: Selezionare la posizione dell'origine dati in base alla posizione in cui sono stati salvati i file. Può trattarsi di un intero volume o solo di una cartella o sottocartella specifica del volume e selezionare **Avanti**.
5. **Configurazioni**: Configurare il modo in cui l'origine dati acquisisce le informazioni dai file e quali file include nelle scansioni:
   - **Filtraggio dei file**: Configurare i file da includere nelle scansioni:

- Nella sezione **supporto tipi di file**, scegliere di includere tutti i tipi di file o selezionare singoli tipi di file da includere nelle scansioni dell'origine dati.
- Nella sezione **filtro ora modifica file**, scegliere di attivare o disattivare l'inclusione dei file in base al tempo di modifica. Se si attiva il filtraggio dell'ora di modifica, selezionare un intervallo di date dall'elenco.

> Se si includono file basati su un intervallo di date di modifica, non appena l'intervallo di date non è soddisfatto (i file non sono stati modificati entro l'intervallo di date specificato), i file saranno esclusi dalla scansione periodica e l'origine dati non includerà questi file.

6. Nella sezione **Permission Aware**, disponibile solo quando l'origine dati selezionata si trova su un volume che utilizza il protocollo SMB, è possibile attivare o disattivare le risposte in base alle autorizzazioni:
   - **Enabled**: Gli utenti del chatbot che accedono a questo connettore riceveranno solo risposte alle query provenienti da origini dati a cui hanno accesso.
   - **Disabled**: Gli utenti del chatbot riceveranno le risposte utilizzando il contenuto di tutte le origini dati integrate.

> Le autorizzazioni del gruppo Active Directory non sono supportate per le origini dati di Amazon Q Business Connector.

7. Selezionare **Aggiungi** per aggiungere questa origine dati ad Amazon Q Business Connector.

Risultato

L'origine dati è integrata nell'indice di Amazon Q Business. Lo stato cambia da "incorporazione" a "incorporata" quando l'origine dati è completamente incorporata.

Dopo aver aggiunto una singola origine dati al connettore, è possibile testarla nell'ambiente chatbot di Amazon Q Business e apportare le modifiche necessarie prima di rendere il servizio disponibile per gli utenti. È inoltre possibile seguire la stessa procedura per aggiungere ulteriori origini dati al connettore.

Somministrare GenAI

Gestire l’infrastruttura GenAI

Puoi visualizzare i dettagli sull’infrastruttura Genai RAG implementata o rimuovere l’infrastruttura chatbot se non ne hai più bisogno.

Consente di visualizzare informazioni sull’infrastruttura

È possibile visualizzare informazioni sull’infrastruttura del chatbot.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dal menu di navigazione di fabbrica del carico di lavoro, selezionare **ai**.
3. Selezionare la scheda **infrastruttura**.
4. Visualizzare le informazioni relative all’infrastruttura, che includono dettagli sui seguenti componenti:
    - Impostazioni AWS
    - Impostazioni dell’infrastruttura
    - Il motore ai
    - Il database vettoriale

Rimuovere l’infrastruttura

Se non hai più bisogno dell’infrastruttura chatbot implementata per uno o più chatbot, puoi rimuoverla da workload Factory.

> Tutti i chatbot che sono stati distribuiti su questa infrastruttura saranno disabilitati e tutta la cronologia chat sarà eliminata.

Questa operazione rimuove dalla fabbrica dei workload solo i link all’infrastruttura ai, ma non tutti i componenti da AWS. Devi eliminare manualmente i seguenti componenti dell’infrastruttura da AWS:

- Istanza della macchina virtuale
- Endpoint privati
- Volume nel file system FSX per ONTAP che contiene i database ai
- Il ruolo IAM
- La policy
- Il gruppo di protezione

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dal menu di navigazione di fabbrica del carico di lavoro, selezionare **ai**.
3. Selezionare la scheda **infrastruttura**.
4. Selezionare ••• e selezionare **Rimuovi infrastruttura chatbot**.

5. Confermare che si desidera eliminare l'infrastruttura e selezionare **Rimuovi**.

Risultato

I componenti dell'infrastruttura del chatbot sono rimossi dalla fabbrica del carico di lavoro.

Gestire le knowledge base GenAI

Dopo aver creato una knowledge base, è possibile visualizzare i dettagli della knowledge base, modificare la knowledge base, integrare fonti di dati aggiuntive o eliminare la knowledge base.

Consente di visualizzare informazioni su una knowledge base

È possibile visualizzare informazioni sulle impostazioni di una knowledge base e sull'origine dati integrata.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dal menu di navigazione di fabbrica del carico di lavoro, selezionare **ai**.
3. Selezionare la knowledge base che si desidera visualizzare.

   Se definito, gli avviamenti di conversazione attualmente in uso vengono visualizzati nel riquadro di destra.

4. Per visualizzare i dettagli della knowledge base, selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

   Il menu **azioni** consente di gestire la knowledge base se si desidera apportare modifiche.

Modificare una knowledge base

È possibile aggiornare una knowledge base modificando alcune impostazioni oppure aggiungere o rimuovere origini dati.

Ogni volta che si aggiungono, modificano o rimuovono origini dati dalla knowledge base, è necessario sincronizzare l'origine dati in modo che venga riindicizzata nella knowledge base. La sincronizzazione è incrementale, quindi Amazon Bedrock elabora solo gli oggetti nel volume FSX per ONTAP che sono stati aggiunti, modificati o eliminati dall'ultima sincronizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera aggiornare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

4. Selezionare il menu **azioni** e selezionare **Modifica knowledge base**.

5. Nella pagina Modifica knowledge base è possibile modificare il nome della knowledge base, la descrizione, il modello di incorporazione, il modello di chat, l'abilitazione delle barriere di protezione dei dati, scegliere se creare automaticamente o manualmente gli avviatori di conversazione e il criterio di snapshot utilizzato per il volume che contiene la knowledge base.

   Se si utilizza la modalità Manuale per avviare una conversazione, è anche possibile cambiare gli avviatori di conversazione.

   > Ogni scansione della Knowledge base, che include l'integrazione, i costi. Se le barriere di sicurezza dei dati sono abilitate dopo la creazione di una knowledge base, la knowledge base viene nuovamente sottoposta a scansione e i costi sono sostenuti. Allo stesso modo, se si modificano i modelli di chat, GenAI eseguirà una nuova scansione delle origini dati associate (con costi aggiuntivi).

6. Selezionare **Salva** dopo aver apportato le modifiche.

Proteggere una knowledge base con le snapshot

Puoi proteggere i dati della tua knowledge base eseguendo e ripristinando snapshot dei volumi della tua knowledge base. È possibile eseguire il ripristino da uno snapshot per tornare alla versione precedente della knowledge base in qualsiasi momento.

Le snapshot possono essere più veloci ed efficienti in termini di storage rispetto ai backup, consentendo inoltre di proteggere ogni knowledge base utilizzando una policy di protezione differente. Di seguito sono elencati alcuni degli scenari in cui le snapshot possono essere utili:

- Perdita o danneggiamento accidentale di dati
- Recupero da dati errati inseriti nella knowledge base
- Test di diverse origini dati o strategie di suddivisione e rapido ritorno al termine del test

Creare un'istantanea del volume di una knowledge base

È possibile salvare lo stato di una knowledge base eseguendo un'istantanea manuale del volume della knowledge base.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina di inventario delle Knowledge base, selezionare la knowledge base che si desidera proteggere.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

4. Selezionare il menu **azioni** e selezionare **istantanea > Crea nuova istantanea**.
5. Facoltativamente, selezionare **Definisci nome istantanea** e immettere un nome personalizzato per l'istantanea.

   La definizione di un nome personalizzato può aiutare a determinare meglio il contenuto di uno snapshot se è necessario ripristinarlo in futuro.

6. Selezionare **Crea**.

   Viene creata un'istantanea della knowledge base.

Ripristinare uno snapshot di un volume della Knowledge base

È possibile ripristinare uno snapshot manuale o pianificato di un volume della Knowledge base in qualsiasi momento.

> Non è possibile ripristinare uno snapshot utilizzando l'interfaccia utente dei carichi di lavoro ai generativi se il database archiviato sul volume è corrotto o è stato eliminato. Come soluzione alternativa, puoi ripristinare la snapshot utilizzando sul cluster ONTAP in cui è ospitato il "CLI ONTAP" volume.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera ripristinare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

4. Selezionare il menu **azioni** e selezionare **istantanea > Ripristina istantanea**.

   Viene visualizzata la finestra di dialogo di selezione delle istantanee, in cui è possibile visualizzare un elenco delle istantanee create per questa knowledge base.

5. (Facoltativo) deselezionare l'opzione **Sospendi scansioni in esecuzione e pianificate dopo il ripristino dell'istantanea** se si desidera che le scansioni dell'origine dati pianificate e in esecuzione continuino dopo il ripristino dell'istantanea.

   Questa opzione è attivata per impostazione predefinita per garantire che la scansione non avvenga mentre la base di conoscenza è in uno stato parzialmente ripristinato o che una scansione non aggiorni una knowledge base appena ripristinata con i dati meno recenti.

6. Selezionare dall'elenco l'istantanea che si desidera ripristinare.
7. Selezionare **Restore** (Ripristina).

Clonare una knowledge base

È possibile creare una nuova knowledge base a partire da un'istantanea della Knowledge base. Ciò è utile se la knowledge base originale è danneggiata o persa.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera ripristinare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di

incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

4. Selezionare il menu **azioni** e selezionare **istantanea > Clona knowledge base**.

   Viene visualizzata la finestra di dialogo Clona.

5. In alternativa, deselezionare l'opzione **Pause running and scheduled scans after cloning the snapshot** se si desidera che le scansioni dell'origine dati pianificate e in esecuzione continuino dopo la clonazione dello snapshot.

   Questa opzione è attivata per impostazione predefinita per garantire che la scansione non avvenga mentre la base di conoscenza è in uno stato parzialmente ripristinato o che una scansione non aggiorni una knowledge base appena ripristinata con i dati meno recenti.

6. Selezionare dall'elenco lo snapshot che si desidera clonare.
7. Selezionare **continua**.
8. Immettere un nome per la nuova knowledge base.
9. Scegliere una SVM del file system e un nome del volume per la nuova knowledge base.
10. Selezionare **Clone**.

Aggiungere ulteriori origini dati a una knowledge base

È possibile incorporare ulteriori origini dati nella knowledge base per inserirle con dati aggiuntivi dell'organizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base in cui si desidera aggiungere l'origine dati.
3. Selezionare ••• e selezionare **Aggiungi origine dati**.
4. **Selezionare un file system**: Selezionare il file system FSX per ONTAP in cui risiedono i file di origine dati e selezionare **Avanti**.
5. **Selezionare un volume**: Selezionare il volume in cui risiedono i file di origine dati e selezionare **Avanti**.

   Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory, che includono il dominio, l'indirizzo IP, il nome utente e la password.

6. **Selezionare un'origine dati**: Selezionare la posizione dell'origine dati in base alla posizione in cui sono stati salvati i file. Può trattarsi di un intero volume o solo di una cartella o sottocartella specifica del volume e selezionare **Avanti**.
7. **Configurazioni**: Configurare il modo in cui l'origine dati acquisisce le informazioni dai file e quali file include nelle scansioni:

   - **Definisci origine dati**: Nella sezione **strategia di Chunking**, definisci il modo in cui il motore GenAI divide il contenuto dell'origine dati in blocchi quando l'origine dati è integrata con una knowledge base. È possibile scegliere una delle seguenti strategie:
     - **Chunking a più frasi**: Organizza le informazioni dalla vostra origine dati in blocchi definiti dalle frasi. È possibile scegliere quante frasi compongono ciascun blocco (fino a 100).
     - **Chunking basato su sovrapposizione**: Organizza le informazioni dall'origine dati in blocchi definiti dai caratteri che possono sovrapporsi a blocchi adiacenti. È possibile scegliere la

dimensione di ciascun frammento in caratteri e la sovrapposizione di ciascun frammento con i frammenti adiacenti. È possibile configurare una dimensione del frammento compresa tra 50 e 3000 caratteri e una percentuale di sovrapposizione compresa tra 1 e 99%.

> La scelta di un'elevata percentuale di sovrapposizione può aumentare notevolmente i requisiti di archiviazione con solo lievi miglioramenti nella precisione di recupero.

- **Filtraggio dei file**: Configurare i file da includere nelle scansioni:
  - Nella sezione **supporto tipi di file**, scegliere di includere tutti i tipi di file o selezionare singoli tipi di file da includere nelle scansioni dell'origine dati.

    Se si includono immagini o file PDF, BlueXP workload factory per GenAI analizza il testo nelle immagini (incluse le immagini nei documenti PDF) e ciò comporta un costo maggiore.

  Quando si includono dati di testo da immagini, GenAI non è in grado di mascherare informazioni di identificazione personale (PII) dall'immagine, poiché i dati di testo acquisiti vengono inviati dall'ambiente ad AWS. Tuttavia, una volta memorizzati i dati, tutte le PII vengono mascherate nel database GenAI.

> La scelta di includere i file di immagine nelle scansioni è correlata al modello di chat della Knowledge base. Se si includono file di immagine nelle scansioni, il modello di chat deve supportare le immagini. Se qui sono selezionati i tipi di file immagine, non è possibile passare dalla knowledge base a un modello di chat che non supporta i file immagine.

- Nella sezione **filtro ora modifica file**, scegliere di attivare o disattivare l'inclusione dei file in base al tempo di modifica. Se si attiva il filtraggio dell'ora di modifica, selezionare un intervallo di date dall'elenco.

> Se si includono file basati su un intervallo di date di modifica, non appena l'intervallo di date non è soddisfatto (i file non sono stati modificati entro l'intervallo di date specificato), i file saranno esclusi dalla scansione periodica e l'origine dati non includerà questi file.

8. Nella sezione **Permission Aware**, disponibile solo quando l'origine dati selezionata si trova su un volume che utilizza il protocollo SMB, è possibile attivare o disattivare le risposte in base alle autorizzazioni:
   - **Enabled**: Gli utenti del chatbot che accedono a questa knowledge base riceveranno solo risposte alle query provenienti da origini dati a cui hanno accesso.
   - **Disabled**: Gli utenti del chatbot riceveranno le risposte utilizzando il contenuto di tutte le origini dati integrate.
9. Selezionare **Aggiungi** per aggiungere questa origine dati alla knowledge base.

Risultato

L'origine dati è integrata nella knowledge base.

Sincronizzare le origini dati con una knowledge base

Le origini dati vengono sincronizzate automaticamente con la knowledge base associata una volta al giorno, in modo che eventuali modifiche all'origine dati vengano riflesse nel chatbot. Se si apportano modifiche a una delle origini dati e si desidera sincronizzare immediatamente i dati, è possibile eseguire una sincronizzazione su richiesta.

La sincronizzazione è incrementale, quindi Amazon Bedrock elabora solo gli oggetti nelle origini dati che sono stati aggiunti, modificati o eliminati dall'ultima sincronizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera sincronizzare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.
4. Selezionare il menu **azioni** e selezionare **Esegui scansione ora**.

   Viene visualizzato un messaggio che indica che le origini dati sono in fase di scansione e un messaggio finale al termine della scansione.

Risultato

La knowledge base è sincronizzata con le origini dati collegate e qualsiasi chatbot attivo inizierà a utilizzare le informazioni più recenti provenienti dalle origini dati.

Consente di sospendere o riprendere una sincronizzazione pianificata

Se si desidera sospendere o riprendere la successiva sincronizzazione (scansione) delle origini dati, è possibile farlo in qualsiasi momento. Potrebbe essere necessario sospendere la successiva sincronizzazione pianificata se si intende apportare modifiche a un'origine dati e non si desidera che la sincronizzazione avvenga durante la finestra di modifica.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla scheda Knowledge base e connettori, selezionare la knowledge base per la quale si desidera sospendere o riprendere le scansioni.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.
4. Selezionare il menu **azioni** e selezionare **scansione > Pausa scansione pianificata** o **scansione > Riprendi scansione pianificata**.

   Viene visualizzato un messaggio che indica che la successiva scansione pianificata è stata sospesa o ripresa.

Valutare i modelli di chat prima di creare una knowledge base

È possibile valutare i modelli di chat di base disponibili prima di creare una knowledge base per individuare il modello più adatto alla propria implementazione. Poiché il supporto dei modelli varia in base alla regione di AWS, fare riferimento a "Questa pagina di documentazione di AWS" per verificare quali modelli è possibile utilizzare nelle regioni in cui si prevede di implementare la propria knowledge base.

> Questa funzionalità è disponibile solo quando non è stata creata alcuna knowledge base — quando nella pagina di inventario della Knowledge base non esiste alcuna knowledge base.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina di inventario delle Knowledge base, vedrai l'opzione per selezionare il modello di chat sul lato

destro della pagina per il chatbot.

3. Selezionare il modello di chat dall'elenco e inserire una serie di domande nell'area degli annunci per vedere come risponde il chatbot.
4. Provare più modelli per individuare il modello più adatto all'implementazione.

Risultato

Utilizzate questo modello di chat quando create la vostra knowledge base.

Annullare la pubblicazione della knowledge base

Dopo aver pubblicato la knowledge base in modo che possa essere integrata con un'applicazione chatbot, puoi annullare la pubblicazione se desideri disabilitare l'applicazione chatbot dall'accesso alla knowledge base.

L'annullamento della pubblicazione della knowledge base impedisce il funzionamento di qualsiasi applicazione di chat. L'endpoint API univoco al quale la knowledge base è stata accessibile è disabilitato.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera annullare la pubblicazione.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione e l'elenco di tutte le origini dati incorporate.
4. Selezionare il menu **azioni** e selezionare **Annulla pubblicazione**.

Risultato

La knowledge base è disabilitata e non è più accessibile da un'applicazione chatbot.

Eliminare una knowledge base

Se non hai più bisogno di una knowledge base, puoi eliminarla. Quando si elimina una knowledge base, questa viene rimossa dalla fabbrica del carico di lavoro e il volume che contiene la knowledge base viene eliminato. Tutte le applicazioni o i chatbot che utilizzano la knowledge base smetteranno di funzionare. L'eliminazione di una knowledge base non è reversibile.

Quando si elimina una knowledge base, è inoltre necessario dissociarla da qualsiasi agente a cui è associata per eliminare completamente tutte le risorse associate alla knowledge base.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina dell'inventario delle Knowledge base, selezionare la knowledge base che si desidera eliminare.
3. Selezionare ••• e selezionare **Gestisci knowledge base**.
4. Selezionare il menu **azioni** e selezionare **Elimina knowledge base**.
5. Nella finestra di dialogo Elimina knowledge base, confermare che si desidera eliminarla e selezionare **Elimina**.

Risultato

La knowledge base viene rimossa dalla fabbrica del carico di lavoro e il volume associato viene eliminato.

Gestione di Amazon Q Business Connector

Dopo aver creato un connettore per Amazon Q Business, è possibile visualizzare i dettagli del connettore, modificare il connettore, integrare altre origini dati o eliminare il connettore.

Consente di visualizzare informazioni su un connettore

È possibile visualizzare informazioni sulle impostazioni di un connettore e sulle origini dati integrate.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dal menu di navigazione di fabbrica del carico di lavoro, selezionare **ai**.
3. Selezionare il connettore che si desidera visualizzare.
4. Per visualizzare i dettagli del connettore, selezionare ••• e selezionare **Gestisci connettore**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

   Il menu **azioni** consente di gestire il connettore se si desidera apportare modifiche.

Modificare un connettore

È possibile aggiornare un connettore modificando alcune impostazioni oppure aggiungere o rimuovere origini dati.

Ogni volta che si aggiungono, modificano o rimuovono origini dati dal connettore, GenAI deve inviare le informazioni dell'origine dati ad Amazon Q Business in modo che vengano riindicizzate. La sincronizzazione è incrementale, quindi Amazon Q Business elabora solo gli oggetti nel volume FSX per ONTAP che sono stati aggiunti, modificati o eliminati dall'ultima sincronizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina di inventario Knowledge base e connettori, selezionare il connettore che si desidera aggiornare.
3. Selezionare ••• e selezionare **Gestisci connettore**.

   Questa pagina visualizza lo stato pubblicato, lo stato di incorporazione delle origini dati, la modalità di incorporazione, l'elenco di tutte le origini dati incorporate e altro ancora.

4. Selezionare il menu **azioni** e selezionare **Modifica connettore**.
5. Nella pagina Modifica connettore è possibile modificare il nome del connettore, la descrizione, il modello di incorporazione, l'abilitazione delle barriere di protezione dei dati e il criterio snapshot utilizzato per il volume che contiene il connettore.

Ogni scansione dell'origine dati, che include l'inclusione, implica un costo. Se si abilitano le barriere di sicurezza dei dati dopo la creazione di un connettore, l'origine dati viene nuovamente sottoposta a scansione e i costi sono sostenuti.

6. Selezionare **Salva** dopo aver apportato le modifiche.

Aggiungere ulteriori origini dati a un connettore

È possibile incorporare ulteriori origini dati nel connettore per popolarlo con ulteriori dati dell'organizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina di inventario Knowledge base e connettori, selezionare il connettore in cui si desidera aggiungere l'origine dati.
3. Selezionare ••• e selezionare **Aggiungi origine dati**.
4. **Selezionare un file system**: Selezionare il file system FSX per ONTAP in cui risiedono i file di origine dati e selezionare **Avanti**.
5. **Selezionare un volume**: Selezionare il volume in cui risiedono i file di origine dati e selezionare **Avanti**.

   Quando si selezionano i file archiviati utilizzando il protocollo SMB, è necessario immettere le informazioni di Active Directory, che includono il dominio, l'indirizzo IP, il nome utente e la password.

6. **Selezionare un'origine dati**: Selezionare la posizione dell'origine dati in base alla posizione in cui sono stati salvati i file. Può trattarsi di un intero volume o solo di una cartella o sottocartella specifica del volume e selezionare **Avanti**.
7. **Configurazioni**: Configurare il modo in cui l'origine dati acquisisce le informazioni dai file e quali file include nelle scansioni:

   - **Definisci origine dati**: Nella sezione **strategia di Chunking**, definisci il modo in cui il motore GenAI divide il contenuto dell'origine dati in blocchi quando l'origine dati è integrata con una knowledge base. È possibile scegliere una delle seguenti strategie:
     - **Chunking a più frasi**: Organizza le informazioni dalla vostra origine dati in blocchi definiti dalle frasi. È possibile scegliere quante frasi compongono ciascun blocco (fino a 100).
     - **Chunking basato su sovrapposizione**: Organizza le informazioni dall'origine dati in blocchi definiti dai caratteri che possono sovrapporsi a blocchi adiacenti. È possibile scegliere la dimensione di ciascun frammento in caratteri e la sovrapposizione di ciascun frammento con i frammenti adiacenti. È possibile configurare una dimensione del frammento compresa tra 50 e 3000 caratteri e una percentuale di sovrapposizione compresa tra 1 e 99%.

       La scelta di un'elevata percentuale di sovrapposizione può aumentare notevolmente i requisiti di archiviazione con solo lievi miglioramenti nella precisione di recupero.

   - **Filtraggio dei file**: Configurare i file da includere nelle scansioni:
     - Nella sezione **supporto tipi di file**, scegliere di includere tutti i tipi di file o selezionare singoli tipi di file da includere nelle scansioni dell'origine dati.

       Se si includono immagini o file PDF, BlueXP workload factory per GenAI analizza il testo nelle immagini (incluse le immagini nei documenti PDF) e ciò comporta un costo maggiore.

Quando si includono dati di testo da immagini, GenAI non è in grado di mascherare informazioni di identificazione personale (PII) dall'immagine, poiché i dati di testo acquisiti vengono inviati dall'ambiente ad AWS. Tuttavia, una volta memorizzati i dati, tutte le PII vengono mascherate nel database GenAI.

> La scelta di includere i file di immagine nelle scansioni è correlata al modello di chat della Knowledge base. Se si includono file di immagine nelle scansioni, il modello di chat deve supportare le immagini. Se qui sono selezionati i tipi di file immagine, non è possibile passare dalla knowledge base a un modello di chat che non supporta i file immagine.

- Nella sezione **filtro ora modifica file**, scegliere di attivare o disattivare l'inclusione dei file in base al tempo di modifica. Se si attiva il filtraggio dell'ora di modifica, selezionare un intervallo di date dall'elenco.

> Se si includono file basati su un intervallo di date di modifica, non appena l'intervallo di date non è soddisfatto (i file non sono stati modificati entro l'intervallo di date specificato), i file saranno esclusi dalla scansione periodica e l'origine dati non includerà questi file.

8. Nella sezione **Permission Aware**, disponibile solo quando l'origine dati selezionata si trova su un volume che utilizza il protocollo SMB, è possibile attivare o disattivare le risposte in base alle autorizzazioni:
   - **Enabled**: Gli utenti del chatbot che accedono a questa knowledge base riceveranno solo risposte alle query provenienti da origini dati a cui hanno accesso.
   - **Disabled**: Gli utenti del chatbot riceveranno le risposte utilizzando il contenuto di tutte le origini dati integrate.
9. Selezionare **Aggiungi** per aggiungere questa origine dati alla knowledge base.

Risultato

L'origine dati è integrata nel connettore.

Sincronizzare le origini dati con un connettore

Le origini dati vengono sincronizzate automaticamente con il connettore associato una volta al giorno, in modo che eventuali modifiche all'origine dati vengano riflesse in Amazon Q Business. Se si apportano modifiche a una delle origini dati e si desidera sincronizzare (eseguire la scansione) immediatamente i dati, è possibile eseguire una sincronizzazione su richiesta.

La sincronizzazione è incrementale, quindi Amazon Q Business elabora solo gli oggetti nelle origini dati che sono stati aggiunti, modificati o eliminati dall'ultima sincronizzazione.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla scheda Knowledge base e connettori, selezionare il connettore che si desidera sincronizzare.
3. Selezionare ••• e selezionare **Gestisci connettore**.
4. Selezionare il menu **azioni** e selezionare **Esegui scansione ora**.

   Viene visualizzato un messaggio che indica che le origini dati sono in fase di scansione e un messaggio finale al termine della scansione.

Risultato

Il connettore è sincronizzato con le origini dati collegate e Amazon Q Business inizierà a utilizzare le informazioni più recenti provenienti dalle origini dati.

Consente di sospendere o riprendere una sincronizzazione pianificata

Se si desidera sospendere o riprendere la successiva sincronizzazione (scansione) delle origini dati, è possibile farlo in qualsiasi momento. Potrebbe essere necessario sospendere la successiva sincronizzazione pianificata se si intende apportare modifiche a un'origine dati e non si desidera che la sincronizzazione avvenga durante la finestra di modifica.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina di inventario del connettore, selezionare il connettore per il quale si desidera sospendere o riprendere le scansioni.
3. Selezionare ••• e selezionare **Gestisci connettore**.
4. Selezionare il menu **azioni** e selezionare **scansione > Pausa scansione pianificata** o **scansione > Riprendi scansione pianificata**.

   Viene visualizzato un messaggio che indica che la successiva scansione pianificata è stata sospesa o ripresa.

Eliminare un connettore

Se non è più necessario un connettore, è possibile eliminarlo. Quando si elimina un connettore, questo viene rimosso dalla fabbrica del carico di lavoro e il volume che contiene il connettore viene eliminato. L'eliminazione di un connettore non è reversibile.

Quando si elimina un connettore, è necessario scollegarlo da tutti gli agenti a cui è associato per eliminare completamente tutte le risorse associate al connettore.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Nella pagina di inventario Knowledge base e connettori, selezionare il connettore che si desidera eliminare.
3. Selezionare ••• e selezionare **Gestisci connettore**.
4. Selezionare il menu **azioni** e selezionare **Elimina connettore**.
5. Nella finestra di dialogo Elimina connettore, confermare che si desidera eliminarlo e selezionare **Elimina**.

Risultato

Il connettore viene rimosso dalla fabbrica del carico di lavoro e il volume associato viene eliminato.

Gestire le origini dati Genai

Dopo aver creato una knowledge base o un connettore utilizzando origini dati sul file system FSX per ONTAP, è possibile visualizzare i dettagli dell'origine dati, aggiornare o modificare il contenuto dell'origine dati, modificare le impostazioni dell'origine dati o eliminare l'origine dati.

Consente di visualizzare informazioni su un'origine dati

È possibile visualizzare le informazioni sul contenuto di un'origine dati ed è possibile visualizzarne lo stato di integrazione con la knowledge base o il connettore. Poiché le origini dati sono associate a una knowledge base o a un connettore, prima di poter visualizzare i dettagli delle origini dati è necessario scegliere la knowledge base o il connettore.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console" .
2. Dal menu di navigazione di fabbrica del carico di lavoro, selezionare **ai**.
3. Selezionare la knowledge base o il connettore in cui risiede l'origine dati, quindi selezionare ••• e selezionare **Gestisci knowledge base** o **Gestisci connettore**.

   Nella parte inferiore della pagina sono elencate le origini dati associate.

4. Espandere ogni riga selezionando i ⌄ per visualizzare informazioni dettagliate su ciascuna origine dati, come il file system FSX per ONTAP, il volume e il percorso in cui risiede l'origine dati.

   Elenca inoltre le informazioni di inclusione e indica se l'origine dati è attualmente incorporata nella knowledge base o nel connettore.

Modificare le impostazioni dell'origine dati

È possibile modificare le informazioni relative a un'origine dati integrata con una knowledge base o un connettore. La maggior parte delle informazioni viene risolta dopo aver aggiunto un'origine dati, ma è possibile apportare modifiche ad alcune configurazioni (ad esempio definizione dei blocchi o consapevolezza delle autorizzazioni).

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina dell'inventario delle Knowledge base, selezionare la knowledge base in cui risiede l'origine dati, quindi selezionare ••• e selezionare **Gestisci knowledge base**.

   Nella parte inferiore della pagina sono elencate le origini dati che fanno parte di questa knowledge base.

3. Nella riga dell'origine dati che si desidera modificare, selezionare ••• e selezionare **Modifica origine dati**.
4. Nella pagina Modifica origine dati, selezionare ⌄ per espandere la riga per la definizione del blocco.
5. Aggiornare le impostazioni per la strategia e la configurazione del chunking e la consapevolezza delle autorizzazioni (per i volumi SMB) e selezionare **Salva**.

Risultato

Le impostazioni dell'origine dati vengono aggiornate e il sistema ai sincronizza l'origine dati in modo che venga riindicizzata nella knowledge base.

Aggiornare il contenuto di un'origine dati esistente

È possibile modificare il contenuto di un'origine dati in qualsiasi momento per aggiungere o aggiornare i dati dell'organizzazione. Se questa origine dati viene utilizzata attivamente in una knowledge base, è necessario sincronizzare l'origine dati in modo che venga riindicizzata nella knowledge base. La sincronizzazione è incrementale, quindi Amazon Bedrock elabora solo gli oggetti nel volume FSX per ONTAP che sono stati aggiunti, modificati o eliminati dall'ultima sincronizzazione.

Le origini dati vengono sincronizzate automaticamente con la knowledge base una volta al giorno, in modo che eventuali modifiche all'origine dati vengano riflesse nel chatbot. Se si apportano modifiche a un'origine dati e si desidera sincronizzare immediatamente i dati, è possibile "eseguire una sincronizzazione su richiesta".

Eliminare un'origine dati

Se non è più necessario che un'origine dati faccia parte della knowledge base, è possibile eliminarla.

Fasi

1. Accedere a workload Factory utilizzando uno dei "esperienze di console".
2. Dalla pagina dell'inventario della Knowledge base, selezionare la knowledge base in cui risiede l'origine dati, quindi selezionare ••• e selezionare **Gestisci knowledge base**.

   Nella parte inferiore della pagina sono elencate le origini dati che fanno parte di questa knowledge base.
3. Nella riga dell'origine dati che si desidera eliminare, selezionare ••• e selezionare **Elimina origine dati**.
4. Nella finestra di dialogo Elimina origine dati, confermare che si desidera eliminarla e selezionare **Conferma**.

Risultato

L'origine dati viene rimossa dalla knowledge base e il sistema ai rimuove le informazioni indicizzate relative a questa origine dati dalla knowledge base. Le informazioni provenienti da tale fonte di dati non saranno più disponibili per i chatbot che utilizzano la knowledge base.

Conoscenza e supporto

Registrati per ricevere supporto per BlueXP workload Factory per GenAI

La registrazione del supporto è necessaria per ricevere supporto tecnico specifico per la fabbrica di carichi di lavoro BlueXP e per le sue soluzioni e servizi di storage. È necessario registrarsi per il supporto dalla console BlueXP , che è una console basata su Web separata da workload Factory.

La registrazione per il supporto non attiva il supporto NetApp per un file service provider cloud. Per il supporto tecnico relativo a un file service del cloud provider, alla sua infrastruttura o a qualsiasi soluzione che utilizzi il servizio, fare riferimento a "Assistenza" nella documentazione del workload factory relativa a quel prodotto.

"Amazon FSX per ONTAP"

Panoramica sulla registrazione del supporto

La registrazione dell'abbonamento al supporto con l'account ID (il numero di serie 960xxxxxxxx a 20 cifre situato nella pagina Support Resources di BlueXP) funge da singolo ID di abbonamento al supporto. Ogni abbonamento al supporto a livello di account BlueXP deve essere registrato.

La registrazione abilita funzionalità come l'apertura di ticket di supporto e la generazione automatica dei casi. La registrazione viene completata aggiungendo account del sito di supporto NetApp a BlueXP come descritto di seguito.

Registra il tuo account per il supporto NetApp

Per registrarti e attivare il diritto al supporto, un utente del tuo account deve associare un account del sito di supporto NetApp al proprio login BlueXP. La modalità di registrazione al supporto NetApp dipende dal fatto che si disponga già di un account NetApp Support Site (NSS).

Cliente esistente con un account NSS

Se sei un cliente NetApp con un account NSS, devi semplicemente registrarti per ricevere supporto tramite BlueXP.

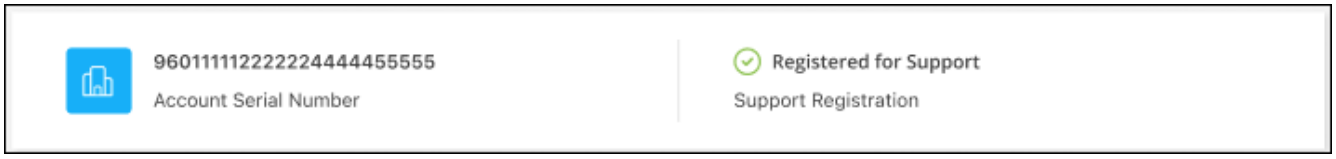
Fasi

1. Nella parte superiore destra della console di fabbrica del carico di lavoro, selezionare **Guida > supporto**.

   La selezione di questa opzione apre la console BlueXP, una nuova scheda del browser e carica la dashboard di supporto.

2. Nella parte superiore destra della console BlueXP, selezionare l'icona Impostazioni e selezionare **credenziali**.
3. Selezionare **User Credentials** (credenziali utente).
4. Selezionare **Aggiungi credenziali NSS** e seguire la richiesta di autenticazione del sito di supporto NetApp.
5. Per confermare che la procedura di registrazione è stata eseguita correttamente, selezionare l'icona Guida e selezionare **supporto**.

La pagina **risorse** dovrebbe mostrare che il tuo account è registrato per il supporto.



Si noti che gli altri utenti di BlueXP non visualizzeranno lo stesso stato di registrazione del supporto se non hanno associato un account del sito di supporto NetApp al proprio login BlueXP. Tuttavia, ciò non significa che il tuo account BlueXP non sia registrato per il supporto. Se un utente dell'account ha seguito questa procedura, l'account è stato registrato.

Cliente esistente ma nessun account NSS

Se sei un cliente NetApp con licenze e numeri di serie esistenti ma *no* account NSS, devi creare un account NSS e associarlo al tuo login BlueXP.

Fasi

1. Creare un account per il sito del supporto NetApp completando il "Modulo di registrazione per l'utente del sito di supporto NetApp"
   a. Assicurarsi di selezionare il livello utente appropriato, che in genere è **cliente/utente finale NetApp**.
   b. Assicurarsi di copiare il numero di serie dell'account BlueXP (960xxxx) utilizzato in precedenza per il campo del numero di serie. In questo modo, l'elaborazione dell'account sarà più rapida.
2. Associa il nuovo account NSS al tuo login BlueXP completando la procedura descritta in Cliente esistente con un account NSS.

Novità di NetApp

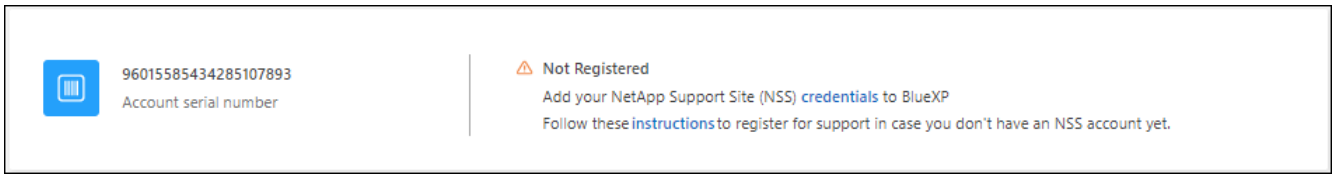
Se sei nuovo di NetApp e non disponi di un account NSS, segui i passaggi riportati di seguito.

Fasi

1. Nella parte superiore destra della console di fabbrica del carico di lavoro, selezionare **Guida > supporto**.

   La selezione di questa opzione apre la console BlueXP, una nuova scheda del browser e carica la dashboard di supporto.

2. Individuare il numero di serie dell'ID account nella pagina risorse di supporto.



3. Accedere a "Sito per la registrazione del supporto NetApp" e selezionare **non sono un cliente NetApp registrato**.
4. Compilare i campi obbligatori (con asterischi rossi).
5. Nel campo **Product Line**, selezionare **Cloud Manager**, quindi selezionare il provider di fatturazione appropriato.
6. Copia il numero di serie del tuo account dal punto 2 precedente, completa il controllo di sicurezza, quindi

conferma di aver letto la Global Data Privacy Policy di NetApp.

Viene immediatamente inviata un'e-mail alla casella di posta fornita per finalizzare questa transazione sicura. Controllare le cartelle di spam se l'e-mail di convalida non arriva in pochi minuti.

7. Confermare l'azione dall'interno dell'e-mail.

   La conferma invia la tua richiesta a NetApp e ti consiglia di creare un account NetApp Support Site.

8. Creare un account per il sito del supporto NetApp completando il "Modulo di registrazione per l'utente del sito di supporto NetApp"
   a. Assicurarsi di selezionare il livello utente appropriato, che in genere è **cliente/utente finale NetApp**.
   b. Assicurarsi di copiare il numero di serie dell'account (960xxxx) utilizzato in precedenza per il campo del numero di serie. In questo modo, l'elaborazione dell'account sarà più rapida.

Al termine

NetApp dovrebbe contattarti durante questo processo. Si tratta di un esercizio di assunzione per i nuovi utenti.

Una volta ottenuto l'account per il sito di supporto NetApp, associare l'account al login BlueXP completando la procedura descritta in Cliente esistente con un account NSS.

Risoluzione dei problemi relativi a GenAI

Scopri come risolvere alcuni problemi comuni che potresti incontrare.

Problemi e soluzioni comuni

Se si verifica uno di questi problemi, è possibile utilizzare la procedura nella colonna soluzione alternativa per tentare di risolverlo.

Area	Problema	Causa	Soluzione alternativa
Implementazione	La distribuzione non riesce perché il volume esiste già.	La workload factory di BlueXP per Genai deve creare un nuovo volume durante il processo di implementazione, ma un volume esiste già utilizzando il nome specificato.	Specificare un nome univoco da utilizzare per il nuovo volume e provare a eseguire nuovamente la distribuzione.
Implementazione	L'implementazione non riesce perché la fabbrica del carico di lavoro BlueXP per GenAI non è in grado di montare il volume.	Una o più porte in entrata richieste per FSX per NetApp ONTAP vengono chiuse o filtrate.	Aprire le seguenti porte in entrata:

| Protocollo | Porta | Scopo

| Tutti gli ICMP | Tutto | Eseguire il ping dell'istanza

| HTTPS | 443 | Accesso dal connettore alla LIF di gestione di fsxadmin per inviare chiamate API a FSX

| SSH | 22 | Accesso SSH all'indirizzo IP della LIF di gestione del cluster o di una LIF di gestione dei nodi

| TCP | 111 | Chiamata a procedura remota per NFS

| TCP | 139 | Sessione del servizio NetBIOS per CIFS

| TCP | 161-162 | Protocollo di gestione di rete semplice

| TCP | 445 | Microsoft SMB/CIFS su TCP con frame NetBIOS

| TCP | 635 | Montaggio NFS

| TCP | 749 | Kerberos

| TCP | 2049 | Daemon del server NFS

| TCP | 3260 | Accesso iSCSI tramite LIF dei dati iSCSI

| TCP | 4045 | Daemon di blocco NFS

| TCP | 4046 | Network status monitor per NFS

| TCP | 10000 | Backup con NDMP

| TCP | 11104 | Gestione delle sessioni di comunicazione tra cluster per SnapMirror

| TCP | 11105 | Trasferimento dei dati SnapMirror con LIF intercluster

| UDP | 111 | Chiamata a procedura remota per NFS

| UDP | 161-162 | Protocollo di gestione di rete semplice

| UDP | 635 | Montaggio NFS

| UDP | 2049 | Daemon del server NFS

| UDP | 4045 | Daemon di blocco NFS

| UDP | 4046 | Network status monitor per NFS

| UDP | 4049 | Protocollo NFS rquotad

Manutenzione	Il motore ai non si avvia e viene visualizzato l'errore "errore di istanza del motore ai" nella pagina **Knowledge base**.	L'istanza del motore ai era danneggiata o non esiste.	Selezionare il pulsante **Ricostruisci**. La farm di workload BlueXP per Genai ricostruisce l'infrastruttura e visualizza l'avanzamento della ricostruzione. Al termine, le knowledge base vengono ricollegate all'infrastruttura ricostruita e viene visualizzato l'elenco delle knowledge base.
Manutenzione	Il motore ai non si avvia e viene visualizzato l'errore "l'istanza del motore Genai è ferma" nella pagina **Knowledge base**.	L'istanza del motore ai non è in esecuzione.	USA la Console di gestione AWS o l'interfaccia a riga di comando AWS per avviare l'istanza del motore ai.
Manutenzione	Il motore ai non si avvia e viene visualizzato l'errore "il server del motore Genai non risponde" nella pagina **Knowledge base**.	L'istanza del motore ai non risponde.	Attenersi alla seguente procedura di ripristino: **Fasi** 1. Modificare il gruppo di sicurezza dell'istanza del motore GenAI per abilitare l'accesso SSH all'istanza del motore GenAI. 2. Accedere all'istanza utilizzando SSH. 3. Eseguire il seguente comando: `docker-compose up`

Manutenzione	L'istanza Docker back-end utilizzata dalla farm di workload BlueXP per Genai non si è avviata.	Eliminazione del volume e riavvio dell'istanza EC2.	Attenersi alla seguente procedura di ripristino:
			
			Fasi
			1. Crea un nuovo volume in FSX per NetApp ONTAP. Ad esempio, il nome del volume può essere `netapp_ai` e il percorso del volume può essere `/netapp_ai`.
			2. SSH all'istanza di Amazon EC2.
			3. Elencare i volumi:
			`docker volume list`
			4. Rimuovere il vecchio volume:
			`docker volume rm ec2-user_persistent_folder`
			5. Aprire il `docker-compose.yml` file utilizzando un editor di testo.
			6. Nella `volumes` sezione , modificare il percorso del dispositivo sul nuovo percorso del volume. Ad esempio:

Manutenzione	L'istanza Docker back-end utilizzata dalla farm di workload BlueXP per Genai non si è avviata.	Il volume root è stato eliminato.	Creare un volume con nome e percorso, quindi riavviare l'istanza Docker back-end da Amazon EC2.
Manutenzione	L'istanza Docker back-end utilizzata dalla farm di workload BlueXP per Genai non si è avviata.	Il volume root è stato eliminato.	Creare un volume con nome e percorso, quindi riavviare l'istanza Docker back-end da Amazon EC2.

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

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

Ottieni aiuto con la fabbrica di workload BlueXP per Genai

NetApp fornisce supporto per la fabbrica di carichi di lavoro BlueXP e i suoi servizi cloud in diversi modi. Sono disponibili opzioni complete di supporto autonomo gratuito 24 ore su 24, 7 giorni su 7, come articoli della knowledge base (KB) e un forum della community. La registrazione al supporto include il supporto tecnico remoto via web ticketing.

Ottieni supporto per FSX per ONTAP

Per il supporto tecnico relativo a FSX per ONTAP, alla sua infrastruttura o a qualsiasi soluzione che utilizza il servizio, fare riferimento a "come ottenere assistenza" nella documentazione della workload Factory relativa al prodotto.

"Amazon FSX per ONTAP"

Per ricevere supporto tecnico specifico di workload Factory e delle relative soluzioni e servizi di storage, utilizzare le opzioni di supporto descritte di seguito.

Utilizzare le opzioni di supporto automatico

Queste opzioni sono disponibili gratuitamente, 24 ore su 24, 7 giorni su 7:

- Documentazione

  La documentazione di fabbrica del carico di lavoro attualmente visualizzata.

- "Knowledge base"

  Ricerca nella knowledge base di workload Factory per trovare articoli utili per la risoluzione dei problemi.

- "Community"

  Unisciti alla community dedicata alla fabbrica dei carichi di lavoro per seguire le discussioni in corso o crearne di nuovi.

Crea un caso con il supporto NetApp

Oltre alle opzioni di supporto autonomo sopra descritte, puoi collaborare con uno specialista del supporto NetApp per risolvere eventuali problemi dopo l'attivazione del supporto.

Prima di iniziare

Per utilizzare la funzionalità **Crea un caso**, è necessario prima registrarsi per ricevere assistenza. Associare le credenziali del sito di supporto NetApp all'accesso alla fabbrica del carico di lavoro. "Scopri come registrarti per il supporto".

Fasi

1. Nella parte superiore destra della console di fabbrica del carico di lavoro, selezionare **Guida > supporto**.

   La selezione di questa opzione apre la console BlueXP, una nuova scheda del browser e carica la dashboard di supporto.

2. Nella pagina **risorse**, scegliere una delle opzioni disponibili in supporto tecnico:

   a. Selezionare **Chiamateci** se si desidera parlare con qualcuno al telefono. Viene visualizzata una pagina su netapp.com che elenca i numeri di telefono che è possibile chiamare.

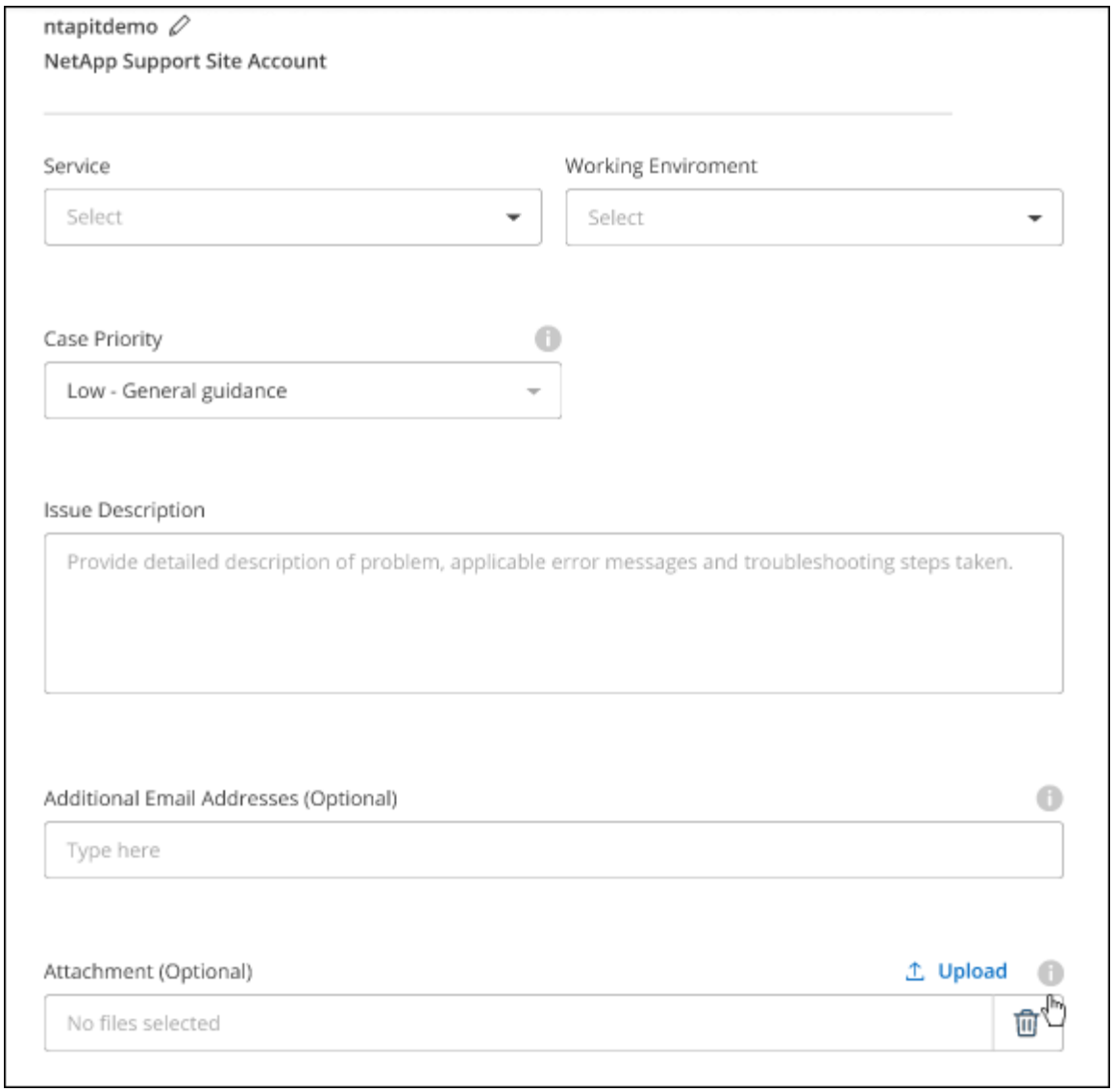
   b. Selezionare **Crea un caso** per aprire un ticket con uno specialista del supporto NetApp:

      - **Servizio**: Selezionare **fabbrica workload**.
      - **Priorità caso**: Scegliere la priorità per il caso, che può essere bassa, Media, alta o critica.

        Per ulteriori informazioni su queste priorità, passare il mouse sull'icona delle informazioni accanto al nome del campo.

      - **Descrizione del problema**: Fornire una descrizione dettagliata del problema, inclusi eventuali messaggi di errore o procedure di risoluzione dei problemi che sono state eseguite.
      - **Indirizzi e-mail aggiuntivi**: Inserisci indirizzi e-mail aggiuntivi se desideri informare qualcun altro del problema.
      - **Allegato (opzionale)**: Carica fino a cinque allegati, uno alla volta.

        Gli allegati sono limitati a 25 MB per file. Sono supportate le seguenti estensioni di file: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx e csv.



Al termine

Viene visualizzata una finestra a comparsa con il numero del caso di supporto. Uno specialista del supporto NetApp esaminerà il tuo caso e ti contatterà al più presto.

Per una cronologia dei casi di supporto, è possibile selezionare **Impostazioni > sequenza temporale** e cercare le azioni "Crea caso di supporto". Un pulsante all'estrema destra consente di espandere l'azione per visualizzare i dettagli.

È possibile che venga visualizzato il seguente messaggio di errore quando si tenta di creare un caso:

"Non sei autorizzato a creare un caso per il servizio selezionato"

Questo errore potrebbe indicare che l'account NSS e la società di registrazione a cui è associato non sono la stessa società di registrazione per il numero di serie dell'account BlueXP (ad es. 960xxxx) o il numero di serie dell'ambiente di lavoro. È possibile richiedere assistenza utilizzando una delle seguenti opzioni:

- Utilizza la chat integrata nel prodotto
- Inviare un caso non tecnico all'indirizzo https://mysupport.netapp.com/site/help

Gestire i casi di supporto (anteprima)

È possibile visualizzare e gestire i casi di supporto attivi e risolti direttamente da BlueXP. Puoi gestire i casi associati al tuo account NSS e alla tua azienda.

La gestione del caso è disponibile come anteprima. Intendiamo perfezionare questa esperienza e aggiungere miglioramenti alle prossime release. Inviaci un feedback utilizzando la chat in-product.

Tenere presente quanto segue:

- La dashboard di gestione dei casi nella parte superiore della pagina offre due visualizzazioni:
  - La vista a sinistra mostra il totale dei casi aperti negli ultimi 3 mesi dall'account NSS dell'utente fornito.
  - La vista a destra mostra il totale dei casi aperti negli ultimi 3 mesi a livello aziendale in base all'account NSS dell'utente.

  I risultati della tabella riflettono i casi correlati alla vista selezionata.
- È possibile aggiungere o rimuovere colonne di interesse e filtrare il contenuto di colonne come priorità e Stato. Altre colonne offrono funzionalità di ordinamento.

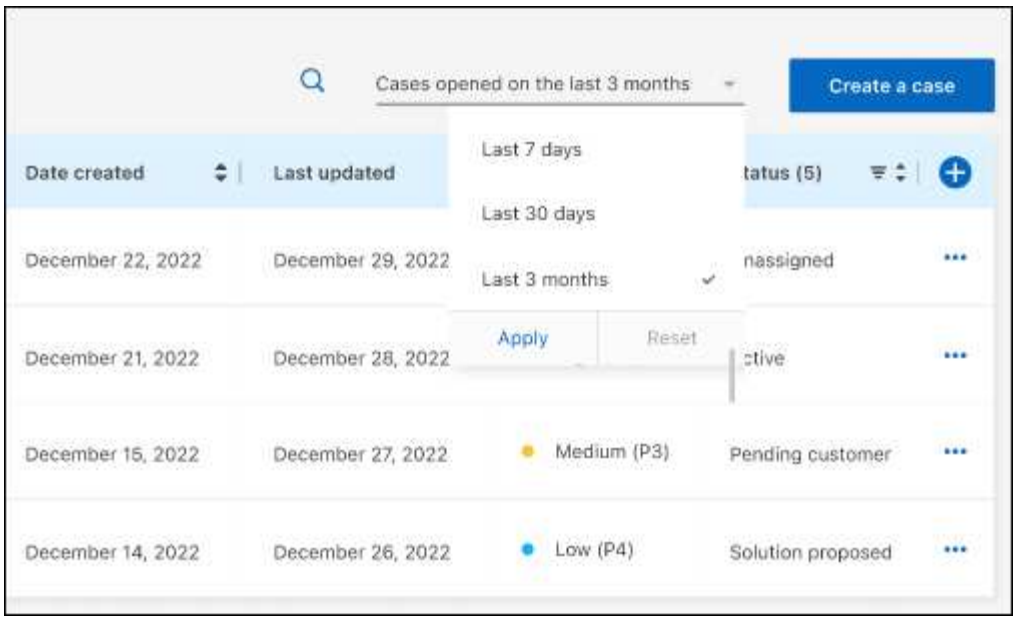
  Per ulteriori informazioni, consulta la procedura riportata di seguito.
- A livello di caso, offriamo la possibilità di aggiornare le note del caso o chiudere un caso che non è già in stato chiuso o in attesa di chiusura.

Fasi

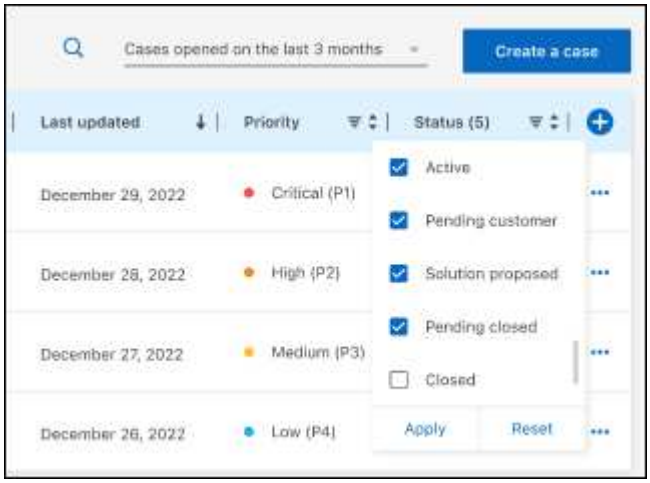
1. Nella parte superiore destra della console di fabbrica del carico di lavoro, selezionare **Guida > supporto**.

   La selezione di questa opzione apre la console BlueXP, una nuova scheda del browser e carica la dashboard di supporto.
2. Selezionare **Gestione casi** e, se richiesto, aggiungere l'account NSS a BlueXP.

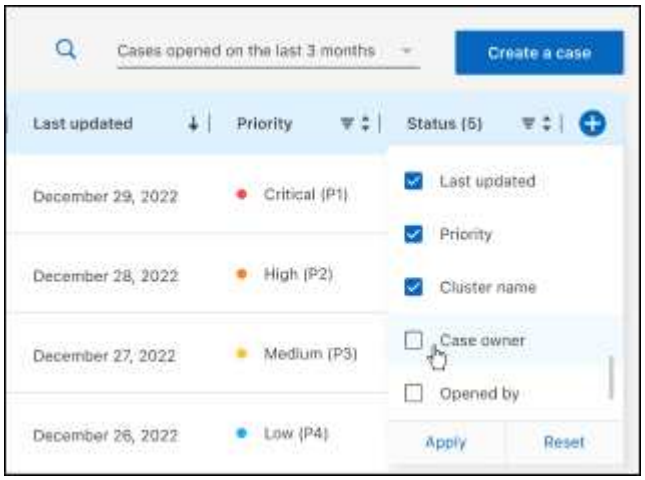
   La pagina **Gestione del caso** mostra i casi aperti relativi all'account NSS associato all'account utente BlueXP. Si tratta dello stesso account NSS visualizzato nella parte superiore della pagina **gestione NSS**.
3. Se si desidera, modificare le informazioni visualizzate nella tabella:
   - In **Organization's Cases** (casi dell'organizzazione), selezionare **View** (Visualizza) per visualizzare tutti i casi associati alla società.
   - Modificare l'intervallo di date scegliendo un intervallo di date esatto o scegliendo un intervallo di tempo diverso.



- Filtrare il contenuto delle colonne.



- Modificare le colonne visualizzate nella tabella selezionando e scegliendo le colonne che si desidera visualizzare.

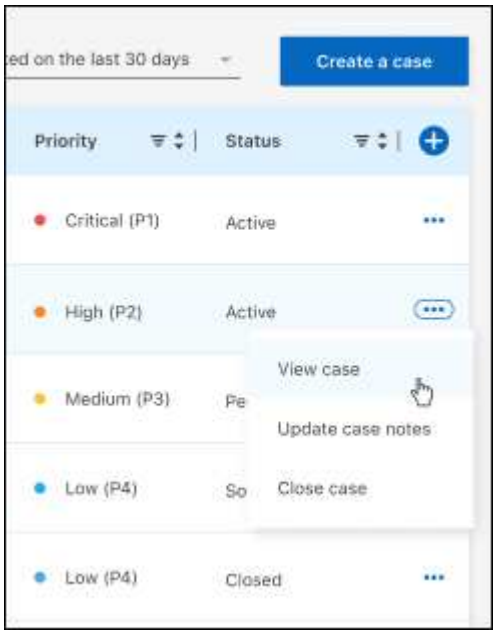


4. Gestire un caso esistente selezionando ••• e selezionando una delle opzioni disponibili:

   - **Visualizza caso**: Visualizza tutti i dettagli relativi a un caso specifico.
   - **Aggiorna note sul caso**: Fornisci ulteriori dettagli sul problema oppure seleziona **carica file** per allegare fino a un massimo di cinque file.

     Gli allegati sono limitati a 25 MB per file. Sono supportate le seguenti estensioni di file: Txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx e csv.

   - **Chiudi caso**: Fornisci i dettagli sul motivo per cui stai chiudendo il caso e seleziona **Chiudi caso**.



Fabbrica di workload BlueXP per avvisi legali Genai

Le note legali forniscono l'accesso a dichiarazioni di copyright, marchi, brevetti e altro ancora.

Copyright

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

Marchi

NETAPP, il logo NETAPP e i marchi elencati nella pagina dei marchi NetApp sono marchi di NetApp, Inc. Altri nomi di società e prodotti potrebbero essere marchi dei rispettivi proprietari.

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

Brevetti

Un elenco aggiornato dei brevetti di proprietà di NetApp è disponibile all'indirizzo:

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

Direttiva sulla privacy

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

Open source

I file di avviso forniscono informazioni sul copyright e sulle licenze di terze parti utilizzate nel software NetApp.

"Fabbrica dei carichi di lavoro BlueXP"

Informazioni sul copyright

Informazioni sul marchio commerciale