Documentation de NetApp Workload Factory pour GenAI

GenAI

NetApp
October 06, 2025

Sommaire

Documentation de NetApp Workload Factory pour GenAI

Notes de mise à jour

Nouveautés de NetApp Workload Factory pour GenAI

Découvrez les nouveautés de la fonctionnalité de charges de travail d'IA générative de Workload Factory.

05 octobre 2025

BlueXP workload factory désormais NetApp Workload Factory

BlueXP a été renommé et repensé pour mieux refléter le rôle qu'il joue dans la gestion de votre infrastructure de données. Par conséquent, BlueXP workload factory a été renommée NetApp Workload Factory.

Prise en charge de l'ajout de sources de données NFS/SMB génériques dans les connecteurs NetApp pour Amazon Q Business

À l'aide de l'API Workload Factory, vous pouvez désormais ajouter une source de données à partir d'un partage NFSv3, NFSv4 ou SMB générique à un connecteur NetApp pour Amazon Q Business. Cela vous permet d'inclure des fichiers stockés sur des volumes hébergés par des systèmes de fichiers autres qu'Amazon Amazon FSx for NetApp ONTAP.

"Créer un connecteur NetApp pour Amazon Q Business"

"Ajouter des sources de données à un connecteur"

Configuration avancée du chat pour les bases de connaissances

Vous pouvez désormais configurer des paramètres de chat avancés applicables au modèle de chat de la base de connaissances, tels que la longueur de la réponse, la température, les paramètres de raisonnement, etc. Certains de ces paramètres, tels que les paramètres de récence et de temps de modification, les paramètres de récupération avancés et l'invite système, ne sont disponibles qu'à l'aide de l'API Workload Factory.

"Créer une base de connaissances GenAI"

La sélection du type d'inférence est désormais prise en charge pour les modèles d'intégration, de discussion et de reclassement

Si le modèle d'intégration, de chat ou de reclassement que vous avez choisi dispose de paramètres d'inférence, vous pouvez désormais sélectionner un type d'inférence. Cela vous permet de mieux adapter les performances du chatbot et les besoins en ressources à vos besoins.

"Créer une base de connaissances GenAI"

03 août 2025

Stockage sécurisé des résultats de données structurées

Si les résultats de la requête du chatbot contiennent des données structurées, GenAI peut stocker les résultats dans un compartiment Amazon S3. Lorsque ces résultats sont stockés dans un bucket S3, vous pouvez les télécharger à l'aide du lien de téléchargement dans la session de chat.

"Créer une base de connaissances GenAI"

Disponibilité du serveur MCP

NetApp fournit désormais un serveur Model Context Protocol (MCP) avec NetApp Workload Factory pour GenAI. Vous pouvez installer le serveur localement pour permettre aux clients MCP externes de découvrir et de récupérer les résultats de requête à partir d'une base de connaissances GenAI.

"Serveur NetApp Workload Factory GenAI MCP"

29 juin 2025

Prise en charge des sources de données hébergées sur des systèmes de fichiers NFS/SMB génériques

Vous pouvez désormais ajouter une source de données à partir d'un partage SMB ou NFS générique. Cela vous permet d'inclure des fichiers stockés sur des volumes hébergés par des systèmes de fichiers autres qu'Amazon FSx pour NetApp ONTAP.

"Ajouter des sources de données à une base de connaissances"

"Ajouter des sources de données à un connecteur"

03 juin 2025

Tracker disponible pour les opérations de surveillance et de suivi

La fonctionnalité de surveillance Tracker est désormais disponible dans GenAI. Vous pouvez utiliser Tracker pour surveiller et suivre la progression et l'état des opérations en attente, en cours et terminées, consulter les détails des tâches et sous-tâches opérationnelles, diagnostiquer les problèmes ou les échecs, modifier les paramètres des opérations ayant échoué et relancer les opérations ayant échoué.

"Surveillez les opérations de charge de travail avec Tracker dans NetApp Workload Factory"

Choisir un modèle de reranker pour une base de connaissances

Vous pouvez désormais améliorer la pertinence des résultats de requêtes reclassés en sélectionnant un modèle de reranking spécifique à utiliser avec une base de connaissances. GenAI prend en charge les modèles Cohere Rerank et Amazon Rerank.

"Créer une base de connaissances GenAI"

04 mai 2025

Prise en charge du connecteur NetApp pour Amazon Q Business

Cette version de GenAI introduit la prise en charge de NetApp Connector pour Amazon Q Business, vous permettant de créer des connecteurs pour Amazon Q Business. Profitez rapidement et facilement de l'assistant IA Amazon Q Business avec moins de configuration initiale qu'en construisant une base de connaissances GenAI pour le socle Amazon.

"Créer un connecteur NetApp pour Amazon Q Business"

Prise en charge améliorée des modèles de chat

GenAI prend désormais en charge les modèles de chat supplémentaires suivants pour les bases de connaissances :

- "Modèles d'IA Mistral"
- "Modèles texte Amazon Titan"
- "Modèles META Llama"
- "Modèles Jamba 1.5"
- "Modèles Cohere Command"
- "Modèles DeepSeek"

GenAI prend en charge les modèles de chaque fournisseur pris en charge par Amazon Bedrock : "Modèles de base pris en charge dans Amazon Bedrock"

"Créer une base de connaissances GenAI"

Mise à jour de la terminologie des autorisations

L'interface utilisateur et la documentation de Workload Factory utilisent désormais « lecture seule » pour faire référence aux autorisations de lecture et « lecture/écriture » pour faire référence aux autorisations d'automatisation.

02 mars 2025

Améliorations du chatbot intégrées

Vous pouvez désormais copier des questions et des réponses directement dans le presse-papiers, ajuster la taille de la fenêtre de discussion et modifier son titre. De plus, les réponses de discussion peuvent maintenant inclure des tableaux, qui sont également copyables.

"Testez une base de connaissances GenAI"

Aide à la citation de réponse par chat

Les réponses à la discussion incluent maintenant des citations qui répertorient les fichiers et les segments de données utilisés pour générer la réponse.

"Testez une base de connaissances GenAI"

Prise en charge améliorée des types de fichiers

Cette version de GenAI fournit une prise en charge améliorée des fichiers :

- Les modèles de chat offrent une prise en charge CSV améliorée. Cela permet d'obtenir des réponses plus utiles lors de l'interrogation de données à partir de fichiers CSV.
- GenAI peut désormais ingérer des fichiers Apache parquet à partir de sources de données.
- GenAI prend désormais en charge l'ingestion de fichiers Microsoft Word DOCX contenant des images. Les images intégrées aux documents DOCX sont numérisées et les informations textuelles des images intégrées sont incluses dans les réponses aux requêtes de la base de connaissances.

"Formats de fichiers de source de données pris en charge"

02 février 2025

Prise en charge des modèles de base Amazon Nova

GenAI prend désormais en charge les modèles de base Amazon Nova. Amazon Nova micro, Amazon Nova Lite et Amazon Nova Pro sont pris en charge.

"Exigences GenAI"

Filtrage de type de fichier pour les sources de données

GenAI prend désormais en charge la sélection de types de fichiers spécifiques à inclure dans l'analyse de la source de données lorsque vous ajoutez une source de données.

"Ajoutez des sources de données à la base de connaissances"

Filtrage de date de modification de fichier pour les sources de données

GenAI prend désormais en charge le filtrage des fichiers à inclure dans l'analyse de la source de données par date de modification lorsque vous ajoutez une source de données. Vous pouvez choisir une plage de dates de modification pour les fichiers inclus.

"Ajoutez des sources de données à la base de connaissances"

Prise en charge des fichiers image et prise en charge améliorée des fichiers PDF

GenAI prend désormais en charge l'amélioration des réponses aux requêtes de la base de connaissances grâce à des informations tirées des images et des descriptions de graphiques, ainsi que du texte du document, ce qui permet d'obtenir des réponses plus riches et de meilleure qualité. GenAI peut désormais numériser des fichiers image et des images dans des fichiers PDF (également appelés prise en charge de fichiers multimodaux). Si vous choisissez de numériser des images ou des fichiers PDF, le texte des images (y compris les images intégrées aux documents PDF) est numérisé dans la source de données et les informations issues des numérisations sont incluses dans les réponses aux requêtes de la base de connaissances.

"Ajoutez des sources de données à la base de connaissances"

Recherche hybride et support de redirection

GenAI peut désormais améliorer considérablement la pertinence et l'exactitude des résultats de recherche grâce à la recherche hybride et au réclassement des résultats. La recherche hybride combine les points forts de la recherche classique basée sur des mots-clés avec des techniques de recherche sémantique basées sur des vecteurs denses et avancées. Les résultats standard de la recherche par mot-clé sont complétés par des correspondances étroites et la nuance linguistique, ce qui améliore la pertinence. GenAI affine ensuite ces résultats en utilisant des modèles de classement avancés, tels que Cohere Rerank et Amazon Rerank, et donne les résultats les plus pertinents. Cette fonctionnalité est disponible pour les bases de connaissances nouvellement créées.

"En savoir plus sur NetApp Workload Factory pour GenAI"

05 janvier 2025

Nom de snapshot personnalisé

Vous pouvez maintenant fournir un nom d'instantané pour un instantané ad hoc.

"Protégez une base de connaissances avec des instantanés"

Nom d'instance de moteur d'IA personnalisé

Vous pouvez maintenant attribuer un nom personnalisé à l'instance du moteur d'IA pendant le déploiement.

"Déployez l'infrastructure GenAI"

Reconstruction de l'infrastructure GenAI corrompue ou manquante

Si votre instance de moteur d'IA est corrompue ou supprimée d'une manière ou d'une autre, vous pouvez laisser Workload Factory la reconstruire pour vous. Workload Factory rattache automatiquement vos bases de connaissances à l'infrastructure une fois la reconstruction terminée, afin qu'elles soient prêtes à être utilisées.

"Dépannage"

01 décembre 2024

Cloner une base de connaissances à partir d'un instantané

NetApp Workload Factory pour GenAI prend désormais en charge le clonage d'une base de connaissances à partir d'un instantané. Cela permet une récupération rapide des bases de connaissances et la création de nouvelles bases de connaissances avec des sources de données existantes, et aide à la récupération et au développement des données.

"Cloner une base de connaissances"

Détection et réplication des clusters ONTAP sur site

Découvrez et répliquez les données du cluster ONTAP sur site vers un système de fichiers FSx pour ONTAP afin qu'elles puissent être utilisées pour enrichir les bases de connaissances de l'IA. Tous les flux de travail de découverte et de réplication sur site sont possibles à partir du nouveau menu **On-Premises ONTAP** dans l'inventaire de stockage.

"Découvrez un cluster ONTAP sur site"

3 novembre 2024

Masquez les informations à caractère personnel avec des limites de données

La charge de travail Generative AI introduit la fonctionnalité de garde-fous de données, optimisée par la classification de la console NetApp . La fonctionnalité de protection des données identifie et masque les informations personnelles identifiables (PII), vous aidant ainsi à maintenir la conformité et à renforcer la sécurité de vos données organisationnelles sensibles.

"Créer une base de connaissances GenAI"

"En savoir plus sur la classification de la console NetApp"

29 septembre 2024

Prise en charge des snapshots et des restaurations pour les volumes de la base de connaissances

Vous pouvez désormais protéger les données de vos workloads d'IA générative en utilisant une copie instantanée d'une base de connaissances. Vous pouvez ainsi protéger vos données contre les pertes accidentelles ou tester les modifications apportées aux paramètres de la base de connaissances. Vous pouvez à tout moment restaurer la version précédente du volume de la base de connaissances.

"Prenez un instantané d'un volume de la base de connaissances"

"Restaurer un snapshot d'un volume de la base de connaissances"

Mettre en pause les analyses planifiées

Vous pouvez maintenant interrompre les analyses de sources de données planifiées. Par défaut, les workloads d'IA génératifs analysent chaque source de données tous les jours afin d'ingérer de nouvelles données dans chaque base de connaissances. Si vous ne souhaitez pas que les dernières modifications soient ingérées (pendant le test ou lors de la restauration d'un instantané, par exemple), vous pouvez interrompre les analyses programmées et les reprendre à tout moment.

"Gérer les bases de connaissances"

Les volumes de protection des données sont désormais pris en charge pour les bases de connaissances

Lors de la sélection d'un volume de la base de connaissances, vous pouvez désormais choisir un volume de protection des données faisant partie d'une relation de réplication NetApp SnapMirror. Cela vous permet de stocker des bases de connaissances sur des volumes déjà protégés par la réplication SnapMirror.

"Identifiez les sources de données à intégrer dans votre base de connaissances"

1er septembre 2024

Stratégies de chaquage supplémentaires

Les workloads d'IA génératifs prennent désormais en charge le traitement en plusieurs phrases et le processus de suppression basé sur les chevauchements pour les sources de données.

Volume dédié pour chaque base de connaissances

Les workloads d'IA génératifs créent désormais un volume Amazon FSX pour NetApp ONTAP dédié à chaque nouvelle base de connaissances, en activant des règles Snapshot individuelles pour chaque base de connaissances et en améliorant la protection contre les défaillances et les empoisonnements de données.

4 août 2024

Intégration des journaux Amazon CloudWatch

Les workloads d'IA génératifs sont désormais intégrés aux journaux Amazon CloudWatch, ce qui vous permet de surveiller les fichiers journaux des workloads d'IA génératifs.

Exemple d’application chatbot

L’exemple d’application NetApp Workload Factory GenAI vous permet de tester l’authentification et la récupération à partir de votre base de connaissances NetApp Workload Factory publiée en interagissant directement avec elle dans une application de chatbot Web.

7 juillet 2024

Version initiale de Workload Factory pour GenAI

La version initiale permet de développer une base de connaissances personnalisée en intégrant les données de votre entreprise. La base de connaissances est accessible par une application de chatbot pour vos utilisateurs. Cette fonctionnalité garantit des réponses précises et pertinentes aux questions spécifiques à l’entreprise, améliorant ainsi la satisfaction et la productivité de tous vos utilisateurs.

En savoir plus sur NetApp Workload Factory pour GenAI

En savoir plus sur NetApp Workload Factory pour GenAI

NetApp Workload Factory pour GenAI vous permet d'intégrer les systèmes de fichiers Amazon FSx for NetApp ONTAP aux modèles de base GenAI. Cela vous offre un stockage hautes performances avec un riche ensemble de fonctionnalités de protection, de sécurité et d'optimisation des coûts pour vos ensembles de données IA.

Qu'est-ce que NetApp Workload Factory pour GenAI ?

NetApp Workload Factory pour GenAI vous permet d'utiliser vos sources de données d'entreprise sur Amazon FSx for NetApp ONTAP avec des applications Generative AI. En utilisant la génération augmentée par récupération (RAG), vous pouvez rapidement connecter des sources de données à des modèles de base disponibles via Amazon Bedrock ou Amazon Q Business pour développer des applications basées sur l'IA générative telles que des assistants virtuels, des chatbots de questions-réponses, la synthèse de documents, la création de contenu, etc.

L'utilisation de l'IA générative avec vos données d'entreprise vous permet de mettre à profit vos propres connaissances et votre expertise, au lieu de vous appuyer sur l'intelligence du modèle basée sur les données publiques sur lesquelles les modèles ont été entraînés. L'utilisation de RAG pour personnaliser les modèles garantit des réponses précises et pertinentes aux questions spécifiques à l'entreprise, améliorant ainsi la productivité et l'efficacité des utilisateurs de vos applications à l'aide de Generative ai.

Le développement d'une application GenAI adaptée aux données de votre entreprise vous permet de tirer parti de vos propres connaissances et de votre expertise. Cette fonctionnalité de personnalisation garantit des réponses précises et pertinentes aux questions spécifiques à l'entreprise, améliorant ainsi la satisfaction et la productivité de tous vos utilisateurs.

Si vous "créer une base de connaissances", GenAI ingère des données à partir de vos sources de données, stocke les résultats vectorisés dans une base de données et vous donne un contrôle total sur l'utilisation des données ingérées pour répondre aux requêtes. Cette approche nécessite une configuration initiale plus importante, mais vous permet de choisir différents modèles de conversation pour différents résultats. Si vous "définir un connecteur NetApp pour Amazon Q Business", les données de vos sources de données sont ingérées par Amazon Q Business et stockées dans un index. Cette approche réduit les besoins en configuration initiale, tout en limitant le contrôle sur les résultats.

Pour plus d'informations sur Workload Factory, reportez-vous au "Présentation de Workload Factory" .

Avantages de l'utilisation de GenAI pour créer des applications d'IA générative

NetApp Workload Factory pour GenAI simplifie le processus de déploiement de l'infrastructure nécessaire à la création d'applications d'IA générative à l'aide de la génération augmentée par récupération (RAG). Plus précisément, GenAI offre les avantages suivants :

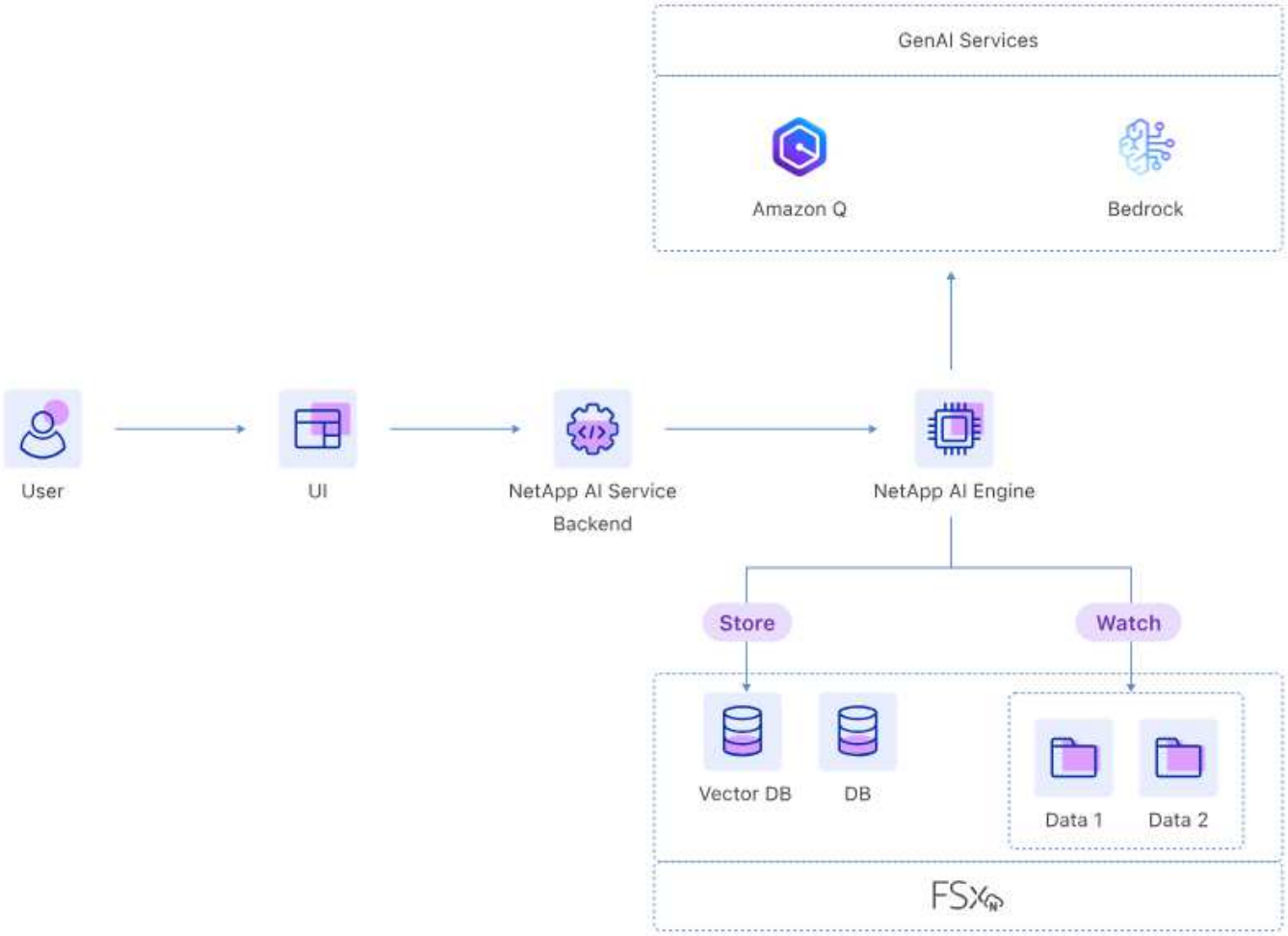
- Sans avoir besoin d'une connaissance approfondie de l'infrastructure des données, des bases de données et des modèles linguistiques, les administrateurs et développeurs INFORMATIQUES peuvent accélérer le développement d'applications en utilisant l'automatisation fournie par GenAI. Les administrateurs et développeurs de données peuvent facilement et rapidement créer des bases de connaissances d'entreprise qui intègrent les données non structurées de votre entreprise à utiliser par les applications d'IA

générative.

- Améliorez la sécurité en préservant les autorisations utilisateur dans les fichiers intégrés aux bases de connaissances pour garantir le maintien de la sécurité et de la confidentialité des données. Une application, telle qu'un chatbot, peut être développée pour fournir uniquement aux utilisateurs authentifiés les réponses basées sur les données auxquels ils ont accès.
- Préservez la confidentialité et la sécurité de vos données d'entreprise sur votre compte client AWS, là où vos données d'entreprise ne sont jamais exposées en externe.
- Accélérez le développement d'applications GenAI, comme un chatbot Q&A, à l'aide d'frameworks open source, comme LangChain, en utilisant l'API GenAI, pour provisionner et gérer des bases de connaissances et des connecteurs, discuter avec une base de connaissances, et stocker et récupérer l'historique de chat.
- Améliorez la protection et la disponibilité des données en déployant l'infrastructure de données d'IA générative sur les systèmes de fichiers FSX pour NetApp ONTAP et en exploitant les fonctionnalités ONTAP telles que la haute disponibilité, les snapshots pour la protection et la restauration des données locales, SnapMirror pour la reprise d'activité et SnapVault pour la sauvegarde de votre infrastructure de données.
- Réduisez les coûts globaux de stockage de l'infrastructure de données d'IA générative en exploitant les fonctionnalités d'efficacité des données ONTAP, telles que la déduplication, la compression et la compaction des données, le Tiering des données et le provisionnement fin.
- Obtenez des résultats de haute qualité à partir de vos données grâce aux fonctionnalités hybrides de recherche et de réorganisation fournies par GenAI. La recherche hybride combinée au nouveau classement améliore considérablement la pertinence des résultats de recherche. Ces fonctionnalités sont disponibles via Amazon AWS et dépendent de la région.

Fonctionnement de GenAI

GenAI utilise les données privées de votre entreprise pour compléter l'intelligence du modèle (en fonction des données sur lesquelles il a été entraîné) afin de fournir des réponses personnalisées aux questions posées par les utilisateurs de votre entreprise. Vous devez d'abord déployer l'infrastructure nécessaire à une structure RAG, puis créer une base de connaissances ou définir un connecteur à l'aide des sources de données et des modèles de base de votre entreprise disponibles via Amazon Bedrock ou Amazon Q Business, puis connecter une application (comme un chatbot Q&A) à la base de connaissances ou au connecteur.



Comment NetApp Workload Factory pour GenAI aide à créer des applications d'IA générative

GenAI permet de créer des applications d'IA générative à l'aide de RAG de l'une des manières suivantes :

- Déploie l'infrastructure requise pour le framework de génération accélérée des données de récupération (RAG) afin de travailler avec les sources de données sur les systèmes de fichiers FSX pour ONTAP et Amazon Bedrock ou Amazon Q Business. L'infrastructure inclut l'instance du moteur NetApp GenAI pour la gestion des données, une base de données vectorielle intégrée (LanceDB) et le stockage sur votre système de fichiers FSX pour ONTAP pour la base de données vectorielle.
- Permet de connecter les sources de données aux codages et aux modèles de langue disponibles via Amazon Bedrock ou Amazon Q Business pour intégrer les sources de données et récupérer les réponses pour les requêtes utilisateur. Les sources de données, ainsi que les modèles et leur configuration, sont présentés sous la forme de bases de connaissances FSX pour ONTAP.
- Ingère les données source dans la base de connaissances ou le connecteur pour intégrer les fichiers source dans les partages SMB et les exportations NFS sur les systèmes de fichiers FSX pour ONTAP, ainsi que pour stocker les autorisations de fichiers dans les partages SMB.
- Crée automatiquement des questions de base de conversation en fonction du contenu des bases de connaissances.
- Fournit un simulateur de discussion permettant aux administrateurs de données de tester la discussion avec les bases de connaissances.
- Fournit une interface de connecteur simple pour vous permettre de connecter GenAI à Amazon Q

Business, rapidement et facilement en utilisant les fonctionnalités de cet assistant d'IA.

Outils pour utiliser NetApp Workload Factory

Vous pouvez utiliser NetApp Workload Factory avec les outils suivants :

- **Console Workload Factory** : La console Workload Factory fournit une vue visuelle et holistique de vos applications et projets.
- * Console NetApp * : la console NetApp fournit une expérience d'interface hybride afin que vous puissiez utiliser Workload Factory avec d'autres services de données NetApp .
- **Demandez-moi** : utilisez l'assistant IA Ask me pour poser des questions et en savoir plus sur Workload Factory sans quitter la console Workload Factory. Accédez à Demandez-moi depuis le menu d'aide de Workload Factory.
- **CloudShell CLI** : Workload Factory inclut une CLI CloudShell pour gérer et exploiter les environnements AWS et NetApp sur plusieurs comptes à partir d'une seule CLI basée sur un navigateur. Accédez à CloudShell depuis la barre supérieure de la console Workload Factory.
- **API REST** : utilisez les API REST de Workload Factory pour déployer et gérer vos systèmes de fichiers FSx for ONTAP et d'autres ressources AWS.
- **CloudFormation** : utilisez le code AWS CloudFormation pour effectuer les actions que vous avez définies dans la console Workload Factory afin de modéliser, provisionner et gérer les ressources AWS et tierces de la pile CloudFormation dans votre compte AWS.
- **Fournisseur Terraform NetApp Workload Factory** : utilisez Terraform pour créer et gérer les workflows d'infrastructure générés dans la console Workload Factory.

Le coût

L'utilisation de la fonctionnalité GenAI de Workload Factory est gratuite.

Toutefois, vous devrez payer les ressources AWS que vous déployez pour prendre en charge l'infrastructure d'IA générative. Par exemple, vous paierez à AWS pour Amazon Bedrock ou Amazon Q Business, le système de fichiers FSX pour ONTAP et la capacité de stockage, ainsi que l'instance EC2 du moteur GenAI.

Certaines opérations multimodales, telles que la numérisation d'images pour des informations de texte, peuvent utiliser davantage de ressources et donc entraîner un coût plus élevé. Certaines opérations de configuration, telles que la modification des paramètres d'une base de connaissances, peuvent entraîner une nouvelle analyse des sources de données, et les analyses des sources de données peuvent également entraîner un coût plus élevé.

Licences

Aucune licence spéciale n'est requise auprès de NetApp pour utiliser les fonctionnalités d'IA de Workload Factory.

Régions

Workload Factory est pris en charge dans toutes les régions commerciales où FSx for ONTAP est pris en charge."Afficher les régions Amazon prises en charge."

Les régions AWS suivantes ne sont pas prises en charge :

- régions de Chine

- Régions GovCloud (États-Unis)
- Nuage secret
- Nuage top secret

Composants du moteur NetApp GenAI

Lorsque vous déployez l'infrastructure GenAI, Workload Factory crée une instance EC2 pour le moteur GenAI. Il crée également un rôle IAM, un groupe de sécurité et des points de terminaison privés pour cette instance. Vous souhaiterez peut-être comprendre plus en détail ces composants que Workload Factory crée dans votre environnement AWS.

Type d'instance EC2

m5.large

Rôle IAM

L'instance du moteur GenAI a besoin d'autorisations pour envoyer des segments de données au modèle d'intégration sur Amazon Bedrock et pour communiquer avec le système interne du service NetApp ai. Le rôle IAM comprend les autorisations suivantes :

Autorisations de rôle IAM

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

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

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

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

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

Groupe de sécurité

Les règles sortantes sont ouvertes à tout le trafic, tandis que les règles entrantes sont complètement fermées.

Terminaux privés

Si le VPC cible ne les possède pas déjà, Workload Factory crée des points de terminaison privés pour l'instance EC2 du moteur GenAI afin qu'elle puisse communiquer avec les services AWS suivants :

- Sol d'Amazon
  - le socle
  - durée de fonctionnement du système de roche
  - bedrock-agent-runtime
- Amazon Elastic Container Registry (ECR)
  - API
  - docker
- AWS Systems Manager (SSM)
  - ssm

    - ec2messages
    - ssmmessages
- Amazon FSX pour NetApp ONTAP
- Amazon CloudWatch

Utilisez GenAI pour construire des bases de connaissances pour Amazon Bedrock

Commencez

Démarrage rapide des bases de connaissances GenAI

Commencez à créer une base de connaissances ou Amazon Q Business Connector à l'aide des données de votre entreprise présentes sur les systèmes de fichiers Amazon FSX pour NetApp ONTAP. Une application telle qu'un chatbot accède à cette base de connaissances ou à ce connecteur pour fournir des réponses organisationnelles aux utilisateurs finaux.

Connectez-vous à Workload Factory

Vous aurez besoin de "créer un compte avec Workload Factory" et connectez-vous en utilisant l'un des "expériences de la console" .

Configurez votre environnement pour répondre aux exigences GenAI

Vous aurez besoin d'identifiants AWS pour déployer l'infrastructure AWS, d'un système de fichiers FSX for ONTAP déployé et découvert, de la liste des sources de données que vous souhaitez intégrer dans votre base de connaissances ou votre connecteur, de l'accès au service d'IA Amazon Bedrock ou à l'application d'entreprise Amazon Q, et bien plus encore.

"En savoir plus sur les exigences GenAI".

Identifiez le système de fichiers FSX pour ONTAP qui contient les sources de données

Les sources de données que vous intégrerez dans votre base de connaissances peuvent se trouver dans un seul système de fichiers FSX pour ONTAP ou dans plusieurs systèmes de fichiers FSX pour ONTAP. Si ces systèmes se trouvent dans des VPC différents, ils doivent être accessibles au sein du même réseau, ou les VPC doivent être associés et utiliser le même compte région et AWS que le moteur d'IA.

"Découvrez comment identifier les sources de données".

Déployez l'infrastructure GenAI

Lancez l'assistant de déploiement d'infrastructure pour déployer l'infrastructure GenAI dans votre environnement AWS. Ce processus déploie une instance EC2 pour le moteur NetApp GenAI et un volume sur un système de fichiers FSX pour ONTAP afin de contenir les bases de données du moteur NetApp ai. Le volume est utilisé pour stocker la base de données vectorielle utilisée par la base de connaissances.

"Découvrez comment déployer l'infrastructure de la base de connaissances".

Et la suite

Vous pouvez maintenant créer une base de connaissances pour fournir des réponses axées sur l'organisation

aux utilisateurs finaux.

Exigences de la base de connaissances GenAI

Assurez-vous que Workload Factory et AWS sont correctement configurés avant de créer votre base de connaissances. Cela inclut vos informations de connexion AWS, un système de fichiers FSx for ONTAP déployé contenant les sources de données que vous souhaitez intégrer dans votre base de connaissances, l'accès au service Amazon Bedrock AI, et bien plus encore.

Exigences GenAI de base

GenAI présente des exigences générales que votre environnement doit satisfaire avant de commencer.

Connexion et compte Workload Factory

Vous aurez besoin de "créer un compte avec Workload Factory" et connectez-vous en utilisant l'un des "expériences de la console" .

Identifiants et autorisations AWS

Vous devez ajouter les informations d'identification AWS à Workload Factory avec des autorisations de lecture/écriture, ce qui signifie que vous utiliserez Workload Factory en mode *lecture/écriture* pour GenAI.

Les autorisations *Basic* mode et *Read-Only* mode ne sont pas prises en charge pour le moment.

Lors de la configuration de vos identifiants, le fait de sélectionner les autorisations comme indiqué ci-dessous vous permet de bénéficier d'un accès complet pour gérer les systèmes de fichiers FSX pour ONTAP, ainsi que pour déployer et gérer l'instance EC2 GenAI et d'autres ressources AWS nécessaires à votre base de connaissances et à votre chatbot.

"Découvrez comment ajouter des informations d'identification AWS à Workload Factory"

Exigences de la base de connaissances GenAI

Si vous prévoyez de travailler avec des bases de connaissances, assurez-vous que votre environnement répond aux exigences suivantes.

Sol d'Amazon

Amazon Bedrock vous permet d'utiliser des modèles de base et offre les fonctionnalités nécessaires pour créer des applications d'IA générative.

Avant de commencer à utiliser NetApp Workload Factory pour GenAI, vous devez configurer Amazon Bedrock. Votre déploiement GenAI doit se trouver dans une région AWS sur laquelle Amazon Bedrock est activé.

- "Documentation AWS : configurez Amazon Bedrock"
- "Documentation AWS : régions et modèles pris en charge pour les bases de connaissances pour Amazon Bedrock"

GenAI classe par défaut les résultats de recherche pour en améliorer la pertinence. Pour obtenir les meilleurs résultats, assurez-vous que votre configuration de modèle de base Amazon Bedrock inclut l'accès à un modèle de classement, tel que Cohere Rerank ou Amazon Rerank, si disponible dans votre région.

Modèle d'intégration

Vous devez activer le modèle d'intégration que vous prévoyez d'utiliser avant de créer votre base de connaissances. Les modèles d'intégration suivants sont pris en charge :

- Codes Titan G1 - texte
- Texte d'intégration Titan v2
- Embedding multimodal G1 Titan
- Intégrer l'anglais
- Intégrer multilingue

"En savoir plus sur Amazon Titan"

Modèle de chat

Vous devez activer le modèle de chat de base que vous prévoyez d'utiliser avant de créer votre base de connaissances. Étant donné que la prise en charge des modèles varie selon les régions AWS, reportez-vous à la section "La documentation AWS" pour vérifier les modèles que vous pouvez utiliser dans les régions où vous prévoyez de déployer votre base de connaissances.

GenAI prend en charge divers modèles de Anthropic, Amazon, Mistral ai, Meta, Jamba, et Cohere.

En savoir plus sur l'utilisation de ces modèles dans Amazon Bedrock :

- "Claude d'anthropic dans le sol d'Amazon"
- "Prise en main d'Amazon Nova dans la console Amazon Bedrock"
- "Modèles d'IA Mistral"
- "Modèles texte Amazon Titan"
- "Modèles META Llama"
- "Modèles Jamba"
- "Modèles Cohere Command"

FSX pour le système de fichiers ONTAP

Vous avez besoin d'au moins un système de fichiers FSX pour ONTAP :

- Un système de fichiers sera utilisé (ou créé, s'il n'existe pas) par le moteur NetApp GenAI pour stocker la base de données vectorielle utilisée par la base de connaissances.

  Ce système de fichiers FSX pour ONTAP doit utiliser des volumes FlexVol. Les volumes FlexGroup ne sont pas pris en charge.

- Un ou plusieurs systèmes de fichiers contiendront les sources de données que vous intégrerez dans votre base de connaissances.

  Un système de fichiers FSX pour ONTAP peut être utilisé pour ces deux raisons ou vous pouvez utiliser plusieurs systèmes de fichiers FSX pour ONTAP.

- Vous devez connaître la région AWS, le VPC et le sous-réseau où réside le système de fichiers AWS FSX pour ONTAP. Le système de fichiers doit se trouver dans une région AWS sur laquelle Amazon Bedrock est activé.
- Vous devrez tenir compte des paires clé/valeur que vous souhaitez appliquer aux ressources AWS

faisant partie de ce déploiement (facultatif).

- Vous devrez connaître les informations relatives aux paires de clés qui vous permettront de vous connecter en toute sécurité à l'instance du moteur NetApp ai.

"Découvrez comment déployer et gérer les systèmes de fichiers FSX pour ONTAP"

Identifier les sources de données à ajouter à une base de connaissances ou à un connecteur

Identifiez ou créez les documents (sources de données) qui résident dans votre système de fichiers FSX pour ONTAP que vous intégrerez dans votre base de connaissances. Ces sources de données permettent à la base de connaissances de fournir des réponses précises et personnalisées aux questions des utilisateurs en fonction des données pertinentes pour votre entreprise.

Nombre maximum de sources de données

Le nombre maximum de sources de données prises en charge est de 10.

Emplacement des sources de données

Les sources de données peuvent être stockées dans un seul volume, ou dans un dossier au sein d'un volume, sur un partage SMB ou une exportation NFS sur un système de fichiers Amazon FSX pour NetApp ONTAP. Les sources de données peuvent également être stockées dans des volumes Amazon FSX pour NetApp ONTAP dans une relation de protection des données NetApp SnapMirror.

Vous ne pouvez pas sélectionner des documents individuels au sein d'un volume ou d'un dossier. Par conséquent, vous devez vous assurer que chaque volume ou dossier contenant des sources de données ne contient pas de documents externes qui ne doivent pas être intégrés à votre base de connaissances.

Vous pouvez ajouter plusieurs sources de données dans chaque base de connaissances, mais elles doivent résider sur des systèmes de fichiers FSX pour ONTAP accessibles depuis votre compte AWS.

La taille de fichier maximale pour chaque source de données est de 50 Mo.

Protocoles pris en charge

Les bases de connaissances prennent en charge les données de volumes qui utilisent des protocoles NFS ou SMB/CIFS. Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory afin que la base de connaissances puisse accéder aux fichiers de ces volumes. Ceci inclut le domaine Active Directory, l'adresse IP, le nom d'utilisateur et le mot de passe.

Lorsque vous stockez votre source de données sur un partage (fichier ou répertoire) accessible via SMB, les données sont uniquement accessibles aux utilisateurs ou groupes de chatbot qui disposent des autorisations d'accès à ce partage. Lorsque cette « fonctionnalité de reconnaissance des autorisations » est activée, le système d'IA compare le courrier électronique de l'utilisateur dans auth0 aux utilisateurs autorisés à afficher ou utiliser les fichiers sur le partage SMB. Le chatbot fournira des réponses en fonction des autorisations utilisateur pour les fichiers intégrés.

Par exemple, si vous avez intégré 10 fichiers (sources de données) dans votre base de connaissances et que 2 fichiers sont des fichiers de ressources humaines qui contiennent des informations restreintes, seuls les utilisateurs de chatbot qui sont authentifiés pour accéder à ces 2 fichiers recevront des réponses du chatbot qui incluent des données de ces fichiers.

Formats de fichiers de source de données pris en charge

Les formats de fichiers de source de données suivants sont actuellement pris en charge avec les bases de connaissances Workload Factory GenAI.

Format de fichier	Extension
Apache Parquet [1]	.parquet
Valeurs séparées par des virgules filefootnote:clause de non-responsabilité[]	.csv
Format d'échange graphique	.gif
JPEG	.jpg or.jpeg
JSON et JSONPFootnote:Disclaimer[]	.json
Démarque	.md
Microsoft Word	.doc ou .docx
Texte brut	.txt
Format de document portable	.pdf
Carte graphique réseau portable	.png
Image WebP	.webp

Déployez l'infrastructure GenAI

Vous devez déployer l'infrastructure GenAI pour la structure RAG dans votre environnement avant de pouvoir créer des bases de connaissances, des connecteurs et des applications FSX pour ONTAP pour votre entreprise. Les principaux composants de l'infrastructure sont le service Amazon Bedrock, une instance de machine virtuelle pour le moteur NetApp GenAI et un système de fichiers FSX pour ONTAP.

L'infrastructure déployée peut prendre en charge plusieurs bases de connaissances, chatbots et connecteurs. Vous n'aurez donc généralement besoin d'effectuer cette tâche qu'une seule fois.

Détails de l'infrastructure

Votre déploiement GenAI doit se trouver dans une région AWS sur laquelle Amazon Bedrock est activé. "Afficher la liste des régions prises en charge"

L'infrastructure comprend les composants suivants.

Service Amazon Bedrock

Amazon Bedrock est un service entièrement géré qui vous permet d'utiliser les modèles de base (FMS) des plus grandes entreprises d'IA via une seule API. Il fournit également les fonctionnalités nécessaires pour créer des applications d'IA générative sécurisées.

"En savoir plus sur Amazon Bedrock"

Amazon Q Business

Amazon Q s'appuie sur Amazon Bedrock pour fournir un assistant d'IA générative entièrement géré que vous pouvez utiliser pour répondre à des questions et générer du contenu à partir d'informations issues de vos sources de données.

"En savoir plus sur Amazon Q Business"

Machine virtuelle pour le moteur NetApp GenAI

Le moteur NetApp GenAI est déployé au cours de ce processus. Il fournit la puissance de traitement permettant d'acquérir les données de vos sources de données, puis d'écrire ces données dans la base de données vectorielle.

FSX pour le système de fichiers ONTAP

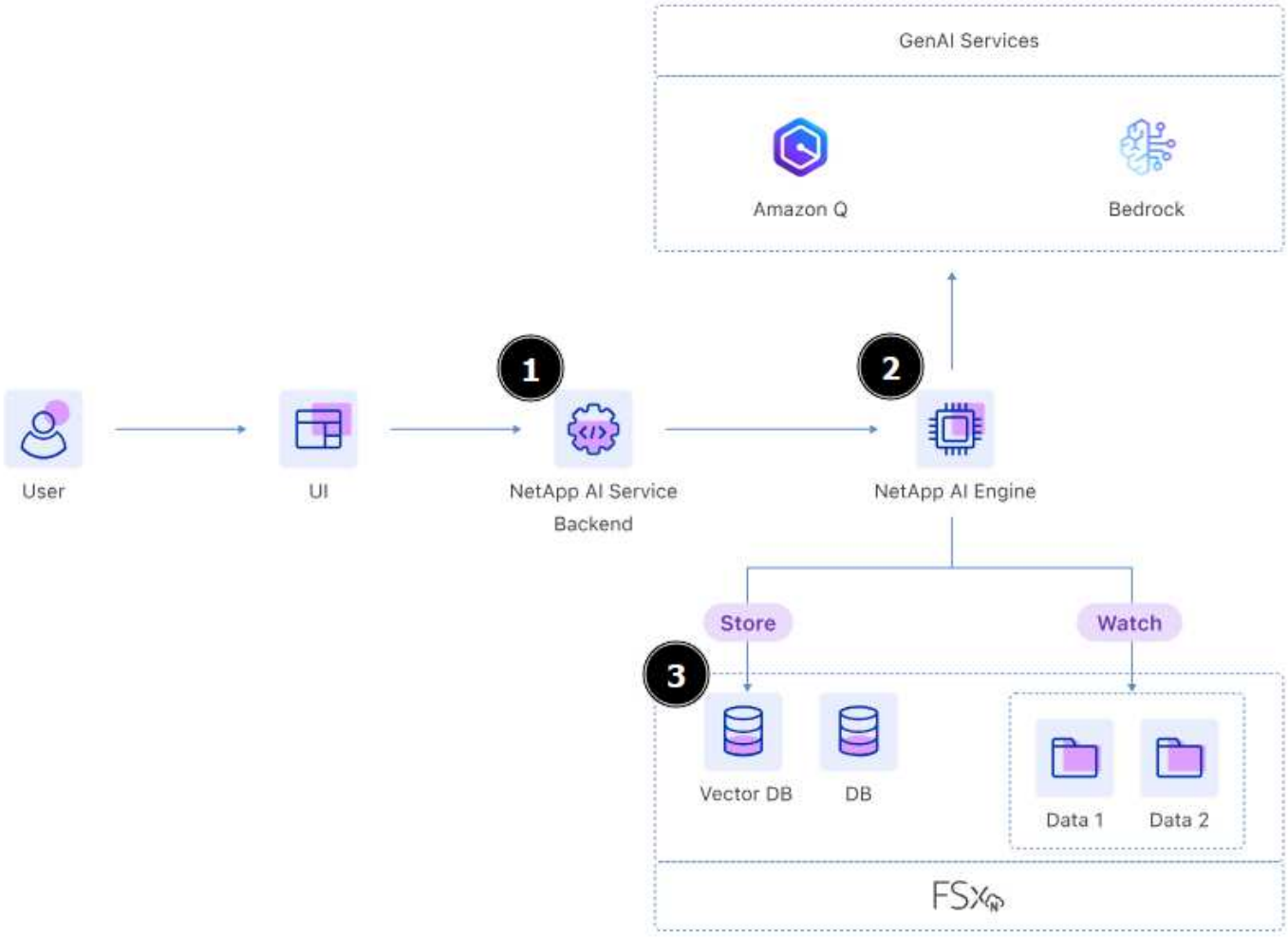
Le système de fichiers FSX pour ONTAP fournit le stockage de votre système GenAI.

Un seul volume est déployé qui contiendra la base de données vectorielle qui stocke les données générées par le modèle de base basé sur vos sources de données.

Les sources de données que vous intégrerez dans votre base de connaissances peuvent résider sur le même système de fichiers FSX pour ONTAP ou sur un autre système.

Le moteur NetApp GenAI surveille ces volumes et interagit avec eux.

L'image suivante montre l'infrastructure GenAI. Les composants numérotés 1, 2 et 3 sont déployés au cours de cette procédure. Les autres éléments doivent être en place avant de commencer le déploiement.



Déployez l'infrastructure GenAI

Vous devrez saisir vos identifiants AWS et sélectionner le système de fichiers FSX pour ONTAP pour déployer l'infrastructure de génération augmentée de récupération.

Avant de commencer

Avant de commencer cette procédure, assurez-vous que votre environnement répond aux conditions requises pour les bases de connaissances ou les connecteurs, selon ce que vous choisissez.

- "Exigences de la base de connaissances"
- "Exigences relatives au connecteur"

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Examinez le diagramme d'infrastructure et sélectionnez **Suivant**.
4. Complétez les éléments de la section **paramètres AWS** :
   a. **Informations d'identification AWS** : sélectionnez ou ajoutez les informations d'identification AWS qui fournissent des autorisations pour déployer les ressources AWS.
   b. **Emplacement** : sélectionnez une région AWS, un VPC et un sous-réseau.

Le déploiement GenAI doit se trouver dans une région AWS sur laquelle le système Amazon Bedrock est activé. "Afficher la liste des régions prises en charge"

5. Complétez les éléments de la section **Paramètres de l'infrastructure** :
   a. **Tags** : saisissez les paires clé/valeur de balise que vous souhaitez appliquer à toutes les ressources AWS qui font partie de ce déploiement. Ces balises sont visibles dans la console de gestion AWS et dans la zone d'informations sur l'infrastructure de Workload Factory et peuvent vous aider à suivre les ressources de Workload Factory.
6. Complétez la section **connectivité** :
   a. **Paire de clés** : sélectionnez une paire de clés qui vous permet de vous connecter en toute sécurité à l'instance du moteur NetApp GenAI.
7. Compléter la section **ai Engine** :
   a. **Nom de l'instance** : si vous le souhaitez, sélectionnez **Définir le nom de l'instance** et saisissez un nom personnalisé pour l'instance du moteur d'IA. Le nom de l'instance apparaît dans la console de gestion AWS et dans la zone d'informations sur l'infrastructure de Workload Factory et peut vous aider à suivre les ressources de Workload Factory.
8. Sélectionnez **déployer** pour commencer le déploiement.

Si le déploiement échoue avec une erreur d'informations d'identification, vous pouvez obtenir plus de détails sur l'erreur en sélectionnant les liens hypertexte dans le message d'erreur. Vous pouvez voir une liste d'autorisations manquantes ou bloquées, ainsi qu'une liste d'autorisations dont la charge de travail GenAI a besoin pour pouvoir déployer l'infrastructure GenAI.

Résultat

Workload Factory commence à déployer l'infrastructure du chatbot. Ce processus peut prendre jusqu'à 10 minutes.

Au cours du processus de déploiement, les éléments suivants sont configurés :

- Le réseau est configuré avec les terminaux privés.
- Le rôle IAM, le profil d'instance et le groupe de sécurité sont créés.
- L'instance de machine virtuelle du moteur GenAI est déployée.
- Amazon Bedrock est configuré pour envoyer des journaux à Amazon CloudWatch, en utilisant un groupe de journaux avec le préfixe `/aws/bedrock/`.
- Le moteur GenAI est configuré pour envoyer des journaux à Amazon CloudWatch Logs, à l'aide d'un groupe de journaux portant le nom `/netapp/wlmai/<tenancyAccountId>/randomId` , où `<tenancyAccountID>` est le "ID de compte de la console NetApp" pour l'utilisateur actuel.

Créer une base de connaissances GenAI

Après avoir déployé l'infrastructure d'IA et identifié les sources de données que vous intégrerez dans votre base de connaissances à partir de vos magasins de données FSx for ONTAP , vous êtes prêt à créer la base de connaissances à l'aide de Workload Factory. Dans le cadre de cette étape, vous définirez également les caractéristiques de l'IA et créerez des déclencheurs de conversation.

Avant de continuer, assurez-vous que votre environnement est conforme aux "de formation"bases de connaissances for.

Description de la tâche

Les bases de connaissances ont deux modalités d'intégration de données : *mode public* et *mode entreprise*.

Mode public

Une base de connaissances peut être utilisée sans intégrer les sources de données de votre organisation. Dans ce cas, une application intégrée à la base de connaissances ne fournira que des résultats provenant d'informations accessibles au public sur Internet. Il s'agit d'une intégration de *mode public*.

Mode entreprise

Dans la plupart des cas, vous voudrez intégrer les sources de données de votre organisation dans la base de connaissances. Il s'agit d'une intégration *Enterprise mode* car elle fournit des connaissances de votre entreprise.

Les sources de données de votre organisation peuvent contenir des informations personnelles identifiables (PII). Pour protéger ces informations sensibles, vous pouvez activer les *garde-fous de données* lors de la création et de la configuration des bases de connaissances. Les garde-fous de données, optimisés par NetApp Data Classification, identifient et masquent les informations personnelles identifiables, les rendant inaccessibles et irrécupérables.

"En savoir plus sur la classification des données NetApp" .

NetApp Workload Factory pour GenAI ne masque pas les informations personnelles sensibles (SPii). Se référer à"types de données personnelles sensibles" pour plus d'informations sur ce type de données.

Les garde-fous de données peuvent être activés ou désactivés à tout moment. Si vous activez les garde-fous de données, Workload Factory analyse l'intégralité de la base de connaissances à partir de zéro, ce qui entraîne un coût.

Créer et configurer la base de connaissances

La base de connaissances définit des caractéristiques telles que les modèles d'IA de Bedrock et le format d'intégration que vous souhaitez utiliser pour créer votre base de connaissances.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans le menu Bases de connaissances et connecteurs, sélectionnez la liste déroulante **Créer un nouveau** et choisissez **Base de connaissances NetApp GenAI pour Bedrock**.
4. Sur la page Créer une base de connaissances NetApp GenAI, configurez les paramètres de la base de connaissances :

Détails de la base de connaissances

1. **Nom** : saisissez le nom que vous souhaitez utiliser pour la base de connaissances.
2. **Description** : saisissez une description détaillée de la base de connaissances.
3. **Bedrock** : choisissez la région dans laquelle Amazon Bedrock est disponible pour votre compte AWS.

Ingestion

1. **Modèle d'intégration**:
   - Choisissez un modèle d'intégration à utiliser pour la base de connaissances. Le modèle d'intégration définit comment vos données seront converties en intégrations vectorielles pour la base de connaissances. Workload Factory prend en charge les modèles suivants :
   - Codes Titan G1 - texte
   - Texte d'intégration Titan v2
   - Embedding multimodal G1 Titan
   - Intégrer l'anglais
   - Intégrer multilingue

     Notez que vous devez avoir déjà activé le modèle d'incorporation à partir d'Amazon Bedrock.

   "En savoir plus sur Amazon Titan"

   - Le cas échéant, sélectionnez le type d'inférence qui correspond à la configuration du modèle d'intégration sélectionné.
2. **Protections de données** : Choisissez si vous souhaitez activer ou désactiver les protections de données. "Découvrez les garde-fous de données, optimisés par la classification des données NetApp" .

   Les conditions préalables suivantes doivent être remplies pour activer les rails de données.

   - Un compte de service est requis pour communiquer avec NetApp Data Classification. Vous devez disposer du rôle *Administrateur de l'organisation* sur votre compte de location NetApp Console pour la création d'un compte de service. Un membre disposant du rôle d'administrateur de l'organisation peut effectuer toutes les actions dans le ."Découvrez comment ajouter un rôle à un membre dans la console NetApp"
   - Le moteur d'IA doit avoir accès à la"Point de terminaison de l'API de la console NetApp" .
   - Vous devrez effectuer les opérations suivantes comme décrit dans"Documentation sur la classification des données NetApp" :
     i. Créer un agent de console
     ii. Assurez-vous que votre environnement peut répondre aux conditions préalables
     iii. Déployer la classification des données NetApp

   > La fonction Data guardrails n'est pas prise en charge lors de l'acquisition de fichiers de données structurés tels que CSV, JSON, JSONP ou parquet.

Paramètres de chat et de récupération

1. **Modèle de chat** :
   - Choisissez parmi différents modèles de chat intégrés à Amazon Bedrock. Notez que vous devez déjà avoir activé le modèle de chat depuis Amazon Bedrock.
   - Le cas échéant, sélectionnez le type d'inférence qui correspond à la configuration du modèle sélectionné.
2. **Paramètres de chat**:

- Choisissez une température pour le chatbot pour configurer le caractère aléatoire et la créativité des réponses. Une température plus basse entraîne des réponses plus prévisibles, et une température plus élevée entraîne des réponses plus variées.
- Choisissez une longueur de réponse maximale pour configurer le niveau de détail des réponses. Des longueurs de réponse plus longues utilisent plus de jetons de réponse et peuvent entraîner un coût plus élevé.

3. **Mode réflexion** : lorsque le mode réflexion est activé, le chatbot prendra plus de temps pour traiter les requêtes et les résultats seront généralement plus précis. Lorsque vous activez le mode de réflexion, vous pouvez contrôler le nombre de jetons de raisonnement utilisés lors de la génération de résultats. L'utilisation de davantage de jetons de raisonnement peut conduire à des réponses plus précises, mais peut entraîner un coût plus élevé.
4. **Reclassement** : Activez ou désactivez le reclassement, ce qui peut améliorer la pertinence et la qualité des résultats de requête. Choisissez un modèle de chat standard ou un modèle de reclassement spécialisé à utiliser pour le reclassement. Les options du modèle Reranker ne sont affichées que si elles sont disponibles dans votre région. Sélectionnez le type d'inférence qui correspond à la configuration du modèle sélectionné.
5. **Amorces de conversation** : indiquez si vous souhaitez fournir jusqu'à quatre invites de démarrage de conversation qui s'affichent aux utilisateurs qui interagissent avec un chatbot utilisant cette base de connaissances. Nous vous recommandons d'activer ce paramètre.

   Si vous activez les amorces de conversation, le mode automatique est sélectionné par défaut. Le « mode manuel » ne peut être activé qu'après avoir ajouté des sources de données à votre base de connaissances. "Apprenez à modifier les paramètres de la base de connaissances".

Définitions de stockage

1. *Système de fichiers FSx pour ONTAP * : lorsque vous définissez une nouvelle base de connaissances, Workload Factory crée un nouveau volume Amazon FSx for NetApp ONTAP pour la stocker. Choisissez un nom de système de fichiers existant et une SVM (également appelée VM de stockage) où le nouveau volume sera créé.
2. **Politique de capture instantanée** : choisissez une politique de capture instantanée dans la liste des politiques existantes définies dans l'inventaire de stockage Workload Factory. Des instantanés récurrents de la base de connaissances seront automatiquement créés à une fréquence basée sur la politique d'instantanés que vous sélectionnez.
3. **Seau S3** : si les résultats de la requête du chatbot contiennent des données structurées, GenAI peut stocker les résultats dans un compartiment S3. Pour utiliser cette fonctionnalité, activez le paramètre **Activer le compartiment S3** et choisissez un compartiment S3 associé à votre compte dans la liste. Lorsque ces résultats sont stockés dans un bucket S3, vous pouvez les télécharger à l'aide du lien de téléchargement dans la session de chat.

   Si la règle de snapshot dont vous avez besoin n'existe pas, vous pouvez "création d'une règle de snapshots" la faire sur la machine virtuelle de stockage qui contient le volume.

4. Sélectionnez **Créer la base de connaissances** pour ajouter la base de connaissances à GenAI.

   Un indicateur de progression s'affiche pendant la création de la base de connaissances.

   Une fois la base de connaissances créée, vous avez la possibilité d'ajouter une source de données à votre nouvelle base de connaissances ou de mettre fin au processus sans ajouter de source de données. Nous vous recommandons de sélectionner **Ajouter une source de données** et d'ajouter une ou plusieurs sources de données maintenant.

Ajoutez des sources de données à la base de connaissances

Vous pouvez ajouter une ou plusieurs sources de données pour remplir la base de connaissances avec les données de votre organisation.

Description de la tâche

Le nombre maximum de sources de données prises en charge est de 10.

Étapes

1. Après avoir sélectionné **Ajouter une source de données**, sélectionnez le type de source de données que vous souhaitez ajouter :
    - Ajouter un système de fichiers FSx pour ONTAP (utiliser les fichiers d’un volume FSx pour ONTAP existant)
    - Ajouter un système de fichiers (utiliser des fichiers provenant d’un partage SMB ou NFS générique)

Ajouter un système de fichiers FSx pour ONTAP

1. **Sélectionnez un système de fichiers** : sélectionnez le système de fichiers FSX pour ONTAP dans lequel résident vos fichiers source de données et sélectionnez **Suivant**.
2. **Sélectionnez un volume** : sélectionnez le volume sur lequel vos fichiers de source de données résident et sélectionnez **Suivant**.

   Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory, notamment le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
3. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.
4. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

   Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

5. Dans la section **permission Aware**, disponible uniquement lorsque la source de données que vous avez sélectionnée se trouve sur un volume qui utilise le protocole SMB, vous pouvez activer ou désactiver les réponses sensibles aux autorisations :
   - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
   - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
6. Sélectionnez **Ajouter** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers NFS générique

1. **Sélectionnez un système de fichiers** : saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données, choisissez le protocole NFS pour le partage réseau et sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom de l'exportation NFS et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner l'exportation dans son intégralité ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :
   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50

et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

> Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

- **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
  - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

    Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

  Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

> Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

> Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers SMB générique

1. **Sélectionnez le système de fichiers**:
   a. Saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données.
   b. Choisissez le protocole SMB pour le partage réseau.
   c. Saisissez les informations Active Directory, qui incluent le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
   d. Sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom du partage SMB et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner le partage entier ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

   - **Conscient des autorisations** : Activer ou désactiver les réponses conscientes des autorisations :
     - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
     - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

     Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

     Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

   - Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de

l'heure de modification, sélectionnez une plage de dates dans la liste.

> Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Résultat

La source de données commence à être intégrée à votre base de connaissances. L'état passe de « intégration » à « intégrée » lorsque la source de données est complètement intégrée.

Après avoir ajouté une seule source de données à la base de connaissances, vous pouvez la tester localement dans la fenêtre du simulateur de chatbot et apporter les modifications requises avant de rendre le chatbot disponible pour vos utilisateurs. Vous pouvez également suivre les mêmes étapes pour ajouter des sources de données supplémentaires à la base de connaissances.

Testez une base de connaissances GenAI

Une fois la base de connaissances créée, vous pourrez la tester localement à l'aide du simulateur de chatbot et apporter les modifications requises avant de mettre la base de connaissances à la disposition de vos utilisateurs via une application de chatbot.

Description de la tâche

Vous testez votre base de connaissances pour vous assurer qu'elle fonctionne comme vous l'attendez, et vous pouvez personnaliser les amorces de conversation que vous voulez être disponible par défaut pour les utilisateurs chatbot de cette base de connaissances. Le simulateur de chatbot s'exécute sur toutes les sources de données qui ont été intégrées dans la base de connaissances.

Vous pouvez tester une base de connaissances en discutant avec vos sources de données intégrées dans le simulateur de chatbot. Notez qu'aucune interaction ou information n'est capturée dans la base de données vectorielle GenAI lors du test local de la base de connaissances.

Vous effectuerez la plupart de vos tests dans Workload Factory avant de déployer la base de connaissances dans une application pour vos utilisateurs. Si vous devez apporter des modifications à votre source de données ou au fonctionnement du chatbot, vous souhaiterez le faire maintenant avant de publier votre base de connaissances.

> Vous pouvez redimensionner et retitle la fenêtre du simulateur de chatbot, et copier des questions et des réponses dans le presse-papiers.

Voici quelques-unes des tâches que vous voudrez effectuer pour tester votre chatbot :

- Saisissez un grand nombre de questions pertinentes pour votre entreprise afin de vous assurer que les réponses sont conformes aux attentes.
- Personnalisez les amorces de conversation que vous souhaitez rendre disponibles par défaut pour vos utilisateurs dans l'application chatbot.
- Assurez-vous que le contenu attribué qui est fourni en bas des réponses du chatbot contient les références

correctes.

Étapes

1. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez tester.

   Le simulateur de chatbot s'affiche dans le volet droit. Si elles sont définies, les amorces de conversation existantes s'affichent également.

2. Dans le champ de saisie chatbot, saisissez une invite ou une question, puis sélectionnez pour voir comment votre chatbot répond avec vos connaissances organisationnelles.

   - Vous pouvez voir les sources utilisées pour produire la réponse en développant la liste **sources** sous la réponse. Cette option fournit une liste des fichiers utilisés pour générer la réponse. Vous pouvez afficher et copier les blocs de données utilisés à partir de chaque chemin de fichier et de volume vers chaque fichier en passant le curseur sur le nom du fichier.
   - Si des tableaux sont inclus dans la réponse, vous pouvez trier les données dans chaque colonne et copier chaque tableau dans le presse-papiers.
   - Si les résultats des réponses contiennent des données structurées et que la fonctionnalité **S3 Bucket** est activée pour la base de connaissances, GenAI stocke les résultats dans un bucket S3. Vous pouvez télécharger les résultats du bucket en utilisant le lien **Télécharger les résultats** dans la session de chat.

3. Si vous devez mettre à jour l'une de vos sources de données afin que votre base de connaissances fournisse des réponses plus ciblées, effectuez ces modifications maintenant, puis testez à nouveau la base de connaissances.

Activez l'authentification externe pour une base de connaissances GenAI

Activez l'authentification pour une base de connaissances afin que la validation des tokens et les listes de contrôle d'accès soient requises lors de l'utilisation des terminaux API pour intégrer une base de connaissances à une application chatbot. Lorsque vous activez l'authentification, vous configurez les paramètres d'un jeton Web JSON qui sera utilisé pour les demandes d'API à une base de connaissances à partir de clients chatbot.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances pour laquelle vous souhaitez activer l'authentification.
4. Sélectionnez et sélectionnez **gérer la base de connaissances**.
5. Sélectionnez le menu **actions** et sélectionnez **gérer les paramètres d'authentification**.
6. Configurer l'authentification :

   a. Sélectionnez **Activer les paramètres d'authentification**.

   b. Fournissez les informations requises. Des exemples sont fournis, mais vous devez obtenir les valeurs

de ces champs auprès de votre fournisseur d'authentification :

- **Algorithmes** : l'algorithme de signature utilisé par votre fournisseur d'authentification.
- **Public** (Facultatif) : chaîne contenant le destinataire du jeton (parfois une URL).
- **Émetteur** : chaîne qui identifie le fournisseur ayant émis le jeton.

  Par exemple, Amazon Cognito utilise des chaînes d'émetteur au format suivant :

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

  Où `<region>` est la région AWS contenant le pool d'utilisateurs et `<UserPoolID>` votre ID de pool d'utilisateurs. Vous pouvez récupérer votre ID de pool d'utilisateurs à l'aide de la commande suivante :

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

- **JWKS URI** : chaîne URI qui fournit les clés publiques nécessaires pour vérifier les signatures de ce token.

  Par exemple, Amazon Cognito utilise des chaînes URI JWKS au format suivant :

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

+
Où `<region>` est la région AWS contenant le pool d'utilisateurs et `<UserPoolID>` votre ID de pool d'utilisateurs. Vous pouvez récupérer votre ID de pool d'utilisateurs à l'aide de la commande suivante :

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

7. Sélectionnez **Enregistrer**.

Résultat

L'authentification pour la base de connaissances est désormais active et vous pouvez utiliser des terminaux d'API pour interagir avec la base de connaissances et intégrer la base de connaissances à une application chatbot.

Publiez une base de connaissances GenAI et affichez le terminal unique

Une fois votre base de connaissances créée et testée localement, vous pouvez publier la base de connaissances afin qu'elle puisse être intégrée à une application chatbot qui permettra aux utilisateurs d'interroger la base de connaissances.

Description de la tâche

La publication de la base de connaissances vous permet de l'utiliser dans des applications de chat. L'action de publication déclenche l'API Workload Factory pour générer et publier des points de terminaison uniques. Après la publication, la base de connaissances devient accessible aux applications de chat et les points de terminaison de l'API sont prêts pour l'intégration.

Chaque base de connaissances que vous publiez possède des points finaux uniques.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez publier.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration et la liste de toutes les sources de données intégrées.

5. Sélectionnez le menu **actions** et sélectionnez **publier**.

   Workload Factory publie la base de connaissances. Sur la page de détails de la base de connaissances, le statut passe de **Non publié** à **Publié**.

   Vous pouvez maintenant obtenir des détails sur le point de terminaison unique de la base de connaissances.

6. En regard de l'état publié, sélectionnez **Afficher**.

   Des détails sur la manière d'accéder à la base de connaissances à l'aide de l'API Workload Factory sont affichés.

7. Dans la boîte de dialogue **Afficher infos publiées**, copiez les noeuds finaux API que vous pouvez utiliser pour intégrer la base de connaissances à une application.

   Pour en savoir plus sur les noeuds finaux API, accédez au "Documentation de l'API" et sélectionnez **ai > externe**.

   Avant de pouvoir utiliser ces noeuds finaux, vous devez obtenir un jeton d'utilisateur auprès de votre fournisseur d'authentification.

Résultat

Vous disposez désormais d'une base de connaissances publiée et du point de terminaison unique que vous pouvez utiliser pour intégrer la base de connaissances à une application chatbot.

Utilisez l'application de chatbot externe GenAI

Une fois que vous avez configuré, activé et publié une base de connaissances, les développeurs d'applications externes peuvent configurer et exécuter l'exemple d'application de chatbot open source fourni par NetApp pour interagir avec votre base de connaissances et apprendre à utiliser l'API Workload Factory pour créer leurs propres applications d'IA génératives.

Étapes

1. "Créer une base de connaissances".
2. "Activez l'authentification" pour la base de connaissances que vous avez créée.

   Cela permet à la base de connaissances d'authentifier les requêtes d'API et de valider les tokens et les listes de contrôle d'accès requises lors de l'utilisation des noeuds finaux de l'API.

   > Les applications de chat externes qui s'intègrent à cette base de connaissances devront utiliser le même fournisseur d'authentification (émetteur) que vous configurez dans les paramètres d'authentification de la base de connaissances.

3. "Publier la base de connaissances" Pour activer l'accès API pour les applications externes.

   Après la publication d'une base de connaissances, les terminaux de l'API sont accessibles en externe. Vous pouvez intégrer la base de connaissances à une application de chat externe (comme l'exemple d'application chatbot).

4. Téléchargez l'exemple de package d'application chatbot sur "GitHub".
5. Installez et exécutez l'application chatbot en suivant les instructions du fichier README inclus dans le package.
6. Accédez à "http://localhost:9091" pour vous connecter à l'application.

   L'exemple d'application chatbot s'affiche.

En savoir plus >>

"Documentation de l'API Workload Factory"

Créez une application GenAI basée sur RAG

Une fois que vous avez créé votre base de connaissances et testé votre chatbot, vous êtes prêt à configurer l'application qui permettra à vos utilisateurs d'interroger le chatbot.

"Découvrez comment créer une application d'IA basée sur RAG sur FSX pour ONTAP"

Ce que vous pouvez faire ensuite avec GenAI

Maintenant que vous avez créé une base de connaissances à l'aide des données de votre entreprise et que vous l'avez déployée pour vos utilisateurs, vous pouvez gérer la base de connaissances, les sources de données et l'infrastructure RAG, y compris les systèmes de fichiers FSX pour ONTAP.

Voici quelques-unes des tâches que vous pouvez effectuer pour gérer les composants de votre base de connaissances :

- Mettez à jour le contenu de vos sources de données, ou ajoutez de nouvelles sources de données, et synchronisez ces modifications avec votre base de connaissances et votre chatbot.
- Gérez les paramètres de la source de données, notamment la stratégie de suppression et la reconnaissance des autorisations (pour l'accès aux fichiers SMB).

- Gérez les paramètres de votre base de connaissances, notamment le modèle de chat et les amorces de discussion.
- Annulez la publication d'une base de connaissances ou republiez-la après avoir effectué des modifications.
- Sauvegardez et protégez les données importantes sur votre système de fichiers FSX pour ONTAP pour vous assurer que vos données de la base de connaissances et d'autres composants d'infrastructure sont toujours disponibles.

Pour plus d'informations sur la gestion de votre système de fichiers FSx for ONTAP , accédez au "Documentation de Workload Factory pour Amazon FSx for NetApp ONTAP" pour afficher les capacités de sauvegarde et de protection que vous pouvez utiliser.

[1] la fonctionnalité Data guardrails n'est pas prise en charge lors de l'acquisition de fichiers de données structurés dans des bases de connaissances.

Utilisez GenAI pour créer des connecteurs pour Amazon Q Business

Commencez

Démarrage rapide pour les connecteurs GenAI

Commencez à créer un connecteur NetApp pour Amazon Q Business à l'aide des données de votre organisation qui existent sur les systèmes de fichiers Amazon FSx pour NetApp ONTAP. Après avoir créé un connecteur, les utilisateurs finaux peuvent accéder à l'assistant Amazon Q Business pour obtenir des réponses axées sur l'organisation à leurs questions.

Connectez-vous à Workload Factory

Vous aurez besoin de "créer un compte avec Workload Factory" et connectez-vous en utilisant l'un des "expériences de la console" .

Configurez votre environnement pour répondre aux exigences GenAI

Vous aurez besoin d'informations d'identification AWS pour déployer l'infrastructure AWS, d'un système de fichiers FSX for ONTAP déployé et découvert, de la liste des sources de données que vous souhaitez intégrer à votre connecteur, de l'accès à l'application Amazon Q Business et bien plus encore.

"En savoir plus sur les exigences GenAI".

Identifiez le système de fichiers FSX pour ONTAP qui contient les sources de données

Les sources de données que vous intégrerez dans votre connecteur peuvent être situées dans un seul système de fichiers FSX pour ONTAP ou dans plusieurs systèmes de fichiers FSX pour ONTAP. Si ces systèmes se trouvent dans des VPC différents, ils doivent être accessibles au sein du même réseau, ou les VPC doivent être associés et utiliser le même compte région et AWS que le moteur d'IA.

"Découvrez comment identifier les sources de données".

Déployez l'infrastructure GenAI

Lancez l'assistant de déploiement d'infrastructure pour déployer l'infrastructure GenAI dans votre environnement AWS. Ce processus déploie une instance EC2 pour le moteur NetApp GenAI et un volume sur un système de fichiers FSX pour ONTAP afin de contenir les bases de données du moteur NetApp ai. Le volume est utilisé pour stocker des informations sur le connecteur.

"Découvrez comment déployer l'infrastructure GenAI".

Et la suite

Vous pouvez maintenant créer un connecteur pour Amazon Q Business afin de fournir des réponses axées sur l'organisation aux utilisateurs finaux.

Exigences relatives au connecteur GenAI

Assurez-vous que Workload Factory et AWS sont correctement configurés avant de créer un connecteur NetApp pour Amazon Q Business.

Exigences GenAI de base

GenAI présente des exigences générales que votre environnement doit satisfaire avant de commencer.

Connexion et compte Workload Factory

Vous aurez besoin de "créer un compte avec Workload Factory" et connectez-vous en utilisant l'un des "expériences de la console" .

Identifiants et autorisations AWS

Vous devez ajouter les informations d'identification AWS à Workload Factory avec des autorisations de lecture/écriture, ce qui signifie que vous utiliserez Workload Factory en mode *lecture/écriture* pour GenAI.

Les autorisations en mode *de base* et en mode *lecture seule* ne sont pas prises en charge pour le moment.

Lors de la configuration de vos identifiants, le fait de sélectionner les autorisations comme indiqué ci-dessous vous permet de bénéficier d'un accès complet pour gérer les systèmes de fichiers FSX pour ONTAP, ainsi que pour déployer et gérer l'instance EC2 GenAI et d'autres ressources AWS nécessaires à votre base de connaissances et à votre chatbot.

"Découvrez comment ajouter des informations d'identification AWS à Workload Factory"

Configuration requise pour NetApp Connector pour Amazon Q Business

Assurez-vous que votre environnement répond aux exigences spécifiques suivantes pour les connecteurs Amazon Q Business.

Application Amazon Q Business

Vous devez créer une application Amazon Q Business ou utiliser une application existante.

- Assurez-vous que l'application existe dans l'une de vos régions AWS.
- Assurez-vous que vous avez "créé un index" pour l'application.
- Assurez-vous que l'application n'est pas en état d'échec.

FSX pour le système de fichiers ONTAP

Vous avez besoin d'au moins un système de fichiers FSX pour ONTAP :

- Un système de fichiers sera utilisé (ou créé, s'il n'existe pas) par le moteur NetApp GenAI pour stocker des informations sur le connecteur.

  Ce système de fichiers FSX pour ONTAP doit utiliser des volumes FlexVol. Les volumes FlexGroup ne sont pas pris en charge.

- Un ou plusieurs systèmes de fichiers contiennent les sources de données que vous allez ajouter à votre connecteur.

  Un système de fichiers FSX pour ONTAP peut être utilisé pour ces deux raisons ou vous pouvez utiliser plusieurs systèmes de fichiers FSX pour ONTAP.

- Vous devez connaître la région AWS, le VPC et le sous-réseau où réside le système de fichiers AWS FSX pour ONTAP.
- Vous devrez tenir compte des paires clé/valeur que vous souhaitez appliquer aux ressources AWS faisant partie de ce déploiement (facultatif).
- Vous devrez connaître les informations relatives aux paires de clés qui vous permettront de vous connecter en toute sécurité à l'instance du moteur NetApp ai.

"Découvrez comment déployer et gérer les systèmes de fichiers FSX pour ONTAP"

Identifier les sources de données à ajouter à un connecteur

Identifiez ou créez les documents (sources de données) qui résident dans votre système de fichiers FSX pour ONTAP que vous allez intégrer à votre connecteur. Ces sources de données permettent à Amazon Q Business de fournir des réponses précises et personnalisées aux questions des utilisateurs en fonction des données pertinentes pour votre entreprise.

Nombre maximum de sources de données

Le nombre maximum de sources de données prises en charge est de 10.

Emplacement des sources de données

Les sources de données peuvent être stockées dans un seul volume, ou dans un dossier au sein d'un volume, sur un partage SMB ou une exportation NFS sur un système de fichiers Amazon FSX pour NetApp ONTAP. Les sources de données peuvent également être stockées dans des volumes Amazon FSX pour NetApp ONTAP dans une relation de protection des données NetApp SnapMirror.

Vous ne pouvez pas sélectionner des documents individuels au sein d'un volume ou d'un dossier. Par conséquent, vous devez vous assurer que chaque volume ou dossier contenant des sources de données ne contient pas de documents externes qui ne doivent pas être intégrés à votre base de connaissances.

Vous pouvez ajouter plusieurs sources de données dans chaque connecteur, mais elles doivent résider sur des systèmes de fichiers FSX pour ONTAP accessibles depuis votre compte AWS.

La taille de fichier maximale pour chaque source de données est de 50 Mo.

Protocoles pris en charge

Les connecteurs prennent en charge les données des volumes qui utilisent les protocoles NFS ou SMB/CIFS. Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory pour que le connecteur puisse accéder aux fichiers de ces volumes. Ceci inclut le domaine Active Directory, l'adresse IP, le nom d'utilisateur et le mot de passe.

Lorsque vous stockez votre source de données sur un partage (fichier ou répertoire) accessible via SMB, les données sont uniquement accessibles aux utilisateurs ou groupes de chatbot qui disposent des autorisations d'accès à ce partage. Lorsque cette « fonctionnalité de reconnaissance des autorisations » est activée, le système d'IA compare le courrier électronique de l'utilisateur dans auth0 aux utilisateurs autorisés à afficher ou utiliser les fichiers sur le partage SMB. Le chatbot fournira des réponses en fonction des autorisations utilisateur pour les fichiers intégrés.

Par exemple, si vous avez intégré 10 fichiers (sources de données) dans votre connecteur et que 2 fichiers

sont des fichiers de ressources humaines qui contiennent des informations restreintes, seuls les utilisateurs de chatbot qui sont authentifiés pour accéder à ces 2 fichiers recevront des réponses du chatbot qui incluent des données de ces fichiers.

Lorsque vous ajoutez des sources de données à Amazon Q Business Connector, seules les autorisations utilisateur s'appliquent aux fichiers de source de données. Les autorisations de groupe ne sont pas appliquées.

Si un fichier de votre source de données manque de texte (par exemple, une image sans texte), Amazon Q Business ne l'indexe pas mais consigne une entrée dans les journaux Amazon CloudWatch en notant l'absence de texte.

Formats de fichiers de source de données pris en charge

Les formats de fichiers de source de données suivants sont actuellement pris en charge avec NetApp Connector pour Amazon Q Business.

Format de fichier	Extension
Fichier de valeurs séparées par des virgules	.csv
JSON et JSONP	.json
Démarque	.md
Microsoft Word	.docx
Texte brut	.txt
Format de document portable	.pdf
Microsoft PowerPoint	.ppt ou .pptx
Langage de balisage hypertexte	.html
Langage de balisage extensible	.xml
XSLT	.xslt
Microsoft Excel	.xls
Format texte riche	.rtf

Déployez l'infrastructure GenAI

Vous devez déployer l'infrastructure GenAI pour la structure RAG dans votre environnement avant de pouvoir créer des bases de connaissances, des connecteurs et des applications FSX pour ONTAP pour votre entreprise. Les principaux composants de l'infrastructure sont le service Amazon Bedrock, une instance de machine virtuelle pour le moteur NetApp GenAI et un système de fichiers FSX pour ONTAP.

L'infrastructure déployée peut prendre en charge plusieurs bases de connaissances, chatbots et connecteurs. Vous n'aurez donc généralement besoin d'effectuer cette tâche qu'une seule fois.

Détails de l’infrastructure

Votre déploiement GenAI doit se trouver dans une région AWS sur laquelle Amazon Bedrock est activé. "Afficher la liste des régions prises en charge"

L’infrastructure comprend les composants suivants.

Service Amazon Bedrock

Amazon Bedrock est un service entièrement géré qui vous permet d’utiliser les modèles de base (FMS) des plus grandes entreprises d’IA via une seule API. Il fournit également les fonctionnalités nécessaires pour créer des applications d’IA générative sécurisées.

"En savoir plus sur Amazon Bedrock"

Amazon Q Business

Amazon Q s’appuie sur Amazon Bedrock pour fournir un assistant d’IA générative entièrement géré que vous pouvez utiliser pour répondre à des questions et générer du contenu à partir d’informations issues de vos sources de données.

"En savoir plus sur Amazon Q Business"

Machine virtuelle pour le moteur NetApp GenAI

Le moteur NetApp GenAI est déployé au cours de ce processus. Il fournit la puissance de traitement permettant d’acquérir les données de vos sources de données, puis d’écrire ces données dans la base de données vectorielle.

FSX pour le système de fichiers ONTAP

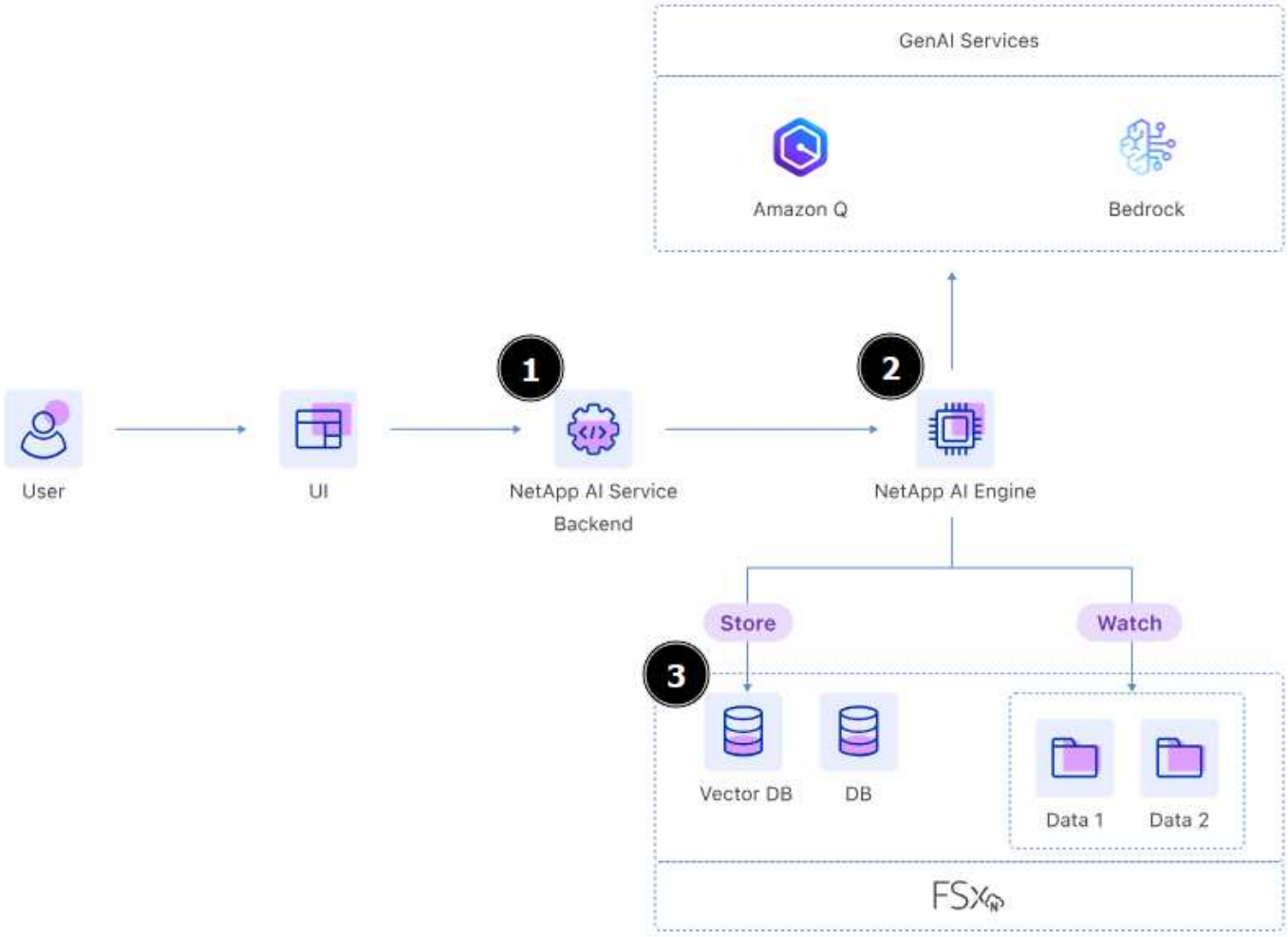
Le système de fichiers FSX pour ONTAP fournit le stockage de votre système GenAI.

Un seul volume est déployé qui contiendra la base de données vectorielle qui stocke les données générées par le modèle de base basé sur vos sources de données.

Les sources de données que vous intégrerez dans votre base de connaissances peuvent résider sur le même système de fichiers FSX pour ONTAP ou sur un autre système.

Le moteur NetApp GenAI surveille ces volumes et interagit avec eux.

L’image suivante montre l’infrastructure GenAI. Les composants numérotés 1, 2 et 3 sont déployés au cours de cette procédure. Les autres éléments doivent être en place avant de commencer le déploiement.



Déployez l'infrastructure GenAI

Vous devrez saisir vos identifiants AWS et sélectionner le système de fichiers FSX pour ONTAP pour déployer l'infrastructure de génération augmentée de récupération.

Avant de commencer

Avant de commencer cette procédure, assurez-vous que votre environnement répond aux conditions requises pour les bases de connaissances ou les connecteurs, selon ce que vous choisissez.

- "Exigences de la base de connaissances"
- "Exigences relatives au connecteur"

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Examinez le diagramme d'infrastructure et sélectionnez **Suivant**.
4. Complétez les éléments de la section **paramètres AWS** :
   a. **Informations d'identification AWS** : sélectionnez ou ajoutez les informations d'identification AWS qui fournissent des autorisations pour déployer les ressources AWS.
   b. **Emplacement** : sélectionnez une région AWS, un VPC et un sous-réseau.

Le déploiement GenAI doit se trouver dans une région AWS sur laquelle le système Amazon Bedrock est activé. "Afficher la liste des régions prises en charge"

5. Complétez les éléments de la section **Paramètres de l'infrastructure** :
   a. **Tags** : saisissez les paires clé/valeur de balise que vous souhaitez appliquer à toutes les ressources AWS qui font partie de ce déploiement. Ces balises sont visibles dans la console de gestion AWS et dans la zone d'informations sur l'infrastructure de Workload Factory et peuvent vous aider à suivre les ressources de Workload Factory.
6. Complétez la section **connectivité** :
   a. **Paire de clés** : sélectionnez une paire de clés qui vous permet de vous connecter en toute sécurité à l'instance du moteur NetApp GenAI.
7. Compléter la section **ai Engine** :
   a. **Nom de l'instance** : si vous le souhaitez, sélectionnez **Définir le nom de l'instance** et saisissez un nom personnalisé pour l'instance du moteur d'IA. Le nom de l'instance apparaît dans la console de gestion AWS et dans la zone d'informations sur l'infrastructure de Workload Factory et peut vous aider à suivre les ressources de Workload Factory.
8. Sélectionnez **déployer** pour commencer le déploiement.

> Si le déploiement échoue avec une erreur d'informations d'identification, vous pouvez obtenir plus de détails sur l'erreur en sélectionnant les liens hypertexte dans le message d'erreur. Vous pouvez voir une liste d'autorisations manquantes ou bloquées, ainsi qu'une liste d'autorisations dont la charge de travail GenAI a besoin pour pouvoir déployer l'infrastructure GenAI.

Résultat

Workload Factory commence à déployer l'infrastructure du chatbot. Ce processus peut prendre jusqu'à 10 minutes.

Au cours du processus de déploiement, les éléments suivants sont configurés :

- Le réseau est configuré avec les terminaux privés.
- Le rôle IAM, le profil d'instance et le groupe de sécurité sont créés.
- L'instance de machine virtuelle du moteur GenAI est déployée.
- Amazon Bedrock est configuré pour envoyer des journaux à Amazon CloudWatch, en utilisant un groupe de journaux avec le préfixe `/aws/bedrock/`.
- Le moteur GenAI est configuré pour envoyer des journaux à Amazon CloudWatch Logs, à l'aide d'un groupe de journaux portant le nom `/netapp/wlmai/<tenancyAccountId>/randomId` , où `<tenancyAccountID>` est le "ID de compte de la console NetApp" pour l'utilisateur actuel.

Créer un connecteur NetApp pour Amazon Q Business

Une fois que vous avez déployé l'infrastructure d'IA et identifié les sources de données que vous utiliserez à partir de vos magasins de données FSx for ONTAP, vous êtes prêt à définir un connecteur NetApp pour Amazon Q Business.

Avant de continuer, assurez-vous que votre environnement est conforme à "de formation"pour Amazon Q Business.

Description de la tâche

Les sources de données de votre organisation peuvent contenir des informations personnelles identifiables (PII). Pour protéger ces informations sensibles, vous pouvez activer les *garde-fous de données* lors de la définition d'un connecteur. Les garde-fous de données, optimisés par NetApp Data Classification, identifient et masquent les informations personnelles identifiables, les rendant inaccessibles et irrécupérables.

"En savoir plus sur la classification des données NetApp" .

NetApp Workload Factory pour GenAI ne masque pas les informations personnelles sensibles (SPII). Se référer à"types de données personnelles sensibles" pour plus d'informations sur ce type de données.

Les garde-fous de données peuvent être activés ou désactivés à tout moment. Si vous activez les garde-fous de données, Workload Factory analyse l'intégralité de la source de données à partir de zéro, ce qui peut entraîner un coût.

Définir un connecteur

Créez un connecteur NetApp pour Amazon Q Business. Ce connecteur permet la communication entre les API et les sources de données entre GenAI et Amazon Q Business.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans le menu Bases de connaissances et connecteurs, sélectionnez la liste déroulante **Créer un nouveau** et choisissez **Connecteur Amazon Q Business**.
4. Sur la page définir le connecteur, configurez les paramètres du connecteur :
   a. **Nom** : saisissez le nom que vous souhaitez utiliser pour le connecteur.
   b. **Description** : saisissez une description détaillée du connecteur.
   c. **Amazon Q** : région et nom d'application de l'instance Amazon Q Business que vous souhaitez intégrer.
   d. **Protections de données** : Choisissez si vous souhaitez activer ou désactiver les protections de données. "Découvrez les garde-fous de données, optimisés par la classification des données NetApp" .

      Les conditions préalables suivantes doivent être remplies pour activer les rails de données.

      - Un compte de service est requis pour communiquer avec NetApp Data Classification. Vous devez disposer du rôle *Administrateur de l'organisation* sur votre compte de location NetApp Console pour la création d'un compte de service. Un membre disposant du rôle d'administrateur d'organisation peut effectuer toutes les actions dans la console NetApp ."Découvrez comment ajouter un rôle à un membre dans la console NetApp"
      - Le moteur d'IA doit avoir accès à la"Point de terminaison de l'API de la console NetApp" .
      - Vous devrez effectuer les opérations suivantes comme décrit dans"Documentation sur la classification des données NetApp" :
        A. Créer un agent de console
        B. Assurez-vous que votre environnement peut répondre aux conditions préalables
        C. Déployer la classification des données NetApp

Lorsque vous activez la fonction Data guardrails, GenAI traite les fichiers .txt, .md, .csv, .docx et .PDF en ingérant uniquement le texte brut (à l'exception de l'image intégrée ou du texte multimédia) et en masquant les données privées ou sensibles. Tous les autres types de fichiers sont traités normalement sans masquer les données privées ou sensibles.

e. *Système de fichiers FSx pour ONTAP * : lorsque vous définissez un nouveau connecteur NetApp pour Amazon Q Business, Workload Factory crée un nouveau volume Amazon FSx for NetApp ONTAP pour stocker les informations du connecteur. Choisissez un système de fichiers existant et une SVM (également appelée VM de stockage) où le nouveau volume sera créé.

f. **Politique de capture instantanée** : choisissez une politique de capture instantanée dans la liste des politiques existantes définies dans l'inventaire de stockage Workload Factory. GenAI crée automatiquement des instantanés récurrents du volume stockant les informations du connecteur à une fréquence basée sur la politique d'instantané que vous sélectionnez.

   Si la règle de snapshot dont vous avez besoin n'existe pas, vous pouvez "création d'une règle de snapshots" la faire sur la machine virtuelle de stockage qui contient le volume.

5. Sélectionnez **Créer un connecteur** pour intégrer Amazon Q Business à GenAI.

   Un indicateur de progression s'affiche pendant la création du connecteur.

   Une fois le connecteur créé, vous avez la possibilité d'ajouter une source de données au connecteur afin qu'Amazon Q Business ingère vos données et les ajoute à son index. Nous vous recommandons de sélectionner **Ajouter une source de données** et d'ajouter une ou plusieurs sources de données maintenant.

Ajoutez des sources de données au connecteur

Vous pouvez ajouter une ou plusieurs sources de données pour remplir l'index Amazon Q Business avec les données de votre organisation.

Description de la tâche

- Le nombre maximum de sources de données prises en charge est de 10.
- Reportez-vous "Documentation Amazon Q Business" au pour connaître les restrictions de service spécifiques de l'indice Amazon Q Business.

Étapes

1. Après avoir sélectionné **Ajouter une source de données**, la page **Sélectionner un système de fichiers** apparaît.
2. **Sélectionnez un système de fichiers** : sélectionnez le système de fichiers FSX pour ONTAP dans lequel résident vos fichiers source de données et sélectionnez **Suivant**.
3. **Sélectionnez un volume** : sélectionnez le volume sur lequel vos fichiers de source de données résident et sélectionnez **Suivant**.

   Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory, notamment le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.

4. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

5. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.
     - Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

     Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

6. Dans la section **permission Aware**, disponible uniquement lorsque la source de données que vous avez sélectionnée se trouve sur un volume qui utilise le protocole SMB, vous pouvez activer ou désactiver les réponses sensibles aux autorisations :
   - **Activé**: Les utilisateurs du chatbot qui accèdent à ce connecteur n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
   - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.

     Les autorisations de groupe Active Directory ne sont pas prises en charge pour les sources de données Amazon Q Business Connector.

7. Sélectionnez **Ajouter** pour ajouter cette source de données à Amazon Q Business Connector.

Résultat

La source de données est intégrée dans l'indice Amazon Q Business. L'état passe de « intégration » à « intégrée » lorsque la source de données est complètement intégrée.

Après avoir ajouté une seule source de données au connecteur, vous pouvez la tester dans l'environnement chatbot Amazon Q Business et apporter les modifications requises avant de mettre le service à la disposition de vos utilisateurs. Vous pouvez également suivre les mêmes étapes pour ajouter des sources de données supplémentaires au connecteur.

Administration et contrôle

Gérer l’infrastructure GenAI

Vous pouvez afficher les détails sur votre infrastructure RAG GenAI déployée ou supprimer l’infrastructure de chatbot si vous n’en avez plus besoin.

Afficher des informations sur l’infrastructure

Vous pouvez afficher des informations sur l’infrastructure du chatbot

Étapes

1. Connectez-vous à Workload Factory en utilisant l’un des"expériences de la console" .
2. Dans la mosaïque workloads d’IA, sélectionnez **déployer et gérer**.
3. Sélectionnez le menu **Infrastructure**.
4. Afficher des informations sur l’infrastructure, qui incluent des détails sur les composants suivants :
    - Paramètres AWS
    - Paramètres de l’infrastructure
    - Le moteur ai
    - La base de données vectorielle

Déposer l’infrastructure

Si vous n’avez plus besoin de l’infrastructure de chatbot que vous avez déployée pour un ou plusieurs chatbots, vous pouvez la supprimer de Workload Factory.

> Tous les chatbots déployés sur cette infrastructure seront désactivés et tous les historiques de discussion seront supprimés.

Cette opération supprime uniquement les liens vers l’infrastructure AI de Workload Factory ; elle ne supprime pas tous les composants d’AWS. Vous devrez supprimer manuellement les composants d’infrastructure suivants d’AWS :

- L’instance de VM
- Terminaux privés
- Volume du système de fichiers FSX pour ONTAP qui contient les bases de données d’IA
- Le rôle IAM
- La politique
- Le groupe de sécurité

Étapes

1. Connectez-vous à Workload Factory en utilisant l’un des"expériences de la console" .
2. Dans la mosaïque workloads d’IA, sélectionnez **déployer et gérer**.
3. Sélectionnez le menu **Infrastructure**.

4. Sélectionnez et sélectionnez **Supprimer l'infrastructure chatbot**.
5. Confirmez que vous souhaitez supprimer l'infrastructure et sélectionnez **Supprimer**.

Résultat

Les composants de l'infrastructure du chatbot sont supprimés de Workload Factory.

Gérer les bases de connaissances GenAI

Après avoir créé une base de connaissances, vous pouvez afficher les détails de la base de connaissances, modifier la base de connaissances, intégrer d'autres sources de données ou supprimer la base de connaissances.

Afficher des informations sur une base de connaissances

Vous pouvez afficher des informations sur les paramètres d'une base de connaissances et sur la source de données intégrée.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Sélectionnez la base de connaissances que vous souhaitez afficher.

   Si elle est définie, les amorces de conversation actuellement utilisées s'affichent dans le volet de droite.

4. Pour afficher les détails de la base de connaissances, sélectionnez et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

   Le menu **actions** vous permet de gérer la base de connaissances si vous souhaitez apporter des modifications.

Modifier une base de connaissances

Vous pouvez mettre à jour une base de connaissances en modifiant certains paramètres ou ajouter ou supprimer des sources de données.

Chaque fois que vous ajoutez, modifiez ou supprimez des sources de données de la base de connaissances, vous devez synchroniser la source de données afin qu'elle soit réindexée dans la base de connaissances. La synchronisation est incrémentielle. Amazon Bedrock ne traite donc que les objets de votre volume FSX pour ONTAP qui ont été ajoutés, modifiés ou supprimés depuis la dernière synchronisation.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez mettre à jour.
4. Sélectionnez et sélectionnez **gérer la base de connaissances**.

Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

5. Sélectionnez le menu **actions** et sélectionnez **Modifier la base de connaissances**.
6. Dans la page Modifier la base de connaissances, vous pouvez modifier le nom de la base de connaissances, la description, le modèle d'intégration, le modèle de discussion, l'activation des fonctionnalités, choisir si les démarreurs de conversation sont créés automatiquement ou manuellement et la stratégie de capture instantanée utilisée pour le volume qui contient la base de connaissances.

   Si vous utilisez le mode Manuel pour les amorces de conversation, vous pouvez également modifier les amorces de conversation ici.

   > Chaque analyse de la base de connaissances, qui inclut l'intégration, coûts. Si des rails de protection des données sont activés après la création d'une base de connaissances, celle-ci est analysée à nouveau et entraîne des coûts. De même, si vous modifiez les modèles de chat, GenAI analyse à nouveau les sources de données associées (ce qui engendre un coût).

7. Sélectionnez **Enregistrer** après avoir effectué vos modifications.

Protégez une base de connaissances avec des instantanés

Vous pouvez protéger les données de votre base de connaissances en prenant et en restaurant des instantanés des volumes de votre base de connaissances. Vous pouvez à tout moment restaurer à partir d'un instantané pour revenir à la version précédente de la base de connaissances.

Les copies Snapshot peuvent être plus rapides et plus efficaces en termes de stockage que les sauvegardes et vous permettent de protéger chaque base de connaissances à l'aide d'une règle de protection différente. Voici quelques-uns des scénarios dans lesquels les snapshots peuvent être utiles :

- Perte ou corruption accidentelle des données
- Récupération à partir de données incorrectes ingérées dans la base de connaissances
- Test de différentes sources de données ou stratégies de suppression, et rétablissement rapide une fois les tests terminés

Prenez un instantané d'un volume de la base de connaissances

Vous pouvez enregistrer l'état d'une base de connaissances en prenant un instantané manuel du volume de la base de connaissances.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez protéger.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

5. Sélectionnez le menu **actions** et sélectionnez **instantané > Créer un nouvel instantané**.

6. Si vous le souhaitez, sélectionnez **définir le nom de l'instantané** et entrez un nom personnalisé pour l'instantané.

   La définition d'un nom personnalisé peut vous aider à mieux déterminer le contenu d'un instantané si vous devez le restaurer ultérieurement.

7. Sélectionnez **Créer**.

   Un instantané de la base de connaissances est créé.

Restaurer un snapshot d'un volume de la base de connaissances

Vous pouvez à tout moment restaurer un snapshot manuel ou planifié d'un volume de la base de connaissances.

> Vous ne pouvez pas restaurer un snapshot à l'aide de l'interface utilisateur des workloads Generative ai si la base de données stockée sur le volume est corrompue ou supprimée. Pour contourner ce problème, vous pouvez restaurer le snapshot à l'aide de "INTERFACE DE LIGNE DE COMMANDES DE ONTAP" sur le cluster ONTAP où le volume est hébergé.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez restaurer.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

5. Sélectionnez le menu **actions** et sélectionnez **instantané > Restaurer instantané**.

   La boîte de dialogue de sélection d'instantané s'affiche, dans laquelle vous pouvez afficher la liste des instantanés créés pour cette base de connaissances.

6. (Facultatif) désélectionnez l'option **suspendre l'exécution et les analyses programmées après la restauration de l'instantané** si vous souhaitez que les analyses de source de données planifiées et en cours d'exécution continuent après la restauration de l'instantané.

   Cette option est activée par défaut pour garantir qu'une analyse ne se produit pas lorsque la base de connaissances est en état partiellement restauré ou qu'une analyse ne met pas à jour une base de connaissances récemment restaurée avec des données plus anciennes.

7. Sélectionnez l'instantané à restaurer dans la liste.
8. Sélectionnez **Restaurer**.

Cloner une base de connaissances

Vous pouvez créer une nouvelle base de connaissances à partir d'un instantané de la base de connaissances. Ceci est utile si la base de connaissances d'origine est corrompue ou perdue.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez restaurer.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

5. Sélectionnez le menu **actions** et sélectionnez **instantané > Cloner la base de connaissances**.

   La boîte de dialogue clone s'affiche.

6. Si vous le souhaitez, désélectionnez l'option **interrompre l'exécution et les analyses planifiées après le clonage de l'instantané** si vous souhaitez que les analyses de source de données planifiées et en cours d'exécution continuent après le clonage de l'instantané.

   Cette option est activée par défaut pour garantir qu'une analyse ne se produit pas lorsque la base de connaissances est en état partiellement restauré ou qu'une analyse ne met pas à jour une base de connaissances récemment restaurée avec des données plus anciennes.

7. Sélectionnez l'instantané à cloner dans la liste.
8. Sélectionnez **Continuer**.
9. Entrez un nom pour la nouvelle base de connaissances.
10. Choisir une SVM de système de fichiers et un nom de volume pour la nouvelle base de connaissances.
11. Sélectionnez **Clone**.

Ajouter des sources de données supplémentaires à une base de connaissances

Vous pouvez intégrer d'autres sources de données dans votre base de connaissances pour les remplir avec des données d'organisation supplémentaires.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances dans laquelle vous souhaitez ajouter la source de données.
4. Sélectionnez ••• et sélectionnez **Ajouter une source de données**.
5. Sélectionnez le type de source de données que vous souhaitez ajouter :
   - Ajouter un système de fichiers FSx pour ONTAP (utiliser les fichiers d'un volume FSx pour ONTAP existant)
   - Ajouter un système de fichiers (utiliser des fichiers provenant d'un partage SMB ou NFS générique)

Ajouter un système de fichiers FSx pour ONTAP

1. **Sélectionnez un système de fichiers** : sélectionnez le système de fichiers FSX pour ONTAP dans lequel résident vos fichiers source de données et sélectionnez **Suivant**.
2. **Sélectionnez un volume** : sélectionnez le volume sur lequel vos fichiers de source de données résident et sélectionnez **Suivant**.

   Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory, notamment le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
3. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.
4. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

   Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

5. Dans la section **permission Aware**, disponible uniquement lorsque la source de données que vous avez sélectionnée se trouve sur un volume qui utilise le protocole SMB, vous pouvez activer ou désactiver les réponses sensibles aux autorisations :
   - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
   - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
6. Sélectionnez **Ajouter** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers NFS générique

1. **Sélectionnez un système de fichiers** : saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données, choisissez le protocole NFS pour le partage réseau et sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom de l'exportation NFS et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner l'exportation dans son intégralité ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :
   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50

et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

> Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

- **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
  - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

    Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

  Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

  > Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

  > Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers SMB générique

1. **Sélectionnez le système de fichiers**:
   a. Saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données.
   b. Choisissez le protocole SMB pour le partage réseau.
   c. Saisissez les informations Active Directory, qui incluent le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
   d. Sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom du partage SMB et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner le partage entier ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

   - **Conscient des autorisations** : Activer ou désactiver les réponses conscientes des autorisations :
     - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
     - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

     Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

     Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

   - Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de

l'heure de modification, sélectionnez une plage de dates dans la liste.

> Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Résultat

La source de données est intégrée à votre base de connaissances.

Synchronisez vos sources de données avec une base de connaissances

Les sources de données sont automatiquement synchronisées avec la base de connaissances associée une fois par jour, de sorte que les modifications éventuelles des sources de données soient répercutées dans le chatbot. Si vous modifiez l'une de vos sources de données et que vous souhaitez synchroniser les données immédiatement, vous pouvez effectuer une synchronisation à la demande.

La synchronisation est incrémentielle. Amazon Bedrock ne traite donc que les objets de vos sources de données qui ont été ajoutés, modifiés ou supprimés depuis la dernière synchronisation.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez synchroniser.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.
5. Sélectionnez le menu **actions** et sélectionnez **Rechercher maintenant**.

   Vous verrez un message indiquant que vos sources de données sont en cours d'analyse et un message final lorsque l'analyse est terminée.

Résultat

La base de connaissances est synchronisée avec les sources de données jointes et tout chatbot actif commencera à utiliser les informations les plus récentes provenant de vos sources de données.

Mettre en pause ou reprendre une synchronisation planifiée

Si vous souhaitez interrompre ou reprendre la prochaine synchronisation (analyse) des sources de données, vous pouvez le faire à tout moment. Vous devrez peut-être interrompre la prochaine synchronisation programmée si vous allez apporter des modifications à une source de données et que vous ne souhaitez pas que la synchronisation se produise pendant la fenêtre de modification.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans le menu Bases de connaissances et connecteurs, sélectionnez la base de connaissances pour

laquelle vous souhaitez suspendre ou reprendre les analyses.

4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.
5. Sélectionnez le menu **actions** et sélectionnez **Scan > Pause scan planifié** ou **Scan > reprendre scan planifié**.

   Un message vous indique que la prochaine analyse programmée a été interrompue ou reprise.

Évaluez les modèles de chat avant de créer une base de connaissances

Vous pouvez évaluer les modèles de chat de base disponibles avant de créer une base de connaissances afin de déterminer le modèle le mieux adapté à votre implémentation. Étant donné que la prise en charge des modèles varie selon les régions AWS, reportez-vous à la section "Cette page de documentation AWS" pour vérifier les modèles que vous pouvez utiliser dans les régions où vous prévoyez de déployer votre base de connaissances.

> Cette fonctionnalité n'est disponible que lorsqu'aucune base de connaissances n'a été créée — lorsqu'aucune base de connaissances n'existe dans la page d'inventaire des bases de connaissances.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Sur la page d'inventaire des bases de connaissances, vous verrez l'option de sélectionner le modèle de chat sur le côté droit de la page pour le Chatbot.
4. Sélectionnez le modèle de chat dans la liste et entrez un ensemble de questions dans la zone d'invite pour voir comment le chatbot répond.
5. Essayez plusieurs modèles pour identifier le modèle le mieux adapté à votre implémentation.

Résultat

Utilisez ce modèle de chat lorsque vous créez votre base de connaissances.

Annulez la publication de votre base de connaissances

Après avoir publié votre base de connaissances afin qu'elle puisse être intégrée à une application chatbot, vous pouvez annuler la publication si vous souhaitez désactiver l'application chatbot pour l'accès à la base de connaissances.

L'annulation de la publication de la base de connaissances empêche toute application de chat de fonctionner. Le point de terminaison unique de l'API auquel la base de connaissances était accessible est désactivé.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez annuler la publication.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration et la liste de toutes les sources de données intégrées.

5. Sélectionnez le menu **actions** et sélectionnez **Annuler la publication**.

Résultat

La base de connaissances est désactivée et n'est plus accessible par une application chatbot.

Supprimer une base de connaissances

Si vous n'avez plus besoin d'une base de connaissances, vous pouvez la supprimer. Lorsque vous supprimez une base de connaissances, elle est supprimée de Workload Factory et le volume qui contient la base de connaissances est supprimé. Toutes les applications ou chatbots utilisant la base de connaissances cesseront de fonctionner. La suppression d'une base de connaissances n'est pas réversible.

Lorsque vous supprimez une base de connaissances, vous devez également dissocier la base de connaissances de tous les agents auxquels elle est associée afin de supprimer entièrement toutes les ressources associées à la base de connaissances.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances que vous souhaitez supprimer.
4. Sélectionnez ••• et sélectionnez **gérer la base de connaissances**.
5. Sélectionnez le menu **actions** et sélectionnez **Supprimer la base de connaissances**.
6. Dans la boîte de dialogue Supprimer la base de connaissances, confirmez que vous souhaitez la supprimer et sélectionnez **Supprimer**.

Résultat

La base de connaissances est supprimée de Workload Factory et son volume associé est supprimé.

Gérer les connecteurs Amazon Q Business

Après avoir créé un connecteur pour Amazon Q Business, vous pouvez afficher les détails du connecteur, modifier le connecteur, intégrer d'autres sources de données ou supprimer le connecteur.

Afficher des informations sur un connecteur

Vous pouvez afficher des informations sur les paramètres d'un connecteur et les sources de données intégrées.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances et des connecteurs, sélectionnez le connecteur que vous souhaitez afficher.
4. Pour afficher les détails du connecteur, sélectionnez ••• et sélectionnez **gérer le connecteur**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste

de toutes les sources de données intégrées, etc.

Le menu **actions** vous permet de gérer le connecteur si vous souhaitez apporter des modifications.

Editer un connecteur

Vous pouvez mettre à jour un connecteur en modifiant certains paramètres ou ajouter ou supprimer des sources de données.

Chaque fois que vous ajoutez, modifiez ou supprimez des sources de données du connecteur, GenAI doit envoyer les informations de la source de données à Amazon Q Business afin qu'elles soient réindexées. La synchronisation est incrémentielle. Amazon Q Business ne traite donc que les objets de votre volume FSX pour ONTAP qui ont été ajoutés, modifiés ou supprimés depuis la dernière synchronisation.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire bases de connaissances et connecteurs, sélectionnez le connecteur à mettre à jour.
4. Sélectionnez ••• et sélectionnez **gérer le connecteur**.

   Cette page affiche l'état publié, l'état d'intégration des sources de données, le mode d'intégration, la liste de toutes les sources de données intégrées, etc.

5. Sélectionnez le menu **actions** et sélectionnez **Modifier le connecteur**.
6. Dans la page Modifier le connecteur, vous pouvez modifier le nom du connecteur, la description, le modèle d'intégration, l'activation des rails de données et la stratégie de snapshot utilisée pour le volume qui contient le connecteur.

   Chaque analyse de source de données, qui inclut l'intégration, implique un coût. Si vous activez les rails de données après la création d'un connecteur, la source de données est à nouveau analysée et génère des coûts supplémentaires.

7. Sélectionnez **Enregistrer** après avoir effectué des modifications.

Ajoutez des sources de données supplémentaires à un connecteur

Vous pouvez intégrer d'autres sources de données dans votre connecteur pour les remplir avec des données d'organisation supplémentaires.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire bases de connaissances et connecteurs, sélectionnez le connecteur dans lequel vous souhaitez ajouter la source de données.
4. Sélectionnez ••• et sélectionnez **Ajouter une source de données**.
5. Sélectionnez le type de source de données que vous souhaitez ajouter :
   - Ajouter un système de fichiers FSx pour ONTAP (utiliser les fichiers d'un volume FSx pour ONTAP existant)

◦ Ajouter un système de fichiers (utiliser des fichiers provenant d’un partage SMB ou NFS générique)

Ajouter un système de fichiers FSx pour ONTAP

1. **Sélectionnez un système de fichiers** : sélectionnez le système de fichiers FSX pour ONTAP dans lequel résident vos fichiers source de données et sélectionnez **Suivant**.
2. **Sélectionnez un volume** : sélectionnez le volume sur lequel vos fichiers de source de données résident et sélectionnez **Suivant**.

   Lorsque vous sélectionnez des fichiers stockés à l'aide du protocole SMB, vous devez entrer les informations Active Directory, notamment le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
3. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.
4. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

   Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

5. Dans la section **permission Aware**, disponible uniquement lorsque la source de données que vous avez sélectionnée se trouve sur un volume qui utilise le protocole SMB, vous pouvez activer ou désactiver les réponses sensibles aux autorisations :
   - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
   - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
6. Sélectionnez **Ajouter** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers NFS générique

1. **Sélectionnez un système de fichiers** : saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données, choisissez le protocole NFS pour le partage réseau et sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom de l'exportation NFS et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner l'exportation dans son intégralité ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :
   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50

et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

> Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

- **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
  - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

    Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

  Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

> Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

- Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de l'heure de modification, sélectionnez une plage de dates dans la liste.

> Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Ajouter un système de fichiers SMB générique

1. **Sélectionnez le système de fichiers**:
   a. Saisissez l'adresse IP ou le nom de domaine complet de l'hôte du système de fichiers où résident vos fichiers source de données.
   b. Choisissez le protocole SMB pour le partage réseau.
   c. Saisissez les informations Active Directory, qui incluent le domaine, l'adresse IP, le nom d'utilisateur et le mot de passe.
   d. Sélectionnez **Suivant**.
2. **Sélectionnez une source de données** : sélectionnez l'emplacement de la source de données en fonction de l'emplacement d'enregistrement des fichiers. Il peut s'agir d'un volume entier, ou simplement d'un dossier ou d'un sous-dossier spécifique dans le volume, et sélectionnez **Suivant**.

Dans certains cas, vous devrez peut-être saisir manuellement le nom du partage SMB et sélectionner « Récupérer les répertoires » pour afficher les répertoires disponibles. Vous pouvez sélectionner le partage entier ou seulement certains dossiers.

3. **Configurations** : configurez la façon dont la source de données ingère les informations de vos fichiers et les fichiers qu'elle inclut dans les analyses :

   - **Définir la source de données** : dans la section **Stratégie de partage**, définissez la façon dont le moteur GenAI divise le contenu de la source de données en blocs lorsque la source de données est intégrée à une base de connaissances. Vous pouvez choisir l'une des stratégies suivantes :
     - **Chunking à plusieurs phrases** : organise les informations de votre source de données en blocs définis par des phrases. Vous pouvez choisir combien de phrases composent chaque morceau (jusqu'à 100).
     - **Chunking basé sur le chevauchement** : organise les informations de votre source de données en blocs définis par des caractères qui peuvent chevaucher des blocs voisins. Vous pouvez choisir la taille de chaque bloc en caractères et la quantité de chaque bloc qui chevauche les blocs adjacents. Vous pouvez configurer une taille de bloc comprise entre 50 et 3000 caractères et un pourcentage de chevauchement compris entre 1 et 99 %.

       Le choix d'un pourcentage de chevauchement élevé peut considérablement augmenter les besoins de stockage avec seulement de légères améliorations de la précision de récupération.

   - **Conscient des autorisations** : Activer ou désactiver les réponses conscientes des autorisations :
     - **Activé** : les utilisateurs du chatbot qui accèdent à cette base de connaissances n'obtiennent que les réponses aux requêtes des sources de données auxquelles ils ont accès.
     - **Désactivé** : les utilisateurs du chatbot recevront des réponses en utilisant le contenu de toutes les sources de données intégrées.
   - **Filtrage de fichiers** : configurez les fichiers inclus dans les analyses :
     - Dans la section **prise en charge des types de fichiers**, choisissez soit d'inclure tous les types de fichiers, soit de sélectionner des types de fichiers individuels à inclure dans les analyses de sources de données.

       Si vous incluez des images ou des fichiers PDF, NetApp Workload Factory for GenAI analyse le texte dans les images (y compris les images dans les documents PDF), ce qui entraîne un coût plus élevé.

     Lors de l'inclusion de données texte à partir d'images, GenAI ne peut pas masquer les informations à caractère personnel (PII) de l'image car les données texte numérisées sont envoyées de votre environnement vers AWS. Cependant, une fois les données stockées, toutes les PII sont masquées dans la base de données GenAI.

     Votre choix d'inclure des fichiers image dans les analyses est lié au modèle de chat de la base de connaissances. Si vous incluez des fichiers image dans les numérisations, le modèle de chat doit prendre en charge les images. Si des types de fichiers d'image sont sélectionnés ici, vous ne pouvez pas passer de la base de connaissances à un modèle de chat qui ne prend pas en charge les fichiers d'image.

   - Dans la section **filtre de temps de modification de fichier**, choisissez d'activer ou de désactiver l'inclusion de fichiers en fonction de leur heure de modification. Si vous activez le filtrage de

l'heure de modification, sélectionnez une plage de dates dans la liste.

> Si vous incluez des fichiers basés sur une plage de dates de modification, dès que la plage de dates n'est pas satisfaite (les fichiers n'ont pas été modifiés dans la plage de dates spécifiée), les fichiers seront exclus de l'analyse périodique et la source de données n'inclura pas ces fichiers.

4. Sélectionnez **Ajouter une source de données** pour ajouter cette source de données à votre base de connaissances.

Résultat

La source de données est intégrée à votre connecteur.

Synchronisez vos sources de données avec un connecteur

Les sources de données sont automatiquement synchronisées avec le connecteur associé une fois par jour, de sorte que les modifications apportées à la source de données soient répercutées dans Amazon Q Business. Si vous apportez des modifications à l'une de vos sources de données et que vous souhaitez synchroniser (analyser) les données immédiatement, vous pouvez effectuer une synchronisation à la demande.

La synchronisation est incrémentielle. Amazon Q Business ne traite donc que les objets de vos sources de données qui ont été ajoutés, modifiés ou supprimés depuis la dernière synchronisation.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans le menu Bases de connaissances et connecteurs, sélectionnez le connecteur que vous souhaitez synchroniser.
4. Sélectionnez ••• et sélectionnez **gérer le connecteur**.
5. Sélectionnez le menu **actions** et sélectionnez **Rechercher maintenant**.

   Vous verrez un message indiquant que vos sources de données sont en cours d'analyse et un message final lorsque l'analyse est terminée.

Résultat

Le connecteur est synchronisé avec les sources de données jointes et Amazon Q Business commencera à utiliser les informations les plus récentes de vos sources de données.

Mettre en pause ou reprendre une synchronisation planifiée

Si vous souhaitez interrompre ou reprendre la prochaine synchronisation (analyse) des sources de données, vous pouvez le faire à tout moment. Vous devrez peut-être interrompre la prochaine synchronisation programmée si vous allez apporter des modifications à une source de données et que vous ne souhaitez pas que la synchronisation se produise pendant la fenêtre de modification.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.

3. Sur la page d'inventaire des connecteurs, sélectionnez le connecteur pour lequel vous souhaitez interrompre ou reprendre les numérisations.
4. Sélectionnez ••• et sélectionnez **gérer le connecteur**.
5. Sélectionnez le menu **actions** et sélectionnez **Scan > Pause scan planifié** ou **Scan > reprendre scan planifié**.

   Un message vous indique que la prochaine analyse programmée a été interrompue ou reprise.

Supprimer un connecteur

Si vous n'avez plus besoin d'un connecteur, vous pouvez le supprimer. Lorsque vous supprimez un connecteur, il est supprimé de Workload Factory et le volume qui contient le connecteur est supprimé. La suppression d'un connecteur n'est pas réversible.

Lorsque vous supprimez un connecteur, vous devez également dissocier le connecteur de tous les agents auxquels il est associé pour supprimer entièrement toutes les ressources associées au connecteur.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire bases de connaissances et connecteurs, sélectionnez le connecteur à supprimer.
4. Sélectionnez ••• et sélectionnez **gérer le connecteur**.
5. Sélectionnez le menu **actions** et sélectionnez **Supprimer le connecteur**.
6. Dans la boîte de dialogue Supprimer le connecteur, confirmez que vous souhaitez le supprimer et sélectionnez **Supprimer**.

Résultat

Le connecteur est supprimé de Workload Factory et son volume associé est supprimé.

Gérez les sources de données GenAI

Après avoir créé une base de connaissances ou un connecteur à l'aide de sources de données sur votre système de fichiers FSX pour ONTAP, vous pouvez afficher les détails de la source de données, mettre à jour ou modifier le contenu de la source de données, modifier les paramètres de la source de données ou supprimer la source de données.

Afficher des informations sur une source de données

Vous pouvez afficher des informations sur le contenu d'une source de données et afficher son état d'intégration avec la base de connaissances ou le connecteur. Étant donné que les sources de données sont associées à une base de connaissances ou à un connecteur, vous devez d'abord choisir la base de connaissances ou le connecteur avant de pouvoir afficher les détails de la source de données.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des "expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.

3. Sélectionnez la base de connaissances ou le connecteur où réside la source de données, puis sélectionnez **gérer la base de connaissances** ou **gérer le connecteur**.

   La partie inférieure de la page répertorie les sources de données associées.

4. Développez chaque ligne en sélectionnant pour afficher des informations détaillées sur chaque source de données, telles que le système de fichiers FSX pour ONTAP, le volume et le chemin d'accès à la source de données.

   Il répertorie également les informations d'intégration et indique si cette source de données est actuellement intégrée dans la base de connaissances ou le connecteur.

Modifier les paramètres de la source de données

Vous pouvez modifier les informations relatives à une source de données que vous avez intégrée à une base de connaissances ou à un connecteur. La plupart des informations sont corrigées après l'ajout d'une source de données, mais vous pouvez apporter des modifications à certaines configurations (telles que la définition de la suppression ou la reconnaissance des autorisations).

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire des bases de connaissances, sélectionnez la base de connaissances où réside la source de données, puis sélectionnez et sélectionnez **gérer la base de connaissances**.

   La partie inférieure de la page répertorie les sources de données qui font partie de cette base de connaissances.

4. Dans la ligne de la source de données que vous souhaitez modifier, sélectionnez et sélectionnez **Modifier la source de données**.
5. Dans la page Modifier la source de données, sélectionnez pour développer la ligne pour la définition de bloc.
6. Mettez à jour les paramètres de la stratégie et de la configuration de chunking, ainsi que la prise en compte des autorisations (pour les volumes SMB) et sélectionnez **Enregistrer**.

Résultat

Les paramètres de la source de données sont mis à jour et le système ai synchronise la source de données afin qu'elle soit réindexée à la base de connaissances.

Mettre à jour le contenu d'une source de données existante

Vous pouvez modifier le contenu d'une source de données à tout moment pour ajouter ou mettre à jour vos données organisationnelles. Si cette source de données est activement utilisée dans une base de connaissances, vous devez synchroniser la source de données afin qu'elle soit réindexée dans la base de connaissances. La synchronisation est incrémentielle. Amazon Bedrock ne traite donc que les objets de votre volume FSX pour ONTAP qui ont été ajoutés, modifiés ou supprimés depuis la dernière synchronisation.

Les sources de données sont synchronisées automatiquement une fois par jour avec la base de connaissances. Les modifications apportées aux sources de données sont donc répercutées dans le chatbot. Si vous apportez des modifications à une source de données et que vous souhaitez synchroniser les données immédiatement, vous pouvez "effectuez une synchronisation à la demande".

Supprimer une source de données

Si vous n'avez plus besoin d'une source de données pour faire partie de votre base de connaissances, vous pouvez la supprimer.

Étapes

1. Connectez-vous à Workload Factory en utilisant l'un des"expériences de la console" .
2. Dans la mosaïque workloads d'IA, sélectionnez **déployer et gérer**.
3. Dans la page d'inventaire de la base de connaissances, sélectionnez la base de connaissances où réside la source de données, puis ••• sélectionnez et sélectionnez **gérer la base de connaissances**.

   La partie inférieure de la page répertorie les sources de données qui font partie de cette base de connaissances.

4. Sur la ligne de la source de données à supprimer, sélectionnez ••• et sélectionnez **Supprimer la source de données**.
5. Dans la boîte de dialogue Supprimer la source de données, confirmez que vous souhaitez la supprimer et sélectionnez **confirmer**.

Résultat

La source de données est supprimée de la base de connaissances et le système ai supprime de la base de connaissances les informations indexées sur cette source de données. Les informations issues de cette source de données ne seront plus disponibles pour les chatbots qui utilisent la base de connaissances.

Surveillez les opérations de charge de travail avec Tracker dans NetApp Workload Factory

Surveillez et suivez l'exécution des opérations de charge de travail et surveillez la progression des tâches avec Tracker dans NetApp Workload Factory.

Description de la tâche

NetApp Workload Factory fournit Tracker, une fonctionnalité de surveillance, qui vous permet de surveiller et de suivre la progression et l'état des opérations de charge de travail, d'examiner les détails des tâches et sous-tâches d'opération et de diagnostiquer les problèmes ou les pannes.

Plusieurs actions sont disponibles dans Tracker. Vous pouvez filtrer les travaux par période (24 heures, 7 jours, 14 jours ou 30 jours), charge de travail, état et utilisateur, rechercher des travaux à l'aide de la fonction de recherche et télécharger la table des travaux sous forme de fichier CSV. Vous pouvez actualiser le Tracker à tout moment et réessayer rapidement une opération ayant échoué ou modifier les paramètres d'une opération ayant échoué et réessayer l'opération.

Tracker prend en charge deux niveaux de surveillance en fonction de l'opération. Chaque tâche, telle que le déploiement du système de fichiers, affiche la description de la tâche, l'état, l'heure de début, la durée de la tâche, l'utilisateur, la région, la ressource proxy, l'ID de la tâche et toutes les sous-tâches associées. Vous pouvez afficher les réponses de l'API pour comprendre ce qui s'est passé pendant l'opération.

Niveaux de tâche de suivi avec des exemples

- Niveau 1 (tâche) : assure le suivi du déploiement du système de fichiers.
- Niveau 2 (sous-tâche) : effectue le suivi des sous-tâches liées au déploiement du système de fichiers.

État de l'opération

L'état d'opération dans le Tracker est le suivant : *in Progress*, *Success* et *failed*.

Fréquence de fonctionnement

La fréquence des opérations est basée sur le type de travail et le planning de travail.

Conservation des événements

Les événements sont conservés dans l'interface utilisateur pendant 30 jours.

Suivez et surveillez les opérations

Suivez et surveillez les opérations dans la console Workload Factory avec Tracker.

Étapes

1. Connectez-vous à l'aide de l'un des "expériences de la console".
2. Dans le menu de charge de travail, sélectionnez **Administration** puis **Tracker**.
3. Dans Tracker, utilisez les filtres ou effectuez une recherche pour affiner les résultats des emplois. Vous pouvez également télécharger un rapport sur les emplois.

Afficher une demande d'API

Affichez la demande API dans la zone de code pour une tâche dans Tracker.

Étapes

1. Dans Tracker, sélectionnez une tâche.
2. Sélectionnez le menu Actions, puis sélectionnez **Afficher la demande API**.

Réessayez une opération ayant échoué

Réessayez une opération ayant échoué dans le Tracker. Vous pouvez également copier le message d'erreur d'une opération ayant échoué.

Vous pouvez réessayer une opération ayant échoué jusqu'à 10 fois.

Étapes

1. Dans Tracker, sélectionnez une opération ayant échoué.
2. Sélectionnez le menu Actions, puis sélectionnez **Réessayer**.

Résultat

L'opération est lancée à nouveau.

Modifiez et réessayez une opération ayant échoué

Modifiez les paramètres de l'opération ayant échoué et relancez l'opération en dehors du Tracker.

Étapes

1. Dans Tracker, sélectionnez une opération ayant échoué.
2. Sélectionnez le menu Actions, puis sélectionnez **Modifier et réessayer**.

Vous êtes redirigé vers la page d'opération où vous pouvez modifier les paramètres et réessayer l'opération.

Résultat

L'opération est lancée à nouveau. Accédez à Tracker pour afficher l'état de l'opération.

Connaissances et support

Inscrivez-vous pour bénéficier de l’assistance NetApp Workload Factory pour GenAI

L’enregistrement du support est requis pour bénéficier d’un support technique spécifique à NetApp Workload Factory et à ses solutions et services de stockage. Vous devez vous inscrire pour bénéficier de l’assistance à partir de la console NetApp , qui est une console Web distincte de Workload Factory.

L’inscription au support n’active pas la prise en charge NetApp pour un service de fichiers de fournisseur cloud. Pour obtenir une assistance technique relative à un service de fichiers de fournisseur cloud, à son infrastructure ou à toute solution utilisant le service, reportez-vous à « Obtenir de l’aide » dans la documentation Workload Factory de ce produit.

"Amazon FSX pour ONTAP"

Présentation de l’inscription au support

L’enregistrement de votre abonnement au support d’ID de compte (votre numéro de série 960xxxxxxxxx à 20 chiffres situé sur la page Ressources de support de la console NetApp) sert d’ID d’abonnement au support unique. Chaque abonnement de support au niveau du compte NetApp doit être enregistré.

L’inscription permet des fonctionnalités telles que l’ouverture de tickets d’assistance et la génération automatique de cas. L’enregistrement est terminé en ajoutant des comptes NetApp Support Site (NSS) à la console NetApp comme décrit ci-dessous.

Enregistrez votre compte pour bénéficier du support NetApp

Pour vous inscrire au support et activer le droit au support, un utilisateur de votre compte doit associer un compte de site de support NetApp à sa connexion à la console NetApp . La manière dont vous vous inscrivez au support NetApp dépend du fait que vous possédez déjà ou non un compte NetApp Support Site (NSS).

Client existant avec un compte NSS

Si vous êtes un client NetApp avec un compte NSS, il vous suffit de vous inscrire pour bénéficier de l’assistance via la console NetApp .

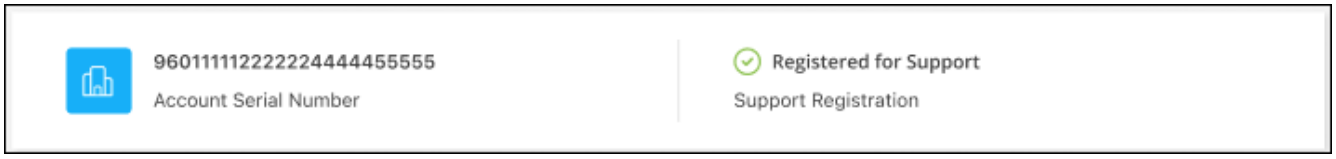
Étapes

1. Dans le coin supérieur droit de la console Workload Factory, sélectionnez **Aide > Support**.

   La sélection de cette option ouvre la console NetApp dans un nouvel onglet de navigateur et charge le tableau de bord de support.

2. Dans le menu de la console NetApp , sélectionnez **Administration**, puis **Informations d’identification**.
3. Sélectionnez **informations d’identification utilisateur**.
4. Sélectionnez **Ajouter des informations d’identification NSS** et suivez l’invite authentification du site de support NetApp (NSS).
5. Pour confirmer que le processus d’enregistrement a réussi, sélectionnez l’icône aide et sélectionnez **support**.

La page **Ressources** doit indiquer que votre compte est enregistré pour le support.



Notez que les autres utilisateurs de la console NetApp ne verront pas ce même statut d'enregistrement de support s'ils n'ont pas associé un compte de site de support NetApp à leur connexion à la console NetApp . Cependant, cela ne signifie pas que votre compte NetApp n'est pas enregistré pour bénéficier du support. Tant qu'un utilisateur du compte a suivi ces étapes, votre compte a été enregistré.

Client existant mais aucun compte NSS

Si vous êtes un client NetApp existant avec des licences et des numéros de série existants mais pas de compte NSS, vous devez créer un compte NSS et l'associer à votre connexion à la console NetApp .

Étapes

1. Créez un compte sur le site de support NetApp en complétant le "Formulaire d'inscription de l'utilisateur du site de support NetApp"
   a. Veillez à sélectionner le niveau d'utilisateur approprié, qui est généralement **client/utilisateur final NetApp**.
   b. Assurez-vous de copier le numéro de série du compte NetApp (960xxxx) utilisé ci-dessus pour le champ du numéro de série. Cela accélérera le traitement du compte.
2. Associez votre nouveau compte NSS à votre connexion à la console NetApp en suivant les étapes ci-dessousClient existant avec un compte NSS .

Découvrez la toute nouvelle gamme NetApp

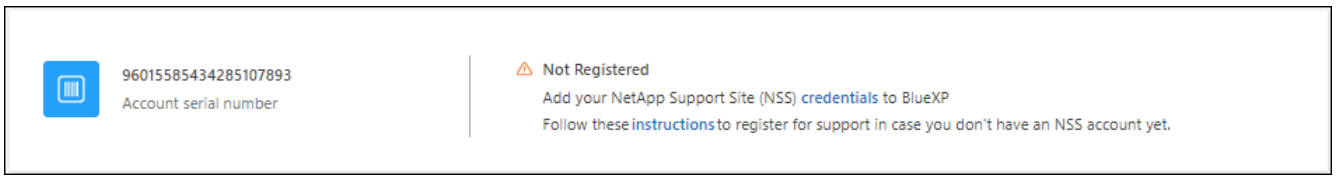
Si vous êtes nouveau chez NetApp et que vous ne disposez pas d'un compte NSS, effectuez chacune des étapes ci-dessous.

Étapes

1. Dans le coin supérieur droit de la console Workload Factory, sélectionnez **Aide > Support**.

   La sélection de cette option ouvre la console NetApp dans un nouvel onglet de navigateur et charge le tableau de bord de support.

2. Recherchez le numéro de série de votre identifiant de compte sur la page des ressources de support.



3. Accédez à "Site d'inscription au support NetApp" et sélectionnez **Je ne suis pas un client NetApp enregistré**.
4. Remplissez les champs obligatoires (ceux avec des astérisques rouges).
5. Dans le champ **Product Line**, sélectionnez **Cloud Manager**, puis votre fournisseur de facturation applicable.

6. Copiez le numéro de série de votre compte à l'étape 2 ci-dessus, vérifiez sa sécurité, puis lisez la Déclaration de confidentialité des données NetApp.

   Un e-mail est immédiatement envoyé à la boîte aux lettres fournie pour finaliser cette transaction sécurisée. Assurez-vous de vérifier vos dossiers de courrier indésirable si l'e-mail de validation n'arrive pas dans quelques minutes.

7. Confirmez l'action à partir de l'e-mail.

   La confirmation de la soumission de votre demande à NetApp et vous recommande de créer un compte sur le site de support NetApp.

8. Créez un compte sur le site de support NetApp en complétant le "Formulaire d'inscription de l'utilisateur du site de support NetApp"
   a. Veillez à sélectionner le niveau d'utilisateur approprié, qui est généralement **client/utilisateur final NetApp**.
   b. Veillez à copier le numéro de série du compte (960xxxx) utilisé ci-dessus pour le champ Numéro de série. Le traitement du compte sera ainsi accéléré.

Une fois que vous avez terminé

NetApp devrait vous contacter au cours de ce processus. Il s'agit d'un exercice d'intégration unique pour les nouveaux utilisateurs.

Une fois que vous avez votre compte de site de support NetApp , associez le compte à votre connexion à la console NetApp en suivant les étapes ci-dessousClient existant avec un compte NSS .

Dépannage GenAI

Apprenez à contourner certains problèmes courants que vous pourriez rencontrer.

Problèmes courants et solutions

Si vous rencontrez l'un de ces problèmes, vous pouvez suivre les étapes de la colonne solution de contournement pour essayer de le résoudre.

De service	Problème	Cause	Solution de contournement
Déploiement	Le déploiement échoue, car le volume existe déjà.	NetApp Workload Factory for GenAI doit créer un nouveau volume pendant le processus de déploiement, mais un volume existe déjà avec le nom que vous avez spécifié.	Spécifiez un nom unique à utiliser pour le nouveau volume et essayez à nouveau de déployer.
Déploiement	Le déploiement échoue car NetApp Workload Factory pour GenAI ne parvient pas à monter le volume.	Un ou plusieurs des ports entrants requis pour FSX pour NetApp ONTAP sont fermés ou filtrés.	Ouvrez les ports entrants suivants :

| Protocole | Port | Objectif

| Tous les protocoles ICMP | Tout | Envoi d'une requête ping à l'instance

| HTTPS | 443 | Accès depuis le connecteur à la LIF de gestion fsxadmin pour envoyer des appels API à FSX

| SSH | 22 | Accès SSH à l'adresse IP du LIF de gestion de cluster ou d'un LIF de gestion de nœud

| TCP | 111 | Appel de procédure à distance pour NFS

| TCP | 139 | Session de service NetBIOS pour CIFS

| TCP | 161-162 | Protocole de gestion de réseau simple

| TCP | 445 | Microsoft SMB/CIFS sur TCP avec encadrement NetBIOS

| TCP | 635 | Montage NFS

| TCP | 749 | Kerberos

| TCP | 2049 | Démon du serveur NFS

| TCP | 3260 | Accès iSCSI via le LIF de données iSCSI

| TCP | 4045 | Démon de verrouillage NFS

| TCP | 4046 | Surveillance de l'état du réseau pour NFS

| TCP | 10000 | Sauvegarde avec NDMP

| TCP | 11104 | Gestion des sessions de communication intercluster pour SnapMirror

| TCP | 11105 | Transfert de données SnapMirror à l'aide de LIF intercluster

| UDP | 111 | Appel de procédure à distance pour NFS

| UDP | 161-162 | Protocole de gestion de réseau simple

| UDP | 635 | Montage NFS

| UDP | 2049 | Démon du serveur NFS

| UDP | 4045 | Démon de verrouillage NFS

| UDP | 4046 | Surveillance de l'état du réseau pour NFS

| UDP | 4049 | Protocole NFS rquotad

Maintenance	Le moteur ai ne démarre pas et vous voyez l'erreur "erreur d'instance de moteur ai" sur la page **bases de connaissances**.	L'instance du moteur ai a été corrompue ou n'existe pas.	Sélectionnez le bouton **Reconstruire**. NetApp Workload Factory pour GenAI reconstruit l'infrastructure et affiche la progression de la reconstruction. Une fois terminé, vos bases de connaissances sont reconnectées à l'infrastructure reconstruite et la liste des bases de connaissances s'affiche.
Maintenance	Le moteur ai ne démarre pas et vous voyez l'erreur "l'instance du moteur GenAI est arrêtée" sur la page **bases de connaissances**.	L'instance du moteur ai n'est pas en cours d'exécution.	Utilisez la console de gestion AWS ou l'interface de ligne de commande AWS pour démarrer l'instance du moteur ai.
Maintenance	Le moteur ai ne démarre pas et vous voyez l'erreur "le serveur du moteur GenAI ne répond pas" sur la page **bases de connaissances**.	L'instance du moteur ai ne répond pas.	Procédez comme suit pour la restauration : **Étapes** 1. Modifiez le groupe de sécurité de l'instance du moteur GenAI pour activer l'accès SSH à l'instance du moteur GenAI. 2. Connectez-vous à l'instance à l'aide de SSH. 3. Exécutez la commande suivante : `docker-compose up`

Maintenance	L’instance Docker backend utilisée par NetApp Workload Factory pour GenAI n’a pas pu démarrer.	Le volume a été supprimé et l’instance EC2 a été redémarrée.	Procédez comme suit pour la restauration : **Étapes** 1. Créez un volume sur FSX pour NetApp ONTAP. Par exemple, le nom du volume peut être `netapp_ai` et le chemin du volume peut être `/netapp_ai`. 2. SSH vers l’instance Amazon EC2. 3. Lister les volumes : ``` docker volume list ``` 4. Supprimez l’ancien volume : ``` docker volume rm ec2-user_persistent_folder ``` 5. Ouvrez le `docker-compose.yml` fichier à l’aide d’un éditeur de texte. 6. Dans la `volumes` section, modifiez le chemin d’accès du périphérique sur le nouveau chemin d’accès du volume. Par exemple :

Maintenance	L’instance Docker backend utilisée par NetApp Workload Factory pour GenAI n’a pas pu démarrer.	Le volume root a été supprimé.	Créez un volume avec un nom et un chemin, puis redémarrez l’instance Docker back-end à partir d’Amazon EC2.
Maintenance	L’instance Docker backend utilisée par NetApp Workload Factory pour GenAI n’a pas pu démarrer.	Le volume root a été supprimé.	Créez un volume avec un nom et un chemin, puis redémarrez l’instance Docker back-end à partir d’Amazon EC2.

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

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

Obtenez de l’aide avec NetApp Workload Factory pour GenAI

NetApp fournit un support pour Workload Factory et ses services cloud de diverses manières. De nombreuses options d’auto-assistance gratuites sont disponibles 24h/24 et 7j/7, telles que des articles de la base de connaissances (KB) et un forum communautaire. Votre inscription au support inclut une assistance technique à distance via un ticket web.

Bénéficiez du support de FSX pour ONTAP

Pour obtenir une assistance technique relative à FSx for ONTAP, à son infrastructure ou à toute solution utilisant le service, reportez-vous à « Obtenir de l’aide » dans la documentation Workload Factory de ce produit.

"Amazon FSX pour ONTAP"

Pour bénéficier du support technique spécifique à Workload Factory et à ses solutions et services de stockage, utilisez les options de support décrites ci-dessous.

Utilisation d’options de support en libre-service

Ces options sont disponibles gratuitement, 24 heures sur 24, 7 jours sur 7 :

- Documentation

  La documentation de Workload Factory que vous consultez actuellement.

- "Base de connaissances"

  Recherchez dans la base de connaissances Workload Factory pour trouver des articles utiles pour résoudre les problèmes.

- "Communautés"

  Rejoignez la communauté Workload Factory pour suivre les discussions en cours ou en créer de nouvelles.

Créez un dossier de demande de support auprès du support NetApp

Outre les options d'auto-support mentionnées ci-dessus, vous pouvez travailler avec un spécialiste du support NetApp pour résoudre tous les problèmes après avoir activé le service de support.

Avant de commencer

Pour utiliser la fonctionnalité **Créer un dossier**, vous devez d'abord vous inscrire au support. Associez vos informations d'identification du site de support NetApp à votre connexion Workload Factory. "Découvrez comment vous inscrire à de l'aide" .

Étapes

1. Dans le coin supérieur droit de la console Workload Factory, sélectionnez **Aide > Support**.

   La sélection de cette option ouvre la console NetApp dans un nouvel onglet de navigateur et charge le tableau de bord de support.

2. Sur la page **Ressources**, choisissez l'une des options disponibles sous support technique :

   a. Sélectionnez **appelez-nous** si vous souhaitez parler avec quelqu'un au téléphone. Vous serez dirigé vers une page netapp.com qui répertorie les numéros de téléphone que vous pouvez appeler.

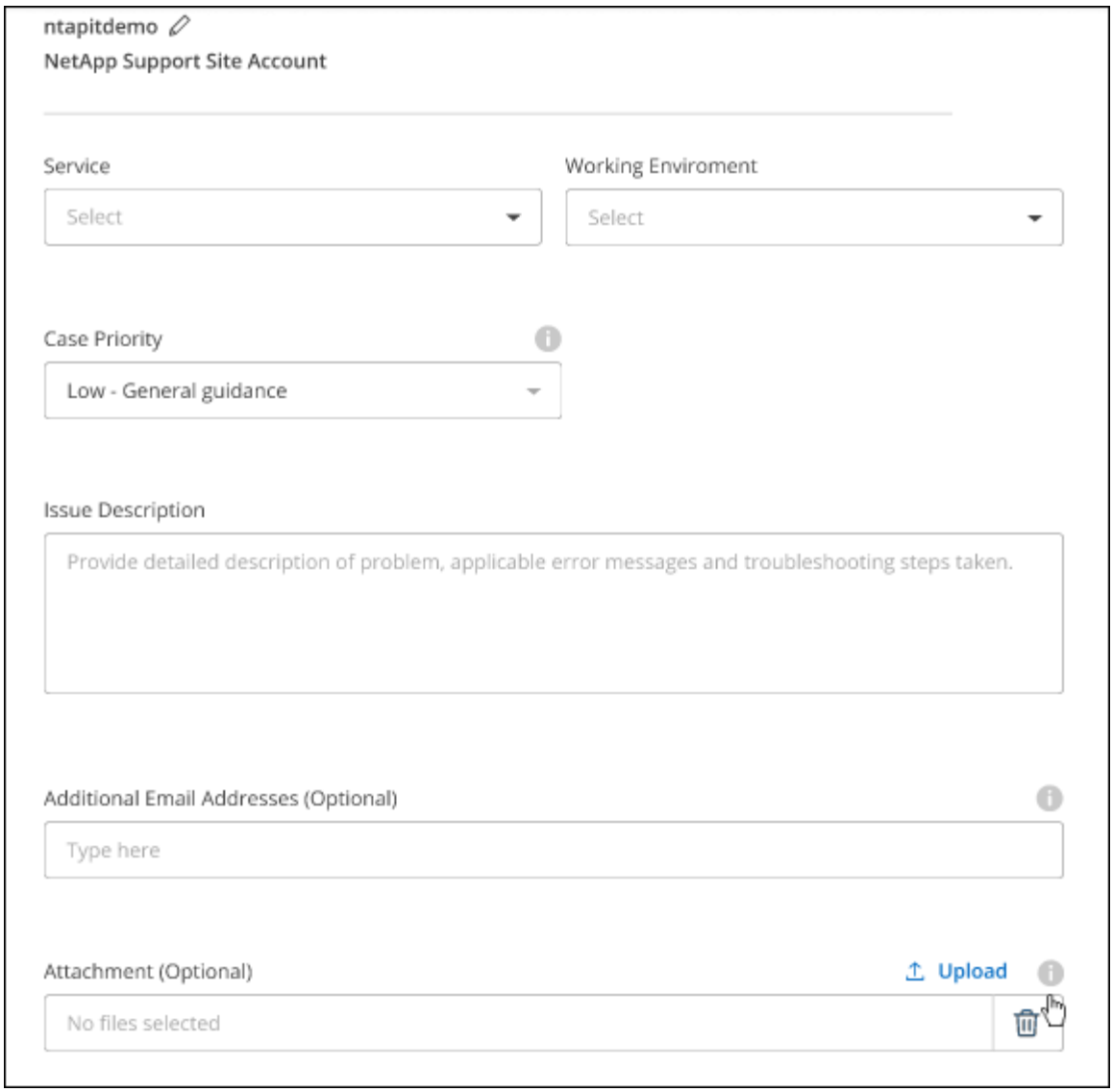
   b. Sélectionnez **Créer un cas** pour ouvrir un ticket avec un spécialiste du support NetApp :

      - **Service** : sélectionnez **Workload Factory**.
      - **Priorité du cas** : choisissez la priorité du cas, qui peut être faible, Moyen, élevé ou critique.

        Pour en savoir plus sur ces priorités, passez votre souris sur l'icône d'information située à côté du nom du champ.

      - **Description du problème** : fournir une description détaillée de votre problème, y compris les messages d'erreur ou les étapes de dépannage applicables que vous avez effectués.
      - **Adresses e-mail supplémentaires**: Entrez des adresses e-mail supplémentaires si vous souhaitez informer quelqu'un d'autre de ce problème.
      - **Pièce jointe (facultatif)** : téléchargez jusqu'à cinq pièces jointes, une à la fois.

        Les pièces jointes sont limitées à 25 Mo par fichier. Les extensions de fichier suivantes sont prises en charge : txt, log, PDF, jpg/JPEG, rtf, doc/docx, xls/xlsx et csv.



Une fois que vous avez terminé

Une fenêtre contextuelle contenant votre numéro de dossier de support s'affiche. Un spécialiste du support NetApp va étudier votre dossier et vous recontacterons très rapidement.

Pour un historique de vos dossiers de support, vous pouvez sélectionner **Paramètres > Chronologie** et rechercher les actions nommées "Créer un dossier de support". Un bouton situé à l'extrême droite vous permet de développer l'action pour afficher les détails.

Il est possible que vous rencontriez le message d'erreur suivant lors de la création d'un dossier :

« Vous n'êtes pas autorisé à créer un dossier pour le service sélectionné »

Cette erreur peut signifier que le compte NSS et la société d'enregistrement à laquelle il est associé ne sont pas la même société d'enregistrement pour le numéro de série du compte NetApp Console (c'est-à-dire. 960xxxx) ou le numéro de série du système. Vous pouvez demander de l'aide en utilisant l'une des options suivantes :

- Utilisez le chat du produit
- Pour soumettre un dossier non technique, rendez-vous sur https://mysupport.netapp.com/site/help

Gestion de vos dossiers de demande de support (aperçu)

Vous pouvez afficher et gérer les cas de support actifs et résolus directement depuis la console NetApp . Vous pouvez gérer les cas associés à votre compte NSS et à votre entreprise.

La gestion des dossiers est disponible en tant qu'aperçu. Nous prévoyons d'affiner cette expérience et d'ajouter des améliorations dans les prochaines versions. Envoyez-nous vos commentaires à l'aide de l'outil de chat In-Product.

Notez ce qui suit :

- Le tableau de bord de gestion des dossiers en haut de la page propose deux vues :
  - La vue de gauche affiche le nombre total de dossiers ouverts au cours des 3 derniers mois par le compte NSS utilisateur que vous avez fourni.
  - La vue de droite affiche le nombre total de dossiers ouverts au cours des 3 derniers mois au niveau de votre entreprise en fonction de votre compte NSS utilisateur.

  Les résultats du tableau reflètent les cas liés à la vue que vous avez sélectionnée.
- Vous pouvez ajouter ou supprimer des colonnes d'intérêt et filtrer le contenu des colonnes telles que priorité et Statut. D'autres colonnes offrent uniquement des fonctions de tri.

  Pour plus d'informations, consultez les étapes ci-dessous.
- Au niveau de chaque dossier, nous offrons la possibilité de mettre à jour les notes de dossier ou de fermer un dossier qui n'est pas déjà à l'état fermé ou en attente fermée.

Étapes

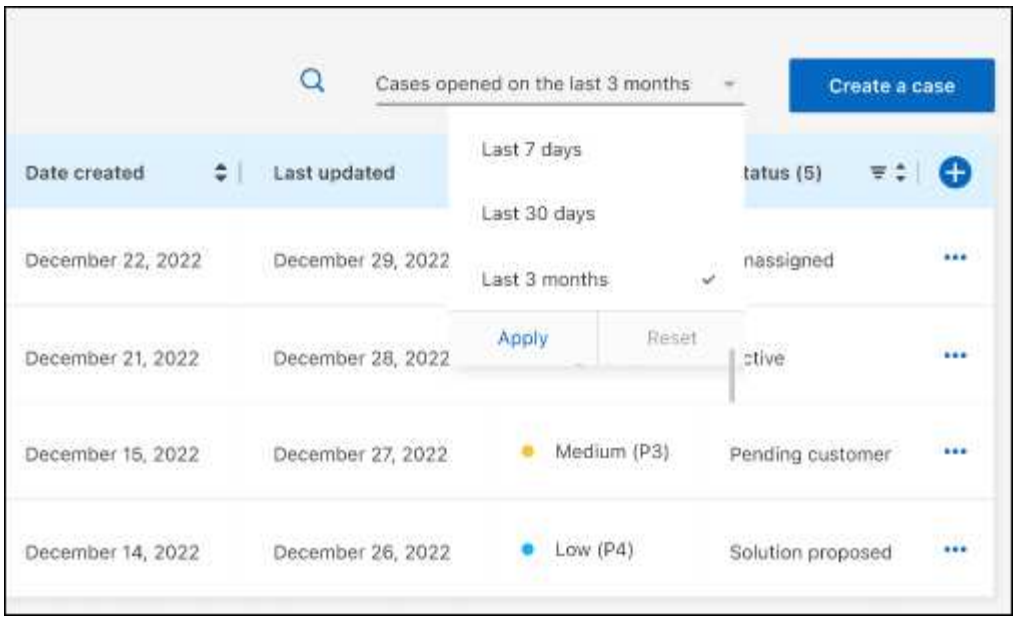
1. Dans le coin supérieur droit de la console Workload Factory, sélectionnez **Aide > Support**.

   La sélection de cette option ouvre la console NetApp , un nouvel onglet de navigateur et charge le tableau de bord de support.

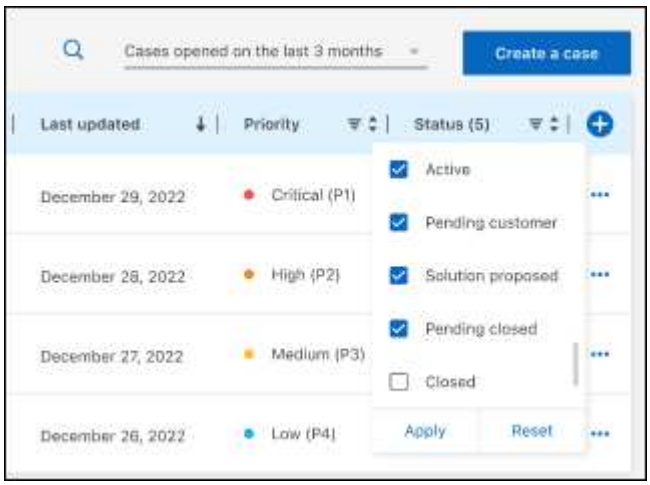
2. Sélectionnez **Gestion des cas** et si vous y êtes invité, ajoutez votre compte NSS à la console NetApp .

   La page **Gestion des cas** affiche les cas ouverts liés au compte NSS associé à votre compte utilisateur NetApp Console. Il s'agit du même compte NSS qui apparaît en haut de la page **Gestion NSS**.

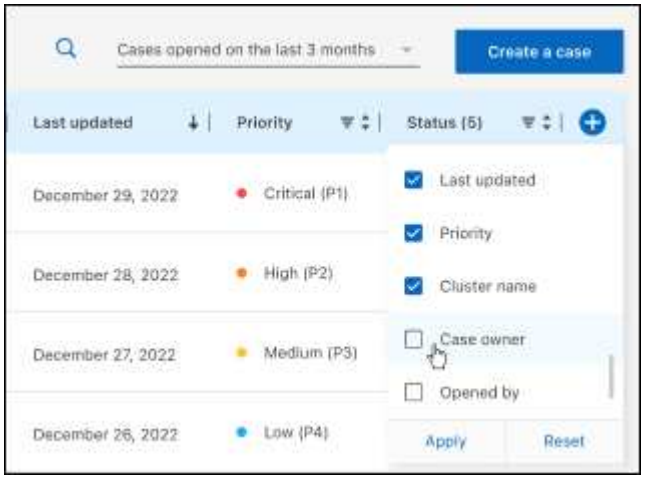
3. Modifiez éventuellement les informations qui s'affichent dans le tableau :
   - Sous **cas de l'organisation**, sélectionnez **Afficher** pour afficher tous les cas associés à votre société.
   - Modifiez la plage de dates en choisissant une plage de dates exacte ou en choisissant une autre période.



◦ Filtrez le contenu des colonnes.



◦ Modifiez les colonnes qui apparaissent dans le tableau en sélectionnant, puis en [Icône plus qui apparaît dans le tableau] choisissant les colonnes que vous souhaitez afficher.



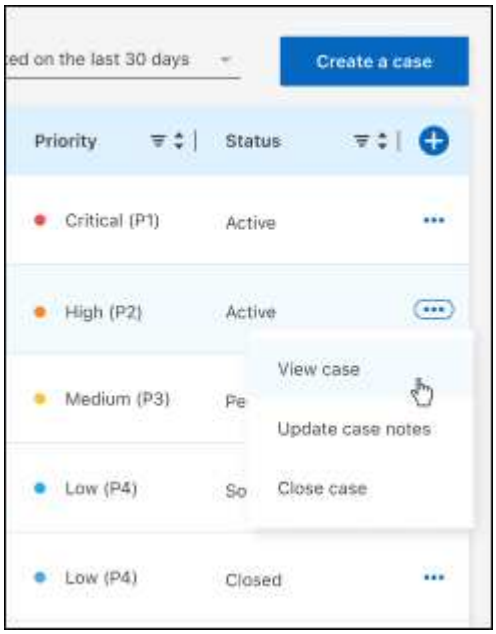
4. Gérer un dossier existant en sélectionnant et en [Icône avec trois points qui apparaît dans la dernière

colonne du tableau] sélectionnant l'une des options disponibles :

- **Voir cas**: Afficher tous les détails sur un cas spécifique.
- **Mettre à jour les notes de cas** : fournir des détails supplémentaires sur votre problème ou sélectionner **Télécharger les fichiers** pour joindre jusqu'à cinq fichiers.

  Les pièces jointes sont limitées à 25 Mo par fichier. Les extensions de fichier suivantes sont prises en charge : txt, log, PDF, jpg/JPEG, rtf, doc/docx, xls/xlsx et csv.

- **Fermer le cas** : fournissez des détails sur la raison pour laquelle vous fermez le cas et sélectionnez **Fermer le cas**.



Mentions légales de NetApp Workload Factory pour GenAI

Les mentions légales donnent accès aux déclarations de copyright, aux marques, aux brevets, etc.

Droits d’auteur

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

Marques déposées

NetApp, le logo NETAPP et les marques mentionnées sur la page des marques commerciales NetApp sont des marques commerciales de NetApp, Inc. Les autres noms de sociétés et de produits peuvent être des marques commerciales de leurs propriétaires respectifs.

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

Brevets

Vous trouverez une liste actuelle des brevets appartenant à NetApp à l’adresse suivante :

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

Politique de confidentialité

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

Source ouverte

Les fichiers de notification fournissent des informations sur les droits d’auteur et les licences de tiers utilisés dans le logiciel NetApp.

"Usine de charges de travail NetApp"

Informations sur le copyright

Informations sur les marques commerciales