NetApp Workload Factory for GenAI ドキュメント

GenAI

NetApp
October 06, 2025

目次

NetApp Workload Factory for GenAI ドキュメント 1
リリースノート 2
NetApp Workload Factory for GenAIの新機能 2
2025年10月5日 2
2025年8月3日 2
2025年6月29日 3
2025年6月3日 3
2025年5月4日 3
2025年3月2日 4
2025年2月2日 5
2025年1月5日 5
2024年12月1日 6
2024年11月3日 6
2024年9月29日 7
2024年9月1日 7
2024 年 8 月 4 日 7
2024年7月7日 8
GenAI向けNetAppワークロードファクトリーの詳細 9
GenAI向けNetAppワークロードファクトリーの詳細 9
NetApp Workload Factory for GenAI とは何ですか? 9
GenAIを使用してジェネレーティブAIアプリケーションを作成するメリット 9
GenAIの仕組み 10
NetApp Workload Factory for GenAIが生成AIアプリケーションの構築にどのように役立つか 11
NetApp Workload Factory を使用するためのツール 12
コスト 12
ライセンス 12
地域 12
NetApp生成AIエンジンのコンポーネント 13
生成AIを使用してAmazon Bedrock用のナレッジベースを構築 20
はじめに 20
生成AIナレッジベースのクイックスタート 20
生成AIナレッジベースの要件 21
ナレッジベースまたはコネクタに追加するデータソースを特定する 23
GenAIインフラの導入 24
生成AIナレッジベースの作成 27
ナレッジベースの作成と設定 27
ナレッジベースへのデータソースの追加 30
生成AIナレッジベースのテスト 34
生成AIナレッジベースの外部認証のアクティブ化 36

生成AIナレッジベースを公開し、一意のエンドポイントを表示する 37
生成AI外部サンプルチャットボットアプリケーションを使用する 38
詳細 38
RAGベースの 生成AIアプリケーションの作成 39
生成AIで次にできること 39
生成AIを使用してAmazon Q Business用のコネクタを作成する 40
はじめに 40
生成AIコネクタのクイックスタート 40
生成AI Connectorの要件 41
コネクタに追加するデータソースを特定する 42
GenAIインフラの導入 43
Amazon Q Business 用の NetApp コネクタを作成する 46
コネクターの定義 47
コネクタへのデータソースの追加 48
管理と監視 50
GenAIインフラの管理 50
インフラに関する情報を表示する 50
インフラストラクチャの削除 50
生成AIナレッジベースの管理 51
ナレッジベースに関する情報を表示する 51
ナレッジベースの編集 51
スナップショットでナレッジベースを保護 52
ナレッジベースへのデータソースの追加 54
データソースとナレッジベースを同期する 58
ナレッジベースを作成する前にチャットモデルを評価する 59
ナレッジベースの非公開 60
ナレッジベースの削除 60
Amazon Q Business Connectorの管理 61
コネクタに関する情報を表示する 61
コネクタの編集 61
コネクタへの追加データソースの追加 62
データソースをコネクタと同期する 66
コネクタの削除 67
生成AIデータソースを管理します。 68
データソースに関する情報を表示する 68
データソース設定の編集 68
既存のデータソースの内容を更新する 69
データソースを削除する 69
NetApp Workload FactoryのTrackerでワークロード操作を監視する 69
運用の追跡と監視 70
View APIヨウキユウ 70

失敗した処理を再試行する . . . 70
失敗した処理を編集して再試行してください . . . 71
知識とサポート . . . 72
NetApp Workload Factory for GenAIのサポートに登録する . . . 72
サポート登録の概要 . . . 72
NetAppサポートのアカウントを登録する . . . 72
生成AIのトラブルシューティング . . . 74
一般的な問題と解決策 . . . 74
NetApp Workload Factory for GenAI のサポートを受ける . . . 78
FSx for ONTAPのサポートを利用する . . . 78
セルフサポートオプションを使用します . . . 78
ネットアップサポートと一緒にケースを作成します . . . 79
サポートケースの管理（プレビュー） . . . 81
NetApp Workload Factory for GenAI の法的通知 . . . 84
著作権 . . . 84
商標 . . . 84
特許 . . . 84
プライバシーポリシー . . . 84
オープンソース . . . 84

NetApp Workload Factory for GenAI ドキュメント

リリースノート

NetApp Workload Factory for GenAIの新機能

Workload Factory の Generative AI ワークロード機能の新機能について説明します。

2025年10月5日

BlueXP workload factoryが**NetApp**ワークロードファクトリーに

BlueXP は、データ インフラストラクチャの管理における役割をより適切に反映するために、名前が変更され、再設計されました。その結果、 BlueXP workload factoryの名前がNetAppワークロード ファクトリーに変更されました。

Amazon Q Business の**NetApp**コネクタに汎用 **NFS/SMB** データソースを追加するためのサポート

Workload Factory API を使用すると、汎用 NFSv3、NFSv4、または SMB 共有から Amazon Q Business のNetAppコネクタにデータ ソースを追加できるようになりました。これにより、Amazon FSx for NetApp ONTAP以外のファイルシステムでホストされているボリュームに保存されているファイルを含めることができます。

"Amazon Q Business 用の NetApp コネクタを作成する"

"コネクタにデータソースを追加する"

ナレッジベースの高度なチャット設定

応答の長さ、温度、推論設定など、ナレッジ ベースのチャット モデルに適用可能な高度なチャット設定を構成できるようになりました。これらの設定の一部 (最新性や変更時間の設定、高度な取得設定、システム プロンプトなど) は、Workload Factory API を使用する場合にのみ利用できます。

"生成AIナレッジベースの作成"

埋め込み、チャット、再ランキングモデルで推論タイプの選択がサポートされるようになりました

選択した埋め込み、チャット、または再ランク付けモデルに推論設定がある場合は、推論タイプを選択できるようになりました。これにより、チャットボットのパフォーマンスとリソース要件をニーズに合わせてより適切に調整できるようになります。

"生成AIナレッジベースの作成"

2025年8月3日

構造化データ結果の安全な保存

チャットボットのクエリ結果に構造化データが含まれている場合、GenAI はその結果を Amazon S3 バケットに保存できます。これらの結果が S3 バケットに保存されると、チャット セッション内のダウンロード リンクを使用してダウンロードできます。

"生成AIナレッジベースの作成"

MCP サーバーの可用性

NetApp は現在、 NetApp Workload Factory for GenAI に Model Context Protocol (MCP) サーバーを提供しています。サーバーをローカルにインストールすると、外部の MCP クライアントが GenAI ナレッジ ベースからクエリ結果を検出して取得できるようになります。

"NetApp Workload Factory GenAI MCP サーバー"

2025年6月29日

汎用 **NFS/SMB** ファイルシステムでホストされるデータソースのサポート

汎用SMBまたはNFS共有からデータソースを追加できるようになりました。これにより、Amazon FSx for NetApp ONTAP以外のファイルシステムでホストされているボリュームに保存されているファイルも含めることができます。

"ナレッジベースにデータソースを追加する"

"コネクタにデータソースを追加する"

2025年6月3日

運用の監視と追跡に使用可能なトラッカー

GenAIでTracker監視機能が利用可能になりました。Trackerを使用すると、保留中、進行中、完了済みの操作の進行状況とステータスを監視および追跡したり、操作タスクとサブタスクの詳細を確認したり、問題や失敗を診断したり、失敗した操作のパラメータを編集したり、失敗した操作を再試行したりできます。

"NetApp Workload FactoryのTrackerでワークロード操作を監視する"

知識ベースの再ランク付けモデルを選択する

ナレッジベースで使用する特定のリランカーモデルを選択することで、再ランク付けされたクエリ結果の関連性を高めることができます。GenAIは、Cohere RerankモデルとAmazon Rerankモデルをサポートしています。

"生成AIナレッジベースの作成"

2025年5月4日

Amazon Q Business 向け **NetApp** コネクタのサポート

GenAI のこのリリースでは、NetApp Connector for Amazon Q Business のサポートが導入され、Amazon Q Business 用のコネクタを作成できるようになりました。Amazon Bedrock用の 生成AIナレッジベースを構築するよりも、初期設定が少なくて済み、Amazon Q Business AIアシスタントをすばやく簡単に活用できます。

"Amazon Q Business 用の NetApp コネクタを作成する"

強化されたチャットモデルサポート

生成AIでは、ナレッジベース用に次の追加チャットモデルがサポートされるようになりました。

- "Mistral AIモデル"
- "Amazon Titanテキストモデル"
- "Meta Llamaモデル"
- "Jamba 1.5モデル"
- "Cohereコマンドモデル"
- "ディープシークモデル"

生成AIは、Amazon Bedrockがサポートする各プロバイダのモデルをサポートしています。"Amazon Bedrockでサポートされる基盤モデル"

"生成AIナレッジベースの作成"

権限に関する用語を更新

Workload Factory のユーザー インターフェースとドキュメントでは、読み取り権限を示すために「読み取り専用」を使用し、自動化権限を示すために「読み取り/書き込み」を使用するようになりました。

2025年3月2日

組み込みチャットボットの機能強化

質問と回答をクリップボードに直接コピーしたり、チャットウィンドウのサイズを調整したり、タイトルを変更したりできるようになりました。さらに、チャット応答にテーブルを含めることができるようになりました。テーブルはコピー可能です。

"生成AIナレッジベースのテスト"

チャット応答引用のサポート

チャットの応答に、応答を生成するために使用されたファイルとデータのチャンクをリストする引用が含まれるようになりました。

"生成AIナレッジベースのテスト"

ファイル形式のサポートの強化

このリリースの 生成AIでは、ファイルサポートが強化されています。

- チャットモデル機能CSVサポートが改善されました。これにより、CSVファイルからデータを照会するときに、より便利な応答が可能になります。
- 生成AIは、データソースからApache Parquetファイルを取り込むことができるようになりました。
- 生成AIでは、画像を含むMicrosoft Word DOCXファイルの取り込みがサポートされるようになりました。DOCXドキュメント内に埋め込まれた画像がスキャンされ、埋め込まれた画像からのテキストインサイトがナレッジベースのクエリへの応答に含まれます。

"サポートされるデータソースファイル形式"

2025年2月2日

Amazon Nova基盤モデルのサポート

生成AIはAmazon Nova基盤モデルをサポートするようになりました。Amazon Nova Micro、Amazon Nova Lite、およびAmazon Nova Proがサポートされています。

"GenAIの要件"

データソースのファイルタイプフィルタリング

生成AIでは、データソースを追加するときに、データソーススキャンに含める特定のファイルタイプの選択がサポートされるようになりました。

"ナレッジベースへのデータソースの追加"

データソースのファイル変更日のフィルタリング

生成AIでは、データソースの追加時に変更日によってデータソーススキャンに含めるファイルのフィルタリングがサポートされるようになりました。インクルードされるファイルの変更日の範囲を選択できます。

"ナレッジベースへのデータソースの追加"

画像ファイルのサポートと**PDF**ファイルのサポートの強化

生成AIでは、画像やグラフの説明、ドキュメントテキストからのインサイトを使用してナレッジベースのクエリへの応答を強化できるようになり、より豊かで高品質な回答が得られるようになりました。生成AIでは、PDFファイル内の画像ファイルおよび画像をスキャンできるようになりました（マルチモーダルファイルサポートとも呼ばれます）。画像またはPDFファイルのスキャンを選択した場合は、画像のテキスト（PDFドキュメントに埋め込まれた画像を含む）がデータソースにスキャンされ、スキャンからのインサイトがナレッジベースのクエリへの応答に含まれます。

"ナレッジベースへのデータソースの追加"

ハイブリッド検索と再検索のサポート

生成AIでは、ハイブリッド検索を使用して結果をリランク付けすることで、検索結果の関連性と精度を大幅に向上させることができます。ハイブリッド検索は、従来のキーワードベース検索の強みと高度な高密度ベクトルベースのセマンティック検索技術を組み合わせたものです。標準的なキーワード検索結果は、近い一致と言語的なニュアンスで強化され、関連性が向上します。その後、生成AIはCohere RerankやAmazon Rerankなどの高度なリランキングモデルを使用してこれらの結果をさらに絞り込み、最も関連性の高い結果を返します。この機能は、新しく作成されたナレッジベースで使用できます。

"GenAI向けNetAppワークロードファクトリーの詳細"

2025年1月5日

カスタム**Snapshot**名

アドホックスナップショットのスナップショット名を指定できるようになりました。

"スナップショットでナレッジベースを保護"

カスタム**AI**エンジンインスタンス名

導入時にAIエンジンインスタンスにカスタム名を付けることができるようになりました。

"GenAIインフラの導入"

破損または欠落している 生成**AI**インフラの再構築

AI エンジン インスタンスが破損したり、何らかの理由で削除されたりした場合は、Workload Factory で再構築することができます。ワークロード ファクトリーは、再構築が完了するとナレッジ ベースをインフラストラクチャに自動的に再接続し、すぐに使用できるようにします。

"トラブルシューティング"

2024年12月1日

Snapshotからナレッジベースをクローニング

NetApp Workload Factory for GenAI では、スナップショットからのナレッジ ベースのクローン作成がサポートされるようになりました。これにより、ナレッジ ベースの迅速な回復と既存のデータ ソースを使用した新しいナレッジ ベースの作成が可能になり、データの回復と開発に役立ちます。

"ナレッジベースの複製"

オンプレミスの**ONTAP**クラスタの検出とレプリケーション

オンプレミスのONTAPクラスターデータを検出して FSx for ONTAPファイル システムに複製し、AI ナレッジベースの強化に使用できるようにします。すべてのオンプレミス検出およびレプリケーション ワークフローは、ストレージ インベントリの新しい **On-Premises ONTAP** メニューから実行できます。

"オンプレミスの ONTAP クラスタを検出"

2024年11月3日

個人識別情報をデータガードレールでマスクする

Generative AI ワークロードでは、 NetAppコンソール分類を活用したデータ ガードレール機能が導入されています。データ ガードレール機能は、個人を特定できる情報 (PII) を識別してマスクし、コンプライアンスを維持し、組織の機密データのセキュリティを強化するのに役立ちます。

"生成AIナレッジベースの作成"

"NetAppコンソールの分類について学ぶ"

2024年9月29日

ナレッジベースボリュームの**Snapshot**とリストアのサポート

ナレッジベースのポイントインタイムコピーを作成することで、ジェネレーティブAIワークロードのデータを保護できるようになりました。これにより、偶発的な損失からデータを保護したり、ナレッジベースの設定の変更をテストしたりできます。以前のバージョンのナレッジベースボリュームはいつでもリストアできます。

"ナレッジベースボリュームのスナップショットの作成"

"ナレッジベースボリュームのスナップショットのリストア"

スケジュール済みスキャンを一時停止

スケジュールされたデータソーススキャンを一時停止できるようになりました。デフォルトでは、ジェネレーティブAIワークロードは各データソースを毎日スキャンし、各ナレッジベースに新しいデータを取り込みます。最新の変更を取り込みたくない場合（テスト中やスナップショットのリストア中など）は、スケジュールされたスキャンを一時停止していつでも再開できます。

"ナレッジベースの管理"

ナレッジベースでのデータ保護ボリュームのサポート

ナレッジベースボリュームを選択する際に、NetApp SnapMirrorレプリケーション関係の一部であるデータ保護ボリュームを選択できるようになりました。これにより、SnapMirrorレプリケーションですでに保護されているボリュームにナレッジベースを格納できます。

"ナレッジベースに統合するデータソースを特定する"

2024年9月1日

その他のチャンキング戦略

ジェネレーティブAIワークロードで、データソースに対してマルチセンテンスチャンキングとオーバーラップベースのチャンキングがサポートされるようになりました。

ナレッジベースごとの専用ボリューム

ジェネレーティブAIワークロードでは、新しいナレッジベースごとに専用のAmazon FSx for NetApp ONTAPボリュームが作成されるようになりました。これにより、ナレッジベースごとに個別のSnapshotポリシーが有効になり、障害やデータポイズニングに対する保護が強化されます。

2024 年 8 月 4 日

Amazon CloudWatch Logsの統合

ジェネレーティブAIワークロードがAmazon CloudWatch Logsと統合され、ジェネレーティブAIワークロードのログファイルを監視できるようになりました。

チャットボットアプリケーションの例

NetApp Workload Factory GenAI サンプル アプリケーションを使用すると、Web ベースのチャットボット アプリケーションで直接対話して、公開されたNetApp Workload Factory ナレッジ ベースからの認証と取得をテストできます。

2024年7月7日

GenAI向けワークロードファクトリーの初期リリース

最初のリリースには、組織のデータを埋め込むことによってカスタマイズされたナレッジベースを開発する機能が含まれています。ナレッジベースには、ユーザー用のチャットボットアプリケーションからアクセスできます。この機能により、組織固有の質問に対する正確で適切な回答が保証され、すべてのユーザーの満足度と生産性が向上します。

GenAI向けNetAppワークロードファクトリーの詳細

GenAI向けNetAppワークロードファクトリーの詳細

NetApp Workload Factory for GenAI を使用すると、Amazon FSx for NetApp ONTAPファイルシステムを GenAI 基盤モデルと統合できます。これにより、AI データセット向けの豊富な保護、セキュリティ、コスト最適化機能を備えた高性能ストレージが提供されます。

NetApp Workload Factory for GenAI とは何ですか?

NetApp Workload Factory for GenAI を使用すると、Amazon FSx for NetApp ONTAP上のエンタープライズデータソースを Generative AI アプリケーションで使用できるようになります。検索拡張生成 (RAG) を利用すると、Amazon Bedrock または Amazon Q Business 経由で利用可能な基盤モデルにデータソースをすばやく接続し、仮想アシスタント、Q&A チャットボット、ドキュメント要約、コンテンツ作成などの生成 AI を活用したアプリケーションを開発できます。

ジェネレーティブAIを組織のデータとともに使用すると、モデルがトレーニングされたパブリックデータに基づくモデルのインテリジェンスだけに頼るのではなく、独自の知識と専門知識を活用できます。RAGを使用してモデルをカスタマイズすると、組織固有の質問に対する正確で適切な応答が保証され、ジェネレーティブAIを使用してアプリケーションのユーザーの生産性と効率が向上します。

組織のデータに合わせてカスタマイズされたGenAIアプリケーションを開発することで、独自の知識と専門知識を活用できます。このカスタマイズ機能により、組織固有の質問に対する正確で適切な回答が保証され、すべてのユーザーの満足度と生産性が向上します。

"ナレッジベースの作成"生成AIでは、データソースからデータを取り込み、ベクトル化された結果をデータベースに格納し、取り込まれたデータを使用してクエリに応答する方法を完全に制御できます。この方法では、より初期設定が必要ですが、結果に応じて異なるチャットモデルを選択できます。"Amazon Q Business用のNetAppコネクタを定義する"データソースからのデータはAmazon Q Businessによって取り込まれ、インデックスに保存されます。この方法では初期設定の手間は少なくて済みますが、結果を制御することはできません。

ワークロードファクトリーの詳細については、 "ワークロードファクトリーの概要" 。

GenAIを使用してジェネレーティブAIアプリケーションを作成するメリット

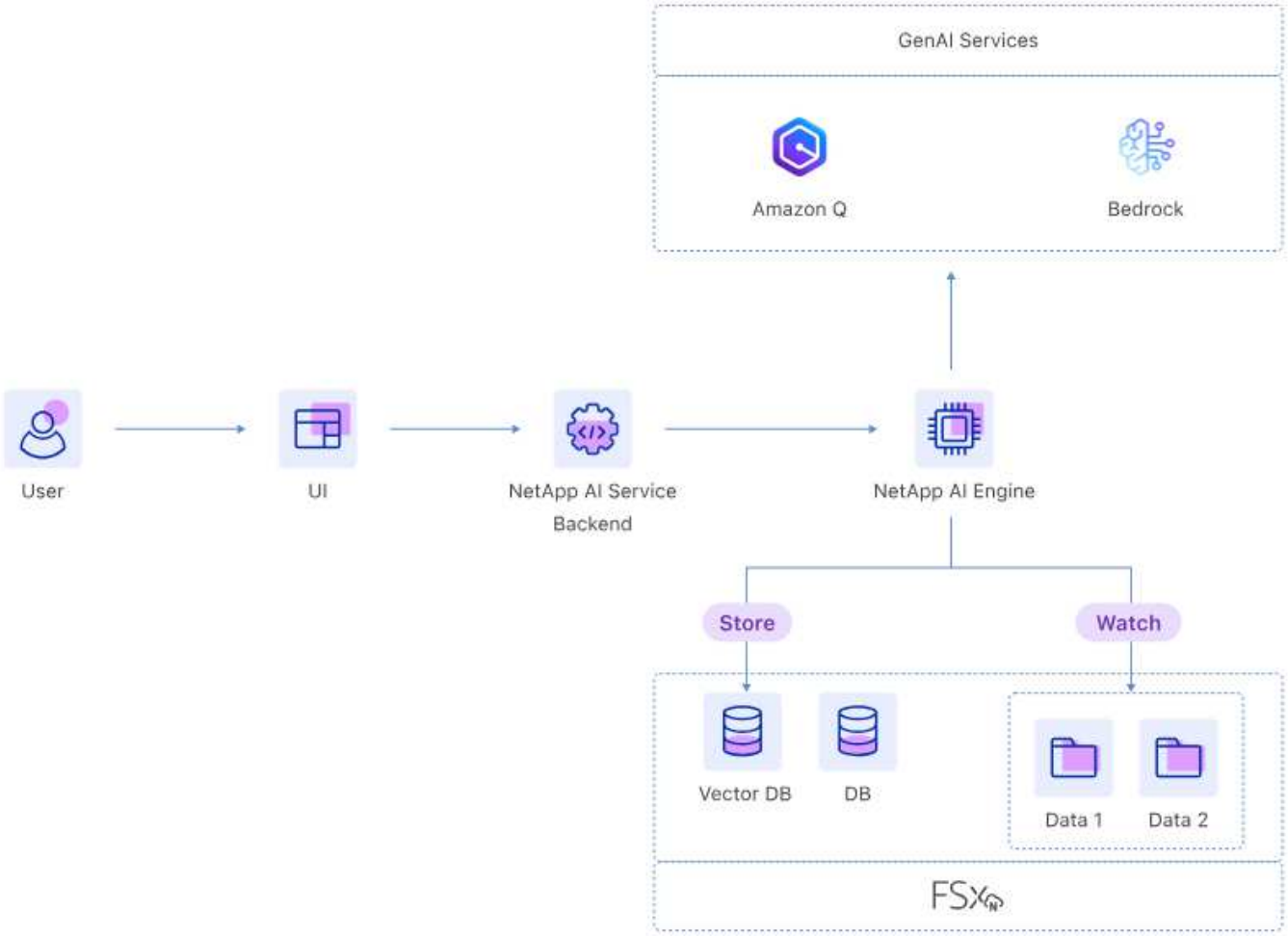
NetApp Workload Factory for GenAI は、検索拡張生成 (RAG) を使用して生成 AI アプリケーションを構築するために必要なインフラストラクチャを導入するプロセスを簡素化します。具体的には、GenAI には次のような利点があります。

- IT管理者や開発者は、データインフラ、基盤、言語モデルに関する深い知識がなくても、GenAIが提供する自動化を活用してアプリケーション開発を加速できます。データ管理者と開発者は、生成AIアプリケーションで使用する組織の非構造化データを埋め込むエンタープライズナレッジベースを簡単かつ迅速に作成できます。
- ナレッジベースに埋め込まれたファイルにユーザー権限を保持して、データのセキュリティとプライバシーを維持することで、セキュリティを強化します。チャットボットなどのアプリケーションは、ユーザーがアクセスできるデータに基づいて、認証されたユーザーのみに回答を提供するように開発できます。

- 組織のデータが外部に公開されることがないAWSカスタマーアカウント内で、エンタープライズデータのプライバシーとセキュリティを確保します。
- LangChainなどのオープンソースフレームワークを使用したQ&Aチャットボットなどの 生成AIアプリケーションの開発を加速 生成AI APIを活用して、ナレッジベースやコネクタのプロビジョニングと管理、ナレッジベースとのチャット、チャット履歴の保存と取得を行います。
- FSx for NetApp ONTAPファイルシステムに生成型AIデータインフラを導入し、高可用性、ローカルデータの保護とリカバリのためのスナップショット、ディザスタリカバリのためのSnapMirror、データインフラのバックアップのためのSnapVaultなどのONTAP機能を活用することで、データの保護と可用性の体制を強化します。
- データの重複排除、圧縮とコンパクション、データ階層化、シンプロビジョニングなどのONTAPのデータ効率化機能を活用することで、生成型AIデータインフラの全体的なストレージコストを削減します。
- 生成AIが提供するハイブリッド検索機能とリランク機能を使用して、データから高品質の結果を得ることができます。ハイブリッド検索とリランク付けを組み合わせることで、検索結果の関連性が大幅に向上します。これらの機能はAmazon AWSから利用でき、リージョンによって異なります。

GenAIの仕組み

生成AIは、組織のプライベートデータを使用してモデルのインテリジェンスを補完し、組織内のユーザーからの質問に対するカスタマイズされた回答を提供します。まずRAGフレームワークに必要なインフラストラクチャをデプロイしてから、Amazon BedrockまたはAmazon Q Businessで利用可能な組織のデータソースと基盤モデルを使用してナレッジベースを構築するかコネクタを定義し、次にアプリケーション（Q&Aチャットボットなど）をナレッジベースまたはコネクタに接続します。



NetApp Workload Factory for GenAIが生成AIアプリケーションの構築にどのように役立つか

GenAIは、次の方法でRAGを使用したジェネレーティブAIアプリケーションの構築を支援します。

- FSx for ONTAPファイルシステムやAmazon BedrockまたはAmazon Q Business上のデータソースと連携するために、検索拡張生成（RAG）フレームワークに必要なインフラを導入します。このインフラには、データ管理用のNetApp生成AIエンジンインスタンス、組み込みベクターデータベース（LanceDB）、FSx for ONTAPファイルシステム上のベクターデータベース用ストレージが含まれています。
- データソースをAmazon BedrockまたはAmazon Q Businessで利用可能な埋め込みおよび言語モデルに接続して、データソースを埋め込み、ユーザークエリの応答を取得できます。データソースは、モデルとその構成とともに、FSx for ONTAPナレッジベースとして提供されます。
- ソースデータをナレッジベースまたはコネクタに取り込んで、ソースファイルをSMB共有に埋め込み、FSx for ONTAPファイルシステムでNFSエクスポートを行うとともに、SMB共有内のファイルに対するファイル権限を格納します。
- ナレッジベースのコンテンツに基づいて、会話スターターの質問を自動的に作成します。
- データ管理者がナレッジベースでチャットをテストするためのチャットシミュレータを提供します。
- シンプルなコネクタインターフェイスを提供するため、このAIアシスタントの機能を迅速かつ簡単に利用して、生成AIとAmazon Q Businessを接続できます。

NetApp Workload Factory を使用するためのツール

NetApp Workload Factory は次のツールで使用できます。

- **Workload Factory** コンソール: Workload Factory コンソールは、アプリケーションとプロジェクトの視覚的かつ全体的なビューを提供します。
- * NetAppコンソール*: NetAppコンソールはハイブリッド インターフェイス エクスペリエンスを提供するため、Workload Factory を他のNetAppデータ サービスと一緒に使用できます。
- 質問する: Ask me AI アシスタントを使用すると、Workload Factory コンソールを離れることなく、質問したり、Workload Factory について詳しく知ることができます。Workload Factory のヘルプ メニューから「Ask me」にアクセスします。
- **CloudShell CLI**: Workload Factory には、単一のブラウザベースの CLI からアカウント全体の AWS およびNetApp環境を管理および操作するための CloudShell CLI が含まれています。Workload Factory コンソールの上部バーから CloudShell にアクセスします。
- **REST API**: Workload Factory REST API を使用して、FSx for ONTAPファイルシステムやその他の AWS リソースをデプロイおよび管理します。
- **CloudFormation**: AWS CloudFormation コードを使用して、Workload Factory コンソールで定義したアクションを実行し、AWS アカウントの CloudFormation スタックから AWS およびサードパーティのリソースをモデル化、プロビジョニング、管理します。
- **Terraform NetApp Workload Factory** プロバイダー: Terraform を使用して、Workload Factory コンソールで生成されたインフラストラクチャ ワークフローを構築および管理します。

コスト

Workload Factory の GenAI 機能の使用には料金はかかりません。

ただし、生成型AIインフラをサポートするには、導入したAWSリソースに料金を支払う必要があります。たとえば、Amazon BedrockまたはAmazon Q Business、FSx for ONTAPファイルシステムとストレージ容量、生成AIエンジンEC2インスタンスの料金をAWSに支払います。

テキスト情報のための画像のスキャンなど、一部のマルチモーダル操作では、より多くのリソースが使用されるため、コストが高くなります。ナレッジベースの設定の変更など、一部の設定処理ではデータソースが再スキャンされたり、データソーススキャンのコストが高くなることがあります。

ライセンス

Workload Factory の AI 機能を使用するために、 NetAppからの特別なライセンスは必要ありません。

地域

Workload Factory は、FSx for ONTAPがサポートされているすべての商用リージョンでサポートされています。"サポートされている Amazon リージョンを表示します。"

次の AWS リージョンはサポートされていません。

- 中国地域
- GovCloud（米国）リージョン
- シークレットクラウド

- トップシークレットクラウド

NetApp生成AIエンジンのコンポーネント

GenAI インフラストラクチャをデプロイすると、Workload Factory によって GenAI エンジン用の EC2 インスタンスが作成されます。また、このインスタンスの IAM ロール、セキュリティ グループ、プライベート エンドポイントも作成されます。 Workload Factory が AWS 環境に作成するこれらのコンポーネントについて、さらに詳しく理解したい場合があります。

EC2インスタンスタイプ

m5.large

IAMロール

GenAIエンジンインスタンスには、データのチャンクをAmazon Bedrockの埋め込みモデルに送信し、NetApp AIサービスバックエンドと通信するための権限が必要です。IAMロールには次の権限が含まれています。

IAMロールノケンゲン

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

セキュリティグループ

発信ルールはすべてのトラフィックに対してオープンであり、着信ルールは完全にクローズされます。

プライベートエンドポイント

ターゲット VPC にまだプライベートエンドポイントがない場合、Workload Factory は GenAI エンジン EC2 インスタンスのプライベートエンドポイントを作成し、次の AWS サービスと通信できるようにします。

- Amazon Bedrock
  - 岩盤
  - Bedrock -ランタイム
  - Bedrock-agentランタイム
- Amazon Elastic Container Registry（ECR）
  - API
  - Docker です

- AWSシステムマネージャ（SSM）
  - SSM
  - ec2messages
  - ssmメツセエシ
- NetApp ONTAP 対応の Amazon FSX
- Amazon CloudWatch

生成AIを使用してAmazon Bedrock用のナレッジベースを構築

はじめに

生成AIナレッジベースのクイックスタート

Amazon FSx for NetApp ONTAPファイルシステムに存在する組織のデータを使用して、ナレッジベースまたはAmazon Q Business Connectorの作成を開始します。チャットボットなどのアプリケーションは、このナレッジベースまたはコネクタにアクセスして、組織に焦点を当てた応答をエンドユーザーに提供します。

1 ワークロードファクトリーにログイン

必要となるのは "Workload Factoryでアカウントを設定する"次のいずれかを使用してログインします "コンソールエクスペリエンス"。

2 生成AIの要件を満たす環境のセットアップ

AWSインフラの導入、導入および検出されたFSx for ONTAPファイルシステム、ナレッジベースまたはコネクタに統合するデータソースのリスト、Amazon Bedrock AIサービスまたはAmazon Q Businessアプリケーションへのアクセスなどには、AWSのクレデンシャルが必要です。

"生成AI要件の詳細"です。

3 データソースが格納されている**FSx for ONTAP**ファイルシステムを特定する

ナレッジベースに統合するデータソースは、単一のFSx for ONTAPファイルシステムに配置することも、複数のFSx for ONTAPファイルシステムに配置することもできます。これらのシステムが異なるVPCにある場合は、同じネットワーク内でアクセス可能であるか、またはAIエンジンと同じリージョンとAWSアカウントを使用してVPC間でピア関係を確立しておく必要があります。

"データソースを特定する方法"です。

4 生成AIインフラの導入

インフラ導入ウィザードを起動して、AWS環境に 生成AIインフラを導入します。このプロセスでは、NetApp生成AIエンジンのEC2インスタンスと、FSx for ONTAPファイルシステムにNetApp AIエンジンのデータベースを格納するボリュームを導入します。ボリュームは、ナレッジベースで使用されるベクターデータベースを格納するために使用されます。

"ナレッジベースインフラストラクチャの導入方法"です。

次のステップ

ナレッジベースを構築して、組織に焦点を当てた回答をエンドユーザーに提供できるようになりました。

生成AIナレッジベースの要件

ナレッジベースを構築する前に、Workload Factory と AWS が適切にセットアップされていることを確認してください。これには、AWS ログイン認証情報、ナレッジベースに統合するデータソースを含むデプロイ済みの FSx for ONTAPファイルシステム、Amazon Bedrock AI サービスへのアクセスなどが含まれます。

生成AIの基本要件

生成AIには、作業を開始する前に環境が満たす必要のある一般的な要件があります。

Workload Factoryのログインとアカウント

必要となるのは "Workload Factoryでアカウントを設定する"次のいずれかを使用してログインします "コンソールエクスペリエンス"。

AWS のクレデンシャルと権限

Workload Factory に読み取り/書き込み権限を持つ AWS 認証情報を追加する必要があります。つまり、Workload Factory を GenAI の 読み取り/書き込み モードで使用することになります。

現時点では、_basic_modeおよび_read-only_mode権限はサポートされていません。

クレデンシャルを設定する際に、以下に示すように権限を選択すると、FSx for ONTAPファイルシステムの管理、GenAI EC2インスタンスおよびナレッジベースとチャットボットに必要なその他のAWSリソースの導入と管理を行うためのフルアクセスが提供されます。

"Workload FactoryにAWS認証情報を追加する方法を学ぶ"

生成AIナレッジベースの要件

ナレッジベースを使用する場合は、環境が次の要件を満たしていることを確認してください。

Amazon Bedrock

Amazon Bedrockを使用すると、基盤モデルを使用でき、生成型AIアプリケーションを構築するための機能を提供します。

NetApp Workload Factory for GenAI を使い始める前に、Amazon Bedrock をセットアップする必要があります。 GenAI のデプロイメントは、Amazon Bedrock が有効になっている AWS リージョンに存在している必要があります。

- "AWSのドキュメント：Amazon Bedrockのセットアップ"
- "AWSのドキュメント：Amazon Bedrockのナレッジベースでサポートされるリージョンとモデル"

生成AIは、検索結果の関連性を向上させるために、デフォルトで検索結果を再ランク付けします。最良の結果を得るには、Amazon Bedrock Foundationモデルの構成に、Cohere RerankやAmazon Rerankなどのリランクモデルへのアクセスが含まれていることを確認してください（地域で利用可能な場合）。

埋め込みモデル

ナレッジベースを作成する前に、使用する予定の埋め込みモデルを有効にする必要があります。次の埋め込みモデルがサポートされています。

- Titan埋め込みG1 -テキスト
- Titan埋め込みテキストv2
- Titan Multimodal Embedding G1
- 英語を埋め込む
- 多言語を埋め込む

  "Amazon Titanの詳細"

チャットモデル

ナレッジベースを作成する前に、使用する予定の基本チャットモデルを有効にする必要があります。サポートされるモデルはAWSのリージョンによって異なるため、ナレッジベースを導入するリージョンで使用できるモデルを確認するには、を参照し "AWSのドキュメント" てください。

生成AIは、Anthropic、Amazon、Mistral AI、Meta、Jamba、Cohereのさまざまなモデルをサポートしています。

Amazon Bedrockでのこれらのモデルの使用について詳しくは、以下をご覧ください。

- "Amazon BedrockのAnthropic's Claude"
- "Amazon Bedrockコンソールで Amazon Novaを使い始める"
- "Mistral AIモデル"
- "Amazon Titanテキストモデル"
- "Meta Llamaモデル"
- "Jambaモデル"
- "Cohereコマンドモデル"

FSx for ONTAPファイルシステム

少なくとも1つのFSx for ONTAPファイルシステムが必要です。

- 1つのファイルシステムが、ナレッジベースで使用されるベクターデータベースを格納するために、NetAppの 生成AIエンジンによって使用されます（存在しない場合は作成されます）。

  このFSx for ONTAPファイルシステムでは、FlexVolボリュームを使用する必要があります。FlexGroupボリュームはサポートされません。

- 1つ以上のファイルシステムには、ナレッジベースに統合するデータソースが含まれています。

  1つのFSx for ONTAPファイルシステムを両方の目的に使用することも、複数のFSx for ONTAPファイルシステムを使用することもできます。

- AWS FSx for ONTAPファイルシステムが配置されているAWSリージョン、VPC、サブネットを把握しておく必要があります。ファイルシステムがAmazon Bedrockが有効になっているAWSリージョンにある必要があります。
- この導入に含まれるAWSリソースに適用するタグのキーと値のペアを検討する必要があります（オプション）。

- NetApp AIエンジンインスタンスに安全に接続するためのキーペア情報を知っておく必要があります。

"FSx for ONTAPファイルシステムの導入と管理の方法をご確認ください"

ナレッジベースまたはコネクタに追加するデータソースを特定する

ナレッジベースに統合するFSx for ONTAPファイルシステムにあるドキュメント（データソース）を特定または作成します。これらのデータソースを使用すると、組織に関連するデータに基づいて、ナレッジベースでユーザクエリに対する正確でパーソナライズされた回答を提供できます。

データソースの最大数

サポートされるデータソースの最大数は10です。

データソースの場所

データソースは、単一のボリュームに格納することも、ボリューム内のフォルダに格納することも、Amazon FSx for NetApp ONTAPファイルシステム上のSMB共有やNFSエクスポートに格納することもできます。また、NetApp SnapMirrorデータ保護関係にあるAmazon FSx for NetApp ONTAPボリュームにデータソースを保存することもできます。

ボリュームまたはフォルダ内の個 々 のドキュメントを選択することはできません。したがって、データソースを含む各ボリュームまたはフォルダに、ナレッジベースと統合すべきではない無関係なドキュメントが含まれていないことを確認する必要があります。

各ナレッジベースには複数のデータソースを追加できますが、それらはすべてAWSアカウントからアクセス可能なFSx for ONTAPファイルシステム上にある必要があります。

各データソースの最大ファイルサイズは50MBです。

サポートされるプロトコル

ナレッジベースでは、NFSプロトコルまたはSMB / CIFSプロトコルを使用するボリュームのデータがサポートされます。SMBプロトコルを使用して保存されているファイルを選択する場合は、Active Directory情報を入力して、ナレッジベースがこれらのボリューム上のファイルにアクセスできるようにする必要があります。これには、Active Directoryドメイン、IPアドレス、ユーザ名、パスワードが含まれます。

SMB経由でアクセスされる共有（ファイルまたはディレクトリ）にデータソースを格納する場合、その共有にアクセスする権限を持つチャットボットのユーザまたはグループのみがデータにアクセスできます。この「権限認識機能」が有効になっている場合、AIシステムはAuth0内のユーザのEメールを、SMB共有上のファイルの表示または使用を許可されているユーザと比較します。チャットボットは、埋め込まれたファイルのユーザー権限に基づいて回答を提供します。

たとえば、10個のファイル(データソース)をナレッジベースに統合し、そのうちの2つが制限された情報を含む人事ファイルである場合、これら2つのファイルへのアクセスを認証されたチャットボットユーザーのみが、それらのファイルからのデータを含むチャットボットから応答を受け取ります。

サポートされるデータソースファイル形式

現在、Workload Factory GenAI ナレッジ ベースでは次のデータ ソース ファイル形式がサポートされています。

ファイル形式	エクステンション
Apache Parquet [1]	寄木細工
カンマ区切り値file [disclaimer]	.csv
グラフィック交換フォーマット	.gif
JPEG	.jpg or.jpeg
JSONおよびJSONP[1]	.json
マークダウン	.md
Microsoft Word	.docまたは.docx
プレーンテキスト	.txt
ポータブルドキュメントフォーマット	.pdf
ポータブルネットワークグラフィックス	.png
WebP画像	.webp

GenAIインフラの導入

組織のFSx for ONTAPナレッジベース、コネクタ、アプリケーションを構築する前に、生成AI Infrastructure for RAGフレームワークを環境に導入する必要があります。主要なインフラコンポーネントは、Amazon Bedrockサービス、NetApp生成AIエンジンの仮想マシンインスタンス、FSx for ONTAPファイルシステムです。

展開されたインフラストラクチャは、複数のナレッジベース、チャットボット、コネクタをサポートできるため、通常はこのタスクを一度だけ実行する必要があります。

インフラの詳細

GenAI環境は、Amazon Bedrockが有効になっているAWSリージョンに配置する必要があります。 "サポートされているリージョンの一覧を表示する"

インフラストラクチャは、次のコンポーネントで構成されています。

Amazon Bedrockサービス

Amazon Bedrockは、業界をリードするAI企業の基盤モデル（FMS）を単一のAPIで使用できるフルマネージドサービスです。また、セキュアなジェネレーティブAIアプリケーションの構築に必要な機能も提供します。

"Amazon Bedrockの詳細"

Amazon Q Business

Amazon QはAmazon Bedrock上に構築されており、フルマネージドのジェネレーティブAIアシスタントを使用して、データソースからの情報に基づいて質問に答えたりコンテンツを生成したりできます。

"Amazon Q Businessの詳細"

NetApp GenAIエンジン用の仮想マシン

このプロセスでは、NetApp GenAIエンジンがデプロイされます。データソースからデータを取り込み、そのデータをベクターデータベースに書き込む処理能力を提供します。

FSx for ONTAPファイルシステム

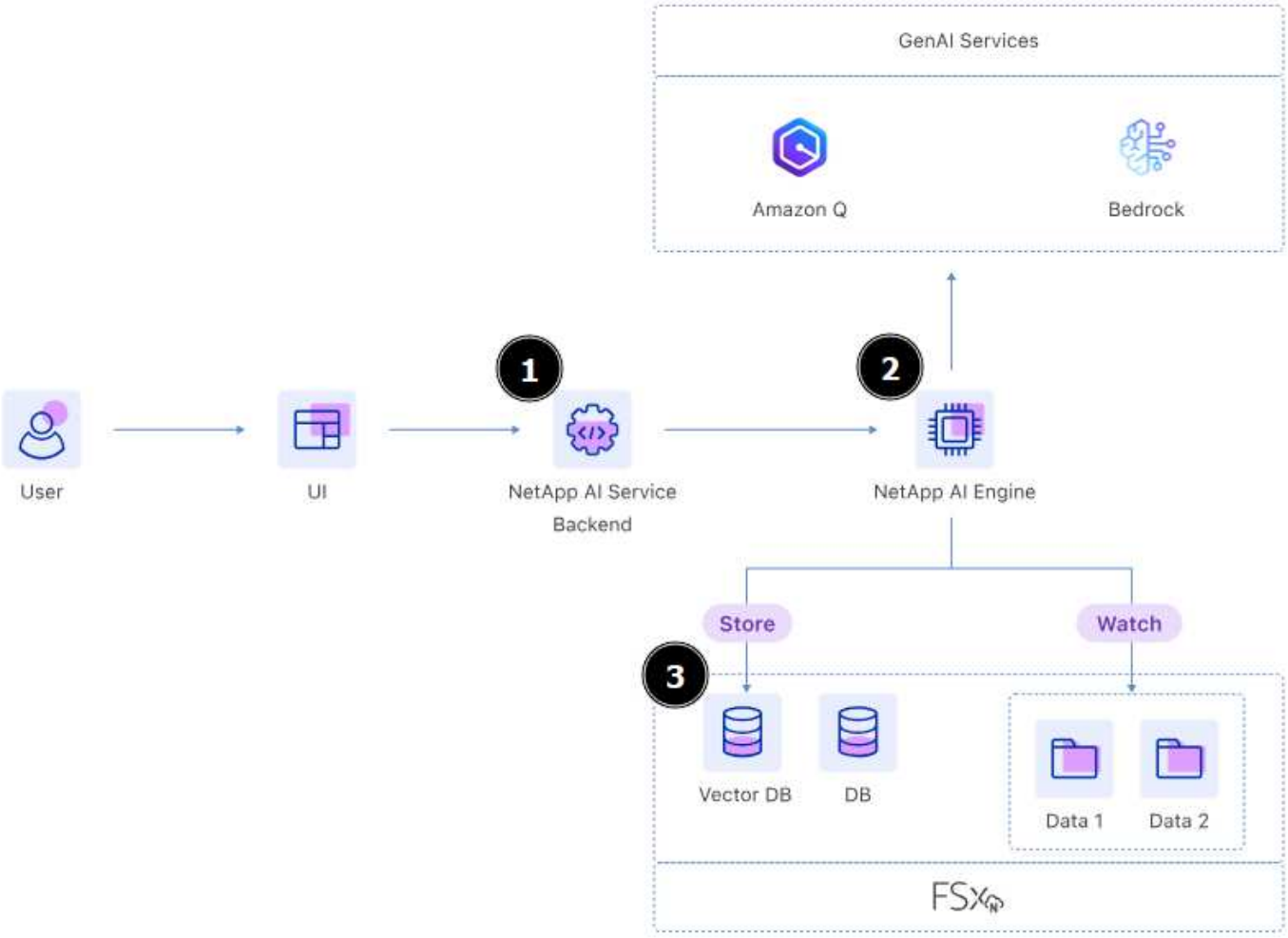
FSx for ONTAPファイルシステムは、GenAIシステムにストレージを提供します。

データソースに基づく基本モデルによって生成されたデータを格納するベクターデータベースを含む単一のボリュームが導入されます。

ナレッジベースに統合するデータソースは、同じFSx for ONTAPファイルシステムに配置することも、別のシステムに配置することもできます。

NetApp GenAIエンジンは、これらのボリュームの両方を監視し、相互作用します。

次の図は、GenAIインフラストラクチャを示しています。この手順では、番号1、2、3のコンポーネントを展開します。展開を開始する前に、他の要素が配置されている必要があります。



GenAIインフラの導入

AWSのクレデンシャルを入力し、FSx for ONTAPファイルシステムを選択して、Retrieval-Augmented Generation（RAG）インフラを導入する必要があります。

開始する前に

この手順を開始する前に、使用している環境がナレッジベースまたはコネクタの要件を満たしていることを確認してください。

- "ナレッジベースの要件"
- "コネクタの要件"

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. インフラの図を確認し、* Next *を選択します。
4. [AWS settings]セクションの項目を入力します。
   a. * AWSクレデンシャル*：AWSリソースを導入する権限を提供するAWSクレデンシャルを選択または追加します。
   b. 場所：AWSのリージョン、VPC、サブネットを選択します。

      GenAI環境は、Amazon Bedrockが有効になっているAWSリージョンに配置する必要があります。 "サポートされているリージョンの一覧を表示する"
5. [インフラストラクチャー設定]セクションの項目を入力します。
   a. タグ: このデプロイメントの一部であるすべての AWS リソースに適用するタグのキーと値のペアを入力します。これらのタグは、AWS マネジメントコンソールと Workload Factory 内のインフラストラクチャ情報領域に表示され、Workload Factory リソースを追跡するのに役立ちます。
6. [Connectivity]*セクションに入力します。
   a. キーペア: NetApp GenAIエンジンインスタンスに安全に接続できるキーペアを選択します。
7. 「AIエンジン*」セクションを完了します。
   a. インスタンス名: オプションで、*インスタンス名の定義*を選択し、AI エンジン インスタンスのカスタム名を入力します。インスタンス名は AWS マネジメントコンソールと Workload Factory 内のインフラストラクチャ情報領域に表示され、Workload Factory リソースを追跡するのに役立ちます。
8. 導入*を選択して導入を開始します。

> 導入がクレデンシャルエラーで失敗した場合は、エラーメッセージ内のハイパーリンクを選択すると、エラーの詳細を確認できます。見つからないかブロックされている権限のリストと、生成AIワークロードが 生成AIインフラを導入するために必要な権限のリストを確認できます。

結果

Workload Factory がチャットボット インフラストラクチャの展開を開始します。このプロセスには最大 10 分かかる場合があります。

導入プロセスでは、次の項目が設定されます。

- ネットワークはプライベートエンドポイントとともにセットアップされます。
- IAMロール、インスタンスプロファイル、およびセキュリティグループが作成されます。

- GenAIエンジンの仮想マシンインスタンスが導入されている。
- Amazon Bedrockは、プレフィックスを持つロググループを使用して、Amazon CloudWatch Logsにログを送信するように構成されて `/aws/bedrock/`います。
- GenAIエンジンは、次の名前のロググループを使用してAmazon CloudWatch Logsにログを送信するように設定されています。 `/netapp/wlmai/<tenancyAccountId>/randomId` 、 どこ `<tenancyAccountID>`は "NetAppコンソールアカウントID"現在のユーザーに対して。

生成AIナレッジベースの作成

AI インフラストラクチャをデプロイし、FSx for ONTAPデータストアからナレッジ ベースに統合するデータ ソースを特定したら、Workload Factory を使用してナレッジ ベースを構築する準備が整います。このステップでは、AI の特性を定義し、会話のきっかけも作成します。

作業を進める前に、環境がナレッジベースのを満たしていることを確認し"要件"てください。

タスクの内容

ナレッジベースには、_パブリックモード_と_エンタープライズモード_という2つのデータ統合モダリティがあります。

パブリックモード

ナレッジベースは、組織のデータソースを統合しなくても使用できます。この場合、ナレッジベースに統合されたアプリケーションは、インターネット上で公開されている情報からのみ結果を提供します。これは_public mode_integrationと呼ばれます。

Enterpriseモード

ほとんどの場合、組織のデータソースをナレッジベースに統合する必要があります。これは、エンタープライズからの知識を提供するため、_Enterprise mode_integrationと呼ばれます。

組織のデータ ソースには、個人を特定できる情報 (PII) が含まれている場合があります。この機密情報を保護するために、ナレッジ ベースを作成および構成するときに、データ ガードレール を有効にすることができます。 NetApp Data Classification を活用したデータ ガードレールは、PII を識別してマスクし、アクセス不能かつ回復不可能な状態にします。

"NetAppデータ分類について学ぶ" 。

NetApp Workload Factory for GenAI は、機密個人情報 (SPii) をマスクしません。参照"機密性の高い個人データのタイプ"このタイプのデータの詳細については、こちらをご覧ください。

データ ガードレールはいつでも有効化または無効化できます。データ ガードレールの有効化を切り替えると、Workload Factory はナレッジ ベース全体を最初からスキャンするため、コストが発生します。

ナレッジベースの作成と設定

ナレッジベースは、ナレッジベースの作成に使用するBedrock AIモデルや埋め込み形式などの特性を定義します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジ ベースとコネクタ] メニューから、[新規作成] ドロップダウンを選択し、[Bedrock 用のNetApp GenAI ナレッジ ベース] を選択します。
4. NetApp GenAI ナレッジ ベースの作成ページで、ナレッジ ベースの設定を構成します。

ナレッジベースの詳細

1. 名前:ナレッジベースに使用する名前を入力します。
2. 説明:ナレッジベースの詳細な説明を入力します。
3. **Bedrock**: AWS アカウントで Amazon Bedrock が利用できるリージョンを選択します。

摂取

1. 埋め込みモデル:

   - ナレッジベースに使用する埋め込みモデルを選択します。埋め込みモデルは、データが知識ベースのベクトル埋め込みに変換される方法を定義します。 Workload Factory は次のモデルをサポートしています。
   - Titan埋め込みG1 -テキスト
   - Titan埋め込みテキストv2
   - Titan Multimodal Embedding G1
   - 英語を埋め込む
   - 多言語を埋め込む

     Amazon Bedrockから埋め込みモデルを有効にしておく必要があります。

   "Amazon Titanの詳細"

   - 該当する場合は、選択した埋め込みモデルの構成に一致する推論タイプを選択します。
2. データ ガードレール: データ ガードレールを有効にするか無効にするかを選択します。"NetApp Data Classification を活用したデータ ガードレールについて学ぶ" 。

   データガードレールを有効にするには、次の前提条件を満たす必要があります。

   - NetApp Data Classification と通信するには、サービス アカウントが必要です。サービス アカウントを作成するには、 NetAppコンソール テナンシー アカウントで 組織管理者 ロールを持っている必要があります。組織管理者の役割を持つメンバーは、 内のすべてのアクションを完了できます。"NetAppコンソールでメンバーにロールを追加する方法を学びます"
   - AIエンジンは、"NetAppコンソールAPIエンドポイント" 。
   - 以下の手順に従ってください。"NetAppデータ分類ドキュメント" :
     i. コンソールエージェントを作成する
     ii. 環境が前提条件を満たしていることを確認する

iii. NetAppデータ分類を導入

CSV、JSON、JSONP、Parquetなどの構造化データファイルを取り込む場合、データガードレール機能はサポートされません。

チャットと検索の設定

1. チャットモデル:
   - Amazon Bedrock に統合されているさまざまなチャットモデルから選択します。 Amazon Bedrock からチャット モデルをすでに有効にしておく必要があることに注意してください。
   - 該当する場合は、選択したモデルの構成に一致する推論タイプを選択します。
2. チャット設定:
   - チャットボットの温度を選択して、応答のランダム性と創造性を設定します。温度が低いと、より予測可能な応答が得られ、温度が高いと、より多様な応答が得られます。
   - 応答の最大長を選択して、応答の詳細度を設定します。応答の長さが長くなると、より多くの応答トークンが使用され、コストが高くなる可能性があります。
3. 思考モード: 思考モードを有効にすると、チャットボットはクエリの処理に時間がかかり、結果は通常より正確になります。思考モードを有効にすると、結果を生成するときに使用する推論トークンの数を制御できます。推論トークンを多く使用すると、応答の精度は上がりますが、コストが高くなる可能性があります。
4. 再ランキング: 再ランキングを有効または無効にします。これにより、クエリ結果の関連性と品質が向上します。再ランク付けに使用する標準チャット モデルまたは特殊な再ランク付けモデルを選択します。Reranker モデル オプションは、お住まいの地域で利用可能な場合にのみ表示されます。選択したモデルの構成に一致する推論タイプを選択します。
5. 会話の開始：このナレッジベースを使用するチャットボットと対話するユーザーに表示される会話の開始プロンプトを最大4つ指定するかどうかを選択します。この設定を有効にすることをお勧めします。

   会話開始機能を有効にすると、デフォルトで「自動モード」が選択されます。「手動モード」は、ナレッジベースにデータソースを追加した後にのみ有効にできます。"ナレッジベースの設定を変更する方法について説明します。"です。

ストレージ定義

1. **FSx for ONTAP**ファイルシステム: 新しいナレッジベースを定義すると、Workload Factory によって、それを保存するための新しいAmazon FSx for NetApp ONTAPボリュームが作成されます。新しいボリュームを作成する既存のファイル システム名と SVM (ストレージ VM とも呼ばれます) を選択します。
2. スナップショット ポリシー: Workload Factory ストレージ インベントリで定義されている既存のポリシーのリストからスナップショット ポリシーを選択します。ナレッジ ベースの定期的なスナップショットは、選択したスナップショット ポリシーに基づいた頻度で自動的に作成されます。
3. **S3** バケット: チャットボットのクエリ結果に構造化データが含まれている場合、GenAI はその結果を S3 バケットに保存できます。この機能を使用するには、*S3 バケットのアクティブ化*設定を有効にし、リストからアカウントに関連付けられている S3 バケットを選択します。これらの結果が S3 バケットに保存されると、チャット セッション内のダウンロード リンクを使用してダウンロードできます。

   必要なSnapshotポリシーが存在しない場合は "Snapshot ポリシーを作成します"、ボリュームを含むStorage VM上で実行できます。

4. [ナレッジベースの作成]*を選択して、ナレッジベースをGenAIに追加します。

   ナレッジベースの作成中は、進行状況インジケータが表示されます。

   ナレッジベースを作成したら、新しいナレッジベースにデータソースを追加するか、データソースを追加せずにプロセスを終了するかを選択できます。[データソースの追加]*を選択し、ここで1つ以上のデータソースを追加することをお勧めします。

ナレッジベースへのデータソースの追加

1つまたは複数のデータソースを追加して、組織のデータをナレッジベースに入力できます。

タスクの内容

サポートされるデータソースの最大数は10です。

手順

1. データ ソースの追加 を選択した後、追加するデータ ソースの種類を選択します。
   - FSx for ONTAP ファイルシステムを追加する (既存の FSx for ONTAP ボリュームのファイルを使用)
   - ファイルシステムを追加する（一般的な SMB または NFS 共有のファイルを使用）

FSx for ONTAP ファイルシステムを追加する

1. ファイルシステムを選択：データソースファイルが存在するFSx for ONTAPファイルシステムを選択し、* Next *を選択します。
2. ボリュームを選択：データソースファイルが格納されているボリュームを選択し、*[次へ]*を選択します。

   SMBプロトコルを使用して保存されているファイルを選択する場合は、ドメイン、IPアドレス、ユーザ名、パスワードなどのActive Directory情報を入力する必要があります。

3. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。
4. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。

   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。

     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。

     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

     画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

     画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

   - [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

5. 権限対応*セクション（選択したデータソースがSMBプロトコルを使用するボリューム上にある場合にのみ表示）で、権限対応の応答を有効または無効にできます。
   - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
   - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
6. [追加]*を選択して、このデータソースをナレッジベースに追加します。

汎用**NFS**ファイルシステムを追加する

1. ファイル システムを選択: データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力し、ネットワーク共有の NFS プロトコルを選択して、次へ を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

場合によっては、NFSエクスポート名を手動で入力し、「ディレクトリを取得」を選択して利用可能なディレクトリを表示する必要があります。エクスポート全体を選択するか、エクスポートから特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。
     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

> 画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

> 変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

汎用**SMB**ファイルシステムを追加する

1. ファイルシステムを選択:
   a. データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力します。
   b. ネットワーク共有の SMB プロトコルを選択します。
   c. ドメイン、IP アドレス、ユーザー名、パスワードを含む Active Directory 情報を入力します。
   d. 「 * 次へ * 」を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

> 場合によっては、SMB共有名を手動で入力し、「ディレクトリの取得」を選択して利用可能なディレクトリを表示する必要があります。共有全体を選択するか、共有内の特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

- 権限認識: 権限認識応答を有効または無効にします。
  - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
  - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
- ファイルフィルタリング:スキャンに含めるファイルを設定します。
  - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

    画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

  画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

  画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

  変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

結果

データソースがナレッジベースに埋め込まれ始めます。データソースが完全に埋め込まれると、ステータスが「埋め込み」から「埋め込み」に変わります。

単一のデータソースをナレッジベースに追加したら、チャットボットシミュレータウィンドウでローカルにテストし、必要な変更を加えてから、ユーザーがチャットボットを使用できるようにします。同じ手順に従って、ナレッジベースにデータソースを追加することもできます。

生成AIナレッジベースのテスト

ナレッジベースを作成したら、チャットボットシミュレータを使用してローカルでテス

トし、必要な変更を加えてから、チャットボットアプリケーションを使用してユーザーがナレッジベースを利用できるようになります。

タスクの内容

ナレッジベースをテストして、期待どおりに動作することを確認します。また、このナレッジベースのチャットボットユーザーがデフォルトで使用できるようにする会話のスターターをカスタマイズできます。チャットボットシミュレータは、ナレッジベースに埋め込まれたすべてのデータソースに対して実行されます。

チャットボットシミュレータで埋め込みデータソースとチャットすることで、ナレッジベースをテストできます。ナレッジベースをローカルでテストする場合、インタラクションやインサイトはGenAIベクターデータベースにキャプチャされないことに注意してください。

ユーザー向けのアプリケーションにナレッジ ベースを展開する前に、ほとんどのテストは Workload Factory 内で実行します。データ ソースまたはチャットボットの操作に変更を加える必要がある場合は、ナレッジ ベースを公開する前に変更を行う必要があります。

チャットボットシミュレータウィンドウのサイズを変更してタイトルを変更したり、質問や回答をクリップボードにコピーしたりできます。

チャットボットをテストするために実行するタスクには、次のものがあります。

- 回答が期待どおりであることを確認するために、組織に関連する多数の質問を入力します。
- チャットボットアプリケーションでユーザーがデフォルトで使用できるようにする会話のスターターをカスタマイズします。
- チャットボットの回答の下部に表示される属性コンテンツに、正しい参照が含まれていることを確認してください。

手順

1. ナレッジベースインベントリページで、テストするナレッジベースを選択します。

   右側のペインにチャットボットシミュレータが表示されます。定義されている場合は、既存のカンバセーションスターターも表示されます。

2. チャットボット入力フィールドで、プロンプトまたは質問を入力し、を選択して、 組織の知識を使用してチャットボットがどのように応答するかを確認します。

   - 回答の下の*ソース*リストを展開すると、回答の作成に使用されたソースを確認できます。これにより、アンサーの生成に使用されるファイルのリストが表示されます。ファイル名にカーソルを合わせると、各ファイルおよびボリュームパスで使用されているデータチャンクを表示およびコピーできます。
   - 回答に表が含まれている場合は、各列のデータを並べ替え、各表をクリップボードにコピーできます。
   - 回答結果に構造化データが含まれており、ナレッジベースで **S3** バケット 機能が有効になっている場合、GenAI は結果を S3 バケットに保存します。チャット セッション内の結果のダウンロード リンクを使用して、バケットから結果をダウンロードできます。

3. ナレッジベースでより焦点を絞った回答が得られるようにデータソースを更新する必要がある場合は、それらの変更を今すぐ行ってから、ナレッジベースを再テストします。

生成AIナレッジベースの外部認証のアクティブ化

APIエンドポイントを使用してナレッジベースとチャットボットアプリケーションを統合するときにトークンの検証とACLが必要になるように、ナレッジベースの認証をアクティブにします。認証を有効にするときは、チャットボットクライアントからナレッジベースへのAPI要求に使用されるJSON Webトークンの設定を構成します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、認証を有効にするナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。
5. メニューを選択し、[認証設定の管理]*を選択します。
6. 認証をセットアップします。

    a. [認証設定をアクティブにする]*を選択します。

    b. 必要な情報を入力します。ここでは例を示しますが、これらのフィールドの値は認証プロバイダから取得する必要があります。

    - アルゴリズム:認証プロバイダが使用する署名アルゴリズム。
    - * Audience *（オプション）：トークンの受信者（URLの場合もあります）を含む文字列。
    - **Issuer**:トークンを発行したプロバイダを識別する文字列。

      たとえば、Amazon Cognitoでは、次の形式の発行者文字列が使用されます。

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

      `<region>`はユーザプールを含むAWSリージョン、はユーザプール `<UserPoolID>` IDです。次のコマンドを使用して、ユーザプールIDを取得できます。

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

    - **JWKS URI**:このトークンの署名を検証するために必要な公開鍵を提供するURI文字列。

      たとえば、Amazon Cognitoでは、次の形式のJWKS URI文字列が使用されます。

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

    +

`<region>`はユーザプールを含むAWSリージョン、はユーザプール `<UserPoolID>` IDです。次のコマンドを使用して、ユーザプールIDを取得できます。

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

7. [保存（ Save ）] を選択します。

結果

ナレッジベースの認証がアクティブになり、APIエンドポイントを使用してナレッジベースと対話し、ナレッジベースとチャットボットアプリケーションを統合できます。

生成AIナレッジベースを公開し、一意のエンドポイントを表示する

ナレッジベースをローカルで構築してテストしたら、ナレッジベースを公開して、ユーザーがナレッジベースを照会できるようにするチャットボットアプリケーションにナレッジベースを統合できます。

タスクの内容

ナレッジ ベースを公開すると、チャット アプリケーションで使用できるようになります。公開アクションは、Workload Factory API をトリガーして、一意のエンドポイントを生成および公開します。公開後、ナレッジ ベースはチャット アプリケーションからアクセスできるようになり、API エンドポイントを統合できるようになります。

公開する各ナレッジベースには、固有のエンドポイントがあります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、公開するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、およびすべての埋め込みデータソースのリストが表示されます。

5. [アクション（Actions）]メニューを選択し、[パブリッシュ（Publish）]を選択します。

   Workload Factory がナレッジ ベースを公開します。ナレッジ ベースの詳細ページで、ステータスが 未公開 から 公開 に変わります。

   これで、ナレッジベースの一意のエンドポイントに関する詳細を取得できます。

6. [公開済み]ステータスの横にある*[表示]*を選択します。

   Workload Factory API を使用してナレッジ ベースにアクセスする方法の詳細が表示されます。

7. [公開情報の表示]ダイアログボックスから、ナレッジベースとアプリケーションの統合に使用できるAPIエ

ンドポイントをコピーします。

APIエンドポイントの詳細については、に移動し "APIドキュメント" 、* AI > External *を選択します。

これらのエンドポイントを使用する前に、認証プロバイダからユーザトークンを取得する必要があります。

結果

これで、公開されたナレッジベースと、ナレッジベースとチャットボットアプリケーションを統合するために使用できる一意のエンドポイントが作成されました。

生成AI外部サンプルチャットボットアプリケーションを使用する

ナレッジ ベースを設定、アクティブ化、公開すると、外部のアプリケーション開発者は、 NetAppが提供するオープン ソースのサンプル チャットボット アプリケーションを設定および実行して、ナレッジ ベースを操作し、Workload Factory API を使用して独自の生成 AI アプリケーションを作成する方法を学習できるようになります。

手順

1. "ナレッジベースの作成"です。
2. "認証をアクティブ化" をクリックします。

   これにより、ナレッジベースでAPI要求を認証できるようになり、APIエンドポイントを使用する際にトークンの検証とACLが必要になります。

   > このナレッジベースと統合する外部チャットアプリケーションは、ナレッジベースの認証設定で設定したものと同じ認証プロバイダ(発行者)を使用する必要があります。

3. "ナレッジベースの公開" 外部アプリケーションのAPIアクセスを有効にします。

   ナレッジベースが公開されると、APIエンドポイントに外部からアクセスできるようになり、ナレッジベースを外部チャットアプリケーション（例のチャットボットアプリケーションなど）と統合できます。

4. からサンプルのチャットボットアプリケーションパッケージをダウンロードし "GitHub"ます。
5. パッケージに含まれているREADMEファイルの指示に従って、チャットボットアプリケーションをインストールして実行します。
6. を参照し "http://localhost:9091" てアプリケーションにログインします。

   チャットボットアプリケーションの例が表示されます。

詳細

"ワークロードファクトリー API ドキュメント"

RAGベースの 生成AIアプリケーションの作成

ナレッジベースを構築してチャットボットをテストしたら、ユーザーがチャットボットを照会できるようにするアプリケーションをセットアップする準備が整いました。

"FSx for ONTAPでRAGベースのAIアプリケーションを作成する方法をご確認ください"

生成AIで次にできること

エンタープライズデータを使用してナレッジベースを作成し、ユーザ向けに導入したので、ナレッジベース、データソース、RAGインフラ（FSx for ONTAPファイルシステムを含む）を管理できます。

ナレッジベースコンポーネントを管理するために実行できるタスクには、次のものがあります。

- データソースのコンテンツを更新するか、新しいデータソースを追加し、それらの変更をナレッジベースやチャットボットと同期します。
- チャンキング戦略や権限対応（SMBファイルアクセスの場合）など、データソースの設定を管理します。
- チャットモデルや会話のスターターなど、ナレッジベースの設定を管理します。
- ナレッジベースをアンパブリッシュするか、変更後に再パブリッシュします。
- FSx for ONTAPファイルシステム上の重要なデータをバックアップして保護し、ナレッジベースデータやその他のインフラコンポーネントを常に利用できるようにします。

FSx for ONTAPファイルシステムの管理については、 "Amazon FSx for NetApp ONTAP のワークロードファクトリーのドキュメント"使用できるバックアップおよび保護機能を表示します。

[1] 構造化データファイルをナレッジベースに取り込む場合、データガードレール機能はサポートされていません。

生成AIを使用してAmazon Q Business用のコネクタを作成する

はじめに

生成AIコネクタのクイックスタート

Amazon FSx for NetApp ONTAP ファイルシステムに存在する組織のデータを使用して、Amazon Q Business 用の NetApp コネクタの作成を開始します。コネクタを作成すると、エンドユーザーはAmazon Q Businessアシスタントにアクセスして、質問に対する組織中心の回答を得ることができます。

1 ワークロードファクトリーにログイン

必要となるのは "Workload Factoryでアカウントを設定する"次のいずれかを使用してログインします "コンソールエクスペリエンス"。

2 生成AIの要件を満たす環境のセットアップ

AWSインフラの導入、導入および検出されたFSx for ONTAPファイルシステム、コネクタに統合するデータソースのリスト、Amazon Q Businessアプリケーションへのアクセスなどには、AWSのクレデンシャルが必要です。

"生成AI要件の詳細"です。

3 データソースが格納されている**FSx for ONTAP**ファイルシステムを特定する

コネクタに統合するデータソースは、単一のFSx for ONTAPファイルシステムに配置することも、複数のFSx for ONTAPファイルシステムに配置することもできます。これらのシステムが異なるVPCにある場合は、同じネットワーク内でアクセス可能であるか、またはAIエンジンと同じリージョンとAWSアカウントを使用してVPC間でピア関係を確立しておく必要があります。

"データソースを特定する方法"です。

4 生成AIインフラの導入

インフラ導入ウィザードを起動して、AWS環境に 生成AIインフラを導入します。このプロセスでは、NetApp生成AIエンジンのEC2インスタンスと、FSx for ONTAPファイルシステムにNetApp AIエンジンのデータベースを格納するボリュームを導入します。ボリュームは、コネクタに関する情報を格納するために使用されます。

"生成AIインフラの導入方法をご確認ください"です。

次のステップ

Amazon Q Business用のコネクタを作成して、組織に焦点を当てた回答をエンドユーザーに提供できるようになりました。

生成AI Connectorの要件

Amazon Q Business 用のNetAppコネクタを作成する前に、Workload Factory と AWS が適切にセットアップされていることを確認してください。

生成AIの基本要件

生成AIには、作業を開始する前に環境が満たす必要のある一般的な要件があります。

Workload Factoryのログインとアカウント

必要となるのは "Workload Factoryでアカウントを設定する"次のいずれかを使用してログインします "コンソールエクスペリエンス"。

AWS のクレデンシャルと権限

Workload Factory に読み取り/書き込み権限を持つ AWS 認証情報を追加する必要があります。つまり、Workload Factory を GenAI の 読み取り/書き込み モードで使用することになります。

現時点では、_基本_モードと_読み取り専用_モードの権限はサポートされていません。

クレデンシャルを設定する際に、以下に示すように権限を選択すると、FSx for ONTAPファイルシステムの管理、GenAI EC2インスタンスおよびナレッジベースとチャットボットに必要なその他のAWSリソースの導入と管理を行うためのフルアクセスが提供されます。

"Workload FactoryにAWS認証情報を追加する方法を学ぶ"

Amazon Q Business 向け **NetApp** コネクタの要件

Amazon Q Business Connectorの次の特定の要件を環境が満たしていることを確認します。

Amazon Q Businessアプリケーション

Amazon Q Businessアプリケーションを作成するか、既存のアプリケーションを使用する必要があります。

- アプリケーションがいずれかのAWSリージョンに存在することを確認します。
- アプリケーションにがあることを確認し "インデックスを作成しました"ます。
- アプリケーションが失敗した状態でないことを確認します。

FSx for ONTAPファイルシステム

少なくとも1つのFSx for ONTAPファイルシステムが必要です。

- NetAppの 生成AIエンジンは、コネクタに関する情報を格納するために1つのファイルシステムを使用（存在しない場合は作成）します。

  このFSx for ONTAPファイルシステムでは、FlexVolボリュームを使用する必要があります。FlexGroupボリュームはサポートされません。

- 1つまたは複数のファイルシステムには、コネクタに追加するデータソースが含まれます。

  1つのFSx for ONTAPファイルシステムを両方の目的に使用することも、複数のFSx for ONTAPファイルシステムを使用することもできます。

- AWS FSx for ONTAPファイルシステムが配置されているAWSリージョン、VPC、サブネットを把握しておく必要があります。
- この導入に含まれるAWSリソースに適用するタグのキーと値のペアを検討する必要があります（オプション）。
- NetApp AIエンジンインスタンスに安全に接続するためのキーペア情報を知っておく必要があります。

"FSx for ONTAPファイルシステムの導入と管理の方法をご確認ください"

コネクタに追加するデータソースを特定する

コネクタに統合するFSx for ONTAPファイルシステムに存在するドキュメント（データソース）を特定または作成します。これらのデータソースを使用すると、Amazon Q Businessは、組織に関連するデータに基づいて、ユーザークエリに対して正確でパーソナライズされた回答を提供できます。

データソースの最大数

サポートされるデータソースの最大数は10です。

データソースの場所

データソースは、単一のボリュームに格納することも、ボリューム内のフォルダに格納することも、Amazon FSx for NetApp ONTAPファイルシステム上のSMB共有やNFSエクスポートに格納することもできます。また、NetApp SnapMirrorデータ保護関係にあるAmazon FSx for NetApp ONTAPボリュームにデータソースを保存することもできます。

ボリュームまたはフォルダ内の個 々 のドキュメントを選択することはできません。したがって、データソースを含む各ボリュームまたはフォルダに、ナレッジベースと統合すべきではない無関係なドキュメントが含まれていないことを確認する必要があります。

各コネクタには複数のデータソースを追加できますが、すべてのデータソースをAWSアカウントからアクセスできるFSx for ONTAPファイルシステムに配置する必要があります。

各データソースの最大ファイルサイズは50MBです。

サポートされるプロトコル

コネクタは、NFSまたはSMB / CIFSプロトコルを使用するボリュームのデータをサポートします。SMBプロトコルを使用して保存されているファイルを選択する場合は、コネクタがこれらのボリューム上のファイルにアクセスできるように、Active Directory情報を入力する必要があります。これには、Active Directoryドメイン、IPアドレス、ユーザ名、パスワードが含まれます。

SMB経由でアクセスされる共有（ファイルまたはディレクトリ）にデータソースを格納する場合、その共有にアクセスする権限を持つチャットボットのユーザまたはグループのみがデータにアクセスできます。この「権限認識機能」が有効になっている場合、AIシステムはAuth0内のユーザのEメールを、SMB共有上のファイルの表示または使用を許可されているユーザと比較します。チャットボットは、埋め込まれたファイルのユーザー権限に基づいて回答を提供します。

たとえば、コネクタに10個のファイル(データソース)を統合し、そのうちの2個が制限された情報を含む人事ファイルである場合、これら2つのファイルへのアクセスを認証されたチャットボットユーザーのみが、それ

らのファイルからのデータを含むチャットボットから応答を受け取ります。

Amazon Q Business Connectorにデータソースを追加する場合、データソースファイルにはユーザー権限のみが適用されます。グループ権限は適用されません。

データソース内のファイルにテキストがない場合(テキストのない画像など)、Amazon Q Businessはインデックスを作成しませんが、テキストがないことを示すエントリをAmazon CloudWatch Logsに記録します。

サポートされるデータソースファイル形式

NetApp Connector for Amazon Q Business では現在、次のデータ ソース ファイル形式がサポートされています。

ファイル形式	エクステンション
カンマ区切り値ファイル	.csv
JSONとJSONP	.json
マークダウン	.md
Microsoft Word	.docx
プレーンテキスト	.txt
ポータブルドキュメントフォーマット	.pdf
Microsoft PowerPoint	.pptまたは.pptx
ハイパーテキストマークアップ言語	.html からのアクセスが可能です
拡張マークアップ言語	.xml に保存されます
XSLT	.XSLT
Microsoft Excel	.xls
リッチテキスト形式	.rtf

GenAIインフラの導入

組織のFSx for ONTAPナレッジベース、コネクタ、アプリケーションを構築する前に、生成AI Infrastructure for RAGフレームワークを環境に導入する必要があります。主要なインフラコンポーネントは、Amazon Bedrockサービス、NetApp生成AIエンジンの仮想マシンインスタンス、FSx for ONTAPファイルシステムです。

展開されたインフラストラクチャは、複数のナレッジベース、チャットボット、コネクタをサポートできるため、通常はこのタスクを一度だけ実行する必要があります。

インフラの詳細

GenAI環境は、Amazon Bedrockが有効になっているAWSリージョンに配置する必要があります。 "サポートされているリージョンの一覧を表示する"

インフラストラクチャは、次のコンポーネントで構成されています。

Amazon Bedrockサービス

Amazon Bedrockは、業界をリードするAI企業の基盤モデル（FMS）を単一のAPIで使用できるフルマネージドサービスです。また、セキュアなジェネレーティブAIアプリケーションの構築に必要な機能も提供します。

"Amazon Bedrockの詳細"

Amazon Q Business

Amazon QはAmazon Bedrock上に構築されており、フルマネージドのジェネレーティブAIアシスタントを使用して、データソースからの情報に基づいて質問に答えたりコンテンツを生成したりできます。

"Amazon Q Businessの詳細"

NetApp GenAIエンジン用の仮想マシン

このプロセスでは、NetApp GenAIエンジンがデプロイされます。データソースからデータを取り込み、そのデータをベクターデータベースに書き込む処理能力を提供します。

FSx for ONTAPファイルシステム

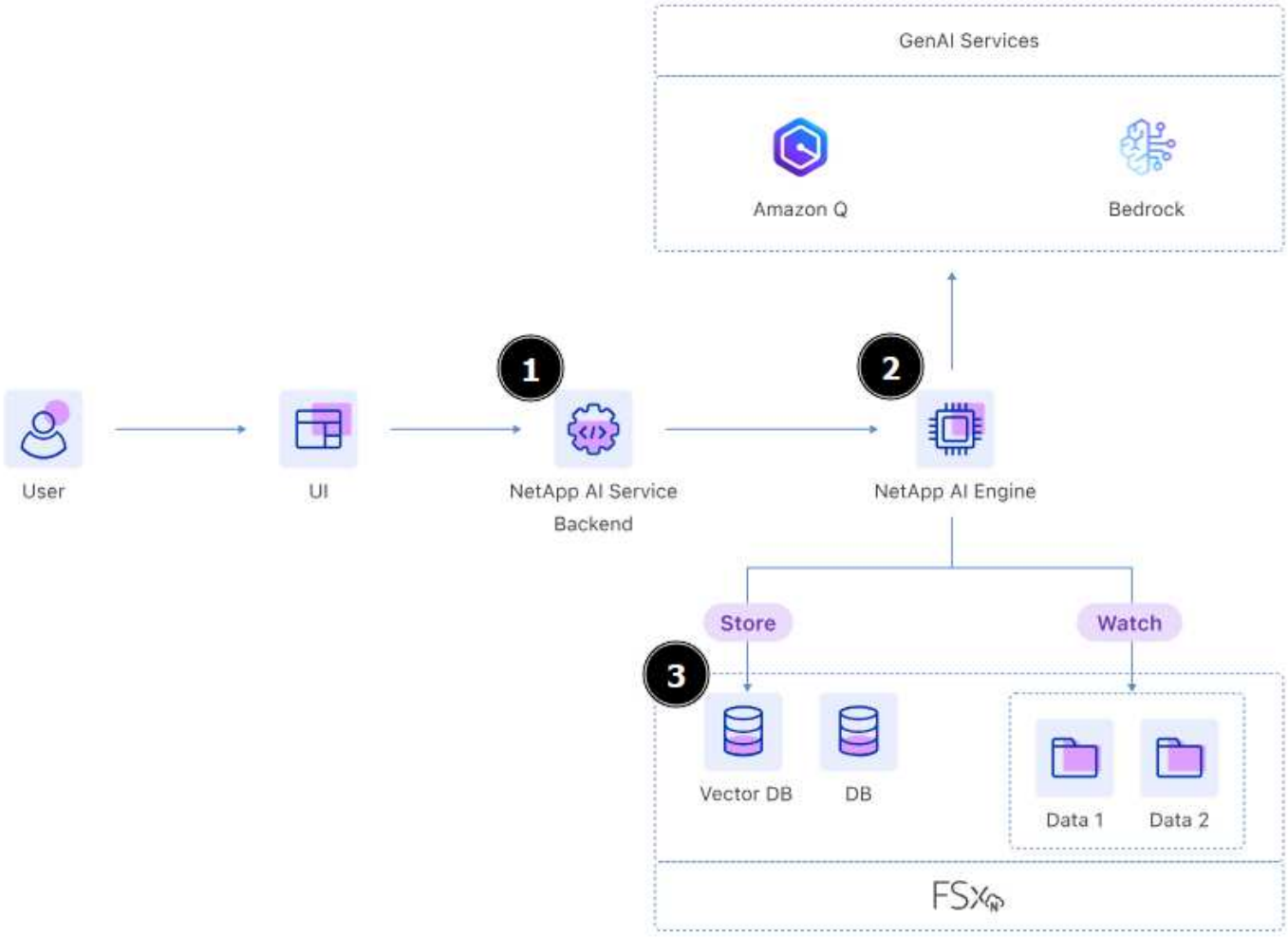
FSx for ONTAPファイルシステムは、GenAIシステムにストレージを提供します。

データソースに基づく基本モデルによって生成されたデータを格納するベクターデータベースを含む単一のボリュームが導入されます。

ナレッジベースに統合するデータソースは、同じFSx for ONTAPファイルシステムに配置することも、別のシステムに配置することもできます。

NetApp GenAIエンジンは、これらのボリュームの両方を監視し、相互作用します。

次の図は、GenAIインフラストラクチャを示しています。この手順では、番号1、2、3のコンポーネントを展開します。展開を開始する前に、他の要素が配置されている必要があります。



GenAIインフラの導入

AWSのクレデンシャルを入力し、FSx for ONTAPファイルシステムを選択して、Retrieval-Augmented Generation（RAG）インフラを導入する必要があります。

開始する前に

この手順を開始する前に、使用している環境がナレッジベースまたはコネクタの要件を満たしていることを確認してください。

- "ナレッジベースの要件"
- "コネクタの要件"

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. インフラの図を確認し、* Next *を選択します。
4. [AWS settings]セクションの項目を入力します。
   a. * AWSクレデンシャル*：AWSリソースを導入する権限を提供するAWSクレデンシャルを選択または追加します。

b. 場所：AWSのリージョン、VPC、サブネットを選択します。

GenAI環境は、Amazon Bedrockが有効になっているAWSリージョンに配置する必要があります。 "サポートされているリージョンの一覧を表示する"

5. [インフラストラクチャー設定]セクションの項目を入力します。
   a. タグ: このデプロイメントの一部であるすべての AWS リソースに適用するタグのキーと値のペアを入力します。これらのタグは、AWS マネジメントコンソールと Workload Factory 内のインフラストラクチャ情報領域に表示され、Workload Factory リソースを追跡するのに役立ちます。
6. [Connectivity]*セクションに入力します。
   a. キーペア: NetApp GenAIエンジンインスタンスに安全に接続できるキーペアを選択します。
7. 「AIエンジン*」セクションを完了します。
   a. インスタンス名: オプションで、*インスタンス名の定義*を選択し、AI エンジン インスタンスのカスタム名を入力します。インスタンス名は AWS マネジメントコンソールと Workload Factory 内のインフラストラクチャ情報領域に表示され、Workload Factory リソースを追跡するのに役立ちます。
8. 導入*を選択して導入を開始します。

導入がクレデンシャルエラーで失敗した場合は、エラーメッセージ内のハイパーリンクを選択すると、エラーの詳細を確認できます。見つからないかブロックされている権限のリストと、生成AIワークロードが 生成AIインフラを導入するために必要な権限のリストを確認できます。

結果

Workload Factory がチャットボット インフラストラクチャの展開を開始します。このプロセスには最大 10 分かかる場合があります。

導入プロセスでは、次の項目が設定されます。

- ネットワークはプライベートエンドポイントとともにセットアップされます。
- IAMロール、インスタンスプロファイル、およびセキュリティグループが作成されます。
- GenAIエンジンの仮想マシンインスタンスが導入されている。
- Amazon Bedrockは、プレフィックスを持つロググループを使用して、Amazon CloudWatch Logsにログを送信するように構成されて `/aws/bedrock/`います。
- GenAIエンジンは、次の名前のロググループを使用してAmazon CloudWatch Logsにログを送信するように設定されています。 `/netapp/wlmai/<tenancyAccountId>/randomId` 、 どこ `<tenancyAccountID>`は "NetAppコンソールアカウントID"現在のユーザーに対して。

Amazon Q Business 用の NetApp コネクタを作成する

AI インフラストラクチャをデプロイし、FSx for ONTAP データストアから使用するデータソースを特定したら、Amazon Q Business 用の NetApp コネクタを定義する準備が整います。

続行する前に、ご使用の環境がfor Amazon Q Businessの要件を満たしていることを確認し"要件"てください。

タスクの内容

組織のデータ ソースには、個人を特定できる情報 (PII) が含まれている可能性があります。この機密情報を保護するために、コネクタを定義するときに データ ガードレール を有効にすることができます。 NetApp Data Classification を活用したデータ ガードレールは、PII を識別してマスクし、アクセス不能かつ回復不可能な状態にします。

"NetAppデータ分類について学ぶ" 。

NetApp Workload Factory for GenAI は、機密個人情報 (SPII) をマスクしません。参照"機密性の高い個人データのタイプ"このタイプのデータの詳細については、こちらをご覧ください。

データ ガードレールはいつでも有効化または無効化できます。データ ガードレールの有効化を切り替えると、Workload Factory はデータ ソース全体を最初からスキャンするため、コストが発生する可能性があります。

コネクターの定義

Amazon Q Business用のNetAppコネクタを作成します。このコネクタにより、GenAIとAmazon Q Business間のAPIおよびデータソース通信が可能になります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. 「ナレッジベースとコネクタ」メニューから、「新規作成」ドロップダウンを選択し、「**Amazon Q** ビジネス コネクタ」を選択します。
4. [コネクタの定義（Define Connector）]ページで、コネクタの設定を構成します。
   a. 名前:コネクタに使用する名前を入力します。
   b. 説明：コネクタの詳細な説明を入力します。
   c. * Amazon Q *：統合するAmazon Q Businessインスタンスのリージョンとアプリケーション名。
   d. データ ガードレール: データ ガードレールを有効にするか無効にするかを選択します。"NetApp Data Classification を活用したデータ ガードレールについて学ぶ" 。

      データガードレールを有効にするには、次の前提条件を満たす必要があります。

      - NetApp Data Classification と通信するには、サービス アカウントが必要です。サービス アカウントを作成するには、 NetAppコンソール テナンシー アカウントで 組織管理者 ロールを持っている必要があります。組織管理者ロールを持つメンバーは、 NetAppコンソールですべてのアクションを完了できます。"NetAppコンソールでメンバーにロールを追加する方法を学びます"
      - AIエンジンは、"NetAppコンソールAPIエンドポイント" 。
      - 以下の手順に従ってください。"NetAppデータ分類ドキュメント" :
        A. コンソールエージェントを作成する
        B. 環境が前提条件を満たしていることを確認する
        C. NetAppデータ分類を導入

データガードレール機能を有効にすると、生成AIは.txt、.md、.csv、.docx、および.pdfファイルを処理し、プレーンテキスト（埋め込み画像またはメディアテキストを除く）のみを取り込み、プライベートデータまたは機密データをマスキングします。他のすべてのファイルタイプは、プライベートデータや機密データをマスキングすることなく正常に処理されます。

e. **FSx for ONTAP**ファイルシステム: Amazon Q Business 用の新しいNetAppコネクタを定義すると、Workload Factory によってコネクタ情報を保存するための新しいAmazon FSx for NetApp ONTAPボリュームが作成されます。新しいボリュームを作成する既存のファイル システムと SVM (ストレージ VM とも呼ばれます) を選択します。

f. スナップショット ポリシー: Workload Factory ストレージ インベントリで定義されている既存のポリシーのリストからスナップショット ポリシーを選択します。 GenAI は、選択したスナップショット ポリシーに基づいて、コネクタ情報を格納するボリュームの定期的なスナップショットを自動的に作成します。

必要なSnapshotポリシーが存在しない場合は "Snapshot ポリシーを作成します"、ボリュームを含むStorage VM上で実行できます。

5. [コネクタの作成]*を選択して、Amazon Q Businessと 生成AIを統合します。

コネクタの作成中は、進行状況インジケータが表示されます。

コネクタが作成されたら、Amazon Q Businessがデータを取り込んでインデックスに追加するように、コネクタにデータソースを追加するオプションがあります。[データソースの追加]*を選択し、ここで1つ以上のデータソースを追加することをお勧めします。

コネクタへのデータソースの追加

1つ以上のデータソースを追加して、組織のデータをAmazon Q Businessインデックスに入力できます。

タスクの内容

- サポートされるデータソースの最大数は10です。
- Amazon Q Businessインデックスの特定のサービス制限については、を参照して "Amazon Q Businessドキュメント"ください。

手順

1. データ ソースの追加*を選択すると、*ファイル システムの選択 ページが表示されます。
2. ファイルシステムを選択：データソースファイルが存在するFSx for ONTAPファイルシステムを選択し、* Next *を選択します。
3. ボリュームを選択：データソースファイルが格納されているボリュームを選択し、*[次へ]*を選択します。

SMBプロトコルを使用して保存されているファイルを選択する場合は、ドメイン、IPアドレス、ユーザ名、パスワードなどのActive Directory情報を入力する必要があります。

4. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。
5. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。

- ファイルフィルタリング:スキャンに含めるファイルを設定します。
  - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。
  - [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

> ⓘ 変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

6. 権限対応*セクション（選択したデータソースがSMBプロトコルを使用するボリューム上にある場合にのみ表示）で、権限対応の応答を有効または無効にできます。
   - 有効:このコネクタにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
   - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。

> ⓘ Active Directoryグループ権限は、Amazon Q Business Connectorデータソースではサポートされていません。

7. このデータソースをAmazon Q Business Connectorに追加するには、*[追加]*を選択します。

結果

データソースはAmazon Q Businessインデックスに埋め込まれています。データソースが完全に埋め込まれると、ステータスが「埋め込み」から「埋め込み」に変わります。

コネクタに単一のデータソースを追加した後、Amazon Q Businessチャットボット環境でテストし、ユーザーがサービスを利用できるようにする前に必要な変更を加えることができます。同じ手順に従って、コネクタに追加データソースを追加することもできます。

管理と監視

GenAIインフラの管理

デプロイされたGenAI RAGインフラストラクチャの詳細を表示したり、不要になった場合はチャットボットインフラストラクチャを削除したりできます。

インフラに関する情報を表示する

チャットボットインフラストラクチャに関する情報を表示できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. *インフラストラクチャ*メニューを選択します。
4. 次のコンポーネントの詳細など、インフラに関する情報を表示します。
    - AWS設定
    - インフラ設定
    - AIエンジン
    - ベクターデータベース

インフラストラクチャの削除

1 つ以上のチャットボット用にデプロイしたチャットボット インフラストラクチャが不要になった場合は、Workload Factory から削除できます。

> このインフラストラクチャに展開されているすべてのチャットボットが無効になり、すべてのチャット履歴が削除されます。

この操作では、Workload Factory から AI インフラストラクチャへのリンクのみが削除され、AWS からすべてのコンポーネントが削除されるわけではありません。次のインフラストラクチャ コンポーネントを AWS から手動で削除する必要があります。

- VMインスタンス
- プライベートエンドポイント
- AIデータベースを含むFSx for ONTAPファイルシステム上のボリューム
- IAMロール
- ポリシー
- セキュリティグループ

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。

2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. *インフラストラクチャ*メニューを選択します。
4. を選択し ••• 、*チャットボットインフラストラクチャの削除*を選択します。
5. インフラを削除することを確認し、*[削除]*を選択します。

結果

チャットボット インフラストラクチャ コンポーネントが Workload Factory から削除されます。

生成AIナレッジベースの管理

ナレッジベースを作成したら、ナレッジベースの詳細の表示、ナレッジベースの変更、追加のデータソースの統合、ナレッジベースの削除を行うことができます。

ナレッジベースに関する情報を表示する

ナレッジベースと統合されているデータソースの設定に関する情報を表示できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. 表示するナレッジベースを選択します。

   定義されている場合は、現在使用されているカンバセーションスタータが右側のペインに表示されます。

4. ナレッジベースの詳細を表示するには、を選択し ••• て*［ナレッジベースの管理］*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

   [アクション]メニューを使用すると、変更を加える場合にナレッジベースを管理できます。

ナレッジベースの編集

一部の設定を変更してナレッジベースを更新したり、データソースを追加または削除したりできます。

ナレッジベースにデータソースを追加、変更、または削除するたびに、データソースを同期してナレッジベースに再インデックス化する必要があります。同期は差分で行われるため、Amazon Bedrockは前回の同期後に追加、変更、または削除されたFSx for ONTAPボリューム内のオブジェクトのみを処理します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、更新するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

5. メニューを選択し、[ナレッジベースの編集]*を選択します。
6. [ナレッジ ベースの編集] ページでは、ナレッジ ベースの名前、説明、埋め込みモデル、チャット モデル、機能の有効化を変更したり、会話の開始点を自動で作成するか手動で作成するかを選択したり、ナレッジ ベースを含むボリュームに使用するスナップショット ポリシーを選択したりできます。

   会話の開始に手動モードを使用している場合は、ここでも会話の開始を変更できます。

   埋め込み、コストを含むすべてのナレッジベーススキャン。ナレッジベースの作成後にデータガードレールを有効にすると、ナレッジベースは再度スキャンされ、コストがかかります。同様に、チャットモデルを変更すると、生成AIは関連するデータソースを再スキャンします(コストが発生します)。

7. 変更が完了したら、*[保存]*を選択します。

スナップショットでナレッジベースを保護

ナレッジベースボリュームのスナップショットを作成および復元することで、ナレッジベースデータを保護できます。スナップショットから復元して、ナレッジベースの以前のバージョンにいつでも戻すことができます。

スナップショットは、バックアップよりも高速でストレージ効率に優れており、異なる保護ポリシーを使用して各ナレッジベースを保護できます。スナップショットが役立つシナリオには、次のようなものがあります。

- 偶発的なデータ損失や破損
- ナレッジベースに取り込まれた不正なデータからのリカバリ
- さまざまなデータソースまたはチャンク戦略をテストし、テストが完了したら迅速にリバートする

ナレッジベースボリュームのスナップショットの作成

ナレッジベースボリュームのスナップショットを手動で作成することで、ナレッジベースの状態を保存できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、保護するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

5. メニューを選択し、[Snapshot]>[Create new snapshot]*を選択します。
6. 必要に応じて、*[Define snapshot name]*を選択し、Snapshotのカスタム名を入力します。

   カスタム名を定義すると、将来スナップショットをリストアする必要がある場合に、スナップショットの

内容をより正確に判断するのに役立ちます。

7. 「 * Create * 」を選択します。

   ナレッジベースのスナップショットが作成されます。

ナレッジベースボリュームのスナップショットのリストア

ナレッジベースボリュームの手動またはスケジュールされたスナップショットは、いつでもリストアできます。

> ボリュームに格納されているデータベースが破損しているか削除されている場合は、Generative AI Workloads UIを使用してSnapshotをリストアすることはできません。対処方法として、ボリュームがホストされているONTAPクラスタでを使用してSnapshotをリストアし"ONTAP CLI"ます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、復元するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

5. メニューを選択し、[スナップショット]>[スナップショットのリストア]*を選択します。

   スナップショット選択ダイアログが表示され、このナレッジベース用に作成されたスナップショットのリストが表示されます。

6. （オプション）Snapshotのリストア後にスケジュール済みおよび現在実行中のデータソーススキャンを続行する場合は、*[Pause running and scheduled scans after restoring the snapshot]*オプションの選択を解除します。

   このオプションはデフォルトで有効になっており、ナレッジベースが部分的に復元された状態のときにスキャンが実行されないようにしたり、新しく復元されたナレッジベースが古いデータで更新されないようにします。

7. リストアするSnapshotをリストから選択します。
8. [* Restore] を選択します。

ナレッジベースの複製

ナレッジベーススナップショットから新しいナレッジベースを作成できます。これは、元のナレッジベースが破損したり失われたりした場合に便利です。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。

3. ナレッジベースインベントリページで、復元するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

5. メニューを選択し、[スナップショット]>[ナレッジベースのクローニング]*を選択します。

   クローンダイアログが表示されます。

6. 必要に応じて、Snapshotのクローニング後にスケジュール済みおよび現在実行中のデータソーススキャンを続行する場合は、*[Pause running and scheduled scans after cloning the snapshot]*オプションの選択を解除します。

   このオプションはデフォルトで有効になっており、ナレッジベースが部分的に復元された状態のときにスキャンが実行されないようにしたり、新しく復元されたナレッジベースが古いデータで更新されないようにします。

7. クローニングするSnapshotをリストから選択します。
8. 「* Continue *」を選択します。
9. 新しいナレッジベースの名前を入力します。
10. 新しいナレッジベースのファイルシステムSVMとボリューム名を選択します。
11. 「* Clone *」を選択します。

ナレッジベースへのデータソースの追加

追加のデータソースをナレッジベースに埋め込んで、追加の組織データをナレッジベースに取り込むことができます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジベース]インベントリページで、データソースを追加するナレッジベースを選択します。
4. を選択 ••• し、*[Add data source]*を選択します。
5. 追加するデータ ソースの種類を選択します。
   - FSx for ONTAP ファイルシステムを追加する (既存の FSx for ONTAP ボリュームのファイルを使用)
   - ファイルシステムを追加する（一般的な SMB または NFS 共有のファイルを使用）

FSx for ONTAP ファイルシステムを追加する

1. ファイルシステムを選択：データソースファイルが存在するFSx for ONTAPファイルシステムを選択し、* Next *を選択します。
2. ボリュームを選択：データソースファイルが格納されているボリュームを選択し、*[次へ]*を選択します。

   SMBプロトコルを使用して保存されているファイルを選択する場合は、ドメイン、IPアドレス、ユーザ名、パスワードなどのActive Directory情報を入力する必要があります。

3. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。
4. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。

   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。

     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。

     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

     画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

     画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

   - [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

5. 権限対応*セクション（選択したデータソースがSMBプロトコルを使用するボリューム上にある場合にのみ表示）で、権限対応の応答を有効または無効にできます。
   - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
   - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
6. [追加]*を選択して、このデータソースをナレッジベースに追加します。

汎用**NFS**ファイルシステムを追加する

1. ファイル システムを選択: データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力し、ネットワーク共有の NFS プロトコルを選択して、次へ を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

場合によっては、NFSエクスポート名を手動で入力し、「ディレクトリを取得」を選択して利用可能なディレクトリを表示する必要があります。エクスポート全体を選択するか、エクスポートから特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。
     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

> ⓘ 画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

> ⓘ 変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

汎用**SMB**ファイルシステムを追加する

1. ファイルシステムを選択:
   a. データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力します。
   b. ネットワーク共有の SMB プロトコルを選択します。
   c. ドメイン、IP アドレス、ユーザー名、パスワードを含む Active Directory 情報を入力します。
   d. 「 * 次へ * 」を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

> ⓘ 場合によっては、SMB共有名を手動で入力し、「ディレクトリの取得」を選択して利用可能なディレクトリを表示する必要があります。共有全体を選択するか、共有内の特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

- 権限認識: 権限認識応答を有効または無効にします。
  - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
  - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
- ファイルフィルタリング:スキャンに含めるファイルを設定します。
  - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

    画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

  画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

  画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

  変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

結果

データソースはナレッジベースに統合されます。

データソースとナレッジベースを同期する

データソースは関連付けられたナレッジベースと1日に1回自動的に同期されるため、データソースの変更がチャットボットに反映されます。いずれかのデータソースに変更を加え、データをすぐに同期する場合は、オンデマンド同期を実行できます。

同期は増分処理であるため、Amazon Bedrockは前回の同期以降に追加、変更、または削除されたデータソース内のオブジェクトのみを処理します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、同期するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。
5. メニューを選択し、[今すぐスキャン]*を選択します。

   データソースがスキャンされていることを示すメッセージが表示され、スキャンが完了すると最終的なメッセージが表示されます。

結果

ナレッジベースは添付されたデータソースと同期され、アクティブなチャットボットはデータソースからの最新情報を使用し始めます。

スケジュールされた同期を一時停止または再開する

データソースの次の同期（スキャン）を一時停止または再開する場合は、いつでも実行できます。データソースに変更を加え、変更ウィンドウ中に同期を行わないようにする場合は、次のスケジュールされた同期を一時停止しなければならないことがあります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. 「ナレッジ ベースとコネクタ」メニューから、スキャンを一時停止または再開するナレッジ ベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。
5. [Actions]メニューを選択し、[Scan]>[Pause scheduled scan]*または[Scan]>[Resume scheduled scan]*を選択します。

   次のスケジュールされたスキャンが一時停止または再開されたことを示すメッセージが表示されます。

ナレッジベースを作成する前にチャットモデルを評価する

ナレッジベースを作成する前に、利用可能な基本チャットモデルを評価して、実装に最適なモデルを確認できます。サポートされるモデルはAWSのリージョンによって異なるため、ナレッジベースを導入するリージョンで使用できるモデルを確認するには、を参照し "AWSのドキュメントページ" てください。

> この機能は、ナレッジベースが作成されていない場合（ナレッジベースインベントリページにナレッジベースが存在しない場合）にのみ使用できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースのインベントリページから、チャットボットのページの右側にチャットモデルを選択する

オプションが表示されます。

4. リストからチャットモデルを選択し、プロンプト領域に質問のセットを入力して、チャットボットがどのように応答するかを確認します。
5. 複数のモデルを試して、実装に最適なモデルを確認してください。

結果

ナレッジベースを作成するときは、そのチャットモデルを使用します。

ナレッジベースの非公開

ナレッジベースを公開してチャットボットアプリケーションと統合できるようにした後、チャットボットアプリケーションからナレッジベースへのアクセスを無効にする場合は、ナレッジベースを非公開にすることができます。

ナレッジベースを非公開にすると、チャットアプリケーションの動作が停止します。ナレッジベースにアクセスできた一意のAPIエンドポイントが無効になります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジベース]インベントリページで、非公開にするナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、およびすべての埋め込みデータソースのリストが表示されます。

5. [アクション（Actions）]メニューを選択し、*[パブリッシュ解除（Unpublish）]*を選択

結果

ナレッジベースが無効になり、チャットボットアプリケーションからアクセスできなくなります。

ナレッジベースの削除

ナレッジ ベースが不要になった場合は、削除できます。ナレッジ ベースを削除すると、そのナレッジ ベースは Workload Factory から削除され、そのナレッジ ベースを含むボリュームも削除されます。ナレッジベースを使用しているアプリケーションやチャットボットはすべて動作を停止します。ナレッジベースの削除は元に戻せません。

ナレッジベースを削除する場合は、ナレッジベースに関連付けられているすべてのリソースを完全に削除するために、ナレッジベースと関連付けられているエージェントの関連付けも解除する必要があります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、削除するナレッジベースを選択します。
4. を選択し ••• 、*[ナレッジベースの管理]*を選択します。

5. [アクション]メニューを選択し、*[ナレッジベースの削除]*を選択します。
6. [ナレッジベースの削除]ダイアログで、削除することを確認し、*[削除]*を選択します。

結果

ナレッジ ベースは Workload Factory から削除され、それに関連付けられたボリュームも削除されます。

Amazon Q Business Connectorの管理

Amazon Q Business用のコネクタを作成したら、コネクタの詳細の表示、コネクタの変更、追加データソースの統合、コネクタの削除を行うことができます。

コネクタに関する情報を表示する

コネクタと統合されているデータソースの設定に関する情報を表示できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジ ベースとコネクタ インベントリ ページから、表示するコネクタを選択します。
4. コネクターの詳細を表示するには、を選択し…て*コネクターの管理*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

   [アクション（Actions）]*メニューを使用すると、変更を加える場合にコネクターを管理できます。

コネクタの編集

一部の設定を変更してコネクタを更新したり、データソースを追加または削除したりできます。

コネクタからデータソースを追加、変更、または削除するたびに、生成AIはデータソース情報をAmazon Q Businessに送信してインデックスを再作成する必要があります。同期は増分処理であるため、Amazon Q Businessは前回の同期後に追加、変更、または削除されたFSx for ONTAPボリューム内のオブジェクトのみを処理します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジベースとコネクタ]インベントリページで、更新するコネクタを選択します。
4. を選択し…、*コネクターの管理*を選択します。

   このページには、パブリッシュされたステータス、データソースの埋め込みステータス、埋め込みモード、すべての埋め込みデータソースのリストなどが表示されます。

5. [アクション]メニューを選択し、*[コネクタの編集]*を選択します。

6. [Edit Connector]ページでは、コネクタの名前、説明、埋め込みモデル、データガードレールの有効化、およびコネクタを含むボリュームで使用されるSnapshotポリシーを変更できます。

   埋め込みを含むすべてのデータソーススキャンにはコストがかかります。コネクタの作成後にデータガードレールを有効にすると、データソースが再度スキャンされ、コストが発生します。

7. 変更を行ったら、*[保存]*を選択します。

コネクタへの追加データソースの追加

追加データソースをコネクタに埋め込んで、追加の組織データをコネクタに取り込むことができます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジベースとコネクタ]インベントリページで、データソースを追加するコネクタを選択します。
4. を選択 ••• し、*[Add data source]*を選択します。
5. 追加するデータ ソースの種類を選択します。
   - FSx for ONTAP ファイルシステムを追加する (既存の FSx for ONTAP ボリュームのファイルを使用)
   - ファイルシステムを追加する（一般的な SMB または NFS 共有のファイルを使用）

FSx for ONTAP ファイルシステムを追加する

1. ファイルシステムを選択：データソースファイルが存在するFSx for ONTAPファイルシステムを選択し、* Next *を選択します。
2. ボリュームを選択：データソースファイルが格納されているボリュームを選択し、*[次へ]*を選択します。

   SMBプロトコルを使用して保存されているファイルを選択する場合は、ドメイン、IPアドレス、ユーザ名、パスワードなどのActive Directory情報を入力する必要があります。

3. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。
4. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。

   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。

     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。

     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

     画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

     画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

   - [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

5. 権限対応*セクション（選択したデータソースがSMBプロトコルを使用するボリューム上にある場合にのみ表示）で、権限対応の応答を有効または無効にできます。
   - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
   - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
6. [追加]*を選択して、このデータソースをナレッジベースに追加します。

汎用**NFS**ファイルシステムを追加する

1. ファイル システムを選択: データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力し、ネットワーク共有の NFS プロトコルを選択して、次へ を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

場合によっては、NFSエクスポート名を手動で入力し、「ディレクトリを取得」を選択して利用可能なディレクトリを表示する必要があります。エクスポート全体を選択するか、エクスポートから特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

       オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

   - ファイルフィルタリング:スキャンに含めるファイルを設定します。
     - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

       画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

> 画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

> 変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

汎用**SMB**ファイルシステムを追加する

1. ファイルシステムを選択:
   a. データ ソース ファイルが存在するファイル システム ホストの IP アドレスまたは FQDN を入力します。
   b. ネットワーク共有の SMB プロトコルを選択します。
   c. ドメイン、IP アドレス、ユーザー名、パスワードを含む Active Directory 情報を入力します。
   d. 「 * 次へ * 」を選択します。
2. データソースを選択：ファイルを保存した場所に基づいてデータソースの場所を選択します。これは、ボリューム全体、またはボリューム内の特定のフォルダまたはサブフォルダにすることができ、* Next *を選択します。

> 場合によっては、SMB共有名を手動で入力し、「ディレクトリの取得」を選択して利用可能なディレクトリを表示する必要があります。共有全体を選択するか、共有内の特定のフォルダのみを選択するかを選択できます。

3. 設定:データソースがファイルから情報を取り込む方法と、スキャンに含めるファイルを設定します。
   - データソースの定義：*チャンク戦略*セクションで、データソースがナレッジベースと統合されている場合に、生成AIエンジンがデータソースのコンテンツをチャンクに分割する方法を定義します。次のいずれかの方法を選択できます。
     - **Multi-sentence chunking**:データソースの情報をセンテンス定義のチャンクに編成します。各チャンクを構成する文の数を選択できます(最大100)。
     - オーバーラップベースのチャンク:データソースからの情報を文字定義のチャンクに編成し、隣接するチャンクとオーバーラップすることができます。各チャンクのサイズを文字単位で選択し、各チャンクが隣接するチャンクとどの程度重なるかを選択できます。チャンクサイズは50～3000文字、オーバーラップパーセンテージは1～99%の範囲で設定できます。

オーバーラップ率を高く設定すると、取得精度がわずかに向上するだけで、ストレージ要件が大幅に増加します。

- 権限認識: 権限認識応答を有効または無効にします。
  - 有効:このナレッジベースにアクセスするチャットボットのユーザーは、アクセス権を持つデータソースからのクエリに対する応答のみを取得します。
  - 無効:チャットボットのユーザーは、統合されたすべてのデータソースからコンテンツを使用して応答を受信します。
- ファイルフィルタリング:スキャンに含めるファイルを設定します。
  - [ファイルタイプのサポート]セクションで、すべてのタイプのファイルを含めるか、データソーススキャンに含めるファイルタイプを個別に選択します。

    画像または PDF ファイルを含めると、 NetApp Workload Factory for GenAI は画像内のテキスト (PDF ドキュメント内の画像を含む) を解析するため、コストが高くなります。

  画像のテキストデータを含めると、スキャンされたテキストデータが環境からAWSに送信されるため、生成AIは画像の個人識別情報(PII)をマスクできません。ただし、データが保存されると、すべてのPIIは 生成AIデータベースでマスクされます。

  画像ファイルをスキャンに含めるかどうかは、ナレッジベースチャットモデルに関連しています。画像ファイルをスキャンに含める場合は、チャットモデルで画像がサポートされている必要があります。ここで画像ファイルタイプが選択されている場合、画像ファイルをサポートしていないチャットモデルにナレッジベースを切り替えることはできません。

- [ファイル変更時刻フィルタ]*セクションで、ファイルの変更時刻に基づいてファイルを含めるかどうかを選択します。変更時刻のフィルタリングを有効にする場合は、リストから日付範囲を選択します。

  変更日の範囲に基づいてファイルをインクルードした場合、日付範囲が満たされない（指定した日付範囲内でファイルが変更されていない）とすぐに、ファイルは定期スキャンから除外され、データソースにはこれらのファイルは含まれません。

4. このデータ ソースをナレッジ ベースに追加するには、[データ ソースの追加] を選択します。

結果

データソースがコネクタに統合されます。

データソースをコネクタと同期する

データソースは関連付けられているコネクタと1日に1回自動的に同期されるため、データソースの変更がAmazon Q Businessに反映されます。いずれかのデータソースに変更を加え、データをすぐに同期（スキャン）する場合は、オンデマンド同期を実行できます。

同期は増分処理であるため、Amazon Q Businessは前回の同期以降に追加、変更、または削除されたデータソース内のオブジェクトのみを処理します。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. 「ナレッジベースとコネクタ」メニューから、同期するコネクタを選択します。
4. を選択し、*コネクターの管理*を選択します。
5. メニューを選択し、[今すぐスキャン]*を選択します。

   データソースがスキャンされていることを示すメッセージが表示され、スキャンが完了すると最終的なメッセージが表示されます。

結果

コネクタは接続されているデータソースと同期され、Amazon Q Businessはデータソースからの最新情報を使用し始めます。

スケジュールされた同期を一時停止または再開する

データソースの次の同期（スキャン）を一時停止または再開する場合は、いつでも実行できます。データソースに変更を加え、変更ウィンドウ中に同期を行わないようにする場合は、次のスケジュールされた同期を一時停止しなければならないことがあります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [Connector]インベントリページで、スキャンを一時停止または再開するコネクタを選択します。
4. を選択し、*コネクターの管理*を選択します。
5. [Actions]メニューを選択し、[Scan]>[Pause scheduled scan]*または[Scan]>[Resume scheduled scan]*を選択します。

   次のスケジュールされたスキャンが一時停止または再開されたことを示すメッセージが表示されます。

コネクタの削除

コネクタが不要になった場合は削除できます。コネクタを削除すると、そのコネクタは Workload Factory から削除され、そのコネクタを含むボリュームも削除されます。コネクタを削除すると元に戻すことはできません。

コネクタを削除する場合は、コネクタに関連付けられているすべてのエージェントからコネクタの関連付けを解除して、コネクタに関連付けられているすべてのリソースを完全に削除する必要があります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. [ナレッジベースとコネクタ]インベントリページで、削除するコネクタを選択します。
4. を選択し、*コネクターの管理*を選択します。

5. [アクション（Actions）]メニューを選択し、[コネクタの削除（Delete Connector）]を選択します。
6. [コネクタの削除（Delete Connector）]ダイアログで、削除することを確認し、*[削除（Delete）]*を選択します。

結果

コネクタはワークロード ファクトリーから削除され、関連付けられているボリュームも削除されます。

生成AIデータソースを管理します。

FSx for ONTAPファイルシステムでデータソースを使用してナレッジベースまたはコネクタを作成したら、データソースの詳細の表示、データソースの内容の更新または変更、データソースの設定の編集、データソースの削除を行うことができます。

データソースに関する情報を表示する

データソースのコンテンツに関する情報を表示したり、ナレッジベースまたはコネクタを使用してデータソースの埋め込みステータスを表示したりできます。データソースはナレッジベースまたはコネクタに関連付けられているため、データソースの詳細を表示するには、まずナレッジベースまたはコネクタを選択する必要があります。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。 "コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. データソースが存在するナレッジベースまたはコネクタを選択し、**[ナレッジベースの管理]*または[コネクタの管理]*を選択します。

   ページの下部には、関連するデータソースが表示されます。

4. を選択して各行を展開し、 FSx for ONTAPファイルシステム、ボリューム、データソースが配置されているパスなど、各データソースに関する詳細情報を表示します。

   また、埋め込み情報と、そのデータソースが現在ナレッジベースまたはコネクタに埋め込まれているかどうかも表示されます。

データソース設定の編集

ナレッジベースまたはコネクタと統合したデータソースに関する情報を編集できます。ほとんどの情報は、データソースを追加した後で修正されますが、一部の構成（チャンキング定義やパーミッションアウェアネスなど）に変更を加えることができます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースインベントリページで、データソースが存在するナレッジベースを選択し、 *[ナレッジベースの管理]*を選択します。

   ページの下部には、このナレッジベースの一部であるデータソースが一覧表示されます。

4. 編集するデータソースの行で、を選択し ••• て*[データソースの編集]*を選択します。
5. [Edit data source]ページで、を選択し ⌄ てチャンク定義の行を展開します。
6. チャンキング戦略と構成、および権限の認識（SMBボリュームの場合）の設定を更新し、*[保存]*を選択します。

結果

データソースの設定が更新され、AIシステムがデータソースを同期して、ナレッジベースにインデックスが再作成されます。

既存のデータソースの内容を更新する

データソースの内容はいつでも変更して、組織のデータを追加または更新できます。このデータソースがナレッジベースでアクティブに使用されている場合は、ナレッジベースに再インデックスされるようにデータソースを同期する必要があります。同期は差分で行われるため、Amazon Bedrockは前回の同期後に追加、変更、または削除されたFSx for ONTAPボリューム内のオブジェクトのみを処理します。

データソースは1日に1回自動的にナレッジベースと同期されるため、データソースの変更がチャットボットに反映されます。データソースに変更を加え、データをすぐに同期したい場合は、を使用できます "オンデマンドで同期を実行する"。

データソースを削除する

データソースをナレッジベースに含める必要がなくなった場合は、そのデータソースを削除できます。

手順

1. 次のいずれかを使用してWorkload Factoryにログインします。"コンソールエクスペリエンス" 。
2. [AI Workloads]タイルで、*[Deploy & manage]*を選択します。
3. ナレッジベースのインベントリページで、データソースが存在するナレッジベースを選択し、 ••• *[ナレッジベースの管理]*を選択します。

   ページの下部には、このナレッジベースの一部であるデータソースが一覧表示されます。
4. 削除するデータソースの行で、を選択し ••• て*[データソースの削除]*を選択します。
5. [データソースの削除]ダイアログで、削除することを確認し、*[確認]*を選択します。

結果

データソースがナレッジベースから削除され、このデータソースに関するインデックス化された情報がAIシステムによってナレッジベースから削除されます。そのデータソースからの情報は、ナレッジベースを使用しているチャットボットでは利用できなくなります。

NetApp Workload FactoryのTrackerでワークロード操作を監視する

NetApp Workload Factory の Tracker を使用して、ワークロード操作の実行を監視および追跡し、タスクの進行状況を監視します。

タスクの内容

NetApp Workload Factory には監視機能の Tracker が用意されており、ワークロード操作の進行状況とステータスを監視および追跡したり、操作タスクとサブタスクの詳細を確認したり、問題や障害を診断したりできます。

Trackerでは、いくつかのアクションを使用できます。期間（過去24時間、7日、14日、または30日）、ワークロード、ステータス、およびユーザでジョブをフィルタリングしたり、検索機能を使用してジョブを検索したり、ジョブテーブルをCSVファイルとしてダウンロードしたりできます。Trackerはいつでも更新でき、失敗した処理をすばやく再試行したり、失敗した処理のパラメータを編集して処理を再試行したりできます。

トラッカーは、操作に応じて2つのレベルの監視をサポートします。ファイルシステムの導入などの各タスクには、タスクの説明、ステータス、開始時間、タスク期間、ユーザー、地域、プロキシリソース、タスクID、および関連するすべてのサブタスクが表示されます。API応答を表示して、処理中に発生した問題を把握することができます。

Trackerタスクレベルと例

- レベル1（タスク）：ファイルシステムの展開を追跡します。
- レベル2（サブタスク）：ファイルシステムの展開に関連するサブタスクを追跡します。

処理のステータス

Trackerの動作ステータスは、進行中、成功、_失敗_です。

動作周波数

処理の頻度は、ジョブタイプとジョブスケジュールに基づいて決まります。

イヘントホシ

イベントはユーザインターフェイスに30日間保持されます。

運用の追跡と監視

Tracker を使用して、Workload Factory コンソールで操作を追跡および監視します。

手順

1. いずれかを使用してログインし"コンソールエクスペリエンス"ます。
2. ワークロード メニューから、管理 を選択し、次に トラッカー を選択します。
3. Tracker では、フィルターまたは検索を使用してジョブの結果を絞り込みます。求人レポートをダウンロードすることもできます。

View APIヨウキユウ

TrackerのタスクのコードボックスでAPIリクエストを表示します。

手順

1. トラッカーでタスクを選択します。
2. アクション メニューを選択し、*API リクエストの表示*を選択します。

失敗した処理を再試行する

Trackerで失敗した操作を再試行します。失敗した処理のエラーメッセージをコピーすることもできます。

失敗した処理は10回まで再試行できます。

手順

1. トラッカーで、失敗した操作を選択します。
2. アクション メニューを選択し、[再試行] を選択します。

結果

処理が再開されます。

失敗した処理を編集して再試行してください

失敗した操作のパラメータを編集し、Trackerの外部で操作を再試行してください。

手順

1. トラッカーで、失敗した操作を選択します。
2. アクション メニューを選択し、*編集して再試行*を選択します。

   処理ページにリダイレクトされ、パラメータを編集して処理を再試行できます。

結果

処理が再開されます。[Tracker]に移動して処理のステータスを確認します。

知識とサポート

NetApp Workload Factory for GenAIのサポートに登録する

NetApp Workload Factory とそのストレージ ソリューションおよびサービスに固有のテクニカル サポートを受けるには、サポート登録が必要です。Workload Factory とは別の Web ベースのコンソールであるNetAppコンソールからサポートに登録する必要があります。

サポートに登録しても、クラウド プロバイダー ファイル サービスに対するNetAppサポートは有効になりません。クラウド プロバイダーのファイル サービス、そのインフラストラクチャ、またはサービスを使用するソリューションに関連するテクニカル サポートについては、その製品の Workload Factory ドキュメントの「ヘルプの取得」を参照してください。

"Amazon FSx for ONTAP"

サポート登録の概要

アカウント ID サポート サブスクリプション (NetAppコンソールの [サポート リソース] ページにある 20 桁の 960xxxxxxxx シリアル番号) を登録すると、単一のサポート サブスクリプション ID として機能します。各NetAppアカウント レベルのサポート サブスクリプションを登録する必要があります。

登録すると、サポート チケットの開設やケースの自動生成などの機能が有効になります。登録は、以下の説明に従ってNetAppコンソールにNetAppサポート サイト (NSS) アカウントを追加することで完了します。

NetAppサポートのアカウントを登録する

サポートに登録し、サポート資格を有効にするには、アカウント内の 1 人のユーザーがNetAppサポート サイト アカウントをNetAppコンソール ログインに関連付ける必要があります。NetAppサポートに登録する方法は、 NetAppサポート サイト (NSS) アカウントをすでにお持ちかどうかによって異なります。

NSSアカウントをお持ちの既存のお客様

NSS アカウントをお持ちのNetApp のお客様の場合は、 NetAppコンソールからサポートに登録するだけです。

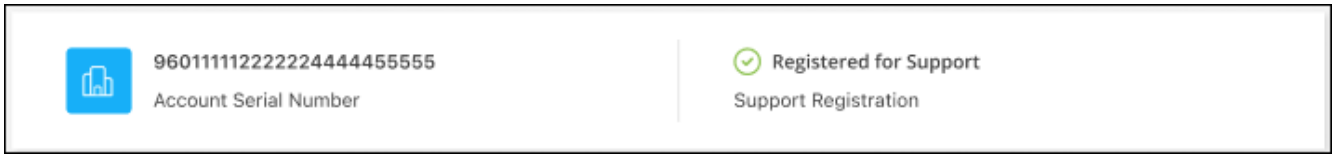
手順

1. Workload Factory コンソールの右上で、[ヘルプ] > [サポート] を選択します。

   このオプションを選択すると、新しいブラウザ タブでNetAppコンソールが開き、サポート ダッシュボードが読み込まれます。

2. NetAppコンソール メニューから、管理 を選択し、資格情報 を選択します。
3. [ユーザクレデンシャル]*を選択します。
4. [NSSクレデンシャルの追加]*を選択し、NetApp Support Site（NSS）認証プロンプトに従います。
5. 登録プロセスが正常に完了したことを確認するには、[ヘルプ]アイコンを選択し、*[サポート]*を選択します。

[リソース]ページに、アカウントがサポートに登録されていることが表示されます。



NetAppコンソール ログインにNetAppサポート サイト アカウントを関連づけていない場合、他のNetAppコンソール ユーザーには同じサポート登録ステータスが表示されません。ただし、これはNetAppアカウントがサポートに登録されていないことを意味するものではありません。アカウント内の 1 人のユーザーがこれらの手順を実行していれば、アカウントは登録済みになります。

NSSアカウントを持たない既存のお客様

既存のNetApp顧客であり、既存のライセンスとシリアル番号を持っているものの、NSS アカウントを持っていない場合は、NSS アカウントを作成し、それをNetAppコンソール ログインに関連付ける必要があります。

手順

1. NetAppサポートサイトのアカウントを作成するには、 "NetApp Support Site ユーザー登録フォーム"
   a. 適切なユーザレベルを選択してください。通常は*ネットアップのお客様/エンドユーザ*がこれに該当します。
   b. 上記で使用したNetAppアカウントのシリアル番号 (960xxxx) を必ずシリアル番号フィールドにコピーしてください。これにより、アカウント処理が高速化されます。
2. 以下の手順を実行して、新しいNSSアカウントをNetAppコンソールログインに関連付けます。NSSアカウントをお持ちの既存のお客様 。

ネットアップのソリューションを初めて導入する場合は

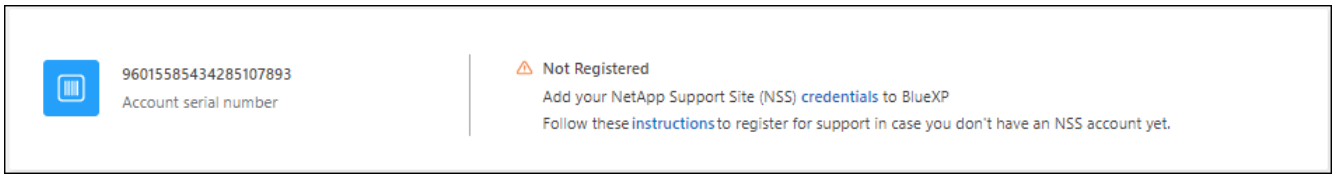
ネットアップ製品を初めてご利用になり、NSSアカウントをお持ちでない場合は、以下の手順に従ってください。

手順

1. Workload Factory コンソールの右上で、[ヘルプ] > [サポート] を選択します。

   このオプションを選択すると、新しいブラウザ タブでNetAppコンソールが開き、サポート ダッシュボードが読み込まれます。

2. [Support Resources]ページでアカウントIDのシリアル番号を確認します。



メニューのスクリーンショット。サポートは最初に表示されるオプションです"]

3. [I am not a registered NetApp Customer]に移動して "ネットアップサポート登録サイト" 選択します。
4. 必須フィールドに入力します（赤いアスタリスクのフィールド）。
5. [製品ライン（Product Line ）]フィールドで、[Cloud Manager *]を選択し、該当する課金プロバイダーを

選択します。

6. 上記の手順2からアカウントのシリアル番号をコピーし、セキュリティチェックを完了して、ネットアップのグローバルデータプライバシーポリシーを確認します。

   この安全なトランザクションを完了するために、メールボックスに電子メールがすぐに送信されます。確認メールが数分で届かない場合は、必ずスパムフォルダを確認してください。

7. Eメールからアクションを確認します。

   確認ではネットアップにリクエストが送信され、NetApp Support Site アカウントを作成することを推奨します。

8. NetAppサポートサイトのアカウントを作成するには、 "NetApp Support Site ユーザー登録フォーム"

   a. 適切なユーザレベルを選択してください。通常は*ネットアップのお客様/エンドユーザ*がこれに該当します。

   b. シリアル番号フィールドには、上記のアカウントのシリアル番号（960xxxx）を必ずコピーしてください。これにより、アカウント処理が高速化されます。

終了後

このプロセスについては、ネットアップからご連絡ください。これは、新規ユーザ向けの1回限りのオンボーディング演習です。

NetAppサポートサイトのアカウントを取得したら、以下の手順を実行して、アカウントをNetAppコンソールのログインに関連付けます。NSSアカウントをお持ちの既存のお客様 。

生成AIのトラブルシューティング

発生する可能性のある一般的な問題を回避する方法について説明します。

一般的な問題と解決策

これらのいずれかの問題がある場合は、[Workaround]列の手順を使用して解決できます。

面積	問題	原因	回避策
導入	ボリュームがすでに存在するため導入に失敗します。	NetApp Workload Factory for GenAI は、デプロイメント プロセス中に新しいボリュームを作成する必要がありますが、指定した名前を使用したボリュームがすでに存在します。	新しいボリュームに使用する一意の名前を指定してから、導入を再試行してください。
導入	NetApp Workload Factory for GenAI がボリュームをマウントできないため、デプロイメントは失敗します。	FSx for NetApp ONTAPに必要な1つ以上のインバウンドポートが閉じているか、フィルタリングされています。	次のインバウンドポートを開きます。

| プロトコル | ポート | 目的

| すべての ICMP | すべて | インスタンスの ping を実行します

| HTTPS | 443 | fsxadmin管理LIFへのコネクタからアクセスし、API呼び出しをFSXに送信します

| SSH | 22 | クラスタ管理 LIF またはノード管理 LIF の IP アドレスへの SSH アクセス

| TCP | 111 | NFS のリモートプロシージャコール

| TCP | 139 | CIFS の NetBIOS サービスセッション

| TCP | 161-162 | 簡易ネットワーク管理プロトコル

| TCP | 445 | NetBIOS フレーム同期を使用した Microsoft SMB over TCP

| TCP | 635 | NFSマウント

| TCP | 749 | Kerberos

| TCP | 2049 | NFSサーバデーモン

| TCP | 3260 | iSCSI データ LIF を介した iSCSI アクセス

| TCP | 4045 | NFSロックデーモン

| TCP | 4046 | NFS のネットワークステータスモニタ

| TCP | 10000 | NDMP を使用したバックアップ

| TCP | 11104 | SnapMirror のクラスタ間通信セッションの管理

| TCP | 11105 | クラスタ間 LIF を使用した SnapMirror データ転送

| UDP | 111 | NFS のリモートプロシージャコール

| UDP | 161-162 | 簡易ネットワーク管理プロトコル

| UDP | 635 | NFSマウント

| UDP | 2049 | NFSサーバデーモン

| UDP | 4045 | NFSロックデーモン

| UDP | 4046 | NFS のネットワークステータスモニタ

| UDP | 4049 | NFS rquotadプロトコル

メンテナンス	AIエンジンが起動せず、*ナレッジベース*ページに「AIエンジンインスタンスエラー」というエラーが表示されます。	AIエンジンインスタンスが破損しているか、存在しません。	*再構築*ボタンを選択します。 NetApp Workload Factory for GenAI はインフラストラクチャを再構築し、再構築の進行状況を表示します。完了すると、ナレッジ ベースが再構築されたインフラストラクチャに再接続され、ナレッジ ベースのリストが表示されます。
メンテナンス	AIエンジンが起動せず、*ナレッジベース*ページに「The生成AI engine instance is stopped」というエラーが表示されます。	AIエンジンインスタンスが実行されていません。	AWS Management ConsoleまたはAWS CLIを使用して、AIエンジンインスタンスを起動します。
メンテナンス	AIエンジンが起動せず、*ナレッジベース*ページに「The生成AI engine server is not responding」というエラーが表示されます。	AIエンジンインスタンスが応答していません。	次のリカバリ手順を実行します。 手順 1. 生成AIエンジンインスタンスのセキュリティグループを変更して、生成AIエンジンインスタンスへのSSHアクセスを有効にします。 2. SSHを使用してインスタンスにログインします。 3. 次のコマンドを実行します。 `docker-compose up`

メンテナンス	NetApp Workload Factory for GenAI で使用されるバックエンド Docker インスタンスの起動に失敗しました。	ボリュームが削除され、EC2インスタンスが再起動されました。	次のリカバリ手順を実行します。
			
			手順
			1. FSx for NetApp ONTAP上に新しいボリュームを作成します。たとえば `netapp_ai`、ボリューム名は、ボリュームパスはになります `/netapp_ai`。
			2. Amazon EC2インスタンスにSSHで接続します。
			3. ボリュームを表示します。
			`docker volume list`
			4. 古いボリュームを削除します。
			`docker volume rm ec2-user_persistent_folder`
			5. `docker-compose.yml`テキストエディタを使用してファイルを開きます。
			6. `volumes`セクションで、デバイスパスを新しいボリュームパスに変更します。例：

メンテナンス	NetApp Workload Factory for GenAI で使用されるバックエンド Docker インスタンスの起動に失敗しました。	ルートボリュームが削除されました。	名前とパスを指定してボリュームを作成し、Amazon EC2からバックエンドDockerインスタンスを再起動します。
メンテナンス	NetApp Workload Factory for GenAI で使用されるバックエンド Docker インスタンスの起動に失敗しました。	ルートボリュームが削除されました。	名前とパスを指定してボリュームを作成し、Amazon EC2からバックエンドDockerインスタンスを再起動します。

NetApp Workload Factory for GenAI のサポートを受ける

NetApp は、Workload Factory とそのクラウド サービスをさまざまな方法でサポートします。ナレッジベース (KB) 記事やコミュニティ フォーラムなど、広範な無料のセルフサポート オプションが 24 時間 365 日ご利用いただけます。サポート登録には、Webチケットによるリモートテクニカルサポートも含まれます。

FSx for ONTAPのサポートを利用する

FSx for ONTAP、そのインフラストラクチャ、またはサービスを使用するソリューションに関連するテクニカル サポートについては、その製品の Workload Factory ドキュメントの「ヘルプの取得」を参照してください。

"Amazon FSx for ONTAP"

Workload Factoryおよびそのストレージソリューションおよびサービスに固有のテクニカルサポートを受けるには、以下に説明するサポートオプションを使用してください。

セルフサポートオプションを使用します

次のオプションは、 1 日 24 時間、週 7 日間無料でご利用いただけます。

- ドキュメント

  現在表示している Workload Factory のドキュメント。

- "ナレッジベース"

  Workload Factory ナレッジベースを検索して、問題のトラブルシューティングに役立つ記事を見つけます。

- "コミュニティ"

  Workload Factory コミュニティに参加して、進行中のディスカッションをフォローしたり、新しいディスカッションを作成したりしてください。

```
"addr=svm-
0f166b85a890
d8a72.\
fsx.us-east-
1.amazonaws.
com,nolock,s
device:
':/netapp_ai
' # Path to
```

ネットアップサポートと一緒にケースを作成します

上記のセルフサポートオプションに加え、サポートを有効にしたあとで問題が発生した場合は、ネットアップサポートの担当者と相談して解決できます。

始める前に

*ケースの作成*機能を使用するには、まずサポートに登録する必要があります。NetAppNetAppサイトの資格情報を Workload Factory ログインに関連付けます。"サポートに登録する方法について説明します" 。

手順

1. Workload Factory コンソールの右上で、[ヘルプ] > [サポート] を選択します。

   このオプションを選択すると、新しいブラウザ タブでNetAppコンソールが開き、サポート ダッシュボードが読み込まれます。

2. [**Resources**]ページで、[Technical Support]で次のいずれかのオプションを選択します。

   a. 電話で誰かと話をしたい場合は、*[電話]*を選択します。netapp.comのページに移動し、電話番号が表示されます。

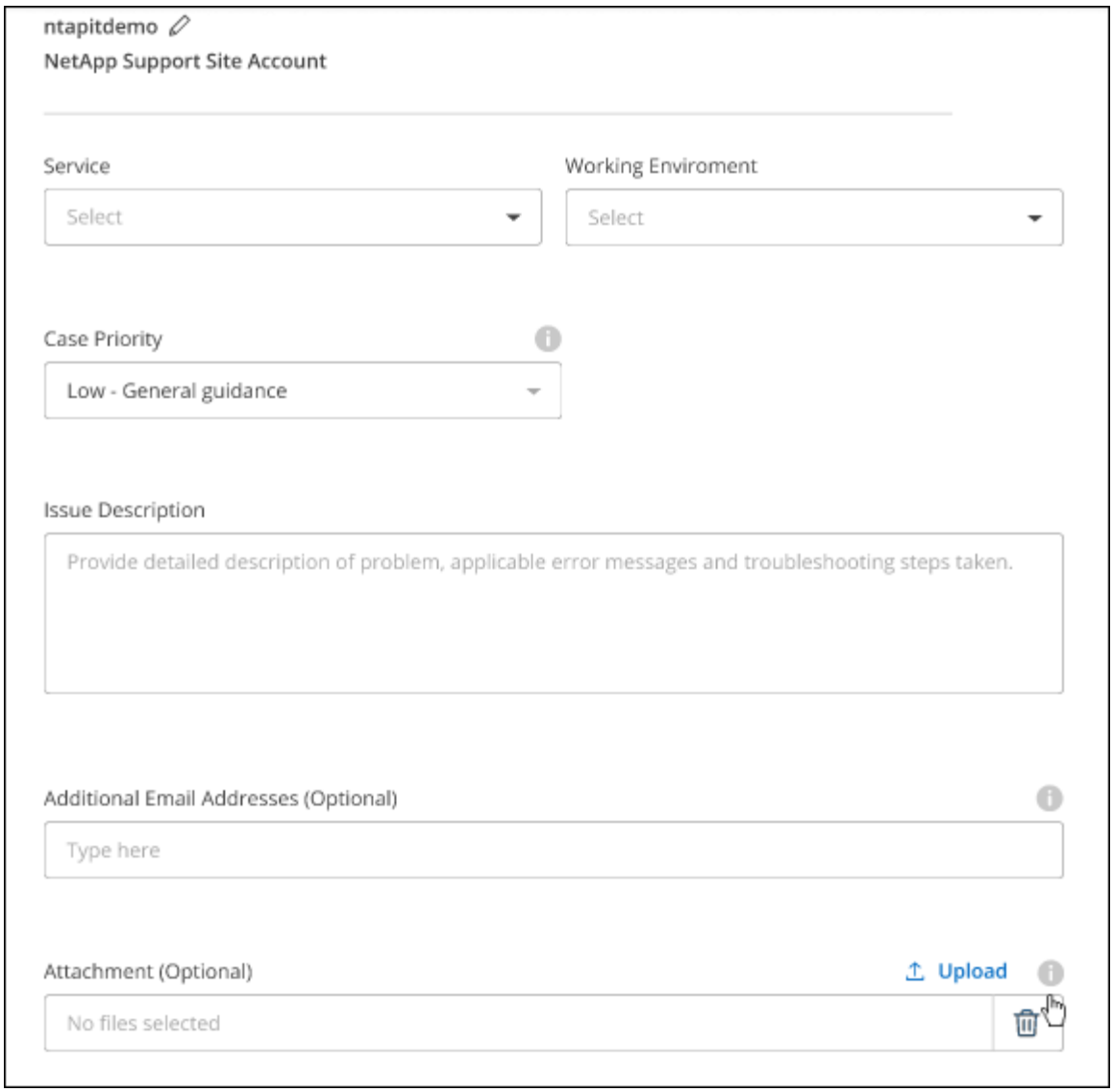
   b. [ケースの作成]*を選択して、NetAppサポートスペシャリストとのチケットをオープンします。

      - * Service ： Workload Factory *を選択します。
      - ケース優先度：ケースの優先度を選択します。優先度は、[低]、[中]、[高]、[クリティカル]のいずれかになります。

        これらの優先度の詳細を確認するには、フィールド名の横にある情報アイコンの上にマウスポインタを合わせます。

      - *事象の説明 *：実行したエラーメッセージやトラブルシューティング手順など、問題の詳細な概要 を入力します。
      - その他のメールアドレス：この問題 を他のユーザーに知らせる場合は、追加のメールアドレスを入力します。
      - 添付ファイル（オプション）：一度に1つずつ、最大5つの添付ファイルをアップロードできます。

        添付ファイルはファイルあたり25 MBに制限されています。サポートされているファイル拡張子は、txt、log、pdf、jpg/jpeg、rtf、 doc/docx、xls/xlsx、およびcsv。



終了後

ポップアップにサポートケース番号が表示されます。ネットアップのサポート担当者がケースを確認し、すぐに対応させていただきます。

サポートケースの履歴を確認するには、*[設定]>[タイムライン]*を選択し、「サポートケースの作成」というアクションを検索します。右端のボタンをクリックすると、アクションを展開して詳細を表示できます。

ケースを作成しようとすると、次のエラーメッセージが表示される場合があります。

"選択したサービスに対してケースを作成する権限がありません"

このエラーは、NSS アカウントとそれに関連付けられているレコード会社が、 NetAppコンソール アカウントのシリアル番号のレコード会社と同じではないことを意味している可能性があります (つまり、960xxxx) またはシステムのシリアル番号。次のいずれかのオプションを使用してサポートを求めることができます。

- 製品内のチャットを使用します
- テクニカル以外のケースを https://mysupport.netapp.com/site/help

サポートケースの管理（プレビュー）

アクティブなサポート ケースと解決済みのサポート ケースをNetAppコンソールから直接表示および管理できます。NSS アカウントおよび会社に関連付けられたケースを管理できます。

ケース管理はプレビューとして使用できます。今後のリリースでは、この点をさらに改良し、機能を強化する予定です。製品内のチャットでご意見をお寄せください。

次の点に注意してください。

- ページ上部のケース管理ダッシュボードには、次の2つのビューがあります。
  - 左側のビューには、指定したユーザNSSアカウントによって過去3カ月間にオープンされたケースの総数が表示されます。
  - 右側のビューには、ユーザのNSSアカウントに基づいて、過去3カ月間にオープンしたケースの総数が会社レベルで表示されます。

  テーブルの結果には、選択したビューに関連するケースが反映されます。
- 目的の列を追加または削除したり、[優先度]や[ステータス]などの列の内容をフィルタリングしたりできます。他の列には、並べ替え機能だけがあります。

  詳細については、以下の手順を参照してください。
- ケースごとに、ケースノートを更新したり、ステータスが「Closed」または「Pending Closed」でないケースをクローズしたりすることができます。

手順

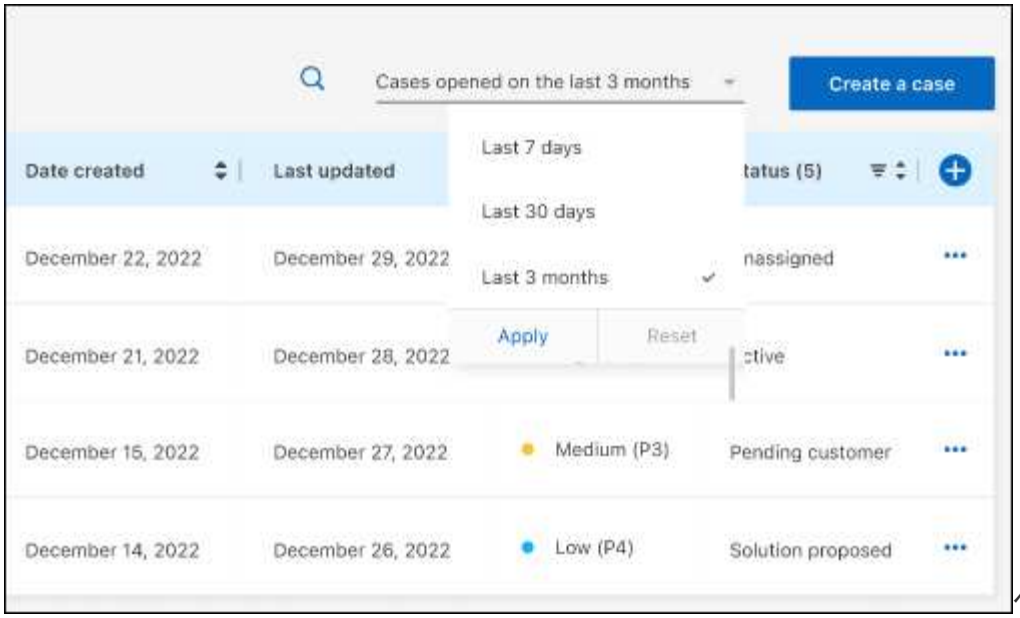
1. Workload Factory コンソールの右上で、[ヘルプ] > [サポート] を選択します。

   このオプションを選択すると、 NetAppコンソールの新しいブラウザ タブが開き、サポート ダッシュボードが読み込まれます。

2. *ケース管理*を選択し、プロンプトが表示されたら、NSS アカウントをNetAppコンソールに追加します。

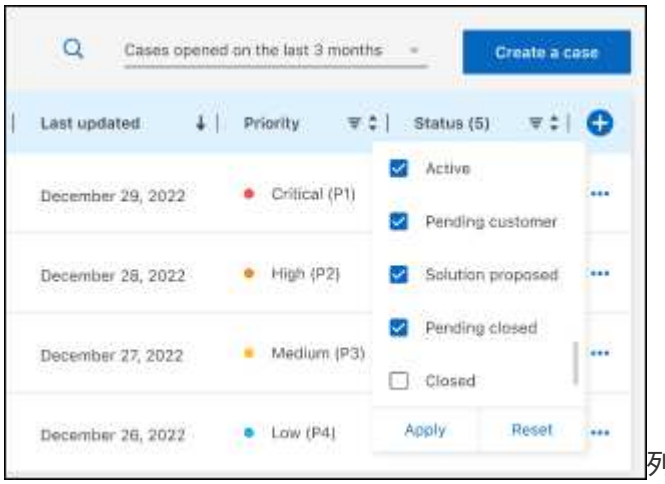
   ケース管理 ページには、 NetAppコンソール ユーザー アカウントに関連付けられている NSS アカウントに関連するオープン ケースが表示されます。これは、**NSS** 管理 ページの上部に表示される NSS アカウントと同じです。

3. 必要に応じて、テーブルに表示される情報を変更します。
   - [Organization’s Cases]*で[View]*を選択すると、会社に関連付けられているすべてのケースが表示されます。
   - 正確な日付範囲を選択するか、別の期間を選択して、日付範囲を変更します。



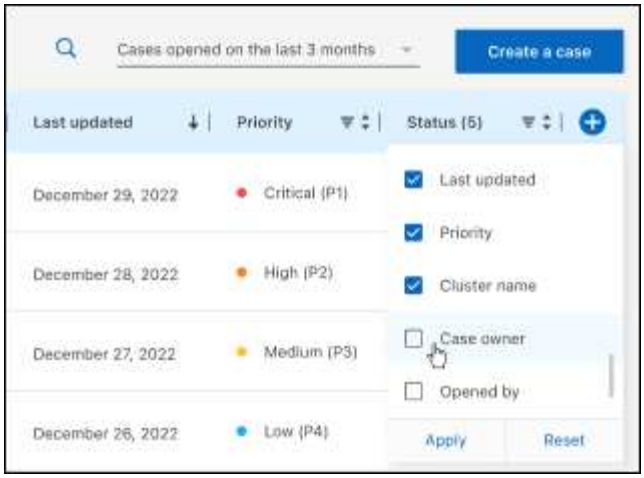
ページのテーブルの上にあるオプションのスクリーンショット。正確な日付範囲、または過去7日、30日、または3カ月を選択できます。"]

- 列の内容をフィルタリングします。



列のフィルタオプションのスクリーンショット。[Active]や[Closed]など、特定のステータスに一致するケースを除外できます。"]

- テーブルに表示される列を変更するには、を選択し [テーブルに表示されるプラスアイコン] 、表示する列を選択します。

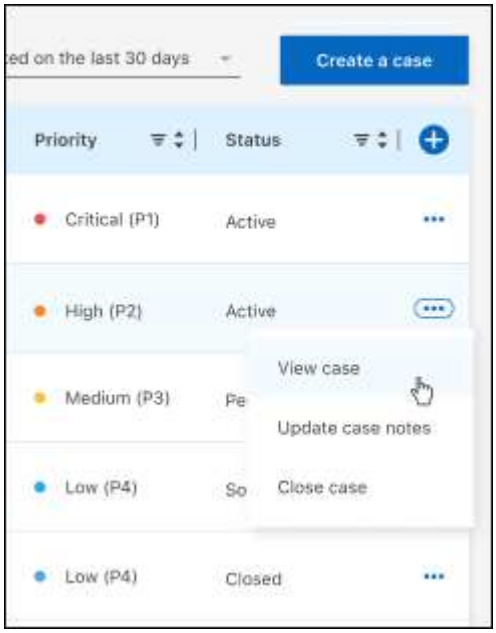


4. 使用可能なオプションのいずれかを選択して、既存のケースを管理し [テーブルの最後の列に表示される3つの点を持つアイコン] ます。

   - ケースの表示:特定のケースの詳細を表示します。
   - ケースノートの更新：問題の詳細を入力するか、*ファイルのアップロード*を選択して最大5つのファイルを添付します。

     添付ファイルはファイルあたり25 MBに制限されています。サポートされているファイル拡張子は、txt、log、pdf、jpg/jpeg、rtf、 doc/docx、xls/xlsx、およびcsv。

   - ケースをクローズ：ケースをクローズする理由の詳細を入力し、*ケースをクローズ*を選択します。



NetApp Workload Factory for GenAI の法的通知

法的通知では、著作権に関する声明、商標、特許などにアクセスできます。

著作権

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

商標

NetApp、NetAppのロゴ、およびNetAppの商標ページに記載されているマークは、NetApp、Inc.の商標です。その他の会社名および製品名は、それを所有する各社の商標である場合があります。

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

特許

NetAppが所有する特許の最新リストは、次のサイトで参照できます。

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

プライバシーポリシー

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

オープンソース

通知ファイルには、ネットアップソフトウェアで使用されるサードパーティの著作権およびライセンスに関する情報が記載されています。

"NetAppワークロード ファクトリー"

著作権に関する情報

商標に関する情報