BlueXP workload factory for GenAI documentation

GenAI

NetApp
August 04, 2025

Table of Contents

BlueXP workload factory for GenAI documentation

Release notes

What's new with BlueXP workload factory for GenAI

Learn what's new with the Generative AI workloads capability of workload factory.

03 August 2025

Secure storage for structured data results

If chatbot query results contain structured data, GenAI can store the results in an Amazon S3 bucket. When these results are stored in an S3 bucket, you can download them using the download link within the chat session.

Create a GenAI knowledge base

MCP server availability

NetApp now provides a Model Context Protocol (MCP) server with BlueXP workload factory for GenAI. You can install the server locally to enable external MCP clients to discover and retrieve query results from a GenAI knowledge base.

NetApp workload factory GenAI MCP server

29 June 2025

Support for data sources hosted on generic NFS/SMB filesystems

You can now add a data source from a generic SMB or NFS share. This enables you to include files that are stored on volumes hosted by filesystems other than Amazon FSx for NetApp ONTAP.

Add data sources to a knowledge base

Add data sources to a connector

03 June 2025

Tracker available for monitoring and tracking operations

The Tracker monitoring capability is now available in GenAI. You can use Tracker to monitor and track the progress and status of pending, ongoing, and completed operations, review details for operation tasks and subtasks, diagnose any issues or failures, edit parameters for failed operations, and retry failed operations.

Monitor workload operations with Tracker in BlueXP workload factory

Choose a reranker model for a knowledge base

You can now increase the relevance of reranked query results by selecting a specific reranker model to use with a knowledge base. GenAI supports the Cohere Rerank and Amazon Rerank models.

Create a GenAI knowledge base

04 May 2025

Support for NetApp Connector for Amazon Q Business

This release of GenAI introduces support for NetApp Connector for Amazon Q Business, enabling you to create connectors for Amazon Q Business. Quickly and easily take advantage of the Amazon Q Business AI assistant with less initial configuration than building a GenAI knowledge base for Amazon Bedrock.

Create a NetApp Connector for Amazon Q Business

Enhanced chat model support

GenAI now supports the following additional chat models for knowledge bases:

- Mistral AI models
- Amazon Titan text models
- Meta Llama models
- Jamba 1.5 models
- Cohere Command models
- Deepseek models

GenAI supports the models from each provider that Amazon Bedrock supports:
Supported foundation models in Amazon Bedrock

Create a GenAI knowledge base

Updated permissions terminology

The workload factory user interface and documentation now use "read-only" to refer to read permissions and "read/write" to refer to automate permissions.

02 March 2025

Embedded chatbot enhancements

You can now copy questions and responses directly to the clipboard, adjust the size of the chat window, and change its title. Additionally, chat responses can now include tables, which are also copyable.

Test a GenAI knowledge base

Chat response citation support

Chat responses now include citations that list the files and chunks of data that were used to generate the response.

Test a GenAI knowledge base

Enhanced file type support

This release of GenAI provides enhanced file support:

- Chat models feature improved CSV support. This enables more useful responses when querying data from

CSV files.

- GenAI can now ingest Apache Parquet files from data sources.
- GenAI now supports ingesting Microsoft Word DOCX files that include images. Images embedded within DOCX documents are scanned, and text insights from the embedded images are included in responses to knowledge base queries.

Supported data source file formats

02 February 2025

Support for Amazon Nova foundation models

GenAI now supports the Amazon Nova foundation models. Amazon Nova Micro, Amazon Nova Lite, and Amazon Nova Pro are supported.

GenAI requirements

File type filtering for data sources

GenAI now supports selecting specific file types to include in the data source scan when you add a data source.

Add data sources to the knowledge base

File modification date filtering for data sources

GenAI now supports filtering files to include in the data source scan by modification date when you add a data source. You can choose a modification date range for the included files.

Add data sources to the knowledge base

Support for image files and enhanced support for PDF files

GenAI now supports enhancing responses to knowledge base queries with insights from images and graph descriptions, as well as document text, leading to richer and higher quality answers. GenAI can now scan image files and images within PDF files (also known as multi-modal file support). If you choose to scan images or PDF files, the text from the images (including images embedded in PDF documents) is scanned into the data source and insights from the scans are included in the responses to knowledge base queries.

Add data sources to the knowledge base

Hybrid search and rerank support

GenAI can now significantly enhances the relevance and accuracy of search results by using hybrid search and re-ranking the results. Hybrid search combines the strengths of traditional keyword-based search with advanced dense vector-based semantic search techniques. The standard keyword search results are augmented with close matches and linguistic nuance, enhancing relevance. GenAI then refines these results further by using advanced re-ranking models, such as Cohere Rerank and Amazon Rerank, and returns the most relevant results. This capability is available for newly created knowledge bases.

Learn about BlueXP workload factory for GenAI

05 January 2025

Custom snapshot name

You can now provide a snapshot name for an ad-hoc snapshot.

Protect a knowledge base with snapshots

Custom AI engine instance name

You can now give a custom name to the AI engine instance during deployment.

Deploy the GenAI infrastructure

Rebuild corrupted or missing GenAI infrastructure

If your AI engine instance becomes corrupted or is somehow deleted, you can let workload factory rebuild it for you. Workload factory automatically reattaches your knowledge bases to the infrastructure after rebuilding is complete, so that they are ready to use.

Troubleshooting

01 December 2024

Clone a knowledgebase from a snapshot

BlueXP workload factory for GenAI now supports cloning a knowledge base from a snapshot. This enables quick recovery of knowledge bases and creation of new knowledge bases with existing data sources, and helps with data recovery and development.

Clone a knowledge base

On-premises ONTAP cluster discovery and replication

Discover and replicate on-premises ONTAP cluster data to an FSx for ONTAP file system so that it can be used to enrich AI knowledge bases. All on-premises discovery and replication workflows are possible from the new **On-Premises ONTAP** tab in the Storage inventory.

Discover an on-premises ONTAP cluster

3 November 2024

Mask Personal Identifiable Information with data guardrails

The Generative AI workload introduces the data guardrails feature, powered by BlueXP classification. The data guardrails feature identifies and masks Personal Identifiable Information (PII) helping you maintain compliance and strengthen security for your sensitive organizational data.

Create a GenAI knowledge base

Learn about BlueXP classification

29 September 2024

Snapshot and restore support for knowledge base volumes

You can now protect your Generative AI workloads data by taking a point-in-time copy of a knowledge base. This enables you to protect your data against accidental loss or test changes to the settings of the knowledge base. You can restore the previous version of the knowledge base volume at any time.

Take a snapshot of a knowledge base volume

Restore a snapshot of a knowledge base volume

Pause scheduled scans

You can now pause scheduled data source scans. By default, Generative AI workloads scans each data source daily to ingest new data into each knowledge base. If you don't want the latest changes to be ingested (during testing or while restoring a snapshot, for example) you can pause the scheduled scans and resume them at any time.

Manage knowledge bases

Data protection volumes now supported for knowledge bases

When selecting a knowledge base volume, you can now choose a data protection volume that is part of a NetApp SnapMirror replication relationship. This enables you to store knowledge bases on volumes that are already protected by SnapMirror replication.

Identify the data sources to integrate in your knowledge base

1 September 2024

Additional chunking strategies

Generative AI workloads now supports multi-sentence chunking and overlap-based chunking for data sources.

Dedicated volume for each knowledge base

Generative AI workloads now creates a dedicated Amazon FSx for NetApp ONTAP volume for each new knowledge base, enabling individual snapshot policies for each knowledge base and improved protection against failures and data poisoning.

4 August 2024

Amazon CloudWatch Logs integration

Generative AI workloads is now integrated with Amazon CloudWatch Logs, enabling you to monitor Generative AI workloads log files.

Example chatbot application

The NetApp workload factory GenAI sample application enables you to test authentication and retrieval from your published NetApp workload factory knowledge base by interacting directly with it in a web-based chatbot application.

7 July 2024

Initial release of the workload factory for GenAI

The initial release includes the capability to develop a knowledge base that is customized by embedding your organization’s data. The knowledge base can be accessed by a chatbot application for your users. This capability ensures accurate and relevant responses to organization-specific questions, enhancing the satisfaction and productivity for all of your users.

Learn about BlueXP workload factory for GenAI

Learn about BlueXP workload factory for GenAI

BlueXP Workload factory for GenAI enables you to integrate Amazon FSx for NetApp ONTAP file systems with GenAI foundation models. This provides you with high performance storage with a rich set of protection, security, and cost optimization features for your AI datasets.

What is BlueXP workload factory for GenAI?

BlueXP workload factory for GenAI enables you to use your enterprise data sources on Amazon FSx for NetApp ONTAP with Generative AI applications. Utilizing retrieval-augmented generation (RAG), you can quickly connect data sources to foundation models available via Amazon Bedrock or Amazon Q Business to develop Generative AI powered applications such as virtual assistants, Q&A chatbots, document summarization, content creation, etc.

Using Generative AI with your organizational data enables you to leverage your own knowledge and expertise, not rely on just the model's intelligence based on public data the models were trained on. Using RAG to customize the models ensures accurate and relevant responses to organization-specific questions, enhancing the productivity and efficiency for the users of your applications using Generative AI.

Developing a GenAI application that is tailored to your organization's data enables you to leverage your own knowledge and expertise. This customization capability ensures accurate and relevant responses to organization-specific questions, enhancing the satisfaction and productivity for all of your users.

If you create a knowledge base, GenAI ingests data from your data sources, stores the vectorized results in a database, and gives you full control over how to use the ingested data to answer queries. This approach requires more initial configuration, but enables you to choose different chat models for different results. If you define a NetApp Connector for Amazon Q Business, data from your data sources is ingested by Amazon Q Business and stored in an index. This approach requires less initial configuration, but gives you less control over the results.

For more information about workload factory, refer to the workload factory overview.

Benefits of using GenAI to create generative AI applications

BlueXP workload factory for GenAI simplifies the process to deploy infrastructure needed to build Generative AI applications using retrieval-augmented generation (RAG). Specifically, GenAI provides the following benefits:

- Without needing a deep knowledge of data infrastructure, foundation and language models, IT administrators and developers can accelerate application development by utilizing the automation provided by GenAI. Data administrators and developers can easily and quickly create enterprise knowledge bases that embed your organization's unstructured data to be used by generative AI applications.
- Enhance security by preserving user permissions in files embedded in the knowledge bases to ensure that data security and privacy is maintained. An application, such as a chatbot, can be developed to provide only the authenticated users with answers based on data the users have access to.
- Keep your enterprise data private and secure within your AWS customer account where your organizational data is never externally exposed.

- Accelerate development of GenAI applications such as a Q&A chatbot using open-source frameworks such as LangChain utilizing the GenAI API to provision and manage knowledge bases and connectors, chat with a knowledge base, and store and retrieve chat history.
- Improve data protection and availability posture by deploying the generative AI data infrastructure on FSx for NetApp ONTAP file systems and taking advantage of ONTAP features such as high availability, snapshots for local data protection and recovery, SnapMirror for disaster recovery, and SnapVault for backing up your data infrastructure.
- Reduce overall storage costs for generative AI data infrastructure by taking advantage of ONTAP data efficiency features such as data deduplication, compression and compaction, data tiering, and thin provisioning.
- Get high quality results from your data with the hybrid search and re-rank features provided by GenAI. Hybrid search combined with re-ranking greatly improve the relevance of search results. These features are available through Amazon AWS and are region-dependent.

How GenAI works

GenAI uses your organization's private data to complement the model's intelligence (based on the data it was trained on) to provide customized answers to questions asked by your users in your organization. You first deploy the infrastructure needed for a RAG framework, then build a knowledge base or define a connector using your organization's data sources and foundation models available via Amazon Bedrock or Amazon Q Business, and then connect an application (such as a Q&A chatbot) to the knowledge base or connector.

How BlueXP workload factory for GenAI helps to build Generative AI applications

GenAI helps to build generative AI applications using RAG in the following ways:

- Deploys the required infrastructure for retrieval-augmented generation (RAG) framework to work with data sources on FSx for ONTAP file systems and Amazon Bedrock or Amazon Q Business. The infrastructure includes the NetApp GenAI Engine instance for managing data, an embedded vector database (LanceDB), and storage on your FSx for ONTAP file system for the vector database.
- Helps connect the data sources to embeddings and language models available via Amazon Bedrock or Amazon Q Business for embedding data sources and retrieving the responses for user queries. The data sources, along with models and their configuration, are presented as FSx for ONTAP knowledge bases.
- Ingests source data into the knowledge base or connector to embed source files on SMB shares and NFS exports on FSx for ONTAP file systems along with storing file permissions for files within SMB shares.
- Automatically builds conversation starter questions based on the content in knowledge bases.
- Provides a chat simulator for data administrators to test chatting with knowledge bases.
- Provides a simple connector interface so you can connect GenAI with Amazon Q Business, quickly and easily utilizing the capabilities of this AI assistant.

Tools to use workload factory

You can use BlueXP workload factory with the following tools:

- **Workload factory console**: The workload factory console provides a visual, holistic view of your applications and projects.
- **BlueXP console**: The BlueXP console provides a hybrid interface experience so that you can use BlueXP workload factory along with other BlueXP services.

- **Ask me**: Use the Ask me AI assistant to ask questions and learn more about workload factory without leaving the workload factory web UI. Access Ask me from the workload factory help menu.
- **CloudShell CLI**: Workload factory includes a CloudShell CLI to manage and operate AWS and NetApp environments across accounts from a single, browser-based CLI. Access CloudShell from the top bar of the workload factory console.
- **REST API**: Use the workload factory REST APIs to deploy and manage your FSx for ONTAP file systems and other AWS resources.
- **CloudFormation**: Use AWS CloudFormation code to perform the actions you defined in the workload factory console to model, provision, and manage AWS and third-party resources from the CloudFormation stack in your AWS account.
- **Terraform BlueXP workload factory provider**: Use Terraform to build and manage infrastructure workflows generated in the workload factory console.

Cost

There is no cost for using the GenAI capability of workload factory.

However, you'll need to pay for AWS resources that you deploy in order to support the generative AI infrastructure. For example, you will pay AWS for Amazon Bedrock or Amazon Q Business, FSx for ONTAP file system and storage capacity, and the GenAI engine EC2 instance.

Some multi-modal operations, such as scanning images for text information, can use more resources and therefore incur a higher cost. Some configuration operations, such as changing settings for a knowledgebase, can cause data sources to be re-scanned, and data source scans can also incur a higher cost.

Licensing

No special licenses are needed from NetApp to use the AI capabilities of workload factory.

Components of the NetApp GenAI engine

When you deploy the GenAI infrastructure, workload factory creates an EC2 instance for the GenAI engine. It also creates an IAM role, security group, and private endpoints for this instance. You might want to understand more details about these components that workload factory creates in your AWS environment.

EC2 instance type

m5.large

IAM role

The GenAI engine instance needs permissions to send chunks of data to the embedding model on Amazon Bedrock and to communicate with the NetApp AI Service Backend. The IAM role includes the following permissions:

IAM role permissions

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

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

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

```
        "ec2messages:FailMessage",
        "ec2messages:SendReply"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "qbusiness:ListWebExperiences",
        "qbusiness:GetApplication",
        "qbusiness:CreateDataSource",
        "qbusiness:DeleteDataSource",
        "qbusiness:ListIndices",
        "qbusiness:StartDataSourceSyncJob",
        "qbusiness:StopDataSourceSyncJob",
        "qbusiness:ListDataSourceSyncJobs",
        "qbusiness:BatchPutDocument",
        "qbusiness:BatchDeleteDocument"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogGroups"
      ],
      "Resource": "*",
      "Effect": "Allow"
    },
    {
      "Action": [
        "logs:DescribeLogStreams",
        "logs:PutLogEvents",
        "logs:CreateLogStream",
        "logs:CreateLogGroup"
      ],
      "Resource": [
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*:log-stream:*",
        "arn:aws:logs:*:*:log-group:/netapp/wlmai/*"
      ],
      "Effect": "Allow"
    },
    {
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
```

```
        ],
        "Resource": "*",
        "Effect": "Allow"
      },
      {
        "Action": [
          "kms:Decrypt",
          "kms:GenerateDataKey"
        ],
        "Resource": "*",
        "Effect": "Allow"
      }
    ]
}
```

Security group

The outbound rules are open to all traffic, while the inbound rules are completely closed.

Private endpoints

If the target VPC doesn't already have them, workload factory creates private endpoints for the GenAI engine EC2 instance so that it can communicate with the following AWS services:

- Amazon Bedrock
  - bedrock
  - bedrock-runtime
  - bedrock-agent-runtime
- Amazon Elastic Container Registry (ECR)
  - API
  - docker
- AWS Systems Manager (SSM)
  - ssm
  - ec2messages
  - ssmmessages
- Amazon FSx for NetApp ONTAP
- Amazon CloudWatch

Use GenAI to build knowledge bases for Amazon Bedrock

Get started

Quick start for GenAI knowledge bases

Get started creating a knowledge base or Amazon Q Business connector using your organization's data that exists on Amazon FSx for NetApp ONTAP file systems. An application such as a chatbot will access this knowledge base or connector to provide organization-focused responses to end users.

Log in to workload factory

You'll need to set up an account with workload factory and log in using one of the console experiences.

Set up your environment to meet GenAI requirements

You'll need AWS credentials to deploy the AWS infrastructure, a deployed and discovered FSx for ONTAP file system, the list of data sources that you want to integrate in your knowledge base or connector, access to the Amazon Bedrock AI service or Amazon Q Business application, and more.

Learn more about GenAI requirements.

Identify the FSx for ONTAP file system that contains the data sources

The data sources that you'll integrate in your knowledge base can be located on a single FSx for ONTAP file system, or on multiple FSx for ONTAP file systems. If these systems are in different VPCs, they must either be accessible within the same network, or the VPCs must be peered and using the same region and AWS account as the AI engine.

Learn how to identify data sources.

Deploy the GenAI infrastructure

Launch the infrastructure deployment wizard to deploy the GenAI infrastructure in your AWS environment. This process deploys an EC2 instance for the NetApp GenAI engine, and a volume on an FSx for ONTAP file system to contain the NetApp AI Engine databases. The volume is used to store the vector database used by the knowledge base.

Learn how to deploy the knowledge base infrastructure.

What's next

You can now build a knowledge base to provide organization-focused responses to end users.

GenAI knowledge base requirements

Ensure that workload factory and AWS are set up properly before you build your knowledge base. This includes having your AWS log in credentials, a deployed FSx for ONTAP file system that contains the data sources you want to integrate in your knowledge base, access to the Amazon Bedrock AI service, and more.

Basic GenAI requirements

GenAI has general requirements that your environment needs to meet before you get started.

Workload factory login and account

You'll need to set up an account with workload factory and log in using one of the console experiences.

AWS credentials and permissions

You need to add AWS credentials to workload factory with read/write permissions, which means you'll be using workload factory in *read/write* mode for GenAI.

Basic mode and *Read-only* mode permissions are not supported at this time.

When setting up your credentials, selecting permissions as shown below provides you with full access to manage FSx for ONTAP file systems and to deploy and manage the GenAI EC2 instance and other AWS resources needed for your knowledge base and chatbot.

Learn how to add AWS credentials to workload factory

GenAI knowledge base requirements

If you plan to work with knowledge bases, ensure your environment meets the following requirements.

Amazon Bedrock

Amazon Bedrock enables you to use foundation models and it provides the capabilities to build generative AI applications.

Before you get started with BlueXP workload factory for GenAI, you must set up Amazon Bedrock. Your GenAI deployment must be in an AWS region that has Amazon Bedrock enabled.

- AWS documentation: Set up Amazon Bedrock
- AWS documentation: Supported regions and models for Knowledge bases for Amazon Bedrock

GenAI re-ranks search results by default to improve result relevancy. For the best results, ensure that your Amazon Bedrock foundation model configuration includes access to a re-rank model, such as Cohere Rerank or Amazon Rerank, if available in your region.

Embedding model

You must enable the embedding model that you plan to use before creating your knowledge base. The following embedding models are supported:

- Titan Embeddings G1 - Text
- Titan Embedding Text v2
- Titan Multimodal Embeddings G1

- Embed English
- Embed Multilingual

  Learn more about Amazon Titan

Chat model

You must enable the foundational chat model that you plan to use before creating your knowledge base. Since model support varies by AWS region, refer to the AWS documentation to verify which models you can use in the regions where you plan to deploy your knowledge base.

GenAI supports various models from Anthropic, Amazon, Mistral AI, Meta, Jamba, and Cohere.

Learn more about using these models in Amazon Bedrock:

- Anthropic's Claude in Amazon Bedrock
- Getting started with Amazon Nova in the Amazon Bedrock console
- Mistral AI models
- Amazon Titan text models
- Meta Llama models
- Jamba models
- Cohere Command models

FSx for ONTAP file system

You need a minimum of one FSx for ONTAP file system:

- One file system will be used (or created, if it doesn't exist) by the NetApp GenAI engine to store the vector database used by the knowledge base.

  This FSx for ONTAP file system must use FlexVol volumes. FlexGroup volumes are not supported.

- One or more file systems will contain the data sources that you'll be integrating into your knowledge base.

  One FSx for ONTAP file system can be used for both of these purposes, or you can use multiple FSx for ONTAP file systems.

- You'll need to know the AWS region, VPC, and subnet where the AWS FSx for ONTAP file system resides. The file system must be in an AWS region that has Amazon Bedrock enabled.
- You'll need to consider the tag key/value pairs that you want to apply to the AWS resources that are part of this deployment (optional).
- You'll need to know the key pair information that allows you to securely connect to the NetApp AI engine instance.

Learn how to deploy and manage FSx for ONTAP file systems

Identify data sources to add to a knowledge base or connector

Identify, or create, the documents (data sources) that reside on your FSx for ONTAP file system that you'll integrate in your knowledge base. These data sources enable the

knowledge base to provide accurate and personalized answers to user queries based on data that is relevant to your organization.

Maximum number of data sources

The maximum number of supported data sources is 10.

Location of data sources

Data sources can be stored in a single volume, or in a folder within a volume, on an SMB share or NFS export on an Amazon FSx for NetApp ONTAP file system. Data sources can also be stored on Amazon FSx for NetApp ONTAP volumes that are in a NetApp SnapMirror data protection relationship.

You can't select individual documents within a volume or folder, therefore, you should ensure that each volume or folder that contains data sources does not contain extraneous documents that shouldn't be integrated with your knowledge base.

You can add multiple data sources into each knowledge base, but they all need to reside on FSx for ONTAP file systems that are accessible from your AWS account.

The maximum file size for each data source is 50 MB.

Supported protocols

Knowledge bases support data from volumes that use either NFS or SMB/CIFS protocols. When selecting files stored using the SMB protocol, you'll need to enter the Active Directory information so that the knowledge base can access the files on those volumes. This includes the Active Directory Domain, IP address, user name, and password.

When storing your data source on a share (file or directory) accessed over SMB, the data is only accessible by chatbot users or groups who have the permissions to access that share. When this "permission-aware capability" is enabled, the AI system will compare the user email in auth0 to the users allowed to view or use the files on the SMB share. The chatbot will provide answers based on user permissions for the embedded files.

For example, if you have integrated 10 files (data sources) into your knowledge base, and 2 of the files are human resources files that contain restricted information, only chatbot users who are authenticated to access those 2 files will received responses from the chatbot that include data from those files.

Supported data source file formats

The following data source file formats are currently supported with workload factory GenAI knowledge bases.

File format	Extension
Apache Parquet [1]	.parquet
Comma-separated values file[1]	.csv
Graphics Interchange Format	.gif
JPEG	.jpg or.jpeg
JSON and JSONP[1]	.json
Markdown	.md

File format	Extension
Microsoft Word	.doc or .docx
Plain text	.txt
Portable Document Format	.pdf
Portable Network Graphics	.png
WebP image	.webp

Deploy the GenAI infrastructure

Unresolved directive in knowledge-base/deploy-infrastructure.adoc - include::_include/deploy-infrastructure.adoc[]

Create a GenAI knowledge base

After you've deployed the AI infrastructure and identified the data sources that you'll integrate in your knowledge base from your FSx for ONTAP datastores, you are ready to build the knowledge base using workload factory. As part of this step, you'll also define the AI characteristics and create conversation starters.

Ensure that your environment meets the requirements for knowledge bases before proceeding.

About this task

Knowledge bases have two data integration modalities - *public mode* and *Enterprise mode*.

Public mode

A knowledge base can be used without integrating data sources from your organization. In this case, an application integrated with the knowledge base will only provide results from publicly available information on the internet. This is known as a *public mode* integration.

Enterprise mode

In most cases you'll want to integrate data sources from your organization into the knowledge base. This is known as an *Enterprise mode* integration because it provides knowledge from your enterprise.

Data sources from your organization may contain Personally Identifiable Information (PII). To safeguard this sensitive information, you can enable *data guardrails* when creating and configuring knowledge bases. Data guardrails, powered by BlueXP classification, identifies and masks PII, making it inaccessible and irretrievable.

Learn about BlueXP classification.

BlueXP workload factory for GenAI does not mask sensitive personal information (SPii). Refer to types of sensitive personal data for more information about this type of data.

Data guardrails can be enabled or disabled at any time. If you switch data guardrails enablement, workload factory scans the entire knowledge base from scratch, which incurs a cost.

Create and configure the knowledge base

The knowledge base defines characteristics such as the Bedrock AI models and embedding format that you want to use to create your knowledge base.

Steps

1. Log in to workload factory using one of the console experiences.
2. In the AI workloads tile, select **Deploy & manage**.
3. From the Knowledge bases & Connectors tab, select the **Create New** dropdown and choose **NetApp GenAI knowledge base for Bedrock**.
4. On the Define knowledge base page, configure the knowledge base settings:

    a. **Name**: Enter the name you want to use for the knowledge base.

    b. **Description**: Enter a detailed description for the knowledge base.

    c. **Embedding model**: The embedding model defines how your data will be converted into vector embeddings for the knowledge base. Workload factory supports the following models:

    - Titan Embeddings G1 - Text
    - Titan Embedding Text v2
    - Titan Multimodal Embeddings G1
    - Embed English
    - Embed Multilingual

    Note that you must have already enabled the embedding model from Amazon Bedrock.

    Learn more about Amazon Titan

    d. **Chat model**: Choose from various chat models that are integrated in Amazon Bedrock. Note that you must have already enabled the chat model from Amazon Bedrock.

    e. **Reranking**: Enable or disable reranking, which can improve the relevance and quality of query results. Choose a standard chat model or a specialized reranker model to use for reranking. Reranker model options are only shown if they are available in your region.

    f. **Data guardrails**: Choose whether you want to enable or disable data guardrails. Learn about data guardrails, powered by BlueXP classification.

    The following prerequisites must be met to enable data guardrails.

    - A service account is required to communicate with BlueXP classification. You must have the *Organization admin* role on your BlueXP tenancy account for service account creation. A member who has the Organization admin role can complete all actions in BlueXP. Learn how to add a role to a member in BlueXP
    - The AI engine must have access to the BlueXP API endpoint.
    - You'll need to do the following as described in BlueXP classification documentation:
        1. Create a BlueXP Connector
        2. Ensure that your environment can meet the prerequisites
        3. Deploy BlueXP classification

The data guardrails feature is not supported when ingesting structured data files such as CSV, JSON, JSONP, or Parquet.

f. **Conversation starters**: Choose whether you want to provide up to four conversation starter prompts that are displayed to users who interact with a chatbot that uses this knowledge base. We recommend that you enable this setting.

   If you activate conversation starters, "Automatic mode" is selected by default. "Manual mode" can be enabled only after you've added data sources to your knowledge base. Learn how to modify knowledge base settings.

g. **FSx for ONTAP file system**: When you define a new knowledge base, Workload factory creates a new Amazon FSx for NetApp ONTAP volume to store it. Choose an existing file system name and SVM (also called a storage VM) where the new volume will be created.

h. **Snapshot policy**: Choose a snapshot policy from the list of existing policies defined in the workload factory storage inventory. Recurring snapshots of the knowledge base will automatically be created at a frequency based on the snapshot policy you select.

i. **S3 Bucket**: If chatbot query results contain structured data, GenAI can store the results in an S3 bucket. To use this feature, enable the **Activate S3 Bucket** setting and choose an S3 bucket that is associated with your account from the list. When these results are stored in an S3 bucket, you can download them using the download link within the chat session.

   If the snapshot policy you need doesn't exist, you can create a snapshot policy on the storage VM that contains the volume.

5. Select **Create knowledge base** to add the knowledge base to GenAI.

   A progress indicator appears while the knowledge base is created.

   After the knowledge base is created, you have the option to add a data source to your new knowledge base or to end the process without adding a data source. We recommend that you select **Add data source** and add one or more data sources now.

Add data sources to the knowledge base

You can add one or more data sources to populate the knowledge base with your organization's data.

About this task

The maximum number of supported data sources is 10.

Steps

1. After you select **Add data source**, select the type of data source you want to add:
   - Add FSx for ONTAP file system (use files from an existing FSx for ONTAP volume)
   - Add file system (use files from a generic SMB or NFS share)

Unresolved directive in knowledge-base/create-knowledgebase.adoc - include::_include/add-data-source-kb.adoc[]

Result

The data source starts to be embedded into your knowledge base. The status changes from "Embedding" to

"Embedded" when the data source is completely embedded.

After you add a single data source to the knowledge base, you can test it locally in the chatbot simulator window and make any required changes before you make the chatbot available to your users. You can also follow the same steps to add additional data sources to the knowledge base.

Test a GenAI knowledge base

After you create the knowledge base, you'll be able to test it locally using the chatbot simulator and make any required changes before you make the knowledge base available to your users through a chatbot application.

About this task

You test your knowledge base to make sure it performs as you expect, and you can customize the conversation starters that you want to be available by default for chatbot users of this knowledge base. The chatbot simulator runs against all the data sources that have been embedded in the knowledge base.

You can test a knowledge base by chatting with your embedded data sources in the chatbot simulator. Note that none of the interaction or insights are captured in the GenAI vector database when testing the knowledge base locally.

You'll perform most of your testing within workload factory before you deploy the knowledge base in an application for your users. If you need to make changes to your data source or the chatbot operation, you'll want to do it now before you publish your knowledge base.

You can resize and retitle the chatbot simulator window, and copy questions and responses to the clipboard.

Some of the tasks you'll want to perform to test your chatbot are:

- Enter a large number of questions that are relevant to your organization to make sure the answers are as expected.
- Customize the conversation starters that you want to be available by default for your users in the chatbot application.
- Make sure that the attributed content that is provided at the bottom of the chatbot answers contain the correct references.

Steps

1. From the Knowledge bases inventory page, select the knowledge base that you want to test.

   The chatbot simulator appears in the right pane. If defined, existing conversation starters are displayed as well.

2. In the chatbot entry field, enter a prompt or question and select ➤ to see how your chatbot responds with your organizational knowledge.

- You can see the sources used to produce the answer by expanding the **Sources** list under the response. This provides a list of files used to generate the answer. You can view and copy the data chunks used from each file and volume path to each file by hovering over the file name.
- If tables are included in the answer, you can sort the data in each column, and copy each table to the clipboard.
- If answer results contain structured data and the **S3 Bucket** feature is enabled for the knowledge base, GenAI stores the results in an S3 bucket. You can download the results from the bucket using the **Download results** link within the chat session.

3. If you need to update any of your data sources so that your knowledge base provides more focused answers, make those changes now and then retest the knowledge base.

Activate external authentication for a GenAI knowledge base

Activate authentication for a knowledge base so that token validation and ACLs are required when using the API endpoints to integrate a knowledge base with a chatbot application. When you activate authentication, you configure settings for a JSON Web Token that will be used for API requests to a knowledge base from chatbot clients.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base for which you want to activate authentication.
3. Select ••• and select **Manage knowledge base**.
4. Select the **Actions** menu and select **Manage authentication settings**.
5. Set up authentication:
   a. Select **Activate authentication settings**.
   b. Provide the required information. Examples are provided, but you should obtain the values for these fields from your authentication provider:
      - **Algorithms**: The signing algorithm that your authentication provider uses.
      - **Audience** (Optional): A string containing the intended recipient of the token (sometimes a URL).
      - **Issuer**: A string that identifies the provider that issued the token.

        For example, Amazon Cognito uses issuer strings with the following format:

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

        Where `<region>` is the AWS region containing the user pool, and `<UserPoolID>` is your user pool ID. You can retrieve your user pool ID using the following command:

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

- **JWKS URI**: The URI string that provides public keys needed to verify signatures of this token.

  For example, Amazon Cognito uses JWKS URI strings with the following format:

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

  Where `<region>` is the AWS region containing the user pool, and `<UserPoolID>` is your user pool ID. You can retrieve your user pool ID using the following command:

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

6. Select **Save**.

Result

Authentication for the knowledge base is now active, and you can use API endpoints to interact with the knowledge base and integrate the knowledge base with a chatbot application.

Publish a GenAI knowledge base and view the unique endpoint

After you've built and tested your knowledge base locally, you can publish the knowledge base so that it can be integrated with a chatbot application that will enable your users to query the knowledge base.

About this task

Publishing the knowledge base enables you to use it in chat applications. The publish action triggers the workload factory API to generate and publish unique endpoints. After publishing, the knowledge base becomes accessible for chat applications, and the API endpoints are ready for integration.

Each knowledge base that you publish has unique endpoints.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to publish.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, and the list of all embedded data sources.

4. Select the **Actions** menu and select **Publish**.

   Workload factory publishes the knowledge base. On the details page for the knowledge base, the status changes from **Unpublished** to **Published.**

   You can now get details about the unique endpoint for the knowledge base.

5. Next to the published status, select **View**.

   Details about how to access the knowledge base using the workload factory API is displayed.

6. From the **View published info** dialog box, copy the API endpoints that you can use to integrate the knowledge base with an application.

   To learn more about the API endpoints, go to the API documentation and select **AI > External**.

   Before you can use these endpoints, you need to obtain a user token from your authentication provider.

Result

You now have a published knowledge base and the unique endpoint that you can use to integrate the knowledge base with a chatbot application.

Use the GenAI external example chatbot application

After you configure, activate, and publish a knowledge base, external application developers can configure and run the open source example chatbot application provided by NetApp to interact with your knowledge base and to learn how to use the workload factory API to create their own generative AI applications.

Steps

1. Create a knowledge base.
2. Activate authentication for the knowledge base that you created.

   This enables the knowledge base to authenticate API requests, and makes token validation and ACLs required when using the API endpoints.

   > External chat applications that integrate with this knowledge base will need to use the same authentication provider (issuer) that you configure in the authentication settings for the knowledge base.

3. Publish the knowledge base to enable API access for external applications.

   After a knowledge base is published, the API endpoints are accessible externally, and you can integrate the knowledge base with an external chat application (such as the example chatbot application).

4. Download the example chatbot application package from GitHub.
5. Install and run the chatbot application by following the instructions in the README file included in the package.
6. Browse to http://localhost:9091 to log in to the application.

   The example chatbot application appears.

Learn more

workload factory API documentation

Create a RAG-based GenAI application

After you build your knowledge base and test your chatbot, you are ready to set up the application that will enable your users to query the chatbot.

Learn how to create a RAG-based AI application on FSx for ONTAP

What you can do next with GenAI

Now that you've created a knowledge base using your enterprise data and deployed it for your users, you can manage the knowledge base, data sources, and the RAG infrastructure, including FSx for ONTAP file systems.

Some of the tasks you can perform to manage your knowledge base components are:

- Update the content of your data sources, or add new data sources, and sync those changes with your knowledge base and chatbot.
- Manage your data source settings, including the chunking strategy and permission awareness (for SMB file access).
- Manage your knowledge base settings, including the chat model and conversation starters.
- Unpublish a knowledge base or republish it after making changes.
- Back up and protect the important data on your FSx for ONTAP file system to make sure your knowledge base data and other infrastructure components are always available.

For information about managing your FSx for ONTAP file system, go to the workload factory for Amazon FSx for NetApp ONTAP documentation to view the backup and protection capabilities you can use.

[1] The data guardrails feature is not supported when ingesting structured data files into knowledge bases.

Use GenAI to create connectors for Amazon Q Business

Get started

Quick start for GenAI connectors

Get started creating a NetApp Connector for Amazon Q Business using your organization's data that exists on Amazon FSx for NetApp ONTAP file systems. After you create a connector, end users can access the Amazon Q Business assistant for organization-focused responses to their questions.

Log in to workload factory

You'll need to set up an account with workload factory and log in using one of the console experiences.

Set up your environment to meet GenAI requirements

You'll need AWS credentials to deploy the AWS infrastructure, a deployed and discovered FSx for ONTAP file system, the list of data sources that you want to integrate in your connector, access to the Amazon Q Business application, and more.

Learn more about GenAI requirements.

Identify the FSx for ONTAP file system that contains the data sources

The data sources that you'll integrate in your connector can be located on a single FSx for ONTAP file system, or on multiple FSx for ONTAP file systems. If these systems are in different VPCs, they must either be accessible within the same network, or the VPCs must be peered and using the same region and AWS account as the AI engine.

Learn how to identify data sources.

Deploy the GenAI infrastructure

Launch the infrastructure deployment wizard to deploy the GenAI infrastructure in your AWS environment. This process deploys an EC2 instance for the NetApp GenAI engine, and a volume on an FSx for ONTAP file system to contain the NetApp AI Engine databases. The volume is used to store information about the connector.

Learn how to deploy the GenAI infrastructure.

What's next

You can now create a connector for Amazon Q Business to provide organization-focused responses to end users.

GenAI connector requirements

Ensure that workload factory and AWS are set up properly before you create a NetApp Connector for Amazon Q Business.

Basic GenAI requirements

GenAI has general requirements that your environment needs to meet before you get started.

Workload factory login and account

You'll need to set up an account with workload factory and log in using one of the console experiences.

AWS credentials and permissions

You need to add AWS credentials to workload factory with read/write permissions, which means you'll be using workload factory in *read/write* mode for GenAI.

Basic mode and *read-only* mode permissions are not supported at this time.

When setting up your credentials, selecting permissions as shown below provides you with full access to manage FSx for ONTAP file systems and to deploy and manage the GenAI EC2 instance and other AWS resources needed for your knowledge base and chatbot.

Learn how to add AWS credentials to workload factory

Requirements for NetApp Connector for Amazon Q Business

Ensure your environment meets the following specific requirements for Amazon Q Business connectors.

Amazon Q Business application

You need to create an Amazon Q Business application or use an existing one.

- Ensure that the application exists in one of your AWS regions.
- Ensure that you have created an index for the application.
- Ensure that the application is not in a failed state.

FSx for ONTAP file system

You need a minimum of one FSx for ONTAP file system:

- One file system will be used (or created, if it doesn't exist) by the NetApp GenAI engine to store information about the connector.

  This FSx for ONTAP file system must use FlexVol volumes. FlexGroup volumes are not supported.

- One or more file systems will contain the data sources that you'll be adding to your connector.

  One FSx for ONTAP file system can be used for both of these purposes, or you can use multiple FSx for ONTAP file systems.

- You'll need to know the AWS region, VPC, and subnet where the AWS FSx for ONTAP file system resides.
- You'll need to consider the tag key/value pairs that you want to apply to the AWS resources that are part of this deployment (optional).

- You'll need to know the key pair information that allows you to securely connect to the NetApp AI engine instance.

Learn how to deploy and manage FSx for ONTAP file systems

Identify data sources to add to a connector

Identify, or create, the documents (data sources) that reside on your FSx for ONTAP file system that you'll integrate in your connector. These data sources enable Amazon Q Business to provide accurate and personalized answers to user queries based on data that is relevant to your organization.

Maximum number of data sources

The maximum number of supported data sources is 10.

Location of data sources

Data sources can be stored in a single volume, or in a folder within a volume, on an SMB share or NFS export on an Amazon FSx for NetApp ONTAP file system. Data sources can also be stored on Amazon FSx for NetApp ONTAP volumes that are in a NetApp SnapMirror data protection relationship.

You can't select individual documents within a volume or folder, therefore, you should ensure that each volume or folder that contains data sources does not contain extraneous documents that shouldn't be integrated with your knowledge base.

You can add multiple data sources into each connector, but they all need to reside on FSx for ONTAP file systems that are accessible from your AWS account.

The maximum file size for each data source is 50 MB.

Supported protocols

Connectors support data from volumes that use either NFS or SMB/CIFS protocols. When selecting files stored using the SMB protocol, you'll need to enter the Active Directory information so that the connector can access the files on those volumes. This includes the Active Directory Domain, IP address, user name, and password.

When storing your data source on a share (file or directory) accessed over SMB, the data is only accessible by chatbot users or groups who have the permissions to access that share. When this "permission-aware capability" is enabled, the AI system will compare the user email in auth0 to the users allowed to view or use the files on the SMB share. The chatbot will provide answers based on user permissions for the embedded files.

For example, if you have integrated 10 files (data sources) into your connector, and 2 of the files are human resources files that contain restricted information, only chatbot users who are authenticated to access those 2 files will received responses from the chatbot that include data from those files.

When you add data sources to an Amazon Q Business connector, only user permissions apply to data source files. Group permissions are not applied.

If a file in your data source lacks text (for example, a text-free image), Amazon Q Business does not index it but logs an entry in Amazon CloudWatch Logs noting the absence of text.

Supported data source file formats

The following data source file formats are currently supported with NetApp Connector for Amazon Q Business.

File format	Extension
Comma-separated values file	.csv
JSON and JSONP	.json
Markdown	.md
Microsoft Word	.docx
Plain text	.txt
Portable Document Format	.pdf
Microsoft PowerPoint	.ppt or .pptx
Hypertext Markup Language	.html
Extensible Markup Language	.xml
XSLT	.xslt
Microsoft Excel	.xls
Rich Text Format	.rtf

Deploy the GenAI infrastructure

Unresolved directive in connector/deploy-infrastructure.adoc - include::_include/deploy-infrastructure.adoc[]

Create a NetApp Connector for Amazon Q Business

After you've deployed the AI infrastructure and identified the data sources that you'll use from your FSx for ONTAP datastores, you are ready to define a NetApp Connector for Amazon Q Business.

Ensure that your environment meets the requirements for Amazon Q Business before proceeding.

About this task

Data sources from your organization might contain Personally Identifiable Information (PII). To safeguard this sensitive information, you can enable *data guardrails* when defining a connector. Data guardrails, powered by BlueXP classification, identifies and masks PII, making it inaccessible and irretrievable.

Learn about BlueXP classification.

BlueXP workload factory for GenAI does not mask sensitive personal information (SPII). Refer to types of sensitive personal data for more information about this type of data.

Data guardrails can be enabled or disabled at any time. If you switch data guardrails enablement, workload factory scans the entire data source from scratch, which can incur a cost.

Define a connector

Create a NetApp Connector for Amazon Q Business. The connector enables API and data source communication between GenAI and Amazon Q Business.

Steps

1. Log in to workload factory using one of the console experiences.
2. In the AI workloads tile, select **Deploy & manage**.
3. From the Knowledge bases & Connectors tab, select the **Create New** dropdown and choose **Amazon Q Business connector**.
4. On the Define Connector page, configure the connector settings:

    a. **Name**: Enter the name you want to use for the connector.

    b. **Description**: Enter a detailed description for the connector.

    c. **Amazon Q**: The region and application name for the Amazon Q Business instance you want to integrate.

    d. **Data guardrails**: Choose whether you want to enable or disable data guardrails. Learn about data guardrails, powered by BlueXP classification.

    The following prerequisites must be met to enable data guardrails.

    - A service account is required to communicate with BlueXP classification. You must have the *Organization admin* role on your BlueXP tenancy account for service account creation. A member who has the Organization admin role can complete all actions in BlueXP. Learn how to add a role to a member in BlueXP
    - The AI engine must have access to the BlueXP API endpoint.
    - You'll need to do the following as described in BlueXP classification documentation:
        1. Create a BlueXP Connector
        2. Ensure that your environment can meet the prerequisites
        3. Deploy BlueXP classification

    > When you enable the data guardrails feature, GenAI processes .txt, .md, .csv, .docx, and .pdf files by ingesting only plain text (excluding embedded image or media text) and masking any private or sensitive data. All other file types are processed normally without masking private or sensitive data.

    a. **FSx for ONTAP file system**: When you define a new NetApp Connector for Amazon Q Business, workload factory creates a new Amazon FSx for NetApp ONTAP volume to store the connector information. Choose an existing file system and SVM (also called a storage VM) where the new volume will be created.

    b. **Snapshot policy**: Choose a snapshot policy from the list of existing policies defined in the workload factory storage inventory. GenAI automatically creates recurring snapshots of the volume storing the connector information at a frequency based on the snapshot policy you select.

    If the snapshot policy you need doesn't exist, you can create a snapshot policy on the storage VM that contains the volume.

5. Select **Create connector** to integrate Amazon Q Business with GenAI.

A progress indicator appears while the connector is created.

After the connector is created, you have the option to add a data source to the connector so that Amazon Q Business ingests your data and adds it to its index. We recommend that you select **Add data source** and add one or more data sources now.

Add data sources to the connector

You can add one or more data sources to populate the Amazon Q Business index with your organization's data.

About this task

- The maximum number of supported data sources is 10.
- Refer to the Amazon Q Business documentation for specific service restrictions of the Amazon Q Business index.

Steps

1. After you select **Add data source**, the **Select a file system** page appears.
   Unresolved directive in connector/define-connector.adoc - include::_include/add-data-source-connector.adoc[]

Result

The data source is embedded into the Amazon Q Business index. The status changes from "Embedding" to "Embedded" when the data source is completely embedded.

After you add a single data source to the connector, you can test it in the Amazon Q Business chatbot environment and make any required changes before you make the service available to your users. You can also follow the same steps to add additional data sources to the connector.

Administer and monitor

Manage the GenAI infrastructure

You can view details about your deployed GenAI RAG infrastructure or remove the chatbot infrastructure if you no longer need it.

View information about the infrastructure

You can view information about the chatbot infrastructure.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the workload factory navigation menu, select **AI**.
3. Select the **Infrastructure** tab.
4. View information about the infrastructure, which includes details about the following components:
   - AWS settings
   - Infrastructure settings
   - The AI engine
   - The vector database

Remove the infrastructure

If you no longer need the chatbot infrastructure that you deployed for one or more chatbots, you can remove it from workload factory.

> All chatbots that have been deployed on this infrastructure will be disabled and all chat history will be deleted.

This operation only removes the links to the AI infrastructure from workload factory; it does not remove all the components from AWS. You'll need to manually delete the following infrastructure components from AWS:

- The VM instance
- Private endpoints
- The volume on the FSx for ONTAP file system that contains the AI databases
- The IAM role
- The policy
- The security group

Steps

1. Log in to workload factory using one of the console experiences.
2. From the workload factory navigation menu, select **AI**.
3. Select the **Infrastructure** tab.
4. Select ••• and select **Remove chatbot infrastructure**.

5. Confirm that you want to delete the infrastructure and select **Remove**.

Result

The chatbot infrastructure components are removed from workload factory.

Manage GenAI knowledge bases

After you create a knowledge base, you can view the knowledge base details, modify the knowledge base, integrate additional data sources, or delete the knowledge base.

View information about a knowledge base

You can view information about the settings for a knowledge base and the data source that are integrated.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the workload factory navigation menu, select **AI**.
3. Select the knowledge base that you want to view.

   If defined, the conversation starters that are currently being used display in the right pane.

4. To view knowledge base details, select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

   The **Actions** menu enables you to manage the knowledge base if you want to make any changes.

Edit a knowledge base

You can update a knowledge base by changing some settings, or you can add or remove data sources.

Each time you add, modify, or remove data sources from the knowledge base, you must sync the data source so that it is re-indexed to the knowledge base. Syncing is incremental, so Amazon Bedrock only processes the objects in your FSx for ONTAP volume that have been added, modified, or deleted since the last sync.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to update.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

4. Select the **Actions** menu and select **Edit knowledge base**.
5. In the Edit knowledge base page, you can change the knowledge base name, description, embedding model, chat model, feature enablement, choose whether conversation starters are created automatically or manually, and the snapshot policy used for the volume that contains the knowledge base.

   If you use Manual mode for conversation starters, you can change conversation starters here as well.

Every knowledge base scan, which includes embedding, costs. If data guardrails is enabled after a knowledge base has been created, then the knowledge base gets scanned again and incurs costs. Similarly, if you change chat models, GenAI will re-scan the associated data sources (incurring a cost).

6. Select **Save** after you have made your changes.

Protect a knowledge base with snapshots

You can protect your knowledge base data by taking and restoring snapshots of your knowledge base volumes. You can restore from a snapshot to revert to the previous version of the knowledge base at any time.

Snapshots can be faster and more storage-efficient than backups, and enable you to protect each knowledge base using a different protection policy. Some of the scenarios where snapshots can be useful are:

- Accidental data loss or corruption
- Recovering from incorrect data being ingested into the knowledge base
- Testing different data sources or chunking strategies, and quickly reverting when the testing is complete

Take a snapshot of a knowledge base volume

You can save the state of a knowledge base by taking a manual snapshot of the knowledge base volume.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to protect.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

4. Select the **Actions** menu and select **Snapshot > Create new snapshot**.
5. Optionally, select **Define snapshot name** and enter a custom name for the snapshot.

   Defining a custom name can help you better determine the contents of a snapshot if you need to restore it in the future.

6. Select **Create**.

   A snapshot of the knowledge base is created.

Restore a snapshot of a knowledge base volume

You can restore a manual or scheduled snapshot of a knowledge base volume at any time.

You cannot restore a snapshot using the Generative AI workloads UI if the database stored on the volume is corrupt or has been deleted. As a workaround, you can restore the snapshot using the ONTAP CLI on the ONTAP cluster where the volume is hosted.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to restore.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

4. Select the **Actions** menu and select **Snapshot > Restore snapshot**.

   The snapshot selection dialog appears, where you can see a list of the snapshots that have been created for this knowledge base.

5. (Optional) Deselect the **Pause running and scheduled scans after restoring the snapshot** option if you want scheduled and currently running data source scans to continue after the snapshot is restored.

   This option is enabled by default to ensure that a scan doesn't happen while the knowledge base is in a partially restored state, or that a scan doesn't update a freshly restored knowledge base with older data.

6. Select the snapshot you want to restore from the list.
7. Select **Restore**.

Clone a knowledge base

You can create a new knowledge base from a knowledge base snapshot. This is useful if the original knowledge base is corrupted or lost.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to restore.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

4. Select the **Actions** menu and select **Snapshot > Clone knowledge base**.

   The clone dialog appears.

5. Optionally, deselect the **Pause running and scheduled scans after cloning the snapshot** option if you want scheduled and currently running data source scans to continue after the snapshot is cloned.

   This option is enabled by default to ensure that a scan doesn't happen while the knowledge base is in a partially restored state, or that a scan doesn't update a freshly restored knowledge base with older data.

6. Select the snapshot you want to clone from the list.
7. Select **Continue**.
8. Enter a name for the new knowledge base.
9. Choose a filesystem SVM and volume name for the new knowledge base.
10. Select **Clone**.

Add additional data sources to a knowledge base

You can embed additional data sources in your knowledge base to populate it with additional organization data.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base where you want to add the data source.
3. Select ••• and select **Add data source**.
4. Select the type of data source you want to add:
   - Add FSx for ONTAP file system (use files from an existing FSx for ONTAP volume)
   - Add file system (use files from a generic SMB or NFS share)

Unresolved directive in knowledge-base/manage-knowledgebase.adoc - include::_include/add-data-source-kb.adoc[]

Result

The data source is integrated into your knowledge base.

Synchronize your data sources with a knowledge base

Data sources are synchronized with the associated knowledge base automatically once a day so that any data source changes are reflected in the chatbot. If you make changes to any of your data sources and you'd like to synchronize the data immediately, you can perform an on-demand synchronization.

Syncing is incremental, so Amazon Bedrock only processes the objects in your data sources that have been added, modified, or deleted since the last sync.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to synchronize.
3. Select ••• and select **Manage knowledge base**.
4. Select the **Actions** menu and select **Scan now**.

   You'll see a message that your data sources are being scanned, and a final message when the scan is complete.

Result

The knowledge base is synchronized with the attached data sources and any active chatbot will start using the newest information from your data sources.

Pause or resume a scheduled synchronization

If you want to pause or resume the next synchronization (scan) of the data sources, you can do so at any time. You might need to pause the next scheduled synchronization if you are going to make changes to a data source and don't want the synchronization happening during the change window.

Steps

1. Log in to workload factory using one of the console experiences.

2. From the Knowledge bases & Connectors tab, select the knowledge base for which you want to pause or resume scans.
3. Select ••• and select **Manage knowledge base**.
4. Select the **Actions** menu and select **Scan > Pause scheduled scan** or **Scan > Resume scheduled scan**.

   You'll see a message that the next scheduled scan has either been paused or resumed.

Evaluate chat models before creating a knowledge base

You can evaluate the available foundational chat models before creating a knowledge base so you can see which model works best for your implementation. Since model support varies by AWS region, refer to this AWS documentation page to verify which models you can use in the regions where you plan to deploy your knowledge base.

> This functionality is available only when no knowledge bases have been created — when no knowledge bases exist in the Knowledge bases inventory page.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, you'll see the option to select the chat model on the right side of the page for the Chatbot.
3. Select the chat model from the list and enter a set of questions in the prompt area to see how the chatbot responds.
4. Try multiple models to see which model is best for your implementation.

Result

Use that chat model when you create your knowledge base.

Unpublish your knowledge base

After you've published your knowledge base so that it can be integrated with a chatbot application, you can unpublish it if you want to disable the chatbot application from accessing the knowledge base.

Unpublishing the knowledge base stops any chat applications from working. The unique API endpoint at which the knowledge base was accessible is disabled.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to unpublish.
3. Select ••• and select **Manage knowledge base**.

   This page displays the published status, embedding status of the data sources, embedding mode, and the list of all embedded data sources.

4. Select the **Actions** menu and select **Unpublish**.

Result

The knowledge base is disabled and is no longer accessible by a chatbot application.

Delete a knowledge base

If you no longer need a knowledge base, you can delete it. When you delete a knowledge base, it is removed from workload factory and the volume that contains the knowledge base is deleted. Any applications or chatbots that are using the knowledge base will stop working. Deleting a knowledge base is not reversible.

When you delete a knowledge base, you should also disassociate the knowledge base from any agents it is associated with to fully delete all resources associated with the knowledge base.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases inventory page, select the knowledge base that you want to delete.
3. Select ••• and select **Manage knowledge base**.
4. Select the **Actions** menu and select **Delete knowledge base**.
5. In the Delete knowledge base dialog, confirm that you want to delete it and select **Delete**.

Result

The knowledge base is removed from workload factory and its associated volume is deleted.

Manage Amazon Q Business connectors

After you create a connector for Amazon Q Business, you can view the connector details, modify the connector, integrate additional data sources, or delete the connector.

View information about a connector

You can view information about the settings for a connector and the data sources that are integrated.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the workload factory navigation menu, select **AI**.
3. Select the connector that you want to view.
4. To view connector details, select ••• and select **Manage connector**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

   The **Actions** menu enables you to manage the connector if you want to make any changes.

Edit a connector

You can update a connector by changing some settings, or you can add or remove data sources.

Each time you add, modify, or remove data sources from the connector, GenAI needs to send the data source information to Amazon Q Business so that it is re-indexed. Syncing is incremental, so Amazon Q Business only processes the objects in your FSx for ONTAP volume that have been added, modified, or deleted since the last sync.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases & Connectors inventory page, select the connector that you want to update.
3. Select ••• and select **Manage connector**.

   This page displays the published status, embedding status of the data sources, embedding mode, the list of all embedded data sources, and more.

4. Select the **Actions** menu and select **Edit connector**.
5. In the Edit connector page, you can change the connector name, description, embedding model, data guardrails enablement, and the snapshot policy used for the volume that contains the connector.

   Every data source scan, which includes embedding, incurs a cost. If you enable data guardrails after a connector has been created, then the data source gets scanned again and incurs costs.

6. Select **Save** after you have made changes.

Add additional data sources to a connector

You can embed additional data sources in your connector to populate it with additional organization data.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases & Connectors inventory page, select the connector where you want to add the data source.
3. Select ••• and select **Add data source**.
4. Select the type of data source you want to add:
   - Add FSx for ONTAP file system (use files from an existing FSx for ONTAP volume)
   - Add file system (use files from a generic SMB or NFS share)

Unresolved directive in connector/manage-connector.adoc - include::_include/add-data-source-kb.adoc[]

Result

The data source is integrated into your connector.

Synchronize your data sources with a connector

Data sources are synchronized with the associated connector automatically once a day so that any data source changes are reflected in Amazon Q Business. If you make changes to any of your data sources and you'd like to synchronize (scan) the data immediately, you can perform an on-demand synchronization.

Syncing is incremental, so Amazon Q Business only processes the objects in your data sources that have been added, modified, or deleted since the last sync.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases & Connectors tab, select the connector that you want to synchronize.
3. Select ••• and select **Manage connector**.

4. Select the **Actions** menu and select **Scan now**.

   You'll see a message that your data sources are being scanned, and a final message when the scan is complete.

Result

The connector is synchronized with the attached data sources and Amazon Q Business will start using the newest information from your data sources.

Pause or resume a scheduled synchronization

If you want to pause or resume the next synchronization (scan) of the data sources, you can do so at any time. You might need to pause the next scheduled synchronization if you are going to make changes to a data source and don't want the synchronization happening during the change window.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the connector inventory page, select the connector for which you want to pause or resume scans.
3. Select ••• and select **Manage connector**.
4. Select the **Actions** menu and select **Scan > Pause scheduled scan** or **Scan > Resume scheduled scan**.

   You'll see a message that the next scheduled scan has either been paused or resumed.

Delete a connector

If you no longer need a connector, you can delete it. When you delete a connector, it is removed from workload factory and the volume that contains the connector is deleted. Deleting a connector is not reversible.

When you delete a connector, you should also disassociate the connector from any agents it is associated with to fully delete all resources associated with the connector.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the Knowledge bases & Connectors inventory page, select the connector that you want to delete.
3. Select ••• and select **Manage connector**.
4. Select the **Actions** menu and select **Delete connector**.
5. In the Delete connector dialog, confirm that you want to delete it and select **Delete**.

Result

The connector is removed from workload factory and its associated volume is deleted.

Manage GenAI data sources

After you create a knowledge base or a connector using data sources on your FSx for ONTAP file system, you can view the data source details, update or change the data source contents, edit data source settings, or delete the data source.

View information about a data source

You can view information about the contents of a data source and you can view its embedding status with the knowledge base or connector. Since data sources are associated with a knowledge base or connector, you'll need to choose the knowledge base or connector first before you can view the data source details.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the workload factory navigation menu, select **AI**.
3. Select the knowledge base or connector where the data source resides, and then select ••• and select **Manage knowledge base** or **Manage connector**.

   The bottom part of the page lists the associated data sources.
4. Expand each row by selecting the ⌄ to view detailed information about each data source, such as the FSx for ONTAP file system, the volume, and the path where the data source resides.

   It also lists the embedding information and whether that data source is currently embedded in the knowledge base or connector.

Edit data source settings

You can edit information about a data source that you've integrated with a knowledge base or connector. Most of the information is fixed after you've added a data source, but you can make changes to the some of the configuration (such as chunking definition or permission awareness).

Steps

1. Log in to workload factory using one of the console experiences.
2. From the knowledge bases inventory page, select the knowledge base where the data source resides, and then select ••• and select **Manage knowledge base**.

   The bottom part of the page lists the data sources that are part of this knowledge base.
3. In the row of the data source that you want to edit, select ••• and select **Edit data source**.
4. In the Edit data source page, select ⌄ to expand the row for chunk definition.
5. Update the settings for the chunking strategy and configuration, and permission awareness (for SMB volumes) and select **Save**.

Result

The data source settings are updated and the AI system synchronizes the data source so that it is re-indexed to the knowledge base.

Update the contents of an existing data source

You can change the contents of a data source at any time to add or update your organizational data. If this data source is being actively used in a knowledge base, you must sync the data source so that it is re-indexed to the knowledge base. Syncing is incremental, so Amazon Bedrock only processes the objects in your FSx for ONTAP volume that have been added, modified, or deleted since the last sync.

Data sources are synchronized with the knowledge base automatically once a day so that any data source changes are reflected in the chatbot. If you make changes to a data source and you'd like to synchronize the

data immediately, you can perform an on-demand synchronization.

Delete a data source

If you no longer need a data source to be part of your knowledge base, you can delete it.

Steps

1. Log in to workload factory using one of the console experiences.
2. From the knowledge base inventory page, select the knowledge base where the data source resides, and then select ••• and select **Manage knowledge base**.

   The bottom part of the page lists the data sources that are part of this knowledge base.
3. In the row of the data source that you want to delete, select ••• and select **Delete data source**.
4. In the Delete data source dialog, confirm that you want to delete it and select **Confirm**.

Result

The data source is removed from the knowledge base and the AI system removes the indexed information about this data source from the knowledge base. Any information from that data source will no longer be available to chatbots that are using the knowledge base.

Monitor workload operations with Tracker in BlueXP workload factory

Monitor and track the execution of workload operations and monitor task progress with Tracker in BlueXP workload factory.

About this task

Workload factory provides Tracker, a monitoring feature, so you can monitor and track the progress and status of workload operations, review details for operation tasks and subtasks, and diagnose any issues or failures.

Several actions are available in Tracker. You can filter jobs by time frame (last 24 hours, 7 days, 14 days, or 30 days), workload, status, and user; find jobs using the search function; and download the jobs table as a CSV file. You can refresh Tracker at any time, and quickly retry a failed operation or edit parameters for a failed operation and try the operation again.

Tracker supports two levels of monitoring depending on the operation. Each task, such as file system deployment, displays the task description, status, start time, task duration, user, region, proxy resource, task ID, and all related sub tasks. You can view API responses to understand what happened during the operation.

Tracker task levels with examples

- Level 1 (task): Tracks file system deployment.
- Level 2 (sub task): Tracks the sub tasks related to the file system deployment.

Operation status

Operation status in Tracker is as follows *in progress*, *success*, and *failed*.

Operation frequency

Operation frequency is based on the job type and the job schedule.

Events retention

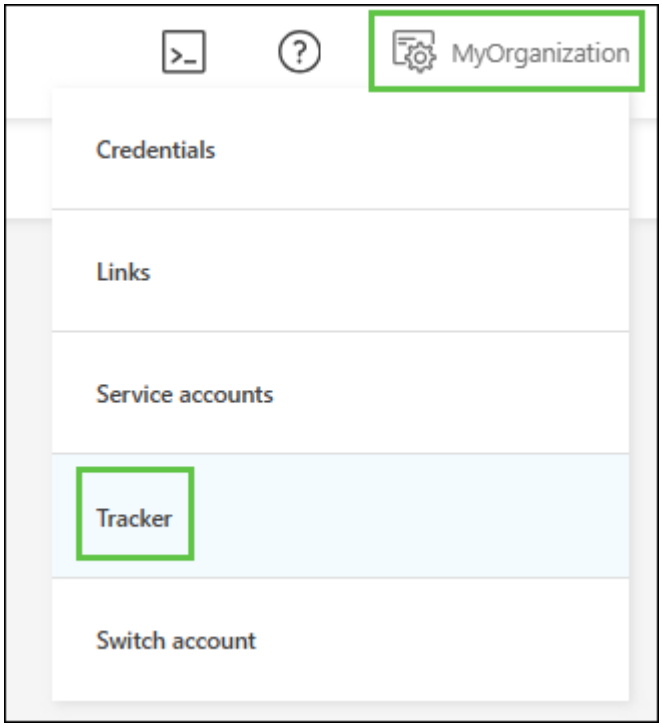
Events are retained in the user interface for 30 days.

Track and monitor operations

Track and monitor operations in BlueXP with Tracker.

Steps

1. Log in using one of the console experiences.
2. In the workload, select the account settings menu and then select **Tracker**.



3. In the Tracker tab, use the filters or search to narrow job results. You can also download a jobs report.

View API request

View the API request in the Codebox for a task in Tracker.

Steps

1. In Tracker, select a task.
2. Select the three-dot menu and then select **View API request**.

Retry a failed operation

Retry a failed operation in Tracker. You can also copy the error message of a failed operation.

You can retry a failed operation up to 10 times.

Steps

1. In Tracker, select a failed operation.

2. Select the three-dot menu and then select **Retry**.

Result

The operation is re-initiated.

Edit and retry a failed operation

Edit the parameters of the failed operation and retry the operation outside Tracker.

Steps

1. In Tracker, select a failed operation.
2. Select the three-dot menu and then select **Edit and retry**.

   You are redirected to the operation page where you can edit the parameters and retry the operation.

Result

The operation is re-initiated. Go to Tracker to view the status of the operation.

Knowledge and support

Register for support for BlueXP workload factory for GenAI

Support registration is required to receive technical support specific to BlueXP workload factory and its storage solutions and services. You must register for support from the BlueXP console, which is a separate web-based console from workload factory.

Registering for support does not enable NetApp support for a cloud provider file service. For technical support related to a cloud provider file service, its infrastructure, or any solution using the service, refer to "Getting help" in the workload factory documentation for that product.

Amazon FSx for ONTAP

Support registration overview

Registering your account ID support subscription (your 20 digit 960xxxxxxxx serial number located on the Support Resources page in BlueXP) serves as your single support subscription ID. Each BlueXP account-level support subscription must be registered.

Registering enables capabilities like opening support tickets and automatic case generation. Registration is completed by adding NetApp Support Site (NSS) accounts to BlueXP as described below.

Register your account for NetApp support

To register for support and activate support entitlement, one user in your account must associate a NetApp Support Site account with their BlueXP login. How you register for NetApp support depends on whether you already have a NetApp Support Site (NSS) account.

Existing customer with an NSS account

If you’re a NetApp customer with an NSS account, you simply need to register for support through BlueXP.

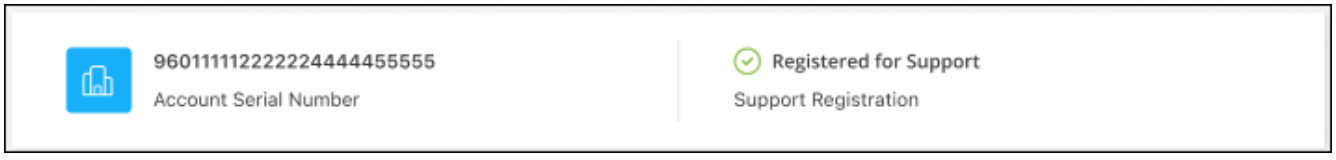
Steps

1. In the upper right of the workload factory console, select **Help > Support**.

   Selecting this option opens the BlueXP console a new browser tab and loads the Support dashboard.

2. In the upper right of the BlueXP console, select the Settings icon, and select **Credentials**.
3. Select **User Credentials**.
4. Select **Add NSS credentials** and follow the NetApp Support Site (NSS) Authentication prompt.
5. To confirm that the registration process was successful, select the Help icon, and select **Support**.

   The **Resources** page should show that your account is registered for support.



Note that other BlueXP users will not see this same support registration status if they have not associated a NetApp Support Site account with their BlueXP login. However, that doesn't mean that your BlueXP account is not registered for support. As long as one user in the account has followed these steps, then your account has been registered.

Existing customer but no NSS account

If you're an existing NetApp customer with existing licenses and serial numbers but *no* NSS account, you need to create an NSS account and associate it with your BlueXP login.

Steps

1. Create a NetApp Support Site account by completing the NetApp Support Site User Registration form
   a. Be sure to select the appropriate User Level, which is typically **NetApp Customer/End User**.
   b. Be sure to copy the BlueXP account serial number (960xxxx) used above for the serial number field. This will speed up the account processing.
2. Associate your new NSS account with your BlueXP login by completing the steps under Existing customer with an NSS account.

Brand new to NetApp

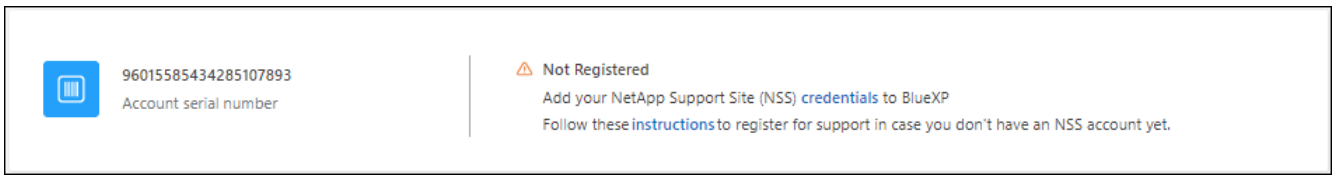
If you are brand new to NetApp and you don't have an NSS account, follow each step below.

Steps

1. In the upper right of the workload factory console, select **Help > Support**.

   Selecting this option opens the BlueXP console a new browser tab and loads the Support dashboard.

2. Locate your account ID serial number from the Support Resources page.



3. Navigate to NetApp's support registration site and select **I am not a registered NetApp Customer**.
4. Fill out the mandatory fields (those with red asterisks).
5. In the **Product Line** field, select **Cloud Manager** and then select your applicable billing provider.
6. Copy your account serial number from step 2 above, complete the security check, and then confirm that you read NetApp's Global Data Privacy Policy.

   An email is immediately sent to the mailbox provided to finalize this secure transaction. Be sure to check your spam folders if the validation email doesn't arrive in few minutes.

7. Confirm the action from within the email.

   Confirming submits your request to NetApp and recommends that you create a NetApp Support Site account.

8. Create a NetApp Support Site account by completing the NetApp Support Site User Registration form
   a. Be sure to select the appropriate User Level, which is typically **NetApp Customer/End User**.

b. Be sure to copy the account serial number (960xxxx) used above for the serial number field. This will speed up the account processing.

After you finish

NetApp should reach out to you during this process. This is a one-time onboarding exercise for new users.

Once you have your NetApp Support Site account, associate the account with your BlueXP login by completing the steps under Existing customer with an NSS account.

GenAI troubleshooting

Learn how to work around some common problems you might encounter.

Common issues and solutions

If you have one of these issues, you can use the steps in the Workaround column to try to resolve it.

Area	Issue	Cause	Workaround
Deployment	Deployment fails because the volume already exists.	BlueXP workload factory for GenAI needs to create a new volume during the deployment process, but a volume already exists using the name you have specified.	Specify a unique name to use for the new volume, and try deploying again.

<table>
<tr><th>Area</th><th>Issue</th><th>Cause</th><th>Workaround</th></tr>
<tr><td>Deployment</td><td>The deployment fails because BlueXP workload factory for GenAI is unable to mount the volume.</td><td>One or more of the inbound ports required for FSx for NetApp ONTAP are closed or filtered.</td><td>Open the following inbound ports:
<table><tr><th>Protocol</th><th>Port</th><th>Purpose</th></tr><tr><td>All ICMP</td><td>All</td><td>Pinging the instance</td></tr><tr><td>HTTPS</td><td>443</td><td>Access from the Connector to fsxadmin management LIF to send API calls to FSx</td></tr><tr><td>SSH</td><td>22</td><td>SSH access to the IP address of the cluster management LIF or a node management LIF</td></tr><tr><td>TCP</td><td>111</td><td>Remote procedure call for NFS</td></tr><tr><td>TCP</td><td>139</td><td>NetBIOS service session for CIFS</td></tr><tr><td>TCP</td><td>161-162</td><td>Simple network management protocol</td></tr><tr><td>TCP</td><td>445</td><td>Microsoft SMB/CIFS over TCP with NetBIOS framing</td></tr><tr><td>TCP</td><td>635</td><td>NFS mount</td></tr><tr><td>TCP</td><td>749</td><td>Kerberos</td></tr><tr><td>TCP</td><td>2049</td><td>NFS server daemon</td></tr><tr><td>TCP</td><td>3260</td><td>iSCSI access through the iSCSI data LIF</td></tr><tr><td>TCP</td><td>4045</td><td>NFS lock daemon</td></tr><tr><td>TCP</td><td>4046</td><td>Network status monitor for NFS</td></tr><tr><td>TCP</td><td>10000</td><td>Backup using NDMP</td></tr><tr><td>TCP</td><td>11104</td><td>Management of intercluster communication sessions for SnapMirror</td></tr><tr><td>TCP</td><td>11105</td><td>SnapMirror data transfer using intercluster LIFs</td></tr></table></td></tr>
</table>

Area	Issue	Cause	Workaround
Maintenance	The AI engine fails to start, and you see the error "AI engine instance error" on the **Knowledge bases** page.	The AI engine instance was corrupted or doesn’t exist.	Select the **Rebuild** button. BlueXP workload factory for GenAI rebuilds the infrastructure and displays the rebuild progress. When complete, your knowledge bases are reconnected to the rebuilt infrastructure and the list of knowledge bases is displayed.
Maintenance	The AI engine fails to start, and you see the error "The GenAI engine instance is stopped" on the **Knowledge bases** page.	The AI engine instance is not running.	Use the AWS Management Console or the AWS CLI to start the AI engine instance.
Maintenance	The AI engine fails to start, and you see the error "The GenAI engine server is not responding" on the **Knowledge bases** page.	The AI engine instance is not responding.	Use the following recovery steps: **Steps** 1. Modify the GenAI engine instance security group to enable SSH access to the GenAI engine instance. 2. Log in to the instance using SSH. 3. Run the following command: `docker-compose up`

Area	Issue	Cause	Workaround
Maintenance	The backend Docker instance used by BlueXP workload factory for GenAI failed to start.	The volume was deleted and the EC2 instance was restarted.	Use the following recovery steps: **Steps** 1. Create a new volume on FSx for NetApp ONTAP. For example, the volume name can be `netapp_ai` and the volume path can be `/netapp_ai`. 2. SSH to the Amazon EC2 instance. 3. List the volumes: ``` docker volume list ``` 4. Remove the old volume: ``` docker volume rm ec2-user_persistent_folder ``` 5. Open the `docker-compose.yml` file using a text editor. 6. In the `volumes` section, change the device path to the new volume path. For example: ``` volumes: persistent_folder: driver_opts: type: 'nfs' o: "addr=svm-0df66b96a890d8a72.\ fs-0d673008aaca12bc3.\ fsx.us-east-1.amazonaws.com,nolock,soft,rw" device: ':/netapp_ai' # Path to new volume ```
Maintenance	The backend Docker instance used by BlueXP workload factory for GenAI failed to start.	The root volume was deleted.	Create a volume with a name and path, and then restart the backend Docker instance from Amazon EC2.

Area	Issue	Cause	Workaround
Maintenance	The backend Docker instance used by BlueXP workload factory for GenAI failed to start.	The root volume was deleted.	Create a volume with a name and path, and then restart the backend Docker instance from Amazon EC2.

Get help with BlueXP workload factory for GenAI

NetApp provides support for BlueXP workload factory and its cloud services in a variety of ways. Extensive free self-support options are available 24x7, such as knowledgebase (KB) articles and a community forum. Your support registration includes remote technical support via web ticketing.

Get support for FSx for ONTAP

For technical support related to FSx for ONTAP, its infrastructure, or any solution using the service, refer to "Getting help" in the workload factory documentation for that product.

Amazon FSx for ONTAP

To receive technical support specific to Workload Factory and its storage solutions and services, use the support options described below.

Use self-support options

These options are available for free, 24 hours a day, 7 days a week:

- Documentation

  The workload factory documentation that you're currently viewing.

- Knowledge base

  Search through the workload factory knowledge base to find helpful articles to troubleshoot issues.

- Communities

  Join the workload factory community to follow ongoing discussions or create new ones.

Create a case with NetApp support

In addition to the self-support options above, you can work with a NetApp Support specialist to resolve any issues after you activate support.

Before you get started

To use the **Create a Case** capability, you must first register for support. associate your NetApp Support Site credentials with your workload factory login. Learn how to register for support.

Steps

1. In the upper right of the workload factory console, select **Help > Support**.

   Selecting this option opens the BlueXP console a new browser tab and loads the Support dashboard.

2. On the **Resources** page, choose one of the available options under Technical Support:

   a. Select **Call Us** if you'd like to speak with someone on the phone. You'll be directed to a page on netapp.com that lists the phone numbers that you can call.

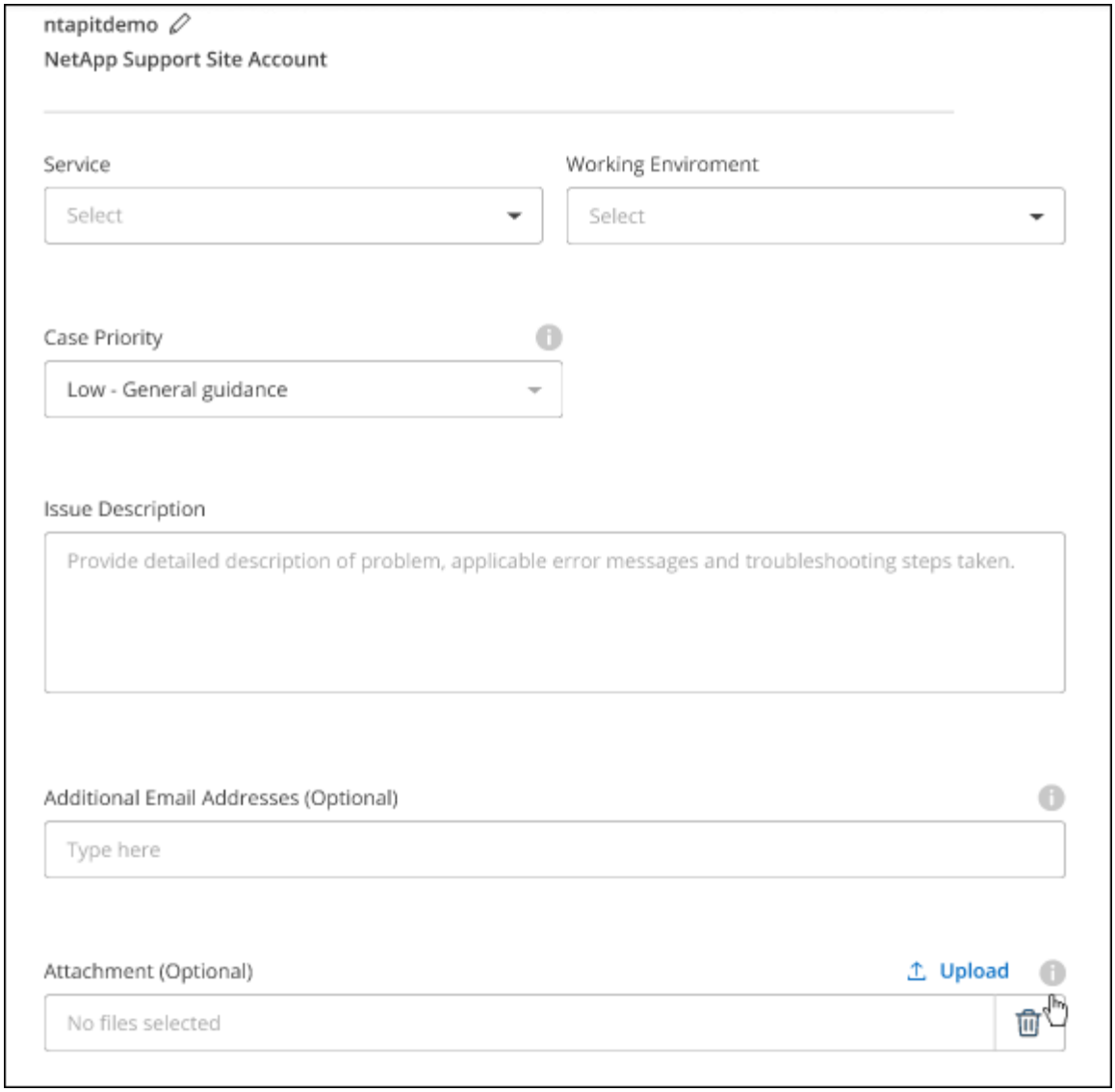
   b. Select **Create a Case** to open a ticket with a NetApp Support specialist:

      - **Service**: Select **Workload Factory**.
      - **Case Priority**: Choose the priority for the case, which can be Low, Medium, High, or Critical.

        To learn more details about these priorities, hover your mouse over the information icon next to the field name.

      - **Issue Description**: Provide a detailed description of your problem, including any applicable error messages or troubleshooting steps that you performed.
      - **Additional Email Addresses**: Enter additional email addresses if you'd like to make someone else aware of this issue.
      - **Attachment (Optional)**: Upload up to five attachments, one at a time.

        Attachments are limited to 25 MB per file. The following file extensions are supported: txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx, and csv.



After you finish

A pop-up will appear with your support case number. A NetApp Support specialist will review your case and get back to you soon.

For a history of your support cases, you can select **Settings > Timeline** and look for actions named "create support case." A button to the far right lets you expand the action to see details.

It's possible that you might encounter the following error message when trying to create a case:

"You are not authorized to Create a Case against the selected service"

This error could mean that the NSS account and the company of record it's associated with is not the same company of record for the BlueXP account serial number (ie. 960xxxx) or the working environment serial number. You can seek assistance using one of the following options:

- Use the in-product chat
- Submit a non-technical case at https://mysupport.netapp.com/site/help

Manage your support cases (Preview)

You can view and manage active and resolved support cases directly from BlueXP. You can manage the cases associated with your NSS account and with your company.

Case management is available as a Preview. We plan to refine this experience and add enhancements in upcoming releases. Please send us feedback by using the in-product chat.

Note the following:

- The case management dashboard at the top of the page offers two views:
  - The view on the left shows the total cases opened in the past 3 months by the user NSS account you provided.
  - The view on the right shows the total cases opened in the past 3 months at your company level based on your user NSS account.

  The results in the table reflect the cases related to the view that you selected.
- You can add or remove columns of interest and you can filter the contents of columns like Priority and Status. Other columns provide just sorting capabilities.

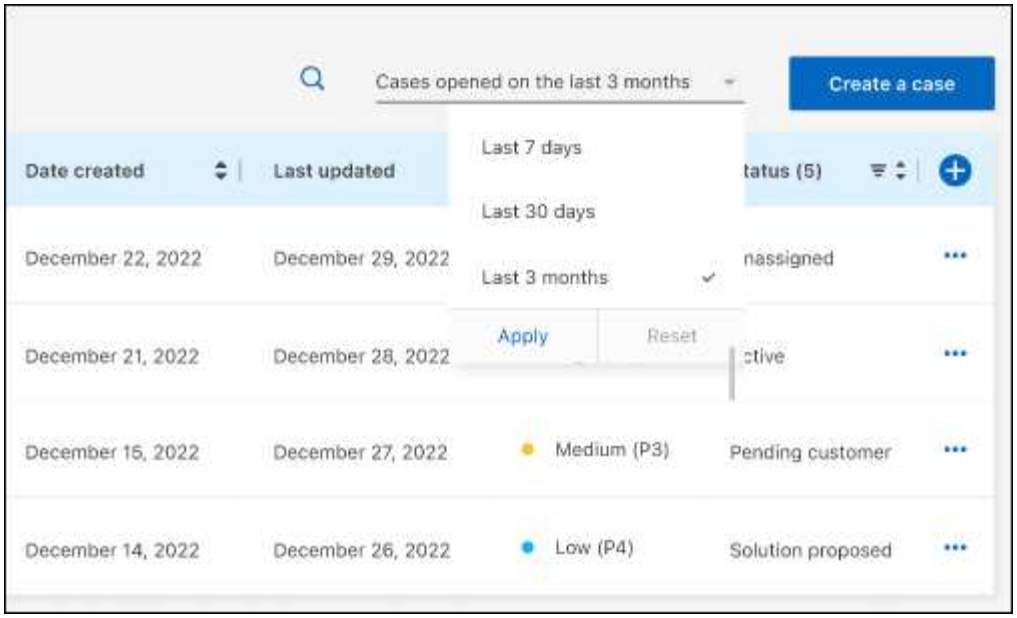
  View the steps below for more details.
- At a per-case level, we offer the ability to update case notes or close a case that is not already in Closed or Pending Closed status.

Steps

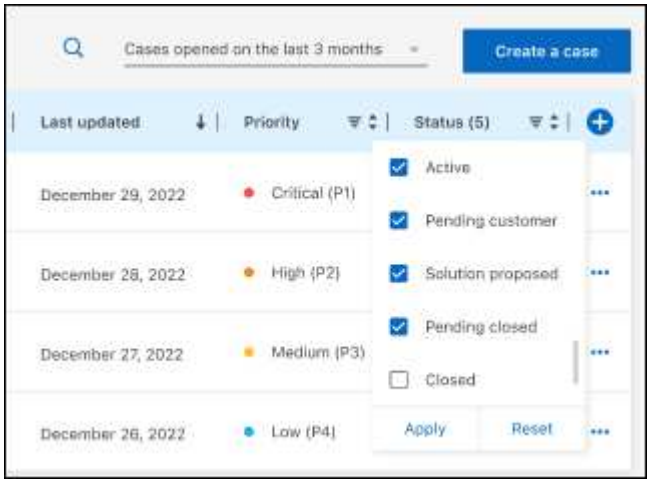
1. In the upper right of the workload factory console, select **Help > Support**.

   Selecting this option opens the BlueXP console a new browser tab and loads the Support dashboard.
2. Select **Case Management** and if you're prompted, add your NSS account to BlueXP.

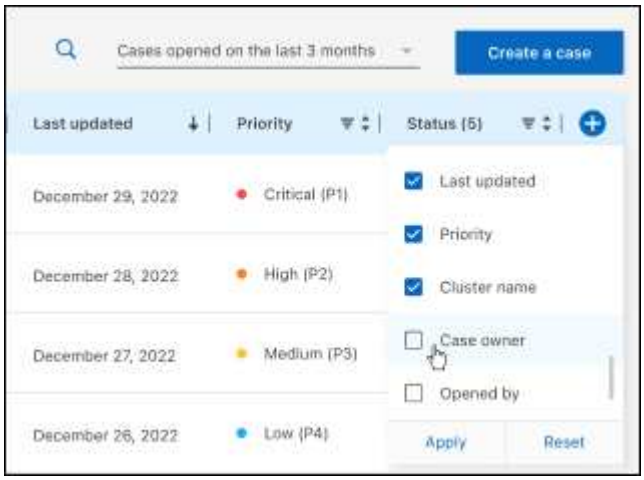
   The **Case management** page shows open cases related to the NSS account that is associated with your BlueXP user account. This is the same NSS account that appears at the top of the **NSS management** page.
3. Optionally modify the information that displays in the table:
   - Under **Organization's cases**, select **View** to view all cases associated with your company.
   - Modify the date range by choosing an exact date range or by choosing a different time frame.



- Filter the contents of the columns.



- Change the columns that appear in the table by selecting ⊕ and then choosing the columns that you'd like to display.

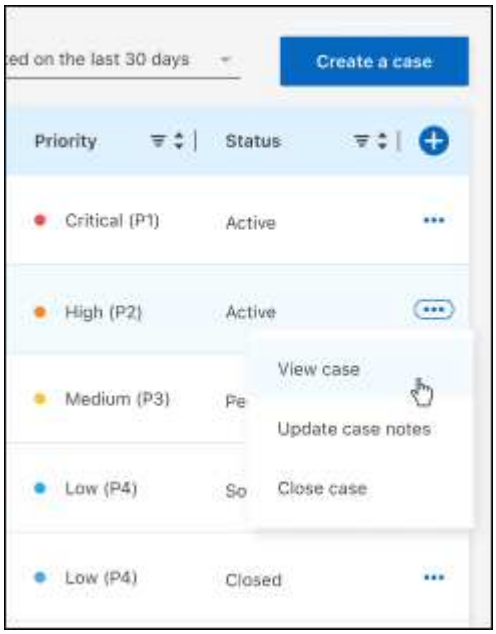


4. Manage an existing case by selecting ••• and selecting one of the available options:

   - **View case**: View full details about a specific case.
   - **Update case notes**: Provide additional details about your problem or select **Upload files** to attach up to a maximum of five files.

     Attachments are limited to 25 MB per file. The following file extensions are supported: txt, log, pdf, jpg/jpeg, rtf, doc/docx, xls/xlsx, and csv.

   - **Close case**: Provide details about why you're closing the case and select **Close case**.



BlueXP workload factory for GenAI legal notices

Legal notices provide access to copyright statements, trademarks, patents, and more.

Copyright

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

Trademarks

NETAPP, the NETAPP logo, and the marks listed on the NetApp Trademarks page are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

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

Patents

A current list of NetApp owned patents can be found at:

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

Privacy policy

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

Open source

Notice files provide information about third-party copyright and licenses used in NetApp software.

BlueXP workload factory

Copyright information

Trademark information