Release notes

GenAI

NetApp
October 06, 2025

This PDF was generated from https://docs.netapp.com/us-en/workload-genai/whats-new.html on October 06, 2025. Always check docs.netapp.com for the latest.

Table of Contents

Release notes

What's new with NetApp Workload Factory for GenAI

Learn what's new with the Generative AI workloads capability of Workload Factory.

05 October 2025

BlueXP workload factory now NetApp Workload Factory

BlueXP has been renamed and redesigned to better reflect the role it has in managing your data infrastructure. As a result, BlueXP workload factory has been renamed to NetApp Workload Factory.

Support for adding generic NFS/SMB data sources in NetApp connectors for Amazon Q Business

Using the Workload Factory API, you can now add a data source from a generic NFSv3, NFSv4, or SMB share to a NetApp Connector for Amazon Q Business. This enables you to include files that are stored on volumes hosted by filesystems other than Amazon FSx for NetApp ONTAP.

Create a NetApp Connector for Amazon Q Business

Add data sources to a connector

Advanced chat configuration for knowledge bases

You can now configure advanced chat settings that are applicable to the chat model for the knowledge base such as the response length, temperature, reasoning settings, and more. Some of these settings, such as recency and modification time settings, advanced retrieval settings, and system prompt are available only using the Workload Factory API.

Create a GenAI knowledge base

Inference type selection now supported for embedding, chat, and reranking models

If your chosen embedding, chat, or reranking model has inference settings, you can now select an inference type. This enables you to better tune the chatbot performance and resource requirements to your needs.

Create a GenAI knowledge base

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 NetApp 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 NetApp 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 NetApp 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

NetApp 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** menu 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 NetApp Console 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 NetApp Console 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.

Copyright information

Copyright © 2025 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.