NETAPP BLUEXP WORKLOAD FACTORY FOR GENAI PROJECTS

THE CHALLENGE

In late 2022, the unveiling of ChatGPT sent shockwaves throughout the tech world, revealing the transformative potential of generative AI (GenAI). Large language models (LLMs) and chatbots began to reshape how organizations envisioned productivity and innovation. It quickly became evident that the real competitive edge lay not merely in leveraging more data, but in integrating publicly available information with proprietary insights to generate ultra-relevant, high-quality outcomes. Retrieval-augmented generation (RAG) emerged as the right approach—enriching foundational LLMs with an organization's exclusive data to create customized, context-aware responses.

However, deploying RAG effectively is complex. Enterprises must manage enormous datasets while facing workforce constraints for continuous evaluation, indexing, and contextualization. Strict data privacy, sensitivity, and governance standards add further complexity. Ultimately, RAG's success depends on seamlessly integrating internal expertise with external data to transform raw information into secure, actionable intelligence that drives innovation.

THE OPPORTUNITY

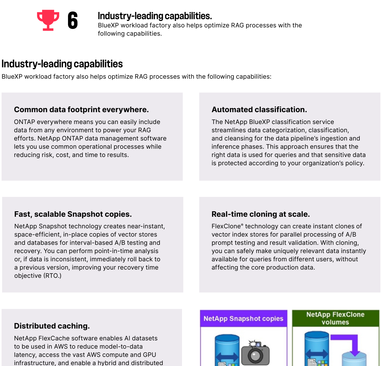
NetApp® BlueXP™ workload factory can power your GenAI efforts by helping you set up the right foundation for RAG operations such as natural language search, summarization, and chatting. By automatically connecting Amazon FSx for NetApp ONTAP with Amazon Bedrock via API, BlueXP workload factory enables you to innovate faster and smarter by securely augmenting public data with your proprietary data during inferencing runs. This ability gives you increased relevance and pinpoint accuracy from your LLM workloads, and it unlocks the full potential of using your NetApp ONTAP® resident data in your GenAI workflows.

THE SOLUTION

Workload Factory offers a seamless, secure solution for deploying a RAG-powered GenAI ecosystem by integrating FSx for ONTAP with Amazon Bedrock and Amazon Q Business through a dedicated data connector. Amazon Bedrock leverages fully managed Knowledge Bases to automatically ingest, index, and retrieve your data, embedding proprietary insights alongside public information. This integration enhances the relevance and precision of natural language search, summarization, and conversational AI outputs, as the Knowledge Bases dynamically feed curated content into the foundation models.

Simultaneously, Amazon Q Business extends enterprise-level querying capabilities via its robust data connector, ensuring that key organizational data is readily available for context-aware responses. Workload Factory not only manages the complexities of data ingestion—from handling diverse FSx for ONTAP sources and Windows file permissions to creating, storing, and retrieving embeddings—but also incorporates advanced sensitive data filtering. This proactive measure identifies and excludes sensitive or regulated information before ingestion, ensuring compliance and safeguarding privacy. With integrated workflow automation and a comprehensive RAG reference architecture document, BlueXP workload factory empowers developers, SREs, and cloud/data administrators to build expansive, secure knowledge bases, fully leveraging your data assets to fuel innovative GenAI applications.

To further support GenAI initiatives, the BlueXP workload factory includes workflow automation, providing a detailed RAG reference architecture document, allowing customers to gain a deeper understanding of the setup, and offering guidance for customization to meet their specific requirements.



Benefits

1 Seamless GenAI infrastructure deployment and management.
Streamlines the setup and management of your RAG-powered GenAI ecosystem by automating the entire data workflow. Using Amazon Bedrock, proprietary FSx for ONTAP data is automatically ingested, indexed, and transformed into vectorized formats, enhancing natural language search, summarization, and conversational AI capabilities. At the same time, a robust data connector enables Amazon Q Business to ingest FSx for ONTAP data, both ensuring enterprise-level querying without the added vectorization step.

2 Enhanced security for GenAI applications.
Your organization can enforce strict data permissions for your proprietary data by leveraging ONTAP's integrated security controls, such as Microsoft Active Directory and Windows ACLs. All data, whether vectorized via Amazon Bedrock or ingested through Amazon Q Business, is securely stored in an encrypted AWS environment (account and VPC), while existing data protection mechanisms (backup, NetApp Snapshot™, SnapVault®, SnapMirror®) further protect your GenAI application from exposing sensitive data.

3 Robust proprietary data foundations.
Build expansive, domain-specific knowledge bases with ONTAP's industry-leading data management. Amazon Bedrock uses vectorized knowledge bases to enrich RAG workflows, ensuring precise and context-aware responses. Simultaneously, Amazon Q Business enables the seamless ingestion of FSx for ONTAP data via its connector, contributing to a comprehensive and responsive data ecosystem.

4 Efficient data mobility.
Leveraging NetApp FlexClone® and SnapMirror® technology, efficient data copies bring data closer to processing environments. This setup reduces inference latency and transfer costs, ensuring that data processed via Amazon Bedrock and ingested by Amazon Q Business is readily accessible and optimized for performance, benefiting large enterprises with distributed data estates.

5 Protect sensitive information.
Advanced classification using NetApp® BlueXP™ ensures sensitive or regulated information is identified and masked before ingestion. This proactive filtering guarantees that both Amazon Bedrock and Amazon Q Business processes handle data in compliance with privacy and governance standards, mitigating the risk of exposing sensitive information.

6 Industry-leading capabilities.
BlueXP workload factory also helps optimize RAG processes with the following capabilities.

Industry-leading capabilities

BlueXP workload factory also helps optimize RAG processes with the following capabilities:

Common data footprint everywhere.
ONTAP everywhere means you can easily include data from any environment to power your RAG efforts. NetApp ONTAP data management software lets you use common operational processes while reducing risk, cost, and time to results.

Automated classification.
The NetApp BlueXP classification service streamlines data categorization, classification, and cleansing for the data pipeline's ingestion and inference phases. This approach ensures that the right data is used for queries and that sensitive data is protected according to your organization's policy.

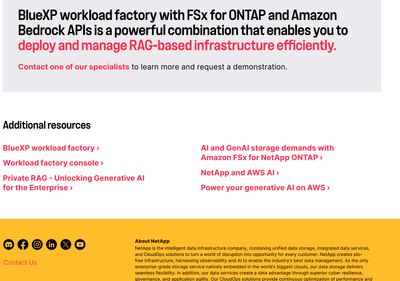
Fast, scalable Snapshot copies.
NetApp Snapshot technology creates near-instant, space-efficient, in-place copies of vector stores and databases for interval-based A/B testing and recovery. You can perform point-in-time analysis or, if data is inconsistent, immediately roll back to a previous version, improving your recovery time objective (RTO.)

Real-time cloning at scale.
FlexClone® technology can create instant clones of vector index stores for parallel processing of A/B prompt testing and result validation. With cloning, you can safely make uniquely relevant data instantly available for queries from different users, without affecting the core production data.

Distributed caching.
NetApp FlexCache software enables AI datasets to be used in AWS to reduce model-to-data latency, access the vast AWS compute and GPU infrastructure, and enable a hybrid and distributed GenAI infrastructure while keeping costs contained.

Use case: GenAI-powered Q&A chat application with Amazon Bedrock

Provide quick, context-aware, and personalized answers to employee user queries based on company data that the user has access to. This application enables platform and development teams to quickly architect, design, develop, test, and deploy a GenAI-powered chatbot that is linked to secure enterprise data sources hosted on FSx for ONTAP or on premises. Data is kept secure and centrally located on NetApp, while ensuring that chatbots take user permissions into consideration.



1. The ingestion service connects to source shares/exports and watches for changes.
2. The embedding service connects to an embedding model.
3. Vector embeddings and metadata are stored on an FSx for ONTAP volume.
4. The user (admin) is authenticated via Active Directory.
5. The user interacts with the source data through chat.
6. Relevant vectors from documents are retrieved and used as prompts for the language model.
7. The language/chat model constructs the response to the user.

GenAI-powered Q&A application with Amazon Q Business

Delivers rapid, context-aware, and personalized answers to employee inquiries by leveraging company data that users are authorized to access. With Amazon Q Business at its core, the platform utilizes a dedicated data connector to ingest data directly from FSx for ONTAP or on-premises systems.

Unlike solutions such as those using Bedrock, which utilize vectorized knowledge bases, Amazon Q Business ingests raw data. It focuses on streamlined enterprise query processing, allowing users to interact with proprietary data in real time while maintaining robust security and compliance.

BlueXP workload factory with FSx for ONTAP and Amazon Bedrock APIs is a powerful combination that enables you to deploy and manage RAG-based infrastructure efficiently.

Contact one of our specialists to learn more and request a demonstration.

Additional resources

BlueXP workload factory ›
Workload factory console ›
Private RAG - Unlocking Generative AI for the Enterprise ›
AI and GenAI storage demands with Amazon FSx for NetApp ONTAP ›
NetApp and AWS AI ›
Power your generative AI on AWS ›