Vector stores in Azure Machine Learning (preview)

Article • 07/18/2024

ⓘ Important

This feature is currently in public preview. This preview version is provided without a service-level agreement, and we don't recommend it for production workloads. Certain features might not be supported or might have constrained capabilities.

For more information, see Supplemental Terms of Use for Microsoft Azure Previews .

This article describes vector indexes in Azure Machine Learning that you can use to perform retrieval-augmented generation (RAG). A vector index stores embeddings that are numerical representations of *concepts* (data) converted to number sequences. Embeddings enable large language models (LLMs) to understand the relationships between the concepts. You can create vector stores to connect your data with LLMs like GPT-4, and retrieve the data efficiently.

Azure Machine Learning supports two vector stores that contain your supplemental data used in a RAG workflow:

Expand table

Vector store	Description	Features and usage
Faiss	Open source library	- Use local file-based store - Incur minimal costs - Support vector-only data - Support development and testing
Azure AI Search	Azure PaaS resource	- Store text data in search index

	- Host large number of indexes with single service - Support enterprise-level business requirements - Access hybrid information retrieval

The following sections explore considerations for working with these vector stores.

Faiss library

Faiss is an open source library that provides a local file-based store. The vector index is stored in the Azure storage account of your Azure Machine Learning workspace. To work with Faiss, you download the library and use it as a component of your solution. Because the index is stored locally, the costs are minimal.

You can use the Faiss library as your vector store and perform the following actions:

- Store vector data locally, with no costs for creating an index (only storage cost)
- Build and query an index in memory
- Share copies for individual use, and configure hosting of the index for an application
- Scale with underlying compute loading index

Azure AI Search

Azure AI Search (formerly Cognitive Search) is a dedicated Azure PaaS resource that you create in an Azure subscription. The resource supports information retrieval over your vector and textual data stored in search indexes. A prompt flow can create, populate, and query your vector data stored in Azure AI Search. A single search service can host a large number of indexes, which can be queried and used in a RAG pattern.

Here are some key points about using Azure AI Search for your vector store:

- Support enterprise level business requirements for scale, security, and availability.

- Access hybrid information retrieval. Vector data can coexist with nonvector data, which means you can use any of the features of Azure AI Search for indexing and queries, including hybrid search and semantic reranking.

- Keep in mind that vector support is in preview. Currently, vectors must be generated externally and then passed to Azure AI Search for indexing and query encoding. The prompt flow handles these transitions for you.

To use AI Search as a vector store for Azure Machine Learning, you must have a search service. After the service exists, and you grant access to developers, you can choose **Azure AI Search** as a vector index in a prompt flow. The prompt flow creates the index on Azure AI Search, generates vectors from your source data, sends the vectors to the index, invokes similarity search on AI Search, and returns the response.

Related content

- Create vector index in Azure Machine Learning prompt flow (preview)
- Vectors in Azure AI Search

Feedback

Was this page helpful? Yes No

Provide product feedback | Get help at Microsoft Q&A