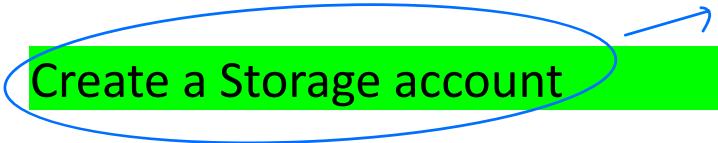


Prereqs for RAG with Azure AI Search

- 
- Azure
- i. Create a Storage account *acasdatastore*
 - ii. Create Embedding Deployment *embeddingacasrag*
 - iii. Create Chats Deployment *chatmodelrag*
 - iv. Create Azure AI Search Service *azureaisearchrag*
 - v. Using the chats Deployment
 - i. Upload the documents.
 - vi. Ingestion > preprocessing > Indexing

RAG

PREREQUISITES

1. Storage Account

Create a storage account ...

Basics **Advanced** **Networking** **Data protection** **Encryption** **Tags** **Review + create**

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#).

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * Q8551
Resource group * nadtest001 [Create new](#)

Instance details

Storage account name * tjdatalogue
Region * (US) East US [Deploy to an Azure Extended Zone](#)

Preferred storage type [Choose preferred storage type](#)
This helps us provide relevant guidance. It doesn't restrict your storage to this resource type. [Learn more](#)

Performance * Standard: Recommended for most scenarios (general-purpose v2 account)
 Premium: Recommended for scenarios that require low latency.

[Previous](#) [Next](#) [Review + create](#)

Redundancy

↳ Stores copies of Data.

2. Deploy Embedding Model

Deployments → Model

Model = 3 large / small / Ada.

Deploy text-embedding-ada-002

Deployment name * Model Name: text-embedding-ada-tj

Deployment type: Global Standard

Global Standard: Pay per API call with the highest rate limits. Learn more about [Global deployment types](#). Data might be processed globally, outside of the resource's Azure geography, but data storage remains in the AI resource's Azure geography. Learn more about [data residency](#).

Deployment details

Model version	AI resource
2	OpenAIAd01
Capacity	Resource location
120K tokens per minute (TPM)	East US
Content safety	Version upgrade policy
DefaultV2	Model version will not be automatically upgraded

[Customize](#) [Deploy](#) [Cancel](#)

Deployment info	
Name	Provisioning state
text-embedding-ada-tj	Succeeded
Deployment type	Created on
Global Standard	2025-10-01T18:53:17.7994573Z
Created by	Modified on
[REDACTED]	Oct 2, 2025 12:25 AM
soft.com	
Modified by	Version upgrade policy
[REDACTED]	Model version will not be automatically upgraded
Rate limit (Tokens per minute)	Rate limit (Requests per minute)
120,000	720
Model name	Model version
text-embedding-ada-002	2
Life cycle status	Date created
GenerallyAvailable	Apr 3, 2023 5:30 AM
Date updated	Model retirement date
Apr 3, 2023 5:30 AM	Apr 30, 2026 5:30 AM

3. Chat Completion Deployment

The screenshot shows the 'Chat playground' setup page. On the left, there's a sidebar with various options like Home, Get started, Model catalog, Playgrounds, Chat (which is highlighted with a pink circle), Assistants, Video, Audio, Images, Tools, Shared resources, Deployments, Quota, Guardrails + Controls, Risks + alerts, Data files, and Assistant vector stores. The main area is titled 'Setup' and has a 'Deployment' section. It shows a dropdown menu with 'my-first-gpt (version:2025-04-14)' selected. Below it, there's a text input field for giving model instructions and context, containing the placeholder 'You are an AI assistant that helps people find information.' At the bottom, there are buttons for 'Apply changes', 'Generate system prompt', and '+ Add section'. A handwritten note 'Chat deployment' is written next to the deployment section.

4. Create AI Search

Azure portal → AI Search

The screenshot shows the 'Create a search service' wizard. It has tabs for Basics, Scale, Networking, Tags, and Review + create. Under Basics, there are sections for Project details (Subscription: QBSS1, Resource group: Create new) and Instance Details (Service name: ai-search-tj, Location: (US) East US). There's also a 'Pricing tier' section with a note about 15 GB/Partition. To the right, there's a 'Select Pricing Tier' table:

Sku	Offering	Indexes	Indexers	Vector quota	Total storage
F	Free	3	3	25 MB	50 MB
B	Basic	15	15	5 GB/Partition	15 GB/Partition
S	Standard	50	50	35 GB/Partition	160 GB/Partition
S2	Standard	200	200	150 GB/Partition	512 GB/Partition
S3	Standard	200	200	300 GB/Partition	1 TB/Partition
S3HD	High-density	1000	0	300 GB/Partition	1 TB/Partition
L1	Storage Optimized	10	10	150 GB/Partition	2 TB/Partition
L2	Storage Optimized	10	10	300 GB/Partition	4 TB/Partition

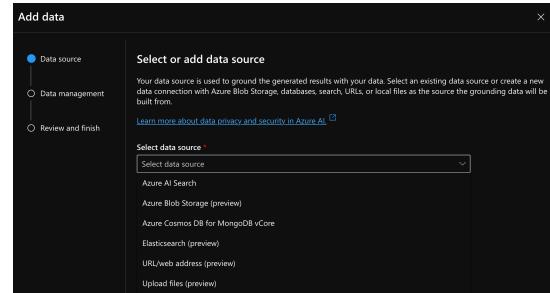
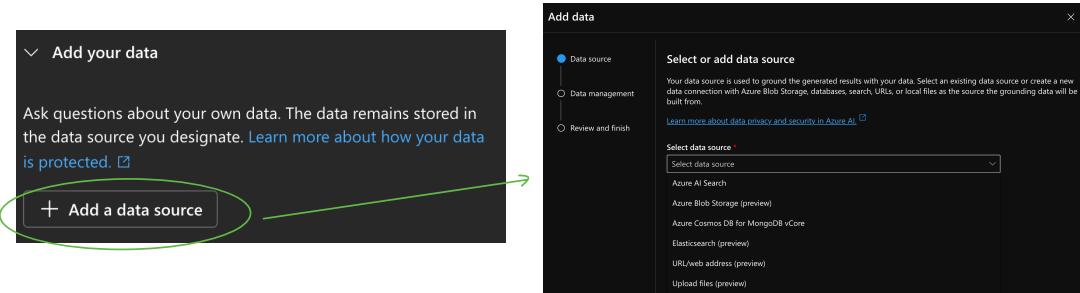
Higher storage limits are available for new services in this region at no additional cost.

So now we have ;

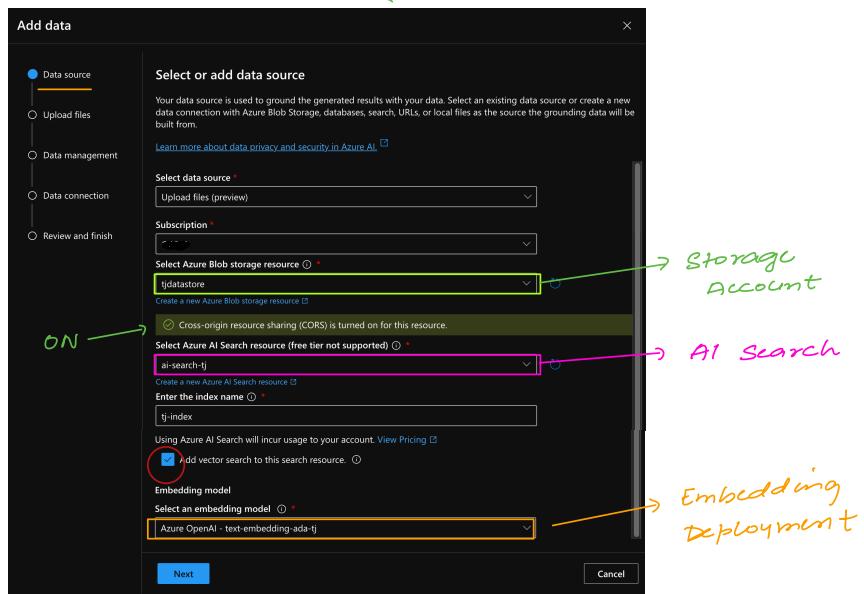
- Storage Account
- Embedding Deployment
- Azure AI Search
- Chat Model Deployment

5.

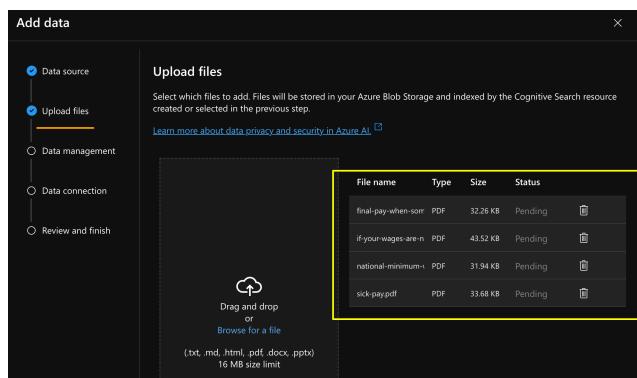
ADD DATA



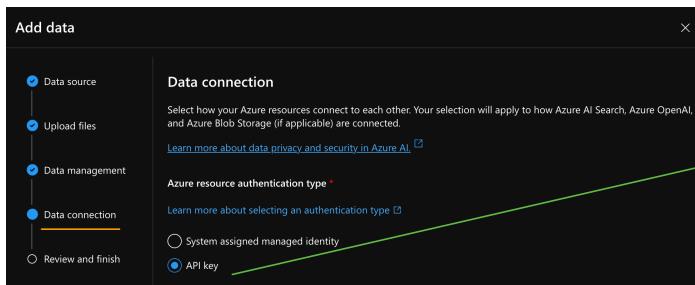
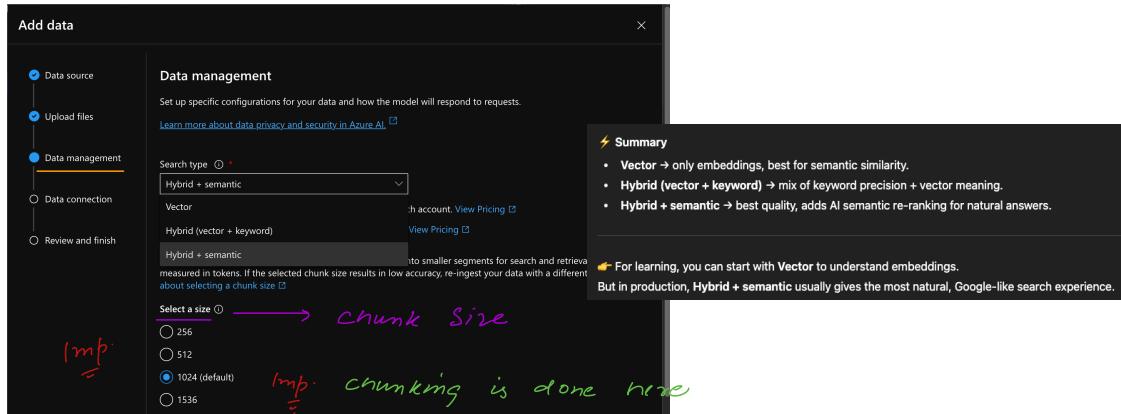
Data Sources



Upload Files

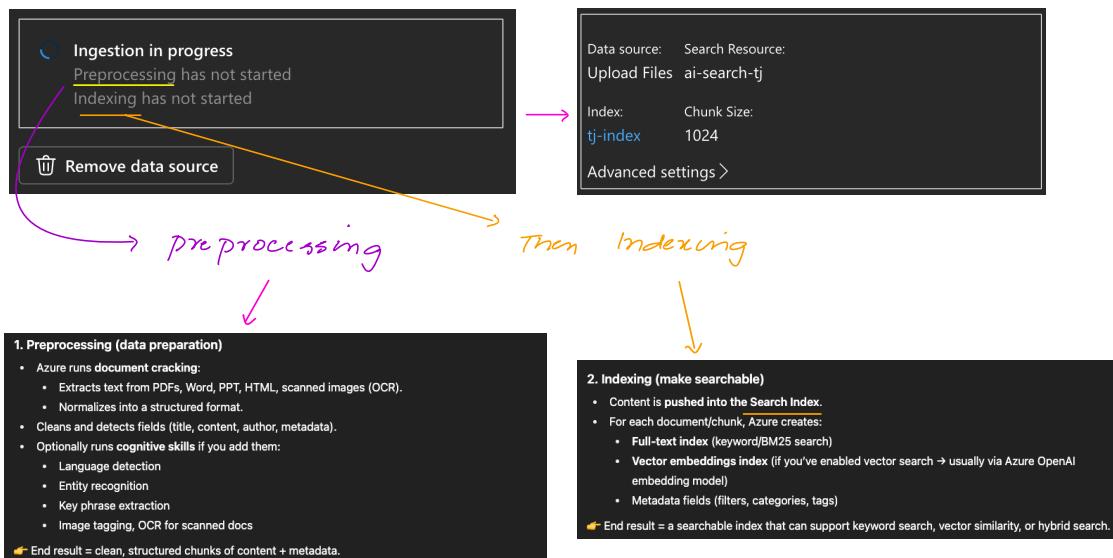


Data Management



Data Ingestion

Once Review & finish ; Data Ingestion starts



Q: what indexing method is used in Azure AI Index?

HNSW → Hierarchical Navigable Small world Graph. (Graph based)

→ It uses ANN (approx. nearest neighbours).

Key Difference in Indexing Context		
Feature	KNN	ANN
Type of search	Exact	Approximate
Speed	Slow for large datasets	Fast
Accuracy	100%	Usually <100% (tunable)
Memory & CPU	High for large datasets	Optimized with index structures
Index structures	Usually none (flat)	HNSW, IVF, PQ, etc.

Imp: Q: what is Hybrid Search?

Hybrid = Both Keyword + vector search
 BM25 HNSW

Hybrid search combines:	How Hybrid Works
<p>1. Vector Search (Semantic / Embedding-based)</p> <ul style="list-style-type: none"> Goal: Find semantically similar items. Index used: ANN (like HNSW, IVF, or PQ). Input: Query embedding vs document embeddings. Output: Candidate vectors that are "close" in semantic space. <p>2. Keyword Search (Exact / Lexical)</p> <ul style="list-style-type: none"> Goal: Match query terms exactly or partially. Method: BM25 (based on bag-of-words). Input: Query terms vs document terms. Output: Ranked documents based on term frequency, IDF, and document length. 	<p>1. Query comes in.</p> <p>2. Two parallel searches:</p> <ul style="list-style-type: none"> ANN search for embeddings → returns semantically similar items. BM25 search for keywords → returns lexically matching items. <p>3. Merge / rerank: Combine the results from both methods using a scoring function.</p> <ul style="list-style-type: none"> Example: weighted sum of BM25 score + ANN similarity score.

Working of BM25 and HNSW

BM25
• Type: Keyword-based search (bag-of-words).
• How it works: Scores documents based on term frequency, inverse document frequency, and document length.
• Example: $TF-IDF \rightarrow$ more frequent words \downarrow unique words \rightarrow Rank high

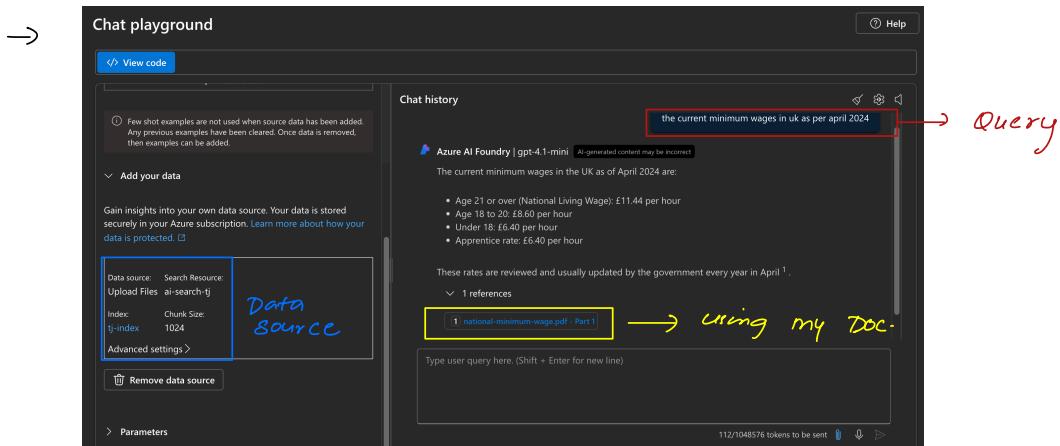
HNSW (ANN)	Graph - Semantic Aware
<ul style="list-style-type: none"> Type: Approximate nearest neighbor search for vectors. How it works: Builds a graph of vectors, traverses efficiently to find nearest neighbors in high-dimensional space. Example: <ul style="list-style-type: none"> Query embedding: vector of "data cleaning" Returns documents semantically similar even if keywords differ, e.g., "Python preprocessing methods" 	

Key Difference
• BM25 → exact keyword match.
• HNSW → semantic similarity using embeddings.

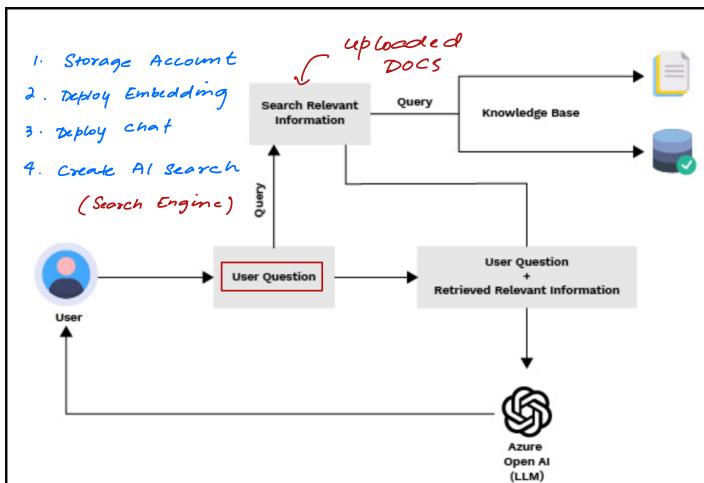
Q1 why indexing?

For fast and efficient retrieval of data.
Filtering etc.

CHAT PLAYGROUND WITH OWN DATA

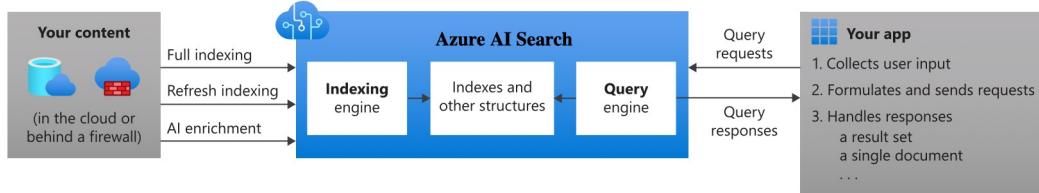


Complete process



Azure AI Search

Azure AI Search ([formerly known as "Azure Cognitive Search"](#)) provides secure information retrieval at scale over userowned content in traditional and generative AI search applications.



Data Chunking

Instead of treating the entire document as a single unit, chunking divides it into smaller segments, typically based on paragraphs, sections, or sentences.

AI Powered Indexing:

Utilizes AI to automatically extract and enrich data from a variety of sources including documents, images, and media files.

Semantic Search:

Uses advanced machine learning models to understand the **intent** behind queries, providing more relevant results.

Cognitive Skills:

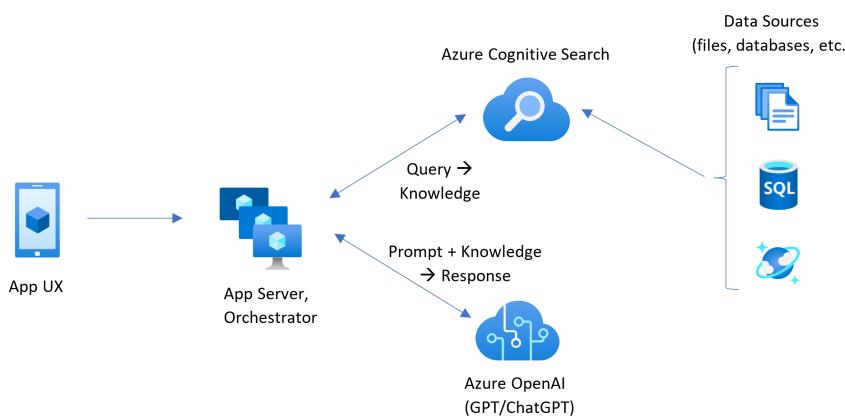
Integrates predefined cognitive skills such as optical character recognition (**OCR**), language **detection**, and entity recognition or allows the creation of custom skills.

Scalable Infrastructure:

Automatically scales to accommodate data volume and query traffic, ensuring efficient performance without manual intervention.

Integration with Azure Ecosystem:

Works seamlessly with other Azure services, such as Azure AI, Azure Functions, and more, for comprehensive data processing and handling.



How vector search works in Azure AI Search

