Integrated Vectorization (IV)
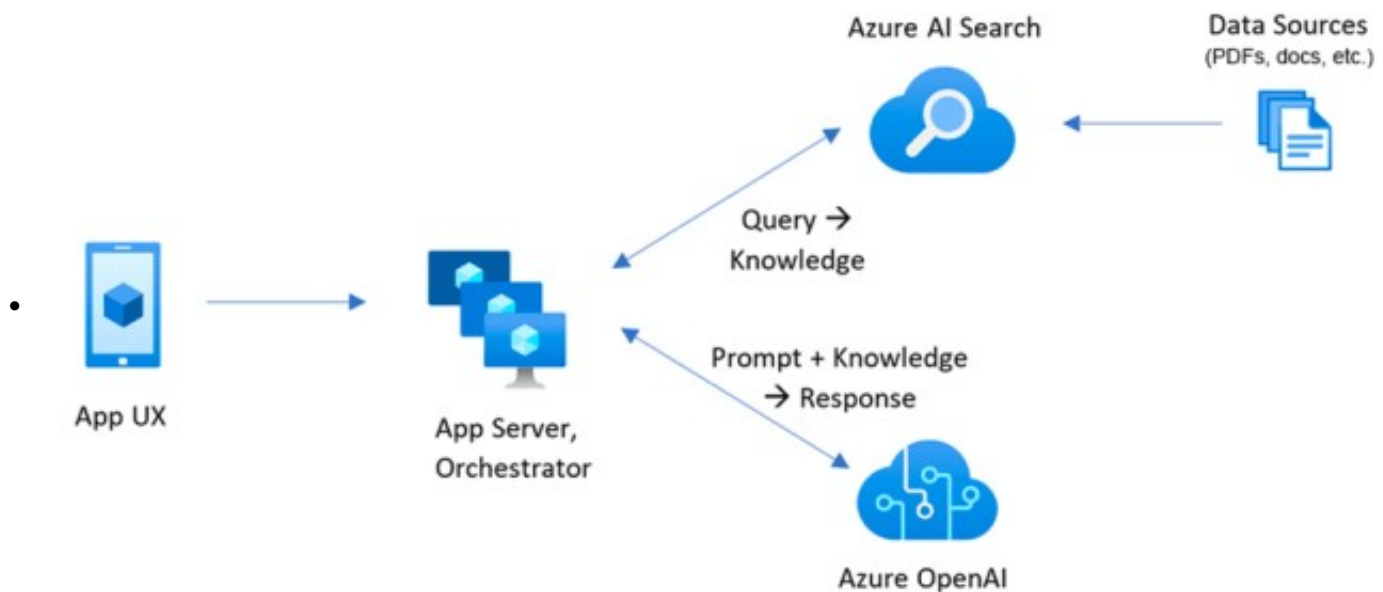
Friday, November 29, 2024   10:24 AM

Info

- https://singhrajeev.com/2024/03/18/getting-started-with-integrated-vectorization/
- https://techcommunity.microsoft.com/blog/azure-ai-services-blog/announcing-the-public-preview-of-integrated-vectorization-in-azure-ai-search/3960809
- Automates most of the chunking and vectorization process
- We cannot use IV on existing index

Example Chat App

- Consider a chat app at ABC company that uses HR documents, employee handbook etc to answer questions.
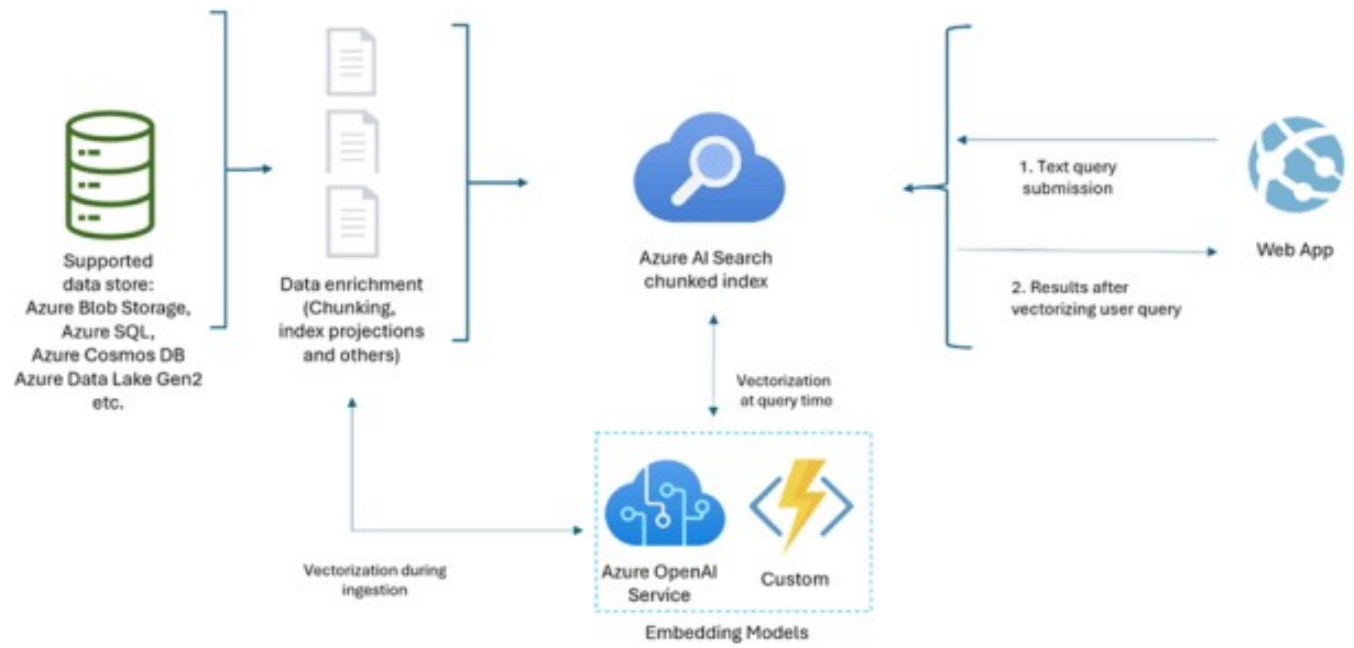   - https://github.com/Azure-Samples/azure-search-openai-demo/tree/main

-

- To push data to AI Search we have to do manual indexing and chunking

Manual Indexing & Chunking

- Create index if doesn't exist
- Upload PDFs to blob storage
- Split PDF to chunk of texts
- Upload chunks to AI Search (compute embeddings if using vectors)

What is Integrated Vectorization

- Allows chunking and vectorization of data during ingestion thru built in pull indexers and vectorization thru vectorizers.
- If we have an deployed embedding model, IV allows us to automate chi=unking and vectorization during data ingestion from various azure sources like Blob, SQL, Cosmos etc

The diagram shows a data flow: Supported data store (Azure Blob Storage, Azure SQL, Azure Cosmos DB, Azure Data Lake Gen2 etc.) → Data enrichment (Chunking, index projections and others) → Azure AI Search chunked index → Web App. Embedding Models (Azure OpenAI Service, Custom) handle Vectorization during ingestion and Vectorization at query time. The Web App submits (1) Text query submission and receives (2) Results after vectorizing user query.

- Users need not maintain separate data chunking and vectorization pipeline, reducing overhead