Microsoft Azure

Chat with your Data-Vision AI Manage AI Apps & Agents with Azure AI Foundry

Priyanshi Singh,
Sr. Technical Specialist, Microsoft

Chris Baudo,
Prin Technical Specialist, Microsoft

Microsoft Azure

Azure AI Foundry

Copilot Studio	Model Catalog				
Visual Studio	Foundational models	Open-source models		Task models	Industry models
GitHub	Azure OpenAI Service	Azure AI Search	Azure AI Agent Service	Azure AI Content Safety	Azure Machine Learning
Azure AI Foundry SDK	Evaluations	Customization		Governance	Monitoring
	Observability				

Azure AI Foundry ecosystem

Copilot Studio

Visual Studio

GitHub

Azure AI Foundry SDK

Design with the best models

Azure OpenAI Service	Phi models	Microsoft Research	Snowflake models	Hugging Face models	Mistral AI models	Meta Llama models	Rockwell	Bayer
Cohere models	Databricks models	Nvidia models	Jais G42 models	Nixtla TimeGen1	NTT data model	Sight machine	Saifr	DeepSeek
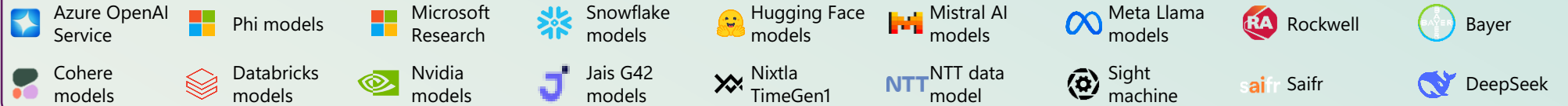
Customize with a comprehensive agent toolchain

1P offerings

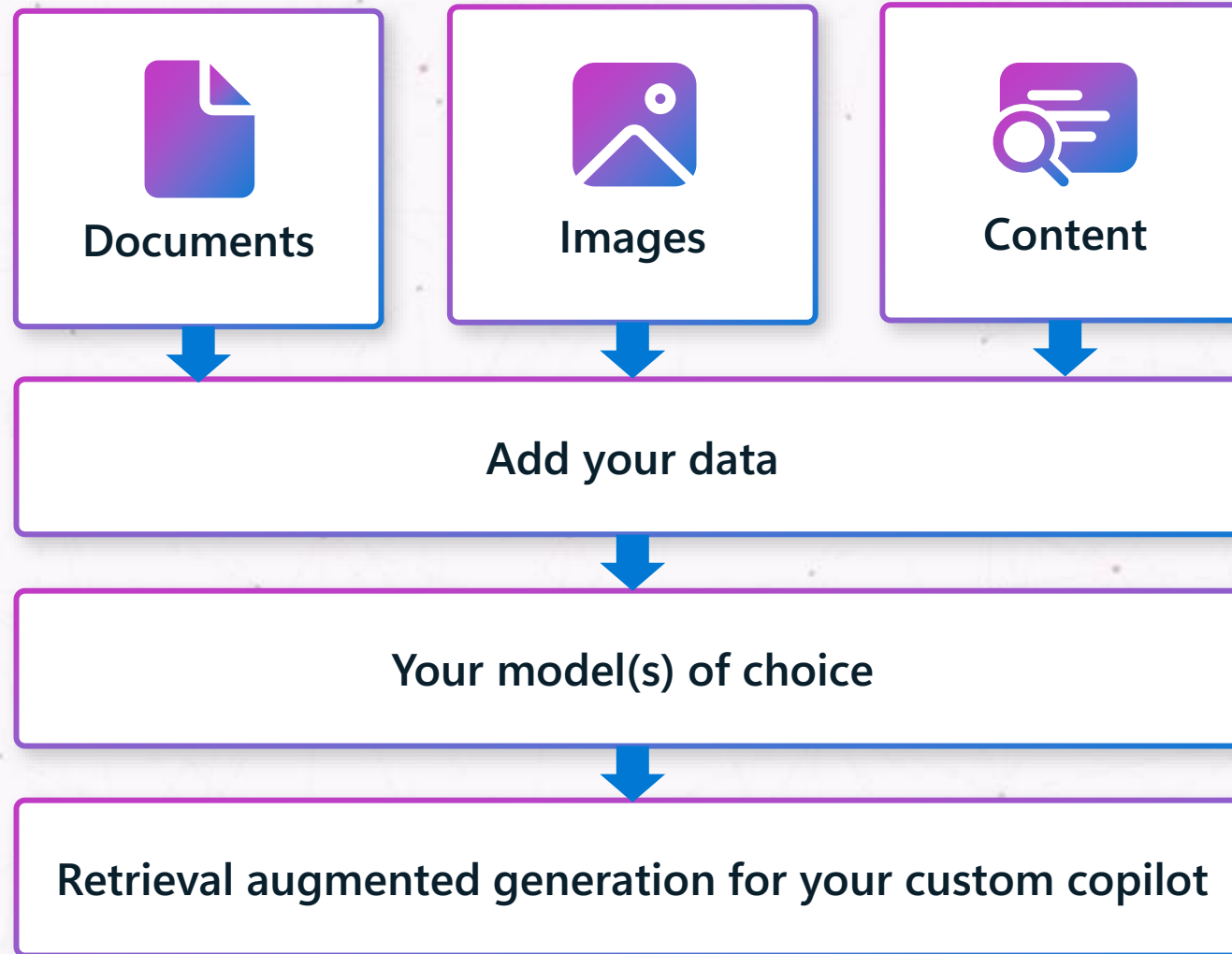|					
Azure AI Agent Service	Azure AI Search	Microsoft Fabric	Azure SQL Database	Microsoft Bing	Microsoft Blob Storage	Microsoft SharePoint
Azure Redis	Cosmos DB	Azure Data Lake Storage	Azure Logic Apps	Azure Functions	Event Hub	Dynamic Sessions
Container Apps	Azure Kubernetes	Azure AI Speech	Azure AI Language	Azure AI Translation	Azure AI Vision	Azure AI Doc. Intelligence
Azure AI Content Und.	Semantic Kernel	AutoGen	Azure ML	Prompty	Fine-tuning & distillation	Tracing & debugging
Evaluation	Experiments	AI Toolkit for VS Code				

3P offerings

Langchain	LlamaIndex
Llama Stack	Code Interpreter
OpenAI Actions	Gretel
Weights & Biases	Scale
Statsig	Arize AI

Manage performance in production

Azure Monitor	Azure App Config	Microsoft Cost Management	ClearML	Dataloop	GitHub Actions
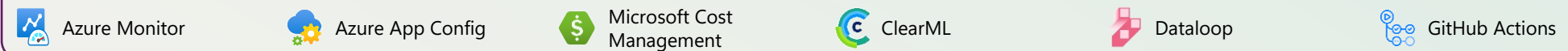
Safeguard with Trustworthy AI

Azure AI Content Safety	Microsoft Defender	Entra ID	Azure Policy	Azure Application Gateway	Azure Confidential Computing	Azure Backup
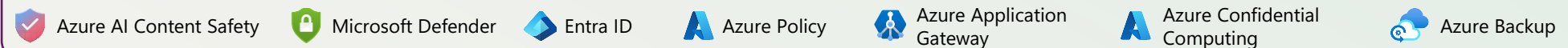
Retrieval Augmented Generation

Retrieval Augmented Generation orchestration

User question

Orchestration workflow
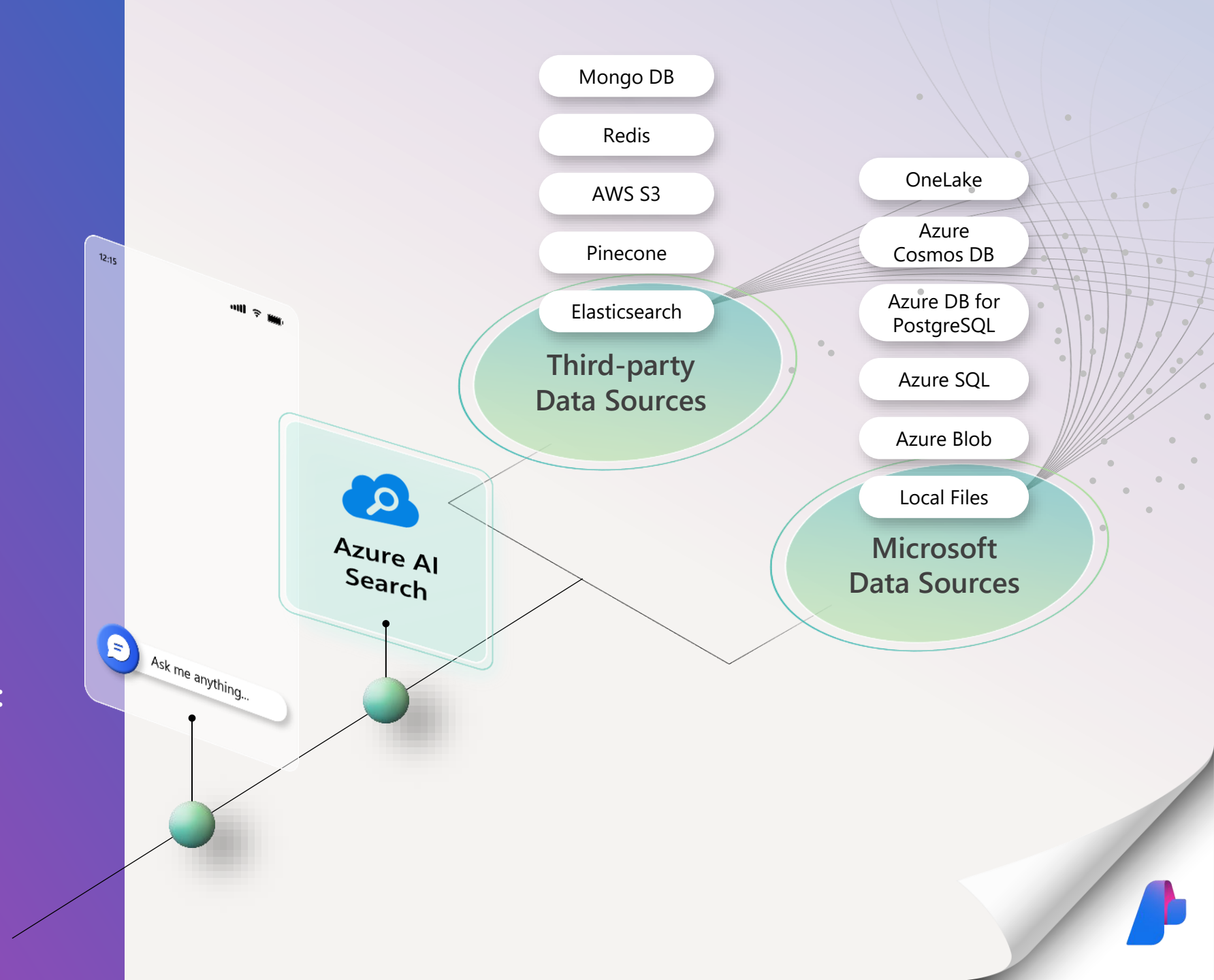
Query my data → Search Index

Add results to prompt

Query model → Model(s) of choice

Send results
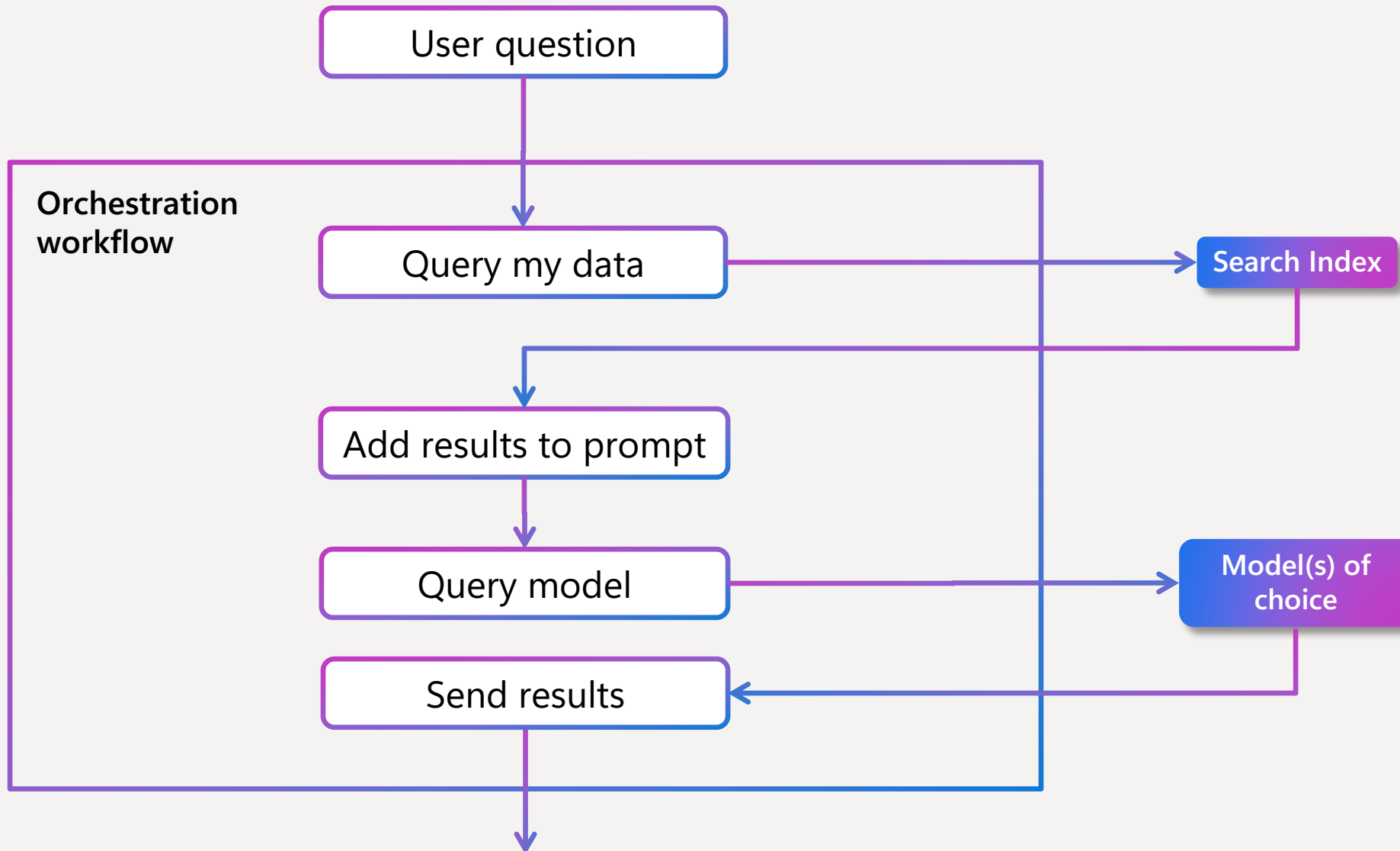
Wherever your data resides, you can bring it to life through Azure AI Search

The Azure AI Foundry difference:

- State-of-the-art ranking
- Automatic indexing
- All data welcome

Azure AI Search

Third-party Data Sources
- Mongo DB
- Redis
- AWS S3
- Pinecone
- Elasticsearch

Microsoft Data Sources
- OneLake
- Azure Cosmos DB
- Azure DB for PostgreSQL
- Azure SQL
- Azure Blob
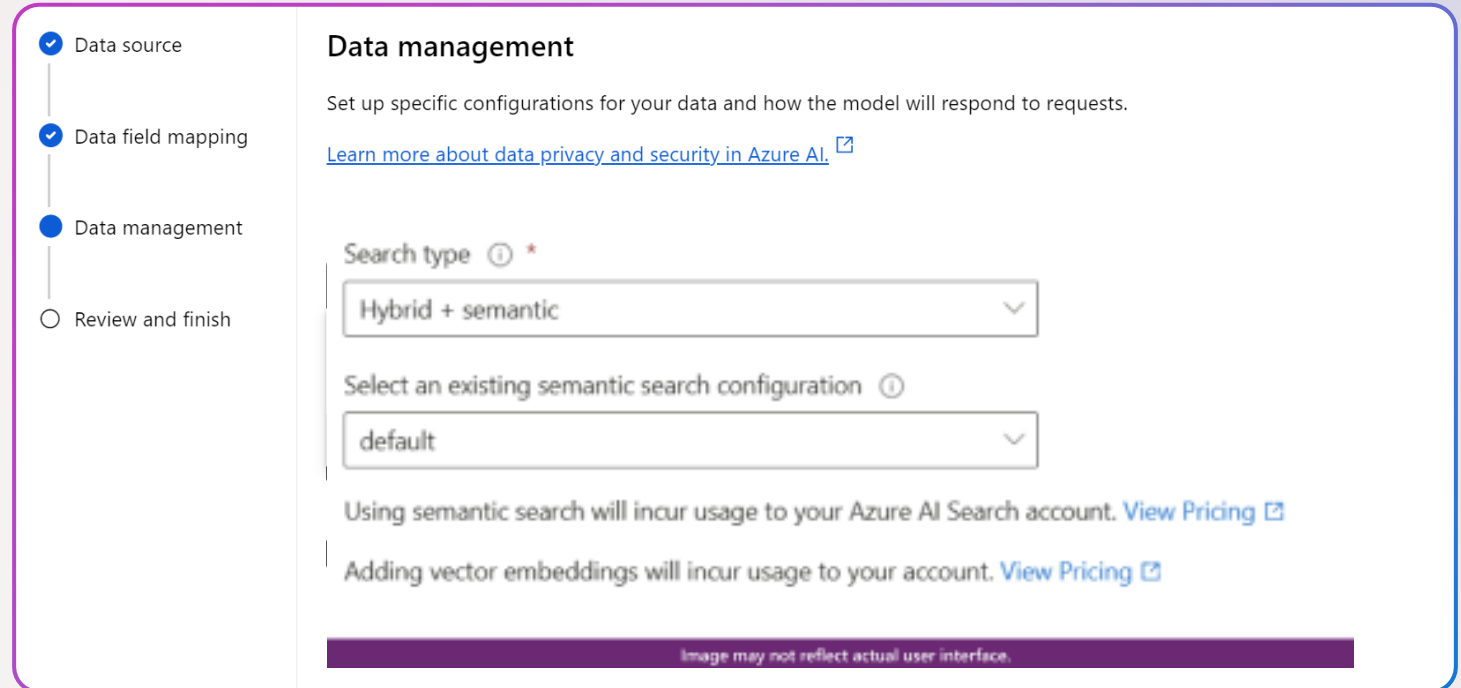- Local Files

Ask me anything...

Azure AI Search

Azure AI Search is a battle-tested retrieval system purpose-built to deliver superior retrieval augmented generation (RAG). Use AI to collect, enrich and surface all aspects of your data.

- Ingest and search across all your data, no matter the type, volume or source

- Streamlined RAG pipeline, from import to chunking to embedding.

- Enterprise-ready foundation cost-optimized for performance and scale.

- Enable hybrid search (vector + keyword) results using RRF (Reciprocal Rank Fusion) and multi-vector retrieval for multi-modal search

- Improve retrieved results immediately with reranking and query rewriting.

- Tailor RAG experiences by utilizing customizable features such as granular access control and relevance tuning

Customize your models

Data source

Data field mapping

Data management

Review and finish

Data management

Set up specific configurations for your data and how the model will respond to requests.

Learn more about data privacy and security in Azure AI.

Search type ⓘ *

Hybrid + semantic

Select an existing semantic search configuration ⓘ

default

Using semantic search will incur usage to your Azure AI Search account. View Pricing

Adding vector embeddings will incur usage to your account. View Pricing

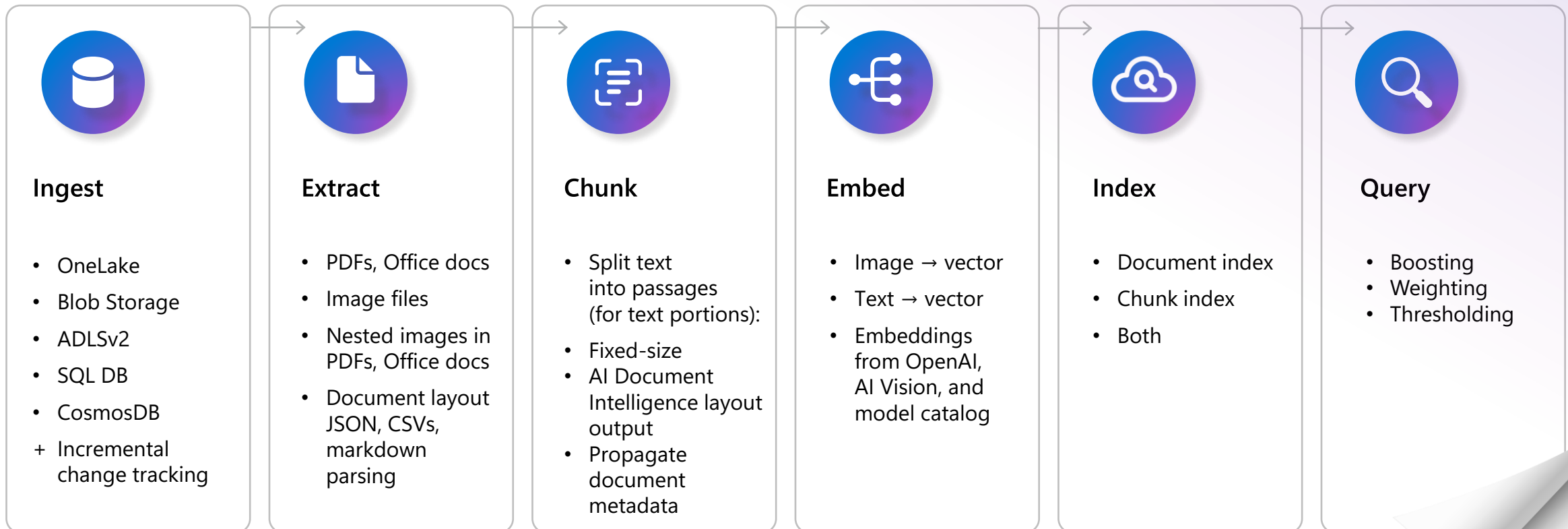Image may not reflect actual user interface.

💡 **Easily enable search best practices through the UI: like being able to select hybrid search + reranking in one click.**

Streamlined RAG data pipeline
End to end, integrated data prep and processing

Ingest

- OneLake
- Blob Storage
- ADLSv2
- SQL DB
- CosmosDB
- + Incremental change tracking

Extract

- PDFs, Office docs
- Image files
- Nested images in PDFs, Office docs
- Document layout JSON, CSVs, markdown parsing

Chunk

- Split text into passages (for text portions):
- Fixed-size
- AI Document Intelligence layout output
- Propagate document metadata

Embed

- Image → vector
- Text → vector
- Embeddings from OpenAI, AI Vision, and model catalog

Index

- Document index
- Chunk index
- Both

Query

- Boosting
- Weighting
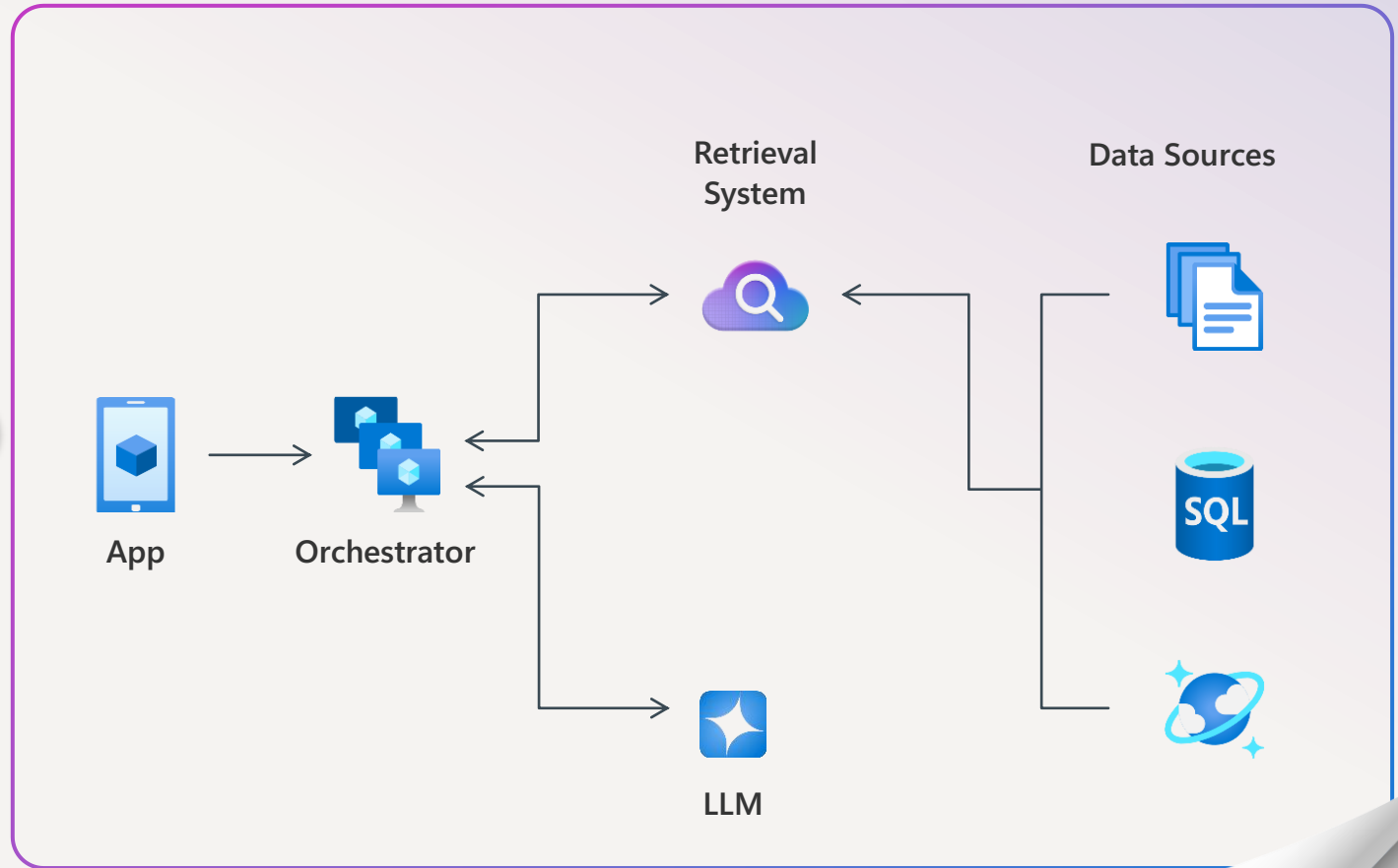- Thresholding

Retrieval Augmented Generation
Combine reasoning with knowledge
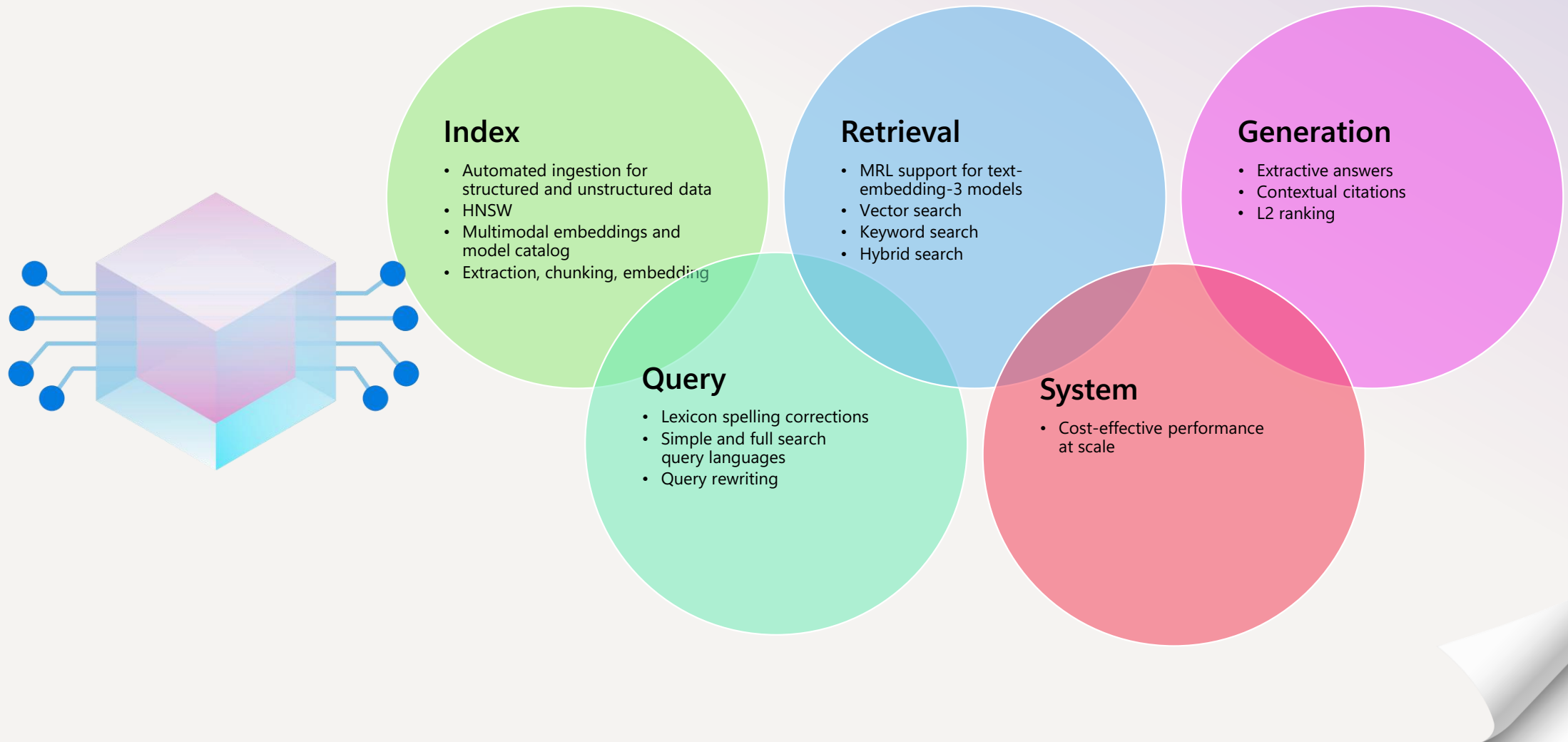
Elements

- Retrieval system collects externalized knowledge
- Language model provides reasoning to provided knowledge and prompt

Techniques retrieval system brings

- Querying: spelling corrections, multi-lingual query support
- Retrieval: keyword, vector, hybrid search
- Augmentation: L2 ranking,
- Generation: Extractive answers, citations
- Indexing: chunking, embedding, automated data import

App → Orchestrator → Retrieval System / LLM → Data Sources

Azure AI Search features across the RAG pipeline

Index
- Automated ingestion for structured and unstructured data
- HNSW
- Multimodal embeddings and model catalog
- Extraction, chunking, embedding

Retrieval
- MRL support for text-embedding-3 models
- Vector search
- Keyword search
- Hybrid search

Generation
- Extractive answers
- Contextual citations
- L2 ranking

Query
- Lexicon spelling corrections
- Simple and full search query languages
- Query rewriting

System
- Cost-effective performance at scale

Integrate structured & unstructured data

- **Azure AI Search**
- **Blob storage**
- **Local files/folders**
- **Storage URLs including OneLake in Microsoft Fabric**
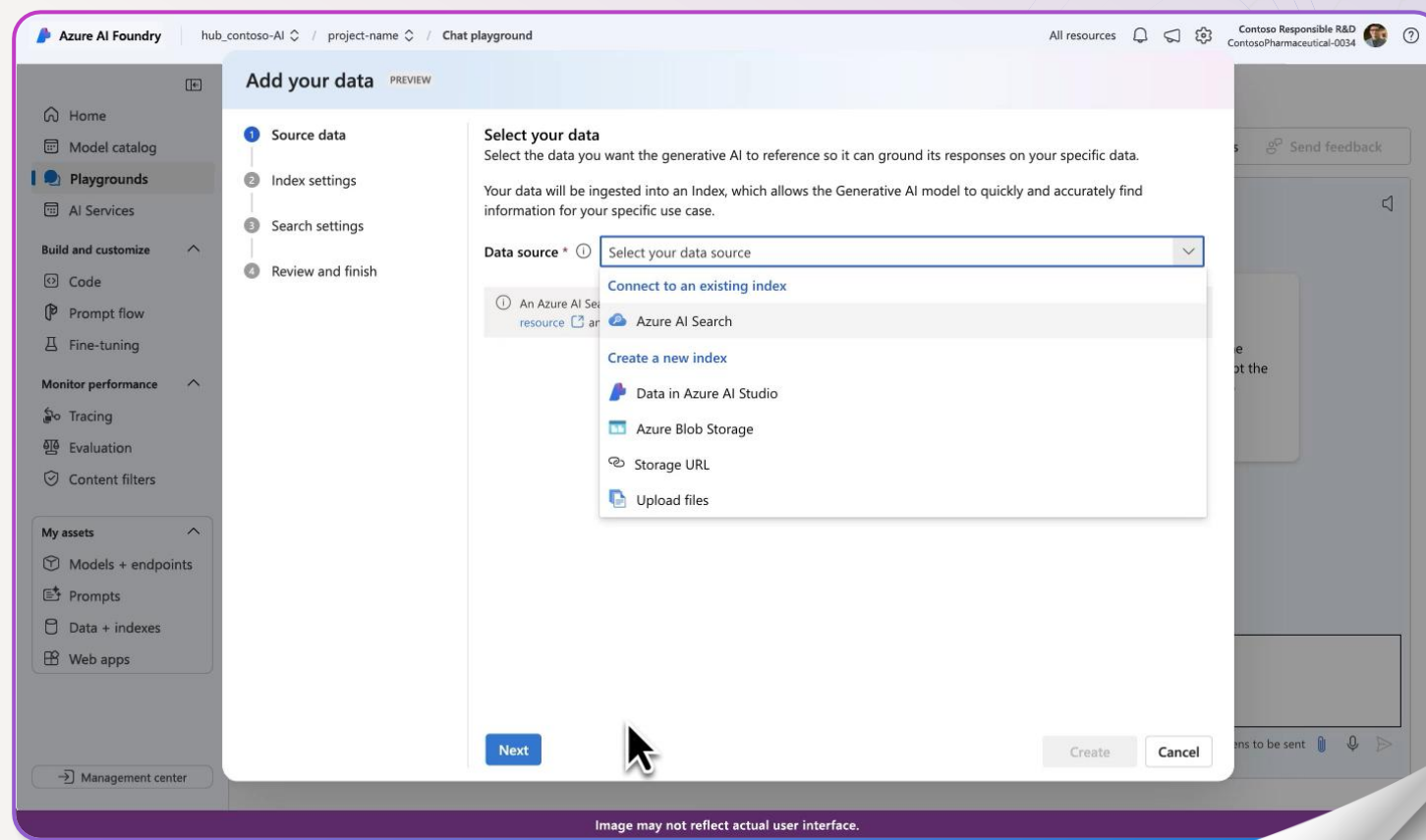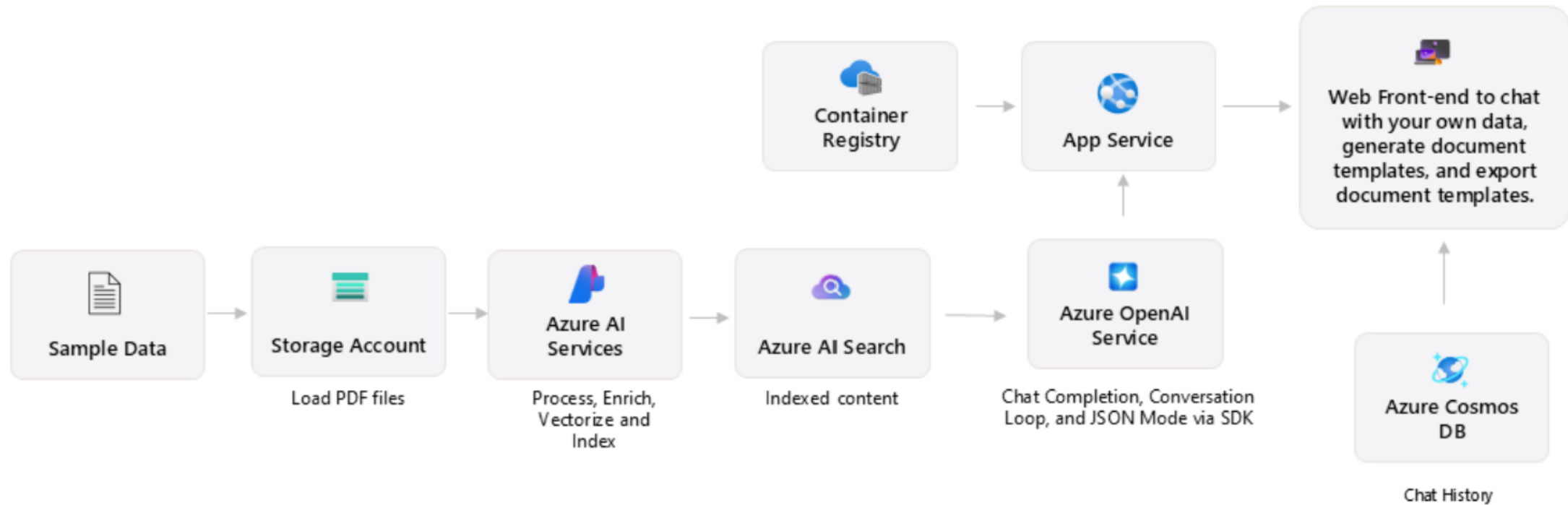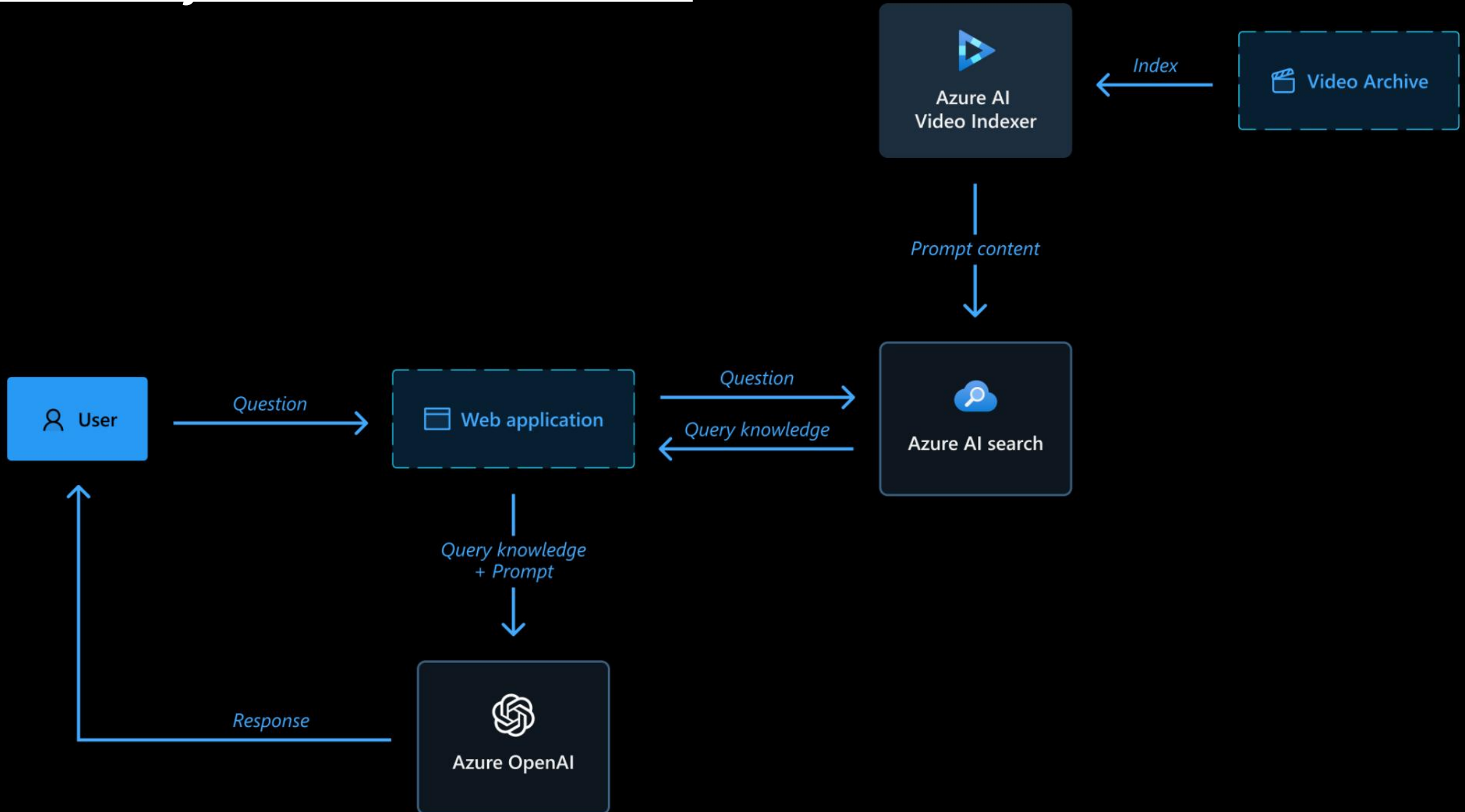 - Azure Databricks
  - S3 buckets via Amazon S3 Shortcuts

Image may not reflect actual user interface.

Solution architecture

Chat With your Data Solution

| Sample Data | → | Storage Account | → | Azure AI Services | → | Azure AI Search | → | Azure OpenAI Service | | Container Registry | → | App Service | → | Web Front-end to chat with your own data, generate document templates, and export document templates. |

Load PDF files — Storage Account

Process, Enrich, Vectorize and Index — Azure AI Services

Indexed content — Azure AI Search

Chat Completion, Conversation Loop, and JSON Mode via SDK — Azure OpenAI Service

Azure Cosmos DB

Chat History

Chat With your Data Solution- Videos

Track 2 GitHub Resources:

- 1. CA AI Hackathon