Enhancing apps with Azure AI capabilities - Semantic Search and Voice Search

Azure Identify Tech Stack for Backend

1. Core Data + Indexing

* Azure Cosmos DB (SQL API): Serves as the primary data store. Azure AI Search uses this as the base to generate the index. Cost: Free tier available.
* Azure AI Search: Acts as the search engine and indexing service for the database. Used by the agent in combination with the LLM model. Free tier includes up to 3 indexes, allowing indexing for up to 3 databases.
* Azure Blob Storage (Optional): Used for staging files that the database can read. This is optional and not required currently. Cost: Pay-as-you-go for storage and read/write operations.

1. Vectorization (to enable semantic similarity)

Integrated Vectorization in Azure AI Search: Automatically chunks and embeds data during indexing, enabling vector or hybrid queries. To use this, enable vectorization while creating the index and specify the embedding endpoint. Cost: Pay-as-you-go based on the embedding model used.

1. Agent & Query Translation layer

The agent acts as a worker that follows predefined goals and steps to execute tasks and perform searches. Cost: Agent container is free; however, models/tools used (e.g., token-based LLMs) incur charges.

* Azure AI Foundry Agent Service: Search Agent
* Azure OpenAI service: Provide LLM
* Agentic Retrieval: First 1000 semantic queries is free monthly if semantic rerank is enable

1. API & Runtime

Azure Functions: Serverless backend for search APIs, index triggers, and health checks. Accepts user queries, invokes the agent, performs the search, and returns results. Cost: Free on the consumption plan.

Azure API Management: Acts as a secure gateway for your functions, offering key management, rate limiting, and transformation. Used to expose stable endpoints (e.g., /api/search) to apps and teammates during testing. Cost: First 1 million calls/month are free on the consumption plan.

1. Security & Networking

Handles identity, secrets, and private access for application code and API calls. Cost: Free.

1. Ops & Observability

Azure Monitor + Application Insights: Provides traces, logs, metrics, and dashboards for functions, agent calls, and search latency. Cost: First 5 GB/month of logs are free.

Connecting the Pieces

1. Data is stored in Cosmos DB.
2. Azure AI Search Indexer reads the data; integrated vectorization creates embeddings and searchable fields.
3. User sends a query to the APIM endpoint, which triggers an Azure Function.
4. The function invokes the Agent Service, which uses Azure OpenAI to translate the query from natural language to a structured hybrid query.
5. The function executes the query on Azure AI Search and returns the results.
6. Application Insights and Azure Monitor capture metrics such as latency, queries per second (QPS), and token usage.

