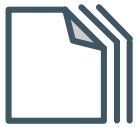
CosmosAIGraph Hybrid RAG Approach

User Inputs in an example AI Conversation:

1. What is the Python Flask Library	DB RAG
2. What are its dependences	Graph RAG
3. What are the alternatives that use async processing	Vector RAG
4. Who is the author	DB RAG
5. What other libraries did she write	Graph Rag
6. Display a graph of all her libraries and their dependencies	Graph RAG

User Input



RAG Data

In-Memory RDF Graph Database, loaded from vCore

Cosmos DB vCore w/Vector data and search

Application Logic:
- Determine Intent & RAG Strategy from User Intent
- Identifiy Entities
- Generate SPARQL query if Graph RAG
- Generate vCore query if DB RAG
- Vectorize user input if Vector RAG
- Execute the DB query to get Docs List
- Fetch Documents per List from Cosmos DB
- Craft the Prompt with the Document RAG Data
- Invoke the LLM with input & RAG data in prompt
- Parse the LLM response, and present in the web UI

Prompt

LLM

Output
JSON, TXT, etc

Repo: aka.ms/caig