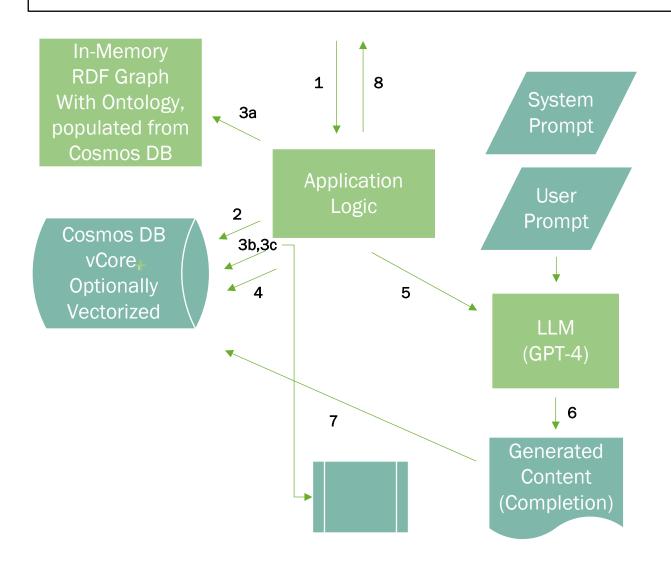
## Retrieval Augmented Generation (RAG) Pattern with either Graph or Vector Search

## **User Input or Prompts:**

- 1) What is the bill of materials for the Python Flask Library?
- 2) Which of these libraries support Asynchronous processing?

(graph search) (vector search)



## **Logical Processing Steps:**

- 1) User Prompt
- 2) Check Prompt History
- 3) Identify System Prompt Data
- 3a) SPARQL Graph Query, and/or
- 3b) Vector Search Query, and/or
- 3c) Cosmos DB Query
- 4) Read System Prompt Data
- 5) Invoke LLM with Prompts
- 6) A Completion is Returned
- 7) Persist Prompt and Completion
- 8) Display Response in UI

Note: The SPARQL query may also be *generated* with per the User prompt and the OWL (Ontology).

