

**Data ingestion:**

**Pdf or any file format 🡪 text 🡪** chunk (word,sentence,semantic,proposition, llm based and summary based on user query)🡪 save chunks with metadata(mogoDB, SQLite etc.)🡪 embedding(Siamese llm)

**Community wise Chunks clustering**: KNN, SOM, cosine similarity on chunks embedding, summarized chunks then cluster.

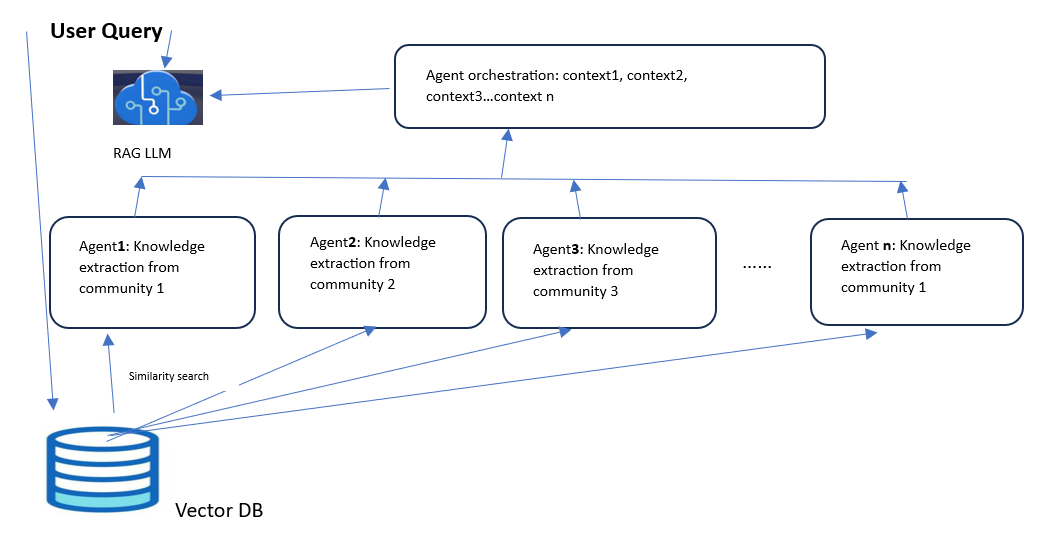
**Vector DB to store Embeddings:** multivector approach for hybrid search, milvus, pinecone, Neo4j to improve accuracy and search space. save embeddings for similar community/clusters for docs/chuncks

**RAG (Retrieval Augmented Generation)**

**Retrieval:** similarity search with cosine similarity from all communities and generate the partial context. If no coherence context found, then agent-based retrieval approach will be used. Below the explain how agent improve to augment the response by LLM by providing coherence context from all clusters/communities.

**GenAI Multi Agent based RAG to retrieve complex information:**

Agent is responsible to create coherence context and optimize the overall performance of the RAG system.

****

**Augmentation:** Giving the user query and the coherence context generated by agents’ orchestration helps LLM to augment user query’s responses.

**Generation:** LLM generates coherence response based on augmented response which will be me structured and accurate. That can’t be possible without agent support.