RAG Pipeline Flow

We can decompose the RAG pipeline into 6 services

1.) Document Ingestion Service - Responsible for ingesting documents of different types (pdf, txt, doc etc). This service should process a document into chunks, assign a unique file_id to the document and store other metadata(like document type, time, document_name etc)

2.) Model Service - Responsible for turning the chunks into embeddings

3.) Vector DB Service - Responsible for storing the embeddings in the VectorDB. The VectorDB will have different 'Collections' in them, each Collection corresponds to a Document Set.

4.)Query Expansion Service - Responsible for applying suitable query expansion techniques to the user query before it is sent to the Search service.

5.)Search Service - Responsible for taking in the transformed query and running Hybrid Search to retrieve the relevant context

6.)Answer Generation - Responsible for using the retrieved context and a LLM call to generate the file answer. The answer service should also return the Document names and the pages of the Documents where the relevant context was retrieved from