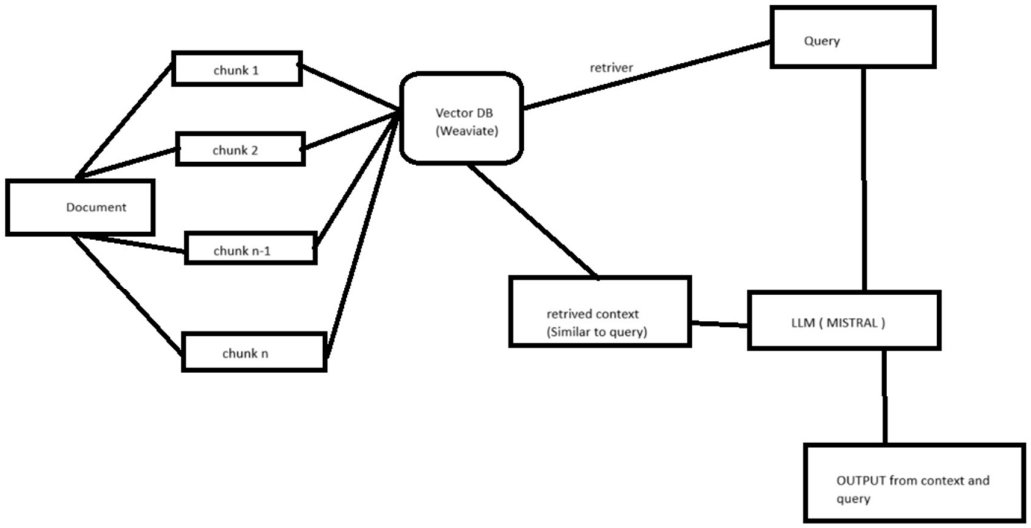
RAG PIPELINE DOCUMENTATION (PART – 1)

MODEL ARCHITECTURE :

1) DATA LOADER - > Using Langchain PDFLOADER
2) EMBEDDING AND STORING IN DB : using weaviate db instance for cloud storage and efficient retrieval
3) RETRIVAL OF RELEVENT CHUNKS : passing query to database to retrieve relevant chunks .
4) Creating Prompt.
5) LLM : Mistral AI for Output generation using prompt which provide the context along with query

ARCHITECTURE :



Retrieval Process : Graph based Approximate Nearest Neighbour Search

Response generation :

The retrieved context along with the query in form of prompt is sent to LLM which generates the response according to the document relevance.

Sample queries and output

1)

```
output = rag_chain.invoke("what is the overall assignment work")
index = output.find('Answer_rag')
print(output[index:])
```

✓ 3.2s Python

```
Answer_rag :

The overall assignment work involves developing a Retrieval-Augmented Generation (RAG) model for a Question Answering (QA) bot problem statement. The m

The task requirements include implementing an RAG-based model, using a vector database like Pinecone DB to store and retrieve document embeddings effic
```

2)

```
output = rag_chain.invoke("Do you know how RAG works ?")
index = output.find('Answer_rag')
print(output[index:])
```

✓ 3.0s Python

```
Answer_rag :

RAG stands for Retrieval-Augmented Generation. It is a model for question answering (QA) bot that uses a vector database like Pinecone DB and a generat

The RAG-based model can handle questions related to the provided document or dataset
```

3)

```
output = rag_chain.invoke("How many parts are there is assignment?")
index = output.find('Answer_rag')
print(output[index:])
```

✓ 3.0s Python

```
Answer_rag :

The assignment has two parts. Part 1 involves developing a Retrieval-Augmented Generation (RAG) model for Question Answering (QA) bot problem statement
```

4)

```
output = rag_chain.invoke(" How much stipend will I be paid  ?")
index = output.find('Answer_rag')
print(output[index:])
```

✓ 1.5s

```
Answer_rag :

I don't know. The context does not provide information about the stipend amount.
```