BGSW
GEN-AI Hackathon
SPRINT-2

Team details

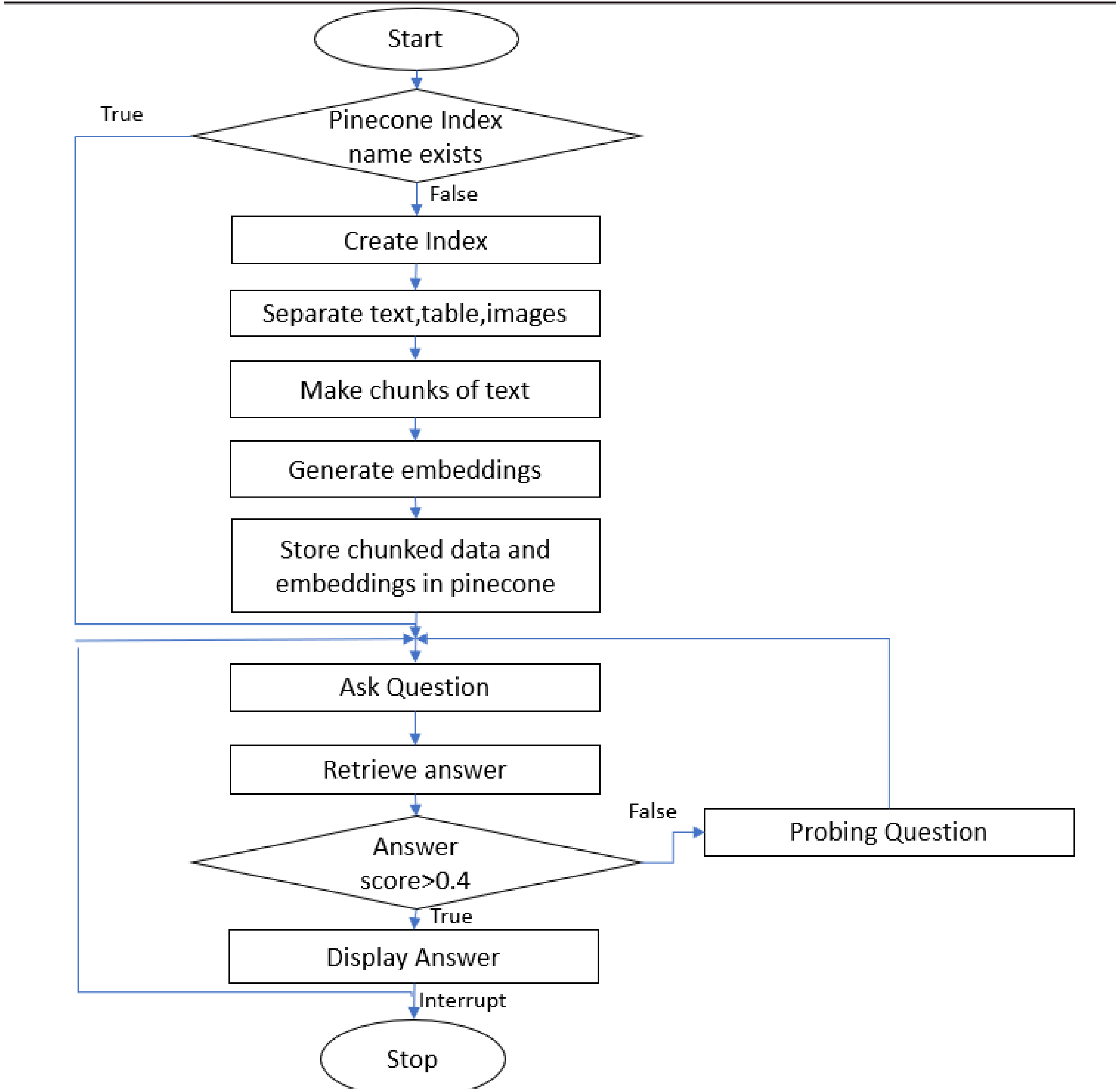
Team Name-Code Crafters

	Name	Institute
1	Abhijeet Jadhav	KLE Technological University
2	Dheeraj Hegde	
3	Poonam Shettar	

Overview

- Storage and retrieval for multiple pdfs is carried out.
- Pinecone is used for storing the pdf data, known for its low latency performance.
- The PDF content has been divided into text, tables, and images.
- Each part has generated embeddings and is stored in Pinecone.
- Text in the PDF data is chunked to a specific length to expedite processing.
- Question answering is carried out through streamlit interface.

Process Flow



Database -Pinecone

- Pinecone is a serverless vector database that helps to build AI applications faster and cheaper.
- Low latency vector search.
- Supports vector search, metadata filters, keyword boosting and integration with various cloud platforms.

Embeddings for text –Sentence Transformers

- State-of-art python framework for sentence, text and image embeddings.
- Model used – "all-MiniLM-L6-v2“
- It maps sentences and paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

Embeddings for image - clip

- Model uses a VIT-B/32 Transformer architecture as an image encoder.
- It generates 512- dimension embeddings .
- Input image is resized to 224x224.

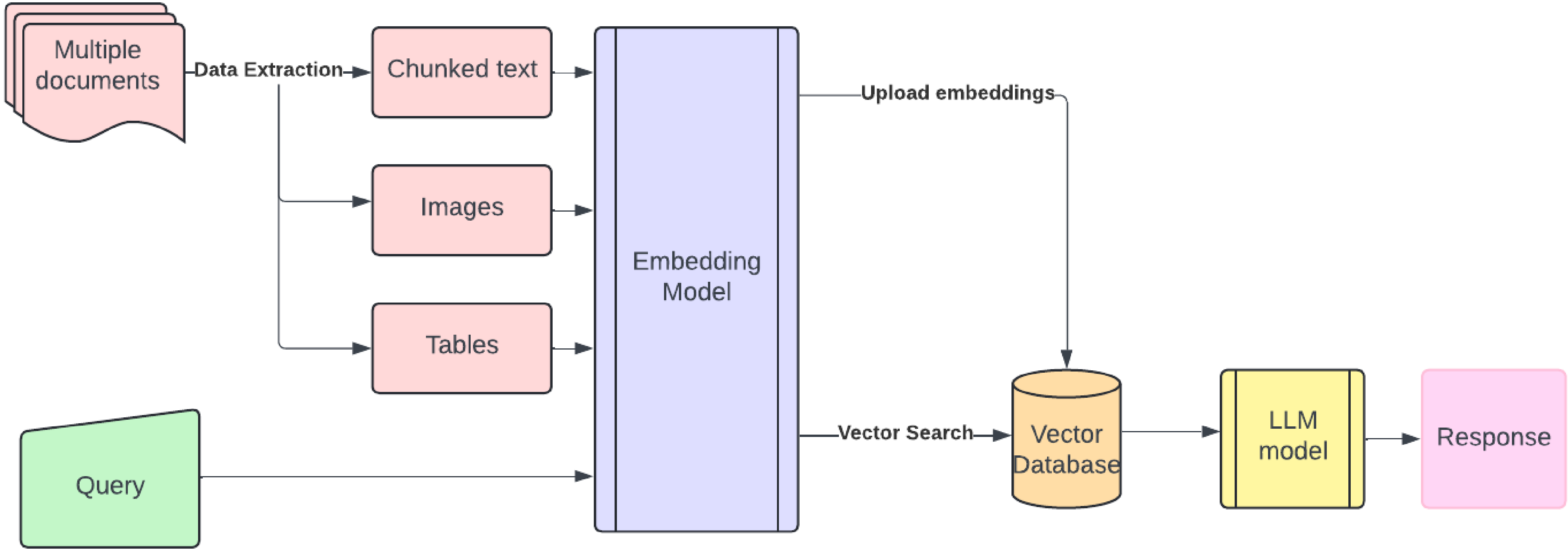
Embeddings for table –Sentence Transformers

- State-of-art python framework for table embeddings.
- Model used – "deepset/all-mpnet-base-v2-table”.
- It maps sentences & paragraphs to a 768 dimensional dense vector space and can be used for tasks like clustering or semantic search.

User Interface- Streamlit

- An open source framework for turning python scripts into web apps.
- To create interactive data apps

Block Diagram



Merits

- Able to generate and store embeddings for multiple documents.
- Able to successfully segment and generate embeddings for text, tables and images utilizing the state-of-art embedding models for each of them separately.
- Able to retrieve and answer queries from multiple documents.
- Able to generate probing questions.
- Chunking reduces time taken for storing in the database.
- Inclusion of embeddings and metadata makes the retrieval faster.
- Inclusion of only open-source models.

Considerations

- Probing questions limit is set to 2 which can be adjusted based on needs.

Thank You