BGSW
GEN-AI Hackathon
SPRINT-1

Team details

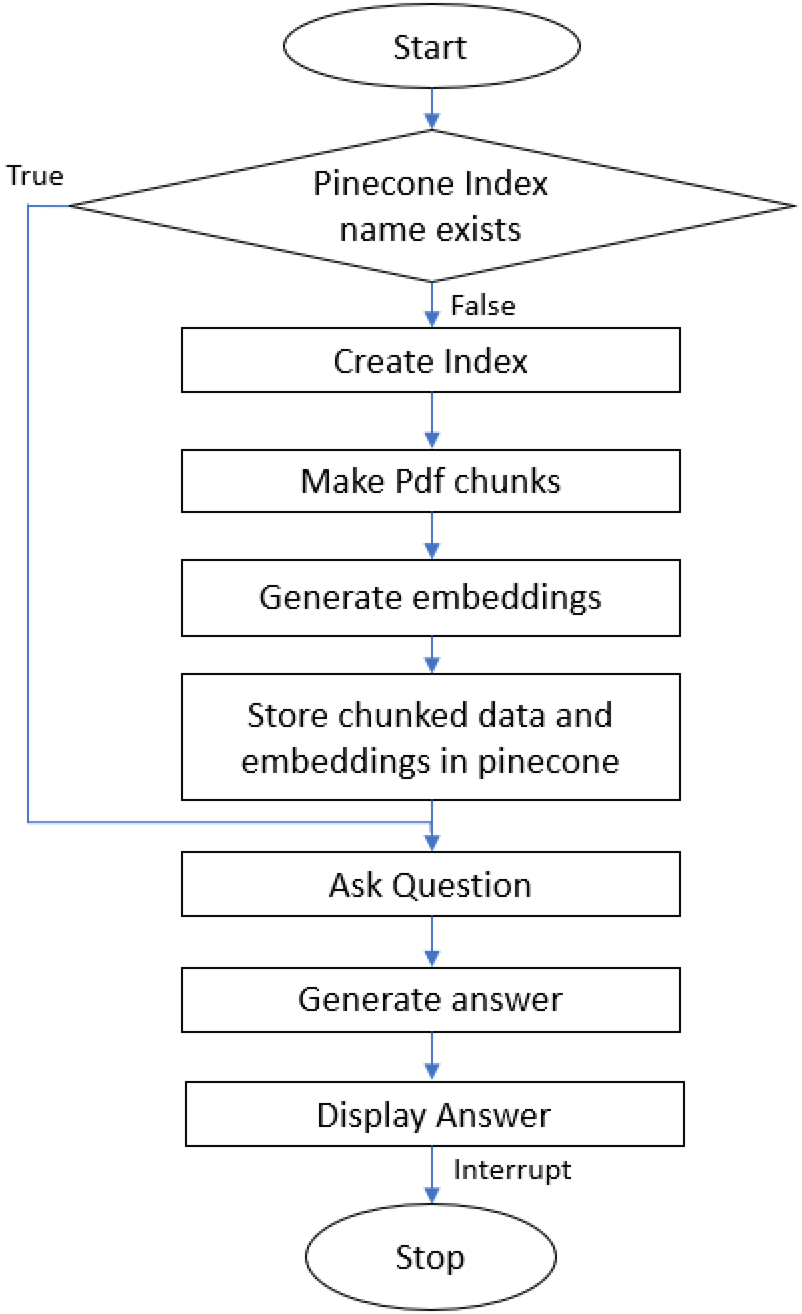
Team Name-Code Crafters

	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.
- PDF data is chunked to expedite processing.
- The chunked data is stored in the vector database(pinecone) along with its generated embeddings.
- 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 –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

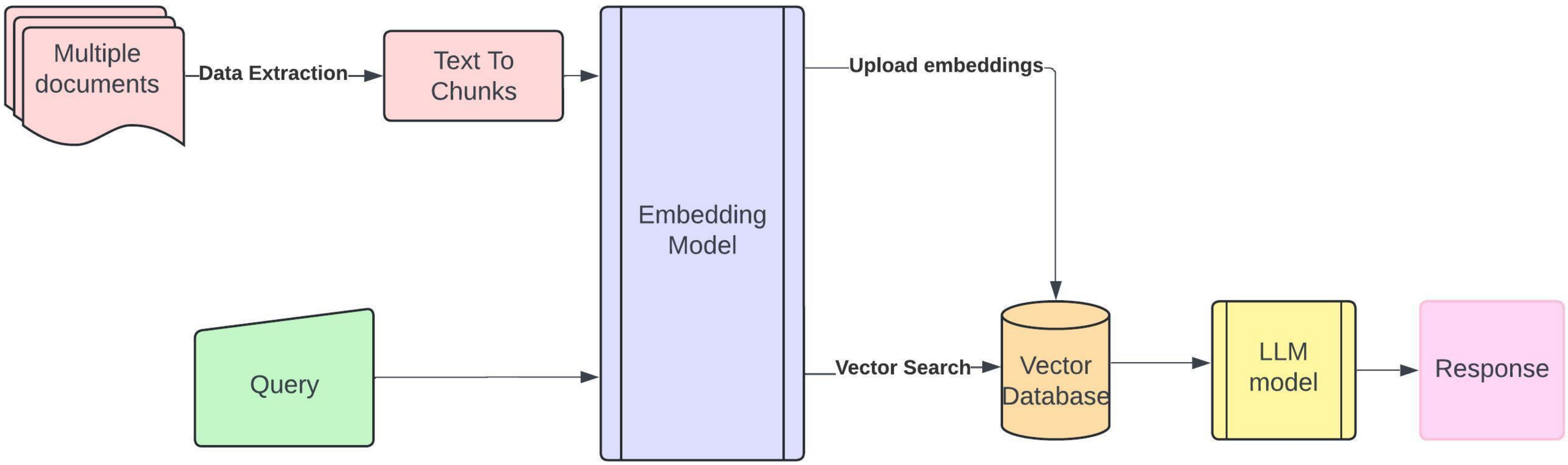
Question Answering-Bart Model

- Bart uses a standard seq2seq/machine translation architecture with a bidirectional encoder (like BERT) and a left-to-right decoder.
- Bart model from hugging-face library.

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 retrieve and answer queries from multiple documents.
- Chunking reduces time taken for storing in the database.
- Inclusion of embeddings makes the retrieval faster.

Current Limitation

- Probing questions were removed to enhance the accuracy of the generated answers. They will be included in the next sprint.

Thank You