Design a system like ChatGPT for your company :
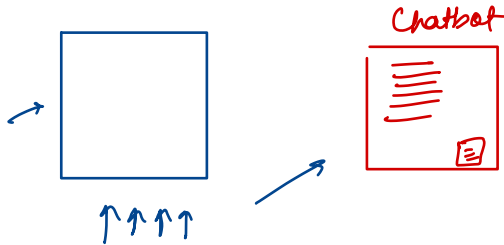
Software → chatbot →

Chatbot

Technical Support
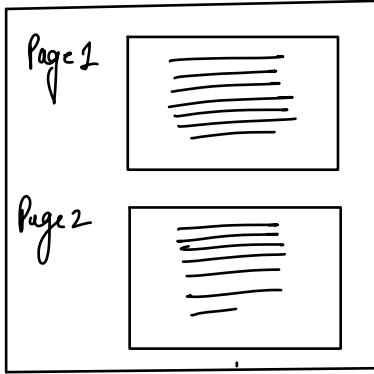
→ TAT ✓

→ 1 ticket → 2 hours
  2 × 20$
  → 40$

Chatbot

→ Take the prompt / Query / Question

→ Keeping the context

→ 50 chats At Max

→ 1 active chat (No concurrent chats)

→ Question based latency

  (Easy Question < 5 sec
   process related Que > 10 sec < 20 sec)

→ Moderation (OPENAI)

CRAWLER

Company → Diff pages

File System
├→ AWS S3
├→ HDFS
├→ GFS

| xyz | ...

Read

Get hyperlinks

all the outgoing links

Adding it in Queue
→ As we process the pages, the length
of the Queue would increase

Page 1

Page 2

EMBEDDINGS

Vectors
(n-dimensional)

① ├→ tokenisation
② ├→ cleaning
③ ├→ OpenAI, BERT
④ ├→ words → Numbers
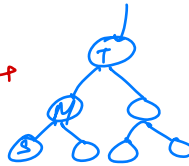
Vector DB

→ Stores & searches embedding efficiently
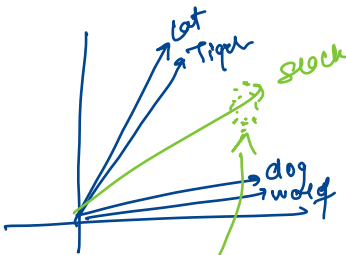
[FAISS] ←

↳ Store pre-computed vectors

↳ index them]
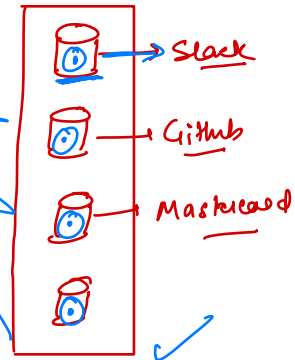
[What is Slack?]

embeddings →

Slack

Github

Mastercard

HNSW →

IVF → Inverted file index

Search on the basis of relevant cluster

Cat
Tiger
stock

dog
wolf

. . . .

Similarity search

Consine, inner product distance

⟹ rerank (?)

```
                    |
                    ↓
    ┌─────────────────────────┐
    │  Response               │
    │        Generation       │
    └─────────────────────────┘
              Transformer
```
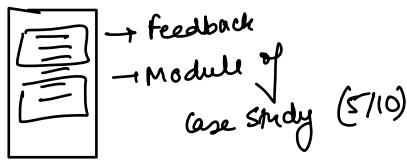
During a Query

1. Each query embedding to all or subset of shards

2. Each shard will do a ANN (Approximate nearest neighbour)
   & returns <u>top k</u> values

3. Results are merged & reranked globally to
   produce <u>top K</u> (finalick)

Horizontal Scaling
    └→ More data ? → Add shards.

Pinecone , Weaviate , Milvus } parallelism
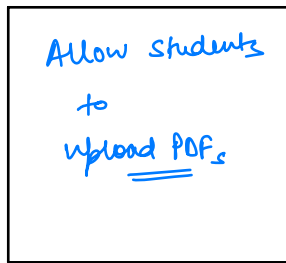   Query Routing

→ TA → TA

→ Feedback
→ Module of
Case Study (5/10)

Turn Around time for the Case Study evaluation

→ pdf format (1000s)
→ previous submission to train your model

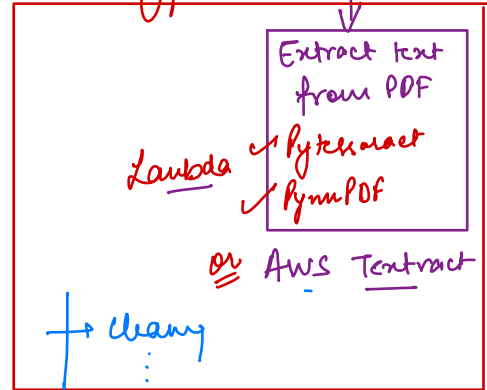
Allow students
to
upload PDFs

Student frontend
(Streamlit)

→ **AWS S3**

pipeline that is
triggered when
a new PDF is
uploaded

AWS
Lambda

PDF processing
trigger

Extract text
from PDF

Lambda ↦ Pytextract
↦ PymuPDF

or AWS Textract

↦ cleany
...

Data preprocessing

Langchain → Hugging Embed
→ OpenAI
Embed

[Sagemaker] Endpoint
↦ text-embedding small
↦ — large

OpenAI

endpoint

⇓

→ **FAISS** ✓

→ **Pinecone** ✓

Vector DBs
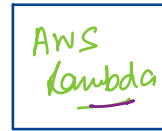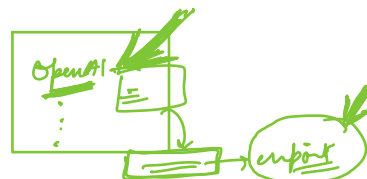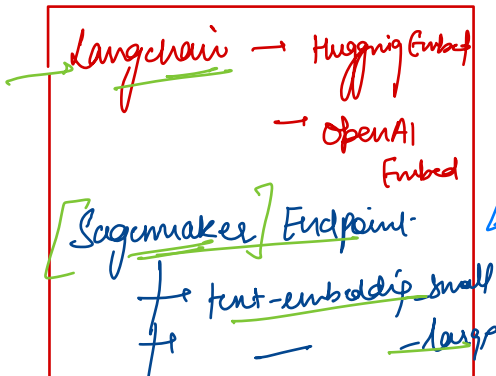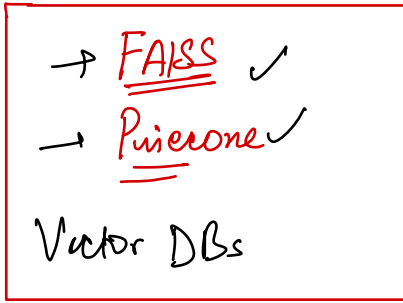
Pg Vector   Aws Aurora ✓

⇓ AWS SQS (Queue the text for evaluation)

✓ Aws Bedrock

[RAG Application]   ≡ ≡

https://aws.amazon.com/bedrock/knowledge-bases/

⇓
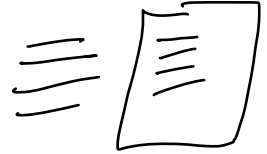
optional → Moderation

System Prompt

"You are an expert business consultant and a fair grader. Your task is to evaluate the provided business case based on the following criteria.

Evaluation Rubric:
1. **Problem Statement Clarity (20 marks):** Is the problem clearly defined, compelling, and supported by evidence?
2. **Solution Viability & Innovation (30 marks):** Is the proposed solution innovative, practical, and does it address the identified problem effectively?
3. **Market Analysis & Target Audience (20 marks):** Is the market size and target audience well-researched? Are competitive advantages clear?
4. **Business Model & Financial Projections (20 marks):** Is the business model sound? Are financial projections realistic and well-justified?
5. **Overall Presentation & Structure (10 marks):** Is the case well-organized, coherent, and professionally presented?

Instructions:
-   Analyze the student's business case thoroughly.
-   Assign a mark for each criterion out of the total specified.
-   Provide constructive feedback for each criterion, highlighting strengths and areas for improvement.
-   Conclude with an overall mark and general feedback.
-   Refer to the provided examples of previously evaluated cases and feedback to inform your assessment, focusing on similar good practices or common pitfalls.

Student Business Case:
<Insert extracted and relevant chunks of student's business case here>

Relevant Previous Cases/Feedback (for context/learning):
<Insert relevant chunks of previous cases and their feedback from vector DB here>

Output Format:
Evaluation Results
Student ID: [Student ID]
Business Case Title: [Title from case]

Marks
-   Problem Statement Clarity: [Score]/20
-   Solution Viability & Innovation: [Score]/30
-   Market Analysis & Target Audience: [Score]/20
-   Business Model & Financial Projections: [Score]/20
-   Overall Presentation & Structure: [Score]/10
Total Marks: [Total Score]/100

Feedback
Problem Statement Clarity:
[Specific feedback for this criterion]

Solution Viability & Innovation:
[Specific feedback for this criterion]

Market Analysis & Target Audience:
[Specific feedback for this criterion]

Business Model & Financial Projections:
[Specific feedback for this criterion]

Overall Presentation & Structure:
[Specific feedback for this criterion]

Overall Feedback:
[General constructive feedback]
"

(handwritten annotation) NO overwriting of system prompt
NO abusive

⇓

Storage

NoSQL -> DynamoDB

SQL → Redshift

S3 → Athena   ⊢ Student