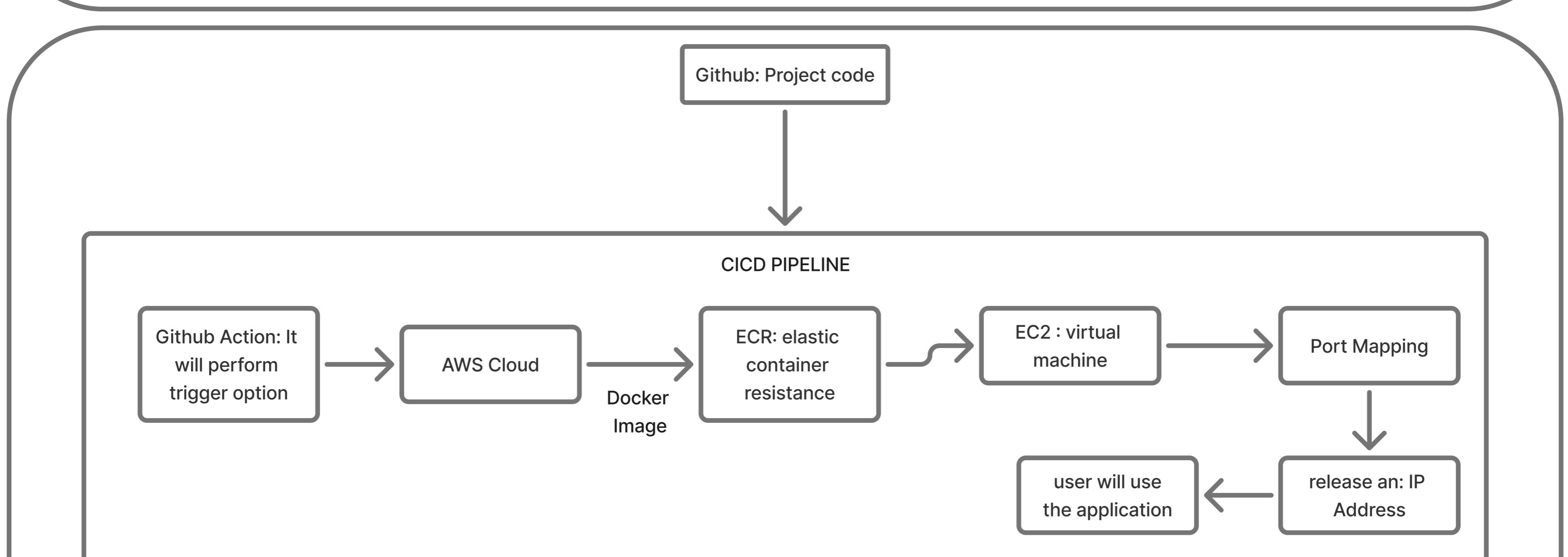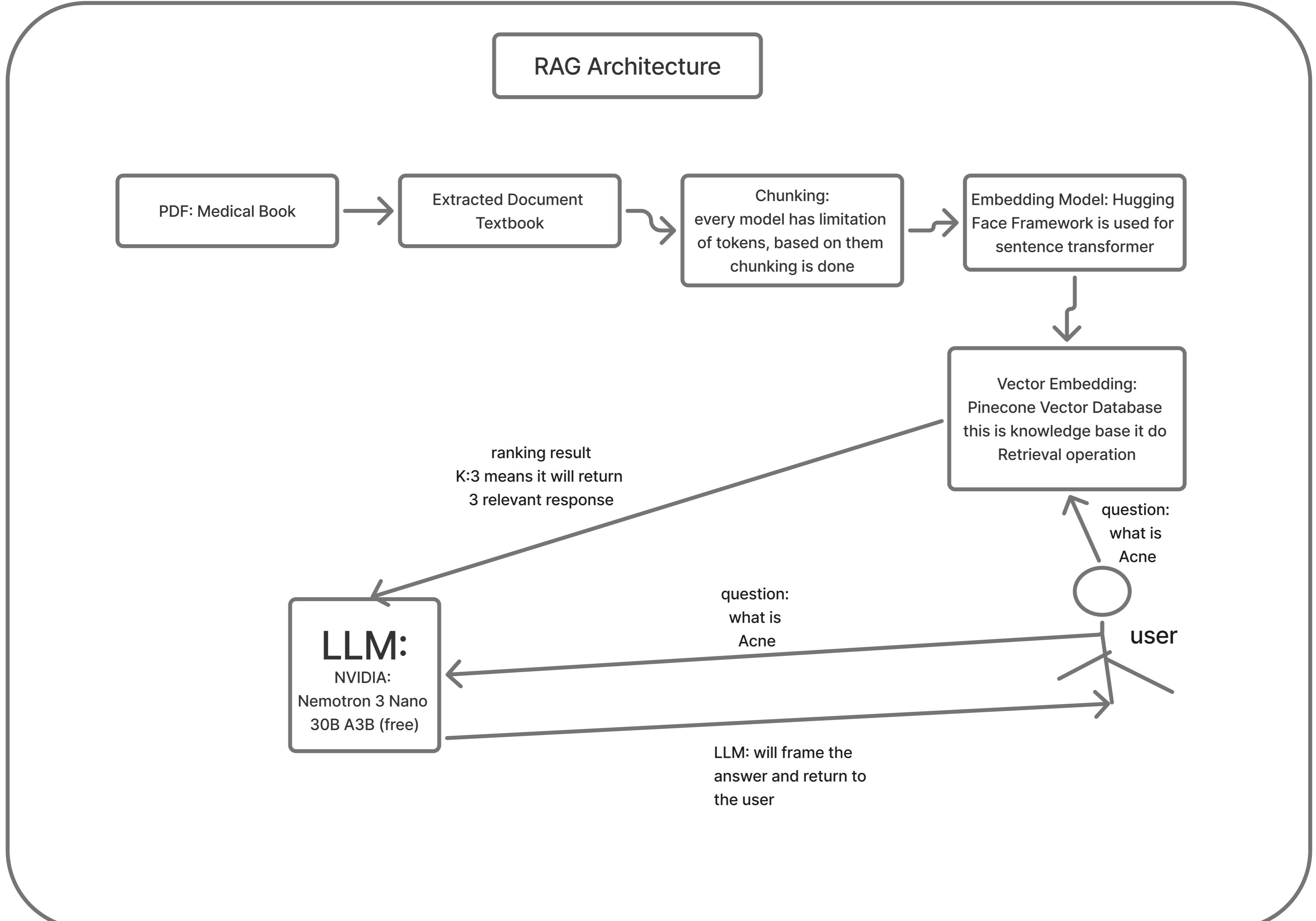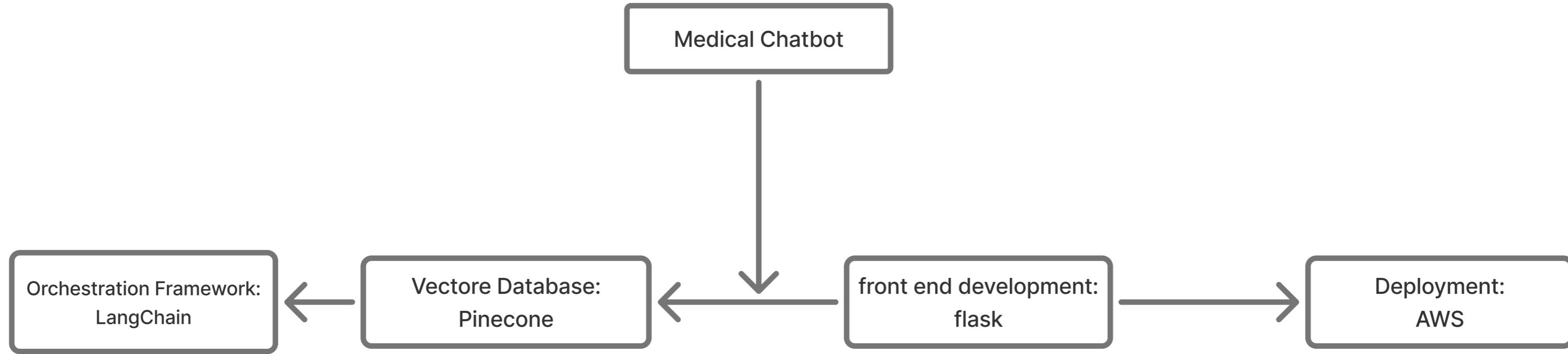#TechStack
1. PYTHON: Coding Language
2. LLM: NVIDIA: Nemotron 3 Nano 30B A3B (free from open router)
3. LANG CHAIN: Hugging Face Framework
4. VECTOR DATABASE: Pinecone
5. Frontend & Backend: Flask
6. Deploymemt of CICD pipline: AWS

Medical Chatbot

Orchestration Framework: LangChain ← Vectore Database: Pinecone ← Medical Chatbot → front end development: flask → Deployment: AWS

RAG Architecture

PDF: Medical Book → Extracted Document Textbook → Chunking: every model has limitation of tokens, based on them chunking is done → Embedding Model: Hugging Face Framework is used for sentence transformer

Vector Embedding: Pinecone Vector Database this is knowledge base it do Retrieval operation

ranking result
K:3 means it will return 3 relevant response

LLM:
NVIDIA: Nemotron 3 Nano 30B A3B (free)

question: what is Acne

question: what is Acne

user

LLM: will frame the answer and return to the user

Github: Project code

CICD PIPELINE

Github Action: It will perform trigger option → AWS Cloud → Docker Image → ECR: elastic container resistance → EC2 : virtual machine → Port Mapping

release an: IP Address → user will use the application

#STEP
1. Github Repo
2. Project Template
3. Project Setup
4. Jupiter Notebook Experiment
5. Modular Code
6. Web App
7. Deployment