LlamaLearn

CS-GY 9223 CLOUD COMPUTING

FINAL PROJECT

Team members:
Ajay Krishnan Gopalan - ag8172
Dhanesh Baalaji Srinivisan - ds7636
Shohna Kanchan - sk11239
Devyani Bairagya - db4922

PROBLEM STATEMENT

Challenge Addressed

Regular search engines and databases can have a hard time grasping customized user queries, leading to unrelated search results.

Proposed Solution - RAG Framework

The approach involves chunking and vectorizing text content for optimized indexing in OpenSearch, coupled with intelligent query processing and response generation using LLMs.

AWS COMPONENTS

Lambda

Opensearch

API Gateway

EC2

EKS

ECR

DynamoDB

Textract

S3

Cognito

CloudWatch

ARCHITECTURE

01

Modular Design with Containerization

Lambda functions, microservices, and EC2 instances in clusters to ensure **flexibility** and **scalability**. This facilitates seamless integration of new text embedding techniques or LLM models through containerization.

02

Lambda Functions as System Backbone

System resilience through Lambdas, designed for specific tasks or managing outputs from microservices. **Decoupling** reduces inter-dependencies, making functions **independent modules for easy deployment and configuration.**

03

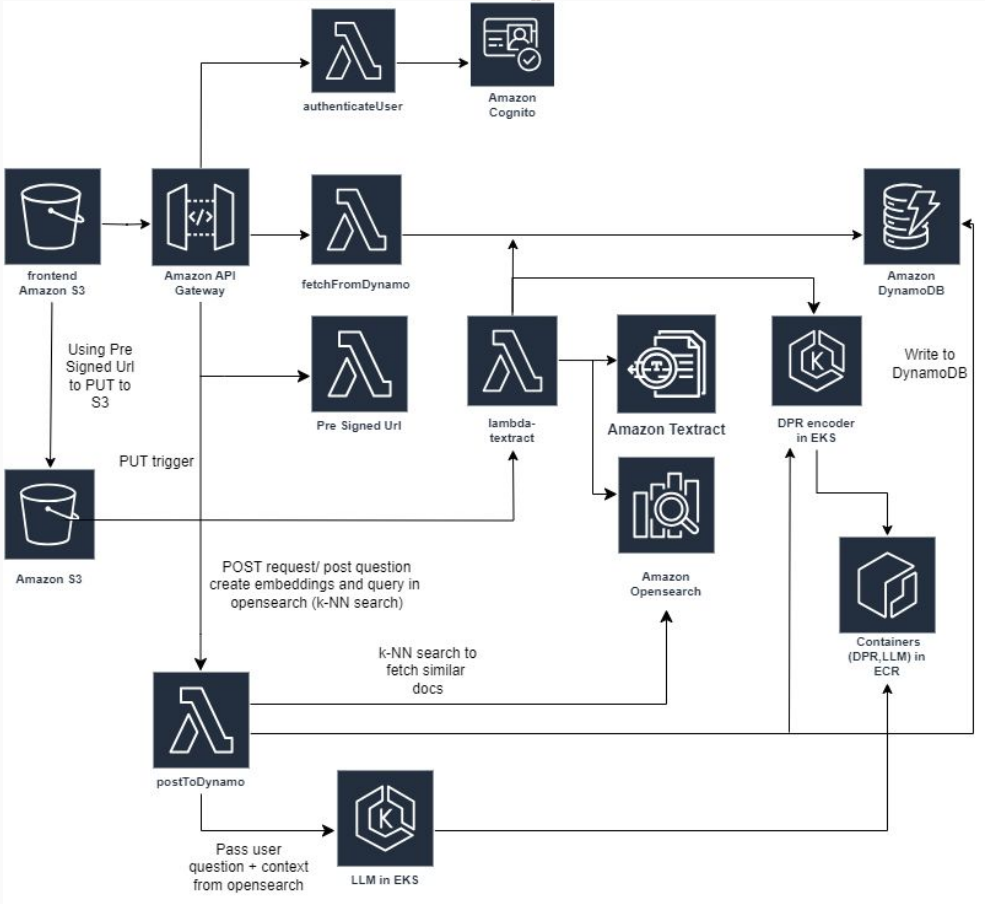
EC2 Clusters for Robust Computing

High availability and load balancing achieved through the strategic use of EKS clusters, providing robust computational resources. This ensures efficient processing and meets demands for a modern, scalable architecture.

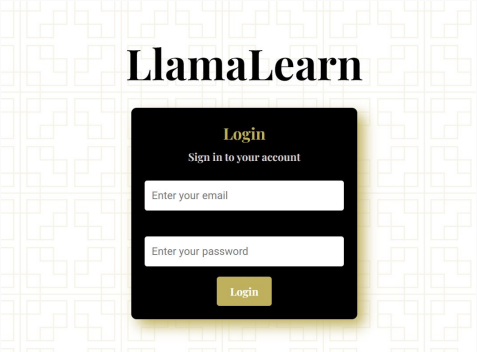
04

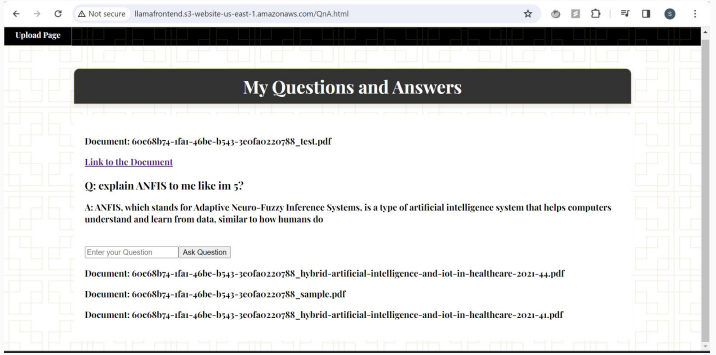
Future-Ready with Streamlined Setup

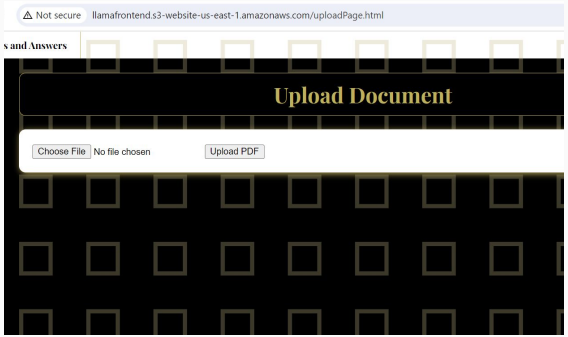
The architecture is ready for future demands. The utilization of modular components allows for effortless deployment, initialization, and configuration, showcasing unparalleled efficiency in setup processes.


authenticateUser
Amazon Cognito
frontend Amazon S3
Amazon API Gateway
fetchFromDynamo
Amazon DynamoDB
Using Pre Signed Url to PUT to S3
Pre Signed Url
lambda-textract
Amazon Textract
DPR encoder in EKS
Write to DynamoDB
PUT trigger
Amazon S3
POST request/ post question create embeddings and query in opensearch (k-NN search)
Amazon Opensearch
Containers (DPR,LLM) in ECR
k-NN search to fetch similar docs
postToDynamo
Pass user question + context from opensearch
LLM in EKS

Visual Overview







Thanks!

Any questions?