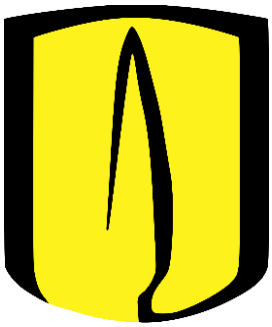
DESARROLLO DE SOLUCIONES CLOUD



Proyecto 1 – Entrega 1

Integrantes:

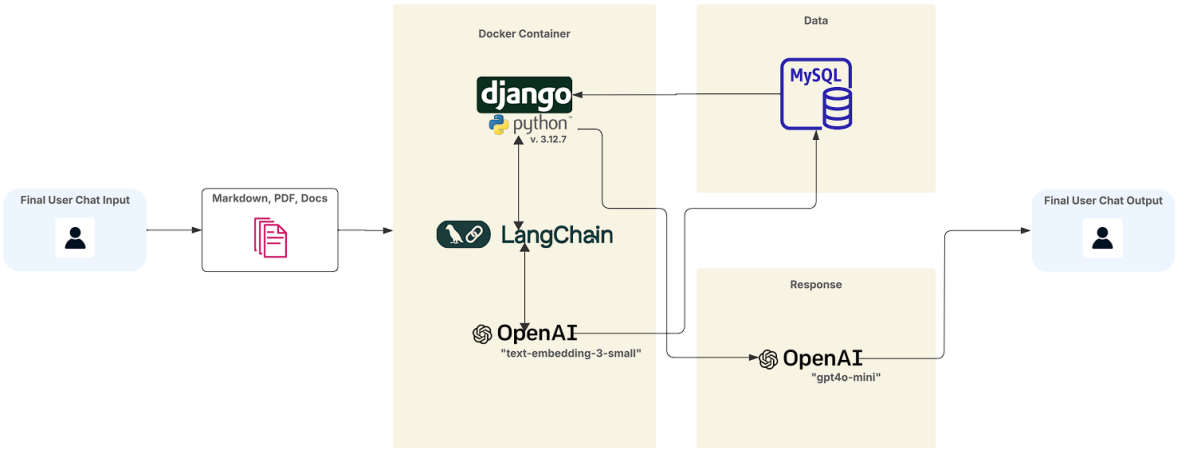
Esteban Caicedo Graciano

Steban Nicolas Tibata Castañeda

Nicolas Miguel Murillo Cristancho

Documentación de arquitectura

Diagrama de arquitectura y despliegue:



La arquitectura sigue un patrón MVC, constando de front (django) langchain (que cumple el papel del controlador) y finalmente el api, siendo el back y la última capa de código.

Este API a su vez llama al api de OpenAi y usa la base de datos MySQL de maner local.

La app es desplegada en una máquina con 2 vCPU, 4 GiB RAM y 20 GiB en almacenamiento

Servicios: la clave de la arquitectura para la escalabilidad y la concurrencia:

La aplicación usa los servicios del API de OpenAI **(**Según autorización dada por el profesor via email) Por lo que la app se concentra en servir el front y el api, a la vez que manejar todos los datos necesarios en la base de datos MySQL para gestionar los chunks y el RAG que posteriormente usará el modelo al responder cada prompt del usuario.

Lo anterior permite una gran concurrencia ya que al tercerizar el uso del modelo este se podrá llamar de manera paralela por cada usuario. Es decir, que en la capa en donde se ejecuta el modelo, cada solicitud no dependerá de otras solicitudes concurrentes. La concurrencia únicamente podría verse afectada por falta de capacidad de cómputo al ejecutar el front o el API, que son componentes mucho más ligeros que la máquina virtual moverá sin problema, estimamos, para varios de cientos de usuarios al tiempo.

Esto se medirá en las pruebas de carga, en caso de no ser posible, para garantizar la concurrencia se deberá aumentar la capacidad de la máquina mediante AWS y/o crear nuevas máquinas con mayor capacidad (memoria, cpus, etc.) e incluir balanceadores de carga.

Limitaciones:

1. La autenticación no ofrece métodos avanzados como autenticación en dos pasos, recuperación de cuenta, etc.
2. Solo se soportan archivos de formato word, pdf, markdown y txt
3. El número de palabras en el input/output está limitado a 96.000 palabras
4. La disponibilidad está sujeta a disponibilidad del API del proveedor