Servidor de IA para CENS

Informe Técnico-Económico

El costo de la instancia en Amazon **lightsail** es **$164** al mes, pero la performance no es adecuada, es muy lenta. El Tiempo de respuesta es muy alto, mayor a 3-5 minutos.

La recomendación es utilizar una instancia en EC2 con GPU, esta tiene un costo por hora de $1.013, se puede limitar el costo limitando las horas de uso, en producción puede ser 12 horas o 24 horas, trasladando el costo al cliente o utilizar activadores por uso.

Amazon AWS EC2 - Costos Variables

Precio Ubuntu ($/hr.)	$1.013

Item	Desarrollo	Producción 1	Producción 2
Restricciones	Horas mínimas	Disponibilidad 12 horas.	Disponibilidad 24 hrs.
horas/dia	65	12	24
días/mes	20	24	30
Monto	**$122**	**$292**	**$729**

Arquitectura propuesta (eficiente y privada)

Usar disparadores para utilizar la instancia sólo cuando hay actividad y/o por periodos.

Componentes

1. **Frontend**: Aplicación cliente carga PDFs y solicita resúmenes y etiquetas.

2. **API Gateway + Lambda** (mínimo costo): recibe la solicitud y actúa como “disparador”.

3. **Lambda**:
   - Revisa si la instancia EC2 está apagada.
   - La enciende (si es necesario).
   - Espera hasta que esté disponible.
   - Redirige la solicitud al backend (Ollama).

4. **Instancia EC2 `g5.xlarge`**: (**Servidor de IA de CENS**)
   - Solo se inicia si hay trabajo que hacer.
   - Procesa el PDF con el modelo `llama3` local.
   - Devuelve resumen y etiquetas clave.

5. **Lambda o cron**: monitorea inactividad y **apaga la instancia** si nadie la usa tras X minutos.

Resp. Luis Ivan Umpire Alvarez.