IA Local vs Cloud

Compliance global y qué modelo se adapta a cada normativa

Si tu sistema procesa datos de clientes y usás un modelo en la nube, esos datos están saliendo de tu infraestructura. No es teórico, es lo que pasa cada vez que tu agente le manda un prompt a un proveedor externo. El problema no es que la nube sea mala, es que la mayoría de las empresas no chequean si el flujo cumple con la normativa que aplica a su industria.

Este documento es un resumen ejecutivo: qué pide cada normativa global que importa, cuándo cloud sigue siendo viable, cuándo local es la única opción, y qué modelos open-source corren en hardware razonable.

1. La diferencia que cambia todo

Un modelo en la nube procesa cada conversación a través de internet. Los datos viajan al proveedor, se procesan en su infraestructura, y según el contrato pueden quedar logueados, usarse para entrenar modelos futuros, o estar sujetos a leyes del país donde está el datacenter.

Un modelo local corre 100% dentro de tu infraestructura. Los datos nunca salen. La desventaja: requiere inversión en hardware (GPU, RAM, almacenamiento) y alguien que mantenga el stack. La ventaja: el costo marginal por inferencia tiende a cero, y el control sobre los datos es total.

Cuándo importa esta decisión

- Procesás datos de salud, financieros, biométricos o de menores
- Tu industria está regulada (banca, seguros, salud, gobierno)
- Tenés clientes en la UE, en California, o sujetos a HIPAA
- Firmaste cláusulas de confidencialidad que prohíben subprocesadores no listados
- El volumen de tokens mensual ya supera lo que costaría amortizar hardware propio

2. Normativas globales que aplican

No es una guía legal. Es el mapa que uso para decidir arquitectura antes de hablar con el equipo legal del cliente. Si alguno de estos marcos aplica a tu negocio, la decisión cloud vs local deja de ser técnica y pasa a ser de cumplimiento.

Normativa	Aplica a	Restricción clave para IA
GDPR (UE)	Cualquier empresa que procese datos de residentes UE, esté donde esté	Transferencia internacional de datos restringida. Subprocesadores deben estar listados y aprobados. Derecho al olvido y a no ser sujeto de decisiones automatizadas. Multas hasta 4% facturación global.
HIPAA (US, salud)	Proveedores de salud, aseguradoras y cualquier subprocesador que toque PHI	Requiere BAA (Business Associate Agreement) firmado con cada proveedor que procese datos de salud. La mayoría de APIs públicas de IA no firman BAA en sus tiers gratuitos o estándar.
SOC 2 (Type II)	Empresas SaaS B2B, especialmente vendiendo a US enterprise	No es ley, es certificación. Auditan cómo manejás datos de clientes, incluyendo subprocesadores. Cloud está OK si el proveedor también es SOC 2 y está documentado en tu registro de subprocesadores.
CCPA / CPRA (California)	Empresas con clientes en California sobre cierto umbral de ingresos o registros	Derecho a opt-out de venta o compartición de datos personales. Si tu proveedor de IA usa los datos para entrenar, eso puede contar como compartición.
EU AI Act	Sistemas de IA usados en UE, clasificados por nivel de riesgo	Sistemas de alto riesgo (RRHH, crédito, salud, educación, biometría) requieren documentación, trazabilidad de datos y supervisión humana. Implementación gradual hasta 2027.
PCI DSS	Cualquier sistema que toque datos de tarjetas de crédito	Datos de tarjetas no deberían pasar por modelos de IA, punto. Si tu agente atiende consultas de pagos, asegurate que los números nunca lleguen al prompt.

Nota: Esto no reemplaza asesoría legal. Si tu producto opera en una industria regulada, validá la arquitectura con un compliance officer antes de mover algo a producción.

3. Qué modelo elegir según el caso

La regla simple: si los datos que procesás están sujetos a una de las normativas de arriba, empezás por descartar opciones cloud y después ves cuál te queda. Si no hay normativa que aplique, cloud sigue siendo la opción de menor fricción para arrancar.

Escenario	Cloud OK	Local recomendado	Por qué
Atención al cliente general (FAQ, soporte L1)	Sí	Opcional	Datos no sensibles. Cloud reduce time-to-market.
Procesamiento de historias clínicas o PHI	No (sin BAA)	Sí	HIPAA exige control total del flujo. Local elimina el subprocesador.
Análisis de contratos o documentos confidenciales B2B	Condicional	Recomendado	Cláusulas de NDA suelen prohibir subprocesadores no listados.
Datos personales de residentes UE	Solo con DPA y datacenter UE	Sí, si hay dudas	GDPR es estricto con transferencias internacionales.
Datos financieros o de tarjetas	No para datos de tarjeta	Sí para análisis	PCI DSS no contempla LLMs como entorno seguro.
Producto interno (uso solo de empleados)	Sí	Si hay IP sensible	Riesgo bajo salvo que se manejen secretos comerciales.

4. Modelos locales y hardware mínimo

Lo que sigue son modelos open-source que corrí o evalué para casos de comunicación y agentes. El hardware listado es lo mínimo para que el modelo sea usable en producción, no lo mínimo para que arranque. La diferencia importa: un modelo cargado con swap a disco te da 1 token por segundo y eso no es producto, es demo.

Todos los modelos asumen cuantización GGUF Q4_K_M o equivalente, que es el sweet spot entre calidad y consumo de VRAM para self-hosting. El runtime asumido es llama.cpp, Ollama o vLLM según el caso.

Modelo	Caso de uso	VRAM mínima	RAM sistema	Notas
Llama 3.1 8B Instruct	Agentes de chat, FAQ, clasificación, RAG simple	8 GB	16 GB	Buen balance. Corre en una RTX 3060 12GB sin drama.
Llama 3.3 70B Instruct	Razonamiento complejo, agentes con tool calling sofisticado	48 GB	64 GB	Requiere 2x RTX 3090/4090 o una A6000. Calidad cercana a GPT-4 turbo.
Mistral Small 3 24B	Agentes de producción, atención al cliente con contexto	16 GB	32 GB	Sweet spot calidad/costo. RTX 4090 24GB lo corre cómodo.
Qwen 2.5 14B	Multilingüe (incluye español rioplatense decente), tool calling	12 GB	32 GB	Mejor que Llama 8B en español. Buena licencia para uso comercial.
Phi-4 14B	Razonamiento, tareas estructuradas, baja latencia	10 GB	16 GB	Microsoft. Muy fuerte en reasoning para su tamaño.
Whisper Large v3	Speech-to-text local (transcripción de llamadas)	10 GB	16 GB	Reemplaza la STT de OpenAI en pipelines de voz.
Piper TTS	Text-to-speech local, voz sintética	CPU	4 GB	Corre en CPU, latencia muy baja. Calidad menor a ElevenLabs pero suficiente para IVR.

Hardware base recomendado para arrancar self-hosting de un modelo de 8B-14B en producción: una RTX 4090 24GB, 64GB de RAM DDR5, 1TB NVMe, Ubuntu 22.04. Inversión inicial alrededor de USD 3000-3500. Para 70B vas a doble GPU o a una A6000 (USD 5000-7000 solo la GPU).

5. Cuándo cloud sigue siendo la mejor decisión

Local no es la respuesta correcta para todos los casos. Si tu producto no procesa datos regulados, no firmaste NDAs estrictos, y tu volumen todavía es bajo, montar infraestructura propia es over-engineering.

Cloud sigue ganando cuando:

- Estás validando un producto y todavía no sabés si va a tener tracción
- El volumen mensual de tokens te cuesta menos de USD 500 al mes
- Necesitás capacidades específicas que solo están en frontier models (Claude Opus, GPT-5 Pro, Gemini 2 Pro) y los modelos open-source todavía no llegan ahí
- No tenés equipo para mantener infraestructura propia
- Tu caso de uso permite firmar un DPA estándar con OpenAI/Anthropic/Google y eso te alcanza para compliance

Híbrido como tercera opción

En la mayoría de los proyectos reales que implementé, la arquitectura termina siendo híbrida: el modelo local maneja datos sensibles (clasificación de PII, transcripción, embeddings) y el modelo en la nube se usa solo para tareas donde la calidad importa y los datos ya fueron anonimizados o redactados. Es más laburo de arquitectura, pero te da lo mejor de ambos mundos.

Cómo decido en la práctica

Antes de elegir arquitectura, mi checklist es siempre el mismo: qué normativa aplica, qué firma el cliente con sus usuarios, qué subprocesadores tiene listados, y cuál es el costo proyectado a 12 meses de cloud vs amortización de hardware. Con eso sobre la mesa, la decisión se cae sola en el 80% de los casos. El otro 20% requiere conversación con legal y eso ya no es problema técnico.

Gonzalo Gomez · Especialista en IA aplicada a sistemas de comunicación y automatización · @ggomezdev