MEDAI · TOOLDATA SPA

Arquitectura de MedAI — resumen

Síntesis para TI del despliegue on-premise, VPC o air-gapped

Activo de confianza · MedAI en pre-lanzamiento · julio de 2026

MedAI (Tooldata SpA) es una plataforma de **IA clínica de apoyo** para clínicas y hospitales de Chile, con arquitectura **human-in-the-loop**: cada salida es un **borrador** que un profesional revisa, edita, valida y **firma**. Se despliega **en la infraestructura del cliente** (on-premise o VPC privada) para que los datos clínicos permanezcan **bajo su control**. No emite diagnósticos autónomos ni reemplaza el juicio médico.

> Documento de síntesis del *whitepaper* de arquitectura de referencia. MedAI está en **pre-lanzamiento**: describe el **diseño objetivo**; algunos componentes están en construcción y las cifras son **referenciales**, se ajustan en cada implementación.

Principios de diseño

- **Soberanía de datos.** En on-premise/air-gapped, los datos clínicos **no salen de la red del cliente**; la inferencia ocurre localmente. En VPC privada, operan en un **entorno dedicado bajo control del cliente**.
- **Sin entrenamiento con datos del cliente.** Los datos no se usan para entrenar modelos de uso general ni de terceros, ni se comparten.
- **Human-in-the-loop.** Toda salida es un borrador de apoyo; la validación y firma de un profesional habilitado es obligatoria y en él permanece la responsabilidad clínica.
- **Integración por estándares, no reemplazo.** Se conecta al stack existente (HIS/EMR, PACS/RIS) mediante estándares médicos.
- **Air-gapped opcional.** Puede operar sin salida a internet; las actualizaciones llegan como paquetes firmados que el cliente importa de forma controlada.
- **Trazabilidad por diseño.** Cada salida queda registrada con autoría, versión de modelo y de plantilla, y marca de tiempo.

Componentes clave

- **API Gateway** — ingreso único con TLS, autenticación **SSO**, *rate limiting* y enrutamiento.
- **Orquestador MedAI** — lógica de negocio, gestión de plantillas/prompts, **cola human-in-the-loop**, validación/firma y trazabilidad.
- **Servidor de inferencia (GPU local)** — ejecuta los modelos localmente (modelos cuantizados).
- **Conectores de integración** — **DICOM (C-STORE/C-FIND)**, **HL7 v2**, **FHIR** y REST hacia PACS/RIS/HIS.
- **PostgreSQL** — metadatos, cola de revisión, **auditoría** y versionado; almacenamiento de objetos en la infra del cliente.
- **Gestor de actualizaciones firmadas** — importa y **verifica releases firmados** (cosign/sigstore), con opción de *rollback*.
- *Complementan:* federación de **identidad** (SAML/OIDC con el IdP del cliente) y **observabilidad** (logs/métricas) dentro del perímetro del cliente.

Modelos de despliegue

Modelo	En una línea
On-premise	Servidor o clúster del cliente; máximo control; los datos no salen de su red; soporta air-gapped.
VPC privada	Entorno dedicado y aislado en la nube elegida por el cliente, **bajo su control**.
Híbrido	Combinación según sede/carga (p. ej. inferencia local + gestión centralizada).
Air-gapped	Sin salida a internet; actualizaciones exclusivamente por paquete firmado.

Integración (diseño por estándares)

MedAI se **diseña para integrarse** al flujo existente mediante estándares médicos, no como integraciones ya probadas en producción:

- **Imagenología:** DICOM **C-STORE / C-FIND** hacia PACS/RIS.
- **Sistemas clínicos: HL7 v2** y **FHIR** hacia HIS/EMR; **REST** para sistemas personalizados.
- **Identidad:** inicio de sesión único **SSO (SAML/OIDC)** federado con el IdP del cliente (AD/Entra/Keycloak).

Seguridad

- **Cifrado** en tránsito (TLS) y en reposo (AES).
- **RBAC + SSO** (SAML/OIDC) con principio de mínimo privilegio.
- **Sin egress** en modo air-gapped; segregación de entornos.
- **Auditoría y versionado** de todas las salidas (trazabilidad completa; qué versión generó cada resultado).
- **Actualizaciones firmadas y verificadas** antes de aplicar, con *rollback*; respaldos y plan de recuperación.

Dimensionamiento de referencia

Cifras referenciales para planificar. El dimensionamiento definitivo se valida en cada implementación según volumen real, modelos, latencia objetivo y alta disponibilidad.

Escenario	Carga aprox.	CPU	RAM	GPU (VRAM)	Almacenamiento
Piloto — lenguaje clínico	hasta ~300 epicrisis/día, baja concurrencia	8 vCPU	32 GB	1× GPU 24 GB (LLM ~7–14B cuantizado)	200 GB SSD
Producción — lenguaje clínico	clínica mediana, concurrencia media	16 vCPU	64 GB	1–2× GPU 48 GB	1 TB SSD
Imagenología (piloto)	volumen acotado por modalidad	16 vCPU	64 GB	1× GPU 48 GB (modelos de visión)	1 TB+ SSD

Alta disponibilidad: duplica los recursos del plano crítico (×2 nodos/GPU, almacenamiento replicado). Imagenología y procesamiento DICOM tienden a requerir más VRAM y almacenamiento; se dimensionan caso a caso.

Marco normativo (Chile)

Diseñado para apoyar el cumplimiento de la **Ley 21.719** (protección de datos; plena vigencia el **1 de diciembre de 2026**; datos de salud = **sensibles**), considerando el secreto médico y la ficha clínica (**Ley 20.584**) y la interoperabilidad HL7 FHIR (**Ley 21.668**). Chile está incorporando el software como dispositivo médico (**SaMD**) al control sanitario del **ISP**; evaluamos el alcance de cada caso de uso con asesoría regulatoria. Se aportan insumos como contrato de encargo (DPA) y plantilla de evaluación de impacto (EIPD/DPIA).

Siguiente paso

- Solicitar el **whitepaper completo** de arquitectura de referencia (principios, diagrama, seguridad, empaquetado y actualizaciones air-gapped).
- Agendar una **revisión técnica** para dimensionar un piloto y validar la integración con su HIS/EMR y PACS/RIS.

Contacto: Tooldata SpA — MedAI · Documento de arquitectura de referencia, julio de 2026. Sujeto a evolución del producto.