MEDAI · TOOLDATA SPA

Arquitectura de referencia on-premise

Despliegue de MedAI en la infraestructura del cliente — on-premise, VPC privada o air-gapped

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

Naturaleza del documento. Describe la **arquitectura de referencia (diseño objetivo)** con la que MedAI se despliega en la infraestructura del cliente. MedAI está en **pre-lanzamiento**: algunos componentes están en construcción y este documento es a la vez una guía de diseño y un activo de due-diligence técnica para clínicas y hospitales. Las cifras de dimensionamiento y rendimiento son **referenciales** y se ajustan en cada implementación.

1. Principios de diseño

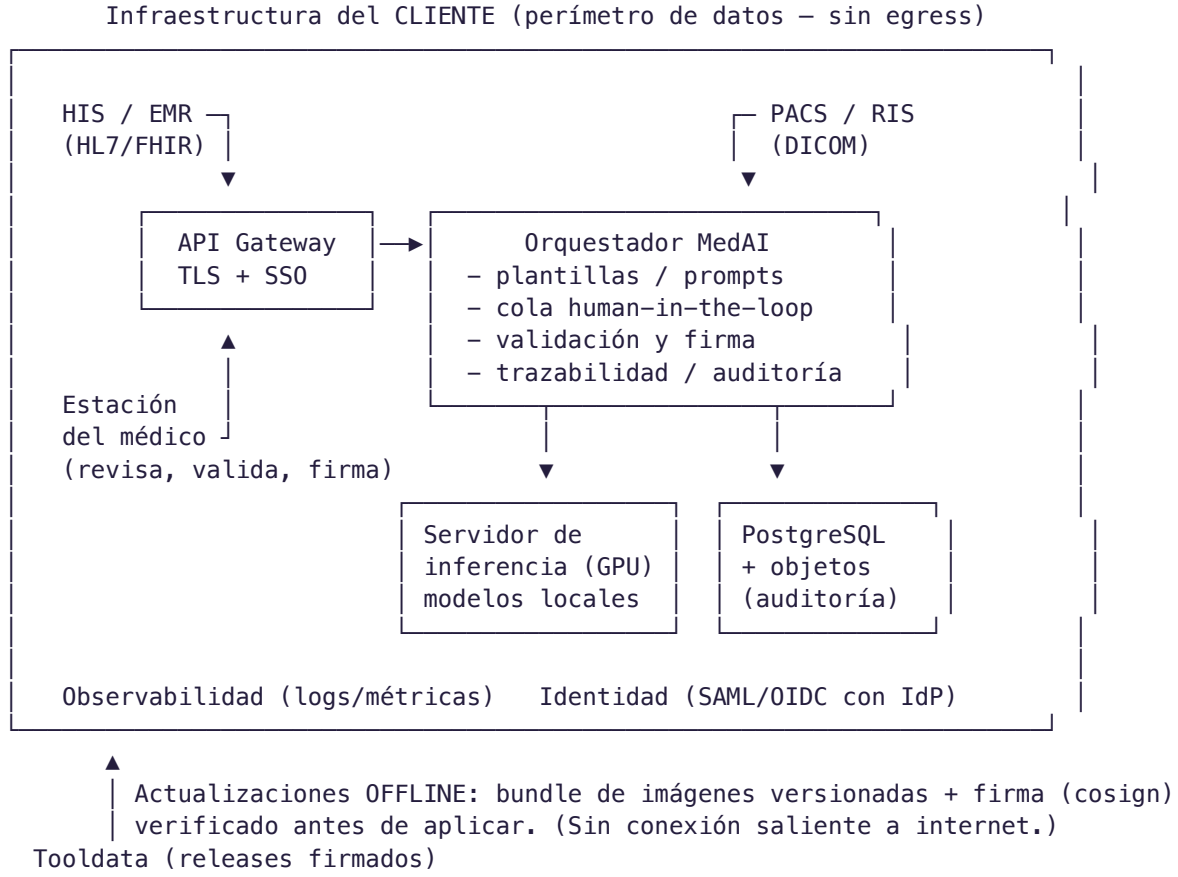
1. **Soberanía de datos:** los datos clínicos **no salen** de la infraestructura del cliente. La inferencia ocurre localmente.
2. **Sin entrenamiento con datos del cliente:** los datos no se usan para entrenar modelos de uso general ni de terceros.
3. **Human-in-the-loop:** toda salida es un **borrador** que un profesional revisa, valida y firma.
4. **Integración, no reemplazo:** MedAI se conecta al stack existente (HIS/EMR, PACS/RIS) por estándares.
5. **Operable sin internet (air-gapped):** las actualizaciones se entregan como paquetes firmados que el cliente importa de forma controlada.
6. **Trazabilidad por diseño:** cada salida queda registrada con autoría, versión de modelo y de plantilla.

2. Componentes lógicos

Componente	Función	Tecnología de referencia
API Gateway	Ingreso único; TLS; autenticación SSO; rate limiting; enrutamiento	Reverse proxy (nginx/Envoy)
Orquestador MedAI	Lógica de negocio, gestión de plantillas/prompts, cola human-in--the-loop, validación/firma, trazabilidad	Servicio de aplicación
Servidor de inferencia	Ejecuta los modelos localmente sobre GPU (modelos cuantizados)	vLLM / Triton / Ollama
Conectores de integración	DICOM (C-STORE/C-FIND), HL7 v2, FHIR, REST hacia PACS/RIS/HIS	Adaptadores estándar
Base de datos	Metadatos, cola de revisión, registros de auditoría y versionado	PostgreSQL
Almacenamiento de objetos	Artefactos temporales, imágenes (si aplica)	S3-compatible (MinIO) en infra del cliente

Componente	Función	Tecnología de referencia
Identidad	Federación con el IdP del cliente	SAML / OIDC (AD/Entra/Keycloak)
Observabilidad	Logs estructurados, métricas, salud de servicios — dentro del perímetro del cliente	Prometheus / Grafana / logs
Gestor de actualizaciones	Importación y verificación de releases firmados (offline)	Imágenes versionadas + firma (cosign/sigstore)

3. Diagrama de referencia



4. Modelos de despliegue

- **On-premise:** servidor o clúster del cliente. Máximo control; soporta air-gapped.
- **Nube privada (VPC):** entorno dedicado gestionado, aislado, en la nube elegida por el cliente.
- **Híbrido:** combinación según sede/carga (p. ej. inferencia local + gestión centralizada).
- **Air-gapped:** sin salida a internet; actualizaciones exclusivamente por paquete firmado.

5. Flujo de datos (ejemplo: lenguaje clínico)

1. El HIS/EMR envía el contexto clínico (HL7/FHIR) al **API Gateway** (autenticado por SSO).
2. El **Orquestador** arma la solicitud con la plantilla correspondiente y la envía al **servidor de inferencia local**.
3. El modelo genera un **borrador** (epicrisis/resumen + propuestas CIE-10).
4. El borrador entra en la **cola human-in-the-loop**; el profesional **revisa, edita y firma**.
5. El documento validado se devuelve al HIS/ficha clínica; queda **registro de auditoría** (autor, versión de modelo y plantilla, marca de tiempo).

En ningún paso los datos clínicos abandonan la infraestructura del cliente.

6. Seguridad (resumen; el detalle está en el Dossier de seguridad de MedAI, documento aparte)

- Cifrado en tránsito (TLS) y en reposo (AES).
- RBAC + SSO (SAML/OIDC); principio de mínimo privilegio.
- Segregación de entornos; sin egress en modo air-gapped.
- Registros de auditoría y versionado de salidas.
- Retención configurable, enmascaramiento/anonimización donde aplique.
- Respaldos y plan de recuperación; rollback por versión.

7. Empaquetado y entrega

Escenario	Empaquetado
Piloto (un nodo con GPU)	`docker-compose` de referencia (ver `docker-compose.reference.yml`)
Producción / alta disponibilidad	**Helm chart** sobre el Kubernetes del cliente
Appliance de referencia	Especificación de hardware llave en mano (ver `sizing.md`)

8. Actualizaciones offline (air-gapped)

1. Tooldata publica un **release versionado** (imágenes de contenedor) y su **firma** (cosign/sigstore).
2. El cliente recibe el **bundle** (descarga controlada o medio físico).
3. El cliente **verifica la firma** antes de importar (cadena de confianza).
4. Importa las imágenes y **aplica** la actualización; puede **revertir** a la versión anterior.

No se requiere conexión saliente a internet en ningún momento.

9. Operación y niveles de servicio

La disponibilidad y los tiempos de soporte se definen **por contrato y según el dimensionamiento** acordado (arquitectura de HA, redundancia, ventana de soporte). En pre-lanzamiento no se comprometen cifras de uptime sin un diseño de HA validado para cada cliente.

10. Requisitos de dimensionamiento

Ver **`sizing.md`** para una guía referencial de CPU/RAM/GPU por volumen y modalidad.

Documento de arquitectura de referencia. Junio de 2026. Sujeto a evolución del producto.