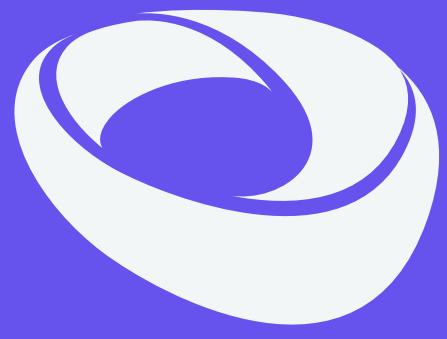


MISO
Maestría en Ingeniería de Software

Proyecto: SaludTech de los Alpes

2025

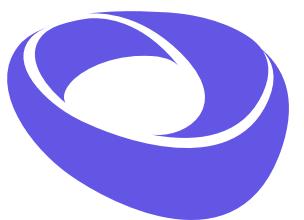
Descripción del proyecto del curso



SaludTech de los Alpes

Contexto

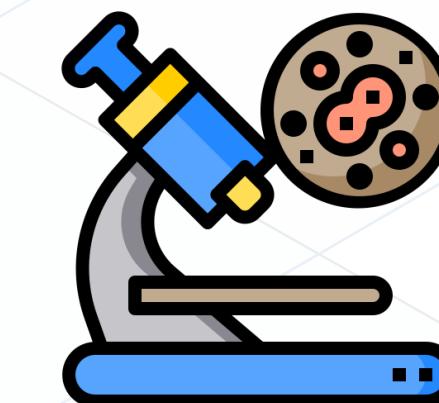
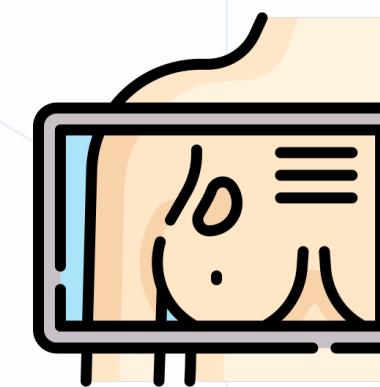
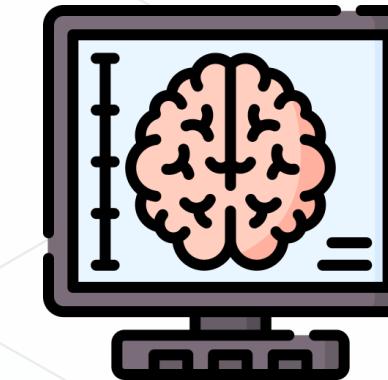
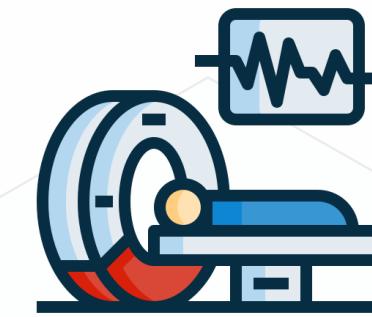
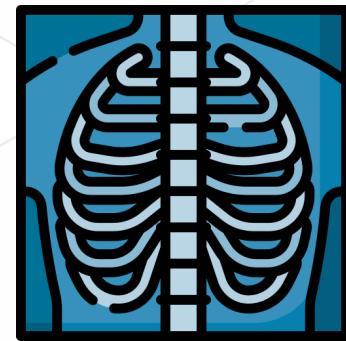
- **SaludTech de los Alpes (STA)** es una compañía colombiana constituida en 2021 que se desenvuelve en la industria del **Vertical AI**. Su enfoque principal es recaudar, procesar y distribuir imágenes médicas y diagnósticos anonimizados a compañías de IA y desarrolladores. Dichas imágenes y metadatos sirven como fuente de entrenamiento para modelos de IA.
- La compañía inició operaciones en Bogotá, siendo capaz de recolectar datos de unos más de 100 centros de salud (hospitales, clínicas y laboratorios). En el transcurso de los últimos 3 años, se han expandido a otros países de Sudamerica: Argentina, Ecuador, Chile y Brasil, siendo este último su mayor fuente de recolección de datos.
- A finales del año 2024, STA recibió financiación para poder comenzar su expansión global por el resto de Latinoamérica y Estados Unidos.



SaludTech de los Alpes



Contexto



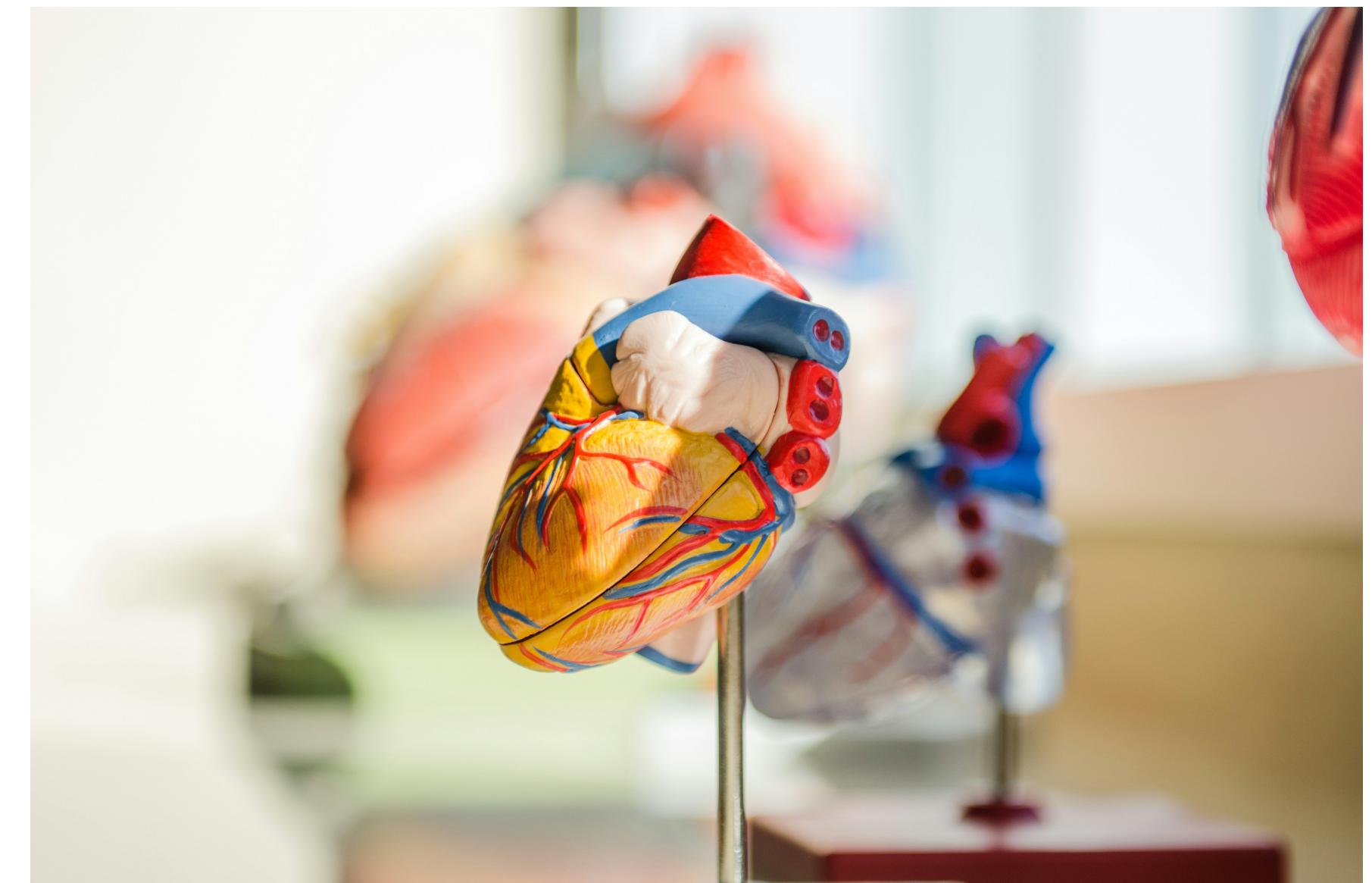
- Modalidad

- La compañía es capaz de procesar diferentes tipos de imágenes, categorizadas por la modalidad, región anatómica y patología/ condición.
- En caso de la modalidad encontramos categorías como:
 - Rayos X
 - Tomografías
 - Resonancia magnética
 - Ultra sonido
 - Mamografía
 - Escaneo TEP
 - Histopatología

Contexto

- Región Anatómica

- Toda imagen tiene una correlación con una parte del cuerpo. Entre las categorías más relevantes encontramos:
 - **Cabeza y cuello:** Cerebro, conducto seno, tiroides, etc.
 - **Tórax:** Pecho, corazón, pulmones, etc.
 - **Abdomen:** Hígado, pancreas, riñón, etc.
 - **Musculoesquelético:** Huesos, articulaciones, tejido blando, etc.
 - **Pélvis:** Órganos reproductivos, vejiga, etc
 - **Cuerpo completo:** Biopsias



Contexto



- Patología o condición

- Las imágenes pueden venir con algunos metadatos acerca de la presencia de ciertas condiciones o enfermedades. Esta información normalmente se encuentra en un archivo de texto plano que hacen parte del diagnóstico dado por un profesional de la salud.
- En algunos casos dichos archivos de diagnóstico pueden ser también imágenes o PDFs, lo que implica que deben ser procesados correctamente para poder ser usados como entrada o punto de dato.
- Algunas de los etiquetas más importante incluyen:
 - Normal vs Anormal
 - Benigno vs Maligno
 - Enfermedades concretas (i.e. Neumonía, infarto, fractura, tumor, etc)
 - Condiciones inflamatorias (i.e arthritis, infecciones, etc)
 - Anomalías congénitas

Contexto

- Etiquetado y caracterización

- Para un correcto uso de las imágenes, los usuarios deben tener la capacidad de entender todas las posibles características y metadatos que pueden tener una implicación en el entrenamiento en los modelos de IA. A continuación se presentan las más importantes:

- Demografía:

- **Grupo de edad:** Neonatal, pediatrico, adulto, geriatrico.
- **Sexo:** Masculino, femenino, intersexual.
- **Etnicidad:** Latino, caucásico, afro, asiatico, etc.

- Atributos de la imagen:

- Resolución
- Contraste
- 2D vs 3D
- Fase del escáner (i.e pre-tratamiento, post-tratamiento, seguimiento)



Contexto



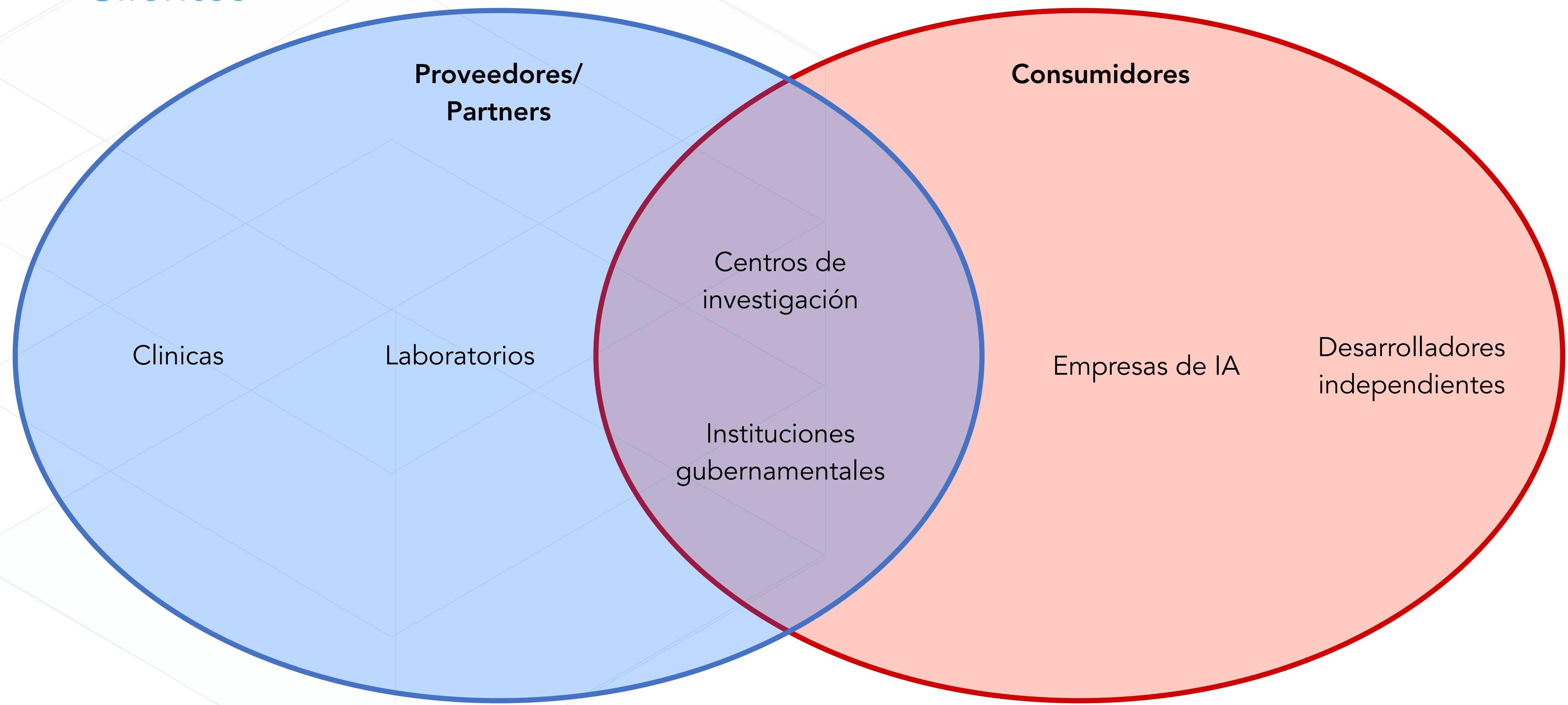
- Metadatos

- Gran parte del entrenamiento de un modelo de salud de IA, no es solamente entender un resultado si no también el contexto del mismo. En el caso de imágenes médicas, éstas en muchos casos vienen con cierta información adicional:
 - **Entorno clínico:** Ambulatorio, interno, UCI, etc
 - **Síntomas o notas clínicas:** Fiebre, estornudo, dolor, etc
 - **Historial del paciente:** Fumador, diabético, condiciones previas, etc.
 - **Contexto procesal:** Pre-operatorio, post-operatorio, exámen de rutina, etc.

Servicios

- Los servicios de STA se pueden dividir en dos grandes categorías:
 - **Data Partnership:** Es la posibilidad de que los proveedores de salud puedan proveer datos anonimizados a STA y obtener un pago por ello. En este caso lo que se propone es un pago acuerdo al nivel de uso de los datos.
 - **Desarrolladores de IA:** STA cuenta con 3 diferentes productos diseñados para servir como datos de entrenamiento de modelos de IA.
 - **STA Standard:** Es una herramienta que sigue un modelo de suscripción para el auto-servicio de obtención de datos.
 - **STA Pro:** Provee un ambiente en la nube completamente administrable por el cliente. A diferencia de la versión standard, este producto permite que los clientes pueden ejecutar incluso consultas SQL en bases de datos totalmente dedicadas al cliente (modelo multi-tenant).
 - **STA Enterprise:** No solo incluye los productos y servicios de standard y pro, sino que además brinda expertos dedicados para ayudar a crear los modelos, entrenamientos e investigación que el cliente necesite.

Clients



Flujo de trabajo

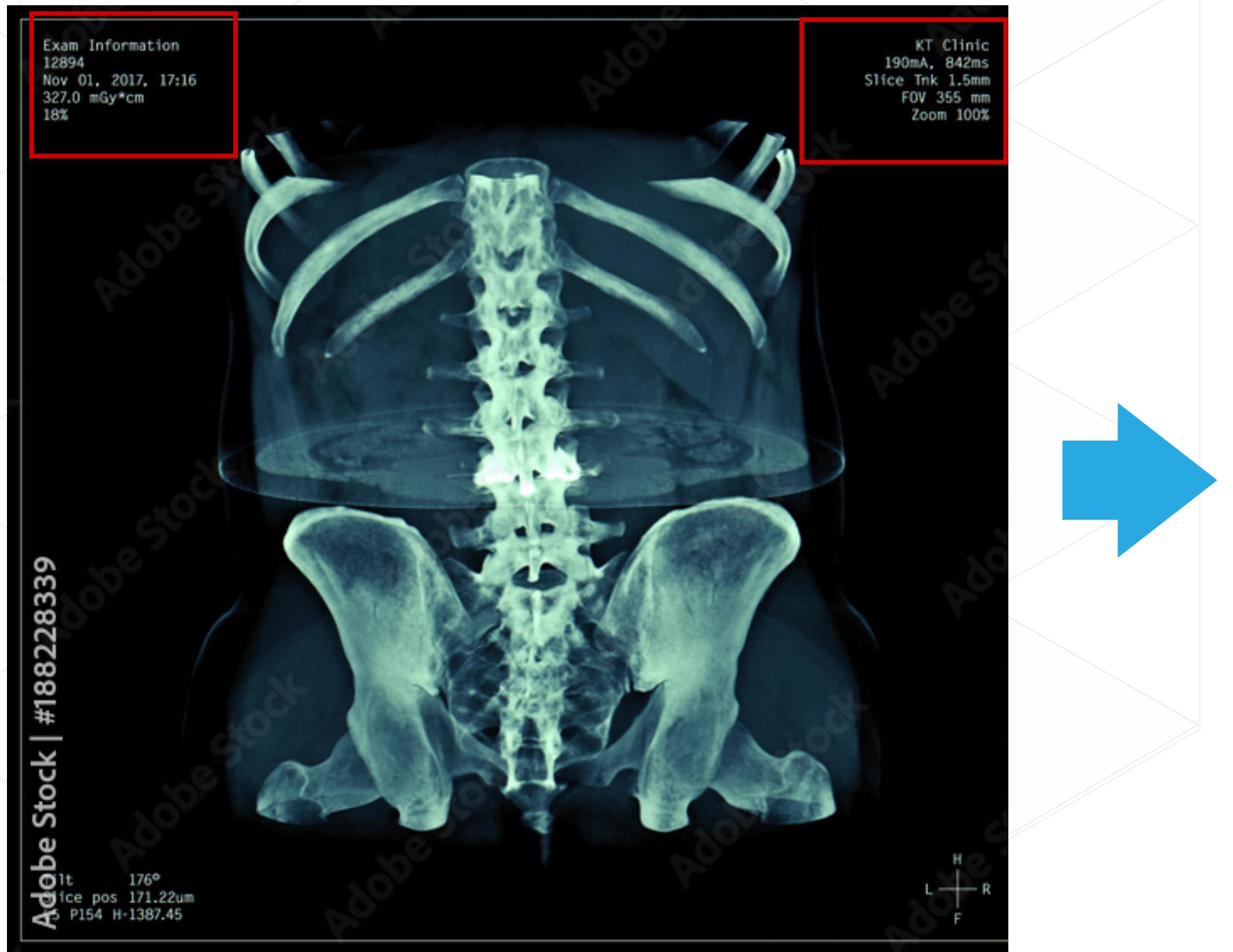
Data Partnership

1. Centros de salud y laboratorios normalmente cuentan con su propia infraestructura la cual en la mayoría de casos es on-premise. Ello implica que uno de los primeros pasos que el equipo de STA debe ejecutar es la importación de dichos datos a la infraestructura en la nube de STA. Sin embargo, debido a las restricciones en términos de privacidad y seguridad (como por ejemplo HIPAA), se debe garantizar que todo dato que llegue a la nube privada de STA se encuentre anonimizado. Este primer paso se puede describir en el siguiente orden:
 - a. El equipo de STA crea un ambiente completamente privado en su nube para almacenar los datos. Este modelo multi-tenant aísla recursos computacionales físicos y virtuales (eso quiere decir que nunca se podría encontrar datos de dos partner en la misma base de datos o siendo procesados por una misma maquina).
 - b. Una vez con el ecosistema privado de nube, se deben preparar los datos a los sistemas en la nube. Dado que se desea que el proceso sea ligero y sin involucrar expertos, en muchos casos los datos pueden venir un Google Drive, Dropbox, o del data center privado del centro de salud.
 - c. Como se desea evitar datos sensibles en las imágenes, el equipo de datos ha preparado unos scripts que permiten quitar información sensible de las imágenes DICOM (estándar de transmisión de imágenes médicas y datos entre hardware de propósito médico). Estos Scripts normalmente corren en la infraestructura del partner o si los datos ya se encuentran en algún servicio en la nube como Drive o Dropbox, son cargados temporalmente en la nube privada, procesados y posteriormente eliminados (los archivos crudos).

***** Nota:** Tenga en cuenta que la combinación modalidad y región anatómica puede ser un universo completo. Es decir, anonimizar radiografías del pie no es lo mismo que radiografías del cuello, y es aún peor si se compara con otra modalidad como ultra sonido. Lo anterior implica, que estos "scripts" cuentan con cierto nivel de entrenamiento para poder hacer anonimización de forma eficaz y segura.

Flujo de trabajo

Data Partnership



Anonimizado



Flujo de trabajo

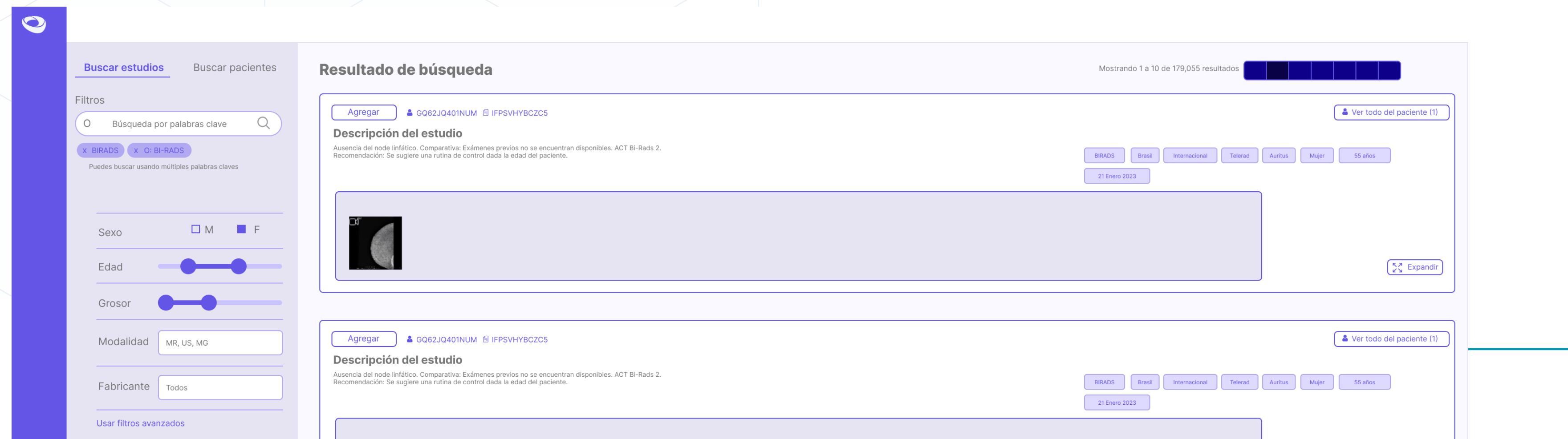
Data Partnership

2. Una vez con los datos en la nube privada del partner, el equipo de investigadores y científicos de datos toman los datos y vuelven a verificar si los datos no contienen ningún tipo de información sensible. Tenga en cuenta que la cantidad de datos capturados pueden ser de cientos de miles a millones de imágenes. En promedio un centro de salud puede proveer 4TB, donde algunos rozan en las cientos de terabytes. Dado lo anterior, acá se ejecuta una mezcla de verificación manual y semi-automatizada. Lo principal de esta etapa no es solamente la verificación sino el mapeo de los datos. Es decir, se espera que al final se puedan agrupar las imágenes y archivos de diagnóstico de una manera en que puedan ser procesados para la generación de etiquetas y metadatos.
3. En la siguiente etapa se ejecutan diferentes Pipelines y modelos acorde a los clusters/agrupaciones en las que se dejaron los datos en la etapa anterior. Ello implica, que se cuenta con modelos de IA entrenados exclusivamente para ciertos tipos de exámenes (modalidad), parte del cuerpo y patológico. Estos pipelines suelen ejecutarse Ad-Hoc, pero algunos partners que cargan información de forma más seguida pueden que necesitan un proceso más automatizado o reactivo.
4. Una vez se ejecutan los pipelines estos generan data frames con la imagen anonimizada y con datos estructurados sobre el diagnóstico. Estos archivos estructurados en formato parquet son almacenados en la nube privada del partner.
5. Por último, se quiere crear conexión entre pacientes, diagnósticos e imágenes. Lo anterior quiere decir, que el proceso de anonimización es capaz de crear un token único para un paciente, haciendo posible que aún con datos ofuscados podamos tener históricos clínicos completos a lo largo de los años sin saber la identidad del paciente, pero identificándolo de forma única en el sistema. En este paso, se genera dicha conexión y se persiste en la base de datos consolidada que usan los diferentes servicios.

Flujo de trabajo

Desarrolladores de IA

1. Un desarrollador, empresa de IA, laboratorio o empresa gubernamental crea una cuenta en STA, donde se le muestra la opción de seleccionar un plan standard (acceso inmediato), pro (se propone hasta un máximo de una semana) y Enterprise (el cual sugiere agendar con un agente de cuenta).
2. En el caso de seleccionar standard, el usuario será inmediatamente redireccionado a una UI donde puede usar diferentes filtros para buscar los datos que necesita. El usuario puede buscar fácilmente millones de imágenes anonimizadas de diferentes modalidades y proveedores. Los resultados de la búsqueda incluyen etiquetas DICOM seleccionadas, informes de radiología e imágenes thumbnails.
3. Una vez el usuario haya agregado los registros deseados puede descargar los datos. Este proceso puede durar un par de horas y se le informa por medio de correo electrónico una vez los datos están listos. Los archivos suelen encontrarse en formato JSON o DICOM, de acuerdo a lo que el usuario seleccione.



The screenshot displays the STA search interface. On the left, there is a sidebar with a purple header containing the STA logo. Below the header, there are tabs for "Buscar estudios" (selected) and "Buscar pacientes". The sidebar also includes a "Filtros" section with a search bar for "Búsqueda por palabras clave" and several filter buttons: "X BIRADS", "X O: BI-RADS", "Sexo" (with checkboxes for M and F), "Edad" (with a slider), "Grosor" (with a slider), "Modalidad" (set to MR, US, MG), "Fabricante" (set to Todos), and a link to "Usar filtros avanzados".

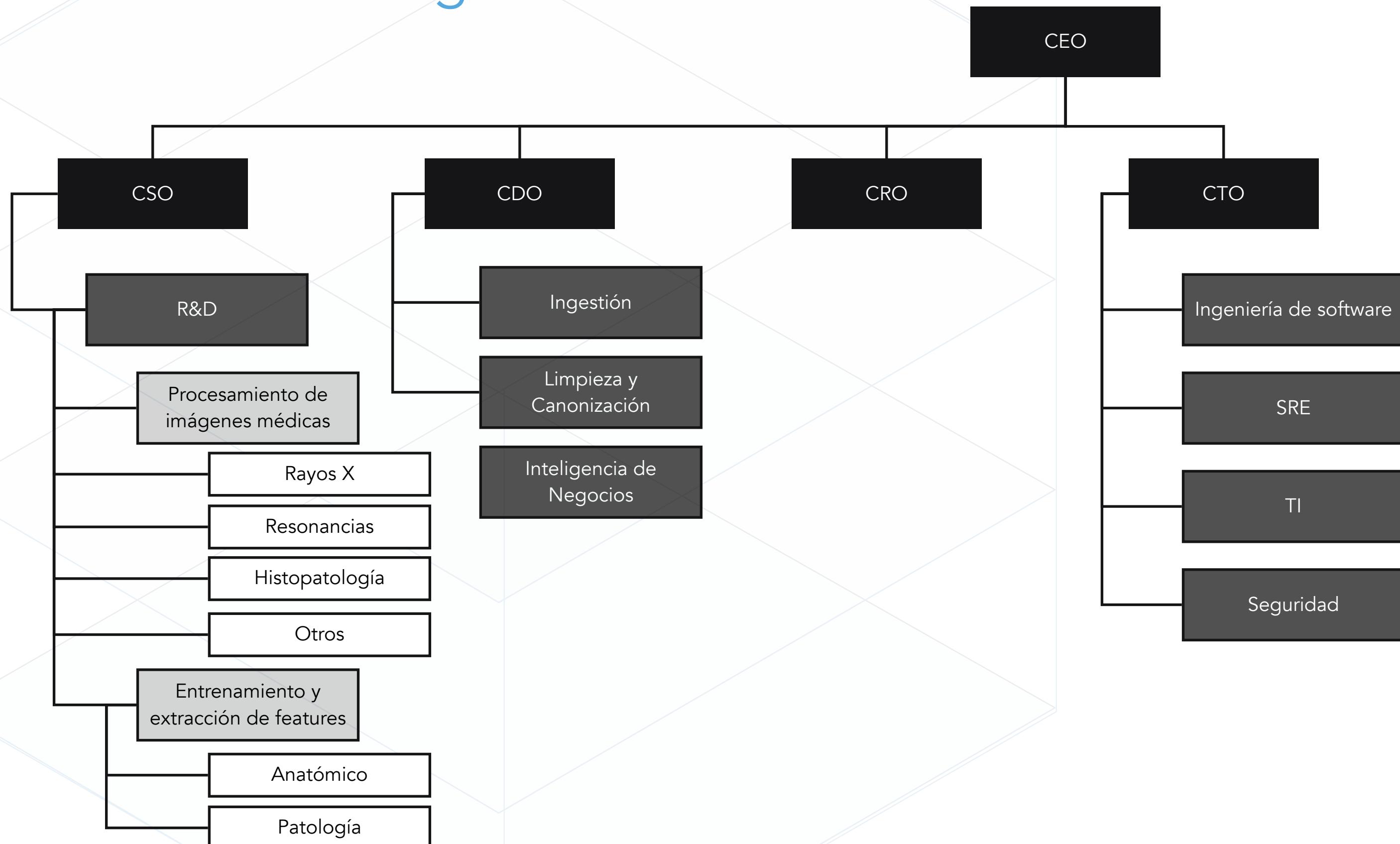
The main area is titled "Resultado de búsqueda" and shows two study results. Each result card includes a "Agregar" button, a patient ID (e.g., GQ62JQ401NUM), a study ID (e.g., IFPSVHYBCZC5), and a "Descripción del estudio" section. The first result notes "Ausencia del nódulo linfático. Comparativa: Exámenes previos no se encuentran disponibles. ACT Bi-Rads 2." and "Recomendación: Se sugiere una rutina de control dada la edad del paciente." It also lists "BIRADS", "Brasil", "Internacional", "Telerad", "Auritus", "Mujer", and "55 años". A thumbnail image of a breast ultrasound scan is shown, along with a "Expandir" button. The second result is identical in structure. At the top right of the main area, it says "Mostrando 1 a 10 de 179,055 resultados".

Flujo de trabajo

Desarrolladores de IA

4. En el caso de seleccionar la opción Pro, el equipo de ventas se contactará con el usuario y comenzarán la creación del ambiente cloud virtual de forma manual. Una vez el ambiente esta creado, se comunica al cliente y se le proveen las credenciales de acceso. En este ambiente virtual el usuario tiene la posibilidad de:
 - a. Ejecutar consultas SQL en una base de datos columnar totalmente dedicada al usuario.
 - b. Entrenamiento sobre los diferentes datasets y formas de utilizarlo
 - c. Acceso a los mismos servicios de la opción standard
5. Si el usuario selecciona la versión Enterprise, similar al paso anterior, un equipo de ventas se contactará con el usuario y, adicional a la creación del ambiente virtual, se comenzará un proceso de entendimiento de los requerimientos específicos e identificación de los recursos que dicho cliente necesita.

Estructura organizacional



- * Esta es una compañía donde la mayoría de sus empleados tienen un background fuerte técnico y académico. Por lo que puede asumir que en el caso de crecer el equipo, está proporción de equipo técnico se va a mantener.

- La compañía como un todo tiene un tamaño de más de **50 empleados**. Con la nueva inversión se desea **triplicar este tamaño**.

- **Convención:**

- **CSO:** Chief Scientific Officer
- **CDO:** Chief Data Officer
- **CRO:** Chief Revenue Officer
- **CTO:** Chief Technology Officer
- **CEO:** Chief Executive Officer

Estadísticas

Item	Valor
Número de centros de salud y laboratorios suscritos como data partners	328
Tamaño aproximado de archivos crudos por data partner	4TB
Cantidad promedio de número de archivos por data partner	325 000
Cantidad de clientes Standard	1347
Cantidad de clientes Pro	145
Cantidad de clientes Enterprise	23

- * Tenga en cuenta que estos son valores actuales, pero para la estrategia global se espera que estos números cambien drásticamente.

Estrategia y visión

Expansión a Estados Unidos

- Como se mencionó anteriormente, la compañía desea expandir operaciones en Latinoamérica y Estados Unidos (US), siendo éste el mercado con mayor potencial de crecimiento, tanto en volumen de datos como de clientes.
- La expansión a US involucra grandes retos, siendo la seguridad y administración de los datos los más relevantes. Desde el año 1996, la ley federal de Estados Unidos, HIPAA, establece normas para la protección y confidencialidad de los datos médicos. Lo anterior agrega un peso sobre los ingenieros en términos de seguridad y calidad de datos. Algunos de los requerimientos más importantes a tener en cuenta son:
 - Los datos no pueden salir del territorio estadounidense, ello implica que todo tipo de procesamiento y almacenamiento debe hacerse en USA. Tenga en cuenta, que casi todo país en el mundo tiene reglas similares.
 - Se debe garantizar la privacidad de los datos, por lo que absolutamente nadie debería estar en la capacidad de consumir de forma no anonimizada los datos de pacientes. Ello implica una alta seguridad en los procesos internos de ingeniería e investigación.
- La junta directiva de STA, desea que muchos de los procesos Ad-hoc o incluso manuales puedan ser completamente automatizados, desde la anonimización, hasta la ingestión, canonización y presentación de los datos. Actualmente, los sistemas edge están desarrollados a la medida de los partners de datos, lo que hace excesivamente lento y costoso el proceso de ingestión y anonimización. Tenga en cuenta que el Mercado Totalmente Abordable (TAM) en USA es de aproximadamente 600 mil centros de salud, lo que implicaría un crecimiento explosivo (se pronostica poder cubrir un 5% de dicho mercado).
- Como se mencionó anteriormente, cada nuevo país involucra retos en el manejo de zonas horarias, latencia de servicios, y diferencias en las leyes acerca del manejo de los datos. Por tal motivo, el sistema debe estar diseñado de una manera en que permita a equipos de tecnología e investigación trabajar de forma independiente evitando cuellos de botella en desarrollo y despliegue de servicios.

Objetivo

- Su equipo fue contratado para diseñar e implementar la arquitectura para soportar la expansión global a múltiples países con alta escalabilidad y seguridad.
- Para ello, se espera que en los próximos 2 meses usted y su equipo presenten un **documento de arquitectura de solución**.
- El equipo de ingeniería desea también validar sus decisiones de diseño. Por ende, se debe construir POC/experimentos para soportar sus decisiones y poderlas presentar al equipo de ingeniería.
- Su documento debe considerar los diferentes puntos de vista y la estructura de equipo que necesita para lograr el cometido.

