Los aspectos principales en el diseño de un entorno de IA listo para la producción

Red Hat

Índice

Los datos son un recurso fundamental para las empresas

El estado del mercado de la inteligencia artificial empresarial

La inteligencia artificial generativa dejó de ser un experimento y se convirtió en una herramienta cotidiana para muchas empresas.

Los equipos la utilizan para resumir contenido, crear código y contenido e interactuar con los datos de manera más natural. Los directivos de las empresas esperan que la inteligencia artificial generativa los ayude a mejorar los resultados para los clientes, el personal y en todas las operaciones, no solo para responder preguntas específicas o crear memes divertidos.

La inteligencia artificial generativa se basa en los datos y las aplicaciones que las empresas ya poseen y les permite:

- convertir grandes volúmenes de contenido no estructurado en conocimiento que se pueda buscar y volver a utilizar;
- ayudar a los desarrolladores, los analistas y los escritores a crear y perfeccionar el código, los informes y el contenido con mayor rapidez;
- personalizar las experiencias digitales de los clientes y los empleados en todos los canales;
- automatizar las decisiones rutinarias y los flujos de trabajo que siguen políticas claras;
- mejorar la productividad de los equipos de desarrollo, operaciones y negocios.

Las investigaciones recientes del sector muestran que este cambio ya está en marcha. Según IDC, más de la mitad de las empresas encuestadas ya ejecutan varias aplicaciones o servicios mejorados con inteligencia artificial generativa en la producción, y se espera que el gasto interanual en inteligencia artificial aumente aproximadamente un tercio entre 2025 y 2029, hasta alcanzar alrededor de USD 1,3 billones en 2029[1]. Para la mayoría de las empresas, la inteligencia artificial generativa está volviéndose parte de los productos y servicios principales.

[1] Whitepaper de IDC. "Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and $1.3 Trillion in 2029, According to IDC". 26 de agosto de 2025.

Al mismo tiempo, las empresas esperan el siguiente paso: la inteligencia artificial con agentes (agentic AI). En lugar de concebir la inteligencia artificial generativa como un solo chatbot o asistente, se utilizan agentes de inteligencia artificial que pueden llamar a herramientas, interactuar con aplicaciones y coordinar tareas de varios pasos. En la práctica, este enfoque puede cambiar la forma en que diseñas y operas el software, desde el autoservicio del cliente y las operaciones de TI hasta los flujos de trabajo empresariales complejos.

IDC informa que más de la mitad de las empresas ya ejecutan pruebas de concepto o casos prácticos iniciales para la inteligencia artificial con agentes y que casi un tercio de las aplicaciones que incorporan inteligencia artificial dependerán de ella para fines de 2026[2]. Actualmente, las empresas lo consideran un plan estratégico para avanzar.

Para aprovechar estos beneficios, necesitas flexibilidad en el modo y el lugar en los que ejecutas la inteligencia artificial.

En la actualidad, muchas empresas planifican adoptar una infraestructura de inteligencia artificial híbrida que combina las nubes públicas con los entornos locales exclusivos. IDC señala que la combinación híbrida de nube pública e infraestructura local se ha convertido en la estrategia de infraestructura digital más común, y que la mayoría de los responsables de la toma de decisiones creen que sus cargas de trabajo de inteligencia artificial requieren una implementación híbrida[3].

Con una plataforma híbrida y abierta, las empresas pueden:

- mantener bajo control los datos y los modelos confidenciales;
- cumplir con los requisitos de privacidad y soberanía de los datos;
- elegir entre varias opciones de hardware;
- optar entre una amplia variedad de modelos open source;
- seguir aprovechando la capacidad de ajuste de la nube cuando la necesitan.

En este ebook, analizaremos los pasos fundamentales para diseñar una plataforma de inteligencia artificial lista para la producción, los aspectos clave a los que se enfrentarán las empresas durante el proceso y la manera en que Red Hat® AI Enterprise ofrece una solución unificada para diseñarla.

[2] Whitepaper de IDC. "Agentic AI Impact on Digital Infrastructure Strategies". Documento n.º US53418526, octubre de 2025. (Se requiere su compra).

[3] Whitepaper de IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025". Documento n.º US53418426, octubre de 2025. (Se requiere su compra).

Casos prácticos de la inteligencia artificial en distintos sectores

Salud
- Mayor eficiencia clínica
- Aumento de la precisión y la velocidad de diagnóstico
- Mejora de los resultados clínicos de los pacientes

Telecomunicaciones
- Mayor comprensión del comportamiento de los clientes
- Mejora de la experiencia de los clientes
- Optimización del rendimiento de la red 5G

Seguros
- Procesamiento automatizado de las reclamaciones
- Servicios de seguros basados en el uso
- Asistencia en el cálculo de riesgos.

Servicios financieros
- Servicios personalizados de atención al cliente
- Mejora del análisis de riesgos
- Detección de operaciones fraudulentas y lavado de dinero

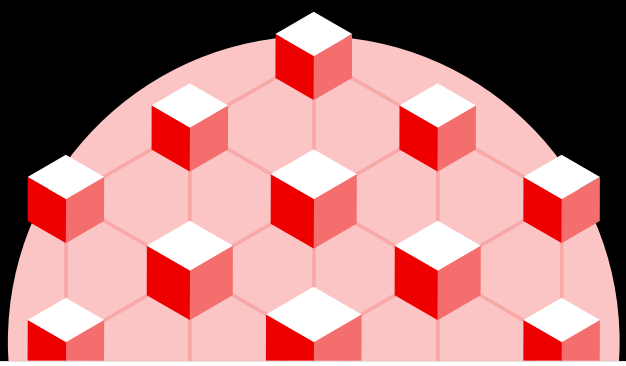
Sector automotor
- Capacidad para ofrecer tecnologías de conducción autónoma
- Predicción de las necesidades de mantenimiento
- Mejora de las cadenas de suministro

Energía
- Previsión de los requisitos de mantenimiento.
- Optimización de la seguridad y las operaciones de campo
- Simulación y predicción de yacimientos más rápidas

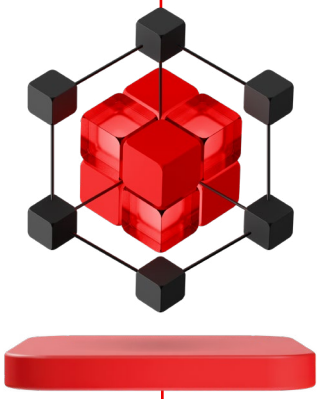
Los pilares de la inteligencia artificial empresarial

En este ebook, analizamos el funcionamiento conjunto de los distintos tipos de inteligencia artificial en una arquitectura empresarial.

- **La inteligencia artificial generativa** utiliza modelos de lenguaje de gran tamaño (LLM) para generar texto, código y otro tipo de contenido a partir de datos y peticiones, de modo que los equipos puedan trabajar más rápido y experimentar con mayor facilidad.
- **La inteligencia artificial predictiva** utiliza datos históricos y en tiempo real para estimar los resultados futuros, como la demanda, el riesgo o el estado de los equipos, para que las empresas puedan actuar antes y con más confianza.
- **La inteligencia artificial con agentes** utiliza agentes de inteligencia artificial que no solo pueden responder a una pregunta, sino que pueden activar herramientas, conectarse a aplicaciones y coordinar flujos de trabajo de varios pasos para alcanzar un objetivo.
- **La inferencia de inteligencia artificial** es la fase del tiempo de ejecución de la producción en la que los modelos aplican lo que aprendieron a los datos nuevos y reales para generar predicciones, recomendaciones o acciones. Se puede ejecutar en un entorno híbrido: en las instalaciones, en la nube o en el extremo de la red.



Diseña una plataforma de IA lista para la producción



El diseño de aplicaciones y agentes de inteligencia artificial generativa es un proceso constante que va más allá de la sola creación de modelos de inteligencia artificial. Estos son los pasos principales del ciclo de vida de la inteligencia artificial:

1. Define tu caso práctico, establece objetivos empresariales para tu iniciativa de inteligencia artificial y obtén el apoyo de las partes interesadas y los directivos.

2. Elige el lugar en el que deseas que se ejecuten las plataformas de experimentación e implementación de modelos: en las instalaciones o en la nube.

3. Elige el modelo de inteligencia artificial que mejor se adapte a tus necesidades. Opta por modelos open source para no tener que depender de un solo proveedor.

4. Personaliza o adapta los modelos que elijas a tus datos propietarios mediante la generación aumentada por recuperación (RAG).

5. Implementa tu modelo en un servidor de inferencia.

6. Diseña aplicaciones o cargas de trabajo con inteligencia artificial generativa.

7. Una vez que cuentes con un entorno de trabajo, amplía y automatiza el flujo de trabajo con la inteligencia artificial con agentes.

8. Supervisa y gestiona los modelos en función de la seguridad y según sea necesario.

Podrás ejecutar este proceso de manera más efectiva con una arquitectura de inteligencia artificial abierta y adaptable, la cual requiere varias tecnologías y funciones clave:

- **Acceso a modelos fronterizos de peso abierto.** Proporcionan a las empresas un punto de partida.
- **Herramientas de GenAIOps y DevOps.** Permiten que los ingenieros de inteligencia artificial, los analistas de datos, los ingenieros de machine learning (aprendizaje automático) y los desarrolladores de aplicaciones creen, implementen y gestionen modelos, agentes y aplicaciones de inteligencia artificial.
- **Acceso a herramientas de perfeccionamiento de modelos y funciones de RAG.** Con ellas, las empresas pueden personalizar los modelos con datos empresariales privados y adaptarlos a los casos prácticos específicos de cada área.
- **Tiempos de ejecución de inferencia.** Mejoran notablemente el rendimiento, la productividad y la latencia.
- **Elementos básicos para los agentes de inteligencia artificial.** Con ellos, las empresas pueden gestionar, controlar y proteger su implementación en la producción.
- **Aceleradores de red, almacenamiento e informática.** Permiten agilizar la preparación de los datos, la personalización de los modelos y las tareas de inferencia.
- **Endpoints de infraestructura.** Brindan recursos para todas las etapas de las operaciones de inteligencia artificial, tanto en los entornos locales, virtuales y del extremo como en los de nube privada, pública e híbrida.

En este ebook, analizamos los aspectos clave para diseñar una arquitectura de inteligencia artificial efectiva.

> La inferencia es el tiempo de ejecución de producción de la inteligencia artificial. Un modelo no es útil hasta que tiene una API y ofrece contenido. Ese contenido se distribuye a través de la inferencia.
>
> **Chris Wright**
> Director de tecnología de Red Hat[4]

[4] Miller, Ron. "Red Hat's CTO sees AI as next step for company's open approach". Fastforward, 11 de noviembre de 2025.

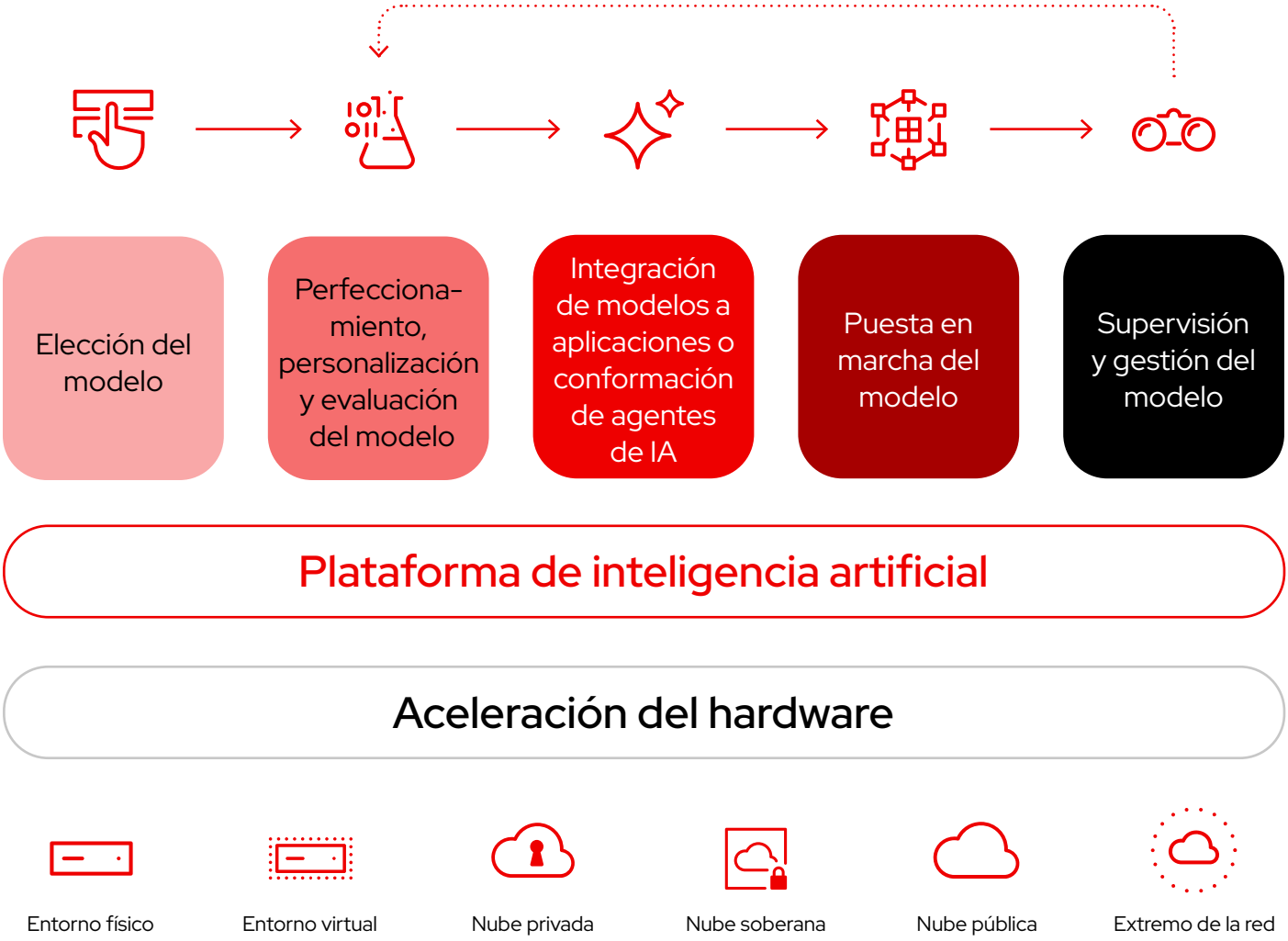


Figura 1. *Los elementos de una arquitectura de inteligencia artificial.*

Desafíos de la implementación de la inteligencia artificial

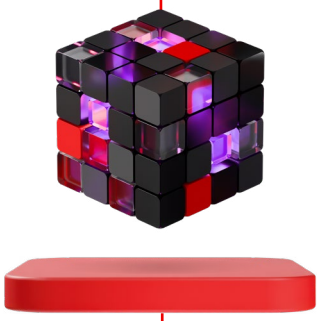
Las empresas se enfrentan a la presión de elegir, diseñar y ofrecer soluciones de inteligencia artificial que ofrezcan una ventaja competitiva. A la hora de poner en práctica y ampliar las implementaciones de inteligencia artificial se deben superar varios obstáculos:

- **Costo del modelo.** Puede resultar costoso ejecutar modelos de gran tamaño e inferencias a gran escala. Las empresas deben optimizar los modelos y la inferencia para contener los costos informáticos y, al mismo tiempo, ofrecer aplicaciones precisas y con capacidad de respuesta.
- **Complejidad de la alineación.** El entrenamiento y el perfeccionamiento de los modelos, así como la creación de canales de RAG, son tareas complejas que requieren un uso intensivo de unidades de procesamiento de gráficos (GPU). Para pasar de la etapa de experimentación a la de producción con mayor rapidez, las empresas pueden simplificar la personalización de los datos empresariales e involucrar a los especialistas en la materia y los desarrolladores.
- **Control y uniformidad.** Los servicios de inteligencia artificial predefinidos limitan el control sobre el hardware, los datos y el control. Elige un enfoque híbrido para poder seleccionar los modelos y la infraestructura sin perder la propiedad de los datos, el ciclo de vida y la capacidad de ajuste de las implementaciones.

Para abordar estos desafíos, se necesita una plataforma de inteligencia artificial híbrida y abierta que proporcione herramientas uniformes para la optimización, la personalización y el control de los modelos en todos los entornos.

Aspectos importantes de la plataforma de inteligencia artificial

Los contenedores y su organización



Contenedores

Los contenedores son unidades básicas de software que empaquetan las aplicaciones con todas sus dependencias. Los contenedores simplifican los procesos de diseño de las aplicaciones y permiten que se implementen en los distintos entornos sin necesidad de realizar cambios.

Importancia para la inteligencia artificial

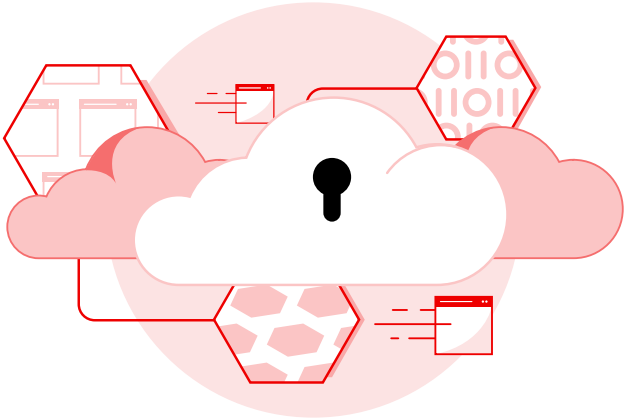
Los ingenieros de inteligencia artificial y los desarrolladores de aplicaciones necesitan acceder a sus herramientas y recursos preferidos para ser más productivos. A la vez, los equipos de operaciones de TI deben garantizar que los recursos estén actualizados, cumplan con los estándares del sector y se utilicen de forma segura.

Los contenedores suelen ser la mejor opción para implementar LLM y aplicaciones con inteligencia artificial generativa, ya que empaquetan los servidores del modelo, las dependencias y la configuración en una unidad repetible que facilita la gestión de la implementación de la producción, el ajuste y las actualizaciones.

También te permiten implementar una amplia selección de herramientas de inteligencia artificial en entornos híbridos de manera uniforme. Para favorecer la transparencia, los equipos pueden modificar y compartir imágenes de contenedores de forma constante a través de las funciones de control de versiones, las cuales permiten hacer un seguimiento de los cambios. Mientras tanto, el aislamiento de los procesos y el control de los recursos mejoran la protección ante las amenazas.

Sugerencias y prácticas recomendadas

Busca una plataforma de contenedores flexible y de alta disponibilidad que incluya funciones de seguridad integradas y optimice la forma en que implementas, gestionas y trasladas los contenedores en tu entorno. Elige una plataforma open source que pueda integrarse a una gran variedad de tecnologías para adquirir más flexibilidad y acceder a más posibilidades.



Organización en contenedores

La organización en contenedores implica gestionar su creación, implementación y ciclo de vida en todo el entorno.

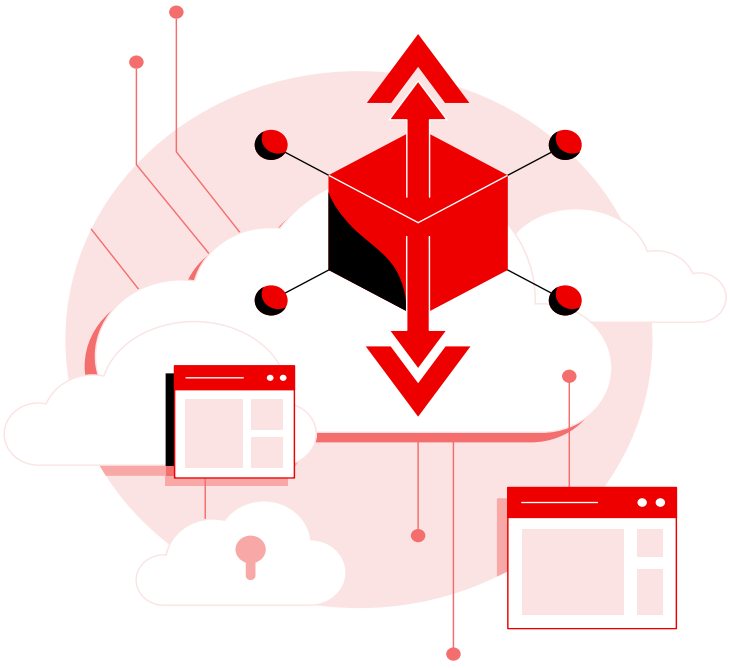
Importancia para la inteligencia artificial

La adopción de los contenedores es solo el primer paso; luego, necesitas una forma de implementarlos, gestionarlos y ajustarlos de manera eficiente. Puedes utilizar un motor de organización en contenedores para administrar el ciclo de vida de manera uniforme. Estas herramientas brindan acceso concentrado a los recursos informáticos, de almacenamiento y de redes en los diferentes entornos en las instalaciones, en el extremo de la red y en la nube. También ofrecen programación unificada de las cargas de trabajo, controles de la arquitectura multiempresa y la aplicación de cuotas.

Sugerencias y prácticas recomendadas

Elige un entorno de organización de contenedores basado en Kubernetes para aprovechar una tecnología de open source líder y no tener que depender de un solo proveedor de nube. Busca una plataforma que ofrezca controles sólidos de arquitectura multiempresa, acceso basado en funciones y gestión de políticas para controlar las cargas de trabajo de inteligencia artificial de manera uniforme. Prioriza las opciones con un amplio ecosistema de operadores e integraciones para que puedas estandarizar la forma en que implementas, ajustas y gestionas los servicios de inteligencia artificial en los entornos híbridos.

Se prevé que, para el año 2027, más del 75 % de las implementaciones de inteligencia artificial utilicen la tecnología de contenedores como entorno informático fundamental, en contraste con el porcentaje obtenido en 2024, que fue inferior al 50 %.[5]



[5] Gartner. "Magic Quadrant for Container Management", 10 de septiembre de 2024.

Aspectos importantes de la plataforma de inteligencia artificial

Gestión de aplicaciones y genAIOps

Gestión del ciclo de vida de las cargas de trabajo de inteligencia artificial

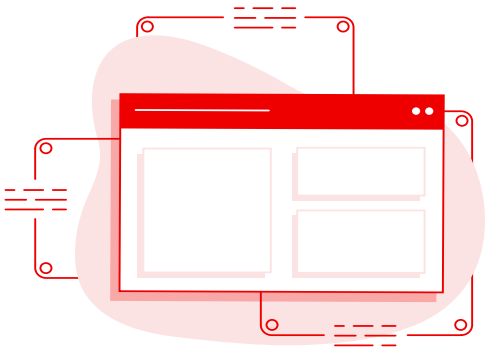
La gestión del ciclo de vida de las cargas de trabajo de inteligencia artificial se centra en la forma en que se implementan, ajustan y administran las herramientas y los servicios que impulsan los casos prácticos de inteligencia artificial.

Importancia para la inteligencia artificial

Los entornos de inteligencia artificial son complejos por naturaleza. Los elementos de gestión del ciclo de vida de las cargas de trabajo de inteligencia artificial, como los notebooks, los entornos de trabajo, los canales y los endpoints de distribución de modelos, deben organizarse en contenedores para facilitar el control y la gestión. Los equipos de operaciones de TI pueden automatizar las tareas comunes del ciclo de vida, como la configuración, el aprovisionamiento y las actualizaciones, para mejorar la precisión y reducir el esfuerzo manual. Los analistas de datos, los ingenieros de inteligencia artificial y los desarrolladores de aplicaciones pueden solicitar entornos de inteligencia artificial aprobados previamente de un catálogo sin tener que iniciar una solicitud de seguimiento con el equipo de TI. La automatización también permite que el personal deje de ocuparse de las tareas repetitivas y dedique su tiempo a las actividades estratégicas de mayor valor.

Sugerencias y prácticas recomendadas

La gestión efectiva del ciclo de vida de las cargas de trabajo de inteligencia artificial comienza con imágenes de notebooks y entornos de trabajo de inteligencia artificial seleccionadas que incluyen bibliotecas de inteligencia artificial y machine learning de uso común para que los equipos partan de una base segura y respaldada, en lugar de hacerlo en entornos creados para una tarea puntual. Las empresas deben proporcionar entornos de notebook que se puedan ejecutar desde un explorador e integrarse a Git para que los equipos puedan colaborar en los experimentos y realizar un seguimiento de los cambios en el código y el modelo a lo largo del tiempo.



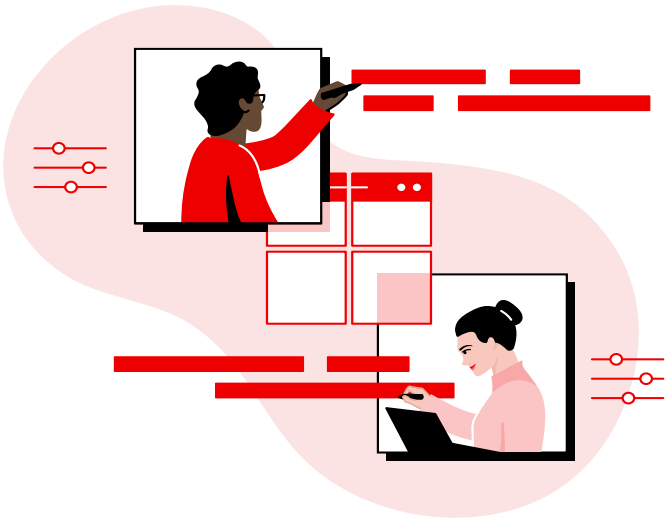
Prácticas de GenAIOps y MLOps

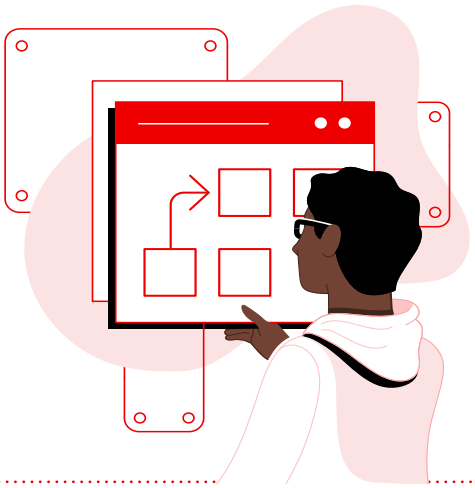
Las prácticas de GenAIOps y MLOps combinan las herramientas, las plataformas y los procesos necesarios para poner en funcionamiento la inteligencia artificial a gran escala.

Importancia para la inteligencia artificial

Las empresas deben desarrollar e implementar los modelos de inteligencia artificial y las aplicaciones que los utilizan de manera rápida y eficiente. Para poder hacerlo correctamente, es esencial que los equipos trabajen juntos.

Al igual que en los enfoques DevOps, en genAIOps y MLOps se fomenta la colaboración entre los ingenieros de inteligencia artificial, los equipos de operaciones de TI y los desarrolladores de aplicaciones para agilizar la creación, el entrenamiento, la implementación y la gestión de los modelos de inteligencia artificial generativa, los agentes de inteligencia artificial y las aplicaciones impulsadas por inteligencia artificial. La automatización, generalmente en forma de canales de integración y distribución continuas (CI/CD), favorece la implementación de cambios rápidos, graduales y repetitivos para acelerar los ciclos de vida del desarrollo de los modelos y las aplicaciones.



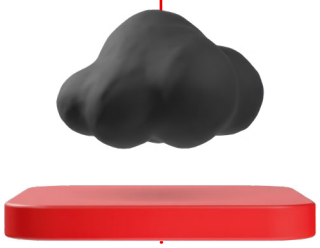


Prácticas de GenAIOps y MLOps

Los enfoques de genAIOps y MLOps no se centran solo en la tecnología, sino que también tienen en cuenta al personal y los procesos. Aplica las prácticas de genAIOps y MLOps en todo el ciclo de vida de la inteligencia artificial. Utiliza la automatización en tus plataformas y herramientas, junto con las tecnologías de open source, como Kubeflow, para crear flujos de trabajo y canales de CI/CD.

Aspectos importantes de la plataforma de inteligencia artificial

Plataforma de nube híbrida



Las plataformas de nube híbrida proporcionan la base para desarrollar, implementar y gestionar la inteligencia artificial en los entornos locales, del extremo de la red y de la nube. Además, te permiten diseñar en función de la inteligencia artificial soberana y la inteligencia artificial privada desde el principio, de modo que puedas decidir qué cargas de trabajo se ejecutan en las nubes públicas y cuáles permanecen en las instalaciones o en los entornos de nube privada que tú controlas.

Importancia para la inteligencia artificial

Los modelos, los agentes, el software y las aplicaciones de inteligencia artificial requieren una infraestructura adaptable para el desarrollo y la implementación. Cuando adoptas una plataforma de nube híbrida uniforme, puedes desarrollar, perfeccionar, probar, implementar y gestionar las aplicaciones y los modelos de inteligencia artificial de la misma forma en todas las partes de tu infraestructura, lo que te brinda más flexibilidad.

Además, esta plataforma es compatible con las estrategias de inteligencia artificial soberana y privada, ya que te permite mantener los modelos y los datos confidenciales en regiones específicas o incluso en entornos desconectados para cumplir con los requisitos de residencia de datos, privacidad y cumplimiento normativo, sin perder el acceso a los servicios de nube pública. Las funciones de autoservicio agilizan la distribución de los recursos y, al mismo tiempo, mantienen el control de la TI.

Por último, con una plataforma uniforme, cuentas con una base para las integraciones tecnológicas de los proveedores externos, las comunidades open source y todas las herramientas personalizadas que utilizas.

Sugerencias y prácticas recomendadas

Elige una plataforma centrada en la seguridad que admita la aceleración del hardware, un amplio ecosistema de herramientas de inteligencia artificial y de desarrollo de aplicaciones, y funciones integradas de gestión de operaciones y genAIOps.

Verifica que ofrezca controles sólidos sobre las políticas de permanencia de los datos en entornos locales, la ubicación de los modelos y el acceso, de modo que puedas ejecutar cargas de trabajo de inteligencia artificial soberana y privada en las instalaciones o en nubes privadas, sin perder el acceso a las nubes públicas.

Las plataformas open source ofrecen más oportunidades de integración y flexibilidad, lo cual impulsa la innovación rápida a través del desarrollo basado en la comunidad, así como las funciones de autoservicio para agilizar la distribución de los recursos y, al mismo tiempo, mantener el control de la TI.

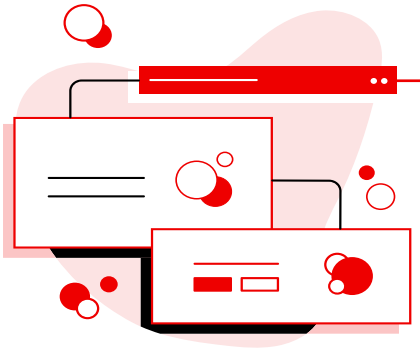
La estrategia de arquitectura de infraestructura digital más común es una combinación híbrida de nube pública e infraestructura local exclusiva[3].

[3] Whitepaper de IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025". Documento n.° US53418426, octubre de 2025. (Se requiere su compra).

Aspectos importantes de la plataforma de inteligencia artificial

Personalización y ajuste de los modelos

Las aplicaciones modernas que utilizan la inteligencia artificial requieren modelos que reflejen los datos, las limitaciones empresariales y los flujos de trabajo específicos de cada empresa. Alinea un modelo abierto o fronterizo con tu información propietaria para pasar de respuestas de conocimiento general a resultados precisos en función de tu área de trabajo.



Importancia para la inteligencia artificial

La inteligencia artificial generativa y la inteligencia artificial con agentes dependen de modelos que comprenden la terminología, los datos y el contexto real.

La alineación ayuda a mantener la precisión y la relevancia al basar los modelos en tus datos privados. Mejora la eficiencia al reducir el costo de la inferencia y evitar el sobredimensionamiento innecesario. Además, refuerza la supervisión y el control, ya que te permite implementar la lógica empresarial, las reglas de seguridad y los requisitos de cumplimiento normativo directamente en el comportamiento del modelo. También respalda la capacidad de ajuste al brindar procesos uniformes para actualizar, volver a entrenar y crear versiones de los modelos a medida que evolucionan tus datos.

La personalización también favorece a las estrategias de inteligencia artificial soberana y privada, lo que permite que las empresas entrenen y distribuyan modelos completamente dentro de entornos controlados para cumplir con los requisitos normativos, de privacidad y de residencia de datos.

Sugerencias y prácticas recomendadas

Adopta flujos de trabajo modulares que comiencen con la RAG, el perfeccionamiento, la ingeniería de peticiones (prompt engineering) y las capas de políticas en función de tus necesidades, en lugar de depender de un solo método. Utiliza modelos abiertos para evitar la dependencia del proveedor y mantener la capacidad de perfeccionar, cuantificar y evaluar los modelos de forma transparente. Solicita la colaboración de especialistas en la materia para garantizar que los modelos reflejen el contexto empresarial real y la precisión de los datos. Optimiza el modelo para la inferencia desde el principio aplicando técnicas como la cuantización, la destilación y los tiempos de ejecución eficientes para controlar los costos y la latencia. Además, debes mantener un control sólido con conjuntos de datos de versiones, ejecuciones de entrenamiento, pesos de modelos e indicadores de evaluación para mantener la capacidad de reproducción y un control sólido.

Aspectos importantes de la plataforma de inteligencia artificial

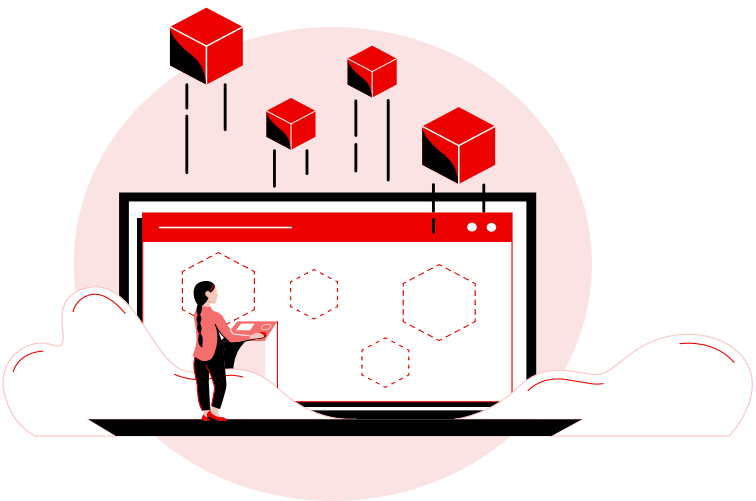
La inferencia de inteligencia artificial a gran escala

La ejecución de la inteligencia artificial en la producción requiere inferencias rápidas, eficientes y confiables. Una vez que los modelos se entrenan o alinean, la inferencia es la etapa en la que procesan datos nuevos, devuelven predicciones, generan contenido o activan acciones dentro de una aplicación o flujo de trabajo.

A medida que las empresas adoptan la inteligencia artificial generativa y la inteligencia artificial con agentes, la inferencia se convierte en un factor fundamental de costos y rendimiento, sobre todo considerando que las aplicaciones pasan de interacciones de una sola consulta a tareas continuas de varios pasos ejecutadas por agentes de inteligencia artificial.

Importancia para la inteligencia artificial

La inferencia determina directamente la experiencia del usuario, el rendimiento de las aplicaciones y los costos operativos. Las cargas de trabajo de inteligencia artificial generativa e inteligencia artificial con agentes suelen requerir respuestas rápidas, solicitudes en paralelo y un rendimiento uniforme en varios entornos, desde los centros de datos hasta la nube pública y los sitios del extremo de la red.



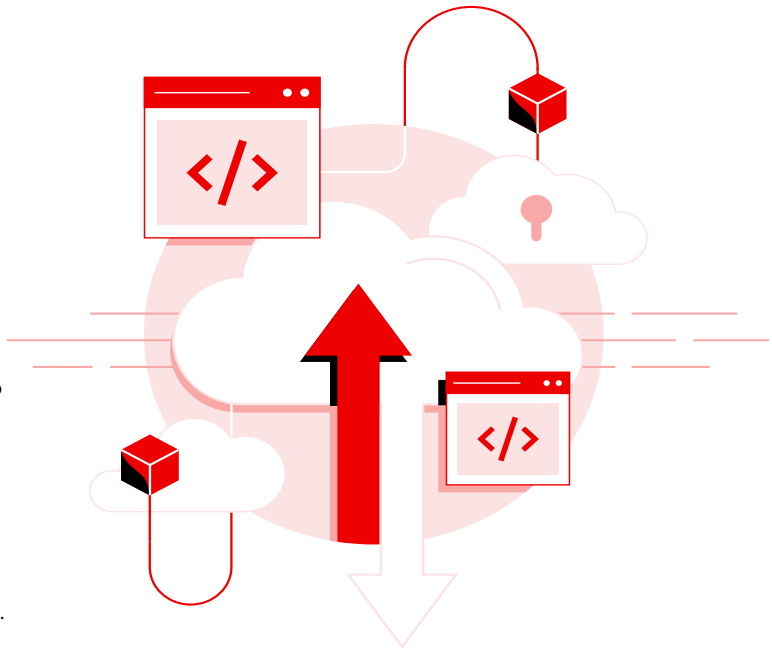
Si los tiempos de ejecución de inferencia son eficientes, se reducen los costos de la GPU y la unidad central de procesamiento (CPU), se mejora la latencia de las tareas interactivas y se satisfacen las necesidades de ajuste de los agentes de inteligencia artificial que realizan llamadas a herramientas, utilizan interfaces de programación de aplicaciones (API) y coordinan flujos de trabajo de varios pasos. Al optimizar la inferencia, también se respaldan las estrategias de inteligencia artificial soberana y privada, ya que las empresas pueden ejecutar inferencias cerca de los datos confidenciales, en las instalaciones o en nubes privadas, mientras mantienen un rendimiento predecible.

Sugerencias y prácticas recomendadas

Selecciona tiempos de ejecución de inferencia optimizados que se adapten a tu tipo de modelo y entorno de implementación, ya sea para los LLM, los modelos multimodales, los modelos predictivos o las cargas de trabajo con agentes. Prioriza los tiempos de ejecución y la infraestructura que admitan la capacidad de ajuste dinámica, tanto horizontal como vertical, para satisfacer las exigencias impredecibles de los LLM interactivos y la inferencia basada en agentes.

Utiliza técnicas de cuantización, destilación y optimización de modelos o asóciate con proveedores que tengan experiencia en esos enfoques para reducir los costos y mejorar la latencia. Combina estas optimizaciones con tecnologías ampliamente adoptadas, como vLLM para la inferencia de LLM de alto rendimiento, y marcos de inferencia distribuida nuevos, como llm-d, que separan el proceso de inferencia para ajustar cada fase de forma independiente.

Implementa la inferencia dentro de los contenedores para empaquetar las dependencias y ajustar la capacidad de manera uniforme en los entornos híbridos. Ubica los endpoints de inferencia donde residen tus datos y aplicaciones para reducir el movimiento y mantener el control, especialmente en los casos de inteligencia artificial soberana y privada. Por último, supervisa el rendimiento de los modelos a lo largo del tiempo y actualiza las versiones a medida que cambian las distribuciones de datos para mantener la precisión y la confiabilidad a gran escala.



90 %

de los responsables de la toma de decisiones cree que la inteligencia artificial será un factor importante en su presupuesto de infraestructura digital y opciones tecnológicas durante 2026[3].

[3] Whitepaper de IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025". Documento n.° US53418426, octubre de 2025. (Se requiere su compra).

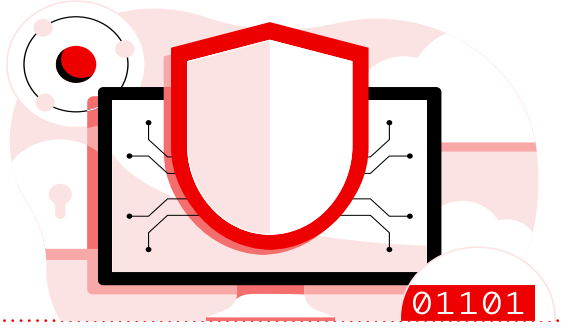
Aspectos importantes de la plataforma de inteligencia artificial

Seguridad de la inteligencia artificial

Los sistemas de inteligencia artificial deben comportarse de manera confiable y predecible y ajustarse a las políticas empresariales. A medida que las empresas pasan de la fase de experimentación a la de producción, la seguridad de la inteligencia artificial se vuelve fundamental, en especial cuando se implementan flujos de trabajo autónomos y de inteligencia artificial generativa y con agentes, que pueden llevar a cabo acciones en lugar de solo brindar sugerencias.

Importancia para la inteligencia artificial

La seguridad se centra en mantener los modelos y los agentes de inteligencia artificial dentro de límites definidos para cumplir con los requisitos empresariales, legales y éticos. Los resultados imprecisos, los desajustes de los modelos, el manejo inseguro de los datos y las acciones involuntarias pueden generar un riesgo operativo real. La inteligencia artificial generativa y los sistemas con agentes también incorporan nuevos desafíos para la seguridad, como las alucinaciones, la ejecución de herramientas sin autorización, el aumento de privilegios y la falta de uniformidad del razonamiento en las tareas de varios pasos. Si tus prácticas de seguridad son sólidas, mantendrás la confianza, protegerás los datos confidenciales y evitarás acciones dañinas o irreversibles. En los sectores regulados, los controles de seguridad son esenciales para el cumplimiento normativo y la preparación para las auditorías en los entornos híbridos y locales.



Sugerencias y prácticas recomendadas

Adopta un enfoque de seguridad en capas que incluya medidas de protección basadas en políticas, filtros de contenido y controles de ejecución de herramientas para los flujos de trabajo con agentes. Valida y prueba los modelos con cierta regularidad para detectar desajustes o una disminución de la precisión. Ejecuta las cargas de trabajo confidenciales en entornos privados o locales para mantener el control sobre la exposición de los datos y el comportamiento del modelo, en consonancia con las estrategias de inteligencia artificial soberana y privada. Utiliza marcos de evaluación de modelos para supervisar el sesgo, la solidez y la confiabilidad. Busca mejorar tus modelos y datos con herramientas que los almacenen en registros que cumplan con los requisitos de los contenedores estándares (OCI) y proporcionen cadenas de suministro seguras. Las tecnologías ampliamente adoptadas, como vLLM para la inferencia de LLM, y las tecnologías distribuidas más recientes, como llm-d, pueden ayudarte a reducir los costos y ajustar la implementación de tu proyecto de inteligencia artificial. Por último, crea versiones y documenta tus modelos, conjuntos de datos y políticas para que puedas hacer un seguimiento de las decisiones y gestionar un control uniforme en todo el ciclo de vida de la inteligencia artificial.

Diseña una base abierta y flexible para la IA

Red Hat AI Enterprise es una plataforma de inteligencia artificial integrada para desarrollar e implementar modelos, agentes y aplicaciones de inteligencia artificial eficientes y rentables en entornos de nube híbrida, y forma parte de la cartera de productos de Red Hat AI.

Unifica los ciclos de vida de las aplicaciones y los modelos de inteligencia artificial para aumentar la eficiencia operativa, agilizar la distribución y reducir los riesgos al proporcionar un entorno de desarrollo listo para usar con funciones de nivel empresarial.

Se trata de una stack completa de inteligencia artificial, impulsada por Red Hat OpenShift, que ha sido probada y cuenta con soporte, y mejora la interoperabilidad y garantiza la continuidad empresarial. Incluye funciones esenciales como el perfeccionamiento de modelos, la inferencia de alto rendimiento y la gestión de flujos de trabajo de inteligencia artificial con agentes. Esto brinda la flexibilidad para utilizar cualquier modelo y hardware e implementarlo donde sea necesario, sin dejar de cumplir con los requisitos de ubicación de los datos. Red Hat AI Enterprise es compatible con todos los entornos híbridos, por lo que los equipos pueden planificar la capacidad, las GPU y los proyectos futuros de inteligencia artificial con confianza.

Red Hat AI Enterprise incluye tecnología del proyecto open source llm-d, que lanzó Red Hat con colaboradores como IBM, NVIDIA, Google y AMD, entre otros. LIm-d mejora la rentabilidad al separar las fases de precarga y decodificación de la inferencia para que cada una pueda ajustarse de manera diferente. Su equilibrador de carga que reconoce las inferencias distribuye las solicitudes en función de las colas de tokens, lo que mejora los tiempos de respuesta y, en algunos casos, dirige las cargas de trabajo de precarga a las CPU.

Agiliza la obtención de resultados.
Implementa una stack de inteligencia artificial para empresas en la infraestructura que elijas, con herramientas preconfiguradas, implementaciones automatizadas y funciones de observabilidad integradas. De esta manera, los desarrolladores y los ingenieros de inteligencia artificial pueden centrarse en diseñar y distribuir aplicaciones basadas en inteligencia artificial desarrolladas en la nube y con agentes.

Aumenta la eficiencia operativa.
Optimiza y automatiza los flujos de trabajo, desde las confirmaciones de cambios en el código hasta el establecimiento de flujos de trabajo de canales de inteligencia artificial y la implementación de modelos. Así, las operaciones de TI pueden ofrecer un rendimiento uniforme y obtener más beneficios de la infraestructura actual con la asignación inteligente de recursos y la gestión integrada del ciclo de vida.

Reduce los riesgos.
Reduce el riesgo de la adopción de la inteligencia artificial empresarial con una stack de inteligencia artificial integrada, probada y con soporte completo que mejora la interoperabilidad entre todos los modelos, los sistemas de hardware y los entornos de nube híbrida. Utiliza esta base para abordar los requisitos normativos y de residencia de datos, de modo que puedas ajustar la inteligencia artificial con confianza.

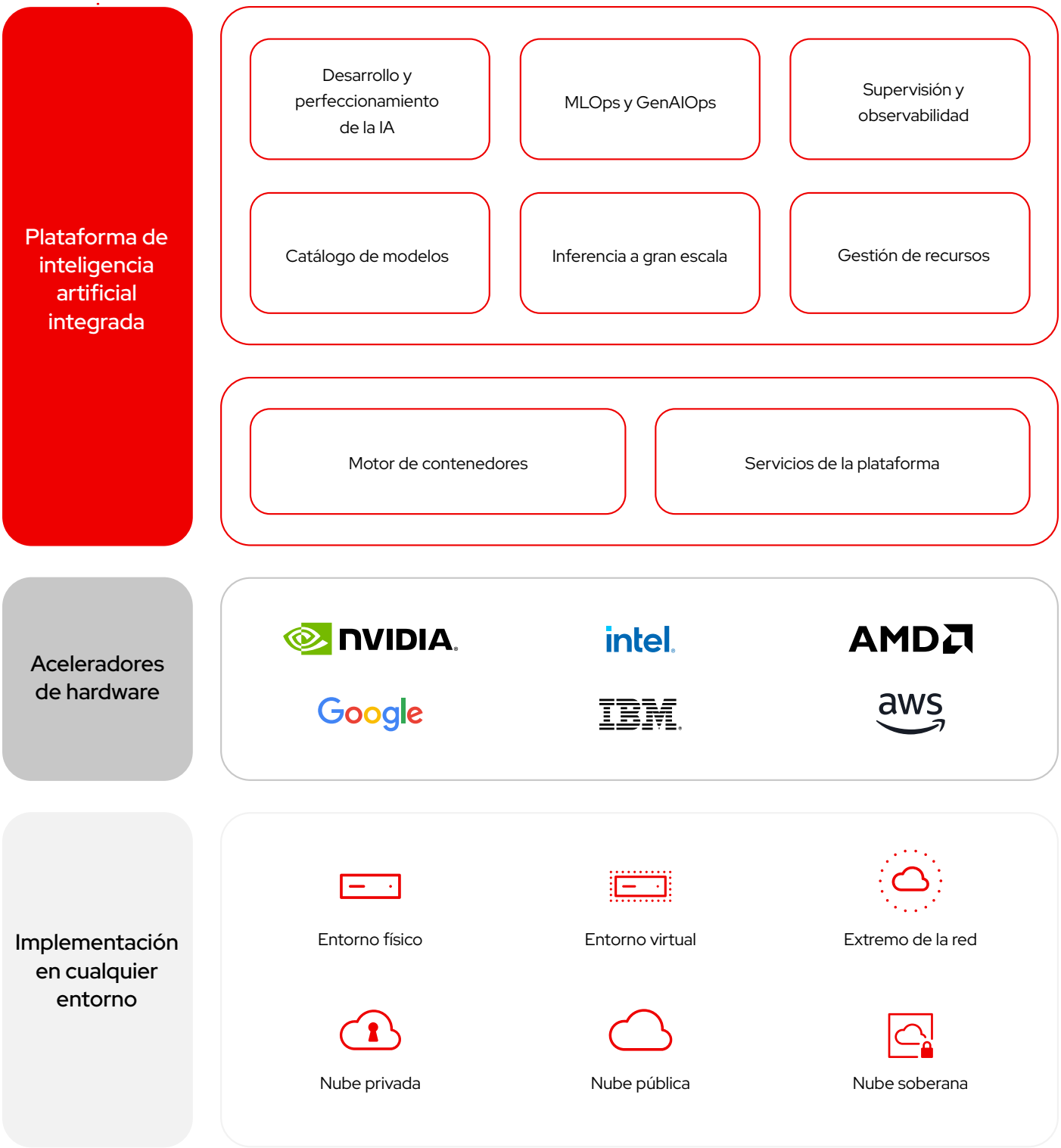
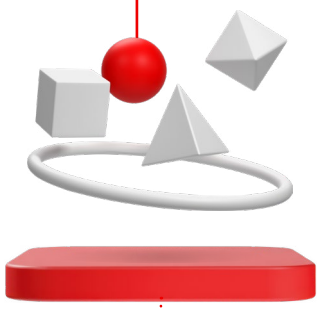


Figura 2. *Los elementos de una plataforma de inteligencia artificial integrada.*

Más opciones y flexibilidad con un ecosistema de partners certificados en IA/ML



El panorama de las herramientas y la tecnología de inteligencia artificial sigue evolucionando rápidamente, por lo que es difícil mantenerse al día con los avances y, al mismo tiempo, garantizar la estabilidad y la confiabilidad dentro de tu entorno de TI.

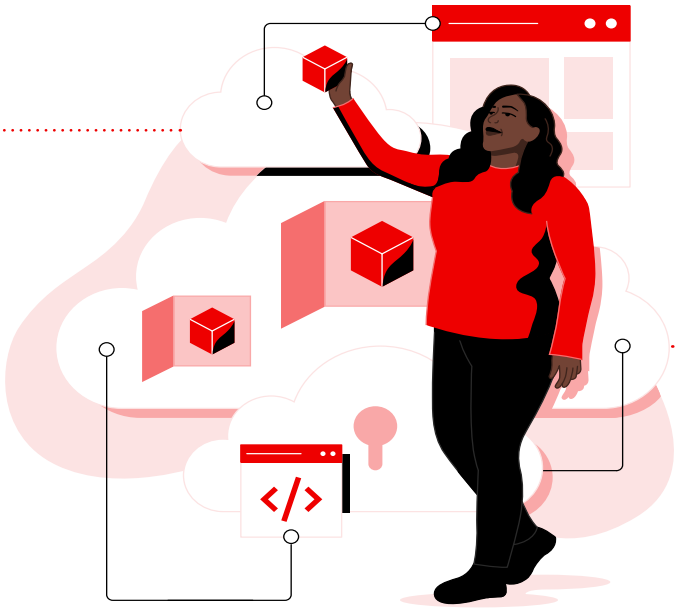
Gracias a las asociaciones con NVIDIA, AMD, Intel y los partners tecnológicos de inteligencia artificial, Red Hat AI Enterprise ofrece una plataforma integral de inteligencia artificial para empresas que se adapta a la nube híbrida y ofrece una implementación más rápida, mayor eficiencia y soporte para la nube híbrida. Los programas de validación y certificación de Red Hat garantizan que el hardware se utilice por completo, mientras que la gestión optimizada de las cargas de trabajo garantiza el uso eficiente de la GPU, lo cual aumenta el rendimiento y los beneficios para los clientes.

La presencia de Red Hat en el ecosistema Hugging Face y en el catálogo de servidores Model Context Protocol (MCP) brinda a los clientes acceso a una biblioteca cada vez más grande de modelos validados y herramientas preintegradas que se ejecutan de manera uniforme con Red Hat AI Enterprise. Al mismo tiempo, gracias a las asociaciones con varios proveedores de aceleradores, las empresas pueden aprovechar las GPU y el hardware especializado en los entornos híbridos. Puedes elegir con confianza los partners, los modelos, las herramientas y las tecnologías que mejor se adapten a tus necesidades, con la seguridad de que funcionarán juntos de manera confiable y contarán con el respaldo de los servicios, el soporte y la capacitación de especialistas para ayudarte a diseñar y ajustar los flujos de trabajo de inteligencia artificial con éxito.

Partners destacados

NVIDIA es líder mundial en tecnología de inteligencia artificial y ofrece las plataformas y las arquitecturas de GPU más avanzadas para la inteligencia artificial y la informática acelerada, por lo que ofrece un rendimiento y una eficiencia energética de primer nivel en el entrenamiento y la implementación de modelos de inteligencia artificial de gran tamaño. NVIDIA y Red Hat tienen una asociación estratégica centrada en ofrecer una infraestructura y un software optimizados y desarrollados en la nube para la inteligencia artificial. En combinación con Red Hat AI Enterprise, los clientes reciben un conjunto completo, optimizado y desarrollado en la nube de sistemas de software de análisis de datos e inteligencia artificial, así como el hardware para ejecutarlos. Los sistemas Red Hat AI Enterprise y NVIDIA DGX ofrecen capacidad de gestión de TI para la infraestructura de inteligencia artificial. El operador NVIDIA GPU Operator automatiza la gestión de todos los elementos de software de NVIDIA necesarios para implementar las GPU.

AMD y Red Hat trabajan juntas para ampliar las opciones de inteligencia artificial y virtualización en los entornos de nube híbrida. La colaboración incorpora los procesadores AMD EPYC y los aceleradores AMD Instinct a Red Hat AI Enterprise, lo cual brinda a las empresas mayor flexibilidad para ejecutar las cargas de trabajo de inferencia, ajuste y entrenamiento de inteligencia artificial en sistemas de hardware rentables y de alto rendimiento. Las iniciativas conjuntas de ingeniería y certificación garantizan que los sistemas con tecnología de AMD se ejecuten de manera uniforme con Red Hat AI Enterprise, lo cual incluye la compatibilidad con las imágenes de contenedores optimizadas, los operadores de Kubernetes y el rendimiento validado para los canales de inteligencia artificial acelerados por GPU.



Intel y Red Hat trabajan en conjunto para ofrecer una infraestructura definida por software y plataformas estándares del sector que mejoran la agilidad y la flexibilidad del centro de datos. La distribución de Intel del kit de herramientas OpenVINO optimiza y convierte los modelos de inteligencia artificial en motores de inferencia de alto rendimiento que pueden expandirse automáticamente a miles de nodos en Red Hat AI Enterprise.

Casos de éxito

Turkish Airlines

Turkish Airlines utiliza Red Hat AI para modernizar las operaciones y ser pionera en la innovación impulsada por la inteligencia artificial en el sector de la aviación. Al adoptar una plataforma de inteligencia artificial abierta y flexible de manera estandarizada, la aerolínea agiliza el desarrollo de modelos, mejora los servicios para los pasajeros y agiliza la toma de decisiones operativas, lo cual demuestra que la inteligencia artificial híbrida puede transformar una de las redes de aerolíneas más grandes del mundo.

Más información

DenizBank

DenizBank utiliza Red Hat AI para acelerar la innovación de la inteligencia artificial en todo su ecosistema de banca digital. Al modernizar su infraestructura de inteligencia artificial con una plataforma abierta y adaptable, el banco acelera la experimentación, mejora la confiabilidad del modelo y ofrece experiencias más inteligentes a los clientes, lo que demuestra cómo la inteligencia artificial híbrida ayuda a las instituciones financieras a avanzar más rápido mientras mantiene una estrategia estricta de seguridad y control.

Más información

AGESIC

AGESIC, el organismo gubernamental digital de Uruguay, utiliza Red Hat AI para estandarizar y ajustar la inteligencia artificial en más de 180 entidades públicas. La plataforma híbrida de inteligencia artificial respalda las prácticas de MLOps, refuerza la seguridad y ayuda a los equipos a diseñar, implementar y controlar las aplicaciones de inteligencia artificial que mejoran los servicios para los ciudadanos.

Más información

¿Todo listo para aprovechar al máximo los datos?

La inteligencia artificial transforma casi todos los aspectos de las empresas. De la mano de Red Hat, puedes diseñar un entorno de inteligencia artificial listo para la producción que agilice el desarrollo y la distribución de las aplicaciones inteligentes para acompañar tus objetivos empresariales.

Obtén más información sobre la manera en que Red Hat AI Enterprise puede ayudarte a diseñar una plataforma unificada para la inteligencia artificial.