IA en IBM Power

La plataforma creada para la IA empresarial

Highlights

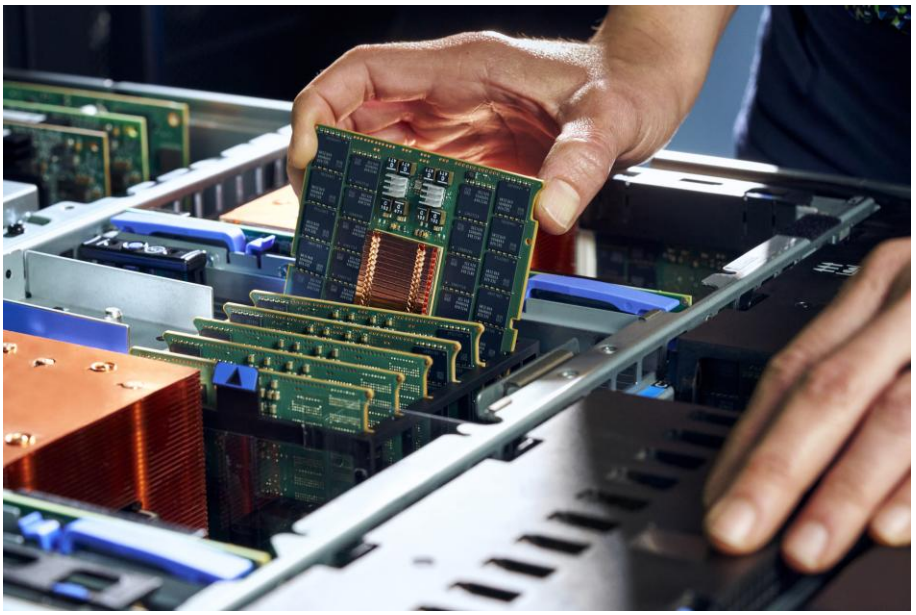
Acelere la IA de manera eficiente: ejecute modelos de IA con alto rendimiento, simplifique las arquitecturas de las soluciones y consiga economías de escala.

Orqueste la IA con flexibilidad: consuma infraestructura de nube híbrida perfectamente, benefíciese del consumo elástico de recursos y combine software de IA empresarial y de código abierto.

Proteja la IA y los datos: minimice la exposición y los riesgos mediante la convergencia de la IA con los datos, asegure las cargas de trabajo de la IA en todas las capas y proteja los datos mediante un cifrado acelerado.

El 85 % de los ejecutivos afirman que la IA propiciará la innovación en los modelos de negocio y el 89 % sostienen que impulsará la innovación en productos y servicios[1]. Hoy en día, los clientes esperan experiencias fluidas y respuestas oportunas a sus preguntas, y las empresas que no logran satisfacer estas expectativas corren el riesgo de quedarse atrás. Se prevé que la inversión en IA generativa crezca significativamente en los próximos años, pero aún representa una pequeña parte del gasto total en IA, y muchos responsables de la toma de decisiones en las empresas coinciden en que el escalado de la IA redunda en una diferenciación competitiva.

A medida que la IA, y especialmente la IA generativa, pasa de la fase de concepción a la de aplicación, las empresas deben elegir una infraestructura adecuada, que sea confiable y ofrezca flexibilidad híbrida e información de confianza. IBM Power proporciona una plataforma agilizada, flexible y protegida diseñada para las cargas de trabajo de IA empresarial. Además, los clientes de IBM® Power® disponen de datos valiosos en sus sistemas IBM Power, lo que les ayuda a obtener insights confiables de sus datos empresariales y aprovechar las ventajas que ofrece la IA.



El 42 %

Más consultas por lotes por segundo en los servidores Power S1022 que en los servidores basados en procesadores x86 comparados durante picos de carga de 40 usuarios simultáneos al utilizar LLM

51 %

Menor costo total de propiedad en un periodo de 3 años ejecutando inferencias paralelas en Cloud Pak for Data en Power S1022 frente a un servidor comparable basado en un procesador x86

¿Por qué usar IA en IBM Power?

Acelere la eficacia de la IA
El hardware y el software optimizados para IA habilitan a los clientes para que aceleren las cargas de trabajo de la IA de manera eficiente sin requerir que los científicos de datos alteren su código, optimizando así directamente desde un principio.

- **Rendimiento mejorado**: el hardware IBM Power incorpora funciones optimizadas para las cargas de trabajo de IA, incluido un acelerador en chip denominado acelerador matemático de matrices (MMA). Junto con la gran capacidad de memoria de IBM Power y el alto paralelismo, estos diferenciadores ofrecen una aceleración eficiente y rentable para las cargas de trabajo de IA. En el caso de los modelos de lenguaje de gran tamaño (LLM), los clientes pueden procesar hasta un **42 % más de consultas por lotes por segundo**[1] en los servidores IBM® Power S1022 que en los servidores basados en procesador x86 comparados durante picos de carga de 40 usuarios simultáneos4 y **disfrutar de una latencia de inferencia inferior a un segundo**[2].
- **Ejecute la IA en una plataforma sostenible y de alto rendimiento**: los servidores basados en procesadores Power10 mejoran la postura de sostenibilidad al proporcionar un **39 % más de inferencia por vatio** en comparación con los servidores basados en procesadores x86[3].
- **Mejora económica**: los clientes pueden aprovechar las capacidades de inferencia paralela y la mayor utilización de la plataforma IBM Power para obtener un **51 % menos de costo total de propiedad en un periodo de 3 años** ejecutando inferencias paralelas en Cloud Pak for Data en IBM Power S1022 en comparación con un servidor comparable basado en un procesador x86[4].
- **Escale soluciones de IA** con un ecosistema en crecimiento.

Orqueste la flexibilidad de la IA
IBM Power proporciona a los clientes la opción de crear y ejecutar su carga de trabajo de IA donde y como sea necesario al brindar:

- **Una infraestructura híbrida sin fricciones** construida para ser congruente en todas las capas (infraestructura, sistema operativo, virtualización y software), ya sea on premises, en una nube privada/gestionada o en la nube pública.
- **Un modelo de consumo flexible** con licencias de pago por uso para software de infraestructura
- y plataforma, independientemente de donde se ejecute la carga de trabajo.
- **Una combinación de software empresarial y/o de código abierto para IA** que ofrece la posibilidad de elegir los componentes básicos para crear las cargas de trabajo de IA que mejor se adapten a sus necesidades empresariales.

Proteja la IA y los datos
A los clientes empresariales les preocupa la seguridad, el riesgo, las vulnerabilidades y el cumplimiento de la normativa. Todas ellas son áreas de creciente interés.
Los modelos de IA pueden procesar datos sensibles a gran escala y, por tanto, los datos deben protegerse mediante mecanismos adecuados de gobernanza y seguridad de datos.

- Simplifique el cifrado y garantice la seguridad de extremo a extremo con **capacidades de cifrado de memoria transparente en** en Power sin perjudicar el rendimiento a través de características de hardware que ofrezcan una experiencia de usuario fluida.
- Minimice la latencia y consolide la criptografía sin tener que enviar datos a aceleradores fuera del dispositivo **mediante la aceleración de algoritmos criptográficos en chip**, que permite que algoritmos como el Estándar de cifrado avanzado (AES), SHA2 y SHA3 se ejecuten con rapidez en los servidores basados en procesadores Power10 y Power11.
- Proteja su aplicación y sus datos con **aislamiento de máquinas virtuales seguro** con órdenes de vulnerabilidades y exposiciones comunes (CVE) de magnitud inferior que los hipervisores relacionados con servidores basados en procesadores x86.
- Perfiles de cumplimiento de seguridad y actualizaciones en tiempo real: las capacidades de **PowerSC** ayudan a los clientes a gestionar, monitorear, elaborar informes y visualizar de forma centralizada la seguridad y el cumplimiento para facilitar las auditorías de cumplimiento de la normativa, incluido el RGPD.

Elección flexible

del mejor software empresarial y de código abierto combinado con una plataforma híbrida sin fricciones (on-prem y en la nube).

Capacidades de IA en IBM Power

IBM Cloud Pak for Data

Las soluciones de IA empresarial de IBM para Power incluyen IBM® Cloud Pak for Data. IBM Cloud Pak for Data es un conjunto modular de componentes de software integrados para el análisis, organización y gestión de los datos. IBM Cloud Pak for Data en Power contiene una amplia gama de componentes de watsonx, Apache, Db2 y Red Hat que ayudan a acelerar las tareas de análisis de datos dentro de Cloud Pak for Data. A medida que sigamos creciendo, se pondrán a disposición otras funciones y servicios de Cloud Pak for Data en Power.

Soluciones de código abierto

IBM Power ofrece capacidades de IA de código abierto compatibles con las comunidades y las empresas. Las soluciones de IA de código abierto en Power se proporcionan a través de RocketCE y Rocket IA Hub for Power. RocketCE es un paquete de herramientas de IA de código abierto optimizadas para Power que aprovecha la aceleración en chip. Está disponible a través del canal público Anaconda de Rocket Software (https://anaconda.org/rocketce/repo). Rocket IA Hub for Power es un conjunto integrado y disponible sin coste de las mejores herramientas de plataforma de IA de código abierto optimizadas para Power, como Katib, Kubeflow, Kubeflow Pipelines, KServe y RocketCE. Todas las herramientas se entregan como imágenes de contenedor que se utilizan dentro de entornos basados en Kubernetes, como Red Hat OpenShift. Todas las herramientas están integradas a través de Kubeflow y optimizadas para aprovechar las capacidades de hardware de IA únicas de la plataforma Power.

IBM watsonx

Mientras el mercado continúa adoptando modelos fundacionales para los casos de uso de IA generativa, Power está preparado para ofrecer capacidades de IA generativa con watsonx y bien posicionado para brindar capacidades de inferencia de modelos fundacionales. Estas capacidades permitirán a los clientes implementar casos de uso de IA generativa para mejorar la experiencia del cliente, aumentar la productividad y optimizar los procesos empresariales. La IA generativa aprovecha la aceleración en chip de Power para ofrecer una experiencia diferenciada a los clientes de IBM. A medida que el mercado continúa evolucionando y los requisitos informáticos cambian, la misión de Power es proporcionar a los clientes una plataforma que pueda satisfacer sus demandas de una manera rentable, sostenible, resiliente y segura. IBM tiene la intención de poner watsonx.data a disposición de Power para ayudar a los clientes a acceder, transformar y gestionar datos empresariales confidenciales a escala para la IA y los analytics.

Spyre en Power

IBM tiene la intención de incorporar el IBM® Spyre Accelerator a las ofertas de Power11 para proporcionar capacidades adicionales de computación de IA. Los procesadores Power11 y el Spyre Accelerator están diseñados para permitir que la infraestructura de próxima generación escale las exigentes cargas de trabajo de IA para las empresas. El Spyre Accelerator es un acelerador de nivel empresarial diseñado específicamente que ofrece capacidades escalables para modelos de IA complejos y casos de uso de IA generativa. El nuevo acelerador cuenta con 32 núcleos aceleradores individuales integrados, y cada Spyre Accelerator está montado en una tarjeta PCIe.

Red Hat OpenShift

Además de las cargas de trabajo que soportan las iniciativas de IA de los clientes de IBM, los avances en Red Hat OpenShift tendrán un impacto en la arquitectura de la solución de IA. Un ejemplo de ello es la introducción de capacidades OpenShift, como la compatibilidad con computación multiarquitectura (MAC). Las MAC permiten a los clientes disponer de clústeres multiarquitectura OpenShift con nodos de cálculo x86 e IBM Power. Estas capacidades permiten a los clientes de IBM implementar las cargas de trabajo donde tenga sentido, aprovechando los puntos fuertes y las ventajas de las distintas arquitecturas y beneficiándose de una flexibilidad optimizada en el despliegue de las cargas de trabajo.

Conclusión

Ahora, los clientes de IBM Power tienen acceso a una suite de capacidades de IA que aprovechan la aceleración en chip de Power, responden a la necesidad de soluciones tanto empresariales como de código abierto y se dirigen a los principales impulsores del mercado de la IA en la actualidad. Estas capacidades permiten a los clientes de IBM hacer frente a los retos empresariales actuales más destacados, extrayendo insights aplicables en la práctica de sus cada vez más numerosos datos multimodales.

IBM tiene la convicción de que la IA, especialmente la IA generativa, debe adaptarse a las necesidades de la empresa.
Esta convicción se respalda a través de los siguientes principios:

- **Abierta:** proporcione una base fundamentada en las mejores tecnologías de código abierto, permitiendo así que los clientes innoven rápidamente con acceso a una comunidad abierta y múltiples modelos.
- **Fidedigna**: proporcione seguridad y protección de datos mediante estricta gobernanza y ética para satisfacer las crecientes exigencias regulatorias y de cumplimiento.
- **Dirigida**: diseñe para casos de uso específicos de la empresa que aporten un nuevo valor comercial al cliente.
- **Facultativa**: permita que los clientes aporten sus propios modelos y datos, creen sus soluciones de IA y amplíen su escala en toda la empresa para su adopción generalizada.

Estos principios se proporcionan a través de capacidades que abarcan el ciclo de vida completo de la IA:

- watsonx: la plataforma de IA y datos
- Infraestructura para IA (en la nube u on-prem): IBM Cloud, IBM Power, IBM® Z e IBM® Storage

Más información
Para obtener más información sobre IBM Power, póngase en contacto con su representante de IBM o con un IBM Business Partner, o visite www.ibm.com/mx-es/power.

1. Comparación basada en pruebas internas de IBM de inferencia de preguntas y respuestas utilizando el modelo PrimeQA (https://github.com/primeqa, basado en los modelos de Dr. Decr y ColBERT). Resultados válidos a 22 de agosto de 2023 y obtenidos en condiciones de laboratorio. Los resultados individuales pueden variar en función del tamaño de la carga de trabajo, el uso de subsistemas de almacenamiento y otras condiciones. La comparación se basa en el rendimiento total en puntuación (inferencias) por segundo en IBM Power S1022 (1x20 núcleos/512 GB) con SMT 8 frente a sistemas basados en Intel Xeon Platinum 8468V (1x48 núcleos/512 GB). La prueba se ejecutó con entornos Python y Anaconda, incluyendo paquetes de Python 3.10 y PyTorch 2.0. Las bibliotecas de Python utilizadas están optimizadas para plataformas Power e Intel. Configuración: tamaño del lote = 60 con 40 usuarios simultáneos. El torch.set_num_threads(*int*) se ha optimizado en una variedad de niveles de carga.

IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html): 6.26 consultas inferidas por segundo con 40 usuarios simultáneos. Sistema x86 comparado: Supermicro SYS-221H-TNR (https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr): 4.4 consultas inferidas por segundo con 40 usuarios simultáneos.
Modelos ajustados por IBM en un corpus de datos internos de IBM.

2. Basado en pruebas internas de IBM de inferencia de preguntas y respuestas utilizando modelos PrimeQA (basados en los modelos Dr. Decr y ColBERT). Resultados válidos a 31 de agosto de 2023 y obtenidos en condiciones de laboratorio. Los resultados individuales pueden variar en función del tamaño de la carga de trabajo, el uso de subsistemas de almacenamiento y otras condiciones. Basado en resultados para un IBM Power S1022 (2x20 núcleos 2.9-4 GHz/512 GB) usando un chip NUMA alineado de 10 núcleos LPAR. Las pruebas se realizaron con entornos Python y Anaconda, incluyendo paquetes de Python 3.10 y PyTorch 2.0. Las bibliotecas de Python utilizadas son bibliotecas optimizadas para la plataforma Power. Configuración: SMT 2, torch.set_num_threads(*16*); tamaño del lote = 1.

IBM Power S1022 (https://www.redbooks.ibm.com/abstracts/redp5675.html)
Modelos PrimeQA: https://github.com/primeqa
Modelos ajustados por IBM en un corpus de datos internos de IBM.

3. Basado en pruebas internas de IBM de componentes de ciencia de datos (WML, WSL, Analytic Engine) de Cloud Pak for Data 4.8 en OpenShift 4.12. Resultados válidos a 17/11/2023 y obtenidos en condiciones de laboratorio. Los resultados individuales pueden variar en función del tamaño de la carga de trabajo, el uso de subsistemas de almacenamiento y otras condiciones. 2. La carga de trabajo imita un flujo lógico de detección de fraudes en tiempo real. JMeter se utiliza para enviar transacciones de tarjetas de crédito para diferentes combinaciones de identificación de usuario y número de tarjeta. La aplicación de inferencias que se ejecuta como microservicios en el espacio de despliegue Cloud Pak for Data extrae el identificador de usuario y el número de tarjeta de crédito y los utiliza para buscar seis transacciones anteriores de la misma combinación de usuario y tarjeta en la base de datos Db2, que también se ejecuta en el clúster de Cloud Pak for Data. A continuación, los datos recuperados de la base de datos se combinan con la nueva entrada y se pasan al modelo LSTM para determinar si la última transacción es fraudulenta o no. La puntuación (valor entre 0 y 1) se devuelve al cliente de JMeter como un indicador de si esa transacción es susceptible de ser un fraude o no. 3. La medida utilizada para ambos sistemas, Power e Intel, es el resultado de rendimiento (puntuación/segundo) notificado por JMeter al ejecutar 192 subprocesos actuales (un subproceso representa a un usuario) en 96 endpoints de inferencia.

4. Power10 S1022 tiene un total de 40 núcleos físicos y 2 TB de RAM (tipo de máquina 9105-22A). Hay 7 LPAR en este sistema, incluidos tres nodos maestros de dos núcleos y 32 GB de RAM cada uno, tres nodos de trabajo de diez núcleos y 490 GB de RAM cada uno, y un nodo bastión de cuatro núcleos y 128 GB de RAM. Se utilizan unidades NVME locales de 800 GB como unidades de arranque para cada nodo, y una NVMe de 1.6 TB para el almacenamiento del servidor NFS que se ejecuta en el nodo bastión. Hay un adaptador Ethernet de 100G virtualizado a través de SRIOV, y cada LPAR ocupa el 10 % del ancho de banda de la red. Cada LPAR se ejecutó con un rango de frecuencia de CPU de 3.20 GHz a 4.0 GHz. Los tres nodos de trabajo funcionan en modo SMT 4, mientras que los nodos maestro y bastión funcionan en modo SMT 8. 5. El sistema Intel es Xeon Platinum 8468V con 96 núcleos físicos y 2 TB de RAM. El host KVM ocupa dos núcleos y 32 GB de RAM, lo que admite siete invitados KVM en este sistema, incluidos tres nodos maestros de cuatro núcleos y 32 GB de RAM cada uno, tres nodos de trabajo de 24 núcleos y 490 GB de RAM cada uno, y un nodo bastión de cuatro núcleos y 128 GB de RAM. Se utilizan unidades NVME locales de 1.6 GB como unidades de arranque para estos nodos, y una NVMe de 1.6 TB para el almacenamiento NFS en el nodo bastión. Hay un adaptador Ethernet de 100G virtualizado a través de SRIOV. Cada invitado KVM se ejecutó con un rango de frecuencia de CPU de 2.40 GHz a 3.8 GHz. Todos los nodos son invitados RHEL CoreOS KVM que se ejecutan en el servidor con hiperprocesamiento habilitado. Los precios se basan en: Power S1022 (véase la página 4). Precios típicos de Intel x86 estándar de la industria (ejemplo en la página 5) https://www.synnexcorp.com/us/govsolv/pricing/ y precios del software de IBM disponibles en https://www.ibm.com/mx-es/downloads/cas/DLBOWBPK

Producido en los
Estados Unidos de América
Agosto de 2025