Articulos

# Combinación de IA con un enfoque de datos confiables en IBM Power para impulsar resultados de negocio

Por Bargav Balakrishnan | Vicepresidente, Banca y Modernización de la Industria, IBM

October 04, 2023

Los datos alimentan la inteligencia artificial (IA), por eso, la infraestructura en la que se ejecuta la IA es esencial. En un momento crucial para la implementación de la inteliencia artificial, los y las líderes empresariales que exploran las muchas ventajas que promete la IA primero deben preguntarse: "¿Cómo puede mi equipo crear, entrenar e implementar sistemas de IA en mi empresa? ¿Tenemos la infraestructura adecuada que pueda ofrecer soporte para las altas demandas computacionales y de memoria de las cargas de trabajo de IA?"

Rápidamente descubrirán que no existe una solución única para la infraestructura de IA y, en cambio, deben alinear la infraestructura adecuada con la tarea de IA correspondiente. Las consideraciones no solo incluyen tareas de inteligencia artificial, tamaño y escala de modelos, sino también seguridad, privacidad, resiliencia, conformidad normativa y políticas. Las cargas de trabajo de IA formarán cada vez más la columna vertebral de las cargas de trabajo críticas para la misión y, por lo tanto, requerirán una infraestructura resiliente por diseño.

Con el reciente lanzamiento de IBM watsonx, la plataforma de inteligencia artificial y datos de IBM para empresas, ayudamos a los clientes a crear una ventaja competitiva al proporcionar herramientas, como un SDK de SAP planificado para watsonx (disponible en el primer trimestre de 2024) en todo el ciclo de vida de la IA que puede escalar la IA en sus organizaciones.

IBM Power está diseñado para IA y cargas de trabajo avanzadas, lo que permite a las empresas inferir e implementar algoritmos de IA en sus datos y transacciones más confidenciales que residen en los sistemas Power. Por ejemplo, los clientes pueden aprovechar la aceleración de IA que viene con cada núcleo Power10 para procesar hasta un 42 % más de consultas, por lotes, por segundo en IBM Power S1022 que un servidor x86 comparado durante la carga máxima de 40 usuarios simultáneos, cuando cada uno utiliza modelos de IA de lenguaje grande. [1] Además, la aceleración de Power10 permite a los clientes inferir modelos de lenguaje grandes en menos de un segundo en servidores IBM Power S1024 con 2×12 núcleos. [2]

**Funcionalidades impactantes basadas en IA**

Ya sea que los clientes necesiten integrar sus datos en tejidos de datos y plataformas de IA o implementar modelos de IA como la IA generativa cerca de sus datos, IBM Power puede ayudar a las empresas a abordar las preocupaciones sobre el tiempo de comercialización de soluciones basadas en IA con un enfoque adecuado para su propósito.

Para muchos de nuestros clientes, ese enfoque incluye SAP HANA en IBM Power para ofrecer un rendimiento récord, escalado, accesible y mayor tiempo de actividad. [3] Planeamos poner a disposición SAP ABAP SDK para watsonx, que está destinado a facilitar el consumo de servicios watsonx dentro de los entornos SAP ABAP que se ejecutan en las instalaciones, en la nube y sobre el motor ABAP independiente en la Plataforma de Tecnología Empresarial de SAP. El SDK también está diseñado para permitir a los clientes de RISE con SAP crear extensiones en entornos SAP ABAP. Y para nuestros clientes que normalmente ejecutan cargas de trabajo de SAP HANA en IBM Power, pueden inferir cerca de sus datos en los sistemas Power y, a través del SDK, aprovechar watsonx en entornos ABAP.

Como ejemplo de un posible caso de uso, los clientes podrían aprovechar el poder de watsonx para simplificar el análisis de parámetros operativos preventivos derivados de sistemas SAP. Esto podría ayudar a identificar patrones que indiquen posibles fallas de activos y establecer un contexto en torno al estado de los dispositivos. Aprovechando las funcionalidades predictivas, un sistema puede anticipar y prever fallas de manera proactiva, lo que permite a las organizaciones tomar medidas preventivas que ayuden a evitar costosos tiempos de inactividad e interrupciones.

Nuestros clientes pueden lograr resultados empresariales impactantes porque IBM Power10, con aceleración integrada y gran memoria, proporciona una plataforma escalable y segura para integrar IA en los flujos de trabajo de transacciones de los clientes y en las experiencias del cliente final. Hemos optimizado IBM Power para las bibliotecas de IA más comunes disponibles por medio del socio de IA/ML de IBM, Rocket Software, a través de RocketCE, que continuará respaldando aplicaciones de IA para capitalizar la innovación de Power10.

También tenemos la intención de expandir nuestro portafolio con Rocket Software en el cuarto trimestre de 2023 con la adición de Rocket AI Hub para IBM Power, un conjunto integrado de herramientas de plataforma de IA de código abierto, como Kubeflow. Rocket AI Hub para IBM Power estará disponible y también tendrá una opción de complemento de soporte comercial.

Hoy, anunciamos la disponibilidad de IBM Power10 en IBM Power Virtual Server, comenzando en centros de datos selectos en los Estados Unidos y expandiéndose a otras geografías a finales de este año. Esto continuará expandiendo las funcionalidades de computación de la flota y ofrecerá más opciones al implementar cargas de trabajo críticas para el negocio en IBM Cloud.

Los clientes con requisitos de rendimiento exigentes o que tienen software licenciado por núcleo se beneficiarán del rendimiento adicional del procesador Power10 en IBM Cloud. Además, los clientes existentes que buscan Power10 en IBM Cloud para alinearse con sus entornos locales para el desarrollo de aplicaciones, pruebas y/o copia de seguridad y recuperación de desastres, o nuevos clientes que adoptan IBM Power Virtual Server por primera vez, pueden sentirse cómodos sabiendo que pueden elegir la tecnología más actual de IBM junto con los sistemas operativos más recientes**.**

"IBM ha sido un proveedor de confianza desde hace mucho tiempo para nuestros sistemas centrales, y estamos

emocionados de explorar la nueva funcionalidad que IBM está incorporando a la plataforma Power10 que permite enfoques de software modernos específicos para machine learning", dijo Ben Metz, Director digital y tecnológico en Jack Henry, una destacada compañía de servicios de tecnología financiera. "Con inferencia en tiempo real, IBM Power tiene el potencial de acercar los insights y la toma de decisiones a nuestros datos de misión crítica, que es una parte esencial de nuestra estrategia de modernización tecnológica".

Los clientes también pueden beneficiarse del soporte de Multi Architecture Cluster (MAC), permitiéndoles combinar nodos de trabajador de IBM Power y x86 en un único clúster de Red Hat OpenShift. El uso de MAC ayuda a los clientes a alinear la tarea de IA correcta con la infraestructura adecuada, abordando la necesidad de que cada tarea de IA se ejecute en una sola plataforma para que puedan crear aplicaciones donde sea necesario y llevarlas rápidamente a producción. Adicionalmente, se espera que IBM Cloud Pak for Data 4.8 amplíe el soporte para IBM Power en el cuarto trimestre trayendo los componentes más recientes de ciencia de datos (como Watson Machine Learning, Watson Studio, Analytics Engine Powered by Apache Spark, Data Refinery y Decision Optimization) a Power para clientes empresariales.

**Trabajar con nuestros asociados del ecosistema para llevar funcionalidades de IA en IBM Power a los clientes**

Para que la IA tenga un impacto significativo, a menudo se requiere un esfuerzo de equipo. Sobre la base de la relación entre MuleSoft de Salesforce e IBM, MuleSoft e IBM están en conversaciones para admitir Anypoint Flex Gateway en IBM Power en el primer trimestre de 2024. MuleSoft Flex Gateway es un gateway de API basado en Envoy, diseñado para gestionar y asegurar las API que se ejecutan en cualquier lugar. Al respaldar Flex Gateway en IBM Power, los clientes pueden habilitar sus aplicaciones basadas en Power con API para obtener un acceso seguro con aplicaciones dentro de su empresa." Con esto, los clientes pueden modernizar, compartir y conectar datos de misión crítica en aplicaciones Power con aplicaciones empresariales para mejorar los modelos de IA utilizando watsonx.

**Aproveche al máximo sus datos con IA en IBM Power**

Como dije al principio, no existe una solución única para su infraestructura de IA. He descubierto que los clientes se benefician más cuando co-creamos y alineamos la infraestructura correcta con la tarea correcta de IA en cuestión.

El asociado perfecto en esta ruta es IBM Consulting. Están trabajando directamente con clientes y asociados globales para co-crear lo que vendrá en IA para impulsar la transformación empresarial. Estos consultores dedicados aportarán experiencia en el dominio para ayudar a los clientes a aprovechar herramientas para crear modelos de watsonx en entornos SAP y también pueden ayudar a los clientes que están considerando ejecutar cargas de trabajo de SAP HANA en IBM Power o IBM Power Virtual Server para acelerar las implementaciones de aplicaciones.

*Descubra más acerca de lo que IBM Consulting puede hacer por usted*

---

*Las declaraciones relativas a la orientación e intención futuras de IBM están sujetas a cambios o a su retiro sin*

*previo aviso y representan únicamente metas y objetivos.*

*[1] Comparación basada en pruebas internas de IBM de inferencia de preguntas y respuestas utilizando modelos PrimeQA (https://github.com/primeqa, basados en los modelos Dr. Decr y ColBERT). Resultados válidos a partir del 22 de agosto de 2023 y realizados bajo condiciones de laboratorio. Los resultados individuales pueden variar según el 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 sistemas IBM Power S1022 (1×20 núcleos/512 GB) frente a sistemas basados en Intel Xeon Platinum 8468V (1×48 núcleos/512 GB). La prueba se ejecutó con entornos Python y Anaconda, incluidos paquetes de Python 3.9 y PyTorch 2.0. Las bibliotecas de Python utilizadas están optimizadas para ambas plataformas, Power e Intel. Configuración: OMP-NUM-THREADS = 4, tamaño de lote = 60 y 40 usuarios concurrentes. IBM Power S1022, 6.26 consultas por lote inferenciadas por segundo con 40 usuarios concurrentes, Sapphire Rapids 8468V, 4.4 consultas por lote inferenciadas por segundo con 40 usuarios concurrentes. Sistema IBM S1022 Power: https://www.redbooks.ibm.com/abstracts/redp5675.html Sistema x86 comparado: Sistema Supermicro SYS-221H-TNR: https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr Modelos afinados 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 (https://github.com/primeqa, basados en los modelos Dr. Decr y ColBERT). Resultados válidos a partir del 22 de agosto de 2023 y realizados bajo condiciones de laboratorio. Los resultados individuales pueden variar según el tamaño de la carga de trabajo, el uso de subsistemas de almacenamiento y otras condiciones. El resultado extrapolado para un IBM Power S1024 (2×12 núcleos 3.4-4 GHz/512 GB) se basa en un tiempo de inferencias medido de 1.008 segundos en IBM Power S1022 (2×12 núcleos 2.9-4 GHz/512 GB). La prueba se ejecutó con entornos Python y Anaconda, incluidos paquetes de Python 3.9 y PyTorch 2.0. Las bibliotecas de Python utilizadas son bibliotecas optimizadas para la plataforma Power. Configuración: OMP-NUM-THREADS = 32, Tamaño del lote = 1.*

*Sistema IBM S1024 Power: https://www.redbooks.ibm.com/abstracts/redp5675.html*

*Modelos ajustados por IBM en un corpus de datos internos de IBM: https://github.ibm.com/systems-cto-innovation/ai-on-ibm-systems/tree/master/primeqa/inference*

*[3] Comparación basada en sistemas de un solo socket (IBM Power E1080 3.55, 4 GHz, 120 núcleos, AIX y Superdome Flex 280 2.90 GHz, Intel Xeon Platinum 8380H) utilizando resultados publicados en www.spec.org/cpu2017/results/ a partir del 2 de septiembre de 2021. SPEC® y los nombres de referencia®2017_int_base y SPECrate®2017_int_peak son marcas comerciales registradas de Standard Performance Evaluation Corporation. Para conocer más acerca de SPEC CPU 2017, consulte www.http:/spec.org/cpu2017/.*

**Article Categories**

Inteligencia Artificial | Nube Híbrida