Enterprise Strategy Group™
de TechTarget

DOCUMENTACIÓN TÉCNICA

Habilitación de soluciones con tecnología Ethernet para la GenAI

La importancia de las redes abiertas

Por Bob Laliberte, Analista Principal de Enterprise Strategy Group

Enero de 2024

Contenido

La infraestructura de IA crece rápidamente

A nivel mundial, la GenAI (IA generativa) desencadenó un tsunami de interés y actividad. De hecho, los sitios web de TechTarget experimentaron un crecimiento de más del 900 % en las actividades de búsqueda relacionadas con la GenAI en 2023. Cabe destacar que este no es un simple interés. Los proveedores de servicios fueron los primeros en adoptar esta tecnología, muchos de los cuales ampliaron su portafolio de servicios para incluir ofertas de GPU como servicio, y las grandes empresas están creando una infraestructura privada de GenAI para casos de uso internos, como el análisis de consumidores y la administración de la cadena de suministro y el inventario. De hecho, muchas juntas corporativas y ejecutivos de alto nivel ya desarrollaron iniciativas para aplicar GenAI en sus procesos empresariales. Además, en la conferencia más reciente de Microsoft Ignite, el líder en GenAI y Presidente y Director Ejecutivo de Nvidia, Jensen Huang, predijo que la GenAI tendrá un impacto significativo y afirmó: “Es más grande que la PC. Es más grande que los dispositivos móviles. Será más grande que Internet”.[1]

Según ESG (Enterprise Strategy Group) de TechTarget, es fácil entender por qué las empresas están tan ansiosas por implementar soluciones de GenAI. La investigación de ESG indica que los beneficios esperados de la IA incluirán información valiosa optimizada, ingresos y rentabilidad mejorados, velocidades de toma de decisiones más rápidas, experiencias de cliente optimizadas y eficiencia operacional mejorada.[2]

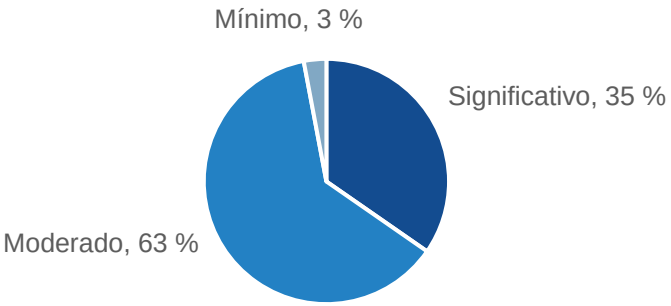
También está claro que estas iniciativas de GenAI requerirán que las empresas adopten infraestructura, software y servicios nuevos para respaldarlas. Pero esos entornos pueden variar enormemente, como señaló Jeff Clarke, Vicepresidente y Director de Operaciones de Dell Technologies. “La GenAI está lejos de ser un modelo único para todos. Requiere una solución integral, la infraestructura adecuada, un plan de datos, software y servicios que funcionen sin inconvenientes para admitir cargas de trabajo en las nubes, las instalaciones y el borde”.

La investigación de ESG demostró que más de 9 de cada 10 empresas (97 %) creen que la infraestructura de IA experimentará un crecimiento significativo o moderado debido a la GenAI (consulte la figura 1).[3] Esto será necesario para admitir los entornos de front-end (usuario) y back-end (GPU) a fin de garantizar entornos de GenAI sólidos.

[1] Fuente: CRN, “Microsoft Ignite 2023: Nvidia CEO Huang Says Microsoft Is Now ‘More Collaborative And Partner-Oriented’”, noviembre de 2023.
[2] Fuente: Resultados completos de la encuesta de Enterprise Strategy Group, *Navigating the Evolving AI Infrastructure Landscape*, realizada en diciembre de 2023.
[3] Ibid.

Figura 1. Crecimiento esperado en el mercado de infraestructura de IA debido a la GenAI

En su opinión, en términos de crecimiento del mercado, ¿qué impacto tendrá la IA generativa en el mercado de infraestructura de IA (es decir, la necesidad de comprar más infraestructura de IA para cumplir con los requisitos de capacitación y mantenimiento de modelos de lenguaje grandes)?



Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

Para reforzar aún más el deseo de adoptar GenAI, las empresas van más allá de investigar el tema y hacen planes para implementar entornos de GenAI, con investigaciones que destacan que la gran mayoría de los encuestados (92 %) planea hacerlo en los próximos 12 meses.[4]

Con ese fin, las empresas necesitan una infraestructura especializada diseñada para manejar los requisitos específicos de la GenAI, especialmente para el entorno de GPU de back-end. Sin embargo, la implementación de tecnología completamente nueva puede presentar desafíos en muchos niveles diferentes.

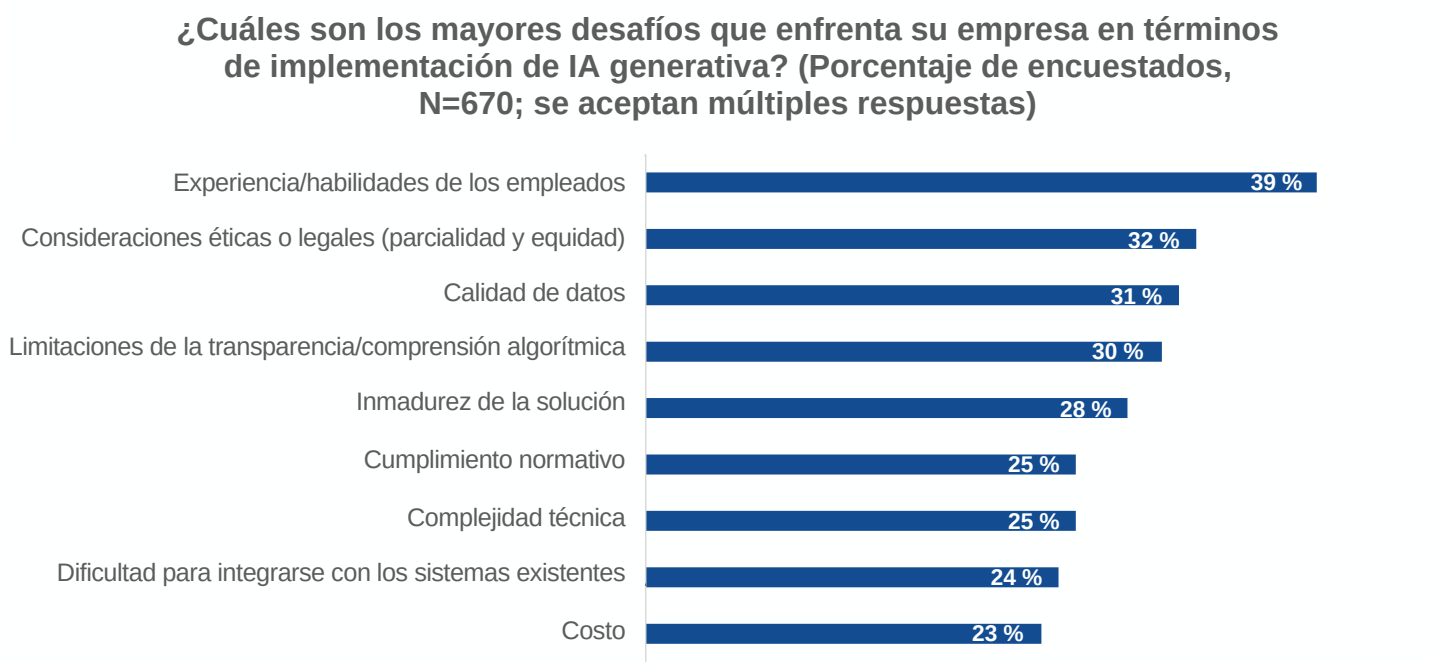
Desafíos de la transición a tecnologías nuevas

La implementación de cualquier tecnología nueva puede ser un desafío para el equipo de TI, incluso si se trata del simple reemplazo de una tecnología existente. Las tecnologías o arquitecturas nuevas pueden ser mucho más difíciles de implementar. Lamentablemente, la GenAI requiere arquitecturas nuevas, las que a su vez requieren infraestructuras nuevas de computación, almacenamiento y red, especialmente para los entornos de GPU de back-end. Esto no solo requerirá más infraestructura, sino también, lo que es más importante, sistemas cuidadosamente diseñados para adaptarse a los requisitos de conectividad masiva en los clústeres de GPU. Las conexiones típicas de 50 GbE (Gigabit Ethernet) o 100 GbE ToR (en la parte superior del rack) con enlaces ascendentes de 400 GbE causarían congestión y demoras significativas para los modelos de lenguaje grandes y pondrían en riesgo toda la iniciativa.

Cuando se les preguntó acerca de los mayores desafíos que enfrentan las empresas al implementar soluciones de IA generativa, los encuestados destacaron varios problemas, incluidos la experiencia y las habilidades de los empleados, la complejidad técnica, la incapacidad de integrarse con sistemas existentes o heredados y el costo, entre muchos otros desafíos relacionados con la calidad de los datos, las consideraciones éticas y la transparencia (consulte la figura 2). [5]

[4] Ibid.
[5] Fuente: Resultados completos de la encuesta de Enterprise Strategy Group, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, realizada en agosto de 2023.

Figura 2. Principales desafíos de la GenAI



Fuente: Enterprise Strategy Group, una división de TechTarget, Inc.

No debería sorprender que el principal desafío sea la falta de habilidades y experiencia, especialmente para una tecnología emergente como la IA generativa. La mayoría de las empresas no tendrán los recursos con las habilidades necesarias para evaluar, diseñar e implementar una infraestructura de GenAI a gran escala, especialmente los entornos de back-end con rendimiento intensivo.

La complejidad técnica también puede afectar las implementaciones de GenAI, ya que algunas soluciones usan tecnología patentada, como las redes InfiniBand, que generalmente se reservan para entornos de HPC (computación de alto rendimiento). Como resultado, hay una cantidad limitada de recursos con los conjuntos de habilidades adecuados. Esto es especialmente cierto para las empresas y los hiperescaladores que se estandarizaron en redes Ethernet. Las soluciones patentadas también pueden ser más difíciles de integrar en cualquier plataforma de monitoreo u orquestación existente, ya que se requieren habilidades, hardware y software adicionales. Otro aspecto que se debe tener en cuenta a la hora de aprovechar una solución patentada son los plazos. Dadas las complicaciones de los últimos años con la cadena de suministro, es posible que las empresas sean reacias a elegir soluciones disponibles de un solo proveedor.

Debido a estos desafíos, las empresas también luchan con los altos costos asociados a la implementación de soluciones nuevas de GenAI, especialmente las patentadas que generan dependencia de un proveedor específico a medida que escalan. El tiempo que se requiere para evaluar y diseñar una solución puede ser bastante largo si no hay diseños ni arquitecturas de referencia.

Las empresas necesitan una infraestructura de GenAI abierta y sólida

Teniendo en cuenta estas consideraciones, las empresas se ven en la necesidad de buscar soluciones abiertas que ayuden a acelerar la implementación de infraestructura de GenAI. Las empresas deberán crear nuevos entornos de front-end que permitan las interacciones de los usuarios a través de una interfaz basada en la web y que se centren en la facilidad de uso y acceso. La infraestructura de back-end es muy diferente de los entornos tradicionales o incluso de HPC y necesitaría admitir LLM (modelos de lenguaje grandes) con tecnología de clústeres de GPU capaces de consumir grandes cantidades de datos. Estos entornos de infraestructura de back-end son fundamentales para el éxito de un proyecto de GenAI.

Idealmente, estas soluciones deberían ser de la siguiente manera:

- **Integrales.** Las empresas que buscan implementar soluciones de GenAI necesitan soluciones completas para entornos de front-end y back-end a fin de acelerar la adopción. Estas soluciones incluirían los recursos adecuados de computación (incluidos los clústeres de GPU), almacenamiento y redes. Además de la infraestructura, estas soluciones requieren herramientas integrales de automatización y monitoreo no solo para la configuración inicial y la administración continua, sino también para ayudar con la optimización del fabric y el ajuste preciso del rendimiento.
- **De alto rendimiento.** Para la red, esto significa implementar fabrics sin bloqueo con entrega confiable, alto ancho de banda y baja latencia. Esta es la razón por la que se creó el UEC (Consorcio Ultra Ethernet) como parte de la Joint Development Foundation de la Fundación Linux, que reúne a empresas para la cooperación en toda la industria en torno al desarrollo de especificaciones de Ethernet y API de software que potencian entornos de IA con rendimiento, escalabilidad, confiabilidad (a través del protocolo RoCE v2, por ejemplo) e interoperabilidad de nivel superior.[6]
- **Previamente probadas y comprobadas.** Para acelerar la adopción de estos nuevos entornos de GenAI, la capacidad de implementar una solución integral que haya sido probada y comprobada para funcionar eficazmente puede ayudar a evitar los inconvenientes comunes de la implementación. Estas soluciones eliminan gran parte del tiempo de investigación, análisis y diseño, lo que permite a las empresas alcanzar sus objetivos y el valor real de sus entornos de GenAI con mayor rapidez.
- **Abiertas y extensibles.** Esto incluiría aprovechar el silicio y los fabrics Ethernet comerciales en lugar de las tecnologías de red patentadas. Los entornos de GenAI requieren tanto rendimiento de red como sea posible, pero a partir de estándares abiertos, no patentados. Para lograr esto, el UEC se asegurará de que Ethernet pueda desempeñar una función importante en los entornos de GenAI. Además, las empresas pueden aprovechar los sistemas operativos de red de código abierto disponibles en el mercado, como SONiC (Software for Open Networking in Cloud). Cabe señalar que los proyectos de SONiC y UEC son auspiciados por la Fundación Linux, lo que facilita la colaboración y la innovación de la industria.

La investigación de Enterprise Strategy Group destaca que las empresas que desean modernizar los centros de datos en las instalaciones mencionaron el uso de soluciones de hiperescala como su acción principal.[7]

- **Complementadas con servicios profesionales.** La capacidad de acelerar el tiempo de creación de valor para las soluciones de GenAI contará con la ayuda de partners que puedan aportar los conocimientos y la experiencia pertinentes. Esto incluiría la capacidad de realizar las evaluaciones adecuadas, crear los diseños e implementar soluciones de manera oportuna. También podría incluir servicios completamente administrados y diseños técnicos o validados.

[6] Consorcio Ultra Ethernet.
[7] Fuente: Informe de investigación de Enterprise Strategy Group, *2023 Technology Spending Intentions Survey*, realizado en noviembre de 2022.

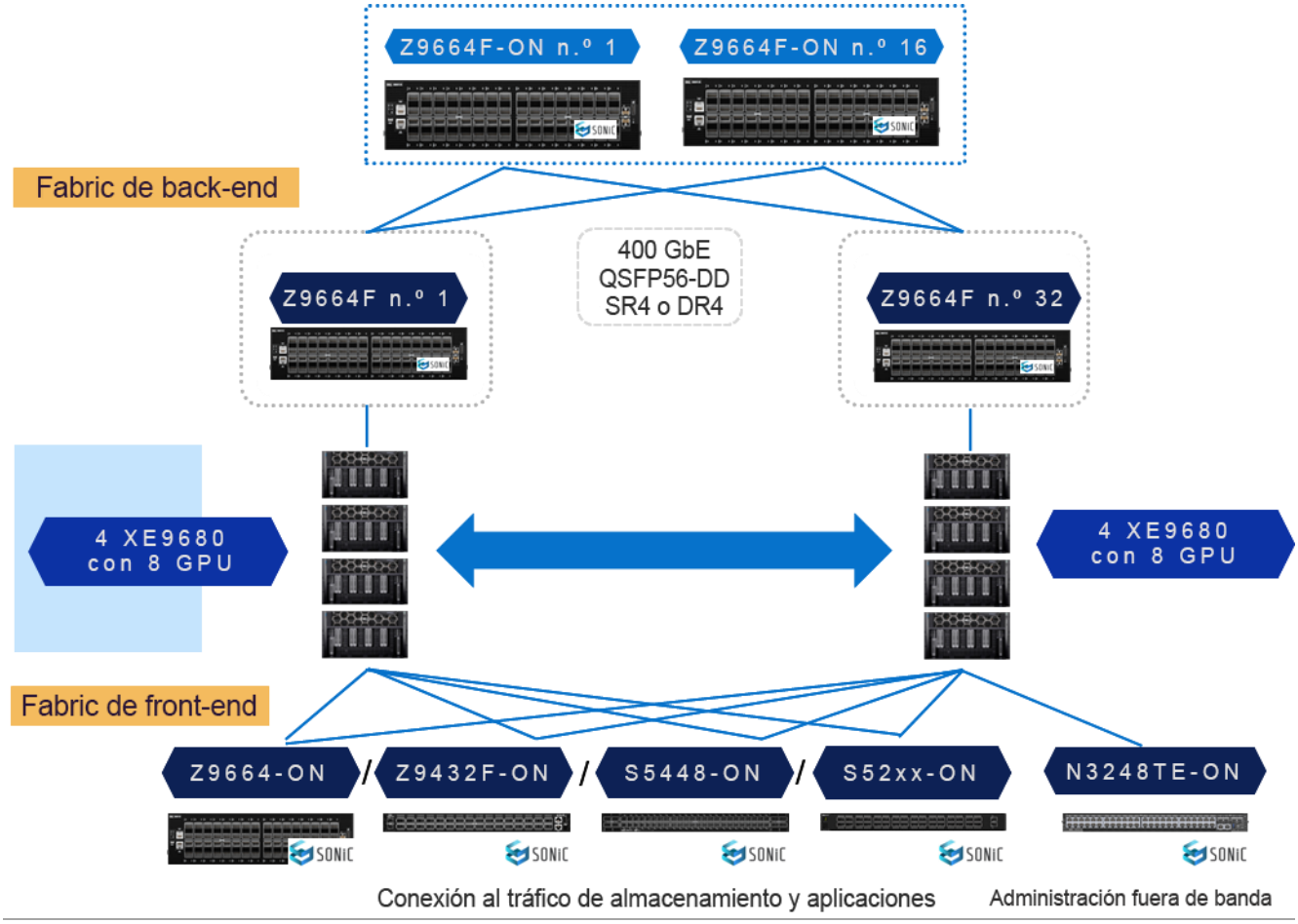
- **Escalables.** Dado que la mayoría de las empresas recién comienza con la GenAI, es posible que las implementaciones iniciales tengan un tamaño limitado, pero deberán ampliarse para adaptarse al aumento de los requisitos. Por lo tanto, será imperativo que la infraestructura de GenAI y, más específicamente, el entorno de red puedan expandirse para satisfacer estas necesidades.
- **Con uso eficiente de la energía.** Las soluciones basadas en GPU requieren grandes cantidades de energía. Por este motivo, las empresas deben tomar todas las medidas posibles para reducir la cantidad de energía consumida. Se debe utilizar tecnología de silicio de última generación que optimiza la relación rendimiento-alimentación. Los switches de mayor velocidad pueden consumir menos espacio en rack, energía y cableado para ofrecer una solución más rentable y ecológica. Además de reducir la alimentación, la capacidad de proporcionar informes de sustentabilidad también ayudará a los equipos de operaciones y administración.
- **Impulsadas por software.** Centrarse en el software acelera el ritmo de la innovación, especialmente en el caso de software desarrollado en entornos abiertos, ya que no se basa en un solo proveedor, sino en decenas de empresas que contribuyen a su innovación.

Dell Technologies ofrece soluciones abiertas con tecnología Ethernet para la GenAI

Dell Technologies ha proporcionado soluciones integrales y de infraestructura abierta para entornos de IA, modelado y HPC durante varios años. Aprovecha su experiencia previa con el fin de habilitar soluciones de infraestructura de GenAI para entornos de front-end (tráfico de aplicaciones, acceso al almacenamiento y red general) y back-end (fabric de GPU) que incluyen computación, almacenamiento y redes.

Una de las claves para habilitar una solución de GenAI de alto rendimiento es un fabric de red de IA abierto y probado, como se muestra en la figura 3.

Figura 3. Soluciones integrales de fabric de red de IA



Fuente: Dell Technologies.

Las soluciones de GenAI de Dell Technologies incluyen lo siguiente:

- **Sistemas de computación modulares.** Basados en los servidores Dell PowerEdge XE y la experiencia de la empresa en el mercado de la IA, el modelado y la HPC, estos servidores están optimizados para la aceleración de dichos entornos. Con opciones de enfriamiento por aire o líquido, así como la cantidad de GPU, junto con un enfoque en la inferencia o la capacitación de LLM, Dell tiene el factor de forma y la solución de alto rendimiento adecuados para satisfacer sus necesidades de computación de GenAI. Los entornos de computación forman parte de una solución de arquitectura y diseño validada para la GenAI.
- **Almacenamiento enfocado en la IA.** Dell tiene una variedad de opciones de almacenamiento disponibles según los requisitos de la carga de trabajo, incluidas las soluciones PowerScale, Elastic Cloud Storage y ObjectScale. El almacenamiento PowerScale OneFS basado en Ethernet permite lecturas y escrituras de streaming para acceder rápidamente a los datos de las cargas de trabajo de IA y mejora la funcionalidad de modelado de IA. Dell menciona que PowerScale se probó en campo con más de 1000 clientes que ejecutan cargas de trabajo de GPU en ellos. Como resultado, existen numerosas soluciones de Dell Validated Design basadas en estas experiencias. La amplia gama de opciones también cuenta con la certificación Energy Star.

- **Fabrics Ethernet de última generación.** Este hardware de red abierta, basado en Dell PowerSwitch y con silicio de última generación, como Broadcom Tomahawk 4, puede proporcionar hasta 51,2 Tb/s con almacenamiento en buffer de paquetes compartidos. Disponibles en el mercado como PowerSwitch de la serie Z, el switch Z9664F-ON de 64 puertos y el switch Z9432F-ON de 32 puertos pueden escalar para admitir miles de nodos. Además, Dell Technologies es miembro del UEC y contribuirá a extender la aplicabilidad de Ethernet para potenciar los entornos de GenAI.
- **Arquitectura basada en software.** Dell Technologies mantiene su compromiso de proporcionar soluciones de redes abiertas para los sistemas operativos de red, la orquestación y el monitoreo en entornos de GenAI. En el caso del sistema operativo de red, Dell Technologies adoptó y reforzó SONiC, lo que proporciona el soporte, la escala y las características globales que requieren las empresas grandes. La versión más reciente de Enterprise SONiC Distribution by Dell Technologies (versión 4.2) proporciona soporte avanzado para entornos de IA que incluye RoCE v2 (RDMA over Converged Ethernet versión 2), hash mejorado y conmutación de corte. La próxima versión 4.3 proporciona mejoras para el balanceo de carga y la asignación. Todas las versiones de SONiC se prueban y se validan en todo el portafolio de la serie Z. Las versiones también se prueban en el ecosistema de partners de aplicaciones de otros fabricantes de Dell.
- **Proporcione servicios para acelerar la adopción y la optimización.** Además del soporte global 24x7, Dell Technologies tiene expertos en servicios profesionales con experiencia comprobada para permitir que las empresas evalúen, diseñen e implementen correctamente soluciones integrales de GenAI. Su capacidad para comprender no solo la red, sino también los dominios de computación y almacenamiento, acelera el proceso de diseño y reduce el riesgo de que surjan problemas de compatibilidad. Estos diseños validados cubren tanto la inferencia como la personalización de modelos, y hay servicios que cubren la preparación y la ingesta de datos para los pipelines de GenAI. Dell también ofrece servicios administrados para operar estos entornos de IA.
- **Enfóquese en la sustentabilidad.** La implementación de entornos de GenAI a escala requiere importantes recursos de energía. Los switches de mayor velocidad de Dell en modo de conexión múltiple requieren menos espacio en rack, alimentación y cableado. Aprovechar la última tecnología de silicio permite que los servidores, las redes y las soluciones de almacenamiento sean lo más eficientes posible en el uso de la energía. Centrarse en la eficiencia energética permite a las empresas reducir los costos y el consumo de energía.

Con estas integraciones, Dell Technologies está bien posicionado para ofrecer soluciones completas de infraestructura de GenAI destinadas a entornos de back-end y front-end.

Conclusión

El aumento en el interés y la actividad de la GenAI motiva a las empresas a evaluar soluciones para sus propios entornos. Sin embargo, debido a su reciente popularidad, la mayoría de los equipos de TI carecen de la pericia o la experiencia para implementar una solución de manera oportuna. Además, para ser justos, estas infraestructuras de GenAI que requieren arquitecturas y tecnologías nuevas son muy complejas. Deben diseñarse cuidadosamente y proporcionar un sistema equilibrado, por lo que tratar de obtener componentes individuales y unirlos puede ser muy arriesgado. Debido a esto, las empresas deben asociarse estratégicamente para adquirir las habilidades y soluciones estrechamente integradas a fin de garantizar un entorno de GenAI exitoso.

Sin embargo, las empresas deben tener cuidado con las soluciones integrales que las limitan a tecnología patentada, especialmente a medida que estos entornos escalan. Las soluciones abiertas pueden proporcionar innovación, flexibilidad y rentabilidad para entornos de GenAI a gran escala. Sin embargo, para garantizar entornos sólidos, también es fundamental asegurarse de que estas soluciones abiertas estén probadas y validadas de forma integral y sean completamente compatibles.

Dell Technologies proporciona soluciones completas de GenAI que incorporan toda la infraestructura y todo el software, incluidas la orquestación y la administración de entornos de front-end y back-end. También incorporan computación, almacenamiento y redes abiertos. Además, las empresas pueden aprovechar los servicios administrados, los servicios profesionales y las arquitecturas y los diseños completamente validados que incluyen el ecosistema de partners de Dell. Estas soluciones integrales y modulares permiten a las empresas acelerar la implementación y el valor de las soluciones de GenAI, junto con reducir el riesgo y garantizar una mayor eficiencia operacional.

Acerca de Enterprise Strategy Group
Enterprise Strategy Group de TechTarget ofrece inteligencia de mercado focalizada y procesable, investigación de la demanda, servicios de asesoría de analistas, orientación estratégica de GTM, validaciones de soluciones y contenido personalizado de apoyo a la compra y venta de tecnología empresarial.

contact@esg-global.com
www.esg-global.com