White Paper

¿Por qué desarrollar e implementar tecnología basada en IA en estaciones de trabajo tiene sentido?

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

LA OPINIÓN DE IDC

La IA ha logrado despuntar como una tecnología importante y elemento diferenciador en todos los sectores; además, el hardware que requiere este tipo de tecnología está evolucionando a toda velocidad. Por lo general, el sector tecnológico presta especial atención al crecimiento exponencial del tamaño de los modelos de IA más avanzados. Se debate sobre las decenas de miles de millones de parámetros, la mejora en la precisión, la ampliación de memoria, las necesidades similares a las de la computaci'on de alto rendimiento (HPC) para el entrenamiento e inferencia de la IA y los racks de servidores acelerados. En realidad, la magnitud extraordinaria que ha adquirido la computación de IA es la excepción a la regla, en particular, en las empresas.

Actualmente, muchas empresas trabajan intensamente en iniciativas de IA, incluida la IA generativa que no requiere superordenadores. De hecho, gran parte del desarrollo en IA (y, cada vez más, las implementaciones de IA, principalmente en entornos EDGE computing o computación en el borde) se ejecuta en realidad en potentes estaciones de trabajo. Las estaciones de trabajo comportan numerosas ventajas para el desarrollo y la implementación de las tecnologías de IA. Gracias a ellas, los científicos o desarrolladores de IA ya no tienen que negociar el tiempo del servidor; además, ofrecen aceleración de la GPU y aunque las GPU basadas en servidor todavía no están fácilmente disponibles en el centro de datos, son muy asequibles en comparación con los servidores, ya que el gasto es único e inferior al no estar sujetas a una facturación que se incrementa velozmente. Tampoco se puede soslayar la comodidad de saber que los datos confidenciales se encuentran almacenados de forma totalmente segura en el entorno local. De esta manera, desarrolladores y científicos se liberarán del estrés que supone acumular costes por simplemente experimentar con modelos de IA.

IDC observa que estos entornos EDGE computing se expanden más rápido que el entorno local o la nube como escenario para la implementación de tecnologías de IA. A este respecto, la función de las estaciones de trabajo como plataformas de inferencia de IA es también cada vez más importante y, en muchas ocasiones, basta con realizar la inferencia en CPU optimizadas por software en lugar de en GPU. Los casos de uso de inferencia de IA en estaciones de trabajo en entornos EDGE aumentan velozmente e incluyen AIOps, respuesta ante desastres, radiología, explotación de crudo y gas, ordenación territorial, atención médica virtual, gestión del tráfico, supervisión de plantas de fabricación y drones.

July 2023, IDC #US51017223

En este documento técnico se contempla el papel cada vez más predominante de las estaciones de trabajo en el desarrollo e implementación de la IA, y se describe brevemente la cartera de estaciones de trabajo para IA que ofrece Dell.

DESCRIPCIÓN GENERAL DE LA SITUACIÓN

El auge de la IA y sus consecuencias en la infraestructura

El número de proyectos de IA que emprenden las organizaciones de todo el mundo aumenta rápidamente. En estos momentos y en todos los sectores, muchas tareas se desempeñan con software impulsado total o parcialmente por modelos de IA. IDC estudia las tecnologías de IA en muchos ámbitos, y una métrica que se debe tener en cuenta es la cantidad que se prevé que deben invertir las empresas y proveedores de servicios de cloud en servidores para el desarrollo y la ejecución de la IA. En 2026, está cantidad ascenderá a 34 600 millones de USD, lo que supone cerca del 22 % del total del gasto en servidores a nivel mundial.

Pero los servidores no son el único factor a tener en cuenta. En las estaciones de trabajo se efectúa mucho trabajo de desarrollo, prototipado y, cada vez con más frecuencia, *implementación* de IA. Las organizaciones, grandes o pequeñas, han visto que pueden captar nuevas oportunidades empresariales introduciendo cierta funcionalidad de IA en sus aplicaciones, algo que ha disparado la experimentación con modelos de IA, y las estaciones de trabajo robustas resultan idóneas para este fin gracias a su disponibilidad inmediata y a su proximidad a los datos.

Cabe preguntarse cómo ha ganado tanto terreno la IA de forma repentina, cuando los algoritmos de IA se llevan implementando desde hace décadas. Esto se debe básicamente a que en los últimos años se han cumplido dos condiciones básicas que han propulsado la red neuronal, un tipo particularmente ventajoso de algoritmo de IA, y son las siguientes: la facilidad de disponibilidad de tipos de datos extensos, baratos y diversos, como los datos no estructurados y semiestructurados; y, por otro lado, el aumento de una computación lineal con un modelo paralelo para procesar estas redes neuronales dentro de un plazo aceptable. Al cumplirse estas dos condiciones esenciales, los científicos de datos han avanzado enormemente en el desarrollo de las redes neuronales, que aprenden automáticamente cómo ejecutar tareas cada vez más impresionantes. Mientras que el aprendizaje automático (ML) sigue siendo importante para datos de texto o numéricos, el aprendizaje profundo (DL) resulta más eficiente en modalidades como el vídeo, el audio o los lenguajes.

Los modelos de aprendizaje automático tradicional se pueden desarrollar generalmente en las CPU de una estación de trabajo, con un máximo de varias decenas de núcleos, pero las redes neuronales requieren procesadores de apoyo para ejecutar el procesamiento en paralelo en miles de núcleos. El principal motivo de esto es que en el aprendizaje automático (ML), la extracción y la clasificación de funciones se realiza a través de un proceso manual, mientas que en el lenguaje profundo (DL) el proceso es automático, lo que hace que el modelo se deba entrenar a través de la repetición constante mediante grandes conjuntos de datos. Actualmente, el procesador de apoyo más común es la GPU, pero las empresas emergentes están comercializando asimismo nuevos procesadores específicos para tecnologías de IA. Este tipo de aceleración, es decir, utilizar un procesador de apoyo discreto para el procesamiento en paralelo, ha revolucionado los mercados de los servidores y las estaciones de trabajo, lo que ha hecho surgir lo que IDC denomina la “computación en paralelo masiva”.

En 2022, los servidores acelerados representaban 21 800 millones de USD en el mercado global, una cifra que llegará a 43 400 millones de USD en 2026, con un 57 % de este total para los servidores

acelerados que ejecutan IA. Al mismo tiempo, la cantidad de GPU discretas que se vendieron para estaciones de trabajo aumentó hasta 6,4 millones en 2022. IDC calcula que el mercado de las estaciones de trabajo para usos de ingeniería de software o científicos, sectores cada vez más dependientes del desarrollo de la IA, aumentará hasta casi 2000 millones de USD para 2026.

Fases del desarrollo de proyectos basados en IA

Como hemos dicho anteriormente, las redes neuronales empezaron a ser factibles gracias a la expansión de los tipos y los volúmenes de datos, así como a la aparición de nuevos enfoques de computación. La primera parte de esta ecuación, que son los volúmenes y los tipos de datos, no es cosa baladí: se cree que hasta el 80 % del trabajo en iniciativas de IA se dedica a la administración y la preparación de los datos. La recopilación, administración y preparación de los datos deben realizarse antes de que pueda comenzar el diseño y el entrenamiento de los modelos. Según IDC, las fases del desarrollo de la IA son las siguientes (véase la Figura 1):

- **Administración de datos:** identificar y administrar los datos importantes para el modelo de IA a partir de los enormes volúmenes de datos que residen en el centro de datos, el perímetro y la cloud que recopila, genera y/o adquiere una organización (estos datos pueden ser de cualquier tipo, basados en eventos o de streaming, y es posible que muchos requieran algún tipo de gobernanza).
- **Preparación de datos:** almacenar los datos (archivos, bloques u objetos) en un almacén de datos o lago de datos, limpiarlos, asegurarse de que estén completos y sean de alta calidad y, posteriormente, pasarlos a un formato que pueda utilizar el modelo de IA (por ejemplo, con Spark o herramientas como Pandas).
- **Elección del modelo:** decidir qué modelo es el que ofrece un rendimiento óptimo para la tarea de IA para la que se ha programado en términos de índice de error y/o rendimiento.
- **Desarrollo de los modelos:** diseñar el modelo de IA mediante infraestructuras como XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn o H2O.
- **Entrenamiento de modelos:** entrenar el modelo en la infraestructura de computación con los procesadores y/o núcleos de procesadores de apoyo suficientes para tareas en paralelo (lo que incluye cada vez con más frecuencia la capacidad de explicar, validar y documentar las decisiones de los modelos con el fin de garantizar la imparcialidad, reconocer responsabilidades y ofrecer transparencia). (Todo ello incluye el prototipado, es decir probar un modelo entrenado mediante la ejecución de la inferencia sobre dicho modelo).
- **Alojamiento y supervisión de modelos:** implementar el modelo en un entorno de producción con el fin de ejecutar la tarea para la que se ha diseñado (normalmente, denominado "inferencia de IA") y supervisar su rendimiento.

Las estaciones de trabajo pueden desempeñar una importante función en cualquiera de estas seis fases en combinación con el centro de datos, la cloud o la infraestructura en el borde.

FIGURA 1

Fases del desarrollo de proyectos basados en IA

Fuente: IDC, 2023

DESARROLLO DE MODELOS DE IA EN LAS ESTACIONES DE TRABAJO

Estaciones de trabajo y ordenadores personales

En general, se sabe que los ordenadores personales (PC) no tienen la potencia suficiente para el desarrollo de la IA. Los científicos de datos y los desarrolladores de IA suelen trabajar en proyectos estratégicamente importantes para sus organizaciones, donde una productividad sin obstáculos es de capital importancia. El rendimiento de las estaciones de trabajo suele ser más predecible que el de los PC, ya que, normalmente, sus componentes ofrecen mayor rendimiento y están optimizadas para el software que se ejecuta en ellas.

Entre los componentes, se incluyen:

- **Procesadores de gran calidad:** un ejemplo son los procesadores escalables Intel Xeon.
- **GPU muy eficaces:** un ejemplo son las GPU profesionales RTX de NVIDIA, como NVIDIA RTX 6000 Ada.
- **Más almacenamiento:** algunas estaciones de trabajo ofrecen hasta 60 TB y la velocidad de las operaciones de E/S suele ser bastante más alta que la de los PC.
- **Más memoria:** las estaciones de trabajo ahora están disponibles con hasta 6 TB de memoria.
- **Refrigeración:** los componentes generan mucho calor y los científicos de datos necesitan estaciones de trabajo con una refrigeración adecuada para evitar el sobrecalentamiento y conseguir un rendimiento óptimo en todo momento.
- **Tarjeta de interfaz de red (NIC):** para los científicos de datos que trabajan con conjuntos de datos voluminosos almacenados en servidores remotos, disponer de una tarjeta de interfaz de red es fundamental para transferir los datos rápida y eficazmente.

- **Pantalla:** es importante disponer de una pantalla de alta calidad para las tareas de visualización de datos, por lo que los científicos de datos deberían buscar monitores de alta resolución, con colores fieles a la realidad y un gran tamaño de pantalla.
- **Memoria con códigos de corrección de errores (ECC):** ECC detecta y corrige los tipos de corrupción de datos internos más comunes, lo que evita las pantallas azules en ejecuciones prolongadas de entrenamiento de la IA a causa de "errores duros" (bits corruptos) o "errores suaves" (cambio de bits que origina valores equivocados). Por otra parte, ECC también garantiza la precisión de los resultados, un requisito esencial en trabajos de vital importancia, como los servicios de atención médica.
- **Tecnología Silicon especializada:** un ejemplo son las unidades de procesamiento de visión Intel Movidius (VPU), que son procesadores de apoyo para operaciones en paralelo para aplicaciones de IA de visión artificial y en el perímetro, las cuales se utilizan en ámbitos como el sector minorista, la seguridad y la automatización industrial. Los FPGA también se utilizan en estaciones de trabajo, por ejemplo, para aplicaciones financieras.
- **Software para optimización:** entre los ejemplos, se incluye OneAPI, que es un modelo de programación basado en los estándares de Intel dirigido a simplificar el desarrollo e implementación de cargas de trabajo centradas en datos en CPU, GPU, FPGA y otros aceleradores o bien, CUDA (Compute Unified Device Architecture), que es la plataforma de computación en paralelo y la interfaz para la programación de aplicaciones de NVIDIA utilizada en la ejecución de cargas de trabajo generales en GPU.

Comparación de las CPU y las GPU para IA

Las estaciones de trabajo se pueden utilizar en diversas fases del desarrollo de la IA y, por lo general, vienen equipadas para varios tipos de funcionalidad. Pese a la importancia de las GPU en el procesamiento en paralelo, las CPU desempeñan una función esencial a la hora de desarrollar un modelo de IA en una estación de trabajo. Al igual que las GPU, las CPU también se pueden utilizar para la manipulación de los datos y, evidentemente, para el desarrollo de modelos de ML tradicionales. Las CPU se utilizan también para la exploración de datos (el proceso de utilizar representaciones visuales de un conjunto de datos para entender las características de los datos).

En el entrenamiento del DL, la función de la CPU en el host se minimiza bastante cuando las GPU se encargan del proceso de entrenamiento real, pero incluso ahí, las CPU siguen empleándose a modo de capa de procesamiento para software crítico, como el sistema operativo o CUDA, así como para la coordinación de procesos entre GPU u otras tecnologías Silicon. Además, las CPU cada vez se utilizan más como motores de inferencia de IA, una nueva función para casos en los que las estaciones de trabajo se utilizan para ejecutar modelos de IA en el entorno de producción. IDC prevé que, para 2024, el gasto en infraestructura para la inferencia de IA superará el gasto en infraestructura para el entrenamiento de la IA y que una parte importante (39 %) de esta inferencia tendrá lugar en las CPU del host.

Estaciones de trabajo y servidores: una relación simbiótica

Para la mayoría de las organizaciones, el pragmatismo es la regla general respecto a cuándo una estación de trabajo, un servidor local, una instancia de cloud o cualquier combinación de estos tres elementos se va a implementar en el desarrollo de IA. Entre las estaciones de trabajo, los servidores y las instancias de cloud se da una relación simbiótica en las diferentes fases de desarrollo de un proyecto de IA.

La ventaja de las estaciones de trabajo frente a los servidores de centro de datos es que los científicos de datos pueden trabajar desde donde quieran, un factor importante en la pandemia, pero también en circunstancias normales. También pueden experimentar con total libertad en sus modelos de IA y realizar iteraciones con la frecuencia que estimen necesaria, ya que la potencia de las estaciones de trabajo modernas, que integran GPU muy eficaces, suele hacer que el proceso de iteración sea más interactivo, lo que ofrece respuestas y resultados en el acto, sin tener que solicitar acceso a los servidores o lidiar con otras restricciones del centro de datos. Además, las estaciones de trabajo ofrecen la flexibilidad necesaria para acercar más los recursos informáticos a los datos, en lugar de a la inversa, lo que ahorra ancho de banda, reduce la congestión de la red y aumenta el rendimiento. Adicionalmente, las estaciones de trabajo se pueden configurar para diferentes necesidades: tareas de ML tradicionales, por ejemplo, o tareas más intensas en lo referente a DL.

Por otra parte, pese a que el mercado de los servidores acelerados ha crecido significativamente, estos no se encuentran todavía disponibles de forma generalizada en los centros de datos empresariales. Cuando se confeccionó este documento técnico, se aceleraba de media el 4 % de los servidores en los centros de datos empresariales, lo que pone de manifiesto que muchas organizaciones no disponen de los medios para desarrollar o ejecutar tecnología de IA en las GPU disponibles en el entorno local. También por este motivo, las estaciones de trabajo representan una alternativa muy práctica para el desarrollo de IA.

Las estaciones de trabajo muy aceleradas ya son lo suficientemente potentes como para llevar a cabo el entrenamiento de DL siempre que el modelo de IA no sea demasiado extenso, lo cual elimina la necesidad de entrenar en servidores. Además, los modelos entrenados en estaciones de trabajo con GPU pueden implementarse tanto en estaciones de trabajo como en servidores sin GPU, de forma que se pueden aprovechar las funcionalidades de inferencia de las CPU. Existen tecnologías de software como DL Boost y oneAPI de Intel que pueden impulsar la inferencia de IA en la CPU, lo que permite que los servidores no acelerados que ya se han implementado en los centros de datos puedan admitir aplicaciones de IA.

Estaciones de trabajo y cloud

El cloud computing ha revolucionado la forma en que las organizaciones conciben la infraestructura, los datos y las aplicaciones. Con una capacidad de ampliación previsible prácticamente sin límites, la cloud permite a los desarrolladores aprovisionar recursos según las necesidades, lo que seguramente aceleraría el ritmo de innovación con menos restricciones. A primera vista, la cloud parece ser el paradigma perfecto para el desarrollo de la IA.

Sin embargo, este no es siempre el caso. De hecho, informes de IDC reflejan que las organizaciones repatrían ciertas cargas de trabajo de la cloud pública a la infraestructura local cada vez con mayor frecuencia. Esto se debe a diversos factores:

- **Disponibilidad de la cloud:** aquellos que hayan confiado en los servicios de cloud habrán sufrido interrupciones, bien sea por problemas del proveedor de cloud o por un fallo en la conectividad de red en algún punto entre el centro de datos hiperescalado y el usuario final. En este tipo de situaciones, los usuarios están a merced del proveedor de servicios para resolver el problema, mientras que la productividad se detiene por completo.
- **Seguridad y cumplimiento:** en muchos sectores, las directivas de gobernanza corporativa dictan dónde se pueden comunicar y almacenar los datos, algo que limita el uso de los servicios de cloud. Normativas gubernamentales como el RGPD en Europa y la Ley de Privacidad de los Consumidores de California también dictan reglas en materia de soberanía de los datos.

- **Coste:** es habitual que las organizaciones subestimen la rapidez con la que pueden aumentar las tasas del servicio de cloud, en especial, para cargas de trabajo que requieren recursos de computación de alto rendimiento y mucha capacidad de almacenamiento. La economía de cloud se basa en medir todos los tipos de consumo de recursos, incluso el envío de datos de vuelta hacia la infraestructura local.
- **Presiones que suscita el método de prueba y error:** la mayoría de las iniciativas de IA comienzan con la ejecución de gran cantidad de pruebas. Hay modelos que fallan, lo cual también forma parte de la fase de desarrollo. Este proceso pasa una factura psicológica que pagan científicos y desarrolladores de IA cuando se acumula la facturación por servicios de cloud sin que puedan mostrar aún resultados ejecutables.

Las estaciones de trabajo pueden abordar este tipo de limitaciones al mismo tiempo que utilizan las tecnologías nativas de cloud, como las arquitecturas basadas en microservicios y la automatización impulsada por API. De esta forma, se obtienen algunos de los beneficios que se apuntaron al comparar las estaciones de trabajo con los servidores de centros de datos:

- **Trabajo desde cualquier lugar:** al eliminar la dependencia de la cloud pública, es posible prescindir de estar conectado. Muchos entornos de alta seguridad están protegidos con capa de aire de las redes públicas y las estaciones de trabajo de IA pueden abordar esta necesidad como ninguna otra tecnología. Los recursos locales también reducen la demanda de sistemas de conectividad de red de alto precio.
- **Localidad de los datos:** la proliferación de los dispositivos de IoT y otros equipos conectados favorece el crecimiento exponencial de los datos en las ubicaciones del perímetro. En muchos casos, resulta razonable alojar los recursos de computación en una estación de trabajo dedicada. De este modo, se solventan muchos de los requisitos de cumplimiento de normas, ya que se limita el movimiento de datos.
- **Libertad para experimentar:** el entrenamiento y la optimización de los modelos de IA es un proceso iterativo que suele incluir algún elemento del método de prueba y error. Los desarrolladores necesitan libertad para desarrollar pruebas sin tener que escatimar por que exista la posibilidad de incurrir en tasas de servicio adicionales. Las estaciones de trabajo ofrecen, asimismo, más flexibilidad para las herramientas personalizadas.

En lo referente al último punto, comparar el precio de una estación de trabajo con una implementación de cloud resulta relativamente sencillo, dado que la mayoría de los proveedores de servicios de cloud ofrecerán inmediatamente previsiones de los costes para cualquier tipo de configuración que desee implementar el usuario final. Por ejemplo, el coste de una única máquina virtual (VM) normal con una tarjeta NVIDIA T4 y una instancia de almacenamiento SSD de 375 GiB que se utilice ocho horas al día y cinco días a la semana asciende a 140 USD con uno de los proveedores de cloud principales. Si duplica las máquinas virtuales, las T4 y las SSD, el coste ascenderá a 365 USD al mes. Si se queda con las dos máquinas virtuales y duplica las T4 a cuatro y el almacenamiento a 4 × 375 GiB, y ejecuta entrenamientos a tiempo completo en el entorno, el coste alcanzará los 2700 USD al mes. Por tanto, sería razonable afirmar que los costes de cloud para el desarrollo de IA pueden dispararse fácilmente a decenas de miles de dólares al año, mucho más que la amortización anual de una estación de trabajo de alto nivel.

PROTOTIPADO DE TECNOLOGÍAS DE IA EN ESTACIONES DE TRABAJO

En comparación con los servidores locales y la cloud, las estaciones de trabajo ofrecen una clara ventaja cuando se trata del prototipado de modelos de IA. Los servidores del centro de datos pueden estar a pleno uso o también resultar esenciales para el prototipado y las pruebas de IA; además, tal como se indicó anteriormente, las instancias de cloud pueden incrementar rápidamente los costes cuando se utilizan sin limitación alguna como entorno de pruebas. Las estaciones de trabajo liberan al científico o desarrollador de IA de tener que negociar el acceso al servidor o de la preocupación de ver cómo se acumulan las facturas de cloud durante la fase del prototipado. Gracias al coste puntual, que es bastante asequible, se disfruta de plena libertad para realizar prototipos en cualquier lugar y momento sin incurrir en gastos adicionales.

IMPLEMENTACIÓN DE MODELOS DE IA EN LAS ESTACIONES DE TRABAJO

Pese a que el desarrollo de modelos de IA en estaciones de trabajo es una estrategia común que se lleva aplicando años, IDC observa un incremento de los casos de uso de *implementación* de modelos de IA en estaciones de trabajo, normalmente en el perímetro o EDGE; es decir, integrar el modelo de IA del entorno de producción en la estación de trabajo mediante la ejecución de inferencias en el modelo de IA. El perímetro se afianza a toda velocidad como ubicación de implementación de IA para servidores (desde 2020 hasta 2024 se ha más que triplicado en términos de gasto anual en hardware) y las estaciones de trabajo no van muy por detrás, porque los usuarios finales se han dado cuenta de las ventajas ofrecidas en el perímetro.

IDC define el perímetro o EDGE como un paradigma de computación distribuido que incluye la implementación de infraestructura y aplicaciones fuera de la cloud centralizada y en centros de datos locales lo suficientemente próximos al punto donde se generan y se consumen los datos. Esto incluye las sucursales y las oficinas remotas, así como las ubicaciones específicas del sector, como las fábricas, los almacenes, los hospitales y los puntos de venta.

Las cargas de trabajo de uso intensivo de datos se implementan en el entorno local o en el perímetro cada vez con mayor frecuencia. El propósito es mitigar las limitaciones inherentes a las clouds públicas, como el tiempo que lleva cargar conjuntos de datos de gran tamaño y los costes variables de entrenar la IA, particularmente en situaciones que requieren muchas pruebas de ciencia de datos.

El informe de IDC pone de manifiesto que las implementaciones de IA en el perímetro avanzan rápidamente con una inversión de 2900 millones de USD en IA por parte de las organizaciones en 2023, una cifra que alcanzará los 6900 millones de USD en 2026 (véase *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage*, IDC #US49671722, septiembre de 2022). Pero además, el perímetro está adquiriendo peso como entorno de implementación para cargas de HPC, como, por ejemplo, en los campos de ingeniería y técnicos; de hecho la inversión actual de las empresas en tales cargas de trabajo ronda los 1000 millones de USD, cifra que llegará a los 2400 millones de USD en 2027 (véase *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs*, IDC #US50525123, abril de 2023). Es en estas áreas donde resulta interesante implementar estaciones de trabajo de IA.

A la hora de implementar un modelo de IA en una estación de trabajo en el perímetro, no siempre es necesario tener GPU de gama alta, como es el caso de las implementaciones de IA. Hay GPU menos avanzadas que pueden llevar a cabo la inferencia de IA y no son pocos los casos en los que ni siquiera hace falta una GPU. En este tipo de situación, las CPU pueden ejecutar correctamente las tareas de inferencia, sobre todo cuando se combinan con optimizaciones como DL Boost de Intel, un conjunto de instrucciones que define una serie de funciones en los microprocesadores de Intel diseñadas para acelerar las cargas de trabajo de IA, incluida la inferencia de IA. Con DL Boost, Intel afirma haber multiplicado por 1,45 el rendimiento de la inferencia en tiempo real de INT8 con procesadores escalables Intel Xeon de 4.ª generación, los cuales admiten DL Boost de Intel en comparación con los de la generación anterior (BERT-Large SQuAD). Esto también favorece la idoneidad de la estación de trabajo para su implementación en el perímetro, donde aspectos como el consumo eléctrico, la movilidad y la gestión térmica no necesitan tanta potencia. Intel Movidius Myriad (M2) se adapta bien en estos requisitos de de potencia gracias a la poca energía que requiere, tan solo 12 W.

Casos de uso para implementar IA en estaciones de trabajo

Existen diversas situaciones que se prestan de manera natural a la implementación de IA en estaciones de trabajo implementadas localmente. Los rasgos comunes son los grandes volúmenes de datos de series temporales generados por ordenador y los datos no estructurados como secuencias de vídeo e imágenes. También existen casos en los que expertos en la materia deben mejorar los modelos de IA con la interpretación humana.

Algunos ejemplos son:

- **AIOps:** a medida que los sistemas de TI van aumentando en complejidad y tamaño, resulta cada vez más necesario pasar de una administración de incidentes reactiva a una supervisión proactiva. Esto se hace especialmente cierto cuando la infraestructura y las aplicaciones se instalan en ubicaciones en el perímetro donde el personal técnico es escaso o inexistente. Al establecer una referencia de rendimiento normal, es posible identificar las anomalías y automatizar las medidas de corrección.
- **Respuestas ante desastres:** en caso de emergencia, las primeras intervenciones deben ir dirigidas a evaluar rápidamente la situación, llevar el seguimiento del equipo esencial y desplegar recursos para prestar asistencia donde más se necesite. Esto suele suceder en entornos sin conectividad de red, que necesitan una estación de trabajo local con capacidad para agregar fuentes de datos, sacar conclusiones con arreglo a los modelos de IA y automatizar las comunicaciones para el personal clave.
- **Radiología:** los avances en la tecnología de creación de imágenes han provocado un aumento del tamaño de los datos generados a partir de un único escáner, lo cual obliga a mantenerlos in situ para analizarlos con prontitud. Los modelos de IA que se han entrenado a partir de millones de casos anteriores pueden identificar patrones con mayor precisión que una persona, lo que aumenta el índice de exactitud.
- **Explotación de crudo y gas:** las empresas petroleras y de gas upstream utilizan una combinación de datos de telemetría, sísmicos y de imágenes para localizar reservas de recursos naturales, seleccionar ubicaciones de perforación y optimizar el rendimiento de los equipos en los procesos de producción. Por lo general, todo esto requiere un análisis de la información en zonas donde solo es posible comunicarse por satélite, algo que resulta muy caro.

- **Investigación oncológica y desarrollo de medicamentos:** los investigadores en hospitales y centros académicos utilizan la IA y el procesamiento del lenguaje natural para ayudar a los oncólogos a determinar cuál es el tratamiento individualizado más efectivo para el cáncer en pacientes. También combinan el aprendizaje automático con la visión artificial para ofrecer a los radiólogos un mejor entendimiento de la evolución de los tumores. Asimismo, utilizan algoritmos para comprender cómo se desarrolla el cáncer y qué tratamientos son más efectivos para combatirlo.
- **Evaluación de las reclamaciones a empresas aseguradoras:** el procesamiento manual de las reclamaciones es laborioso y está sujeto al error humano. Si la IA puede evaluar la validez de la reclamación, los costes se verán reducidos, ya que los peritos del seguro podrán centrarse en aquellos casos que requieran más investigación. Esto aumenta el rendimiento global de la operación sin sacrificar la precisión.
- **Atención médica virtual:** la IA está mejorando los índices de recuperación de los pacientes al individualizar la planificación del tratamiento en función de la información vital en tiempo real que registran los dispositivos que llevan los pacientes. Esta información se combina con la historia clínica del paciente y una base de conocimientos de casos similares. Esto resulta particularmente importante en zonas rurales con mayor dependencia de la atención sanitaria remota.
- **Seguridad en los comercios (antirrobo):** los análisis en tiempo real aplicados a las secuencias de vídeo se utilizan para predecir el comportamiento humano que podría derivar en actividad criminal. Para ello, lo habitual es unir varias fuentes de vídeo con el fin de llevar el seguimiento de los movimientos de una persona en una tienda. Dada la urgencia a la hora de interpretar hechos relevantes, es mejor que este proceso se pueda ejecutar localmente.
- **Gestión del tráfico:** los organismos gubernamentales encargados de las operaciones de transporte utilizan cada vez más la IA para coordinar los semáforos y la señalización digital con el fin de mejorar el flujo de vehículos y salvaguardar la seguridad de los ciudadanos. Para ello, hace falta combinar diversas entradas de dispositivos, incluidas de cámaras de vídeo, así como datos de telemetría de sensores en las vías para optimizar los patrones de tráfico.
- **Supervisión de fábricas:** para los responsables de planta, es de vital importancia poder garantizar el tiempo de actividad de los procesos esenciales y cumplir los programas de producción. Esto se traduce en llevar un mantenimiento predictivo de los equipos fundamentales, detectar de forma automatizada los fallos y optimizar las entradas y salidas en la cadena de suministro del emplazamiento. Se trata de un campo en el que la IA puede ayudar a los operarios a aumentar el rendimiento mientras se respetan los estándares de seguridad.
- **Drones:** el análisis automatizado de las imágenes que capturan los drones permite supervisar una gran variedad de situaciones a gran escala que antes no resultaba posible. Este avance ofrece importantes cambios en la inspección de la infraestructura de las instalaciones de gas y electricidad, las encuestas de las compañías de seguros, las operaciones de búsqueda y rescate, la agricultura de precisión, y la conservación de recursos pesqueros y reservas naturales.
- **Entornos empresariales cotidianos:** los entornos empresariales cotidianos no dejan de mejorar gracias a las herramientas de productividad basadas en IA, como Microsoft Copilot.
- **Energías renovables:** las tecnologías renovables, como los parques eólicos, presas hidroeléctricas y parques solares, requieren supervisión, mantenimiento y recopilación en tiempo real de los datos, que se deben generar y analizar localmente.

ESTACIONES DE TRABAJO DELL PARA IA

Dell ofrece una amplia gama de estaciones de trabajo para diversos niveles de desarrollo y/o implementación de IA, todo ello bajo el paraguas de la marca Data Science Workstation (DSW). En esta sección, se ofrecerá un resumen de las especificaciones y se abordarán las numerosas aplicaciones y perfiles para el uso de IA, como los científicos de datos, además de las ventajas de la tecnología Dell DSW. Estas estaciones de trabajo de ciencia de datos preparadas para IA se han diseñado específicamente para científicos de datos. Las estaciones de trabajo Precision Data Science Workstation más recientes utilizan la funcionalidad de IA para ajustar los dispositivos con el fin de optimizar el rendimiento de las aplicaciones más utilizadas por los científicos de datos. De esta forma, podrán completar el trabajo más importante con mayor rapidez. Además, las estaciones de trabajo Dell Precision se prueban y certifican por parte de ISV independientes con el fin de garantizar que admiten aplicaciones de alto rendimiento, las cuales necesitan los clientes de Dell para completar sus tareas cotidianas.

Aspectos que diferencian las estaciones de trabajo Dell del resto

Las estaciones de trabajo Dell Precision con tecnología de las GPU NVIDIA RTX están diseñadas para ofrecer gran capacidad de ampliación y rendimiento para los procedimientos de análisis y de IA de las organizaciones. Dell Technologies ofrece soluciones de hardware integrales que están optimizadas para ejecutar el software de IA más reciente del sector.

- **Robustez en la configuración del hardware:** las estaciones de trabajo Dell Precision ofrecen una variedad de configuraciones de hardware muy eficaces, entre las que se incluyen los procesadores multinúcleo, la memoria RAM de alta capacidad y las diversas opciones de GPU. Estos componentes proporcionan los recursos informáticos necesarios para realizar tareas de IA, lo que ofrece un entrenamiento y una inferencia eficientes.
- **Capacidad de ampliación y personalización:** las estaciones de trabajo Dell Precision ofrecen capacidad de ampliación y personalización, lo que permite a los usuarios adaptar la configuración de hardware a sus necesidades de IA específicas. Esta flexibilidad garantiza que las estaciones de trabajo se puedan optimizar para las necesidades particulares de las cargas de trabajo de IA.
- **Certificación y optimización:** Dell colabora con NVIDIA para certificar las estaciones de trabajo Precision en lo que respecta a compatibilidad y rendimiento con las GPU de NVIDIA RTX, incluidas las tarjetas gráficas NVIDIA RTX 6000 Ada Generation. Esta certificación garantiza una integración sencilla y un rendimiento optimizado si se utilizan las estaciones de trabajo Dell Precision con las GPU NVIDIA RTX para tareas de IA.
- **Gran capacidad de procesamiento:** las estaciones de trabajo Dell Precision que están equipadas con procesadores Intel ofrecen la potencia informática necesaria para las tareas de IA. Gracias a los procesadores multinúcleo y las altas velocidades de reloj, estas estaciones de trabajo ofrecen el rendimiento necesario para entrenar y ejecutar inferencias en los flujos de trabajo de IA.
- **Soporte a través de software específico y herramientas asociadas:** las estaciones de trabajo Dell Precision vienen precargadas con software y herramientas que admiten el desarrollo y la implementación de IA. Esto incluye pilas de software optimizadas, infraestructura de IA y bibliotecas que aprovechan la eficacia de las GPU NVIDIA RTX, para que los usuarios puedan comenzar proyectos de IA con mayor facilidad.

Además, las tecnologías que se abordan en las siguientes secciones describen otros aspectos importantes en los que las estaciones de trabajo de Dell se diferencian del resto.

Tecnología de memoria fiable

Dell ofrece una tecnología basada en ECC que se denomina Reliable Memory Technology Pro (RMT Pro), diseñada para maximizar el tiempo de actividad. Funciona en combinación con la memoria ECC para detectar y corregir errores de memoria en tiempo real. De acuerdo con Dell, RMT Pro elimina prácticamente los errores de memoria al evitar que se vuelva a acudir a memorias dañadas, incluso aunque DIMM esté en pleno uso. Después de reiniciar el sistema, RTM Pro aislará la zona de la memoria defectuosa y la ocultará al sistema operativo. De esta forma, los científicos de datos y desarrolladores de IA evitarán bloqueos constantes por el uso reincidente de una memoria dañada, lo que supone un gran impulso para la productividad.

Dell Optimizer for Precision

Dell también incluye Dell Optimizer for Precision en la mayoría de las estaciones de trabajo, una solución que ajusta de forma automática los parámetros del sistema para que la estación de trabajo pueda ejecutar diversas aplicaciones comerciales habituales a la mayor velocidad posible. Esto mejora la productividad de los científicos de datos y de los desarrolladores. La herramienta también crea para el departamento de TI informes sobre el rendimiento en tiempo real relativos al uso del procesador, el almacenamiento, la memoria y los gráficos. DOP todavía no se ejecuta en Linux, por lo que es particularmente útil para la implementación de IA, ya que este tipo de implementación se suele realizar con software de código abierto basado en Linux. Dell Optimizer for Precision también ofrece ExpressSign-in, Express Charge (en dispositivos móviles), Intelligent Audio y herramientas de generación de informes y análisis para ayudar a optimizar la estación de trabajo.

DESAFÍOS Y OPORTUNIDADES

Para las empresas

IDC observa una bifurcación en el mercado de IA. Por un lado, las empresas están implementando estrategias de datos para mantenerse competitivas, lo que incluye integrar la IA de forma masiva. A modo de ejemplo, se observa cómo empresas similares han obtenido resultados extraordinarios con ofertas de infraestructura de IA empresarial que se incluyen en la lista de los 100 superordenadores más potentes. Por otro lado, las empresas saben cómo es la realidad de las pequeñas iniciativas de IA, que se prueban en los servidores disponibles del centro de datos o la cloud, por lo general con un presupuesto insuficiente o un hardware poco adecuado.

Para muchas empresas, el primer caso no es relevante y el segundo, desafortunadamente, es lo habitual. Para ellas, la complejidad reside en dotar a los científicos de datos o desarrolladores de IA con las herramientas adecuadas para que puedan entrenar la IA en un plazo razonable sin grandes inversiones económicas en instancias de cloud o servidores de centros de datos acelerados por GPU. IDC opina que estas empresas estarían bien cubiertas si pudieran facilitar a sus científicos y desarrolladores estaciones de trabajo potentes aceleradas por GPU.

Para Dell

Existe en el mercado el concepto equivocado de que el desarrollo e implementación de tecnologías de IA requiere equipos caros de hardware de servidor acelerado, a menudo incluso en un clúster. Es posible que esto sea cierto cuando se trata de algoritmos de IA gigantescos, con miles de millones de parámetros, pero la mayoría de las empresas no desarrollan algoritmos de este tamaño. Usan sus iniciativas de IA de forma práctica, eficaz y factible, y muchas empresas no son conscientes de que estos modelos tan comunes de IA se pueden desarrollar e implementar en estaciones de trabajo. El reto de Dell reside en romper los prejuicios y mostrar al mercado las posibilidades de su cartera de estaciones de trabajo.

Al mismo tiempo, Dell debe asegurarse de que sus estaciones de trabajo ofrecen las prestaciones previstas y no se convierten en un obstáculo tecnológico con el paso del tiempo. Esto requiere una innovación rápida y constante para no defraudar a los usuarios finales que utilizan estas estaciones de trabajo correctamente (es decir, que no intentan ejecutar en ellas un algoritmo con miles de millones de parámetros). Asimismo, aquellos clientes que empiezan a ampliar rápida y repentinamente o tienen algoritmos que están adquiriendo un gran tamaño deben tener en cuenta que es posible realizar una transición sencilla de la estación de trabajo a la línea de servidores de IA de Dell. Evidentemente, es aquí donde reside la oportunidad para Dell, ya que puede ofrecer la solución adecuada para cada cliente, sin importar la magnitud de la iniciativa de IA que deseen abordar.

CONCLUSIÓN

IDC opina que actualmente se subestima el potencial de las estaciones de trabajo como sólidas plataformas para el desarrollo e implementación de tecnologías de IA en muchos casos de uso. Proporcionan a los científicos y desarrolladores de IA una potente plataforma acelerada por GPU que representa un menor gasto de capital (CAPEX) que los servidores, unos gastos de explotación drásticamente menores que las instancias de cloud y una libertad mucho mayor para experimentar con modelos de IA. Las empresas que están desarrollando iniciativas de IA que no necesitan algoritmos de miles de millones de parámetros (como la gran mayoría) deberían plantearse la posibilidad de dotar a sus equipos de IA con estaciones de trabajo para poder desarrollar sin trabas tecnologías de IA y para llevar fácilmente las implementaciones al perímetro.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com