White Paper

# Por que desenvolver e implantar tecnologia de IA em estações de trabalho faz sentido

Sponsored by: Dell Technologies

Peter Rutten Dave McCarthy
July 2023

## OPINIÃO DA IDC

A IA decolou como um importante recurso diferencial em todos os setores, e o hardware necessário para executar a IA está evoluindo rapidamente. O setor de tecnologia está muito concentrado no crescimento exponencial da maioria dos modelos avançados de IA . As discussões são sobre dezenas de bilhões de parâmetros, redução da precisão, expansão da memória, necessidades semelhantes à computação com alto desempenho (HPC) como para treinamento e inferência de IA, e racks de servidores acelerados. Na verdade, essa extraordinária escala de computação de IA é a exceção, especialmente nas empresas.

No momento, muitas empresas trabalham arduamente nas iniciativas de IA, inclusive de IA generativa, que não requerem um supercomputador. De fato, muito do desenvolvimento da IA – e cada vez mais a implementação de IA, especialmente na borda – está ocorrendo em workstations avançadas. As workstations têm diversas vantagens para o desenvolvimento e a implementação da IA. Elas dispensam cientistas ou desenvolvedores de IA da tarefa de negociar o tempo do servidor, fornecem aceleração da GPU, embora as GPUs baseadas em servidor ainda não estejam facilmente disponíveis no datacenter, são extremamente acessíveis e representam um tamanho menor e despesa única, em vez de uma conta que se acumula rapidamente para uma instância de nuvem, , além de existir o conforto de saber que os dados confidenciais estão protegidos com segurança local *(on-premises*). Desse modo, elas também aliviam os cientistas ou desenvolvedores da ansiedade quanto aos custos enquanto estão apenas experimentando modelos de IA.

A IDC presencia um cenário de implementação de IA mais rápido na borda do que local *(on-premises)* ou na nuvem. Aqui também, as workstations desempenham um papel cada vez mais essencial como plataformas de inferência de IA, geralmente sem exigir GPUs, mas realizando a inferência em CPUs otimizadas por software. Os casos de uso de inferência de IA na borda em workstations estão crescendo rapidamente. Como exemplo podemos citar AIOps, resposta a desastres, radiologia, exploração de petróleo e gás, gestão de terras, telemedicina, gestão de tráfego, monitoramento de fábricas e drones.

Este white paper analisa o a crescente importância das workstations no desenvolvimento e na implementação da IA, além de abordar rapidamente o portfólio de workstations para IA da Dell.

## VISÃO GERAL DA SITUAÇÃO

### A explosão da IA e o impacto na infraestrutura

A quantidade de projetos de IA em que as organizações no mundo todo estão engajadas aumenta rapidamente. Em todos os setores, muitas tarefas já são realizadas por software parcial ou totalmente orientado por um modelo de IA. A IDC monitora a IA em muitos níveis, e uma métrica que é útil considerar é o valor que as empresas e os provedores de serviços em nuvem pretendem investir em servidores para desenvolver e executar a IA. Até 2026, isso será equivalente a US$ 34,6 bilhões, o que representa aproximadamente 22% do gasto total em servidores no mundo todo.

No entanto, os servidores não compõem todo o cenário. Está acontecendo muita preparação, desenvolvimento, prototipação e crescente *implementação* da IA nas workstations. À medida que as organizações, sejam elas de pequeno ou grande porte, descobriram que oportunidades de negócios podiam ser percebidas infundindo os aplicativos com um pouco de IA, a experimentação com modelos de IA dispararam, e as workstations robustas passaram a ser ideais para essa finalidade, com sua disponibilidade imediata e proximidade de dados.

Como a IA passou repentinamente a ser tão predominante, visto que os algoritmos de IA são implementados há décadas? Isso ocorre principalmente porque duas condições fundamentais para impulsionar um tipo especificamente bem-sucedido de algoritmo de IA, a rede neural, foram constatadas nos últomos anos: a fácil disponibilidade de um vasto, , barato e diversificado tipo de dados, como dados não estruturados ou semiestruturados, e o aumento da computação linear com um modelo paralelo para processar essas redes neurais dentro de um período aceitável. Com essas duas condições básicas atendidas, os cientistas de dados fizeram avanços tremendos no desenvolvimento de redes neurais que aprendem automaticamente a executar tarefas cada vez mais impressionantes. Embora o aprendizado de máquina (ML) tradicional continue sendo relevante para dados textuais ou numéricos, a aprendizagem profunda (DL-*deep learning*) é mais eficiente para vídeo, áudio, linguagens etc.

Os modelos tradicionais de aprendizado de máquina geralmente podem ser desenvolvidos nas CPUs das workstations, que têm, em sua maioria, dezenas de núcleos. No entanto, as redes neurais exigem coprocessadores para paralelizar o processamento em milhares de núcleos. O principal motivo é: no ML, a extração e a classificação de recursos é um processo manual, enquanto na DL, é automático, exigindo que o modelo seja treinado por meio da constante repetição usando grandes conjuntos de dados. Atualmente, o coprocessador mais comum é a GPU, mas novos processadores específicos para IA desenvolvidos por startups também estão sendo disponibilizados. Esse tipo de aceleração, usando um coprocessador dedicado para processamento em paralelo, revolucionou os mercados de servidores e workstations, dando início ao que a IDC chamada de computação amplamente paralela.

Em 2022, os servidores acelerados constituíam um mercado mundial de US$ 21,8 bilhões, alcançando US$ 43,4 bilhões até 2026, com 57% do total representando servidores acelerados para executar a IA. Ao mesmo tempo, o número de GPUs dedicadas vendidas para uso em workstations aumentou para 6,4 milhões em 2022. A IDC estima que o mercado de workstations usadas para fins científicos ou

de engenharia de software, cada vez mais orientado pelo desenvolvimento da IA, aumentará para aproximadamente US$ 2 bilhões até 2026.

## Estágios do desenvolvimento da IA

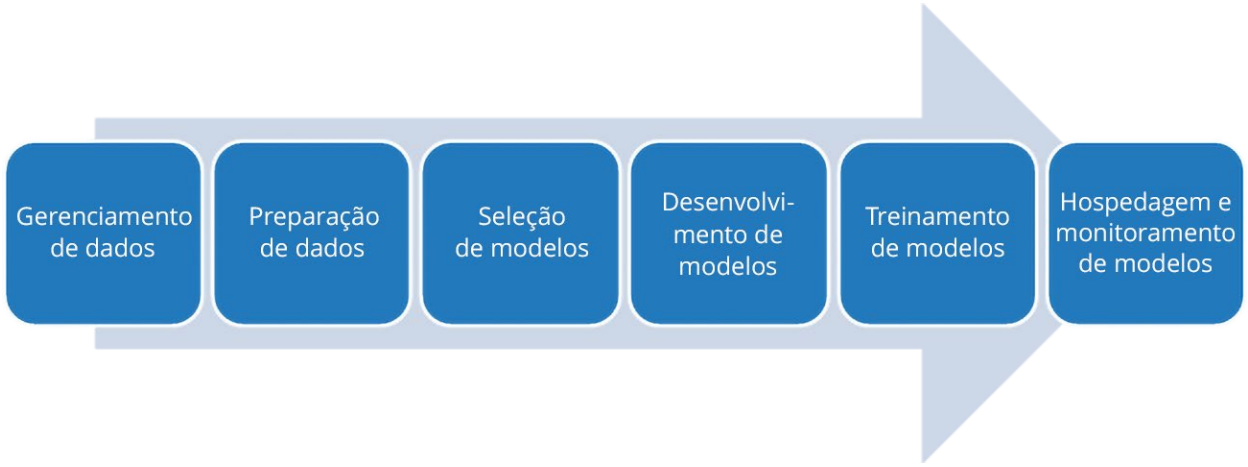
Conforme mencionado, as redes neurais se tornaram viáveis graças aos tipos e volumes de dados em expansão e às novas abordagens para a computação. A primeira parte dessa equação, os volumes e tipos de dados, não é trivial – em algumas contas, 80% do esforço em uma iniciativa de IA de aprendizagem profunda reside no gerenciamento e na preparação de dados. Os dados precisam ser ingeridos, gerenciados e preparados antes do início do design e do treinamento de modelos. De acordo com a IDC, a seguir estão os estágios do desenvolvimento da IA (veja a Figura 1):

- **Gerenciamento de dados:** Identificar e gerenciar os dados relevantes para o modelo de IA com base em grandes volumes de dados no data center, na borda e na nuvem que uma organização ingere, gera e/ou adquire. Esses dados podem ser de qualquer tipo, orientados a eventos ou de streaming, e muitos deles podem exigir algum tipo de governança.
- **Preparação de dados:** Armazenar os dados (arquivo, bloco ou objeto) em um repositório de dados ou data lake, limpá-los, garantindo que estejam completos e sejam de alta qualidade, e transformá-los em uma forma que possa ser usada no modelo de IA – por exemplo, com Spark ou ferramentas como Pandas
- **Seleção de modelos:** Decidir qual modelo tem o desempenho ideal da tarefa de IA para o qual está programado em termos de taxa de erro e/ou desempenho
- **Desenvolvimento de modelos:** Criar o modelo de IA usando estruturas, como XGBoost, LightGBM, GLM, Keras, TensorFlow, PyTorch, Caffe, RuleFit, FTRL, Snap ML, scikit-learn ou H2O
- **Treinamento de modelos:** Treinar o modelo na infraestrutura de computação com núcleos de processador e/ou coprocessador suficientes para paralelização (aumentando também, cada vez mais, a capacidade de explicar, validar e documentar as decisões de um modelo para garantir a legitimidade, a responsabilidade e a transparência). Isso inclui a prototipação – testar o modelo treinado executando a inferência nele.
- **Hospedagem e monitoramento de modelos:** Implementar o modelo em um ambiente de produção para executar a tarefa para a qual ele foi desenvolvido, algo normalmente chamado de “inferência de IA”, e monitorar o desempenho

As workstations podem desempenhar um importante papel nesses seis estágios em combinação com a infraestrutura de datacenter, nuvem ou borda.

## FIGURA 1

**Estágios do desenvolvimento da IA**




Fonte: IDC, 2023

## DESENVOLVENDO MODELOS DE IA EM WORKSTATIONS

### Workstations equiparadas aos computadores pessoais

Geralmente, compreende-se que os computadores pessoais (PCs) não são poderosos o suficiente para o desenvolvimento da IA. Os cientistas de dados e os desenvolvedores de IA geralmente se envolvem em projetos estrategicamente importantes para as organizações, e a produtividade sem impedimentos é de extrema importância. As workstations tendem a ter um desempenho mais previsível do que os PCs, pois geralmente são criadas com componentes de melhor desempenho e otimizados para o software que executam.

Esses componentes incluem:

- **Processadores de alto nível:** Um exemplo são os processadores Intel Xeon Scalable.
- **GPUs poderosas:** Um exemplo são as GPUs profissionais RTX da NVIDIA, como NVIDIA RTX 6000 Ada.
- **Mais armazenamento:** Algumas workstations podem oferecer 60 TB, e as velocidades de I/S tendem a ser significativamente mais altas do que em PCs.
- **Mais memória:** Já existem workstations com mais de 6 TB de memória.
- **Resfriamento:** Componentes de alto-desempenho geram muito calor, e os cientistas de dados precisam de uma workstation com resfriamento adequado para evitar superaquecimento e manter o desempenho ideal.
- **Placa de interface de rede (NIC):** Para cientistas de dados que trabalham com grandes conjuntos de dados armazenados em servidores remotos, uma placa de interface de rede de alta velocidade é essencial para transferir dados rapidamente e com eficiência.
- **Monitor:** Um monitor de alta qualidade é importante para tarefas de visualização de dados, e os cientistas de dados devem buscar um monitor com alta resolução, precisão de cores e grande tamanho de tela.

- **Memória de Código de Correção de Erros (ECC):** O ECC detecta e corrige os tipos mais comuns de corrupção de dados internos, evitando telas azuis durante um longo treinamento de IA causadas por um erro de hardware (bit incorreto) ou erro de software (bit trocado, causando valores incorretos). O ECC também garante a precisão de resultados, um requisito crítico para trabalhos essenciais à vida, como cuidados de saúde.
- **Chip especializado:** Um exemplo são as Unidades de processamento visual (VPUs) Intel Movidius, que são coprocessadores de processamento paralelo para aplicativos de IA de visão computacional e borda, usados em configurações de varejo, segurança e automação industrial, por exemplo. FPGAs também são usadas em workstations para aplicações financeiras, por exemplo.
- **Software de otimização:** Exemplos incluem a OneAPI, que é o modelo de programação baseado em padrões da Intel para simplificar o desenvolvimento e a implementação de cargas de trabalho centradas em dados em CPUs, GPUs, FPGAs e outros aceleradores, ou CUDA, que é a plataforma de computação paralela e a interface de programação de aplicativos da NVIDIA para executar cargas de trabalho gerais em GPUs.

### *CPUs versus GPUs para IA*

As workstations podem ser usadas em vários estágios do desenvolvimento da IA e geralmente são equipadas para diversas capacidades. Apesar da ênfase nas GPUs para o processamento paralelo, as CPUs desempenham um papel crítico ao desenvolver um modelo de IA em uma workstation. Assim como as GPUs, as CPUs também podem ser usadas para manipulação de dados e, naturalmente, para desenvolver modelos tradicionais de ML. As CPUs também são usadas para exploração de dados – o processo de usar representações visuais de um conjunto de dados para entender as características deles.

No treinamento da DL, a função da CPU do host é ligeiramente reduzida à medida que as GPUs assumem o controle durante o processo real de treinamento, mas, mesmo assim, as CPUs continuam servindo como a camada de processamento para softwares essenciais, como sistema operacional ou CUDA, e para processos de orquestração entre as GPUs ou com outros chips. Além disso, as CPUs assumem cada vez mais um novo papel como mecanismos de inferência de IA nos casos em que uma workstation é usada para executar um modelo de IA em produção. A IDC espera que, até 2024, os gastos com a infraestrutura para a inferência de IA ultrapassem os gastos com a infraestrutura de IA para treinamento e que uma parte importante (39%) dessa inferência ocorra nas CPUs host.

## Workstations equiparadas a servidores: um relacionamento simbiótico

Para a maioria das organizações, pragmatismo é a regra de ouro para quando uma workstation, um servidor no local, uma instância de nuvem ou qualquer combinação desses três é implementada para o desenvolvimento da IA. Há um relacionamento simbiótico entre workstations, servidores e instâncias de nuvem para diferentes estágios do desenvolvimento de um projeto de IA.

A vantagem das workstations versus servidores de data center é que os cientistas de dados podem trabalhar de onde querem – um fator importante na pandemia recente, mas também em circunstâncias normais. Eles também podem testar livremente seus modelos de IA, iterando-os sempre que considerarem necessário, pois o poder das workstations modernas, com GPUs poderosas, geralmente permite que o processo iterativo seja mais interativo, oferecendo feedbacks e resultados instantâneos, sem precisar solicitar acesso a servidores ou encontrar outras restrições de data center. E as workstations proporcionam a flexibilidade de aproximar o computador dos dados e não o contrário, economizando largura de banda, reduzindo o congestionamento da rede e aumentando o throughput. Além disso, as workstations podem ser configuradas para diferentes necessidades: tarefas tradicionais de ML, por exemplo, ou tarefas mais intensas de DL.

Embora tenha ocorrido um crescimento significativo no mercado de servidores acelerados, eles ainda não estão amplamente disponíveis nos data centers empresariais. No momento em que este white paper foi escrito, em média, 4% dos servidores nos data centers empresariais foram acelerados, o que significa que muitas organizações não conseguem desenvolver ou executar a IA em GPUs no local já disponíveis. Também por esse motivo, workstations aceleradas são uma alternativa útil para o desenvolvimento da IA.

Workstations altamente aceleradas agora são tão potentes que conseguem realizar o treinamento de DL, desde que o modelo de IA não seja excessivamente grande, eliminando a necessidade de treinamento nos servidores. E os modelos treinados em workstations com GPUs podem ser implementados em workstations ou em servidores sem GPUs, impulsionando os recursos de inferência nas CPUs. As tecnologias de software, como DL Boost e oneAPI da Intel, podem impulsionar a inferência de IA na CPU, permitindo que servidores não acelerados e já implementados nos data centers suportem os aplicativos de IA.

## Workstations equiparadas à nuvem

A cloud computing revolucionou o modo como as organizações pensam em infraestrutura, dados e aplicativos. Com a promessa de uma escalabilidade quase sem limites, a nuvem permite que os desenvolvedores provisionem recursos sob demanda, acelerando potencialmente o ritmo da inovação com menos restrições. Ao pé da letra, a nuvem parece ser o paradigma perfeito para o desenvolvimento da IA.

No entanto, esse não é sempre o caso. Na verdade, a pesquisa da IDC mostrou que as organizações cada vez mais repatriam determinadas cargas de trabalho da nuvem pública para infraestrutura local *(on-premises)*. Isso ocorre por diversos fatores:

- **Disponibilidade da nuvem:** Quem já contou com os serviços em nuvem já passou por um episódio de interrupção, seja por problemas no provedor de serviços em nuvem em si, seja por um lapso na conectividade de rede entre o data center de hiperescala e o usuário final. Nessas situações, os usuários ficam à mercê do provedor de serviço para resolver o problema, enquanto a produtividade fica paralisada.
- **Segurança e conformidade:** Em muitos setores, as políticas de governança corporativa ditam onde os dados podem ser comunicados e armazenados, o que limita o uso dos serviços em nuvem. As regulamentações de governo, como GDPR na Europa e a Lei de Privacidade do Consumidor da Califórnia (CCPA), também reforçam as regras sobre a soberania dos dados.
- **Custo:** É comum que as organizações subestimem a velocidade na qual as taxas dos serviços em nuvem podem aumentar, especialmente para cargas de trabalho que exijam recursos de computação de alto desempenho e grandes quantidades de armazenamento. A economia da nuvem se baseia em medir todos os tipos de consumo de recursos, inclusive saída de dados de volta para a infraestrutura local.
- **Pressão de teste e erro:** A maioria das iniciativas de IA iniciam com uma quantidade significativa de testes, em que os modelos que falham fazem parte do processo de desenvolvimento. Nesse processo, os cientistas e desenvolvedores de IA pagam uma "taxa psicológica" quando as contas da nuvem se acumulam sem que eles possam mostrar resultados executáveis.

As workstations podem atender a essas limitações enquanto ainda utilizam as tecnologias nativas da nuvem, como arquiteturas baseadas em microsserviços e automação orientada à API. Isso possibilita alguns dos benefícios no que se refere à comparação de workstations com servidores de data center:

- **Trabalhe de qualquer lugar:** Ao remover a dependência da nuvem pública, cenários desconectados agora são possíveis. Muitos ambientes de alta segurança têm uma lacuna de ar de redes públicas, e as workstations de IA podem atender exclusivamente a essa necessidade. Os recursos locais também reduzem a demanda por uma conectividade de rede de alto custo.
- **Localidade dos dados:** A proliferação dos dispositivos de IoT e outros equipamentos conectados contribui para um crescimento exponencial dos dados em locais de borda. Em muitas situações, faz sentido colocar os recursos de computação junto a uma workstation dedicada. Isso também resolve muitos requisitos de conformidade, limitando o movimento dos dados.
- **Teste livremente:** O treinamento e a otimização dos modelos de IA é um processo iterativo, que geralmente inclui alguns elementos de teste e erro. Os desenvolvedores precisam da liberdade de conduzir testes sem fazer concessões, devido ao potencial de taxas de serviço adicionais. As workstations também oferecem mais flexibilidade para ferramentas personalizadas.

A respeito do último ponto, comparar o preço de uma workstation com uma implementação de nuvem é relativamente fácil, pois a maioria dos provedores de serviços em nuvem fornecerão estimativas de custo instantâneas de qualquer configuração que um usuário final deseje implementar. Por exemplo, o custo de uma única máquina virtual (VM) regular com uma NVIDIA T4 e uma instância de armazenamento SSD de 375 GiB usada oito horas por dia, cinco dias por semana, é de US$ 140 em um grande provedor de serviços em nuvem. Com o dobro de VMs, T4s e SSDs, o custo vai para US$ 365 por mês. Continue com duas VMs, mas dobrando as T4s para quatro e o armazenamento para quatro de 375 GiB e faça um treinamento em período integral no ambiente e o custo sobe para US$ 2.700 por mês. Desse modo, é razoável afirmar que os custos da nuvem para o desenvolvimento da IA podem aumentar de maneira vertiginosa para dezenas de milhares de dólares ao ano, substancialmente mais do que a depreciação anual de uma workstation de última geração.

## PROTOTIPANDO A IA EM WORKSTATIONS

Em comparação com os servidores no local e a nuvem, as workstations fornecem uma vantagem distinta no que se refere à prototipação de modelos de IA. Os servidores no data center podem estar sendo totalmente utilizados ou serem muito essenciais para a prototipação e os testes de IA e, conforme já discutido, as instâncias de nuvem podem levar rapidamente a um excesso de custos quando usadas de modo liberal como ambiente de teste. Com as workstations, o cientista ou desenvolvedor de IA fica livre de precisar negociar acesso ao servidor ou da incômoda preocupação de acumular contas da nuvem durante o estágio de prototipação. Os custos únicos e baixos oferecem a liberdade de prototipar de qualquer lugar e a qualquer momento, sem custos adicionais.

## IMPLEMENTANDO MODELOS DE IA EM WORKSTATIONS

Embora o desenvolvimento de modelos de IA em uma workstation tenha sido uma estratégia comum há anos, a IDC vê um aumento dos casos de uso para *implementar* um modelo de IA em uma workstation, normalmente na borda — em outras palavras, colocar o modelo de IA em produção na workstation executando a inferência no modelo de IA. A borda está crescendo com rapidez como um local de implementação de IA para servidores — mais do que o triplo de 2020 a 2024 em termos de gastos anuais de hardware — e as workstations não ficam muito atrás, já que os usuários finais descobriram suas vantagens na borda.

A IDC define a borda como um paradigma de computação distribuído que inclui a implementação de infraestrutura e aplicativos fora da nuvem centralizada e em data center no local. Isso inclui escritórios remotos e filiais, bem como locais específicos do setor, como fábricas, depósitos, hospitais e lojas de varejo.

As cargas de trabalho com muitos dados e computação estão sendo cada vez mais implementadas no local ou em locais de borda. Isso é feito para diminuir as limitações inerentes às nuvens públicas, como o tempo necessário para o upload de grandes quantidades de dados e os custos variáveis de conduzir o treinamento da IA, especialmente em situações que exigem uma quantidade significativa de experimentação de ciência de dados.

A pesquisa da IDC mostra que a borda é um cenário de implementação de IA que cresce rapidamente, com as organizações investindo US$ 2,9 bilhões em computação de IA na borda em 2023, com um aumento para US$ 6,9 bilhões em 2026 (consulte *Worldwide AI Hardware Forecast, 2022-2026: Strong Market Growth for AI Compute and Storage,* IDC nº US49671722, de setembro de 2022). Além disso, a borda está ganhando tração como opção de implementação para cargas de trabalho de HPC, como engenharia e área técnica, com as empresas investindo atualmente cerca de US$ 1 bilhão nessas cargas de trabalho na borda, com um aumento para US$ 2,4 bilhões até 2027 (consulte *Worldwide High-Performance Computing Server Forecast, 2023-2027: Enterprise Will Overtake HPC Labs,* IDC nº US50525123, de abril de 2023). Nessas áreas, faz sentido implementar uma workstation de IA.

Ao implementar um modelo de IA em uma workstation na borda, nem sempre é necessário ter GPUs de última geração, como no caso do desenvolvimento de IA. GPUs mais simples podem realizar a inferência de IA e, em poucos casos, as GPUs nem são necessárias. Nessas situações, as CPUs podem realizar corretamente a tarefa de inferência, em especial quando usadas com otimizações, como o Intel DL Boost, um conjunto de recursos de instrução nos microprocessadores Intel criados para acelerar as cargas de trabalho da IA, inclusive inferência de IA. Com o Intel DL Boost, a Intel afirma que observou um aumento de 1,45 vez na saída de inferência em tempo real de INT8 com o processadores Intel Xeon Scalable de 4ª geração oferecendo Intel DL Boost versus a geração anterior (BERT-Large SQuAD). Isso também ajuda a tornar a workstation mais adequada para implementação na borda, onde as considerações como energia, mobilidade e gerenciamento térmico demandam tensões mais baixas. O Intel Movidius Myriad (M2) se enquadra bem nesse range de consumo de energia  de alimentação graças ao pequeno consumo energético  de 12 W.

### Casos de uso para implementar a IA em workstations

Há diversas situações que naturalmente se direcionam para a implementação da IA em workstations implementadas no local *(on-premises).* Características comuns são grandes volumes de dados de séries temporais gerados por máquina e dados não estruturados, como streaming de vídeo e imagens.

Há também situações em que os especialistas no assunto aumentam os modelos de IA com interpretação humana.

Exemplos:

- **AIOps:** À medida que os sistemas de TI aumentam em escala e complexidade, há uma maior necessidade de migrar de um gerenciamento reativo de incidentes para um monitoramento proativo. Isso vale especialmente para infraestruturas e aplicativos que são distribuídos para locais de borda em que há uma equipe técnica pequena ou nenhuma. Modelando uma base de desempenho normal, é possível identificar anomalias e automatizar as etapas de remediação.
- **Resposta a desastres:** Em uma emergência, os socorristas podem avaliar uma situação rapidamente, acompanhar o equipamento crítico e implantar recursos para ajudar os mais necessitados. Isso geralmente acontece em um ambiente sem conectividade de rede, necessitando de uma workstation local que possa agregar feeds de dados, inferir modelos de IA e automatizar comunicações para equipes importantes.
- **Radiologia:** Os avanços na tecnologia de imagem levaram a um aumento no tamanho dos dados gerados em um único scan, o que exige que eles continuem no local para poderem ser analisados rapidamente. Modelos de IA treinados com milhares de exemplos anteriores podem identificar padrões de modo mais preciso do que o olho humano, aumentando as taxas de precisão.
- **Exploração de petróleo e gás:** Importantes empresas de petróleo e gás usam uma combinação de dados sísmicos, de telemetria e de imagens para localizar reservas de recursos naturais, selecionar locais de perfuração e otimizar o desempenho de equipamentos no processo de produção. Isso normalmente requer a análise de informações em áreas em que somente a comunicação cara com satélites está disponível.
- **Pesquisa sobre câncer e desenvolvimento de medicamentos:** Pesquisadores em hospitais e centros de pesquisa usam a IA e o processamento de linguagem natural para ajudar os oncologistas a determinar o tratamento mais eficaz e individualizado para seus pacientes com câncer. Eles combinam o aprendizado de máquina com a visão do computador para oferecer aos radiologistas um melhor entendimento de como o tumor dos pacientes está progredindo. Também usam algoritmos para entender melhor como o câncer se desenvolve e quais tratamentos funcionam melhor para combatê-lo.
- **Avaliações de solicitação de seguros:** O processamento manual de pedidos exige muito trabalho e está propenso a erro humano. A IA que pode avaliar a validade das solicitações e reduzir os custos, permitindo que os reguladores de seguros se concentrem em casos que precisam de maior investigação. Isso aumenta o resultado geral da operação sem sacrificar a precisão.
- **Telemedicina:** A IA está aumentando as taxas de recuperação de pacientes, personalizando planos de tratamento individuais com base nos sinais vitais em tempo real provenientes de dispositivos vestíveis. Essas informações se combinam a registros do histórico de pacientes e a uma base de conhecimento de casos semelhantes. Isso é especialmente importante para áreas rurais que contam mais com o sistema de saúde remoto.
- **Segurança em varejo (antirroubo):** A lógica analítica em tempo real aplicada a streamings de vídeo está sendo usada para prever o comportamento humano que pode levar a atividades criminosas. Isso normalmente requer costurar vários feeds de vídeo para acompanhar os movimentos de uma pessoa dentro de uma loja. Dada a natureza impreterível da identificação de um evento material, esse processo funciona melhor localmente.
- **Gerenciamento de tráfego:** As entidades governamentais que são responsáveis pelas operações de transporte usam cada vez mais a IA para coordenar os semáforos e a sinalização

digital para melhorar o fluxo de veículos e manter a segurança dos cidadãos. Isso requer uma combinação de entradas, inclusive videocâmeras e telemetria para sensores em estradas de modo a otimizar os padrões do trânsito.

- **Monitoramento na manufatura:**Garantir o tempo ativo de processos críticos e o andamento dos cronogramas de produção são tarefas primordiais de um gerente de manufatura. Isso se traduz na manutenção preditiva dos principais equipamentos, na detecção automática de defeitos e na otimização da cadeia de suprimentos dentro e fora do local. Essa é uma área em que a IA pode ajudar operadores humanos a aumentar o desempenho, mantendo os padrões de segurança.
- **Drones:** A análise automática de imagens capturadas por drones permite monitorar diversas condições em grande escala, como nunca antes foi possível. Isso tem um impacto importante na inspeção de infraestrutura de utilitários de gás e eletricidade, pesquisas de seguros, esforços de busca e salvamento, agricultura de precisão e manutenção da preservação na pesca e na vida selvagem.
- **Ambientes de escritório:** Os ambientes cotidianos de escritório estão sendo cada vez mais aperfeiçoados com as ferramentas de produtividade baseadas em IA, como o Microsoft Copilot.
- **Energia renovável:** Locais de energia renovável, como parques de energia eólica, usinas hidrelétricas e parques de energia solar, exigem monitoramento, manutenção e coleta de dados em tempo real, que devem ser gerados e analisados localmente.

## WORKSTATIONS DELL PARA IA

A Dell oferece diversas workstations para vários níveis do desenvolvimento e/ou da implementação da IA, tudo com respaldo da marca Data Science Workstation (DSW). Esta seção descreve brevemente as especificações e fala sobre as diversas personas/aplicações da IA, como cientistas de dados, bem como os benefícios da tecnologia DSW da Dell. As workstations de ciências de dados prontas para IA foram desenvolvidas especificamente para cientistas de dados. As mais recentes workstations de ciência de dados Precision utilizam a funcionalidade da IA para ajustar os dispositivos para o desempenho otimizado dos aplicativos que os cientistas de dados mais usam. Isso permite realizar o trabalho mais importante de modo mais rápido. Além disso, as workstations Dell Precision são testadas e certificadas por ISVs independentes para garantir o suporte a aplicativos de alto desempenho necessário para que os clientes da Dell realizem suas tarefas do dia a dia.

### Como as workstations da Dell se destacam

As workstations Dell Precision com GPUs NVIDIA RTX foram projetadas para fornecer forte escalabilidade e desempenho para as iniciativas de lógica analítica e IA da organização. A Dell Technologies oferece soluções de hardware completas e otimizadas para executar o software de IA mais recente do setor:

- **Configuração de hardware robusta:** As workstations Dell Precision oferecem diversas configurações poderosas de hardware, inclusive processadores multi-core, RAM de alta capacidade e várias opções de GPU. Esses componentes contam com os recursos computacionais necessários para tarefas de IA, permitindo o treinamento e a inferência eficientes.
- **Escalabilidade e capacidade de personalização:** As workstations Dell Precision são escaláveis e podem ser personalizadas, permitindo que os usuários adaptem a configuração do hardware

de acordo com seus requisitos específicos de IA. Essa flexibilidade garante que a workstation possa ser otimizada para as necessidades específicas das cargas de trabalho de IA.

- **Certificação e otimização:** A Dell colabora com a NVIDIA para certificar as workstations Precision quanto à compatibilidade e ao desempenho com GPUs NVIDIA RTX, inclusive placas NVIDIA RTX 6000 Ada Generation. Essa certificação garante a integração perfeita e o desempenho otimizado ao usar as workstations Dell Precision com GPUs NVIDIA RTX para tarefas de IA.
- **Poderosa capacidade de processamento:** As workstations Dell Precision equipadas com processadores Intel fornecem a potência necessária para tarefas de IA. Com os processadores multi-core e clock de alta velocidade, as workstations entregam o desempenho necessário para treinamento e inferência em fluxos de trabalho de IA.
- **Suporte de software e ferramentas:** As workstations Dell Precision vêm pré-carregadas com software e ferramentas que suportam o desenvolvimento e a implementação da IA. Isso inclui pilhas de software otimizadas, estruturas de IA e bibliotecas que aproveitam as GPUs NVIDIA RTX, facilitando os primeiros passos dos usuários em projetos de IA.

Além disso, as tecnologias abordadas nas seções que se seguem são outras áreas importantes em que as workstations Dell se destacam.

### *Tecnologia confiável de memória*

A Dell fornece uma tecnologia além do ECC chamada Reliable Memory Technology Pro (RMT Pro), desenvolvida para ajudar a maximizar o tempo ativo. Ela funciona em conjunto com a memória ECC para detectar e corrigir erros de memória em tempo real. De acordo com a Dell, a RTM Pro quase que elimina os erros de memória, impedindo que uma memória ruim seja revisitada novamente, mesmo quando o DIMM continua em uso total. Após a reinicialização do sistema a RTM Pro isola a área de memória com defeito e a oculta do sistema operacional. Como resultado, os cientistas de dados e desenvolvedores de IA não têm o problema de falhas contínuas, pois a memória ruim continua sendo endereçável – um importante impulso para a produtividade.

### *Dell Optimizer for Precision*

A Dell também inclui o Dell Optimizer for Precision na maioria de suas workstations, que ajusta automaticamente as configurações do sistema para que a workstation execute vários aplicativos comerciais populares com a máxima velocidade possível. Isso aumenta a produtividade dos cientistas de dados ou desenvolvedores. A ferramenta também cria relatórios de desempenho em tempo real para a TI sobre o processador, oarmazenamento, amemória e a utilização gráfica. O DOP ainda não funciona no Linux e, portanto, é mais útil para implementar IA, visto que desenvolver a IA tende a ser uma tarefa realizada em um software de código aberto baseado em Linux. O Dell Optimizer for Precision também fornece os recursos ExpressSign-in, Express Charge (em dispositivos móveis), Intelligent Audio e ferramentas de relatórios e lógica analítica para ajudar a ajustar a workstation.

## DESAFIOS/OPORTUNIDADES

## Para as empresas

A IDC está notando uma bifurcação no mercado de IA. Por um lado, as empresas estão implementando estratégias de dados para continuarem competitivas, incluindo integração ampla da IA. Como exemplo, elas são apresentadas com parceiros que fizeram um trabalho extraordinário, usando as ofertas de infraestrutura de IA da empresa que são realmente registradas nos 100 melhores supercomputadores.

Por outro lado, as empresas percebem a realidade diária das pequenas iniciativas de IA sendo testadas em servidores disponíveis no datacenter ou na nuvem, geralmente com orçamento insuficiente e hardware de baixo desempenho.

Para muitas empresas, o primeiro cenário não é relevante, e o segundo, muito real. Para elas, o desafio é oferecer aos cientistas de dados e/ou desenvolvedores de IA as ferramentas certas para realizar o treinamento da IA de modo rápido, sem gastar muito dinheiro em instâncias de nuvem ou servidores de data center acelerados por GPU. A IDC acredita que essas empresas serão bem atendidas ao fornecer aos cientistas e desenvolvedores workstations poderosas aceleradas por GPU.

## Para a Dell

Há um mal entendido no mercado de que o desenvolvimento e a implementação da IA requer hardware caro e de servidor acelerado, muitas vezes até um cluster. Isso pode ser válido para os maiores algoritmos, com bilhões de parâmetros, mas a maioria das empresas não desenvolve esses algoritmos grandes. Elas estão fazendo algo útil, impactante e gerenciável com a iniciativa de IA, e muitas empresas não percebem que esses modelos de IA em escala comum podem ser desenvolvidos – e implementados – em workstations. O desafio da Dell é quebrar o preconceito e orientar o mercado sobre as possibilidades com seu portfólio de workstations.

Ao mesmo tempo, a Dell deve garantir que as workstations solucionem gargalos de tecnologia ao longo do tempo, em vez de se tornarem um. Isso significa uma inovação contínua e rápida de modo a nunca desapontar os usuários finais que estão usando as workstations de modo adequado (em outras palavras, que não estão tentando executar um algoritmo de parâmetros multibilionário). Isso também significa que, para os clientes estão começando a escalar muito rapidamente ou para aqueles cujos algoritmos já estão ficando muito grandes, há uma transição perfeita da workstation para a linha de servidores de IA da Dell. Nisto, naturalmente, também está a oportunidade para a Dell: ter a solução certa para cada cliente, independentemente do tamanho da iniciativa de IA em que estão trabalhando.

## CONCLUSÃO

A IDC acredita que as workstations estão sendo atualmente subestimadas como a potência de desenvolvimento e implementação de IA para muitos casos de uso. Elas proporcionam aos cientistas e desenvolvedores de IA uma plataforma acelerada por GPU poderosa que representa menor Capex em comparação com os servidores, reduzindo drasticamente o Opex em comparação com as instâncias de nuvem e com muito mais liberdade para experimentar os modelos de IA. As empresas que estão desenvolvendo iniciativas de IA que não exigem algoritmos com bilhões de parâmetros (e a maioria não faz isso) devem considerar fortalecer suas equipes de IA com workstations para o desenvolvimento da IA sem restrições e fácil implementação na borda.

## 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

---