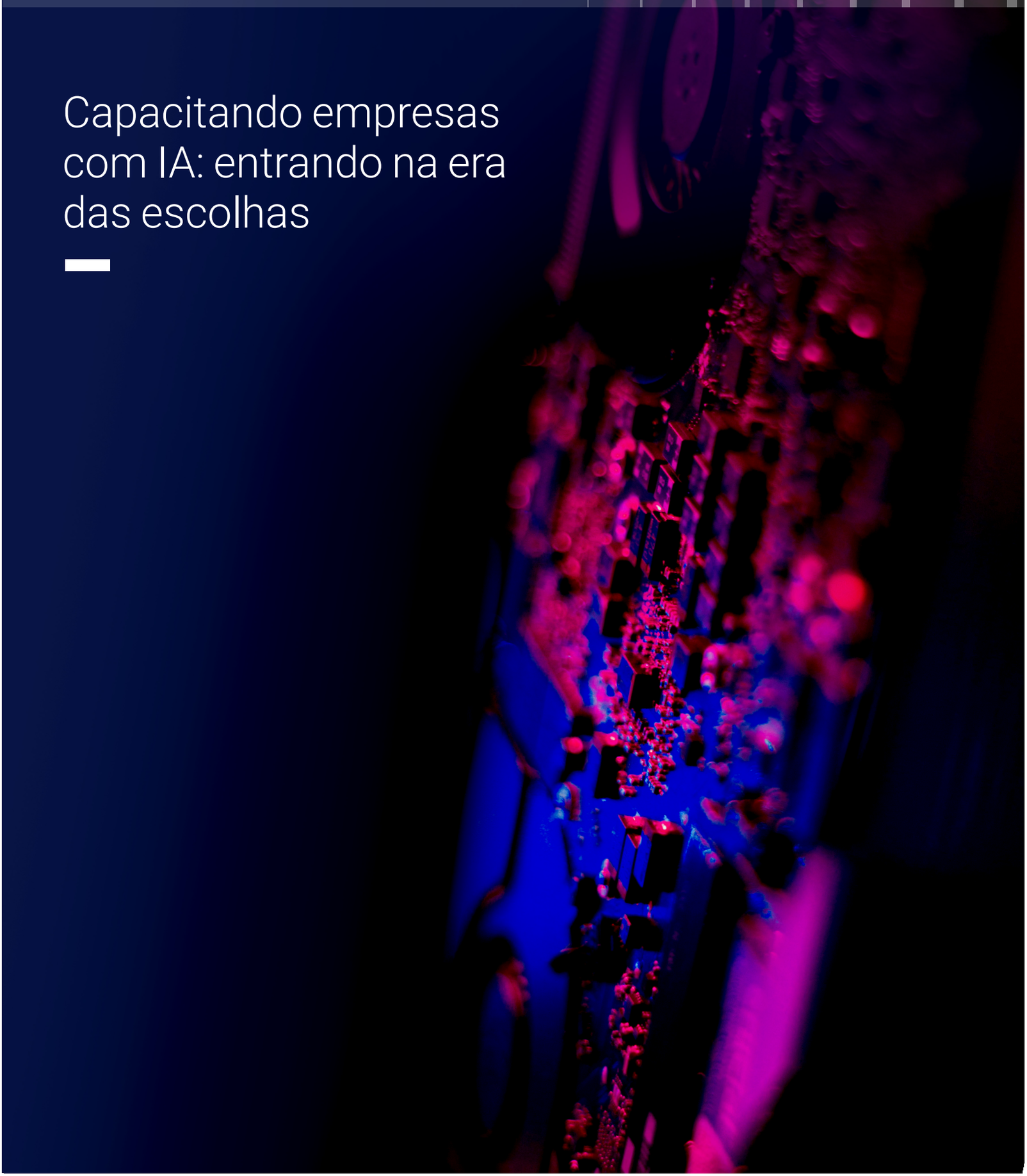
DELL Technologies

AMD
together we advance_

2024 | White paper

Capacitando empresas com IA: entrando na era das escolhas



Sumário

DELL Technologies

AMD
together we advance_

2024 | White paper

A oportunidade de transformar os setores com IA

Hoje, graças à IA, não há uma maior oportunidade de transformar sua empresa para o futuro da inovação. Dados coletados da Accenture Vision Technology 2023 mostram que 98% dos executivos globais concordam que os modelos básicos de IA desempenharão um papel importante nas estratégias de sua organização nos próximos três a cinco anos.[1]

A IA se tornou incrivelmente útil para empresas de áreas como varejo, saúde e serviços financeiros devido à sua capacidade de aumentar a eficiência das tarefas, impulsionar a inovação e melhorar os processos de tomada de decisões. No entanto, apesar das vantagens, devido a alguns equívocos comuns, ainda se nota uma barreira à entrada quando se trata de integrar a IA.

Para começar, é preciso ter uma equipe de desenvolvedores de IA:

Embora o conhecimento especializado em ciência de dados ainda seja valioso para desenvolver soluções avançadas de IA e entender os princípios subjacentes, ele não é mais um pré-requisito. Houve uma proliferação de ferramentas de IA fáceis de usar, plataformas como Hugging Face e modelos específicos para tarefas que abstraem grande parte da complexidade envolvida no desenvolvimento de soluções de IA.

É necessário gastar dezenas de milhões em hardware para obter resultados:

Esse equívoco prejudica gravemente a diversidade dos recursos de IA disponíveis atualmente. Embora, muitas vezes, esses recursos bastante conhecidos sejam avançados e recebam bastante suporte, eles nem sempre são a escolha mais adequada ou econômica para todas as empresas.

É necessário trabalhar incansavelmente para adquirir aceleradores:

Embora os aceleradores se destaquem em cargas de trabalho com uso intenso de IA, as empresas podem não precisar de tanta capacidade de computação para aplicativos de IA. E, simplesmente, esperar um período excessivamente longo para ter acesso aos principais aceleradores do mercado também não é uma atitude realista. Em muitos casos, as CPUs otimizadas para IA podem de fato entregar o desempenho e a eficiência necessários para produzir análises e decisões assistidas por IA em tempo real, além de serem uma solução muito mais econômica e adaptável.

[1] Accenture, 30 de março de 2023, "Accenture Technology Vision 2023: Generative AI to Usher in a Bold New Future for Business, Merging Physical and Digital Worlds", https://newsroom.accenture.com/news/2023/accenture-technology-vision-2023-generative-ai-to-usher-in-a-bold-new-future-for-business-merging-physical-and-digital-worlds

Felizmente, o ambiente de IA está evoluindo. Juntas, a **Dell** e a **AMD** criaram uma parceria para desmistificar tudo isso, tornando as tecnologias e ferramentas de IA acessíveis a uma maior variedade de usuários, com uma infraestrutura completa projetada para atender às demandas atuais de IA.

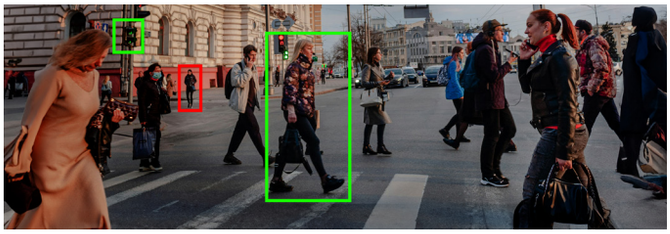
É possível começar com um modelo já otimizado, uma pilha de software confiável e um sistema de hardware versátil, tudo disponível abertamente por meio da parceria entre a Dell e a AMD. Ter acesso a aceleradores cada vez mais escassos, a um grupo substancial de engenheiros de IA qualificados ou a recursos para implementar enormes clusters em nuvem não é mais um requisito para aproveitar a IA.

A colaboração entre a **Dell** e a **AMD** oferece um ecossistema unificado de hardware e software, projetado para permitir que os desenvolvedores criem soluções completas de IA que incorporam o aprendizado por transferência, o ajuste fino e a inferência de maneira fácil e eficiente. Com o suporte da **Hugging Face**, agora, nós temos um portfólio crescente de modelos executados em servidores Dell PowerEdge com processadores AMD EPYC™ ou aceleradores AMD Instinct™ MI300X para que os desenvolvedores possam fazer o ajuste fino, aplicar o aprendizado por transferência e implementá-lo para fins de inferência. Os investimentos em AMD ROCm™ e AMD ZenDNN™, bem como as parcerias com as estruturas PyTorch, Tensorflow e ONNX Runtime, são os ativadores fundamentais dos desenvolvedores de IA aplicada que experimentam a democratização da IA. O diagrama de pilha abaixo detalha os componentes que compõem o ecossistema unificado de IA da Dell e da AMD.

IA nos setores

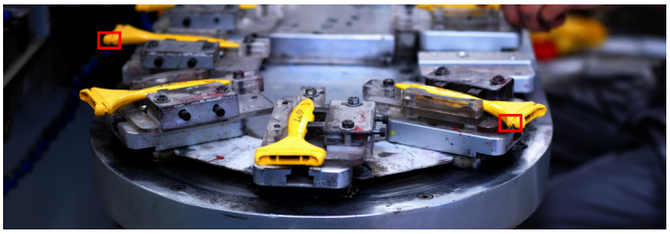
Com a diversificação de recursos e a ênfase na inovação de código aberto, a IA está migrando para muitos setores diferentes, inclusive atendimento ao cliente, finanças e serviços bancários, área da saúde e varejo, entre outros. No entanto, em todos esses setores, a IA permite coletivamente que as organizações desbloqueiem o potencial de seus próprios dados exclusivos e reinventem seus fluxos de trabalho de IA abordando os seguintes recursos principais: análise de dados, automação, personalização e análise preditiva. Além disso, as bibliotecas AMD ROCm e ZenDNN aceleram esses fluxos de trabalho de IA para entregar resultados quase em tempo real.

Veja abaixo exatamente como a IA influencia vários setores.



Automotivo

A IA é usada para detecção de objetos, rastreamento de faixas e tomada de decisões em veículos autônomos. Ela também pode prever quando o componente de um veículo está propenso a falhar, o que permite a manutenção proativa e reduz o tempo de inatividade.



Produção e indústria

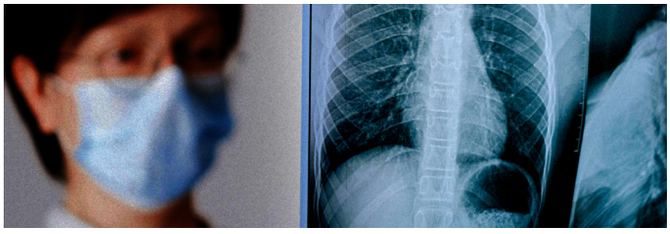
É possível usar IA na produção e na indústria para fins de manutenção preditiva, controle de qualidade, otimização de processos e gerenciamento de cadeia de suprimentos, o que resulta em mais eficiência e em menos tempo de inatividade.

Varejo

A IA pode analisar o comportamento do cliente para fazer recomendações personalizadas de produtos, melhorando o engajamento do cliente e as vendas. Ela também pode otimizar os níveis de estoque prevendo a demanda e minimizando o excesso ou a falta de estoque.

Serviços financeiros

É possível usar IA no setor financeiro e de serviços bancários para fins de detecção de fraudes, avaliação de riscos, serviço de atendimento ao cliente e análise de investimentos, o que resulta em segurança aprimorada e a uma tomada de decisões mais fundamentada.



Área da saúde

É possível usar IA na área da saúde para uma variedade de aplicações, como análise de imagens médicas, diagnóstico de doenças, planejamento de tratamento personalizado e descoberta de medicamentos, o que resulta em melhores resultados para os pacientes e custos reduzidos.

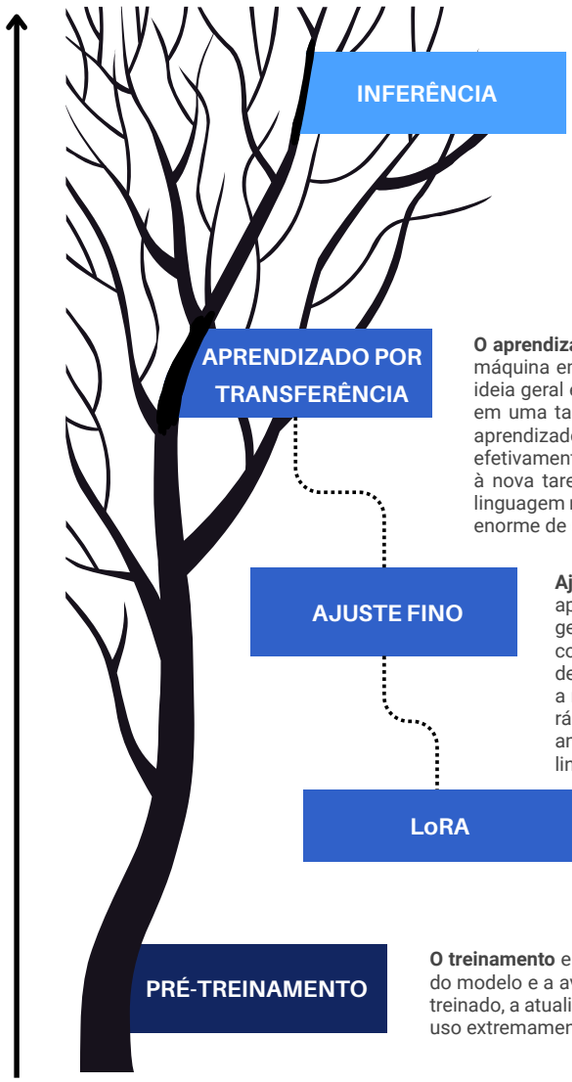
Automação de serviços

Os chatbots alimentados por IA podem lidar com consultas de clientes e oferecer suporte, reduzindo a necessidade de intervenção humana. A IA também pode automatizar tarefas repetitivas, como entrada de dados ou processamento de documentos, o que melhora a eficiência e reduz os erros.

O que os responsáveis pelas decisões de TI devem considerar

COMO COMEÇAR: ANÁLISE DA IA

Antes de abordar estes casos de uso, vamos analisar mais detalhadamente o ciclo de vida da IA. O ciclo de vida da IA (Inteligência Artificial) se refere às fases envolvidas no desenvolvimento, na implementação e na manutenção de um sistema de IA. Embora as metodologias e as terminologias específicas possam variar, um ciclo de vida típico de IA sempre inclui treinamento e inferência de modelos.



Inferência é o processo de usar um modelo treinado de IA para fazer previsões, classificações ou decisões com base em dados novos e inéditos. Nesta fase do ciclo de vida da IA, o modelo treinado é implementado e usado em um ambiente de produção para realizar a tarefa específica para a qual foi projetado. Durante a inferência, o modelo processa os dados de entrada usando as relações e os padrões aprendidos na fase de treinamento e produz um resultado, como uma previsão, uma classificação ou um texto em cargas de trabalho de IA generativa (GenAI). Em seguida, é possível usar esse resultado para várias aplicações, como recomendar produtos a um usuário, detectar anomalias em um sistema ou reconhecer objetos em uma imagem. No processamento de linguagem natural, a geração aumentada de recuperação (RAG) melhora a precisão ao implementar modelos, permitindo que os modelos acessem e aproveitem fontes de conhecimento externas, o que permite que eles entendam e respondam melhor à entrada de linguagem natural em uma ampla variedade de contextos.

O aprendizado por transferência, também conhecido como retreinamento, é uma técnica de aprendizado de máquina em que um modelo treinado em uma tarefa é redirecionado a uma segunda tarefa relacionada. A ideia geral é usar em uma nova tarefa que não tem muitos dados o conhecimento que um modelo aprendeu em uma tarefa com muitos dados rotulados de treinamento disponíveis. Em vez de iniciar o processo de aprendizado do zero, nós começamos com padrões aprendidos ao resolver uma tarefa relacionada, atualizando efetivamente os parâmetros do modelo treinado para a tarefa relacionada para que eles se adequem melhor à nova tarefa. O aprendizado por transferência é usado principalmente em tarefas de processamento de linguagem natural e visão computacional, como análise de sentimentos, para reduzir a quantidade geralmente enorme de capacidade computacional necessária para essas tarefas.

Ajuste fino no ciclo de vida de IA se refere ao processo de treinar ainda mais um modelo de aprendizado de máquina pré-treinado em um novo conjunto de dados ou tarefa. Essa técnica geralmente é usada no aprendizado por transferência, em que um modelo treinado em um grande conjunto de dados de uma tarefa específica é adaptado a uma tarefa relacionada ou a um conjunto de dados menor. O ajuste fino permite a transferência de conhecimento do modelo pré-treinado para a nova tarefa ou o novo conjunto de dados, o que muitas vezes resulta em uma convergência mais rápida e no melhor desempenho em comparação com o treinamento de um modelo do zero. Ele é amplamente utilizado em várias aplicações de IA, como visão computacional, processamento de linguagem natural e reconhecimento de fala.

LoRA (Low Rank Adaptation, Adaptação de Baixo Posto) é uma nova técnica para fazer o ajuste fino de modelos de aprendizagem profunda. Ela funciona reduzindo o número de parâmetros treináveis e permite a troca eficiente de tarefas, fazendo com que o processo de ajuste fino faça um uso menos intenso de computação e tornando-o mais acessível, com um hardware amplamente disponível.

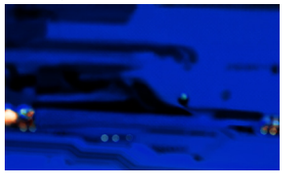
O treinamento envolve a seleção da arquitetura de modelo ou do algoritmo apropriados, o ajuste dos parâmetros do modelo e a avaliação interna do desempenho do modelo. Dependendo do tamanho do modelo que está sendo treinado, a atualização de parâmetros em cada iteração do processo de treinamento pode ser um processo que faz uso extremamente intenso de dados e computação.

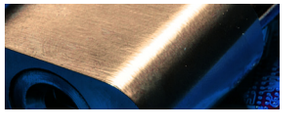
Figura 1: O ciclo de vida da IA

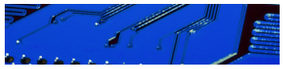
ESCOLHAS ESSENCIAIS

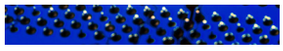
| Desempenho

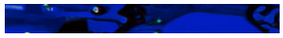
Em muitas dessas aplicações reais, a tomada de decisões em tempo real ou quase em tempo real é essencial para o sucesso. Por exemplo, é necessário identificar imediatamente atividades fraudulentas em transações financeiras ou sinistros de seguros para evitar perdas financeiras e proteger os ativos de negócios. Em um cenário de produção, é necessário monitorar dinamicamente os defeitos da linha de montagem ou as condições da fábrica para fins de garantia da qualidade. Efetivamente, é necessário otimizar o processador que lida com sua carga de trabalho de inferência para processar fluxos de dados recebidos com rapidez e eficiência. Os servidores Dell PowerEdge unidos aos processadores AMD EPYC são uma combinação versátil e adequada para lidar com cargas de trabalho de inferência de borda, bem como tarefas que envolvem computação com alto desempenho, computação em nuvem e lógica analítica de Big Data.











| Segurança de dados

A segurança de dados é crucial para o sucesso dos sistemas de IA, especialmente aqueles que utilizam IA generativa, e é uma grande preocupação para os líderes de tecnologia que pretendem incorporar a IA em suas operações. Geralmente, os sistemas de IA dependem de grandes volumes de dados, que podem incluir informações confidenciais e sensíveis, como dados pessoais, dados financeiros ou informações exclusivas. Proteger esses dados é essencial para impedir o acesso não autorizado ou o roubo de dados, bem como para garantir a precisão, a confiabilidade e a consistência dos modelos e das previsões de IA.

Computação confidencial é uma tecnologia que facilita o processamento de dados em uma área segura, protegendo-os contra acesso não autorizado ou manipulação por partes não autorizadas, inclusive provedores de serviços em nuvem e outros usuários.[2] A criptografia e outras medidas de segurança são usadas para isolar os dados durante o processamento. O AMD Infinity Guard, um conjunto de recursos sofisticados de segurança integrados aos processadores AMD EPYC, oferece suporte à computação confidencial utilizando o Secure Encrypted Virtualization (SEV), que criptografa máquinas virtuais (VMs) usando uma chave conhecida apenas pelo processador. Esses serviços têm o intuito de oferecer ambientes de execução confiável baseados em hardware usando o AMD SEV-Secure Nested Paging (SEV-SNP), que aprimora as proteções de convidados para ajudar na defesa contra ameaças externas.

O aprendizado federado é outro método para manter a segurança de dados. Ele treina um modelo central em dispositivos ou servidores descentralizados.[3] Em vez de transferir todos os dados para uma localização central, cada dispositivo treina o modelo localmente, e apenas as atualizações do modelo são compartilhadas. Essa abordagem preserva a privacidade e permite o aprendizado colaborativo sem compartilhar dados brutos. A plataforma de IA federada da Dell Technologies permite a execução de processos computacionais, IA e algoritmos de ML em conjuntos de dados na borda da rede à medida que vão sendo coletados, compartilhando com outros dispositivos de borda, data centers ou a nuvem apenas modelos matemáticos, metadados e resultados de consultas pela rede. Essa troca aprimora os resultados, permitindo a extração quase em tempo real de informações úteis de grandes conjuntos de dados distribuídos, sem revelar os dados nem qualquer propriedade intelectual.

[2] Advanced Micro Devices, Inc. 30 de agosto de 2023, "AMD shares the technical details of technology Powering Innovative Confidential Computing Leadership Cloud Offerings", https://www.AMD.com/en/newsroom/press-releases/2023-8-30-AMD-shares-the-technical-details-of-technology-pow.html
Advanced Micro Devices, Inc., 2021, Resumo da solução "Data Center Solutions, Confidential Computing", https://www.AMD.com/content/dam/AMD/en/documents/EPYC-business-docs/solution-briefs/confidential-computing-solution-brief.pdf

[3] Analytics Vidhya, dezembro de 2023, "Federated Learning: A Beginner's Guide", https://www.analyticsvidhya.com/blog/2021/05/federated-learning-a-beginners-guide/#:~:text=Federated%20learning%20works%20by%20training,learning%20without%20sharing%20raw%20data
Dell Technologies, 2021, Resumo da solução "A federated learning platform for real-time artificial intelligence", https://www.Delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/dt-sb-analytics-anywhere.pdf

DIMENSIONAMENTO DA SUA SOLUÇÃO

| Equilíbrio entre custo e inovação

Encontrar o equilíbrio certo entre custo e inovação garante que as soluções de IA não sejam apenas financeiramente viáveis, mas também impactantes, o que agrega valor real para empresas e usuários. Um componente-chave para encontrar esse equilíbrio está em identificar um hardware que resolva seus casos de uso e se integre facilmente à infraestrutura existente. No mercado moderno de hardware de IA, o aumento da demanda por aceleradores de vários setores, além das restrições de capacidade de produção, dos desafios logísticos e da escassez de semicondutores, estão contribuindo para a escassez de aceleradores.

No entanto, as CPUs já são um componente padrão na maioria dos data centers, tornando a integração mais simples e econômica em comparação com a adição de um hardware de acelerador totalmente novo. As CPUs otimizadas para IA podem aproveitar o software e as ferramentas existentes, o que reduz a necessidade de uma grande troca de ferramentas ou de novos treinamentos. As CPUs também oferecem mais flexibilidade e eficiência para uma ampla variedade de tarefas para além da IA, permitindo um uso mais versátil de recursos no data center. Atualizar seu data center com servidores Dell PowerEdge que rodam processadores AMD EPYC permitirá que você atenda às suas cargas de trabalho existentes e, ao mesmo tempo, permaneça pronto para avançar em direção às maiores inovações e eficiências orientadas pela IA.

| Simplicidade e flexibilidade

A simplicidade e a flexibilidade de seu sistema de IA são essenciais para a criação de soluções de IA eficazes, adaptáveis e escaláveis em longo prazo. Ter acesso a uma suíte de otimizações e estruturas de software que complementam o hardware aprimora o desempenho sem que você gaste tempo e esforço extras com a integração entre plataformas. Essas qualidades são especialmente importantes para lidar com cargas de trabalho mistas de IA, que envolvem uma combinação de diferentes tipos de tarefas de IA, como treinamento, inferência e processamento de dados.

A AMD e a Dell Technologies lidam com cargas de trabalho mistas de IA por meio de uma combinação de soluções de hardware e software. Os processadores AMD EPYC oferecem capacidade de computação com alto desempenho, com recursos como multithread simultâneo (SMT) e grande número de núcleos, o que permite o processamento paralelo eficiente de cargas de trabalho de IA. Esses processadores são otimizados para tarefas de IA, oferecendo um sólido desempenho para cargas de trabalho de treinamento e inferência. Os servidores Dell PowerEdge, equipados com processadores AMD EPYC, oferecem uma plataforma escalável e flexível para implementar cargas de trabalho de IA. Além disso, a suíte Dell OpenManage Software oferece ferramentas de gerenciamento para otimizar a alocação de recursos e o monitoramento de desempenho das cargas de trabalho mistas de IA.

A AMD também oferece o Unified Inference Frontend (UIF), que aproveita as versões aprimoradas para desempenho de cada uma das pilhas de software atuais e se baseia na biblioteca AMD ZenDNN para processadores AMD EPYC, na pilha AMD ROCm de código aberto para aceleradores AMD Instinct e em uma pilha de software para SoCs adaptáveis AMD. A AMD ROCm também foi projetada para funcionar com uma ampla variedade de CPUs e aceleradores AMD, inclusive produtos profissionais e em nível de consumidor.

| Garantia de explicabilidade

A IA explicável desempenha um papel fundamental na garantia de transparência, confiabilidade e eficácia das aplicações de inteligência artificial. A IA explicável apresenta informações sobre como os modelos de IA tomam decisões, esclarecendo os fatores subjacentes e os processos de raciocínio. Essa transparência é crucial para ganhar a confiança das partes interessadas, especialmente em domínios sensíveis como área da saúde, finanças e justiça criminal, em que as decisões afetam diretamente a vida dos indivíduos.

Os sistemas de IA com **envolvimento humano** aproveitam a inteligência humana para aprimorar o desempenho da IA e reduzir as tendências algorítmicas. Ao integrar a supervisão humana, esses sistemas podem lidar com situações complexas e ambíguas com mais eficácia, garantindo que as soluções de IA se alinhem às normas éticas e sociais. Além disso, o envolvimento humano permite a adaptação e o refinamento contínuos dos modelos de IA com base em feedback real, promovendo a melhoria iterativa e a confiabilidade em longo prazo. Essas abordagens são essenciais para o desenvolvimento de sistemas de IA explicáveis, responsáveis e inclusivos que atendam aos melhores interesses da sociedade.

Cenários reais

A Scalers AI colaborou com a Dell e a AMD para demonstrar os recursos dos servidores Dell PowerEdge equipados com processadores AMD. Confira como essas tecnologias são aproveitadas para treinamento, aprendizado por transferência e inferência em cenários de varejo e da área da saúde.

VAREJO

A Scalers AI criou a solução de referência Retail Inventory Management, um sistema projetado para monitorar e gerenciar os níveis de estoque nas prateleiras de varejo por meio da implementação de um modelo de IA para detecção de objetos. Essa solução de referência utiliza o modelo SSD_MobileNet_V2 para identificar e reconhecer produtos nas prateleiras das lojas. Em última análise, isso proporciona contagens automáticas de estoque e monitoramento preciso dos níveis de estoque. O modelo passou por aprendizado por transferência usando o conjunto de dados de imagens SKU110K, composto por 23 mil imagens da Roboflow. Ao aproveitar os algoritmos de aprendizado de máquina e visão computacional, o sistema pode detectar quando os itens estão sem estoque ou acabando, emitindo alertas à equipe da loja para fazer a reposição ou o reabastecimento em tempo hábil.

Esta solução utiliza o servidor Dell PowerEdge R7615 com o processador AMD EPYC 9354P de 32 núcleos.

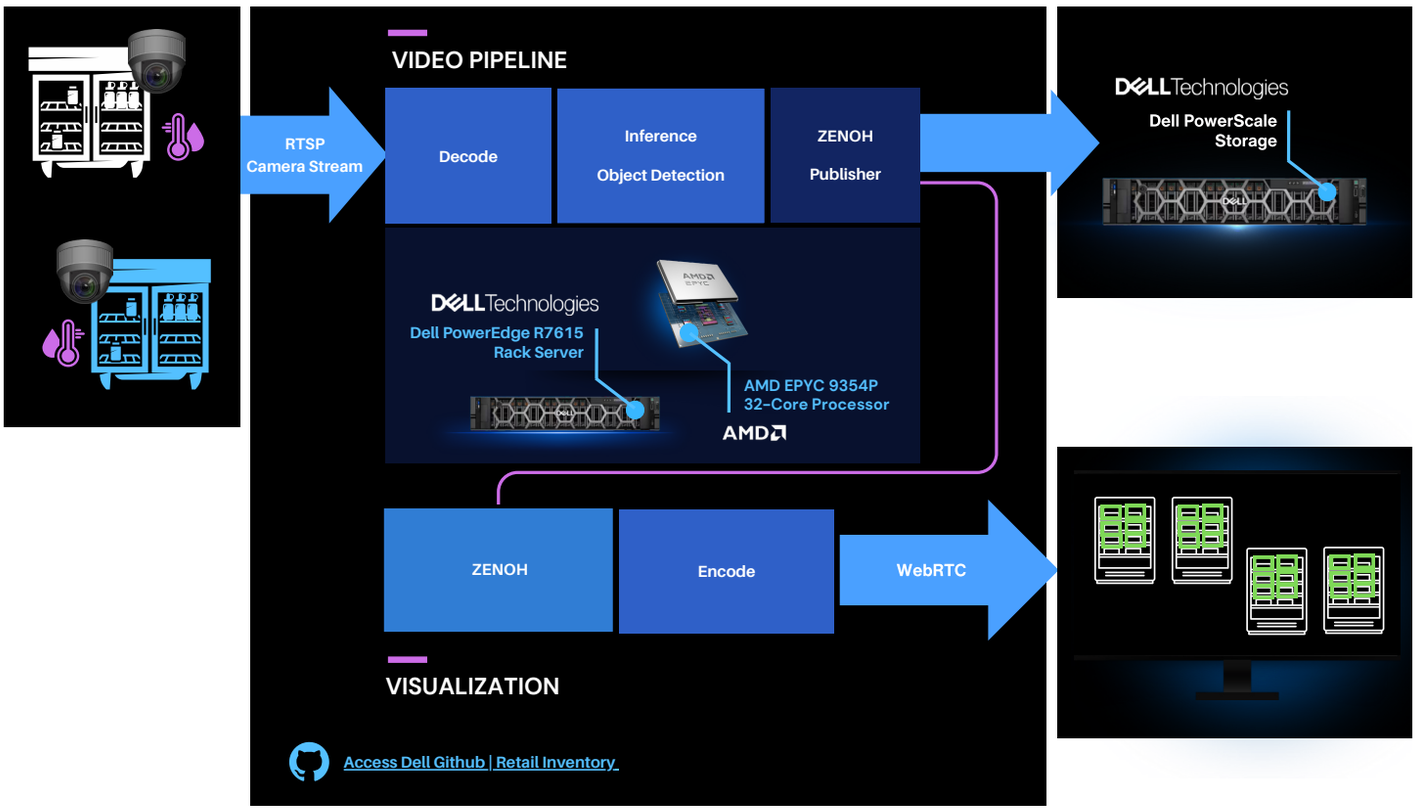


Figura 2: Diagrama de arquitetura da solução de referência Retail Inventory Management

ÁREA DA SAÚDE

As imagens médicas alimentadas por IA são imensamente valiosas por sua capacidade de aprimorar a área da saúde ao melhorar a precisão e a eficiência dos diagnósticos e oferecer aos profissionais de saúde informações precisas sobre condições que podem ser difíceis de detectar a olho nu. Ao automatizar a análise de imagens médicas, a IA reduz o tempo necessário para o diagnóstico, permitindo decisões de tratamento mais rápidas e, em última análise, aprimorando os resultados dos pacientes.

A Scalers AI explorou os recursos do servidor Dell PowerEdge R7625 equipado com processadores AMD EPYC 9554 de 64 núcleos para criar uma solução de imagens médicas alimentada por IA para detecção de pneumonia. Usando algoritmos avançados e técnicas de aprendizado de máquina para analisar imagens médicas, como radiografias ou tomografias, a solução ajuda a aumentar a velocidade e a precisão do diagnóstico de pneumonia em pacientes. Por fim, isso introduz uma camada adicional de revisão assistida por computador, criando potencial para ajudar os profissionais da área da saúde a lidar com grandes volumes de dados de imagem com mais eficiência.

Essa solução de referência utiliza o modelo ResNet50 para analisar radiografias de tórax obtidas do conjunto de dados do NIH Clinical Center. Seu objetivo primário é detectar a presença ou a ausência de pneumonia, realizando basicamente uma classificação binária. O modelo foi treinado usando o conjunto de dados DICOM para radiografia do NIH Clinical Center, envolvendo aprendizado por transferência com a arquitetura ResNet50.

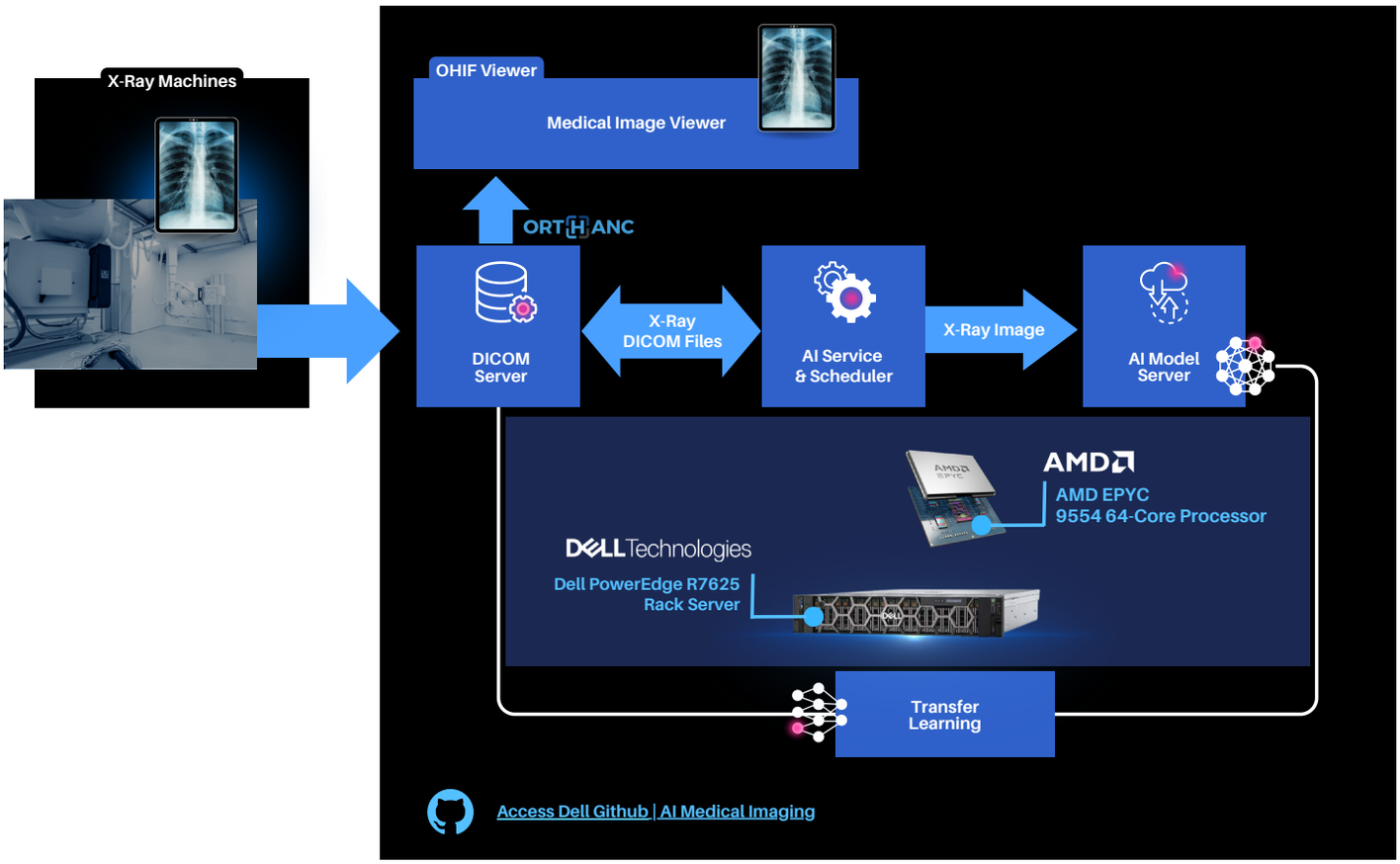
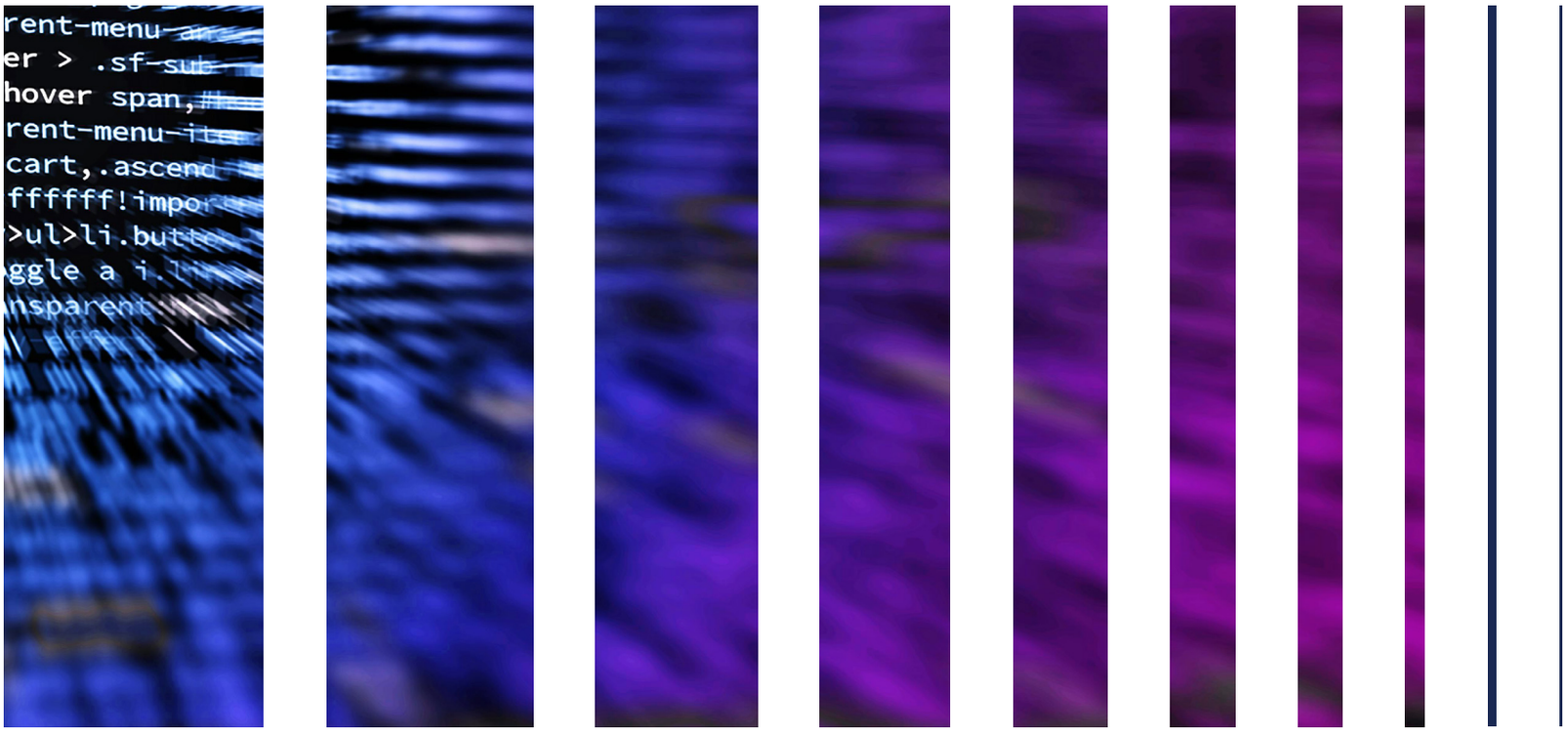


Figura 3: Diagrama de arquitetura da solução Medical AI Imaging

Nossas soluções

A IA É PARA TODOS: DELL E AMD DEMOCRATIZANDO A IA

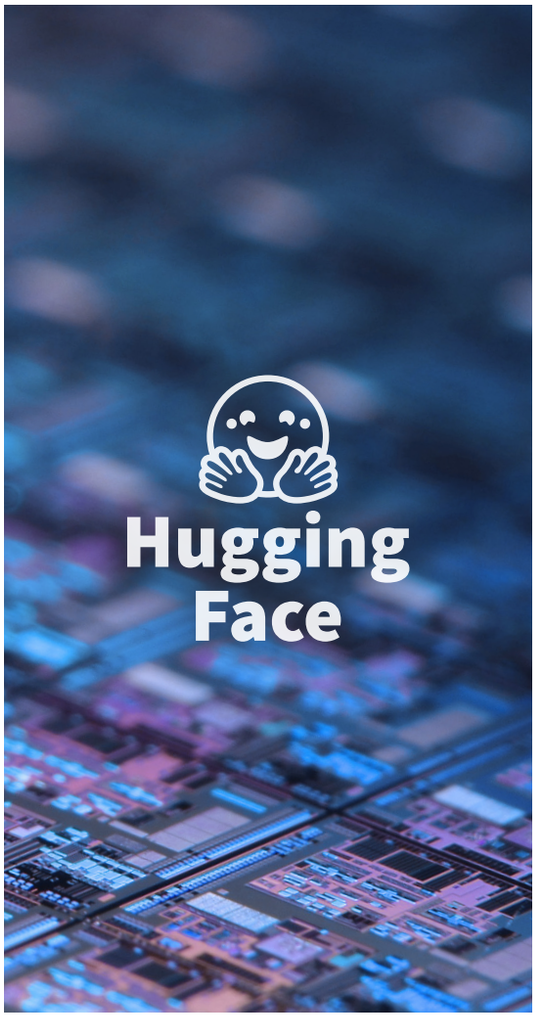
Essa colaboração estabelece a base para a democratização da IA, que é essencial para fomentar a inovação e promover a inclusão no ecossistema de IA. A Dell e a AMD estão alcançando esse resultado capacitando indivíduos e organizações a aproveitar a IA e vencer desafios exclusivos de seus respectivos campos com uma suíte acessível de servidores avançados e equipados com tecnologias de CPUs e aceleradores AMD de última geração. Os servidores Dell PowerEdge com os aceleradores AMD Instinct MI300X podem lidar com grandes cargas de trabalho de IA, como treinamento e ajuste fino de grandes modelos de linguagem (LLMs), enquanto os servidores Dell PowerEdge equipados com processadores AMD EPYC se destacam ao lidar com cargas de trabalho de inferência de borda. Além da plataforma de hardware subjacente, a AMD também oferece a biblioteca de software ZenDNN para a otimização da inferência de aprendizagem profunda em CPUs AMD e a biblioteca de software AMD ROCm para aprimorar os recursos de treinamento, ajuste fino e inferência dos aceleradores AMD Instinct. Todas essas opções estão perfeitamente vinculadas no Unified Inferencing Model (UIF) da AMD, por meio do qual os usuários podem criar soluções completas de IA, com flexibilidade na escolha de estruturas de software, otimizações de software e plataformas de hardware.

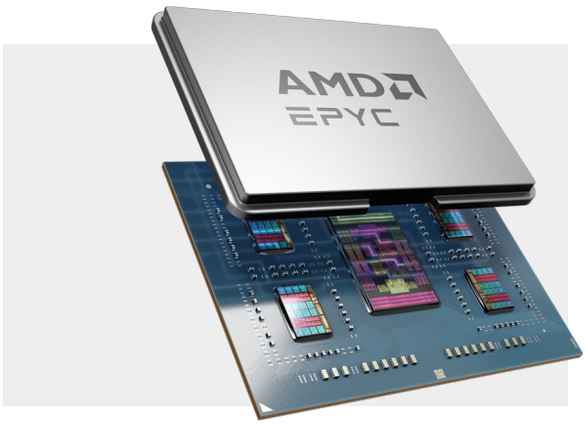


COLABORAÇÃO DA HUGGING FACE

As empresas ansiosas por adotar a IA podem começar aproveitando os modelos preexistentes ou os fluxos de trabalho de IA personalizados para suas necessidades específicas diretamente da Hugging Face, uma plataforma de código aberto dedicada à ciência de dados e ao aprendizado de máquina. A AMD fez uma colaboração com a Hugging Face com o objetivo compartilhado de entregar um desempenho de alto nível da Transformers, adicionando otimizações de software específicas da AMD a estruturas e bibliotecas de software que já se integram perfeitamente às plataformas AMD. A Hugging Face está colaborando ativamente com a equipe de engenharia da AMD para otimizar os principais modelos para o máximo desempenho, incorporando a AMD ROCm à biblioteca Transformers e aprimorando a Optimum-AMD, uma biblioteca projetada especificamente para plataformas AMD, para ajudar os usuários da Hugging Face a utilizá-los com alterações mínimas de código.

Recentemente, a Dell Technologies também uniu forças com a Hugging Face para simplificar o processo para que as empresas desenvolvam, façam o ajuste fino e apliquem seus próprios modelos de IA generativa (GenAI) de código aberto usando a comunidade Hugging Face, tudo em produtos e serviços de infraestrutura Dell líderes do setor. Um novo portal da Dell está sendo desenvolvido na plataforma Hugging Face, que incluirá contêineres e scripts personalizados e dedicados para ajudar os usuários a implementar com segurança e facilidade os modelos de código aberto disponíveis na Hugging Face usando servidores e sistemas de armazenamento de dados da Dell. Agora, as empresas podem aproveitar ao máximo os recursos da Hugging Face para implementar modelos diretamente em servidores Dell PowerEdge com processadores AMD e criar soluções completas de IA usando seus próprios dados exclusivos.



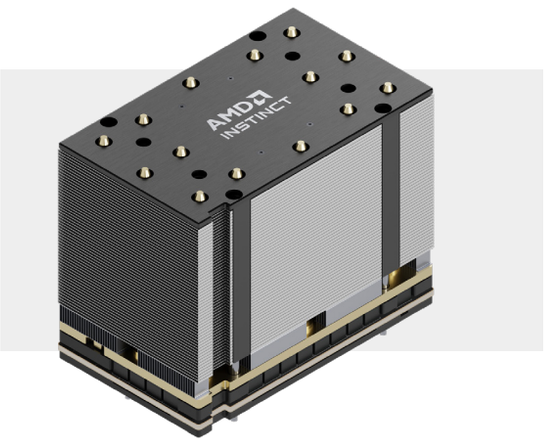


PROCESSADORES AMD EPYC

A AMD oferece os avanços tecnológicos necessários para data centers modernos baseados em nuvem por meio de seus processadores AMD EPYC. Esses processadores são um sistema em chip (SoC) projetado do zero para atender com eficiência às demandas dos data centers atuais e futuros. Os processadores AMD EPYC série 9000 equipam o data center com até 128 núcleos, 256 threads, 12 canais de memória que oferecem suporte até 6 TB de memória por soquete e 128 vias PCIe de 5ª geração. Tudo isso é combinado com a solução de segurança de servidor x86 incorporada ao hardware, que é pioneira no setor. Ao integrar recursos essenciais de computação, memória, E/S e segurança no SoC, os processadores AMD EPYC produzem um desempenho de nível superior e promovem um menor custo total de propriedade (TCO).

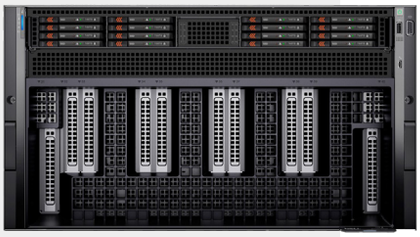
ACELERADORES AMD INSTINCT MI300X

O acelerador AMD Instinct MI300X, integrado à moderna arquitetura AMD CDNA 3, oferece eficiência e desempenho líderes do setor para as aplicações que fazem uso mais intenso de IA e HPC. Ele é equipado com 304 unidades de computação de alto desempenho e apresenta funções específicas de IA, como suporte a novos tipos de dados e decodificação de fotos e vídeos, bem como uma memória HBM3 inigualável de 192 GB em um só acelerador.



PLATAFORMA DE SOFTWARE DE CÓDIGO ABERTO AMD ROCM 6

A plataforma de software de código aberto AMD ROCm 6 é otimizada para maximizar a computação com alto desempenho (HPC) e o desempenho das cargas de trabalho de IA dos aceleradores AMD Instinct MI300X. Ela também estende o suporte a aceleradores AMD Instinct MI300X, garantindo a compatibilidade com estruturas de software do setor. A plataforma AMD ROCm encapsula diversos drivers, ferramentas de desenvolvimento e APIs que facilitam a programação de aceleradores desde o nível do kernel até os aplicativos do usuário final e pode ser adaptada para se alinhar a seus requisitos específicos. Ela é especialmente ideal para aplicações de computação com alto desempenho (HPC), inteligência artificial (IA) e computação científica. Além disso, a plataforma AMD ROCm oferece suporte à computação com vários aceleradores, inclusive acesso direto remoto à memória (RDMA) para comunicação entre servidores e nós.



PORTFÓLIO DE SERVIDORES DELL POWEREDGE

O investimento da Dell na AMD cria uma opção essencial no mercado para democratizar a IA, como é evidenciado por suas quatro plataformas de servidor com EPYC e seu principal servidor em rack Dell PowerEdge XE9680 com aceleradores AMD Instinct MI300X. A geração mais recente de servidores PowerEdge alimentada por processadores AMD EPYC aumenta a agilidade nos negócios e o time-to-market, com a capacidade de oferecer suporte a cargas de trabalho transformadoras como bancos de dados e lógica analítica, virtualização, armazenamento definido por software, infraestrutura de desktop virtual (VDI), conteinerização, computação com alto desempenho (HPC), IA e aprendizado de máquina (ML). Seus servidores em rack com um soquete (CPU única) oferecem um equilíbrio de baixo custo entre desempenho e capacidade de armazenamento e foram projetados para crescer perfeitamente com sua empresa, enquanto os servidores em rack com dois soquetes (CPU dupla) acomodam cargas de trabalho mais exigentes com uma ampla variedade de recursos.

O servidor em rack Dell PowerEdge XE9680 é uma sólida potência de processamento de dados, desenvolvida especificamente para tarefas de IA. Ele oferece suporte a oito aceleradores, o que o torna perfeito para cargas de trabalho de inferência e treinamento de aprendizado de máquina (ML)/aprendizagem profunda (DL), principalmente para treinar grandes modelos de linguagem (LLMs). Equipado com 8 aceleradores MI300X, cada um com 192 GB de memória com grande largura de banda (HBM3) de 5,3 TB/s, o que resulta em uma capacidade total de HBM3 de 1,5 TB por servidor e mais de 21 petaflops de desempenho FP16, o servidor em rack Dell PowerEdge XE9680 com aceleradores AMD Instinct MI300X está pronto para ampliar ainda mais a acessibilidade da IA generativa para as empresas. Isso permite que eles treinem modelos maiores, minimizem o espaço ocupado pelo data center, reduzam o TCO e obtenham uma vantagem competitiva.

Resumo

O ritmo acelerado das inovações impulsionado pela IA está revolucionando as cargas de trabalho de data center com mais rapidez do que em qualquer outra transformação tecnológica. Para oferecer suporte a esses avanços tecnológicos, a Dell e a AMD estão trabalhando rumo a um ecossistema de IA mais inclusivo, inovador e desenvolvido com ética que incentive desenvolvedores de todos os setores a colaborar em recursos de código aberto e impulsionar a inovação da IA generativa atual. Independentemente de sua solução de IA atender a seus requisitos de desempenho nos processadores AMD EPYC ou em servidores equipados com aceleradores AMD Instinct, nós oferecemos a flexibilidade para executar cargas de trabalho de IA em todas as nossas plataformas de hardware. Assim, você aproveita o melhor que a Dell e a AMD têm a oferecer.

REFERÊNCIAS

Imagens da AMD: AMD.com, AMD Partner Resource Library, **https://www.amd.com/en/partner/resources/resource-library.html**

Imagens da Dell: **Dell.com**