**Sustentabilidade da IA**

Introdução

Considere um cenário em que sua organização planeja escalar as operações de IA. No entanto, a gerência está preocupada com o aumento dos custos de energia e o impacto ambiental negativo que quaisquer mudanças na infraestrutura do data center causarão. Como profissional de TI, como você ameniza as preocupações levantadas pela gerência?

Com os insights deste treinamento, você poderá propor uma abordagem equilibrada para a atualização do data center. Usando os princípios da IA ​​Verde, você poderá fornecer uma solução que ofereça resultados de alto desempenho, ao mesmo tempo em que conduz sua organização em direção a uma estratégia de IA sustentável e econômica que contribua para a responsabilidade ambiental.

Ao final deste treinamento, você será capaz de:

* Aplique técnicas como poda de modelos e quantização para modelos de IA com eficiência energética.
* Avalie as compensações entre soluções de IA locais e baseadas na nuvem para otimizar custos.
* Crie um plano para integrar aceleradores de IA e fontes de energia renováveis ​​para minimizar o impacto ambiental.

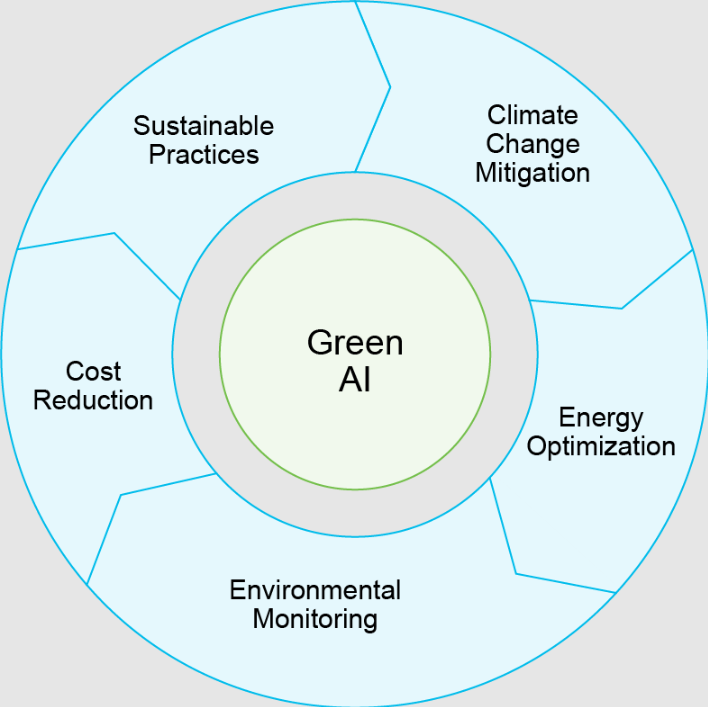
**IA Verde vs. IA Vermelha**

Com o crescimento da IA, uma questão que muitas vezes permanece inexplorada é como ela afeta e continuará afetando a sustentabilidade ambiental. O consumo de energia de um data center de IA é realmente sem precedentes. Veja a plataforma Nvidia Blackwell como exemplo. Um cluster Blackwell com 8 GPUs requer até 15.000 watts de energia, com uma média de 1.875 watts por GPU. A GPU da geração anterior, a H100, consome apenas 700 watts.

Em pequena escala, esses números podem não parecer muitos, mas quando você considera que os novos data centers de IA têm 300.000 GPUs ou mais, os requisitos de energia se tornam astronômicos. 300.000 GPUs podem consumir mais da metade da energia que uma usina nuclear média pode produzir.

**IA verde**

Você pode ver como os requisitos da IA ​​podem impactar o mundo. IA Verde refere-se a um movimento na comunidade de inteligência artificial para reduzir o impacto ambiental por meio do uso de tecnologias de IA. Essas tecnologias se concentram no desenvolvimento de modelos de IA mais eficientes em termos de energia, que promovam a sustentabilidade ambiental.



A IA Verde alinha o avanço tecnológico com a responsabilidade ambiental. À medida que a IA se integra cada vez mais em diversos setores, a necessidade de mitigar sua pegada ambiental se torna mais premente. Ao adotar práticas de IA Verde, as organizações podem garantir que o crescimento da IA ​​contribua positivamente para as metas globais de sustentabilidade.

Os benefícios da IA ​​Verde incluem o seguinte:

* **Práticas sustentáveis:** a IA verde pode ajudar a garantir que os recursos sejam usados ​​de forma eficiente e que as práticas da economia circular sejam respeitadas.
* **Mitigação das mudanças climáticas:** com a aplicação de gestão inteligente de energia e práticas agrícolas aprimoradas, a IA Verde pode ajudar a reduzir os efeitos das mudanças climáticas.
* **Otimização de energia:** a IA verde pode ajudar a reduzir o consumo de energia em edifícios inteligentes, além de reduzir o desperdício e as emissões de carbono.
* **Monitoramento ambiental:** ao monitorar o ambiente em tempo real, a Green AI pode fornecer recomendações para governança ambiental.
* **Redução de custos:** ao reduzir o consumo de energia, a IA Verde pode ter um impacto positivo na redução de custos.

Apesar de sua importância, a implementação da IA ​​Verde enfrenta diversos desafios. Entre eles, destacam-se:

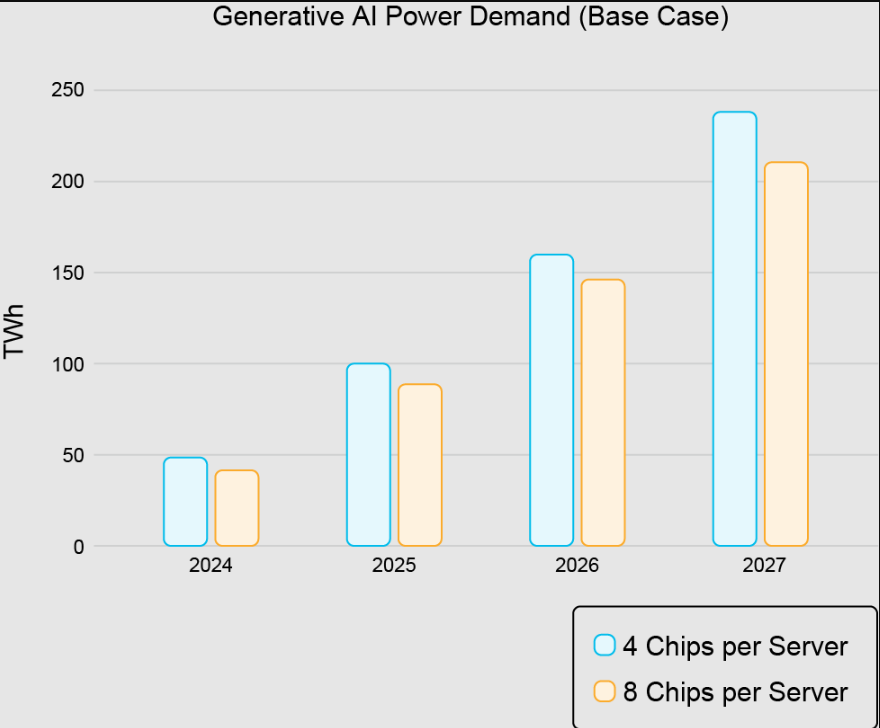
* A necessidade de I&D significativa para criar modelos energeticamente eficientes
* O potencial aumento de custos associado às tecnologias mais ecológicas
* A complexidade da integração de práticas de IA Verde em diversos setores e infraestruturas existentes

Para superar esses desafios, várias estratégias podem ser empregadas:

* **Priorizar o Design de Modelos com Eficiência Energética:** Focar no desenvolvimento e na implantação de modelos de IA que exijam menos poder computacional e energia. Essa abordagem inclui o uso de técnicas como poda de modelos, quantização e a adoção de modelos menores e mais eficientes sempre que possível.
* **Uso de fontes de energia renováveis:** a transição de data centers e operações de IA para operar com fontes de energia renováveis ​​pode reduzir significativamente a pegada de carbono associada às cargas de trabalho de IA.
* **Incorporação de Métricas de Sustentabilidade:** Incorpore a sustentabilidade como uma métrica fundamental nos processos de desenvolvimento de IA. Essa estratégia pode envolver a avaliação do impacto ambiental dos modelos de IA durante a fase de projeto e a priorização de modelos que atendam aos padrões de eficiência energética.
* **Promovendo P&D Colaborativo:** Envolva-se em colaborações intersetoriais para impulsionar a inovação em tecnologias de IA com eficiência energética. A parceria com instituições acadêmicas, líderes do setor e organizações ambientais pode acelerar o desenvolvimento da IA ​​Verde.
* **Implementando infraestrutura com eficiência energética:** invista em hardware e infraestrutura projetados para minimizar o consumo de energia, como processadores com eficiência energética, sistemas de resfriamento e soluções de armazenamento de dados.

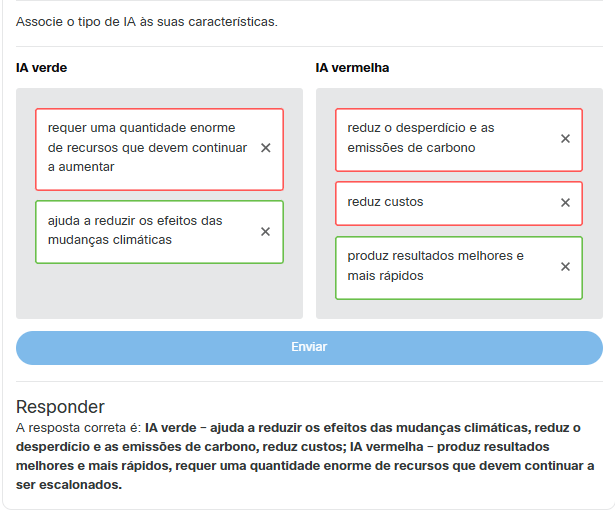
**IA vermelha**

Ao contrário da IA ​​Verde, o objetivo da IA ​​Vermelha é produzir resultados melhores e mais rápidos, com pouca preocupação com o impacto ambiental. A filosofia da IA ​​Vermelha é investir mais potência na IA para ajudar a aumentar a precisão. Para ilustrar esse ponto, o Morgan Stanley prevê que as necessidades de energia da IA ​​poderão superar o consumo total de energia dos data centers em 2022 até 2027.



A IA Vermelha consome muita energia, o que lhe rendeu uma reputação negativa. No entanto, apesar de todos os pontos negativos da IA ​​Vermelha, é impossível ignorar o fato de que seus modelos oferecem maior precisão e velocidade. Além disso, o lado positivo da má reputação da IA ​​Vermelha é que ela atraiu mais atenção e elevou o debate sobre a IA Verde.

À medida que a IA continua a crescer, a batalha entre a IA Vermelha e a IA Verde provavelmente continuará. É fundamental que os desenvolvedores trabalhem juntos para alcançar um equilíbrio que resulte em uma IA eficiente e ética.

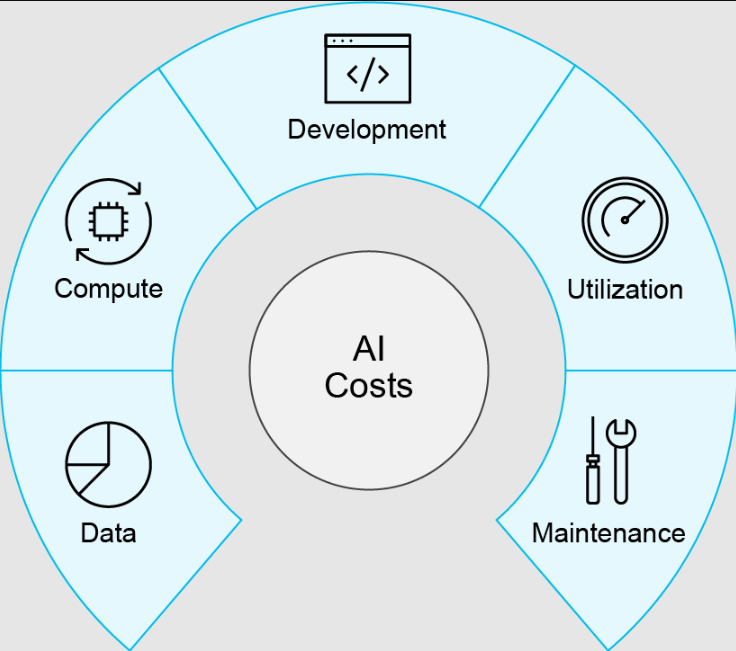


Otimização de custos

No início de 2024, a UPS fez um anúncio que causou preocupação para muitas pessoas. Para controlar o aumento dos custos, a UPS decidiu demitir 12.000 pessoas e substituí-las por sistemas habilitados para IA. Muitos temem que essa tendência continue, com resultados potencialmente prejudiciais. Outros acreditam que se trata apenas de uma mudança no ambiente de trabalho, semelhante à introdução do tear no início do século XIX. Independentemente de como você veja essa mudança, uma coisa é certa: as empresas estão buscando usar a IA para otimizar os custos operacionais.

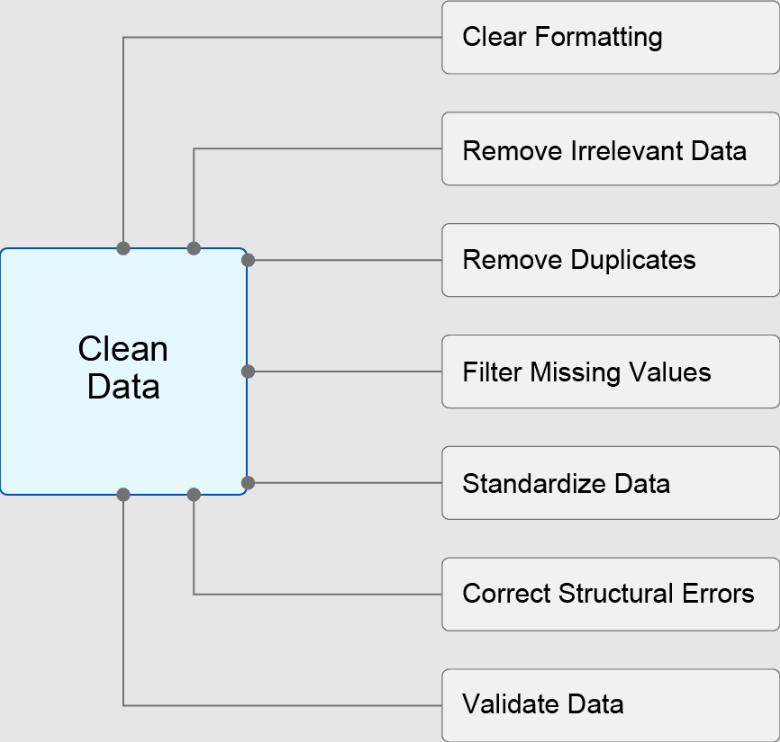
E embora a IA possa ajudar as organizações a reduzir custos operacionais, sua implementação também pode ser bastante dispendiosa. Muitos fatores afetam os custos da IA, como infraestrutura, energia, treinamento e desenvolvimento. Além disso, há custos de capital humano. A implantação e o uso adequados da IA ​​exigem habilidades especializadas que atualmente exigem altos salários.

Devido a todos esses fatores, antes de implementar uma estratégia de IA, é importante considerar ações que minimizem o custo de implementação. Os custos de IA normalmente se enquadram em cinco categorias: dados, computação, desenvolvimento, utilização e manutenção.



**Custos de dados**

A aquisição e o processamento necessários para produzir dados de alta qualidade às vezes não são considerados quando as organizações decidem explorar uma solução de IA pela primeira vez. Quando a maioria das pessoas usa um serviço de IA para geração de texto, como OpenAI, Claude, Bard ou um dos inúmeros serviços de bate-papo com IA, elas não param para considerar a enorme quantidade de tempo, energia e custo investidos na criação desses sistemas. Estima-se que o custo de desenvolvimento do ChatGPT da OpenAI seja superior a 100 milhões de dólares. Espera-se que esses custos continuem a aumentar, com algumas previsões de que cheguem a mais de 100 bilhões de dólares (cerca de US$ 310 por pessoa nos EUA).



Felizmente, as organizações podem não precisar construir e treinar modelos de IA inteiros, pois muitas vezes podem ajustar modelos pré-treinados ao seu caso de uso específico. Por exemplo, uma IA eficaz requer dados limpos e de alta qualidade. Os dados brutos devem ser limpos manualmente, por meio de automação, por meio de um serviço ou talvez por uma combinação desses métodos. A preparação de dados precisos, confiáveis ​​e limpos facilitará o ajuste fino de um modelo existente, garantindo que ele gere inferências de qualidade, aumente a eficiência e reduza o custo geral

**Custos de computação**

Outra consideração importante da IA ​​é o custo do poder computacional que ela requer. Uma das primeiras decisões que as organizações que implementam IA precisam tomar é se usarão uma solução de computação local ou em nuvem. Cada solução tem suas vantagens e desvantagens, portanto, ambas devem ser cuidadosamente consideradas.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Ao determinar qual tipo de implantação será mais econômico, você também precisa equilibrar essa decisão com os outros requisitos da sua organização. Por exemplo, se você deseja usar uma solução em nuvem por ser a mais econômica considerando a utilização esperada, mas sua empresa tem uma política de que nenhuma informação interna pode ser armazenada em servidores de terceiros, talvez uma solução híbrida seja a mais apropriada.

**Custos de desenvolvimento**

Ao escolher um caminho de desenvolvimento de IA, você tem essencialmente duas opções: uma solução pré-construída ou uma solução personalizada. Os custos associados a essas opções podem variar significativamente com base em diversos fatores. Os custos de desenvolvimento abrangem não apenas o investimento financeiro, mas também o tempo, os recursos e os riscos potenciais envolvidos na criação e manutenção de uma solução de IA. Esses custos são cruciais porque impactam diretamente a eficácia e a sustentabilidade geral da implantação de IA.

As principais questões que devem orientar a avaliação dos custos de desenvolvimento incluem:

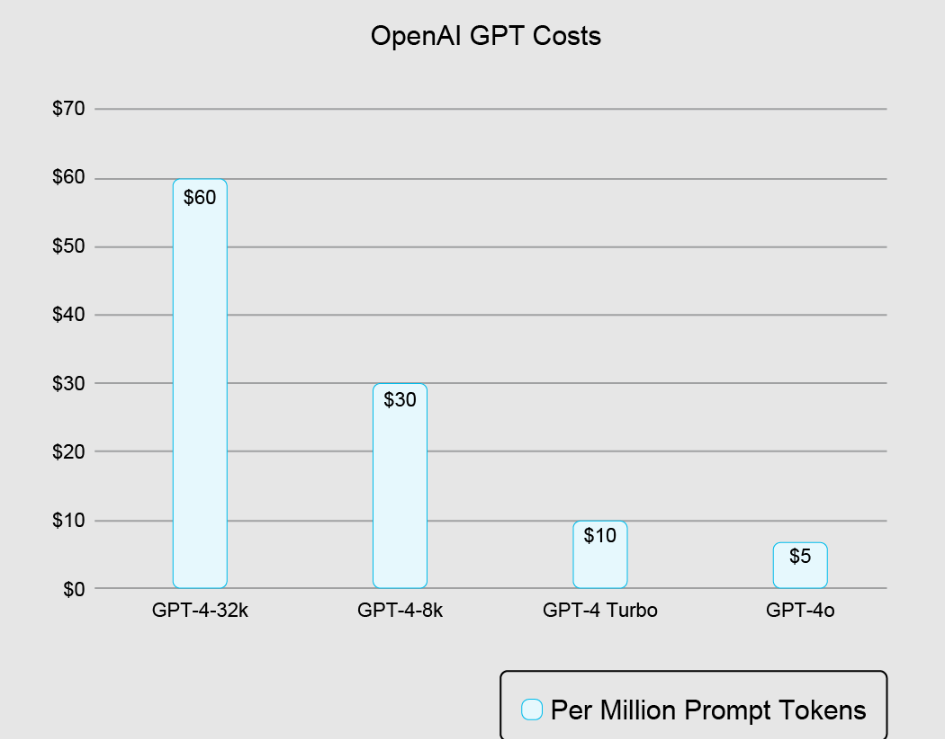
* Qual é a complexidade do sistema de IA? Sistemas mais complexos exigem mais tempo de desenvolvimento e expertise, o que pode aumentar os custos.
* Quanta personalização é necessária? O grau de personalização afetará os custos iniciais e contínuos.
* Qual é o tamanho do conjunto de dados? Conjuntos de dados maiores podem exigir uma infraestrutura mais robusta, o que influencia as considerações de custo.
* Quais são os requisitos de escalabilidade? A capacidade da solução de IA de se adaptar às necessidades da organização pode ter implicações significativas em termos de custos.
* Quais são os requisitos de manutenção e atualização? A manutenção contínua e a necessidade de atualizações regulares também podem impactar o custo total de propriedade.

Compreender e definir esses custos é vital para tomar uma decisão informada entre uma solução de IA pré-desenvolvida ou personalizada. Os custos de desenvolvimento devem ser vistos como um parâmetro que influencia o design e a implantação de sistemas de IA garantindo que estejam alinhados ao orçamento e aos objetivos de longo prazo da organização

**Custos de Utilização**

Os custos de utilização da IA, as despesas incorridas durante a fase de inferência dos modelos de IA, podem, por vezes, ultrapassar todos os outros custos associados, especialmente ao utilizar serviços de IA em nuvem. Esses serviços em nuvem normalmente cobram com base no número de tokens processados ​​pela sua solução de IA. Alguns provedores oferecem taxas mensais fixas com um limite para o uso total, geralmente projetadas para soluções com menor utilização prevista.

A figura a seguir mostra os custos associados às diferentes versões do modelo GPT-4 da OpenAI. O GPT-4 é um modelo de linguagem amplamente utilizado que alimenta diversas aplicações, de chatbots a ferramentas de geração de conteúdo. As diferentes versões na figura refletem capacidades e custos variados, com modelos de ponta como o GPT-4-32k oferecendo recursos mais abrangentes a um preço mais alto.

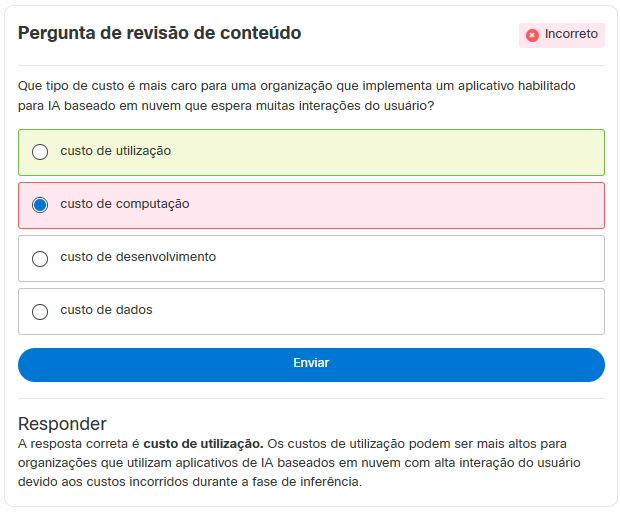


A boa notícia é que, à medida que os modelos e o hardware melhoram, o custo por token está diminuindo. Para referência, um token de prompt é uma unidade de entrada de texto processada pelo modelo; saídas mais detalhadas ou complexas geralmente usam mais tokens. O GPT-4-32k da OpenAI custa atualmente US$ 60 por 1 milhão de tokens de prompt, enquanto o GPT-4o custa apenas US$ 5 por milhão de tokens de prompt.

Escolher o modelo correto para suas necessidades pode ter um impacto sério no custo geral da solução.

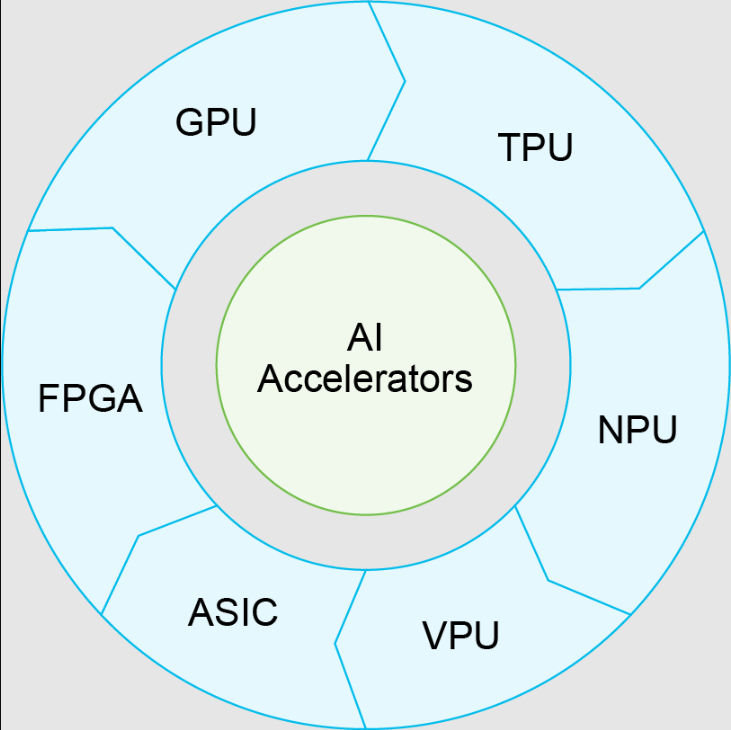
**Custos de manutenção**

Os custos de manutenção de IA geralmente andam de mãos dadas com o caminho de desenvolvimento escolhido. Soluções pré-construídas geralmente incluem manutenção no plano, portanto, o custo já está incluído e é previsível. No entanto, soluções personalizadas podem frequentemente incorrer em custos adicionais para itens como atualizações de conjuntos de dados, upgrades de modelos e complementos de recursos, só para citar alguns. Uma solução personalizada também exige um orçamento para custos de manutenção, garantindo que o sistema permaneça funcional, atualizado e capaz de atender às necessidades em evolução.



**Aceleradores de IA**

O poder computacional da IA ​​necessário para atingir velocidades de processamento mais rápidas continua a crescer. Uma abordagem para atender aos requisitos atuais de computação de IA é implementar aceleradores de IA, hardware especializado projetado especificamente para acelerar os tempos de processamento de aplicativos habilitados para IA. Ao adicionar aceleradores de IA a outros recursos computacionais, eles adicionam potência adicional ao sistema. Os aceleradores de IA são projetados para processar cargas de trabalho de IA especificamente, ao contrário das CPUs, que processam diversas cargas de trabalho e tarefas do sistema



Aceleradores de IA também ajudam a levar o processamento de IA para a borda da rede, onde os dados são gerados e processados ​​localmente. Essa estratégia reduz a latência e o uso de largura de banda, minimizando a necessidade de enviar dados para servidores centralizados para processamento. Frequentemente, a carga de trabalho de processamento de IA é realizada na nuvem e os resultados são enviados para a borda. Embora esse processo funcione bem com frequência, ele não funciona bem quando é necessária uma latência extremamente baixa. Em uma aplicação de IA, como carros autônomos, em que decisões em frações de segundo precisam ser tomadas, mover o processamento para a borda reduz drasticamente a latência e a dependência da conectividade de rede.

Tipos de aceleradores de IA

Existem vários tipos diferentes de aceleradores de IA, cada um com sua própria função e propósito. A única característica que todos eles têm em comum é que oferecem muito mais eficiência na carga de trabalho de processamento de IA do que uma CPU tradicional.

Os tipos de aceleradores de IA incluem o seguinte:

* **Unidades de Processamento Gráfico (GPUs):** Como o nome indica, as Unidades de Processamento Gráfico (GPUs) foram inicialmente projetadas para aliviar a carga de trabalho gráfica da CPU. Elas podem lidar com várias operações simultaneamente, o que é conhecido como processamento paralelo, tornando-as candidatas perfeitas para processamento de IA. As GPUs existem há muito tempo; o termo foi cunhado pela primeira vez na década de 1980, mas foi a GeForce 256, lançada pela Nvidia em 1999, que o introduziu na computação convencional.
* **Unidades de Processamento Tensor (TPUs):** As TPUs foram desenvolvidas pelo Google e criadas para acelerar cargas de trabalho de aprendizado de máquina. São circuitos integrados específicos para aplicações (ASICs) otimizados para trabalhar com uma estrutura de aprendizado de máquina chamada TensorFlow. As TPUs são usadas para treinamento de modelos por empresas como Google, Apple e Lyft.
* **Unidades de Processamento Neural (NPUs):** As NPUs são projetadas para processar rapidamente cálculos de redes neurais. Elas podem realizar tarefas como processamento de linguagem natural, reconhecimento de imagens e outras que exigem altos recursos computacionais. As NPUs são mais frequentemente encontradas em dispositivos de ponta e usadas para aplicações que variam de celulares a carros autônomos.
* **Unidades de Processamento de Visão (VPUs):** Assim como as NPUs, as VPUs são frequentemente encontradas na borda. Elas são projetadas especificamente para tarefas de visão computacional, tornando-as ideais para tarefas de reconhecimento visual, como vigilância, carros autônomos, drones e inspeção de peças.
* **Circuitos Integrados de Aplicação Específica (ASICs):** Os ASICs são circuitos integrados (CI) projetados e fabricados para executar uma tarefa específica e somente essa tarefa, o que os torna extremamente eficientes e de alto desempenho.
* **Matrizes de Portas Programáveis ​​em Campo (FPGAs):** FPGAs são um tipo de circuito integrado (CI), mas, diferentemente dos ASICs, podem ser configurados após a fabricação. Essa capacidade permite que o mesmo tipo de FPGA seja usado para diferentes tarefas e seja ajustado para fornecer melhor desempenho, potência e outros requisitos de aplicação.
* Benefícios dos aceleradores de IA

A maioria dos aceleradores pode realizar processamento paralelo. Esse tipo de processamento permite que múltiplas tarefas computacionais sejam executadas simultaneamente, proporcionando maior velocidade e desempenho. Os aceleradores de IA também oferecem a flexibilidade de executar diversas cargas de trabalho de IA no mesmo hardware, permitindo que se adaptem a uma variedade de tarefas de IA e a requisitos em constante mudança.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

À medida que a demanda aumenta, aceleradores de IA adicionais podem ser adicionados a um sistema e trabalhar em paralelo com os aceleradores existentes para escalar. Alguns sistemas podem ter centenas de milhares de GPUs.

Como os aceleradores de IA são projetados para tarefas específicas, eles podem oferecer um grau maior de precisão à solução. Essa precisão, aliada a velocidades mais rápidas, pode fornecer resultados com maior eficiência do que dispositivos computacionais comuns que utilizam CPUs.

Por fim, todos esses benefícios: velocidade, flexibilidade, escalabilidade, precisão e eficiência, fornecem uma maneira econômica de gerenciar e dimensionar cargas de trabalho de IA.

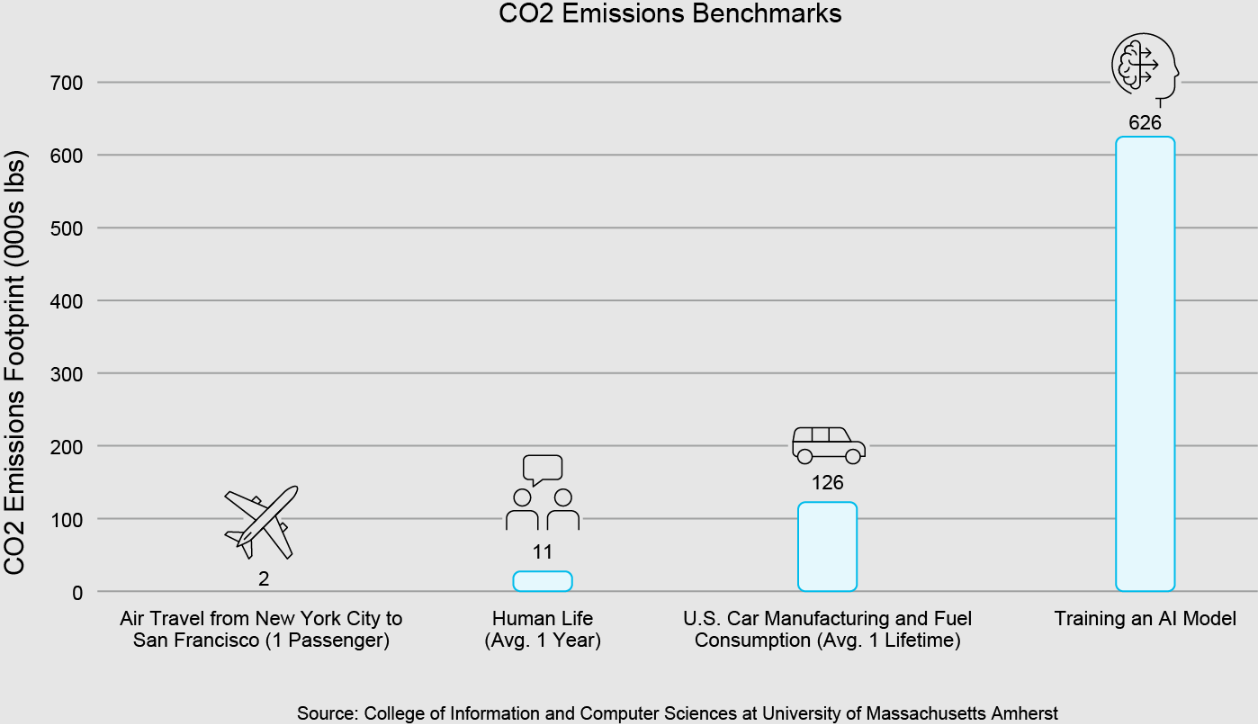
Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

**Energia e refrigeração**

A energia necessária para operar um novo data center de treinamento de modelos de IA tem o potencial de ser maior do que a de uma usina nuclear média. Empresas de tecnologia que se comprometeram a reduzir sua pegada de carbono, na verdade, a aumentaram. Por exemplo, relatórios mostram que a pegada da Microsoft aumentou quase 30% desde 2020, e muitos apontam a IA como a fonte desse aumento.

Algumas pessoas esperam que a IA possa eventualmente ajudar na busca pela sustentabilidade ambiental a longo prazo. No entanto, neste momento, a IA pode estar causando mais mal do que bem. De acordo com um estudo, treinar um modelo de IA produz mais CO2 do que uma pessoa média de 55 anos produz desde o nascimento.



É fundamental que as organizações que trabalham para o avanço da IA ​​o façam de maneira ambientalmente responsável. As três principais áreas nas quais as organizações podem reduzir os requisitos de energia e refrigeração gerados pela IA são a otimização de hardware, a eficiência de software e a gestão de data centers.

**Otimização de Hardware**

É menos eficiente usar uma CPU para uma tarefa que um acelerador de IA pode realizar. Existem vários tipos de hardware especializado, como VPUs, ASICs, TPUs, NPUs e outros, que têm requisitos de energia menores do que processadores de uso geral

Linha do tempo

O conteúdo gerado por IA pode estar incorreto.

O gerenciamento dinâmico de recursos é outro método para reduzir o consumo de energia. Um método chamado limitação de energia impõe limites à quantidade de energia que um dispositivo pode usar, evitando que altos níveis de energia sejam usados ​​durante os horários de pico de carga. Outro método, a Voltagem Dinâmica com Escala de Frequência (DVFS), reduz o consumo de energia ajustando dinamicamente a voltagem e a frequência dos processadores com base na carga de trabalho atual. Por fim, a Memória Computacional de Acesso Aleatório (CRAM) ainda está em fase de pesquisa, mas os resultados são extremamente promissores. A CRAM processa dados dentro das células de memória, eliminando a necessidade de transferi-los entre os módulos lógicos e de memória. Os testes iniciais mostram uma redução de 2.500 vezes no consumo de energia em relação aos métodos tradicionais.

**Eficiência de software**

A eficiência do software de IA refere-se à capacidade de algoritmos, modelos e aplicativos de IA de executar tarefas com o uso otimizado de recursos, como poder computacional. A principal maneira de alcançar a eficiência do software com IA é por meio da otimização do modelo de IA.

Tabela

O conteúdo gerado por IA pode estar incorreto.

Os principais aspectos da otimização do modelo incluem as seguintes técnicas:

* **Quantização:** envolve o ajuste da precisão dos números do modelo, como usar inteiros de 8 bits em vez de inteiros de 32 bits.
* **Redução de modelo:** a prática de remover parâmetros de rede neural não utilizados para ajudar a reduzir o tamanho do modelo, mantendo o desempenho.
* **Destilação do Conhecimento:** Treina um modelo menor para se comportar como um modelo maior. O modelo menor é considerado o aluno, enquanto o modelo maior é o professor. Como o modelo do aluno é menor, ele consumirá menos energia.
* Gestão de Data Center
* Há vários fatores a serem considerados ao tentar reduzir o consumo de energia em um data center de IA. Não existe um método único que possa ser implementado para alcançar uma solução adequada, mas, ao implementar estrategicamente vários métodos, é possível obter uma redução significativa no consumo de energia.

Gráfico, Gráfico de pizza

O conteúdo gerado por IA pode estar incorreto.

Para gerenciar um data center com eficácia, uma organização precisa ter processos para monitorar o consumo de energia. O monitoramento vai além da simples verificação do consumo total de energia no final do mês. Em vez disso, o monitoramento inclui a implementação de sistemas que monitoram continuamente o consumo de energia em cada dispositivo para otimizar o uso dinamicamente com base na carga de trabalho.

As organizações também devem considerar a origem da energia que obtêm, o que pode ser mais importante do que a quantidade que consomem. Por exemplo, um data center alimentado por uma usina termelétrica a carvão produzirá uma emissão de CO2 muito maior do que um movido a energia solar.

Por fim, para que os data centers funcionem de forma confiável, seus componentes devem ser mantidos refrigerados. Dispositivos projetados para IA produzem muito calor. As estimativas variam, mas a maioria concorda que o custo do resfriamento de dispositivos específicos para IA pode ser responsável por 30 a 40% do consumo de energia de um data center.

Métodos comuns de resfriamento de data center e suas vantagens e desvantagens incluem o seguinte:

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

Texto

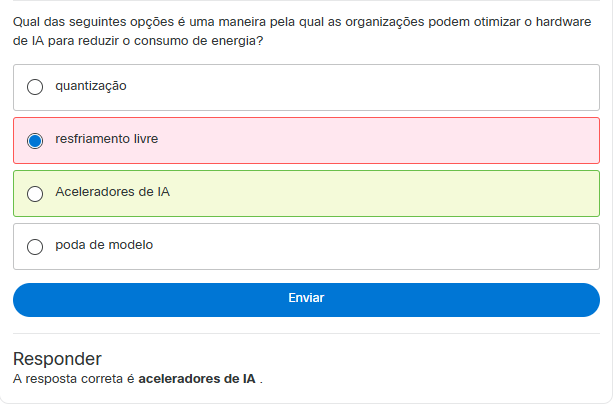
O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.



Resumo

Considere as perguntas nas seguintes áreas ao refletir sobre o conteúdo do treinamento e aplicar os conceitos que aprendeu em seu próprio ambiente:

* **Implementação de IA Verde** : Como sua organização pode priorizar o desenvolvimento e a implantação de modelos de IA com eficiência energética? Que medidas podem ser tomadas para integrar métricas de sustentabilidade aos seus processos de desenvolvimento de IA?
* **Otimização de Custos** : Considerando as cinco categorias de custos de IA (dados, computação, desenvolvimento, utilização e manutenção), qual área você acha que sua organização poderia otimizar mais? Que ações específicas poderiam ser tomadas para reduzir custos nessa área?
* **Aceleradores de IA e Gerenciamento de Energia** : Como sua organização pode aproveitar os aceleradores de IA para melhorar a eficiência do processamento? Quais estratégias podem ser implementadas para otimizar a energia e o resfriamento em seus data centers de IA e reduzir o impacto ambiental?