**Considerações sobre o design da infraestrutura de IA**

**Design de infraestrutura de IA**

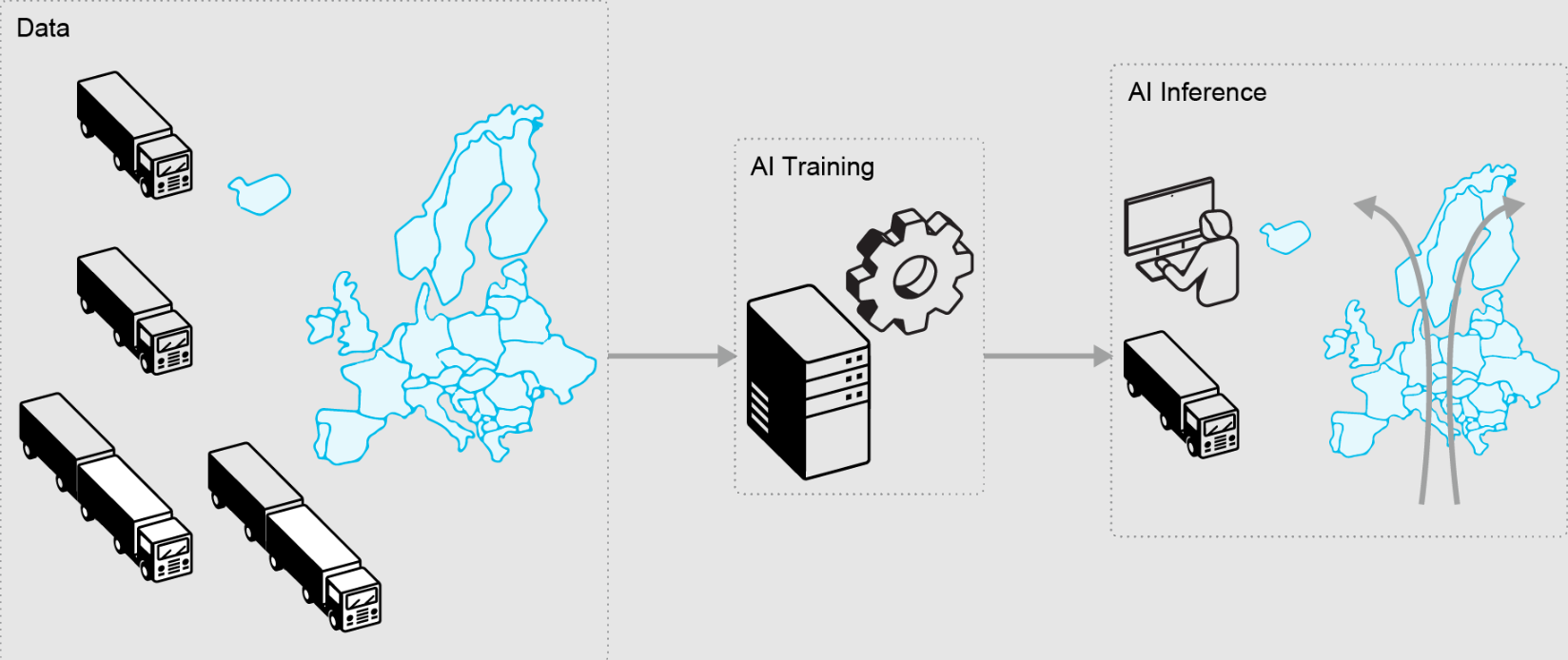
Introdução

A Inteligência Artificial (IA) está se tornando um divisor de águas para muitos setores, especialmente na otimização de operações e na melhoria da eficiência. Esta atividade se concentra em você aconselhar uma empresa sobre como decidir sobre uma infraestrutura de IA. O objetivo desta atividade é fornecer orientações sobre como criar um ambiente de IA eficiente, econômico e capaz de resolver problemas do mundo real, como reduzir ineficiências operacionais e aprimorar a tomada de decisões.

**Descrição do Projeto**

Você trabalha para uma empresa que é uma grande empresa no setor de transporte, gerenciando uma vasta frota de caminhões que operam em diversas regiões da Europa.

Apesar do sucesso, sua empresa de transporte, com uma receita anual de US$ 230 milhões e 1.200 funcionários, enfrenta diversos desafios relacionados a ineficiências operacionais. Alguns desses desafios incluem consumo de combustível, otimização de rotas e programação de manutenção. Reconhecendo o potencial da IA ​​para lidar com essas questões, a empresa está ansiosa para estabelecer uma prática de IA do zero. Seu objetivo é usar a IA para otimizar a gestão de sua frota, reduzir custos, melhorar os prazos de entrega e aprimorar a eficiência operacional geral.



Seu papel

Você é consultor de uma empresa de transporte. Sua função é orientar a liderança da empresa sobre o melhor caminho a seguir na construção de uma infraestrutura de IA. Você fornecerá recomendações sobre como configurar e gerenciar a infraestrutura necessária para dar suporte às iniciativas de IA. Sua orientação incluirá a avaliação de diferentes opções de infraestrutura de IA, a identificação dos principais investimentos em hardware e software e a proposta de estratégias para maximizar a eficiência e minimizar os custos. Além disso, você aconselhará sobre práticas de gerenciamento de dados para garantir a conformidade com as regulamentações e explorará tecnologias de ponta que podem ser integradas às operações da empresa. Seus insights serão cruciais para ajudar sua empresa a usar a IA para lidar com suas ineficiências e atingir seus objetivos de negócios.

Atividade 1: Tipo de carga de trabalho de IA

Nesta atividade, você será responsável por planejar a infraestrutura de IA necessária para sua empresa de transporte, à medida que ela explora a criação de capacidades internas para o treinamento de modelos de IA. O CTO propôs a ideia de desenvolver modelos de IA em larga escala internamente, o que pode ser um grande avanço para a empresa. O e-mail a seguir é um convite para discutirmos o caminho a seguir.

*Convite para reunião:*

***Assunto:****Convite para Reunião — Planejamento de Infraestrutura de IA para Treinamento de Modelo Interno*

*Querida equipe,*

*Espero que esta mensagem o encontre bem. Como parte dos nossos esforços contínuos para otimizar* as operações da empresa *usando Inteligência Artificial, estou considerando uma meta ambiciosa para a nossa empresa. Quero explorar a possibilidade de construir uma infraestrutura de IA que nos permita treinar nossos próprios modelos de IA, semelhante a como algumas empresas líderes de tecnologia desenvolveram suas próprias versões de IA.*

*A pauta desta reunião inclui:*

* *Ao discutir os requisitos de infraestrutura para treinar modelos de IA em larga escala, me disseram que o LLaMA de código aberto é o caminho a seguir.*
* *Avaliar os custos associados à construção e manutenção dessa infraestrutura.*
* *Decidir sobre a estratégia mais viável e econômica*

*Sua experiência e contribuição serão valiosas para nos ajudar a tomar uma decisão informada. Aguardo ansiosamente nossa conversa.*

*Atenciosamente,*

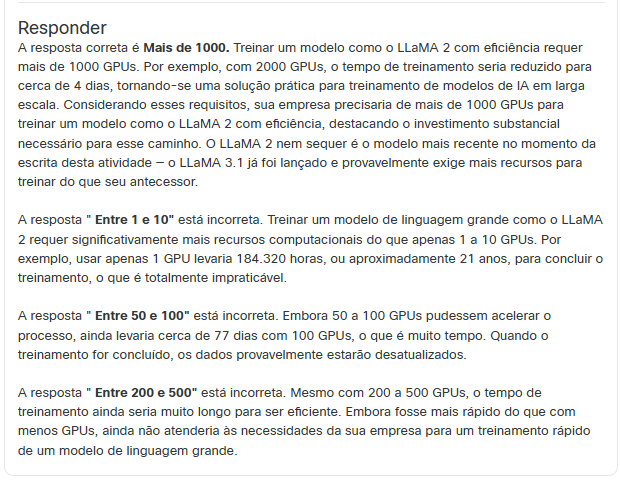
*Diretor de Tecnologia*

**Pergunta de revisão de conteúdo**

Considerando o tamanho da sua empresa e a receita atual, você percebe que o CTO pode não entender completamente a escala da infraestrutura de IA necessária.

No início da reunião, o CTO pergunta quantas GPUs seriam necessárias para treinar um modelo de linguagem grande como o LLaMA 2 da Meta do zero. Qual seria a melhor estimativa entre as respostas possíveis?

1. Parte superior do formulário
2. Entre 1 e 10
3. Entre 50 e 100
4. Entre 200 e 500
5. Mais de 1000



Parte inferior do formulário

**Pergunta de revisão de conteúdo**

Após a discussão anterior, o CTO agora compreende os altos custos e requisitos de infraestrutura associados ao treinamento de um grande modelo de linguagem. O CTO agora pergunta se existem alternativas viáveis.

O que sua empresa poderia considerar para usar IA sem a despesa significativa de treinamento do zero?

1. Parte superior do formulário
2. A empresa deve prosseguir com o treinamento em IA a qualquer custo para permanecer competitiva.
3. A empresa pode se concentrar no uso de modelos pré-treinados e personalizá-los para aplicações específicas.
4. A empresa deve abandonar completamente a IA devido aos altos custos.
5. A empresa deve investir apenas em ferramentas básicas de análise de dados.

Texto, Carta

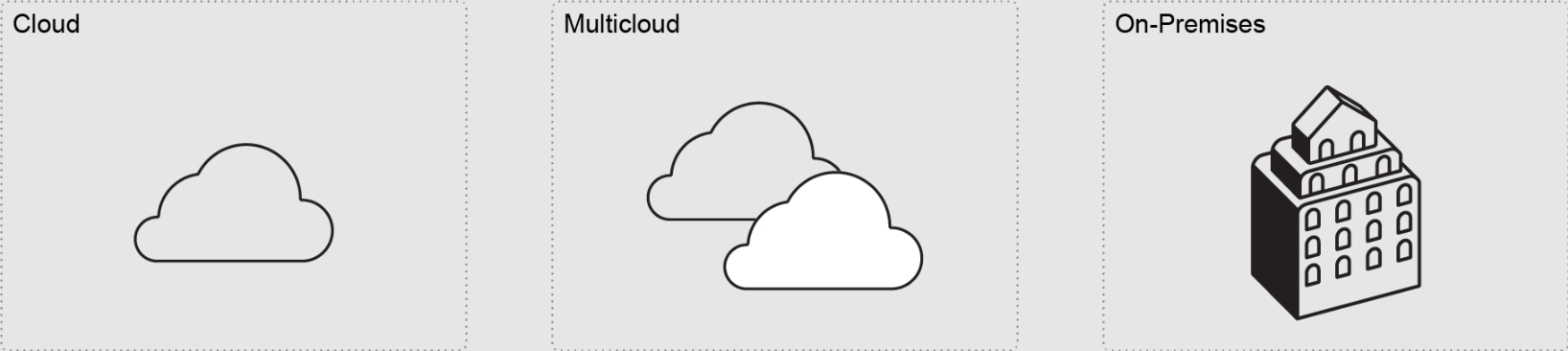
O conteúdo gerado por IA pode estar incorreto.

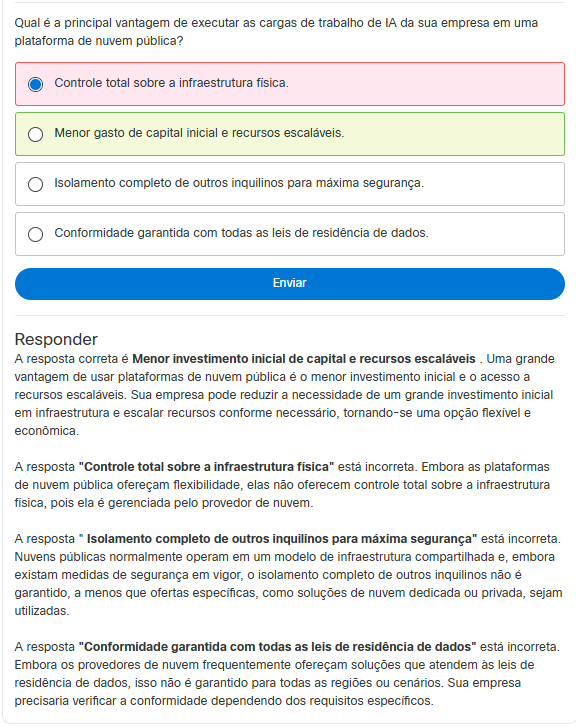
Parte inferior do formulário

A maioria das empresas hoje opta por usar modelos de IA prontos em vez de construir os seus próprios do zero. Treinar um modelo do zero exige muito tempo, dinheiro e poder de computação, o que muitas empresas consideram muito caro. Empresas como Meta e Google já oferecem modelos pré-treinados de código aberto e alta qualidade que podem ser adaptados a diferentes necessidades de negócios sem a necessidade de reconstrução. Com opções como RAG e ajuste fino, as empresas ainda podem personalizar esses modelos para suas necessidades específicas, economizando tempo e dinheiro. Isso permite que empresas como a sua implementem soluções de IA rapidamente sem investir em infraestrutura cara, tornando a IA acessível e econômica.

Atividade 2: Nuvem vs. Local

*O CTO da* sua empresa *quer avançar com o uso de IA para aprimorar as operações da empresa por meio de modelos pré-treinados. Embora a empresa tenha um pequeno data center, ela está considerando usar a nuvem pública para executar suas cargas de trabalho de IA.*





Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Sua empresa decidiu executar suas cargas de trabalho de IA localmente após analisar suas necessidades e custos de soberania de dados. Como as regulamentações rigorosas na Europa, incluindo o Regulamento Geral sobre a Proteção de Dados (GDPR), exigem que os dados sejam processados ​​e armazenados no país em que foram coletados, eles precisavam manter seus dados locais para cumprir a norma. Além disso, uma análise de custos mostrou que gerenciar sua própria infraestrutura de IA lhes daria melhor controle sobre as despesas de longo prazo. Ao gerenciar hardware e software eles mesmos, eles poderiam reduzir custos em comparação com a dependência de provedores de nuvem, o que poderia se tornar caro à medida que o uso de IA aumentasse. Esse foco em conformidade legal e economia de custos os levou a escolher uma solução local.

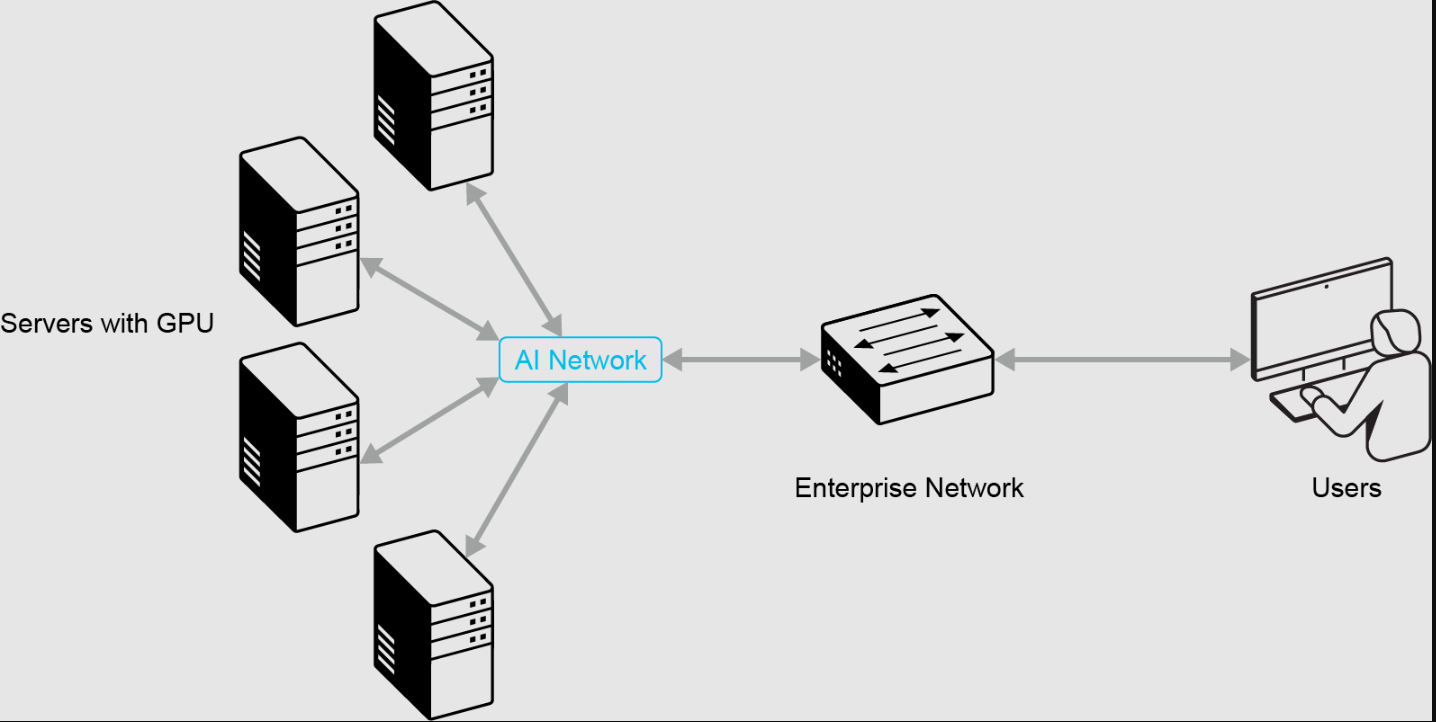
Atividade 3: A Escolha da Rede

Após extensas discussões com a equipe de dados da sua empresa, foi tomada a decisão de implantar o modelo LLaMA 3.1 com oito bilhões de parâmetros para melhorar a eficiência operacional e aprimorar a tomada de decisões baseada em dados. De acordo com o Hugging Face em <https://huggingface.co/> , diferentes modelos exigem quantidades variadas de memória, dependendo de sua arquitetura e das tarefas específicas que executam. Por exemplo, o modelo LLaMA 3.1 8B para inferência pode exigir até 16 GB de memória por instância, enquanto o ajuste fino pode exigir até 60 GB de memória por instância. O Hugging Face é uma plataforma onde a comunidade de aprendizado de máquina colabora em modelos, conjuntos de dados e aplicações.

Os requisitos de memória para executar modelos de IA dependem de vários fatores:

* **Tamanho do modelo (número de parâmetros)** : modelos maiores com mais parâmetros exigem mais memória para armazenar e processar dados.
* **Precisão (FP16 vs. FP32)** : A precisão usada durante a computação afeta o uso da memória. FP16 (ponto flutuante de 16 bits) usa menos memória que FP32 (ponto flutuante de 32 bits), mas maior precisão fornece respostas de melhor qualidade.
* **Tamanho do lote:** tamanhos de lote maiores aumentam a memória necessária para armazenar ativações e cálculos intermediários.
* **Comprimento da sequência:** modelos que lidam com sequências maiores de dados precisam de mais memória para processamento.

Considerando essas considerações, sua empresa não quer ficar excessivamente limitada pela escolha de um modelo específico. Em vez disso, optou por começar com servidores equipados com GPUs NVIDIA A100-80, cada uma oferecendo 80 GB de VRAM. Essa escolha proporciona à sua empresa ampla capacidade de memória para executar confortavelmente a maioria dos modelos de IA em uma única GPU, garantindo flexibilidade e escalabilidade em suas implantações de IA.



Além de selecionar GPUs potentes, sua empresa também deve garantir que sua infraestrutura de IA possa lidar com o tráfego significativo de dados gerado por esses modelos. Cargas de trabalho de IA, especialmente aquelas que envolvem modelos grandes como o LLaMA 3.1, exigem soluções de rede robustas para evitar gargalos que podem prejudicar o desempenho. Para manter a comunicação eficiente entre os servidores e garantir que os dados fluam de forma rápida e fluida durante o treinamento e a inferência, sua empresa precisa de uma rede capaz de fornecer baixa latência e alta taxa de transferência. Isso será fundamental para otimizar suas cargas de trabalho de IA e evitar atrasos no processamento de dados, contribuindo, em última análise, para um melhor desempenho geral do sistema e para a experiência do usuário.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Atividade 4: Escolha da Plataforma e Sustentabilidade

Após a avaliação abrangente das soluções de rede, sua empresa decidiu executar suas cargas de trabalho de IA localmente, utilizando os recursos robustos do Cisco HyperFabric para latência ultrabaixa e interconectividade de alta taxa de transferência entre nós. Essa decisão foi influenciada pela necessidade de uma rede escalável e de alto desempenho que simplifique o gerenciamento, dado o tamanho limitado da equipe de rede da sua empresa. Com a infraestrutura de rede instalada, a peça final do quebra-cabeça da implantação de IA envolve a seleção dos servidores apropriados e do hardware associado. À medida que sua empresa avança com essa estratégia local, é crucial considerar os aspectos de consumo de energia e sustentabilidade das opções de servidores para garantir a eficiência operacional e o alinhamento com as metas de sustentabilidade corporativa

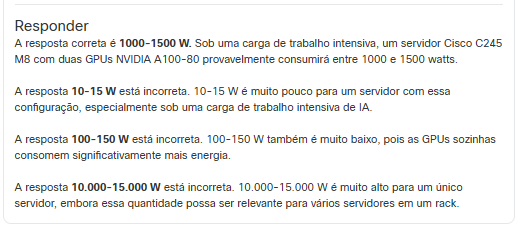
Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Atividade 5: Considerações sobre energia

Sua empresa decidiu prosseguir com a implantação de servidores Cisco para suas cargas de trabalho de IA. O plano inicial envolve o uso de cinco servidores Cisco C245 M8, cada um equipado com duas GPUs NVIDIA H100-80. Olhando para o futuro, sua empresa pretende expandir essa configuração para um total de 20 servidores. Um técnico de data center propôs um plano em que todos os 20 servidores serão instalados em um único rack para otimizar o uso do espaço.





O consumo de energia é um fator crítico para cargas de trabalho de IA, especialmente em ambientes com disponibilidade limitada de energia em racks. Muitos clientes possuem racks com disponibilidade de energia limitada a 8-10 kW. A instalação de 20 desses servidores em um único rack pode resultar em um consumo total de energia de 20-30 kW, dependendo da configuração do servidor.

Caso o cliente não tenha energia suficiente disponível no rack, ele tem algumas opções. Uma opção é aumentar o fornecimento de energia ao rack, embora isso muitas vezes não seja viável, especialmente em instalações de colocation onde a capacidade de energia é fixa. Como alternativa, o cliente pode precisar adotar um layout menos eficiente, distribuindo os servidores em vários racks que podem não ser totalmente utilizados. Essa solução, embora garanta energia e refrigeração adequadas, pode levar a maiores necessidades de espaço e custos potencialmente mais altos devido ao uso ineficiente do espaço do rack.

A captura de tela a seguir mostra as opções de configuração selecionadas para a Calculadora de Energia do Cisco UCS. Esta configuração inclui componentes-chave para um servidor que lidará com cargas de trabalho de IA:

* **Processador** : Duas CPUs AMD 9654, cada uma rodando a 2,4 GHz, com 96 núcleos.
* **Memória** : 12 pentes de 64 GB DDR5-5600 RDIMM, totalizando 768 GB.
* **Fonte de alimentação** : Duas fontes de alimentação, cada uma com potência nominal de 2300 W.
* **Controlador de armazenamento dedicado** : um controlador RAID otimizado para inicialização usando SSDs M.2.
* **Armazenamento** : Dois SSDs SATA M.2 de 240 GB para armazenamento.
* **Adaptador** : Uma placa Ethernet Cisco UCS VIC 15235 Dual Port 40/100/200/400Gb.
* **GPU** : Duas GPUs NVIDIA H100, cada uma consumindo 350 W e fornecendo 80 GB de memória para processamento de IA.

Esta configuração é apenas um exemplo, e o consumo real de energia varia de acordo com os componentes selecionados. No entanto, é importante observar que as GPUs são os componentes que mais consomem energia.

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

A captura de tela a seguir mostra as estimativas de consumo de energia para o servidor configurado:

* **Fator de Carga de Trabalho do Sistema (W)** : Estima-se que o servidor consuma 1.239,3 watts a 50% de utilização, o que é uma boa média para a maioria das cargas de trabalho de IA. Isso dá uma ideia aproximada de como será o consumo de energia em condições normais.
* **Potência Máxima (W)** : O consumo máximo de energia do servidor é estimado em 1.887,1 watts. Embora seja improvável que o servidor sempre atinja esse nível, é aconselhável planejar isso ao projetar o data center para garantir que haja energia e refrigeração suficientes disponíveis.
* **Energia ociosa (W)** : quando o servidor não está processando ativamente cargas de trabalho, ele consome 599,3 watts de energia.

Os resultados também mostram outras informações importantes:

* **Requisitos de resfriamento:** O servidor precisará de 4.228,8 BTU/h de resfriamento.
* **Emissões de GEE:** As emissões estimadas de gases de efeito estufa são de 4,77 toneladas de CO2 por ano para este servidor.

Este cálculo de potência é apenas um exemplo, pois os números variam dependendo dos componentes exatos utilizados. No entanto, na maioria das cargas de trabalho de IA, as placas de GPU são as que mais consomem energia, portanto, devem ser o foco ao planejar a energia e o resfriamento do data center. É essencial planejar tanto o consumo médio de energia quanto a potência máxima que o servidor pode consumir, para evitar problemas com limites de energia ou superaquecimento, especialmente ao executar vários servidores em um único rack.

Interface gráfica do usuário, Aplicativo, Tabela

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

O conteúdo gerado por IA pode estar incorreto.

Ao implantar cargas de trabalho de IA com servidores de alto desempenho, é crucial considerar o aumento do consumo de energia e a consequente geração de calor. Esses servidores, especialmente quando equipados com GPUs de alta potência como a NVIDIA H100-80, geram quantidades significativas de calor, exigindo recursos de resfriamento adequados para manter as temperaturas operacionais ideais e evitar superaquecimento. O resfriamento adequado não apenas garante a confiabilidade e a longevidade do servidor, mas também aumenta a eficiência energética, reduzindo a energia necessária para os sistemas de resfriamento. O resfriamento insuficiente pode levar ao superaquecimento, à redução do desempenho do servidor ou até mesmo a falhas de hardware, tornando-se um fator crítico no planejamento de data centers para cargas de trabalho de IA.

A Potência Térmica de Design (TDP) é a quantidade máxima de calor gerada por um chip ou componente de computador. A GPU NVIDIA H100 de 80 GB, por exemplo, tem um TDP de 350 watts, o que significa que gera uma quantidade significativa de calor durante a operação, especialmente em cargas de trabalho de IA, onde a GPU é muito utilizada. O resfriamento adequado é essencial para garantir que a GPU tenha um desempenho ideal sem superaquecimento.

Os servidores Cisco UCS foram projetados em estreita colaboração com a NVIDIA para lidar com as altas demandas térmicas de GPUs modernas, como a H100. Essa cooperação garante que os servidores sejam equipados com soluções de resfriamento avançadas, capazes de dissipar o calor gerado pelas placas de GPU. Os servidores Cisco UCS utilizam designs de fluxo de ar eficientes e mecanismos de resfriamento otimizados para manter as placas de GPU em temperaturas operacionais seguras, mesmo sob cargas de trabalho intensas. Essa abordagem integrada ajuda a manter o desempenho da GPU e prolonga a vida útil do hardware, evitando potenciais problemas de superaquecimento que podem levar à degradação do desempenho ou falha do hardware.

Ao garantir o resfriamento adequado, os servidores Cisco UCS podem suportar efetivamente as altas demandas de energia dessas GPUs, tornando-os ideais para cargas de trabalho de IA em ambientes de data center.

**Resumo**

Os principais pontos de aprendizado desta atividade incluem a compreensão de quão caro e intensivo em recursos é treinar grandes modelos de IA e por que o uso de modelos pré-treinados pode economizar custos. Você também aprenderá sobre as compensações entre usar infraestrutura em nuvem ou local, com foco na flexibilidade e evitando a dependência excessiva de um único fornecedor. A rede é importante para a IA, necessitando de conexões rápidas e confiáveis, e servidores de alto desempenho exigem um planejamento cuidadoso de energia e resfriamento. Por fim, a sustentabilidade e o gerenciamento fácil da infraestrutura, como o uso de ferramentas da Cisco, são essenciais para uma implantação bem-sucedida da IA.