**Posicionamento de carga de trabalho de IA e interoperabilidade.**

Introdução

Como profissional de TI, arquiteto de soluções, cientista de dados ou engenheiro de dados, é sua responsabilidade implantar cargas de trabalho nos ambientes mais econômicos e de alto desempenho, garantindo que seus aplicativos funcionem sem problemas e sem tempo de inatividade. A capacidade de se adaptar rapidamente às demandas em constante mudança e mover cargas de trabalho para onde elas são necessárias o diferenciará como um líder de TI proativo e engenhoso.

Este treinamento foi desenvolvido especificamente para profissionais como você, responsáveis ​​por gerenciar e otimizar cargas de trabalho de IA em diversos ambientes. Seja lidando com data centers locais, nuvens públicas ou modelos híbridos, o conteúdo deste treinamento fornecerá o conhecimento e as ferramentas para navegar por essas complexidades com eficácia.

Neste treinamento, você aprenderá a mover cargas de trabalho de IA dinamicamente para otimizar custos, aprimorar o desempenho e garantir a conformidade com os requisitos regulatórios. Você explorará estratégias para implementar ambientes multicloud para evitar a dependência de fornecedores e aplicar os melhores recursos de diferentes provedores de nuvem.

Ao final deste treinamento, você será capaz de:

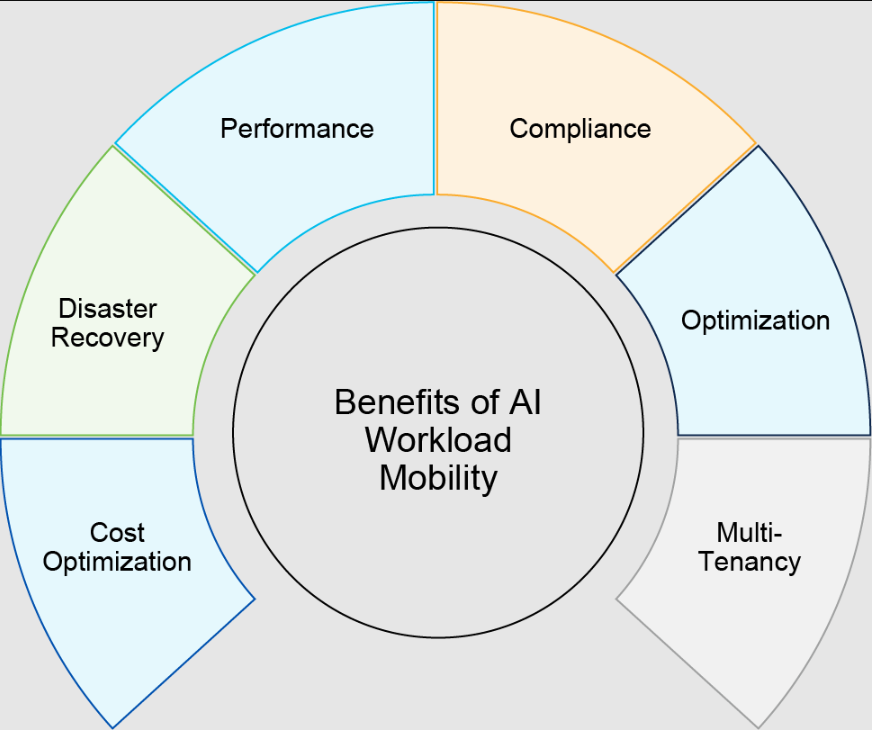
* Aplique técnicas de mobilidade de carga de trabalho para mover dinamicamente cargas de trabalho de IA entre diferentes ambientes, garantindo otimização de custos e continuidade dos negócios.
* Avalie estratégias multicloud para selecionar os melhores provedores de nuvem para tarefas específicas de IA, melhorando o desempenho e a conformidade.
* Crie um plano robusto para mitigar os riscos de dependência de fornecedores, garantindo que sua organização permaneça flexível e adaptável às mudanças nas necessidades dos negócios.

As habilidades que este treinamento fornece capacitarão você a otimizar suas cargas de trabalho de IA, garantindo que sua organização permaneça competitiva e resiliente em um cenário tecnológico em rápida evolução.

**Mobilidade de Carga de Trabalho. (Workload Mobility)**

A capacidade de mover dinamicamente "máquinas virtuais", ou servidores, entre diferentes hosts físicos, dentro ou entre data centers, não apenas alterou os requisitos de rede, mas também o paradigma operacional dos data centers. Hoje, a maioria das grandes organizações possui vários data centers globalmente e conexões de nuvem pública a partir de seus data centers. Essa implantação permite que as equipes de data center operem os ambientes como se fossem um provedor de serviços para sua organização. Essas equipes podem provisionar recursos dinamicamente com base nas mudanças nas necessidades operacionais. Torna-se irrelevante para o usuário onde o aplicativo reside se as expectativas de desempenho e continuidade dos negócios são atendidas.

Essa flexibilidade com os recursos do data center continuou com o crescimento da IA ​​e da mobilidade de cargas de trabalho. A mobilidade de cargas de trabalho de IA envolve a transferência do modelo de IA, de um aplicativo de inferência de IA ou de dados consumidos ou produzidos a partir de um modelo. Esses métodos de movimentação podem ser manuais ou automatizados, dependendo dos serviços incorporados à plataforma. Por exemplo, o Kubernetes e a VMware possuem recursos para mobilidade de cargas de trabalho manuais e automatizadas incorporados em seus produtos.



A ideia de mover cargas de trabalho dinamicamente é comum e por um bom motivo. A mobilidade da carga de trabalho da IA ​​faz sentido comercialmente pelos seguintes motivos:

* **Otimização de Custos** – Mover cargas de trabalho entre ambientes pode reduzir as despesas operacionais, utilizando recursos mais baratos (como energia, capacidade computacional, etc.) fora dos horários de pico ou aproveitando as variações de preço entre provedores de nuvem pública. Projetar para cargas de trabalho móveis também pode evitar o aprisionamento em nuvem pública, também conhecido como aprisionamento em fornecedor.
* **Recuperação de Desastres e Balanceamento de Carga** – Em caso de falha parcial ou total do data center, a movimentação de cargas de trabalho garante a continuidade dos negócios. As operadoras também têm a opção de balancear a demanda entre os data centers ou garantir que a aplicação seja hospedada na região mais próxima dos usuários. Esse recurso é importante quando a proximidade dos recursos muda conforme os fusos horários ficam online e offline durante seus respectivos horários comerciais.
* **Desempenho** – À medida que as cargas de trabalho se aproximam das fontes de usuários, a latência pode ser reduzida, proporcionando melhores experiências ao usuário. As organizações também podem usar serviços elásticos que dimensionam automaticamente as cargas de trabalho, tanto vertical quanto horizontalmente, com base na demanda.
* **Conformidade** – Muitas organizações estão sujeitas a regulamentações rígidas que determinam onde os dados podem ser armazenados e acessados. Mover essas cargas de trabalho pode ajudar a atender a requisitos específicos de segurança ou auditoria, transferindo-as para ambientes com padrões ou certificações de segurança mais elevados.
* **Otimização** – As cargas de trabalho podem ser movidas para ambientes mais adequados às suas necessidades específicas. No caso de cargas de trabalho de IA, compreender as demandas das estruturas de treinamento e inferência permite que os operadores otimizem o posicionamento. Com o advento da computação de ponta, as cargas de trabalho de IA são cada vez mais implantadas na borda, mais próximas de onde os dados são gerados. E como o treinamento de modelos de IA exige muitos recursos, a inferência de IA é mais adequada para locais de ponta, onde as aplicações podem ser movidas para diferentes locais de ponta dependendo das demandas em tempo real.
* **Multilocação** – Utilizando técnicas de virtualização oferecidas por soluções SDN, os recursos de IA dentro das fabrics podem ser alocados e atribuídos a diferentes unidades de negócios. A capacidade de mover tarefas de treinamento ou aplicações de inferência entre locais ou dentro de um data center multilocatário proporciona flexibilidade. Isso permite que as operadoras maximizem os recursos disponíveis e suportem cargas de trabalho de IA simultâneas dentro de um cluster onde os recursos são atribuídos e instanciados virtualmente.

**Métodos de mobilidade de carga de trabalho de IA**

Tanto a mobilidade manual quanto a automatizada da carga de trabalho com IA são opções adequadas, dependendo das necessidades específicas de uma organização. Cada método tem suas vantagens, desvantagens e um conjunto de ferramentas para mover cargas de trabalho.

**Métodos Manuais**

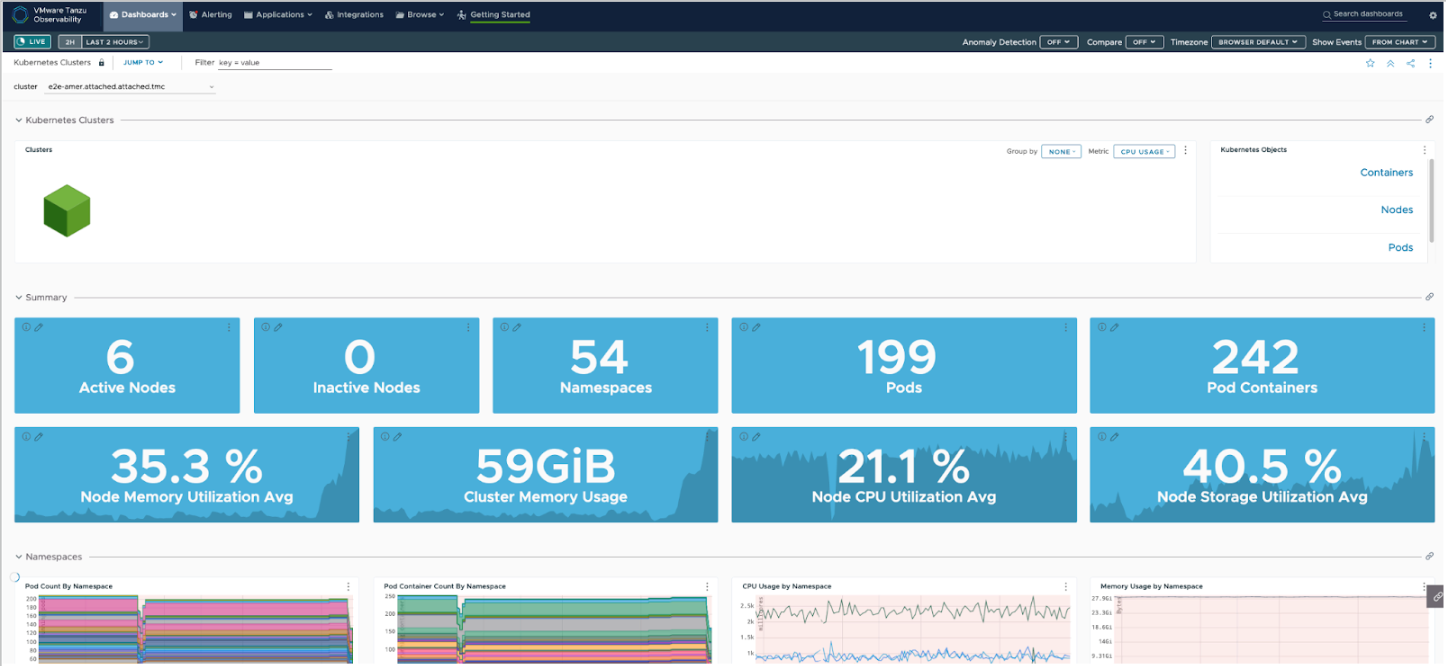
Abordagens manuais para mobilidade de carga de trabalho proporcionam maior controle e personalização, mas ao custo de tempo e potenciais erros. Métodos manuais utilizam ferramentas de transferência de dados, como o Secure Copy Protocol (SCP), ou até mesmo serviços de armazenamento em nuvem, como o AWS S3 ou outros serviços de armazenamento em nuvem pública. A capacidade de implantar e mover manualmente aplicativos em contêineres está incorporada em plataformas como Docker, Rancher e Kubernetes. Da mesma forma, plataformas de hipervisor, como a VMware, também oferecem suporte a opções de migração manual que utilizam tecnologias como vMotion ou snapshots manuais. Também é possível extrair manualmente o código para implantar modelos de repositórios como GitHub ou GitLab para diferentes ambientes.

**Métodos Automatizados**

Abordagens automatizadas para mobilidade de carga de trabalho oferecem eficiência e escalabilidade, mas exigem investimento em ferramentas e expertise. Diversas soluções oferecem diversas opções de automação.

Tecnologias de orquestração incluem plataformas como o Kubernetes, que automatiza a implantação, o escalonamento e o gerenciamento de cargas de trabalho de IA em contêineres. Ferramentas como o Kubeflow estendem o Kubernetes especificamente para as nuances dos fluxos de trabalho de IA. O OpenShift da Red Hat, desenvolvido no Kubernetes, é uma ferramenta de orquestração de contêineres e gerenciamento de clusters. E o Apache Mesos é uma plataforma semelhante de código aberto para orquestração que pode gerenciar cargas de trabalho de IA em sistemas distribuídos. O Apache Airflow é outra ferramenta de orquestração de fluxo de trabalho de código aberto para criação, agendamento e monitoramento programático de pipelines de dados. O Airflow é usado para automatizar processos de ETL (extração, transformação, carregamento) para ingestão e pré-processamento de dados e é particularmente valioso para gerenciar fluxos de trabalho complexos.

* O Apache Spark, por outro lado, é um poderoso sistema de computação distribuída de código aberto que fornece uma estrutura de processamento de dados em memória. É amplamente utilizado para tarefas de ETL em larga escala. Ele suporta um conjunto muito grande de transformações de dados e é usado para preparar os dados que serão inseridos no processo de treinamento de um modelo de IA/ML. O Spark também se integra bem a bibliotecas como a MLlib, que fornece um conjunto abrangente de ferramentas para gerenciar algoritmos, transformação de recursos, avaliação e ajuste de modelos e muito mais. Ele pode ser integrado a diversas fontes de dados, como AWS Simple Storage Service (S3), Hadoop Distributed File System (HDFS) ou outras ferramentas de big data.
* Provedores de nuvem pública oferecem conjuntos de ferramentas voltados para as comunidades de ciência de dados e IA/ML para serviços de ETL. Essas ferramentas incluem o Glue da AWS, o Dataflow da GCP ou o Data Factory do Azure. Há também fornecedores como Databricks e Snowflake, que oferecem plataformas de análise de dados baseadas em nuvem e data warehouse como serviço.
* O VMware Tanzu fornece ferramentas para implantar e gerenciar cargas de trabalho de IA em ambientes híbridos, nos quais as organizações possuem e operam seus próprios data centers, além de utilizar serviços de nuvem pública. O painel do Tanzu fornece uma visão geral da utilização de recursos nos níveis de cluster, nó, namespace e contêiner. O painel fornece dados que podem ser usados ​​para utilização de recursos e planejamento de capacidade em clusters, nós e pods.



O Google Anthos oferece um recurso semelhante, facilitando a mobilidade entre provedores locais, GCP (Google Cloud Platform) e outros provedores de nuvem. O Azure Arc gerencia cargas de trabalho de IA em data centers, locais de borda e nuvens públicas.

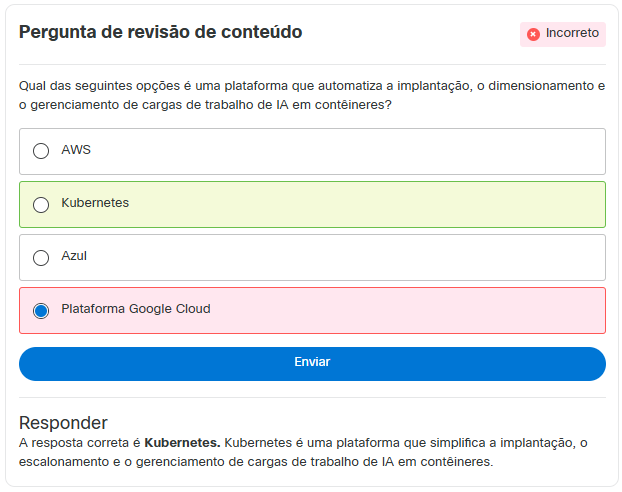
Em relação à borda, o NetApp Data Fabric oferece uma estrutura unificada de gerenciamento de dados e a capacidade de mover cargas de trabalho entre o local, a borda e a nuvem. O NVIDIA EGX também oferece suporte à implantação de cargas de trabalho de IA na borda e à mobilidade por meio da integração em redes de nuvem híbrida.

Plataformas arquitetônicas sem servidor, também conhecidas como Função como Serviço (FaaS), como AWS Lambda, GCP Functions ou Azure Functions, podem permitir que cargas de trabalho de inferência de IA sejam acionadas automaticamente com base em eventos. O FaaS facilita a mobilidade do código que executa microsserviços sem intervenção manual.

Algumas plataformas de AIOps, como Splunk ou Dynatrace, podem automatizar a movimentação de cargas de trabalho com base em análises em tempo real do desempenho e da disponibilidade da infraestrutura. Algumas ferramentas de gerenciamento de modelos de IA, como TensorFlow Serving ou TorchServe, também oferecem suporte para automatizar a implantação de modelos entre ambientes locais, na nuvem e na borda. O TorchServe foi desenvolvido para modelos baseados em PyTorch, que podem ser usados ​​para implantar e gerenciar modelos em diferentes ambientes.

**Observação**

É importante garantir que os serviços de rede estejam configurados para permitir a mobilidade da carga de trabalho de IA. Tecnologias como VXLAN EVPN e SD-WAN permitem a instanciação de serviços definidos por software sob demanda, o que ajuda a isolar e priorizar com segurança os fluxos de dados em grandes áreas geográficas.

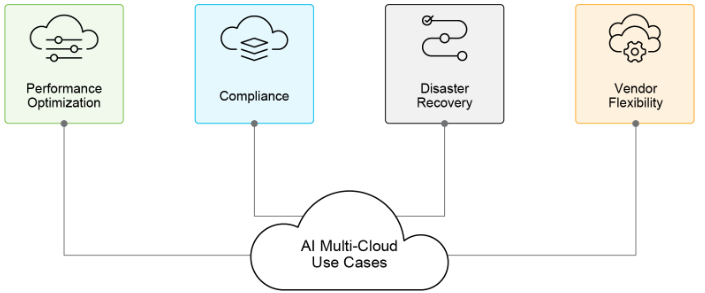


**Implementação Multi-Cloud (Multi-Cloud Implementation)**

O uso de múltiplos provedores de nuvem para cargas de trabalho de IA também está se tornando cada vez mais comum e viável para as empresas. Os serviços em nuvem são flexíveis e escaláveis, oferecendo acesso rápido a tecnologias de ponta. A decisão de usar serviços em nuvem em vez de construir um data center local para treinamento de modelos de IA ou aplicações de inferência depende de vários fatores.

**Casos de uso de IA multi-nuvem**

Com uma implantação de IA multicloud, cada provedor de nuvem pode se destacar em uma área específica. Ao selecionar a melhor nuvem para cada tarefa específica, as organizações podem otimizar o desempenho, a latência e o custo. Por exemplo, um provedor pode oferecer ferramentas de IA e aprendizado de máquina superiores, enquanto outro pode fornecer os melhores serviços de computação sem servidor.



Uma estratégia de IA multicloud ajuda a garantir que as organizações atendam a regulamentações diversas. Ter múltiplas nuvens permite que as organizações utilizem provedores de nuvem com data centers em localizações geográficas específicas, garantindo a conformidade com as leis de residência e soberania de dados.

Estratégias multi-nuvem oferecem recursos de recuperação de desastres, particularmente úteis para aplicações de inferência de IA com requisitos de tempo de atividade exigentes. Múltiplos provedores aumentam a resiliência e garantem a continuidade das operações em caso de interrupção ou falha em um dos ambientes de nuvem.

As empresas costumam usar estratégias multicloud para evitar a dependência de um único fornecedor. Distribuir cargas de trabalho entre várias nuvens reduz a dependência de um único fornecedor, oferecendo maior flexibilidade e poder de negociação. As organizações também podem aproveitar as diferenças de preço entre fornecedores ou entre regiões.

No entanto, como a multicloud está se tornando mais comum nas operações gerais de TI, é muito menos comum usar vários provedores de nuvem para treinar modelos de IA. Em casos raros, partes de um modelo podem ser treinadas em paralelo. O uso de ambientes de diferentes provedores proporciona otimização de custos com base nos pontos fortes de cada um para facilitar o treinamento de modelos específicos. Lembre-se de que visão computacional, processamento de linguagem natural, detecção de fraudes, recomendações de usuários, modelos de linguagem de grande porte e muito mais têm requisitos diferentes para computação, armazenamento e rede. As ofertas de hardware, os contratos de nível de serviço e os preços variam entre os principais provedores.

Além disso, uma implantação multicloud de IA exige que você garanta que suas ferramentas de IA, pipelines de dados e fluxos de trabalho sejam compatíveis entre os provedores e funcionem perfeitamente em diferentes ambientes. Por exemplo, ao depender de um serviço pré-configurado de propriedade de um único provedor, essa oferta de serviço funcionará apenas dentro do provedor de nuvem que a oferece.

**Gerenciamento de dados multi-nuvem de IA**

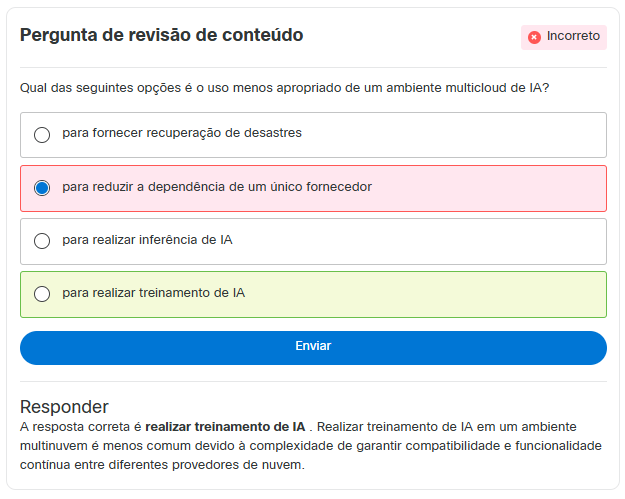
O gerenciamento eficaz de dados é essencial em um ambiente multinuvem, especialmente para cargas de trabalho de IA que dependem fortemente de grandes conjuntos de dados.

Interface gráfica do usuário, Texto, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

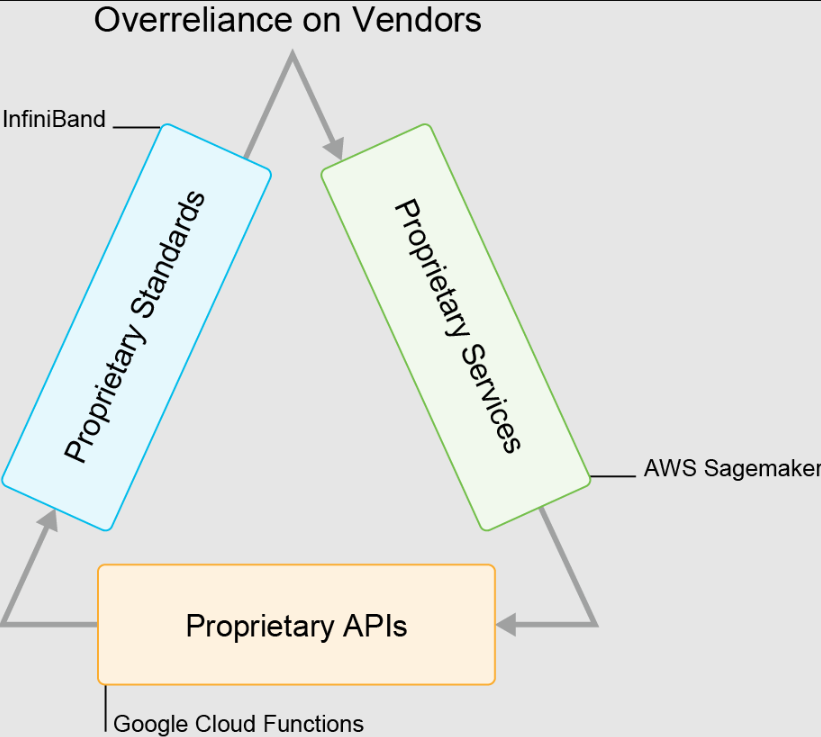
A lista a seguir descreve as diretrizes para garantir portabilidade, consistência, segurança e eficiência de custos de dados em diversas plataformas de nuvem:

* Portabilidade e interoperabilidade de dados
  1. **Sincronização de dados** : garanta a consistência dos dados entre várias nuvens por meio de ferramentas de sincronização, como Apache Kafka, AWS DataSync e Google Cloud Dataflow, para ajudar a manter a integridade da sincronização.
  2. **Abstração de armazenamento** : use ferramentas como o NetApp Cloud Volumes, uma solução de armazenamento que oferece suporte a diversos ambientes e, ao mesmo tempo, fornece gerenciamento perfeito de todos eles.
* Estratégias de Armazenamento de Dados
  1. **Localização de Dados** : Armazene os dados na região da nuvem mais próxima da carga de trabalho de inferência de IA para minimizar a latência e os custos de transferência de dados. Utilize soluções de armazenamento multirregionais oferecidas por provedores de nuvem, garantindo que os dados sejam replicados entre regiões, se necessário.
  2. **Soluções de armazenamento híbrido** : considere usar modelos de armazenamento híbrido que permitam que os dados residam no local e na nuvem para fornecer flexibilidade no gerenciamento de dados e na conformidade.
* Consistência e sincronização de dados
  1. **Bancos de dados distribuídos** : use bancos de dados distribuídos que suportam ambientes multicloud, como Google Spanner ou AWS Aurora, para garantir a consistência dos dados em todas as plataformas de nuvem.
  2. **Ferramentas de sincronização de dados** : empregue ferramentas de sincronização de dados que possam gerenciar a consistência de dados em diferentes sistemas de armazenamento em nuvem, garantindo que todos os nós de inferência de IA tenham acesso aos dados mais atualizados.
* Segurança e conformidade de dados
  1. **Criptografia** : Garanta que os dados sejam criptografados tanto em trânsito quanto em repouso em todos os ambientes de nuvem. Use padrões de criptografia consistentes e serviços de gerenciamento de chaves para manter a segurança.
  2. **Gerenciamento de conformidade** : implemente ferramentas de governança de dados que ajudem a aplicar políticas de conformidade em vários ambientes de nuvem para garantir que os requisitos de residência de dados sejam atendidos e que as trilhas de auditoria para acesso e modificações de dados sejam concluídas.
  3. **Controle de acesso** : utilize soluções de gerenciamento de identidade e acesso (IAM) que podem operar em várias nuvens, garantindo que o acesso aos dados e serviços de inferência de IA seja rigorosamente controlado e consistente em todos os ambientes.
* Otimização de custos
  1. **Custos de Transferência de Dados** : Esteja ciente dos custos associados à transferência de dados entre ambientes de nuvem. Minimize a movimentação desnecessária de dados e considere o uso de técnicas de compactação de dados para reduzir o tamanho das transferências.
  2. **Armazenamento em camadas** : implemente estratégias de armazenamento em camadas que coloquem dados acessados ​​com frequência em armazenamento de alto desempenho e arquivem dados menos críticos em soluções de armazenamento de longo prazo e econômicas.



**Riscos de bloqueio de fornecedor. (Vendor Lock-In Risks)**

O aprisionamento de fornecedor ocorre quando uma empresa se torna excessivamente dependente de um único fornecedor para seus produtos ou serviços, dificultando a migração para outro sem incorrer em custos significativos, interrupções operacionais ou perda de dados. Muitas vezes, padrões abertos são embelezados, criando um uso proprietário de um determinado protocolo. Quando uma equipe se torna fluente e dependente de um determinado conjunto de recursos ou ferramentas, torna-se cada vez mais difícil migrar para uma nova ferramenta.



No caso de data centers habilitados para IA, o aprisionamento de fornecedores pode ocorrer quando uma empresa depende fortemente do hardware e software de IA, das ferramentas de IA ou dos serviços de nuvem de um determinado fornecedor para construir e operar sua infraestrutura de IA. Por exemplo, a NVIDIA detém um controle de mercado significativo com o InfiniBand para definir padrões e impulsionar inovações em ambientes de computação de alto desempenho para conectar recursos de rede. Esse domínio a torna a primeira escolha para muitas organizações que constroem infraestrutura específica para IA. Embora existam outros players no mercado de redes, a posição da NVIDIA com o InfiniBand é amplamente incontestável, tornando-o essencialmente um padrão "proprietário".

Dependências e limitações do fornecedor

As dependências de fornecedores podem surgir em diferentes contextos: a tecnologia em uso, os formatos de dados ou API, as habilidades e o conhecimento da equipe ou até mesmo os detalhes de contratos de longo prazo. A lista a seguir descreve diversas dependências de fornecedores de IA e o impacto que elas têm nas organizações que implementam um data center habilitado para IA.

**Explore as dependências e limitações do fornecedor**

Clique em cada aba para saber mais sobre cada uma delas.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Texto

O conteúdo gerado por IA pode estar incorreto.

Interface gráfica do usuário, Texto, Aplicativo, Email

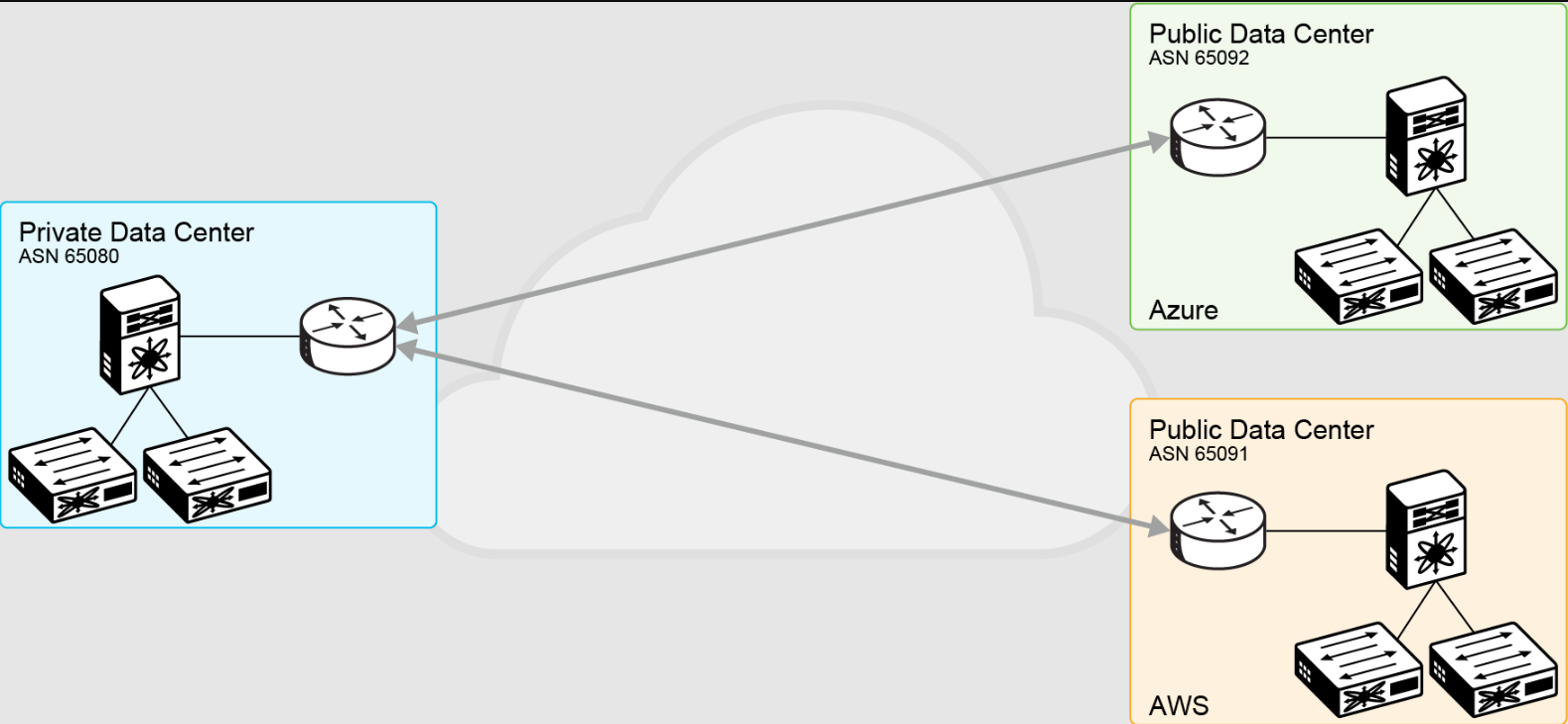
O conteúdo gerado por IA pode estar incorreto.

**Mitigação de bloqueio de fornecedor - Vendor Lock-In Mitigation.**

À medida que as necessidades dos negócios mudam, a capacidade de adaptação a diferentes plataformas ou provedores torna-se relevante. Ao mitigar proativamente o aprisionamento de fornecedores, as organizações que possuem aplicativos habilitados para IA podem manter maior controle sem serem restringidas pelas limitações ou preços de um único fornecedor.

**Interoperabilidade do centro de dados de IA**

A interoperabilidade desempenha um papel vital na mitigação da dependência de fornecedores, pois permite que diferentes sistemas, aplicativos e serviços funcionem em conjunto perfeitamente, independentemente da localização dos data centers. Seja implantando data centers privados, utilizando a nuvem pública ou um modelo híbrido, é importante garantir que as tecnologias utilizadas no data center privado suportem as mesmas plataformas de software oferecidas pelos outros data centers ou provedores de nuvem.



Quando implementados corretamente, os serviços de rede fornecidos por um provedor de nuvem emulam um ambiente de data center privado. Essas instâncias de "data centers privados virtuais" hospedados na nuvem parecem data centers remotos para as equipes organizacionais que gerenciam ambientes híbridos e multicloud. O aprisionamento de fornecedores pode ser mitigado pela implementação de redundância, seja na forma de um cenário de balanceamento de carga ativo-ativo ou de um cenário de redundância ativo-em espera entre provedores. Ao monitorar continuamente os custos e o desempenho entre provedores, as organizações podem obter o melhor valor e permanecer protegidas contra o excesso de dependência.

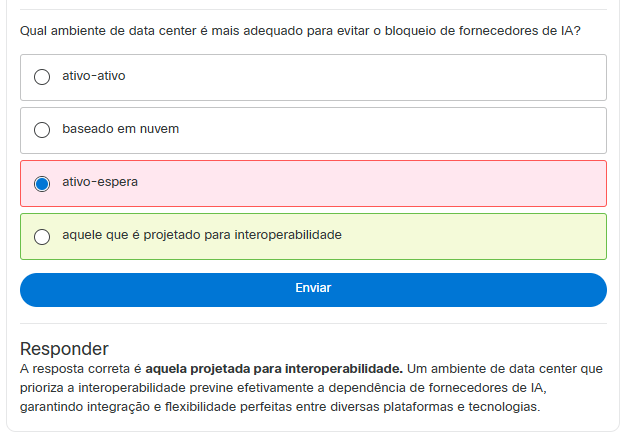
Evitar a dependência de fornecedores em data centers de IA exige uma estratégia proativa. A interoperabilidade garante que os sistemas de uma organização sejam compatíveis com uma ampla gama de tecnologias, permitindo fácil migração, integração e colaboração entre diferentes plataformas.

Diagrama

O conteúdo gerado por IA pode estar incorreto.

Considere as seguintes diretrizes para promover a interoperabilidade do data center habilitado para IA:

* **Use padrões abertos** – priorize plataformas e ferramentas que sigam padrões suportados por vários fornecedores e que sejam testados para serem interoperáveis.
* **Use soluções de código aberto** – Muitas ferramentas de código aberto são incorporadas às soluções dos fornecedores. Por exemplo, PyTorch e TensorFlow são comumente usados ​​em plataformas de fornecedores para treinar modelos de aprendizado profundo. A funcionalidade do PyTorch será a mesma, independentemente de estar presente em um conjunto de ferramentas NVIDIA ou Red Hat.
* **Desenvolver expertise interna** – Invista em pessoal com expertise em um conjunto diversificado de tecnologias de data center de IA. Por exemplo, entender RDMA e RoCEv2 é valioso em qualquer data center de IA, tradicional ou moderno. Capacite equipes em diversas plataformas e ferramentas. Desenvolva equipes de data center de IA que reforcem os valores de código aberto e a cultura organizacional necessária para manter um ambiente multifornecedor.
* **Use Camadas de Abstração** – Desvincule a infraestrutura física dos aplicativos em execução. Por exemplo, implemente soluções que permitam substituir o hardware de um fornecedor por outro, dentro da mesma estrutura, sem alterar a plataforma operacional e suas ferramentas. Ferramentas como o Terraform abstraem os detalhes de configuração específicos de cada fornecedor e permitem que as operadoras empreguem uma abordagem de Infraestrutura como Código. Quando aliada à automação, a Infraestrutura como Código facilita a rápida implantação e reimplantação de serviços de rede em data centers e provedores de nuvem.



Resumo

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

* Como sua organização pode se beneficiar da implementação da mobilidade da carga de trabalho de IA e quais ferramentas ou plataformas específicas seriam mais adequadas para sua infraestrutura atual?
  1. Considere as tecnologias de orquestração manual e automatizada para mobilidade de carga de trabalho que permitem a movimentação dinâmica de cargas de trabalho de IA em diferentes ambientes para otimizar custos, garantir recuperação de desastres, melhorar o desempenho e atender aos requisitos de conformidade.
* Quais são os riscos potenciais de dependência de fornecedor na configuração atual do seu data center de IA e quais medidas você pode tomar para mitigar esses riscos e garantir maior flexibilidade e controle?
  1. Considere as dependências de tecnologia, formatos de dados, capital humano e contratos de longo prazo que podem dificultar a troca de provedores.
* Como uma estratégia multicloud pode melhorar seu gerenciamento de carga de trabalho de IA e quais desafios você pode enfrentar ao implementar tal estratégia?
  1. Considere os requisitos de conformidade e as necessidades de desempenho da sua organização.