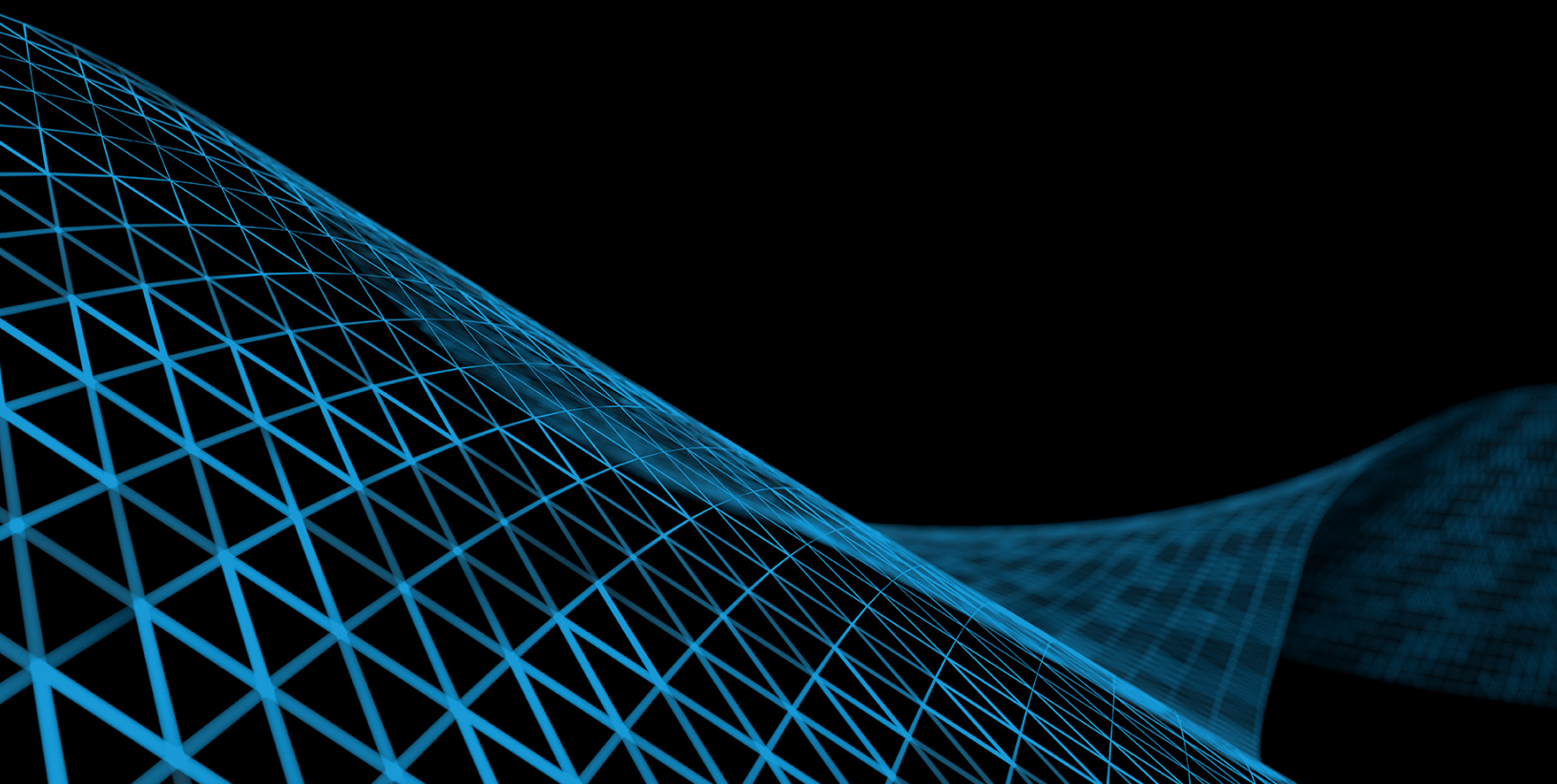
TechTarget | Custom Media
E-Book

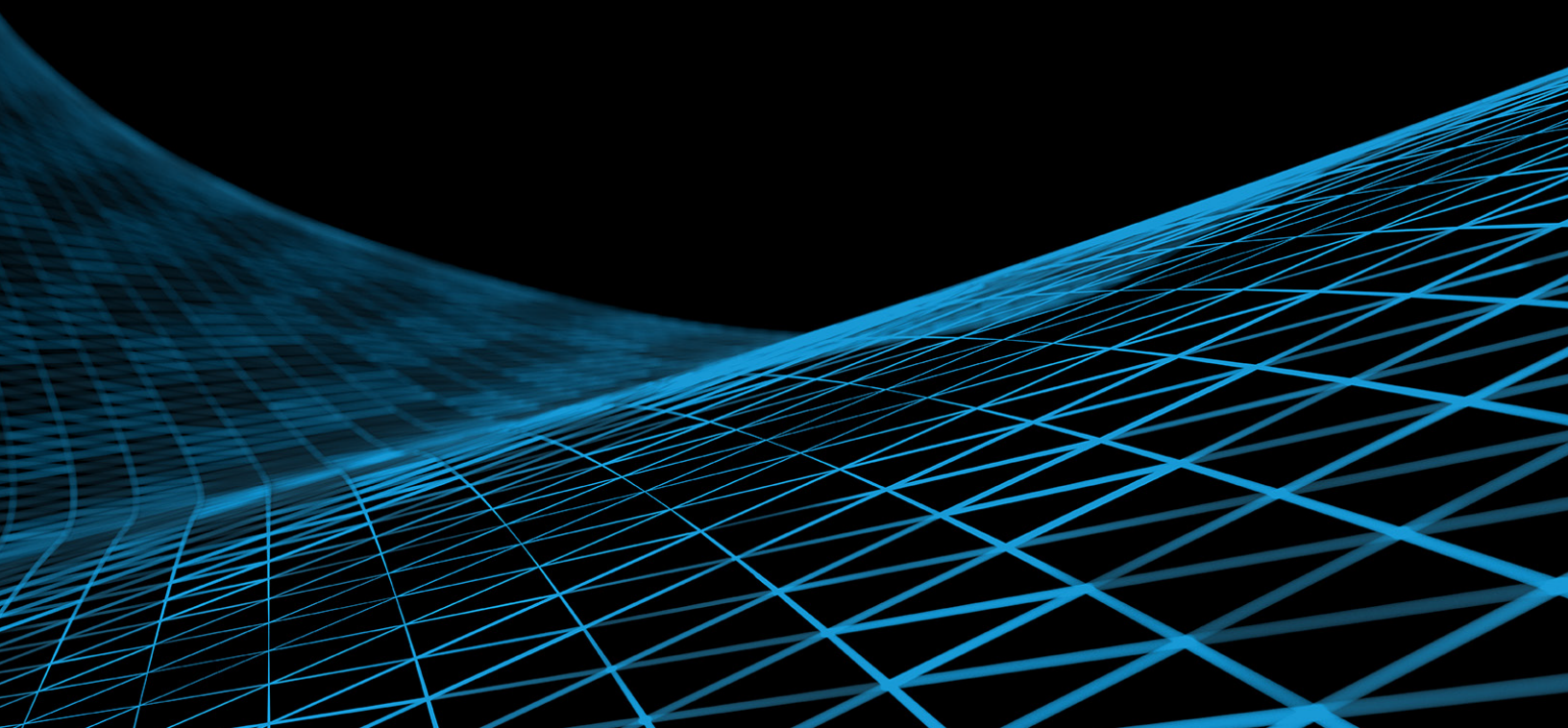
Como a IA está transformando data centers, edge, nuvem e muito mais

O impacto transformador da IA não pode ser superestimado. Ela está mudando a natureza de como a edge computing e a nuvem agregam valor às organizações de todos os setores, regiões e casos de uso. Este eBook apresenta uma perspectiva prática sobre o papel cada vez mais essencial da IA na nuvem distribuída.

Akamai



Sumário



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Introdução

Por que este eBook é importante

Se você é responsável pelas decisões de TI, membro da diretoria executiva, parte interessada na linha de negócios ou membro do Conselho, é muito provável que esteja gastando cada vez mais tempo lendo, falando e pensando sobre o impacto da inteligência artificial em sua organização e, até mesmo, na sua vida pessoal e profissional.

A IA progrediu a passos largos no caminho do desenvolvimento de mercado muito mais rápido que a maioria das tecnologias mais modernas que vieram antes dela. Ela é uma das poucas tecnologias importantes dos últimos anos em que o entusiasmo de mercado inicial realmente foi justificado e na qual foi possível observar resultados tangíveis. Vamos compartilhar alguns números que ilustram esse importante fato posteriormente.

No entanto, este eBook não apenas coloca o crescimento e o potencial da IA em um contexto real. Ele analisa a IA por meio de uma lente importante: a IA perto dos usuários e na nuvem distribuída. Muito se falava sobre a importância de ter os dados perto dos usuários e na nuvem; agora, entramos na era da inferência descentralizada da IA e da IA na nuvem distribuída.

O que nos traz de volta ao título: por que este eBook é importante. Oferecemos quatro motivos pelos quais responsáveis pela tomada de decisões, influenciadores e controladores precisam ler este eBook:

1. **A edge computing e a computação em nuvem são as novas fronteiras da inovação de IA e da criação de valor.** Ao criar e implantar recursos de IA fora do data center, as organizações podem colher os benefícios de novos insights, modelos de negócios e soluções que acompanham o mesmo caminho que o mercado de TI mais amplo tomou na última década.
2. **A IA é um gigante da governança e regulação.** Se você pensou que era difícil seguir diretrizes de uso e práticas de “IA responsável” quando a IA era implantada de forma geral em sandboxes locais, considere agora a conformidade, a proteção de dados e os problemas de segurança da IA perto dos usuários e na nuvem distribuída. Sua organização precisa blindar essa área ou então correrá o risco de enfrentar muitos problemas.
3. **O impacto da IA sobre os padrões e práticas da força de trabalho em constante mudança será ainda maior do que o esperado.** Embora quase não haja dúvidas de que muitas funções de TI adotarão um caminho mais direcionado à IA e deixarão funcionários humanos de lado, tirar a IA do data center criará oportunidades empolgantes e até mesmo inspiradoras para trabalhadores administrativos e de TI.
4. **Não há tempo a perder.** As iniciativas de IA estão indo para o topo da lista de prioridades estratégicas de todas as organizações. Se sua organização não aproveitou totalmente a IA no local para testar novos casos de uso e saber mais sobre o que a IA pode fazer, você com certeza ficou para trás. Mas aproximar a IA dos usuários e na nuvem distribuída representa uma nova chance.

Mais um motivo pelo qual acreditamos que você encontrará valor neste eBook: no final, oferecemos conselhos práticos para que sua organização possa aproveitar ao máximo a mudança para a IA na nuvem distribuída.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Capítulo 1

A revolução da IA e como ela já mudou a dinâmica do mercado

Como a IA já existe há décadas, é justo chamar esta era atual de "Revolução da IA"? Ou a IA está simplesmente evoluindo, embora em um ritmo acelerado, com base em suas formas e funções anteriores? Essas não são perguntas retóricas: nosso setor tende a falar sobre revoluções técnicas que surgem aparentemente da noite para o dia, que não foram detectadas pela massa, apenas por especialistas técnicos que já fizeram experiências com isso em laboratórios.

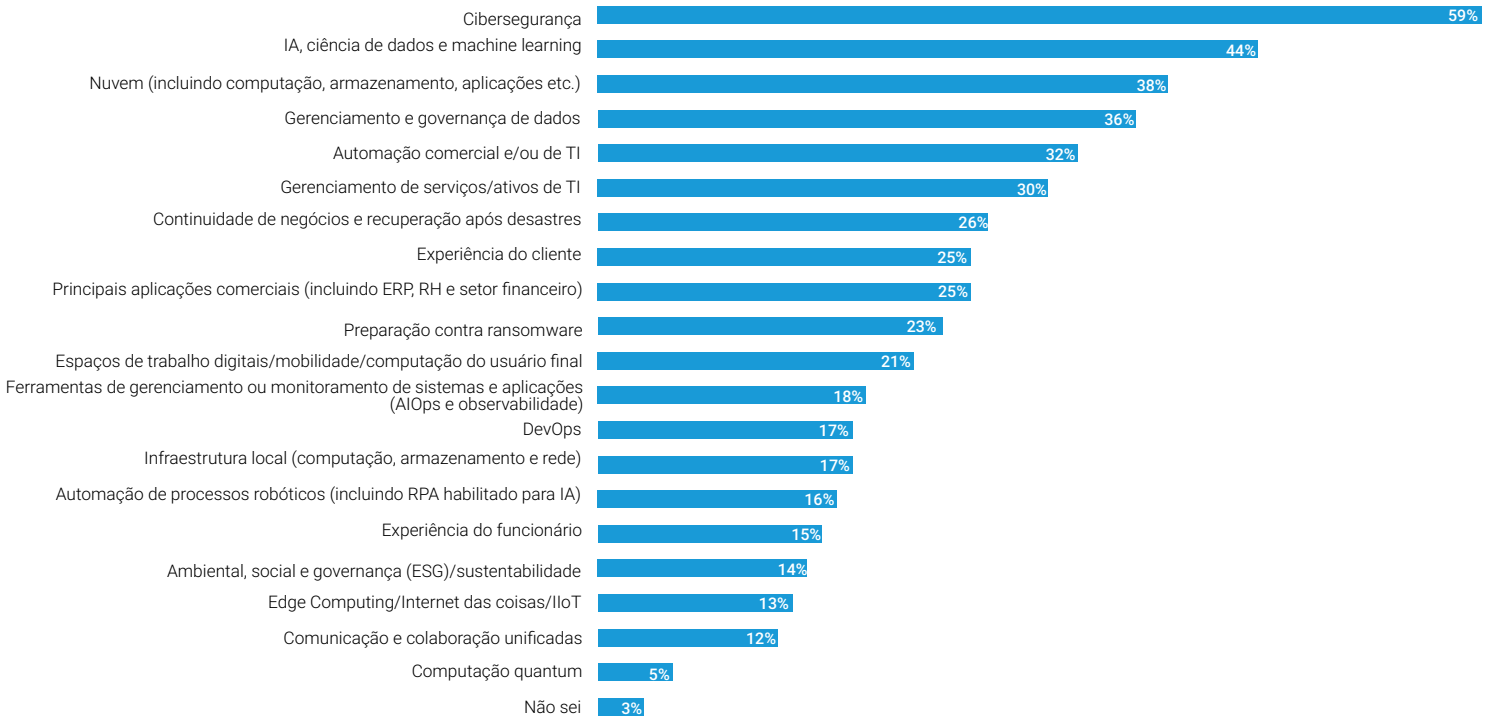
Mas sim, é justo chamar isso de revolução da IA por várias razões. Primeiro, não é possível negar o crescimento impressionante do mercado de IA, que acompanhou a importância que líderes de organizações agora colocam na tecnologia. Suas taxas de crescimento e o impacto no mercado atual ultrapassam muito os movimentos tecnológicos considerados marcos importantes no desenvolvimento do setor da TI, como computadores pessoais, redes locais, computação em nuvem e Internet das coisas.

Esses pontos de dados ilustram o impacto, a importância e o potencial da revolução da IA:

- Os gastos globais com hardware, software e serviços relacionados à IA ultrapassaram os US$ 600 bilhões até o final de 2024 e vão disparar para mais de US$ 2,7 trilhões até 2032, o que representa uma taxa de crescimento anual composta de mais de 20%.[1]
- Cerca de 97% dos líderes comerciais afirmam que suas organizações estão obtendo um retorno positivo sobre seus investimentos em IA.[2]
- Em apenas um ano, a porcentagem de CEOs que colocaram a IA como uma das prioridades tecnológicas de suas empresas pulou de 4% para 24%.[3]
- Como mostrado na figura abaixo, nos últimos dois anos, quase metade das empresas dos Estados Unidos afirmam que a IA é cada vez mais importante para o futuro da organização, sendo superada apenas pela cibersegurança como uma iniciativa corporativa.[4]

Figura 1. A segurança lidera as principais iniciativas tecnológicas, incluindo IA, nuvem e gerenciamento de dados

Quais das iniciativas tecnológicas abrangentes a seguir se tornaram significativamente mais importantes para o futuro da sua organização nos últimos dois anos? (Porcentagem de participantes: N = 1.351, salvo exceções para múltiplas respostas)



1 Fonte: Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032", dezembro de 2024.
2 Fonte: Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks", EY.com, 10 de dezembro de 2024.
3 Fonte: LinkedIn, Gartner's 2024 CEO survey reveals AI as top strategic priority, maio de 2024.
4 Fonte: Resultados da pesquisa completa do Enterprise Strategy Group, *2025 Technology Spending Intentions Survey*, dezembro de 2024.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

Na verdade, o aumento na adoção da IA e no investimento de capital em IA superou o que era investido na computação em nuvem, um dos segmentos mais vitais e de crescimento mais rápido do cenário da TI. O motivo pelo qual o crescimento da IA está ofuscando o crescimento da nuvem é bem simples: embora a computação em nuvem represente uma maneira poderosa e altamente alavancável de desenvolver e fornecer serviços de TI, a IA está mudando a forma como trabalhamos, vivemos e interagimos em quase todos os aspectos de nossas vidas.

Além disso, a rápida adoção da IA expôs limitações das arquiteturas de computação em nuvem tradicionais. Podemos citar como exemplo os modelos de nuvem centralizados, que apresentam dificuldades para atender às demandas de desempenho intenso da natureza em tempo real e com uso intensivo de dados da IA. De modo geral, a IA mudou drasticamente as expectativas para arquiteturas em nuvem quando se trata de problemas como desempenho, custos, habilidades pessoais, segurança e governança de dados.

O que está impulsionando o crescimento estratosférico do mercado de IA e o impacto no mercado geral? Um dos principais fatores é que, no último ano, as organizações passaram a usar modelos de IA de forma muito mais exigente e sofisticada. Devido ao rápido aumento das vantagens e da facilidade de uso que os modelos de IA proporcionam, além do incrível retorno sobre o investimento que muitas empresas notaram, surgiram casos de uso importantes para reforçar o pensamento inicial sobre como usar a IA e novas formas de pensar sobre como se beneficiar dos modelos de IA.

Por exemplo, considere o seguinte:

- **A previsão de negócios** sempre foi um componente essencial de organizações bem-sucedidas, e a IA tornou essa função mais precisa, oportuna e prática. Atualmente, a previsão de tendências futuras com ferramentas como modelos de IA preditiva é muito mais avançada do que há uma década, quando as organizações adotaram a análise de Big Data, porque os modelos de IA atuais são reforçados com os próprios dados das organizações. Isso permite que executivos de negócios, e não apenas cientistas de dados, analisem grandes volumes de dados com mais rapidez e precisão do que com modelos tradicionais.
- **O resumo de documentos** costumava ser território dos trabalhadores de “colarinho branco”, mas a GenIA (IA generativa) mudou as regras do jogo. Os documentos "long-form" ou de conteúdo longo, ou seja, aqueles com conteúdo complexo e altamente técnico, são revisados, analisados e resumidos automaticamente em uma fração do tempo. Isso não só gera insights melhores, mas também permite que tanto profissionais de TI quanto de negócios possam ser alocados para trabalhos mais inovadores e estratégicos, melhorando a produtividade e a experiência do funcionário.
- **A modelagem de rotatividade** agora é um elemento essencial no gerenciamento do relacionamento com o cliente. Ela ajuda as organizações a tomar decisões melhores e mais relevantes contextualmente para prever como, quando e por que os clientes podem abandonar a empresa (ou por que podem aumentar seu compromisso de vendas).
- **Os chatbots com inteligência artificial** estão transformando o atendimento ao cliente, o gerenciamento de serviços de TI e as funções de recursos humanos, permitindo um autoatendimento inteligente em tempo real em muitas solicitações comuns. Esses chatbots são usados até mesmo para casos de uso de engenharia altamente técnica e ciência da computação, como para geração de código e modelagem de novos recursos.

Esses casos de uso ajudam a gerar muitas outras aplicações de IA, fluxos de trabalho e casos de uso, como detecção de ameaças em cibersegurança, análise de dados/inteligência de negócios, análise competitiva, prototipagem rápida, gerenciamento de conformidade e muito mais.

É importante ressaltar que nenhuma tecnologia nova, nem mesmo uma tão promissora e oportuna quanto a IA, está livre de desafios no desenvolvimento, na implantação e no gerenciamento. Esses desafios podem ser técnicos, como a falta de qualidade de dados, a criação da infraestrutura certa ou a garantia do alto desempenho necessário para conduzir LLMs (modelos de linguagem grande). Eles também podem estar relacionados a negócios ou riscos, como a governança de dados, a garantia de uso responsável da IA ou a superação da grande e crescente lacuna de habilidades de IA.

Mas as evidências preliminares indicam que as organizações compreendem e estão lidando com esses desafios com os recursos corretos e com o apoio do Conselho e da diretoria executiva. No próximo capítulo, explicamos um dos principais desenvolvimentos que tornam a IA mais do que um projeto científico: a mudança do foco da IA do treinamento de modelos para a inferência.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Capítulo 2

Mudança de paradigma da IA: do treinamento à inferência

O treinamento de modelos de IA foi um fator essencial no desenvolvimento do mercado e o foco inicial das organizações nas iniciativas de IA. As organizações rapidamente perceberam que a falta de qualidade de dados, ou talvez a falta do tipo e volume corretos de dados, afetou o desenvolvimento de modelos de IA. Uma pesquisa do Enterprise Strategy Group da Informa TechTarget apontou que 37% das organizações citaram problemas de qualidade de dados como o principal desafio na implementação da IA generativa, um problema atrás apenas da falta de habilidades em IA.[5]

As organizações perceberam que aproveitar seus próprios dados era a melhor maneira de lidar com problemas de qualidade de dados e criar LLMs mais precisos, responsivos e contextualmente cientes. Isso permite a criação de um modelo de IA ao alcance das organizações. Essa abordagem muitas vezes é melhor do que comprar modelos de IA comerciais prontos para uso que foram treinados com dados de terceiros, que podem gerar viés nos dados, erros ou falta de recursos de personalização. Como prática padrão, a maioria das organizações focava o treinamento de LLMs em ambientes de nuvem centralizada, uma abordagem eficaz para a onda inicial de desenvolvimento de LLM.

O treinamento de modelos de IA foi aperfeiçoado e estabilizado em termos de qualidade e integridade. As organizações já estão alocando menos recursos para seus modelos de treinamento internos, principalmente pensando que a otimização e os ajustes da IA reduzem a necessidade de uma infraestrutura com intensa capacidade de computação e centrada na GPU.

Mike Leone, diretor de prática do Enterprise Strategy Group para IA, análise e inteligência comercial, ofereceu uma perspectiva independente sobre essa transição do treinamento para a inferência:

"Conforme os modelos pré-treinados e aplicações de IA em tempo real se tornam mais aperfeiçoados e proeminentes, a inferência agora está desempenhando um papel central na entrega eficiente e econômica do valor prometido pela IA", disse. "Essa mudança de foco permite implantações escaláveis em todos os setores, e uns dos maiores impulsionadores disso são os avanços contínuos nas pilhas de hardware e software. Os fornecedores têm um foco maior em dimensionar e otimizar corretamente a infraestrutura de IA, melhorar a eficiência do modelo e gerenciar os custos operacionais contínuos. É claro que, à medida que a inferência democratiza a disponibilidade da IA para as massas, ela introduz desafios como escalabilidade, preocupações com privacidade e análises regulatórias."

A transição do treinamento de modelos de IA para aplicações de IA orientadas por inferência e casos de uso de alto impacto já está em andamento. Armand Ruiz, vice-presidente da IBM para plataformas de IA, sintetizou de forma poderosa: "Na minha opinião, a inferência dominará as cargas de trabalho de IA. [...] Uma vez que um modelo é treinado, idealmente, ele não precisa mais ser treinado. A inferência, no entanto, é um processo em andamento."[6]

Quando ocorrem transições de mercado, elas inevitavelmente são marcadas por oscilações desproporcionais sobre como e onde o capital é investido. Por exemplo, uma estimativa apontou que cerca de 15% dos gastos com chips de silício de IA são dedicados ao treinamento de IA, 45% para a inferência de IA baseada em data center e os 40% restantes para a inferência baseada na edge.[7]

5 Fonte: Resultados da pesquisa completa do Enterprise Strategy Group, "*The State of the Generative AI Market: Widespread Transformation Continues*", setembro de 2024.

6 Fonte: LinkedIn, Armand Ruiz, dezembro de 2024.

7 Fonte: Jonathan Goldberg, "The AI chip market landscape: Choose your battles carefully", TechSpot.com, 30 de maio de 2023.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


Outra pesquisa apontou para uma tendência similar para a inferência, e um estudo do Google apontou o oposto: o Google indicou que 60% de seu consumo de energia centrado em machine learning é usado para inferência, e os 40% restantes são usados para treinamento.[8]

Não é surpresa que essa mudança do treinamento de modelo de IA para a inferência de IA tenha implicações significativas para a infraestrutura de IA e as arquiteturas de nuvem. Essa transformação tem importância especial para as organizações que buscam o ambiente e a arquitetura em nuvem ideais para usar no desenvolvimento, na implantação e no gerenciamento de aplicações de IA.

Os modelos de nuvem centralizados tradicionais ainda funcionam bem para muitas cargas de trabalho comerciais, mas as demandas de desempenho da IA colocam demandas intensas na arquitetura em nuvem em escala. Esses desafios se traduzirão em problemas frustrantes de latência, em que há uma lacuna perceptível no tempo entre as consultas e respostas.

A infraestrutura de computação também é um ponto-chave que as organizações precisam considerar ao mudar do treinamento de modelos para a inferência de IA. As GPUs centradas em IA receberam reconhecimento por sua capacidade de oferecer a potência necessária para treinar LLMs porque oferecem alto desempenho computacional e largura de banda de memória. Em comparação, a inferência exige uma mentalidade de infraestrutura focada na baixa latência e no uso eficiente de recursos.

No próximo capítulo, analisaremos com mais atenção como as organizações podem ajudar a atender às demandas de desempenho da inferência de IA, colocando recursos de inferência mais perto do usuário e em uma arquitetura em nuvem distribuída.

8 Fonte: “*The AI boom could use a shocking amount of electricity*”, Scientific American, 13 de outubro de 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Capítulo 3

A ascensão da inferência descentralizada de IA e a nuvem distribuída

Como explicado no capítulo anterior, a arquitetura em nuvem centralizada tradicional impõe limites e desafios problemáticos para organizações que buscam otimizar seu uso de IA na criação de aplicações inovadoras. É importante considerar e se preparar para resolver problemas como latência, limitações de largura de banda da rede, garantir o desempenho em escala e complexidade de gerenciamento conforme os requisitos de IA mudam do treinamento de modelo para a inferência.

As aplicações de inferência de IA modernas e de última geração exigem a capacidade de desenvolver, implantar e aproveitar a inferência onde os dados estão. Cada vez mais, isso significa perto dos usuários ou em uma arquitetura de nuvem distribuída ágil, escalável e flexível.

Antes de nos aprofundarmos, vamos esclarecer algumas definições.

- *A inferência de IA descentralizada* é um paradigma para o desenvolvimento de fluxos de trabalho de IA que abrangem data centers centralizados (a nuvem) e dispositivos fora da nuvem que estão mais próximos de itens físicos e seres humanos (a edge). Isso contrasta com a prática mais comum na qual as aplicações de IA são desenvolvidas e executadas inteiramente na nuvem, o que passou a ser chamado de *inteligência artificial em nuvem*. Também difere de abordagens de desenvolvimento de IA mais antigas, nas quais as pessoas criavam algoritmos de IA em desktops e depois os implantavam em outros desktops ou hardwares específicos para tarefas, como ler números de verificação.
- A Akamai, fornecedora líder em plataformas para segurança na nuvem e entrega de conteúdo, define *nuvem distribuída* como "uma forma de computação em nuvem na qual uma empresa utiliza infraestrutura de nuvem pública em várias localizações geográficas, enquanto as operações, a governança e as atualizações são gerenciadas centralmente por um único provedor de serviços de nuvem pública. A distribuição de serviços de nuvem pode ajudar as organizações a atingir objetivos de desempenho e tempo de resposta de aplicações, edge computing, exigências regulamentares ou outras demandas que podem ser atendidas com o fornecimento de serviços de nuvem de locais mais próximos aos usuários finais e dispositivos. O uso da computação em nuvem distribuída foi impulsionado pela ascensão das redes de IoT (Internet das coisas), inteligência artificial e outros casos de uso que processam grandes quantidades de dados em tempo real."

A distribuição de serviços em nuvem pode ajudar as organizações a atingir objetivos de desempenho de aplicação e tempo de resposta, edge computing, exigências normativas ou outras demandas.

Tanto a inferência de IA descentralizada quanto a nuvem distribuída são essenciais para o desenvolvimento e a utilização da inferência de IA. Elas representam formas modernas, eficientes, seguras e confiáveis de superar os desafios que já mencionamos com alto desempenho e em escala, uso eficiente de recursos de computação, armazenamento e rede, bem como a capacidade de superar os problemas de latência associados a grandes conjuntos de dados, compostos muitas vezes de grandes quantidades de dados não estruturados.

Como isso funciona na vida real? Uma empresa buscava tornar a personalização o carro-chefe da sua experiência do cliente e precisava garantir que estava captando dados de todo o espectro de dispositivos e fontes de dados. Essa empresa decidiu usar um chatbot de IA com LLM que distribuiu dados localizados para locais próximos aos usuários. Isso reduziu a latência e melhorou os tempos de resposta, resultando em desempenho muito melhor e um suporte ao cliente mais rápido e personalizado.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

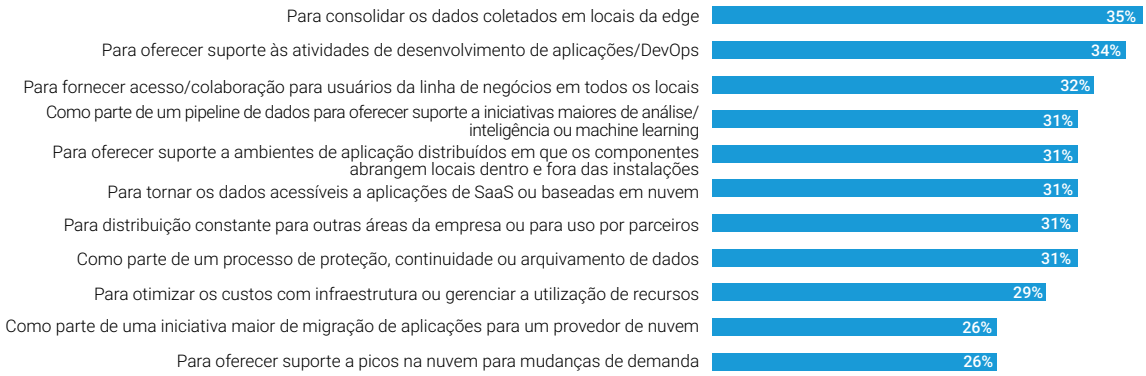

Organizações que buscam aproveitar as oportunidades crescentes com aplicações de inferência de IA adotam essa abordagem cada vez com mais frequência. Para resolver problemas de latência e de desempenho influenciados por limitações da infraestrutura de computação, as cargas de trabalho de IA estão se aproximando cada vez mais de fontes de dados utilizadas por uma variedade maior de usuários finais por meio de inferência descentralizada de IA e arquiteturas em nuvem distribuídas.

Por exemplo, esse modelo é amplamente usado em setores marcados por instalações altamente distribuídas e geograficamente dispersas, incluindo varejo, telecomunicações e mídia. Por conta da sua capacidade de superar desafios de latência e desempenho em escala associados aos requisitos complexos da inferência de IA, cada vez mais essa abordagem está se tornando o modelo de arquitetura padrão para a inferência de IA.

Um dos principais impulsionadores da mudança da inferência de IA para um modelo de nuvem distribuída é o crescente requisito das organizações de usar a edge para mover dados entre data centers e CSPs (provedores de serviços de nuvem). O Enterprise Strategy Group descobriu que 35% das organizações que regularmente transferem dados entre os data centers e a infraestrutura dos CSPs fazem isso para consolidar os dados coletados em locais de edge, um dos principais motivos para esse movimento de dados.[9]

Figura 2. Principais impulsionadores da edge para a transferência de dados entre data centers e serviços de nuvem pública

Você indicou que sua organização transfere dados regularmente entre seus data centers e serviços de nuvem pública. Com que finalidade sua organização transfere dados entre os data centers e os serviços de nuvem pública? (Porcentagem de entrevistados: N = 170, múltipla escolha)



9 Fonte: Relatório de pesquisa do Enterprise Strategy Group, Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud, novembro de 2023.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

Capítulo 4

Por que a inferência de IA descentralizada é o futuro

Criar e implantar inferência de IA com dados de locais da edge é fundamental para ajudar as organizações a aproveitar ao máximo a inferência. Com tantos dados criados e localizados em vários dispositivos da edge, incluindo smartphones, dispositivos móveis com propósito específico, sinalização digital e dispositivos de Internet das coisas, as organizações devem adotar uma mentalidade direcionada à edge para alcançar o processamento de dados em tempo real que é essencial para a inferência de IA.

A transição para uma inferência de IA descentralizada para se alinhar com os requisitos de inferência de IA foi rápida e profunda. E pesquisas indicam que os investimentos em sistemas de inferência perto dos usuários aumentarão de US$ 27 bilhões em 2024 para quase US$ 270 bilhões em 2032.[10]

A inferência de IA descentralizada é caracterizada por diversas considerações relevantes que ajudam as organizações a desenvolver e implantar aplicações de inferência de IA, como processamento de IA nativo do dispositivo, estruturas de segurança integradas e otimizadas para dispositivos da edge, baixo consumo de energia e baixa latência. Esses recursos resultam em desempenho mais rápido em escala, insights de dados mais precisos e profundos e investimentos mais baixos em largura de banda. Isso, por sua vez, permite que membros de vários locais geográficos de uma organização usem a inferência de IA para tomar decisões mais inteligentes e em tempo real mais próximas de onde os dados são criados, armazenados, processados e acessados.

A Bloomfield é um exemplo de uma organização inovadora que usou inferência de IA descentralizada para melhorar operações essenciais e simplificar o gerenciamento de dados. A empresa precisava gerenciar grandes volumes de dados de suas câmeras inteligentes distribuídas, de maneira eficiente, econômica e segura para os agricultores que realizam operações com acesso limitado à rede. Usando um modelo arquitetônico de inferência de IA descentralizada, a Bloomfield aproveitou a inferência de IA para ajudar a determinar a melhor forma de organizar, armazenar e apresentar os dados aos agricultores. Isso resultou em um gerenciamento simplificado da integridade das plantações, melhor rendimento da produção e tomada de decisões mais rápida e precisa longe de um data center local.

A capacidade da inferência de IA descentralizada de oferecer suporte a aplicações de inferência de IA em locais distribuídos e diversificados faz dela a escolha natural para:

- **Aplicações de medicina de cabeceira ou de medicina remota,** em que a tomada de decisões é feita pelos médicos instantaneamente usando smartphones, dispositivo de Internet das coisas e outros dispositivos portáteis leves, mas poderosos. A inferência de IA descentralizada também é ideal para a tomada de decisões em tempo real usando dispositivos inteligentes na área da saúde, como bombas de infusão, marca-passos, monitores cardíacos e análises de medicamentos.
- **Cidades inteligentes** nas quais os sistemas da edge controlam fluxos de tráfego, monitoram o comportamento do sistema de serviços públicos, simplificam o registro de eleitores, melhoram a segurança pública e muito mais.
- **Veículos autônomos** que usam câmeras inteligentes e dados de sensores a bordo para avaliar opções de rota, monitorar e melhorar o uso de combustível, documentar alterações de GPS e comunicar-se com autoridades legais e agências de segurança pública.
- **Varejo,** que oferece suporte a diversas aplicações de inferência de IA, desde vigilância por vídeo IP e gerenciamento de inventário até prevenção contra roubo e padrões de tráfego dos clientes na loja.

À medida que as empresas buscam maneiras de reduzir a latência entre consultas em tempo real e suas respostas, a inferência de IA descentralizada será um requisito essencial para uma fonte mais eficiente, mais bem informada e mais precisa de tomada de decisões comerciais. As aplicações de IA não precisarão mandar ou solicitar dados de ou para data centers distantes. Em vez disso, elas processarão e compartilharão dados em uma nuvem globalmente distribuída.

O chefe de engenharia de uma grande empresa de tecnologia de publicidade destacou: “Passamos muito tempo otimizando os modelos de IA. Agora, percebemos que a inferência não deveria ser uma solução de última hora.”

As organizações não têm mais o luxo de decidir se devem ou não adotar a inferência de IA como um componente importante em sua estratégia comercial. Elas precisam assumir uma postura de maior urgência e usar a inferência de IA descentralizada para realizar inferência, pois muitos de seus concorrentes já estão dando passos largos em direção ao uso da inferência de IA perto dos usuários.

10 Fonte: Fortune Business Insights, “Edge AI Market Size … 2024-2032”, dezembro de 2024.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Capítulo 5

Uma visão estratégica sobre inteligência artificial, edge e nuvem distribuída: para onde estão apontando e o que isso significa

Avaliar, escolher e colaborar com um parceiro de tecnologia confiável é um elemento fundamental para o sucesso de uma iniciativa de inferência de IA de uma organização, mas não é fácil de se alcançar.

É fundamental devido a um fato inegável: a grande escassez de profissionais qualificados em IA quando se trata de infraestrutura de IA, arquitetura em nuvem, casos de uso sobre inferência, otimização de aplicações e implantação impecável.

Mas também é complicado de se alcançar devido à falta de candidatos adequados. Conforme observado acima, um empreendimento bem-sucedido de inferência de IA deve combinar perfeitamente o conhecimento técnico, a experiência em domínio, a compreensão da segurança, da governança e do gerenciamento de riscos, e atenção total aos detalhes da implementação e do gerenciamento contínuos.

Para desenvolver e implantar a inferência de IA em toda a estrutura física e virtual da empresa, no núcleo, na edge, na nuvem e além, as organizações precisam se alinhar com um fornecedor que já tenha feito isso anteriormente. Ter um parceiro que entende os elementos técnicos e comerciais de um projeto de IA bem-sucedido é essencial.

A Akamai, fornecedora líder de soluções sofisticadas de computação em nuvem, segurança e entrega de conteúdo, tem um histórico estabelecido de criação, implementação e gerenciamento de iniciativas de IA, desde o treinamento de modelos até a inferência. Sua experiência prática com cargas de trabalho complexas com uso intensivo de computação prepara a Akamai para trabalhar com diversos tipos de clientes com diferentes casos de uso, setores e regiões geográficas.

Para ajudar as organizações a lidar facilmente com seus requisitos de inferência de IA, a Akamai implantou com sucesso, de forma consistente, confiável e segura, uma arquitetura em nuvem distribuída para permitir que as organizações processem cargas de trabalho de IA mais próximas do usuário final. Isso tem sido inestimável na redução da latência e na garantia de alto desempenho em escala, ambos essenciais para cargas de trabalho de inferência de IA com uso intenso de computação.

Um dos principais benefícios da arquitetura em nuvem distribuída da Akamai é a capacidade de os desenvolvedores se concentrarem no desenvolvimento neutro da plataforma em uma estrutura nativa em nuvem com base em padrões de nuvem aberta.

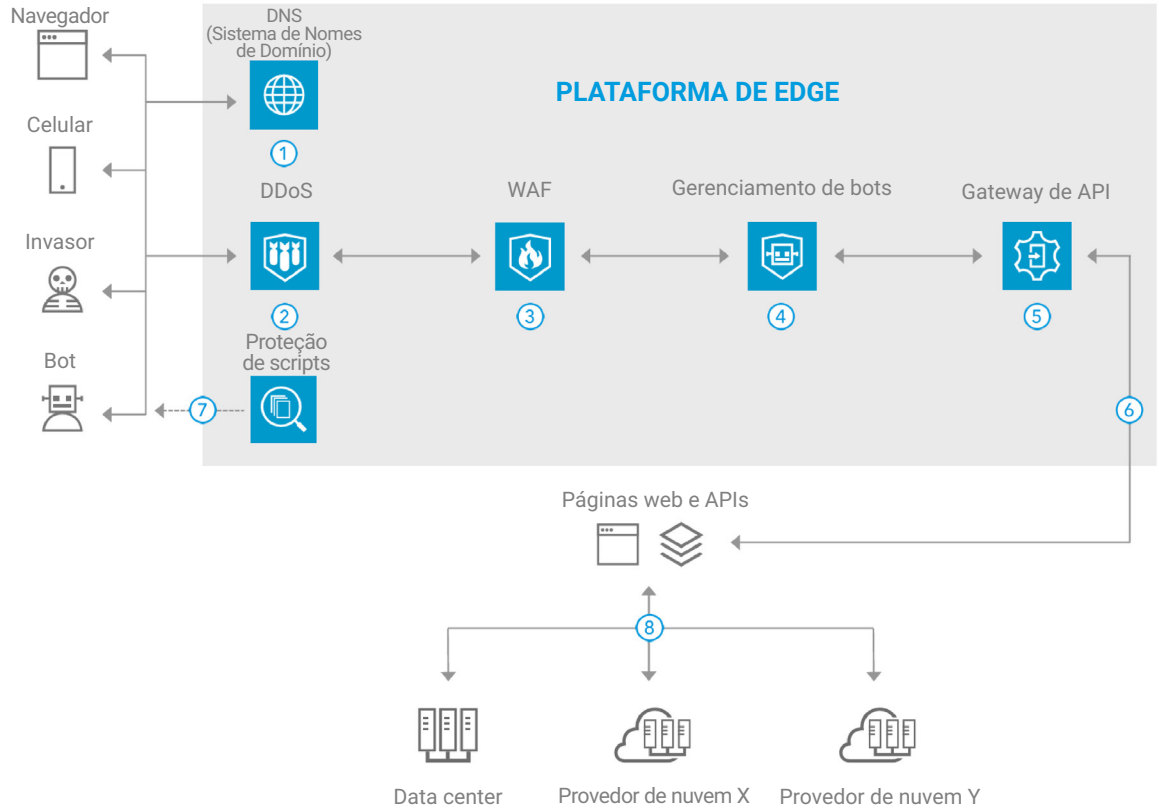
Para alcançar alto desempenho em escala, a Akamai usa várias técnicas diferentes de armazenamento em cache, integradas ao software de código aberto, para garantir taxa de transferência suficiente e limitar a latência. Por exemplo, muitos clientes da Akamai que usam a arquitetura em nuvem distribuída relataram uma redução de 50% nos tempos de resposta de chatbots voltados para o cliente, um caso de uso de inferência muito importante e popular. Esse tipo de resposta quase em tempo real permite que as organizações de diversos setores, incluindo varejo, saúde, serviços financeiros e de alta tecnologia, forneçam um serviço de atendimento ao cliente mais rápido, personalizado para os requisitos exclusivos de cada cliente.

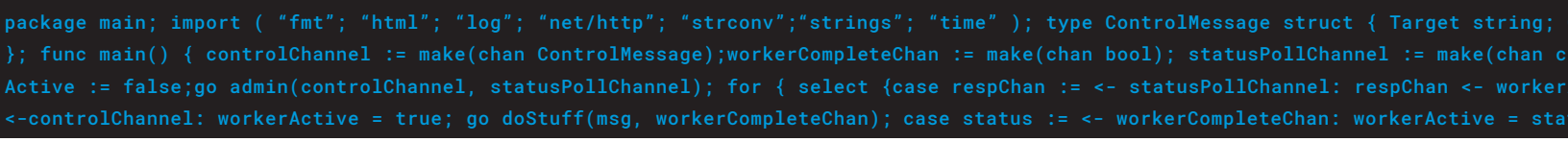
```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

A Akamai também oferece uma arquitetura de referência de IA (consulte a Figura 3 abaixo) que permite que as organizações iniciem rapidamente e gerem alto valor econômico e operacional. Essa arquitetura de referência é abrangente porque por meio de um conjunto familiar de navegadores e dispositivos definidos pelo usuário, as organizações podem implantar uma solução de IA empresarial do data center principal até a edge e nuvem e vice-versa. Isso beneficia todas as partes da empresa: desenvolvedores, especialistas em TI, gerentes de infraestrutura, arquitetos de nuvem e interessados na linha de negócios.

Figura 3.

Com os rápidos avanços em tecnologia e o aumento da complexidade e do valor das cargas de trabalho de IA, as organizações devem agir com rapidez e foco para superar a concorrência. É essencial usar a inferência de IA em seu máximo potencial, adotando a inferência de IA descentralizada e uma arquitetura em nuvem distribuída, especialmente quando planejada, desenvolvida, implantada e gerenciada por um líder reconhecido, como a Akamai.





Conclusão

Quais são os próximos passos para sua organização

Foi dito que a IA mudou as regras do jogo para a excelência em TI e transformação dos negócios. Mas a chave para otimizar o valor dos investimentos em IA de uma organização é a transição rápida do treinamento de modelos de IA para a inferência de IA.

Para aproveitar melhor todas as vantagens que a inferência de IA tem a oferecer, as organizações devem desenvolver e implantar suas aplicações de inferência e cargas de trabalho mais próximas de onde o trabalho é feito e onde profissionais estratégicos podem extrair resultados notáveis do sistema. Isso significa passar de um modelo de computação em nuvem centralizado para um modelo de nuvem distribuída, amplamente moldado pela IA na edge.

Conforme apontado anteriormente neste eBook, há muitos problemas que as organizações devem considerar e diversos desafios para entender e superar. Isso pode parecer assustador para as organizações que ainda não priorizaram as iniciativas de IA, mas há etapas importantes que elas podem e devem percorrer para se antecipar e se manter à frente.

“As organizações que estão correndo atrás da IA não devem ter medo de se perder em busca da tecnologia”, destacou Leone, do Enterprise Strategy Group. “O foco deve ser a criação da base correta identificando problemas comerciais específicos que a IA pode solucionar, avaliando a infraestrutura de dados e as habilidades internas existentes e identificando parceiros fundamentais que podem trazer o conhecimento específico necessário. As empresas devem ser estratégicas para identificar um ou dois projetos piloto de alto impacto e casos de uso que possam demonstrar valor claro usando dados de qualidade existentes.

Saber o que medir e como medir é fundamental para qualquer iniciativa bem-sucedida de inferência de IA.

E mesmo com tudo isso acontecendo, ela deve simultaneamente desenvolver talentos internos e estabelecer estruturas de governança para garantir o uso responsável. O segredo é evitar a implementação apressada e focar em etapas estratégicas que desenvolvam recursos sólidos e, de preferência, duradouros.”

Para isso, temos quatro recomendações que você e seus colegas devem entender e colocar em prática o mais rápido possível:

1. **Preste atenção à escolha de caso de uso.** Essa é uma primeira etapa essencial, já que é preciso criar confiança em sua capacidade de implantar a inferência de IA corretamente. Também é importante criar confiança por parte das partes interessadas da sua empresa e dos aprovadores financeiros com algumas vitórias iniciais.
2. **As opções de infraestrutura são importantes.** Obviamente, como a inferência de IA tem uso intenso de recursos de computação, é uma decisão fácil usar a GPU líder do mercado para o núcleo do seu sistema. Mas há muitas opções excelentes de GPUs e outras infraestruturas de IA. Reserve um tempo para escolher corretamente (ou confie na experiência de um parceiro que já fez isso antes).
3. **Obtenha a adesão da gerência.** Concentre-se nas duas principais preocupações de todo membro do Conselho e da diretoria executiva: qual é a nossa oportunidade e como ela se equilibra com os riscos em potencial? O projeto e os casos de uso devem estar alinhados com as principais metas comerciais, como melhorar a experiência do cliente, identificar e explorar as oportunidades de mercado mais rapidamente e criar um processo de teste beta mais preciso. Ao fazer isso, identifique e recrute patrocinadores de alto nível de grupos como TI, linha de negócios e membros do Conselho.
4. **Defina e entregue sucesso.** Saber o que medir e como medir é fundamental para qualquer iniciativa bem-sucedida de inferência de IA. Escolha métricas (ou trabalhe com seu parceiro para desenvolver métricas personalizadas) que sejam realistas, relevantes e robustas o suficiente para gerar resultados expressivos, principalmente depois de explorar as oportunidades mais acessíveis com os primeiros casos de uso de inferência de IA.

Este conteúdo foi solicitado pela Akamai e produzido pela TechTarget Inc.