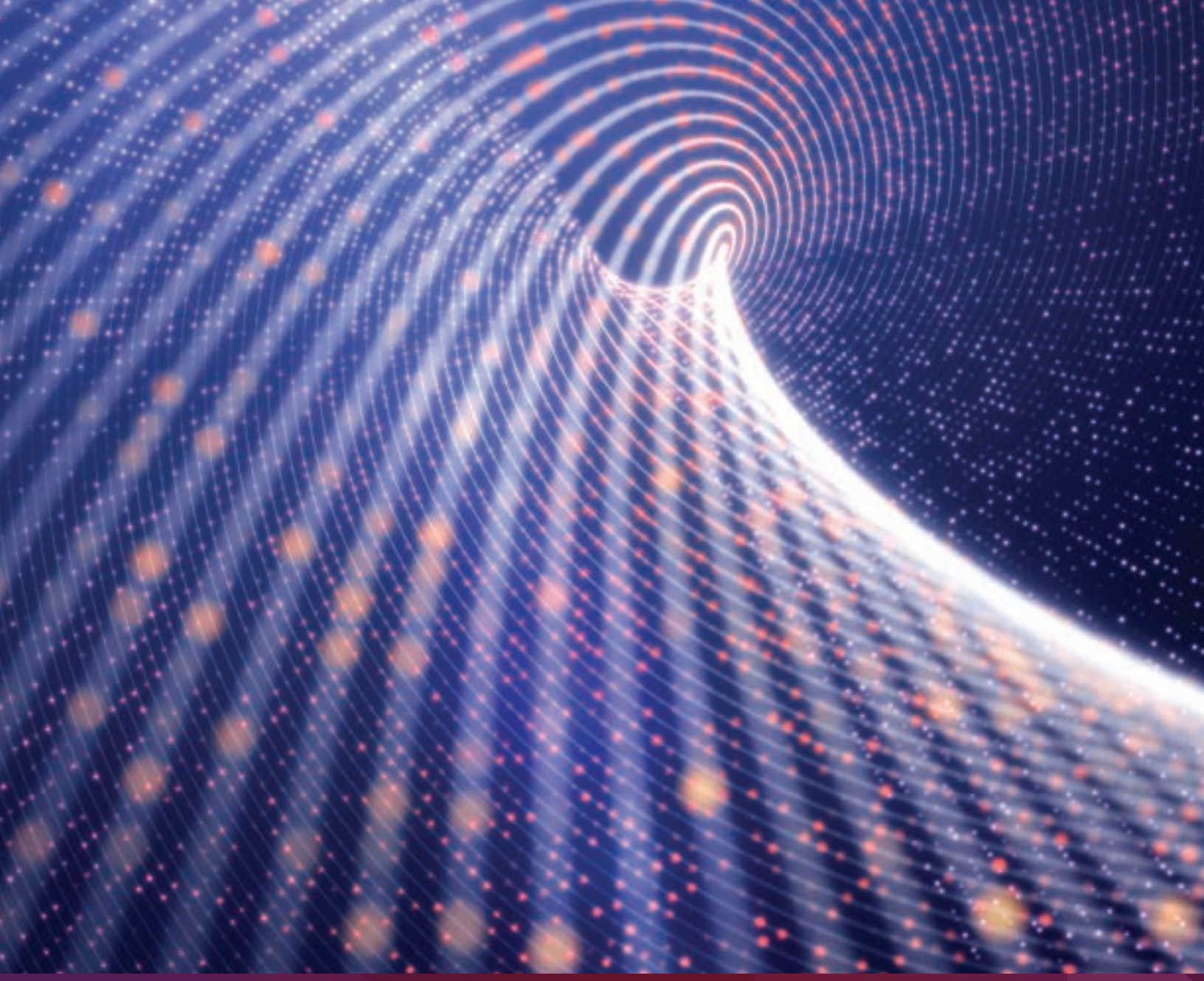


Acelere os fluxos de trabalho de IA com as Soluções Aceleradas da Lenovo e NVIDIA®

Um guia completo para empresas que desejam aproveitar o poder da IA generativa em seus data centers privados, utilizando soluções otimizadas para IA da Lenovo e NVIDIA.

Lenovo | NVIDIA

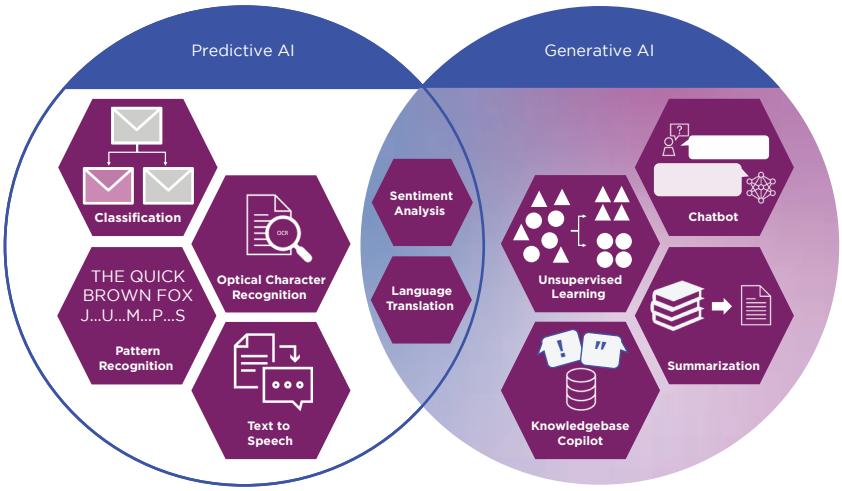
Do conceito à realidade, em um clique, com IA generativa

A era digital desencadeou uma explosão de dados, com cerca de 330 exabytes produzidos diariamente[1] — equivalente a uma chamada de vídeo de 75 milhões de anos[2] ou quase 300 milhões de anos de música.[3] Essa abundância de dados sem precedentes, impulsionada por tecnologias de big data, Internet das Coisas (IoT) e nossos dispositivos conectados, desencadeou uma revolução rápida e transformadora na inteligência artificial (IA). Modelos de IA, treinados em conjuntos massivos de dados, agora podem resolver problemas complexos e criar possibilidades antes inimagináveis.

Até recentemente, a IA realizava, principalmente, tarefas preditivas, como prever vendas futuras, detectar irregularidades de desempenho ou analisar mercados financeiros. No entanto, desde o final de 2022, houve uma mudança dramática com o surgimento da IA generativa (GenAI).

A empolgação em torno da GenAI decorre de sua capacidade de criar novo conteúdo indistinguível do trabalho humano, como texto, imagens, áudio, vídeo, música, arte e código. A GenAI demonstra o potencial de transformar muitas indústrias e aspectos de nossas vidas, criando novos produtos e serviços comerciais, apoiando novas descobertas científicas e melhorando desde a educação até o entretenimento, o produto interno bruto e nossa existência como um todo.

A IA preditiva concentra-se em entender dados históricos e fazer previsões precisas



A IA generativa cria novos dados com base em padrões e tendências aprendidos nos dados de treinamento

Um dos avanços-chave na GenAI foram os Modelos de Linguagem de Grande Escala (LLMs), popularizados pelo Open AI ChatGPT e pelo Google Bard. Os LLMs são treinados em texto e código para aprender os padrões e estruturas da linguagem humana por meio de aprendizado de máquina e aprendizado profundo. Eles são usados para escrever textos, traduzir idiomas, compor conteúdo criativo e muito mais. A velocidade e a qualidade de saída capturaram a imaginação do mundo, gerando uma gama incrivelmente ampla de casos de uso bem-sucedidos e uma série de histórias na mídia beirando a ficção científica.

O crescimento exponencial da IA

$150 bilhões

A indústria global de IA foi avaliada em $150 bilhões em 2023 e prevê-se que cresça a uma taxa de crescimento anual composta (CAGR) de 36,8%, de 2023 a 2030.[4]

34,3%

34,3% dos trabalhadores se identificam como usuários regulares de GenAI em quatro setores: serviços financeiros (42%), varejo (30%), indústrias avançadas (32%) e saúde (33%).[5]

$1,3 trilhão

Até 2032, a indústria de GenAI crescerá para $1,3 trilhão, com hardware sendo o maior segmento, em $640 bilhões, e a receita de software atingindo $280 bilhões.[6]

70%

A GenAI tem o potencial de automatizar até 70% das atividades empresariais até 2030.[7]

O processo criativo reimaginado

Os modelos de GenAI utilizam redes neurais para analisar dados existentes e gerar conteúdo novo e original. Eles são treinados usando aprendizado não supervisionado ou semi-supervisionado, permitindo que aproveitem grandes quantidades de dados não rotulados. Os modelos podem aprender a distribuição de probabilidade subjacente dos dados, permitindo-lhes produzir saídas realistas e diversas. Eles também podem ser condicionados a um prompt ou texto inicial para criar conteúdo sobre um tópico ou estilo específico.

Aqui está um exemplo simplificado de como um modelo de GenAI pode funcionar para compor texto:

1. O modelo é treinado em um conjunto de dados grande.
2. O modelo aprende os padrões, estruturas e distribuição de probabilidade dos dados.
3. O modelo recebe um prompt, que pode ser texto, imagem, áudio ou vídeo.
4. O modelo mostra a distribuição de probabilidade para gerar tokens (palavras, imagens, código etc.) como saída.

Impactos Transformadores para Todas as Indústrias

A GenAI já está causando um impacto significativo em uma ampla gama de indústrias, incluindo:

Indústria	Aplicações	Caso de Uso	Saídas Geradas
Serviços Financeiros	Conformidade regulatória	Automatização de tarefas de conformidade, como coleta de dados, avaliação de riscos e geração de relatórios finais.	Coleta e organização automática de dados de transações financeiras e comportamento do cliente. Avaliação de fatores de risco e geração de relatórios de conformidade de acordo com regulamentação. Redução de verificações manuais e minimização de erros.
Varejo	Atendimento ao cliente	Chatbots automatizados para perguntas de clientes, reclamações e recomendações.	Interação com clientes em tempo real, abordando perguntas comuns sobre produtos, pedidos e serviços com disponibilidade 24/7, reduzindo a workload dos agentes de atendimento ao cliente humanos.
Manufatura	Cadeia de suprimentos	Otimização de níveis de inventário, rotas de entrega e gestão de fornecedores.	Previsão de necessidades futuras de inventário com base em tendências passadas e demanda atual. Otimização de rotas de entrega usando dados em tempo real para reduzir o consumo de combustível e acelerar os tempos de entrega. Automação da comunicação com fornecedores para pedidos e reposição, resultando em uma cadeia de suprimentos mais eficiente e econômica.
Saúde	Diagnóstico e planejamento de tratamento	Análise automatizada de dados de pacientes para diagnósticos precoces e precisos de doenças.	Processamento e análise de diversos dados médicos, incluindo registros eletrônicos de saúde, exames laboratoriais e imagens médicas para identificar padrões indicativos de doenças específicas, possibilitando diagnósticos mais rápidos e precisos. Fornecimento de opções de tratamento personalizadas baseadas em dados.

Os pioneiros na GenAI estão obtendo uma vantagem competitiva por meio de eficiências operacionais de custo, aumento de vendas, desenvolvimento mais rápido de produtos e melhor controle de qualidade, tudo possibilitado por uma maior produtividade, personalização, insight e automação.

Iniciando com a GenAI

Existem três opções de modelos de IA, que variam em custo, complexidade e valor. As organizações provavelmente usarão os três tipos, muitas vezes várias instâncias de cada um, com modelos otimizados e treinados para departamentos, funções ou aplicações específicas, por exemplo, um chatbot para serviços ao cliente, um modelo de geração de imagem para desenvolvimento de produtos, um modelo de geração de texto e tradução para vendas e marketing, um modelo de geração de código para TI. Os tipos básicos de modelos são:

- **Opção #1 — Modelo de propósito geral:**
  Frequentemente chamado de GenAI como Serviço, esta é uma opção pronta com base em um modelo de fundação pré-treinado e será usado por todas as empresas. É pago conforme o uso e a implantação mais direta da GenAI, mas oferece pouca personalização e controle. Projetado para casos de uso genéricos, em que a informação está prontamente disponível e pouco contexto organizacional é necessário. Exemplo: OpenAI ChatGPT/Open AI API ou Stable Diffusion.

- **Opção #2 — Modelo moderadamente personalizado:**
  Um modelo de base GenAI treinado nos dados da empresa; um processo chamado de ajuste fino. Esta opção oferece mais personalização e controle, com investimento inicial em infraestrutura e desenvolvimento, e custos contínuos de manutenção. É ideal quando o modelo GenAI requer dados exclusivos para melhorar respostas a necessidades específicas. Exemplo: Um chatbot baseado em Meta Llama 2 e treinado nos dados de uma empresa, como o Chatbot Pré-Treinado da Lenovo, com informações de manuais de serviço, especificações técnicas e garantias da Lenovo.

- **Opção #3 — Modelo extensivamente personalizado:**
  Um modelo treinado do zero, em um conjunto de dados único adaptado a um caso de uso específico, oferecendo completa personalização e controle. Este modelo apresentará custos iniciais para construção e desenvolvimento. Se projetado para uma aplicação específica, pode se beneficiar de um menor Custo Total de Propriedade (TCO), por meio da redução de requisitos contínuos de computação e custos gerais. Ideal para casos de uso únicos que dependem fundamentalmente de dados proprietários. Exemplo: Um sistema de descoberta de medicamentos treinado em dados proprietários ou uma plataforma de dados financeiros privados como o BloombergGPT.

Uma vez construídas, as organizações podem implementar eficientemente múltiplas aplicações personalizadas de GenAI em departamentos, alterando a fonte de dados e desbloqueando benefícios para todas as partes interessadas. Com uma abordagem estratégica para a GenAI, as organizações simplificarão tarefas complexas, impulsionarão a automação, amplificarão a inovação e aumentarão a produtividade, resultando em melhores resultados de negócios e, em última análise, uma vantagem competitiva.

Inovação Empresarial com um Modelo GenAI Privado

A jornada começa com dados

Os dados são essenciais para construir um modelo GenAI personalizado. Modelos GenAI podem ter bilhões de parâmetros, e a hospedagem privada oferece confiança na segurança e treinamento de dados, permitindo que as organizações configurem os dados para obter o máximo valor. Além disso, oferece controle sobre o processo de inferência e as saídas do modelo para garantir o uso responsável e ético.

Uma análise comparativa de modelos privados e públicos

Organizações que desejam aproveitar a GenAI podem utilizar modelos GenAI públicos, privados ou ambos. Modelos públicos são de fácil acesso, disponíveis prontos para uso e, geralmente, oferecem capacidades de geração de texto, imagem, vídeo ou código. Modelos privados são hospedados em uma plataforma própria ou na nuvem. Esses modelos, muitas vezes, são desenvolvidos usando um modelo de base público, com aplicações personalizadas projetadas para fornecer saídas e resultados adaptados, maior controle sobre os dados e medidas de segurança aprimoradas.

Característica	Modelo GenAI de base	Modelo GenAI personalizado
Dados	Treinado em dados públicos com aplicações pré-construídas	Treinado ou ajustado com dados proprietários com aplicações pré-construídas ou personalizadas
Acesso	Acessível publicamente	Privado para a organização que o gerencia
Custo	Pago conforme o uso	Investimento inicial e desenvolvimento contínuo para manutenção
Expertise	Não requer expertise em IA	Requer expertise em IA
Customização	Customização limitada	Pode ser personalizado para atender a informações e habilidades específicas do domínio
Segurança	Menos seguro, pois os dados são compartilhados com o público	Mais seguro, pois os dados não são compartilhados com o público
Gestão	Como serviço	Gestão e propriedade privadas

Uma organização precisará de um modelo privado se:

- **A privacidade dos dados for imperativa:** Um provedor de serviços de saúde poderia usar um modelo GenAI para desenvolver um sistema de diagnóstico de pacientes treinado em dados de pacientes. Isso permitiria ao provedor fornecer diagnósticos mais precisos e personalizados aos pacientes.
- **Informações transacionais atualizadas forem necessárias:** Um varejista poderia usar um modelo GenAI para desenvolver um sistema de precificação dinâmica atualizado com os últimos dados de vendas. Isso permitiria ao varejista otimizar os preços e maximizar os lucros.
- **Houver uma oportunidade de alavancar dados proprietários e propriedade intelectual:** Um provedor de serviços financeiros poderia usar um modelo GenAI para desenvolver uma estratégia de negociação de investimentos exclusiva treinada em seus dados e pesquisas.

Muitas grandes organizações estão construindo modelos GenAI privados e personalizados para impulsionar o valor comercial, apesar dos custos e desafios. Modelos privados com aplicações personalizadas proporcionarão retornos tangíveis à propriedade intelectual, apoiarão os requisitos de proteção e conformidade de dados cada vez mais rigorosos e estabelecerão a base para uma vantagem competitiva sustentável a longo prazo.

Alavanque a expertise para velocidade e economia

Os modelos GenAI privados são complexos. O design e a implementação requerem expertise profunda em IA, aprendizado de máquina, ciência de dados e nas tecnologias subjacentes. Com a rápida trajetória dos desenvolvimentos em IA, todas as equipes de projetos GenAI devem incluir indivíduos com um entendimento intrínseco das últimas tendências em IA e experiência prática em implementação. A orientação de especialistas é acelerar a tomada de decisões e o desenvolvimento, reduzindo o potencial de erros custosos.

Do conceito à vantagem competitiva

Planejar a construção de um modelo GenAI sofisticado começa com a compreensão dos objetivos comerciais e das capacidades de dados. Embora esse processo seja multifacetado e possa exigir recursos significativos, os seis passos a seguir podem levar ao sucesso do projeto:

1. **Identificar o problema comercial e o resultado desejado**
2. **Obter e analisar dados**
3. **Construir o caso de negócios**
4. **Planejar infraestrutura, design do modelo e implementação**
5. **Construir e treinar o modelo**
6. **Implementar, monitorar o desempenho e aprimorar conforme necessário**

1. Identificar o problema comercial e o resultado desejado

Comece identificando claramente o problema comercial a ser resolvido e o resultado desejado. Uma declaração de problema bem definida dará direção ao projeto, garantirá alinhamento com os objetivos estratégicos e incentivará o comprometimento da liderança e dos interessados.

Um fabricante poderia utilizar um modelo GenAI para otimizar cronogramas de produção, aproveitando dados internos como previsões de demanda, níveis de estoque e disponibilidade de máquinas para minimizar custos e maximizar a produção. O resultado desejado poderia ser reduzir os custos de produção em 10% ou aumentar a produção em 5%.

2. Obter e analisar dados

A qualidade e relevância dos dados são cruciais para construir modelos privados. As empresas devem utilizar dados internos e proprietários para obter uma vantagem.

Por exemplo, empresas financeiras podem usar dados internos (dados históricos de negociação, comportamento do cliente, fatores de risco) para construir modelos que prevejam a rotatividade de clientes, impulsionem a receita, detectem possíveis fraudes e reduzam a responsabilidade e a exposição.

Uma vez que os dados foram obtidos e analisados, uma estratégia de dados para IA pode ser desenvolvida. Essa estratégia deve definir os objetivos do projeto de IA, os dados que serão utilizados e como os dados serão gerenciados e governados.

3. Construir o caso de negócios

O caso de negócios é essencial após a identificação dos resultados comerciais e das competências de dados. Um caso de negócios deve planejar o processo de design, desenvolvimento e implementação, permitindo conversas internas produtivas sobre o investimento necessário, riscos envolvidos e retornos projetados do projeto. Um caso de negócios robusto irá:

- Definir objetivos e expectativas claras
- Identificar riscos e desafios
- Incentivar o financiamento/investimento do projeto
- Esboçar os entregáveis e o cronograma do projeto
- Identificar e alinhar stakeholders, parceiros e especialistas
- Planejar a gestão de riscos e mudanças
- Possibilitar o acompanhamento do progresso e a medição do sucesso

4. Planejar software, infraestrutura, design e implementação

Os modelos GenAI são intensivos em computação, exigindo recursos computacionais significativos e energia para treinar e implementar. Por exemplo, o BloombergGPT levou 1,3 milhão de horas de tempo de GPU para treinar.[8]

Software

O software necessário para desenvolver, implementar e gerenciar um modelo GenAI inclui:

- **Modelos fundamentais:** LLMs, modelos de imagem, modelos de vídeo etc., que podem ser ajustados para tarefas específicas, como geração de texto, tradução e conclusão de código (por exemplo, GPT-3, PaLM e LLaMA).
- **Ferramentas de curadoria e treinamento de dados:** Ferramentas de curadoria auxiliarão na limpeza, preparação e organização dos dados. Ferramentas de treinamento permitirão o desenvolvimento do modelo por meio de aprendizado de máquina e profundo.
- **Ferramentas de inferência:** Com recursos como otimização de modelo, processamento em lote e redução de latência.
- **Guardrails:** Guardrails de tópicos e segurança garantem que os modelos de IA permaneçam no caminho certo, impedindo que se desviem para áreas indesejadas, garantindo respostas precisas e apropriadas, restringindo conexões a aplicativos de terceiros confiáveis.

Além desses itens essenciais, muitos outros aplicativos de software podem auxiliar na construção de um GenAI, como ferramentas de monitoramento, otimização e depuração de modelos, além de guardrails para ajudar a mitigar os riscos de viés e uso indevido.

Infraestrutura

A infraestrutura para implementações privadas pode ser em local próprio, co-localização, nuvem ou edge. A melhor escolha para um projeto específico dependerá de fatores como o tamanho e a complexidade do modelo, o desempenho e escalabilidade desejados e o orçamento.

- **Implementação local:** O modelo é implementado nos servidores da empresa em armazenamento gerenciado internamente. Isso dá à empresa mais controle sobre o modelo e seus dados. No entanto, a implementação local pode ser mais cara e complexa do que a implementação na nuvem.
- **Implementação na nuvem:** O modelo é implementado em uma plataforma de nuvem como serviço. A implementação na nuvem geralmente é mais acessível, mais fácil de gerenciar e mais eficaz em termos de custo para dimensionar do que a implementação local; no entanto, essa opção pode não ser adequada para todas as organizações.
- **Edge computing:** Servidores compactos e robustos de edge podem ser usados para implementar modelos GenAI mais próximos à fonte de dados, reduzindo a latência, acelerando a inferência para o aplicativo e melhorando o desempenho.

Computação acelerada

As GPUs são fundamentais para avançar a GenAI, fornecendo a potência computacional, a capacidade de processamento paralelo e a aceleração de hardware necessárias para treinar e implementar modelos complexos de maneira eficiente.

Considere os seguintes fatores ao escolher hardware de computação:

- **Desempenho:** GenAI e LLMs são extremamente intensivos em computação. O hardware deve fornecer o desempenho necessário para treinar, implementar e executar o modelo prontamente.
- **Custo:** O hardware de computação acelerada exigirá um investimento considerável, portanto, selecione hardware custo-eficaz para as necessidades específicas do projeto. Considere o Custo Total de Propriedade (TCO) e os custos iniciais.
- **Escalabilidade:** O hardware deve ser escalável para atender às crescentes necessidades do modelo à medida que é treinado e implementado.

Rede e conectividade

Os requisitos de rede e conectividade para implementações GenAI privadas variam dependendo da infraestrutura específica escolhida. O desempenho da rede é necessário para a importação de dados de treinamento, gerenciamento de modelos e segurança. Muitos desafios podem ser superados com os switches, ethernet e hardware InfiniBand mais recentes, além de otimizações como compressão de redes neurais.

Segurança

Os modelos GenAI são vulneráveis a uma variedade de ameaças de segurança, incluindo:

- **Ataques adversários:** Tentativas de manipular ou enganar um modelo fornecendo entradas cuidadosamente elaboradas para fazer com que o modelo gere saídas incorretas ou prejudiciais.
- **Envenenamento de dados:** Inserir dados maliciosos nos dados de treinamento de um modelo GenAI. Isso pode fazer com que o modelo aprenda padrões incorretos ou enviesados.
- **Roubo de modelo:** Copiar ou distribuir não autorizadamente um modelo GenAI. Isso permite que atacantes usem o modelo para fins maliciosos.

Além de construir um modelo privado, implemente medidas de segurança adequadas para proteger os modelos dessas ameaças. Isso pode incluir:

- **Segurança de dados:** Garantir que os dados de treinamento sejam seguros e protegidos contra acesso não autorizado usando hardware, software e controles de segurança de rede mais recentes.
- **Monitoramento de modelo:** Monitorar o desempenho do modelo para detectar quaisquer anomalias que possam indicar um ataque.
- **Reforço do modelo:** Usar treinamento adversarial e validação de entrada para tornar o modelo mais resistente a ataques.

5. Construir e Treinar o Modelo

Uma vez que a infraestrutura está no lugar e o design do modelo está completo, o modelo pode ser construído e treinado. Esse processo é demorado e intensivo em computação, mas pode ser acelerado ao aproveitar a potência de GPU e redes otimizadas. Alguns algoritmos de treinamento comuns incluem:

- **Aprendizado supervisionado:** O modelo é treinado em um conjunto de dados rotulados, onde cada ponto de dados tem uma saída conhecida. O modelo aprende a prever os resultados de novos pontos de dados, identificando padrões no conjunto de dados rotulados.
- **Aprendizado não supervisionado:** No aprendizado não supervisionado, o modelo é treinado em um conjunto de dados não rotulados, onde os pontos de dados não têm saídas conhecidas. O modelo aprende padrões nos dados sem conhecimento prévio das saídas desejadas.
- **Aprendizado por reforço:** No aprendizado por reforço, o modelo aprende a realizar uma tarefa por tentativa e erro. O modelo é recompensado por resultados desejados e penalizado por saídas indesejadas.

Uma vez que o modelo é treinado, seu desempenho pode ser testado e avaliado.

6. Implementar, Monitorar o Desempenho e Refinar conforme Necessário

A implementação de produção envolve integrar o modelo à infraestrutura da empresa, aos sistemas existentes e aos processos.

Após a implementação, o modelo deve ser monitorado quanto ao desempenho e oportunidades de aprimoramento, conforme necessário.

Técnicas de otimização incluem retreinamento com novos dados ou com novos hiperparâmetros. Para mais informações, consulte a arquitetura de referência técnica da Lenovo aqui.

A GenAI está transformando todas as indústrias

Como a GenAI privada e aplicações personalizadas estão impactando finanças, varejo, manufatura e saúde

A GenAI está trazendo benefícios significativos para o mundo dos negócios. Aplicações personalizadas da GenAI em modelos privados lideram o caminho, com impactos que incluem tomadas de decisões mais precisas em finanças, aumento de vendas e retenção de clientes no varejo, garantia de qualidade e eficiência de custos na manufatura, e aceleração do desenvolvimento de medicamentos e cuidados aprimorados ao paciente na área da saúde. Em todas as indústrias, a GenAI está impulsionando melhorias tangíveis, destacando seu papel fundamental no ambiente de trabalho moderno.

Os impactos da implementação da GenAI:

Finanças e Serviços Financeiros

- **Melhoria de desempenho:** Insights inteligentes aprimoram a tomada de decisões, levando a um desempenho otimizado.
- **Melhoria da satisfação do cliente:** Respostas automatizadas e personalizadas levam a uma resolução mais rápida e precisa de problemas.
- **Eficiência operacional aprimorada:** Com a IA lidando com tarefas rotineiras, agentes humanos podem focar em consultas de clientes mais complexas.

Leia mais

Varejo

- **Aumento de vendas:** Recomendações personalizadas de produtos, frequentemente, resultam em taxas de conversão mais altas.
- **Melhoria do engajamento do cliente:** Consultores virtuais de produtos e estratégias omnicanal oferecem uma experiência de cliente contínua e envolvente.
- **Retenção de clientes:** Experiências personalizadas tornam os clientes mais propensos a retornar.

Leia mais

Manufatura

- **Garantia de qualidade:** Inspeções visuais automatizadas reduzem a taxa de defeitos e melhoram a qualidade do produto.
- **Segurança aprimorada:** O monitoramento em tempo real pode identificar rapidamente perigos de segurança, reduzindo acidentes de trabalho.
- **Redução de custos:** A automação minimiza a necessidade de inspeções manuais, resultando em economias significativas.

Leia mais

Saúde

- **Velocidade na pesquisa:** A GenAI acelera o ritmo no qual novos medicamentos podem ser desenvolvidos e levados ao mercado.
- **Qualidade dos cuidados:** Modelos preditivos avançados aprimoram os planos de tratamento do paciente, melhorando os resultados na área da saúde.
- **Segurança de dados:** Modelos privados personalizados da GenAI garantem que dados sensíveis do paciente permaneçam seguros e em conformidade com regulamentações.

Leia mais

Lenovo | NVIDIA

Finanças e Serviços Financeiros

A indústria financeira sempre esteve na vanguarda dos avanços tecnológicos. As instituições financeiras buscam continuamente maneiras de superar seus concorrentes em um mercado caracterizado por alta competição. Adotar tecnologias inovadoras oferece a elas uma vantagem competitiva e as alinha com as necessidades e expectativas em constante evolução de seus clientes.

Tendências da Indústria

- **De $200 bilhões a $340 bilhões** podem ser economizados anualmente pela GenAI para bancos e o setor de serviços financeiros.[7]
- **Uma redução de até 20%** nas taxas de perda e inadimplência pode ser alcançada com a otimização do risco de crédito.[9]

A banca de varejo está mudando significativamente à medida que os clientes demandam serviços personalizados e convenientes. Segundo a Deloitte,[10] os clientes esperam mais de seus bancos — mais tecnologia, orientação, suporte e experiências entre canais. A pesquisa da McKinsey[11] reforça isso, afirmando que os bancos que oferecem uma melhor experiência digital lideram em satisfação do cliente e se destacam em métricas financeiras-chave.

História Destacada de Aplicação da GenAI: Automação Inteligente

A indústria financeira utiliza a automação inteligente dentro desse cenário competitivo e centrado no cliente. Essa tecnologia é implantada para aprimorar a experiência do cliente e otimizar as operações de centrais de atendimento. Modelos personalizados privados da GenAI tornaram-se a solução preferida em uma indústria onde a privacidade dos dados é não negociável e os dados proprietários são abundantes.

Esses modelos privados de IA são treinados em conjuntos de dados internos, garantindo que dados financeiros e pessoais sensíveis nunca deixem os limites organizacionais. Eles podem automatizar consultas rotineiras de clientes, direcionar chamadas para departamentos apropriados e fornecer análises de dados em tempo real para monitorar a qualidade do serviço, tudo mantendo rigorosos protocolos de segurança de dados.

Aplicações Adicionais Selecionadas na Indústria

Aplicação na indústria	Descrição	Impacto
IA de pesquisa de documentos empresariais	Otimiza a recuperação de informações, avaliando várias fontes de dados e resumindo os resultados. Também gera relatórios para otimizar tarefas de escritório.	Aumenta a eficiência organizacional, tornando as informações prontamente acessíveis.
Assistente bancário de IA	Personaliza a experiência do cliente, incorporando funcionalidades de prevenção de fraudes, conformidade e gerenciamento de riscos de crédito.	Eleva a satisfação e a confiança do cliente, gerenciando riscos de maneira eficaz.
Perspectivas de investimento	Utiliza Processamento de Linguagem Natural (PLN) para analisar pesquisas de negociação e resumir fluxos de dados em tempo real.	Acelera os processos de tomada de decisão em negociação e investimento, potencialmente resultando em retornos mais altos.

Varejo

A indústria do varejo passou por transformações dramáticas nas últimas duas décadas. Inovações em tecnologia, mudanças no comportamento do consumidor e o surgimento de canais digitais remodelaram o cenário varejista, tornando-o mais complexo e repleto de novas oportunidades. A indústria varejista está recorrendo à tecnologia alimentada por IA em resposta a essas mudanças.

Tendências da Indústria

- **Até 40%** dos varejistas e marcas em todo o mundo estão na fase de experimentação do GenAI.[12]
- **Até 59%** de melhoria na lucratividade até 2035, usando soluções de varejo baseadas em IA.[13]

A ascensão meteórica do comércio eletrônico tem sido uma das mudanças mais significativas no varejo. De acordo com o Statista,[14] as vendas online representaram 23% de todas as vendas no varejo em 2023, em comparação com 20% em 2022. O varejo é um mercado dinâmico e competitivo, com varejistas utilizando tecnologia para melhorar margens e fidelidade do cliente.

História em destaque sobre aplicação GenAI: Marketing hiper-personalizado

A GenAI está transformando as indústrias de varejo e comércio eletrônico, possibilitando estratégias de marketing hiper-personalizado. Modelos GenAI privados personalizados analisam dados do cliente para gerar experiências de compra altamente personalizadas, incluindo recomendações de produtos, serviços de consultoria virtual de produtos e automação inteligente omnicanal.

Recomendações de produtos personalizadas são criadas ao entender o comportamento, preferências e padrões de compra do cliente, aumentando a probabilidade de compra. Consultores virtuais de produtos envolvem os clientes em tempo real, orientando-os na seleção de produtos e sugerindo itens complementares. A automação inteligente omnicanal garante uma experiência consistente e personalizada em todas as plataformas, enquanto automatiza interações com o cliente, como consultas de serviço e acompanhamentos pós-compra.

Outras aplicações selecionadas para a indústria

Aplicação na Indústria	Descrição	Impacto
Concierge de Funcionários	Sistema alimentado por IA projetado para auxiliar funcionários em várias tarefas, como agendamento e recuperação de informações.	Melhora a produtividade e a satisfação no trabalho dos funcionários.
Serviço ao Cliente Personalizado	Utiliza IA para oferecer serviços de atendimento ao cliente personalizados, com base no comportamento e preferências individuais.	Aumenta a satisfação do cliente e promove a fidelidade.
Gestão da Cadeia de Suprimentos	Otimiza rotas de transporte, prevê tempos de entrega e identifica possíveis interrupções; comunica-se com partes interessadas e clientes.	Melhora a eficiência da cadeia de suprimentos, reduz custos e mitiga riscos.

Manufatura e Indústrias Avançadas

A indústria de manufatura enfrentou vários desafios recentemente, desde a globalização e aumento da concorrência até cadeias de suprimentos complexas e rigorosa conformidade regulatória. O setor também enfrenta custos operacionais crescentes e ameaças crescentes de cibersegurança. Diante desse cenário, a adoção de práticas inovadoras e sustentáveis tornou-se crucial para a sobrevivência e o crescimento.

Tendências da Indústria	
	• **Até 27%** dos fabricantes já estão investindo em tecnologias GenAI.[15] • **Até 45%** de redução no tempo de inatividade com manutenção preditiva.[16]

Muitos fabricantes estão recorrendo à tecnologia para aliviar algumas dessas pressões. Mais da metade das empresas de manufatura planejam aumentar o uso de aplicativos da Internet das Coisas (IoT), automação, gerenciamento de estoques e manutenção preditiva nos próximos anos.[17] Além disso, a indústria enfrenta uma escassez global de habilidades, tornando a retenção de mão de obra uma questão crítica. Três quartos dos fabricantes citam isso como um desafio, sendo que um em cada três executivos afirma que a retenção de funcionários de alto desempenho é uma prioridade estratégica para 2023 e além.[18]

História em destaque sobre aplicação GenAI: Dados sintéticos para detecção de defeitos

A indústria de manufatura utiliza GenAI e dados sintéticos para treinar modelos avançados de visão computacional para várias aplicações, incluindo detecção de defeitos de produtos. No passado, obter dados do mundo real era caro ou logisticamente desafiador. No entanto, com dados sintéticos, os modelos GenAI são treinados em dados do mundo real disponíveis para entender as características estatísticas antes de gerar dados que se assemelham de perto às suas atribuições. Isso fornece um recurso valioso para o treinamento de modelos de visão computacional, com vantagens significativas de custo, melhoria na precisão do modelo e, em última análise, resultando em melhor qualidade e segurança do trabalhador.

Os fabricantes estão usando GenAI para criar imagens sintéticas de produtos com diferentes tipos de defeitos para treinar modelos a identificar e sinalizar problemas em situações do mundo real.

Aplicações Adicionais Selecionadas na Indústria

Aplicação na Indústria	Descrição	Impacto
Colaboração de Design com Inteligência Artificial	IA facilita processos de colaboração e revisão em tempo real entre equipes de engenharia, melhorando o design do produto.	Acelera o tempo de chegada ao mercado e otimiza o design.
LLM para Desenvolvimento de Produtos	Máquinas de Aprendizado ao Longo da Vida são treinadas para auxiliar no desenvolvimento interno de produtos, adaptando-se continuamente a novos dados.	Aumenta a inovação e reduz os ciclos de desenvolvimento de produtos.
Detecção de Ameaças Internas Impulsionada por IA	Algoritmos de IA monitoram o comportamento interno da rede para detectar anomalias que poderiam indicar ameaças internas e agem de acordo.	Reforça a cibersegurança e protege a propriedade intelectual.

Saúde e Farmacêutica

A indústria de saúde enfrenta muitos desafios, incluindo aumento da concorrência, elevação dos custos operacionais, interrupção global da cadeia de suprimentos e muitos obstáculos regulatórios. Diante dessas complexidades, a saúde conta com a tecnologia para otimizar operações, permitir eficiências e impulsionar resultados aprimorados.

Tendências da Indústria

- Até 2025, a Gartner espera que mais de **30%** de novos medicamentos e materiais sejam descobertos sistematicamente usando GenAI.[19]
- Até **30** vezes mais rápidas e **99%** de precisão em mamografias usando técnicas de tradução GenAI.[20]

Os dados desempenham um papel crítico na área da saúde, representando 30% dos dados mundiais e crescendo a uma taxa de 36% a cada ano, de acordo com a OCDE.[21] O investimento em software de IA na saúde deve atingir US$ 11,6 bilhões até 2026, destacando a mudança da indústria para soluções baseadas em tecnologia.[22]

História em destaque de aplicação GenAI: Simulação Molecular

O setor de Biofarmácia utiliza GenAI para diversas aplicações a fim de acelerar a pesquisa e desenvolvimento. Isso inclui Simulação Molecular, Biologia Estrutural, Treinamento/Desenvolvimento de Modelos Bio GenAI, Inferência/Design Bio GenAI em Imagens Biomédicas e Evidências do Mundo Real por Meio de Modelagem Preditiva.

O GenAI personalizado analisa grandes conjuntos de dados para identificar padrões, fazer previsões e sugerir possíveis desenvolvimentos de medicamentos ou caminhos de tratamento para pacientes. Esses modelos são particularmente hábeis em lidar com os grandes volumes de dados complexos e sensíveis típicos da pesquisa em saúde.

Aplicações Adicionais Selecionadas na Indústria

Aplicação na Indústria	Descrição	Impacto
Atendimento de Saúde Personalizado com IA	Algoritmos impulsionados por IA desenvolvem planos de tratamento personalizados, com base em dados de saúde individuais.	Aprimora os resultados do paciente e melhora a eficiência do tratamento.
Imagens Médicas Aprimoradas por IA	Algoritmos de IA melhoram a qualidade e detalhes de imagens médicas, auxiliando em diagnósticos e tratamentos.	Acelera os processos de diagnóstico e aprimora o cuidado ao paciente.
Análise Genômica Assistida por IA	IA analisa dados genômicos para identificar padrões e anomalias, auxiliando em pesquisas e medicina personalizada.	Avanços em genômica podem levar a descobertas em diversas condições médicas.

Lenovo leva a IA aos Dados

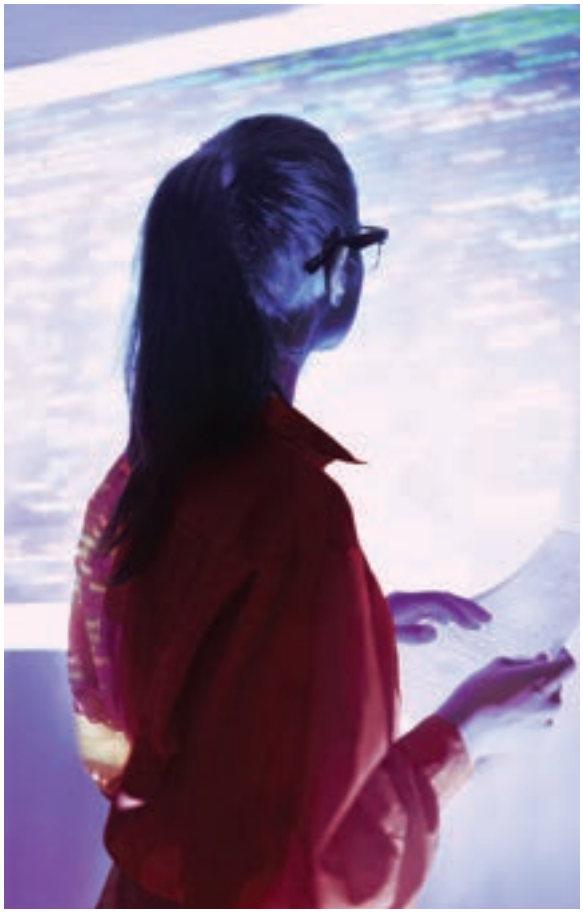
Realize o potencial do GenAI com a Lenovo e a NVIDIA

A Lenovo e a NVIDIA uniram forças para fornecer uma ampla gama de soluções e serviços projetados para impulsionar a adoção global do GenAI e ajudar os clientes a realizar seu pleno potencial. Desfrute de uma vantagem competitiva e de um futuro acelerado com soluções e serviços rápidos, seguros, escaláveis e abrangentes, respaldados por know-how, líder na indústria e serviços profissionais.

Em colaboração, a Lenovo e a NVIDIA oferecem o portfólio de IA mais abrangente, experiência comprovada e suporte consultivo, desbloqueando o poder do GenAI para cada setor e ajudando os clientes a trabalhar em direção a um futuro mais inteligente e rápido.

Aumente as oportunidades de receita, otimize produtividade e custos e mitigue riscos com um modelo privado GenAI alimentado pela Lenovo e pelo portfólio de soluções e serviços prontos para IA da NVIDIA:

- **Aproveite soluções inovadoras de IA:** Tire vantagem do compromisso de US$ 100 milhões da Lenovo com IA e alcance resultados de próximo nível. Com mais de 150 soluções prontas para IA de quase 50 parceiros de IA, o programa Lenovo AI Innovators oferece o tempo mais rápido para a solução em cada setor.
- **Otimize a infraestrutura de IA:** Implante um portfólio líder em IA e ofereça IA onde o negócio precisa com o menor Custo Total de Propriedade — do edge para a nuvem. Impulsione o desempenho inovador com uma solução privada GenAI respaldada por servidores acelerados, rede rápida, armazenamento confiável e plataforma de software de IA da NVIDIA.
- **Possibilite a descoberta de IA:** Trabalhe com os especialistas em IA da Lenovo e da NVIDIA para obter o máximo valor, reduzindo os riscos do projeto. A Lenovo está ultrapassando limites na vanguarda da IA há quase uma década. Beneficie-se do Lenovo AI Discover Lab, workshops de avaliação de IA e um comitê de IA impulsionando a adoção de IA para clientes em todos os continentes.



Adote a IA mais rapidamente com as Soluções Lenovo NVIDIA

Inovação em Soluções de IA

Em parceria com inovadores em IA, como DeepBrain — vídeos de IA para vendas e atendimento ao cliente, Chooch — visão computacional de IA para controle de qualidade e segurança, e Edgebricks — infraestrutura de IA para implementação mais rápida de IA, a Lenovo oferece um ecossistema de soluções empresariais de IA personalizáveis por meio do programa AI Innovators. O programa proporciona IA para operações de ponta a ponta, incluindo reconhecimento de áudio, previsão, segurança e assistentes virtuais para todas as indústrias, como finanças, varejo, manufatura e saúde.

O portfólio de infraestrutura otimizado para IA mais abrangente do setor

A Lenovo lançou um portfólio de soluções de infraestrutura de IA projetadas para alimentar desempenho de alto nível do edge à nuvem, atendendo às crescentes demandas do mercado para o GenAI. A pilha de computação acelerada Lenovo e NVIDIA permite que todas as indústrias aproveitem o poder da IA, proporcionando o desempenho, escala e eficiência necessários para executar a próxima onda de aplicativos. É uma plataforma de pilha completa que possibilita inovação e criatividade para resolver os desafios mais difíceis do mundo.

Lenovo | NVIDIA

Infraestrutura otimizada para IA

Servidores de Data Center

Lenovo ThinkSystem SR675 V3

O Lenovo ThinkSystem SR675 V3 é um versátil servidor de rack 3U rico em GPU que suporta oito GPUs de largura dupla, incluindo as novas GPUs Tensor Core NVIDIA H100 e L40S ou a oferta NVIDIA HGX H100 4-GPU com NVLink e refrigeração híbrida líquida-ar Lenovo Neptune.

O servidor oferece desempenho otimizado para IA, Computação de Alto Desempenho (HPC) e workloads gráficos, permitindo que os usuários extraiam insights mais profundos e impulsionem a inovação utilizando aprendizado de máquina e aprendizado profundo.

Servidores de Data Center

Lenovo ThinkSystem SR670 V2

Alimentado pelas novas GPUs Tensor Core NVIDIA H100 e L40S ou a oferta NVIDIA HGX H100 4-GPU, o Lenovo ThinkSystem SR670 V2 é otimizado para IA, HPC e workloads gráficas para uma ampla variedade de aplicações, incluindo GenAI.

O SR670 V2 oferece uma solução de nível empresarial, acelerando workloads em produção e maximizando o desempenho do sistema. As indústrias de varejo, manufatura, serviços financeiros e saúde aproveitam o SR670 V2 para aprimorar processos e impulsionar a inovação.

Servidor Edge

Lenovo ThinkEdge SE455 V3

O servidor Lenovo ThinkEdge SE455 V3 traz uma nova abordagem modular para edge computing e poder de processamento pronto para IA, armazenamento e rede mais próximos de onde os dados são gerados.

Como o servidor otimizado para edge principal da Lenovo, o SE455 V3 com GPUs NVIDIA L40 ou L4 é ideal para workloads grandes e exigentes de IA no edge, apresentando desempenho líder de mercado e sustentabilidade incorporada.

Componentes de Suporte

NVIDIA ConnectX-7

A NVIDIA ConnectX-7 SmartNIC é otimizada para fornecer rede acelerada para moderna nuvem, inteligência artificial e workloads empresariais tradicionais. O ConnectX-7 fornece um amplo conjunto de capacidades de rede, armazenamento e segurança definidas por software e aceleradas por hardware, permitindo que as organizações modernizem e protejam suas infraestruturas de TI.

Infraestrutura como Serviço (IaaS)

O Lenovo TruScale for AI ajuda as empresas a alcançar desempenho premium com um investimento inicial limitado. O modelo IaaS oferece acesso imediato à implementação de IA e uma conexão única para mais de 150 soluções prontas para IA da Lenovo, acelerando a transformação inteligente. O TruScale oferece serviços escaláveis de ponta a ponta, desde a implementação até a gestão e a atualização, proporcionando aos clientes um modelo de pagamento mensal previsível.

Plataforma de Inferência de IA da NVIDIA

A plataforma de inferência de IA da NVIDIA oferece uma pilha completa e uma variedade de produtos, infraestrutura e serviços, incluindo a NVIDIA AI Enterprise, para fornecer o desempenho, eficiência e responsividade que são críticos para impulsionar a próxima geração de inferência de IA — na nuvem, no data center, no edge ou em dispositivos embarcados.

NVIDIA AI Enterprise

O software de nível empresarial que alimenta a plataforma de IA da NVIDIA, a NVIDIA AI Enterprise acelera a ciência de dados. Simplifica o desenvolvimento e a implementação de IA generativa pronta para produção, visão computacional, IA de fala e muito mais. Empresas que conduzem seus negócios com IA contam com a NVIDIA AI Enterprise para melhorar a produtividade das equipes de IA e obter insights empresariais mais rapidamente.

NVIDIA NeMo

Parte da NVIDIA AI Enterprise, o NVIDIA NeMo permite que organizações construam modelos linguísticos grandes (LLMs) personalizados a partir do zero, personalizem modelos pré-treinados e os implementem em escala. O NeMo inclui estruturas de treinamento e inferência, kits de ferramentas de proteção, ferramentas de curadoria de dados e modelos de IA pré-treinados.

Possibilitando a Descoberta de IA

Adote a IA mais rapidamente com especialistas em IA, workshops e melhores práticas. O Lenovo AI Discover Lab fornece acesso a cientistas de dados da Lenovo, arquitetos de IA e engenheiros para ajudar a explorar, implementar e escalar soluções de IA. O serviço orienta os clientes para os parceiros de software mais apropriados e infraestrutura otimizada para IA, compartilhando conhecimentos desenvolvidos a partir dos investimentos e inovações combinados em IA da Lenovo e NVIDIA.

A Lenovo oferece workshops de avaliação para impulsionar a adoção de IA e orientação responsável de IA para ajudar as organizações a compreender e abordar considerações de privacidade, uso justo, diversidade, equidade, inclusão e acessibilidade por meio do Comitê de IA Responsável da Lenovo.

Em parceria, a Lenovo e a NVIDIA estão ajudando os clientes a aproveitar o valor de seus dados para implementar soluções de IA projetadas com rapidez, transformando organizações com resultados mais previsíveis.

Converse com os especialistas da Lenovo e NVIDIA para iniciar mais rapidamente sua jornada de IA.

Entre em contato com o Laboratório de Descoberta de IA da Lenovo em AIDiscover@lenovo.com para agendar uma consulta.

Lenovo e NVIDIA

Em parceria com a NVIDIA, a Lenovo está desenvolvendo tecnologias que estão transformando o mundo e compartilhando sua experiência combinada por meio de serviços profissionais para criar uma sociedade mais eficiente, conectada e digital. Ao projetar e desenvolver o portfólio mais completo do mundo de dispositivos e infraestrutura inovadores prontos para IA, e impulsionar a adoção por meio de serviços de consultoria e suporte, a Lenovo e a NVIDIA estão liderando uma Transformação Inteligente — para criar melhores experiências e oportunidades para milhões de clientes em todo o mundo.

A aceleração da IA depende de uma infraestrutura acelerada e de software poderoso, e a NVIDIA oferece aceleração onde quer que seja necessário — em data centers, desktops, laptops e nos supercomputadores mais rápidos do mundo. À medida que as empresas se tornam cada vez mais orientadas por dados, a demanda por tecnologia de IA cresce. A IA proporciona às equipes empresariais o poder, as ferramentas e os algoritmos para trabalhar de forma eficaz, desde chatbots de atendimento ao cliente até comunicação hiperpersonalizada e otimização de produção automatizada.

Lenovo e NVIDIA trazem soluções inovadoras e infraestruturas inteligentes para resolver os desafios mais significativos de hoje e amanhã. Juntas, equipamos pesquisadores, pioneiros e visionários centrados em dados em todas as indústrias com as ferramentas para evoluir, transformar e implementar soluções empresariais de IA para oferecer uma Tecnologia mais Inteligente para Todos.

Saiba mais

Vamos começar

Nunca houve um momento melhor para investir em IA

Entre em contato hoje

Aproveite a experiência e o poder da Lenovo e NVIDIA

Agende uma oficina estratégica de negócios de IA

Entre em contato com a equipe da Lenovo para iniciar sua jornada de IA

Fale com os especialistas em IA

Referências

[1] Statista, 2021, Volume de dados/informações criados, capturados, copiados e consumidos em todo o mundo de 2010 a 2020, com previsões de 2021 a 2025

[2] Departamento de Ciência da Computação da Universidade de Chicago, 2020, Globus atinge a marca de um exabyte em gerenciamento de dados de pesquisa

[3] Scality, O que é um exabyte, afinal?

[4] MarketsandMarketsTM Research Private Ltd., Mercado de Inteligência Artificial (IA) por Oferta (Hardware, Software), Tecnologia (ML (Aprendizado Profundo (LLM, Transformadores (GPT 1, 2, 3, 4)), PNL, Visão Computacional), Função Empresarial, Vertical e Região — Previsão Global até 2030

[5] McKinsey & Company, 2023, Qual é o futuro da IA generativa? Uma visão inicial em 15 gráficos

[6] Bloomberg, 2023, A IA generativa se tornará um mercado de US$ 1,3 trilhão até 2032, revela pesquisa

[7] McKinsey Digital, 2023, O potencial econômico da IA generativa: A próxima fronteira da produtividade

[8] Lenovo, 2023, Arquitetura de Referência para IA Generativa Baseada em Modelos de Linguagem Grandes (LLMs)

[9] Avenga, 2022, Inteligência Artificial (IA) para Gerenciamento de Risco de Crédito em Bancos

[10] Deloitte, 2023, Perspectivas bancárias e de mercados de capitais para 2024

[11] McKinsey & Company, 2023, Cinco maneiras de impulsionar o crescimento liderado por experiências na área bancária

[12] IDC, 2023, Como Varejistas e Marcas Estão Aproveitando a IA Generativa

[13] Accenture, 2023, 6 Benefícios Cruciais da IA para o Varejo (+ Casos de Uso de Principais Marcas)

[14] Statista, 2023, Comércio eletrônico como porcentagem das vendas no varejo mundial de 2015 a 2027

[15] IDC, 2023, Como a IA Generativa Está Impactando as Indústrias

[16] McKinsey & Company, 2021, A Internet das Coisas: Acompanhando uma oportunidade acelerada

[17] Lumen Technologies, 2021, Edge Computing: Serviços para Manufatura

[18] Deloitte, 2023, Perspectivas da indústria de manufatura para 2023

[19] Gartner, 2023, Além do ChatGPT: O Futuro da IA Generativa para Empresas

[20] CNA, 2017, IA, Robôs e ENXAMES: QUESTÕES, PERGUNTAS E ESTUDOS RECOMENDADOS

[21] OCDE, 2021, A importância do aumento do acesso a dados de saúde de alta qualidade

[22] Lenovo, 2023, Movendo a IA da Ideia para a Execução