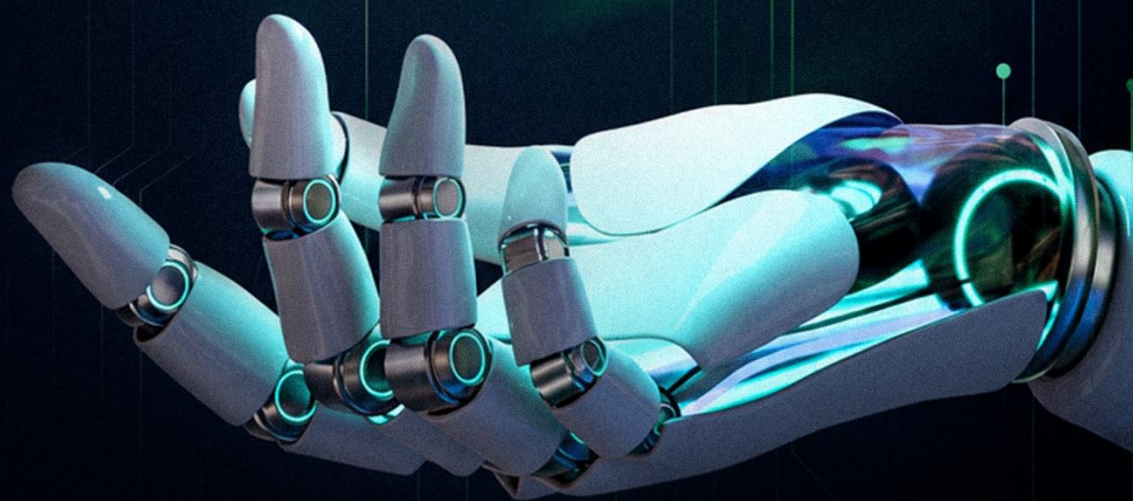


Deloitte.
Together makes progress



A nova equação da IA

Valor, custo e soberania na decisão de arquitetura

Uma nota do autor:

À medida que a inteligência artificial se expande nas organizações, as decisões de arquitetura ficam mais complexas. Elas precisam equilibrar o valor gerado, a previsibilidade de custos e o grau de soberania necessário para controlar riscos regulatórios, jurídicos, geopolíticos e de proteção da propriedade intelectual. O tema se torna multidisciplinar conforme ganha escala, mas essas dimensões ainda costumam ser tratadas separadamente. Assim, este estudo adota uma visão integrada das variáveis que envolvem o uso da inteligência artificial e apresenta uma metodologia para apoiá-las de forma consistente. Foi concebido para líderes de tecnologia, finanças e riscos, bem como para todas as áreas cuja relevância estratégica cresce à medida que a inteligência artificial se torna um elemento central nas organizações. Trata-se do primeiro de três estudos complementares: o segundo detalhará a camada de abstração responsável por traduzir políticas em operações, e o terceiro será dedicado às fábricas de IA.

SOBRE O ESTUDO E SUA METODOLOGIA

Este estudo executivo tem como objetivo apoiar decisões de arquitetura de inteligência artificial por meio da análise integrada de valor, custo e soberania. A metodologia reúne referências de mercado, evidências setoriais e regulatórias e experiência prática, aplicando critérios comuns aos diferentes arranjos de consumo e controle considerados. As comparações são realizadas com base em faixas e premissas declaradas, considerando benefícios esperados, custo total, exposições, controles e condições de mudança e saída. Para demonstrar a aplicação do método, o estudo apresenta um caso ilustrativo composto, que distingue os dados de referência das hipóteses adotadas. Os resultados refletem a aplicação desses critérios e premissas, sem estabelecer preferência prévia por fornecedor, tecnologia, arquitetura ou modelo de contratação.

A equação em uma página

Com a expansão da inteligência artificial nas organizações, as decisões de arquitetura passam a abranger, além dos aspectos técnicos, seus efeitos sobre o orçamento, a geração de valor e o grau de soberania requerido.

Essas três dimensões costumam ser tratadas isoladamente: a área de tecnologia controla o custo; as áreas de negócio respondem pela captura de valor; as áreas de segurança, jurídico, dados, compliance e operação identificam as exposições que precisam ser controladas. Como cada dimensão é analisada separadamente, projetos pilotos avançam sem um retorno claro, contratos crescem sem previsibilidade e escolhas de arquitetura são feitas antes da total compreensão do grau de soberania necessário para cada tipo de uso.

Para cada família de aplicações, o método analisa a decisão arquitetural a partir de três perguntas: que valor essa família pode gerar, quanto custa sustentá-la e que grau de soberania é considerado necessário para controlar suas exposições. Como as respostas variam de acordo com o uso, diferentes famílias podem requerer abordagens distintas: algumas priorizam velocidade e acesso a modelos de fronteira; outras, previsibilidade, governança e maior controle. O método parte dessa distinção.

A aplicação dessa lógica resulta em seis implicações práticas:

- **A soberania é analisada por família, não apenas pelo portfólio.** No método proposto, o custo adicional de controle é comparado à redução das exposições relevantes, à preservação de alternativas futuras e aos custos de mudança. A aplicação uniforme de controles a todo o portfólio pode resultar na adoção de proteções semelhantes para famílias com perfis de exposição distintos; por isso, a metodologia associa o grau de soberania às características de cada família.
- **A avaliação de valor utiliza uma linha de base e considera o valor de opção.** Sem uma medida inicial da tarefa, a análise de impacto fica limitada à percepção sobre o piloto. Além da eficiência imediata, algumas famílias têm finalidade exploratória, voltada à experimentação, à inovação e ao desenvolvimento de capacidades com potencial de geração de valor futuro, sob limites definidos de gasto, exposição e governança.
- **A redução do custo unitário não implica, necessariamente, menor orçamento total.** Preços inferiores podem ampliar o uso, habilitar novos casos e aumentar a autonomia dos agentes. Como resultado, o gasto agregado pode crescer mesmo quando o custo por unidade diminui.
- **A arquitetura é avaliada também diante de mudanças de cenário.** Preços, modelos, regras e exposições mudam em prazos diferentes, e nenhuma previsão cobre todos os casos. Quando cada família tem premissas declaradas e gatilhos de revisão, parte das mudanças pode ser absorvida por parâmetros de política, contingência e substituição;

outras podem exigir adaptação, migração ou revisão contratual. A explicitação dessas dependências pode ampliar a capacidade de adaptação e reduzir a necessidade de redesenho integral.

- **A camada de abstração integra a governança de IA.** A política definida para cada família torna-se operacional quando aplicada às chamadas correspondentes. A camada de abstração, apresentada adiante, concentra a execução das regras, a medição do consumo e a trilha de auditoria.
- **No Brasil, a análise incorpora condições locais.** A participação de fontes renováveis pode influenciar a avaliação de cargas intensivas de inferência quando a energia está disponível no local e no prazo requeridos. Incertezas relacionadas a conexão, infraestrutura e incentivos tributários são tratadas como variáveis de risco, e não como premissas garantidas.

Três variáveis, uma decisão

Neste estudo, a unidade de análise é formada por famílias de aplicações que compartilham perfil de exposição, complexidade, volume, necessidade de experimentação e requisitos técnicos. Esse nível intermediário permite considerar diferenças que podem ser ocultadas tanto pela avaliação isolada de cada aplicação quanto pela adoção de uma única arquitetura corporativa.

A escolha de unidade importa. Uma aplicação isolada pode ser detalhada demais para orientar investimento, mas a arquitetura corporativa tende a ser ampla demais para capturar diferenças relevantes. Neste método, a família de aplicações é adotada como unidade de comparação que permite medir custo, estimar valor e definir o grau de soberania necessário para controlar as exposições daquele conjunto.

As diferentes famílias de aplicações apresentam requisitos e prioridades específicos. Algumas dependem de velocidade, acesso a modelos de fronteira e espaço para experimentar, enquanto outras requerem mais previsibilidade, supervisão, proteção de dados e rota de saída. Há ainda famílias em que agentes elevam consumo e exposição ao mesmo tempo. Nesses casos, valor, custo e soberania devem ser considerados de forma integrada no processo decisório.

O estudo The GenAI Divide: State of AI in Business 2025 indica, na amostra analisada, que 95% das organizações não obtiveram retorno verificável de suas iniciativas de inteligência artificial generativa.¹ Em paralelo, três quartos dos líderes globais pesquisados declararam preocupação com o risco geopolítico associado aos dados, e dois terços informaram que já revisam suas estratégias.² Esses resultados evidenciam a presença das dimensões de valor, custo e soberania na agenda executiva e a necessidade de analisá-las de forma integrada.

No referencial proposto, o ajuste de valor, custo e soberania às características de cada família permite avaliar de forma integrada espaço para experimentação, exposição e capacidade de adaptação a diferentes cenários.

A seção seguinte aborda o aspecto mais tangível da decisão: os arranjos de consumo. Antes de comparar custo, valor e soberania, é preciso entender o que se contrata quando se consome inteligência artificial — quem fornece o modelo, quem opera a computação, quem assume a variância do consumo e onde ficam os pontos de controle.

Os arranjos de consumo da IA

O mercado de inteligência artificial é altamente dinâmico, mas as decisões de contratação continuam a girar em torno de algumas perguntas fundamentais: Quem fornece o modelo? Quem opera a computação? Quem absorve a variância do consumo? Onde ficam os pontos de controle? Essas questões ajudam a compreender os diferentes arranjos, porque revelam como se distribuem as duas camadas centrais da IA: o modelo, que produz a inteligência, e a computação, que processa a demanda.

A1 — API direta ao provedor de modelo. A organização consome o modelo diretamente do fornecedor e paga pelo uso. É um arranjo frequentemente utilizado para testar rapidamente e acessar os lançamentos do provedor sem intermediação. Em contrapartida, a aplicação tende a se acoplar à interface, aos termos e à operação desse fornecedor. A troca pode exigir adaptação técnica, nova avaliação e revisão de controles. Serve a pilotos, testes e experimentação rápida quando essa dependência é aceitável.

A2 — Serviços de IA na nuvem. O modelo é escolhido no catálogo do provedor de nuvem e executado dentro do perímetro contratado. O arranjo oferece integração com a plataforma, controles comuns e acesso a diferentes modelos sob uma relação operacional e contratual compartilhada. Essa variedade pode reduzir a vinculação a um único fornecedor de modelos, embora permaneça condicionada ao catálogo, ao ritmo de atualização e às condições de controle da plataforma. A flexibilidade decorre da combinação de serviços dentro do ecossistema, acompanhada de vinculação operacional e tecnológica a esse ambiente.

Nos dois primeiros arranjos, a organização compra conveniência. Nos dois seguintes, assume maior controle, governança e responsabilidade.

A3 — Modelo auto-hospedado em nuvem. A organização contrata capacidade computacional e executa nela um modelo próprio, aberto ou ajustado ao domínio. O controle aumenta; a responsabilidade por operação, segurança, versionamento e confiabilidade também.

A4 — Modelo em infraestrutura própria. Computação, modelo e operação ficam no perímetro da organização. É o arranjo de maior controle e maior compromisso: capital, equipe, energia, ciclo de vida tecnológico e risco de ociosidade entram na conta.

Há ainda uma escolha transversal: quem opera a fábrica de IA. A fábrica é o conjunto que transforma computação em inteligência — modelos, inferência, treinamento, plataformas e aplicações. No vocabulário global do mercado, é a AI factory. Ela pode estar em nuvem, infraestrutura própria, colocation ou combinação desses ambientes. E pode ser operada pela própria organização ou por um terceiro.

A5 — Fábrica de IA como serviço. A organização consome a fábrica como serviço de um operador especializado. A posse da infraestrutura pode variar; o resultado é acesso à capacidade de fábrica sem construir e operar tudo sozinha. Fábricas próprias, consorciadas e como serviço são o tema do terceiro estudo desta série.

FIGURA 1
Posse e operação nos arranjos de consumo

	CUSTO DA COMPUTAÇÃO	DONO DO MODELO	QUEM OPERA
A1 · API direta ao provedor de modelo contrato direto com o fornecedor; a aplicação tende a se acoplar à interface, aos termos e à operação	ao provedor do modelo	provedor	provedor
A2 · Serviços de IA na nuvem catálogo do provedor de nuvem; variedade de modelos sob uma relação operacional e contratual comum	ao provedor de nuvem	catálogo do provedor	provedor
A3 · Modelo auto-hospedado em nuvem capacidade contratada; modelo próprio, aberto ou ajustado ao domínio	por capacidade (GPU/h)	organização	organização
A4 · Modelo em infraestrutura própria computação, modelo e operação no perímetro da organização	capital próprio	organização	organização
A5 · Fábrica de IA como serviço posse e operação deixam de precisar estar na mesma organização	infraestrutura da organização operada por um terceiro		infraestrutura do operador consumida como serviço

A IA embarcada em software de terceiros fica fora da taxonomia; o inventário do método a captura.

Os arranjos tornam comparáveis as formas de consumir inteligência artificial. Os quatro primeiros mostram onde estão o modelo, a computação e quem assume a responsabilidade econômica, enquanto o quinto introduz uma decisão adicional: possuir a infraestrutura e contratar um terceiro para operá-la, ou consumir uma fábrica cuja infraestrutura também

pertence ao operador. Assim, propriedade e operação não precisam ser responsabilidade da mesma organização.

A computação na borda pode aparecer em qualquer um desses arranjos quando latência, residência de dados ou processamento em tempo real são determinantes. Esse formato tem avançado e se mostrado uma opção relevante para aplicações que precisam processar dados mais perto de sua origem.

A partir daqui, é importante separar duas decisões: o arranjo de consumo e o regime de controle. O primeiro descreve como o modelo e a computação são contratados e quem assume a operação e a variância econômica. O segundo estabelece as salvaguardas oferecidas em termos de dados, operação, jurisdição e tecnologia. Embora relacionadas, essas dimensões não são equivalentes: um mesmo arranjo pode estar associado a diferentes regimes de controle. Essa distinção permite avaliar cada família de aplicações a partir de uma mesma lógica, considerando o valor que habilita, o custo que impõe e o grau de soberania que sustenta.

A primeira variável a ser calculada é o custo, pois é a mais tangível no orçamento. Mas, sua visão, isoladamente, é limitada; sua capacidade de orientar a decisão depende da análise conjunta da tarefa, da variância do consumo e do regime de controle que cada arranjo determina.

O custo: do token à tarefa

O custo da inteligência artificial parece simples quando aparece como preço por token. Na prática, porém, ele raramente se resume a esse indicador. Os modelos de contratação combinam consumo variável, créditos, franquias, pacotes, descontos, capacidade provisionada e compromissos corporativos. Nesse contexto, a pergunta fundamental não é apenas quanto custa um token, mas como o consumo se traduz em variabilidade, previsibilidade financeira e compromisso de longo prazo.

Para compreender os modelos de custo, o primeiro passo é distinguir token de tarefa. O token permite comparar preços entre modelos e arranjos; a tarefa, comparar resultados. O mesmo volume de tokens pode gerar entregas substancialmente diferentes dependendo do modelo utilizado. Com isso, uma mesma tarefa pode consumir quantidades muito distintas de tokens em função do desenho da aplicação, do uso de cache, do roteamento entre modelos, do contexto carregado e da autonomia dos agentes. Por este motivo, o método considera as duas medidas: a do token, para avaliar eficiência econômica, e a da tarefa, para governar custos a partir do valor efetivamente entregue ao negócio. Antes de comparar configurações, é preciso avaliar se a arquitetura atual já eliminou desperdícios evidentes. Direcionar cada solicitação ao menor modelo capaz de resolvê-la, prática conhecida como roteamento em cascata, pode reduzir o custo por tarefa em múltiplos. O mesmo vale para cache de contexto, processamento assíncrono, redução de contexto e quantização, isto é, a redução da precisão numérica do modelo. A sequência importa: primeiro, otimiza-se a base de consumo; depois, avaliam-se as diferentes configurações e o prêmio associado a maiores níveis de proteção, discutido nas seções seguintes.

Cada arranjo transforma a incerteza de consumo em um compromisso diferente. Nas APIs diretas e nos serviços de IA na nuvem, a contratação tende a ser simples. A cobrança pode assumir diferentes formatos como token, pacote, franquia ou contrato corporativo, mas em todos os casos o orçamento permanece sensível a oscilações de volume, de comportamento de uso e de preços. A introdução de agentes amplia essa incerteza, uma vez que parte do consumo passa a ser decidida em tempo real durante a operação. No modelo auto-hospedado, a organização contrata capacidade, e a viabilidade econômica depende da taxa de utilização dos recursos contratados. Na infraestrutura própria, a lógica econômica se inverte: o custo fixo aumenta, o custo marginal cai e a variabilidade dá lugar à previsibilidade, com risco de ociosidade. A fábrica consumida como serviço combina elementos dos dois lados: previsibilidade contratada, sem compromisso integral de capital, com parte relevante da operação transferida a um terceiro.

A expansão da inteligência artificial também acrescenta uma dimensão de custo que não se atribui facilmente a uma tarefa ou família de aplicações: a atração, o desenvolvimento e a

retenção de talentos. Essa dimensão ganha relevância econômica à medida que a organização amplia sua atuação em áreas que requerem profissionais especializados, como desenvolvimento de modelos, operação de plataformas, segurança e governança. Ao mesmo tempo, os talentos reunidos e desenvolvidos nesse percurso ampliam a capacidade de experimentar, inovar e conduzir a transformação. Como seus custos e benefícios se distribuem entre diferentes iniciativas e se materializam ao longo do tempo, essa dimensão exige uma abordagem própria de mensuração. Por isso, embora reconhecida neste estudo, ela não integra o cálculo por família e será aprofundada nos demais estudos da série.

Com agentes, parte do consumo é decidida em tempo real durante a execução; controlar a variabilidade torna-se parte do custo.

A escolha entre modelos de fronteira ou modelos menores especializados atravessa todos esses arranjos. Os compactos, ajustados ao domínio, por exemplo, podem entregar resultado suficiente com menor custo e mais controle. A decisão, portanto, não é apenas qual deles utilizar, mas que combinação de modelo, capacidade e regime de controle sustenta cada família de aplicações.

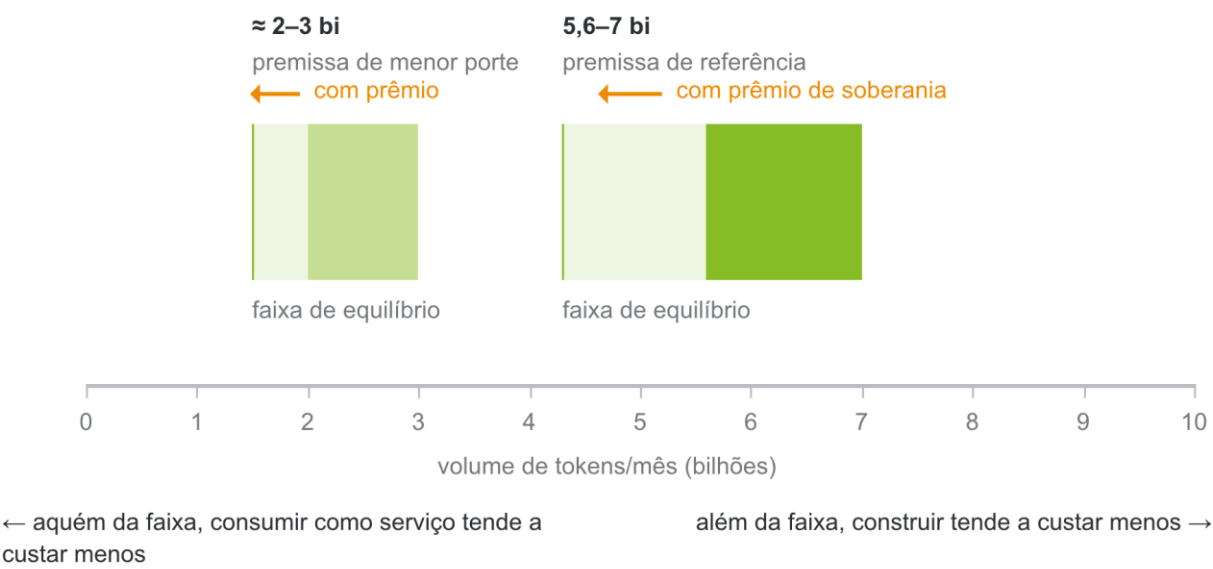
A faixa móvel do equilíbrio

A decisão entre consumir IA como serviço e contratar ou construir capacidade dedicada é frequentemente resumida a uma pergunta: quanto vale migrar de um modelo para outro? A resposta, porém, não se traduz em um valor único. O ponto de equilíbrio depende da combinação entre tecnologia disponível, utilização, custos, estrutura comercial do serviço e exigências de controle. Este estudo, adota as faixas de referência, em vez de um único ponto de equilíbrio, tomando como base as análises apresentadas pela Deloitte em *The Pivot to Tokenomics: Navigating AI's New Spend Dynamics* (2026).³

FIGURA 2

A faixa móvel do equilíbrio

O ponto de equilíbrio depende da premissa de hardware – e os requisitos de soberania o antecipam na medida do prêmio.



O ponto de equilíbrio é o volume em que o custo unitário da infraestrutura própria iguala o preço do serviço compatível com o controle exigido.

Faixa da direita: nó de aceleradores de geração atual, com custo total de propriedade completo. Faixa da esquerda: nó único de menor porte servindo modelo ajustado ao domínio, comparado ao serviço equivalente na tarefa. Deslocamentos ilustrativos, proporcionais ao prêmio de cada caso; premissas na nota 3.

A deflação tecnológica move a faixa primeiro. Conforme o Artificial Intelligence Index Report (2025), entre novembro de 2022 e outubro de 2024, o custo de inferência para um nível fixo de desempenho caiu de US\$ 20,00 para US\$ 0,07 por milhão de tokens, uma redução superior a 280 vezes. O hardware melhora em ritmo próprio: cada dólar gasto em capacidade de processamento compra cerca de 30% a mais de desempenho por ano.⁵ Ou seja, um ponto de equilíbrio estimado hoje pode precisar ser revisto em poucos trimestres.

O uso segue o mesmo comportamento. Quando a inferência fica mais barata, novas aplicações passam a fazer sentido e o consumo total tende a crescer. A passagem do assistente para o agente acelera esse movimento, porque cada interação pode gerar novas chamadas, novos contextos e novas decisões em tempo de execução. O preço unitário cai, mas a conta total pode subir.

No método proposto, uma família de aplicações que requer maior controle é comparada à alternativa compatível com o grau de soberania definido para seu perfil, e não apenas ao serviço convencional de menor preço. Quando essa alternativa incorpora prêmio de soberania, a faixa de equilíbrio da capacidade dedicada ou própria pode ocorrer antes do que indicaria a comparação com o serviço convencional.

Esse deslocamento depende do **prêmio de soberania**: o sobrecusto da configuração que oferece mais proteção do que a referência econômica convencional, para volume e resultado equivalentes. A Comissão Europeia identificou prêmio médio próximo de 15% em

serviços soberanos comparados à região convencional equivalente do mesmo provedor; em licitações analisadas sob confidencialidade, as diferenças variaram de cerca de 10% abaixo a 12% acima para níveis semelhantes de soberania.²⁶ A ordem de grandeza orienta a análise, mas o prêmio precisa ser recalculado para cada degrau, geografia, operador, configuração e família.

Neste estudo, a faixa de equilíbrio é avaliada com base no volume das famílias de aplicações que requerem o nível de controle oferecido pela capacidade dedicada, considerando o custo da alternativa compatível, e não necessariamente o volume total da organização ou o arranjo de menor preço do portfólio. Quanto maior a eficiência na alocação de cargas, menor tende a ser o custo médio e mais distante pode ficar o ponto de equilíbrio. Nessa lógica, a otimização pode reforçar a atratividade econômica do consumo como serviço, enquanto a capacidade dedicada é analisada em função dos requisitos de controle de determinadas cargas.

No referencial adotado, o ponto de equilíbrio funciona como gatilho de revisão, e não como decisão automática de investimento. Em um mercado que evolui rapidamente, a capacidade própria é tratada como uma alternativa estratégica que pode ser avaliada e contratada em etapas, de acordo com volume, utilização, exposição e capacidade operacional. Enquanto as condições para uma eventual migração são avaliadas, o conjunto de opções pode incluir consumo de IA como serviço, fábrica como serviço ou capacidade consorciada. A análise considera o momento, a escala e o grau de soberania associados a cada alternativa.

O ponto de equilíbrio alerta para o momento de mudar de configuração. Determinar se a família gera valor compatível com o que consome é, frequentemente, um aspecto difícil da equação.

O valor: da percepção à métrica

Entre as três variáveis da equação, o custo costuma ser a mais visível. Ele se materializa em contratos, faturas, consumo de tokens e capacidade contratada. O valor, por sua vez, é mais difícil de observar: precisa ser atribuído a uma tarefa, comparado a uma linha de base e reconhecido por quem responde pelo resultado. Somente nesse momento ele se torna visível nos resultados do negócio, nos indicadores gerenciais ou na percepção do cliente. Essa assimetria ajuda a explicar parte da frustração atual com a inteligência artificial: a adoção avança, mas os benefícios nem sempre se traduzem em resultados financeiros, ganhos de eficiência ou na melhoria da experiência do cliente.

O estudo The GenAI Divide: State of AI in Business 2025 apontou que, na amostra analisada e apesar dos investimentos entre US\$ 30 bilhões e US\$ 40 bilhões em inteligência artificial generativa empresarial, 95% das organizações não obtiveram retorno verificável.¹ Segundo o levantamento, iniciativas de IA apresentaram maior dificuldade de gerar resultados sustentáveis quando não foram acompanhadas por desenho de processos, métricas claras e responsabilização pela captura de valor. Na mesma amostra, os investimentos se concentraram nas funções de contato com o cliente, enquanto os ganhos mais consistentes foram observados no backoffice. O relatório também registrou taxas de sucesso superiores em iniciativas conduzidas com parceiros externos em comparação com construções internas, resultado que deve ser interpretado nos limites metodológicos e amostrais indicados na nota correspondente.

O estudo State of AI in the Enterprise 2026: The Untapped Edge, conduzido pela Deloitte, revela que um quarto das organizações relata o impacto transformador da adoção de IA, o dobro do ano anterior, e 84% ampliaram seus investimentos. A adoção de IA agêntica, por sua vez, deve saltar de 23% para 74% em dois anos. A principal lacuna está na sustentação desse avanço: apenas 21% das organizações dispõem de governança madura para agentes autônomos, e o risco mais citado é privacidade e segurança de dados.⁶ A adoção da tecnologia avança mais rapidamente do que a governança dos agentes autônomos e a avaliação de seus impactos. Por isso, valor e soberania precisam evoluir juntos: medir valor permite capturar impacto; governar a soberania possibilita escalar a IA sem escalar sua exposição além do necessário.

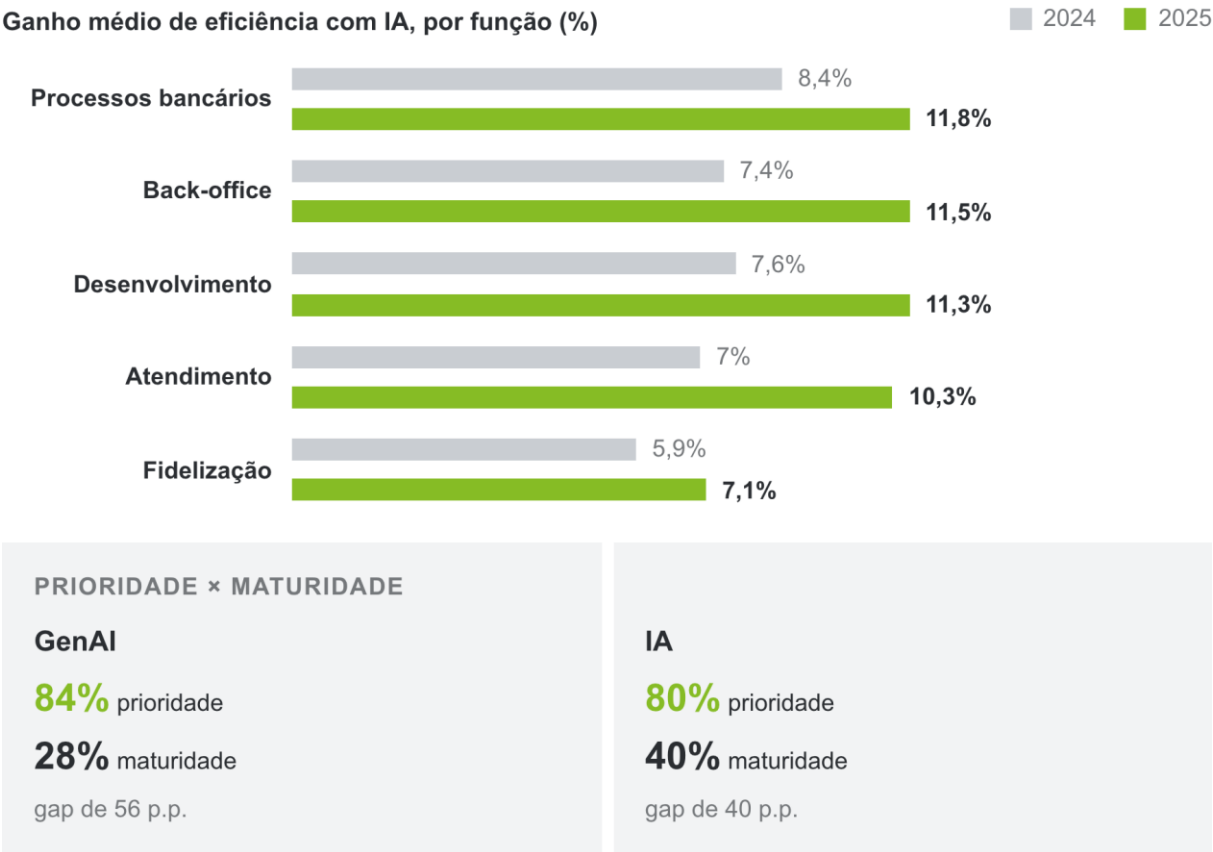
A série que a literatura global raramente tem

A Pesquisa Febraban de Tecnologia Bancária, conduzida com a Deloitte, oferece uma visão que mede os ganhos de eficiência dos bancos após a adoção de inteligência artificial e que acompanha sua evolução no setor. Entre 2024 e 2025, os ganhos médios cresceram em todas as funções medidas — backoffice, desenvolvimento de aplicações, processos

bancários e atendimento ao cliente. A parcela de bancos com ganhos elevados, de 20% ou mais, também aumentou, com avanço especialmente claro em backoffice e atendimento.

Os dados permitem observar a evolução dos ganhos ao longo do tempo. A medição recorrente transforma percepções pontuais em uma série passível de acompanhamento. A mesma pesquisa indica diferenças entre prioridade declarada e maturidade reportada: inteligência artificial e IA generativa receberam prioridade média ou alta de 84% dos bancos, enquanto 40% e 28%, respectivamente, relataram níveis de maturidade médios ou altos nessas áreas.⁴

FIGURA 3
A evidência bancária brasileira sobre medição e captura de valor



prioridade média ou alta; maturidade média ou alta

Fonte: Febraban e Deloitte, Pesquisa Febraban de Tecnologia Bancária 2026. Amostras equalizadas de 18 a 21 instituições nas séries de eficiência e de 25 em prioridade e maturidade; dados autorreportados.

Na série analisada para o setor bancário, a medição recorrente permite acompanhar a evolução dos ganhos reportados após a adoção de inteligência artificial.

A evidência de que o valor se torna mais observável quando medido de forma recorrente orienta o método deste estudo. Enquanto a medição amadurece, a contribuição econômica pode integrar a avaliação como julgamento estruturado, com prazo definido para conversão em métrica. A avaliação inicial de cada família registra os benefícios esperados e as respectivas premissas — horas liberadas, processos substituídos, receita incremental e capacidade estratégica habilitada. A redução de perdas esperadas é registrada separadamente no risco residual, para evitar dupla contagem.

Para medir melhor, é preciso primeiro nomear melhor. Neste estudo, os benefícios da inteligência artificial são observados por quatro componentes: **produtividade**, quando libera horas de trabalho; **custo evitado**, quando substitui processos ou reduz esforço operacional; **receita incremental**, quando cria ofertas, conversões ou vendas adicionais; e **valor de opção**, quando preserva a possibilidade de experimentar, inovar e desenvolver capacidades que ainda não têm retorno imediato, mas podem abrir caminhos de valor no futuro. A redução de perdas esperadas não é somada como benefício: ela aparece como diminuição do risco residual. Essa separação preserva a conexão entre valor e soberania sem contar duas vezes o mesmo efeito econômico.

A transição da percepção para a mensuração do valor é rápida quando sustentada por disciplina. A hipótese de valor é documentada, calibrada com referências externas, benchmarks setoriais e experiências internas e passa a ser monitorada em relação a uma linha de base previamente definida, com critérios de reconhecimento de ganhos alinhados entre as áreas de negócio e finanças. A lógica de priorização também precisa ser explícita: famílias que não demonstrem geração de valor perdem prioridade. Se, após dois ciclos de avaliação, a hipótese de valor não evoluir para métricas concretas, sua nota deve ser rebaixada nos ciclos subsequentes. A exceção são famílias sustentadas por valor de opção, como capacidades exploratórias e plataformas habilitadoras. Nesses casos, o acompanhamento ocorre por marcos de capacidade e a revisão exige deliberação de comitê. Esse agrupamento permite distinguir as aplicações experimentais das demais e submetê-las a uma governança própria, compatíveis ao seu potencial estratégico e estágio de maturidade.

Essa disciplina busca viabilizar a avaliação conjunta do valor com as demais variáveis da equação. A próxima seção examina o nível de soberania necessário e o prêmio associado aos controles adotados para reduzir cada exposição.

O espectro da soberania

Neste estudo, soberania é definida como o conjunto de decisões que relaciona exposições a controles. Em inteligência artificial, o conceito corresponde à escolha das camadas que requerem maior controle, daquelas que podem permanecer compartilhadas e do custo adicional associado a essa configuração.

A soberania é avaliada em quatro dimensões: **dados** abrangem requisitos de residência, acesso e legislação aplicável; **operação** define quem administra o ambiente, com qual nível de autonomia e sob quais mecanismos de controles; **jurisdição** determina o alcance legal sobre os agentes que operam a solução, do âmbito global ao perímetro nacional ou a um regime especial; **tecnologia** contempla propriedade, portabilidade e capacidade de substituição dos componentes subjacentes. Uma organização pode buscar elevada soberania em dados sem necessariamente operar o mesmo nível de controle sobre a operação, a jurisdição e a tecnologia.

A soberania costuma ser tratada como um atributo nacional: um Estado e as empresas que nele operam desenvolvem, treinam e utilizam inteligência artificial sob as próprias leis, infraestrutura e estruturas de governança de dados.⁶ Essa definição é relevante para a formulação de políticas públicas, mas oferece pouca granularidade para decisões empresariais. Na prática, as organizações não escolhem sua jurisdição de origem, escolhem o nível de controle que desejam exercer sobre cada família de aplicações e o custo associado a essa decisão. Ao adotar a perspectiva empresarial, a soberania deixa de ser um conceito geopolítico e se torna uma decisão de arquitetura, governança e investimento.

A estrutura tecnológica que sustenta a inteligência artificial demonstra a razão pela qual a soberania deve ser tratada como um espectro. Algumas dependências são difíceis de substituir no curto prazo. Em 2024, um único fabricante concentrava cerca de 92% do mercado de GPUs para data centers.⁷ Na fabricação avançada de semicondutores, uma única foundry respondia por cerca de 70% da receita das dez maiores do setor.⁸ A mesma dinâmica de concentração se repete na infraestrutura de nuvem e na capacidade computacional. Na Europa, os três provedores globais de nuvem somam 70% do mercado regional, enquanto todos os provedores europeus, reunidos, detêm 15%.⁹ Nos Estados Unidos, por sua vez, concentra-se cerca de 75% do desempenho agregado de supercomputadores de IA registrado na maior base pública do setor.¹⁰ Ou seja, diferentes níveis de dependência tendem a permanecer ao longo da cadeia. A decisão arquitetônica consiste em identificar quais dependências são aceitáveis, quais exigem mitigação e em quais camadas o investimento em maiores controles gera valor proporcional ao custo adicional.

FIGURA 4
A cadeia da IA concentra, camada por camada

Aplicações & agentes	a camada que a organização constrói	fragmentado
Modelos	fronteira e modelos abertos	poucos provedores concentram a fronteira
Computação de IA	desempenho agregado de supercomputadores	75%  EUA, na maior base pública do setor
Nuvem	três provedores globais, mercado europeu	70% 
Aceleradores (GPUs)	mercado de data center, 2024	92% 
Fabricação avançada	receita das dez maiores foundries	70% 
Energia & território	geração firme e capacidade instalada	o gargalo físico da corrida

Os dois valores de 70% medem bases distintas: participação dos três provedores globais no mercado europeu de nuvem e participação de uma única foundry na receita das dez maiores do setor. O valor de 75% é a participação na base de dados pública da Epoch AI.

Neste referencial, a decisão sobre soberania é analisada a partir da definição das exposições consideradas aceitáveis e daquelas para as quais são requeridos controles adicionais.

A regulação e a geopolítica passaram a influenciar também o desenho da arquitetura. As regras avançam em ritmos e direções diferentes: a União Europeia ajustou o cronograma do AI Act em 2026, a França alterou requisitos de hospedagem de dados sensíveis do Estado, o Brasil consolidou o regime de transferências internacionais e reconheceu a União Europeia como jurisdição adequada, enquanto seu marco legal de IA segue em tramitação, e a Coreia do Sul pôs em vigor uma lei de IA com alcance extraterritorial.^{11, 12, 13, 14, 15} Nesse contexto, a metodologia considera diferentes requisitos de soberania, sem pressupor uma abordagem única para todos os casos.

No sistema financeiro brasileiro, a regra já produz efeitos operacionais. A contratação de processamento e armazenamento no exterior depende de acordo de troca de informações entre o Banco Central e o supervisor do país de destino. Os ambientes de pagamentos instantâneos e de liquidação, por sua vez, passaram a exigir instância dedicada em nuvem. A norma não impõe residência de dados em território nacional, ao contrário do que o mercado costuma presumir.¹³

A jurisdição também alcança quem opera o ambiente. Provedores estão sujeitos às leis das jurisdições às quais se vinculam, e eventuais determinações legais ou administrativas podem afetar dados, operações ou acesso a informações. Os Estados Unidos oferecem um exemplo desse tipo de alcance jurisdicional, mas a mesma lógica pode ser observada em diferentes mercados. Provedores podem contestar ordens consideradas inadequadas, publicar relatórios de transparência e oferecer controles contratuais, criptográficos e de gestão de chaves que reduzem o risco residual. Esses mecanismos importam na comparação, mas não retiram o operador do alcance jurídico de sua jurisdição de origem.¹⁶ Os controles de exportação de chips seguem lógica semelhante, com restrições definidas por licenciamento caso a caso.¹⁷ A exposição, portanto, não decorre apenas do lugar onde o dado repousa, mas também das leis e restrições que alcançam o operador e a tecnologia.

Governos e empresas respondem com investimentos em capacidade. A maior iniciativa privada americana anunciou a intenção de investir US\$ 500 bilhões em quatro anos;¹⁸ a União Europeia mobiliza €200 bilhões e fixou a meta de ao menos triplicar sua capacidade de data centers;¹⁹ o Brasil estruturou seu plano nacional com até R\$ 23 bilhões até 2028.²⁰ A capacidade efetivamente disponível, contudo, dependerá da execução desses investimentos e de sua conversão em infraestrutura operacional.

A demanda sugere que diferentes níveis de soberania já carregam valor econômico mensurável. Os investimentos globais em nuvem soberana devem alcançar US\$ 80 bilhões em 2026,²¹ e projeta-se que 65% dos governos adotem requisitos formais de soberania até 2028.²² Na mesma direção, a receita associada a projetos de IA soberana do principal fornecedor de aceleradores superou US\$ 30 bilhões no último ano fiscal,²³ enquanto 77% das organizações já consideram o país de origem na seleção de fornecedores.⁶ No Brasil, a Pesquisa Febraban oferece uma evidência adicional dessa tendência: entre 2024 e 2025, o gasto dos bancos com nuvem privada cresceu 87%, frente a 23% na nuvem pública.⁴ No setor mais regulado da economia, as infraestruturas associadas a maior controle avançaram em ritmo quase quatro vezes superior ao das alternativas mais compartilhadas.

Esses movimentos indicam que a soberania influencia decisões de investimento, contratação e arquitetura. Para comparar as alternativas, o estudo utiliza um referencial comum que considera os controles oferecidos, as exposições remanescentes e o custo adicional associado. A seção seguinte apresenta esse referencial: a escada dos regimes de controle, organizada em três zonas.

A escada, em três zonas

A escala de soberania estabelece uma linguagem comum para comparar ofertas que o mercado descreve com nomes diferentes. Ela organiza sete regimes de controle, que vão da nuvem pública comum à infraestrutura própria, e permite perguntar, para cada família de aplicações, quais exposições permanecem, quais foram mitigadas e qual o custo adicional associado a cada decisão. Os degraus são analisados sob quatro dimensões: dados, operação, jurisdição e tecnologia. A progressão indica aumento predominante de controle, sem pressupor evolução uniforme em todas as dimensões. Para facilitar a leitura, os regimes são agrupados em três zonas; o percurso completo aparece na Figura 5.

Ainda não há uma taxonomia consolidada para arquiteturas de soberania. O mercado utiliza rótulos variados, e cada fornecedor organiza suas ofertas de modo próprio. A escada traduz esse vocabulário em regimes de arquitetura. Certificações, níveis de garantia e requisitos de compras públicas continuam relevantes, mas funcionam como evidências dentro de cada degrau.²⁴ Cada família define o conjunto mínimo de controles que exige nas quatro dimensões. Uma oferta só entra na comparação quando demonstra, em contrato e em operação, que cumpre esse mínimo. O rótulo ajuda a localizar a opção; os controles demonstrados determinam sua viabilidade.

Conveniência e escala (N0–N1). No N0, a organização consome um serviço público internacional e preserva escala, velocidade de acesso e baixo compromisso operacional. A infraestrutura compartilhada também sustenta ciclos frequentes de atualização, amplitude de serviços, segurança, resiliência e investimentos em conformidade que seriam difíceis de reproduzir isoladamente. Esses benefícios vêm acompanhados de maior exposição em dados, jurisdição, operação e dependência tecnológica. O N1 acrescenta residência local, controles contratuais e criptografia, sem eliminar a ligação jurídica, operacional e tecnológica com o fornecedor global. A zona atende experimentação, cargas de baixa sensibilidade e usos que valorizam o acesso contínuo à escala e à inovação do ecossistema. Para famílias que exigem maior autonomia ou controles específicos, os mesmos atributos podem ser insuficientes.

Perímetros progressivos (N2–N4). O N2 combina residência de dados e operação local em um perímetro regional dedicado, com entidade estabelecida no território e pessoal residente sujeito aos controles definidos. O N3 instala a infraestrutura do provedor no data center da organização. O perímetro físico aumenta, a escala compartilhada diminui e a economia passa a depender da utilização sustentada. O N4 é um regime conceitual no qual a tecnologia permanece licenciada por um fornecedor global, enquanto a operação e a relação jurídica passam a uma entidade nacional, com capital, pessoal e responsabilidade locais. Sua presença na escada não implica que exista, em cada mercado, uma oferta

contratável com catálogo, capacidade, resiliência, certificações e ritmo de atualização equivalentes aos dos provedores globais. A ordem expressa a progressão predominante do controle jurídico e operacional. Isso não torna o N4 superior ao N3 em todas as dimensões: o N3 preserva maior domínio do perímetro físico. A zona atende famílias que precisam de mais controle sem assumir integralmente uma infraestrutura própria; a dependência tecnológica permanece nos três regimes.

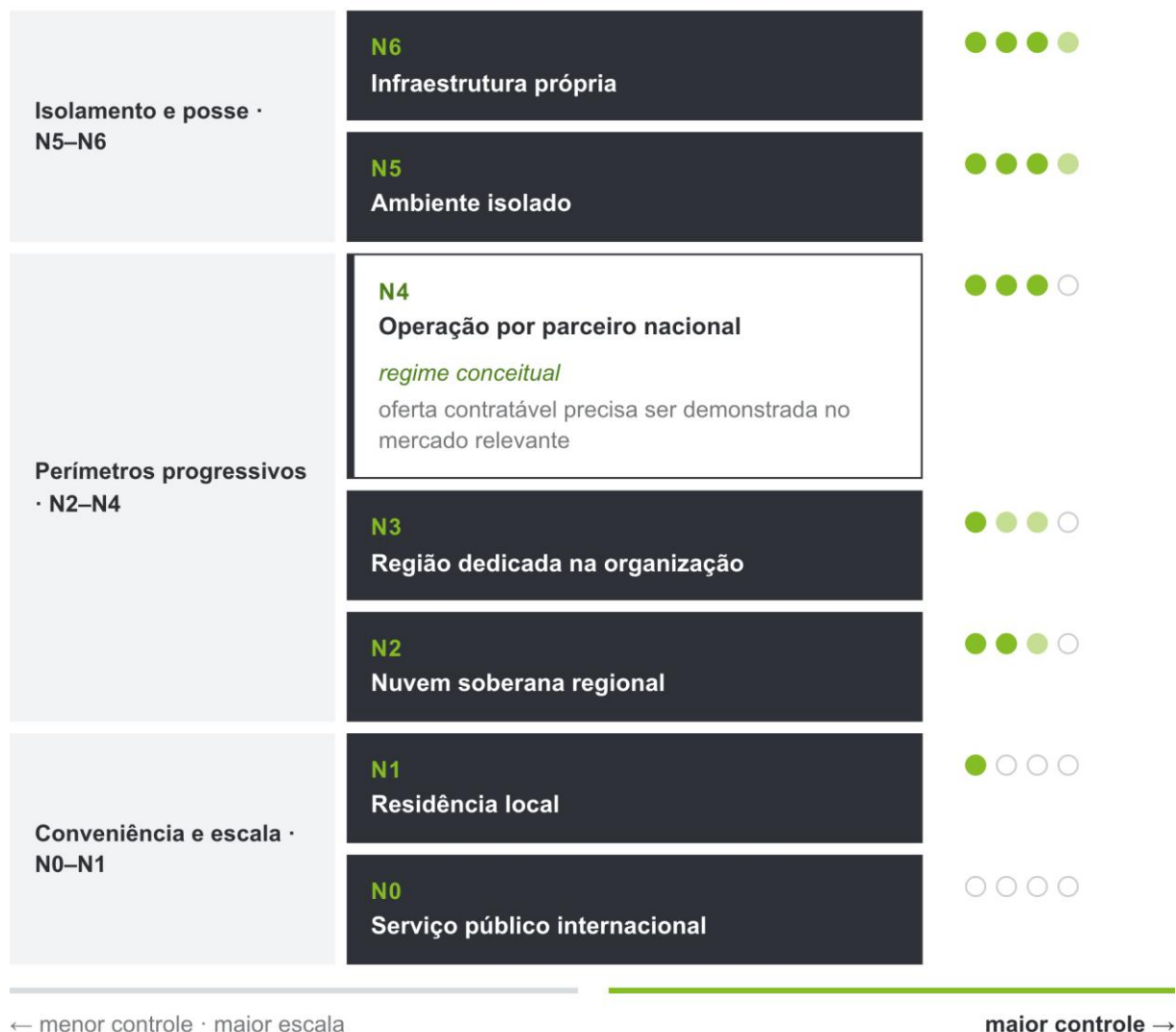
Isolamento e posse (N5–N6). O N5 prioriza isolamento: a infraestrutura fica desconectada da internet, as atualizações são auditadas e o acesso depende de pessoal credenciado. É um regime voltado a defesa, inteligência e cargas classificadas. O N6 prioriza posse e controle institucional: computação, operação e perímetro pertencem à organização ou a uma estrutura criada para esse fim, cativa, consorciada ou comercial. Um ambiente próprio também pode incorporar isolamento, mas essa não é a característica que define o degrau. O N6 representa o maior grau de controle operacional e institucional da escada, sem equivaler a autonomia tecnológica plena. Aceleradores, software de baixo nível e partes da cadeia de suprimento continuam globais.

FIGURA 5

Três zonas, sete degraus

● controla ● parcial ○ não controla

ordem dos pontos: dados · operação · jurisdição · tecnologia



Os degraus ordenam controle, não custo. A progressão indica aumento predominante de controle, sem pressupor avanço simultâneo nas quatro dimensões.

Os degraus representam regimes de contratação dentro de um ecossistema tecnológico compartilhado. Os grandes provedores globais de nuvem dispõem de tecnologia para grande parte da escada, da residência local ao ambiente isolado. O licenciamento para uma operação nacional é uma possibilidade arquitetural adicional, mas sua disponibilidade e suas capacidades precisam ser demonstradas no mercado relevante. A escolha, em última instância, estabelece quem controla cada camada, sob qual jurisdição e com quais mecanismos de supervisão e prestação de contas.

A nuvem soberana do governo federal brasileiro, disponível a mais de 250 órgãos desde junho de 2025 e com mais de R\$ 1 bilhão em investimentos previstos, utiliza tecnologia de fornecedores globais instalada em data centers de empresas estatais. O governo a descreve como uma combinação de soberania de dados e soberania operacional.²⁵ No referencial da escada, o caso ilustra uma configuração intermediária que amplia controles específicos sem eliminar todas as dependências tecnológicas ou jurisdicionais.

O aumento do controle não segue uma trajetória linear de custos. Regiões dedicadas, ambientes instalados no perímetro da organização e capacidade própria combinam custos fixos, escala e responsabilidades de maneiras diferentes. No futuro, consórcios ou operadores nacionais poderão deslocar essa economia ao agregar demanda, mas sua viabilidade dependerá de oferta efetivamente contratável, capacidade disponível, resiliência, amplitude tecnológica, poder de investimento e condições de saída. Até que esses requisitos sejam demonstrados, tais alternativas devem entrar como cenários potenciais, não como equivalentes disponíveis às plataformas globais. A comparação considera apenas opções viáveis no horizonte da decisão e não presume que um degrau mais alto será sempre mais caro.

A escada organiza os regimes de contratação e controle e permite avaliar cada oferta pelos controles que oferece, pelo custo adicional que incorpora e pelas exposições que permanecem. O próximo passo é integrar essa régua às dimensões de custo e valor.

Da equação à arquitetura

Com custo, valor e soberania definidos, a decisão pode ser reunida em uma única regra por família de aplicações. O perfil de exposição estabelece o grau mínimo de soberania e elimina as configurações que não entregam os controles exigidos. Entre as alternativas compatíveis, a escolha recai sobre a que preserva mais valor líquido no horizonte considerado. A ordem importa: ela evita que uma alternativa barata, mas incompatível, entre na comparação, e que controles sem efeito relevante sejam tratados como valor. O quadro a seguir apresenta a equação e a ordem em que ela se aplica.

QUADRO A equação

Antes da conta, o filtro: entram na comparação apenas as configurações que cumprem o grau mínimo de soberania da família.

	Valor líquido	o que a configuração preserva ao fim do horizonte
=	Benefícios	valor: produtividade, custo evitado, receita incremental e valor de opção
−	Custo total de propriedade	custo: com o prêmio de soberania já incorporado
−	Risco residual	soberania: a exposição que permanece após os controles
−	Mudança e saída	reversibilidade: migração, encerramento e rota de saída

Entre as configurações compatíveis, vence a que preserva mais valor líquido no horizonte considerado.

A CONTA SE APLICA EM TRÊS MOVIMENTOS

1 Compatibilidade RESTRINGE O CONJUNTO	O perfil de exposição da família define os controles mínimos em dados, operação, jurisdição e tecnologia, e elimina as configurações que não os entregam.
2 Custo e exposição AVALIA AS ELEGÍVEIS	Cada configuração elegível é avaliada pelo custo total de propriedade, pelo risco residual e pelos custos de mudança e saída.
3 Benefícios COMPARA COM O VALOR	O resultado é confrontado com os benefícios esperados da família.

Duas regras evitam dupla contagem

- a redução de perdas esperadas aparece como diminuição do risco residual, nunca como benefício;
- o valor de opção é reconhecido na família ou no portfólio, nunca nos dois níveis.

A soberania entra duas vezes na conta: como filtro de elegibilidade, antes da conta, e como prêmio dentro do custo total.

O prêmio de soberania, já incorporado ao custo total, permite comparar o custo adicional dos controles à redução do risco residual da família. Seu efeito sobre a faixa de equilíbrio e sua variação entre os degraus foram tratados nas seções anteriores.

No método proposto, a migração para níveis mais elevados de controle é analisada a partir da relação entre o custo adicional, a redução das exposições identificadas, a preservação de alternativas futuras e os benefícios esperados para a família de aplicações.

Nas famílias sensíveis, esse prêmio pode ser menor do que a perda esperada associada a riscos regulatórios, operacionais ou reputacionais. Nas famílias de baixa exposição, pode consumir valor sem reduzir risco material. Ou seja, a análise depende das premissas da família, não de uma preferência geral por mais ou menos soberania.

O efeito no portfólio decorre da combinação dessas escolhas, ponderadas pelo volume e pelo valor de cada família. Assim, a organização concentra salvaguardas onde a exposição as justifica e preserva escala e experimentação onde controles adicionais não se pagam.

A equação transforma o diagnóstico em decisão estratégica. Aplicá-la ao portfólio inteiro é o trabalho do método, organizado em cinco passos.

O método: cinco passos

O método organiza um inventário fragmentado em uma política de arquitetura por família de aplicações, com mecanismos de execução e revisão. A aplicação envolve responsabilidades coordenadas: tecnologia avalia os arranjos viáveis e suas implicações operacionais; finanças compara gasto variável, compromisso contratual e capital imobilizado; risco define a exposição aceitável, o risco residual e o grau de soberania requerido; e negócio responde pela hipótese e pela captura de valor. À alta administração cabe deliberar sobre os conflitos entre essas perspectivas antes da consolidação da arquitetura.

Na prática, o trabalho percorre cinco passos. Cada um produz uma entrega verificável, e o último devolve telemetria ao primeiro, formando um ciclo contínuo de decisão. A Figura 6 resume o percurso.

FIGURA 6

O método em cinco passos

Cada passo produz uma entrega verificável; o quinto devolve telemetria ao primeiro e o ciclo recomeça.

1 Medir	Linha de base de custo, consumo, benefícios, exposições, controles e rotas de saída.	ENTREGA linha de base de valor, custo, exposição e controle
2 Classificar	Famílias que podem compartilhar uma decisão: exposição, volume, experimentação, requisitos e autonomia.	ENTREGA mapa de famílias, com responsáveis e horizonte
3 Diagnosticar	Configurações incompatíveis eliminadas; as elegíveis comparadas por custo, risco residual e benefícios.	ENTREGA mapa de desalinhamentos e prioridades
4 Definir desenho e política	Por família: zona e degrau, arranjo, operador, limites e rota de saída. Depois, consolidação do portfólio.	ENTREGA política do portfólio, com limites e gatilhos
5 Implantar	Regras aplicadas em cada chamada pela camada de abstração; contratos, controles e migração.	ENTREGA política implantada e telemetria contínua

ASSESSMENT · PASSOS 1–4

IMPLANTAÇÃO · PASSO 5

telemetria de custo, valor, desempenho e exposição realimenta a linha de base

Os gatilhos de revisão — volume, preço, regulação, exposição, desempenho dos modelos, disponibilidade de capacidade e maturidade da medição de valor — reabrem o ciclo a qualquer momento.

1 - Medir. O primeiro passo consolida a base fragmentada entre tecnologia, contratos, risco e áreas usuárias. A linha de base registra custo por token e por tarefa, volume, variância, modelos e arranjos em uso, benefícios observados e alavancas de otimização já capturadas. Além de identificar as exposições atuais, os controles existentes, o risco residual, os compromissos contratuais e as rotas de saída disponíveis. Em aplicações agênticas, mede decisões tomadas em tempo de execução, permissões, memória, ferramentas e consumo por ação. Sem essa análise, a organização pode migrar cedo demais, construir capacidade sem volume suficiente ou pagar um prêmio de soberania sem demonstrar o risco que ele reduz. Entrega: linha de base de valor, custo, exposição e controle.

2 - Classificar. As aplicações em produção, piloto e avaliação, inclusive a IA embarcada em softwares de terceiros, são agrupadas em famílias para as quais uma decisão arquitetural comum é coerente. O agrupamento considera perfil de exposição, volume projetado,

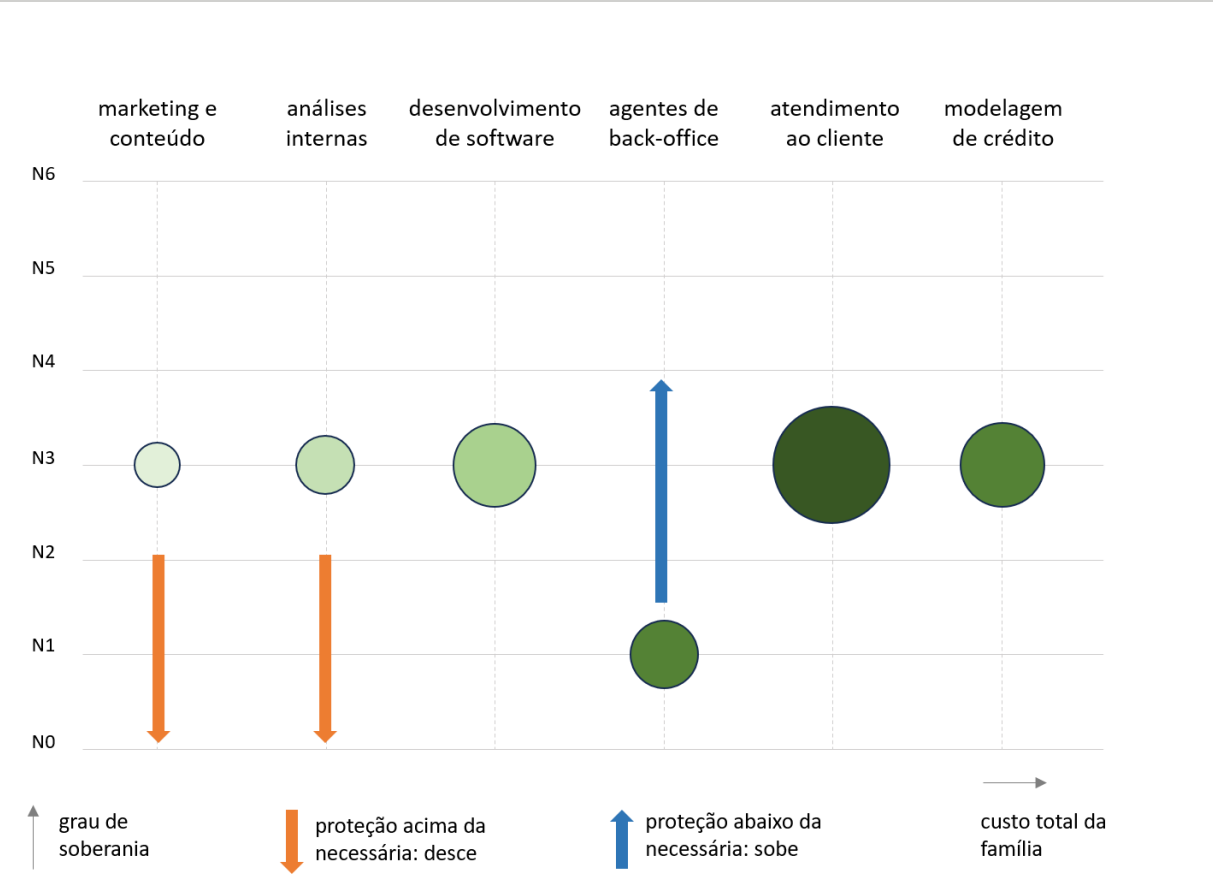
necessidade de experimentação e requisitos técnicos, sem ocultar diferenças relevantes. Quando treinamento, ajuste ou avaliação exigem condições diferentes da inferência, a família permanece como unidade de decisão, mas sua configuração é definida por etapa do ciclo de vida do modelo. A autonomia dos agentes entra como atributo da família: quanto maior a capacidade de acessar sistemas, acumular contexto e executar ações, maiores tendem a ser o consumo projetado e o grau mínimo de soberania. Cada família recebe responsáveis de negócio e tecnologia, horizonte de análise e cadência de revisão. Entrega: mapa de famílias, com atributos, responsáveis e horizonte definidos.

3 - Diagnosticar. Para cada família, o perfil de exposição define o grau mínimo de soberania e elimina as configurações incompatíveis. As alternativas elegíveis são comparadas por custo total, risco residual, custos de mudança e benefícios esperados. O diagnóstico mostra onde há controle insuficiente e/ou proteção excessiva, quanto custa corrigir cada desvio e que valor essa correção preserva. A prioridade combina materialidade da exposição, valor líquido e urgência, sem reduzir a equação ao plano custo × soberania usado como representação visual na Figura 7. Entrega: mapa de desalinhamentos e prioridades de correção.

FIGURA 7

Diagnóstico do portfólio no plano custo × soberania

Caso ilustrativo · cor = benefícios esperados · tamanho = volume de tokens · setas = movimento proposto



Representação visual do diagnóstico; a decisão considera também benefícios esperados, risco residual e custos de mudança e saída.

Fonte: análise Deloitte. Composto ilustrativo, calibrado por benchmarks públicos e pela experiência da prática.

4 - Definir desenho e política. O passo começa no desenho de cada família. Entre as configurações compatíveis, a organização escolhe a zona e o degrau, o arranjo de consumo, o operador, os limites de ação e gasto e a rota de saída. Em famílias agênticas, a política inclui permissões, memória, ferramentas autorizadas, supervisão humana e trilha de auditoria.

As escolhas são, depois, consolidadas no portfólio, porque plataforma, segurança, rede, observabilidade, contratos, energia e equipes são compartilhados. A consolidação recalcula custos fixos, concentração, ociosidade e capacidade operacional, e também incorpora restrições de hardware, memória, energia e disponibilidade, que afetam preço, prazo e escala. Por fim, ficam explícitos os gatilhos de revisão: volume, preço, regulação, exposição geopolítica, desempenho dos modelos, disponibilidade de capacidade e maturidade da medição de valor. Entrega: política do portfólio, com configurações autorizadas, limites e gatilhos de revisão.

5 - Implantar. A política se converte em mecanismos técnicos, contratuais e organizacionais. A camada de abstração, apresentada adiante, aplica em cada chamada as regras de acesso, consumo e registro e restringe o encaminhamento aos destinos autorizados. Contratos, controles de segurança, responsabilidades operacionais, planos de migração e rotas de

saída completam a implantação. A telemetria de custo, desempenho, valor e exposição retorna à linha de base, e o ciclo recomeça a cada período orçamentário ou sempre que um gatilho é acionado. Entrega: política implantada, controles operantes e telemetria contínua.

O método foi concebido para permitir decisões distintas por família mantendo uma lógica comum de portfólio: mede, classifica, diagnostica, define, implanta e revisa.

A seção seguinte aplica o ciclo completo a um portfólio concreto e mostra como decisões distintas por família se convertem em uma política comum de arquitetura.

O método em campo

O método se concretiza quando encontra um portfólio formado em momentos diferentes, com aplicações em produção, pilotos em expansão e capacidades ainda em avaliação. O caso a seguir é um composto ilustrativo de uma instituição financeira de médio porte. Foi calibrado pela série setorial da Febraban, pelas faixas de prêmio e de equilíbrio discutidas neste estudo e pela experiência da prática. Os números servem para demonstrar o percurso; em uma aplicação real, premissas e resultados precisam ser recalculados com os dados da organização.

1 - Medir. A instituição identifica 34 aplicações de inteligência artificial e outras sete embarcadas em softwares de terceiros. A linha de base consolida volume, custo por tarefa, modelos, contratos, benefícios observados, controles e rotas de saída. O portfólio havia crescido sem arquitetura comum e, depois de um parecer de risco sobre a aplicação de crédito, todo o conjunto foi levado a um perímetro dedicado com consumo por catálogo. Para 24 meses, o custo projetado alcançava R\$ 41 milhões. Dois pilotos agênticos respondiam pela maior parte da variância, enquanto famílias de baixa exposição pagavam o mesmo prêmio de soberania.

A medição também separa o que já era observado do que ainda dependia de hipótese. Custos, volumes, contratos e controles eram dados disponíveis. Benefícios de produtividade estavam instrumentados apenas em atendimento e desenvolvimento; nas demais famílias, entraram como hipóteses documentadas, com responsáveis e prazo para conversão em métrica. A redução de perdas permaneceu no risco residual, sem ser somada aos benefícios.

2 - Classificar. As 41 aplicações são reunidas em seis famílias: marketing e conteúdo; desenvolvimento de software; análises internas sem dados pessoais; atendimento ao cliente; agentes de backoffice; e crédito. Cada família recebe responsáveis de negócio e tecnologia, horizonte de 24 meses, volume projetado, perfil de exposição e grau de autonomia. A classificação mostra que o padrão único escondia necessidades muito diferentes de experimentação, escala e controle.

3 - Diagnosticar. O perfil de exposição elimina as configurações incompatíveis e revela três tipos de desalinhamento. Marketing e conteúdo, desenvolvimento de software e análises internas sem dados pessoais tinham proteção acima da necessária: pagavam o prêmio integral para reduzir uma exposição residual pouco relevante. Os agentes de backoffice apresentavam o problema oposto: consumo decidido em tempo de execução, permissões amplas e acesso a sistemas produziam exposição superior à suportada pelos controles contratados. Crédito ocupava o grau adequado de soberania, mas em arranjo econômico desfavorável, dedicado a uma única organização.

FIGURA 8

Do inventário à configuração das famílias

34 aplicações e sete softwares com IA embarcada · seis famílias · R\$ 41 milhões → cerca de R\$ 33 milhões em 24 meses

2 · CLASSIFICAR	3 · DIAGNOSTICAR	4 · DEFINIR
41 cargas → seis famílias de aplicações	três tipos de desalinhamento	degrau · arranjo — antes → depois
41 cargas inventariadas		
crédito	Proteção acima da necessária famílias de baixa exposição pagando o prêmio integral	crédito N3 · A2 → oferta soberana compartilhada <i>mantém o grau, muda o arranjo</i>
agentes de back-office	Controles insuficientes agentes com consumo, permissões e acesso acima do contratado	agentes de back-office N1 · A1 → N2 · A2 <i>sobe e recebe política própria</i>
atendimento ao cliente		atendimento ao cliente N3 · A2 → N2 · A2 <i>desce um degrau, mantém o catálogo</i>
desenvolvimento de software	Arranjo econômico desfavorável crédito no grau adequado, em configuração dedicada individual	desenvolvimento e análises N3 · A2 → N0–N1 <i>sem dados pessoais</i>
análises internas sem dados pessoais		marketing e conteúdo N3 · A2 → N0 · A1 <i>prêmio sem alvo</i>
marketing e conteúdo		

DE ONDE VÊM OS R\$ 8 MILHÕES

R\$ 2 mi	R\$ 4 mi	R\$ 2 mi
otimização do consumo	prêmio retirado de três famílias	mudança de arranjo em crédito

Fonte: análise Deloitte. Composto ilustrativo de instituição financeira de médio porte; premissas declaradas no corpo.

4 - Definir desenho e política. Cada desalinhamento recebe uma resposta própria. As três famílias de baixa exposição retornam ao serviço convencional, com limites de uso e rota de saída. Atendimento ao cliente desce um degrau, mas preserva o consumo por catálogo e os controles necessários sobre dados pessoais. Os agentes de backoffice sobem de degrau e recebem política específica para permissões, memória, ferramentas, supervisão humana, gasto por ação e trilha de auditoria. Crédito mantém o grau de soberania exigido, mas

substitui a configuração dedicada individual por uma oferta soberana compartilhada já contratável, com controles demonstrados e menor compromisso fixo.

A consolidação recalcula o portfólio com plataforma, segurança, contratos, concentração e capacidade operacional compartilhados. Dos R\$ 8 milhões de redução projetada, cerca de R\$ 2 milhões vêm da otimização do consumo efetivo, com roteamento, cache e redução de contexto. Aproximadamente R\$ 4 milhões decorrem da retirada do prêmio de soberania das três famílias de baixa exposição e do ajuste de atendimento. Os R\$ 2 milhões restantes resultam da mudança de arranjo em crédito, já descontado o custo adicional de governança dos agentes. O custo de 24 meses recua de R\$ 41 milhões para cerca de R\$ 33 milhões, uma redução próxima de 19%.

A política consolidada define configurações autorizadas, limites e três gatilhos de revisão. O primeiro é o volume agregado das famílias de alto valor, que pode aproximar a capacidade dedicada da faixa de equilíbrio. O segundo é uma mudança na regulação de crédito que altere o grau mínimo de soberania. O terceiro é a variação do preço do serviço soberano em relação às alternativas elegíveis. O orçamento liberado financia a instrumentação de valor e os controles dos agentes, em vez de permanecer imobilizado em proteção indiferenciada.

5 - Implantar. A instituição converte a política em mecanismos técnicos, contratuais e organizacionais. A camada de abstração aplica por família as regras de acesso, consumo e registro e restringe o encaminhamento aos destinos autorizados. Os contratos incorporam compromissos de portabilidade e saída. As áreas responsáveis validam permissões, controles e planos de migração antes da mudança de cada carga. A telemetria passa a registrar custo por tarefa, qualidade, consumo por ação dos agentes, benefícios reconhecidos e eventos de exposição.

No primeiro ciclo de revisão, os dados de operação retornam à linha de base. Se o consumo agêntico superar o limite, a política restringe ferramentas ou exige aprovação humana. Se o ganho de atendimento não se confirmar, sua hipótese de valor é rebaixada. Se o volume de crédito atingir a faixa de equilíbrio, a organização reabre a comparação entre ofertas soberanas contratáveis, fábrica como serviço e capacidade própria. Operadores nacionais ou consórcios só entram quando demonstrarem capacidade, controles e condições comerciais compatíveis com a família. O caso, assim, percorre o ciclo completo: mede, classifica, diagnostica, define, implanta — e revisa.

No cenário ilustrativo, a redução de 19% decorre das premissas e dos ajustes aplicados às famílias analisadas, e não constitui expectativa de resultado para outras organizações. O portfólio composto reunia famílias com controles acima dos requisitos definidos, agentes com controles abaixo do perfil de exposição e uma aplicação crítica com grau de soberania compatível, mas estrutura econômica distinta das demais alternativas elegíveis. A aplicação por família demonstra como o método reorganiza salvaguardas, experimentação e controles

observáveis. Resultados, custos e configurações devem ser recalculados em aplicações reais com base nos dados, contratos e requisitos específicos de cada organização.

Aspectos críticos na decisão. O método busca aumentar a consistência da decisão, mas não elimina resultados desfavoráveis diante de mudanças de contexto ou de premissas. Esses desvios podem ocorrer quando custo, exposição e valor de opção deixam de ser reavaliados de forma integrada. Três situações concentram os casos recorrentes observados no referencial.

Adotar controles adicionais sem redução proporcional da exposição. Essa situação pode ocorrer de duas formas. Na primeira, a classificação da oferta substitui a demonstração dos controles: residência física é interpretada como imunidade jurídica, operação local como autonomia plena ou um degrau intermediário como controle integral. Na segunda, os requisitos da aplicação mais sensível são estendidos a todo o portfólio. Em ambos os casos, o custo pode aumentar sem redução equivalente do risco residual. O método prevê evidências contratuais e operacionais para cada dimensão de controle e calibração do grau de soberania por família.

Reduzir uma dependência e criar outra. A mudança de arquitetura pode ampliar o controle em uma dimensão e restringir alternativas em outra. Uma fábrica própria pode reduzir a exposição ao provedor de nuvem e, simultaneamente, ampliar a vinculação à arquitetura de aceleração, ao software proprietário, à cadeia de semicondutores e a capacidades profissionais especializadas. Em 2024, um fabricante concentrava cerca de 92% do mercado de GPUs para data centers; no último relatório anual do fornecedor de maior participação, dois clientes responderam por 36% da receita.^{7, 27} Situação semelhante pode ocorrer quando contratos longos, dados pouco portáteis ou uma regra transitória são incorporados ao desenho como condições permanentes. Esses fatores podem ampliar a estabilidade no curto prazo e reduzir a flexibilidade de mudança. O método considera alternativas de compra e operação, rotas de saída testadas, compromissos escalonados e gatilhos de revisão para preço, volume, regulação e disponibilidade tecnológica.

Limitar a evolução para estabilizar o presente. A busca por previsibilidade pode induzir todas as famílias para configurações maduras e tratar acesso à fronteira, modelos locais e experimentação como escolhas incompatíveis. O efeito precisa ser medido na tarefa. Em algumas famílias, a distância para os modelos mais avançados determina qualidade e valor. Em outras, modelos menores, locais ou especializados entregam resultado suficiente com custo menor e mais controle. Um relatório técnico de um desenvolvedor brasileiro, por exemplo, mostra um modelo nacional atrás nos exames gerais e à frente no domínio jurídico, com preço de entrada aproximadamente duas vezes menor.²⁸ O erro está em congelar uma dessas respostas como padrão. Com o tempo, a proteção excessiva aparece como atraso de aprendizado, perda de opcionalidade e menor capacidade de capturar o próximo ciclo de valor. A prevenção está em preservar famílias experimentais com limites

próprios de gasto e exposição, medir desempenho na tarefa e revisar a configuração conforme modelos e capacidades evoluem.

As três situações apontam para uma disciplina comum: o grau de soberania é avaliado pela demonstração dos controles, pela preservação de alternativas de mudança e pela manutenção de espaço para aprendizado. O método trata a dimensão institucional dessa disciplina por meio de responsáveis, entregas e gatilhos de revisão. A dimensão operacional requer um ponto comum para aplicar e medir, em cada chamada, as regras definidas para cada família.

A camada de abstração: onde a política vira operação

A implantação do método combina contratos, responsabilidades, controles e um componente técnico comum. A camada de abstração ocupa esse último papel: fica entre as aplicações e os provedores de inteligência, identifica a família associada a cada solicitação e aplica as regras de acesso, consumo e registro antes de encaminhá-la a um destino autorizado. Um *AI gateway* resolve parte dessa função. A Figura 9 mostra sua posição entre o consumo e os provedores.

FIGURA 9

A camada de abstração

Entre as aplicações e os provedores de inteligência: identifica a família de cada solicitação e aplica as regras antes de encaminhá-la.



A camada aplica as regras definidas no quarto passo do método. O desenho detalhado, a escolha dos componentes e a implantação são tratados no segundo estudo desta série.

A camada reúne quatro capacidades. A primeira é **aplicar a política**: modelos, provedores, regiões, dados, permissões, ferramentas e limites de gasto autorizados para cada família tornam-se regras executáveis. A segunda é **definir o perímetro de consumo**: a política estabelece quais provedores, modelos, regiões e configurações cada família pode utilizar, além das condições de contingência e substituição. A aplicação, a orquestração ou o harness escolhe, dentro desse conjunto autorizado, a alternativa adequada à tarefa. A terceira é **medir**: consumo, custo por tarefa, qualidade, latência, decisões dos agentes e eventos de exposição passam a ser observados em um ponto comum. A quarta é **registrar e auditar**: os destinos utilizados, os controles aplicados e as exceções formam uma trilha verificável.

A camada permite combinar capacidades distintas entre provedores dentro de uma política comum, preservando governança, capacidade de monitoramento e alternativas de mudança.

Essas capacidades podem contribuir para aumentar a resiliência porque concentram a resposta operacional às mudanças de preço, modelo, disponibilidade e regra. Parte das adaptações pode ocorrer pela atualização do perímetro autorizado, das condições de contingência e dos parâmetros da política. Outras continuam a exigir avaliação, ajuste de aplicação, migração de dados, revisão contratual ou mudança de controles. A camada não elimina esse trabalho; reduz sua dispersão e torna explícitas as dependências que precisam ser tratadas.

A camada também precisa de fronteiras claras com componentes vizinhos. A orquestração coordena fluxos e decide como uma tarefa será executada. O harness envolve o modelo durante a execução e fornece contexto, memória, ferramentas e permissões. A camada de abstração define quais recursos cada família pode consumir e sob quais condições; a aplicação, a orquestração ou o harness seleciona, dentro desse perímetro, a alternativa adequada à tarefa. Em aplicações agênticas, as interfaces se encontram: a camada aplica tetos de gasto, destinos autorizados e requisitos de registro; a orquestração decide a sequência de ações permitida pela política.

Seu desenho deve seguir os mesmos princípios do método: critérios comuns entre provedores, portabilidade, capacidade de monitoramento, segurança, disponibilidade e separação entre política e implementação. A camada pode ser implantada no ambiente da organização, em serviço gerenciado ou em uma combinação dos dois, desde que preserve o controle das regras, dos dados de telemetria e das rotas de saída. A fábrica de IA aparece para a camada como um provedor de inteligência entre outros, sujeito às condições definidas para cada família. O desenho detalhado, a escolha dos componentes e a implantação serão tratados no segundo estudo desta série.

Uma base comum não implica política idêntica em todas as geografias. A camada aplica a mesma lógica de decisão, mas parametriza provedores, regiões, controles e evidências conforme jurisdição, regulação, energia, tributos e disponibilidade local. Esses coeficientes alteram a comparação entre consumir, contratar, consorciar e construir. A seção seguinte leva a equação ao contexto brasileiro.

A equação no Brasil

No Brasil, a equação recebe coeficientes próprios: energia, infraestrutura digital, tributação, regulação e demanda setorial alteram o custo das configurações, o grau de soberania viável e a escala necessária para sustentar capacidade local. O país parte de ativos relevantes: o Plano Brasileiro de Inteligência Artificial prevê R\$ 23 bilhões entre 2024 e 2028,²⁰ o Brasil concentra cerca de metade da capacidade de data centers em operação na América Latina²⁹ e sua matriz elétrica tem elevada participação de fontes renováveis. Esses ativos ampliam o conjunto de opções, mas não tornam nenhuma delas automaticamente superior.

Energia altera o custo, desde que esteja disponível no lugar e no prazo necessários. A participação de fontes renováveis pode reduzir a exposição a carbono e aumentar a atratividade de cargas intensivas de inferência para operadores e clientes com metas ambientais. A vantagem se materializa apenas quando a energia pode ser contratada, conectada e entregue com confiabilidade. Capacidade de geração, acesso à rede, transmissão, prazo de conexão e localização da demanda precisam entrar no custo total. Sem expansão e coordenação da infraestrutura elétrica, os benefícios de uma matriz limpa tendem a permanecer limitados na prática.

A infraestrutura instalada reduz a barreira de entrada, mas capacidade regional não equivale à capacidade elegível. A presença de data centers, conectividade e operadores cria uma base econômica para consumir, contratar ou consorciar capacidade no país. Ainda assim, a decisão por família depende de requisitos concretos: aceleradores disponíveis, densidade energética, latência, certificações, operador, jurisdição, portabilidade e tempo de implantação. Para uma organização, a vantagem relevante é a capacidade que atende a esses requisitos, não o estoque agregado do mercado.

A tributação move a faixa de equilíbrio e precisa ser tratada por cenários. O regime especial para data centers voltou ao debate legislativo pelo Projeto de Lei nº 278/2026, aprovado na Câmara dos Deputados e ainda em tramitação no Senado. Enquanto não houver regra vigente e regulamentada, tributos sobre equipamentos e serviços críticos permanecem no custo total de propriedade.³⁰ A comparação entre construir, contratar e consumir deve testar pelo menos dois cenários, com e sem incentivo, e condicionar o compromisso de capital à confirmação das premissas.

Regulação e operação local alteram o grau de soberania, sem impor uma resposta única ao portfólio. Em setores regulados, determinadas famílias já carregam requisitos específicos de processamento, auditoria, isolamento e supervisão. Essas regras delimitam o conjunto de configurações elegíveis para essas cargas. Para as demais, a definição do modelo operacional deve considerar requisitos de risco, resiliência, valor de negócio e eficiência econômica.

A demanda regulada pode formar escala e dar visibilidade ao valor. O setor bancário oferece o sinal mais concreto. Seu orçamento de tecnologia foi de R\$ 46,8 bilhões em 2025 e deve alcançar R\$ 50,4 bilhões em 2026. Entre 2024 e 2025, o gasto com inteligência artificial avançou de R\$ 596 milhões para R\$ 826 milhões, enquanto a migração para nuvem passou de R\$ 3,0 bilhões para R\$ 3,9 bilhões.⁴ Embora insuficiente para justificar sozinho uma fábrica ou um regime soberano, o volume cria as condições necessárias para consolidar demanda, otimizar a contratação de capacidade e gerar ganhos recorrentes. A eficiência surge quando famílias de cargas compatíveis capturam economias de escala sem a necessidade de controles uniformes.

A consequência muda conforme a posição de cada organização. Para um operador de infraestrutura, energia, território e demanda podem sustentar investimento em capacidade. Para uma empresa usuária, esses mesmos ativos só geram vantagem se reduzirem o custo de uma configuração elegível, ampliarem o controle necessário ou preservarem alternativas de fornecimento. Para o país, hospedar máquinas aumenta capacidade local; para a organização, soberania continua dependendo de dados, operação, jurisdição e tecnologia.

A decisão executiva no Brasil

A equação brasileira não aponta para uma arquitetura nacional única, mas exige que cada família incorpore coeficientes locais ao mesmo cálculo de valor, custo e soberania. Energia disponível e incentivo confirmado reduzem custo. Regulação e jurisdição elevam o piso de controle. Infraestrutura elegível e demanda agregável ampliam as alternativas de escala. Incertezas de conexão, tributação e execução entram como risco e gatilho de revisão.

Para os executivos, três perguntas resumem a decisão: Se uma camada crítica ficar indisponível ou restrita em 90 dias, qual rota de continuidade pode ser acionada? Que parcela do gasto está em consumo variável, compromisso contratual ou capital imobilizado, e que exposição cada parcela reduz? Quem aceita o risco residual e quem responde pela captura do valor? As respostas conectam o contexto brasileiro ao método e preparam a passagem da análise para a ação.

Da análise à decisão

A aplicação pode começar em escala controlada. Uma amostra do portfólio deve reunir aplicações em produção, pilotos que ganharam tração, cargas sensíveis, agentes em expansão, contratos herdados e iniciativas que ainda precisam de espaço para aprender. Esse recorte é suficiente para estabelecer a linha de base, formar famílias, identificar desalinhamentos e testar a equação com dados reais.

A primeira aplicação não precisa definir a arquitetura definitiva. Precisa produzir decisões verificáveis: onde ampliar controle, onde recuperar escala, onde preservar experimentação e quais gatilhos justificam uma mudança de rota. O resultado é uma política inicial por família, acompanhada de responsáveis, limites e telemetria para o próximo ciclo.

A decisão entre diferentes fornecedores e arranjos torna-se mais consistente quando orientada por critérios comuns e por uma visão integrada de portfólio, capaz de converter considerações tecnológicas e operacionais em prioridades, caminhos arquiteturais, critérios de investimento e gatilhos de revisão. Os próximos estudos desta série aprofundam a execução: o segundo trata do desenho e da implantação da camada de abstração; o terceiro examina fábricas de IA próprias, consorciadas e como serviço, além da agenda de energia associada a essas escolhas.

Sob o referencial deste estudo, a análise de arquitetura não parte da premissa de que a alternativa mais sofisticada, de menor custo ou de maior controle seja adequada em termos absolutos. A avaliação considera o ajuste entre valor, custo e soberania para cada família de aplicações, com revisão periódica das premissas segundo critérios de capital, risco e estratégia.

Autor

Paulo André Domingos

Partner, Artificial Intelligence & Data

Deloitte Brasil

paulodomingos@deloitte.com

Agradecimentos a João Zaiden pelas contribuições na redação deste estudo, e à Renata Muramoto e Fernando Fujihara pelo apoio ao longo do processo.

Explore mais

A economia da IA: a virada para a tokenomics. O estudo global da Deloitte que estrutura o lado do custo desta equação: padrões de compra, TCO por token e pontos de equilíbrio.

Pesquisa Febraban de Tecnologia Bancária 2026. Conduzida com a Deloitte: orçamentos, prioridades e ganhos medidos de inteligência artificial no setor que mais investe em tecnologia no Brasil.

A camada de abstração. O segundo estudo desta série, em preparação: da política à operação.

As fábricas de IA. O terceiro estudo desta série, em preparação: próprias, consorciadas e como serviço.

Nota de divulgação

A Deloitte presta serviços profissionais relacionados a alguns dos arranjos e instrumentos abordados neste estudo, conforme informado na abertura. Para assegurar transparência e consistência, a metodologia aplica critérios comuns às diferentes alternativas analisadas, incluindo tecnologias de provedores globais de nuvem, operadores especializados, infraestrutura própria e soluções nacionais ou consorciadas, quando efetivamente disponíveis e capazes de demonstrar requisitos compatíveis com a família de aplicações avaliada.

As análises, comparações e conclusões refletem a aplicação da metodologia, das premissas e das fontes indicadas no estudo. Não representam avaliação definitiva, recomendação individualizada, preferência por fornecedor, tecnologia ou modelo de contratação, nem garantia de desempenho, viabilidade, economia ou resultado. Cada organização deve realizar sua própria avaliação, considerando objetivos, circunstâncias, riscos e requisitos específicos.

Nota editorial: As informações regulatórias, legislativas e de mercado foram verificadas até 20 de agosto de 2026 e estão sujeitas a alterações. Projeções, estimativas e declarações prospectivas citadas são de responsabilidade das instituições referidas e não constituem garantia de resultados futuros. Este material contém informações de caráter geral e não constitui aconselhamento jurídico, tributário, financeiro, contábil, regulatório, tecnológico ou de investimento, nem substitui a consulta a profissionais qualificados.

Notas

1. Aditya Challapally et al., "The GenAI Divide: State of AI in Business 2025", MIT Project NANDA, versão preliminar v0.1, julho de 2025. Base: mais de 300 iniciativas públicas revisadas, 52 entrevistas estruturadas e 153 executivos pesquisados em quatro conferências setoriais, entre janeiro e junho de 2025; amostra de conveniência, sem revisão por pares. O indicador de 95% refere-se a organizações que não obtêm retorno de suas iniciativas de IA generativa; o relatório registra separadamente que 60% das organizações avaliaram ferramentas corporativas, 20% chegaram ao piloto e 5% alcançaram produção. Os achados secundários citados no corpo constam do mesmo relatório: concentração dos investimentos em funções de contato com o cliente, com ganhos mais consistentes em backoffice, e taxas de sucesso superiores em iniciativas conduzidas com parceiros externos em comparação com construções internas.
2. Kyndryl, "Cloud Readiness Report 2025", novembro de 2025. Base: 3.700 líderes seniores em 21 países.
3. Deloitte, "The Pivot to Tokenomics: Navigating AI's New Spend Dynamics", 2026. Estudo global de custo. A faixa deste estudo deriva das duas inflexões anuais da Figura 5 da fonte: 67 bilhões de tokens por ano, quando o custo por token da fábrica de IA fica abaixo da API, e 84 bilhões, quando fica abaixo também dos provedores especializados de computação, ou 5,6 e 7,0 bilhões por mês. A comparação da fonte é entre alternativas da mesma classe (like-for-like): modelo aberto da classe de 70 bilhões de parâmetros em servidor de geração corrente, contra a API da mesma classe, a US\$ 3,98 por milhão de tokens na simulação; nó único servindo modelo de menor porte ajustado ao domínio, comparado ao serviço equivalente na tarefa. As faixas usadas neste estudo são sensibilidades paramétricas sobre essa base, e não fronteiras universais de mercado.
4. Febraban e Deloitte, "Pesquisa Febraban de Tecnologia Bancária 2026", 34ª edição, junho de 2026, Volume 1. Série setorial brasileira. Amostras equalizadas de 18 a 21 instituições nas séries de eficiência e de 25 nas de prioridade e maturidade; a quebra do orçamento por tema tem amostra de 16 instituições e o orçamento total, de 21; dados autorreportados. A pesquisa mede sete funções em duas séries: ganho médio de eficiência e percentual de bancos com ganhos elevados, definidos como 20% ou mais.
5. Stanford HAI, "AI Index Report 2025", capítulo 1, Research and Development, subseção Inference Cost. A melhora de cerca de 30% ao ano refere-se ao preço-desempenho do hardware, medido em capacidade computacional por dólar, e constitui série própria, de horizonte distinto da série de inferência.
6. Deloitte, "State of AI in the Enterprise 2026: The Untapped Edge", janeiro de 2026. Pesquisa global de adoção; base de 3.235 líderes em 24 países. Impacto transformador e ampliação de investimento, p. 10; uso de IA agêntica, p. 17; governança de agentes autônomos, p. 20; risco mais citado, p. 21; definição de soberania como atributo de país, p. 15; país de origem na seleção de fornecedor e composição local da pilha, p. 16.
7. Joaquin Fernandez, "The leading generative AI companies", IoT Analytics, 4 de março de 2025. Participação da NVIDIA no mercado de GPUs para data center em 2024, dimensionado pela fonte em US\$ 125 bilhões, com escopo restrito a sistemas destinados a data center.
8. TrendForce, "AI Demand Drives 4Q25 Global Top 10 Foundries Revenue Up 2.6% QoQ; Samsung Gains Share and Tower Moves Up in Rankings", 12 de março de 2026. Participação de 70,4% na receita das dez maiores foundries no quarto trimestre de 2025.
9. Synergy Research Group, "European Cloud Providers' Local Market Share Now Holds Steady at 15%", 24 de julho de 2025. Amazon, Microsoft e Google somam 70% do mercado europeu de infraestrutura de nuvem.
10. K. F. Pilz, J. Sanders, R. Rahman e L. Heim, "Trends in AI Supercomputers", arXiv:2504.16026, abril de 2025; e Epoch AI, "The US hosts the majority of GPU cluster performance, followed by China", 5 de junho de 2025. O percentual é participação dentro da base de dados da Epoch AI, que a própria organização estima cobrir entre 10% e 20% do desempenho agregado mundial em março de 2025.
11. Regulamento (UE) 2026/1744 do Parlamento Europeu e do Conselho, de 8 de julho de 2026, que altera o Regulamento (UE) 2024/1689 quanto à simplificação da aplicação de regras harmonizadas em matéria de inteligência artificial, em vigor desde 27 de julho de 2026. As obrigações de transparência do artigo 50 permanecem com aplicação em 2 de agosto de 2026, com prazo estendido até 2 de dezembro de 2026 para sistemas já colocados no mercado.
12. França, Décret n° 2026-272, de 14 de abril de 2026, de aplicação do artigo 31 da lei SREN n° 2024-449. O artigo 2° admite a qualificação pelo referencial nacional ou certificação da União Europeia ou de Estado do Espaço Econômico Europeu de nível ao menos equivalente. O prazo de dezoito meses é a duração máxima de derrogação para projetos anteriores à entrada em vigor.

13. ANPD, Resolução CD/ANPD nº 19, de 23 de agosto de 2024, com cláusulas-padrão contratuais obrigatórias desde 23 de agosto de 2025; e Resolução CD/ANPD nº 32, de 26 de janeiro de 2026, que reconhece a União Europeia como organismo internacional com grau de proteção adequado para os fins do artigo 33, inciso I, da Lei nº 13.709/2018, alcançando os Estados-membros, Islândia, Liechtenstein e Noruega, com dispensa de mecanismos adicionais de transferência e exclusão das finalidades de segurança pública, defesa nacional, segurança do Estado e investigação e repressão penal. No sistema financeiro: Resolução CMN nº 5.274 e Resolução BCB nº 538, de 18 de dezembro de 2025, que alteram a Resolução CMN nº 4.893/2021 e a Resolução BCB nº 85/2021, com prazo de adequação encerrado em 1º de março de 2026, instituindo trilhas de auditoria do processamento fim a fim, testes de intrusão anuais realizados com independência por terceiro contratado e isolamento físico e lógico dos ambientes do Pix e do Sistema de Transferência de Reservas, com instância dedicada em serviços de nuvem contratados. O regime de contratação no exterior permanece no artigo 16 da Resolução CMN nº 4.893/2021, que exige acordo para troca de informações entre o Banco Central e as autoridades supervisoras dos países de prestação e definição prévia, em contrato, dos países e regiões; na ausência do acordo, autorização prévia do Banco Central, com antecedência mínima de sessenta dias. Não há exigência de residência de dados em território nacional.
14. Brasil, PL 2338/2023, aprovado no Senado Federal em dezembro de 2024 e em tramitação na Câmara dos Deputados.
15. República da Coreia, Framework Act on the Development of Artificial Intelligence and Establishment of Trust (Lei nº 20676), em vigor desde 22 de janeiro de 2026, com alcance extraterritorial previsto no artigo 4º. Alterada pela Lei nº 21311, de 20 de janeiro de 2026, cujos dispositivos dependentes de regulamentação entraram em vigor em 21 de julho de 2026, junto com a alteração do decreto de execução.
16. Estados Unidos, CLOUD Act (Clarifying Lawful Overseas Use of Data Act), 18 U.S.C. § 2713; e U.S. Department of Justice, "Promoting Public Safety, Privacy, and the Rule of Law Around the World: The Purpose and Impact of the CLOUD Act", abril de 2019.
17. U.S. Department of Commerce, Bureau of Industry and Security, comunicado de 13 de maio de 2025; U.S. Government Accountability Office, decisão B-337935, de 12 de maio de 2026.
18. OpenAI, "Announcing The Stargate Project", 21 de janeiro de 2025. A formulação original declara a intenção de investir US\$ 500 bilhões em quatro anos, com US\$ 100 bilhões em implantação imediata.
19. Comissão Europeia, InvestAI, fevereiro de 2025; AI Continent Action Plan, COM(2025) 165, abril de 2025; e Cloud and AI Development Act, consulta pública de 2025. Os €200 bilhões são meta de mobilização, e não capital executado.
20. Brasil, MCTI, Plano Brasileiro de Inteligência Artificial (PBIa), 2024–2028.
21. Gartner, "Gartner Says Worldwide Sovereign Cloud IaaS Spending Will Total \$80 Billion in 2026", 9 de fevereiro de 2026 (US\$ 80,4 bilhões em 2026, alta de 35,6% sobre 2025). Complementarmente: Gartner, "Forecast Analysis: Sovereign Cloud IaaS, Worldwide", 25 de outubro de 2024.
22. Gartner, "Gartner Reveals Top Technologies Shaping Government AI Adoption", 9 de setembro de 2025. Predição também destacada em Deloitte, "A new era of self-reliance: Navigating technology sovereignty", TMT Predictions 2026, 18 de novembro de 2025.
23. NVIDIA Corporation, teleconferência de resultados do quarto trimestre e do ano fiscal de 2026, 25 de fevereiro de 2026. Declaração da diretora financeira sobre o crescimento e o volume da receita com projetos de IA soberana. Trata-se de divulgação da administração em teleconferência de resultados, e não de segmento reportado em demonstração financeira auditada; a companhia não divulga o critério de classificação dessa receita.
24. Comissão Europeia, "Cloud Sovereignty Framework", versão 1.2.1, outubro de 2025; e "Cloud Sovereignty Framework — Implementation guidance", 1º de junho de 2026. O vetor de quatro dimensões usado neste estudo é síntese autoral informada por esses referenciais.
25. Brasil, Ministério da Gestão e da Inovação em Serviços Públicos, "Gestão, Serpro e Dataprev lançam serviços de nuvem de governo para órgãos federais", 10 de junho de 2025; e Agência Gov, sobre a contratação de tecnologia de provedores globais instalada nos data centers das estatais. A fonte oficial informa investimento previsto superior a R\$ 1 bilhão e disponibilidade a mais de 250 órgãos federais, com tecnologias de AWS, Google, Huawei e Oracle operadas pelas empresas públicas.
26. Comissão Europeia, Commission Staff Working Document — Impact Assessment Report accompanying the proposal for a Cloud and AI Development Act, SWD(2026) 502 final, 3 de junho de 2026, p. 24 (CELEX 52026SC0502). A avaliação compara serviços de uma oferta soberana europeia de um provedor global de nuvem com a região convencional equivalente do mesmo provedor; em processos competitivos analisados sob confidencialidade, registra diferenças de aproximadamente -10% a +12% para níveis semelhantes de soberania.
27. NVIDIA Corporation, Formulário 10-K, exercício encerrado em 25 de janeiro de 2026, protocolado na U.S. Securities and Exchange Commission em 25 de fevereiro de 2026, accession nº 0001045810-26-000021. Dois clientes diretos responderam por 22% e 14% da receita total, contra 12% e 11% no exercício anterior.

- 28.** Maritaca AI, "Sabiá-3 Technical Report", arXiv:2410.12049, outubro de 2024; e "Sabiá-4 Technical Report", arXiv:2603.10213, março de 2026. Os resultados citados são do relatório de 2026: média de 86,6 no conjunto de treze exames brasileiros, contra 86,1 da geração anterior da fronteira e 92,9 e 95,0 dos modelos correntes de maior desempenho; e 97,4 no benchmark de conhecimento jurídico brasileiro, contra 88,6 do melhor concorrente. Os preços de entrada declarados no relatório são de US\$ 0,93 por milhão de tokens para o modelo nacional, contra US\$ 1,75 a US\$ 2,00 para os modelos de fronteira comparados; a razão é cálculo sobre a tabela publicada, e não afirmação do relatório.
- 29.** JLL, "Latin America Data Center Report — Year-end 2025", 23 de fevereiro de 2026, com o percentual brasileiro reproduzido por Forbes Brasil e CBIC. Apurações independentes convergem na mesma ordem de grandeza: Moody's Ratings estima 49% sobre base regional de 2.340 MW instalados, e a CBRE, em recorte mais estreito (São Paulo, inventário wholesale dos quatro maiores mercados), registra 51,4% no primeiro trimestre de 2026, na mesma ordem de grandeza.
- 30.** Brasil, MP 1.318/2025 (caducada em fevereiro de 2026) e PL 278/2026, aprovado na Câmara dos Deputados e em tramitação no Senado Federal.



A Deloitte refere-se a uma ou mais empresas da Deloitte Touche Tohmatsu Limited (“DTTL”), sua rede global de firmas-membro e suas entidades relacionadas (coletivamente, a “organização Deloitte”). A DTTL (também chamada de “Deloitte Global”) e cada uma de suas firmas-membro e entidades relacionadas são legalmente separadas e independentes, que não podem se obrigar ou se vincular mutuamente em relação a terceiros. A DTTL, cada firma-membro da DTTL e cada entidade relacionada são responsáveis apenas por seus próprios atos e omissões, e não entre si. A DTTL não fornece serviços para clientes. Por favor, consulte www.deloitte.com/about para saber mais.

A Deloitte oferece serviços profissionais de ponta para quase 90% das empresas listadas na Fortune Global 500® e milhares de outras organizações. Nossas pessoas entregam resultados mensuráveis e duradouros que ajudam a reforçar a confiança pública nos mercados de capitais e permitir que os clientes se transformem e prosperem. Com seus 180 anos de história, a Deloitte está hoje em mais de 150 países e territórios. Saiba como os cerca de 470 mil profissionais da Deloitte em todo o mundo geram um impacto que importa em www.deloitte.com.

© 2026. Para mais informações, contate a Deloitte Global.