Principais considerações sobre o desenvolvimento de um **ambiente de IA** pronto para produção

Red Hat

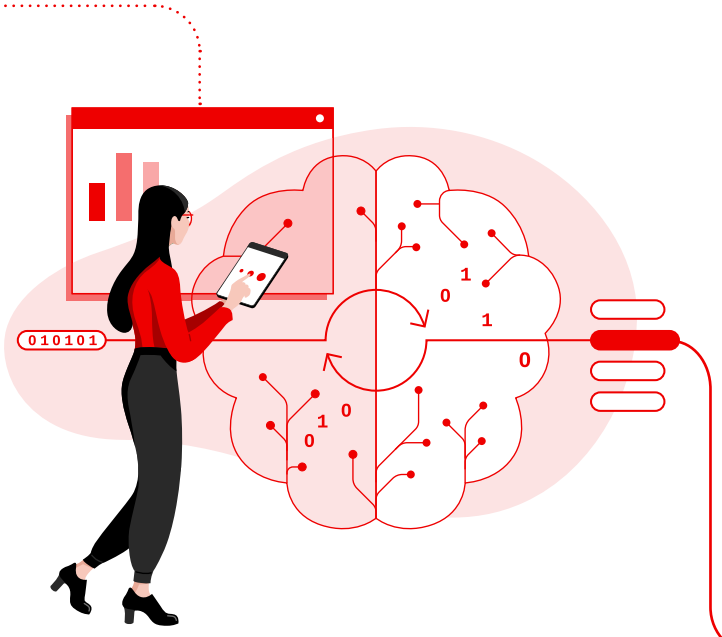
Sumário

Os dados são recursos empresariais essenciais

O estado atual do mercado de IA empresarial

A inteligência artificial generativa (gen IA) deixou de ser um experimento e se tornou uma ferramenta do dia a dia de muitas organizações.

As equipes a usam para resumir conteúdo, obter ajuda com o código e criar conteúdo, além de interagir com os dados de forma mais natural. Em escala empresarial, os líderes esperam que a gen IA os ajude a melhorar os resultados para clientes, funcionários e operações, em vez de apenas responder a perguntas específicas ou criar memes divertidos.



Com base nos dados e aplicações existentes, a gen IA pode ajudar as organizações a:

- Transformar grandes volumes de conteúdo não estruturado em conhecimento pesquisável e reutilizável.
- Ajudar desenvolvedores, analistas e redatores a criar e refinar códigos, relatórios e conteúdos com mais rapidez.
- Personalizar experiências digitais para clientes e funcionários em todos os canais.
- Automatizar decisões e fluxos de trabalho rotineiros que seguem políticas claras.
- Aumentar a produtividade das equipes de desenvolvimento, operações e negócios.

Pesquisas recentes do setor mostram que essa mudança já está em curso. A IDC relata que mais de 50% das organizações pesquisadas já implementaram diversas aplicações ou serviços aprimorados por gen IA na produção. A expectativa é que os gastos anuais com IA entre 2025 e 2029 cresçam aproximadamente um terço, chegando a cerca de US$ 1,3 trilhão até 2029.[1] Para a maioria das empresas, a gen IA está se tornando parte das principais soluções e serviços.

[1] Whitepaper da IDC. "Agentic AI to Dominate IT Budget Expansion Over Next Five Years, Exceeding 26% of Worldwide IT Spending, and $1.3 Trillion in 2029, According to IDC." 26 de agosto de 2025.

Ao mesmo tempo, as organizações estão visando a próxima etapa: agentic AI. Em vez de tratar a gen IA apenas como um chatbot ou assistente, a agentic AI usa agentes de IA que podem acionar ferramentas, interagir com aplicações e coordenar tarefas de várias etapas. Na prática, essa abordagem pode mudar a forma de desenvolver e operar softwares, como self-service de clientes, operações de TI e fluxos de trabalho empresariais complexos.

A IDC relata que mais de 50% das organizações já executam provas de conceito ou casos de uso iniciais da agentic AI, sendo que quase um terço das aplicações com IA dependerão dessa tecnologia até o fim de 2026.[2] As empresas agora a tratam como um caminho estratégico para o futuro.

Para agregar esse valor, você precisa de flexibilidade em como e onde executa a IA.

Agora, muitas organizações planejam uma infraestrutura de IA híbrida, combinando nuvens públicas com ambientes on-premise dedicados. A IDC destaca que uma combinação híbrida de nuvem pública e infraestrutura on-premise se tornou a estratégia de infraestrutura digital mais comum, e que a maioria dos tomadores de decisão acredita que suas cargas de trabalho de IA exigem implantação híbrida.[3]

Com uma plataforma híbrida open source, as organizações podem:

- Manter dados e modelos confidenciais sob controle.
- Atender aos requisitos de privacidade e soberania de dados.
- Escolher entre uma variedade de opções de hardware.
- Escolher entre uma grande variedade de modelos open source.
- Aproveitar a escalabilidade da nuvem quando necessário.

Este e-book aborda as principais etapas para desenvolver uma plataforma de IA pronta para produção, os principais pontos a serem considerados pela organização ao longo do processo e como o Red Hat® AI Enterprise oferece uma solução unificada para ajudar nesse desenvolvimento.

[2] Whitepaper da IDC. "Agentic AI Impact on Digital Infrastructure Strategies." Documento nº US53418526, outubro de 2025. (compra obrigatória)

[3] Whitepaper da IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Documento nº US53418426, outubro de 2025. (compra obrigatória)

Casos de uso da IA em diferentes setores

Setor de saúde
- Aumentar a eficiência clínica.
- Aprimorar a rapidez e a precisão dos diagnósticos.
- Melhorar os resultados de pacientes.

Telecom
- Obter insights sobre o comportamento de clientes.
- Aprimorar a experiência do cliente.
- Otimizar o desempenho da rede 5G.

Seguradoras
- Automatizar o processamento de solicitações.
- Oferecer serviços de seguro com base no uso.
- Auxiliar no cálculo de riscos.

Serviços financeiros
- Personalizar serviços voltados aos clientes.
- Aperfeiçoar a análise de risco.
- Detectar fraudes e lavagem de dinheiro.

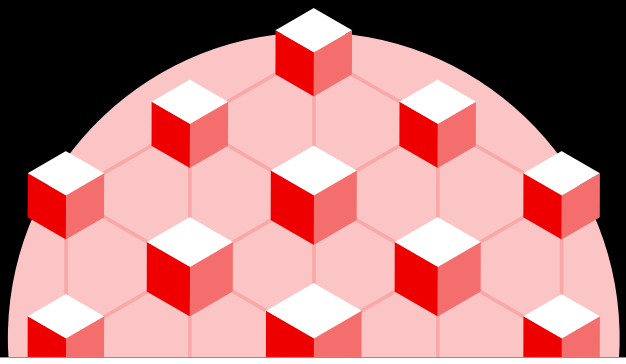
Automotivo
- Auxiliar a direção autônoma.
- Prever necessidades de manutenção.
- Aprimorar as cadeias de suprimentos.

Energia
- Prever requisitos de manutenção.
- Otimizar a segurança e as operações de campo.
- Acelerar a previsão e a simulação de reservatórios.

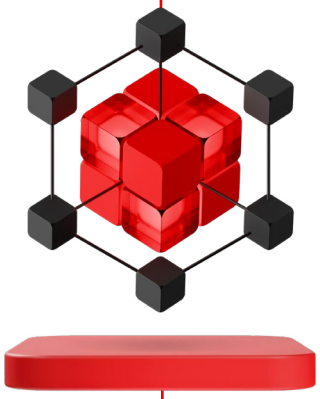
Componentes básicos da IA empresarial

Neste e-book, você descobre como diferentes tipos de IA funcionam em conjunto em uma arquitetura empresarial.

- **Gen IA**: usa Large Language Model (LLM) para gerar textos, códigos e outros conteúdos a partir de dados e prompts, para as equipes trabalharem com mais rapidez e experimentarem com mais facilidade.
- **IA preditiva**: usa dados históricos e em tempo real para estimar resultados futuros, como demanda, riscos ou integridade do equipamento, para as organizações agirem com antecedência e mais confiança.
- **Agentic AI**: usa agentes de IA que não só respondem a uma simples pergunta, mas podem acionar ferramentas, conectar a aplicações e coordenar fluxos de trabalho de várias etapas para atingir um objetivo.
- **Inferência de IA**: a fase de runtime da produção, quando os modelos aplicam o que aprenderam a novos dados reais para gerar previsões, recomendações ou ações. A inferência pode ser executada em um ambiente híbrido: on-premise, na nuvem ou na edge.



Desenvolva uma plataforma de IA pronta para produção



O desenvolvimento de aplicações e agentes de IA com a gen IA é um processo iterativo que vai além de apenas criar modelos de IA. As principais etapas do ciclo de vida da IA são:

1. Definir o caso de uso, estabelecer metas de negócios para a iniciativa de IA e obter a adesão dos stakeholders e da liderança.
2. Escolher onde você quer a execução das suas plataformas de implantação e experimentação de modelos: on-premise ou na nuvem.
3. Escolher o modelo de IA mais adequado às suas necessidades. Evite a dependência de fornecedor escolhendo modelos open source.
4. Personalizar ou alinhar os modelos escolhidos com seus dados proprietários usando Geração Aumentada de Recuperação (RAG).
5. Implantar seu modelo em um servidor de inferência.
6. Desenvolver cargas de trabalho e aplicações com tecnologia de gen IA.
7. Após criar um ambiente de trabalho, amplie e automatize o fluxo de trabalho com a agentic AI.
8. Monitorar e gerenciar modelos em grande escala e com foco na segurança.

Com uma arquitetura de IA open source e adaptável, você executa esse processo com mais eficiência. Essa arquitetura requer muitas tecnologias e recursos essenciais:

- **Acesso a modelos open-weight de ponta**: oferece às organizações um ponto de partida.
- **Ferramentas GenAIOps e DevOps**: permitem que engenheiros de IA, cientistas de dados, engenheiros de machine learning (ML) e desenvolvedor de aplicações criem, implantem e gerenciem modelos de IA, agentes de IA e aplicações com IA.
- **Acesso a ferramentas de ajuste de modelos, como ajuste fino e recursos de RAG**: para personalizar modelos com dados de empresas privadas e alinhar a casos de uso contextualizados.
- **Runtimes de inferência**: permitem oferecer o melhor desempenho, taxa de transferência e latência.
- **Componentes base dos agentes de IA**: para gerenciar, governar e proteger a implementação na produção.
- **Aceleradores de computação, armazenamento e rede**: para acelerar a preparação de dados, a personalização de modelos e as tarefas de inferência.
- **Endpoints de infraestrutura**: disponibilizam recursos em ambientes locais e virtuais, na edge e em ambientes de nuvens privadas, públicas e híbridas, em todos os estágios das operações de IA.

Neste e-book, você encontra uma análise das principais considerações sobre o desenvolvimento de uma arquitetura de IA eficaz.

> A inferência é o runtime da produção para a IA. Um modelo não é útil para você até ter uma API e oferecer conteúdo. Esse conteúdo é oferecido por meio da inferência.
>
> **Chris Wright**
> CTO da Red Hat[4]

[4] Miller, Ron. "Red Hat's CTO sees AI as next step for company's open approach." Fastforward, 11 de novembro de 2025.

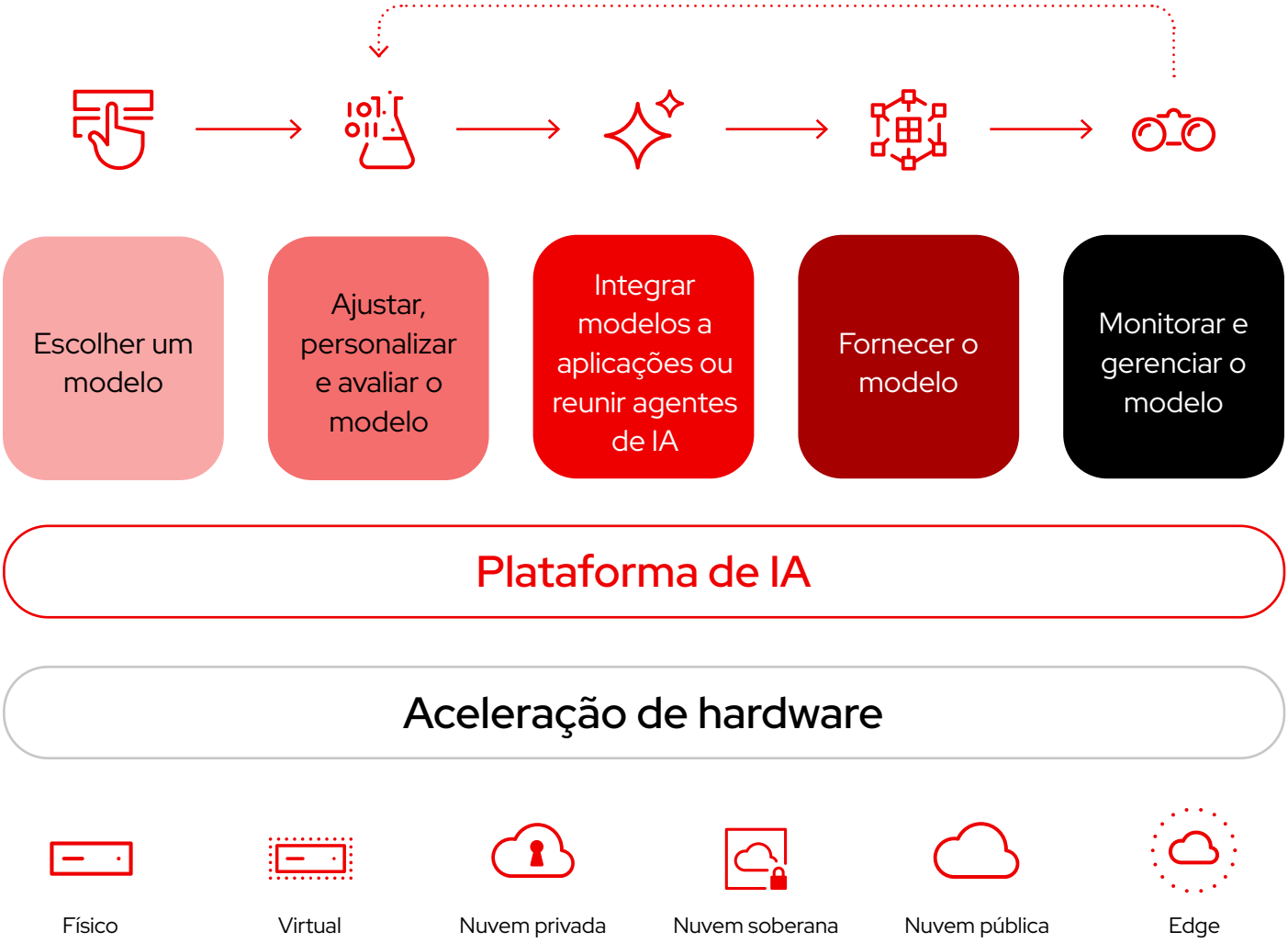


Figura 1. *Os componentes da arquitetura de IA.*

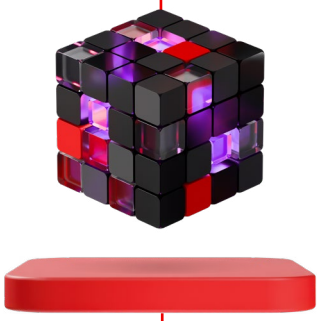
Desafios de implantar a IA

As empresas estão sendo pressionadas a escolher, desenvolver e entregar soluções de IA que ofereçam uma vantagem competitiva. Há vários desafios que impedem a operacionalização e a escala de implantações de IA:

- **Custo do modelo:** a execução de grandes modelos e inferência em grande escala pode ser cara. As organizações devem otimizar os modelos e a inferência para conter os custos computacionais, sem deixar de oferecer aplicações precisas e responsivas.
- **Complexidade do alinhamento:** o treinamento e o ajuste de modelos, além da criação de pipelines de RAG, são processos complexos que exigem um alto nível de processamento da unidade de processamento gráfico (GPU). As organizações podem simplificar a personalização dos dados empresariais e envolver desenvolvedores e especialistas no assunto para passar mais rápido dos experimentos à produção.
- **Controle e consistência:** os serviços de IA prontos limitam o controle sobre hardware, dados e governança. Escolha uma abordagem híbrida para selecionar os modelos e a infraestrutura, mantendo a propriedade dos dados, do ciclo de vida e da escalabilidade das implantações.

Para superar esses desafios, é necessário adotar uma plataforma híbrida de IA e open source que ofereça ferramentas consistentes para otimização, personalização e governança de modelos em todos os ambientes.

Containers e orquestração de containers



Containers

Um container é uma unidade básica de software que empacota as aplicações com suas respectivas dependências. Os containers simplificam o processo de desenvolvimento de aplicações e facilitam a implantação delas em ambientes diferentes, sem a necessidade de alterá-las.

Por que eles são importantes para a IA?

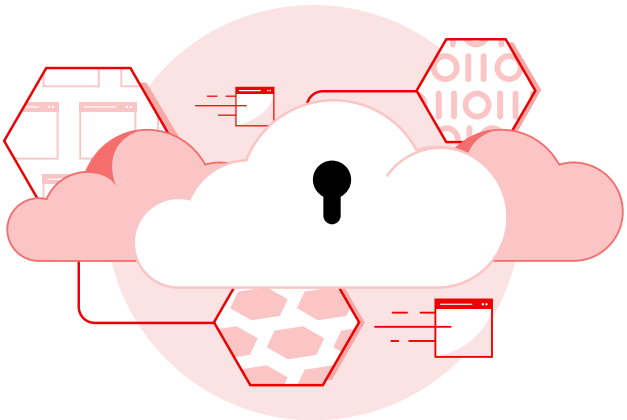
Engenheiros de IA e desenvolvedores de aplicações precisam ter acesso a suas ferramentas e recursos preferidos para serem mais produtivos. Ao mesmo tempo, as equipes de operações de TI precisam assegurar que os recursos estejam atualizados, em conformidade e possam ser usados de maneira segura.

Muitas vezes, os containers são a melhor opção para implantar LLMs e aplicações com tecnologia de gen IA porque eles reúnem servidores de modelos, dependências e configurações em uma unidade repetível. Isso facilita o gerenciamento de distribuição, escalonamento e atualizações da produção.

Os containers permitem implantar uma ampla seleção de ferramentas de IA em ambientes híbridos com consistência. As equipes podem modificar e compartilhar imagens de containers com recursos para controle de versão de forma iterativa, a fim de monitorar mudanças e assegurar a transparência. Além disso, o isolamento de processos e o controle de recursos aprimora a proteção contra ameaças.

Práticas recomendadas

Procure uma plataforma de aplicações em container flexível e altamente disponível. Ela deve incluir funcionalidades integradas de segurança e simplificar a implantação, o gerenciamento e a transferência de containers em todo o ambiente. Escolha uma plataforma open source que possa ser integrada a uma ampla variedade de tecnologias para ganhar mais flexibilidade e opções.



Orquestração de containers

A orquestração de containers envolve o gerenciamento da criação, da implantação e do ciclo de vida dos containers em todo o ambiente.

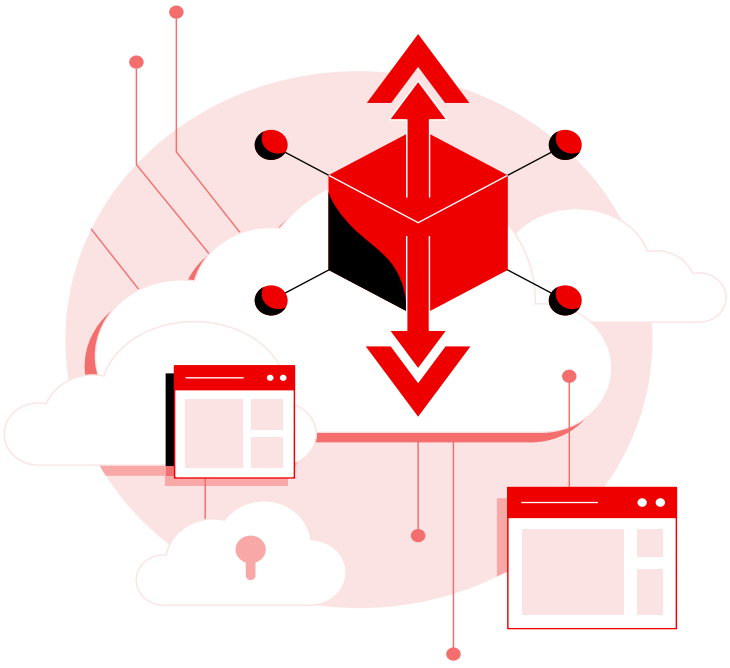
Por que isso é importante para a IA?

Após adotar a tecnologia de containers, você precisará de uma maneira de implantar, gerenciar e escalar seus containers com eficiência. Com um mecanismo de orquestração de containers, você pode administrar o ciclo de vida dos containers com consistência. Normalmente, essas ferramentas centralizam o acesso aos recursos de computação, armazenamento e rede em ambientes locais, de nuvem ou na edge. Elas também oferecem funcionalidades unificadas de programação de cargas de trabalho, controles de multilocação e imposição de cotas.

Práticas recomendadas

Selecione um ambiente de orquestração de containers com base em Kubernetes para aproveitar uma tecnologia open source líder do setor e evitar a dependência de uma nuvem proprietária. Escolha uma plataforma que ofereça controles robustos de multilocação, acesso baseado em função e gerenciamento de políticas para você controlar as cargas de trabalho de IA com consistência. Priorize opções com um amplo ecossistema de operadores e integrações para padronizar o modo de implantar, escalar e gerenciar serviços de IA em ambientes híbridos.

A previsão é de que, até 2027, mais de 75% de todas as implantações de IA usarão a tecnologia de containers como ambiente de computação subjacente, um aumento em relação aos menos de 50% em 2024.[5]



[5] Gartner. "Magic Quadrant for Container Management", 10 de setembro de 2024.

Considerações sobre a plataforma de IA

Gerenciamento de aplicações e genAIOps

Gerenciamento do ciclo de vida das cargas de trabalho de IA

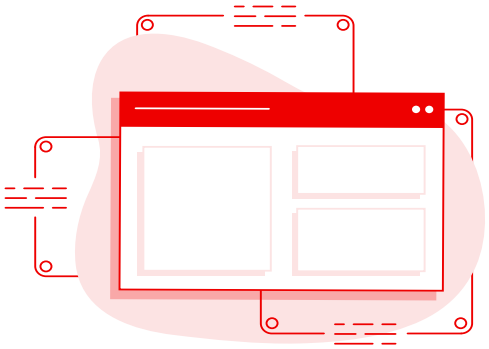
O gerenciamento do ciclo de vida das cargas de trabalho de IA se concentra em como você implanta, escala e administra as ferramentas e os serviços que impulsionam seus casos de uso de IA.

Por que isso é importante para a IA?

Os ambientes de IA são tipicamente complexos. Os componentes de gerenciamento do ciclo de vida das cargas de trabalho de IA, como notebooks, workbenches, pipelines e endpoints de model serving, devem ser inseridos em containers para permitir controle e gerenciamento simplificados. As equipes de operações de TI podem automatizar tarefas comuns do ciclo de vida, como configuração, provisionamento e atualizações, para aprimorar a precisão e reduzir o esforço manual. Os cientistas de dados, engenheiros de IA e desenvolvedores de aplicações podem solicitar ambientes de IA pré-aprovados a partir de um catálogo sem precisar abrir tickets com a equipe de TI. A automação também redireciona o tempo da equipe de tarefas repetitivas para atividades estratégicas de maior valor.

Práticas recomendadas

O gerenciamento eficaz do ciclo de vida das cargas de trabalho de IA começa com imagens selecionadas de workbenches e notebooks que incluem bibliotecas de IA/ML comumente usadas. Dessa forma, as equipes começam com uma base segura e com suporte, em vez de ambientes ad hoc. Para as equipes poderem colaborar em experimentos e acompanhar as alterações no código e no modelo ao longo do tempo, as organizações devem oferecer ambientes de notebook com base em navegador e integração Git.



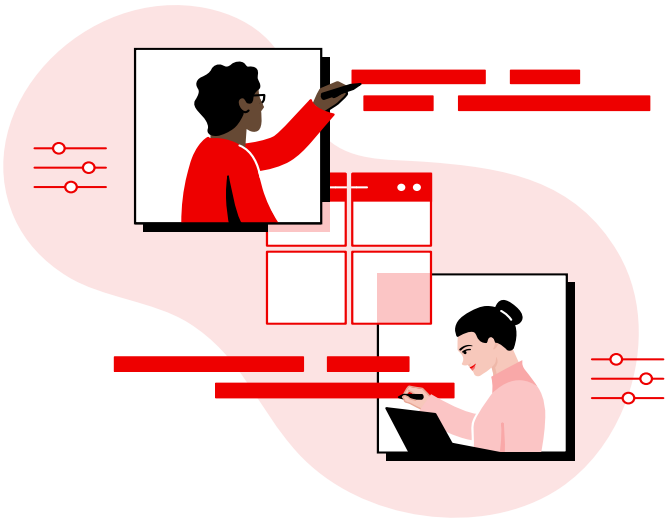
Práticas de GenAIOps e MLOps

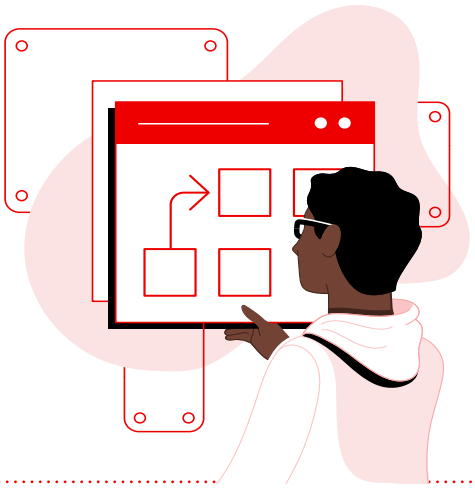
As práticas de GenAIOps e MLOps reúnem ferramentas, plataformas e processos necessários para operacionalizar a IA em grande escala.

Por que isso é importante para a IA?

As organizações precisam desenvolver e implantar modelos de IA, além de aplicações que usam esses modelos, com rapidez e eficiência. A colaboração entre equipes é fundamental para o sucesso desses esforços.

Semelhante ao DevOps, as abordagens de genAIOps e MLOps promovem a colaboração entre engenheiros de IA, desenvolvedores de aplicações e operações de TI para acelerar a criação, o treinamento, a implantação e o gerenciamento de modelos de gen IA, agentes de IA e aplicações com IA. A automação, geralmente na forma de pipelines de integração e entrega contínuas (CI/CD), possibilita a realização de alterações rápidas, incrementais e iterativas, para acelerar os ciclos de vida do desenvolvimento de aplicações e modelos.





Práticas de GenAIOps e MLOps

GenAIOps e MLOps não se resumem só à tecnologia, as pessoas e os processos também têm funções importantes. Aplique as práticas de genAIOps e MLOps em todo o ciclo de vida da IA. Use a automação, além de outras tecnologias open source, como o Kubeflow, nas plataformas e ferramentas para criar fluxos de trabalho e pipelines de CI/CD.

Considerações sobre a plataforma de IA

Plataforma de nuvem híbrida

Uma plataforma de nuvem híbrida oferece uma base para desenvolver, implantar e gerenciar a IA em ambientes locais, de nuvem ou na edge. Ela também permite projetar desde o início para a IA soberana e a IA privada. Dessa forma, você pode decidir quais cargas de trabalho serão executadas nas nuvens públicas e quais permanecerão nos ambientes de nuvem privada ou on-premise que você controla.

Por que isso é importante para a IA?

É necessário ter uma infraestrutura escalável para o desenvolvimento e a implantação de modelos, agentes, softwares e aplicações de IA. Com uma plataforma de nuvem híbrida consistente, é possível desenvolver, ajustar, testar, implantar e gerenciar modelos e aplicações de IA da mesma maneira e em todas as partes da infraestrutura, gerando mais flexibilidade.

Ela também viabiliza estratégias de IA soberana e IA privada. Isso permite manter dados e modelos confidenciais em regiões específicas ou até em ambientes desconectados, para atender a requisitos de conformidade, privacidade e residência de dados, sem perder a conexão oportuna com serviços em nuvem pública. Além disso, as funcionalidades de self-service podem acelerar a entrega de recursos, sem negligenciar o controle da TI.

Por fim, uma plataforma consistente oferece a base para integrações com tecnologias de outros fornecedores, comunidades open source e qualquer ferramenta personalizada utilizada por você.

Práticas recomendadas

Selecione uma plataforma que priorize a segurança, seja compatível com a aceleração de hardware, ofereça um amplo ecossistema de ferramentas de IA e desenvolvimento de aplicações e possa ser integrada a recursos de gerenciamento de operações e genAIOps.

Busque controles robustos de política para localização de dados, posicionamento de modelos e acesso. Assim, você pode executar cargas de trabalho de IA privada e soberana on-premise ou em nuvens privadas, enquanto mantém a conexão com nuvens públicas quando necessário.

Ao escolher uma plataforma open source, você ganha mais integração e flexibilidade, acelera a inovação com o apoio da comunidade e agiliza a entrega de recursos com funcionalidades de self-service, sem perder o controle da TI.

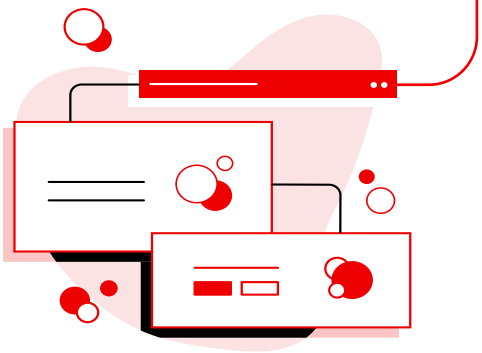
A estratégia arquitetônica de infraestrutura digital mais comum é a combinação híbrida de nuvem pública e infraestrutura on-premise dedicada.[3]

[3] Whitepaper da IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Documento nº US53418426, outubro de 2025. (compra obrigatória)

Considerações sobre a plataforma de IA

Personalização e alinhamento de modelos

As aplicações modernas com IA exigem modelos que reflitam os dados, fluxos de trabalho e restrições empresariais específicos de uma organização. Ao alinhar um modelo de ponta ou open source às suas informações proprietárias, você passa de respostas genéricas a resultados precisos e alinhados ao contexto do seu domínio.



Por que isso é importante para a IA?

A gen IA e a agentic AI dependem de modelos que entendam sua terminologia, dados e contexto do mundo real.

O alinhamento ajuda a manter a precisão e a relevância vinculando modelos aos seus dados privados. Ele melhora a eficiência reduzindo o custo da inferência e evitando o superdimensionamento desnecessário. Ele também fortalece a governança e o controle, permitindo que você implemente a lógica de negócios, as regras de segurança e os requisitos de conformidade diretamente no comportamento do modelo. Além disso, ele oferece suporte à escalabilidade com processos consistentes para atualizar, retreinar e controlar a versão de modelos conforme os dados evoluem.

A personalização também facilita estratégias de IA soberana e IA privada. Dessa forma, as organizações podem treinar e disponibilizar modelos integralmente nos ambientes controlados, a fim de atender às exigências regulatórias, de privacidade e de residência dos dados.

Práticas recomendadas

Adote fluxos de trabalho modulares que comecem com RAG, ajuste fino, engenharia de prompt e camadas de políticas com base nas suas necessidades, em vez de depender de um único método. Use modelos open source para evitar a dependência de fornecedor e manter a capacidade de fazer ajuste fino, quantizar e avaliar modelos com transparência. Inclua especialistas no assunto para assegurar que os modelos reflitam o contexto real dos negócios e a precisão dos dados. Otimize seu modelo para inferência com antecedência, aplicando técnicas como quantização, destilação e runtimes eficientes para controlar o custo e a latência. Além disso, use conjuntos de dados em versões, execuções de treinamento, pesos de modelos e métricas de avaliação para manter a reprodutibilidade e uma governança sólida.

Considerações sobre a plataforma de IA

Inferência de IA em grande escala

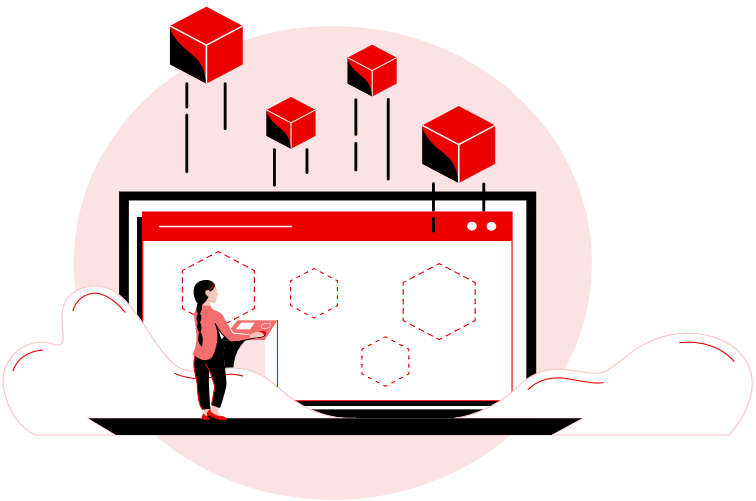
A execução da IA na produção depende de inferência rápida, eficiente e confiável. Após o treinamento ou alinhamento dos modelos, a inferência é a fase em que eles processam novos dados, retornam previsões, geram conteúdo ou desencadeiam ações em uma aplicação ou fluxo de trabalho.

Conforme as organizações adotam a gen IA e a agentic AI, a inferência passa a ser um fator crítico de custo e desempenho, especialmente porque as aplicações evoluem de interações pontuais para tarefas contínuas e de várias etapas executadas por agentes de IA.

Por que isso é importante para a IA?

A inferência influencia diretamente a experiência do usuário, o desempenho da aplicação e o custo operacional. Geralmente, as cargas de trabalho de gen IA e agentic AI exigem respostas rápidas, solicitações paralelas e taxa de transferência consistente em muitos ambientes, como data centers, nuvem pública e edge.

Com runtimes de inferência eficientes, é possível reduzir o custo da GPU e da CPU, melhorar a latência de tarefas interativas e atender às necessidades de escalabilidade de agentes de IA que acionam ferramentas, usam interfaces de programação de aplicações (APIs) e coordenam fluxos de trabalho de várias etapas. A otimização da inferência também favorece as estratégias de IA soberana e IA privada, permitindo que as organizações executem inferência próxima a dados confidenciais, on-premise ou em nuvens privadas, com desempenho previsível.

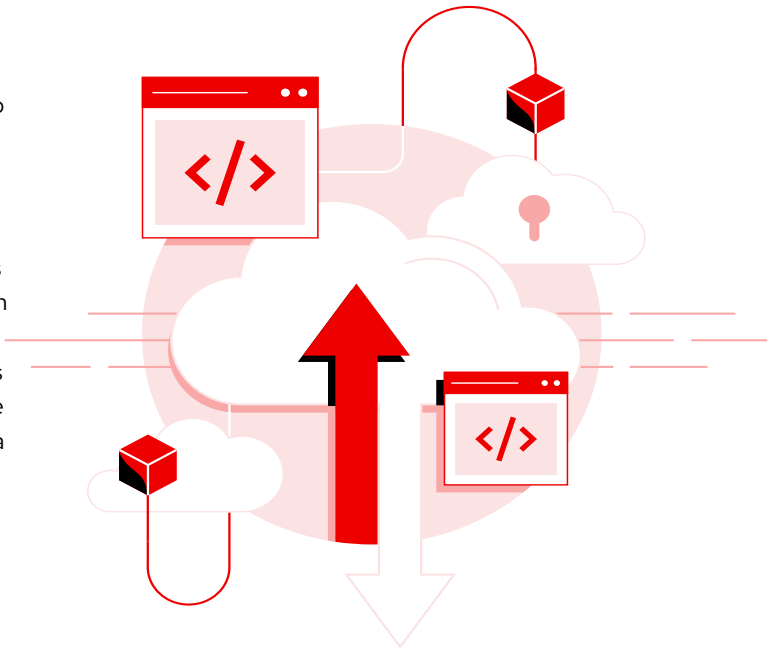


Práticas recomendadas

Escolha runtimes de inferência otimizados que se adequem ao seu tipo de modelo e ambiente de implantação, seja para LLMs, modelos multimodais, modelos preditivos ou cargas de trabalho agentic. Priorize runtimes e infraestruturas compatíveis com escalabilidade dinâmica, horizontal e vertical, para atender às demandas imprevisíveis de LLM interativo e inferência baseada em agente.

Use técnicas ou faça parcerias com fornecedores que tenham expertise em abordagens como quantização, destilação e otimização de modelos para reduzir custos e melhorar a latência. Combine essas organizações com tecnologias amplamente adotadas, como o vLLM para inferência de LLM de alta taxa de transferência e novos frameworks de inferência distribuída, como o llm-d, que desagregam o processo de inferência para escalar cada fase independentemente.

Implante inferência nos containers para agrupar dependências e escalar com consistência nos ambientes híbridos. Aplique endpoints de inferência onde seus dados e aplicações são armazenados para reduzir a movimentação dos dados e manter o controle, especialmente em cenários de IA soberana e privada. Por fim, para manter a precisão e a confiabilidade em grande escala, monitore o desempenho dos modelos ao longo do tempo e atualize as versões conforme as distribuições de dados mudarem.



90%
dos tomadores de decisão acreditam que a IA será um motivador essencial para o orçamento da infraestrutura digital e as escolhas de tecnologia em 2026.[3]

[3] Whitepaper da IDC. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Documento nº US53418426, outubro de 2025. (compra obrigatória)

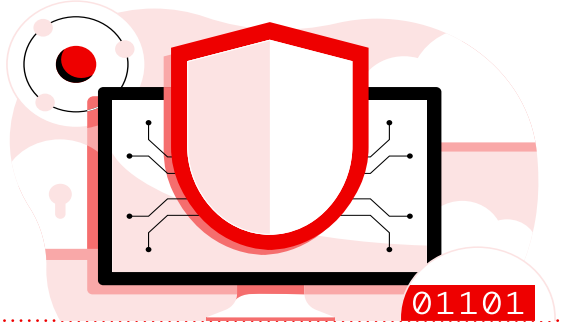
Considerações sobre a plataforma de IA

Segurança de IA

Os sistemas de IA devem se comportar de forma confiável, previsível e alinhada às políticas organizacionais. Conforme as empresas migram da experimentação para a produção, a segurança da IA se torna um requisito essencial, especialmente na implantação de gen IA, agentic AI e fluxos de trabalho autônomos, que podem executar ações, não só oferecer sugestões.

Por que isso é importante para a IA?

A segurança tem como foco manter os modelos e agentes de IA dentro dos limites definidos para atender aos requisitos empresariais, legais e éticos. Resultados imprecisos, desvios do modelo, tratamento de dados precário ou ações não intencionais podem gerar riscos operacionais reais. Os sistemas agentic e de gen IA também apresentam novas questões de segurança, como alucinações, execução de ferramentas não aprovadas, escalonamento de privilégios e raciocínio inconsistente em tarefas de várias etapas. Com práticas robustas de segurança, você mantém a confiança, protege os dados confidenciais e impede ações prejudiciais ou irreversíveis. Em setores regulamentados, os controles de segurança são essenciais para assegurar a conformidade e a preparação para auditoria em ambientes híbridos e on-premise.



Práticas recomendadas

Adote uma abordagem de segurança em camadas que inclua medidas de segurança com base em políticas, filtros de conteúdo e controles de execução de ferramentas para fluxos de trabalho agentic. Para detectar desvios ou degradação da precisão, valide e teste modelos regularmente. Execute cargas de trabalho confidenciais em ambientes privados ou on-premise para manter o controle sobre a exposição dos dados e o comportamento do modelo, alinhando-se às estratégias de IAs soberana e privada. Use frameworks de avaliação de modelos para monitorar vieses, robustez e confiabilidade. Amplie seus modelos e dados usando ferramentas que os armazenem em registros compatíveis com containers padrão (OCI) e ofereçam cadeias de suprimentos seguras. Tecnologias amplamente adotadas, como o vLLM para inferência de LLM, e novas tecnologias distribuídas, como o llm-d, podem ajudar a reduzir custos e escalar a implantação do projeto de IA. Por fim, controle a versão e documente seus modelos, conjuntos de dados e políticas para monitorar decisões e gerenciar uma governança consistente em todo o ciclo de vida da IA.

Tenha uma base aberta e flexível para IA

O **Red Hat AI Enterprise** é uma plataforma de IA integrada que faz parte do portfólio do Red Hat AI. Ele é usado para desenvolvimento e implantação eficientes e econômicos de modelos, agentes e aplicações de IA em ambientes em nuvem híbrida.

Ele unifica os ciclos de vida de aplicações e modelos de IA para melhorar a eficiência operacional, acelerar a entrega e mitigar riscos, oferecendo um ambiente de desenvolvimento pronto para uso com recursos de nível empresarial.

Desenvolvida com a tecnologia do Red Hat OpenShift, essa plataforma é um stack de IA completo, testado e comprovado que aprimora a interoperabilidade e garante a continuidade dos negócios. Ela inclui recursos essenciais, como ajuste de modelos, inferência de alto desempenho e gerenciamento de fluxos de trabalho de agentic AI. Com isso você tem a flexibilidade necessária para oferecer suporte a qualquer modelo, usar qualquer hardware e implantar em qualquer ambiente, atendendo aos requisitos de localização de dados. O Red Hat AI Enterprise é compatível com ambientes híbridos. Dessa forma, as equipes podem planejar capacidade, GPUs e futuros projetos de IA com confiança.

O Red Hat AI Enterprise inclui tecnologia do projeto open source llm-d, lançado pela Red Hat com colaboradores como a IBM, NVIDIA, Google, AMD e outros. Esse projeto llm-d melhora a eficiência de custos, separando as fases de pré-preenchimento e decodificação da inferência para o escalonamento de cada uma ser diferente. Ele apresenta um balanceador de carga com reconhecimento de inferência que encaminha as solicitações com base nas filas de token, melhorando os tempos de resposta e, em alguns casos, direcionando as cargas de trabalho de pré-preenchimento para as CPUs.

Acelerar o time to value (TTV).
Implante um stack de IA empresarial na sua infraestrutura preferida, com ferramentas pré-configuradas, implantações automatizadas e observabilidade integrada. Dessa forma, os desenvolveres e engenheiros de IA podem se concentrar em desenvolver e entregar aplicações agentic nativas em nuvem e com IA.

Aumentar a eficiência operacional.
Otimize e automatize fluxos de trabalho, dos commits de código à criação dos fluxos de trabalho de pipeline da IA por meio da implantação de modelos. Com isso, as operações de TI podem oferecer desempenho consistente e obter mais valor da infraestrutura existente com alocação inteligente de recursos e gerenciamento integrado do ciclo de vida.

Reduzir riscos.
Reduza o risco da adoção da IA empresarial com um stack de IA integrado, testado e com suporte completo, que aprimora a interoperabilidade em todos os modelos, hardwares e ambientes de nuvem híbrida. Para escalar a IA com confiança, use essa base para atender às exigências regulatórias e de residência de dados.

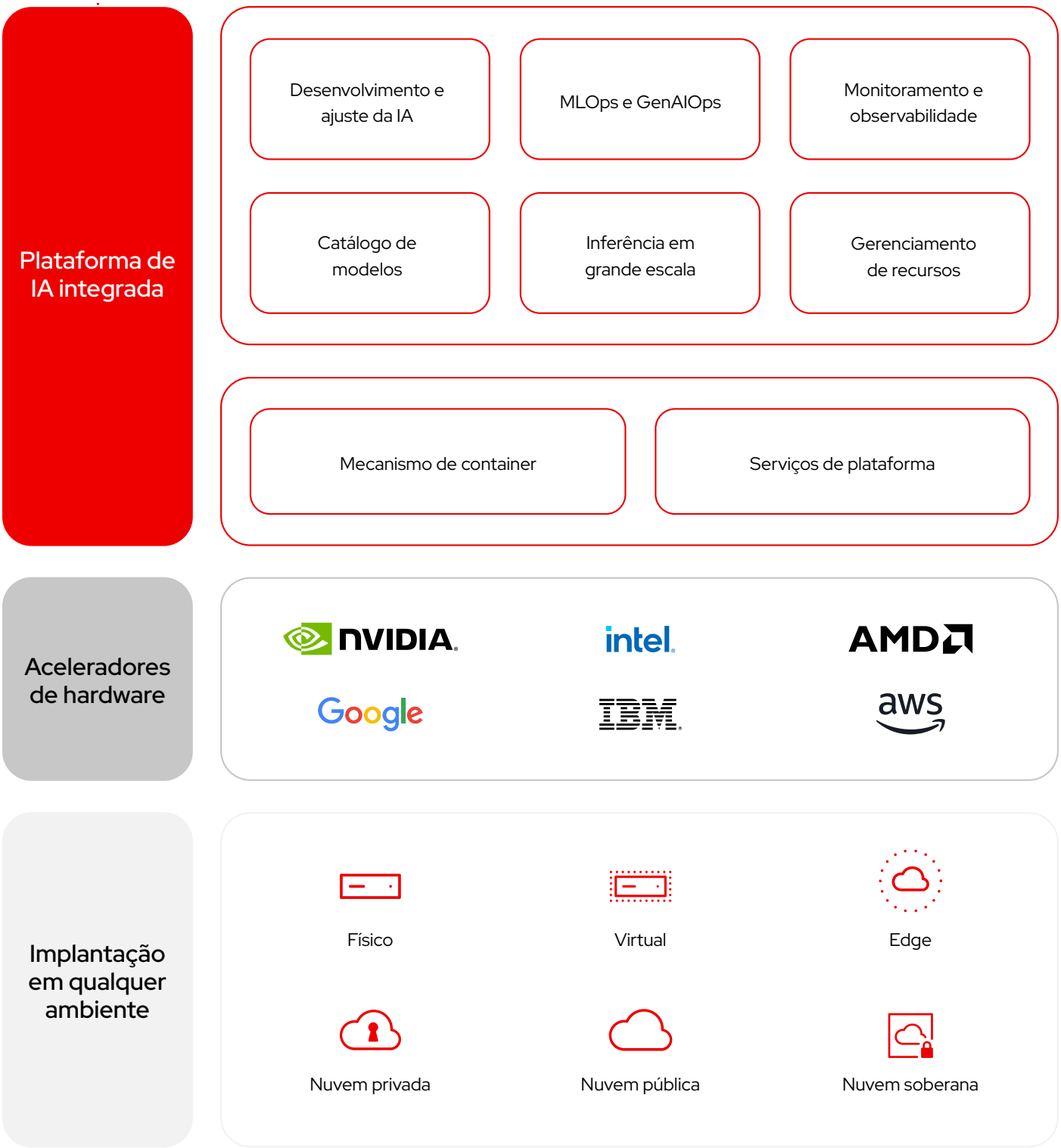
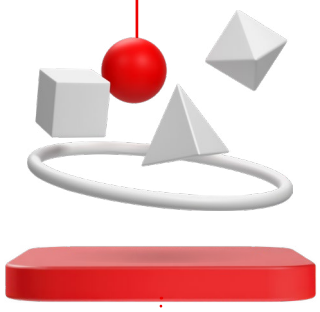


Figura 2. *Os componentes de uma plataforma de IA integrada.*

Tenha opções e flexibilidade com um ecossistema de parceiros certificados de IA/ML



O cenário de ferramentas e tecnologias de IA continua em rápida evolução, o que torna um desafio acompanhar os avanços e manter a estabilidade e a confiabilidade no ambiente de TI.

Por meio de parcerias com a NVIDIA, AMD, Intel e parceiros de tecnologia de IA, o Red Hat AI Enterprise oferece uma plataforma de IA empresarial de ponta a ponta que escala na nuvem híbrida, oferecendo implantação rápida, eficiência aprimorada e suporte à nuvem híbrida. Os programas de validação e certificação da Red Hat garantem a utilização completa do hardware, e o gerenciamento otimizado da carga de trabalho assegura o uso eficiente da GPU, maximizando o desempenho e o valor para os clientes.

A presença da Red Hat no ecossistema da Hugging Face e o catálogo de servidores do Model Context Protocol (MCP) oferecem aos clientes acesso a uma crescente biblioteca de modelos validados e ferramentas pré-integradas que são executados consistentemente com o Red Hat AI Enterprise. Ao mesmo tempo, as parcerias com diversos fornecedores de aceleradores ajudam as organizações a aproveitar as GPUs e o hardware especializado nos ambientes híbridos. Você pode escolher com confiança os parceiros, modelos, ferramentas e tecnologias que melhor se adequam às suas necessidades. Eles funcionarão bem juntos e contam com serviços especializados, suporte e treinamento para ajudar você a desenvolver e escalar fluxos de trabalho de IA com sucesso.

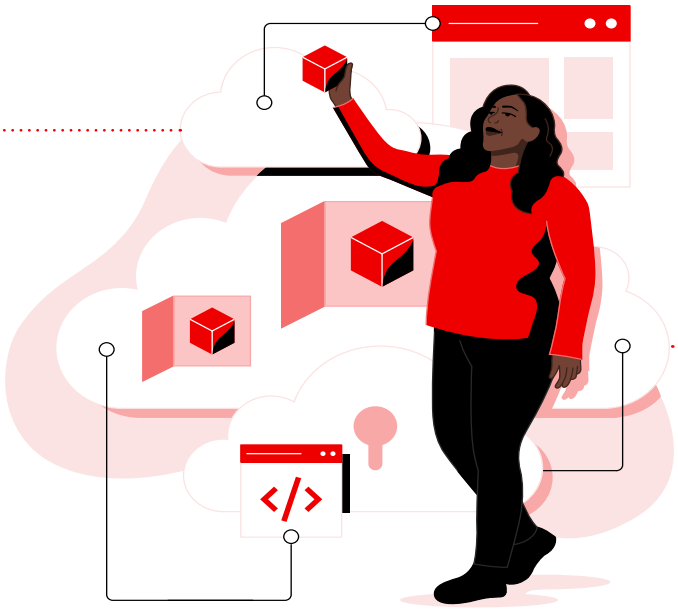
Destaques do parceiro

NVIDIA

A NVIDIA é líder global em tecnologia de IA. Ela oferece a melhor arquitetura e plataforma de GPU para IA e computação acelerada, oferecendo desempenho e eficiência energética de ponta para treinamento e implantação de grandes modelos de IA. A parceria estratégica da NVIDIA e da Red Hat tem como foco disponibilizar software e infraestrutura nativa em nuvem para IA. Com o Red Hat AI Enterprise, os clientes recebem um conjunto completo, otimizado e nativo em nuvem de softwares de IA e data analytics, além do hardware necessário para executá-lo. Os sistemas Red Hat AI Enterprise e NVIDIA DGX proporcionam capacidade de gerenciamento para a infraestrutura de IA. O NVIDIA GPU Operator automatiza o gerenciamento de todos os componentes de software da NVIDIA necessários para provisionar GPUs.

AMD

A AMD e a Red Hat trabalham juntas para ampliar as opções de IA e virtualização nos ambientes de nuvem híbrida. Essa colaboração integra processadores AMFD EPYC e aceleradores AMD Instinct ao Red Hat AI Enterprise. Isso permite que as organizações tenham mais flexibilidade para executar cargas de trabalho de treinamento, ajuste e inferência de IA em hardware de alto desempenho e econômico. Os esforços conjuntos de engenharia e certificação garantem que os sistemas com tecnologia AMD funcionem consistentemente com o Red Hat AI Enterprise, incluindo suporte para imagens de container otimizadas, operadores de Kubernetes e desempenho validado para pipelines de IA acelerados por GPU.



intel

A Intel e a Red Hat trabalham juntas para oferecer infraestruturas de software e plataformas padrão da indústria para melhorar a agilidade e flexibilidade do data center. A distribuição das ferramentas OpenVINO da Intel otimiza e converte modelos de IA em mecanismos de inferência de alto desempenho que podem escalar automaticamente para milhares de nós no Red Hat AI Enterprise.

Sucesso em ação

Turkish Airlines

A Turkish Airlines usa o Red Hat AI para modernizar as operações e inovar com o uso de IA no setor aéreo. Com a padronização em uma plataforma de IA escalável e open source, a companhia aérea acelera o desenvolvimento de modelos, aprimora o serviço aos passageiros e otimiza a tomada de decisões operacional, provando que a IA híbrida pode transformar uma das maiores redes aéreas do mundo.

Leia mais

Denizbank

O Denizbank usa o Red Hat AI para acelerar a inovação com IA no seu ecossistema bancário digital. Com a modernização da infraestrutura de IA usando uma plataforma escalável e open source, o banco agiliza a experimentação, aprimora a confiabilidade dos modelos e oferece experiências mais inteligentes aos clientes. Isso demonstra como a IA híbrida ajuda as instituições financeiras a avançarem mais rapidamente, mantendo uma postura rigorosa de segurança e governança.

Leia mais

AGESIC

AGESIC, a agência governamental digital do Uruguai, usa o Red Hat AI para padronizar e escalar a IA em mais de 180 entidades públicas. A plataforma híbrida de IA viabiliza práticas de MLOps, fortalece a segurança e ajuda as equipes a desenvolver, implantar e governar aplicações de IA que melhoram os serviços para os cidadãos.

Leia mais

Tudo pronto para aproveitar melhor seus dados?

A IA está transformando praticamente todos os aspectos dos negócios. A Red Hat pode ajudar você a desenvolver um ambiente de IA pronto para produção que acelera o desenvolvimento e a entrega de aplicações inteligentes. Assim, sua empresa fica mais perto de alcançar as metas de negócios.

Descubra mais sobre como o Red Hat AI Enterprise ajuda você a desenvolver uma plataforma unificada para a IA.