Red Hat
OpenShift AI

Uma plataforma aberta para modelos de IA na nuvem híbrida

Destaques

Operacionalize e escale a inferência de IA e a agentic AI em uma plataforma de aplicações comprovada.

Melhore a eficiência operacional em projetos de inteligência artificial e machine learning com uma experiência do usuário consistente, que facilita o trabalho de cientistas de dados, engenheiros de dados, desenvolvedores de aplicações e equipes de DevOps.

Obtenha a flexibilidade da nuvem híbrida criando, treinando, implantando e monitorando cargas de trabalho de inteligência artificial e machine learning on-premise, na nuvem ou na edge.

Adote aplicações inteligentes e IA generativa

A inteligência artificial (IA), o machine learning (ML) e o deep learning (DL) vêm influenciando muito os esforços de modernização de aplicações em diversos negócios e setores. A necessidade de inovar, extrair valor estratégico e obter novos insights a partir de dados vem expandindo o uso de aplicações nativas em nuvem com IA, MLOps e metodologias de GenAIOps. Os desafios nesse admirável mundo novo podem ser complexos, como, por exemplo, o rápido aumento dos custos dos modelos ao passar para a produção, a personalização complexa, as rígidas restrições de implantação e as operações necessárias para acompanhar o ritmo da inovação. As empresas precisam de soluções que reduzam os custos de inferência, simplifiquem a escala e o monitoramento e se adaptem às constantes mudanças.

O Red Hat® AI acelera o desenvolvimento e a implementação de soluções de IA empresarial em ambientes de nuvem híbrida. Ele funciona como uma plataforma abrangente para gerenciar todo o ciclo de vida de inteligência artificial e machine learning, oferecendo recursos de MLOps e GenAIOps. O Red Hat AI tem como foco quatro pilares principais:

- Aumentar a eficiência com inferência rápida, flexível e eficiente.
- Simplificar a experiência de conexão de modelos a dados.
- Acelerar a inovação da agentic AI.
- Garantir flexibilidade e consistência ao escalar a IA na nuvem híbrida.

O Red Hat OpenShift® AI, baseado no Red Hat OpenShift, uma plataforma de aplicações em nuvem híbrida líder do setor, é a principal solução no portfólio do Red Hat AI. A plataforma de IA oferece aos engenheiros, cientistas de dados e desenvolvedores uma base de inteligência artificial e machine learning de alta performance para o desenvolvimento e a implantação de modelos generativos e preditivos e aplicações de IA em grande escala. Com ele, as empresas podem experimentar uma variedade de ferramentas, colaborar e acelerar o time to market, tudo no mesmo lugar. O Red Hat OpenShift AI combina o ambiente self-service que desenvolvedores e cientistas de dados desejam com a confiança de que a TI empresarial precisa.

Desenvolva, treine, teste e implante rapidamente

O Red Hat OpenShift AI é uma plataforma de MLOps flexível e escalável desenvolvida com tecnologias open source, oferecendo recursos confiáveis e operacionalmente consistentes para as equipes experimentarem, disponibilizarem modelos e entregarem aplicações inovadoras. O OpenShift AI acelera a entrega de aplicações com IA, ajudando as organizações a migrar de pilotos iniciais para implantações operacionalmente robustas com mais velocidade e controle.

A plataforma oferece uma experiência de interface de usuário (UI) integrada com ferramentas para desenvolver, treinar, ajustar, implantar e monitorar modelos de IA preditiva e generativa. Você pode implantar modelos em ambientes de nuvem híbrida, com ênfase específica em disponibilizar uma área de ocupação controlada e protegida para IA privada e soberana. Com essa abordagem, dados confidenciais e modelos de IA permanecem dentro dos limites geográficos ou organizacionais designados, atendendo a rigorosos requisitos de conformidade e regulatórios.

Red Hat
OpenShift AI

Previsões de analistas de gen IA

"Espera-se que a IA seja um fator muito importante para impulsionar os orçamentos de infraestrutura digital em 2026 conforme as organizações trabalham para adequar os requisitos de carga de trabalho e dados às opções de infraestrutura híbrida. 90% dos tomadores de decisão acreditam que a IA será um motivador importante no orçamento da infraestrutura digital e das escolhas tecnológicas deles em 2026."[1]

Simplifique a adoção da IA

Como um complemento do Red Hat OpenShift, o OpenShift AI fornece uma plataforma projetada para aumentar a adoção e a confiança nas iniciativas de IA, combinando comunidades open source com um ecossistema de IA robusto. Isso oferece mais flexibilidade e liberdade para escolher a tecnologia de inteligência artificial e machine learning ideal para sua organização. Os usuários podem desenvolver modelos preditivos ou começar com um modelo externo de gen IA e aprimorá-lo com geração aumentada de recuperação (RAG), usando um dos vários servidores de modelo da plataforma. A plataforma oferece acesso rápido a modelos de terceiros otimizados e validados, como Llama, Mistral, DeepSeek e Granite, que funcionam com eficiência em vLLM, disponíveis no repositório do Red Hat AI no Hugging Face. Com o catálogo, os usuários podem explorar esses modelos e incluir os seus próprios.

Aprimore a consistência operacional entre as equipes

O Red Hat OpenShift AI oferece uma experiência de usuário consistente que permite que cientistas de dados, engenheiros de IA, desenvolvedores e equipes de DevOps colaborem efetivamente para entregar soluções de IA oportunas. Ele oferece acesso self-service a fluxos de trabalho colaborativos, aceleração da unidade de processamento gráfico (GPU) e operações otimizadas, possibilitando uma entrega consistente de soluções de IA em grande escala em ambientes de nuvem híbrida e na edge da rede.

As operações de TI se beneficiam das configurações simplificadas e dos fluxos de trabalho mais automatizados em uma plataforma verificada que pode escalar vertical ou horizontalmente com pouco esforço, ao mesmo tempo em que oferece melhor governança e segurança.

Aproveite a flexibilidade da nuvem híbrida

O Red Hat OpenShift AI permite treinar, implantar e monitorar cargas de trabalho de inteligência artificial e machine learning em vários ambientes (nuvem, data centers on-premise ou ambientes isolados) para atender aos requisitos regulatórios, de segurança e de dados. A plataforma é compatível com diversos aceleradores de IA de fornecedores como NVIDIA, AMD e Intel. Esse recurso pode ser expandido para criar um ambiente de GPU como serviço, possibilitando às organizações gerenciar, particionar e programar recursos de GPU centralmente, além de oferecer observabilidade detalhada do uso.

Agentic e gen IA

Para projetos de gen IA, experiências de usuário dedicadas são oferecidas por componentes como o AI hub (prévia para desenvolvedores), uma experiência de dashboard para engenheiros de plataforma, consolidando as implantações de catálogo, registro e modelo para configurar e implantar modelos e servidores MCP. O gen AI studio (prévia para desenvolvedores) oferece endpoints de ativos de IA e um ambiente onde engenheiros de IA e desenvolvedores de aplicações podem acessar, experimentar, comparar, avaliar e testar modelos implantados e servidores MCP.

O OpenShift AI acelera a agentic AI ao fornecer uma camada de API unificada e uma base flexível e escalável. O suporte à API Llama Stack e ao MCP (prévia de tecnologia) inclui uma implementação empresarial da API Llama Stack, oferecendo um único ponto de entrada padronizado para vários recursos de IA.

Outras ferramentas incluem avaliação de LLM (LM Eval) e benchmark de LLM para ajudar nas implantações de inferências no mundo real. O compressor de LLM disponibiliza algoritmos para reduzir o tamanho dos modelos personalizados de uma organização, usando métodos semelhantes aos usados pela Red Hat para criar modelos validados e otimizados no repositório do Red Hat AI no Hugging Face.

1. *IDC Tech Supplier. "AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025." Documento nº US53418426, outubro de 2025. (Exige login do cliente.)*

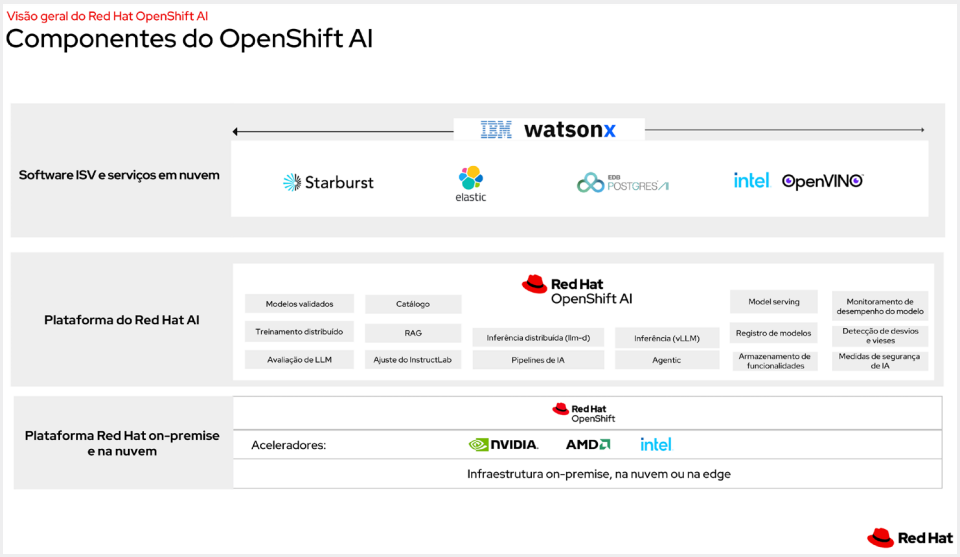

Figura 1. Red Hat OpenShift AI e componentes opcionais de parceiros.

Diversos outros recursos e ferramentas essenciais do Red Hat OpenShift AI oferecem uma base sólida para o desenvolvimento de soluções com IA:

- **Desenvolvimento e personalização de modelos:** cientistas de dados podem conduzir experimentos exploratórios diretamente na interface do JupyterLab, que oferece imagens de notebooks prontas para uso seguras e equipadas com as principais bibliotecas Python. Para projetos de gen IA, o OpenShift AI viabiliza a Geração Aumentada de Recuperação (RAG) e o treinamento distribuído do InstructLab, oferecendo ferramentas de alinhamento para contribuir mais eficientemente com habilidades e conhecimento nos modelos de gen AI.
- **Model serving:** o Red Hat OpenShift AI oferece uma variedade de frameworks usando o KServe como mecanismo principal para model serving, a fim de simplificar a implantação de machine learning preditivo ou modelos fundamentais para ambientes de produção. Para LLMs que exigem escalabilidade máxima, o OpenShift AI oferece serviço em paralelo com runtimes de vLLM. O llm-d fornece um framework para otimizar a inferência de LLMs ao desagregar o pipeline em serviços modulares, que permite escalabilidade automática inteligente e roteamento eficiente de solicitações.
- **Pipelines de IA:** o Red Hat OpenShift AI oferece um componente de pipelines que permite orquestrar tarefas de IA e desenvolver fluxos completos usando uma interface gráfica. As empresas podem organizar processos como preparação de dados, desenvolvimento de modelos e disponibilização para inferência em etapas conectadas.
- **Monitoramento de modelos:** profissionais com perfil de operações podem monitorar o funcionamento e o desempenho dos servidores e dos modelos implantados com o Red Hat OpenShift AI. Os usuários podem ter acesso a visualizações prontas para uso para obter métricas de desempenho e operações ou integrar dados com outros serviços de observabilidade.
- **Cargas de trabalho distribuídas:** permitem que as equipes acelerem o processamento de dados, o treinamento, o ajuste e a disponibilização de modelos. Com esse recurso, é possível priorizar e distribuir a execução de tarefas, além de otimizar o uso dos nós disponíveis. O suporte avançado de GPUs ajuda a atender às demandas de carga de trabalho dos modelos fundamentais.

- **Medidas de segurança de IA, detecção de desvios e vieses:** o Red Hat OpenShift AI oferece ferramentas que ajudam os cientistas de dados e os engenheiros de IA a monitorar se os modelos se mantêm justos e imparciais, tanto durante o treinamento quanto nas implantações em ambientes reais. As medidas de segurança de IA oferecem um framework personalizável que implementa controles de proteção essenciais, ajudando a garantir que os modelos sejam transparentes, justos e confiáveis para uso na produção. Ferramentas de detecção de desvios analisam a distribuição de dados de entrada para modelos de ML em produção, identificando quando os dados em tempo real divergem significativamente dos utilizados no treinamento.
- **Catálogo e registro:** o Red Hat OpenShift AI oferece um catálogo interno de modelos e um selecionado onde os engenheiros de plataforma podem descobrir, comparar e avaliar modelos otimizados de gen IA. Ele também fornece um registro central que ajuda cientistas de dados e engenheiros de IA a compartilhar, modificar, implantar e monitorar artefatos de modelo, metadados e modelos de IA preditiva e generativa.
- **Armazenamento de funcionalidades:** gerencie funcionalidades de dados limpos e bem definidos para modelos de ML, melhorando o desempenho e acelerando os fluxos de trabalho.

Ferramentas para todo o ciclo de vida da IA

O Red Hat OpenShift oferece os serviços e software necessários para as organizações treinarem e implantarem seus modelos e movê-los para produção com sucesso (veja a Figura 2).

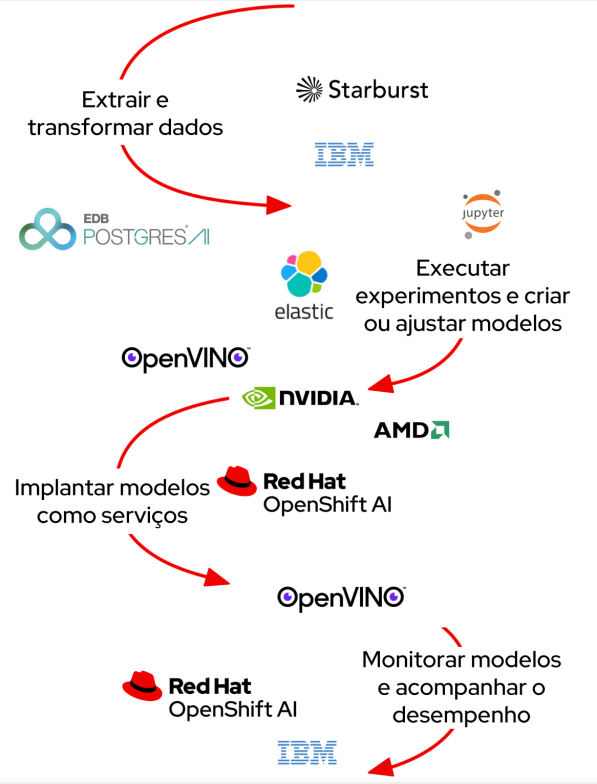

Figura 2. Operacionalização típica de um modelo de ciclo de vida.

Red Hat OpenShift AI

O dashboard do Red Hat OpenShift AI centraliza o acesso a aplicações e à documentação, facilitando a adoção da plataforma. Tutoriais inteligentes para iniciantes estão disponíveis no dashboard e trazem as práticas recomendadas para componentes típicos e software integrados de parceiros. Eles ajudam os cientistas de dados a aprender e começar com mais rapidez. As seções a seguir descrevem as ferramentas de parceiros de tecnologia integradas ao Red Hat OpenShift AI. Algumas ferramentas exigem licença adicional do parceiro.

Starburst

Starburst

O Starburst acelera o processo de analytics ao tornar rápido e fácil para suas equipes aproveitarem seus dados para melhorar o funcionamento do negócio. Oferecido como uma solução autogerenciada ou um serviço totalmente gerenciado, o Starburst democratiza o acesso aos dados, trazendo insights abrangentes. O Starburst foi criado no Trino open source (antigamente chamado de PrestoSQL), o principal mecanismo de linguagem de consulta estruturada (SQL) de processamento paralelo massivo (MPP). Desenvolvido e operado por especialistas em Trino, o Starburst oferece liberdade para consultar conjuntos de dados onde quer que estejam, sem a necessidade de mover informações.

O Starburst se integra ao armazenamento em nuvem escalável e aos serviços de computação oferecidos pelo Red Hat OpenShift, resultando em uma forma estável, segura, eficiente e econômica de consultar todos os dados da empresariais. Os benefícios incluem:

- **Automação:** os operadores do Starburst e do Red Hat OpenShift contam com configuração, ajuste e gerenciamento automáticos de clusters.
- **Alta disponibilidade e redução de escala gradual:** o balanceador de carga do Red Hat OpenShift pode manter serviços como o Trino em um estado de atividade contínua.
- **Escalabilidade elástica:** o Red Hat OpenShift pode expandir automaticamente o nó de trabalho do Trino com base na carga de consulta.

Hewlett Packard Enterprise

HPE Machine Learning Data Management Software

As empresas precisam de soluções de gerenciamento de dados que facilitem desde experimentos em laptops até implantações essenciais. O HPE Machine Learning Data Management Software (anteriormente conhecido como Pachyderm) permite às equipes de ciência de dados desenvolver pipelines de ML em containers orientados por dados, com escalabilidade e linhagem de dados assegurada pelo controle de versão automático. Projetado para resolver problemas de ciência de dados reais, o HPE Machine Learning Data Management Software oferece uma base de dados que permite automatizar e escalar o ciclo de vida de machine learning com reprodutibilidade. Com casos de uso que vão de dados não estruturados a data warehouses, processamento de linguagem natural, extração, transformação e carregamento (ETL) de vídeo e imagem, serviços financeiros e ciências biológicas, o HPE Machine Learning Data Management Software oferece:

- Controle de versão de dados automatizado, o que garante às equipes alto desempenho no acompanhamento das variações nas informações.
- Pipelines em containers e orientados por dados, que aumentam a eficiência do processamento e reduzem os custos.
- Linhagem de dados imutável, com registro fixo das atividades e recursos no ciclo de vida de machine learning.
- Um console que proporciona uma visualização intuitiva do seu gráfico acíclico direcionado (DAG) e ajuda com a depuração e a reprodutibilidade dos dados.
- Suporte ao Jupyter notebook por meio da extensão JupyterLab Mount para criar uma interface point-and-click para os dados versionados.

- Administração empresarial com ferramentas robustas para implantar e administrar o HPE Machine Learning Data Management Software em grande escala e para diferentes equipes.

NVIDIA acelera a implantação de soluções de IA

Conforme aplicações de inteligência artificial e machine learning se tornam cada vez mais essenciais para o sucesso dos negócios, as organizações precisam de plataformas que consigam processar cargas de trabalho complexas, otimizar o uso de hardware e oferecer escalabilidade. Processamento escalável, data analytics e treinamento e inferência de machine learning representam tarefas computacionais com uso intensivo de recursos. O software da NVIDIA possibilita acelerar todos os aspectos da ciência de dados de ponta a ponta usando os recursos de processamento paralelo de GPUs.

Ele aprimora o gerenciamento e o desempenho de GPUs da NVIDIA no ambiente do Red Hat OpenShift, permitindo que as aplicações de IA usem todo o potencial do hardware e software de IA da NVIDIA. Juntos, o NVIDIA NIM e o Red Hat OpenShift AI possibilitam uma melhor alocação de recursos, maior eficiência e execução mais produtiva das cargas de trabalho de IA.

Ferramentas Intel OpenVINO

As ferramentas Intel OpenVINO aceleram o desenvolvimento e a implantação de aplicações de inferência de DL com alto desempenho em plataformas Intel. Com essas ferramentas, você pode adotar, otimizar e ajustar quase todos os modelos de rede neural, além de executar diversas inferências de IA usando o ecossistema de ferramentas de desenvolvimento OpenVINO.

- **Modelar:** os desenvolvedores de software têm a flexibilidade de usar os próprios modelos de DL. Para acelerar o time to market, eles também podem usar modelos pré-treinados e pré-otimizados disponibilizados pela colaboração da Intel com a Hugging Face e suas ferramentas OpenVINO. O OpenVINO é compatível com PyTorch, ONNX, TensorFlow e outros formatos de modelo conhecidos.
- **Otimizar:** as ferramentas OpenVINO oferecem diversas maneiras de converter modelos para maior conveniência e desempenho, ajudando os desenvolvedores de software a obter uma execução de modelos de IA mais rápida e eficiente. Os desenvolvedores podem executar inferência a partir dos formatos PyTorch, ONNX, TensorFlow, TensorFlow Lite, JAX ou PaddlePaddle, sem precisar converter os modelos. A conversão para o formato OpenVINO IR oferece desempenho ideal, que pode ser otimizado ainda mais usando as funcionalidades de compactação e quantização de pesos disponíveis no Neural Network Compression Framework do OpenVINO. As mesmas funcionalidades também reduzem a área de ocupação do armazenamento e runtime.
- **Implantar:** o mecanismo de inferência OpenVINO Runtime é uma interface de programação de aplicações (API) projetada para ser integrada às suas aplicações e acelerar o processo de inferência. Com a abordagem "escreva uma vez, implante em qualquer lugar", você pode executar tarefas de inferência com eficiência em diversos hardwares Intel, como unidades centrais de processamento (CPUs), GPUs, NPU e FPGAs. A biblioteca de extensões OpenVINO GenAI simplifica a implantação de cargas de trabalho de gen IA, em muitos casos reduzindo o código necessário para apenas 3 a 5 linhas. O OpenVINO Model Server oferece diversas funcionalidades para cenários agentic e de model serving, reduzindo ainda mais o esforço do desenvolvimento.

EDB

A EDB Postgres AI é uma plataforma avançada e inteligente projetada processar cargas de trabalho transacionais, analíticas e de IA, oferecendo flexibilidade incomparável para dados on-premise ou em qualquer ambiente de nuvem. Como líder global em soluções empresariais de banco de dados Postgres, a EDB oferece uma plataforma open source empresarial de IA e dados soberanos que acelera a produção de projetos de IA em até três vezes. Com a integração do Red Hat OpenShift AI, a EDB Postgres AI permite que os usuários desenvolvam bases de conhecimento de IA robustas para Geração Aumentada de Recuperação (RAG), unificando dados, modelos e aplicações de IA

em uma plataforma de IA soberana full-stack implantável em qualquer lugar. Essa transformação de dados operacionais essenciais em um ativo pronto para IA pode aumentar a eficiência em até 30% e simplificar o uso de dados privados, inclusive os não estruturados, para consolidar resultados de modelos na base de conhecimento de uma organização.

Elastic

O Elastic Search AI Platform (desenvolvido no ELK Stack[2]) combina a precisão da busca com a inteligência da IA, permitindo que usuários criem protótipos e integrem com LLMs mais rápido, além de usar gen IA para desenvolver aplicações escaláveis e econômicas. O Elastic Search AI Platform permite desenvolver aplicações de geração aumentada de recuperação (RAG) avançadas, resolver proativamente problemas de observabilidade e enfrentar ameaças complexas de segurança O Elasticsearch pode ser implantado onde quer que as aplicações estejam: on-premise, no provedor de nuvem de sua escolha ou em ambientes isolados.

O Elastic se integra a modelos de embedding a partir do ecossistema, incluindo Red Hat OpenShift AI, Hugging Face, Cohere, OpenAI e outros, por meio de uma chamada de API direta. Essa abordagem garante um código limpo para gerenciamento de inferência híbrida em cargas de trabalho de RAG, com funcionalidades que incluem:

- Fragmentação, conectores e rastreadores web para ingestão de diversos conjuntos de dados na sua camada de pesquisa.
- Pesquisa semântica com Elastic Learned Sparse EncodeR (ELSER), modelo de ML integrado, e o modelo de embedding E5, possibilitando a pesquisa vetorial multilíngue.
- Segurança em nível de documento e de campo, implementando permissões e direitos que mapeiam o controle de acesso baseado em função (RBAC) da sua empresa.

Com o Elastic Search AI Platform, você faz parte de uma comunidade global de desenvolvedores, onde colaboração e inspiração estão sempre ao seu alcance. Encontre a comunidade da Elastic no Slack, em nossos fóruns de discussão ou nas redes sociais.

Conclusão

Com o Red Hat OpenShift AI, as organizações podem experimentar, colaborar e acelerar a jornada de aplicações com tecnologia de IA. Cientistas de dados e engenheiros de IA têm a flexibilidade de usar o Red Hat OpenShift AI para desenvolver e implantar modelos na nuvem híbrida. As equipes de operações de TI e os engenheiros de plataforma devem contar com os recursos de MLOps e GenAIOps para acelerar a implantação de modelos em produção. Os serviços self-service para desenvolvedores, engenheiros de IA e cientistas de dados, inclusive o acesso a GPUs, impulsiona a inovação em plataformas de aplicações que já estão em atividade e conquistaram a total confiança das empresas. O Red Hat OpenShift AI continua a disponibilizar uma plataforma abrangente, confiável e consistente, oferecendo diferenciais exclusivos em inferência eficiente, agentic AI e operações escaláveis de nuvem híbrida, com o apoio de um ecossistema de parceiros robusto.

Mais informações

Veja por onde começar na página do Red Hat OpenShift AI.

Sobre a Red Hat

A Red Hat ajuda os clientes a definir padrões entre diferentes ambientes e a desenvolver aplicações nativas em nuvem, além de integrar, automatizar, proteger e gerenciar ambientes complexos com serviços de consultoria, treinamento e suporte premiados.

2 *O ELK Stack é composto por Elasticsearch, Kibana, Beats e Logstash.*

facebook.com/redhatinc
@redhatbr
linkedin.com/company/red-hat-brasil

AMÉRICA LATINA
+54 11 4329 7300
latammktg@redhat.com

BRASIL
+55 11 3629 6000
marketing-br@redhat.com

br.redhat.com
#2876428_1125