ORACLE

7 Perguntas sobre Infraestrutura de Dados para o Sucesso com a IA

Ao abordar esses pontos-chave, os líderes podem simplificar a adoção da IA, maximizar o ROI e impulsionar a inovação nos negócios

Índice

Seus dados estão prontos para a IA?

Por Jeffrey Erickson
Redator sênior

Imagine os mesmos grandes modelos de linguagem, ou LLMs, que podem explicar a física quântica e resumir romances franceses, adquirindo profundo conhecimento sobre os dados operacionais e a base de conhecimento específicos da sua empresa. Agora, recuperar, combinar e contextualizar esses dados deixa de ser uma tarefa na fila de um analista e passa a ser uma conversa em linguagem natural entre um usuário corporativo e um agente de IA que pode revelar insights e, então, tomar medidas. Os recursos de TI são liberados e sua organização passa a ser mais orientada por dados.

Parece bom, não é? Por isso, não é de se admirar que as empresas de tecnologia mais ambiciosas, talentosas e bem financiadas do mundo — e as startups que concorrem com elas — estejam correndo para colocar as mãos no potencial da IA generativa. Mesmo em meio a todo esse entusiasmo, é difícil exagerar o quanto os modelos de IA generativa podem multiplicar o valor dos seus dados, alterando fundamentalmente a forma como as pessoas em toda a sua organização os acessam e utilizam.

A base para tudo isso é uma infraestrutura voltada para as necessidades da IA generativa — com a combinação certa de ferramentas personalizadas, LLMs fundamentais, técnicas e sistemas de computação para fornecer respostas rápidas e oferecer suporte a casos de uso tão variados quanto detecção de anomalias e identificação de objetos. O sucesso geralmente depende de ter potência suficiente para lidar com operações complexas de IA — pense em GPUs interconectadas por uma rede de clusters capaz de atingir desempenho ultraelevado e latência de microssegundos. Também requer bancos de dados vetoriais; uma maneira de aproveitar os agentes de IA para retrieval-augmented generation, ou RAG, que combina suas diversas bases de conhecimento empresarial com LLMs; e interfaces de agentes de IA fáceis de usar.

> É difícil exagerar o quanto os modelos de GenAI podem multiplicar o valor dos seus dados.

O que é Agentic AI?

Agentic AI, ou IA autônoma, refere-se à inteligência artificial capaz de compreender e responder a informações, assim como buscar ativamente objetivos.

Principais características da Agentic AI

Comportamento proativo
A IA pode iniciar ações em vez de simplesmente reagir a comandos externos.

Adaptação
A IA pode aprender com a experiência, aceitar feedback e ajustar-se para melhor atingir seus objetivos.

Comportamento orientado para objetivos
A IA tem objetivos específicos que procura alcançar e pode traçar os passos para chegar lá.

Autonomia
A IA pode tomar decisões e agir de forma independente, dentro de determinados parâmetros.

A promessa da IA generativa é substancial. O mesmo se aplica aos investimentos em infraestrutura necessários para chegar lá.

Você provavelmente já ouviu falar de estatísticas impressionantes sobre a expansão dos data centers para dar suporte à IA. Grande parte desses gastos tem sido feita por provedores de nuvem em hiperescala que buscam oferecer experiências de IA excepcionais, com algumas organizações visionárias também preparando suas infraestruturas de dados para dar suporte ao que está por vir.

Então, o que você pode fazer para se preparar agora? Abaixo estão sete perguntas que muitos arquitetos de TI estão fazendo enquanto trabalham para concretizar a promessa da IA agora e no futuro.

1 Como nos beneficiaremos de uma série de casos de uso da IA generativa, incluindo alguns que ninguém pensou ainda?

Alguns projetos de IA — pense em marketing hiperpersonalizado, chatbots de atendimento ao cliente realmente bons, auxiliares de codificação que aumentam a produtividade — muitas vezes justificam o custo da infraestrutura de suporte. E a lista de casos de uso potenciais cresce a cada dia, porque, francamente, se você pode imaginar, provavelmente pode fazer. Isso se deve em grande parte aos LLMs avançados que agora estão incorporados em aplicativos empresariais, como ERP, HCM e SCM, e amplamente utilizados em áreas de pesquisa intensiva, como robótica, saúde, genômica e engenharia aeroespacial. Em outras palavras, em quase todos os lugares.

O segredo é complementar esses LLMs para que sua IA se torne a principal especialista em seu negócio. Ao fornecer acesso a bibliotecas de dados históricos, operacionais, financeiros e de documentação, você possibilita que a IA generativa atue como um multiplicador de força de várias maneiras, para muitas pessoas em toda a sua empresa.

E como seria isso? Pode significar que os executivos possam "conversar" com os dados, aprofundar-se neles e obter novas informações de inteligência empresarial na hora. Pode significar que os agentes de IA se tornem parceiros no desenvolvimento de software e em sessões de brainstorming de P&D, e até mesmo criadores extremamente rápidos de protótipos e maquetes. Pode significar que os vendedores possam gerenciar mais leads com eficiência, pois a prospecção, as comunicações e os detalhes do processo alcançam novos níveis de automação, graças à IA.

Esses são apenas alguns exemplos. Uma sessão de brainstorming em que suas equipes de produto, finanças, RH e jurídico compartilhem suas ideias provavelmente resultará em casos de uso criativos de IA generativa e um ponto de partida sólido para um plano de infraestrutura de dados.

Quer mais ideias de empresas do mesmo setor? Confira uma série de casos de uso reais tão variados quanto análise genética e transmissão esportiva.

Ajuste fino e RAG: duas maneiras de ajudar a IA a “entender” seus dados

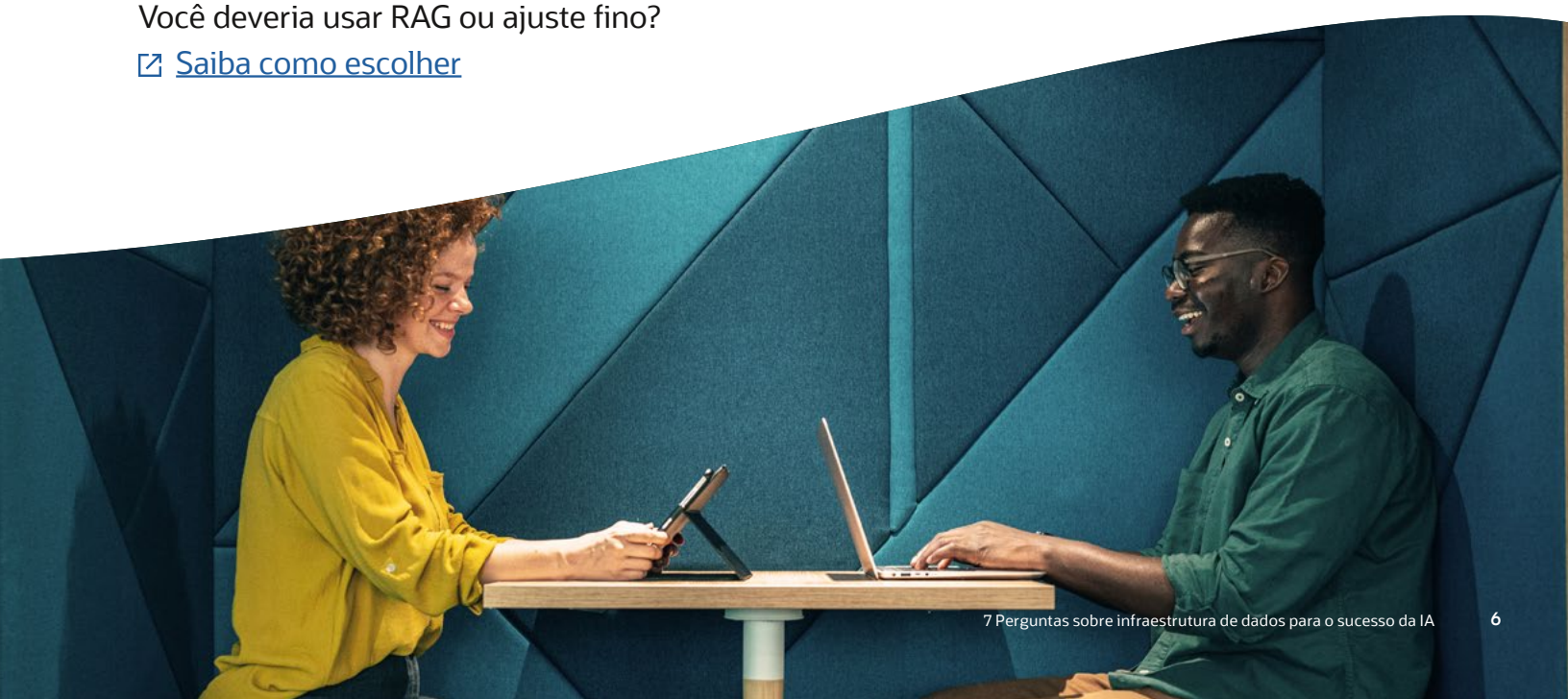
Tanto o ajuste fino quanto o RAG ajudam os modelos de IA generativa a fornecer respostas mais relevantes para o contexto e adaptadas à sua organização. Veja como eles se diferem.

Ajuste fino significa pegar um modelo de uso geral, como o Command da Cohere ou o Llama 3 da Meta, e submetê-lo a rodadas adicionais de treinamento em um conjunto de dados menor e específico do domínio. O ajuste ajuda o modelo a ter um melhor desempenho em tarefas específicas, pois ele foi adaptado às nuances e à terminologia de um domínio, como codificação, finanças ou saúde. As desvantagens são o custo e a necessidade de uma infraestrutura baseada em GPU para dar suporte ao processo.

RAG é uma estrutura arquitetônica que ajuda os modelos de IA de uso geral a fornecer resultados úteis para organizações específicas, fornecendo dados selecionados e relevantes ao LLM à medida que ele formula respostas às consultas. O resultado é um sistema de IA que combina a fluência linguística de um LLM com insights de seus dados internos para fornecer respostas direcionadas e mais adequadas ao contexto. Ao contrário do ajuste fino do modelo de IA, o RAG funciona sem modificar o modelo subjacente. O LLM também “esquecerá” os dados fornecidos a ele após a conclusão das consultas, reduzindo uma fonte potencial de vazamento de dados.

Você deveria usar RAG ou ajuste fino?

Saiba como escolher



2 Como disponibilizaremos todos os nossos dados não estruturados e semiestruturados para a IA?

O processamento de dados não estruturados e semiestruturados para uso pela IA é um processo de várias etapas que envolve coleta e ingestão de dados, armazenamento e processamento — limpeza, extração de recursos, normalização, segmentação e, possivelmente, outras etapas. Só então as imagens, documentos e arquivos de áudio e vídeo em seu data lake estarão preparados para ajuste fino, vetorização e RAG.

Dados para ajuste fino: O ajuste fino de modelos de IA generativa para tarefas específicas pode exigir a apresentação de novos dados específicos da disciplina. Por exemplo, se você estiver ajustando um modelo de IA generativa para produzir relatórios médicos, precisará de um conjunto de dados de relatórios médicos relevantes e de alta qualidade para ajudar o modelo a aprender a terminologia e o contexto adequados.

Dados para resultados contínuos de IA: Um sistema de gerenciamento de dados pronto para IA precisa fornecer aos seus LLMs acesso a uma ampla variedade de armazenamentos de dados, como documentos JSON e dados relacionais e semiestruturados. Ele precisa aplicar eficientemente incorporações vetoriais aos dados, armazenar as incorporações como vetores em um banco de dados e consultar o banco de dados para permitir uma pesquisa vetorial eficiente e o processamento de linguagem natural. Isso pode exigir coleta contínua de dados e um pipeline de integração que também inclua uma arquitetura RAG para ajudar a melhorar a precisão e a relevância dos resultados do seu modelo de IA.

Como fica quando esses sistemas estão em funcionamento? Uma equipe de vendas, por exemplo, poderia salvar o áudio de todas as chamadas e deixar que a IA analisasse o sentimento do cliente enquanto cria resumos. Chega de rabiscos indecifráveis enquanto os vendedores tentam interagir e tomar notas ao mesmo tempo. Chega de detalhes esquecidos em uma solicitação de cotação para o próximo pedido do cliente. É a mesma atenção aos detalhes, mas com muito mais facilidade, precisão e eficiência.

Uma chave para o sucesso: Um banco de dados multimodal unificado que permite habilitar todos esses processos sem mover, ressincronizar e proteger novamente os dados em sistemas especializados. Esse modelo de dados unificado pode acelerar as operações em ordens de magnitude, permitindo pesquisas em todos os tipos de dados, incluindo vetores de IA, sem que a TI precise integrar silos distintos. Isso pode levar a resultados mais ricos, pois os modelos de IA consideram relações sutis entre muitas informações aparentemente não relacionadas, mas críticas.

3 Como protegeremos os dados confidenciais para a IA, de preferência com uma única estratégia de governança e controle de acesso?

Manter a segurança e a privacidade dos dados fornecidos às suas ferramentas de IA generativa é um princípio essencial do projeto de infraestrutura, tanto no ajuste fino do modelo e no RAG quanto durante o uso diário.

Dependendo das suas necessidades, os dados para sua plataforma de IA generativa podem ser anonimizados, criptografados ou mascarados. Os dados destinados ao uso na geração de resultados de IA também precisam ser rigidamente controlados, levando em consideração a função do solicitante e empregando processos de identificação multifatorial. Algumas organizações optam por executar inferências em suas próprias cópias de modelos de IA hospedados em infraestruturas dedicadas, ou mesmo em data centers locais, para proteger ainda mais os dados privados.

Você pode procurar ajuda junto aos seus fornecedores de tecnologia e órgãos normativos. Por exemplo, a Oracle está entre as mais de 280 organizações que colaboram com o National Institute of Standards and Technology (NIST) em seu Artitificial Intelligence Safety Institute Consortium[1]. O objetivo do AISIC é "desenvolver diretrizes e normas baseadas na ciência e empiricamente comprovadas para a medição e política de IA, estabelecendo as bases para a segurança da IA em todo o mundo". Os membros trabalharão para desenvolver diretrizes, ferramentas, métodos, protocolos e melhores práticas para ajudar a comunidade a desenvolver e implantar a IA com segurança.

Existem outros esforços semelhantes centrados na IA também.

[1] *NIST Artificial Intelligence Safety Institute Consortium*

Arquitetura de IA generativa: banco de dados convergentes versus vários bancos de dados de uso único

Com um banco de dados multimodal, normalmente não é necessário realizar ETL e movimentação de dados extensivos para aproveitar as vantagens da IA generativa. Em vez disso, você pode oferecer suporte a diversos projetos e cargas de trabalho de IA usando uma única plataforma de dados, reduzindo significativamente a complexidade e a sobrecarga de gerenciamento, ao mesmo tempo em que minimiza o risco envolvido na movimentação de dados. Um banco de dados multimodal não significa que os dados devem estar em um único formato ou modelo; significa apenas que você pode escolher sua arquitetura de dados para cada tipo de aplicativo.

4 Como as equipes colaborarão em nossa estratégia de IA generativa, mantendo os padrões em vigor?

Nas mãos certas, um modelo pronto para uso pode se tornar uma potência de IA personalizada que somente sua organização poderia criar. As pessoas querem aproveitar as vantagens da IA. Oferecer a elas as ferramentas necessárias para isso aumenta muito as chances de que os fluxos de trabalho de IA generativa se tornem parte da estratégia da sua empresa, e não projetos pontuais e paralelos de TI.

Uma maneira de maximizar seus esforços é um centro de excelência, que oferece transparência e promove a consistência entre os departamentos. Um centro de excelência consolida as melhores práticas, ferramentas e técnicas relacionadas à IA generativa e lidera o caminho para usá-las na resolução de problemas de negócios.

Aprenda a criar um centro de excelência em IA na sua empresa

Uma tecnologia essencial para essa colaboração será uma plataforma de ciência de dados que incentive a colaboração nas etapas técnicas de ajuste fino e implementação de modelos. Seu provedor de nuvem em hiperescala terá uma plataforma de IA generativa onde os cientistas de dados poderão trabalhar e aprimorar modelos básicos poderosos da Cohere, Meta e outros fornecedores. Essas plataformas oferecem aos cientistas de dados e equipes de TI um local para armazenar, trocar e potencialmente reutilizar modelos, conjuntos de dados e rótulos de dados entre serviços.

Os hiperescaladores também oferecem catálogos de modelos e fornecem LLMs aperfeiçoados com os dados e a infraestrutura de computação necessários para executá-los. Agora, as unidades de negócios podem aproveitar os sucessos anteriores da IA em outras partes da organização.

Com um centro de excelência e uma plataforma líder em ciência de dados, você pode promover uma cultura de aprendizado e melhoria contínuos. Provedores como a Oracle oferecem serviços totalmente gerenciados, como o Oracle Cloud Infrastructure (OCI) Generative AI, para integrar perfeitamente LLMs em uma ampla gama de casos de uso. Você também pode criar modelos personalizados, ajustando os modelos básicos com seu próprio conjunto de dados.

5 Como podemos combinar os pontos fortes de nossos provedores de nuvem para maximizar a disponibilidade de dados para IA?

Talvez sua organização esteja explorando o Google Gemini ou o Copilot da Microsoft. Ou você pode decidir hospedar modelos de código aberto na AWS ou modelos básicos da Cohere na OCI. Agora, novos acordos entre a Oracle e outros hiperescaladores abrem um amplo mundo de possibilidades para empresas cujas estratégias de governança de dados corporativos são construídas em torno da tecnologia do Oracle Database, seja on-premises ou na nuvem. Você pode adotar qualquer um desses modelos com baixa latência e sobrecarga de gerenciamento, graças às relações multicloud inovadoras entre a Oracle e o Azure, o Google Cloud e a AWS, que possibilitam usar os serviços do Oracle Database não apenas na OCI, mas também nos data centers de cada provedor.

O sucesso com esses casos de uso de IA generativa ainda não concebidos pode ser mais provável quando você pode usar exatamente os modelos de IA que deseja, executando-os onde quiser, sem comprometer o acesso fácil e seguro aos seus dados. As equipes podem aproveitar os melhores serviços para tarefas específicas, mantendo a segurança e a resiliência e reduzindo custos.

6 Como vamos adquirir, gerenciar e custear os sistemas necessários para o ajuste fino e a inferência?

O ajuste fino e a implementação da IA generativa são tarefas que exigem muito poder de computação — cada interação pode envolver cálculos complexos em conjuntos de dados massivos, nos quais um LLM contendo bilhões de parâmetros utiliza sistemas de computação especializados para alimentar sua manipulação e análise de informações. Quanto mais complexo for o modelo, mais poder computacional será necessário para processar os dados e fornecer resultados relevantes. Você precisará de um plano para manter baixos os custos de computação tanto para o ajuste fino do modelo quanto para a inferência, ao mesmo tempo em que fornece resultados de alta qualidade.

Em vez de arcar com os custos e desenvolver o conhecimento necessário para criar um sistema interno, muitas organizações optam por plataformas populares de ciência de dados, como as disponibilizadas pelos principais provedores de nuvem, para tornar os conjuntos de dados limpos e relevantes para uma tarefa.

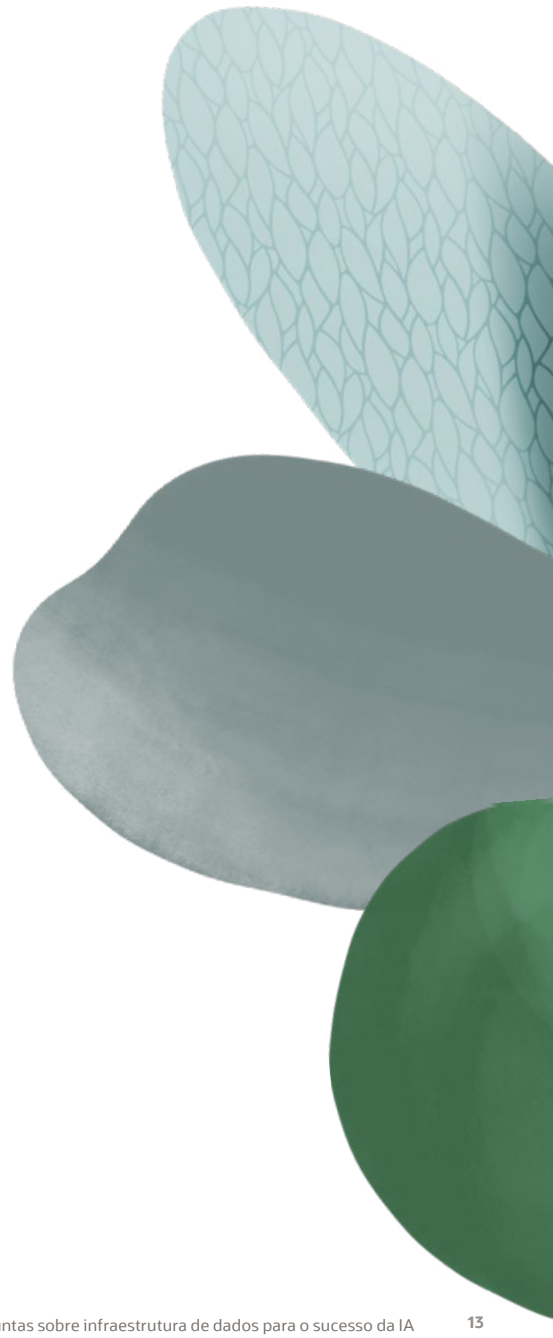
Em seguida, entenda os serviços de infraestrutura de IA dos seus provedores de nuvem. Muitas vezes, você pode obter acesso imediato ou integração com ferramentas populares de código aberto, bem como modelos básicos de IA generativa e outras tecnologias necessárias.

Tenha em mente que a comunidade de código aberto e os hiperescaladores estão empenhados em ajudar a reduzir os custos da IA generativa. É na nuvem que você encontrará as poderosas GPUs e redes de cluster necessárias para acelerar as cargas de trabalho de IA, bem como recursos como autoescalonamento, instâncias spot, instâncias reservadas, agendamento otimizado de tarefas, distribuição eficiente da carga de trabalho, multilocação e otimização de modelos. Essas técnicas maximizam a utilização de recursos — ao combinar o poder de computação com as demandas exatas da carga de trabalho em determinado momento, as empresas pagam apenas pelo que usam.

O ajuste fino também está sendo, digamos, ajustado

Um método conhecido como T-Few pode reduzir a duração do treinamento e os recursos computacionais necessários em comparação com os métodos convencionais de ajuste fino, mantendo ainda assim uma alta precisão.

Por fim, você vai querer comprovar que seus modelos e infraestrutura estão funcionando com eficiência máxima. Você pode monitorar e otimizar o processo de inferência por meio do serviço de IA generativa do seu provedor de nuvem ou por meio de estruturas eficientes de serviço de modelos, como TensorFlow Serving ou TorchServe.



Temos apoio executivo para o nosso plano de IA? Quais grupos trabalharão conosco?

Sua estratégia de IA provavelmente começará com seus sistemas empresariais atuais. Os fornecedores de CRM, HCM, ERP e outras aplicações essenciais estão incorporando recursos de IA generativa e agentes de IA em fluxos de trabalho comuns. Pense no reconhecimento inteligente de documentos para ajudar a processar faturas de fornecedores mais rapidamente e com muito menos trabalho manual, narrativas generativas para trazer uma compreensão mais profunda dos relatórios e análises e uma ampla gama de automações inteligentes em áreas como análise de risco e desempenho. Isso indicará o caminho, mas você precisará de uma estratégia mais ampla para incorporar a IA aos seus fluxos de trabalho exclusivos. Isso pode envolver o centro de excelência discutido anteriormente.

A introdução de serviços de IA generativa e agentes de IA nas operações da sua empresa exigirá o apoio e os recursos de toda a organização, incluindo executivos, chefes de departamento, equipes de TI, jurídicas e de compliance e, é claro, as pessoas que usarão a nova tecnologia. Ao trabalhar com as partes interessadas em cada unidade funcional para desenvolver o caso de negócio, com foco em ganhos de produtividade e vantagens competitivas, você pode desenvolver uma abordagem coordenada.

Uma vez que você tenha decidido quais fluxos de trabalho de IA usar inicialmente, é hora de envolver suas equipes mais amplas de TI e ciência de dados para entender que tipo de arquitetura você precisará. Algumas perguntas a serem feitas incluem: Onde os LLMs ficarão? Quem pagará por eles? Como os dados fluirão — precisamos de novos bancos de dados vetoriais ou arquiteturas RAG? Os LLMs precisam ser ajustados para um departamento ou tarefa específica? De onde virão o armazenamento de dados e o poder de computação? Nosso projeto de rede e outras decisões arquitetônicas nos proporcionarão a baixa latência e o alto rendimento de que precisamos para a inferência contínua de IA?

Prepare uma proposta detalhada com cronogramas, etapas importantes e requisitos de recursos, além de um resumo dos riscos, incluindo privacidade de dados, segurança e compliance, e considerações de custo. Considere como seus provedores confiáveis de serviços em nuvem podem ajudar.

> Tenha em mente que nenhum modelo de IA será eficaz sem um fluxo constante de dados limpos e bem organizados.

Portanto, considere a implementação da IA como uma extensão da sua estratégia geral de dados. Para promover a adesão da diretoria, as equipes vão querer confirmar que seus planos de IA estejam alinhados com a estratégia geral de negócios e o plano de governança de dados da organização, sem trazer complexidade desnecessária com a transferência de dados e a adição de recursos de gerenciamento.

Soluções da Oracle

Quando sua infraestrutura de dados precisa oferecer ajuste fino e inferência rápidos e econômicos dos modelos de IA generativa mais avançados do mundo, a Oracle pode ajudar.

Oracle Cloud Infrastructure

Oracle AI Infrastructure

A OCI oferece serviços completos de IA e inovações de IA generativa de última geração, tudo em uma infraestrutura de IA de primeira linha. Você pode escolher modelos pré-construídos ou aproveitar um serviço de IA generativa que permite escolher entre LLMs de código aberto ou proprietários e, em seguida, ajustar os modelos e aprimorá-los com seus próprios dados empresariais. Além disso, você tem acesso ao sistema de gerenciamento de banco de dados líder do setor.

Saiba mais

Oracle Database 23ai

O Oracle Database 23ai é um banco de dados multimodal para transações e análises que oferece um banco de dados vetorial integrado, além de JSON, dados relacionais, gráficos e espaciais, bem como aprendizado de máquina no banco de dados. Você também pode fazer processamento vetorial e executar o Oracle Database em plataformas otimizadas exclusivamente, disponíveis apenas na OCI e como serviços OCI nos data centers de nossos parceiros multicloud.

Saiba mais

Serviços de OCI Data Science e de IA generativa

O OCI Data Science é um serviço em nuvem que permite aos cientistas de dados criar, treinar, implantar e gerenciar modelos de machine learning (ML) de forma colaborativa com suas estruturas de código aberto favoritas, ou eles podem aproveitar o ML no banco de dados. O serviço de IA generativa da OCI é um serviço totalmente gerenciado para integrar os modelos de linguagem de sua escolha em uma ampla gama de casos de uso.

Saiba mais

Na OCI, você terá um ajuste fino, inferência e processamento em lote mais rápidos e econômicos. A OCI oferece a escala e o desempenho necessários para executar grandes cargas de trabalho de IA com mais rapidez, sem custos excessivos de computação.

Como a Oracle pode te ajudar

Relacionamentos multicloud exclusivos podem ajudar a facilitar o compartilhamento de dados.

Organizações em todo o mundo utilizam os serviços Oracle Database executados na OCI para criar e executar rapidamente aplicativos no Oracle Exadata Database Service e aproveitar os recursos do Oracle Autonomous Database. As relações multicloud permitem que as organizações que executam aplicativos em outras nuvens de hiperescala obtenham os mesmos benefícios do Oracle Database com a simplicidade, segurança e baixa latência de um único ambiente operacional.

Crie e execute aplicações rapidamente na AWS, usando o Exadata Database Service e o Autonomous Database com serviços como o Amazon Bedrock; no Azure, usando o Exadata Database Service e o Autonomous Database com serviços como o Azure OpenAI; e no Google Cloud, usando o Exadata Database Service e o Autonomous Database com serviços como o Vertex AI e os modelos básicos Gemini do Google Cloud.

Saiba mais

Fale com a gente

Ligue 0800-891-4433 ou +55 11 5189 3137 (Brasil) ou acesse oracle.com/br

Fora da América do Norte, encontre o seu escritório local em oracle.com/br/corporate/contact