



ALÉM DA LLM: COMO PREPARAR SUA ARQUITETURA PARA AGENTES DE IA

RAG, agentes e arquiteturas para IA corporativa

Guia executivo para transformar modelos de linguagem em aplicações inteligentes, conectadas aos dados, sistemas e processos do negócio com a GotoBiz e AWS.

MENSAGEM CENTRAL

A LLM é apenas uma camada da solução. O valor corporativo surge quando contexto, conhecimento, ferramentas, segurança e operação trabalham de forma integrada.

PARA QUEM	O QUE ENTREGA	PRÓXIMO PASSO
CIOs, CTOs e líderes de aplicações, arquitetura, dados, TI e inovação	Visão prática de RAG, agentes, integrações, arquitetura e caminho para produção	Webinar gratuito sobre modernização de aplicações com AWS e IA Generativa

Agenda da jornada

01 MATERIAL ENVIADO Fundação de dados para IA	02 WEBINAR CONCLUÍDO GenAI, RAG e agentes
03 ♀ ESTE GUIA Modernização de aplicações	04 PRÓXIMO ENCONTRO AWS Transform, Kiro e Bedrock



1. A fundação está pronta: Agora é preciso preparar as aplicações.

Nosso primeiro material desta jornada mostrou que a IA depende de dados confiáveis, organizados, governados e acessíveis. Essa fundação continua indispensável. Porém, para que a IA deixe de ser apenas uma interface de perguntas e respostas, ela precisa ser incorporada às aplicações, às integrações e aos processos da empresa.

Aplicações legadas nem sempre foram desenhadas para consumir modelos generativos, recuperar conhecimento em tempo real, acionar APIs com segurança, sustentar agentes com memória e identidade ou registrar decisões e ações. É por isso que a modernização passa a ser a segunda frente da jornada: transformar a base de dados e conhecimento em experiências inteligentes que possam operar com segurança e escala.

PONTO DE PARTIDA

A pergunta deixa de ser apenas “qual modelo usar?” e passa a ser: qual arquitetura permitirá que a IA compreenda contexto, acesse fontes confiáveis, interaja com sistemas e gere valor de forma sustentável?

Da informação organizada à inteligência incorporada

Quando documentos, regras e históricos já estão organizados, o próximo passo é torná-los utilizáveis pelas aplicações. O [KnowledgeOps AI](#) ajuda a transformar documentos, tickets, dados operacionais e histórico real em conhecimento estruturado, governado e rastreável. Para levar esse contexto às aplicações, os projetos de [Consulting & Professional Services](#) podem apoiar o desenho da arquitetura, das integrações e dos casos de uso, conectando estratégia, implementação e transição para a operação.

LLM isolada x aplicação inteligente

LLM isolada	Aplicação inteligente
Responde com conhecimento geral	Utiliza dados e contexto da empresa
Depende principalmente do prompt	Recupera fontes, aplica regras e preserva contexto
Não executa ações	Interage com APIs, ferramentas e sistemas
Baixa rastreabilidade	Registra fontes, ações e decisões
Piloto pontual	Serviço integrado à operação
Avaliação subjetiva	Métricas de qualidade, custo e risco



2. A anatomia de uma aplicação inteligente

Uma aplicação corporativa de IA não é formada apenas pelo modelo. Ela combina diferentes camadas que conectam experiência, conhecimento, ações e operação. Quanto mais crítica for a tarefa, maior deve ser a clareza sobre responsabilidades, permissões, limites e mecanismos de supervisão.

1. Experiência

Usuários, canais, aplicações, copilotos e interfaces conversacionais.

2. Orquestração Inteligente

Prompts, agentes, memória, planejamento, regras e validação humana.

3. Conhecimento e contexto

RAG, Knowledge Bases, busca semântica, bancos vetoriais, histórico e grafos.

4. Ações e integrações

APIs, sistemas corporativos, automações, workflows, function calling e ferramentas.

5. Fundação tecnológica

Modelos, dados, aplicações, infraestrutura e arquitetura cloud.

CAMADA TRANSVERSAL

Segurança, identidade, governança, observabilidade, avaliação, custos e compliance atravessam todas as camadas. O modelo produz linguagem; a arquitetura transforma linguagem em contexto, decisão e ação.

Na camada de conhecimento, o [KnowledgeOps AI](#) contribui para organizar e disponibilizar fontes corporativas de forma governada. Já a integração entre modelos, APIs, aplicações e processos pode ser conduzida com apoio de [Consulting & Professional Services](#), reduzindo o risco de iniciativas desconectadas, difíceis de manter ou sem caminho claro para produção.



3. Oito abordagens para levar GenAI ao ambiente corporativo

As abordagens abaixo não competem necessariamente entre si. Em muitos casos, elas se combinam para resolver diferentes partes do mesmo problema. A decisão deve considerar contexto, criticidade, atualização das informações, necessidade de ação, segurança e custo operacional.

Abordagem	Necessidade	O que resolve	Ponto de atenção
RAG	Conteúdo atualizado	Recupera contexto antes da resposta.	Qualidade da recuperação.
Knowledge Bases	Organizar fontes	Estrutura documentos e dados para consulta.	Atualização e permissões.
Bases vetoriais	Buscar por significado	Encontra conteúdo semanticamente relacionado.	Chunking e embeddings.
Prompt engineering	Padronizar respostas	Define instruções, contexto, formato e critérios.	Manutenção e testes.
Fine-tuning	Especializar comportamento	Ajusta padrões recorrentes do modelo.	Custo e necessidade real.
Tool use	Interagir com sistemas	Permite chamar APIs, funções e aplicações.	Autorização e validação.
Agentes	Executar fluxos	Planeja etapas, utiliza ferramentas e memória.	Limites, supervisão e logs.
GraphRAG	Entender relações	Explora conexões entre entidades e contextos.	Modelagem e manutenção.

Não é uma escolha única

Um agente de suporte pode usar RAG para consultar runbooks, uma Knowledge Base para organizar documentos, tool use para abrir chamados, memória para manter contexto e observabilidade para registrar todo o fluxo. O papel da arquitetura é combinar apenas os componentes que o caso exige, evitando tanto a simplicidade excessiva quanto a complexidade sem valor.

Nessa etapa, [Consulting & Professional Services](#) pode apoiar a seleção das abordagens, o desenho técnico e a implementação dos primeiros casos, conectando as decisões de arquitetura aos objetivos, restrições e capacidades reais da organização.



4. Do problema à arquitetura: uma matriz de decisão

A escolha da arquitetura deve partir do resultado esperado. Casos que exigem apenas consulta a conteúdo podem começar com RAG. Cenários que precisam interagir com sistemas, tomar decisões em etapas ou executar ações controladas tendem a demandar ferramentas, agentes e mecanismos adicionais de governança.

Cenário	RAG	Fine-tuning	Tool use	Agente	GraphRAG
Consulta a políticas e manuais	Alto	Baixo	Baixo	Médio	Baixo
Atendimento com histórico	Alto	Médio	Alto	Alto	Médio
Classificação especializada	Médio	Alto	Baixo	Baixo	Baixo
Automação de processos	Médio	Baixo	Alto	Alto	Baixo
Análise de relações e riscos	Alto	Baixo	Médio	Médio	Alto
Assistente de desenvolvimento	Médio	Médio	Alto	Alto	Baixo
Diagnóstico operacional	Alto	Baixo	Alto	Alto	Médio

Perguntas para orientar a decisão

- A resposta depende de informações atualizadas?
- O sistema precisa apenas responder ou também executar ações?
- Existem APIs disponíveis e governadas?
- A tarefa possui etapas e critérios bem definidos?
- O resultado exige aprovação humana?
- A resposta precisa citar ou rastrear as fontes utilizadas?
- Qual é o custo aceitável por interação e por processo concluído?

Antes de incorporar GenAI a uma aplicação, o [Well-Architected Review by GotoBiz](#) pode ajudar a avaliar riscos, dependências, segurança, confiabilidade, desempenho, custos e prioridades de evolução da arquitetura. Essa leitura reduz a chance de modernizar sem compreender o ponto de partida ou de levar limitações antigas para uma nova camada de IA.



5. Onde a arquitetura se transforma em valor

Casos de uso ganham consistência quando são descritos como uma jornada completa: entrada, contexto, arquitetura, ação e resultado. Essa visão evita iniciativas centradas apenas na interface e ajuda a identificar quais dados, integrações, controles e capacidades operacionais serão necessários.

A seguir alguns casos de uso que ilustram esse cenário:

Atendimento ao cliente

Entrada: Solicitação do cliente | **Contexto:** Contratos, políticas, histórico e produtos

Arquitetura: RAG + Knowledge Base + CRM | **Ação:** Responder, registrar interação ou abrir demanda

Valor esperado: Menor tempo de resposta e maior consistência

TI e operações

Entrada: Alerta ou chamado | **Contexto:** Logs, CMDB, incidentes e runbooks

Arquitetura: RAG + agente + ferramentas | **Ação:** Recomendar correção ou executar ação controlada

Valor esperado: Diagnóstico mais rápido e padronizado

Jurídico

Entrada: Contrato ou consulta | **Contexto:** Cláusulas, políticas e pareceres

Arquitetura: RAG + GraphRAG + validação humana | **Ação:** Identificar riscos e comparar versões

Valor esperado: Agilidade com rastreabilidade

Desenvolvimento

Entrada: Requisito, incidente ou código | **Contexto:** Repositórios, arquitetura e padrões

Arquitetura: Agente + ferramentas de desenvolvimento | **Ação:** Analisar, documentar, testar ou propor evolução

Valor esperado: Redução de débito técnico e ciclo mais rápido

Nesses casos de uso, o [KnowledgeOps AI](#) pode organizar contratos, políticas, runbooks, tickets e histórico em uma base pesquisável e rastreável. Nos cenários que exigem integração com CRM, ITSM, repositórios ou APIs, [Consulting & Professional Services](#) pode apoiar a construção e a modernização das aplicações. Quando esses fluxos passam a executar automações recorrentes, o [Automation Care](#) contribui para sustentação, monitoramento, governança e evolução contínua.



6. De piloto à produção: o que não pode faltar

Uma demonstração pode funcionar com poucos usuários, dados selecionados e supervisão constante. Produção exige uma estrutura capaz de sustentar qualidade, segurança, desempenho, custos e evolução ao longo do tempo.

Valor	Dados e conhecimento	Arquitetura	Segurança e governança	Operação
• Caso de uso prioritário • Usuário e processo definidos • Resultado mensurável • Critérios de sucesso	• Fontes confiáveis • Atualização e qualidade • Permissões • Rastreabilidade	• Integrações • Escalabilidade e resiliência • Memória e estado • Fallback e contingência	• Identidade e autorização • Limites de ação • Proteção de dados • Auditoria e human-in-the-loop	• Observabilidade • Avaliação contínua • Latência e custo • Suporte e melhoria

Produção exige operação contínua

Quando aplicações e agentes entram em produção, a empresa precisa acompanhar disponibilidade, desempenho, falhas de integração, qualidade das respostas e consumo. A plataforma [LiveCloud](#) pode apoiar a visibilidade operacional do ambiente, enquanto [LiveCloud FinOps](#) ajuda a conectar consumo, custos, eficiência e decisões de otimização. Para automações e agentes que passam a executar tarefas recorrentes, o [Automation Care](#) contribui com governança, monitoramento e evolução planejada.

MENSAGEM-CHAVE

Publicar um agente não significa colocá-lo em produção. Produção exige saber o que ele pode consultar, quais ações pode executar, como seu desempenho será medido e quando uma pessoa deve assumir a decisão.

7. Roadmap: da fundação à aplicação inteligente

A modernização para IA não precisa acontecer de uma única vez. Uma abordagem por ondas permite gerar valor inicial, validar hipóteses e criar padrões reutilizáveis antes de ampliar escopo e criticidade.

1. Priorizar	Selecionar casos por valor, risco e viabilidade.	Backlog e critérios de sucesso
2. Preparar	Mapear dados, documentos, APIs, aplicações e permissões.	Mapa de fontes e dependências
3. Arquitetar	Escolher RAG, agentes, ferramentas e controles.	Arquitetura-alvo e plano de integração
4. Validar	Executar um piloto com métricas de qualidade, segurança e custo.	Evidências e decisão de escala
5. Operar e evoluir	Monitorar, governar, otimizar e ampliar o caso.	Modelo operacional contínuo

Conectando a jornada

A evolução pode começar pela organização do conhecimento com o [KnowledgeOps AI](#), avançar para avaliação arquitetural com o [Well-Architected Review by GotoBiz](#) e seguir para desenho, desenvolvimento e modernização por meio de [Consulting & Professional Services](#). À medida que a aplicação entra em produção, da plataforma [LiveCloud](#), [LiveCloud FinOps](#) e [Automation Care](#) ajudam a sustentar visibilidade, eficiência econômica e evolução operacional.

PRÓXIMO PASSO

Aplicações antigas, integrações frágeis e alto débito técnico podem limitar a adoção de GenAI e agentes. O próximo passo da jornada é avaliar como modernizar o ambiente para incorporar inteligência de maneira sustentável.



8. Seu ambiente está preparado para agentes de IA?

Modernizar aplicações para IA exige alinhar arquitetura, código, dados, integrações, segurança e operação. No próximo SmartMeeting, especialistas da GotoBiz mostrarão como AWS e IA Generativa podem acelerar essa evolução e preparar o ambiente para a nova geração de agentes inteligentes.

Descubra como modernizar aplicações IA Generativa

O desafio não é criar agentes. É fazê-los operar com segurança e escala.

QUERO PARTICIPAR DO SMARTMEETING

