## Módulo 05 - Usando bases de conhecimento

A Geração Aumentada por Recuperação (RAG) é uma técnica que melhora o desempenho dos modelos de base (FMs) ao incorporar informações externas. Embora esses modelos sejam treinados com grandes volumes de dados para tarefas como responder perguntas e traduzir textos, seu conhecimento pode ser limitado ou desatualizado. O RAG resolve isso ao buscar dados atualizados em fontes externas antes de gerar respostas.

### Casos de uso do RAG:

* Respostas a perguntas
* Geração de conteúdo
* IA Conversacional
* Análise de dados e insights
* Verificação de fatos
* Geração de conteúdo personalizado

### Arquitetura RAG:

* Ingestão de dados

O cerne da abordagem RAG da IA ​​generativa é a incorporação de dados externos além dos dados de treinamento originais do FM. Esses dados externos, extraídos de diversas fontes, como APIs, bancos de dados ou repositórios de documentos, expandem o conhecimento e as informações disponíveis para o sistema de IA generativa.

* Recuperar informações relevantes

Após a inserção dos dados no repositório de vetores, a próxima etapa é realizar uma pesquisa de relevância. A consulta do usuário é convertida em uma representação vetorial e, em seguida, uma pesquisa semântica é realizada nos bancos de dados de vetores. Isso permite que o sistema identifique e recupere as informações mais relevantes que correspondam ao conteúdo semelhante à consulta do usuário.

* Aumente o prompt de FM

A etapa final é ampliar a entrada (ou prompts) do usuário, incorporando os dados relevantes recuperados em contexto. Esse processo utiliza técnicas de engenharia de prompts para se comunicar efetivamente com o FM. O prompt ampliado permite que o modelo base gere respostas precisas e relevantes às consultas do usuário, utilizando os dados externos e a biblioteca de conhecimento criada anteriormente.

* Geração

O prompt aumentado é então passado como entrada para o modelo base para gerar a resposta.

### Desafios ao construir aplicativos RAG:

* Atualizando dados externos
* Escalabilidade
* Relevância e precisão
* Preconceito e justiça
* Avaliação e métricas

### Bases de conhecimento do Amazon Bedrock:

O Amazon Bedrock Knowledge Bases simplifica a configuração e a implementação do RAG automatizando várias etapas desse processo.

O Amazon Bedrock Knowledge Bases oferece suporte a dois métodos de API diferentes:

* **Recuperar** : Recuperar consulta uma base de conhecimento para buscar informações relevantes para uma solicitação do usuário. Isso normalmente é usado se você quiser personalizar a parte de geração do RAG.
* **RetreiveAndGenerate** : RetrieveAndGenerate vai um passo além e usa as informações recuperadas para complementar o prompt do FM. A resposta do FM cita apenas fontes na base de conhecimento que sejam relevantes para a consulta. O gerenciamento de contexto de sessão é integrado, para que seu aplicativo possa suportar facilmente conversas multi-turn.

