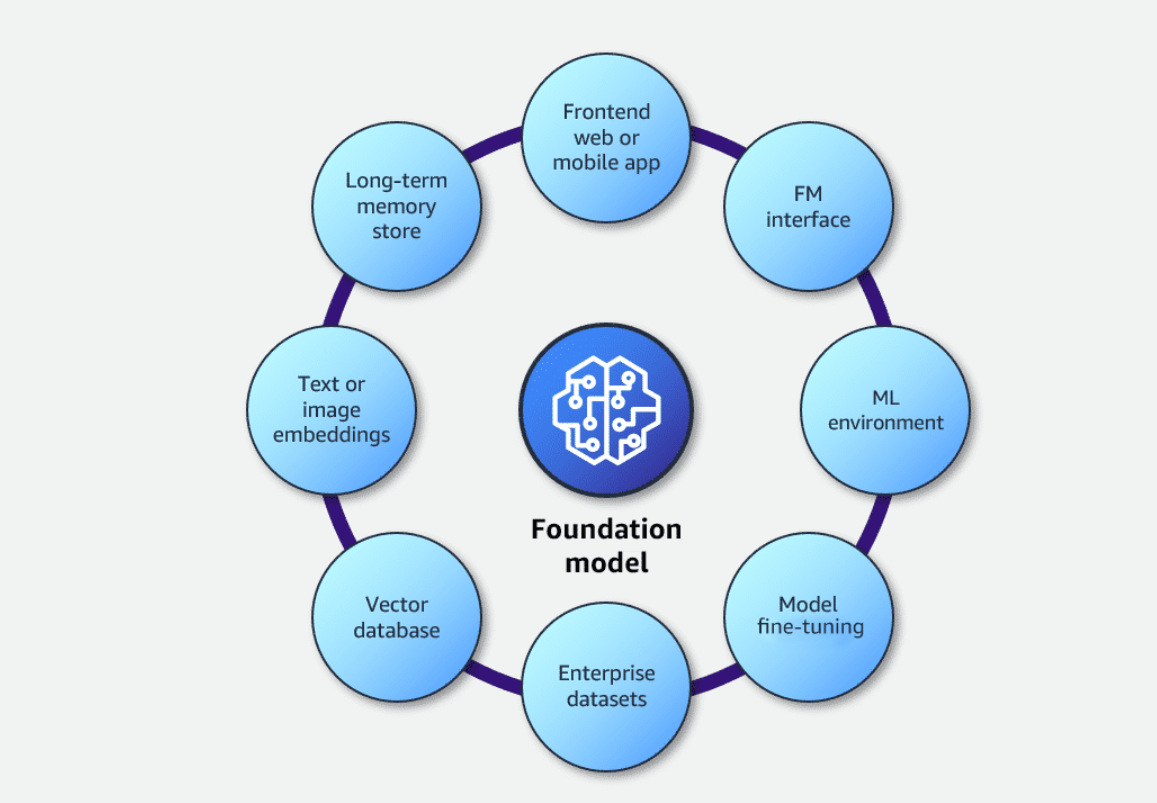
## Módulo 03 - Componentes do aplicativo:



O diagrama mostra como funciona um aplicativo de IA generativa:

* Foundation Model (Modelo Fundamental) é o núcleo do sistema de IA.

Os conjuntos de dados corporativos são vetorizados por um FM de incorporação e armazenados em um banco de dados vetorial para facilitar a pesquisa de documentos de alto desempenho, semelhante a uma consulta do usuário.

O Amazon Bedrock fornece modelos de incorporação, como o Amazon Titan Embeddings G1, um modelo de texto que pode converter texto em embeddings. Além disso, há o Cohere Embed, que oferece suporte para inglês e é multimodal — modelos de incorporação de texto para inglês e vários idiomas. Outro modelo disponível é a Incorporação Multimodal Titan, que permite a incorporação de imagens e textos.

Exemplo de banco de dados vetoriais da Amazon:  
 - Amazon OpenSearch Service (provisionado)  
 - Amazon OpenSearch sem servidor  
 - Extensão pgvector no Amazon Relational Database Service (Amazon RDS) para PostgreSQL  
 - Extensão pgvector no Amazon Aurora PostgreSQL - Edição compatível

RAG - é uma técnica que combina IA generativa com mecanismos de recuperação de documentos. Ele usa dados corporativos atualizados (armazenados em bancos vetoriais) para gerar respostas mais precisas, superando as limitações de conhecimento do modelo base.

Ajuste Fino - é quando você modifica permanentemente um modelo de IA com dados rotulados (exemplos bem definidos), para que ele melhore em tarefas específicas ou aprenda sobre um domínio (área de conhecimento) particular.

Personalização (aprendizado baseado em prompts)  
 Adaptação de domínio (pré-treinamento contínuo)

Padrões de arquiteturas que podem ser implementadas com o Amazon Bedrock e o LangChain:

* Geração de texto ou código
* Resumo de texto
* Resposta a perguntas
* Assistentes de IA

LangChain - Framework de código aberto usado para construir aplicativos com LLMs. Ajuda a orquestrar prompts, respostas e integração com outros componentes (ex: bancos de dados, APIs).

* Frontend (site ou app) é a interface onde os usuários interagem com a IA.

O frontend atua como ponte inteligente entre o usuário e a IA, cuidando da comunicação, tratamento de dados e experiência do usuário.

* ML Environment é onde o modelo é treinado.
* Model Fine-tuning é o ajuste do modelo para tarefas específicas.
* Enterprise Datasets são os dados usados para treinar o modelo.
* Vector Database armazena dados em forma de vetores, facilitando buscas.
* Embeddings são representações numéricas de textos ou imagens.
* Long-term Memory Store permite que o modelo "lembre" de interações passadas.