Spring AI – Visão Geral

Spring AI é um projeto do ecossistema Spring que fornece abstrações e integrações para facilitar o uso de Inteligência Artificial Generativa em aplicações Java. Ele segue os mesmos princípios do Spring Framework: simplicidade, desacoplamento e produtividade.

O objetivo principal do Spring AI é permitir que desenvolvedores criem aplicações baseadas em LLMs (Large Language Models) sem depender diretamente de SDKs específicos de provedores, como OpenAI, Azure OpenAI, Ollama ou outros.

Principais Conceitos

ChatClient: abstração central para interação com modelos de linguagem. Ele encapsula prompts, mensagens do sistema, histórico de conversa e opções de inferência.

VectorStore: responsável por armazenar embeddings vetoriais e permitir buscas por similaridade semântica. É fundamental para arquiteturas RAG (Retrieval Augmented Generation).

DocumentReader e TextSplitter: usados para ingestão de dados. Permitem ler arquivos (PDF, TXT, HTML) e dividir o conteúdo em chunks apropriados para indexação.

Spring AI e RAG

Uma das aplicações mais comuns do Spring AI é a implementação de RAG. Nesse padrão, dados externos são recuperados de um VectorStore e utilizados como contexto adicional para o modelo de linguagem.

O fluxo típico envolve: upload de documentos, extração de texto, divisão em chunks, geração de embeddings, armazenamento vetorial e, por fim, recuperação baseada em similaridade durante a inferência.

Com isso, é possível criar assistentes corporativos, buscadores semânticos e sistemas de perguntas e respostas altamente contextualizados.