

LangChain: Criando chatbots inteligentes com RAG

Pipelines para Dados Complexos

Instrutor(a): Leonardo Pena




MERGULHE EM TECNOLOGIA_



Seja bem vindo (a)!





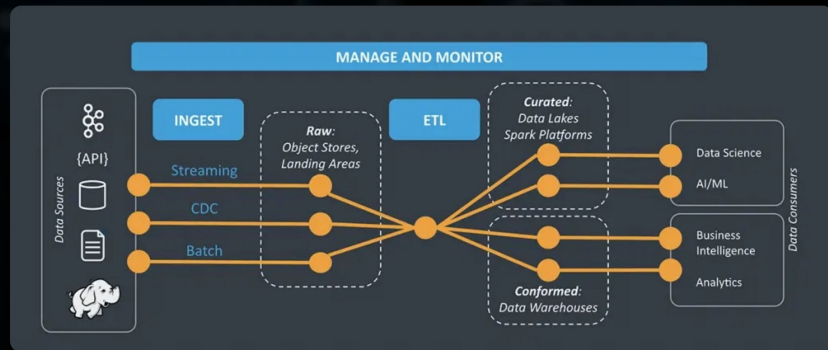
Aula

Pipelines para Dados

Complexos

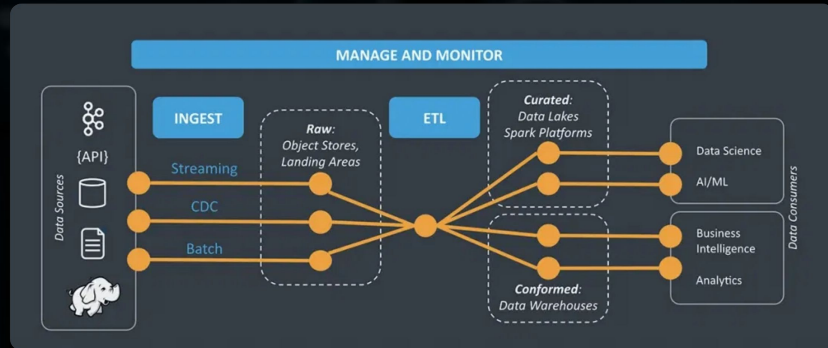
O que é um Pipeline de Ingestão?

- Um **pipeline de ingestão** é o processo completo pelo qual dados brutos de diversas fontes são processados para uso em sistemas RAG.
- O processo inclui: **carregamento** de dados brutos, **divisão em chunks**, **transformação**, **conversão em embeddings** e **indexação** em armazenamento vetorial.



O que é um Pipeline de Ingestão?

- Pipelines bem projetados garantem que os dados sejam **processados de forma consistente**, mantendo a qualidade e o contexto necessários para respostas precisas.
- A eficiência do sistema RAG depende diretamente da **qualidade do pipeline de ingestão**, que deve ser adaptado aos tipos específicos de dados processados.



O que é um Pipeline de Ingestão?

- A arquitetura de um pipeline de dados segue geralmente o modelo **ETL (Extract, Transform, Load)**, adaptado para as necessidades específicas de sistemas RAG.

The ETL Process Explained



O que é um Pipeline de Ingestão?



Extract: Coleta de dados de múltiplas fontes

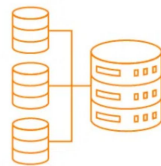


Transform: Processamento, chunking e enriquecimento



Load: Indexação em armazenamento vetorial

The ETL Process Explained



Extract

Retrieves and verifies data from various sources



Transform

Processes and organizes extracted data so it is usable



Load

Moves transformed data to a data repository

O que é um Pipeline de Ingestão?

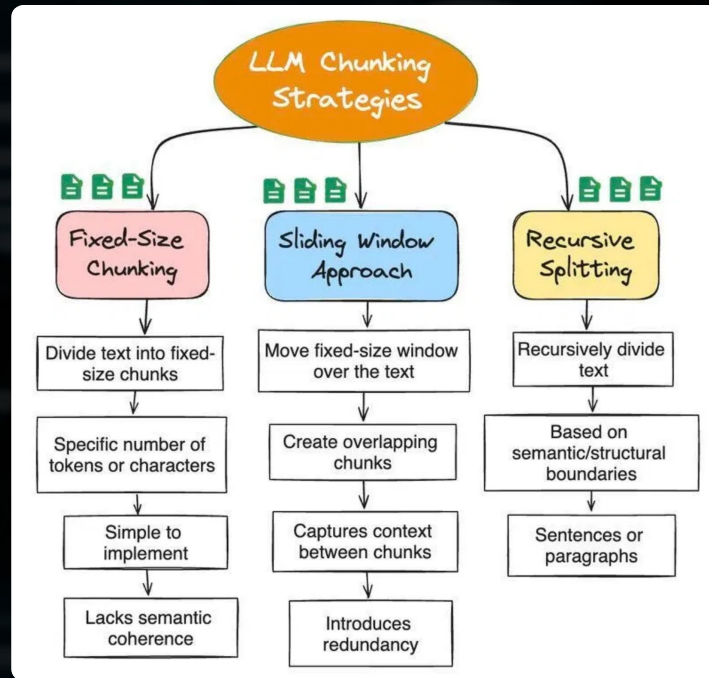
- Uma arquitetura bem estruturada proporciona **escalabilidade**, **manutenção simplificada**, **monitoramento eficiente** e **adaptabilidade** a novos tipos de dados.

The ETL Process Explained



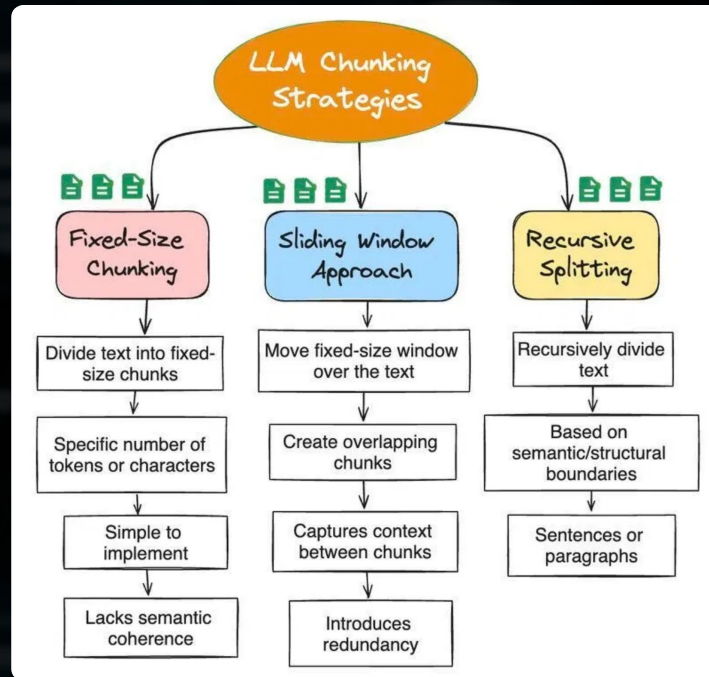
Chunking Adaptativo

Chunking adaptativo é a técnica de dividir documentos longos em pedaços (chunks) de forma inteligente, preservando o contexto semântico.



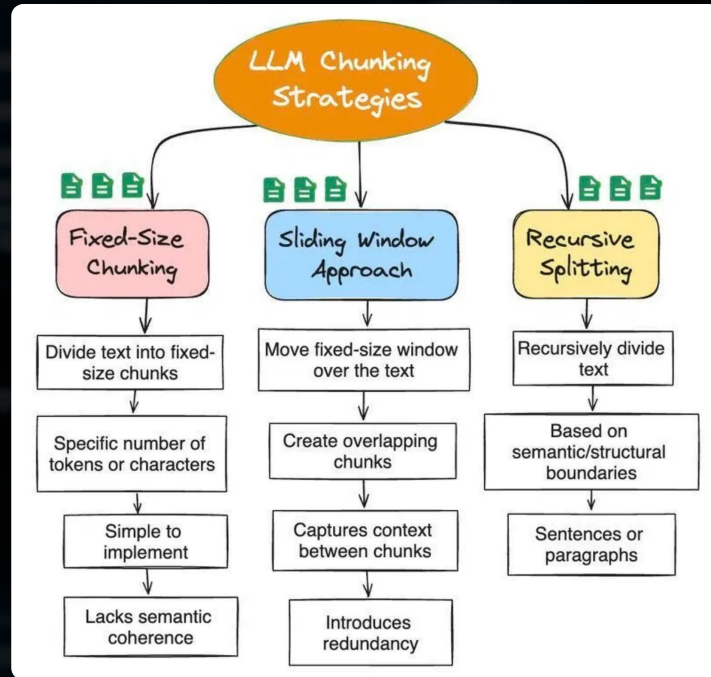
Chunking Adaptativo

- 🧩 **Fixed-Size Chunking:** Divisão por número fixo de tokens ou caracteres
- 📁 **Sliding Window:** Sobreposição (overlap) para manter contexto entre chunks
- 👥 **Recursive Splitting:** Divisão baseada na estrutura semântica do texto



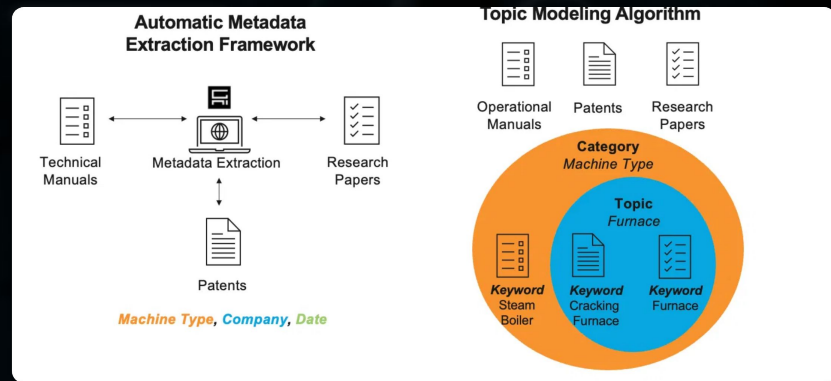
Chunking Adaptativo

A estratégia de chunking deve ser **adaptada ao tipo de dado**: textos contínuos, tabelas, código-fonte ou documentos estruturados requerem abordagens diferentes.



Extração de Metadados

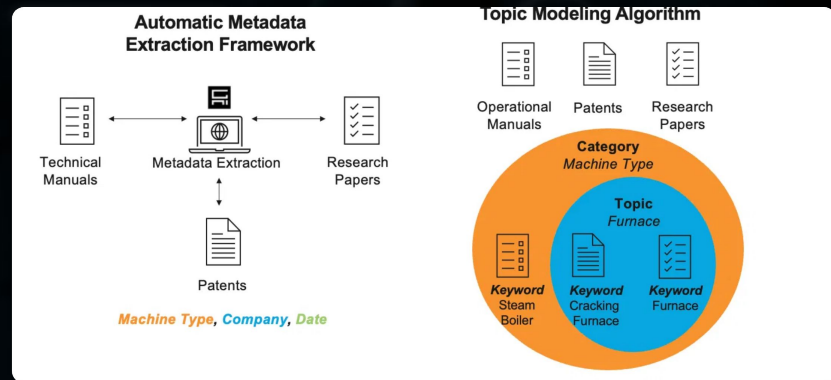
A extração de metadados é a etapa do pipeline que identifica e captura informações adicionais sobre os documentos, enriquecendo o contexto disponível.



Extração de Metadados

- 👤 **Autoria:** Autor, organização, credenciais
- 📅 **Temporal:** Data de criação, modificação, publicação
- 🏷️ **Categórico:** Tópicos, tags, classificações a
- 🔗 **Estrutural:** Hierarquia, seções, relacionamentos

Metadados bem extraídos permitem filtragem precisa durante a busca e contextualização aprimorada das respostas geradas pelo LLM.



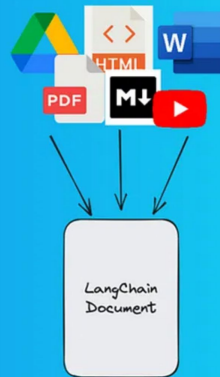
Ferramentas para Dados Complexos

Lidar com dados variados e complexos exige **ferramentas especializadas** para extração, processamento e transformação eficientes.



LangChain Indexes

Document Loaders



Ferramentas para Dados Complexos



LangChain Loaders

Interface unificada para carregar dados de múltiplas fontes, como PDFs, CSVs e páginas web.



Unstructured

Biblioteca focada em extrair texto e metadados de arquivos complexos como PDFs e e-mails, preservando sua estrutura original.



OCR (Tesseract)

Reconhecimento Óptico de Caracteres, indispensável para extrair texto de documentos escaneados ou imagens.



LangChain Indexes

Document Loaders

