



Data Science
Academy



Engenharia de Dados Para IA

Conhecimentos, técnicas e ferramentas para preparar os dados que alimentam LLMs, RAG e Agentes de IA.

👉 Modelos estão ao alcance de todos. O diferencial está nos dados.

POR QUE IMPORTA

Modelos estão ao alcance de todos. O diferencial está nos dados.

Engenharia de Dados Para IA reúne conhecimentos, técnicas e ferramentas para preparar dados que:



Treinam e ajustam modelos

Datasets limpos, representativos e versionados.



Alimentam sistemas de RAG

Documentos em chunks, embeddings e índices de busca.



Dão contexto aos Agentes

Informação atualizada e confiável no momento da decisão.

O QUE MUDA

Na IA, quem consome o dado é um modelo

Dimensão	Engenharia tradicional	Engenharia para IA
Consumidor	Analistas, dashboards e BI	Modelos de ML, LLMs, RAG e agentes
Tipo de dado	Estruturado (tabelas)	Não estruturado somado ao estruturado
Unidade de trabalho	Linhas e colunas	Chunks, embeddings e features
Qualidade	Completude e consistência	Também relevância, viés e vazamento
Atualização	Lote diário costuma bastar	Contexto em minutos ou em tempo real
Rastreabilidade	Linhagem até o relatório	Linhagem até o modelo e cada resposta

Os fundamentos continuam valendo: SQL, modelagem e pipelines confiáveis seguem como base, agora com novas camadas por cima.

CONHECIMENTOS

Base sólida de dados, somada a ML e LLMs



Base de dados

- | SQL e Python avançados, também para texto, JSON e vetores
- | Arquitetura lakehouse e camadas bronze, prata e ouro
- | Processamento distribuído e streaming



IA e recuperação

- | Fundamentos de ML: treino, validação, overfitting e data leakage
- | LLMs: tokens, janela de contexto e embeddings
- | Busca vetorial e híbrida e arquiteturas RAG



Operação e governança

- | Governança, LGPD, mascaramento e linhagem
- | Cloud, Docker e Kubernetes
- | Orquestração e monitoramento de pipelines

PIPELINE

Da fonte ao modelo, um ciclo que se realimenta



Em todas as etapas: testes de qualidade, versionamento e linhagem

TÉCNICAS

Oito técnicas que sustentam dados prontos para IA

01

Ingestão e parsing

Extrair texto e estrutura de PDFs, HTML e imagens, com metadados.

02

Limpeza e privacidade

Remover ruído e duplicatas e mascarar dados pessoais, conforme a LGPD.

03

Chunking com metadados

Trechos que respeitam a estrutura, com fonte, data e permissões.

04

Embeddings e indexação

Índices versionados e busca híbrida: vetorial e por palavra-chave.

05

Feature store

A mesma feature no treino e na inferência evita o training-serving skew.

06

Qualidade e data contracts

Testes automáticos e contratos entre quem produz e quem consome.

07

Versionamento e linhagem

Saber que dados treinaram cada modelo e originaram cada resposta.

08

Avaliação e dados sintéticos

Datasets rotulados e casos raros gerados com validação humana.

FERRAMENTAS

Uma stack organizada em camadas

Ingestão

Airbyte, Fivetran, Apache Kafka, Debezium

Parsing

Docling, Unstructured, Apache Tika

Processamento

Apache Spark, Polars, DuckDB, Apache Flink

Transformação

dbt, SQLMesh

Lakehouse

Apache Iceberg, Delta Lake, Databricks, Snowflake, Microsoft Fabric

Bancos vetoriais

pgvector, Qdrant, Milvus, Weaviate, Pinecone

Frameworks RAG

LlamaIndex, LangChain

Orquestração

Apache Airflow, Dagster, Prefect

Qualidade

Great Expectations, Soda, Monte Carlo

Versionamento

DVC, lakeFS, OpenLineage

Feature store

Feast, Hopsworks

Em destaque, as camadas novas trazidas pela IA. A escolha depende de volume, nuvem adotada e maturidade do time.

PRÓXIMOS PASSOS

Pratique com um pipeline de ponta a ponta



Com testes de qualidade em cada etapa, esse projeto já demonstra as competências que o mercado procura.



Comece pelo simples: PostgreSQL com pgvector, Python e um orquestrador resolvem boa parte dos projetos básicos de RAG antes que um banco vetorial dedicado se torne necessário.

CONCLUSÃO

Modelos mudam a cada trimestre. Dados e boa engenharia continuam gerando valor.

Conheça a Formação AI Data Engineer 4.0:

www.datascienceacademy.com.br