# RAG na sua base transacional com MongoDB Atlas

Gabriel Limoni


DEVPIRA + pecege

# Gabriel Limoni

- Formado em ADS pelo IFSP
- Dev e Tech Lead na Liven
- Resolvedor de problemas - *geralmente com TI*
- Baterista
- Vice-campeão de basquete do Inter Pira 2017 e 18 pelo IFSP
- Sócio do melhor studio de pilates de São Carlos ***@annamaria.pilates***

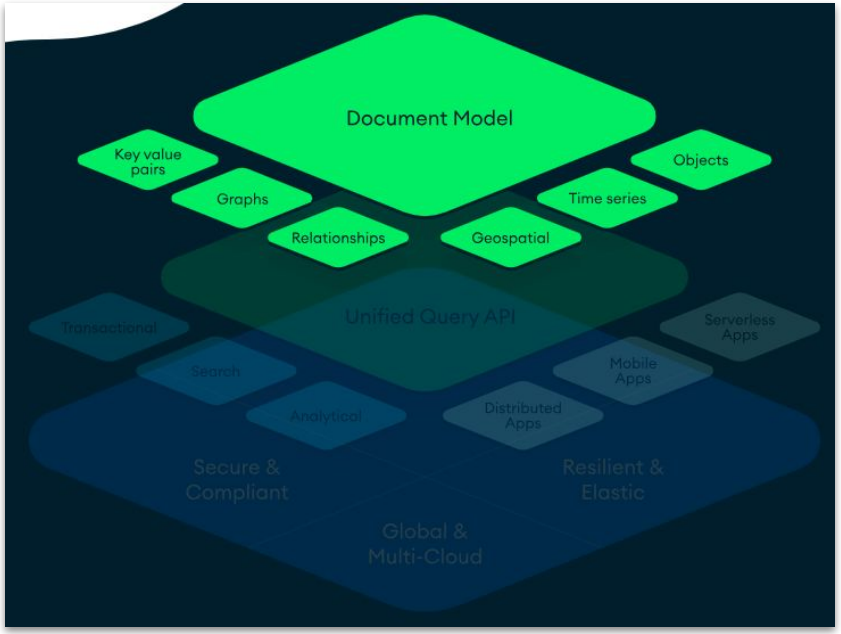

PILATES
PILATES

# MongoDB

- Banco de dados NoSQL orientado a documento

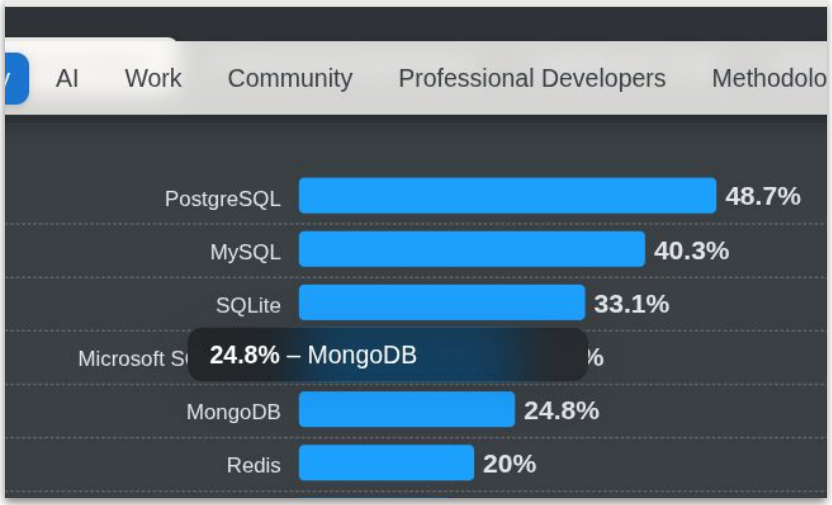
# MongoDB

- Banco de dados NoSQL orientado a documento
- Aplicável em diversos paradigmas
  - Documento
  - Chave-valor
  - Grafo
  - Time-series
  - etc...



# MongoDB

- Banco de dados NoSQL orientado a documento
- Aplicável para diversos paradigmas
  - Documento
  - Chave-valor
  - Grafo
  - Time-series
  - etc...
- NoSQL mais utilizado (stackoverflow 2024)



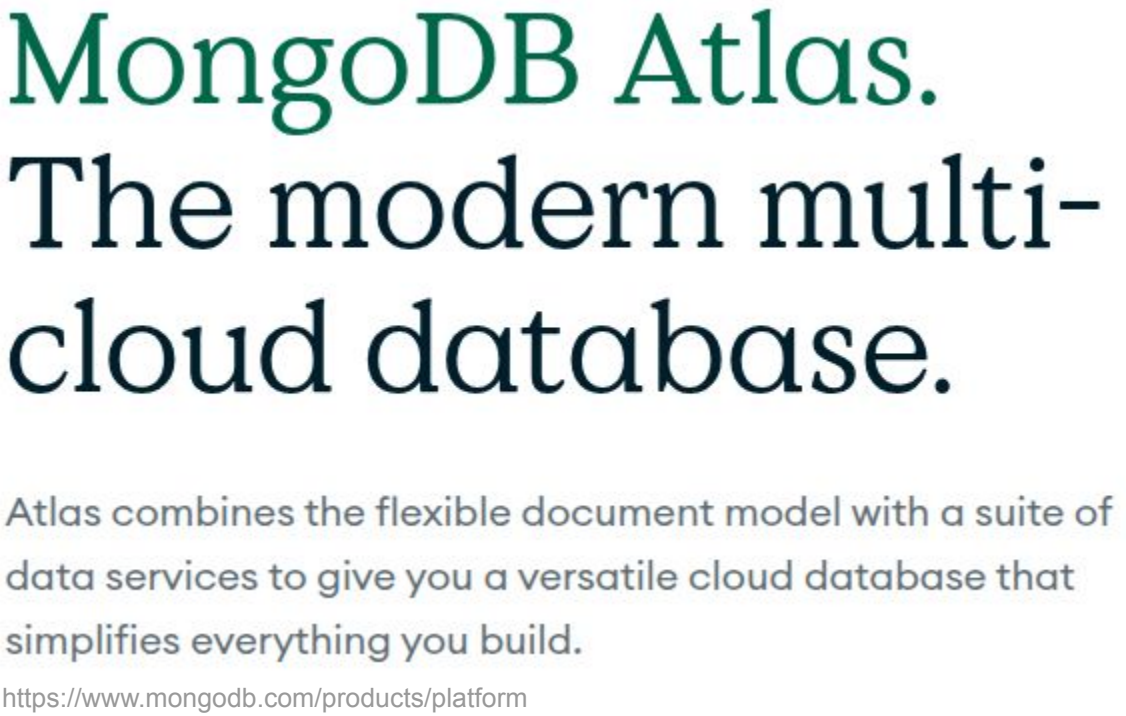

https://survey.stackoverflow.co/2024/technology

# MongoDB

- Banco de dados NoSQL orientado a documento


- Óbvio que tem coisa pra IA. Quem não tem?

# Atlas


MongoDB Atlas.
The modern multi-cloud database.
Atlas combines the flexible document model with a suite of data services to give you a versatile cloud database that simplifies everything you build.
https://www.mongodb.com/products/platform

# Atlas - Data Services

### Database

A multi-cloud database service built for resilience, scale, data privacy, and security.

Learn more >

### Search

Build relevance-based search 4x faster and for 77% lower cost than alternative search solutions.

Learn more >

### Vector Search

Build intelligent applications powered by semantic search and generative AI over any type of data.

Learn more >

### Charts

Bring your data to life and get real-time insights with embeddable dashboards and visualizations.

Learn more >

### Atlas CLI

Create and manage your MongoDB Atlas database from the command line.

Learn more >

### Data Federation

Seamlessly query, transform, and aggregate data from one or more Atlas databases and AWS S3 buckets.

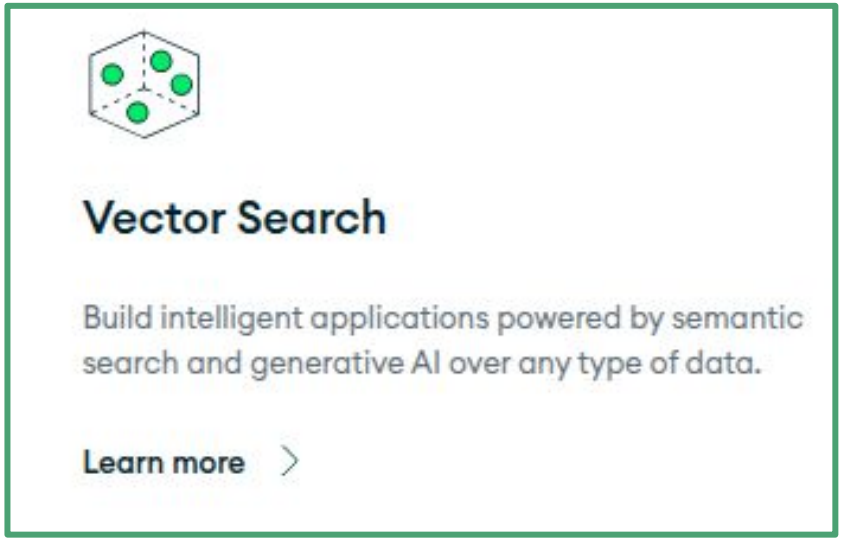
Learn more >

### Online Archive

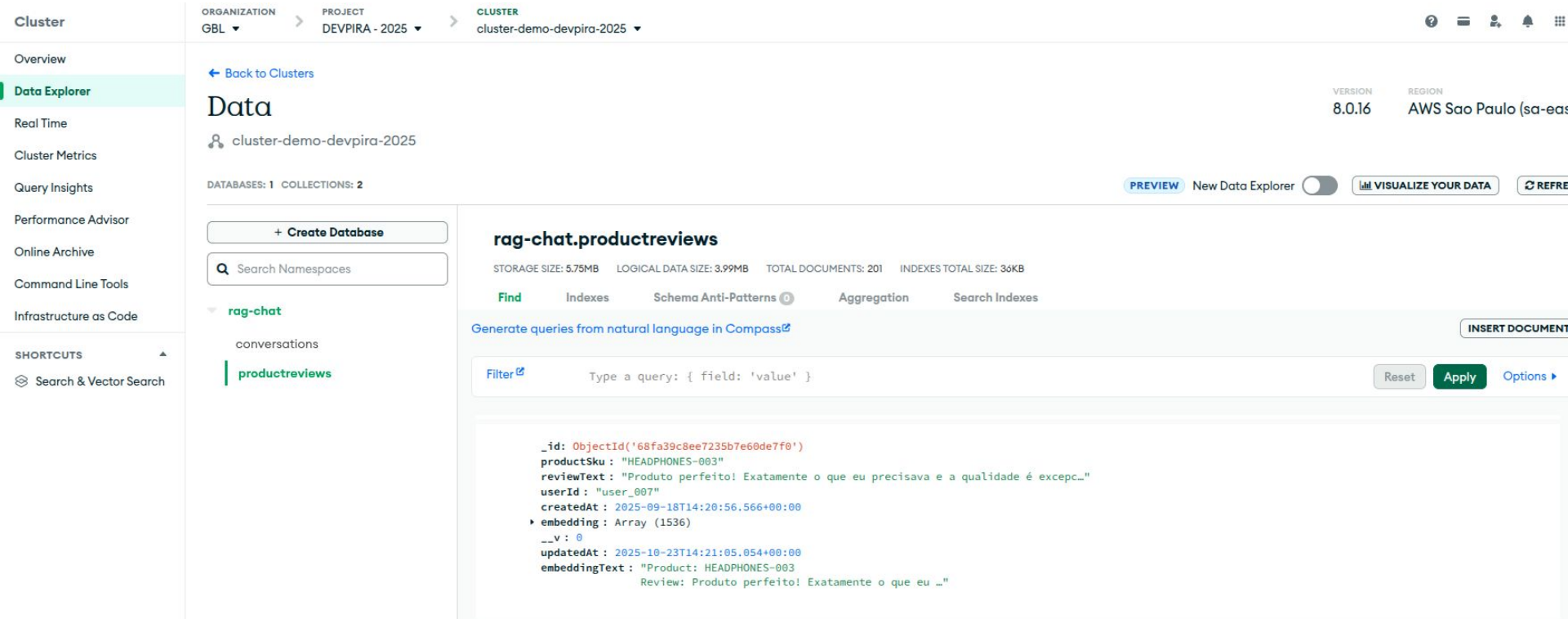
Archive data from Atlas clusters to fully managed object storage and query it through a single endpoint.

Learn more >

# Atlas - *Vector Search*



# Aplicação demo usando MongoDB Atlas



# **L***arge* **L***anguage* **M***odel*

- Modelo de IA treinado em **grandes volumes de texto**
- Capaz de **processar linguagem natural**
- Baseado em arquiteturas modernas, como **Transformers**
- Aprende padrões, contexto e **relações semânticas** das palavras
- Utilizado para tarefas como chatbots, resumo, tradução, análise de sentimentos, entre outros
- Não é um banco de dados, mas sim um **modelo estatístico** que prevê palavras e ideias
- Pode ser especializado ou **aumentado via RAG** para acesso a conhecimento atualizado
- Requer **alto poder computacional** para treinamento
- Necessita de **curadoria e alinhamento** para reduzir erros e vieses

# **L***arge* **L***anguage* **M***odel* - bullet points para sala de aula

- Modelo de IA treinado em **grandes volumes de texto**
- Capaz de **processar linguagem natural**
- Baseado em arquiteturas modernas, como **Transformers**
- Aprende padrões, contexto e **relações semânticas** das palavras
- Utilizado para tarefas como chatbots, resumo, tradução, análise de sentimentos, entre outros
- Não é um banco de dados, mas sim um **modelo estatístico** que prevê palavras e ideias
- Pode ser especializado ou **aumentado via RAG** para acesso a conhecimento atualizado
- Requer **alto poder computacional** para treinamento
- Necessita de **curadoria e alinhamento** para reduzir erros e vieses

# O que é uma **LLM**? - by Chat GPT

*Uma LLM é um modelo de IA capaz de gerar e compreender linguagem com alta precisão usando padrões aprendidos de grandes volumes de texto.*

# O que é uma **LLM**? - by Chat GPT

*Uma LLM é um modelo de IA capaz de gerar e* ***compreender*** *linguagem com* ***alta precisão*** *usando padrões aprendidos de grandes volumes de texto.*

# **LLM** - compreendem mesmo?

compreende mesmo ou apenas processa?

Ótima pergunta — e é importante ser preciso aqui.

**LLMs não *compreendem* linguagem no sentido humano.**
Elas **processam padrões** e **preveem a próxima palavra** com base em correlações aprendidas em grandes volumes de dados.

Então, uma frase mais correta seria:

**"Uma LLM é um modelo de IA que processa linguagem ao identificar padrões em grandes quantidades de texto para gerar respostas úteis."**

# **LLM** - respostas sempre úteis?

as respostas são sempre úteis?

Direto ao ponto: **não, as respostas não são sempre úteis.**

LLMs fazem um ótimo trabalho na maior parte do tempo, mas têm limitações claras:

- **Podem errar** quando falta contexto ou quando o assunto é muito específico.
- **Podem soar confiantes mesmo quando estão erradas** — esse é o famoso *alucinar*.
- **Podem interpretar a pergunta de forma incompleta** e entregar algo fora do alvo.
- **Dependem totalmente da qualidade do prompt** e do conhecimento disponível no treinamento ou no RAG.

# **LLM** - desafios/problemas

quais os problemas que ocorre com as as LLMs? Liste alguns poucos pra eu colocar no slide, por favor

Aqui estão **poucos e diretos problemas** de LLMs para usar no slide:

- **Alucinações:** respostas incorretas apresentadas com confiança.
- **Falta de compreensão real:** opera por padrões, não por entendimento.
- **Vieses dos dados:** pode reproduzir preconceitos presentes no treinamento.
- **Sensibilidade ao contexto:** pequenas mudanças no prompt podem gerar respostas muito diferentes.
- **Dependência de dados limitados:** não sabe fatos posteriores ao seu treinamento (a menos que use RAG).

# **LLM** - hype e bolha?

**Em resumo:**

> *Existe hype, pode haver bolha financeira, mas LLMs já são uma tecnologia fundamental — o "estouro" apenas ajustaria o mercado, não eliminaria o avanço.*

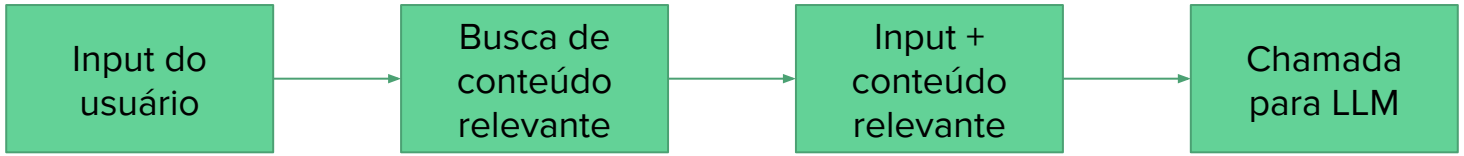
# **LLM** - by ChatGPT *(pós bate papo)*

*Uma LLM é um modelo que processa linguagem com base em padrões aprendidos, podendo gerar respostas valiosas, mas também* ***sujeitas a erros, vieses e limitações que exigem uso crítico****.*

# **R***etrieval-***A***ugmented* **G***eneration*

- Técnica que permite uma LLM **incorporar novas informações** no seu processamento
- Dispensa a necessidade de **retreinar a LLM** a cada nova informação
- Geralmente implementado com **busca semântica** em bases vetoriais

*Antes de chamar uma LLM, **busca conteúdo relevante***
*em uma ou mais base de dados e **joga pra dentro do prompt***




https://arxiv.org/pdf/2406.04369

# Fluxo de RAG típico

**Pré-processamento**

1. Geração de **embedding** do conteúdo que será usado para aumentar o contexto *(é muito comum gerar os embeddings de documentos PDF como manuais, artigos e outros documentos de referência)*
2. Armazenamento dos embeddings em uma **base vetorial**

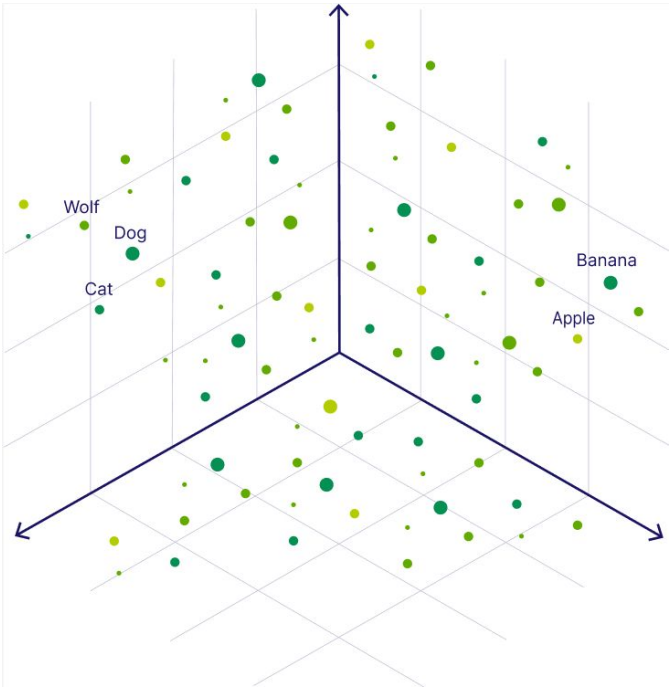
**Processamento em tempo de execução**

1. Input do usuário é transformado em um embedding
2. Consulta na base vetorial pelo embedding gerado
3. Conteúdo obtido é usado para aumentar o prompt

** podem ser realizadas queries híbridas - apenas semântica pode não ser suficiente*

# Banco de dados vetorial

- Vetor
  - lista de números de tamanho fixo
  - cada item no array representa uma dimensão
  - podem haver centenas ou milhares de dimensões a depender da complexidade da informação
  - pode ser uma palavra, frase, documento completo, imagem, áudio, etc...
- Banco de dados
  - armazenamento dos vetores
  - queries por proximidade dos vetores
  - diversos algoritmos disponíveis
  - ***matemática braba***




https://weaviate.io/blog/what-is-a-vector-database

# Usando RAG nas reviews de produtos

**Implementação**

- Geração de embedding da review + SKU do produto assim que adicionada na base (pré-processamento)
- Aplicação web Chat GPT-like para conversação aplicando RAG (runtime)

**Casos de uso**

- Q&A sobre as avaliações de um determinado produto
- Oportunidades de melhoria
- Comparação de sentimentos entre produtos
- etc...

# DEMO *(AO VIVO SE DEUS QUISER)*

devpira-2025.limoni.dev.br

# Concluindo

- O bagulho é complexo e **requer profissionais especializados** para um bom resultado em produção
- Talvez desse pra **resolver de um jeito mais simples** sem ser com vector search a depender do caso de uso
- Só **busca semântica pode não resolver** - queries híbridas podem ser necessárias
- Uso de tools e **grafos mais elaborados** geralmente é necessário
- Todo dia surge uma **coisa nova**
- **Arquitetura** para pré-processamento e runtime
- **Custo!**
- **Hype?**
- **Bolha?**


?
AI

Nosso objetivo sempre é melhorar e pra isso gostaríamos de contar com sua ajuda.

UM FUTURO CHEIO DE POSSIBILIDADES

INTELIGÊNCIA, SUSTENTABILIDADE E COMUNIDADE

# O que você achou?

DEVPIRA + pecege