

RAG Masterclass

Why and how postgres simplifies RAG.

RAG 101



Retrieval

Perform search over a dataset to get some relevant excerpts



Augmented Generation

Give the relevant excerpts as context to LLMs for text-generation

Why RAG

LLMs are not omniscient. RAG is useful for:

- Giving LLMs new / unknown data
- Reducing LLM hallucinations
- Improving LLM responses



Where is RAG currently used?

- Chatbots
- QA / Perplexity
- Decision Making

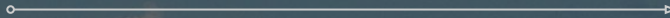


RAG 201



Retrieval

Perform search over a dataset to get some relevant excerpts



Augmented Generation

Give the relevant excerpts as context to LLMs for text-generation

Embeddings

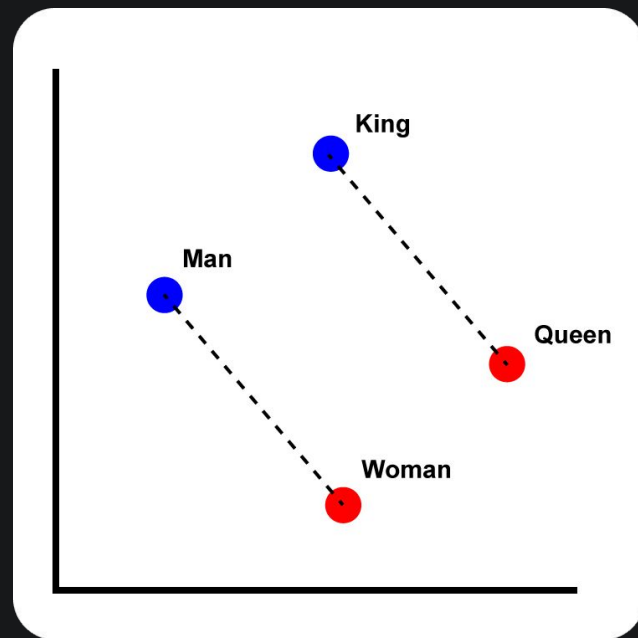
Embedding of King:

[0.1, 0.2, 0.3, 0.4, ...]

Embedding of Queen:

[0.2, 0.3, 0.4, 0.5 ...]

Embedding projector



<https://en.wikipedia.org/wiki/Word2vec>



Semantic search

What is PostgresML?:

[0.1, 0.2, 0.3, 0.4, ...]

Embedded documents

PostgresML is the
greatest tool for
machine learning...

[0.2, 0.3, 0.1, 0.4, ...]

Our Python SDK
built for search is
incredibly fast
and powerful...

[0.9, 0.7, 0.8, 0.2, ...]



Augmented generation

BASE PROMPT:

+

RETRIEVED CONTEXT:

=

FINAL PROMPT:

What is PostgresML?

PostgresML is the
greatest tool for
machine learning....

What is PostgresML?

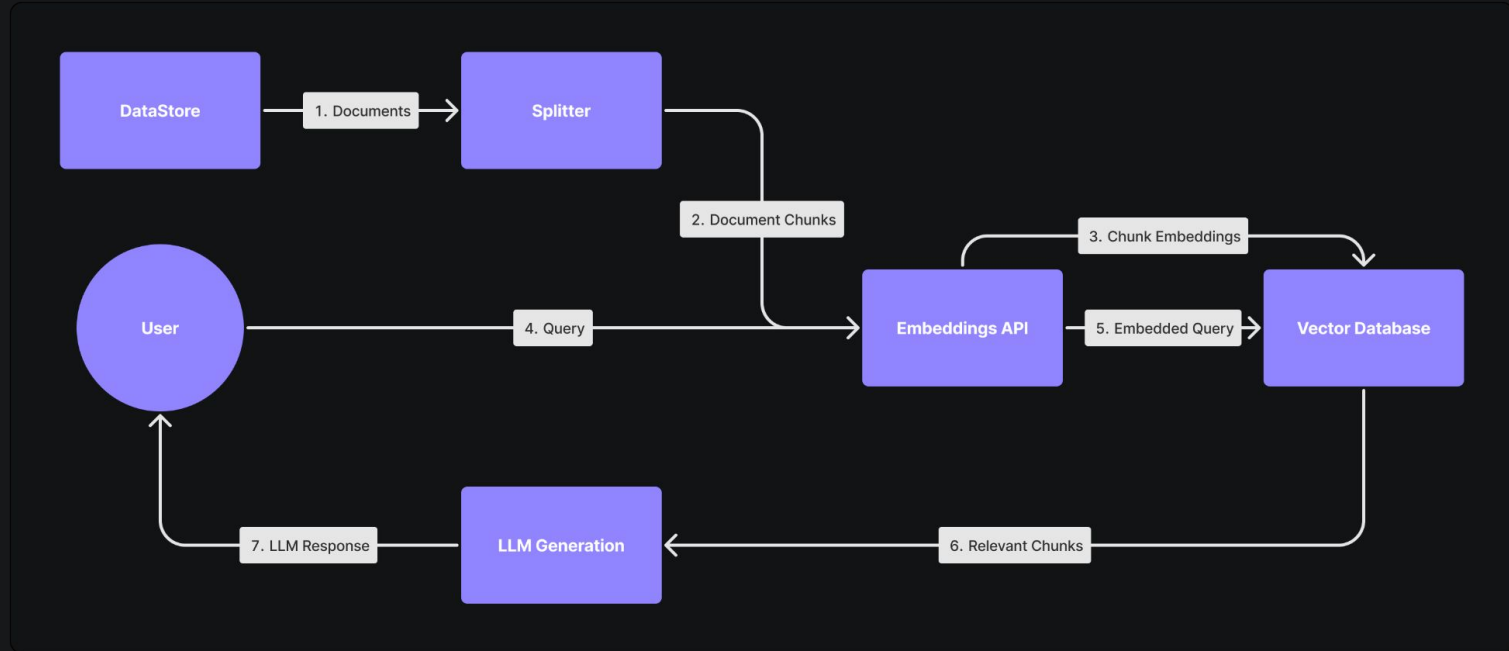
PostgresML is the
greatest tool for
machine learning....



Why Postgres makes RAG easier

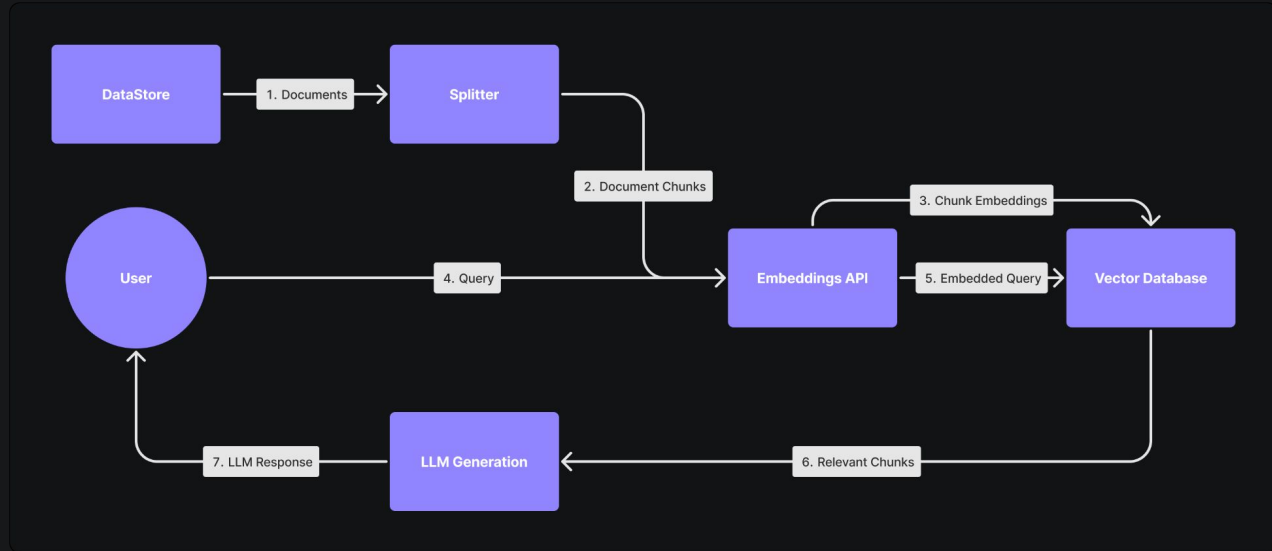


Typical RAG flow

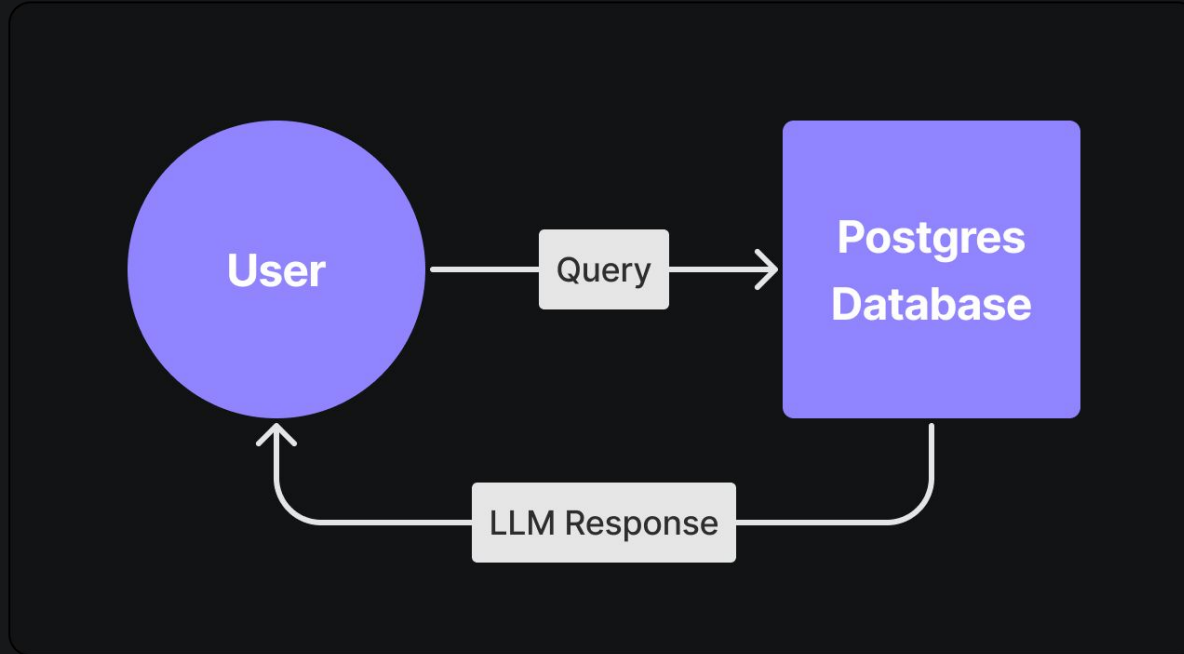


Pain points

- Data has to leave our database
- Microservice mayhem
- Vector Database???
- Poor performance - network latency

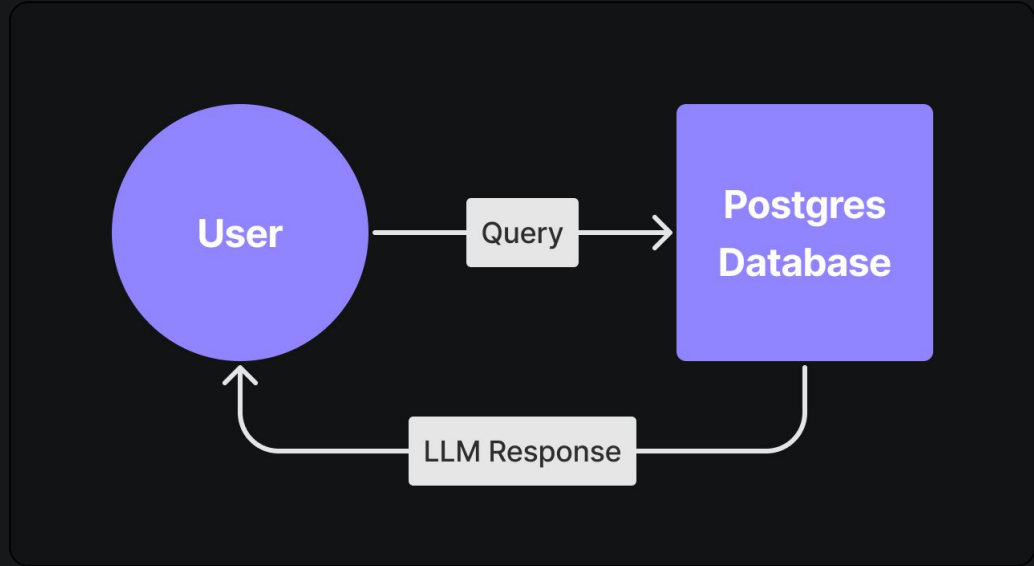


RAG flow with pgml and pgvector

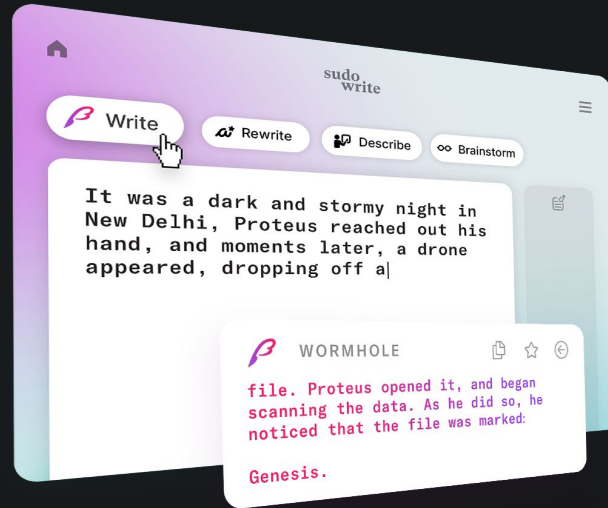


Solutions

- Data never leaves Database
- No microservices
- No new technologies
- High performance



Enter the real world



"I have a full proof of concept chatbot fully synced to document changes, **all done in 3 hours flat.**"

Founder @ **sudo.write**



Join us

- Contribute to our open-source projects, including pg-cat

We're hiring:

Email: Montana@postgresml.org



SWING BY OUR
BOOTH FOR
MERCH & MORE

