PostgresML

Maximizing RAG Performance

With LlamaIndex On PostgresML

Build better with LlamaIndex + PostgresML

5/30/2024

RAG 101

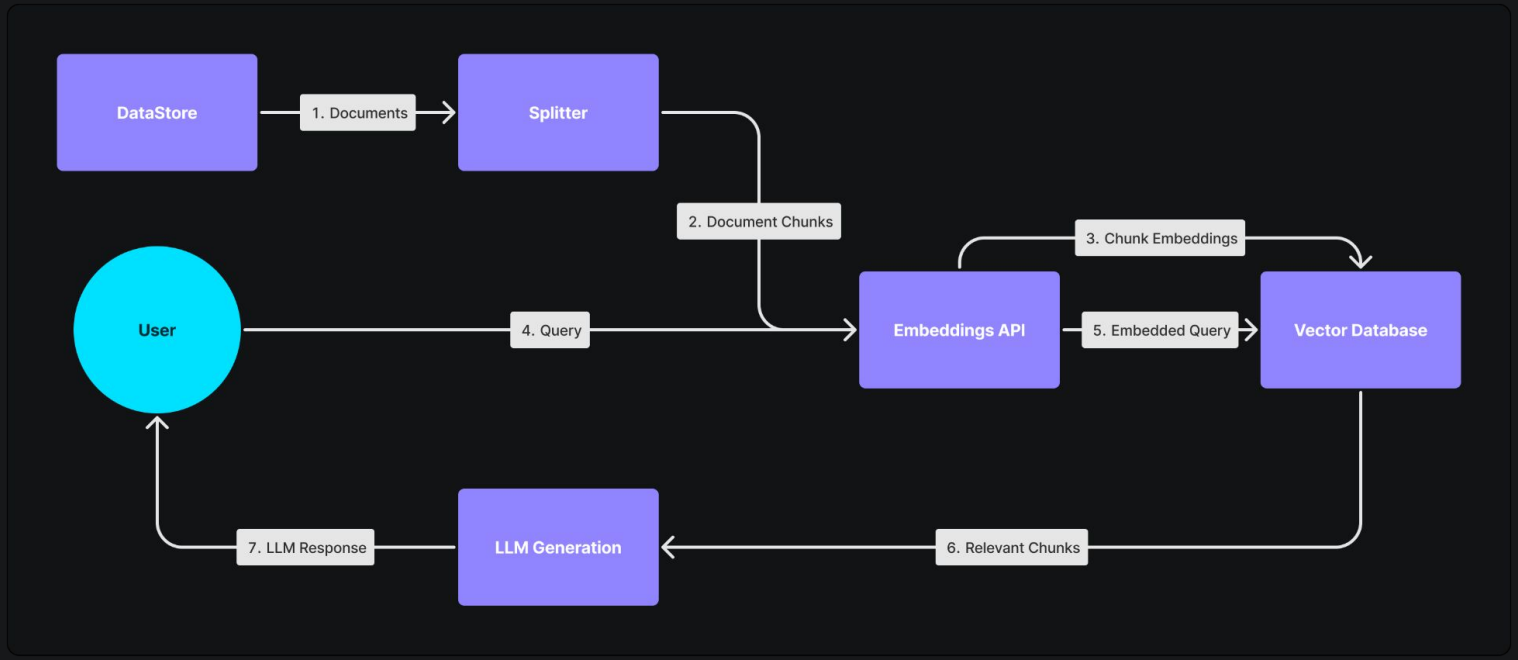
Retrieval

Perform search over a dataset to get some relevant excerpts

Augmented Generation

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

Typical RAG flow



What Does RAG Cost?

Developer Cost

New technologies to learn

Financial Cost

Three services to pay for

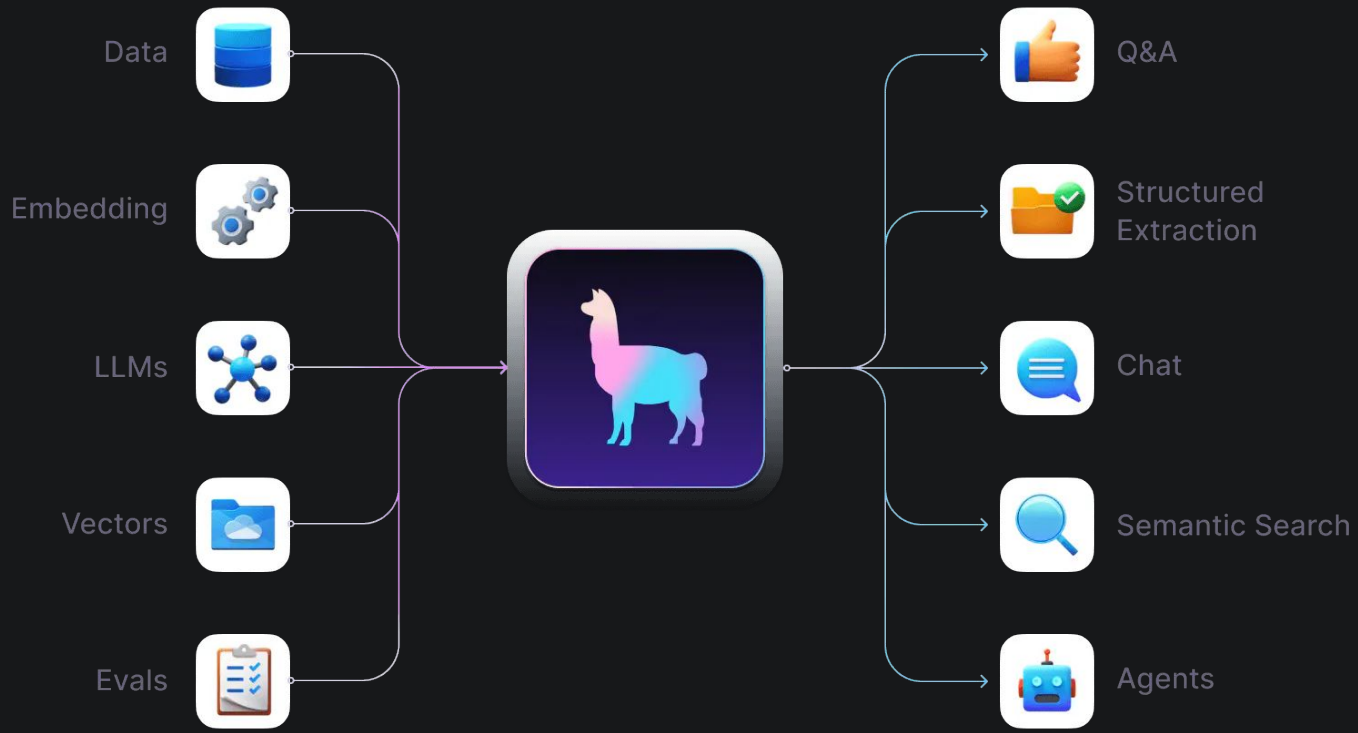
Scaling Cost

Many points of failures

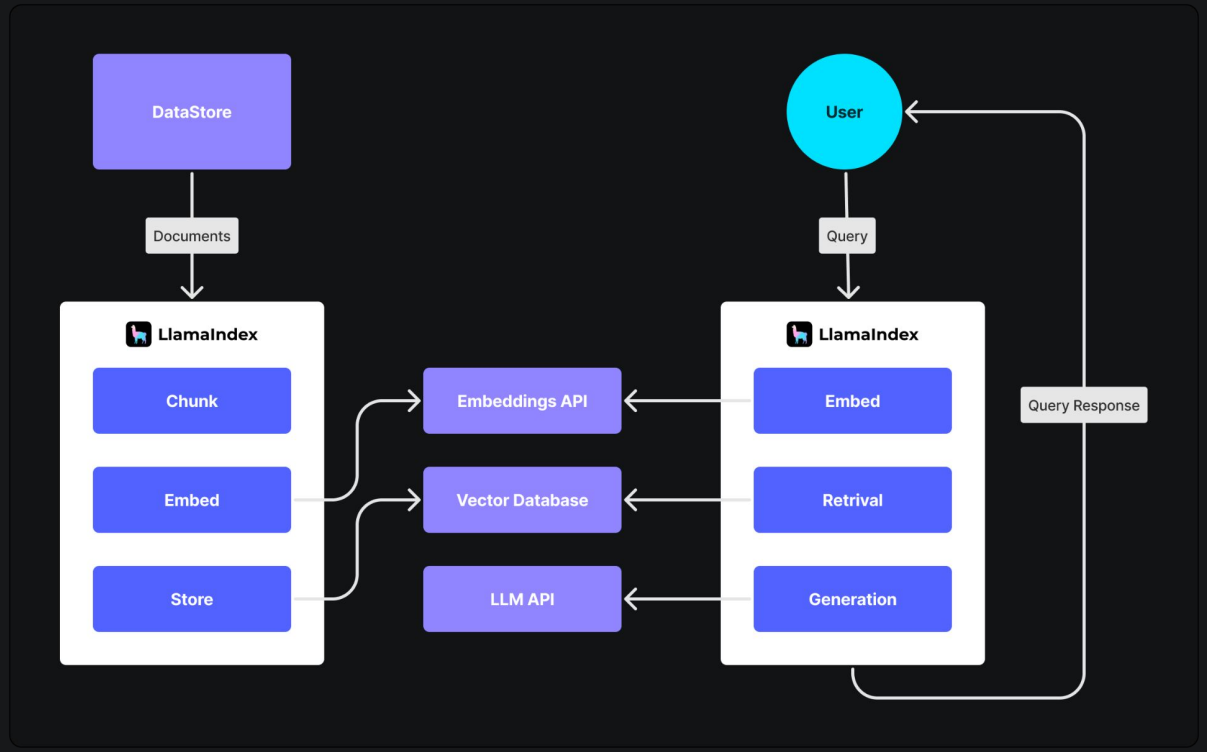
User Cost

Poor performance & privacy

LlamaIndex


Data
Embedding
LLMs
Vectors
Evals
Q&A
Structured Extraction
Chat
Semantic Search
Agents

LlamaIndex Abstractions



What Does RAG With LlamaIndex Cost? 🧐

Developer Cost

Fewer technologies to learn

Financial Cost

Three services to pay for

Scaling Cost

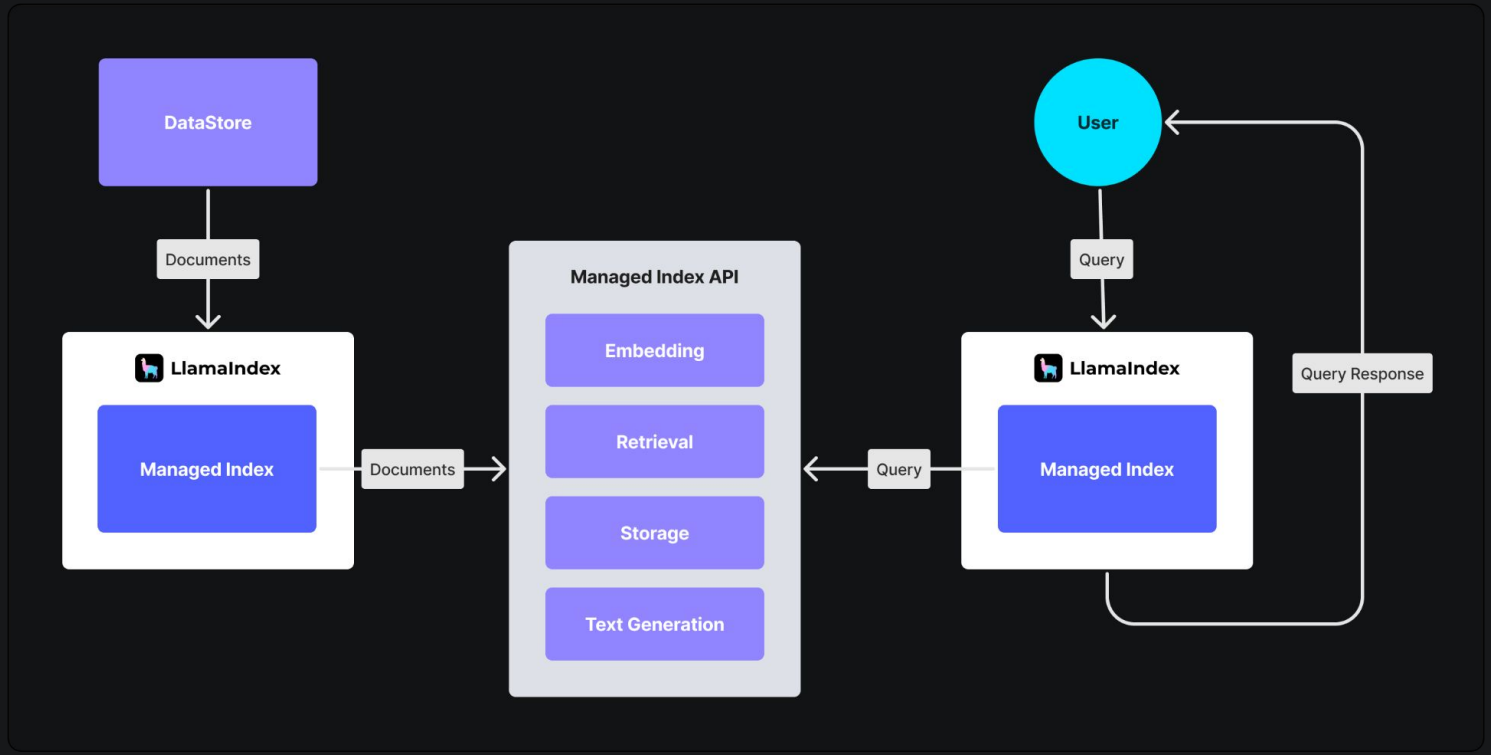
Many points of failures

User Cost

Poor performance & privacy

Let’s see this in code!

Managed Index

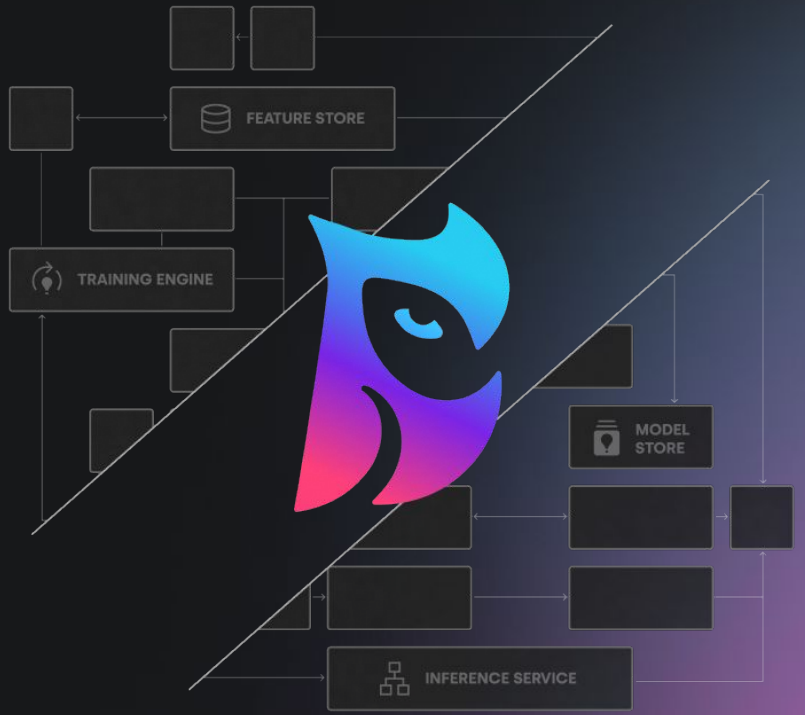


PostgresML

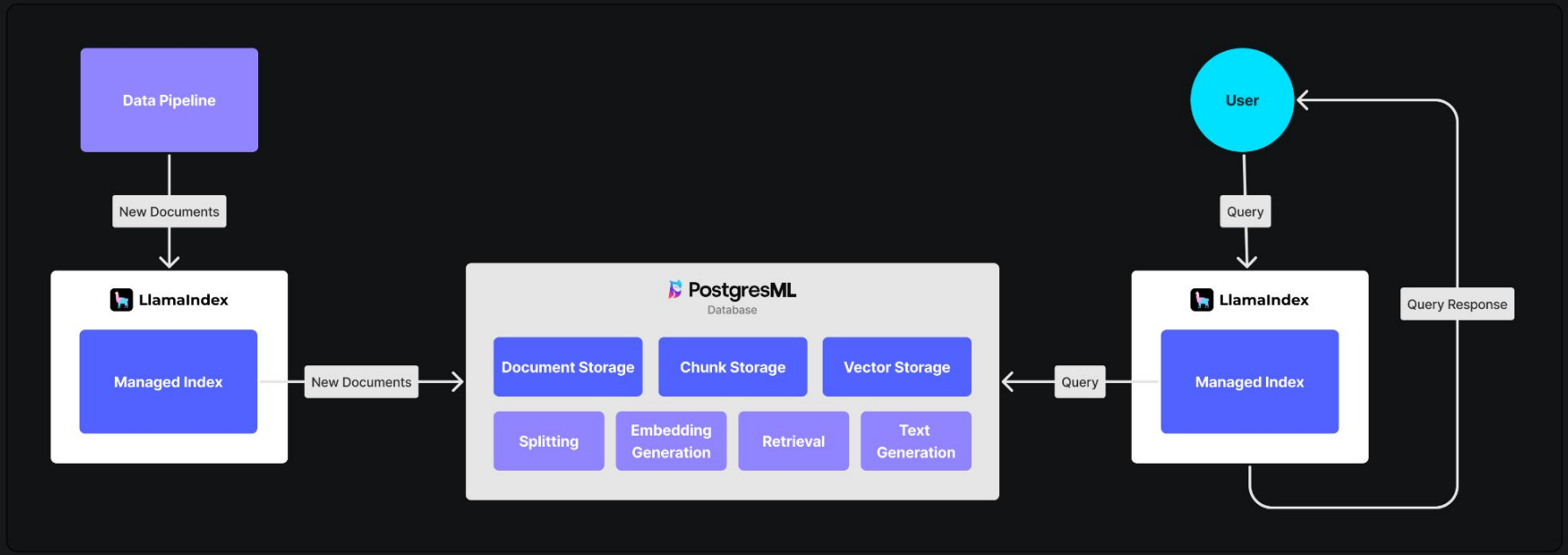
Less is More with PostgresML.

An open source machine learning platform built inside Postgres that provides:

- Embedding Generation, Storage and Retrieval
- Text Generation
- ...and More



How PostgreML Work With LlamaIndex



Let’s see this in some more code!

How PostgresML Improves 😝

Developer Cost

Fewer technologies to learn

Financial Cost

One service

Scaling Cost

One point of failure

User Cost

None

Join us

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

We’re hiring:

Email: **Montana@postgresml.org**