zilliz

RAG without OpenAI - Milvus, Mixtral, OctoAI, and LangChain

Yujian Tang | Zilliz

Speaker

Yujian Tang

Senior Developer Advocate, Zilliz

yujian@zilliz.com
https://www.linkedin.com/in/yujiantang
https://www.twitter.com/yujian_tang

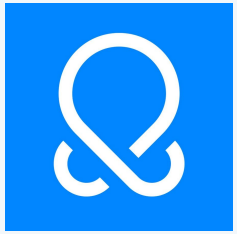
CONTENTS

01

Overview

Tech Stack

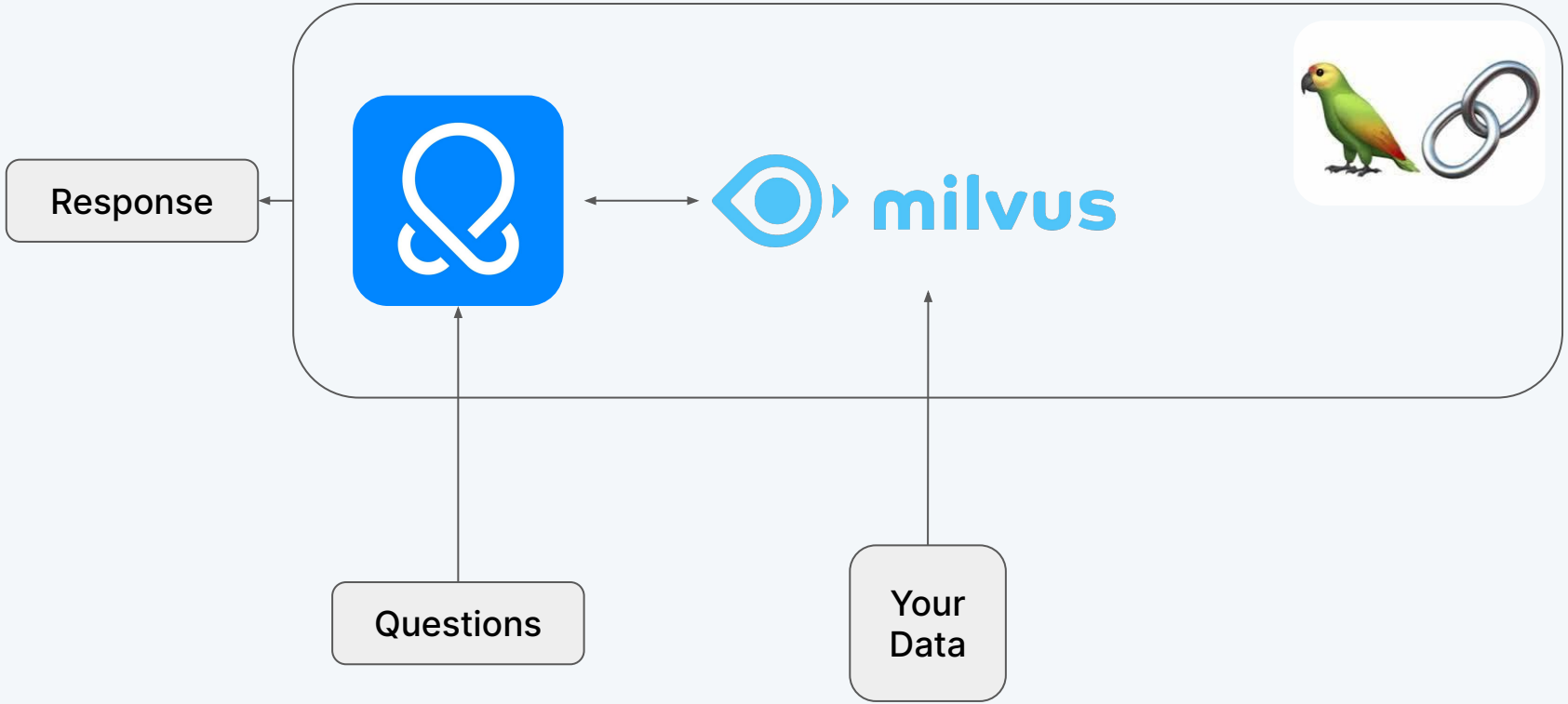
Milvus



Mixtral via OctoAI

LangChain

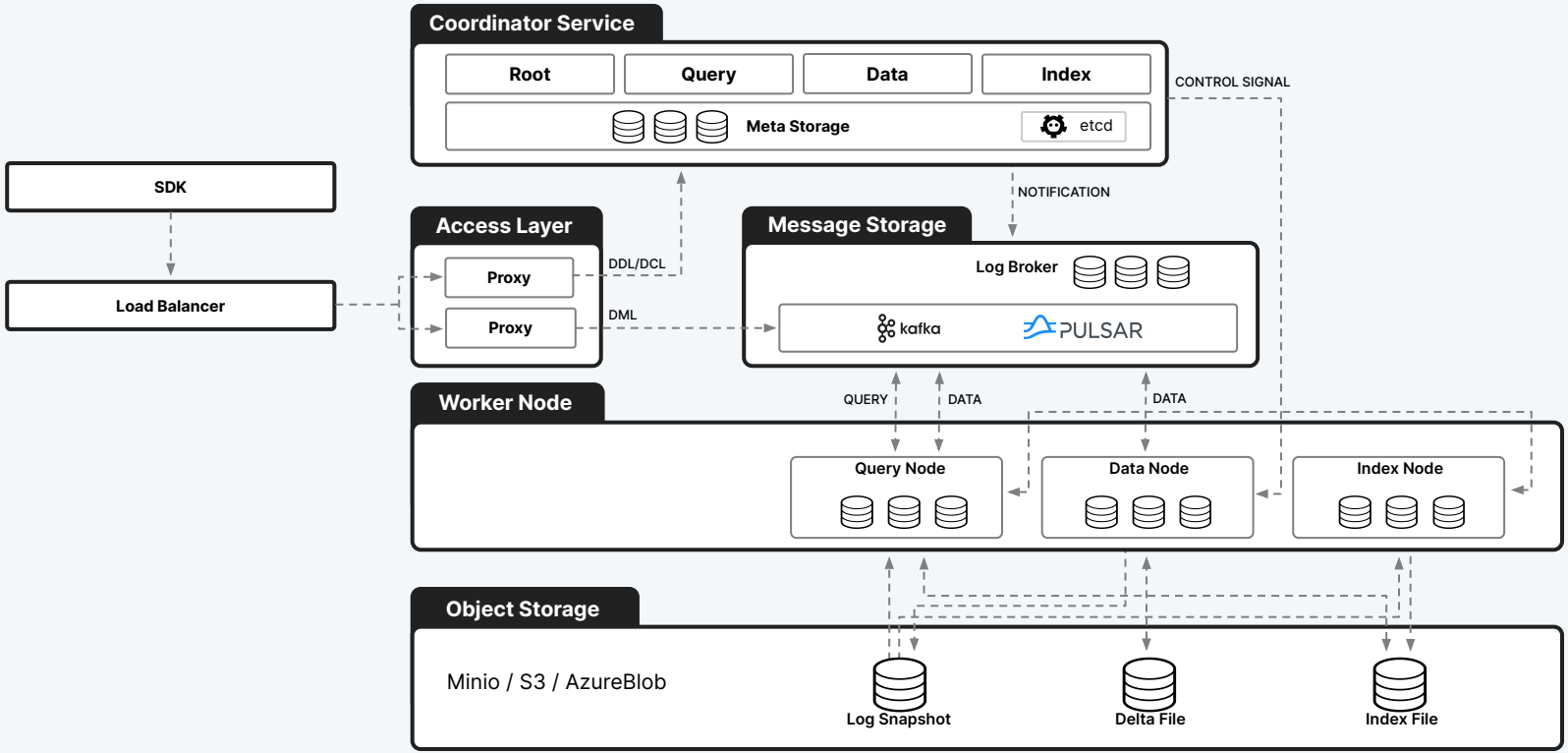
Architecture



02

Milvus

High-level overview of Milvus' Architecture

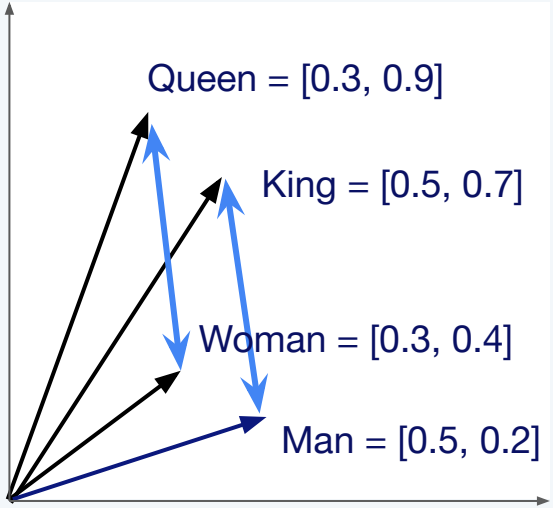


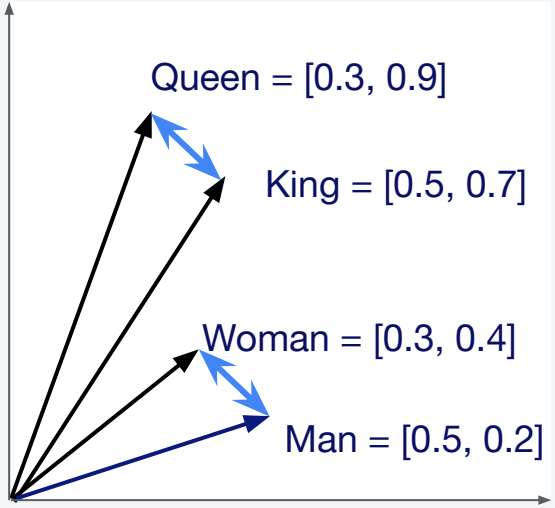
What Does Vector Data Look Like?

"id": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-9275854
"embedding": [-0.042092223,-0.0154002765,-0.014588429,-0.031147376,0.03801204,0.013369046,0
"date": "2023-06-01"
"paragraph": "We define an anomaly as follows:"
"reading_time": "11"
"subtitle": "A guide for the implementation of an anomaly..."
"publication": "Towards Data Science"
"responses": "1"
"article_url": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-
"title": "Detection of Credit Card Fraud with an Autoencoder"
"claps": "229"

↑ Hide 6 fields Vector search

Semantic Similarity





Queen - Woman + Man = King

Queen = [0.3, 0.9]
\- Woman = [0.3, 0.4]

[0.0, 0.5]
\+ Man = [0.5, 0.2]

King = [0.5, 0.7]

Image from Sutor et al

Milvus lets us compare unstructured data at scale

03

Mixtral

What is Mixtral?

- Mixture of Experts
- 8 “experts”
- 7B parameters each
- English, French, Italian, German and Spanish
- 32k Tokens
- 2 experts active at a time

04

OctoAI

What is OctoAI?

- Generative AI as API endpoints
- Media Gen
  - Stable Diffusion, Control Net
- Text Gen
  - Mixtral, Gemma, Smaug, Llama
- Compute
  - Bring Your Own Model

What is OctoAI?

Text Gen via Mixtral

Hyperparameters:

- Temperature
- Top P
- Max Tokens

05

LangChain

What is LangChain

- LLM Chaining Framework
- Plugins with most popular tools
- Focus on “chaining” results or “orchestration”

06

Demo

Give Milvus a Star!

Start building
with Zilliz Cloud today!
zilliz.com/cloud