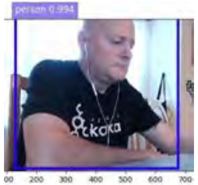
Enriching Generative AI as Events in Real-Time Streaming Pipelines

Tim Spann
Principal Developer Advocate

May 2024



CONF42

© 2024 Tim Spann All rights reserved.

https://github.com/tspannhw

DZone REFCARDS TREND REPORTS

Top IoT Experts

Tim Spann

Principal Developer Advocate, Cloudera

https://github.com/tspannhw/SpeakerProfile/ Tim Spann is a Principal Developer Advocate in Data In Motion for Cloudera. He works with Apache NiFi, Apache Pulsar, Apache...

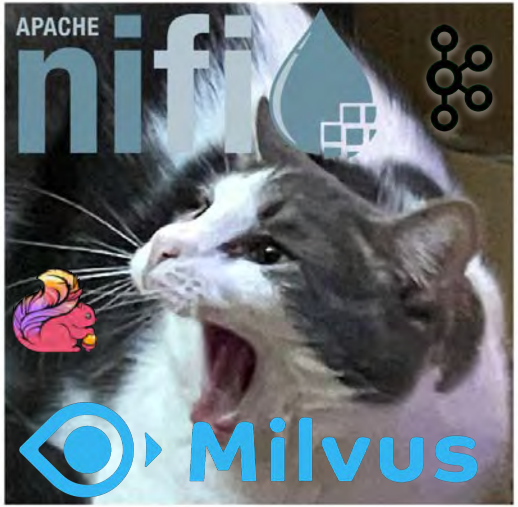
Timothy Spann

Principal Developer Advocate

https://medium.com/@tspann

© 2024 Tim Spann All rights reserved

FLaNK-AIM Stack Weekly



towhee

https://bit.ly/32dAJft

https://www.meetup.com/futureofdata-princeton/

This week in Milvus, Towhee, Attu,Apache NiFi, Apache Flink, Apache Kafka, ML, AI, Apache Spark, Apache Iceberg, Python, Java, LLM, GenAI, Vector DB and Open Source friends.

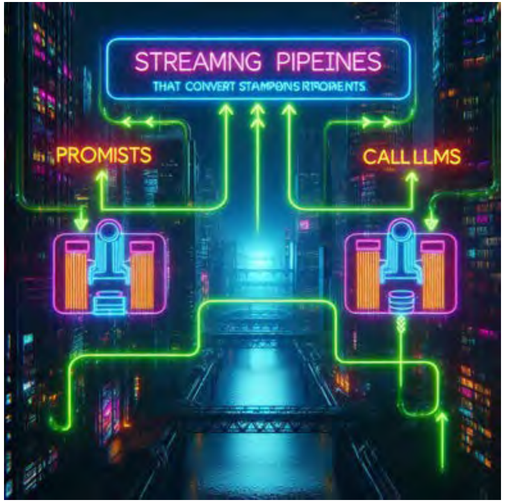
© 2024 Tim Spann All rights reserved.

towhee

Milvus

Let’s build streaming pipelines that convert streaming events into prompts and call LLMs and process the results.

LET'S BUILD STREAMING PIPINES
LLMS
CONVER THE STREAMING EVETS INTO PROMTS
CACLL LLMS AND POCESTHE RESULTS.



© 2024 Tim Spann All rights reserved.

Unstructured Data is Everywhere

Unstructured data is any data that does not conform to a predefined data model.

By 2025, IDC estimates there will be 175 zettabytes of data globally (that's 175 with 21 zeros), with **80% of that data being unstructured**.

Text

Images

Video

and more!

BEFORE MILVUS

© 2024 Tim Spann All rights reserved.

Milvus

https://milvus.io/milvus-demos/reverse-image-search/

© 2024 Tim Spann All rights reserved.

milvus

LF AI & DATA

Easy Setup

Pip-install to start coding in a notebook within seconds.

Reusable Code

Write once, and deploy with one line of code into the production environment

Integration

Plug into OpenAI, Langchain, LlmaIndex, and many more

Feature-rich

Dense & sparse embeddings, filtering, reranking and beyond

Milvus is **an open-source vector database** for **GenAI** projects. Pip-install on your laptop, plug into popular AI dev tools, and push to production with a single line of code.

27K+	2,600+	25M+	250+
GitHub Stars	Forks	Downloads	Contributors

We’ve built technologies for various types of use cases

Index Types

Offer a wide range of **15 indexes** support, including popular ones like HNSW, PQ, Binary, Sparse, DiskANN and GPU index

Empower developers with tailored search optimizations, catering to performance, accuracy and cost needs

Search Types

Support multiple types such as **top-K ANN, Range ANN, sparse & dense, multi-vector, grouping,** and metadata **filtering**

Enable query flexibility and accuracy, allowing developers to tailor their information retrieval needs

Multi-tenancy

Enable **multi-tenancy** through collection and partition management

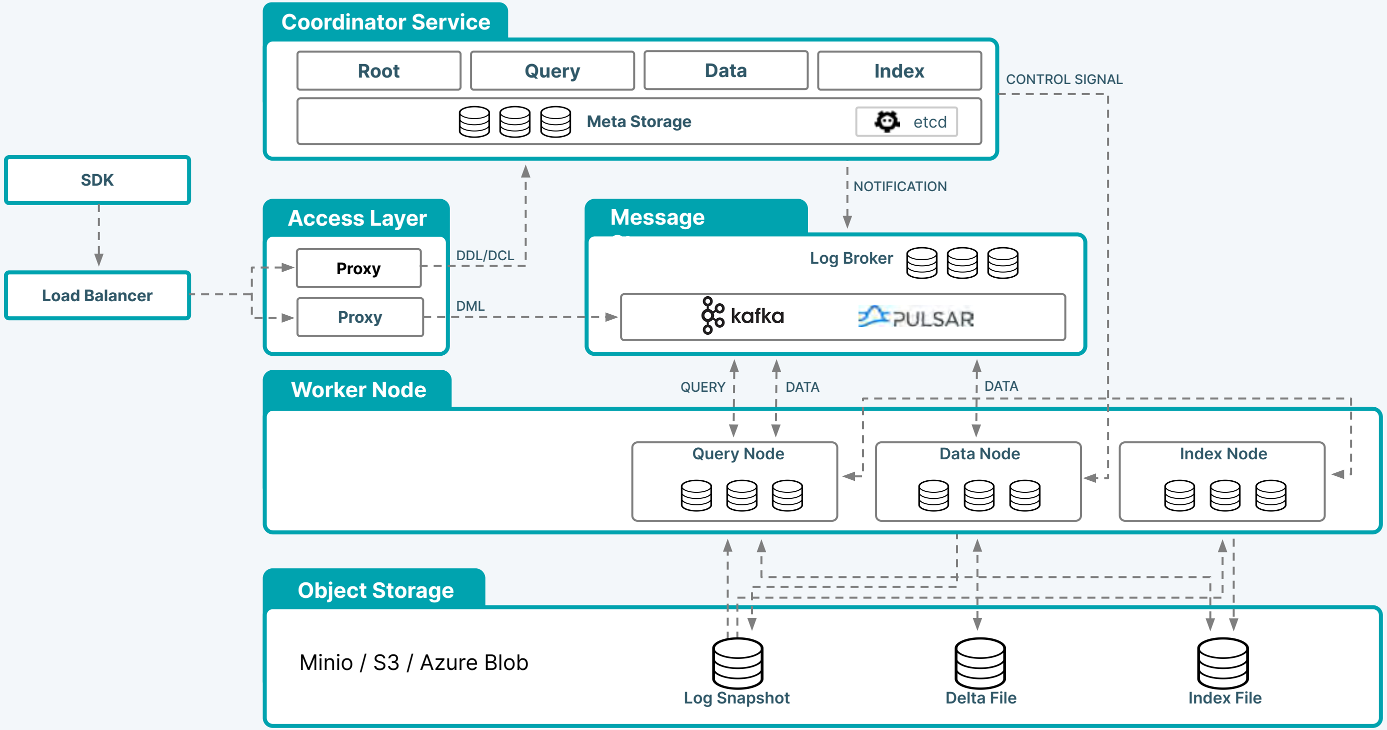
Allow for efficient resource utilization and customizable data segregation, ensuring secure and isolated data handling for each tenant

Compute Types

Designed for various compute powers, such as **AVX512, Neon for SIMD, quantization cache-aware optimization** and **GPU**

Leverage strengths of each hardware type, ensuring high-speed processing and cost-effective scalability for different application needs

Milvus’ fully distributed architecture is designed scalability and performance



Common AI Use Cases

LLM Augmented Retrieval

Expand LLMs' knowledge by incorporating external data sources into LLMs and your AI applications.

Recommender System

Match user behavior or content features with other similar behaviors or features to make effective recommendations.

Text/ Semantic Search

Search for semantically similar texts across vast amounts of natural language documents.

Image Similarity Search

Identify and search for visually similar images or objects from a vast collection of image libraries.

Video Similarity Search

Search for similar videos, scenes, or objects from extensive collections of video libraries.

Audio Similarity Search

Find similar audios from massive amounts of audio data to perform tasks such as genre classification, or recognize speech.

Molecular Similarity Search

Search for similar substructures, superstructures, and other structures for a specific molecule.

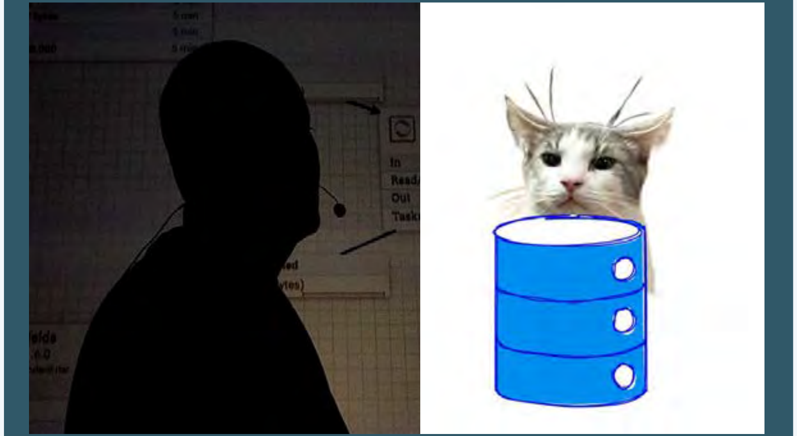
Question Answering System

Interactive QA chatbot that automatically answers user questions

Multimodal Similarity Search

Search over multiple types of data simultaneously, e.g. text and images

Milvus Features



Scalable and Elastic Architecture

Diverse Index Support

Versatile Search Capabilities

Tunable Consistency

Multi-Tenancy

Hardware-Accelerated Compute Support

Python, Java, Golang, NodeJS

Milvus Lite, K8, Zilliz Cloud, Docker

GEN AI

DataFlow Pipelines Can Help

External Context Ingest

Ingesting, routing, clean, enrich, transforming, parsing, chunking and vectorizing structured, unstructured, semistructured, binary data and documents

Prompt engineering

Crafting and structuring queries to optimize LLM responses

Context Retrieval

Enhancing LLM with external context such as Retrieval Augmented Generation (RAG)

Roundtrip Interface

Act as a Discord, REST, Kafka, SQL, Slack bot to roundtrip discussions

UNSTRUCTURED DATA WITH NIFI

- **Archives** - tar, gzipped, zipped, …
- **Images** - PNG, JPG, GIF, BMP, …
- **Documents** - HTML, Markdown, RSS, PDF, Doc, RTF, Plain Text, …
- **Videos** - MP4, Clips, Mov, Youtube URL…
- **Sound** - MP3, …
- **Social / Chat** - Slack, Discord, Twitter, REST, Email, …
- **Identify Mime Types, Chunk Documents, Store to Vector Database**
- **Parse Documents -** HTML, Markdown, PDF, Word, Excel, Powerpoint

NiFi 2.0.0 Features

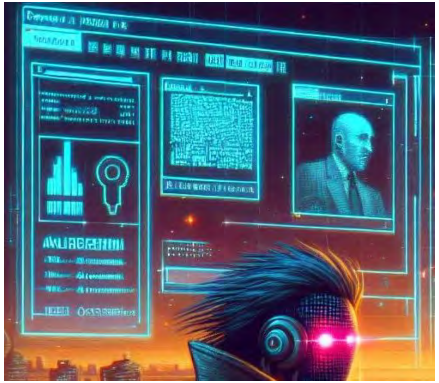
- Python Integration
- Parameters
- JDK 21+
- JSON Flow Serialization
- Rules Engine for Development Assistance
- Run Process Group as Stateless
- flow.json.gz

https://cwiki.apache.org/confluence/display/NIFI/NiFi+2.0+Release+Goals

https://medium.com/cloudera-inc/getting-ready-for-apache-nifi-2-0-5a5e6a67f450

nifi

Python Processors



Address To Lat/Long

- Python 3.10+
- geopy Library
- Nominatim
- OpenStreetMaps (OSM)
- openstreetmap.org/copyright
- Returns as attributes and JSON file
- Works with partial addresses
- Categorizes location
- Bounding Box

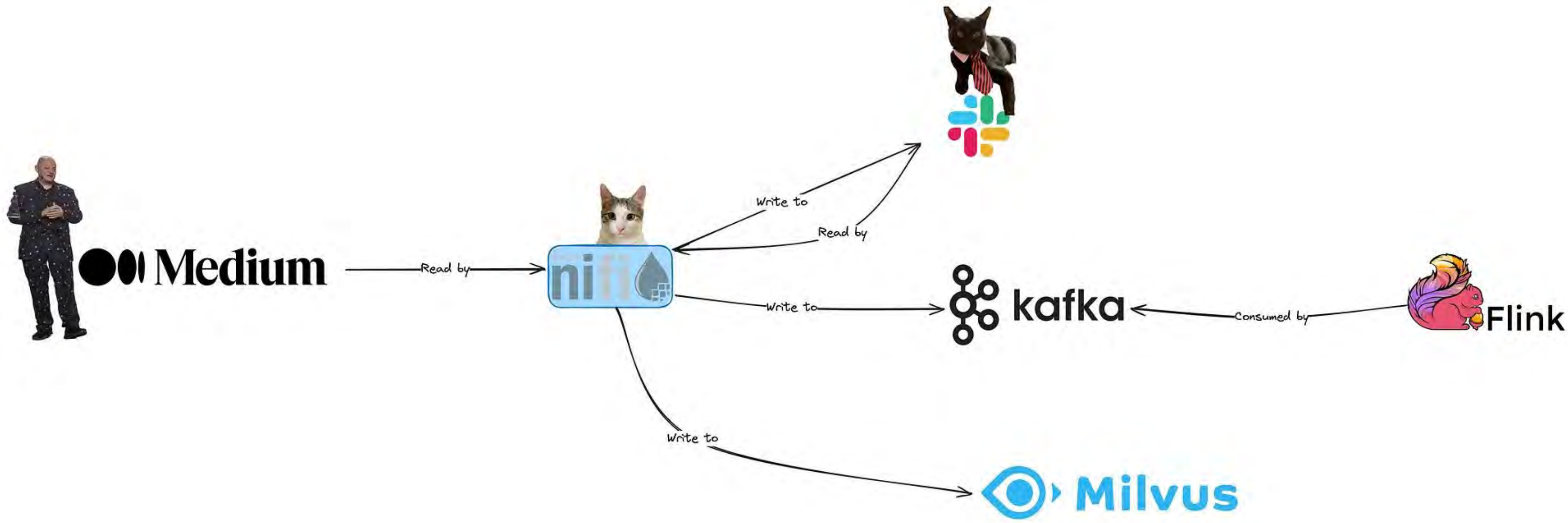
https://github.com/tspannhw/FLaNKAI-Boston

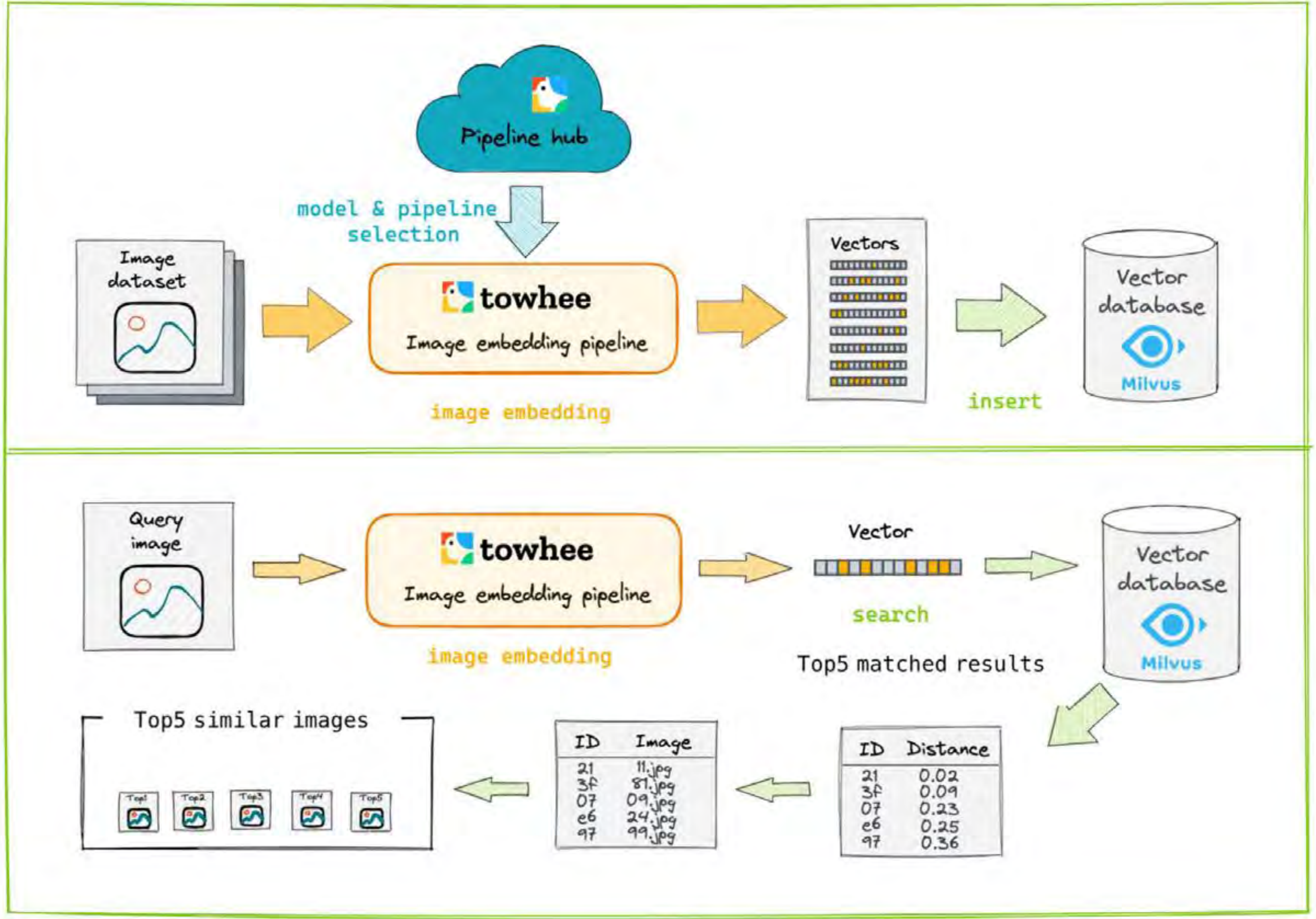
DEMOS

CYBERPUNK
CYBEERPUNK CODELLESS DITATRASE NIFI STREAM

GEN AI NEEDS STREAMING
imgflip.com

Building a Milvus Connector For NiFi



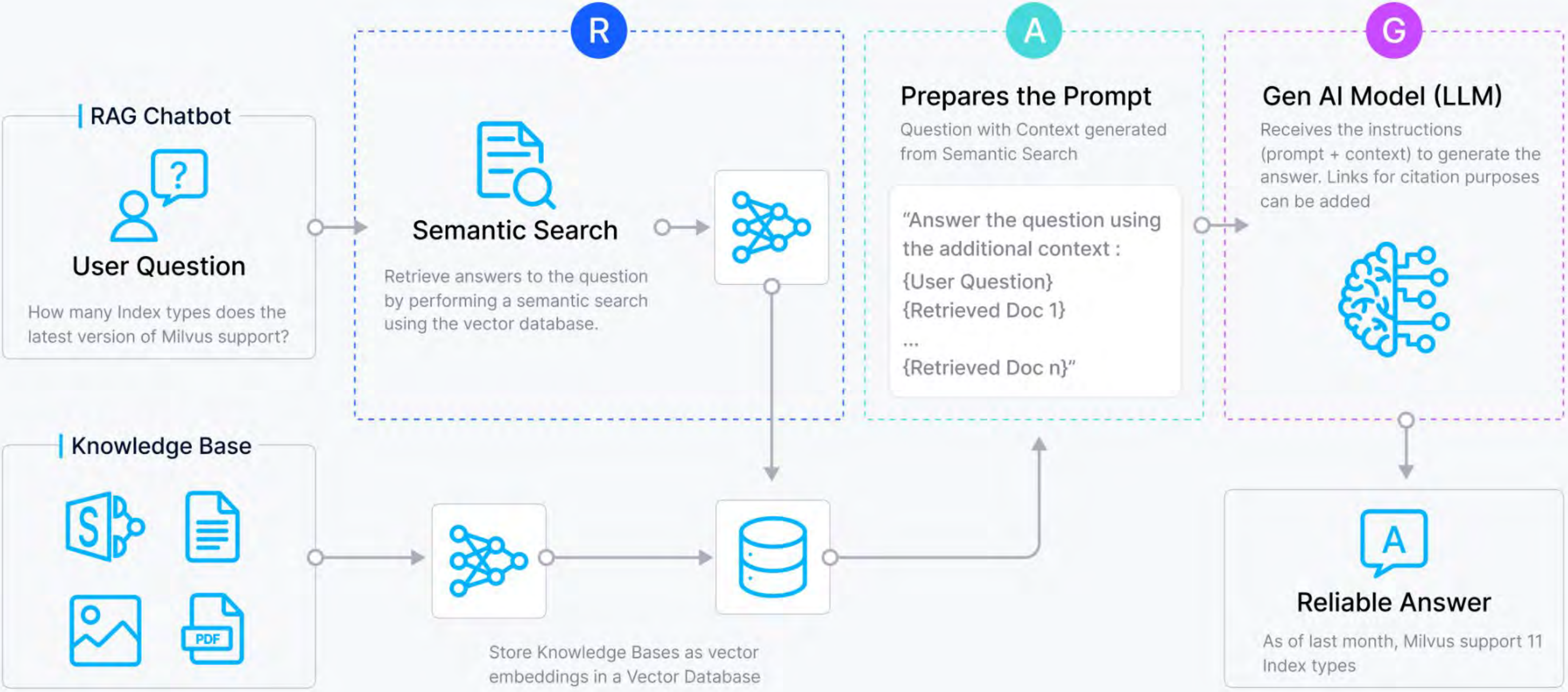


ID	Distance
21	0.02
3f	0.09
07	0.23
e6	0.25
97	0.36

ID	Image
21	11.jpg
3f	81.jpg
07	09.jpg
e6	24.jpg
97	99.jpg

milvus

Retrieval-Augmented Generation RAG Chatbot



RAG Chatbot

User Question

How many Index types does the latest version of Milvus support?

R

Semantic Search

Retrieve answers to the question by performing a semantic search using the vector database.

A

Prepares the Prompt

Question with Context generated from Semantic Search

"Answer the question using the additional context :
{User Question}
{Retrieved Doc 1}
...
{Retrieved Doc n}"

G

Gen AI Model (LLM)

Receives the instructions (prompt + context) to generate the answer. Links for citation purposes can be added

Knowledge Base

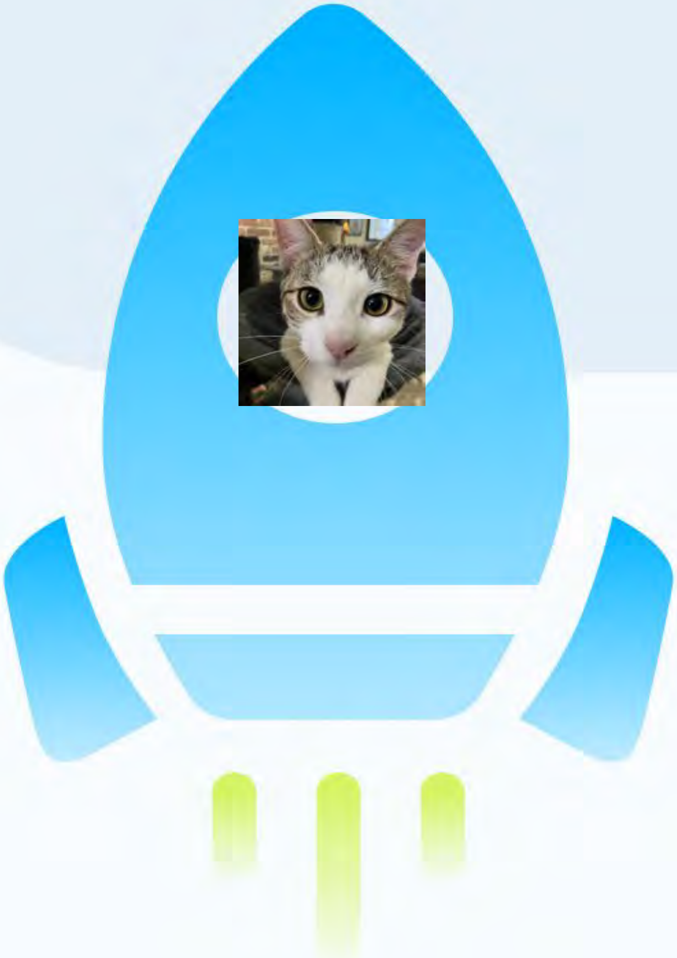
Store Knowledge Bases as vector embeddings in a Vector Database

Reliable Answer

As of last month, Milvus support 11 Index types

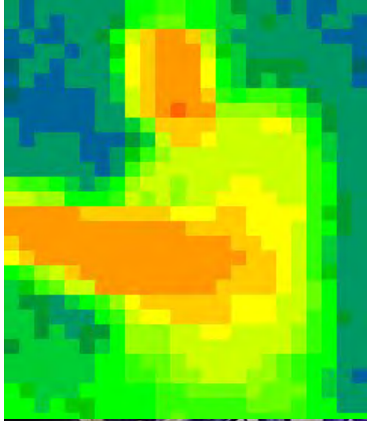
milvus

How To Get Started With Milvus



SPIRIT
Business
Cat
ADULT
Size Costume
ONE SIZE FITS MOST

AUTOMATE EVERYTHING
imgflip.com



Milvus

THANK YOU

REAL-TIME EVENTS

Why Use It?

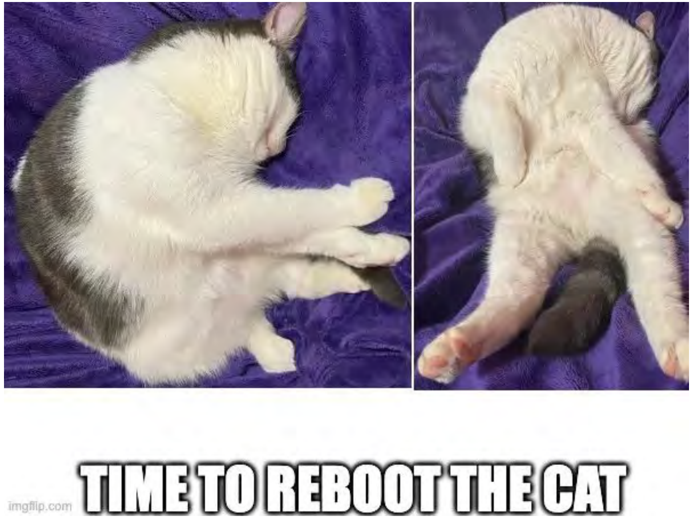
Open Source

Fast

Many Indexes

© 2024 Tim Spann All rights reserved.

Why?



LF AI & Data Foundation Graduate Project

Scalability and tunability to handle growing data volumes

Multi-tenancy and data isolation for efficient resource use and privacy

A comprehensive suite of APIs for diverse programming languages

User-friendly interfaces that simplify interaction with complex data.

© 2024 Tim Spann All rights reserved.