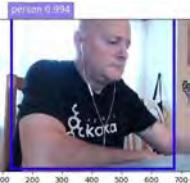


Enriching Generative AI as Events in Real-Time Streaming Pipelines

Tim Spann
Principal Developer Advocate

May 2024



CONF42
© 2024 Tim Spann All rights reserved.

FLaNK-AIM Stack Weekly



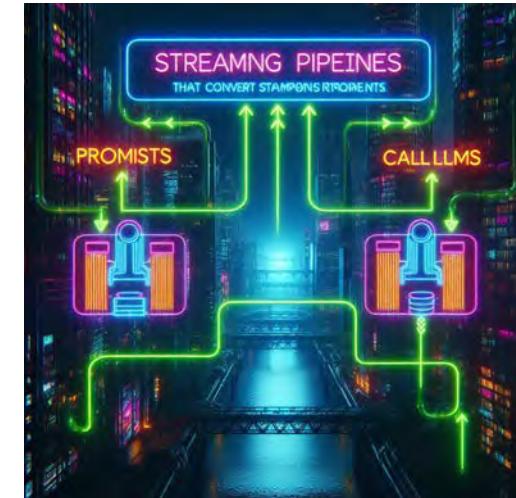
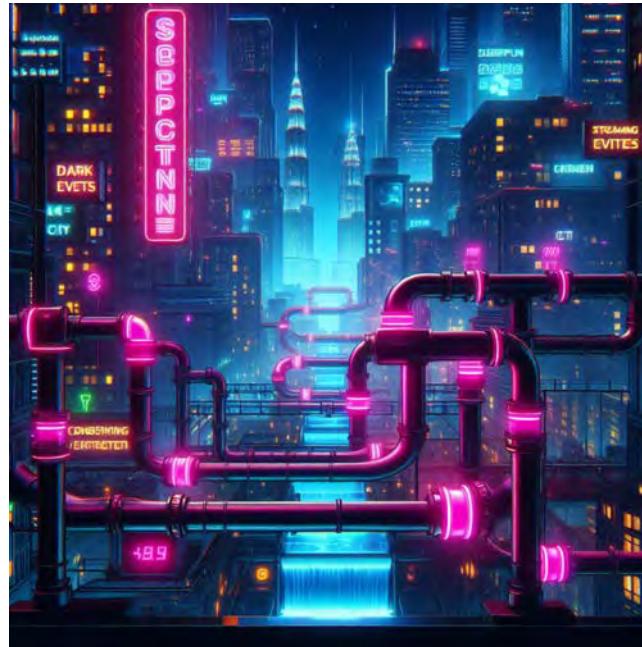
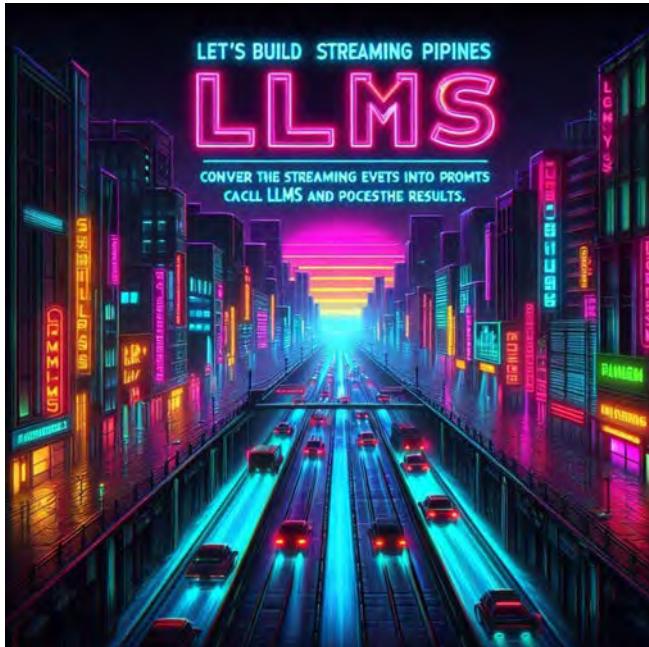
<https://bit.ly/32dAJft>

<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.



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



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!

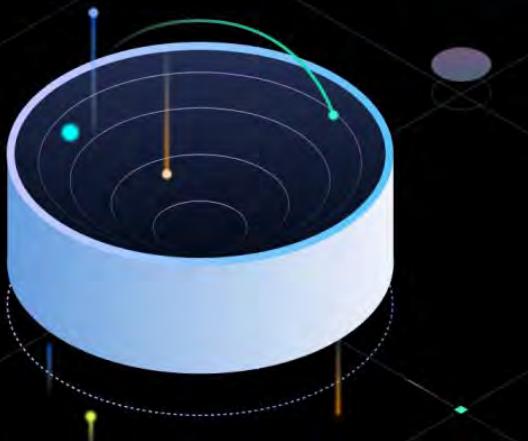
BEFORE MILVUS





DLF AI
& DATA

Vector database built for scalable
similarity search



igflip.com
JAKE-CLARK.TUMBLR

The Milvus logo, featuring a stylized eye icon followed by the word "Milvus" in a large, bold, blue sans-serif font.

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

© 2024 Tim Spain All rights reserved.



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, LlmalIndex, 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+

GitHub Stars

2,600+

Forks



25M+

Downloads



250+

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

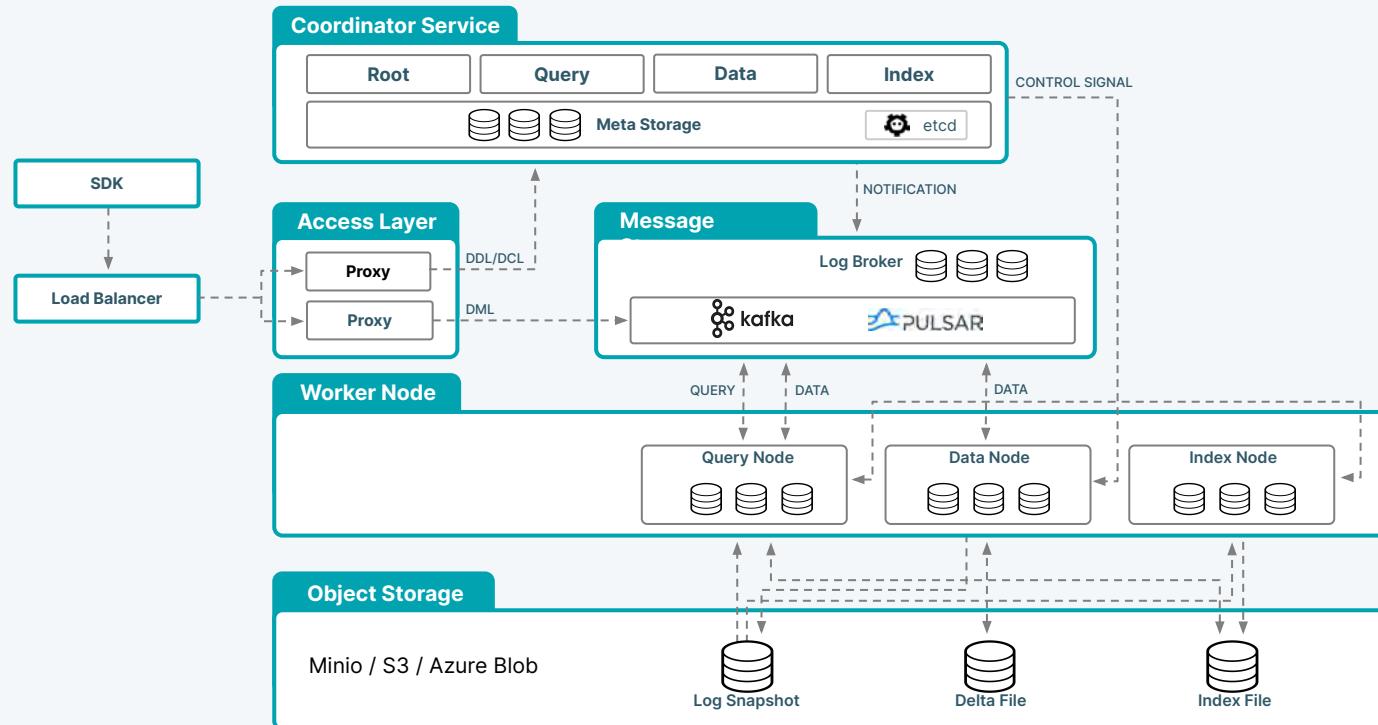


Compute Types

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

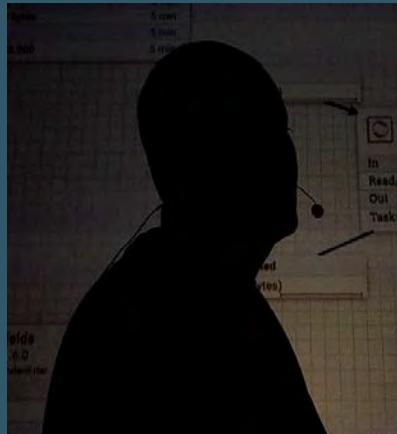
Interactive QA chatbot that automatically answers user questions



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



MILVUS



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>



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



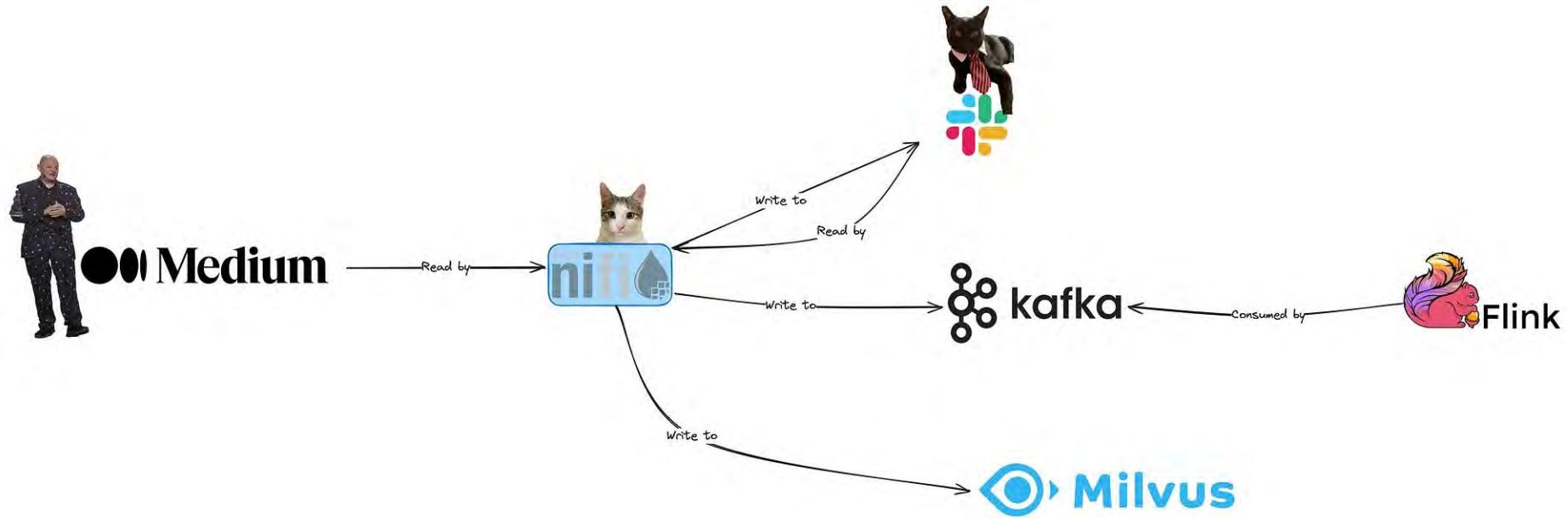
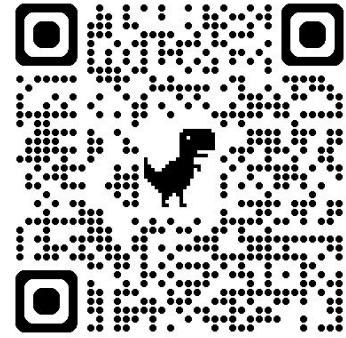
<https://github.com/tspannhw/FLaNKAI-Boston>

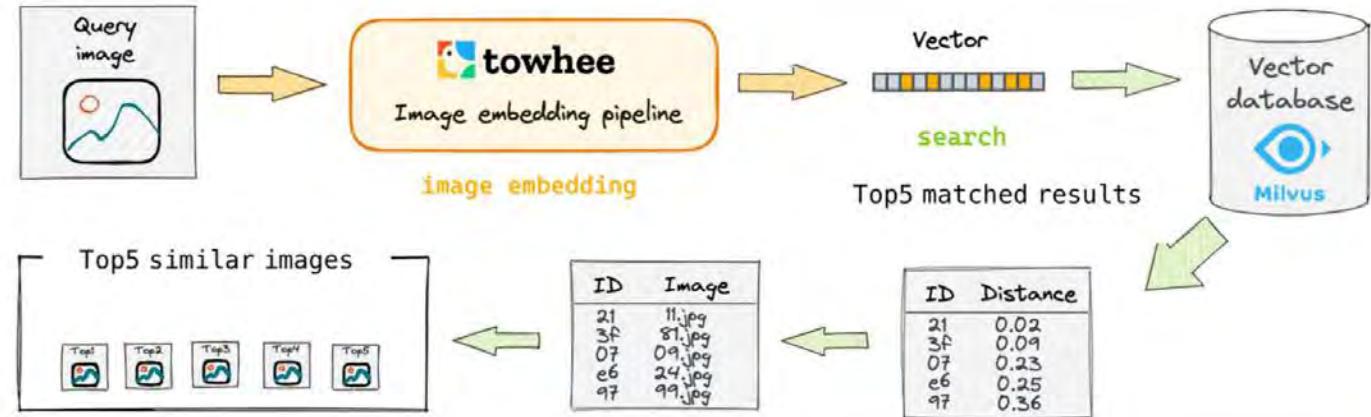
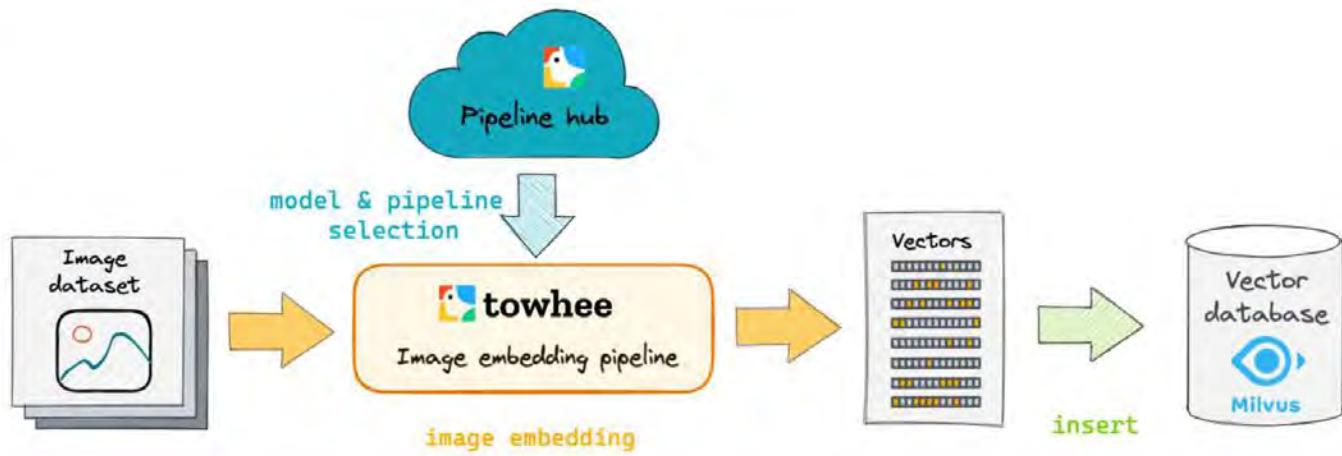


DEMOS

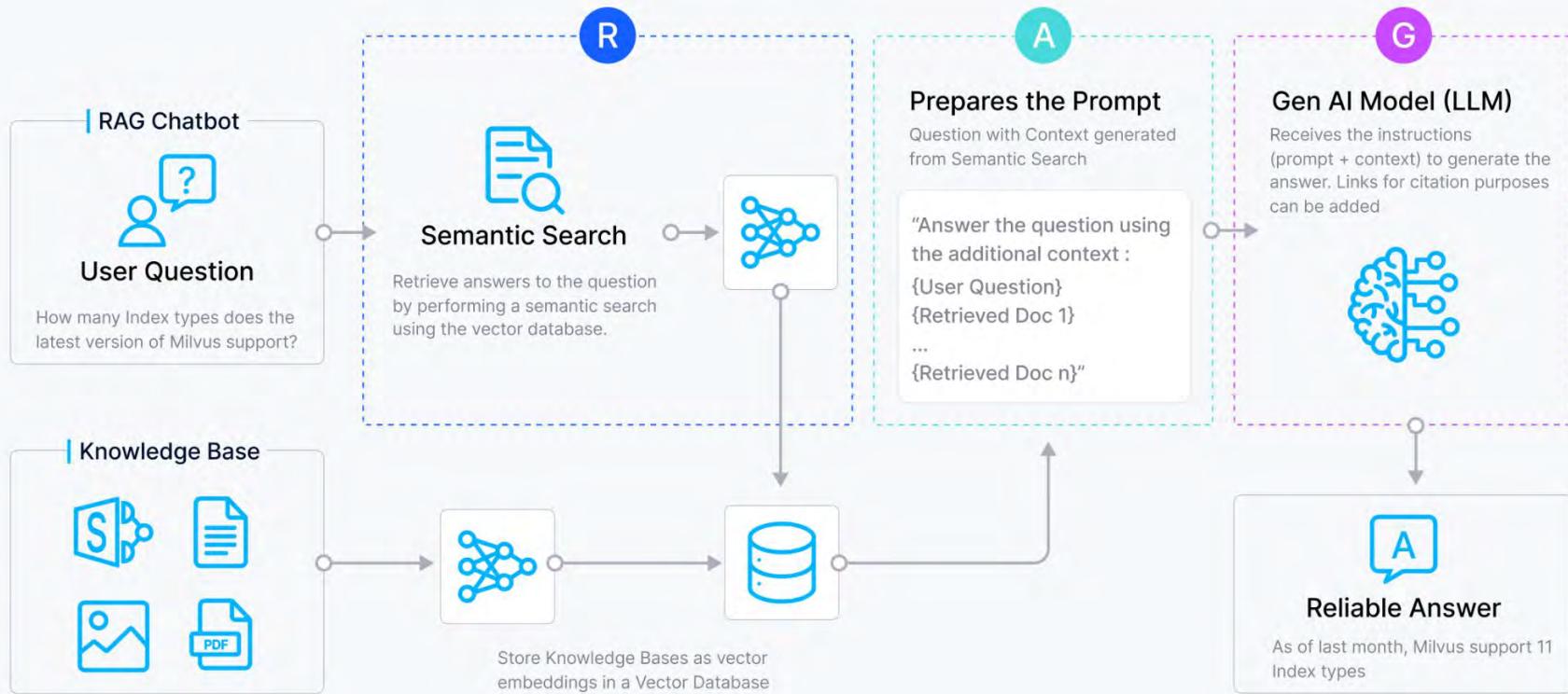


Building a Milvus Connector For NiFi



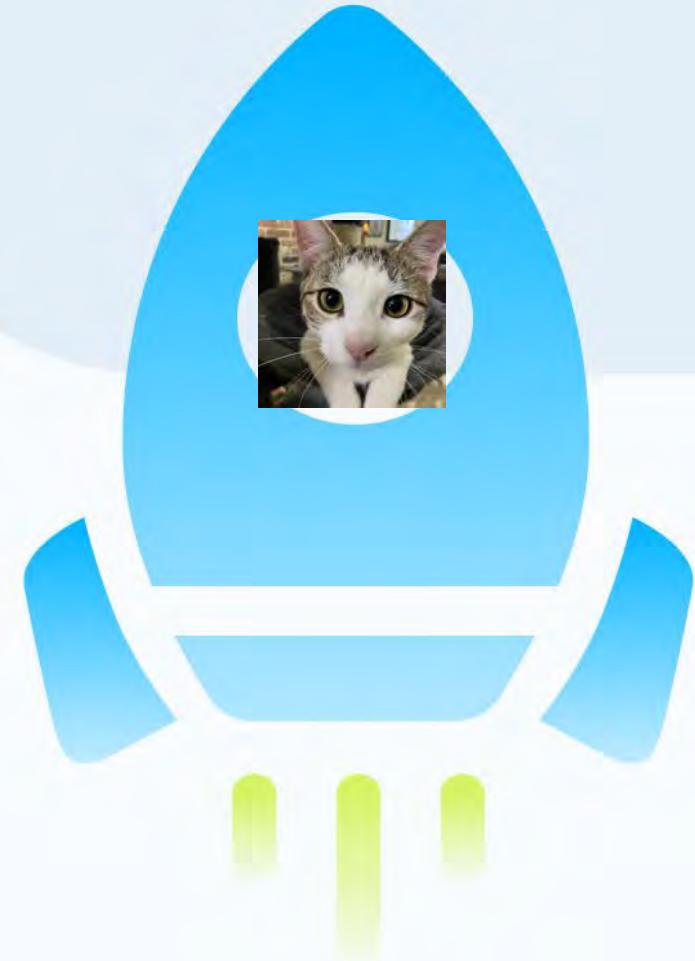
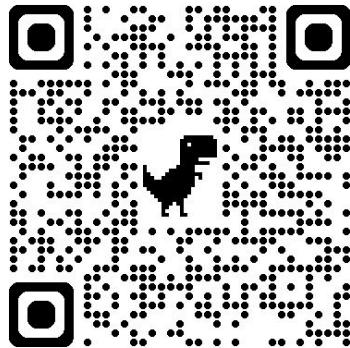


Retrieval-Augmented Generation RAG Chatbot





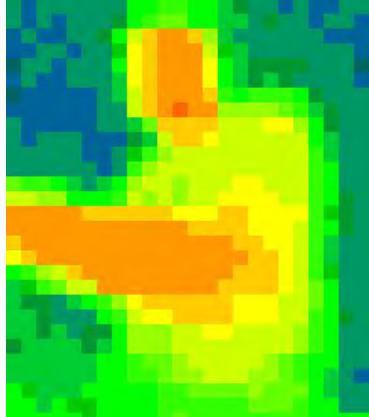
How To Get Started With Milvus







THON^oYOU



REAL-TIME EVENTS



Why Use It?



Open Source

Fast

Many Indexes



Why?



imgflip.com

TIME TO REBOOT THE CAT

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.