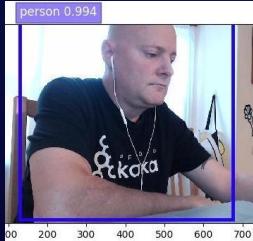




Building Real-Time Pipelines

Tim Spann
Principal Developer Advocate

June 12, 2024



Speaker



Tim Spann

Principal Developer Advocate, Zilliz

tim.spann@zilliz.com

<https://www.linkedin.com/in/timothyspann/>

<https://x.com/paasdev>

<https://github.com/tspannhw>



Unstructured Data Meetup @ New York



<https://www.meetup.com/unstructured-data-meetup-new-york/>

<https://www.meetup.com/pro/unstructureddata/>

From Unstructured Data to Vector Databases to ML to Generative AI to Deep Learning to Data Science



FLaNK-AIM Stack Weekly



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

<https://www.meetup.com/unstructured-data-meetup-new-york/>

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.



DLF 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, 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.



27.5K+

GitHub Stars

2,700+

Forks



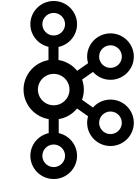
25M+

Downloads



250+

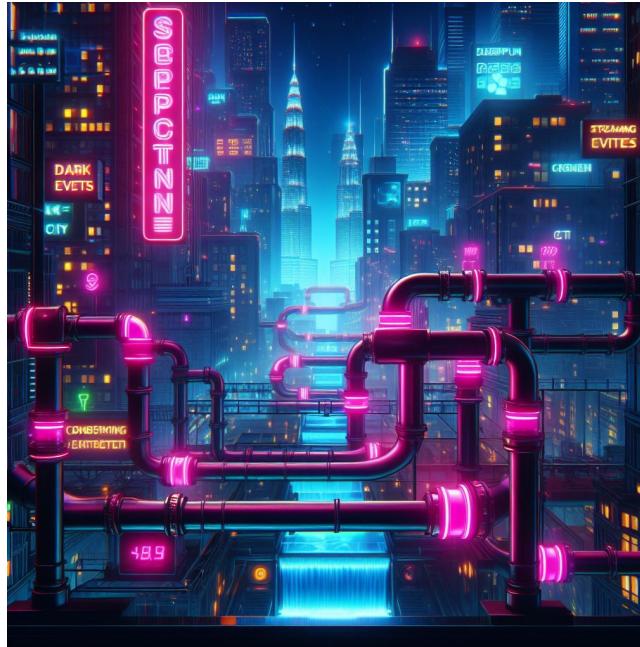
Contributors



Milvus



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.



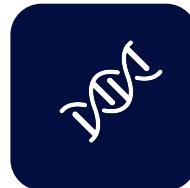
Text



Images



Video



and more!

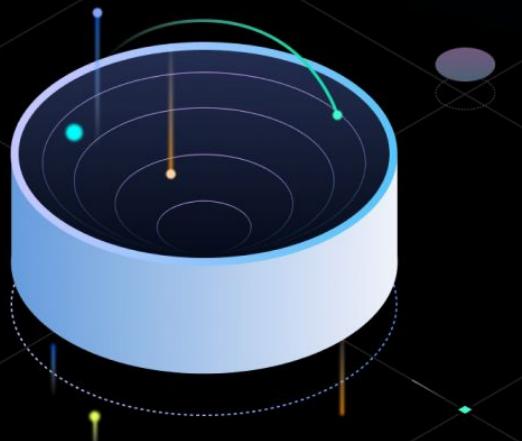
BEFORE MILVUS





DLF AI
& DATA

Vector database built for scalable
similarity search



Milvus

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

© 2024 Tim Spain All rights reserved.

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

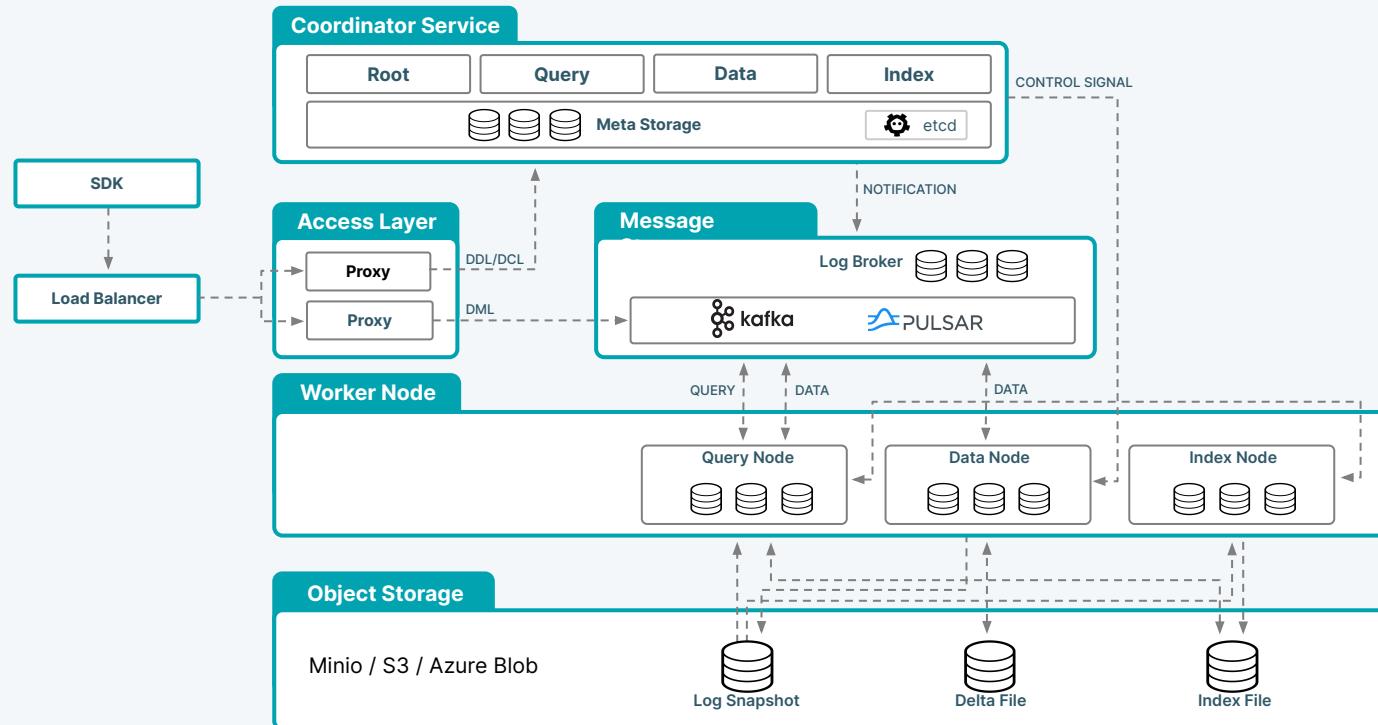


Compute Types

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

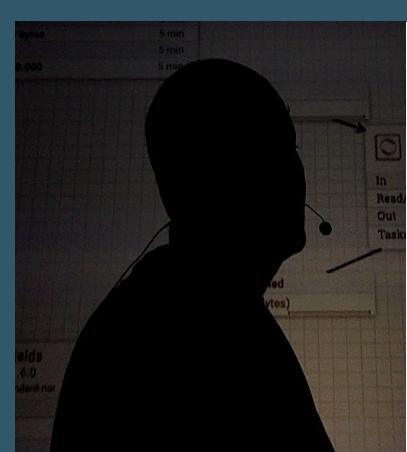
Interactive QA chatbot that automatically answers user questions



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**





TIME TO REBOOT THE CAT

imgflip.com

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>



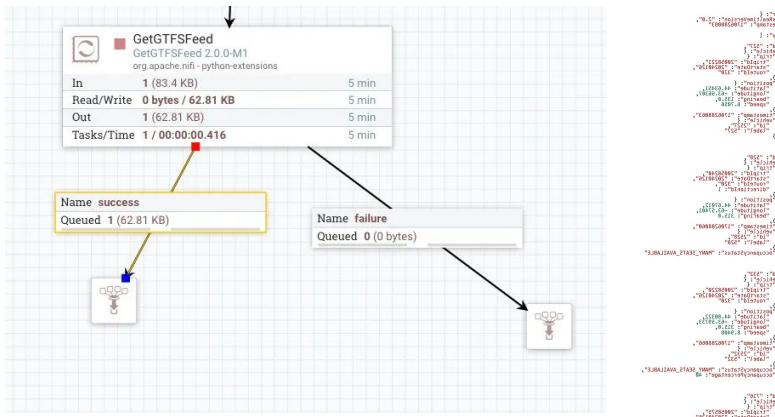
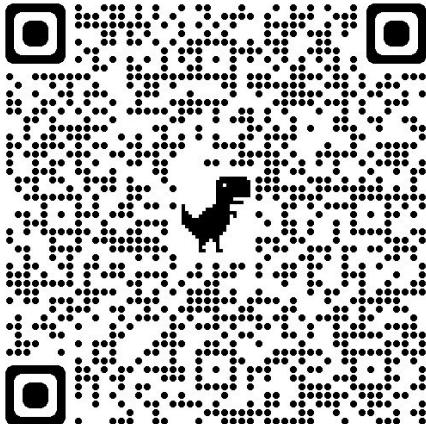
Python Processors





Get GTFS Data

- Python 3.10+
 - GTFS from Transit URL
 - Alerts, Trip Updates or Vehicle Positions
 - Returns JSON
 - `google.transit` and `google.protobuf`





Get Compound GTFS Data

- Python 3.10+
- GTFS to JSON

trip_update

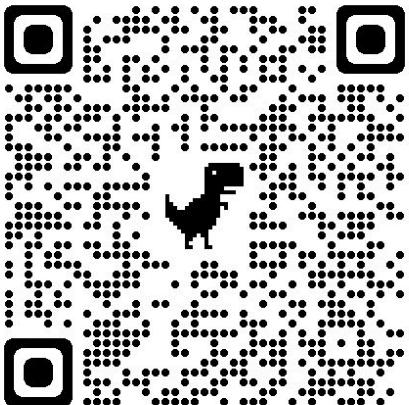
trip_update

vehicle

vehicle

alert

Reference parameter...



Processor Details | GetGTFSCompoundFeed 2.0.0-M2

Running

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMM

Required field

Property	Value
URL for GTFS Feed	Sensitive value set
API Key for header (MTA)	Sensitive value set
API Key for header name ex: (MTA)	x-api-key
Type for GTFS Feed	vehicle

<https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetGTFSCompoundFeed.py>



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



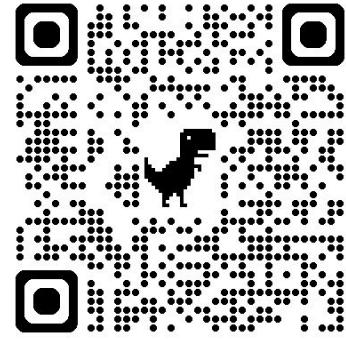
<https://github.com/tspannhw/FLaNKAI-Boston>



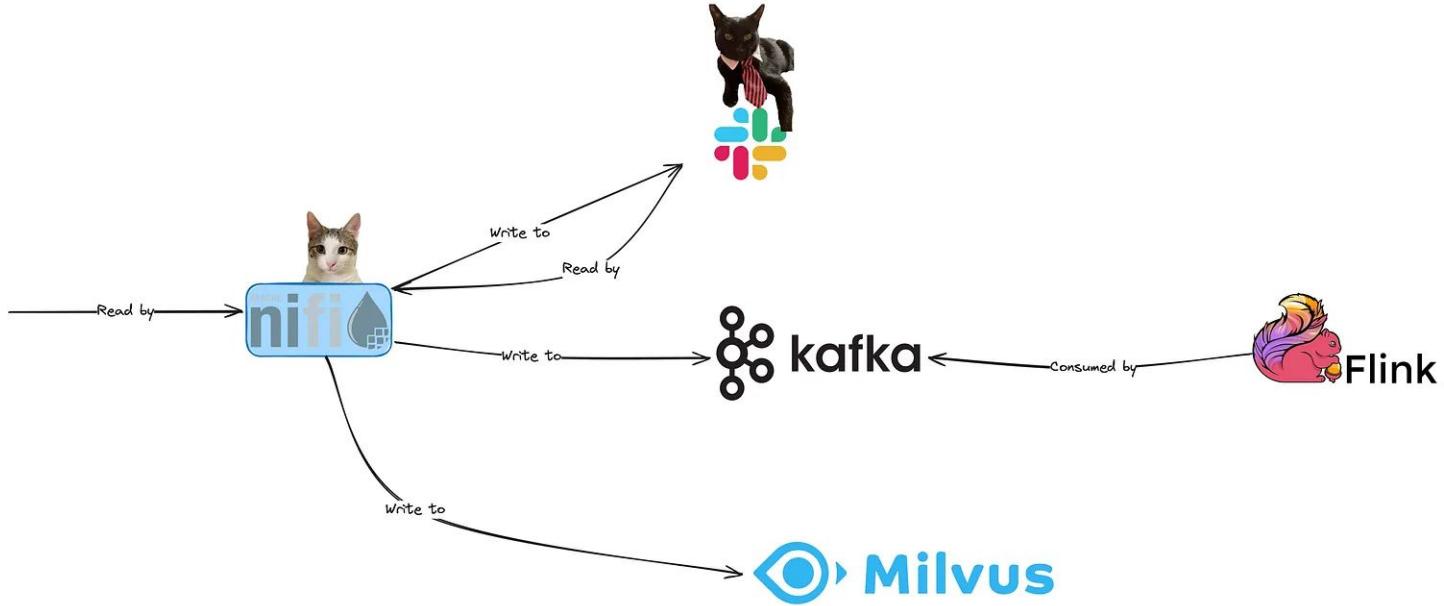
DEMOS

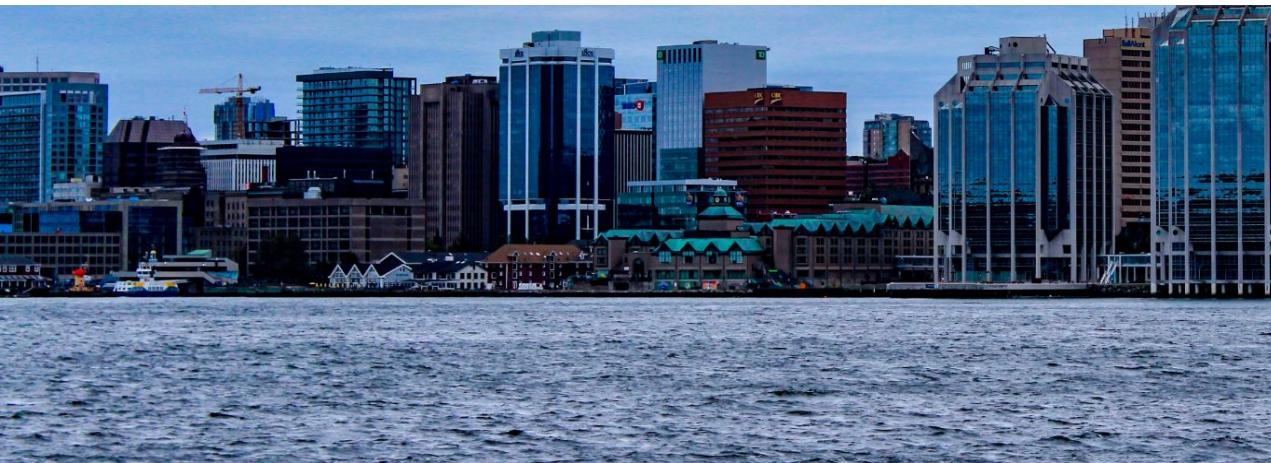
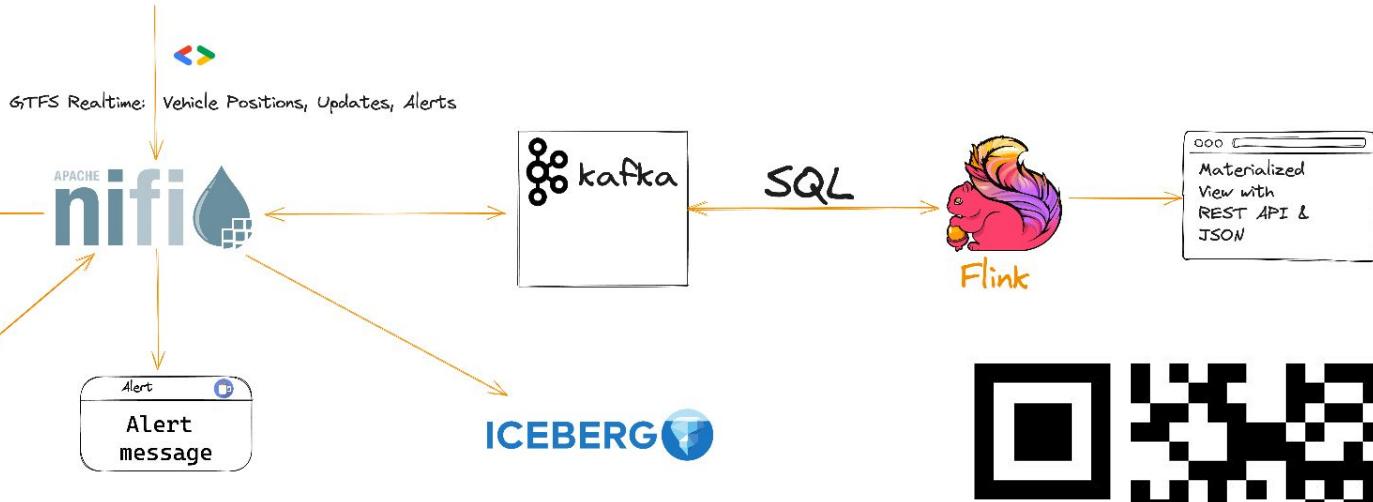


Building a Milvus Connector For NiFi



Medium

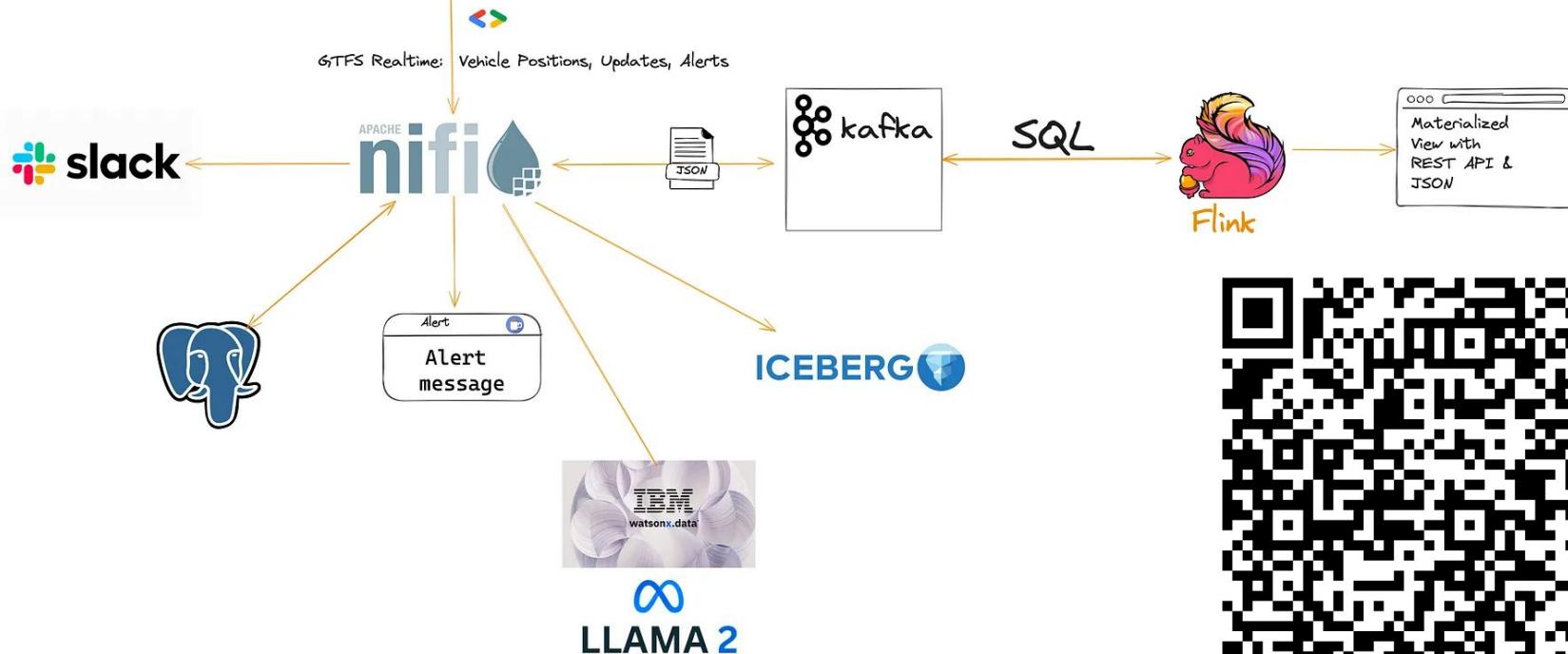


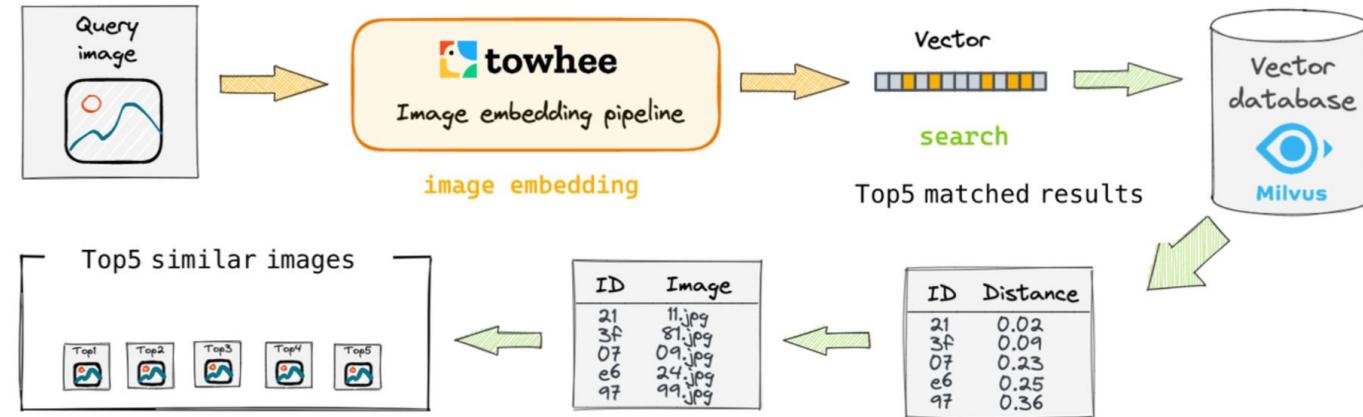
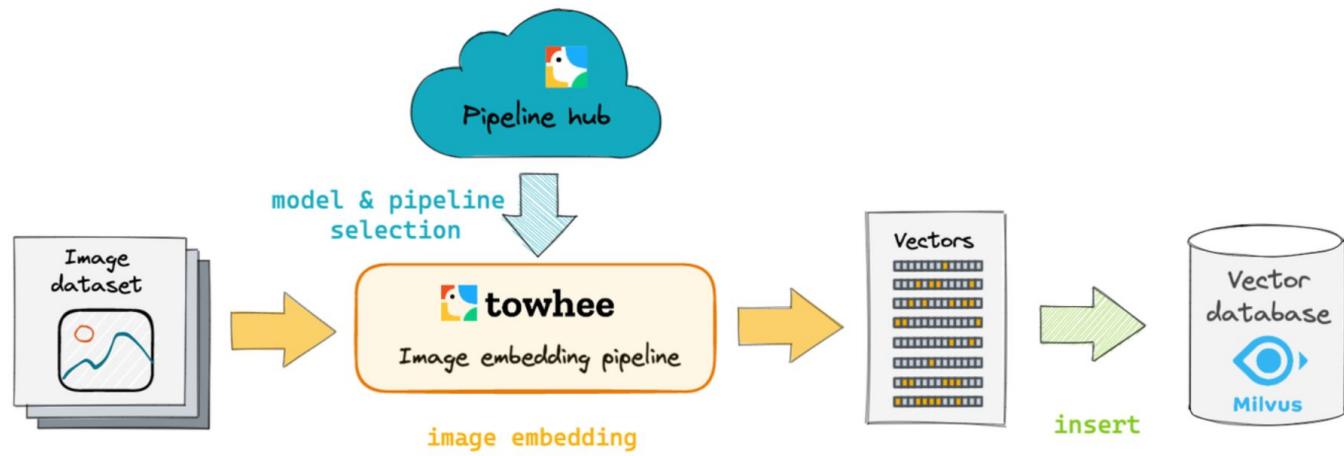




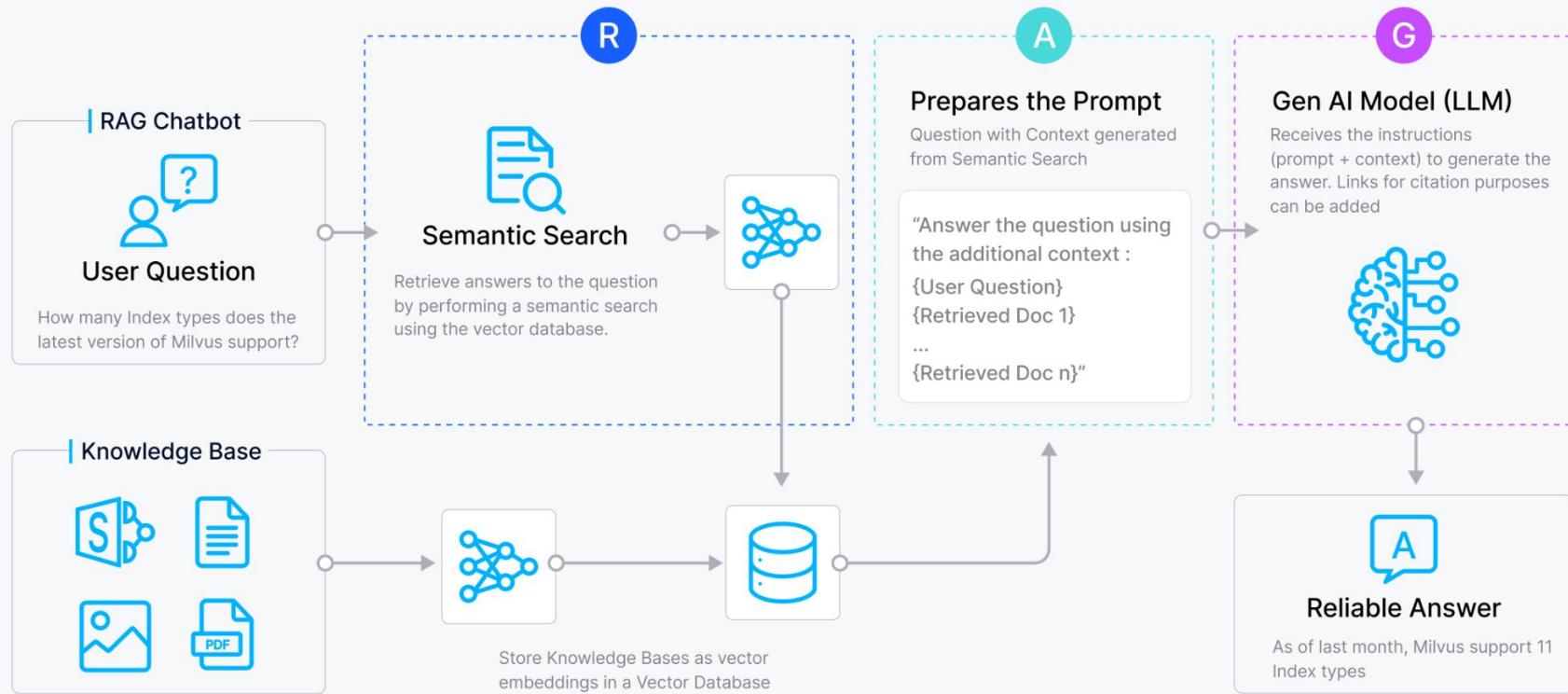
<https://github.com/MobilityData/mobility-database-catalogs/>

Every Transit System



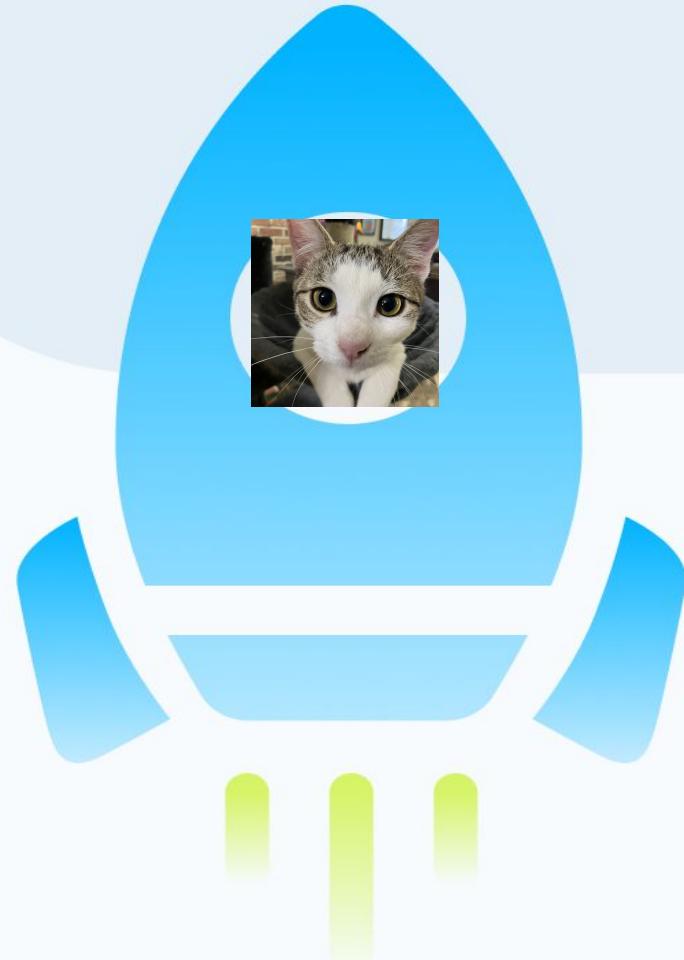
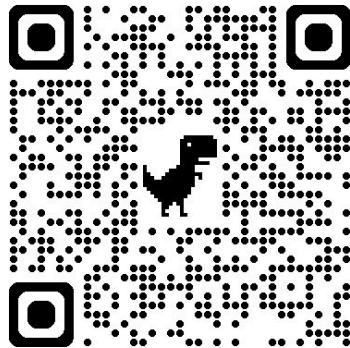


Retrieval-Augmented Generation RAG Chatbot



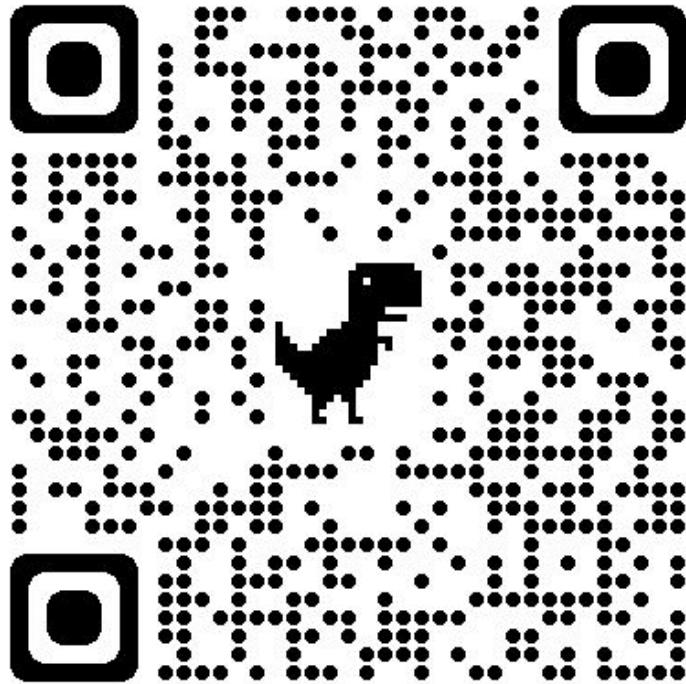


How To Get Started With Milvus





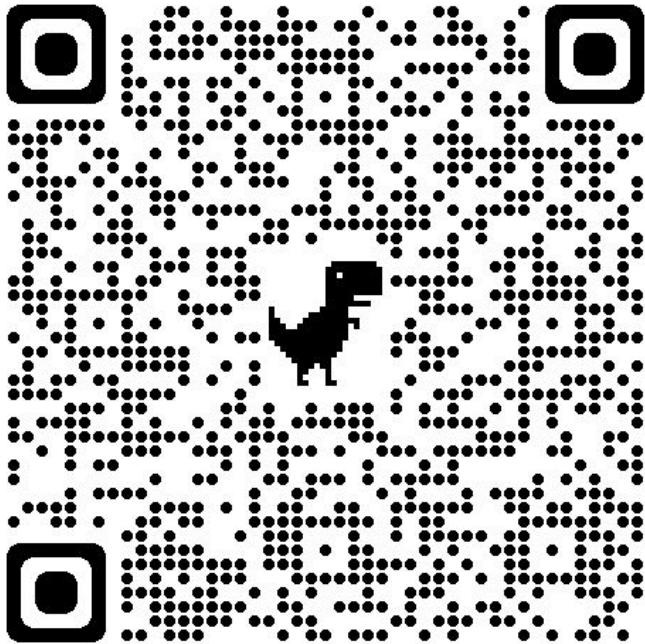
Irish Transit



<https://medium.com/@tspann/real-time-irish-transit-analytics-ea76164c9595>

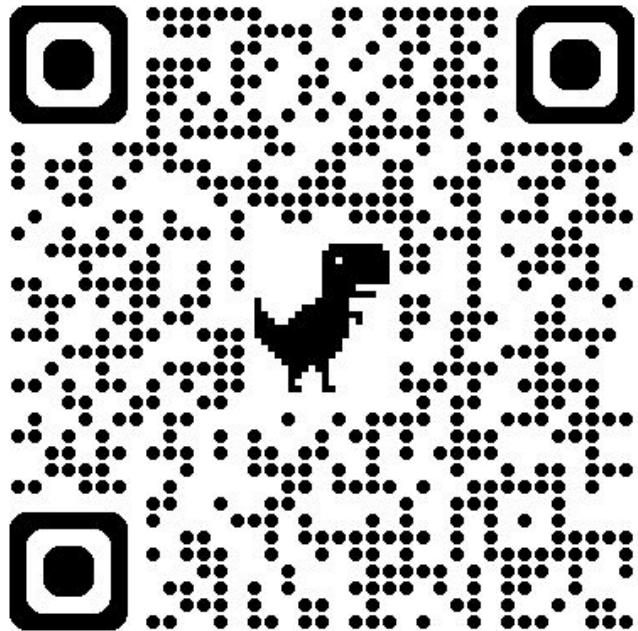
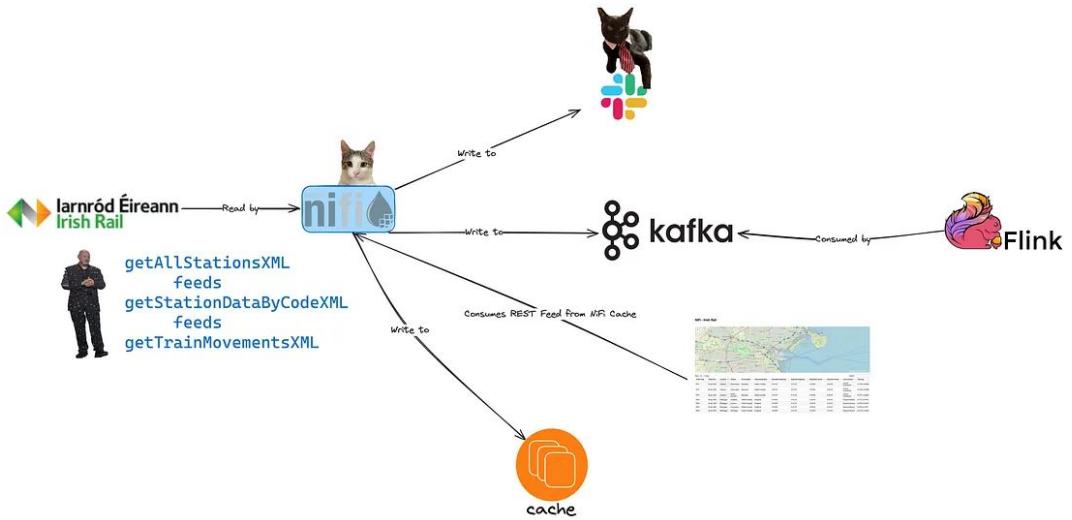


Street Cameras



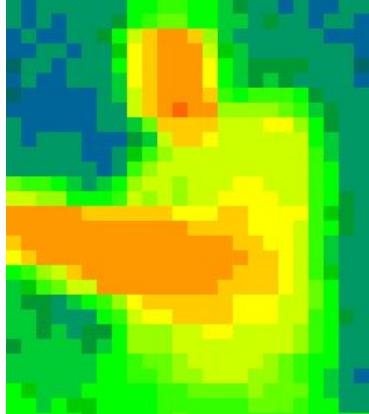
<https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-v8-with-python-and-nifi-to-minio-s3-3277e73723ce>

Irish Rail Example



<https://medium.com/cloudera-inc/events-streams-tiows-and-maps-zza8d2/cayb4>





IS THIS ENOUGH DATA?



DZone REF CARDS 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...



imgflip.com

Timothy Spann
Principal Developer Advocate

<https://medium.com/@tspann>



© 2024 Tim Spann All rights reserved