How to Handle Mixed Source-Type Information in GenAI?

Aleksei Kolesnikov

Staff Software Engineer

Imagine we have a variety of unstructured content like videos, text, images, and audio.

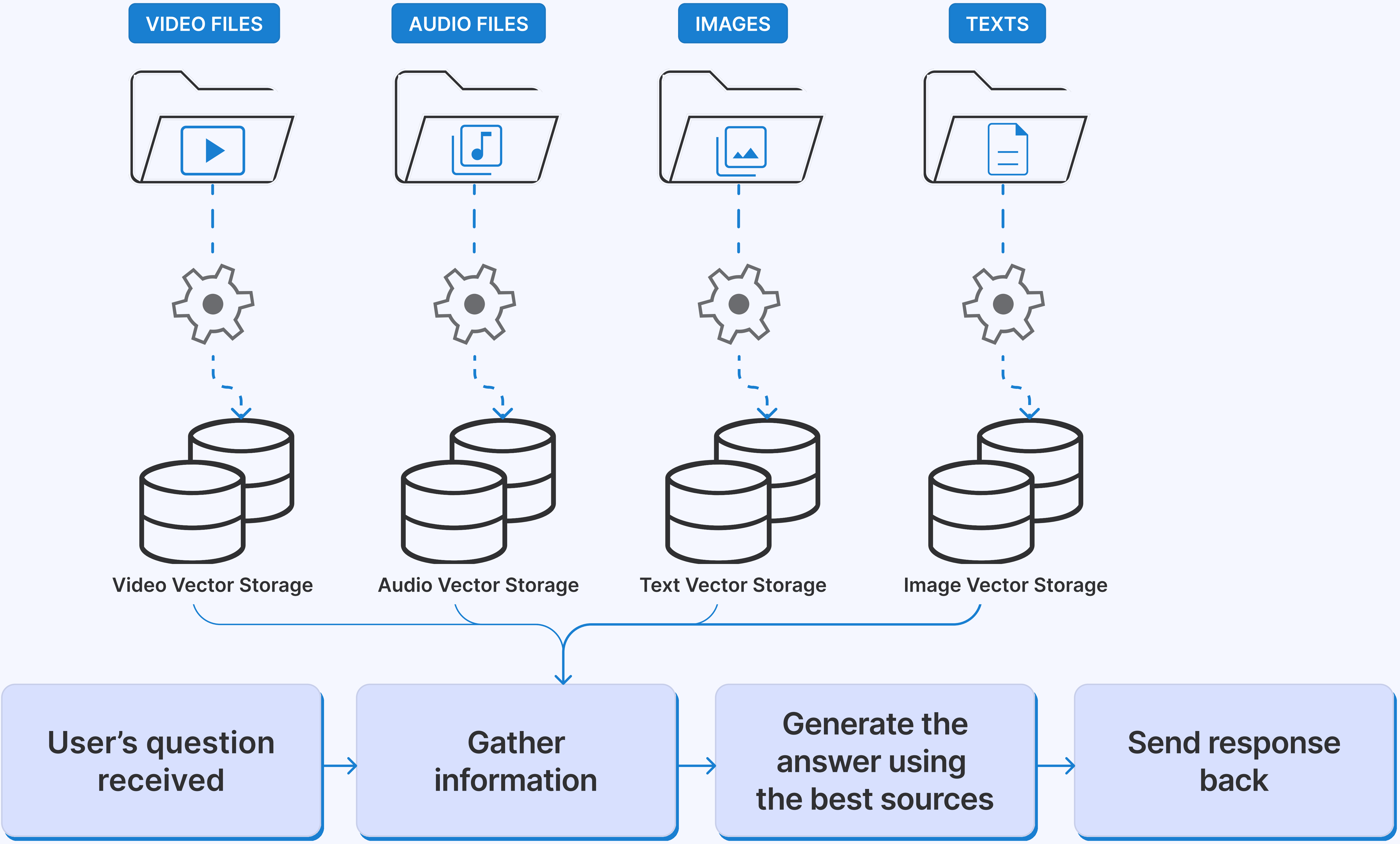
To answer queries, we need to draw information from all these sources.

So, how do we go about this?

There are two methods of achieving this:

1. The basic approach with multiple vector storages
2. The enhanced with one multi-modal vector storage

Basic Multi-Modal with Retrieval Augmented Generation



VIDEO FILES

AUDIO FILES

IMAGES

TEXTS

Video Vector Storage

Audio Vector Storage

Text Vector Storage

Image Vector Storage

User’s question received

Gather information

Generate the answer using the best sources

Send response back

Aleksei Kolesnikov
Staff Software Engineer

In a straightforward scenario, you can turn to embedding algorithms to create vectors from varied data and store them separately.

This raises a few questions:

1 Which algorithms can help extract information from images, audio, and videos?

2 Would using different embedding algorithms alter my responses?

3 Is it possible to consolidate vectors from different sources into one repository?

Let's answer on them!

1 There is plenty of algorithms!

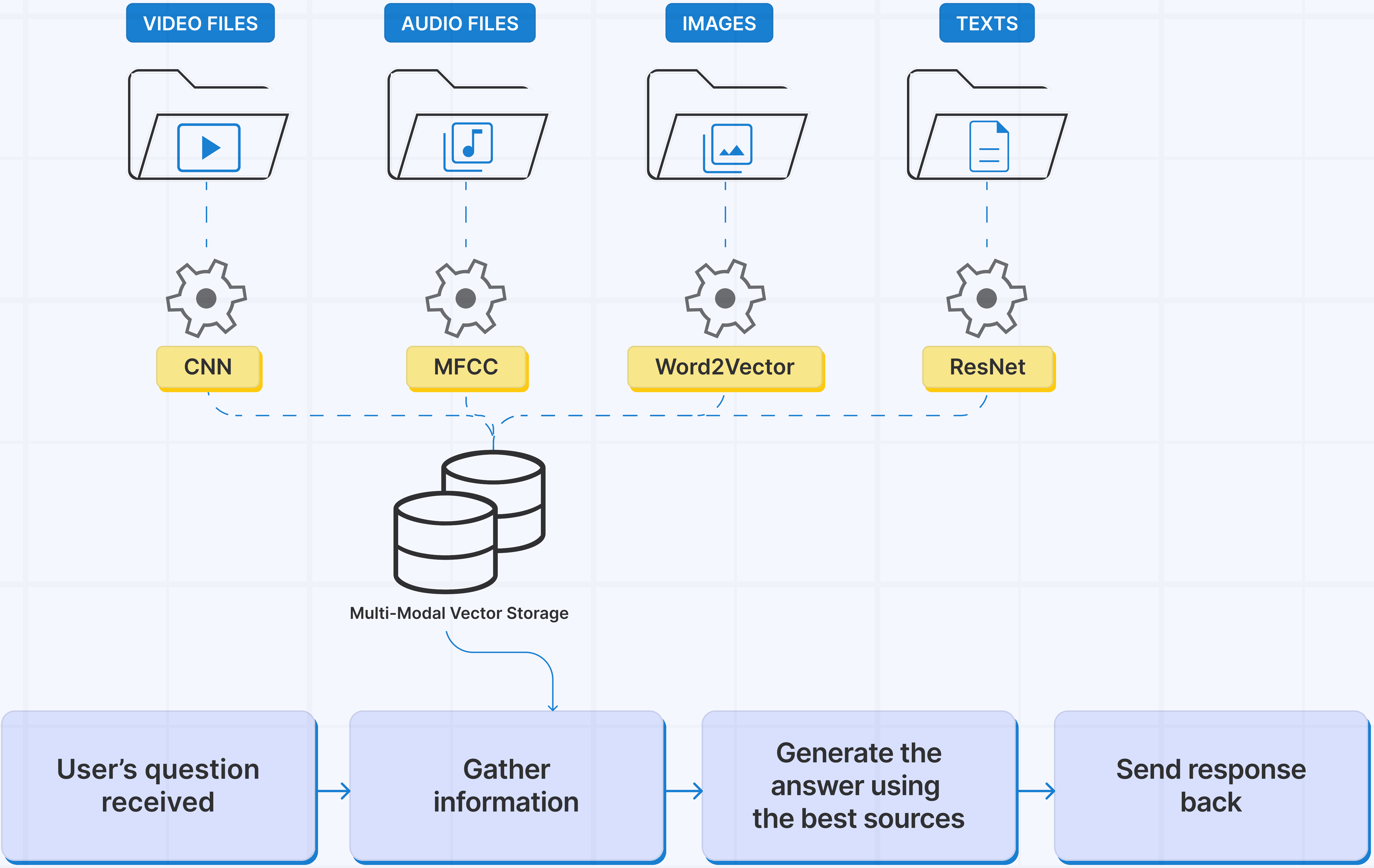
- For text is quite obvious to go with **Word2Vec** or **BERT**,
- For video it could be Convolutional Neural Network (**CNN**) or Supervised Contrastive Learning (**CNN-SCL**).
- For audio files the most commonly used is **MFCC** (Mel Frequency Cepstral Coefficients).
- For images **ResNet** could be used.

2 Of course it affects genAI responses! Using different algorithms can result in different embeddings, leading to varying responses generated by the multi-modal RAG paradigm!

3 Definitely Possible using Multi-modal index! By creating a combined representation for each piece of multi-modal data, search and retrieval processes can be greatly enhanced. **But it's a challenging process.**

Recommended articles:
https://arxiv.org/pdf/2306.08789.pdf
https://arxiv.org/pdf/2104.08108.pdf

Basic Multi-Modal with Retrieval Augmented Generation



VIDEO FILES

AUDIO FILES

IMAGES

TEXTS

CNN

MFCC

Word2Vector

ResNet

Multi-Modal Vector Storage

User's question received

Gather information

Generate the answer using the best sources

Send response back

Aleksei Kolesnikov
Staff Software Engineer

How to scan newly added files? Push or Pull (Scan) strategies!

→ With the **Push approach**, event-triggered approach could be leveraged, such that each time a new file is added, it's automatically processed into a vector.

→ Alternatively, using the **Pull approach**, scanning process could be setup and it will process files in the storage at periodic intervals, like every hour, for instance. Used in Semantic Search.

The final question might be the following

Can the Multi-modal RAG approach function synergistically with other RAG or FLARE paradigms?

The response is **YES!**

Numerous articles and statistics attest to enhanced accuracy and relevance in responses.

However, be aware that this could come with a considerable performance cost and seamless integration issues.

Good article was published on **llamaindex**:

https://docs.llamaindex.ai/en/stable/examples/evaluation/multi_modal/multi_modal_rag_evaluation.html