

Graph RAG: Leveraging the power of graphs to enhance retrieval

Prashanth Rao

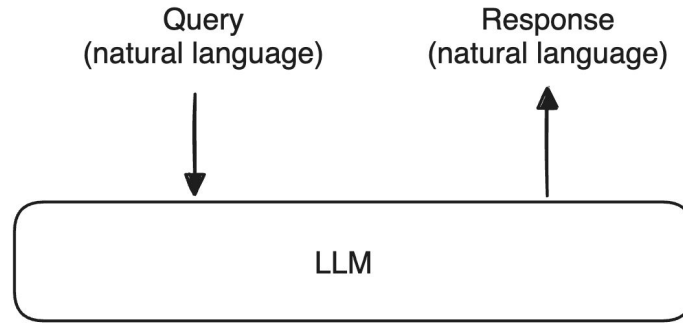
AI Engineer, Kùzu Inc. 🇨🇦

kuzudb.com

GDG Surrey DevFest

Surrey, BC | 26 Oct 2024

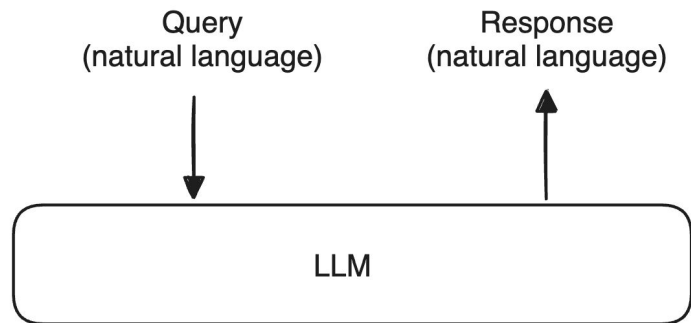
“Chat with an LLM” (e.g., ChatGPT)



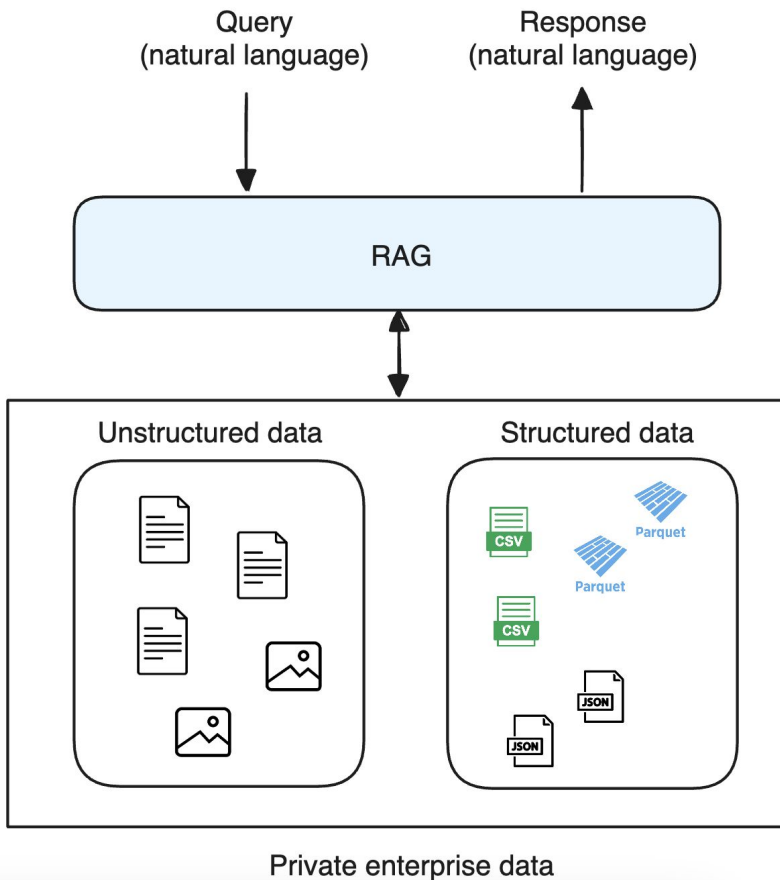
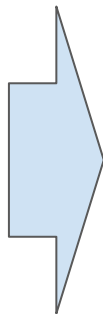
- Cannot easily retrieve from private enterprise data
- This is a retrieval problem

RAG: Retrieval in the age of LLMs

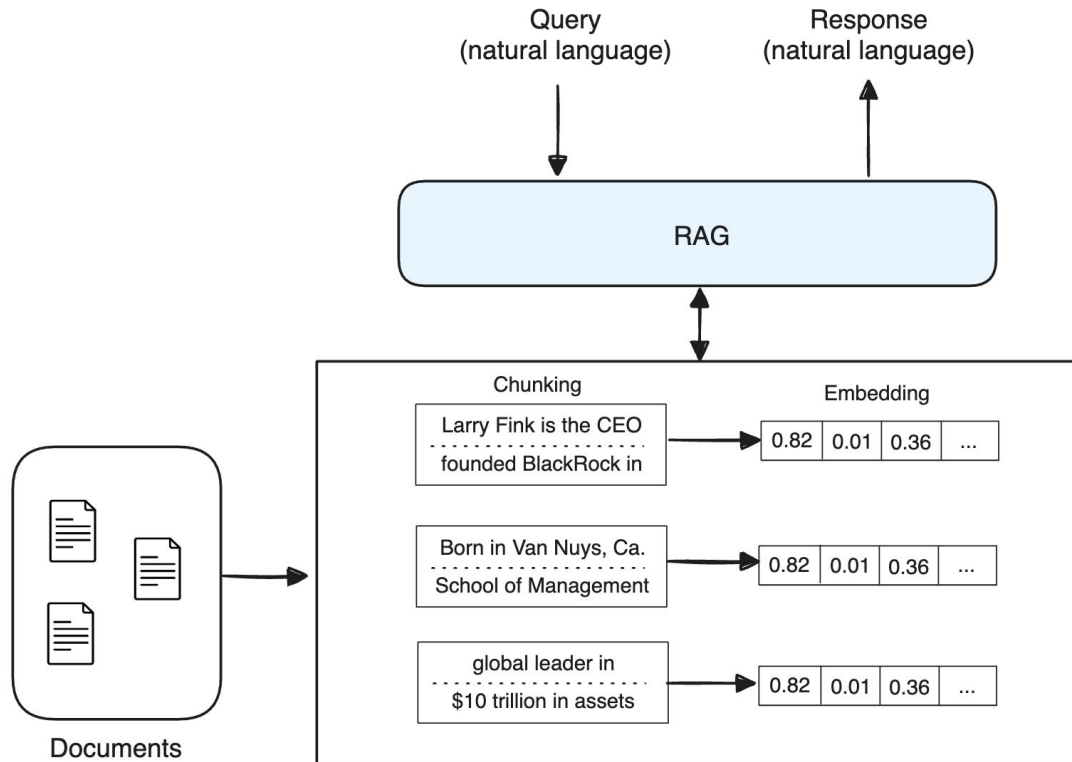
“Chat with an LLM”



Cannot easily retrieve from
private enterprise data



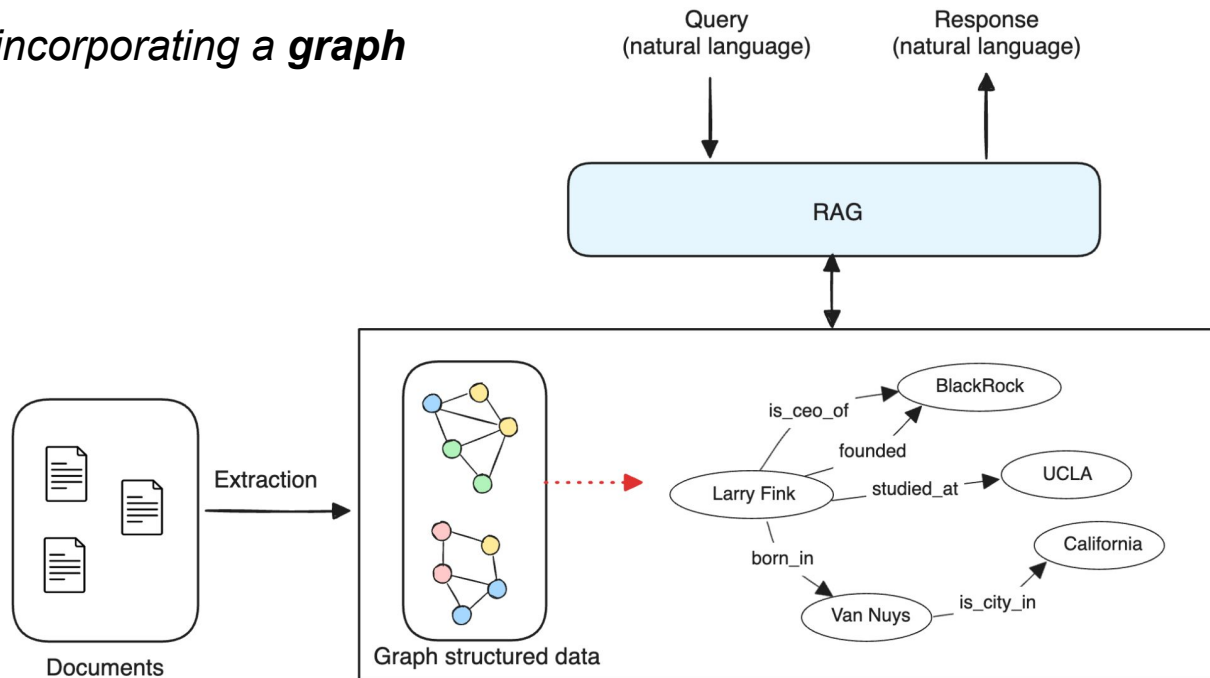
A deeper look at traditional RAG



Vector database (retrieve top-k documents)

What is Graph RAG?

*Extends traditional RAG by incorporating a **graph** as part of the retrieval step*

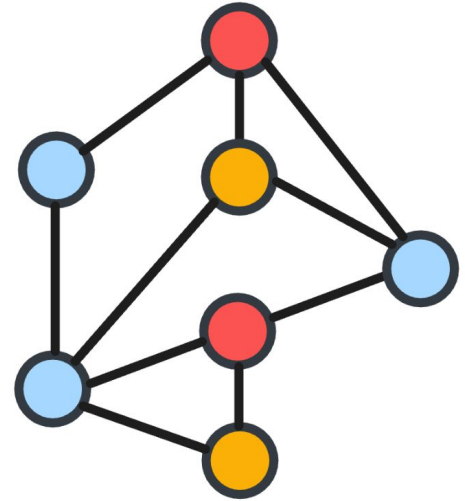


In any system that uses this approach:

- Question 1: What is the graph? i.e., what are its nodes and edges?
- Question 2: How is the retrieval process different from traditional RAG?

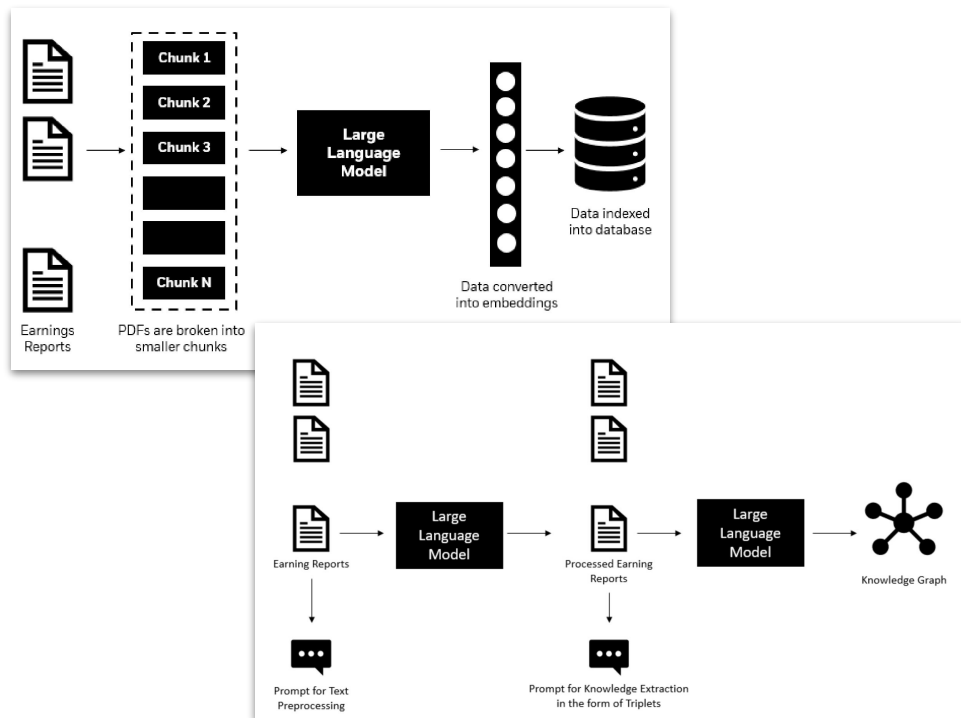
Why enhance unstructured data with a graph?

- Graphs are **object-oriented** in nature – they represent entities or objects in the real world via nodes, and how they are connected via edges
- Graphs capture relationships between entities **explicitly**
 - Traversing the vicinity of an entity to get added context is *natural and easy*
- A graph data model is a good fit to **add structure** to related entities extracted from unstructured data
- Importantly, graph triples/edges <subject, predicate, object>, can be represented as **simple sentences** (useful to generate context)



Do graphs measurably improve RAG in practice? KUZU

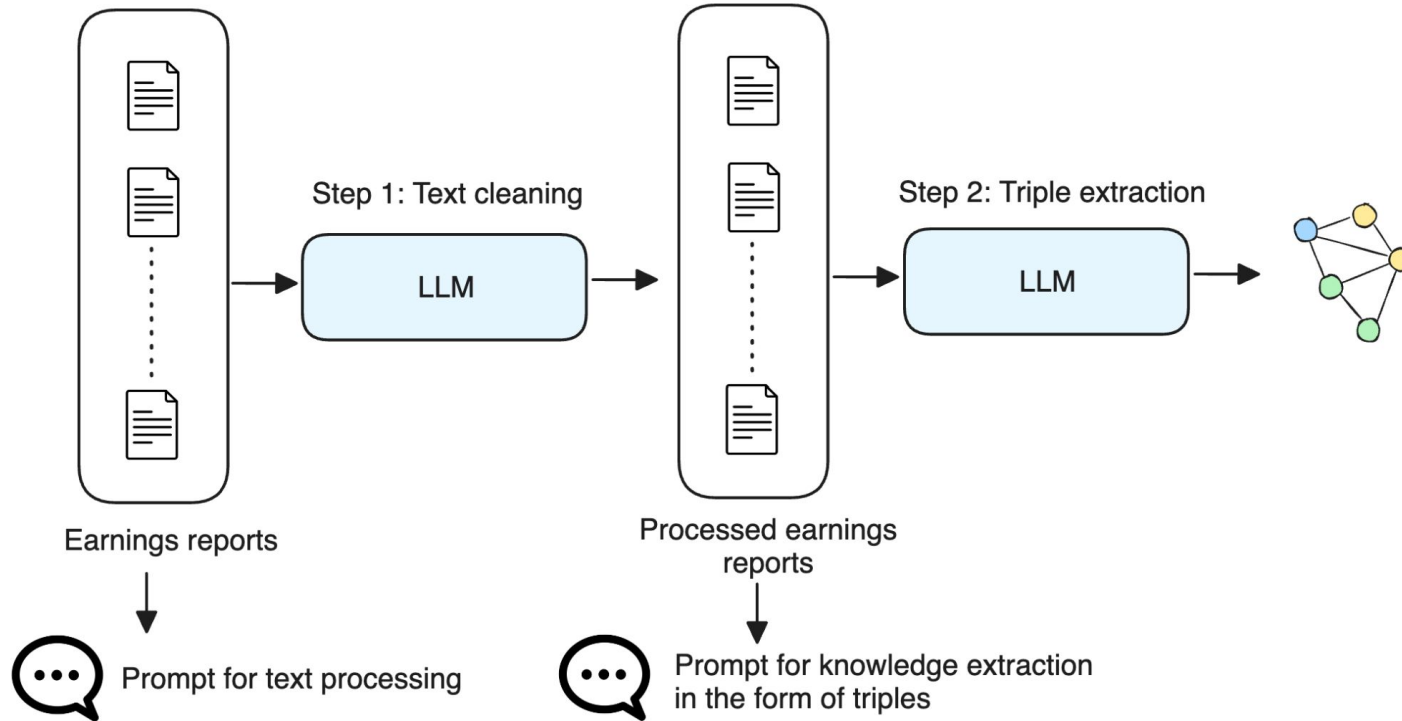
HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction (BlackRock & Nvidia), Aug 2024



Evaluation: Hybrid RAG system **does better overall** than systems that were based on vector retrievals or graph retrievals alone

Unpacking BlackRock's Hybrid RAG (1)

Question 1: What is the graph? What do its nodes and edges represent?



Unpacking BlackRock's Hybrid RAG (2)

Example of summarization and triple extraction

Chunk 1

Larry Fink is the CEO and co-founder of BlackRock, the world's largest asset management firm, established in 1988 ...

Processed chunk 1

Larry Fink is the CEO and co-founder of BlackRock.
BlackRock was established in 1988.

<Larry Fink, is_ceo_of, BlackRock >
<Larry Fink, founded, BlackRock >
<BlackRock, founded_in, 1988 >

Chunk 2

Born in Los Angeles, California, in 1952, Fink grew up in Van Nuys and later earned his MBA from UCLA's Anderson School of Management ...

Step 1: Text processing



Processed chunk 2

Larry Fink was born in Los Angeles, California.
Larry Fink earned his MBA from UCLA

Step 2: Triple extraction



<Larry Fink, born_in, Los Angeles >
<Los Angeles, is_city_in, California >
<Larry Fink, graduated_from, UCLA >

⋮

⋮

Chunk n

...
10.0 trillions of dollars in asset management ...
...

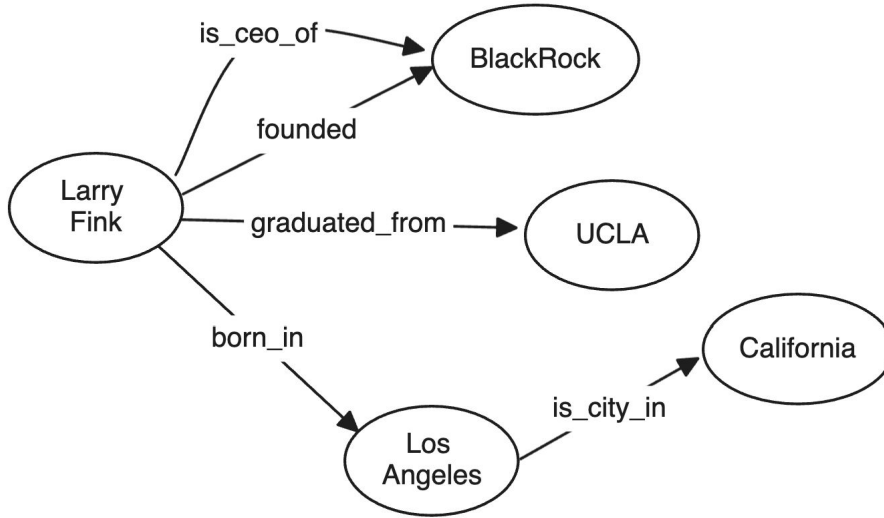
Processed chunk n

...
BlackRock manages 10.5 trillion dollars in assets.
...

<BlackRock, asset_value, 10.5 trillion >

Unpacking BlackRock's Hybrid RAG (3)

Recall: Graphs can model simple sentences



Chunk 1

<Larry Fink, is_ceo_of, BlackRock >
<Larry Fink, founded, BlackRock >

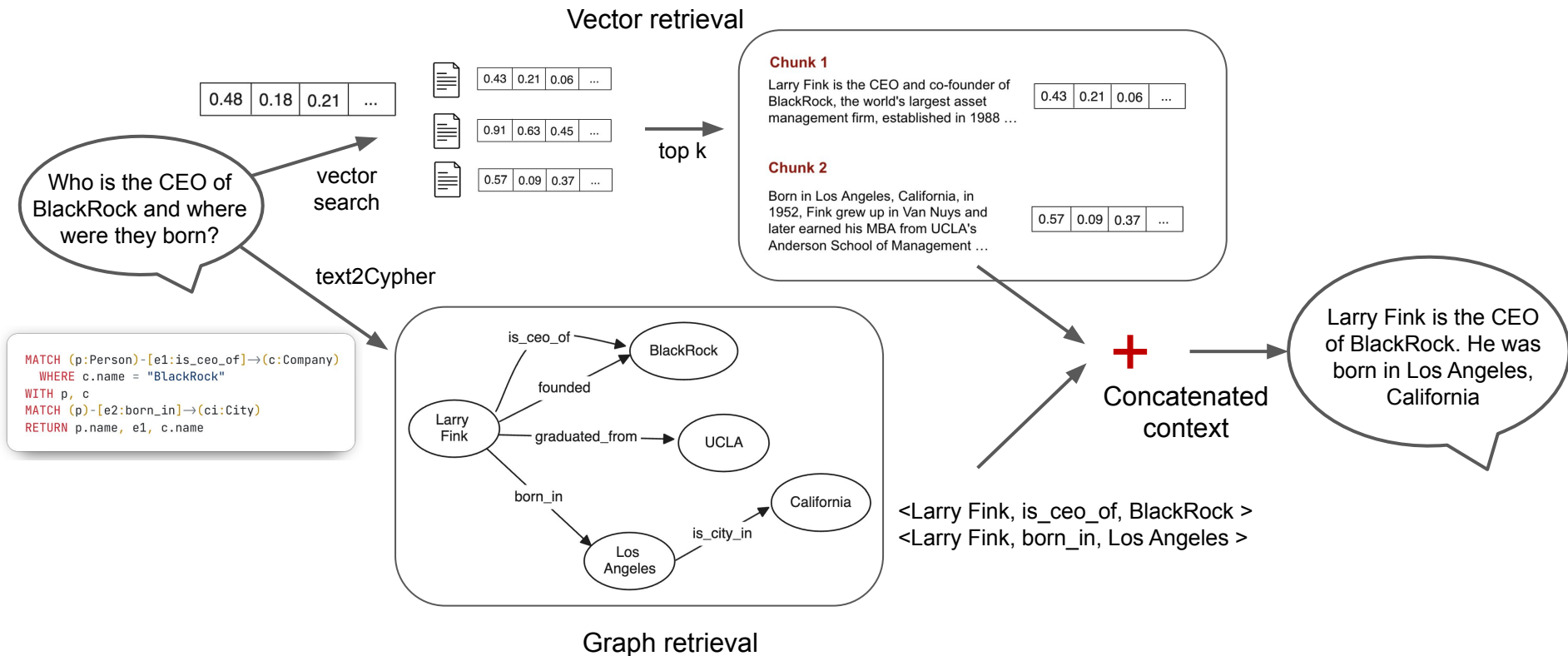
Chunk 2

<Larry Fink, born_in, Los Angeles >
<Los Angeles, is_city_in, California >
<Larry Fink, graduated_from, UCLA >

- Benefit 1: Information in disparate chunks are now **directly connected**
- Benefit 2: Triples are a form of capturing the **essence of text chunks** in **very simple sentences**
- Benefit 3: Put the triples into a graph DB where you can query it using a declarative **query language**

Unpacking BlackRock's Hybrid RAG (4)

Question 2: How is the retrieval process different from traditional RAG?



Outcomes

Q1	Who are the founders of BlackRock? Return the names as a numbered list.
Q2	Where did Larry Fink graduate from?
Q3	When was Susan Wagner born?
Q4	How did Larry Fink and Robert Kapito meet?

Vector RAG

Q1	✗
Q2	✓
Q3	✗
Q4	✓

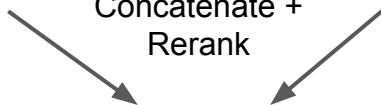
Graph RAG

Q1	✓
Q2	✓
Q3	✓
Q4	✗

Hybrid RAG



Concatenate +
Rerank



Q1	✓
Q2	✓
Q3	✓
Q4	✓

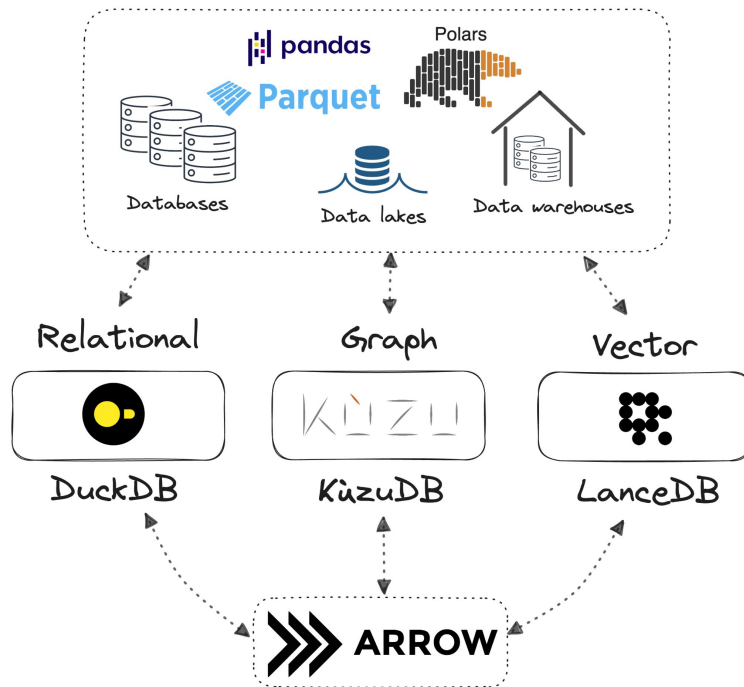
*Neither retrieval method is inherently better or worse than the other – on average, **combining** them can provide value!*

Let's go through some code!

<https://github.com/kuzudb/google-devfest-graph-raq>

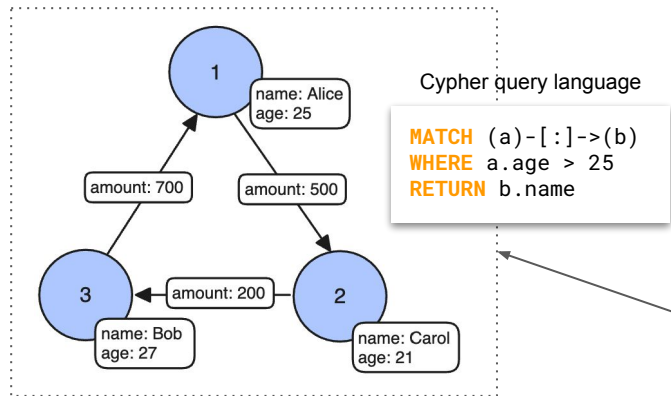
Databases are evolving alongside RAG

- Embeddability + ease of setup + interoperability + permissive licensing
- These characteristics do **not** preclude scalability or performance



Usability features of Kùzu

Property graph data model & RDF wrapper

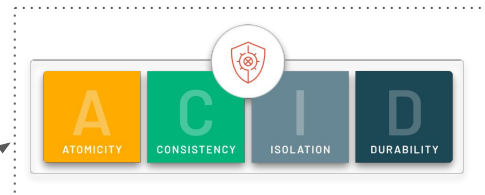


Embedded (similar philosophy to DuckDB, LanceDB)

```
import kuzu

db = kuzu.Database("db")
conn = kuzu.Connection(db)
res = conn.execute("MATCH (a)-[:]->(b)")
print(res.get_as_df())
```

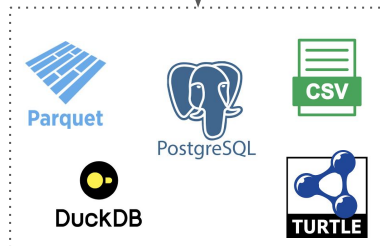
ACID transactions



Integrations with ML/AI frameworks



Usability features



Permissively licensed



Interoperable with many formats

Learn more at <https://kuzudb.com>

- Graphs can help *explicitly model* relationships between entities in your data
- To retrieve *factual* information, a graph can help store manually gathered data in a structured, maintainable fashion
- For better retrieval from the graph, keep the following in mind
 - Graph quality is important: improves the *retrieval* outcome
 - The choice of LLM matters: improves the *quality of Cypher generation*
 - Prompts matter: Provide *schema* in the prompt to improve Cypher generation
- The vector embedding and graph data pipelines can be *built and tuned independently*
- Design concrete **evaluation** strategies using a suite of representative questions in your domain

Thank you!

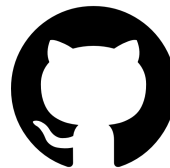


Kùzu is **open source**

Code: <https://github.com/kuzudb/google-devfest-graph-raq>



Join our Discord!



github.com/kuzudb/kuzu



[@kuzudb](https://twitter.com/kuzudb)



We're [Kùzu Inc.](#)
on LinkedIn!