neo4j

Beyond Vectors

Alison Cossette - Developer Advocate Data Science

Session Overview

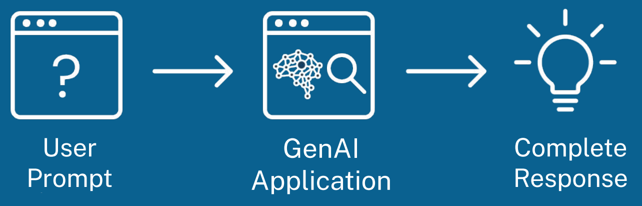
1. **Data Sources**
   The right sources for the right questions
2. **Data Understanding**
   Unstructured EDA
3. **Data Quality**
   Makings of a quality GenAI dataset
4. **Application Understanding**
   Learnings from application use

Data Sources

The right data for the right question.

A Generative AI application
uses an LLM
to provide **responses**
to **user prompts**

(aka ChatGPT)

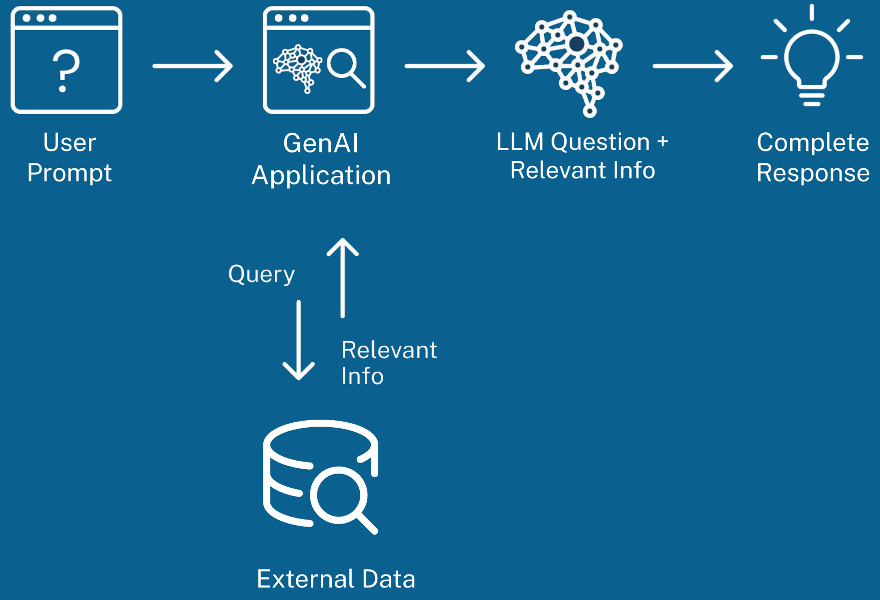


4

RAG augments the LLM by intercepting a **user's prompt**,

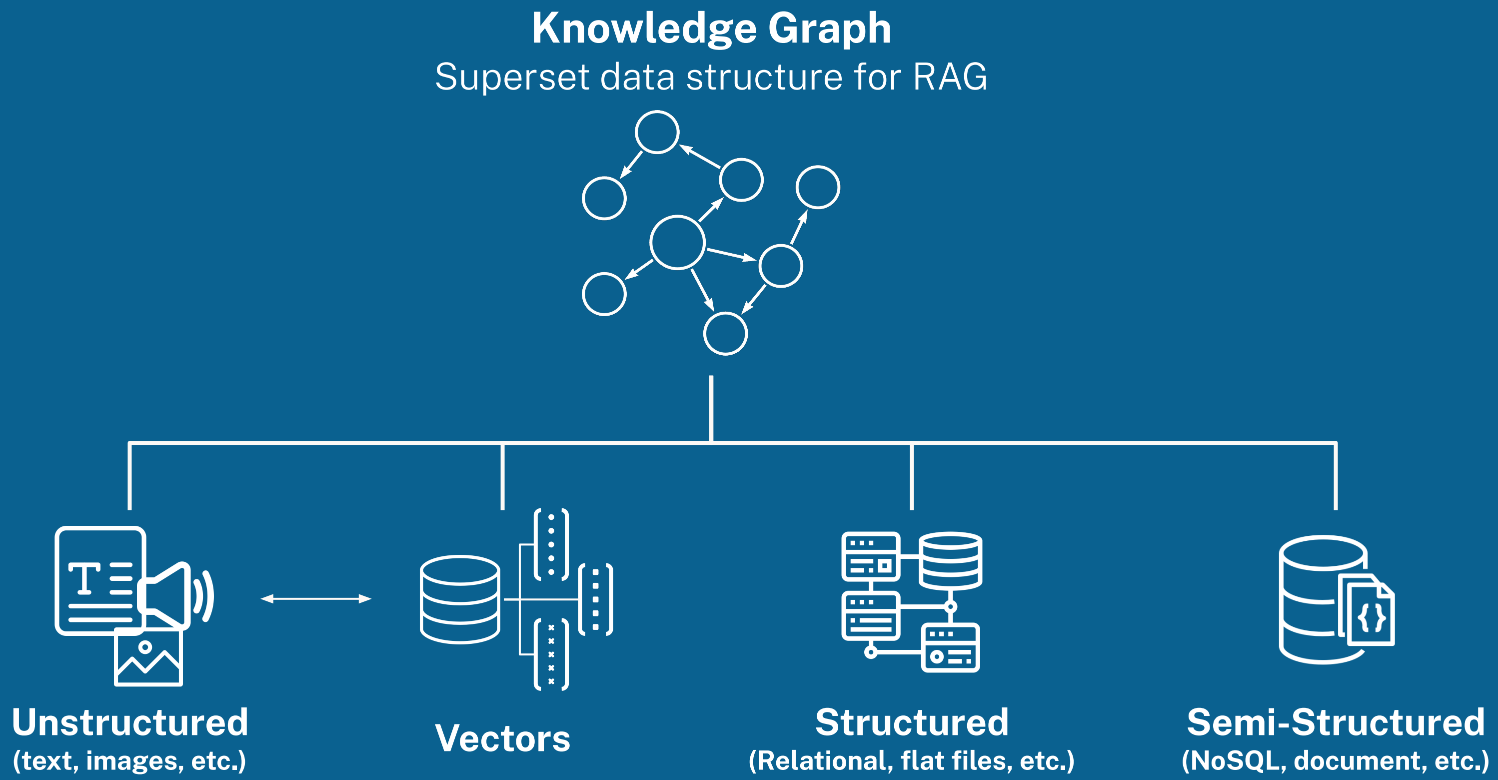
then making a **query to external data**,

then passing relevant results from the query back to the LLM for a **complete, curated response.**



5

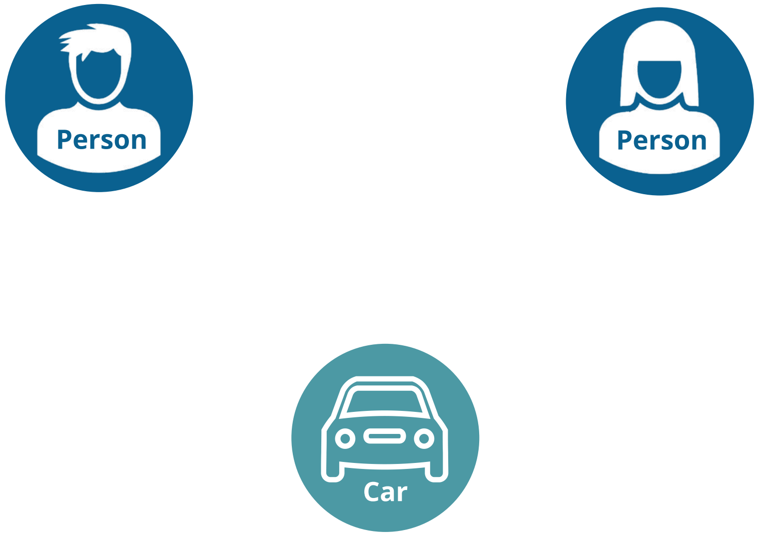
What to use for External Data?



6

Neo4j Graph Components

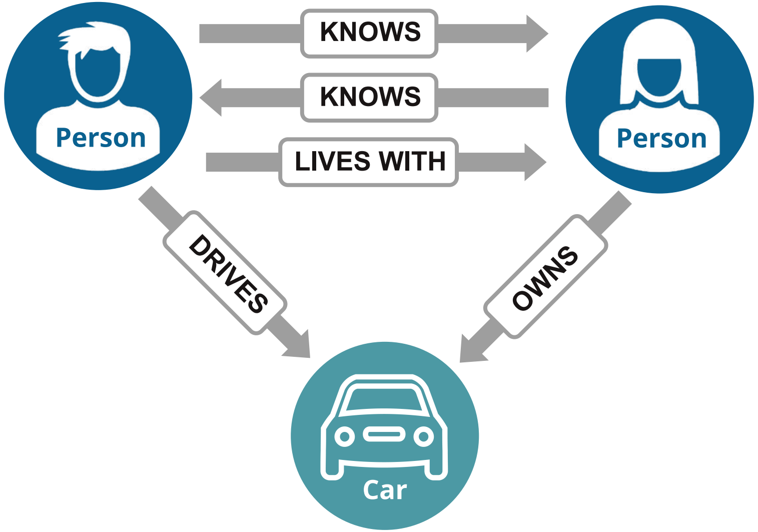
Nodes represent entities in the graph



Neo4j Graph Components

Nodes represent entities in the graph

Relationships represent associations or interactions between nodes

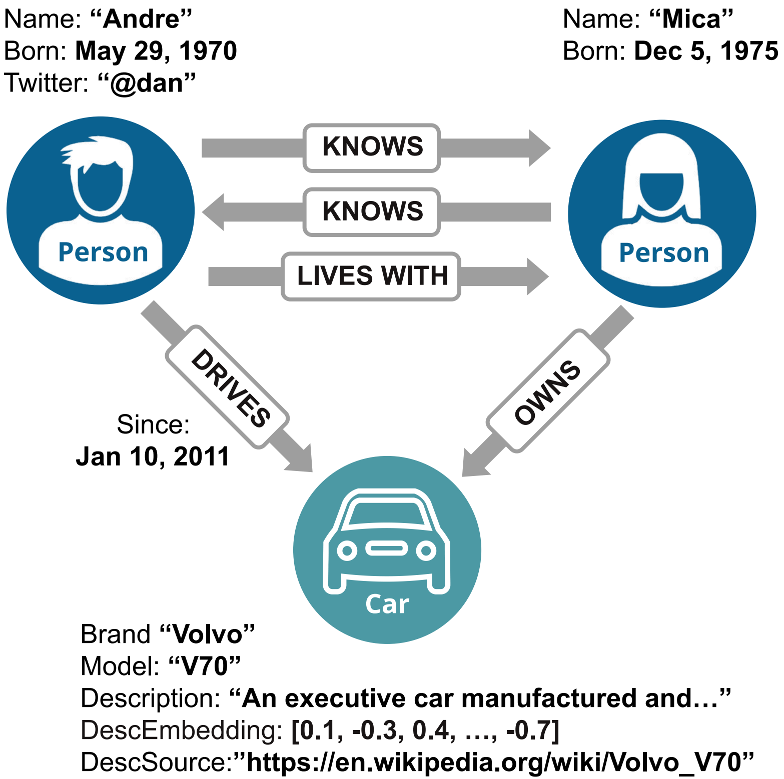


Neo4j Graph Components

Nodes represent entities in the graph

Relationships represent associations or interactions between nodes

Properties represent attributes of nodes or relationships including vectors



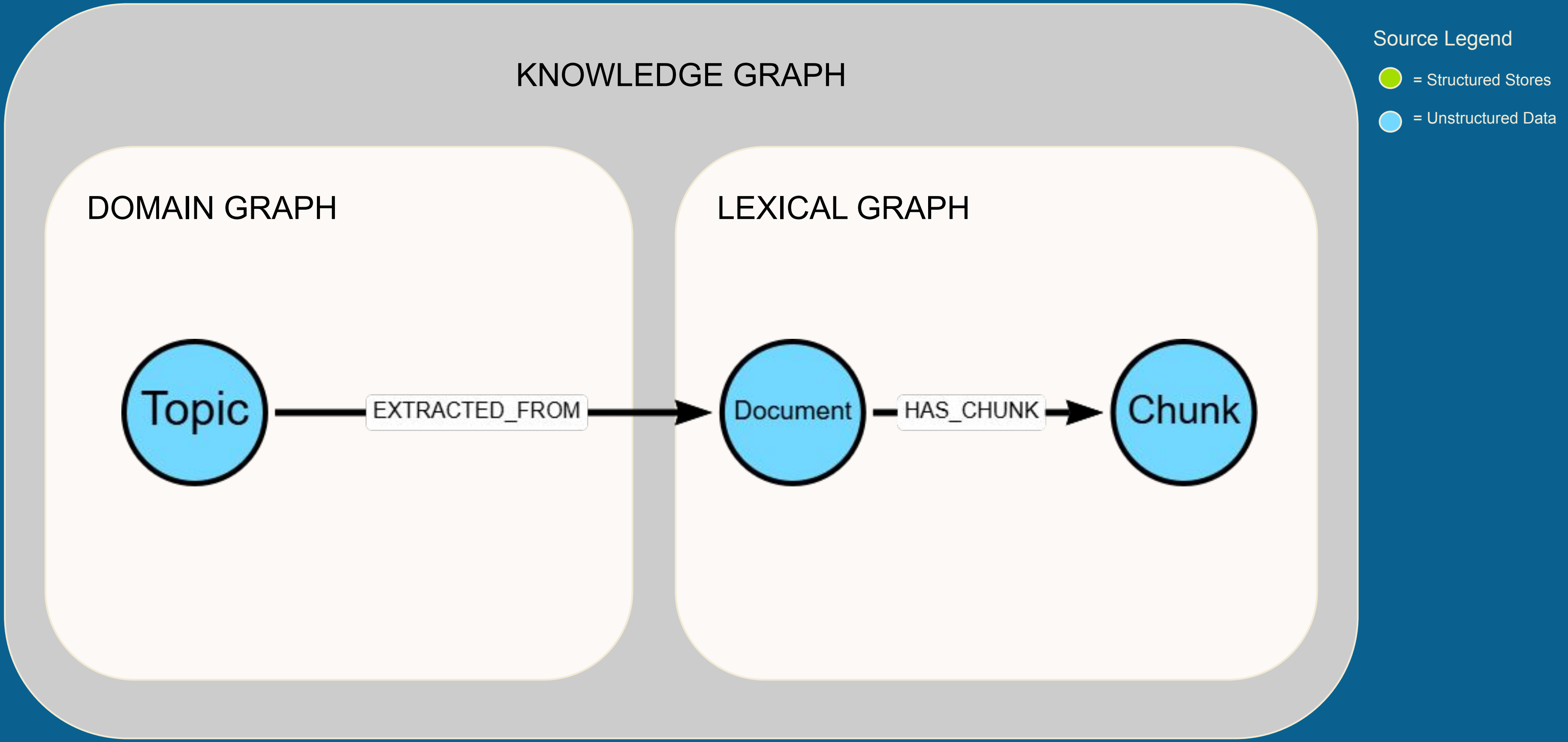
Knowledge Graphs

KNOWLEDGE GRAPH

LEXICAL GRAPH

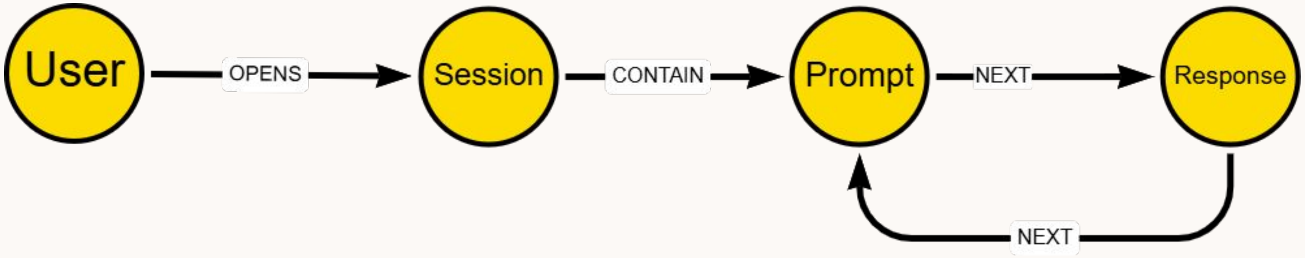
Source Legend

- = Structured Stores
- = Unstructured Data
- = Application


KNOWLEDGE GRAPH
DOMAIN GRAPH
LEXICAL GRAPH
Topic
EXTRACTED_FROM
Document
HAS_CHUNK
Chunk
Source Legend
= Structured Stores
= Unstructured Data

KNOWLEDGE GRAPH

MEMORY GRAPH

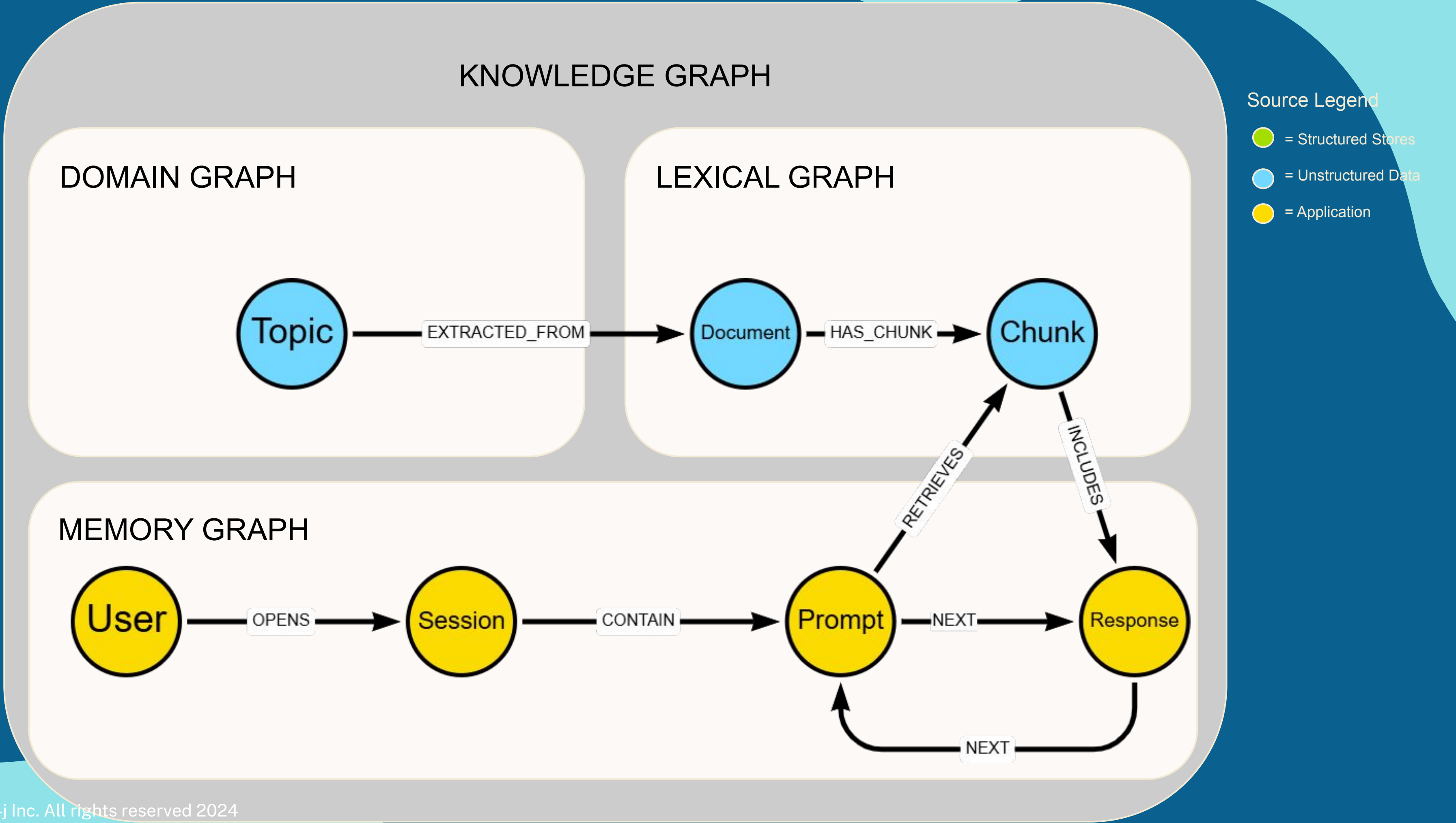


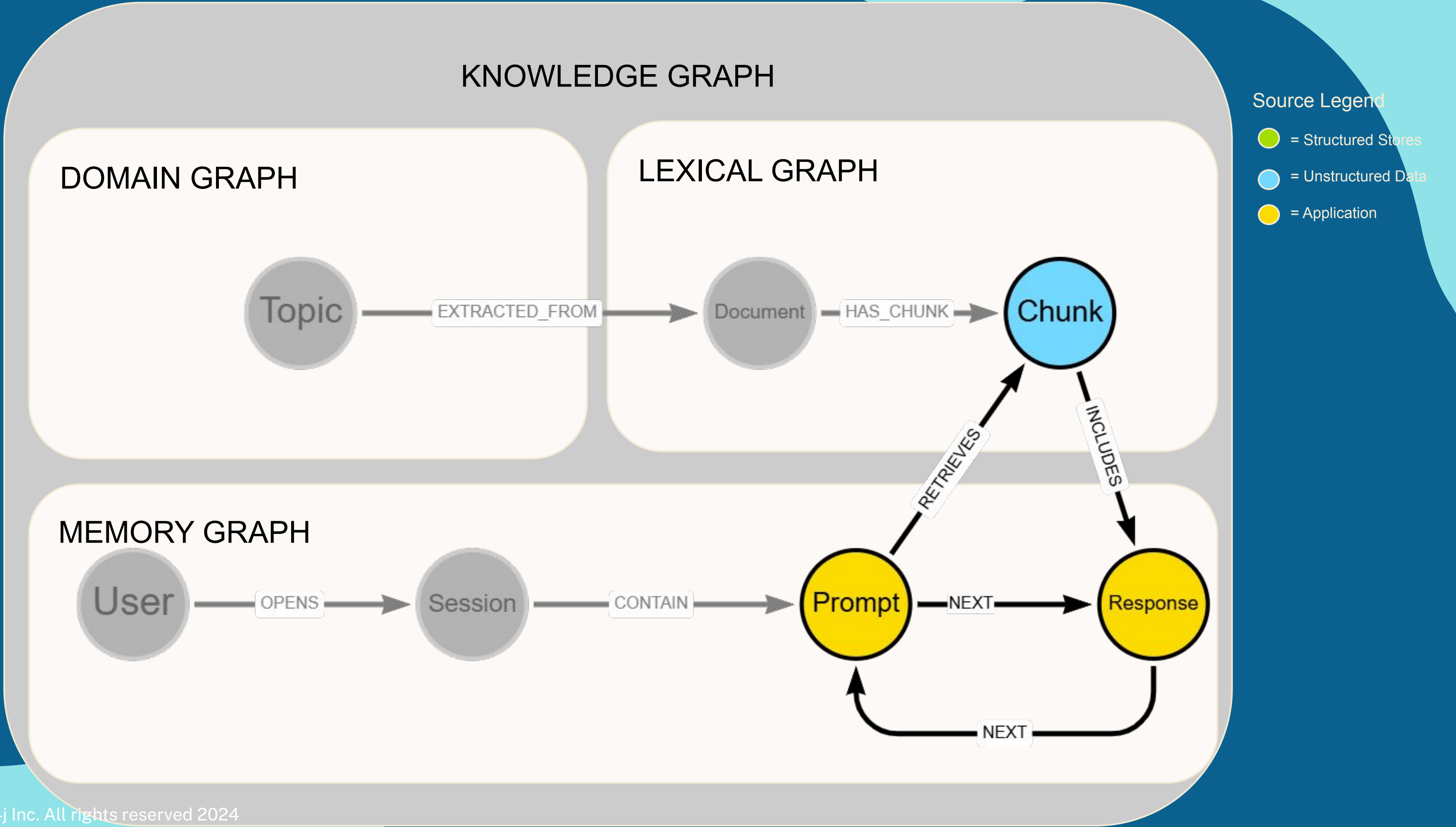
Source Legend

= Structured Stores

= Unstructured Data

= Application


KNOWLEDGE GRAPH
DOMAIN GRAPH
LEXICAL GRAPH
MEMORY GRAPH
Topic
EXTRACTED_FROM
Document
HAS_CHUNK
Chunk
RETRIEVES
INCLUDES
User
OPENS
Session
CONTAIN
Prompt
NEXT
Response
NEXT
Source Legend
= Structured Stores
= Unstructured Data
= Application



KNOWLEDGE GRAPH
DOMAIN GRAPH
LEXICAL GRAPH
MEMORY GRAPH
Topic
EXTRACTED_FROM
Document
HAS_CHUNK
Chunk
RETRIEVES
INCLUDES
User
OPENS
Session
CONTAIN
Prompt
NEXT
Response
NEXT
Source Legend
= Structured Stores
= Unstructured Data
= Application

Data Understanding

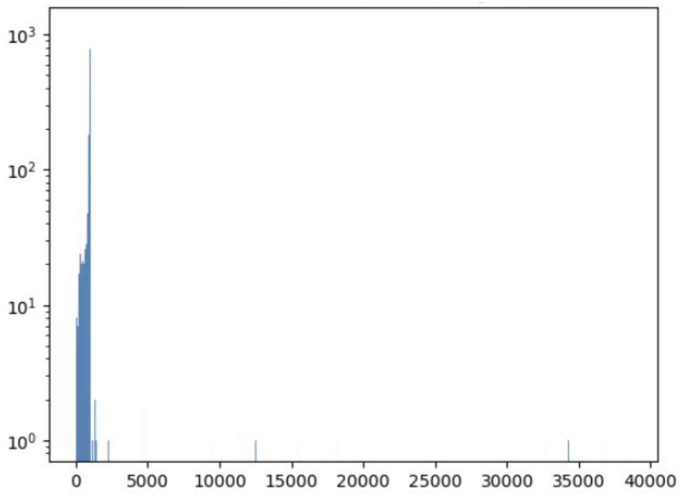
Exploring your unstructured data

Grounding Data Sources

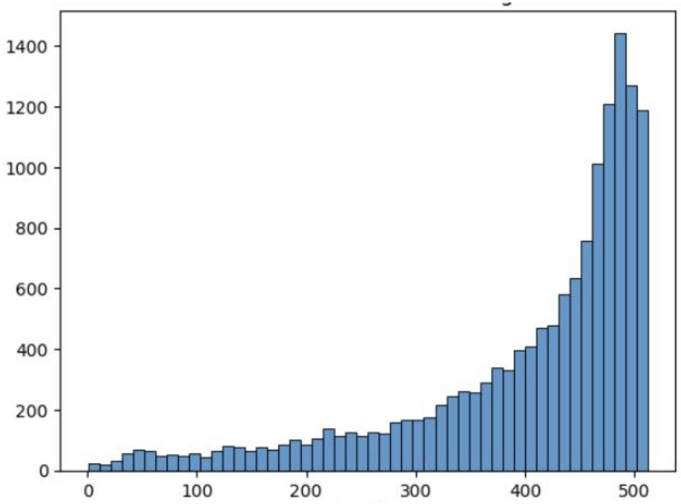
- Dataset (all public data sources)
  - 1,150 documents of official Neo4j documentation
  - Developer blogs
  - Support knowledge base
  - Github
- Split this text into 15,000 embedded text ‘chunks’
  - 512 chunk size
  - LangChain Recursive Text Splitter
  - Embeddings via GCP
  - URL of each chunk for LLM citation

EDA on Source Documents

EDA on source documents and document chunks is a critical step before generating embeddings and loading them into the database.



Text length distribution with chunking strategy errors

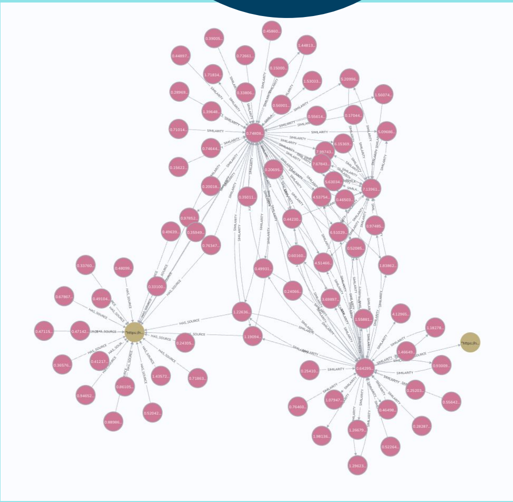


Text length distribution with corrected chunking strategy

EDA with Graphs and GDS

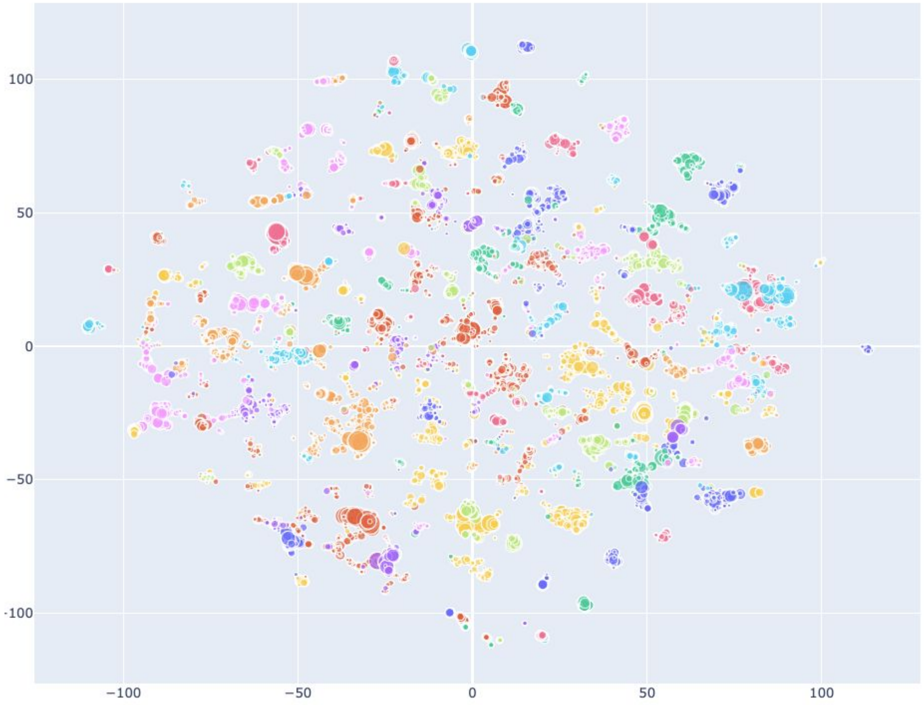
CONNECT **CLUSTER** **CURATE**

- **KNN Similarity** create relationships between the most similar chunk
- **Community Detection** and creates clusters based on similarity relationships
- Curate the grounding data set via techniques that work **at scale**



SImilarity Graph of Context Document Chunks (red) with Source URLs (gold)

Analyzing Context with Graphs and GDS

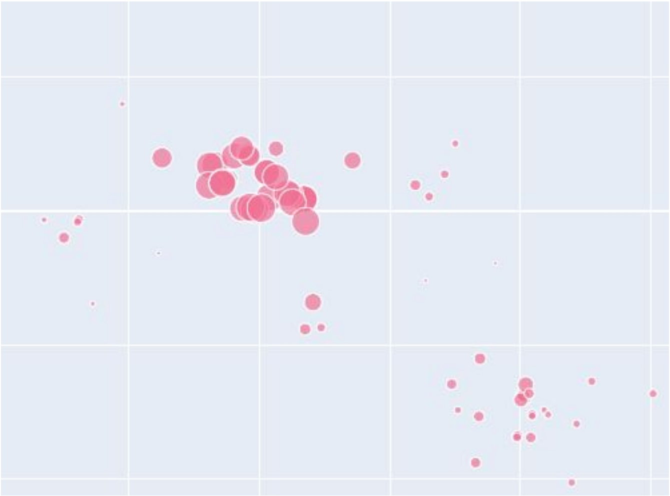


2D Visualization of Context Document Node Embeddings

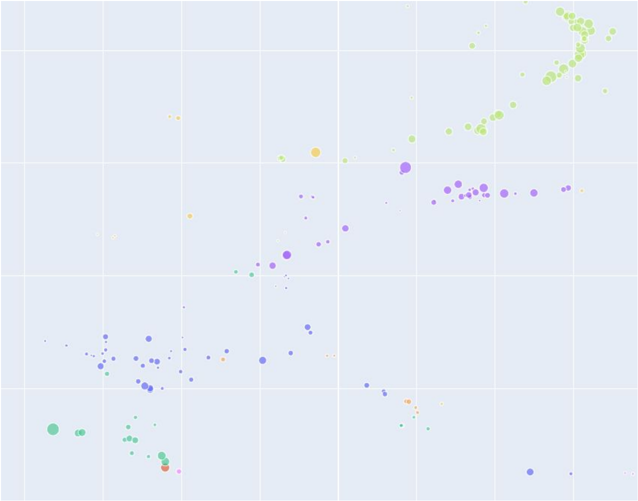
Clean

Embedding Visualization Detail

Our grounding document embeddings are generally distinct, but some communities overlap, which may warrant further analysis.



Single-Community Document Cluster



Cluster of Overlapping Document Communities

Data Quality

Curation for AI Success

Elements of High-Quality Grounding Data

Relevant	Related to the problem the LLM is solving and the questions you expect users to ask.
Augmenting	Fills known gaps in the LLM's 'knowledge', due to data being non-public our outside the training window.
Reliable	Contains accurate information, whether from inside or outside of the organization.
Clean	Is generally free of errors or noise, especially if generated from notebooks, websites, repos, etc...
Efficient	Does not contain duplicates, or near-duplicate, 'chunks' that take up valuable context limits.

Clean

Identifying Text Errors

Combining graph and traditional statistics helps us identify outliers or data quality issues.

- Traditional: Text length, word count, and word length
- Graph: Community, community size, and PageRank score

community	size	med_textLen	med_wordCount	med_avgWordLen	med_pageRank
14015	44	372.0	35.0	8.38	2.236358
755	30	507.0	78.0	5.50	1.893640
7117	51	512.0	1.0	512.00	1.811893
4506	25	479.0	46.0	8.16	1.677685
8299	22	422.5	68.0	5.22	1.603973
12142	27	465.0	61.0	6.71	1.498172
4035	43	407.0	70.0	4.83	1.495224
4701	51	373.0	50.0	6.88	1.468466
1455	22	422.0	56.0	6.77	1.466139
10877	37	421.0	40.0	7.37	1.397775

Graph Communities with additional statistics;
Note: Outlier average word length in community 7117

Clean

Investigating Text Chunk Outliers

text	text_len	word_count	avg_word_len	community	pageRank
{"payload":{"allShortcutsEnabled":false,"fileTree":{"":{"items":[{"name":"algorithms","path":"algorithms","contentType":"directory"},{"name":"embeddings","path":"embeddings","contentType":"directory"},{"name":"README.md","path":"README.md","contentType":"file"},{"name":"gds-resources.md","path":"gds-resources.md","contentType":"file"},{"name":"graph-data-modeling.md","path":"graph-data-modeling.md","contentType":"file"},{"name":"graph-eda.md","path":"graph-eda.md","contentType":"file"},{"name":"graphs-llms.	512	1	512.0	7117	2.415697
{"payload":{"allShortcutsEnabled":false,"fileTree":{"":{"items":[{"name":"algorithms","path":"algorithms","contentType":"directory"},{"name":"embeddings","path":"embeddings","contentType":"directory"},{"name":"README.md","path":"README.md","contentType":"file"},{"name":"gds-resources.md","path":"gds-resources.md","contentType":"file"},{"name":"graph-data-modeling.md","path":"graph-data-modeling.md","contentType":"file"},{"name":"graph-eda.md","path":"graph-eda.md","contentType":"file"},{"name":"graphs-llms.	512	1	512.0	7117	2.413580
{"payload":{"allShortcutsEnabled":false,"fileTree":{"":{"items":[{"name":"algorithms","path":"algorithms","contentType":"directory"},{"name":"embeddings","path":"embeddings","contentType":"directory"},{"name":"README.md","path":"README.md","contentType":"file"},{"name":"gds-resources.md","path":"gds-resources.md","contentType":"file"},{"name":"graph-data-modeling.md","path":"graph-data-modeling.md","contentType":"file"},{"name":"graph-eda.md","path":"graph-eda.md","contentType":"file"},{"name":"graphs-llms.	512	1	512.0	7117	2.288317

Highly-Similar Text Chunks

Efficient

Because we persisted the KNN Similarity relationships, we can query them just like any other object in the graph:

- Identify all text chunks with at least one 99%+ similarity relationship
- Identify communities with the highest average similarity scores

community	average_similarity	node_count	relationship_count
4213	0.991812	17	272
4702	0.986700	49	1207
11148	0.986366	17	272
6979	0.985077	71	1553
11180	0.982547	17	272

Highly-Similar Text Chunks

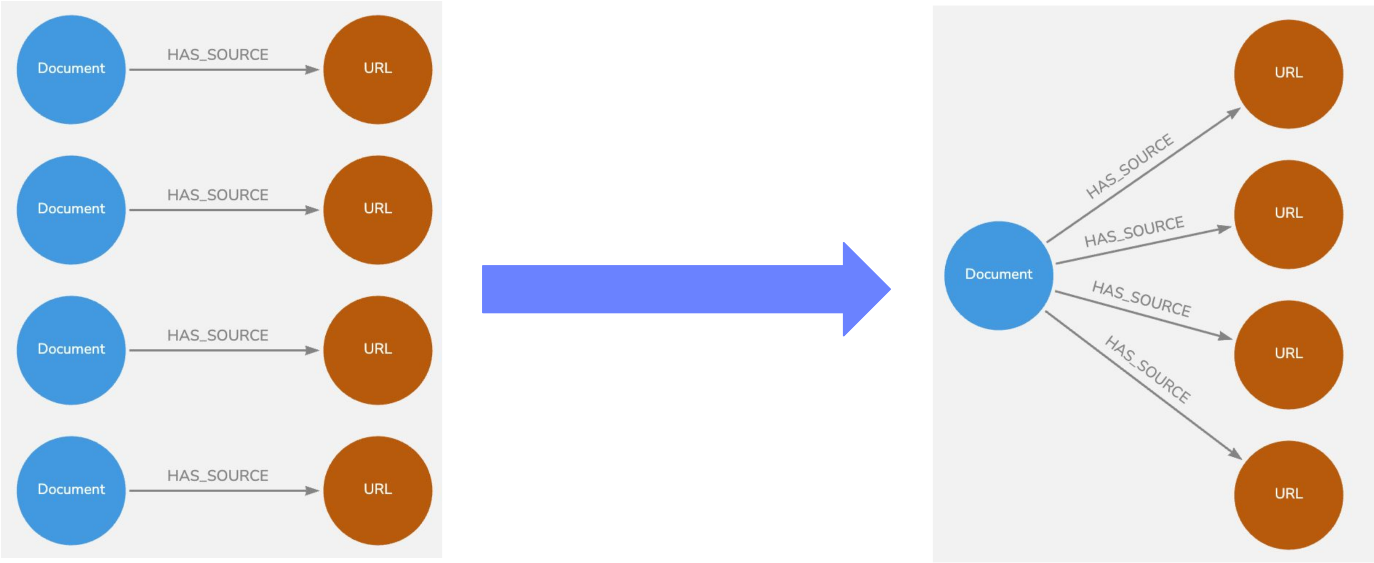
Efficient

url	text	community
https://neo4j.com/docs/graph-data-science/current/machine-learning/node-embeddings/graph-sage/	Heterogeneous relationships Heterogeneous relationships fully supported. The algorithm has the ability to distinguish between relationships of different types. Heterogeneous relationships Heterogeneous relationships allowed. The algorithm treats all selected relationships similarly regardless of their type. Weighted relationships Weighted trait. The algorithm supports a relationship property to be used as weight, specified via the relationshipWeightProperty configuration parameter.	4702
https://neo4j.com/docs/graph-data-science/current/algorithms/harmonic-centrality/	Heterogeneous relationships Heterogeneous relationships fully supported. The algorithm has the ability to distinguish between relationships of different types. Heterogeneous relationships Heterogeneous relationships allowed. The algorithm treats all selected relationships similarly regardless of their type. Weighted relationships Weighted trait. The algorithm supports a relationship property to be used as weight, specified via the relationshipWeightProperty configuration parameter.	4702
https://neo4j.com/docs/graph-data-science/current/algorithms/bellman-ford-single-source/	Heterogeneous relationships Heterogeneous relationships fully supported. The algorithm has the ability to distinguish between relationships of different types. Heterogeneous relationships Heterogeneous relationships allowed. The algorithm treats all selected relationships similarly regardless of their type. Weighted relationships Weighted trait. The algorithm supports a relationship property to be used as weight, specified via the relationshipWeightProperty configuration parameter.	4702
https://neo4j.com/docs/graph-data-science/current/algorithms/modularity-optimization/	Heterogeneous relationships Heterogeneous relationships fully supported. The algorithm has the ability to distinguish between relationships of different types. Heterogeneous relationships Heterogeneous relationships allowed. The algorithm treats all selected relationships similarly regardless of their type. Weighted relationships Weighted trait. The algorithm supports a relationship property to be used as weight, specified via the relationshipWeightProperty configuration parameter.	4702
https://neo4j.com/docs/graph-data-science/current/algorithms/closeness-centrality/	Heterogeneous relationships Heterogeneous relationships fully supported. The algorithm has the ability to distinguish between relationships of different types. Heterogeneous relationships Heterogeneous relationships allowed. The algorithm treats all selected relationships similarly regardless of their type. Weighted relationships Weighted trait. The algorithm supports a relationship property to be used as weight, specified via the relationshipWeightProperty configuration parameter.	4702

Efficient

Collapsing Duplicate Nodes

We can use `apoc.nodes.collapse()` to combine duplicate nodes into a single node with all prior relationships pointing to the single node.



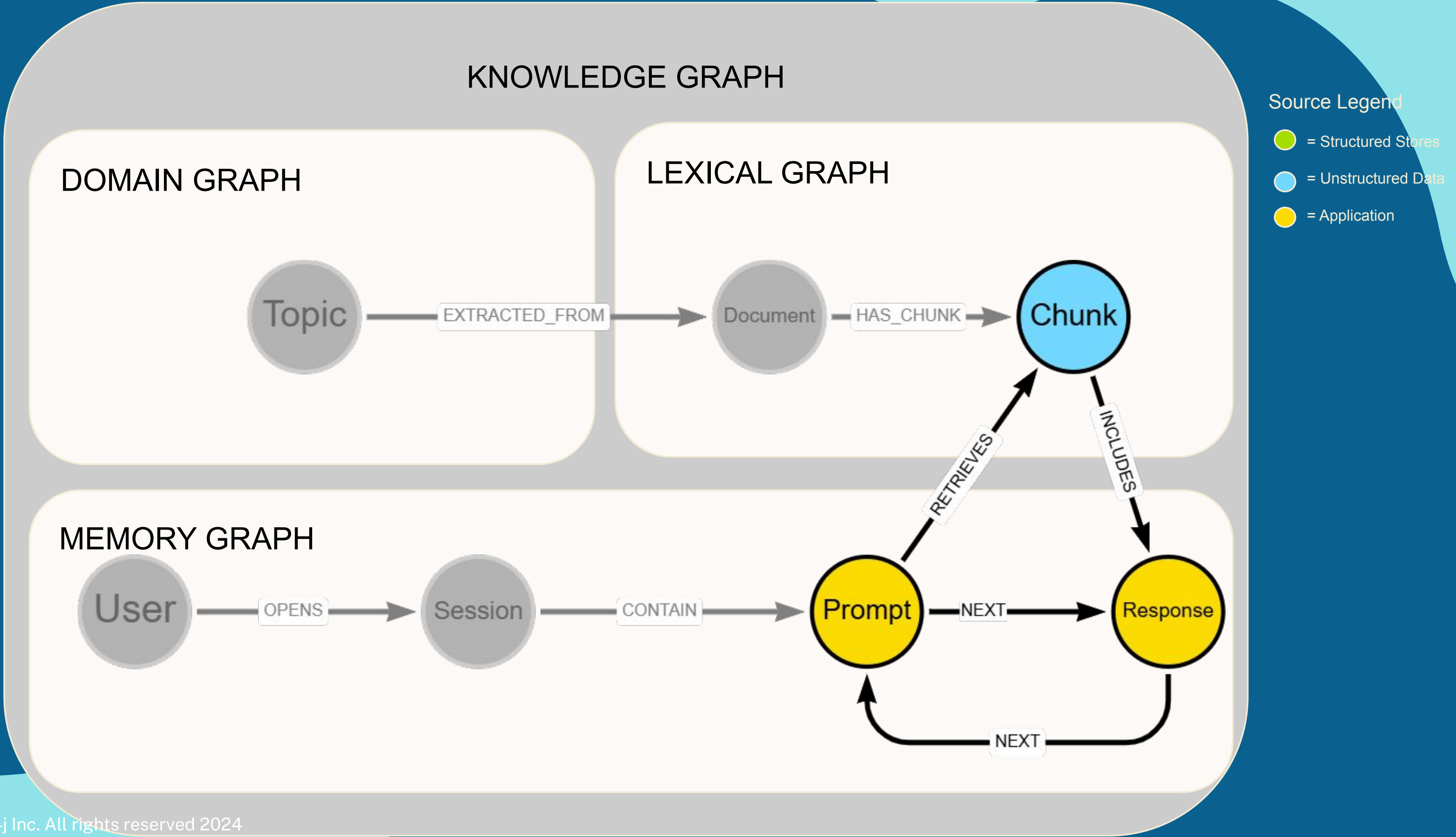
Application Understanding

What your application can teach you.

Graphs Enable Explainable AI with LLMs

- Production
    - How the LLM use grounding documents
    - How they produce answers will become more and more important

- Knowledge Graphs and GDS enable Explainable AI by:
    - Logging user interactions in the same database as the context
    - Visualizing conversations with context
    - Providing tools to analyze LLM performance and identify opportunities for improvement


KNOWLEDGE GRAPH
DOMAIN GRAPH
LEXICAL GRAPH
MEMORY GRAPH
Topic
EXTRACTED_FROM
Document
HAS_CHUNK
Chunk
RETRIEVES
INCLUDES
User
OPENS
Session
CONTAIN
Prompt
NEXT
Response
NEXT
Source Legend
= Structured Stores
= Unstructured Data
= Application

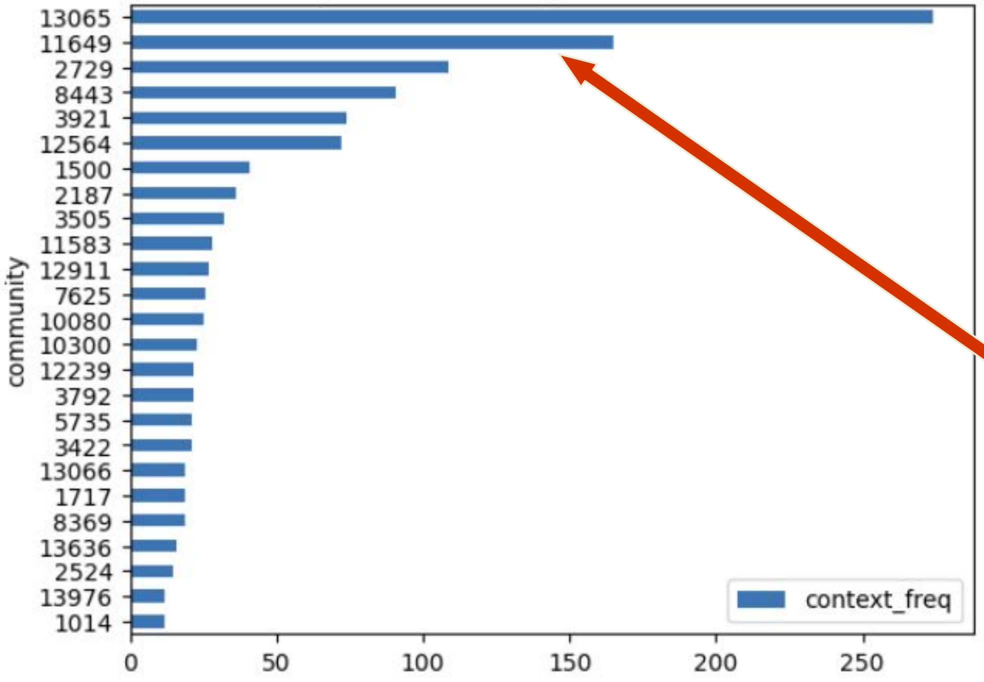
Most Frequently Used Grounding Text

Relevant

context_count	text
10.0	GDS Degree Centrality algorithm is useful for creating statistics that can support calculating ratios and identifying outliers. The same could also be performed using Cypher (and, if a large graph, apoc.periodic.iterate()). However, one of the benefits of using GDS and Graph Projections is that we can create a single projection and run multiple algorithms on it.</p>\n \nFor example, if we were analyzing a financial transaction network we may want to identify customers who had the most transactions. We
9.0	Neo4j Data Connectors Apache Kafka, Apache Spark, and BI tools Cypher Query Language Powerful, intuitive, and graph-optimized Solutions Use Cases Fraud detection, knowledge graphs and more Generative AI Back your LLMs with a knowledge graph for better business AI Learn More
9.0	want you to act as an experienced graph data scientist who works at Neo4j. A customer asks you how large language models (LLMs) like ChatGPT can assist with graph data science, specifically using Neo4j Graph Data Science algorithms. How would you advise this customer to explore integrating LLMs into their graph data science workflows? What would likely be the easiest or most impactful ways in which an LLM can make them more productive and effective?</p>\n<h2 tabindex=\"-1\" dir=\"auto\"><a
9.0	Perhaps you are a data scientist, or you aspire to become one. Graph analytics and data science offer a wide variety of algorithms that can enhance your analytical toolbox and help you find meaningful insights into highly-connected datasets. In this section, I will show how easily you can integrate graph algorithms into your analytical workflows. Neo4j offers a Python client for Neo4j Graph Data Science library that seamlessly allows you to execute graph algorithms using only Python code.

Most Frequent Document Communities

Relevant

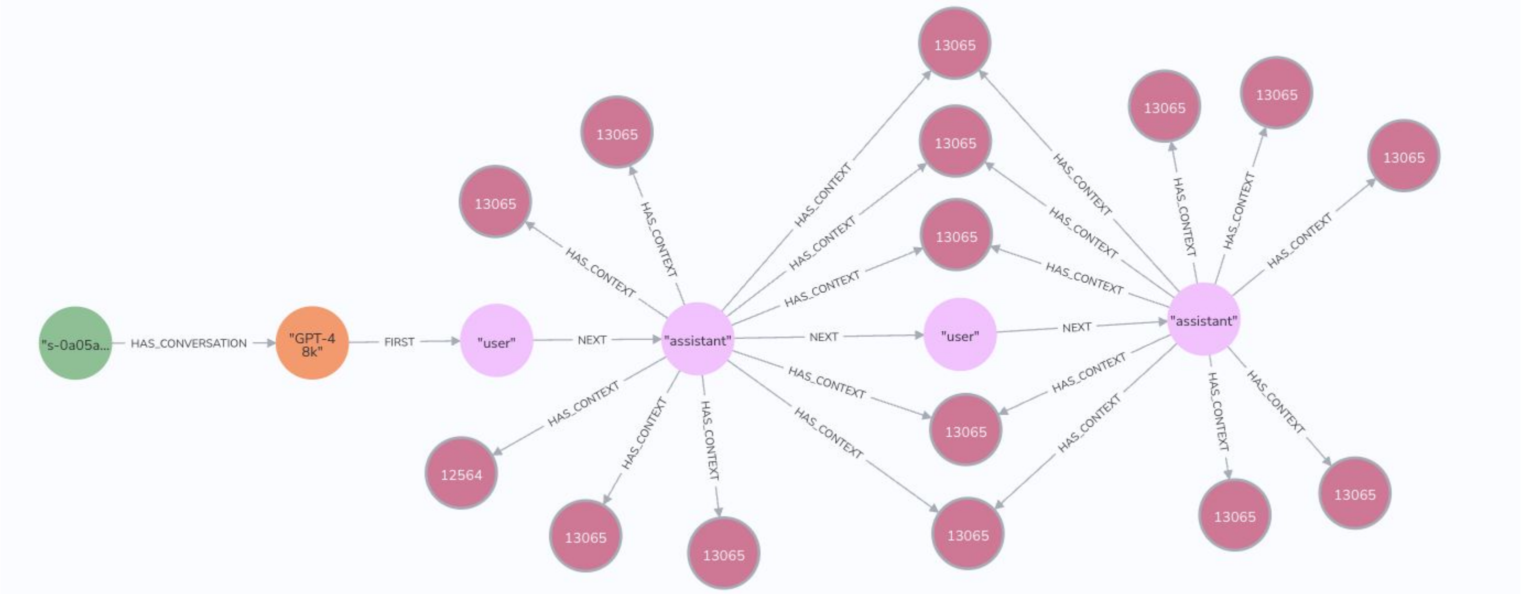


- Combining Document usage frequency with previously identified Document Communities, we can see which of these Communities are the most frequent sources of Context Documents

- The second most frequent Community (11649) comprises text chunks from blogs by Tomaz Bratanic

Logging and Visualizing Conversations

Graphs enable logging of LLM conversations in the same database as the context documents and with defined relationships.

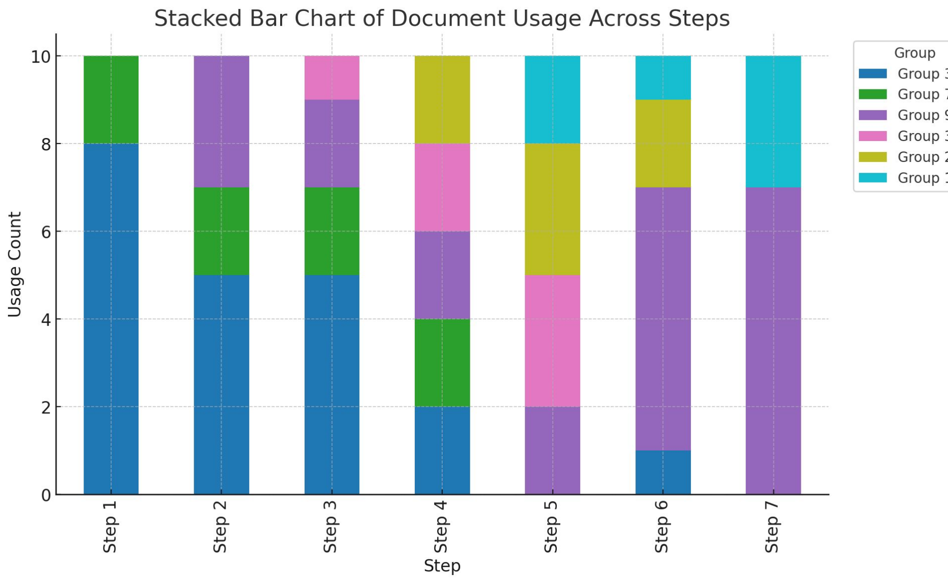


Graph of an actual conversation between an Agent Neo user and the ChatGPT-4 LLM. Context Documents are labeled with their GDS Community.

Visualizing Document Usage

Relevant

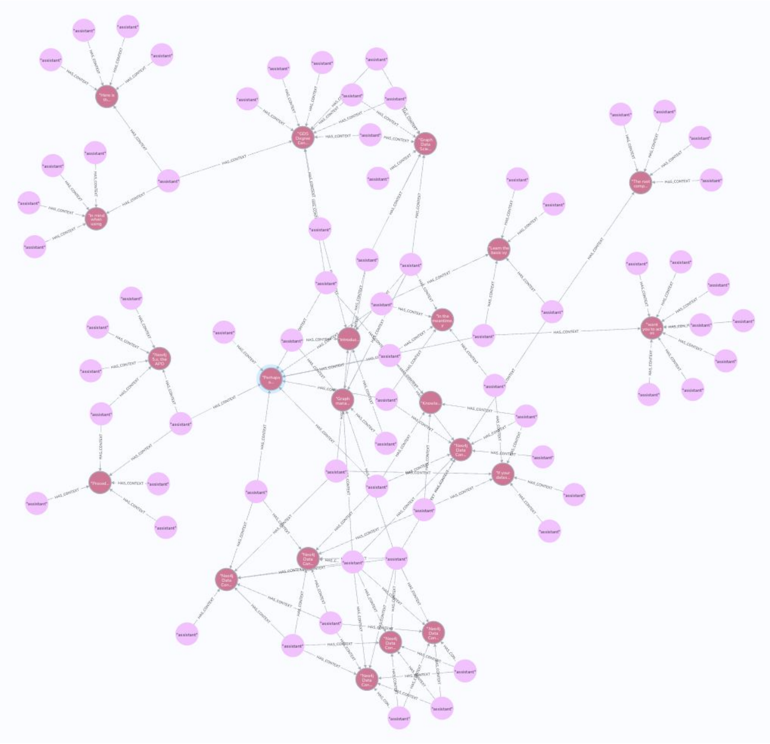
- The heat map depicts how frequently documents are used during a single conversation:
  - X-axis represents LLM messages
  - Y-axis represents individual documents
  - Color depicts document use frequency count (1x to 3x)
- Documents are re-used throughout conversations



Visualizing Context Document Usage

- Graphs enable us to visualize the most frequently used context Documents along with the associated LLM responses

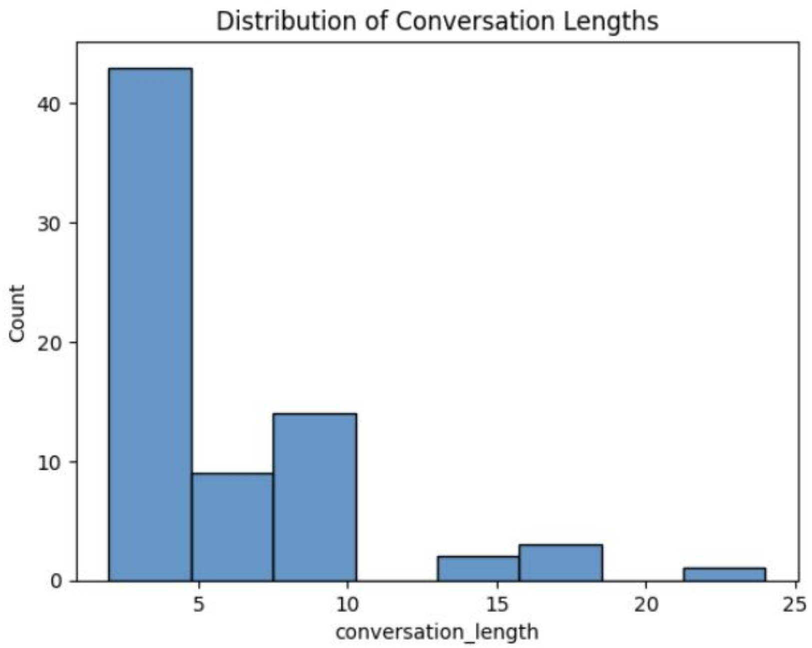
- Natural clusters form in the graph even among the most frequently used Documents



LLM Responses (pink) and
Most Frequently Used Context Documents (red)

Conversation Lengths

- Because conversations are logged as graphs it is easy to measure the length and store it as a new property on each Conversation node.
- Most users have been experimenting with our tool, so we expect conversation lengths to be shorter.
- As users are more comfortable with these applications, we expect conversation lengths to increase.



Relevant, Augmenting, Clean

Inspect Communities of LLM Responses

We can use similar approaches to inspect each of the LLM response communities via traditional and GDS-based statistics.

community	size	med_numDocs	med_pageRank	med_textLen	med_wordCount	ratings_Good	ratings_Bad	NotRated
14725	25	6.0	1.009269	1911.0	246.0	15	2	8
14501	16	5.0	0.950335	2001.0	246.5	4	3	9
14699	14	10.0	0.862450	1778.0	219.5	6	2	6
14818	13	10.0	0.980136	1963.0	244.0	6	1	6
14685	11	10.0	0.800114	2089.0	246.0	5	4	2

Graph enable Explainable AI

Document Usage

How the LLM is using the various documents in your dataset

Application/ Agent Logic

You can track graph traversals to explain chain of actions

User Experience

Visualizing conversations with context.

Applications

Analyze performance and identify opportunities for improvement

Session Summary

1. **Data Sources**
The right sources for the right questions

2. **Data Understanding**
Unstructured EDA

3. **Data Quality**
Makings of a quality GenAI dataset

4. **Application Understanding**
Learnings from application use

Thank you!

For more information or questions about grounding your LLM application with a knowledge graph, please contact us via alison.cossette@neo4j.com

