

A person with glasses is shown in profile, pointing at a screen. The background is dark with many colorful, out-of-focus light spots (bokeh) in shades of blue, yellow, and purple. The person is wearing a light blue shirt.

Introducing Graph RAG

Enhancing language models with external knowledge

Guido Schmutz (Senior Manager & Architect)

 **accenture**

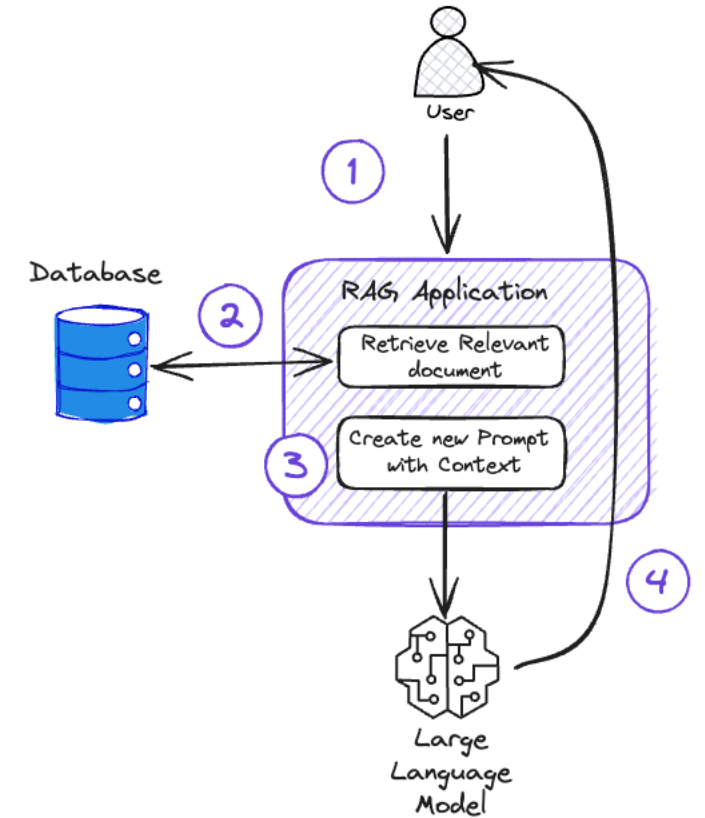
Retrieval Augmented Generation (RAG)

A software design pattern for integrating GenAI applications with custom data sources, like a database

1. RAG augments the LLM by intercepting a **user's prompt**
2. a **query to the database (knowledge base)** is made
3. using the query results as **context** (a.k.a. **grounding**) for the user's prompt, a **new prompt** is created that is passed to the LLM
4. LLM is used to **create complete, curated response**

Challenges:

- What is the **right context** to enhance question?
- What **database/data model** to use as the knowledge base?
- Which **large language model (LLM)** to use to generate the answer?



Provide LLM with data **it is not aware of** (either because of time or data privacy)

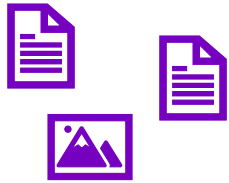
No need to fine-tune the LLM to customize results => save time and money

Using local LLM's, context (question and additional data) **will not leave the company**

The Sources of Data for (Graph) RAG

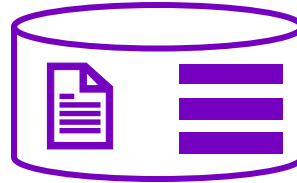
Each with different access patterns, supporting different kind of questions

Pure Text



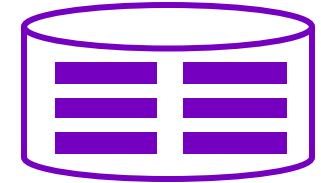
Unstructured data in PDFs,
plain text files, or images

**Mixed
Text + Data**



Structured data together with
long form text
Highly structured documents

Pure Data



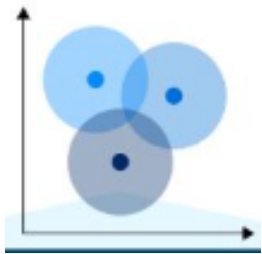
Structured data in a database

The Sources of Data for (Graph) RAG

Each with different access patterns, supporting different kind of questions

Pure Text

Vector Search



Find relevant documents plus context for
information search

Mixed
Text + Data

Search & Pattern Matching



Expand context and rank relevance for
information discovery

Pure Data

Graph Queries



Directly query the knowledge graph for
information query

Graph RAG includes a **graph database** as a source of the contextual information sent to the LLM

Demo Use Case – IIHF Official Rulebook

Let's see if we can ask questions about the rules of ice hockey



No matter where ice hockey is played, the object of the game is the same – **to put the puck into the opponent's goal**

The comprehensive IIHF rule book **ensures a consistent set of rules for all**, maintaining a fair and uniform standard of play globally, preserving the "language" of the game.



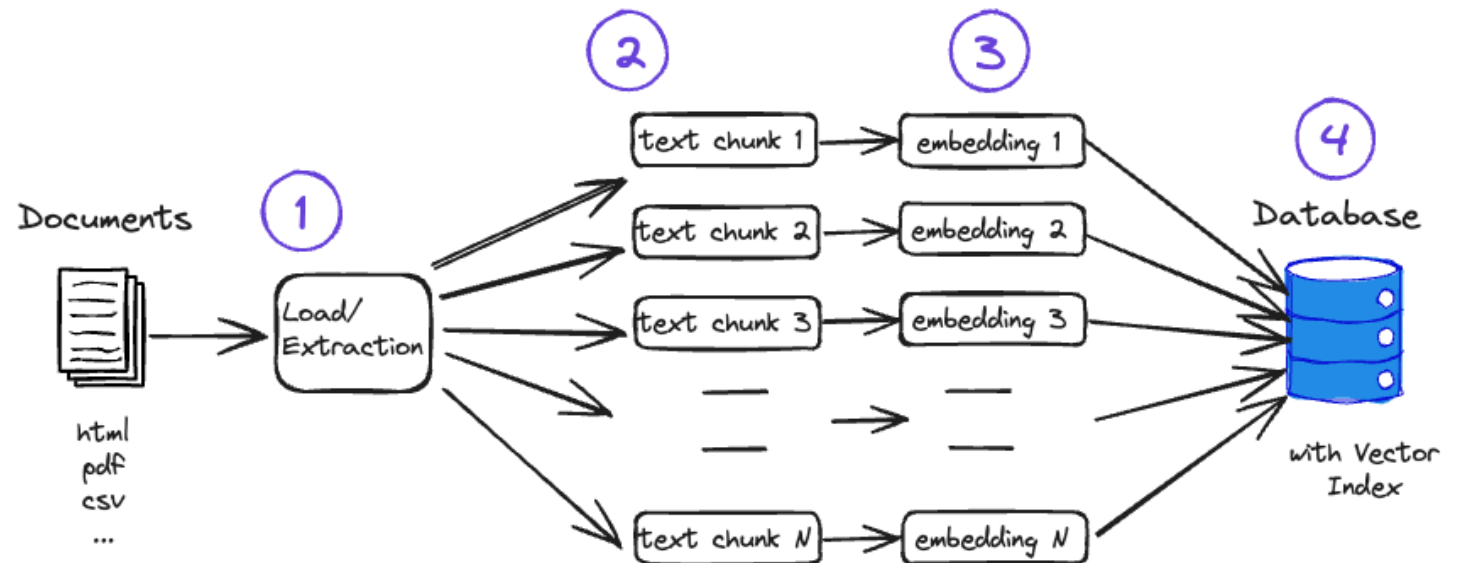
Preprocessing

An LLM has limited context window, therefore we have to provide it with the right information

1. **Load** the document(s) and **extract** the information (PDF) as text
2. **Split text data** into **right-sized chunks** to improve information retrieval and because LLM has a limited context window
3. Create **embeddings** for **each chunk**
4. **Store embedding** and corresponding text in a **vector index** in the database of your choice

Challenges :

- How **extract information** from PDF?
- What is the right **chunking strategy** and **size**?
- Which **embedding model** to use?
- Which **(vector) database** to use?
- How to **model the (knowledge) database**?



Preprocessing– Extract text from PDF

Convert PDF to a markup language such HTML or Markdown to simplify further processing

This is **known to be a hard problem** and I did not get a satisfying result with the options I have tried so far

- [Unstructured.io](#)
- [LLMSherpa](#)
- [LLamaParse](#) (example)

At the end I decided **to manually convert** the PDF into a markdown document

```
#### 3.2. PENALTY BOX

Each Rink must be provided with benches or seats to be known as the "Penalty Box".
Separate Penalty Boxes shall be provided for each Team and they shall be situated on
the opposite sides directly across the ice from their Players' Benches. Teams must
use the "Penalty Box" opposite their Players' Bench and must use the same "Penalty
Box" for the duration of a game.
Each "Penalty Box" should be at least 4.0 m in length and 1.50 m in width and shall
be separated from the spectators by a Protective Glass to afford the necessary
protection for the Players.
Each Penalty Box must be of the same size and quality, offering no advantage to
either Team in any manner and must have only one door for both entry and exit and
must be operated only by the "Penalty Box Attendant".
Only the Penalty Box Attendant, penalized Skaters, and Game Officials are allowed
access to the Penalty Boxes.

### RULE 4 SIGNAL AND TIMING DEVICES

#### 4.1. SIGNAL DEVICES

Each Rink must be provided with a suitable sound device that will sound
automatically at the conclusion of each period of play. Should the sound device fail
to sound automatically when time expires, the determining factor as to whether the
period has ended shall be the Game Clock.

#### 4.2. TIMING DEVICES

Each Rink shall be provided with some form of electronic game clock for the purpose
of keeping the spectators, Players, Team Personnel and Game Officials accurately
informed as to all time elements at all stages of the game including the time
remaining to be played in any period and the time remaining to be served by
penalized Players on each Team.
```

```
3.2. PENALTY BOX

Each Rink must be provided with benches or seats to be known as the "Penalty Box".

Separate Penalty Boxes shall be provided for each Team and they shall be situated on the opposite sides directly across the ice from
their Players' Benches. Teams must use the "Penalty Box" opposite their Players' Bench and must use the same "Penalty Box" for
the duration of a game.
Each "Penalty Box" should be at least 4.0 m in length and 1.50 m in width and shall be separated from the spectators by a Protective
Glass to afford the necessary protection for the Players.

Each Penalty Box must be of the same size and quality, offering no advantage to either Team in any manner and must have only one
door for both entry and exit and must be operated only by the "Penalty Box Attendant".
Only the Penalty Box Attendant, penalized Skaters, and Game Officials are allowed access to the Penalty Boxes.
→ For more information refer to Appendix VI – Infographics

RULE 4 SIGNAL AND TIMING DEVICES
---
4.1. SIGNAL DEVICES
# PLAYING AREA

Each Rink must be provided with a suitable sound device that will sound automatically at the conclusion of each period of play. Should the sound device fail
to sound automatically when time expires, the determining factor as to whether the period has ended shall be the Game Clock.
← For more information refer to IIHF Sport Regulations.

3.2. PENALTY BOX

Each Rink must be provided with benches or seats to be known as the "Penalty Box".

Separate Penalty Boxes shall be provided for each Team and they shall be situated on the opposite sides directly across the ice
from their Players' Benches. Teams must use the "Penalty Box" opposite their Players' Bench and must use the same "Penalty Box"
for the duration of a game.

Each "Penalty Box" should be at least 4.0 m in length and 1.50 m in width and shall be separated from the spectators by a Protec
tive Glass to afford the necessary protection for the Players.

Each Penalty Box must be of the same size and quality, offering no advantage to either Team in any manner and must have only one
door for both entry and exit and must be operated only by the "Penalty Box Attendant".

Only the Penalty Box Attendant, penalized Skaters, and Game Officials are allowed access to the Penalty Boxes.

→ For more information refer to Appendix VI – Infographics

RULE 4 SIGNAL AND TIMING DEVICES

4.1. SIGNAL DEVICES

Each Rink must be provided with a suitable sound device that will sound automatically at the conclusion of each period of play.
Should the sound device fail to sound automatically when time expires, the determining factor as to whether the period has ended
shall be the Game Clock.

» For more information refer to IIHF Technology Codes & Regulations.

4.2. TIMING DEVICES

Each Rink shall be provided with some form of electronic game clock for the purpose of keeping the spectators, Players, Team Per
sonnel and Game Officials accurately informed as to all time elements at all stages of the game including the time remaining
to be played in any period and the time remaining to be served by penalized Players on each Team.
```

Preprocessing – Chunking Strategy and Size

Breaking up data into smaller pieces or chunks is a crucial step in a RAG solution

IIHF OFFICIAL RULE BOOK 2023/24 – SECTION 02

28

TEAMS

RULE 8 INJURED PLAYERS

8.1. INJURED PLAYER

When a Player is injured or compelled to leave the ice during a Game, they may retire from the Game and be replaced by a substitute, but play must continue without the Teams leaving the ice.

During the play, if an injured Player wishes to retire from the ice and be replaced by a substitute, they must do so at the Players' Bench and not through any other exit leading from the Rink. This is not a legal Player change and therefore when a violation occurs, a Bench-minor Penalty shall be imposed.

If a penalized Player has been injured, they may proceed to the Dressing Room without taking a seat in the Penalty Box. The penalized Team shall immediately put a substitute Player in the Penalty Box, who shall serve the penalty until the injured Player is able to return to the game. They would replace their Teammate in the Penalty Box at the next stoppage of play. For violation of this rule, a Bench Minor Penalty shall be imposed.

Should the injured penalized Player who has been replaced in the Penalty Box return to their Players' Bench prior to the expiration of their penalty, they shall not be eligible to play until their penalty has expired. This includes coincidental penalties when their substitute is still in the Penalty Box awaiting a stoppage in play.

The injured Player must wait until their substitute has been released from the Penalty Box before they are eligible to play. If, however, there is a stoppage of play prior to the expiration of their penalty, they must then replace their Teammate in the Penalty Box and is then eligible to return once their penalty has expired.

When a Player is injured so that they cannot continue play or go to their Players' Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

In the case where it is obvious that a Player has sustained a serious injury, the Referee and/or Linesperson may stop the play immediately. Where an injury has occurred to a Player and there is a stoppage of play, a Team Doctor (or other Medical Personnel) may go onto the ice to attend to the injured Player without waiting for the Referee's consent.

When play has been stopped by the Referee or Linesperson due to an injured Player, or whenever an injured Player is attended to on the ice by the Coach or Medical Personnel, such Player must be substituted for immediately. This injured Player cannot return to the ice until play has resumed.

When play is stopped for an injured Player, the ensuing "face-off" shall be conducted at the Face-off Spot in the zone nearest the location of the puck when the play was stopped.

When the injured Player's Team has control of the puck in the Attacking Zone, the "face-off" shall be conducted at the nearest Face-off Spot in the Neutral Zone.

When the injured Player is in their Defending Zone and the attacking Team is in "possession of the puck" in the Attacking Zone, the "face-off" shall be conducted at the nearest Face-off Spot in the defending Team's zone.

A player who lies on the ice either feigning an injury or refusing to get up off the ice will be issued a Minor Penalty.

Chunking

ChunkViz v0.1

Language Models do better when they're focused.

One strategy is to pass a relevant subset (chunk) of your full data. There are many ways to chunk text.

This is an tool to understand different chunking/splitting strategies.

[Explain like I'm 5...](#)

Player without waiting for the Referee's consent.

When play has been stopped by the Referee or Linesperson due to an injured Player, or whenever an injured Player is attended to on the ice by the Coach or Medical Personnel, such Player must be substituted for immediately. This injured Player cannot return to the ice until play has resumed.

When play is stopped for an injured Player, the ensuing "face-off" shall be conducted at the Face-off Spot in the zone nearest the location of the puck when the play was stopped.

When the injured Player's Team has control of the puck in the Attacking Zone, the "face-off" shall be

Upload.txt

Splitter: Recursive Character Text Splitter

Chunk Size: 600

Chunk Overlap: 0

Total Characters: 3271

Number of chunks: 8

Average chunk size: 408.9

When a Player is injured or compelled to leave the ice during a Game, they may retire from the Game and be replaced by a substitute, but play must continue without the Teams leaving the ice.

During the play, if an injured Player wishes to retire from the ice and be replaced by a substitute, they must do so at the Players' Bench and not through any other exit leading from the Rink. This is not a legal Player change and therefore when a violation occurs, a Bench-minor Penalty shall be imposed.

If a penalized Player has been injured, they may proceed to the Dressing Room without taking a seat in the Penalty Box. The penalized Team shall immediately put a substitute Player in the Penalty Box, who shall serve the penalty until the injured Player is able to return to the game. They would replace their Teammate in the Penalty Box at the next stoppage of play.

For violation of this rule, a Bench Minor Penalty shall be imposed.

Should the injured penalized Player who has been replaced in the Penalty Box return to their Players' Bench prior to the expiration of their penalty, they shall not be eligible to play until their penalty has expired. This includes coincidental penalties when their substitute is still in the Penalty Box awaiting a stoppage in play.

The injured Player must wait until their substitute has been released from the Penalty Box before they are eligible to play. If, however, there is a stoppage of play prior to the expiration of their penalty, they must then replace their Teammate in the Penalty Box and is then eligible to return once their penalty has expired.

When a Player is injured so that they cannot continue play or go to their Players' Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

<https://chunkviz.up.railway.app/>

Preprocessing – Which Embedding Model?

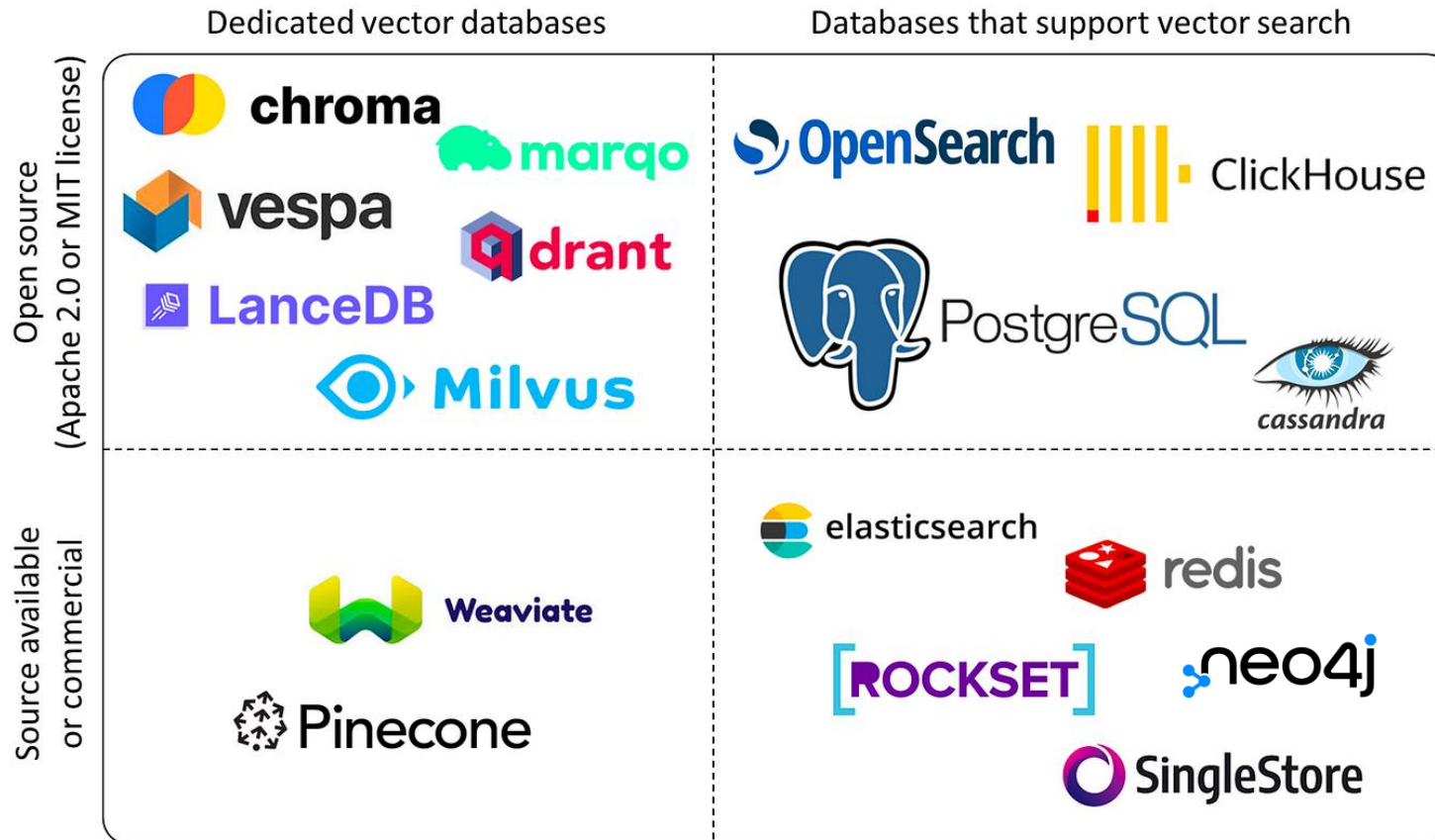
Decided to use **mxbai-embed-large-v1** as it is part of Ollama => **max tokens = 512** must be taken into account when choosing chunk size => for each chunk we create an embedding

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
8	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26
9	google-gecko.text-embedding-ada-002	1200	4.47	768	2048	66.31	81.17	47.48
10	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14
11	gte-large-en-v1.5	434	1.62	1024	8192	65.39	77.75	47.96
12	LLM2Vec-Meta-Llama-3-supervised	7505	27.96	4096	8192	65.01	75.92	46.45
13	LLM2Vec-Mistral-supervised	7111	26.49	4096	32768	64.8	76.63	45.54
14	echo-mistral-7b-instruct-latest	7111	26.49	4096	32768	64.68	77.43	46.32
15	mxbai-embed-large-v1	335	1.25	1024	512	64.68	75.64	46.71
16	UAE-Large-V1	335	1.25	1024	512	64.64	75.58	46.73
17	text-embedding-3-large			3072	8191	64.59	75.45	49.01

<https://huggingface.co/spaces/mteb/leaderboard>

Preprocessing – Which Vector DB to use?

Using a graph database (Neo4J) combines vector searches with graph searches and traversal



Source: <https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c>

Demo Use Case - Toolstack used



Neo4J

popular native graph database

- Native Labeled Property Graph
 - Cypher Query Language
 - Graph Algorithms
- Provides Vector Indices



Ollama

Running LLMs locally

- Provides access to many high-quality LLMs
- Can be installed on Mac, Windows or Linux
- Uses GPU if available



Langchain

Simplify the creating of apps leveraging LLMs

- Powerful tool for building LLM apps
- provides components, such as prompt templates, chains, agents, memory, and document loaders
- Integrates with various vector stores



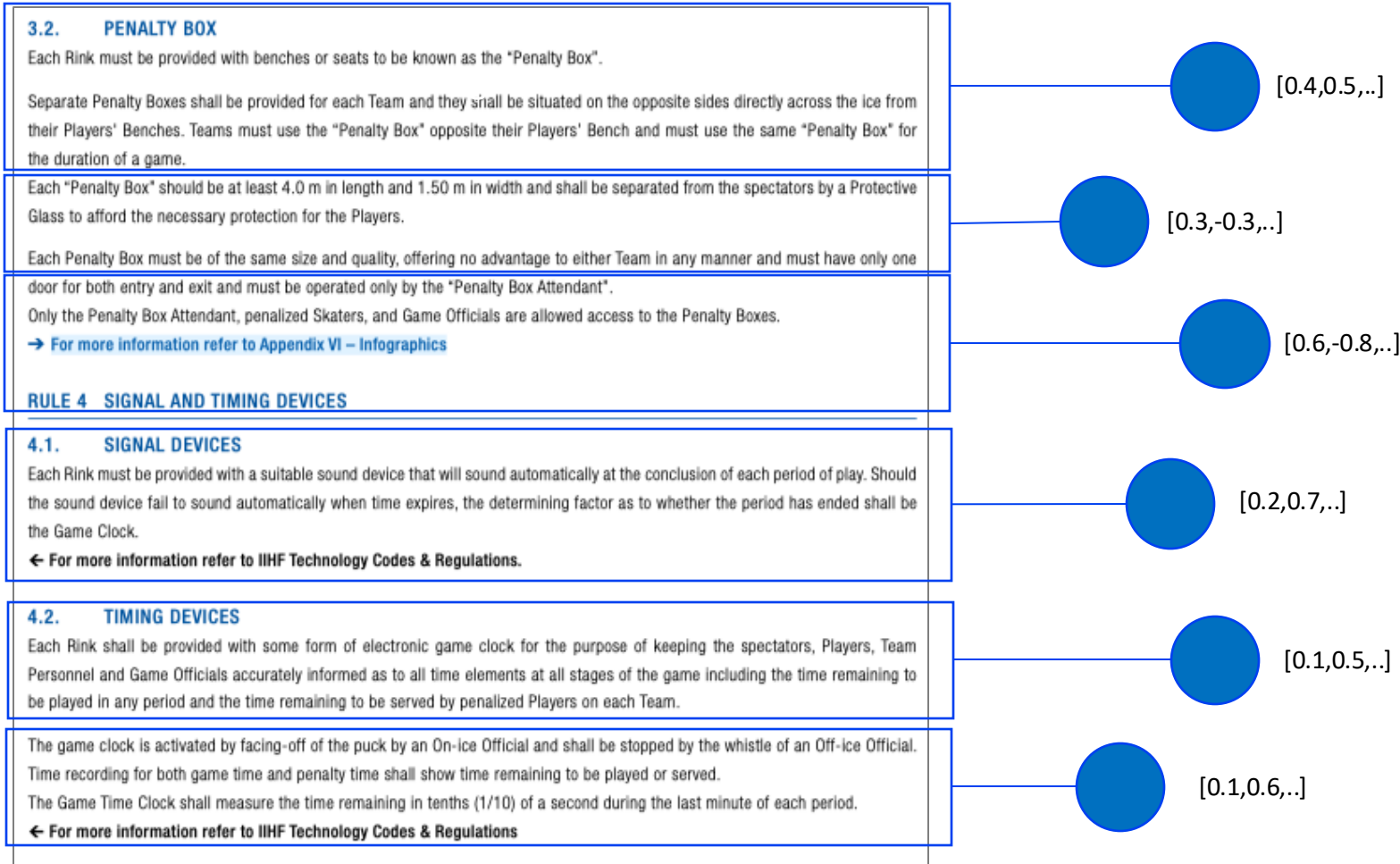
Platys

Accelerator for setting up modern data platforms

- Generate docker compose stacks
- To be used for Proof-of-Concepts (PoC) or trainings
- Very lightweight, runs on a single machine
 - Open Source

Preprocessing – Data Preparation

Traditional RAG with chunking on full text



Preprocessing – Data Preparation

Using RecursiveCharacterTextSplitter to chunk on the full text

```
[3]: # Splitting text into chunks using the RecursiveCharacterTextSplitter
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 600,
    chunk_overlap = 0,
    length_function = len,
    is_separator_regex = False,
)
```

Text splitter demonstration

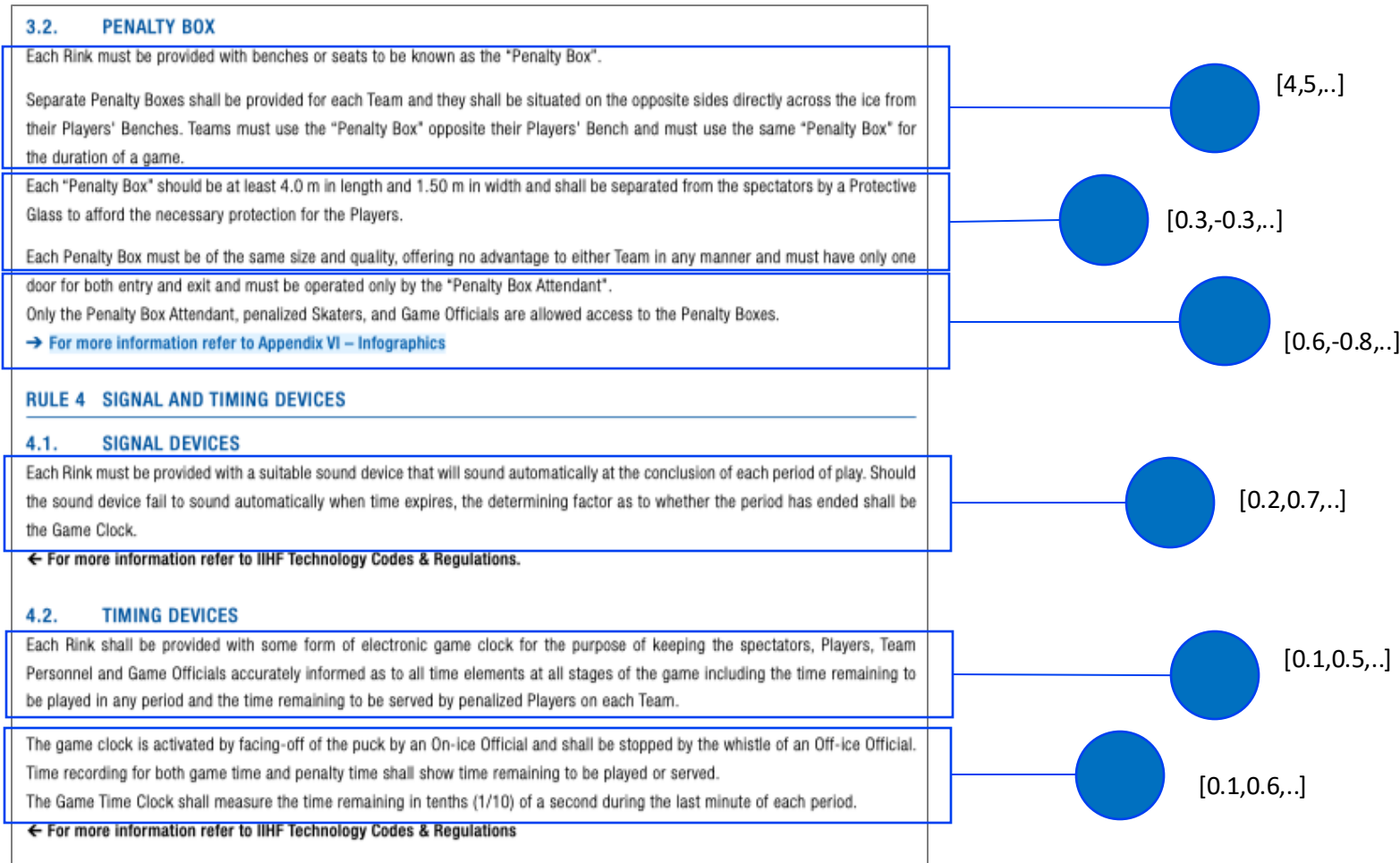
You can see what the text splitter will do by splitting up the `page_content`.

```
[4]: text_chunks = text_splitter.split_text(documents[0].page_content)
text_chunks[19]
```

```
[4]: 'The Goal posts shall be of an approved design and material, extending vertically 1.22 m above the surface of the ice and set 1.83 m apart measured from the inside of the posts. A crossbar of the same material as the Goal posts shall extend from the top of one post to the top of the other. The Goal posts and crossbar shall be painted in red color and all other exterior surfaces shall be painted in white color.\n\n#### 2.2 GOAL NETS'
```

Preprocessing – Data Preparation

Traditional RAG with consideration of document structure for chunking



Preprocessing – Data Preparation

Using MarkdownHeaderTextSplitter with RecursiveCharacterTextSplitter to chunk text of section

Markdown Header Text Splitter

We first use the Markdown Header Text splitter to split on the structure of the markdown document (using Header 1 - 4).

```
print (documents[0].metadata["source"])
print (len(documents))

headers_to_split_on = [
    ("#", "header1"),
    ("##", "header2"),
    ("###", "header3"),
    ("####", "header4"),
]

markdown_splitter = MarkdownHeaderTextSplitter(headers_to_split_on=headers_to_split_on, strip_header=True)
md_header_splits = markdown_splitter.split_text(documents[0].page_content)

/data-transfer/iihf/rulebook.md
1

md_header_splits[0]
```

```
Document(page_content='No matter where ice hockey is played, the object of the game is the same – to e
puck into the opponent's goal. Beyond that, ice hockey across the globe is subject to certain vari
This makes the rules of the game extremely important. These rules must be followed all times, in all
ies, in all age categories, for the game to be enjoyed by everyone. \nHockey's speed is one of the
es that makes it so exciting. But this skill and excitement must be balanced with fair play and resp
\nIt is, therefore, important to make a clear separation between the purpose of all the elements of
e and to use these respectfully. These distinctions can be taught at an early age or whenever one be
show interest in the game. And this is why hockey development begins with parents and coaches, those
most influential in guiding a person, old or young, into playing the game properly and within the ru
\nThe IIHF Championship program encompasses 81 Member National Associations, five age and gender cat
over 30 international ice hockey tournaments, including the Olympic Winter Games. \nThe extensivene
he program is acknowledged in the rule book. The goal is to provide everyone one set of rules from w
work. This presents a fair and leveled standard of play. It is a means of keeping the game's "langua
same regardless of where it is played.', metadata={'header1': 'IIHF Official Rulebook 2023/24', 'hea
'Welcome'})
```

Recursive Character Text Splitter

and now also use the Recursive Character Text Splitter to further split the text blocks.

```
# Char-level splits
from langchain_text_splitters import RecursiveCharacterTextSplitter

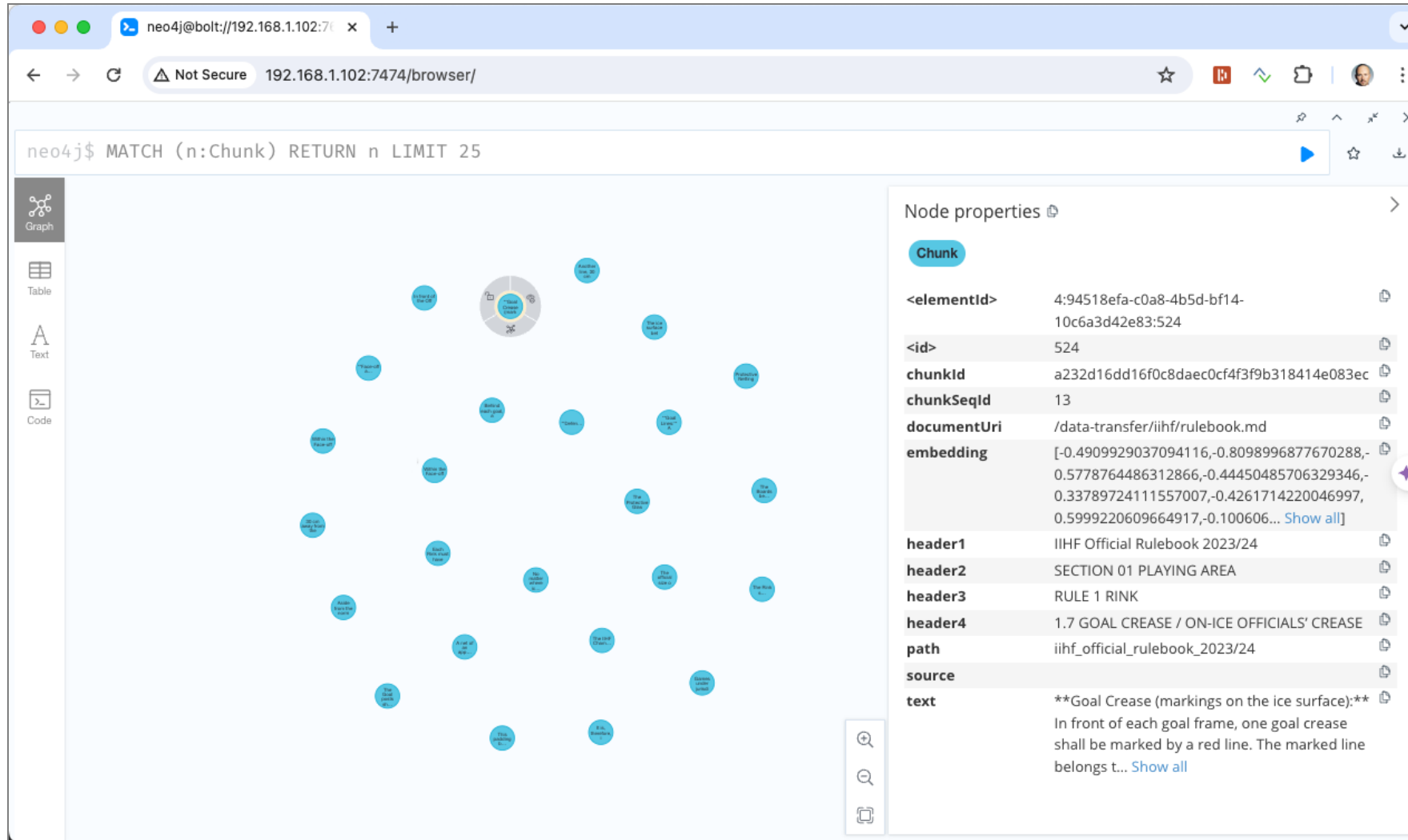
chunk_size = 600
chunk_overlap = 0
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size=chunk_size,
    chunk_overlap=chunk_overlap,
    length_function = len,
    is_separator_regex = False,
)

# Split
chunks = text_splitter.split_documents(md_header_splits)
chunks[0]
```

```
Document(page_content='No matter where ice hockey is played, the object of the game is the same – to put th
e puck into the opponent's goal. Beyond that, ice hockey across the globe is subject to certain variations.
This makes the rules of the game extremely important. These rules must be followed all times, in all countr
ies, in all age categories, for the game to be enjoyed by everyone. \nHockey's speed is one of the qualiti
es that makes it so exciting. But this skill and excitement must be balanced with fair play and respect.',
metadata={'header1': 'IIHF Official Rulebook 2023/24', 'header2': 'Welcome'})
```

Preprocessing – Data Preparation

Chunks with the embeddings as loaded into Neo4j



The screenshot shows the Neo4j Browser interface. The top bar indicates the connection to 'neo4j@bolt://192.168.1.102:7474'. The address bar shows '192.168.1.102:7474/browser/'. The main query area contains the Cypher query: `neo4j$ MATCH (n:Chunk) RETURN n LIMIT 25`. The left sidebar shows navigation options: Graph, Table, Text, and Code. The central area displays a graph view with blue circular nodes representing data chunks. The right sidebar shows the 'Node properties' for a selected chunk.

Node properties

- Chunk**
- <elementId>**: 4:94518efa-c0a8-4b5d-bf14-10c6a3d42e83:524
- <id>**: 524
- chunkId**: a232d16dd16f0c8daec0cf4f3f9b318414e083ec
- chunkSeqId**: 13
- documentUri**: /data-transfer/iihf/rulebook.md
- embedding**: [-0.4909929037094116,-0.8098996877670288,-0.5778764486312866,-0.44450485706329346,-0.33789724111557007,-0.4261714220046997,0.5999220609664917,-0.100606... [Show all](#)]
- header1**: IIHF Official Rulebook 2023/24
- header2**: SECTION 01 PLAYING AREA
- header3**: RULE 1 RINK
- header4**: 1.7 GOAL CREASE / ON-ICE OFFICIALS' CREASE
- path**: iihf_official_rulebook_2023/24
- source**
- text**: ****Goal Crease (markings on the ice surface):****
In front of each goal frame, one goal crease shall be marked by a red line. The marked line belongs to... [Show all](#)

Testing RAG Chain

Of the Traditional RAG Strategy

```
# Create a langchain vector store from the existing Neo4j knowledge graph.
neo4j_vector_store = Neo4jVector.from_existing_graph(
    embedding=embeddings_api,
    url=NEO4J_URI,
    username=NEO4J_USERNAME,
    password=NEO4J_PASSWORD,
    index_name=VECTOR_INDEX_NAME,
    node_label=VECTOR_NODE_LABEL,
    text_node_properties=[VECTOR_SOURCE_PROPERTY],
    embedding_node_property=VECTOR_EMBEDDING_PROPERTY,
)

# RAG prompt
prompt = hub.pull("rlm/rag-prompt-llama")

# Create a retriever from the vector store
retriever = neo4j_vector_store.as_retriever(search_kwargs={'k': 3})

# Create a chatbot Question & Answer chain from the retriever
#chain = RetrievalQAWithSourcesChain.from_chain_type(
#    chat_api, chain_type="stuff", retriever=retriever
#)

chain = RetrievalQA.from_chain_type(
    chat_api,
    chain_type="stuff",
    retriever=retriever,
    verbose=True,
    chain_type_kwargs={"prompt": prompt, "verbose": True}
)

chain_traditional = prettifyChain(chain)
```

```
#prettyVectorSearch("what is the size of the rink?")
chain_traditional("what happens if player is injured")
```

> Entering new RetrievalQA chain...

> Entering new StuffDocumentsChain chain...

> Entering new LLMChain chain...

Prompt after formatting:

Human: [INST]<<SYS>> You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.<</SYS>>
>

Question: what happens if player is injured

Context:

text: When a Player is injured so that they cannot continue play or go to their Players' Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

Answer: [/INST]

> Finished chain.

> Finished chain.

> Finished chain.

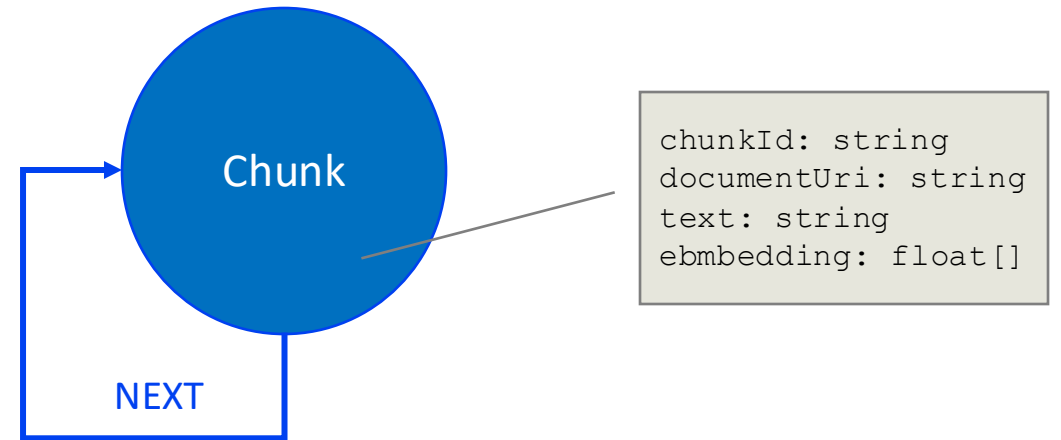
[/INST]<<SYS>> If a player is injured and cannot continue playing or go to their bench, play will not stop until their team gains control of the puck. If they're already in control when the injury occurs, play will only be stopped if they're in a scoring position. Otherwise, play will continue.

Preprocessing – Data Preparation (II)

Traditional RAG enhanced with Linked List of chunks => first step of Graph RAG

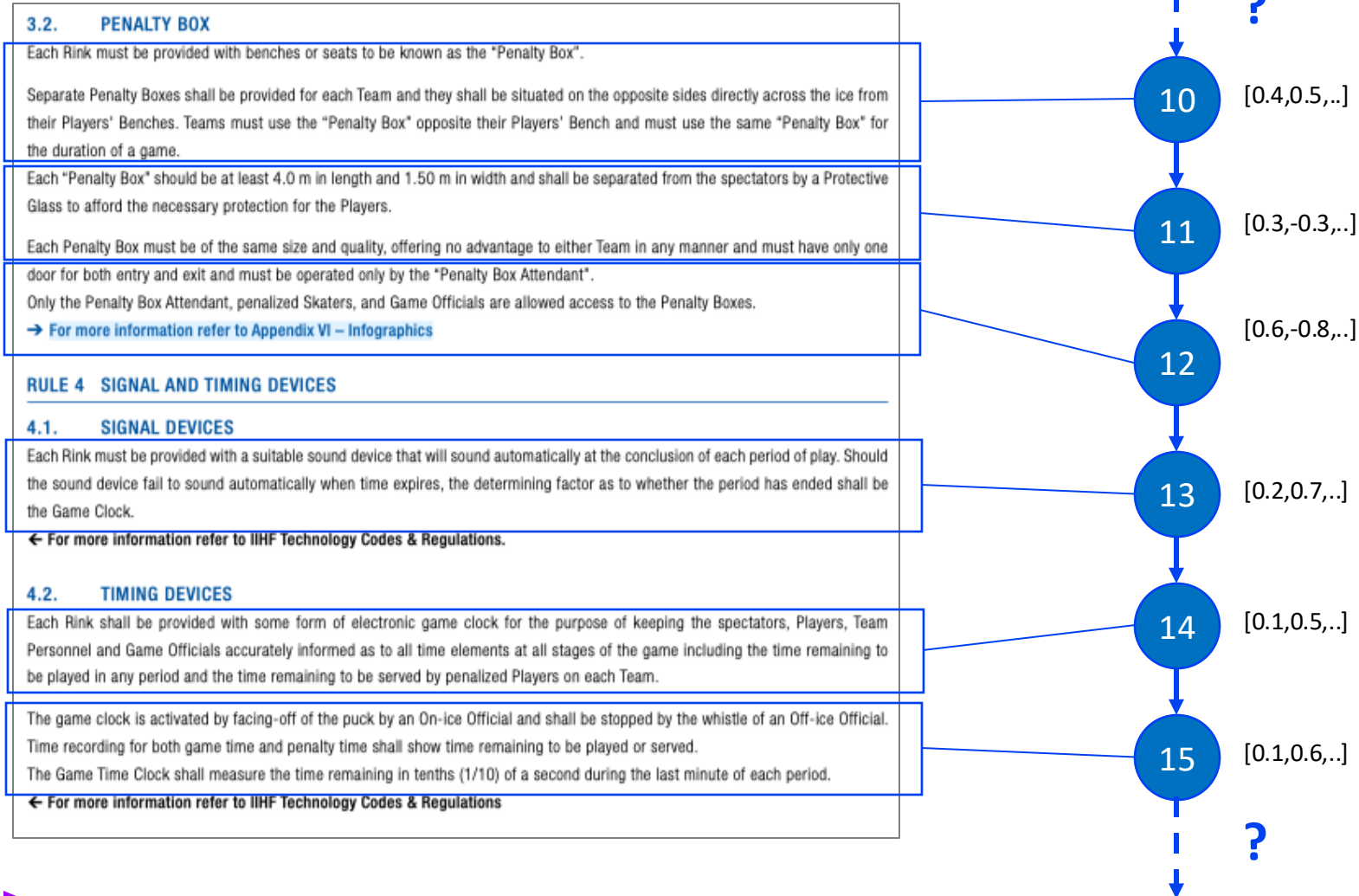
Benefits:

- **Vector similarity** search to find relevant text
- **Expand context** window with previous/next chunks
- Enable **paging** through text
- Chunking **no longer** requires **the chunks to overlap**



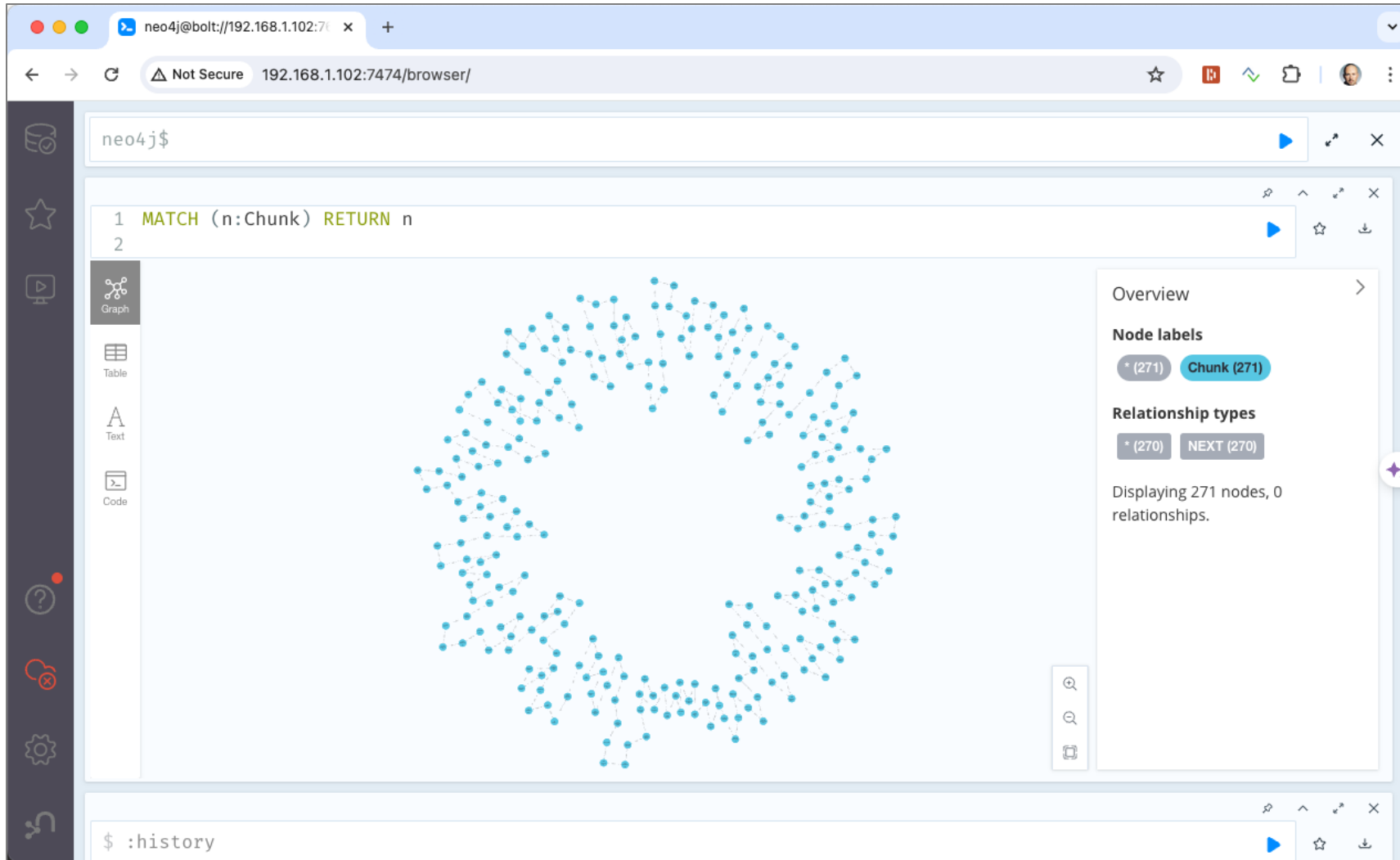
Preprocessing – Data Preparation (II)

Traditional RAG enhanced with Linked List of chunks



Preprocessing – Data Preparation

Chunks as linked list loaded into Neo4J



Testing RAG Chain

Of the Traditional RAG Strategy with Query Window

```
retrieval_query_window = """
OPTIONAL MATCH window=
  (:Chunk)-[:NEXT*0..1]->(node)-[:NEXT*0..1]->(:Chunk)
WITH node, score, window as longestWindow
ORDER BY node, length(window) DESC
WITH nodes(longestWindow) as chunkList, node, score
UNWIND chunkList as chunkRows
WITH collect(chunkRows.text) as textList, node, score
WITH apoc.text.join(textList, " \n ") as text,
score,
node {.source} AS metadata
RETURN text, score, metadata ORDER BY score DESC LIMIT 1
"""

vector_store_window = Neo4jVector.from_existing_index(
    embedding=embeddings_api,
    url=NEO4J_URI,
    username=NEO4J_USERNAME,
    password=NEO4J_PASSWORD,
    database="neo4j",
    index_name=VECTOR_INDEX_NAME,
    text_node_property=VECTOR_SOURCE_PROPERTY,
    retrieval_query=retrieval_query_window
)

# Create a retriever from the vector store
retriever_window = vector_store_window.as_retriever(search_kwargs={'k': 1})

# Create a chatbot Question & Answer chain from the retriever
chain_window = prettifyChain(RetrievalQA.from_chain_type(
    chat_api,
    chain_type="stuff",
    retriever=retriever_window,
    chain_type_kwargs={"verbose": True}
))
```

The injured Player must wait until their substitute has been released from the Penalty Box before they are eligible to play. If, however, there is a stoppage of play prior to the expiration of their penalty, they must then replace their Teammate in the Penalty Box and is then eligible to return once their penalty has expired.

When a Player is injured so that they cannot continue play or go to their Player's Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

In the case where it is obvious that a Player has sustained a serious injury, the Referee and/or Linesperson may stop the play immediately. Where an injury has occurred to a Player and there is a stoppage of play, a Team Doctor (or other Medical Personnel) may go onto the ice to attend to the injured Player without waiting for the Referee's consent.

The injured Player must wait until their substitute has been released from the Penalty Box before they are eligible to play. If, however, there is a stoppage of play prior to the expiration of their penalty, they must then replace their Teammate in the Penalty Box and is then eligible to return once their penalty has expired.

When a Player is injured so that they cannot continue play or go to their Player's Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

When a Player is injured so that they cannot continue play or go to their Player's Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

In the case where it is obvious that a Player has sustained a serious injury, the Referee and/or Linesperson may stop the play immediately. Where an injury has occurred to a Player and there is a stoppage of play, a Team Doctor (or other Medical Personnel) may go onto the ice to attend to the injured Player without waiting for the Referee's consent.

When a Player is injured so that they cannot continue play or go to their Player's Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

Human: what happens if player is injured

> Finished chain.

> Finished chain.

According to the context, when a Player is injured so that they cannot continue play or go to their Player's Bench:

- * The play shall not be stopped until the injured Player's Team has secured control of the puck.
- * If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.
- * Additionally, if it is obvious that a Player has sustained a serious injury:
- * The Referee and/or Linesperson may stop the play immediately.
- * A Team Doctor (or other Medical Personnel) may go onto the ice to attend to the injured Player without waiting for the Referee's consent.

Preprocessing – Data Preparation (III)

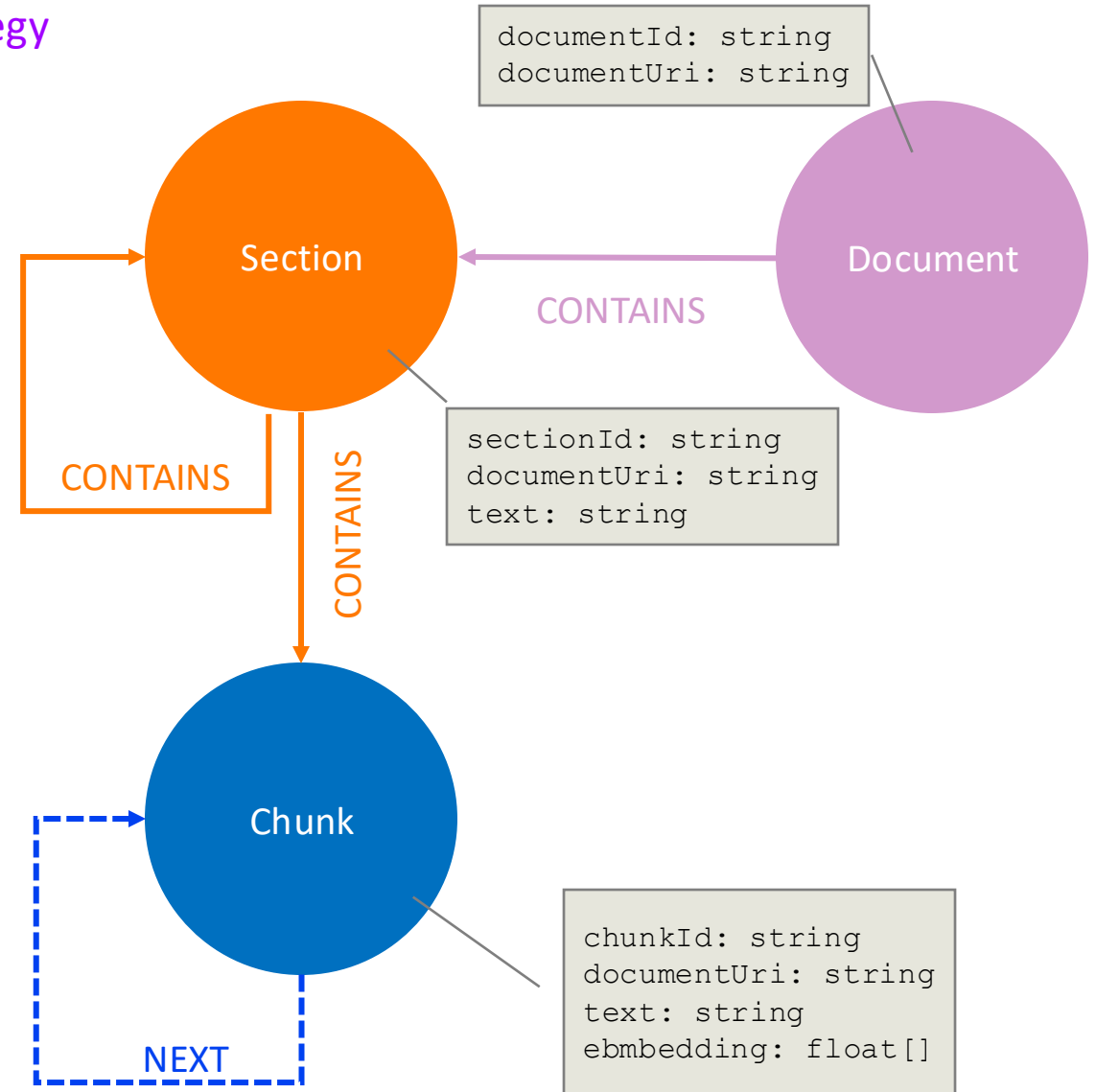
Enhance graph with structure to support Parent Retriever strategy

Create Section nodes for each Section (Header 1 – 4) and a Document node for the document

Connect from Section to all Chunk Nodes

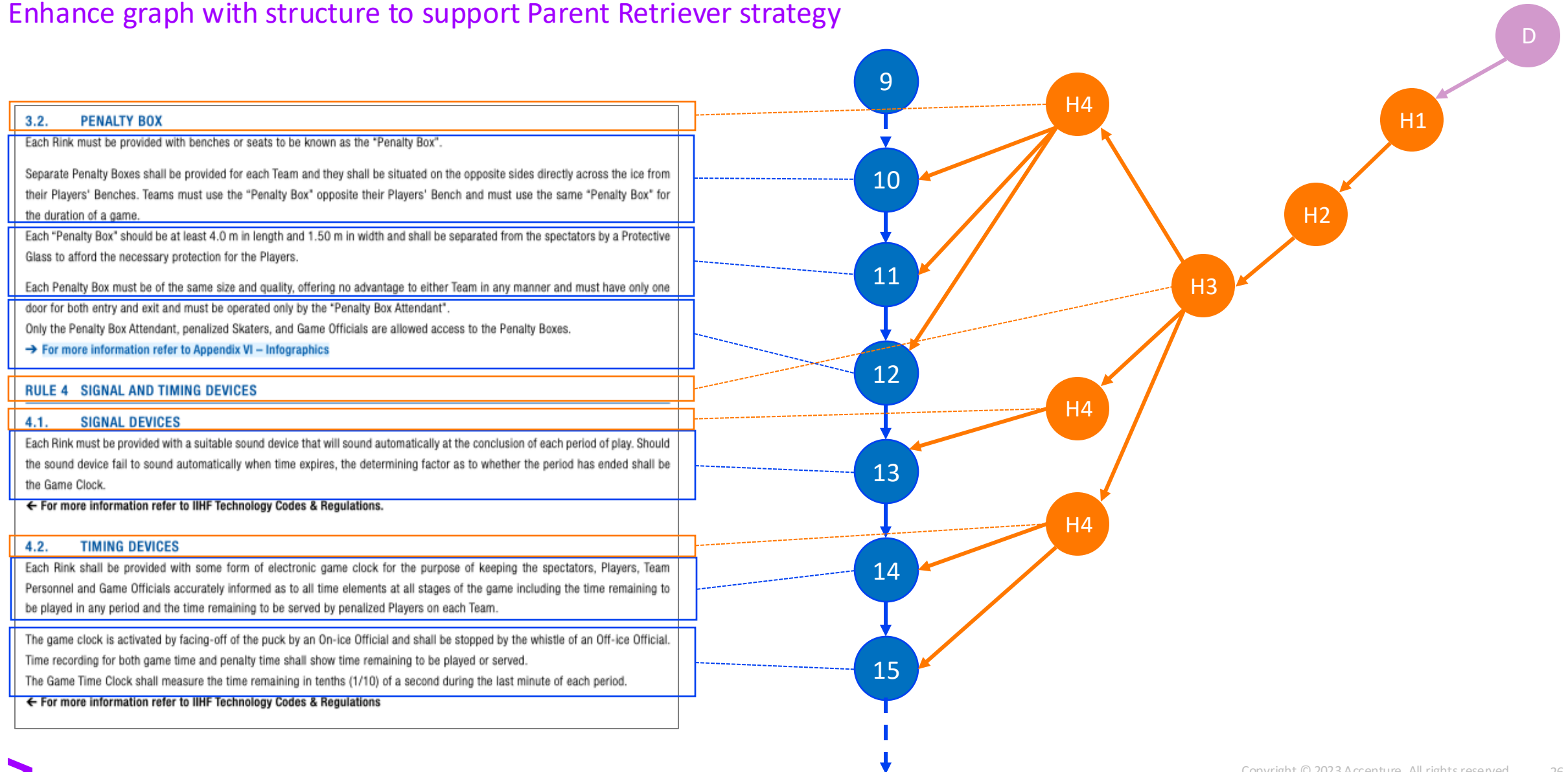
Benefits:

- Get all "sibling" chunks by traversing up from the most relevant chunk to the section and back to all other nodes
- We can also do that ordered so that the NEXT relationship is no longer needed
- Enhanced context with more relevant information



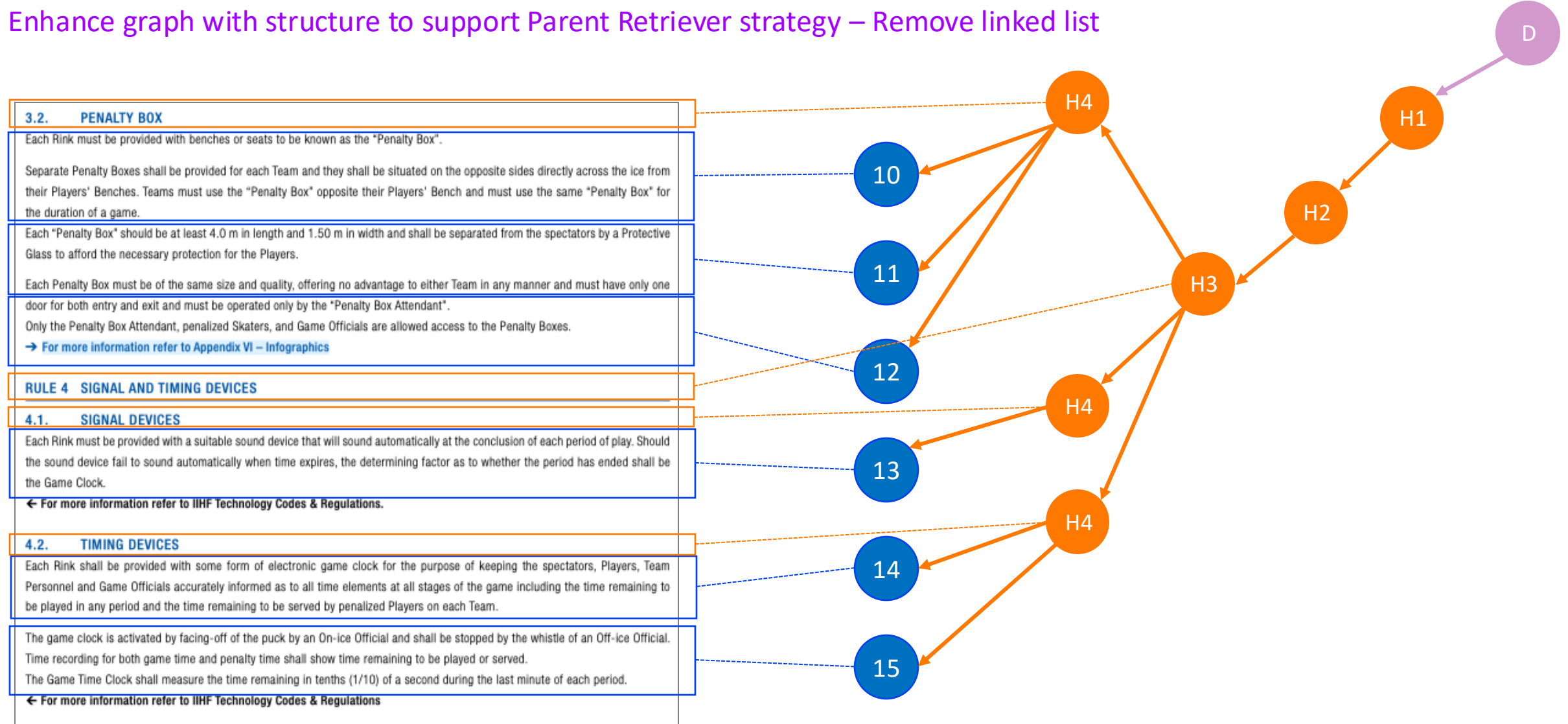
Preprocessing – Data Preparation (III)

Enhance graph with structure to support Parent Retriever strategy



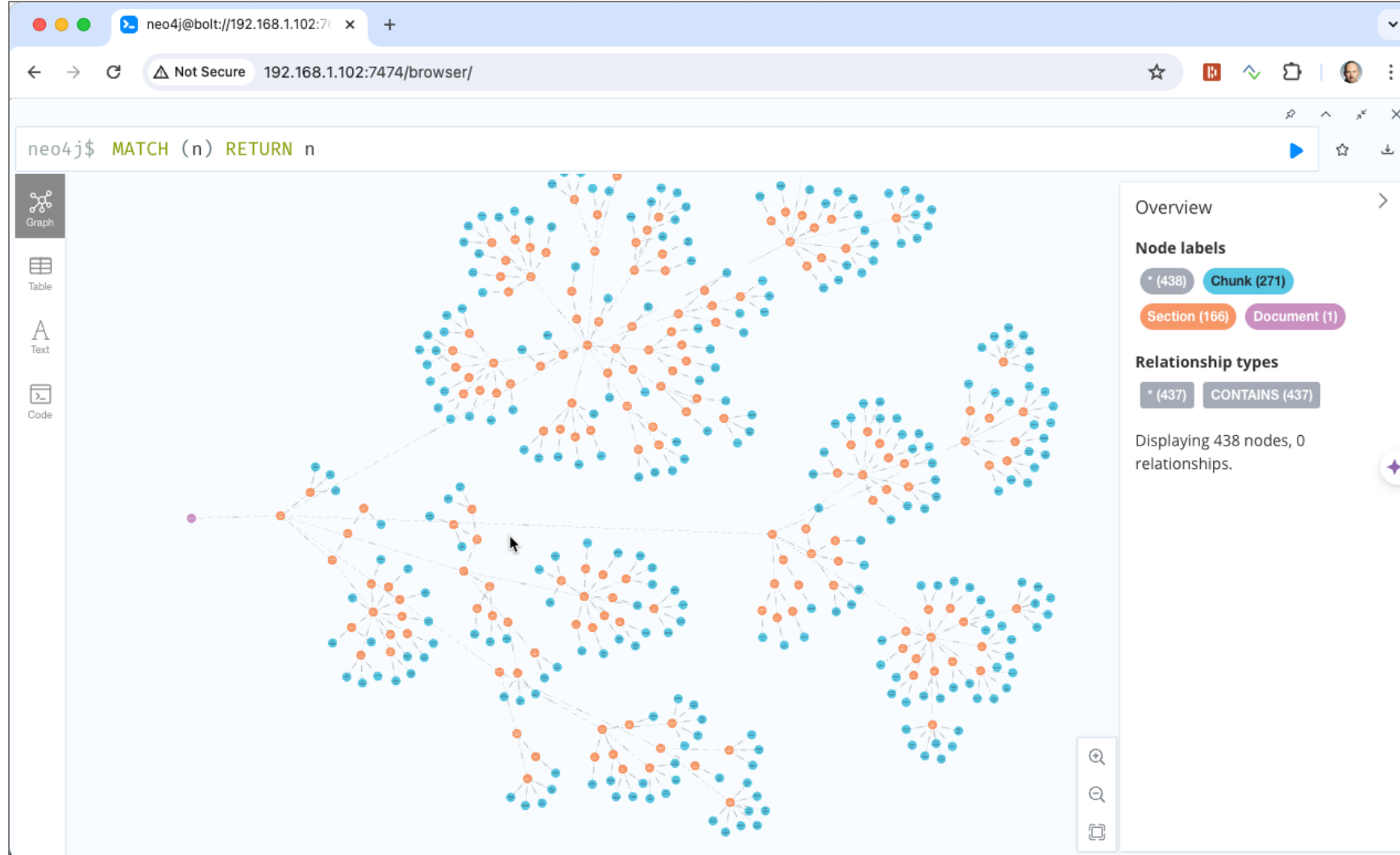
Preprocessing – Data Preparation (III)

Enhance graph with structure to support Parent Retriever strategy – Remove linked list



Preprocessing – Data Preparation (III)

Graph with Chunks, Section and Document



Testing RAG Chain

Of the Traditional RAG Strategy with Query Window

```
retrieval_query_parent = """"

    WITH node AS chunk, score AS score ORDER BY score
    OPTIONAL MATCH (chunk:Chunk)<--[:CONTAINS]-(parent)
    OPTIONAL MATCH (parent)-[:CONTAINS]->(s:Chunk)
    WITH chunk, s, score ORDER BY s.chunkSeqId ASC
    WITH collect(s.text) as textList, chunk.text as text, score AS score
    RETURN apoc.text.join(textList, " \n ") as text,
    score, {} AS metadata
    ORDER BY score desc

""""

vector_store_parent = Neo4jVector.from_existing_index(
    embedding=embeddings_api,
    url=NEO4J_URI,
    username=NEO4J_USERNAME,
    password=NEO4J_PASSWORD,
    database="neo4j",
    index_name=VECTOR_INDEX_NAME,
    text_node_property=VECTOR_SOURCE_PROPERTY,
    retrieval_query=retrieval_query_parent
)

# Create a retriever from the vector store
retriever_parent = vector_store_parent.as_retriever(search_kwargs={'k': 1})

# Create a chatbot Question & Answer chain from the retriever
chain_parent = prettifyChain(RetrievalQA.from_chain_type(
    chat_api,
    chain_type="stuff",
    retriever=retriever_parent,
    chain_type_kwargs={"verbose": True}
))
```

s' Bench, the play shall not be stopped until the injured Player's Team has secured control of the puck. If the Player's Team is in "control of the puck" at the time of injury, play shall be stopped immediately unless their Team is in a scoring position.

In the case where it is obvious that a Player has sustained a serious injury, the Referee and/or Linesperson may stop the play immediately. Where an injury has occurred to a Player and there is a stoppage of play, a Team Doctor (or other Medical Personnel) may go onto the ice to attend to the injured Player without waiting for the Referee's consent.

When play has been stopped by the Referee or Linesperson due to an injured Player, or whenever an injured Player is attended to on the ice by the Coach or Medical Personnel, such Player must be substituted for immediately. This injured Player cannot return to the ice until play has resumed.

When play is stopped for an injured Player, the ensuing "face-off" shall be conducted at the Face-off Spot in the zone nearest the location of the puck when the play was stopped.

When the injured Player's Team has control of the puck in the Attacking Zone, the "face-off" shall be conducted at the nearest Face-off Spot in the Neutral Zone.

When the injured Player is in their Defending Zone and the attacking Team is in "possession of the puck" in the Attacking Zone, the "face-off" shall be conducted at the nearest Face-off Spot in the defending Team's zone.

A Player who lies on the ice either feigning an injury or refusing to get up off the ice will be issued a Minor Penalty.

Human: what happens if a player is injured

> Finished chain.

> Finished chain.

According to the context, when a Player is injured or compelled to leave the ice during a Game: 1. The Player may retire from the Game and be replaced by a substitute. 2. Play must continue without the Teams leaving the ice. 3. If an injured Player wishes to retire from the ice and be replaced by a substitute, they must do so at the Players' Bench and not through any other exit leading from the Rink. 4. A violation of this rule results in a Bench-minor Penalty being imposed. Additionally: 1. If a penalized Player has been injured, they may proceed to the Dressing Room without taking a seat in the Penalty Box. 2. The penalized Team must immediately put a substitute Player in the Penalty Box, who shall serve the penalty until the injured Player is able to return to the game. 3. The injured Player can replace their teammate in the Penalty Box at the next stoppage of play. It's also important to note that: 1. If an injured penalized Player returns to their Players' Bench prior to the expiration of their penalty, they shall not be eligible to play until their penalty has expired. 2. The injured Player must wait until their substitute has been released from the Penalty Box before they are eligible to play. In summary, when a player is injured during a game: * Play continues without stopping * The injured player may retire and be replaced by a substitute * If the injured player was penalized, their team must put a substitute in the penalty box * The injured player cannot return to play until their penalty has expired or they replace their teammate in the penalty box.

Preprocessing – Data Preparation (IV)

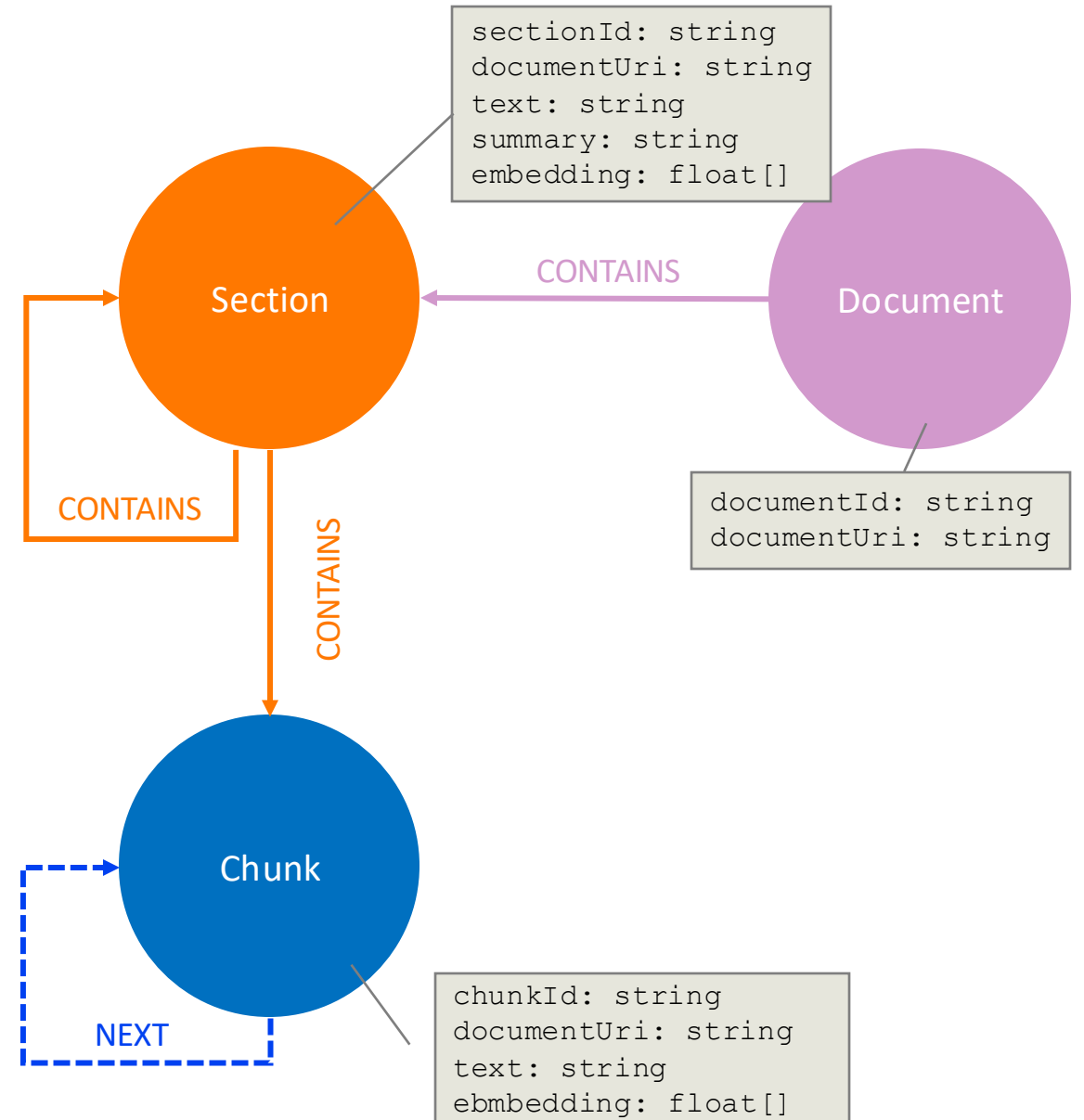
Parent Retriever Strategy by enhancing sections with summary

Use LLM to generate summaries for all chunks of a given section

Add the summary and its embeddings to the Section node

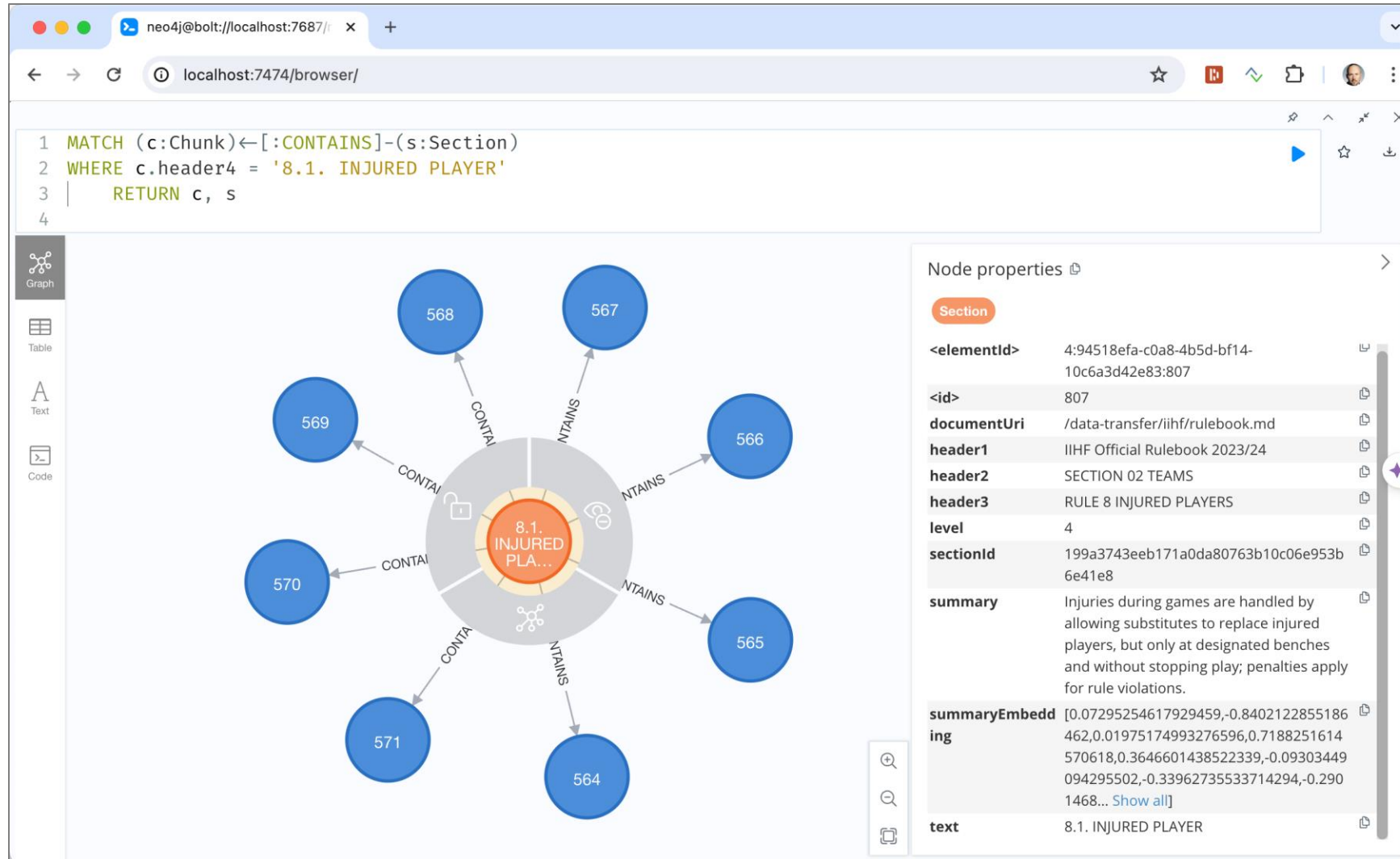
Benefits:

- Summary has its own embeddings and can be another path to find the most relevant section



Preprocessing – Data Preparation (IV)

Section nodes enhanced with with summary and its embeddings



Preprocessing – Data Preparation (V)

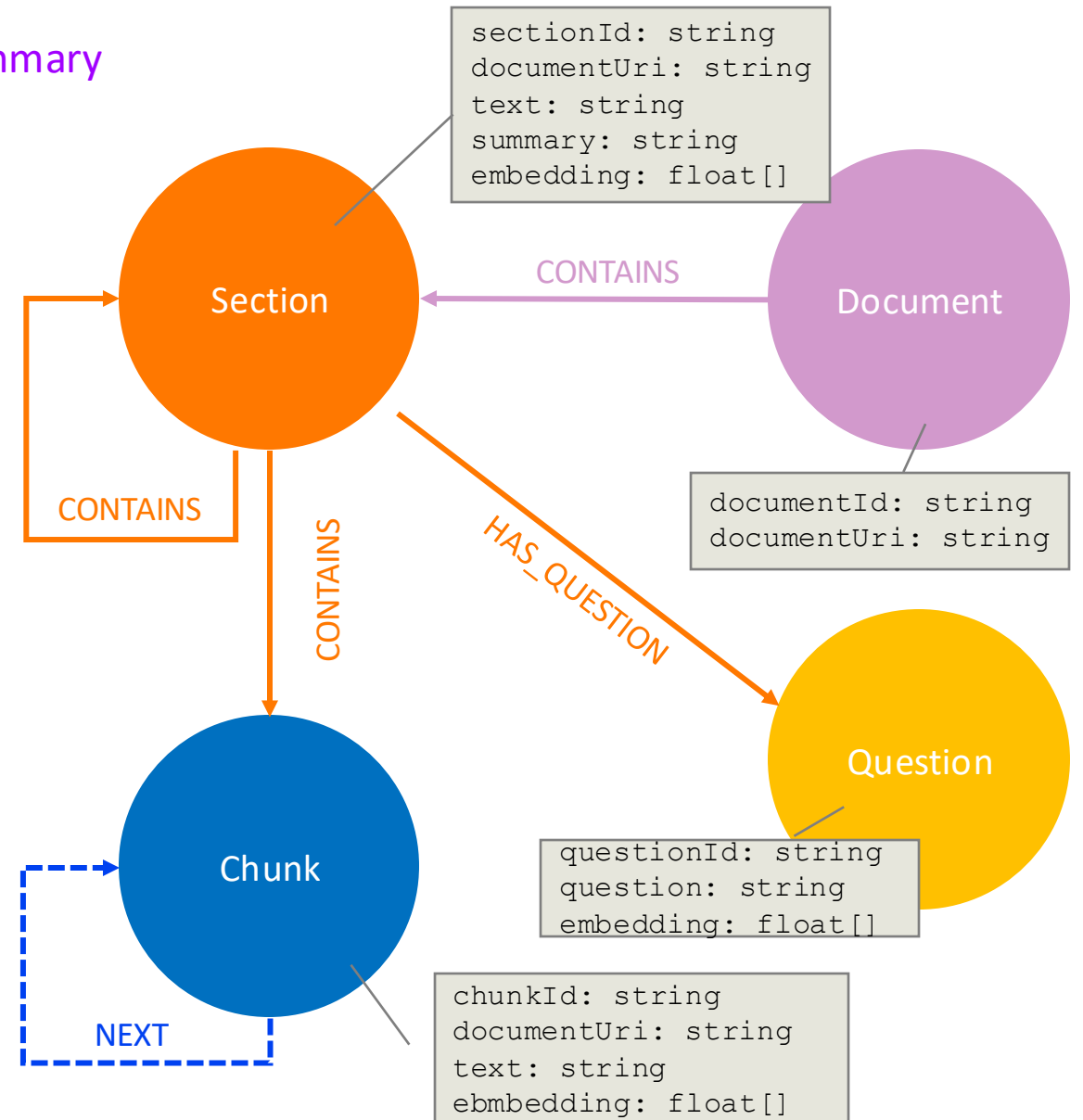
Hypothetical Questions Strategy by enhancing sections with summary

Use the LLM to generate hypothetical questions somebody might ask and add it to the graph

Calculate and add an embedding to each question

Benefits:

- Each question has its own embeddings and can be another path to find the most relevant section



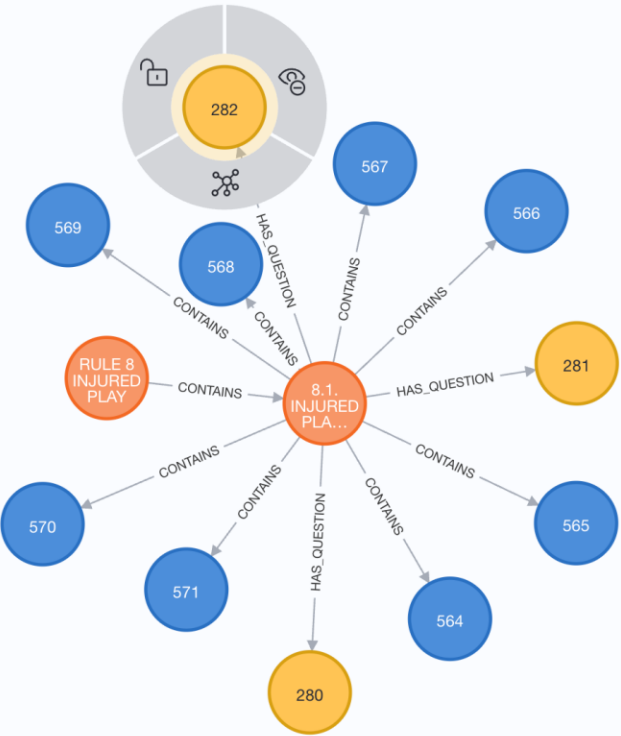
Preprocessing – Data Preparation (IV)

Graph enhanced with Question nodes for each generated hypothetical question

neo4j@bolt://localhost:7687/ x +

localhost:7474/browser/

```
neo4j$ MATCH (c:Chunk)←[:CONTAINS]-(s:Section) WHERE c.header4 = '8.1. INJURED PLAYER' RETURN c, s
```



Node properties

Question

<elementId> 4:94518efa-c0a8-4b5d-bf14-10c6a3d42e83:282

<id> 282

embedding [0.30488285422325134,-0.6791223287582397,-0.936174750328064,0.8059087991714478,-0.08675538748502731,-0.3923969566822052,0.6886387467384338,0.3049145340... [Show all](#)]

question If a player feigns an injury and lies on the ice, can they still be replaced by a substitute, or will they receive a minor penalty instead?

questionId a08557c8-9408-4966-b1bb-3e8aee6eef1d

Summary & Outlook & Ideas

Thanks to the Neo4J people for all the material that served as inspiration

Summary

- a very fast moving field with a vibrant community (blogs, youtube, discord)
- Parsing and extracting the structure of a PDF document is hard
- With knowledge graph we can enhance context by relevant information and increase quality of the responses from the LLM
- Everything works locally -> **no data left** my machine

Outlook & Ideas

- Further invest into PDF extraction techniques and libraries
- More testing and evaluation
- Build a UI with streamlit
- Further enhance knowledge graph with entity extraction (Entities as nodes)
- Use a separate Embedding node for all embeddings on all nodes



References

- Chunk Visualizer on Hugging Face: https://huggingface.co/spaces/m-ric/chunk_visualizer
- Chunk Viz: <https://chunkviz.up.railway.app/>
- Neo4J: <https://neo4j.com/>
- Ollama: <https://ollama.com/>
- Langchain: <https://www.langchain.com/>
- Platys: <https://github.com/trivadispf/platys> and <https://github.com/TrivadisPF/platys-modern-data-platform>
- Github: <https://github.com/gschmutz/generative-ai-demo>

Thank you!