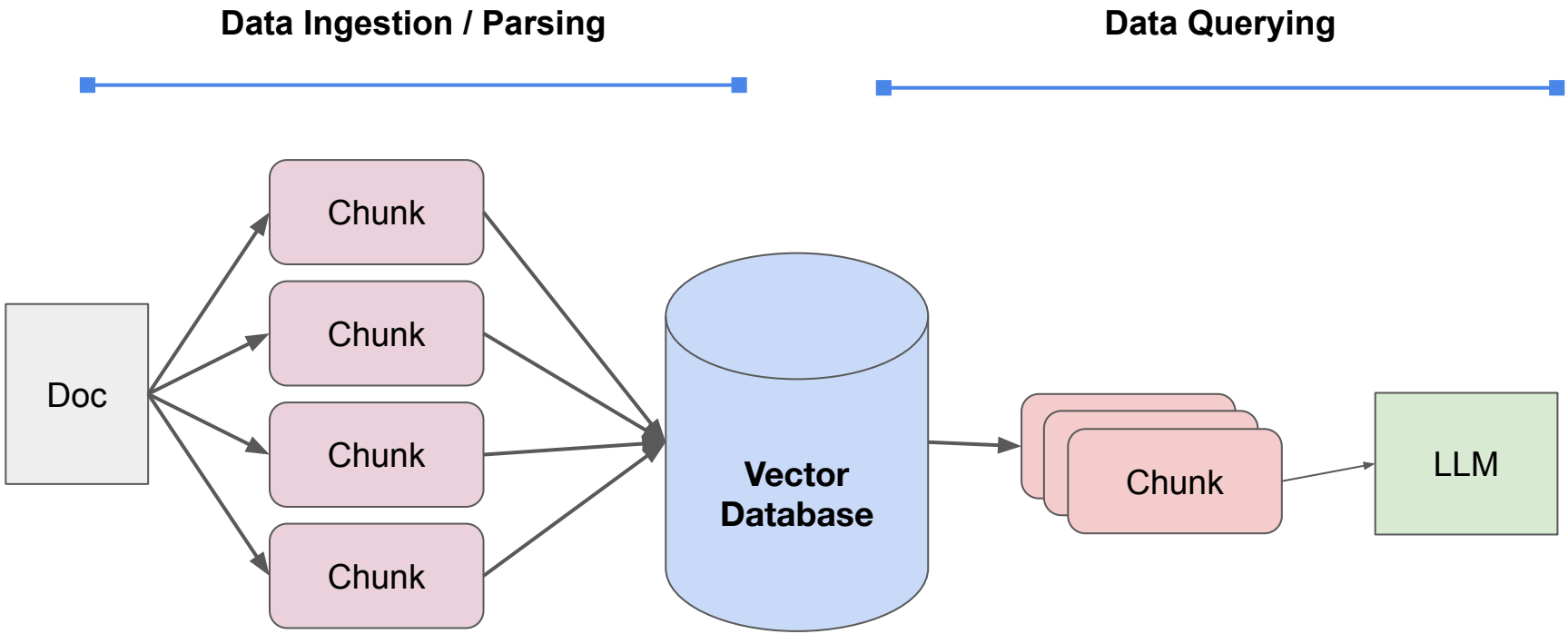
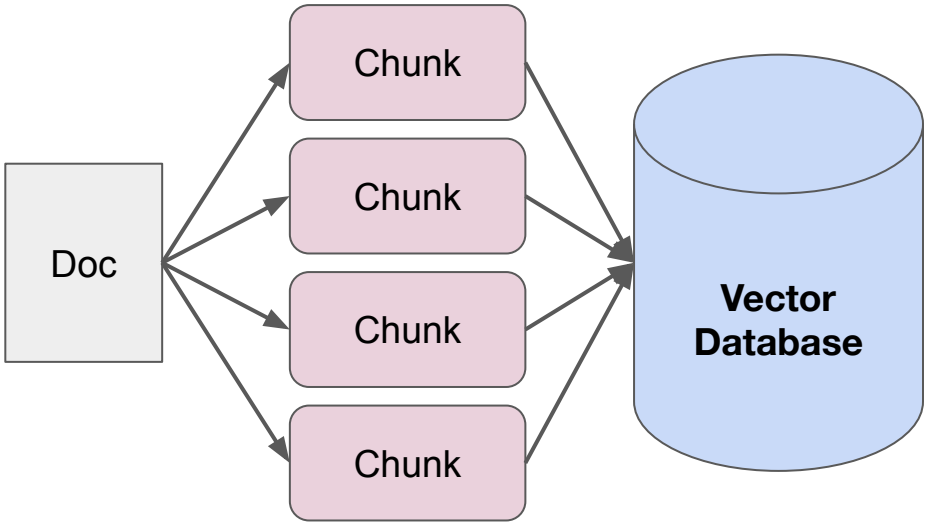
Practical Data Considerations for building Production-Ready LLM Applications

Jerry Liu, LlamaIndex co-founder/CEO

Naive RAG Stack for building a QA System



Current RAG Stack (Data Ingestion/Parsing)



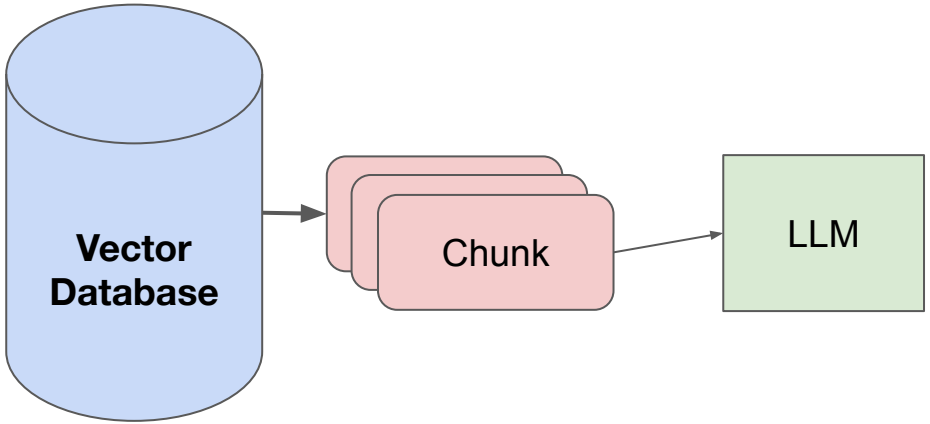
Naive State:

- Split up document(s) into even chunks.
- Each chunk does not contain parent context.
- All chunks are stored in the same collection in a vector database.

Current RAG Stack (Querying)

Naive State:

- Find top-k most similar chunks from vector database collection
- Plug into LLM response synthesis module



Challenges with Naive RAG (Response Quality)

- When RAG fails, the most common reason is bad retrieval
  - If the retrieved results are bad, there's no way the LLM can synthesize a proper response without hallucinating!
- The most common retrieval method is top-k embedding lookup

Challenges with Naive RAG (Response Quality)

- Causes of bad retrieval quality
  - Each chunk does not have awareness of parent context or related context
  - The query assumes a certain traversal structure that top-k embedding lookup doesn’t utilize.
  - The data is redundant or out of date

Challenges with Naive RAG (System Concerns)

There are also system-level considerations with this stack

- How do you deal with updates in the source document?
  - How do you update stored chunks in the vector database?

Key Lessons

To improve your RAG stack,

Improve the way you define **state**, not just the retrieval algorithm!

Pick a Good Parser

There's general text splitting strategies to split unstructured documents.

- Character splitting
- Token splitting
- Sentence splitting

Also great out of the box file parsers (LlamaHub, Unstructured)

But you may inevitably have to write your own parser for your domain-specific use case.

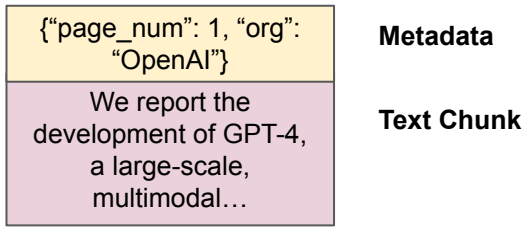
- SEC document
- Supreme Court filing
- Instacart web page

Augmenting Chunks with Context

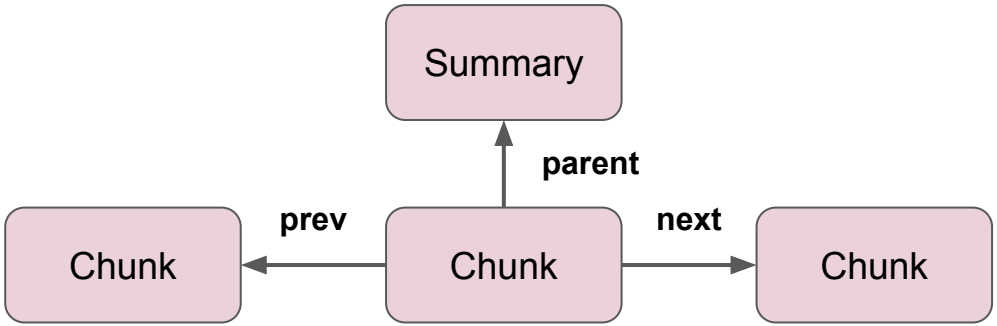
- One of the reasons embedding retrieval fails is that relevant context chunks do not match the query embedding

Different Context Augmentation Strategies

Injecting Metadata



Defining Node Relationships



Simple use case: adding page numbers to PDF's allows for in-line citations

Stream response with page citation

```
response = query_engine.query("What was the impact of COVID? Show statements in bullet form and show page
response.print_response_stream()
```

```
• Decreased demand for our platform leading to decreased revenues and decreased earning opportunities for
drivers on our platform (Page 6)
• Establishing new health and safety requirements for ridesharing and updating workplace policies (Page
6)
• Cost-cutting measures, including lay-offs, furloughs and salary reductions (Page 18)
• Delays or prevention of testing, developing or deploying autonomous vehicle-related technology (Page 1
8)
• Reduced consumer demand for autonomous vehicle travel resulting from an overall reduced demand for trav
el (Page 18)
• Impacts to the supply chains of our current or prospective partners and suppliers (Page 18)
• Economic impacts limiting our or our current or prospective partners' or suppliers' ability to expend r
esources on developing and deploying autonomous vehicle-related technology (Page 18)
• Decreased morale, culture and ability to attract and retain employees (Page 18)
• Reduced demand for services on our platform or greater operating expenses (Page 18)
• Decreased revenues and earnings (Page 18)
```

Simple use case: adding page numbers to PDF's allows for in-line citations

Inspect source nodes

```
for node in response.source_nodes:
    print('-----')
    text_fmt = node.node.text.strip().replace('\n', ' ')[:1000]
    print(f"Text:\t {text_fmt} ...")
    print(f'Metadata:\t {node.node.extra_info}')
    print(f'Score:\t {node.score:.3f}')
```

```
-----
Text:    Impact of COVID-19 to our BusinessThe ongoing COVID-19 pandemic continues to impact commu
nities in the United States, Canada and globally. Since the pandemic began in March 2020,go
vernments and private businesses - at the recommendation of public health officials - have
enacted precautions to mitigate the spread of the virus, including travelrestrictions and soc
ial distancing measures in many regions of the United States and Canada, and many enterpris
es have instituted and maintained work from homeprograms and limited the number of employees on s
ite. Beginning in the middle of March 2020, the pandemic and these related responses caused decreased dem
and for ourplatform leading to decreased revenues as well as decreased earning opportunities for drivers
on our platform. Our business continues to be impacted by the COVID-19pandemic. Although we have seen so
me signs of demand improving, particularly compared to the dema ...
Metadata:        {'page_label': '6'}
Score:   0.823
-----
Text:    storing unrented and returned vehicles. These impacts to the demand for and operations of the di
fferent rental programs have and may continue to adversely affectour business, financial condi tion and r
esults of operation.• The COVID-19 pandemic may delay or prevent us, or our current or prospective partne
rs and suppliers, from being able to test, develop or deploy autonomousvehicle-related technology, incl
uding through direct impacts of the COVID-19 virus on employee and contractor health; reduce
d consumer demand forautonomous vehicle travel resulting from an overall reduced demand for travel; s
helter-in-place orders by local, state or federal governments negatively impactingoperations, including
our ability to test autonomous vehicle-related technology; impacts to the supply chains of our current or
prospective partners and suppliers;or economic impacts limiting our or our current or prospectiv
e partners' or suppliers' ability to expend resources o ...
Metadata:        {'page_label': '18'}
Score:   0.811
-----
```

Using LLMs for Automatic Metadata Extraction

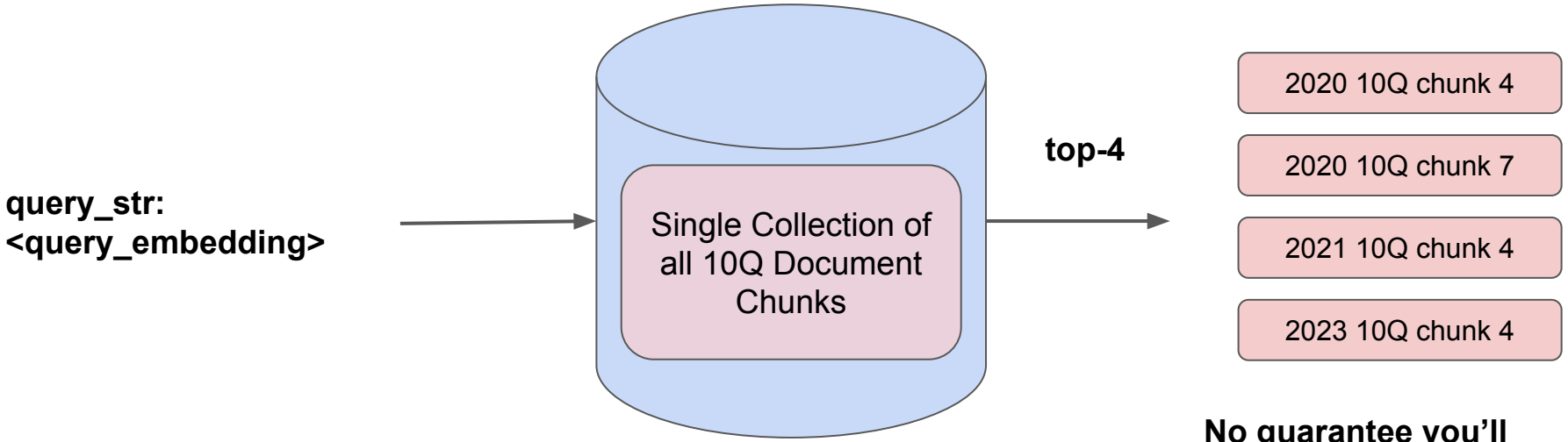
```
print(
    "LLM sees:\n",
    (uber_nodes + lyft_nodes)[9].get_content(metadata_mode=MetadataMode.LLM),
)
```

```
LLM sees:
 [Excerpt from document]
page_label: 65
file_name: 10k-132.pdf
document_title: Uber Technologies, Inc. 2019 Annual Report: Revolutionizing Mobility and L
ogistics Across 69 Countries and 111 Million MAPCs with $65 Billion in Gross Bookings
questions_this_excerpt_can_answer:

1. What is Uber Technologies, Inc.'s definition of Adjusted EBITDA?
2. How much did Adjusted EBITDA change from 2017 to 2018?
3. How much did Adjusted EBITDA change from 2018 to 2019?
Excerpt:
-----
See the section titled “Reconciliations of Non-GAAP Financial Measures” for our definition
and a
reconciliation of net income (loss) attributable to  Uber Technologies, Inc. to Adjusted E
BITDA.

  Year Ended December 31,   2017 to 2018   2018 to 2019
(In millions, exce pt percenta ges)  2017   2018   2019   % Chan ge  % Chan ge
Adjusted EBITDA ..............................  $ (2,642) $ (1,847) $ (2,725)  30%  (4
8)%
-----
```

Defining the right indexes over your data

Question: “Can you tell me about Google’s R&D initiatives from 2020 to 2023?”

Dumping chunks to a single collection doesn’t work.



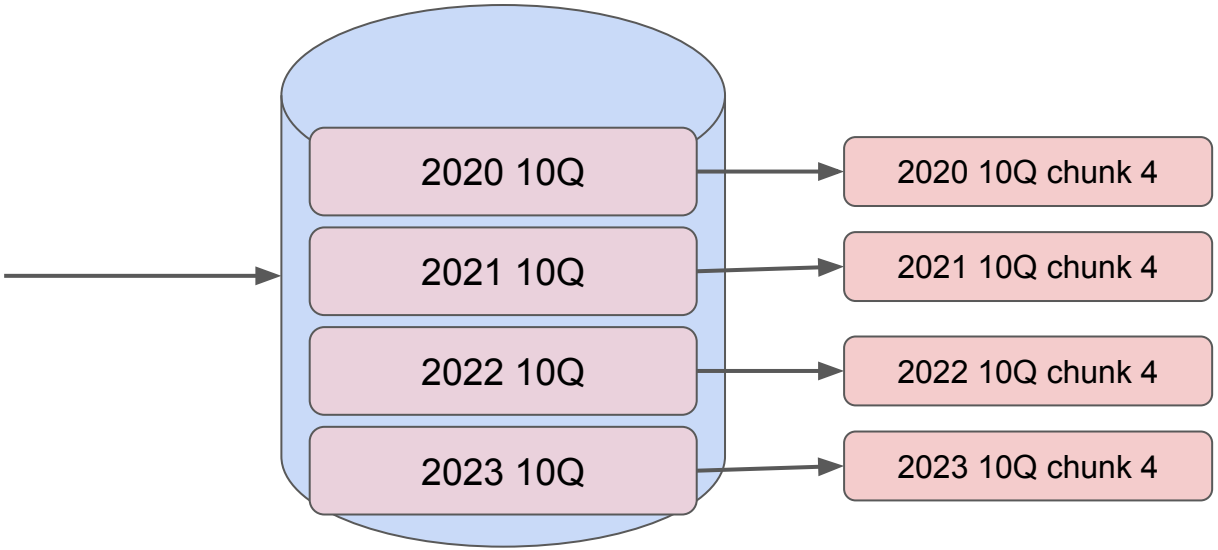
Defining the right indexes over your data

Question: “Can you tell me about Google’s R&D initiatives from 2020 to 2023?”

Here, we separate and tag the documents.

query_str:
<query_embedding>

Metadata tags:
<metadata_tags>



Handling Source Document Updates

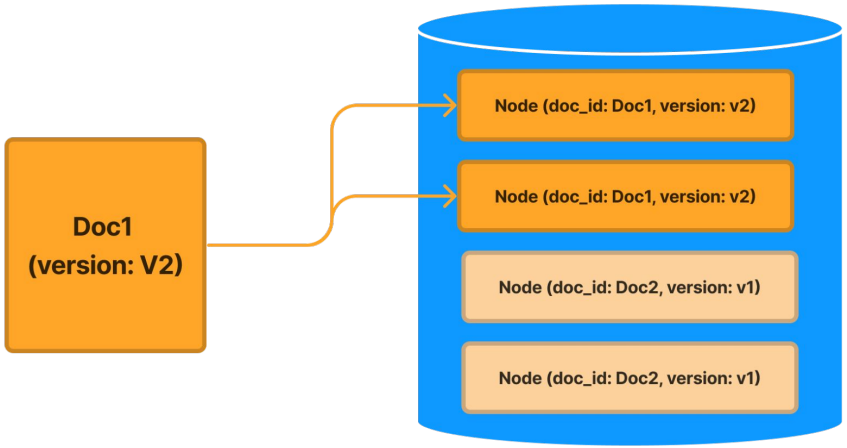
If source document changes, you need to update nodes in the vector database.

Different frequencies:

- Periodic update
- Real-time update

Ideally you don’t need to re-update the entire collection!

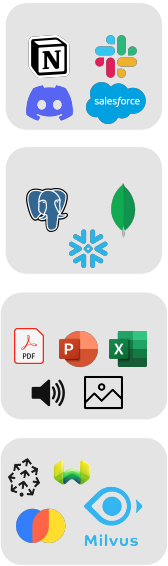
Handling Source Document Updates



- Maintain a hash per document. Pass source doc_id to Nodes.
- The doc_id → hash mapping can be stored in any document store.
- Sync data changes ♻️: if source data updates, then the hash changes. Update vector database

LlamaIndex: A data framework for LLM applications

- Data Management and Query Engine for your LLM application
- Offers components across the data lifecycle: ingest, index, and query over data



Data Ingestion
(LlamaHub 🦙)

- Connect your existing data sources and data formats (API's, PDF's, docs, SQL, etc.)

Data Structures

- Store and index your data for different use cases. Integrate with different db's.

Retrieval and Query Interface

- Given an input prompt, retrieve relevant context and synthesize a knowledge-augmented output.

Data Solutions in LlamaIndex

Define/customize metadata: https://gpt-index.readthedocs.io/en/latest/how_to/customization/custom_documents.html

Automatic metadata extraction: https://gpt-index.readthedocs.io/en/latest/how_to/index/metadata_extraction.html

Document Comparisons:
https://gpt-index.readthedocs.io/en/latest/examples/query_engine/sub_question_query_engine.html

Handling Document Updates:
https://gpt-index.readthedocs.io/en/latest/how_to/index/usage_pattern.html#handling-document-update