zilliz

A Beginners Guide to Building a RAG App Using Milvus

Stephen Batifol | Zilliz

Speaker

Stephen Batifol

Developer Advocate, Zilliz

stephen.batifol@zilliz.com
https://www.linkedin.com/in/stephen-batifol/
https://twitter.com/stephenbtl

zilliz

RAG

(Retrieval Augmented Generation)

zilliz

Basic Idea

Use RAG to **force** the **LLM to work with your data**

by injecting it via a vector database like **Milvus**

zilliz

Vector DB for RAG

Vector Databases provide the ability to **inject your data** via

semantic similarity

Considerations include: scale, performance, and flexibility

zilliz

LLMs are Stochastic

LLMs predict future tokens (a-la RNNs)

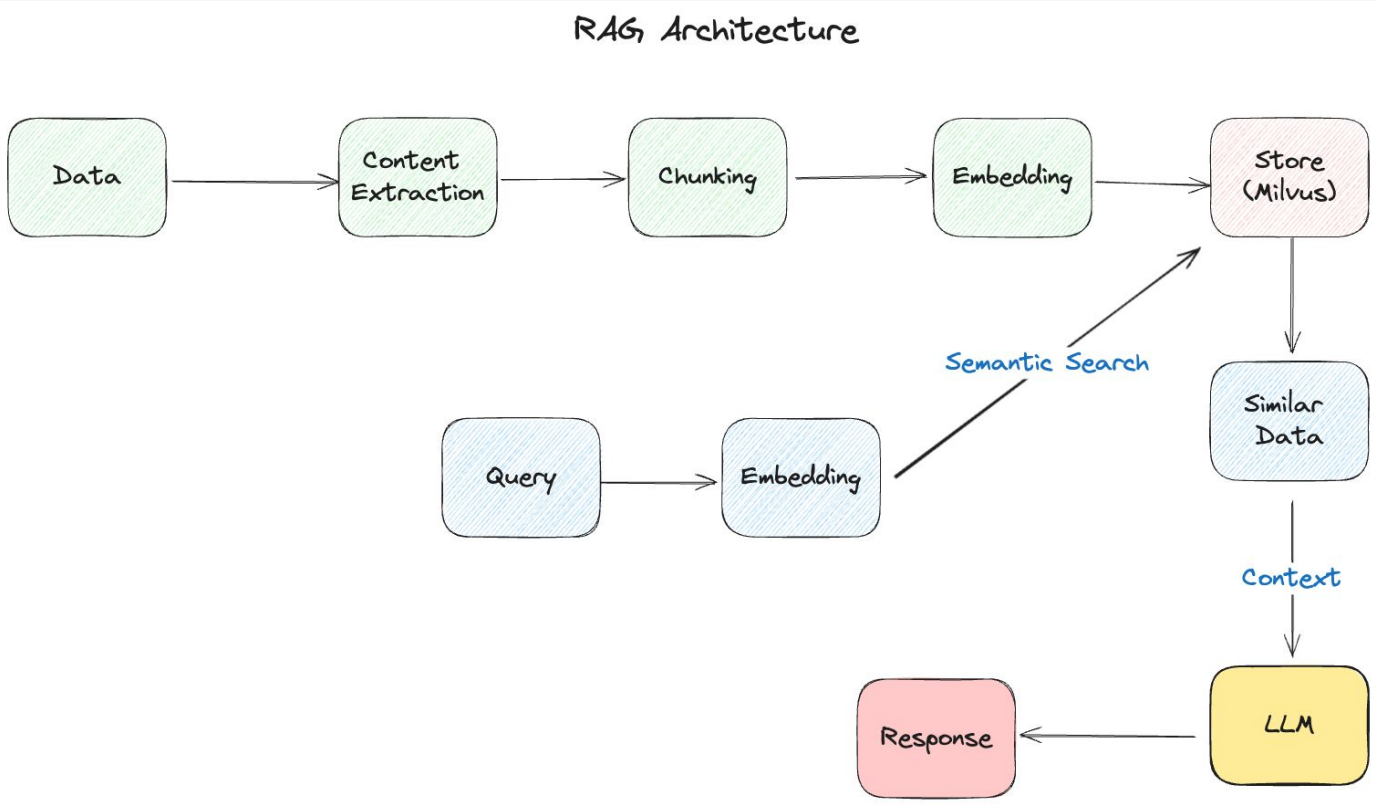
- “Milvus is the world ’s most popular vector ___”
- {“database”: 0.86, “search”: 0.11, “embedding”, 0.01, …}

Downside: outdated input data could be cause for hallucination

- Plausible-sounding but factually incorrect responses

zilliz

Basic RAG Architecture



zilliz

01

Tech Stack

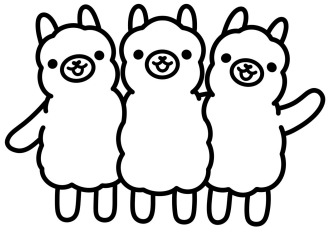
Tech Stack

LangChain

milvus

SBERT.net

Meta



ollama run llama3

zilliz

Langchain

- Framework for building LLM Applications
- Focus on retrieving data and integrating with LLMs
  - Loading the Data
  - Chunk & Chunk Overlap
- Integrations with most popular tools

LangChain

zilliz

Ollama

- Run quantized LLMs Locally
- Embeddings Models

zilliz

Milvus

1. Cloud Native, Distributed System Architecture
2. True Separation of Concerns
3. Scalable Index Creation Strategy with 512 MB Segments

zilliz

Embeddings Models

zilliz

02

Embeddings

Examining Embeddings

Picking a model

What to embed

Metadata

 |

zilliz

Embeddings Strategies

Level 1: Embedding Chunks Directly

Level 2: Embedding Sub and Super Chunks

Level 3: Incorporating Chunking and Non-Chunking Metadata

zilliz

Metadata Examples

Chunking

- Paragraph position
- Section header
- Larger paragraph
- Sentence Number
- ...

Non-Chunking

- Author
- Publisher
- Organization
- Role Based Access Control
- ...

zilliz

What your data looks like

Text:

“preferences of customers and prospective customers with respect to remote or hybrid working, as a result of the COVID-19 pandemic, leading to a parallel delay, or potentially permanent change, in receiving the corresponding revenue; •our projected financial information, anticipated growth rate, and market opportunity; •our ability to maintain the listing of our Class A Common Stock and Warrants on the NYSE; •our public securities’ potential liquidity and trading;”

Vector:

[-0.09975282847881317,-0.02853492833673954,-0.0478860922157764433,0.012315821833908558,-0.004004416521638632,0.08756010979413986,0.013248161412775517,0.010704956017434597,-0.06194952502846718,0.021150749176740646,0.02453230880200863,0.03979797288775444,-0.032914288341999054,-0.011855324730277061,...]

zilliz

Takeaway:

Your embeddings strategy depends on your **accuracy**,

cost, and **use case** needs

zilliz

03

Chunking

zilliz

Chunking Considerations

Chunk Size

Chunk Overlap

Character Splitters

zilliz

Examples

Chunk Size=50, Overlap=0

```
Text is: Table of Contents
UNITED STATES
SECURITIES

Text is: AND EXCHANGE COMMISSION
Washington, D.C.

Text is: 20549
FORM 10-Q
(Mark

Text is: One)
☒ QUARTERLY REPORT

Text is: PURSUANT TO SECTION 13 OR

Text is: 15(d) OF THE SECURITIES EXCHANGE

Text is: ACT OF 1934
For the quarterly period ended

Text is: June 30,

Text is: 2023
OR
☐ TRANSITION REPORT

Text is: PURSUANT TO SECTION 13 OR
```

Examples

Chunk Size=128, Overlap=20

```
Text is: Table of Contents
UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 10-Q
(Mark One)
☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the quarterly period ended June 30, 2023
OR
☐ TRANSITION REPORT

Text is: the quarterly period ended June 30, 2023
OR
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from  _____  to _____
Commission file number 001-39419
WEWORK INC.
(Exact name of registrant as specified in its charter)
Delaware

Text is: INC.
(Exact name of registrant as specified in its charter)
Delaware 85-1144904
(State or other jurisdiction of incorporation or
organization)(I.R.S. Employer Identification No.)
12 East 49th Street, 3rd floor
New York, NY
(Address of principal executive offices)10017
(zip code)
(646) 389-3922
Registrant's telephone number, including area

Text is: code)
(646) 389-3922
Registrant's telephone number, including area code
Securities registered pursuant to Section 12(b) of the Act:
Title of each class Trading Symbol(s) Name of each exchange on which registered
Class A common stock, $0.0001 per share WE The New York Stock Exchange
Redeemable warrants, exercisable for shares of Class A common stock at
an exercise
```


zilliz

Examples

Chunk Size=256, Overlap=50

```
Text is: Table of Contents
UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 10-Q
(Mark One)
☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the quarterly period ended June 30, 2023
OR
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from  _____  to _____
Commission file number 001-39419
WEWORK INC.
(Exact name of registrant as specified in its charter)
Delaware 85-1144904
(State or other jurisdiction of incorporation or
organization)(I.R.S. Employer Identification No.)
12 East 49th Street, 3rd floor
New York, NY
(Address of principal executive offices)10017
(zip code)
(646)

Text is: or other jurisdiction of incorporation or
organization)(I.R.S. Employer Identification No.)
12 East 49th Street, 3rd floor
New York, NY
(Address of principal executive offices)10017
(zip code)
(646) 389-3922
Registrant's telephone number, including area code
Securities registered pursuant to Section 12(b) of the Act:
Title of each class Trading Symbol(s) Name of each exchange on which registered
Class A common stock, $0.0001 per share WE The New York Stock Exchange
Redeemable warrants, exercisable for shares of Class A common stock at
an exercise price of $11.50 per shareWE WS The New York Stock Exchange
Class A Common Stock Purchase Rights – The New York Stock Exchange
Securities registered pursuant to section 12(g) of the Act:
None
Indicate by check mark whether the registrant: (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934
```

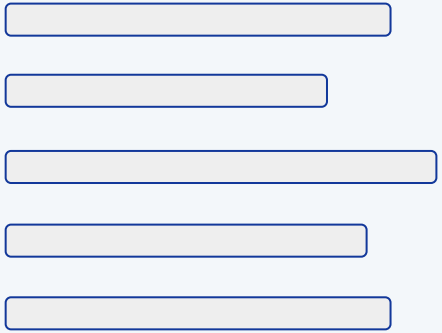
Examples

SemanticChunker

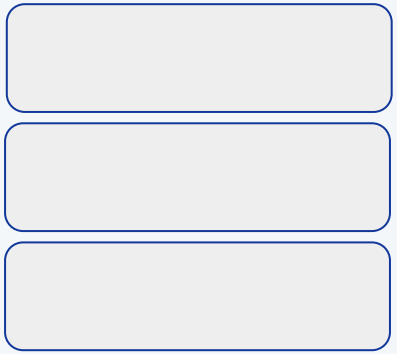
```
Text chunk: ['Table of Contents\nUNITED STATES\nSECURITIES AND EXCHANGE COMMISSION\nWashington, D.C. 20549\nFORM 10-Q\n(Mark One)\n☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIE
XCHANGE ACT OF 1934\nFor the quarterly period ended June 30, 2023\nOR\n☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934\nFor the transition period from  ___
_____  to ___________\nCommission file number 001-39419\nWEWORK INC.', "(Exact name of registrant as specified in its charter)\nDelaware 85-1144904\n(State or other jurisdiction of incorporation or\
ganization)(I.R.S. Employer Identification No.)\n12 East 49th Street, 3rd floor\nNew York, NY\n(Address of principal executive offices)10017\n(zip code)\n(646) 389-3922\nRegistrant's telephone numbe
including area code\nSecurities registered pursuant to Section 12(b) of the Act:\nTitle of each class Trading Symbol(s) Name of each exchange on which registered\nClass A common stock, $0.0001 per s
e WE The New York Stock Exchange\nRedeemable warrants,", 'exercisable for shares of Class A common stock at\nan exercise price of $11.50 per shareWE WS The New York Stock Exchange\nClass A Common St
Purchase Rights — The New York Stock Exchange\nSecurities registered pursuant to section 12(g) of the Act:\nNone\nIndicate by check mark whether the registrant: (1) has filed all reports required to
filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the registrant was required to file\nsuch reports); and (2) has be
subject to such filing requirements for the past 90 days.', 'Yes  x   No  o\nIndicate by check mark whether the registrant has submitted electronically and posted on its corporate web site, if any,
ery Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter)\nduring the preceding 12 months (or for such shorter period that the r
strant was required to submit and post such files).Yes  x   No  o \nIndicate by check mark whether the registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, or a sm
er reporting company.', 'See the definitions of “large accelerated filer,” “accelerated filer” and “smaller reporting company”\nin Rule 12b-2 of the Exchange Act. (Check one):\nLarge accelerated fil
o Accelerated filer x\nNon-accelerated filer  o Smaller reporting company o\nEmerging growth company o\n                \nIf an emerging growth company, indicate by check mark if the registrant has
cted not to use the extended transition period for complying with any new or revised financial accounting standards provided pursuant to Section 13(a) of the Exchange\nAct.', "o\nIndicate by check m
whether the registrant is a shell company (as defined in Rule 12b-2 of the Act).     Yes ☐    No  x\nIndicate the number of shares outstanding of each of the issuer's classes of common stock, as of
latest practicable date: As of August 7, 2023, there were 2,110,280,756 shares of Class A common stock, par value $0.0001 per share, and 19,938,089\nshares of Class C common stock, par value $0.0001
r share, issued and outstanding."]

Text chunk: ['Table of Contents\nWeWork Inc.\nFORM 10-Q\nTHREE AND SIX MONTHS ENDED JUNE 30, 2023\nTABLE OF CONTENTS\nPage\nPart I - Financial Information\nCautionary Note Regarding Forward-Looking
tements 3\nItem 1.', 'Financial Statements and Supplementary Data 6\nCondensed Consolidated Balance Sheets 7\nCondensed Consolidated Statements of Operations 8\nCondensed Consolidated Statements of
prehensive Loss 9\nCondensed Consolidated Statement of Changes in Convertible Preferred Stock, Noncontrolling Interests, and Equity 10\nCondensed Consolidated Statements of Cash Flows 12\nNotes to t
Condensed Consolidated Financial Statements 13\nItem 2. Management’ s Discussion and Analysis of Financial Condition and Results of Operations 70\nItem 3. Quantitative and Qualitative Disclosures Ab
Market Risk 113\nItem 4. Controls and Procedures 114\nPart II - Other Information\nItem 1.', 'Legal Proceedings 115\nItem 1A. Risk Factors 116\nItem 1B. Unresolved Staf f Comments 116\nItem 2. Unreg
ered Sales of Equity Securities and Use of Proceeds 116\nItem 3. Defaults Upon Senior Securities 117\nItem 4. Mine Safety Disclosures 117\nItem 5. Other Information 117\nItem 6. Exhibits 119\nSignat
s 119']
```


zilliz

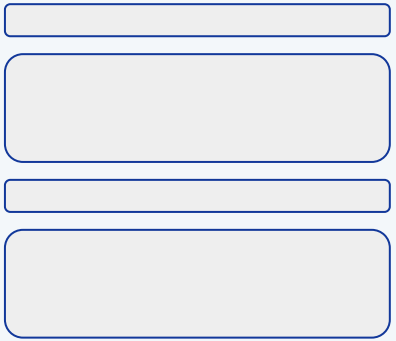
How Does Your Data Look?



Conversation Data



Documentation Data



Lecture or Q/A Data

zilliz

Takeaway:

Your chunking strategy depends on **what your data looks like** and **what you need from it.**

zilliz

Demo!

zilliz

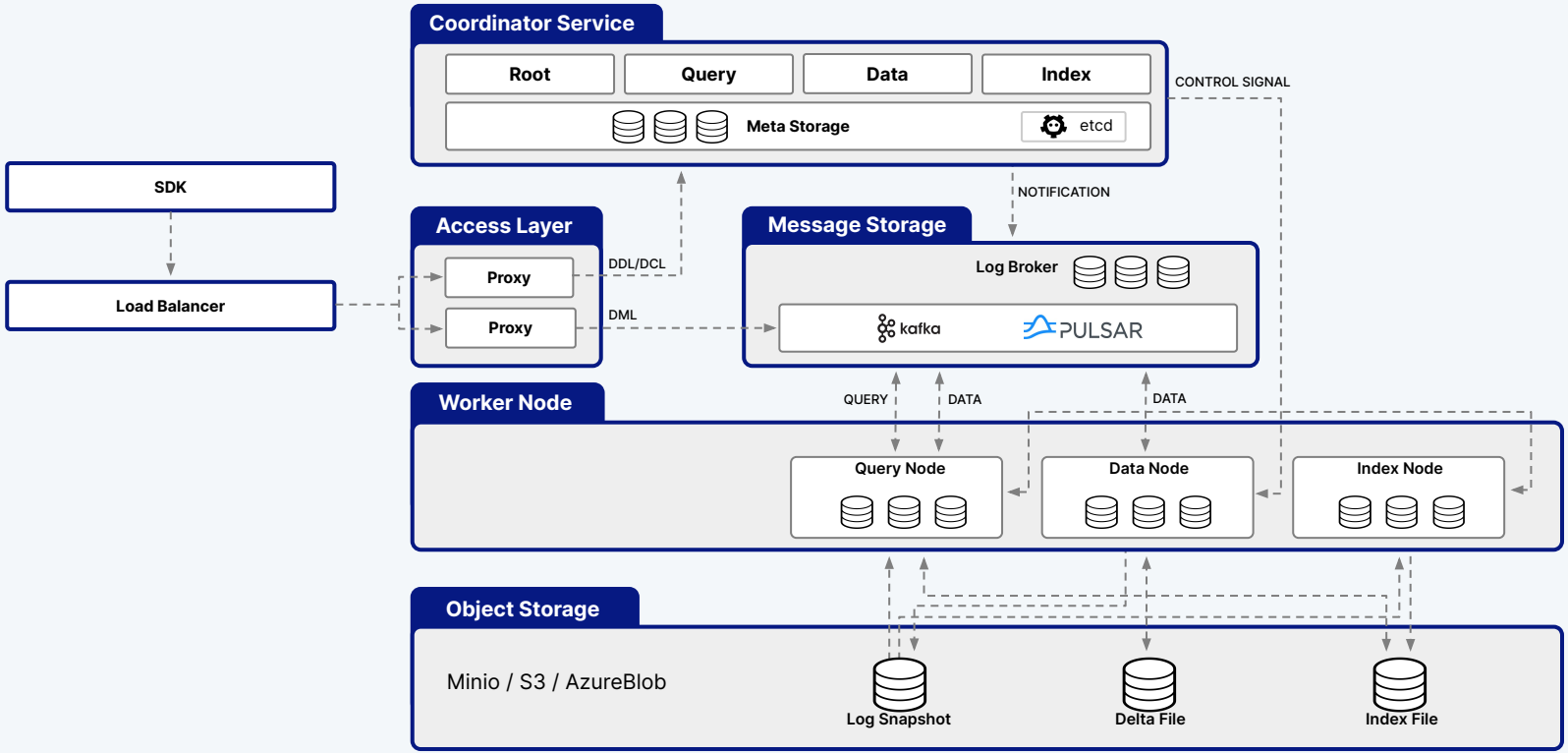
Questions?

Give Milvus a Star!

Chat with me on Discord!

zilliz

Milvus Architecture



 |