



Production -scale, Private RAG Pipelines with LanceDB



Chang She
CEO / Cofounder, LanceDB


Data & AI Summit
June 12, 2024





Chang She

CEO / Co - founder LanceDB

 @changhiskhan

- Data tools (~ 2 decades)
- Pandas
- Big Data / RecSys
- LanceDB:
The Database for Multimodal AI

Agenda

- Why RAG?
- Productionizing RAG in the Enterprise
- DBRX + LanceDB + Spark
- ~~Text-only~~ -> Multimodal

RAG

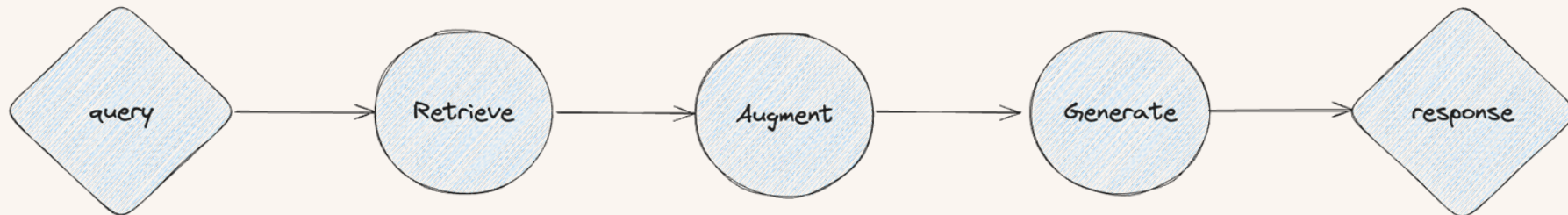
Retrieval Augmented Generation

- Extend model knowledge
- Reduce hallucination
- Works in conjunction with fine -tuning

RAG

Retrieval Augmented Generation

- Chat
- Search
- What's next for RAG?
Enterprise decision
support



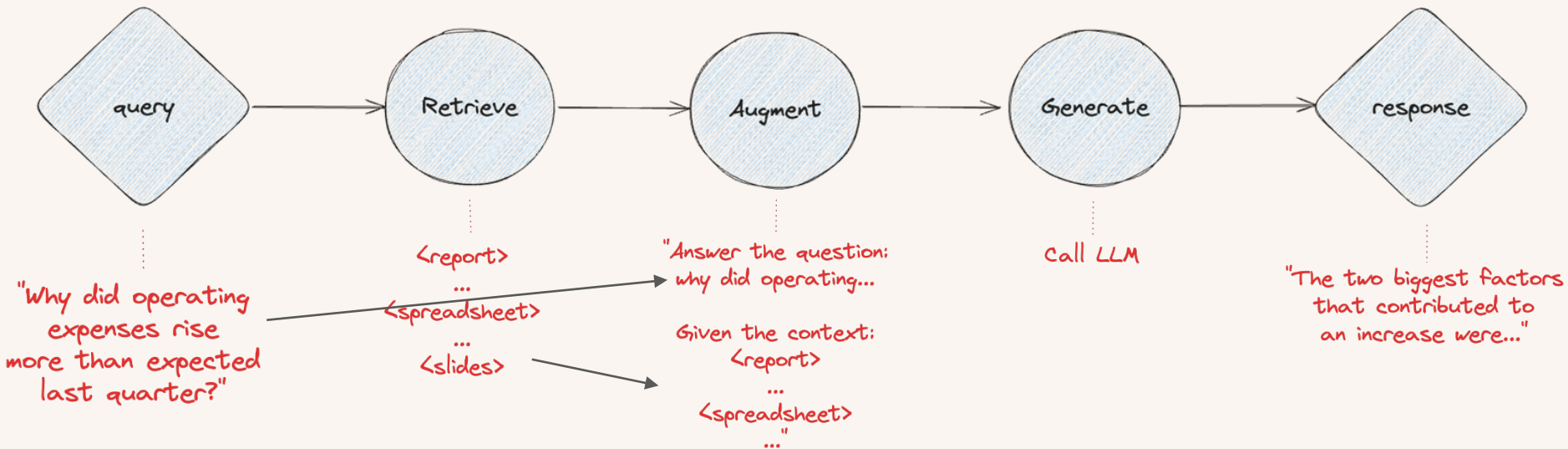
Production

Pitfalls for RAG in Production

- Scale
- Retrieval quality
- Data privacy
- Vendor lock -in

Simplified RAG

PrivateCo FP&A App



Scale

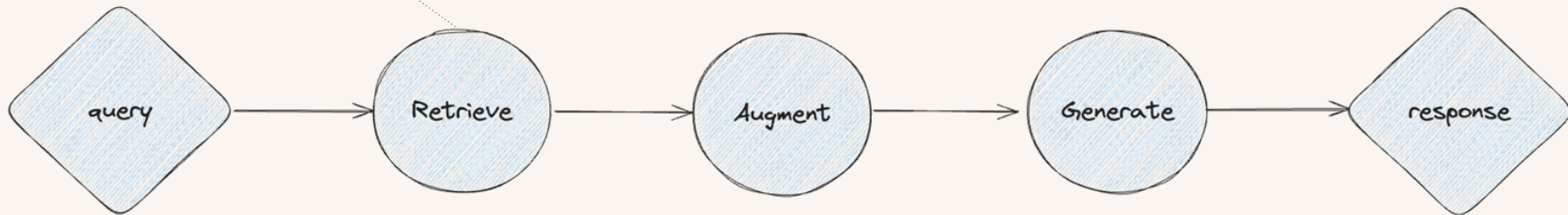
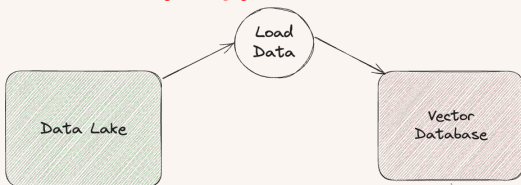
Scale sneaks up on you in production

- Num tables
- Num of vectors
- Queries per second (QPS)
- Update frequency

Unnecessary copy / slow / \$\$\$

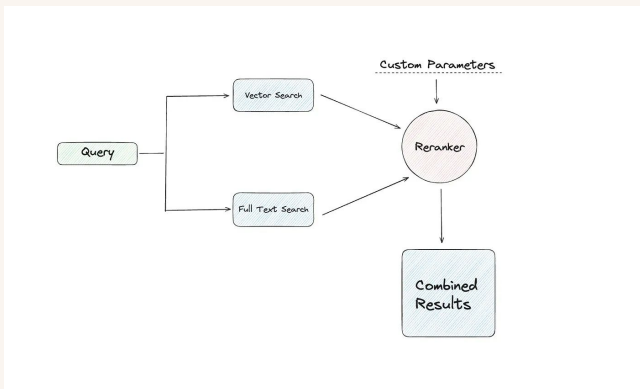
Simplified RAG

PrivateCo FP&A App



Quality

More than just vector search



- Different retrieval modes
- Hybrid + reranking
- Fine-tuning embeddings
- Composability + customizability

Privacy

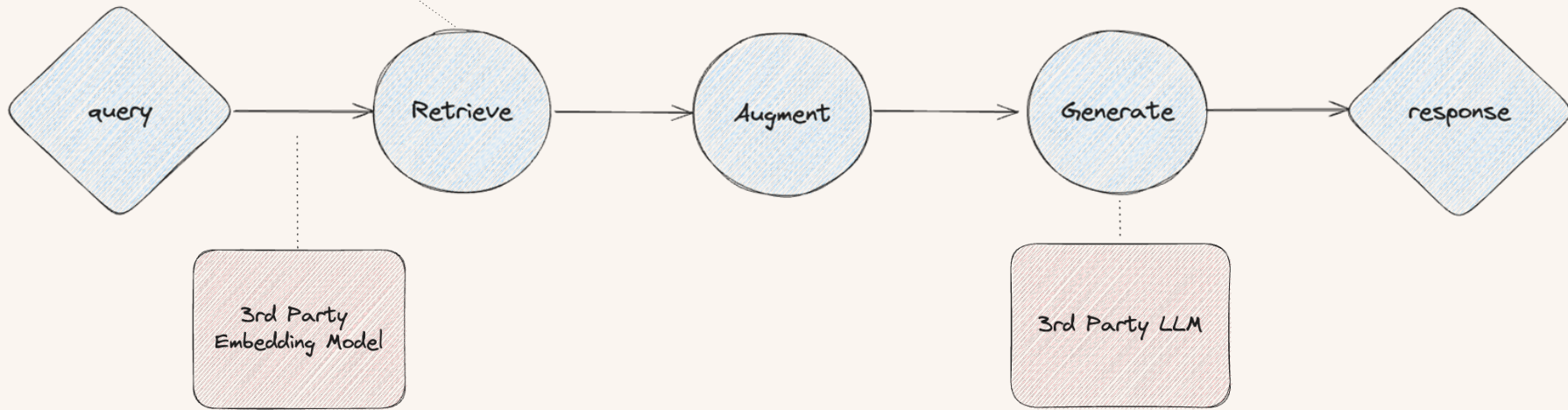
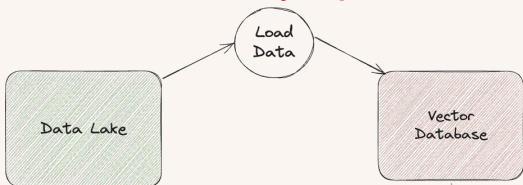
RAG context is valuable

- Models will commoditize
- Data is enterprise value
- Data should stay on -
premise

Simplified RAG

PrivateCo FP&A App

Unnecessary copies



Searches leaving premises

Financials leaving premises!

Lock-in

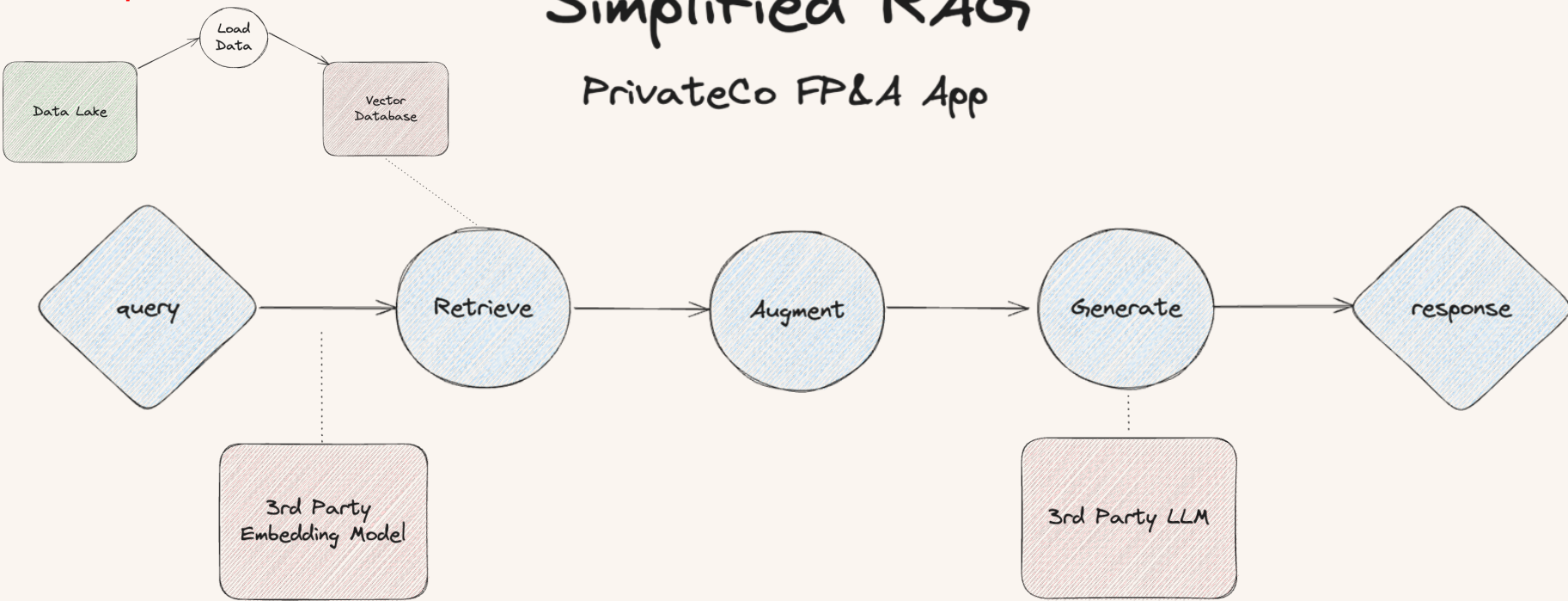
Long-term flexibility is important

- Storage should be open
- ~~Walled-garden~~ -> data lake
- RAG components should be easily swappable

Complex + Slow + \$\$\$\$

Simplified RAG

PrivateCo FP&A App

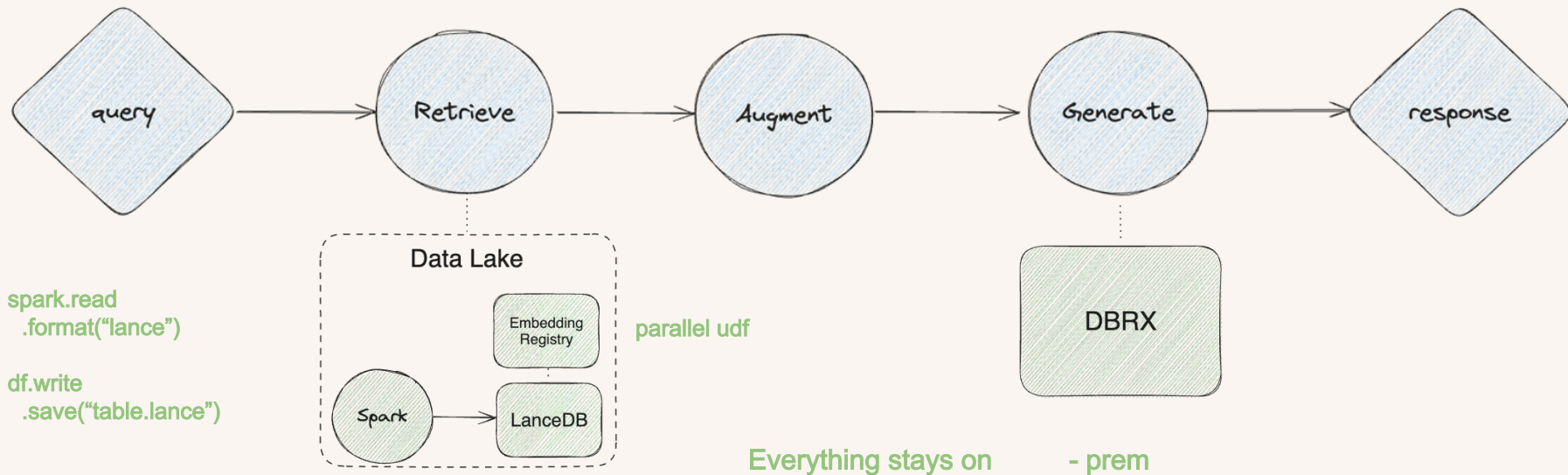


Searches leaving premises

Financials leaving premises!

Simplified RAG

PrivateCo FP&A App



LanceDB

Database for Multimodal AI



At LanceDB, we Lancelot

Effortless scale

- Billions of vectors @ 10K+ QPS
- High recall + low latency
- 10x more efficient than alternatives

Enterprise ready

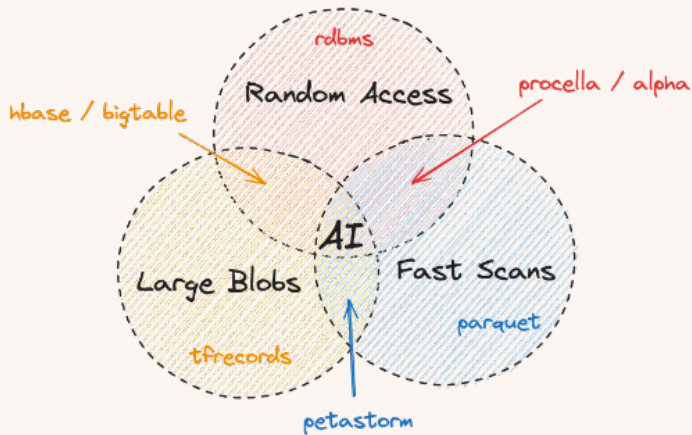
- BYOC
- SOC2
- Data Lake integration

Multimodal

- Data: image, audio, video
- Queries: vector, fts, sql, hybrid, reranking
- Workloads: EDA, Search, Training

Lance Format

Foundation for AI data



CAP Theorem for AI Data

Core to LanceDB

- Optimized for AI data
- Indexing (vector, fts, fast filters)
- Schema evolution

Benefits

- Compute-storage separation
- Unify EDA, Search, Training
- Fast training io (save GPU \$\$\$)

Multimodal

Ironically this slide contains only text

- We all have five senses!
- Image, audio, video, etc can be more expressive
- Bigger scale and more complicated RAG
- But also way more compelling

Conclusion

- RAG in production comes with new challenges
 - Scale
 - Retrieval quality
 - Data privacy
 - Vendor lock -in
- LanceDB is great for Enterprise
 - Composable RAG pipeline
 - Everything stays on -premise
 - Hyper scale at 10x efficiency

Thank you!



Discord

- Format <https://github.com/lancedb/lance>
- DB <https://github.com/lancedb/lancedb>
- Examples
<https://github.com/lancedb/vectordb-recipes>

contact@lancedb.com