SingleStore™

Dedicated Memory + Retrieval Architecture for LLMs in Enterprise Settings

Arthur Poon

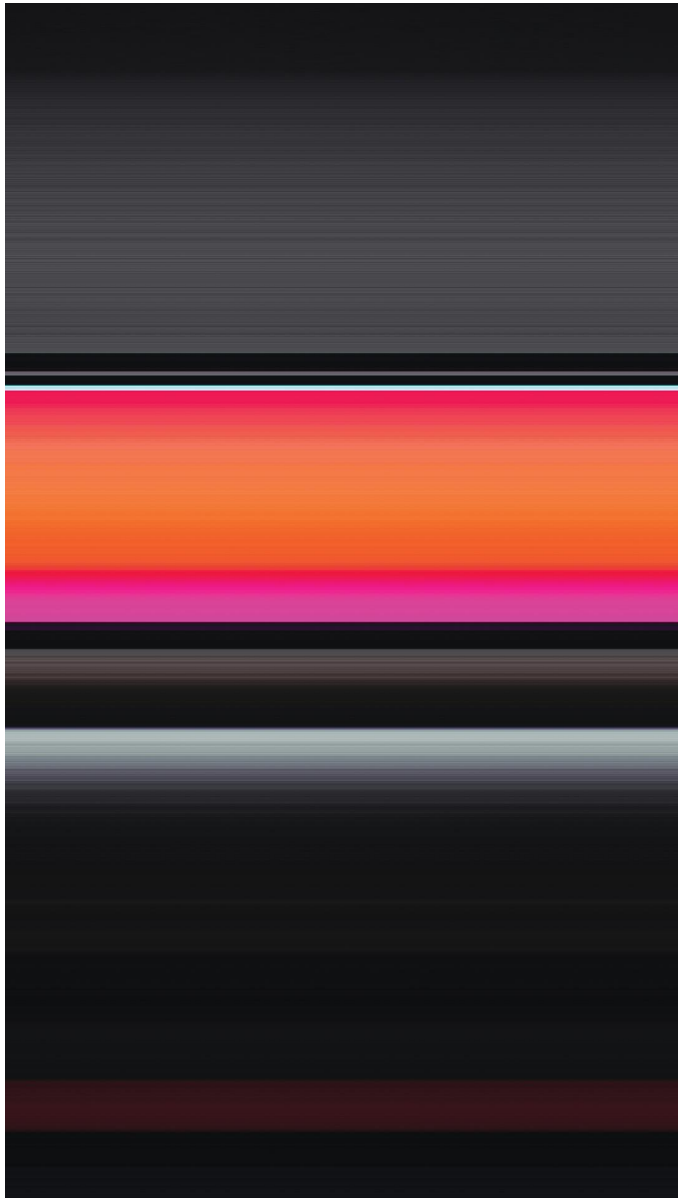
Agenda

The promise of agentic AI

The technical requirements of agentic AI

Complexities of the enterprise data stack

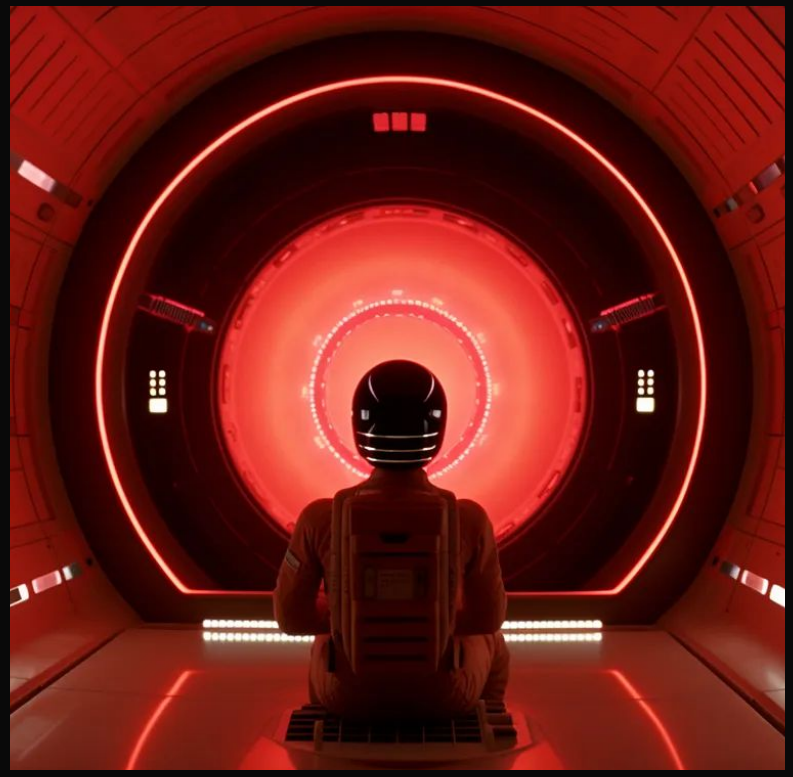
SingleStore’s innovations and future roadmap



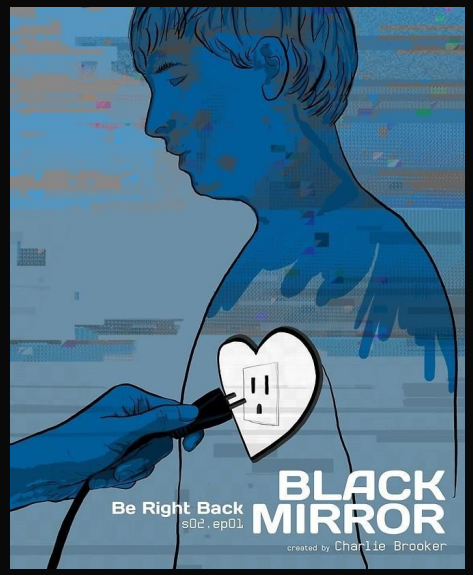
The Promise of Agentic AI


JOAQUIN PHOENIX AMY ADAMS ROONEY MARA
OLIVIA WILDE AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY


NEUROMANCER
WILLIAM GIBSON




ES GIBT NICHTS MENSCHLICHERES
ALS DEN WILLEN ZU ÜBERLEBEN
EX_MACHINA


Be Right Back
s02.ep01
BLACK
MIRROR
created by Charlie Brooker

Today: Human Agency

two common workflows in enterprises today:

Knowledge search

“How does the Digital Services Act (DSA) in Europe affect our company's content moderation policies and practices if we operate a social media platform or online marketplace in Europe?”

Root-Cause Analysis

Company experienced a 35% drop in revenue in Asia Pacific, our hypotheses are X, Y, Z → go produce some analysis in Tableau/Power BI.

Tomorrow: AI Agency

In the agentic paradigm, insights are pre-emptive not reactive, getting answers for questions you didn't even know to ask.

In the future:

~~Knowledge search~~

~~"How does the Digital Services Act (DSA) in Europe affect our company's content moderation policies and practices if we operate a social media platform or online marketplace in Europe?~~

Knowledge feed

Agent: "The Digital Services Act is currently in EU council review, this could affect your digital marketing strategies and specifically X, Y, Z campaigns."

~~Diagnostic analytics~~

~~"We experienced a 35% drop in revenue in Asia Pacific, our hypotheses are X, Y, Z" → go produce some analysis in Tableau/Power BI.~~

Preemptive analytics

Agent: "Competitor Y's acquisition of local distributor in Malaysia could decrease our selling volume by 15% across Thailand, Indonesia and Vietnam due to their ability to access a logistics network through their acquired target"

What Is Needed for Agentic AI?

Expectations of Agentic AI (Among Many)

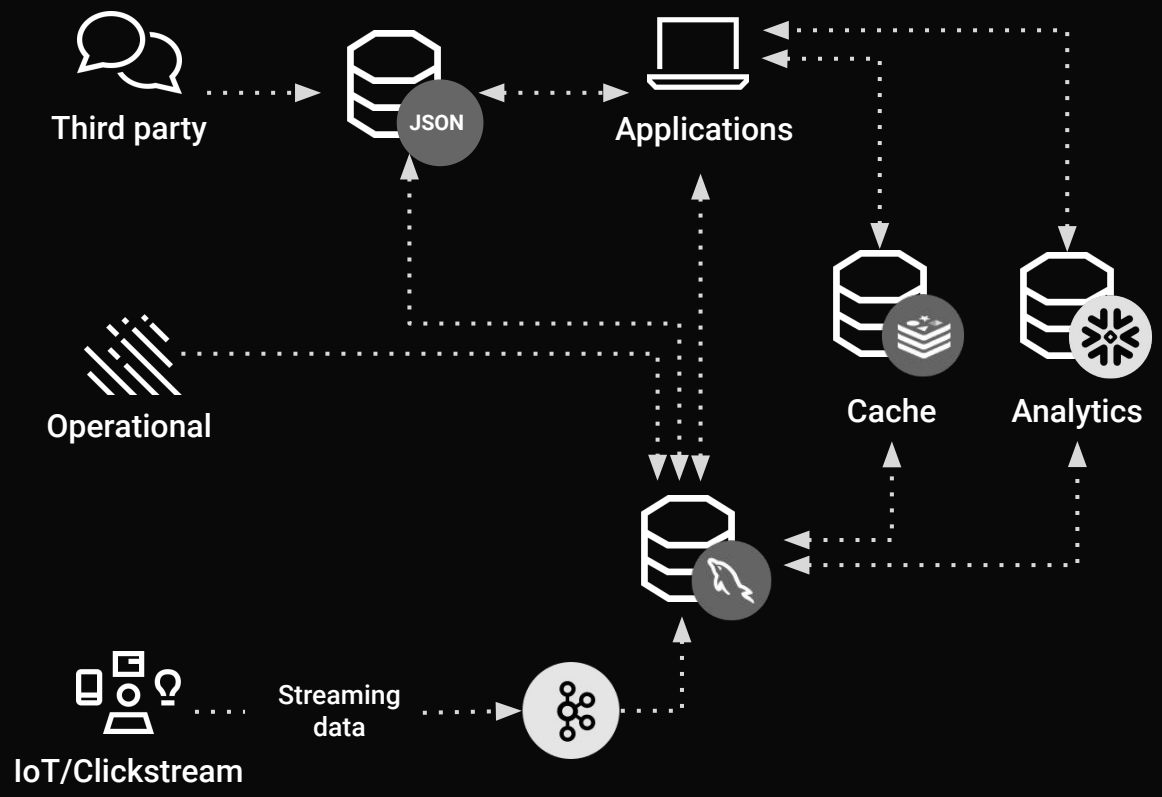
User expectations	Technical requirements
Constantly consuming all sorts of data to develop live context →	Real-time ingestion of multi-modal data
Instantly relate consumed data to everything it knows →	On-demand embeddings
Remember everything that it's learned →	Unlimited data store
Quickly and accurately identify what data is relevant to current situation →	Fast queries + search + graph engine

What Agentic AI Needs

	Technical requirements
Dedicated memory	Real-time ingestion of multi-modal data
	On-demand embeddings
	Unlimited data store
Retrieval architecture	Fast queries + search + graph engine

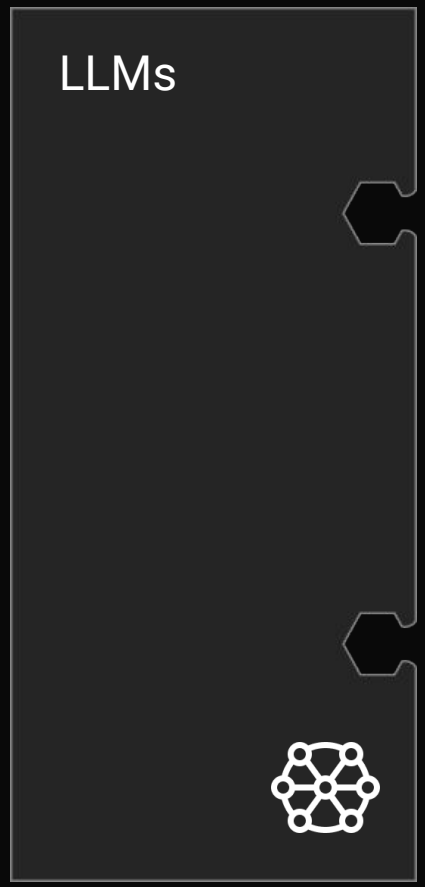
How Can This Be Brought to Enterprise?

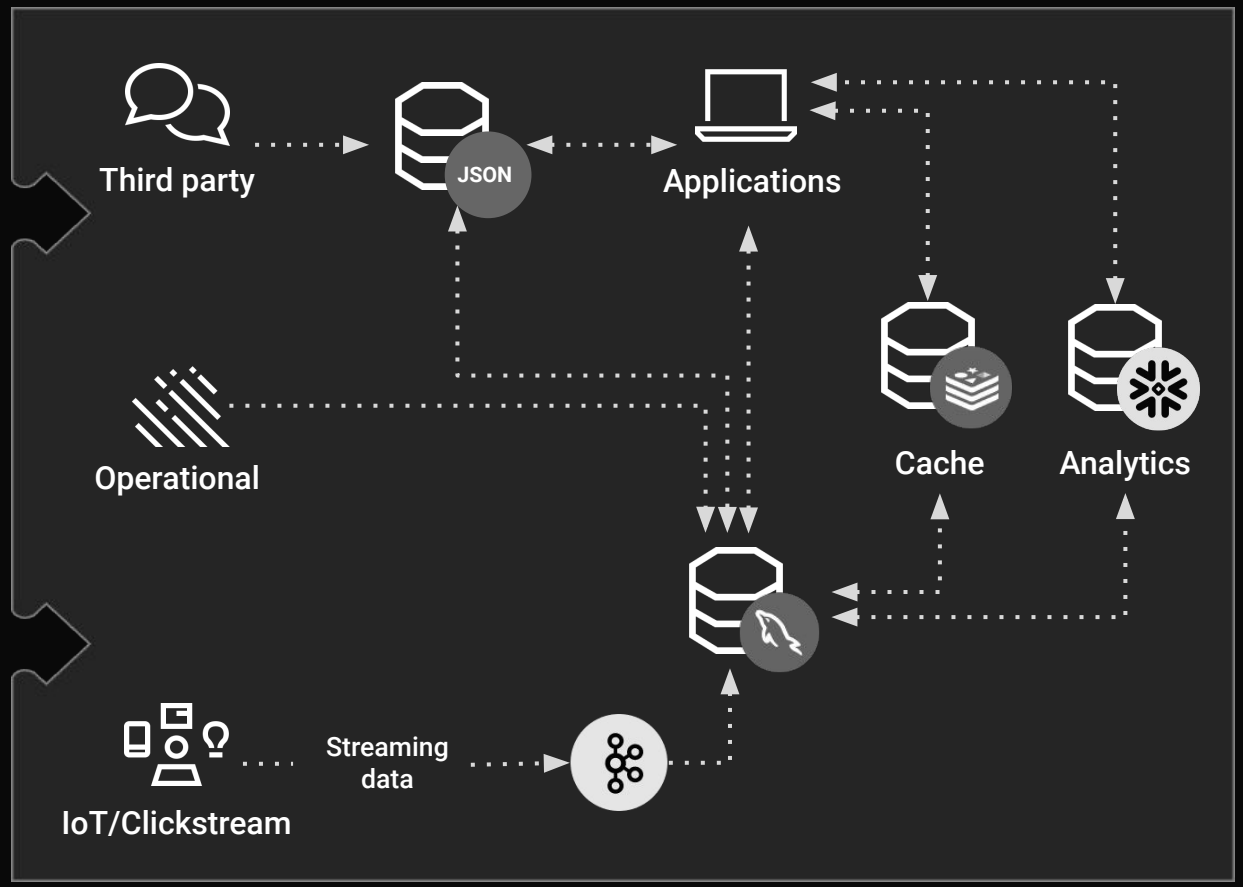
Most Enterprises Look Like This.



SingleStore

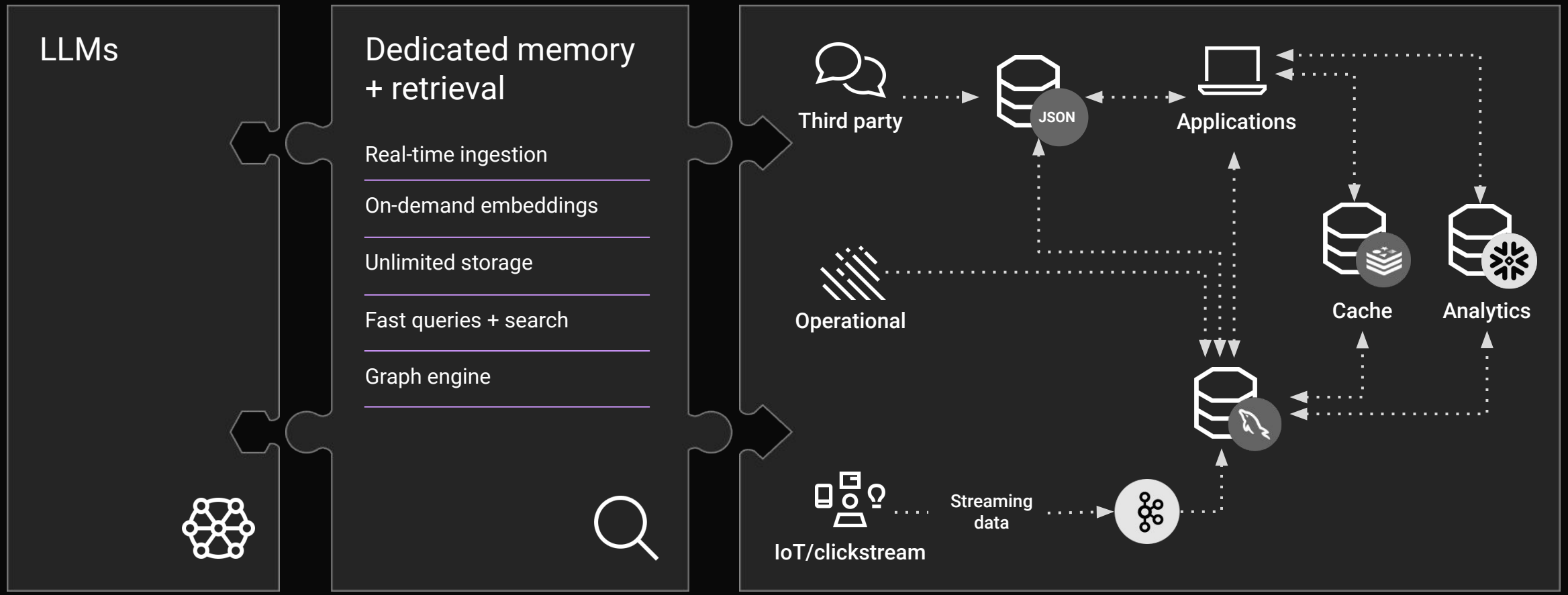
How do you add an LLM?





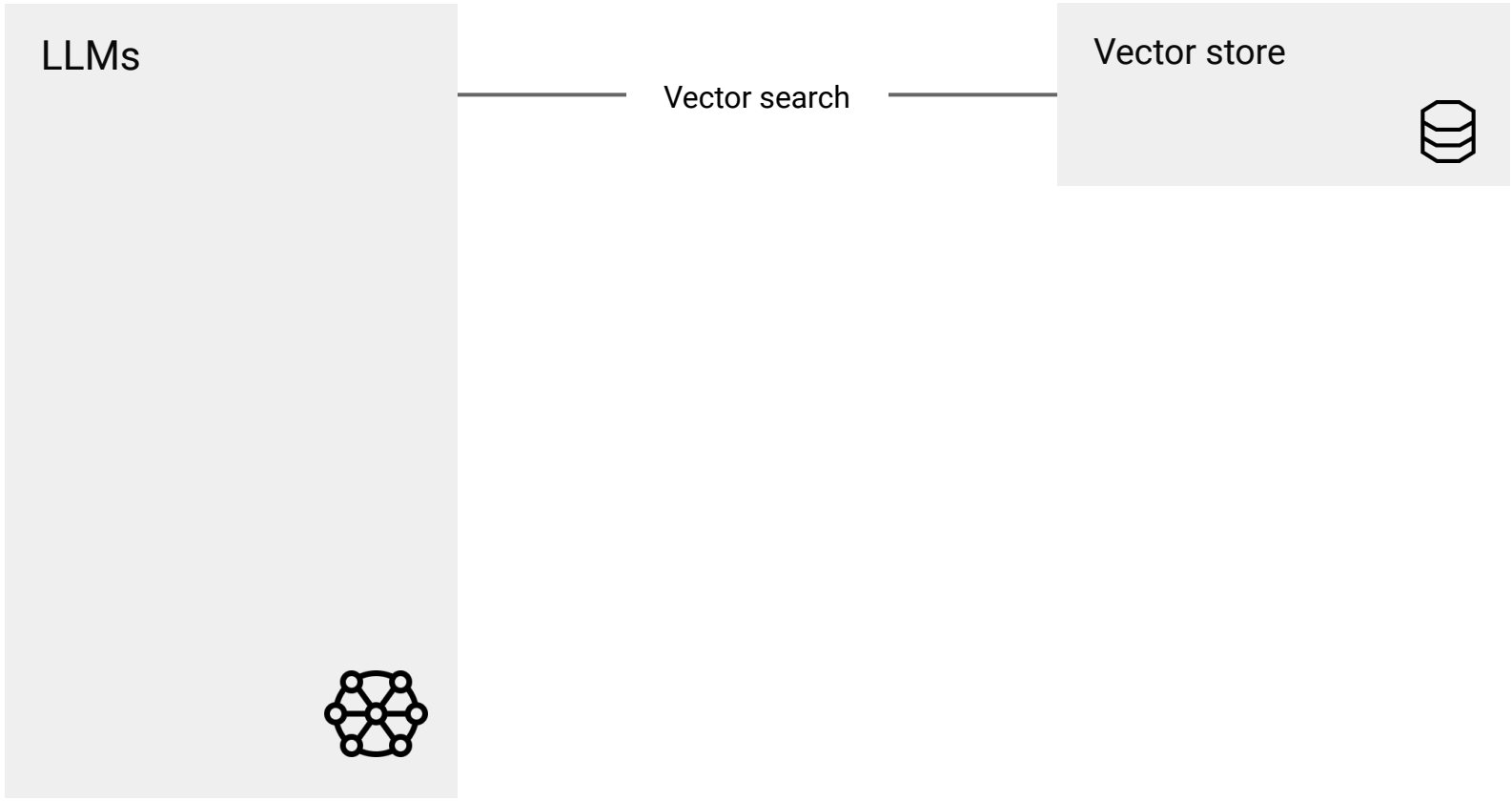
SingleStore

How do you add an LLM?
How do you create dedicated memory and retrieval?



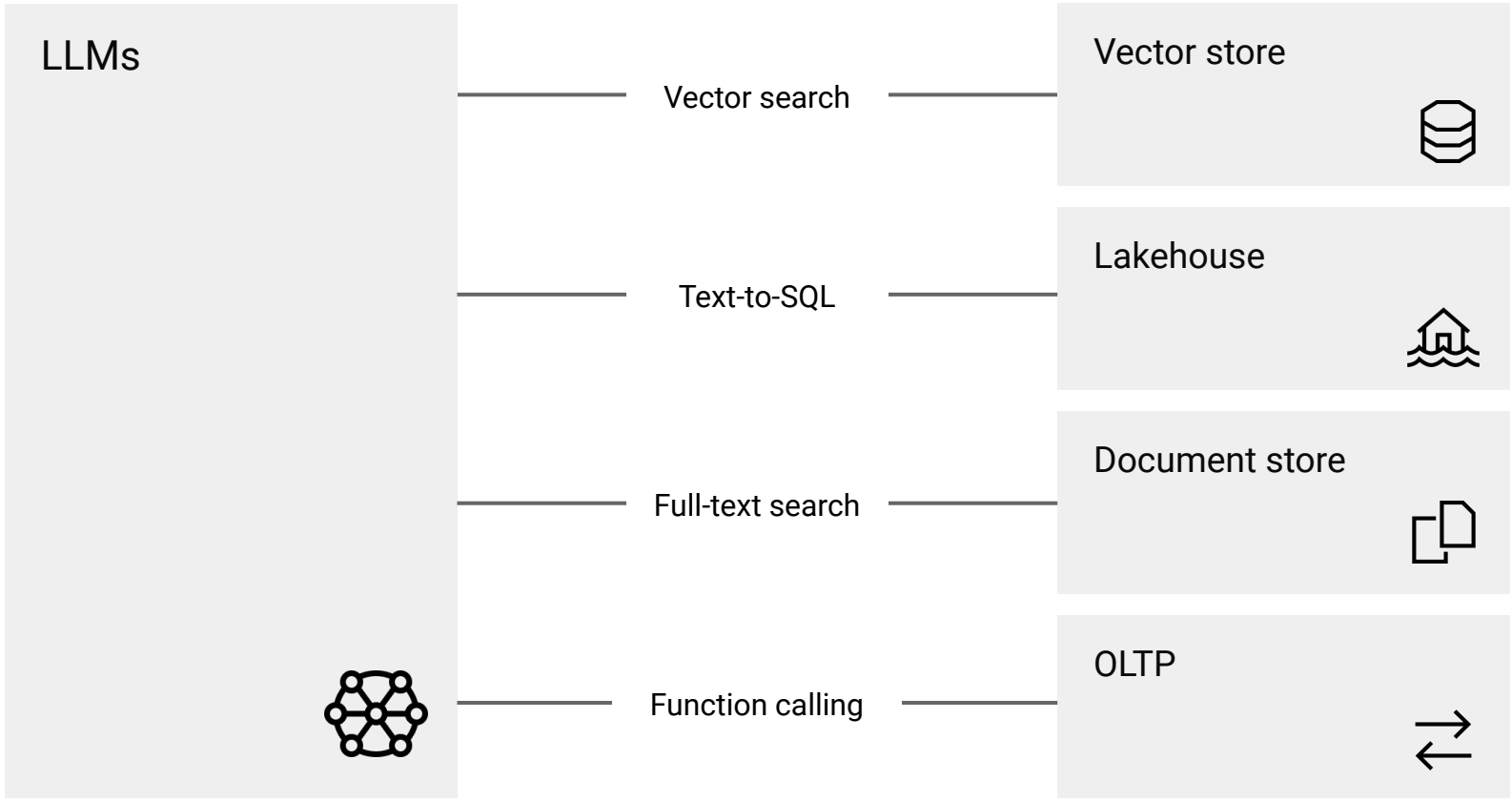
SingleStore

The Data You Need Doesn’t Live in a Vector Store



The Data You Need Doesn’t Live in a Vector Store

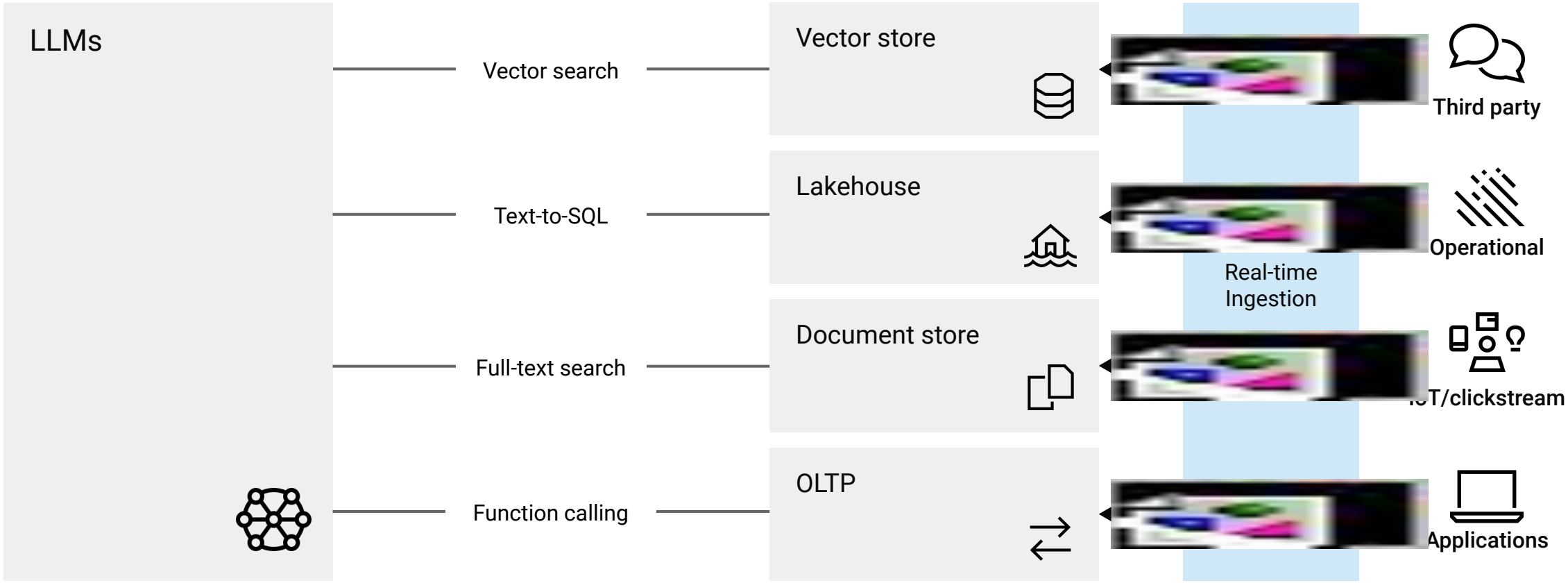
…and this is just for retrieval



SingleStore™

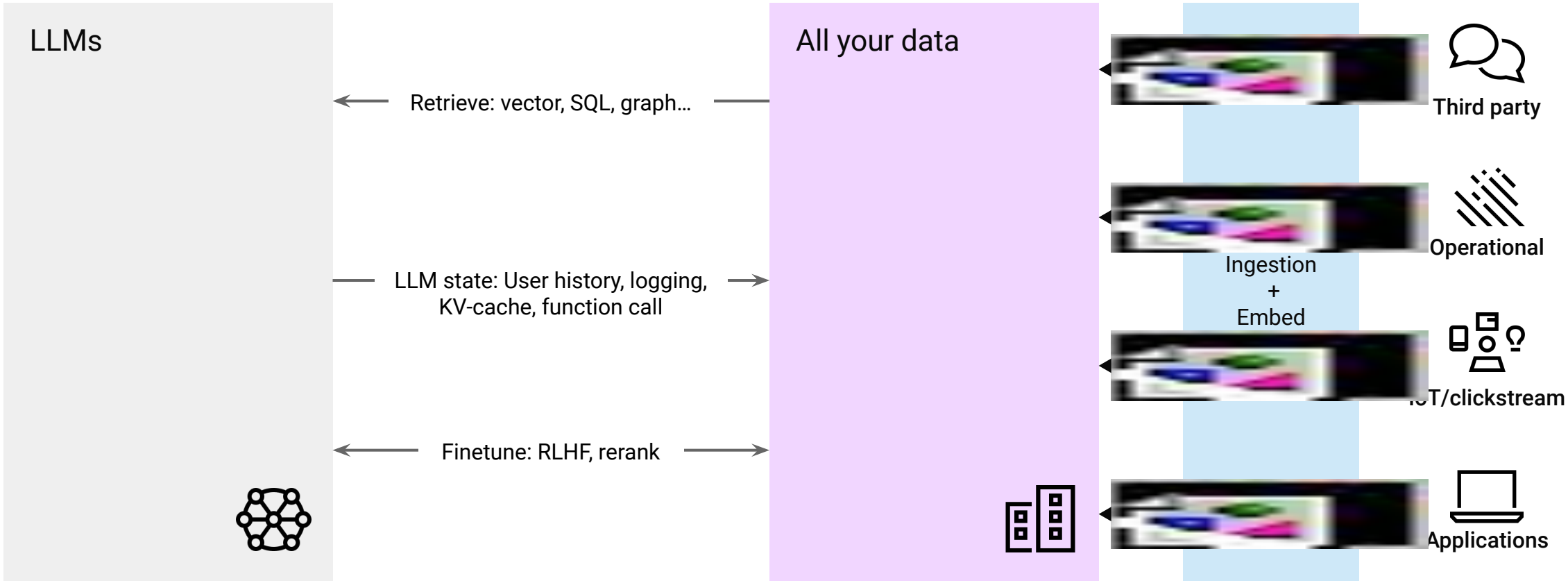
The Data You Need Doesn't Live in a Vector Store

…this is if you want to utilize real-time feeds



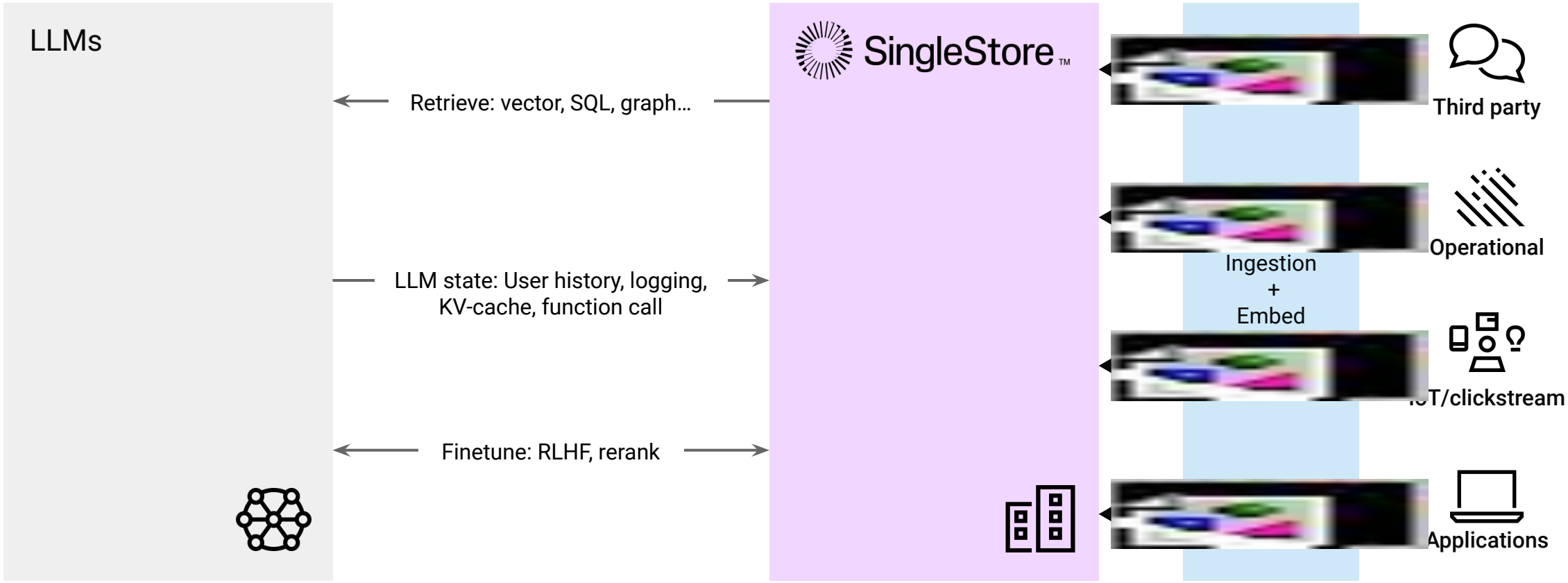
SingleStore™

Wouldn’t It Be Nice…



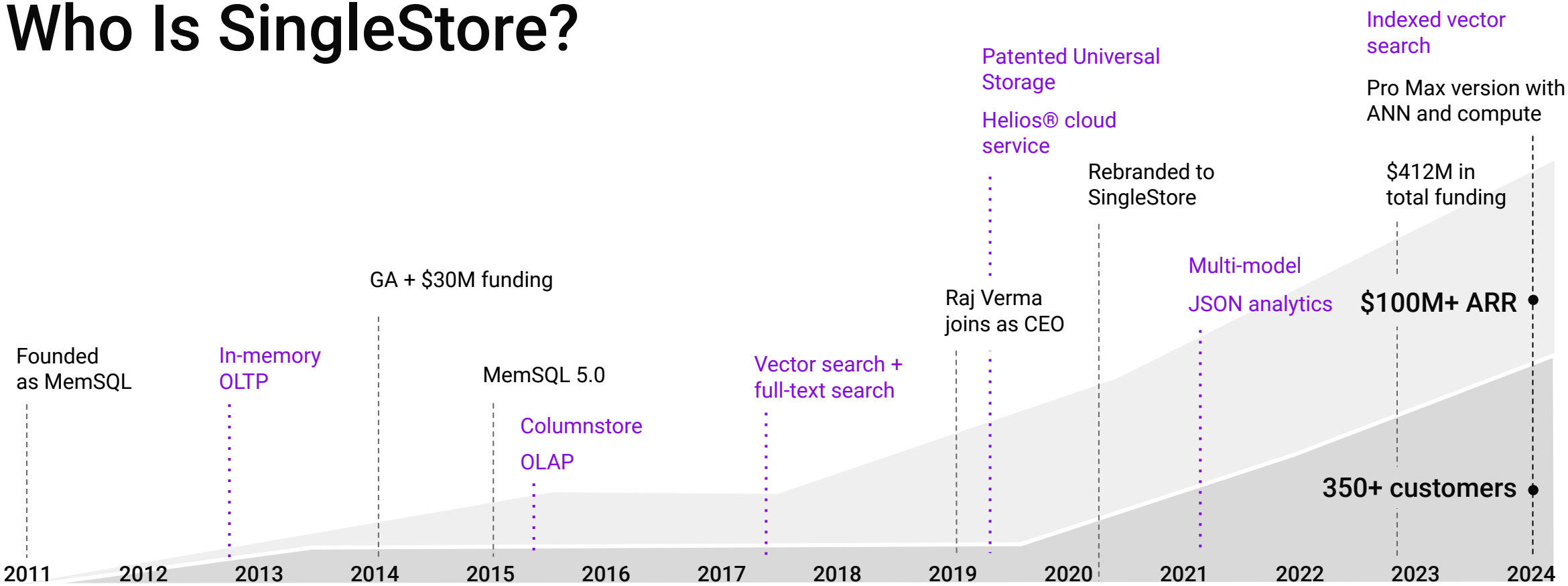
SingleStore

SingleStore was Built for this Use-Case



SingleStore Today

Who Is SingleStore?



Investors

Accel · GLYNN CAPITAL · khosla ventures · SANABIL INVESTMENTS · IBM · G/ · Prosperity7 VENTURES · DELL Technologies CAPITAL · iqt IN-Q-TEL · Goldman Sachs · INSIGHT PARTNERS · R|E|V Funded by RELX · Hewlett Packard Enterprise

SingleStore

Our Historical Strengths: Extensive Ingestion Integrations

1

Integrations with whole enterprise data + fast writes

Parallelized real-time data ingestion from S3, Kafka, HDFS and Iceberg tables

Change data capture (CDC) from MySQL and MongoDB®

Native integration with lakehouse

Enterprise-level security + consistency (ACID, etc.)

2

Bottomless storage

Separation of storage and compute for SingleStore

Always-on Continuous backup and log archive

Data files (snapshots, logs and blobs) are persisted to blob storage

3

High concurrency, low latency queries

High concurrency workloads

Millisecond query performance on relational data, JSON, time-series, vector, full-text search, geospatial and more

Sub MS aggregates for ML and analytics

Our Historical Strengths: Durable Petabyte Scale Storage

1

Integrations with whole enterprise data + fast writes

Parallelized real-time data ingestion from S3, Kafka, HDFS and Iceberg tables

Change data capture (CDC) from MySQL and MongoDB®

Native integration with lakehouse

Enterprise-level security + consistency (ACID, etc.)

2

Bottomless storage

Separation of storage and compute for SingleStore

Always-on Continuous backup and log archive

Data files (snapshots, logs and blobs) are persisted to blob storage

3

High concurrency, low latency queries

High concurrency workloads

Millisecond query performance on relational data, JSON, time-series, vector, full-text search, geospatial and more

Sub MS aggregates for ML and analytics

Our Historical Strengths: Millisecond Query and Aggregation

1

Integrations with whole enterprise data + fast writes

Parallelized real-time data ingestion from S3, Kafka, HDFS and Iceberg tables

Change data capture (CDC) from MySQL and MongoDB®

Native integration with lakehouse

Enterprise-level security + consistency (ACID, etc.)

2

Bottomless storage

Separation of storage and compute for SingleStore

Always-on Continuous backup and log archive

Data files (snapshots, logs and blobs) are persisted to blob storage

3

High concurrency, low latency queries

High concurrency workloads

Millisecond query performance on relational, JSON, time-series, vector, full-text, geospatial and more data types

Sub MS aggregates for ML and analytics

Progress

SingleStore Today	SingleStore 8.9
Real-time ingestion of multi-modal data	
	On-demand embeddings
Unlimited data store	
Fast queries + search	Graph Engine

Native Embeddings in Engine, On-Demand

```
SELECT product_name, product_description,
array_cosine_similarity(embedding('A sleek and
powerful laptop with a high-resolution
display, fast processor, and long battery
life, perfect for productivity and
entertainment on the go.',
'text-embedding-ada-002'),
product_description_embeddings) as similarity
FROM ecommerce.products
WHERE product_name != 'Ultrabook Pro'
  AND category = 'Laptops'
ORDER BY similarity DESC;
```

```
# Get the most similar item from database
for each suggestion from LLM:
   pipeline = [
        {
            "$vectorSearch": {
                "index": "item_index",
                "path": "item_embedding",
                "query": {
                    "$generateEmbeddings": {
                        "source": suggestion,
                        "Model":"openai/text-embedding-3-large"
                    }
                },
                "numCandidates": 1,
                "limit": 1
            },
        },
        {
            "$project": {
                "item": 1,
            }
        }
    ]
    result = collection.aggregate(pipeline)
    for doc in result: print(doc)
```

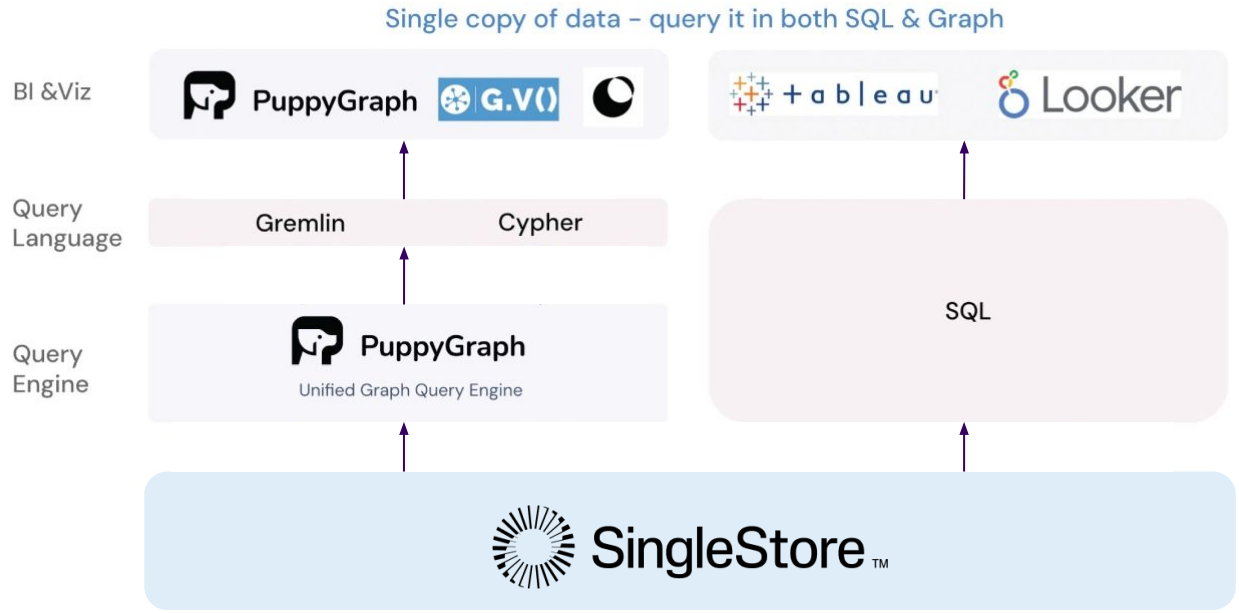
Graph Capabilities Out-of-the Box

Klarna, Microsoft, Meta etc. use knowledge graphs to add relational meaning to data to augment retrieval

Complexities of graph databases today:

- Develop and manage intricate ETL pipelines to morph your data into graph-compatible formats
- Unfamiliar computational and horizontal scaling needs
- Another additional component of your data stack

You can perform graph queries directly on your SQL tabular data in SingleStore, with zero ETL



We Are Built to Serve Enterprises

Default workspace group settings with config manager	Create and update default cluster configurations, and apply them to new or existing workloads seamlessly
Global firewall policies	Manage global firewall policies and apply them to workloads to improve security and reduce administrative overhead
SCIM — Okta + Azure AD	System for Cross-domain Identity Management allows SingleStore users and permissions to be centrally managed in Okta or Azure AD
Project-level settings within organizations	Simplified cluster management, administration and monitoring from central Portal account, centralized user RBAC + SSO, single pane of consumption data
Centralized private networking	Deploy a single private networking link per cloud region enable hundreds or thousands of workloads to access SingleStore without leaving your internal CSP network

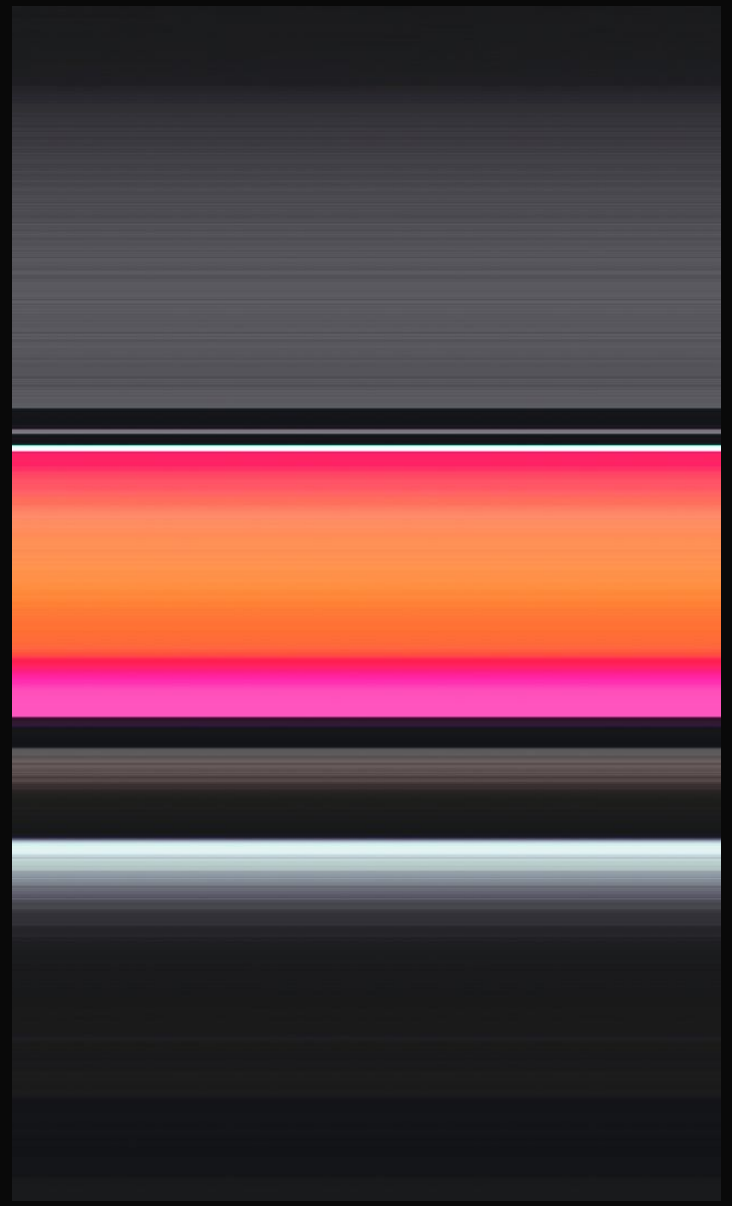
We Are Built to Serve Enterprises

Find more customers at [Made on SingleStore]

SingleStore™

Please reach out!

Email us: ai@singlestore.com



SingleStore™

Thank You