Google Cloud | SingleStore™

Dev Day
SingleStore + Google

Bill Scolinos, Solutions Engineer

Try it Today for a Chance to Win!

Anyone who tries out the SingleStore tutorial today will be entered for a chance to win new Meta Ray Bans Smart Glasses or branded AirPods Pro.

Simply click this link, sign up for a SingleStore Free Trial, try out the features, and we'll announce a winner by the end of today's session!

bit.ly/devday-boston-raffle

Join us for our exclusive Happy Hour!

After our workshop, unwind and network with fellow developers over cold beverages and appetizers atShy Bird, just a short walk from the Google office:

Shy Bird Kendall Square
390 Third St
Cambridge, MA 02142

RAG to Production

1

Rich context for RAG

Data freshness to provide recency and knowledge graphs for better accuracy

Accuracy + richness of context

2

Multi-modal RAG

Ability to store audio, video, images in addition to structured data in one place

Real-time AI

3

RAG with analytics

Ability to provide richness of data back to the LLM and also for better evals

Accuracy + richness of response

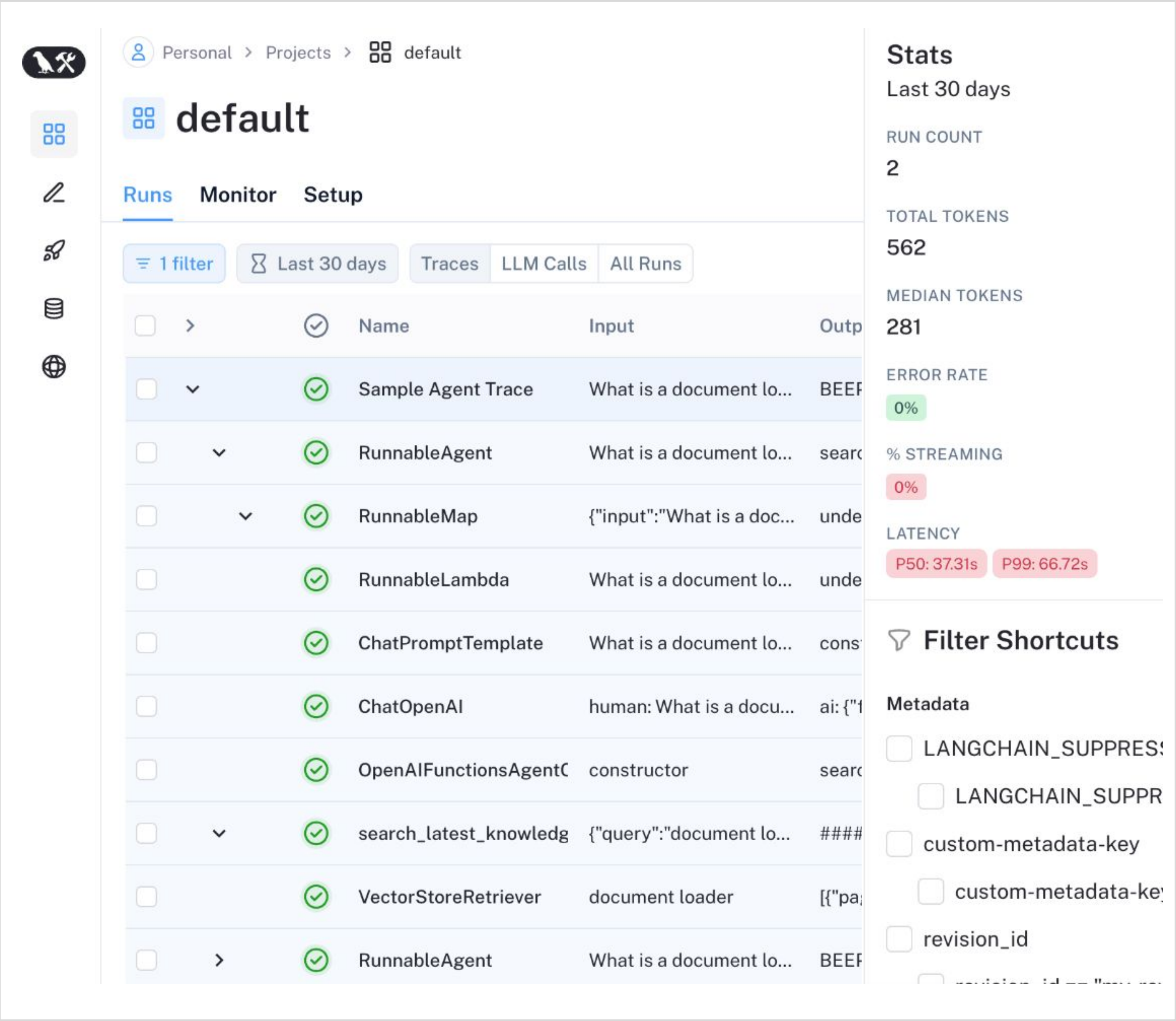
Advanced RAG

Evaluations
(Scoring RAG and LLM responses)

Guardrails
(Vertex AI)

Explainability
(LangSmith)

Re-training on the fly
(LangSmith)



Personal > Projects > default

default

Runs Monitor Setup

1 filter | Last 30 days | Traces | LLM Calls | All Runs

□			Name	Input	Outp
□	⌄	✓	Sample Agent Trace	What is a document lo...	BEEF
□	⌄	✓	RunnableAgent	What is a document lo...	searc
□	⌄	✓	RunnableMap	{"input":"What is a doc...	unde
□		✓	RunnableLambda	What is a document lo...	unde
□		✓	ChatPromptTemplate	What is a document lo...	cons
□		✓	ChatOpenAI	human: What is a docu...	ai: {"f
□		✓	OpenAIFunctionsAgent(	constructor	searc
□	⌄	✓	search_latest_knowledg	{"query":"document lo...	####
□		✓	VectorStoreRetriever	document loader	[{"pa
□	›	✓	RunnableAgent	What is a document lo...	BEEF

Stats
Last 30 days

RUN COUNT
2

TOTAL TOKENS
562

MEDIAN TOKENS
281

ERROR RATE
0%

% STREAMING
0%

LATENCY
P50: 37.31s P99: 66.72s

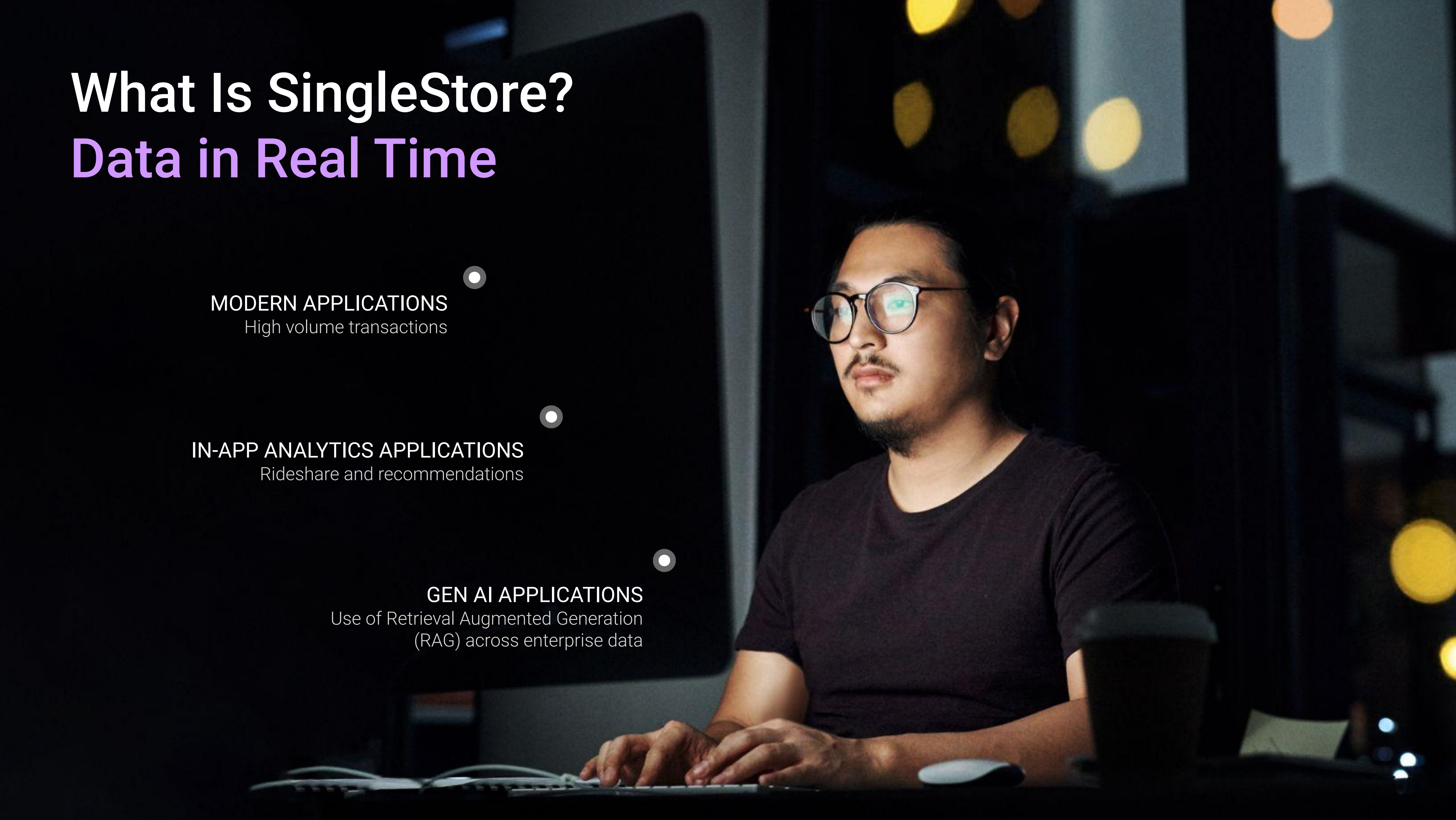
Filter Shortcuts

Metadata

- □ LANGCHAIN_SUPPRES
  - □ LANGCHAIN_SUPPR
- □ custom-metadata-key
  - □ custom-metadata-ke
- □ revision_id

SingleStore + Google Vertex AI

What Is SingleStore?
Data in Real Time

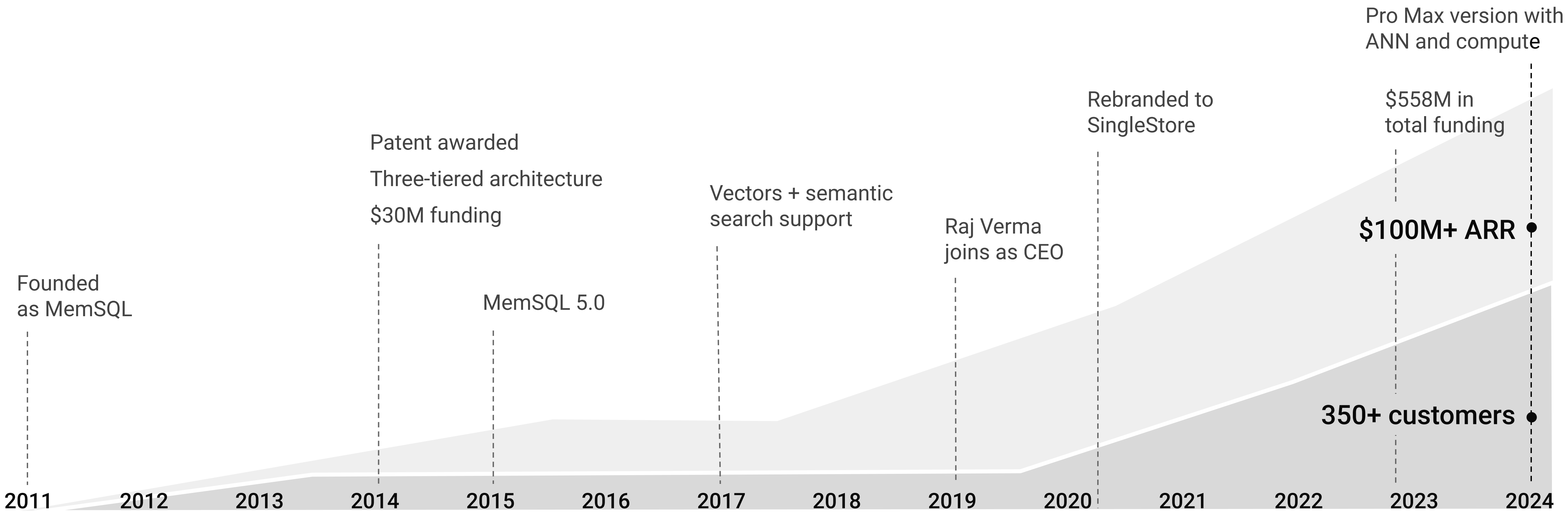


MODERN APPLICATIONS
High volume transactions

IN-APP ANALYTICS APPLICATIONS
Rideshare and recommendations

GEN AI APPLICATIONS
Use of Retrieval Augmented Generation (RAG) across enterprise data

Who Is SingleStore?



Founded as MemSQL

Patent awarded

Three-tiered architecture

$30M funding

MemSQL 5.0

Vectors + semantic search support

Raj Verma joins as CEO

Rebranded to SingleStore

$558M in total funding

Pro Max version with ANN and compute

$100M+ ARR

350+ customers

Investors

Accel

GLYNN CAPITAL

khosla ventures

سنابل للاستثمار SANABIL INVESTMENTS

IBM

iqt IN-Q-TEL

Goldman Sachs

INSIGHT PARTNERS

R | E | V Funded by RELX

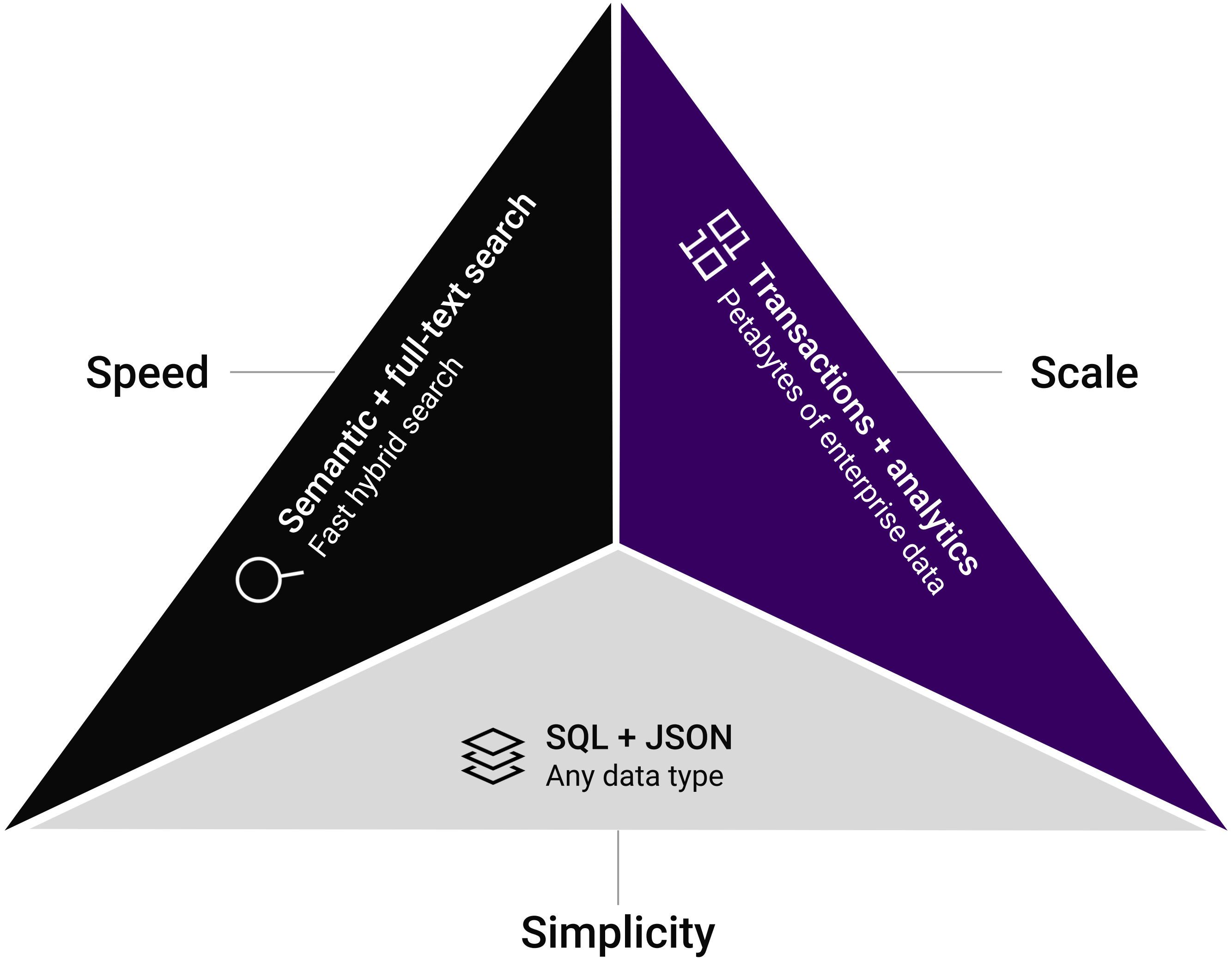
Hewlett Packard Enterprise

GV

Prosperity7 VENTURES

DELL Technologies CAPITAL

SingleStore Tenets



Google Cloud | SingleStore™

SingleStore
Key product capabilities

Universal Storage
Patented single table type for transactions and analytics. Compatibility with ANSI SQL, MySQL + MariaDB ecosystem; NoSQL

Pure speed
Fast transactions, analytics, vectorization and query compilation

Fast ingestion
SingleStore Pipelines — load data with updates

Bottomless storage
Separation of storage + compute in a unified database

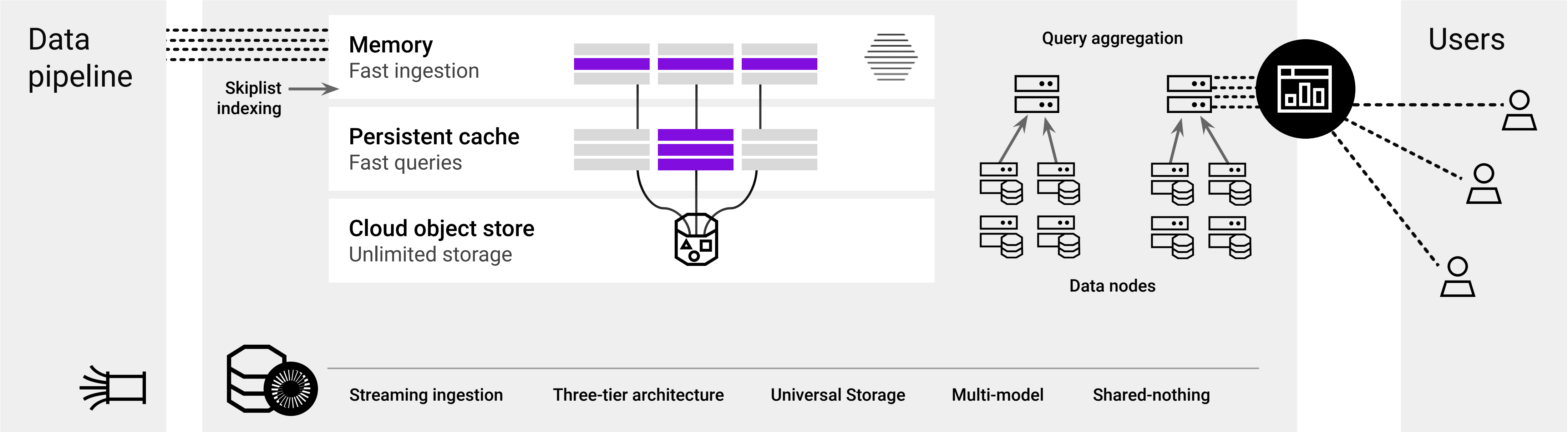
MongoDB® API
SingleStore Kai™ powers up to 1,000x faster analytics on JSON

Multi-model
Relational foundation with support for various data types, including vectors

Run anywhere
Hybrid, multi-cloud, SaaS, on-premises, Kubernetes operator

Intelligence
Build generative AI and other apps with **Notebooks** on top of SingleStore

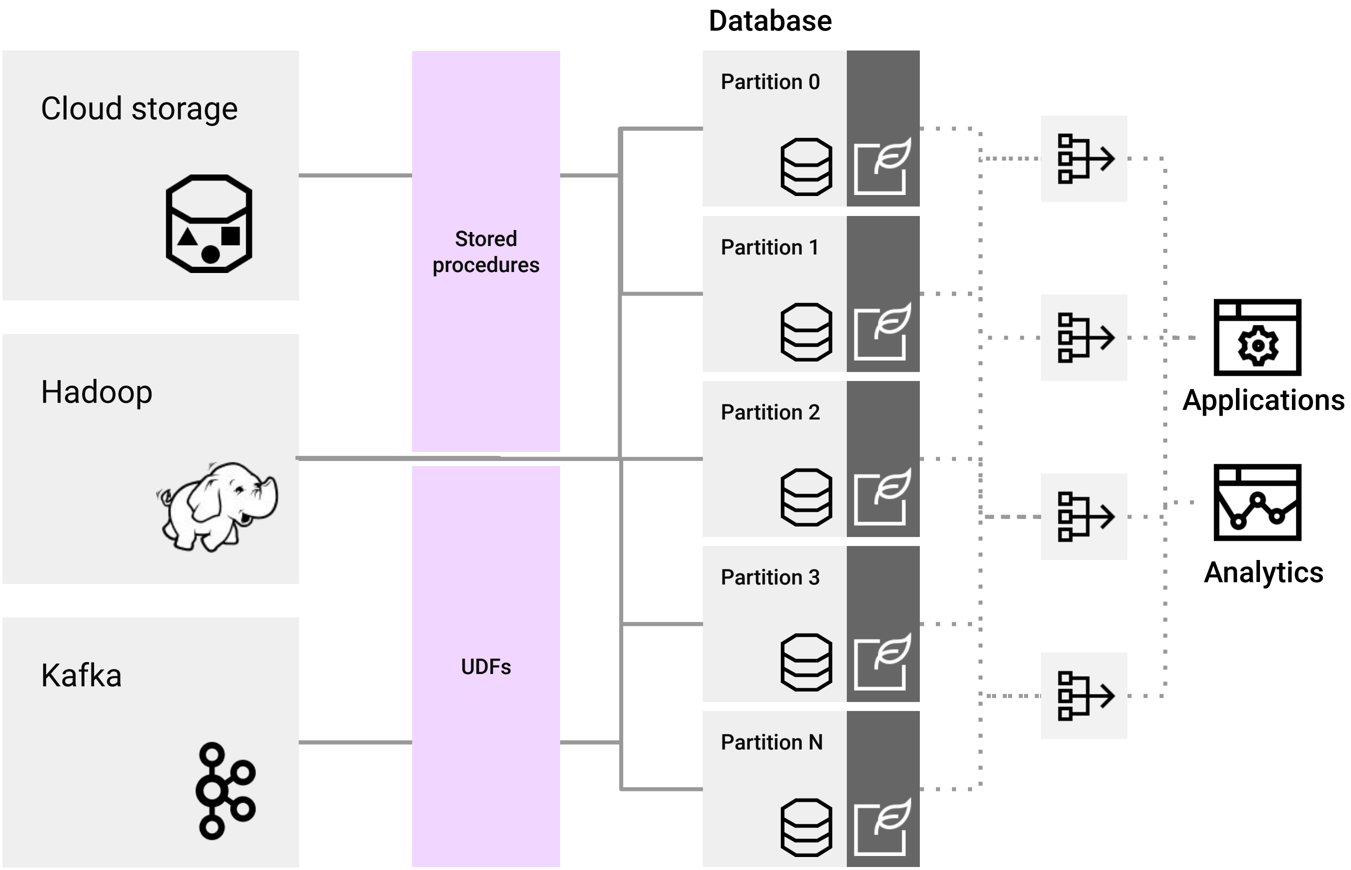
SingleStore: Under the Hood



Pipelines: Real-Time Ingest

Product overview

- Millions of records / sec with immediate availability
- Built-in component of database
- Transactional consistency
- Exactly-once semantics
- Native integrations with Kafka, cloud storage, HDFS
- Ingest to stored procedures to shape/modify data
- Pipelines support JSON, Avro, Parquet and CSV data formats.



Workspaces Overview

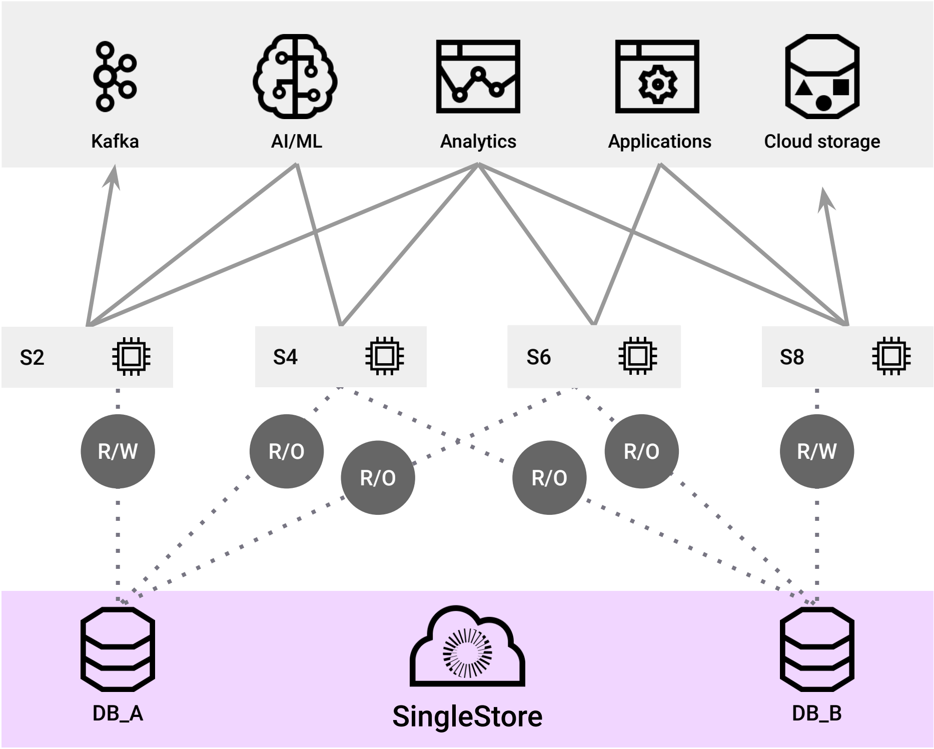
Isolated workloads on shared data

Compute operates on shared data, without one workload affecting the performance of others

- Allow granular scalability and isolation of compute resources
- Eliminate cost of moving and maintaining data between multiple workloads
- Scale ingest and compute workloads independently
- Isolate internal and customer facing real-time applications simultaneously on shared data

Separation of compute + storage

Separate write transactions from read-only workloads (analytics) each with its own dedicated compute resources — without data duplication.



Each workspace is isolated and can scale compute independently

Searching + Contextualizing

Simple semantic search

```
stmt = f"""
        SELECT
            text,
            DOT_PRODUCT_F64(JSON_ARRAY_PACK_F64(%s),
embedding) AS score
        FROM {table_name}
        ORDER BY score DESC
        LIMIT %s
    """.format(table_name=table_name)
```

Hybrid search

```
SET @promptEmbedding = '[...]' :> vector(1536) :>
blob;
SELECT
  products.id,
  MATCH(products.description) AGAINST ('blue') AS
ft_score_color,
  title_v <*> @promptEmbedding AS v_score_title,
  description_v <*> @promptEmbedding AS
v_score_description,
  v_score_title + v_score_description AS score
FROM products
JOIN product_sku ON products.id =
product_sku.product_id
JOIN product_sizes ON product_sku.product_size_id =
product_sizes.id AND product_sizes.value = 'xs'
WHERE ft_score_color AND (v_score_title >= 0.75 OR
v_score_description >= 0.75)
AND price BETWEEN 100 AND 1000
GROUP BY products.id
ORDER BY score DESC
LIMIT 5;
```

SingleStore AI Capabilities

Core engine

MySQL wire compatibility

Scale out (distributed SQL)

SQL compiled to machine code

Universal storage table type

Three-layer storage architecture

Enterprise-grade availability, security

Point-in-Time Recovery (PITR)

Projections and branching

Hybrid search

ANN with PQ, IVF, HNSW

Multiple indices on one column

Exact KNN search

Vector datatype

JLucene under-the-hood for exact keyword search

Ecosystem

CDC in from MySQL, MongoDB®

Pipeline with Kafka

Connector in LangChain, Llamaindex, Flowise, etc.

SDKs for CRUD in different languages

Optimize Tradeoffs on Your Own Terms

IVF-PQ (ANN)

+ Low index build time
+ Cost (smaller index size)
- Recall

HNSW-FLAT (ANN)

+ Recall
+ Speed (performant at high dimensionality)
- Large index size

kNN

+ Recall
+ Cost (no added size)
- Slow at scale (vs ANN)

Vector Range Search using ANN index

```
select id, txt, vec <*> @query_vec as score
from t
where score > 0.70;
```

What it's good for:

- Answer, is there anything even close to this query vector?
    - "even close" might mean "vector_col <*> @query_vec > 0.60"
        - can be faster than a "top K" query

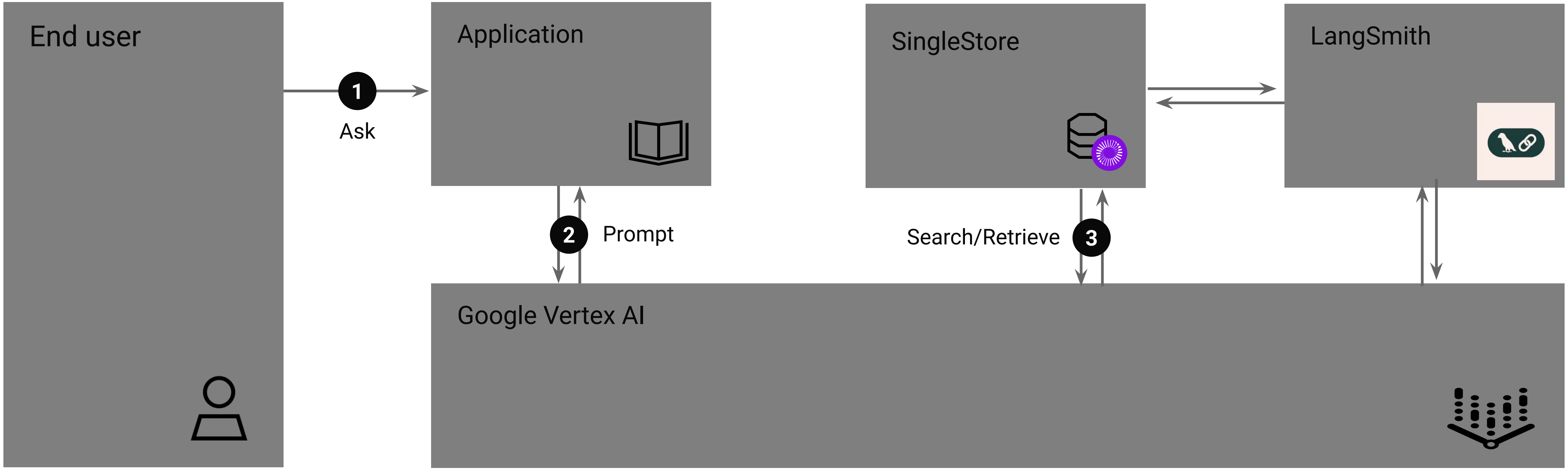
- Find "all the really close stuff"
    - "vector_col <*> @query_vec > 0.95"
    - works in one shot, easier to program that running top 5, top 10, top 20 in a loop

15-min break

Demo: Production Grade RAG with SingleStore + Vertex AI

Production-ready RAG Architecture

1 min



5 min

Production Grade RAG

Scenario

- **Input:** User question regarding contract data
- **Output:** Response to user question
- **Goal 1:** Provide the user with fast responses to their questions
- **Goal 2:** Log all relevant information

Why SingleStore + GCP?

ML Model Deployment on Vertex AI

Streaming ingest through Pipelines

Fast Vector Search in SingleStore

Large context window with Gemini

Google Cloud | SingleStore™

Thank You