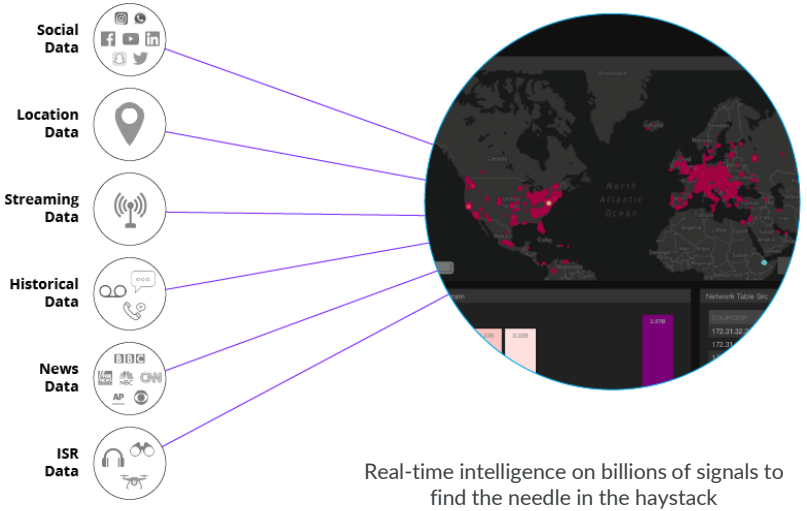

Kinetica
The real-time accelerated database for analytics and generative AI
kinetica

# Kinetica - Born out of National Defense need...

Kinetica's origins were with the US Department of Defense in 2009 to find & track entities/bad actors to prevent terrorist attacks worldwide.

Mission Data Requirements

- Need to Collect data from over 200 sources (time series and spatial data) in near real-time and make it available immediately for query
- Whilst maintaining performance of queries as data is simultaneously ingested
- To create an operational view of the entity **in real time.**




Real-time intelligence on billions of signals to find the needle in the haystack

MERITECH GREATPOINT make no little plans Canvas VENTURES citi VENTURES

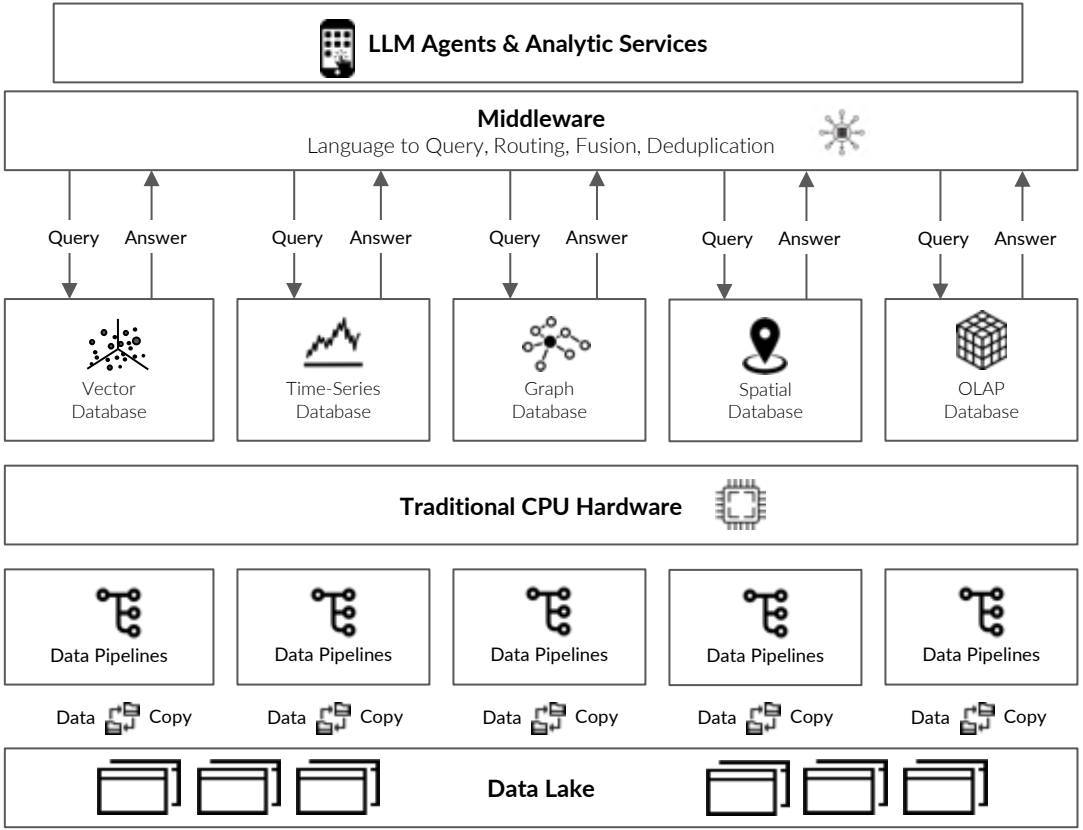
**Real-time data and compute use to be a nice to have in the enterprise.**

**Generative AI has made it a requirement.**

# Moving from Reports and Dashboards to Co-Pilots and Agents has changed the demands on the enterprise data stack

| 2022 | Now |
| --- | --- |
| Tolerance for older data and long running queries | **Zero tolerance** any delay of data or query response delivers a broken experience |
| Careful planning and data engineering to support a curated query inventory for the report or dashboard | Ad-hoc and unpredictable queries being generated from user conversations or agent to agent interactions |

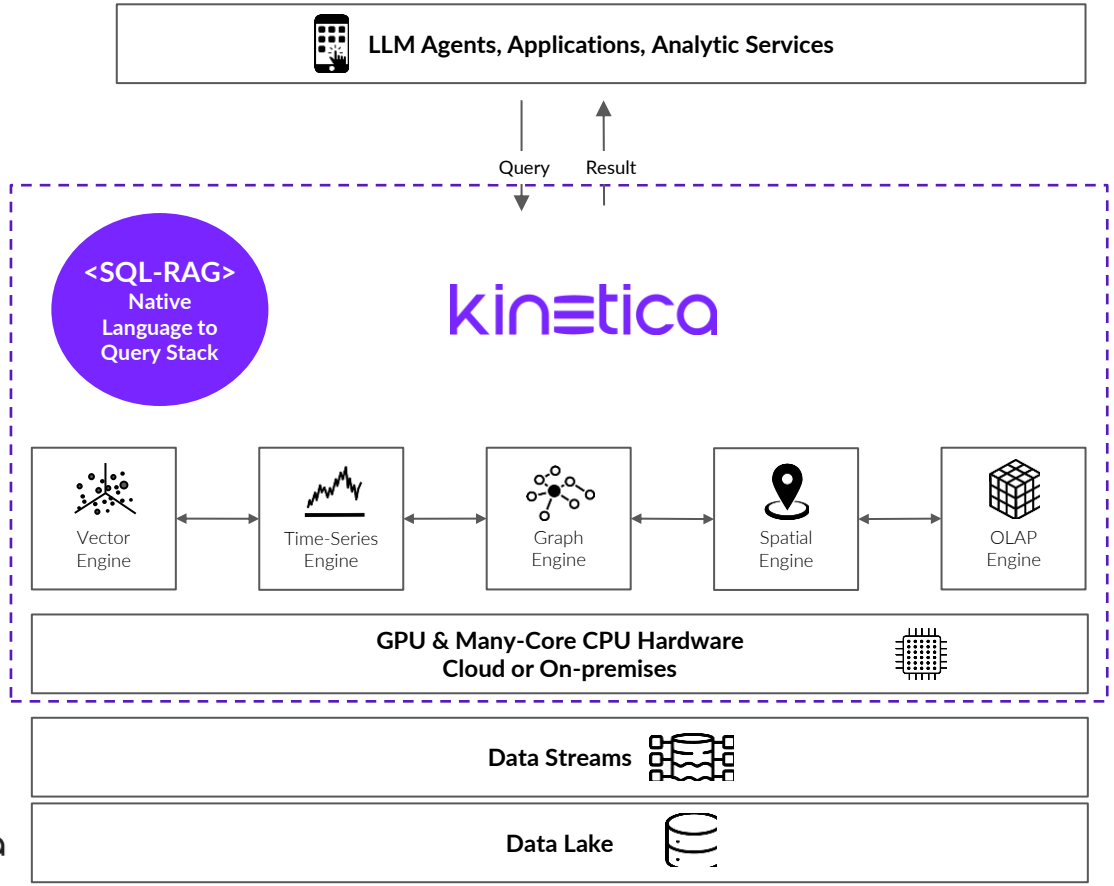
# Current Real-Time Enterprise Architectures aren't ready to power seamless retrieval for co-pilots and agents




## Common Pain Points

- **Data latency**
- **Synchronization between single domain products**
- **Infrastructure Sprawl & Data Duplication**
- **Query Performance**

# What Real Time Retrieval Looks Like With Kinetica




**Translate language to SQL with tight integration into compute layer with a single API**

**Converged analytic engines to answer any question**

**Next generation parallelism removes the need to build data pipelines, allowing unknown questions to be asked**

**Query real-time and historical data**

# Kinetica

The unified real-time retrieval engine for Generative AI and Analytics

## Real-time

- Minimize data latency for relational queries and vector similarity search
- Ingest high volume data streams and execute queries simultaneously

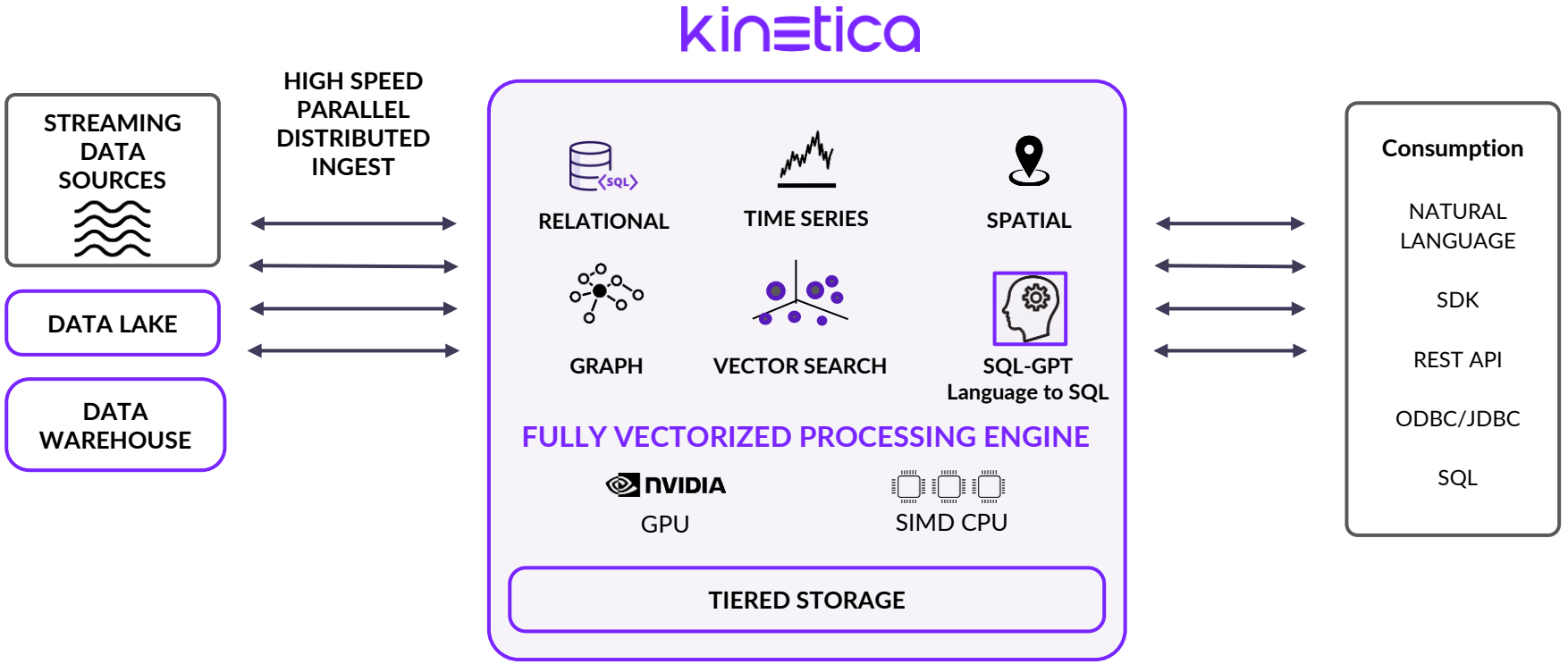
## Many-core Accelerated

- Execute complex analytical and vector search queries in sub-seconds
- Answer ad-hoc multi-modal analytic queries and accelerate vector search, even without an index

## Multi-Modal RAG

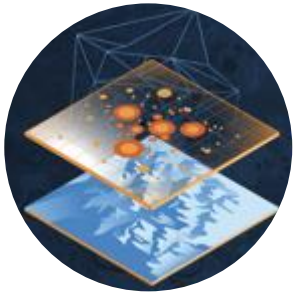
- **SQL**: Generate accurate SQL at enterprise scale with Kinetica Native LLM and Metadata Catalog
- **Graph**: Vector Search Weighted Traversal and Matching
- **Best in class Real-time Vector Search**

# Kinetica Overview - Architecture




*The **Only GPU Accelerated Database** that is distributed and **PostgreSQL** Compliant*

# Next Level Analytic Capabilities



**Spatial**

- Geo-joins
- ST_Geometry suite
- Entity Tracks
- Server Side Visualization
- And more

**Time-Series**

- AS-OF joins
- Aggregations
- Window functions
- Interpolations
- And more

**Knowledge Graph**

- Adjacency
- Centrality
- MSDO
- Shortest Path
- And more

**AI**

- Vector Search
- Embedding generation
- NL2SQL
- ML Inference
- And more

# Common Vector RAG Pitfalls in the Enterprise

**Readily available embedding models are ill-suited for creating vector embeddings of structured relational data. Match accuracy drops dramatically as you increase data causing LLMs to give highly misleading responses.**

**Creating vector embeddings of operational sources expose extreme data latency issues in most vector databases meaning LLMs are often hours or days behind ground truth.**

**Many vector search engines lack the fine-grained data access controls to limit which vectors a user can find matches for.**

# Real-Time Vector Similarity Search with Kinetica

## Minimal data latency

Kinetica's innovative internal index orchestration delivers near zero data latency and outstanding query throughput

## Horizontal scaling

Scale to handle billions of embeddings

## Enterprise Grade Security

Fine-grained data access controls to limit which vectors a user can find matches for that easily integrate with enterprise identity systems


## GPU accelerated

NVIDIA's RAPIDS cuVS and CAGRA for outstanding non-indexed and indexed search performance

## Pg Vector Compliant

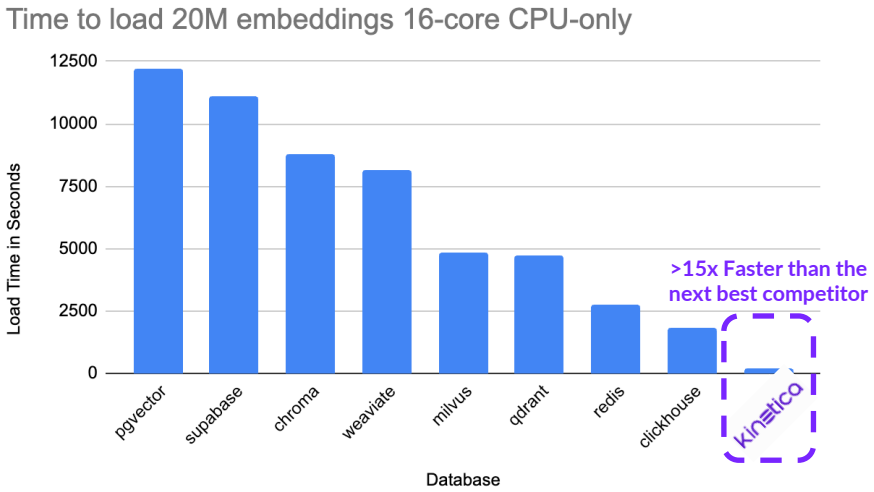
Fully integrated with SQL for powerful
hybrid analytics

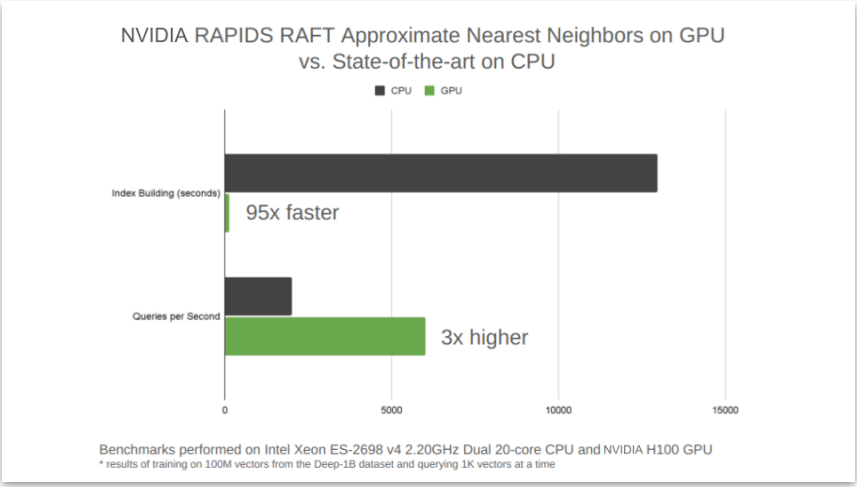
## Vector Search for Graph and Relational Query

The only engine that allows you to do both vector search weighted graph and relational query

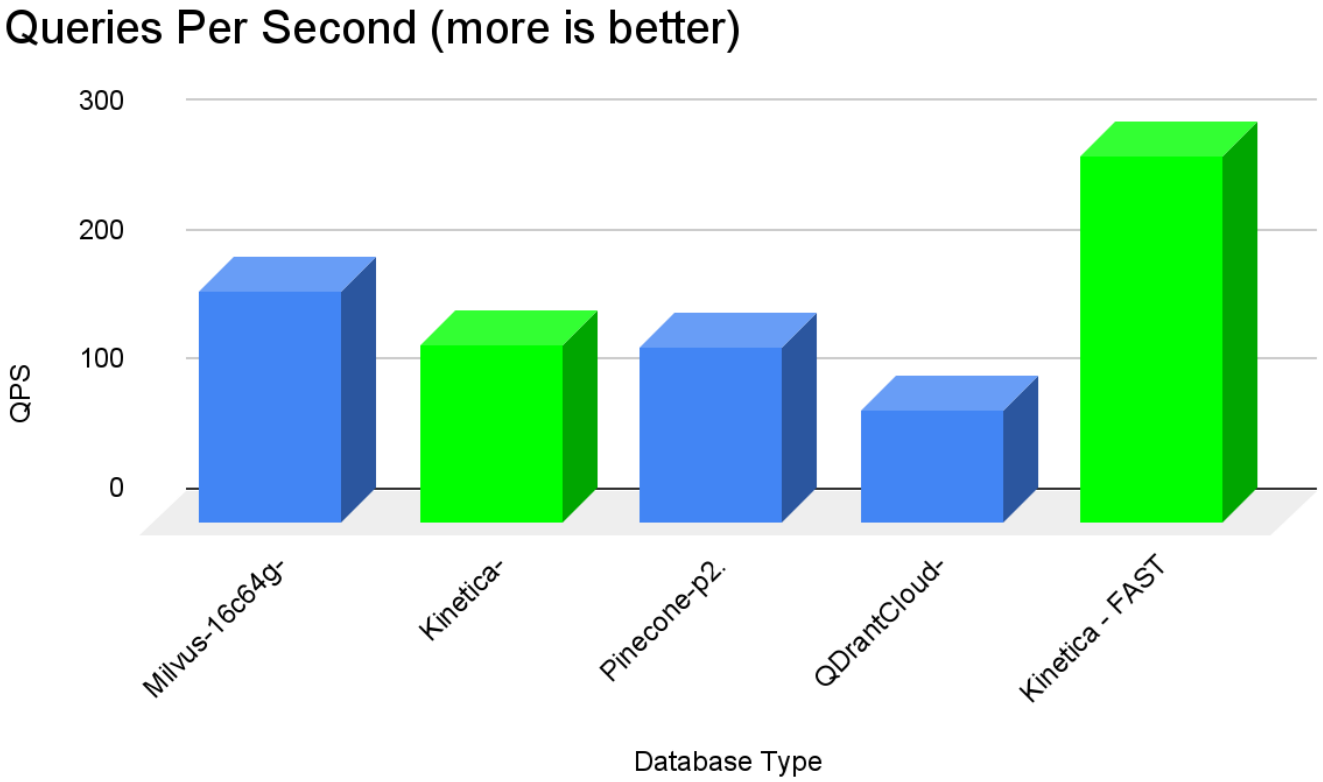
# Real-Time Vector Search

The only CPU and GPU accelerated vector search that offers scale and speed



# VectorDBBenchmark - Query Latency - 10m 768 Dim - Low Filter




| Database Type | QPS | Recall |
|---|---|---|
| Milvus-16c64g-hnsw | 178 | 92% |
| **Kinetica-Cagra** | **138** | **99%** |
| **Kinetica-FAST HNSW** | **457** | **95%** |
| Pinecone-p2.x1-8node | 136 | 92% |
| QDrantCloud-4c16g-5node | 87 | 89% |

# Common SQL-RAG Pitfalls in the Enterprise

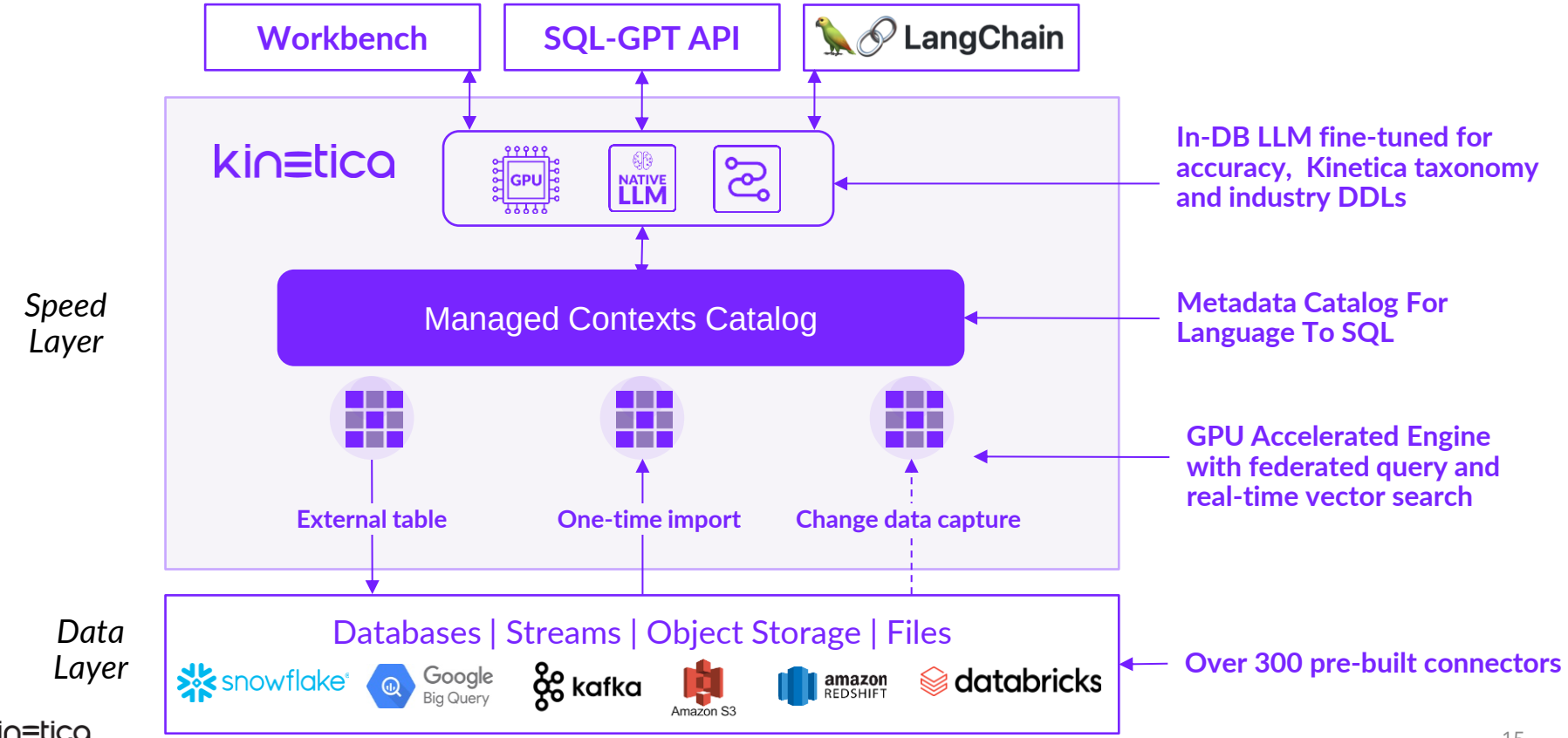
**Enterprises quickly exceed the context token limit when attempting to supply all necessary meta-data in the LLM context.**

**Existing meta-data systems are ill suited to capture and make available for query/retrieval unique elements required in language-to-sql workload**

**Meta-data that should not be seen by user leaks into context and is included in LLM response.**

**Minor variances in SQL dialects across databases make it difficult for LLM to generate syntactically correct SQL across a number of data sources in federated query scenarios.**

# Kinetica Real-Time RAG Architecture



# Telco Network Troubleshooting

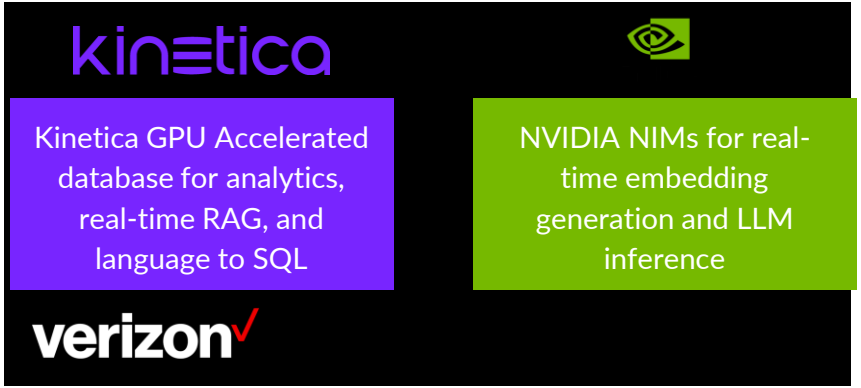
## Faster troubleshooting and root cause analysis of network issues

### Reduce Troubleshooting Time
Automated RCA identification streamlines error detection

### Increase Call Quality
Proactive AI interventions improve overall call performance

### Enhance Customer Satisfaction
Reduced downtime leads to better call quality and customer satisfaction.

### Boost Efficiency
Human language querying simplifies analysis, making it more accessible.




| Data Source | Retrieval Type | Data Lifecycle |
|---|---|---|
| Network PCAP | SQL | Real-Time |
| System Telemetry | SQL | Real-Time |
| 3GPP and other Knowledge Documents | Vector | Static |

# Air Domain Intelligence at US Air Force

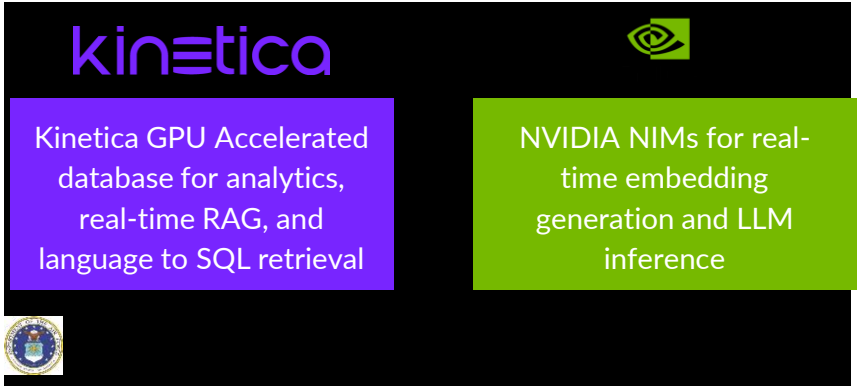
## Anomaly detection and analysis co-pilot

### Automated Threat Monitoring

Leverage multiple data sources to understand new attack vectors and analyze airspace radar activity for potential signal.

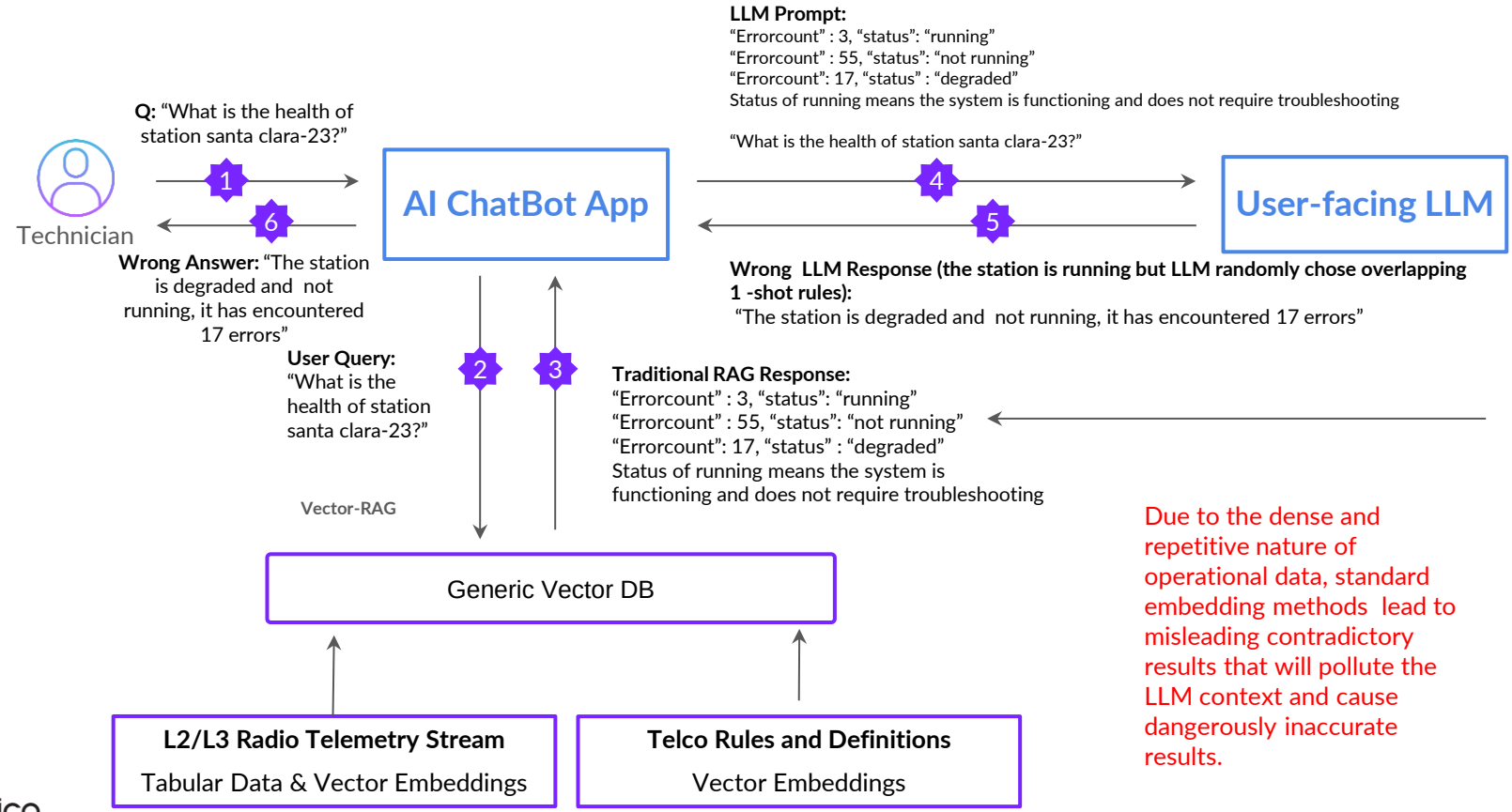
### Force Multiplier

Human language querying of relational and vector data simplifies analysis, making more data more accessible for analysts.



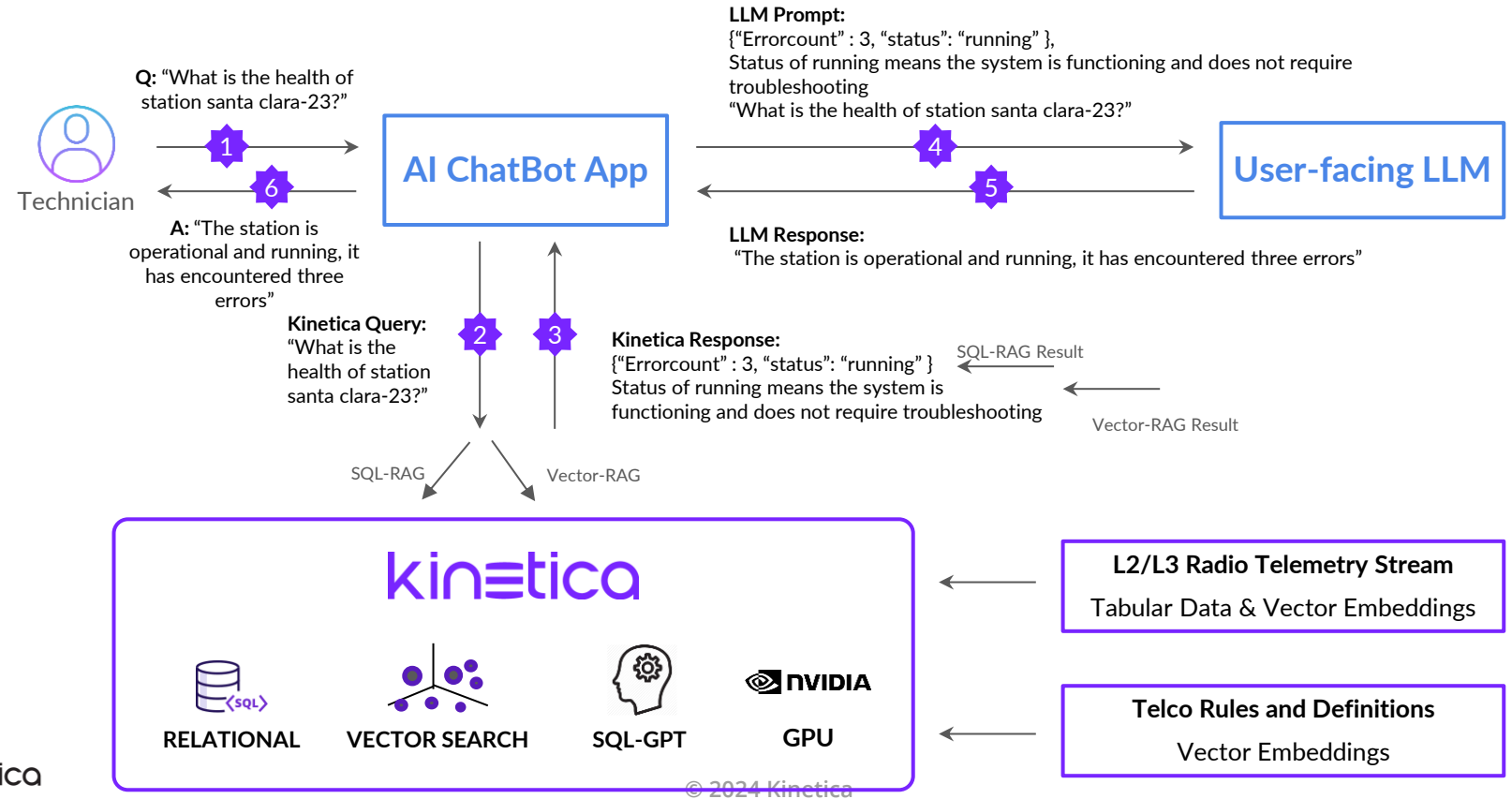

| Data Source | Retrieval Type | Data Lifecycle |
|---|---|---|
| Military Airspace Telemetry | SQL | Real-Time |
| Commercial Industry Airspace Telemetry | SQL | Real-Time |
| Intelligence Knowledge-Base Documents Sources | Vector | Real-Time |

# Dangerous Failure: Network Telemetry Co-Pilot with traditional RAG (no Kinetica)




Due to the dense and repetitive nature of operational data, standard embedding methods lead to misleading contradictory results that will pollute the LLM context and cause dangerously inaccurate results.

# Empowering Success: Network Telemetry Co-Pilot with Kinetica RAG



# Powering The Next Generation Fraud and AML analytic and alerting platforms

BANK OF AMERICA

Real-time Trading

Connected Car

**> Anti-Money Laundering**

Network Optimization

Federated Vector Search

Real-time ✓

Spatial ✓

Relational ✓

Graph ✓

**500 Million** Records Daily

2s Data Latency SLA

2s Query Latency SLA

# Low latency time-series risk analytics at Global Scale

**> Real-time Risk**

Connected Car

Entity Tracking & Logistics

Network Optimization

Federated Vector Search


Real-time ✓

Time-series ✓

Relational ✓

**5 Billion** Records Daily

1s Data Latency SLA

2s Query Latency SLA

# Powering The Edge: Advanced Autonomous Drones

LOCKHEED MARTIN

Real-time Trading

Connected Car

**> Autonomous Edge**

Network Optimization

Federated Vector Search


- Real-time ✓
- Spatial ✓
- Relational ✓
- Vector Search ✓

**Hundreds of Millions of** Records per Flight

< 1s Data Latency SLA

< 1s Query Latency SLA

kinetica

# Powering Generative AI across data lake with real-time vector search

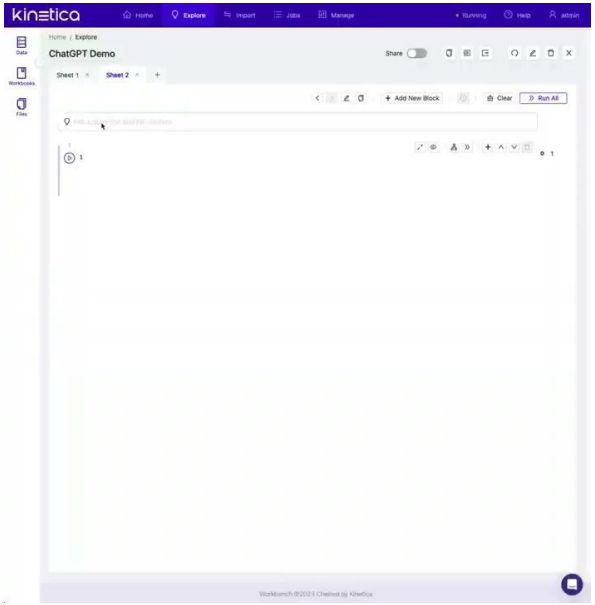

Real-time Trading

Connected Car

Entity Tracking & Logistics

Network Optimization

**> Real-Time Vector Search**




Real-time ✓

Time-series ✓

Relational ✓

Vector Search ✓

**10 million vectors** Daily

1s Data Latency SLA

2s Query Latency SLA

# Powering Generative AI query operations and visualization for real-time sensor data

AIRFORCE

Real-time Trading

Connected Car

**> Entity Tracking & Logistics**

Network Optimization

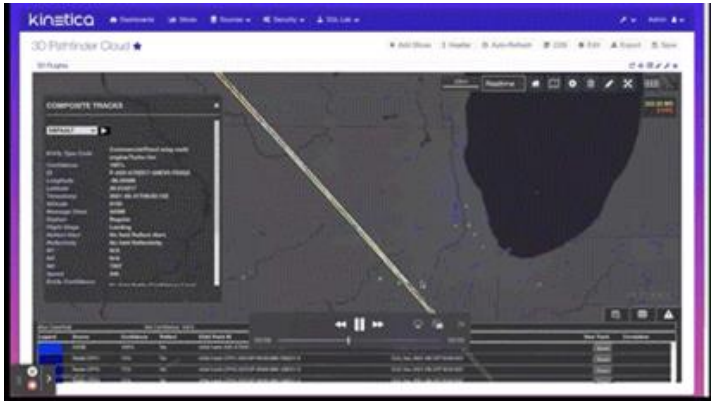
Federated Vector Search

Real-time ✓

Spatial ✓

Relational ✓

Vector Search ✓




**3 Billion** Records Daily

1s Data Latency SLA

2s Query Latency SLA

# Only Kinetica Has This Critical Combination of Capabilities

| | Real-time | OLAP and SQL | Vector Storage and Search | Graph | Time Series | Spatial | GPU Support & Acceleration | Scalability |
|---|---|---|---|---|---|---|---|---|
| Vector DBs (Pinecone, Weaviate, Qdrant…) | | | CPU Only | | | | | |
| Analytic DBs (Clickhouse, Singlestore) | | | CPU Only | | | | | |
| **Kinetica** | | | CPU,GPU | | | | | |

# Demo


Q & A
kinetica