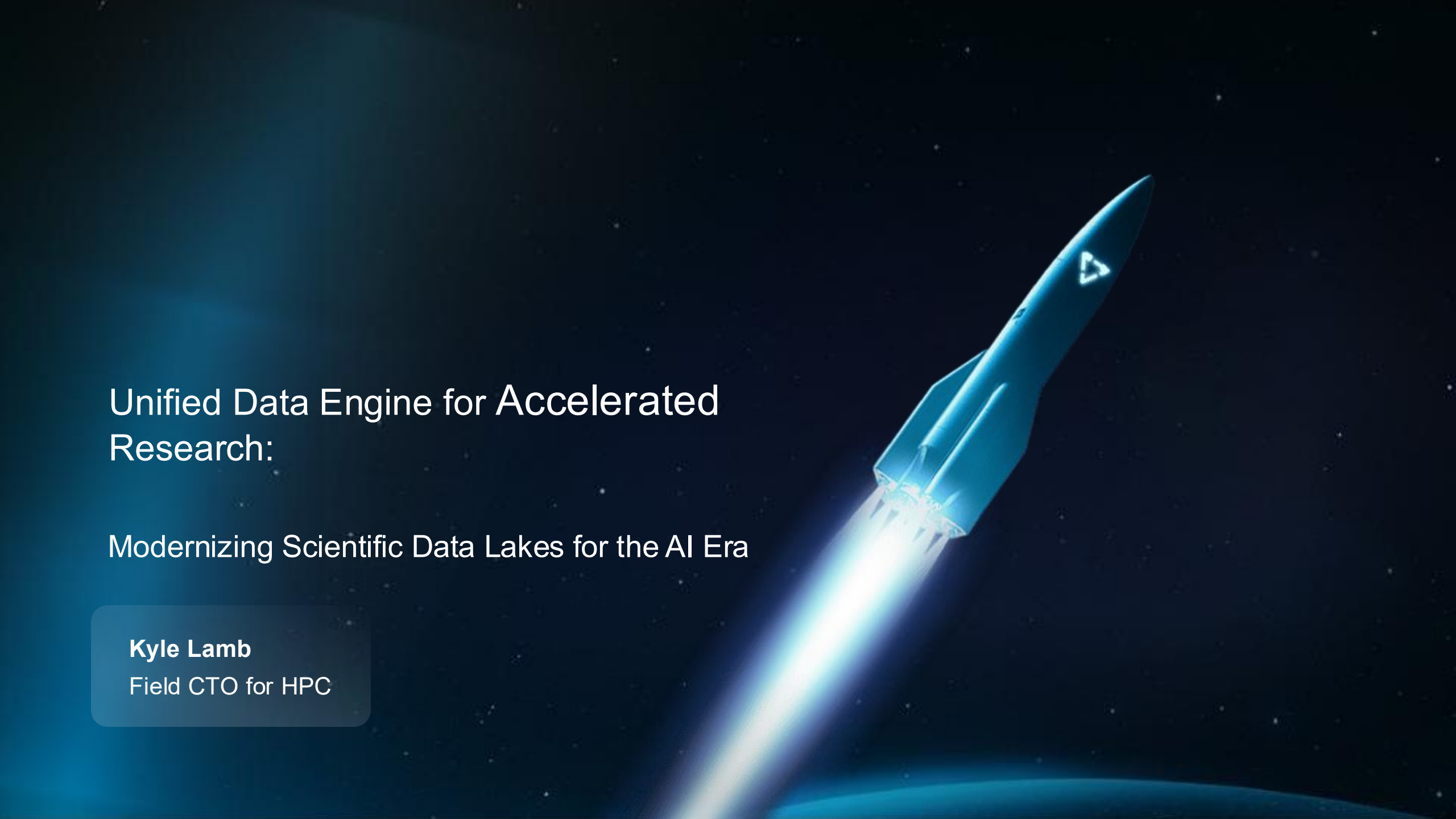


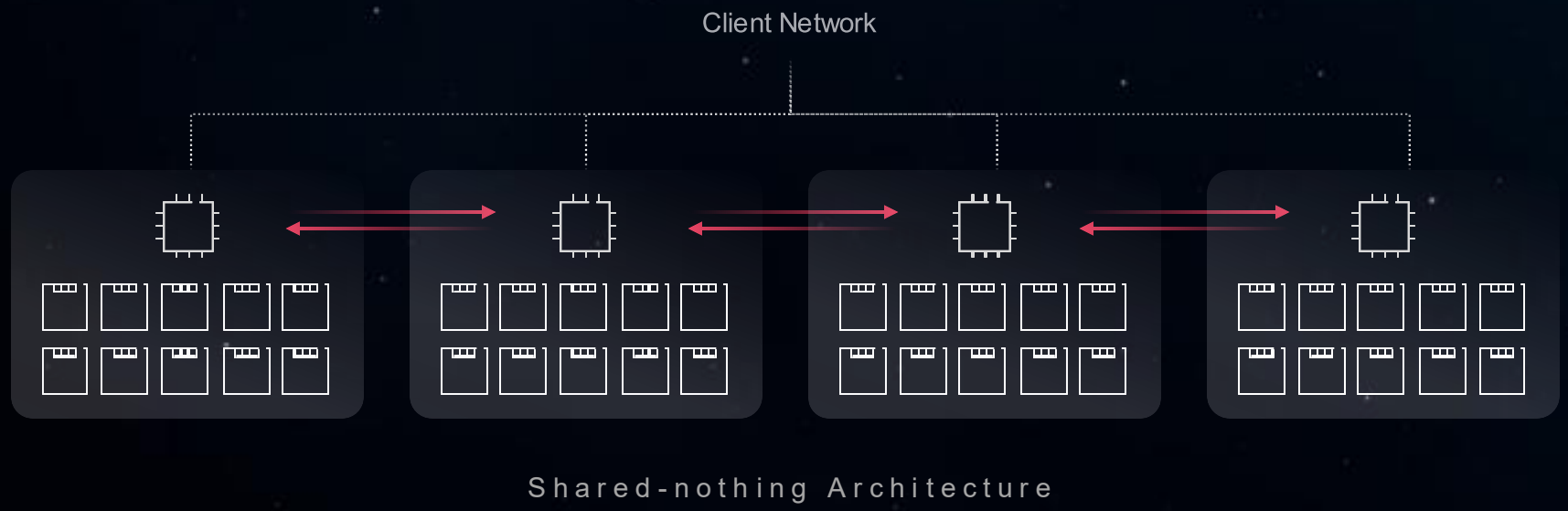
Unified Data Engine for Accelerated Research:

Modernizing Scientific Data Lakes for the AI Era

Kyle Lamb
Field CTO for HPC

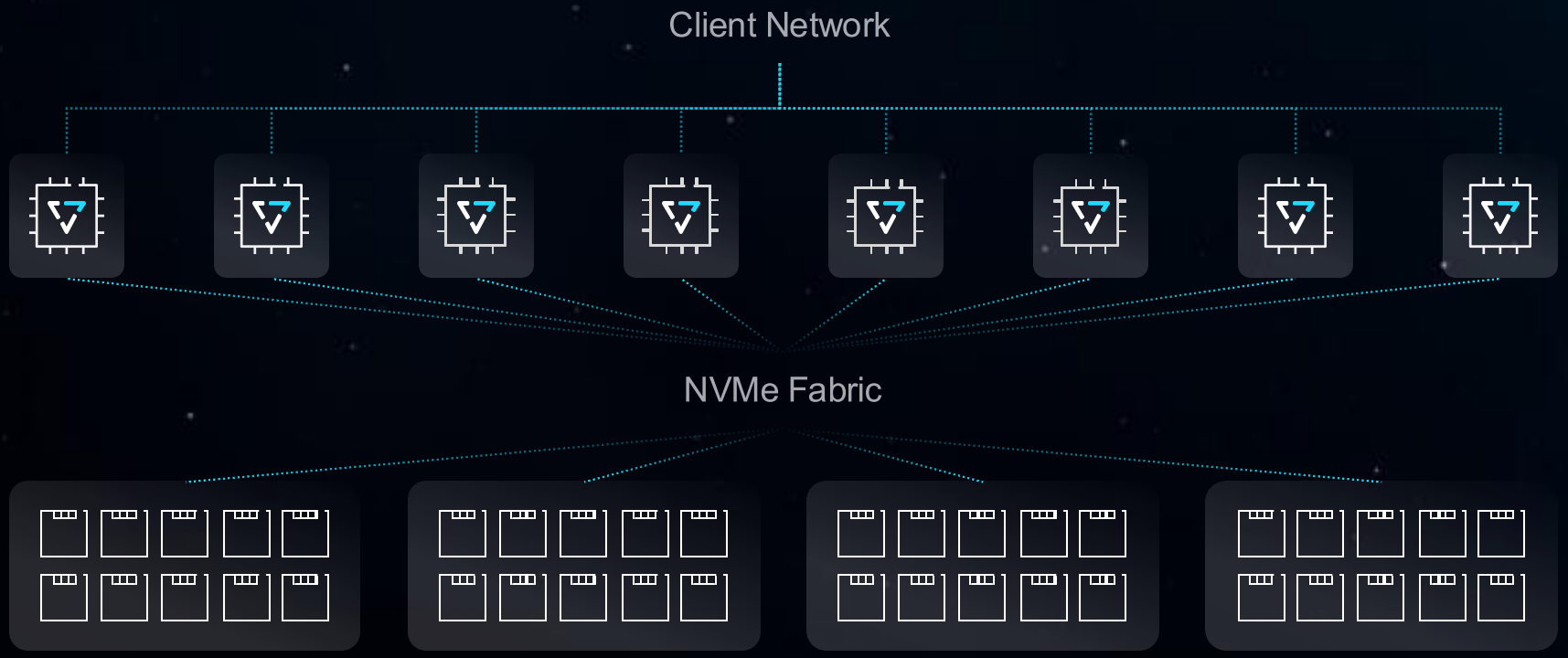
NAS

Legacy Architecture



East-west cluster traffic slows transactions as clusters scale. Not designed for NVMe.

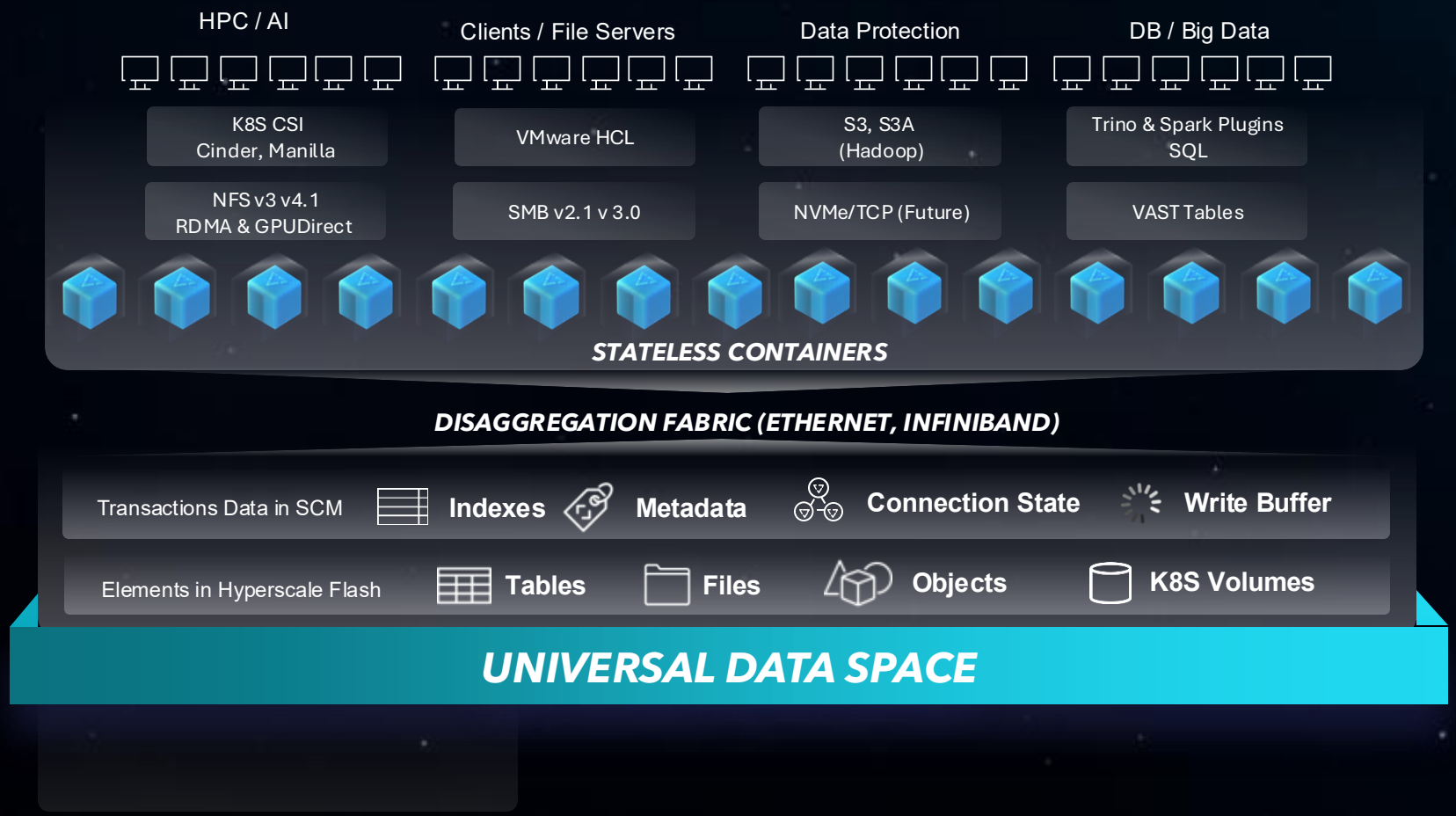
Parallel Architecture Enhancements



Disaggregated, Shared-Everything (DASE) Architecture

Data Platform

Disaggregated, Shared Everything (DASE)

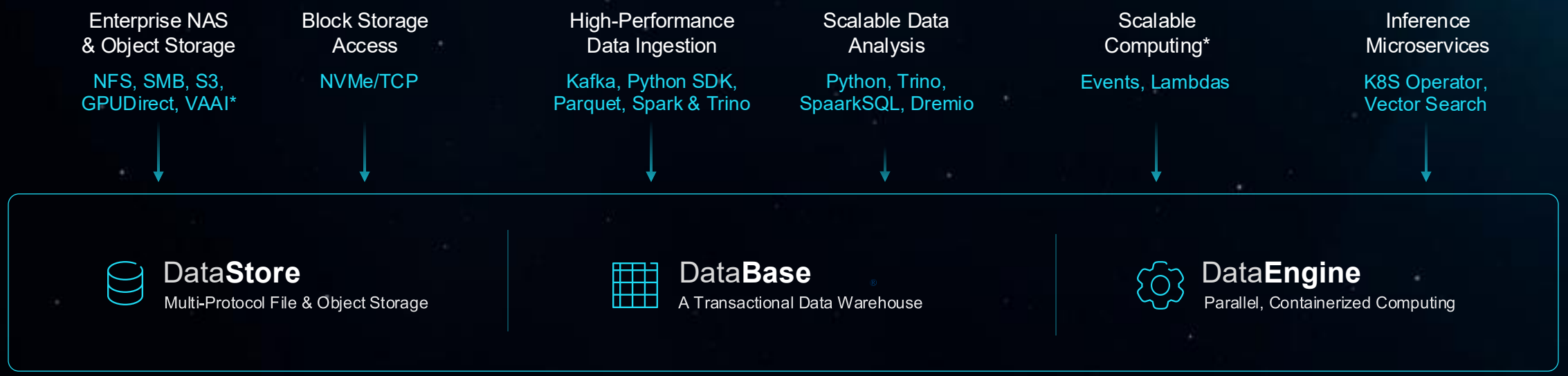


Asymmetric Architecture

Scalability and Resilience

Data Platform

Combined Storage allows for Cost savings across many Platforms

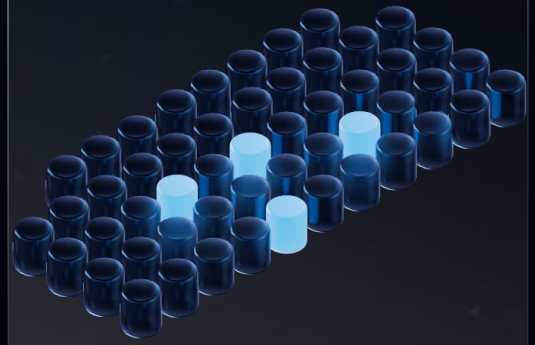


* (1H'25)

Capacity Cost Savings

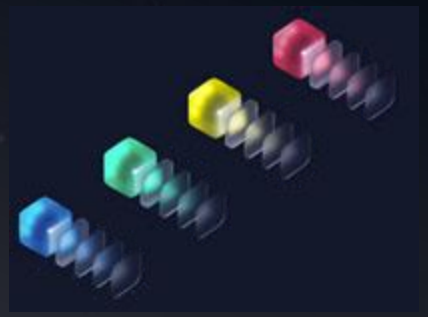
Write Shaping with Large scale Write Buffers

Low Cost QLC



Locally-Decodable Codes:

EC provides low overhead data protection



Neres Neighbour Algorithm Data Reduction

Global De-dupe with fine grained resolution

Significant Data Reduction Ratio DRR Achieved through use of Large Write Buffer

Averaging 3:1 DRR

Reduces Cost of Flash by 3X

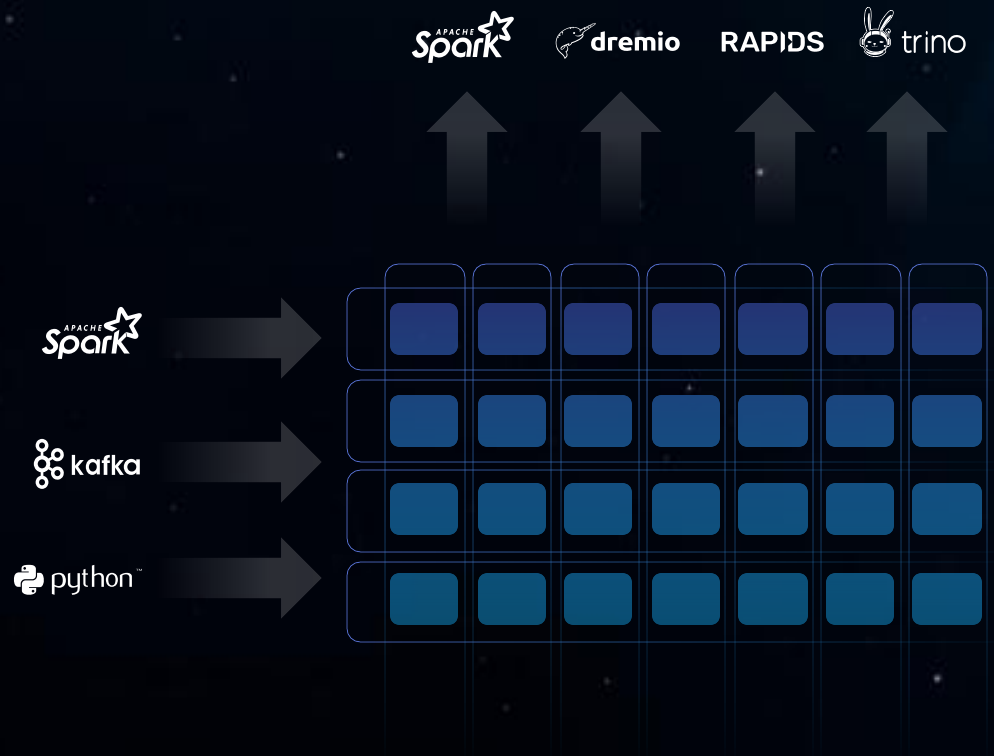
Architecture allows for Large Scale Data Lake

Parallel Analytics

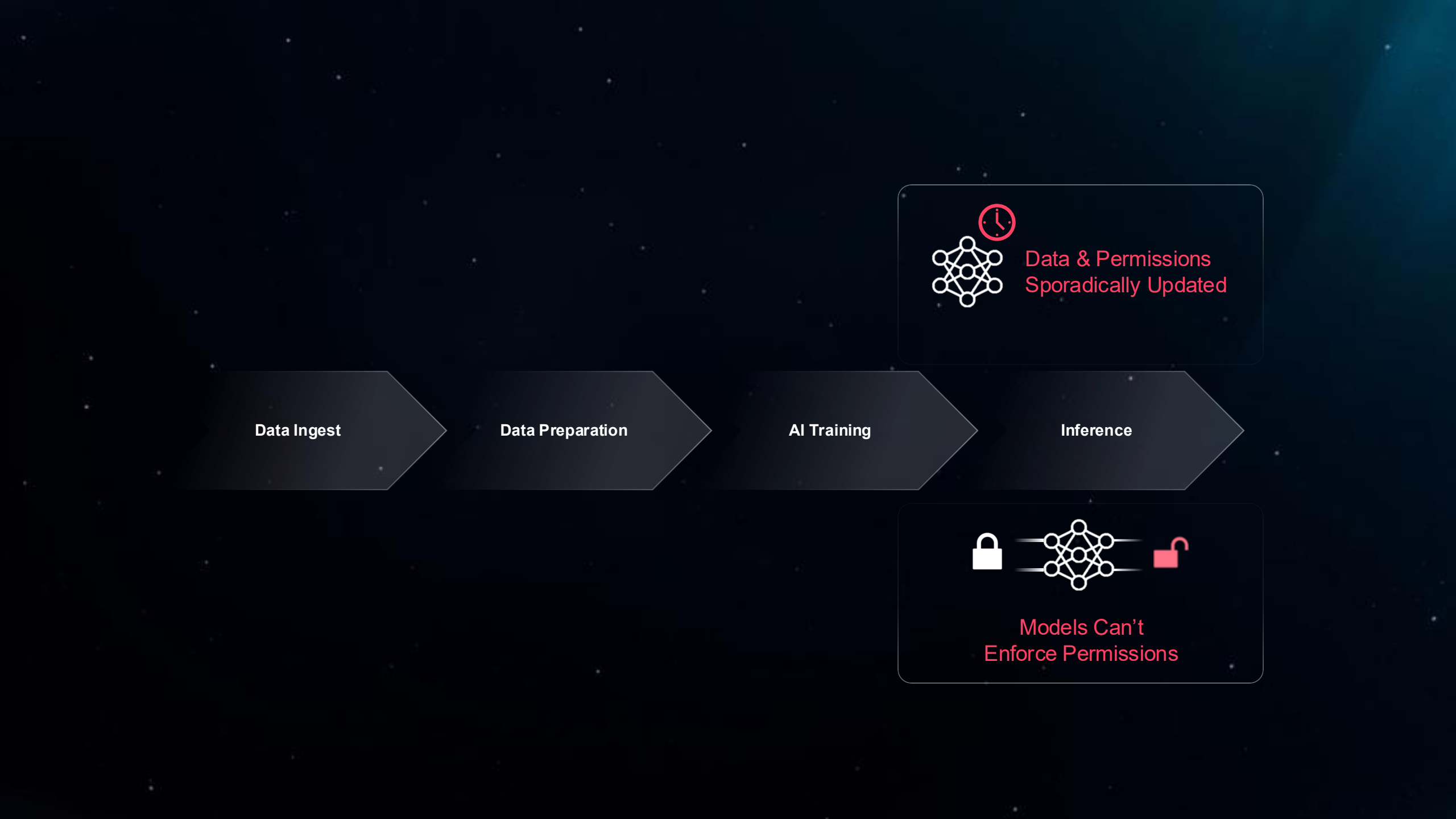
- NVMe-Optimized Columnar Structures
- >20x faster than Parquet for selective operations
- Query via Arrow, Python or Query engine (written in C++)

Parallel Ingest

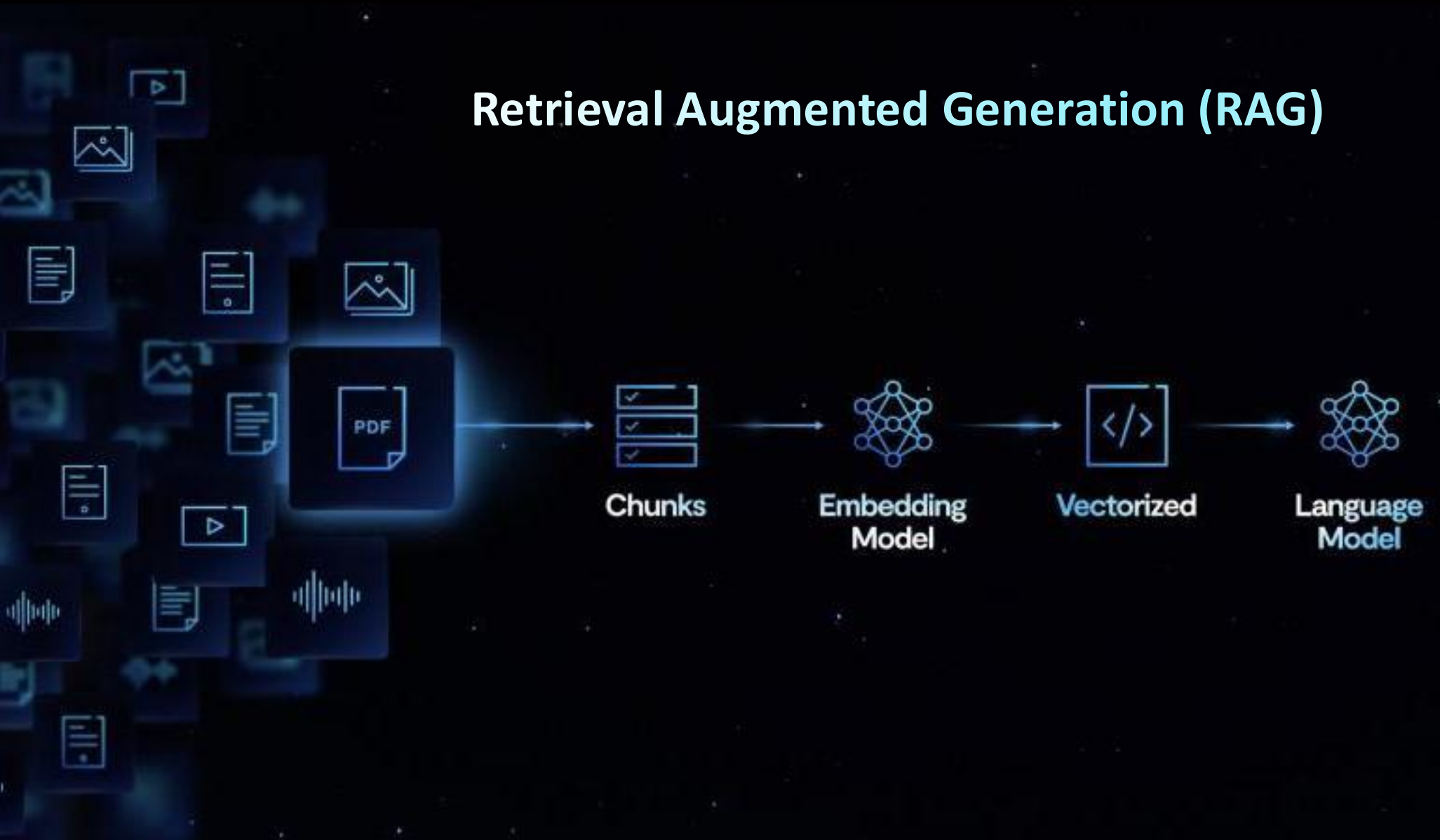
- Ingests records in row form to a deep write buffer
- 7M batched inserts per second, **per server**
- DASE enables linear transactions, 100Ms at scale

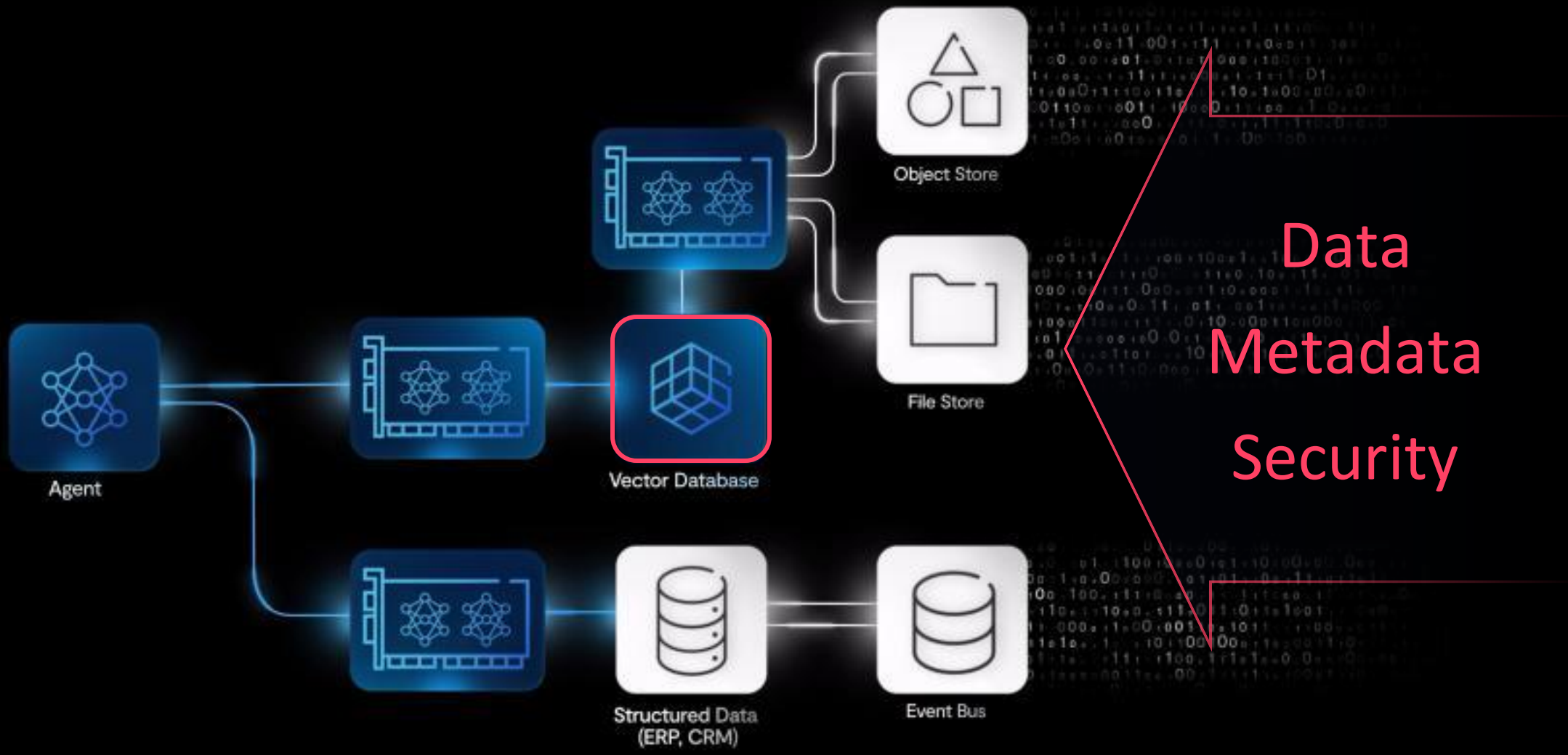


Insight Engine: Preview


Data & Permissions
Sporadically Updated
Data Ingest
Data Preparation
AI Training
Inference
Models Can’t
Enforce Permissions

Retrieval Augmented Generation (RAG)




Object Store
File Store
Agent
Vector Database
Structured Data
(ERP, CRM)
Event Bus
Data
Metadata
Security

1

Vector **Space** Scalability

Trillions of vectors.

Not breaking the (memory) bank.

2

Vector **Search** Scalability

Agentic AI needs GPU-speed.

Search time must be GPU time.

3

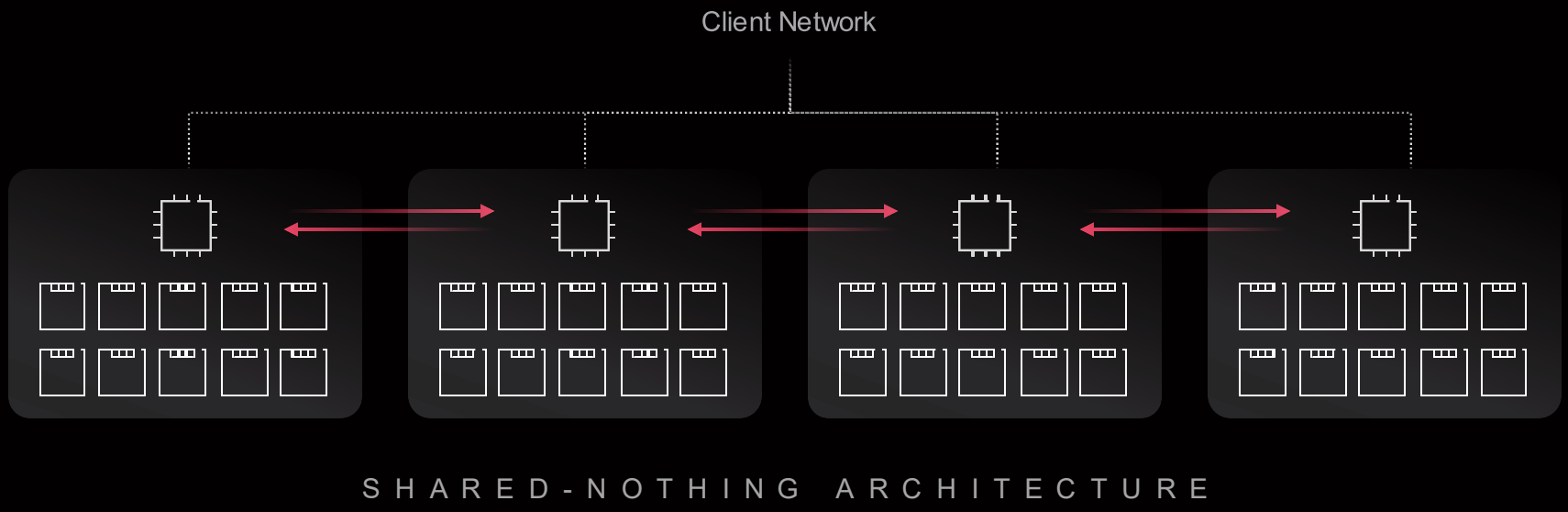
Vector **Insert** Scalability

Data is always flowing.

Agentic AI needs real-time info.

Why are Classical Systems not good for Real-Time Inference?

Legacy Database & Storage Architectures…



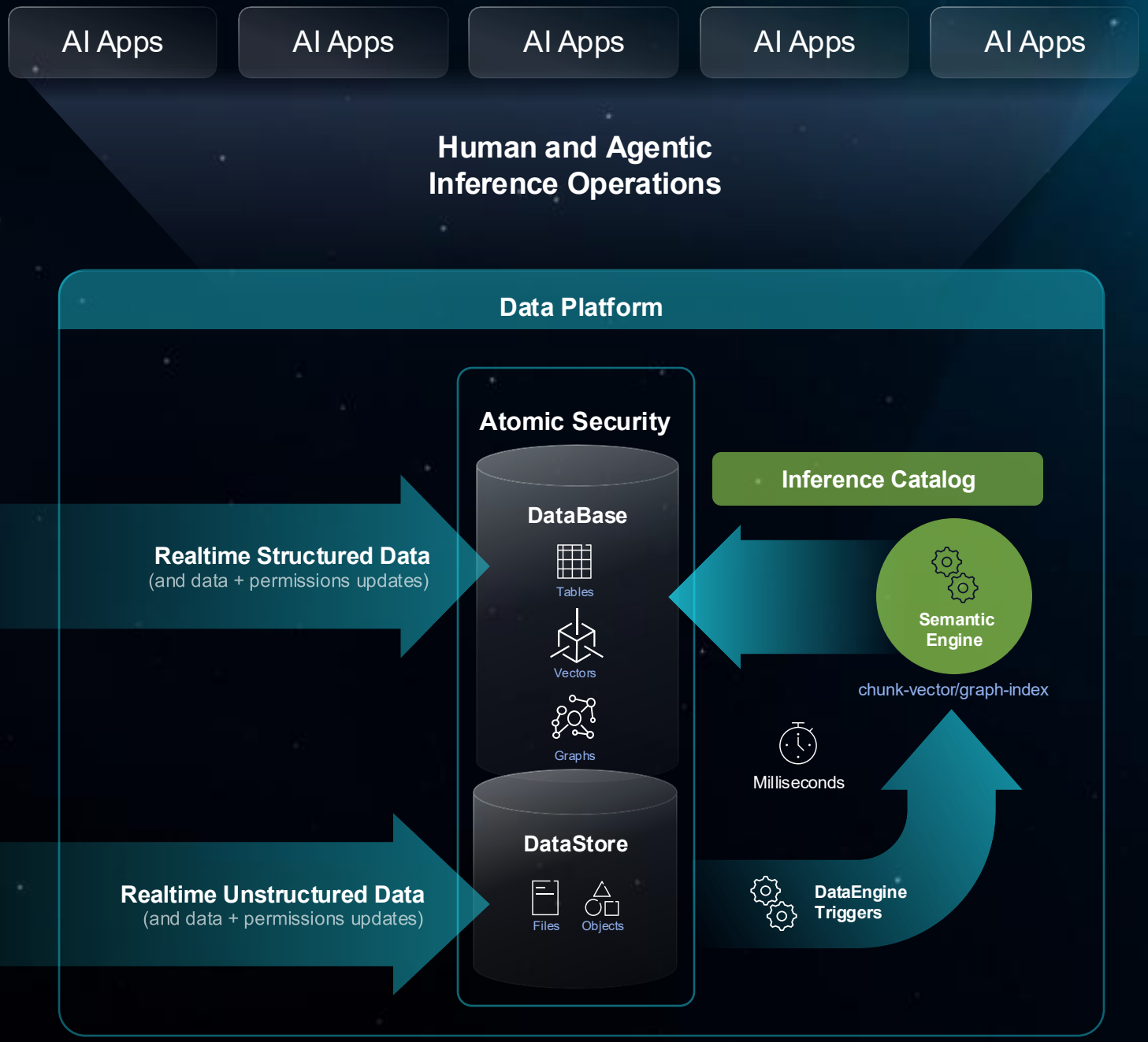
East-West Cluster Traffic Slows Databases, File Systems & Even Event Busses @ Scale

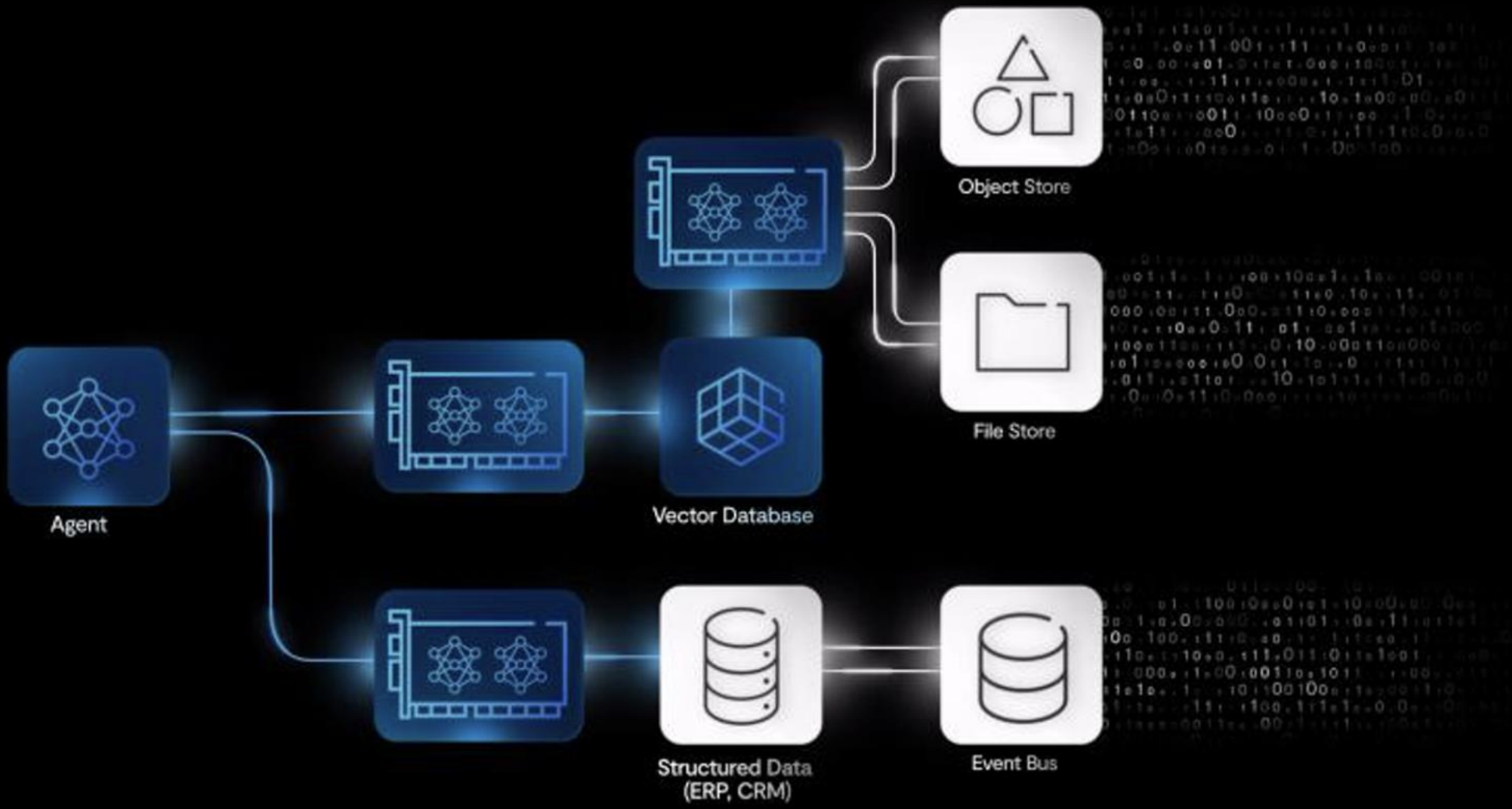
Transactions & Deep Search Suffer

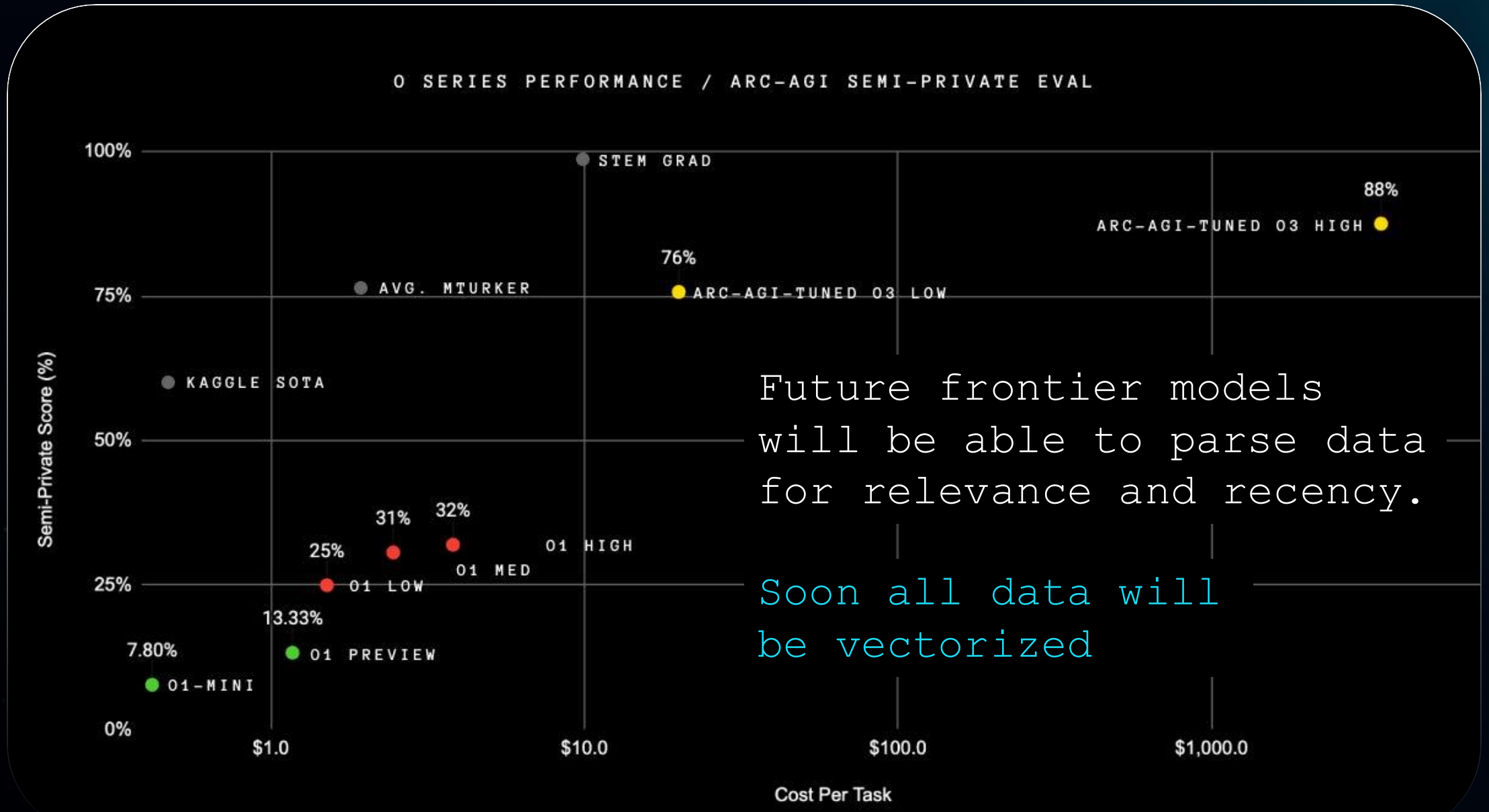
Automatic

Inference

Real-Time Data Flow




Object Store
File Store
Vector Database
Agent
Structured Data
(ERP, CRM)
Event Bus



Future frontier models will be able to parse data for relevance and recency.

Soon all data will be vectorized

Global and Unified: Policy-Based Vector Access

```
{
  "Statement":[
     {
        "Effect": "RowColumnSecurity",
        "Resource":  ["my_db/my_schema/*"],
         ...
        "RowFilters": [{"query_engine":"*"),
                        "row_filter": (col_groups IN (current_groups()) AND
                                       current_user IN col_users)
                       }]
        ...
        "RowFilterByPathAuth": "True",
     },
     ...
}
```

ROW_ID	vastdb_path_auth	users	groups	…	vector
1234	/path/to/original/file1	*	Marketing, Sales	…	[1,2,3,4]
1235	[/path/to/original/file2, /path/to/original/file4]	Amy, John	HR, Finance	…	[5,6,7,8]

Vast Insight Engine Value

Ultimately Simple

everything prepackaged

VAST's InsightEngine will have everything needed to capture, index and make context available for retrieval.

Recent & Relevant Data

Linearly transactional

VAST's InsightEngine is the only solution to support real-time workflows, with real-time data & security insights.

Massively-Scalable

exabyte-scale

VAST's InsightEngine sits on the VAST DASE architecture to enable data & vectors at any scale, with fast search.

Enterprise Grade AI

security & data mgmt. built-in

VAST's InsightEngine is built upon a foundation that features robust unified data security, audit & governance.

In the future: Pre-Packaged Workflows

VAST's InsightEngine will be configured to automatically process documents, videos, genomics, etc.

BROWN + VAST

Campus-Wide Enterprise File Services At The Speed of Brown Research

Storage for a Diverse Set of Requirements Across Campus
Brown consolidated their enterprise and HPC environment into one multi-tenant system

Quality of Service to Protect Against Noisy Neighbors
CNode and network-based isolation enables HPC jobs to run in parallel with enterprise workloads

Robust Data Governance to Enable PII Data Storage
Audit logs enable Brown to store HIPAA-compliant and regulated data

Performance Goes to HPC, Capacity Goes to Enterprise
Amortize everything across multiple users, thanks to quality of service

VAST Brown customer interview here.

PIXAR + VAST

The Data Foundation for the Next Generation of Film Animation

Accelerating Next-Generation Volumetric Rendering

Pixar moved to VAST when new applications compelled the need for all-flash rendering systems

Always-On Infrastructure with Enterprise Data Services

High-performance snapshots, database services for user and admin auditing

Data Reduction Efficiency, Even For Media Formats

1.5:1 data reduction & 2.7% erasure coding overhead enable a 60% footprint reduction vs Isilon

A Spectrum of Studio Use Cases

Rendering, AI research on VAST Data sets, used for film mastering to replace SAN storage

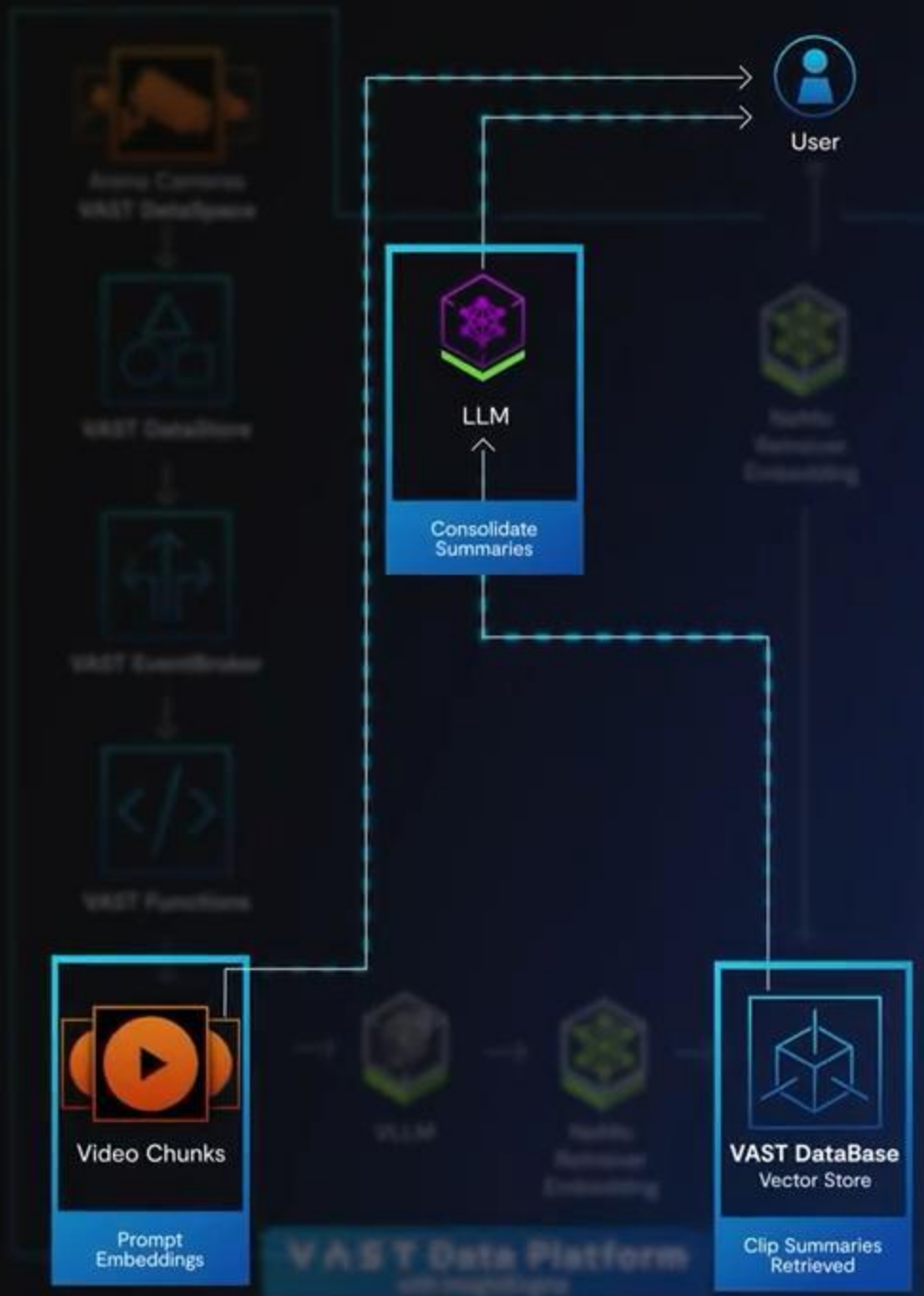
Modernizing and Simplifying Digital Asset Management

All Archive Content is Online in Real-Time
Elevated 10s Of PBs Of Data From LTO Libraries Into VAST

Scalable Infrastructure to Support NHL's Workflows
All Of NHL's Content Production Pipeline Runs On VAST

Data Intelligence, Built-In
VAST Data Catalog Provides Built-in Metadata Management

Video Search: Finding Hockey Clips



POWERED BY VAST

Footage provided courtesy of National Hockey League

TACC + VAST

Simple, Accelerated Data Infrastructure for Exascale Academic Computing

All-Flash, Embarrassingly-Parallel File Services
DASE is even more parallel than parallel file systems; all-flash makes even badly-formed I/O OK

Always-On Infrastructure during Expansions and Rolling Upgrades
Changes expectation of how scalable systems should work, snapshots make for easy backups

Data Reduction is Game Changing
At 2:1 Data Reduction – TACC can afford to consolidate and collapse storage tiers

VAST Data Catalog Enables Simple User Management
No need for complex tree walking, data purge policies are enabled by VAST's fast + easy catalog

"VAST is smooth and graceful for what we use to call bad I/O patterns, which we now just call I/O patterns."

Dan Stanzione – Executive Director