SNIA DEVELOPER CONFERENCE

SDC 25

By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Towards Unified Knowledge Platforms: Evolving Storage Systems for Generative and Agentic AI

Annmary Justine K – Senior Research Engineer
Srikant Varadan – Distinguished Technologist

Evolution of Intelligence in Storage Systems

Early AI Systems

2000s

- **Passive Storage (Pre 2000s)**
  - Function: Passive data repositories
  - Block/File/Object storage
  - No awareness of data content

Big Data

2000s–2010s

- **Smart Storage**
  - Function: Enhanced data management
  - Metadata tagging
  - Analytics integration

Deep Learning, ML

2010s–2020s

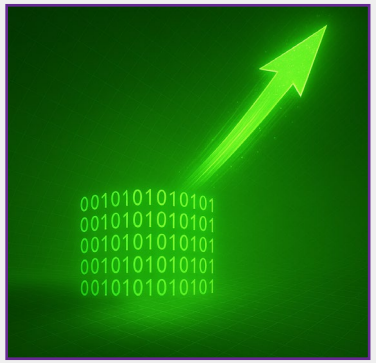
- **Intelligent Storage**
  - Function: Context-aware and policy-driven
  - AI/ML for predictive analytics
  - Automated tiering and optimization
  - Integration with data lakes and warehouses

Generative AI

Present

- **Insight Engines for AI Agents**
  - Function: Active data interpretation and delivery
  - Semantic search
  - Natural language interfaces
  - Real-time insights for AI agents
  - Autonomous decision-making, generative AI, multi-modal agents

Catalysts for Storage Transformation



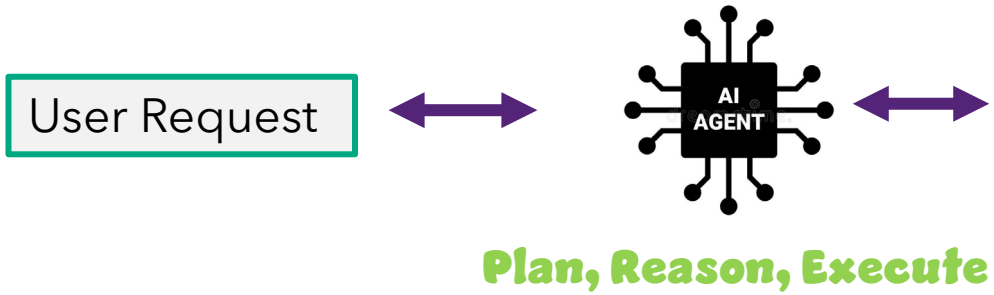
Advancements in

- Generative AI – Transformer Architecture, Reasoning Models
- Hardware and Compute
- Software Stack – Frameworks for Agentic and Generative AI workflows

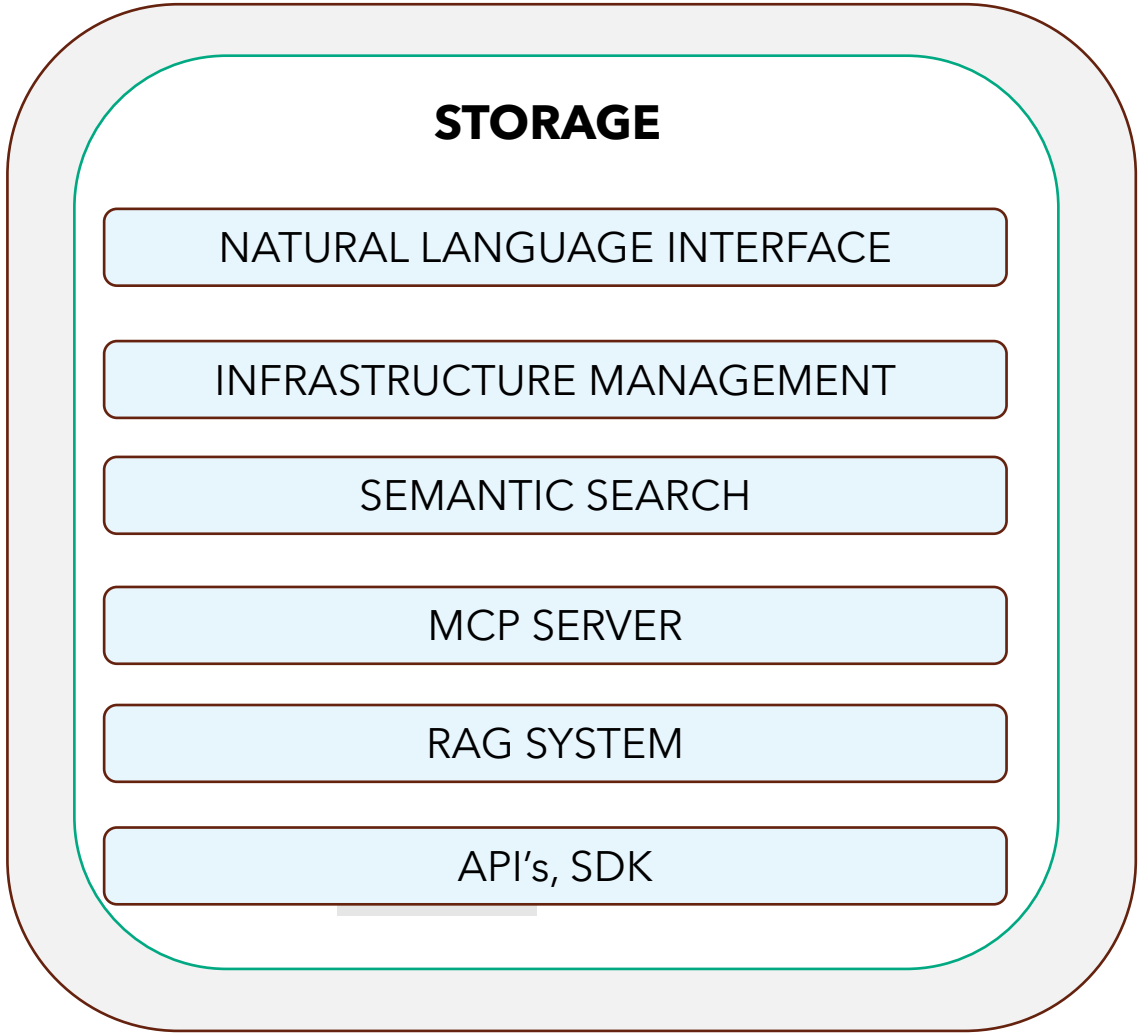
AI Driven Innovations - Storage Centric

Data Centric - Agentic Storage Systems

Automated, intelligent orchestration of workflows
Reduces complexity for end users
Elevates storage to Knowledge platforms



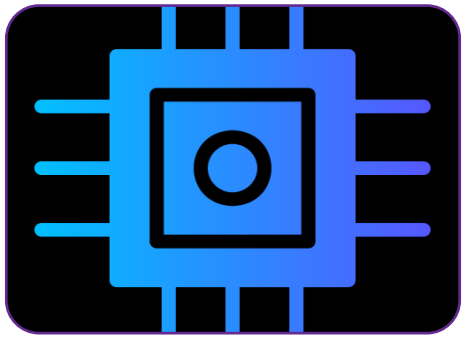
Storage administration, Read and write data,
Search data and Metadata

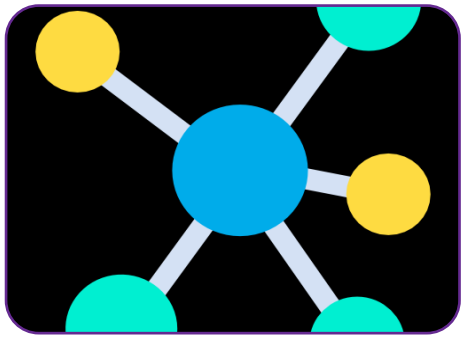


Unlocking Data Value - Agentic Storage Systems

- Retrieval-Augmented Generation – LLM provides context specific responses . Context provided from the data stored in the storage
- Natural Language Interfaces - Enables users to query the metadata using natural language
- Semantic Search – Search not limited by keyword matching, but includes semantic search
- Agentic infrastructure management - Enables management of infrastructure through agents. Eg – create bucket, create policy etc
- Agentic Workflows, MCP Servers, SDK's - Programmatic access to metadata.

Agentic AI workload Characteristics





Storage

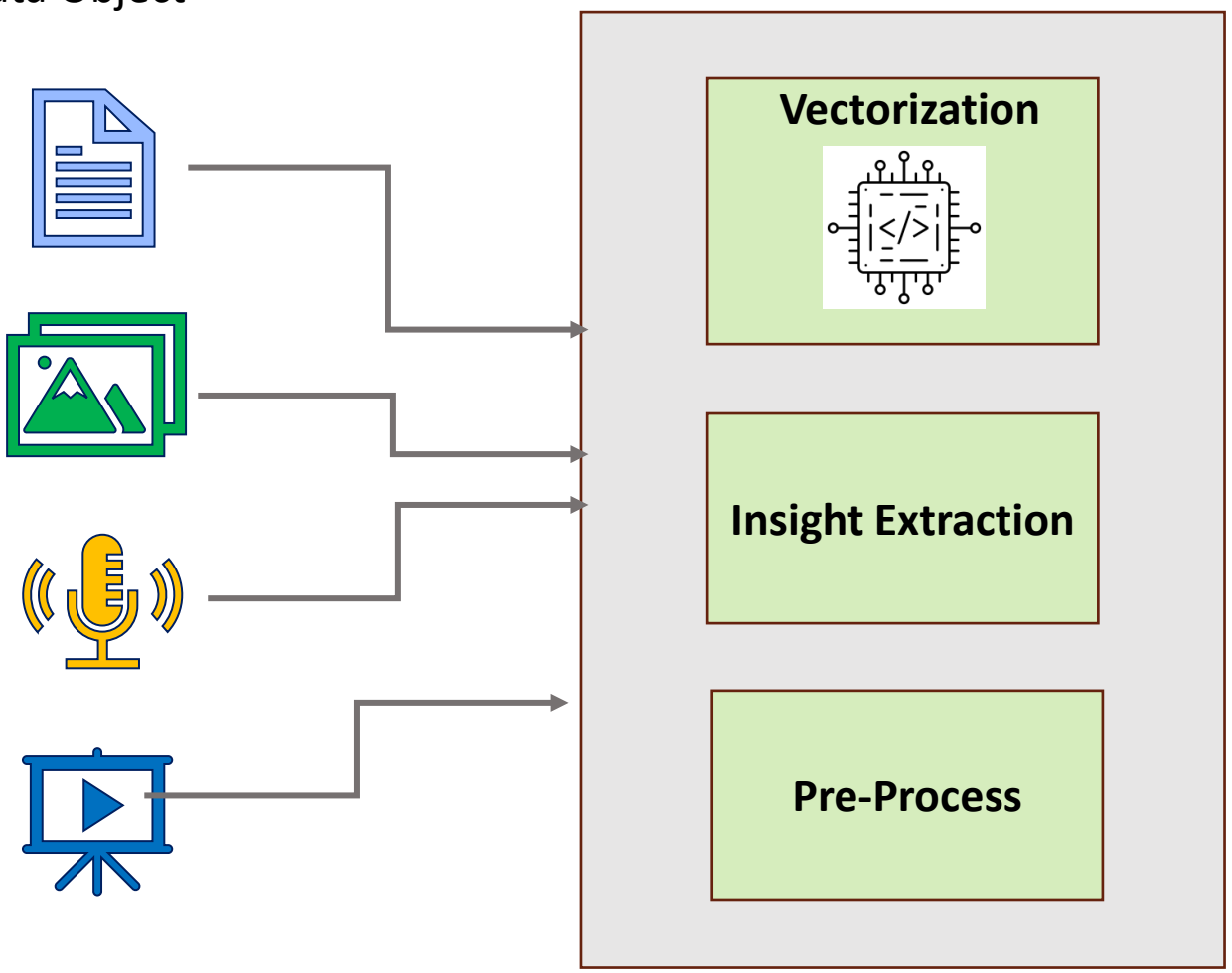
Compute

Networking

Data Transformations leading to Enriched data

Governed, Structured, Contextualized, and Accessible data

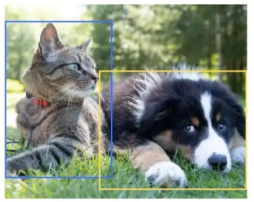
Data Object



Vector Embeddings

0.06	1.12	0..89	1.23	34.2	11.23	67.4

Enriched Metadata

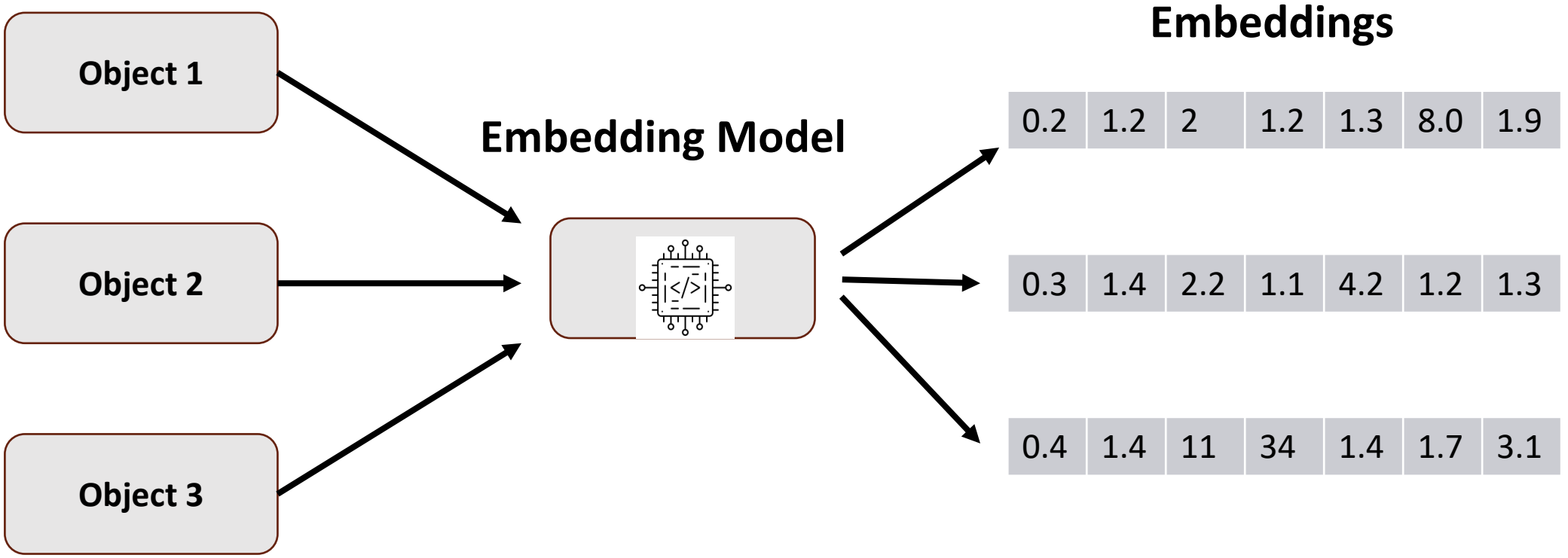


ImageId:a12be56, tags : Cat, Dog

Structured Data

Storage Considerations for Agentic AI Workloads

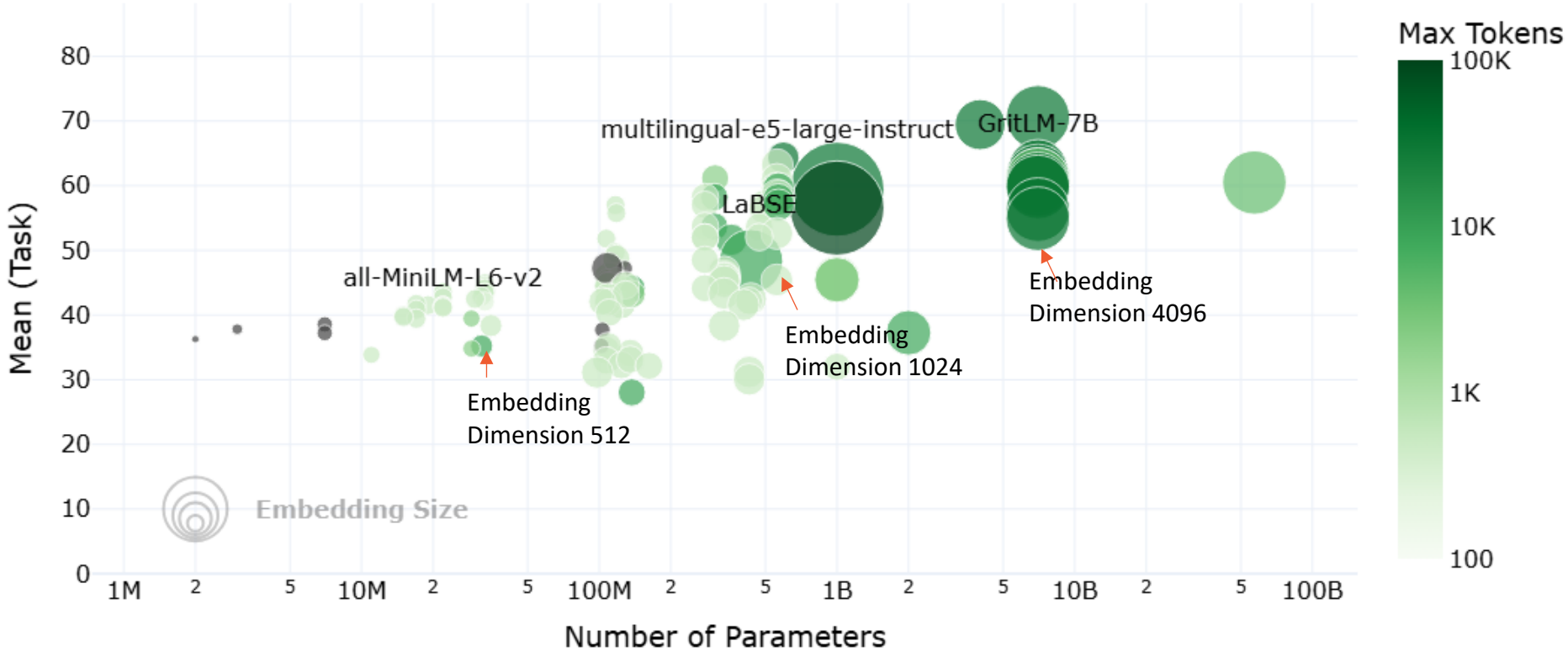
Embeddings and Embedding Models



Embeddings Power - RAG, Semantic Search

Storage Considerations for Agentic AI Workloads

Embedding Model Sizes



Source MTEB leaderboard (Sep 8,2025)

Storage Considerations for Agentic AI Workloads

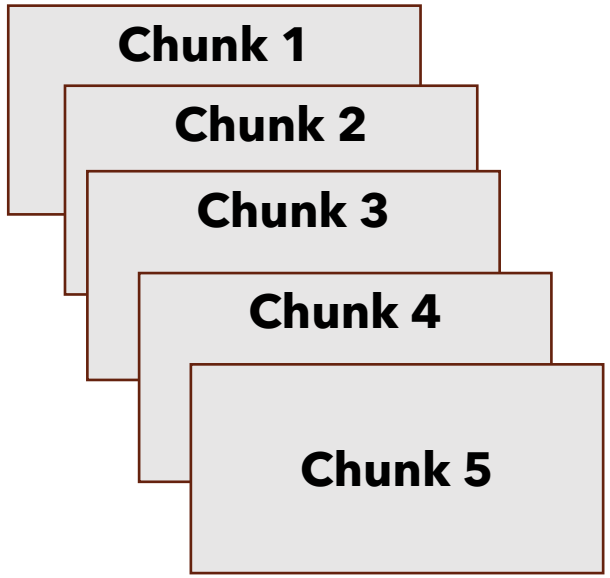
Why Embedding Dimensions Matters ?

*Storage Bloat ~**10 - 20X**

or lower-precision floating-point types (like float16) instead of float32, leverage database compression, and store data efficiently in specialized vector databases or cloud services. Embeddings are fixed-length arrays of floating-point numbers efficiently in specialized vector databases or cloud services. Embeddings are fixed-length arrays of floating-point numbers

1
2
3

10 KB



5 Chunks (each 500 tokens)

0.2	1.2	2	1.2	1.3	8.0	1.9
0.3	1.4	2.2	1.1	4.2	1.2	1.3
0.4	1.4	11	34	1.4	1.7	3.1
2.3	1.2	4.5	2.2	1.5	1.3	1.1
3.4	1.4	2.2	1.3	1.3	2.4	1.1

= 5* 4096 Dimensions * 4 Bytes * 1.5 ~ 120 KB

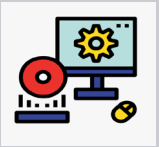
*Size depends on – Embedding dimension, precision

Compute Considerations for Agentic AI Workloads

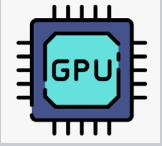
Why Embedding Inference matters ?

Compute Intensive

Typical throughput on single on prem *GPU – 30,000 – 120,000 **Request /sec for 8B models .



Performance depends on the hardware and software stack



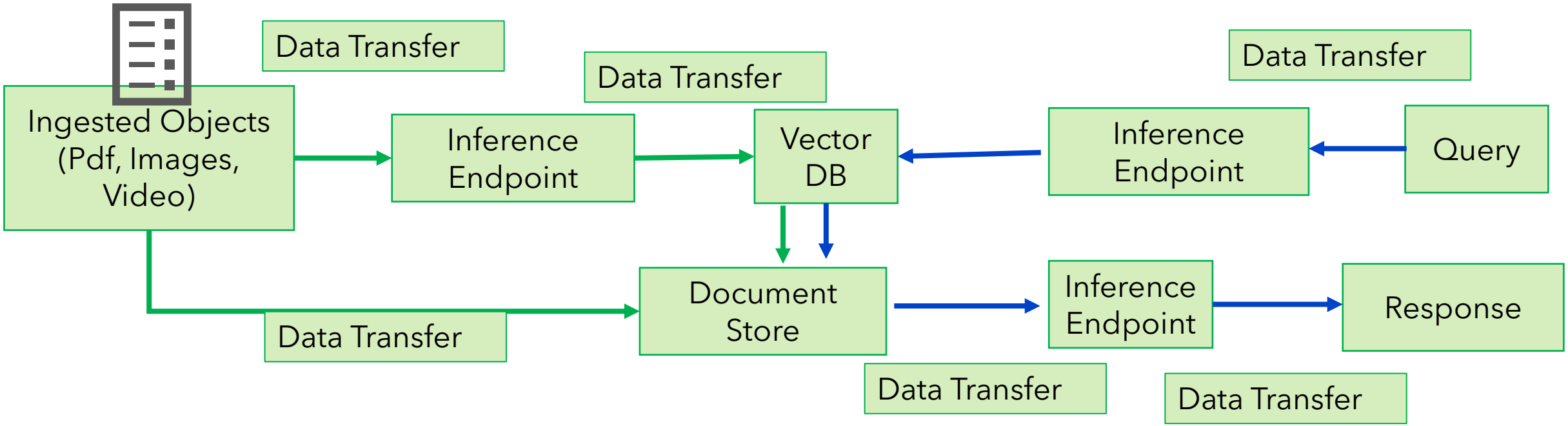
GPU vs CPU

*H100

** 500 Tokens per Request

Network Considerations for Agentic AI Workloads

East –West Traffic



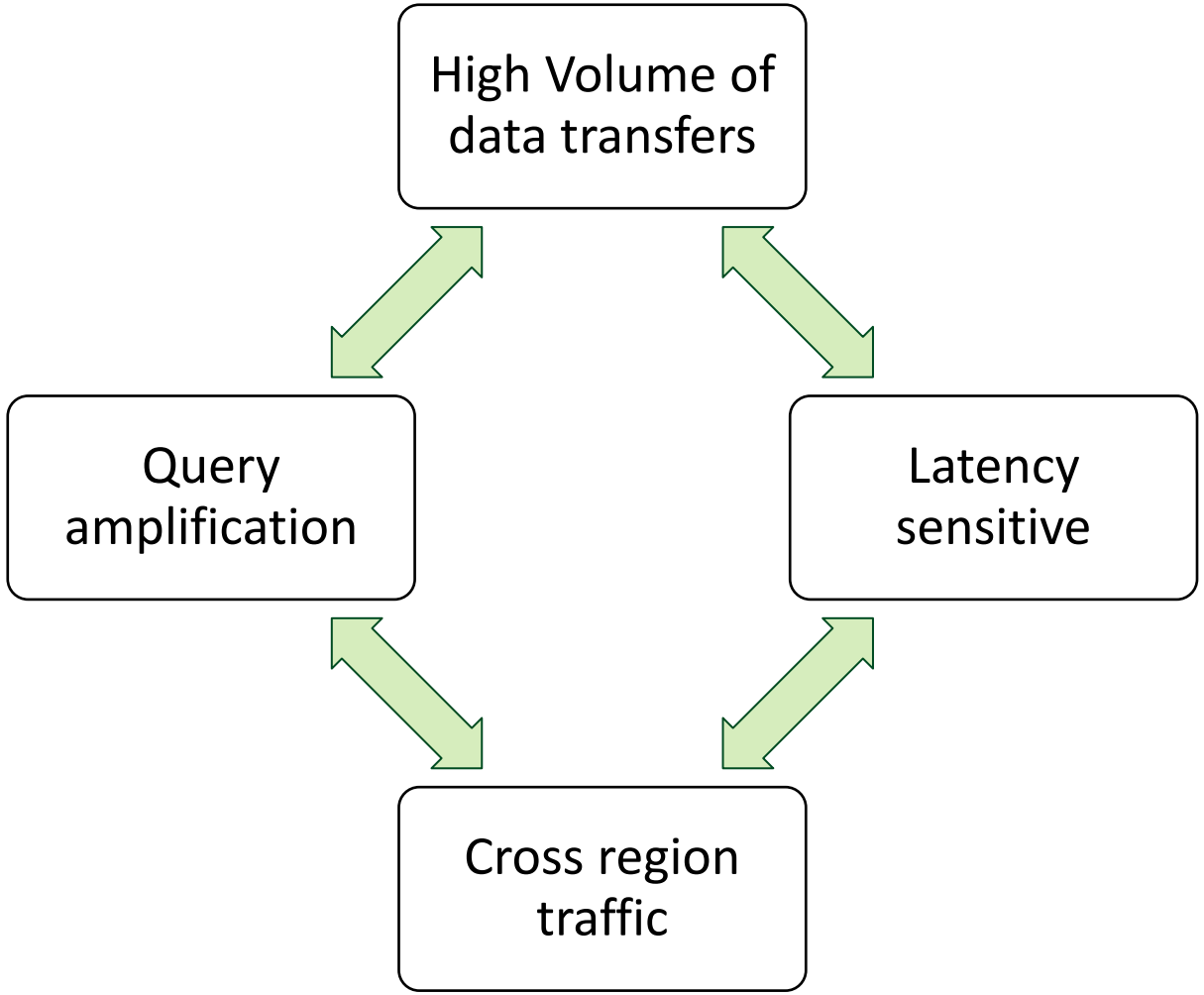
High Bandwidth

Latency sensitive

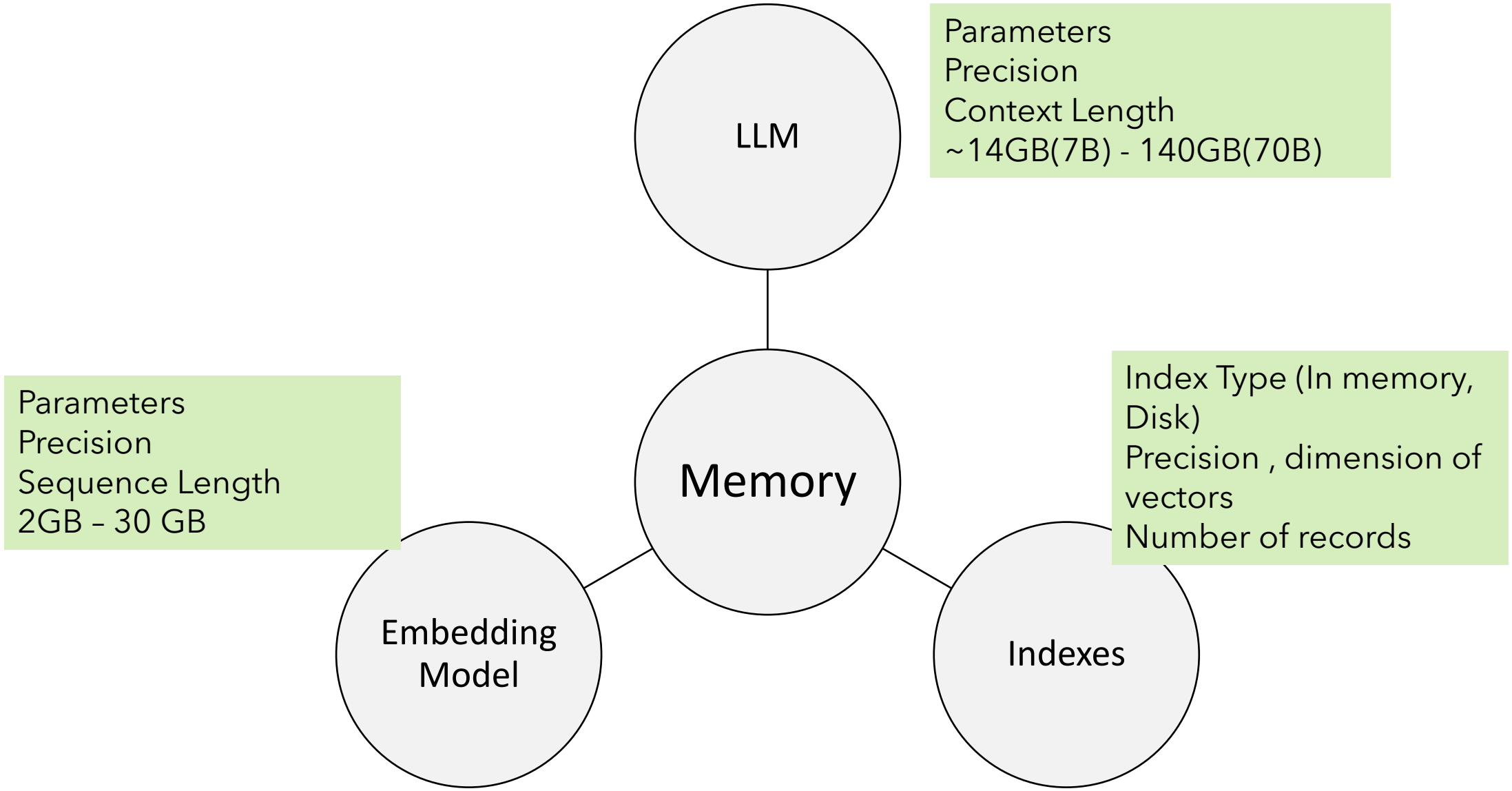
Ingest

Query

Network Considerations for Agentic AI Workloads



Memory Considerations for Agentic AI Workloads



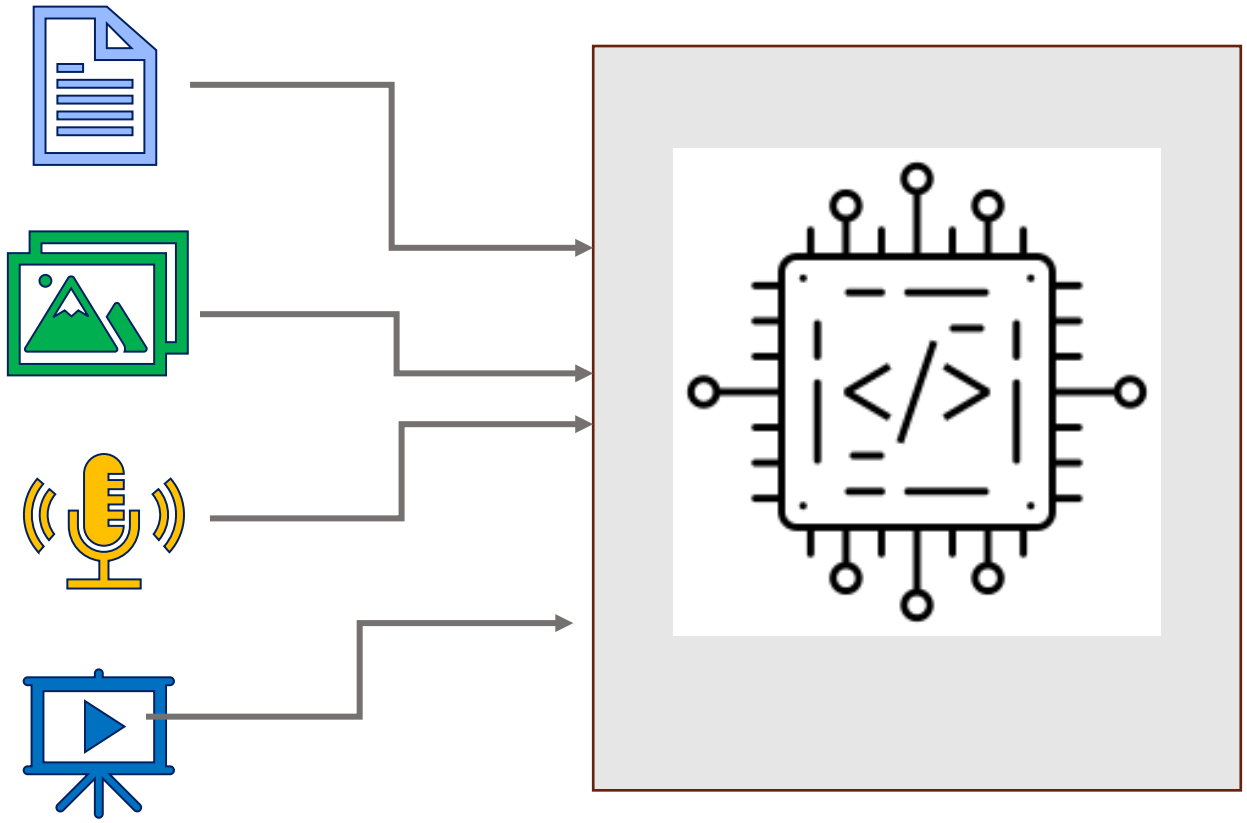
Data Challenges

- Multimodal data
- Realtime inference
- Strong data governance

Multimodal data

Data Object

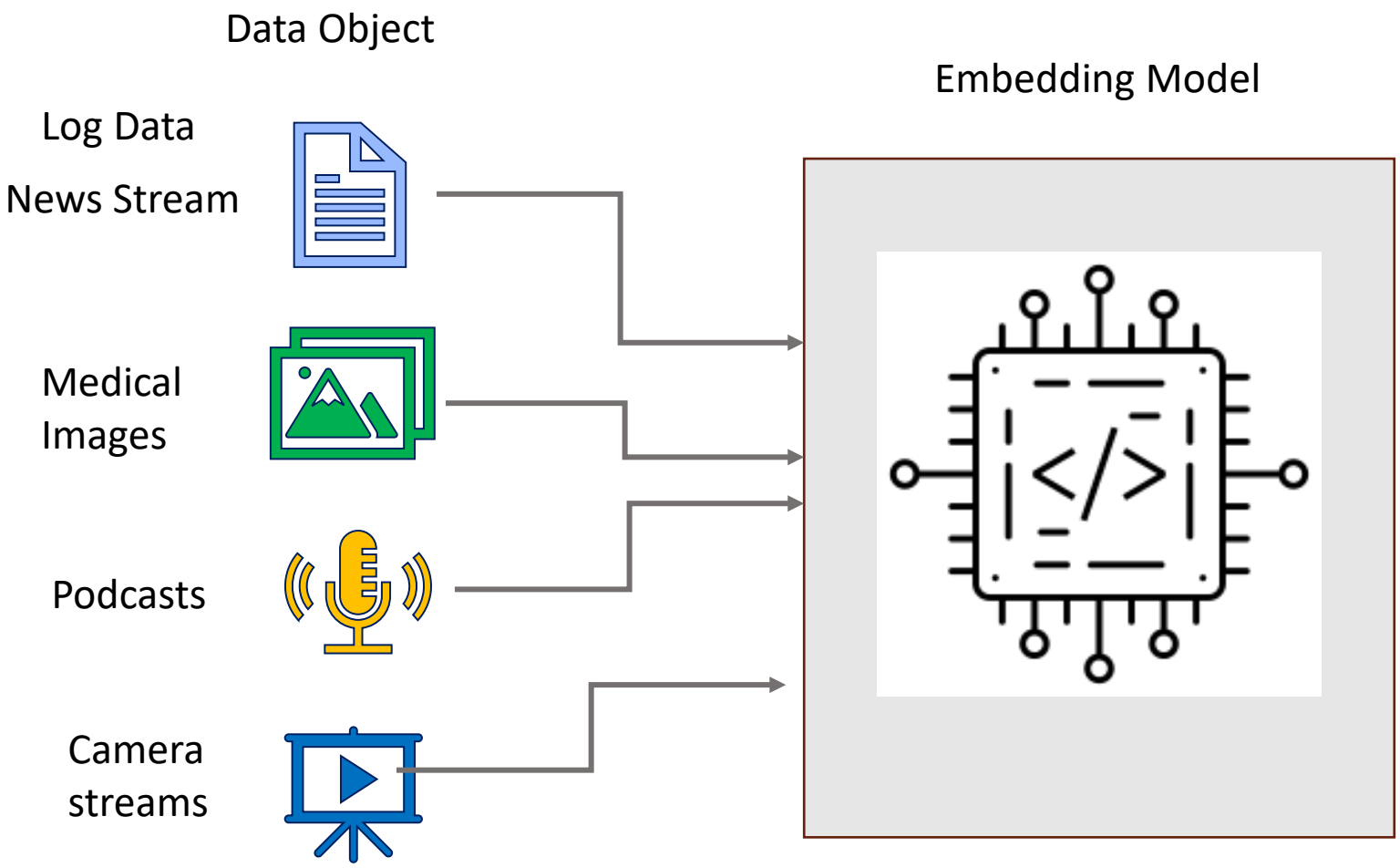
Embedding Model



Data as Vector Embeddings

0.06	1.12	0..89	1.23	34.2	11.23	67.4
0.25	3.47	7.93	1.78	9.82	11	23.12
0.61	1.45	0.77	1.09	2.45	3.78	23.56

Realtime data

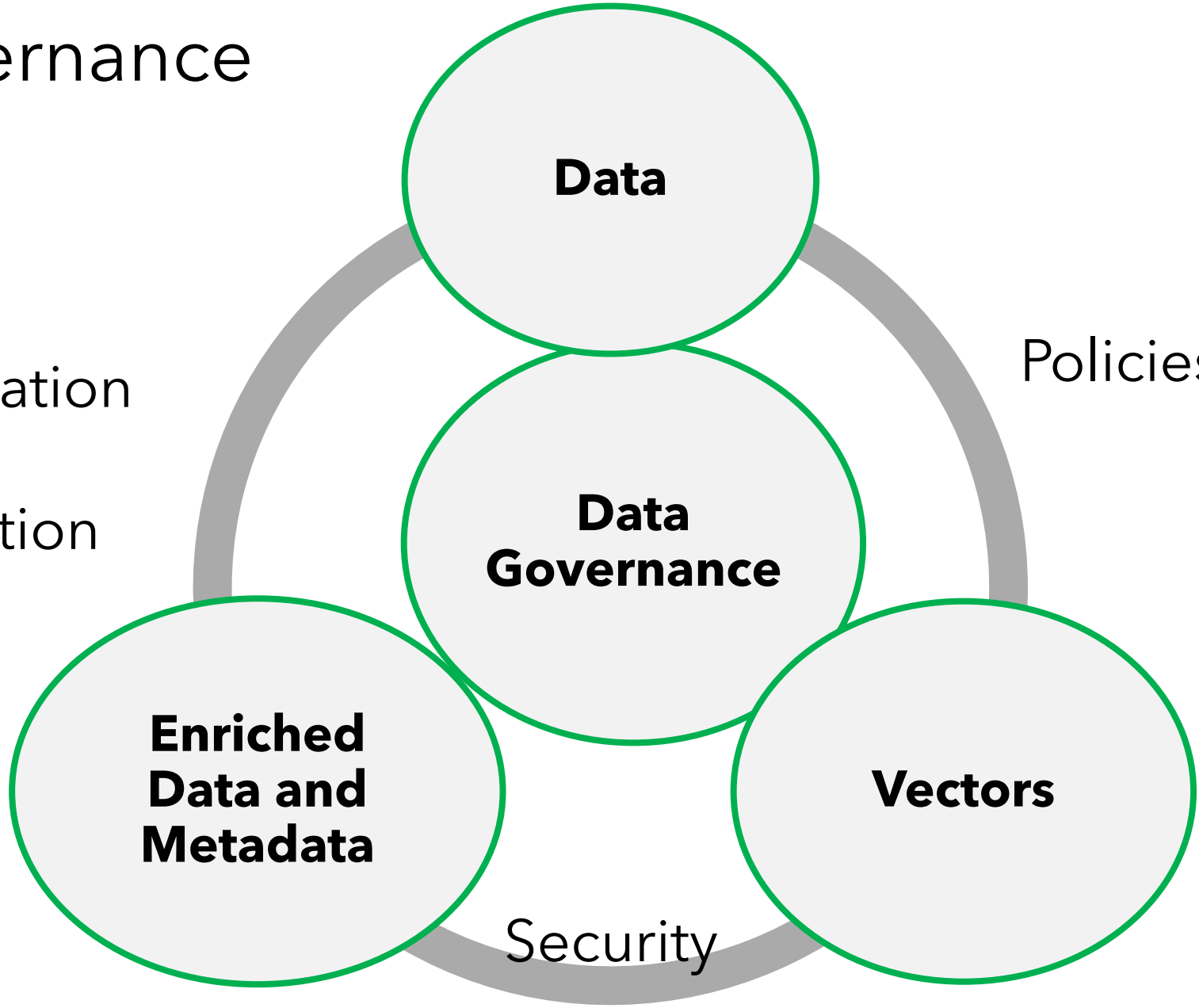


Data as Vector Embeddings

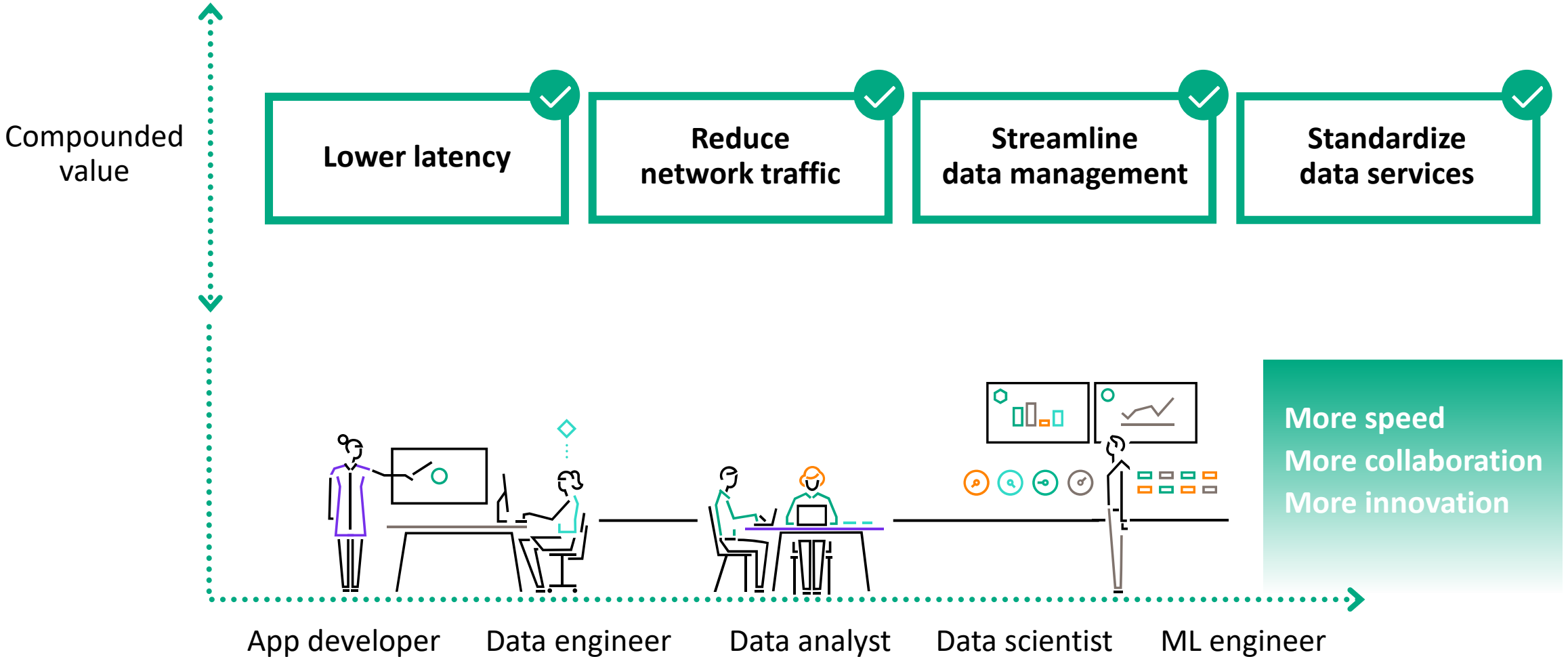
0.06	1.12	0..89	1.23	34.2	11.23	67.4
0.25	3.47	7.93	1.78	9.82	11	23.12
0.61	1.45	0.77	1.09	2.45	3.78	23.56

Data Governance

Authentication
and
Authorization



Storage level data intelligence enables instant delivery of RAG based LLMs



HPE Alletra Storage MP X10000

Software-defined for hybrid clouds

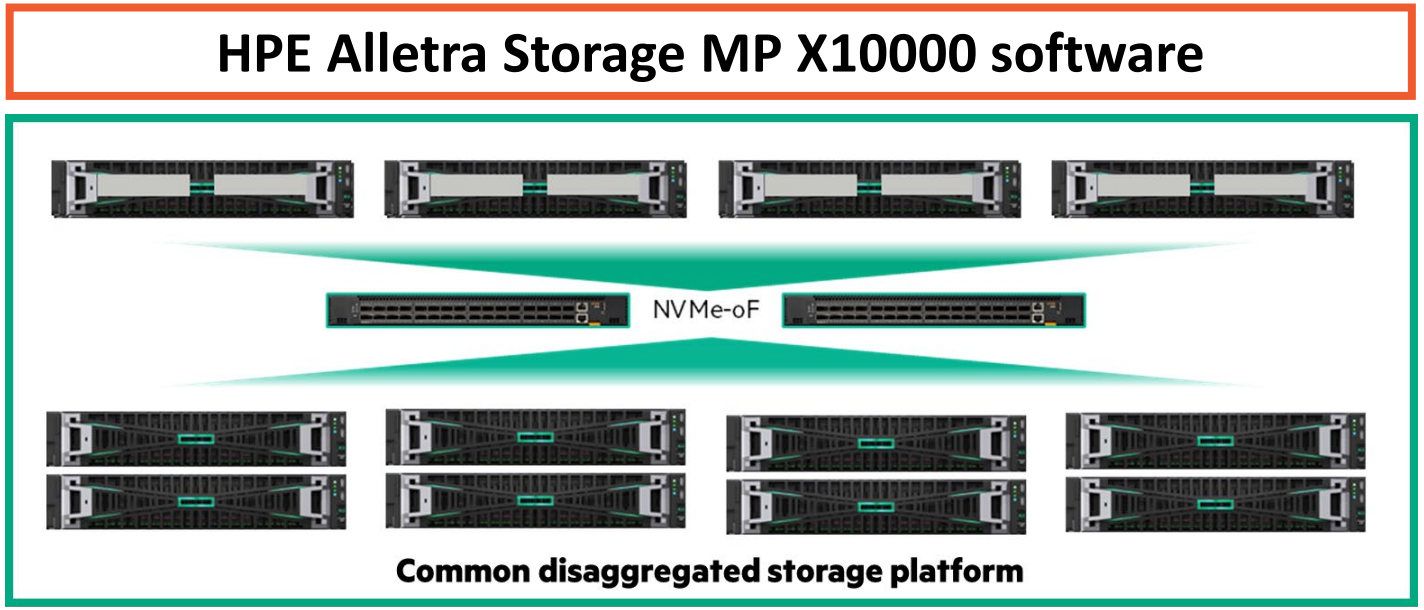
- Containerized architecture with Kubernetes orchestration

Reliable industry standard data access

- Native S3 API
- Fault tolerant design with no single points of failure

Performance and efficiency at AI scale

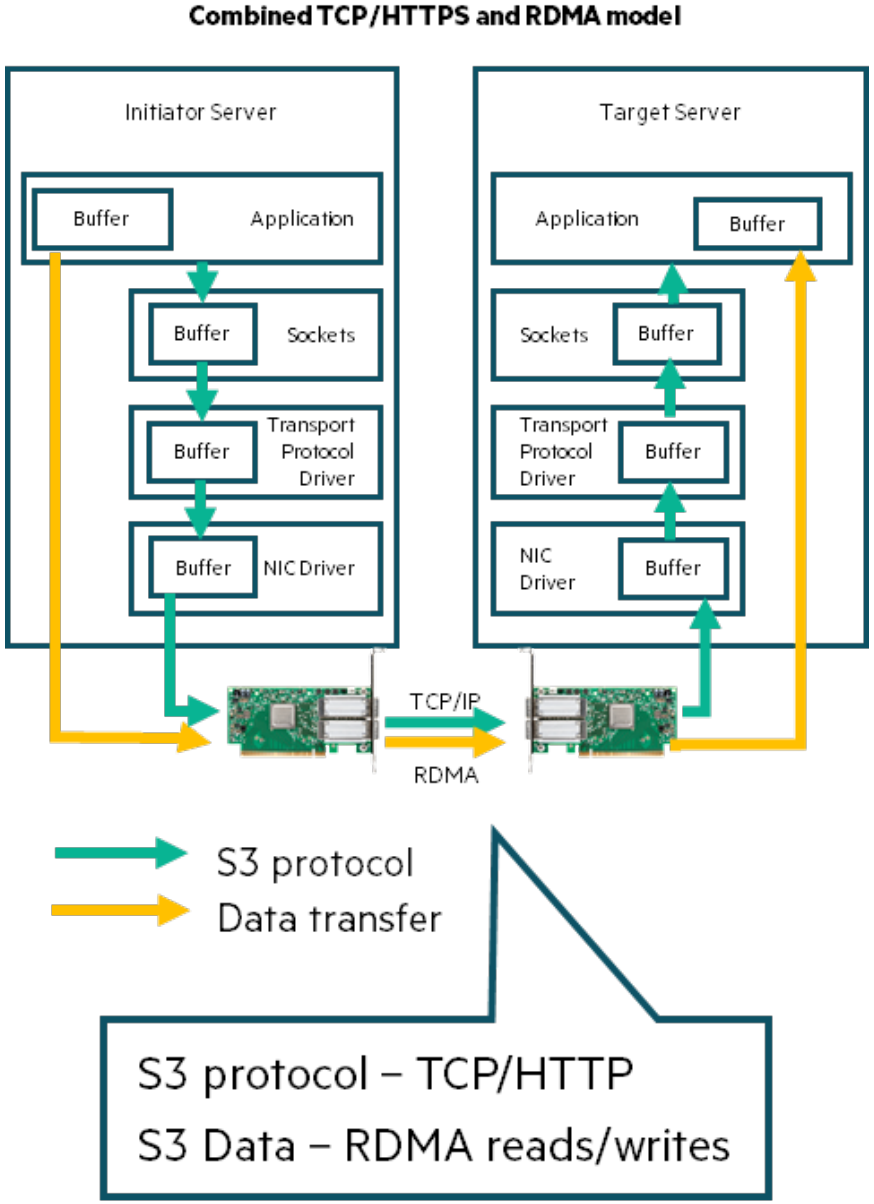
- Key value store
- Optimized for TLC and QLC flash
- In-line data reduction and erasure coding
- Triple Parity RAID for enhanced capacity efficiency, data protection, and security
- Up to 60x data reduction and significant performance gains with StoreOnce integration



Right-size with ease

- HPE Alletra Storage MP
- Shared everything disaggregated storage
- Independent scaling of compute and storage
- Start small and scale big

Storage Performance is Essential

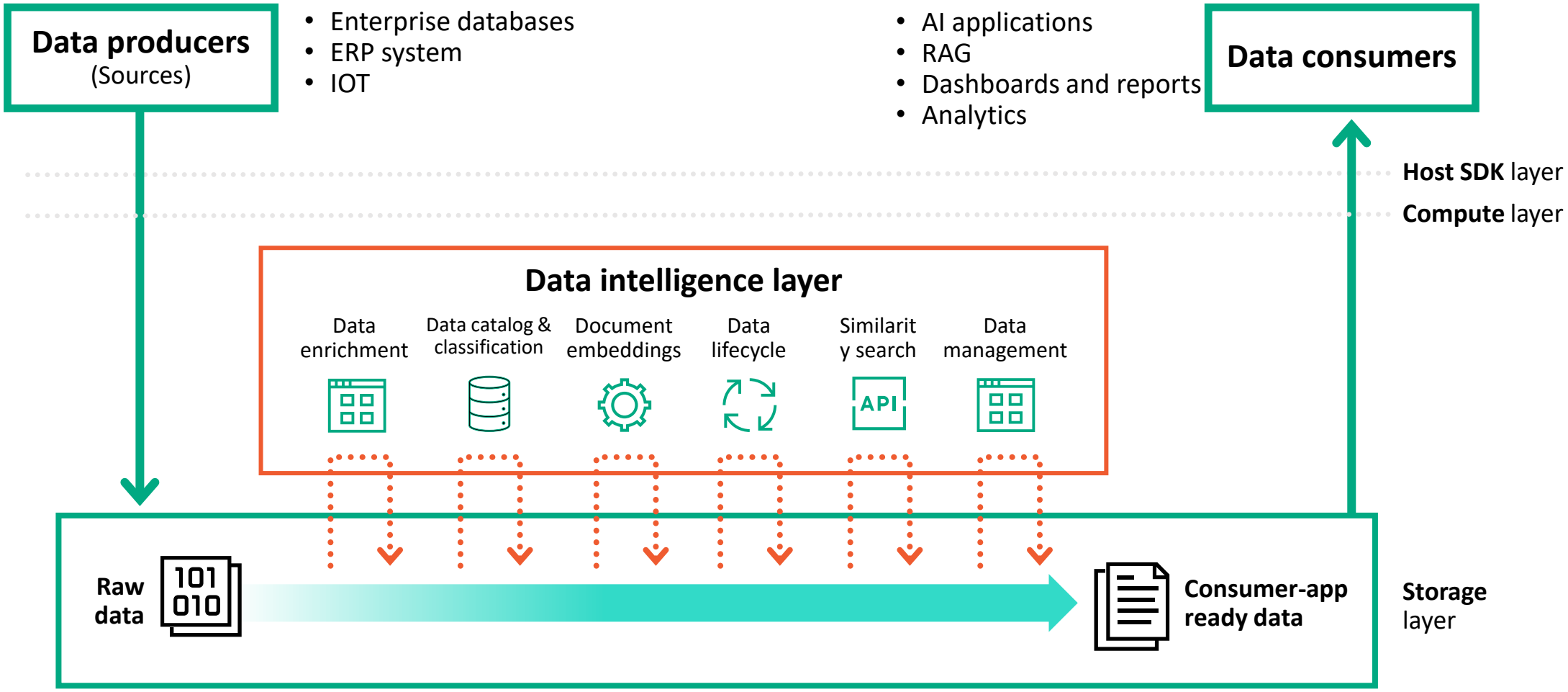


GPU Direct and RDMA for S3 Object

- Collaboration between NVIDIA and HPE to enable RDMA and GDS for Object
- S3 still uses HTTP/TCP for all protocol transactions, RDMA is used for Data transfers
- Initial performance results look promising,
  - ~2X throughput of HTTP for large read workloads
  - ~80% reduction in latency
  - 1% CPU Utilization
- Serves use cases beyond AI, including “World’s Fastest Restore” with Commvault

Unlock data value with inline data intelligence

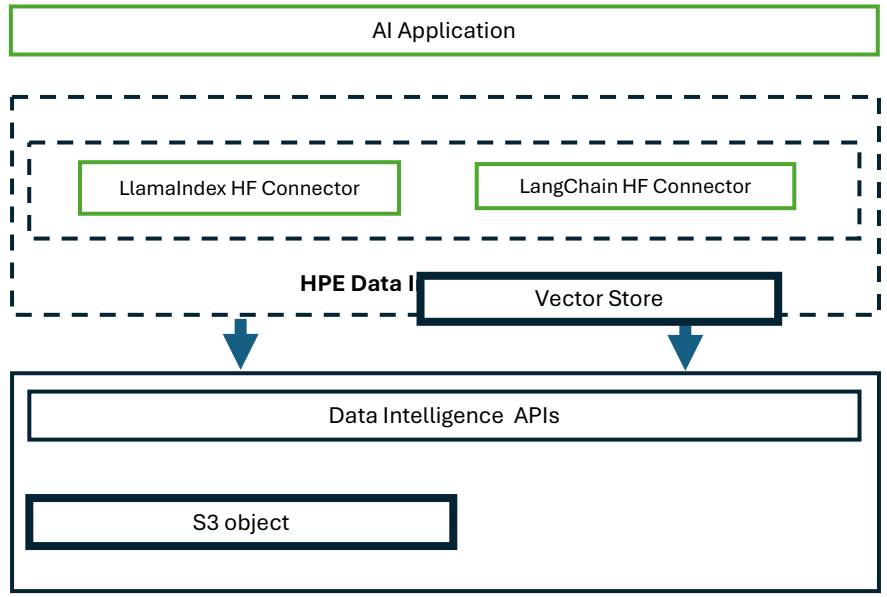
Including NVIDIA GPUDirect for S3 and the X10000 Data Intelligence API and SDK



HPE Data Intelligence SDK

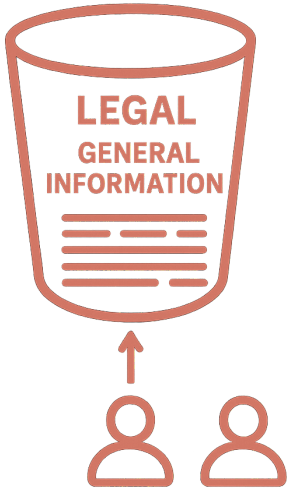
Open source repo:
https://github.com/hpe-storage/pydi-client

- Flexible representation of intelligence metadata
  - Decouples applications from underlying structure of metadata
- Automatic management of metadata lifecycle
  - Keeps metadata in sync with object data
  - Integrated with S3 event notifications API
- Security & Privacy
  - Common model for AuthN (IAM, Active Directory) + AuthZ (user/bucket policies), shared across data & metadata
- Ecosystem fit
  - Built using widely adopted frameworks such as LlamaIndex, LangChain

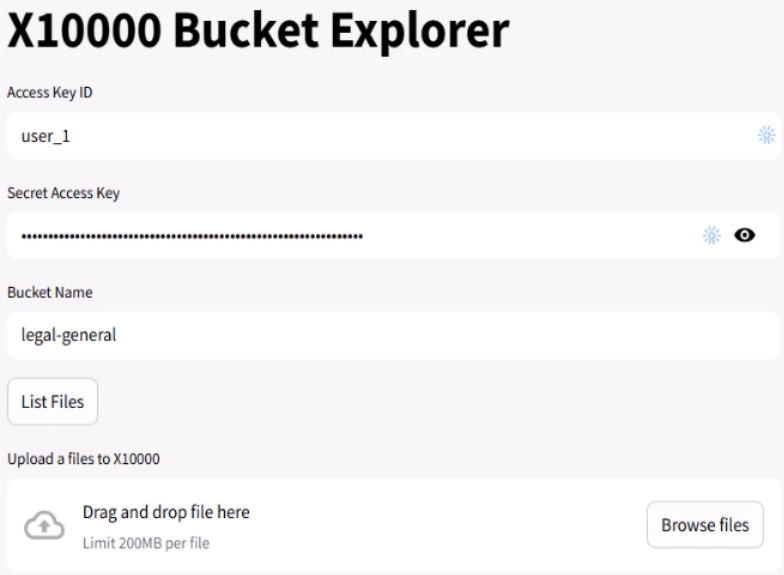


Demo:

Real time data update!

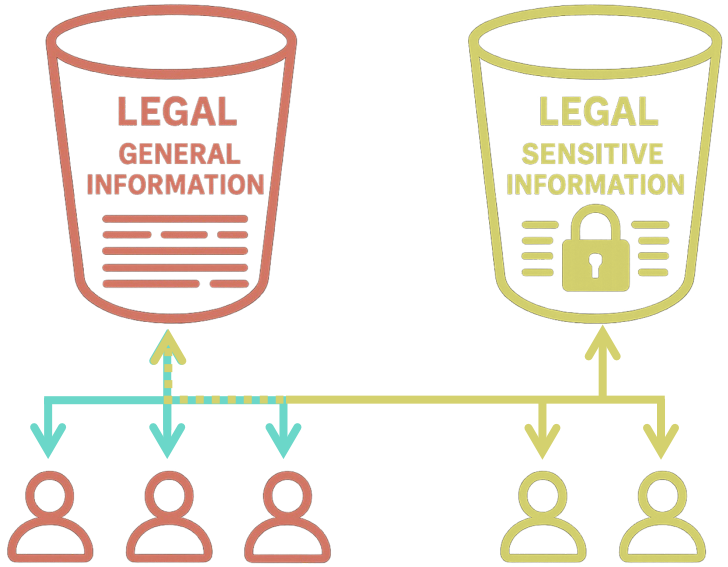


1. Add Policy doc
2. Replace existing Policy with NEW
3. Instant Update of content and vectors
4. Delete Policy
5. Instantly updated

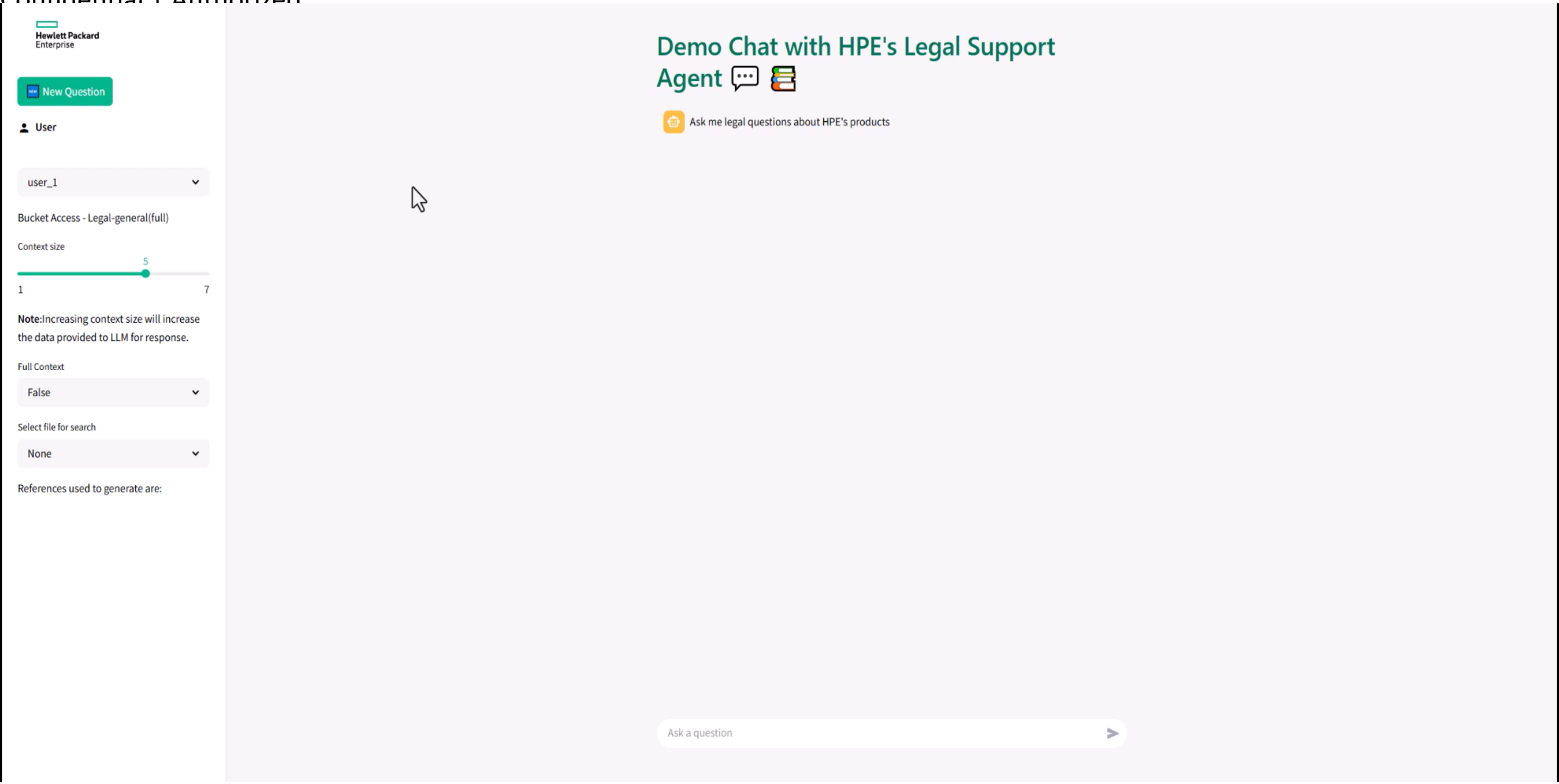

X10000 Bucket Explorer
Access Key ID
user_1
Secret Access Key
Bucket Name
legal-general
List Files
Upload a files to X10000
Drag and drop file here
Limit 200MB per file
Browse files

Demo:

RAG Security



1. Some users don't have access to sensitive data
2. Other users have access to both
3. IAM access is enforced at data layer, vector layer and application



Questions ?

New Frontiers for Agentic AI

