

Cloud Storage Considerations for Retrieval Augmented Generation (RAG) in AI Applications

Daniel Falkner – Senior Product Manager – Azure Object Storage
Scott Hoag – Principal Product Manager – Azure Object Storage

Agenda



Object Storage: Workloads over the years

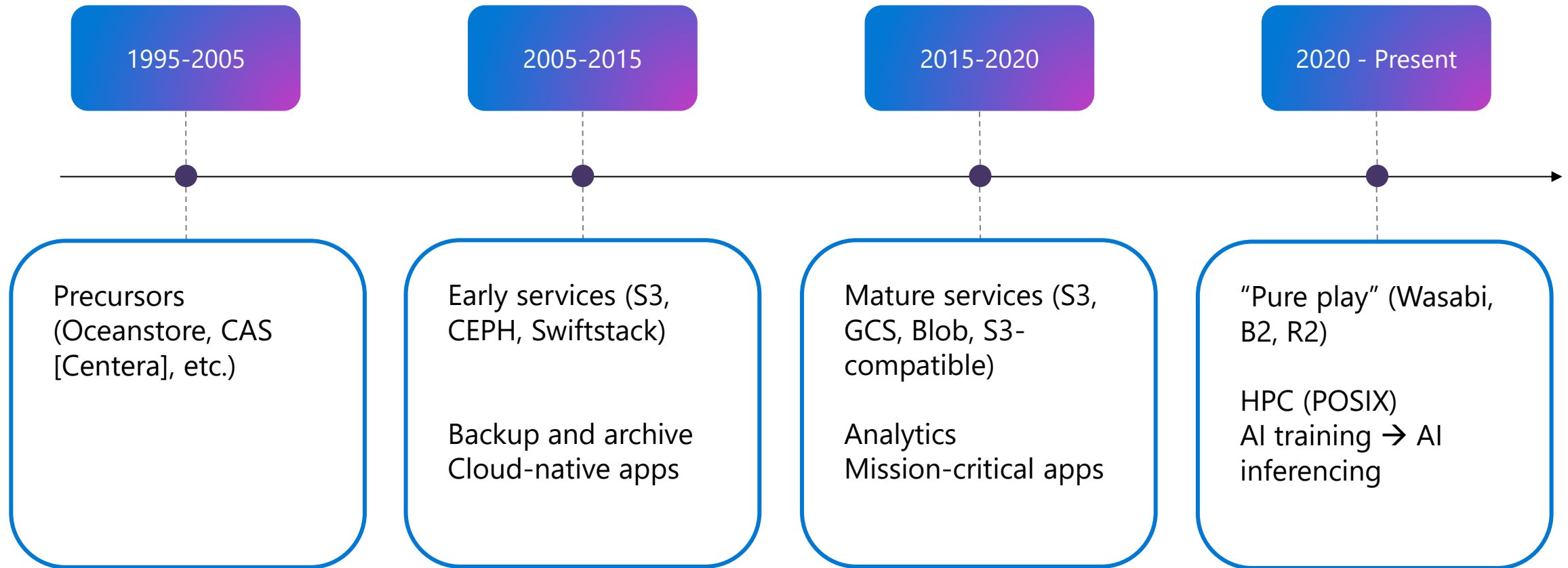


AI Workloads and Storage requirements



Retrieval Augmented Generation (RAG)

Object Storage: Workloads over the years



AI Workloads & Storage Requirements

Azure Storage portfolio

Durable, highly available, massively scalable

Block storage

Services

Azure Disk Storage
Azure Elastic SAN
Azure Container Storage

Unique capabilities

Ultra Disk and Premium SSD v2
Azure Elastic SAN
Shared disks

Object storage

Services

Azure Blob
Data Lake Storage

Unique capabilities

Premium Blob
Multi-protocol access (e.g. NFS, HDFS)
Azure Managed Lustre

File storage

Services

Azure Files
Azure NetApp Files

Unique capabilities

Native NetApp File Storage
Azure File Sync

Capacity

100s of trillions of objects across
many exabytes of data

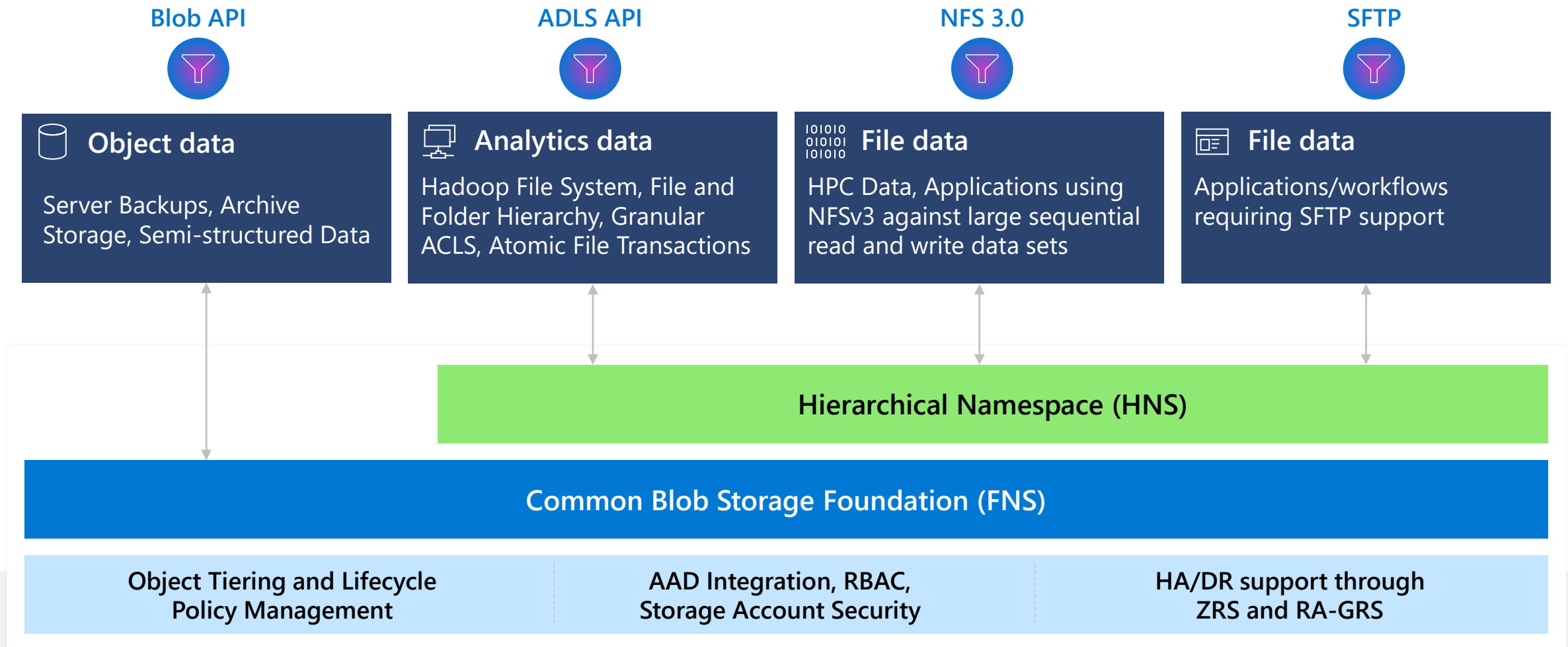
Throughput

> 500 Tbps average
(> 150 exabytes per month)

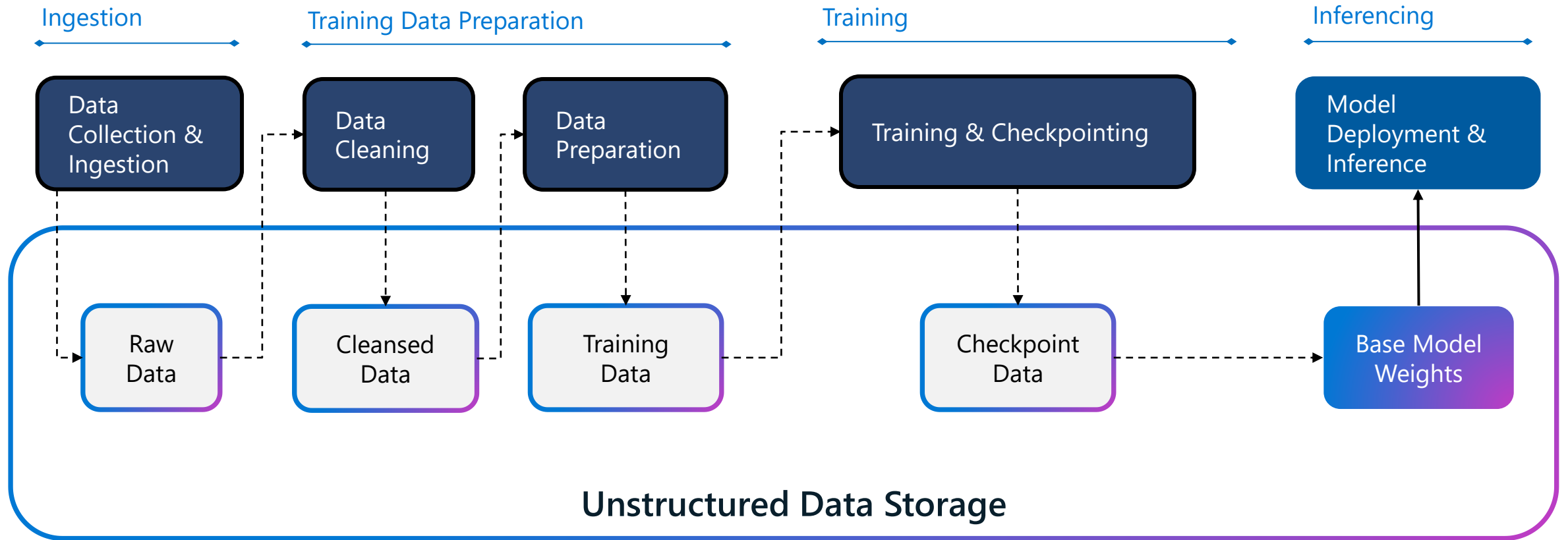
IOPS

> 500M tps
(> 1 quadrillion per month)

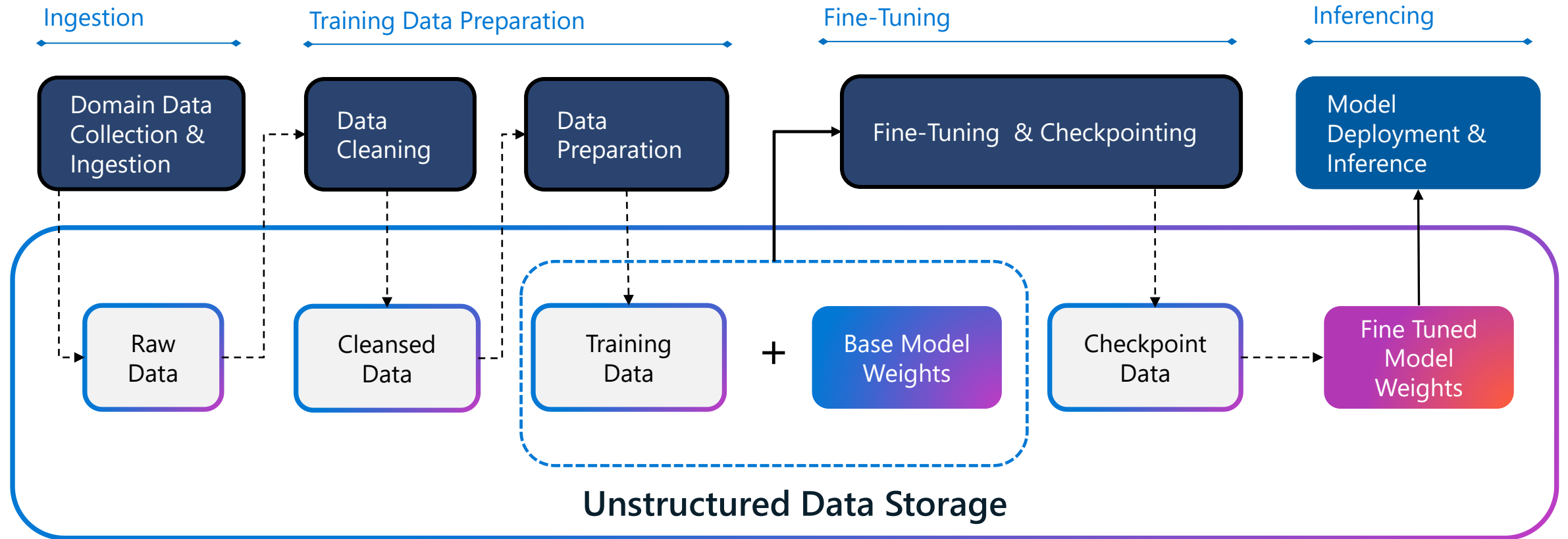
Azure Blob Storage: Multi-protocol, single storage platform



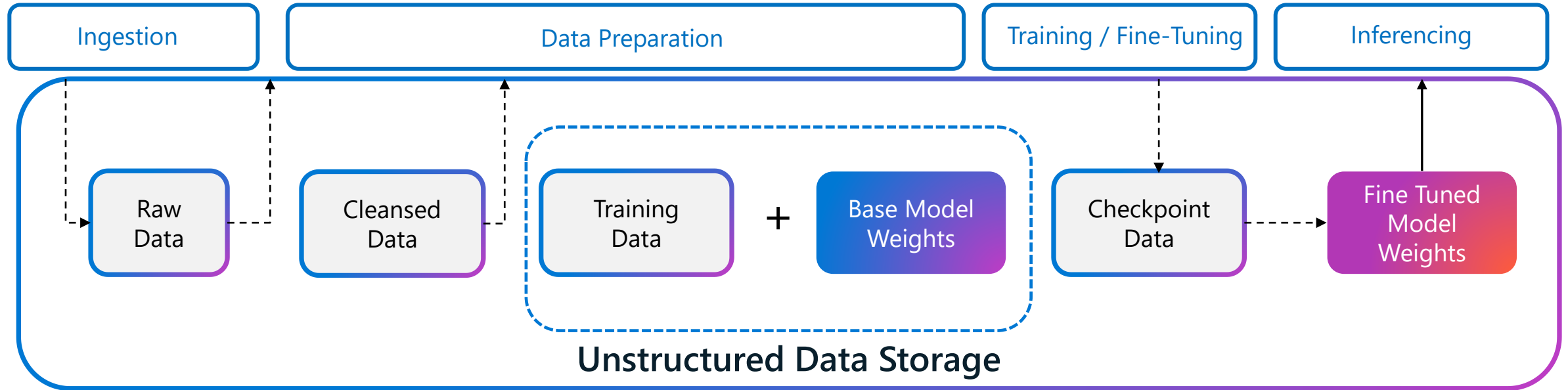
AI Pipeline – Storage centric view



AI Fine-Tuning – Storage centric view



AI Pipeline - Storage Requirements



Requirements

Training / Fine-Tuning

- **Ingestion:** Bring raw training data to Azure
- **Data Preparation:** Integration with Spark, MosaicML, etc.
- **Training/Fine-Tuning:** Data to GPU nodes, checkpoints to storage. Integration with PyTorch and other ML frameworks
- **Data Management:** Secure & cost-efficient retention

Deployment/Inference

- **Deployment:** Model distribution and load times
- **Data Management:** Model versioning, retention of inference inputs and outputs

RAG with Object Storage

Bringing domain knowledge to LLMs



Prompt engineering

In-context learning



Fine-tuning

Learn new skills



Retrieval augmentation (RAG)

Learn new facts

RAG: Retrieval-Augmented Generation

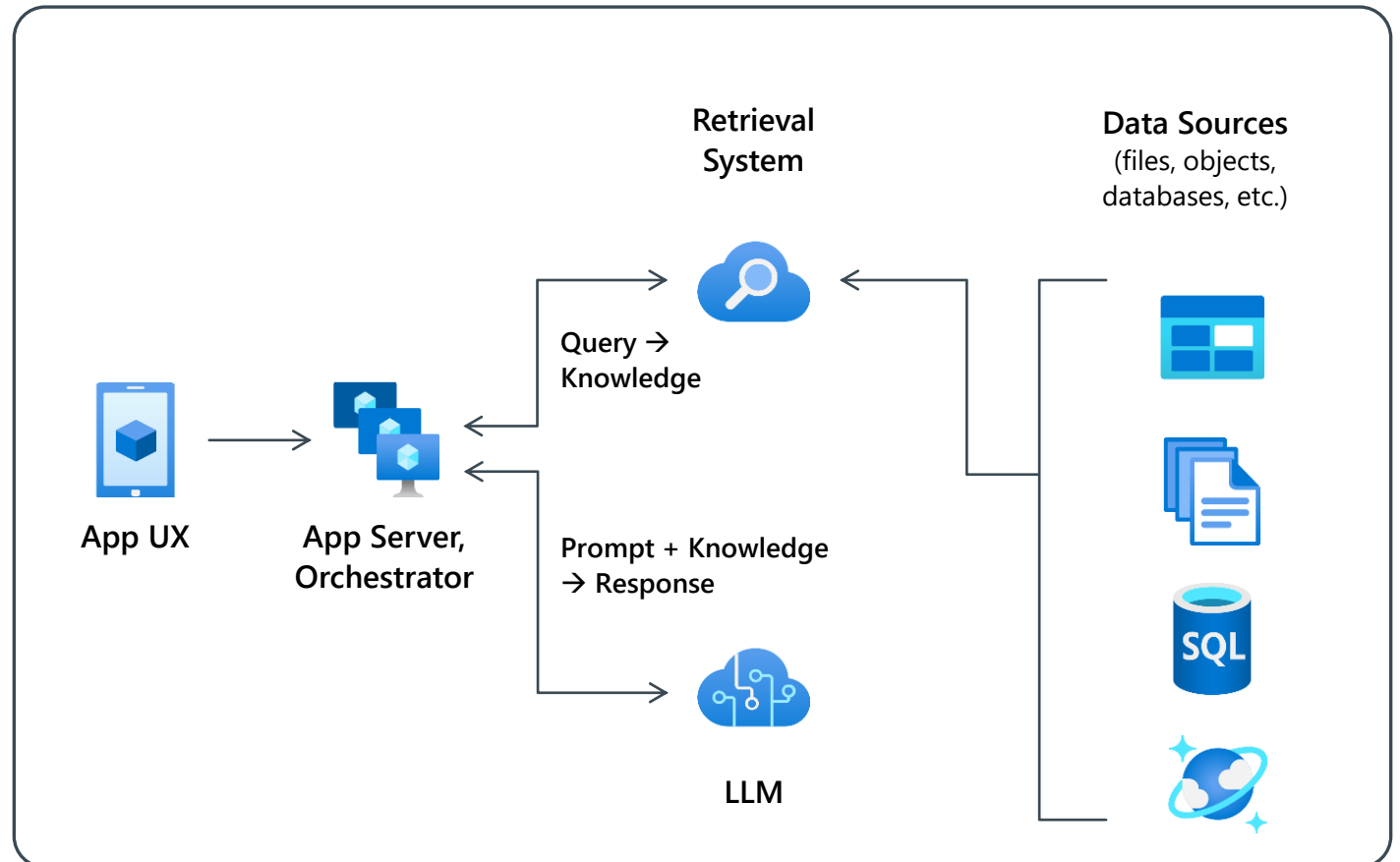
Combine reasoning + knowledge

Key Elements

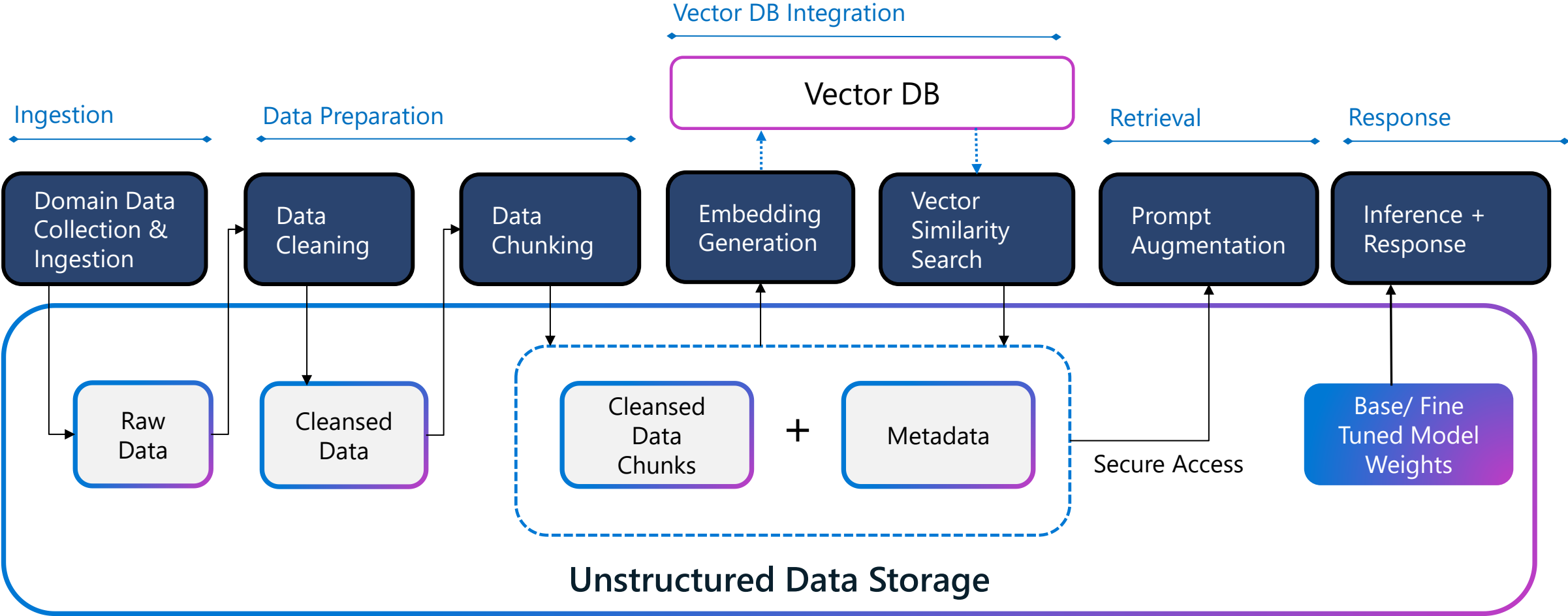
- Externalized knowledge
- Orchestrator drives interaction
- Prompts = instructions + context + grounding data

Achieving quality results

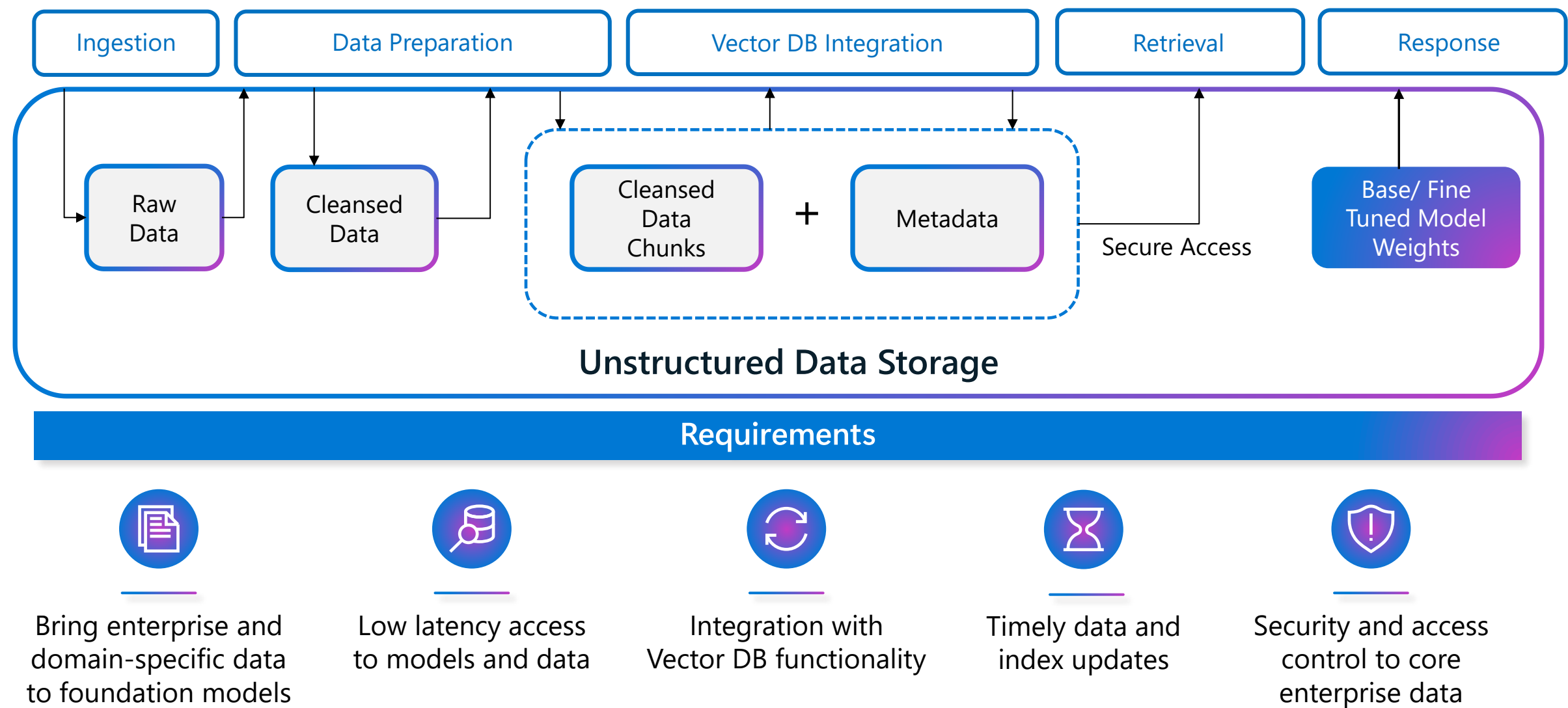
- Different workflows for different tasks
- Evaluation & RAI



RAG Pipeline – Storage centric view



RAG Pipeline – Storage Requirements



RAG with Blob Storage



Multi-Protocol

Unified storage
for heterogenous
updates



Low latency access

Premium Blob
Storage



Vector DB Integration

Azure AI search,
flexibility to BYO

Multiple indexes,
dev-focused
SDK/tools



Freshness

Blob Change
Feed

Change
notifications



Security

Azure Entra ID
integration

RBAC and ABAC

SSD-backed Premium Blob Storage for RAG

Latency (400KiB PDFs)	
Operation	Premium Blob (SSD) vs Standard (Hot) Blob (HDD)
PutBlob (Data Ingest + Chunk Writes)	Premium ~2x faster
GetBlob (Vectorization)	Premium ~3x faster
GetBlob (Retrieval)	Premium ~3x faster

Premium delivers ~3x faster RAG performance with 65% savings on Transactions!

Blob Storage: Scaled accounts



Increase Capacity,
Bandwidth, IOPS to meet
hyperscale customer asks



Scale across a much
larger hardware footprint



Eliminate the need for
sharding large workloads
across accounts



No change to pricing,
not a new product/SKU,
goal is to auto-scale

Workloads



Storage

Acct

Acct

Acct

Acct

Slice 1

Scaled

Account

Slice 2

Acct

Acct

Storage

Key Takeaways

Leveraging Object Storage for building AI Apps ...

Ideal for AI Training and Fine Tuning

Scalable

to Exabytes of data and many Tbps of throughput

Cost-effective

with storage tiers and automated lifecycle management

Integrated

with analytics engines for data preparation

Interoperable

GPU Node-local POSIX-like mount points

Accelerates building RAG based LLM Apps

Secure

with identity-based AuthN, AuthZ, RBAC

Interoperable

with vector DBs and orchestrators with SDK/tools

Low-Latency Access

with SSD-backed object storage

Freshness

with change feed and change notifications

Q&A

Resources

- [Building AI applications that leverage your data in object storage | BRK216](#)
- [Open AI accelerates AI innovation with Microsoft Azure Blob Storage scaled accounts](#)
- [Training and fine-tuning your large AI models using Azure Blob Storage](#)
- [Bring your data to AI applications – RAG with Azure Blob Storage](#)
- [SNIA SDC 2024 - Supercharging OpenAI Training with Azure Blob Storage](#)
- [OpenAI transforms AI model development with Azure Blob Storage | Microsoft Customer Stories](#)