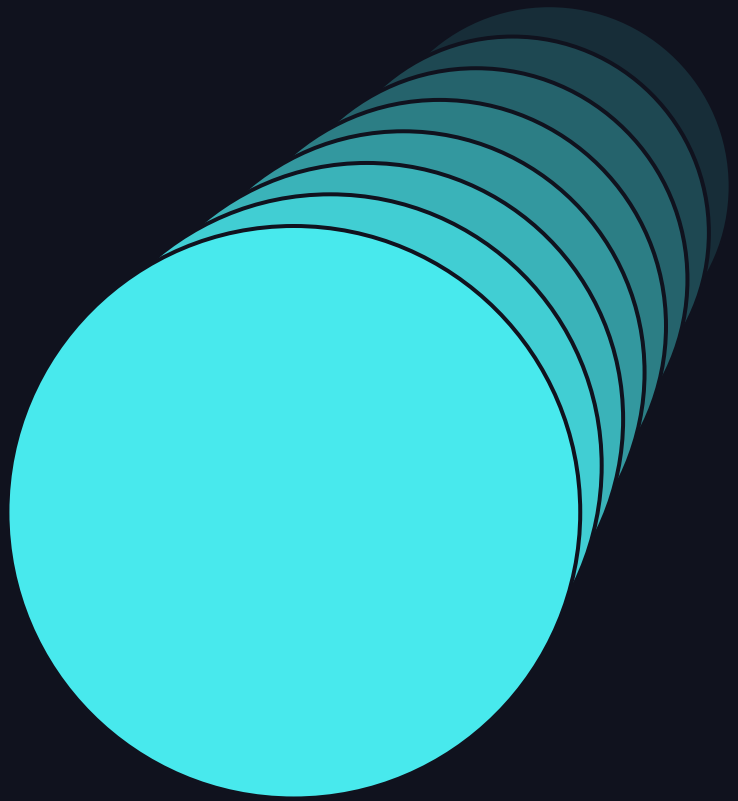
DATA+AI SUMMIT BY databricks

Powering scalable analytics and AI with Azure Data Lake Storage



Jeff King, Microsoft
Saurabh Sensharma, Microsoft

Agenda

AI Workloads and Storage requirements

Training & Fine Tuning

Retrieval Augmented Generation (RAG)

Bringing domain knowledge to LLMs

Prompt engineering

In-context learning

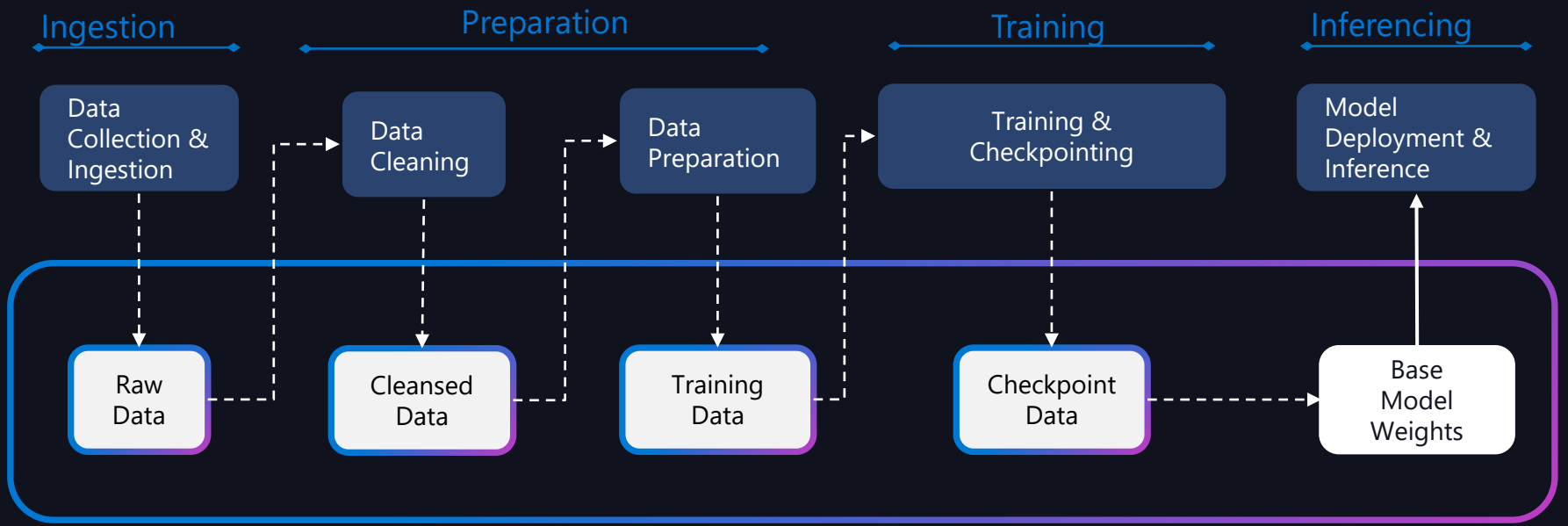
Fine-tuning

Learn new skills

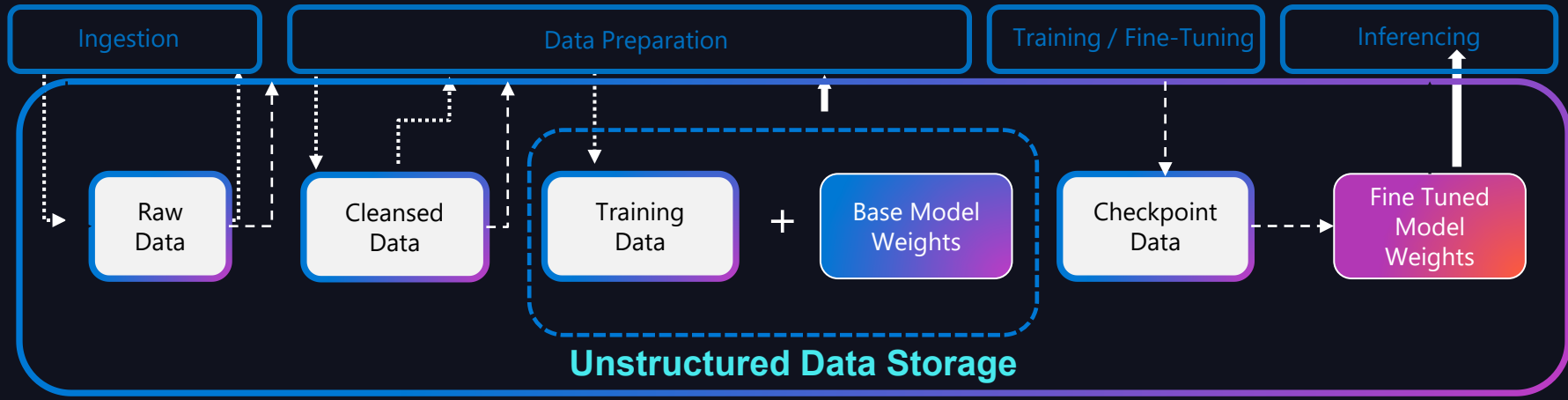
Retrieval augmentation (RAG)

Learn new facts

Analytics + AI Training Pipeline



AI Pipeline - Storage Requirements



Requirements

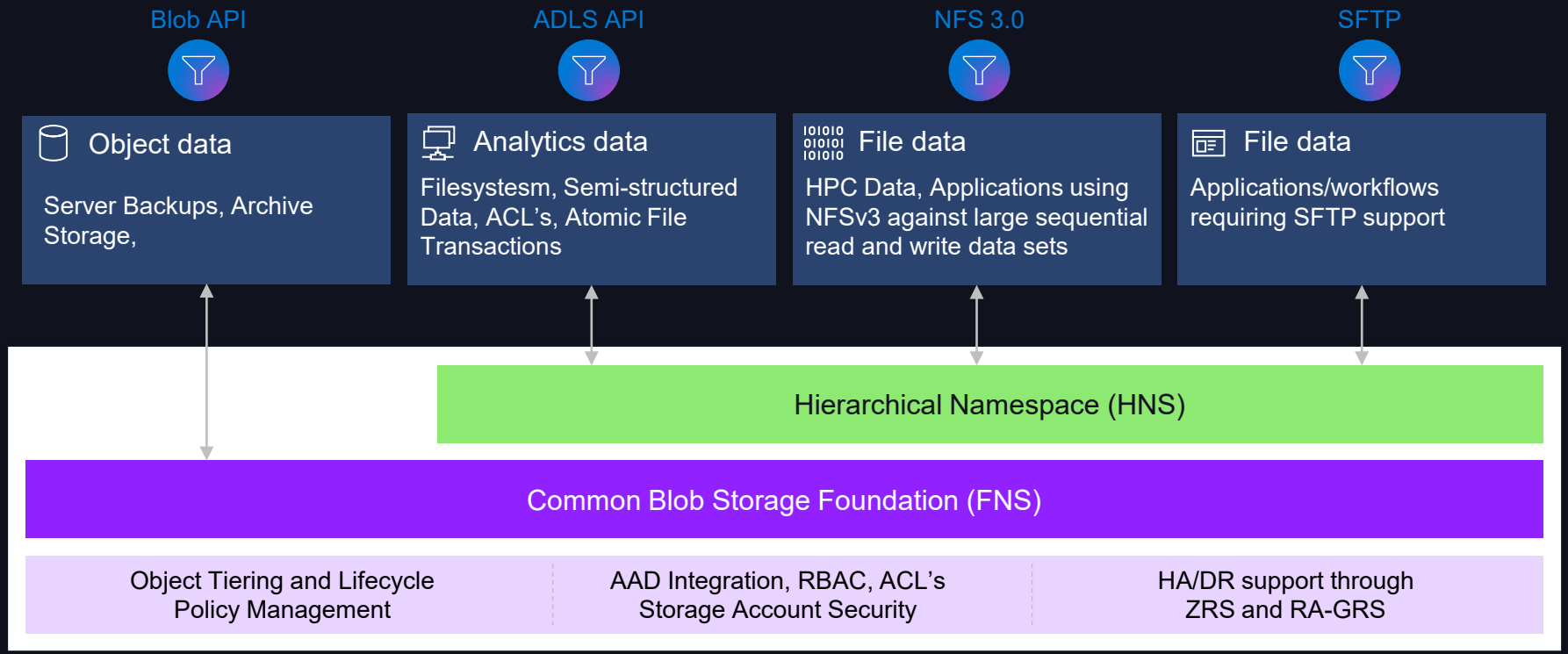
Training / Fine-Tuning

- **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

Azure Storage: Multi-protocol, single platform



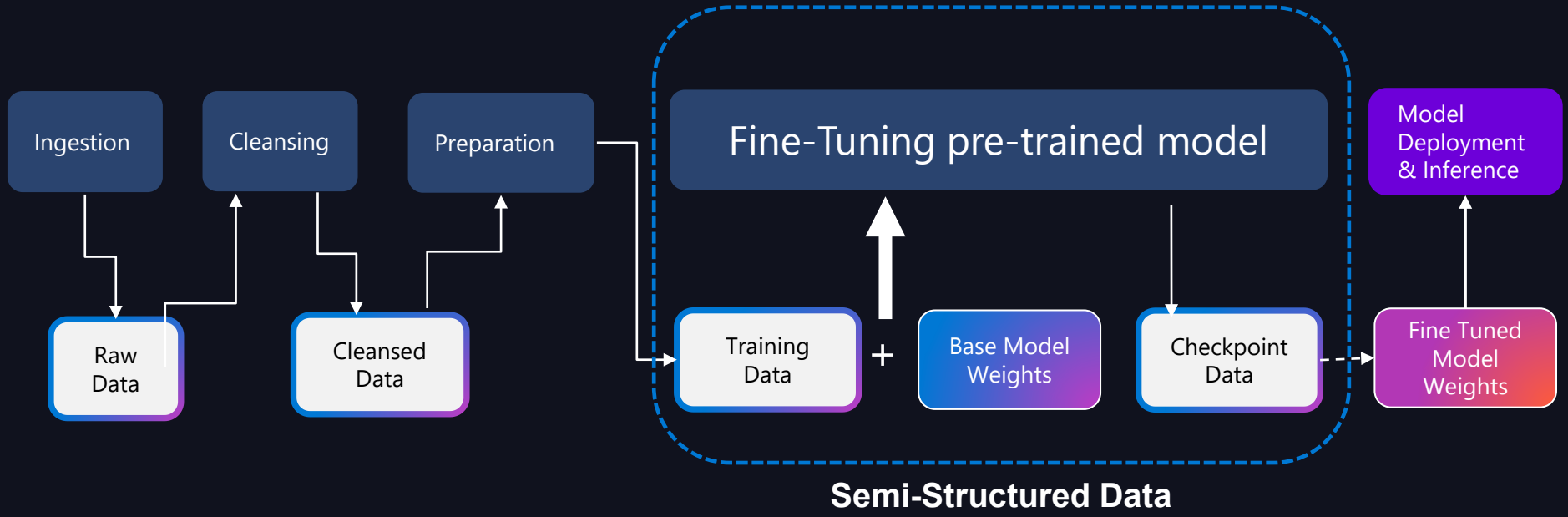
Azure Storage Solutions

		Azure Standard Blob/ADLS	Azure Premium Blob/ADLS	Azure Premium Files	Azure NetApp Files	Azure Managed Lustre
Performance & scale	Capacity	**100s of PiB**	**10s of PiB**	100 TiB	500 TiB	**1 PiB**
	Bandwidth	**>1 Tbps**	100s of Gbps	10 Gbps	10 Gbps	**512 Gbps**
	IOPS	100s of KTps	100s of KTps	100 KTps	800 KTps	**> 100 KTps**
	Latency	<100 ms	**<10 ms**	3-5 ms	**<1 ms**	**<2 ms**
Protocols		**REST** HDFS **NFSv3** SFTP **FUSE** CSI	**REST** HDFS **NFSv3** SFTP **FUSE** CSI	REST **NFSv4.1 (or) SMB** CSI	**NFSv3 & SMB3*** **NFSv4.1 & SMB*** CSI	**Lustre** CSI

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

Training & Fine-tuning with ADLS

Fine-Tuning Pipeline



Training and Fine-Tuning

Scalable

PBs of data, Tbps of R/W throughput

Cost Effective

Tiered storage for long-term retention

Multi-Protocol

Integrates well with analytics engines

Optimized for node-local access

BlobFuse2 with caching

Data Management

Blob index tags, automated lifecycle management

Demo: Blobfuse2 perf

Blobfuse2

GA + update

High Throughput access to Blob Storage

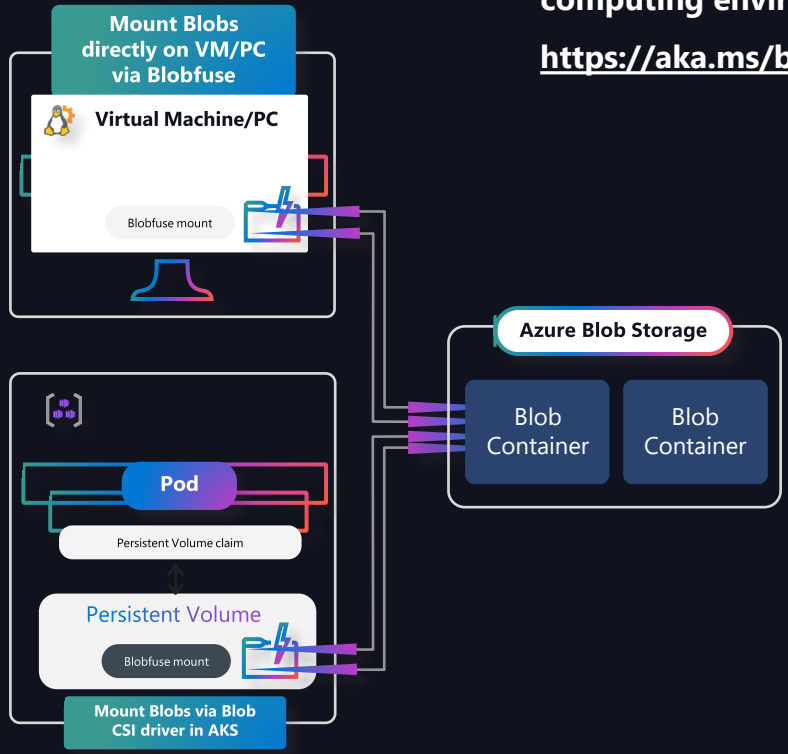
Easy to install and work with PiB scale data

Open-sourced & supported by Microsoft

Secure access to data

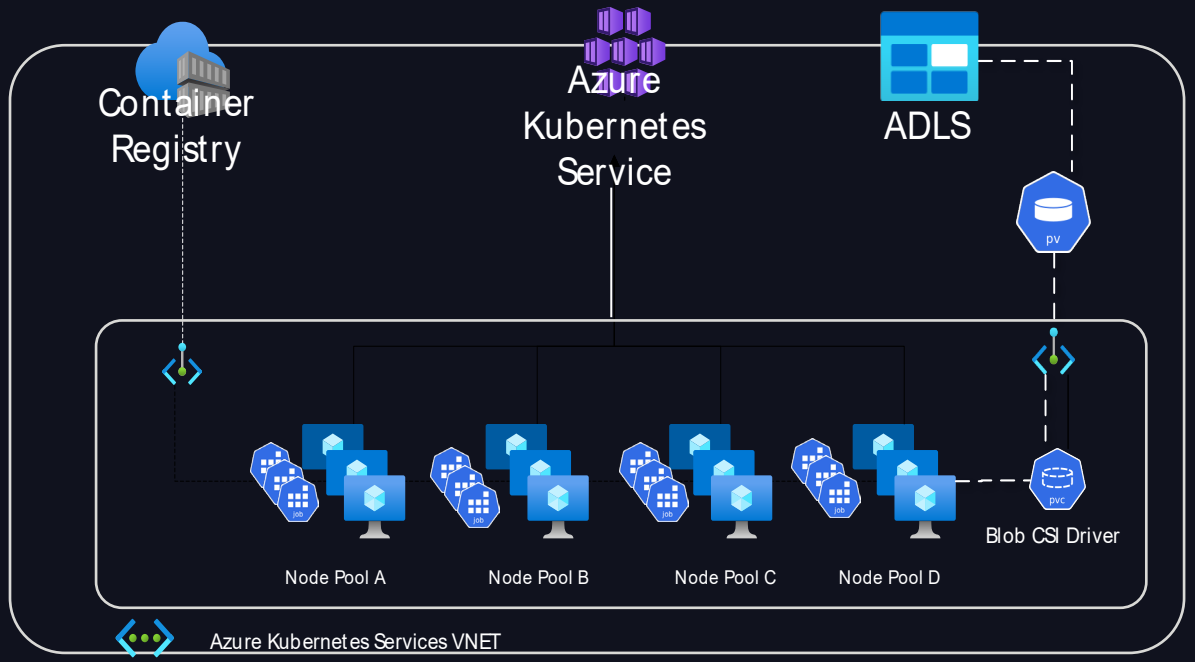
Virtual File System Driver to mount blob storage in your computing environment

https://aka.ms/blobfuse



IOR Benchmark test for HPC systems

Objective Maximizing available storage bandwidth using Blobfuse2

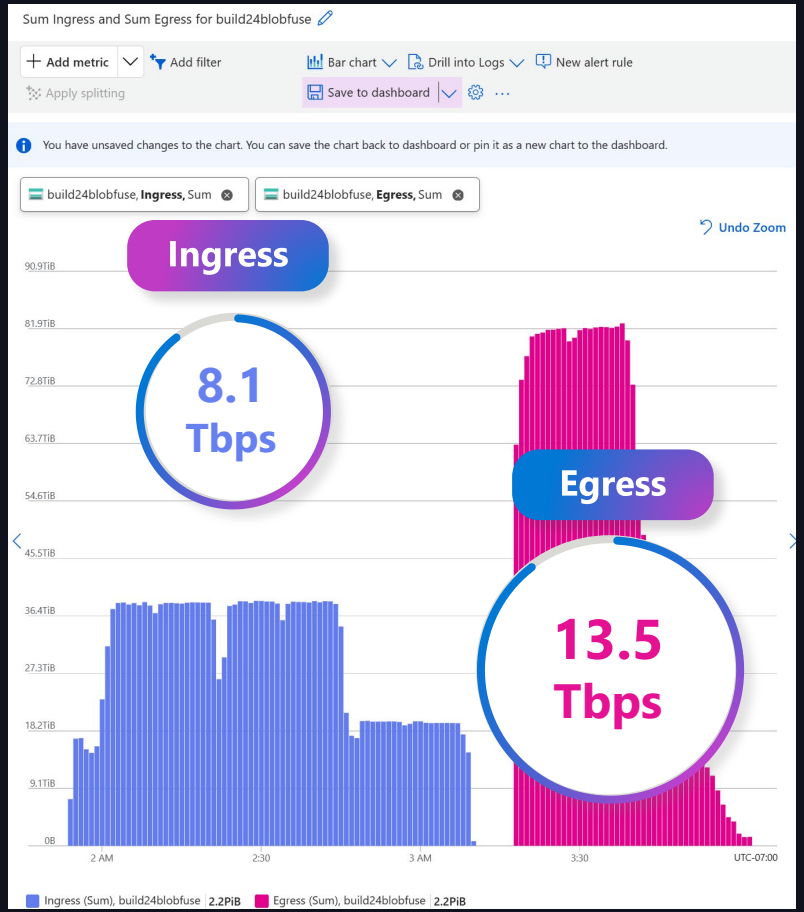


Setup

- D96ds and D96ads Spot VMs
- 350 AKS Pods
- 16,800 cores
- Blobfuse2 via AKS CSI driver
- > 2 PiB working data

https://github.com/hpc/ior

IOR Benchmark Results



Setup

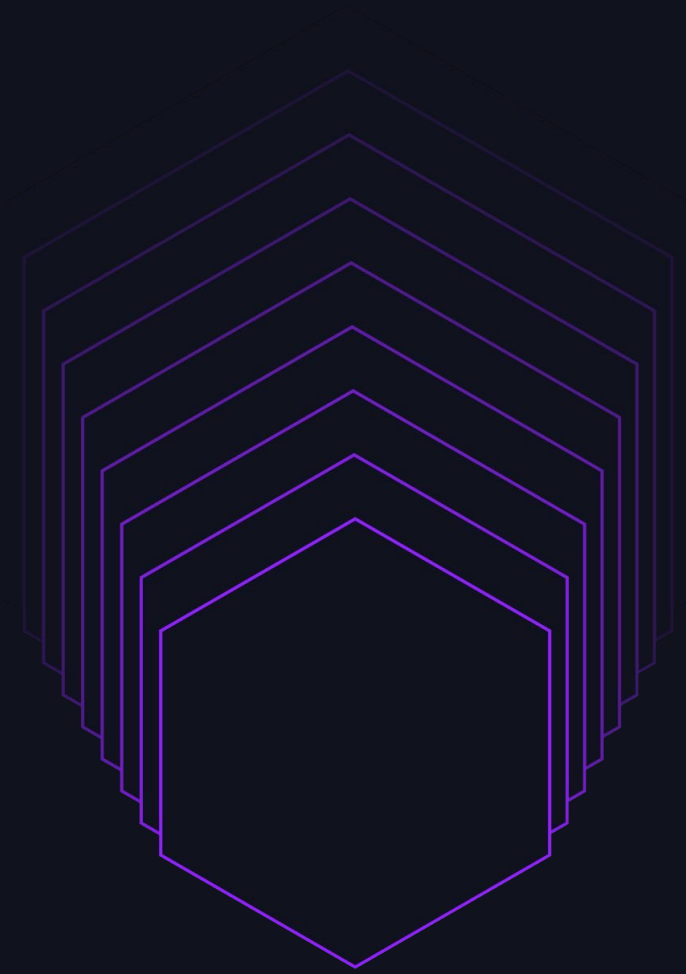
- D96ds and D96ads Spot VMs
- 350 AKS Pods
- 16,800 cores
- Blobfuse2 via AKS CSI driver
- > 2 PiB working data

Demo: Fine Tuning with Spot VM's

Special thanks to Wolfgang De Salvador!

Demo

RAG



RAG: Retrieval-Augmented Generation

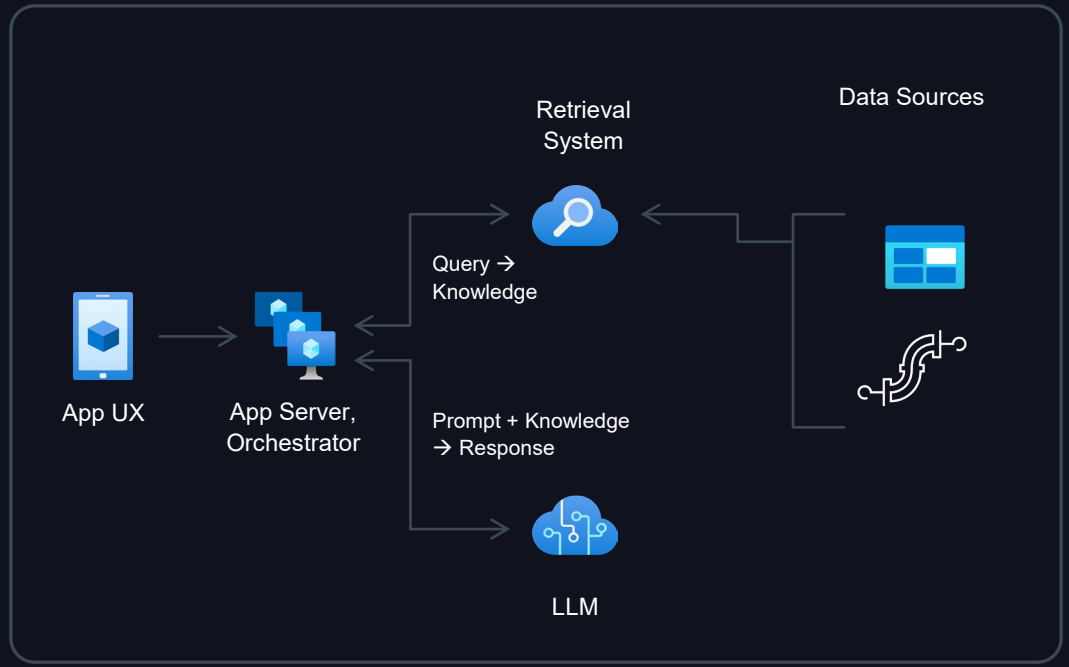
Combine reasoning + knowledge

Key Elements

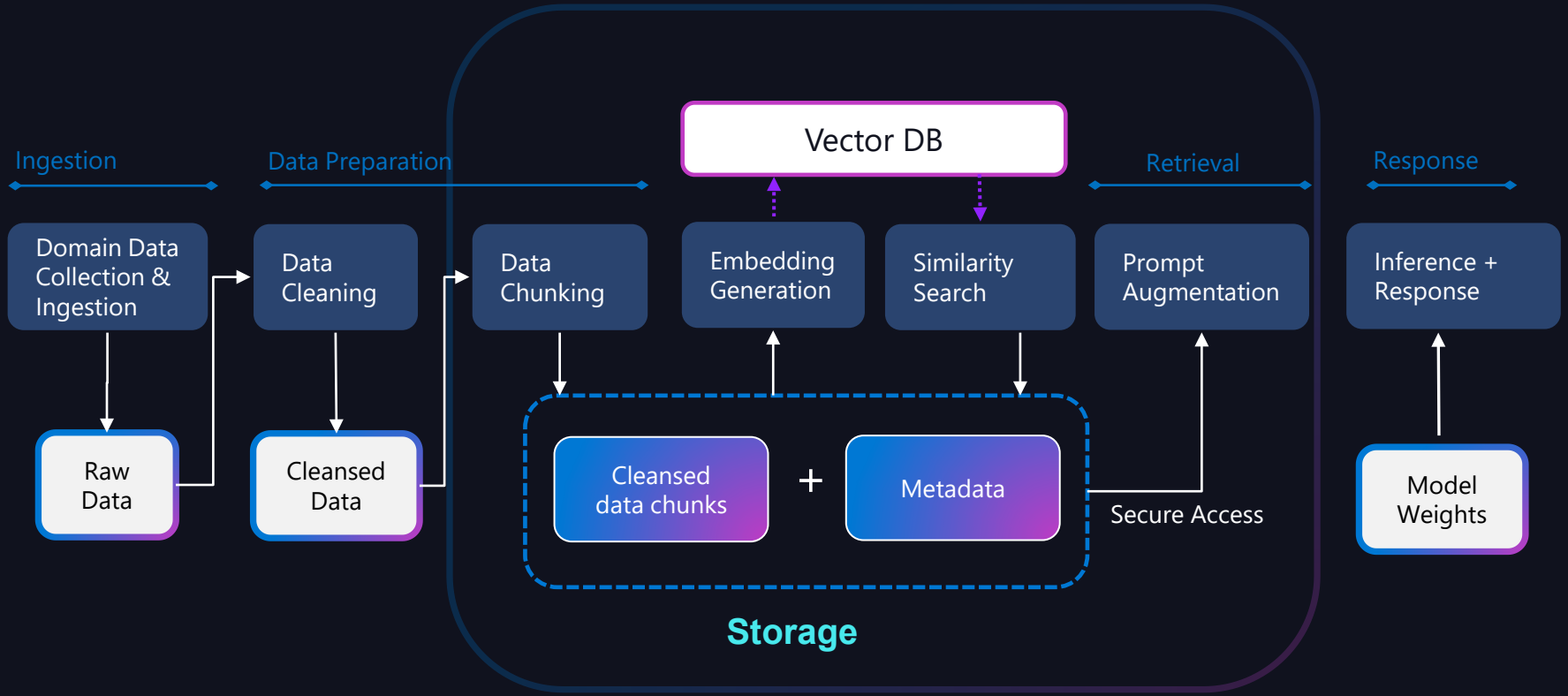
- Externalized knowledge
- Orchestrator drives interaction
- Prompts =
instructions + context + grounding data

Achieving quality results

- Different workflows for different tasks
- Evaluation & RAI

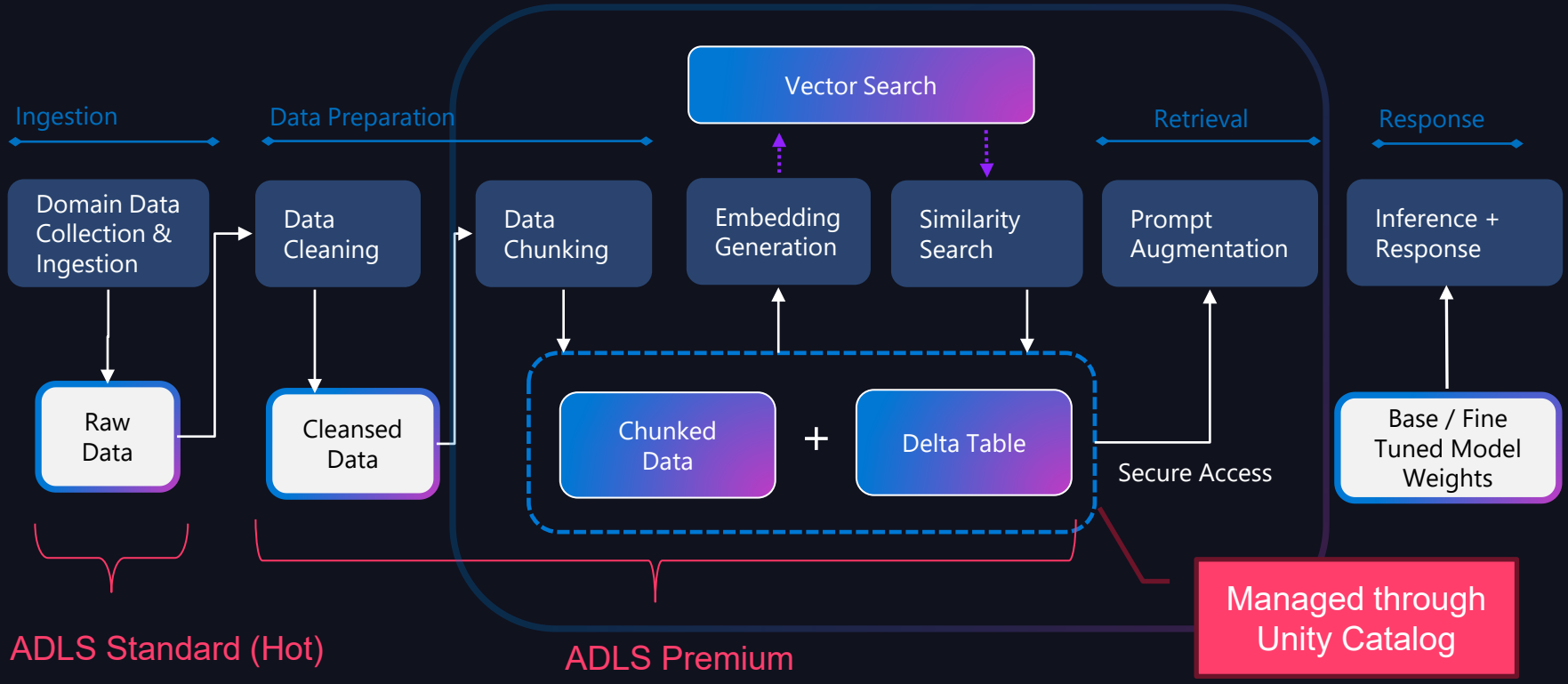


RAG



Demo: RAG walkthrough

RAG (Demo implementation)



RAG on ADLS + Azure Databricks

Multi-Protocol

Unified storage for heterogenous updates

Low latency access

Premium ADLS

Vector DB Integration

Vector Search, flexibility to BYO

Multiple indexes, dev-focused SDK/tools

Freshness

Change Feed

Change notifications

Security

Azure Entra ID

ACL

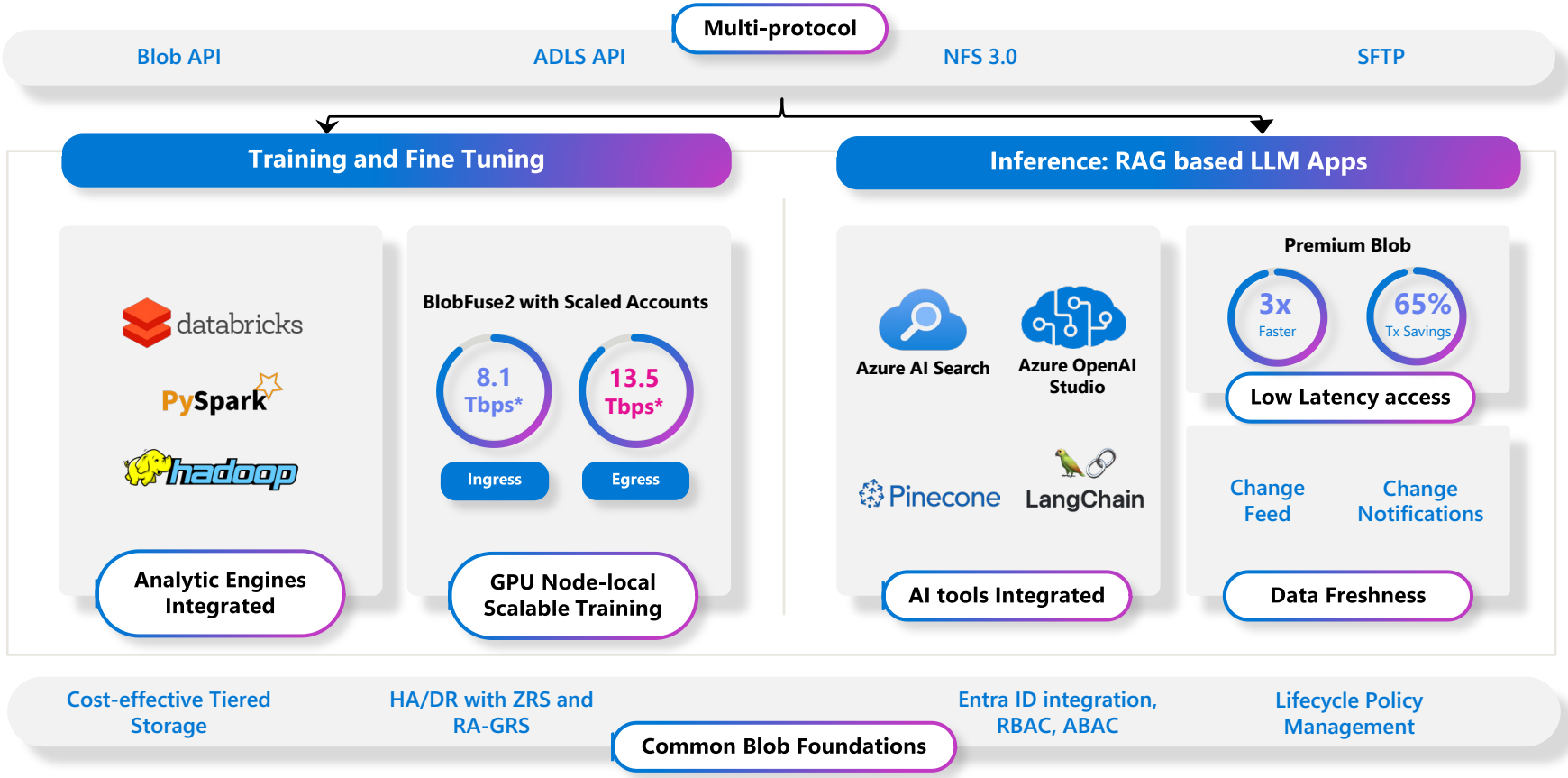
RBAC

ABAC

RLS, CLS

Demo 2: Automatically processing change

Azure Blob Storage is great for building AI Apps…



*Peak throughput results from IOR Benchmark test for HPC systems: https://github.com/hpc/ior

Key Takeaways

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

BlobFuse2 for seamless use in GPU clusters

Accelerates building RAG based LLM Apps

Integrated

ADLS, Databricks, OAI,

Interoperable

with vector DBs and orchestrators with SDK/tools

Secure

with Entra ID integration, RBAC and ABAC

Low-Latency Access

with Premium Blob Storage

Freshness

with Blob change feed and change notifications

Related sessions

Type	Title	Date/Time	Location	Speaker
Customer Breakout	**Scaling Real-Time Healthcare Data Processing for the Veterans Affairs**	Wednesday 5:10 – 5:50	SL2 Room 202	Kash Sabba – MSFT Spencer Schaefer - VA
Microsoft Breakout	**Confidential Computing in Azure Databricks**	Thursday 12.30 – 1.10	South Esplanade Room 159	Lindsey Allen - MSFT

Get access to the session demos on GitHub!

Aka.ms/XStore-DAIS2024