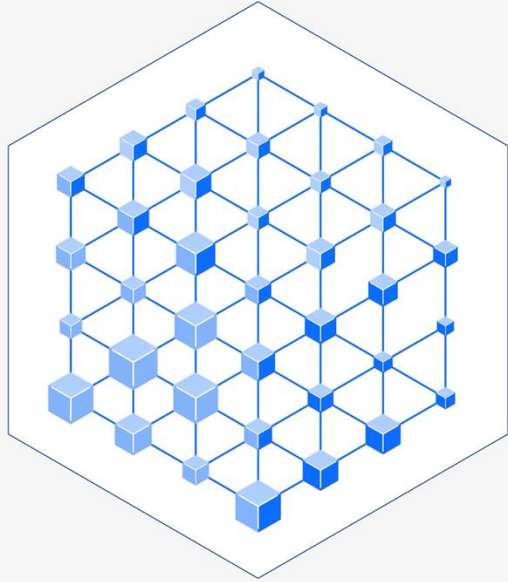


# Agentic AI in Action

Ted Hoover

Product manager , Storage for Data AI





# What is agentic AI?

Agentic AI is a **framework for accomplishing goals** with limited supervision that **consists of AI agents**.

In multiagent systems, **each agent performs a specific subtask** that's required to reach the goal.

# AI – Value Evolution



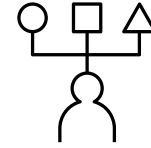
## AI Assistant

Information retrieval



## AI Agent

Perform tasks with oversight



## Multi-agent Platform

Autonomous action-taking



Reactive

Generates content

Prescriptive tasks

Single-step processes

Human interaction required

Fixed flow at build time

Proactive

Makes operational decisions

Goal-oriented

Multi-step processes

Less or no human interaction required

Reflective and Self-correcting

# Technical evolution: from standalone models to agentic systems

~2022

## Large Language Models (LLMs)

- Models that predict the next word
- Pitfalls:
  - Limited by what they were trained on
  - Hallucinations (weather)
  - Bad at certain things (e.g. math)

*Instead of trying to put all the knowledge inside the model, design systems on top of the model*

~2023

## Compound AI Systems (fixed flows, e.g. RAG)

- Way to infuse new knowledge without retraining the model
- Uses multiple interacting components - calls to models, retrievers, or external tools (e.g. guard rails)
- Reduces risk of hallucinations
- Pitfalls:
  - Needs set up at build time (“fixed flow”) which can limit use cases

*What if, instead of responding to a question, AI can accomplish a task or goal?*

~2024+

## Agentic Systems

- A program whose **execution logic is controlled by an LLM**
- Instead of “generating content” to return to the user, it performs actions (“has **agency**”) on behalf of the user
- Can define how to solve, then reflect, and update the plan (vs. fixed systems that are set up at build time)

Generates content

Takes action

More reliable

Less flexible

Less reliable

More flexible

Expect to see combinations of fixed flows and autonomous agents

# Why do we need a new AI storage?

## 1 Need fast and cost-effective storage for distributed inferencing

- Accelerate TTFT with high performance storage
- Improve GPU efficiency by preserve KV data (tokens)

Production service data from Moonshot AI

Source: <https://www.usenix.org/system/files/fast25-qin.pdf>

Use case	Input tokens processed <i>per hour</i>	Cache hit rate	Unique KV data generated per day (projection)
Conversational Chat Bot	144M	40%	<b>662 TB</b>
Tool & Agent	203M	60%	<b>864 TB</b>

## 2 Current method to search data is very inefficient Only a small fraction (<1%) of the enterprise data is used in Gen A

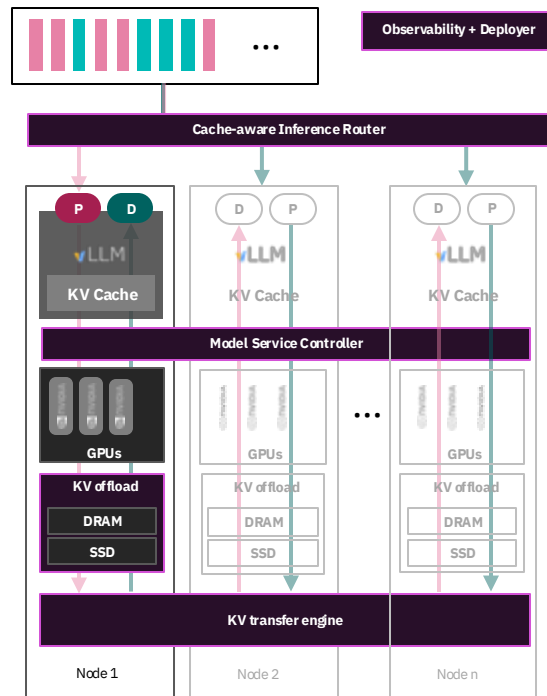
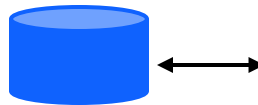
- Data is copied multiple times – from source to lakehouse to data processor to vector database
- Too many copies of the data, too much data transfer, loss of security access control
- *All the data* gets reprocessed every time – no awareness of data changes

1

# Fast storage accelerates and cost-optimizes AI inference

Storing and retrieving KV Cache from external storage offers many advantages:

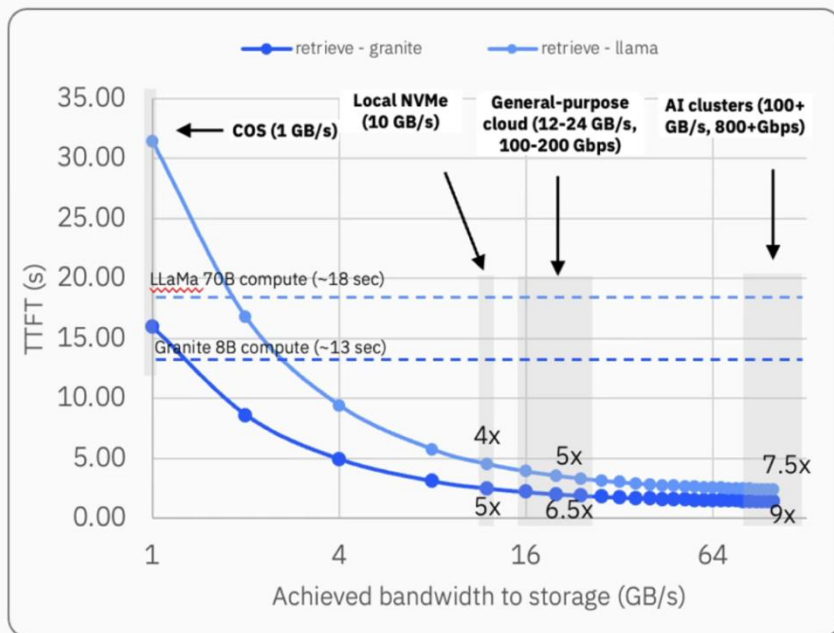
- **Performance:** retrieving KV data from high-I/O storage dramatically reduces time-to-first-token
- **Scalability:** Persistence and reuse of larger volumes of KV data, beyond what fits in local compute resources
- **Cost efficiency:** Replacing GPU cycles with storage capacity can reduce overall costs
- **Persistence:** Ability to restore history and context across sessions, including GPU context-switching



# Fast storage accelerates and cost-optimizes AI inference

## Performance (TTFT)

**TTFT** speed-up by KV cache retrieval from storage  
vs. re-computation (Modeling)



**Assumptions:** H100 GPU; 128k context; average prefix match hit of 75% (varies by prefix match hit)

## Scalability

Production service data from Moonshot AI

Source: <https://www.usenix.org/system/files/fast25-qin.pdf>

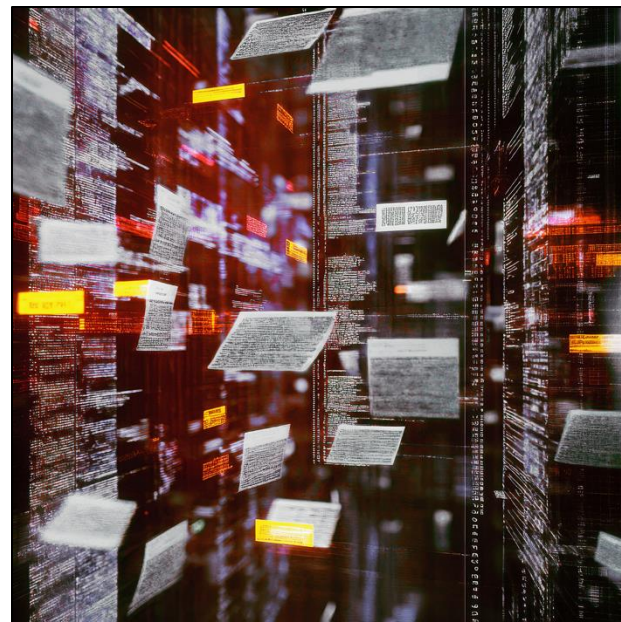
Use case	Input tokens processed per hour	Cache hit rate	Unique KV data generated per day (projection)
Conversational Chat Bot	144M	40%	<b>662 TB</b>
Tool & Agent	203M	60%	<b>864 TB</b>

Details: LLaMa 70B (v3), H800, *average* input seq. length ~ 12k tokens

Total DRAM ~ 1 TB per node, typical DGX fits ~ 60 TB NVMe per node

## 2 The 1% problem

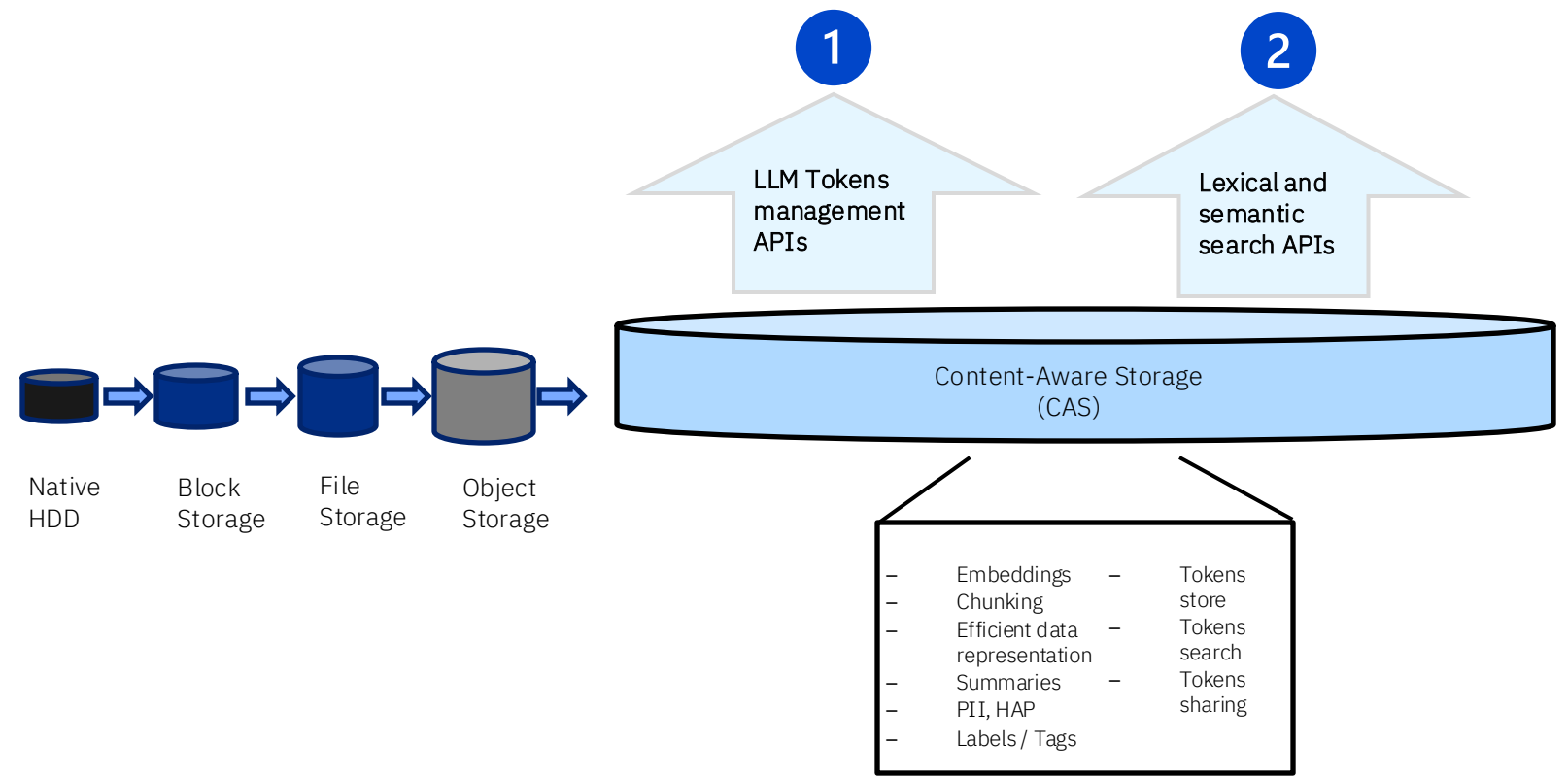
- Organizations are swamped with unstructured data
- But less than 1% of all enterprise data was used to train major large language models
- Retrieval augmented generation (RAG) improves inferencing by incorporating near real-time data, but it's costly, complex, and time consuming
- Other issues include data copying, security vulnerabilities, and operational challenges





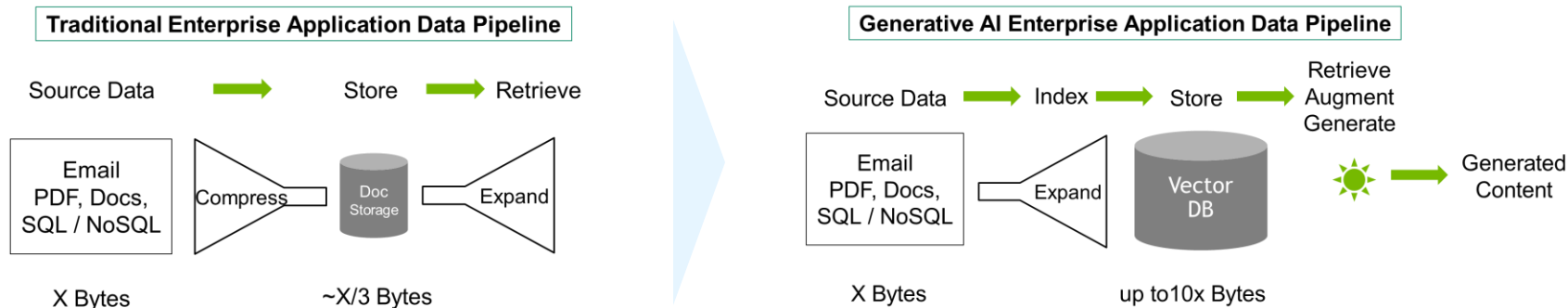
# AI storage: IBM content aware storage

## Reimagining storage systems and consumption of Gen AI data



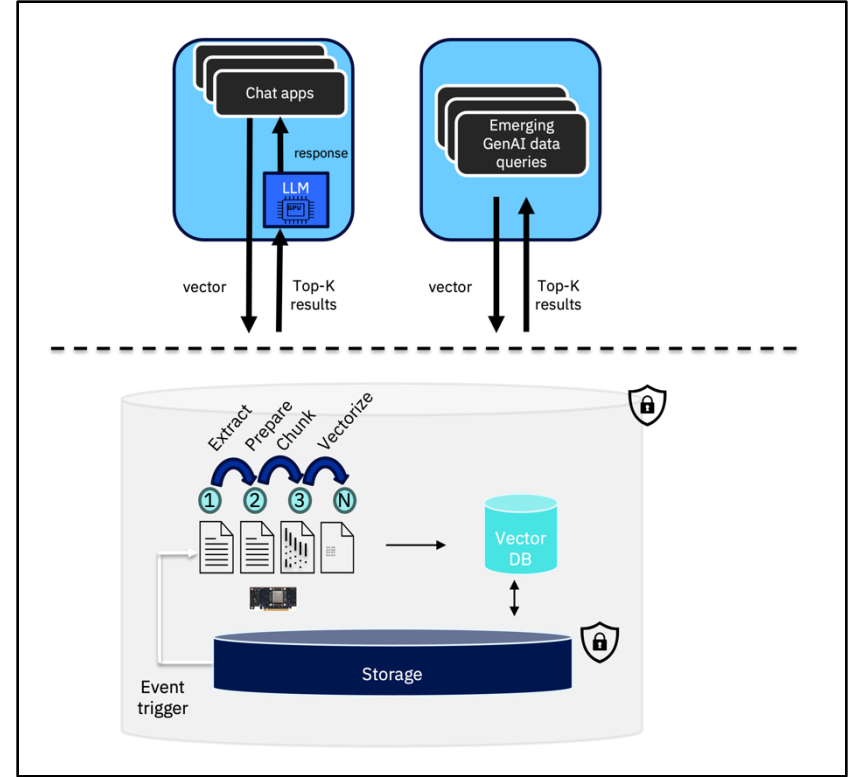
# Rationale for building with a storage-first mentality

- Scalability: capacity multiplier effect – vectorization process generates data derivatives
- Security: data access controls – leverage RBAC policies enforced by the storage system
- Data gravity: bring AI to the data – process the data closer to where it resides (e.g., NVIDIA GDS)



# Foundational technologies for content-aware storage

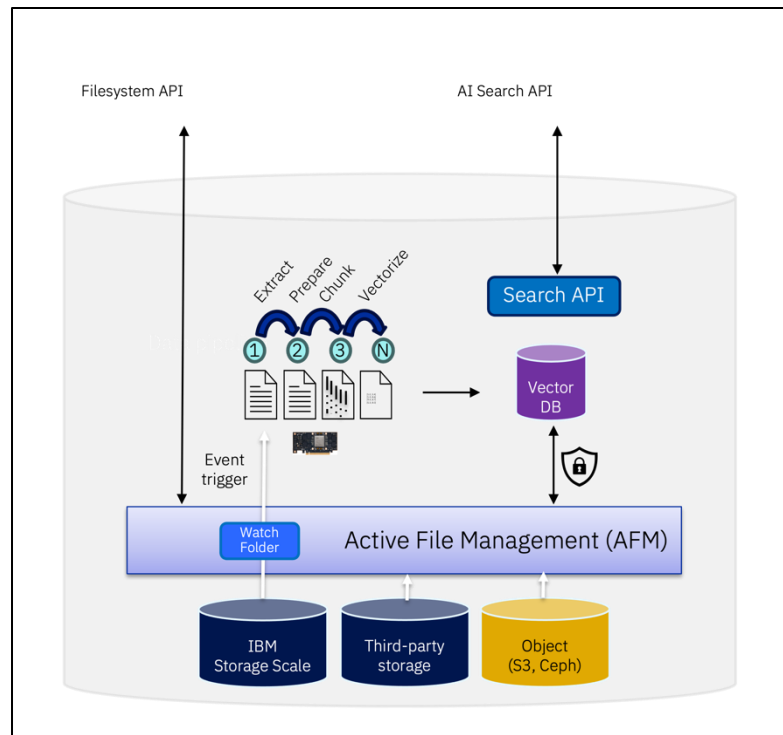
- Content-aware storage leverages:
  - AI optimized storage – IBM Storage Scale
  - AI data pipelines, such as NVIDIA NIMs
  - Vector databases and metadata model
  - Hardware accelerators (such as GPUs)
  - Storing and retrieving KV Cache from external storage



*Content-aware storage leverages AI storage and data processing pipelines*

# Advantages of IBM CAS: leave your data in place

- Storage Scale's active file management abstracts other storage systems, including legacy IBM and third-party systems
- Automatically detects data changes for incremental processing
- Enterprise grade security: no unnecessary data copies, persistent ACLs, consistent encryption



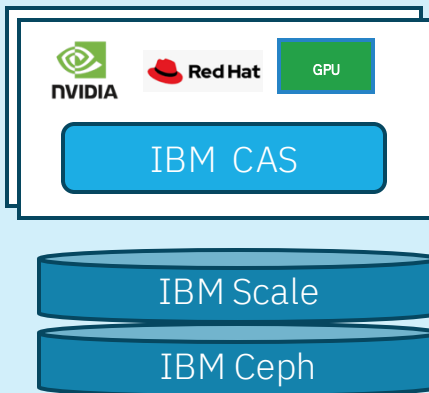
# IBM Fusion CAS : Hybrid Cloud inferencing storage services

(EDGE, Fusion HCI, GPU servers, clouds)

## IBM Fusion CAS– Single name space for distributed inferencing

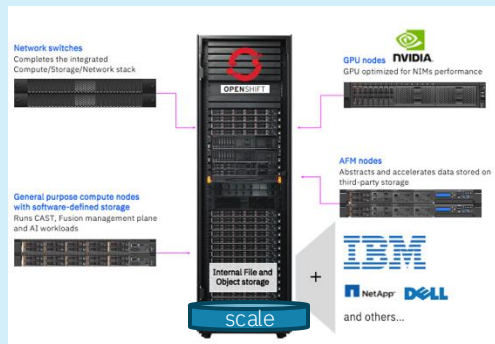
### Inferencing SDS

- Running CAS SDS on GPU servers on prem (including DGX, HGX)
- AI CSP: AI factories, Neo cloud

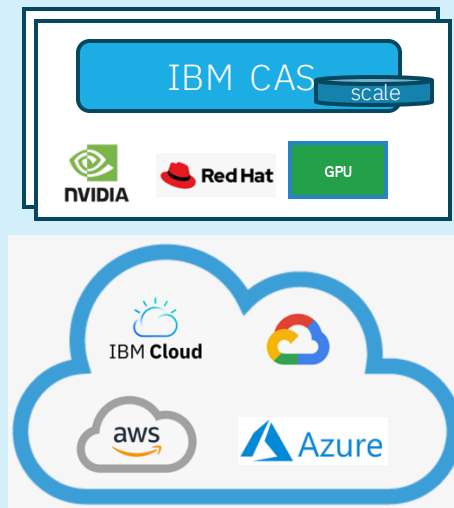


### Inferencing appliance

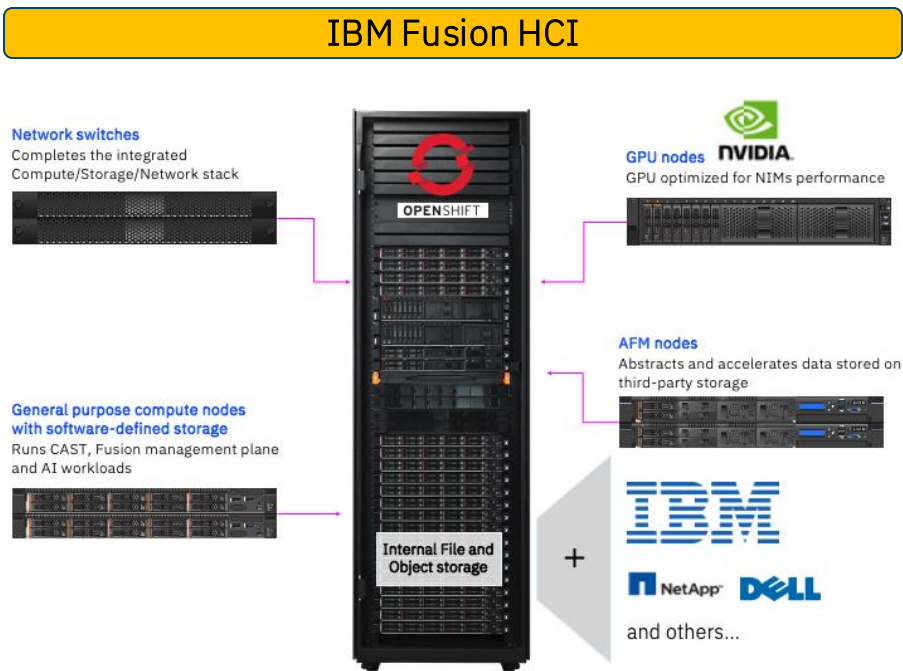
- Turn-Key inferencing server
- IBM Fusion HCI
- Server partners



### Inferencing Storage aaS



# IBM Fusion HCI delivers CAS as a turnkey inferencing appliance



## Simple

- Fully integrated with inferencing ecosystem
- Turn-key, all-in-one
- Zero to inferencing in weeks not months

## Efficient

- Connects to IBM and third-party storage
- Unleash data without copying/moving

## Enterprise-grade

- HA/DR/backup built-in
- Automated Day-2 operations
- Global data platform

# Use Case - Cyber Resiliency

“How do I *really* know when I recover my data that it’s a good copy and not corrupted?”

- Large Financial Services company

## PROVABILITY

- Lack of confidence that running a single tool truly ensures data integrity
- Manual verification is often required
- Organizations are unaware of how recoverable their mission critical data is



## DYNAMIC + COMPLEX ENVIRONMENTS

- Microservices (especially those with persistent storage) **complicate the boundaries and scope** of an application
- Updating of data integrity checks as **applications evolve**
- Ensuring **compatibility** of recovery environment with application/production



## SCALE OF RECOVERY

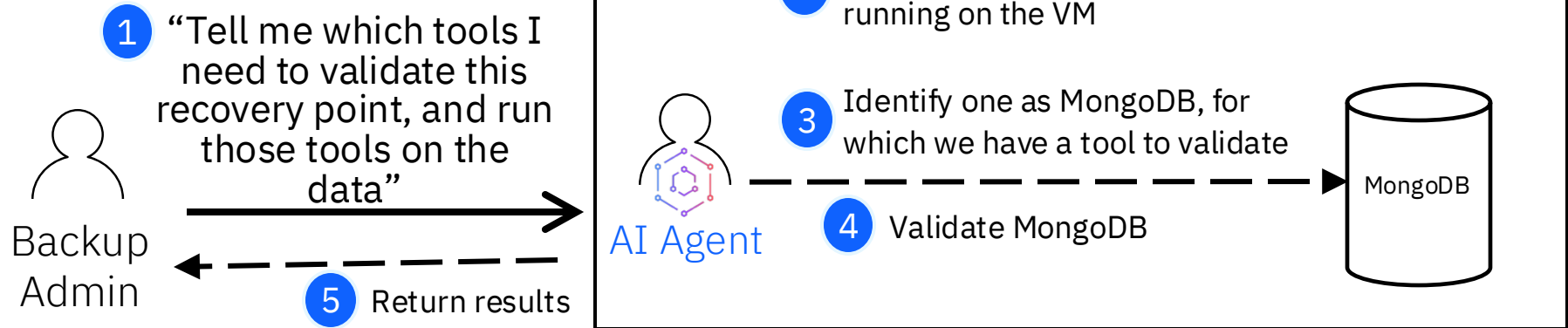
- Disconnect between **infrastructure** and **application** teams that need to collaborate on data recovery planning
- Constraints on **staffing and expertise**
- **Regulations** (e.g. DORA) require reports on provability of recovery
- Cyber Resiliency **recovery testing is not performed** (time, cost)



*All of this is expensive at scale, making it hard to prove to stakeholders that your data and applications are recoverable.*

# Agentic Use Case – Cyber Recovery Data Validation

## Interactive flow:



## Autonomous flow:

Schedule runs of the AI agent with preloaded prompts and email the Admin with summarized results



## Agentic AI in Action

- Agentic AI is emerging, and it has potential to change business operations in significant ways. Agentic AI will take some time to mature to multi-agent collaboration.
- Data and storage optimization has a significant impact to the efficiency, performance and accuracy of Agentic AI.
- IBM Content Aware Storage is optimized for Agentic AI platform from data ingestion to distributed inferencing.



