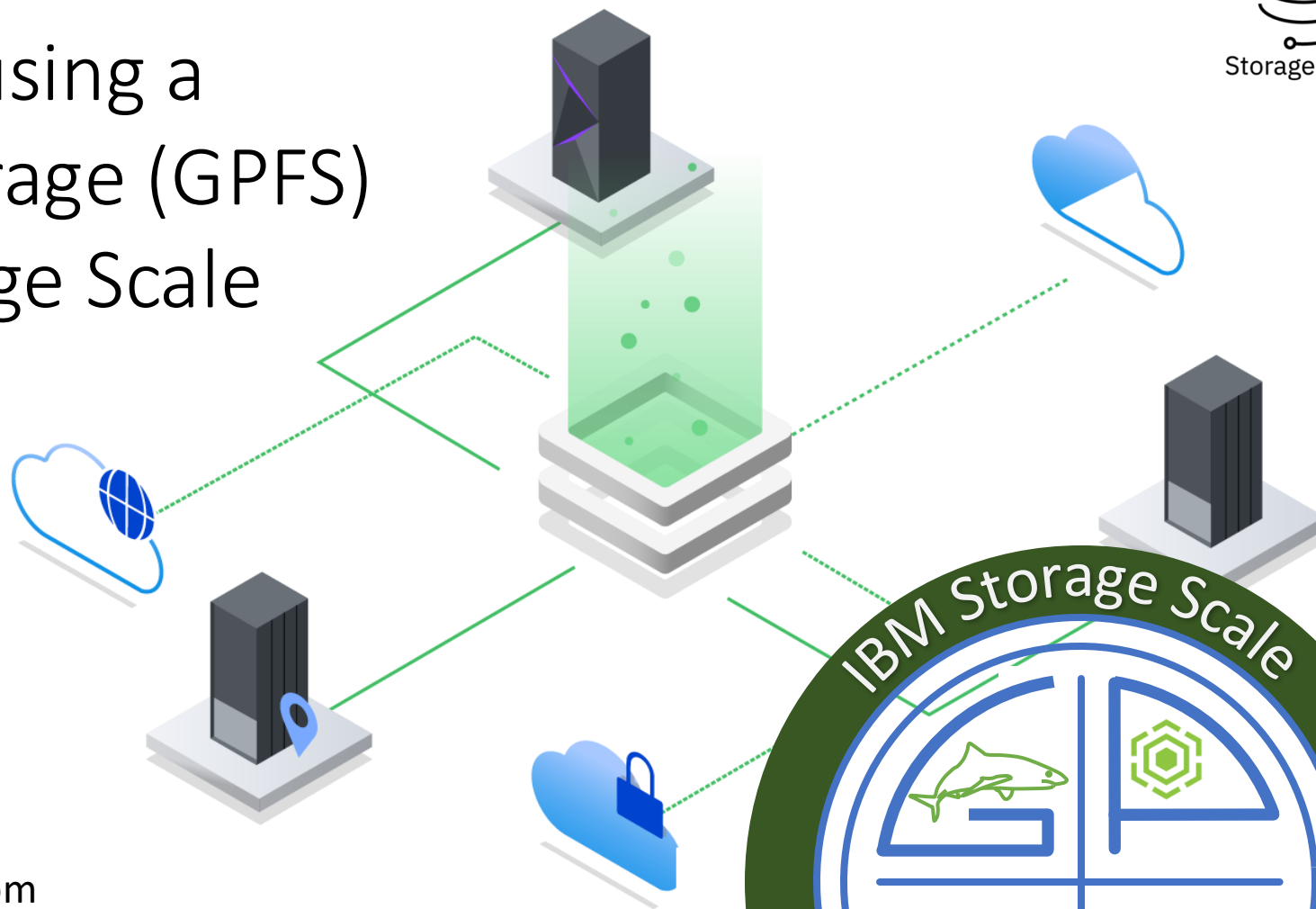


Data and AI workflows using a Global Platform For Storage (GPFS) Services with IBM Storage Scale



Jan Krol – jjkrol@ibm.com
Principal Storage Sales Specialist
Storage for Data and AI



Lindsay Todd, PhD – lindsay@us.ibm.com
Principal Storage Technical Specialist
Advanced Technology Group



Agenda



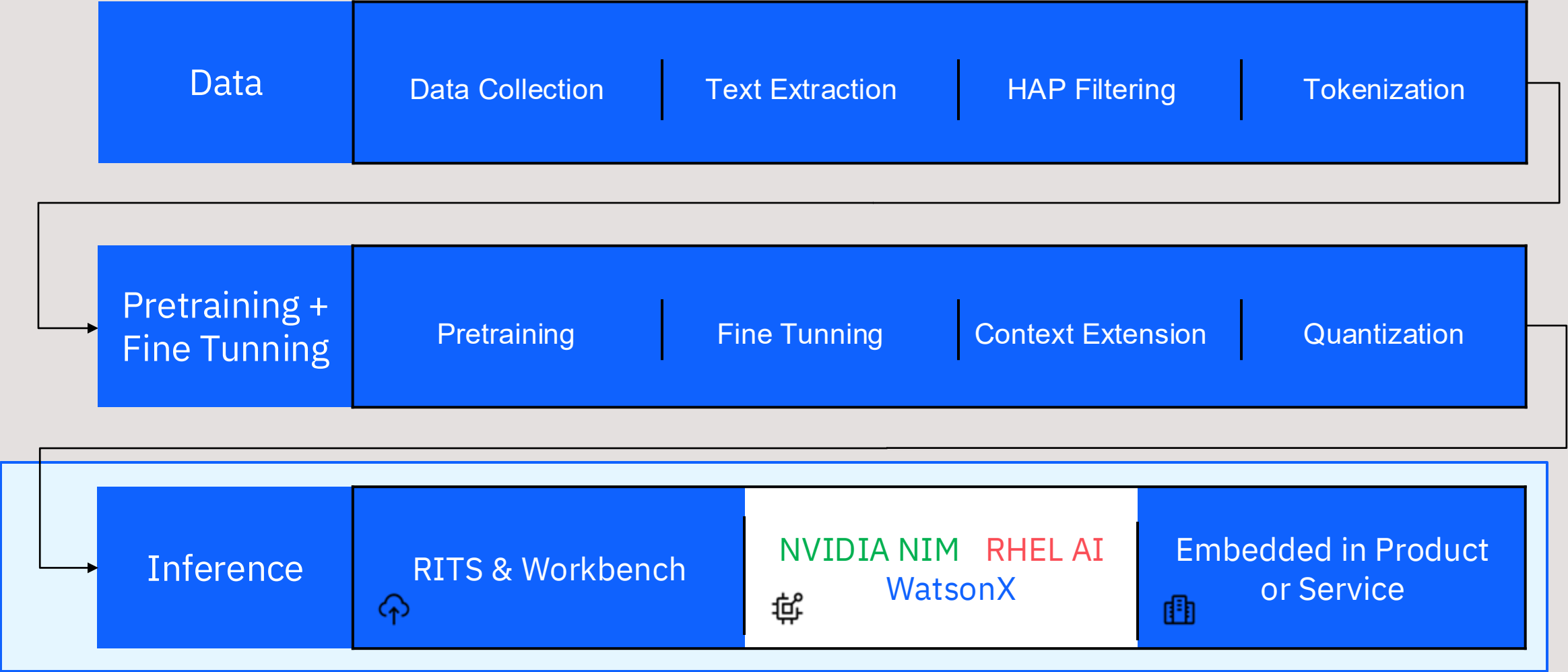
IBM

Inferencing



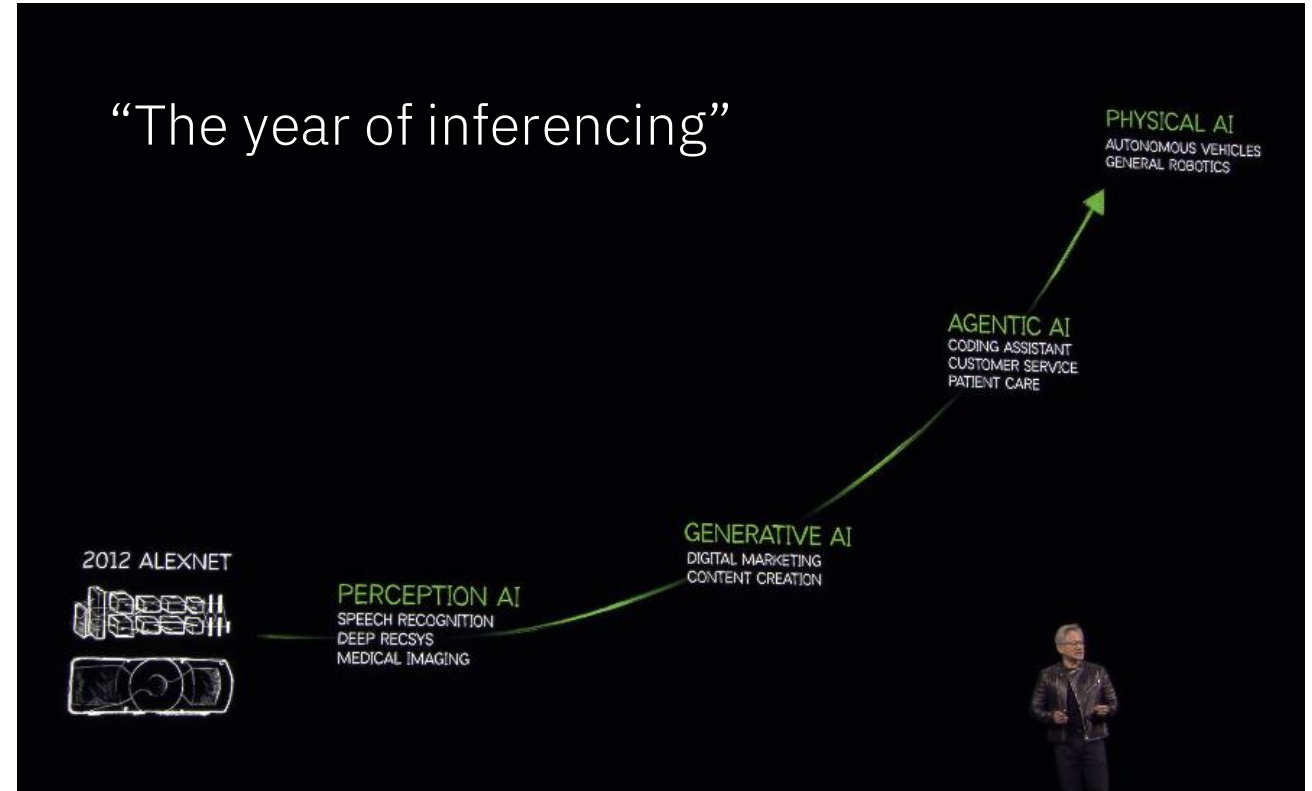
IBM.

AI Workflow - Inference



From training AI models to inferencing

- AI infrastructure requirements are shifting from *training* AI models to *running* them – inferencing.
- At NVIDIA GTC in March, Jensen Huang called 2025 “the year of inferencing.”



Two key elements to inferencing

1. Data ingestion process

- Natural language processing is used to parse, chunk, and embed the *meanings* of text.

2. Inferencing operations

- Queries are processed by a large language model, enhanced via retrieval augmented generation (RAG).

Why is it hard?

Scanned documents

Invoice account: 501-C02567
Terms of payment: End of Month + 60 days
Due date: 31/01/2018
Method of payment: 501-PS047141 from 28/11/2017
Delivery: 501-PS047141 from 28/11/2017

Order account: 501-S00479
Order number: 501-S00479
Customer: ARROW FINANCE HQ
VAT Reg No: NA-GB-230

Invoice address: Gulf Business Machines Abu Dh
W.L.L. P.O. Box 37543
Abu Dhabi
United Arab Emirates

Delivery address: Gulf Business Machines Abu
W.L.L. P.O. Box 37543
Abu Dhabi
United Arab Emirates

Invoice No: 501-SPI045340
Tax Point Date: 28/11/2017

Item number	Description	Quantity	List price	Discount %
ARW_INT_BM_SFS3	IBM FlashSystem V9000 Storage Implementation	1.00	2,200.00	

Sales tax code: Serv 123456789, 3rd city
Amount origin: 2,200.00
VAT amount: 0.00

EXPORT SERVICES

Handwritten notes: GIBP/11/320.40, 20/800/800, S. Arin

Noise artifacts

Variety of tables

Fig.7: Sensitivity to price and volume movements

	Net Profit	
	FY14F	FY15F
For each +/-1% chg in coal price		
Indo Tembung	-0.40%	-1.89%
Adaro	-0.50%	-1.83%
Hulu	-1.00%	-1.83%
For each +/-1% chg in volume		
Indo Tembung	-1.00%	-1.10%
Adaro	-1.10%	-1.10%
Hulu	-1.00%	-1.20%

	Previously
FCCR (limitation on indebtedness)	3x
Debt/EBITDA	4x
Bekasi Power debt basket	5mn
General debt basket	10mn
Permitted investment basket	-
Interest reserve account establishment	Yes

Table with visual clues only

Table with graphic lines

Multi-row, multi-column headers

	Three months ended September 30			N
(in thousands)	2015	2014	Change	
Salaries and benefits	171,506	153,123	15.1%	
Employee share purchase plan	24,101	19,879	21.0%	
Employee profit share	32,974	19,039	73.2%	
Share-based payment plans	4,123	4,348	(5.2%)	
Total	230,786	196,499	17.4%	

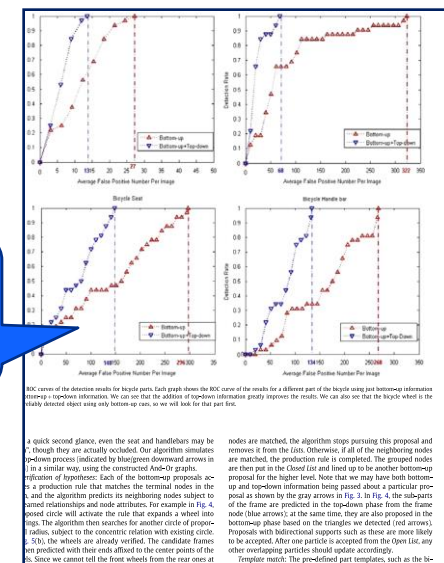
Presentation on the Condensed Consolidated Statement of Earnings:

	2015	2014	Change
Airport operations	30,916	28,145	9.8%
Tighten conditions and average pool changes	67,546	59,251	13.8%
Sales and distribution	10,496	17,312	39.4%
Marketing, general and administration	27,063	22,818	18.6%
Infra	36,116	33,828	6.8%
Maintenance	15,737	15,256	3.2%
Employee profit share	32,974	19,039	73.2%
Total	230,786	196,499	17.4%

Nested row headers

Tables with Textual content

Various elements



Line Charts
Histogram
Pie Charts

ORACLE

SERVICES AGREEMENT

The Services Agreement (the "Agreement") is between Oracle Corporation, Redwood City, California 94065 ("Oracle") and IBM Corporation, Kingston, NY 12401 ("IBM").

I. Services

Oracle will provide to Client, in the United States, the Services specified on a Work Order, under the terms of this Agreement.

II. Definitions

2.1. "Work Order" shall mean Oracle's standard form for ordering Services (entitled "Work Order" or "Order Form") and shall specify the Services and applicable fees. Each Work Order shall be governed by the terms of this Agreement and shall reference the Effective Date specified below.

2.2. "Services" shall mean work performed by Oracle for Client pursuant to a Work Order, agreed to by the parties, under this Agreement. The schedule for Services will be agreed upon by the parties, subject to availability of Oracle personnel.

III. Charges, Payment, and Taxes

3.1. Fees for Services

Unless otherwise expressly specified in the applicable Work Order, Services shall be provided on a time and material ("T&M") basis at Oracle's T&M rates current when the Services are performed. If a dollar limit is stated in the applicable Work Order for T&M Services, the limit shall be deemed an estimate for Client's budgeting and Oracle's resource scheduling purposes; after the limit is expended, Oracle will continue to provide the Services on a T&M basis, if a Work Order for continuation of the Services is signed by the parties.

3.2. Incidental Expenses

Client shall reimburse Oracle for reasonable travel, communications, and out-of-pocket expenses incurred in conjunction with the Services.

Fuzzy Text & Skew

Per capita poultry consumption

Country	Chicken consumption (kg/capita/year)	GDP
Malaysia	37.3	
Singapore	36.2	
Thailand	12.5	
China	9.2	
Philippines	8.4	
Vietnam	7.2	
Indonesia	6.1	
India	2.3	

Colored background

Level	Description	Consolidation	Third Party Property	Third Party Property
Stream or flood	you had entered into a contract of sale to purchase your vehicle, or contract to sell your vehicle, and the vehicle was damaged by the flood.	✓	✓	✓
Third party attempted theft	We will cover your vehicle if the loss is caused by a third party who attempted theft. We will cover the loss if the vehicle was damaged by the flood.	✓	✓	✓

Signatures

Logos

WAIVER OF BIO REQUEST

Signature of [Name] [Date]

Signature of [Name] [Date]

Parsing example

S3 Deep Archive Storage on Diamondback



Reduce the Cost of Archiving Data, Without Impacting Data Access

IBM S3 Deep Archive on IBM Diamondback provides S3 Glacier storage class access at up to 85% savings compared to AWS S3 Glacier hosted storage.

■ Highlights

Interoperable with all S3 Glacier Flexible Retrieval storage class supporting applications.

No tape skills required.

Flexible Retrieval performance at a fraction the cost of S3 Deep Archive storage classes.

The rising costs of storing data continues to be a challenge. "many data center managers will be forced to use tape..." according to Furthur Market Research, "as ultra-low-cost, sustainable storage alternatives."¹ Some data center managers resist tape as difficult to deploy and operate due to the need for specific tape software and skills. IBM S3 Deep Archive flips the rhetoric with simple deployment and a standardized interoperable interface, without sacrificing the ultra-low-cost storage position.

IBM S3 Deep Archive brings S3 Glacier Deep Archive Flexible retrieval classes on-premises at a fraction of the cost of even AWS S3 Glacier Deep Archive. Approximately 80% of organizations have S3 skills.² IBM S3 Deep Archive is delivered configured and ready for deployment as an S3 target for all deep archive data.

- S3 low-cost, secure, durable availability zone
- No supporting application modifications
- Subscription like capacity, without the monthly charges
- Zero egress fees

Table 1. Total Acquisition Cost Compare

	IBM S3 Deep Archive	
	9PB	27PB
Total Acquisition Cost per TB	\$22.56	\$11.89
5-year \$/TB year	\$4.51	\$2.37
AWS S3 Glacier Flexible Retrieval \$/TB year	\$43.20	
Savings compared to AWS	89%	94%

Unstructured text

S3 Deep Archive Storage on Diamondback

Reduce the Cost of

Archiving Data, Without

Impacting Data Access

T56000

IBM S3 Deep Archive on IBM Diamondback

provides S3 Glacier storage class access at up

to 85% savings compared to AWS S3 Glacier

hosted storage.

Highlights

Interoperable with all S3 Glacier Flexible Retrieval

storage class supporting applications.

No tape skills required.

Flexible Retrieval performance at a fraction the cost of S3 Deep Archive storage classes.

The rising costs of storing data continues to be a challenge. "many data center managers will be forced to use tape..." according to Furthur Market Research, "as ultra-low-cost, sustainable storage alternatives."¹ Some

data center managers resist tape as difficult to deploy and operate due to the need for specific tape software and skills. IBM S3 Deep Archive flips the rhetoric with simple deployment and a standardized interoperable interface, without sacrificing the ultra-low-cost storage position.

IBM S3 Deep Archive brings S3 Glacier Deep Archive Flexible retrieval classes on-premises at a fraction of the cost of even AWS S3 Glacier Deep Archive. Approximately 80% of organizations have S3 skills.² IBM S3 Deep Archive is delivered configured and ready for deployment as an S3 target for all deep archive data.

- S3 low-cost, secure, durable availability zone
- No supporting application modifications
- Subscription like capacity, without the monthly charges
- Zero egress fees

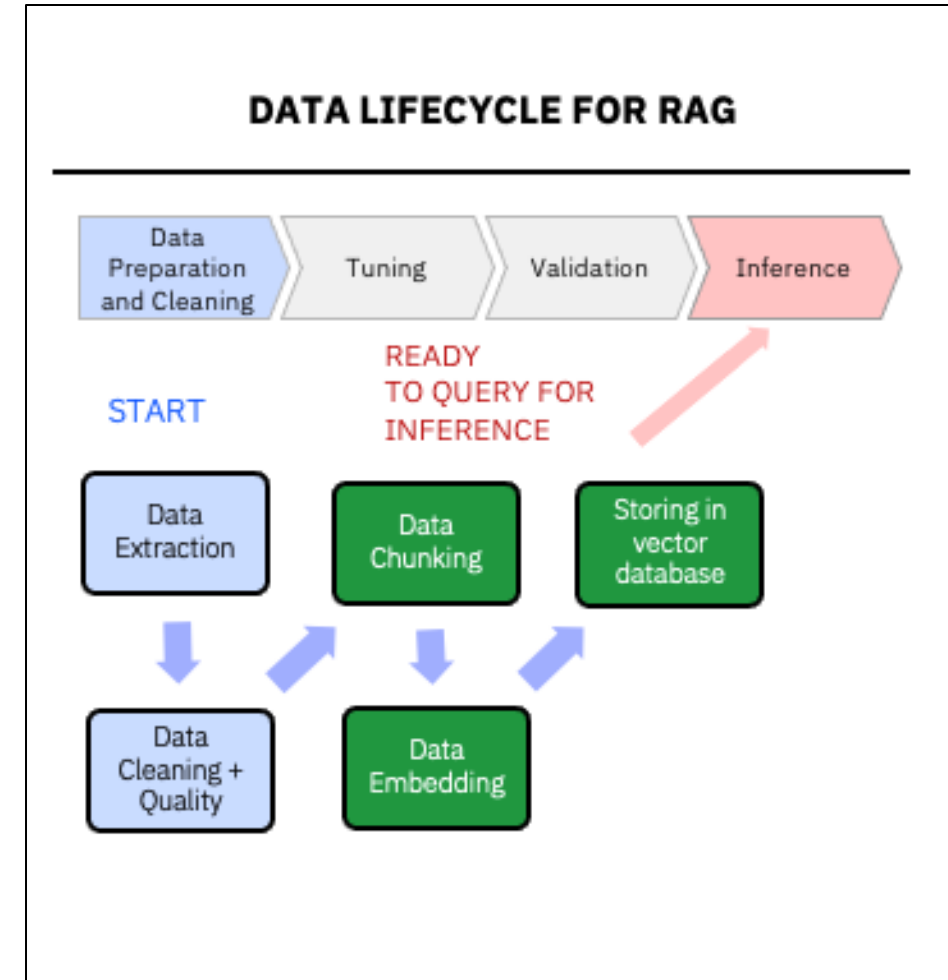
Table 1. Total Acquisition Cost Compare

Table with the following rows:
Total Acquisition Cost per TB, IBM S3 Deep Archive: 9PB = \$22.56, IBM S3 Deep Archive: 27PB = \$11.89
5-year \$/TB year, IBM S3 Deep Archive: 9PB = \$4.51, IBM S3 Deep Archive: 27PB = \$2.37
AWS S3 Glacier Flexible Retrieval \$/TB year, (IBM S3 Deep Archive: 9PB, IBM S3 Deep Archive: 27PB) = \$43.20
Savings compared to AWS, IBM S3 Deep Archive: 9PB = 89%, IBM S3 Deep Archive: 27PB = 94%

Table

Limitations in existing RAG solutions

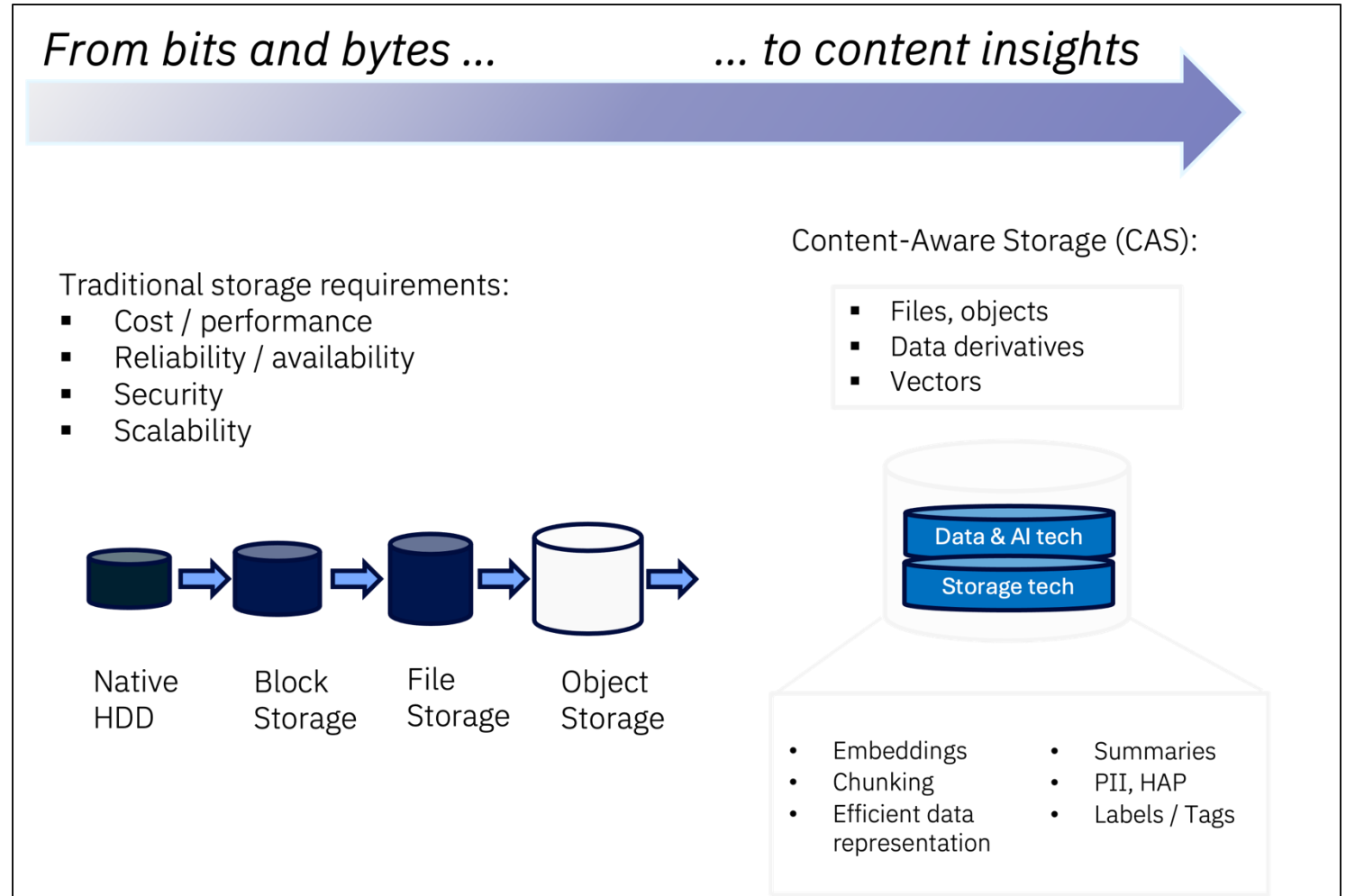
- Only a fraction of enterprise data is indexed.
- It's typically indexed daily or weekly in batch mode, not real time.



The solution? Content-aware infrastructure

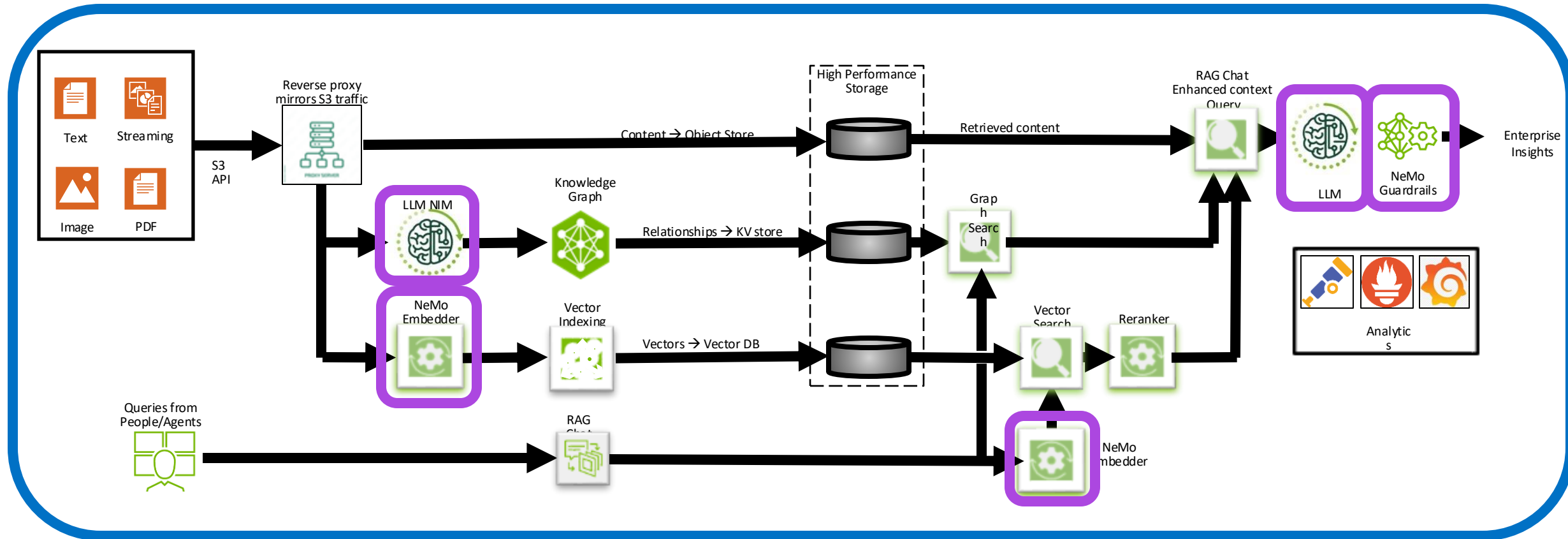
- A new paradigm that leverages bringing vector processing closer to the storage layer.
- From “bringing data to AI” to “bringing AI to data”.

Scale can use one of its introspection capabilities – “Cluster Watch Folders” – to automatically trigger tokenization and indexing of documents and artifacts as they are ingested.



Storage Appliances to Become AI Data Platforms

Simultaneously Store, Embed, Index, Graph - Makes Information Instantly Searchable Findable



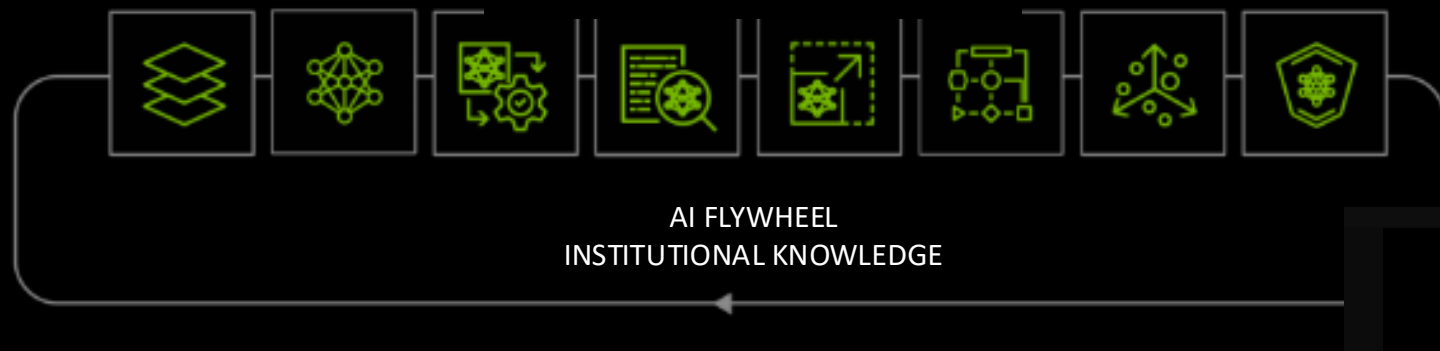
NIMs

Blueprint

NVIDIA NIM Optimized Inference Microservices

Accelerated runtime for generative AI

NVIDIA NIM BLUEPRINTS

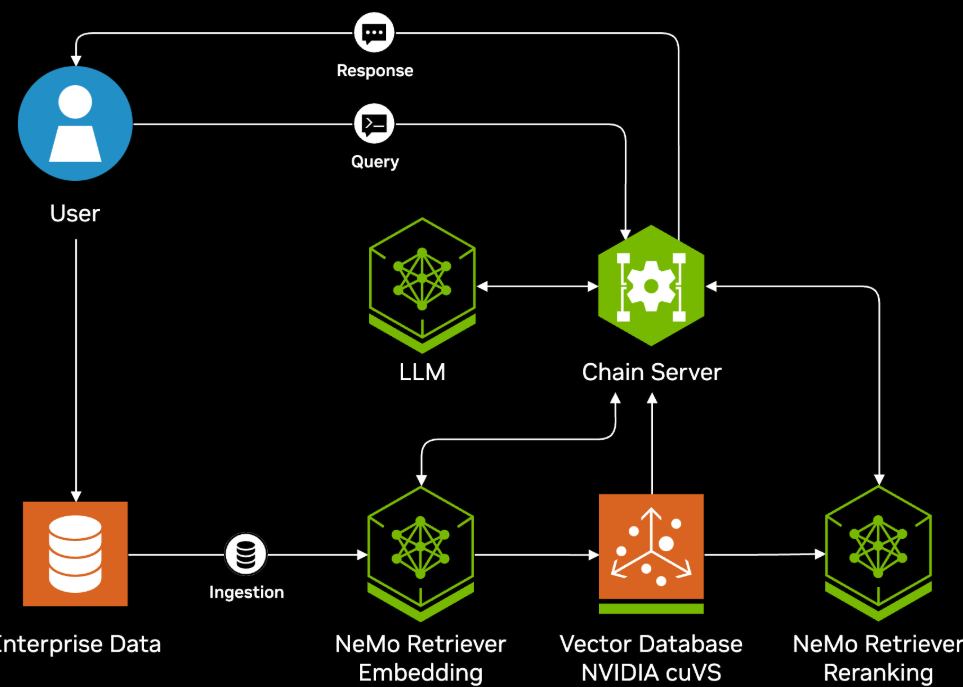


able Deploy anywhere with security and control of AI applications and data

Support for custom models

Domain specific code

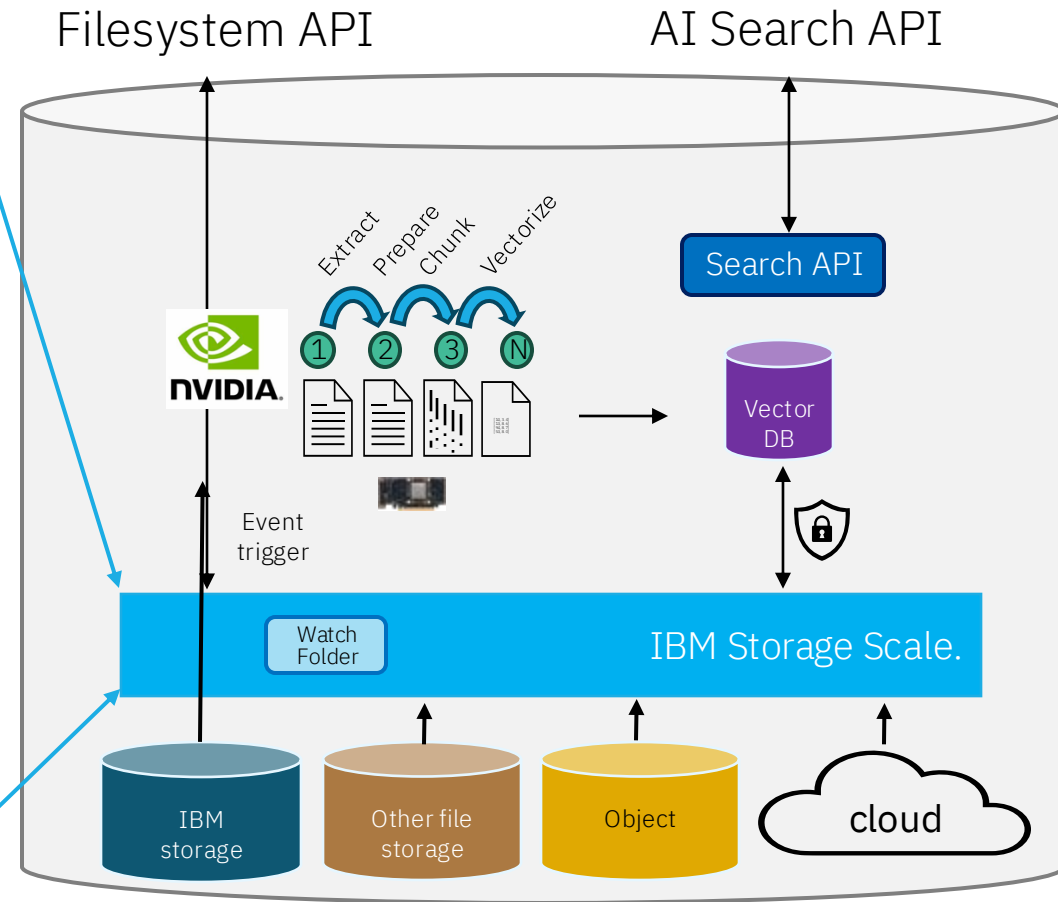
Optimized inference engines



IBM AI optimized storage provides CAS differentiation

IBM AI optimized storage and AI runtime

- **Enterprise grade secure:** Consistent ACL and encryptions
- **Efficient:** Detect data change for incremental data processing
- **Support your legacy storage:** Connect to heterogeneous storage systems, including legacy unstructured data storage
- **Accelerate and scale:** GPU-optimized storage solution

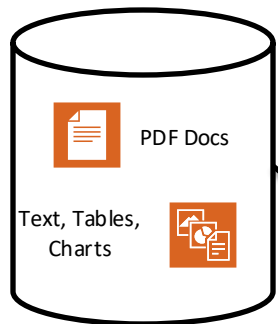


Example : IBM AI optimized storage provides CAS differentiation

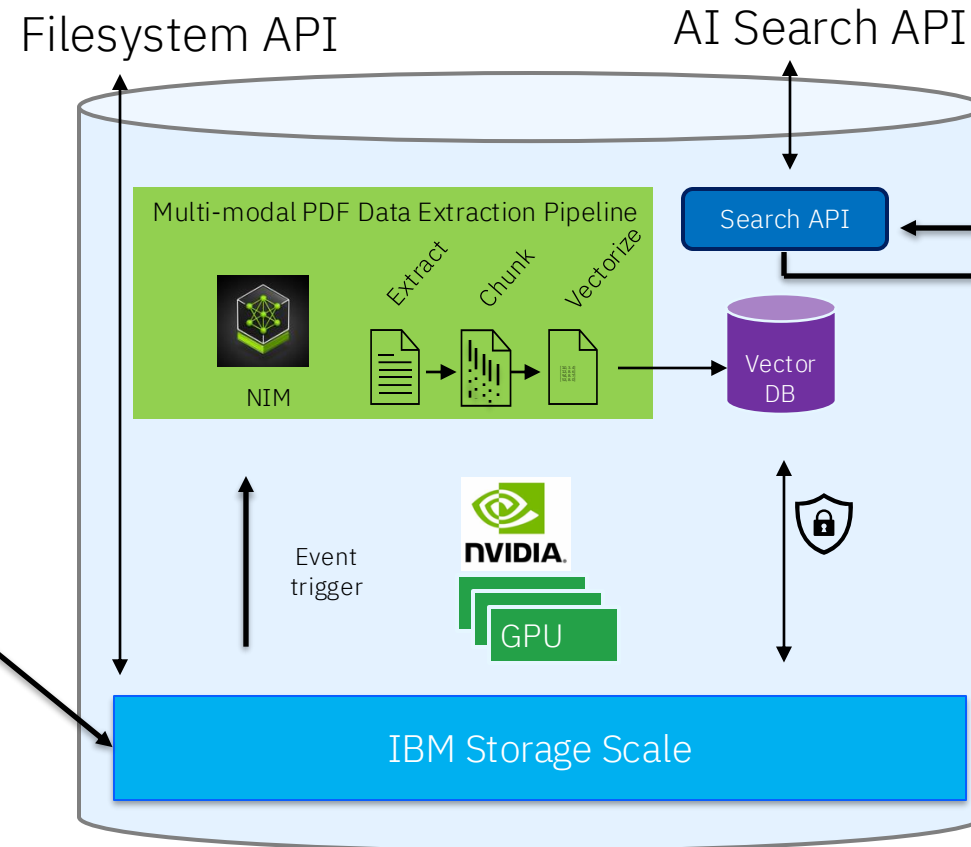
- ✓ **Simple:** Enable existing Enterprise data for use with all legacy storage into RAG storage. No data copy needed
- ✓ **Efficient:** Optimize RAG processing by 10x or more by only ingesting and processing changed data for use with AI virtual assistants
- ✓ **Secure:** Enforce data source access control in RAG pipeline

1. Customer imports documents into their secure S3 bucket (no changes required)

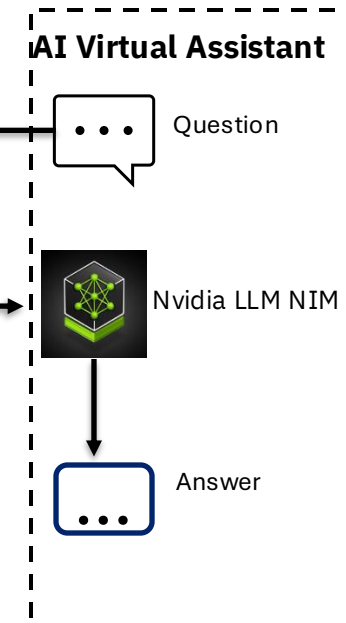
Existing Enterprise S3 Buckets



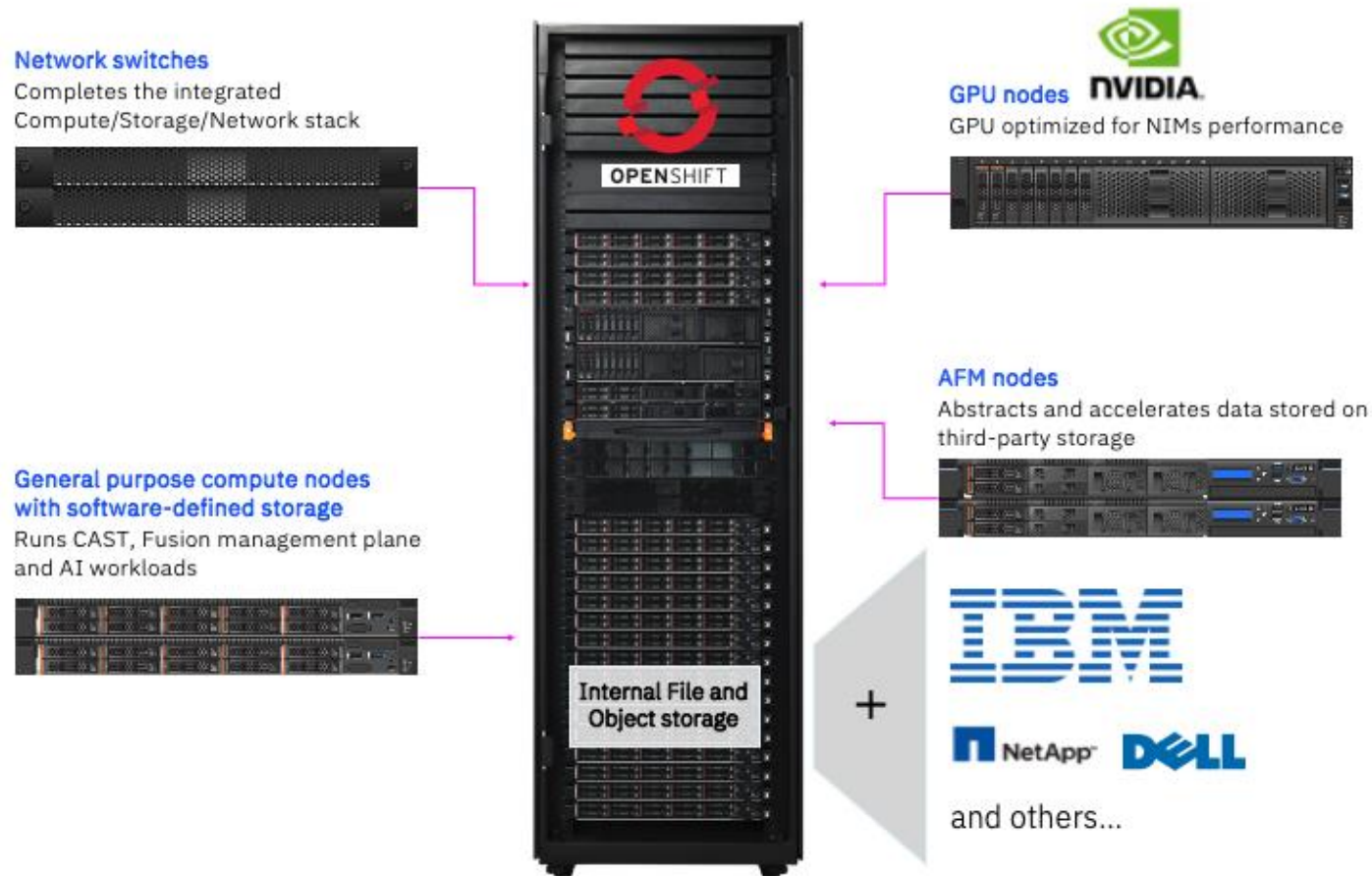
2. Data is automatically processed for use in virtual assistant



3. **Instantaneously**, virtual assistant can answer questions on the newly imported documents



IBM has integrated CAS with IBM Fusion HCI to provide a turn-key engineered experience



Simple

- Turn-key, all-in-one
- Zero to inferencing in weeks not months

Works w/ legacy data

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

Enterprise grade

- HA/DR/backup built-in
- Automated Day-2 operations

BACKUP

Archive

and

Data-resiliency



IBM.

Trends driving tiered data storage requirements

3100%

There has been an eruption in the amount of data created annually versus a decade ago.¹



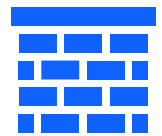
149

Exponential increase in the number of foundation models released.²



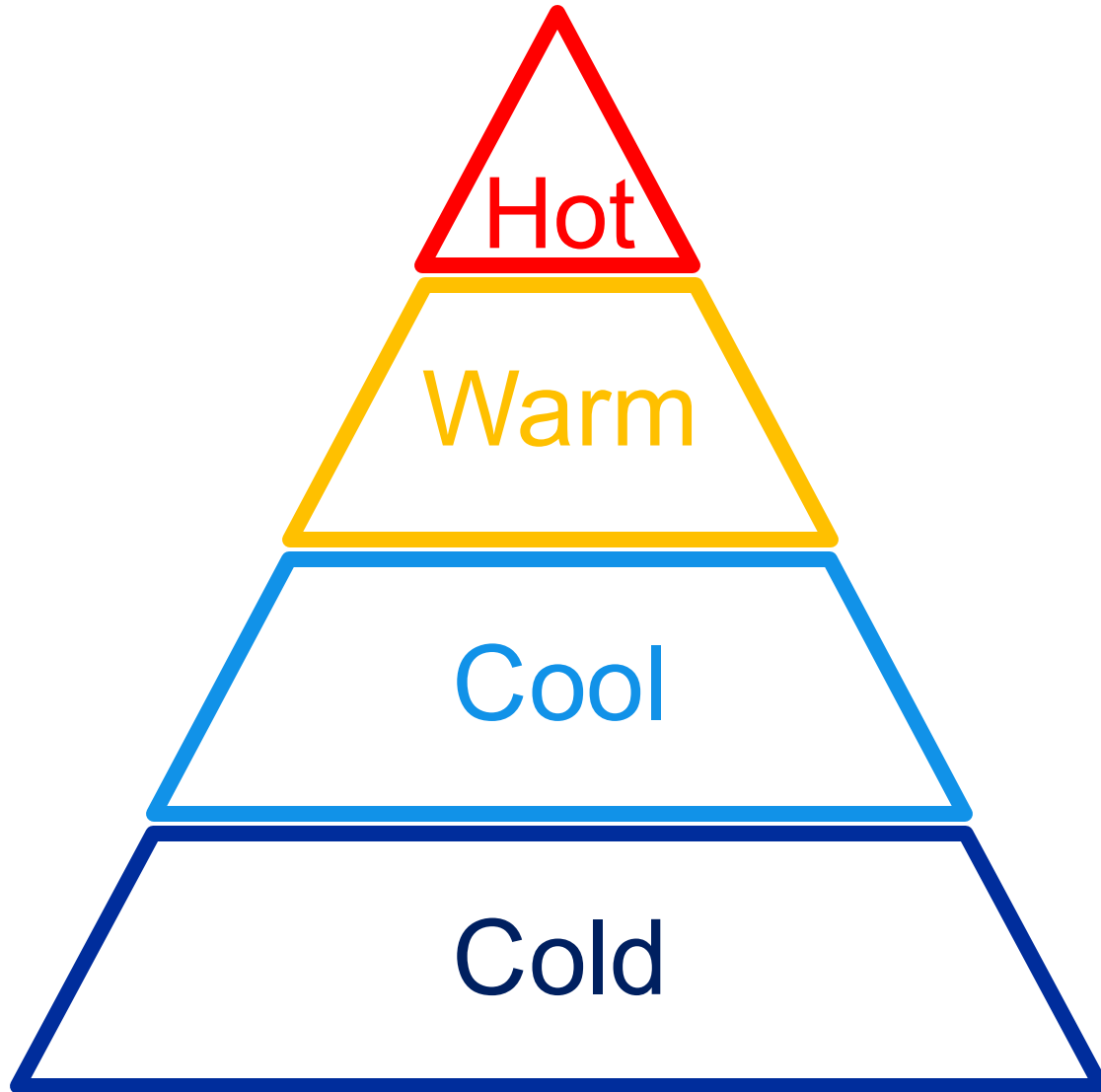
Factory

The concept of an 'AI Factory' relies upon the gathering and aggregation of a range of internal and external data sources.



Speed isn't
always the name
of the game.

Covering Every Tier with IBM Storage



High-performance SSS 6000 delivering 1PB usable NVMe and 310 GB/s per 4U for AI inferencing and training

Capacity model SSS 6000 with HDDs delivering up to 90GB/s with automated tiering for tokenization and ready access

Storage Ceph providing scalable object storage for data extraction, collection, and aggregation

Diamondback securing 27PB per rack for cloud-like deep archive S3 storage ready to recall

Data Resiliency – because things go wrong!

Data Resiliency is the ability to keep all the data available that an organization needs to continue functioning, even in the face of disruptions such as:

- Natural disasters
- Human-caused disasters
- “Oops!”
- Cyberattacks

Disasters and cyberattacks have different characteristics, leading to different recovery strategies.

Category	Disaster Recovery	Cyberattack Recovery
Recovery Point	Known Point in time	Not known...yet
Recovery Time	RPO/RTO	Trusted and Verified first
Nature of Disaster	Flood, power outage, weather	Targeted
Impact of Disaster	Regional	Global
Recovery	Failback	Based on situation
Data to Recover	Known	Unknown
Topology	Connected Data Centers	Isolated and away from production
Data Volume	Comprehensive, all data	Super selective
Probability	Low	High

Oops!

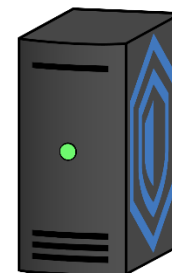
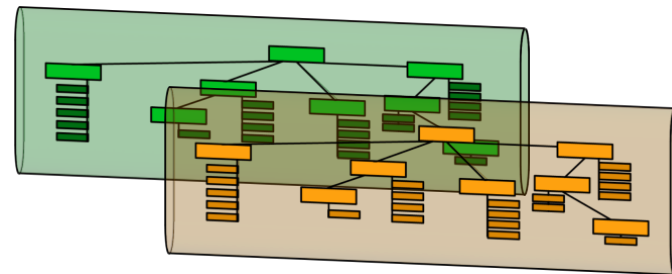
Backing up the Scale File System

Backing up the user data of a Scale file system requires also backing up the metadata associated with each file and directory:

- Create, access, and modify times
- Size of the file
- Mode bits and owner, owning group
- Immutability attributes
- Access Control Lists
- Extended attributes

IBM Storage Protect does incremental backups and stores all these metadata, and Scale can leverage its parallelism to speed this up.

Snapshots provide a readonly replica of the file system, useful for backups or self-service recovery.



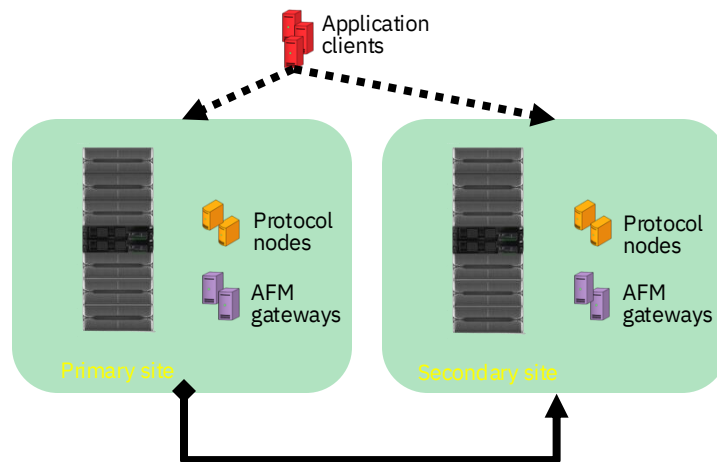
Replication – because disasters happen

A “**stretched cluster**” provides active/active synchronous replication for high availability.

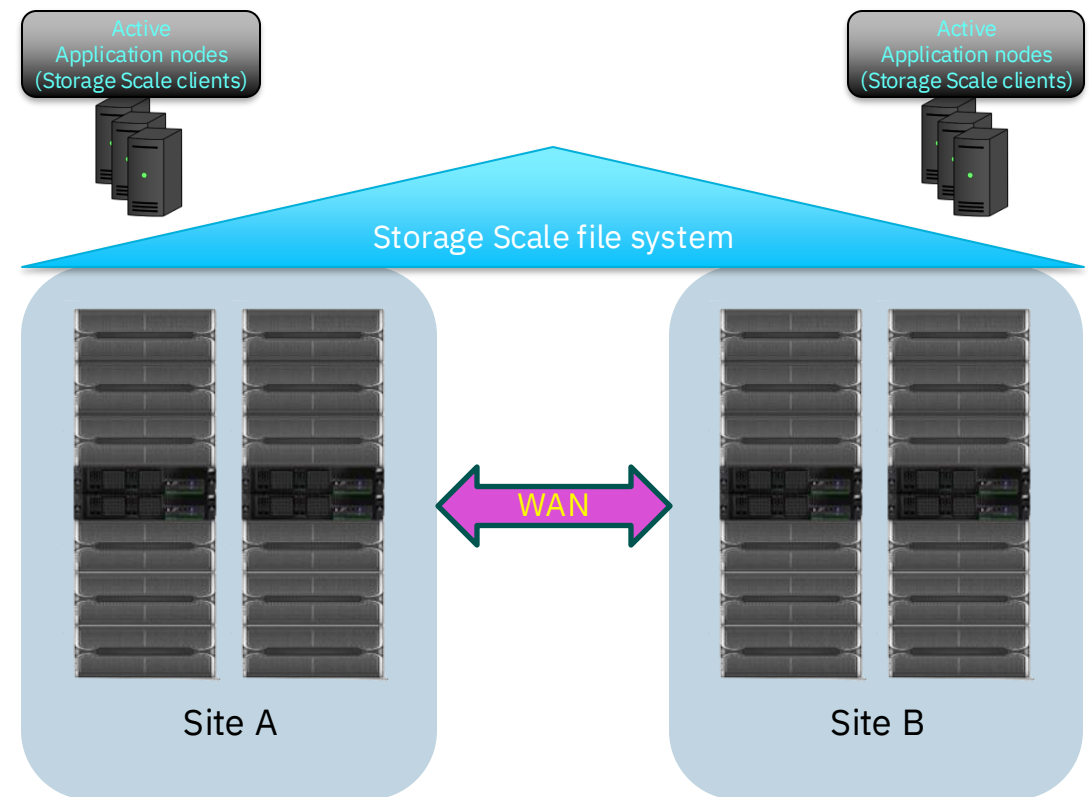
- Latency impact of WAN and replicated writes
- Design network with readReplicaPolicy (default, local, fastest) in mind

Active File Management (AFM) for DR provides an active/passive asynchronous replication.

- Which AFM mode to use? (RO, LU, SW, IW, DR)
- RPO for DR (can we make it on time...)



AFM-DR for active/passive Disaster Recovery



Stretched cluster architecture for highly-available active/active resiliency

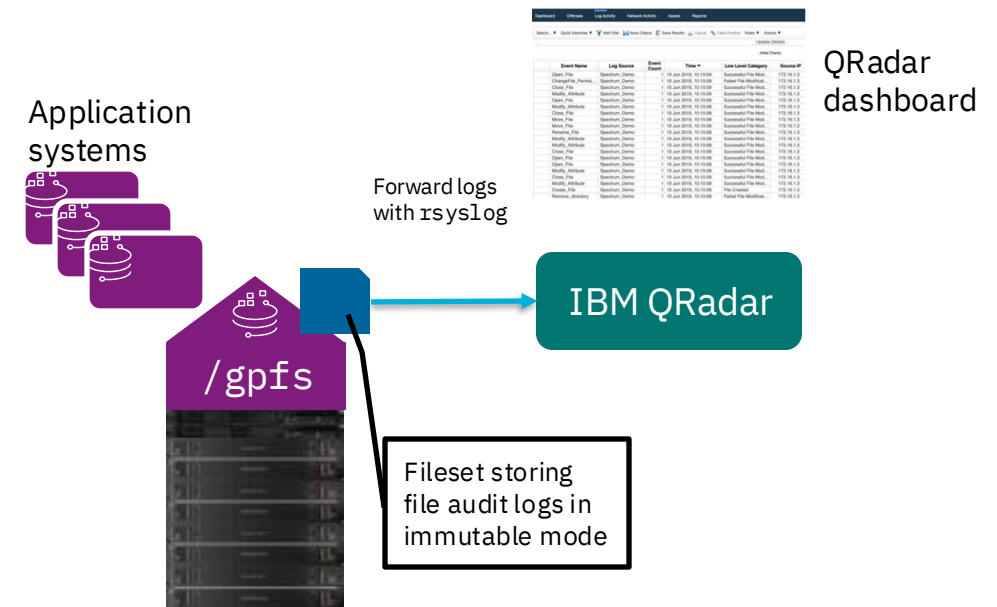
Auditing and Watch Folders – because intruders happen

Storage Scale's File Audit Logging feature records file access events, with audit information, into a JSON files stored in an immutable (append-only) fileset.

Scale's Watch Folders queue file access events to a Kafka topic.

These events can then be parsed by other tools, perhaps a SIEM such as IBM QRadar. These tools can report and react to unusual activity.

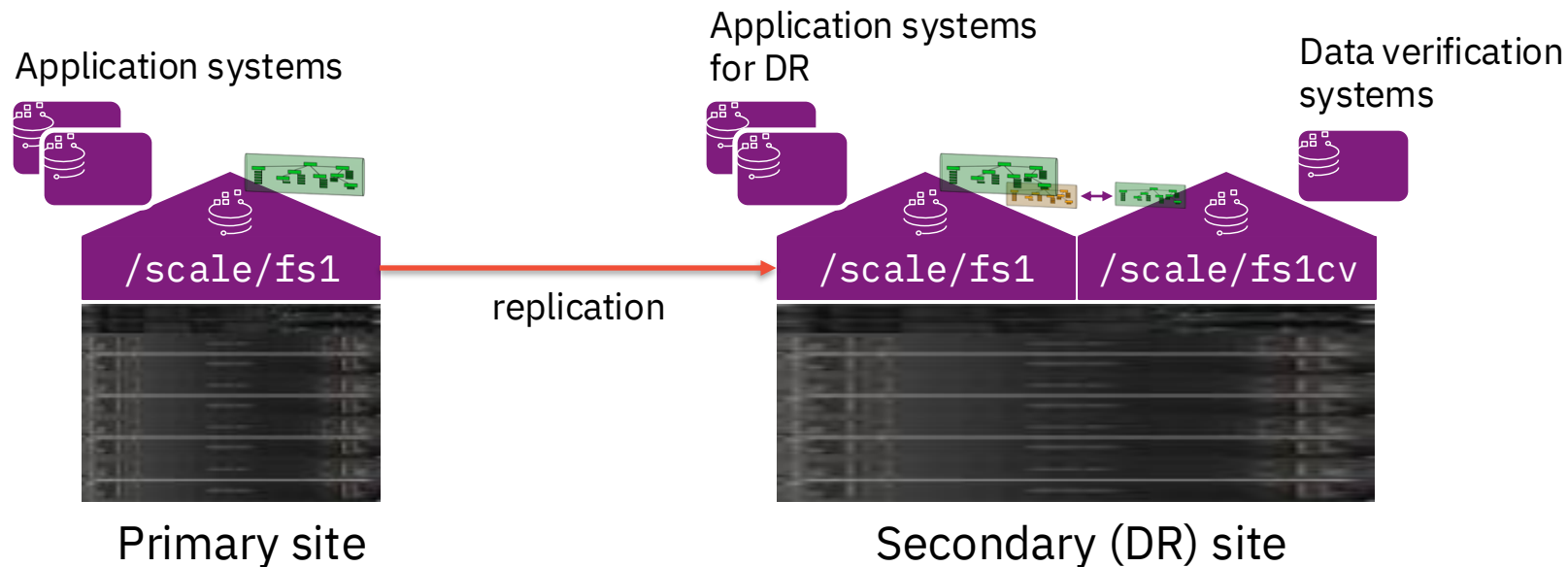
- For example, QRadar could be configured to automatically unlike filesets being attacked, or perhaps change file system ACLs, if an unexpected file access is detected.



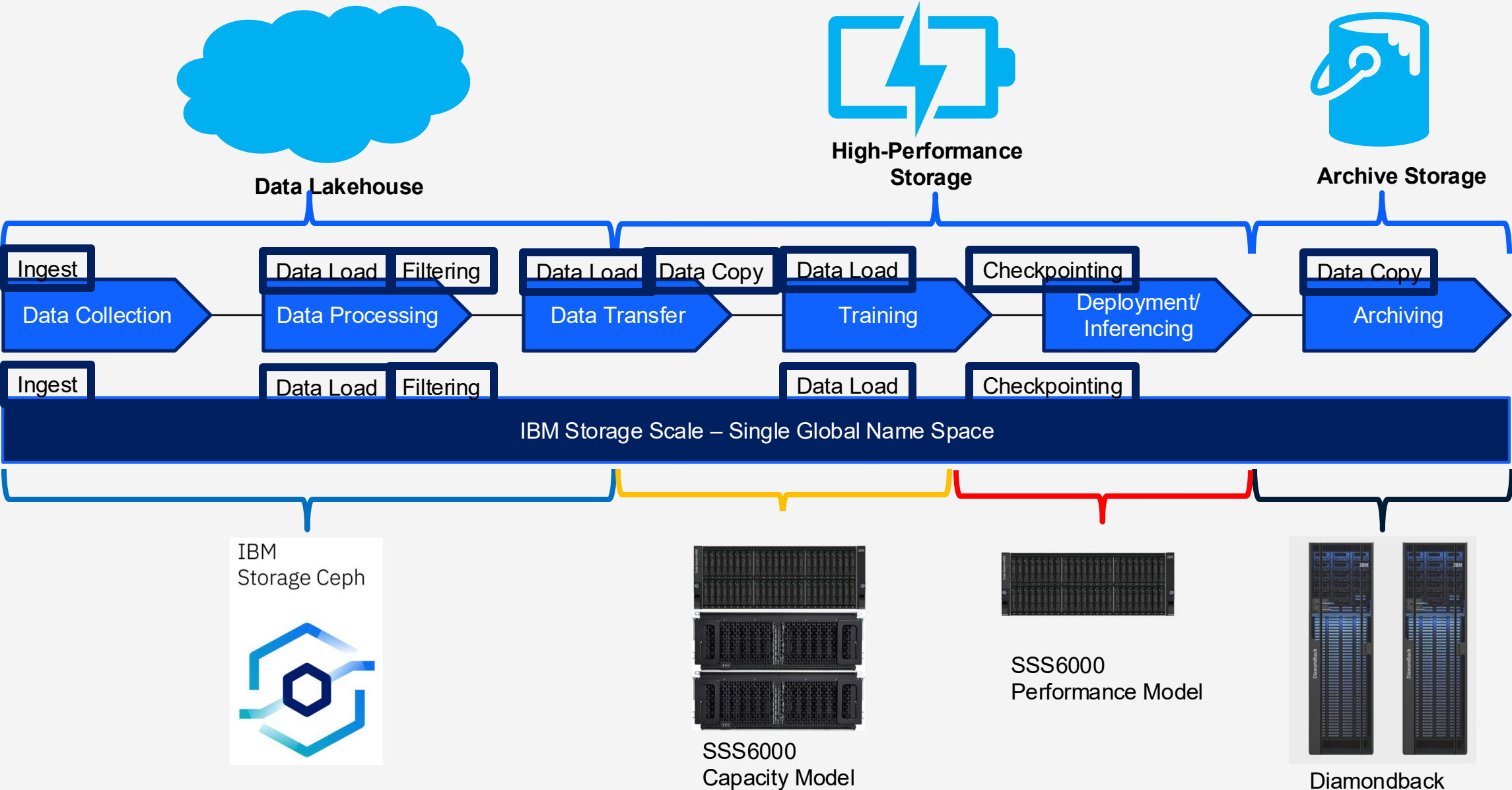
Using AFM-DR to build an air-gapped cyber vault – because cyber-events happen

AFM-DR creates a read-only copy of the primary data at the secondary site.

AFM-LU caches can be made of peer snapshots or Safeguarded Copies of the secondary fileset.



Data Management Architectures



Hands-on Storage Scale and Scale System Workshop – August 20-21

If you are actively considering Storage Scale for your organization, come join us for this very technical interactive workshop designed to let you:

- Gain a deep overview of IBM Storage Scale and the Storage Scale System – how it works, how to use it.
- Dive into advanced features and functions, including:
 - Information Lifecycle Management (ILM)
 - Cluster Export Services (NFS, SMB, S3)
 - Replication strategies for caching, HA, and DR
 - Networking best practices
 - Using storage-rich servers
 - Data-resiliency with Scale
 - Content-aware storage
 - Architecting Storage Scale solutions to solve your business problems
- Participate in hands-on labs

August 20-21, 2025
IBM Silicon Valley Lab (SVL)
Executive Briefing Center,
555 Bailey Avenue
San Jose, CA 95141

Customers and Business Partners: If interested, please speak to your IBM representative and see if they can nominate you to attend.

Business Partners and IBMers: Nominate your clients at <https://ibm.biz/Bdnw28>

Questions about workshop content: Reach out to Lindsay Todd, lindsay@us.ibm.com

Strategic technical collaboration between IBM storage and NVIDIA

