

IBM Content Aware Storage

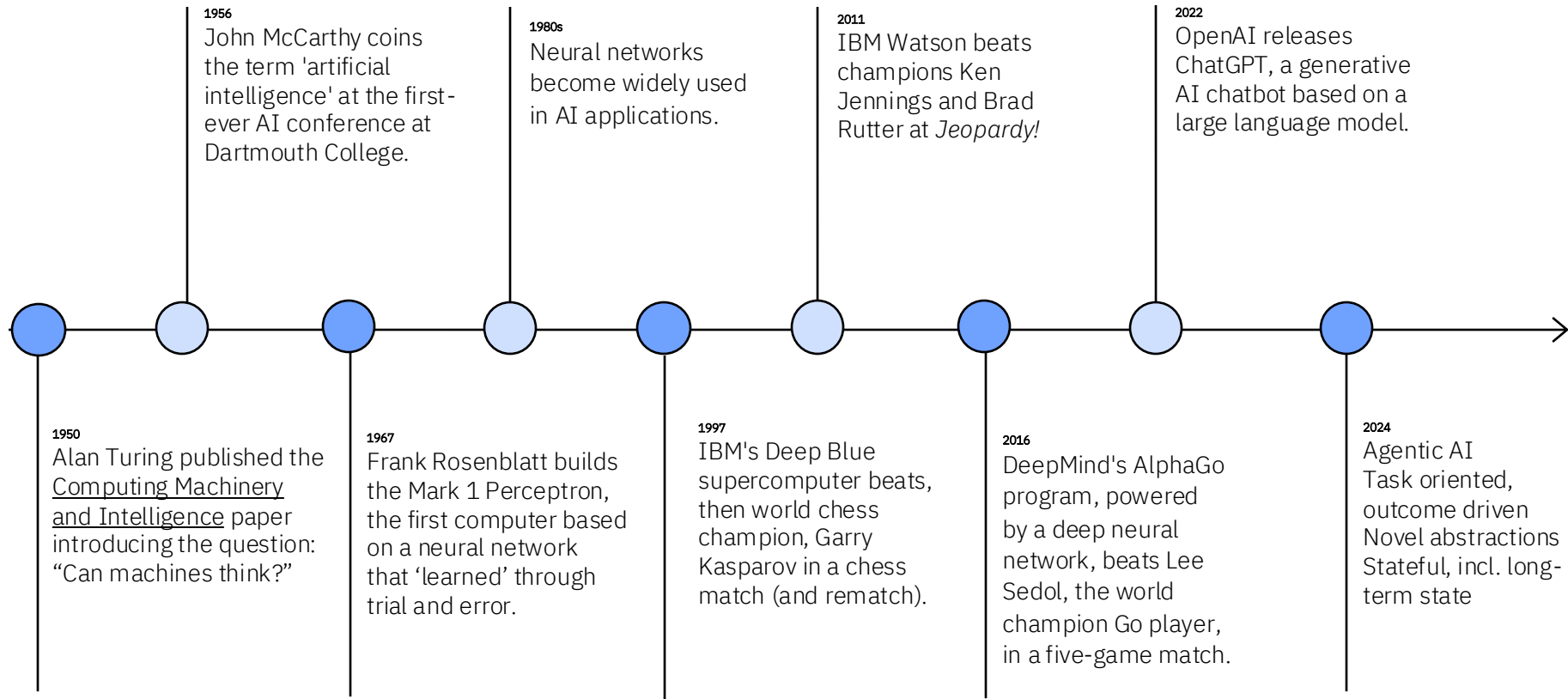
*Accelerate enterprise AI
with IBM Storage*



Vincent Hsu IBM Fellow, CTO & VP of Storage
Matheen Raza, Principle Product Marketing Manager



AI milestones

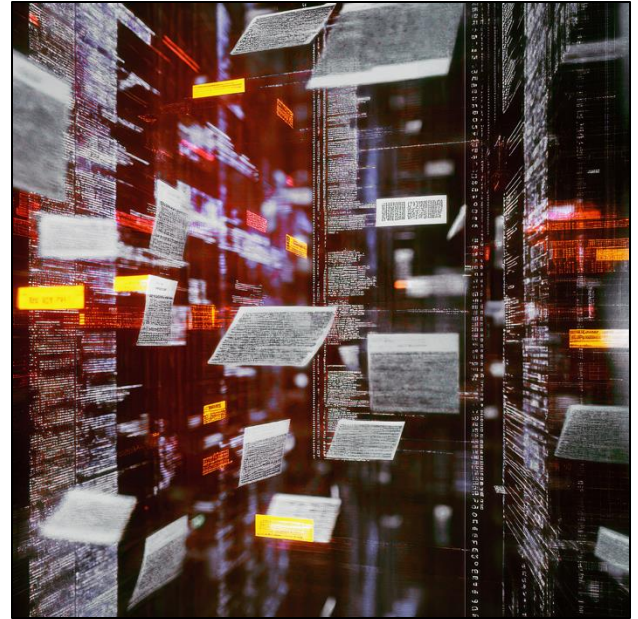


What are the biggest storage trends in 2025?

1. Enterprise storage platforms can serve generative AI (GenAI) workloads such as retrieval-augmented generation (RAG) inference and small-scale fine-tuning of language models, eliminating the need for a separate data lake.
2. Nearline SSD flash storage provides the opportunity to replace hybrid hard-disk drive (HDD) arrays with cost-efficient quad-level cell (QLC) flash solutions that improve overall performance, space efficiency and carbon emissions.
3. Cyberstorage solutions can be an additional layer of active defense at the storage layer to augment the traditional practice of deploying security at the network or application layers.
4. Integrated data intelligence enables unstructured data to become queryable data repositories to allow retrieval of data at a subobject level to support data analytics and AI and GenAI workflows.
5. Hybrid cloud storage presents the opportunity to optimize costs and enhance flexibility by seamlessly integrating on-premises and cloud resources. This approach supports global data growth, simplified management and the facilitation of hybrid cloud data workflows such as AI and GenAI.

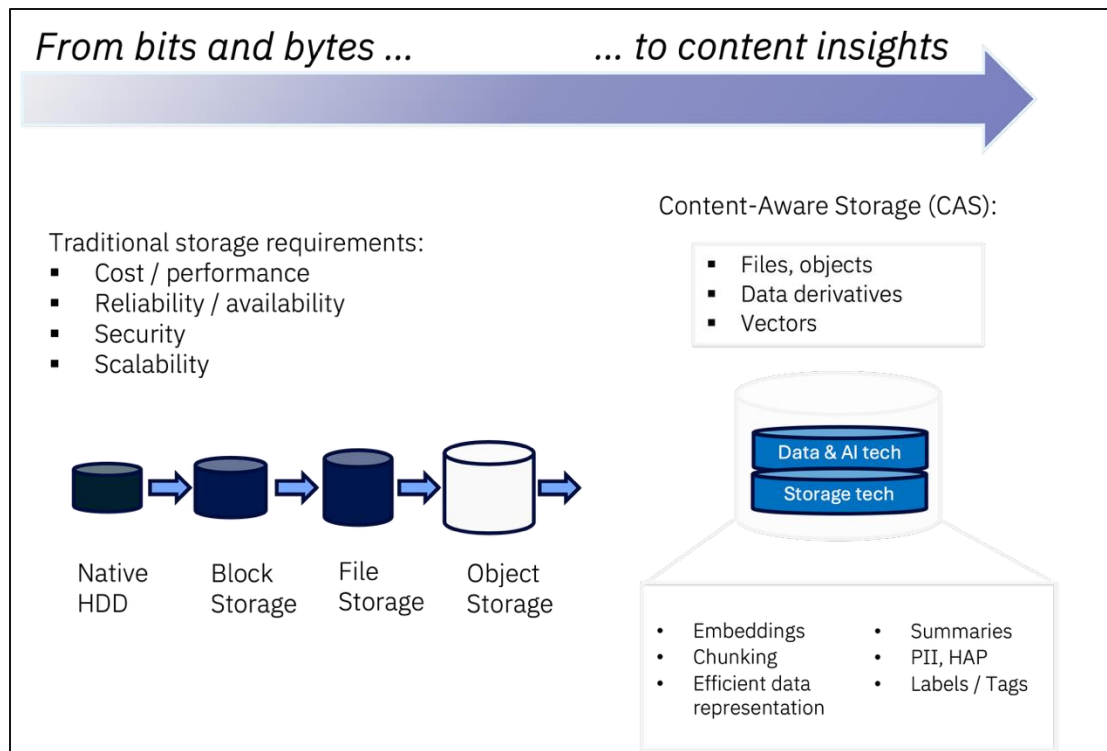
The 1% problem

- Organizations are swamped with unstructured data
- But less than 1% of all enterprise data was used to train major large language models
 - Data is copied multiple times – from source to lakehouse to data processor to vector database
 - Too many copies of the data, too many data transfers, loss of security access control
 - *All the data* gets reprocessed every time – no awareness of data changes



Content-aware storage: from “moving data to AI” to “moving AI to data”

- Leveraging AI technology inside storage to enable the data content awareness (aka queryable storage)
- Leveraging content awareness to accelerate inferencing efficiency

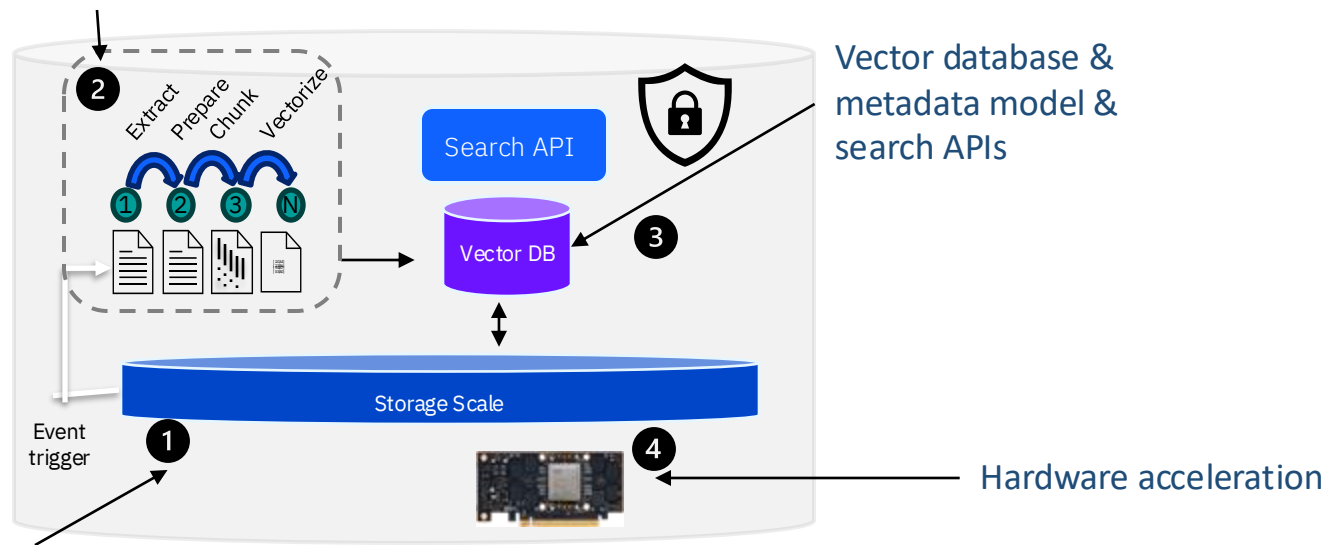


Building a content-aware storage

Data processing pipelines

(NVIDIA NIMs)

- Transfer multi-modal data to vectors



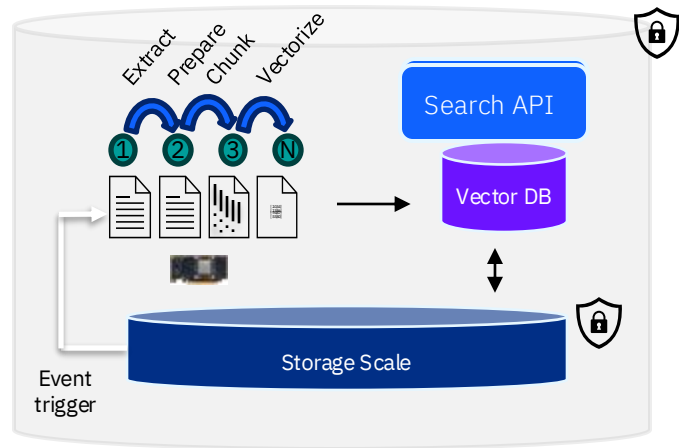
AI optimized storage

- run time and metadata management
- high performance storage
- storage virtualization
- Efficient tier storage
- Search APIs

Content aware storage simplify and secure data ingestion and search

IBM Content Aware storage offers the following advantage

- **Support legacy storage** : Connect to heterogeneous storage systems, including legacy unstructured data storage
- **Efficient usage of resource** : only process the changed data (vs re-processing all the data all the time)
- **Enterprise grade security** :
 - No unnecessary replicas of data
 - consistent ACL across raw data and embedding
 - Consistent encryption



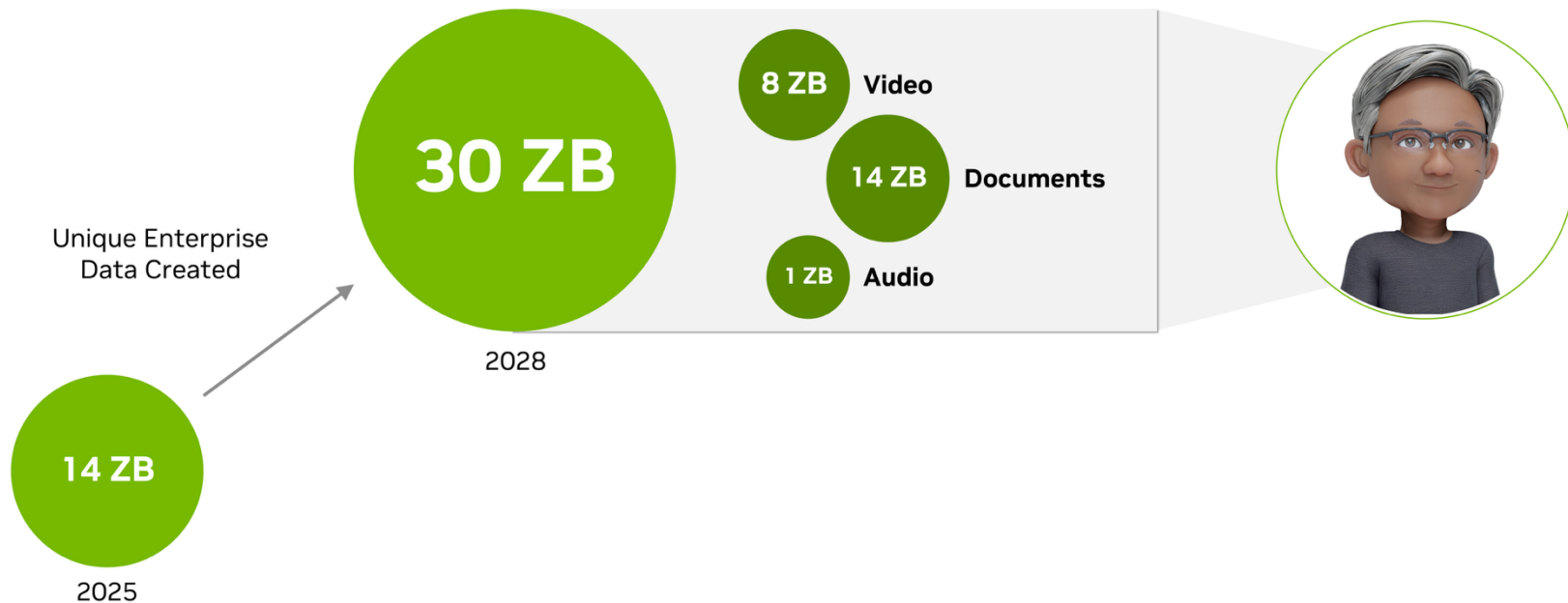


NVIDIA NeMo Retriever **Unlock Insights from Enterprise Data**

Sean Sodha, Senior Product Manager

The Amount of Enterprise Data is Massive & Growing

AI Agents Turn Knowledge into Action



Challenges

Difficult to take a RAG pipeline from proof-of-concept to production

Accuracy



Accurate generations require retrieval systems tuned to match the data

Data Security



Sending sensitive data to remote endpoints is inherently insecure

Disaggregated



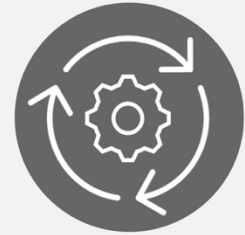
System builders must piece together and integrate many components

Cost



Automated LLM transactions make transaction costs unpredictable

Innovation Velocity



New models and techniques appear every day

NVIDIA NeMo Retriever Accelerates RAG Applications

NVIDIA open, commercial microservices power enterprise RAG pipelines turning data into knowledge



Optimized Inference Engines,
Built on NVIDIA NIM



State-of-the-art, customizable
models, fine-tuned for accuracy



Flexible and modular
deployment

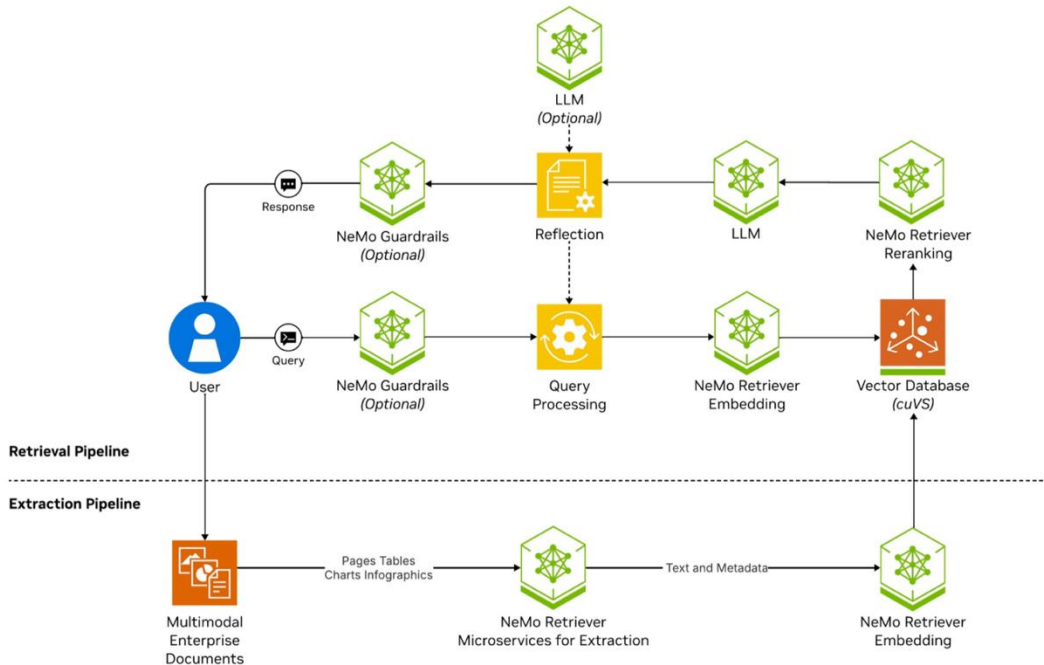


Accelerated vector search



Production Ready

Try RAG Blueprint at build.nvidia.com



Multimodal Data Extraction for Enterprise RAG at Scale

Unlocks knowledge from trillions of documents with 15X higher extraction throughput



Multimodal

Unlocks the next level of indexable enterprise data from text, tables, charts & infographics.



High Accuracy

World-class accuracy with high throughput extraction for RAG pipelines.



High Throughput

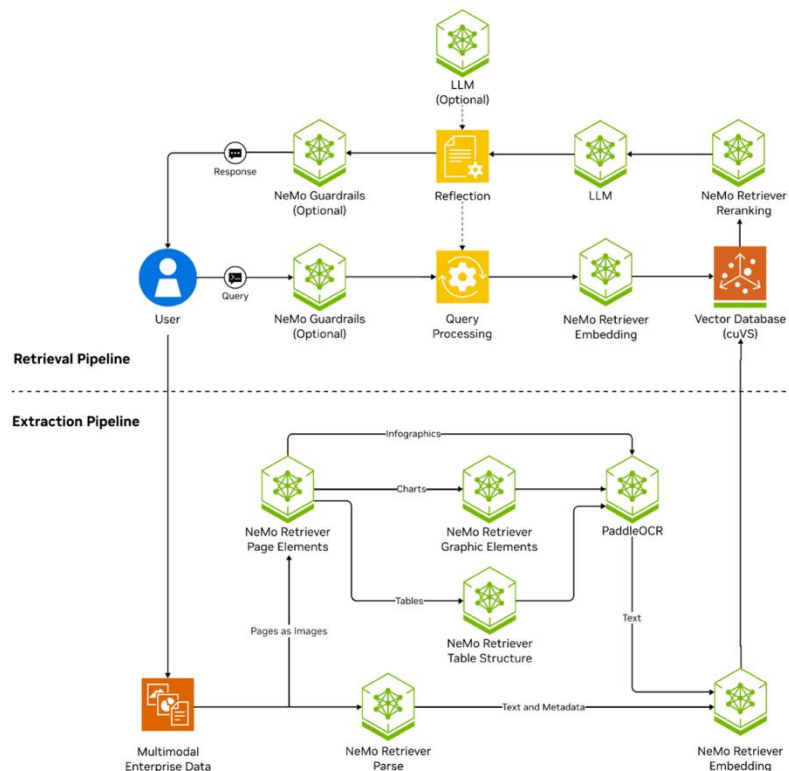
Quickly parse, extract, and embed data from complex documents at scale.



Customizable & Decomposable

Hyper-modular components for flexible deployments

See RAG Blueprint at build.nvidia.com



NeMo Retriever Multimodal Data Extraction Microservices

State-of-the-art data extraction for petabytes of PDFs created annually

Customizable & Scalable

Document Ingestion—any format,
with any modality, of any size



Supports docx, pptx, png, jpg,
infographics

Future: html, xlsx



Extract text, structured charts,
tables

Future: flow charts, block diagrams,
infographics



Customizable extraction operations



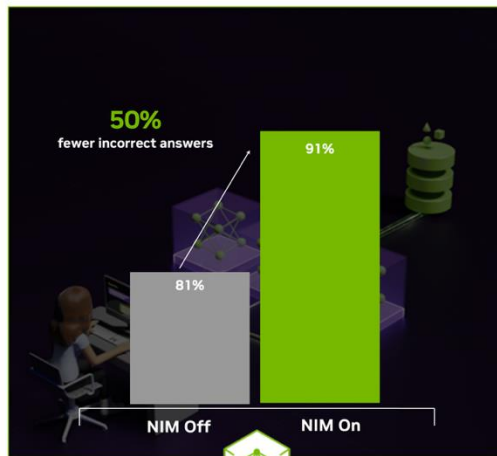
GPU accelerated linear scaling



Built on NVIDIA NIM

See [Performance Benchmarks](#)

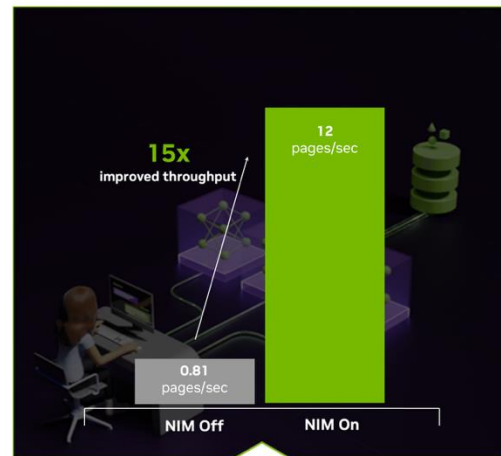
Multimodal Retrieval Accuracy



NeMo Retriever Extraction Recall@5 Accuracy
Retrieval of Enterprise Documents

Evaluated on publicly available dataset of PDFs consisting of text, charts, tables, and infographics.
NIM On: nemoretriever-page-elements-v2, nemoretriever-table-structure-v1, nemoretriever-graphic-elements-v1, paddle-ocr
NIM Off: open-source alternative: HW - 1xH100

Higher Throughput



Multimodal Data Extraction Throughput
Extraction of Enterprise Documents

Pages per second, evaluated on publicly available dataset of PDFs consisting of text, charts, and tables.
NIM On: nv-yolox-structured-image-v1, nemoretriever-page-elements-v1, nemoretriever-graphic-elements-v1, nemoretriever-table-structure-v1, PaddleOCR, nv-llama3.2-embedqa-1b-v2
NIM Off: open-source alternative; HW - 1xH100

Technical evolution: from standalone models to agentic systems

AI that can create for you



Models

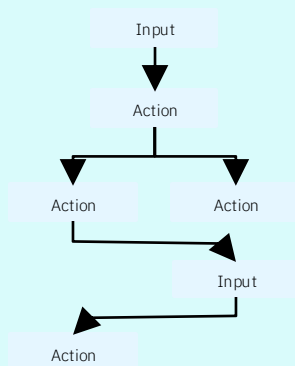
- Problem-solving
- Logical thinking
- Pattern matching

AI that can chat for you



Assistants

- Information retrieval
- Prescriptive tasks
- Single-step processes



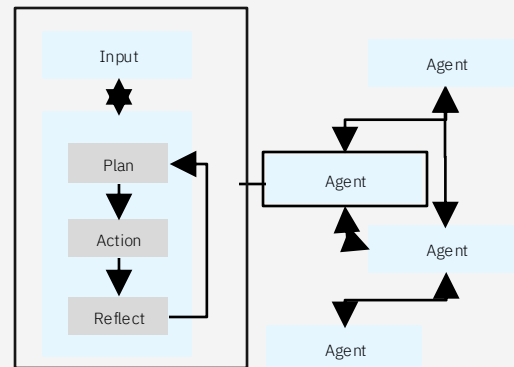
Traditional assistants

AI that can do for you



Agents

- Multi-step processes
- Autonomous actions
- Self-corrections

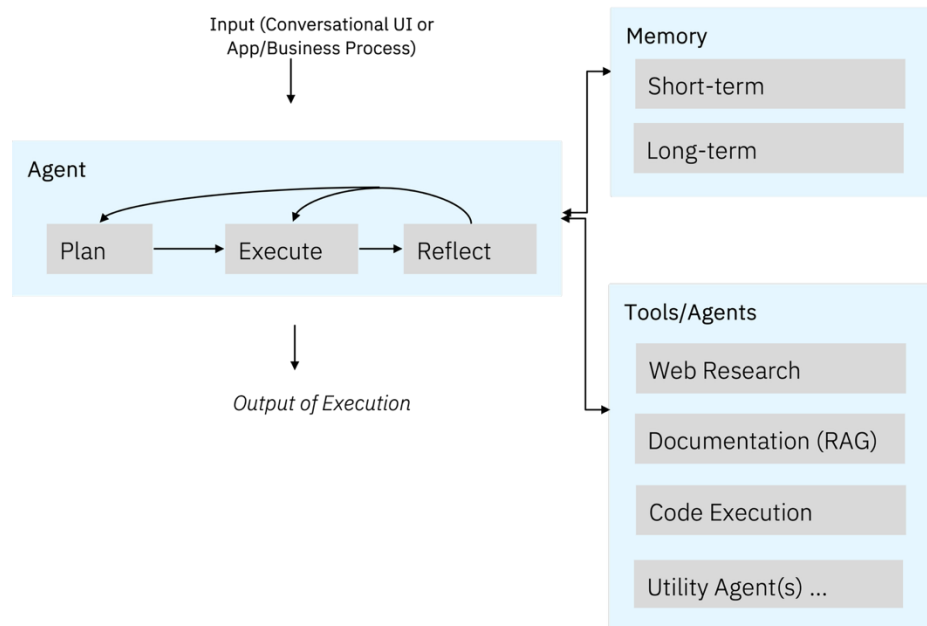


Single-agent assistants

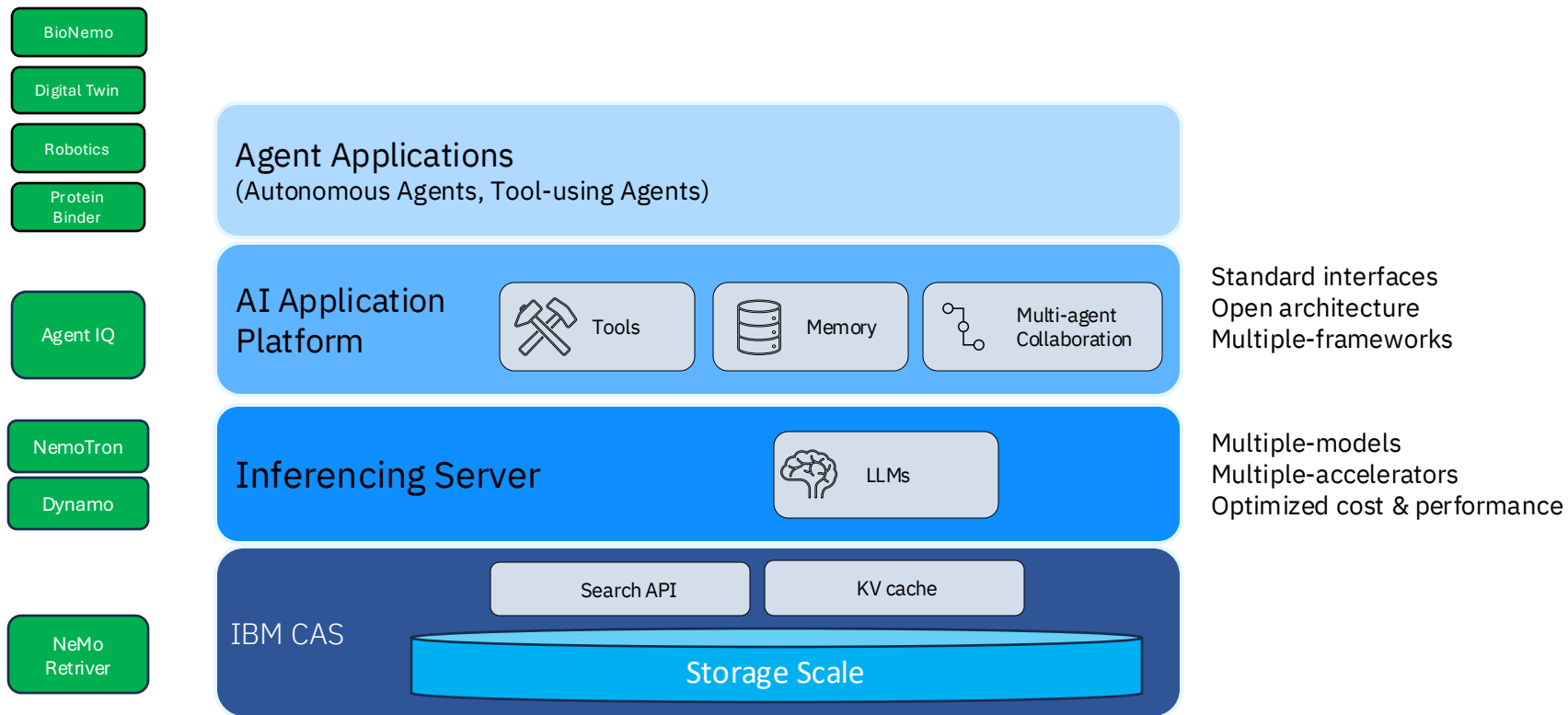
Multi-agent assistants

AI Agents

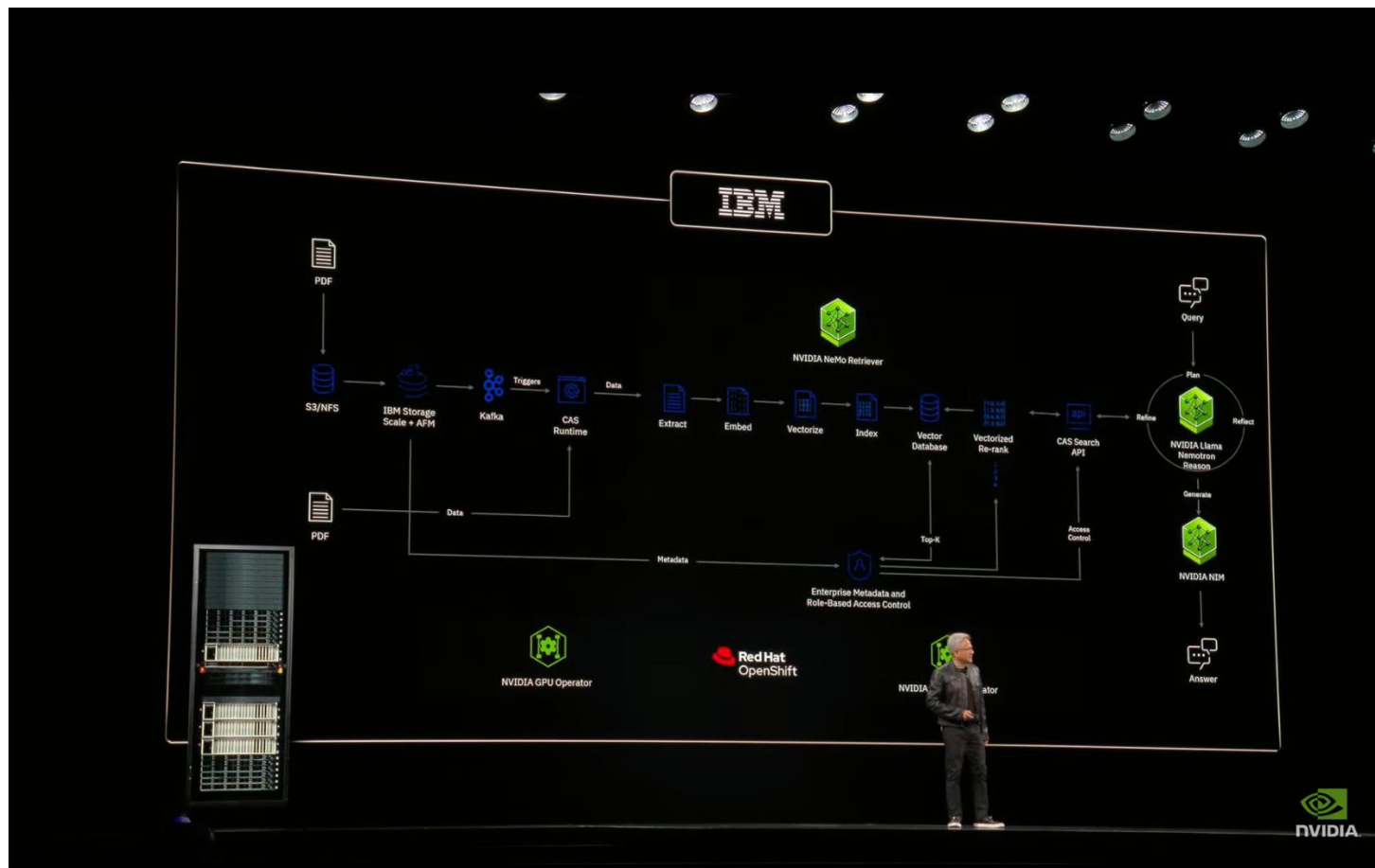
An AI agent is an autonomous system that can use tools and collaborate with other agents to plan and act on tasks. After it acts, the agent reflects on the results of its actions, learning iteratively and refining its approach to better align with its defined objectives.



AI Application Platform



IBM CAS Scale at GTC Taipei on 05/19



IBM Content Aware Storage

Hybrid Cloud Enterprise inferencing storage services (EDGE, Fusion HCI, GPU servers, clouds)

Inferencing SDS :

- Running CAS SDS on GPU servers (including DGX, HGX)

Inferencing appliance

Turn-Key inferencing server
IBM Fusion HCI

Inferencing StorageaaS

IBM CAS – Single name space for distributed inferencing



ESS6000

Network switches
Completes the integrated
Compute/Storage/Network stack



General purpose compute nodes
with software-defined storage
Runs CAS, Fusion management plane
and AI workloads



GPU nodes
NVIDIA
GPU optimized for AI/ML performance

AFM nodes
Abstracts and accelerates data stored on
third-party storage



NetApp DELL
and others...



IBM Cloud



aws

Azure

