Redbooks
ibm.com/redbooks

IBM Data Highway Blueprint for Life Sciences: Built for AI-Ready Scientific Data

Tran Nguyen
Andy Tran
Meghan Grable
Matthew Klos

IBM

Introduction

Purpose of this document

The focus is on addressing the full scientific data lifecycle, from high-speed instrument ingest through analysis and long-term retention. The architecture is built on IBM Storage Scale and IBM Fusion HCI, extended by IBM's data management capabilities for cataloging, content awareness, and agent access: IBM Fusion Data Catalog (FDC), Content-Aware Storage (CAS), and the Model Context Protocol (MCP) Server. It also integrates with the NVIDIA AI Data Platform (AIDP), along with partner and open-source technologies for protocol access, data movement, and governance.

Who should read this document

This document is intended for the following audiences:

- **Storage and infrastructure architects** evaluating or designing shared data platforms for life sciences HPC, AI, and research environments.
- **IT leaders and decision-makers** at research institutions, hospitals, genomic centers, and pharmaceutical organizations seeking to modernize scientific data infrastructure.
- **Data engineers and platform teams** responsible for implementing data ingest, tiering, lifecycle management, and AI readiness for scientific workloads.
- **Research computing and HPC teams** looking to connect instrument ingest, compute clusters, and AI workflows to a unified data platform.
- **AI and data science teams** requiring governed, discoverable, and content-aware scientific data for model training, RAG pipelines, and agentic AI workflows.

You should have working knowledge of enterprise storage concepts and familiarity with life sciences research workflows. Experience with IBM Storage Scale, Red Hat OpenShift, or NVIDIA GPU infrastructure is helpful but not required.

Section overview

Executive summary

An overview of the life sciences data challenge, IBM's data highway architecture, and the key technologies that enable scientific data to flow from instrument ingest to AI-driven analysis and long-term governance.

Introduction

A detailed examination of the four scientific data challenges (volume, velocity, variety, and veracity), the problems with traditional research workflows, how AI increases data management complexity, the concept of data gravity, and the IBM life sciences data highway architecture.

Reference architecture overview

A component-level walkthrough of the full architecture, covering the compute and AI platform (IBM Fusion HCI), the shared data platform layer (IBM Storage Scale), intelligent data management services (FDC, CAS, MCP Server, AIDP), the scientific data awareness maturity model, and an end-to-end workflow example.

Key use cases and customer benefits

Real-world application patterns across microscopy, genome sequencing, clinical imaging, AI-driven research, and data discovery — illustrated with six anonymized customer architecture examples and a summary of the platform's key benefits.

Guidelines and design considerations

Practical design guidance covering ingest, storage, compute, data management, the awareness layer, high availability, multi-institutional collaboration, data privacy and compliance, and operational considerations.

Conclusion

A summary of IBM's four core architectural capabilities — Abstraction, Acceleration, Access, and Awareness — and how the integrated platform transforms scientific data into a strategic, AI-ready research asset.

Authors

Tran Nguyen is a Technical Product Manager working in storage solutions at IBM, with a focus on developing reference architectures and driving technical initiatives for modern data and AI workloads. Her work combines technical ownership and execution across solution architecture, process development, and technical enablement. She drives cross-functional collaboration to translate technical requirements into practical solutions and guide projects from early concepts through implementation. Her background spans systems engineering, software engineering, hardware and software integration, automation, and technical research. Tran holds a degree in Materials Science & Engineering and Electrical Engineering & Computer Science from the University of California, Berkeley.

Andy Tran is a Storage Client Solutions Engineer, focused on designing reference architectures for storage, hybrid cloud, and AI infrastructure in research and data-intensive environments. He has led architecture discovery with account teams, built and documented validated lab deployments and partner integrations, and developed technical enablement material for field teams and business partners. His background includes software engineering, distributed systems, cloud infrastructure, and AI research. Andy holds a degree in Computer Science from the University of California, Riverside.

Contributors

Matthew Klos is a Senior Solutions Architect working across many industries, including financial services, research organizations, higher education, automotive, healthcare and life sciences, and many more. He has been administering and designing scale and scale system solutions for over 10 years internally and externally. Matt holds a degree in Information Technology from Marist University.

Meghan Grable is a Technical Product Manager on IBM's Client Solutions Engineering team, where she leads strategic customer engagements, first-of-a-kind solution validations, and proof-of-concept initiatives that help shape the future of IBM Storage. With a background in product management, cyber resilience, and data protection, she specializes in translating complex customer challenges into scalable technical solutions. Meghan combines expertise in enterprise design thinking, product strategy, and emerging technologies to help customers and partners accelerate innovation and business outcomes.

Feedback

This document reflects practical experience implementing these technologies in production environments. If you encounter issues not covered here, or if you have suggestions for improving the guidance, your feedback helps make this resource more valuable for the community.

Contact IBM Redbooks

Executive summary

Life sciences: IBM storage built for scientific data

Life sciences organizations are experiencing an unprecedented explosion of data from scientific instruments: microscopes, sequencers, beamlines, mass spectrometers, and clinical imaging systems. These instruments can generate hundreds of gigabytes per experiment, and a single research institution can produce petabytes per year. The problem is no longer only storing this data. The problem is moving it, processing it, archiving it and finding it again later for re-use.

The challenge

Most life sciences environments rely on fragmented data workflows rather than a unified data lifecycle. Instruments write data to the Windows workstations; researchers manually copy it to shared storage and then copy it again to compute environments for analysis. Over time, archives fill up with files that are difficult to search, govern, or reuse.

Traditional storage can support part of this cycle, but it was not designed to handle the full flow of ingesting, processing, managing, and reusing scientific data at scale. As a result, time to results suffers. Duplicate data piles up. AI projects stall because teams cannot easily find the right dataset or move it into the right workflow.

The solution

IBM addresses this bottleneck through an end-to-end architecture for scientific data called the *data highway*. This architecture enables scientific data to flow more efficiently from ingest to processing to long-term management, supporting both traditional HPC and AI-driven workloads.

Key components of this architecture include IBM Storage Scale, Tuxera Fusion SMB, IBM Fusion HCI, and IBM's intelligent data management capabilities, including IBM Fusion Data Catalog (FDC), Content-Aware Storage (CAS), Model Context Protocol (MCP) Server, and the AI Data Platform (AIDP).

The architecture at a glance

Together, these capabilities enable organizations to transform scientific data from a storage challenge into a reusable research asset that supports AI, analytics, collaboration, and long-term governance.

IBM Storage Scale is the shared file system foundation. Windows workstations, Linux servers, S3-based applications, containers, and GPU workloads can access the same data without creating unnecessary copies.

Tuxera Fusion SMB supports high-performance data ingest from instrument workstations, including Windows-based environments where SMB is the standard access method. This pattern was validated at a large cancer research institution for multi-GB/s write performance.

IBM Fusion HCI provides the container-native platform for AI, analytics, and modern application workloads next to the data.

IBM FDC, CAS, MCP Server, and AIDP capabilities add the data awareness layer: cataloging metadata, vectorizing content, enabling semantic search, and allowing AI agents to act on what they find.

Key terms and technologies

This paper assumes no prior familiarity with the technologies it references. The following terms appear throughout:

Term	Definition
Data catalog	A searchable inventory of an organization's data. A catalog records what datasets exist, where they live, and what they contain — owners, tags, and other metadata — so researchers can find data by searching rather than by browsing file systems.
HPC (high-performance computing)	Clusters of many servers working in parallel on large scientific workloads such as genomics pipelines, imaging analysis, and simulation. HPC systems mount the shared data platform directly.
IBM Storage Scale	A high-performance parallel file system (formerly IBM Spectrum Scale, and GPFS before that) that presents one namespace — a single directory tree spanning every tier and site — across flash, disk, tape, and cloud, with the same data accessible simultaneously through multiple protocols. It is the shared data platform referenced throughout this paper. Appliance versions are branded IBM Elastic Storage System (ESS); IBM Storage Scale System 6000 is the current generation building block.
Access protocols (POSIX, NFS, SMB, S3)	The standard interfaces through which applications read and write data. POSIX is the native interface for Linux and HPC applications; NFS and SMB are network file-sharing protocols for Linux and Windows respectively; S3 is the object storage interface popularized by the cloud, with IBM Cloud Object Storage (COS) as IBM's implementation. Storage Scale serves all of them against the same data.
Tuxera Fusion SMB and SMB Direct	A high-performance implementation of the SMB protocol for Storage Scale from IBM partner Tuxera, which lets Windows-based instruments and workstations write directly to the platform. SMB Direct adds RDMA — networking that moves data straight between systems' memory over InfiniBand or high-speed Ethernet — for multi-gigabyte-per-second ingest.
IBM Fusion and Fusion HCI	IBM Fusion is IBM's software suite for running data services on Red Hat OpenShift, IBM's Kubernetes-based container platform. Fusion HCI is the hyperconverged appliance version: pre-integrated racks of compute, storage, and networking, with optional GPUs, that arrive ready to run OpenShift and AI workloads on premises.
IBM Fusion Data Catalog (FDC)	IBM's data catalog for unstructured data. It builds a global metadata index across file, object, and cloud storage, and adds classification, tagging, and policy controls on top of that index.
Content-Aware Storage (CAS)	A Storage Scale capability that reads the content of stored files and converts it into vector embeddings — numeric representations of meaning — so AI systems can search data by what it says rather than by what it is named. It is accelerated by NVIDIA software.
NVIDIA AI Data Platform (AIDP)	NVIDIA's reference design for AI infrastructure, pairing its accelerated computing with partner storage platforms. IBM integrates Storage Scale and Content-Aware Storage with it. DGX and HGX, referenced alongside it, are NVIDIA's integrated GPU server systems.

Term	Definition
Model Context Protocol (MCP)	An open standard that lets AI assistants and agents connect to external tools and data sources in a consistent way. In this architecture, an MCP server is how AI agents reach the catalog and storage services.
Agentic AI	AI systems that take multi-step actions on their own — searching for data, retrieving it, and operating on it through tools — rather than only answering a single prompt. The awareness layer in this paper is what gives such agents governed access to scientific data.
IBM watsonx	IBM's portfolio of AI products for building, tuning, and governing AI models and assistants.
GPUDirect Storage	An NVIDIA technology that moves data directly from storage into GPU memory, bypassing the server's CPU, so GPUs are fed at full speed.
Active File Management (AFM)	Storage Scale's built-in replication and caching between sites. It keeps data consistent across campuses, datacenters, or partner institutions, and powers the multi-site examples in this paper.
IBM Storage Protect, Diamondback, and TS4500	IBM's data protection stack. Storage Protect is backup and recovery software; Diamondback and TS4500 are tape libraries used for low-cost, long-term, and air-gapped retention; IBM Guardium Key Lifecycle Manager (GKLM) manages the encryption keys.
Retrieval-augmented generation (RAG)	An AI technique in which a model first retrieves relevant documents or data and then generates its answer from what it found. RAG quality depends directly on how findable and well-described the underlying data is.
iRODS	The Integrated Rule-Oriented Data System, an open-source platform for policy-based research data management. Rules automate data placement, retention, and access, with strong provenance tracking.
Globus	A data transfer and sharing service widely used by research institutions to move large datasets between sites and collaborators securely.

Introduction

The scientific data challenge

Life sciences organizations are generating data at a rate that traditional research data workflows were never designed to support. Advances in microscope, genomic sequencing, clinical imaging, beamline research, and simulation technologies have dramatically increased both the volume of data being produced and the speed at which it is generated. What was once measured in terabytes is now routinely measured in petabytes.

Scientific data presents four major infrastructure challenges: volume, velocity, variety, and veracity.

Volume

Research institutions are producing data at unprecedented scale. For example, a U.S.-based national research laboratory's upgraded Advanced Photon Source generates hundreds of PB/yr, while some genomic sequencing centers can produce up to 300 TB/day. As instrument capabilities continue to improve, petabyte-scale environments are becoming the norm rather than the exception.

Velocity

Scientific instruments can generate data faster than researchers can move or process. A single imaging session may produce hundreds of gigabytes of data, while large-scale experiments can generate terabytes within hours. In many environments, the bottleneck is no longer the instrument, but the workflow required to move data between systems.

Variety

Scientific data exists in many forms, including microscopy images, genomic sequences, clinical records, simulation outputs, structured databases, and electronic laboratory notebooks. These datasets often need to be analyzed together despite being stored in different formats and accessed by different applications.

Veracity

Scientific data must remain trustworthy throughout its lifecycle. Researchers need to know where data originated, which instruments generated it, who modified it, and how it has been used. In regulated environments, maintaining data provenance, access controls, retention policies, and auditability is critical for scientific integrity and regulatory compliance.

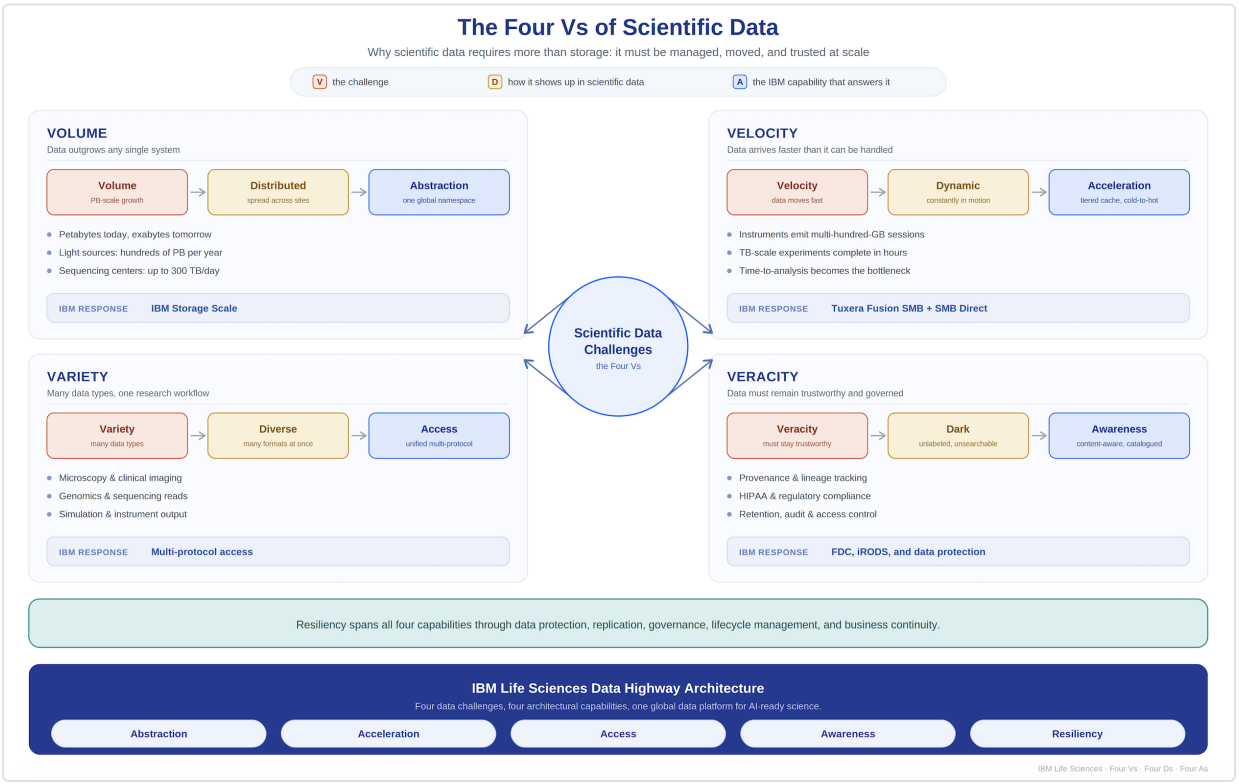

Figure: The Four Vs of Scientific Data — how Volume, Velocity, Variety, and Veracity map to IBM's five architectural capabilities: Abstraction, Acceleration, Access, Awareness, and Resiliency

These four challenges point directly to four capabilities. The V's describe what makes scientific data difficult; the four A's describe what a platform must do about it: Abstraction (one namespace across every tier and site), Acceleration (ingest and access at instrument speed), Access (every protocol against the same data), and Awareness (data that is

findable, governed, and AI-ready). The remainder of this paper traces that shift from managing the V's to operating on the A's.

The traditional workflow problem

Despite advances in scientific instruments, many research environments still rely on workflows that were designed for much smaller datasets.

A typical workflow often includes the following steps:

1. Researchers capture data on a specific instrument.
2. Data lands on a local disk via a Windows workstation.
3. Researchers copy the data to shared storage.
4. Data gets copied again to a compute cluster for processing.
5. Results get written somewhere, archived, and eventually lost.
6. This leads to inconsistency with data and future searches.
7. Researchers risk data loss at every step.

Why AI makes the problem worse

The shift toward AI-driven research has increased these challenges.

Modern AI workflows such as retrieval-augmented generation (RAG), vector search, NVIDIA AIDP integrations, and agentic AI systems require more than access to storage. They require data that can be discovered, understood, governed, and reused at scale.

Simply storing files is no longer enough. AI systems need metadata, context, lineage, and searchable content in order to identify the right datasets and generate meaningful results. As organizations adopt AI, the focus shifts from storing data to managing it as an active research asset.

These requirements do not depend on which AI model an organization chooses. Some institutions will run general-purpose open models; others will adopt models trained specifically for biomedical research. The architecture supports either path: the data layer's job is to deliver governed, content-aware data to whichever model is selected. Comparing the two approaches, and validating that choice for scientific use, remains an evaluation each institution should plan for.

Data gravity

As a dataset grows, it develops gravity. The larger and more active it becomes, the more expensive it is to move, because latency rises and effective throughput falls the farther compute sits from the data. At petabyte scale this makes copy-based workflows progressively unsustainable, so the practical answer is to bring compute and AI to the data rather than repeatedly relocating the data to compute. This principle is what the data highway is built on. A shared Storage Scale foundation lets HPC, AI, and analytic workloads run in place against a single copy, instead of staging duplicates across systems.

Infrastructure requirements for modern scientific data

A modern life sciences environment requires more than capacity. It requires infrastructure that supports the entire data lifecycle.

Key requirements of a modern life sciences environment include the following features:

- A parallel system capable of ingesting multi-GB/s data streams from scientific instruments while simultaneously serving HPC and AI workloads.

- Multi-protocol access that allows Windows, Linux, S3 applications, containers, and GPU environments to access the same data without creating unnecessary copies.
- Data cataloging and search capabilities that make PB-scale datasets discoverable and reusable.
- Lifecycle management, governance, and data protection capabilities that support retention policies, compliance requirements, and long-term preservation of scientific data.

The life sciences data highway IBM architecture

We coined the term *data highway* because scientific data moves through multiple stages of its lifecycle, from instrument ingest to analysis, collaboration, retention, and reuse. The name is an analogy rather than an industry term. Instruments serve as the on-ramp, the shared data platform is the roadway, multi-protocol access and compute are the exit lanes, and governance and protection operate along the entire route. Traditional environments often create disconnected storage silos between these stages. The goal of the architecture is to provide a continuous path where data can move efficiently without repeated copies, manual transfers, or loss of context.

The life sciences data highway backed by IBM solutions is designed to address the full scientific data lifecycle, from instrument ingest to AI-driven analysis and long-term retention. Rather than treating storage as a standalone repository, the architecture creates a unified platform where data can be ingested, processed, governed, discovered, and reused without unnecessary movement or duplication.

The architecture combines high-performance storage, multi-protocol access, AI infrastructure, metadata services, and data governance capabilities into a single workflow-structured platform.

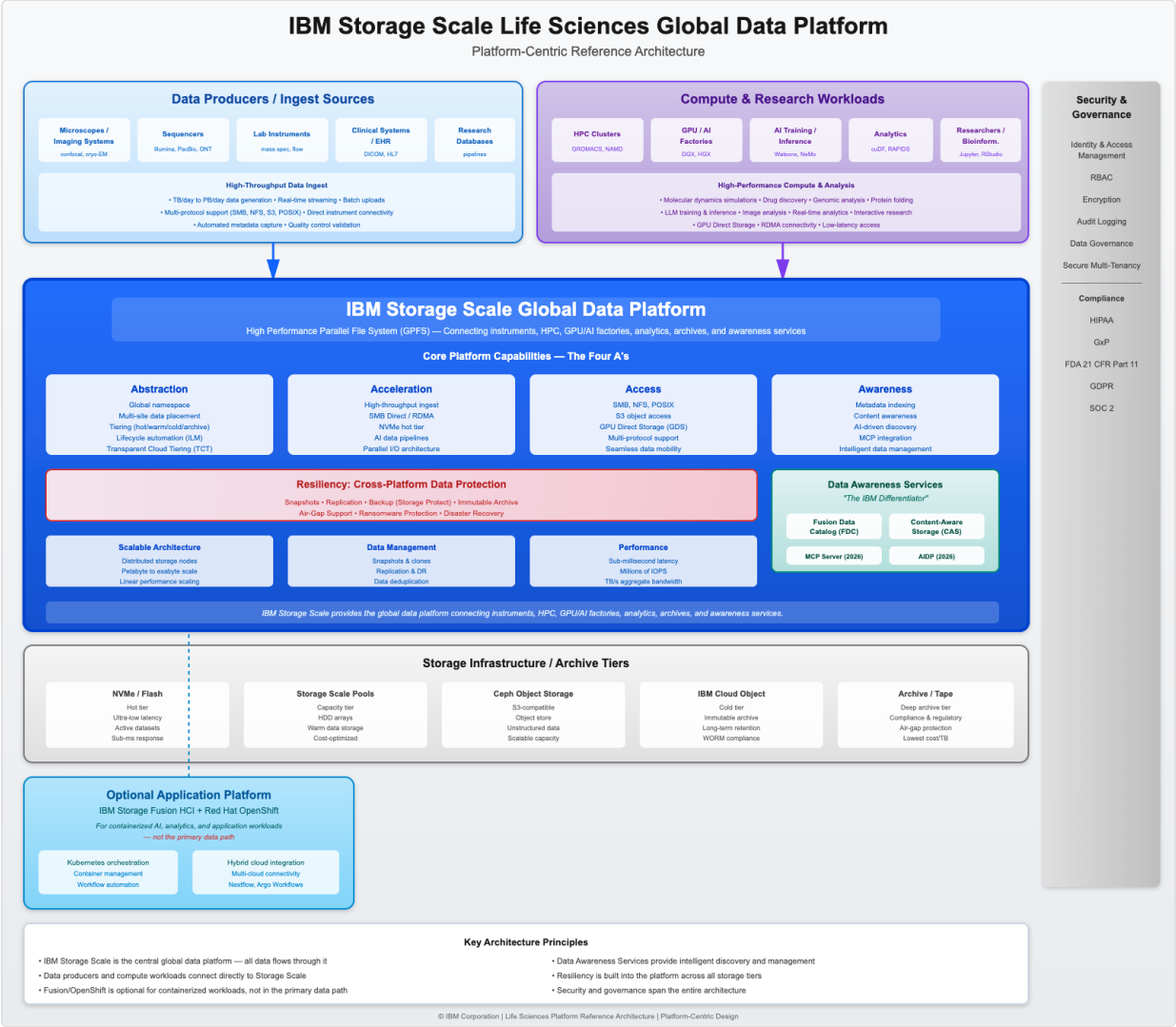

Figure: IBM Storage Scale Life Sciences Global Data Platform — platform-centric reference architecture connecting instruments, HPC, GPU/AI factories, analytics, archives, and awareness services through a single IBM Storage Scale foundation

Reference architecture overview

This section describes the components and layers that make up the IBM Data Highway Blueprint for Life Sciences. The architecture is designed to connect scientific instruments, high-performance storage, AI compute, and data management services into a unified platform, enabling data to move from ingest to AI-driven analysis without duplication or manual intervention. Each layer is covered in turn, followed by a summary of how the pieces work together and an end-to-end workflow example.

Compute and AI platform

IBM Storage Fusion HCI provides the compute and application platform layer. Built on Red Hat OpenShift, it combines compute, storage integration, networking, virtualization, and GPU acceleration into a pre-validated platform for scientific workloads.

Key Core Components / Capabilities	What it does
Bare-metal OpenShift	Kubernetes platform for AI, analytics, watsonx, and containerized scientific workloads
OpenShift Virtualization	Supports traditional VM-based applications such as Ansys and other GPU-heavy software not yet containerized
Scale Integration	High-throughput access to shared scientific datasets, including GPU Direct Storage support
NVIDIA GPU nodes	DGX/HGX-class compute for AI training, inference, and advanced analytics

The IBM and NVIDIA solution combination is complementary by design. NVIDIA provides accelerated compute for AI and advanced analytics, while IBM provides the high-performance data layer needed to efficiently support those workloads with scientific data. IBM Storage Scale supports NVIDIA GPUDirect Storage for high-throughput data access, alongside CAS that can leverage NVIDIA-accelerated capabilities.

Data platform layer

IBM Fusion HCI and IBM Storage technologies work together to connect accelerated compute with a high-performance shared data layer. Fusion HCI provides the OpenShift-based compute and application environment, while IBM Storage Scale provides high-throughput access to shared scientific data across instrument, HPC, AI, and analytics workflows. Together, they create a unified scientific data platform that supports high-speed ingest, shared access, intelligent data management, metadata services, AI integration, and long-term data management.

The goal is to combine parallel storage, optimized instrument ingest, multi-protocol access, and intelligent data management capabilities so that researchers and AI agents can discover, understand, govern, and act on scientific data throughout its lifecycle.

The architecture consists of five connected layers:

1. Data Ingest
2. Shared Storage
3. Multi-Protocol Access
4. Data Awareness & Cataloging
5. AI and Scientific Computing

Each layer builds on the previous one, allowing scientific data to move through its lifecycle without requiring manual transfers or duplicate copies.

Key Core Components / Capabilities	What it does
IBM Storage Scale	Parallel file system that scales from TB to EB. Single namespace across all storage tiers. All other components in the architecture read from and write to Scale.
Tuxera Fusion SMB	High-performance SMB for Windows-based instrument workstations. Supports SMB Direct (RDMA) for accelerated ingest workflows validated in large-scale research environments. Replaces Samba on Storage Scale protocol nodes for improved throughput and reduced infrastructure footprint.
Multi-Protocol Access	Concurrent access via POSIX, NFS, SMB, S3, and GPU Direct Storage on the same files, eliminating duplicate copies across HPC, AI, and analytics workloads.
IBM Fusion Data Catalog (FDC)	Indexes metadata across Scale, tagging files, driving policy-based tiering and lifecycle management.
IBM Content-Aware Storage (CAS)	Creates vector representations of file content to support semantic search, AI retrieval workflows, and content-aware data management.
Model Context Protocol Server (MCPS)	Exposes storage, metadata, and CAS through MCP, enabling AI assistants and agents to interact with enterprise data using natural language workflows.
AI Data Platform (AIDP)	Launching 2026; integration layer between IBM storage and NVIDIA AIDP.
Data protection	Storage Protect / archivable COS / Atempo Miria for replication.
Existing Tools (Globus, iRODS)	Globus: inter-institutional transfer of data. iRODS: governance and federation.

> **Note:** Tuxera Fusion SMB, Globus, and iRODS represent ecosystem integrations that complement the IBM Storage platform and may be deployed based on customer requirements.

Why AI requires intelligent data management

Traditional storage platforms focus on where data is stored. The challenge in modern life sciences is increasingly understanding what the data is, where it came from, and how it can be reused once it has been stored. As datasets grow into the petabyte range, organizations often accumulate large amounts of "dark data" — datasets that exist but are difficult to discover and understand. Researchers may know the data exists but struggle to find the right experiment, image, sequencing run, or simulation result.

The awareness layer addresses this challenge by transforming stored data into discoverable and reusable research assets. FDC indexes metadata across datasets, CAS analyzes file content to provide additional context, and MCP Server exposes these services to AI applications and agents. Together, these services help make scientific data searchable, understandable, and AI-ready, enabling researchers to quickly identify relevant datasets, improve dataset reuse, and accelerate scientific discovery.

Scientific data awareness maturity

Data awareness is not a single product or capability. It develops progressively as organizations improve how they discover, govern, understand, and act on scientific data. The maturity model below illustrates this progression from metadata discovery to AI-enabled orchestration.

These levels are complementary rather than mutually exclusive. Organizations may adopt capabilities incrementally based on their existing tools, governance requirements, and AI objectives.

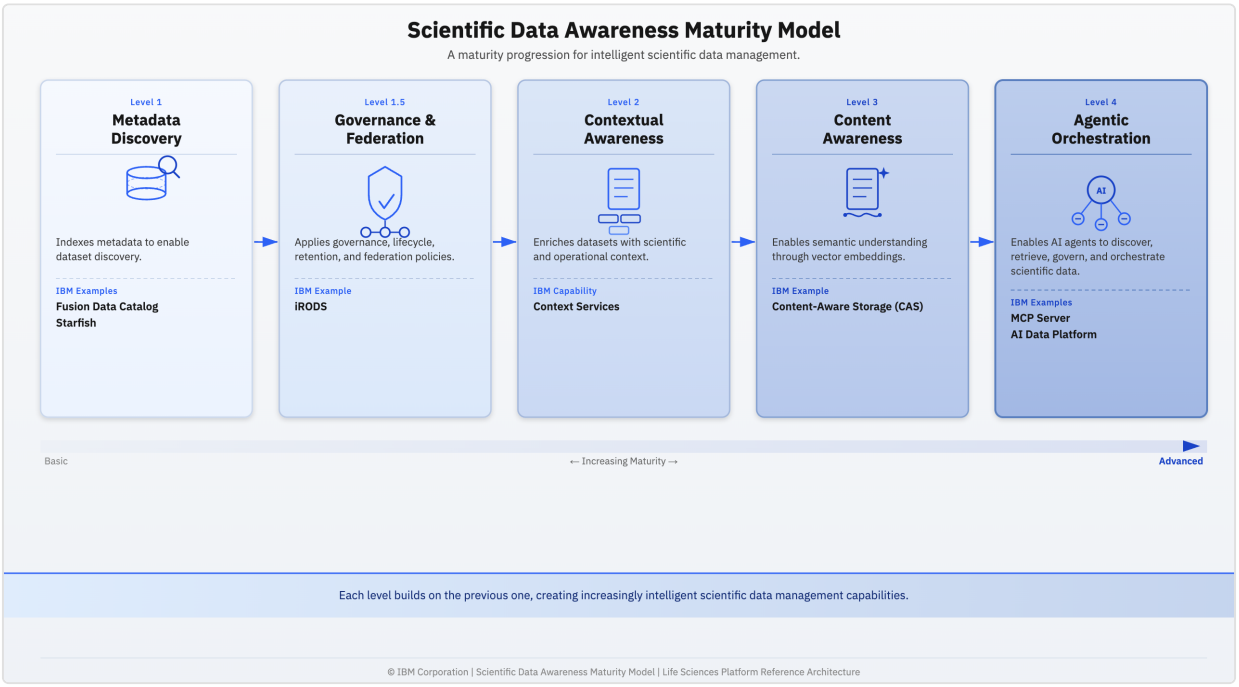

Figure: Scientific Data Awareness Maturity Model — five-level progression from metadata discovery and governance federation through contextual and content awareness to agentic AI orchestration

Level 1: Metadata discovery

At the foundational level, metadata discovery capabilities index scientific data and make datasets easier to search, locate, and understand. Tools such as IBM FDC enable researchers, applications, and AI workflows to find datasets through metadata rather than manually browsing file systems.

These capabilities can also support automated discovery and real-time indexing across file, object, and cloud sources, along with classification, tagging, and metadata enrichment to improve visibility into distributed scientific data.

Level 1.5: Governance and federation

At this level, policy-based governance and federation build on metadata discovery by introducing automated controls for how data is managed across its lifecycle. These frameworks can apply policies for data placement, retention, access, provenance, and movement across departments, institutions, or research sites while maintaining a consistent view of distributed data. Open-source systems such as iRODS exemplify this approach, particularly in environments where provenance, reproducibility, and policy-driven data management are important.

Level 2: Contextual awareness

At this level, technical metadata is enriched with scientific and operational context, such as the experiment, instrument, research project, or clinical workflow associated with a dataset. This allows users and AI systems to understand not only where data is located, but also why it was created and how it should be used.

Level 3: Content awareness

CAS analyzes the contents of files and creates vector representations that support semantic search, AI retrieval, and deeper understanding of unstructured scientific data.

Level 4: Agentic orchestration

At the highest level, AI agents can use metadata, context, content awareness, policies, and infrastructure insight to take action on data. This may include locating datasets, initiating caching, moving data between tiers, applying retention policies, preparing data for AI workflows, or retrieving archived data through natural-language requests.

How the pieces work together

The Life Sciences architecture is built around IBM Storage Scale as the central data platform. IBM Fusion HCI provides a container-native platform for AI inferencing, agentic workflows, metadata services, and modern applications, while larger HPC clusters and NVIDIA AIDP environments can consume the same data through Storage Scale. This allows organizations to place workloads where they are most appropriate while maintaining a unified data foundation.

Fusion HCI brings the GPU compute and the OpenShift Virtualization for legacy apps. Storage Scale brings the file system. The data management capabilities provide metadata discovery, governance, content understanding, and AI orchestration that transform stored data into an active research asset.

Researchers do not need to reorganize their existing workflows. Instruments continue writing through the same Windows-based workflows, while existing tools such as Globus and iRODS continue to operate as expected. This architecture adds high-performance ingest, multi-protocol access, and data awareness capabilities on top of the existing environment.

Together, these components create a unified platform where scientific data can move from instrument ingest to AI-driven analysis without requiring researchers to manually relocate or duplicate datasets.

The following core components are organized by category:

- Ingest
  - Instruments → Windows Workstation
  - Tuxera Fusion SMB (Partner Integration)
  - CES Protocol Nodes
  - SMB Direct (RDMA)
- Storage
  - IBM Storage Scale
  - Global Namespace
  - Multi-Protocol Access
- Processing and AI
  - IBM Fusion HCI (Containers, VMs, Inference)
  - HPC Clusters (Scientific Computing)
  - NVIDIA AI Data Platform (AIDP)
  - IBM watsonx / Foundation Models
- Awareness
  - IBM Fusion Data Catalog (FDC)
  - IBM Content-Aware Storage (CAS)
  - MCP Server

- Protection
  - IBM Storage Protect
  - IBM COS
  - Atempo Miria
  - Air-Gapped Snapshots
- Existing Ecosystem Integrations
  - Globus for inter-institutional transfer
  - iRODS for governance and federation

End-to-end workflow example

1. The microscope captures the image. The Windows workstation writes the file via SMB Direct at multi-GB/sec to a high-performance SMB protocol node.
2. IBM Storage Scale lands the file on the NVMe hot tier.
3. Metadata services index the dataset. CAS enriches the file content to support semantic search and downstream AI workflows.
4. A researcher's AI pipeline running on Fusion HCI reads the file through the AIDP integration layer, which connects Storage Scale and CAS to NVIDIA-based AI infrastructure. The pipeline runs analysis or inference. Results are written back to IBM Storage Scale.
5. After a period of inactivity, IBM Storage Scale's tiering policy moves the file from NVMe to the HDD capacity tier.
6. Replication services copy the file to a disaster recovery site. Backup services create an immutable archive in IBM Cloud Object Storage.
7. Months later, an AI agent receives a plain-language request to find similar imaging. The agent uses AI orchestration services to query metadata and content-awareness services, returning matching files.
8. Globus is used to ship the result set to a collaborating institution.

Key use cases and customer benefits

This section outlines how the IBM Data Highway Blueprint applies to real-world life sciences workloads, and the tangible benefits it delivers to research and clinical organizations. The use cases span imaging, genomics, clinical workflows, and AI — each illustrating how the architecture's storage, data management, and AI layers work together to accelerate scientific outcomes.

Use cases across life science applications

Microscopy and imaging research

Tuxera Fusion SMB lands the files that microscopes write — multi-hundred GB files per session — on Scale at multi-GB/sec rates. Researchers can begin analysis immediately without waiting for manual data transfers between systems.

Scientific data generated by microscopy instruments is ingested directly into IBM Storage Scale through high-performance SMB, where it becomes immediately available to GPU clusters, HPC resources, and AI workflows. Cross-site replication provides disaster recovery, while an emerging metadata and content-awareness layer supports future semantic search and AI-assisted research.

Genome sequencing

Sequencing centers generate 150 to 300 TB/day. With Storage Scale multi-protocol access, sequencing pipelines (POSIX), variant analysis, and clinical interpretation can read the same data. This reduces duplicate datasets and allows bioinformatics, clinical, and research teams to work from a single source throughout the sequencing pipeline.

Clinical imaging and diagnostics

Hospitals that need access to imaging governed by HIPAA can run AI inferencing on IBM Fusion HCI without sending data outside. Sensitive patient imaging remains within the organization's environment while enabling AI-assisted diagnostics, research, and workflow automation.

Data discovery, preparation, and AI-driven research

As scientific datasets grow into the petabyte range, researchers often spend significant time locating, organizing, validating, and preparing data before meaningful analysis can begin. Data may be distributed across microscopy systems, genomic repositories, clinical records, imaging archives, and simulation outputs, making it difficult to identify the most relevant datasets for new studies.

By providing centralized access, metadata awareness, and intelligent data discovery across distributed repositories, this blueprint helps researchers reduce time spent on data wrangling and preparation. Scientists can more quickly identify relevant datasets, prepare them for downstream analysis, and integrate them into AI, machine learning, and advanced analytics workflows.

Reducing the effort required to find and prepare data enables researchers to focus more time on scientific discovery, hypothesis testing, and accelerating research outcomes.

Data-driven AI research and simulations

Modern life sciences organizations are increasingly combining traditional scientific computing with AI-driven research workflows. Beamline facilities, genomic research centers, pharmaceutical companies, and academic institutions often need to support HPC simulations, AI model training, AI inferencing, and advanced analytics against the same datasets.

By providing a shared global data platform, IBM Storage Scale enables HPC environments, GPU clusters, AI factories, and analytics platforms to access the same data without creating duplicate copies. Researchers can move seamlessly from data generation and simulation to model training, inferencing, and scientific discovery while maintaining a consistent view of data throughout the research lifecycle.

This approach accelerates AI-driven research initiatives, improves resource utilization, and reduces the operational complexity associated with managing separate storage environments for HPC and AI workloads.

Example customer architectures

This section shows six representative architectures based on interviews with research institutions running IBM Storage Scale today. Each diagram reads the same way: data generation and ingest on the left, the shared IBM Storage Scale platform in the center, scientific compute on the right, and protection, replication, and archive at the bottom. Solid blue elements are in production. Dashed purple elements are planned or under evaluation. Institutions are anonymized; capacities and technologies are as described by each organization.

AI-enabled microscopy research platform

A large research institution runs the microscopy and imaging research pattern described above. Instruments write directly to IBM Storage Scale through Tuxera Fusion SMB, so researchers can begin analysis without manual transfers. The same platform feeds GPU and HPC clusters for imaging, genomics, and AI workloads, and replicates to a dedicated disaster recovery site.

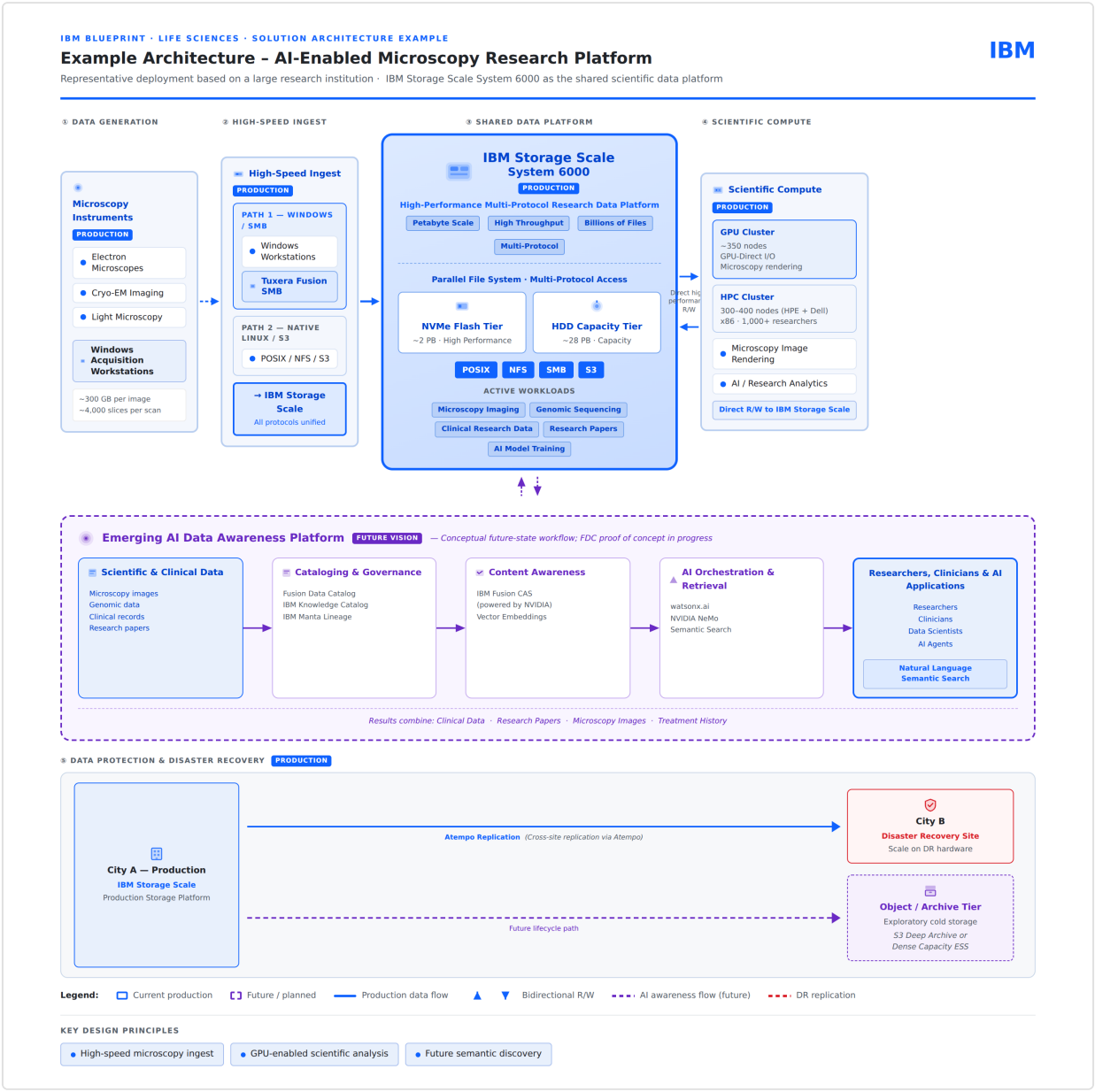

Figure 1: AI-Enabled Microscopy Research Platform — large pediatric research institution with IBM Storage Scale System 6000, Tuxera Fusion SMB ingest, GPU and HPC compute, and planned AI data awareness using IBM FDC, CAS, and watsonx

Research archive lifecycle platform

A biomedical research institute runs a complete IBM data lifecycle at a single site. Data moves from IBM FlashSystem staging to ESS 3500 primary storage, then through IBM Storage Protect, IBM Storage Archive, and IBM Diamondback to TS4500 tape. Starfish provides the data discovery and preparation capability described above, with IBM Fusion Data Catalog under evaluation alongside it.

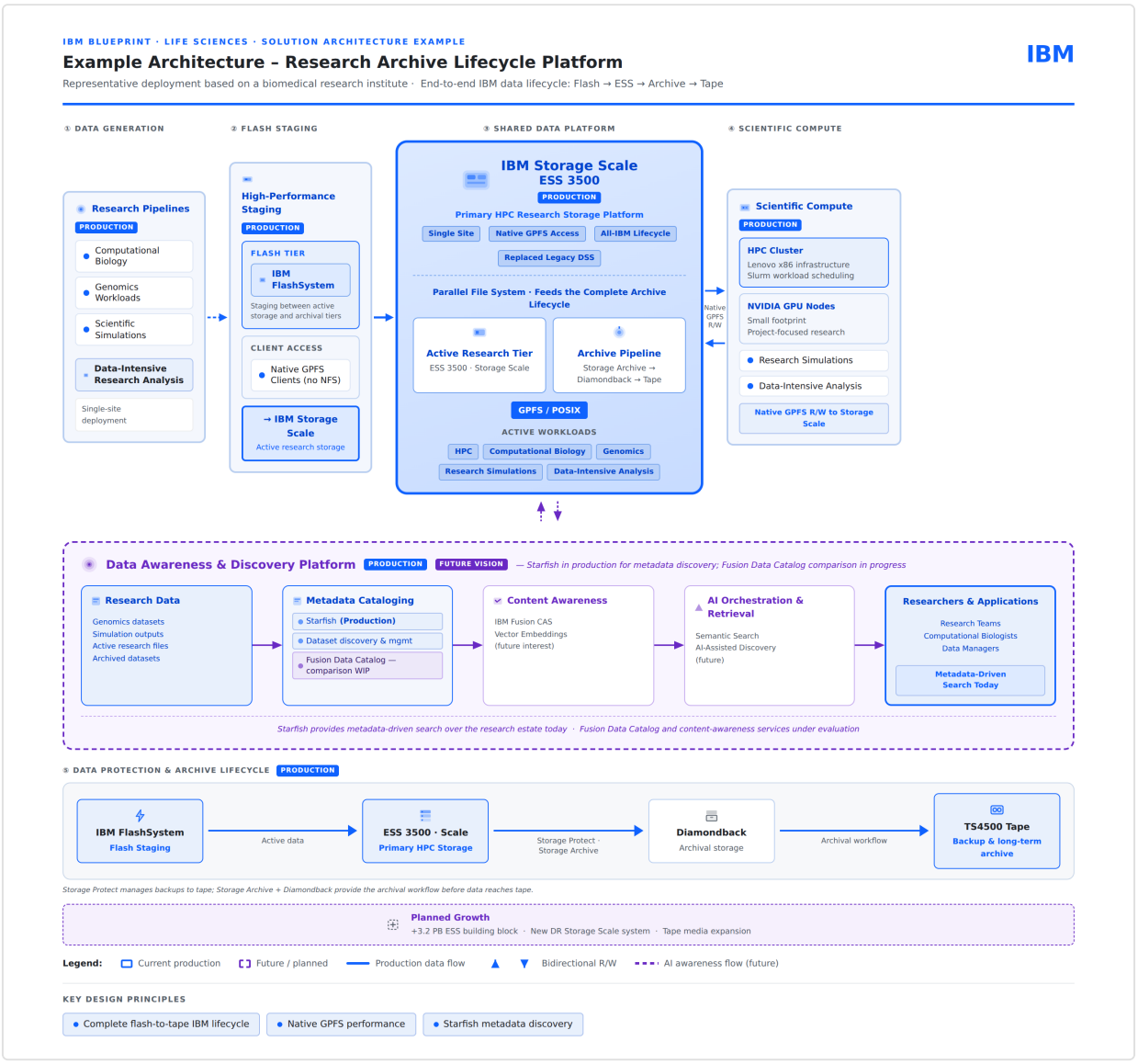

Figure 2: Research Archive Lifecycle Platform — biomedical research institute with end-to-end IBM data lifecycle from IBM FlashSystem flash staging through ESS 3500 HPC storage to TS4500 tape archive via Storage Protect, Storage Archive, and Diamondback

PB-scale bioinformatics HPC and archive platform

A bioinformatics institute went into production in early 2026 with approximately 31 PB of Storage Scale System 6000 capacity for HPC scratch and project storage. A 110+ PB tape archive sits behind an object storage ingestion layer, and AFM replicates data between campus datacenters. This is the genome sequencing pattern at archive scale. A cataloging and governance layer is recommended as a next step in this blueprint.

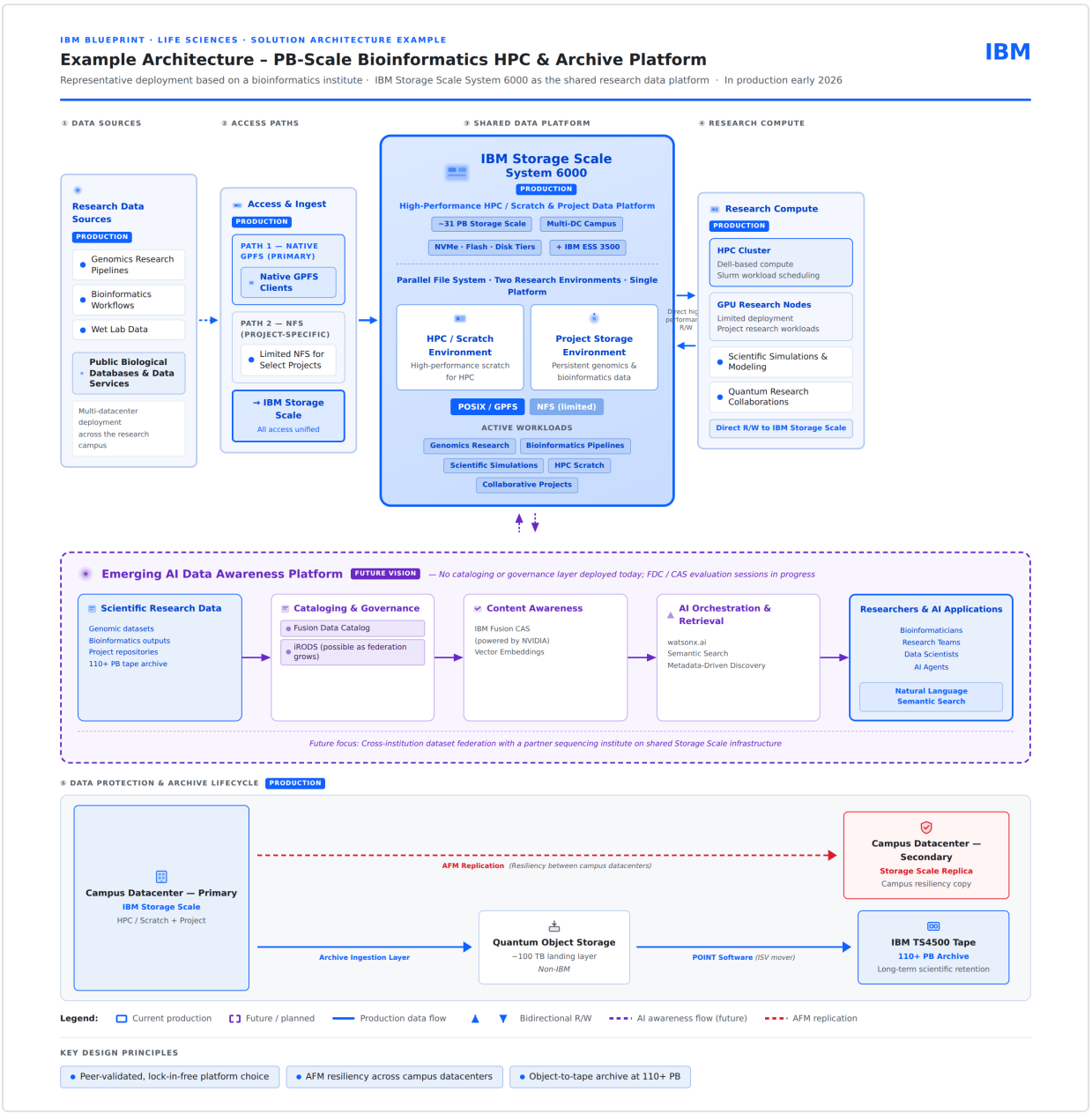

Figure 3: PB-Scale Bioinformatics HPC and Archive Platform — major bioinformatics institute with 31 PB IBM Storage Scale System 6000, AFM inter-datacenter replication, and 110+ PB TS4500 tape archive

Governed genomics research data platform

A large-scale genomic sequencing institute manages roughly 61 PB of Scale storage across a primary campus and a dedicated recovery site. iRODS provides institutional separation, governance, and workflow orchestration on Storage Scale System 6000, supporting multi-institutional collaboration. AFM replication to the secondary site, tape backup, and GKLM encryption protect the platform.

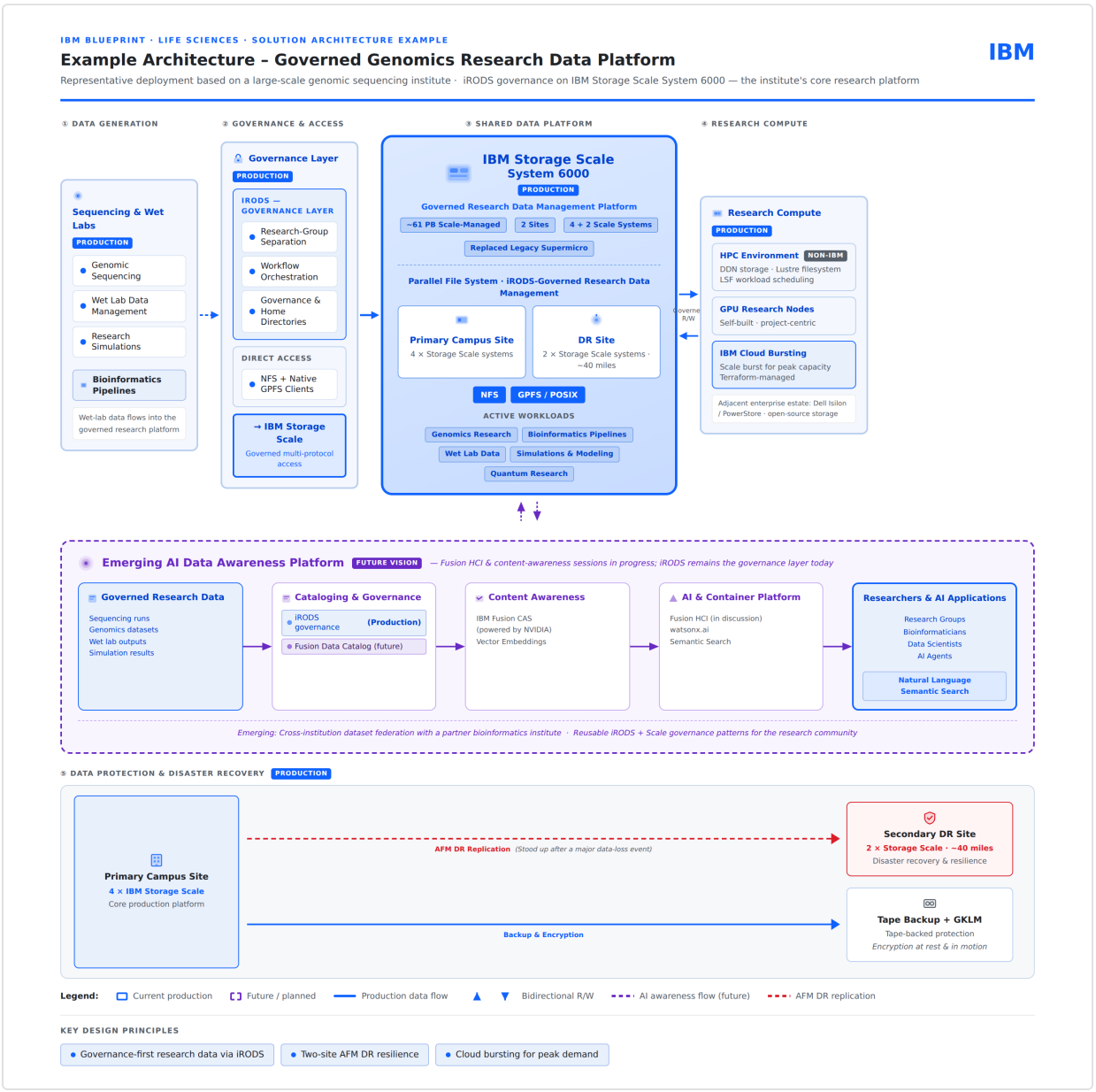

Figure 4: Governed Genomics Research Data Platform — large-scale genomic sequencing institute with 61 PB IBM Storage Scale System 6000, iRODS governance and federation, two-site AFM DR replication, and GKLM-encrypted tape backup

Multi-site research data fabric

A research university extends multi-institutional collaboration to continental scale. A 15 to 16 PB home cluster acts as the single point of truth, all-NVMe IBM Storage Scale System 6000 provides HPC scratch, and AFM caches reach partner institutions more than 1,000 miles away, with every access protocol served at record scale. IBM Fusion Data Catalog and Content-Aware Storage are planned on a Red Hat OpenShift and IBM Fusion HCI foundation that also supports a sovereign AI build.

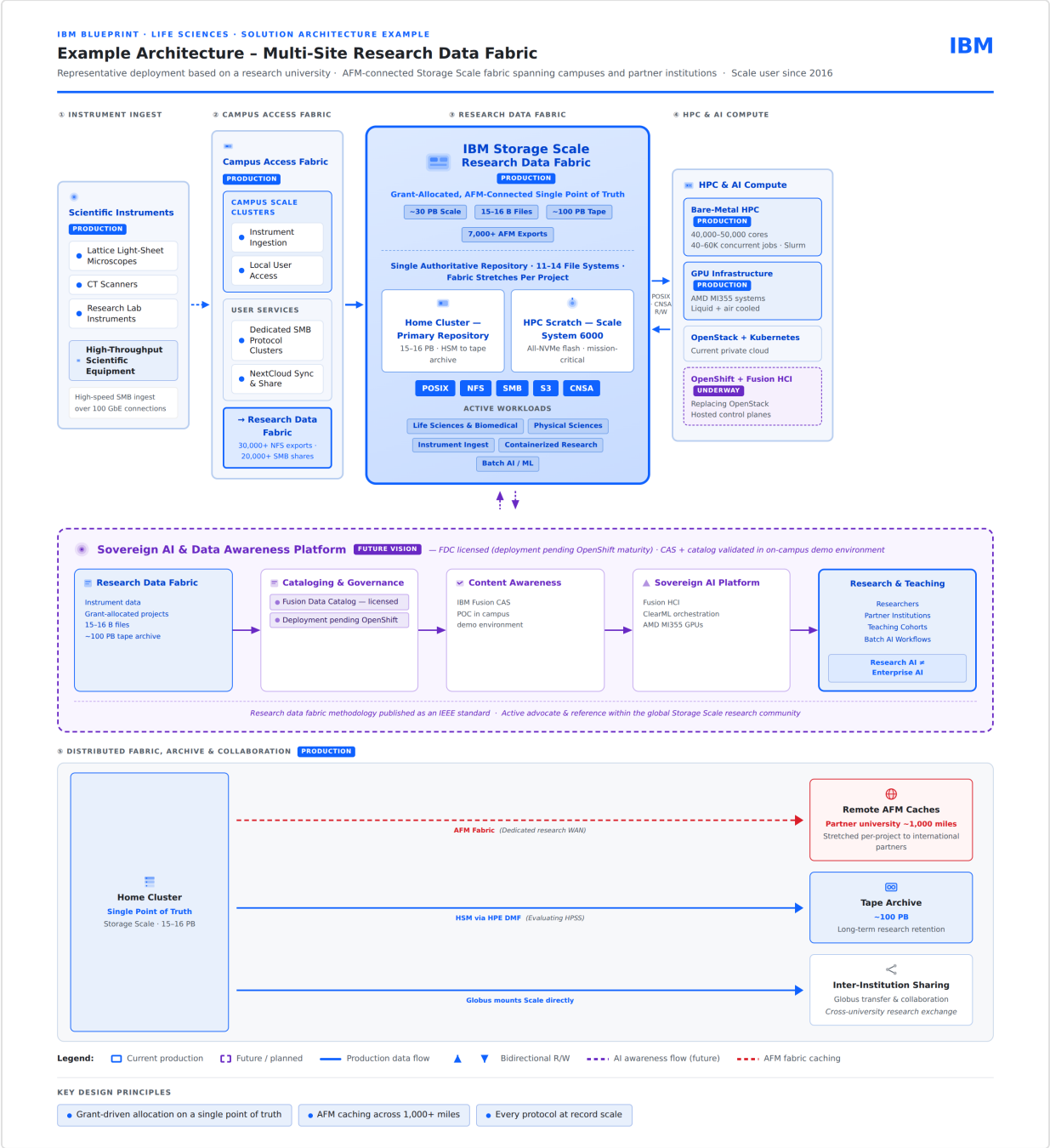

Figure 5: Multi-Site Research Data Fabric — research university with 30 PB IBM Storage Scale, long-range AFM caching to partner institutions over 1,000 miles away, all-NVMe HPC scratch, 100 PB tape archive, and planned sovereign AI platform on IBM Fusion HCI

Metro-resilient research HPC and AI platform

A medical research college runs a synchronously replicated stretch cluster across three metro sites: two data copies, three metadata copies, one namespace. The platform serves HPC, Parallel Works, and researcher home directories; Tuxera provides SMB access. Five GPU-equipped IBM Fusion HCI racks host an OpenShift AI Model-as-a-Service proof of concept, demonstrating the on-premises AI pattern from the clinical imaging and AI research use cases above.

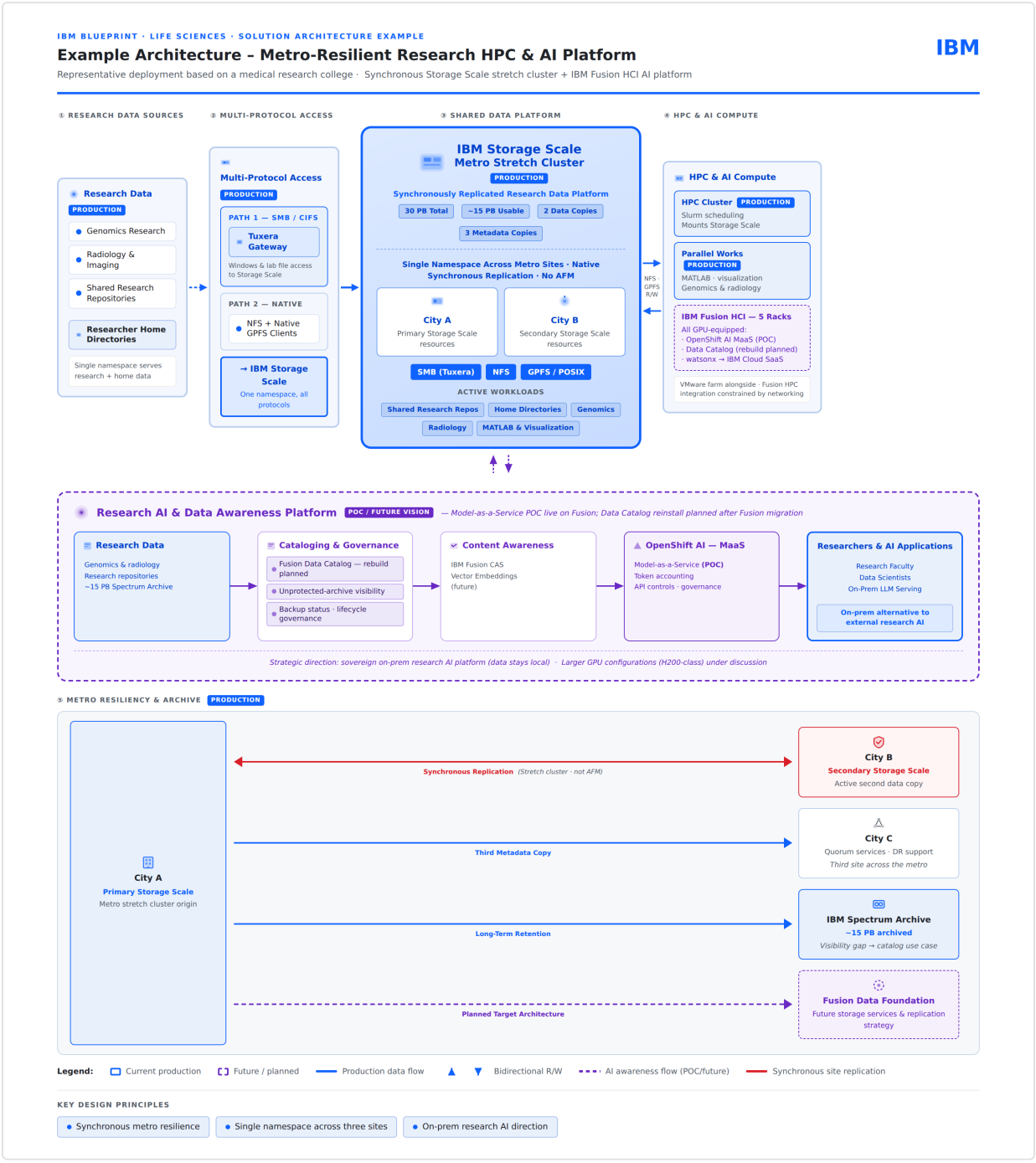

Figure 6: Metro-Resilient Research HPC and AI Platform — medical research college with synchronous three-site IBM Storage Scale stretch cluster, Tuxera SMB gateway, five IBM Fusion HCI GPU racks running OpenShift AI Model-as-a-Service, and 15 PB IBM Spectrum Archive

Customer benefits

Elimination of copy-based workflows

Scientific data can be ingested, processed, analyzed, and archived on a shared platform rather than being repeatedly copied between instrument workstations, storage systems, and compute environments. This reduces unnecessary data movement, storage duplication, and the operational overhead associated with managing multiple copies of the same dataset.

High-speed ingest further reduces the traditional local write-then-copy pattern at the instrument, allowing downstream analysis to begin as soon as data becomes available on the shared platform.

Greater data visibility and reuse

Beyond providing high-performance access to scientific data, the architecture incorporates metadata awareness, content awareness, AI integration, and lifecycle management. Capabilities such as FDC, CAS, the AIDP, and MCP Server enable data to be discovered and understood through metadata, content, and AI-assisted workflows.

This improves the ability of researchers and applications to locate existing datasets, understand their context, and reuse valuable scientific data rather than allowing it to become difficult to find after its initial use.

Governance without unnecessary data movement

Because data can remain on the shared platform throughout more of its lifecycle, organizations can apply governance, protection, and lifecycle policies without continually moving datasets between disconnected environments. For on-premises deployments, sensitive or regulated scientific data can also remain within the organization's governed infrastructure while still supporting HPC and AI workflows.

Guidelines and design considerations

This section provides practical guidance for architects and engineers deploying the IBM Data Highway Blueprint. A scientific data pipeline is only as strong as the decisions made at each of its layers — from how data arrives off the instrument through to how it is protected, governed, and made available to AI workloads. Poor design at any point in the pipeline creates bottlenecks, data silos, or gaps in governance that are difficult to correct after deployment.

The following subsections follow the flow of data through the architecture and outline the components that should be considered during implementation and planning of the data pipeline, as well as further considerations for data sharing and operations.

Ingest design

Ingest design addresses how scientific data enters the platform, including the following considerations:

- Protocol node sizing relative to ingest throughput requirements
- Tuxera Fusion SMB deployment options
- Network requirements: InfiniBand or RDMA-capable NICs
- SMB Direct configuration and Windows workstation prerequisites

Storage design

Storage design covers the following tiering, sizing, and capacity planning considerations:

- ESS 6000 building block configurations
- Data tier planning for cost vs. performance constraints downstream
- NVMe / HDD capacity tier based on workloads
- Policy-based tiering rules
- File system sizing for projected data growth

Compute design

Compute Design addresses GPU sizing, workload placement, and application deployment considerations:

- GPU node sizing for AI inferencing workloads

- Container vs. Virtual Machine decisions for legacy applications
- HPC and Fusion HCI workload allocation

Data management

Data Management covers the following metadata, integration, and configuration considerations:

- Metadata tagging strategy
- FDC and iRODS integration and deployment
- Globus integration for data sharing and transfer across the pipeline

Awareness layer

Awareness layer addresses the services that transform stored data into an AI-queryable asset:

- CAS sizing for vectorization workloads
- Vector database capacity planning based on file content volume
- MCP Server deployment and AI agent access
- AIDP integration with NVIDIA AIDP

High availability and scalability

High availability and scalability covers redundancy and cluster sizing to ensure the pipeline remains resilient as data volumes and workload demands grow:

- Node redundancy and transparent failover
- Multi-site replication and disaster recovery
- Distributed storage cluster sizing
- Active File Management for multi-site caching

Sample configurations

Sample configurations provides the following reference sizing anchors:

- Starter deployment: under 5 PB, single rack
- Mid-tier research deployment: approximately 30 PB
- Large institution deployment: 100 PB or greater, multi-rack

Multi-institutional collaboration

Multi-institutional collaboration and data privacy and compliance address the governance, regulatory, and cross-institution considerations that are particularly important in clinical and regulated research environments.

Universities, national labs, and research hospitals can all share data through policy-based governance. Researchers can securely share datasets across institutions while maintaining governance policies, metadata context, and data provenance.

Data privacy and compliance

Data privacy and compliance addresses the following concerns:

- Life sciences data carries regulatory weight that general enterprise data does not.

- Clinical and patient-related data falls under privacy regulations such as HIPAA; genomic datasets are frequently controlled-access; validated environments operate under GxP and 21 CFR Part 11.
- Cross-border collaboration introduces obligations under regulations such as GDPR.

Partner channel

Federal and DOE deployments with certified IBM partners.

Operational considerations

Operational Considerations covers the monitoring, capacity, and service lifecycle considerations to keep the full stack healthy over time:

- Monitoring and observability across the full stack: Storage Scale health, protocol node performance, FDC indexing throughput, CAS vectorization queue depth, and MCP Server activity.
- Capacity planning for both primary storage and the supporting services that scale with it.
- Metadata growth in the FDC catalog as file count grows — plan for catalog database expansion and query performance over time.
- Vector growth in the CAS vector database as files are ingested and re-vectorized — size the vector index storage independently from the file data it represents.
- Service lifecycle management for the awareness layer: rolling upgrades, schema evolution, and recovery procedures for the catalog and vector store.
- Cost attribution and chargeback across storage tiers, compute, and AI services for organizations with internal billing models.

Conclusion

Life sciences organizations have traditionally focused on solving the challenges of volume, velocity, variety, and veracity. As scientific data environments become increasingly distributed, dynamic, diverse, and affected by growing amounts of dark data, organizations require more than high-performance storage alone. They require intelligent data management that enables AI-driven scientific workflows across the complete lifecycle of scientific data — from ingest and analysis to governance, protection, discovery, and reuse.

The IBM life sciences blueprint addresses these challenges through four core architectural capabilities:

- Abstraction: Providing a data platform that unifies across storage tiers, locations, and environments.
- Acceleration: Enabling high-speed ingest, movement, and access to scientific datasets.
- Access: Delivering concurrent multi-protocol access for researchers, applications, high performance computing (HPC) environments, and AI workloads.
- Awareness: Delivering intelligent data management through metadata discovery, contextual understanding, content awareness, and AI-driven orchestration.

These capabilities are delivered through an integrated IBM data platform that includes the following components:

- IBM Storage Scale as the foundation
- Tuxera Fusion SMB for high-performance ingest from Windows-based instruments
- Fusion HCI for AI, analytics, and HPC workloads
- Fusion Data Catalog (FDC), Content-Aware Storage (CAS), MCP Server, and AI Data Platform (AIDP) for intelligent data management and AI-enabled scientific workflows
- Storage Protect, IBM Cloud Object Storage, and Atempo Miria for data protection and lifecycle management
- Third-party governance integrations for collaboration, governance, and federation

Many storage architectures focus primarily on capacity and performance. The IBM life sciences data blueprint extends beyond traditional storage by combining high-performance infrastructure with intelligent data management, enabling organizations to transform scientific data into a strategic research asset that can be discovered, governed, protected, shared, and reused throughout its lifecycle. As AI-driven research continues to evolve, organizational success will depend not only on storing data, but on understanding it, managing it intelligently, and enabling AI to act on it — to completely transform life sciences research.

Abbreviations

Abbreviation	Definition
AFM	Active File Management — Storage Scale's built-in replication and caching capability between sites
AIDP	AI Data Platform — NVIDIA's reference design for AI infrastructure, paired with IBM Storage Scale
CAS	Content-Aware Storage — IBM Storage Scale capability that creates vector embeddings from file content to support semantic search
CES	Clustered Export Services — the protocol node layer in IBM Storage Scale that serves SMB, NFS, and S3
COS	Cloud Object Storage — IBM's S3-compatible object storage service
DGX	NVIDIA's integrated GPU server system for AI training and inference
ESS	Elastic Storage System — IBM's appliance version of IBM Storage Scale
FDC	Fusion Data Catalog — IBM's data catalog for unstructured data that indexes metadata across file, object, and cloud storage
GDPR	General Data Protection Regulation — EU data privacy regulation relevant to cross-border research collaboration
GKLM	IBM Guardium Key Lifecycle Manager — manages encryption keys for IBM storage platforms
GPFS	General Parallel File System — the original name for IBM Storage Scale
GPU	Graphics Processing Unit — used for AI training, inference, and advanced analytics workloads
HCI	Hyperconverged Infrastructure — pre-integrated compute, storage, and networking in a single appliance (e.g., IBM Fusion HCI)
HGX	NVIDIA's modular GPU server platform for high-density AI workloads
HIPAA	Health Insurance Portability and Accountability Act — US regulation governing the privacy and security of patient health information
HPC	High-Performance Computing — clusters of servers working in parallel on large scientific workloads
HSM	Hierarchical Storage Management — automated policy-based data migration between storage tiers
IBM	International Business Machines Corporation
iRODS	Integrated Rule-Oriented Data System — open-source platform for policy-based research data management

Abbreviation	Definition
MCP	Model Context Protocol — open standard allowing AI assistants and agents to connect to external tools and data sources
MCPS	Model Context Protocol Server — the IBM implementation of an MCP server that exposes storage, metadata, and CAS services to AI agents
NFS	Network File System — standard network file-sharing protocol for Linux environments
NVMe	Non-Volatile Memory Express — high-speed flash storage interface used for hot-tier scientific data
ONT	Oxford Nanopore Technologies — a genomic sequencing platform
POSIX	Portable Operating System Interface — native file access interface for Linux and HPC applications
RAG	Retrieval-Augmented Generation — AI technique in which a model retrieves relevant data before generating an answer
RBAC	Role-Based Access Control — access management approach assigning permissions based on user roles
RDMA	Remote Direct Memory Access — networking technology that moves data directly between system memory over InfiniBand or high-speed Ethernet
S3	Simple Storage Service — object storage interface popularized by AWS and used by IBM Cloud Object Storage
SMB	Server Message Block — network file-sharing protocol used by Windows-based instruments and workstations
SOC 2	Service Organization Control 2 — security compliance framework relevant to cloud and data services
TB	Terabyte — unit of data storage equal to 1,000 gigabytes
PB	Petabyte — unit of data storage equal to 1,000 terabytes
EB	Exabyte — unit of data storage equal to 1,000 petabytes
VM	Virtual Machine — software-based emulation of a physical computer, supported by OpenShift Virtualization on Fusion HCI
WORM	Write Once Read Many — storage policy that prevents modification or deletion of data after it is written, used for compliance archiving

Notices

This information was developed for products and services offered in the US. This material might be available from IBM® in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing IBM Corporation North Castle Drive, MD-NC119 Armonk, NY 10504-1785 United States of America

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

Trademarks

The following terms are trademarks of the International Business Machines Corp.

AI Attribution

This work was primarily human-created. AI was used to make stylistic edits, such as changes to structure, wording, and clarity. AI was used to make content edits, such as changes to scope, information, and ideas. AI was prompted for its contributions, or AI assistance was enabled. AI-generated content was reviewed and approved. The following model(s) or application(s) were used: IBM Bob and Microsoft CoPilot.

Version History

Version 1.0 - Initial Publication

Document Number: MD260039

Title: IBM Data Highway Blueprint for Life Sciences: Built for AI-Ready Scientific Data

Authors:

- Tran Nguyen
- Andy Tran

Contributors:

- Matthew Klos
- Meghan Grable

Initial Release Content:

- Section 1: Executive Summary
- Section 2: Introduction
- Section 3: Key Use Cases & Customer Benefits
- Section 4: Reference Architecture Overview
- Section 5: Guidelines & Design Considerations
- Section 6: Conclusion

Publication Date: 2025

Summary: Initial publication providing a reference architecture and deployment guidance for IBM's life sciences data highway — an end-to-end platform for ingesting, processing, governing, and reusing scientific data at scale. Covers IBM Storage Scale, Tuxera Fusion SMB, IBM Fusion HCI, IBM Fusion Data Catalog, Content-Aware Storage, MCP Server, and NVIDIA AI Data Platform integrations, with six real-world customer architecture examples.