IBM Storage Scale

Software-defined storage for building a global data platform for AI, analytics, and hybrid cloud

Highlights

Global data abstraction services ensure seamless connectivity across multiple sources and locations.

Extensive protocol support, including NFS, SMB, S3, HDFS, POSIX, NVIDIA GPUDirect Storage, and CSI, with new support for NFS nconnect and SMB Multichannel.

Robust data resilience safeguards against ransomware and evolving cyber threats.

IBM Storage Scale System as appliance is also available, offering rapid deployment and building-block expandability.

Organizations worldwide are restructuring their data resources to modernize and capitalize on the opportunities presented by artificial intelligence (AI). However, this transformation comes with several key challenges:

- More data: Enterprises are generating and storing data at unprecedented rates, a trend that shows no signs of slowing down.
- More locations: Modern data strategies rely on distributed storage architectures to optimize performance, cost, and resilience.
- More formats: Organizations must manage a mix of structured data (SQL databases), semi-structured data (web pages, social media posts, and log files), and unstructured data (text, video, audio, and IoT sensor data).

A significant portion of this data is unstructured, generated by AI/ML workloads, analytics, data lakes, IoT, cloud-native applications, and backup and archive solutions. To ensure accessibility across geographically distributed applications, services, and devices, this data must reside in scalable, distributed file and object storage systems.

IBM Storage Scale is designed to address these evolving data demands. It provides global data abstraction services, enabling seamless data connectivity across multiple sources and locations. This capability extends to both IBM and non-IBM storage environments, making it an ideal solution for heterogeneous infrastructures. Built on a massively parallel file system, IBM Storage Scale can be deployed across multiple hardware platforms, including x86, IBM Power, IBM Z, ARM-based POSIX clients, virtual machines, and Kubernetes environments.

Unlocking AI Potential with Content-Aware Storage
Very little enterprise data has been indexed for generative AI applications, which prevents AI assistants from providing accurate, up-to-date answers. The content-aware storage capabilities in Storage Scale address this challenge by extracting the semantic meaning hidden inside unstructured data so that AI assistants can automatically generate smarter answers. Storage Scale enriches data using embedded compute and data pipelines that minimize data movement and latency to help reduce costs and improve performance.

For the 17th consecutive year, IBM is recognized in the 2025 Gartner® Magic Quadrant™ for Enterprise Storage Platforms, listing IBM Storage Scale as a Leader for unstructured data.

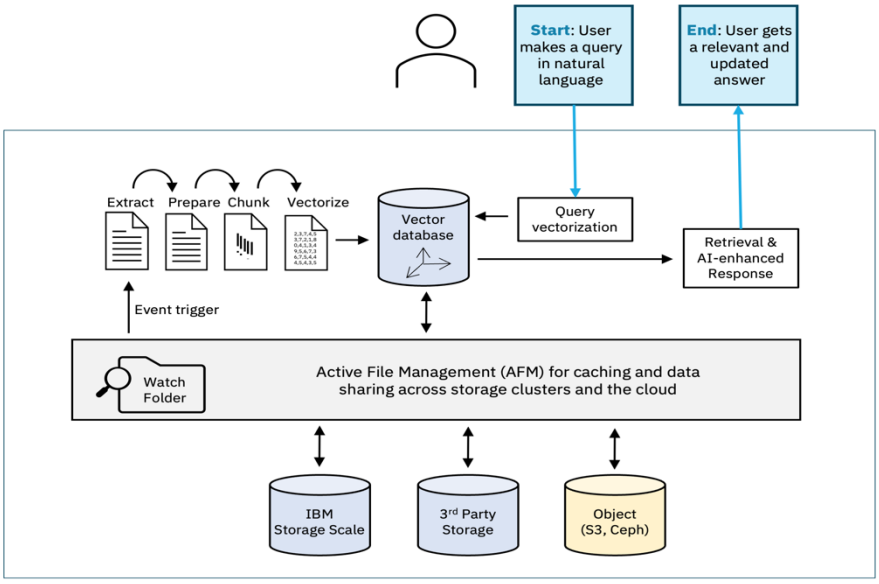

Figure 1. Storage Scale enhances AI-driven workflows by integrating Active File Management (AFM) and Content-Aware Storage to streamline data processing and retrieval.

Storage Scale automates data extraction, vectorization, and storage updates, enabling seamless retrieval via a vector database. When users submit natural language queries, AI enhances search results for optimized responses. The table below summarizes how Storage Scale handles data updates and AI-driven query processing.

Step	Data update process based on storage changes (Storage Event Trigger)	Query processing and response generation (User-initiated Trigger)
1	A file is created, modified, or deleted in storage	User submits a search query in natural language
2	The event trigger detects the change and activates the processing pipeline	The query is converted into a vector representation for similarity search
3	The changed data is extracted, prepared, chunked, and vectorized.	The storage system retrieves the most relevant data from the vector database
4	The processed data is stored in the vector database for future retrieval	Relevant data is retrieved, and an AI model enhances the response
5	Updated data is made available for AI-driven retrieval and response generation	User receives a relevant and updated answer

The new release introduces asynchronous notifications within the Content-Aware Storage capability, enabling faster, event-driven data ingestion into AI inferencing workflows, reducing latency and improving operational efficiency.

New Data Acceleration Tier (DAT)
Built on an NVMeoF-based architecture, DAT enhances the performance and efficiency of IBM Storage Scale by addressing IOPS bottlenecks that limit real-time AI workloads. It is engineered to handle the small, random I/O patterns typical of AI inferencing, delivering consistent, predictable performance across large-scale deployments.

Working in conjunction with IBM Storage Scale System 6000, DAT provides extreme IOPS and ultra-low latency for AI inferencing workloads, enabling faster data access and decision-making across hybrid environments.

DAT offers the following key enhancements:

- High-performance design delivering up to 28 million IOPS and 340 GB/s throughput in NVMeoF configurations.
- Optimized for small and random I/O patterns to improve responsiveness in AI inferencing and real-time analytics.
- Seamless hybrid integration with existing IBM Storage Scale environments, supporting both dedicated acceleration tiers and mixed clusters.
- NVMe-optimized data path that enhances data ingestion and retrieval efficiency at scale.

The main use cases for IBM Storage Scale include GPU-accelerated AI, big data analytics and data lakehouses, high-performance computing (HPC) for scientific simulations and complex computations, IT modernization, and backup and archiving. These workloads benefit from automated tiering, advanced data placement, and secure long-term data retention enabled by the platform's global data architecture.

Together with Content-Aware Storage, DAT forms a unified data foundation that combines performance acceleration with intelligent data management for AI-driven workloads.

Scalable File and Object Storage Services

Storage Scale provides a flexible, scalable approach to data management, allowing organizations to start with Base Data Services and seamlessly deploy and consume Abstraction and Advanced Data Services as needed. This enables efficient, AI-optimized, and resilient storage tailored to evolving business demands.

Base Data Services
Base Data Services provide the foundation for scalable, high-performance data management, enabling seamless access across AI, analytics, and HPC workloads. With automated data tiering, advanced caching, and multi-location accessibility, these services optimize storage efficiency, streamline operations, and enhance resource utilization. Built-in data protection, governance, and access control ensure security and reliability, enabling organizations to manage large-scale datasets with confidence across hybrid and multicloud environments.

Abstraction Data Services

Abstraction Data Services provide a differentiated capability to enhance AI economics by abstracting external storage and unifying diverse storage systems under a single global namespace. This enables simultaneous access via multiple protocols, streamlining data management across heterogeneous environments. By delivering a consistent, high-performance, and seamless experience, these services optimize both new and existing storage infrastructures, ensuring efficiency and scalability for AI-driven workloads.

Base Data Services	Abstraction Data Services	Advanced Data Services
Base File System • Multi-protocol access • Quotas, QoS, Snapshots • Remote mount • ILM policies • Scale-to-scale caching • Single-site tiering	***Abstract External Storage*** • Differentiated capability to improve AI economics • Different storage systems in a single global namespace, accessible via multiple protocols • Consistent, high-performance, seamless experience for new or existing storage	***Enterprise Resiliency*** • Encryption, Compression • Disaster Recovery (DR), High Availability (HA), Multi-site tiering • Safeguarded copy ***AI Data Services:*** • Content-Aware Storage • Fusion Data Catalog ***Erasure Code Services*** • Storage-rich servers

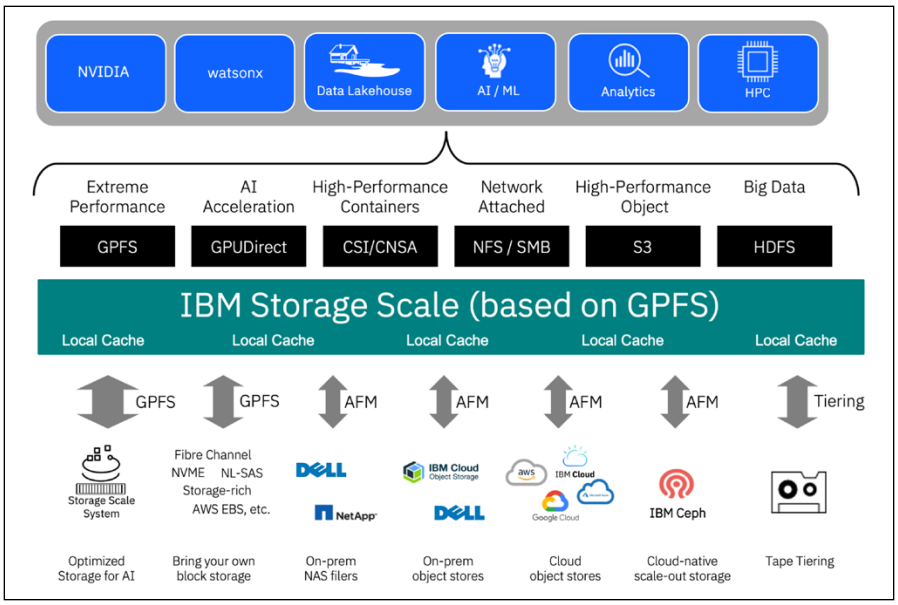

Figure 2. IBM Storage Scale provides a global data platform for your organization's geographically dispersed devices, data sources, and workloads.

Figure 3. IBM Storage Scale System 6000 is a hardware appliance that allows you to deploy IBM Storage Scale on thousands of nodes with TB/s performance, low latency, and up to 28 million IOPS and 340 GB/s throughput in NVMe deployments.

IBM Storage Scale System

Storage Scale is also available as an appliance, IBM Storage Scale System, for streamlined, rapid deployment complete with IBM support services. This option is designed for organizations wanting to build high-performance global data storage capabilities in their own data centers or co-location facilities.

Optimized for AI / NVIDIA workloads

IBM Storage Scale System 6000 and IBM Storage Scale deliver extreme performance for GPU-accelerated workloads. The expanded NVIDIA integration adds CNSA support for GPUDirect Storage, enhanced Base Command Manager connectivity, and NVIDIA Nsight support for advanced observability and debugging. Additional NVIDIA certifications and validated reference architectures align with NVIDIA BasePOD, SuperPOD, and Grace Blackwell platforms, ensuring high performance and compatibility across AI training and inferencing environments.

For more information

Find further details about IBM Storage Scale by contacting your IBM representative or IBM Business Partner, or visit ibm.com/products/storage-scale.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
October 2025