White Paper

AI-Ready Data Storage Infrastructure — Definition, Taxonomy, Ontology, and Future Outlook

Sponsored by: NetApp

Phil Goodwin Ashish Nadkarni Dave Pearson Carol Sliwa
Johnny Yu
August 2025

IDC OPINION

Artificial intelligence (AI) — despite the hype — is poised to become one of the most impactful technology evolutions of this decade. Only a short time ago, the industry was abuzz with generative AI (GenAI), followed shortly by retrieval-augmented generation (RAG) AI and now agentic AI. One must also not forget that these newer AI modalities stand on the shoulders of predictive, interpretive, behavioral, and other, more mature AI methodologies. AI is becoming a broad umbrella term where additional specificity is needed to frame the discussion. Simply calling a product feature or capability "AI" will neither adequately describe it to the IT buyer nor distinguish it in a marketplace where AI claims proliferate.

Because of the transformative promise of AI, IT buyers are investing heavily in AI projects and the hardware infrastructure needed to support them. Unfortunately, IDC research indicates that many of these efforts fall short; less than half (44%) of AI pilot projects advance into production. While there may be many reasons for the low success rate, we believe organizations must approach AI projects from a data-centric perspective. Clearly, the quality of data and timeliness are foundational to successful AI learning systems; without these, an AI project cannot succeed. IT organizations are still learning how to properly provision and deploy infrastructure assets to support the various types of AI workloads. As IT buyers search for the optimal infrastructure, IT suppliers must not only align their solutions with the particular requirements but also articulate the differentiating advantages they bring to the table.

AI-ready data storage infrastructure (AI-RDSI) requires a combination of hardware and software that builds upon traditional technologies. We believe AI-RDSI will involve "embedded" AI. Embedded AI refers to AI technologies used internally to enhance system use, performance, reliability, and operational efficiency, and it will be specific to

the hardware or software in which it is embedded. Examples include AI agents, which perform system functional tasks that heretofore were human specific. This level of AI is different from workload AI, which refers to the ability to optimize data storage in support of AI workloads, such as LLMs, agents, and the like.

IT executives understand the necessity of investing in AI projects and the infrastructure to support them. According to IDC research, tech leaders anticipate doubling overall AI budgets and targeted GenAI budgets over the next two years (source: IDC's *Future Enterprise Resiliency and Spending [FERS] Survey, Wave 10,* October 2024). Nevertheless, budgets are not unlimited, and as noted previously, organizations still struggle to bring AI projects from pilot to production. Consequently, expenditures and results are monitored closely at the highest levels of the organization.

According to IDC's October 2024 *Future Enterprise Resiliency and Spending Survey, Wave 10,* 35.1% of respondents said AI-enabled capabilities (i.e., embedded AI) would have the most impact on their team's data management efficiency and effectiveness over the next 18 months. In addition, respondents also cited data management as one of the top inhibitors to the use of GenAI.

It is unlikely that any single supplier can offer everything for everyone or every use case. Thus vendors must be prepared to operate within an ecosystem of partners and competitors to provide a full-stack AI infrastructure offering. Some market participants will be system suppliers and others software-only vendors. Regardless, we expect competition to be robust based on clear differentiation. Those IT organizations best able to best leverage AI for operational purposes and to support customers in their AI journey with AI-ready data storage infrastructure will be best positioned to accelerate growth in the coming decade. Hot products today may become obsolete in such a dynamic market. Extensible architectures, agile development, and responsive research and development will be at a premium during this time.

SITUATION OVERVIEW

Defining AI-Ready Data Storage Infrastructure

AI-ready data storage infrastructure is defined as:

> The hardware, software, and services necessary to prepare, ingest, store, manage, protect, secure, govern, and move data to address the requirements of artificial intelligence applications. AI-RDSI also encompasses service levels pertaining to AI workloads, including performance and system availability; data quality–related attributes, such as trust and provenance; and technologies for the disposition of data post-analysis.

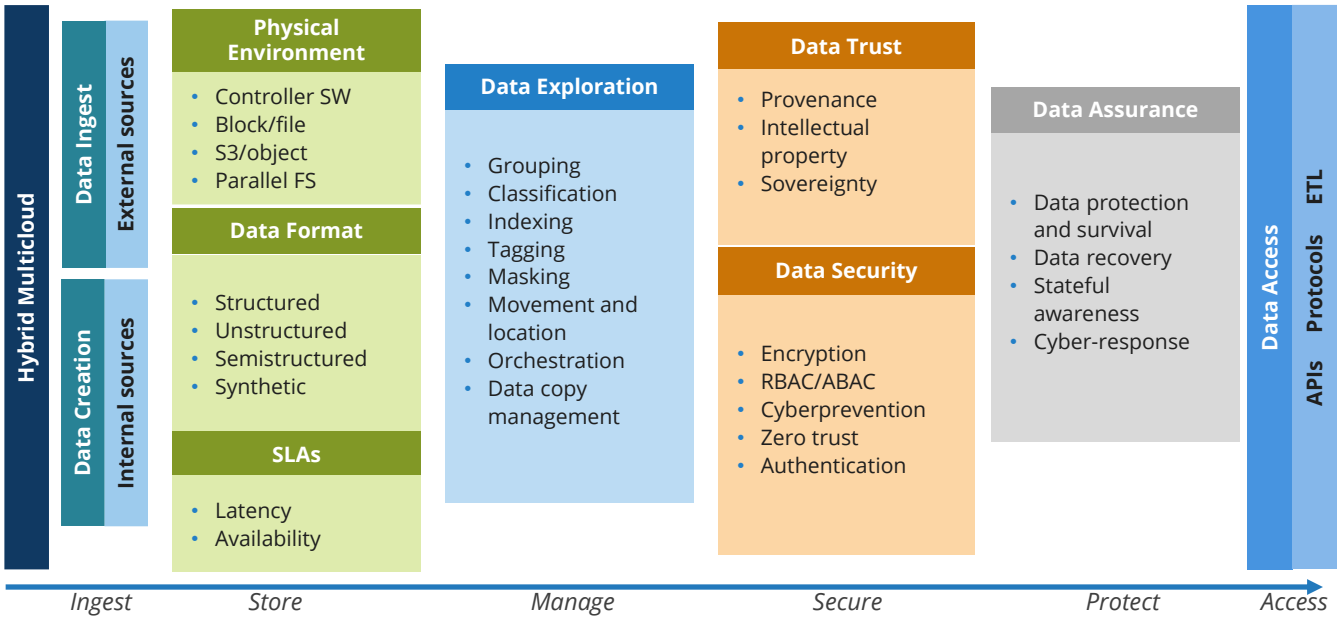
Broadly speaking, data enters an organization's computer systems in one of two ways: either as "organic" data generated by internal operations or data ingested from an external source. IDC's Global DataSphere research shows that 92.3% of stored data in 2023 was unstructured/semistructured, with the remaining 6.7% being structured. However, IDC forecasts that structured data will grow at a 49.3% CAGR through 2028, with unstructured/semistructured data growing at a 21.4% CAGR through the same period. By 2028, IDC forecasts 17.7% of data stored will be structured. Furthermore, 48% of data is on premises, 29% in the cloud, 19% at the edge, and the other 4% in "other" locations. While data is trending toward the cloud, it is a gradual shift. All of this data, regardless of format or location, may be of value to AI workloads.

Data Logistics and AI Workloads

As data enters the organization, it becomes subject to a process described by IDC as "data logistics." Using a package logistics analogy, data logistics is the process of taking data from its origin to destination with guaranteed delivery, quality, security, and timeliness. Data logistics provides the foundation for AI-RDSI, yet AI workloads demand additional levels of rigor. Figure 1 illustrates the data logistical journey for data in AI environments and the technological elements along that journey. Readers should observe the progressive arrow at the bottom in Figure 1. This arrow highlights the journey the data makes from the time it is ingested (external sources) or created (internal sources). This ingestion may be from on-premises workloads or those in the cloud. From that time, the data must be stored according to policies and workload performance requirements. Moreover, data must be classified, indexed, and labeled to ensure proper governance. Security, data trust, and protection are foundational to AI-ready data storage infrastructure.

FIGURE 1

AI-Enabling Data Storage Infrastructure: Ingest to Access



Source: IDC, 2025

When considering the AI-RDSI definition, it is important to also consider what is *not* included. For example, while many of these elements are common to data life-cycle management (DLM), the full scope of DLM is beyond the scope of this study. In addition, data management for AI models themselves is out of the scope of this study.

Readers will also note that data exploration–type activities, data governance, provenance, and the like will also take place when data is ingested into AI workloads (i.e., data management within the AI workload itself). These activities are separate and distinct from AI-RDSI capabilities. This is an important distinction, as IT teams may conflate the two without realizing that both may be required and may be compelled to use two tools that conflict and cause integration challenges.

Single Source of Truth Requirement

IDC research has found that IT organizations deal with an average of 6.4 data silos per organization. Our research has further found that these IT teams must manage 13 copies of data, which may be spread across primary storage, secondary storage, cloud, and edge storage. Multiple copies of data may be created for various reasons, including protection (backup), test/dev, analytics, and archive. All are valid and necessary for operations.

Multiple copies of data present specific challenges for AI workloads. While data timeliness may be less important in data lakes and data analytics applications, it can be highly important for AI. AI accuracy depends on data accuracy and data timeliness to properly learn and respond to changing requirements in a timely manner — sometimes in real time. Ingesting data from obsolete copies may result in AI learning modules "going backward." Thus having a copy data management (CDM) capability within AI-RDSI is vital for AI workloads. Knowing which copy is most current, pruning obsolete copies, or even managing a "golden copy" can help ensure that AI models are constantly working from a single source of truth.

AI-Enabling Data Infrastructure

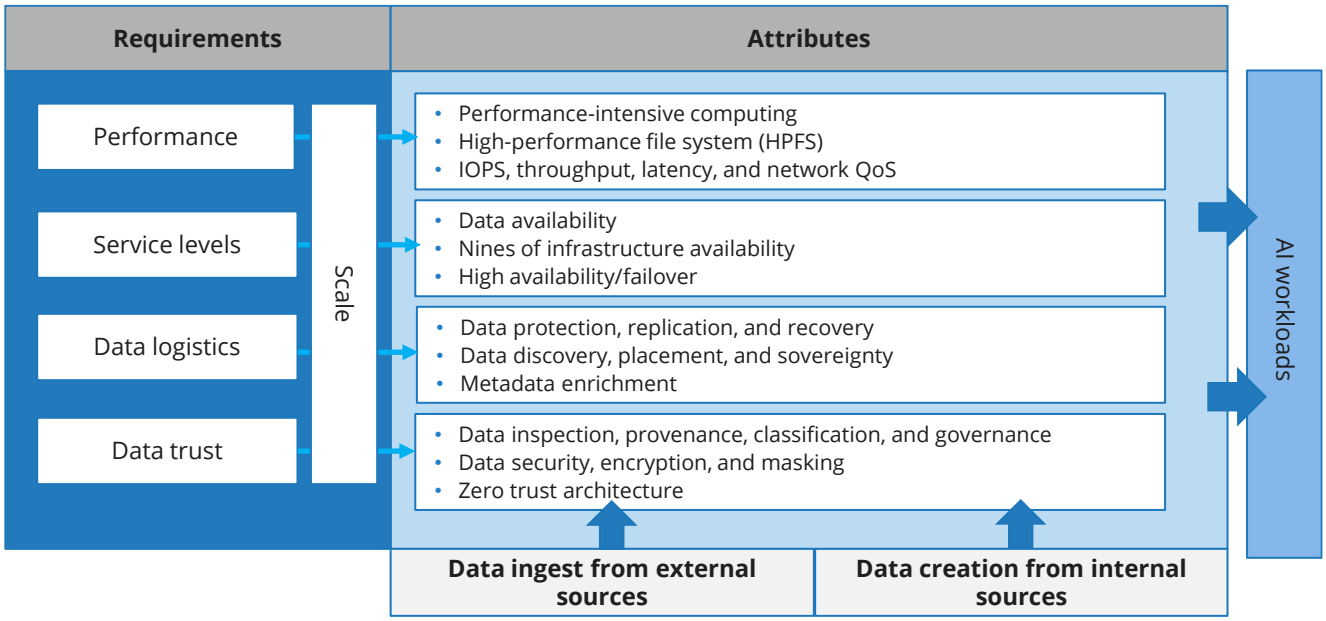
AI workloads may be compute intensive, using a combination of CPUs and GPUs that may number in the dozens or hundreds. To optimize these expensive resources, infrastructure teams must deploy storage systems able to match these requirements. Data delivery to the compute resources should not cause idle compute cycles to occur. Figure 2 shows many of the necessary storage characteristics.

From Figure 2, it can be noted that four primary attributes of data infrastructure exist. A detailed explanation of these attributes is as follows:

- **Performance:** End-to-end storage system performance must consider AI workload needs, including data throughput, input/output operations per second (IOPS), latency, necessary network bandwidth, and the demands of performance-intensive computing (PIC). Achieving high throughput may require the use of technologies such as parallel file systems or parallel NFS (pNFS), and for necessary IOPS and low latency, the use of flash storage or storage-class memory. Aspects of AI workloads with moderate or lower performance needs may incorporate technologies such as tiered storage, object stores, or hard disk drives (HDDs) for cost efficiency.
- **Service levels:** Service-level requirements go hand in hand with performance, but they are more directed toward data availability. Common service levels include "nines" of data availability or total uptime. Five-nines (99.999%) uptime will be a common requirement for AI workloads as downtime will be extremely disruptive.
- **Data logistics:** Data logistics policy engines ensure that data will be delivered to the right place at the right time for AI optimization. This will include location while ensuring adherence to sovereignty requirements.
- **Data trust:** The desire to feed as much data as possible to AI models for optimized learning must be balanced against the need for data quality. Data trust is core to data quality by having appropriate policies and procedures in place to reduce data contamination or tampering.

FIGURE 2

AI-Enabling Data Infrastructure



Source: IDC, 2025

FUTURE OUTLOOK

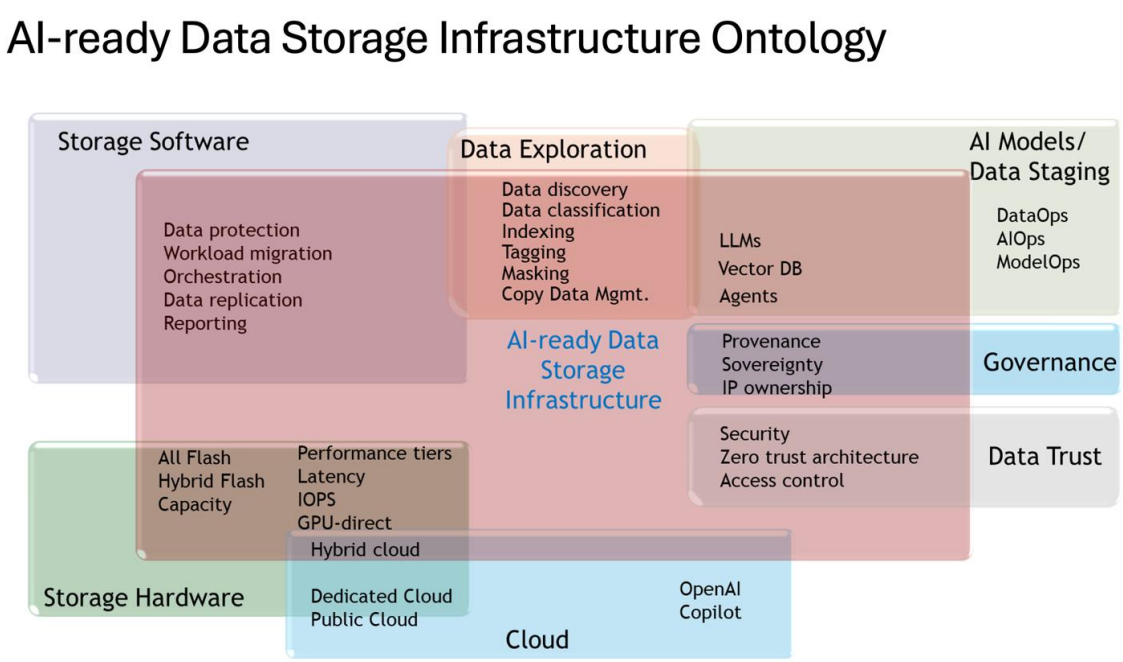
As AI-RDSI is delivered to IT consumers, it can be categorized according to hardware, software, and AI-specific taxonomies. Some IT providers, such as storage systems vendors, will provide capabilities for all three taxonomies. These systems will necessarily be hardware specific, though they may incorporate software from other sources. Independent software vendors (ISVs) will provide many of the capabilities and attempt to do so on a hardware-agnostic basis.

It is unlikely that any single vendor will be able to deliver every necessary capability for AI-RDSI. Thus it is useful to look at the totality of the solution and how the components fit together. Figure 3 shows the AI-ready data storage infrastructure ontology.

Ontological models describe the relationship between elements. Figure 3 illustrates the relationship between the elements of AI-RDSI, regardless of source, platform, or delivery mechanism. This ontological view assembles the major components of AI-RDSI. Each component represents an element of the total solution. IT teams can use an ontology as the basis for establishing an AI-ready data storage infrastructure.

FIGURE 3

AI-Ready Data Storage Infrastructure Ontology



Source: IDC, 2025

Figure 4 details the necessary hardware functionality for AI-RDSI.

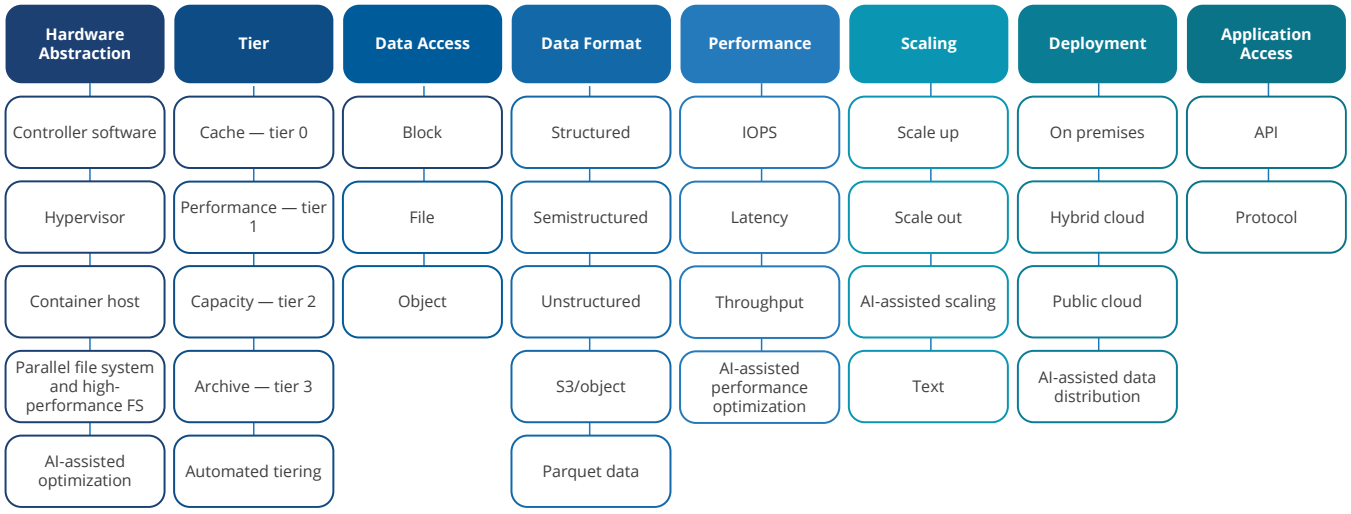
This AI-RDSI hardware taxonomy includes eight categories of componentry:

- **Hardware abstraction:** Virtual infrastructure is well proven to provide greater flexibility of workloads, workload migration, and data location.
  - **AI implications:** AI (and ML) can be used to assist hardware deployment configuration and dynamic performance optimization based on policies or SLAs, fault prediction/detection, isolation, and correction. AI-driven dynamic resource allocation can apply necessary resources, and AI workloads drive performance requirements and balance them with other workload requirements.
- **Tiering:** Storage infrastructure uses up to four tiers of storage types that may require a variety of media technologies such as NAND flash, storage-class memory, and HDDs, with each tier having specific performance requirements.

    - **AI implications:** AI capabilities will be able to predictively allocate resources based on workload requirements or other factors and move data to the appropriate tier in order to provide optimum performance.
- **Data access:** AI-ready storage infrastructure must be able to support structured and unstructured data and protocols/interfaces for block, file, and object storage. Each access method has its uses for specific applications.
- **Data format:** Similar to data access requirements, AI-ready storage infrastructure must be able to support all types of structured, unstructured, and semi-structured data as well as specific file types for high-performance computing and data lakes.
- **Performance:** The various stages of AI workloads' data life cycles create performance characteristics demands on data infrastructure in terms of IOPS, latency, and throughput. Balancing these requirements can involve certain trade-offs.
    - **AI implications:** AI capabilities should be able to move data to the appropriate storage tier or location in order to gain the necessary performance characteristics to meet any workload type.
- **Scaling:** Various workloads may require dynamic scaling for a variety of reasons. The proliferation of data, especially unstructured, in both AI ingest and output can lead to massive and irregular capacity requirements. Moreover, storage system architectures may be either scale up or scale out.
    - **AI implications:** AI capabilities should be able to predictively and dynamically move data to the architecture that best fits the workload requirements.
- **Deployment:** Most organizations used a combination of on premises, private cloud, public cloud, and hybrid cloud to support their workloads. AI-RDSI systems must work in this ecosystem.
    - **AI implications:** AI capabilities can optimize storage placement based on complex requirements of performance, cost, data location, privacy, regulatory compliance, data sovereignty requirements, governance, security, and data protection.
- **Application access:** Storage systems must support the breadth of data access protocols.

FIGURE 4

AI-Ready Storage System Functionality



Source: IDC, 2025

Figure 5 shows the software taxonomy for AI-RDSI.

The AI-RDSI software taxonomy includes four categories:

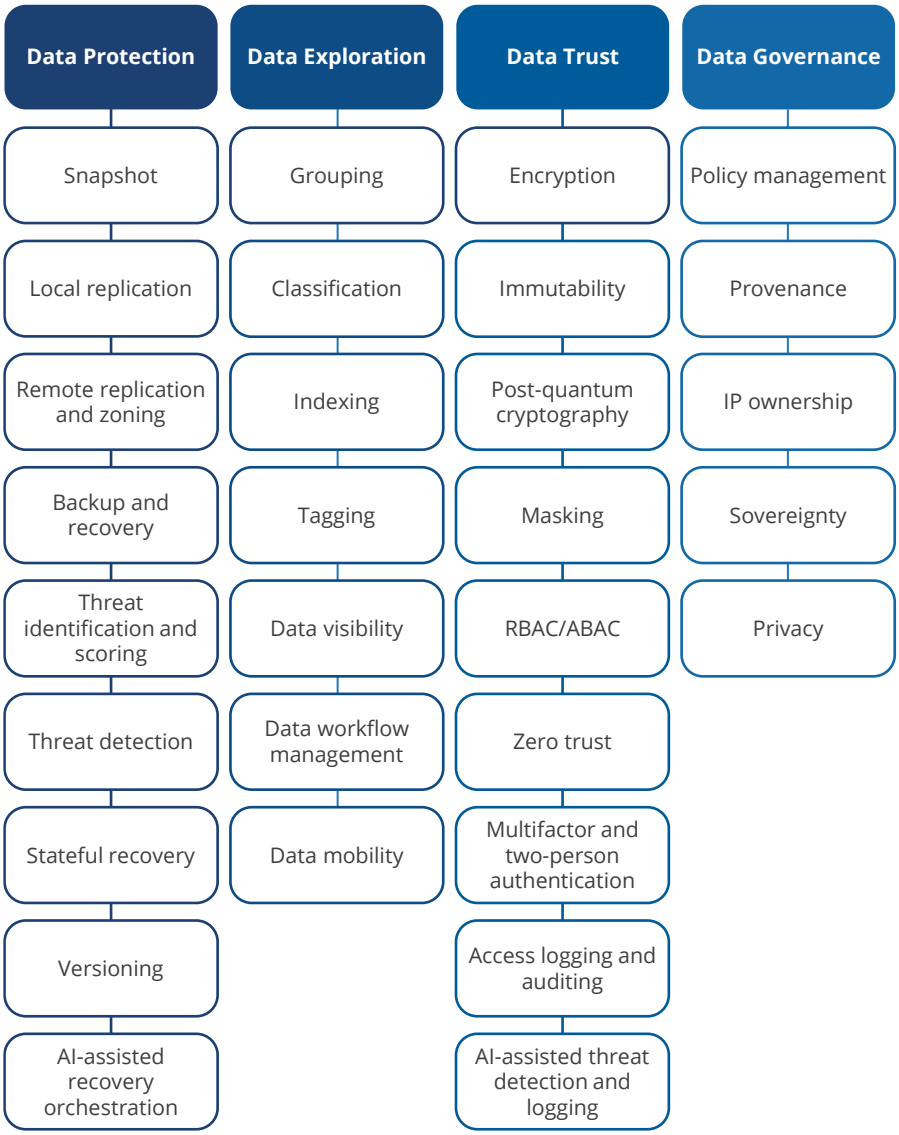
- **Data protection:** Data protection for AI environments begins with the fundamentals of backup and recovery. There are several specific related capabilities needed for AI.
  - **AI implication for data recovery:** AI workloads require stateful protection of AI data stores. These data stores (e.g., LLMs, vector DBs) may gather data in real time from multiple sources. Should one of these sources inject data that should not be ingested, such as proprietary IP, sensitive or prohibited data, and even malicious data, IT teams must have the ability to recover to the point of data injection without losing data from other sources.
  - **AI implications for recovery orchestration:** With the complexity of data recovery, especially where ransomware attacks are involved that affect subsets of data, the time-consuming effort to manually determine the best recovery point and recovery method can delay recovery for days or even weeks. AI-assisted recovery orchestration can help determine exactly what data needs to be recovered and the fastest method of doing so. AI-assisted recovery can also factor in SLA requirements to adjust infrastructure and backup jobs to meet those SLAs.
  - **AI implications for threat detection:** Although threat detection should occur in the network and primary storage systems, it is also necessary in

secondary backup storage. According to IDC research, ransomware attackers attack secondary data before primary data in almost half of all attacks. Thus indicators of data compromise may first appear in backups. AI anomaly detection should be able to correlate seemingly unrelated events and detect attack activity that may not be detectable through conventional signature-based detection. AI should also be able to assist in the detection of malware in data stores.

- **Data exploration:** Data quality is central to AI workload accuracy and effectiveness, and data exploration is central to data quality.
  - **AI implications for data grouping, indexing, and tagging:** Data must be accurately identified and classified as it is stored to ensure that the correct data is fed into AI models. It may be the key to determining the "single source of truth" for AI data feeds in the face of data silos and data redundancy. Data exploration may be AI-driven but isn't necessarily so.
  - **AI implications for data workflow management:** Processing and feeding data into vector DBs or LLMs will involve various workloads to move data and transform and convert it to different store types and so on. Using AI to orchestrate these workflows based on policy engines can reduce manual effort and improve accuracy.
- **Data trust:** Data trust is essential to data accuracy and AI workloads. Data trust is related to data security. Key pillars include data encryption, immutability, multifactor authentication (MFA), and role-based access control (RBAC).
  - **AI implications:** As with secondary storage mentioned previously, AI can assist in threat detection on primary storage to alert on anomalous behavior. AI can also be used for automated and adaptive incident response.
- **Data governance:** Data governance drives data exploration activities. The proper handling of data determines if the data must be masked, encrypted, moved, retained within sovereign boundaries, tracked for provenance, and the like.
  - **AI implications:** Proper data governance feeds data reliability and trust. Governance informs the policy engines that drive data exploration.

FIGURE 5

AI-Ready Data Storage Software Taxonomy



Source: IDC, 2025

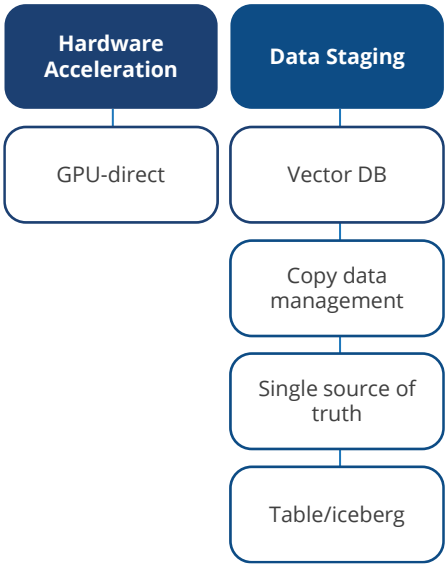
Figure 6 shows optimization for the AI workload taxonomy. In detail:

- **Hardware acceleration:** AI computing infrastructure resources can be quite expensive; optimization is important for the best possible AI project ROI.
  - **AI implications:** Matching storage system performance to computing resources requirements can be accomplished, in part, with GPU-direct data access.

- **Data staging:** The goal of AI-RDSI is to provide the best quality data to feed AI workloads. Data staging, then, is the last step in the process.
  - **AI implications:** Because of the data challenges associated with silos and multiple data copies, copy data management utilities fed by data exploration engines can help ensure the single source of truth for the AI workloads. CDM manages the many copies of data, whether snapshots or clones/mirrors, regardless of location. Some CDM systems can also create "golden copies" of data that provide a virtual view of data to facilitate consistent use across multiple workloads.

FIGURE 6

Optimization for AI Workload Taxonomy



Source: IDC, 2025

ACTIONS TO CONSIDER

Advice to the IT Buyer

- **Characterize AI workloads.** Not all workloads are the same. Very large LLMs analyzing billions of data points will require different infrastructure capabilities from smaller, focused agentic AI models. One size does not fit all, and different vendors will target different workloads as their "sweet spot."
- **Identify the "single source of truth."** The age-old adage of "garbage in, garbage out" applies to AI as much as prior technologies. Data quality and data

currency are vital, and AI developers need access to them. Copy data management, data classification, and tagging may play critical roles in reducing silos and defining the single source of truth.

- **Modernize data storage infrastructure.** To support AI workloads and to maximize AI project success, investing in AI-ready data storage infrastructure will be necessary. Moreover, as systems reach the end of their useful lives, tech refreshes emphasizing AI readiness will help position the organization to be responsive to evolving data requirements.
- **Factor in the extent of the data estate.** The majority of organizations are hybrid multicloud with on-premises and private cloud data repositories as well as multiple public cloud environments. These repositories are often geographically distributed, perhaps globally. This expanse of repositories can lead to data silos, which inhibit data leverage accuracy. AI-RDSI will address these issues using a common data plane across repositories.
- **Look for embedded AI.** Embedded AI, or AI within a solution, can offer many benefits. These may include AI-driven data discovery, classification, and advanced data handling. Other embedded AI systems may drive dynamic infrastructure configuration, workload management, SLA attainment, and so on. These embedded AI capabilities will be core differentiators between solutions.
- **Treat data like a product.** IDC predicts that by 2026, data silos will be reduced by 50% at large enterprises. *Da*ta-as-a-product solutions enable a single enterprise data catalog accessible to various data teams, thereby enhancing collaboration. Data products provide access, assign ownership, and drive business value.

ABOUT IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology, IT benchmarking and sourcing, and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives. Founded in 1964, IDC is a wholly owned subsidiary of International Data Group (IDG, Inc.).

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com