The Future of Data

Suparna Bhattacharya, HPE Labs
Thomas Coughlin, Coughlin Associates
Lance Evans, Hewlett Packard Enterprise
Paolo Faraboschi and **Eitan Frachtenberg**, HPE Labs
Gary Grider, Los Alamos National Lab
Dejan Milojicic, HPE Labs
Sreenivas Rangan Sukumar, Hewlett Packard Enterprise
Alex Veprinsky, Hewlett Packard Enterprise Storage

With artificial intelligence (AI), data is not the past; it's the future. Its value, size, and relevance are meant to continuously fuel every aspect of our lives.

Data has always been important, but even more so today. In the era of artificial intelligence (AI), data also appears as models, tags, voice, video, sensor readings, artificially created data, and much more.

In this article, we discuss the historical and new use of storage and data, with special focus on AI and agentic AI, privacy, and security. We cover storage architecture, media, and its economics. We conclude with a summary and outlook.

HISTORICAL AND MODERN USE

Classical storage workloads include serial and random reading from and writing to one or more storage devices at different operation sizes. Other attributes of storage simplify storage management, like enforcing immutability. There are many access methods for data—at the lowest level, it is file, object, and block, and at higher levels, it is key-value, tabular data, etc. Data can also be local to the producer/consumer or remote via a network, which can have performance and cost implications. Bringing data closer improves both cost and performance. Another important characteristic of data storage is its resiliency, from backup copies on different devices to redundancy schemes like erasure, which distributes redundant data with different overheads and performance characteristics.

Commodity storage devices each have a sweet spot in density/bandwidth/input/output operations per second (IOPS) for read, write, and mixed workloads. Since the 1960s, these economic sweet spots have given rise to multiple storage devices serving multiple simultaneous workloads. One can separate this use of storage devices into two types:

- concurrent, which involves many tasks accessing storage on many different functions, like database queries running separately or users accessing different files
- truly parallel, which is many tasks all accessing storage for a single function, like many processes

Digital Object Identifier 10.1109/MC.2025.3603404
Date of current version: 29 October 2025

EDITOR **DEJAN MILOJICIC**
HPE Labs;
dejan.milojicic@hpe.com

all writing the same, possibly distributed, data set in parallel to storage.

Each of these categories holds the use of multiple storage devices.

Large sets of users sharing large amounts of data gave rise to scale-out network attached storage in the 1990s, which handled most concurrent workloads well. The high-performance computing (HPC) community's uses encompass both concurrent and truly parallel use cases. The latter takes two forms: many tasks, each reading/writing their own file, and many tasks all writing to one or a few files. The first creates many files, often in tightly synchronized parallel access, stressing the file metadata creation/query capabilities, while the second can be harder to control to ensure protecting the producers/consumers from ordering issues. HPC systems have classically not used local storage because the large-scale distributed nature of the applications makes the compute node the most likely part to fail, so local-only storage is more vulnerable. More recently, AI workloads that run the gamut from concurrent to purely parallel can also leverage storage local to the compute to gain speed and efficiency.

Dominant use-cases that demand true parallelism have emerged and evolved with the advent of cloud computing. For cloud workloads, the cloud providers have their own semi-custom solutions that aren't really available outside of their sphere of influence. With AI factories added to the mix, we now have far more large-scale workloads, including true parallel and many highly concurrent workloads at immense scales. AI factories move fast, so completely custom solutions aren't possible, but highly tuned products are favored. AI factories and cloud computing both have remote storage needs as well. With the need for HPC/simulation workflows feeding synthetic data to models and becoming agents for models, many sites now have the same plethora of workloads that clouds and AI factories have. In addition, many smaller sites are doing more on-premise work in AI, partially due to GPU availability but also due to data movement costs, giving rise to far more sites needing more scalable storage than ever.

With this recent increase in workload breadth and number of sites, especially commercial sites, that need scaled storage, there has been a rise in the number of scalable storage solutions. Most of the new solutions are proprietary, with one growing category in particular, the parallel network file system (PNFS). It has fully open standards, and it is mostly open source.[1] There are at least five commercial offerings in the market. Most of these solutions are trying to cover the entirety of the wide breadth of workloads to varying degrees of success.

Bringing data closer improves both cost and performance.

MEDIA

Digital storage media technologies continue evolving to meet the capacity and performance needs of AI as well as traditional workflows. Nonvolatile memory devices are beginning to augment traditional volatile memory, such as dynamic random access memory (DRAM), to enable lower power consumption and increased capacity, particularly for embedded computing applications.

Magnetic tape and hard disk drives (HDDs), both using magnetic recording, continue to advance, with the linear tape-open (LTO) consortium introducing 30-terabyte (TB) native LTO-10 tapes, and with IBM's enterprise tapes available with up to 50TB native. The LTO roadmap shows close to 700TB capacities expected later in the 2030s, and with compression, this represents over 1 petabyte (PB) per tape cartridge. Magnetic tape and optical disks are the least expensive current media, although they are often the highest latency storage, leading to predominantly archival use cases.

HDDs are currently available with up to 36TB using heat-assisted magnetic recording (HAMR), and shingled magnetic recording, SMR. 40+TB HDDs will be available by 2026, with 60+TB drives in the next couple of years. 100+TB HDDs are expected by the 2030s.

NAND flash, the technology inside solid-state drives (SSDs), continues to advance in capacity and performance. Because of the diminishing cost advantage with 3D layers, that growth has slowed, but other ways to increase capacity have appeared, such as denser cells in the layers and more bits per cell. The latter is behind the shift to quad-level cells (QLCs), from tri-level cells, with dense packaging providing up to 256TB and projections for 1PB SSD capacities using the enterprise and data center standard E3 form factor (EDSFF).

Enterprise SSDs are still about six times more expensive than HDDs on raw storage capacity and are expected to remain so with the expansion of HAMR HDDs. Although SSDs are now the primary storage in data centers supporting DRAM in AI and other applications, colder storage still largely relies on HDDs and may remain so for many years. This leads to a hierarchy

of digital storage, optimizing performance and costs, where magnetic tape is still used for even colder data and archiving data. Figure 1 shows Coughlin Associates' projections for annual shipping capacity for SSDs, HDDs, and magnetic tape historically and out to 2030.

There are also recent optical disk startups, mostly going after the archiving and digital preservation markets, and some DNA storage startups going after the same markets.

In addition, nonvolatile memory technologies, such as magnetic RAM (MRAM), resistive RAM (RRAM), ferroelectric RAM, and phase change memory, are maturing and could replace current nonvolatile and volatile memories.

For embedded devices, NOR flash, widely used for code storage, can't shrink below 28 nanometers, and MRAM and RRAM are offered by the major semiconductor foundries as a replacement for NOR flash for smaller feature chips. MRAM has longer endurance and much more symmetric read and write speeds than NOR flash.

MRAM and RRAM are also starting to augment SRAM for embedded memory applications, when more memory is needed in a given die area and where power consumption is important, such as battery-based applications. Increasing use of these memories in embedded applications will reduce the costs and improve the yields of the nonvolatile memories, and could lead to replacing other volatile memories, such as DRAM, in the future.

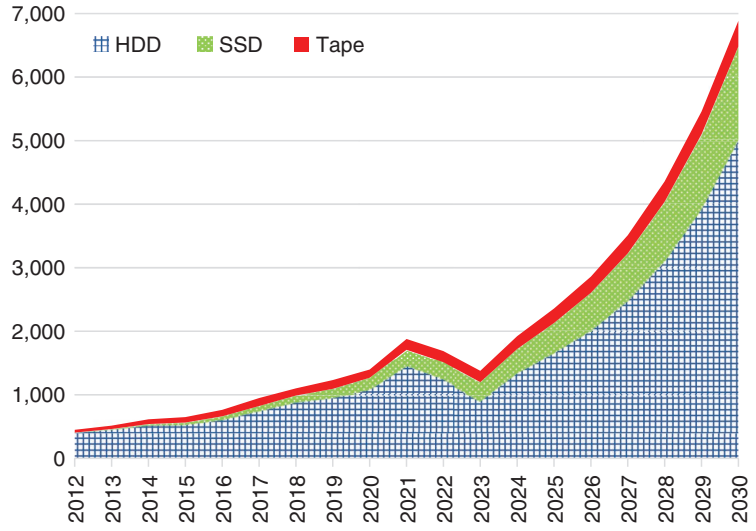

FIGURE 1. Historic and projected capacity of shipped HDD, SSD, and tapes in exabytes. Data obtained from Coughlin Associates. More information at: https://tomcoughlin.com/product/Digital-Storage-Technology-Newsletter/.

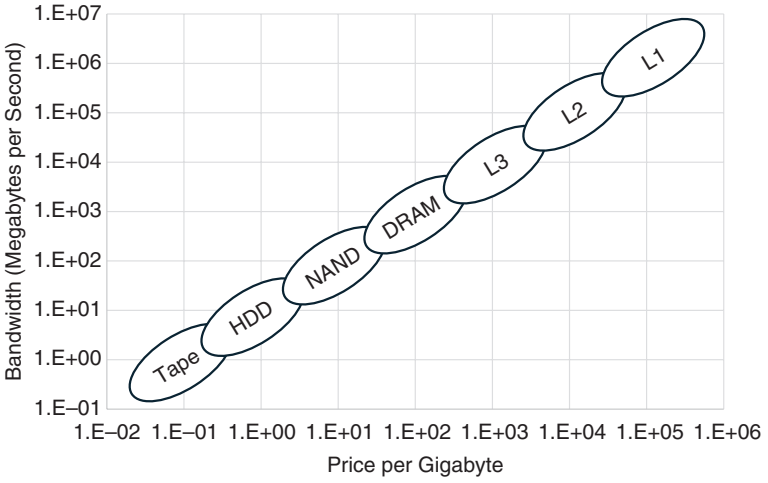

FIGURE 2. Media bandwidth as a function of the price per Gigabyte. Derived from the graph provided by Jim Handy. It provides a visual view of the tradeoffs between storage capacity costs and performance, here given as bandwidth.

STORAGE ECONOMICS

From 56 kilobytes on paper tape at 10 bytes/sec bandwidth in the 1950s, to today's commercial 256TB nonvolatile memory express SSDs that achieve 10 gigabytes (GB)/sec, the industry has increased storage capacity and bandwidth by over nine orders of magnitude in 75 years (Figure 2). The storage industry has ambitions to increase by another order of magnitude in the next three years. Random IOPS have only enjoyed an improvement of six orders of magnitude in that same time frame.

There has always existed a tradeoff between density and random IOPS, and that continues today. Another important trade space in data storage devices is bandwidth versus capacity. Data centers often employ several digital storage and memory technologies in a hierarchy to store the most frequently accessed data on fast but expensive memory or storage and less frequently accessed data on slower but less expensive memory and storage. This allows the optimization of performance and cost for data center operations.

Power consumption is rapidly becoming an important factor in storage and memory, and this could lead to changes in the distribution of the storage hierarchy. In addition, SSDs are so much more reliable than HDDs, so at large capacities, the economics of the HDD-based tiers is beginning to erode.

There is a movement to hybrid clouds for cost control and to use the

best applications available for data analysis. In AI, local storage will likely increase in demand to contain domain-specific or private data for retrieval-augmented generation (RAG)-based inference .

The emerging AI factory market, combined with smaller commercial sites moving AI processing to on-premise sites, creates a much bigger market for scalable storage solutions. There is finally profit to be made in selling scalable solutions, including software. This was largely not true five years ago, with dominant solutions like Lustre at the high end of scalable storage solutions, and cloud offerings with their own solutions. This has led to proprietary solutions that try to cover a broad set of workloads at scale. for example, object stores, for which there is no real standard, and scalable file systems, with true parallelism at some scale.

Proprietary solutions vary widely in their value, much like cloud solutions. Until the PNFS solutions started to emerge, open standards-based/open source solutions were losing ground, but at least five commercial PNFS-based solutions are entering the market. The health of the scalable storage industry will depend on a mix of proprietary/cloud and more open and standards-based solutions that allow more flexibility.

Explosion of metadata operations, fueled by the AI workflows, emerges as a driver for the new class of storage requirements—a need for high IOPS on a small subset of data with very fine, almost memory-like granularity. This is especially evident with compute-local storage, used as an extension of memory, which emerges as part of the new AI factory architecture. This, in turn, drives the storage industry to develop new classes of storage, such as storage class memory devices of various types, and new storage protocols, leveraging memory semantics, such as Compute Express Link, as opposed to the traditional block device protocols.

Moving data around is one of the biggest factors contributing to an increase in energy consumption and latency. AI workflows can require the movement of large amounts of data—as data may not necessarily be processed where it is created and stored. To mitigate the cost of data movement, domain-specific processing close to memory and storage, offloading various AI functions, is increasing in popularity. Enabling efficient manipulation of multiple variable-length data pieces requires a departure from the standard fixed block access data semantics, similar to IBM mainframe count-key-data storage systems. Additionally, most data lakes use columnar data layouts and work in row groups, which lend themselves well to variable-length-supporting storage devices.

STORAGE ARCHITECTURE

Trends in data volume and SSD technology are driving storage system changes against industry inertia, data longevity, and slow software maturation cycles. File systems typically have multidecade lifespans (see Figure 3).

Our dependence on continuous, easy, fast access to every stored datum we ever generated has never been higher. Data growth exceeds declining capacity costs, dragging IT infrastructure spending upward. This presents scaling problems in any storage system challenged to maintain, reference, and curate data, compounding IT maintenance and administration stresses. Ease-of-use and continuous operation become essential.

Cold and disaster-recovery data are rarely accessed, while hot data are ever-hotter, driven by new AI training, inference, virtualization, and simulation workflows. Scaled apps from the HPC realm are now common, data breach threats, multitenancy, and sporadic high IO frequency neighbors demand dedicated resources. All these trends challenge on-premise, private cloud, and public cloud infrastructures alike to hide complexity from user apps.

This requirement mix drives the evolution of disk and tape software, while solid-state media demands a stepwise hardware and software revolution. Disk and tape work better with multimegabyte sequential streams that hide giant latencies. Flash devices are opposite, with 10–30 times the throughput with much smaller, highly-concurrent input–output (I/O), delivering microsecond-level data access latency. Flash requires new data path software to handle this, and it forces reanalysis of enabling hardware. Just three enterprise-grade flash SSDs can already saturate a 400 GB network. QLC SSDs now reach 120TB in one SSD, about four times that of the largest available hard disks, and E3.L devices double that.[2] Systems holding dozens of devices are too expensive, with giant blast radii. High availability is even more challenging with each doubling of bus and network bandwidth. Flash now uses two times HDD power,

> Nonvolatile memory devices are beginning to augment traditional volatile memory.

FIGURE 3. Life span of well-known file systems. (Source: Wikipedia pages for each file system.)

and faster writes will demand even more. And processing of flash I/O is already daunting. Finally, while liquid cooling is trending for GPUs, it is now also spilling to flash.[3,4]

How to address all these disruptive flash requirements? Such solutions will allow compute nodes to each be equipped with relatively small capacity and to work on small fractions of the overall data using a shared-nothing architecture, eventually staging/flushing the data to a common storage tier. This can work when the use cases and capacity requirements are known at purchase, such as the lifetime training of a specific large model. But data movement, management of each node's data, secure multitenancy, varying workloads, reliability, and other evolving requirements leave storage stranded, unused, wasted, or insufficient if not overprovisioned. It is vulnerable to attack, often without resiliency, and results in isolated data. Network-shared approaches overcome drawbacks of local flash while adding complexity, scale, and potentially performance issues. These challenges demand newly minted software over a proper hardware topology; several have pursued including single- and multihop I/O paths, each with single- and multilayer networks. Other evolving requirements leave storage stranded, unused, wasted, or insufficient if not overprovisioned. It is vulnerable to attack, often without resiliency, and results in isolated data. Network-shared approaches overcome drawbacks of local flash while adding complexity, scale, and potentially performance issues. These challenges demand newly minted software over a proper hardware topology; several have pursued including single- and multihop I/O paths, each with single- and multilayer networks.

The most efficient architectures employ the simplest single-layer networks with single-hop I/O paths between apps and devices. Many small nonredundant nodes internally balanced with single-digit flash devices and high-speed network interfaces will be typical, whether in a bespoke blade[5] or a commodity server.[6] Nodes will provide networked storage-protocol-serving targets, with data shared across them redundantly to cover device and node losses.

To provide trusted environments to untrusted multitenant workloads, consumer-side endpoints will be provisioned within network targets, and common data interfaces [for example, block, file, object, key value (KV), etc.] will be served securely from there.[7,8] Eventually, both client- and server-side data services may be delegated to network endpoints to provide ubiquitous, secure, accelerated, efficient, highly-scalable flash storage and data services across datacenters, regions, or the globe.

> Power consumption is rapidly becoming an important factor in storage and memory.

The result will be a set of two tiers, each split into two layers. The first will be optional compute-local flash as a caching layer in cases where it fits, coupled directly across a high-speed network to a highly scalable networked shared layer built from multiple flash types on a sea of simple nonredundant blades or commodity servers. The second tier remains much as today, a bulk data store comprised first of spinning disk, and secondarily of robotic tape. The tiers within and among on-premise and/or cloud data centers will be enabled by purpose-built software, data paths suitable for each media type and use case, lashed to one another through advanced schedulers that drive parallel data movement among them.

DATA AND AI

AI has already significantly reshaped data organization and storage requirements. AI workloads demand high-throughput, low-latency access to massive data sets, often in the peta-to-exabyte range. Data organization must prioritize efficient retrieval and processing, leading to the adoption of tiered storage systems that balance cost, speed, and scalability. AI also requires data to be organized for rapid iteration, with metadata-rich systems to track data lineage and preprocessing steps (see also "Data for AI training and inference at scale" section). The need for distributed AI frameworks further drives a combination of high-performance parallel file systems and S3-like object storage.

Unlike enterprise applications, which often focus on structured data in relational databases, AI demands unstructured/semi-structured data (for example, images, video, text, or sensor data) with random-access patterns. Enterprise storage systems also emphasize data consistency and integrity, which are less critical for AI's fault-tolerant processes based on iterative improvements (in training) or requiring a search-like flow (in inference).

Compared to HPC, which prioritizes high bandwidth with predictable access patterns and resilience (checkpoint/restart) support, AI workloads require frequent, irregular data access across very large data sets, especially during model training.

Traditional big-data analytics, such as those using Hadoop or Spark, share similarities with AI but are optimized for batch processing and lack real-time, iterative processing support for model training, and were built for different frameworks and physical media (spinning hard disks).

AI's emphasis on rapid data ingestion and preprocessing also contrasts with cloud workloads, which prioritize cost efficiency and relaxed consistency models over raw performance. Cloud storage often struggles with the low-latency needs of AI, requiring

hybrid solutions that combine cloud scalability with the performance that comes from tightly coupling compute and data.

AI workloads thus require a hybrid approach, blending HPC's performance, cloud's scalability, and big data's flexibility, necessitating specialized storage systems that combine high-performance all-flash arrays and distributed data organization. These demands have already spurred several innovations, such as AI-optimized data lakes and specialized hardware support, like GPU-direct storage.

Data for AI training and inference at scale

The data preparation stage for AI training resembles big-data processing, but with unique nuances. Data must be cleaned, normalized, and transformed into machine learning (ML)-friendly formats, often requiring extensive preprocessing to reach high-quality labeled data sets.

Tools like TensorFlow Data Validation, Apache Airflow, and DVC (dvc.org) are used in data validation, workflow orchestration, and versioning. Frameworks like Apache Spark or Dask handle large-scale data processing, while data lakes (for example, Delta Lake) provide centralized repositories for AI-ready data. AI data prep emphasizes iterative refinements, data augmentation, and integration with compute frameworks, requiring systems that support high IOPS and low-latency access.

Inference at scale imposes distinct storage requirements and a balance of speed, scalability, and cost-efficiency, tailored to the deployment environment. Inference prioritizes low-latency, high-concurrency access to smaller, often real-time data streams and caching tiers. Storage systems must support rapid retrieval of model weights and input data, often using in-memory key-value stores for sub-millisecond latency. Edge inference, common in Internet of Things (IoT) and autonomous systems, requires lightweight storage with minimal footprints, embedded databases, or local caches. Cloud inference leverages object storage for model artifacts and input data, optimized for cost and latency, and efficiently supporting containerized environments.

What comes next: multimodal, live, and reasoning AI

Multimodal AI, processing text, time-series, images, audio, and video (and in the future, possibly DNA, proteins, polymers, weather, etc.), will significantly amplify data and storage demands by 2030. Storage architectures will evolve toward AI-optimized data lakes with advanced metadata management to support cross-modal data retrieval and preprocessing. For inference, multimodal AI will necessitate edge and cloud storage solutions that support real-time data fusion, requiring hybrid systems with enhanced caching mechanisms and data compression to manage the increased data volume and variety efficiently.

The rise of "live AI" assistants, constantly active for the end user, will transform storage requirements by prioritizing real-time, low-latency data access. When billions of personal AI assistants generate continuous streams from user interactions, sensors, and contextual inputs, storage systems stress high concurrency and scalability. Edge storage becomes critical, with lightweight, embedded databases or in-memory caches deployed on AI endpoints or IoT hubs. Cloud storage adds long-term data retention and model updates, using scalable object stores. Data privacy concerns will also drive encrypted and decentralized storage solutions to balance user-specific data while enabling seamless access for live AI applications.

Reasoning workloads and the pursuit of artificial general intelligence (AGI) will introduce unprecedented challenges due to their need for iterative inference on contextually rich data sets, and the combination with other tools like Internet search. Reasoning systems require iterative, dynamic data access for complex tasks, hypothesis generation, and multistep problem-solving. This will drive demand for advanced indexing and search capabilities to handle knowledge graphs and ontologies. Storage tiers will evolve to prioritize hot and fresh data for active reasoning while archiving less critical data to cost-efficient cold storage.

> Moving data around is one of the biggest factors contributing to an increase in energy consumption and latency.

Nonfunctional requirements

New AI workloads, including multimodal AI, live AI assistants, and reasoning tasks trending toward AGI, also introduce nonfunctional requirements, some of which we outline next.

AI systems require data management systems to evaluate and address *bias mitigation and fairness,* and the corresponding storage systems must support mechanisms to detect and flag biased data sets. This involves integrating metadata frameworks that tag data sets with bias indicators, enabling AI pipelines to filter or reweight data during preprocessing. Future storage solutions will likely incorporate built-in AI-driven analytics to monitor and audit data sets.

Data tagging, lineage, and provenance are critical nonfunctional requirements for AI systems, particularly for reproducibility and regulatory compliance. Storage systems must support

rich metadata schemas to tag data with attributes that enable traceability. Object stores could even integrate blockchain-inspired ledgers or DVC-like tools to ensure tamper-proof lineage, increasing storage overhead but enhancing trust and compliance.

AI workloads also require *data formats with varying precision and accuracy needs*. High-precision formats are critical for training, while lower-precision formats suffice for inference to optimize performance and memory footprint. Storage systems will dynamically manage these formats, supporting seamless conversion and compression without data loss, format-agnostic abstractions, and compression algorithms.

> The rise of "live AI" assistants, constantly active for the end user, will transform storage requirements by prioritizing real-time, low-latency data access.

DATA FOR AGENTIC AI AND AGENTIC AI FOR DATA

The emergence of large-scale AI foundation models and the progression of agentic AI are changing the ways in which data are created, consumed, and managed. A significant proportion of the world's digitally recorded data and knowledge expressed in both human and programming languages is now incorporated in models and accessed through them rather than through original sources. Agentic AI systems based on these foundation models are becoming remarkably good at connecting the dots, searching, synthesizing, and interpreting the meaning associated with different pieces of data, often replacing the need for human annotation. The intriguing power of generative AI models comes with inherent reliability limitations, such as the potential for the generation of plausible, imaginative content whose accuracy, safety, and ethical implications needs to be verified by additional means when embedded as part of a larger system or solution. This evolution impacts almost every stage of data processing and management, from how data are perceived, transformed, organized, validated, connected, transferred, saved, and protected.

Agentic AI systems augment parametric knowledge (encoded in the generative AI-based foundation models trained on enormous Internet-scale data) with external nonparametric knowledge and tools (where tools could include both traditional and AI software). This external knowledge includes short-term and long-term agentic memory representations, often derived from multiple iterations of search/retrieval, processing, and generation, resembling human memory-like constructs, such as semantic, procedural, and episodic memory.

New forms of data hierarchies are emerging to support this paradigm, with different indexing structures (for example, vector stores for embeddings, KV caches, graphs), caching/retention policies, and associated optimizations applicable for representations at different levels. Challenges in managing and sharing the growing need for longer LLM context memory (currently in the order of a million tokens) for stateful agents have spurred many variations of long-term agentic memory architectures. This area is ripe for standardization opportunities, especially with the rising horizon (length) of tasks achievable by AI systems and the need for sharing this context across multiple agents.

Data management systems are beginning to leverage agentic AI capabilities too, intelligently applying it at various stages of data curation and preprocessing to feed this hierarchy, improving the time to value (time to inference) of incoming data. This trend leads to new flavors of scaling challenges when integrating a variety of multimodal data sources and high velocity data or sharing representations and metadata structures across agentic systems.

The notion of data veracity and lineage is also evolving as AI-generated data across various modalities can be hard to distinguish from "real data" and attributed to specific sources.[9] Data provenance, end-to-end observability, and lineage metadata tracking in AI are necessary for ensuring reproducibility, quality, and trustworthiness (including transparency, explainability, and bias mitigation), which are prerequisites for regulatory compliance and meeting responsible AI standards. However, the current state of practice reflects a critical need to address the pervasive gaps in traceability[10,11] both when it comes to large-scale training data collection for foundation models (considered a "crisis of data set provenance" as per the Data Provenance Initiative https://www.dataprovenance.org/); and during the deployment of complex distributed agentic AI workflows at scale (given the fragmentation of tools resulting in stranded or missing lineage and further exacerbated by the sheer scale and iterative nature of these workflows).

Embedding lineage tracking and observability capabilities directly within standard ecosystem components, such as model context protocol layers, along with a decentralized (git-like) lineage tracking framework, such as common metadata framework,[12] can enable the persistence of the corresponding metadata structures alongside data. AI agents for automatic lineage annotation, verification, and governance are also emerging and could potentially lower the barrier to adoption, but must be coupled with new approaches for tagging, protecting, and certifying data authenticity for digital data as the parallel universe of data generated by multimodal and physical AI systems becomes pervasive.

SECURITY AND PRIVACY

The security and privacy of our data urge an increasing share of our attention. With every website you visit, every show you watch, every photo you post, every coffee shop you leave your phone number with, and every credit card transaction, you leave a digital breadcrumb trail with identifying information. This information can be collected and collated to compose a rich profile of your explorations, with a similarly rich potential for exploitation. The spectrum spans from the seemingly innocuous targeted advertising, through deepfakes, sprawling cybercrime, and even deadly ramifications, as evidenced by recent wars. It is no wonder then that people care increasingly more about the privacy and security of their data, so it behooves us to ask how these will evolve with the other trends discussed in this article.

Three exponential trends combine to exacerbate the security and privacy challenge: the growth of data production, the growth of data procurement, and the growth of data processing.

On the production side, most of the added bytes are expected to come from video and social media, Web browsing, and other sources of personal information.[13] More and more companies are interested in collecting and processing this data, growing at an estimated rate of 10%–30% per year.[14,15] Social media monitoring is one of the most popular data collection applications, and the raw data are increasingly augmented with labels created automatically or through outsourcing. The increasing availability of storage and processing power to crunch this voluminous data into actionable insights is driving the parallel growth we observe in datacenter capacity.

If the probability of a catastrophic data breach is a function of opportunity, means, and motive, then as data availability, processing power, and lucrativeness grow exponentially, this probability approaches certainty. So, what are we to do about this looming disaster?

One approach is to resign to the new reality and do nothing. Many people already assume that their digital lives have become nonprivate and impossible to protect. Perhaps they're hoping that their ordinary data are unattractive within a sea of bigger fish; perhaps they remain purposefully ignorant of the risks; or perhaps they practice some form of digital abstinence with the private data they care about most. Either way, current data privacy practices are inadequate for much of the population, so a path of inaction is ill-advised.[16] The survey by Pew Research Center also showed strong support for a second approach to protecting personal information: regulatory action.[16]

> Those who own data will be in control.

There have been some significant efforts to regulate the collection and use and private information, such as the European Union's General Data Protection Regulation. These efforts are still a far cry from solving security issues: First, they are not universal, so data that is prohibited from being collected in one country can still be collected in another. Second, they sometimes conflict with the regulatory body's own interests. Since governments are some of the largest collectors and users of personal data, they could contribute more to the problem than to the solution. And third, regulation tends to respond slowly to technological changes, trying to catch the horse after it has left the barn. Sometimes, it is even at odds with technological innovations that do preserve privacy, such as cryptocurrencies.

Which brings us to the third actionable approach: developing technology to harden security and privacy. Although it may be naïve to expect the tech industry to curb its own appetite for personal data, progress in the field is nevertheless continuing at a rapid pace. Innovations in encryption, differential privacy, secure multiparty computation, zero-knowledge proofs, and anonymization are trying to fulfill the promise of abundant data without the price of privacy loss.[17] In this regard, we as technologists have a responsibility to educate ourselves on contemporary topics in privacy and security and keep them foremost in our minds, even if not directly related to our work. It is both our burden and our opportunity to foster technological innovation that brings about benefits, not disasters, to humankind.

Data continues to be the most critical asset in the world. Everything depends on data. AI has even further amplified its value. Its traditional characteristics, such as scale, retention and endurance, and protection, are now extended with privacy and sovereignty, bias, and lineage. However, data does not come without cost. Its continuous scaling in terms of size, source, and use may elevate data movement to the costliest energy consumer.

Because of its relevance, data will remain a critical asset equally for individuals, companies, and states. Those who own data will be in control. Despite its exponential growth, we predict that media, storage architectures, and governance of data will continue to evolve and support the data of the future.

REFERENCES

1. "Creating high-performance parallel file systems with NFS," Hammerspace, Redwood City, CA, USA, Tech. Brief, Feb. 2024. [Online]. Available: https://hammerspace.com/creating-high-performance-parallel-file-systems-with-nfs/
2. "KIOXIA announces industry's first 245.76 TB NVMe SSD built for the

demands of generative AI environments," *KIOXIA*, Jul. 21, 2025. [Online]. Available: https://americas.kioxia.com/en-us/business/news/2025/ssd-20250721-1.html
3. B. Behlendorf and O. Faaland, "Rabbit storage for El Capitan: Fast I/O through big, pointy teeth," Lawrence Livermore Nat. Lab., Livermore, CA, USA, May 2023. [Online]. Available: https://www.opensfs.org/wp-content/uploads/Fast-IO-El-Capitan-Rabbits.revised.pdf
4. "Media kit: Data center - D7-PS1010 E1.S," *Solidigm Newsroom*, Mar. 18, 2025. [Online]. Available: https://news.solidigm.com/en-WW/media_kits/235611/
5. "Unstructured data storage." Pure Storage FlashBlade//SS. Accessed: Aug. 11, 2025. [Online]. Available: https://www.purestorage.com/products/unstructured-data-storage/flashblade-s.html
6. "HPE ProLiant compute DL325 Gen12 QuickSpecs." HPE. Accessed: Aug. 11, 2025. [Online]. Available: https://www.hpe.com/psnow/doc/a50009232enw.pdf
7. "AWS nitro system." AWS. Accessed: Aug. 11, 2025. [Online]. Available: https://aws.amazon.com/ec2/nitro/
8. "BlueField networking platform." NVIDIA. Accessed: Aug. 11, 2025. [Online]. Available: https://www.nvidia.com/en-us/networking/products/data-processing-unit/
9. K. Jaźwińska and A. Chandrasekar, "AI search has a citation problem," *Columbia Journalism Rev.*, Mar. 6, 2025. [Online]. Available: https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php
10. S. Longpre et al., "Bridging the data provenance gap across text, speech and video," 2025, *arXiv:2412.17847*.
11. S. Longpre et al., "Position: Data authenticity, consent, & provenance for AI are all broken: what will it take to fix them?" in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 32,711–32,725.
12. A. J. Koomthanam, A. Tripathy, S. Serebryakov, G. Nayak, M. Foltin, and S. Bhattacharya, "Common metadata framework: Integrated framework for trustworthy artificial intelligence pipelines," *IEEE Internet Comput.*, vol. 28, no. 3, pp. 37–44, May/Jun. 2024, doi: 10.1109/MIC.2024.3377170.
13. F. Duarte, "Amount of data created daily," *Exploding Topics*, Apr. 24, 2025. [Online]. Available: https://explodingtopics.com/blog/data-generated-per-day
14. "Data collection and labeling market size, share & trends analysis report by data type (text, image/video, audio), by vertical (automotive, government, healthcare, BFSI, retail & e-commerce), by region, and segment forecasts, 2025 - 2030," Grand View Res., San Francisco, CA, USA, 2024. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market
15. T. Navio, "Data analytics market analysis, size, and forecast 2025-2029: North America (US and Canada), Europe (France, Germany, and UK), Middle East and Africa (UAE), APAC (China, India, Japan, and South Korea), South America (Brazil), and rest of world (ROW)," Technavio, London, U.K., Jan. 2025. [Online]. Available: https://www.technavio.com/report/data-analytics-market-industry-analysis
16. M. Faverio, "Key findings about Americans and data privacy," *Pew Research Center*, Oct. 18, 2023. [Online]. Available: https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy
17. Q. Razi, R. Piyush, A. Chakrabarti, A. Singh, V. Hassija, and G. S. S. Chalapathi, "Enhancing data privacy: A comprehensive survey of privacy-enabling technologies," *IEEE Access*, vol. 13, pp. 40,354–40,385, 2025, doi: 10.1109/ACCESS.2025.3546618. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10908383/

SUPARNA BHATTACHARYA is a Fellow and vice president at HPE Labs, Bangalore, Karnataka 560048, India. Contact her at suparna.bhattacharya@hpe.com.

THOMAS COUGHLIN is president of Coughlin Associates, San Jose, CA 95124 USA. Contact him at tom@tomcoughlin.com.

LANCE EVANS is chief architect of HPC storage at Hewlett Packard Enterprise, Ft Collins, CO 80528 USA. Contact him at lance.evans@hpe.com.

PAOLO FARABOSCHI is a Fellow, vice president, and director of the AI Research Lab at HPE Labs, Milpitas, CA 95035 USA. Contact him at paolo.faraboschi@hpe.com.

EITAN FRACHTENBERG is a master technologist at HPE Labs, Milpitas, CA 95035 USA. Contact him at eitan.frachtenberg@hpe.com.

GARY GRIDER is a high-performance computing division leader at Los Alamos National Lab, Los Alamos, NM 87545 USA. Contact him at ggrider@lanl.gov.

DEJAN MILOJICIC is a Fellow and vice president at HPE Labs, Milpitas, CA 95035 USA. Contact him at dejan.milojicic@hpe.com.

SREENIVAS RANGAN SUKUMAR is a senior distinguished technologist at Hewlett Packard Enterprise, Bellevue, WA 98005 USA. Contact him at sreenivas.sukumar@hpe.com.

ALEX VEPRINSKY is a Fellow and vice president at Hewlett Packard Enterprise Storage, San Jose, CA 95002 USA. Contact him at alex.veprinsky@hpe.com.