IBM Storage solutions for advanced driver assistance systems and autonomous driving

Build your end-to-end data pipeline with IBM Storage delivering extreme scalability with high performance.

Highlights

Enables design and deployment of end-to-end data pipelines:

- IBM Spectrum Scale file storage for scalable high performance

- IBM Cloud Object Storage for global-scale efficiency and durability

- IBM Spectrum Discover for modern metadata management

Eases management by integrating capacity and performance tiers

Minimizes data silos and copies for improved data governance

Offers software-defined storage as a service, software-only or integratedd

Matches business needs with perpetual, monthly and utility pricing

The automobile is quickly morphing from an isolated, largely mechanical piece of equipment to one of the most technically sophisticated and connected platforms on the planet. From entertainment and navigation to driver assistance and crash avoidance, today's car is vastly different from those of a few years ago. The huge opportunity to share the future of that connected car, especially around autonomous driving, is drawing the interests of both technology and auto manufacturing companies.

These connected-cars initiatives are categorized into two primary categories:

- Advanced driver assistance systems (ADAS), which includes services such as navigation, remote diagnostics and collision warning
- Autonomous driving (AD), which includes services such as automatic parking and autopilot

The one thing these initiatives all have in common is data—miles and miles of data. Data sources include sensor data, weather data, satellite data, behavioral and other personal data, diagnostic data, and more. Each connected car generates data—from a few megabytes to sometimes gigabytes per day. This data volume can reach terabytes per car per day when that car is a test vehicle used to train AD models. These connected-cars initiatives generate storage demand for approximately 200 exabytes of data daily.

Build an end-to-end data pipeline with IBM Storage

Storage can be categorized based on data pipeline stages: edge storage; transient storage; global ingest tier; storage for fast ingest/real-time analytics; storage for extract, transform, load (ETL) or data processing; storage for Hadoop or Spark data lakes; storage for machine learning/deep learning; and archive storage. As data flows through the data pipeline, storage requirements for each stage can vary, with each potentially having unique demands for scalability, performance and location of the storage that supports it.

The following figure illustrates how IBM Storage offerings enable you to build an end-to-end data pipeline serving the varying storage requirements of each stage in the data pipeline.
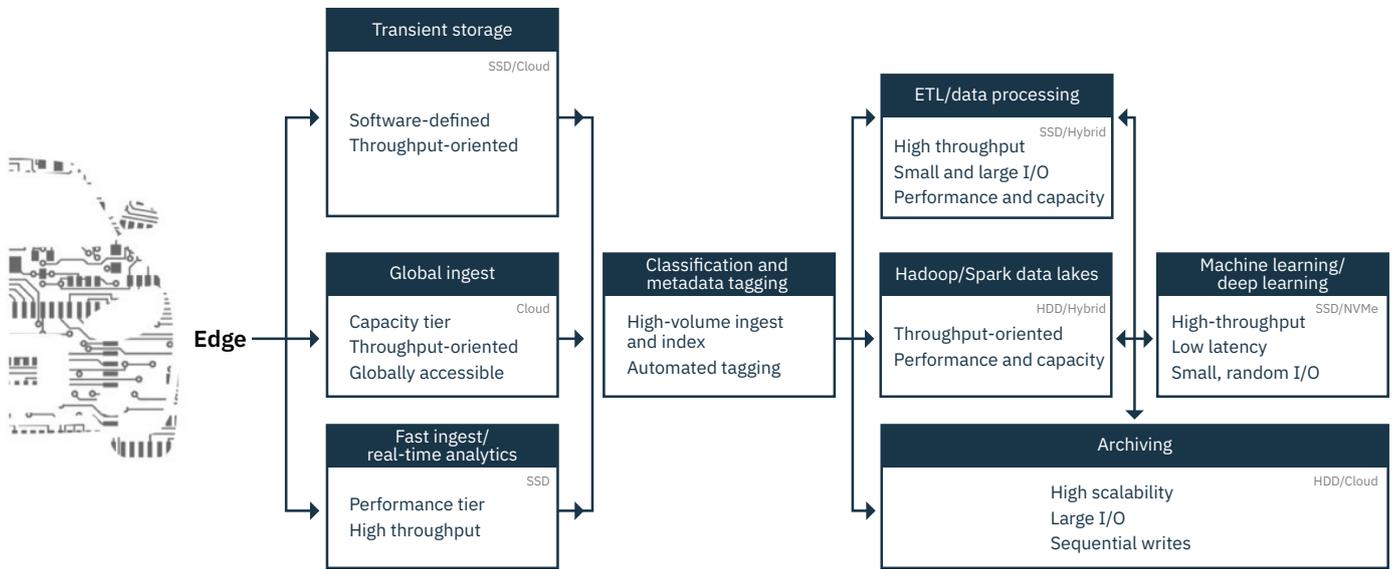

Figure 1: Data pipeline for ADAS and AD.

IBM Spectrum Scale file storage

Developing and testing AD systems requires the storage and analysis of more data now than ever before. Clients who can deliver insights faster while managing rapid infrastructure growth will be poised to be industry leaders. To deliver these insights, the underlying storage must support both new big data and traditional applications with security, reliability and high performance. To handle massive unstructured data growth, the solution must scale seamlessly while matching data value to the capabilities and costs of different storage tiers and types. IBM Spectrum Scale, part of the IBM Spectrum Storage family, meets these challenges and more.

IBM Spectrum Scale is a high-performance file storage solution for managing data with the distinctive ability to perform analytics in place with comprehensive support for data access protocols including POSIX, NFS, SMB, HDFS and S3/Object. It can provide a single namespace for all this data, offering a single point of management with an intuitive graphical user interface. IBM Spectrum Scale offers system scalability, high availability and reliability with no single point of failure in large storage infrastructure.

As Figure 2 illustrates, the IBM Spectrum Scale file system can be used for file (POSIX, NFS, SMB), object (S3, SWIFT) and/or Hadoop Distributed File System (HDFS) for in-place analytics workloads.

Proven high-performance and extreme scalability

IBM Spectrum Scale is a well-established file storage system with large deployments in the field storing billions of files and petabytes of data.

It is also known for its high-performance characteristics. 50 percent of file storage systems delivering top SPEC SFS benchmark results run on IBM Spectrum Scale software. One of America's newest and fastest supercomputers, SUMMIT, also runs on IBM Spectrum Scale.
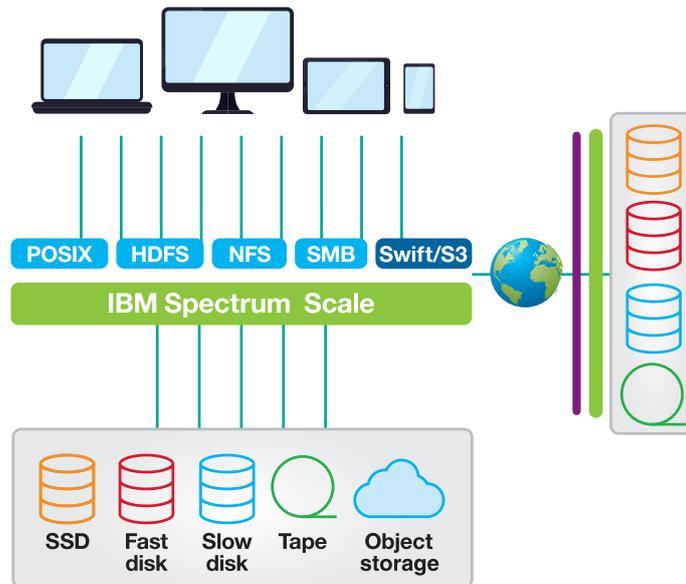

Figure 2: IBM Spectrum Scale provides a common namespace with intelligent storage tiering to build scalable data lakes.

Advanced data management

ADAS development requires contractual and regulatory commitments for test-data retention. Keeping tens, and soon hundreds of petabytes in high-performance data storage is a requirement during the simulation and validation phases of development. This data must be retained for multiple decades, and service contracts commonly mandate restoration and re-simulation times measured in days.

IBM Spectrum Scale can help improve performance, lower costs, add resiliency and simplify collaboration with algorithmic and policy-driven data movement including copying and caching. IBM Spectrum Scale catalogs data across multiple storage pools, including the cloud. It tracks usage profiles, storage latency and a broad range of standard and custom metadata from which data movement policies can be constructed.

Armed with both awareness of data usage and its underlying storage, IBM Spectrum Scale curates data across multiple storage tiers, including tape and cloud. The powerful, data-aware intelligence engine can create optimized, tiered storage pools by grouping devices—flash, solid-state drives (SSDs), disks or tape—based on performance, location or cost. Migration polices transparently move data from one storage pool to another without changing the file's location in the directory structure. Automated analysis of data usage patterns can help raise data to higher performance tiers as needed. The information lifecycle management tools built into IBM Spectrum Scale help simplify data management by providing additional control over data placement. These tools include storage pooling and a high-performance, rule-based policy engine.

Remove data-related bottlenecks for HiL/SiL/MiL testing

Slow storage negatively impacts applications, delays tasks and wastes expensive infrastructure. IBM Spectrum Scale can speed time to results and help maximize storage utilization by providing parallel data access, a requirement for hardware-in-the-loop (HiL), software-in-the-loop (SiL) and machine-in-the-loop (MiL) testing. Further, shared disks and storage-rich servers improve scalability for high-performance workloads. IBM Spectrum Scale is based on a parallel file system with intelligence in the client, spreading the load across all storage cluster nodes, including individual files. In traditional scale-out network-attached storage, a file can only be accessed through one node at a time by each client. This parallel file system architecture allows IBM Spectrum Scale to seamlessly handle tens of thousands of clients, billions of files and yottabytes of data.

Empower global collaboration

IBM Spectrum Scale enables low-latency read and write access to data from anywhere in the world using Active File Management (AFM), distributed routing and advanced caching technology. AFM expands the IBM Spectrum Scale global namespace across geographical distances, providing fast read and write performance with automated namespace management. As data is written or modified at one location, all other locations get the same data with minimal delays. AFM leverages the inherent scalability of IBM Spectrum Scale, providing a high-performance, location-independent solution that masks network failures and hides wide-area latencies and outages. These game-changing capabilities accelerate project schedules and improve productivity for globally distributed teams.

Tiering across storage layers—simplify data management at scale

IBM Spectrum Scale includes integrated management tools and an intuitive graphical user interface to help manage data at scale. The file system can span multiple storage environments and data centers across the world to eliminate data silos and "filer sprawl." IBM Spectrum Scale can intelligently spread data across multiple heterogeneous storage devices—optimizing available storage utilization, reducing administration and delivering high performance where needed. The software includes multiple deployment and configuration options that accommodate current NFS filers, block storage and storage-rich servers in a global namespace with universal access.

Integration with Hadoop and Spark workloads

IBM Spectrum Scale supports Hortonworks Hadoop workloads and the HDFS without requiring any changes to applications. With the IBM Spectrum Scale Hadoop connector, multiple IBM Spectrum Scale clusters or other HDFS repositories can be federated into a single HDFS instance. IBM Spectrum Scale reduces the need to move data, simplifying the deployment and workflow of Hadoop, Apache Spark and related packages.

IBM Elastic Storage Server

IBM Elastic Storage Server is a pre-integrated system powered by IBM Spectrum Scale software. The system includes IBM Power servers, flash and hard drive storage, and IBM Spectrum Scale software. It also comes with an efficient erasure-coding implementation that provides superior data protection and reliability over traditional hardware RAID implementations.

IBM Cloud Object Storage

Data storage is a vital component of ADAS and AD development. Automobile manufacturers now require systems that are capable of handling the huge amount of data generated by "smart" cars. Onboard computers, and the myriad of sensors and other tech powering AD, are creating petabyte-scale volumes of unstructured data.

Data storage needs to be highly reliable and must not compromise data processing speed. In reality, data must be processed and communicated to the car or driver quickly, to help avoid potentially serious or even fatal traffic incidents.

Through the use of the dispersed-storage technologies of IBM Cloud Object Storage, ADAS and AD development infrastructure can quickly and securely store and efficiently process massive amounts of data. The possibilities from managing effective, dispersed storage with IBM Cloud Object Storage in autonomous cars are numerous—saving lives, lowering insurance premiums, easing congestion, reducing pollution and ultimately increasing efficiency.

What is dispersed storage?

IBM Cloud Object Storage uses an innovative approach for cost-effectively storing large volumes of unstructured data while helping ensure security, availability and reliability. IBM Cloud Object Storage uses Information Dispersal Algorithms (IDAs) and IBM SecureSlice to separate data into unrecognizable "slices" that are distributed through network connections to local storage nodes or across the world. This collection of distributed storage appliances creates the IBM Cloud Object Storage System. With IBM Cloud Object Storage dispersed-storage technology, transmission and data storage are inherently private and secure. No complete copy of the data resides in any single storage node, and only a subset of nodes must be available to fully retrieve the networked data.

What is information dispersal?

At the heart of IBM Cloud Object Storage is a technology called information dispersal, the practice of using erasure codes to create redundancy for transferring and storing data.

Your data is online and accessible

In the configuration below, up to six systems can be down—
including an entire site. Data remains online and available.
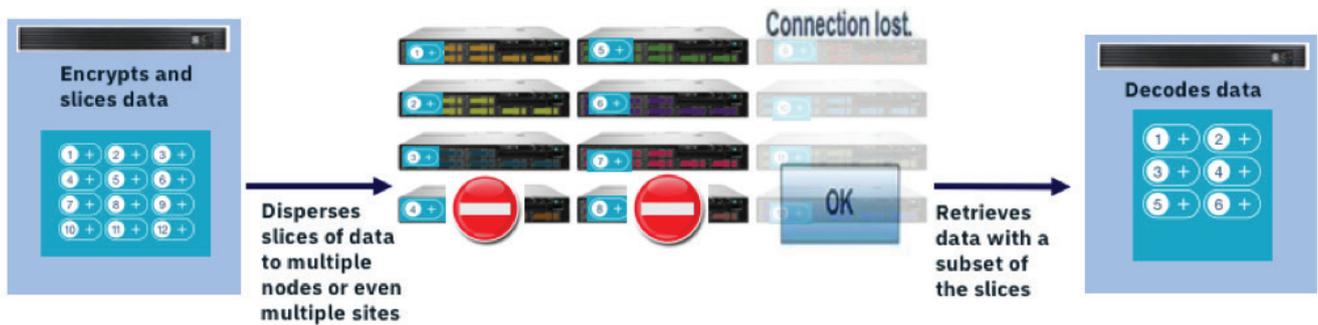

Figure 3: Data dispersal allows recovery in the event of system failure by reassembling slices of previously distributed data.

An erasure code is a Forward Error Correction (FEC) code that transforms a message of k symbols into a longer message with n symbols such that the original message can be recovered from a subset of the n symbols (k symbols).

An IDA can be made from FEC code. With an IDA, the coded data is additionally split into multiple segments, which can then be stored on different devices or media to attain a high degree of failure independence. For example, using FEC alone on files on a computer is less likely to help if that computer's hard drive fails, but with an IDA to disperse data across machines, those machines can tolerate multiple failures without losing the ability to reassemble that data.

Built-in single-site or multi-site failure tolerance of IBM Cloud Object Storage

With built-in dispersed-storage technology, only a subset of slices is required to retrieve data. This allows IBM Cloud Object Storage to tolerate appliance failures both within a single site and across multiple sites.

Benefits of IBM Cloud Object Storage for ADAS and AD development infrastructure

Traditional storage protection technologies such as RAID are simply inadequate when it comes to protecting digital information from data loss at petabyte-scale and beyond, as is required for ADAS and AD development. Traditional storage architectures are not designed to scale to petabyte range, as they are less reliable and secure.

IBM Cloud Object Storage is best used for "warm" data—data that must be active and ready for use, such as data in:

- Backup and archive that keeps data in an efficient and protected repository for easy, secure and fast access
- Large big data repositories for analysis with online scalable capacity
- Cloud native applications that are portable and developed for cloud scale and access from anywhere

ADAS and AD with IBM Cloud Object Storage
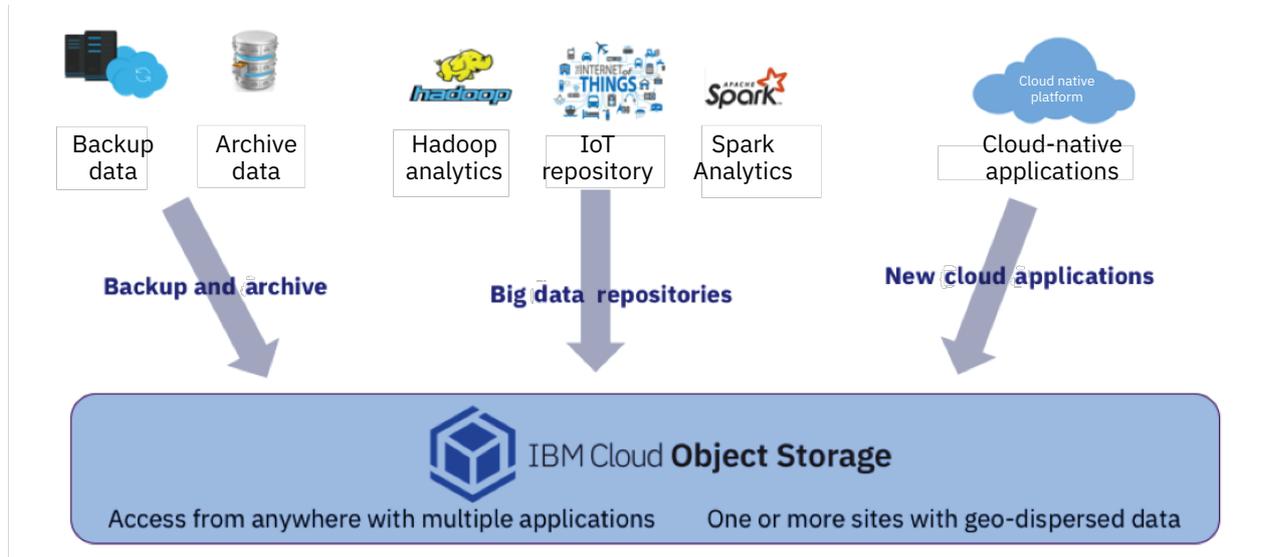
Keeping data warm for easy access

Figure 4: IBM Cloud Object Storage allows convenient warm-data access.

There are five areas where IBM Cloud Object Storage provides critical benefits:

- **Availability:** Data is always available regardless of whether there is planned or unplanned downtime.
- **Scalability:** Systems can grow from terabytes to petabytes to exabytes.
- **Security:** Data confidentiality is maintained even when multiple drives, servers, containers or locations are compromised.
- **Cost effectiveness:** The need for costly replication is eliminated, significantly lowering the total cost of ownership for storage systems at the petabyte level and beyond.
- **Efficiency:** Administrators can each manage petabytes of storage.

IBM Spectrum Archive for cost-effective tape storage for ADAS and AD development infrastructure

IBM Spectrum Archive is a member of the IBM Spectrum Storage software portfolio designed to address data storage inefficiencies by changing storage economics with a layer of intelligent software.

IBM Spectrum Archive is a cost-effective solution for retaining large amounts of data generated by smart cars. It is an ideal infrastructure solution for autonomous vehicle developers that provides an easy way to move data from test vehicles to cost-effective IBM tape drives and libraries within a tiered storage infrastructure. By using tape libraries instead of disks for Tier 2 and Tier 3 long-term storage, ADAS and AD development infrastructure can improve efficiency and reduce costs related to storing growing amounts of data.

IBM Spectrum Archive seamlessly integrates with the scalability, manageability and performance of IBM Spectrum Scale to offer an ideal ADAS and AD development infrastructure storage solution.

Using the IBM Linear Tape File System (LTFS) format, IBM Spectrum Archive provides direct, intuitive and graphical access to data stored in IBM tape drives and libraries using Linear Tape-Open (LTO) Ultrium generation 7, 6 and 5 tape cartridges, as well as in IBM 3592 cartridges in libraries with IBM TS1150 and IBM TS1140 tape drives. It eliminates the need for additional tape management and software to access data.

With IBM Spectrum Archive, accessing data stored on an IBM tape cartridge—instead of on disk—is transparent. IBM Spectrum Scale keeps a pointer to the data on tape and, if requested, retrieves the data without user or operator intervention. By leveraging the capabilities of IBM Spectrum Scale, IBM Spectrum Archive enables high performance and reliable access across the storage infrastructure. In turn, IBM Spectrum Archive enables IBM Spectrum Scale installations to add extensive capacity with lower media, floor space and power costs. And with policy-based migration, archive capacity can be expanded without impacting data availability.

IBM Spectrum Archive enables organizations to improve the cost-effectiveness of their cloud and analytics infrastructure by enabling:

- Operational storage tiers with tape, rather than storing cold data on costly disk storage.
- Storage of digital assets for the long term, so assets can be referenced and monetized for years to come.
- Creation of copies of data on operational storage, improving the efficiency and cost effectiveness of a tiered storage infrastructure.

Exceptional data insights for exabyte-scale unstructured ADAS and AD data using IBM Spectrum Discover

ADAS and AD systems comprise numerous physical sensors including radar, LIDAR, ultrasonic sensors, photonic mixer devices (PMDs), cameras and night vision devices. These devices monitor near and far fields in every direction, generating massive amounts of unstructured data when the vehicle is in motion. Correctly processing that data helps ensure vehicle, driver, passenger and pedestrian safety based on factors such as traffic, weather and unforeseen dangerous driving conditions. But deriving useful insights from so much data requires the ability to identify, classify and curate this unstructured data for analysis with performance and efficiency at scale.

IBM Spectrum Discover, a member of the IBM Spectrum Storage software portfolio, is modern metadata management software designed to provide data insights for unstructured data at exabyte scale, making it well suited for the massive amounts of data generated by ADAS and AD systems. And IBM Spectrum Discover is flexible enough to support multiple use cases including large-scale analytics, data governance and storage optimization.

How does IBM Spectrum Discover provide exceptional data insight and flexibility at scale?

- **Automatic metadata ingest:** IBM Spectrum Discover automatically ingests and indexes system metadata from both IBM Spectrum Scale and IBM Cloud Object Storage both by scanning existing stored data and in response to live events as files and objects are added or modified.
- **Beyond system-level metadata:** IBM Spectrum Discover enables users to create and manage custom metadata tags that can be applied to data based on user-defined policies. These custom tags enable data managers to classify data using domain-specific tags that can, for example, associate data with the specific sensor that collected it, or with a specific event or condition. Enriching data with such meaningful semantic labels at the time of ingest, or as soon as possible thereafter, makes it easy to curate and prepare specific, highly-targeted data for subsequent operations such as analytics or operational model training, thereby reducing time to results and accuracy.
- **Extensible platform:** IBM Spectrum Discover provides an extensible platform for integration with IBM solutions and open-source software via its Action Agent application programming interface (API). The combination of API, custom tags and policy engine can be used to support additional use cases such as content inspection, or to orchestrate data movement and more complex workflows.

Why IBM Spectrum Discover?

IBM Spectrum Discover delivers significant added value to IBM Spectrum Scale and IBM Cloud Object Storage implementations for ADAS and AD infrastructure. It enables organizations to better manage their data and storage environments in several ways:

- More effectively analyze data and uncover hidden value by identifying relevant data needed for upstream analytics processing, enabling data scientists and researchers to generate results faster and with greater accuracy.
- Improve governance to mitigate risk and improve data quality by identifying data that needs to be retained to meet internal policies, or data that presents a risk to the organization and should be removed from the storage environment. Its rapid search capabilities across large-scale environments enables faster responses to requests for legal discover or regulatory audit.
- Optimize storage infrastructure and improve utilization by identifying data that can be moved to lower-cost storage, data that no longer needs to be retained by the organization, or redundant data. This helps reduce cost as capacity can be reclaimed for new data being created. IT can more efficiently allocate storage resources based on the value or importance of the data itself.

An example use case for IBM Spectrum Discover in ADAS and AD workflow

- Define custom tags. For example, "vehicle model," "traffic condition," and "geo-location."
- Add a custom action agent that inspects a proprietary file format and extracts information relevant to ADAS and AD analytics.
- Define a policy. For example, "Fleet-Information." Associate this policy with the custom tags and action agent defined above, and schedule to run it now.
- Capture incoming auto fleet data in IBM Cloud Object Storage, which sends notifications to IBM Spectrum Discover.
- With IBM Spectrum Discover, run the Fleet-Information policy in response to these notifications.
- Using the action agent, extract information from incoming files and apply it as custom tags to the incoming data with IBM Spectrum Discover.
- Use IBM Spectrum Discover to identify relevant data using the custom metadata tags.
- Pull this data into a high-performance tier to conduct large-scale analytics, then further enrich the data by using IBM Spectrum Discover to tag the data with new custom tags derived from the analytics results.
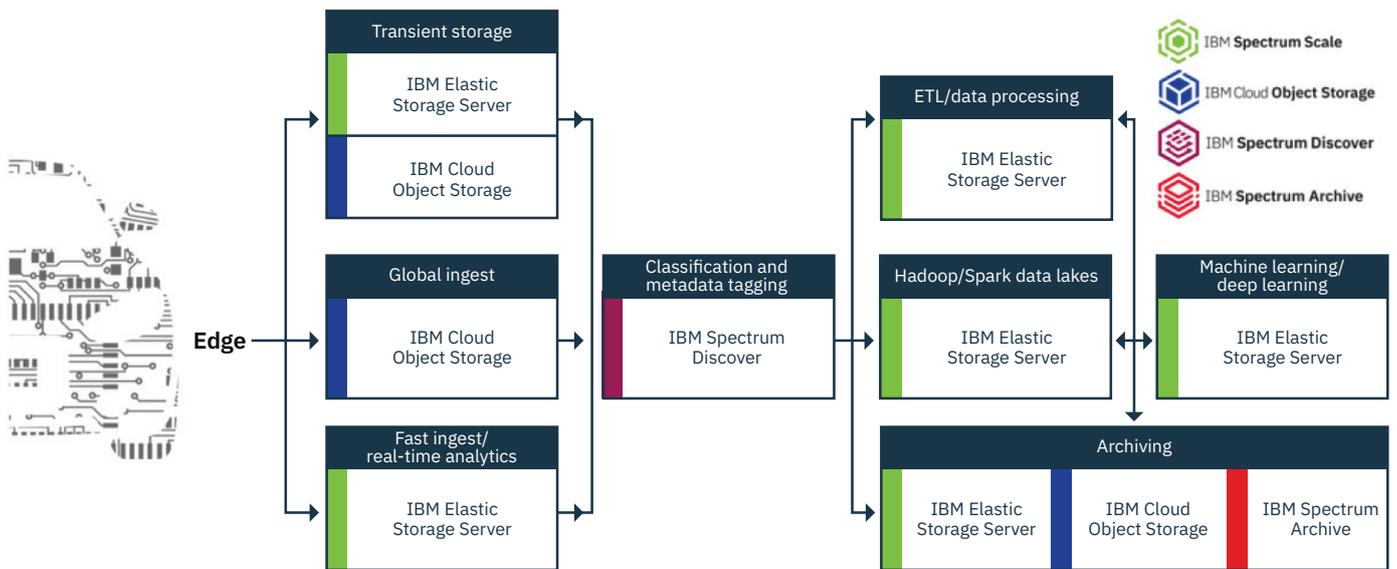

Figure 5: The IBM Spectrum Storage portfolio of intelligent storage software and IBM Elastic Storage Server can handle every stage of the data pipeline.

Why IBM?

With a long history of innovation in storage products and data management, IBM has the proven experience to help enterprises architect and deploy complex, future-oriented storage infrastructure. IBM storage solutions do more than provide a trusted resting place for data; they offer user interface and management innovations that simplify configuration and expand accessibility though ongoing metadata analysis and augmentation. The IBM Spectrum Storage portfolio of products also facilitates the seamless movement of data for business-appropriate, cost-efficient storage that balances speed, cost, and reliability across storage tiers. Together, these make IBM storage innovations well suited for today's demanding high-volume data staging, archiving and analysis.

For more information

To learn more about the IBM Spectrum Storage, please contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/storage

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing