IBM Spectrum Storage for AI with Power Systems

IBM Cloud Private for Data, IBM Watson Machine Learning Accelerator and IBM Elastic Storage Server: The Ideal Hybrid Multicloud AI Solution for Enterprises

Highlights

- Deploy enterprise-grade AI in hours, with a platform that can take you from initial prototype to widespread proliferation.
- Exploit unique data architecture of accelerated Power servers and turn deep learning performance into value.
- Leverage Large Model Support and Distributed Deep Learning to support models at higher resolution, up to almost any scale.
- Proliferate your AI deployment to multiple users and lines of business with multi-tenancy and role-based access controls. Protect them with end-to-end security and IBM technical support.
- Efficient scaling in terms of compute, storage and network capacity.

We live in a data-driven world. Successful enterprises are extracting information and intelligence from all the data being collected to better understand their customers and to create and deliver more valuable products and services for them. In the current disruptive market environment, data is driving change to business models.

Data is the life-blood of artificial intelligence and deep learning (AI and DL). As these technologies mature and applications proliferate, they will generate vast amounts of data and with them comes new storage challenges. Large datasets are required to train AI and DL algorithms to deliver accurate decisions. This, in turn, drives significant storage demands.

AI can unlock the potential in all data— internal, external, structured, unstructured, voice, and visual— and make it work together. With enterprise-grade AI, organizations can make better operational decisions, understand customer wants and needs, communicate in real time, and optimize business process, infused with the cognitive ability to understand, reason and learn.

An IBM Watson Machine Learning Accelerator and IBM Elastic Storage Server solution helps make deep learning easier, faster and extremely scalable for organizations by bringing together some of the most popular open source frameworks for deep learning, with development and management tools in a single package.

A tremendous amount of data in different formats is required to be prepared for insights, and this data needs to be hosted. This requires fast access to data to simplify enterprise grade end-to-end deep learning.

IBM Cloud Private for Data is an open, cloud-native information architecture for AI. With this integrated, fully governed team platform, you can keep your data secure at its source and add preferred data and analytics microservices flexibly. Watson Machine Learning Accelerator on IBM Power System Accelerated Compute Server (AC922) allows enterprises to spend less time on data preparation, implementation and integration, allowing more time to train neural networks for accurate results. IBM Elastic Storage Server is a modern implementation of software-defined storage based on IBM Spectrum Scale software that can easily scale out to handle petabytes or exabytes of data.

What is Watson Machine Learning Accelerator?

Watson Machine Learning Accelerator is a package of software distributions for many of the major deep learning (DL) software frameworks for model training, such as TensorFlow, Caffe, Chainer, Torch, and Theano, and their associated libraries, including CUDA Deep Neural Network (cuDNN), and nvCaffe. There are extensions that take advantage of accelerators. For example, nvCaffe is an NVIDIA extension to Caffe enabling it to work on graphical processing units (GPU). Similarly, IBM has our own extension to Caffe, called IBM Caffe. Furthermore, the IBM Watson Machine Learning Accelerator solution is optimized for performance by using the NVLink-based IBM Power System AC922 server for high performance computing. The stack also comes with supporting libraries, such as Deep Learning GPU Training System (DIGITS), OpenBLAS, Bazel, and NVIDIA Collective Communications Library (NCCL).

As part of the Watson Studio and Watson Machine Learning family, Watson Machine Learning Accelerator also comes with IBM Spectrum Conductor version 2.3.0 and IBM Spectrum Conductor Deep Learning Impact version 1.2.0 to give data scientists everything they need to build a distributed deep learning environment in hours rather than days or weeks and to easily manage it as the environment grows.

What is IBM Cloud Private for Data?

IBM Cloud Private™ for Data is a new kind of data and analytics platform with built-in governance. It simplifies and unifies how you collect, organize and analyze data to accelerate the value of data science and AI. This multicloud platform delivers a broad range of core data microservices, with the option to add more from a growing services catalog. Experience greater flexibility, security and control, and the benefits of the cloud without having to move your data.

IBM Cloud Private for Data features:

- **Single, unified platform:** Speed, time to value with a single platform that integrates data management, data governance and analysis for greater efficiency and improved use of resources. Enables self-service collaboration across teams.
- **Built-in data governance:** Efficiently respond to changing regulations with embedded, sophisticated governance capabilities; these include automated discovery and classification of data, masking of sensitive data, data zones and data lifecycle management.
- **Cloud-native agility:** Accelerate application development and deployment with a multicloud data platform that is agile, resilient and portable. Benefit from Kubernetes containerization to provision and scale services in minutes, instead of months, inside a more secure, governed environment.

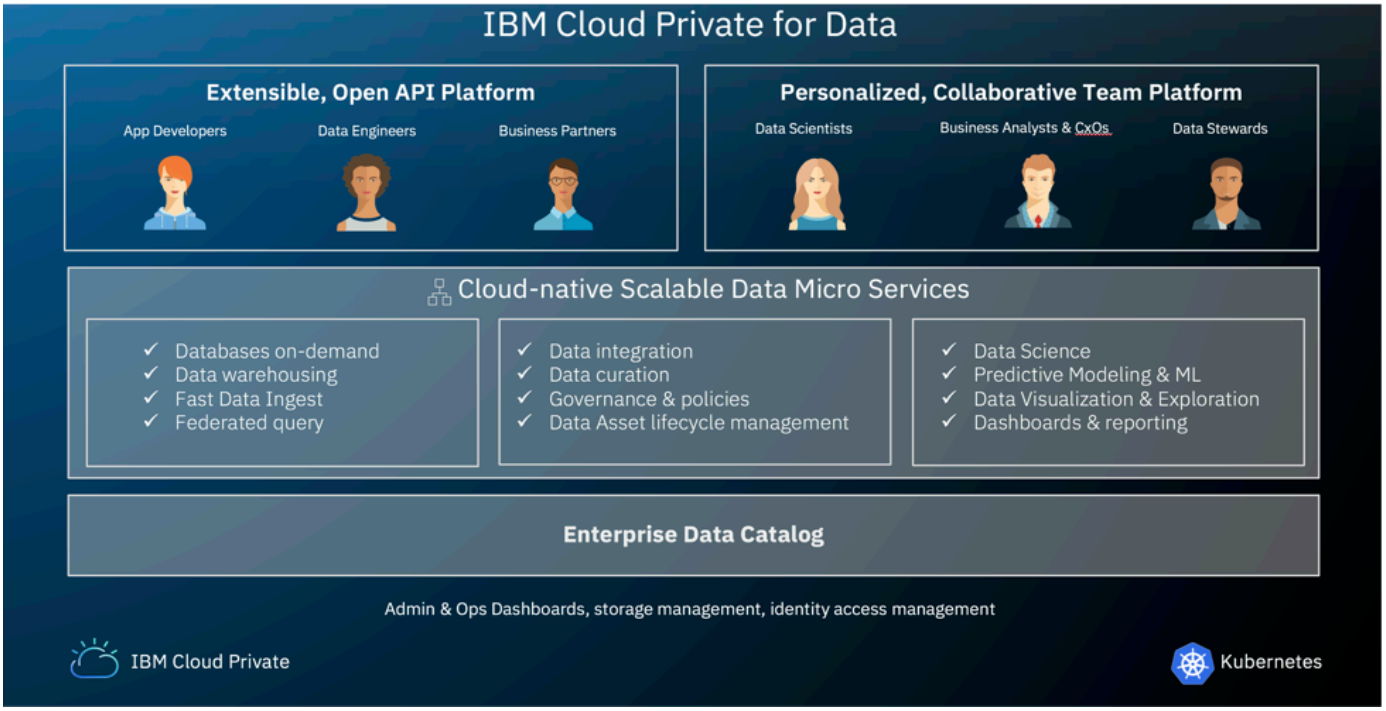

Figure 1: IBM Cloud Private for Data

- **Industry Accelerators:** Benefit from a broad ecosystem of complementary hardware, software and services through a growing service catalog. Provision preferred data services and rapidly customize data workflows to your individual needs.

- **Industry leading data virtualization:** Query data easily and more securely across multiple sources, on cloud or on premises. Exploit the combined processing power of these sources for massive query acceleration and achieve the speed and scalability your business needs for today's and tomorrow's workloads.

- **AI-ready:** Manage end-to-end data workflows to help ensure that data is easily accessible for AI. Make sure that your data is high-quality to deliver accurate, automated insights and decisions. Seamlessly build and manage machine learning models across development and production in a collaborative environment.

- **Extensible APIs and ecosystem:** Use best practices to your advantage to accelerate implementations and deliver significant business value. With IBM Cloud Private for Data, benefit from built-in models and accelerators for various industries including finance, insurance, healthcare, energy, utilities and more.

What is IBM Elastic Storage Server?

IBM Elastic Storage Server™ is a modern implementation of software-defined storage, combining IBM Spectrum Scale software with IBM Power8® processor-based I/O-intensive servers and robust, dual-ported storage enclosures. IBM Spectrum Scale is the parallel file system at the heart of IBM Elastic Storage Server. IBM Spectrum Scale scales system throughput as it grows while still providing a single namespace. This ability eliminates data silos, simplifies storage management and delivers high performance. By consolidating storage requirements across the enterprise IT infrastructure onto IBM Elastic Storage Server, enterprises effectively reduce inefficiencies, lower acquisition costs and support demanding workloads such as AI and DL.

The capabilities of IBM Elastic Storage Server include:

- **Declustered data:** IBM Spectrum Scale RAID distributes client data, redundancy information and spare space uniformly across disks. This distribution reduces the rebuild or disk-failure recovery process overhead compared to traditional RAID. Critical rebuilds of failed multi-terabyte drives full of data can be accomplished in minutes rather than hours or even days as is often the case when using traditional RAID technology.

- **Data redundancy:** IBM Spectrum Scale RAID supports highly reliable 2-fault-tolerant Reed-Solomon-based parity codes (erasure coding) as well as three-way and four-way replication.

- **Tuned performance:** Software-defined IBM Spectrum Scale RAID software, explicitly coupled with large memory cache in the IBM Power server, allows IBM Elastic Storage Server to mask the inefficiencies and long latency times of nearline-SAS drives with low-latency flash storage, while still leveraging the high density of the drives themselves.

- **Simplified management:** The intuitive graphical user interface (GUI) of software and systems for management and monitoring of IBM Elastic Storage Server also integrates into IBM Spectrum Control™

- **Superior streaming performance:** The system can deliver more than 36 GB/s sustained performance.

- **Scalability and extensibility with multi-site and cloud support:** Multiple deployment options for software-defined storage to scale in performance and capacity while still providing a single namespace. This means installations can start small and grow as data needs expand.

An AI future needs storage with unlimited scale

Large datasets are required to train AI and DL algorithms to deliver accurate decisions. This, in turn, drives significant storage demands. For example, five years of continuous speech data was needed to allow computers to talk and billions of miles of driving data are in the works for autonomous driving. Managing these datasets requires storage systems that can scale without limits.

After the AI algorithms are trained, they will start generating their own data. The original dataset will expand and improve with use. For that to happen, data must be given context through metadata. But humans can't manually add context to each piece of data; the sheer amount of data would take weeks or months for a human to analyze. Artificial intelligence systems, however, can process these growing datasets in a matter of minutes. Thus, the use of AI to improve AI will further increase the demand for data storage.

IBM Elastic Storage Server is easy and seamless and scales on demand. Storing petabytes of data is of no value to businesses unless that data can be accessed, retrieved and analyzed quickly. Sustained streaming performance of data can reach 40 GB/s in each building block and grow as more blocks are added to a configuration. By combining the superior data movement capability of IBM Spectrum Scale RAID, a complete

storage solution can be deployed to support AI and DL workloads.

Best-in-class AI and DL storage architecture

AI and DL learn from many different data types, which require varying performance capabilities. As a result, systems must include the right mix of storage technologies. A hybrid architecture is needed to meet the simultaneous needs for scale and performance. Also, for datasets that grow without limits, a parallel-access architecture is essential. Without it, the systems will develop bottlenecks that limit data growth.

IBM Elastic Storage Server provides superior low latency performance for AI and DL workloads. With support for multiple 10GbE, 40GbE or 100GbE Ethernet and Enhanced Data Rate (EDR) or (fourteen data rate) FDR InfiniBand, IBM Elastic Storage Server has the parallel architecture to deliver high data throughput to meet the demands of AI and DL workloads.

How does Watson Machine Learning Accelerator simplify adoption of AI and DL?

Today, organizations are using DL to develop powerful new analytics capabilities spanning multiple usage patterns from computer vision, object detection, and improved human computer interaction through natural language processing (NLP) to sophisticated anomaly detection capabilities. However, there are enormous difficulties when organizations try to expand their area of DL or start working on the development of DL, such as performance issues caused by hardware limitations and time-consuming processes in each framework regarding setup, tuning, upgrading and others.

The main objective of Watson Machine Learning Accelerator is to reduce the time required to develop, train and deliver production-grade AI systems by making it easy with an enterprise-ready software distribution for DL and AI that simplifies the development and training experience.

Watson Machine Learning Accelerator provides an end-to-end, deep learning platform for data scientists. This includes complete lifecycle management from installation and configuration; data ingest and preparation; building, optimizing, and distributing the training model; to moving the model into production. Watson Machine Learning Accelerator enables you to iterate quickly through the training cycle using more data to continuously improve the model over time.

Watson Machine Learning Accelerator provides many optimizations that accelerate performance, improve resource utilization, and reduce installation, configuration, and management complexities, such as the following:

- Distributed deep learning architecture that simplifies the process of training deep learning models across a cluster for faster time to results.
- Large model support that helps increase the amount of memory available for deep learning models up to 16 GB or 32 GB per network layer, enabling more complex models with larger, more high-resolution data inputs.
- Enhanced data ingest, preparation, and transformation tools, using Apache Spark for data management.
- Powerful model development tools, including real-time training visualization and runtime monitoring of accuracy and hyper-parameter search and optimization, for faster model development.
- Ready-to-use deep learning frameworks (TensorFlow, Caffe-BVLC and IBM Caffe) are included.
- Multitenant architecture designed to run deep learning, high-performance analytics, and other long-running services and frameworks on shared resources.

Key benefits of the AI and DL solution

- **Extreme scalability with parallel file system architecture:** IBM Elastic Storage Server with IBM Spectrum Scale software is a parallel architecture. No single storage node can become a bottleneck, and every node in the cluster can serve both data and metadata, enabling a single IBM Spectrum Scale file system to store billion of files. This architecture enables enterprises to grow their AI and DL environments seamlessly as the data grows. Additionally, among the most useful and valuable attributes of IBM Elastic Storage Server are its ability to run diverse and demanding workloads, and the ability to tier infrequently accessed data to active archive.
- **A global namespace that can span multiple AI and DL environments and geographical areas:** Using IBM Spectrum Scale global namespace, enterprises can create active, remote data copies and enable real-time, global collaboration. This namespace enables global organizations to form data lakes around the world and host their distributed data under a single namespace. Spectrum Scale also enables multiple AI and DL cluster environments to access a single file system while still providing all the required data isolation semantics. The Transparent Cloud Tiering feature of Spectrum Scale can archive data into a S3/SWIFT compatible cloud object storage system, such as IBM Cloud™ Object Storage or Amazon S3, by using the powerful Spectrum Scale Information Lifecycle Management (ILM) policies.

- **A reduced data center footprint with the industry's best in-place analytics:** IBM Spectrum Scale has the most comprehensive support for data access protocols. It supports data access by using NFS, Object, POSIX and HDFS (Hadoop Distributed File System) API. This eliminates the need to maintain separate copies of the same data for traditional applications and for analytics. Moreover, enterprises can leverage Spectrum Scale POSIX support with flash-based IBM Elastic Storage Server to achieve super-fast ingest from their existing data lakes.

- **True software-defined shared storage:** With IBM Elastic Storage Server, enterprises can control cluster sprawl and grow storage independently of the compute infrastructure. IBM Elastic Storage Server uses erasure coding to eliminate the need for the three-way replication for data protection.

- **IBM hardware advantage:** A key advantage for IBM Elastic Storage Server is to lower capacity requirements. IBM Power Systems™ servers along with IBM Elastic Storage Server offer the most optimized hardware stack for running AI and DL workloads. Enterprises can enjoy up to three times reduction of storage and compute infrastructure by moving to IBM Elastic Storage Server compared to commodity scale-out x86 systems.

To support the security and regulatory compliance requirements of organizations, IBM Spectrum Scale offers Federal Information Processing Standard (FIPS) compliant data encryption for secure data at rest, policy-based tiering/ILM, cold data compression, disaster recovery, snapshots, backup and secure erase.

Optimized infrastructure and efficient data management

Figure 2 illustrates the AI workflow as a process cycle. Once you produce a trained neural network model, you go back and retrain the model with new data to keep it current and to improve its accuracy. Training ML and DL models requires huge quantities of data. Both training and inferencing are compute-intensive and require high performance for fast execution. AI applications push the limits on thousands of GPU cores or thousands of CPU servers. AI and DL workloads require a new class of accelerated infrastructure primarily based on GPUs. For the linear math computations needed for training neural network models, a single system configured with GPUs is significantly more powerful than a cluster of non-accelerated systems.

Parallel compute demands parallel storage. While the training phase requires large data stores, inferencing has less need for it. The inference models are often stored in a DevOps-style repository where they benefit from ultra-low-latency access. While the training phase is set once the execution model has been developed based on the data and the workload has moved to the inferencing stage, re-training of the model is often needed as new or modified data becomes available. In some cases, the real-time nature of the application may require near constant re-training and updating of the model. Enterprises may also benefit from re-training the model over time with additional data sources and to capitalize on new insights.

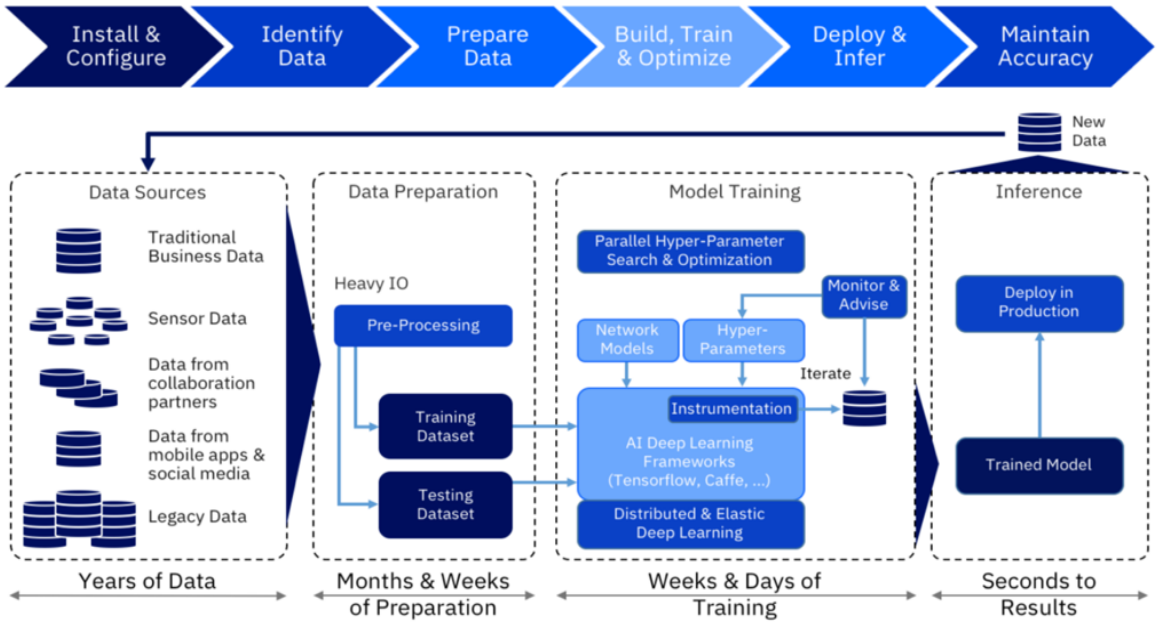

Figure 2: AI Workflow

Enterprises cannot support cutting-edge applications like AI on legacy infrastructure that is challenged to meet such requirements for scale, elasticity, compute power, performance, and data management. Organizations are now using different infrastructure solutions and approaches to support the data pipeline for AI, a process that generally leads to data silos. Some create duplicate copies of the data for the pipeline to avoid disturbing stable applications. Instead, enterprises need to adopt infrastructure that is dynamically adaptable, scalable and intelligent (self-configurable, self-optimizing, and self-healing). Such an infrastructure is tuned for various data formats and access, can process and analyze the required volumes of data, and provides the performance necessary to support faster compute calculations and decision making, while also managing risks and reducing the overall cost of AI deployments.

Simplifying deployment with IBM Power Accelerated Computing Platform (IBM ACP)

The IBM® Power® Accelerated Computing Platform (IBM Power ACP) system, based upon the world's smartest and most powerful compute installations at Oak Ridge and Lawrence Livermore National Labs, is a complete offering of IBM Power System Accelerated Compute Server (AC922); IBM Elastic Storage™ Server (ESS); networking, development and runtime software; and professional services, designed to help any nascent organization support AI and HPC workload.

The key to successfully deploying these IBM Power ACP systems is consultation and configuration by IBM Systems Lab Service Engineers. An IBM Power ACP system will be delivered to your location, rack-mounted and integrated, with the operating systems and required software preloaded and configured as defined in the implementation design workshop. One size never fits all, and IBM wants to accelerate your time-to-success by ensuring the deployed solution is quickly installed and moved into production. IBM Power ACP can provide the world's most powerful servers customized for each client's unique environment.

By combining IBM Cloud Private for Data, with IBM Watson Machine Learning Accelerator, IBM Elastic Storage Server powered with IBM Spectrum Scale software and IBM Power Systems Accelerated Computing Platform, enterprises can quickly deploy an optimized, high performance, and fully supported multicloud infrastructure for AI/ML/DL workloads. This approach lessens the challenges of pulling a solution together from open source. All the components of the IBM solution are sourced and supported by IBM. With this solution, enterprises can simplify the development experience, reduce the time required for training AI models, and reduce time to model accuracy and insights.

For more information

To learn more about IBM Spectrum Storage for AI, or about IBM Watson Machine Learning Accelerator, IBM Elastic Storage Server, IBM Power System AC922 server, or IBM Power Accelerated Computing Platform, please contact your IBM representative or IBM Business Partner, or visit:

IBM Spectrum Storage for AI:
https://www.ibm.com/it-infrastructure/storage/ai-infrastructure

IBM Watson ML Accelerator:
ibm.com/us-en/marketplace/deep-learning-platform

IBM Elastic Storage Server:
ibm.com/us-en/marketplace/ibm-elastic-storage-server

IBM Power System AC922:
ibm.com/us-en/marketplace/power-systems-ac922

IBM Corporation
IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
March 2019