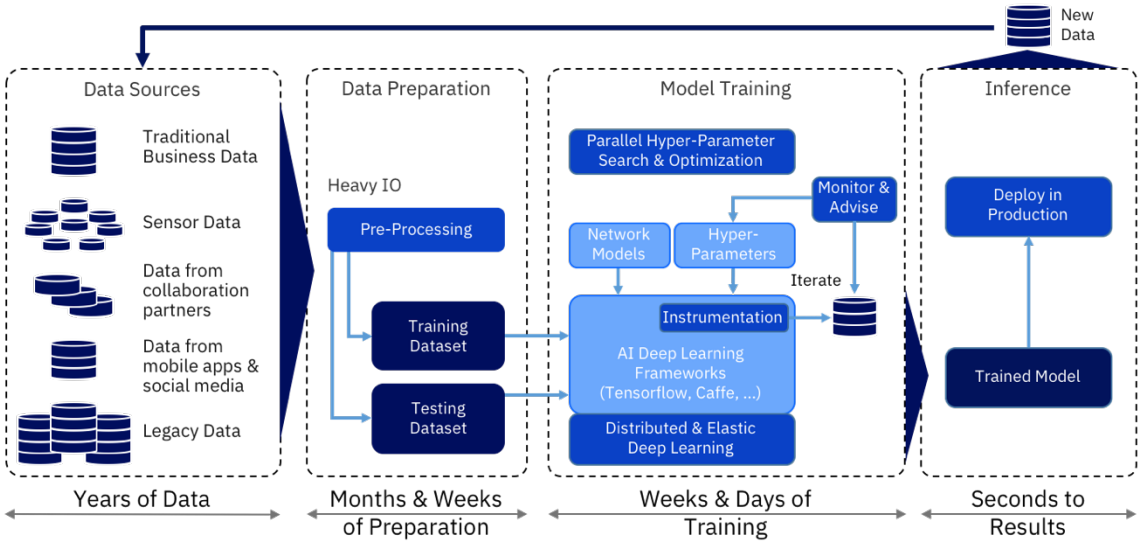
IBM Systems
Solution Brief

IBM AI Infrastructure Reference Architecture

Organizations are using Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) to develop powerful new analytics spanning multiple usage patterns, from computer vision and object detection, to improved human computer interaction through natural language processing (NLP) and very sophisticated anomaly detection capabilities. To achieve these potentials for business transformation and scientific breakthroughs, AI initiatives must meet two criteria 1) accuracy in results and 2) timeliness in providing value. Developing the right AI infrastructure is critical in achieving those objectives.



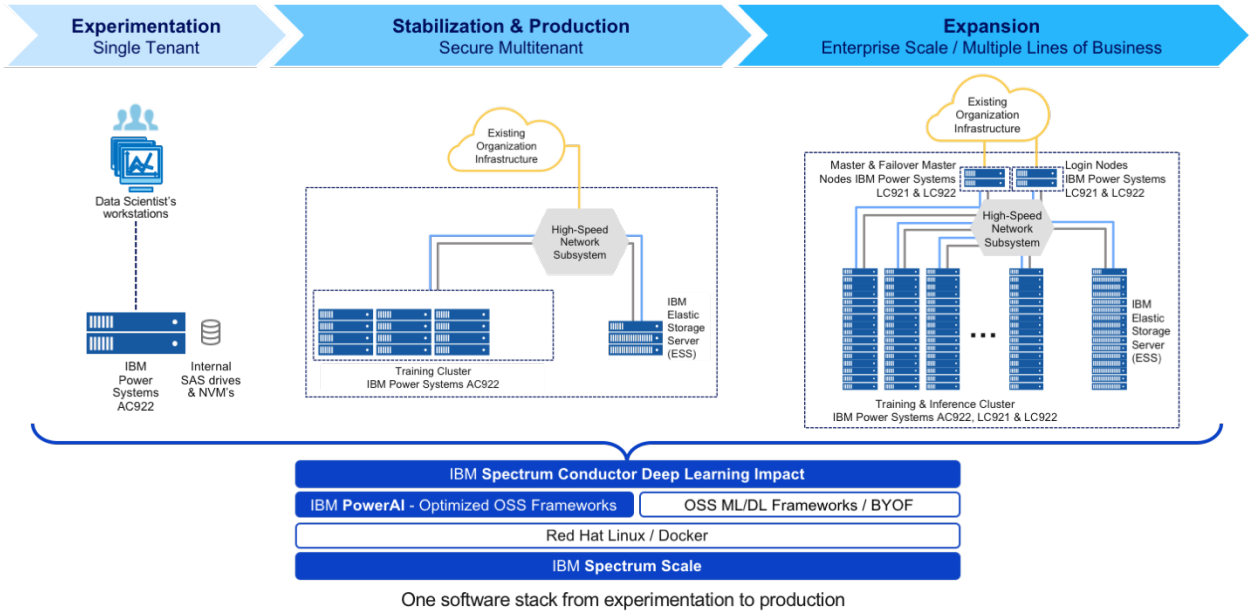
But, implementing AI infrastructure can be challenging. It is built on a complex mix of emerging and rapidly changing technologies and requires high performance resources not commonly found in enterprise environments. Steep data science learning curves and the complexity of open source frameworks means it can take weeks to get up and running. IT departments are challenged with providing the computing power and storage capacity needed for timely AI data management, preparation, and model training.

The IBM AI Infrastructure Reference Architecture is the result of working with 100's of customers deploying on-premise and hybrid cloud AI solutions. Through simplifying the complexity of AI initiatives, speeding the time to model accuracy and providing efficient high-performance, enterprise-grade infrastructure, this reference architecture delivers cost-effective, accurate and timely AI value.

85016785USEN-00

IBM AI Infrastructure Reference Architecture

This reference architecture is intended to be used as a reference by data scientists and IT professionals who are defining, deploying and integrating AI solutions into an organization. This document describes an architecture that will facilitate a productive proof of concepts (PoC) and allow growth into a multitenant, production system that allows for sustained growth to enterprise scale, while integrating the solution into an organization's existing IT infrastructure.



One software stack from experimentation to production

Features of the Reference Architecture

Accelerated servers - GPU accelerated servers delivering 2 to 3 times higher throughput and performance than commodity servers to speed model training and optimization.

Accelerated servers – speed model training and optimization with GPU accelerated servers that deliver 2 to 3 times higher performance than commodity Linux servers.

Storage resource connectors – simplify and maintain ongoing connections to multiple data sources

Multitenancy – reduce wasted time, cost and administrative overhead and provides access to a larger cluster by supporting multiple data scientists, frameworks and applications on a common shared compute cluster.

Resource sharing while maximizing the utilization of server and GPU resources.

Elastic Distributed Training – dynamically assign GPUs to models while training speeding time to results. GPUs can be added and removed without needing to stop the training enables GPU sharing and provides resiliency to failures

Enterprise Suitability

This reference architecture describes systems that are deployed in enterprise environments including regulated and many top financial institutions worldwide. This enterprise suitability is built around the physical infrastructure (i.e., servers, storage and network), and the software platform (i.e., OS, storage management, workflow, resource and workload management). Six characteristics of t enterprise suitability.

Reliability and Availability

Reliability and availability are cornerstones for deploying into an enterprise environment, especially as it transitions out of development and the business begins to rely on it as mission critical. The technologies used in this Reference Architecture are production proven by decades of implementation including highly scalable, reliable and redundant storage, server infrastructure, distributed computing and storage software management stack and an OS already trusted by organizations.

Heterogeneous Environments

The computing and storage management platforms can take advantage of and manage existing non-accelerated IBM Power Systems servers, x86 based servers and non-IBM storage, extending ROI and reducing administrative overhead.

Security

The technologies used in this reference architecture use the latest security protocols and have been subjected to extensive security scanning and penetration testing, allowing them to be deployed into many regulated organizations including major financial and governmental institutions. End-to-end security, from data acquisition and preparation to training and inference implements authentication, authorization, impersonation and encryption.

Scalability

The reference architecture is designed to lay the foundation for efficient scaling in terms of compute, storage and network capacity. The approach is to start small, with a couple of servers and existing storage, and then scale by adding compute and GPU or storage as needed. Additional compute, storage and software can be added without the need for planned down time and scale in near linear fashion with little degradation in performance.

Support and Maintenance

All components specified in this reference architecture acquired from IBM (i.e., servers, storage, networks, OS, software including AI frameworks) are fully backed by IBM levels 1 - 3 support and services providing the convenience of a single point of contact.

Cloud-ready

The computing and storage software specified in this reference architecture are cloud ready for on-premises, cloud and hybrid cloud configurations. When additional resources are needed to handle peak loads or if the data is already located in the cloud, additional cloud-based compute and storage resources can be easily accessed.

TCO and ROI

AI requires a new approach to infrastructure and the technologies presented in the reference architecture are designed to maximize the value of new investments and to take advantage of existing infrastructure.

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America


Please Recycle

85016785USEN-00