Accelerate Your AI Journey with

a Hyperconverged Data and Analytics Platform

Sponsored by IBM and Intel

Ritu Jyoti

February 2020

INTRODUCTION AND OVERVIEW

By 2024, with proactive, hyperspeed operational changes and market reactions, artificial intelligence (AI)-powered enterprises will respond to customers, competitors, regulators, and partners 50% faster than their peers. These digital transformation (DX) initiatives will be supported by AI capabilities, providing timely critical insights, richer and immersive user experiences, and improved business outcomes.

IDC forecasts that global AI spending will reach $97.9 billion by 2023, driven mostly by deployments in banking, retail, and manufacturing. However, AI adoption has been slow. Automated customer service agents, IT automation, sales process recommendation, and automation are the current top uses, but we expect to see automated human resources, digital assistants for enterprise knowledge workers, regulatory intelligence, and advanced digital simulation emerge as the fastest growing use cases over the next five years.

AI will be a true differentiator, with services that run from edge to core to cloud and hybrid and multicloud deployments as the new norm. Organizations that master AI will take off; those that don't will dwindle. Effectively applying machine learning (ML) for business benefit requires:

1. **ML training.** ML training consists of the steps required to build the ML model and can include model generation, model build, and model fit.

2. **ML inference.** ML inference (i.e. prediction, scoring, or model serve) generates the insights that need to be integrated into a business use case, creating an ML business application that ultimately generates customer value.

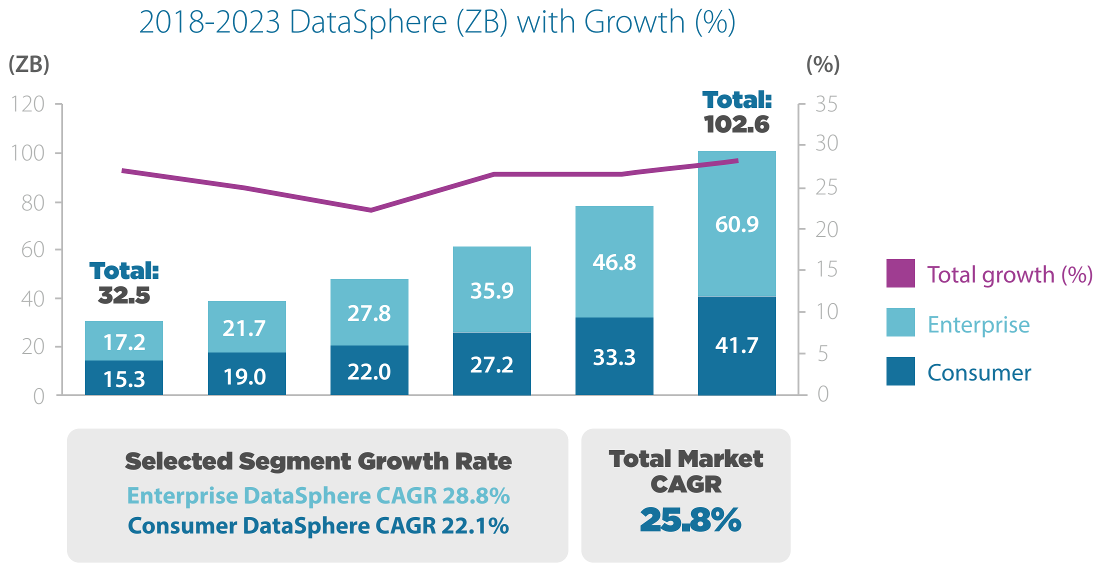
ML training and ML inference do not exist in isolation. There is always a cycle that connects them. Models generated by ML training need to be sent to ML inference, and immediately or eventually, the experiences of live data must be used to further optimize the model in the next round of ML training. ML inference is integrated into the business use case, creating an application that generates customer value. In recent history, this cycle could span months or even years, and as such, it was almost possible to forget it even existed. However, now with advances in training algorithms, powerful hardware, and scalable analytic engines, runtimes for each phase have been substantially reduced. An AI-optimized hyperconverged data and analytics platform can truly accelerate the AI journey and enable faster realization of business value.

SITUATION OVERVIEW

The Global DataSphere Is Exploding, But Little Is Useful for AI

Successful digital transformation relies on converting data into actionable insights, and this increasing dependence on data-powered insights is contributing to a new era of the data age. In fact, IDC forecasts that by 2023, the Global DataSphere — all data created and consumed worldwide — will grow to 102.6ZB (see Figure 1). All this data has the potential to unlock unique user experiences and countless new business opportunities.

Figure 1. Worldwide Global DataSphere (102.6ZB by 2023)



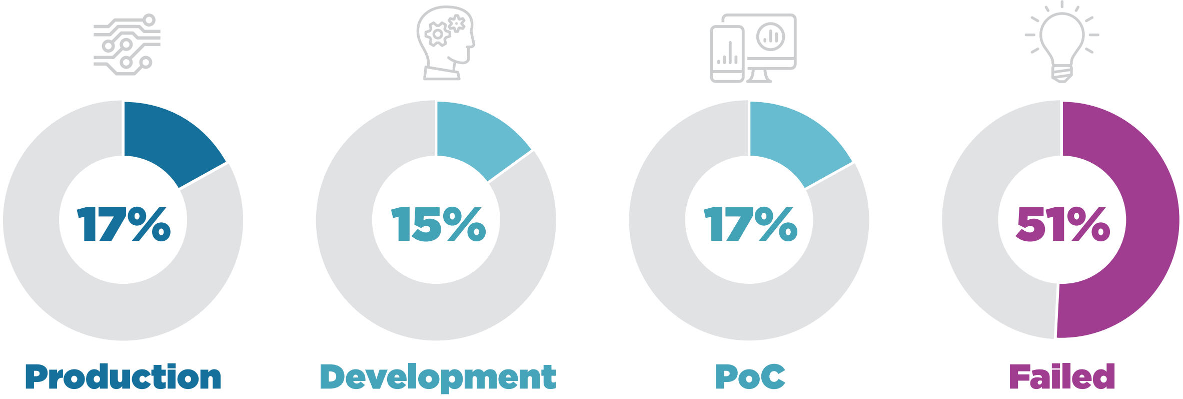
Source: *IDC, Worldwide Global DataSphere Forecast, 2019–2023: Consumer Dependence on the Enterprise Widening, #US44615319*

However, the explosion in data creation is not the only defining aspect of the new data age. It is also important to understand that just 27% of the data created will be "useful if tagged." What's more, only 44% of the "useful if tagged" data is tagged, and only 21% of the "tagged" data is analyzed. From there, only 15% of the "analyzed" data is fed into AI systems. Calculate all these cascading percentages, and the problematic result is that less than 1% of the global datasphere is currently used by AI systems; the remainder is dormant or dark data, which is not currently being used for insights or decision making. And because the shortage of usable data is so severe, businesses are creating synthetic data (a repository of data that is generated programmatically) to build the repositories needed to train ML models.

AI Adoption Trends and Challenges

AI is a true competitive differentiator; it improves business agility and accelerates time to market with newer products and services. According to IDC's *Global Artificial Intelligence (AI) Survey* conducted in May 2019, only 17% of all AI initiatives are in production, another 15% are in development, and 17% are in proof of concept stage (see Figure 2). In contrast, over half, or 51%, have failed.

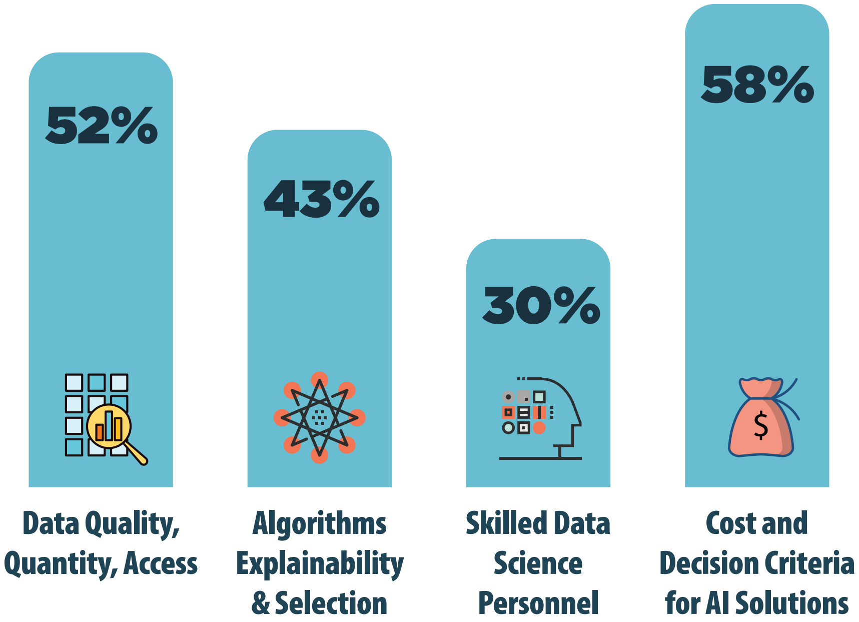
Figure 2. Current Reality of AI Deployments



Source: Global Artificial Intelligence (AI) Survey IDC, May 2019

Cost and AI solution decision criteria challenges, data quality and access, challenges with algorithms, and data science skills shortage are the key factors holding businesses back from implementing AI (see Figure 3).

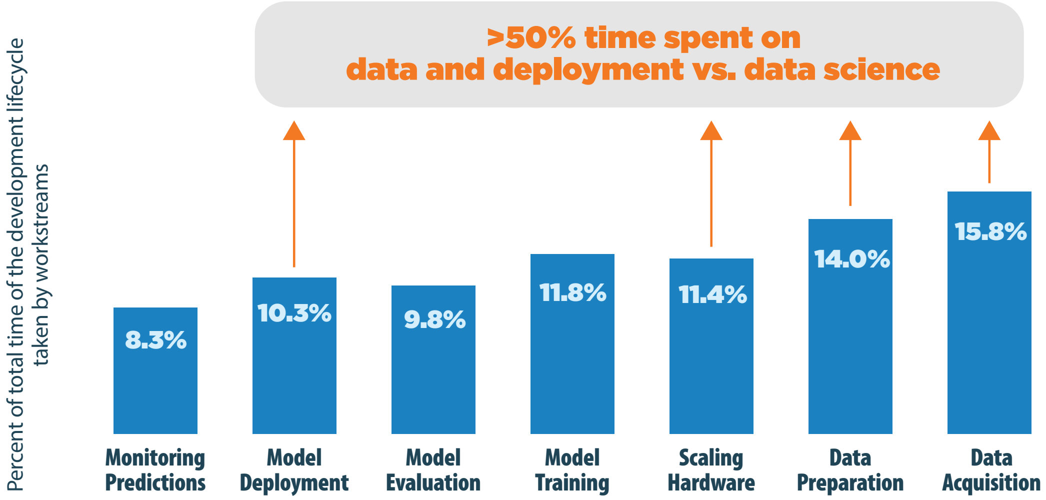
Figure 3. Top Factors Holding Back AI Deployments



Source: Global Artificial Intelligence (AI) Survey, IDC May 2019

In another IDC study, businesses report spending more than 50% of their time on data preparation and deployment, as opposed to actual data science (see Figure 4). Scaling infrastructure and performance inhibit realization of business value.

Figure 4. Data and Deployment Tasks are Time Consuming



>50% time spent on data and deployment vs. data science

Source: IDC, 2019

The New Competitive Frontier: MLOps

In order to meet the challenges associated with AI deployments, businesses are innovating with the emerging practice of MLOps (an amalgam of "machine learning" and information technology "operations"). MLOps focuses on collaboration and communication between data scientists and operations professionals, and it spans the entire ML/deep learning (DL) lifecycle, from experimentation to production, including:

- Data integration and cataloging
- Data drift and concept drift support
- AutoML
- Production monitoring of compliance and safeguards

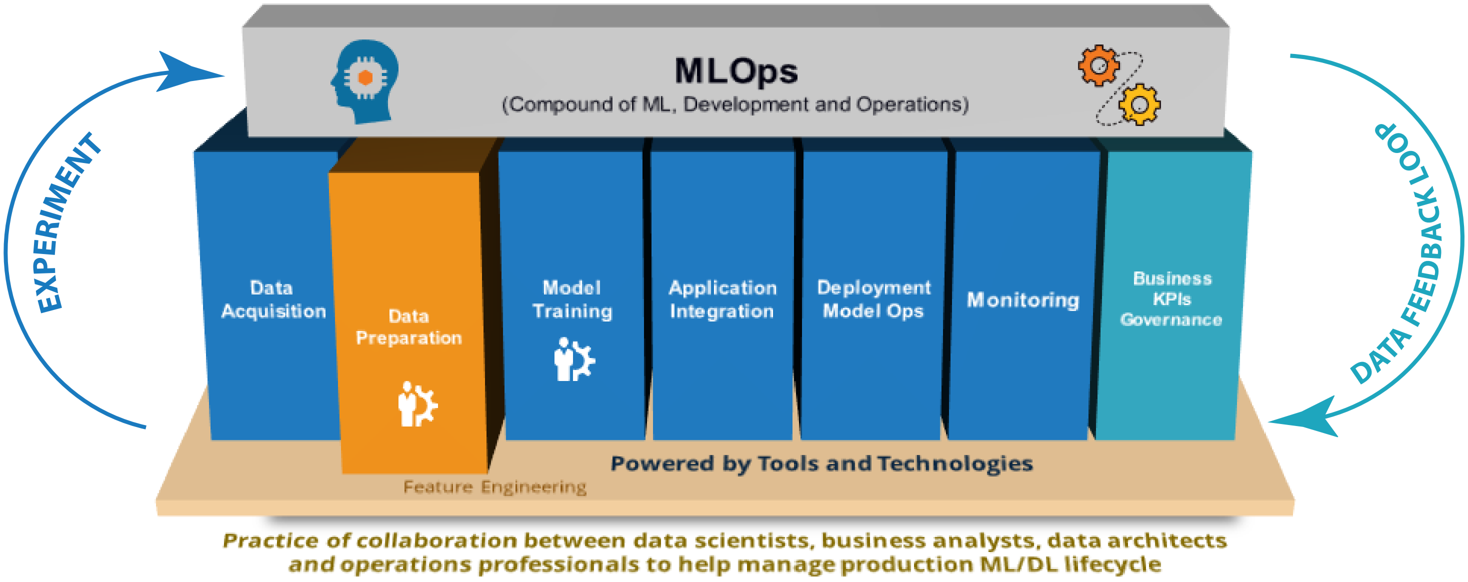
Bias remediation and transparency support

Operationalization of models

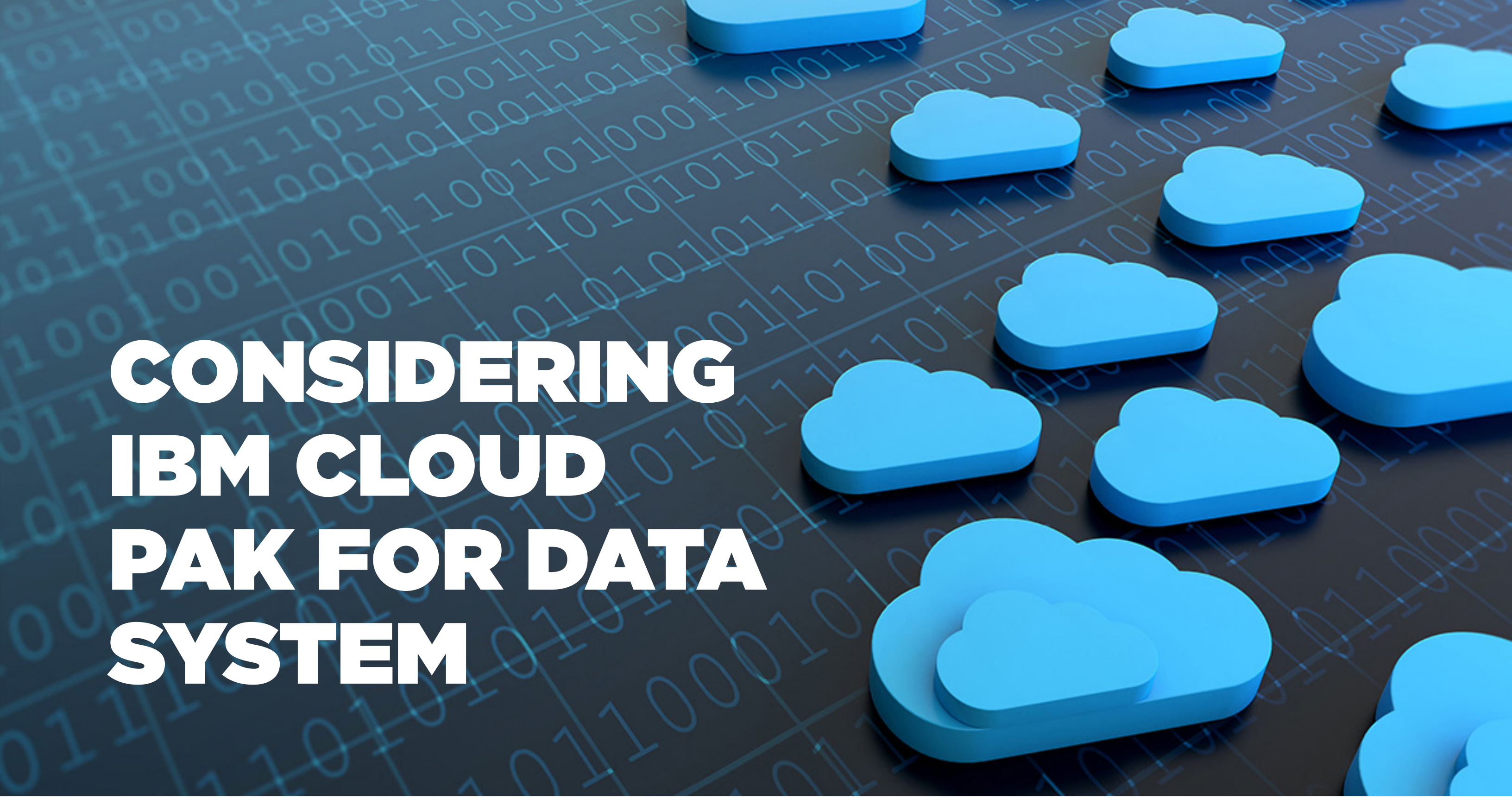
Security and access controls

Powered by tools and technologies, MLOps offers container support for hybrid, multicloud, and edge deployments. It creates an end-to-end optimized elastic stack that supports the ability to start small and scale capacity and performance, while also maintaining consistent and predictable performance. As a result, effective MLOps keeps costs in check, simplifies management, and accelerates time to value (see Figure 5).

Figure 5. MLOps Spans the ML/DL Lifecycle


MLOps
(Compound of ML, Development and Operations)
EXPERIMENT
DATA FEEDBACK LOOP
Data Acquisition
Data Preparation
Model Training
Application Integration
Deployment Model Ops
Monitoring
Business KPIs Governance
Powered by Tools and Technologies
Feature Engineering
Practice of collaboration between data scientists, business analysts, data architects and operations professionals to help manage production ML/DL lifecycle

Source: IDC, 2019

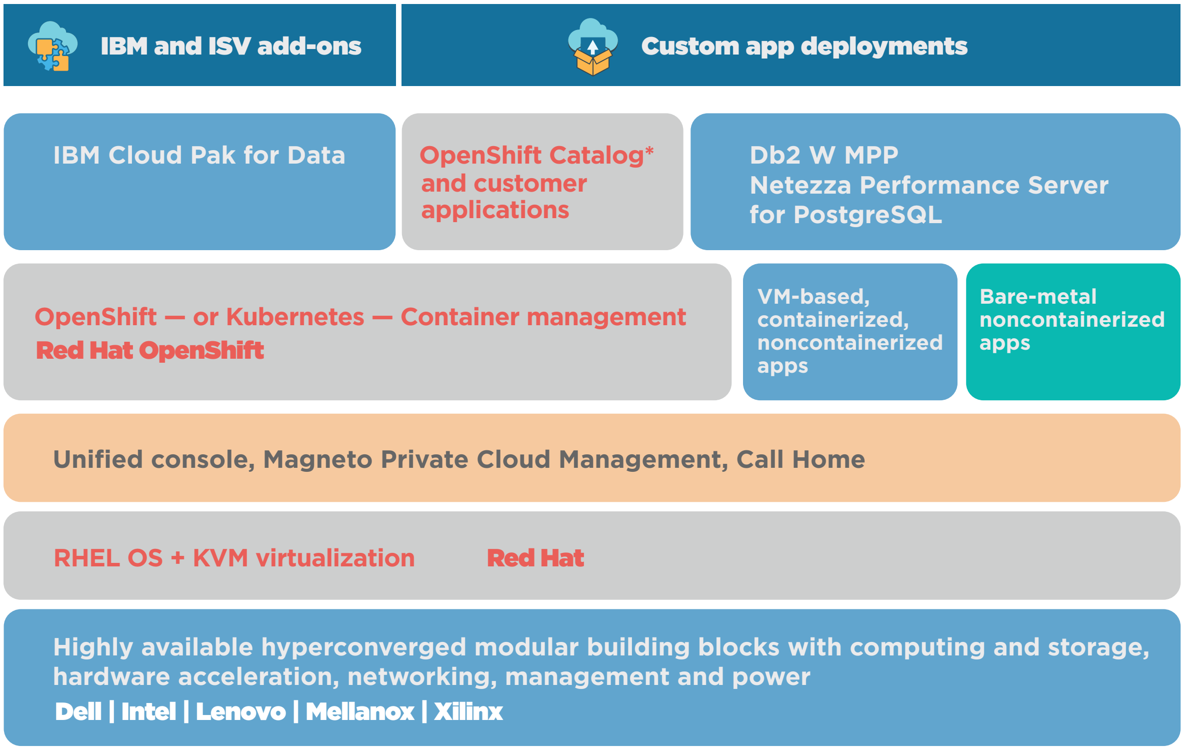


CONSIDERING IBM CLOUD PAK FOR DATA SYSTEM

The IBM Cloud Pak for Data System is designed to make the MLOps goal of simplified AI lifecycle management and increased collaboration attainable. As a hyperconverged, cloud-native data and AI platform, IBM Cloud Pak for Data System provides a pre-configured, governed, and secure environment to collect, organize, and analyze data (see Figure 6). The platform gives organizations the capabilities to take advantage of a broad set of data and AI services and integrate them into applications to accelerate time to value, time to insight, and time to market.

IBM Cloud Pak for Data System is built on the Red Hat OpenShift Container Platform and combines storage, compute, networking, and software into plug-and-play nodes. This hyperconverged architecture simplifies software and hardware management and can speed private cloud deployment to a matter of hours. A flexible pay-as-you-go model can help businesses keep their costs in check, with the ability to scale-out compute performance as well as storage capacity.

Figure 6. IBM Cloud Pak for Data System Architecture


IBM and ISV add-ons
Custom app deployments
IBM Cloud Pak for Data
OpenShift Catalog* and customer applications
Db2 W MPP
Netezza Performance Server for PostgreSQL
OpenShift — or Kubernetes — Container management
Red Hat OpenShift
VM-based, containerized, noncontainerized apps
Bare-metal noncontainerized apps
Unified console, Magneto Private Cloud Management, Call Home
RHEL OS + KVM virtualization
Red Hat
Highly available hyperconverged modular building blocks with computing and storage, hardware acceleration, networking, management and power
Dell | Intel | Lenovo | Mellanox | Xilinx

** Requires customer to bring container platform license*

Built on Red Hat

With IBM Cloud Pak for Data System, organizations can collect, organize, and analyze data for use in AI-infused applications. The IBM Cloud Pak for Data System is a multicloud data and AI platform delivering an information architecture for AI that offers flexibility, security, and control, along with the benefits of the cloud without having to move data. It helps you build, run and manage AI/ML models with Watson Studio, available as part of the base components in IBM Cloud Pak for Data.

IBM's end-to-end data science toolkit can help data scientists of all skill levels to:

- Prepare data
- Build AI models
- Train ML/DL, either through an interactive or batch paradigm
- Deploy and manage the lifecycle of models

- Enable GPU acceleration to train models as well as leverage GPUs and field programmable gate arrays (FPGAs) for inferencing
- Scale to enterprise-wide deployments of AI models

Core Capabilities of IBM Cloud Pak for Data System

The IBM Cloud Pak for Data System provides a modular approach to compute, network, and storage on standard hardware. Its core capabilities include:

- Red Hat OpenShift support which is certified across IBM Cloud Pak for Data services
- Open source governance capabilities for managing risks and accelerating open source-based AI projects delivery
- Automated development of AI models supported by AutoAI
- Built-in data science and machine learning
- High performance analytics (powered by Netezza Performance Server for PostgreSQL, which is 100% compatible with Netezza) in "cloud-in-a-box" setup to take advantage of hyper-converged modularity
- Visual application building, near real-time visual debugging, and support for Red Hat AMQ Streams
- New industry-specific accelerators
- New offerings called packages which includes IBM Cloud Pak for Data entitlements that are required to run the service
- New licensing model for IBM Cloud Pak for Data that enables businesses to purchase licenses to IBM software in a cloud-centric model aligned with a Red Hat subscription model to deliver a consistent buying experience across both product portfolios
- Extensible third-party services like Figure Eight to help annotate training data and fuel machine learning initiatives

Intel Can Speed AI Results

Intel delivers an innovative and flexible portfolio of processors and accelerators across the entire data pipeline. Many of the most prevalent AI and analytics frameworks have been optimized for Intel Xeon Scalable Processors to significantly boost performance. These are available as optimized software extension services, with direct access from IBM Cloud Pak for Data System software.

Intel Xeon processors, accelerators and workload optimized software add-ons can help accelerate AI results and time to actions. These technologies can help speed development and insights with optimized frameworks and help collect and organize data faster, with a modernized, high-performance and secure platform that protects data.

Intel Data Center SSDs using the NVMe interface are part of the IBM Cloud Pak for Data solution, providing the throughput and consistently low latency required to support applications like analytics. Underlying the IBM Cloud Pak for Data software, the Red Hat OpenShift Container Platform benefits from the close collaboration between Intel and Red Hat, with investments in co-engineering and optimizations using Intel's libraries and tools to ensure the software can take advantage of hardware enhanced technologies.

Adaptable as Business Needs Change

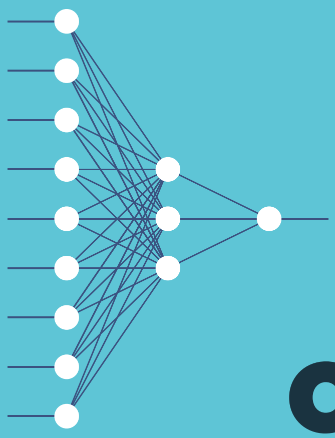
Designed as "plug-and-play," IBM Cloud Pak for Data System is also designed to provide "plug-and-grow" capabilities, allowing bare metal compute and storage nodes to be added, recognized, and provisioned to help meet evolving business needs. Built-in data virtualization enables access to new data sources via a single, unified console that eliminates having to move data and is designed to provide a seamless user experience of data and AI capabilities.

CHALLENGES AND OPPORTUNITIES

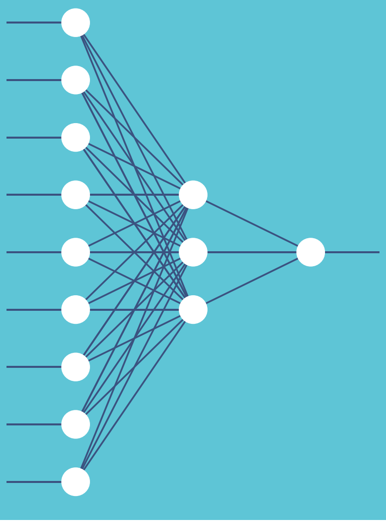
Unrealistic business expectations from AI and a lack of data science skills are noted as some of the key inhibitors to AI adoption. IDC recommends IBM partner with the breadth of system integrators including the newer breed of AI-focused professional services players to truly simplify the end-to-end AI build, run and deploy experience for businesses globally.

Today's AI applications will touch every aspect of our lives — including transportation, finance, retail, healthcare, smart manufacturing, education, and services industries. AI technologies will be at the forefront of digitally connected cars, smart manufacturing, and medical image recognition. To support edge inferencing, constrained footprints and often rugged environments, IBM could consider ruggedized offerings that integrate operational technology and customer technology functions. The inherent efficiency of HCI architecture, with combined compute/storage/networking in each node, could provide IBM the ability to scale down as well to meet edge inferencing requirements, as well as other space-constrained environments. In addition, to ensure that the offering is power and performance efficient to support varied workload requirements, IBM could look to unite heterogeneous pieces of computing power – incorporating discrete accelerators like FPGAs, GPUs, ASICs, and ASSPs.

IDC applauds IBM on offering a cloud-centric subscription-based licensing model for software and a flexible pay-as-you-go capacity model. As AI technologies mature and organizations build their trust with AI by relying on outcome-based-pricing, so that they can pay for results, not technology, IBM could explore the feasibility of offering outcome-based licensing models.



CONCLUSION



Every business wants to be more agile and accelerate time to market, thus it can be tempting to dive headlong into AI initiatives. However, the benefits of AI don't happen by magic. Instead, the benefits of AI are the result of strategic planning and a relentless commitment to data management.

Remember: Data is foundational to AI, and every step of the AI lifecycle is crucial. You cannot progress with AI or ML without data, so you must ensure that you understand and manage the lifecycle of that data. When your data is managed properly, AI can absolutely transform the abilities and possibilities for an organization.

But how can businesses improve data management and deploy AI initiatives when there is such a shortage of data scientists to help? In response to the current data science talent gap, we encourage businesses to explore AI-powered self-service data preparation and automated/assisted feature engineering that enable existing teams to start utilizing the power of data science today. These applications can either gather the data itself or seamlessly integrate with existing data systems. Having the right data is crucial for "feature generation," part of "feature engineering" i.e., creating and selecting the features or attributes to use in a predictive model. In addition, it is critical to embrace the practice and discipline of collaboration across data engineers, data architects, data scientists, AI app developers, and MLOps personas.

AI is emerging as a key business differentiator, running everywhere from edge to core to cloud. By exploiting the power of an AI-optimized and scalable, integrated solution, you can streamline workflows for your teams, support easy access to data, and help scale and ensure consistent performance along with microservices-oriented development in a secure and governed fashion. With an end-to-end solution that is open, interoperable, and standard — enabling easy portability of AI initiatives — your business will have the capabilities to unlock the true potential of AI.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC
ANALYZE THE FUTURE

Global Headquarters

5 Speen Street

Framingham, MA 01701

USA

508.872.8200

Twitter: @IDC

idc-community.com

www.idc.com

Copyright Notice