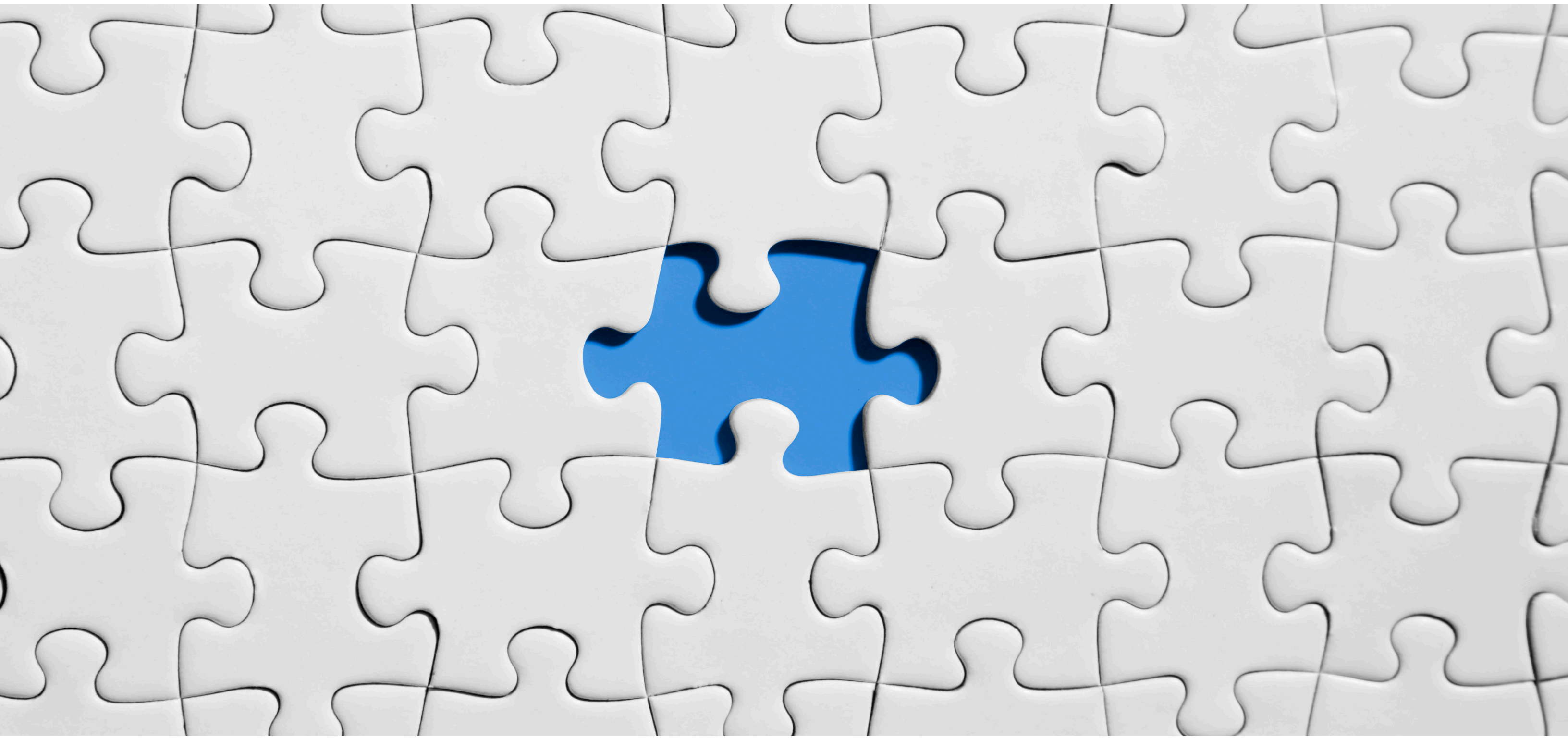
AI ROLE IN IT BEFORE AND DURING COVID-19



Mikhail Gloukhovtsev

Sr. Solutions Architect

Digital Solutions, Cloud & IoT

Orange Business Services

Mikhail.gloukhovtsev@orange.com

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the author. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

1. Introduction

Applied Artificial Intelligence (AI) transforms IT services (operations [AIOps], security, storage [intelligent IO-pattern-based performance optimization and data life cycle management], AI-based intelligent full-stack monitoring, and AI-powered chatbots for helpdesk services, etc.) by making them “smart”. In this article I consider how AI-driven IT service management creates a new paradigm in developing IT service strategy for companies and organizations. Integrating AI into IT services has become a strategic priority for IT service management, but it is a complex process and I review the challenges. I describe new solutions developed by Dell Technologies to make this transition easier and reduce complexity. I consider various technologies for building AI-powered infrastructure platforms and also discuss AI and Machine Learning as a Service (AIaaS/MLaaS), which is a good choice for some companies.

AI adoption in IT services is accelerated by digital transformation changing the business models of companies in various industry verticals. AI’s role in IT operations grows as digital transformation significantly increases the volume and variety of data generated by IT operations and needed to be analyzed so that IT services can be efficiently supported. The impact of the COVID-19 pandemic on AI adoption has been multifaceted. The growing economic fallout of the pandemic has forced some companies to reduce their budgets for new technologies, including AI applications to IT services. At the same time, COVID-19 has dramatically accelerated digital transformation and stimulated migration to cloud services, allowing companies to accommodate remote work quickly, run more workflows if data centers face staff reductions and enable them to scale-up quickly, in particular for COVID-19-affected workloads. Efforts on deployment of AI-based automation in IT services, which would require several years in the pre-COVID time, have taken months or even just weeks during the pandemic.

In the wake of COVID-19, cloud migration has become the key component of IT service strategy and I show that AI plays the major role in making cloud migration successful. AI enables efficiency of cloud-operations (CloudOps) by intelligently automating them and optimizing their cost. AIOps also helps in managing multi-cloud environments that become the key component of an enterprise cloud strategy.

As the move to remote working caused by COVID-19 has elevated the role of Edge Computing, I also consider the emerging concept of Edge AI. All in all, I hope my article will help readers develop AI-powered IT service strategies in the post-COVID-19 era.

2. A Bird's-eye View on AI Applications in IT

2.1 How Can AI Be Defined?

Various definitions for AI and Machine Learning (ML), which is an application of AI, fall in a broad range depending on their application area. Some of them focus on designing intelligent computer systems with features associated with human behavior intelligence. Other definitions of AI/ML emphasize abilities to make decisions autonomously based on ML models. AI can also be defined as the branch of computer science studying knowledge representations which are symbol-based rather than number-based and use heuristic methods for processing information. Of course, the AI definition changes to reflect AI development. Since this article considers AI applications for IT services, I prefer the AI and ML definitions given by Gartner: “AI applies advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automate decisions, and take actions” and “Advanced ML algorithms are composed of many technologies (such as deep learning, neural networks and natural language processing), used in unsupervised and supervised learning, that operate guided by lessons from existing information.”[1]

The main goals of AI can be defined[2] as:

(1) Creation of expert systems that exhibit intelligent behavior, learn, demonstrate, explain, and advise users.

(2) Implementation of human intelligence features in machines that can understand, think, learn, and behave like humans.

ML can be seen as the applied AI and it is based on the assumption that we can feed data to machines and they will learn on their own. In other words, ML is the way of getting computers to act without being explicitly programmed. ML is also referred to as predictive analytics because it is used for making predictions.

Aside from ML, there are various other approaches used to build AI systems. An example is evolutionary computation that is inspired from natural evolution and uses algorithms reflecting random mutations and combinations between generations to create optimal solutions. Another example is expert systems using rules to mimic behavior of a human expert in a specific domain.

2.2 How AI Shapes IT Evolution

While AI is at the core of technology and service deployment in every industry, it plays a paramount role in IT. Today AI/ML makes possible the development and deployment of IT systems and services that were impossible earlier. The very nature of AI and IT causes rapid evolution of these technologies. AI transforms IT services into intelligent ones with new functionalities that cannot be feasible without AI/ML. Before we consider the main applications of AI/ML in IT services and how their roles change during the COVID-19 pandemic, let us try to categorize the key areas of AI/ML applications in IT.

ML and Process Automation. In contrast to hardware-driven, robotic automation based on automating manual or other tasks, AI performs frequent, high-volume, computerized tasks. In this automation type, human inquiry is still essential to define the automation process and its goals. Integrating AI in the DevOps process reduces deployment time as developers need not wait until the last stage of deployment (Section 6). AI-enhanced DevOps-automated deployment process assures software quality by detecting and eliminating bugs during the processes of development.

IT Service Optimization by Adding Intelligence. AI/ML adds intelligence to existing or new products. As we see in this article, application of AI/ML to IT Ops results in creating new services areas such as AIOps (Section 6), ModelOps (Section 8), MLOps (Section 9), and even NoOps (Section 10).

AI Takes Software Development to the Next Level. In addition to low-code development (Section 16), AI might be able to automatically produce code – the feature referred to as "program synthesis." The main benefit of low-code AI software development is speed.

The importance of fast AI development has grown during the COVID-19 pandemic (Section 17.5). For example, the New York City Health Department wanted to track how COVID-19 had been spread through the city. For this purpose, they needed to build an online portal for collecting data on individuals with COVID-19 and others who were in contact with them. Using fast low-code AI development was critical and the COVID-19 Engagement Portal was up and running in just three days.[3]

Autonomous Compute. Autonomous Compute is self-deploying, self-provisioning, and self-healing infrastructure driven by AI. It is exemplified by the Dell PowerOne solution.[4] Its components in compute (PowerEdge MX), storage (PowerMax) and networking (PowerSwitch) are tied together via PowerOne Fabric. They are managed by an automation engine named the PowerOne Controller (Section 15).

More Accurate Data Analysis. We all know about the fast and almost exponential growth of the amount of data collected and stored. Analysis of the diverse large data sets is done by using Deep Learning (DL). DL is a subset of ML and it uses neural networks that are capable of parsing bigger data sets. Training DL models requires a lot of data because they can only learn from the data. The larger the data volume used in training, the more accurate DL models become. DL enables us to get more information, parsing more data than has ever been possible. It has been made possible by breakthroughs in ML algorithm development (Section 3.1) and new hardware AI-accelerators (Section 3.2).

Cybersecurity. There is no surprise that AI and ML have become critical technologies in information security because of their ability to quickly analyze millions of events and identify many different types of threats – from malware exploiting zero-day vulnerabilities to pinpointing abnormal behavior that may be associated with a phishing attack or downloading of malicious code (Section 12).

2.3 AI Architecture

There are various lists of AI key components and they differ in types of components and their numbers. To achieve artificial intelligence capabilities, AI systems combine and use mainly ML and other types of data analytics methods. Figure 1 illustrates most of the AI component domains. For this article, we assume that any AI system consists mainly of the following components[2]:

- Machine Learning
- AI programming language
- Knowledge representation
- Problem-solving (mainly by heuristic search)
- AI hardware

A typical AI technology stack is presented in Table 1.

Training and Inferencing. Training and inferencing are the principal tasks of AI.[2] The training is a data- and compute-intensive process that prepares AI models for production applications. In the training phase, a known data set goes through an untrained artificial neural network. Artificial neural networks are created to simulate the human brain with neuron nodes interconnected like a web. The more layers of neurons in the neural network, the deeper the network is. A DL algorithm (DL was discussed in Section 2.2) uses a neural network to solve a particular problem.

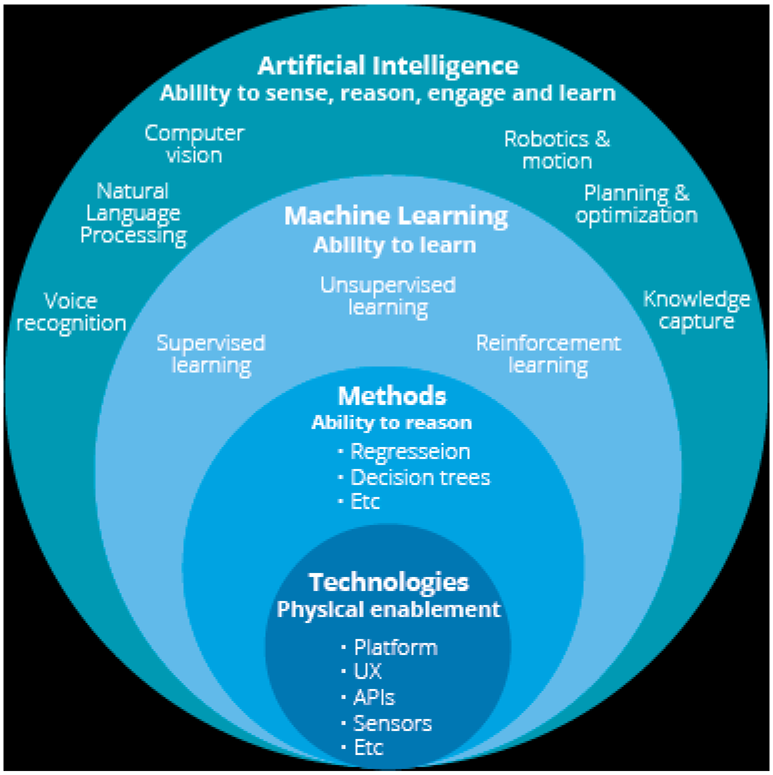


Figure 1. Main AI Component Domains (Ref.2).

Inferencing is the term describing use of a trained neural network to generate insights.[5] The inferencing phase that takes place after the training phase requires only a fraction of the processing power needed for the training (see Fig. 2). Therefore, instead of the powerful processing system required for training the neural network, significantly fewer compute resources are needed for an inference server whose only purpose is to execute a trained AI model.

Technology	Stack	Definition
Services	Solution and use case	Integrated solutions that include training data, models, hardware, and other components (e.g. voice-recognition systems)
Training	Data types	Data presented to AI systems for analysis
Platform	Methods	Techniques for optimizing weights given to model inputs
	Architecture	Structured approach to extract features from data (e.g. convolutional or recurrent neural networks)
	Algorithm	A set of rules that gradually modifies the weights given to certain model inputs within the neural network during training to optimize inference
	Framework	Software packages to define architectures and invoke algorithms on the hardware through the interface
Interface	Interface systems	Systems within the framework that determine and facilitate communication pathways between software and underlying hardware
Hardware	Head node	Hardware unit that orchestrates and coordinates computations among accelerators
	Accelerator	Silicon chip designed to perform highly parallel operations required by AI; also enables simultaneous computations

Table 1. AI Technology Stack (Adapted from Ref. 6)

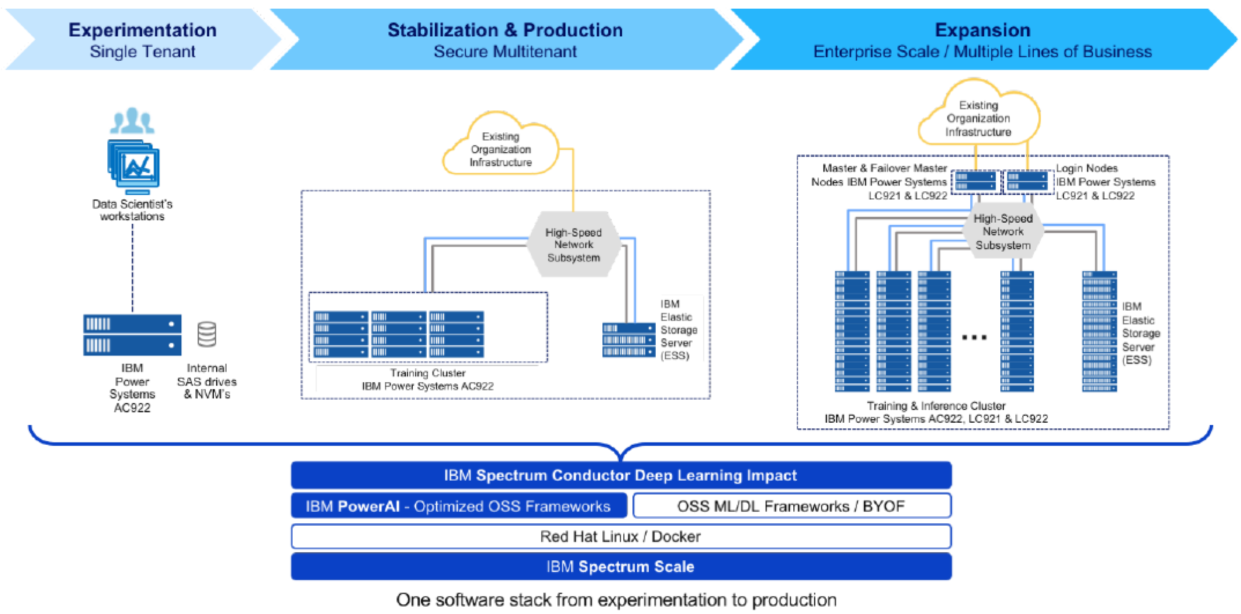


Figure 2. IBM AI Infrastructure Reference Architecture (Ref. 7)

AI Infrastructure Reference Architecture. IBM suggests the AI Infrastructure Reference Architecture with a single software stack from experimentation to production (Fig. 2).[7] The AI infrastructure architecture includes GPU-accelerated servers having 2 to 3 times higher throughput and performance than commodity servers (for graphics processing unit (GPU) overview, see Section 3.2.1), storage resource connectors, and Elastic Distributed Training to dynamically assign GPUs to models. The ability to add and remove GPUs without needing to stop the training enables GPU sharing and provides resiliency to failures. Multitenancy can support multiple teams of data scientists, frameworks and applications on a common shared compute cluster.

3. Quantum Leap in AI: Massively Parallel Methods and AI Accelerators

There are many excellent overviews of AI history and I refer readers to Refs. 8-10. They vary in the date of the birth of AI technology. Some AI history studies that define AI as any machine capable of performing a task that would typically require human intelligence even claim that evidence of AI can be found in ancient Egypt and Greece.[9,10] While Talos, who would defend Crete from invaders, can be considered as a "automaton" being that exemplifies primitive forms of artificial intelligence,[9] we will focus our attention on much more recent events that can be considered as the AI revolution.

There have been several milestones in AI development in this century. The AI greatest successes in the early 21st century are results of two breakthrough developments: (1) massively parallel methods for ML algorithms, enabling them to process large data sets and (2) AI hardware and software accelerators.

3.1 Massively Parallel Methods

The seminal paper published by Andrew Ng's group at Stanford University in 2008 presented general principles for massively parallelizing unsupervised learning tasks using GPUs.[11] The authors have shown that these principles can be applied to scaling up learning algorithms for both Deep Belief Networks (DBNs) and sparse coding. Their implementation of DBN learning is up to 70 times faster than a dual-core CPU implementation. This allowed the research team to reduce the time required to learn a four-layer DBN with 100 million free parameters from several weeks to around a single day. This work has brought practicality into DL studies by making training on huge volumes of data efficient.

The use of cross-GPU parallelization in the AlexNet DL model, which is an 8-layer convolutional neural network (CNN) designed by Alex Krizhevsky, resulted in winning Imagenet's image classification contest with accuracy of 84% in 2012. This paper[12] was like a seismic shift in the DL study landscape as the accuracy of earlier models (74%) trailed far behind. In the following years, this solution allowed several teams to build CNN architectures capable of exceeding human level accuracy.

3.2 AI Accelerators

An AI accelerator is a specialized hardware accelerator based on so-called AI chips or software created to accelerate artificial intelligence apps. We start our review with hardware AI accelerators including graphics processing units (GPUs), field-programmable gate arrays (FPGAs), certain types of application-specific integrated circuits (ASICs), vision processing unit (VPU), and tensor processing unit (TPU).

Different AI tasks require different types of AI accelerators. For example, GPUs are most often used for training of AI models. Inferencing phase (see Section 2.3) is mostly done by using FPGAs. ASICs, which are often termed an SoC (system-on-chip), can be designed for either AI training or inference phases. Because of their unique features, AI accelerators are tens or even thousands of times faster and more efficient than CPUs for training and inference of AI algorithms.[13] The state-of-the-art AI accelerators are also dramatically more cost-effective than CPUs.

3.2.1 Graphics Processing Units

As discussed in Section 3.1, Andrew Ng and his team determined[12] in 2008 that GPUs executing a large number of operations in parallel rather than sequentially can increase the speed of DL processes by 70 times. GPUs are efficient in the matrix/vector computations in ML. Originally GPUs were designed for image-processing applications, mainly in gaming, that benefited from parallel computation. In 2012, the AlexNet model (Section 3.1), based on cross-GPU parallelization, accelerated the use of GPUs for training AI systems, and GPUs became dominant in AI applications by 2017.[13] GPUs are also sometimes used for inference.

GPUs are optimized for training AI and DL models as they can use thousands of cores to process tasks in parallel. Additionally, computations in DL use very large amounts of data and they benefit from using GPUs' memory bandwidth. Nvidia, which is the leader in the GPU market, recently announced a new GPU - A100 Tensor Core GPU.[14] A100, based on the NVIDIA Ampere Architecture, provides up to 20 times higher performance compared with the previous generation GPU (Volta). Multi-instance GPU technology enables partitioning A100 into seven GPU instances to dynamically adapt to workload changes. A100 GPU with 80 GB memory capacity can deliver the memory bandwidth of 2 TB/s so that it can run very large ML models. NVIDIA Tesla GPUs are in the Dell PowerEdge C4140 servers that are compute nodes in the Dell Ready Solutions for AI - Deep Learning presented in Figure 3. A high-level overview of the Dell Ready Solutions for AI is shown in Fig. 4.

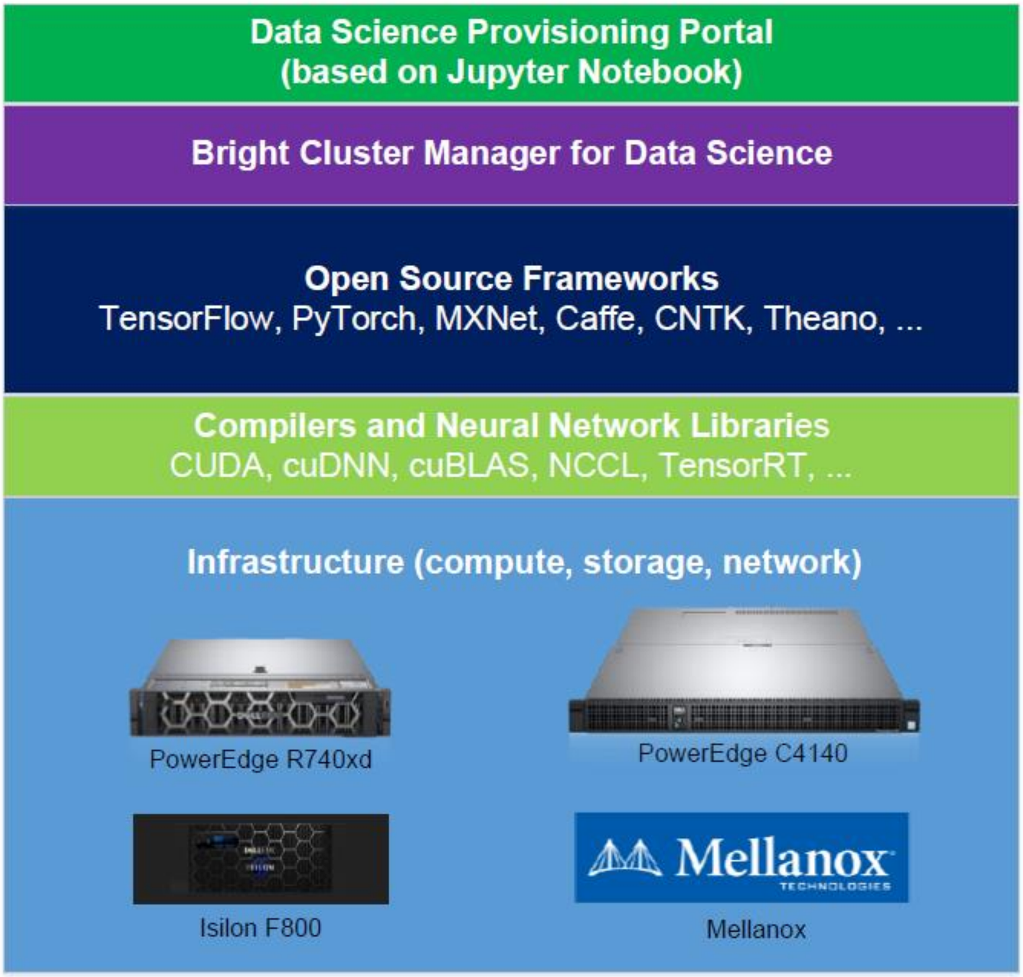


Figure 3. Overview of Dell EMC Ready Solutions for AI - Deep Learning (Ref.15)

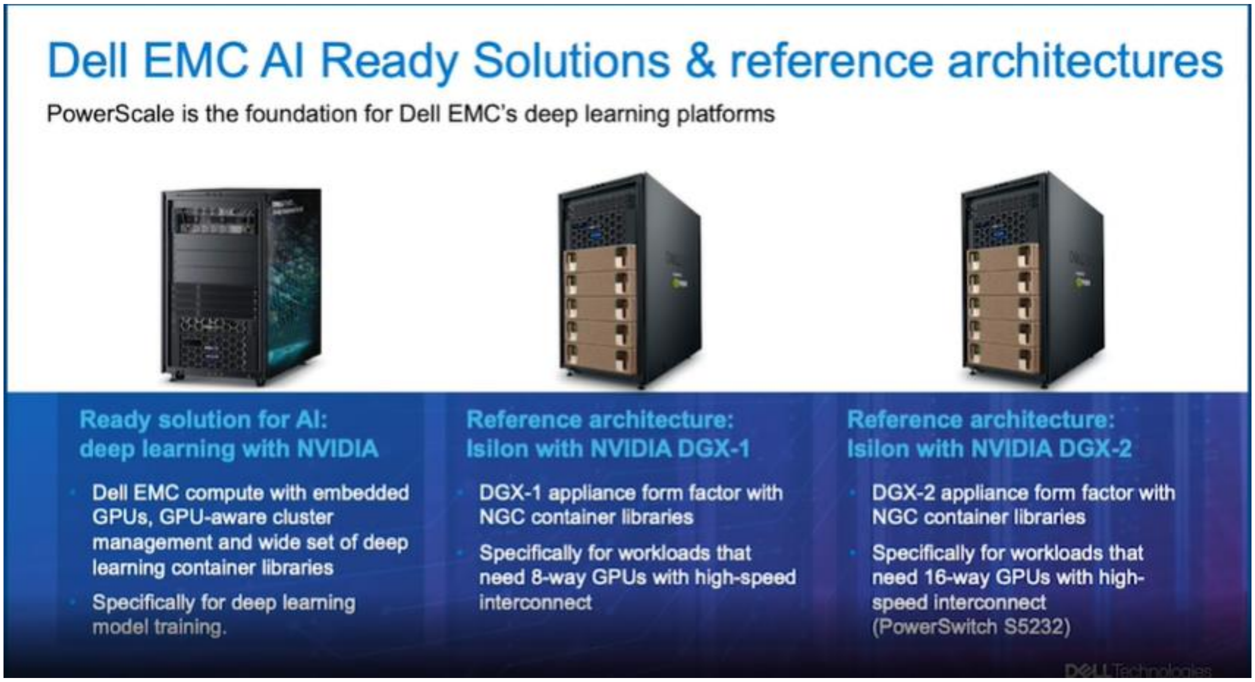


Figure 4. Dell EMC Ready Solutions for AI and References Architectures (Ref. 16)

3.2.2 Field-Programmable Gate Array (FPGA)

Field programmable gate arrays (FPGAs) can be described as integrated circuits with a programmable hardware fabric. FPGAs are called "field-programmable" because they can be configured after manufacturing, unlike GPUs or ASICs. This capability gives FPGAs several advantages over ASICs (Section 3.2.4), which have hardwired circuitry customized to specific algorithms and require a long development time as well as a significant investment for design and fabrication. Hence, FPGAs offer a combination of speed, programmability, and flexibility without the cost and complexity of developing custom ASICs. These FPGA features enable designers to test algorithms quickly and get to market fast.[17]

FPGAs can deliver better performance than GPUs in DL applications for which low latency is critical. By using FPGAs to accelerate search ranking, Microsoft's Bing search engine realized a 10-fold reduction in the latency produced by Bing's intelligent search models.[18] Some modern FPGAs integrate programmable logic with hard processor cores into a single device called a System on Chip (SoC) for increased computing performance. This hybrid architecture combines the ease of programming a processor along with the flexibility and performance of a programmable logic fabric. Compared to using a stand-alone processor and a stand-alone FPGA, SoC FPGA is cheaper, consumes less power, and is easier to design.

3.2.3 Vision Processing Unit (VPU)

A vision processing unit (VPU) is a specialized processor that is made to function as an AI accelerator by supporting tasks like image processing.[19] VPUs are built for parallel processing and are used on low-power platforms for object and facial recognition, security access and ML applications in areas such as visual retail, security, and safety. VPU technology enables intelligent cameras and AI appliances with DL network, as well as computer vision-based applications.

3.2.4 Application-Specific Integrated Circuit (ASIC)

ASICs as AI hardware accelerators use lower precision arithmetic to accelerate calculations. Leading ASICs typically provide greater efficiency than FPGAs, although FPGAs are more customizable than ASICs, as we discussed in Section 3.2.2. ASICs have various advantages for AI acceleration and speed is the main one. For example, Intel offers Nervana Neural Network Processors (NPP) (NNP-T1000 for ML training and NNP-I1000 for inference) that provide a significant parallelization of AI tasks.[20]

3.2.5 Tensor Processing Unit (TPU)

A tensor processing unit (TPU), also called a TensorFlow processing unit, is an AI accelerator ASIC developed by Google for ML in 2016.[21] The Google TPUs are available in two types: cloud TPU and edge TPU. Cloud TPUs forming TPU Pods in Google data centers can be accessed from a Google Colab notebook. Each TPU core comprises scalar, vector, and matrix units. Multiple TPUs in a TPU Pod configuration are connected through a dedicated high-speed network to create a larger pool of TPU cores and TPU memory for running ML workloads. An edge TPU is an ASIC designed by Google for high performance ML inferencing on low-power edge devices. Edge TPUs combining custom hardware, open source software, and AI algorithms are used for AI solution deployment at the edge.[22]

3.2.6 AI Accelerator Benchmarks

There are various benchmarks focusing on different aspects of AI accelerator performance. It is well known that AI chips typically provide a 10-fold to 1,000-fold improvement in efficiency and speed relative to CPUs, with GPUs and FPGAs on the lower end of this range and ASICs at the higher end (Sections 3.2.1-3.2.3). Better performance for AI accelerators is not always measured in speed because in some cases the key metric is lower power consumption or heat generation. Some vendors use TOPS/W (Tera-operations per second if hypothetically one Watt is consumed) as a benchmark for AI inference chips.[23] With growing AI adoption, AI applications and environments in which they run become very diverse.[24] As a result, a single metric like TOPS/W may not be able to serve as a complete benchmark.

3.2.7 GPU Virtualization

GPU virtualization enables GPU acceleration in virtual machines (VMs). For example, VMware vSphere uses DirectPath I/O for virtualizing GPUs. It bypasses the ESXi hypervisor and provides direct guest OS access to a GPU. DirectPath I/O allows a VM to utilize one or more full GPUs for ML and HPC workloads.[25] NVIDIA vGPU enables sharing of vGPUs across multiple virtual machines.[26] NVIDIA vGPU-sharing allows for multiple VMs to access a single physical GPU. GPU virtualization by using VMware vSphere Bitfusion technology[27] enables Dell Technologies to bring GPU-as-a-Service (GPUaaS)[28] to the AI accelerator market as part of the Dell EMC Ready Solutions for AI.

3.2.8 SW AI Accelerators

AI software accelerators are offered by some vendors. For example, the Intel distribution of OpenVINO Toolkit (Open Visual Inference and Neural network Optimization) allows developers to implement DL inference solutions quickly and efficiently across Intel architectures and AI accelerators.[29] The toolkit also enables CNN-based DL (Section 3.1) inference to run on the edge.

DL startup DeepCube launched the only software-based AI inference accelerator that improves AI performance on intelligent edge devices.[30] DeepCube's proprietary framework, which can be deployed on top of any existing hardware (CPU, GPU, ASIC) in both datacenters and edge devices, enables over 10-fold speed improvement.

3.3 AI OS

AI-based Operating System like classical OS provides software and hardware management and common OS services. Intelligence in management of the system is what differentiates AI OS.[31] Various AI techniques such as perceptive intelligence, context-specific search, context priming, etc., enable AI OS to efficiently use parallelization of processes and improve memory management and security. AI OS has the capability to learn and evolve like Element AI OS, described as a cognitive operating system providing enterprise applications such as risk management and data compliance.[32] Another example is the aiOS, an AI-based OS for SD-WLANs.[33] The aiOS implements ML models for user-adaptive frame-length selection in SD-WLANs and can improve the aggregated network throughput by up to 55%. According to some opinions in the industry, OSes based on AI represent the future of operating systems.[34]

4. Cloud and Edge Computing Solutions in AI/ML

As discussed in Section 17.9, the COVID-19 pandemic has caused many organizations to accelerate their migrations to public cloud solutions. AIaaS offerings by Amazon, Microsoft, and Google have grown significantly during the pandemic.[35-37] As cloud computing evolves into cloud-edge model and then into distributed cloud, let us take a quick look at which environments suit better for AI training and inference.

Generally, DL training that is a compute-intensive process is carried out in the cloud. Indeed, the cloud is an ideal location for ML/DL training that benefits from cloud services providing large data stores, multiple servers or serverless platforms. Furthermore, the cloud is cost effective because it allows GPUs and other expensive AI accelerator hardware to train multiple AI models.

AI algorithms for inference work with less data but must generate responses quickly. Inference non-time-critical workloads can also be done in the cloud. However, to reduce the latency, more commonly inference is carried out on a device where the analyzed data resides. Indeed, a self-driving car does not have time to send images to the cloud for processing and get the results back when it detects an object in the road. In addition, sending AI data to the cloud requires addressing privacy and security issues. Hence, the edge is the best choice for inference. The AI/ML future is distributed intelligence using both local devices and cloud services.

5. Neuromorphic Computing: The Next-Level Artificial Intelligence

Many general-purpose computers today are based on the von Neumann architecture, separating memory and processing. In the von Neumann architecture, the speed of moving data between the memory and CPU is limited by the bus throughput and this causes the so-called von Neumann bottleneck. This bottleneck can be potentially overcome in a neuromorphic architecture that mimics the central nervous system's information processing architecture.[38] The computational building blocks in neuromorphic systems are logically analogs of neurons. While von Neumann computing systems are largely serial, neuromorphic systems try to emulate brains and use massively parallel computing. Developers of neuromorphic systems also want to model better fault-tolerance of brains. The greatest driving force for investments in neuromorphic system development is the promise these systems hold for AI. Neuromorphic systems are also likely to help in creating new types of AI applications as they are more comfortable with ambiguity[38] whereas the current generation AI systems are mostly rules-based and have difficulties with noisy and uncertain data.

Neuromorphic computing chips such as Intel's Loihi,[40] IBM's TrueNorth,[41] and Qualcomm's Zeroth[42] processors are compact, portable, and energy efficient. They do not replace GPUs, FPGAs, ASICs, and other AI-accelerators chip architectures. Instead neuromorphic computing AI accelerators will probably be used for specialized event-driven workloads in which asynchronous spiking neural networks that are the key component of many neuromorphic architectures have an advantage.

6. Artificial Intelligence for IT Operations (AIOps)

6.1 What Is AIOps?

We will consider how AIOps (Artificial Intelligence for IT Operations) transforms IT operations. Gartner, who suggested the term AIOps, states that[43] "AIOps combines big data and machine learning to automate IT operations processes, including event correlation, anomaly detection and causality determination." AIOps can be described in detail as the application of ML and data science to IT Operations. AIOps solutions combine Big Data and ML functionality to directly and indirectly enhance primary IT operations functions (monitoring, automation and IT service management [ITSM]): (1) availability and performance monitoring; (2) event correlation and analysis; (3) ITSM and automation, as shown in Fig. 5. The three domains comprising AIOps – Monitoring (Observer), Engage (ITSM), and Act (Automation) – are presented in Fig. 5. The AIOps workflow starts with collecting and curating large amounts of data by using Big Data tools. Then this data is fed into analytical models for analysis and event correlation to deliver proactive and predictive insights into problems. Finally, AIOps provides recommendations for actions and action automation. Hence, AIOps solutions that are service-centric and provide actionable intelligence enable organizations to transform their IT operations and, as a result, improve the ROI for their IT infrastructure investments.

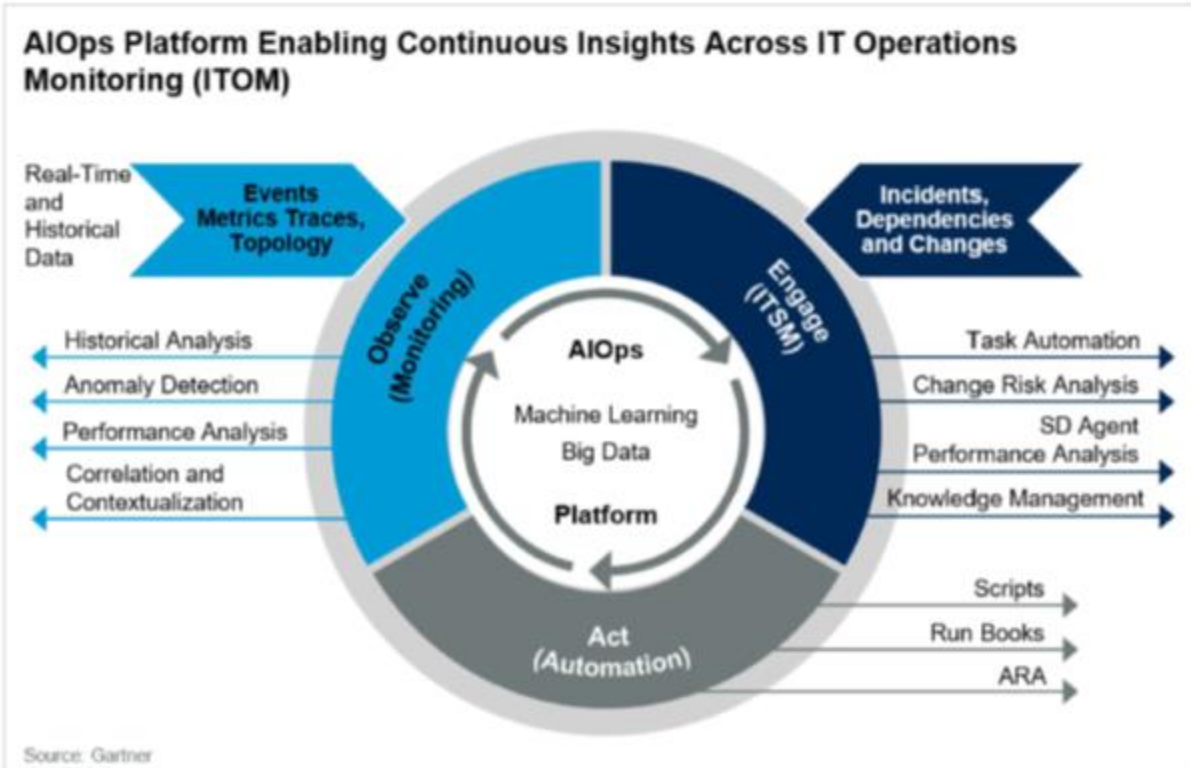


Figure 5. Gartner's Visualization of the AIOps Platform (Ref.44)

6.2 AIOps Taxonomy

There are several classifications of AIOps that depend on what IT service functions are considered as the principal ones. For example, IBM classifies AIOps into Simplified AIOps, Dynamic AIOPs, and Proactive AIOps.[45]

Simplified AIOps provides automated noise reduction and automated incident remediation. ML reduces manual effort by automating many IT Ops decisions. Alert noise can be filtered out by using advanced seasonal behavior analysis. Simplified AIOps automatically analyzes patterns in operational data to identify automation opportunities.

Dynamic AIOps gives insights for probable cause identification. It helps with rapid problem resolution and reduction of mean time to repair (MTTR).

Proactive AIOps facilitates incident prevention by using ML for proactive management of apps and infrastructure. As a result, operations teams can take corrective actions to avoid service outages.

Gartner distinguishes AIOps capabilities into the following categories[44]:

Domain-agnostic AIOps is a general-purpose AIOps platform relying mostly on monitoring tools for data collection and being applied for the broadest use cases.

Domain-centric AIOps has a relatively narrow set of use cases. The vendors of domain-centric AIOps zero in on a single domain (for example, APM, network, endpoint security). While domain-centric tools provide a deep view into a specific domain, they lack the ability to generate a correlated, end-to-end view across domains. Some vendors combine the domains into hybrid-sources domain-centric solutions.

Do-it-yourself (DIY) AIOps is built by mixing and matching the components from multiple providers and open-source projects. Building an AIOps platform is not a simple project, as developing the solution stack requires integration of various layers such as data collection and streaming, data aggregation, visualization tools, etc. Consequently, DIY AIOps deployments are limited by the availability of in-house skills to build and support them.

6.2 AIOps Solutions Architecture

The AIOps architecture amalgamates ML and Big Data technologies to bring IT Ops to the next level. The AIOps architecture should be able to work with multiple monitoring tools and analyze the large data volume and wide range of various data types that are generated by the infrastructure and application monitoring. The AIOps reference architecture developed by IBM[46] is shown in Fig. 6. The leading vendors of AIOps platforms are listed in Ref.44.

As seen from Fig. 6, the AIOps workflow is a multistep process. In Step 1, the IT Ops processes (for example, service requests, incidents, problem changes, and configuration changes) serve as valuable data sources. At the same time, these processes can benefit from getting back the insights provided by integrating the AI/ML outputs into each process. In the next step, data from as many data sources as possible are streamed in real-time into a data lake. The larger the number of data sources, the more accurate the AI models are. The data is curated and organized in Step 3 to mitigate against the problem of “Garbage in, garbage out.”

Structured data is placed into a data mart in Step 4 and IT-related dashboards can be created. In Step 5, structured and unstructured data in the data lake is used to build and test different DL and AI models. The AI core services provide a framework to initiate corrective and preemptive actions.

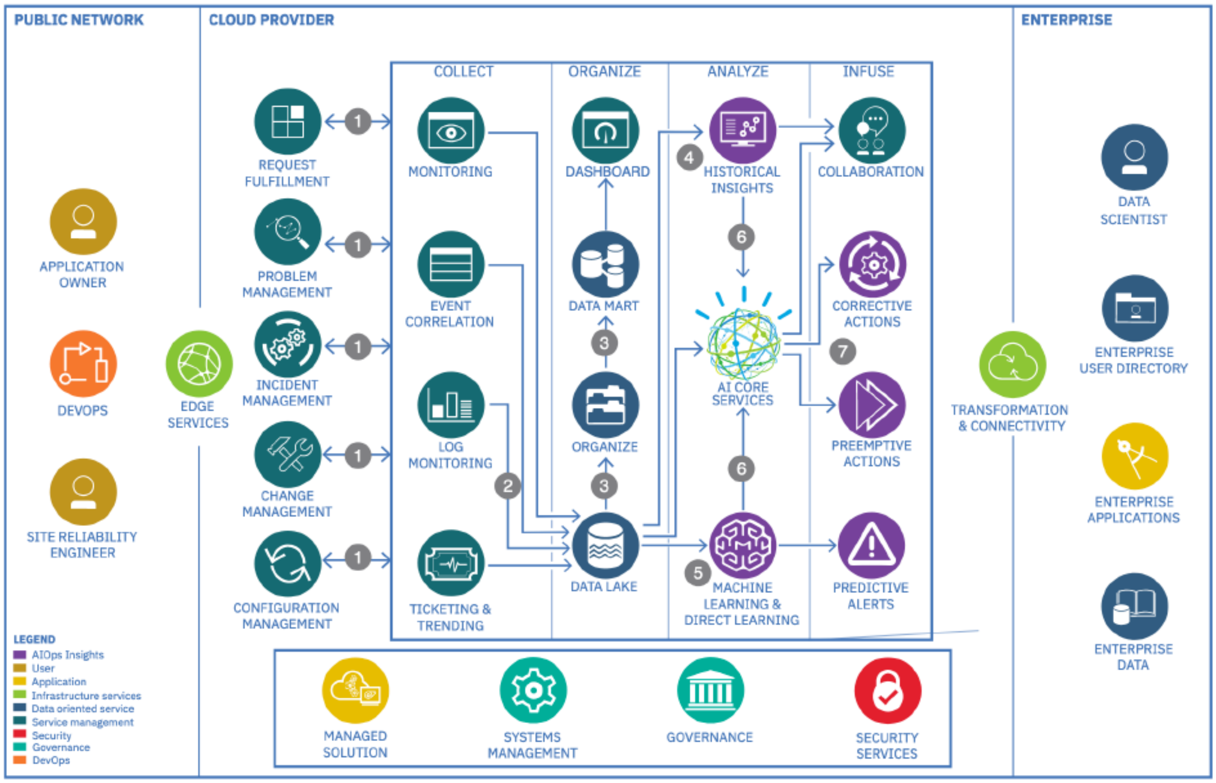


Figure 6. AIOps Operations Reference Architecture (Ref.46)

As seen in Fig. 5, there are three main AIOps domains: (1) Observe (Monitoring); (2) Engage (ITSM), and (3) Act (Automation). Let us take a quick look at them.

6.3 AIOps Solutions for Monitoring and Observability

Modern IT services are based on multicloud environments. To meet business agility requirements, these environments are dynamic and scalable. Consequently, they are also complex and change frequently. As a result, traditional monitoring of these environments is no longer effective. Just displaying data on dashboards without providing event-correlation, root cause analysis automated by using analytics, pattern-based predictions and proactive actionable recommendations is not enough. Traditional monitoring tools require manual configurations and the manual efforts do not scale in multi-cloud environments.

AIOps comes to the stage because AIOps analysis reveals anomalies, shows topology-based microservice relationships, and uses contextual intelligence to provide a complete operational view for service assurance. In other words, AIOps enables observability characterized by Gartner[47] as the evolution of traditional monitoring solutions to meet the demands of cloud-native technologies. In contrast to conventional approaches to monitoring, observability means looking at the complete service stack. The combination of AIOps and observability is of great value to DevOps teams since it can automate detection, diagnosis and remediation of problems with the speed and accuracy required in their CI/CD environments. Observability does not substitute for monitoring. On the contrary, they are complementary, and observability with AIOps automates monitoring.

6.4 AIOps Solutions for ITSM

AIOps orchestrates key IT service management (ITSM) activities ("Engage", Fig. 5), such as asset management, change management, and incident management. Previously these activities were manual and they cannot be efficient after the transition to agile multi-cloud environments. AIOps enables more efficient change management. The use of an AI-backed change risk prediction engine can streamline the change management process by minimizing the need for approvals as recommended by ITIL4 framework.[48] Indeed, ITIL4 recognizes that not all changes need Change Advisory Board (CAB) approval. AIOps can accurately flag changes that require a CAB review so that business disruption risks are minimized. Certain actions can be taken without human intervention as AI technology permits changes to be made automatically to avoid predicted service issues.

IT help desk tasks that traditionally have been high-volume but low-value ITSM functions can be automated by an AI-enabled ITSM solution. The AIOps solution will review previous help desk ticket history to determine how a newly submitted request will be managed. Emerging AI Service Management (AISM) solutions bring ITSM to the next level by delivering scalable, proactive, and secure IT services.[49]

6.5 AIOps Solutions for IT Service Automation

AIOps takes IT Ops management one step further towards automated self-driving IT operations management (ITOM).[44] AI-based automation (Fig. 5, "Act") of problem triage and remediation is an example (Fig. 7). AI tools help in IT workflow automation and provide more sophisticated automation capabilities in orchestration systems. For example, AIOps-driven automation enables realization of the full benefits of SDN and NFV.[50]

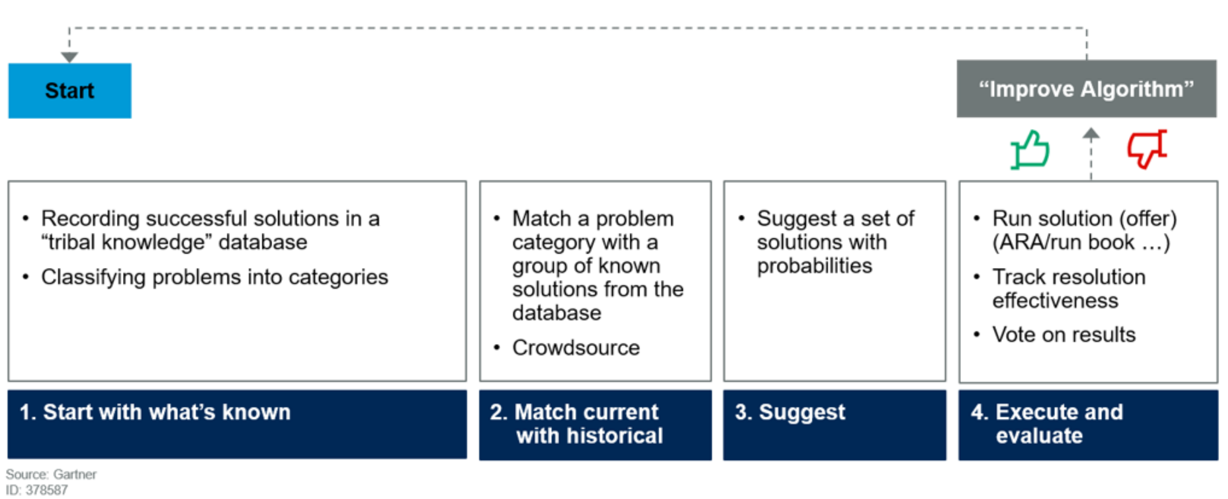


Figure 7. AI-Assisted Automation: Problem Triage and Remediation (Ref.50).

7. AIOps-driven Transformation from DevOps to AI DevOps

One of the key applications of AI to IT Ops is the adaptation of AI methods and technologies to DevOps. This enhancement to DevOps is called AI DevOps. Similarly, AIDev is considered as the adaptation of AI for application development processes. Thus, there are intersections of AI DevOps with AIOps and AIDev in the areas where including AI in operations and development is relevant for DevOps purposes (Fig. 8).

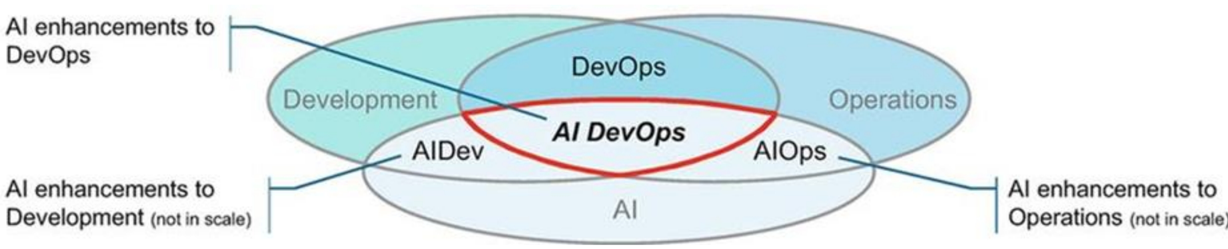


Figure 8. Relationship of AI DevOps with AIOps and AIDev (Ref.51).

The AI DevOps model (Fig. 9) can simplify, accelerate, and automate the DevOps lifecycle by using AI and data science methods. The enhancement of DevOps by AI enables risk assessments in code and in builds as part of application release automation and detection of potential security issues. AI DevOps can be leveraged to improve both the development and operations domains of DevOps. For example, the development team can use AI for segmentation and improved automation of test cases whereas the IT Ops team can apply AI for baseline determination and anomaly detection in software subsystems.

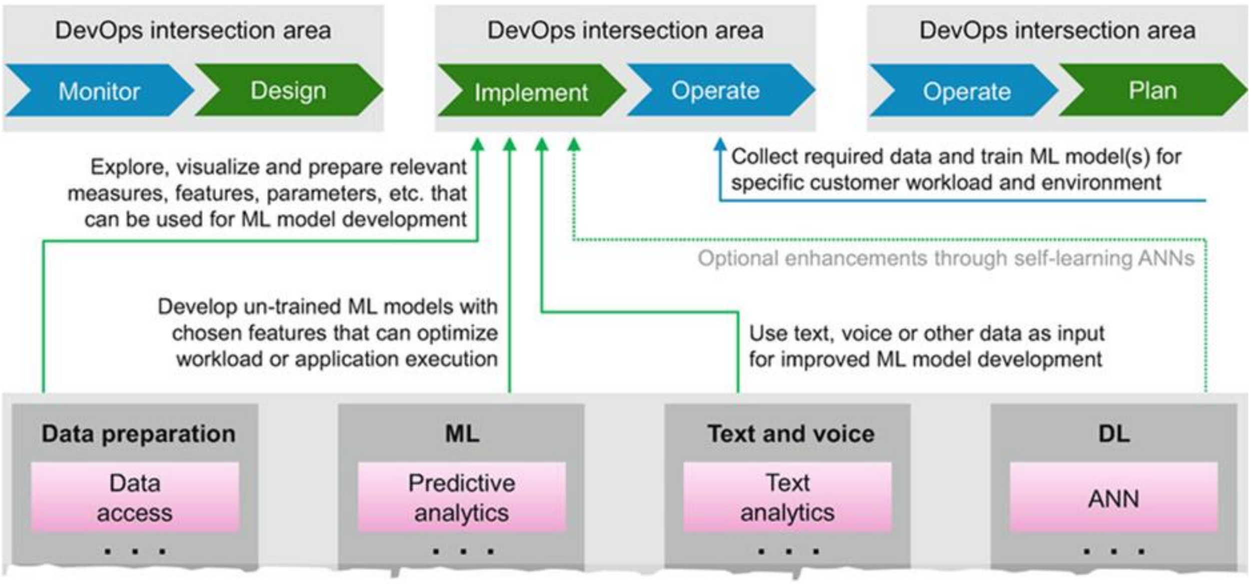


Figure 9. AI DevOps Model (Ref.51).

8. ModelOps

Use of analytics models plays a critical role in applying AIOps. As 50% of models never make it into production,[52] IT organizations have understood that a special operational model is needed. This has resulted in establishing ModelOps (Model Operations), a holistic approach to automate deployment to production, monitoring, governance and continuous improvement of data analytics models. ModelOps can be seen as a DevOps variation.[53] Indeed, while DevOps focuses on application development, ModelOps is designed for deployment of analytics models.[54] ModelOps, which is seen as AI model operationalization, is crucial for continuous delivery, efficient development, and deployment of AI models.

Increasing use of multi-cloud IT service environments has led to development of multi-cloud ModelOps to manage the end-to-end lifecycles of models and applications across clouds.[55] Accelerating AI and ML adoption speeds up the use of analytics models and, as a result, the ModelOps role in moving models through the analytics lifecycle becomes more important. By providing effective model lifecycle management, ModelOps can help companies maximize AI benefits for the business.

9. MLOps

While ModelOps primarily provides the governance and lifecycle management of all AI and decision making models (Section 8), MLOps (Machine Learning Operations), a subset of ModelOps, focuses on the operationalization of ML models only.[56] Like the DevOps or DataOps methodologies, the goal of MLOps is to enhance effective collaboration and communication between data scientists and Ops teams for the management of production ML/DL lifecycle. MLOps bridges the gap between ML applications and the CI/CD pipelines by adding the discipline and efficiency of DevOps into AI/ML practices.[57] As seen in Fig. 10, MLOps is at the confluence of ML, data engineering, and DevOps, three tightly interwoven components of MLOps.

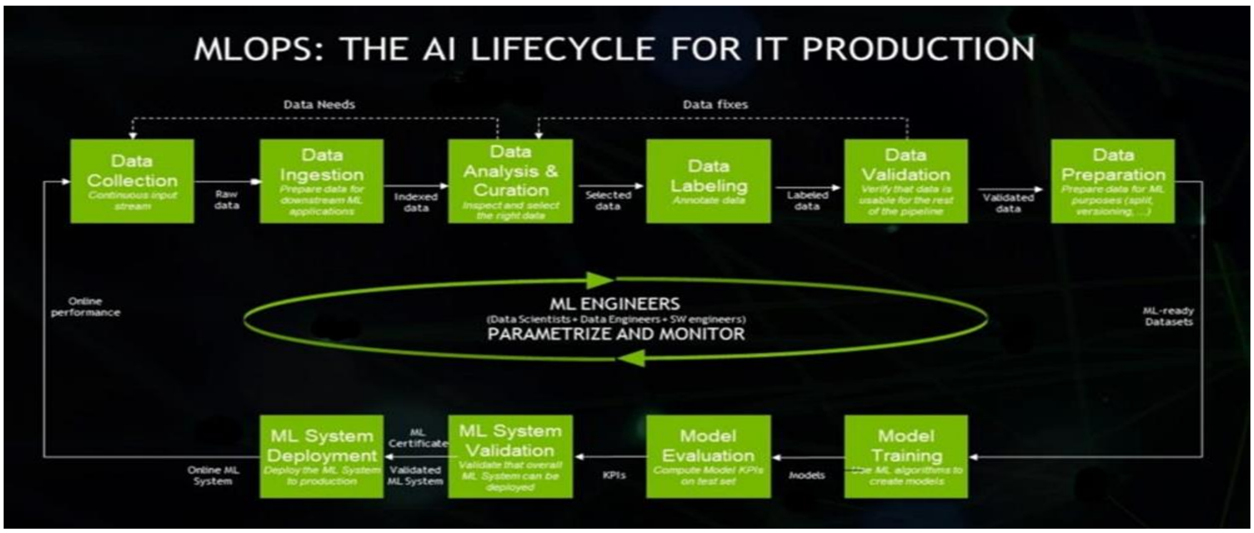


Figure 10. MLOps Governance of AI/ML Lifecycle (Ref.58).

MLOps manages the entire lifecycle of ML, from data collection and ingestion, data validation and preparation, model development, training, evaluation, validation, and deployment (Fig. 10). Hence, MLOps enables ML technologies to generate business value by swiftly and reliably deploying ML technology into production.

10. NoOps

The growing adoption of AIOps contributes to the morphing of DevOps into the NoOps methodology. While in DevOps the development team and Ops team work together through the entire application lifecycle, in the NoOps scenario, these teams never need to interact to get their jobs done. The goal of the NoOps concept is to automate the operational environment to a kind of hyper-automation that no Ops team is needed to manage.[59]

NoOps is different from outsourcing IT operations to managed IT service providers. Powered by the convergence of sophisticated ML and hyper-automation of cloud computing, NoOps envisions a fully automated IT environment that is entirely abstracted from the underlying hardware infrastructure. In the NoOps environment the developer commits the code into the repository and every IT system component required by the application to run will be deployed with no involvement of the Ops team. Delivering infrastructure-as-code enables NoOps to improve service resilience as it better controls the infrastructure configurations and application changes.

Function-as-a-Service (FaaS). NoOps implementation is based on Function-as-a-Service (FaaS) solutions, which are serverless computing services, for example AWS Lambda and Azure Functions. FaaS provides automated scaling for ML applications and enables them to effectively use microservice-based design.

Heuristic automation. A new approach to automation with the goal of developing hyper-automation for NoOps incorporates "heuristic learning".[60] The idea behind heuristic learning is to build analytical models, producing a solution for complex optimization problems quickly although the solution may be not the best one. This will allow intelligent heuristic automation systems not only to automate incidents, changes, or releases, but also to be able to apply knowledge-based automation to more complex decision-making problems such as capacity management or service continuity management without human intervention.

NoDev NoOps NoIT. As IT Ops methodologies keep moving in the direction of ever greater automation, naturally the question arises: what will NoOps evolve into? There are discussions about Intelligent Ops (IntOps) – the concept of running applications by AI-trained machines.[61] The other polemic principles are "NoDev NoOps NoIT," also called “KnowIT” principles.[62]

11. CloudOps

Adopting AI-based cloud operations (AIOps for CloudOps) in the era of the “Cloud First” IT strategies provides an automated way for IT teams to have better analytics, event reduction, and faster decision-making capabilities to resolve the anomalies and root cause events.[63,64] While providing service agility and flexibility, cloud operations bring some challenges, including security compliance issues, alert fatigues, and analytics-driven monitoring requirements. AI-powered CloudOps enables IT teams to address these cloud management challenges. It helps in automating cloud service request fulfilment, incident remediation, and cloud service provisioning. CloudOps also provides service improvement as required by the ITIL framework. Indeed, AI-based monitoring and management systems are self-learning: the longer they run, the better they function. The automation of cloud-native application deployment, the use of cloud infrastructure as code (IaC), and patch management automation are possible with the AI-based CloudOps tools.

12. Applying AI to Cybersecurity

The complexity and dynamics of modern hybrid-cloud and multi-cloud environments present many challenges to overwhelmed IT security teams. In some cases, it is just not possible for humans to respond adequately to a large-scale dynamic enterprise attack. AI/ML can be used by IT security teams to shrink the attack surface instead of constantly chasing malicious activity.[64] AI-based security systems provide much needed analysis and threat identification so that IT security teams can act to reduce breach risk and improve security posture. AI can identify and prioritize risk, instantly detect any malware on a network, and guide incident response.

While IT security techniques that depend on signatures can detect only about 90% of threats, AI-based network security solutions do not rely on signatures and can identify zero-day attacks. ML methods combined with threat intelligence can detect patterns of malicious behavior and discover sophisticated threat actors. AI platforms automate SecOps tasks to accelerate security threat analysis and predict how and where a security breach can happen. This enables security teams to respond proactively rather than reactively.

13. AI in Network Management

In traditional networks, incident management is post factum. In contrast, AI-driven networks use real-time predictive analytics and allow network administrators to identify issues proactively. Such networks are also capable of dynamically optimizing the processes with a feedback loop powered by AI. For example, Intent-Based Networking (IBN) uses a combination of AI, ML, and network orchestration to provide secure optimization and automation of network processes.[65] Fig. 11 shows the four key functional blocks of IBN and how AI provides the feedback loop. AI can accelerate the path from Intent into Translation and Activation (Fig. 11). Then AI examines network and behavior data in the Assurance step. Activation uses the AI-provided insights to create more intelligent actions for improved performance, reliability, and security. AI-powered IBN is a self-healing and self-improving network. Network administrators of IBNs can focus on the overall intent of the network rather than spending time to design and implement policies.

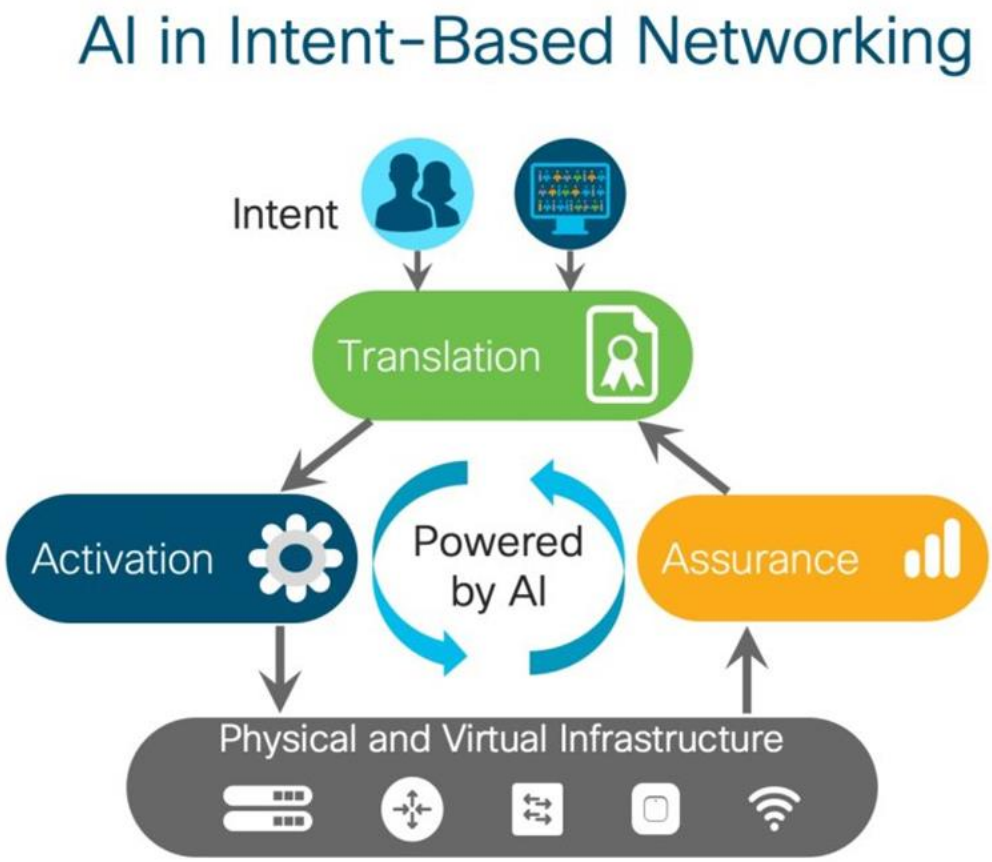
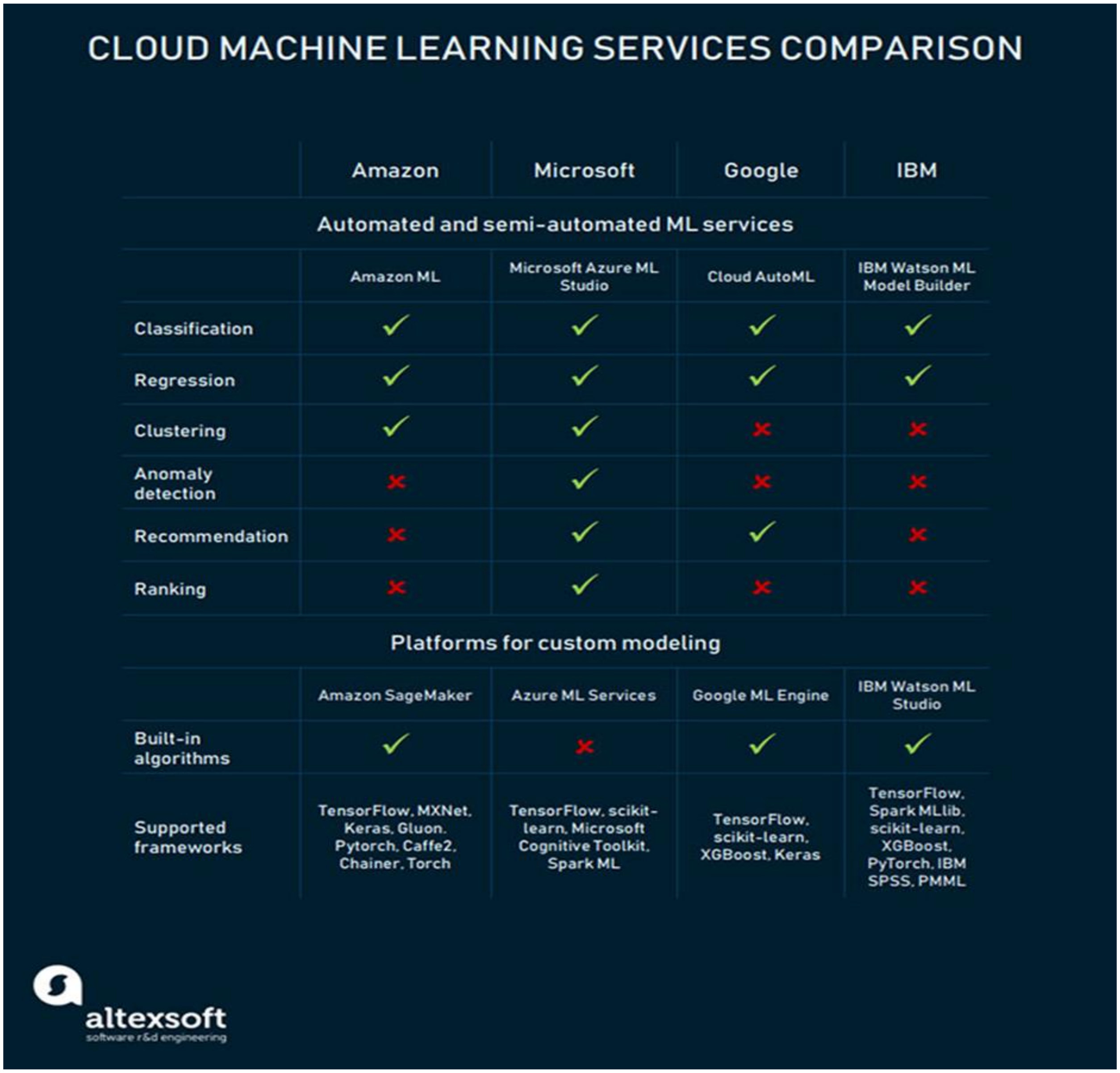


Figure 11. Conceptual View of Intent-Based Networking (Ref.65)

AI plays a crucial role by providing a path towards a zero-touch network that is a software-defined network (SDN) platform utilizing automation and AI.[66] In such a network, all network operations are automated, requiring no operator steps beyond the instantiation of intent. AI systems can autonomously come up with predictions, recommendations, and decisions.

14. AI/ML-as-a-Service

Machine learning as a service (MLaaS) is offered by all the major cloud service providers: (1) Amazon Machine Learning; (2) Azure Machine Learning; (3) Google Cloud Machine Learning; (4) IBM Watson Machine Learning. While the capabilities and main characteristics of these MLaaS platforms vary (Fig. 12), typically MLaaS includes services for data transformation, predictive analytics, data visualization, and advanced machine learning algorithms.



CLOUD MACHINE LEARNING SERVICES COMPARISON

	Amazon	Microsoft	Google	IBM
Automated and semi-automated ML services				
	Amazon ML	Microsoft Azure ML Studio	Cloud AutoML	IBM Watson ML Model Builder
Classification	✓	✓	✓	✓
Regression	✓	✓	✓	✓
Clustering	✓	✓	✗	✗
Anomaly detection	✗	✓	✗	✗
Recommendation	✗	✓	✓	✗
Ranking	✗	✓	✗	✗
Platforms for custom modeling				
	Amazon SageMaker	Azure ML Services	Google ML Engine	IBM Watson ML Studio
Built-in algorithms	✓	✗	✓	✓
Supported frameworks	TensorFlow, MXNet, Keras, Gluon, Pytorch, Caffe2, Chainer, Torch	TensorFlow, scikit-learn, Microsoft Cognitive Toolkit, Spark ML	TensorFlow, scikit-learn, XGBoost, Keras	TensorFlow, Spark MLlib, scikit-learn, XGBoost, PyTorch, IBM SPSS, PMML

Figure 12. MLaaS Comparison (Ref.67)

The main benefit of MLaaS is that customers can get started with ML applications quickly without need to build in-house infrastructure nor invest heavily in storage systems, servers, and software licenses.[68] However, MLaaS platforms are not free of downsides that keep some companies away from using them. For example, MLaaS solutions might not fit the specific needs of the company. Using ready-made MLaaS solutions, an organization does not develop its in-house expertise and this makes solution customization difficult. Finally, an organization relying on MLaaS becomes dependent on the service provider, which can change its product lists, pricing options, and service features.

15. How AI Makes IT Infrastructure Intelligent and Autonomous

AI-enabled intelligent infrastructure, which is self-deploying, self-provisioning, and self-healing is becoming central to computing, whether it is cloud computing or cloud-like environment on-premises. To implement this infrastructure, vendors focus on various aspects. Some vendors develop predictive analytics with proactive support, whereas others concentrate on application-level performance and availability. AI-driven intelligent infrastructures automate routine infrastructure management tasks and the time-consuming identification and resolution of application performance problems. As a result, intelligent infrastructures contribute to adoption of AIOps (Section 6) to free IT teams to work on projects that create new business value. Enabling infrastructures to autonomously optimize themselves for application availability, performance, and total cost of ownership (TCO) is the ultimate goal. Due to the size limitations of this article, I will briefly consider only a few examples of AI-based intelligent infrastructure solutions.

Dell CloudIQ is a free, software as a service (SaaS) solution that combines ML and human intelligence for efficient management of Dell EMC's storage systems (Unity, SC Series, XtremIO, PowerMax/VMAX, and PowerStore).[69] CloudIQ generates predictive ML analytics to look at historical data of systems and to offer best practices of Dell Technologies (Fig. 13). Consequently, IT Ops teams get faster insights and take quicker actions to prevent issues.

Dell PowerOne is an autonomous infrastructure integrating PowerEdge compute, PowerMax storage, PowerSwitch networking and VMware virtualization via PowerOne Fabric into a single system having a built-in AI engine to automate its lifecycle.[4] This advanced AI-based automation engine called PowerOne Controller enables autonomous operations for PowerOne. The automation engine makes deployment of workload-ready clusters possible with just a few clicks.

Dell PowerStore has AI/ML-driven systems self-management.[70] Its onboard AI engine automates initial volume placements based on available capacity, drive wear, and other factors. PowerStore helps in managing capacity planning by estimating "out of capacity" date and recommending volume migration plans for approval. PowerStore support plans include the above-mentioned CloudIQ cloud-based analytics.

Figure 13. CloudIQ Features and Benefits (Ref.69).

Dell PowerMax storage array is the next generation system in the VMAX enterprise product line. It has the real-time ML engine, which uses predictive analytics and pattern recognition for intelligent storage automation.[71] The ML engine provides PowerMax storage intelligence by analyzing data from Dell EMC's installed base of VMAX3 and VMAX/all-flash arrays. Dell PowerMax relies on a proprietary AI/ML-powered system for intelligent storage tiering.

Dell Precision Optimizer 5.0 is a hardware system optimizer that uses AI for optimizing system configuration to increase performance of specific applications.[72] It captures data from an application's background behavior and uses the data sets for training ML models to enable automatic adjustment of system configurations such as CPU, memory and storage to provide the most optimal settings.

16. Low-Code/No-Code & AI

Low-Code/No-Code application development is a visual development methodology based on model-driven logic and drag and drop elements to create applications.[73] It can be seen as "development democratization" as it allows "citizen developers" to utilize AI building blocks to generate an AI engine without writing complex code. The key benefit of low-code/no-code platforms like Mendix, OutSystems, Appian, and Caspio is speed of application development and it has been extremely important during the COVID-19 pandemic (Section 17.5). According to the forecast for 2020-2030,[74] the global market for low-code platforms will have revenue of $187 billion by 2030 and 31.1% CAGR, during the forecast period. There is also a growing interest to AI-driven programming, called program synthesis, that may be able to automatically generate code.[75]

17. How COVID-19 Changes AI/ML Role in IT Services

17.1 COVID-19 Transforms IT Service Strategies

The onset of the COVID-19 pandemic has pushed every industry worldwide to rapidly change the business model to respond appropriately to the outbreak. IT service strategies have been evaluated and some technology development projects were postponed or scrapped completely, whereas others became key strategies to address the business changes and the need to support a suddenly remote workforce. The pandemic has become a wake-up call for many companies as it showed how inefficient and outdated their legacy IT platforms really were. Remote support of some of them was very difficult or even not possible at all. At the same time, the percentage of employees working remotely has increased three times since the beginning of the pandemic, according to the Forrester report.[76] Work from home has become a "new normal" for enterprise IT services. This "anywhere-plus-office hybrid" model created new business and information technology requirements associated with a broader process of digital transformation.

Indeed, COVID-19 has become an accelerator for digital transformation of all businesses. IT teams scrambled to provide the resources required to shift to an essentially fully remote workforce. This has driven an investment in AI-based automation (Section 6) and public cloud services. As a result, in a matter of months, IT management had to make a decade's worth of changes in IT services.

Cybersecurity and data protection have also been key focus areas to address in IT response to COVID-19 as cybercrime activity increased by exploiting new vulnerabilities created by remote and distributed IT operations. Even before the pandemic, the critical role of AI in cybersecurity was well known (Section 12). COVID-19 has given AI-based cybersecurity an even higher priority.[77]

17.2 AIOps in the COVID-19 Time

The effects of COVID-19 on IT operations supported by a remote IT workforce are dramatic. Some organizations have seen incident volume increases of three to five times. AIOps that can make sense of the chaos (Section 6) has become a must-have tool in the COVID-19 time. In effect, AIOps has gained an unintentional proof-of-concept during COVID-19.

AIOps provides the agility necessary to support modern, software-defined IT services, the development of automated processes, and accommodation of non-traditional ways of working. Applications of AIOps amid the COVID-19 outbreak open ways for revamping IT operational infrastructure to address the increase in security threats.[78]

17.3 MLOps and COVID-19

One of the COVID-19 impacts on MLOps (Section 9) is the so-called concept drift. Of course, AI models are subject to change and the COVID-19 pandemic has just demonstrated that relying on the pre-pandemic training data caused many AI models to go haywire. For example, in the beginning of the pandemic most of the AI predictive models were limited by Chinese samples, which might not be generalizable. The problems with the absence of historical training COVID data have been exaggerated by noise inserted in big-data sets by social media and other online traffic. This noise should be filtered out; otherwise it may result in generating so-called lucky good fit models.[79]

Monitoring a model for concept drift plays a key role in applying MLOps in the pandemic time.[80] Indeed, concept drift in MLOps has reached unprecedented levels due to the COVID-19 pandemic when IT operations models dramatically changed by migrating to hybrid-cloud solutions, aggressive automation, addressing the increase in cybersecurity attack surfaces, etc. Implementing MLOps practices allows IT services to be more resilient to external volatile events, like rapid IT platform changes, regulatory changes, and other unforeseen external events caused by the COVID-19 pandemic.

17.4 How COVID-19 Is Speeding NoOps Implementations

We have already discussed how AIOps contributes to the morphing of DevOps into the NoOps concept (Section 10). Covid-19 is speeding up this transformation. It appears that in the post-COVID world there will be a growing business efficiency divide between organizations that use NoOps and as a result are becoming more agile in providing IT services and those which do not embrace the NoOps movement.

Layoffs caused by COVID-19 have reduced human resources in every department and IT is not an exception. This means the changes in IT operations need to be more than just rearranging the deck chairs on the Titanic. So, the NoOps concept comes on the scene with a promise to automate the IT operational environment to the point that a team is not needed to manage it.[81]

The hope brought by NoOps is that automating the deployment, monitoring and management of applications will eventually result in environments that no longer require the presence of humans in the IT operation center. COVID-19 can require dramatic changes in IT applications. As infrastructure configurations and application changes are well known be the top reasons for downtime, NoOps can shine in IT service delivery again by providing an infrastructure-as-code operational model, enabling the environment to be more easily stabilized and helping reduce some of the root causes of IT service outages. While NoOps model may look like an extreme at the present time, nevertheless NoOps allows IT teams to spend more of their time delivering new digital services, improving existing applications and contributing more time to broader business transformation efforts required by COVID-19.

17.5 The Bloom of AI-Based Low-Code/No-Code Development in the Time of COVID-19

New online apps and services have become urgently needed during the COVID-19 pandemic. For example, hospitals are looking for new mobile functionalities to track COVID-19 response and schools have started using apps for tracking virtual classroom attendance. The apps need to be developed quickly and organizations are quickly adopting low-code and AI to allow employees with no coding skills to build apps.[82,83]

The COVID-19 pandemic has forced app development teams to change how they work when the app development speed offered by low-code AI becomes a must-have. The low-code development for the New York City Health Department without writing any code is an example.[84] A similar “no-code” software development project has been done for the COVID-19 Support Hub in Washington, DC. Another example is the COVID-19 Risk Assessment App, which enables people everywhere to assess their risk for COVID-19, that has been developed and deployed in days instead of month-long coding sprints.[85]

At biotech company Moderna, which has successfully brought its coronavirus vaccine to the market, the improvement of onboarding became a key challenge. Using a low-code platform, all onboarding resources were put together in one place and it helped bring new workers up to speed quickly.[86] Responding to COVID-19, Geisinger Healthcare used the low-code platform developed by Quick Base to create an app allowing the healthcare provider to manage a significant drop in outpatient visits and a parallel increase of in-patient and ICU admissions.[86] It took less than a day for another healthcare provider in the U.S. to develop an app by using low-code tools so that doctors could enroll patients affected by COVID-19 in clinical trial treatments.

While in the pre-COVID era, low-code platforms were not on the radars of many CIOs, the pandemic has given them higher visibility. Development teams start using ML models to make test automation smarter, and natural language processing (NLP) is used for test case reviews. The advent of the GPT-3 (Generative Pre-Trained Transformer-3) from OpenAI has huge potential. For example, the use cases for GPT-3 APIs may include translating natural language into code for websites and solving complex medical question-and-answer problems.[87]

17.6 Edge Applications of AI for COVID-19

Mask detection, social distancing and field COVID testing have created a need for edge AI software and hardware. GPUs give better precision for biometric facial recognition algorithms. Nvidia, the leading GPU provider, has offered new edge software and AI models that are expected to play a vital role in biometric applications in the healthcare industry during the COVID-19 pandemic and in the post-COVID time.[88] For enhancing hospital public safety and patient monitoring, Nvidia also provided pre-trained AI models with support for vision, speech and NLP using Nvidia EGX Edge AI hardware.

17.7 Move to AI-First Cybersecurity in Times of COVID-19

Cyber criminals see COVID-19 as an opportunity to spread malware, ransomware, spam, scam, phishing, etc. (Fig. 14).[89] The sudden shift to remote work caused by the COVID pandemic has forced IT Ops teams to quickly adjust in various ways to maintain uptime and stability of critical IT services. Amidst this crisis, AIOps has emerged as a lifeline, as it facilitates remote collaboration, streamlines incident management, and accelerates threat detection and resolution. For that reason, AIOps has become the foundation of virtual Network Operations Centers (NOCs) enabling IT Network Ops teams to communicate and collaborate effectively during COVID-19.[90] Virtual NOCs will likely play an important role in post-COVID cybersecurity as well.

Figure 14. COVID-19 Related Security Threats (Ref.89)

Global network traffic patterns for telecommunications and web-based applications have changed dramatically during the COVID-19 pandemic. Companies using rules-based network monitoring systems have really struggled to keep up with these changes. In the beginning of the pandemic, ML models trained on pre-pandemic data sets could not predict unprecedented paradigm shifts in network usage in the COVID-19 time. This is like the concept drift discussed in Section 17.3. However, once more COVID-aware network and cybersecurity data sets become available for ML training, ML systems eventually catch up to the new network usage patterns so that they can determine new network threats.

COVID has accelerated the transition to the AI-First strategy in cybersecurity by implementing the so-called Third Wave AI (the term was suggested by DARPA [the Defense Advanced Research Projects Agency]).[91] This type of AI identifies malicious and abnormal behavior within a company's network, and it can manage a shift from historical procedures and network data to the COVID-time patterns. Third Wave AI systems can constantly adapt their baseline to reflect the rapid changes that they can handle better than humans.[92] For example, at the beginning of the pandemic, many conventional rule-based security systems went haywire and sent security teams storms of false alarms. AI-first security platforms can monitor the unexpected network behavior by utilizing behavioral analysis. Third-wave AI systems can identify network traffic pattern anomaly by detecting brand new attacks which are occurring more frequently during the COVID-19 pandemic.

17.8 How COVID-19 Causes Transformation to AI-Driving Data Centers

COVID-19's impact on data centers is what can be expected: creating team shifts meeting social distance requirements; staff shortages, as some people have to be off work for self-isolation; and difficulties in performing some routine tasks remotely. A need to defer some maintenance because of COVID-19 can affect IT service availability. Indeed, even missing one maintenance on a generator can impact its warranty and lead to a failure that could have been prevented by performing the scheduled maintenance service. Replacement of failed hardware can be a challenge because COVID made data center equipment supply chains unreliable.

All these circumstances start driving the use of AI to efficiently manage power and cooling systems, automate predictive maintenance, and balance workload distribution in enterprise data centers. Monitoring the health of servers, storage, and networking by AI systems allows them to proactively predict if equipment is likely to fail. AI systems can automate workload deployment to achieve the most efficient consumption of the infrastructure resources in real time.[93] The AI help in creating highly automated, secure, and to some extent self-healing data centers requiring little human intervention has become extremely important for providing IT services in the COVID era. The pandemic has also become a driving force for organizations to move from their own data centers to colocation services[94] and to public clouds.

17.9 COVID-19 as a Catalyst for AI-Guided Cloud Migration

Dell Technologies CEO Michael Dell said the COVID-19 pandemic caused many organizations to accelerate their migrations to public cloud solutions since cloud service elasticity is capable of meeting unexpected spikes in service demand.[95] However, using only public clouds is not likely to be the ultimate OpEx-driven way of doing business and the IT leader expects customers to eventually define their long-term strategy for IT services to utilize advantages of multi-cloud-based services. Multi-cloud services provide agility, better business continuity, and efficient cost optimization.

Migrations to the cloud helped companies reinvent the way they conduct their businesses in the time of COVID-19. The need for AI services has grown and many cloud providers offer AIaaS and MLaaS (Section 14). As a result, the global cloud market recorded significantly larger growth of the healthcare segment in 2020. AI is at the heart of many healthcare methods to fight COVID-19. For example, the National Institutes of Health has launched the Medical Imaging and Data Resource Center (MIDRC) to use AI and medical imaging in COVID-19 patient treatment.[96] Cloud-based AI applications can analyze complicated medical images to help radiologists in COVID-19 diagnostics.

All this is driving migration of IT services used by healthcare organizations to the cloud.[97] However, it is a great challenge to smoothly move IT apps to a cloud-centric IT model within a short period of time. Cloud migration projects of enterprise level are generally complex and involve planning transformations of a large number of applications, many of which are business critical. To be successful, cloud migrations should rely on automating IT operations and application performance monitoring; on automating root-cause analysis and remediation; and on automating performance baselining and configuration management. This “automate everything” approach must leverage AI/ML.[98,99] AI-based solutions can also be applied for improving CI/CD pipelines to meet migration deadlines. AIOps tools (Section 6) can automate the cloud migration process by leveraging AI to redefine the configuration required for the new platform in the cloud.

18. What Is the Place of AI for IT Services in the Post COVID-19 World?

The pandemic has transformed nearly everything. At the time of writing this article, COVID-19 vaccines authorized by the FDA have already become available and the post-pandemic time does not look to be too far away. However, its shapes are not clear. It is hardly expected that people's lives and businesses will return to what they were in the pre-COVID world. On the contrary, we will see many things become "new normal" and the use of AI/ML in IT services is not an exception. The changes are multi-dimensional and if I try making predictions on the post-pandemic roles of AI/ML in IT services, I will be on thin ice.

Business recovery from the COVID-19 fallout will depend on how successful the organization has been in the digital transformation of its business. The increasing reliance on digital solutions will elevate the role of IT services. The significant growth of IT service automation during COVID-19 will continue in the post-COVID time.[100] The development of AIOps and MLOps has increasing momentum and will capitalize on the value these Ops models have demonstrated during the pandemic.[101] IT strategy reviews will start to zero in on NoOps. Data centers will keep adding more automation and remote monitoring in the post-COVID future.

AI-guided cloud migrations and the use of cloud-provided AI tools including AIaaS will continue at an accelerated pace. AI has cemented its foothold in IT services during the pandemic. Adoption of cloud computing will be accompanied by the growth of edge computing and transformation of cloud computing in the distributed cloud platform to meet 5G and remote workforce requirements. To address cybersecurity threats caused by COVID-19, organizations put effective AI-based defenses in place.[102] This salient trend will carry on and AI-driven cybersecurity automation will progress to a higher level.[103]

19. Conclusion

AI applications to IT services are a very broad area with diverse applications and use cases, and considering all of them is not possible in a limited size article. AI has proven itself as a key enabler of intelligent self-managing infrastructure, AIOps, MLOps, NoOps, hyper-automation, and low-code/no-code application development accelerated time-to-value, particularly in the COVID-19 time. While I have had to leave AI-driven intelligent storage services, AI-powered database services, and AIOps-based management of 5G infrastructure out of the article scope because of length limitations, I intend to review them in my next articles.

I hope the readers will find my article helpful in understanding how AI transforms traditional IT services and how they can contribute to this complex multidirectional process.

20. References

1. http://www.gartner.com/en/information-technology/glossary
2. E. Hechler, M. Oberhofer, and T. Schaeck. Deploying AI in the Enterprise — IT Approaches for Design, DevOps, Governance, Change Management, Blockchain, and Quantum Computing. Apress, 2020.
3. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361109
4. http://www.delltechnologies.com/en-sa/blog/unveiling-dell-emc-s-powerone-autonomous-infrastructure
5. http://www2.deloitte.com/fi/fi/pages/technology/articles/part1-artificial-intelligence-defined.html
6. G. Batra, Z. Jacobson, S. Madhav, A. Queirolo, and N. Santhanam. Artificial-intelligence hardware: New opportunities for semiconductor companies. McKinsey & Company, 2018.
7. AI Infrastructure Reference Architecture. IBM. Doc. 87016787USEN-00. 2018.
8. http://learn.g2.com/history-of-artificial-intelligence
9. http://ahistoryofai.com/antiquity/
10. http://www.futurity.org/artificial-intelligence-greek-myths-1999792
11. R. Raina, A. Madhavan, and A. Y. Ng. Largescale deep unsupervised learning using graphics processors. Proceedings of the 26th Annual International Conference on Machine Learning. ICML 2009. New York, NY, USA: ACM: 873–880.
12. A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS, 2012, pp. 1106–1114.
13. S. M. Khan and A. Mann. AI Chips: What They Are and Why They Matter. Center for Security and Emerging Technology, 2020.
14. http://www.nvidia.com/en-us/data-center/a100
15. R. Xu, F. Han, and Q. Ta. Dell EMC Ready Solutions for AI – Deep Learning with NVIDIA. Dell EMC Reference Architecture. 2019.
16. T. Henson. AI Architectures: Platform Considerations from Proof-of-Concepts to Large Scale Deployments. Dell Technologies World, 2020.

17. http://www.intel.com/content/www/us/en/artificial-intelligence/programmable/fpga-gpu.html
18. http://www.eweek.com/cloud/microsoft-uses-intel-fpgas-for-smarter-bing-searches
19. http://www.intel.com/content/www/us/en/products/processors/movidius-vpu.html
20. http://www.sdxcentral.com/articles/news/intel-reveals-ai-asics-claims-3-5b-in-ai-revenue-in-2019/2019/11/
21. http://cloud.google.com/tpu/docs/system-architecture
22. http://cloud.google.com/edge-tpu
23. http://medium.com/@aron.kirschen/why-tops-w-is-a-bad-unit-to-benchmark-next-gen-ai-chips-573b9152f9ae
24. http://semiengineering.com/the-murky-world-of-ai-benchmarks
25. http://blogs.vmware.com/ml-ai/2020/01/31/gpu-virtualization-for-machine-learning-workloads
26. http://www.nvidia.com/en-us/data-center/virtualization
27. http://docs.vmware.com/en/VMware-vSphere-Bitfusion/index.html
28. GPU-as-a-Service (GPUaaS). Solution Brief. Dell Technologies, 2020.
29. http://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit.html
30. http://www.businesswire.com/news/home/20200512005105/en/DeepCube-Launches-Industry-First-Deep-Learning-Software-Accelerator-to-Enable-Real-World-AI-Deployments.
31. http://medium.com/datadriveninvestor/ai-operating-system-7809e7fc4e49
32. http://www.enterpriseai.news/2020/10/14/blackswan-unveils-an-ai-os
33. E. Coronado, A. Thomas, S. Bayhan and R. Riggio. aiOS: An Intelligence Layer for SD-WLANs. NOMS. IEEE/IFIP Network Operations and Management Symposium, 2020.
34. R. Snehal and A. Shetty. A Study on Artificial Intelligence Operating System. International Journal of Latest Trends in Engineering and Technology. Special Issue SACAIM, 2017, pp. 59-61.
35. http://aws.amazon.com/machine-learning
36. http://azure.microsoft.com/en-us/overview/ai-platform
37. http://cloud.google.com/products/ai
38. http://www.zdnet.com/article/what-is-neuromorphic-computing-everything-you-need-to-know-about-how-it-will-change-the-future-of-computing
39. http://www.informationweek.com/big-data/ai-machine-learning/advancing-ai-with-neuromorphic-computing-platforms

40. http://newsroom.intel.com/press-kits/neuromorphic-computing
41. http://www.research.ibm.com/articles/brain-chip.shtml
42. http://www.qualcomm.com/news/onq/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing
43. http://www.gartner.com/en/information-technology/glossary/aiops-artificial-intelligence-operations
44. C. Rich, P. Prasad, and S. Ganguli. Market Guide for AIOps Platforms. Gartner, Nov. 2019.
45. http://www.ibm.com/cloud/architecture/architectures/sm-aiops/overview
46. http://www.ibm.com/cloud/architecture/architectures/sm-aiops/reference-architecture
47. Innovation Insight for Observability. Gartner. Published 28 September 2020 - ID G00720189.
48. http://www.numerify.com/blog/how-has-itil-4-evolved-change-management-and-how-can-ai-powered-analytics-propel-its-objectives
49. http://www.prnewswire.com/news-releases/aisera-launches-ai-service-management-aism-platform-with-astonishingly-high-auto-resolution-rates-300994811.html
50. http://www.cisco.com/c/dam/en/us/products/collateral/cloud-systems-management/crosswork-network-automation/white-paper-c11-741691.pdf
51. E. Hechler et al., Deploying AI in the Enterprise. https://doi.org/10.1007/978-1-4842-6206-1_7
52. http://www.sas.com/en_us/insights/articles/analytics/modelops.html
53. http://www.sas.com/en_us/solutions/operationalizing-analytics/modelops-approach.html
54. http://www.modelop.com
55. http://www.ibm.com/cloud/machine-learning/modelops
56. F. Choudhary, S. Vashisth, A. Chandrasekaran, and E. Brethenoux. Innovation Insight for ModelOps. Gartner, 6 August 2020.
57. http://www.bmc.com/blogs/mlops-machine-learning-ops
58. http://blogs.nvidia.com/blog/2020/09/03/what-is-mlops
59. http://www.cio.com/article/3407714/what-is-noops-the-quest-for-fully-automated-it-operations.html
60. http://www.techradar.com/news/world-of-tech/management/no-humans-required-heuristic-automation-and-the-it-system-which-runs-itself-1285259
61. http://www.itpro.com/development/devops/356330/what-is-noops

62. H. Badakhchani. NoDev NoOps NoIT: Principles governing the ideology, methodology and praxeology of informed IT decision making. Independently Published, Jul 23, 2017.
63. http://www.infoworld.com/article/3373850/ai-cloudops-is-coming-whether-you-like-it-or-not.html
64. http://www.balbix.com/insights/artificial-intelligence-in-cybersecurity
65. http://blogs.cisco.com/networking/improving-networks-with-ai
66. B. Koley. The Zero Touch Network. Google Technical Infrastructure. CNSM 2016.
67. O. Kolisnykov. http://www.topbots.com/comparing-machine-learning-as-a-service-platforms
68. http://www.topbots.com/comprehensive-guide-to-mlaas
69. D. Barboza, S. Sharpe, and S. Morgan. Dell EMC CloudIQ: A Detailed Review. Technical White Paper. Dell Technologies, 2020.
70. http://www.delltechnologies.com/en-us/storage/powerstore-storage-appliance.htm
71. http://www.delltechnologies.com/en-us/storage/powermax.htm
72. http://www.delltechnologies.com/en-us/work-at-full-speed/optimizer.htm
73. http://towardsdatascience.com/no-code-low-code-ai-new-business-models-and-future-of-data-scientists-a536beb8d9e3
74. http://www.researchandmarkets.com/reports/5184624/low-code-development-platform-market-research
75. http://towardsdatascience.com/program-synthesis-in-ai-991590b4ba7f
76. http://go.forrester.com/blogs/the-state-of-remote-work-during-Covid-19-an-update-on-forresters-research
77. J. Webster and K. Macomber. Enterprise IT Response to COVID-19. The Evaluator Group, October 2020
78. http://devops.com/aiops-greasing-the-wheel-of-digital-security-amid-Covid-19/
79. S. Whitelaw, M. A. Mamas, E. Topol, and H. G C Van Spall. Applications of digital technology in COVID-19 pandemic planning and response. www.thelancet.com/digital-health Vol 2, August 2020.
80. http://www.iguazio.com/blog/concept-drift-and-the-impact-of-Covid-19-on-data-science/
81. http://devops.com/revisiting-2020-tech-predictions-in-light-of-Covid-19
82. http://www.salesforce.com/blog/low-code-apps-development-Covid19
83. http://www.forbes.com/sites/forbestechcouncil/2020/07/28/how-our-health-crisis-proved-low-code-is-here-to-stay
84. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7361109

85. http://www.alphasoftware.com/blog/low-code-helps-battle-the-coronavirus-and-it-can-help-your-enterprise-as-well
86. http://www.ciodive.com/news/low-code-implementation-coronavirus-2020/580024/
87. http://industrywired.com/openai-brings-a-new-hope-through-the-Covid-19-pandemic-with-gpt-3
88. http://www.steatite-embedded.co.uk/edge-ai-computing-from-steatite/
89. http://cujo.com/Covid-19-related-threats
90. http://www.orange-business.com/en/webinars/leading-partners-interview-moogsoft-how-aiops-virtual-noc-provide-better-collaboration
91. http://artificialintelligence-news.com/2018/09/28/darpa-third-wave-artificial-intelligence/
92. http://www.dogtownmedia.com/will-Covid-19-spur-the-shift-to-ai-first-cybersecurity/
93. http://www.networkworld.com/article/3568354/how-ai-can-create-self-driving-data-centers.html
94. http://www.datacenterknowledge.com/industry-perspectives/Covid-19-accelerating-enterprise-data-center-migration-offsite-multitenant
95. http://www.crn.com/news/data-center/michael-dell-Covid-19-caused-public-cloud-fast-route-but-multi-cloud-will-resonate
96. http://www.nih.gov/news-events/news-releases/nih-harnesses-ai-Covid-19-diagnosis-treatment-monitoring
97. http://www.networkworld.com/article/3596189/with-Covid-19-hanging-on-migration-to-the-cloud-accelerates.html
98. http://thenewstack.io/how-ai-makes-cloud-migration-a-whole-lot-easier/
99. H. Sun, M. Vukovic, J. Rofrano, and C. Lin. Advantages and Challenges of Using AI Planning in Cloud Migration. IBM T.J. Watson Research Center, 2019.
100. http://www.plugandplaytechcenter.com/resources/artificial-intelligence-automation-post-Covid-19-era/
101. http://sciencelogic.com/blog/why-you-need-aiops-in-a-post-Covid-world
102. http://www.f5.com/labs/articles/bylines/how-ai-will-automate-cybersecurity-in-the-post-Covid-world
103. http://venturebeat.com/2020/09/06/how-ai-will-automate-cybersecurity-in-the-post-Covid-world/