Digital Systems & Technology

How Applying an AI-Defined Infrastructure Can Boost Data Center Operations

An artificial-intelligence-based infrastructure that uses the data available within the data center to optimize and automate infrastructure operations can enhance operational efficiencies and improve the quality of service offered to the business.

Executive Summary

It is hardly news that artificial intelligence (AI) is entering the mainstream. In fact, Gartner predicts that by 2022 AI will contribute nearly $3.9 trillion in business value globally,[1] and that 40% of all application development projects will have AI developers on their teams. Not surprisingly, Gartner also places AI-based technologies strongly in the top trends for 2019.[2]

Moreover, the use of AI within the data center is growing steadily. In this context, AI is being used to predict and automate many of the tasks that humans currently perform. This concept is known as an AI-defined data center.[3] As automation and the use of code to run the data center through software-defined technologies mature, data center operations are becoming much more efficient, with fewer mistakes due to manual

interventions. The logical next step in increasing performance is the use of an AI-defined infrastructure.[4]

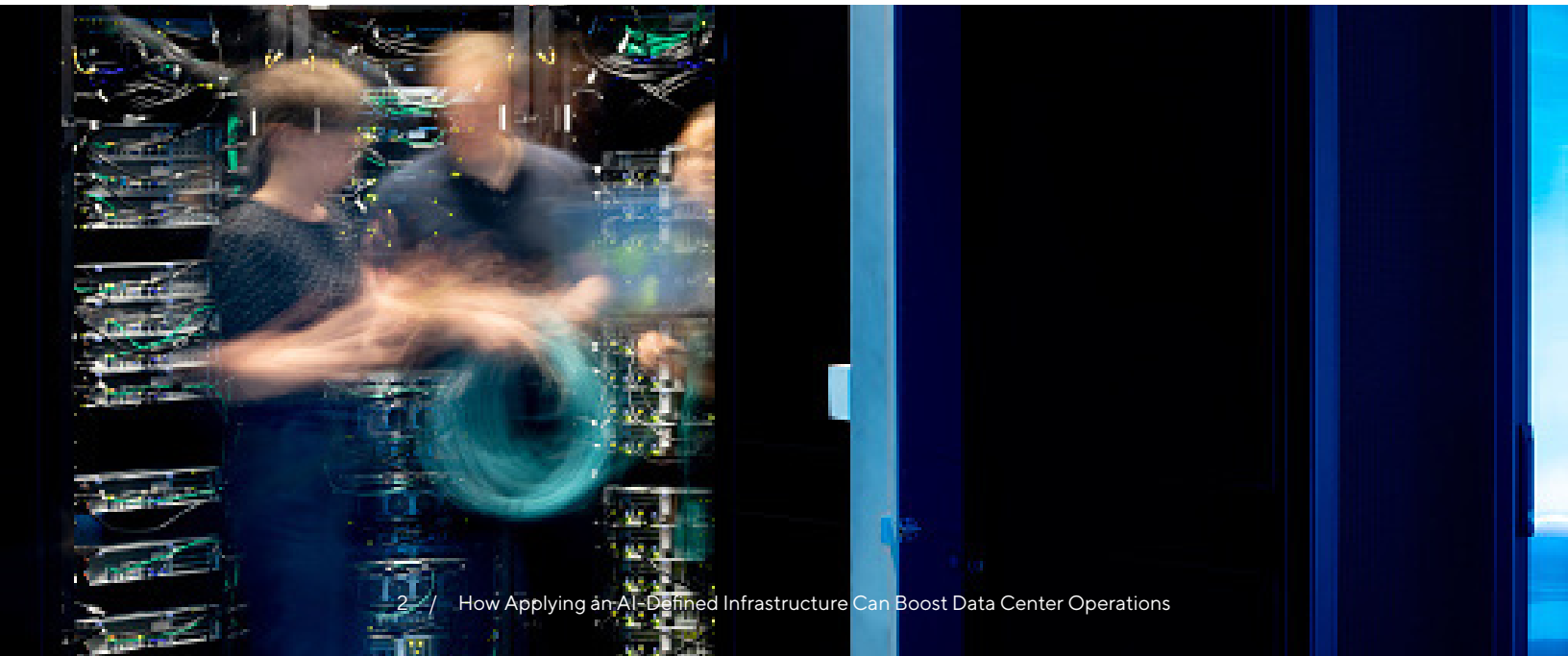
Recent studies by EMC and Intel for Forbes Insights[5] reveal that 70% of organizations classified as leaders in digital transformation believe that data and analytics will become integral to running IT infrastructures within the next two to five years. Organizations such as Hitachi Vantara have already begun incorporating AI technologies into storage elements of its IT infrastructure operations.[6] Other vendors such as Extreme Networks are reportedly readying AI to enhance operations capabilities within networks.

This white paper explores the evolution of the data center and details how an AI-defined infrastructure can help businesses operate more efficiently. It also provides high-level guidance on how IT organizations can implement their own AI-defined data centers.

Benefits of an AI-defined infrastructure

The advantages — primarily operational — of using an AI-defined infrastructure include the following:

- Categorizing incidents automatically highlights the key areas of the data center that operational teams must focus on.
- Using automated IT systems reduces the amount of human error in operational activities.
- AI-based interactions with end users enhances the quality of services provided to the business.
- Automatic assignment of capacity within the data center enables optimum cost utilization for infrastructure.
- Enhanced security and multilingual services of an AI-defined infrastructure allow IT operations to better use global teams.
- Increased data-driven insights into data center operations raise awareness among leadership teams for making efficiency gains.
- Anomaly detection and quick automated responses to threats further enhance the security of data center operations.
- Predictive and preemptive resolution of incidents reduce down time for business applications.



An AI engine for the data center

The heart of an AI-defined data center is its AI engine. Figure 1 illustrates the architecture of AI-defined data centers.

First, data from various sources needs to feed into a central data repository, or data lake. The data from that data lake is then fed into an AI engine, which features a combination of machine learning (ML) and other AI-based capabilities. The engine then processes the data and provides an output, which can be used either for reporting or to alert operations staff to take further steps or perform some activity automatically through automation technologies.

Anatomy of an AI-defined data center

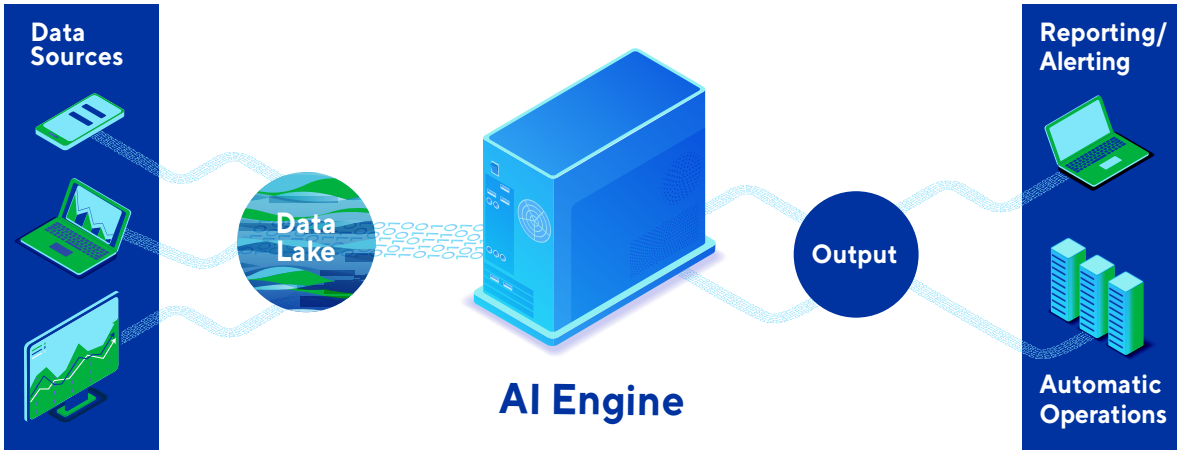


Figure 1

Machine-learning algorithms

The four types of ML algorithms typically used for data learning and prediction are:

1. **Regression:** These algorithms are used to forecast a numerical value based on the pattern observed for the data series. An example: "What will the CPU utilization of my server be at 10 pm tomorrow night?"
2. **Classification:** These algorithms are used to classify a data set into one of several predefined categories. An example: "What is the priority level of the ticket received — p1, p2 or p3?"
3. **Clustering:** These algorithms group data sets automatically into clusters based on similarities between data values. There are several clustering algorithms available, all of which group data values slightly differently. An example: "How do I group these servers based on their similarities in performance?"
4. **Anomaly detection:** These algorithms are used to automatically detect outliers in a data series when the observed values do not follow the expected pattern. An example: "How do I detect a difference in network behavior when a virus has attacked the network?"

Figure 2 shows the mapping of potential data inputs available in a data center to potential outputs of ML algorithms.

Mapping data inputs to outputs through ML algorithms

Category	Data Inputs	Regression	Classification	Clustering	Anomaly Detection	Automatic Operations
End User Computing	• Patch levels • User allocation • App service status	• User base predictions • Predicting patching requirements	• Security risk classifications • User type classifications • App service types classification	• Automatically grouping of users based on feature sets available	• Anomalies in patching statuses over time • Anomalies in user access • Anomalies in services running	• Reporting and planning for patch management • Automatic patch management for failures
Server Operating System	• Patch levels • Server roles • App service status		• Patch level classifications • Classification of server types to business functions • Classification of service types to business processes	• Clustering of servers based on their usage characteristics	• Anomalies in patch levels • Unexpected server roles • Unexpected services running	• Automatic patch management for failures • Services outage management
Server Infrastructure	• CPU, memory, storage, networks utilization • Cooling, power and data center consumption • Serial numbers and asset inventory	• Data center sizing predictions • Utilization predictions	• Utilization levels (high/medium/low)	• Clustering of infrastructure based on usage characteristics	• Unexpected utilization behaviors	• Auto-scaling • Automatic rebooting • Alerting for anomaly analysis.
Storage	• IOPS • Utilization, capacity	• Utilization and capacity predictions	• Utilization levels (high/medium/low) • Classification into storage types	• Grouping of LUNs based on usage parameters	• Unexpected utilization behaviors	• Automatic provisioning of additional storage
Networks	• Bandwidth utilization • Ports and traffic types	• Utilization and capacity predictions	• Utilization levels (high/medium/low)	• Grouping of network zones based on usage	• Unexpected utilization behaviors • Unexpected ports/traffic types	

Figure 2 (*continues on next page*)

Category	Data Inputs	Regression	Classification	Clustering	Anomaly Detection	Automatic Operations
Firewall	• Type of firewall • Source, destination, ports used		• Classification into business/ infrastructure/ others		• Unexpected network transactions	
Security	• Usage anomalies • Authentication • User access • User location • User IP address	• User access predictions	• Classification into user types	• Grouping of usage patterns	• Unexpected user access (location, IP address, authentication attempts, etc.)	• Blocking of insecure transactions • Automatically moving virus workloads to DMZ
Databases	• Usage • Utilization	• Utilization and capacity predictions	• Utilization levels (high/medium/ low)	• Clustering of databases based on usage	• Unexpected utilization behaviors	
Monitoring	• Outage metrics • Logs	• Outage predictions	• Criticality classifications (P1, P2, etc.)	• Clustering outage types	• Nonordinary monitoring alerts	• Automatic ticket generation • Preemptive fixes
Load Balancing	• Uptime • Certificate usage	• Downtime predictions	• Classification into normal/HA/ DR scenarios		• Unexpected behaviors	• Auto-scaling • Invoking DR
Service Desk	• Ticket metrics	• Ticket predictions • Frequently searched topics	• Classification into P1, P2, P3, etc. • Classification into responsible operational team	• Clustering tickets based on features for further analysis	• Increase in tickets after patching or upgrades	• Auto-assigning tickets based on priority levels
Licensing & Software Metering	• Usage metrics • Licensing terms	• Software usage predictions • Licensing calculations (for example, with Java SE, processor-based metric for servers and named user plus XX-based metric for desktops) • 3. Licensing recommendations	• User workload segmentation	• Grouping of servers/ applications resulting in consolidation for reduced TCO • Utilization trends of license types	• Spike in usage – e.g., OSS/FOSS software	

Figure 2 (*continued*)

If the CPU of a virtual machine is expected to spike every weekend, then the AI engine can automatically increase CPU allocation during the weekends and reduce it during the week to optimize costs.

Automatic operations using ML outputs

Organizations have historically used data within data centers to trigger alerts or perform analyses on the health of the infrastructure. With the advent of automation and the public cloud, it is highly possible that these data points can be used to trigger operations within the data center.

For example, if the CPU of a virtual machine is expected to spike every weekend, then the AI engine can automatically increase CPU allocation during the weekends and reduce it during the week to optimize costs.

Figure 2 includes a column called "Automatic Operations," which offers potential use cases that yield the IT operational benefits outlined above. For example, patching environments is currently a major challenge for organizations. Every time software providers release a new security patch, IT operations teams must go through major planning to roll out these patches across their estate. This is often a tedious and time-consuming task of keeping track of all successful patches, identifying the failed patches and replanning the rollout for those failed patches. With the use of AI technologies, it will be possible to automate the patching process, considerably reducing the operational overhead in tracking and managing these patching cycles. Automating patching also makes it possible to introduce more frequent patch cycles, thereby enhancing the security profile.

Other forms of applicable AI

Several other forms of AI can also be utilized within the data center including:

- **Text to speech and speech to text:** This technology allows a system to convert audio captured from a human voice into machine-readable text, and vice versa.
- **Intelligent search:** This technology allows search engines to automatically use meaning extracted from unstructured data for searches. An example: Searching for a specific brand name within a two-hour video sequence that automatically identifies the time stamps where the brand name appears.

With the use of AI technologies, it will be possible to automate the patching process, considerably reducing the operational overhead in tracking and managing these patching cycles.

- **Computer vision (image and video analysis):** This technology allows for the analysis of images and video automatically to extract useful information. For example, identifying images that contain data centers.
- **Advanced text analytics:** This technology allows a system to analyze sentences to extract useful information. For example, Twitter feeds can be analyzed to identify positive and negative sentiments toward a particular topic.
- **Translation services:** These produce translations from one language to another. This is particularly useful in a call center scenario where language no longer must be a barrier to providing good service to the customer.

Figure 3 summarizes the potential input/output types in a data center that utilizes these other forms of AI.

Mapping data inputs to outputs via other forms of AI

Category	Text to Speech	Speech to Text	Intelligent Search	Image & Video Analysis	Advanced Text Analysis	Translation Services
End User Computing		• Modern UI			• Analyze patch levels	
Compute, Storage, Networks			• Advanced technology documentation search		• Analyze patch levels	
Firewall			• Advance search of firewall configurations		• Automatic analysis of firewall configurations	
Security				• Data Center Security		
Databases			• Search for GDPR data			
Monitoring					• Intelligent analysis of monitoring alert • Automatic analysis of log files	• Logs translations
Service Desk	• Chatbots for telephone calls	• Chatbots for telephone calls			• Analysis of ticket type and content • Chatbots	• Global service desk
Licensing & Software Metering					• Automatic analysis of license terms	

Figure 3

Moving forward: Getting started with AI-defined data centers

We recommend the following steps to implement an AI-defined infrastructure:

- **Understand what data is available in the data center.** This is a critical first step that organizations must perform in order to implement an AI-defined infrastructure. The most obvious place to extract data is from existing monitoring solutions, many of which provide API integrations to extract the events. Most infrastructure providers for storage, network and compute also provide data relevant to those technologies which can be tapped into.
- **Clean the available data set to plug any present gaps.** Since the data originates from different sources, it is important to agree on a standard approach for storing it, as well as the associated data architecture. For example, systems may capture date/time or names in a different format than others. Part of the data cleansing stage is to convert the data from each source to a central unified format so that it can be used uniformly by the downstream stages.
- **Implement a data lake to store the cleaned data set centrally.** Depending on how the data will be used for business value, multiple applications will need to be developed — but all referencing the same central data source for information. Therefore, it is important that the data is centrally stored securely with access controls such that downstream applications can make use of the relevant data sets.
- **Research AI models and automation tools that can make use of this data set.** This is a time-consuming activity, and is often very different for each type of business. For example, a retail organization may find that systems are more heavily used at certain times of the year, whereas a manufacturing-based system is less seasonal and is used 24/7. Based on usage variances, different AI models will work better for particular business scenarios, so it is key to understand which model to use for which business process.

- **Train the appropriate AI models.** AI models, particularly involving deep learning, will need to go through a training phase so that the models are trained to increase accuracy of results. This stage should be planned into the work program to allocate for infrastructure and time.
- **Implement an AI engine for the data center.** Once the models are trained to the required accuracy, you can put the capability into production.
- **Design the automation actions using the AI engine's output.** The AI engine outputs can be used in different ways, and this stage of the process (which can be parallel to the previous phase) will be to decide how best to make use of the outputs. For example, if a security anomaly is detected in a particular server, should the server automatically be locked down from the network or is it preferable that a notification be sent to the operations teams so that it can make an appropriate decision?

As AI data center technologies advance, IT organizations must consider ways to yoke the AI and infrastructure concepts and generate further operational value for the business.

Given the dynamic nature of today's business, it is absolutely essential that organizations make use of such technology advances to increase operational efficiency across their IT infrastructure. If not, they will be left behind playing catch-up in an ever-changing data-driven digital world.

Endnotes

1 "Top 10 Strategic Technology Trends for 2019," Gartner Symposium/ITxpo 2018 in Orlando, October 2018, www.gartner.com/en/newsroom/press-releases/2018-10-15-gartner-identifies-the-top-10-strategic-technology-trends-for-2019.

2 "Top 10 Trends Impacting Infrastructure and Operations for 2019," Gartner IT Infrastructure, Operations and Cloud Strategies Conference 2018, December 2018, www.gartner.com/en/newsroom/press-releases/2018-12-04-gartner-identifies-the-top-10-trends-impacting-infras.

3 "Introducing the AI Defined Infrastructure," CIO.COM, September 2017, www.cio.com/article/3227608/artificial-intelligence/introducing-the-ai-defined-infrastructure.html.

4 "5 ways AI can change Infrastructure Management Services," Flat world solutions, www.flatworldsolutions.com/IT-services/articles/5-ways-ai-can-change-infrastructure-management-services.php.

5 "Artificially Intelligent Data Centres: How the C-Suite is Embracing Continuous Change to Drive Value," Forbes Insights, EMC and Intel, December 2018.

6 Hitachi Vantara, "Infrastructure AI Operations at Hitachi Vantara," May 2018.

About the authors

Dr. Naveen Thomas

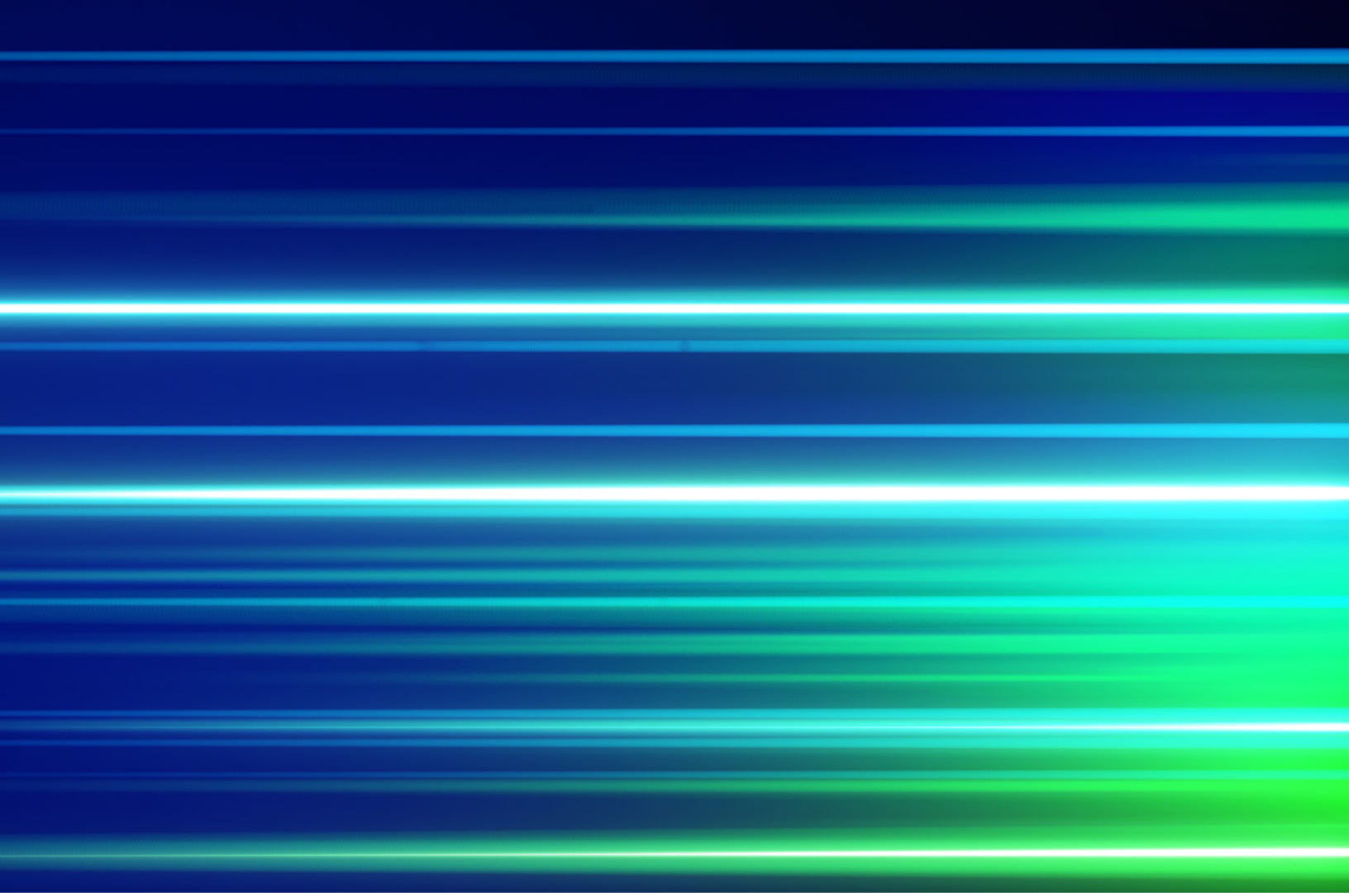
Associate Director,
Cognizant Technology Services

Dr. Naveen Thomas is an Associate Director at Cognizant Technology Services within the company's Digital Systems & Technology line of service. He has over 16 years of consulting experience at several leading consulting firms, focusing predominantly on innovation, cloud technologies and infrastructure consulting. Naveen holds a PhD in computer vision/AI and a master's in computer systems engineering from the University of Bristol. He can be reached at Naveen.Thomas@cognizant.com | www.linkedin.com/in/naveenthomas1980/.

Alec Selvon-Bruce

Senior Director and Lead, Enterprise Computing and Cloud Practice,
Cognizant Technology Services in EMEA

Alec Selvon-Bruce is a Senior Director and Lead for Enterprise Computing and Cloud Practice at Cognizant Technology Services in EMEA, within the company's Digital Systems & Technology line of service. He has over 25 years of professional services experience across leading system integrations and technology providers. Alec drives the innovation councils that provide guidance on infrastructure and cloud strategy. He holds a master's in political studies from the University Of Aberdeen. Alec can be reached at Alec.Selvon-Bruce@cognizant.com | www.linkedin.com/in/alecselvonbruce/.



About Cognizant's Digital Systems & Technology

The Cognizant Digital Systems & Technology line of service works with clients to simplify, modernize and secure IT infrastructure and applications, unlocking the power trapped in their technology environments. We help clients create and evolve systems that meet the needs of the modern enterprise by delivering industry-leading standards of performance, cost savings and flexibility. To learn more, contact us at simplify@cognizant.com. You can also visit us at www.cognizant.com/cognizant-digital-systems-technology, or e-mail us at Inquiry@ cognizant.com.

About Cognizant

Cognizant (Nasdaq-100: CTSH) is one of the world's leading professional services companies, transforming clients' business, operating and technology models for the digital era. Our unique industry-based, consultative approach helps clients envision, build and run more innovative and efficient businesses. Headquartered in the U.S., Cognizant is ranked 195 on the Fortune 500 and is consistently listed among the most admired companies in the world. Learn how Cognizant helps clients lead with digital at www.cognizant.com or follow us @Cognizant.

Cognizant

World Headquarters

500 Frank W. Burr Blvd.
Teaneck, NJ 07666 USA
Phone: +1 201 801 0233
Fax: +1 201 801 0243
Toll Free: +1 888 937 3277

European Headquarters

1 Kingdom Street
Paddington Central
London W2 6BD England
Phone: +44 (0) 20 7297 7600
Fax: +44 (0) 20 7121 0102

India Operations Headquarters

#5/535 Old Mahabalipuram Road
Okkiyam Pettai, Thoraipakkam
Chennai, 600 096 India
Phone: +91 (0) 44 4209 6000
Fax: +91 (0) 44 4209 6060

Codex 4815