White Paper

Key Considerations When Operationalizing an Artificial Intelligence Strategy

Sponsored by: NVIDIA and Digital Realty

Ashish Nadkarni Peter Rutten Sriram Subramanian Shane Rau

September 2021

IDC OPINION

Businesses are quickly realizing the potential for achieving competitive differentiation by introducing artificial intelligence (AI) technologies into their organizations. Externally, they can achieve this competitive differentiation via products and service enhancements. Internally, they can achieve this differentiation via streamlined and efficient business operations. In the end, it is all about time to value of insights creating top-line and bottom-line impact on the business. Not acting quickly and sufficiently on AI investments is a risk that no business can afford to take.

Investing in AI is quite different from procuring and deploying other enterprise off-the-shelf products or services. In fact, it is a highly bespoke endeavor that requires data scientists to work in unison with business analysts, data engineers, DevOps and IT, and app developers to define and implement an AI strategy and launch AI initiatives. Data scientists can develop AI models that map to specific business problems or imperatives; developers with expertise in building custom and cloud-native apps incorporate those models in a production application. So far so good.

Operationalizing AI is another story altogether. Two recent IDC studies provide some insight. One found that more than 30% of the respondents cited a failure rate of two-thirds for their AI projects (source: IDC's *AI Strategies View,* April 2021). Another found that 80% of the respondents cited an average duration of three months to one year spent on building an AI model for deployment. Further, these respondents spend up to a year preparing a completed model for deployment (source: IDC's *AI Infrastructure View,* August 2021). The reason for these outcomes is that most businesses do not develop AI apps in the same manner as conventional enterprise apps. That can present a problem during production rollout.

A suitable plan for operationalizing AI can minimize tensions and disconnects between information technology (IT) operations, data scientists, and developers. It requires organizations to:

- Gain full understanding of critical capabilities required to succeed and especially dealing with challenges associated with managing the life cycle of AI apps and data.
- Invest in innovative software technologies to cross the chasm between developers, data scientists, and IT operations teams.
- Invest in purpose-built infrastructure that can scale to support the burgeoning compute and data persistence requirements of AI apps.
- Implement a process and methodology such as MLOps that fosters collaboration between the business and technical teams, as well as between the development and operations teams.

Businesses often give up on AI initiatives because of failures that stem from a lack of understanding of infrastructure requirements. Organizations must also understand that successful AI initiatives must factor in the influence of data gravity on deployment choices. Meeting the objective of making AI ubiquitous across a global business means that public cloud (based infrastructure) is not always the best deployment option. To eliminate the barriers to a broad and secure AI deployment, businesses must invest in the right infrastructure stack and, importantly, in the right hybrid IT strategy.

SITUATION OVERVIEW

Operationalizing Artificial Intelligence

Businesses have acknowledged that strategic investments in artificial intelligence technologies are key to their future. AI capabilities provide competitive advantage to enterprises through new business models and digitally enabled products and services. They enable businesses to improve user experience, increase productivity, and innovate for the future. There is little doubt that AI apps and algorithms unlock new opportunities that are impossible to achieve with traditional approaches.

AI app development does not follow standard or well-accepted norms found within traditional enterprise app development. AI app development is highly custom in nature. Even when off-the-shelf products exist, businesses must customize them for their specific use case and with desired outcomes in mind. Businesses are more evolving in their AI adoption maturity and are reaching the next stage of AI development where they are deploying AI at production scale and continuously iterating and evolving their production models. However, they are quickly realizing – often the hard way – that introducing AI into production and integrating it with business processes and workflows is not as simple and straightforward as it sounds.

Operationalizing AI is a technology issue and a process issue. More importantly, it is an organizational issue. A common challenge that organizations find when they operationalize an AI strategy is lack of understanding of the critical capabilities required to succeed, and especially dealing with challenges associated with managing the life cycle of AI apps and data. This circumstance means organizations must also invest in an agile process that fosters collaboration between the business and technical teams, as well as between the development and operations teams.

From a technology perspective, there are four areas of investments that organizations must make as part of their AI strategy. They are:

- Software technologies and platforms that deliver base functionality for downstream app developments
- Software technologies that help cross the chasm between developers, data scientists, and IT operations teams
- Purpose-built infrastructure that can scale to support the burgeoning compute and data persistence requirements of AI apps
- Deployment locations for infrastructure to enable ubiquitous consumption and insights across the entire organization

From a process and methodology perspective, the focus of AI life-cycle management is shifting to the operational aspects of the AI model life cycle. Customers need to scale their AI operations, which include workflow collaboration, model prototyping and training, model deployment, model performance evaluation, and ongoing model monitoring. They are also increasingly looking at operationalizing

model life-cycle management through capabilities including model repositories, model diagnostics, model feature store, model delivery, model governance, and model fairness. The deployment of MLOps requires a supporting infrastructure stack that provides MLOps at scale and enables data-labeling capabilities at scale through fully automatic or semiautomatic mechanisms.

Leveraging Infrastructure for the AI-Infused Enterprise

Infrastructure is one of the most misunderstood and underestimated parts of any AI stack. IDC research shows that in the early days of AI, little consensus existed among organizations as to what type of infrastructure would be most suitable for running their AI workloads. As a result, they tried everything and often ended up with unsatisfactory results. For cloud service providers (SPs) and hyperscale datacenters that were implementing AI functionality, the infrastructure was less of a question mark — they leveraged their general scale-out approach to compute and build vast scale-out AI processing compute environments. For enterprises, the use of general-purpose infrastructure at a limited scale can lead to AI initiatives failing in production early.

As organizations are becoming increasingly savvy with developing and running AI, the size and complexity of their AI models have begun to grow dramatically faster than the capabilities of available compute. Workload accelerators, primarily general-purpose graphical processing units (GP-GPUs), have already made their entry, and their use is accelerating, especially for training AI models. More and more businesses now say that they have an enterprisewide AI strategy in which they maximize efficiency across the organization. In addition, they are increasingly running multilayered machine learning (ML) in production, including recurrent, recursive, and convolutional neural networks (CNNs) as well as unsupervised pretrained networks. This momentum of AI in production has developed into an opportunity for higher-end infrastructure stacks to ramp up in the AI training and inferencing markets.

The amount of compute that AI demands is proving to be insatiable. During the AI development, and especially during the model training stage, businesses must find performant infrastructure, more so for ever larger and more complex AI models that power modern enterprises while allowing data scientists to iterate as often as needed without wasting time waiting for training runs to complete.

IDC is seeing enterprise infrastructure for AI model training evolve from standalone accelerated bare metal servers to increasingly high-performance computing (HPC) like tightly connected server clusters. The most popular AI models, like those for natural language processing, may consist of tens of billions of parameters, which means that, just as with modeling and simulation (HPC) workloads, performance is key. IDC calls this implementation a massively parallel computing (MPC) architecture, an emerging computing platform and data management architecture that relies on massive parallelization for processing large volumes of data or executing complex instruction sets in the fastest way. Note that MPC is one technology approach within what IDC calls performance-intensive computing (PIC). PIC combines infrastructure approaches across three major and fast-growing use case categories: modeling and simulation (M&S), artificial intelligence, and Big Data and analytics (BDA).

The demand for more compute is certainly important. Increasingly, organizations are looking for not just the hardware but a complete AI infrastructure stack that combines server hardware, hardware abstraction layers, orchestration layers, AI development layers, and data science layers that seamlessly operate together. This demand has resulted in an explosion of different AI stacks from not just server and storage OEMs but also processor and accelerator manufacturers and other players in the market.

As businesses deploy newly developed AI models into production, the demand for AI inferencing compute is taking off as well. The AI model being inferenced needs to follow close functional coordination with other enterprise applications. Accordingly, this scenario also requires careful consideration and selection of the various infrastructure options, which are not necessarily the same as those for training the model. Lighter and/or virtualized accelerators may well carry the inferencing workload more than sufficiently.

Most important in both cases — AI training and inferencing infrastructure — is that the organizational AI initiatives do not become a costly endeavor with disparity between the investment and the return on that investment.

Delivering Intelligence Anywhere

A key objective for developers and scientists to infuse AI into the business workflow is that the insights will emerge from any stage within them. This fact means that it is essential to take the compute to where the data is collected or stored, thus compressing the time to value of insights. Any business that has tried the converse (i.e., to move data to where compute is) knows very well that data gravity can make gaining insights an arduous endeavor. Common issues include escalating costs and complexities associated with data transit and storage. Like infrastructure, accepted operating models or deployments like public cloud often do not work for AI initiatives.

Businesses that wish to embrace AI globally and gain insights consistently across all their business units and locations often struggle with distributed AI deployments and especially with a cloud-first strategy. With public cloud services, businesses face two primary challenges:

- The provider may have only a single region or zone servicing the data source or locality. Local laws or network limitations may prevent businesses from moving the data from edge locations to that region in a timely manner.
- The provider may not offer the right compute instances or support the right infrastructure stack for running for AI training or inferencing apps in that region. This may lead to subpar performance across deployments.

In short, data gravity when coupled with the limitations of a cloud service provider can present challenges for an organization's intelligence anywhere strategy.

Enabling intelligence requires the right infrastructure in the right place and in the right deployment. Crucially, it works best in a hybrid environment that bridges together various deployments into a common operating model. An important pillar of a hybrid operating model for AI deployments is a partnership with a leading multitenant datacenter provider. Organizations benefit from the provider's capabilities such as:

- A global footprint, which distributes infrastructure closer to the data source or locality
- An ecosystem of leading infrastructure partners that enables the right size and type of infrastructure stack deployed at the appropriate location
- An edge infrastructure network that enables primary data aggregation and analysis, thus compressing time to value
- An edge-to-cloud network that enables data uploads to a public cloud service or central datacenter for secondary analysis

In short, the capabilities of a multitenant datacenter provider can truly enable scaling of an organization's AI implementation. The customer can start small and add capabilities as they introduce new models or increase the breadth or their AI deployment.

Automated machine learning capabilities have made it easier and faster for data scientists and business users to create customized machine learning models as part of bigger AI initiatives. This is made easier by distributed AI infrastructure deployments, which support a ubiquitous process and workflow offering consistent data preparation, workflow collaboration, automated machine learning, and model deployment capabilities to globally distributed teams.

THE DIGITAL REALTY AND NVIDIA SOLUTION

Digital Realty is partnering with NVIDIA as a DGX-Ready colocation provider to enable enterprises to seamlessly deploy globally distributed AI infrastructure. Businesses can access Digital Realty's PlatformDIGITAL® Data Hub solution (which also includes access to NVIDIA's DGX POD™ platforms) across Digital Realty's datacenter footprint of 290+ datacenters worldwide. This partnership is significant because it brings together the synergies of two market leaders (and their ecosystems) into a joint solution:

- Digital Realty supports the datacenter, colocation, and interconnection strategies of customers across the world. Its customers include cloud and information technology service providers, communications and social networking service providers, and enterprises in the financial services, manufacturing, energy, healthcare, and consumer products industries.
- NVIDIA is an enterprise infrastructure provider with deep expertise in AI technologies. Its AI solutions stack based on DGX accelerated computing systems is the validated standard in the industry in terms of performance, scaling, and compressed time to value for AI initiatives powered by a full stack platform that is purpose built and optimized for the unique demands of AI development.

NVIDIA DGX System

NVIDIA has designed the DGX system as a building block for AI infrastructure. The result is that IT organizations can select the DGX A100 system for performance-intensive computing use cases. These systems integrate eight of the world's fastest datacenter accelerators — the NVIDIA A100 Tensor Core GPU — in combination with 2nd Gen AMD EPYC™ processors and are offered as certified reference architecture systems that support scaled infrastructure under the NVIDIA DGX POD™ and NVIDIA DGX SuperPOD™ brands.

NVIDIA takes a "silicon up" approach with its integration. A key differentiator for NVIDIA is the software optimization layer that it bundles with the DGX System. In other words, for NVIDIA the differentiation lies in software, just as much as it does in silicon, including its AI libraries, SDKs, optimized frameworks, and prebuilt models — all of which combine with the DGX hardware platform to deliver highly optimized outcomes for the stated use case. NVIDIA does not directly seek to monetize the software layer but instead places bets on customers procuring the entire system and operationalizing it to support enterprisewide AI. NVIDIA is also clear in its mission in which its systems deliver differentiated outcomes for specific performance-intensive use cases such as AI using the stack certified by NVIDIA and its partners.

Digital Realty PlatformDIGITAL®

Digital Realty's PlatformDIGITAL® is a global datacenter platform designed to host critical infrastructure and interconnect digital ecosystems. It provides a foundation for businesses investing in a digital-first strategy, of which distributed AI is a core component. It is meant for businesses for which the increasing demands of AI and AI-infused apps require a new global datacenter partner that can:

- Solve global coverage, capacity, and ecosystem connectivity needs
- Tailor infrastructure deployments and controls matched to business needs irrespective of datacenter size, scale, location, configuration, or ecosystem interconnections
- Operate deployments as a seamless extension of any global infrastructure with the consistent experience, security, and resiliency business demands
- Enable global distributed workflows at centers of data exchange to remove data gravity barriers and scale digital business

PlatformDIGITAL® assists businesses in solving their data gravity challenges and thus scale the business digitally. It is the physical implementation of a Pervasive Datacenter Architecture (PDx™) Strategy, which is itself a step-by-step approach for developing a new decentralized IT infrastructure architecture for digital businesses that are investing in AI. It enables the business to:

- Plan centers of data exchange zones to remove barriers of data gravity
- Deploy purpose-built datacenter footprints tailored to business needs
- Interconnect to digital ecosystems
- Control and deliver a consistent global operational experience
- Achieve specific performance, resiliency, and security requirements

The PlatformDIGITAL® Solution Model is the physical implementation of the PDx™ Strategy. It is a single global datacenter platform that is meant to solve global coverage, capacity, and connectivity needs of AI development, by bringing users, networks, clouds, controls, systems, and things to the data, which removes barriers of data gravity, creates centers of data exchange to accommodate distributed workflows, and scales digital business. PlatformDIGITAL® comprises four solution areas:

- **Network Hub,** which consolidates and localizes traffic into ingress/egress points to optimize network performance and cost
- **Data Hub,** which localizes data aggregation, staging, analytics, streaming, and data management to optimize data
- **Control Hub,** which hosts adjacent security and IT controls to improve security posture and IT operations
- **SX Fabric,** which adds software-defined networking (SDN) overlay to service chain multicloud and B2B application ecosystems (It connects hubs across metros and regions to enable secure and performant distributed workflows.)

PlatformDIGITAL® Data Hub with NVIDIA DGX

Businesses can direct their IT organizations to access PlatformDIGITAL® Data Hub (with access to NVIDIA's DGX POD™) making it a turnkey bundle that includes the necessary components and services to rapidly plan, deploy, and scale AI infrastructure. The preconfigured solution accommodates a typical enterprise deployment of a Data Hub to solve placement, connectivity, and hosting of critical data infrastructure in proximity to end users, networks, public and dedicated clouds, and endpoints (which includes IoT devices). It enables enterprises to support their AI workflow, optimizing the placement of data and compute at global points of business presence.

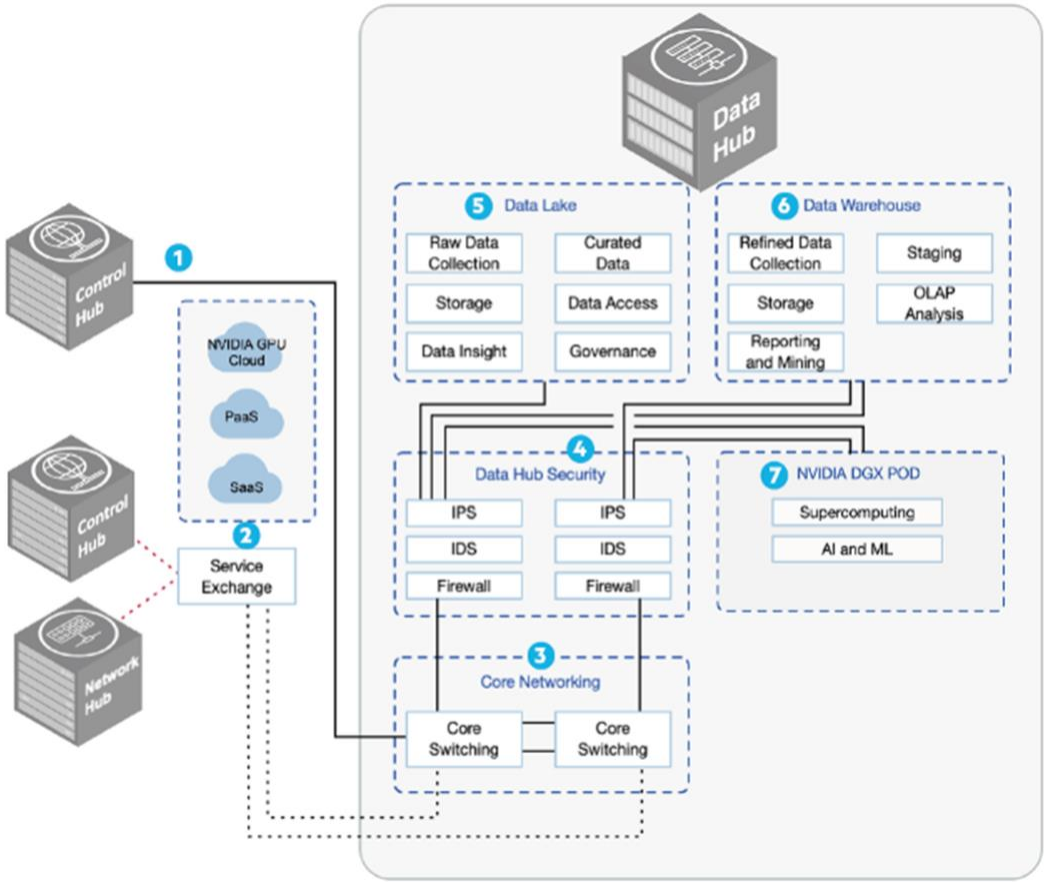
For Digital Realty's customers, the NVIDIA DGX POD™ eliminates the design complexity and lengthy deployment cycle associated with AI infrastructure, giving data scientists and developers the most powerful tools for AI exploration in a platform that scales. Now enterprises can dramatically shorten development and deployment times, and thus, the time to insight from data.

Figure 1 illustrates the Data Hub reference architecture. Key elements of the architecture are:

- The Data Hub is placed near the Control Hub and connects using a Campus Connect or Metro Connect.
- A Control Hub is housed in another metro location and connects back to the Data Hub using Service Exchange. Trusted data from Network Hubs flows to the Data Hub for further analysis and modeling. Service Exchange enables virtual cloud connectivity to the cloud resources supporting hybrid IT, including NVIDIA GPU Cloud.
- The Core Switching infrastructure terminates connectivity into the Data Hub and enables access to the cloud for deep analytics and archival storage.
- Strict controls and logging mechanisms maintain value and sensitivity of enterprise data access.
- Data scientists and developers analyze and curate raw data ingested into the Data Lakes.
- Business professionals and analysts can then access the refined data transferred to the Data Warehouse.
- The NVIDIA DGX POD™ is located directly adjacent to the data store for direct access and enables rapid iteration on AI models and scaling of AI production-ready applications.

FIGURE 1

Digital Realty's PlatformDIGITAL® Data Hub Reference Architecture



Source: Digital Realty, 2021

In summary, the PlatformDIGITAL® Data Hub brings together industry-leading AI infrastructure with a platform that simplifies and speeds innovation across the enterprise.

CHALLENGES/OPPORTUNITIES FOR DIGITAL REALTY AND NVIDIA

The biggest opportunity for Digital Realty and NVIDIA lies in helping their customers make AI ubiquitous, globally. The challenge for Digital Realty and NVIDIA lies in articulating the return on investment of a turnkey solution vis-à-vis other approaches, which may appear cheaper at first but end up costing the business dearly. No one wants their AI initiatives to fail because they cut corners early on during the design and development phase. Thus far, enterprises have struggled to scale AI on three fronts, all related to infrastructure:

- **First, underestimating the need for purpose-built and optimized infrastructure stack for performance-intensive use cases, and specifically AI training and inferencing.** This chain of events starts at the development stage wherein developers and data scientists start the journey with general-purpose infrastructure with a piecemeal or self-built stack. At this stage,

the model is not complex, or it has dummy data with size and complexity that is not representative of real-life scenarios. By the time this initiative moves into the production stage, it is too late, and it fails to deliver the stated business outcomes.

- **Second, not realizing the importance of data gravity when it comes to scaling AI deployments.** Businesses often make the mistake of assuming that a public cloud deployment by way of its global nature is the best deployment model for a global and distributed AI rollout. The fact of the matter is that AI workloads — unlike other workloads — are all about compressing time to value of insights and, that in turn, relies on taking optimized compute as close as possible to the data source or locality.
- **Third, deciding to go it alone. AI — unlike other business initiatives — is a people, process, and technology challenge.** It is further complicated by the fact that approaches to AI development and deployment do not follow well-established norms of mainstream enterprise applications. AI initiatives are no longer garage projects, and partners like Digital Realty and NVIDIA can assist customers in gaining consistent and reliable outcomes as they roll out these initiatives into production.

NVIDIA and Digital Realty by way of their partnership can jointly educate their customers on the importance of an optimized outcomes-based solution like PlatformDIGITAL® Data Hub. They can assist IT and line-of-business decision makers in making the case for an end-to-end solution from trusted partners that are also leaders in their respective markets. Finally, they can assist customers in doing a global rollout, complete with integration of various data sources into the stack.

ESSENTIAL GUIDANCE FOR IT DECISION MAKERS

A recent IDC survey found that only about 25% of the surveyed organizations have an enterprisewide AI strategy and about 50% of enterprises struggled in making AI a top priority. They cited lack of skill set, concerns around ethics and bias, laggard regulations, and unrealistic expectations as adoption inhibitors. It is noteworthy that respondents did not cite infrastructure as an area of concern, which is a mistake and goes to prove IDC's assertion that organizations often make the mistake of deploying AI on general-purpose infrastructure.

IDC recommends that organizations start by implementing an enterprisewide AI vision, which then forms the basis of an "AI first" business and technology strategy. They ought to examine ways to make AI part of their core product and services development workflows and to streamline business operations. This approach must serve as a guide for businesses on the nature and scale of investments they need to make, and this, in turn, dictates the investments in AI infrastructure stacks and vendor partnerships. Further:

- **First, IT and business decision makers must start by accepting that AI software development does not follow norms of other business apps.** They must approach this situation not just as a technology investment but also as a people (skill set) and process (workflow) investment. The good news is that the landscape for AI software and apps is rapidly democratizing, leading to more off-the-shelf options for AI, which is happening through rapid proliferation of AI platforms and services. This democratization will further catalyze the growth of AI-enabled enterprise workloads, and businesses can start now by making changes to their organization.

- **Second, IT decision makers must empower their IT staff to investigate a purpose-built and optimized infrastructure stack that best suits their company's AI strategy.** Working in collaboration with developers and data scientists, they must build an on-ramp for a gradual increase in investments, transitioning from a suitable infrastructure for development and testing to one that can scale for production deployments.
- **Third, IT and business decision makers must carefully reconsider their choice of deployment model and location for their AI infrastructure.** Businesses that would otherwise be cloud first or cloud only must weigh the impact of data gravity and the associated benefit of hybrid IT infrastructure (i.e., investing in a distributed stack located at datacenters operated by a multitenant datacenter provider with a global footprint).
- **Fourth, they should take their time to evaluate the buy versus build approach.** Decision makers must weigh their use cases, in-house skill sets, and custom versus off-the-shelf culture and appetite for infrastructure investment and choose the consumption model accordingly.
- **Finally, fifth, deciding to go it alone.** IDC recommends that businesses partner early on in their AI journey with a suitable infrastructure stack provider and a partner that provides a solution based on that infrastructure provider's stack. It can make an enormous difference with predictable outcomes especially as the implementation scales, gains more complexity, and becomes a core part of the business itself.

CONCLUSION

AI is one of the most consequential and disruptive sets of technologies that humankind has developed in recent times. For companies embarking on a digital transformation journey, the question of adopting AI technologies and infusing them into their business processes is not that of an "if" but a "when." Unfortunately, AI technologies — unlike other enterprise technologies — are rarely a buy and deploy in nature. They require organizations to embrace new skills, new processes, and new deployment models as part of an overall AI strategy, which is itself borne out of a vision to gain unprecedented business differentiation via existing as well as new products and services. Choosing the right infrastructure stack, the right deployment model, and — crucially — the right partners are important decisions for businesses to make when embracing an AI-first strategy.

About IDC

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

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com