

White Paper

Harnessing the Potential of Generative AI in the Enterprise

Sponsored by: Quantiphi

Ritu Jyoti Jennifer Hamel

September 2023

INTRODUCTION

The world economic outlook is uncertain amid financial sector turmoil, high inflation, geopolitical crises and increasing labor shortages. In many developed economies, large numbers of employees from the outsized Baby Boomer generation have retired in recent years while members of the Millennial generation have moved past entry-level roles. These shifts have created talent shortages in the comparatively small Generation X and Generation Z cohorts at the leadership and entry levels, respectively. Labor economists predict such shortages will be a permanent market condition, persisting well into the late 2030s or beyond.

Enterprises around the world are confronting fractured business plans. As per IDC's *Future Enterprise Resiliency & Spending Survey - Wave 2*, March 2023, concerns about an imminent recession continue to swirl but declined slightly compared to the last three months of 2022. IT leaders remain very cautious when it comes to IT spending in 2023, but the effect is more a slowdown in the rate of spend increase or a freeze, rather than outright reductions.

IDC believes that IT leaders in digital businesses will rapidly adapt to these unsettled conditions by transforming into a resilient digital business where value creation is based on the effective use of fast-evolving and innovative technologies. Artificial intelligence (AI) is at the forefront of digital disruption. Most AI systems so far have been used for classification and predictions. Generative AI's ability to be creative has the potential to be a major game changer for businesses. This technology is a branch of computer science that uses unsupervised and

IDC defines a digital business as an organization where value creation is based on the use of digital technologies, including internal and external processes; how an organization engages with customers, citizens, suppliers, and partners; how it attracts, manages, and retains employees; and what products, services, and experiences it provides.

semi-supervised algorithms to enable computers to create new content using previously created content, such as text, audio, video, images, and code, in response to short prompts.

The application of generative AI in enterprises is just starting to unfold along with a world of possibilities. Enterprises can leverage generative AI to drive innovation, automate repetitive tasks, improve decision-making, personalize customer experiences, and boost efficiencies. As such, businesses that can effectively leverage technology are likely to gain a significant competitive advantage.

Generative AI Opportunities

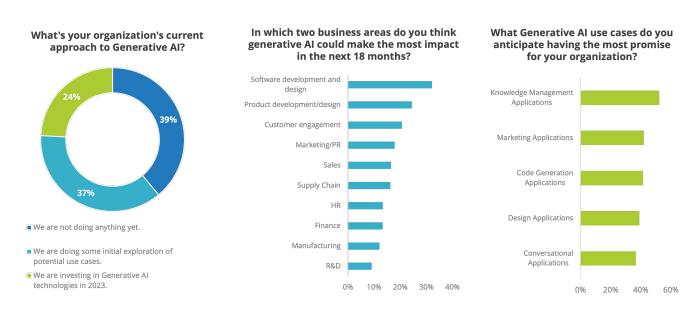
Generative AI powers a class of machine learning models that are trained on diverse data sets and can be adapted or fine-tuned for a wide range of downstream tasks. The large-scale model revolution is driven by the emergence of the transformer model architecture in 2017. This architecture is an improvement on the limitations of recurrent neural networks (RNNs). Transformers can take into account the sequential nature of data while also remembering distant words. It is also parallelizable for training, enabling researchers to leverage GPU improvements.

As we know, it is the early days for generative AI adoption. As per IDC's *Future of Enterprise Resiliency Survey*, March 2023, about one quarter of the respondents globally noted that they are investing in generative AI technologies in 2023, 37% are doing some initial exploration of potential use cases, and the remaining 39% are not doing anything yet (see **Figure 1**).

When asked about the use cases with the most promise for the technology, **knowledge management applications were number one**, **followed by marketing and code generation applications**. Software development and design and product development and design are noted as the top two business areas where generative AI could have the most impact over the next 18 months.

FIGURE 1

Generative Al Use Cases and Investments: Worldwide



Source: Future Enterprise Resiliency & Spending Survey Wave 2, IDC, March, 2023. n= 952

Knowledge management applications employ large language models (LLMs) as a means for managing text-based, image, and video-based knowledge within an organization. This could significantly reduce the labor intensiveness involved in creating structured databases, which has been extremely difficult for many large companies. Marketing focused versions of LLMs can produce blogs, social media posts, web copy, sales emails, ads, and other types of customer-facing content. Text to image models can be used to design and simulate products. GPT-3 has proven to be an effective, if imperfect, generator of computer program code. GPT-3's Codex program (a machine learning model) is specifically trained for code generation and can produce code in a variety of different languages. Code generation applications could also be used by IT operations engineers for Infrastructure as Code (IaC), which is the managing and provisioning of infrastructure using code instead of manual processes.

Generative AI has applications across many industries:

- Healthcare: Generative AI can be used to improve physicians' workflows. Clinicians spend excessive time reviewing voluminous electronic health records (EHRs), which impedes doctor-patient relationships and drives burnout. Generative AI is good at summarization, so could be used to review the medical record and provide a succinct summary relevant to the patient and the treating physician at the point of care. Generative AI can also be used to design proteins and drug discovery.
- Financial Services: Having a clear view of the future is vital for making accurate forecasts and optimization decisions. However, this is made substantially more difficult for events that are rare, or where limited data is available. Financial institutions are using generative AI to create synthetic data for rare market events. This enables them to improve the accuracy of price forecasts and make more confident decisions around portfolio optimization.
- Engineering: Generative design leverages artificial intelligence and machine learning to turn tedious engineering design processes into a sophisticated yet natural interaction between the computer and the engineer. The majority of the topology optimization and simulation is automatically performed by the computing unit. In addition, negative feedback loops are cut short by lowering the barriers to design. As a result, the engineer has more room to tackle challenges that require common sense or that cannot be solved by computers.

Generative AI Challenges

While generative AI holds huge promise for enterprises, it is important to acknowledge that adoption currently has limitations and challenges, and organizations need to embrace it responsibly with appropriate guardrails. AI ethics have been discussed extensively in the past, but the ethical discussion around generative AI is relatively new.

Some of the key generative AI issues include:

Accuracy

LLMs learn complex patterns and relationships between words and phrases using advanced algorithms and architectures. Using this knowledge, they predict the most suitable word or phrase to continue in a given context. However, sometimes these predictions can be erroneous, resulting in unexpected outcomes or "hallucinations" Where the generated text is semantically or grammatically correct but nonsensical in the conversation's context.

Pre-trained large language model applications like ChatGPT are not dynamic. ChatGPT can only fetch data prior to the year 2021 as its training stopped in the year 2021.

According to the Stanford University *Artificial Intelligence Index Report 2022* (TruthfulQA benchmark test), on average, most generative models are truthful only 25% of the time. Before utilizing generative AI tools and products, organizations and individuals must independently assess the truthfulness and accuracy of their generated information.

Misuse

Generative AI could produce misleading, harmful, or inappropriate content in any context. In the educational context, generative AI could generate false or misleading information and present it as fact. Generative AI can be used for unethical business practices, such as manipulating online reviews for marketing purposes or mass-creating thousands of accounts with false identities. Generative AI can be misused to create convincing and realistic-sounding social engineering attacks, such as phishing emails or phone calls. These attacks could be designed to trick individuals into revealing sensitive information, such as login credentials or financial information, or to convince them to download malware.

Deepfakes can be used to generate synthetic media, such as images, videos, and audio. Such Al generated content can be difficult or impossible to distinguish from real media, posing serious ethical implications. Such media may spread misinformation, manipulate public opinion, or even harass or defame individuals.

Biases

Large language models enable human-like speech and text. However, recent evidence suggests that larger and more sophisticated systems are often more likely to absorb underlying social biases from their training data. These Al biases can include sexist, racist, or ableist approaches within online communities. As per the Stanford University *Artificial Intelligence Index Report 2022*, compared to a 117 million parameter model developed in 2018, a 280 billion parameter model created lately demonstrated an enormous 29% increase in toxicity levels. As these systems evolve into even greater powerhouses for Al research and development, there is the potential for increased bias as well.

Copyright Ambiguities

An ethical concern around generative AI is the ambiguities over the authorship and copyright of AI generated content. Generated data can be used for training machine learning models. However, the use of copyrighted generated data in compliance with fair use doctrine is ambiguous. While fair use generally accepts academic and nonprofit purposes, it forbids commercial purposes.

Generative AI outputs may be considered derivative work, which is new work based on someone else's original idea. Anyone can create derivative works, whether they have permissions from the original creator or not. Copyright infringement occurs when using or copying another person's work without permission or attribution which has legal consequences.

Data Security and Privacy

Due to its ability to generate fake photos and images closely resembling the real thing, generative Al may increase identity theft, fraud, and counterfeiting cases. Data privacy issues can arise in highly regulated industries if generative Al involves collecting private information about individuals.

Lack of Skills

Organizations will need new roles (prompt engineers) for prompt tuning, which is a new and rapidly evolving way to use the fixed weights of a large, pre-trained language model for many different tasks. (See the next section for more on prompt tuning.)

Building a Generative AI Enterprise Solution

Technology suppliers and model providers are subscribing to the responsible use of generative Al and are helping developers use best practices by providing such tools as free content filtering, end-user monitoring to prevent misuse, and specialized endpoints to scope API usage.

The latest class of generative AI systems have emerged from foundation models—large-scale, deep learning models trained on massive, broad, unstructured data sets (such as text and images) that cover many topics, making them easier to customize for tasks like analyzing legal contracts or detecting fraud in financial documents.

Developers can adapt the models for a wide range of use cases, with little fine-tuning required for each task. For example, GPT-3.5, the foundation model underlying ChatGPT, has been used to translate text, while scientists used an earlier version of GPT to create novel protein sequences. In this way, the power of these capabilities is accessible to all, including developers who lack specialized machine learning skills and, in some cases, people with no technical background. Using foundation models can also reduce the time for developing new AI applications to a level rarely possible before.

Until recently, fine-tuning by adjusting parameter values was the best way to redeploy one of these pretrained models for specialized tasks. One gathered and labeled examples of the target task and fine-tuned the model instead of training an entirely new model from scratch.

Of late, a simpler, more energy-efficient technique has emerged as foundation models grow relentlessly larger: prompt-tuning. In prompt-tuning, the best cues, or front-end prompts, are fed to the AI model to give it task-specific context. The prompts can be extra words introduced by a human or AI-generated numbers introduced into the model's embedding layer. Like crossword puzzle clues, both prompt types guide the model toward a desired decision or prediction. Prompt-tuning allows a company with limited data to tailor a massive model to a narrow task. It also eliminates the need to update the model's billions (or trillions) of weights, or parameters.

Redeploying an AI model without retraining it can cut computing and energy use by at least 1,000 times, saving thousands of dollars. Prompt-tuning originated with large language models but has since expanded to other foundation models, like transformers that handle other sequential data types, including audio and video. Prompts may be snippets of text, streams of speech, or blocks of pixels in a still image or video. One drawback of prompt-tuning is its lack of interpretability. The AI discovers prompts optimized for a given task but cannot explain why it chose those embeddings. Like deep learning models themselves, soft prompts are opaque.

Role of Professional Services Firms in Enterprise Adoption of Generative Al

With rising public awareness of generative AI capabilities, including the ability to interact with free, web-based chatbots and image generators, organizations are feeling pressure to move faster to get up to speed on generative AI, or risk being left behind by competitors. **Professional services firms remain**

a critical source of expertise, skills, and tools to incorporate AI into digital business strategies and build production-grade solutions that can be delivered at scale. While generative AI seems to have burst on to the scene only in recent months, many services firms have been experimenting with the technology for years, both internally and with clients. Services firms can act as trusted advisors to help organizations cut through hype and hysteria, set reasonable expectations for what generative AI can and should do for their businesses, and develop roadmaps for adopting and managing generative AI solutions across the enterprise.

Advice

Organizations encounter a wide spectrum of challenges in their AI adoption journeys that often require expert advice to navigate. Services firms can assess an organization's AI maturity, readiness, talent, and data needs and assist with creating or refining AI strategies and operating models to achieve specific business objectives. Even organizations with previously established AI programs may find their strategies and governance frameworks need adjustment to consider new implications (ethical, regulatory, or otherwise) of generative AI capabilities and to incorporate appropriate guardrails for developing and using the technology.

Services firms also work with organizations to discover and prioritize use cases, define KPIs for measuring business value, and guide innovation around generative AI solutions. Here, industry and functional domain knowledge gained from experience working with many different customers can accelerate the process of identifying and developing high-impact generative AI use cases.

Services firms also advise on technology selection and reference architectures, which is particularly important given the rapid proliferation of generative AI foundation models and applications in the market, whether open source or commercial. Given the high costs of training foundation models, organizations will need to conduct cost-benefit analyses to determine the appropriate technology solution for their goals and available resources.

Finally, rolling out generative AI solutions at scale will almost certainly require some reengineering of business processes, as well as training and change management for employees who will be interacting with a new and disruptive technology.

Execution

Once a generative AI use case is defined, services firms can assist to bring a solution to fruition through design, implementation, and managed services. The promise of the latest generative AI foundation models is that they have already been trained on a wide variety of data sets and topics and so therefore are easier to customize for particular use cases.

Services firms can support a variety of deployment scenarios, including implementing an off-the-shelf generative AI application from an ISV, designing a custom application around a pretrained foundation model, integrating generative AI capabilities with enterprise intelligence and automation systems through APIs, and prompt-tuning or, if required, fine-tuning foundation models for specialized tasks.

Regardless of how a generative AI application is developed, deploying solutions at enterprise scale will require expertise in a variety of areas common to all AI solutions, including:

- Data ingestion, preparation, and integration
- Platform engineering and implementation

- Model performance monitoring (including human-in-the-loop evaluation of model output and recommendations)
- Ethical and responsible Al adherence
- Infrastructure and cloud management

Organizations differ in how they choose to invest in internal vs. external capabilities to manage Al solutions on an ongoing basis, but many services firms offer flexible, "factory"-style delivery models and Al centers of excellence to provide support as needed.

Accelerators

The fundamental value that professional services firms offer for enterprise adoption of generative AI is in helping customers achieve ROI more quickly than they would on their own. Many services firms have invested heavily in IP, partner ecosystems, and in-demand skills over the past several years in preparation for the inflection point that recent advances in generative AI have brought about in the market.

- IP: Proprietary assets can include pretrained industry- or function-specific models, reusable component repositories, curated and annotated training data sets, developer tools and microservices, and even full-fledged products and platforms. These assets can fill gaps in commercial software products, address specific business domain or technical challenges (such as integrating legacy enterprise systems with new AI capabilities), or industrialize AI solution development and management.
- Partners: Services firms also work closely with an ecosystem of partners, including
 foundational model providers, public cloud platforms, enterprise application and source data
 system providers, third-party data providers, and even academia to access and drive
 innovation that benefits their customers and moves the industry forward.
- Skills: Services firms offer a wide range of skills that organizations need for successful generative AI adoption beyond data science, including digital business strategy, solution architecture, data engineering, platform engineering, technical development, prompt engineering, human-centric UI/UX design, agile innovation, ethics and behavioral science, security and privacy, and legal and compliance expertise. While it is theoretically possible for organizations to build each of these skills internally, collaboration with one or more external services firms will likely be both necessary and more cost-effective to execute on generative AI initiatives.

Considering Quantiphi for Generative Al

Introduction to Quanitiphi's Generative Al Platform

Quantiphi's 'baioniq' is a generative AI platform focused on enhancing the productivity of knowledge workers. The baioniq platform leverages a combination of open source and proprietary foundation models from partners like NVIDIA and Google Cloud that are instruction fine-tuned and domain-adapted to build generative AI applications trained on customer data.

Baioniq offers an enterprise-ready and modular generative Al platform that is designed to help organizations fine-tune LLMs and leverage them to automate domain-specific tasks for boosting the

productivity of their knowledge workers. For example, it provides tools that enable users to retrieve up-to-date information across unstructured data objects such as text, images and tables from their proprietary content management systems in a ChatGPT-like conversational interface.

The platform addresses present issues of LLMs, such as factual inconsistencies, lack of domain understanding, and misalignment with responsible Al principles, by:

- Performing supervised fine-tuning of LLMs on their proprietary data to ensure that the generative responses are specific to their unique domains.
- Performing instruction fine-tuning on LLMs on a broad range of enterprise-specific tasks.
- Generating responses that are grounded in facts while maintaining referenceability to the source of truth.
- Aligning to responsible and ethical Al principles by continuously filtering out unwanted responses through reinforcement learning.

Why Quantiphi

Harnessing the Power of Al Ecosystem

As a professional services provider, Quantiphi operates as vital connective tissue within the Al ecosystem. The company helps customers derive value and realize transformation opportunities by minimizing friction points and enabling access to the ecosystem. Quantiphi has invested in co-development of tools and solutions with a variety of partners, stakeholders, and clients. The company has strategic partnerships with Google Cloud, Nvidia, AWS, Snowflake, and Databricks.

Google is a pioneering contributor to the growth of the global AI space with several early and crucial developments. Since the 2013 ImageNet breakthrough and the introduction of the transformer architecture in 2017, the company has worked to mainstream deep learning, introducing several key AI offerings such as Vertex AI, purpose-built TPUs, the PaLM2 model, Generative AI Studio, and more. Google offers a vertically integrated generative AI technology stack from the hardware to the service API level.

Quantiphi is positioned as a strategic Google Cloud partner with access to Google Cloud's generative Al suite of services, and has achieved CCAI, ML, and Analytics specializations. Named the Breakthrough Partner of the Year for North America in 2021, Quantiphi recently partnered with Google Cloud to develop generative Al features for domain-specific knowledge work transformation. The companies collaborate to combine Google Cloud's technology capabilities with Quantiphi's solution expertise to support customers looking to leverage generative Al for enterprise use cases and keep responsible Al and investment future-proofing at the forefront.

The NVIDIA partnership plays the most integral role in baioniq, as customers of the platform gain access to NVIDIA AI Foundations cloud services, namely the NVIDIA NeMo framework and language model customization service. Quantiphi joined the NVIDIA Partner Network in 2019, is an Elite Service Delivery Partner for NVIDIA DGX AI Compute Systems, and has more than 400 NVIDIA Deep Learning Institute-certified professionals.

In addition to the NeMo framework, Quantiphi brings expertise in other NVIDIA technologies such as the Jetson Edge Al Platform and Omniverse Enterprise to help joint customers create industry-focused Al solutions that work on both the edge and the cloud.

As an AWS Advanced Consulting Partner, named Machine Learning/Artificial Intelligence Partner of the Year in North America in 2022, Quantiphi has achieved competencies, including Machine Learning, Financial Services, Data & Analytics, DevOps, and Migration with over 400+ AWS-certified Professionals.

Quantiphi is also an active contributor to the open-source AI community. The company has a vibrant and active community of technology enthusiasts who participate in some of the most popular open-source solutions such as PyTorch and mxnet, among others.

Quantiphi has partnered with research institutions such as MIT CSAIL to address complex business challenges in a research-oriented environment and accelerate research to commercialization.

Al-first Digital Engineering Services

Quantiphi's Al-first digital engineering services background both drives the company's development of Al platforms such as baioniq and enables its customers' adoption of generative Al solutions. The company uses an industrialized, IP-driven approach to build customized Al solutions using a repository of microagents that perform specialized tasks.

Quantiphi organizes its portfolio of Applied AI services into four "meta-use cases": knowledge and discovery (which includes LLMs and generative AI), automation, experience, and simulation, which correspond with AI-driven capabilities of sight, sound, language, and pattern recognition. These capabilities form the basis of a suite of accelerators aligned by meta-use case and industry verticals. Examples include generative design for product development, generative robotic programming for manufacturing and automotive industries, automated physical notes summarization for healthcare and life sciences, and personalized offer generation for targeted acquisition for banking, financial services and insurance. Qognition.AI, Quantiphi's orchestration platform for MLOps, provides access to the accelerator repository as well as tools to manage the process of assembling, training, validating, and monitoring AI solutions at scale.

In addition to its IP, Quantiphi uses its talent pool of more than 4,000 professionals, including industry analysts, data and platform engineers, and ML engineers, to provide cost-effective AI services to customers in a mix of onsite and offshore delivery. Identifying and aligning the right talent for each project - both engineering and domain expertise - is critical to the success of Quantiphi's engagement model, as is ongoing reskilling and upskilling of talent to meet evolving customer needs, including generative AI. For example, the company's engineers working on conversational AI projects began using transformers upon their inception in 2017, closely followed the research on language models since then, and eventually started leveraging LLMs with the release of GPT-based architectures. Quantiphi's engineering teams also have access to internal notebooks for a range of capabilities from training LLMs from scratch to prompt tuning, P-tuning, task tuning, RLHF, and more.

The company's engagement model takes customers through a four-step framework from ideation (hack it) to pilot (prove it) to production applications (nail it) to a factory-based model for Al solutions (scale it), focusing Al development efforts on use cases that will result in quantifiable business impacts. The framework begins with discovery workshops to educate and brainstorm with business stakeholders around possible Al solutions that could address key priorities and challenges for the

enterprise. Following the prototyping and production phases of AI solution engineering, Quantiphi works with customers to institutionalize AI within their organizations, providing consulting for establishing an AI center of excellence, change management, end-user adoption, and responsible AI.

Responsible AI at Quantiphi

Beyond embedding a responsible AI governance framework into the baioniq platform, Quantiphi also incorporates responsible AI into engineering processes across the ML lifecycle in all AI solution development. The company approach focuses on eight fundamental principles:

- Fairness: Establish diversity in data handling and annotation to mitigate data biases.
- Governance and accountability: Adhere to territory-specific laws and regulations.
- Human centric design: Ensure most impactful decisions are made using human-in-the-loop design principles.
- Safety, robustness, and reliability: Achieve high model performance through continuous learning, monitoring, and feedback loops.
- Scientific rigor: Follow high standards of science-based approaches and collaborate with the scientific community to advance the state of the art.
- Security and privacy: Deploy Al in secure and conducive environments for data collection and storage.
- Socially beneficial: Build AI solutions that deliver long-term societal impact and protect companies and communities from potential threats.
- Transparency and explainability: Enable users to understand logic behind Al-driven decisions.

Quantiphi embeds these principles into its four-step engagement model to help customers to identify and address ethical concerns across the adoption process, enabling customers to create innovative, yet compliant Al solutions.

CHALLENGES AND OPPORTUNITIES

While the market for generative AI solutions is still nascent, few technologies have become so dominant a topic in so many different arenas - consumer, enterprise, academia, public policy, etc. - in such a short period of time. Services firms need to thread the needle of moving quickly to stand out in an intensely competitive field while taking enough time with customers to thoroughly vet opportunities to adopt generative AI solutions in their businesses.

As such, Quantiphi will need to maintain its services-led, consultative approach, even as it positions the baioniq platform as a product. The three-pronged value proposition of platform, partner ecosystem, and Al-first digital engineering services demonstrates the company's intention to do this. For the baioniq platform, Quantiphi has an opportunity to broaden support for prompts and completions lifecycle.

As generative AI adoption grows, organizations will increasingly need help from services firms to address broader business transformation concerns, such as the future of work, reengineering business and IT processes, addressing evolving legal and regulatory compliance requirements, and competing against an influx of new market entrants built around generative AI capabilities. Quantiphi must continue to invest in

scaling its industry-focused consulting teams to engage in these broader conversations with customers, and partner as needed to bring strategy consulting expertise to generative AI engagements.

CONCLUSION

Generative Generative AI promises to make 2023 one of the most exciting years yet for AI. While the buzz today is predominantly about large language and image models, soon we expect more multimodal foundation models to be available. In the near term, we expect generative AI applications to co-exist with existing systems of record and experience. Over time, we expect convergence or a total re-write of digital value engineering. But as with every new technology, business leaders must proceed with their eyes wide open to avoid ethical and practical challenges as noted earlier.

To maximize ROI and accelerate time to value, IDC advises enterprises to focus on the following:

- Create an environment of experimentation for the right/prioritized use cases.
- Assess if the use case necessitates training your own model to win competitive advantage.
- Decide if you want to build or train the model in-house (understanding large budget for computing and need for in-house talent) or partner with a model provider that supports training proprietary models.
- Partner with trusted technology and solution suppliers and service providers, prioritizing those that strive to provide transparency on training datasets and appropriate model usage, or offer their models in open source.
- Prepare for fine-tuning and prompt-tuning skillsets either through hiring, reskilling, or professional services support.
- Develop policies around responsible use of generative AI and inhibit nefarious scenarios.
- Engage in proactive change management to mitigate the impact on the workforce.

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

Copyright Notice

External Publication of IDC Information and Data – Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2023 IDC. Reproduction without written permission is completely forbidden.

