AI and Generative AI Infrastructure Stacks and Deployments

AN IDC CONTINUOUS INTELLIGENCE SERVICE

IDC's *AI and Generative AI Infrastructure Stacks and Deployments* service provides qualitative and quantitative insights on the infrastructure and infrastructure-as-a-service stacks for predictive AI and generative AI (GenAI) workloads. IDC defines an infrastructure stack as an integrated set of hardware and software platforms, systems, and technologies optimized for specific outcomes. IDC defines deployments to include shared and dedicated tenancy, cloud and noncloud deployments, capex and opex procurements, as well as on-premises, collocated, hosted, and cloud services. The service offers analyst perspectives on which infrastructure stacks, deployments, and consumption models are best suited for the myriad of AI and GenAI use cases. Specific focus includes infrastructure needs for AI and GenAI data preparation, AI model training, retraining, fine-tuning, optimization, and AI inferencing.

Markets and Subjects Analyzed

- Commercial and self-built infrastructure stacks for AI and GenAI workloads and use cases
- Compute (processor and accelerator architectures, computing platforms and systems, operating environments, virtualization, and containerization software)
- Storage systems, data persistence mechanisms, organization, access, and connectivity
- Software that enables optimized access to the hardware) as well as the NVIDIA CUDA, OpenCL, and so forth
- Cloud and noncloud approaches for AI and GenAI use cases

Core Research

- Market Taxonomy for Infrastructure and Infrastructure-as-a-Service Stacks for AI and GenAI
- Market Size and Forecast for AI and GenAI Infrastructure and Infrastructure as a Service
- End-User Adoption Trends, Use Cases, and Evolving Application Requirements
- Processor and Coprocessor/Accelerator Trends for AI and GenAI Use Cases
- Commercial and Open Source File, Object, and Block Scale-Up/Scale-Out Platforms and Systems

In addition to the insight provided in this service, IDC may conduct research on specific topics or emerging market segments via research offerings that require additional IDC funding and client investment. To learn more about the analysts and published research, please visit: AI and Generative AI Infrastructure Stacks and Deployments.

Key Questions Answered

1. How big is the infrastructure and infrastructure-as-a-service market for AI and GenAI workloads?
2. What are the infrastructure hardware and software requirements imposed by AI and GenAI workloads?
3. What are some of the data life-cycle challenges associated with AI and GenAI workloads?
4. What are the optimal compute and storage configurations for AI and GenAI workloads?
5. What is the role of accelerated computing (GPUs, FPGAs, ASICs, manycore processors, and emerging acceleration technologies), NVMe, tiering, de-duplication, and compression as they are related to AI and GenAI workloads?

Companies Analyzed

This service reviews the strategies, market positioning, and future direction of providers in the PIC market, including:

Amazon Web Services, AMD, Alibaba, Baidu, Broadcom, Cisco, DDN, Dell, Facebook, Google, Hewlett Packard Enterprise, Hortonworks, Huawei, IBM, Inspur, Intel, Juniper, Lenovo, Microsoft, NetApp, NVIDIA, Oracle, Penguin Computing, Pure Storage, SambaNova Systems, SAP, SAS, Supermicro, Symantec, and Tencent.